National College of Ireland

BSc (Honours) Data Science

BSHDS4

2022/2023

Niamh Daly

X19370553

X19370553@student.ncirl.ie

Racing Against the Elements:

Analysing the Impact of Weather on Formula One Races

Data Science Report

Contents

Executive Summary

Formula One is a data driven sport, where data is continuously being gathered and processed to facilitate fast paced decision making to get the best results. A challenge that Formula One has faced in recent years has been adverse weather conditions causing stoppages to the Formula One races. These stoppages occurring in the form of red flags, safety cars, and virtual safety cars. In some cases, the full race distance was not being completed due to these stoppages, and instead races being completed under a countdown clock as opposed to a lap counter (Emms, 2022). A recent example of this, and the inspiration for the project, is the 2021 Belgian Grand Prix, whereby the race was scheduled to run 44 laps, and ended up being officially concluded after 3 laps due to the torrential wet weather conditions at the circuit (Formula1, 2021). In its wake the event left a heap of disappointed fans and a question as to the impact the weather has on Formula One. It would seem that the weather is having a greater impact on the sport and its ability to perform to the highest standard in these conditions.

This analysis attempts to determine the impact weather has had on Formula One as a sport, utilising historical Formula One and weather statistics, sourced and compiled from a number of resources solely for the purpose of this study.

The Formula One data was sourced from the Ergast Developer API (Ergast Developer API, n.d.), Formula One Wiki (Formula 1 Wiki, n.d.), and Motorsport Stats (Motorsport Stats, n.d.). This data included: lap times, race locations, scheduled race start times, safety car deployment, virtual safety car deployment, red flag data, and many more. The weather data was sourced from Visual Crossing (Visual Crossing, n.d.), which provided detailed historical weather data (precipitation, temperature, windspeed, visibility, cloud cover, dew, etc.). From the available data, the chosen period for the analysis covered the 2005 season until the end of the 2022 season, which included 348 races, at 37 racetracks, across 29 countries.

1.0 Introduction

1.1. Background

Formula One, also known interchangeably as F1, and Formula 1, is said to be the pinnacle of motorsport, known for its high-speed action, cutting-edge technology, and high-profile drivers (Wikipedia, 2022). Formula One is a type of open-wheel car racing that originated in Europe in the 1950s and is now world renowned and increasingly more popular due to the success of the Netflix series 'Drive to Survive' (Richards, 2022). Formula One is a sport that travels around the world. A combination of both permanent racetracks and street circuits serve as venues for Formula One motor races.

F1 has evolved significantly over the years, with new rules and regulations being introduced to improve safety and increase competitiveness amongst the cars. Formula One is a high-speed and high-risk sport but safety in the sport is a top priority for the teams, drivers, and the organizers alike. Over the 73 years since the sport began, there have been significant improvements in the cars, and in the running of the sport in the efforts to make the sport safer (MotorsportWeek, 2022). There are several safety measures in place to be used by the drivers and on the Formula One cars, some of these include:

- Drivers are required to wear fire-resistant materials in their driver suits, gloves, and shoes to protect in the event of a fire (McLaren, n.d.).
- Headrests and neck supports in the cars protect the drivers head and neck in the event of a crash (MotorsportWeek, 2022).
- The cars are required to have a 'Halo', which was initially a controversial addition to the aesthetics of the F1 cars when it was first introduced in 2018, and caused many a debate amongst fans. The Halo is a curved bar made from titanium (a strong lightweight material designed to absorb the energy of an impact) that sits above the drivers head and is attached to the cars chassis. Most drivers and teams have supported the Halo, and it has been credited with saving lives in a number of high-speed accidents since its introduction to the sport (MotorsportWeek, 2022).

In Formula One, safety flags and the safety car are used to help ensure the safety of drivers and track marshals during a race. There are a number of different types of safety flags that are used in Formula One, each with different meanings (Pretorius, n.d.):

For example:

Green Flags: Used to indicate that the track is clear, and that normal racing conditions apply. Green flags are waved at the start of each race after the lights have gone. These flags signify that the race has started and that the drivers are free to race one another and drive without any concern for hazards on track. This flag is usually shown after a period of caution (e.g., after yellow flag) to signal the track is safe to continue racing (Pretorius, n.d.).

Yellow Flags: Yellow flags indicate that there is a potential hazard on the track, such as a car having gone off the track (Pretorius, n.d.). Drivers are required to slow down and reduce their speed and are not allowed to overtake.

Red Flags: These flags indicate that the race or session has been immediately stopped due to a serious incident or the conditions are unsafe on the track and drivers must return to the pitlane (e.g., a lot of debris after an accidents, or large amount of water and the visibility is too poor for racing) (Pretorius, n.d.)

Blue Flags: Blue flags are shown to drivers who are being lapped by another driver (Pretorius, n.d.). It informs the driver that there is a faster driver approaching and that the slower driver who is on a different lap should allow the faster driver to pass.

The safety car and virtual safety cars in Formula One are safety measures that help to ensure that drivers and track marshals can remain safe during a race. The safety car is a specifically equipped vehicle that is deployed and used to lead the Formula One cars around the track at a reduced speed when there is a potential hazard on the track (Wikipedia, 2022). When a safety car is deployed during the race, the race is effectively paused, and the drivers are not allowed to pass each other until the safety car returns to the pits and racing is resumed.

Virtual safety cars (VSC) are a software system that is used to control the speed of cars on the track in the event of a caution period, such as when a car goes off the track or when there is debris out on the track, and it needs to be cleared (Westbrook, 2018). When the VSC is activated, the cars are required to slow down to a predetermined speed and maintain a certain distance between each other. The VSC was introduced in Formula One in 2015 as part of a series of safety improvements that were made in response to several high-profile accidents in the sport (Westbrook, 2018).

Weather can have a significant effect on Formula One races. Different weather conditions can affect the grip and handling of the car, as well as visibility for drivers as mentioned above. Rain would appear to be one of the most significant weather conditions that can affect Formula One. When it rains, regardless of the amount, the track becomes wet, which can reduce the grip and increase the risk of aquaplaning. Drivers need to adjust their driving styles and be more cautious during wet weather conditions. Rain can also affect the visibility for the drivers. Making it harder for the drivers to see the drivers around them and increase the risks of accidents or collisions. Overall, weather can have a significant effect on Formula One races, and teams and drivers must be prepared to adapt to changing conditions in order to be competitive.

The inspiration for the project is the 2021 Belgian Grand Prix, whereby the race was scheduled to run 44 laps, and ended up being officially concluded after 3 laps due to the torrential wet weather conditions at the circuit (Formula1, 2021). In its wake the event left a heap of disappointed fans and a question as to the impact the weather has on Formula. It would seem that the weather is having a greater impact on the sport and its ability to perform to the highest standard in these conditions.

1.2. Aims

The aim of the project is to complete an analysis into the impact that weather has on Formula One races. The goal is to compile a dataset to be used for the analysis by obtaining historical weather and Formula One race data. The data will contain information from different sources on Formula One races and their respective weather conditions. Using different visualisation tools and technologies this data will analyse the impact of weather on:

1. Lap times
2. Safety car usage, cause, and duration
3. Red flag usage, cause, and duration
4. Race length completed (e.g., race not completing the total distance/ number of laps shortened)

Using methods to determine which attributes are the best suited to analysing the impact that weather has on Formula One races.

The project aims to create an interactive dashboard to analyse the weather and race statistics for specific races and circuits.

Use machine learning models to attempt to predict the relationship between the Weather and Formula One races.

1.3. Technology

Several different tools and technologies will be used to carry out what has been set out in the aims of the project. Some of the technologies and tools that are to be used for the project are newly learned for the purpose of carrying out the project.

Conducting research before starting out on the project allowed the tools to be properly utilized and learned where appropriate to complete the task. In the State of the Art in 2.0 below, the available tools and technologies that had been utilised in other analyses were noted within the tables. The following tools and technologies were utilized in the implementation of this project: Python, Jupyter Notebooks, Pandas, Numpy, Scikitlearn, Seaborn, Matplotlib, Dash, Excel, and Visme.

1.3.1. Python

Python is the programming language that was taught and used on the BSc Data Science course; therefore, it was an obvious choice for programming language for this data science project. Through the research carried out on other projects, it was also discovered that Python is a popular programming language among data scientists for data analysis tasks. Python is known for its simple straightforward syntax, which makes it easier to learn and understand (Yıldırım, 2022). Python will be utilized for carrying out most of the main aspects of the project, including compiling, cleaning and creating models for the analysis. The following libraries will be used with Python for carrying out the analysis: Pandas,

Figure 1: Python Logo (Python Software Foundation, 2023)

Numpy, Scikitlearn, Seaborn, and Matplotlib. The web application framework Dash will be used in conjunction with Python as outlined below.

1.3.2. Jupyter Notebooks

Jupyter Notebooks is an open-source web-based interactive development environment (IDE) that allows users to create notebooks that contain live code, equations, and visualisations with explanatory text to accompany the code (Driscoll, n.d.). Jupyter Notebooks are to be used for most of the coding scripts produced from the project that involved Python. The only exception being the dashboard which will be completed in a '.py' and not a '.ipynb' file. Jupyter notebook allows for easy formatting and commenting of code, with the markdown cells being utilized for separating blocks of code for specific functions. Jupyter Notebooks allows for producing easily reproducible code, with table, graphs and results print outs being presented underneath the specific block of code.

Figure 2: Jupyter Notebook Logo (Jupyter, 2023)

1.3.3. Pandas

Pandas is a powerful and widely used Python library that is to be used for the project. Pandas can be used for reading the csv files into the python environment as pandas dataframes in preparation for manipulation and for working with the data for the project. Some of the features used from the Pandas library include: creating the dataframes from the imported csv files, converting columns to their specified format (timedelta, Timestamp, datetime), merging dataframes together to get the complete dataset that was desired for the analysis (left merge, merge, concat), for identifying missing values (isna), and grouping the dataframes to filterspecific data, to name a few of the features made pissible with Pandas in Python. Pandas is designed to make it easy to work with complex datasets, providing a range of tools for cleaning, transforming and manipulating data (Mitchell, 2023).

Figure 3: Pandas Logo (Pandas, 2023)

As the data that will be used for the analysis is complex and is being created from a number of different sources, it was important to select a library that would most effectively be able to deal with large amounts of data. Pandas is a library that will work well with the other tools and technologies that will be used to perform the analyses. The ability to handle DataFrames is vital for the project and is an important feature of the pandas library. Using Pandas to drop rows, select rows and create new columns are all important computations that are made simpler with Pandas.

1.3.4. NumPy

NumPy is a Python library for scientific computing that provides support for large, multidimensional arrays and matrices of numerical data, as well as a large collection of mathematical functions to operate on these arrays (NumPy, n.d.). NumPy provides a wide range of mathematical functions that can be used to perform operations on ndarrays, such as linear algebra operations,

Figure 4: NumPy logo (NumPy, 2023)

statistical functions, and signal processing functions (NumPy, n.d.). NumPy is designed to work with other scientific Python libraries, such as SciPy and Pandas, making it easy to use NumPy arrays in a variety of applications. Numpy is a tool that will be useful for statistical analysis of the datasets and will be utilized for carrying out these analyses on the chosen data for the project. Numpy will be used in the project to concatenate the data from the different sources into one data frame.

1.3.5. Scikit-learn

Scikit-learn is a Python library for machine learning that provides a range of tools for tasks such as classification, regression, clustering, and dimensionality reduction (Scikit-learn, n.d.). It is built on top of NumPy and SciPy, two popular scientific computing libraries, and is designed to be easy to use and efficient. Scikit-learn provides a range of pre-processing tools for preparing data for machine learning, such as functions for scaling, imputing missing values, and encoding categorical variables (Scikit-learn, n.d.). Scikit-learn provides functions for evaluating the performance of machine learning models, as well as functions for selecting the best model based on a given dataset. Scikit-learn works with other scientific Python libraries, such as NumPy and Pandas. Scikit learn will be used for the machine learning tasks in the project, as the aims for the project require machine learning tasks to be carried out.

Figure 5: Scikit learn logo (Wikimedia Commons, 2020)

1.3.6. Seaborn

Seaborn is a Python data visualization library that is built on top of Matplotlib. It is designed to make it easier to create attractive and informative statistical graphics in Python (seaborn, n.d.). Seaborn has pre-designed styles, that make it easier to create plots that are professional looking and are easy to read. Seaborn supports a wide variety of plot types, including line plots, scatter plots, bar plots, error bars, and many others. It also provides functions for creating complex plots, such as heatmaps, violin plots, and cluster maps (seaborn, n.d.). Seaborn provides functions specifically designed for plotting statistical data, such as functions for visualizing the distribution of a dataset, comparing the relationships between different variables, and comparing the performance of different models. Seaborn is another Python library that will be used for data visualisations. Each of the Python visualisation libraries will be utilized for different types of visualisations. The requirements of each of the visualisation tools will become more apparent as the project progresses.

Figure 6: Seaborn Logo (seaborn, n.d.)

1.3.7. Matplotlib

Matplotlib is a Python library for creating static, animated, and interactive visualizations in Python. It is one of the most widely used plotting libraries in Python (Matplotlib, n.d.). Matplotlib has a wide range of plot types including line plots, scatter plots, bar plots, error bars, and many others. Users are also able to

Figure 7: Matplotlib Logo (Matplotlib, n.d.)

create custom plot types if needed with the ability to also customize existing plot types, including being able to control the colours, line styles, and markers of the plots. Matplotlib supports interactive plots, allowing users to zoom, pan, and hover over data points to see more information (Matplotlib, n.d.). These interactive plots can be create using Matplotlib's built-in interactive backends, or they can be embedded in web pages using tools like Plotly. Matplotlib's library will be utilized for creating several visualisations throughout the project and analysis.

1.3.8. Plotly Dash

Dash was the chosen application for the dashboarding capabilities of the project. Plotly Dash is a Python library for building interactive web-based analytics applications. It is built on top of the Plotly.js library, which is a popular JavaScript library for creating interactive charts and plots and is designed to be easy to use and integrate into existing web applications (Tomar, 2021). Plotly Dash provides a range of tools for building dashboard-style applications, which allow users to display multiple plots and charts on the same page and interact with them in real-time. Plotly Dash is a powerful and widely used library for building interactive analytics applications in Python.

Figure 8: Plotly Dash Logo (plotly, n.d.)

1.3.9. Excel

The data was to be compiled from several different sources, some of which were available for download as comma separated value (csv) files which were to be opened with Excel to complete an initial investigation into what was contained within the datasets. Other data was taken from several websites, so they were to be manually input into Excel workbooks and converted into tabular format and the files converted into csv files. Some of the files were going to be manually merged in Excel with the use of VLOOKUPS and pivot tables to assist in corroborating the correct data for use in the analysis.

Figure 9: Excel Logo (Wikimedia Commons, 2023)

1.3.10. Visme

Visme was to be used in the project for the purpose of creating the project poster. The project poster was to provide a summary of the analysis and the aims of the project which was then going to be displayed during the Project Showcase. A subscription will be obtained for the service in order to obtain more advanced features from the service, such as downloading the poster in a higher quality.

Figure 10: Visme Logo (StickPNG, n.d.)

1.4. Structure

The report is structured using the Data Analysis Report Template provided for the Data Science Project. Aspects of the project have been addressed within each of the section headings, with some overlap occurring within 3.0 Data, 4.0 Methodology, 5.0 Analysis, and 6.0 Results. An overview of what is addressed in each of the section has been provided in Table 1 below.

Table 1: Structure of the Report and Headings

Section	**What addressed?**
1.0 Introduction	The introduction provides insight into beginning stages of the project. It provides details on why the project was selected (Background), and the goals of the project upon completion (Aims) Chosen technologies are addressed in this section and their capabilities and the approach to using them are outlined (Technology)
2.0 State of the Art	The research conducted into the project domain is outlined and similar analysis, if any, is addressed. The unique aspects of the project are identified, and the data science methods and technology used in the researched works are described in tabular format.
3.0 Data	The requirements of the datasets for the project are detailed. Sources for the datasets were identified, and the Ethical considerations for the data were discussed. The compilation of each of the datasets are addressed in detail, with summary tables of the original sources of the data and then the final compiled datasets being produced. Exploratory analysis was detailed and addressed, and a description of the dashboard and visualisations are provided with reference to the insights gained from the data compilation.
4.0 Methodology	The KDD methodology is outlined in detail using the 7 KDD step process: 1. ***Selection-*** References the compilation of the datasets, and a description of the final dataset is outlined 2. ***Pre-processing-*** Additional steps are outlined to clean the data and remove noise 3. ***Transformation-*** Steps detailed for creating new features from the data, and formatting of the data for analysis 4. ***Data Mining-*** Descriptions of the models for the project are outlined 5. ***Interpretation-*** Reference to analysing the results to extract knowledge, explored in more detail in the *Results* section of the report 6. ***Evaluation-*** A description of the evaluation method outlined for the models 7. ***Deployment-*** Reference to the project poster for the summary of project being a deployment method
5.0 Analysis	The models from the Methodology section are outlined, and the steps carried out to produce the models are addressed. The decisions made for the modelling are outlined and the State of the Art research is referenced as the inspiration for the chosen approaches.
6.0 Results	Results are presented using tables and visualisation and the outcomes described.

7.0 Conclusions	The project outcome and results are addressed, with reference to the advantages and disadvantages of certain aspects of the project. The strengths and limitations of the whole are addressed.
8.0 Further Development or Research	The potential direction for the project I outlined, with reference to additional insights that could be gained from the project with further development.
9.0 References	The references for the project are detailed in APA referencing style.
10.0 Appendices	Additional resources and appendices are attached, including the proposa and reflective journals.

2.0 State of the Art

Formula One is a competitive sport that like most, has its diehard fans, and with that leads to much discussion around which Formula One Driver, or team is the best in the sport. Because of this, there is evidence of these types of analysis carried out in an attempt to find the best driver or team in F1 (F1 Analysis, n.d.) (Bell, Smith, Sabel, & Jones, 2016). Another popular topic for Formula One analysis, is building predictions for F1 race results (Kuosmanen, 2020) (Stoppels, 2017), championship results (George, 2021) (Sicoie, 2022), and pit stop strategies (Łusiak, 2021) (Stavropoulos, 2018).

In most cases, there is a limited amount of weather data included in conjunction with all of the other F1 data that is being used in the analysis. For example, some reports refer to wanting to include more detailed weather data in their research so that the weather data will have more of a weight on the predictions (George, 2021). Analysing the impact that weather has on Formula One races is not a topic that has been completed before in a research study. The topic has been addressed in numerous videos, articles, and blog posts where the question is asked, how does the weather impact F1 (Collantine, 2022)

The blog posts and articles are topical when a wet Formula One race occurs, most notably a number of write ups were conducted after the wet Belgian grand prix in 2021 and speak on the lack of visibility for the driver and the results of the race having to be stopped because of the rain (Chiu, 2021). There is discussion about the technology f1 has for predicting the weather (Formula One, 2016), and what a wet race means in terms of tyres selections (Media, 2022) (Millward, 2022), visibility (Horton, F1 Officials Determined to Find Safer Ways of Racing in the Rain, 2021), and the cars (MetMatters, 2011). The technology is able to judge where and when rain is approaching specific parts of the track, for passing on the information onto the driver. Predictions on the weather during the race are so precise that they are able to predict a number of variables for the teams during the races, from wind strength and direction to the more vital information of when the first drops of rain will be falling and where it will hit first (Masefield, 2013). Talk has also occurred around the topic of shifting race start times due to the weather, in doing so being able to avoid poor weather forecasts (Khorounzhiy, 2021).

It is also mentioned that the cars in future years are likely to have a smaller profile to produce different spray patters (Horton, F1 Officials Determined to Find Safer Ways of Racing in the Rain, 2021). Which could reduce the risk of races being delayed or called off due to persistent rain striking a Grand Prix race. It was said that the Belgian Grand Prix the track conditions were not completely undriveable however it was the spray that was being produced by the cars that reduced visibility for following cars, and making the conditions extremely dangerous with little visibility, with mention of f1 drivers not being able to see the red light that is visible on the rear of f1 cars during periods of slow driving or when conditions are bad (Cooper, 2021).

Some analysis has been performed into the Formula One calendar, as the Net Zero Carbon goals in F1 2030 has also been introduced, to reduce the carbon emissions produced by freight and travel, by looking at a potential to regionalise the calendar (Chalmers, 2008) (Formula

One, 2022). Creating the ideal calendar by looking at the location of racetracks and reducing travel between circuits.

When looking at other research and analysis that has been completed before it was important to not limit the research to solely Formula One, and to investigate if studies have been completed in other sports events.

Some analysis was completed to carry out a statistical analysis into the advantages and disadvantages of weather on sporting events, with a number of sporting events investigated (Thornes, 1977). Reference is made to Formula One being very sensitive to weather, however it is not included in the statistical analysis as the definition of sport derived for the analysis. Sport was described as involving the human body solely and not taking energy from elsewhere, in the case of F1, the energy is being taken from the car. The study classified the effect of weather on sport in two categories: the effect on the sport itself and the effect on the spectator comfort.

From other research there was evidence of analysis being completed into the revenue from sporting events being influenced by the weather at major sporting events (Dawkins & Stern, n.d.). The research conducted analysed the relationship between rain, wind and temperature with spectator buying patterns with merchandise for example a higher relationship was observed between sales of wind breakers when the temperature and the sunshine hours were low. A loss in revenue at sporting events was attributed to the weather either being too bad, in the case of torrential rain, and in the case that the weather is too good, with really high temperatures preventing spectators from wanting to spend a lot of time outside (Dawkins & Stern, n.d.).

In the case of competing in sporting events, some analysis was computed on the intention to revisit a sporting event, using a number of hypothetical factors and scenarios that included weather conditions, travel time and distance, and also the cost of travelling to the destination (Whitehead & Wicker, 2020). Performance at outdoor endurance sporting events was seen to be affected by temperature and rainfall, with evidence clearly showing that the performance of athletes suffers in warm conditions as it does in rainy conditions.

An analysis was conducted on Boston marathon performance based on the weather conditions, together with sex, country of origin, with the running performance (Knechtle, et al., 2019). The weather variables that were included in the analysis were: Weather conditions, over the hours of the event, air temperature, total precipitations, wet-bulb globe temperature, wind speed, wind direction, and barometric pressure. The effect of the weather on performance was analysed using regression models for subgroups of runners.

In the State of the Art research there were a number of sources used for inspiration and research, some of the sources have been mentioned above, with full details of the research papers, videos, and blog posts being presented in Table 2, Table 3 and Table 4 respectively.

From the State of the Art analysis it was discovered that the chosen analysis has not been previously completed before; making it stand out from the existing works that are available.

Table 2: Summary of State of the Art Analysis Conducted: Research Papers

What the work is about	Similarity/ difference with my own work	Type	Data Science Methods	Technology	Ref.
The inter-relations between weather and sport. Financial impact of weather awareness in sport management	Looks into sport and the effect of weather on the event.	Research Paper	statistics, visualisations, mathematics	research	(Thornes, 1977)
The effect of training satisfaction and weather on revisit intention. Monetary values of weather conditions and training satisfaction. Willingness to travel converted into monetary values using travel costs.	Conducts surveys to attempt to gather data on hypothetical scenarios for return visitation that randomly assigned different travel cost per mile, travel distances, weather forecasts, and respondents' training satisfaction	Research Paper	statistics, visualisations, modelling	surveys	(Whitehead & Wicker, 2020)
The relationship of weather conditions, together with sex and country of origin, with running performance in the Boston Marathon from 1972 to 2018	investigate the role of weather conditions, together with sex and country, on female and male performance	Research Paper	statistics, visualisations		(Knechtle, et al., 2019)
Revenue risk increase due to the uncertainty and the unpredictability of the 'right' kind of weather.	Concluding that there are relationships between the sales of various products and various weather parameters with the Australian Open.	Research Paper	statistics, visualisations, mathematics		(Dawkins & Stern, n.d.)
Case highlights the issues associated with the F1 Australian grand prix. highlights the problems with economic impact studies and the need to focus on the triple bottom line approach by examining the economic, social, and environmental issues associated with the event	The study investigates the issues with the event that receives substantial government funding and therefore the worth of the event receives consistent public scrutiny	Research Paper	statistics, visualisations		(Fairley, Tyler, Kellett, & D'Elia, 2011)
examined the new money generated from	F1 event appears to influence on sports-	Research Paper	statistics, visualisations		(Min Kil, et al., 2016)

Formula One Grand Prix (F1) and the economic impacts of this new money on the host economy using input–output analysis	related industry as well as other industries such as manufacturing industry				
random-coefficient models and (a) finds rankings of who are the best formula 1 (F1) drivers of all time, conditional on team performance; (b) quantifies how much teams and drivers matter; and (c) quantifies how team and driver effects vary over time and under different racing conditions	effects are then allowed to vary by year, track type and weather conditions using complex variance functions	Research Paper	statistics, visualisations, modelling		(Bell, Smith, Sabel, & Jones, 2016)

Table 3: Summary of State of the Art Analysis Conducted: Videos

What the work is about	Similarity/ difference with my own work	Type	Data Science Methods	Technology	Ref.
Using Python to perform some exploratory data analysis and draw insights from the data using charts and visualisations	Using datasets and libraries such as NumPy, pandas, Matplotlib, and seaborn, to answer questions related to Formula 1	Video	statistics, visualisations, modelling	Python	(Simplilearn, 2022)
Talks about the different tyre compounds and the conditions they are raced on.	Safety issues with racing under wet weather conditions	Video	statistics		(Millward, 2022)

Table 4: Summary of State of the Art Analysis Conducted: Blog Posts

What the work is about	Similarity/ difference with my own work	Type	Data Science Methods	Technology	Ref.
Speaks on the 2021 Belgian Grand Prix. Speaking on the event being seen as a 'wasted afternoon'. F1 cars dependent on aerodynamics, with tyres being sensitive to conditions.	*Laps under the safety car, drivers had reduced visibility standing water and rain on track. Visibility described as one of the main issues.*	*Blog Post*	*statistics*		(Chiu, 2021)
The Japanese Grand Prix has been red-flagged and suspended after just two laps of running as heavy rain fell	*The rain began falling long before the race got underway. It built steadily in the run-up to the start.*	*Blog Post*	*statistics*		(Collantine, 2022)

	Safety car was deployed for car recovery, rain intensifying, red flagged the race.				
Track temperature effect on tyres, cloud cover, daytime or night-time, colour of the tarmac, track temperature directly effects tyres.	*Car performance effected by the weather/ temperature. Too hot, too cold, needs to be just right for the optimal results*	*Blog Post*	*statistics*		(Media, 2022)
In recent years, Formula 1 has had seemingly more and more rained-out Fridays and Saturdays at events due to torrential rain. In 2021, Formula managed an entire event that consisted of just three laps behind the Safety Car on race Sunday at Spa.	*The two big issues in wet races are visibility and aquaplaning, and F1 officials believe that visibility is an area that can be improved*	*Blog Post*	*statistics*		(Horton, F1 Officials Determined to Find Safer Ways of Racing in the Rain, 2021) (Horton, 2021)
Interactive visualisations from f1 data, looking into lap times, evolution of f1 cars, f1 drivers, speed of f1 cars, logistics of transporting goods to races	*Working with several visualisations with the F1 data, some are interactive in nature and other are infographic type visualisations*	*Blog Post*	*statistics, visualisations*		(Wang M. , 2022)
Expressing a F1 dataset as an interactive visualisation. "Who is the GOAT in F1", a question deliberately selected for its subjective nature	*Another example of an interactive visualisation displayed as an information visualisation web app.*	*Blog Post*	*statistics, visualisations, modelling*		(Paul, n.d.)
A website for f1 visualisations	*Blog/ website for data presentation*	*Blog Post*	*statistics, visualisations, modelling*		(Formula 1 Trends, n.d.)
To explore the relational data model of Formula 1 historical data by visualization in QlikView	*Exploring F1 data in QlikView, an analytics tool developed by Qlik*	*Blog Post*	*statistics, visualisations, modelling*	*QlikView*	(Wang J. , 2020)
overview of the 2021 season using a series of animated visualizations. Lightning Plots, and other visualisations	*Interactive website created to display the analysis of the formula one 2021 season data*	*Blog Post*	*statistics, visualisations*	*AnyChart*	(AnyChart Team, 2022)
Malaysia is no stranger to spectacular downpours - the race has been red-flagged twice in the last eight	*Looking at how F1 teams, race engineers, strategists, weather watchers, and the*	*Blog Post*	*statistics*		(Formula One, 2016)

years because of torrential rain. how do the teams stay on top of what the weather is doing, not just in Sepang but at different circuits around the world?	*driver deal with deciding on the best course of action for the weather*				
weather has an impact on many sports, but there are very few sports where the stakes are as high as they are in Formula One	*Looking at different factors that effect F1 during weather conditions, tyres, rain, heat, and wind*	*Blog Post*	*statistics*		(MetMatters, 2011)
Speaks on Belgian grand prix nearly always having rain due to its own microclimate at the time of the year the race is held	*Teams are accurately able to judge when a bank of rain is approaching in the pits or on the pit wall to within a matter of minutes and relay that information to the driver over the radio*	*Blog Post*	*statistics*		(Masefield, 2013)
Formula 1 race director speaking on whether it can have the flexibility to bring a grand prix start time earlier into the weekend in case of a poor weather forecast	*Looking into the possibility of changing race weekend start times based on the weather*	*Blog Post*	*statistics*		(Khorounzhiy, 2021)
Development of a 100% sustainable fuel, slashing the use of single-use plastics and reviewing travel and freight logistics – these are just some of the things Formula 1 as a sport is working on as part of its commitment to be Net Zero Carbon by 2030…	*Looking ahead, there are plans to build future F1 calendars to improve freight and travel logistics, so the sport is moving more efficiently around the world.*	*Blog Post*	*statistics*		(Formula One, 2022)
A machine learning approach to predict the winner of the next F1 Grand Prix	*A study into predicting race winners, data collection, analysis and models*	*Blog Post*	*statistics, visualisations, modelling, machine learning*	*Python*	(Nigro, 2020)
There are a number of different factors involved in creating the perfect F1 calendar, because there are so many people to satisfy	*Mentions races ideally being held in the same region to be grouped together for teams to save money and not so many miles are clocked up.*	*Blog Post*	*statistics, visualisation*	*research*	(Chalmers, 2008)

3.0 Data

The dataset used for the project was compiled from a number of different sources in order to carry out the analysis. The research in formula one has not been completed before, and the data for the analysis was therefore not available for download as a single dataset. The data sourcing and compiling is described in detail in the relevant sections below:

3.1. Data Required for the Analysis

The purpose of the project is to analyse the impact of weather on Formula 1 races, so for that there were two main types of data that needed to be collected: Formula 1 race data, and weather data for the race.

As the project was looking at race data specifically it was important to understand what Formula One data was sought after, and what data was available to be used for the project. After research conducted for the state of the art analysis it was evident that there was no complete dataset readily available that would fit the requirements of the project. Therefore, efforts needed to be made to source the correct data that could be used in order to form a dataset that would be suitable for the project.

3.1.1. Formula One Race Data Required

The race data that was required for the project was for data that could be attributed to a race, and aid in understanding the impact of weather on the race. The following types of data were noted so that work could begin on sourcing the data. The formula one data that was needed can be attributed into the three main categories:

- **Location Data:** A Formula 1 race is held at a track, as already mentioned above these tracks can take many forms, being permanent racetracks, street circuits, or temporary tracks. Each of these tracks can be in different cities, countries and continents. Getting location data that could pinpoint the exact location of these tracks would give insight and details that could be used for sourcing the weather data. The type of location information that is needed for the weather data will differ depending on the service that is being used to obtain the weather data, as some services can use more general location such as cities or countries, however some services provide more detail and can give weather data using longitude and latitude values. To ensure that all weather services were considered at this stage it was decided that the some of the following data should be prioritised when getting Formula One data: Name of the circuit for identification, the longitude and latitude coordinates of the circuit, the city, and country of the circuit.

- **Date and Time Data:** Formula One races are mostly held on Sunday, with some exceptions in the past. In order to be able to link weather to a specific race, the date and time of when a race was held needs to be known. If there was no record of the time a race was scheduled, then the only appropriate weather data that could be used is daily average weather statistics. This data would then not be able to show a true investigation into the impact weather has on F1, as the weather averages for a day could be skewed if the weather was unfavourable for a period

during the race, and thus not reflected in the daily averages that would be obtained. Data that can link the weather data back to the time the race was held at was a necessity. Time data that was needed for the race not only included the start time of the race in local time of the region, but also included the year the race was held, the round number that the race was in the calendar season, and the date that the race was held on. All of this data would be useful for obtaining the correct weather data.

- **Race Statistics:** The race data that was required for the analysis was specific to the aim of the project. The aim is to analyse the impact of weather on Formula One races, so therefore data was needed for the happenings during a Formula One race. This data would include a selection of fastest lap times during the race, data on the safety measures used during a race, i.e., the number of safety cars deployed, virtual safety cars deployed, and red flags used. The completed race distance would also be included in these race statistics, meaning the number of laps completed, as from the inspiration the Belgian Grand Prix only ran 3 laps out of the scheduled 44 laps. For the analysis however, driver specific details are not required i.e., which driver won the race, or who completed the fastest lap. These elements are more predictable and expected elements during a race, as a driver will finish first and a driver will get the fastest lap. The focus of the study is on the impact of weather on the more uncertain or unpredictable elements of a F1 race, such as safety car periods, red flags, crashes etc, as these can impact the outcome of the race.
  As the yellow, green, and blue flags are in most cases used as a means of warning the drivers of potential hazards or changed to the race conditions. These flags can cause a short disturbance but do not typically result in a complete stoppage of the race. Red flags however indicate immediate stoppages to the race, which can be due to serious accidents, hazardous track conditions, or other significant events that require the race to be halted for safety reasons. As the project aims to analyse the impact of weather on races then the data related to the red flags would be the only necessary flag data to be sourced.

3.1.2. Weather Data Required

A lot of the Formula One data for the location, and the date and time data are needed because it will aid in obtaining the correct weather data from the appropriate sources.

The most important factor to consider when obtaining the weather data was ensuring that historical data was obtainable. The analysis will be looking at races that have already happened in previous years, so historical weather data was crucial for the ability to complete the project. The weather data will be collected as a separate source from the Formula One data and therefore there needs to be data or identifiers available within the data that can link the weather data back to the Formula One data.

The State of the Art analysis included analyses that had used weather data in marathon performance in competitors (Knechtle, et al., 2019), so this study helped identify some of the attributes that would be useful for a study of the weather. Other studies included in the State of the Art analysis included comments on their future works, stating that they wanted more detailed weather data in order to improve their study (George, 2021).

It was important then to make sure enough efforts were spent on sourcing the correct weather data that would aid in the completion of the project.

- **Location Specific:** Historical weather data that could be collected from a number of specific locations was necessary for the project as the Formula One races take place in a number of locations around the world. Ideally the location data could be collected using longitude and latitude coordinates as that was hopefully to be obtained from the Formula One Data and would provide the most accurate results for the location specific weather data.

- **Date and Time Specific:** Another aspect of the weather data that is needed for the project is date and time data specific to the race locations. The races data will be collected for a number of years, so the weather data needed to be historical data that spanned far enough back into the Formula One seasons. Ideally the weather data would be obtained for hourly increments as that would be more accurate weather data for the races. A race would typically be held over a period of a few hours and not for the whole day, so more detailed and accurate weather statistics would provide a clearer picture into the impact of the weather on the F1 races. As mentioned above, other analysis highlighted that hourly data would have elevated their study (George, 2021). If daily statistics were collected the analysis would be working on averages of weather conditions for a whole day.

- **Weather Statistics:** The weather statistics needed for the project would be specific to what is available from the sources that are can provide location and hourly specific weather data. As the inspiration for the project is the 2021 Belgian Grand Prix, where torrential rain caused the race to be stopped, then statistics on precipitation types and levels would be appropriate to be collected for the project. From the State of the Art analysis, other important weather statistics that would be useful to be included are: temperature, precipitation, wind speed, wind direction, humidity, dew point, cloud cover, solar radiation, etc, as mentioned in a number of studies (Knechtle, et al., 2019) (Thornes, 1977) (Whitehead & Wicker, 2020) (Dawkins & Stern, n.d.).

3.2. Sourcing the Data

Equipped with the ideal requirements for the data and after conducting research into a number of different sources, the following sources were selected as being appropriate for getting the relevant data compiled for the analysis. The Formula One data was to be compiled from three different sources while the weather data was obtained from one

source. Several options were investigated for the weather data but only one was selected as appropriate for the analysis. Some of the available options for weather data were limited in the number of years of historical data years that were available and did not have hourly weather data but had daily averages, making them unsuitable for the project as per the requirements.

The sources for the data are:

- Ergast Developer API
- Formula 1 Wiki
- Motorsport Stats
- Visual Crossing

3.2.1. Ergast Developer API

The Ergast Developer API is an experimental web service which provides a historical record of motor racing data for non-commercial purposes (Ergast Developer API, n.d.).

The Ergast Developer API was the main source for the Formula One data. The data was available to download as csv files and was structured in table format as can be seen in Figure 11 displaying the entity relationship diagram, which outlines the type of data that is available within the dataset. The data was available to download as a zipped folder containing 14 csv files corresponding to each of the tables in the entity relationship diagram. The data is explored in more detail in 3.4.1 The Ergast Developer had data on all the full F1 seasons from its beginning in 1950 until the end of the 2022 season. The Ergast Developer also contained data from the first few races of the 2023 season as it was being updated after each race with the relevant statistics.

An updated version of the data was downloaded after the midpoint submission as there was additional 2022 race data that was not available at the time of the original download as the 2022 f1 season was still underway. The Ergast Developer API had data that covered most of the requirements that had been set out for the data in 3.1.1

Figure 11: Entity Relationship Diagram for Ergast Developer API Data (Ergast Developer API, n.d.)

requirements outline above. The Ergast Developer API was selected as it was the most openly available source for the data and contained a large amount of f1 data. The ethical consideration for this data source is outlined in the 3.3 below, detailing the permissions that are available for the data source.

3.2.2. Formula One Wiki

The Ergast Developer API had the majority of the F1 statistics that were needed for the races, however it did not have information or statistics on the safety car, virtual safety car, or red flag use for a race. These were the main attributes that were to be analysed for the project to analyse the impact of weather on them. Therefore, this data needed to be sourced elsewhere.

The Formula 1 Wiki is Fandom's leading reference on the world's most prestigious motor racing competition, Formula One, in terms of size, detail and relevance (F1 Wiki, n.d.).

Formula One Wiki had two pages that were specific to what was required. A dedicated page with lists detailing the safety car and virtual safety car data, which included details such as the race they were deployed at, the cause of the safety car, the lap the car was deployed and called in, and the number of full laps completed by the safety cars, a similar table was produced for the virtual safety cars. Figure 12 below displays an extract from the Formula One Wiki Safety Car data.

	Grand Prix	Cause	Deployed (lap)	Called in (lap)	Number of full laps
1	1973 Canadian Grand Prix	Accident	33	39	5
2	1993 Brazilian Grand Prix	Accident/Rain	29	38	8
3	1993 British Grand Prix	Stranded car	38	40	1
4	1994 San Marino Grand Prix	Accident	1	6	4
5	1995 Belgian Grand Prix	Rain	28	33	4
6	1996 Argentine Grand Prix	Accident	28	33	4
7	1996 Belgian Grand Prix	Accident	13	18	4
8	1997 Argentine Grand Prix	Accident	1	6	4
9	1997 Belgian Grand Prix	Rain	1 (Race started behind SC)	4	3

Figure 12: Safety Car Data Extract from Formula One Wiki (F1 Wiki, n.d.)

A second page which detailed the red flag usage during F1 races was available on Formula One Wiki. The red flag consisted of the race name and year, the lap that the red flag was issued on, the reason for the incident and the record of which drivers failed to restart because of the red flag. The data on the Formula One Wiki was presented in a tabular format with some merged rows due to the duplication in events where more than one safety car or red flag were deployed, Figure 13 provided is an extract from the Formula One Wiki and displaying the first few instances in the data.

The data contained year and race name values, so had enough information for merging this data with the Ergast Developer Data. Like the Ergast developer API, the Formula One Wiki data was available to be used and the ethical considerations are outlined in 3.3.

N	The race was not restarted.
Y	The race was restarted.
R	The race was resumed with the scheduled distance completed.
S	The race was restarted over a shortened distance.

Event	No.	R	Incident	Failed to Restart
1950 Indianapolis 500	138	N	Rain.	
1971 Canadian Grand Prix	64	N	Mist.	
1973 British Grand Prix	2	Y	Accident involving Jody Scheckter, Mike Hailwood, Jochen Mass, Carlos Pace, Jean-Pierre Beltoise, Andrea de Adamich, Roger Williamson, George Follmer, and Jackie Oliver.	Jody Scheckter, Mike Hailwood, Jochen Mass, Carlos Pace, Jean-Pierre Beltoise, Andrea de Adamich (injury), Roger Williamson, George Follmer, and Jackie Oliver.
1974 Brazilian Grand Prix	32	N	Rain.	

Figure 13: Red Flag Data extract from Formula One Wiki (F1 Wiki, n.d.)

3.2.3. Motorsport Stats

Originally when the data was first sourced, the data from Motorsport Stats was not needed as the Formula One data from the Ergast Developer API, and the Formula One Wiki provided enough of the attributes that were needed for the project. It was on further analysis of the data that it was discovered that the race time records in the Ergast Developer API had errors and inconsistencies within the data. There was no record of the time zone that the time data was being held in, so it was assumed that the times were recorded in GMT. Efforts were then made to convert this time value to the local time of the racetrack, by using the pytz library in python to compute time zone calculations for each of the specified regions. It was also noted that the summertime hours within each of the regions needed to be checked and efforts made to ensure that in the time zone calculation that the addition or subtraction of the hour depending on the circumstance was included.

After these efforts were completed. It was noticed that there were some errors within the time data. Most notably in the time zone conversion results it was noted that some of the races were calculated to have been held at 1am local time in some regions. On further investigation it was noted that some of the times that were recorded in the Ergast Developer API were recorded in local time, and others appeared to be in GMT. For that reason, it was decided to locate a complete source of the data for the times the races were held at in local time to avoid the errors computed using the Ergast Developer API data.

Motorsport Stats is one of the largest repositories of motor racing results and statistics from F1 to WRC (Motorsport Stats, n.d.). Motorsport Stats had records and on their site with details of each of the Formula One races from the selected period of the 2005 season to the end of the 2022 season, which included the start time of the race in local time. It was decided that this data would be used to replace the calculation attempts of local time from the Ergast Developer API data. The time data for each of the races for the dataset was then manually recorded from each of the pages dedicated to each of the races and input into Excel. The details of the data compilation are explored in more detail in 3.4.3. The ethical consideration for the Mortorsport Stats data is outlined in 3.3, as permission was needed before the data could be used from the website.

3.2.4. Visual Crossing

Visual Crossing is the easiest-to-use and lowest-cost source for historical and forecast weather data (Visual Crossing, n.d.).

Visual crossing had an incredibly detailed amount of data that was available for download using longitude and latitude coordinates of the tracks. From the research conducted into a number of different sources for weather data, Visual Crossing was decided to be the most suitable for sourcing the data for the project. In comparison to other weather sources available, Visual Crossing had the option to access daily and hourly weather data from a much larger range of years than some of the other sources that were initially sourced. Visual Crossing had 50 years' worth of historical weather data. Visual Crossing also had a number of the statistics that were mentioned in the State of the Art study such as: temperature, precipitation, humidity, dew, wind gust, wind speed, pressure, solar radiation, etc. (Knechtle, et al., 2019) (Thornes, 1977) (Whitehead & Wicker, 2020) (Dawkins & Stern, n.d.).

Like the Ergast Developer API, the data was available to download as csv files. The weather data could be sourced using longitude and latitude values and it would return results based on the weather stations that were near to the location.

There was a lot of weather data that was required for the project to be downloaded. Visual Crossing allowed a limited number of data downloads that could be performed in a day with the standard account. Visual Crossing offers free academic upgrade to subscription after reaching out with the details of the project and purpose of the project. The academic upgrade gives the same permissions as their corporate subscription level, allowing unlimited records download per month, as opposed to the 1000 records per day limit that is available on the free subscription level. The academic upgrade also gave access to the hourly data that was needed for the project (Visual Crossing, 2021).

3.3. Data Ethical Considerations

It was vital that the data that was sourced for the analysis was available for use and, if required, permission was obtained for the data to be used in the project before the work could begin on the data. When sourcing the data, it was important to check the permission

level on the data or the websites the data was available on because in some cases permission was needed to be granted for using specific data and the steps needed for gaining permission were outlined in detail on the source webpages.

Ergast Developer API: gives permission for personal, non-commercial applications and services including educational and research purposes (Ergast Developer API, n.d.).

Formula One Wiki: All the text on the F1 Wiki is licensed under the Creative Commons Attribution-Share Alike License 3.0 (Unported) (CC-BY-SA) (F1 Wiki, n.d.).

Visual Crossing: Permission was granted with the creation of an account to access the data but with limitations on the access of the data and the downloading capabilities of the account. After upgrading the account to an academic research account, more access was granted for the account (Visual Crossing, 2021). Visual Crossing reached out to request the findings of the project be shared with them upon completion.

Motorsport Stats: A data and analytics division of Motorsport Network, their policy states that unauthorised reproduction or translation of any content (including words, data, information, photos, videos and any other intellectual property) is prohibited (Motorsport Stats, n.d.), and therefore permission was sought out to gain this access to the data. After reaching out to the Motorsport Stats to enquire about getting permission to use the local time data for the 2005 races to the end of the 2022 seasons races, permission was received for the data to be taken from the website and utilized for the purpose of the project.

3.4. Compiling the Data

The largest and most time-consuming aspect of the project was organising the data for compilation, and ensuring the appropriate data was merged into the dataset for the project. Using Python modules and libraries and Excel these data sources were to be merged together to create the final data frame for analysis, and additional pre-processing etc. The data was complex as it was coming from a number of sources, and also in a number of formats.

3.4.1. Ergast Developer API

The data was compiled using the csv files, there were 14 csv files that were downloaded in the zipped folder. Each of these 14 files had a number of instances and attributes within each of the files, a summary of the 14 files is provided in Table 5. In order to decide which files were needed for the project, all of the csv data files were read into the Jupyter notebooks using pandas. The top columns of each of the 14 datasets were shown to determine what attributes from the data frames were best suited for the project and to discover what primary of foreign keys from each of the tables would allow a merge of the chosen columns into the dataframes. The descriptive statistics were computed for each of the 14 dataframes, which showed a count of each of the numeric instances in the data frame.

Table 5: Summary of Ergast Developer API Data Files

	File Size [MB]	File Type	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
circuits	0.01	CSV	9	77	5	4
constructor_results	0.219	CSV	5	12,200	1	4
constructor_standings	0.312	CSV	7	12,961	0	7
constructors	0.018	CSV	5	211	4	1
driver_standings	0.873	CSV	7	33,942	0	7
drivers	0.093	CSV	9	857	6	3
lap_times	16.268	CSV	6	541,113	0	6
pit_stops	0.381	CSV	7	9,773	0	7
qualifying	0.423	CSV	9	9,635	0	9
Races	0.157	CSV	18	1,102	2	16
Results	1.64	CSV	18	25900	0	18
seasons	0.005	CSV	2	74	1	1
sprint_results	0.009	CSV	16	120	0	16
Status	0.003	CSV	2	139	1	1

Three dataframes were selected from the 14 dataframes that were obtained from the Ergast Developer API, the 3 selected had the most appropriate data that met the requirements outlined for the data above. The ‘races’, ‘circuits’, and ‘lap_times’ dataframes were the selected dataframes from the Ergast Developer API data. These three dataframes had primary and foreign keys attributes within the dataframes, which would allow for an accurate merge of the dataframes. The ‘races’ and ‘circuits’ had the ‘circuitId’ common between the dataframes, while the ‘races’ and the ‘lap_times’ had the ‘raceId’ common between the two. These were the keys to be used when merging the dataframes together. All the relevant data will be merged onto the ‘races’ data frame. Before merging the dataframes however some initial cleaning and pre-processing was carried out.

‘races’

As the data from the ‘circuits’ and ‘lap_times’ is being merged onto the ‘races’ data frame, it was important to ensure the data frame was clean of unnecessary attributes and instances that are not to be included within the analysis. Starting with the ‘races’

data frame, 11 attributes were initially dropped from the data frame as they were not relevant to the F1 races, and instead were related to free practices, sprints, and qualifying. As the analysis is looking at only the effect on the races, then the qualifying and sprint data does not need to be included within the data. The sprint races themselves were introduced into Formula One in 2021, with three races held in the 2021 season, and three races held in the 2022 season, as this is a very limited amount of data, it made sense to remove it from the data frame (Coleman, 2023).

The 'races' data frame had a total of 1102 instance which equated to each of the races from the first race in 1950 until the schedule for the 2023 calendar. Using regular expression all instances that did not have a value for the 'time' the race was scheduled for was dropped from the data frame. As the weather will be gathered for specific times around the scheduled time for the race, any races that did not include a value for the time was dropped, as corresponding weather data would be unable to be sourced for the period of time the race was held. The resulting data frame that was left consisted of 371 rows. These 371 instances began in 2005 and included all scheduled races for the 2023 season.

When the data was originally compiled at the beginning of the project, the data was downloaded before the 2022 season had been concluded. It was only when later into the project when the lap time data was being merged onto the data was it discovered that some of the data from the end of the 2022 season was missing from the initial download, and therefore an updated download of the Ergast Developer API data was required. Therefore, there was more recent race schedule data available for 2023 within the data. Because the 2023 scheduled race calendar was included in the data frame, and the 2023 was not yet completed, the decision was to drop these instances, so the dataset only contained the races from 2005 until the end of the 2022 season. A summary of the 'races' data frame after dropping instances and attributes has been provided in Table 6 which can be compared to the original data frame characteristics in Table 5.

Table 6: Summary of 'races' data frame after unnecessary columns attributes and instances dropped

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
Races	7	348	1	6

'circuits'

From the 'circuits' data frame, the 'url' was the only column that was dropped at this stage, as the remaining attributes would provide accurate detail on the circuits location and would be needed for gathering the weather data. The resulting data frame after dropping the 'url' attribute can be seen in Table 7.

Table 7: Summary of the 'circuits' data frame after unnecessary column dropped

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
circuits	77	8	4	4

'lap_times'

In the lap_times data frame, the data had lap times recorded by every race driver during each lap of the race they were present in. The data frame consisted of over 540,000 instances. Each instance corresponded to a lap time achieved by a driver on a specific lap at the specific race. To merge this 'lap_times' data with the 'races' data, which had a single instance for a unique race, the lap time data needed to be presented in a single row, corresponding to a unique race so to merge the data into the correct format. The 'lap_times' data frame also had the position the driver was in when they completed the lap and had the lap time in mm:ss.sss format and in milliseconds.

It was decided to compute the top 5 lap times achieved at each circuit during a specific race, as there were values for every race lap completed by each driver. The analysis is to look at the impact of weather on Formula One races, and therefore comparing the fastest lap times completed under the different weather conditions would be more informative than having every single lap time recorded. The lap times were grouped by the 'raceId' which was the unique identifier for each of the races. The 5 fastest laps for each of the races were then found using lambda and nsmallest. Each of these 5 fastest laps for each race were then placed into 5 new attribute columns that were created ('fast_lap_1', 'fast_lap_2', 'fast_lap_3', 'fast_lap_4' and 'fast_lap_5'), with 'fast_lap_1', being the fastest lap recorded during the race. The fastest lap columns were converted into milliseconds, and then formatting the timedelta as mm:ss.sss

Unnecessary columns and instances were dropped from the data frame, including the 'driverId', 'lap', 'position', etc, as these attributes were not needed for the analysis into the weather and the impact on the races, as these attribute are driver specific and specific to a certain instance during the race, which is unnecessary to keep within the dataframe. These instances that were dropped were the duplicate rows that resulting in getting the top 5 fastest laps for each unique race. The last instance of each of the duplicate rows was kept as only one instance of each race was required for the merge. All other instances with the same attributes were dropped from the data frame. A data frame with the top 5 fastest laps for each race was then ready for merging on the merged race and circuits data. The resulting 'lap_times' data frame after dropping attributes and instances is displayed in Table 8 below**.**

Table 8: Summary of the 'lap_times' data frame after unnecessary attributes and instances dropped

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
lap_times	6	501	0	6

The ‘lap_times’ data frame was also used to create a record for the total number of laps that were completed during a race. This race duration data frame was created on a copy of the ‘lap_times’ data and was computed by taking the max value of the ‘lap’ for each race, using the ‘raceId’ of each race. This data frame ‘max_laps’ was then merged on the ‘races’ data using the ‘raceId’.

Table 9: 'lap_times' data transformed to create a new data frame for race duration

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
max_laps	501	2	0	2

The ‘circuits’ and ‘ ‘lap_times’ dataframes were then merged in two batches onto the ‘races’ data. First the ‘races’ and ‘circuits’ were merged on the ‘circuitId’, and the values of the data frame were sorted by the year and the round number so the first value in the data frame was from 2005 and was round 1 in the season. There were duplicate attribute names within the data frames, and when merged, they took an ‘_x’ or an ‘_y’ at the end of the name. These duplicate columns that were in the data frame did not have the same values, so the attributes were renamed. Both ‘races’ and ‘circuits’ had an attribute called ‘name’, but with different values in the two columns, for example the ‘name’ in the ‘races’ referred to the name of the race (e.g., Australian Grand Prix), whereas the ‘name’ in the ‘circuits’, referred to the circuit name (e.g., Albert Park Grand Prix Circuit), so the two attributes were renamed to reflect their differences. Renamed to ‘raceName’ and ‘circuitName’ respectively.

The ‘lap_times’ data frame was then merged onto the ‘race’ and ‘circuits’ merged data frame. For this merge the data frames could be merged on the ‘raceId’, which was the unique identifier for each of the races. The resulting data frame ‘merged_df’ from these merges is summarised in Table 10. The merge created a data frame with 20 attributes and 348 instances.

Table 10: The resulting data frame from the merge of the 'races', 'circuits' and 'lap_times' data

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
merged_df	20	348	5	15

3.4.2. Formula 1 Wiki

From the Formula 1 Wiki, the safety car, virtual safety car, and red flag data was sourced. As already mentioned above in sourcing the data, the data from the Formula

One Wiki was manually input and format in separate excel files which were saved as csv files to be opened in Python. An example of the data format on the website is provided in Figure 12 and Figure 13 above.

Safety Car Data

The safety car data on the Formula 1 Wiki website had 5 attributes and 301 instances, as seen in Table 11. This data was copied into the excel file and transformed to include more attributes using the data present in the csv file. For example, some of the numerical attributes i.e., 'Deployed (lap)', which had the lap number that the safety car was deployed, also had instances that had categorical data, explaining more detail about the deployment of the safety car. An example of this can be seen in Figure 12 above where the 1997 Belgian Grand Prix safety car was deployed on lap 1, but with the additional information in the column, we can see that the race started behind the safety car, instead of the safety car being deployed during the 1st lap. All the categorical values within the table instances were found using filters and pivot tables in Excel. There were repeated values within the columns, and therefore it was decided to separate these categorical values into new attributes within the data frame. For example, 'Started/ Resumed Behind Safety Car' became one of the new attributes in the data table.

Table 11: Summary of Safety car data from Formula One Wiki

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
Safety car data	301	5	2	3

The next bit of formatting to get the data correctly format in csv was with splitting the year and race name into two separate attributes. These attributes were previously present as a single attribute on the website e.g., '1973 Canadian Grand Prix' under the 'Grand Prix' attribute was changed to '1973' for 'Year' and 'Canadian Grand Prix', for 'Grand Prix'. To convert these into two separate attributes for 'Year' and the 'Grand Prix', Excel text to columns was utilized. As the years were at the beginning of each of the instances, the fixed width was utilised and split the data into the correct columns.

After the data was transformed in Excel, the data was converted into a csv file and then opened using pandas in the Jupyter Notebook. The safety car data from wiki had 5 attributes as seen in Table 11. The safety car data frame had 11 attributes and 301 instances of safety car occurrences after the transformation, as see in Table 12.

Table 12: Sumary of Safety Car Data after processing in Excel

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
safety_car	11	301	7	3

In python the csv data was read using pandas and some exploratory data analysis was computed on the safety car data, which is described in more detail in 3.4 Exploratory Data Analytics below.

After the exploratory data analytics was computed the unnecessary columns from the data frame were dropped as they were deemed not necessary for the analysis. And the remaining columns were renamed in preparation for merging with the merged races data frame. For example, 'Year', was renamed to 'year' to remove the capitalisation, and 'Grand Prix' renamed to 'raceName', to match the format of the merged races data frame.

The values in the data frame were sorted so the years were in ascending order. The first value in the data frame was for 1973. From there the data that was greater than or equal to 2005, and less than 2023 was selected, to maintain the time period that was required for the project. It was later discovered that some of the names of the races were different in the safety car data than were In the races data, and therefore their values did not all merge with the rest of the safety car data. In the safety car data for example the issue was the safety car had a hyphen between some of the 'raceName', for example the 'Emilia-Romagna Grand Prix' was renamed to the 'Emilia Romagna Grand Prix', to remove the hyphen and to allow the corresponding data to be merged onto the data frame. Another example was the 'São Paulo Grand Prix' was renamed to the 'Brazilian Grand Prix' to match the merge.

Similar to the lap time data, the safety car data has some calculations to establish the total number of laps that a safety car completed in a single race, as well as a count of the number of safety cars deployed in a single race. The data in the safety car data frame had a single row dedicated to an instance that the safety car was deployed. In order for the safety car data to correctly merge to the merged data frame, the safety car data needed to be format in a single line per race, as that was the format of the data in the merged data frame.

To do this a new column was added to the safety car data frame called 'sc_totalLaps' which was to take count of the total number of full laps the safety car completed during a race, over each of the deployments of the safety car. This value was obtained by grouping the data frame by year and 'raceName' and selecting to get the sum of the 'sc_fullLaps' column.

The number of times the safety car was deployed during a race was also calculated under 'sc_count'. 'sc_count' was initialised as 1, as each instance was an occurrence of the safety car during a race. The rows the had the same 'year' and 'raceName' were the columns that added an instance to the 'sc_count' attribute. The last occurrence of the 'sc_count' for a 'year' and 'raceName' had the final figure for the total number of safety cars deployed during the race. Because the last values are the instance that was to be kept for merging with the race data, only the last occurrence of the 'year' and 'raceName' in the data frame was kept, and all the others were dropped. The resulting data frame from the transforming and cleaning is described in Table 13.

Table 13: Summary of Safety Car Data after processing and transformation with Python

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
safety_car	4	169	1	3

Virtual Safety Car Data

The virtual safety car data was the second data table that was taken from Formula One Wiki. The same processing steps were taken for the virtual safety car data as was completed in the safety car data. Categorical instances were converted into attribute heading. For the virtual safety car data, only one additional column was created from the categorical instance, as there was only one additional comment within the data as can be seen in the extract of the data from Formula One Wiki table in Figure 14 below. 'Turned into a full SC' was created into an attribute heading and the values were computed. The 'Grand Prix' was also split into 'Year' and 'Grand Prix' as described above for the safety car, using Excel text to columns. A summary of the data after being transformed in Excel is seen in Table 14. The data was then read into Jupyter for analysis.

	Grand Prix	Cause	Deployed (lap)	Ended (lap)	[Collapse] Number of full laps
1	2015 Monaco Grand Prix	Accident	64	64 (turned into a full SC)	0
2	2015 British Grand Prix	Stranded car	34	35	0
3	2015 Hungarian Grand Prix	Accident	43	44 (turned into a full SC)	0
4	2015 Belgian Grand Prix	Stranded car	21	22	0
5	2015 Singapore Grand Prix	Accident	13	15 (turned into a full SC)	1
6	2015 United States Grand Prix	Debris from accident	5	8	2
7		Accident/Stranded car	38	39	0

Figure 14: Extract from Virtual Safety Car Data on Formula One Wiki (F1 Wiki, n.d.)

Table 14: Summary of Virtual Safety car data after processing in Excel

	No. of Attributes	No. of Instances	Categorical Attributes	Numerical Attributes
Virtual Safety Car	7	54	3	4

Exploratory data analysis was computed for the virtual safety car data, which is described in more detail in 3.4 Exploratory Data Analytics described below. After Exploratory analysis, the unnecessary columns were dropped from the data frame.

Again, the attributes were renamed in preparation for merging, and 'Year' was renamed 'year', 'Grand Prix' renamed 'raceName' etc. There was a value of the '2015 United States Grand Prix' where the 'year' and 'raceName' did not correctly split and

the ‘2015’ was present within the ‘year’ and the ‘raceName’ attributes. This value was fixed by copying the value of the previous instance that had the correct name ‘United States Grand Prix’ and applying that value to the row with the error.

The values in the virtual safety car data frame were sorted in ascending order by ‘year’, ‘raceName’, and ‘vsc_startLap’ and like the ‘lap_time’ data and the safety car data, the total laps completed by the virtual safety car were computed. This was calculated by subtracting the start lap of the virtual safety car from the end lap of the virtual safety car. The sum of the total laps by race was computed by grouping the ‘year’ and ‘raceName’. The number of virtual safety cars deployed during each race were computed by adding 1 to the count when the previous row in the data frame has the same ‘year’ and ‘raceName’. Duplicate values were removed from the data by keeping only the last instance of the ‘year’ and ‘raceName’.

Unnecessary columns were dropped from the data frame and the summary of the virtual safety car data before being merged is provided in Table 15. **T**he table can be compared to the data that was provided from Formula One Wiki. After the processing and dropping, and creation of new attributes, the same number of attributes as were provided in the Formula One Wiki site were the result of the compilation.

Table 15: Summary of Virtual Safety Car Data after processing and transformation with Python

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
Virtual Safety Car	5	40	2	3

Red Flag Data

The red flag data was the final data table that was compiled from Formula One Wiki. The data in the Formula One Wiki table was broken down into codes for the race being restarted, not restarting, resuming with scheduled race distance completed, or shortened race distance completed. An extract of the Formula One Wiki table is provided in Figure 13 above.

These values were converted into a status column in Excel. The ‘Event’ attribute was split into ‘Year’ and ‘Country/Circuit’, and the ’Incident’ attribute, was summarised to the description of the red flag reason ‘Desc’ attribute. A summary of the data frame that was read into the Jupyter notebook after processing and transformation in Excel is provided in Table 16 below.

Table 16: Summary of Red Flag data after processing in Excel

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
Red flag data	5	87	5	0

The data types for the attributes were obtained, and the 'Year' and the 'Lap No.' were shown as object values so were converted to integers. Some exploratory analysis was conducted with the attributes in the data, which is explored in more detail below. The attributes were renamed to match the races data for the merge. 'Year' was renamed to 'year' to remove the capitalisation. 'Country/ Circuit' renamed to 'raceName'. 'Status' and 'Desc' were renamed to 'rf_status' and 'rf_desc'. The data for the 2005 season and after were kept, dropping the other red flags from earlier races. Again the 'Emilia-Romagna Grand Prix' needed to be renamed to remove the hyphen.

The total number of red flags per race was calculated using the same method as was used in the safety car and virtual safety car total counts. The duplicate rows were dropped, and the data frame was summarised in Table 17.

Table 17: Summary of Red Flag Data after processing and transformation with Python

	No. of Attributes	**No. of Instances**	**Categorical Attributes**	**Numerical Attributes**
Red flag data	5	25	3	1

Merging on the races data

The safety car, virtual safety car, and red flag data were all joined separately on the races data using the 'raceName' and 'year' attributes. They were joined using a left join on the data. To ensure all the data had merged over, the length of each of the total columns were computed to ensure each value merged correctly. It was at this stage that it was discovered that the 'raceName' for some of the races were incorrect as they had not merged over with the rest of the data, so the corrections were able to be made in the processing of the data frames to avoid this error occurring again. The resulting data frame after all three data frames were merged onto the data is displayed in the Table 18 below.

Table 18: 'races' data frame after 'circuits' and 'lap_times' merged

	No. of Attributes	**No. of Instances**
merged_df	28	348

3.4.3. Motorsport Stats

The data collected from Motorsport Stats was collected by converting the races data frame to a csv and using the 'year' and 'raceName' data to search for the records relating to the specific race and noting the time that was present on the website. The race time on the Motorsport Stats website was displayed in local time, so the column was added into the Excel csv file of the races data for 'localTime'. Each of the 348 races were searched for on the website, and their local time recorded in the Excel in hh:mm:ss.

The local time data was merged to the races merged data. The merge was a left join on 'raceName', 'circuitRef' and 'year'. Adding 'localTime' to the races data frame on the merge. The resulting data frame after the merge is shown in Table 19 below. The table shown in Table 19 included the attributes that were created as a result of computing the "local time" from the Ergast Developer Data, as mentioned in Motorsport Stats 3.2.3 above. One of the values for the local time did not merge on the data: the 'São Paulo Grand Prix' of 2021. So, the local time was confirmed on the Motorsport Stats website and the correctly value was copied form another instance in the data frame and applied to the data frame.

Table 19: Merged dataframe after local times added from Motorsport Stats Data (incl. local time calculation)

	No. of Attributes	**No. of Instances**
merged_df	32	348

After the local time data was merged onto the races data, it was noted that the values for the race times were not always on the hour, as some races started at 10 minutes past the hours or at half past the hour. As the weather data was going to be gathered in hourly increments, it was decided to round the values of local time down to the nearest hour and have that time as the scheduled race start time. With the scheduled race time calculated, the period around the race was calculated, by computing the two hours before the race scheduled time, and the three hours after the race is scheduled to start. From the State of the Art, in an example there was evidence of weather data being collected for the period that the Boston marathon was taking place within (Knechtle, et al., 2019). As there was evidence of Formula One races being started under safety car conditions, it was decided to collect the two hours before the race was scheduled to start, in order to analyse the weather leading up to the scheduled start time, and whether this weather had an impact on the race. The three hours after the race was started were also computed, as there is reference to the F1 regulations stipulating that an F1 race once started must be completed within a total race time of three hours (Horton, 2021). These race times were added to the merged data frame as new instances under the following attribute headings: 'race_time_less_2','race_time_less_1','race_time',race_time_plus_1', 'race_time_plus_2', and 'race_time_plus_3'.

Table 20: Merged dataframe with 6 hour period

	No. of Attributes	**No. of Instances**
merged_df	39	348

3.4.4. Visual Crossing

Originally the data was downloaded in bulk for every race that were present within the same number of years, as this made the download easier to select on the website. The locations longitude and latitude values were input into the Visual Crossing query builder then the time frame selected, finally the data is downloaded. The resulting

data files were too large to work with when read into pandas, therefore a different approach to downloading the data was explored.

The Visual Crossing data was to be produced for each of the races from 2005 to the end of the 2022 season. The weather data in Visual Crossing could be obtained by using the longitude and latitude values for the racetracks, and then selecting a period of time on the calendar and completing the download. As the data was being downloaded in hourly increments for each racetrack during the period being downloaded for, it was important to create smaller batches of the race data in order to get more manageable download of the data. The races merged data 'date' was converted into 'dd/mm/yyyy' format as opposed to the 'yyyy-mm-dd' that was originally the format of the data. The format that was present within the weather data was 'dd/mm/yyyy', so in preparation for merging the weather data with the races merged data.

A csv file was created for the races merged data, to be used for filtering out data and keeping record of what data had been downloaded. For example, the hourly weather data for a single track was downloaded for the month the race was scheduled within, and the year that the race took place. A single racetrack was filtered from the races csv file, 'albert_park' circuit in Australia will be the example. From the filter in applied to the data in Excel the date of each of these races were displayed. There were 16 instances of races the took place at 'albert_park'. These 16 races were carried out over 16 years and all of these races were held between the months of March and April. The longitude and latitude coordinates were obtained from the csv, and then used with the circuit reference to download the hourly weather data for the months of March and April for the first year, in this example the first year being 2005. Then the csv data file was downloaded and put in a folder with the circuit reference name. Each race had a csv file for each year that a race was held at that location. 'albert_park' had 16 csv files, one for each instance of the race over the selected months. An individual download for each of the races were completed, meaning there were approx. 348 csv files divided between their respective folders. Some races took place more than once in the same year, so there would be one csv for that track in that year. The number of files per track ranged from 2 files to 18 files.

After all the files were downloaded for the individual races, a new csv file was created for each of the racetracks. These racetrack csv files contained the weather statistics that were specific to the date that the race was held. Previously the larger files contained a 2-to-3-month period of data, so these files were to combine all of the race day specific hourly weather for each racetrack. Meaning there were 37 csv files, one for each of the racetracks, and within these files were the weather statistics that were manually copied from the each of the approximately 348 files. A summary of the 37 csv track files is available in Table 21.

Table 21: Summary of the CSV race day weather data

	Size of File [kb]	No. of Attributes	No. of Instances
albert_park	0.073	25	409
americas	0.042	25	242
bahrain	0.063	25	432
baku	0.020	25	144
buddh	0.009	25	72
catalunya	0.076	25	456
fuji	0.009	25	48
hockenheimring	0.030	25	217
hungaroring	0.085	25	432
imola	0.021	25	120
indianapolis	0.012	25	72
interlagos	0.067	25	431
istanbul	0.035	25	216
jeddah	0.007	25	48
losail	0.004	25	24
magny_cours	0.018	25	96
marina_bay	0.057	25	312
miami	0.005	25	24
monaco	0.069	25	408
monza	0.085	25	432
mugello	0.005	25	24
nurburgring	0.028	25	168
portimao	0.008	25	49
red_bull_ring	0.052	25	264
ricard	0.020	25	96
rodriguez	0.025	25	173
sepang	0.051	25	313
shanghai	0.058	25	384
silverstone	0.095	25	456
sochi	0.026	25	192
spa	0.073	25	408
suzuka	0.065	25	336
valencia	0.016	25	120
villeneuve	0.062	25	361
yas_marina	0.053	25	336
yeongam	0.014	25	96

These csv files were then read into the Jupyter Notebook using pandas. And some of the attributes that contained little to no data were removed from the dataframes. For example, ‘snow’ and ‘snowdepth’ were removed from the dataframes as there has never been a race where it has snowed (Opong, 2022). Each of the 37 dataframes were combined into one that was to be used for merging the weather data onto the F1 races data. The complete weather data frame of all race day weather contained 8,456 instances and 17 attributes. The attributes have been described on Visual Crossing website (Visual Crossing, n.d.).

Again, like the F1 'lap_times' data, the weather data has one row that corresponds to a single instance, in this case each row is an hour within the selected days. This format is not suitable for merging with the races data as the races data has a singles instance that corresponds to a unique race. Therefore, the data could not be merged completely as is without some processing. As the weather data was needed for each of the 6 hours that were computed in the local time data, it was decided to rename the attributes of the weather data frame and merge the renamed attributes to correspond with each of the hours in the data frame. An example of the first instance of renaming in the data frame is with the following attributes that were renamed to have '_1', as well as merging the time on the corresponding time within the 6 hours period: 'name' was renamed to 'circuitRef'. The remaining attributes 'temp', 'feelslike', 'dew', 'humidity', 'precip', 'precipprob', 'preciptype', 'windgust', 'windspeed', 'winddir', 'cloudcover' 'visibility', 'conditions', and 'icon', were all renamed with the '_1' until '_6' for the last instance of time in the dataframe.

A left merge was then computed on the first instance of time in the races data, i.e., the 'race_time_less_2' hours. The dataframes were merged on the 'circuitRef', 'date', and the time. Each renaming and merge were completed 6 times in total to ensure each of the hours available in the races data frame, have the corresponding weather in the data frame. The races data frame with all of the corresponding merges is described in the Selection 4.1 Selection below.

3.4 Exploratory Data Analytics

The data came from a number of sources, so it was important to get an understanding of what data was available from these different sources. The exploratory data analysis is to be used to summarise the datasets and their main characteristics. Summary statistics were computed for each of the dataframes, including counts of the values within the dataframes, and an analysis into the numeric values within the dataframes. Before decision were made on which data was to be removed from each of the dataframes, the exploratory data analysis provided an insight into the values that could be obtained from the respective data sources.

Seaborn, Matplotlib and Visme were all used for creating static plots to explore the numeric and categorical attributes present within the dataframes. For example, Figure 15 depicts the safety car data before being merged with the races data and before the data between 2005 and 2022 was selected. The status of safety cars for starting/ restarting the race was explored. It can be seen that approximately 20 occurrences of the safety car were started/ resumed behind the safety car.

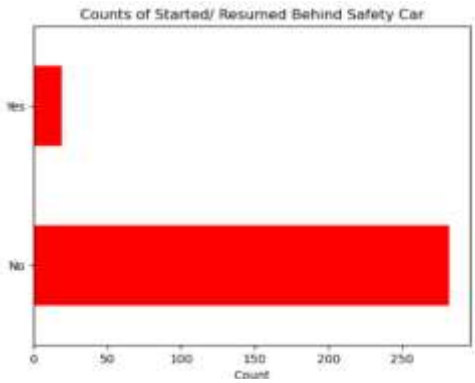


Figure 15: Safety Car Status Count 'Started/ Resumed Behind Safety Car'

Error! Reference source not found. is another example of the exploratory analysis conducted on the safety car data. Exploring the causes of safety cars within Formula One races. The biggest cause of safety cars in the Formula One races, have been described as 'accident', with 'rain' being the third highest cause of safety car occurrences within the Formula One races.

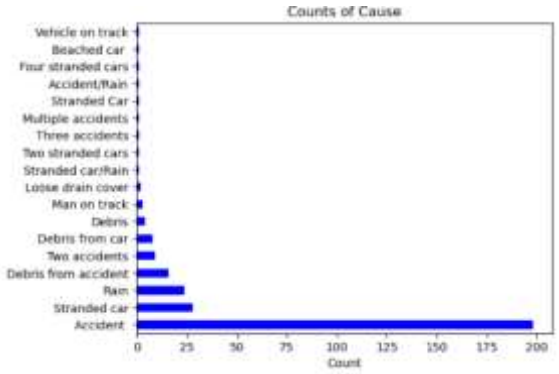


Figure 16: Causes of Safety Cars within Races

In comparison, in **Error! Reference source not found.** the cause of the virtual safety cars has been presented, and rain is not a cause that is mentioned in any of the causes for the virtual safety car. 'Stranded car' being the main cause of the deployment of the virtual safety car that can be derived from the data.

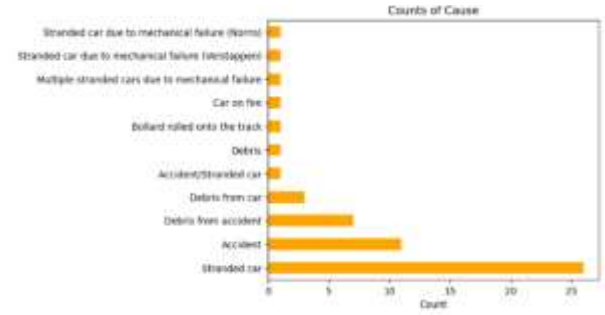


Figure 17: Causes of Virtual Safety Cars

From Figure 18 a breakdown of the tracks that had the most amount of virtual safety cars deployed has been provided. At the lower end of the scale, we see Miami that had one virtual safety car. Miami only has had one formula one race, and this race appeared to have a virtual safety car deployed.

Monaco is the race with the most amount of virtual safety cars that have been deployed, with 6 recorded in the figure. Monaco is one of a few street circuits on the Formula One calendar.

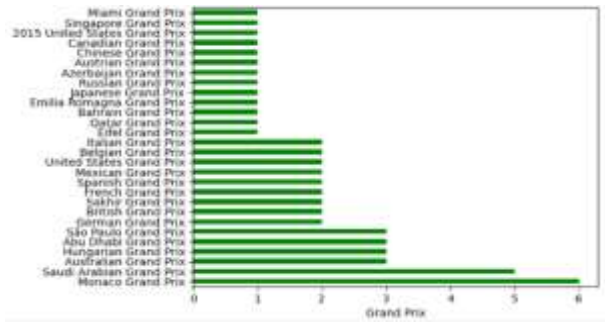


Figure 18: Breakdown of Tracks with number of Virtual Safety Cars

When looking at the Red flag data in Figure 19 it can be seen that Monaco was the third highest track for the number of red flags that have been deployed at the circuit. With 7 red flags deployed at the track, it is only two behind the track with the greatest number of red flags: the British Grand Prix, at the Silverstone circuit.

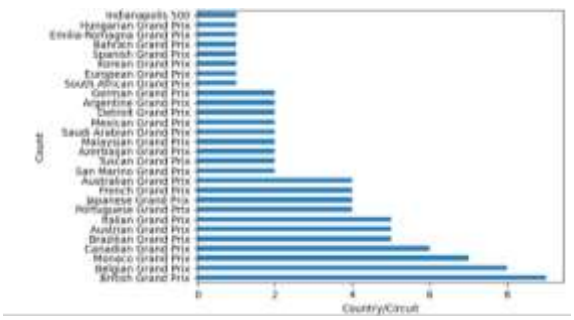


Figure 19: Breakdown of Tracks with number of Red Flags

With information on the tracks that had the greatest number of red flags, it was decided to investigate the main result after a red flag was deployed, Figure 21. The majority of races that had a red flag were restarted, and with the scheduled race distance being completed. The difference between the 'restarted', and 'restarted with scheduled distance completed' was unclear, as there was an option within the graph that displayed the number of red flagged races that were 'not restarted', and 'restarted over a shortened distance'.

Figure 20 displays the breakdown of the races that were not restarted, and the reasons that were associated with the races not restarting. 'Rain' has been attributed to the main reason

these races were failing to resume. ‘Rain and accident’ being the third highest reason for red flagged races to not resume.

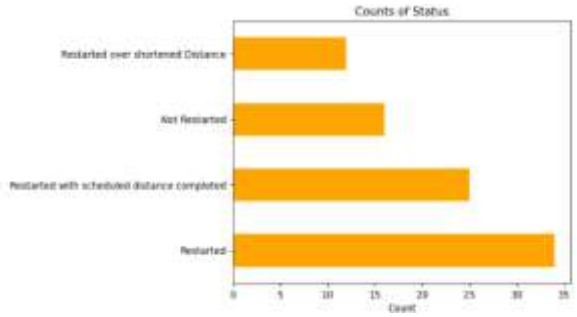


Figure 21: Result of a Red Flag

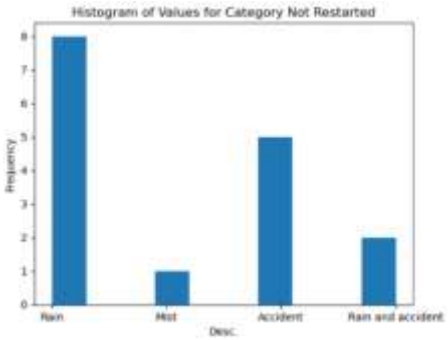


Figure 20: Breakdown of Races 'Not Restarted' after Red flag

Additional exploratory data analytics in the form of a dashboard were produced using Plotly Dash to analyse the data that was within the combined weather and Formula one dataset. The interactive dashboard that was produced for the exploratory analysis can be seen in the extracts in the Figure 22, Figure 23, Figure 24, and Figure 25.

The Dash interactive dashboard uses data from the merged data frame after the missing values have been dealth with and the weather data had been preprocessed and transformed. Details of that dataset can be seen in the Methodology below in Pre-processing and Transformation.

The interactive dashboard allows the user to select a specific year from the 2005 to the 2022 race calendar. When a year is selected, the drop down filter for the Country will be unlocked and the user can select the country from the selected year that they would like to analyse. Finally the track can be selected from the track drop down, as in some years, there were more than one race that took place in the same Country, but at a different racetrack. As the dropdown selection are being made on the dashboard, the map and other visualisation within the dashboard are changing to reflect the user defined selections. The map will zoom into the selected track on the map by the country, and remove the shading from the tracks/ countries that are not selected. The intensity of the blue on the chart refers to the number of races that were present in that location, as can be seen in the key to the right of the map, the darker the blue the higher the number of races, and the lighter the blue, the fewer races that were present at that location Figure 22.

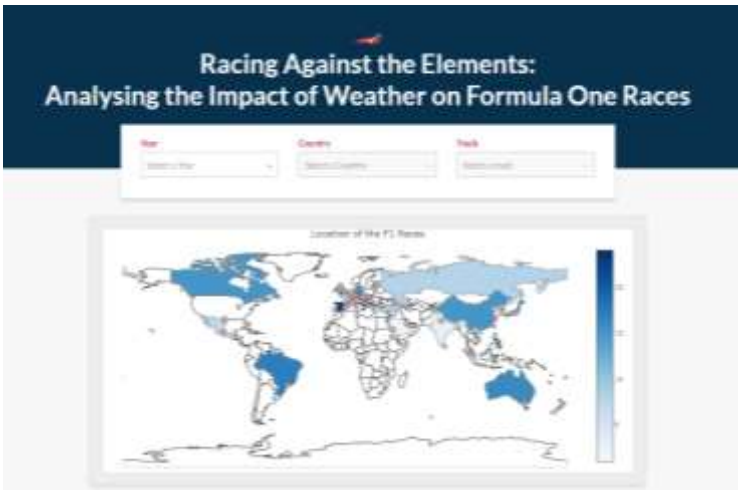


Figure 22: Beginning of Interactive Dashboard (Menu and Map)

After the year, country, and track are selected from the menu dropdowns a number of the race specific statistics for the chosen track and year are displayed. The fastest lap in miliseconds recorded during the race is displayed, along with the race duration in laps, the status of red flags present during the race, the number of safety cars, and virtual safety cars that occurred during the race.

Figure 23: Dashboard Race Specific Statistics (Weather and F1)

The average weather statistics for the 6 hour period around the scheduled race time are also presented in the dashboard: showing the average temperature during the 6 hour period, the average precipitation, average wind speed, and the average visibility are all computed Figure 23.

As well as race specific statistics and visualisation, there were track specific visualisations produced based on the selected track, to show the average precipitation and the fastest lap times that were produced over the years, when there was a race held at that track. The track that is elected will change the values for the titles and the result of the graphs Figure 24. The figures allow a general picture to be derived from the relationship between precipitation and laptimes, and find a year that had a lot of precipitation, which can encourage users to change the values in the dropdown to get the race day statistics form this racetrack. And to see whether weather has an apparent impact on the Formula One races. The average precipitation for the racetracks was presented on a bar chart, with years that had no bar, meaning there was no average precipitation recorded during the race.

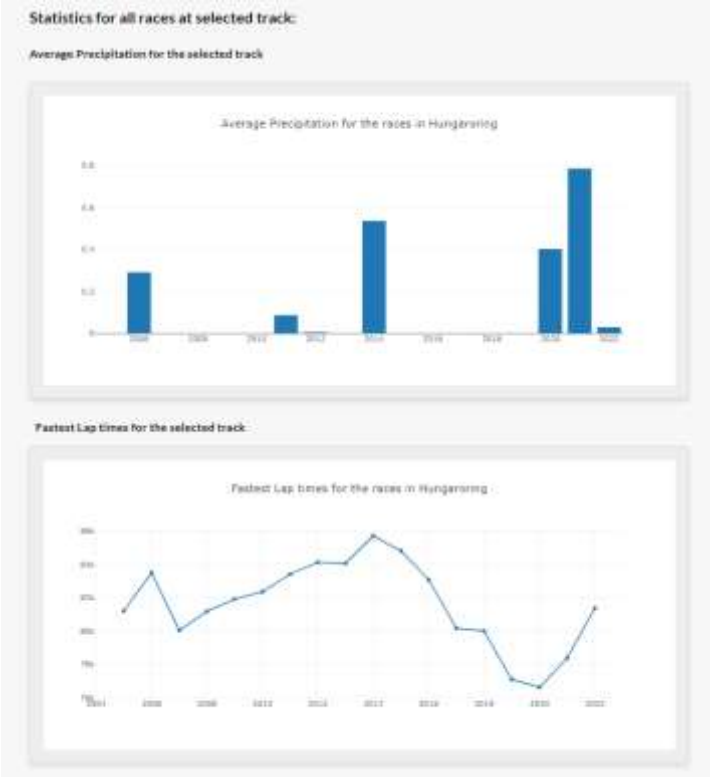


Figure 24: Track Specific Statistics (Lap times, Precipitation)

The fastest laps were presented on a line chart, with the closer the value to the x axis, the faster the lap is, so for example on the chart, the 2014 race at the Hungaroring had the second highest value for precipitation during the race, but the slowest lap recorded on the fastest lap line chart. In contrast the 2021 race at the Hungaroring had the most amount of precipitation recorded, but the third fastest lap recorded at the circuit during an f1 race.

Finally the dashboard displays the race calendar for the chosen year selected, and the order that the races were run. The track map, displays the path that the Formula One seson has travelled in the selected year, in order to travel to each of the race countries in the order that the calendar is followed Figure 25. The overlap that is present within the calenadar can be seen with the pink lines which are travelling back and forth between the different countries and regions.

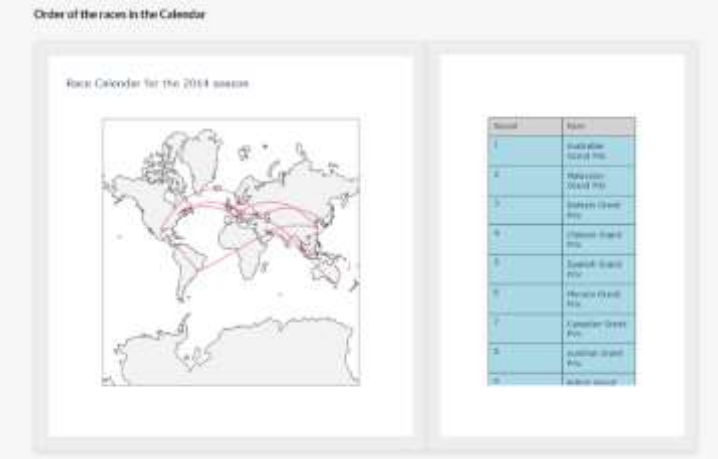


Figure 25: Order of Races in the calendar

4.0 Methodology

KDD- Knowledge Discovery in Databases was the chosen methodology for the project. As the project aim was to investigate the impact of the weather on Formula One. This analysis is completed by analysing the data within the dataset that was created, and extracting insights from the data, that may have been previously unknown, as the scope of the project has not been completed before. KDD is a process that usually involves the 7 main steps: selection, pre-processing, transformation, data mining, interpretation, evaluation, and deployment. Each of these steps are described and outlined below and detailing their relevance to the project. Some of the steps are explored in greater detail in 5.0 and 6.0, within the Analysis and Results section of the report.

4.1. Selection

Data selection involves selecting a relevant subset of data for the analysis (Rajput, n.d.). The sourcing and compilation of the data was described in detail in Data, 3.0 above. The data was selected from a number of sources based on the requirements defined for the project, and also selected based on the data that was available to download with relevant permission obtained, if required. It was important that data selection is the first step within the KDD methodology, and within the project. With the relevant data obtained for use within the project at the beginning of the project, then the missing attributes can be discovered early on in the process and efforts can be made to correct errors or gaps within the data.

For the purpose of the study, and with the available data that was obtainable, the chosen data consisted of weather and formula one race statistics for each Formula One race from the beginning of the 2005 season, until the end of the 2022 season. This data included 348 races, which took place at 17 racetracks across a number of countries around the world. The exploratory analysis was vital for investigating the data that was available within the selected data sources, and also highlighted what data had errors, or was missing data. As mentioned in the Motorsport Stats section in 3.2.3. It was discovered during the exploratory analysis that the time data that was available for the scheduled race times contained inconsistencies within the data. With some of the values appearing

to be recorded in the tracks local time but other being recorded in GMT. And after computing calculations for the changing the time values to local time, some of the races were being recorded as taking place at 1 a.m. which immediately could be flagged as an error. Therefore, when the issue was discovered, it was important that efforts were made to find a suitable replacement for the data. If the data selection and exploratory analysis was not conducted, then the error would not have been discovered. These errors would have impacted the results of the project. The Motorsport Stats data was sourced, and permission granted to use the time data from the website for the 2005 season to the 2022 season.

As the dataset had to be compiled from several sources for the project, it was expected that there were a lot of redundant columns or attributes that were not going to be useful for the analysis. As the data was being compiled, the exploratory analysis provided insight into a number of columns that could be dropped from the dataset. As the project progressed more attributes were cleaned from the analysis, however at this stage of the process, the main noisy data was dropped from the dataset. below in Table 22 is a summary of the data that was compiled at the beginning of the project and the result of the merging of the data sources from the data section above:

Table 22: Summary of the Data compiled from Data 3.0

	No. of Attributes	**No. of Instances**
Data	123	348

4.2. Pre-processing

Pre-processing in the KDD methodology involves cleaning and transforming the data to make it ready for analysis. Pre-processing task can include tasks such as data normalization, missing value handling, and data integration (Rajput, n.d.)**.**

A lot of the noise was removed for the data during the sourcing and compilation of the data into a single data frame. It was not an exhaustive amount of data that was removed, but the instances that were not relevant, i.e., the instances that did not have values for the time the race was scheduled. Attributes that were deemed unnecessary were also dropped from the data during this process.

Summary statistics were computed in the data compiling which detailed counts of the numeric columns in the dataframes. These summary statistics can be used to determine which attributes have missing values present within the data.

Table 23: Summary of the attributes with missing values

	No of attributes with missing values	**Categorical attributes with missing values**	**Numerical attributes with missing values**
Data	40	9	31

Missing value handling had not yet been completed for the data, and so this step was carried out at this stage. The data that was found to be null values within the data frame were dealt with by replacing the null values with the zeros for the numerical values and for the categorical attributes, the null values were dealt with by replacing the values with 0, there were not many null values present within the data, with almost all of the null values coming from the merge between the weather data and the safety car, virtual safety car, and red flag data. As the safety cars, virtual safety cars, and red flags are only occurring in a certain number of the races within the data, then it is not expected to compute values for these races, e.g., using the mean, max or min values to deal with the null values, as these computed values would skew the data, as the weather data would not be changed to reflect these computed values.

The missing values could not be removed from the dataset either, because then the only data that would be present within the data would be the data that corresponded to safety cars, virtual safety cars, and red flag occurrences. Because of this the null values in these numerical instances were replaced with 0. As for example if the total number of safety car, virtual safety cars, or red flags was 0 after the blanks were replaced then that would correctly display that there were 0 safety cars, virtual safety cars or red flags present within the chosen race. Similarly for the categorical attributes within the data, were replaced with categorical values to explain that there was no occurrence of the safety car, virtual safety car, or red flags within the specified race.

From Table 40 out of the 123 attributes had missing values, 9 of those were categorical attributes, with ‘precipitype_1’ to ‘preciptype_6’, ‘vsc_count’, ‘rf_status’ and ‘rf_desc’. Null values in ‘preciptype were replaced with ‘none’, the ‘vsc_count’ null values replaced with ‘No VSC’, and the two red flag attributes replaced with ’No Red Flag During the race’.

4.3. Transformation

Transformation involves transforming the data into a format suitable for data mining and modifying the data to create new features from the data (Rajput, n.d.).

The Data 3.0, details the data transformation for the safety car, virtual safety car, red flag, and race duration data. These attributes were transformed to compute total number of occurrences of the each of the attributes during an F1 race. The transformations were computed by calculating the sum of instances in the data, and in the case of the virtual safety car total laps, by subtracting attributes. The transformations are explained in more detail in the Data section. The transformations were computed to get more detailed information about the occurrences of safety cars, virtual safety cars, and red flags during the races.

The data for the project contained 123 attributes and 348 instances. The weather data that is present within the dataset is transformed into 33 weather attributes for each of the 6 hours for the chosen race period from the 66 columns they were originally. This data is useful as it has been selected for a specific period around the race start time, getting some values for the weather just before the race, at the beginning and at the end of the

race. The race data that has been obtained is not broken down by time, the only true value that is provided is the scheduled start time of the race. Because of this, we do not know at what time during the race was the safety car, virtual safety car, or red flag deployed. We also don't know what time the fastest lap times were recorded at during the race, and therefore having the weather data if split into specific hours is not providing detail that can be attributed to a specific occurrence during the race. Therefore, it was decided to compute min, max, and mean values for each of the 11 unique weather statistics, with the intention of dropping the original 66 weather attributes from the data.

Ordinal Encoding was computed for the categorical Formula One statistics, three of these attributes were transformed using ordinal encoding. These attributes went through ordinal encoding instead or one hot encoding because they were being used as the dependent variables in the modelling and had more than two values. It is recommended that that dependent variables with three or more be encoded with ordinal coding, and one hot encoding be used where two variables are present in the attribute (McCaffrey, 2020). The three attributes that were encoded were: 'vsc_convertedFullSafetyCar', 'rf_status', and 'rf_desc'.

Variance Inflation factor (VIF) is used to provide a measure of multicollinearity among the independent variables of a regression model (The Investopedia Team, 2023). VIF was calculated for each of the weather attributes to determine the multicollinearity between the attributes. Scatter matrices were produced, and the correlation matrix was obtained for the weather attributes. The collinear variables were dropped from the data frame and the remaining 15 attributes for the weather were kept to be used in the data mining. The remaining weather attributes are displayed in Table 24.

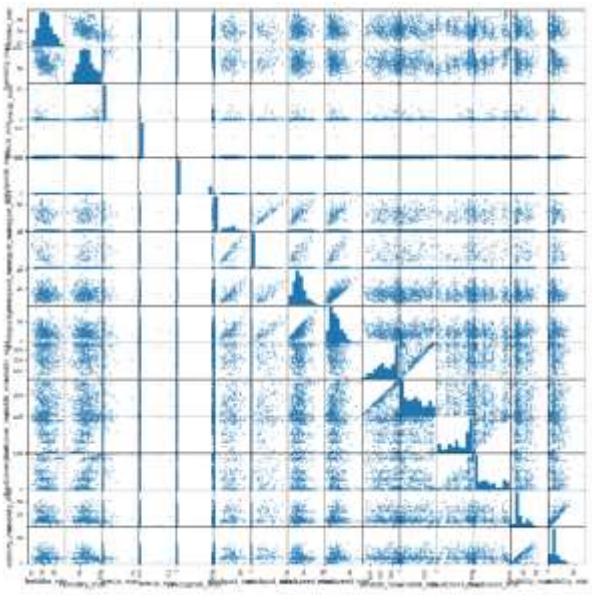


Figure 26: Scatter Matrix for Weather Attributes

Table 24: VIF of remaining weather attributes

	feature	VIF
0	feelslike_min	11.191634
1	humidity_max	19.559063
2	precip_max	1.441852
3	precip_min	1.085617
4	precipprob_max	1.976891
5	windgust_max	2.807709
6	windgust_min	2.306858
7	windspeed_max	18.516160
8	windspeed_min	11.944767
9	winddir_max	9.680285
10	winddir_min	4.044061
11	cloudcover_max	10.724653
12	cloudcover_min	4.328631
13	visibility_max	7.068942
14	visibility_min	6.186480

4.4. Data Mining

Data mining involves applying data mining techniques and algorithms to the data to extract useful information and insights (Rajput, n.d.). Using inspiration from the state-of-the-art analysis, it was decided to use regression models on the data for the project. These models were completed for each of the 10 dependent variables. Models were run with the chosen 15 weather features and F1 predictor variable.

1. *Linear Regression*

   Linear regression is a type of regression method in data mining. This regression method used a straight line to form a link between the predictor variables and the defined dependent variable (Varshney, 2022). Linear regression is used to see whether the predictor variables do a good job at predicting the dependent variable (IBM, n.d.).

2. *Lasso Regression*

   Lasso regression is a regularization technique. It is used to make more accurate predictions using shrinkage (Kumar, 2023). LASSO stands for Least Absolute Shrinkage and Selection Operator. Shrinkage is used in models that have high levels of multicollinearity, as the shrinkage causes the data values to be shrunk towards a central point to the mean.

3. *Ridge Regression*

   Like Lasso regression, Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization to deal with values that suffer from multicollinearity. When multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values (Great Learning Team, 2022). Ridge Regression shrinks the parameters; therefore, it is mostly used to prevent multicollinearity (Jain, 2023).

4. *Random Forest Regression*

   Random forest is a Supervised Machine Learning Algorithm that is used widely in Regression problems. It builds decision trees on different samples and takes their average for regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression (E R, 2023).

5. *Random forest – using grid search*

   Grid search is a hyperparameter tuning technique that involves searching for the optimal combination of hyperparameters by evaluating the performance of the algorithm on a validation set. The technique involves specifying a grid of hyperparameter values to be searched over and then evaluating the performance of the algorithm for each combination of hyperparameters on the validation set. The combination of hyperparameters that yields the best performance on the validation set is then chosen as the optimal hyperparameters for the algorithm (Collins, 2021).

4.5. Interpretation

Interpretation within KDD involves analysing the results and extracting knowledge from the data. The interpretation of the results will be analysed in more detail in Results 6.0, The results will be analysed with a number of graphs of the models, and by examining and interpreting the results of the graphs, to make broad assumptions of the results of the models.

4.6. Evaluation

Evaluation of the models will be conducted with R-squared. R-squared is the coefficient of determination which measures how well a statistical model predicts an outcome. The outcome is represented by the model's dependent variable (Turney, 2022). R-squared will be used to evaluate whether the extracted knowledge from the models is useful and meaningful. The best possible score for R-squared is 1.0 with the potential to be negative as the model can be arbitrarily worse (Scikit Learn). In general, the coefficient of determination normally ranges between 0 and 1. R-squared is said to be a measure of goodness of fit and is the proportion of variance in the dependent variable that is explained by the model (Turney, 2022). In this project, the R-squared will be describing the proportion of variance in the F1 statistics that can be explained by the weather data in the model. R2 is frequently used as a performance metric in regression tasks predicting continuous outcomes.

4.7. Deployment

Deployment is the final phase in the KDD methodology, which involves using the discovered knowledge. For the deployment phase of the KDD methodology a poster will be produced, with a highlight and summary of the findings of the project.

5.0. Analysis

The State of the Art research and analysis provided valuable insight into the types of tools and technologies that were used across different types of analyses. It was used to influence the chosen methods on the data for the project analysis. From Table 2, Table 3 and Table 4 the data science methods were outlined for the relevant studies. Across the different mediums that were analysed: research papers, videos, and blog posts, the most popular data analysis methods were visualisations and statistics. Modelling was the third most popular method employed across some of these studies. Visualisations and interactive dashboards were very popular among the studies, with a number of different tools being utilized for analysing the data.

The first approach that was inspired by the State of the Art analysis that was used in the project were computing statistics. Summary statistics, in the form of descriptive statistics were computed for the datasets when compiling the data in 3.0. The descriptive statistics allowed insights to be obtained, and decisions be made about the data that was present within the dataframes. This analysis discovered the missing values within the data and allowed the handling methods for the data be decided. The missing values were dealt with in 4.2 where it was decided from the summary statistics, to not drop, or replace the missing values with computed values (i.e., minimum, maximum, and average, etc), but deal with the

missing values by replacing them with ‘0’ or in the context of the categorical attributes, ‘none’, or ‘No VSC’, etc.

The next approach that was used for the analysis of the project were static visualisations and an interactive dashboard. The static visualisations were deployed using some of the tools and technologies outlined in Technology Matplotlib, Seaborn, Excel, Visme and Plotly Dash were all visualisation tools used for analysis at different stages of the project. The initial visualisations that were created for the project were used in the processing of the data in Excel, this analysis attempted to discover the attributes that were not as important to include in the overall study and therefore were transformed pr removed from the data. The analysis was completed using pivot tables to determine what values were present within each of the csv data files. Noisy data was removed from the data based on the findings from the pivot tables and the

The visualisations created using matplotlib and seaborn were created to analyse more of the data in the 3.4 Exploratory Data Analytics. These visualisations resulted in insights being obtained from the data. And an analysis into the type of data that could be observed from the results i.e., before the weather data being obtained, the analysis into the F1 data described rain the highest reasons for red flagged races to not resume. This provided confirmation that the weather features that were being sourced needed to include rain/ precipitation statistics.

For the project, the dashboard was computed for analysis as described in 3.4 Exploratory Data Analytics. Plotly Dash was used for the creation of the dashboard for the analysis into the weather and F1 statistics. Dash was a new tool and was learned for the purpose of the project. It provided interesting perspective on the results of precipitation and recorded fastest lap times at tracks, described in Figure 24. For example, the Hungaroring circuit had the second highest value for precipitation during the race in 2014, but the slowest lap recorded between 2005 and 2022. In contrast the 2021 race at the Hungaroring had the most amount of precipitation recorded, but the third fastest lap recorded at the circuit during an f1 race. Showing that even with high levels of precipitation, the lap times will not necessarily be impacted.

The visualisation available on the dashboard provided insights into the overview of the data before conducting modelling on the data. Dash had a wide variety of resources online which helped with creating the dashboard, and then adding the formatting and styling. The dashboard was an effective analysis tool, which allowed a number of graphs and statistics be produced in a simple readable nature, that can help with the users understanding of the data. It also helped with reducing the number of visualisations that needed to be included in the report, as each visualisation was available for all 37 racetracks in one place.

The types of model computed for the analysis are outlined in the Data Mining which detailed each of the types of models that were run on the data. There were 5 types of models ran for each of the F1 statistics being replaced as the dependent variable**.** Regression was a type of model that was run on a number of studies that were used in the State of the Art analysis. Scikit learn was used for the implementation of the models. The data was first split using a train

test split of 80% training and 20% testing. The Train test split technique is used to estimate the performance of machine learning algorithms which are used to make predictions on data not used to train the model (Narang, 2022).

As described in Data Mining there were five variations of regression models that were computed for the project. These models were: Linear Regression, Lasso Regression, Ridge Regression, Random Forest, and Random Forest with Grid Search.

The approaches to carrying out each of the models were completed by creating a machine learning pipeline for each of the formula one statistics, so each of the models were run off of the same data using the same X, and y values. Machine learning pipelines can be created by putting together a sequence of steps involved in training a machine learning model. It can be used to automate a machine learning workflow. The pipeline can involve pre-processing, feature selection, classification/regression, and post-processing (Saeed, 2021). Machine learning pipelines were utilised, so that the models between the formula one statistics could be compared easily with the results from the other models.

Each of the pipelines were defined for the models using Scikitlearn modules for linear models (Scikitlearn, n.d.), pre-processing steps, and standardization techniques could be employed within the pipelines. The models were run with different methods employed. The first group of methods included standardization of the data and also had the fitting the y-intercept set to true. The second time around the data was not scaled, and the fitting the y-intercept was set to false. The results are compared in section 6 below.

The hyperparameters were defined for the grid search random forest model. The hyperparameters were defined for the n estimators the max depth, and the min samples split. Seaborn and matplotlib were utilized for graphing the results of the models. And summary tables of results were produced in Excel. The accuracy of the models was computed using R-squared values.

The F1 statistics were chosen as the 10 dependent variables in order to get the results of predicting the variables based on the selected weather attributes. The weather data attributes that were selected for the models were computed using vif feature selection as described in Transformation above.

The virtual safety car models were completed on a smaller subset of the data from 2015 to 2022, as the virtual safety cars were only introduced into the sport in 2015, and therefore would not be analysed in respect to the whole dataset.

6.0. Results

The models were run, and the results have been presented in a number of figures and tables below. The first batch of models were computed after being standardized. Scikitlearn was used to implement the standardization class. Centring and scaling occur independently on each of the weather attributes by computing the relevant statistics on the samples in the training set (Scikit Learn, n.d.). The standard scaler was fit to the data first before being transformed. The ‘fit_intercept’ parameter was set to True for this selection of models. The intercept of the model was set to true, so that an intercept would be calculated within the

model. Setting the intercept means specifying the value for the intercept for the equation. In some analysis, the regression model only becomes significant when the intercept is removed, and the regression line reduces to Y = bX + error (Sukhadeve, 2016).

The absolute values of the R-squared are presented for the standardized data in Table 25. The values for R-squared have been calculated and the results multiplied by 100 to turn the values into percentages.

Table 25: R-squared absolute values (scaled, and fit_intercept=True)

	Linear Regression	**Lasso**	**Ridge**	**Random Forest**	**Random Forest – Grid Search**
sc_totalLaps	22	0.63	21.43	22.03	19.29
sc_count	6.21	0.47	6.14	3.37	2.57
vsc_convertedFullSafetyCar	4.4	7.1	4.32	17.13	11.75
vsc_totalLapsCombined	10.91	7.23	10.84	6.07	5.29
vsc_count	2.3	3.57	2.25	5.95	4.24
rf_status	9.56	0.24	9.47	11.13	13.89
rf_desc	30.03	4.71	29.93	26.57	26.62
rf_count	20.2	0.06	20.07	19.88	27.12
raceDuration_laps	24.65	0	23.85	38.36	38.52
fast_lap_1	0.07	0.11	0.25	8.68	9.05

R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. Having an R-squared of 100% means that the predictors are able to explain completely the occurrence of the dependent variable (Fernando, 2023). Meaning that if any result of 100% was obtained for one of the F1 attributes, then that attribute would be able to be fully explained by the 15 weather features that are used in the models. The interpretation of R-squared is dependent on the context of the problem.

From the Table 25 it can be seen that the 'raceDuration_laps' has both the highest value (38.52%) and the lowest value (0%) for R-squared within the scaled data. The results that have been produced from the models have an overall average absolute R-squared value of approximately 12%. The random forest, and random forest grid search were the two best performing models with an average value of 15.8% across the two sets of models on the F1 dependent variables. Lasso was the worst performing model with an average value of 2.4% across the models on the standardized models.

The absolute values of the R-squared are presented for the data that was not standardized and fit_intercept was set to false are presented in Table 26.

Table 26: R-squared absolute values (not scaled, and fit_intercept=False)

	Linear Regression	**Lasso**	**Ridge**	**Random Forest**	**Random Forest – Grid Search**
sc_totalLaps	21.67	4.97	3.32	19.79	19.5
sc_count	6.46	1.22	10.06	3.38	2.56
vsc_convertedFullSafetyCar	3.72	18.42	6.1	13.77	10.06

vsc_totalLapsCombined	8.85	4.64	7.02	6	5.18
vsc_count	1.28	0.7	3.27	6.01	3.95
rf_status	9.32	1.48	8.77	11.24	13.89
rf_desc	26.16	1.76	26.3	26.57	26.78
rf_count	20.19	1.59	21.7	20.07	27.32
raceDuration_laps	95.71	62.66	95.84	38.29	38.44
fast_lap_1	36.23	36.38	41.45	8.94	9.06

From Table 26 the R-squared results for the data that was not scaled, and the models that fit_intercept was set to False, are displayed. The lowest values reported for the models occurred for the 'vsc_count' with 0.7%, and the highest value like with the standardized data was seen from the 'raceDuration_laps' with approximately 95.84% . The average absolute value for R-squared improved after not computing the standardization and setting the fit_intercept to False, increasing to approximately 18%. The linear regression and the ridge regression models were the two best performing models using this data. They had an average of approximately 22.6% across the two sets of models. The lasso model was still the lowest performing model over the two groups, however with the non-standardized data the average for the lasso model was 13.38% as opposed to the 2.4% that it was achieving with the standardized data.

Along with the table of results for each of the groups of models, regression plots were produced for each of the models that were run. Presented in Figure 27 and Figure 28 are a sample of the plots that were created with the use of matplotlib and seaborn. The plots depict the actual and predicted values of the models, along with the predicted results and the residuals. The residuals are the difference between the actual value and the value predicted by the model for any given point (Khan Academy, n.d.).

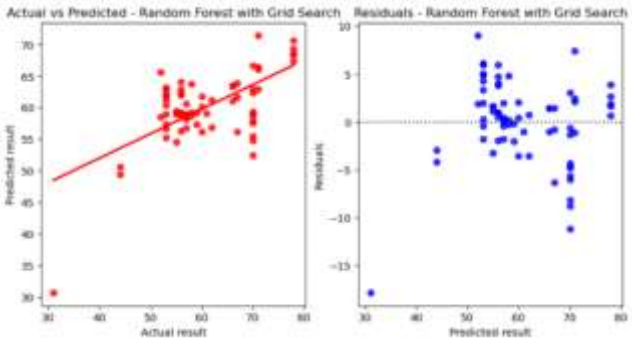


Figure 27: Actual and Predicted values of models (random forest - grid search: raceDuration_laps)

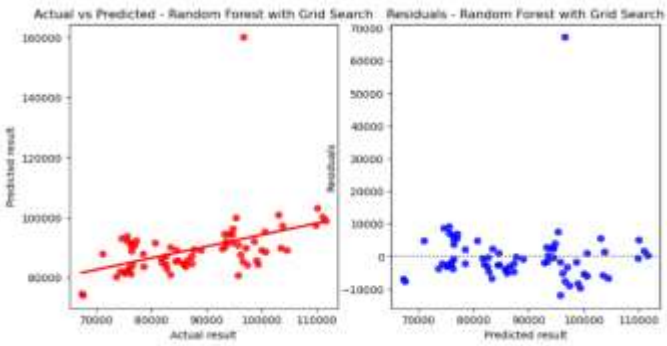


Figure 28: Actual and Predicted values of models (random forest - grid search: fast_lap_1)

The average values for R-squared from both tables were not very large percentages. A rule of thumb for small values of R-squared is that the fraction by which the standard deviation of the errors is less than the standard deviation of the dependent variable is approximately one-half of R-squared (Nau, n.d.). So, the errors for the two groups of models are approximately 6% and 8% respectively.

Determining how big of a R-squared value is good enough depends on the situation the results are being used in. It will depend on the objectives and the needs of the study. There is no research out there in the chosen area of analysis and therefore there is no way to compare what is expected for the results as there are no published literature studies available with an accuracy level to base the results on.

7.0. Conclusions

The aim of the project was to complete an analysis into the impact that weather has on Formula One races. Through State of the Art analysis it was discovered that the project was a unique idea, and the implementation had not been carried out for research purposes before. Because of this, a lot of the project was focused on creating the data set that would be needed in order to carry out the project aims.

The data was compiled from a number of sources and extensive exploratory analysis was performed. New tools and technologies were implemented in order to investigate the insights that could be derived from the weather and Formula One data. The available data spanned from the 2005 grand prix to the end of the 2022 season, which consisted of 348 races, which spanned across 37 racetracks, around the world. The weather data that was combined with the F1 data consisted of 6 hours' worth of weather statistics that spanned around the scheduled race time in local time.

Visualisation tools and technologies were discovered and implemented to interpret the relationship between weather and Formula One race statistics. Contradicting results were found from the visualisation results. Example being the Hungaroring track, in the 2014 race the second highest value for precipitation was recorded during the race, but the slowest lap was recorded on the fastest lap line chart. In contrast the 2021 race at the Hungaroring had the most amount of precipitation recorded, but the third fastest lap recorded at the circuit during an f1 race.

An advantage to this study is the interactive tool produced to provide insights at the touch of a button into the weather statistics during a race, allowing users to select a year, country, and racetrack to view the f1 results and the corresponding weather statistics for that race. A feature also allows users to view all races from that circuit over the years and presents findings on the average precipitation along with the fastest lap times as discussed above.

Using machine learning was another advantage of the study. There were two groups of models that were performed on the same predictor and dependent variables. In one scenario, the data was scaled before modelling, and the fitting of the intercept was set to true. In the

remaining scenario, the data was not scaled, and the intercept was set to false. There was a lot of variation between the results that were produced from the models.

A disadvantage of the study was the limitations of the accurate Formula One data that was able to be obtained. The local time the race was scheduled for being the main cause of inaccuracies within the data. If accurate time data was able to have been sourced for each of the races before 2005, then the rest of the Formula One data would have been able to be included in the study. Bringing the total number of attributes from 348 to 1,102. Having a larger amount of data within the dataset may increase accuracy within the models.

8.0. Further Development or Research

With additional time and resources, the project could take a direction that would allow more years to be used in the data with the hopes of increasing the accuracy of the predictions made by the models. The period of data that was selected for the project was suitable as that was the data that was available to source for the project.

With additional time and resources on this project, the goal would be to improve the prediction model. More approaches would need to be tested to uncover a more accurate prediction, to evaluate the impact of weather on the formula one races.

Perhaps the having more data within the dataset would produce more comprehensive results, by analysing not only the formula one races, but look too at the free practices, qualifying, and sprint races in a formula one race weekend.

9.0. References

AnyChart Team. (2022, May 6). *Visualizing Data on F1, Indian Ocean, Oreo, Conflicts in Legal Amazon — DataViz Weekly.* Retrieved October 27, 2022, from https://www.anychart.com/blog/2022/05/06/visualizing-data-f1-indian-ocean-oreo-conflicts/

Bell, A., Smith, J., Sabel, C. E., & Jones, K. (2016, June). *Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014.* Retrieved October 27, 2022, from https://www.researchgate.net/publication/274080402_Formula_for_success_Multilevel_modelling_of_Formula_One_Driver_and_Constructor_performance_1950-2014

Chalmers, D. (2008, August 18). *How Could F1 Improve Its Calendar?* Retrieved October 27, 2022, from https://bleacherreport.com/articles/48779-how-could-f1-improve-its-calendar

Chiu, N. (2021, September 1). *The real reason why F1 struggles to race in torrential rain conditions.* Retrieved October 27, 2022, from https://racingnews365.com/the-real-reason-why-f1-struggles-to-race-in-torrential-rain-conditions

Coleman, M. (2023, April 28). *How F1 sprint races work — and how the new format will shake things up in Baku*. Retrieved from The Athletic: https://theathletic.com/4460164/2023/04/28/how-formula-one-sprint-races-work/

Collantine, K. (2022, October 9). *Japanese Grand Prix halted after two laps in heavy rain.* Retrieved October 27, 2022, from https://www.racefans.net/2022/10/09/japanese-grand-prix-halted-after-two-laps-in-heavy-rain/

Collins, A. (2021, March 18). *Using Grid Search to Optimize Hyperparameters.* Retrieved from Section: https://www.section.io/engineering-education/grid-search/#:~:text=Grid%20search%20refers%20to%20a,for%20each%20combination%20of%20hyperparameters.

Cooper, A. (2021, August 31). *Gasly: F1 should address spray issue after poor visibility at Spa.* Retrieved from motorsport.com: https://www.motorsport.com/f1/news/gasly-spray-issue-poor-visibility/6657654/#:~:text=Visibility%20rather%20than%20potential%20loss,running%20at%20safety%20car%20speeds.

Dawkins, S., & Stern, H. (n.d.). *MANAGING WEATHER RISK DURING MAJOR SPORTING EVENTS:.* Retrieved October 27, 2022, from https://d1wqtxts1xzle7.cloudfront.net/42812717/Advances_in_Tourism_Climatology20160218-6411-1xiulgz-with-cover-page-v2.pdf?Expires=1667202474&Signature=HDW2rOgR8j5f4idkhG6BNwcxpUQzubP6JJd3koOuQjseg85BVXPJqvASUSWw73-FxPe8nd4rdI-cUdOL7qmEii6yKDqrroFylL8sR00

Driscoll, M. (n.d.). *Jupyter Notebook: An Introduction.* Retrieved from Real Python: https://realpython.com/jupyter-notebook-introduction/

E R, S. (2023, April 26). *Understand Random Forest Algorithms With Examples (Updated 2023).* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Emms, N. (2022, October 3). *CLOCK'S TICKING Why does today's F1 Grand Prix have a countdown clock instead of lap counter?* Retrieved from The Irish Sun: https://www.thesun.ie/sport/9505694/why-does-todays-f1-grand-prix-countdown-clock/

Ergast Developer API. (n.d.). *Ergast Developer API*. Retrieved from API Documentation: http://ergast.com/mrd/

Ergast Developer API. (n.d.). *Terms & Conditions*. Retrieved from Ergast Developer API: http://ergast.com/mrd/terms/

F1 Analysis. (n.d.). *F1 Analysis.* Retrieved from F1 Analysis: https://f1-analysis.com/2021/07/08/who-was-the-best-f1-driver-each-year/

F1 Wiki. (n.d.). *Formula 1 Wiki:About*. Retrieved from F1 Wiki: https://f1.fandom.com/wiki/Formula_1_Wiki:About

F1 Wiki. (n.d.). *Home.* Retrieved from F1 Wiki: https://f1.fandom.com/wiki/Formula_1_Wiki#:~:text=The%20Formula%201%20Wiki%20is,of%20size%2C%20detail%20and%20relevance

F1 Wiki. (n.d.). *Red-flagged races.* Retrieved from F1 Wiki: https://f1.fandom.com/wiki/Red-flagged_races

F1 Wiki. (n.d.). *Safety Car.* Retrieved from F1 Wiki: https://f1.fandom.com/wiki/Safety_Car

Fairley, S., Tyler, D. B., Kellett, P., & D'Elia, K. (2011, May). *The Formula One Australian Grand Prix: Exploring the triple bottom line.* Retrieved October 27, 2022, from https://www.researchgate.net/publication/227425409_The_Formula_One_Australian_Grand_Prix_Exploring_the_triple_bottom_line

Fernando, J. (2023, April 8). *R-Squared: Definition, Calculation Formula, Uses, and Limitations.* Retrieved from Investopedia: https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20values%20range%20from,variable%20you%20are%20interested%20in

Formula 1 Trends. (n.d.). *https://www.f1trends.com/*. Retrieved October 27, 2022, from https://www.f1trends.com/

Formula 1 Wiki. (n.d.). *Home*. Retrieved from Formula 1 Wiki: https://f1.fandom.com/wiki/Formula_1_Wiki

Formula One. (2016, September 28). *Watching the skies - how weather forecasting works in F1.* Retrieved October 27, 2022, from https://www.formula1.com/en/latest/features/2016/9/how-weather-forecasting-works-in-f1.html#:~:text=%E2%80%9CVia%20a%20web%20portal%20we,will%20fall%20and%20so%20on.%E2%80%9D

Formula One. (2022, June 27). *F1 continues push to hit Net Zero Carbon by 2030 target*. (Formula One) Retrieved October 27, 2022, from https://www.formula1.com/en/latest/article.f1-continues-push-to-hit-net-zero-carbon-by-2030-target.7fGtPCNCwOnMFFt9Ys1HAa.html

Formula1. (2021, August 29). *Verstappen takes win and Russell first podium in shortest ever Grand Prix as heavy rain hits Spa.* Retrieved from Formula1: https://www.formula1.com/en/latest/article.verstappen-takes-victory-in-severely-shortened-rain-affected-belgian-gp-as.4AqGhiKQfFaqr7KZjyDjPZ.html

George, W. (2021, August 6). *Formula 1 Championship Predictor.* Retrieved from Medium: https://medium.com/@willgeorge93/formula-1-championship-predictor-a-machine-learning-solution-a86efcb9298

Great Learning Team. (2022, November 16). *What is Ridge Regression?* Retrieved from Great Learning: https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values

Horton, P. (2021, November 8). *F1 Officials Determined to Find Safer Ways of Racing in the Rain.* Retrieved October 27, 2022, from https://www.autoweek.com/racing/formula-1/a38192185/f1-officials-safer-racing-rain/

Horton, P. (2021). *FIA reduces maximum F1 race time to three hours.* Retrieved from Motorsport week: https://www.motorsportweek.com/2020/12/17/fia-reduces-maximum-f1-race-time-to-three-hours/

IBM. (n.d.). *What is Linear Regression?* Retrieved from https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

Jain, S. (2023, May 2). *Lasso & Ridge Regression | A Comprehensive Guide in Python & R (Updated 2023).* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/#13

Jupyter. (2023). *Jupyter Notebook: The Classic Notebook Interface.* Retrieved from Jupyter: https://jupyter.org/

Khan Academy. (n.d.). *Introduction to residuals and least-squares regression.* Retrieved from Khan Academy: https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/xfb5d8e68:residuals/v/regression-residual-intro#:~:text=In%20linear%20regression%2C%20a%20residual,sum%20of%20the%20squared%20residuals.

Khorounzhiy, V. (2021, September 28). *FIA COULD ROUTINELY CHANGE F1 GP TIMES TO AVOID HEAVY RAIN.* Retrieved October 27, 2022, from https://the-race.com/formula-1/fia-could-routinely-change-f1-race-start-times-to-avoid-rain/

Knechtle, B., Di Gangi, S., Rust, C. A., Villiger, E., Rosemann, T., & Nikolaidis, T. P. (2019). *The role of weather conditions on running performance in the Boston Marathon from 1972 to 2018.* Retrieved October 27, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6407773/

Kumar, D. (2023, January 12). *A Complete understanding of LASSO Regression.* Retrieved from Great Learning: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters

Kuosmanen, V. (2020, September 30). *Predicting Formula 1 results with Elo Ratings.* Retrieved from Towards Data Science: https://towardsdatascience.com/predicting-formula-1-results-with-elo-ratings-908470694c9c

Łusiak, M. (2021, May 5). *Predicting F1 race strategy using ML.NET – Part 1 – Bahrain.* Retrieved from mlusiak: https://mlusiak.com/f1-race-strategy-ml-net-bahrain/

Masefield, F. (2013, August 28). *Predicting the Weather for a Grand Prix: How Formula 1 Teams Plan for Rain.* Retrieved October 27, 2022, from https://bleacherreport.com/articles/1752419-predicting-the-weather-for-agrand-prix-how-f1-teams-plan-for-rain#:~:text=Teams%20are%20accurately%20able%20to,via%20an%20internal%20F1%20intranet.

Matplotlib. (n.d.). *Matplotlib logo*. Retrieved from Matplotlib: https://matplotlib.org/stable/gallery/misc/logos2.html

McCaffrey, J. (2020, December 8). *Data Prep for Machine Learning: Encoding.* Retrieved from Visual Studio Magazine: https://visualstudiomagazine.com/articles/2020/08/12/ml-data-prep-encoding.aspx#:~:text=A%20dependent%20variable%20that%20has,%3D%201%2C%20green%20%3D2

Media. (2022, June 16). *What Impact Does The Weather Have On F1?* Retrieved October 27, 2022, from https://f1chronicle.com/what-impact-does-the-weather-have-on-f1/

MetMatters. (2011, September 22). *F1 and the weather.* Retrieved October 27, 2022, from https://www.rmets.org/metmatters/f1-and-weather

Millward, A. (2022, October 3). *DOES F1 FEAR RAIN? Opinions on the Lack of Wet Weather Racing.* Retrieved October 27, 2022, from https://www.youtube.com/watch?v=7XDvttk4stY&ab_channel=AidanMillward

Min Kil, K., Suk-Kyu, K., Jae-Ahm, P., Michael, C., Jae-Gu, Y., & Kyunga, N. (2016, April 28). *Measuring the economic impacts of major sports events: the case of Formula One Grand Prix (F1).* Retrieved October 27, 2022, from https://www.tandfonline.com/doi/abs/10.1080/10941665.2016.1176061?journalCode=rapt20

Mitchell, B. (2023, February 20). *What Is Pandas In Python? A Guide For Beginners.* Retrieved from Coding Dojo Blog: https://www.codingdojo.com/blog/pandas-python-guide

Mitchell-Malm, S. (2021, December 22). *BELGIAN GP F1 FANS GET COMPENSATION – BUT IS IT ENOUGH?* Retrieved from The Race: https://the-race.com/formula-1/belgian-gp-f1-fans-get-compensation-but-is-it-enough/

Motorsport Stats. (n.d.). *About Us*. Retrieved from Motorsport Stats: https://motorsportstats.com/business/about-us

Motorsport Stats. (n.d.). *Motorsport Stats*. Retrieved from Motorsport Stats: https://motorsportstats.com/

Motorsport Stats. (n.d.). *RESULTS*. Retrieved from Motorsport Stats: https://motorsportstats.com/results

Narang, M. (2022, December 13). *Train test split.* Retrieved from Shiksha Online: https://www.shiksha.com/online-courses/articles/train-test-split/

Nau, R. (n.d.). *What's a good value for R-squared?* Retrieved from Statistical forecasting:: https://people.duke.edu/~rnau/rsquared.htm

Nigro, V. (2020, June 11). *Formula 1 Race Predictor.* Retrieved October 27, 2022, from https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da

NumPy. (2023). *PRESS KIT.* Retrieved from NumPy: https://numpy.org/press-kit/

Opong, R. (2022, February 24). *Has F1 Ever Raced In Snow? Could It Happen?* Retrieved from Flow racers: https://flowracers.com/blog/has-f1-raced-in-snow/

Pandas. (2023). *Citing and logo.* Retrieved from Pandas: https://pandas.pydata.org/about/citing.html

Paul, J. J. (n.d.). *Formula 1 Data Vis.* Retrieved October 27, 2022, from https://jasonjpaul.squarespace.com/formula-1-data-vis

plotly. (n.d.). *Logos in Python/v3*. Retrieved from plotly: https://plotly.com/python/v3/logos/

Python Software Foundation. (2023). *The Python Logo.* Retrieved from Python: https://www.python.org/community/logos/

Rajput, A. (n.d.). *KDD Process in Data Mining.* Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Saeed, M. (2021, October 22). *Modeling Pipeline Optimization With scikit-learn.* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/modeling-pipeline-optimization-with-scikit-learn/#:~:text=A%20machine%20learning%20pipeline%20can,regression%2C%20and%20post%2Dprocessing

Scikit Learn. (n.d.). *sklearn.preprocessing.StandardScaler.* Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Scikit Learn. (n.d.). *sklearn.metrics.r2_score.* Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

Scikitlearn. (n.d.). *API Reference.* Retrieved from Scikitlearn: https://scikit-learn.org/stable/modules/classes.html

seaborn. (n.d.). *Citing and logo*. Retrieved from seaborn: https://seaborn.pydata.org/citing.html

Sicoie, H. (2022, January 14). *Machine Learning Framework for Formula 1 Race Winner and Champtionship Standings Predictor.* Retrieved from Arno Academic Publications Online: http://arno.uvt.nl/show.cgi?fid=157635

Simplilearn. (2022, April 11). *Formula 1 Data Analysis Using Python 2022 | 2022 Formula 1 Data Analysis Project | Simplilearn.* Retrieved October 27, 2022, from https://www.youtube.com/watch?v=7v9puRHfelw&ab_channel=Simplilearn

Stavropoulos, V. (2018, May 18). *A Thorough Analysis of the Pit Stop Strategy in Formula 1.* Retrieved from Statathlon: https://statathlon.com/analysis-of-the-pit-stop-strategy-in-f1/

StickPNG. (n.d.). *Download Visme logo transparent PNG*. Retrieved from StickPNG: https://www.stickpng.com/img/icons-logos-emojis/tech-companies/visme-logo

Stoppels, E. (2017). *Predicting Race Results using.* Retrieved from UNIVERSITY OF TWENTE STUDENT THESES: https://essay.utwente.nl/74765/1/FinalThesisEloyStoppelsNoCompany.pdf

Sukhadeve, A. (2016, July 10). *Why is intercept important in regression analysis?* Retrieved from Analytics Insight: https://www.analyticsinsight.net/why-is-intercept-important-in-regression-analysis/#:~:text=The%20Importance%20of%20Intercept%20The,to%20Y%20%3D%20bX%20%2B%20error.

The Investopedia Team. (2023, February 12). *Variance Inflation Factor (VIF).* Retrieved from Investopedia: https://www.investopedia.com/terms/v/variance-inflation-factor.asp

Thornes, J. (1977, July). *The Effect of Weather on Sport.* Retrieved October 27, 2022, from https://www.researchgate.net/profile/John-Thornes/publication/260819737_The_Effect_of_Weather_on_Sport/links/5e1b1aeaa6fdcc28376cf986/The-Effect-of-Weather-on-Sport.pdf

Turney, S. (2022, September 14). *Coefficient of Determination (R^2) | Calculation & Interpretation.* Retrieved from Scribbr: https://www.scribbr.com/statistics/coefficient-of-determination/

Varshney, H. (2022, May 27). *Regression Method in Data Mining Simplified 101.* Retrieved from HEVO: https://hevodata.com/learn/regression-method-in-data-mining/#:~:text=Linear%20Regression%20is%20a%20type,represented%20by%20the%20given%20equation

Visual Crossing. (2021, May 28). *Academic, student and research discounts for weather data*. Retrieved from Visual Crossing: https://www.visualcrossing.com/resources/documentation/weather-data/academic-student-and-research-discounts-for-weather-data/

Visual Crossing. (n.d.). *Weather Data Documentation.* Retrieved from VISUAL CROSSING WEATHER: https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/

Visual Crossing. (n.d.). *Weather Data & API*. Retrieved from Visual Crossing: https://www.visualcrossing.com/

Wang, J. (2020, January 20). *QlikView Visualization of Formula 1 (F1) Relational Data Model.* Retrieved October 27, 2022, from https://towardsdatascience.com/qlikview-visualization-of-formula-1-f1-relational-data-model-82e8b46c9f71

Wang, M. (2022, April 18). *11 fascinating graphs on Formula 1.* Retrieved October 27, 2022, from https://medium.com/visual-analytics-field-notes/11-fascinating-graphs-on-formula-1-acd05bcd3e73

Whitehead, J. C., & Wicker, P. (2020, July). *The effects of training satisfaction and weather on revisiting sport events and their monetary value: The role of attribute non-attendance.* Retrieved October 27, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S2211973620300805

Wikimedia Commons. (2020, September 17). *File:Scikit learn logo small.svg.* Retrieved from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Scikit_learn_logo_small.svg

Wikimedia Commons. (2023, May 3). *File:Microsoft Office Excel (2019–present).svg*. Retrieved from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Microsoft_Office_Excel_(2019%E2%80%93present).svg

Yıldırım, S. (2022, July 18). *Benefits of Learning Python.* Retrieved from LearnPython.com: https://learnpython.com/blog/benefits-of-learning-python/

10.0. Appendices

10.1. Project Proposal

National College of Ireland

Project Proposal

An Analysis into the Effects Weather has on Formula One Races and the Changes Needed to Create the Optimal Race Calendar

31st October 2022

BSc (Honours) Data Science

BSHDS4

2022/2023

Niamh Daly

x19370553

x19370553@student.ncirl.ie

Contents

1.0 Objectives

The objectives of the project are to:

1. Analyse the effect that weather/climate has on the Formula One races within calendar seasons
    a. Getting the statistics for Formula One races under different weather conditions and analysing
        i. Difference in lap times
        ii. Difference in safety car usage
        iii. Difference in safety flag usage
        iv. Difference in race length completed (e.g race not completing the total distance/ number of laps)
        v. Difference in race result

2. Create a race calendar that best suits the weather effects of different regions
    a. Use weather data to determine when the races on a typical race calendar are most suitable (using the results from section 1 above to aid this analysis)

3. Compare different variations of the race calendar in terms of Formula One net zero Carbon goals:
    a. Based on location (distance from one racetrack to another)
    b. Based on climate/weather (results from section 2 above)

4. Create/utilize a website to hold interactive dashboards for the resulting analyses

2.0 Background

I have always had an interest in Formula 1, from the very first essay I completed in primary school about my favourite driver, to my junior certificate art project in secondary school.

I have wanted to have a career in data science because of my interest in Formula 1.

Formula one is a data driven sport, with data constantly being collected and manipulated to make fast paced decisions to get the best results.

A problem that Formula 1 has been facing in recent years is the issue with race stoppages/ inability to complete full race distance base on the weather. For example, the 2021 Belgian Grand Prix, whereby the race was scheduled to run 44 laps, and ended up being concluded after 3 laps due to the wet conditions (Wikipedia, 2022). It would seem that the weather is having a greater impact on the sport and its ability to perform, and I want to analyse this to determine how much of an impact the weather has on Formula 1 as a sport.

If we were to look at Formula One by itself irrespective of the sport sponsorships, obligations to hold specific races at specific times, and economic impacts. Would it be possible to create the ideal Formula 1 race calendar that utilizes trends in weather data to hold the race/ event weekend at a time when the weather is most suitable to get the best

out of the racing event. Would my solutions/results fit into the sports goals to become Net Zero Carbon by 2030 (Formula One, 2022).

3.0 State of the Art

As Formula One is a data driven sport there is a lot of information out there for people to complete different types of analysis. The following research has been conducted on both Formula One analysis and sports and weather analyses, summarised in Table 1 below.

TABLE 1: SUMMARY OF STATE OF THE ART RESEARCH

Research Title/ Website Article	**What the work is about**	**Similarity/ difference with my own work**	**Ref.**
The Effect of Weather On Sport	The inter-relations between weather and sport. Financial impact of weather awareness in sport management	Looks into sport and the effect of weather on the event	(Thornes, 1977)
The effects of training satisfaction and weather on revisiting sport events and their monetary value: The role of attribute non-attendance	The effect of training satisfaction and weather on revisit intention. Monetary values of weather conditions and training satisfaction. Willingness to travel converted into monetary values using travel costs.	Conducts surveys to attempt to gather data on hypothetical scenarios for return visitation that randomly assigned different travel cost per mile, travel distances, weather forecasts, and respondents' training satisfaction	(Whitehead & Wicker, 2020)
The role of weather conditions on running performance in the Boston Marathon from 1972 to 2018	study examined the relationship of weather conditions, together with sex and country of origin, with running performance in the Boston Marathon from 1972 to 2018	investigate the role of weather conditions, together with sex and country, on female and male performance	(Knechtle, et al., 2019)
Managing weather risk during major sporting events: The use of weather derivatives	Revenue risk increase due to the uncertainty and the unpredictability of the 'right' kind of weather.	Concluding that there are relationships between the sales of various products and various weather parameters with the Australian Open.	(Dawkins & Stern, MANAGING WEATHER RISK DURING MAJOR SPORTING EVENTS:)
The real reason why F1 struggles to race in torrential rain conditions	Speaks on the 2021 Belgian Grand Prix. Speaking on the event being seen as a 'wasted afternoon'. F1 cars dependent on aerodynamics, with tyres being sensitive to conditions.	Laps under the safety car, drivers had reduced visibility standing water and rain on track. Visibility described as one of the main issues.	(Chiu, 2021)
Japanese Grand Prix halted after two laps in heavy rain	The Japanese Grand Prix has been red-flagged and suspended after just two laps of running as heavy rain fell	The rain began falling long before the race got underway. It built steadily in the run-up to the start. Safety car was deployed for car recovery, rain intensifying, red flagged the race.	(Collantine, 2022)

DOES F1 FEAR RAIN? Opinions on the Lack of Wet Weather Racing	Talks about the different tyre compounds and the conditions they are raced on.	Safety issues with racing under wet weather conditions	(Millward, 2022)
What Impact Does The Weather Have On F1?	Track temperature effect on tyres, cloud cover, daytime or night-time, colour of the tarmac, track temperature directly effects tyres.	Car performance effected by the weather/ temperature. Too hot, too cold, needs to be just right for the optimal results	(Media, 2022)
F1 Officials Determined to Find Safer Ways of Racing in the Rain	In recent years, Formula 1 has had seemingly more and more rained-out Fridays and Saturdays at events due to torrential rain. In 2021, Formula managed an entire event that consisted of just three laps behind the Safety Car on race Sunday at Spa.	The two big issues in wet races are visibility and aquaplaning, and F1 officials believe that visibility is an area that can be improved	(Horton, 2021)
11 fascinating graphs on Formula 1	Interactive visualisations from f1 data, looking into lap times, evolution of f1 cars, f1 drivers, speed of f1 cars, logistics of transporting goods to races	Working with several visualisations with the F1 data, some are interactive in nature and other are infographic type visualisations	(Wang, 2022)
Formula 1 Data Vis "Who is the GOAT?"	Expressing a F1 dataset as an interactive visualisation. "Who is the GOAT in F1", a question deliberately selected for its subjective nature	Another example of an interactive visualisation displayed as an information visualisation web app.	(Paul, Formula 1 Data Vis)
Formula 1 Trends	A website for f1 visualisations	Blog/ website for data presentation	(Formula 1 Trends, n.d.)
QlikView Visualization of Formula 1 (F1) Relational Data Model	To explore the relational data model of Formula 1 historical data by visualization in QlikView	Exploring F1 data in QlikView, an analytics tool developed by Qlik	(Wang, 2020)
Formula 1 Data Analysis Using Python 2022 | 2022 Formula 1 Data Analysis Project | Simplilearn	Using Python to perform some exploratory data analysis and draw insights from the data using charts and visualisations	Using datasets and libraries such as NumPy, pandas, Matplotlib, and seaborn, to answer questions related to Formula 1	(Simplilearn, 2022)
Visualizing Data on F1, Indian Ocean, Oreo, Conflicts in Legal Amazon — DataViz Weekly	overview of the 2021 season using a series of animated visualizations. Lightning Plots, and other visualisations	Interactive website created to display the analysis of the formula one 2021 season data	(AnyChart Team, 2022)
Watching the skies - how weather forecasting works in F1	Malaysia is no stranger to spectacular downpours - the race has been red-flagged twice in the last eight years because of torrential rain. how do the teams stay on top of what the weather is doing, not just in Sepang but at different circuits around the world?	Looking at how F1 teams, race engineers, strategists, weather watchers, and the driver deal with deciding on the best course of action for the weather	(Formula One, 2016)
F1 and the weather	weather has an impact on many sports, but there are very few sports where the stakes are as high as they are in Formula One	Looking at different factors that effect F1 during weather conditions, tyres, rain, heat, and wind	(MetMatters, 2011)
Predicting the Weather for a Grand Prix: How Formula 1 Teams Plan for Rain	Speaks on Belgian grand prix nearly always having rain due to its own	Teams are accurately able to judge when a bank of rain is approaching in the pits or on the pit wall to within a	(Masefield, 2013)

	microclimate at the time of the year the race is held	matter of minutes and relay that information to the driver over the radio	
FIA could routinely change F1 GP times to avoid heavy rain	Formula 1 race director speaking on whether it can have the flexibility to bring a grand prix start time earlier into the weekend in case of a poor weather forecast	Looking into the possibility of changing race weekend start times based on the weather	(Khorounzhiy, 2021)
F1 continues push to hit Net Zero Carbon by 2030 target	Development of a 100% sustainable fuel, slashing the use of single-use plastics and reviewing travel and freight logistics – these are just some of the things Formula 1 as a sport is working on as part of its commitment to be Net Zero Carbon by 2030…	Looking ahead, there are plans to build future F1 calendars to improve freight and travel logistics, so the sport is moving more efficiently around the world.	(Formula One, 2022)
Formula 1 Race Predictor	A machine learning approach to predict the winner of the next F1 Grand Prix	A study into predicting race winners, data collection, analysis and models	(Nigro, 2020)
The Formula One Australian Grand Prix: Exploring the triple bottom line	Case highlights the issues associated with the F1 Australian grand prix. highlights the problems with economic impact studies and the need to focus on the triple bottom line approach by examining the economic, social, and environmental issues associated with the event	The study investigates the issues with the event that receives substantial government funding and therefore the worth of the event receives consistent public scrutiny	(Fairley, et al., 2011)
Measuring the economic impacts of major sports events: the case of Formula One Grand Prix (F1)	examined the new money generated from Formula One Grand Prix (F1) and the economic impacts of this new money on the host economy using input–output analysis	F1 event appears to influence on sports-related industry as well as other industries such as manufacturing industry	(Min Kil, et al., 2016)
Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950–2014	random-coefficient models and (a) finds rankings of who are the best formula 1 (F1) drivers of all time, conditional on team performance; (b) quantifies how much teams and drivers matter; and (c) quantifies how team and driver effects vary over time and under different racing conditions	effects are then allowed to vary by year, track type and weather conditions using complex variance functions	(Bell, et al., 2016)
How Could F1 Improve Its Calendar?	There a number of different factors involved in creating the perfect F1 calendar, because there are so many people to satisfy	Mentions races ideally being held in the same region to be grouped together for teams to save money and not so many miles are clocked up. Mentions it as a step towards a greener F1	(Chalmers, 2008)

4.0 Data

Below is a table, Table 2, listing the data that will be used for the analysis and how the data will be compiled and accessed:

TABLE 2: SUMMARY OF POTENTIAL DATASETS FOR THE PROJECT

Data Source	**Dataset Description**	**How will be accessed/ Compiled**	**Notes**	**Ref.**
Ergast Developer API	-Lap times -Number of Laps -Pit stop times -Fastest Laps -Date time of races -Circuit names -Location -Country -Latitude and Longitude -Race results -Races per season -etc.	The data can be downloaded from the Ergast Developer API site as CSV files and will need to be clean and compiled for use in the project	Season: Data from 1950 to 2022 season An entity relationship diagram has been provided on the site to depict the dataset. (shown below table as figure 1)	(Ergast Developer API, n.d.)
Wikipedia	-Red flagged races	Compiling the data from Wikipedia	A table with data on races that have been red flagged	(Wikipedia, 2022)
Formula 1 Wiki	-Red flagged races	Compiling the data from F1 Wiki	Data in table format	(F1Wiki, n.d.)
Formula 1 Wiki	-Safety Car/ Virtual Safety Car ->Cause ->Deployed(lap) ->Called in(lap) ->Number of full laps	Compiling the data from F1 Wiki	Data in table format	(F1Wiki, n.d.)
Meteostat Developers	-Historical Weather Data	Can download data in bulk in csv format, hourly data	Gives historical weather data from	(Meteostat Devel, n.d.)
Open Weather Map	-Historical Weather Data	Open weather map api	Archive since March 2019 Student Access Account	(OpenWeather, n.d.)
Visual Crossing Weather	-Historical Weather Data -Longitude and Latitude coordinates -maximum temperature -minimum temperature -temperature (mean temp) -dew -feels like -precipitation -precipitation chance -precipitation cover -precipitation type -snow	Data can be selected in time periods and downloaded as csv files	Have emailed to change my account to a student account to get more access and increase daily downloads limits	(Visual Crossing Weather, n.d.)

	-snow depth -wind speed -wind gust -wind direction -visibility -cloud cover -relative humidity -etc.			

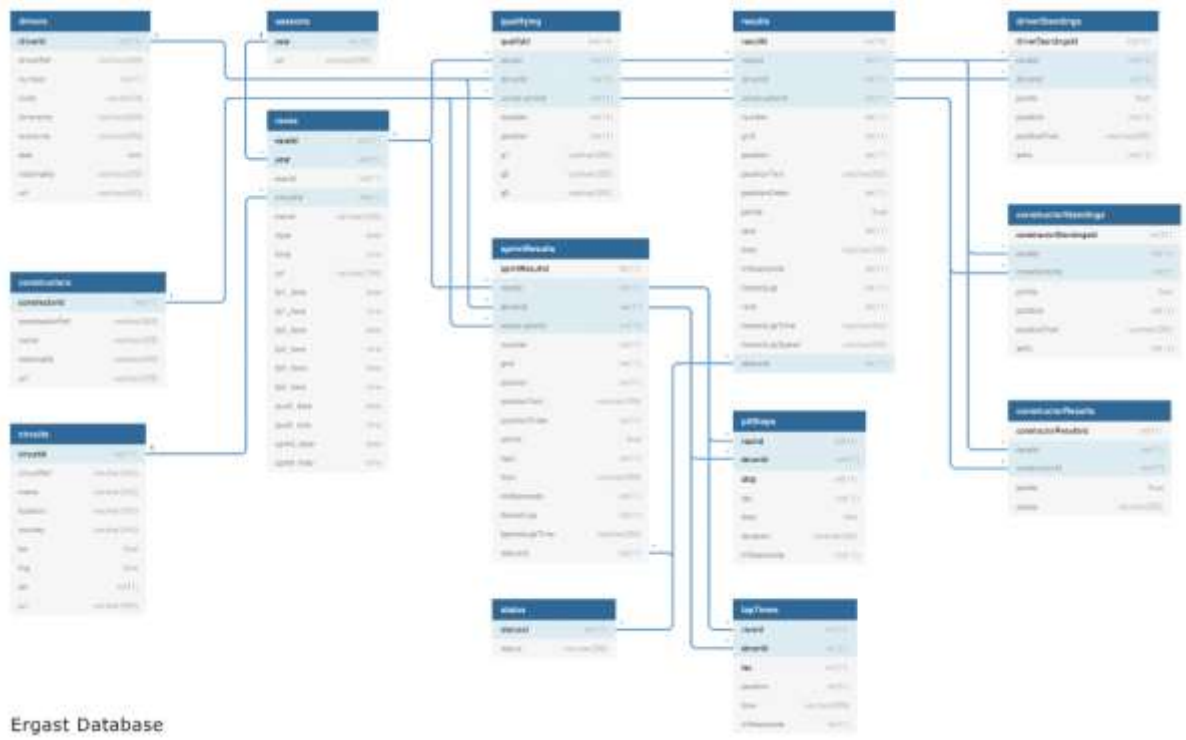


Figure 29: Ergast Entity Relationship Diagram (Ergast Developer API, n.d.)

5.0 Methodology & Analysis

For my analysis I will be using the CRISP-DM methodology. Cross Industry Standard Practice for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process (Nick, 2022).

Within this methodology there are 6 sequential phases:

1. Business understanding – What does the business need?
2. Data Understanding – What data do we have/need? Is it clean?
3. Data Preparation – How do we organize the data for modelling?
4. Modelling – What modelling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

As the project will be dealing with an analysis on a real-world problem, finding the best time to hold Formula 1 events on the race calendar, the CRISP-DM methodology is going to be the best suited methodology for this project.

Phase 6 of the CRISP-DM methodology is particularly useful for this project. The deployment phase is vital for sharing the results. The project report and the plans to create interactive visualisations, hosted on a website are two examples of deployment of the results.

As the project has multiple different objectives it is important that the methodology suits the flexible approach in which the project will be carried out, moving between tasks to complete the different objectives of the project. CRISP-DM can be closely aligned and integrated with Agile (Thurber, 2020), making it a suitable methodology for the project scope.

The project will be broken down into phases/ sprints based on the objectives and the phases within the CRISP-DM methodology. The objectives will become the main tasks that need to be completed with the CRISP-DM methodology expanding on what areas need to be explored.

The first tasks completed so far has focused on discovery and business understanding. Learning and researching what the core business needs for the project are, the availability of data, and the current analysis within the field of Formula 1.

6.0 Technical Details

There will be a use of a multitude of tools to complete the analysis. At this stage plans are to use some of the following tools, with the list expected to expand with further development into the project and requirements for the analysis:

- Python
- R
- SPSS
- Excel
- PowerBI
- Tableau
- SharePoint

I plan to implement machine learning and AI models during the analysis for the project.

7.0 Project Plan

The project plan has been outlined in Figure 2. With 4 main milestones: project pitch, project proposal, midpoint implementation, and final implementation. The Gantt chart was created using the Office Timeline application **Invalid source specified.**.

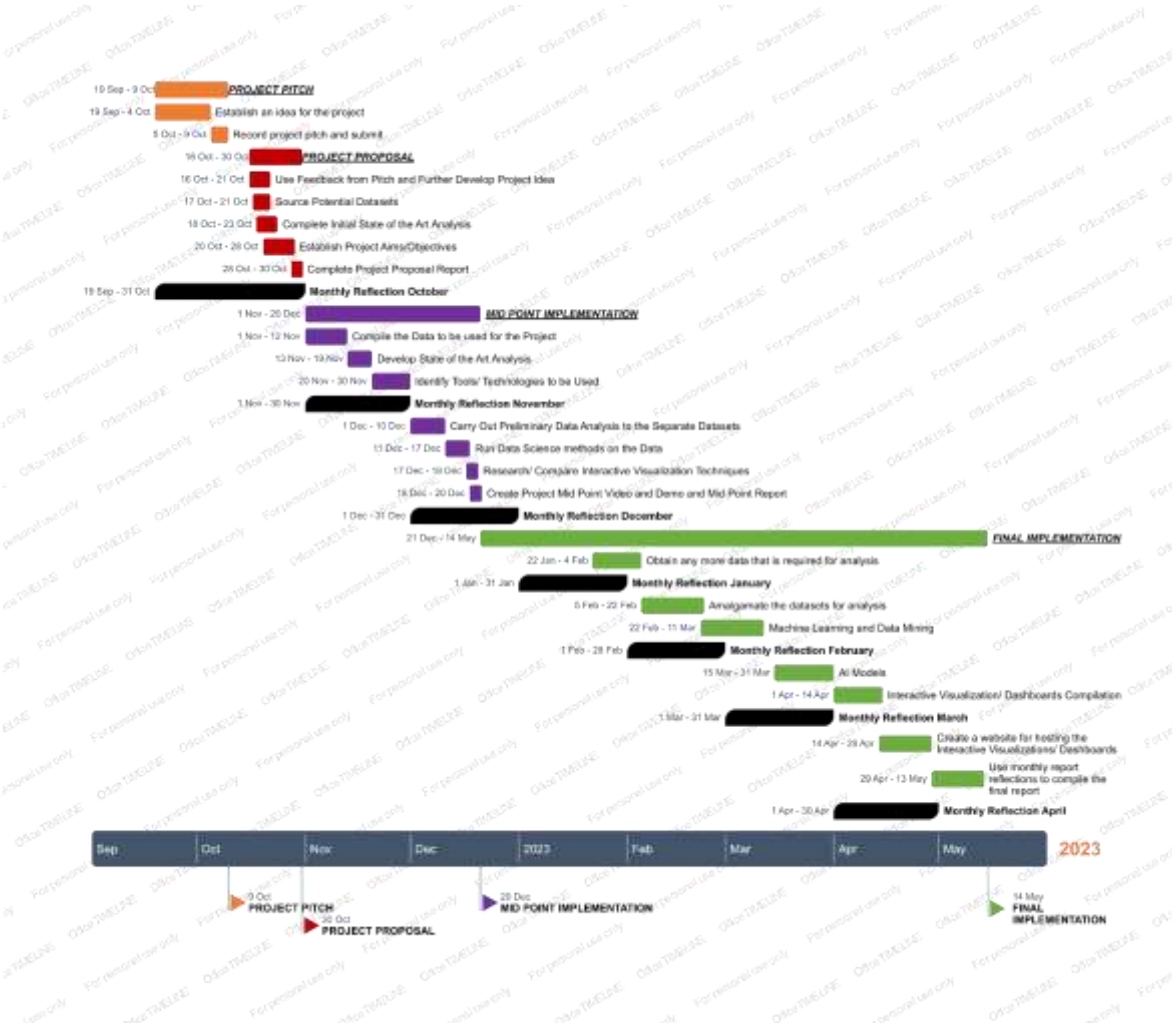


Figure 30: Gantt Chart Created Using Office Timeline ***Invalid source specified.***

References

AnyChart Team, 2022. *Visualizing Data on F1, Indian Ocean, Oreo, Conflicts in Legal Amazon — DataViz Weekly.* [Online]
Available at: https://www.anychart.com/blog/2022/05/06/visualizing-data-f1-indian-ocean-oreo-conflicts/
[Accessed 27 October 2022].

Barretto, L., 2022. *Net Zero Carbon: How Formula 1 is going to meet this ambitious target by 2030.* [Online]
Available at: https://www.formula1.com/en/latest/article.net-zero-carbon-how-formula-1-is-going-to-meet-this-ambitious-target-by-2030.5QsK9NpYbz7pXp7423I3iJ.html
[Accessed 12 December 2022].

Barretto, L., n.d. *Everything you need to know about the F1 Sprint format – including how it works.* [Online]
Available at: https://www.formula1.com/en/latest/article.everything-you-need-to-know-about-f1s-new-sprint-qualifying-format-including.1Lawf6r6Ab0y8ha0satSjx.html
[Accessed 20 December 2022].

Bear, C., 2018. *A short history of crash barrier technology in F1.* [Online]
Available at: https://www.youtube.com/watch?v=FkpvPWB3jMk
[Accessed 10 December 2022].

Bell, A., Smith, J., Sabel, C. E. & Jones, K., 2016. *Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014.* [Online]
Available at:
https://www.researchgate.net/publication/274080402_Formula_for_success_Multilevel_modelling_of_Formula_One_Driver_and_Constructor_performance_1950-2014
[Accessed 27 October 2022].

Chalmers, D., 2008. *How Could F1 Improve Its Calendar?.* [Online]
Available at: https://bleacherreport.com/articles/48779-how-could-f1-improve-its-calendar
[Accessed 27 October 2022].

Chiu, N., 2021. *The real reason why F1 struggles to race in torrential rain conditions.* [Online]
Available at: https://racingnews365.com/the-real-reason-why-f1-struggles-to-race-in-torrential-rain-conditions
[Accessed 27 October 2022].

Collantine, K., 2022. *Japanese Grand Prix halted after two laps in heavy rain.* [Online]
Available at: https://www.racefans.net/2022/10/09/japanese-grand-prix-halted-after-two-laps-in-heavy-rain/
[Accessed 27 October 2022].

Dawkins, S. & Stern, H., n.d. *MANAGING WEATHER RISK DURING MAJOR SPORTING EVENTS:.* [Online]
Available at:
https://d1wqtxts1xzle7.cloudfront.net/42812717/Advances_in_Tourism_Climatology20160218-6411-1xiulgz-with-cover-page-v2.pdf?Expires=1667202474&Signature=HDW2rOgR8j5f4idkhG6BNwcxpUQzubP6JJd3koOuQjseg85BVXPJqvASUSWw73-FxPe8nd4rdI-cUdOL7qmEii6yKDqrroFylL8sR00
[Accessed 27 October 2022].

Ergast Developer API, n.d. *Ergast Developer API.* [Online]
Available at: http://ergast.com/mrd/
[Accessed 27 October 2022].

F1Wiki, n.d. *Red-flagged races.* [Online]
Available at: https://f1.fandom.com/wiki/Red-flagged_races
[Accessed 26 October 2022].

F1Wiki, n.d. *Safety Car.* [Online]
Available at: https://f1.fandom.com/wiki/Safety_Car
[Accessed 27 October 2022].

Fairley, S., Tyler, D. B., Kellett, P. & D'Elia, K., 2011. *The Formula One Australian Grand Prix: Exploring the triple bottom line.* [Online]
Available at:
https://www.researchgate.net/publication/227425409_The_Formula_One_Australian_Grand_Prix_Exploring_the_triple_bottom_line
[Accessed 27 October 2022].

Formula 1 Trends, n.d. *https://www.f1trends.com/.* [Online]
Available at: https://www.f1trends.com/
[Accessed 27 October 2022].

Formula One, 2016. *Watching the skies - how weather forecasting works in F1.* [Online]
Available at: https://www.formula1.com/en/latest/features/2016/9/how-weather-forecasting-works-in-f1.html#:~:text=%E2%80%9CVia%20a%20web%20portal%20we,will%20fall%20and%20so%20on.%E2%80%9D
[Accessed 27 October 2022].

Formula One, 2022. *F1 continues push to hit Net Zero Carbon by 2030 target.* [Online]
Available at: https://www.formula1.com/en/latest/article.f1-continues-push-to-hit-net-zero-carbon-by-2030-target.7fGtPCNCwOnMFFt9Ys1HAa.html
[Accessed 27 October 2022].

Horton, P., 2021. *F1 Officials Determined to Find Safer Ways of Racing in the Rain.* [Online]
Available at: https://www.autoweek.com/racing/formula-1/a38192185/f1-officials-safer-racing-rain/
[Accessed 27 October 2022].

Ingargiola, A., 2015. *What is the Jupyter Notebook?.* [Online]
Available at: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html
[Accessed 18 December 2022].

Jupyter, n.d. *Jupyter.* [Online]
Available at: https://jupyter.org/
[Accessed 2 December 2022].

Khorounzhiy, V., 2021. *FIA COULD ROUTINELY CHANGE F1 GP TIMES TO AVOID HEAVY RAIN.* [Online]
Available at: https://the-race.com/formula-1/fia-could-routinely-change-f1-race-start-times-to-avoid-rain/
[Accessed 27 October 2022].

Knechtle, B. et al., 2019. *The role of weather conditions on running performance in the Boston Marathon from 1972 to 2018.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6407773/
[Accessed 27 October 2022].

Masefield, F., 2013. *Predicting the Weather for a Grand Prix: How Formula 1 Teams Plan for Rain.* [Online]
Available at: https://bleacherreport.com/articles/1752419-predicting-the-weather-for-agrand-prix-how-f1-teams-plan-for-rain#:~:text=Teams%20are%20accurately%20able%20to,via%20an%20internal%20F1%20intranet.
[Accessed 27 October 2022].

Matplotlib, n.d. *Matplotlib: Visualization with Python.* [Online]
Available at: https://matplotlib.org/
[Accessed 25 December 2022].

McLaren, n.d. *F1 PLAYBOOK.* [Online]
Available at: https://www.mclaren.com/racing/f1-playbook
[Accessed 15 December 2022].

Media, 2021. *What is Formula 1?.* [Online]
Available at: https://f1chronicle.com/what-is-formula-1/
[Accessed 25 December 2022].

Media, 2022. *What Impact Does The Weather Have On F1?.* [Online]
Available at: https://f1chronicle.com/what-impact-does-the-weather-have-on-f1/
[Accessed 27 October 2022].

Meteostat Devel, n.d. *Hourlly Data.* [Online]
Available at: https://dev.meteostat.net/bulk/hourly.html#endpoints
[Accessed 27 October 2022].

MetMatters, 2011. *F1 and the weather.* [Online]
Available at: https://www.rmets.org/metmatters/f1-and-weather
[Accessed 27 October 2022].

Microsoft, 2022. *Excel Logo.* [Online]
Available at: https://logos-world.net/excel-logo/
[Accessed 1 December 2022].

Microsoft365, n.d. *Microsoft Excel.* [Online]
Available at: https://www.microsoft.com/en-ie/microsoft-365/excel
[Accessed 1 December 2022].

Millward, A., 2022. *DOES F1 FEAR RAIN? Opinions on the Lack of Wet Weather Racing.* [Online]
Available at: https://www.youtube.com/watch?v=7XDvttk4stY&ab_channel=AidanMillward
[Accessed 27 October 2022].

Min Kil, K. et al., 2016. *Measuring the economic impacts of major sports events: the case of Formula One Grand Prix (F1).* [Online]
Available at:
https://www.tandfonline.com/doi/abs/10.1080/10941665.2016.1176061?journalCode=rapt20
[Accessed 27 October 2022].

MotorsportWeek, 2022. *These Formula 1 Safety Features Keep Drivers Alive.* [Online]
Available at: https://www.motorsportweek.com/2022/07/12/these-formula-1-safety-features-keep-drivers-alive/
[Accessed 17 December 2022].

Nick, H., 2022. *What is CRISP DM?.* [Online]
Available at: https://www.datascience-pm.com/crisp-dm-2/
[Accessed 28 October 2022].

Nigro, V., 2020. *Formula 1 Race Predictor.* [Online]
Available at: https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da
[Accessed 27 October 2022].

NumPy, n.d. *What is NumPy?.* [Online]
Available at: https://numpy.org/doc/stable/user/whatisnumpy.html
[Accessed 7 December 2022].

OpenWeather, n.d. *Weather API.* [Online]
Available at: https://openweathermap.org/api
[Accessed 27 October 2022].

Pandas, 2022. *pandas.* [Online]
Available at: https://pandas.pydata.org/
[Accessed 5 December 2022].

Paul, J. J., n.d. *Formula 1 Data Vis.* [Online]
Available at: https://jasonjpaul.squarespace.com/formula-1-data-vis
[Accessed 27 October 2022].

Pretorius, n.d. *FLAGS IN F1 EXPLAINED.* [Online]
Available at: https://onestopracing.com/flags-in-f1-explained/#:~:text=Red%20means%20the%20session%20is,means%20a%20driver%20is%20disqualified.
[Accessed 26 December 2022].

Python, n.d. *What is Python? Executive Summary.* [Online]
Available at: https://www.python.org/doc/essays/blurb/
[Accessed 30 November 2022].

RAPIDS, n.d. *Build Web Applications With Dash.* [Online]
Available at: https://rapids.ai/plotly.html
[Accessed 29 November 2022].

Richards, G., 2022. *Drive to Survive documentary helping bring in new generation of US F1 fans.* [Online]
Available at: https://www.theguardian.com/sport/2022/may/05/drive-to-survive-documentary-helping-bring-in-new-generation-of-us-f1-fans
[Accessed 15 December 2022].

Schmitt, M., n.d. *Streamlit vs. Dash vs. Shiny vs. Voila vs. Flask vs. Jupyter.* [Online]
Available at: https://www.datarevenue.com/en-blog/data-dashboarding-streamlit-vs-dash-vs-shiny-vs-voila
[Accessed 29 November 2022].

Scikit-learn, n.d. *scikit-learn.* [Online]
Available at: https://scikit-learn.org/stable/
[Accessed 8 December 2022].

seaborn, n.d. *seaborn: statistical data visualization.* [Online]
Available at: https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,introductory%20notes%20or%20the%20paper.
[Accessed 16 December 2022].

Simplilearn, 2022. *Formula 1 Data Analysis Using Python 2022 | 2022 Formula 1 Data Analysis Project | Simplilearn.* [Online]

Available at: https://www.youtube.com/watch?v=7v9puRHfelw&ab_channel=Simplilearn
[Accessed 27 October 2022].

Thornes, J., 1977. *The Effect of Weather on Sport.* [Online]
Available at: https://www.researchgate.net/profile/John-Thornes/publication/260819737_The_Effect_of_Weather_on_Sport/links/5e1b1aeaa6fdcc28376cf986/The-Effect-of-Weather-on-Sport.pdf
[Accessed 27 October 2022].

Thurber, M., 2020. *A Holistic Framework for Managing Data Analytics Projects.* [Online]
Available at: https://www.elderresearch.com/blog/a-holistic-framework-for-managing-data-analytics-projects/#:~:text=CRISP%2DDM%20can%20be%20closely,stages%20within%20an%20agile%20board
[Accessed 27 October 2022].

Tomar, A., 2021. *Dash for Beginners: Create Interactive Python Dashboards.* [Online]
Available at: https://towardsdatascience.com/dash-for-beginners-create-interactive-python-dashboards-338bfcb6ffa4
[Accessed 18 December 2022].

Visual Crossing Weather, n.d. *Weather Data Documentation.* [Online]
Available at: https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/
[Accessed 27 October 2022].

Wang, J., 2020. *QlikView Visualization of Formula 1 (F1) Relational Data Model.* [Online]
Available at: https://towardsdatascience.com/qlikview-visualization-of-formula-1-f1-relational-data-model-82e8b46c9f71
[Accessed 27 October 2022].

Wang, M., 2022. *11 fascinating graphs on Formula 1.* [Online]
Available at: https://medium.com/visual-analytics-field-notes/11-fascinating-graphs-on-formula-1-acd05bcd3e73
[Accessed 27 October 2022].

Westbrook, J. T., 2018. *Here's How Virtual Safety Cars Work In Formula One.* [Online]
Available at: https://jalopnik.com/heres-how-virtual-safety-cars-work-1826237975
[Accessed 18 December 2022].

Whitehead, J. C. & Wicker, P., 2020. *The effects of training satisfaction and weather on revisiting sport events and their monetary value: The role of attribute non-attendance.* [Online]
Available at: https://www.sciencedirect.com/science/article/abs/pii/S2211973620300805
[Accessed 27 October 2022].

Wikipedia, 2022. *2021 Belgian Grand Prix.* [Online]
Available at: https://en.wikipedia.org/wiki/2021_Belgian_Grand_Prix#:~:text=The%20race%20was%20won%20by,George%20Russell%20and%20Lewis%20Hamilton.&text=Planned%20to%20run%2044%20laps,the%20race%20on%20lap%20three
[Accessed 27 October 2022].

Wikipedia, 2022. *Formula One.* [Online]
Available at: https://en.wikipedia.org/wiki/Formula_One
[Accessed 14 December 2022].

Wikipedia, 2022. *List of red-flagged Formula One races.* [Online]
Available at: https://en.wikipedia.org/wiki/List_of_red-flagged_Formula_One_races
[Accessed 26 October 2022].

Wikipedia, 2022. *Safety car.* [Online]
Available at: https://en.wikipedia.org/wiki/Safety_car
[Accessed 14 December 2022].

10.2. Reflective Journals

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 19th September 2022 – 31st October 2022

What?

This month was vital for the beginning stages of my final year project. We began the college semester on the 19th of September with the introduction to the module and learned about the scope of the final year Data Science Project. We learned about the milestones for the project and each of the due dates and marking schemes.

This month the milestones that were due were the project pitch video, the draft project proposal, and the project ethics form.

In preparation for the milestones that were due this month it was important to start the process of determining a project topic/ problem that I wanted my project to be about. After my internship in 3rd year, I knew I wanted my project to be based on a topic within Formula 1, so it was essential for me to understand what previous projects and topics had been covered to determine a unique problem/ task to solve. I began looking at research papers and available data to determine what information was readily available to use for my project idea.

I was developing my idea for the project pitch when the 2022 Japanese Grand Prix was held. This race was red flagged and delayed after just two laps due to the heavy rain that hit Suzuka after the race start. The race subsequently only ran 29 out of the 53 scheduled laps. There was great controversy over the race as it was to be the deciding race in the 2022 drivers' championship.

Project Pitch Video

I formulated my project idea and pitch and was able to record my pitch and submit it for review.

Draft Project Proposal

I was able to meet with my supervisor before the deadline for the draft project proposal and was able to get feedback on my project

I had my first meeting with my supervisor in which he advised of the upcoming deadline and also was given feedback on the datasets I wanted to use.

Project Ethics Form

I gathered my datasets and was able to complete the Project Ethics form including the sources and terms for the datasets.

So What?

When I learned what the scope of the data science project was, I was able to begin working on a project idea.

After watching the 2022 Japanese Grand Prix I knew I wanted to complete my analysis on Formula 1 and the effect the weather has on the races. I was able to find a lot of analyses completed on determining the best driver in formula one and predictions for race results. I knew these were topics I should not investigate as there was a lot of information available for these topics already. The weather and its effect on Formula 1 is a topic that has not been covered in detail before.

Project Pitch Video

The pitch was accepted, and I was assigned my supervisor and was able to meet to get feedback on what I needed to investigate for the project. It was decided that I needed to search for more weather datasets online and determine what resources were available to me to use for the project.

Draft Project Proposal

Because I had completed research, I was able to determine what similar analysis had been completed and what I could do to my project to ensure it was different from what was already out there.

Project Ethics Form

I found my datasets for my analysis and was able to cite the licence agreements and the usage agreements for each of the datasets and data sources.

Now What?

Now that the project pitch video was submitted and approved, and the project proposal and ethics forms have been submitted. I can continue working on my plan for the project and work on the feedback I have received from my supervisor.

The next Project deliverable is the Mid-Point Implementation, Documentation & Video Presentation on December 20th. I will be working towards that deadline for my project objectives and continuing to meet with my supervisor during the arranged sessions.

Project Ethics Form

I can submit a new form next month if there is a requirement or a change in the datasets that I intend to be using for the project.

Now with my data collected and a plan formulated, I can continue with the project and work on developing my project.

Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st November – 31st November

What?

This month I was able to begin working on the project in preparation for the midpoint implementation.

I was able to meet with my supervisor to discuss my weekly progress for the project and to receive feedback and guidance for the next stages to be completed.

For the project proposal I complied several datasets that could have been used for the project. I looked for data from different sources so that I would be able to decide which data sources were the best for the project. I created summary tables with information about the datasets and data sources that I will be using for the project. Not all the data sources that I originally provided in the project proposal were used, e.g., some of the sites for weather data, as I was unable to obtain historical data from these sites.

I also spent time this month further developing my state-of-the-art analysis for the midpoint implementation. I created a summary table for my project proposal state of the art research. I wanted to add to the table and include more details about the state-of-the-art analysis I completed. I went back through my sources and collected more information to be amended to the table. I added a 'type' column to the table (e.g., research paper/blogpost, etc.), a 'Data Science Methods' column (e.g., statistics, visualisation, machine learning, etc.), and a 'Technologies' column (e.g., R, Python, etc.).

I used the state-of-the-art research to Identify some tools and technologies that I could use for the project. I also completed some research to compare similar types of software/ tools to determine what would be best suited for my project.

So What?

Meeting with my supervisor allows me to stay on track and to ensure I am moving the project in the right direction. It is important for keeping myself up to data with the project status and my own progress as the weeks go on.

Going back over the data sources I had collected allowed me to figure out the sources that best suited my project. As I want to look at weather trends at track locations, it is vital for me to have historical weather data for the projects to determine what the weather was like for the races present in the past. Some of my sources did not provide historical weather, or the historical weather was only provided for the past 4 years.

Adding more columns to the state-of-the-art table means I will be able to add more detailed summaries of state-of-the-art sources to the table as I carry out the project.

The technologies column in the state-of-the-art table provided insights into what technologies had been mentioned or used within my research sources, so I was able to carry out my own research on what tools to use, while also considering what had been previously used by others.

Now What?

With the data compiled and the tools and technologies decided, I will be able to continue with the next stages of the Gantt Chart created for the project proposal the next stages for the project are:

- Carrying out the preliminary data analysis on the separate datasets
- Running Data science methods on the data
- Research/ Compare Interactive Visualization Techniques
- Create Project Midpoint Video and Demo and Midpoint report

Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st December – 31st December

What?

At the end of this month the Mid-Point Implementation, Documentation & Video Presentation was due to be submitted.

I was able to continue to meet with my supervisor to discuss my weekly progress for the project and to receive feedback and guidance for the next stages to be completed.

The main deliverables for the Mid-Point Implementation, Documentation & Video Presentation were:

- **Documentation/ Report:** outlining the work completed so far on the project, including the finished proposal and the October and November monthly supervision and reflection journals.
- **Video Presentation:** to be recorded and following the presentation template given, including the questions and answers section required for the presentation.
- **Code and Data Artefacts:** Providing evidence of the code and the data that has been used for the project so far.

I was able to use the Gantt Chart that I created for the proposal to ensure all tasks needed to be completed for the Mid-Point were carried out. In last month's (November) Monthly Supervision & Reflection Journal in the 'Now What?' section I outlined each of the tasks I needed to complete for the Mid-Point submission.

So What?

Meeting with my supervisor allows me to stay on track and to ensure I am moving the project in the right direction. It is important for keeping myself up to data with the project status and my own progress as the weeks go on.

Following the Gantt Chart, I was able to work on and complete the following:

- Carrying out the preliminary data analysis on the separate datasets
- Running Data science methods on the data
- Research/ Compare Interactive Visualization Techniques
- Create Project Midpoint Video and Demo and Midpoint report.

I worked on each of the sections above and was able to outline my findings and work within the documentation. We were required to follow a template for report documentation. The template had various headings that I was able to add information into. This template/report will be the same report I will be adding to for the Final submission in May 2023.

Now What?

With the Mid-Point Implementation, Documentation & Video Presentation completed and submitted. I am now able to move onto the next stages listed in the Gantt Chart.

In the following months I will be continuing to work on the following:

- Writing up the Monthly Supervision & Reflection Journal
- Obtaining any more data that I need for the analysis.
- Amalgamate the separate datasets into one, for a joined analysis.

Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st January – 31st January

What?

With the Mid-Point Implementation, Documentation & Video Presentation completed and submitted. I am now able to move onto the next stages listed in the Gantt Chart.

In the following months I will be continuing to work on the following:

- Writing up the Monthly Supervision & Reflection Journal.
- Obtaining any more data that I need for the analysis.
- Amalgamate the separate datasets into one, for a joined analysis.
- Investigating dashboard capabilities for visualising data.

For the Mid-Point I was able to use the two datasets I have obtained, the Formula 1 datasets and the weather dataset. I carried out a preliminary analysis on these datasets, but for the final implementation I wish to join the two datasets together in order to investigate the effect of the weather on the races.

The goal is to work on obtaining more data for the weather, i.e., daily, and hourly figures for each of the racetracks. Then combine these datasets and investigate the insights obtained from the analyses.

So What?

When I receive the results for the Mid-Point submission, I will be able determine what items need more work, which items are heading in the right direction, and whether I need to add new features to the project.

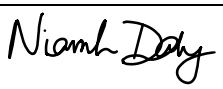
It is important to keep checking my progress against the Gantt Chart I created, to ensure I am on the right track. It is a very effective way of keeping my time management on track.

Now What?

With the Mid-Point submitted, I will be waiting to receive my results so that I can get some feedback on the work that I have already completed and ensure I am on the right track for the final implementation and submissions due in May 2023.

In the following months I will be continuing to work on the following:

- Writing up the Monthly Supervision & Reflection Journal.
- Amalgamate the separate datasets into one, for a joined analysis.
- Carrying out Machine Learning and Data Mining on the datasets.
- Continue to work on investigating dashboard capabilities for the project and the steps needed to implement one for my solution.

Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st February – 28th February

What?

This month I received my results for the Mid-Point, I was able to meet with my supervisor and receive feedback on my results and advice on what needs to be completed next.

This month I have been working on on the following:

- Amalgamate the separate datasets into one, for a joined analysis.
- Carrying out Machine Learning and Data Mining on the datasets.
- Continue to work on investigating dashboard capabilities for the project and the steps needed to implement one for my solution.

I met with my supervisor during weekly meetings to discuss my progress on the project.

I was able to attend the weekly seminars to learn more about different aspects of the project and to learn about the School of Computing 4th Year Full-Time Undergraduate Project Showcase, taking place on the 31st of May 2023.

So What?

With the results and feedback for the Mid-Point submission obtained, I am now able to determine what items need more work in my project, which items are heading in the right direction, and whether I need to add new features to the project.

It is important to keep checking my progress against the Gantt Chart I created, to ensure I am on the right track. It is a very effective way of keeping my time management on track.

Now What?

With the feedback obtained from the Mid-Point and being able to complete items from my Gantt Chart, I am able to move on to the next items on the Chart.

In the following months I will be continuing to work on the following:

- Writing up the Monthly Supervision & Reflection Journal.
- Machine Learning and AI models
- Working on interactive visualisations and dashboard

Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st March – 31st March

What?

This month I have been working on on the following:

- Writing up the Monthly Supervision & Reflection Journal.
- Continuing to work on Machine Learning models.
- Continuing to work on the interactive visualisations and dashboard.
- Beginning to experiment and work with AI models.

I met with my supervisor during weekly meetings to discuss my progress on the project.

So What?

I am still following the Gantt Chart to determine what my next steps are for the project, and what is still to be completed.

It is important to keep checking my progress against the Gantt Chart I created, to ensure I am on the right track. It is a very effective way of keeping my time management on track.

Now What?

In the following months I will be continuing to work on the following:

- Writing up the Monthly Supervision & Reflection Journal.
- Completion of the Interactive Visualisations and dashboard.

• Completion of the Machine learning and AI Models. • Completion of the Final Write up and the additional deliverables for the project. • Taking part in the School of Computing 4th Year Full-Time Undergraduate Project Showcase, taking place on the 31st of May 2023	
Student Signature	

Monthly Supervision & Reflection Journal

Student Name	Niamh Daly
Student Number	X19370553
Course	BSc. (Honours) Data Science (BSHDS4)
Supervisor	Arghir Moldovan

Month: 1st April – 30th April

What?

This month I have been working on on the following:

- Continuing to work on the interactive visualisations and dashboard.
- Working on Machine Learning for the project.
- Working on the report write up.
- Learning more about the project poster deliverable.
- Learning more about the project video and presentation that we must submit.
- Submitted project profile for the project showcase website.

I met with my supervisor during weekly meetings to discuss my progress on the project.

So What?

I am still following the Gantt Chart to determine what my next steps are for the project, and what is still to be completed.

It is important to keep checking my progress against the Gantt Chart I created, to ensure I am on the right track. It is a very effective way of keeping my time management on track.

Learning about what is expected for the project poster and the video presentation allows me to progress in my project and begin to work on these deliverables. These will be the last deliverables to be completed as they will include different pieces of information about each of the aspects of my completed project.

Now What?

In the following weeks I will be finalising work on the project by:

- Completing the programming aspects of the project including the Interactive Visualisations and dashboard and Machine learning.
- Finishing up the Final Write up and the additional deliverables for the project.
- Taking part in the School of Computing 4th Year Full-Time Undergraduate Project Showcase, taking place on the 31st of May 2023

Student Signature