

National College of Ireland  
BSc in Business Information Systems  
2017/2018

Ian Donnelly  
X14111659  
x14111659@student.ncirl.ie

Correlation of Airline Flight delays with Weather Conditions  
Final Report



# Table of Contents

Introduction .....	6
Project Background .....	6
Aims and motivations .....	6
Definitions, Acronyms, and Abbreviations .....	8
Technologies .....	10
MySQL .....	10
MS Excel .....	10
IBM SPSS .....	11
R/R Studio.....	11
Tableau .....	11
Methodology .....	12
Literary Review.....	13
Introduction .....	13
Literature Reviews.....	14
Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs .....	14
Further Investigations into the Causes of Flight Delays .....	15
Analysis of Delay Causality at Newark International Airport .....	16
Microsoft Excel 2016 Data Analysis and Business Modelling: Winter’s Method.....	16
Conclusion .....	17
Bibliography .....	18
Research Question .....	19
Research Title and Question .....	19
Hypothesis .....	19
Normalisation.....	19
Differentiation.....	20
Correlation .....	20
Time Series Analysis.....	21
Methods .....	22

Introduction .....	22
ETL Process .....	22
Data Warehouse.....	23
Datasets Metadata .....	23
Creating the Data Warehouse .....	26
MS Excel.....	30
IBM SPSS.....	32
R/RStudio.....	33
Tableau .....	33
Results.....	34
Kolmogorov-Smirnov.....	37
Mann-Whitney U.....	38
Snow    38	
Wind    40	
Temperature .....	42
Rain    44	
Pearson's Correlation Coefficient.....	46
Snow    46	
Wind    47	
Temperature .....	48
Rain    49	
Holt Winter's Exponential Smoothing.....	50
Tableau .....	55
Snow    55	
Wind    58	
Temperature .....	60
Rain    62	
Testing.....	65
Datawarehouse .....	65
Correlation.....	66
Forecasting.....	67
Future Opportunities .....	69

Further Analysis.....	69
Conclusion.....	71
Appendix.....	72
Executive Summary .....	75
1 Introduction .....	76
1.1 Purpose.....	76
1.2 Project Scope .....	76
1.2.1 Constraints.....	77
1.3 Definitions, Acronyms, and Abbreviations .....	78
1.4 Background.....	79
1.4.1 Motivations.....	79
1.4.2 Similar Studies.....	80
1.5 Aims .....	81
1.6 Technologies .....	81
1.6.1 Services Used.....	81
1.7 Commercialisation.....	81
2 System .....	83
2.1 Functional Requirements .....	83
2.1.1 Requirement 1 <Outputting Data>.....	84
2.1.2 Requirement 2 < Delete Data> .....	86
2.1.3 Requirement 3 < Report Findings> .....	88
2.1.4 Requirement 4 <Use Flight Evidence> .....	91
2.1.5 Requirement 5 <Use Flight Evidence> .....	94
2.2 Non-Functional Requirements.....	96
2.2.1 Data requirements .....	96
2.2.2 User requirements.....	97
2.2.3 Environmental requirements .....	97
2.2.4 Usability requirements.....	97
2.3 Design and Architecture.....	97
2.4 Implementation .....	99
2.5 Testing .....	99

3	Appendix .....	100
3.1	Project Proposal .....	100
	<b>Correlation of Airline Flight delays with external data .....</b>	<b>100</b>
4	Table of Contents .....	101
5	Objectives .....	105
5.1	Lecture's Initial Proposal .....	105
5.2	Proposal .....	105
6	Background .....	107
6.1	Motivations .....	107
6.2	Similar Studies .....	108
7	Technical Approach .....	109
7.1	Development .....	109
7.2	Literature Review .....	109
7.3	Requirements Capture .....	109
7.4	Implementation .....	109
7.5	Project Management .....	110
8	Special Resources Required .....	111
8.1	Software .....	111
8.2	Hardware .....	111
8.3	Documentation .....	111
8.4	Proposed Technologies .....	111
8.5	Services Used .....	111
9	Evaluation .....	112
9.1	Project Plan .....	113
9.2	Monthly Journals .....	114

## **Introduction**

The purpose of this document is to give an in depth statistical analysis of my final year project: Correlation of Airline Flight delays with external data. The document will include a detailed description of the project itself, technologies used in the building and completion of the project including the techniques and methodologies. Within the document there will also be an in-depth analysis of the chosen subject, literary reviews of said chosen subject and the requirements that were found and implemented in order to formulate the findings of the project, including the analytical functions used, their process and the graphical representation of what is being reported. Ultimately this document will act as both a manual for guidance and a reporting tool to highlight the relevance of the findings.

## **Project Background**

The aim of the project is to discover the effects weather conditions have on flight delays and if there is a correlation between severe weather conditions such as heavy snow and the length of the delays on flights. For the purpose of the project, I concentrated my study on Salt Lake City Airport in the American State of Utah. Through the US Department of Transport (USDOT) I was able to acquire the relevant flight data. The data used was all flights over a 9-year period from January 1<sup>st</sup>, 2009 up until December 31<sup>st</sup>, 2017 at Salt Lake City Airport. Together with the flight data I was able to obtain weather data over the same period recorded at Salt Lake City Airport's local weather station from the National Climate Data Centre (NCDC) and map the flights with the weather conditions through the dates of both.

## **Aims and motivations**

Using statistical analysis formulas, the aim of the project is to first compare flight delays at different weather conditions to discover if weather itself effects the flights in a negative way and if so, run a correlation coefficient known as Pearson's R, to discover if there is a correlation between these delays and the change in weather

conditions. After studying the findings, I will then run a time series analysis known as the Holt Winters Exponential Smoothing method, which in theory should allow the ability to accurately forecast the level of delay on future flights at different times of the year based on previous cyclical data. This information if accurately forecasted can be used to inform passengers of expected delays in advanced, prepare airlines and airports in advanced and enable them to implement counteractive methods to help to reduce delay or if unable to positively effect delay time, better prepare for the consequences for example, with more personnel. In order to then validate the precision of the forecast, I will apply the forecast to the historical data and compare it against the actuals over the same period to gain a percentage accuracy level of the level of delays on flights.

## Definitions, Acronyms, and Abbreviations

In the section below, you will see some terms used throughout the project/report and a brief description on their meanings and uses.

- ❖ USDOT: The US Department of Transport was sourced to access the public data on flights used throughout the statistical analysis tests and report.
- ❖ NCDC: The National Climate Data Centre was sourced to access the public data on weather used throughout the statistical analysis tests and report.
- ❖ DB: Database, a structured set of data held in a computer, especially one that is accessible in various ways.
- ❖ DW: Datawarehouse: A store of data accumulated from a range of sources within and used to guide management decisions.
- ❖ R: An open source programming language used for statistical computing and graphics.
- ❖ MySQL: Is an opensource DB management system acquired by Oracle.
- ❖ SQL: Structured Query Language, used for querying and managing data stored in a DW/DB.
- ❖ UI: User Interface, the means by which a user and computer system interact.
- ❖ GUI: Graphical User Interface, is an interface that allows users to interact with a computer system through graphical and visual icons.
- ❖ Tableau: A UI application used for data visualisation.
- ❖ ETL: Extract, Transform and Load, the functions combined into one tool to pull data out of one DB and store in another. Extract is the process of reading data from one DB, Transform is the process needed to in the conversion of the data from the original DB into the form needed to be placed in another and Load is the process of writing the transformed data into the target DB.
- ❖ Entity: A component of data in a DB.
- ❖ ERD: Entity Relationship Diagram, shows the relationships of entity sets stored in a DB.

- ❖ BI: Business Intelligence, comprises the strategies and technologies used by enterprises for the data analysis of business information.
- ❖ MS Excel: Microsoft Excel, a spreadsheet developed by Microsoft that can be used for calculations, graphics and formula.
- ❖ IBM SPSS: Is a software package used for statistical analysis.
- ❖ Data Dumps: A large amount of data transferred from one system or location to another.
- ❖ CSV: Comma Separated Values, a file type used to store data in a tabular format.
- ❖ H0: Null Hypothesis, is the assumed hypothesis. This is the hypothesis we set out to disprove.
- ❖ H1: Alternative hypothesis, is the hypothesis we accept when rejecting the H0.
- ❖ PK: Primary key, is a special column within a relational database or data warehouse designated to uniquely identify all table records.
- ❖ FK: Foreign key, is a field in a table of a database or data warehouse that uniquely identifies a row in another table.
- ❖ APE: Absolute percentage error, is a measure of predicting the accuracy of your forecast.
- ❖ MAPE: Mean absolute percentage error, the mean accuracy of your forecast

## **Technologies**

### **MySQL**

With the project itself being centred around the use and manipulation of data it is no surprise that a DB and subsequently a DW were constructed. This enabled the loading of the data on flight and weather through data dumps. The data was then queried through BI queries to retrieve the relevant data needed to conduct the analysis. Although built to manipulate and query the data, the main use of the DW would be through further development of the report and the implementation of the statistical analysis on other airports, in other climates throughout the world. The DW itself shows the benefits of having the ability to manipulate the data on a small scale compared to one created for not just Salt Lake City but the entire airport infrastructure throughout the United States of America. This could then be used to query flight delays based on climates and weather throughout different part of the country rather than just one airport and the results and findings of such could be very valuable to the likes of airports themselves, airline companies and the USDOT also.

### **MS Excel**

Microsoft Excel played a vital role in the completion of the project and was used from the beginning to the end. It was needed in the initial downloading of the data files from USDOT and NCDC respectively, being downloaded as .CSV file formats to enable their use in Excel. Utilised in the uploading of data into the DW, it was also used in both the ETL process and the statistical analysis in which it yielded a lot of the findings set out in the project itself. This was achieved through the use of the Data Analytics tool pack, building of pivot tables and the writing of formulas to manipulate the data to enable its use. The resulting Excel sheets are then imported into Tableau for graphical representation.

## **IBM SPSS**

SPSS is a statistical analysis package and was used throughout the project to run statistical analysis checks on the data along with descriptive tests to find out the relevant information needed about the statistics to know which tests to conduct on the data for the accurate results. Also used to show visual representations of the findings through charts and graphs, test the data's integrity and confirm the results of other analytical statistics tests conducted, SPSS played a vital role in the project and as a technology helped define the results of the conducted tests and their accuracy which gave confidence to the use of the data in the results for further testing and forecasting.

## **R/R Studio**

The use of R as a programming language allowed for further statistical analysis to be conducted on the data in a necessary way to gain the relevant findings to be used within the report. This was done through the R Studio application which provided the environment for the coding of the language to be implemented to yield the results needed. R was used in the gathering of certain integral data for the forecasting procedure within the report, along with the analysis and graphical representation of the necessary data. R is another technology that played an integral part of the project from the enabling of certain manipulations on the data, to testing the accuracy of the results within the statistical analysis being conducted to the programming used to gather statistical information on the data sets in question.

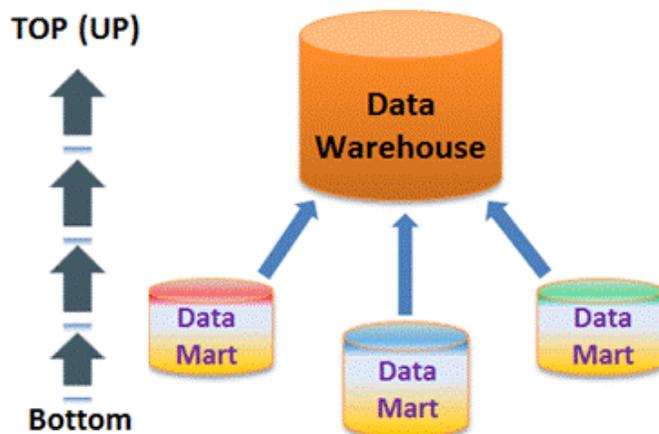
## **Tableau**

Tableau is a data visualisation tool which enables the importing of data through the Excel sheets for graphical representation of the findings. When conducting the statistical analysis, the resulting data from the test conducted is imported into Tableau which can be manipulated to show graphical representation of the findings

through many different graphs and chart types. Through the creation of multiple graphics, a dashboard was created and then hosted on Tableau Public server to create a GUI of the results to be viewed, used and manipulated by the end user to gain a visual understanding of the statistical analysis results retrieve from the tests conducted. Tableau was the technology that pieced together all the results and gave purpose and representation to all the other technologies used throughout the project.

## Methodology

There are many different methodologies in creating a DW, the two most common and main methodologies are Kimball and Inmon's methods. In the design of my DW I chose to use Kimball's approach as I found it more pertinent as it is what's known as a bottom up outlook compared to Inmon's top down approach. This is duly named as in its creation we construct the dimension tables first or data marts, which are then integrated together to create the warehouse.



### Kimball's Data Model Approach

In the image above, we can see a diagram of the Kimball approach. This can be mapped to my own approach as I first built the Airport, Date and Station Dimensions before loading the data from the marts into the FAWFacts centralised table.

# Literary Review

## Introduction

A literary review was carried out on the topic of the project; Correlation of Airline Flight delays with weather conditions. The purpose of the literary review is to carry out research on previous literary already conducted in the field and to provide further understanding of the field in which the topic is associated. The purpose of my literary review is to gather information through research of books, scholarly articles and papers and other sources relevant to the hypothesis of the statistical analysis report, do weather conditions have a negative impact on flight delays? Through extensive research on this topic from such sources as Klein and Kavoussi's study on weathers impact on airport capacity and cost, I will discuss some of my literary findings within the field on the subject and related subjects that impact my hypothesis. I will do this by both acknowledging the work carried out and positive impacts it has within the subject but by also critiquing the works of others where necessary to question the subject and delve further into it gaining a better understanding of the overall field and the questions being asked throughout my report. The aim of my statistical analysis report is to better understand the effects weather has on flights themselves but to also use the information found to better forecast the length of the delays to help improve preparation for and thus reduce the delays going forward if possible, or to at least be as informed as accurately possible in the expectations of the delays ahead. In order to implement a strategy such as this it is important to review the field of study and further writings within the field, which will be done throughout this literary review.

## Literature Reviews

### Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs

This study focuses on the weather conditions and their impact on airports as a whole, the level of delay they cause, but also other factors that cause delays on flights and negatively impact airports and their ability to run a smooth process. This is then taken from the outlook of cost and the impact any delay to services, processes or operations will have on the airport and its airlines in a financially negative way. From a cost benefit position the ability to reduce or stop any delays will increase cost and this is the aim of the research paper, to reduce losses by reducing delays. Within the paper delays are divided into avoidable and unavoidable, the latter being considered for more severe weather conditions such as storms (*Klein, Kavoussi and Lee, 2009*).

The results of the research paper align with my own finding in that severe weather conditions impact flight delays in a negative manner however, the findings of this paper do differ in some respects. The findings of the paper suggest that the biggest weather condition to negatively impact on flights is wind. Although this is accurate it is also dependent on outside influences, such as the geographical location of the airport and the flight path of the aircraft. For example, in my findings the biggest impact on flights is days of bad snow conditions, this is down to the geographical location in which my statistical analysis report was conducted. In Salt Lake City Utah, the climate allows for substantial amount of snow fall each year, which is not something that you could relate to other areas of even the United States of America and furthermore, wind speeds and negative wind conditions would be less drastic than at the geographical location of other airports.

Although the high-level results of wind speeds causing rougher delays on average than any other weather condition, it is not as black and white as this. High wind speeds are a more common weather condition throughout the United States as a whole and occurs more often than that of bad snow conditions which, when occurring tends to have a greater impact than that of wind. This gives further

strength to the benefits of the project and report I am conducting. The building of the DW will enable further use of any/all airports if expanded and will allow the ability to drill down into a specific location, discover the weather conditions which most affect the area, create a similar analytical report of your findings on that specific airport and with this information through forecasting techniques used, be better prepared for upcoming delays in flights at each individual location.

### Further Investigations into the Causes of Flight Delays

This particular study focuses on the flight delays as a whole, the statistics behind delayed flights, what causes them and the financial impact that delayed flight have on airline companies throughout the United States of America. In this report, flight data from the year 1994-2007 was investigated and shown that over the 14-year period, just over 20% of all flights were delayed. A flight qualifies as being delayed if actual departure time is more than 15 minutes later than scheduled departure time. This means that 1 in every 5 flights over the 14-year period departed more than 15 minutes after schedule. According to this study, this level of delay is said to have cost the airline industry in excess of \$3b annually. When delving further into the data, it is found that of that 20%, weather is the leading cause of delayed flights and contributes to almost a third of delays at 32% and 6.4% of all flights being delayed by weather (*Rupp, 2007*).

With weather causing such an impact on flight delays annually, the future implementation and monetising of my project is obvious. With 32% of all delayed flights throughout the United States of America, being caused by weather, which is a third of the \$3b impact delays have on the airline industry, weather delays cost the airline industry in excess of \$1b yearly in the United States of America alone. With the ability to run statistical analysis on individual airports and geographical locations to return bespoke findings and solutions for each airport to better prepare for the forecasted conditions and delays, to strive to reduce and remove the delays, saving the airline industry up to \$1b annually in the United States of America and even more worldwide.

## Analysis of Delay Causality at Newark International Airport

This study focuses on the delay in flights at Newark airport in New York City, during the 3-year period of September 1998 and August 2001. Newark Airport has one of the highest flight delay statistics in the whole of the United States of America, which is why the study was conducted to try to help manage the delays through the understanding of the cause of the delays themselves. This study shows the effects weather has on overall delays such as bad wind conditions causing 14% of all the delayed flights, confirming what was discussed previously. However, what this study also shows is the knock-on effect delayed flights in one airport can have on all airports within the network, as delayed departure flights lead to delayed arrival times and thus affect flights downstream causing a ripple like effect (*Allan et al., 2002*).

Accurately attributing the cause of delays has become increasingly important to the aviation industry, it is quite clear that a substantial amount of delays is caused by weather conditions, but what can be done about it? With the implementation of technologies and statistical analysis studies within my project, on individual airports, we can work towards solving up to 32% of all delayed flights or at least reduce it, but also as this study suggests, prevent the butterfly effect, delayed flights at the departure airport can cause at the destination airport. Increasing the prevention process of delayed flights at one airport, will lower even slightly, the overall average of delayed flights throughout the corresponding network.

## Microsoft Excel 2016 Data Analysis and Business Modelling: Winter's Method

In chapter 63 of this book, the Holt Winter's exponential smoothing method that I researched is explained. This smoothing method is used to help forecast cyclical data based on trends and parameters of the data involved. The example provided in the book takes a look at the housing market in the United States of America during the period of 1987 to 1997. Based on the values within the housing market and trends used, the Holt winter's method can use smoothing techniques to then forecast the housing market prices for the coming months of 1998 and further, in

the case of the example within the book. By studying this chapter, I was then able to apply the Holt Winter's method to my data to try to accurately forecast the level of flight delays of the months ahead, which in my case was January through March 2018 (*Winston, 2016*).

## **Conclusion**

When looking back on the findings of the literary reviews, we can see the major impact weather conditions can have on flights throughout the United States of America on a broad universal scale, causing up to 32% of all delays on flights throughout the country. We can also see the financial implications that occur, for the aviation industry, from delays throughout the country and understand the knock-on effect these delays can have on other airports downstream. My project aims to gain an understanding of the weather implications at each individual airport concentrating on Salt Lake City, to see what weather conditions impact Salt Lake City airport the most, the level of effect these conditions have, use a time series analysis to try and forecast the delays and use this information to work on prevention techniques. This can then be used and adapted on an individual level for each airport.

## **Bibliography**

- Klein, A., Kavoussi, s. and Lee, R. (2009). Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs. pp.1-10.
- Rupp, N. (2007). Further Investigations into the Causes of Flight Delays. [online] pp.1-40. Available at: <http://www.ecu.edu/cs-cas/econ/upload/ecu0707.pdf> [Accessed 11 Feb. 2018].
- Allan, S., Beesley, J., Evans, J. and Gaddy, S. (2002). Analysis of Delay Causality at Newark International Airport, [online] pp.1-11. Available at: <https://pdfs.semanticscholar.org/a256/fe70c6aa05c86b1edb94f3c434b3ebe3ceb5.pdf> [Accessed 11 Feb. 2018].
- Winston, W. (2016). Microsoft excel data analysis and business modeling. [Place of publication not identified]: Microsoft.

## **Research Question**

The ability to ask the correct question is fundamental in the construction of a report like this as the question being asked is the fundamental foundation of the report itself. In this section I will discuss the questions being asked and thus the answers I hope to receive from the statistical analysis report, but also set out the relevant statistical hypothesis to be conducted along the way.

## **Research Title and Question**

When setting out to conduct a statistical analysis test on the weather and flights data, the main objection was to discover if there is a correlation between severe weather conditions and the flight delay times in the same regions. However, when the question is broken down what is actually being asked, there are many questions involved. When drilling down into the question further I must find out first if there is a significant difference between flight delays in comparison to days for example with little to no snow and heavy snowfall, with no rain to heavy rain and with mild or heavy winds etc. Then I must ask is there a cyclical pattern occurring as to the weather and its effects on flights, can it be put down to seasons and times of year for example? Are the weather conditions affecting flights throughout the year, with all different weather types and furthermore, with the findings can I forecast what delays to be expected at various times of the year, throughout different weather conditions based on the strength of the correlation and the cyclical data used?

## **Hypothesis**

### **Normalisation**

The first step in the statistical analysis is to determine whether the data being used is normalised data. Normalisation of data is whether the data being tested is normally distributed. There are multiple indicators to signify if data is normal or not normal, however the conducting of a Kolmogorov-Smirnoff test will give a definitive answer to the hypothesis. The significance of the normalisation of the data is to determine which tests to conduct for accurate results. If the result of the test shows

the data is normal, parametric tests for further analysis are used however, if the data is not normal then non-parametric tests to improve the level accuracy in the findings are conducted. The hypotheses are written as follows:

H0: The data is normal

H1: The data is not normal

### Differentiation

Once the results of the Kolmogorov-Smirnov test have been gathered and have determined the data being dealt with is in fact not normal data, I can begin to conduct non-parametric tests. There are four tests in total conducted, each one used to determine if there is a difference between the effect each weather condition has on the outcome of delayed flights and most importantly is the difference significant. The non-parametric test used is known as the Mann-Whitney U test. A Mann-Whitney U test was conducted on the effects of rain, snow, wind and temperature on the length of flight delays. The hypothesis for all 4 are as follows:

H0: M rain = M no rain

H1: M rain  $\neq$  M no rain

H0: M wind < 10mhp = M wind > 10mph

H1: M wind < 10mhp  $\neq$  M wind > 10mph

H0: M snow = M no snow

H1: M snow  $\neq$  M no snow

H0: M positive 32 = M negative 32

H1: M positive 32  $\neq$  M negative 32

### Correlation

When analysing the findings, it can be seen what kind of impact different weather conditions can have on the level of delays on flights. In order to gain a greater understanding of the level of effect in which the weather impacts the delays a correlation test known as Pearson's Correlation Coefficient or Pearson's R can be

conducted. The Correlation Coefficient measures the strength and direction of a linear relationship between two variables, in this case weather condition and delays. It is important to remember however, no matter how strong the level of correlation between two variables, correlation is not causation. Given that correlation is a test to determine the level of relationship between two variables the closer the R value is to 0 the lower the relationship level. The scale for Pearson's R starts at -1 indicating a strong negative correlation and ranges to +1 indicating a strong positive correlation. The hypothesis is written as follows:

H0:  $P = 0$

H1:  $P \neq 0$

### Time Series Analysis

Time series analysis comprises of procedures for analysing time series data with the intention of extracting significant statistics and other qualities of the data. The procedure of time series forecasting within time series analysis is the use of a model to predict/forecast future values based on previously observed values. The time series forecasting method used for the forecasting of flight delays in my statistical reports is the Holt Winter's Exponential Smoothing method also known as the Triple Exponential Smoothing method. The Holt Winters method is known as the most powerful smoothing method, which is why I chose to implement it within the report, to gain the most accurate forecasts as possible. It is best used for cyclical data or seasonal data, which means it's repeated over time, such as the months of a year used in this paper. Although there is no scientific H0 and H1 for the Holt Winter's method, a question is still being asked on whether or not the average delay of flights in future months can be accurately predict.

# Methods

## Introduction

The methods section will describe the processes and methods used in the completion of the statistical analysis project. From the downloading of the data right up until the reporting of the findings and those used throughout the project will be conveyed in this section. This section will act as a walkthrough of the construction of the statistical analysis report and a guide on the technologies used.

## ETL Process

The ETL process is vitally important in any project involving data, from the construction of a DW to the analytical formulas, clean and precise data is a necessity. Right through the project Excel played an essential role and the ETL process is included in that. The initial data download for both the flight data received from the USDOT and the weather data from the NCDC are major datasets and would need to be cleaned up and transformed before gaining the ability to utilise them in the statistical analysis report.

The first stage in the ETL process is the extraction stage. In the downloading of the data from the NCDC for Salt Lake City, I had to first select the most relevant weather station. Conveniently most airports will have a weather station in the same vicinity to be as precise as possible about the forecasting of weather. This made the selection of the weather station simplistic. Selecting the time period of 10 years ending December 31<sup>st</sup>, 2017 enabled me to investigate the findings on relevant up to date information and on a substantial amount of data over a respectful period of time. When this was selected the download began. For the flight data gathered by the USDOT, the extraction process was not as simplistic. This data which was selected for the same time frame and over the same time period included information on all flights in and out of Salt Lake City airport over the 10-year period. The sheer capacity of data involved in the download created complications in the extraction. The volume of information in the data meant that the datasets were too large to download all at once and had to be download one month at a time, for the whole 10-year period. This led to 120 downloads in the extraction of the flight data and consumed a considerable amount of time.

The transformation process is the most important part of the ETL process, it enables the data extracted from the origin DB/DW to be converted into the correct format to be manipulated in the destination. The first task undertaken in the transformation process was the amalgamation of all 120 files downloaded from the USDOT into one large dataset, also requiring a substantial amount of time and

effort. With the merging of the 120 files on flights the discovery of an obstruction arose, which also lead to a discovery of an unknown interesting fact. With the sizeable amount of information on flights over the 10-year period, after 9 years and 2 months I reached the max capacity of rows that Excel can allow, which is 1,048,576 rows. After maxing out Excels capabilities, the planned 10-year data analysis was reduced to a 9-year period. The ETL process manually conducted in Excel enabled me to investigate the information within the datasets to remove all unnecessary data, which would not be needed in the analytical study conducted. This included the cleaning up and removing of redundant data, the deletion of duplicate rows, spell check errors and the removal of any unnecessary fields/columns retrieved from the downloads, such as model numbers of aircrafts in the flight data. When the cleansing of the data was completed, the file formats were then transformed from .XLSX files into .CSV files

The final method in the ETL process is the loading of the newly transformed data into the destination system. In this case the files were loaded into a DB through MySQL Workbench. This was done through data dumps of large .CSV files with SQL coding. Once loaded in to the DB, the data is then ready to be operated upon.

## **Data Warehouse**

The process of creating the data warehouse began by firstly creating a database to enable the importing and manipulation of data to be configured by the DW.

### **Datasets Metadata**

The two datasets dealt with throughout the entire report was the flight dataset from the USDOT and the weather dataset from the NCDC. The flight dataset consists of data in regard to Salt Lake City airport in Utah and its attributes include:

- ❖ FL\_DATE – Flight Date

Details about the Airlines and the location of the airport:

- ❖ Unique Carrier
- ❖ Airline ID
- ❖ Origin Airport
- ❖ Origin
- ❖ Airport Name
- ❖ State
- ❖ Destination Airport ID
- ❖ Destination Airport name.

It also includes measured data that we will track to discover if there is a correlation in the data, such as:

- ❖ Planned Departure Time
- ❖ Actual Departure Time
- ❖ Departure Delay
- ❖ Taxi out time
- ❖ Wheels off time
- ❖ Wheels on time
- ❖ Taxi In time
- ❖ Planned Arrival Time
- ❖ Actual Arrival Time
- ❖ Arrival Delay
- ❖ If Cancelled
- ❖ If Diverted
- ❖ Planned Elapsed Time
- ❖ Actual Elapsed Time
- ❖ Air time
- ❖ Weather Delay

The dataset for the weather was includes attributes such as:

- ❖ Date

It then also includes information on the weather station in which the data was recorded, in the case of this report, the weather station is the Salt Lake City Airport

Weather Station in Utah:

- ❖ Station
- ❖ Station Name
- ❖ Latitude
- ❖ Longitude
- ❖ Elevation

The measurements of the different types of weather are included in the dataset used and are as follows:

- ❖ AWND
- ❖ FMTM
- ❖ PGTM
- ❖ PRCP
- ❖ Snow
- ❖ SNWD
- ❖ TAVG

- ❖ Tmax
- ❖ Tmin
- ❖ WDF2
- ❖ WDF5
- ❖ WESD
- ❖ WSF2
- ❖ WSF5

Below are images with important information on the meaning of the abbreviations used in the dataset.

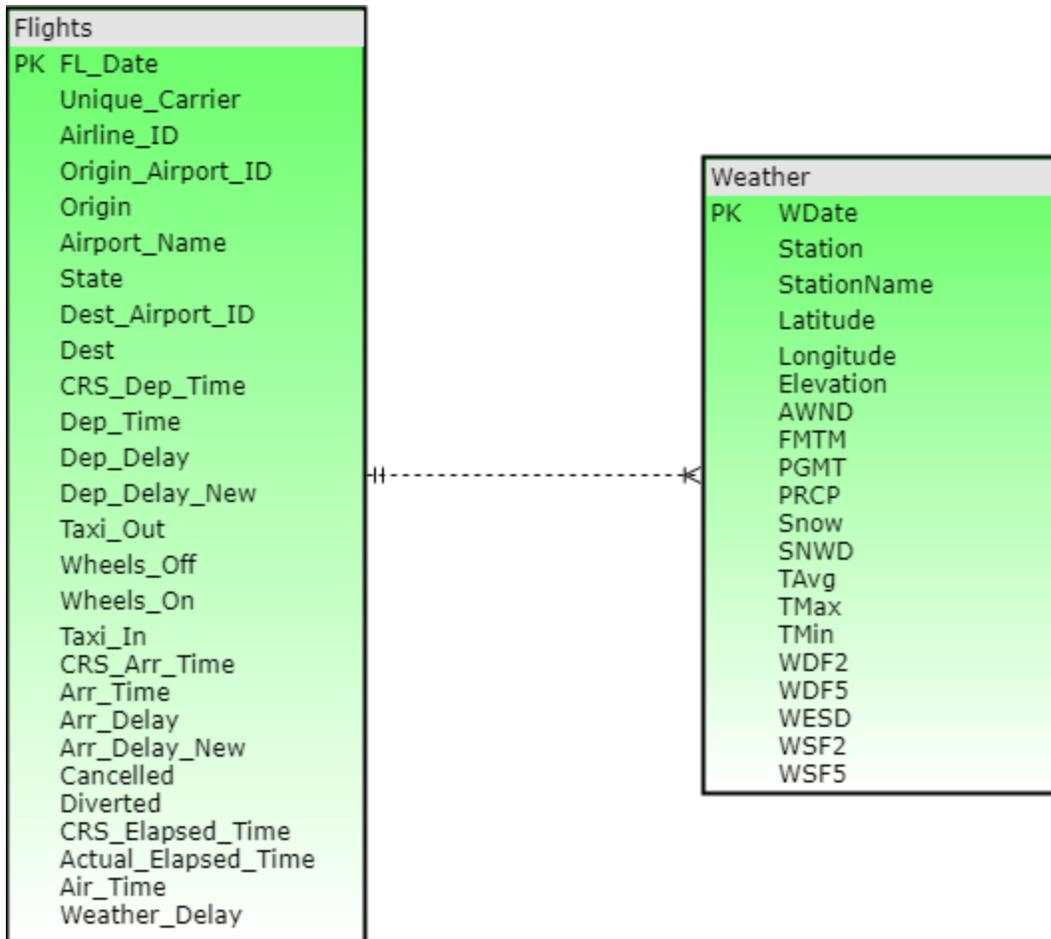
- ☐ ☐ Land
  - ☐ Maximum soil temperatures (SN\*\*)
  - ☐ Minimum soil temperatures (SX\*\*)
- ☐ ☐ Precipitation
  - ☐ Precipitation (PRCP)
  - ☐ Snow depth (SNWD)
  - ☐ Snowfall (SNOW)
- ☐ ☐ Sunshine
  - ☐ Total sunshine for the period (TSUN)
- ☐ ☐ Air Temperature
  - ☐ Average Temperature. (TAVG)
  - ☐ Maximum temperature (TMAX)
  - ☐ Minimum temperature (TMIN)
- ☐ ☐ Water
  - ☐ Water equivalent of snow on the ground (WESD)
- ☐ ☐ Wind
  - ☐ Average wind speed (AWND)
  - ☐ Direction of fastest 2-minute wind (WDF2)
  - ☐ Direction of fastest 5-second wind (WDF5)
  - ☐ Fastest 2-minute wind speed (WSF2)
- ☐ ☐ Wind
  - ☐ Average wind speed (AWND)
  - ☐ Direction of fastest 2-minute wind (WDF2)
  - ☐ Direction of fastest 5-second wind (WDF5)
  - ☐ Fastest 2-minute wind speed (WSF2)
  - ☐ Fastest 5-second wind speed (WSF5)
  - ☐ Peak gust time (PGTM)
  - ☐ Time of fastest mile or fastest 1-minute wind (TF1M)
- ☐ ☐ Weather Type
  - ☐ Weather types (WT\*\*)

## Creating the Data Warehouse

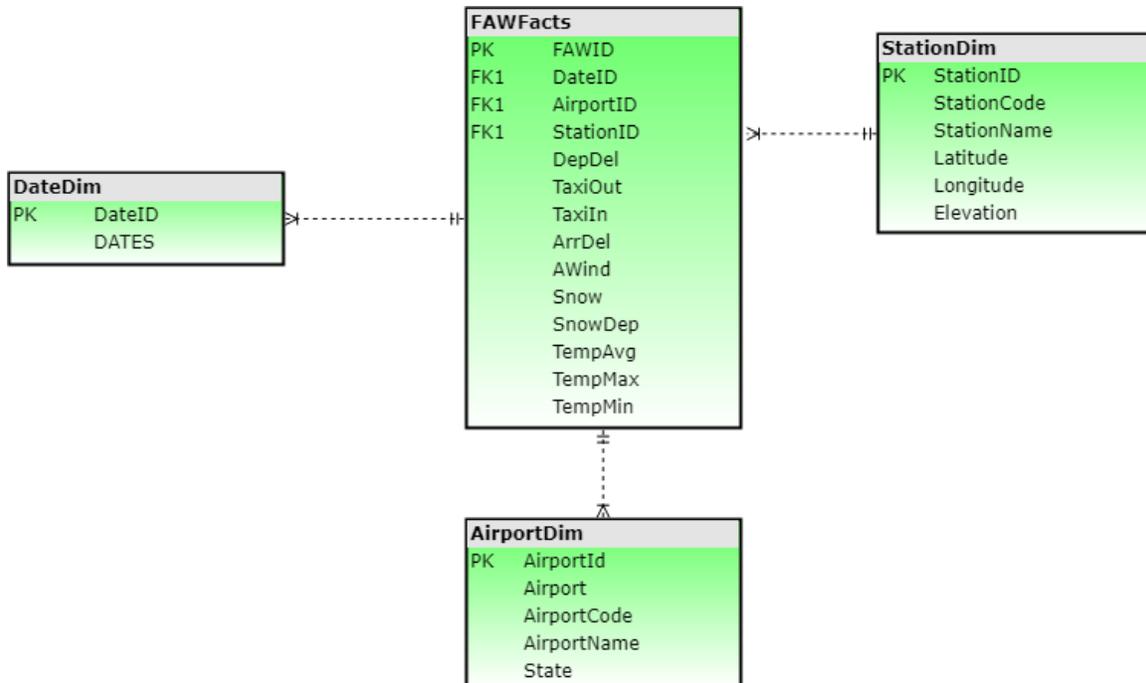
Created in MySQL workbench the database is made up of two large tables, one for flights and the other for weather. This enabled me to code a data dump to import data from .CSV files directly into the DB, since they had already been cleaned up and prepared in Excel throughout the ETL process.

```
45
46 /* Data Dump into table */
47
48 • LOAD DATA LOCAL INFILE 'H:/FinalYearProject/28319994_T_ONTIME/FlightData2.csv'
49 INTO TABLE Flights FIELDS TERMINATED BY ','
50 ENCLOSED BY '"' LINES TERMINATED BY '\n'
51 IGNORE 1 LINES;
52
```

The snippet of SQL code above was written to insert the data from the flight dataset directly into the flight table. There was a similar technique also used to import the data for the weather DB table. The formatting of the Excel files was essential in the data dump. If you see the figure below, it shows a graphic of the DB schema including both tables and each attribute within the tables. In order to successfully import the data from the .CSV file into the individual tables, the file had to be formatted in a way that each column header had to match the order of the attributed within the respective table.



From the two large database tables above, I was then able to create the structure of my star schema, seen in the figure below. The purpose of drawing the star schema was to give structure to the DW and create a visual plan before beginning to code. It also enabled me to get an understanding of what attributes from the DB would be needed and used in the DW. The first thing to figure out when creating the DW was the different tables needed, the central facts table, the date dimension, the airport dimension and the station dimension, these can be seen visually in the image of the star schema below.



As seen above, the DW has a central facts table which consists of the PK of its table and the FKs of the dimension tables, along with the measurement attributes to be used. The dimension tables, Date, Airport and Station consist of their respective PK IDs which enables querying from the facts table directly to an ID in the dimension table where the stored descriptive are, for faster querying and to reduce redundancy of data. The population of the dimension tables was done through the attributes in the DB tables. Firstly, the Date dimension table was populated with the date from the flights table in the DB and a Date ID was then created in the dimension table to store the dates in. The Station dimension table consists of information on the weather station and includes data such as station name, longitude and latitude, such information was loaded into the dimension table from the weather table in the original DB. The Airport Dimension table consists of the data needed to accurately represent the airport being queried in the BI reports. This has such information as Airport Name, the State and unique airport code and was loaded into the dimension table from the flight database table. Along with the information in the Airport and station, the dimension tables were given unique IDs created when coding the table.

When executing queries on the DW the idea is to link dimension tables and fact tables using BI queries to gain relevant information. Such queries similar the one seen below are used.

```
83  
84 • select AVG(DepDel), Awind from FAWFacts where StationID = '1' and DateID = '1';  
85
```

Queries like this aim to return information useful to the needs of the user by accessing the stored data in the DW.

AVG(flights.DEP_DELAY_NEW)	AWIND	StationName	FL_DATE
7.3446	11	SALT LAKE CITY INTERNATIO	2009-01-01

The idea is to return average flight delays on a certain day with the wind speed and also link the station table to show where you are querying as seen above.

Similar to the first query the next shows syntax intended to return a desired result from the DW.

```
83  
84 • select AVG(ArrDel), Snow from FAWFacts where StationID = '1' and DateID = '2';  
85
```

The intention with this query is the return the average arrival delay with the amount of snowfall at station 1 on the 2<sup>nd</sup> day and the desired result is below.

AVG(flights.ARR_DELAY_NEW)	Snow	StationName	FL_DATE
37.2700	2	SALT LAKE CITY INTERNATIO	2009-01-02

Shown above is the average delay of arrival on January the 2<sup>nd</sup> 2009, showing an average of 37-minute delay on flight arrivals and 2 inches of snowfall. The arrival delay is significantly higher on day two than the departure delay on day 1 possibly due to the snow conditions.

The next query is designed to include the overall Taxi-Out average throughout the date selected, average temperature and Airport Dim table in the results section instead of the Arrival delay, snow and Station name.

```
86  
87 • |select AVG(TaxiOut), TempAvg from FAWFacts where AirportID = '1' and DateID = '3';  
88  
89
```

The intention of queries such as this is the return data at a specific airport on a specific day in a specific condition.

AVG(flights.Taxi_Out)	TAVG	Airport_Name	FL_DATE
21.8468	20	Salt Lake City	2009-01-03

In the above query we see that a correctly executed query will return the desired attributes from the desired tables and can link each table to the fact table with ease. This could be better implemented with a fully functioning data warehouse that includes more airports, more weather stations and even over a longer period of time to enable bespoke searches on individual geographical locations at any time throughout the year to make better informed business intelligence decisions.

## MS Excel

Playing a pivotal role in the ETL process was not the only process in which Excel was utilised. Excel was also used in the consolidation of the datasets to create files for the conduction of the statistical descriptive and the running of statistical analysis tests. To consolidate the two large Excel spreadsheets into multiple smaller spreadsheets for statistical analysis, Excel formulas were written to run functions on the data that manipulate it in ways needed by the writer. The creation of these formulas automated the process of building separate spreadsheets for further statistical analysis.

For example, using an AVERAGEIF function, I was able to average up the delays of all the flights in Salt Lake City Airport over each individual day throughout the whole 9 years. I was then able to use an index match function to index each average flight delay and match it to the equivalent data within the weather

spreadsheet by matching the dates and displaying the results in a much smaller and more concise table. Instead of the 1m+ rows in the flight dataset alone, I was able to amalgamate both data sets into daily averages and reduce the rows from 1m+ to just 3288 rows. This allowed me to use the data with less effort, manipulate it in different ways and work from one spreadsheet.

Along with the writing of formulas, the building of pivot tables allowed for easier management of the crucial data. These pivot tables and formulas were used to create the required .CSV files used for importing into not just the DW as discussed earlier, but for importing to both SPSS and R for testing and also to import results to Tableau for visualisation.

Excel was also used for the execution of the time series analysis test to forecast the average flight delays over the months to come. The Holt Winder's method is used as it is an exponential smoothing method that adequately captures demand, seasonal variation and trend over time. Within this method the forecasting can be broken down into three components, which are the base level, the trend component and the seasonal factor. After the breaking down into the three components a smoothing constant can be applied to each one, for this there are three different smoothing factors, alpha, beta and gamma. First, I started out by seeding the seasonal factors, to do this I took the actual demand observation for the first period and subtract the average demand over the course of the first year. I then got the base level, I got this by taking the last month of the 2nd year and divided it by the seasonal indices of the same period. The trend is then calculated by the percentage increase per period in the base. By following the equations below base, trend and seasonality can be followed to the end of the known data within the time period.

$$L_t = \alpha \frac{x_t}{s_{t-c}} + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

$$s_t = \gamma \frac{x_t}{L_t} + (1 - \gamma)s_{t-c}$$

The above formulas were followed in the Holt Winter's calculations to enable the ability to determine an estimated trend alongside the actual trend line using L: time, T: trend and S: Seasonal Component. Using exponential smoothing, seasonality and actuals, the estimated trend line can be calculated to then forecast future data.

### **IBM SPSS**

The statistics software package SPSS was used in the testing of the data both to gather further information on the data in use and to conduct statistical analysis and test differentiation of conditions within the data. The first test ran was the descriptive, this gives further information on the data, such as the mean, median and standard deviation within the data. This enables us to better understand the data and gives us a sense of whether or not the data in use is normally distributed.

Next in SPSS I conducted the Kolmogorov-Smirnov goodness of fit test to confirm the normality of the distribution of the data. Using the above descriptive statistics and further information on the data itself the test is conducted to determine the normality of the data. Once the test of normality was conducted, based on the normality of the data either parametric or non-parametric tests are performed on the comparisons. In SPSS I compared the times of all averaged delays on days with no rainfall compared to days with rainfall, days with no snow compared to days with snow, days below 32 degrees Fahrenheit (0 degrees Celsius) compared to days above and days with 10Mph wind or below in

comparison to days above 10mph winds, to discover if there was a significant difference in delays flights under the different weather conditions.

### ***R/RStudio***

The programming language R was used to determine the level of correlation between the weather conditions and the average delay of flights over a period of time. By importing the relevant .CSV files into the R environment and writing functional code I was able to apply Pearson's Correlation R to the data and retrieve the level of correlation between the weather conditions and the length of delays on flights.

R was also used to assist in the conduction of the Holt Winter's Exponential Smoothing method. It was in R where I was able to write the relevant code to gain the most accurate values for the alpha, beta and gamma parameters that each respective trend will be passed through, in order to gain accurate evaluations and improve the forecasting ability of the equations implemented in the Excel document previously discussed.

### **Tableau**

Tableau is the environment in which I was able to both most accurately, in regards of my results, and simplistically, in regard to the UI of the end user, display the graphical visualisations of the findings. Although each of the previous technologies used were vitally important in the methods throughout the statistical analysis report, it is Tableau where I will bridge it all together. In Tableau I will visualise the outcomes of Excel, R and SPSS, it is through this visualisation that I will aim to show a greater level of explanation and make the reading and understanding of the results of the study undertaken simplistic, clear and easier to understand. In the use of Tableau, I will export the necessary Excel/.CSV spreadsheets with the results on them from the previously used technologies and import them into Tableau to manipulate them in a graphical user interface.

## Results

In the testing section the first method executed was to find the descriptive of the date being used. This was done by importing the .CSV spreadsheet I created, that amalgamated the pertinent information needed from both the flight and the weather dataset into the SPSS application. Below are the most applicable results of the descriptive tests:

		Stat Type		
		Statistic	Std. Error	
Dependent Variables	Statistics			
	AvgDelNew	Mean	7.6506	.10319
		Median	6.5000	
		Skewness	7.341	.043
		Kurtosis	114.180	.085
AvgArrNew	Mean	8.2545	.12694	
	Median	6.7000		
	Skewness	7.601	.043	
	Kurtosis	113.976	.085	
AWND	Mean	7.7823	.06067	
	Median	6.9300		
	Skewness	1.202	.043	
	Kurtosis	1.868	.085	
PRCP	Mean	.0424	.00224	
	Median	.0000		
	Skewness	4.992	.043	
	Kurtosis	35.400	.085	
Snow	Mean	.1119	.01060	
	Median	.0000		
	Skewness	8.582	.043	
	Kurtosis	91.576	.085	
TAVG	Mean	54.485	.3335	
	Median	53.000		
	Skewness	-.028	.043	
	Kurtosis	-.991	.085	

As seen above, the mean, median, skewness and kurtosis of the data columns AvgDelNew (average departure delay), AvgArrNew (average arrival delay), AWND (average wind speed), PRCP (average precipitation/rainfall), Snow (average snowfall) and TAVG (average temperature). This is the information I will concentrate most of the statistical analysis on in order to ascertain if there is in fact correlation between the selected weather conditions and flight delays.

When examining the descriptives for each individual attribute of data we can learn much about the data itself. When we take a look at both the average departure and arrival delay, firstly it is important to keep in mind that a flight is not considered delayed until it is late 15 minutes or more, so when looking at the mean of the average delays of the 9-year period the 7.65 and 8.25 minutes delays for departure and arrival respectively are 7.65 and 8.25 minutes following the first 15. When looking at the average wind speed we must remember that the information being used is from an American data source which uses the imperial system, therefore the information specified is in miles per hour and not kilometres. When measuring snowfall, it is measured in inches and rain is measured in mm. The Fahrenheit scale is used to measure temperature, for the purpose of the data used, it is key to know that 32 degrees in Fahrenheit is 0 degrees Celsius and technically at freezing point.

The descriptives gives information about the data including an indication of whether or not the data that we are dealing with is normal or not. While investigating the data further I was able to get indication that the data itself is not normal data. I came to this conclusion by first comparing the mean of each individual attribute of data with its median, generally if these two pieces of statistical evidence are close, it can indicate normalised data. However, when taking the average departure delay there is over a minute difference between the median at 6.5 and the mean at 7.65. The reading of results sections is very much about knowing your data, when dealing with time a minute difference in averages may not be considered significant, but when the averages are as small as 6-7 minutes respectively there may in fact be a significant difference. When looking at the average arrival delay we also see a difference in regards the mean of 8.25 and the median 6.7. When dealing with the wind statistics we have a mean of

7.78 miles per hour speeds and a median of 6.93 mph. The rainfall and snow show a means of 0.04 and 0.11 respectively, with both showing medians of 0.00. The mean of the average temperature is 54.48 where the median is 53. Each of the attributes are showing difference between the means and medians, however the question is, are these differences significant?

The next indicator of normalised data is the skewness of the data, this tells you if there is a positive(Right) or negative(left) skew in the distribution of the data. If the skewness value is above 1 or below -1, then the distribution of the data is highly skewed. If it is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed and if the skewness is b -0.5 and 0.5, the distribution is approximately symmetric. The closer the level of skewness is to 0 the better the more symmetrical the data. The average temperature has a skewness level of -0.28, which shows the distribution of the skewness of the data is very symmetrical. The skewness of the other attributes are as follows, average departure delay 7.34, average arrival delay 7.6, average wind 1.2, rain 4.99 and snow 8.5. We can see that the symmetric distribution for average temperature does not follow suit in the other attributes, with some showing a very high level of skewness, this indicates again that the data is not normal.

The kurtosis of the data tells you the height and sharpness of the central peak, relative to that of a standard bell curve when charting the data. Comparable to when dealing with skewness the closer the level of kurtosis is to 0, the more evenly distributed the data is. Again, just like with skewness average temperature is the closest to 0, with a value of -0.99, very close to -1 and would indicate non-symmetrical data even though it is the lowest of all the attributes. The kurtosis of the all other attributes are as follows, average departure delay 114, average arrival delay 113, average wind 1.86 rain 35.4 and snow 91.5. Just like when comparing the median and mean and with skewness, the high levels of kurtosis on the attributes tested are further indications that I am dealing with data not normally distributed.

## Kolmogorov-Smirnov

When we set out to conduct the Kolmogorov-Smirnov test we set a null and alternative hypothesis of:

H0: the data is normal

H1: the data is not normal

We aim to discover using this method if the data is normal or not, as stated in the hypotheses. However before conducting the test we must set an alpha value, which we will set at 0.05. An alpha value is the chances of making a type one error while conducted the test. Which means the chances of rejecting the null hypothesis when we should have accepted. The alpha value determines the level of percentage in which an error may have been made. Setting the alpha value at 0.05 means there is a 5% chance of a type one error occurring as in the range 1 is 100%. In other words, there is a 95% chance or 0.95 that we have not made a type one error.

Kolmogorov-Smirnov <sup>a</sup>			
	Statistic	df	Sig.
AvgDelNew	.161	3287	.000
AvgArrNew	.186	3287	.000
AWND	.119	3287	.000
PRCP	.385	3287	.000
Snow	.496	3287	.000
TAVG	.065	3287	.000

As seen in the figure above, which show the results of the Kolmogorov-Smirnov test, when we look at the sig level we see on all 6 tests conducted, the values are 0.000. When the significance value is less than 0.000 it states it as 0.000. This states that the P value for all 6 tests conducted are  $< 0.000$ , which is less than our alpha value of 0.05. Which means at an alpha value of 0.05 we can reject the null hypothesis stating the data are normal, in favour of the alternative hypothesis, the data are not normal. The performing of the test confirms what had been implied from the descriptives, that we are in fact working with not normally distributed data in the case of all 6 data types within the dataset.

## Mann-Whitney U

In the statistical analysis there were four Mann-Whitney U tests conducted overall. Each one to determine the effects each weather condition, snow, wind, temperature and rain had on the delayed flights and if there was a significant difference in compared to flights during times of clearer conditions.

### Snow

As in the test for normalisation we must set out asking a question, in the Mann-Whitney U test we are setting out to see if the average flight delays on days of no snow and the days where it does snow differ significantly.

H0: M snow = M no snow

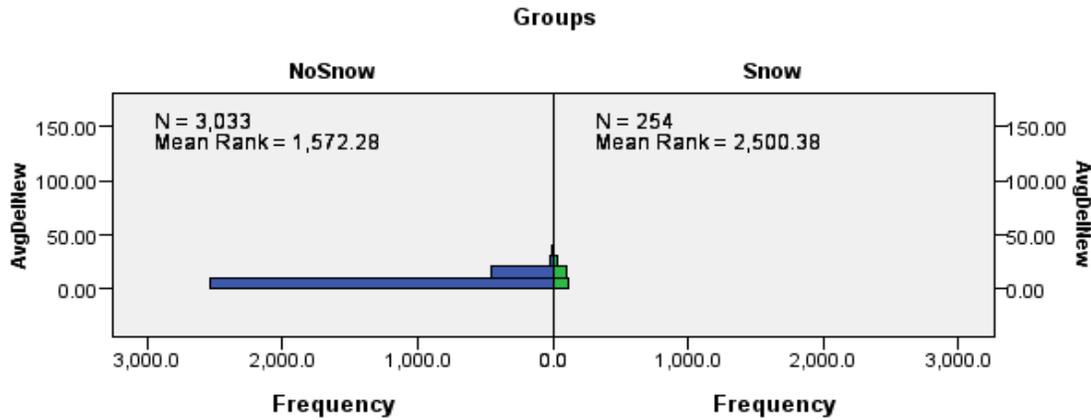
H1: M snow  $\neq$  M no snow

Alpha = 0.05

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of AvgDelNew is the same across categories of Groups.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .05.				

In the figure above, we see the hypothesis test summary, this shows the summary of the results of the conducted Mann-Whitney U test. In the first column it states the null hypothesis and in the second it confirms the analysis that was carried out. In the next column we see the sig. which the level of significance is, in the case of the Mann-Whitney, assigned the letter U. The result of the U value is  $U < 0.000$  which at an alpha value of 0.05, allows us to reject the null hypothesis in favour of the alternative.

## Independent-Samples Mann-Whitney U Test



<b>Total N</b>	3,287
<b>Mann-Whitney U</b>	167,670.500
<b>Wilcoxon W</b>	4,768,731.500
<b>Test Statistic</b>	167,670.500
<b>Standard Error</b>	14,528.740
<b>Standardized Test Statistic</b>	-14.972
<b>Asymptotic Sig. (2-sided test)</b>	.000

In the above figures we see a breakdown of the results that led to the decision to reject the null hypothesis. The top figure above shows information on the data such as the amount of days with no snow, compared to the amount of days with snow, signified by the letter N. The lower of the two figures above enables us to further evaluate the decision by supplying more information. The total N shows the amount of the overall days tested which equals 9-years and the resulting statistics for the Man-Whitney U test including the level of error and the U significance value again. We report our findings as follows:

$$U(3287) = 167670500, P < 0.000$$

Overall the results of the test shows that there is a significant difference in the length of delays on flights on days it snows, compared to days that have no snow.

### Wind

In the Mann-Whitney U test conducted on wind and its effects on delayed flights, we are setting out to see if the average flight delay on days with wind speeds greater than 10mph differ significantly than the days where the average wind speed is less than 10mph.

H0: M wind < 10mhp = M wind > 10mph

H1: M wind < 10mhp ≠ M wind > 10mph

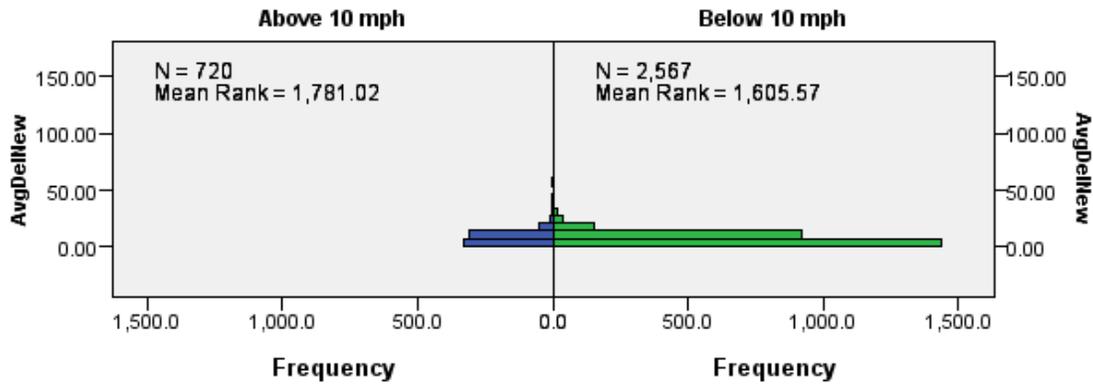
Alpha = 0.05

<b>Hypothesis Test Summary</b>				
	<b>Null Hypothesis</b>	<b>Test</b>	<b>Sig.</b>	<b>Decision</b>
1	The distribution of AvgDelNew is the same across categories of Groups.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .05.				

When examining the figure above, we see the hypothesis test summary of the Mann-Whitney U test conducted. In the first column it states the null hypothesis and in the second it confirms the statistical test carried out. In the third column we see the sig. or the U value. The result of the U value is  $U < 0.000$  which at an alpha value of 0.05, allows us to reject the null hypothesis in favour of the alternative hypothesis.

## Independent-Samples Mann-Whitney U Test

Groups



<b>Total N</b>	3,287
<b>Mann-Whitney U</b>	825,469.000
<b>Wilcoxon W</b>	4,121,497.000
<b>Test Statistic</b>	825,469.000
<b>Standard Error</b>	22,503.707
<b>Standardized Test Statistic</b>	-4.384
<b>Asymptotic Sig. (2-sided test)</b>	.000

In the breakdown of the results in the figures above we see the descriptives that led to the decision to reject the null hypothesis. The first of the two figures above show that there were 720 days with win above 10mph and 2567 days with average wind speeds of below 10mph indicated by the letter N. The second of the two figures above show the total N of 3287 which equals 9-years. The results of the Mann-Whitney U test on average wind speeds can be written as follows:

$$U(3287) = 825469000, P < 0.000$$

The conclusion of the test shows that there is a significant difference in the length of flight delays on days of high winds, compared to days of low wind speed.

### Temperature

A Mann-Whitney U test conducted on the effects of average flight delays caused at extreme low temperatures of 32 degrees Fahrenheit or less compared to that of days with average or high temperatures, above 32 degrees Fahrenheit.

H0: M positive 32 = M negative 32

H1: M positive 32  $\neq$  M negative 32

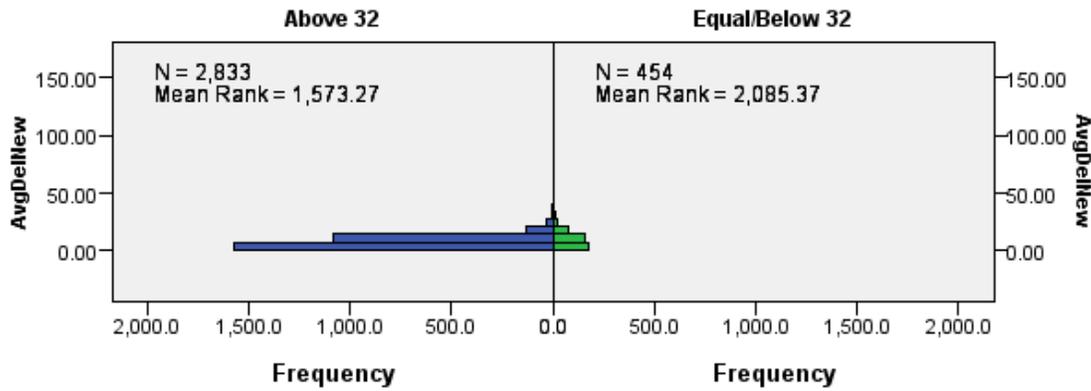
Alpha = 0.05

<b>Hypothesis Test Summary</b>				
	<b>Null Hypothesis</b>	<b>Test</b>	<b>Sig.</b>	<b>Decision</b>
1	The distribution of AvgDelNew is the same across categories of Groups.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .05.				

In the figure above, we see the hypothesis test summary of the Mann-Whitney U test on the affects different temperatures has on delayed flights, confirmed in column two of the above table. In the first column it states the null hypothesis of the test conducted. In the second to last column we see the sig. which the level of significance or what the value of U in a Mann-Whitney test is. The result of the U value is  $U < 0.000$  which at an alpha value of 0.05, states that we can reject the H0 in favour of the H1.

## Independent-Samples Mann-Whitney U Test

Groups



<b>Total N</b>	3,287
<b>Mann-Whitney U</b>	843,473.000
<b>Wilcoxon W</b>	946,758.000
<b>Test Statistic</b>	843,473.000
<b>Standard Error</b>	18,772.672
<b>Standardized Test Statistic</b>	10.674
<b>Asymptotic Sig. (2-sided test)</b>	.000

When delving further into the statistical results of the Mann-Whitney U test we can see the number of days below 32 degrees Fahrenheit and the number of days above in the first of the two figures above. Represented by the field indicated total N we see the period of time over which the test was conducted. The results of the Mann-Whitney U test on level of delays cause by weather at low temperatures compared to days at average to high temperatures are as follows:

$U(3287) = 843473000, P < 0.000$

The U result above states that as we must reject the null hypothesis, we must accept the alternative hypothesis, that there is a significant difference in the average flight delays on days of polarised temperatures.

### Rain

The final Mann-Whitney U test conducted in determining difference in the effects weather conditions has on average flight delays is the effect of rain. When conducting the test, we commenced with the question, do the level of average flight delays differ on days with no rain, compared to days when it does rain. That question is reflected below in the hypothesis:

H0: M rain = M no rain

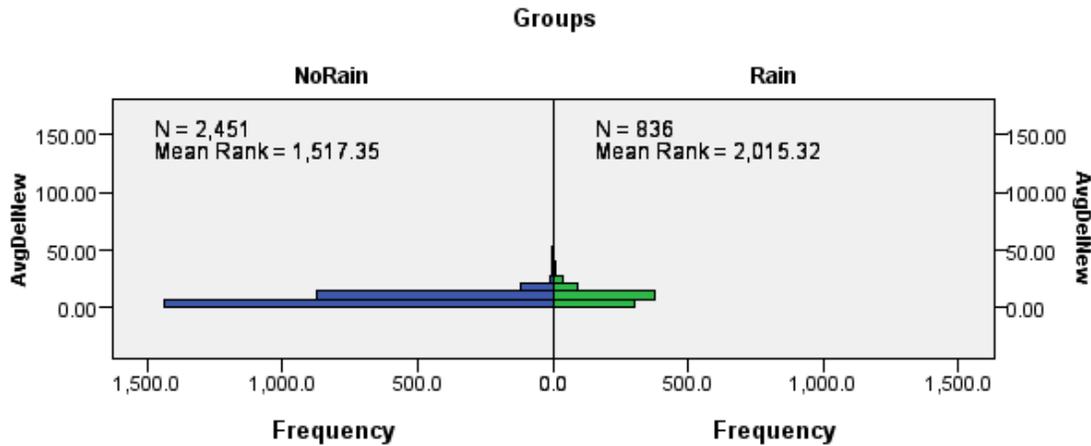
H1: M rain  $\neq$  M no rain

Alpha = 0.05

<b>Hypothesis Test Summary</b>				
	<b>Null Hypothesis</b>	<b>Test</b>	<b>Sig.</b>	<b>Decision</b>
1	The distribution of AvgDelNew is the same across categories of Groups.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .05.				

In the above figure just like previously, we see the hypothesis test summary, this shows the summary of the results of the performed Mann-Whitney U test. In the column to the left it declares the null hypothesis and in the second it confirms the name of the analysis test that was carried out on the data. In the third column we see the U value, labelled sig, of 0.000. This low value allows us at an alpha of 0.05, to reject the stated H0 in favour of the H1.

## Independent-Samples Mann-Whitney U Test



<b>Total N</b>	3,287
<b>Mann-Whitney U</b>	714,090.500
<b>Wilcoxon W</b>	3,719,016.500
<b>Test Statistic</b>	714,090.500
<b>Standard Error</b>	23,694.616
<b>Standardized Test Statistic</b>	-13.101
<b>Asymptotic Sig. (2-sided test)</b>	.000

In the top figure above, there is a breakdown of the results that led to the verdict of rejecting the null hypothesis. It shows information on the data such as the amount of days with no rain, 2451, compared to the amount of days with rain, 836. The bottom figure above enables us to further evaluate the rejection decision by providing more evidence. The total N shows the amount of the overall days tested which 3287 days or 9-years and the resulting statistics for the Man-Whitney U test. It also includes the level of standard error and the sig value at the bottom. We report our findings as follows:

$U(3287) = 714090500, P < 0.000$

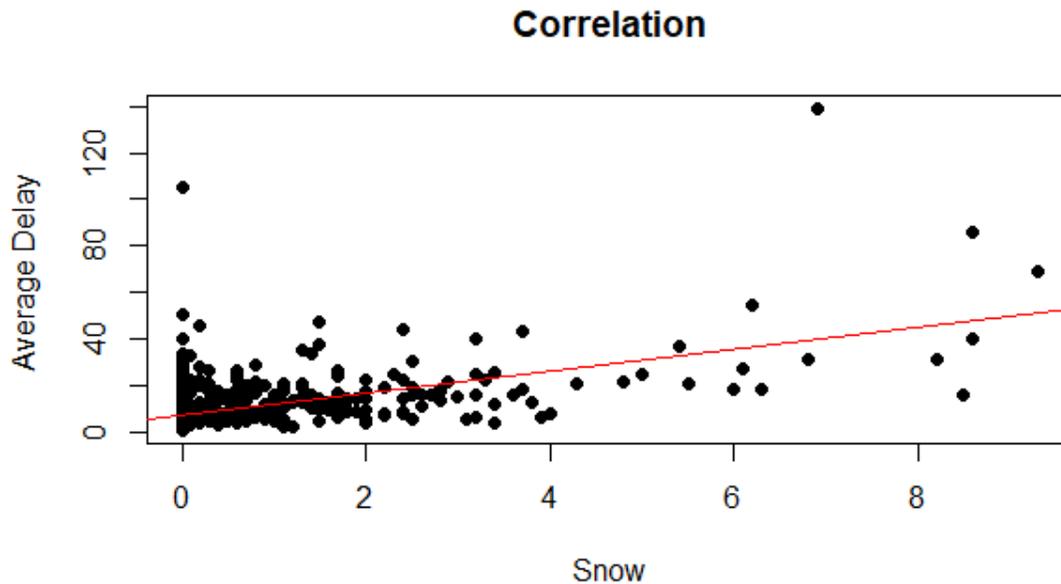
The outcome of the test shows that there is a significant difference in the length of the average flight delay in Salt Lake City airport on days it rains, compared to the days it does not rain.

### **Pearson's Correlation Coefficient**

After conducting the Mann-Whitney U tests on rain, snow, temperature and wind A significant difference in time delay of flights was found in each case. Next, I wanted to carry out a statistical experiment to see if I can discover if there is a correlation between those specific weather conditions and the delay in flights and if there is correlation, is it positive or negative and also how strong is the level of correlation? To do this I used Excel to create a .CSV file with the specific information needed to run Pearson's coefficient and imported it into R. Pearson's Coefficient is measured on a scale of -1 to 1, where negative 1 shows a total negative correlation, 1 shows a total positive correlation and 0 shows no linear correlation. On the correlation scale a correlation is considered strong when it is below -0.5 or above 0.5 respectively.

#### **Snow**

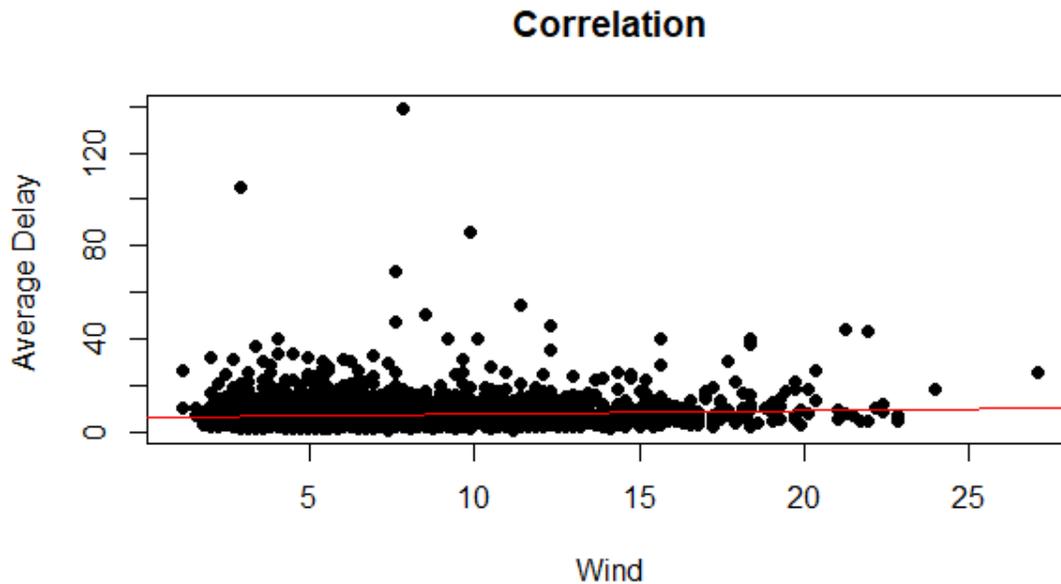
In the hypothesis that was previously stated it suggests that if P is not equal to 0 then there is correlation between the attributes tested. This would suggest that the finding of correlation between two elements is rather simplistic, but the type and level of the correlation is crucial. When using Excel and R to conduct my correlation test on snow, both returned a P value of 0.483. The P value suggests that there is correlation between the snow fall and delayed flights. But what does that value mean? The positive number proposes a positive correlation between flight delays and snow, which suggests as snow goes up, so too does the elapsed delay time of flights. The P value suggests a medium to strong correlation between snowfall and delayed flights.



The figure above shows a graphical representation of the correlation. The cluster of black dots represent the amount of snowfall shown on the x axis plotted against the delayed flights on the y axis. The red trend line shows the correlation between the two. An ascending line suggest positive correlation. The slope of the line shows the strength of the correlation. This confirms that as the levels of snow increase so too does the delay on the flight times and supports the P value in suggesting a medium to strong correlation between both.

#### Wind

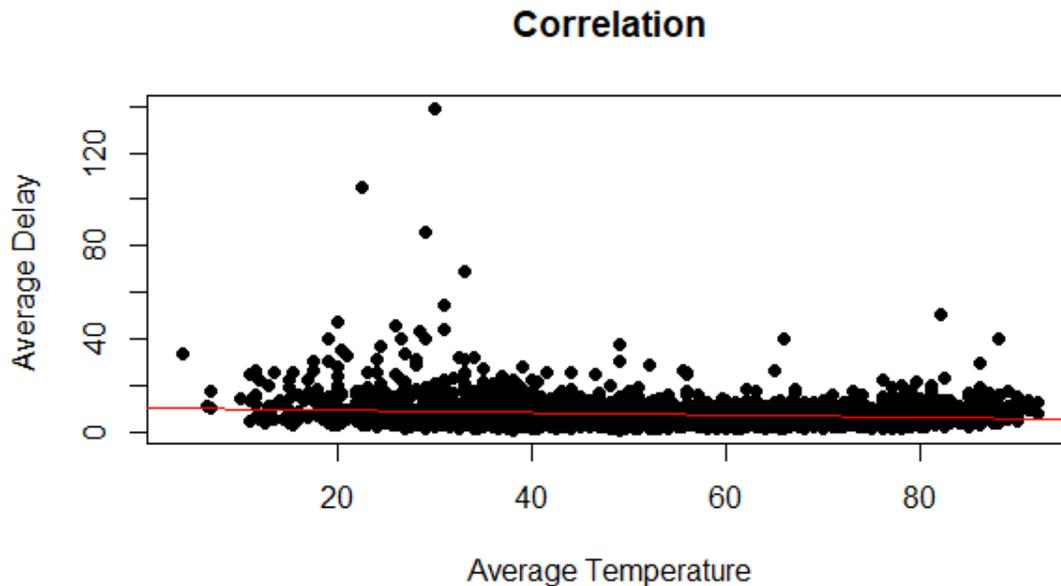
When running the correlation coefficient in R to show the level of correlation between flight delays and average wind speed, a P value of 0.084 was ascertained. The P value suggests that there is very little correlation between the wind and the effects on delayed flights as the closer the P value is to 0 the lower the level of correlation between the examined components. When plotting the graphical representation of results the below graph was produced.



The x axis we can see represents average wind speed and the y axis average delayed flights. The red trend line as we see is slightly rising suggesting that the correlation is positive, however the degrees of the angle of the rise in the slope of the line suggests a weak correlation, which supports the P value of 0.084. This shows that in Salt Lake City airport, although there is a difference in delay time on days with stronger wind speed in comparison of days with low wind speeds, there is little evidence to show that the rise in delayed flights is due to a correlation with the wind conditions themselves.

### Temperature

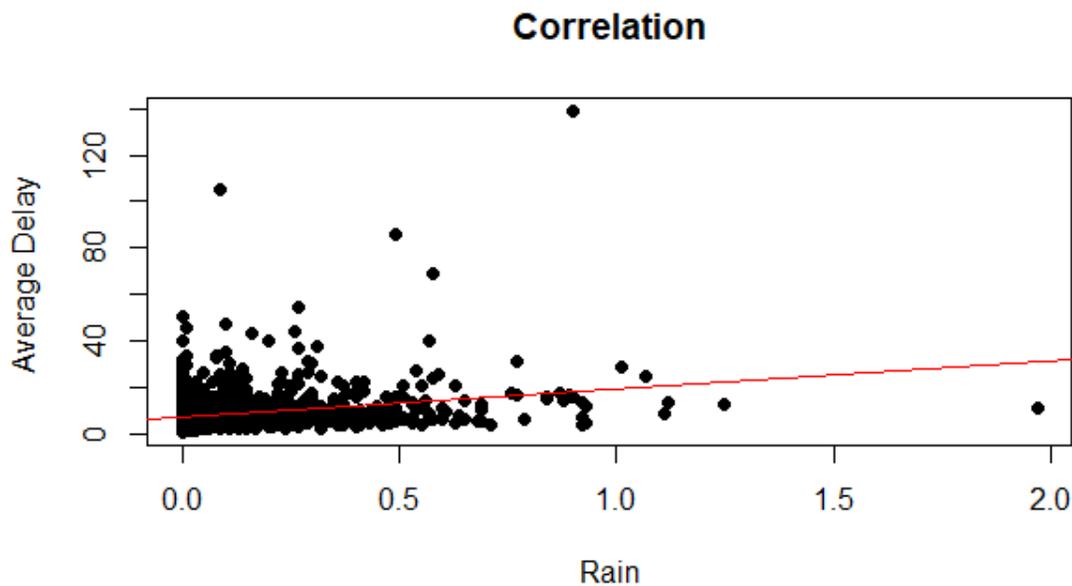
In the conducting of Pearson's Correlation Coefficient on the relationship between flight delays and average temperature and their level of correlation the established P value was -0.155. In the case of temperature, we see a negative number, this would not suggest a lack of correlation but however just a negative correlation between the two. Although there is a negative correlation the P value being so close to 0 suggests a low strength in correlation. A visual representation of the result is seen plotted on the figure below.



In the scatterplot above we see the average temperature on the x axis plotted against the average delay on the y axis. The temperatures ranges from 0 to 100, this is the case as we are using the Fahrenheit measurements taken from the imperial system in the United States of America. The red trend line shows a minor decline in slope which verifies the negative P value. The negative correlation is expected in the case of temperature as it would suggest the higher the temperature the less delays of flights, unlike snow, rain or wind, the lower number would suggest more extreme weather conditions when it comes to temperature. Although the P value shows an expected negative value, the low value shows a weak correlation with the delay times, which would indicate that the temperature at Salt Lake City has little correlation with the time delay on flights.

### Rain

When operating the statistical analysis on rainfall and the effects it has to delays flights through a Pearson's correlation R a P value of 0.26 was determined. The resulting P value would suggest that there is a small positive correlation between the average rainfall in Salt Lake City and the length of delayed flights. The P value again is quite low, just like with temperature and wind and can be seen visualised on the scatterplot graph below.



Similar to the previous scatterplots we see the average delay displayed on the y axis and in this case, we see the average rainfall on the x axis. As we would have gathered from the P value the trend line is rising to show a positive correlation between the amount of rainfall and length of flight delays. The lack of slope on the trend gives visual support to the low P value of 0.26, which shows a weak to medium correlation. Again, although a significant difference was found between days of rainfall and days of no rainfall as regards to delayed flights, the results of the correlation test would suggest that is not a direct correlation between both.

### **Holt Winter's Exponential Smoothing**

In order to carry out the Holt Winters analysis I first created an Excel spreadsheet for the actuals of the datasets. I created a month average flight delay for each month of the 9 years by writing excel formulas to average up each flight delay on each day of the respective months. I then began to gather the information needed to carry out the equations on the data. The first step was to gather the trend in the average delay. To do this I got the monthly average of the first two years individually, divided one by the other and got  $1/12^{\text{th}}$  the power of the answer, I used

12 as we are dealing with months. Next, I gathered the seasonal indices, to get this I started by getting the average of the current month of the previous two years. In the first case I was dealing with January 2011, so I got the average of January 2009 and 2010. I then divided that answer by the average of each month over the two-year period to give some seasonal indices of the previous two years. Alpha, beta and gamma were gathered through R. While working on the Holt Winters method in R, I ran a multiplicative check on the dataset in order to gain the most accurate figures of the alpha, beta and gamma parameters. The initial base value was determined by dividing the seasonal indices by the actuals for the month. The future base values were then calculated using the following formula passed through the alpha parameters determined through R.

Formula bar:  $=\$G\$11*(B26/G14)+(1-\$G\$11)*(C25*D25)$

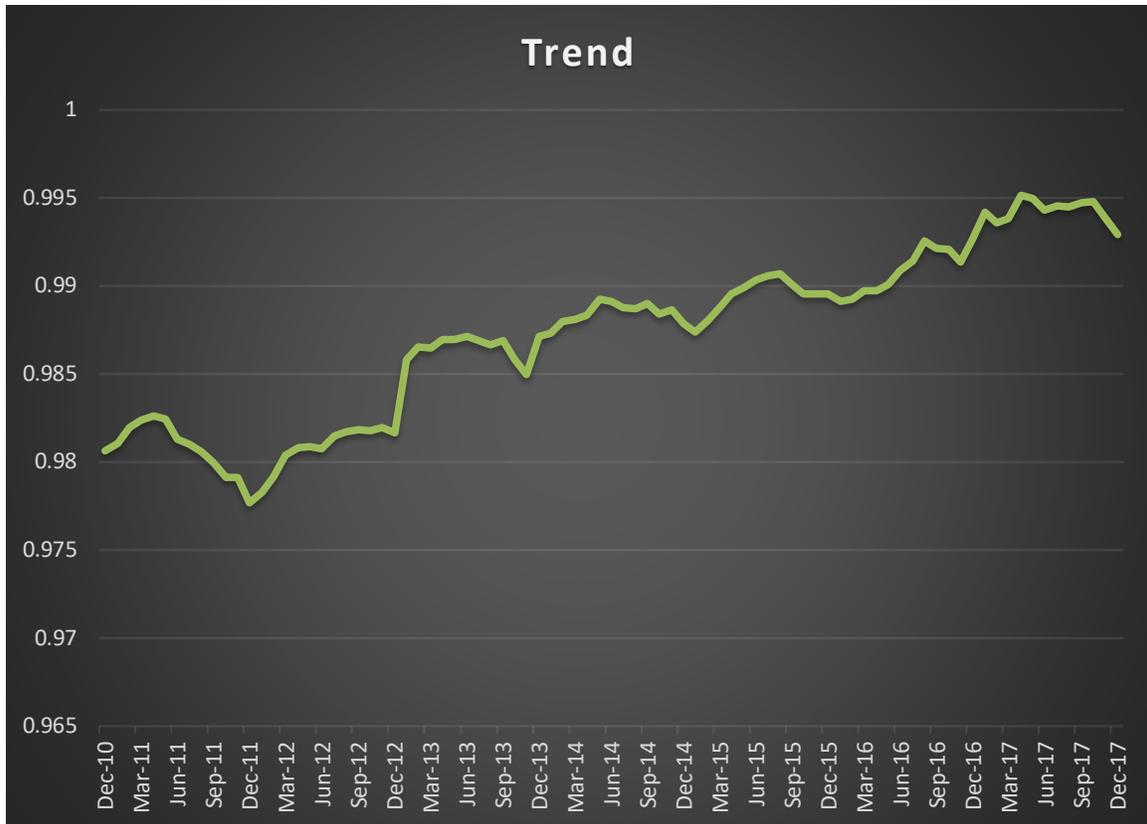
A	B	C	D	E	F	G	H	I
Sep-09	4.44					alpha	beta	gamma
Oct-09	7.52					0.31	0.01	0.40
Nov-09	3.12							
Dec-09	12.72					seasonal indices		
Jan-10	7.22					1.074202209		
Feb-10	6.08					0.805491099		
Mar-10	6.15					0.818037299		
Apr-10	6.50					0.752868701		
May-10	6.99					0.721452445		
Jun-10	8.52					1.133258795		
Jul-10	9.29					1.083109111		
Aug-10	8.31			MAPE	0.1677972	1.005118166		
Sep-10	6.51					0.733317008		
Oct-10	7.40					0.999723463		
Nov-10	10.91	Base	Trend	Forecast	APE	0.939764735		
Dec-10	16.14	8.348721	0.980636			1.933656968		
Jan-11	10.03	$C25*D25$	0.981064	8.7945523	0.1234447	1.114208691		
Feb-11	8.70	9.131334	0.98194	6.7521735	0.2236602	0.86428811		

The trend formula used to determine future trends from the second month onwards is show on the image below. As you can see, this formula is passed through the beta parameter determined in R.

M    X    ✓    fx    = $\$H\$11*(C26/C25)+(1-\$H\$11)*D25$

A	B	C	D	E	F	G	H	I
Sep-09	4.44					alpha	beta	gamma
Oct-09	7.52					0.31	0.01	0.40
Nov-09	3.12							
Dec-09	12.72					seasonal indices		
Jan-10	7.22					1.074202209		
Feb-10	6.08					0.805491099		
Mar-10	6.15					0.818037299		
Apr-10	6.50					0.752868701		
May-10	6.99					0.721452445		
Jun-10	8.52					1.133258795		
Jul-10	9.29					1.083109111		
Aug-10	8.31			MAPE	0.1677972	1.005118166		
Sep-10	6.51					0.733317008		
Oct-10	7.40					0.999723463		
Nov-10	10.91	Base	Trend	Forecast	APE	0.939764735		
Dec-10	16.14	8.348721	0.980636			1.933656968		
Jan-11	10.03	8.544478	D25	8.7945523	0.1234447	1.114208691		
Feb-11	8.70	9.131334	0.98194	6.7521735	0.2236602	0.86428811		
Mar-11	8.41	9.373558	0.982386	7.3348686	0.1277596	0.849671315		
Apr-11	7.46	9.423816	0.982616	6.9327556	0.0701513	0.76818703		
May-11	6.25	9.075666	0.98242	6.6806424	0.068617	0.708407006		

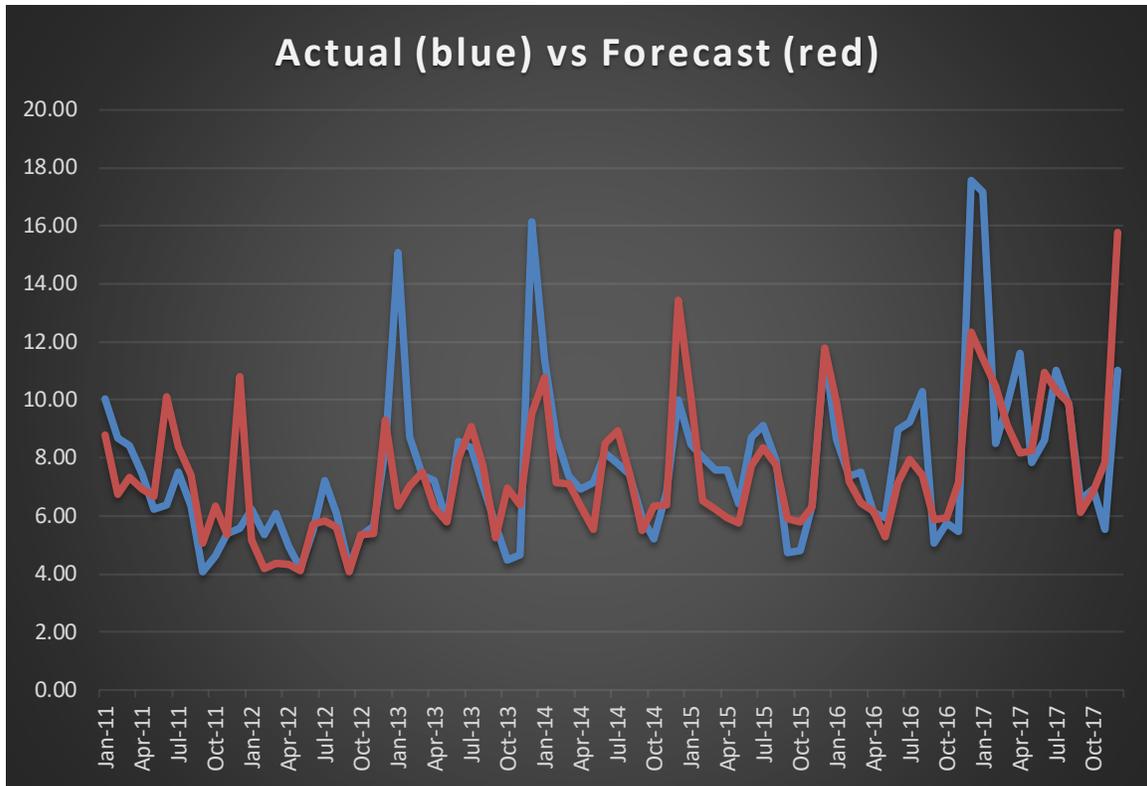
Using this equation, we can then graph the trend in the data over the desired period of time. The line graph below shows the month on month trend over the course of December 2010 to December 2017. As seen in the graph the trend stayed similar throughout the 8 year period showing a slight and steadily incline.



The forecast is then calculated by multiplying the base and trend of the previous month and multiplying the answer by the season indices of the previous month also. The APE value is gathered by getting the absolute value of the actuals of the month minus the forecast and dividing the result by the actuals of the month too. The APE is used to measure the prediction accuracy of your forecast and a score close to 0 is desirable as 0 indicates 100% successful predictions. To evaluate the seasonal indices for the second year and onwards another formula was created as seen in the image below. The seasonal indices are passed through the gamma parameter ascertained in R for more precise answers.

A	B	C	D	E	F	G	H	I
Sep-09	4.44					alpha	beta	gamma
Oct-09	7.52					0.31	0.01	0.40
Nov-09	3.12							
Dec-09	12.72					seasonal indices		
Jan-10	7.22					1.074202209		
Feb-10	6.08					0.805491099		
Mar-10	6.15					0.818037299		
Apr-10	6.50					0.752868701		
May-10	6.99					0.721452445		
Jun-10	8.52					1.133258795		
Jul-10	9.29					1.083109111		
Aug-10	8.31			MAPE	0.1677972	1.005118166		
Sep-10	6.51					0.733317008		
Oct-10	7.40					0.999723463		
Nov-10	10.91	Base	Trend	Forecast	APE	0.939764735		
Dec-10	16.14	8.348721	0.980636			1.933656968		
Jan-11	10.03	8.544478	0.981064	8.7945523	0.1234447	(1-\$I\$11)*G14		
Feb-11	8.70	9.131334	0.98194	6.7521735	0.2236602	0.86428811		
Mar-11	8.41	9.373558	0.982386	7.3348686	0.1277596	0.849671316		

The MAPE shows the mean of the APE, this is the overall accuracy of the predicted forecast. On an average over the 7-year period when comparing the forecast to the actuals, it was returned to have an 84% success rate. This shows a high level of prediction accuracy for such a complex topic such as delayed flights when considering seasonality, weather conditions and all other issues that arise to determine the smooth operating of all processes within the flow of an airport with so many flights on a day to day basis.



The line chart shown above is a graphical representation of the actuals versus the forecasted average delayed flight time each month. The blue line shows the actuals and the red shows the forecasted number over the 7-year period. We can see above, the visuals of how accurate the forecast can be, although forecast can be off in certain points this graph can give support to the 84% accuracy of the forecasting of future average delay time.

### Tableau

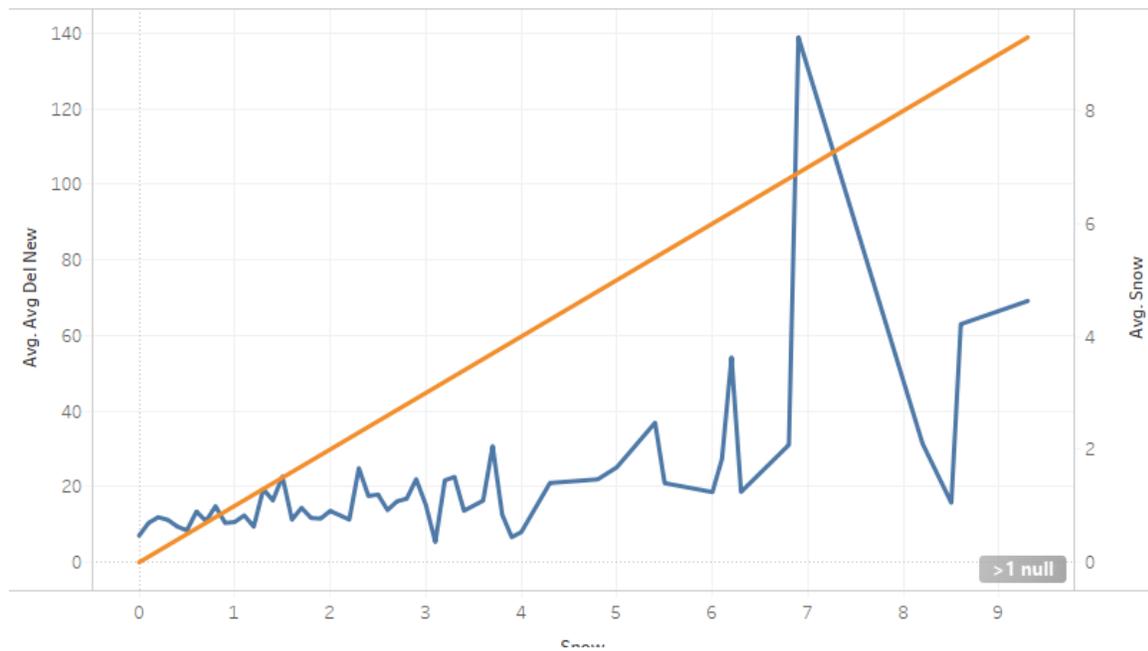
Tableau is the software application used to visualise the data within the analytical study. I have created dashboards that will be made available on Tableau Public to view [here](#). The Tableau Public shows the live Tableau Dashboards hosted on Tableau server.

### Snow

The first results from Tableau we look at is that of the flight delays and the effects they receive from snow. In the first of the 3 graphs on the snow dashboard in Tableau we see two lines on a chart. The blue wave line represents the average

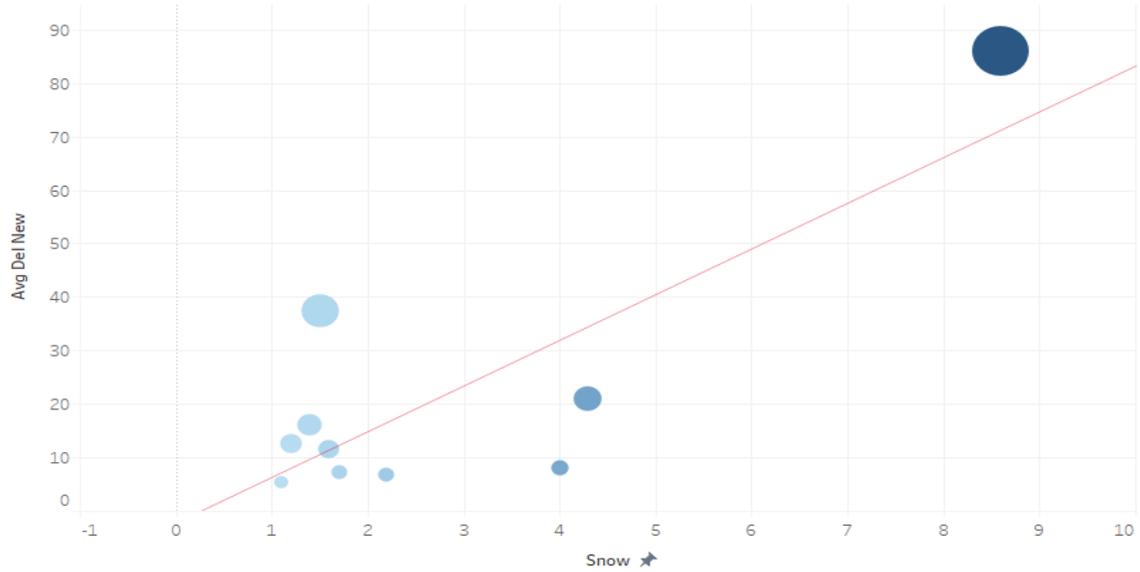
delay and the gold line represents the snowfall. As the gold line rises so does the level of snowfall. This visualisation shows that at low snowfall, there is little fluctuation in the wave of the blue line representing average delay and also relatively short delays. As the gold line increases so too does the level of the delay represented by the blue line.

Snow



The second of the three charts in the dashboard shows the correlation between flight delays and snow. On the x axis we have snow and on the y axis we have average delay. For the sake of the display I have removed all days with 0 snowfall, this removes the cluster of balls representing 0 snowfall for a clearer graph but also increases the correlation trend line as it only shows days affected by snow and shows the increase in flight delay as a relationship with the increase of the amount of snowfall.

SnowCor(no 0)



The concluding graph shows the trend lines for both average delay (top) and snowfall (bottom). When looking at the snowfall we see blue spikes at certain points in almost every year, at these points so too do we see a spike in average delays. In the years 2012 and 2015 however, there are minimal peaks in snow, which is also represented in the delays. This shows a relationship in trends for snowfall and flight delay.

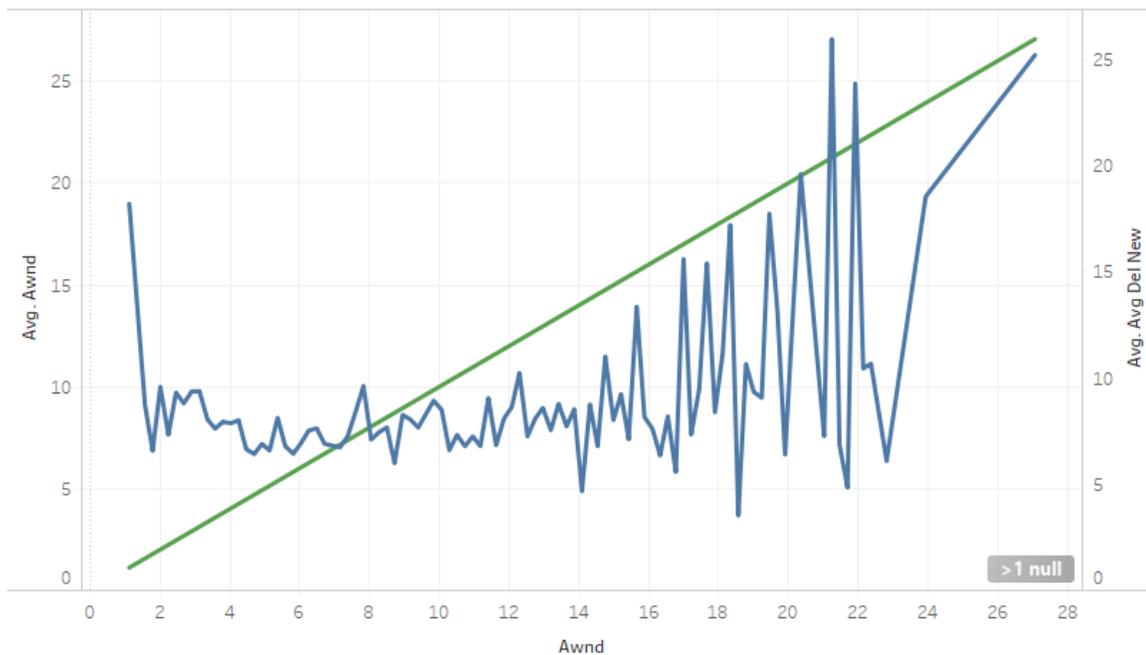
SnowTrend



## Wind

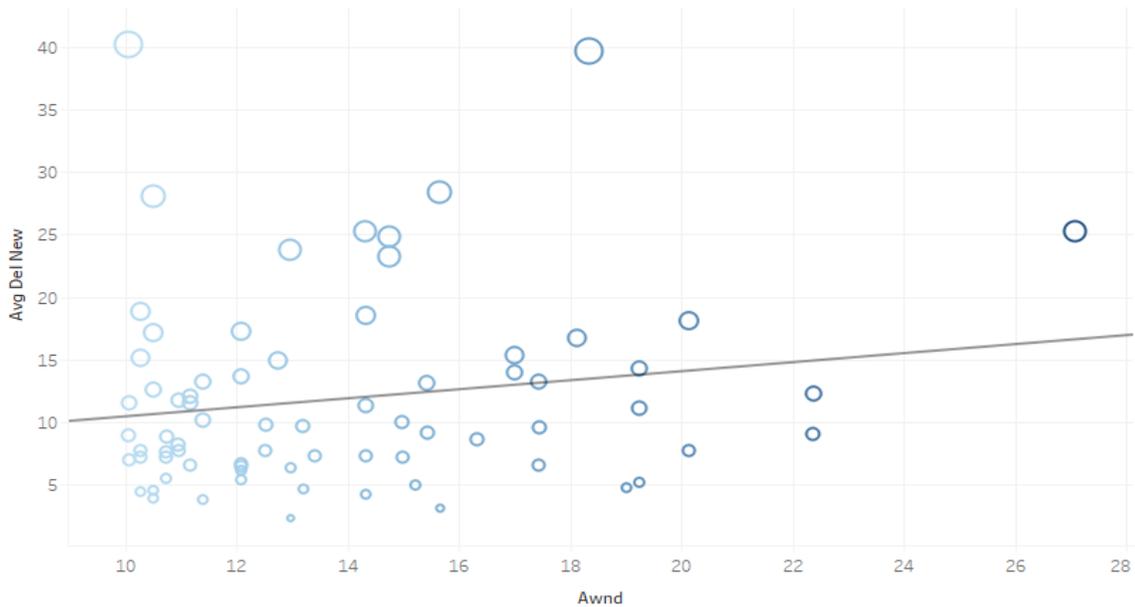
Just like in the section for snow we also have 3 charts showing a graphical representation of the statistical analysis retrieved from Tableau. The immediate graph below shows a line chart representing wind (x axis) and average flight delays (y axis). The blue line below represents average flight delay and the green line wind speed. On the left-hand side of the chart we see a slight variation in delays when the wind is very low which evens itself out throughout the middle. As we move further right and as the wind speed increases, the delay line becomes less stable and the delays times themselves increase.

## Wind



The correlation between wind and delays comes next, for the purpose of the experiment I have removed all of the days with wind speeds of less than 10 mph (similarly to the separation for the Mann-Whitney U test conducted earlier). This removes the ball of cluster as lower days for a nicer visualisation but also shows a more sloped trend line, leading to correlation in the delayed flights as the wind speed surges.

WindCor(+10)



The final graph on the dashboard for wind shows the trend lines of both average delays and average wind speed of a seasonality course throughout the years. From the graph we can see an ascending and descending line on both wind and delays, this implies a relationship of sorts between the weather condition and the length of time the flights will be delayed during those months.

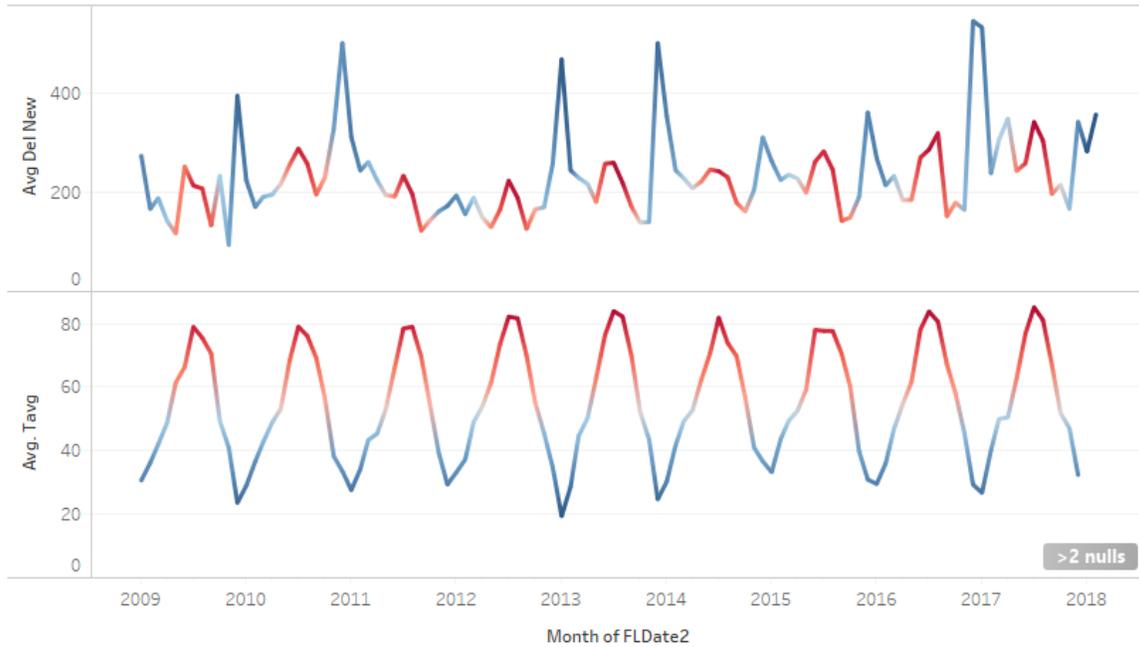
WindTrend



## Temperature

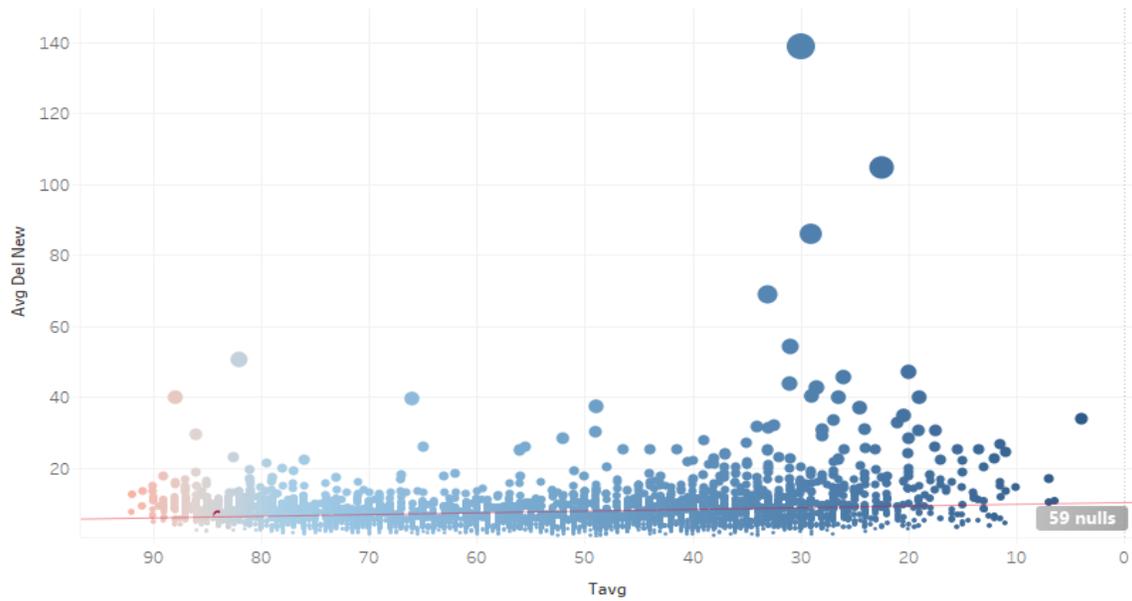
The next section shows the visualisation for temperature as seen below. The first graph shows the seasonality trend of the rise and fall of the temperature in comparison to the length of delayed flights over the same period of time.

TempTrend



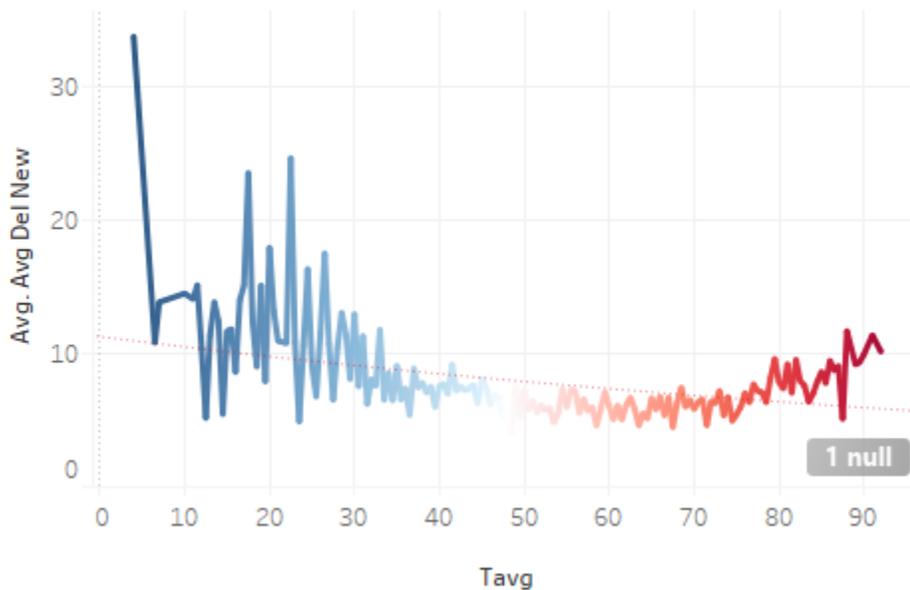
The graph below shows a scatterplot of the average temperature mapped against the average delayed flight over the 9-year period. It also shows the correlation trend line to show a slight positive correlation as the slope of the line slightly ascends.

## TempCor



The final temperature graph below shows the temperature on the x axis and the flight delays on the y axis. The blue end of the graph shows the low temperature and the fluctuation of flight delays is clear at the extreme left. As we see the line smoothens out the further right it does and the higher the temperature becomes showing both less delays and less of a range of delayed times too.

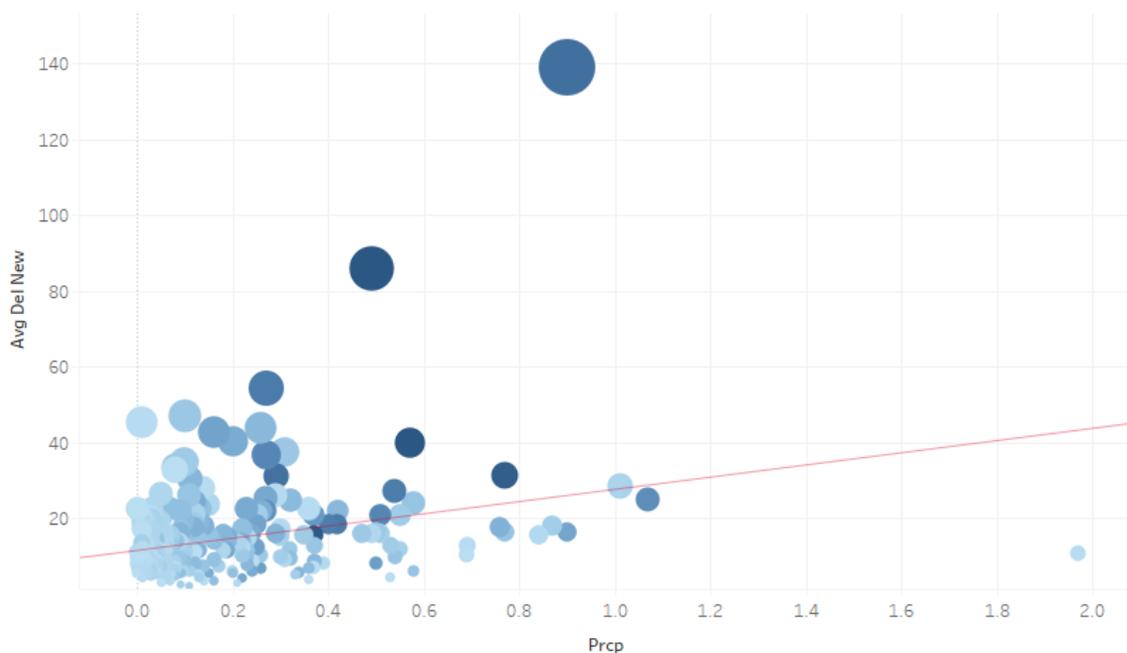
## Temperature



## Rain

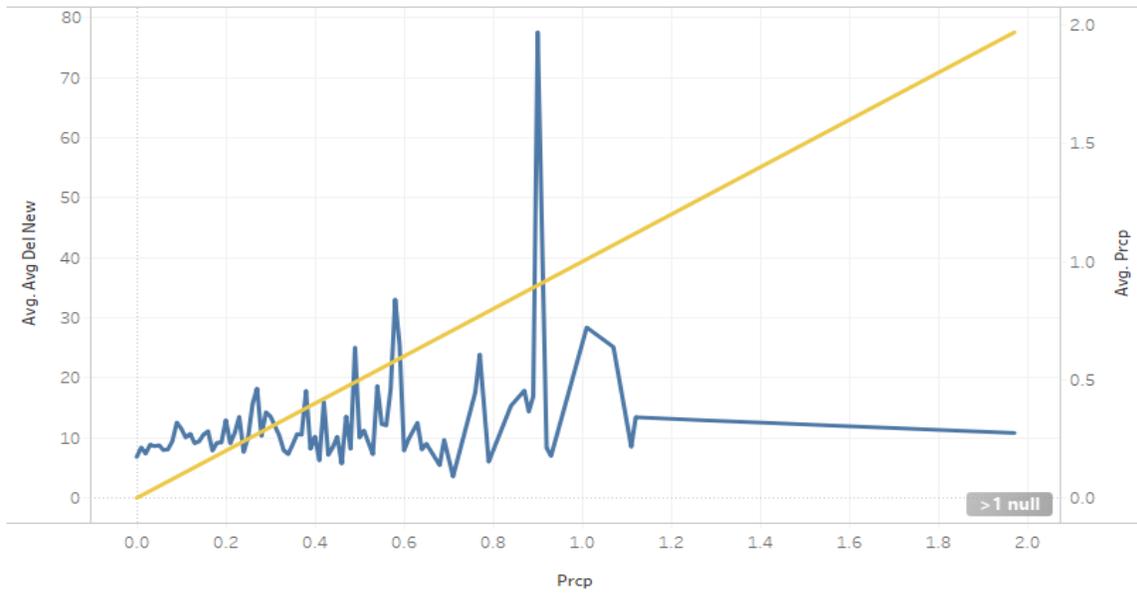
The final section shows the 3 different graphs included in the rain dashboard. The first graph shows the correlation on a scatterplot with average rain on the x axis and average delay on the y axis. The correlation trend line is included to show the positive weak correlation between the two. For the sake of the Tableau graph below all days with 0 rain were removed to show a clearer graphical representation and a more accurate correlation of days that include rain and the impact these days have on flight delays.

RainCor(no 0)



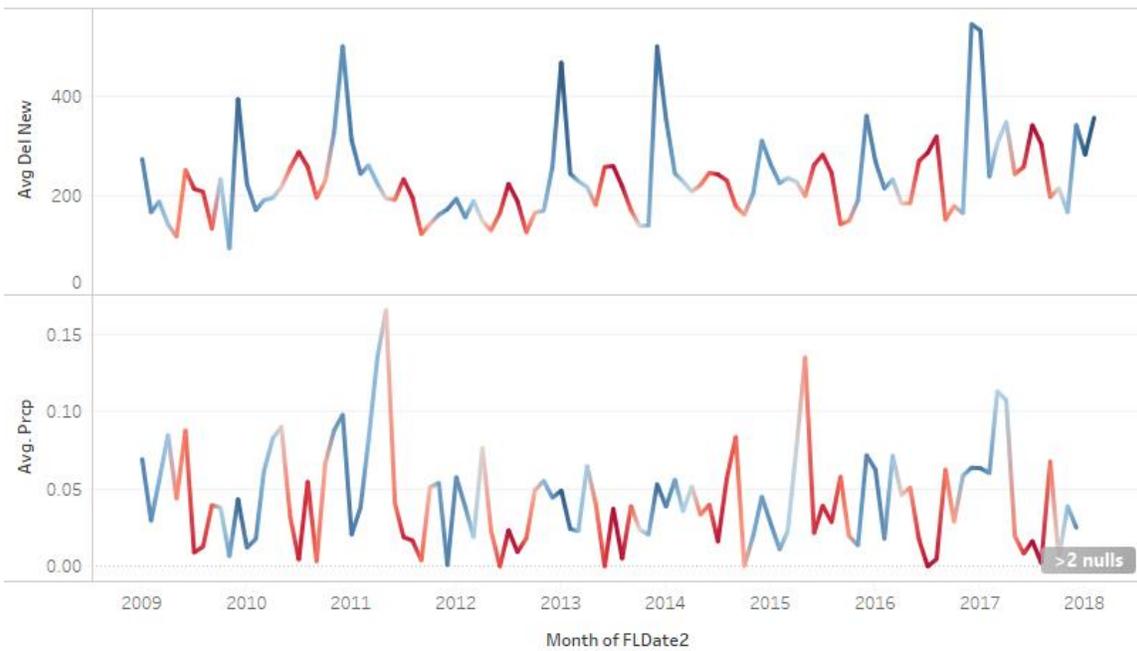
On the graph below we see a wave line in blue, which represents the average delayed flight time and a gold line to represent the level of precipitation. As we see on the chart, as the gold line increases so too does the fluctuation of the flight delays. The average delay line does smooth out at the end which is caused by an outlier in the data. For more accurate data some statisticians remove outliers to hinder the skewness of the data.

### Rain

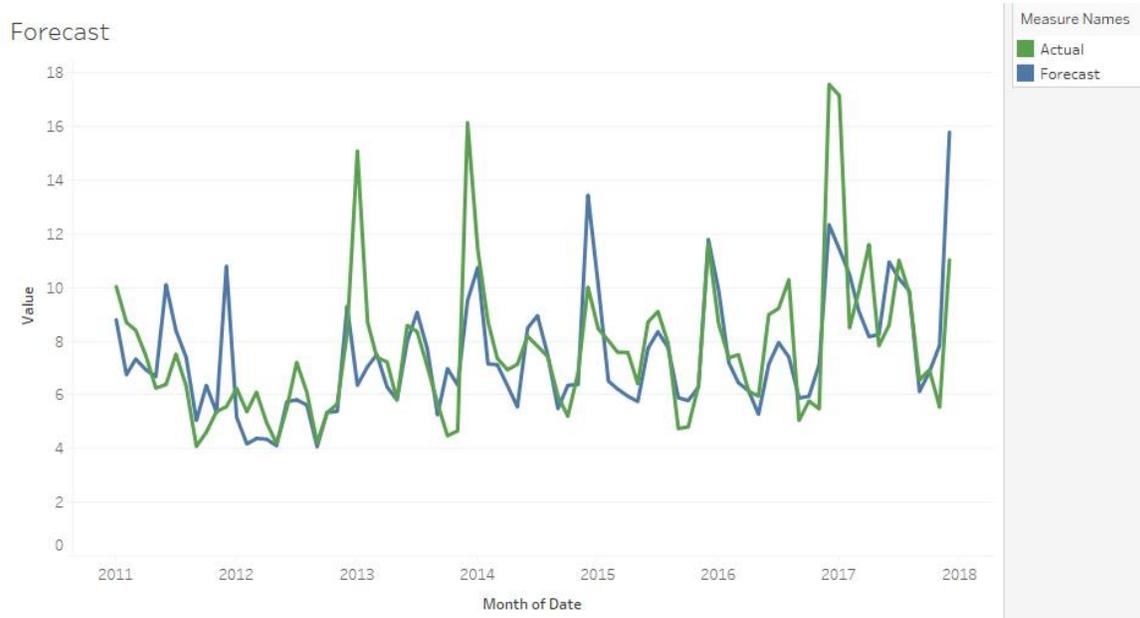


The last graph shows the seasonality trend of the average rain fall (lower) and the average flight delay(upper). For the most part we can see in the graph both delays and rain seem to ascend and descend at similar stages showing a trend in both over the 9-year period.

### RainTrend



The Tableau line chart below shows a graphical representation to support results of the trend forecasted by the Holt Winter's Exponential Smoothing executed in Excel. Through Tableau the results were imported and graphed and the published to Tableau Public with the range of dashboards created throughout the statistical report.



## Testing

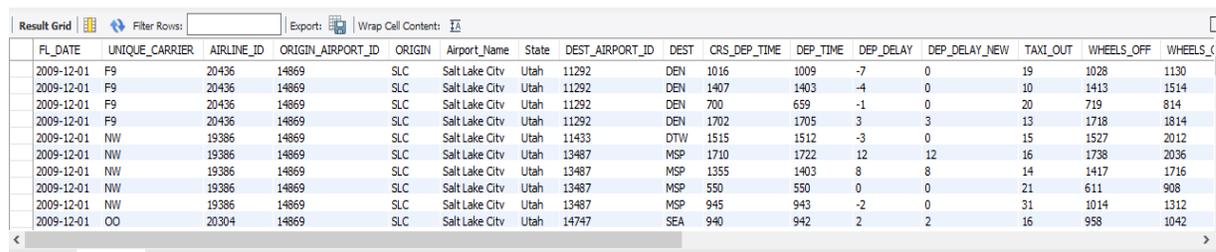
Testing is a vital part of any project, be it the development of technology or the conducting of a statistical analysis report such as this, testing is a way in which we confirm the product/analysis is functioning correctly and also how we can verify the accuracy of the data and results in analytical reports. Testing is important to do throughout the production process and not just at the end, if there is an error in it less costly and more efficient to find it as early along the production timeline as possible.

### Datawarehouse

In the development of the DW testing both of the creation of the DW and the loading of the data was conducted by running some BI queries to make sure the code worked, and the data returned was accurate. The first query was executed to check the creation of the flights table and data.

```
52
53 /* Test queries for successful import */
54
55 • SELECT * FROM Flights;
56 • SELECT * from Flights WHERE FL_Date='2009-12-01';
57
```

The syntax above runs a query on the flights table commanding the return of all data on the date 2009-12-01.



FL_DATE	UNIQUE_CARRIER	AIRLINE_ID	ORIGIN_AIRPORT_ID	ORIGIN	Airport_Name	State	DEST_AIRPORT_ID	DEST	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	DEP_DELAY_NEW	TAXI_OUT	WHEELS_OFF	WHEELS_C
2009-12-01	F9	20436	14869	SLC	Salt Lake City	Utah	11292	DEN	1016	1009	-7	0	19	1028	1130
2009-12-01	F9	20436	14869	SLC	Salt Lake City	Utah	11292	DEN	1407	1403	-4	0	10	1413	1514
2009-12-01	F9	20436	14869	SLC	Salt Lake City	Utah	11292	DEN	700	659	-1	0	20	719	814
2009-12-01	F9	20436	14869	SLC	Salt Lake City	Utah	11292	DEN	1702	1705	3	3	13	1718	1814
2009-12-01	NW	19386	14869	SLC	Salt Lake City	Utah	11433	DTW	1515	1512	-3	0	15	1527	2012
2009-12-01	NW	19386	14869	SLC	Salt Lake City	Utah	13487	MSP	1710	1722	12	12	16	1738	2036
2009-12-01	NW	19386	14869	SLC	Salt Lake City	Utah	13487	MSP	1355	1403	8	8	14	1417	1716
2009-12-01	NW	19386	14869	SLC	Salt Lake City	Utah	13487	MSP	550	550	0	0	21	611	908
2009-12-01	NW	19386	14869	SLC	Salt Lake City	Utah	13487	MSP	945	943	-2	0	31	1014	1312
2009-12-01	OO	20304	14869	SLC	Salt Lake City	Utah	14747	SEA	940	942	2	2	16	958	1042

The above is a snippet of the results retrieved from the DW, which I cross referenced with the Excel file imported into the DW to test for accuracy also. When creating the flights tables similar commands were performed on the table to test the code and the successful import of the .CSV files created in Excel.

```

84
85 /* Test queries for successful import */
86
87 • SELECT * FROM Weather;
88 • SELECT * FROM Weather WHERE Wdate='2010-12-19';

```

This test code commands MySQL to return all data in the Weather table where the date is equal to 2010-12-19. I specifically chose a date different from the first test done on the flight table to check the integrity of another section of the data dumps.

STATION	StationName	Latitude	Longitude	Elevation	Wdate	AWND	FMTM	PGTM	PRCP	Snow	SNWD	TAVG	Tmax	Tmin	WDF2	WDF5	WESD	WSF2	WSF5
USW00024127	SALT LAKE CITY INTERNATIO	41	-112	1288	2010-12-19	11	1859	1858	1	0	0	42	51	32	170	170	0	24	32

The weather table returned the above findings when completing the directive. This data again was cross referenced in the original Excel file to examine the accuracy of the data import and the storage of the DW.

**Correlation**

When carrying out Pearson’s Correlation Coefficient in order to determine the accuracy of the level or correlation between the 4 different weather conditions and the average length of delayed flights, I carried out the test in separate applications to verify the results. I initially carried the test out in Excel to verify the strength of the relationship of each condition along with delays.

	A	B	C	D	E	F	G
1		<i>FLDate2</i>	<i>AvgDelNew</i>	<i>AWND</i>	<i>PRCP</i>	<i>Snow</i>	<i>TAVG</i>
2	FLDate2	1					
3	AvgDelNe	0.1076	1				
4	AWND	0.0247	0.0844	1			
5	PRCP	-0.0176	0.2639	0.1611	1		
6	Snow	-0.0238	0.4831	0.0540	0.3674	1	
7	TAVG	0.0934	-0.1554	0.3386	-0.0903	-0.2132	1
8							

The highlighted cells above show the level of correlation with flight delays and wind (C4), rain (C5), snow (C6) and temperature (C7). In order to verify these numbers, I then ran a correlation test for each in R also.

```

> ##### Testing Correlation
> cor(Correlation$AvgDelNew, Correlation$Snow)
[1] 0.4831415
> cor(Correlation$AvgDelNew, Correlation$AWND)
[1] 0.08439345
> cor(Correlation$AvgDelNew, Correlation$PRCP)
[1] 0.2639435
> cor(Correlation$AvgDelNew, Correlation$TAVG)
[1] -0.1554065
> |

```

The above R results show the exact same stats in regard to the level of correlation for snow (top) wind (2<sup>nd</sup>), rain (3<sup>rd</sup>) and temperature (4<sup>th</sup>) as shown in the Excel figure, confirming the accuracy of the test conducted and the level of correlation between flight delays and each weather condition is considered verification testing.

### Forecasting

In the time series analysis conducted to try to predict the average delays over the 7-year period based on the historical data received from the NCDC and USDOT returned an 84% MAPE rate of accuracy. The Holt Winter's method in itself was used as a form of testing. By running the forecast method into the first 2 months of 2018 (most up to date data available on USDOT) to try to forecast for data not yet known, I was able to make a prediction of the length of the average delays of the course of the first two months of the new year.

	A	B	C	D	E	F	G
	Date	AvgDel			Forecast	APE	
0	Jan-18	9.09	7.094244	0.992194	11.89184719	0.3085838	1.444627587
1	Feb-18	12.72			7.367712944	0.4209129	
2							

The above Excel snippet shows the level of forecasting accuracy for January and February 2018. The APE in column F shows that the forecast was out by 30%, showing a 70% accurate rate of forecasting for January and a 58% accurate rate of forecast for February. This APE rate is significantly lower than the MAPE of 84% accuracy rate tested from the previous 7 years which may indicate other unexpected causes of delay for the initial months of 2018. An overall test accuracy

of 84% when predicting a flight delays, something affected by so many variables is more than desirable.

## **Future Opportunities**

The potential future prospects of having the ability to run statistical analysis not only on delays that weather causes to flights in general, but to have the potential to drill down into each individual airport and conduct the research on low level data to determine what climate conditions the airport is in through geographical location and then determine how the weather specific to that area will affect the delayed flights. This type of airport specific data can be used by airlines themselves to predict staff needs for certain days, weeks or months of the year saving costs on wages alone. The ability to anticipate the delays based on statistical analysis can also save processing times and cost if proper preparation is implemented with this knowledge. The ability for airline companies to forecast with an 84% success rate an issue that costs the aviation industry billions each year, can have amazing potential to better processes throughout airports in the United States of America and worldwide.

### **Further Analysis**

Given the chance to implement a system such as this I would first conduct further analysis on the industry as a whole to better forecast the issues. There are some conditions which we cannot predict that will affect flight delays, such as technical issues and labour strikes, that will skew data throughout the year however another major cause of delayed flights in the United States of America is the strict security procedures. The length of time in which it takes to pass through security at airports has further knock-on effects to flights throughout the country. If conducting further analysis in the field this is where I would begin. By gaining data on the average footfall in the airport in which the study is being conducted, in the case of mine let's take Salt Lake City again. I would then map this against the length of time the average security check takes, with the amount of personnel on duty at the given times of the day to gain an average of the expected time to pass through security and begin a correlation analysis on the results. A multivariate analysis of these conditions along with the results of the effects of weather conditions could have the potential to forecast the average delay to an even greater MAPE percentage

than 84%. If correlation is shown on both, there could be massive potential for the use of this data within the aviation industry.

Further analysis to consider would be the affect weather may have on each aircraft type and the level of delay based on the aircraft. Are there aircrafts affected and slowed down in such conditions more so than others? Are there aircrafts more susceptible to damages in certain climates? Analysis on the relevant data could potentially return the best aircrafts to fly in certain conditions in. This can reduce cost of maintenance on the aircrafts if airline companies can select aircraft types to fly based off conditions. It could also increase reputation and brand by reducing delays if it is indeed proven that delays are higher with certain aircrafts, as planning the aircrafts to be used in different circumstances can up efficiency, lower delay time and make for a happier customer base.

If given the opportunity of taking the statistical analysis further I would also run a similar study on another airport in the United States of America of similar size ideally, I would choose an airport in a location with very different climate conditions. If conducting similar research, I could then compare it against Salt Lake City airport in Utah to see if climate in an area does in fact effect flights and their time delays differently from area to area and airport to airport.

## Conclusion

Throughout the analysis of weather and flight delays I created a DW as a repository to store and manipulate the flight and weather data sets as needed. I created this also to enable the addition of further data to be inserted based on different airports. The data was then examined to test for normality in order to determine which analytical tests I would conduct. Then the Mann-Whitney U tests were conducted on all 4 conditions, all of which returned a significant difference between low and high weather conditions compared to flight delay average times. When investigating the correlation, I then found that although there were differences we could not correlate those difference to the weather itself as 3 out of 4 weather conditions had a weak relationship when it came to levels of correlation. In forecasting the average delays using a time series analysis, I was able to forecast with an 84% accuracy the level of delay for the coming months. As shown in the forecasting for 2018 however some months are as high as 40% off but in contrast other months are 99% accurate.

In the beginning of the report I set out to discover if there was in fact a high level of correlation between the level of flight delays and the weather affecting the flights throughout the year. It is important in statistics not to assume a result as it is not about proving the connection exists, but it is about gathering findings even if the answer to the initial question is that there is no relationship or very little, this is still a very significant finding regarding the aviation industry and in regard my report also.

## Appendix

National College of Ireland  
BSc in Business Information Systems  
2017/2018

Ian Donnelly  
X14111659  
x14111659@student.ncirl.ie

Correlation of Airline Flight delays with external data  
Technical Report



# Table of Contents

Executive Summary .....	75
1 Introduction .....	76
1.1 Purpose .....	76
1.2 Project Scope .....	76
1.2.1 Constraints .....	77
1.3 Definitions, Acronyms, and Abbreviations .....	78
1.4 Background .....	79
1.4.1 Motivations .....	79
1.4.2 Similar Studies .....	80
1.5 Aims .....	81
1.6 Technologies .....	81
1.6.1 Services Used .....	81
1.7 Commercialisation .....	81
2 System .....	83
2.1 Functional Requirements .....	83
2.1.1 Requirement 1 <Requirement 1: Outputting Data> .....	84
2.1.2 Requirement 2 < Delete Data> .....	86
2.1.3 Requirement 3 < Report Findings> .....	88
2.1.4 Requirement 4 <Use Flight Evidence> .....	91
2.1.5 Requirement 5 <Use Flight Evidence> .....	94
2.2 Non-Functional Requirements .....	96
2.2.1 Data requirements .....	96
2.2.2 User requirements .....	97
2.2.3 Environmental requirements .....	97
2.2.4 Usability requirements .....	97
2.3 Design and Architecture .....	97
2.4 Implementation .....	99
2.5 Testing .....	99
3 Appendix .....	100
3.1 Project Proposal .....	100

Correlation of Airline Flight delays with external data .....	100
4 Table of Contents .....	101
5 Objectives .....	105
5.1 Lecture's Initial Proposal .....	105
5.2 Proposal .....	105
6 Background .....	107
6.1 Motivations .....	107
6.2 Similar Studies .....	108
7 Technical Approach .....	109
7.1 Development .....	109
7.2 Literature Review .....	109
7.3 Requirements Capture .....	109
7.4 Implementation.....	109
7.5 Project Management.....	110
8 Special Resources Required .....	111
8.1 Software .....	111
8.2 Hardware .....	111
8.3 Documentation .....	111
8.4 Proposed Technologies .....	111
8.5 Services Used .....	111
9 Evaluation .....	112
9.1 Project Plan .....	113
9.2 Monthly Journals.....	114

## **Executive Summary**

In my project I aim to prove the hypothesis that flights can be affected by and even delayed depending on certain weather conditions. In order to complete this, I will take a look at data from the US Department of Transport (USDOT) about flight times and correlate it with data from the National Climate Data Centre (NCDC) on weather in the area to see the average delayed flight depending on the weather conditions themselves. My aim will be to manipulate the data using time and date in a certain area and link this with flights in the same area at the same time, using the planned time of take-off and the actual time of take-off and see how one was affected by the other. I will take a base point in weather, for example, days with clear weather over a certain period of time and compare it to the same data results over a similar period of time, in harsh weather conditions such as rain, wind, hail and even storms. Using the data taken from the USDOT and the NCDC I intend to build a data warehouse, to store the data, to enable the updating of data and also to enable the querying of the data using relevant business intelligence queries in order to better understand the data and use it to suit needs of the hypothesis. I will then create a detailed statistical analysis document which will enable me to report and visualise my findings in a clear and concise manner.

# 1 Introduction

## 1.1 Purpose

The purpose of this document is to set out the requirements for the development of a data warehouse that will aim to gather evidence to hypothesise that weather conditions can affect flights and cause cancelations, delays and have further impacts on flights, their schedules and their patterns. My intentions behind this project is to try and gather enough evidence to show the affects weather can have on flying, at what point does the severity of the weather begin to affect the flights themselves and the times of year that you are most likely to be affected by the weather conditions when traveling. Using the information of the aircrafts themselves I hope to show which aircrafts are most likely to be affected and how. I will then create a document to report my findings and this can hopefully be used to help understand the effects of the weather on flying and this information will hopefully have a practical use for those traveling and how they may be affected. When the report itself is completed the aim is to have a fully functional technical report that will be usable as an open source resource for those needing further evidence of the affects of weather conditions on flights such as government organisations, airline companies, entrepreneurs and passengers.

## 1.2 Project Scope

The project scope is to develop a data warehouse system that will enable me to upload relevant precise data. To guarantee the data is factual and accurate I will take the data for the flights from the US Department of Transportation (USDOT) and the data on weather from National Climate Data Centre (NCDC). When I gather the accurate data, I will then perform the Extract, Transform and load (ETL) process on the data to enable its use in a structured manor for the data warehouse. I will then use this data by loading it into my data warehouse. The data warehouse will have the ability to be updated with newer data, so I can use as up to date relevant data as possible when creating my final analytical report on the findings of my study. I have been in touch with a few lecturers on the creation of my project

from Oisín Creanor, my Business Intelligence and Data Warehousing lecturer to Eugene O'Loughlin my Business Data Analytics, in order to outline the plan for the project, the requirements needed and the steps to take in developing my idea into a functioning data warehouse and the type of analytical report to create from my findings.

### **1.2.1 Constraints**

Choosing a project like this brought forth many constraints in enabling the completion of all of the tasks involved. The first constraint and the most important was the verification of the data itself. When undertaking a project like this, the relevance and accuracy of the data is vital to its completion. In order to guarantee satisfactory data, I turned to governmental organisations for my data, the USDOT and NCDC respectively. Gathering the data and confirming its authenticity is not the only constraint to occur when working with data warehousing and business intelligence, once the data is found it then starts its own constraints also. When using data from multiple data sources it is next to impossible to guarantee the data will be both structured the same and ready for use in the exact way in which it is needed. Building a strong data warehouse and performing the ETL process on both sets of data to create data to be structured similarly enough to use in the correlation process needed to report relevant, accurate and useable findings in the final analytical report.

There are constraints on the development of the project itself. For this project I will be using MySQL for the development of the data warehouse itself. I will then be using such environments as excel and SPSS to create visualisations on my findings and to report them. Other tools used will be the programming languages VBA and R, these will be used to enable me to run scripts and to automate the findings to enable me to report on them easier.

Further constraints such as scheduling restraints would be those restraints put on us from the college itself to meet deadlines. In order to see a detailed look at the plan to keep to the schedule I have included a Gantt chart in the appendix below.

### **1.3 Definitions, Acronyms, and Abbreviations**

<b>Abbreviation</b>	<b>Description</b>
<b>USDOT</b>	<b>US Department of Transport</b>
<b>NCDC</b>	<b>National Climate Data Centre</b>
<b>DW</b>	<b>Data Warehouse: a store of data accumulated from a range of sources within and used to guide management decisions.</b>
<b>R</b>	<b>Open source programming Language</b>
<b>SQL</b>	<b>Structured Query Language</b>
<b>Entity</b>	<b>A component of data in a database</b>
<b>UI</b>	<b>User Interface</b>
<b>Tableau</b>	<b>A UI application used for data visualization.</b>
<b>ETL</b>	<b>Extract, Transform and Load: functions that are combined into one tool to pull data out of a database and place it into another</b>
<b>ERD</b>	<b>Entity Relationship Diagram: shows the relationships of entity sets stored in a database</b>
<b>VBA</b>	<b>Visual Basic for Applications: A Microsoft event-driven programming language</b>
<b>IBM SPSS</b>	<b>SPSS is a software package used for statistical analysis</b>
<b>GUI</b>	<b>Graphical User Interface</b>
<b>UML</b>	<b>Unified Modelling Language: A standardised modelling language used in visualising a system.</b>
<b>Data Dumps</b>	<b>A large amount of data transferred from one system or location to another.</b>
<b>CSV</b>	<b>comma-separated values file stores tabular data in plain text.</b>

## **1.4 Background**

### **1.4.1 Motivations**

When researching ideas for my final year project I was wanting to come up with an idea that encapsulated my specialisation of my 4-year degree. As I am a Business Information Systems student I wanted to develop a project that would be clearly beneficial to myself as a BIS student and that would also highlight my learnings over the last 4 years. When I struggled to identify the right project that would both act as a capstone to my 4 years at NCI and be beneficial to my future after NCI, I approached Oisin Creanor to allow me to use his proposed idea as I felt it was exactly what I was looking for, when trying to encapsulate all my thoughts on my learnings thus far but it was also something that I found very interesting when reading the proposal.

In developing the idea, I will look to implement many different learnings and modules over the previous years such as Databases, Business Data Analytics and Business Intelligence & Data Warehousing, all of which were two-part modules, an introduction and an advanced module for all three. In creating this project, I feel I am implementing a major part of my degree and my learnings to display my understanding of the last 4 years. I believe this project will allow me to both showcase what I have already learned and drive myself into furthering my teachings to a level needed to complete such a task. As I have been contemplating the option of going on to further studies in Data Science, such as a Masters Degree or PHD, Data Analytics and Data Science is something I find very interesting and I believe this project will allow me to delve further into the field and hopefully find a further love for the topic and drive myself into developing my learning through further study.

As I have already completed a module on Business Intelligence & Data Warehousing and currently undertaking the advanced module, I will use my

knowledge obtained from this implement the Extract, Transform and Load (ETL) aspect of the project. As I am currently undertaking my first module on Data Analytics and will participate in an advanced module on the topic in semester two of my final academic year my aim is to use the skills obtained to analyse the information gained from the both the Department of Transportation and NCDC, after the ETL process to create an in-depth analytical report that will outline my findings in a clear and concise manner.

I am hugely looking forward to creating this report as it will give me the opportunity to showcase my skill set as previously discussed but also give me the opportunity to advance my knowledge in the field and give me the chance to work with and learn methodologies the will be used in the field when undertaking employment after my degree.

#### **1.4.2 Similar Studies**

After my initial research into the field and similar studies carried out in the past on the topic of weather conditions and the impact on flights such as delays has provided me with a lot of information regarding the subject. In researching these studies, it will enable me to get a broader understanding of the topic itself before creating my own statistical analysis report. Although there are similar studies done, I have yet come across one that will aim to prove what I am setting out to prove using the same techniques and methods that I am.

## **1.5 Aims**

In the report my aim is to analyse and build a statistical model to record the impact weather has on flights and their time delays. I aim to visualise this information, by taking accurate data from the US Department of Transportation, which provides detailed departure information and other information on flights throughout the U.S.A. By cross referencing this data with external data from the National Climate Data Centre (NCDC), I aim to create a Data model designed to run queries which will enable me to fact check time and dates of flights, along with time and date of weather in a specific area to prove my hypothesis. Furthermore, using this information I also aim to discover which weather conditions affect flights the most, using the same method and statistical data. I also aim to discover which seasons and/or months of the year it is best to fly in to minimise the risk of flight delay and which are the worst. In this study I also intend to find out which aircrafts are affected mostly by the weather and if size/shape of aircraft has any impact on how the weather can affect the flights themselves. I will then create a visualisation tool which will help to query the data in a user-friendly user interface.

## **1.6 Technologies**

- MySQL
- Excel Functions
- VBA
- IBM SPSS Statistics
- Microsoft Excel
- Tableau
- R Data Science Language

### **1.6.1 Services Used**

- The US Department of Transportation
- NCDC archive of global historical weather

## **1.7 Commercialisation**

The aim of this project is to perform a strongly analytical report to support the theory that flights can be affected and delayed depending on climate and weather.

In the completion of my report I would hope that my findings could be used to further help those in need of evidence in such topics. This can be entrepreneurs in the business of flights/airline or with the idea of developing an app that can use the information within the report. Others can include governmental organisations that such as the Department of Transport and the NCDC to enable better results on their own studies. This can also be used by statistical analysts and even media outlets such as newspapers to help prove a story or they are reporting on. As big data and data mining has become so important in the world today, using studies such as my own analytical report that I will produce at the end of my project, major aircraft companies can look to my report to help make business and managerial decisions that can benefit the company and its bottom line, creating a monetizing and commercial aspect to the study for others. For example, if certain aircrafts are being affected by weather conditions more frequently than others and investment in those aircrafts affected less by weather can both save cost on repairs and also aim to gain revenue through less delayed and cancelled flights.

## **2 System**

### ***2.1 Functional Requirements***

The following functional Requirements are in ranked order depending on importance:

1. Sourcing of accurate data
2. Perform ETL process on data
3. Build data warehouse to store and query data
4. Perform relevant queries to return reportable data
5. Visualise and report the findings
6. Use finding to prove/disprove the hypothesis
7. Use UPI to update data

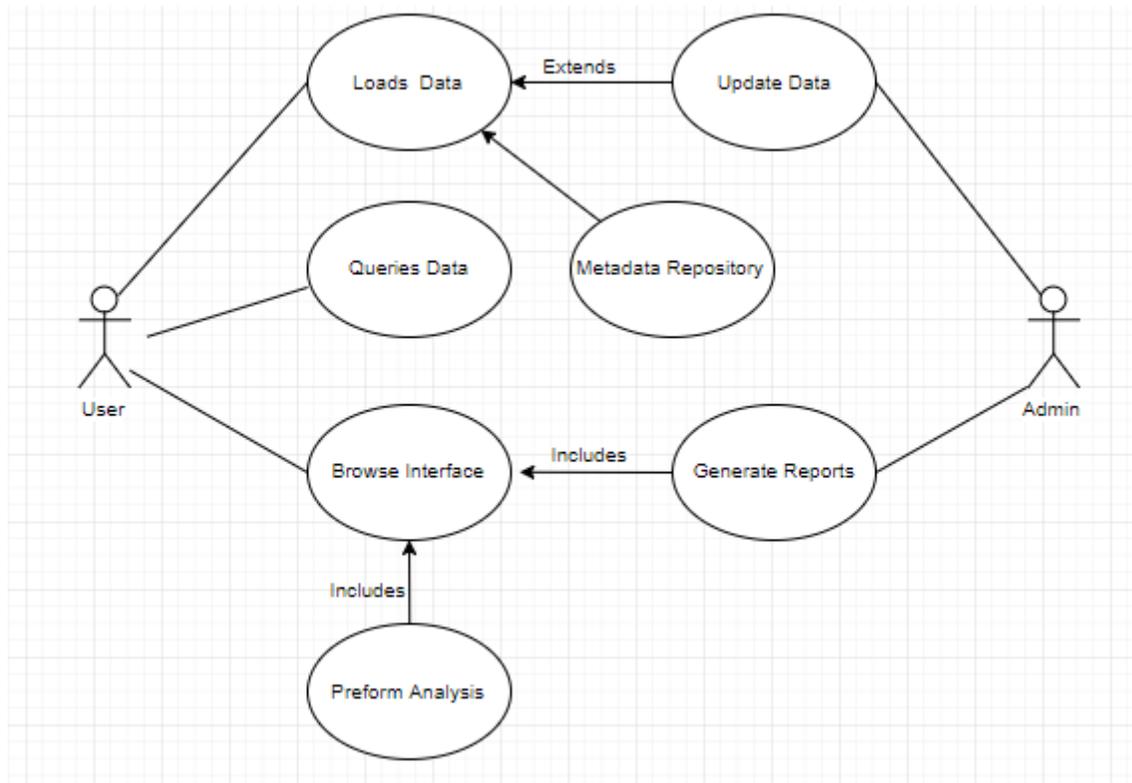
The above list is ranked on importance not in order of how and when to undertake the tasks. Some of the tasks may depend entirely on a task later in the list however may still be of more relevance and importance than its predecessor.

## 2.1.1 Requirement 1 <Outputting Data>

### 2.1.1.1 Description & Priority

The ability to output data is the fundamental requirement of the project. Without the ability to output the data I will not be able to report the findings.

### 2.1.1.2 Use Case Diagram



### 2.1.1.3 Use Case

Output Data from data warehouse.

#### Scope

The scope of this use case is to enable the user to output data using queries on the data warehouse.

#### Description

This use case describes the how the user would query the database to return relevant information.

## **Actor**

Administrator

## **Flow Description:**

### **Precondition**

The system is in initialisation mode, linked to the data warehouse and hosted, ready for use.

### **Activation**

This use case starts when the user opens the interface.

### **Main flow**

1. The system identifies the user authentication details (E1).
2. The user opens the application for the interface of the DW.
3. The user accesses the DW data.
4. The user queries the information using MySQL commands.
5. The system returns data outputs (A1).
6. The user downloads relevant data.

### **Alternate flow**

A1 : <Altered Queries>

1. The user returns to the queries interface.
2. The user re-enters desired queries to generate more valid results.
3. The system returns new data outputs.
4. The user downloads relevant data.

### **Exceptional flow**

E1 : <Access Denied>

1. The system denies access to the user
2. The user runs through security checks to verify authentication
3. The user refreshes access rights and/or password etc.
4. The user reattempts log-in.

### **Termination**

The user logs out of the system.

### **Post condition**

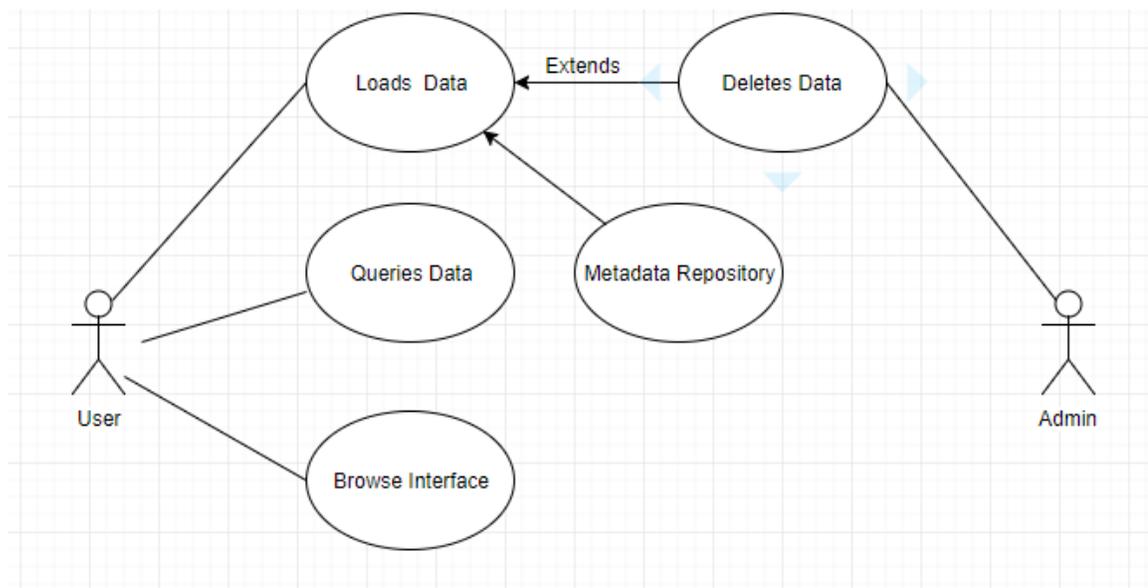
The system goes into a wait state

## 2.1.2 Requirement 2 < Delete Data >

### 2.1.2.1 Description & Priority

The ability to delete data stored in the DW.

### 2.1.2.2 Use Case Diagram



### 2.1.2.3 Use Case

Delete data from data warehouse.

#### Scope

The scope of this use case is to enable the user to delete already saved data in the data warehouse.

#### Description

This use case describes the how the user would delete data stored in the DW.

#### Actor

Administrator

#### Flow Description:

#### Precondition

The system is in initialisation mode, linked to the data warehouse and hosted, ready for use.

### **Activation**

This use case starts when the user opens the interface and access the current data in the DW.

### **Main flow**

1. The system identifies the user authentication details (E1).
2. The user opens the application for the interface of the DW.
3. The user accesses the DW data.
4. The user writes syntax code to remove certain data in the DW.
5. The system prompts a confirmation of the data wanting to delete (A1).
6. The system deletes chosen data.

### **Alternate flow**

A1 : <Altered Queries>

1. The system informs the user that there is currently an active user accessing data wanting to be removed.
2. The system prompts a choice for user to delete anyway.
3. The system deletes data from DW.

### **Exceptional flow**

E1 : <Access Denied>

1. See E1 Access denied on requirements 1: Outputting Data

### **Termination**

The user logs out of the system.

### **Post condition**

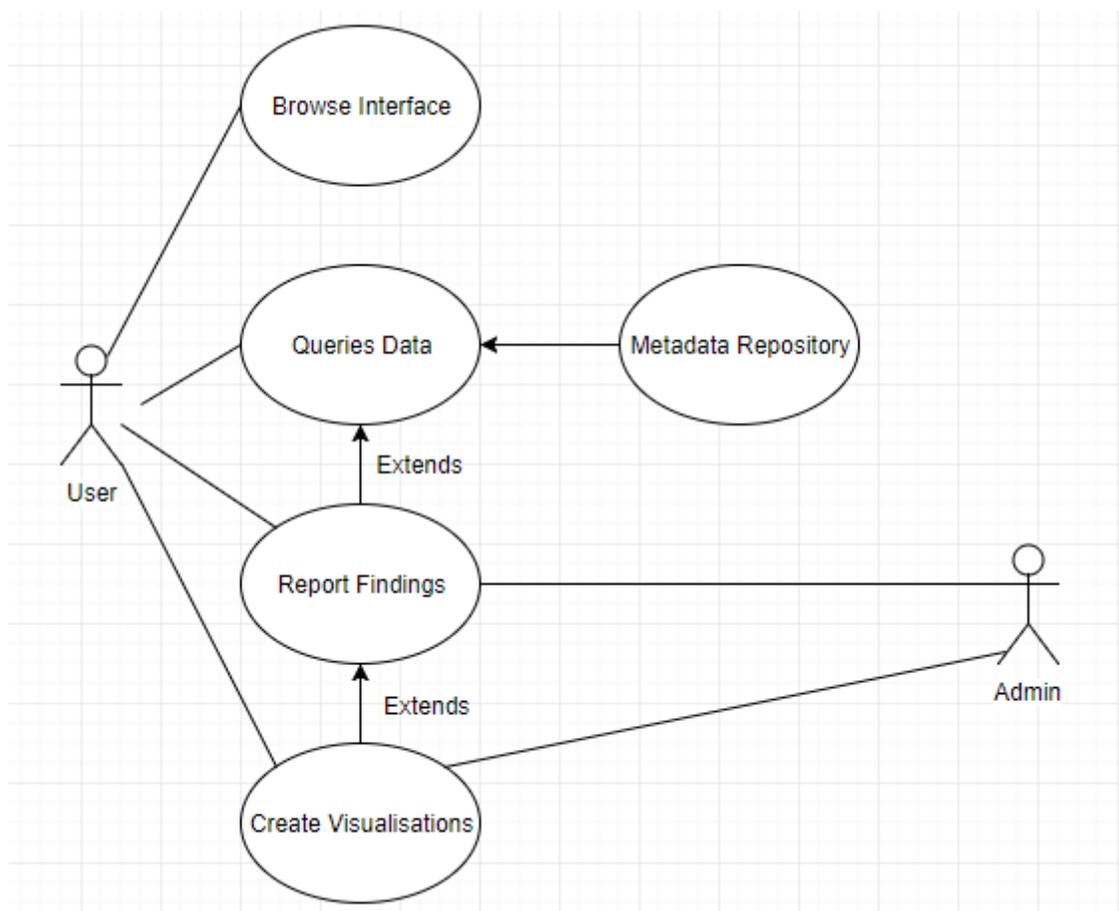
The system goes into a wait state

## 2.1.3 Requirement 3 < Report Findings>

### 2.1.3.1 Description & Priority

This requirement deals with taking of the data queried in the DW and reporting on the findings.

### 2.1.3.2 Use Case Diagram



### 2.1.3.3 Use Case

Report on data from data warehouse.

#### Scope

The scope of this use case is to enable the user to report and draw visualisations from the data acquired from the DW.

### **Description**

This use case describes the how the user would use and manipulate the data to report on their findings.

### **Actor**

Administrator

### **Flow Description:**

### **Precondition**

The system is in initialisation mode, linked to the data warehouse and hosted, ready for use.

### **Activation**

This use case starts when the user opens the interface and access the current data in the DW.

### **Main flow**

1. The system identifies the user authentication details (E1).
2. The user opens the application for the interface of the DW.
3. The user accesses the DW data.
4. The user writes syntax code to remove certain data in the DW.
5. The user takes data from DW and formats it in an external file such as Excel.
6. The user draws visualisations such as graphs to show the data in a more user-friendly manner (A1).
7. The user reports on findings.

### **Alternate flow**

A1 : <Excel functions>

1. The user writes excel function codes to manipulate the data into further metadata.
2. The user uses data for further reporting.

### **Exceptional flow**

E1 : <Access Denied>

1. See E1 Access denied on requirements 1: Outputting Data

**Termination**

The user logs out of the system.

**Post condition**

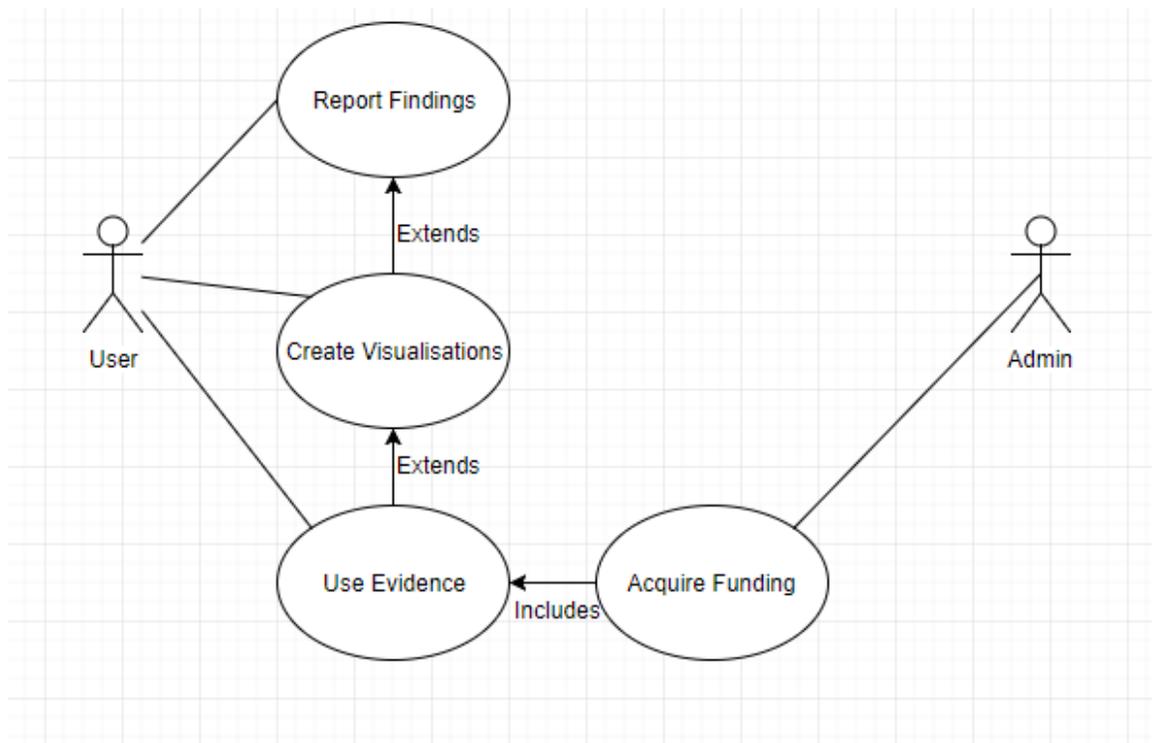
The system goes into a wait state

## 2.1.4 Requirement 4 <Use Flight Evidence>

### 2.1.4.1 Description & Priority

This requirement deals with taking the findings of the report and using it as evidence in a particular topic, in this case an entrepreneur looking for funding for a new flight delay app.

### 2.1.4.2 Use Case Diagram



### 2.1.4.3 Use Case

Report on data from data warehouse.

#### Scope

The scope of this use case is for an entrepreneur to use relevant data needed for convincing evidence in acquiring funding for its new app he would like to develop.

### **Description**

This use case describes the how the user would use and manipulate the data to present it as evidence.

### **Actor**

User

### **Flow Description:**

### **Precondition**

The Report has been delivered and is accessible as open source information.

### **Activation**

This use case starts when the user acquires the report for use.

### **Main flow**

1. The user opens online report.
2. The user searches for information relevant to his app.
3. The user creates own report using evidence from finalized analytical report.
4. The user reports his findings to funder.
5. The user acquires funding (A1).
6. The user Sources the original report.

### **Alternate flow**

A1 : <Denied Funding>

1. The user gets denied funding.
2. The user uses data for further reporting to back up his initial evidence.
3. The user again tries for funding (See main flow) (E1).

### **Exceptional flow**

E1 : <Failed effort>

1. The user is again refused funding and decides to give up.

### **Termination**

The user closes out of report.

**Post condition**

The report is available for access again by the user and others.

## 2.1.5 Requirement 5 <Use Flight Evidence>

### 2.1.5.1 Description & Priority

Level1 Being Critical as the whole project depends on being able to store the datasets within a database, without this the project cannot take place

### 2.1.5.2 Use Case

In this use case it will show how to create and upload data to the DW

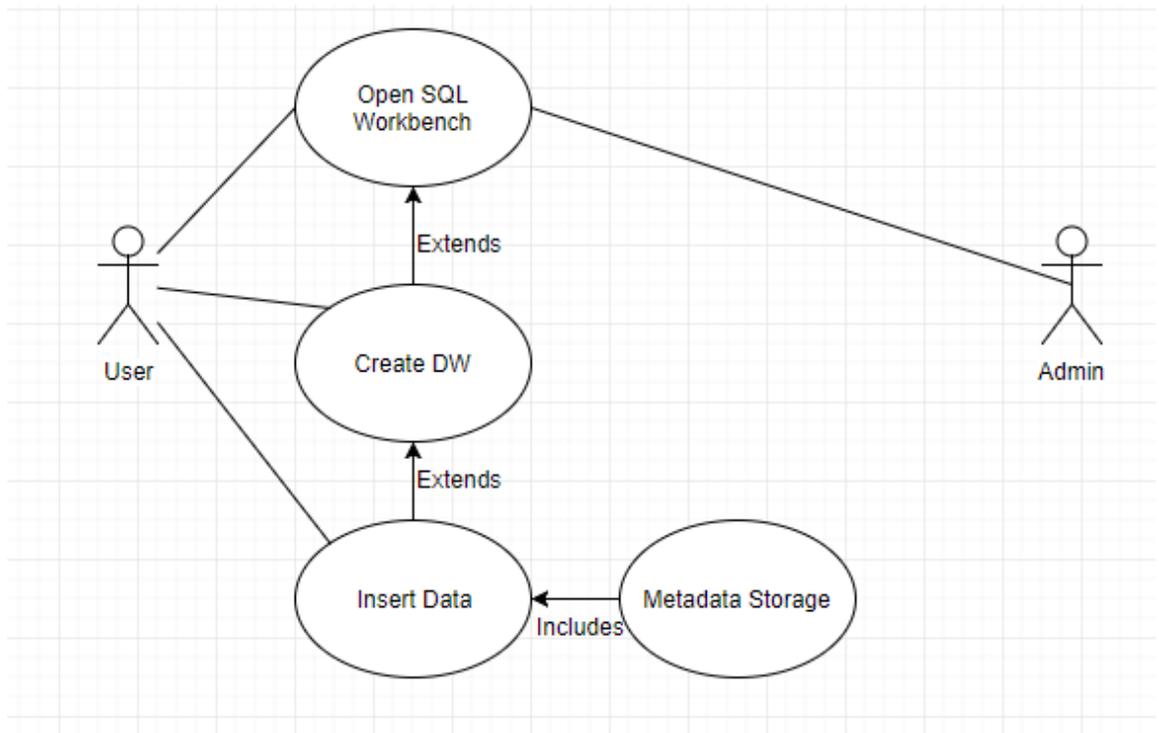
#### Scope

The scope of this use case is to use MySQL create and populate the DW

#### Description

This use case describes the operation of creation of a database and then how to load it with the data.

#### Use Case Diagram



#### Flow Description

#### Precondition

The database that is created must be able to be accessed by user.

#### Activation

This use case starts when the user opens SQL workbench

**Main flow**

1. The user opens the DW application.
2. The user creates the Database.
3. The user opens MS Excel(A1).
4. The user imports the datasets using the SQL functions (E1).
5. The user then closes the DW application.

**Alternate flow**

A1 : <SPSS Import>

1. The user opens IBM's SPSS.
2. The user imports data from IBM SPSS into DW using SWL functions

**Exceptional flow**

E1 : <SQL Function Error>

1. The system prompts error warning.
2. The user checks the error and adjusts functions accordingly.

**Termination**

The user logs out.

**Post condition**

The data is stored within the DW.

## 2.2 Non-Functional Requirements

### 2.2.1 Data requirements

The data requirements for the project will be the most accurate data that I can find on the topic to enable me to create as accurate report as possible. These data I will retrieve from government organisations. I regards the flight data I will use the USDOT official data to show that the data is verified and correct. See below for example of flight data details:

Key	Airport	Place	Country	IATA	ICAO	Latitude	Longitude	Altitude	Time-Z	DST	TZ data	Type	Source
7	Narsarsua	Narssarsss	Greenland	UAK	BGBW	61.1605	-45.426	112	-3	E	America/	airport	OurAirports
8	Godthaab	Godthaab	Greenland	GOH	BGGH	64.1909	-51.6781	283	-3	E	America/	airport	OurAirports
9	Kangerlussuaq	Sondrestrom	Greenland	SFJ	BGSF	67.01222	-50.7116	165	-3	E	America/	airport	OurAirports
10	Thule Air Base	Thule	Greenland	THU	BGTL	76.5312	-68.7032	251	-4	E	America/	airport	OurAirports
21	Sault Ste Marie	Sault Ste Marie	Canada	YAM	CYAM	46.485	-84.5094	630	-5	A	America/	airport	OurAirports
22	Winnipeg	Winnipeg	Canada	YAV	CYAV	50.0564	-97.0325	760	-6	A	America/	airport	OurAirports
23	Halifax	Halifax	Canada	YAW	CYAW	44.6397	-63.4994	144	-4	A	America/	airport	OurAirports
24	St. Anthony	St. Anthony	Canada	YAY	CYAY	51.3919	-56.0831	108	-3.5	A	America/	airport	OurAirports
25	Tofino	Tofino	Canada	YAZ	CYAZ	49.07983	-125.776	80	-8	A	America/	airport	OurAirports
26	Kugaaruk	Pelly Bay	Canada	YBB	CYBB	68.5344	-89.8081	56	-7	A	America/	airport	OurAirports
27	Baie Comore	Baie Comore	Canada	YBC	CYBC	49.1325	-68.2044	71	-5	A	America/	airport	OurAirports
28	CFB Bagotville	Bagotville	Canada	YBG	CYBG	48.3306	-70.9964	522	-5	A	America/	airport	OurAirports
29	Baker Lake	Baker Lake	Canada	YBK	CYBK	64.2989	-96.0778	59	-6	A	America/	airport	OurAirports
30	Campbell	Campbell	Canada	YBL	CYBL	49.9508	-125.271	346	-8	A	America/	airport	OurAirports
31	Brandon	Brandon	Canada	YBR	CYBR	49.91	-99.9519	1343	-6	A	America/	airport	OurAirports
32	Cambridge	Cambridge	Canada	YCB	CYCB	69.1081	-105.138	90	-7	A	America/	airport	OurAirports
33	Nanaimo	Nanaimo	Canada	YCD	CYCD	49.05497	-123.87	92	-8	A	America/	airport	OurAirports
34	Castlegar	Castlegar	Canada	YCG	CYCG	49.2964	-117.632	1624	-8	A	America/	airport	OurAirports
35	Miramichi	Chatham	Canada	YCH	CYCH	47.0078	-65.4492	108	-4	A	America/	airport	OurAirports
36	Charlo	Charlo	Canada	YCI	CYCI	47.9908	-66.3303	132	-4	A	America/	airport	OurAirports

The next dataset needed would be the weather data set from the NCDC. This is the data set that I will use to correlate the data to make my findings as genuine as possible:

STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	MLY-TMIN	MLY-TMAX	MLY-PRCP-NORMAL
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201001	-43	145	55
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201002	4	199	44
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201003	140	318	72
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201004	294	517	98
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201005	417	656	249
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201006	521	743	376
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201007	564	793	322
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201008	542	789	275
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201009	438	685	208
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201010	310	531	169
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201011	165	337	81
GHCND:USC00327027	PETERSBURG 2 N ND US	466.3	48.0355	-98.01	201012	15	187	73

### **2.2.2 User requirements**

ETL is the process of extraction of data from source systems, transforming it into viable data and loading it into the DW. The extraction process involves taking data from the source and placing it into the DW. The transformation of data from the source format into the format needed for the DW is the transform and the load involve the loading of the data into the target DW.

The requirements required by the user will be the ability to load and update data in the DW. When building the DW warehouse one further requirement would be an user-friendly interface to enable easy querying. The DW will need to output data from the entered queries to be used for statistical analysis. This can be done through exporting to Excel, SPSS or Tableau for example.

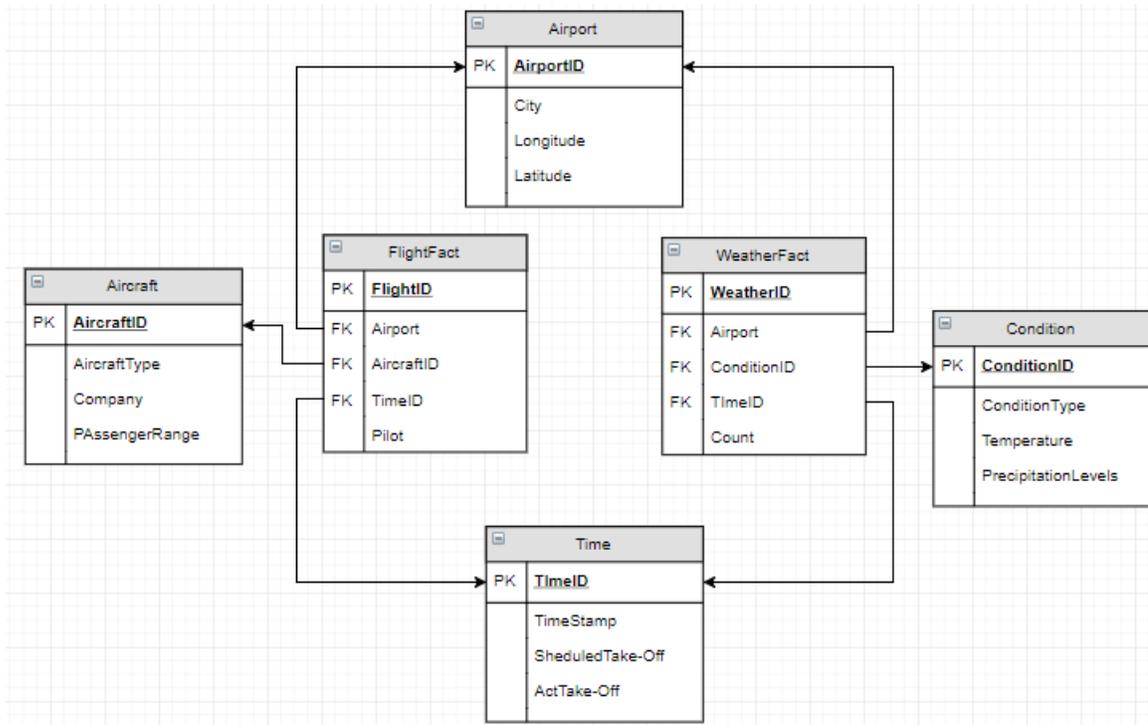
### **2.2.3 Environmental requirements**

The environmental requirements needed for the project will include but are not limited to SQL Workbench for the creation of the DW. For the reporting I will also be using applications such as Tableau, Excel and SPSS for reporting and visualisation respectively.

### **2.2.4 Usability requirements**

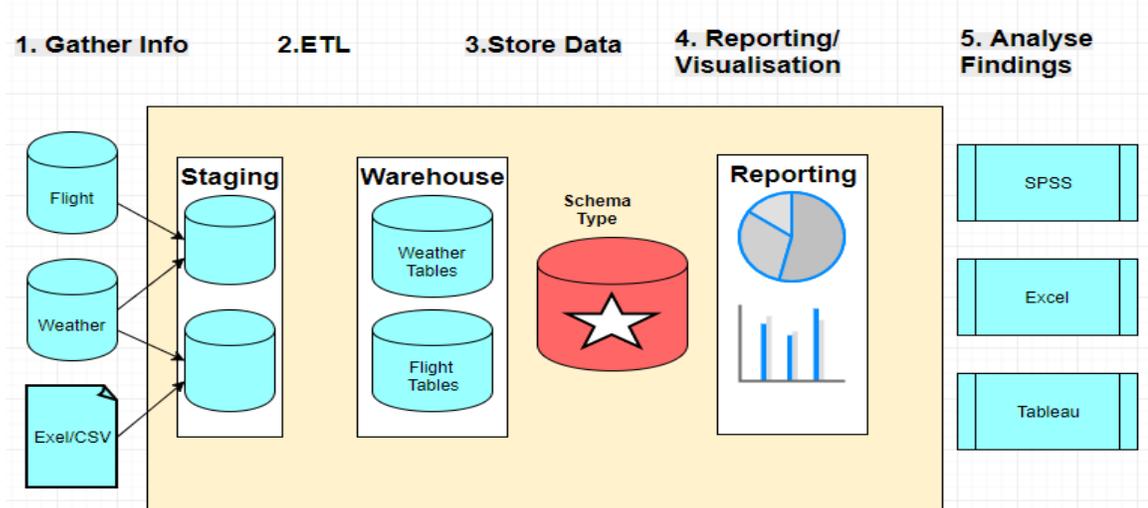
## ***2.3 Design and Architecture***

When building any DW it is vitally important to plan out using diagrams your intended workload. Below is an example Star Schema ERD of what the DW could look like using multiple fact tables, in this case the weather table and the flight table. From these fact tables we have multiple dimensional tables to help both structure the data itself and to link to for further querying. The creation of these diagrams is vitally important in laying out your plans going forward and to formulate the beginning of the ETL process.



The ERD is a preliminary mock-up of how the DW may be structured and is a live diagram that can be changed throughout the process to reflect the DW.

When creating a data warehouse, planning is a major part of the workload. When undertaking the task of creating a DW the mindset of “Fail to prepare, prepare to fail” is a great approach to live by. Before creating a DW, creating the system architect diagram will help show the outlining plan of what is to come.



The system architecture diagram above outlines the process from gathering the requirements and all the information of the project to the analysing and reporting of the data. Outlined above it shows the type of schema being used in the DW, which is a star schema, the platforms in which we will report and analyse the data and the types of data marts needed in the DW itself.

## ***2.4 Implementation***

When implementing the project, it was important to have a DW that was updatable as more information was gathered. Implementing the right schemas to get the best use of the DW will be equally important as the speed of recovery of information and the relevance of the data received is both time and cost effective. In the case of this project a star schema was selected due to its flexibility with data and data marts within a DW.

## ***2.5 Testing***

When undertaking a task of this nature, testing is a major aspect throughout the process of creating and implementing the idea. In the case of testing an agile/iterate process will be taken to test the project, the data and the DW itself at all relevant stages throughout the process. This will be important for both time and cost on a project of this kind. Testing at stages throughout the project reduces the possibility of mistakes being made and only discovered later on in the development, creating a need to go back to an earlier stage and wasting valuable time and money. Testing at an iterative level will create stages to the development cycle and make testing smaller tasks and easier to fix any errors found throughout.

## **3 Appendix**

### **3.1 Project Proposal**



National  
College of  
Ireland

*The college for a  
learning society*

# Project Proposal

## **Correlation of Airline Flight delays with external data**

Ian Donnelly, x14111659, x14111659@student.ncirl.ie

BSc (Hons) in Business Information Systems

15th November 2017

## 4 Table of Contents

Introduction .....	6
Project Background .....	6
Aims and motivations .....	6
Definitions, Acronyms, and Abbreviations .....	8
Technologies .....	10
MySQL .....	10
MS Excel.....	10
IBM SPSS .....	11
R/R Studio .....	11
Tableau .....	11
Methodology.....	12
Literary Review .....	13
Introduction .....	13
Literature Reviews .....	14
Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs.....	14
Further Investigations into the Causes of Flight Delays.....	15
Analysis of Delay Causality at Newark International Airport .....	16
Microsoft Excel 2016 Data Analysis and Business Modelling: Winter’s Method .....	16
Conclusion.....	17
Bibliography .....	18
Research Question .....	19
Research Title and Question .....	19
Hypothesis .....	19
Normalisation .....	19
Differentiation .....	20
Correlation .....	20
Time Series Analysis .....	21
Methods.....	22

Introduction .....	22
ETL Process .....	22
Data Warehouse.....	23
Datasets Metadata.....	23
Creating the Data Warehouse.....	26
MS Excel.....	30
IBM SPSS .....	32
R/RStudio .....	33
Tableau .....	33
Results .....	34
Kolmogorov-Smirnov .....	37
Mann-Whitney U.....	38
Snow <b>38</b>	
Wind <b>40</b>	
Temperature.....	42
Rain <b>44</b>	
Pearson’s Correlation Coefficient .....	46
Snow <b>46</b>	
Wind <b>47</b>	
Temperature.....	48
Rain <b>49</b>	
Holt Winter’s Exponential Smoothing.....	50
Tableau .....	55
Snow <b>55</b>	
Wind <b>58</b>	
Temperature.....	60
Rain <b>62</b>	
Testing.....	65
Datawarehouse.....	65
Correlation .....	66
Forecasting.....	67

Future Opportunities .....	69
Further Analysis .....	69
Conclusion .....	71
Appendix .....	72
Executive Summary .....	75
1 Introduction .....	76
1.1 Purpose .....	76
1.2 Project Scope .....	76
1.2.1 Constraints .....	77
1.3 Definitions, Acronyms, and Abbreviations .....	78
1.4 Background .....	79
1.4.1 Motivations .....	79
1.4.2 Similar Studies .....	80
1.5 Aims .....	81
1.6 Technologies .....	81
1.6.1 Services Used .....	81
1.7 Commercialisation .....	81
2 System .....	83
2.1 Functional Requirements .....	83
2.1.1 Requirement 1 <Outputting Data> .....	84
2.1.2 Requirement 2 < Delete Data> .....	86
2.1.3 Requirement 3 < Report Findings> .....	88
2.1.4 Requirement 4 <Use Flight Evidence> .....	91
2.1.5 Requirement 5 <Use Flight Evidence> .....	94
2.2 Non-Functional Requirements .....	96
2.2.1 Data requirements .....	96
2.2.2 User requirements .....	97
2.2.3 Environmental requirements .....	97
2.2.4 Usability requirements .....	97
2.3 Design and Architecture .....	97
2.4 Implementation .....	99

2.5	Testing.....	99
3	Appendix.....	100
3.1	Project Proposal.....	100
4	Table of Contents.....	101
5	Objectives.....	105
5.1	Lecture's Initial Proposal.....	105
5.2	Proposal.....	105
6	Background.....	107
6.1	Motivations.....	107
6.2	Similar Studies.....	108
7	Technical Approach.....	109
7.1	Development.....	109
7.2	Literature Review.....	109
7.3	Requirements Capture.....	109
7.4	Implementation.....	109
7.5	Project Management.....	110
8	Special Resources Required.....	111
8.1	Software.....	111
8.2	Hardware.....	111
8.3	Documentation.....	111
8.4	Proposed Technologies.....	111
8.5	Services Used.....	111
9	Evaluation.....	112
9.1	Project Plan.....	113
9.2	Monthly Journals.....	114

## 5 Objectives

### 5.1 *Lecture's Initial Proposal*

**Areas of Interest:** Data Warehousing, ETL, structured data analytics

**Proposer:** Oisín Creanor

**Title:** Correlation of Airline Flight delays with external data

**Proposal:** The US Department of Transportation provides detailed departure and other statistics on airline flights in the USA. By combining the data from this source with data from external sources such as weather data by airport/region and public calendars, this project will attempt to identify the dominant cause of airline delays and cancellations, and measure the probable impact of various weather conditions on the on-time status of airline flights in the USA.

### 5.2 *Proposal*

My proposal is to conduct an in-depth statistical analysis on how the weather conditions can impact flight times. When researching for the project I will aim to prove whether the weather conditions play a significant part in what causes a delay on flights in the USA. In doing so I will look at existing data and existing studies as well as other theories in the field in order to identify any existing trends or patterns while also trying to create my own, to prove or disprove my hypothesis. In undertaking this project, I will implement many various aspects of my previous and current learnings at NCI, in the Business Information Systems stream.

In the report my aim is to analyse and build a statistical model to record the impact weather has on flights and their time delays. I aim to visualise this information, by taking accurate data from the US Department of Transportation, which provides detailed departure information and other information on flights throughout the

U.S.A. By cross referencing this data with external data from the National Climate Data Centre (NCDC), I aim to create a Data model designed to run queries which will enable me to fact check time and dates of flights, along with time and date of weather in a specific area to prove my hypothesis. Furthermore, using this information I also aim to discover which weather conditions affect flights the most, using the same method and statistical data. I also aim to discover which seasons and/or months of the year it is best to fly in to minimise the risk of flight delay and which are the worst. I will then create a visualisation tool which will help to query the data in a user-friendly user interface.

## **6 Background**

### **6.1 Motivations**

When researching ideas for my final year project I was wanting to come up with an idea that encapsulated my specialisation of my 4-year degree. As I am a Business Information Systems student I wanted to develop a project that would be clearly beneficial to myself as a BIS student and that would also highlight my learnings over the last 4 years. When I struggled to identify the right project that would both act as a capstone to my 4 years at NCI and be beneficial to my future after NCI, I approached Oisín Creanor to allow me to use his proposed idea as I felt it was exactly what I was looking for, when trying to encapsulate all my thoughts on my learnings thus far but it was also something that I found very interesting when reading the proposal.

In developing the idea, I will look to implement many different learnings and modules over the previous years such as Databases, Business Data Analytics and Business Intelligence & Data Warehousing, all of which were two-part modules, an introduction and an advanced module for all three. In creating this project, I feel I am implementing a major part of my degree and my learnings to display my understanding of the last 4 years. I believe this project will allow me to both showcase what I have already learned and drive myself into furthering my teachings to a level needed to complete such a task. As I have been contemplating the option of going on to further studies in Data Science, such as a Masters Degree or PHD, Data Analytics and Data Science is something I find very interesting and I believe this project will allow me to delve further into the field and hopefully find a further love for the topic and drive myself into developing my learning through further study.

As I have already completed a module on Business Intelligence & Data Warehousing and currently undertaking the advanced module, I will use my

knowledge obtained from this implement the Extract, Transform and Load (ETL) aspect of the project. As I am currently undertaking my first module on Data Analytics and will participate in an advanced module on the topic in semester two of my final academic year my aim is to use the skills obtained to analyse the information gained from the both the Department of Transportation and NCDC, after the ETL process to create an in-depth analytical report that will outline my findings in a clear and concise manner.

I am hugely looking forward to creating this report as it will give me the opportunity to showcase my skill set as previously discussed but also give me the opportunity to advance my knowledge in the field and give me the chance to work with and learn methodologies the will be used in the field when undertaking employment after my degree.

## **6.2 *Similar Studies***

After my initial research into the field and similar studies carried out in the past on the topic of weather conditions and the impact on flights such as delays has provided me with a lot of information regarding the subject. In researching these studies, it will enable me to get a broader understanding of the topic itself before creating my own statistical analysis report. Although there are similar studies done, I have yet come across one that will aim to prove what I am setting out to prove using the same techniques and methods that I am.

## **7 Technical Approach**

### ***7.1 Development***

For the development of my project I will implement many different tools. The data warehouse itself for example will be developed using MySQL. I aim to use such visualisation tools as, Excel, Tableau, and SPSS. In developing the script, I aim to implement both the programming language R and I also aim to use VBA within Excel to automate some of my queries and create more user-friendly ways to display my findings.

### ***7.2 Literature Review***

Considering the vast amount of research that has been done on this topic in the past I intend to use these resources throughout the project to both compare and analyse my findings in regard to previous studies done in the field. But also, to give me a foundation to begin on while also aiming to prove my own hypothesis using my own tools and in my own way.

### ***7.3 Requirements Capture***

The main aspect for the requirements capture will be to determine the authenticity of the data being retrieved. The data downloads will be retrieved from national/governmental agencies and will have negligible risk as far as its reliability. Then I will need to use the proper ETL tools to enable the data to be used and manipulated to suit the process needed in creating the statistical report.

### ***7.4 Implementation***

In the implementation of the project I will first need to identify the information being used and its relevance to the project. After confirming its authenticity, the ETL

process will begin. After the implementation of the data warehouse itself, the next step will be identifying the correct queries to run to remove the most relevant information. After this I will then use my analytical methods to create the report itself and to visualise the findings.

### ***7.5 Project Management***

The best approach I find while managing this project is an agile approach. For many reason some being the ability to reiterate through phases and adapt the project as I go. But because this is a project that will require the continuous upload of new and relevant data and considering the flexibility of the agile approach, I believe it will be the best methodology to use.

## **8 Special Resources Required**

### **8.1 Software**

The software I will need to use will mostly come from the NCI Data Analytics Suite. I will also use other software programs such as Tableau, Excel and SPSS.

### **8.2 Hardware**

No specific hardware needed for this project. The use of any laptop/PC will be sufficient.

### **8.3 Documentation**

The documentation used for this project will be my project proposal and requirements specifications report. I will however also be conducting an in depth statistical analysis report based on my findings and I will research and follow a suitable structure to that of similar reports in the field.

### **8.4 Proposed Technologies**

- MySQL
- Excel Functions
- VBA
- IBM SPSS Statistics
- Microsoft Excel
- Tableau
- R Data Science Language

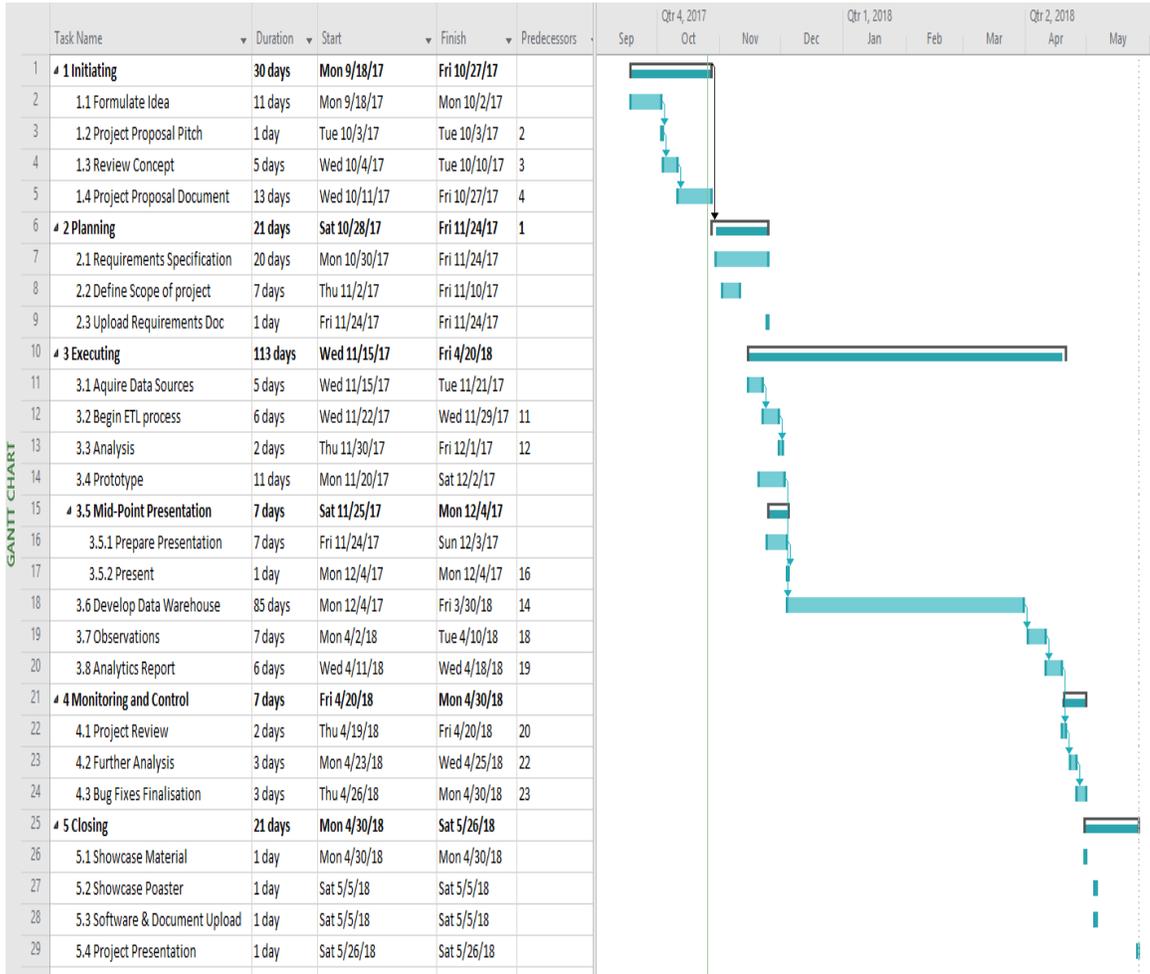
### **8.5 Services Used**

- The US Department of Transportation
- NCDC archive of global historical weather

## **9 Evaluation**

Just like in any project undertaken, testing is an unbelievably important part of the process. Without testing an incomplete, faulty or inaccurate project can be completed. Therefore, at various stages throughout the project I will be testing and re-testing my work and my findings. In completion of all the tests and of the project as a whole. I will then undergo thorough evaluation of my results and findings and document them in a presentable manner and to allow for analysis and critiquing of the project and its findings.

## 9.1 Project Plan



## **9.2 Monthly Journals**

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: September

Achievements: Peace of mind.

At the beginning of the semester I was quite overwhelmed with the daunting task of facing the final year project itself. Such a massive project and here I am struggling to even find an idea for something to do. Most of this month I spent seeking out lecturers help with coming up with an idea. Being in BIS I found it a difficult task to know what was asked of me and the project itself. I felt like the project classes were aimed towards those in the software related streams and the BIS stream was not getting enough information needed to help us with even just the expectations of the project itself. This was not helping the overwhelming feeling. The added anxiety had come from the knowledge that I did have the best part of 3 years to come up with an idea for the project and here I was at the beginning of year 4 with nothing in mind. The project module itself allows for cross over and I could have tried to undertake a project that was more relevant in a software stream however as my capstone project I wanted to create a project that would reflect my 4 years in college and that would show my strengths when out in the industry looking for my first graduate role.

Then came the project proposal pitch, this added to extra anxiety as I went to the pitch with no idea and looking for some inspiration. From talking to the lecturers at the pitch I was helped to formulate an idea or what at least was expected of myself and the BIS stream. I came out of the pitch, although with no idea as of yet, less anxious as I felt I was able to grasp what was expected of me.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: October

Achievements: Formulating an idea.

Still struggling with an idea for the project itself I turned to lecturers in the college again for help. I approached Ron Elliot for advice as he would have lots of experience and information regarding the BIS stream. I approached Eugene O'Loughlin, my data analytics lecturer for help in formulating an idea that would encapsulate my data analytics module as this is a module I am very interested in. I then turned to Oisín Creanor my business intelligence and Data Warehousing lecturer as the further I advanced through the project ideas concept, the more I realised it would be a DW and Business intelligence report that would highlight my achievements and learnings in college most. When speaking to Oisín, I was lucky enough to find that he had suggested some ideas for the list of project ideas by faculty that students could pitch for and hopefully develop. After approaching Oisín over his idea about flights and the impact weather may have on them, I was happy to undertake this task as it seemed very interesting and Oisín gave me permission to carry out his idea as my project.

The rest of October was spent formulating my project proposal to get an understanding of what would be required of me in this project.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: November

Achievements: Technical Report

After finally formulating an idea and working on the research report, November was spent gathering requirements and making a plan of action for the Mid-point technical report due at the end of the month. A lot of time went in to both the planning of the document and then the implementation of it. For the month of November not a lot of time was spent doing much else as the technical report document ended up nearly 7000 words and took up so much time. With the writing of the findings in the requirement spec, to the building of use cases and the formulating of the project to come there was a lot to cover in such a short time. Throughout the work of the technical document time was spent to create the prototype needed also.

In the month of November, I also reached out to my Temporary Supervisor Christian Rusu for a meeting. This was not an ideal situation as for most of the semester I did not know my Temporary supervisor and it put a little added uncertainty to the project itself. After meeting with Christian I was able to get some feedback on my work to date and manipulate it to better match his view on the project and prepare it for upload.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: December

Achievements: Presentation and End of Semester

The mid-point presentation was still to come. This involved the building of the slide set and then the finalising of the mid-point prototype for which, I created an Excel document with a sample of the data to be used in the final report (3-month period). Through this I was able to run formulas and create pivot tables to link both datasets and create graphs showing the link between weather and delayed flights over the period of time. The presentation itself came around and it was an experience. Usually very strong on presentations I walked out for once unsure on how it went. The extreme grilling of questions at the end had me in doubt, although I had the knowledge to stand behind my idea and answer any question related to the project, it could not help but put doubt in my mind however all in all I believe I performed very well and was happy with what I could do.

The rest of December was spent finalising CAs due for other modules and with the amount of time and concentration spent on the final project in the months previous it was nice to take my mind off it temporarily although still aimed elsewhere and well and truly kept busy, it was not what I would call downtime or time of relief.

My December ended uniquely with contact made with lecturers and supervisors in regard to my mid-point result, this was the first time in my college life I was left unsatisfied with my situation and went seeking advice on how to express my concerns. With my supervisor absent for personal reasons the meetings conducted with other lecturers helped with the gaining knowledge of the process of resolving the issue. After receiving my project idea from the list of faculty approved ideas and having the validity of the entire idea questioned by my temporary supervisor so late on really did knock me and my confidence in carrying on.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: January

Achievements: Exams and New Semester

After the Christmas period like every year it was hard to get back into the swing of things, even more so this year with uncertainty about my project and midpoint issues still unresolved. But you are left little time to sit around as the exams come thick and fast. The first half of January just like the end of December was no time to concentrate on the project unfortunately as I was so preoccupied with exam time and studying it left no time or brain capacity for anything else. As the exams came and went, one by one I felt happy with my performances but just like at the end of every other exam period, doubt sets in and worry takes over and you count down the days till results time to finally get some relief and confirmation of your efforts and how they pay off.

The second half of January was beginning of the new semester, learning new module descriptors meeting new lecturers and overall preparation for the months to come. Being the beginning of the final semester, I felt like the end was almost near and the work and effort put in over the last 4 years was soon going to be worth it.

Meeting with Eamon and Anu at the end of January about the issues with the midpoint project deliverables helped to inform them regarding the situation, but as of the end of January there had still not been a resolution.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: February

Achievements: Project Progress

As January ended and February began, concentration went back almost entirely on the project. There were exams and CAs to do along side but nothing like the semester end rush and the January exam times. February was spent with the downloading of the data and the construction of the DW. This was very time consuming as the data had to be downloaded month by month from the USDOT and over a 10-year period took a lot of hours to do. Then The formatting of the data into one excel document took time too. All this time spent on just preparation of the data before the real work on the project could begin made me very happy that I spent the time in February to do it as it was not until I went to download the data in bulk I realised how time consuming it was.

When this was done the coding of the DW came next, all this while researching further literature on the topic, researching the best tests to carry out. February was an extremely productive month in the overall scheme of things. During the coding of the data warehouse the data needed to be formatted and February was spent with a lot of time on the ETL process to load and complete the DW for later use. This all done before the real analysis was to be carried out on the data, not actually fully part of the project so to speak, but so vitally important is the preparation in the implementation of a product/study.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: March

Achievements: CAs and planning

March was a month I really wanted to make the most of, with a two-week reading week I knew March was a time in which I could get a lot done and out of the way. This was true for projects and CAs in other modules even more so than that of the final statistical project. While conversing with Dr O'Loughlin throughout my ABDA module in regard to the best statistical analysis tests to run on the data to find the most relevant information and the most accurate results, I began to create a planned structure of what the statistical side of the project (the main course) was going to take.

The rest of the month was spent completing CAs for other modules and starting/structuring and researching CAs that were not even due for weeks/months to come. This was part of my plan and enabled me to eliminate the worry and concern for other deadlines from other modules throughout the rest of the semester while allowing me the ability to concentrate my efforts on the report in the months leading up to May's upload.

With the showcase profile sign off at the start of April I also used the time in March to organise to get my picture taken for the showcase book.

Student Name: Ian Donnelly

Program: BSc in Business Information Systems

Month: April and May

Achievements: Showcase profile sign off, exams and project development.

April being the penultimate month as anticipated it was very busy when it came to the work to be done on the project. In anticipation I managed to schedule the month off work to concentrate my whole effort when not in college on the project. At the beginning of the month I met in person with Dr O'Loughlin in his office to get further guidance on the statistical analysis tests, the structure of them and any advice I could from him. I really appreciated the help of Dr O'Loughlin over both semesters of 4<sup>th</sup> year as he helped grow my passion for data analysis.

The ending of the semester came early in semester two, which meant so did CA deadlines and lucky for me I had prepared well for this and felt little pressure from the other modules through a greatly implemented process of action. This however did not ring through for the exams which were scheduled at the end of April. With study and pressure from the exams, the project unfortunately took a back seat once again and with so little time left for completion I am glad I had the plan in place from February to relieve myself of the end of semester pressure. The exams came and went, and I was happy with them. It was an unusual feeling after the exams as usually the relief of its over, is such a peaceful moment, however with the 4<sup>th</sup> year project still hanging over my head that was not the feeling of relief it usually is.

With less than 3 weeks to upload the term the business end really rang true. The creation of excel sheets, the running of statistics and writing of R code all began to make the days feel like they were rolling into one VERY long day!!! The results to be visualised in Tableau at times it definitely felt like there was too much to do and not enough time to do it. My schedule at times felt like I had not given myself enough time, but in truth it was just anxiety and stress of the emerging deadline. After conducting all the analysis and having all my findings, the relief did

not begin there. The reporting of it still lay ahead. The document itself can only be compared to a thesis in time effort and length taken to complete. With help and effort from My supervisor I was able to format the document, create the poster and implement any last-minute details to bring the project to a close before time of upload. Now comes the preparation for the Presentation and showcase which means it still doesn't end here.....