National College of Ireland

BSc in Business Information Systems

2017/2018

Robert Kane

X14110831

X14110831@Student.ncirl.ie

A Statistical Study on the Impact of Weather on Crime

Technical Report

Contents

Declaration for Project Submission

Name: Robert Kane
Student ID: x14110831
Supervisor: Lisa Murphy

Confirmation of Authorship

I confirm that I have read the College statement on plagiarism and that the work I have submitted for assessment is entirely my own work.

Signature:

Date: 10/ 5/ 2018

1. Abstract

The purpose of this research is to study the effect weather has on incidents of crime. The research analyses archival weather records and quantitative incidents of crime for the period 2008 through 2017 for the city of Chicago. The study was conducted for a final year project with the aim of identifying trends and patterns in the data in order to investigate if future forecasts could be made.

The methods implemented to conduct this research include the development of a data warehouse for the modelling of historical data, a statistical analysis of reports produced by the data warehouse using data analysis software, and visualisations of the results and findings of the study through Tableau for end user interpretation. A literature review was carried out to examine existing academic hypothesis and scientific findings on the relationship of weather conditions and criminality. This was to ensure the relevancy and academic integrity of the results and forecasts presented in this study.

The studies investigated in the literature review show evidence of a strong positive correlation between weather and rates of crime. These findings were used as a guideline for the development of a statistical model to demonstrate and visualise the effect temperature and weather conditions have on crime. As the relationship is influenced by and dependent on seasonality, an analysis of seasonal time series data using Holt-Winters exponential smoothing methods was implemented to construct a model to present future forecasts. These forecasts could be of significant benefit to law enforcement agencies and emergency response services to assist in resource allocation and preventative strategies.

2. Introduction

2.1. Project Background

This project aims to conduct an in-depth statistical analysis on the impact temperature and weather conditions have on incidents of crime. The purpose of this study is to examine existing academic hypothesis and scientific findings in this area, as well as seeking to identify new links, trends and patterns of my own to form a conclusion and make future forecasts. I propose to achieve this by developing a data warhouse and implementing the advanced methods of data analysis I have studied throughout my final year as a Business Information Systems student at NCI.

For this analysis, I propose to build a statistical model to demonstrate and visualise the effect temperature and weather conditions have on various categories of crime. For me to achieve this and present meaningful, accurate results, the data I use must be accurate in order to provide the base for all transformations and subsequent aggregate queries. For this reason the study will be focused on the American city of Chicago. The Chicago Data Portal has made available to the public, statistics recorded on crime via The Chicago Police Department's open data API. Data obtained through this portal will be mapped to historical weather data obtained from the NCDC archive of global historical weather.

2.2. Motivations for the Development of this Project

As a capstone to my studies as a Business Information Systems student at NCI, I want to produce a substantial project that integrates and synthesizes all the various aspects of my learning and acquired expertise in this field. For this reason I have chosen to undertake a project I feel encapsulates and showcases the disciplines and advanced understanding of modules such as Databases, Business Intelligence & Data Warehousing and Business Data Analysis. The area of data analytics is one that I am extremely interested in and have a keen desire to pursue as a career upon graduation from NCI. I feel this project will allow me the opportunity to pursue and showcase relevant independent academic research in this field of study, as well as utilising the resources available to me at NCI, such as the expertise and guidance of lecturers at NCI in these subject areas.

As I have studied modules such as Advanced Data Analytics in my final year, I feel it is important for me to be produce a technical project that incorporates these analytical skills and allows me to present them in an academic body of research. This project will be a valuable asset for me to be able to show potential employers. The project also gives me an opportunity to learn and develop new skills such as R Studio, SPSS and Tableau. When initialising the project, these are technologies I would have had no prior experience using. Demonstrating the ability to learn and develop new skills is a major motivation and desire to showcase upon the completion of the project.

The most important aspect of this project for me is not the concept, but rather the opportunity to showcase a high level of proficiency in the technologies and methodologies used to develop it. The architectural planning and implementation of a project such as this will allow me to work with technologies that are essential in the field of business intelligence and data analytics. It will also allow me to demonstrate a high level of abstract thinking and an ability to integrate advanced analytical methodologies for business solutions.

Business Information Systems students sit in the middle of technology and business solutions; therefore, it is important for me to demonstrate not just the ability to work with technology proficiently, but to also implement that technology to effectively communicate meaningful analysis, insights and solutions.

3. Literature Review

3.1. Introduction

A comprehensive literature review must be conducted for the successful completion of a project of this nature. A literary review of existing academic hypothesis and scientific findings in this area provided an essential overview of current studies and methods relevant to this research. It is also necessary to provide context and understanding of the research topic by evaluating similar existing works.

My exploration into the relationship of weather and criminality brought numerous existing studies on the subject to my attention. I studied these works continuously throughout the lifecycle of the project as part of the literature review. This allowed me to formulate a unique approach to my analysis both in terms of relevancy and identifying a variance in different methods that have not yet been explored. The literature review will be presented in chronological order of studying and the author is credited using Harvard Style referencing.

A literature review was also necessary to identify the required approaches of data analysis and the correct methodology of implementing them for this project. The literature review, along with advice and guidance from Dr. Eugene O'Loughlin highlighted key areas of great interest and significance for the completion of this project. Consulting with Dr. Eugene O'Loughlin about the project, he suggested I implement a method of seasonal time series data using Holt-Winters exponential smoothing methods.

As this method of time series analysis was not covered on the Advanced Business Data Analysis syllabus, I was provided with appropriate literature and material to study on his recommendation in order to implement the forecasting models used in this project

3.2. Literature Reviews

An Analysis of Relationship Between Weather and Crime in Cleveland, Ohio by Paul Butke and Scott C. Sheridan (2010)

This study focused on the relationship between weather and crime for the period from 1999 through 2004 in the city of Cleveland, Ohio. The author primarily focused the study on the summer months of June – August because these are the months when the most oppressive conditions occur. The results of the research reflected my own preliminary observations of the data used in this research – aggressive crime generally increases linearly as ambient average temperature increases. The research carried out analysis at the city ground level to investigate the spatial patterns of aggressive crime when it is hot compared to when it is cold. Despite the numerous different spatial analysis performed, the results demonstrated a minimal correlation between spatial patterns of crime counts and hotter weather conditions. Rather, the results demonstrated that increases in temperature resulted in similar percentage increases in crime citywide. This was particularly important to discover as my analysis could then focus on ambient temperature and weather conditions alone without spatial patterns as an intrinsic dimension. The aim of the paper was to examine thermal comfort with the focus on the level of oppressiveness of summer weather. Researching this paper was both interesting and beneficial to me as the study brought four main theories on the relationship between weather and crime to my attention. The Negative Affect Escape Model was of particular interest in explaining observations found in my own data. The Negative Affect Escape Model demonstrates that negative affects such as feelings of annoyance and discomfort and crime rate increase as temperature increases up to a certain inflection point. At this point the model predicts a decrease in crime due to a person's escape motives overriding their aggressive motives. This literature review provided an explanation to this inflection point observed in my data analysis. The inflection point can generally be observed at around 30C° in this study, this was something I was eager to have an explanation for. The research supports the hypothesis that hotter temperatures account for higher amounts of crime and the association between crime and weather can be similarly incorporated to assist in determining where and when intervention is most useful. This is beneficial to my study as I intend to develop a forecasting model for predictions.

Effects of Weather on Crime by Rodrigo Murataya and Daniel R. Gutierrez (2013)

This paper was particularly beneficial to my literature review as it touched on a number of different studies conducted on the subject matter of weather and crime. The research paper analysed data from multiple research studies from various areas around the world. According to this research, weather and crime have been found to have a significant correlation. The paper goes on to analyse numerous works that show evidence of a positive relationship between weather and crime. This was a significant finding because the correlation does not seem to be affected by geographical location or population factors – the results and outcomes are fairly similar and are equally observed in many different locations. The literature review section of this paper consolidated a lot of academic research on the subject that gave me an extensive insight into the existing research and scientific findings in this area. The paper was beneficial in assuring me the observations being discovered in my data analysis and visualisations are accurate and conform to similar scientific modelling. The paper focused on the comparative variables of weather conditions and crime and states that researching the relationship between the two will provide law enforcement a better understanding of how the weather affects crime. This research will allow authorities to better prepare their departments during weather conditions that influence certain criminal activity. This was a significant point as it demonstrates that the aim of my project to develop a predictive forecasting model can be achieved and implemented. The study concludes with recommendations to further enhance the accuracy of studies by updating the data that has already been obtained. Updating my obtained data to measure actual records against forecasted figures will verify and enhance the accuracy of my own report.

Prediction of Crime Occurrence from Multi-Model Data Using Deep Learning by Kyeon-Woo and Hang Bong Kang (2017)

This study was of relevant interest to my literature review as the focus of the research was on the prediction of crime occurrence. The research proposed an accurate crime occurrence prediction method by combining multi-model data from multiple domains with environmental context information. The method incorporates past criminal activity records and analyses them based on deep learning to make future forecasts. The ideas and findings are of general interest but present quite a few limitations acknowledged within the study. There is heavy consideration to environmental context information based on the findings of a deep neural network (DNN). As previously discussed, spatial and environmental factors have little influence on crime rates in correlation to ambient temperature and weather conditions. The methodology for the prediction model gathers and considers data based on urban factors as well as spatial and temporal patterns, this results in a lack of crime occurrence report data at all sampling points and therefore results in the prediction model being unbalanced. The limitations of the study are that it is not possible to apply a DNN-based crime occurrence prediction method to regions of insufficient data. The study was interesting and highlighted research in the area of crime prediction and prevention but was heavily focused on environmental data and spatial mapping as the components of analysis.

Chapter 63 (Winters's Method) Microsoft Excel 2016 Data Analysis and Business Modelling by Wayne L. Winston (2016), p645-649

Dr. Eugene O'Loughlin recommended this literary resource to me as a result of a consultation with him for academic guidance and advice concerning aspects of my project. As I demonstrated data visualisations displaying a seasonal correlation in the relationship of crime and weather over a ten-year period in my Tableau workbook, he recommended a particular method of time series analysis that could be implemented in my project for forecasting a time series that exhibits seasonal behaviour. He highlighted a chapter on the analysis of seasonal time series data using Holt-Winters exponential smoothing methods. As this method of time series analysis was not covered on the Advanced Business Data Analysis module, this resource proved extremely helpful in assisting me to implement the forecasting model in this project. Dr. Eugene O'Loughlin recommended I review the literature around this method and learn how to implement it for my project.

This chapter describes in detail methods of predicting future values of a time series. This is usually difficult as the characteristics of a time series are constantly changing. Smoothing and adaptive measures are required for accurate future forecasting and this chapter describes in detail the most powerful smoothing method: Winters's method. The chapter uses an example case to predict housing starts in the United States and details a step-by-step tutorial on the methods of implementing a Holt-Winters exponential smoothing forecast. This literature review was extremely beneficial in assisting my personal learning requirements to achieve complex aspects of this project.

Time Series Forecasting Using Holt-Winters Exponential Smoothing by Prajakta S. Kalekar (2004)

A further literature review of the Holt-Winters exponential smoothing method was required for me to gain a deeper understanding and knowledge of the complex components involved in accurately implementing the method. This paper explained in great detail the components that make up the method, what they mean and how they are attained. The paper provides an in-depth description and explanation of the forecasting equations as well as the initial values to be used for the parameters. This paper was of particular interest as it provided information on the multiplicative model for time series exhibiting multiplicative seasonality. This was vital in assisting me to correctly select the model with the potential to produce the best forecasting results.

3.3. Conclusion

The purpose of conducting this literature review was to investigate and examine existing academic hypothesis and scientific findings on the relationship of weather conditions and criminality. This was to ensure the relevancy and academic integrity of the results and forecasts presented in this study as well as providing me with an overview of the research methodology. The literature review provided me with the assurance that the aims of the project were viable and could be successfully achieved if the correct methodology was followed. The nature of the literature review was comprised of two principle topics – a review of similar studies in this area of research and a literature review to determine suitable methods of data analysis and forecasting. An extensive understanding of these topics was required for this project and a great deal of beneficial information has been obtained from the literature review.

Additional resources were included in conducting the literature review for this project, but as the reviewing of these titles led to the discovery of them being unsuitable, they have not been documented. I have limited the literature review to the resources that directly assisted me in understanding the relevant research area and provided me with an understanding of the methods of data analysis and forecasting to implement in this project.

The key points to take from the literature review are that there seems to be an academic consensus on the positive relationship between crime and weather. Although the various research methods reviewed often incorporate or heavily focus on aspects of environmental spatial factors, psychology or biological behavioural science, the one common theme found in these studies is that crime increases proportionately to ambient temperature. Another key point to take from the literature review is that this relationship is observed across the world regardless of culture, population density or environmental context. This allows me to determine the correlation in the data for this project with confidence and apply it to a complex forecasting model.

3.4. Bibliography

1. Butke, P. and Sheridan, S. (2010). *An Analysis of the Relationship between Weather and Aggressive Crimein Cleveland, Ohio*. [online] researchgate.com. Available at: https://www.researchgate.net/publication/249622050_An_Analysis_of_the_Relationship_between_Weather_and_Aggressive_Crime_in_Cleveland_Ohio [Accessed 20 Nov. 2017].

2. Kalekar, P. (2004). *Time series Forecasting using Holt-Winters Exponential Smoothing*. [online] labs.omniti.com. Available at: https://labs.omniti.com/people/jesus/papers/holtwinters.pdf [Accessed 7 Apr. 2017].

3. Kang, H. and Kang, H. (2017). *Prediction of crime occurrence from multi-modal data using deep learning*. [online] journals.plos.org/plosone. Available at: http://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0176244 [Accessed 8 Feb. 2018].

4. Murataya, R. and Gutiérrez, D. (2013). *Effects of Weather on Crime*. [online] ijhssnet.com. Available at: http://www.ijhssnet.com/journals/Vol_3_No_10_Special_Issue_May_2013/7.pdf [Accessed 27 Jan. 2018].

5. Winston, W. (2016). *Microsoft Excel 2016 data analysis and business modeling*. 1st ed. Microsoft Press, pp.645 - 649.

4. Research Questions and Objectives

4.1. Research Question

The purpose of this study is to conduct an investigation into how weather conditions contribute to the occurrence of crime. As previously discussed in the literature review, a significant number of academic studies examining the relationship between weather and crime have already been conducted but there has been limited effort of pursuing the relationship to any great depth to develop a model to forecast the occurrence of crimes.

This research paper therefore poses two questions:

1. Do weather conditions have a statistically significant effect on the rate at which crime occurs?
2. Can analysis and plotting of this relationship support the development of an accurate forecasting model?

To satisfy these questions the required data necessary to perform a statistical analysis has been gathered from the official sources and structured accordingly. A ten-year overview of the data will be examined to identify any apparent trends and patterns as well as a year-by-year break down to investigate the annual cyclical relationship of the variables. From there, the analysis can then be performed and the findings discussed.

4.2. Hypotheses Overview

In order to test the data and answer the research questions set out above, a set of hypotheses must be declared to make an inference about the findings once the tests have been conducted. The test for correlation will involve stating a null and alternative hypothesis to investigate if a significant statistical correlation can be observed at the chosen significance level. As there is no test statistic for the time series analysis test, a null and alternative hypothesis will not be required. The forecast performance and accuracy will be measured against actual statistics gathered after forecasts have been made and the mean average percentage error will be calculated to determine the accuracy and dependability of the model.

4.3. Hypothesis Testing – Correlation

Is there a statistically significant correlation between weather conditions and crime?

H_0: $\rho = 0$: There is no correlation between weather and crime
H_A: $\rho \neq 0$: There is a correlation between weather and crime

Now that the research questions and hypotheses have been outlined, the following sections of the document will detail the methods implemented to answer them.

5. Method & Implementation

5.1. Introduction

The methodology used to implement the various aspects of this project and answer the research questions will be broken down into individual components and explained. The components used for this research include gathering the required data, the development of a data warehouse, the analysis of the data using the appropriate data analysis software packages, and the development of a Tableau visualisations dashboard for the presentation of results for end user interpretation.

All statistical analysis will be performed and validated in data analysis software packages such as R Studio, IBM SPSS Statistics and Microsoft Excel.

5.2. ETL Processes & Data Requirements

As part of the requirements gathering process, I identified the sources and suitability of the data required for this project. Once I determined the usability of the data, I initiated the process of familiarising myself with the format in order to begin the ETL process for this project.

ETL (Extract, Transform and Load) is a process in data warehousing responsible for the extraction of data from source systems and placing it into the data warehouse. ETL can fundamentally be viewed as the intersection of computer science, management of information systems and data science and as such, a good design of these processes in the early stages of a data warehouse project is essential.

The ETL processes for this project involved downloading the required data and structuring and transforming it in an appropriate way for loading the data warehouse and performing the statistical analysis on.

5.2.1. Data Sources

The following data sources provided the required data sets for the completion of this project. These data sources allow access to accurate historical records of historical crime and weather statistics.

The Chicago Police Department's Open API

Since 2001, The Chicago Data Portal has made available to the public, statistics recorded on crime via The Chicago Police Department's open data API. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

NCDC Archive of Global Historical Weather

Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information. This data includes quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree-days as well as radar data and 30-year climate records.

5.3. Data Warehouse Development

Once the required data had been gathered and the necessary ETL processes had been carried out to cleanse and structure the data, the development of the data warehouse could take place. The intended use for the development of the data warehouse is to identify and generate meaningful queries on the crime and weather data. The output generated from these queries will then be transported and implemented to conduct the statistical analysis in the various data analysis software packages this project utilises. The data warehouse was developed using MySQL Workbench.

5.3.1. Data Warehouse Star Schema

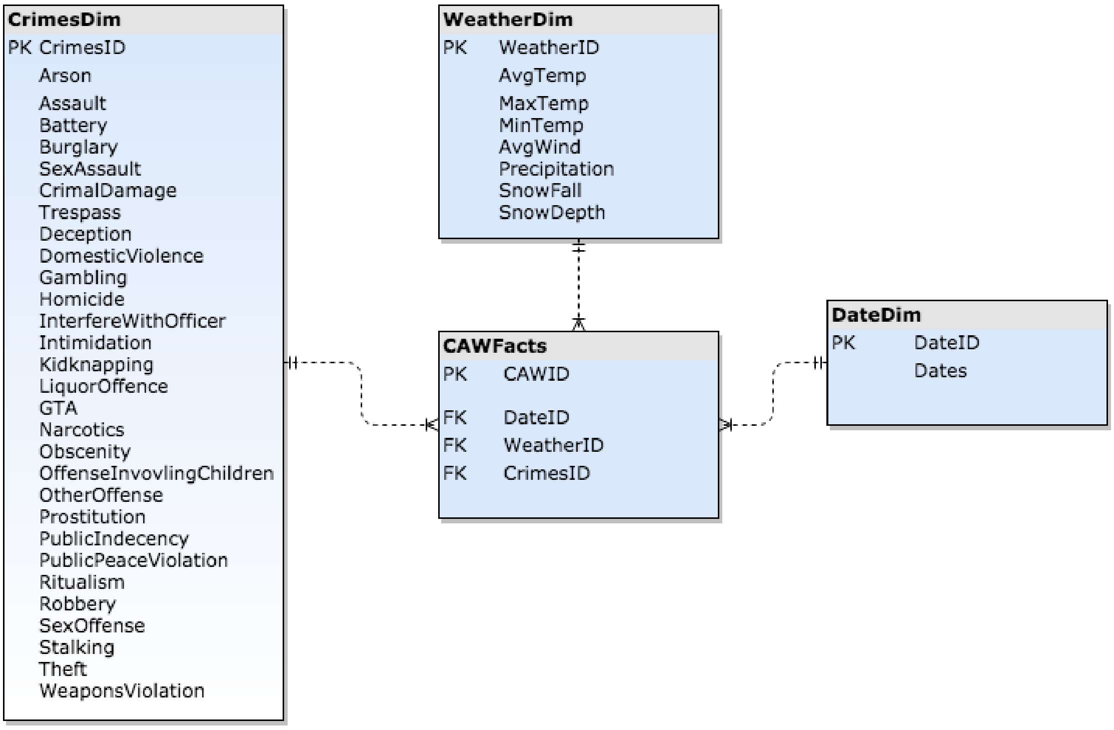
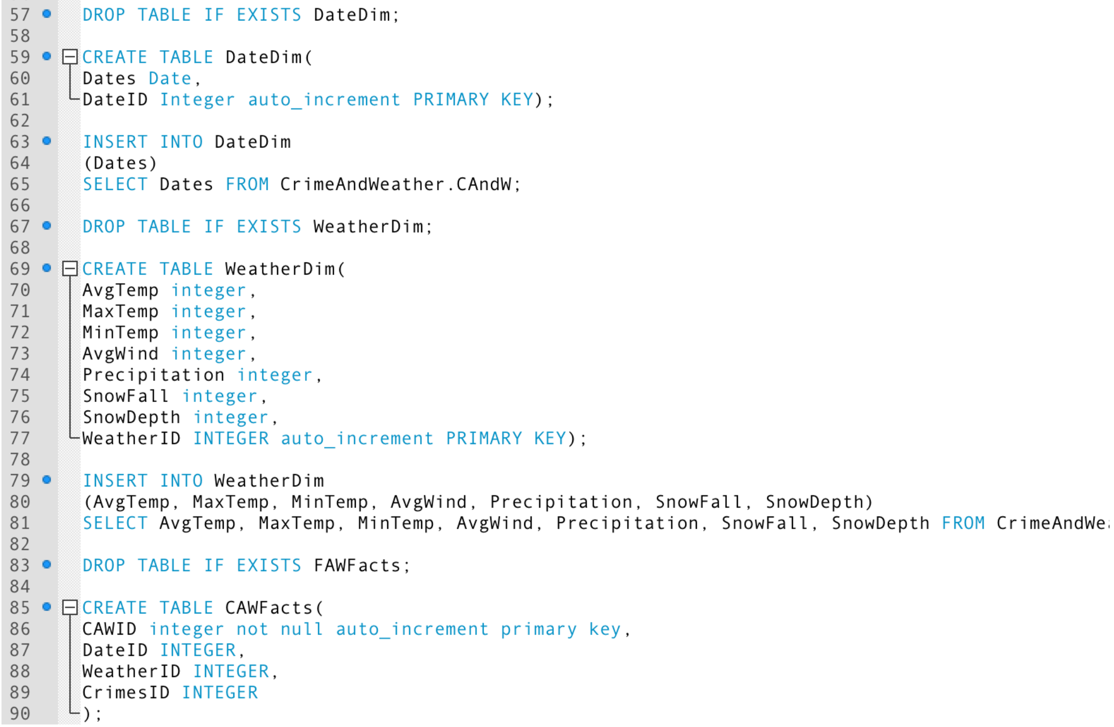


Fig 1: Data warehouse star schema.

Fig 1 demonstrates the star schema used to develop the architecture of the data warehouse.



```
DROP TABLE IF EXISTS DateDim;

CREATE TABLE DateDim(
Dates Date,
DateID Integer auto_increment PRIMARY KEY);

INSERT INTO DateDim
(Dates)
SELECT Dates FROM CrimeAndWeather.CAndW;

DROP TABLE IF EXISTS WeatherDim;

CREATE TABLE WeatherDim(
AvgTemp integer,
MaxTemp integer,
MinTemp integer,
AvgWind integer,
Precipitation integer,
SnowFall integer,
SnowDepth integer,
WeatherID INTEGER auto_increment PRIMARY KEY);

INSERT INTO WeatherDim
(AvgTemp, MaxTemp, MinTemp, AvgWind, Precipitation, SnowFall, SnowDepth)
SELECT AvgTemp, MaxTemp, MinTemp, AvgWind, Precipitation, SnowFall, SnowDepth FROM CrimeAndWe

DROP TABLE IF EXISTS FAWFacts;

CREATE TABLE CAWFacts(
CAWID integer not null auto_increment primary key,
DateID INTEGER,
WeatherID INTEGER,
CrimesID INTEGER
);
```

Fig 2: Data warehouse SQL code

Fig 2 shows the development of the data warehouse.

5.4. Tableau

Data visualisations are inherently a critical element in communicating the results and findings of a project of this nature. Tableau Desktop visualisation suite therefore played an extremely important role in achieving the goals of this project. Tableau was a critical element for the completion of the project not only for presenting the results and findings of the research to the end user for interpretation, but also to perform analysis and generate preliminary visuals in order to identify any apparent trends and patterns emerging.

Visually mapping and assessing the observations found in the data through Tableau allowed for the identification of seasonality. This in turn led to the Holt-Winters method being identified as an ideal forecasting model. Fig 3 demonstrates an overview of weather and crime trends for Chicago for the years 2008 through to 2017 generated in Tableau.

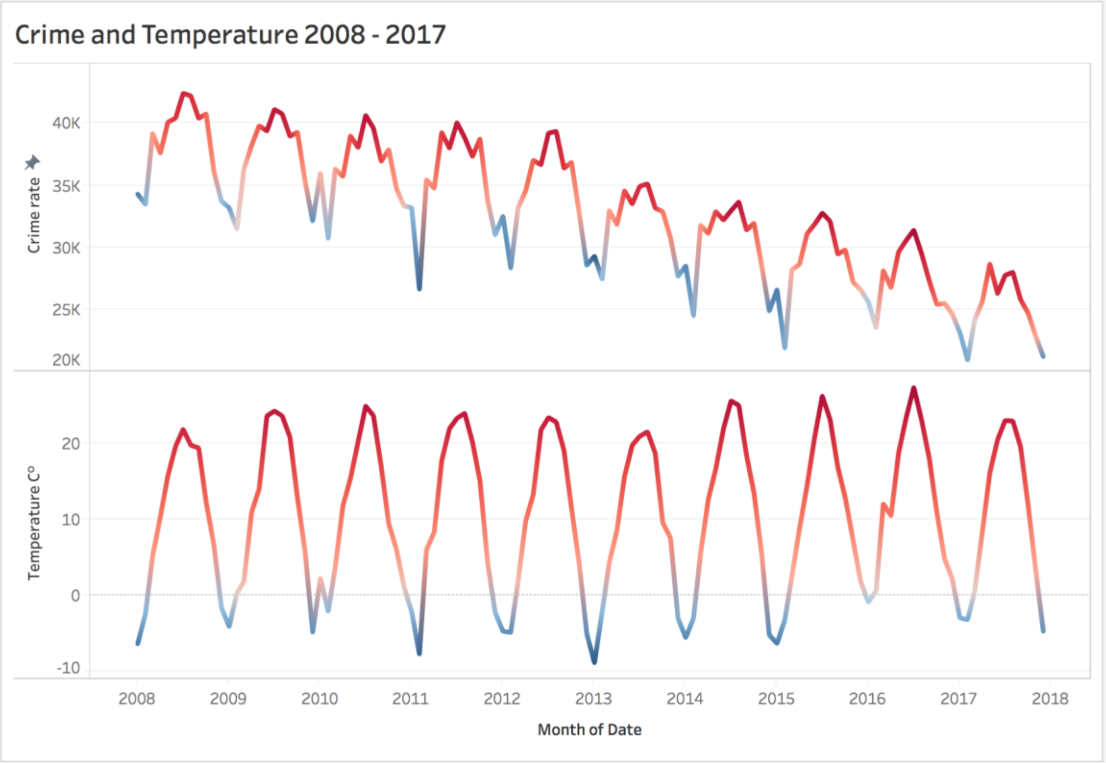
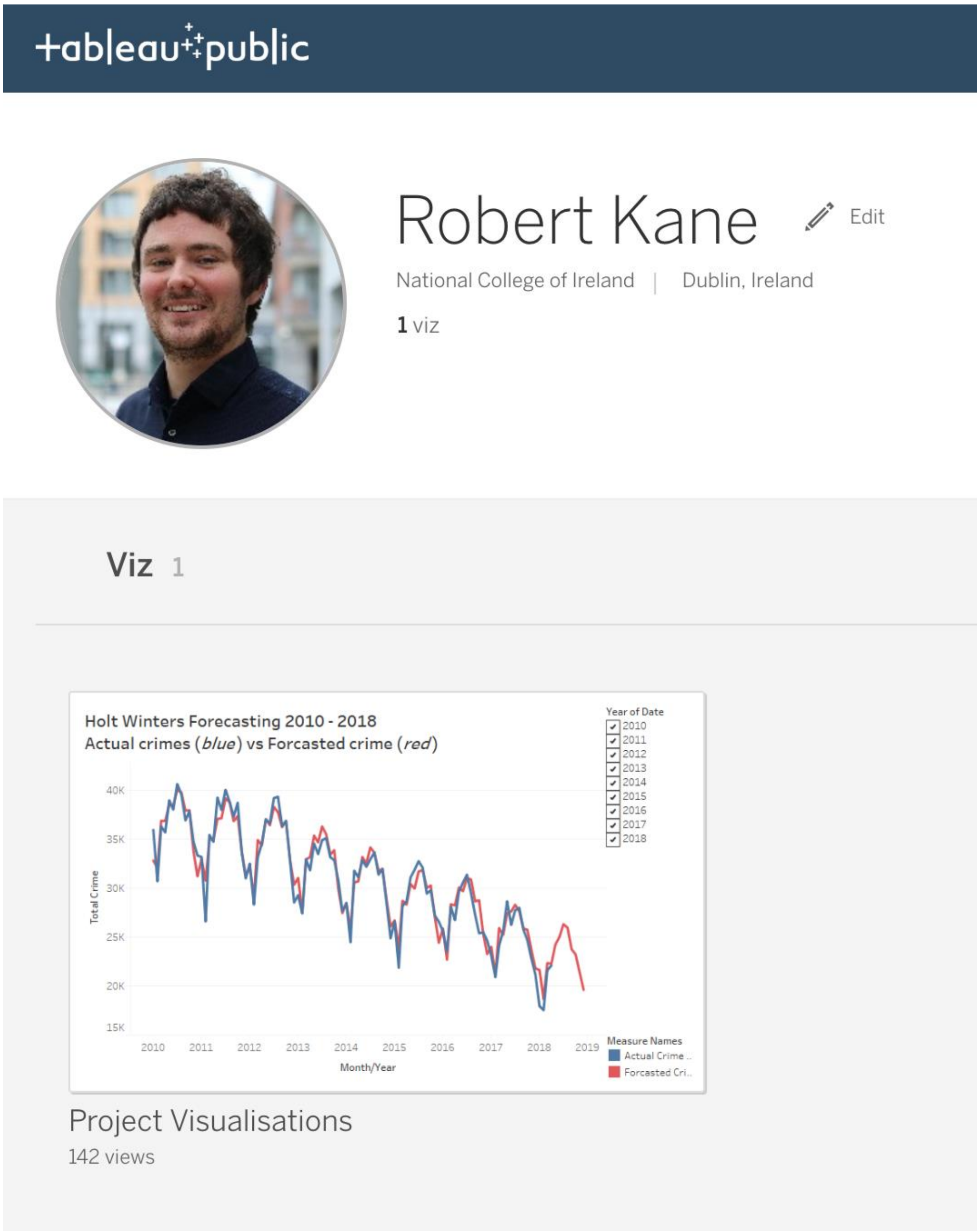


Fig 3: 2008 - 2017 overview of weather and crime trends

In order to achieve the goals of this project and create the required data visualisations, it was necessary to familiarise myself with the Tableau development environment and study various techniques to accurately visualise the findings of the research. Tableau was heavily implemented throughout the lifecycle of the project to investigate correlations, perform analysis and visually represent the results of the statistical analysis output from R Studio, IBM SPSS, Excel and the data warehouse reports. Output from these components were imported into Tableau as a data source and the required visualisations were generated. Further filters and manipulation of the data could then be applied for further emphasis of certain observations.

5.4.1. Tableau Public



+ableau+public

Robert Kane

National College of Ireland | Dublin, Ireland

1 viz

Viz 1

Holt Winters Forecasting 2010 - 2018
Actual crimes (*blue*) vs Forcasted crime (*red*)

Project Visualisations

142 views

Fig 4: Tableau public profile and hosted workbook

As part of the research, a platform was required to host and display the results online. Fig 4 shows the Tableau public profile that allowed for the publication of the workbook online. This is the access point for the end user to explore the results. Any alterations or subsequent development in the Tableau workbook throughout the project could be saved to the Tableau public profile and the changes would be reflected in the publically available workbook.

The Tableau Public profile and workbook can be found at:
https://public.tableau.com/profile/robert.kane#!/

5.5. Pearson's Correlation

5.5.1. Introduction

Pearson's correlation measures the strength of a linear correlation between two variables and the direction of the relationship. To measure the strength of the relationship between the two variables, a correlation coefficient value between the range of +1 and -1 is obtained where a value of +1 indicates a total positive correlation, 0 is no correlation and -1 indicates a total negative correlation.

To determine the strength of the relationship of the variables for this research and to investigate if a statistically significant correlation can be observed, a Pearson's Correlation test will be conducted to test the linear relationship of crime rate and weather conditions for the years 2008 – 2017.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Fig 5: Pearson's Correlation.

The Pearson Correlation test is the most widely used correlation statistic to measure the relationship between linearly related variables. Fig 5 shows the formula used to calculate Pearson's r Correlation where:

r = Correlation coefficient.

N = Number of observations.

Σ xy = Sum of the products of paired scores.

Σ x = Sum of x scores.

Σ y = Sum of y scores.

Σ x2= Sum of squared x scores.

Σ y2= Sum of squared y scores.

5.5.2. Method

To carry out the correlation test for this research, a correlation test to examine the annual cyclical relationship over a ten-year period was carried out, as well as observing the relationship on a year-by-year basis to account for changes over time and to examine the relationship on an annual basis. The hypotheses for the correlation tests are as follows:

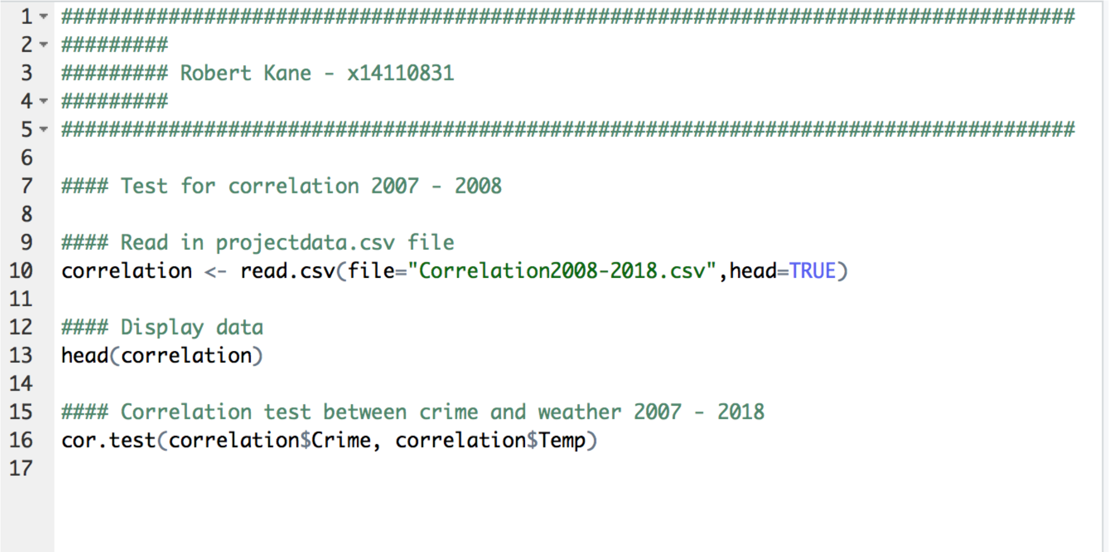
H_0: $\rho = 0$: There is no correlation between weather and crime
H_A: $\rho \neq 0$: There is a correlation between weather and crime

The following methods were implemented to ensure the tests were conducted properly and the obtained Pearson's r correlation test statistics were accurate.

Step 1: The required data to conduct the correlation test was aggregated and structured into the appropriate .csv files. This included aggregating the required data by month and by year in order to have a uniform structure to load and read into analysis software such as Tableau, R Studio and SPSS Statistics.

Step 2: A preliminary visual inspection of the data was developed in Tableau to investigate what kind of relationship could be observed when the variables were plotted on a scatter plot over the ten-year period. The relationship was further investigated by investigating the relationship for each individual year.

Step 3: In order to calculate the strength of the relationship and validate the accuracy of the obtained statistic, the test was carried out in both SPSS and R Studio.



```
#####################################################################################
#########
######### Robert Kane - x14110831
#########
#####################################################################################

#### Test for correlation 2007 - 2008

#### Read in projectdata.csv file
correlation <- read.csv(file="Correlation2008-2018.csv",head=TRUE)

#### Display data
head(correlation)

#### Correlation test between crime and weather 2007 - 2018
cor.test(correlation$Crime, correlation$Temp)

```

Fig 6: Pearson's correlation test performed in R Studio.

Step 4: Pearson's r correlation value was obtained and the relationship between weather and crime was measured against the required critical value in order to accept or reject the null hypothesis. The full test and explanation of the observed results are presented in the results section of this document.

5.6. Holt-Winters Method

5.6.1. Introduction

Triple exponential smoothing, also known as the Holt-Winters method, is a method used to forecast data points in a series that exhibits seasonality, i.e. a trend or pattern that repeats over time. When attempting to select a model for forecasting, the first step was to graph the sequence plots of the time series to be forecasted. The purpose of the sequence plot is to provide a visual overview of the characteristics of the times series. These observations would suggest if the data series for this research was exhibiting certain behavioural components such as trend and seasonality. In order to investigate if these components could be observed, the total crimes time series for the years 2008 through 2017 were plotted alongside the mean average temperature for that month in Tableau. The results of the time series are displayed in Fig 7.

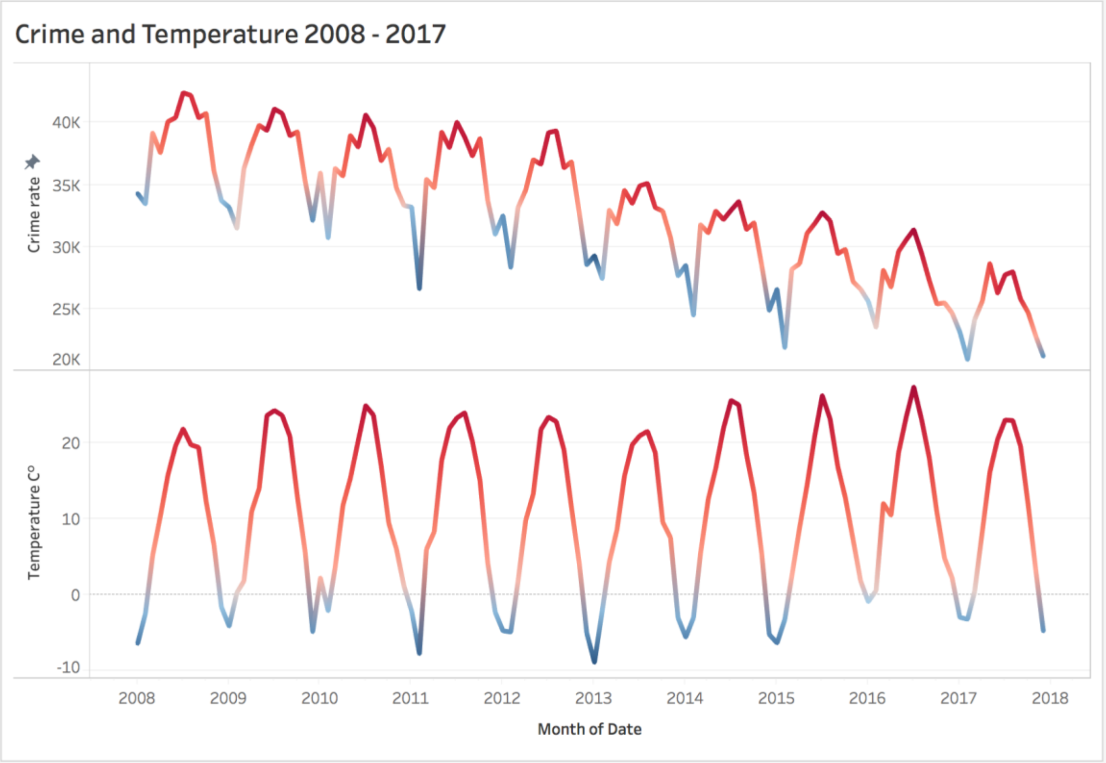


Fig 7: Data series displaying seasonality characteristics

The time series displayed above in Fig 7 demonstrates a clear pattern of seasonality. Overall crime rates in Chicago have been decreasing over the ten-year period. This may be down to any number of external factors such as a decrease in population or better policing methods, but despite this, the behavioural component of seasonality remains a consistent observable factor. Crime rates are generally observed to be 18% - 25% higher in the summer months than in the winter months. The average monthly temperatures for this time period can also be observed to remain consistent.

Now that the time series has been plotted and the components of seasonality have been satisfied, selecting the Holt-Winters method of exponential smoothing for the forecasting model is justified. The next step in implementing this method is specification. The process of specification involves selecting the required variables to be included, selecting the type of equation that best suits these variables and then estimating the values of the parameters for the equation. After the model is specified and the values of the parameters have been determined using R Studio, the performance accuracy of the model will be validated by comparison of the forecasts with historical data. For the purpose of this model and research, the method of error measurement will be MAPE (Mean absolute percentage error). MAPE is a highly accurate performance metric used to assess the accuracy of a forecasting model.

5.6.2. Terminology

Multiplicative Seasonality

As previously discussed, when the time series for this research is plotted, the data demonstrates a clear pattern of seasonality. Seasonality means that characteristics in the data repeat after a certain number of periods. Therefore, the term season represents the time before the cycle repeats.

Taking the data observed in this research, crime statistics are observed to demonstrate an increase of 18% - 25% during the summer months, which equates to an increase factor of 1.18 - 1.25. Overall crime statistics demonstrate a steady decrease over the ten-year period, however the seasonality factor remains proportionately constant. This means that as the absolute number of crimes recorded is decreasing, the seasonal increase for the absolute number of crimes will remain proportionately relative to the observed value. For instance, regardless of the absolute number of crimes recorded in the winter period, we can be confident that the increase in crime for the summer period will be 18% - 25%. In this case, crime rates increase by a factor and the seasonal component is therefore described as multiplicative.

Base, Trend & Seasonality

In order to comprehend the behaviour of the forecasting model implemented in this research, an understanding of the three principal characteristics is required. The three basic characteristics are base, trend and seasonality. The base of a time series refers to the series current level in the absence of seasonality. The trend of a time series is the increase per period in the base. In the case of this study, a trend of 1.2 would indicate an increase of 20%. Finally, seasonality index describes how far above or below an average month you can expect the actual crime total to be.

5.6.3. Method

Holt-Winters method is comprised of three principle equations that update the level, trend and seasonal index of the model.

Formula 1: Base Level

$$\bar{R}_t = \alpha \frac{y_t}{\bar{S}_{t-L}} + (1-\alpha) * (\bar{R}_{t-1} + \bar{G}_{t-1})$$

The first formula determines the new base estimate, which is a simple deseasonalised weighted average of the current season and the last periods base updated by the latest trend estimate.

Formula 2: Trend

$$\bar{G}_t = \beta * (\bar{S}_t - \bar{S}_{t-1}) + (1-\beta) * \bar{G}_{t-1}$$

The estimate of the trend parameter is the smoothed difference between the last two estimates of the deseasonalised base level values.

Formula 3: Seasonality Index

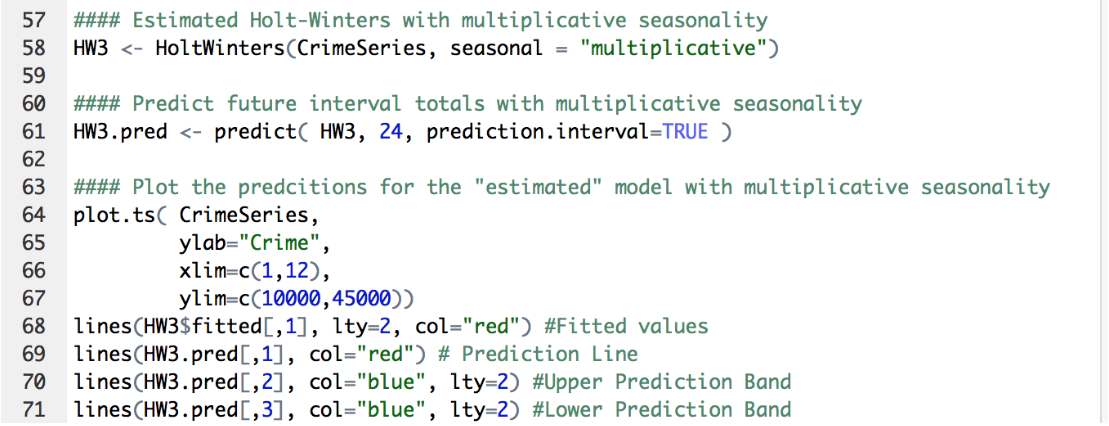
$$\bar{S}_t = \gamma * (y_t / \bar{S}_t) + (1-\gamma) * \bar{S}_{t-L}$$

The seasonality formula updates the seasonal index as a weighted average of the estimated seasonal index determined by the current and previous period's estimate.

The equations account for smoothing parameters *alpha, beta* and *gamma.* The method for deriving the values for these parameters must be estimated in such a way that minimises the mean absolute percentage error value and as such, is best left to analysis software to correctly determine.

Estimating the smoothing constants

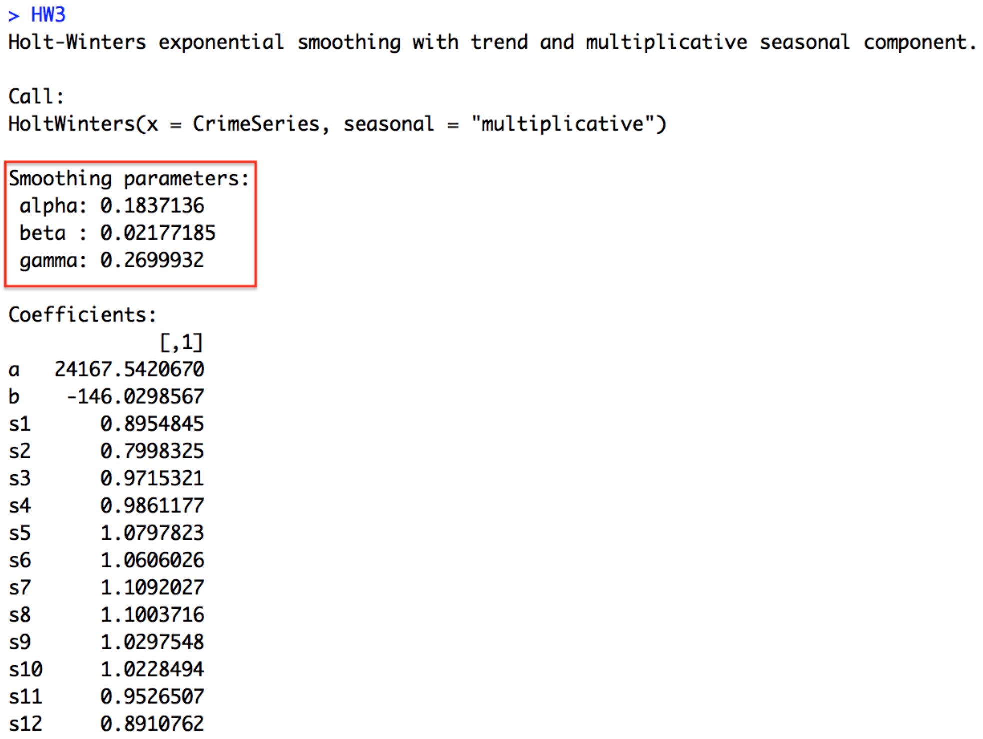
In order to determine the best possible *alpha, beta* and *gamma* values for the forecasting model for this research, a test was conducted in R Studio.



```r
#### Estimated Holt-Winters with multiplicative seasonality
HW3 <- HoltWinters(CrimeSeries, seasonal = "multiplicative")

#### Predict future interval totals with multiplicative seasonality
HW3.pred <- predict( HW3, 24, prediction.interval=TRUE )

#### Plot the predcitions for the "estimated" model with multiplicative seasonality
plot.ts( CrimeSeries,
         ylab="Crime",
         xlim=c(1,12),
         ylim=c(10000,45000))
lines(HW3$fitted[,1], lty=2, col="red") #Fitted values
lines(HW3.pred[,1], col="red") # Prediction Line
lines(HW3.pred[,2], col="blue", lty=2) #Upper Prediction Band
lines(HW3.pred[,3], col="blue", lty=2) #Lower Prediction Band
```

Fig 8: Holt-Winters test in R Studio

Fig 8 shows the R code implemented to calculate the best possible values for the smoothing parameters. The R functions were coded to cater for multiplicative seasonality and a forecast interval of twenty-four periods was projected. In the case of this test, the periods represent months, so a projected forecast of two years is being made.



```
> HW3
Holt-Winters exponential smoothing with trend and multiplicative seasonal component.

Call:
HoltWinters(x = CrimeSeries, seasonal = "multiplicative")

Smoothing parameters:
 alpha: 0.1837136
 beta : 0.02177185
 gamma: 0.2699932

Coefficients:
             [,1]
a   24167.5420670
b    -146.0298567
s1      0.8954845
s2      0.7998325
s3      0.9715321
s4      0.9861177
s5      1.0797823
s6      1.0606026
s7      1.1092027
s8      1.1003716
s9      1.0297548
s10     1.0228494
s11     0.9526507
s12     0.8910762
```

Fig 9: R Code parameter value estimations.

Fig 9 shows the output of the R code to determine the smoothing parameter estimates as:

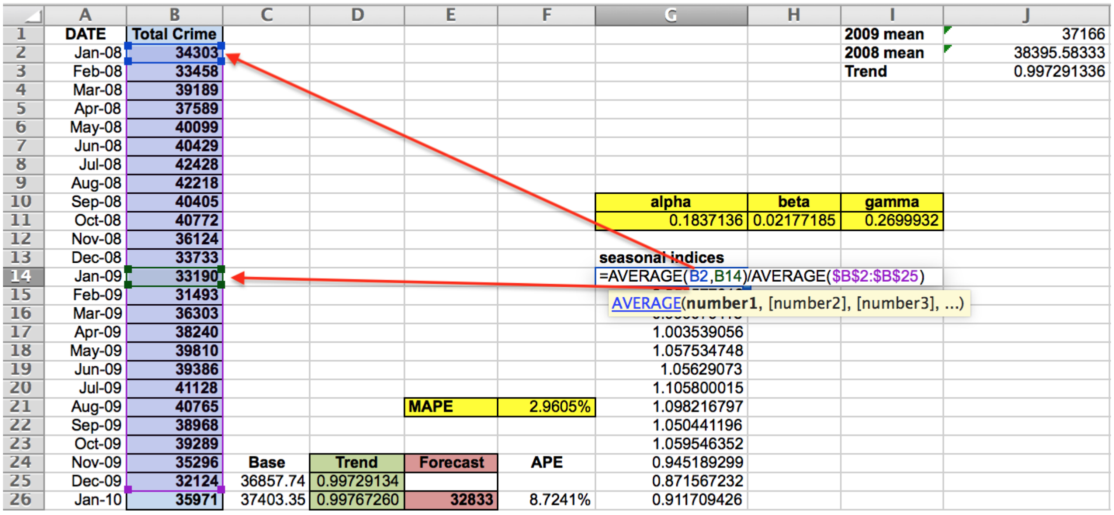
alpha :	0.1837136
beta :	0.02177185
gamma:	0.2699932

Now that these values had been computed, the next steps in the development of the forecasting model can be implemented.

Initializing Holt-Winters Method

To initialize Holt-Winters method, the initial values for base, trend and seasonality must be calculated before the smoothing parameters determined by R Studio are called in. To achieve this, monthly crime totals for the years 2008 and 2009 were used to establish the initial values. The method is detailed below:

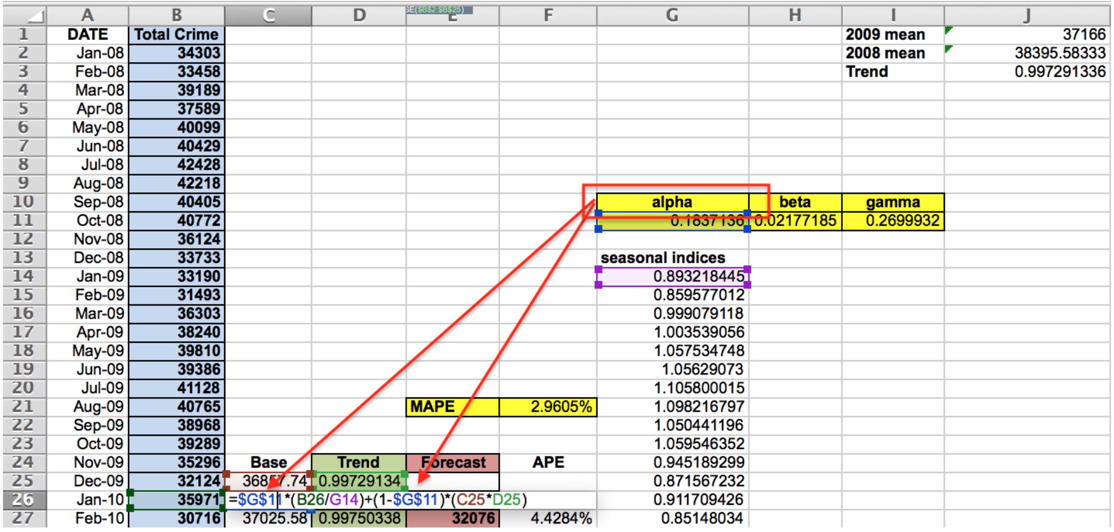
The initial estimate for the first January seasonal index value is obtained by calculating the average January crime rate for 2008 and 2009 divided by the average monthly crime for the entire years 2008 and 2009. Fig 10 shows the calculation being performed in Excel.



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					=AVERAGE(B2,B14)/AVERAGE($B$2:$B$25)			
15	Feb-09	31493					AVERAGE(number1, [number2], [number3], ...)			
16	Mar-09	36303					[illegible]			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	37403.35	0.99767260	32833	8.7241%	0.911709426			

Fig 10: Calculating the seasonal index for January 2009

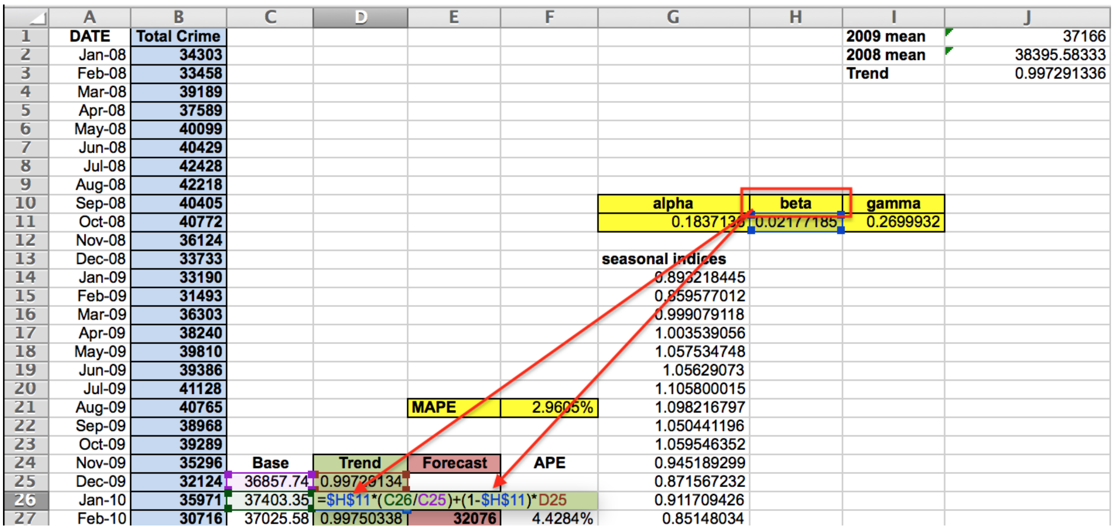
This formula is repeated until the forecast period starting January 2010 is reached. At this point the formula calls in the smoothing parameters *alpha, beta* and *gamma* determined by R Studio to calculate the forecasted values.



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					0.893218445			
15	Feb-09	31493					0.859577012			
16	Mar-09	36303					0.999079118			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	=$G$11*(B26/G14)+(1-$G$11)*(C25*D25)				0.911709426			
27	Feb-10	30716	37025.58	0.99750338	32076	4.4284%	0.85148034			

Fig 11: Alpha value smoothing the new base value.

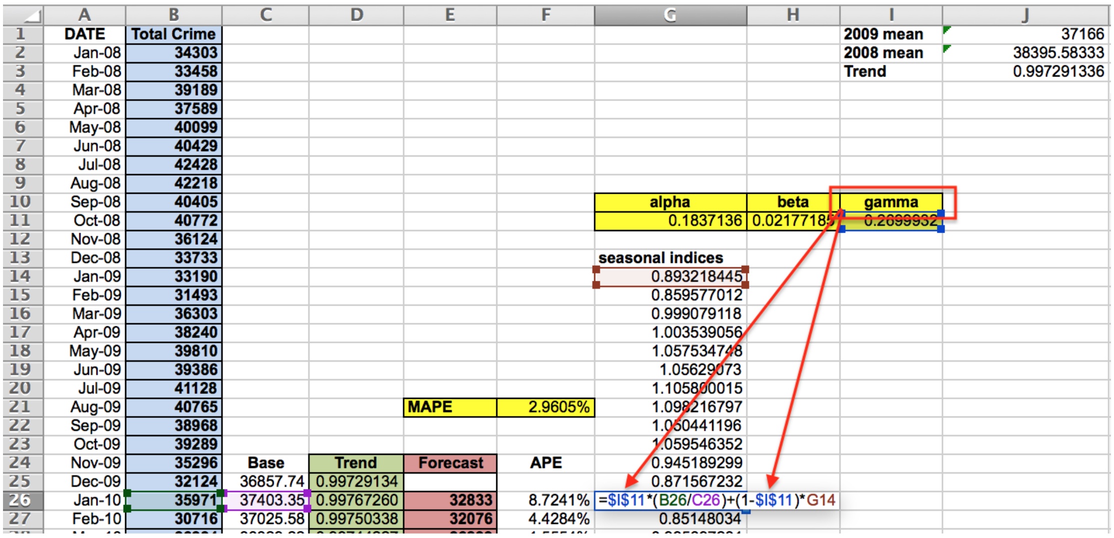
Fig 11 shows the alpha value determined by R Studio being called into the base level formula to smooth the base level of the time series forecast.



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					0.893218445			
15	Feb-09	31493					0.859577012			
16	Mar-09	36303					0.999079118			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	37403.35	=$H$11*(C26/C25)+(1-$H$11)*D25			0.911709426			
27	Feb-10	30716	37025.58	0.99750338	32076	4.4284%	0.85148034			

Fig 12: Beta value smoothing the trend value.

Fig 12 shows the beta value being called in to the trend function to smooth the trend value.



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					0.893218445			
15	Feb-09	31493					0.859577012			
16	Mar-09	36303					0.999079118			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	37403.35	0.99767260	32833	8.7241%	=$I$11*(B26/C26)+(1-$I$11)*G14			
27	Feb-10	30716	37025.58	0.99750338	32076	4.4284%	0.85148034			

Fig 13: Gamma value smoothing the seasonal indices.

Fig 13 shows the gamma value being called in to the seasonal index formula to smooth the seasonality index value.

The next step was to continue the formula throughout the forecasted time series. Once this had been completed and the forecasted values were obtained, the data was gathered from the Excel formula spread sheet and the results of the computations made in R Studio were plotted in the Tableau workbook along with the observed actual values. The accuracy of the predictions is measured against the actual crime totals recorded after the predicted forecast period has come to pass. Crime statistics for the forecasted period will be gathered and aggregated and the accuracy will be measured and expressed as a mean absolute percentage error statistic. The results of the forecast are displayed and discussed in the results section of this document.

5.7. Conclusion

This section of the paper outlined the various components implemented and methodology required to achieve the goals of this project. As they will identify if a statistically significant correlation can be observed in the relationship between weather and crime, the tests conducted are suitable to satisfy the questions for this research. The results of the tests will then be assessed to identify if the variables can be suitably structured to develop an accurate model for forecasting future trends. The interpretation of the results and visualisations of these tests will be presented in the next section of the document.

6. Results

6.1. Pearson’s Correlation Test

6.1.1. Introduction

To investigate if a relationship between incidents of crime and weather could be observed, a number of visualisations were developed plotting the variables. These initial graphs and scatter plots provided a preliminary indication that a strong positive relationship between the two variables could be observed. Fig 14 shows the average number of crimes recorded at each incremental rise in ambient temperature. As the graph shows, a steady increase in the number of crimes can be observed as the average ambient temperature rises. Although this presents strong evidence of a statistically significant correlation, further testing was required to confirm this assumption.

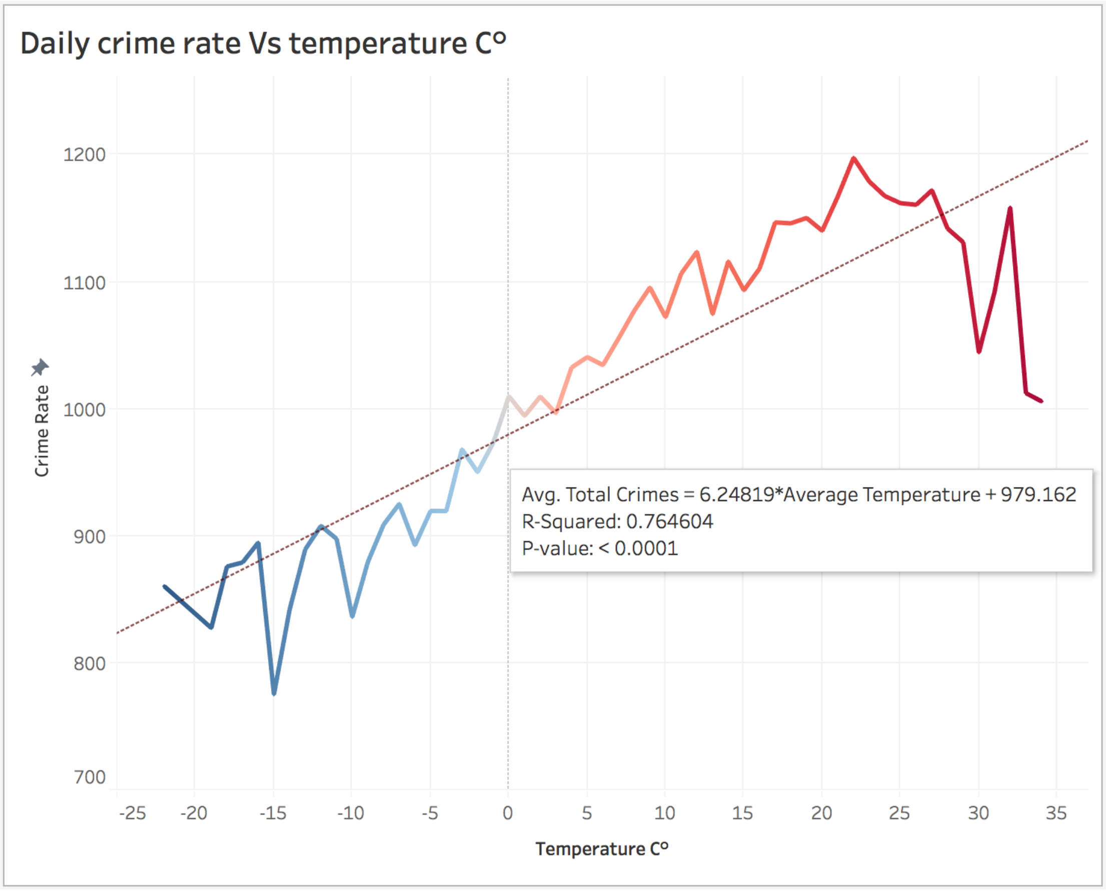


Fig 14: Crime rate rises as temperature rises

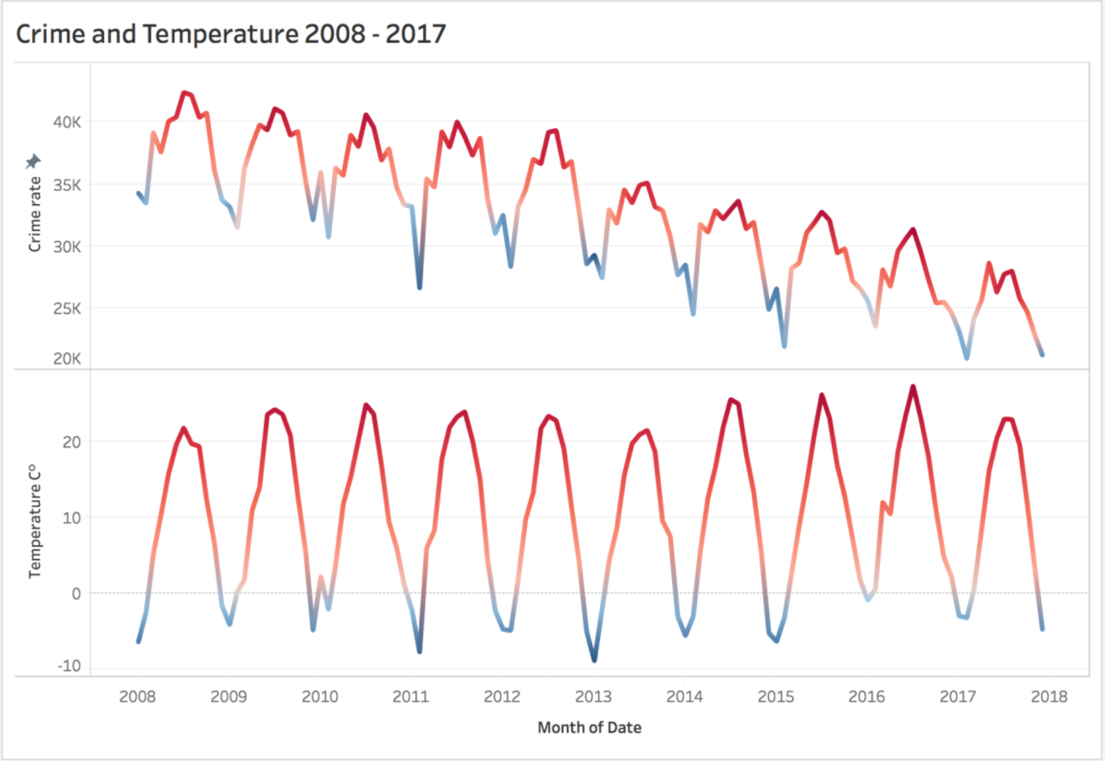
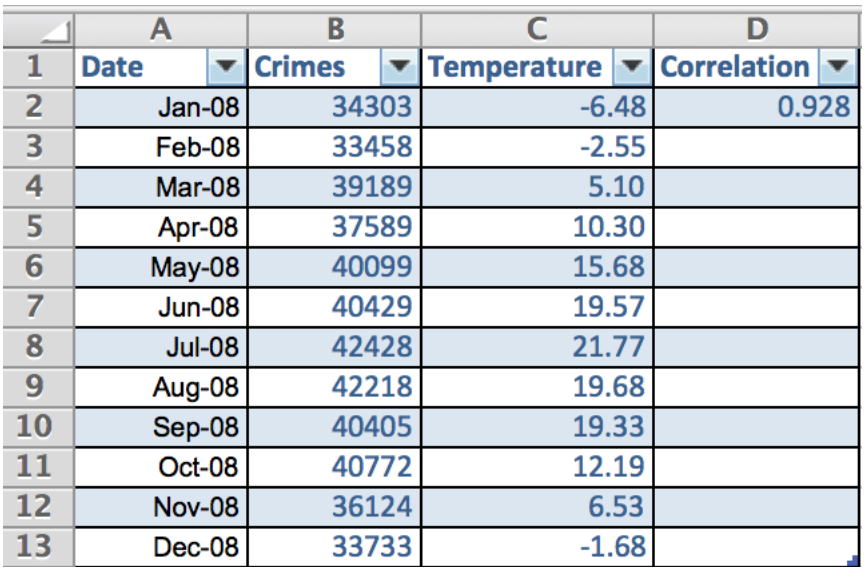


Fig 15: Observable seasonality and correlation

Fig 15 shows the time series for the average number of crimes and temperature by month over the ten-year period of 2008 through to 2017. This plot demonstrates a compelling overview of the seasonal relationship between incidents of crime and temperature. The graph strongly indicates that a statistically significant correlation does exist between the two variables. The trend and pattern of the data for each variable is observed to almost run in sequence with one another. This visualisation presents clear evidence of a correlation between the two, but a Pearson's correlation test was required to determine the exact statistical significance of this correlation.

To examine the statistical significance of the correlation indicated by the visualisations presented above, a Pearson's correlation test was conducted for each year of the study, as well as a correlation test for the overall ten-year period. These tests were carried out in Excel, IBM SPSS Statistics and R Studio, and the results were visualised in Tableau. The results of the tests are presented below.

6.1.2. Test Results



	A	B	C	D
1	Date	Crimes	Temperature	Correlation
2	Jan-08	34303	-6.48	0.928
3	Feb-08	33458	-2.55	
4	Mar-08	39189	5.10	
5	Apr-08	37589	10.30	
6	May-08	40099	15.68	
7	Jun-08	40429	19.57	
8	Jul-08	42428	21.77	
9	Aug-08	42218	19.68	
10	Sep-08	40405	19.33	
11	Oct-08	40772	12.19	
12	Nov-08	36124	6.53	
13	Dec-08	33733	-1.68	

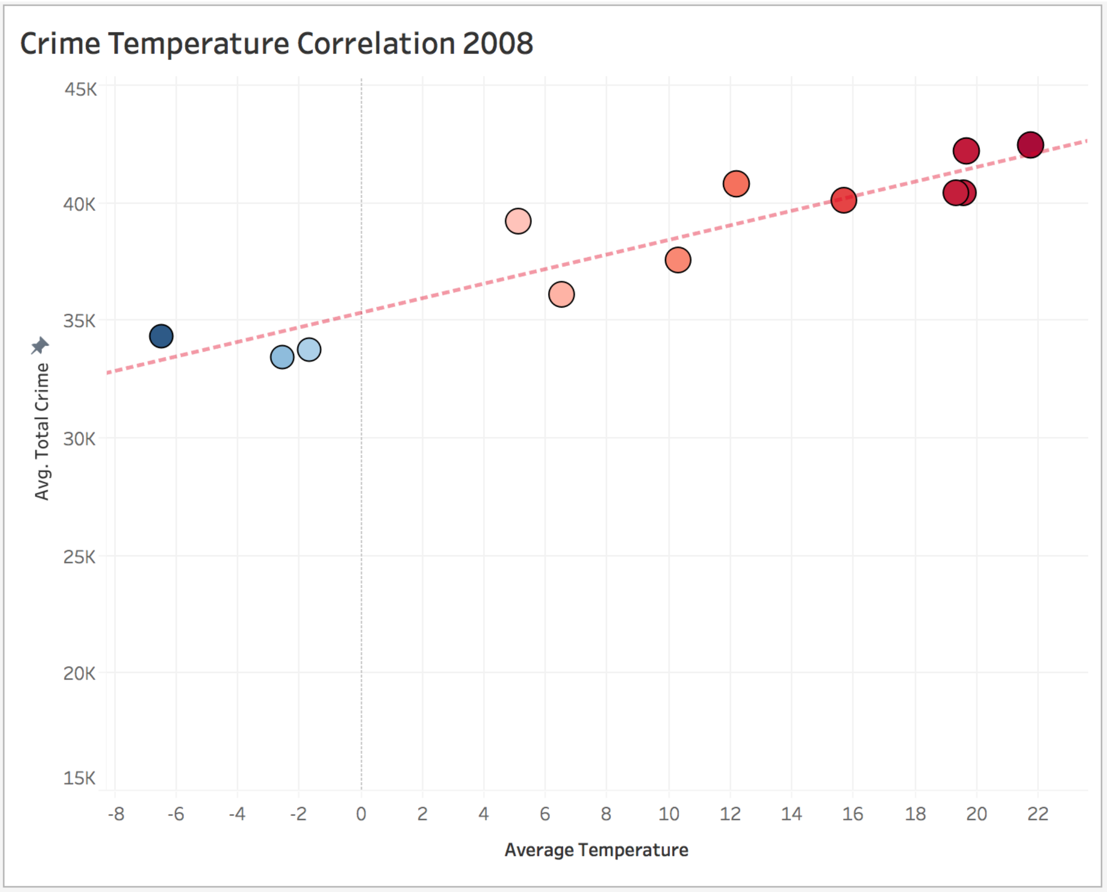
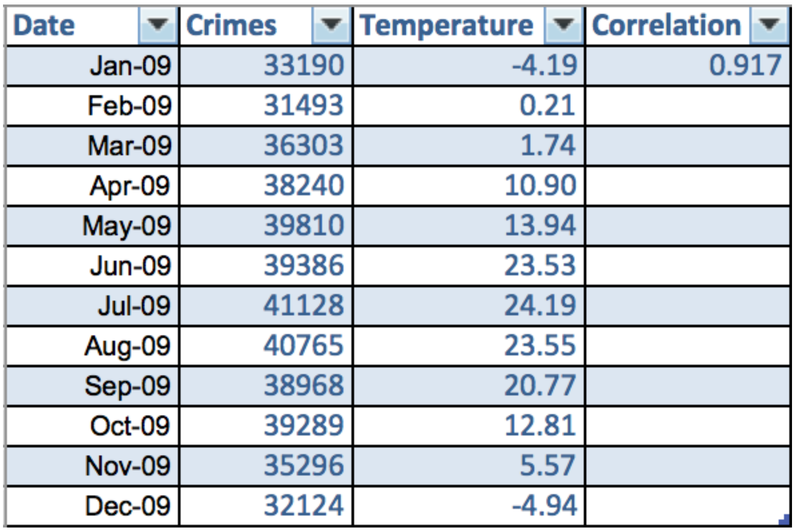


Fig 16: Correlation for 2008



Date	Crimes	Temperature	Correlation
Jan-09	33190	-4.19	0.917
Feb-09	31493	0.21	
Mar-09	36303	1.74	
Apr-09	38240	10.90	
May-09	39810	13.94	
Jun-09	39386	23.53	
Jul-09	41128	24.19	
Aug-09	40765	23.55	
Sep-09	38968	20.77	
Oct-09	39289	12.81	
Nov-09	35296	5.57	
Dec-09	32124	-4.94	

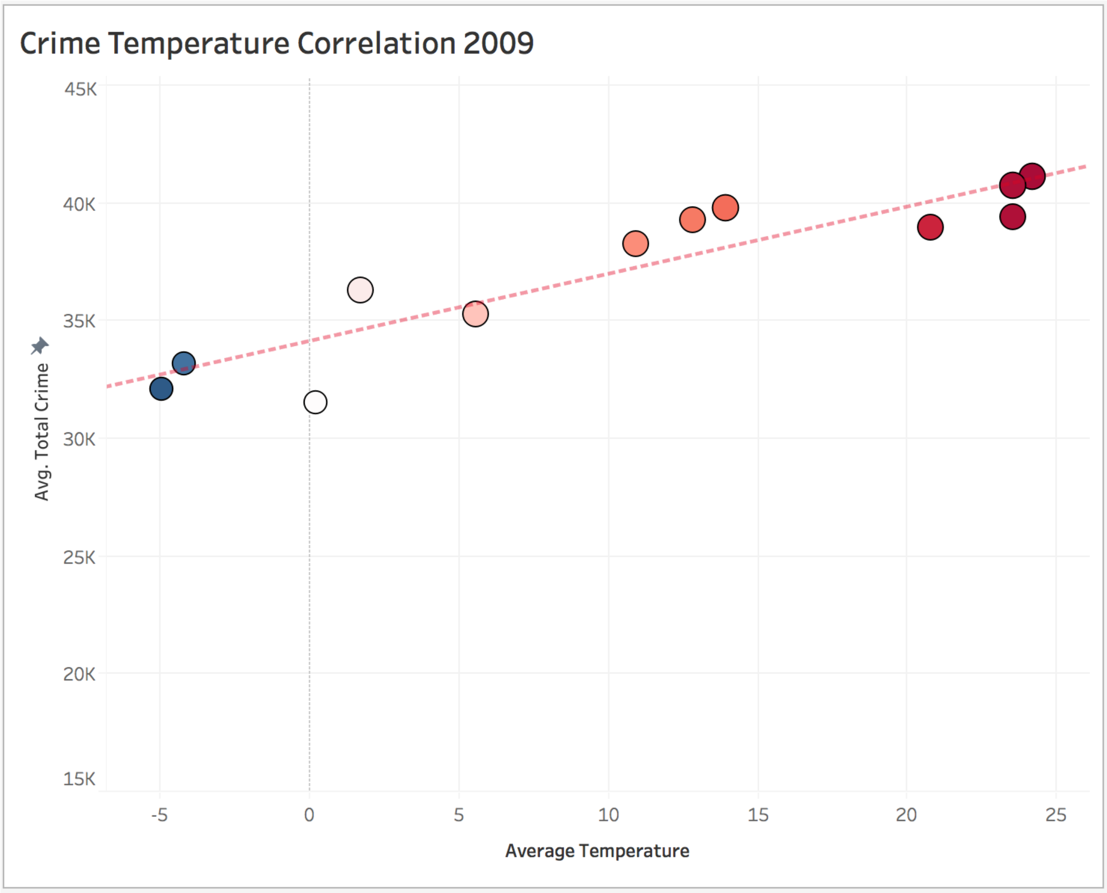
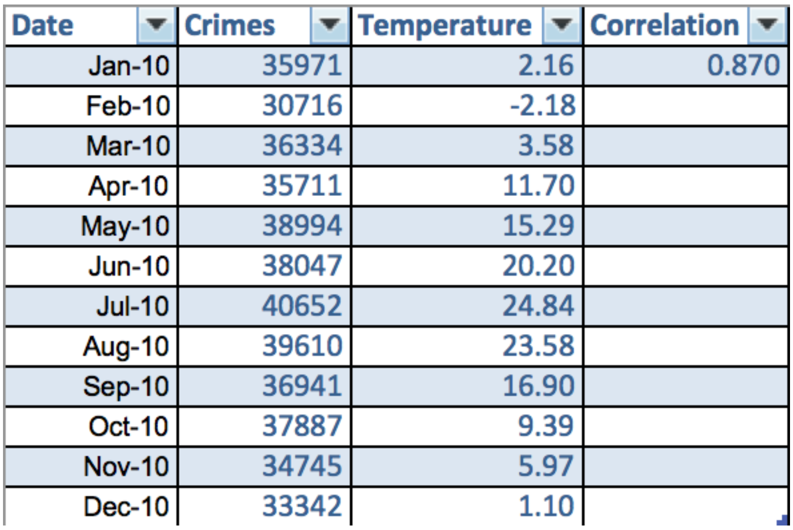


Fig 17: Correlation for 2009



Date	Crimes	Temperature	Correlation
Jan-10	35971	2.16	0.870
Feb-10	30716	-2.18	
Mar-10	36334	3.58	
Apr-10	35711	11.70	
May-10	38994	15.29	
Jun-10	38047	20.20	
Jul-10	40652	24.84	
Aug-10	39610	23.58	
Sep-10	36941	16.90	
Oct-10	37887	9.39	
Nov-10	34745	5.97	
Dec-10	33342	1.10	

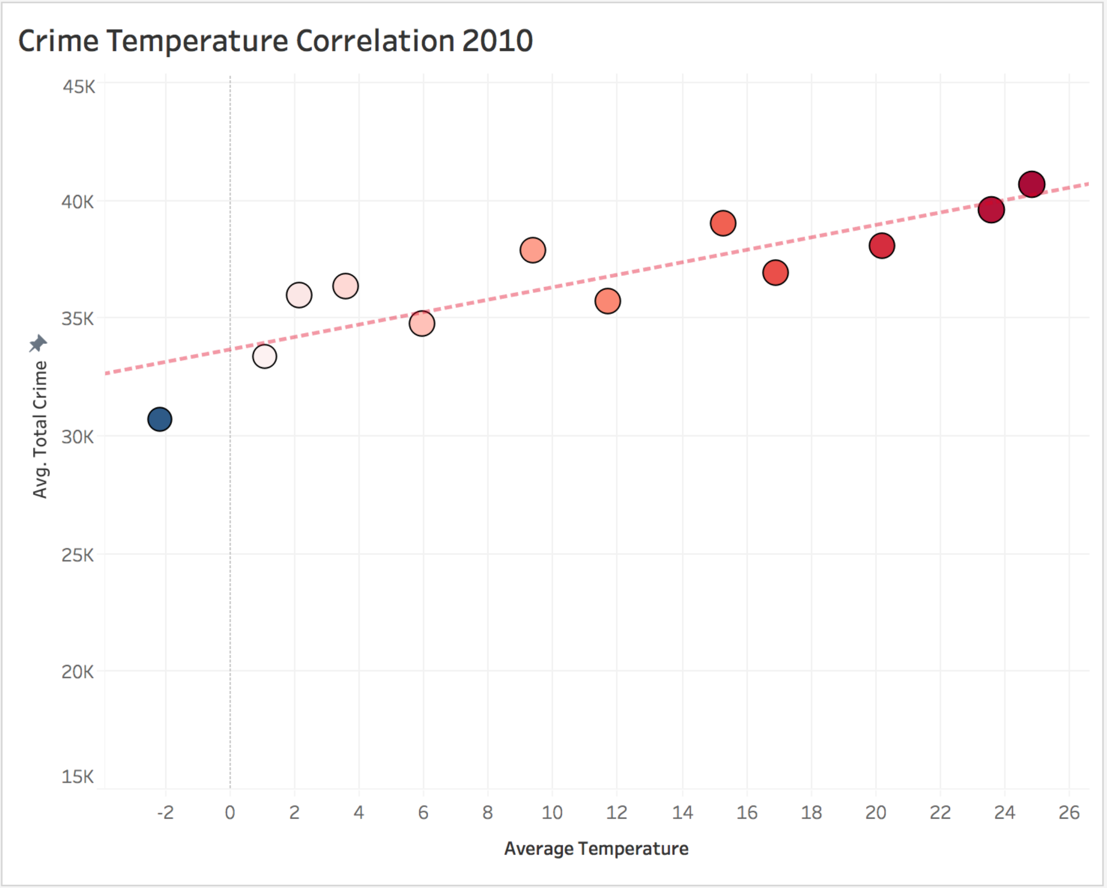
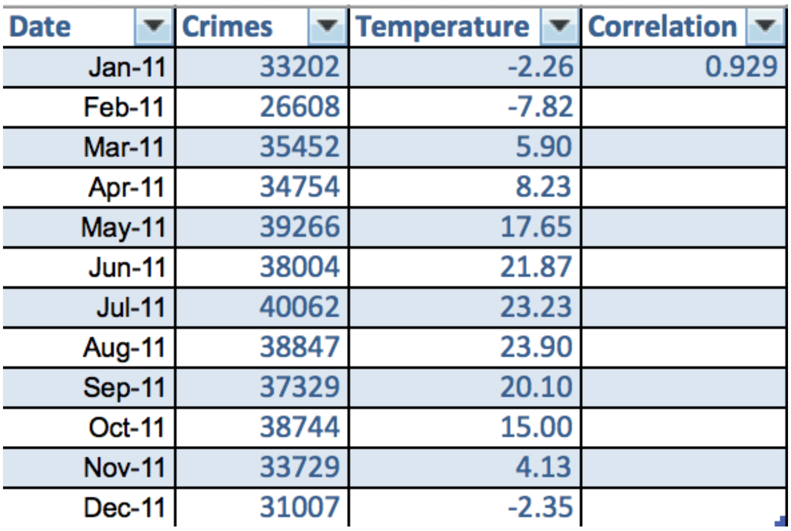


Fig 18: Correlation for 2010



Date	Crimes	Temperature	Correlation
Jan-11	33202	-2.26	0.929
Feb-11	26608	-7.82	
Mar-11	35452	5.90	
Apr-11	34754	8.23	
May-11	39266	17.65	
Jun-11	38004	21.87	
Jul-11	40062	23.23	
Aug-11	38847	23.90	
Sep-11	37329	20.10	
Oct-11	38744	15.00	
Nov-11	33729	4.13	
Dec-11	31007	-2.35	

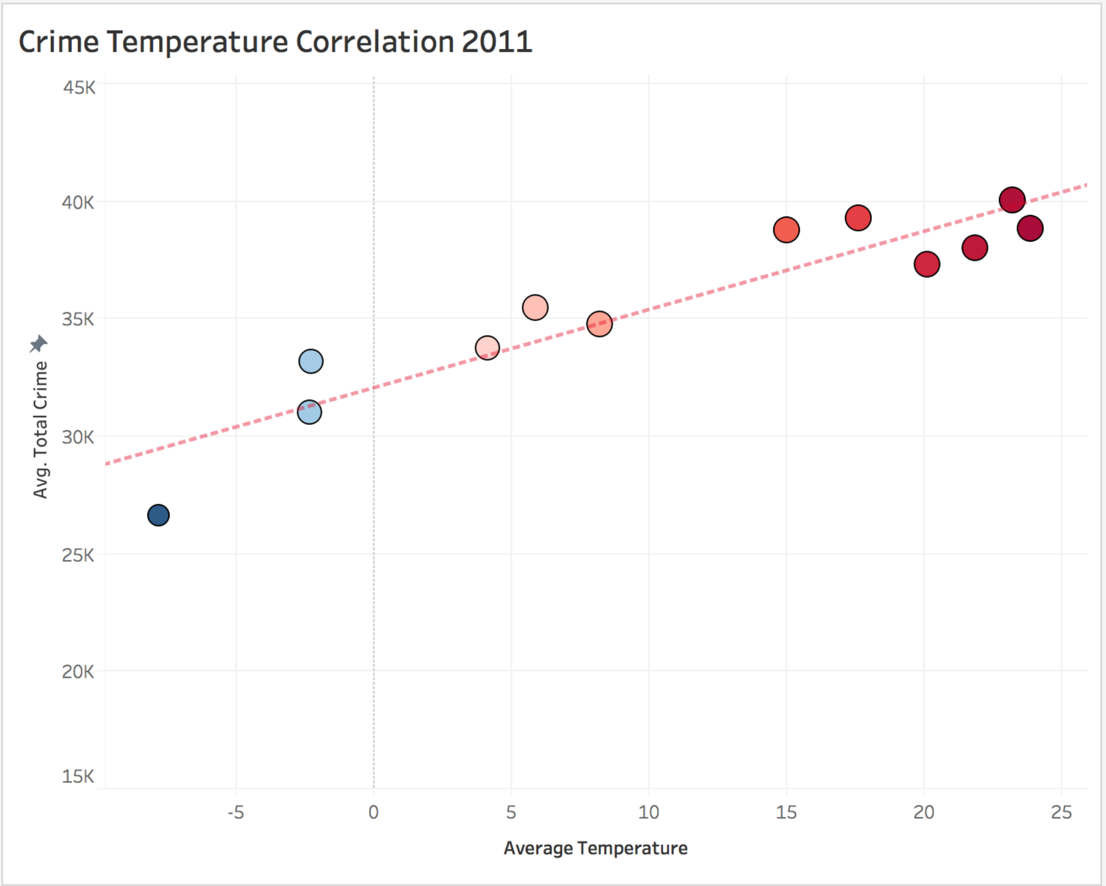
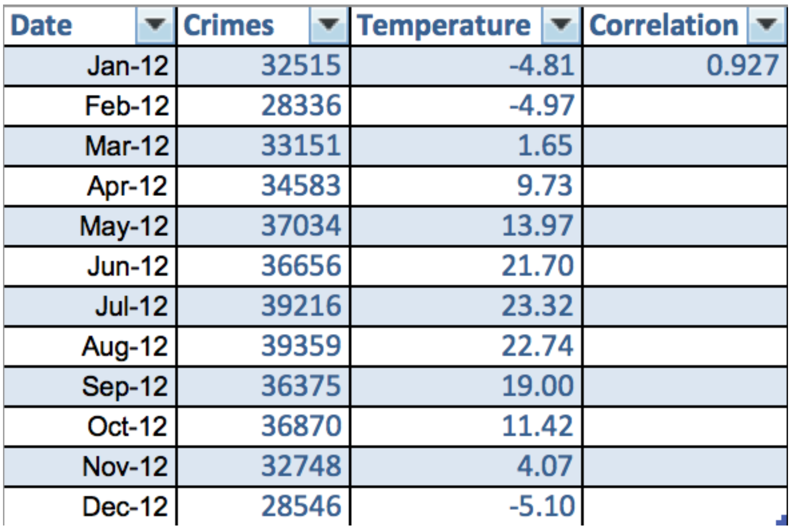


Fig 19: Correlation for 2011



Date	Crimes	Temperature	Correlation
Jan-12	32515	-4.81	0.927
Feb-12	28336	-4.97	
Mar-12	33151	1.65	
Apr-12	34583	9.73	
May-12	37034	13.97	
Jun-12	36656	21.70	
Jul-12	39216	23.32	
Aug-12	39359	22.74	
Sep-12	36375	19.00	
Oct-12	36870	11.42	
Nov-12	32748	4.07	
Dec-12	28546	-5.10	

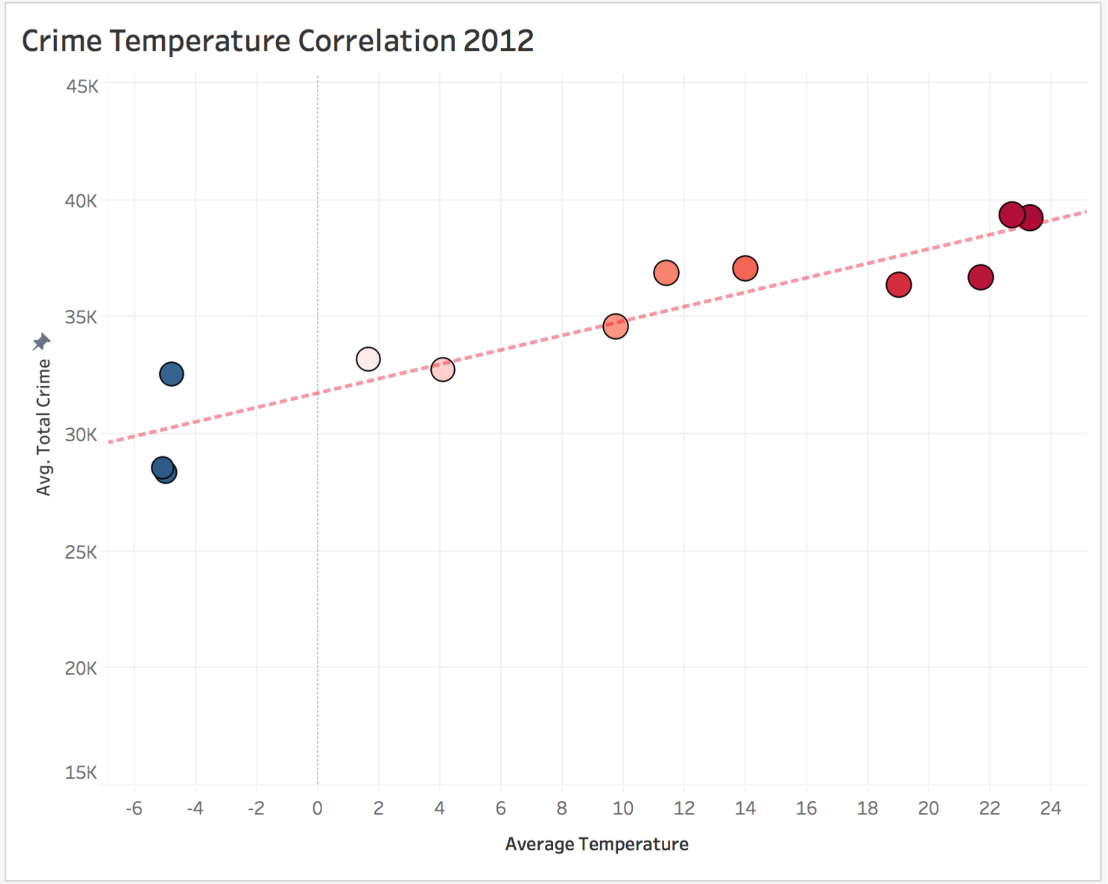
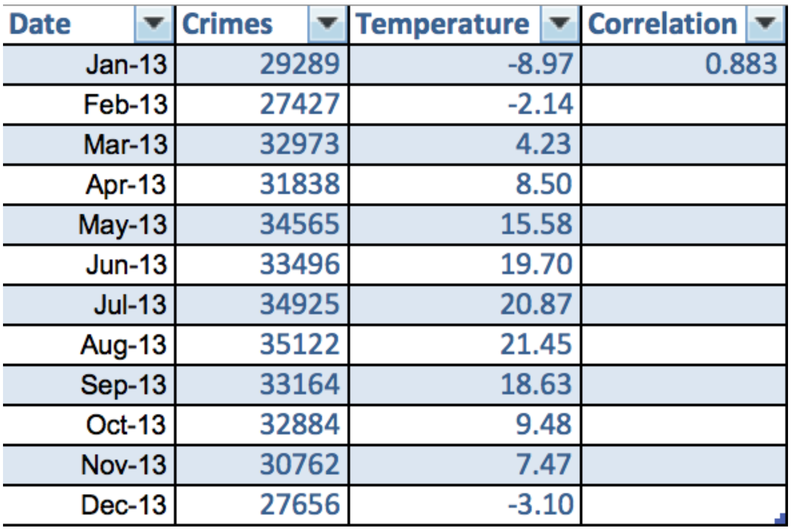


Fig 20: Correlation for 2012



Date	Crimes	Temperature	Correlation
Jan-13	29289	-8.97	0.883
Feb-13	27427	-2.14	
Mar-13	32973	4.23	
Apr-13	31838	8.50	
May-13	34565	15.58	
Jun-13	33496	19.70	
Jul-13	34925	20.87	
Aug-13	35122	21.45	
Sep-13	33164	18.63	
Oct-13	32884	9.48	
Nov-13	30762	7.47	
Dec-13	27656	-3.10	

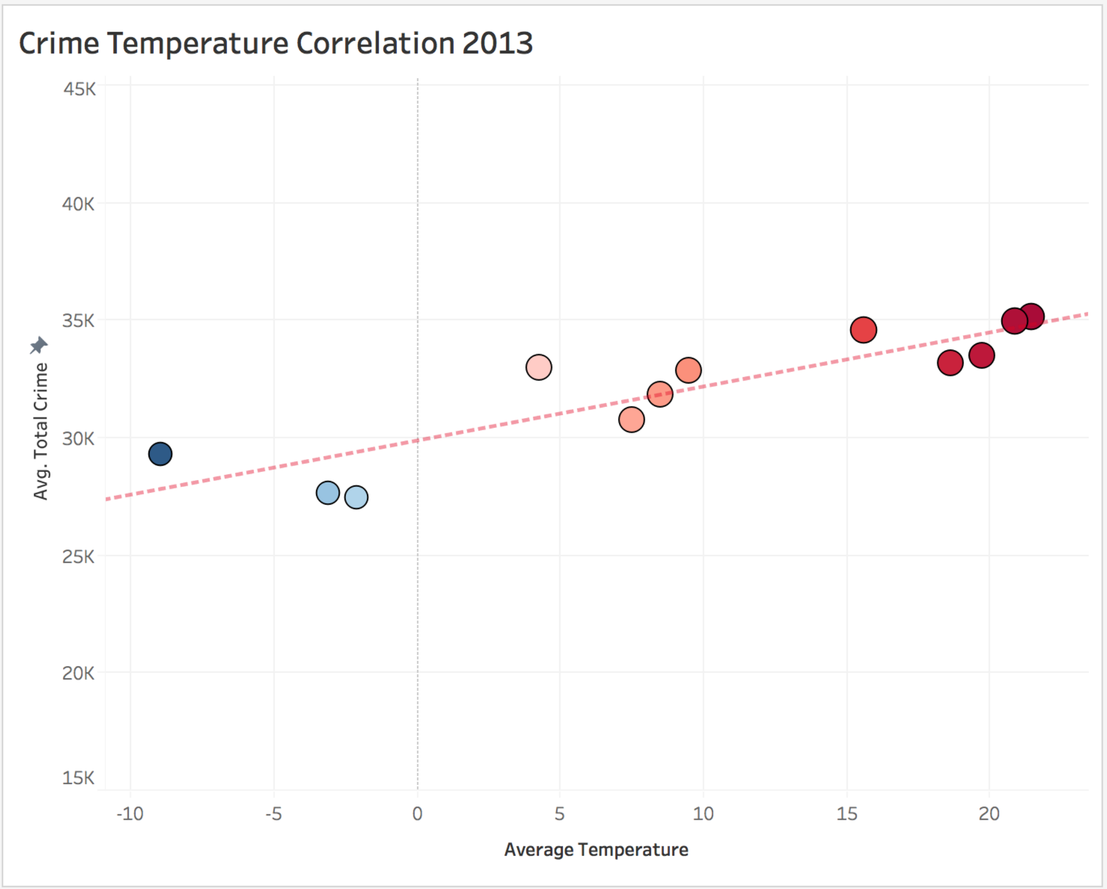
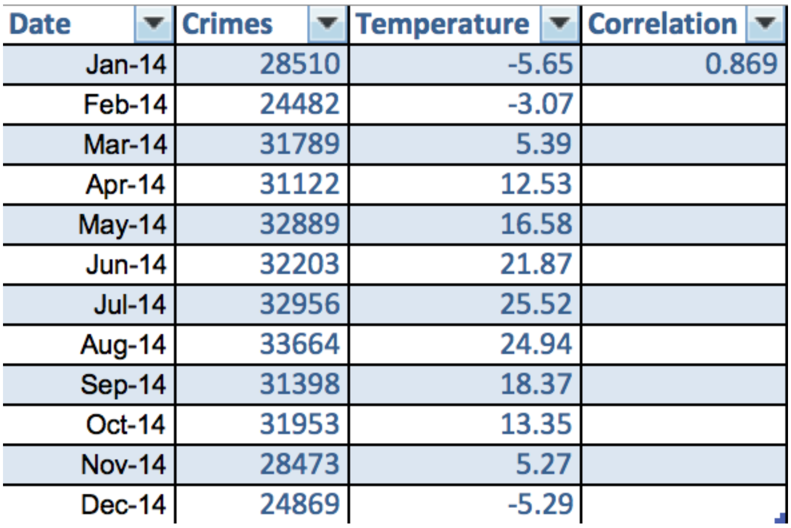


Fig 21: Correlation for 2013



Date	Crimes	Temperature	Correlation
Jan-14	28510	-5.65	0.869
Feb-14	24482	-3.07	
Mar-14	31789	5.39	
Apr-14	31122	12.53	
May-14	32889	16.58	
Jun-14	32203	21.87	
Jul-14	32956	25.52	
Aug-14	33664	24.94	
Sep-14	31398	18.37	
Oct-14	31953	13.35	
Nov-14	28473	5.27	
Dec-14	24869	-5.29	

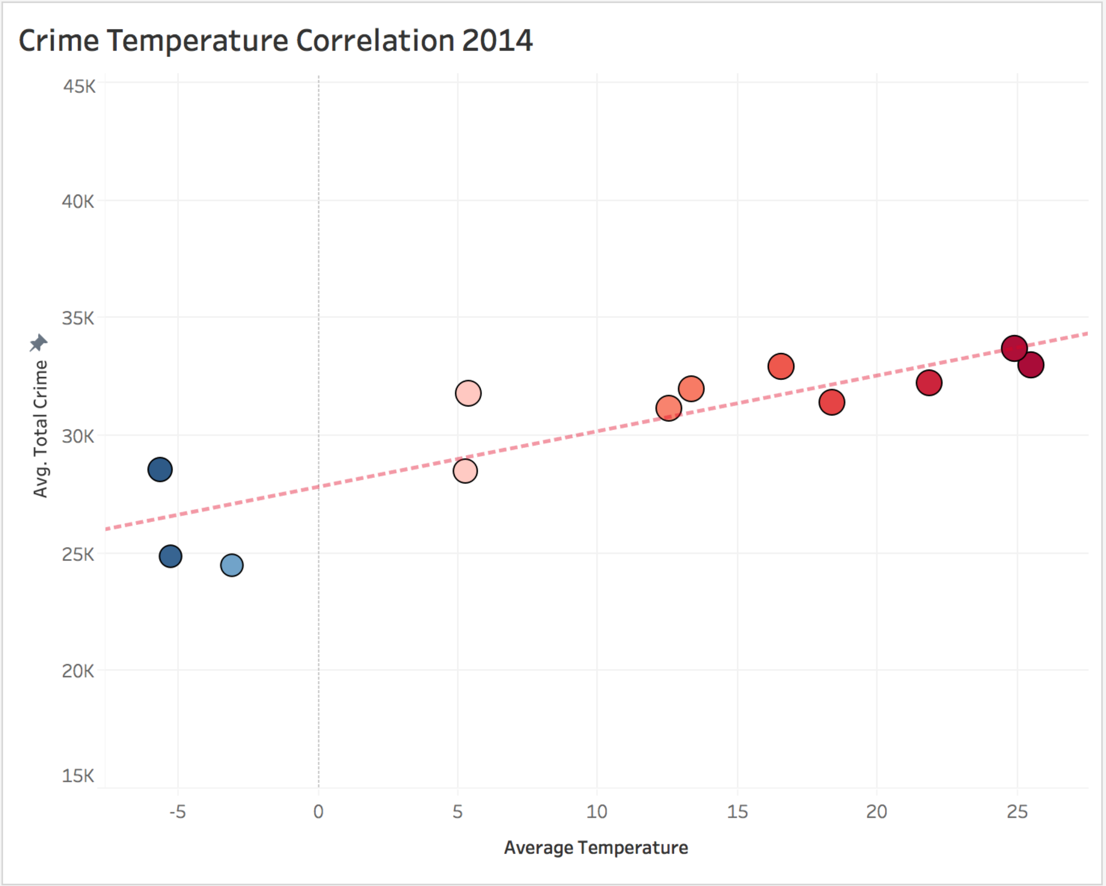
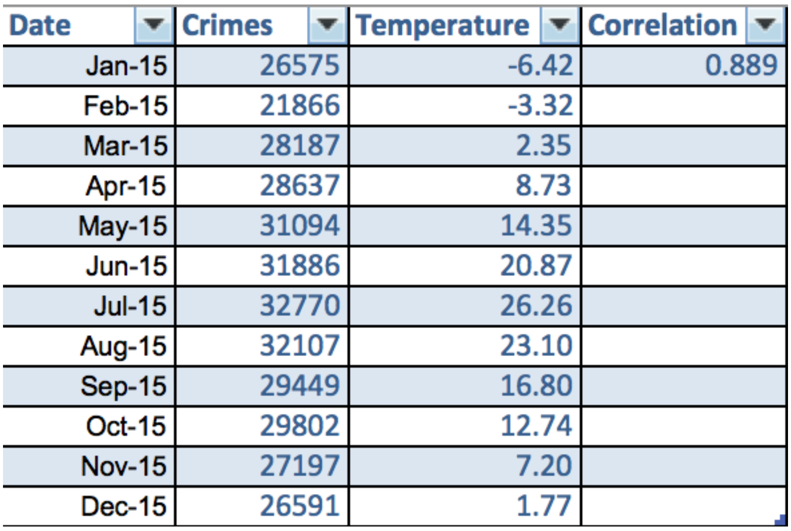


Fig 22: Correlation for 2014



Date	Crimes	Temperature	Correlation
Jan-15	26575	-6.42	0.889
Feb-15	21866	-3.32	
Mar-15	28187	2.35	
Apr-15	28637	8.73	
May-15	31094	14.35	
Jun-15	31886	20.87	
Jul-15	32770	26.26	
Aug-15	32107	23.10	
Sep-15	29449	16.80	
Oct-15	29802	12.74	
Nov-15	27197	7.20	
Dec-15	26591	1.77	

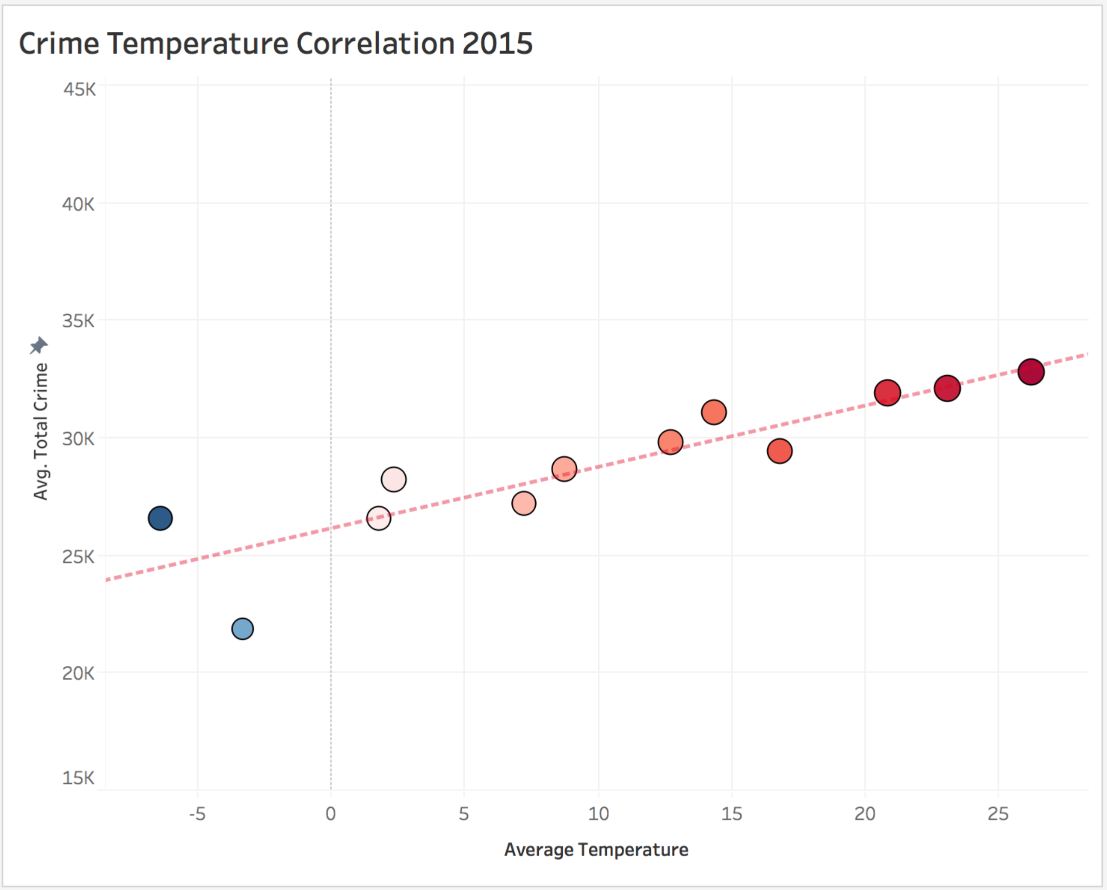
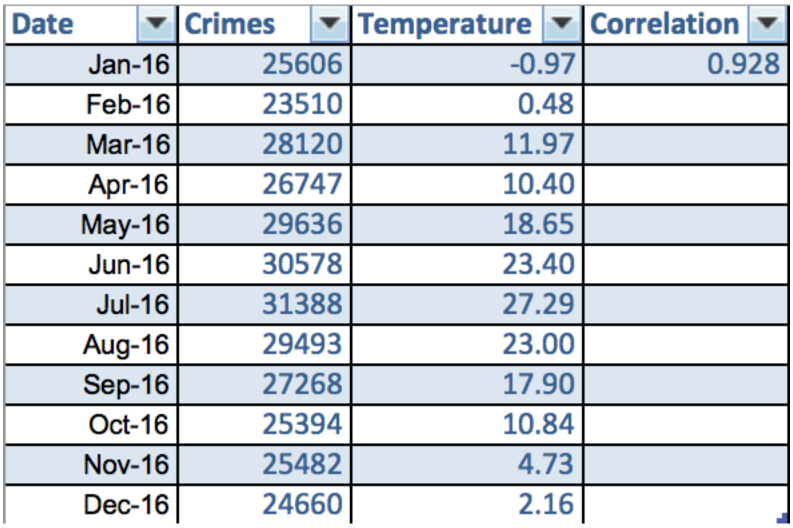


Fig 23: Correlation for 2015



Date	Crimes	Temperature	Correlation
Jan-16	25606	-0.97	0.928
Feb-16	23510	0.48	
Mar-16	28120	11.97	
Apr-16	26747	10.40	
May-16	29636	18.65	
Jun-16	30578	23.40	
Jul-16	31388	27.29	
Aug-16	29493	23.00	
Sep-16	27268	17.90	
Oct-16	25394	10.84	
Nov-16	25482	4.73	
Dec-16	24660	2.16	

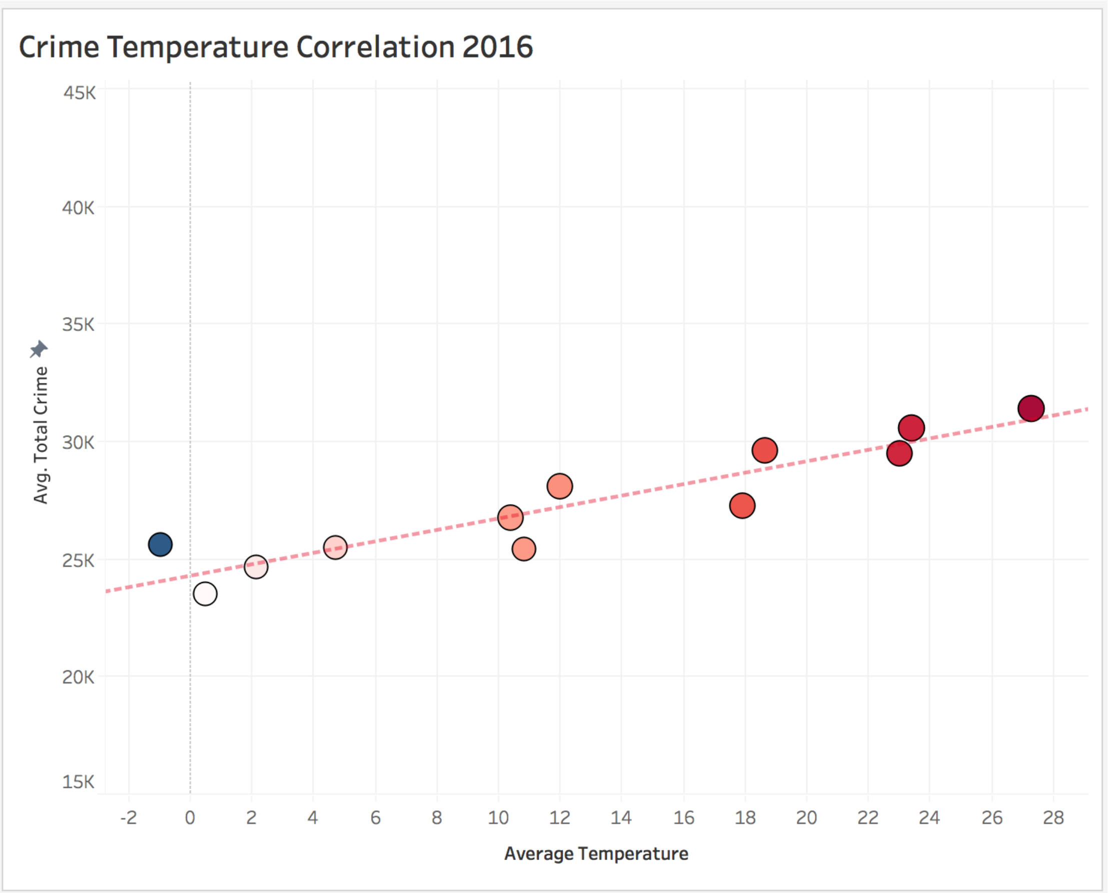
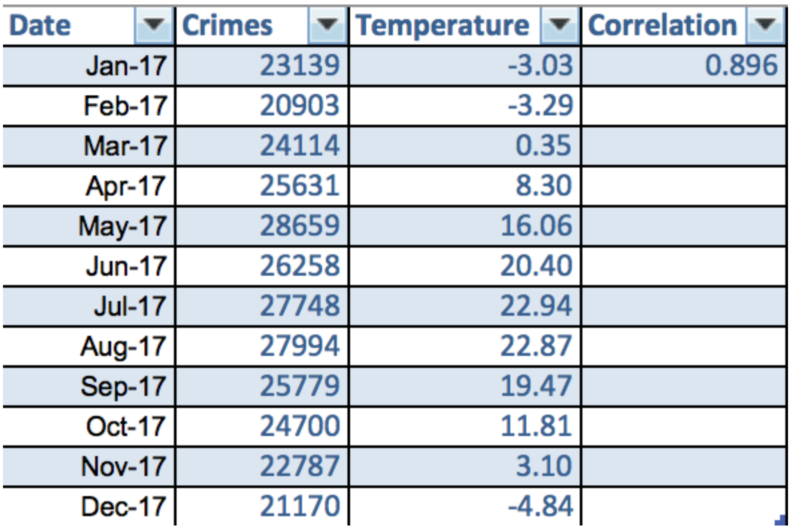


Fig 24: Correlation for 2016



Date	Crimes	Temperature	Correlation
Jan-17	23139	-3.03	0.896
Feb-17	20903	-3.29	
Mar-17	24114	0.35	
Apr-17	25631	8.30	
May-17	28659	16.06	
Jun-17	26258	20.40	
Jul-17	27748	22.94	
Aug-17	27994	22.87	
Sep-17	25779	19.47	
Oct-17	24700	11.81	
Nov-17	22787	3.10	
Dec-17	21170	-4.84	

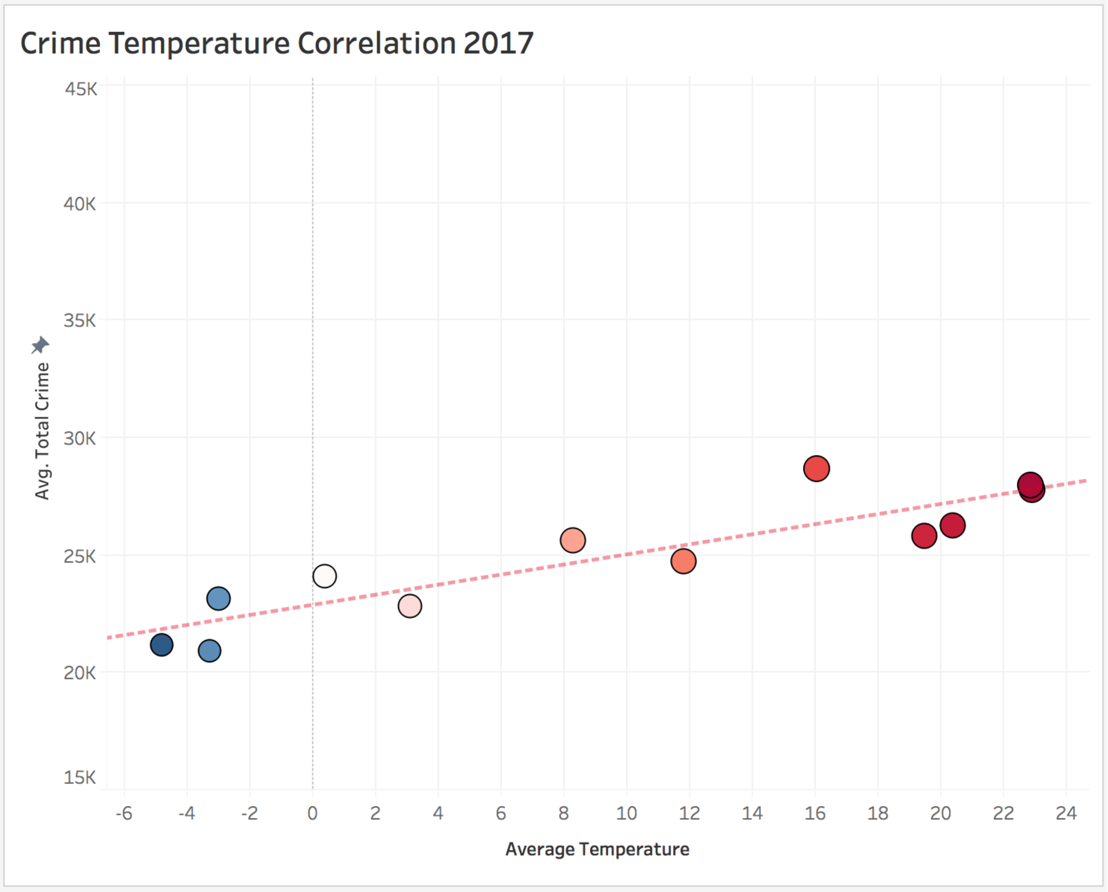


Fig 25: Correlation for 2017

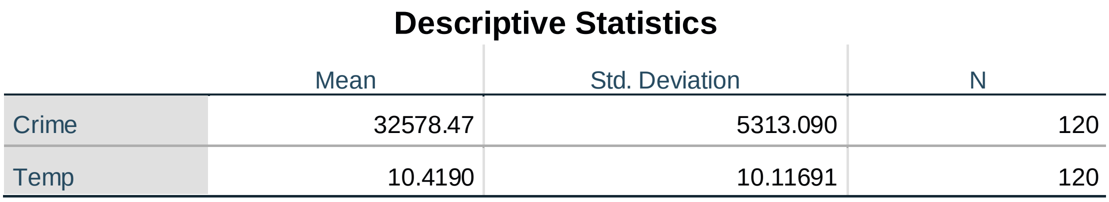
As demonstrated by the above series of scatter plots and correlation tests, each year demonstrates a strong positive correlation between crime and temperature. The scatter plots are observed to be very close to the trend line in all cases and the correlation coefficient values range from ***0.869***: ***0.929*** which is an extremely high correlation statistic.

Table 1: Correlation coefficient for 2008 - 2017

2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
0.928	*0.917*	*0.870*	*0.929*	*0.927*	*0.883*	*0.869*	*0.889*	*0.928*	*0.896*

Table 1 presents the obtained correlation values for each individual year. The results of these correlation tests confirm that weather conditions and crime have a significant positive relationship when observed over an annual period. Further correlation tests were carried out to examine the relationship of individual crimes over the ten-year period of this study.

Table 2: Descriptive Statistics 2008 - 2017

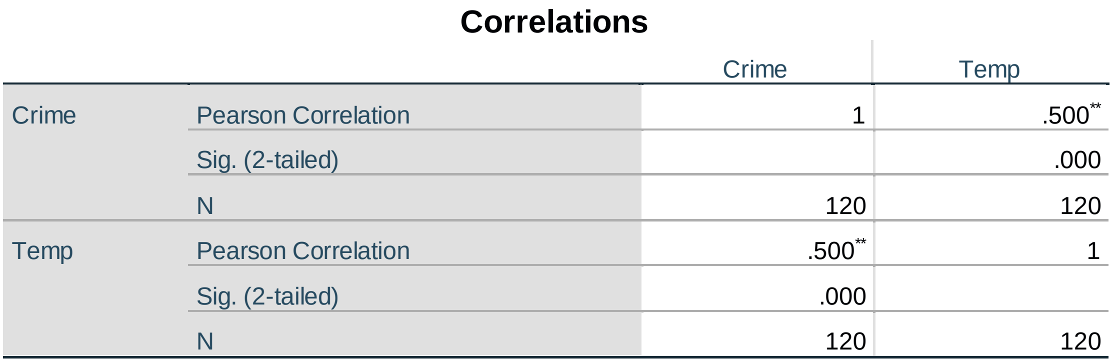


Descriptive Statistics

	Mean	Std. Deviation	N
Crime	32578.47	5313.090	120
Temp	10.4190	10.11691	120

Table 2 demonstrates the descriptive statistics for the sample range used to determine the correlations. The sample size is 120 as there are monthly totals for each month from 2008 - 2017

Table 3: Correlations

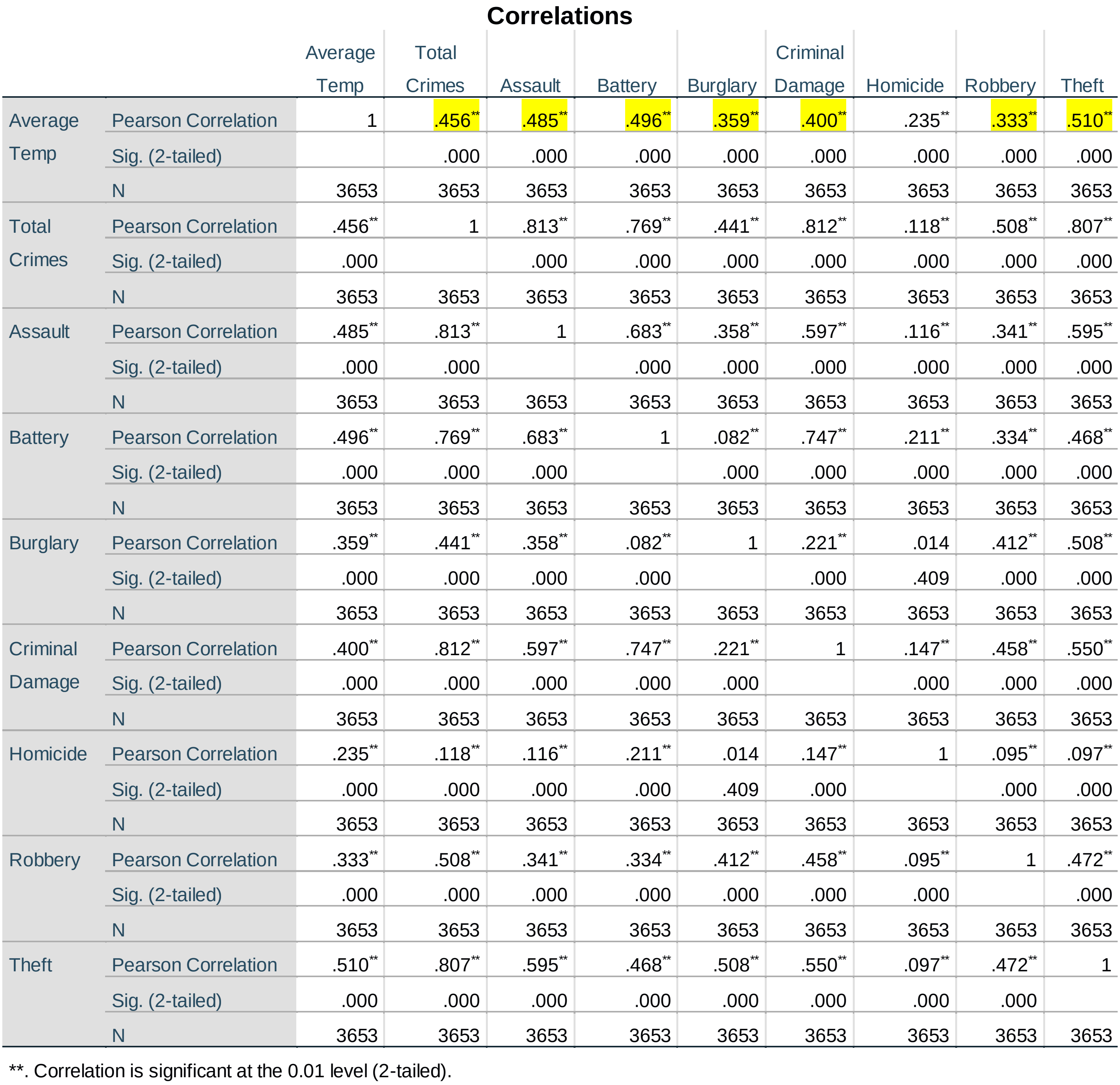


Correlations

		Crime	Temp
Crime	Pearson Correlation	1	.500**
	Sig. (2-tailed)		.000
	N	120	120
Temp	Pearson Correlation	.500**	1
	Sig. (2-tailed)	.000	
	N	120	120

Table 3 shows the correlation matrix for crime and temperature over the ten-year period. The Pearson correlation value obtained for the test is ***r* = .500**. This demonstrates that a statistically significant correlation between the two variables can still be observed over the ten years. A further analysis of the correlation of individual crimes over this period was conducted to identify any particular crimes of a high correlation.

Table 4 Correlation matrix for crimes 2008 - 2017



Correlations

		Average Temp	Total Crimes	Assault	Battery	Burglary	Criminal Damage	Homicide	Robbery	Theft
Average Temp	Pearson Correlation	1	.456**	.485**	.496**	.359**	.400**	.235**	.333**	.510**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Total Crimes	Pearson Correlation	.456**	1	.813**	.769**	.441**	.812**	.118**	.508**	.807**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Assault	Pearson Correlation	.485**	.813**	1	.683**	.358**	.597**	.116**	.341**	.595**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Battery	Pearson Correlation	.496**	.769**	.683**	1	.082**	.747**	.211**	.334**	.468**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Burglary	Pearson Correlation	.359**	.441**	.358**	.082**	1	.221**	.014	.412**	.508**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.409	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Criminal Damage	Pearson Correlation	.400**	.812**	.597**	.747**	.221**	1	.147**	.458**	.550**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Homicide	Pearson Correlation	.235**	.118**	.116**	.211**	.014	.147**	1	.095**	.097**
	Sig. (2-tailed)	.000	.000	.000	.000	.409	.000		.000	.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Robbery	Pearson Correlation	.333**	.508**	.341**	.334**	.412**	.458**	.095**	1	.472**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653
Theft	Pearson Correlation	.510**	.807**	.595**	.468**	.508**	.550**	.097**	.472**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	
	N	3653	3653	3653	3653	3653	3653	3653	3653	3653

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4 shows the correlation matrix of individual crimes correlated with temperature over the ten years. As demonstrated, all the selected categories of aggressive and violent crime have a statistically significant correlation with temperature. The findings of these correlation tests are of great significance to the development of the forecasting model, as we can now demonstrate an observable relationship between the two variables.

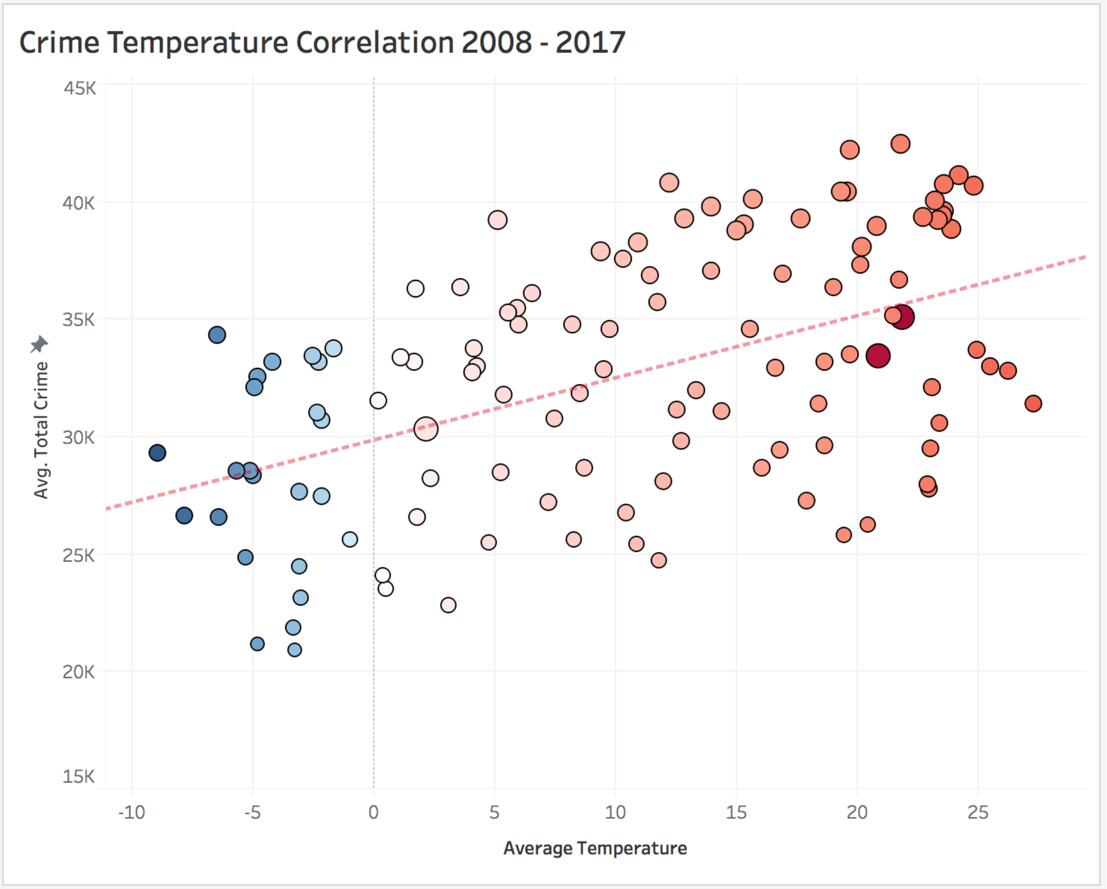
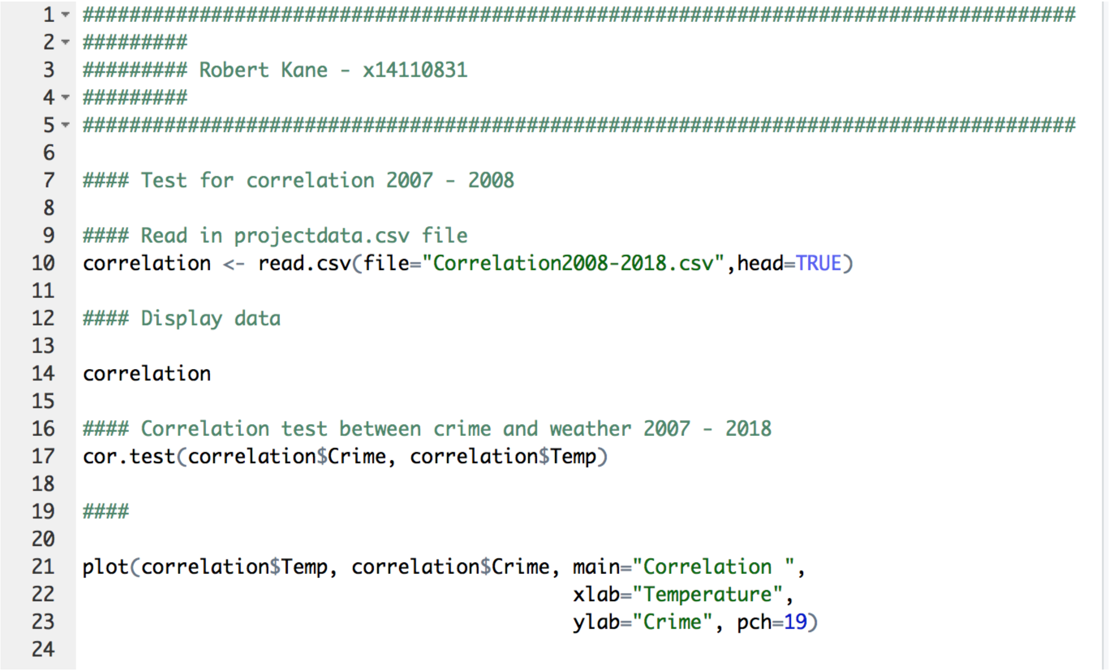


Fig 26: Correlation of crime and weather from 2008 – 2017

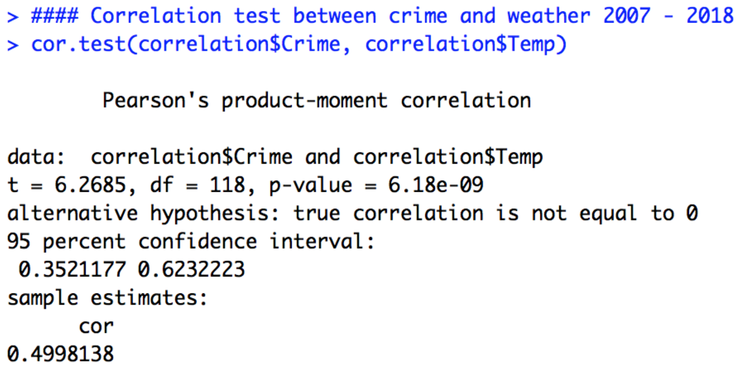
Fig 26 shows the scatter plot of crime and temperature from 2008 – 2017. It is clear from observing the scatter plot that incidents of crime increase as the temperature does. To validate the results of the correlation tests and observations made in the scatter plots, a correlation test was conducted in R Studio to validate the results.



```r
#####################################################################################
#########
######### Robert Kane - x14110831
#########
#####################################################################################

#### Test for correlation 2007 - 2008

#### Read in projectdata.csv file
correlation <- read.csv(file="Correlation2008-2018.csv",head=TRUE)

#### Display data

correlation

#### Correlation test between crime and weather 2007 - 2018
cor.test(correlation$Crime, correlation$Temp)

####

plot(correlation$Temp, correlation$Crime, main="Correlation ",
                                          xlab="Temperature",
                                          ylab="Crime", pch=19)

```

Fig 27: Correlation test carried out on R Studio

Fig 27 displays the R code for the correlation test to be conducted in R Studio. The results of the test are shown below.



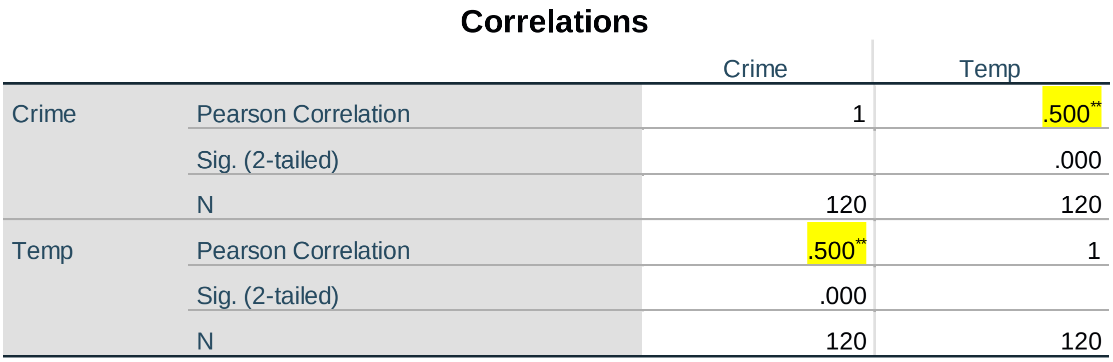
```
> #### Correlation test between crime and weather 2007 - 2018
> cor.test(correlation$Crime, correlation$Temp)

        Pearson's product-moment correlation

data:  correlation$Crime and correlation$Temp
t = 6.2685, df = 118, p-value = 6.18e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3521177 0.6232223
sample estimates:
      cor
0.4998138
```

Fig 28: Correlation test results from R Studio.

As displayed in Fig 28, the output of the correlation test in R is consistent with the decision to reject the null hypothesis (r = 0.499, t = 6.268, with 118 degrees of freedom, and a p-value = 6.18e-09). As temperature increases, so does crime. The 95% confidence interval for the correlation of weather and crime is (0.352, 0.623). As the confidence interval does not contain a value of 0, it is consistent with the decision to reject the null hypothesis for this research.

6.1.3. Conclusion

The results of the correlation test confirm that a strong positive correlation between weather and crime can be observed. It can be concluded that as temperature increases, so do the incidents of crime.



Correlations

		Crime	Temp
Crime	Pearson Correlation	1	.500**
	Sig. (2-tailed)		.000
	N	120	120
Temp	Pearson Correlation	.500**	1
	Sig. (2-tailed)	.000	
	N	120	120

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
0.928	*0.917*	*0.870*	*0.929*	*0.927*	*0.883*	*0.869*	*0.889*	*0.928*	*0.896*

H_A: ρ ≠ 0: There is a correlation between weather and crime

(r = 0.499, t = 6.268, with 118 degrees of freedom, and a p-value = 6.18e-09).

Table 5: Correlation strength interpretation

Size of Correlation	Interpretation
.90 to 1.00	**Very high positive**
.70 to .90	High positive
.50 to .70	**Moderate positive**
.30 to .50	Low positive
.00 to .30	Negligible correlation

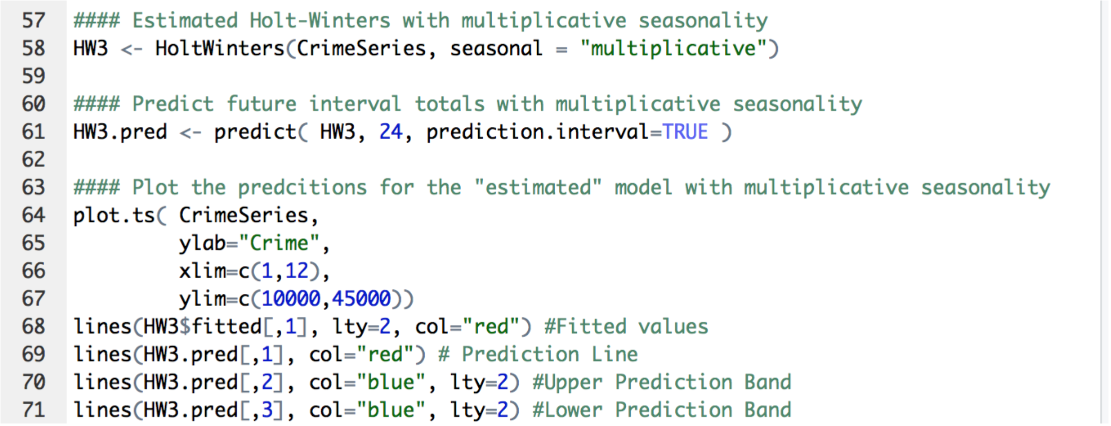
It is clear from the range of annual correlation values that weather conditions and temperature have a very high positive correlation. The results of these tests are significant when considering the development of an accurate forecasting model. Now that we have established and proven a statistically significant correlation between the two variables, the next section of the document will present the results of the forecasting method implemented.

6.2. Holt-Winters Method

6.2.1. Introduction

As previously discussed in the method section of the document, the results of the plotted time series of crime and weather for this research show a seasonality increase factor of 1.18 to 1.25. When paired with the positive results of the correlation test, this indicated an accurate model of prediction could be developed. A Holt-Winters triple exponential smoothing model was developed in Excel and R Studio to determine the required smoothing parameter values and to forecast predicted future values. The results of the model are displayed below.

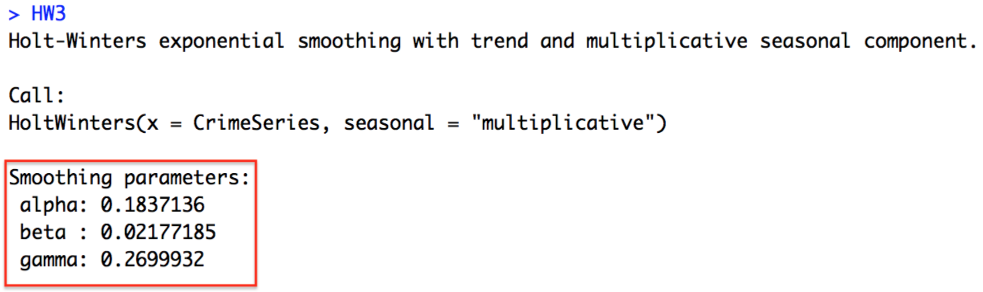
6.2.2. Results



```
#### Estimated Holt-Winters with multiplicative seasonality
HW3 <- HoltWinters(CrimeSeries, seasonal = "multiplicative")

#### Predict future interval totals with multiplicative seasonality
HW3.pred <- predict( HW3, 24, prediction.interval=TRUE )

#### Plot the predcitions for the "estimated" model with multiplicative seasonality
plot.ts( CrimeSeries,
         ylab="Crime",
         xlim=c(1,12),
         ylim=c(10000,45000))
lines(HW3$fitted[,1], lty=2, col="red") #Fitted values
lines(HW3.pred[,1], col="red") # Prediction Line
lines(HW3.pred[,2], col="blue", lty=2) #Upper Prediction Band
lines(HW3.pred[,3], col="blue", lty=2) #Lower Prediction Band
```

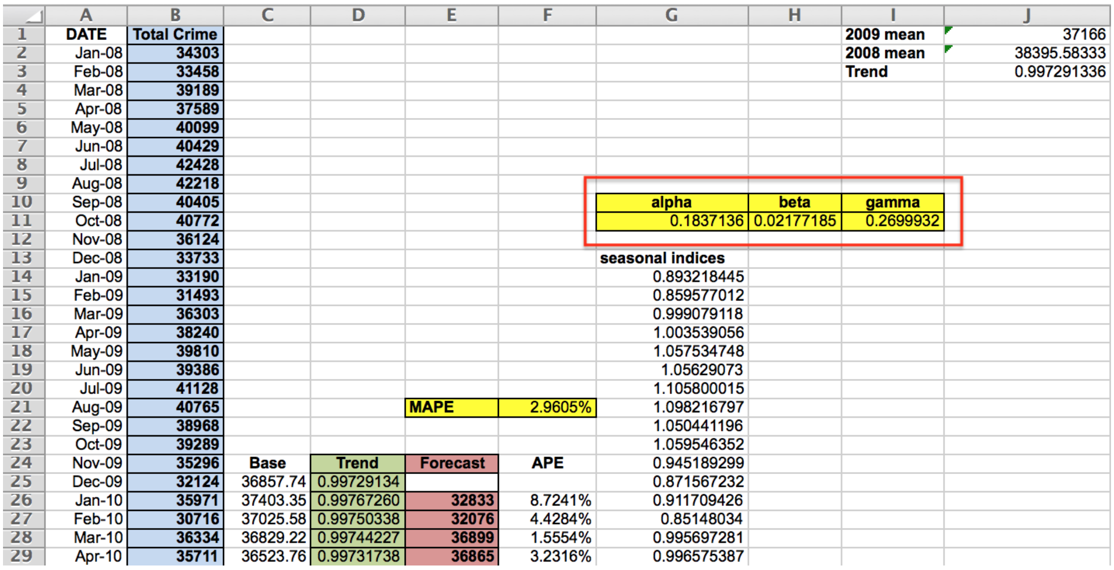
Fig 28: Holt Winters multiplicative seasonality R code

Fig 28 shows the R code used to develop the Holt-Winters forecasting model. The alpha, beta and gamma smoothing parameter values determined by R shown in Fig 29 were taken and applied to the development of the model in Excel.



```
> HW3
Holt-Winters exponential smoothing with trend and multiplicative seasonal component.

Call:
HoltWinters(x = CrimeSeries, seasonal = "multiplicative")

Smoothing parameters:
 alpha: 0.1837136
 beta : 0.02177185
 gamma: 0.2699932
```

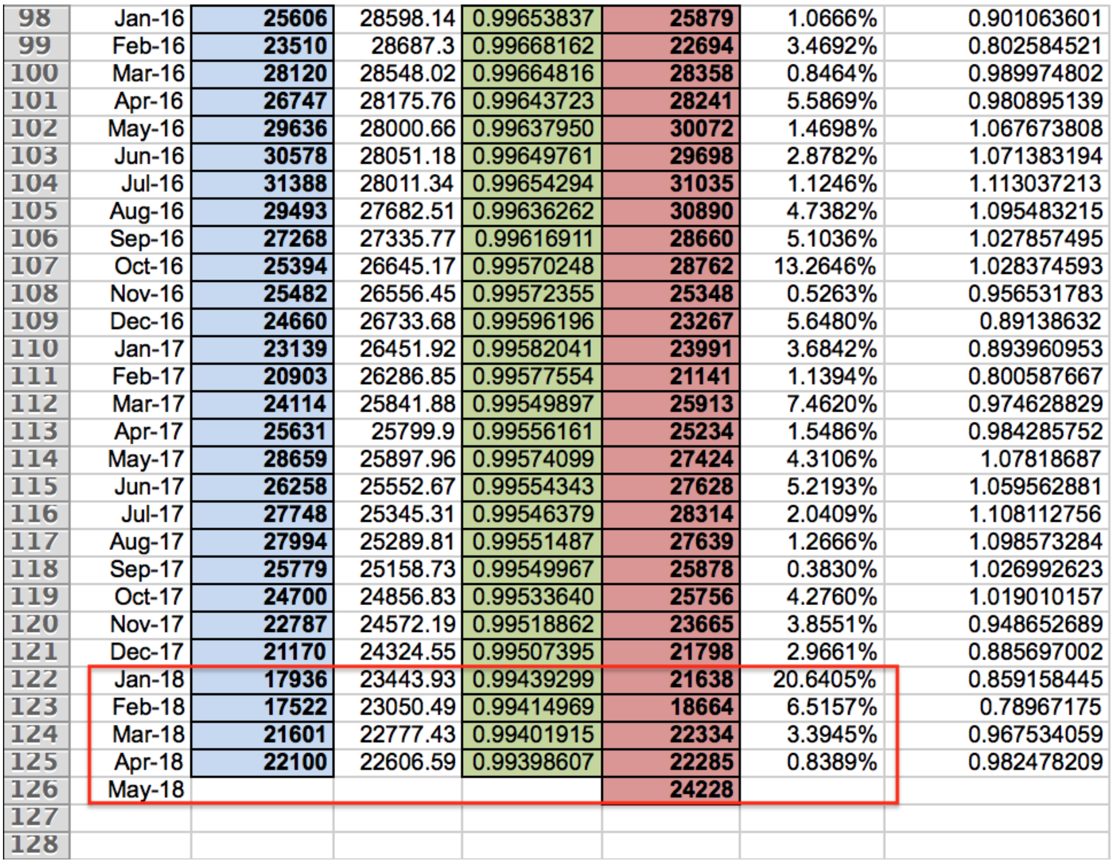
Fig 29: Smoothing parameters determined by R



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					0.893218445			
15	Feb-09	31493					0.859577012			
16	Mar-09	36303					0.999079118			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	37403.35	0.99767260	32833	8.7241%	0.911709426			
27	Feb-10	30716	37025.58	0.99750338	32076	4.4284%	0.85148034			
28	Mar-10	36334	36829.22	0.99744227	36899	1.5554%	0.995697281			
29	Apr-10	35711	36523.76	0.99731738	36865	3.2316%	0.996575387			

Fig 30: Smoothing parameters determined by R taken into Excel

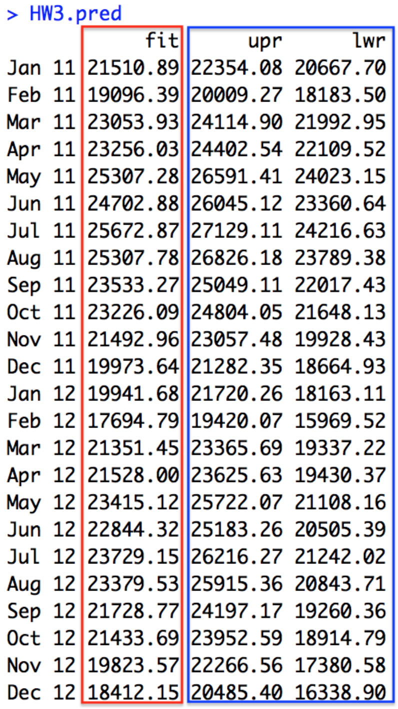
The completed model produced the following results:



98	Jan-16	25606	28598.14	0.99653837	25879	1.0666%	0.901063601
99	Feb-16	23510	28687.3	0.99668162	22694	3.4692%	0.802584521
100	Mar-16	28120	28548.02	0.99664816	28358	0.8464%	0.989974802
101	Apr-16	26747	28175.76	0.99643723	28241	5.5869%	0.980895139
102	May-16	29636	28000.66	0.99637950	30072	1.4698%	1.067673808
103	Jun-16	30578	28051.18	0.99649761	29698	2.8782%	1.071383194
104	Jul-16	31388	28011.34	0.99654294	31035	1.1246%	1.113037213
105	Aug-16	29493	27682.51	0.99636262	30890	4.7382%	1.095483215
106	Sep-16	27268	27335.77	0.99616911	28660	5.1036%	1.027857495
107	Oct-16	25394	26645.17	0.99570248	28762	13.2646%	1.028374593
108	Nov-16	25482	26556.45	0.99572355	25348	0.5263%	0.956531783
109	Dec-16	24660	26733.68	0.99596196	23267	5.6480%	0.89138632
110	Jan-17	23139	26451.92	0.99582041	23991	3.6842%	0.893960953
111	Feb-17	20903	26286.85	0.99577554	21141	1.1394%	0.800587667
112	Mar-17	24114	25841.88	0.99549897	25913	7.4620%	0.974628829
113	Apr-17	25631	25799.9	0.99556161	25234	1.5486%	0.984285752
114	May-17	28659	25897.96	0.99574099	27424	4.3106%	1.07818687
115	Jun-17	26258	25552.67	0.99554343	27628	5.2193%	1.059562881
116	Jul-17	27748	25345.31	0.99546379	28314	2.0409%	1.108112756
117	Aug-17	27994	25289.81	0.99551487	27639	1.2666%	1.098573284
118	Sep-17	25779	25158.73	0.99549967	25878	0.3830%	1.026992623
119	Oct-17	24700	24856.83	0.99533640	25756	4.2760%	1.019010157
120	Nov-17	22787	24572.19	0.99518862	23665	3.8551%	0.948652689
121	Dec-17	21170	24324.55	0.99507395	21798	2.9661%	0.885697002
122	Jan-18	17936	23443.93	0.99439299	21638	20.6405%	0.859158445
123	Feb-18	17522	23050.49	0.99414969	18664	6.5157%	0.78967175
124	Mar-18	21601	22777.43	0.99401915	22334	3.3945%	0.967534059
125	Apr-18	22100	22606.59	0.99398607	22285	0.8389%	0.982478209
126	May-18				24228		
127							
128							

Fig 31: Forecast values for January to May 2018.

The actual and predicted crime values produced by the forecast model are shown in Fig 31. The forecasts have been projected to May 2018 here in Excel and a forecast of 2 years was conducted in R Studio.



```
> HW3.pred
            fit      upr      lwr
Jan 11 21510.89 22354.08 20667.70
Feb 11 19096.39 20009.27 18183.50
Mar 11 23053.93 24114.90 21992.95
Apr 11 23256.03 24402.54 22109.52
May 11 25307.28 26591.41 24023.15
Jun 11 24702.88 26045.12 23360.64
Jul 11 25672.87 27129.11 24216.63
Aug 11 25307.78 26826.18 23789.38
Sep 11 23533.27 25049.11 22017.43
Oct 11 23226.09 24804.05 21648.13
Nov 11 21492.96 23057.48 19928.43
Dec 11 19973.64 21282.35 18664.93
Jan 12 19941.68 21720.26 18163.11
Feb 12 17694.79 19420.07 15969.52
Mar 12 21351.45 23365.69 19337.22
Apr 12 21528.00 23625.63 19430.37
May 12 23415.12 25722.07 21108.16
Jun 12 22844.32 25183.26 20505.39
Jul 12 23729.15 26216.27 21242.02
Aug 12 23379.53 25915.36 20843.71
Sep 12 21728.77 24197.17 19260.36
Oct 12 21433.69 23952.59 18914.79
Nov 12 19823.57 22266.56 17380.58
Dec 12 18412.15 20485.40 16338.90
```

Fig 32: R Code prediction for 24 periods (2 year forecast)

The results of the fitted values with upper and lower band predictions are shown in Fig 32. These are the forecasted crime totals for the years 2018 – 2019. To visualise how these forecasts look in the time series, the values were plotted in R Studio.

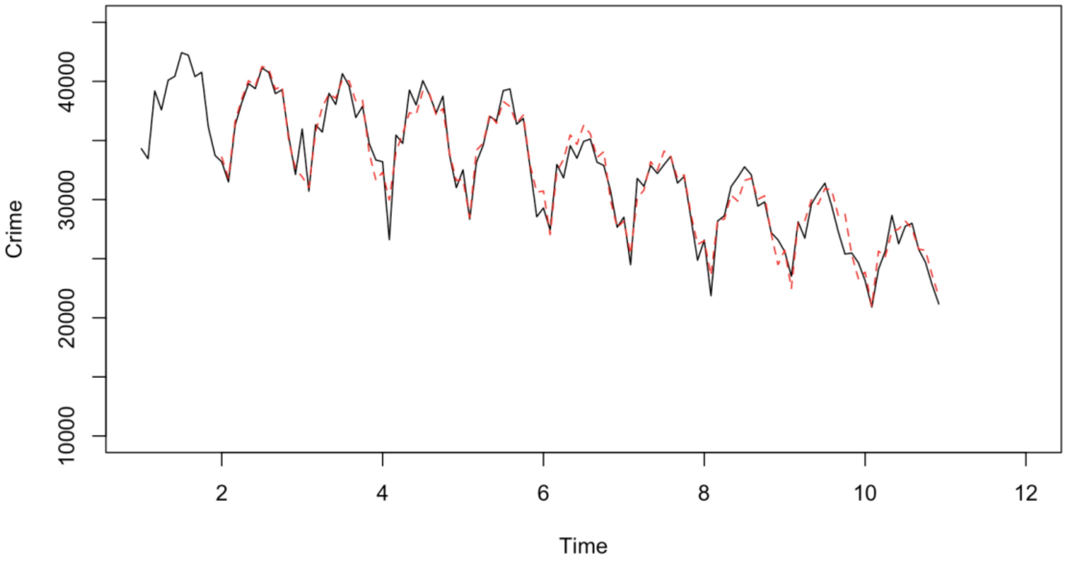


Fig 33: R Code displaying the fitted value forecast estimates

The fitted forecast estimates shown in Fig 33 show the resulting trend analysis for the years 2010 – 2017 having generated the seasonal trend of the time series using the years 2008 and 2009. The time axis represents the number of periods for the test, in this case there are 12 periods as the years 2008 – 2019 are being analysed and forecasted. The results of the forecasted values for 2018 and 2019 with the upper and lower prediction bands are shown below in Fig 34.

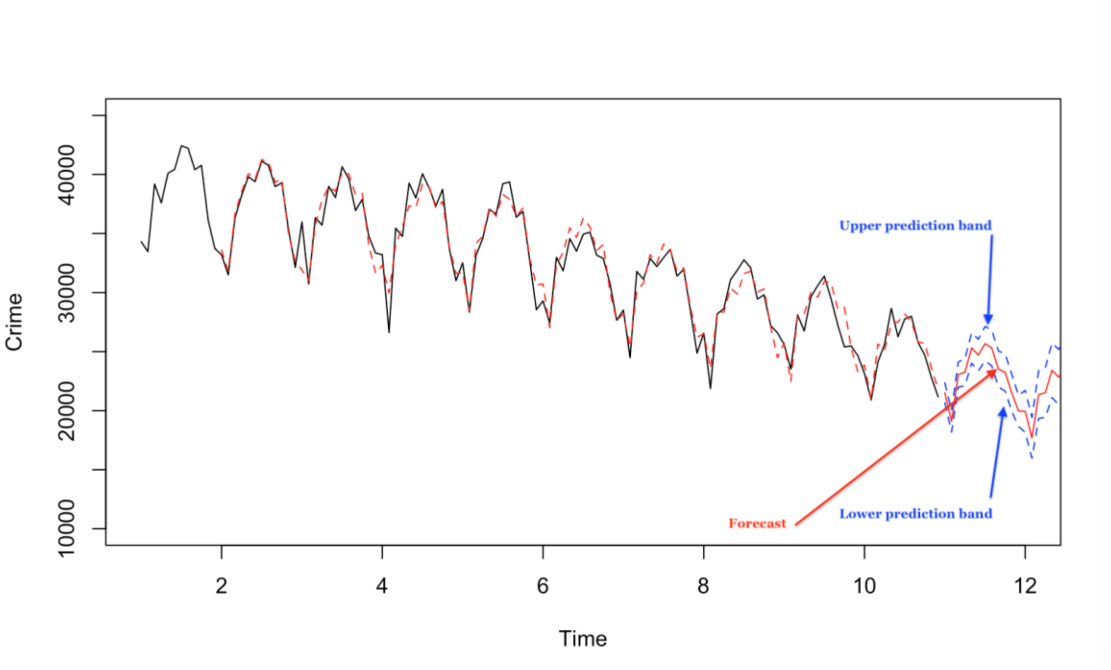
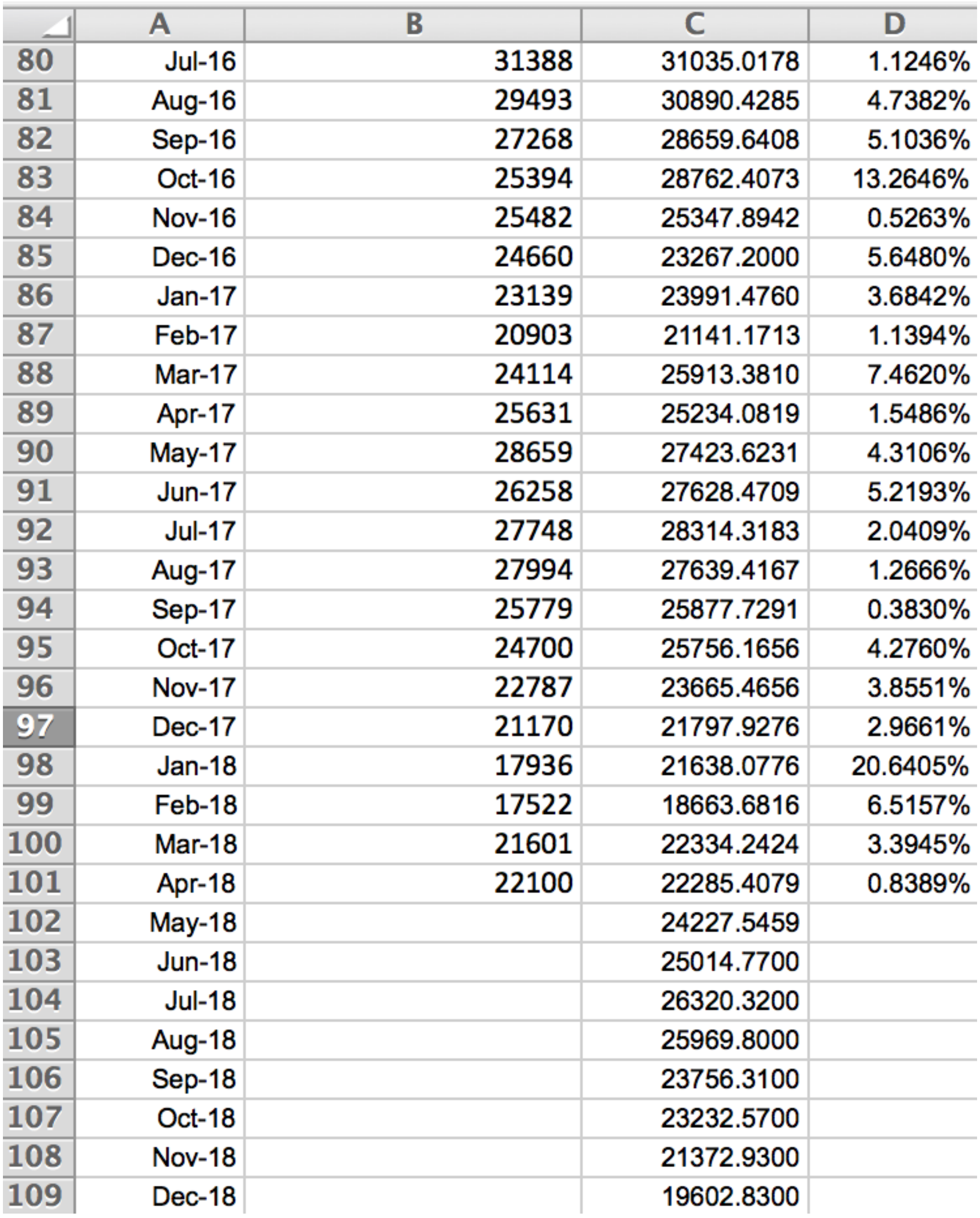


Fig 34: Prediction forecasts for 2018 and 2019

The forecast values were taken from R and compared with the subsequent actual values for that period. These values were obtained from the original data source after each month in 2018 had passed. The forecasts were measured for accuracy against the actual values to determine the mean absolute percentage error of the forecasting model. The results of the forecasts are shown in Table 6.

Table 6: Forecasted crime predictions for 2018

	January	February	March	April
Forecast	21638	18664	22334	22285
Actual	17936	17522	21601	22100
MAPE	20.60	6.51	3.39	0.83



	A	B	C	D
80	Jul-16	31388	31035.0178	1.1246%
81	Aug-16	29493	30890.4285	4.7382%
82	Sep-16	27268	28659.6408	5.1036%
83	Oct-16	25394	28762.4073	13.2646%
84	Nov-16	25482	25347.8942	0.5263%
85	Dec-16	24660	23267.2000	5.6480%
86	Jan-17	23139	23991.4760	3.6842%
87	Feb-17	20903	21141.1713	1.1394%
88	Mar-17	24114	25913.3810	7.4620%
89	Apr-17	25631	25234.0819	1.5486%
90	May-17	28659	27423.6231	4.3106%
91	Jun-17	26258	27628.4709	5.2193%
92	Jul-17	27748	28314.3183	2.0409%
93	Aug-17	27994	27639.4167	1.2666%
94	Sep-17	25779	25877.7291	0.3830%
95	Oct-17	24700	25756.1656	4.2760%
96	Nov-17	22787	23665.4656	3.8551%
97	Dec-17	21170	21797.9276	2.9661%
98	Jan-18	17936	21638.0776	20.6405%
99	Feb-18	17522	18663.6816	6.5157%
100	Mar-18	21601	22334.2424	3.3945%
101	Apr-18	22100	22285.4079	0.8389%
102	May-18		24227.5459	
103	Jun-18		25014.7700	
104	Jul-18		26320.3200	
105	Aug-18		25969.8000	
106	Sep-18		23756.3100	
107	Oct-18		23232.5700	
108	Nov-18		21372.9300	
109	Dec-18		19602.8300	

Fig 35 Forecasted crime values taken from R studio model

The forecasted values obtained in R were recorded alongside the actual values. At the end of each passing month throughout the lifecycle of the project, crime statistics were gathered and evaluated against the predicted forecasts of the model. These results were then documented and imported into Tableau for the development of the required visualisations. The result of the developed model is displayed below.

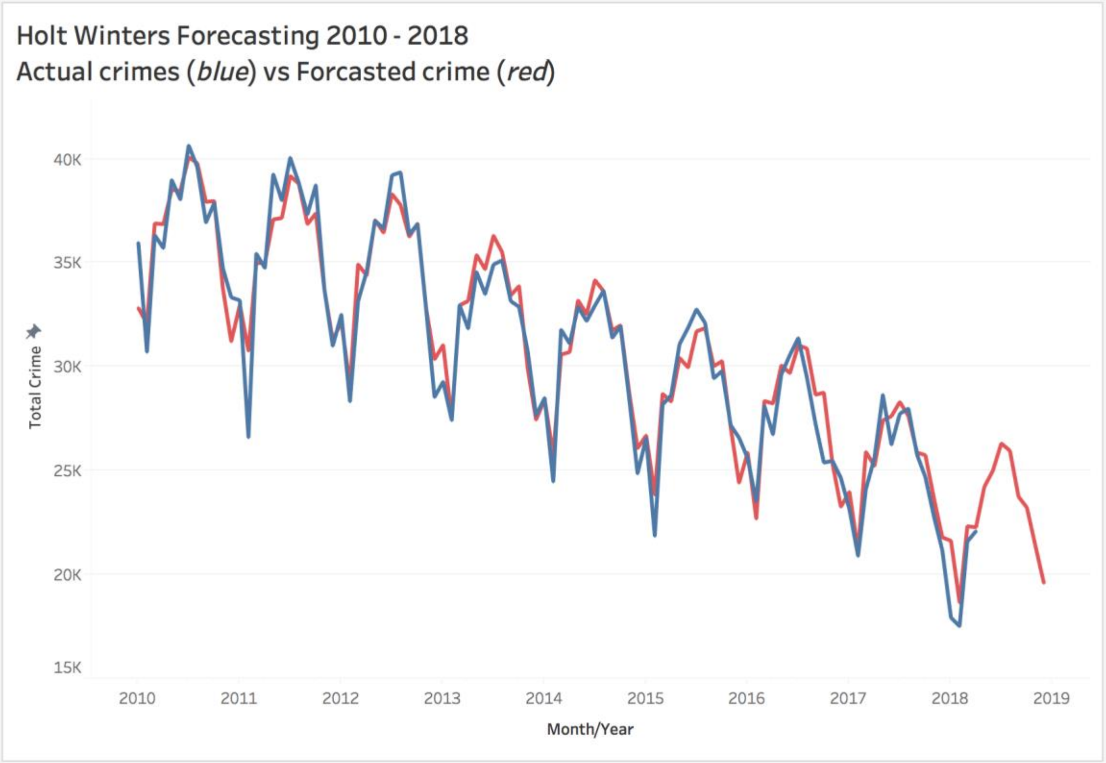
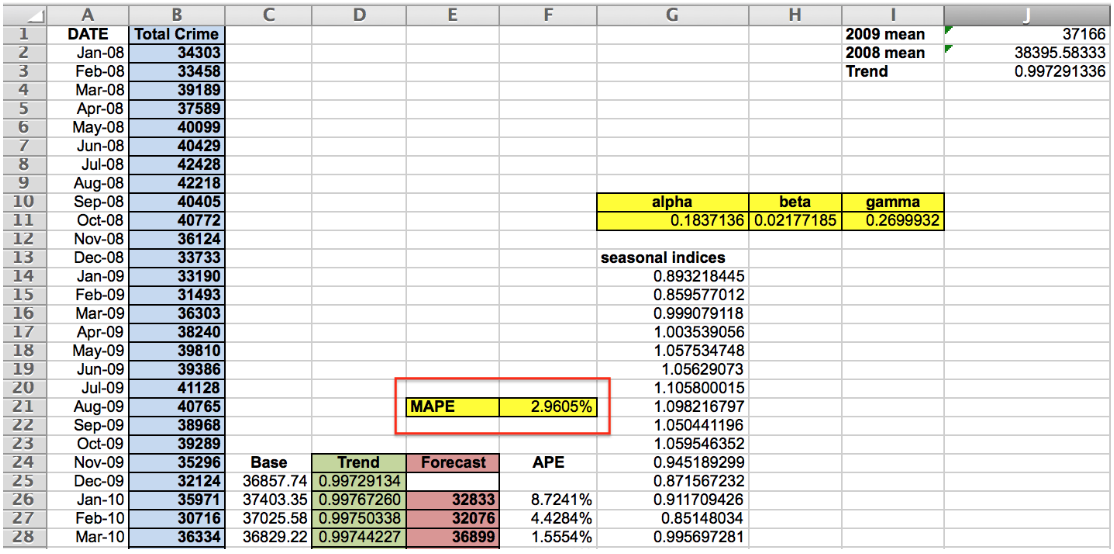


Fig 36: Holt Winters Forecasting 2010 – 2018.

Fig 36 shows the developed forecast model in Tableau. A high degree of accuracy is evident by the close proximity of the forecast value shown in red to that of the actual observed values shown in blue. The model operates at a high level of accuracy producing a mean absolute percentage error of just 2.96% over the time series.



	A	B	C	D	E	F	G	H	I	J
1	DATE	Total Crime							2009 mean	37166
2	Jan-08	34303							2008 mean	38395.58333
3	Feb-08	33458							Trend	0.997291336
4	Mar-08	39189								
5	Apr-08	37589								
6	May-08	40099								
7	Jun-08	40429								
8	Jul-08	42428								
9	Aug-08	42218								
10	Sep-08	40405					alpha	beta	gamma	
11	Oct-08	40772					0.1837136	0.02177185	0.2699932	
12	Nov-08	36124								
13	Dec-08	33733					seasonal indices			
14	Jan-09	33190					0.893218445			
15	Feb-09	31493					0.859577012			
16	Mar-09	36303					0.999079118			
17	Apr-09	38240					1.003539056			
18	May-09	39810					1.057534748			
19	Jun-09	39386					1.05629073			
20	Jul-09	41128					1.105800015			
21	Aug-09	40765			MAPE	2.9605%	1.098216797			
22	Sep-09	38968					1.050441196			
23	Oct-09	39289					1.059546352			
24	Nov-09	35296	Base	Trend	Forecast	APE	0.945189299			
25	Dec-09	32124	36857.74	0.99729134			0.871567232			
26	Jan-10	35971	37403.35	0.99767260	32833	8.7241%	0.911709426			
27	Feb-10	30716	37025.58	0.99750338	32076	4.4284%	0.85148034			
28	Mar-10	36334	36829.22	0.99744227	36899	1.5554%	0.995697281			

Fig 37: Mean absolute percentage error 2.96%

6.2.3. MAPE (Mean Absolute Percentage Error)

The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method. When calculated for the crime and weather statistics for this research, the model performs to an MAPE value of 2.96%. Therefore, the forecasting model performs with over 97% accuracy in the forecast predictions of crime. This is an incredibly high level of prediction accuracy for any forecasting model to perform at. Visual representations of the performance metrics are presented below.

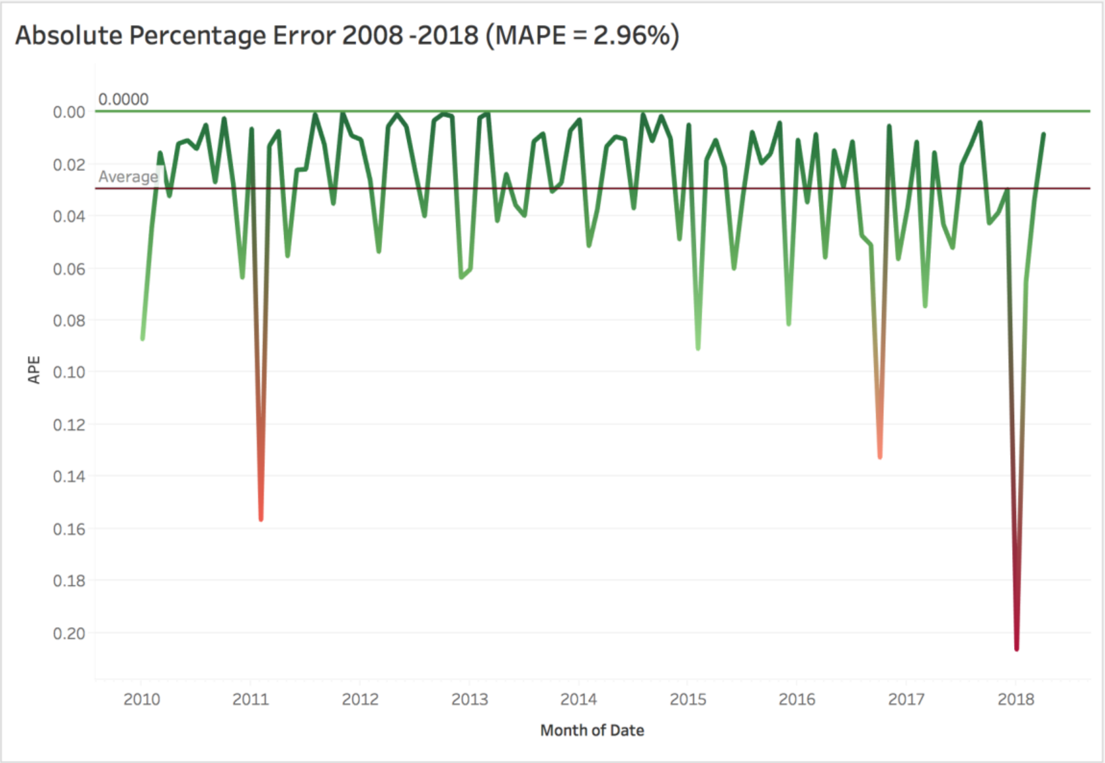


Fig 38: MAPE values plotted.

Fig 38 shows a visual representation of the plotted MAPE values. Optimum performance of the forecasting model is indicated by an MAPE value as close to 0% as possible. As the graph shows, the performance of the forecasting model is extremely accurate with several points reaching a mean absolute percentage error of less than 1%. The overall average MAPE value for the years 2010 – 2018 is calculated at 2.96%. This indicates a level of confidence greater than 97% can be placed in the accuracy of predictions produced by this forecasting model.

6.2.4. Conclusion

The aim of this research was to investigate if a statistically significant correlation between incidents of crime and weather conditions could be identified and used to develop an accurate forecasting model for the prediction of crime. The results of the Holt-Winters forecasting model demonstrate convincing results that this aim has been successfully achieved and implemented. The ability to predict future levels of crime to an accuracy greater than 97% would be of great significance and interest to those who have a practical application or need for the findings such as law enforcement agencies and emergency services.

Assessing the results and performance of the model, there are some evident outliers over the time series. These outliers could be caused by unseasonable or extreme weather conditions that cause the level of crimes to fall short or extend above the forecasted prediction for that period. In spite of this, the accuracy of the model still maintains an extremely high level of performance with a mean absolute percentage error of just 2.6%.

When considering the most recent forecast prediction for the completed month of April 2018, the model forecasted a prediction that came to pass with a MAPE of 0.83%. This means the forecasted prediction accuracy was higher than 99%. In light of the results produced and the performance level of the forecasting model, I am satisfied that my goal of identifying a method to develop a meaningful, accurate model to predict the occurrence of crime has been successfully achieved. The future prospects and implications of this successful model are discussed in the following sections of the document.

7. Discussion

The statistical analysis and research of the relationship between weather and crime is an extremely complex task that incorporates a multitude of both tangible and intangible factors that must be considered. These factors range from geographic location, environmental and spatial surroundings, political and social landscapes as well as the economic climate of the location being studied. Although these factors are complex and often present no quantifiable scale for measurement that allow us to gauge the direct impact they have on the occurrence of crime, one measurable factor remains an observable constant in the behavioural trend and pattern in the occurrence of crime – weather. The analysis of the relationship between weather and crime in Cleveland, Ohio conducted by Paul Butke and Scott. C Sheridan in 2010 concluded that despite numerous different spatial analyses performed, a minimal correlation between spatial patterns of crime counts and hotter weather conditions could be observed. The study demonstrated that increases in temperature resulted in similar percentage increases in crime citywide regardless of other possible influencing factors. This finding is particularly significant to the application of the findings presented in this research as it strengthens the assertion made by the correlation tests that increasing temperature directly influences the occurrence of crime. Based on this observable relationship, this research could successfully develop and implement a highly sophisticated and accurate method of forecasting.

Throughout the lifecycle of the project, literature reviews were conducted to determine existing scientific basis and findings already presented in this field of research. Many of these studies aim to quantify and measure the correlation of a number of influencing factors simultaneously, but few attempt to successfully implement a model of prediction that presents measurements of the results for analysis. I believe the strength and innovation of this research is attributed to the successful development of a forecasting model that presents measured results that show a high level of accuracy. Similar studies identified in the literature review have attempted a model for prediction, but these present limitations in the assumptions and prerequisites for the optimal performance and accuracy of the model. The approach taken in this research puts weight to the scientific observation that ambient temperature directly influences crime. The results of the correlation test confirm the existence of a strong positive relationship between the two variables when examined annually, and a prediction model demonstrating a performance accuracy of over 97% was developed implementing these insights. The significance here is that regardless of other influencing factors, the model demonstrates a high level of sophistication and accuracy for the prediction of crime.

The findings presented in this research have great potential to be applied in conjunction with the findings of similar studies on the effects of weather conditions on crime. A study on the effects of weather on Crime by Rodrigo Murataya and Daniel R. Gutierrez in 2010 concluded that the correlation does not seem to be affected by geographical location or population factors and that the results and outcomes of various tests are fairly similar and are equally observed throughout many different locations around the world. The significance of this in relation to discussing the outcomes of this research is that it implies the forecasting model should theoretically work when transposed and implemented for any number of regions.

With the utilisation of emerging IoT technologies and the increased interconnectivity these innovations afford us, the findings of this study could be applied to the development of a wide scale network. Law enforcement authorities and emergency services are constantly seeking reliable forecasting, detection and preventative strategies to assist in their methods of response. All information that leads to better-informed decisions is inherently valuable; therefore, the findings presented in this study demonstrate great potential to be implemented in these sectors to gain valuable insights for their proposed solutions.

8. Future Prospects and Commercial Aspects

The research aims and objectives of set out at the start of this project have been successfully achieved and implemented. The information and insights produced by the research has concluded with the successful development of a forecasting model that performs with over 97% accuracy. These results have great potential to be incorporated into similar studies and models of analysis.

The results presented in this research could potentially be expanded to accommodate and display statistical analysis for more locations in the future. As previously discussed, the correlation between weather and crime is observed to be a constant despite geographical location and other environmental factors. The conceptual blueprints created and mapped out in this project could be applied to more data sources and cities. The system has the potential to evolve into a concept where the methodologies implemented allow for a decision support system to be developed that focuses on various locations around the world to produce intelligence reports.

Data mining has become big business in recent years and information that enhances the accuracy of decision-making processes and resource allocation is inherently valuable. The results of this project could potentially be of great interest to those who have a practical application or need for the findings. Data is valuable and people are willing to pay for the solutions data analysis provides.

9. Evaluation and Testing

9.1. Evaluation

I believe I have been successful in implementing an innovative approach to this area of research as evidenced by the strength of the prediction model produced. The literature review for this project outlined a number of different approaches that focused varying assessment methods on different variables but none of the studies researched demonstrated a forecasting model case study with verification and validation of results. With the implementation of the project complete and the results obtained, the next step in the project was to conduct extensive testing. The following section will give details of the test approaches carried out as well as the documented evidence of the test results.

9.2. Testing

Verification and Validation tests were performed on all components and test statistics throughout the lifecycle of this project. Validation and accuracy checks were performed on the data, the functional components of the system and the results produced by the test. The results of the systems and validation testing are presented below.

9.2.1. Objectives

The objectives of the test plan are to validate and verify the results of the statistical tests and performance of the integrated components used to produce and visualise them for this project. The components are verified in terms of performance and accuracy to the requirements of the project.

9.2.2. Scope

The scope of the testing will extend to all functional components, requirements and outputs produced by the project. These include the data warehouse, source data, Tableau workbooks, statistical analysis and results produced.

9.2.3. Testing and Evaluation Results Tables

The following tables provide the details of the test carried out for this project:

Test Case 1			
Name	Data Warehouse Functionality testing		
Priority	1	**Date of Test**	1/03/18

Test ID	**1**
Purpose of Test	To Ensure that: It is possible to load data warehouse with required data dumps. It is possible to query the data to produce reports.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and MySQL Workbench through citrix. Test data was .csv files of gathered data.
Method	From MySQL Workbench the tester should: Ensure the data is properly loaded into the data warehouse. Run queries to validate the successful import.
Expected Result	On completing of the above steps, the data warehouse should be populated with data for weather and crime statistics
Actual result	The data was successfully imported and the data warehouse is functioning as expected.
Comments	N/A
Resolution	N/A

Test Case 2			
Name	R Studio Functionality Testing		
Priority	2	**Date of Test**	5/03/18

Test ID	**2**
Purpose of Test	To Ensure that: It is possible set the working directory in R Studio to locate and read in the required data sources for analysis It is possible to perform analysis on the data.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and R Studio. Test data was required for statistical tests.
Method	From R Studio the tester should: Ensure the data is properly loaded and run functions to ensure the directory is connected and functional
Expected Result	On completing of the above steps, R Studio should produce confirmation that the working directory is visible and the appropriate data has been imported for analysis
Actual result	The data was successfully imported and the R Studio is functioning as expected
Comments	N/A
Resolution	N/A

Test Case 3			
Name	Tableau Functionality Testing		
Priority	1	**Date of Test**	10/03/18

Test ID	**3**
Purpose of Test	To Ensure that: It is possible set the data source to locate and extract the required data sources for analysis It is possible to perform visualisations on the data.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and Tableau Desktop 10.5. Test data was required the for development of visualisations
Method	From Tableau the tester should: Ensure the data is properly loaded and run visualisation operations to ensure the data source is connected and functional
Expected Result	On completing of the above steps, Tableau should produce visualisations for plotted variables and the data source is connected and working appropriately.
Actual result	The data sources were successfully connected and extracted and Tableau is functioning as expected
Comments	N/A
Resolution	N/A

Test Case 4			
Name	SPSS Functionality Testing		
Priority	1	**Date of Test**	10/03/18

Test ID	**4**
Purpose of Test	To Ensure that: Data files are loaded into SPSS correctly and the required analysis is produced. It is possible to perform analysis on the data.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and accessing SPSS through Citrix. Test data was required data for statistical tests.
Method	From SPSS the tester should: Ensure the data is properly loaded and run data analysis test to ensure the data is read and functional
Expected Result	On completing of the above steps, SPSS should produce confirmation that the data has been imported and that analysis can be produced on the selected ranges.
Actual result	The data was successfully imported and the SPSS is functioning as expected
Comments	N/A
Resolution	N/A

Test Case 5			
Name	Data Cleansing and ETL process validation		
Priority	1	**Date of Test**	1/02/18

Test ID	**5**
Purpose of Test	To Ensure that: The acquired data for the project is correct and formatted correctly for consumption of various components. That the data is structured correctly.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and Microsoft Excel. Source data was required for the test.
Method	From Excel the tester should: Ensure the data is properly loaded and run functions to ensure the data is correct.
Expected Result	On completing of the above steps, Excel should produce confirmation that the data has been properly cleansed and structured for use in the analysis
Actual result	The data was successfully structured and Excel is functioning as expected
Comments	N/A
Resolution	N/A

Test Case 6			
Name	Test Result Validation 1		
Priority	1	**Date of Test**	15/04/18

Test ID	**6**
Purpose of Test	To Ensure that: The obtained test statistics for the analysis are valid and accurate
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and R Studio, Microsoft Excel and SPSS Statistics Test data was required for statistical tests.
Method	From R Studio the tester should: Ensure the data is properly loaded and run tests to verify the results produced by Excel and SPSS
Expected Result	On completing of the above steps, R Studio should produce confirmation that the obtained test statistic is valid.
Actual result	The test statistic obtained matches the results of the other analysis methods.
Comments	N/A
Resolution	N/A

Test Case 7			
Name	Tableau Visualisations		
Priority	1	**Date of Test**	17/04/18

Test ID	**7**
Purpose of Test	To Ensure that: Visualisations produced in Tableau are saved, the data source has been extracted and the workbook has been saved to Tableau Public for end user interpretation.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and Tableau Desktop 10.5 Test data was required data for visualisations and a hosted Tableau Public profile is required.
Method	From Tableau the tester should: Ensure the data is properly loaded and develop data visualisations to push to the Tableau Public Server.
Expected Result	On completing of the above steps, Tableau should produce confirmation that the workbook is accessible online and that the work produced in Tableau Desktop is hosted in the Tableau Public profile.
Actual result	The workbook was successfully hosted and is accessible online through various tested devices and shared links.
Comments	N/A
Resolution	N/A

Test Case 8			
Name	Tableau Public is hosted and visualisations are accessible		
Priority	1	**Date of Test**	1/05/18

Test ID	**8**
Purpose of Test	To Ensure that: Visualisations produced in Tableau are accessible to the end user.
Test Environment	The test environment is as follows: Client Hardware: MacBook Pro 2014 running High Sierra and Tableau Desktop 10.5 Test data was required data for visualisations and a hosted Tableau Public profile is required.
Method	From Tableau the tester should: Access the hosted Tableau Public workbook and ensure all pages and visualisations display as expected from the workbook.
Expected Result	On completing of the above steps, confirmation that the workbook is accessible online and that the work produced in Tableau Desktop is hosted for user access.
Actual result	The workbook was successfully hosted and is accessible online.
Comments	N/A
Resolution	N/A

9.3. Testing Results and Conclusion

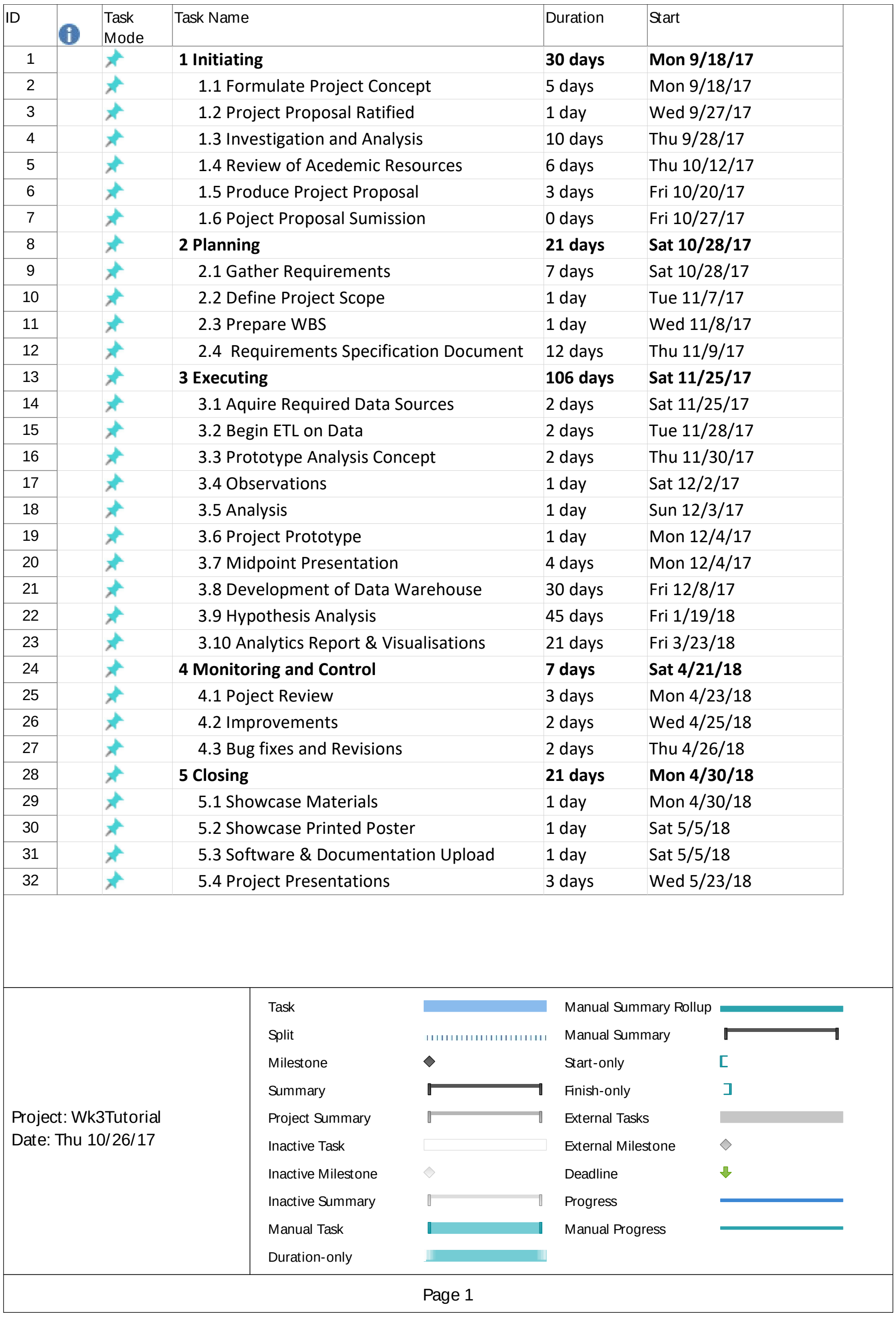
All functional components and obtained output from the various data analysis methods conducted for this project were extensively tested and found to be working optimally as expected. Each component was validated and verified for accuracy and performance. The testing process for the project concluded with no major issues (test case fails) and the findings from the test procedure required no corrective action or further investigation. The test cases were closed and signed off as passed upon completion of the project.

10. Appendices

10.1. Consultation with Dr. Eugene O'Loughlin

To ensure the methods implemented in this research were appropriate and carried out effectively, a consultation meeting was requested to seek the input and advice of Dr. Eugene O'Loughlin. He was extremely helpful in providing insights for proposed methodology and guidance. During the consultation we discussed the aims and objectives of the project and upon examination of the visualisations and work on the data completed to this point, he suggested I look into studying and implementing the Holt-Winters method of exponential smoothing for the forecasting model as my data series exhibited characteristics of seasonality. Dr. O'Loughlin was also very helpful in providing literature and resources to help me implement the aspects discussed. The consultation also provided me with an outline for the structure of my research, how to structure the report and items that would be necessary to document. The advice and guidance gave me great resources and ideas to enhance the quality of the research.

10.2. Project Plan

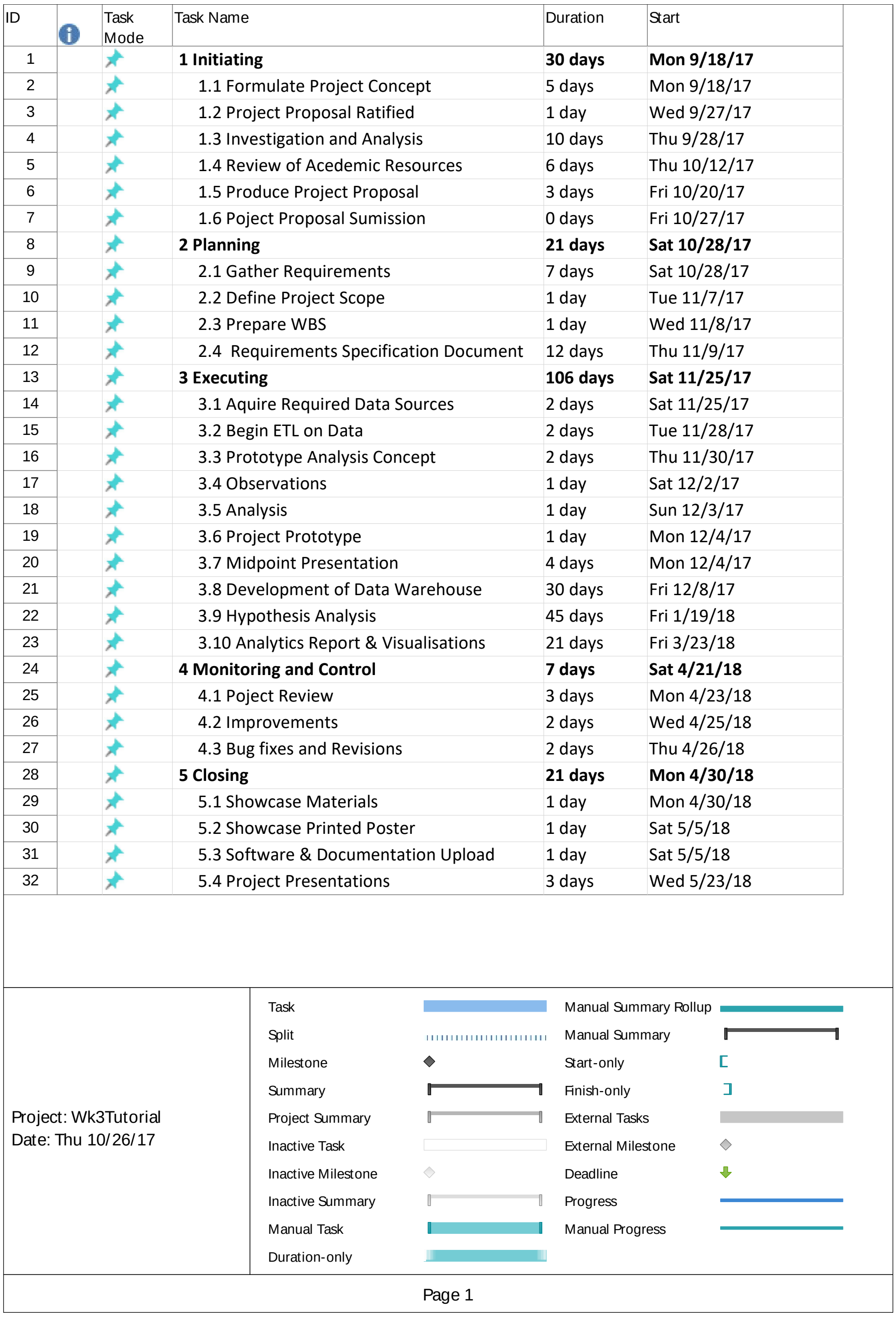


ID	ℹ	Task Mode	Task Name	Duration	Start
1		📌	**1 Initiating**	**30 days**	**Mon 9/18/17**
2		📌	1.1 Formulate Project Concept	5 days	Mon 9/18/17
3		📌	1.2 Project Proposal Ratified	1 day	Wed 9/27/17
4		📌	1.3 Investigation and Analysis	10 days	Thu 9/28/17
5		📌	1.4 Review of Acedemic Resources	6 days	Thu 10/12/17
6		📌	1.5 Produce Project Proposal	3 days	Fri 10/20/17
7		📌	1.6 Poject Proposal Sumission	0 days	Fri 10/27/17
8		📌	**2 Planning**	**21 days**	**Sat 10/28/17**
9		📌	2.1 Gather Requirements	7 days	Sat 10/28/17
10		📌	2.2 Define Project Scope	1 day	Tue 11/7/17
11		📌	2.3 Prepare WBS	1 day	Wed 11/8/17
12		📌	2.4 Requirements Specification Document	12 days	Thu 11/9/17
13		📌	**3 Executing**	**106 days**	**Sat 11/25/17**
14		📌	3.1 Aquire Required Data Sources	2 days	Sat 11/25/17
15		📌	3.2 Begin ETL on Data	2 days	Tue 11/28/17
16		📌	3.3 Prototype Analysis Concept	2 days	Thu 11/30/17
17		📌	3.4 Observations	1 day	Sat 12/2/17
18		📌	3.5 Analysis	1 day	Sun 12/3/17
19		📌	3.6 Project Prototype	1 day	Mon 12/4/17
20		📌	3.7 Midpoint Presentation	4 days	Mon 12/4/17
21		📌	3.8 Development of Data Warehouse	30 days	Fri 12/8/17
22		📌	3.9 Hypothesis Analysis	45 days	Fri 1/19/18
23		📌	3.10 Analytics Report & Visualisations	21 days	Fri 3/23/18
24		📌	**4 Monitoring and Control**	**7 days**	**Sat 4/21/18**
25		📌	4.1 Poject Review	3 days	Mon 4/23/18
26		📌	4.2 Improvements	2 days	Wed 4/25/18
27		📌	4.3 Bug fixes and Revisions	2 days	Thu 4/26/18
28		📌	**5 Closing**	**21 days**	**Mon 4/30/18**
29		📌	5.1 Showcase Materials	1 day	Mon 4/30/18
30		📌	5.2 Showcase Printed Poster	1 day	Sat 5/5/18
31		📌	5.3 Software & Documentation Upload	1 day	Sat 5/5/18
32		📌	5.4 Project Presentations	3 days	Wed 5/23/18

Project: Wk3Tutorial
Date: Thu 10/26/17

Task	▬	Manual Summary Rollup	▬
Split	┈	Manual Summary	┏━┓
Milestone	◆	Start-only	[
Summary	┏━┓	Finish-only	]
Project Summary	┏━┓	External Tasks	▬
Inactive Task	▭	External Milestone	◇
Inactive Milestone	◇	Deadline	⬇
Inactive Summary	┏━┓	Progress	─
Manual Task	▬	Manual Progress	─
Duration-only	▬		

Page 1



ID	ℹ	Task Mode	Task Name	Duration	Start
1		📌	**1 Initiating**	**30 days**	**Mon 9/18/17**
2		📌	1.1 Formulate Project Concept	5 days	Mon 9/18/17
3		📌	1.2 Project Proposal Ratified	1 day	Wed 9/27/17
4		📌	1.3 Investigation and Analysis	10 days	Thu 9/28/17
5		📌	1.4 Review of Acedemic Resources	6 days	Thu 10/12/17
6		📌	1.5 Produce Project Proposal	3 days	Fri 10/20/17
7		📌	1.6 Poject Proposal Sumission	0 days	Fri 10/27/17
8		📌	**2 Planning**	**21 days**	**Sat 10/28/17**
9		📌	2.1 Gather Requirements	7 days	Sat 10/28/17
10		📌	2.2 Define Project Scope	1 day	Tue 11/7/17
11		📌	2.3 Prepare WBS	1 day	Wed 11/8/17
12		📌	2.4 Requirements Specification Document	12 days	Thu 11/9/17
13		📌	**3 Executing**	**106 days**	**Sat 11/25/17**
14		📌	3.1 Aquire Required Data Sources	2 days	Sat 11/25/17
15		📌	3.2 Begin ETL on Data	2 days	Tue 11/28/17
16		📌	3.3 Prototype Analysis Concept	2 days	Thu 11/30/17
17		📌	3.4 Observations	1 day	Sat 12/2/17
18		📌	3.5 Analysis	1 day	Sun 12/3/17
19		📌	3.6 Project Prototype	1 day	Mon 12/4/17
20		📌	3.7 Midpoint Presentation	4 days	Mon 12/4/17
21		📌	3.8 Development of Data Warehouse	30 days	Fri 12/8/17
22		📌	3.9 Hypothesis Analysis	45 days	Fri 1/19/18
23		📌	3.10 Analytics Report & Visualisations	21 days	Fri 3/23/18
24		📌	**4 Monitoring and Control**	**7 days**	**Sat 4/21/18**
25		📌	4.1 Poject Review	3 days	Mon 4/23/18
26		📌	4.2 Improvements	2 days	Wed 4/25/18
27		📌	4.3 Bug fixes and Revisions	2 days	Thu 4/26/18
28		📌	**5 Closing**	**21 days**	**Mon 4/30/18**
29		📌	5.1 Showcase Materials	1 day	Mon 4/30/18
30		📌	5.2 Showcase Printed Poster	1 day	Sat 5/5/18
31		📌	5.3 Software & Documentation Upload	1 day	Sat 5/5/18
32		📌	5.4 Project Presentations	3 days	Wed 5/23/18

Project: Wk3Tutorial
Date: Thu 10/26/17

Task	▬	Manual Summary Rollup	▬
Split	┅	Manual Summary	┣━┫
Milestone	◆	Start-only	[
Summary	┣━┫	Finish-only	]
Project Summary	┣━┫	External Tasks	▬
Inactive Task	▭	External Milestone	◆
Inactive Milestone	◇	Deadline	⬇
Inactive Summary	┣━┫	Progress	━
Manual Task	▬	Manual Progress	━
Duration-only	▬		

Page 1

10.3. Monthly Journals

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: September

Key Achievements:

I was slow to make any meaningful progress on the project in the month of September. I felt anxious about the prospect of committing to a project concept that would ultimately become unviable. Formulating the concept for the project is hugely important and naturally, it was something I was aware of long before the start of the semester. The trouble I initially had was the desire to undertake a project that was specific and relevant to my stream as a BIS student but also encapsulated all aspects of the program. At the project proposal pitch I was given a lot of helpful advice that allowed me to go on a form my project idea.

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: October

Key Achievements:

- Finalised project concept
- Completed and submitted the project proposal

As soon as I came up with the concept for my project I went to discuss the idea with Lisa. At the meeting, Lisa offered lots of encouraging feedback and potential aspects to include in the project. I was able to form an idea of how my prototype would take shape for the Mid-Point presentation from the insights and advice she offered, even at this early stage. This lifted a lot of the pressure I was experiencing at the time off my shoulders and made me feel positive about progressing and completing the proposal document that was due. The remaining time in the month of October was spent researching and completing the proposal document.

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: November

Key Achievements:

- Finalised and submitted Requirements Specification
- Prepared the project prototype for mid-term presentation
- Prepared for mid-term presentation
- Supervisor meeting with Lisa for advice on deliverables and presentation

November was an extremely busy month with important key deliverables to be finalised and submitted for the end of the month. The month of November was spent working on aspects of the requirement specification and ensuring all preliminary components for the mid-term prototype presentation and technical documents were completed. Finalising everything up to required standard for the mid point presentation was quite stressful as there was a lot to get organised and completed, with a lot at stake. To ensure I was on the right path, I had supervisor meetings with Lisa in which she reviewed my work to date, gave advice on the presentation slides and offered recommendations for alterations and inclusions in the technical document before the upload.

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: December

Key Achievements:

- Mid point presentation
- Supervisor meeting with Lisa to prepare for the mid-point presentation

December is always a chaotic month in college as the CA deliverables are all usually within close proximity at the end of the semester. The workload at the start of the month was intensive with a number of CA deliverables due as well as the requirements for the mid point presentation. I successfully achieved the goals for the project on schedule to my project plan and ultimately achieved a grade of 75 with help from Lisa with deliverables before the mid-point presentation. I was ecstatic about as the hard work and planning had paid off with a high result and I was happy that my project was on the right track.

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: January

Key Achievements:

- Started development of data warehouse
- Literature reviews

January began as always – the exams in the first two weeks of the month. Naturally this time took up a large part, if not all of my focus for the duration of the exam period. I always tried to tip away at aspects of the project and remain on schedule, but this is easier said than done during the period of exams. For this period very little progress was made in furthering the development and goals of the project, but literature reviews were sourced and studied after the exam period. I felt this was an ideal time to have somewhat of down time after the exams in which I just read related studies and familiarised myself with the technical approach for a project of this nature. At the end of the month, with the exams completed and classes resumed, developments on technical aspects of the project resumed.

Student Name: Robert Kane

Program: BSc in Business Information Systems

Month: February

Key Achievements:

- Finished and documented initial literature reviews
- Finalised structuring of data for all components
- Set up Tableau to begin exploring the data
- Began testing the data in SPSS
- Began using R Studio in Advanced Business Data Analysis

A lot was achieved in the month of February as the project was now in full swing and the components for development needed to be finalised in order to start initialising the key deliverables. As the second semester had now begun, I was now studying Advanced Business Data Analysis, which introduced the use of R Studio to our studies. I began studying methods and tutorials in both R studio and Tableau this month. These are new technologies I have no previous experience using prior to this project. The challenging aspects of the project are becoming apparent but I am confident my progress in learning the new technologies to achieve the goals of the project is coming along steadily. Dr. Eugene O'Loughlin provides help and advice with learning resources for Tableau, SPSS and R Studio. With each passing week the content covered in the Advanced Business Data Analysis gives me more grounding and understanding in the area of data analytics. The links provided on the module Moodle to learn R Studio and SPSS a great resource and give explanations on how to use the new technologies.

Student Name:	Robert Kane
Program:	BSc in Business Information Systems
Month:	March

Key Achievements:

- Finalised Showcase Profile
- Began structuring and completing the technical document
- Consultation with Dr. Eugene O'Loughlin
- Tableau Hosted

A lot was achieved in the month of March. The required deliverables for the showcase profile were completed and the profile was signed off and passed by Eamon. This is a great milestone as the project really feels like it is entering into the final phase. To get the showcase profile finalised and passed was a great relief and gave a great sense of optimism that the project was on track and nearing completion. The showcase profile is worth a full 3% with an addition 2% for the production of a poster, so this deliverable was crucial and important to complete. The two-week mid term break was spent working intensively on CA deliverables and using the time to study for the upcoming exams later in April. I used this time period to really make a significant dent in the workload of the project. Using the time not in class to really focus on aspects of the project and identify any unforeseen areas that may be problematic. Thankfully any obstacles were overcome and resolutions to the problems were resolved. I had issues with the Tableau product key and initially chose the wrong hosting options. This lead to the initial account I set up for Tableau expiring and being suspended as I had set up the wrong profile type. Once I identified the problems, I was able to rectify the situation and get my Tableau workbook hosted on the proper Tableau Public platform.

I scheduled a consultation meeting with Dr. Eugene O'loughlin this month to seek his advice and guidance on aspects of my project now that it was well into the development stages. I demonstrated my initial findings and Tableau visualisations and he provided insights and advice on methods to implement for my project. He provided learning resources and literature on advanced methods of exponential forecasting that would not be covered on the course. This consultation was of tremendous benefit as Dr O'Loughlin provided me with invaluable advice and guidance on what would become major components for the successful implementation of this project. The following weeks were spent studying the resources and methods given to at the consolation. I began structuring and aggregating the data for the project in such a way that allowed me to implement the methods discussed.

10.4. Proposal & Requirement Specification

Introduction

This project proposes to conduct an in-depth statistical analysis on the impact temperature and weather conditions have on incidents of crime. The purpose of this study is to examine existing academic hypothesis and scientific findings in this area, as well as seeking to identify new links, trends and patterns of my own to form a conclusion and make future forecasts. I propose to achieve this by developing a data warhouse and implementing advanced methods of data analysis I will study throughout my final year as a Business Information Systems student at NCI.

For this analysis, I propose to build a statistical model to demonstrate and visualise the effect temperature and weather conditions have on various categories of crime. For me to achieve this and present meaningful, accurate results, the data I use must be accurate in order to provide the base for all transformations and subsequent aggregate queries. My initial proposal aimed to conduct the study based on Ireland, but my requirements research identified limitations in the data made publically available in the required domains for a study of this nature. Due to the restrictions and limitations on the availability of appropriate data for Ireland, I have chosen to focus the study on the American city of Chicago.

Background

Motivations for the Development of this Project

As a capstone to my studies as a Business Information Systems student at NCI, I want to produce a substantial project that integrates and synthesizes all the various aspects of my learning and acquired expertise in this field. For this reason I have chosen to undertake a project I feel encapsulates and showcases the disciplines and advanced understanding of modules such as Databases, Business Intelligence & Data Warehousing and Business Data Analysis. The area of data analytics is one that I am extremely interested in and have a keen desire to pursue as a career upon graduation from NCI. I feel this project will allow me the opportunity to pursue and showcase relevant independent academic research in this field of study, as well as utilising the resources available to me at NCI, such as the expertise and guidance of lecturers at NCI in these subject areas.

As I will be studying Advanced Data Analytics in the second semester of my final year, I feel it is important for me to be working on a technical project that incorporates these analytical skills and allows me to present them in an academic body of research. This project will be a valuable asset for me to be able to show potential employers. Once I have completed the ETL processing of the data for the study and structured the technical aspects of the data warehouse, I will then be in a position to conduct an in-depth data analytics report to demonstrate and visualise the findings of my study.

The most important aspect of this project for me is not the concept, but rather the opportunity to showcase a high level of proficiency in the technologies and methodologies used to develop it. The architectural planning and implementation of a project such as this will allow me to work with technologies that are essential in the field of business intelligence and data analytics. It will also allow me to demonstrate a high level of abstract thinking and an ability to integrate advanced analytical methodologies for business solutions.

Business Information Systems students sit in the middle of technology and business solutions; therefore, it is important for me to demonstrate not just the ability to work with technology proficiently, but to also implement that technology to effectively communicate meaningful analysis, insights and solutions.

Existing Studies on Weather & Criminality

My initial exploration into the relationship of weather and criminality has brought numerous existing studies on the subject to my attention. I intend to source and extensively study these works continuously throughout the lifecycle of the project as part of the literature review. This will allow me to formulate a unique approach to my analysis both in terms of relevancy and identifying a variance in different methods that have not yet been explored. I feel there will be enough variants in my analysis and use of current, up to date data to justify the uniqueness of my approach and conclusions.

As I study Advanced Business Data Analysis in the second semester, I can receive guidance and recommendations on how new and innovative research techniques can be carried out on the data from discussing the project with my lecturer and project supervisor.

The Aim of the Project

This project aims to develop a data warehouse. The intended use and aim for the development of the data warehouse is to identify and generate meaningful statistical correlations found in data on crime and weather. The output gathered from this data warehouse will then be implemented to conduct an in-depth statistical analysis with the aim of producing a detailed analytics report on the impact weather conditions and temperature have on incidents of crime.

Commercial Potential, Target Market & Further Uses

The commercial potential and practical uses for this project once completed would be for local authorities and governments to strategize methodologies of crime prevention and resource planning. The findings of the study could in theory be applied to any region for use in predictive analysis and pre-emptive measures against incidents of crime. The methods and conclusions of the analysis could be implemented to make future forecasts and assumptions. This insight would be a tool of great interest to those in the field of law enforcement, public safety and social science.

Project Scope & Deliverables

The scope of this project is to develop a data warehouse to use in a statistical analysis of the impact weather conditions and temperatures have on crime.

The system will be used to generate relevant data visualisations and to develop statistical correlations for use in the production of a data analytics report.

Below is a list of key, high level deliverables this project aims to achieve:

- A data warehouse to aggregate data and perform intelligence queries

- A method of loading the data warehouse with relevant data
- A method of storing and filtering the data
- A dashboard to dynamically present data visualisations in an intuitive and meaningful way
- Structured documentation on the data sources, processes and ETL techniques used to load the data warehouse
- The completion of an extensive statistical analysis on the impact of weather conditions and temperature on crime to produce an in-depth data analytics and intelligence report.

Project Constraints

Data Sources/Inputs

The requirements elicitation process required sourcing appropriate data and verifying it was suitable for this project. The main sources of data that will be used for the development of this project are:

- The Chicago Police Department's open data API.
- NCDC archive of global historical weather

These data sources will be used for the loading of data to the data warehouse after ETL processing has been carried out. The generated output and findings of the data warehouse will then be used for the statistical analysis report.

Schedules

Details of the scheduling constraints that are placed upon the requirements elicitation process of this project can be found in a detailed Gantt chart available in the appendix under project plan. This provides details of anticipated dates for project milestones and deliverables for the project.

Software Development Environment

The data warehouse will be developed using MySQL and visualisations of the data will be presented using a Tableau dashboard. Details of these technologies can be found in section 3 of this document.

Tools for Statistical Analysis

The tools I intend to implement for the statistical analysis in this project are Microsoft Excel, IBM SPSS and the R Data Science Language. Details of these resources can be found in section 3 of this document.

High Level Analysis

Challenging Aspects of this Project

The challenging aspects of this project will be for me to overcome the complexity and difficulty of implementing the correct methods of data analysis. As the methods of these calculations will determine the overall accuracy and ultimate relevancy of my conclusions, it is something I must give great consideration to from the outset. Staging and correctly processing the source data for the project will also be a challenging aspect for these reasons also.

Solutions to These Challenges

To overcome these challenges, I intend to conduct extensive research into similar studies of this nature throughout the lifecycle of the project. This will help me identify the correct methodologies to implement. Dr. Eugene O'Loughlin has also kindly offered his guidance and advice should I require it at any stage in the project. I have already spoken with him in relation to identifying these solutions for my project.

Definitions, Acronyms & Abbriviations

The following section explains some of the common terms and language that will be used throughout this document and gives some context to their role and functions in this project.

- **Data Warehouse**
  A data warehouse is a large store of data accumulated from a wide range of sources and used for reporting and data analysis.
- **Tableau**
  Tableau is a data visualisation application that allows users to connect, explore and visualize data. This interface will present the underlying data from the data warehouse in a meaningful and intuitive way to the user.
- **IBM SPSS**
  IBM SPSS stands for Statistical Package for the Social Sciences (SPSS Statistics). It is a software package used for logical batched and non-batched statistical analysis and will be used for the statistical analysis of the output generated by the data warehouse in this project.
- **Microsoft Excel**
  Microsoft Excel is a spreadsheet program included in the Microsoft Office suite of applications. Spreadsheets present tables of values arranged in rows and columns that can be manipulated mathematically using both basic and complex arithmetic operations and functions. Excel will be used in this project for the storing and loading of required data, as well as statistical computation and data analysis.
- **SQL**
  SQL stands for Structured Query Language. It is a language designed to facilitate database manipulation.
- **ERD**
  Stands for Entity-Relationship Diagram. It is a data modelling technique that graphically illustrates an information system's entities and the relationships between those entities. An ERD is a conceptual and representational model of data used to represent the entity framework infrastructure.

- **ETL**

  Stands for Extract, Transform and Load. These are three database functions that are combined in to one tool to take data from one source and place it in another.
- **CSV Files**

  CSV stands for comma-separated values. A CSV file allows data to be saved in a table structure format.
- **DB**

  Database. Stores persistent data for use in the project and data warehouse.
- **GUI**

  Stands for Graphical User Interface. Allows the user to interact with the system. In the case of this project, the Tableau dashboard will be the end user GUI.
- **R**

  R is an open source programming language and software environment for statistical computing and graphics. This will be implemented along with Excel and IBM SPSS to generate statistical models for use in the analytics report.
- **API**

  Set of functions and procedures that allow the creation of applications which access the features or data of an operating system, application, or other service.

Technologies & Special Resources Required

Technical Approach

I propose to develop the data warehouse using MySQL. Data visualisations and analytics will be performed using Tableau and SPSS Statistics. Microsoft Excel will be used throughout the ETL process and data staging phase of the project, as well as for high-level statistical analysis and data manipulations. For statistical computing and graphics, I will use the R programming language. Once development is completed and the data warehouse is fully operational, a Tableau visualisation dashboard will be synchronised to display the output and hosted online.

Special Resources Required

Software

I will require access to the NCI Data Analytics Suite and software programs.

Hardware

There are no special hardware resources required for this project.

Documentation

There are extensive academic studies and analysis conducted in this area as well as documentation on the technology stack I intend to use for this project.

Technologies

MySQL

I propose to develop the data warehouse using MySQL. This is a perfect development environment for a project of this scale and proposed data set size. MySQL offers real time analytics, standard reporting and historical data therefore is ideal for the development of data analytics and OLAP applications.

Tableau

Tableau is a desktop business intelligence and analytics visualisation dashboard that can connect to almost any database. I intend to connect the data warehouse to a Tableau dashboard to serve as a GUI in order to present the underlying data to the end user.

IBM SPSS

The role of SPSS in this project will be to perform statistical analysis and calculations on the output of the data warehouse. Below is an example of what the statistical output from SPSS might look like when generated for the analytics report.

Microsoft Excel

Microsoft Excel will be used to handle the CSV files for the data warehouse input for this project. Excel will also be used as the staging environment when performing ETL and data cleansing processes on the source data. Excel will also serve this project as a tool to perform high-level analysis for the production of the analytics report.

Requirments Elicitation

This section of the document will set out and give a detailed description of the technical requirements and dependencies for the development of all aspects of the project.

User Requirements

As I will be using the system to generate output to produce an analytics report, it was necessary for me to carefully consider and determine the core functionality that is required for the data warehouse to function and serve the project at an optimum level of performance. As such, the requirements elicitation process for this project involved me carrying out academic research and literary reviews of existing statistical studies and systems for me to establish a clear approach and set of required components for a project of this nature.

From this research and requirements elicitation process, I have determined the following list of core functionalities and critical outputs required of the system from an end-user requirements perspective:

- Ability to load data warehouse with data.
- Ability to dynamically present data visualisations
- Ability to use data warehouse output for statistical analysis.

Once the system has been developed and these functionalities provide the required output for the statistical analysis, I will have the necessary data required to produce an in-depth analytics report. I will then use the output generated by the system for analysis using tools such as IBM SPSS.

Requirements Specification

Technical details of both functional and non-functional requirements will be given in the following section of the document. This will give a collective overview of the requirements that are fundamental to the overall completion and delivery of this project.

Functional Requirements

The core functionality required for the successful completion of this project is as follows:

- The data warehouse must have the ability to be loaded with data once cleansing and ETL processes have been performed.
- The system must have the ability to present data visualisations to the user.

Requirement 1: Generate Data Output

- **Description & Priority**

'Generate Data Output' is the main requirement of the system and is essential to the overall architecture of the project as the output generated by the data warehouse will be exported and used for the statistical analysis. The data warehouse output will also be displayed using a data visualisation dashboard.

- **Use Case Diagram**

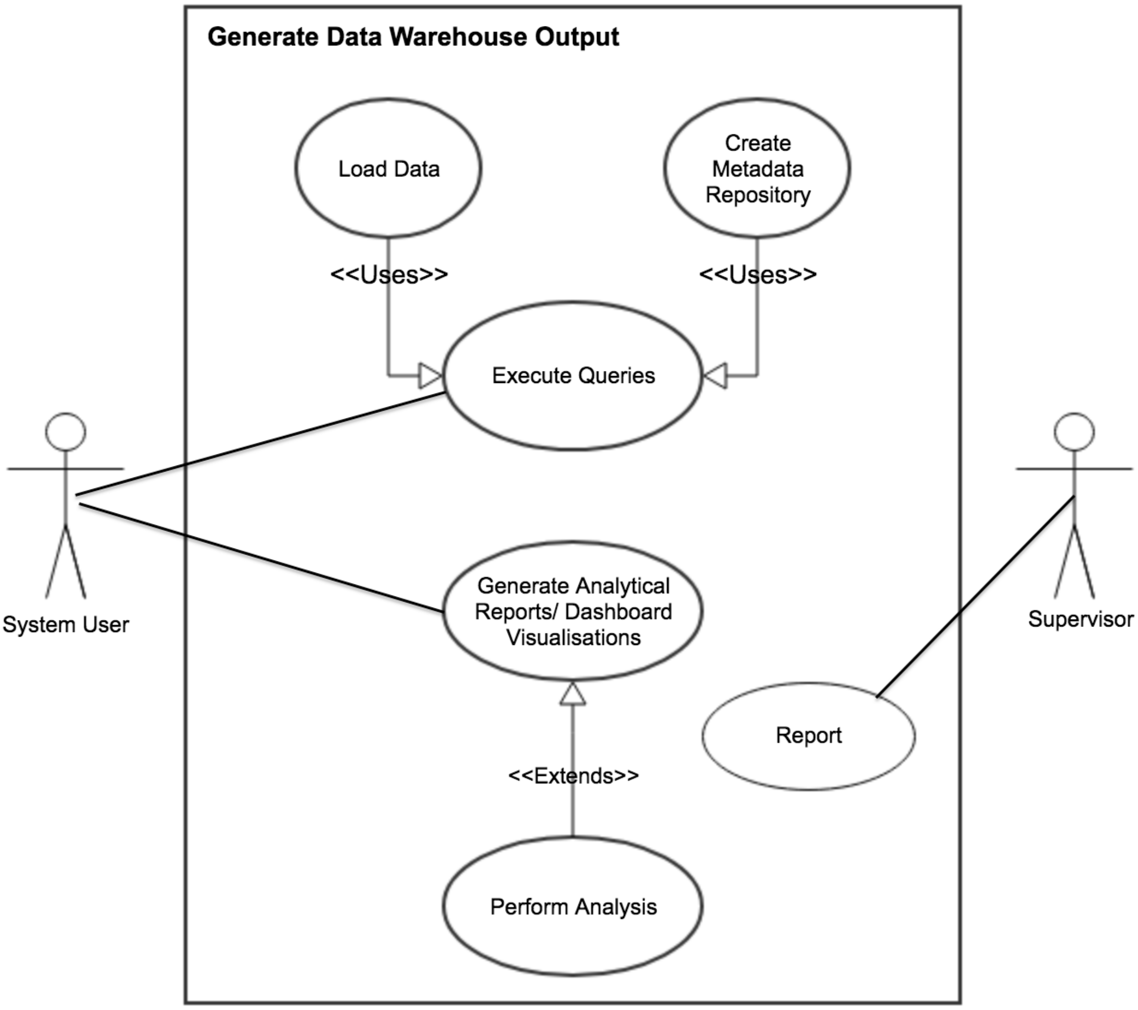

Generate Data Warehouse Output
Load Data
Create Metadata Repository
<<Uses>>
<<Uses>>
Execute Queries
System User
Supervisor
Generate Analytical Reports/ Dashboard Visualisations
Report
<<Extends>>
Perform Analysis

Fig 1: Use Case Diagram for Generating Data Output.

- **Use Case**

Below is the use case for Generating Output:

Description/Scope	This use case describes how to use the data warehouse to generate output from the system for statistical analysis.
Flow Desciption:	
Precondition	The data warehouse has been loaded with data and the development environment is now open.
Activation	This use case starts when the user runs a query on the data warehouse
Main Flow	1. The system loads the data warehouse environment. 2. The user chooses to execute statements and run queries on the data or construct a new statement (see A1). 3. The system runs the queries. 4. The system displays the results of the queries (see E1) 5. The system updates and the user saves the results. 6. The data visualisations dashboard updates to reflect the changes.
Alternate Flow	A1: New Query Constructed 1. The user constructs a new statement. 2. The user executes the statements. 3. The system displays the results.
Exceptional Flow	E1: SQL Syntax/ System Error 1. The system alerts the user that an error has occurred. 2. The user identifies the errors and makes corrections.
Termination	The User exits the data warehouse environment and closes down the system.
Post Condition	The System saves the changes to the schema.

Requirement 2: Generate Data Visualisations

- **Description & Priority**

The dashboard is of critical importance as it presents the value of the underlying system data. Navigation and interaction with the dashboard allows the user to access and retrieve the information they require from the system.

- **Use Case Diagram**

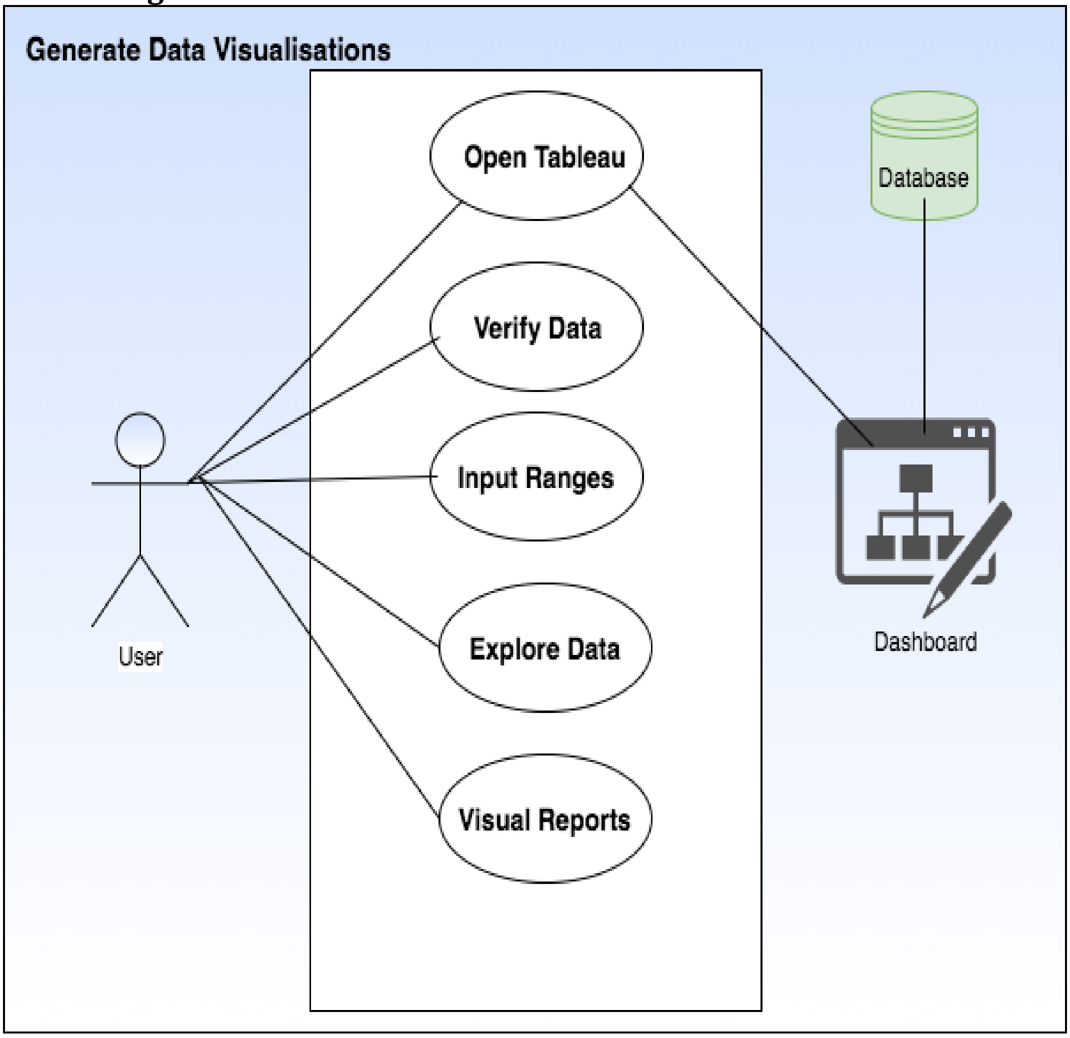


Fig 2: Use Case Diagram for Generating Data Visualisations.

➢ Use Case

Below is a use case for Generate Data Visualisations:

Description/Scope	Allows the user access to the data visualisation dashboard for the system.
Flow Description:	
Pre-Condition	Dashboard must be linked to DW, hosted and accessible.
Activation	The user opens the dashboard interface.
Main Flow	The user performs the desired operations and explores output of the underlying data.
Alternate Flow	The user alters the parameters on the data to dynamically update the desired visualisations.
Exceptional Flow	The system encounters a communication error with the connecting data warehouse.
Termination	The user exits the dashboard.
Post Condition	The user retrieves and saves required output and reports.

Requirement 3: Generate Reports

- **Description & Priority**

The ability to generate reports is of significant importance as it offers real value to the user. This aspect of functionality has commercial potential and further uses beyond the lifecycle of the project to external end users of the system.

- **Use Case Diagram**

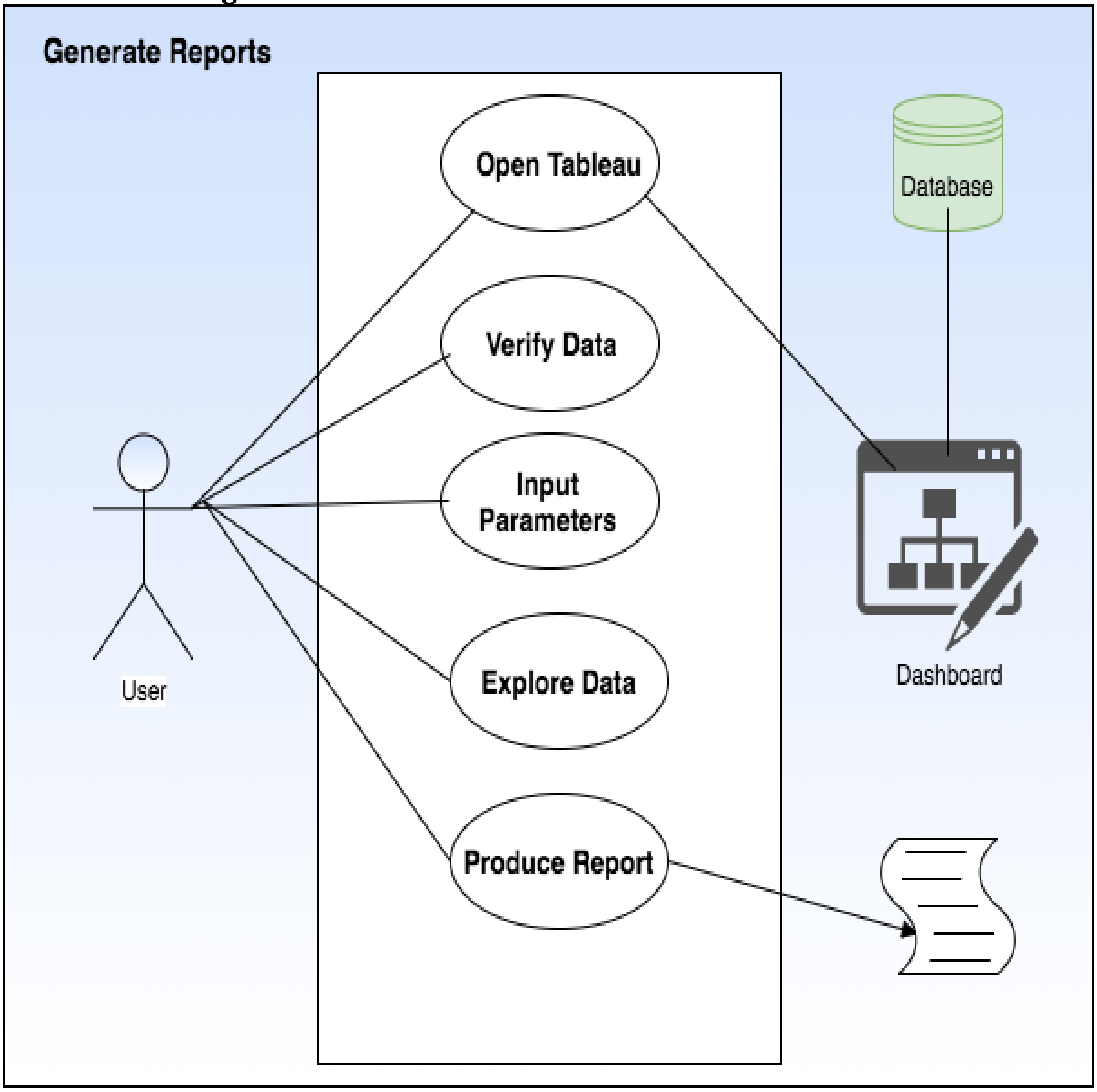


Fig 3: Use Case Diagram for Generating Reports

- **Use Case**

Below is a use case for 'Generate Reports'

Description/Scope	Allows users of the system to produce reports
Flow Description:	
Pre-Condition	Dashboard must be linked to DW, hosted and accessible. The user must have access rights
Activation	The user opens the dashboard
Main Flow	The user performs the desired operations and explores output of the underlying data. The user can enter a range of parameters in order to filter the data to specific dates and focuses. Once the user has the desired data ranges and result, they can export the results of the report.
Termination	The user exits the system
Post Condition	The user retrieves the exported reports and data from the system.

Non Functional Requirements

Non-functional attributes required by the system are detailed below:

Performance / Response Time Requirement

The system must be fluid in response to promote usability for the end user. All actions executed by the user of the visualization dashboard must evoke an instant dynamic reaction to reflect the underlying data and parameters the user has requested.

Availability Requirement

The system must be available to all users required to use it. The system will initially be available through a local environment before the dashboard is hosted online.

Accessibility Requirement

The system should be easy to use, intuitive and extremely navigable with a focus of usability. The system should be highly responsive and fluid to deliver an interactive experience to the user. Strict design principles in accordance with data visualisations should also be adhered to in order to correctly and effectively communicate the underlying data of the system.

Security Requirement

There is no requirement for the system to store sensitive or personal data but the system should still prevent against malicious threats.

Portability Requirement

The system is not currently being designed for portability, but with future development and implementation of required functionality the system could become portable.

Reliability Requirement

The system output must have the utmost integrity and reliability for the results of computations and data aggregations displayed to the user to be accurate and correct. The reliability of these results is fundamental to the success of the project.

Error Checking Requirement

The system must be able to identify errors and alert the user to avoid any incorrect results or data being displayed that could potentially skew the data analysis and conclusions.

10.4.1. ETL Processes & Data requirements

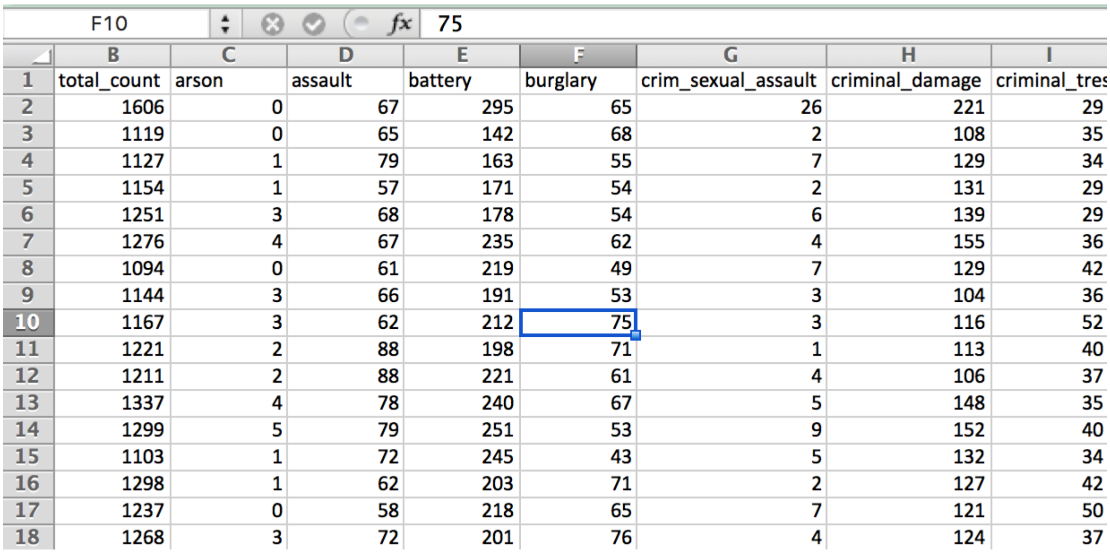
As part of the requirements gathering process, I have identified and investigated the sources and suitability of data required for this project. Once I determined the usability of the data, I initiated the process of familiarising myself with the format in order to begin planning how the data warehouse will consume the data and be developed.

ETL (Extract, Load & Transform) Processes

ETL (Extract, Transform and Load) is a process in data warehousing responsible for the extraction of data from source systems and placing it into the data warehouse. ETL can fundamentally be viewed as the intersection of computer science, management of information systems and data science and as such, a good design of these processes in the early stages of a data warehouse project is essential.

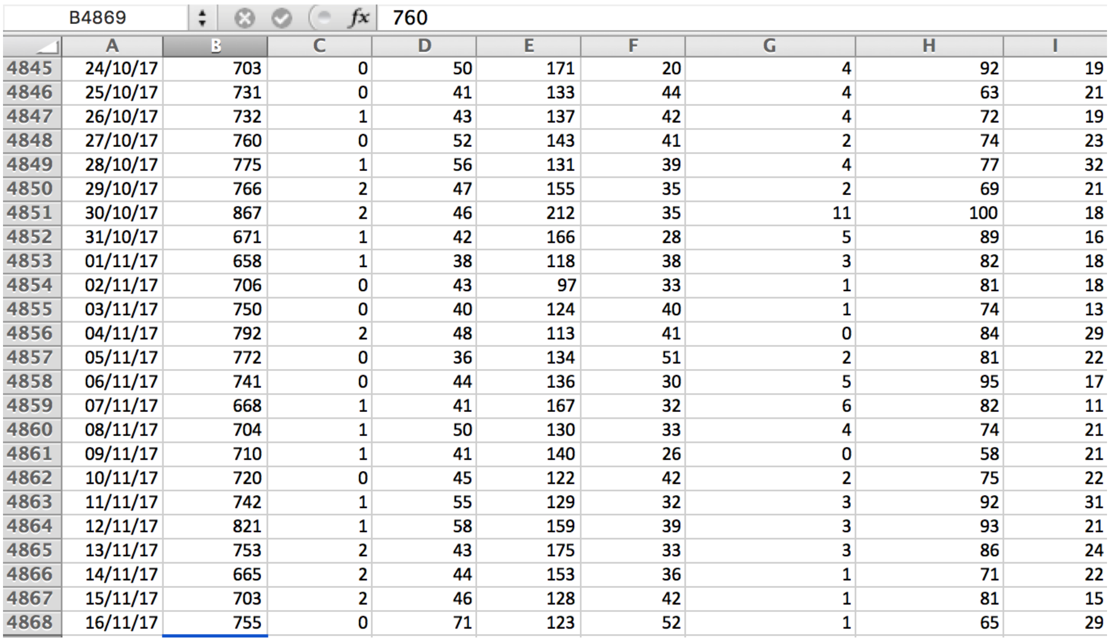
The below screenshots are of the initial ETL processes being carried out on data sources during the requirements gathering stage.

CSV files on crime data:



	B	C	D	E	F	G	H	I
1	total_count	arson	assault	battery	burglary	crim_sexual_assault	criminal_damage	criminal_tres
2	1606	0	67	295	65	26	221	29
3	1119	0	65	142	68	2	108	35
4	1127	1	79	163	55	7	129	34
5	1154	1	57	171	54	2	131	29
6	1251	3	68	178	54	6	139	29
7	1276	4	67	235	62	4	155	36
8	1094	0	61	219	49	7	129	42
9	1144	3	66	191	53	3	104	36
10	1167	3	62	212	75	3	116	52
11	1221	2	88	198	71	1	113	40
12	1211	2	88	221	61	4	106	37
13	1337	4	78	240	67	5	148	35
14	1299	5	79	251	53	9	152	40
15	1103	1	72	245	43	5	132	34
16	1298	1	62	203	71	2	127	42
17	1237	0	58	218	65	7	121	50
18	1268	3	72	201	76	4	124	37

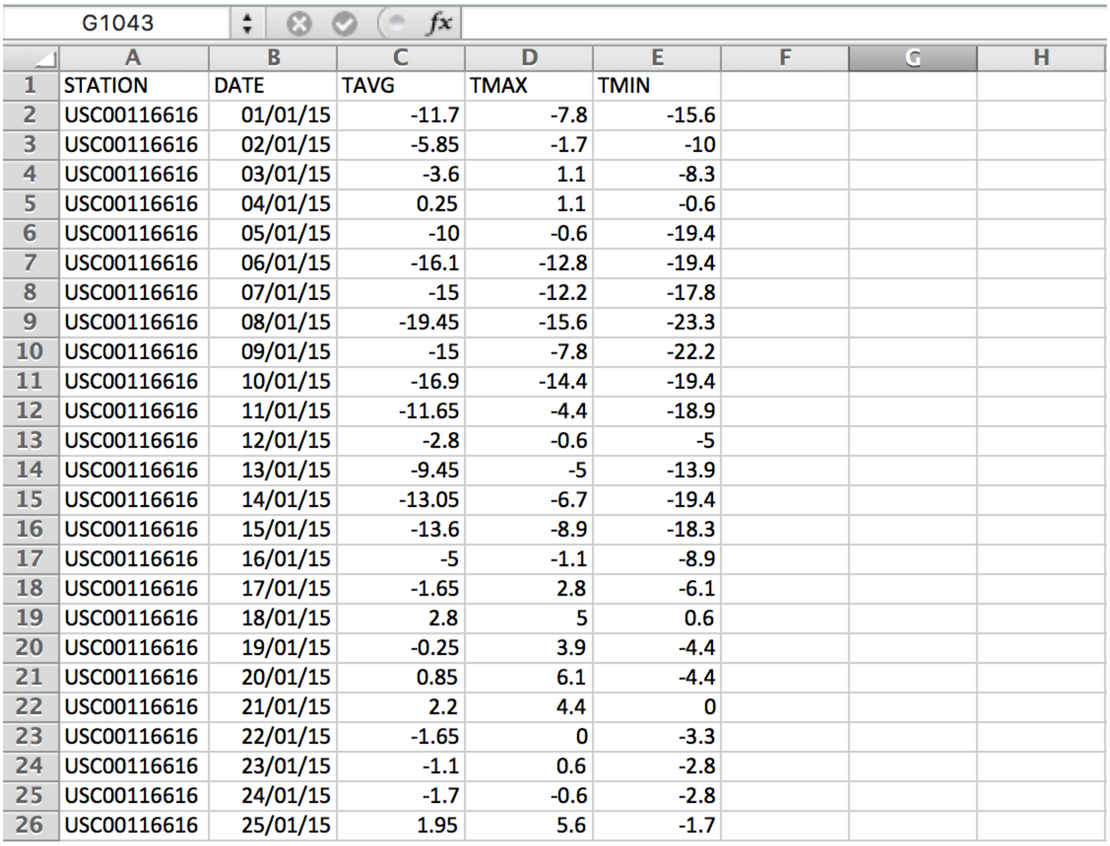
Fig 5: Crime Data Spread sheet



	A	B	C	D	E	F	G	H	I
4845	24/10/17	703	0	50	171	20	4	92	19
4846	25/10/17	731	0	41	133	44	4	63	21
4847	26/10/17	732	1	43	137	42	4	72	19
4848	27/10/17	760	0	52	143	41	2	74	23
4849	28/10/17	775	1	56	131	39	4	77	32
4850	29/10/17	766	2	47	155	35	2	69	21
4851	30/10/17	867	2	46	212	35	11	100	18
4852	31/10/17	671	1	42	166	28	5	89	16
4853	01/11/17	658	1	38	118	38	3	82	18
4854	02/11/17	706	0	43	97	33	1	81	18
4855	03/11/17	750	0	40	124	40	1	74	13
4856	04/11/17	792	2	48	113	41	0	84	29
4857	05/11/17	772	0	36	134	51	2	81	22
4858	06/11/17	741	0	44	136	30	5	95	17
4859	07/11/17	668	1	41	167	32	6	82	11
4860	08/11/17	704	1	50	130	33	4	74	21
4861	09/11/17	710	1	41	140	26	0	58	21
4862	10/11/17	720	0	45	122	42	2	75	22
4863	11/11/17	742	1	55	129	32	3	92	31
4864	12/11/17	821	1	58	159	39	3	93	21
4865	13/11/17	753	2	43	175	33	3	86	24
4866	14/11/17	665	2	44	153	36	1	71	22
4867	15/11/17	703	2	46	128	42	1	81	15
4868	16/11/17	755	0	71	123	52	1	65	29

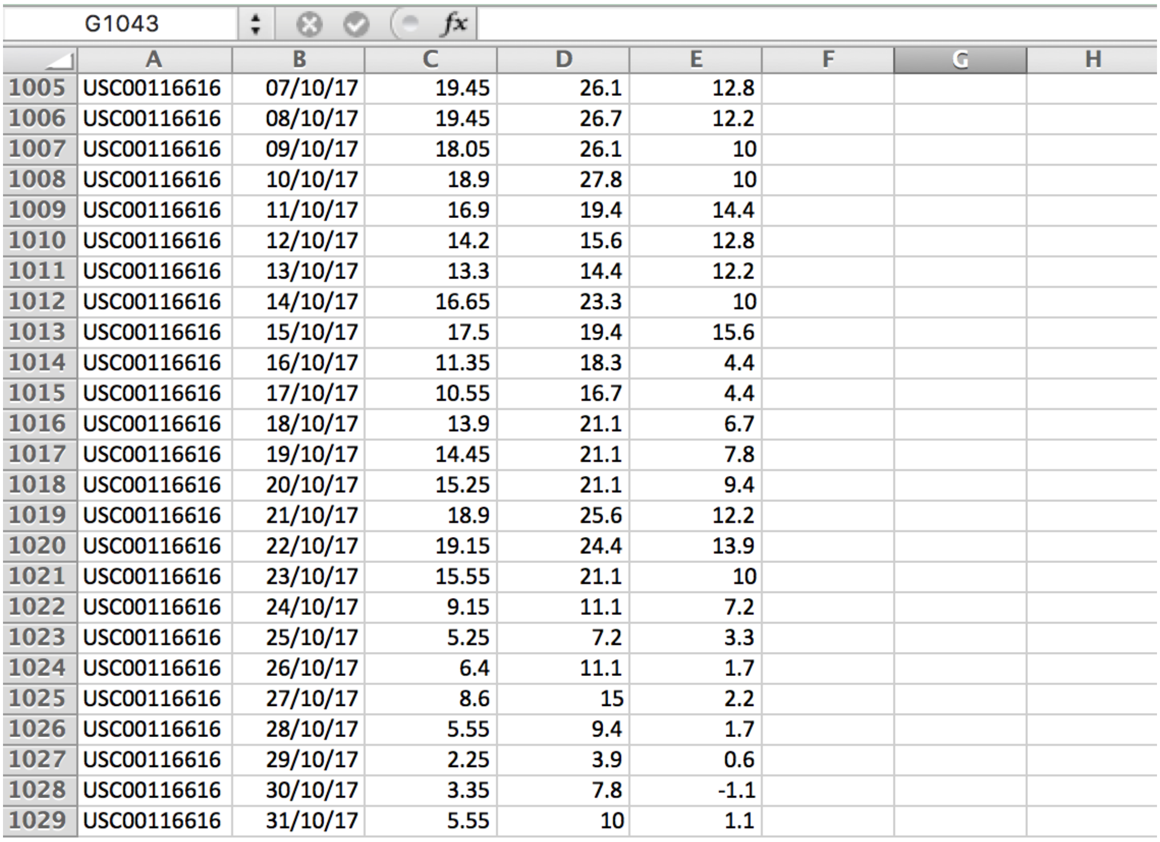
Fig 6: Crime Data Spread sheet

CSV files on weather data:



	A	B	C	D	E	F	G	H
1	STATION	DATE	TAVG	TMAX	TMIN			
2	USC00116616	01/01/15	-11.7	-7.8	-15.6			
3	USC00116616	02/01/15	-5.85	-1.7	-10			
4	USC00116616	03/01/15	-3.6	1.1	-8.3			
5	USC00116616	04/01/15	0.25	1.1	-0.6			
6	USC00116616	05/01/15	-10	-0.6	-19.4			
7	USC00116616	06/01/15	-16.1	-12.8	-19.4			
8	USC00116616	07/01/15	-15	-12.2	-17.8			
9	USC00116616	08/01/15	-19.45	-15.6	-23.3			
10	USC00116616	09/01/15	-15	-7.8	-22.2			
11	USC00116616	10/01/15	-16.9	-14.4	-19.4			
12	USC00116616	11/01/15	-11.65	-4.4	-18.9			
13	USC00116616	12/01/15	-2.8	-0.6	-5			
14	USC00116616	13/01/15	-9.45	-5	-13.9			
15	USC00116616	14/01/15	-13.05	-6.7	-19.4			
16	USC00116616	15/01/15	-13.6	-8.9	-18.3			
17	USC00116616	16/01/15	-5	-1.1	-8.9			
18	USC00116616	17/01/15	-1.65	2.8	-6.1			
19	USC00116616	18/01/15	2.8	5	0.6			
20	USC00116616	19/01/15	-0.25	3.9	-4.4			
21	USC00116616	20/01/15	0.85	6.1	-4.4			
22	USC00116616	21/01/15	2.2	4.4	0			
23	USC00116616	22/01/15	-1.65	0	-3.3			
24	USC00116616	23/01/15	-1.1	0.6	-2.8			
25	USC00116616	24/01/15	-1.7	-0.6	-2.8			
26	USC00116616	25/01/15	1.95	5.6	-1.7			

Fig 7: Weather Data Spread sheet



	A	B	C	D	E	F	G	H
1005	USC00116616	07/10/17	19.45	26.1	12.8			
1006	USC00116616	08/10/17	19.45	26.7	12.2			
1007	USC00116616	09/10/17	18.05	26.1	10			
1008	USC00116616	10/10/17	18.9	27.8	10			
1009	USC00116616	11/10/17	16.9	19.4	14.4			
1010	USC00116616	12/10/17	14.2	15.6	12.8			
1011	USC00116616	13/10/17	13.3	14.4	12.2			
1012	USC00116616	14/10/17	16.65	23.3	10			
1013	USC00116616	15/10/17	17.5	19.4	15.6			
1014	USC00116616	16/10/17	11.35	18.3	4.4			
1015	USC00116616	17/10/17	10.55	16.7	4.4			
1016	USC00116616	18/10/17	13.9	21.1	6.7			
1017	USC00116616	19/10/17	14.45	21.1	7.8			
1018	USC00116616	20/10/17	15.25	21.1	9.4			
1019	USC00116616	21/10/17	18.9	25.6	12.2			
1020	USC00116616	22/10/17	19.15	24.4	13.9			
1021	USC00116616	23/10/17	15.55	21.1	10			
1022	USC00116616	24/10/17	9.15	11.1	7.2			
1023	USC00116616	25/10/17	5.25	7.2	3.3			
1024	USC00116616	26/10/17	6.4	11.1	1.7			
1025	USC00116616	27/10/17	8.6	15	2.2			
1026	USC00116616	28/10/17	5.55	9.4	1.7			
1027	USC00116616	29/10/17	2.25	3.9	0.6			
1028	USC00116616	30/10/17	3.35	7.8	-1.1			
1029	USC00116616	31/10/17	5.55	10	1.1			

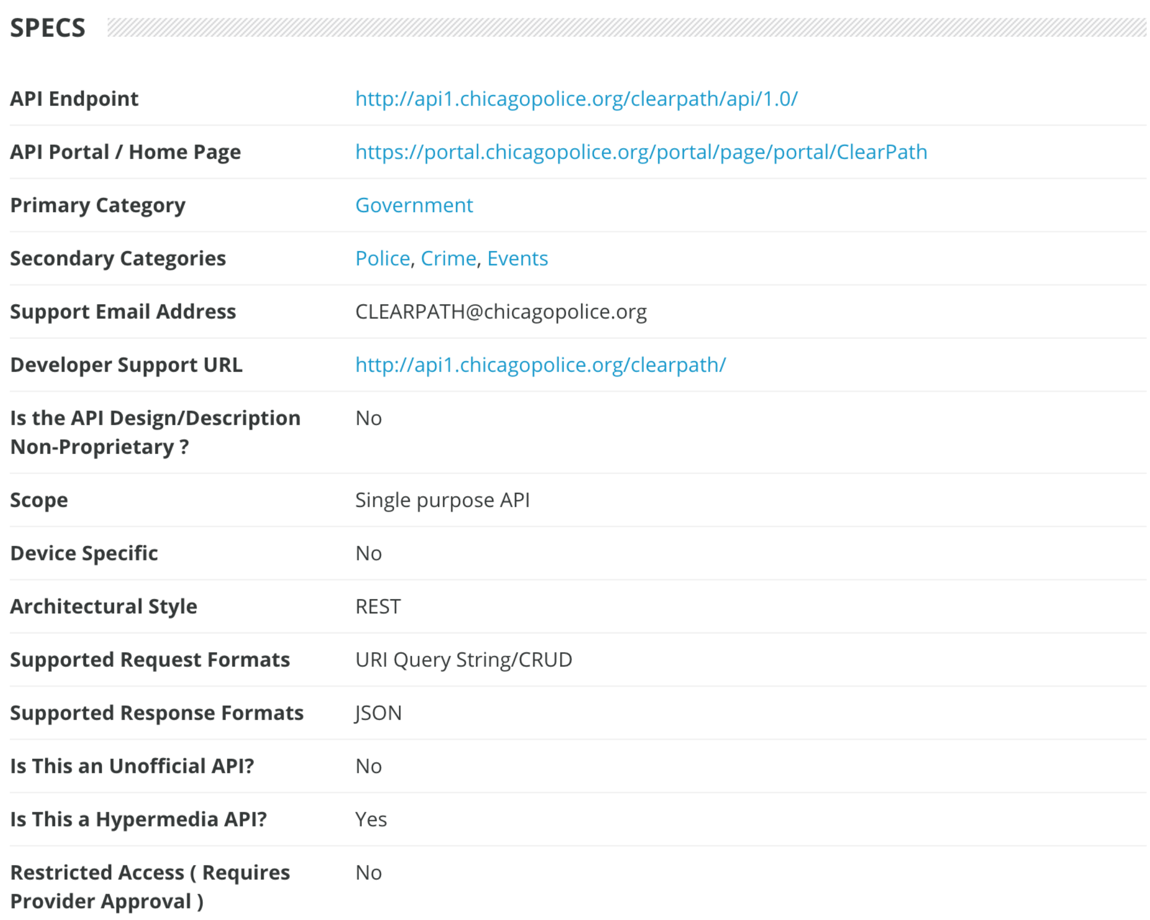
Fig 8: Weather Data Spread sheet

10.4.2. Application Programming Interfaces (API)

This project will implement the following API and data sources:

The Chicago Police Department's Open Data API

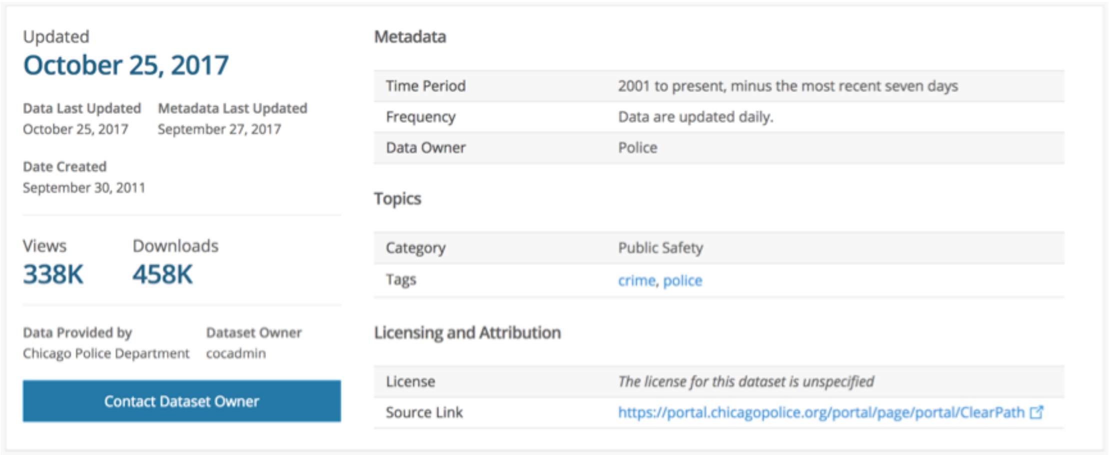
Since 2001, The Chicago Data Portal has made available to the public, statistics recorded on crime via The Chicago Police Department's open data API. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.



SPECS

API Endpoint	http://api1.chicagopolice.org/clearpath/api/1.0/
API Portal / Home Page	https://portal.chicagopolice.org/portal/page/portal/ClearPath
Primary Category	Government
Secondary Categories	Police, Crime, Events
Support Email Address	CLEARPATH@chicagopolice.org
Developer Support URL	http://api1.chicagopolice.org/clearpath/
Is the API Design/Description Non-Proprietary ?	No
Scope	Single purpose API
Device Specific	No
Architectural Style	REST
Supported Request Formats	URI Query String/CRUD
Supported Response Formats	JSON
Is This an Unofficial API?	No
Is This a Hypermedia API?	Yes
Restricted Access (Requires Provider Approval)	No

Fig 9: Chicago Police Department API Specifications



Updated
October 25, 2017

Data Last Updated	Metadata Last Updated
October 25, 2017	September 27, 2017

Date Created
September 30, 2011

Views	Downloads
338K	458K

Data Provided by	Dataset Owner
Chicago Police Department	cocadmin

Metadata

Time Period	2001 to present, minus the most recent seven days
Frequency	Data are updated daily.
Data Owner	Police

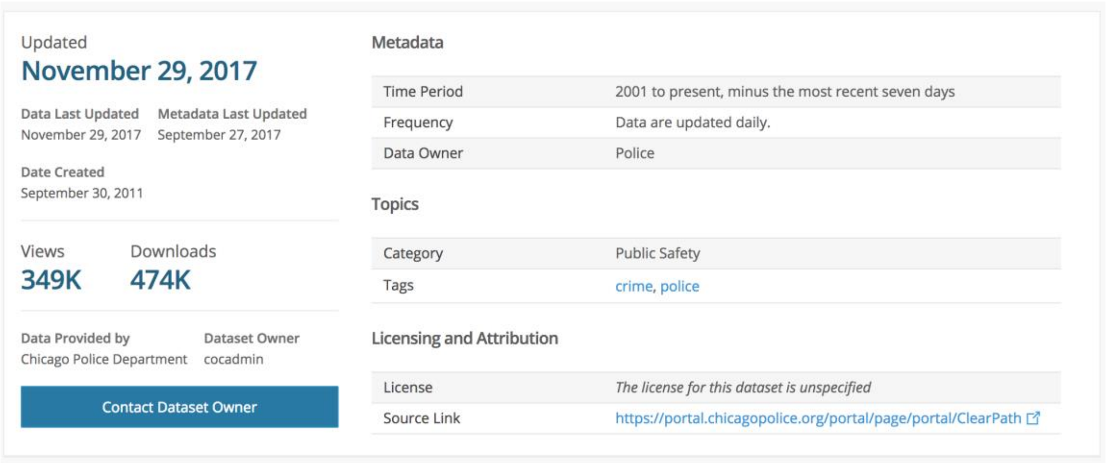
Topics

Category	Public Safety
Tags	crime, police

Licensing and Attribution

License	*The license for this dataset is unspecified*
Source Link	https://portal.chicagopolice.org/portal/page/portal/ClearPath

Fig 10: Meta Data for Crime Data Sets



Updated
November 29, 2017

Data Last Updated
November 29, 2017

Metadata Last Updated
September 27, 2017

Date Created
September 30, 2011

Views
349K

Downloads
474K

Data Provided by
Chicago Police Department

Dataset Owner
cocadmin

Metadata

Time Period	2001 to present, minus the most recent seven days
Frequency	Data are updated daily.
Data Owner	Police

Topics

Category	Public Safety
Tags	crime, police

Licensing and Attribution

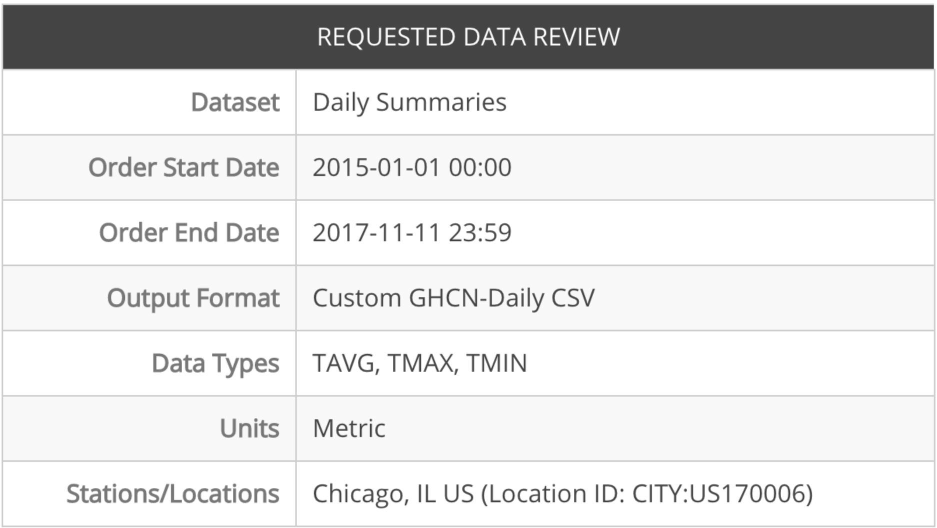
License	*The license for this dataset is unspecified*
Source Link	https://portal.chicagopolice.org/portal/page/portal/ClearPath

Fig 11: Meta Data Showing Updated Data

NCDC Archive of Global Historical Weather

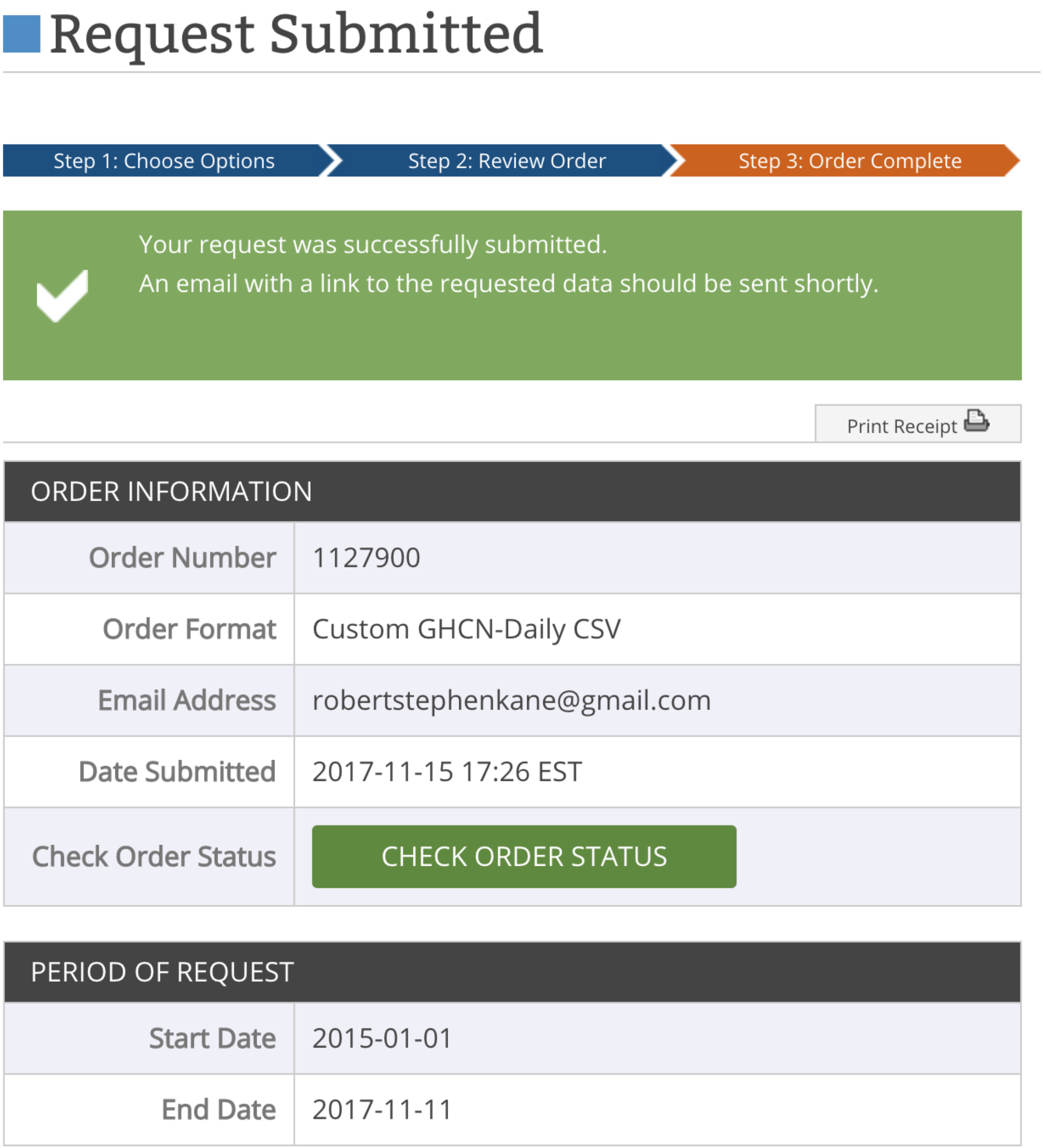
Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information. This data includes quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree-days as well as radar data and 30-year climate normals.

This data will be used in the correlation of crime statistics for the data warehouse and will account for various recordings across season, time of day and day of the week.



REQUESTED DATA REVIEW	
Dataset	Daily Summaries
Order Start Date	2015-01-01 00:00
Order End Date	2017-11-11 23:59
Output Format	Custom GHCN-Daily CSV
Data Types	TAVG, TMAX, TMIN
Units	Metric
Stations/Locations	Chicago, IL US (Location ID: CITY:US170006)

Fig 12: Digital Receipt for Requested Crime Data Set



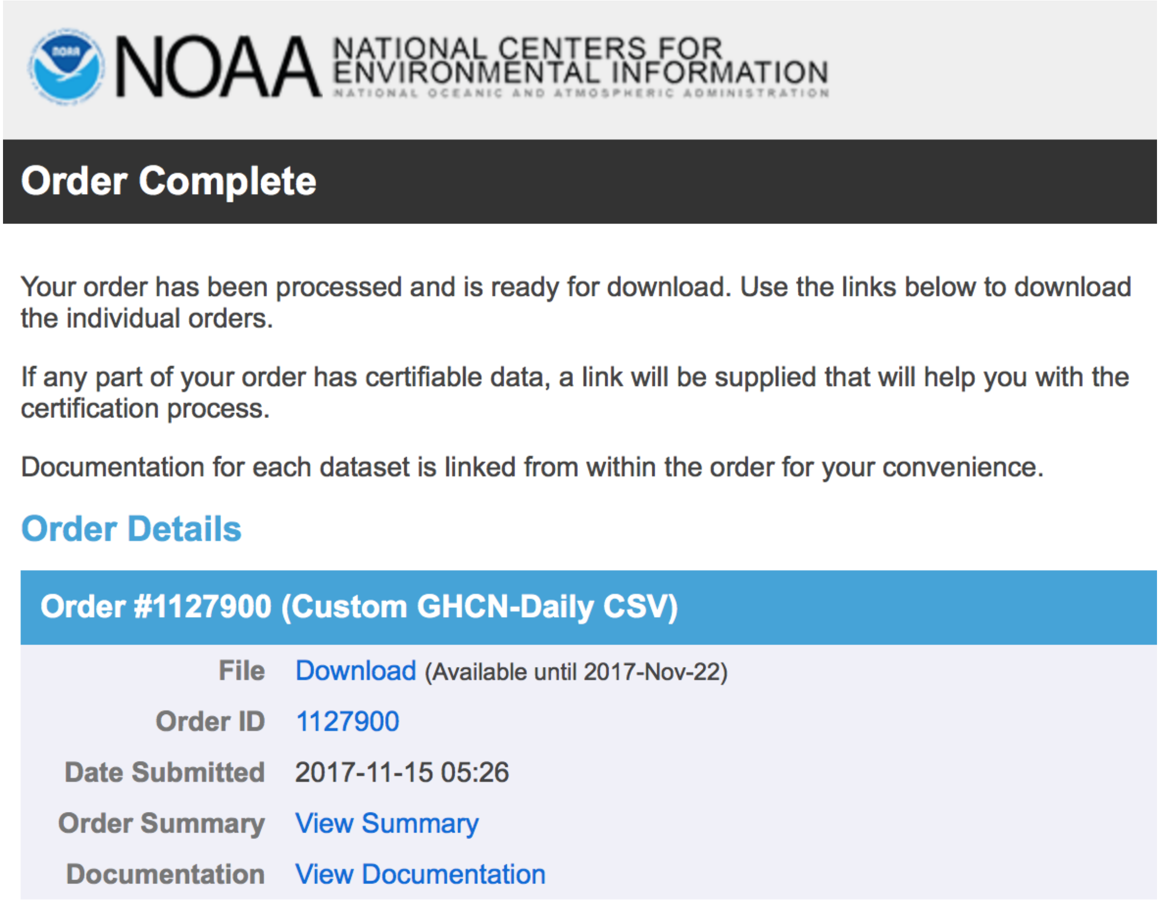
Request Submitted

✓ Your request was successfully submitted.
An email with a link to the requested data should be sent shortly.

ORDER INFORMATION	
Order Number	1127900
Order Format	Custom GHCN-Daily CSV
Email Address	robertstephenkane@gmail.com
Date Submitted	2017-11-15 17:26 EST
Check Order Status	CHECK ORDER STATUS

PERIOD OF REQUEST	
Start Date	2015-01-01
End Date	2017-11-11

Fig 13: Digital Receipt for Requested Weather Data



NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Order Complete

Your order has been processed and is ready for download. Use the links below to download the individual orders.

If any part of your order has certifiable data, a link will be supplied that will help you with the certification process.

Documentation for each dataset is linked from within the order for your convenience.

Order Details

Order #1127900 (Custom GHCN-Daily CSV)

File	Download (Available until 2017-Nov-22)
Order ID	1127900
Date Submitted	2017-11-15 05:26
Order Summary	View Summary
Documentation	View Documentation

Fig 14: CSV File Download Link for Requested Data

System Topology & Architecture

The diagram below is a high level overview of how the proposed system architecture and external components of this project will be combined and implemented to produce the various outputs and data analytics report. Once the project has been developed, users of the system will have the potential to explore the data in a meaningful and intuitive way. This will enable them to apply their own parameters, investigate correlations and produce the reports they require.

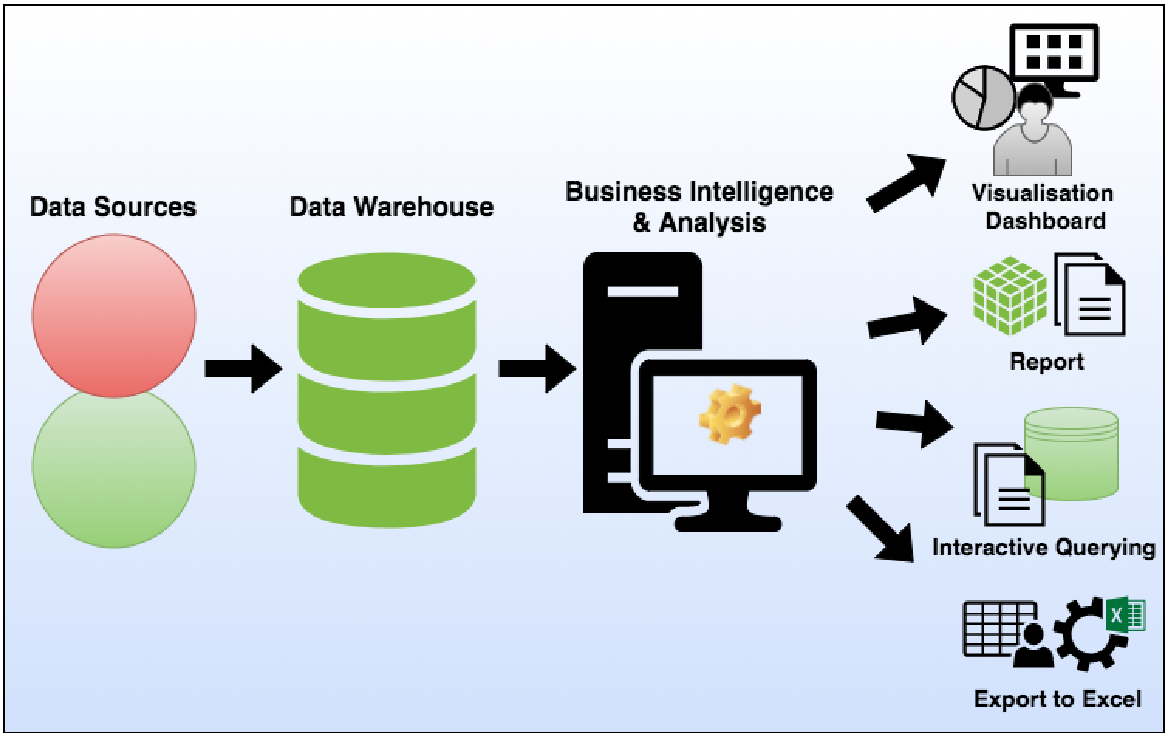


Fig 17: High Level Overview of the Components for the Project

The diagram below shows the system architecture. This demonstrates the system and project workflow from data sources and ETL processes to results presentation and report output.

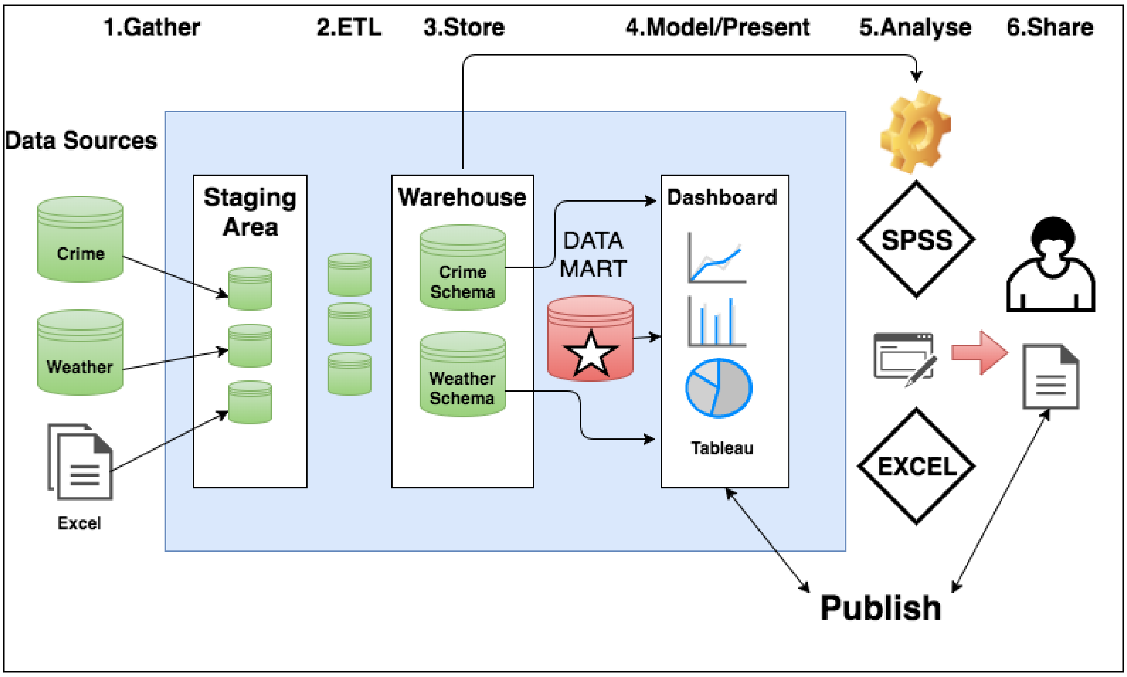


Fig 18: System Architecture Diagram

10.5. Testing

Testing is critical to the delivery of a successful application. The nature of this project would inherently involve testing at each step in development. This is to ensure the architecture of the data warehouse and integrity of the aggregated data queries for use in the analysis is correct and accurate. It also ensures the data being presented to the end user through the dashboard is coherent. An extensive critical evaluation of the analytics report produced will also be conducted to ensure the relevancy and accuracy of the conclusions and forecasts being made.

System Evolution & Further Development

Next Steps of Development

Now that the requirements have been gathered and data has been proved suitably workable for this project, the next steps are to push on with completing all necessary ETL processes and begin development of the data warehouse.

Once the data warehouse has been developed, I can connect it to a Tableau dashboard to visualize the results. I can then begin using the output generated by the system to conduct the statistical analysis.

System Evolution

The system could potentially expand to accommodate and display statistical analysis for more locations upon completion. The conceptual blueprints created and mapped out during the process of this project could be applied to more data sources. The system has the potential to evolve into a concept where the methodologies implemented allow for a hosted site that allow users to focus in on various locations around the world for similar data.

Commercial Aspects For the Results of this Project

Data mining has become big business in recent years. All information that leads to better-informed decisions is inherently valuable and thus commercially valuable. The results of this project could potentially be of great interest to those who have a practical application or need for the findings. Data is valuable and people are willing to pay for data that offers meaningful insight for their proposed solutions.