National College of Ireland

BSHTM4

Data Analysis

2020/2021

Darragh Brennan

17469074

x17469074@student.ncirl.ie

The effect of the climate on the spread of COVID-19 in the United States of America

Technical Report

Contents

Executive Summary

This report provides an analysis and evaluation on the effect of the climate on the spread of the coronavirus in the United States of America. The main datasets that were used for this analysis were “United States COVID-19 Cases and Deaths by State over time” (Data.Gov,2020), and a weather dataset that was found on the NOAA website (National Oceanic and Atmospheric Administration,2020). By analysing these datasets this report attempts to answer the question “Does the climate have a direct effect on the spread of COVID-19 in the United States of America?”.

Methods of analysis include data pre-processing in Excel, cluster analysis, interactive maps, linear regression, and normality tests in RStudio, various non-parametric tests in SPSS and graphs and plots that were created in Tableau. The results of the mentioned analysis methods can be found throughout this report.

The results of the analysis methods show that initially it did appear that the hotter the climate was the more contagious the COVID-19 virus was but after further investigation and as more data was gathered and analysed it then appeared not to be the case and that the virus was actually more contagious the colder the climate was. The report concludes that with the current amount of available data the climate alone does not appear to have a direct impact on the spread of the virus in the USA however more data would be required to give a definite answer.

The report also investigates the fact that the analysis conducted does have limitations. This analysis solely looks at the cases and deaths in each state by month and then looks at the average temperature of the given month in each state. This limits the report as population, population density, vaccine rates or any other factors are not considered in this analysis and because COVID-19 spreads by close contact there is more likely other factors that impact the spread of the virus more than climate. The fact the dataset only began recording data from January 2020 also limits this study as predictions and conclusions are more difficult to make which makes it more difficult to answer the research question as there is no real historic data to analyse.

1.0 Introduction

1.1. Background

This topic was chosen for investigation because of how relevant COVID-19 and climate change are. This analysis was carried out in the hope of being able to produce some valuable information and findings to see if there is in fact a link between climate and the spread of COVID-19 as there is still a lack of understanding when it comes to the COVID-19 virus. COVID-19 was firstly identified in December of 2019 in Wuhan, China. In March 2020, the COVID-19 outbreak was declared a pandemic. At the time of writing this there has been a total of over 158 million confirmed COVID-19 cases and over 3.29 million COVID-19 related deaths worldwide. In the USA alone there has been over 32.5 million confirmed cases and just over 580,000 deaths (World Health Organisation,2020).

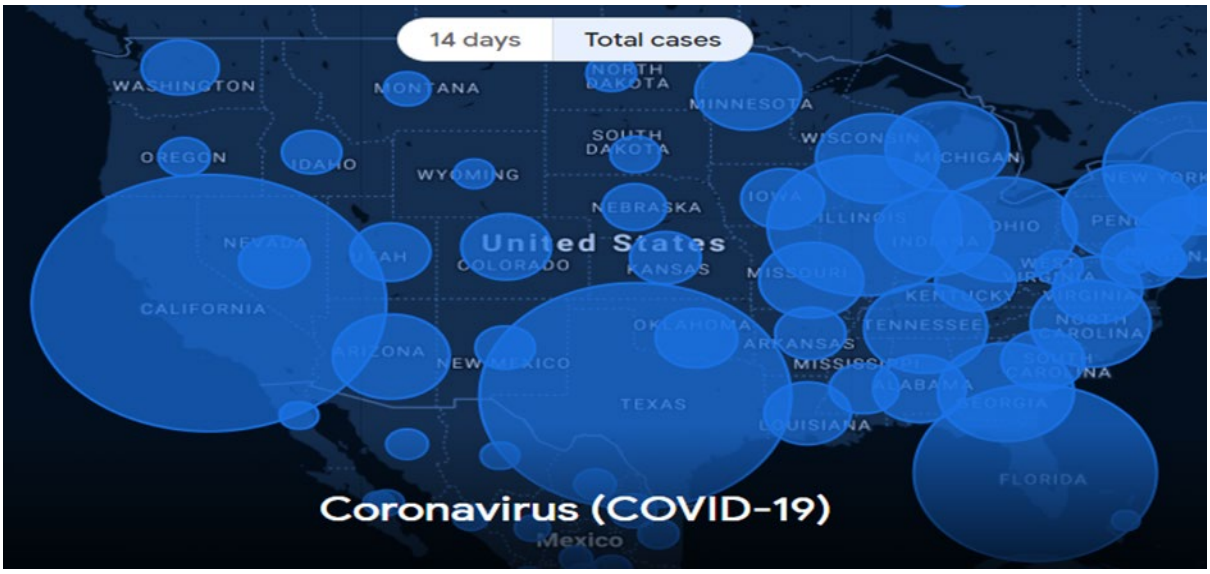


Figure 1: COVID-19 USA total cases map (Google News, 2020).

WHO (World Health Organisation) say that COVID-19 has 2 main modes of transmission, droplet transmission and direct contact with an infected person. “Droplet transmission occurs when a person is in close contact (within 1 m) of someone who has respiratory symptoms (e.g., coughing or sneezing) and is therefore at risk of having his/her mouth, nose or eyes exposed to potentially infective respiratory droplets.” They say “transmission may also occur through fomites in the immediate environment around the infected person. Therefore, transmission of the COVID-19 virus can occur by direct contact with infected people and indirect contact with surfaces in the immediate environment or with objects used on the infected person (e.g., stethoscope or thermometer)” (World Health Organisation,2020).

The reason for choosing the topic of “The effect of the climate on the spread of the coronavirus in the United States of America” was because when looking at the total cases map above it can be seen that the most affected areas in the US were Texas, Florida, and California, these three states are all considered to be warm states and looking at the least affected areas these were states such as Montana, North Dakota and Oregon which are all considered cold states. For these reasons, this topic was chosen for further investigation to see if there was any direct corelation between the climate of a state and COVID-19 cases/deaths.

This study aims to provide information that could help state governors plan for the coming months by analysing and trying to predict when the transmission of the virus is at its highest which could then help them implement guidelines or extra restrictions for certain time periods of the year. For example, if cases/deaths are at their highest point in June because of hot temperatures, then the governor of the state could enforce closures to reduce human interaction or make it mandatory to wear a mask and gloves during these high transmission months.

1.2. Aims/Objectives

The main objective of this project is to try and answer the research question which is “Does the climate have a direct effect on the spread of COVID-19 in the United States of America?”. This project aims to investigate data to see if there is enough evidence to conclude that the climate does or does not have an effect on the spread of the coronavirus in the United States of America. To successfully analyse the data and attempt to answer this question some important tasks had to be carried out.

Firstly, datasets that were most suitable for this research topic were acquired. 2 main datasets were chosen for this analysis, the first was an up-to-date COVID-19 cases/deaths in the USA dataset and the other was a dataset that contained the average temperature in each state by month.

Once the best datasets were chosen, they were then cleaned so that the results of the tests would be accurate, and they were also organised in a manner so that the various tests and visualisations could be carried out. Cleaning the datasets involved removing irrelevant values, dealing with missing values, ensuring there was no duplicated data, and getting rid of useless rows and columns. Organising the datasets involved subsetting and joining the values of the datasets in a specific way so that each of the tests could be carried out.

Next, the information contained in the datasets was analysed using various different tests and was turned into easy to interpret visualisations. Different non-parametric tests were carried out and visualisations such as interactive maps, bar charts, line graphs and cluster plots were created to analyse the datasets to try and extract as much information as possible to help answer the research question.

1.3. Technology

Various technologies were used to achieve the objective of this project. Technologies such as RStudio, Microsoft Excel, Tableau, SPSS and GitHub were used throughout this project.

R studio is the integrated development environment (IDE) that uses the R programming language. RStudio was used to select and pre-process the datasets and carry out different data mining techniques such as clusters analysis’, linear regression, normality tests and interactive maps.

Microsoft excel was used to create CSV files that contain the relevant information from the main and secondary datasets that were also downloaded in the CSV format. For this study it was important to ensure that the datasets were in the correct format so they could be easily imported into RStudio, SPSS and Tableau for the different graphs, plots, clusters, and tests to be created and carried out. Excel was also used to combine and pre-process most of the datasets for easier and quicker analysis in R studio.

Tableau was used to create an interactive dashboard containing various graphs and plots relevant to the study so viewers can clearly see the findings in the form of interactive visualisations.

SPSS was used to carry out the majority of the statistical analysis which was made up of various non-parametric tests such as the Wilcoxon Rank Test, Mann-Whitney U Test and the Friedman Test. These tests were carried out to produce some interesting figures and descriptive statistics that helped to provide an insight into the topic of this project. A time series analysis was also carried out in SPSS to try and make some predictions of future cases and deaths.

GitHub was used for version control purposes of the code and files that are used throughout this project. All files related to this project are stored on GitHub.

1.4. Structure

The structure of the remainder of this document is as follows:

Data: In this section, each dataset that was used throughout this study will be discussed and where each dataset was found, what they consist of and what they were used for will be highlighted.

Methodology: In this section, the data mining framework that was followed when carrying out this analysis will be explained and each of the steps within the framework will be discussed so that this analysis could be easily reproduced on the chosen datasets for this analysis or on different datasets.

Analysis: In this section, the different approaches used for this analysis will be described, the reason for choosing each approach will be explained and any of the decisions made whilst carrying out the different analysis techniques will be justified.

Results: In this section, the results of the various data mining techniques, tests and visualisations that were carried out will be presented and explained.

Conclusions: In this section, a conclusion will be made regarding the research question "Does the climate have a direct effect on the spread of COVID-19 in the United States of America?". Also, the advantages, disadvantages, strengths and limitations of the project will be highlighted and discussed.

Further development or research: In this section, possible extensions or new topics that could be added to improve this analysis will be highlighted.

2.0 Data

A total of 4 datasets were used to carry out this analysis, the two main datasets being “United States COVID-19 Cases and Deaths by State over Time” found on the Data.Gov (Data.Gov,2020) website and a state-wide weather dataset found on the NOAA website (National Oceanic and Atmospheric Administration,2020). The two secondary datasets used were a state abbreviations dataset found on the World Population Review website (World Population Review,2000) and a state latitude and longitude dataset that was found on the Google developers’ website (Google,2020).

The first main dataset was chosen because this project required a dataset that broke down the number of COVID-19 cases/deaths by state since the beginning of the pandemic. At the time of writing, the dataset contains data on 50 states from the 22nd of January 2020 to the 6th of May 2021. However, for this project only the 48 main land states were analysed. Before cleaning and pre-processing was carried out the dataset consisted of 15 columns and over 28,000 rows, some of the columns include things such as date, state, total cases, new cases, total deaths, new deaths and more. This dataset is still being updated daily as new case and death numbers are being confirmed and released. This dataset was perfect for the purpose of this project and provides the best information to help produce the most accurate and relevant results.

The next main dataset that was required for this project was the weather dataset so that the average temperatures for the different states could be obtained. This dataset was difficult to choose as there were many different options, for this reason each of the possible options were explored by carrying out some exploratory analysis techniques in RStudio and Excel. Each of the possible datasets were joined in Excel and then some graphs and tests were ran in RStudio to see which dataset was most suitable for this project.

After carrying out the exploratory analysis techniques it was decided that NOAA’s weather dataset was the most suitable for this project. This dataset contained average temperatures for every month in each state, that were calculated using past state weather reports from 1895 to 2019. This dataset consisted of 8 columns and over 72,000 rows, these columns include things such as location, date, values, mean temperature and more. The reason an average temperature dataset was chosen and not just the previous year’s state temperature reports was because the average temperature dataset would allow for more accurate predictions for the coming years and rules out the chance of the results being negatively impacted by once off temperatures. For these reasons, this dataset was chosen for this project.

The two secondary datasets used for this project were only acquired to aid the interactive map creation. The state abbreviations dataset and the latitude and longitude dataset were required so that markers for each state could be added to the interactive maps.

Each of these datasets were then combined in many different ways so the various tests, visualisations and data mining techniques could be carried out and analysed to see if there was any evidence to help answer the research question.

3.0 Methodology

To successfully analyse the data the Knowledge Discovery in Databases (KDD) methodology was followed. The KDD methodology consists of four main steps which are selection, pre-processing, extraction, and evaluation.

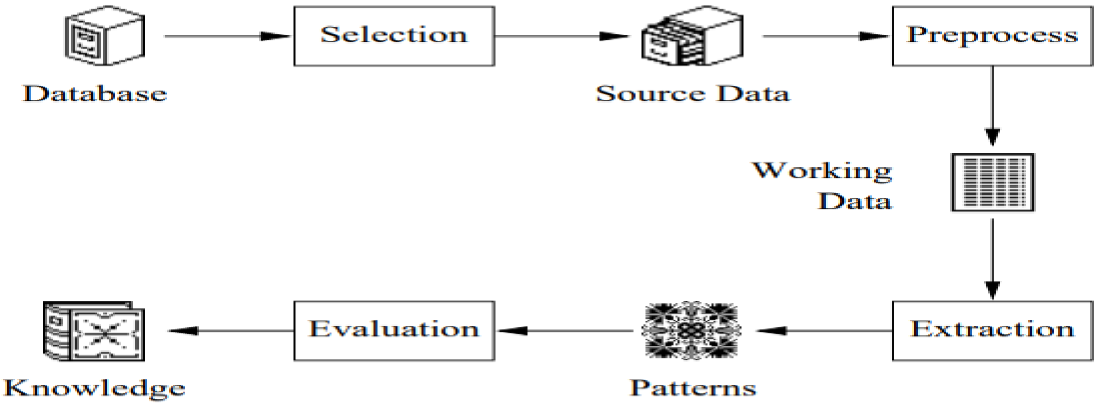


Figure 2: KDD methodology steps (Williams & Huang, 1996).

Selection: “Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection” (Rajput,2019). Firstly, multiple datasets that were relevant to this study were located online and downloaded. Once all the datasets were downloaded, they each had to go through a series of exploratory analysis techniques to find out which datasets would be best for this study. The most suitable datasets were identified and the relevant data within them was selected and segregated using Microsoft Excel and RStudio. The different columns and rows within the datasets had to be carefully selected and segregated into multiple CSV files to allow all of the different tests to be carried out and the different visualisations to be created. Once the selection process was completed the source data was then ready to be pre-processed.

Pre-processing: “The purpose of the Pre-processing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in later stages” (Williams & Huang, 1996). The data was pre-processed using RStudio and Microsoft Excel. Some pre-processing techniques that were carried out were the removal of irrelevant rows and columns, the renaming of columns, data subsetting and some basic aggregation and sampling. If the data is not pre-processed correctly visualisations and calculations of statistics can be negatively impacted, and what should be simple tasks can become difficult down the line. In terms of data pre-processing there was not much that needed to be done other than the techniques mentioned above as the datasets that were selected for this study were complete datasets with no null values and a low number of dimensions. Once the pre-processing techniques were carried out and the working datasets were prepared, the extraction methods could then be executed.

Extraction: The extraction phase is when a variety of software’s are used to explore different types of patterns within the working datasets. RStudio, SPSS and Tableau were the main software’s used for exploring and extracting the different patterns within the data for this study. RStudio was used to perform data mining techniques such as clustering and linear regression, it was used to perform statistical tests such as the Shapiro-Wilk normality test and it was also used to create visualisations. SPSS was used to perform a time series analysis and various non-parametric tests such as the Wilcoxon Rank test, Mann-Whitney U test and the Friedman test. Lastly, Tableau was used to create different visualisations such as plots and graphs and display them in the form of an interactive dashboard. All of the results of these extraction methods and techniques could then be analysed and evaluated.

Evaluation: Evaluation takes place once all the extraction methods are complete. Once the extraction methods were carried out the outputs of the techniques and tests were then analysed, and any interesting findings and patterns within the data were identified. The results were then evaluated in an attempt to answer the research question "Does the climate have a direct effect on the spread of COVID-19 in the United States of America?".

4.0 Analysis

The first thing that had to be done was the datasets that would be most suitable for the purpose of this project had to be selected. The 2 main datasets that were selected for this project were both downloaded in the CSV format from online websites. For this reason, Microsoft Excel was used to initially analyse and ensure that the datasets were suitable for this project and contained the required information. Microsoft Excel was also used for pre-processing, filtering the data and creating multiple modified datasets so that the different data mining techniques and statistical tests could be carried out and analysed further.

Next, the programming language that was most suitable for this study had to be chosen. The 2 languages that were considered for this study were R and python. After researching both languages, it was found that R was built as a statistical language whereas Python provides a more general approach when it comes to data science. For this reason, it was decided that due to the statistical power of R it would be the best programming language to use for this study.

Once the best programming language was selected the most suitable IDE for the language then had to be chosen. Because it was decided that the R programming language would be used, the IDE most suitable for R is RStudio. This IDE contains a large number of packages and allows many different statistical tests, and many exploratory and data mining techniques to be executed. For this study RStudio was used to firstly explore the different dataset options by applying exploratory analysis techniques such as creating different plots and graphs and trying to carry out some statistical tests to identify which datasets would be most suitable for this study. Once the best datasets were identified, RStudio was then used to pre-process and subset the data to create various CSV files that could then be used to create visualisations and carry out different statistical tests in all of the software's used for this study.

Once the data was pre-processed and segregated in RStudio and Microsoft Excel, RStudio was then used to create 2 interactive maps, create a linear regression model, carry out a cluster analysis, and run a Shapiro-Wilk test and create Q-Q plots to see if the data was normally distributed.

SPSS was then used to carry out different non-parametric statistical tests. The first test that was carried out was the Wilcoxon-Rank test, this test was used to compare 2 related samples. The second test that was carried out was the Mann-Whitney U test, this test was used to compare differences between two independent groups and the third and final test that was carried out was the Friedman test, this test was used to test for differences between groups. SPSS was also used to perform a time series analysis to try and make predictions based on the patterns found in the data. The reason SPSS was chosen for the statistical analysis over RStudio is because SPSS is designed specifically for statistical analysis which makes it more user friendly and more powerful than RStudio when it comes to performing statistical tests.

For the last section of the analysis, it was decided that an interactive dashboard would work best to display the various different graphs and plots. Although it is possible to create an interactive dashboard in RStudio it is not the most effective software to use when it comes to making an interactive dashboard as the customisation options are limited, and the data has to be in a very particular format making it time consuming and difficult to create. After researching more effective software's that could be used to create an interactive dashboard the two best options that were found were Apache Superset and Tableau. However, for this study Tableau was identified as the best software to use as it allowed CSV files to be imported directly where the data could then be turned into a huge variety of different graphs and plots with high levels of customisation before being added to an interactive dashboard.

5.0 Results

Throughout this section the results of the different data mining techniques, statistical tests and visualisations will be presented and explained in an attempt to answer the research question.

5.1 RStudio

5.1.1 Interactive maps

The first set of techniques and tests were carried out using RStudio. Firstly, 2 interactive choropleth maps were created, the first map displayed the number of COVID-19 cases in each state and the average temperature for each month and the second map displayed the number of COVID-19 related deaths in each state and the average temperature for each month.

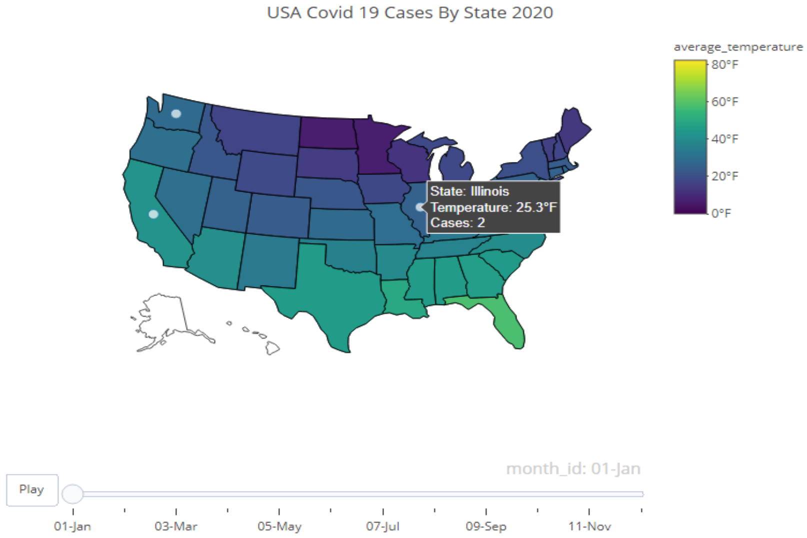


Figure 3: COVID-19 cases across the USA in January 2020.

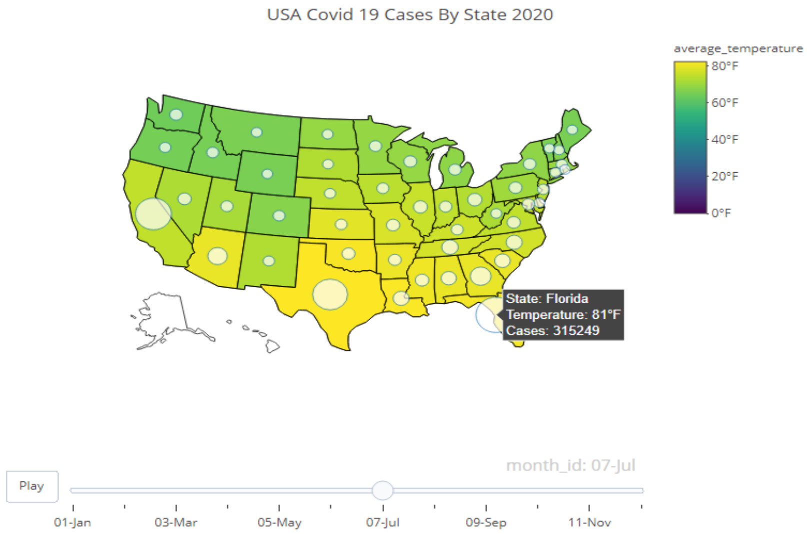
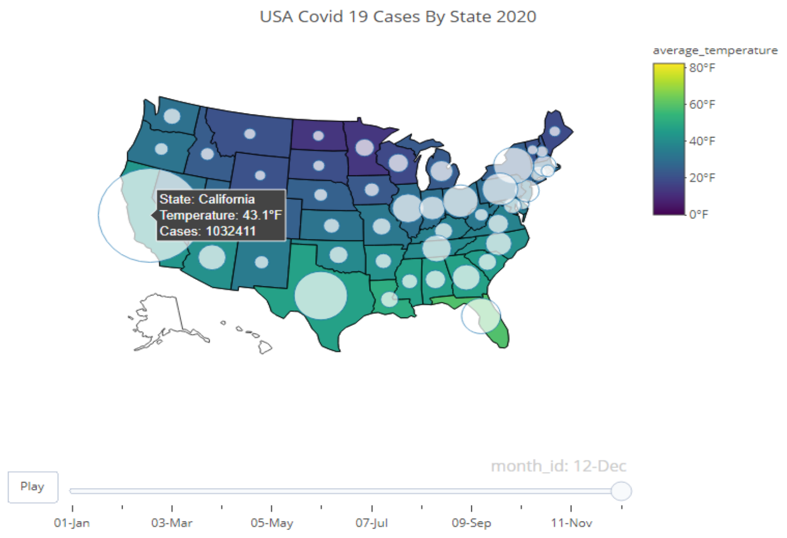


Figure 4: COVID-19 cases across the USA in July 2020.



State: California
Temperature: 43.1°F
Cases: 1032411

Figure 5: COVID-19 cases across the USA in December 2020.

The figures above are images taken from the first interactive choropleth map that displays the number of COVID-19 cases in each state and the average temperature for each month. For the purpose of this report only 3 of the 12 months were selected. The white markers on each of the states represents the number of cases in that state and the colour of the state represents the average temperature in degrees Fahrenheit for the given month. The bigger the marker is the higher number of cases within that state and the more yellow the state is the hotter the average temperature is for that month.

Looking at figure 3 above which shows the number of COVID-19 cases across the USA in January 2020 it can be seen that the average temperature in each state is quite low and that there are virtually no white markers, this was because the severity of the virus was still unknown and testing for the virus was practically non-existent. Large scale testing only began once the virus was declared a national pandemic in March 2020.

Looking at figure 4 then which shows the number of COVID-19 cases across the USA in July 2020, the map is a lot more yellow as the temperatures are higher because it is peak summertime. It also does appear that the number of cases are higher in the states that have a higher average temperature which may indicate that the virus is more contagious in hotter climates.

However, looking at figure 5 above which shows the number of COVID-19 cases across the USA in December 2020 it is seen that the cases are at an all time high across every state. This may indicate that in fact the virus is more contagious in colder climates than in hotter climates. To analyse this further a bar chart was created using Tableau.

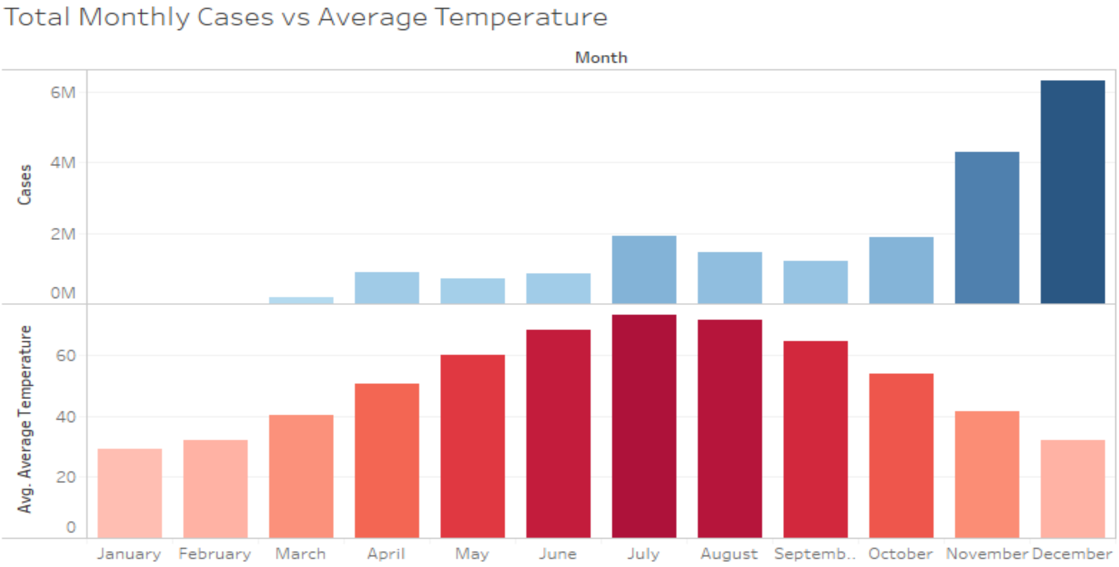


Figure 6: Total monthly cases vs the average state-wide temperature by month.

Looking at figure 6 above, if it were the case that the hotter the climate is the more contagious the virus is then it would be expected that as the temperature increases so would the total number of cases which does seem to be the case in April, May, June and July. It would then be expected that as the average temperature decreases again so would the total number of COVID-19 cases which does appear to happen in August and September but then in October the cases jumped back up to roughly the same numbers as in July and then there was a huge increase of cases in November and December. Because of this huge increase of cases in the colder months this may indicate that the virus is actually more contagious in colder climates than in hotter climates.

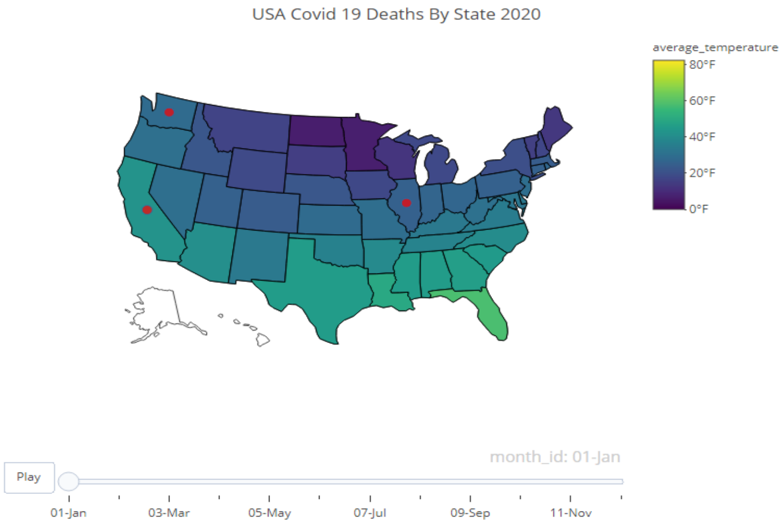


Figure 7: COVID-19 related deaths across the USA in January 2020.

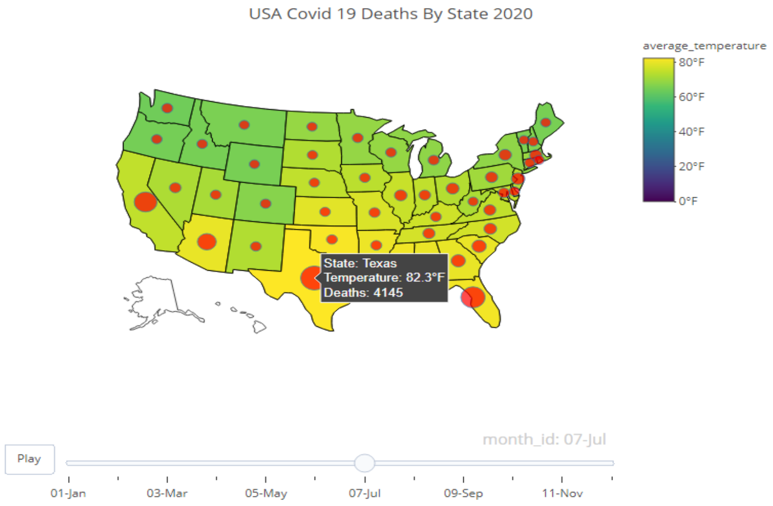


Figure 8: COVID-19 related deaths across the USA in July 2020.

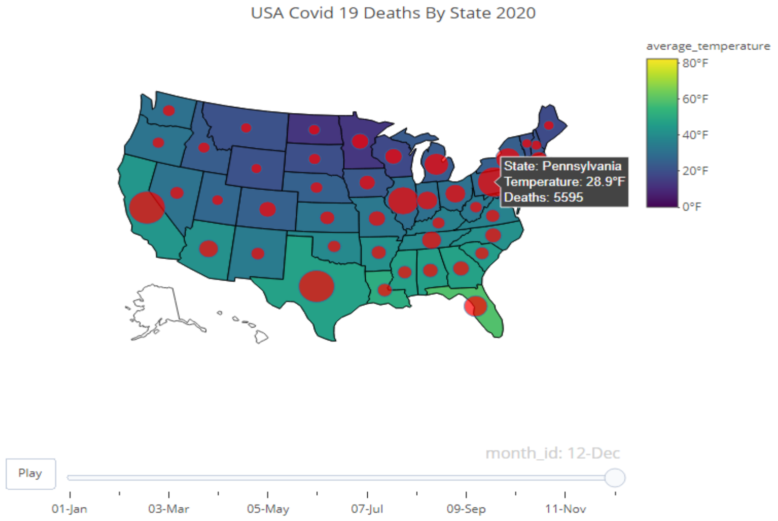


Figure 9: COVID-19 related deaths across the USA in December 2020.

The figures above are images taken from the second interactive choropleth map that displays the number of COVID-19 deaths in each state and the average temperature for each month. The red markers on each of the states represents the number of COVID-19 related deaths in that state and the colour of the state represents the average temperature in degrees Fahrenheit for the given month. The bigger the marker is the higher number of related deaths within that state and the more yellow the state is the hotter the average temperature is for that month.

Looking at the above figures the maps correlate with the cases maps showing virtually no deaths in January 2020 then as the number of cases increase in July so do the number of deaths and finally then when cases are at their all time high in December the number of deaths are also at their highest.

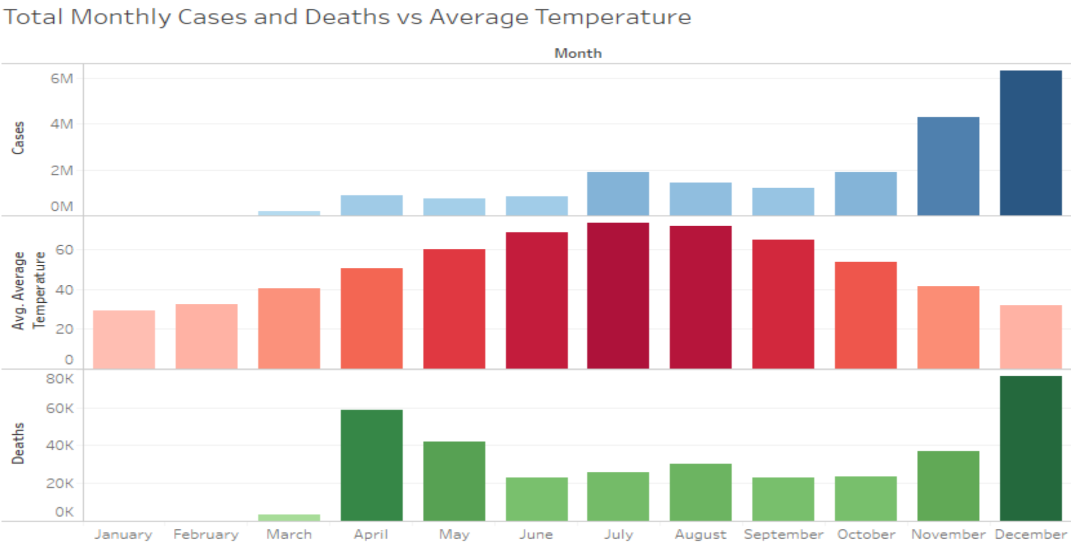


Figure 10: Total monthly cases and deaths vs the average monthly temperature across the USA.

Looking at the graph above helps to identify the correlation between cases and deaths. It can be seen in most instances that the higher the number of cases the higher the number of deaths. Looking at the above figures it appears that the climate does not affect the number of deaths which was expected however the map was created to analyse the relationship between the number of cases and the number of deaths.

5.1.2 Cluster analysis

The second analysis technique that was carried out in RStudio was a K-means cluster analysis. This analysis was carried out to try and identify which states were impacted the most by COVID-19 in general and then in the top three months that the cases were at their highest which were July, November, and December. Each cluster that was created grouped the data based on the number of cases/deaths and the average temperature.

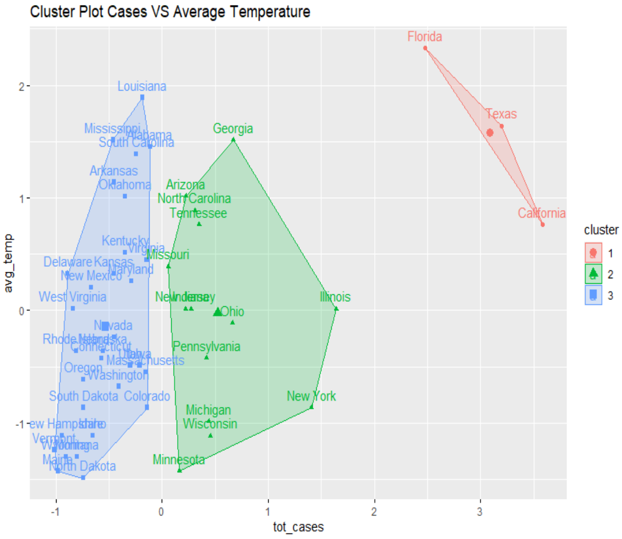


Figure 11: K-means cluster plot grouping the data by the total number of cases and the average annual temperature of each state.

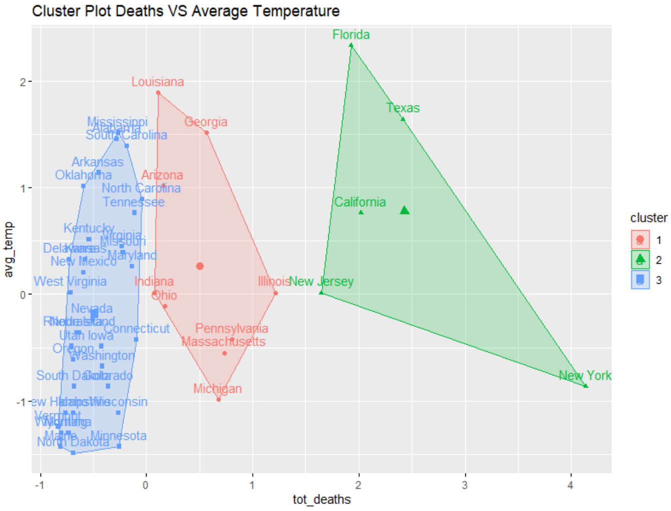
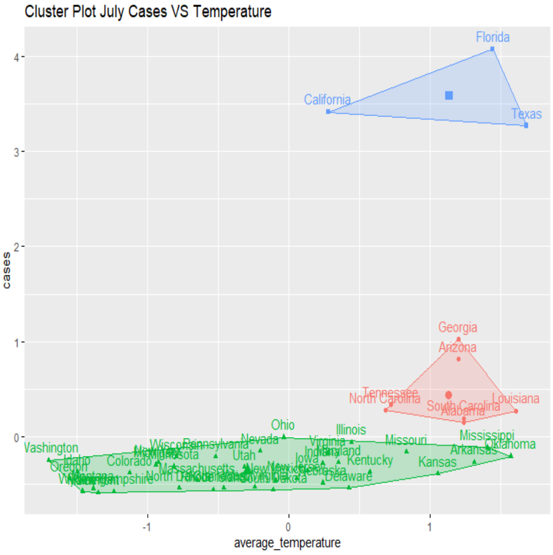


Figure 12: K-means cluster plot grouping the data by the total number of deaths and the average annual temperature of each state.

Looking at figure 11 above which groups the data based on the total number of cases and the average annual temperature of each state it can be seen that there are 3 clusters. Cluster 3 represents the states that had a low number of total cases, cluster 2 represents the states that had a medium number of cases and cluster 1 represents the states that had a very high number of cases. Looking at this figure what was interesting is that in cluster 2 and 3 the states are spread quite consistently across the y axis which represents the average annual temperature of each state and looking at cluster 1 although the states are all in the higher end of the average annual temperature, there is still states in the other clusters that are hotter than 2 of the 3 states in cluster 1. This indicates that climate does not appear to have a direct effect on the spread of the virus.

Looking at figure 12 then which groups the data by the total number of deaths and the average annual temperature of each state it can be seen that there are 3 clusters again. This time cluster 3 represents the states that had a low number of deaths, cluster 2 represents the states that had a high number of deaths and cluster 1 represents the states that had a medium number of deaths. This figure shows that each cluster contains roughly the same states and the states that had the highest number of cases also have the highest number of deaths. This again shows the relationship between the cases in each state and the death rate.

For a better insight, the months that the cases were at their highest which were July, November and December were then analysed.



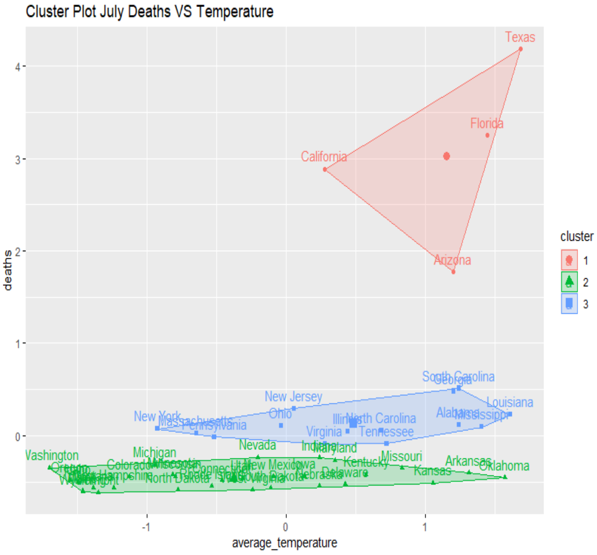
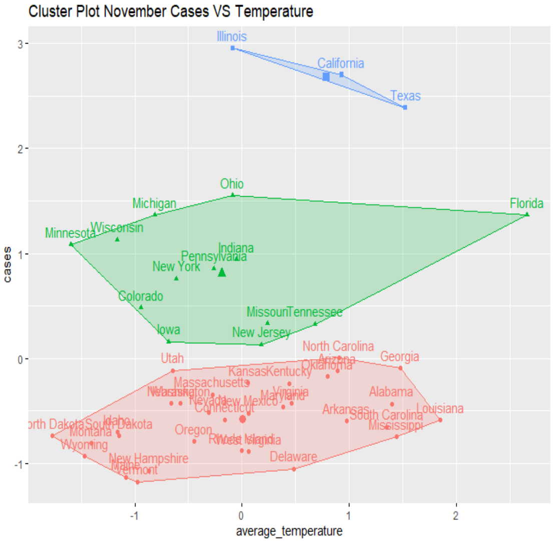


Figure 13: July COVID-19 cases and average temperature.

Figure 14: July COVID-19 related deaths and average temperature.



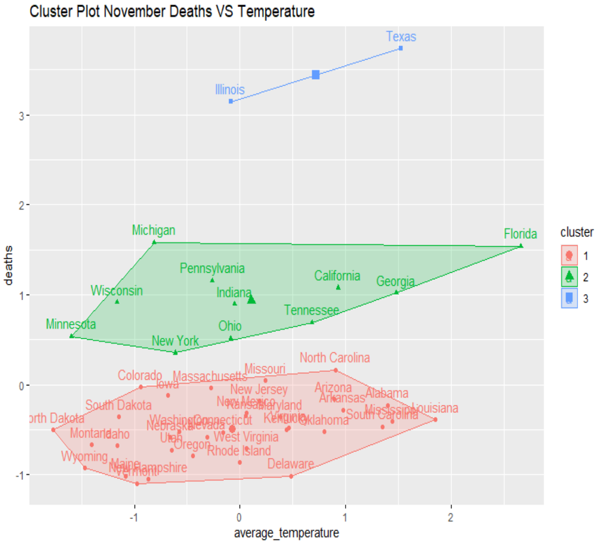
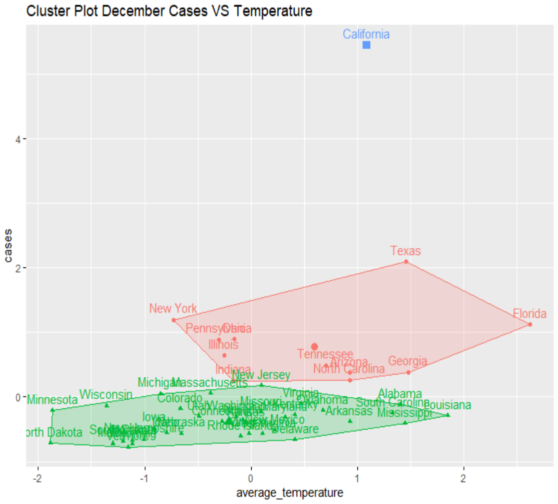


Figure 15: November COVID-19 cases and average temperature.

Figure 16: November COVID-19 related deaths and average temperature.



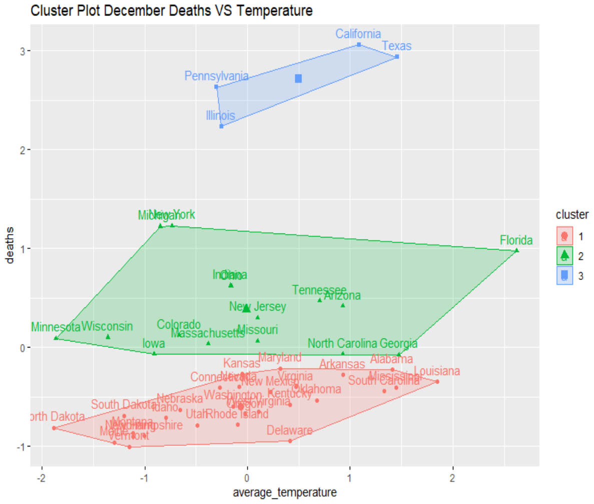


Figure 17: December COVID-19 cases and average temperature.

Figure 18: December COVID-19 related deaths and average temperature.

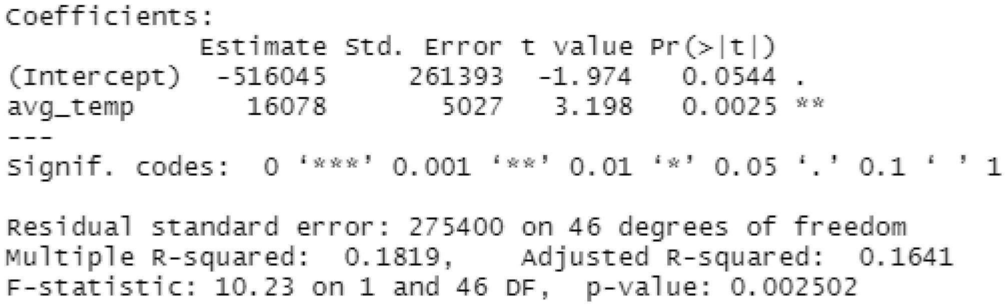
Starting with figure 13 and 14 which look at the month July, it can be seen in figure 13 that the states where the cases are highest are found in cluster 1 and 3. These states all have quite high average temperatures as well which may indicate that the virus is more contagious in a warmer climate. However, cluster 2 also contains states that have very high average temperatures but nowhere near the same number of cases as those in cluster 3 which would then indicate that the climate does not directly impact the spread of the virus. The clusters in figure 14 relatively match those found in figure 13 which again indicates that the number of deaths is higher when there are a higher number of cases.

Moving onto figure 15 and 16 then which look at the month of November, it can be seen in figure 15 that clusters are spread out a lot more and there in no real obvious pattern. Because the states within the cluster are so spread out across both the x and y axis and the central point of each cluster is relatively close to 0 on the x axis which represents the average temperature, this indicates that the average temperature did not impact the number of cases therefore, it appears that the climate did not directly impact the spread of the virus in November. Again, in figure 16 the clusters are nearly identical to those in figure 15 which indicates that the number of deaths is higher when there are a higher number of cases.

Lastly, looking at figure 17 and 18 which look at the month of December, in figure 17 it is seen that cluster 1 and 2 are relatively evenly spread out across the x axis and again the central points of the two clusters are close to 0. Cluster 3 only contains 1 state which is California where the cases were extremely high but the average temperature for California is not very high in December. Both of these findings indicate again that climate does not directly impact the spread of the virus. Figure 18 which looks at the number of COVID-19 related deaths in this case the clusters do not match with the clusters that were found in figure 17. It would be expected that because California had the highest number of cases by a big margin that California would also have the highest number of deaths by a large margin however, Texas, Pennsylvania and Illinois all do not have that many less deaths. Although this is not related to the research question it is still an interesting finding.

5.1.3 Linear regression

The next technique that was carried out in RStudio was linear regression. Linear regression was carried out to see if the average temperature could be used as a predictor variable when it comes to attempting to predict the total number of cases and deaths.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -516045     261393  -1.974   0.0544 .
avg_temp       16078       5027   3.198   0.0025 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 275400 on 46 degrees of freedom
Multiple R-squared:  0.1819,	Adjusted R-squared:  0.1641 
F-statistic: 10.23 on 1 and 46 DF,  p-value: 0.002502
```

Table 1: Summary table of a linear regression model where the intercept is the total number of COVID-19 cases.

Looking at the above summary table it shows that the average temperature is significant when it comes to predicting the dependent variable with a value of p < 0.05. This means that when it comes to predicting the dependent variable which in this case is the total number of cases that the independent variable average temperature can be used. However, this model has a very low adjusted R-squared value which indicates that there is noisy data that may impact the predictive power of this model. Overall, the model above has a p-value of < 0.05, this means that the model is significant and therefore the null hypothesis should be rejected, and this proves that the model is capable of predicting the dependent variable total cases.

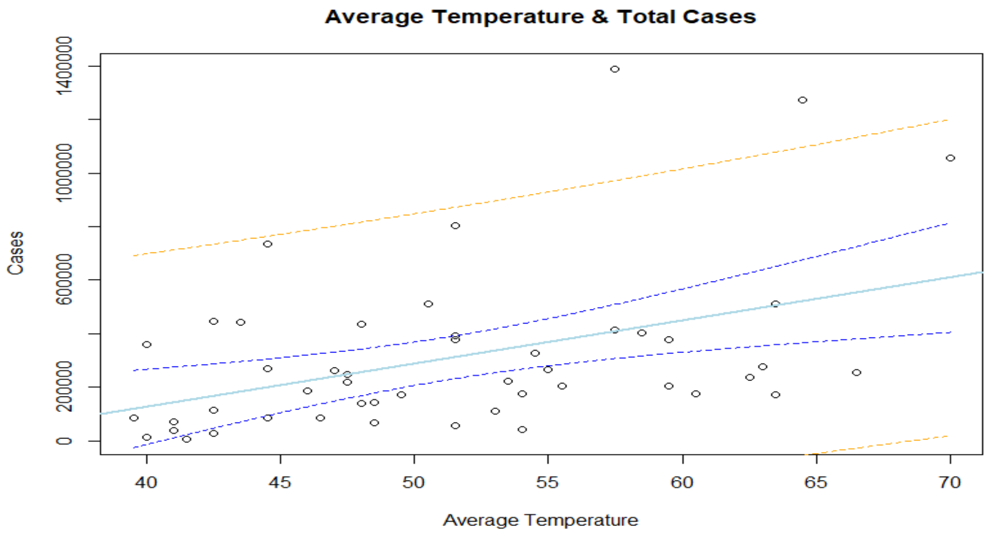
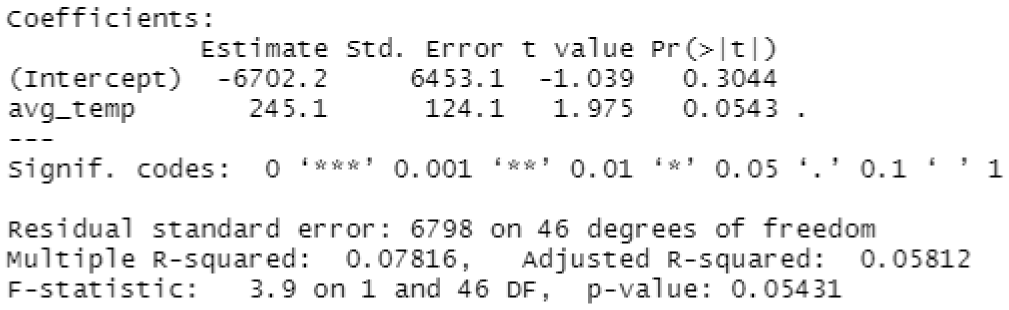


Figure 19: Total cases and average temperature linear regression model plotted.

In the figure above the light blue line represents the prediction line, the two blue dotted line represent the upper and lower confidence intervals and the two yellow lines represent the upper and lower prediction intervals. Looking at this figure it can be seen that the prediction line for this model is increasing, this indicates that as the average temperature increases so do the total number of cases. This would mean that the virus is more contagious in warmer climates than in colder climates. However, it is evident that there is a lot of noisy data which explains the low adjusted R-squared value seen in the previous summary table, this means that the prediction line may not be accurate and may actually fall any where between the two confidence intervals and for this reason the model cannot be relied on.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6702.2     6453.1  -1.039   0.3044
avg_temp       245.1      124.1   1.975   0.0543 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6798 on 46 degrees of freedom
Multiple R-squared:  0.07816,	Adjusted R-squared:  0.05812
F-statistic:   3.9 on 1 and 46 DF,  p-value: 0.05431
```

Table 2: Summary table of a linear regression model where the intercept is the total number of COVID-19 deaths.

The above summary table shows that the average temperature is not significant when it comes to predicting the dependent variable with a value of p > 0.05. This means that the dependent variable which is the total number of deaths cannot be predicted by this independent variable. This model also has an even lower adjusted R-squared value than the last model which indicates that there is a lot of noisy data that may impact the predictive power of this model. Overall, this model above has a p-value of > 0.05, this means that the model is not significant and therefore the null hypothesis should be accepted, as this model is not capable of predicting the dependent variable total cases.

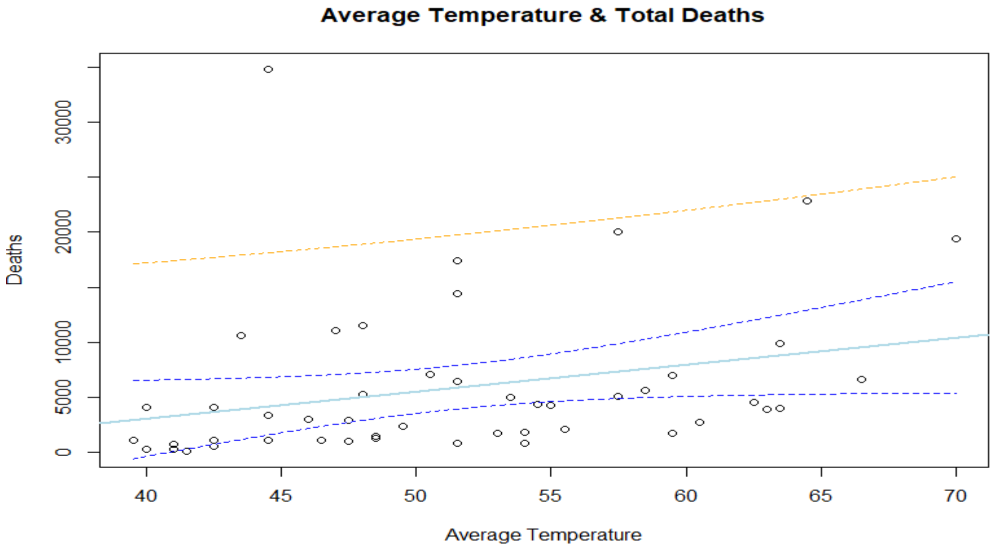
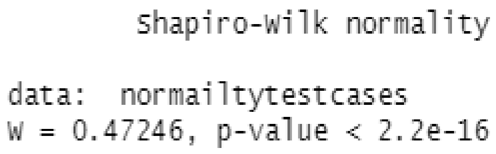


Figure 20: Total deaths and average temperature linear regression model plotted.

The figure above has the same lines as figure 19. Looking at this figure it can be seen that there is a lot of noisy data again and the prediction line is also much flatter especially when it is taken into consideration that the prediction line can fall anywhere between the two confidence intervals. Because the null hypothesis was accepted this means that there were no patterns identified within the data, this indicates that the climate does not impact the number of COVID-19 related deaths.

5.1.4 Testing for normality

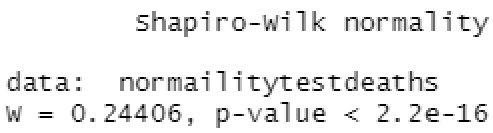
The last test that was carried out in RStudio was a Shapiro-Wilk normality test. This test was carried out to see if the data was normally distributed or not so that it could be identified whether parametric or non-parametric tests had to be carried out. The Shapiro-Wilk test was carried out on the new cases column and the new deaths column. The reason the Shapiro-Wilk test was carried out on these two columns is because these are the 2 most important columns for this study and contain the data that the analytical tests would be carried out on. The Shapiro-Wilk test requires a sample of no more than 5000 rows so 5000 rows were taken from each column and the test was conducted.



```
	Shapiro-Wilk normality test

data:  normailtytestcases
W = 0.47246, p-value < 2.2e-16
```

Table 3: Shapiro-Wilk test new cases column results.

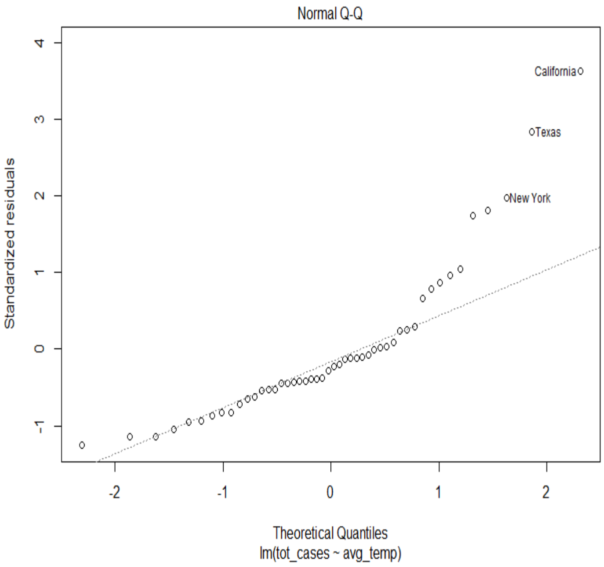


```
	Shapiro-Wilk normality test

data:  normailitytestdeaths
W = 0.24406, p-value < 2.2e-16
```

Table 4: Shapiro-Wilk test new deaths column results.

The results of both tests carried out show p < 0.05, this means that the null hypothesis is rejected and the data is not normally distributed. For further confirmation each of the linear regression models were also plotted.



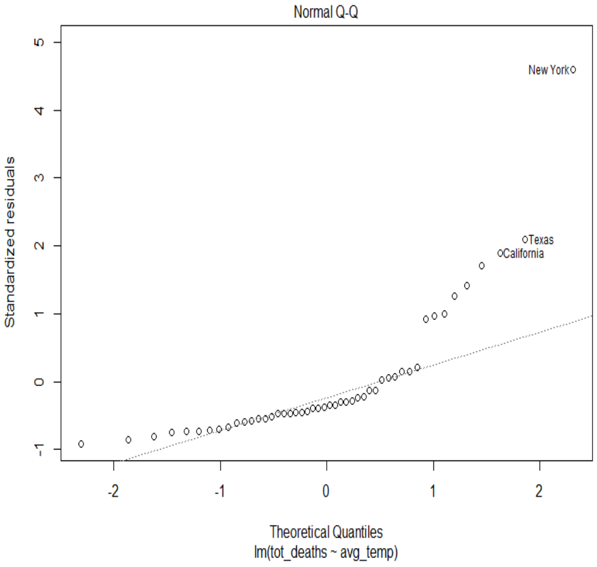


Figure 21: Q-Q plot of the total cases linear regression model.

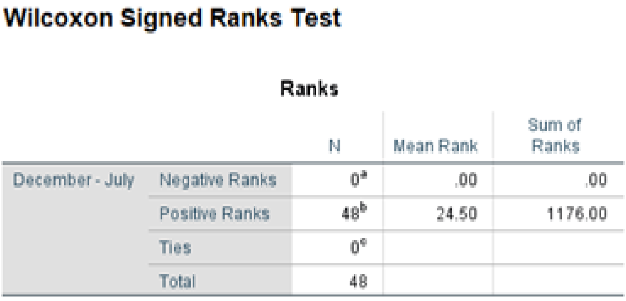
Figure 22: Q-Q plot of the total deaths linear regression model.

Looking at the two Q-Q plots above it can be seen that the majority of the data points do not fall on the line of normality. This further proves that the data is not normally distributed. Because the data is not normally distributed this means that non-parametric tests will have to be carried out on the data.

5.2 SPSS

5.2.1 Wilcoxon Signed Rank test

SPSS was used to carry out different statistical tests. The first test that was carried out was the Wilcoxon Rank test. Because the data is not normally distributed a dependent t-test was not appropriate to use on this data, for this reason a Wilcoxon Rank test was carried out instead. This test was carried out to compare the number of cases in each state in the hottest month in the USA which is July and the coldest month which is December to see if there is any relation between the number of cases when the climate is hot climate or when the climate is cold.



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
December - July	Negative Ranks	0[a]	.00	.00
	Positive Ranks	48[b]	24.50	1176.00
	Ties	0[c]		
	Total	48		

Table 5: Wilcoxon signed rank test results.

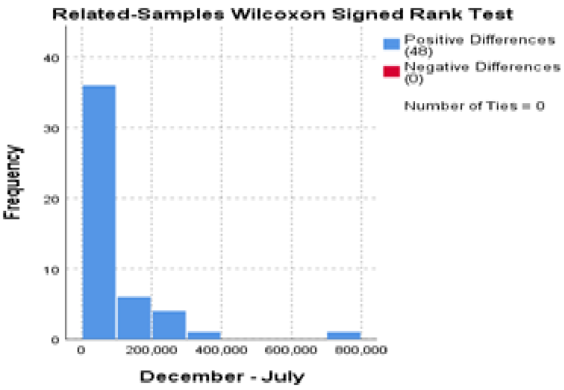
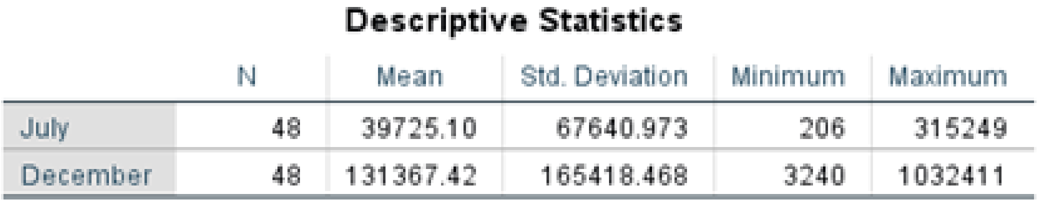


Figure 23: Positive and negative ranks graph.

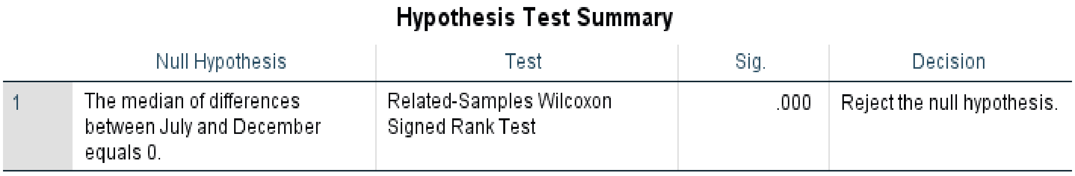


Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
July	48	39725.10	67640.973	206	315249
December	48	131367.42	165418.468	3240	1032411

Table 6: Descriptive statistic results of the Wilcoxon signed ranks test.

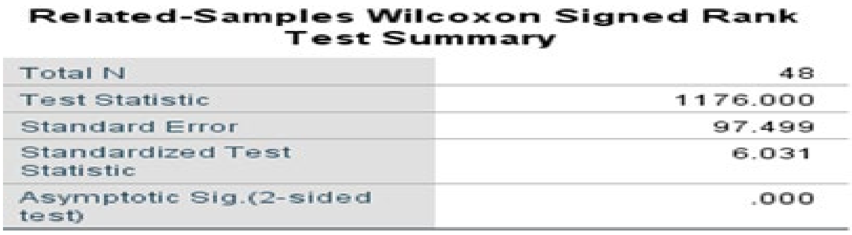
The negative ranks value represents the number of times the number of cases are lower in December than in July and the positive ranks value represents the number of times the number of cases are higher in December than in July. Each rank represents one state and there are 48 ranks because only the 48 main land states were analysed for this study. The ranks show that in every state the number of cases were higher in December than in July. Looking at all the descriptive statistics in table 6 above there is a significant difference between all of the values that show the number of cases in December were way higher than the number of cases in July. Although it is quite clear that there is a significant difference between the 2 months just to ensure that this was the case the Wilcoxon Signed Rank test summary table was analysed.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between July and December equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

Table 7: Hypothesis test summary of the Wilcoxon signed rank test.



Related-Samples Wilcoxon Signed Rank Test Summary

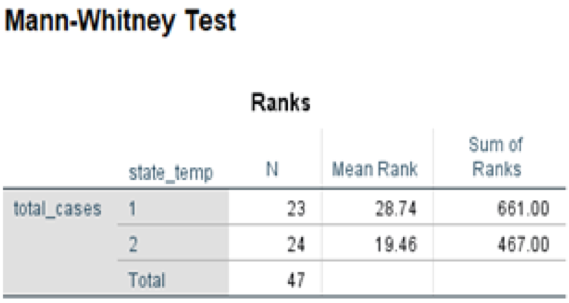
Total N	48
Test Statistic	1176.000
Standard Error	97.499
Standardized Test Statistic	6.031
Asymptotic Sig.(2-sided test)	.000

Table 8: Wilcoxon Signed Rank test summary table.

The results of this test indicate that the July ranks were statistically significantly higher than the December ranks, Z = 1176, $p < 0.05$, therefore the null hypothesis should be rejected as the median of differences between July and December do not equal to 0, and the alternate hypothesis is accepted. This means that it can be concluded that the cases were higher in December in every state and there was a significant difference between how many cases there were in July and how many cases there were in December. Because the number of cases are higher in December in every state this indicates that COVID-19 may be more contagious in colder climates.

5.2.2 Mann-Whitney U test

The Mann-Whitney U test was used to compare the differences between the number of cases in states that had an average annual temperature greater than and less than the mean average annual temperature which was 51.3°F.



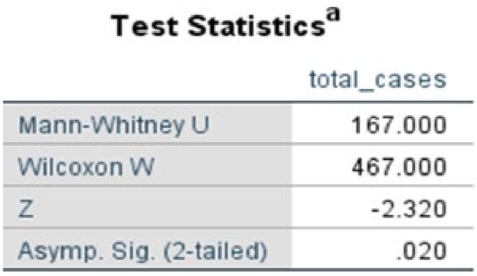
Mann-Whitney Test

Ranks

	state_temp	N	Mean Rank	Sum of Ranks
total_cases	1	23	28.74	661.00
	2	24	19.46	467.00
	Total	47		

Table 9: Mann-Whitney U test ranks.

Looking at the table above, 1 contains the states that had an average annual temperature greater than 51.3°F and 2 contains the states that had an average annual temperature less than 51.3°F. The N value shows the number of states contained within each group so within group 1 there are 23 states and within group 2 there are 24 states. The reason only 47 states are being tested is because the state that was equal to the mean average annual temperature was removed as it did not fit in either group. Looking at the mean rank and the sum of ranks values it can be seen that even though there is an extra state in the group below the mean average annual temperature the values are still lower than the values of the group above the mean average annual temperature. This suggests that the total number of cases was higher in the states with a warmer climate and the cases were lower in states with a colder climate.



Test Statistics[a]

	total_cases
Mann-Whitney U	167.000
Wilcoxon W	467.000
Z	-2.320
Asymp. Sig. (2-tailed)	.020

Table 10: Mann-Whitney U test statistics.

Looking at the results in the test statistics table above, with a U value of 167 and a p value < 0.05 the null hypothesis is rejected, and the alternate hypothesis is accepted, this means that it can be determined that there is a significant difference between the number of cases in states that have a warmer climate and the states that have a colder climate.

5.2.3 Friedman test

The Friedman Test was used to highlight the differences between the number of cases and see if the number of cases was evenly distributed across the 2 overall hottest states which are Florida and Louisiana, the 2 overall coldest states which are Minnesota and North Dakota, and the 2 states that are in the middle which are Indiana and New Jersey.

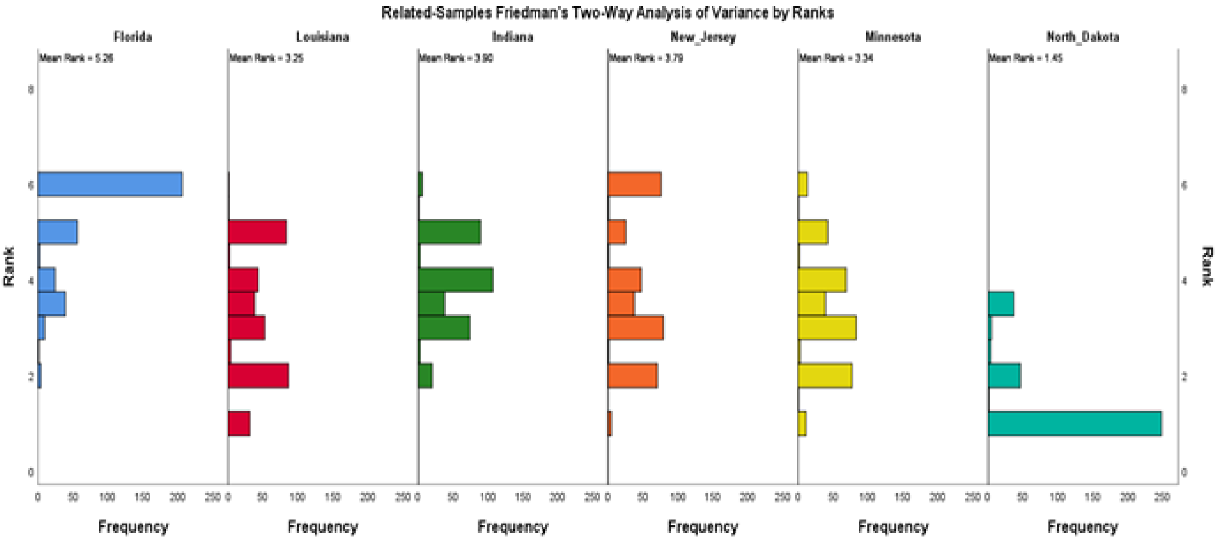
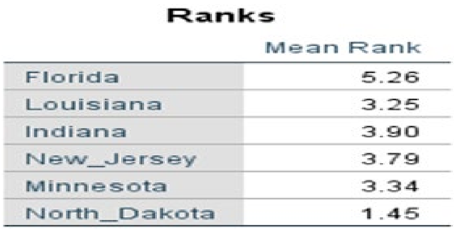


Figure 24: Frequency graphs of the different states.



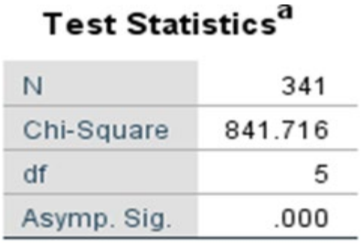
Ranks

	Mean Rank
Florida	5.26
Louisiana	3.25
Indiana	3.90
New_Jersey	3.79
Minnesota	3.34
North_Dakota	1.45

Table 11: Friedman test ranks.

Looking at the above ranks results the higher the mean rank value is the higher the number of cases is in that state and the lower the value is the lower the number of cases is in that state. The state with the highest mean rank is Florida with a value of 5.26, the state with the second highest mean rank is Indiana with a value of 3.90 and the state with the third highest mean rank is New Jersey with a value of 3.79. This means that Florida had the highest number of cases followed by Indiana and then New Jersey. Looking at the other 3 states however what is interesting is that although Louisiana is one of the top two hottest states it has a lower mean rank value than Minnesota which is one of the two coldest states. If the climate did have a direct effect on the spread of COVID-19 it would be expected that the majority of the cases would be found in the two hottest states or in the two coldest states however that is not the case for this test.

By analysing figure 24 above it can also be seen that of the 6 states the cases within the states are relatively evenly distributed in 4 of them however to figure out if the cases are evenly distributed across all of the states the Friedman test statistics have to be calculated.



Test Statistics[a]

N	341
Chi-Square	841.716
df	5
Asymp. Sig.	.000

Table 12: Friedman test statistics.

The test statistic results for the Friedman Test above indicate that there was a statistically significant difference in the number of cases across all of the states, χ2(5) = 841.716, p < 0.05. Therefore, the distribution of the number of cases are not the same across every state and the null hypothesis should be rejected. In other words, this means that the number of cases were higher in some states which was also identified when analysing the ranks in table 11.

5.2.4 Time series analysis

A time series analysis was carried out to try and predict the number of COVID-19 cases there would be in the next month which at the time of writing is May 2021. The time series analysis was carried out on the state with the hottest mean average annual temperature which is Florida, the coldest mean average temperature which is North Dakota and the state in the middle which is Indiana. To evaluate the time series analysis the time series analysis was firstly carried out on the month of January 2021 to predict the number of cases and then the results were compared with the actual case numbers in each of the states. The first state that was analysed was Florida.

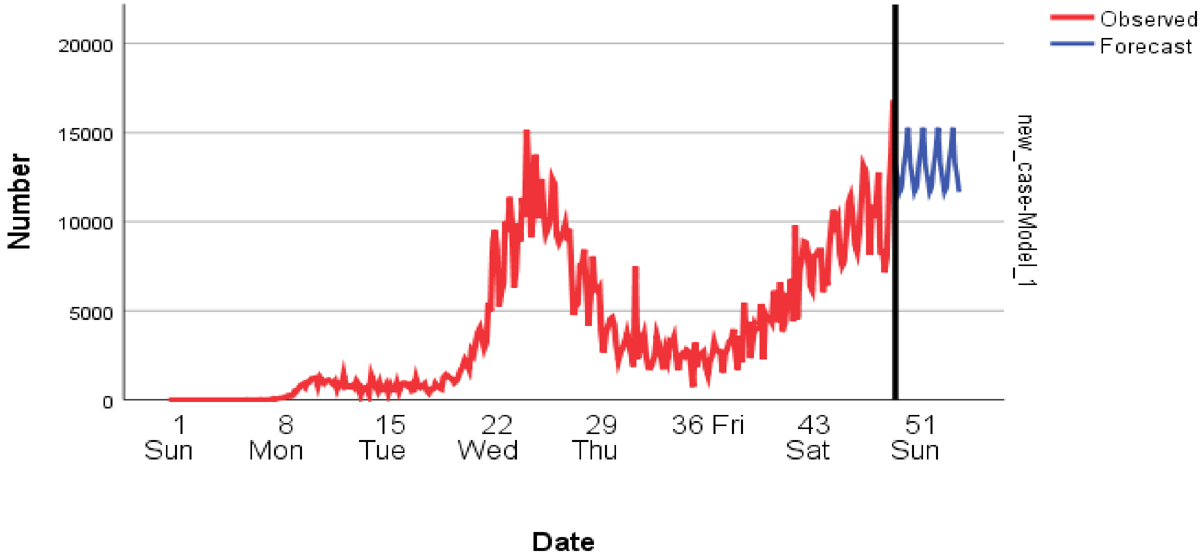


Figure 25: Time series analysis predicting case numbers in Florida for January 2021.

The time series analysis above was created using the expert modeller that is built into SPSS. This model was created to test the time series model to see how accurate it was by attempting to predict the number of cases there would be in January 2021 in Florida. The red line represents the actual data, and the blue line represents the predicted values. This model predicts that the cases will not continue to rise in Florida and the number of cases will level out to between about 11,000 to 15,000 a day. To test this model the actual number of cases in January in Florida were analysed.

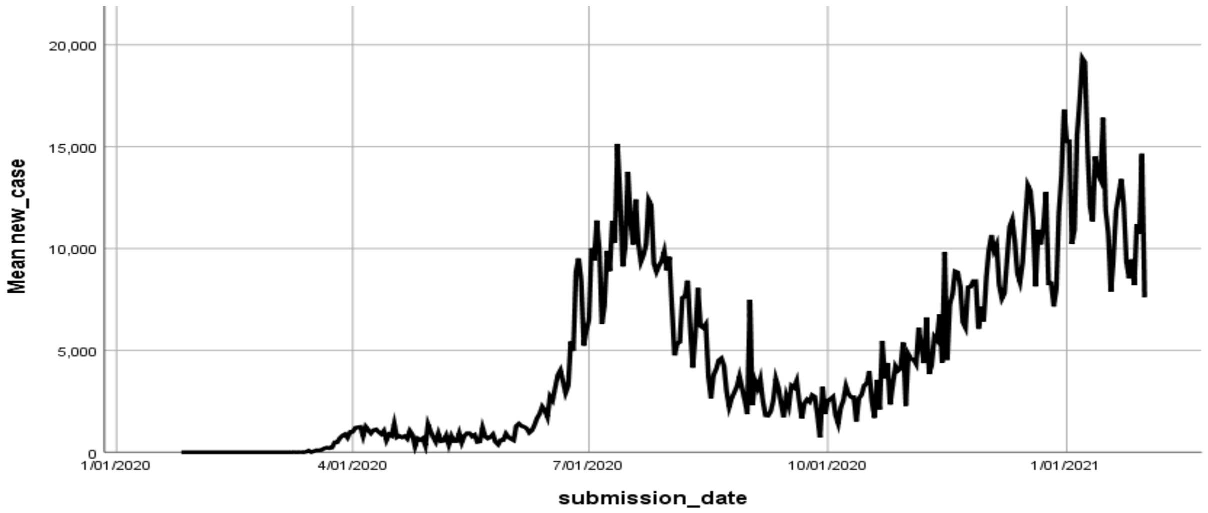


Figure 26: Graph displaying the number of cases in Florida in 2020 and January 2021.

In figure 26 above it can be seen that the cases do not increase which is what the time series model in figure 25 predicted and the number of cases do appear to be levelling out, and although the actual case numbers are between about 7,500 and 15,000 the model did a reasonably good job at predicting the number of cases. The expert modeller was then used to try and predict the number of COVID-19 cases there would be in May 2021.

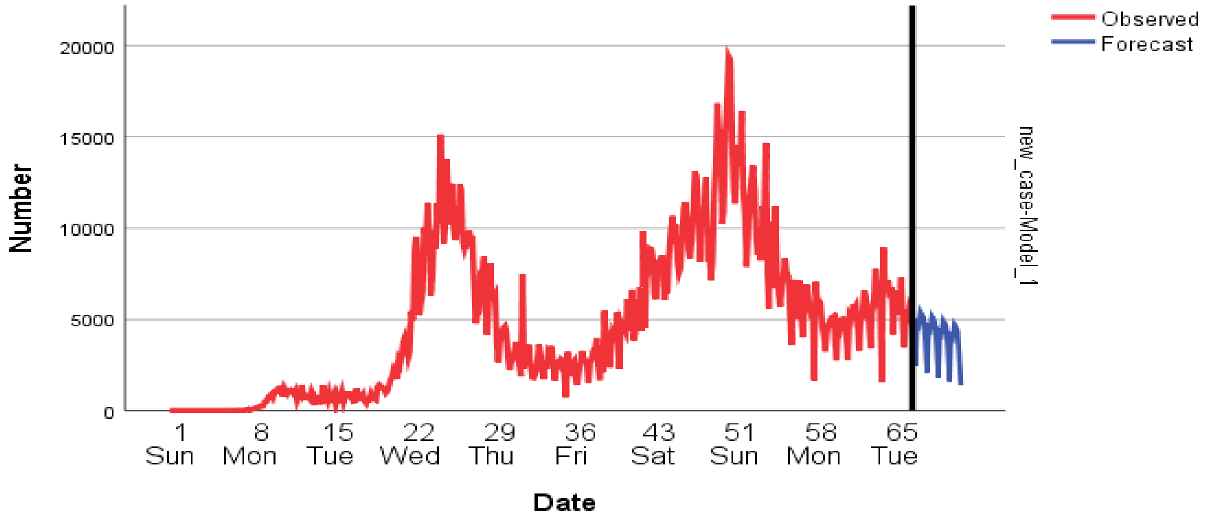
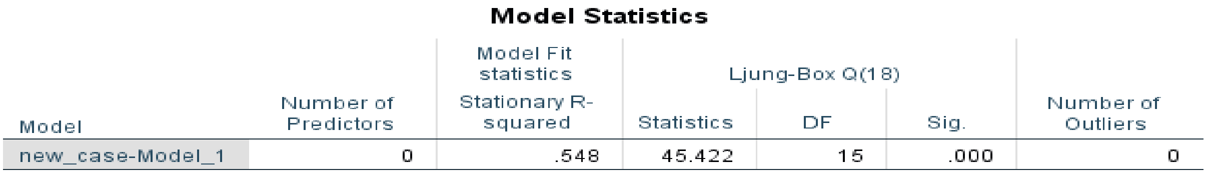


Figure 27: Time series analysis predicting case numbers in Florida for May 2021.

The blue line of the time series model that was created above is the predicted case numbers in Florida in May 2021. The model predicts that the number of cases will continue to drop well below 5,000 cases per day. To further analyse the accuracy of this model the model statistics summary table was analysed.



Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
new_case-Model_1	0	.548	45.422	15	.000	0

Table 13: Florida time series model statistics.

The model statistics above show that there is a p-value < 0.05, therefore, the null hypotheses should be rejected, and this model is significant and capable of predicting the number of cases based on the data provided. The model also has quite a high R-squared value of 0.548, although the optimum R-squared value is between 0.75 and 0.9, the R-squared value for this model shows that there is certainly no overfitting, and the model is well capable of predicting the variable which in this case is the number of cases in Florida in May 2021. The next state that was analysed was Indiana.

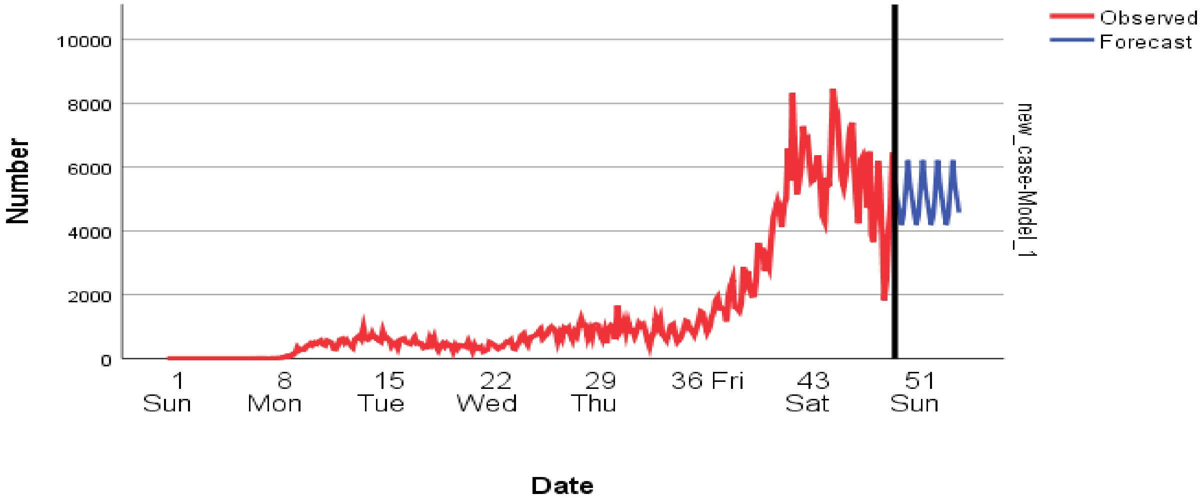


Figure 28: Time series analysis predicting case numbers in Indiana for January 2021.

This time series analysis was also created using the expert modeller built into SPSS to firstly try and predict the number of cases in January 2021 in Indiana to evaluate the model before predicting unknown values. Looking at the above model it can be seen that the model is predicting that the cases level out and stay between roughly 6,000 and 4,000 cases per day throughout January 2021. To test the accuracy of this model the actual results were analysed.

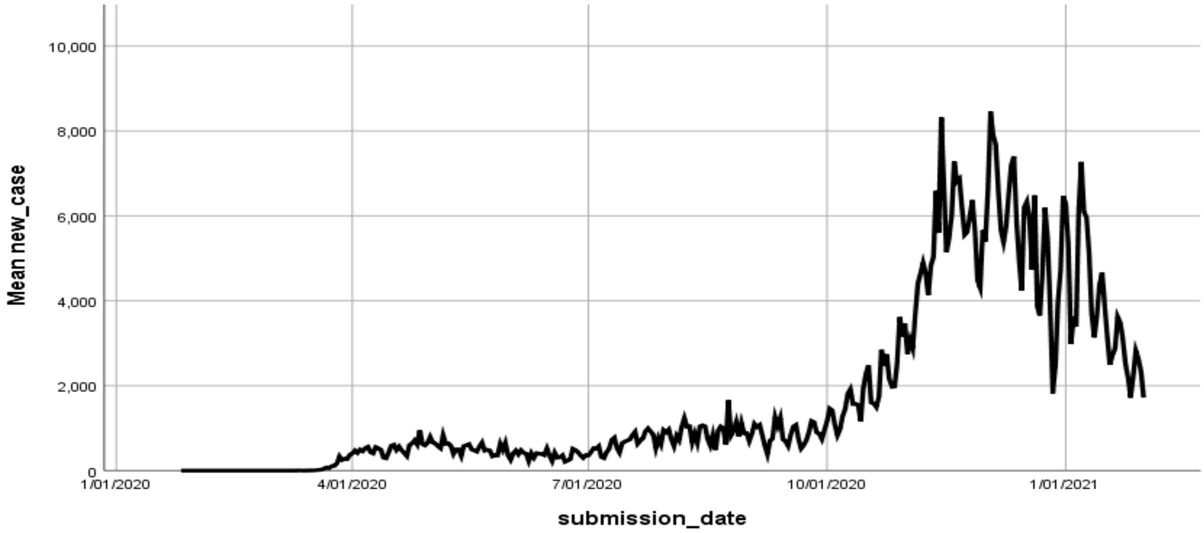


Figure 29: Graph displaying the number of cases in Indiana in 2020 and January 2021.

Looking at figure 29 above this time it can be seen that the cases spike at the beginning of January 2021 before sharply decreasing to between 4,000 and 2,000 cases per day. The model in figure 28 did not predict that the cases would drop to these low numbers, this may be because there is a high level of noisy data and the model was unable to identify any real patterns within the data. To analyse this further the model was used to attempt the number of cases in Indiana in May 2021 and the model statistics summary table was then analysed.

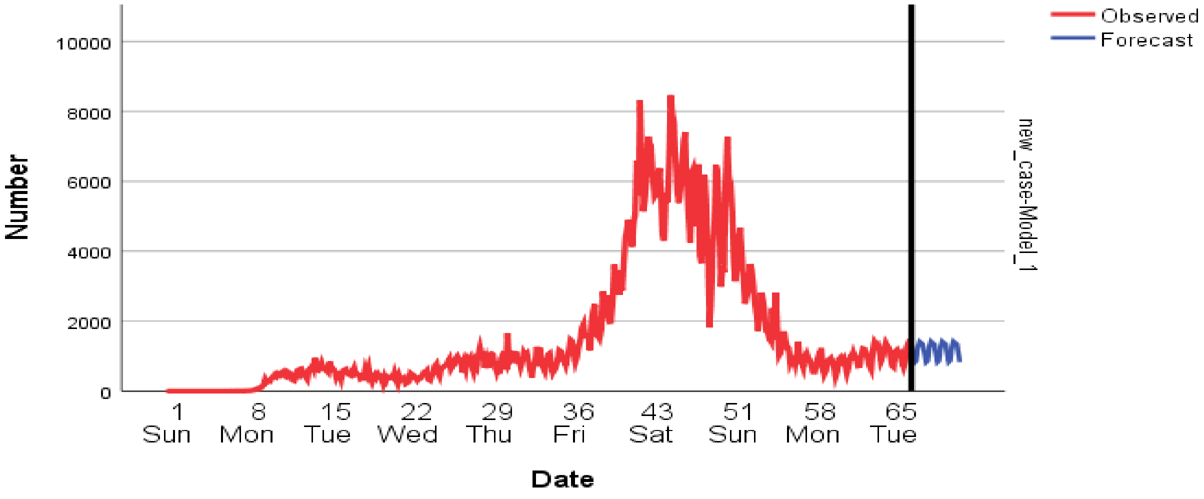
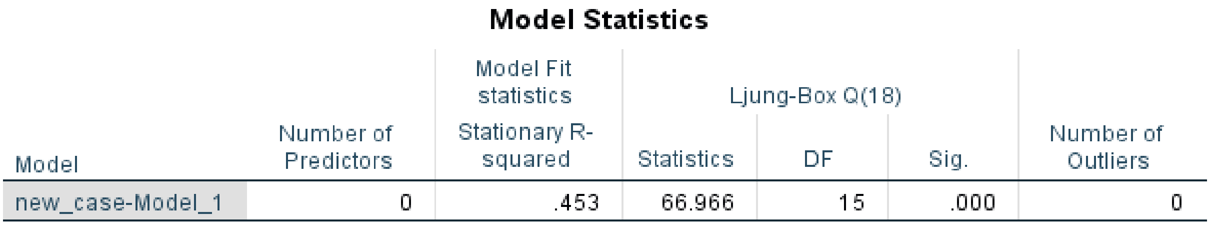


Figure 30: Time series analysis predicting case numbers in Indiana for May 2021.

The time series model above predicts that the case numbers in Indiana will remain between roughly 750 and 1,250 for the month of May with no spikes or increases. However, because the accuracy of this model was off in figure 28 this predicted value may not be reliable therefore the statistics summary table was analysed for a better insight.



Model Statistics

		Model Fit statistics	Ljung-Box Q(18)			
Model	Number of Predictors	Stationary R-squared	Statistics	DF	Sig.	Number of Outliers
new_case-Model_1	0	.453	66.966	15	.000	0

Table 14: Indiana time series model statistics.

The model statistics above show that there is a p-value < 0.05, which again means that the null hypotheses should be rejected, and this model is significant and capable of predicting the number of cases based on the data provided. However, this time the model has a lower R-squared value than the last model with a value of 0.453, this means that this model is not going to be as accurate as the last model as there is more noisy data which makes it harder for the model to find patterns within the data and make predictions. For these reasons the model can be considered but should not be relied upon. The last state that was analysed was the state with the coldest mean average temperature which was North Dakota.

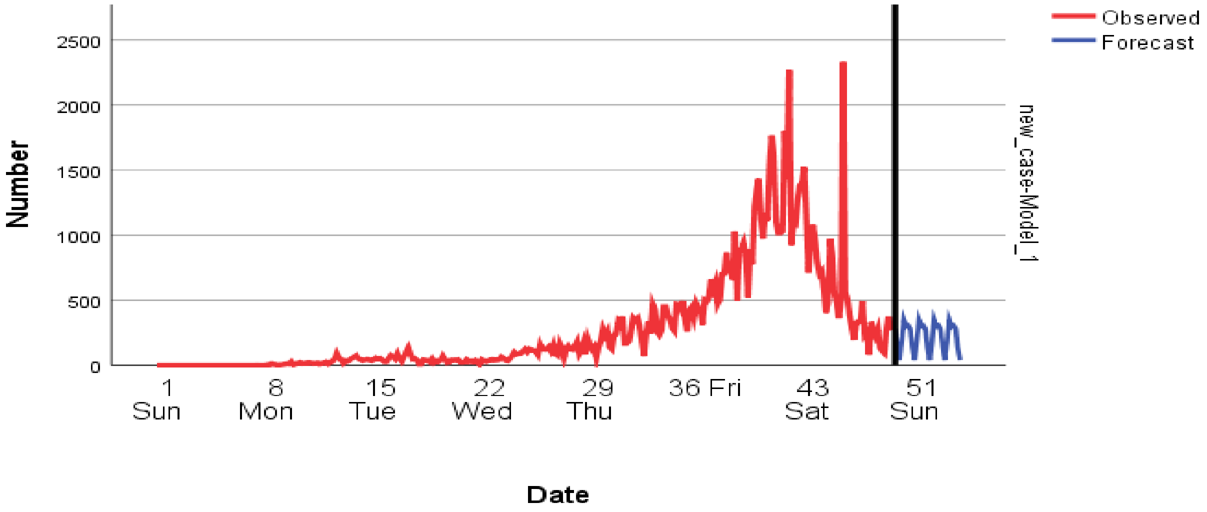


Figure 31: Time series analysis predicting case numbers in North Dakota for January 2021.

Again, this time series model was create using the expert modeller to try and predict the number of cases in North Dakota in January 2021. The model predicted that the case numbers would remain level and vary between roughly 10 cases and 400 cases. To test the accuracy of this model the actual case numbers were analysed.

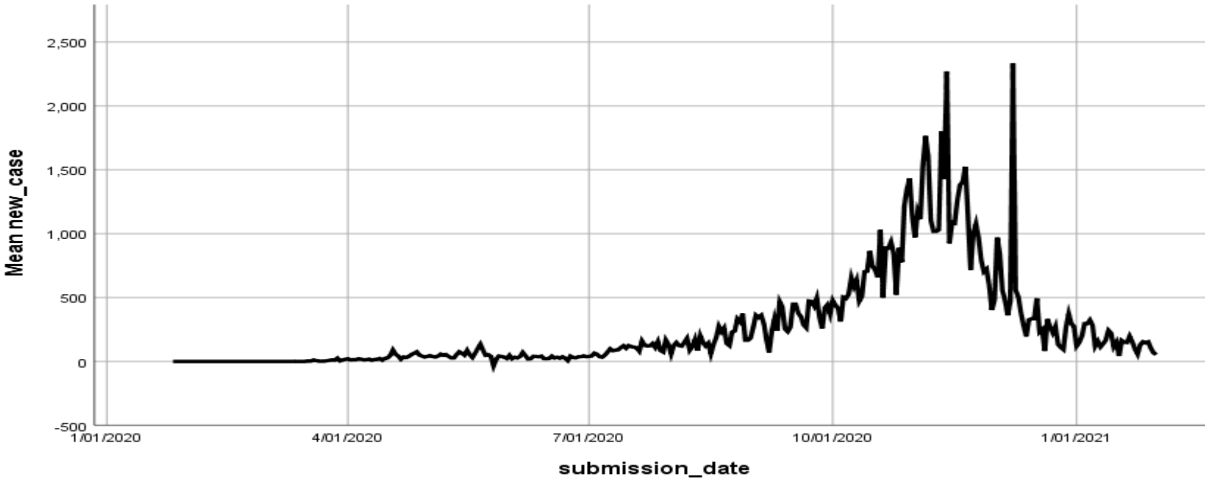


Figure 32: Graph displaying the number of cases in North Dakota in 2020 and January 2021.

Looking at figure 32 above it can be seen that the cases remain between the roughly 10 and 350 which is what the model in figure 31 predicted. This shows that the model is quite accurate and has good predictive power. The model was then used to try and predict the number of cases in North Dakota in May 2021.

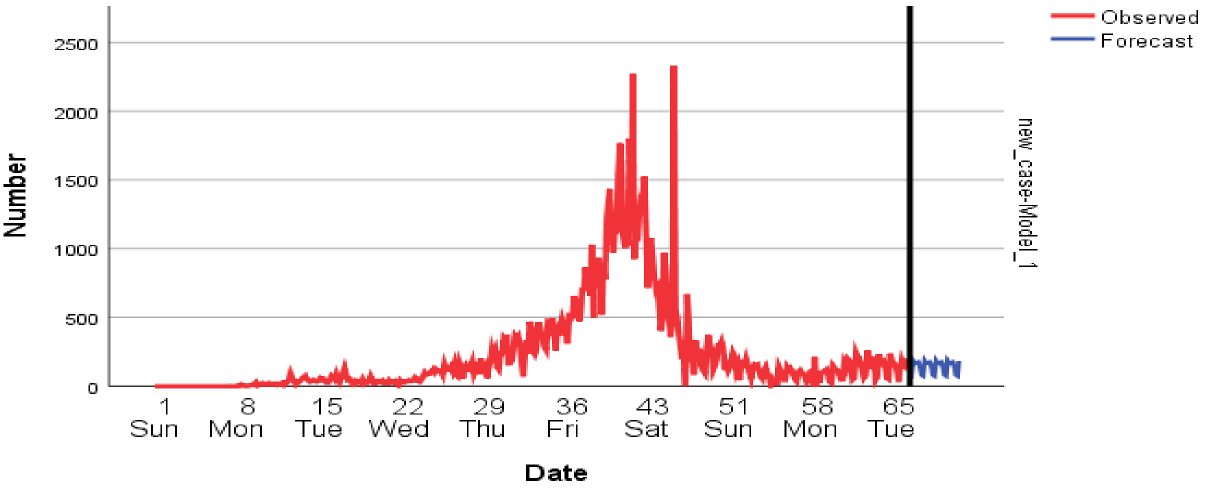
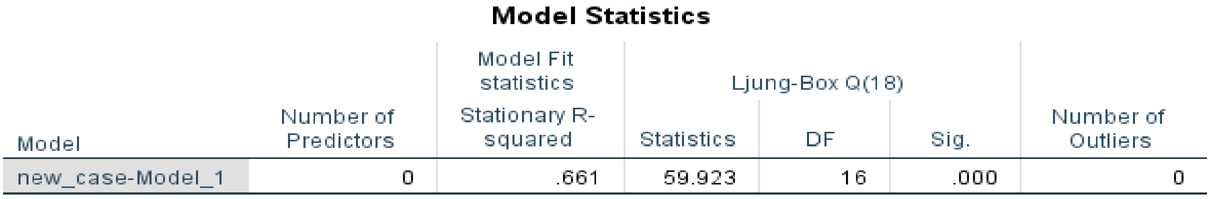


Figure 33: Time series analysis predicting case numbers in North Dakota for May 2021.

This model predicts that the cases will continue to be roughly between 10 and 200 cases a day. Judging by the trend within the data it appears that the model is correct however to analyse the model further the statistics summary table was analysed.



Model Statistics

		Model Fit statistics	Ljung-Box Q(18)			
Model	Number of Predictors	Stationary R-squared	Statistics	DF	Sig.	Number of Outliers
new_case-Model_1	0	.661	59.923	16	.000	0

Table 15: North Dakota time series model statistics.

The model statistics above show that there is a p-value < 0.05, therefore, the null hypotheses should be rejected. This model is significant and capable of predicting the dependent variable which in this case is the number of cases in North Dakota in May 2021, based on the data provided. This model also has the highest R-squared value out of all the models with a value of 0.661, this shows that this model has the least amount of noisy data, and that the model is not overfitted.

When all of these models are analysed together it can also be seen that the cases are highest in the hottest average state which is Florida and lowest in the coldest average state which is North Dakota. This may indicate that the climate does impact the spread of the virus but what is also important to note is that in all of the states even though the temperature is increasing the cases appear to be levelling out and there are no major increases or decreases which indicates that the climate more than likely does not impact the spread of the virus.

5.3 Tableau

Tableau was used to create some various different graphs to help further understand the data and aid this study.

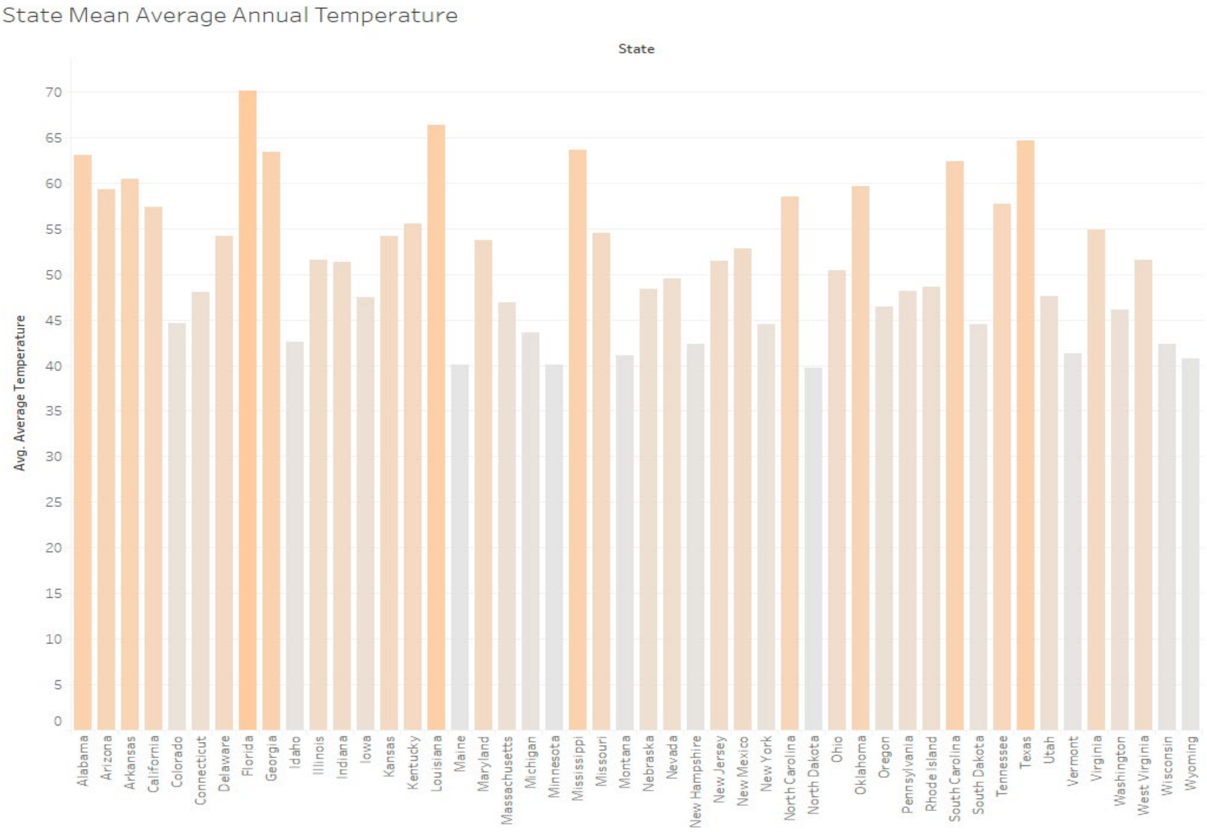
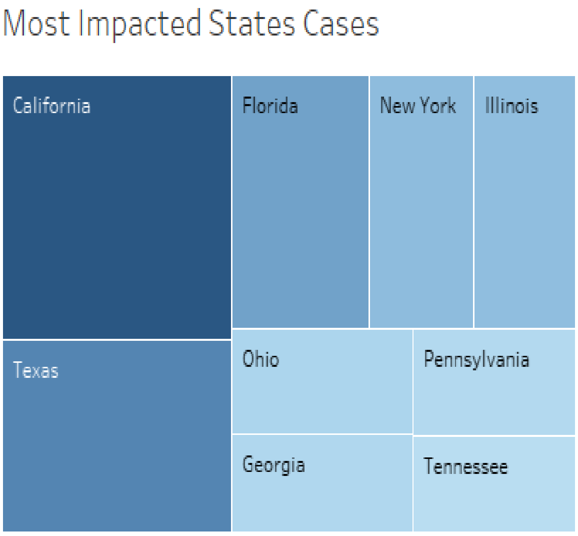


Figure 34: Bar chart displaying the mean average annual temperature of each state.

This bar chart helps to identify which states have the highest average annual temperatures and which states have the coldest average annual temperature. For example, it can be easily seen that Florida, Louisiana, Texas, and Mississippi all have very high average annual temperatures and on the other hand it is easily seen that North Dakota, Vermont, Minnesota, and Maine all have very low average annual temperatures.



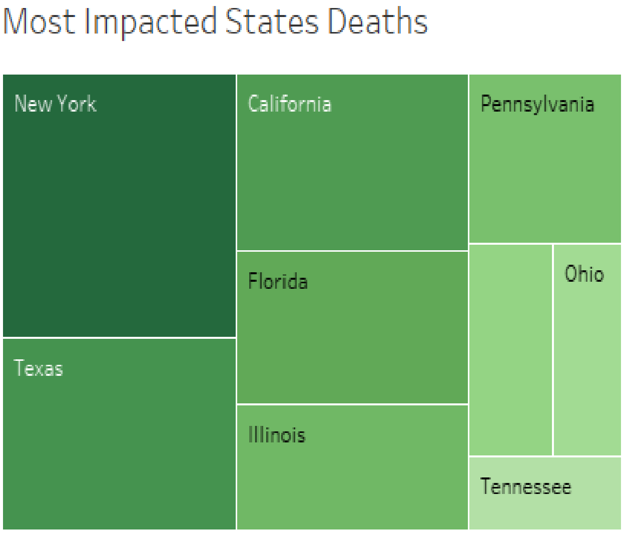


Figure 35: Tree map showing the states with the most annual cases. **Figure 36:** Tree map showing the states with the most annual deaths.

These tree maps make it easy to identify that states that were the most heavily impacted by the virus and when the states present in these graphs are analysed alongside figure 34 it can be seen that more of the hotter and mid temperature states were more heavily impacted than the colder states.

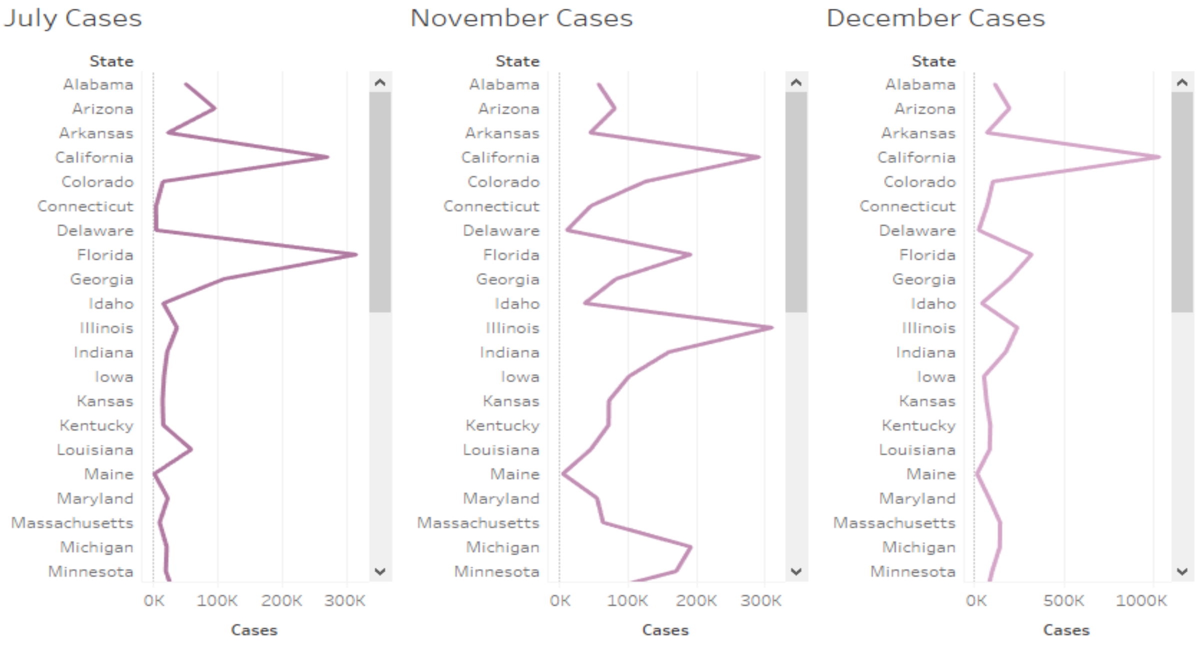


Figure 36: A line graph displaying the number of cases in the three months that had the highest number of cases.

The graph above helps to see which states were the most impacted in each given month, once it has been identified which states were impacted the most the temperature in each of the states can be analysed further see if there is any correlation between the number of cases and the temperature of the given month. All the visualisations were then displayed with each other in the form of an interactive dashboard so that all of the graphs could be found in the same place.

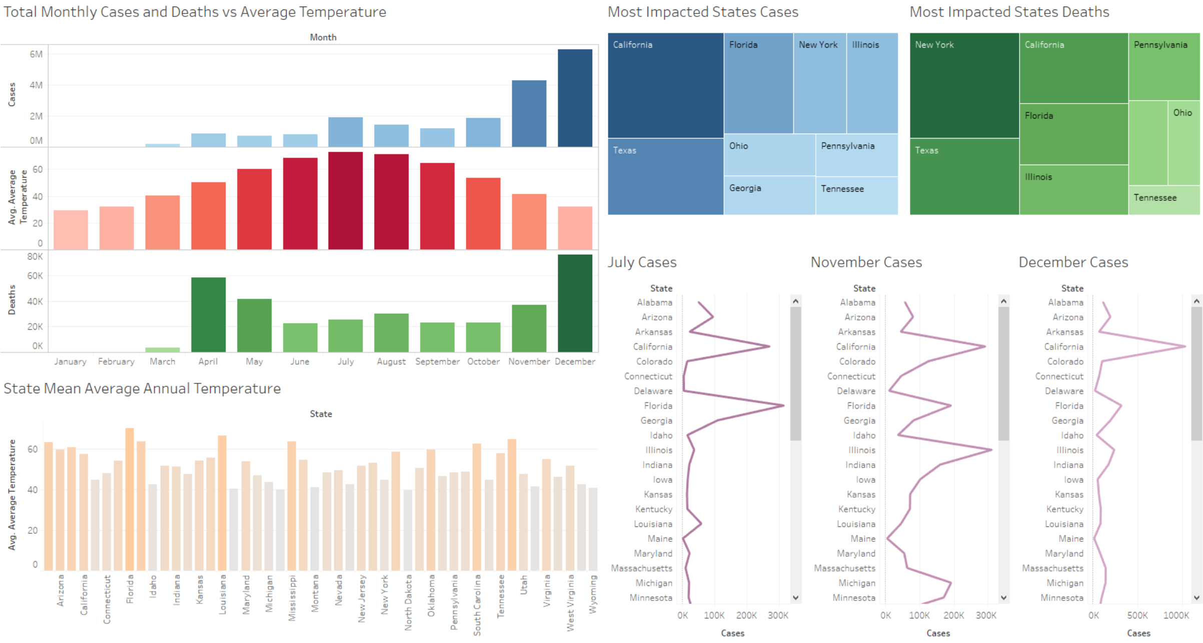


Figure 37: Interactive dashboard created in Tableau.

6.0 Conclusions

After analysing the results of the different analysis techniques and tests carried out above it can be concluded that although it does not appear that the climate has a direct effect on the spread of COVID-19 in the United States of America, there are still some signs that the climate may actually be a factor, but with the data that was available for this study there is not enough evidence to give a definite conclusion or a definite answer to the research question which is “Does the climate have a direct effect on the spread of COVID-19 in the United States of America?”. Because the virus has only been around since the beginning of 2020, there is not enough data to analyse and therefore it is very difficult to give a definite answer to the research question.

In some cases, it appears that the virus spreads more in hot climates and in other cases it appears that the virus spreads more in cold climates. These conflicting results make it appear that the climate does not have a direct effect on the spread of COVID-19 but the only way to provide a definite answer would be to wait for more data to become available and carry out the same analysis techniques again. The other reason a definite answer could not be provided to the research question is because there are limitations to this study.

The first limitation has already been mentioned which is there is not enough historical data to analyse, this makes it hard to identify patterns within the data. Another limitation of this study is that population and population density are not considered. Because the virus spreads by contact population density is very important when it comes to how easily the virus can spread. This means that the states that are showing high numbers of cases is more likely to be because of high population density rather than the climate of the state. Another limitation is that vaccine rates and government lockdowns are not considered, this means that when cases drop it is more than likely not because of the climate alone but it could be because of a boost in the number of vaccines or more strict lockdown measures. Because of these limitations it was not possible to give a definite answer to the research question.

However, there are some other advantages to this study, the main advantage of this study is that states that had a very high number of cases in the most effected months were identified, this can help the state to identify when their cases are highest, and they can then put restrictions in place for these problem months to reduce the number of cases. Another advantage of this study is that all of the techniques that were carried out are all relevant and as new data becomes available the results of the tests will become more accurate.

7.0 Further Development and Research

In terms of further development and research this study would benefit greatly from the introduction of more data. This study could also be carried out with more factors such as population, population density, vaccine rates, lockdown duration and more. With the introduction of more factors, multiple linear regression could be carried out and more in-depth models could be created and analysed to give a better insight into what effects the spread of COVID-19 the most. With more factors being analysed the study would then be able to give more accurate and better results that could help with providing a definite answer to the research question of this study.

If this analysis was to be carried out again a dataset with more factors would be chosen. Initially it was unsure what was required for this study in terms of data as the techniques and technologies were only being taught at the same time this project was being developed. This made it difficult as there was no background knowledge of data analytics and when it came to performing the techniques that were taught throughout the year it was identified that the techniques were limited by the initial choice of data and it was too late to change the data. After completing this project and becoming more familiar with the different techniques and technologies, better datasets would be chosen that are more suitable to the study and so more advanced data mining techniques could then be carried out such as data normalisation, principal component analysis and multiple linear regression. By introducing these techniques, the study would be more in depth and better conclusions could be obtained.

8.0 References

Centers for Disease Control and Prevention, United States COVID-19 Cases and Deaths by State over Time, Data.Gov, United States of America, Data.Gov 2020. Accessed on: December 21, 2020. [Online]. Available: https://catalog.data.gov/dataset/united-states-covid-19-cases-and-deaths-by-state-over-time-845b7.

Google Developers 2012, states.csv, United States of America. Accessed on: December 20, 2020. [Online]. Available: https://developers.google.com/public-data/docs/canonical/states_csv.

Google Developers 2020, Normalization, viewed 21 December 2020, <https://developers.google.com/machine-learning/data-prep/transform/normalization>.

Google News 2020, Corona Virus United States, Google, viewed 08/11/2020, <https://news.google.com/covid19/map?hl=en-IE&mid=%2Fm%2F09c7w0&gl=IE&ceid=IE%3Aen>.

National Centers for Environmental Information, Climate at a glance, NOAA, United States of America, National Oceanic and Atmospheric Administration 2020. Accessed on: December 21, 2020. [Online]. Available: https://www.ncdc.noaa.gov/cag/statewide/mapping.

Rajput, A 2019, KDD Process in Data Mining, GeeksforGeeks, viewed 21 December 2020, <https://www.geeksforgeeks.org/kdd-process-in-data-mining/>.

Williams, G.J & Huang, Z 1996, Modelling the KDD process: A four stage process and four element model, Technical Report TR-DM-96013, CSIRO Division of Information Technology. Available:https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.8799&rep=rep1&type=pdf.

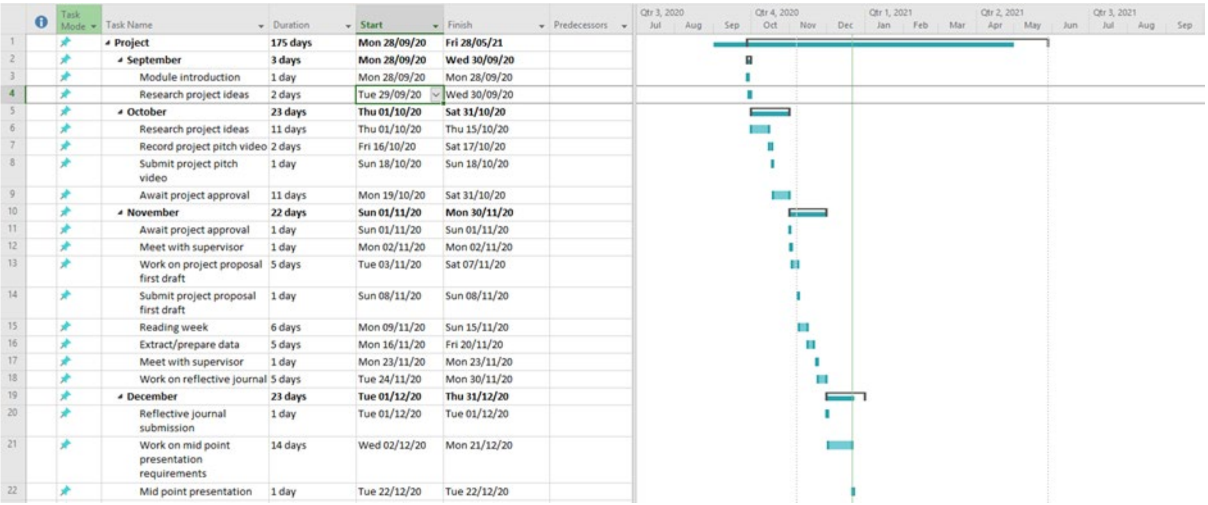
World Health Organization 2020, WHO Coronavirus Disease (COVID-19) Dashboard, WHO, viewed 21 December 2020, <https://covid19.who.int/>.

World Population review 2020, List of State Abbreviations, United States of America. Accessed on: December 20, 2020. [Online]. Available: https://worldpopulationreview.com/states/state-abbreviations.

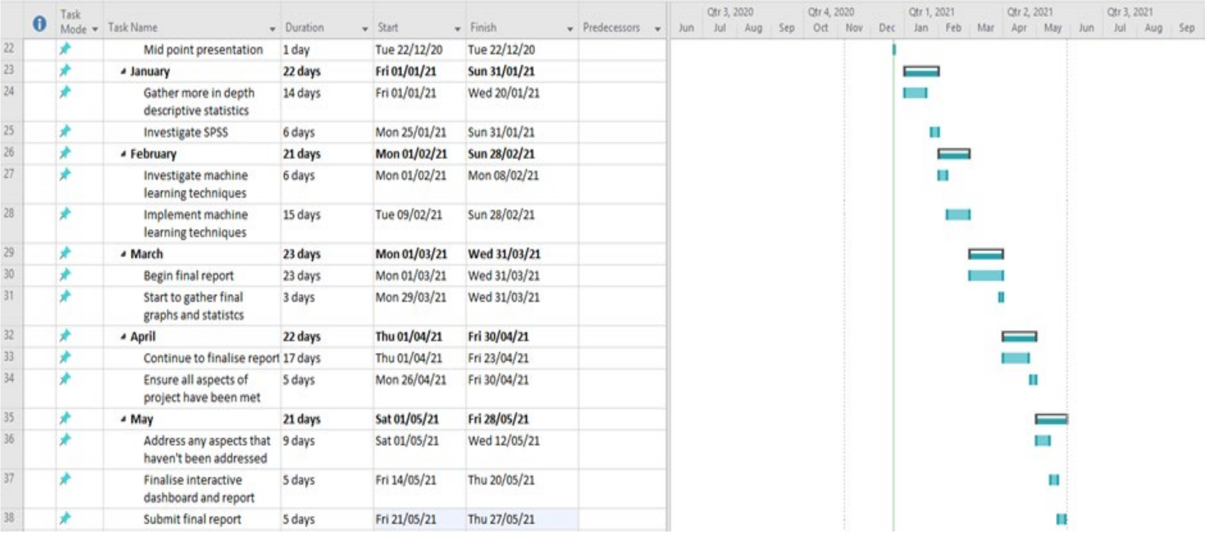
9.0 Appendices

9.1. Project Plan

A Gantt chart was created using Microsoft Project to outline specific deadlines that had to be met to ensure that the project was complete on time and to a high standard.



	ⓘ	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		📌	**◢ Project**	**175 days**	**Mon 28/09/20**	**Fri 28/05/21**	
2		📌	**◢ September**	**3 days**	**Mon 28/09/20**	**Wed 30/09/20**	
3		📌	Module introduction	1 day	Mon 28/09/20	Mon 28/09/20	
4		📌	Research project ideas	2 days	Tue 29/09/20	Wed 30/09/20	
5		📌	**◢ October**	**23 days**	**Thu 01/10/20**	**Sat 31/10/20**	
6		📌	Research project ideas	11 days	Thu 01/10/20	Thu 15/10/20	
7		📌	Record project pitch video	2 days	Fri 16/10/20	Sat 17/10/20	
8		📌	Submit project pitch video	1 day	Sun 18/10/20	Sun 18/10/20	
9		📌	Await project approval	11 days	Mon 19/10/20	Sat 31/10/20	
10		📌	**◢ November**	**22 days**	**Sun 01/11/20**	**Mon 30/11/20**	
11		📌	Await project approval	1 day	Sun 01/11/20	Sun 01/11/20	
12		📌	Meet with supervisor	1 day	Mon 02/11/20	Mon 02/11/20	
13		📌	Work on project proposal first draft	5 days	Tue 03/11/20	Sat 07/11/20	
14		📌	Submit project proposal first draft	1 day	Sun 08/11/20	Sun 08/11/20	
15		📌	Reading week	6 days	Mon 09/11/20	Sun 15/11/20	
16		📌	Extract/prepare data	5 days	Mon 16/11/20	Fri 20/11/20	
17		📌	Meet with supervisor	1 day	Mon 23/11/20	Mon 23/11/20	
18		📌	Work on reflective journal	5 days	Tue 24/11/20	Mon 30/11/20	
19		📌	**◢ December**	**23 days**	**Tue 01/12/20**	**Thu 31/12/20**	
20		📌	Reflective journal submission	1 day	Tue 01/12/20	Tue 01/12/20	
21		📌	Work on mid point presentation requirements	14 days	Wed 02/12/20	Mon 21/12/20	
22		📌	Mid point presentation	1 day	Tue 22/12/20	Tue 22/12/20	



	ⓘ	Task Mode	Task Name	Duration	Start	Finish	Predecessors
22		📌	Mid point presentation	1 day	Tue 22/12/20	Tue 22/12/20	
23		📌	**◢ January**	**22 days**	**Fri 01/01/21**	**Sun 31/01/21**	
24		📌	Gather more in depth descriptive statistics	14 days	Fri 01/01/21	Wed 20/01/21	
25		📌	Investigate SPSS	6 days	Mon 25/01/21	Sun 31/01/21	
26		📌	**◢ February**	**21 days**	**Mon 01/02/21**	**Sun 28/02/21**	
27		📌	Investigate machine learning techniques	6 days	Mon 01/02/21	Mon 08/02/21	
28		📌	Implement machine learning techniques	15 days	Tue 09/02/21	Sun 28/02/21	
29		📌	**◢ March**	**23 days**	**Mon 01/03/21**	**Wed 31/03/21**	
30		📌	Begin final report	23 days	Mon 01/03/21	Wed 31/03/21	
31		📌	Start to gather final graphs and statistcs	3 days	Mon 29/03/21	Wed 31/03/21	
32		📌	**◢ April**	**22 days**	**Thu 01/04/21**	**Fri 30/04/21**	
33		📌	Continue to finalise report	17 days	Thu 01/04/21	Fri 23/04/21	
34		📌	Ensure all aspects of project have been met	5 days	Mon 26/04/21	Fri 30/04/21	
35		📌	**◢ May**	**21 days**	**Sat 01/05/21**	**Fri 28/05/21**	
36		📌	Address any aspects that haven't been addressed	9 days	Sat 01/05/21	Wed 12/05/21	
37		📌	Finalise interactive dashboard and report	5 days	Fri 14/05/21	Thu 20/05/21	
38		📌	Submit final report	5 days	Fri 21/05/21	Thu 27/05/21	

Figure 38: Project plan.

This project plan was created prior to understanding what exactly was required and exactly how long it would take to execute some of the techniques, but the plan was still followed reasonably well. However, due to a high workload throughout the year and a lack understanding on how to use some of the technologies some of the deadlines were pushed back until the technologies were taught in the different modules.

9.2. Project proposal

9.2.1 Objectives

The main purpose of this report is to try and figure out if there is a link between climate and the spread of Covid-19 in the USA. The datasets that will be examined for this study are “United States COVID-19 Cases and Deaths by State over time” (Data.Gov,2020), and a weather dataset that was found on the NOAA website (National Oceanic and Atmospheric Administration,2020).

Using various data mining techniques such as cluster analysis, normalisation, interactive graphs, and descriptive statistics, the plan is to be able to deliver clear and easy to interpret results by producing visualisations and accurate statistics using software’s such as RStudio, Tableau, Excel, and SPSS.

The main research question that this study will aim to answer is “Does the climate have a direct effect on the spread of COVID-19 in the United States of America?”. By carrying out different data mining techniques and statistical tests this study will try to find enough evidence to attempt to answer the research question and present the results in a clear manner.

9.2.2 Background

This topic was chosen for investigation because of how relevant COVID-19 and climate change are. This analysis was chosen in the hope of being able to produce some valuable information and findings to see if there is in fact a link between climate and the spread of COVID-19 as there is still a lack of understanding when it comes to the COVID-19 virus. COVID-19 was firstly identified in December of 2019 in Wuhan, China. In March 2020, the COVID-19 outbreak was declared a pandemic. At the time of writing the project proposal there was a total of over 49.5 million confirmed COVID-19 cases and over 1.2 million COVID-19 related deaths worldwide. In the USA alone there was over 9.9 million confirmed cases and just over 237,000 deaths (World Health Organisation,2020).

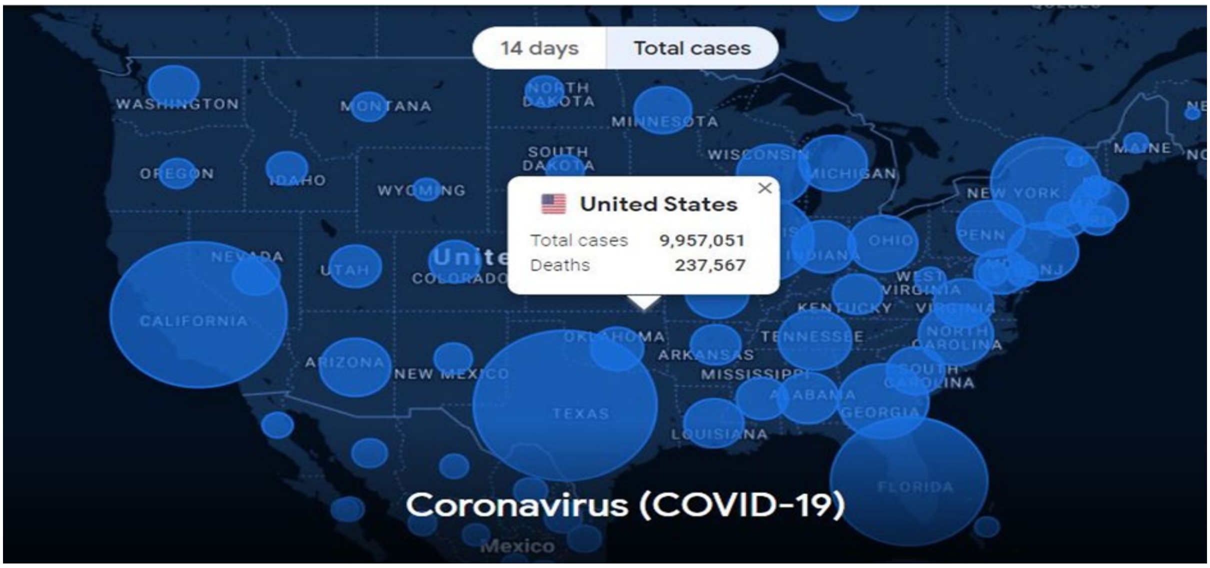


Figure 39: COVID-19 USA total cases map (Google News, 2020).

WHO (World Health Organisation) say that COVID-19 has 2 main modes of transmission, droplet transmission and direct contact with an infected person. “Droplet transmission occurs when a person is in close contact (within 1 m) of someone who has respiratory symptoms (e.g., coughing or sneezing) and is therefore at risk of having his/her mouth, nose or eyes exposed to potentially infective respiratory droplets.” They say “transmission may also occur through fomites in the immediate environment around the infected person. Therefore, transmission of the COVID-19 virus can occur by direct contact with infected people and indirect contact with surfaces in the immediate environment or with objects used on the infected person (e.g., stethoscope or thermometer)” (World health organisation,2020).

The reason that the topic ‘the effect of the climate on the spread of the virus’ was chosen is because whilst looking at the total cases graph above and it was noticed that the most affected areas in the US were Texas, Florida and California, these three states are all considered warm states and I noticed that some of the least affected areas were Montana, North Dakota and Oregon which are all considered cold states. That is why this topic was chosen to be investigated further and see if there is any corelation between the climate of a state and COVID-19 cases/deaths.

The goal of this study is to hopefully be able to provide enough information to answer the research question which in turn would be able to help state governors in the future, predict when the transmission of the virus is at its highest which could help them implement guidelines for certain times of the year. For example, if cases/deaths are at their highest point in June because of hot temperatures, then the governor of the state could enforce closures to reduce human interaction or make it mandatory to wear a mask and gloves during these high transmission months.

9.2.3 Technical Approach

After researching different ideas online, it was decided that this analysis would look at COVID-19 because there is still a lack of understanding of the virus and there are a lot of things that are still unknown about it. This study plans to research different datasets, academic journals, and websites in order to obtain as much information as possible to attempt to answer the research question.

The requirements that were set for this study are:

- To decide on at least two datasets to examine.
- To clean the data and prepare it for investigation.
- To carry out various data mining techniques and statistical tests.
- To visualise the data so the findings are easy to interpret.
- To attempt to answer the main research question.

To examine, clean and carry out various data mining techniques 2 programming languages were analysed to see which would be best to use for this study. The two that had to be chosen between were R and Python. After researching both languages, it was decided that R would be the best programming language to use for this study because R is a statistical programming language and therefore will be more suitable for this study than Python. To visualise the data and carry out different data mining techniques a suitable IDE would have to be chosen and the best IDE to use for R is RStudio. RStudio is equipped with many libraries that make it a great tool for carrying out a wide variety of data mining techniques and make it very easy to create visualisations. For these reasons it was decided that RStudio would be the best IDE to use for this study.

To ensure that this analysis is carried out correctly the KDD methodology will be followed. The KDD methodology involves cleaning, selecting, transforming, and analysing the data. By following this methodology, it will help to ensure that the data is pre-processed correctly so that the different data mining techniques and visualisations can be carried out and analysed effectively.

To carry out this analysis it is planned that a number of different software's will be used such as Tableau, SPSS and Excel as well as RStudio. It is planned that Tableau will be used to produce some visualisations and display them in the form of an interactive dashboard, SPSS will be used to carry out different statistical tests and Excel will be used to select and pre-process the data so that all of the tests can be easily executed in RStudio and SPSS.

9.2.4 Technical Details

As mentioned already, the plan is to use Excel to select and pre-process the data, use the R programming language and the RStudio IDE to carry out data mining techniques and create interactive visualisations, use Tableau to create an interactive dashboard and use SPSS to carry out different statistical tests.

In Excel, it is planned to carry out data pre-processing and data selection techniques so that different CSV files can be created and formatted correctly to allow the different data mining techniques and visualisations to be created and carried out.

In RStudio the plan is to use many different libraries such as "plotly", "car", "rcompanion", "ggplot2" and "dplyr" for example. The plan is to thoroughly analyse the data by carrying out different data mining techniques such as regression, clustering, normalisation and more to produce some interesting insights and create some visualisations to display the data. It is also planned that RStudio will also be used to further pre-process the data when required to allow the data mining techniques and statistical tests to be executed.

It is planned that SPSS will be used to carry out different parametric or non-parametric statistical tests depending on if the data is normally distributed or not. Tests such as the chi square test, Pearsons R test, Friedman test, Wilcoxon Rank test and more can be used to statistically analyse the datasets chosen and hopefully find some interesting results. What also may be useful is to use SPSS to carry out a time series analysis to attempt to predict future case numbers in certain areas. These findings could be interesting and may help answer the research question.

Lastly, it is planned to use Tableau to create some graphs and plots and display them in the form of an interactive dashboard. An interactive dashboard would benefit this study greatly and would provide the reader with a lot of useful information in one location.

9.2.5 Evaluation

In order to validate the data a few different steps will need to be carried out, some of these steps are:

1. Determine the data sample – It is important to ensure that the datasets that are chosen for this study are large enough to provide accurate results to the different data mining techniques that are going to be carried out throughout this project. It is also important to use a large data sample when carrying out tests for accurate results. The 2 main datasets that were chosen for this study are both large enough to obtain some interesting results and provide a thorough analysis.

2. Validate the database – Next, it is important to ensure that the data that was chosen for this study contains enough relevant columns and rows so that the analysis can be carried out and many results can be gathered from the datasets, In the case of the 2 main datasets that were chosen, the first contains 15 columns and over 19,000 rows and the second contains 8 columns and over 72,000 rows at the time of writing. This is more than enough unique ID's to carry out a thorough analysis.

3. Validate the data format – The condition of the data then needs to be determined and a search for incongruent or incomplete counts, duplicate data, incorrect formats, and null field values should be carried out to ensure the data is completely clean before executing any analysis techniques. To validate the data format exploratory techniques will be carried out to analyse the data in RStudio and Microsoft Excel before carrying out any tests.

4. Validate the data mining techniques – Each of the data mining techniques should be validated by carrying out the tests on data that is already known or on training and test models. By validating the data mining techniques, it can then be determined how accurate the results are and if they can be relied on.

9.2.6 References

Centers for Disease Control and Prevention, United States COVID-19 Cases and Deaths by State over Time, Data.Gov, United States of America, Data.Gov 2020. Accessed on: November 08, 2020. [Online]. Available: https://catalog.data.gov/dataset/united-states-covid-19-cases-and-deaths-by-state-over-time-845b7.

Google News 2020, Corona Virus United States, Google, viewed 08/11/2020, <https://news.google.com/covid19/map?hl=en-IE&mid=%2Fm%2F09c7w0&gl=IE&ceid=IE%3Aen>.

National Centers for Environmental Information, Climate at a glance, NOAA, United States of America, National Oceanic and Atmospheric Administration 2020. Accessed on: December 21, 2020. [Online]. Available:https://www.ncdc.noaa.gov/cag/statewide/mapping.

World Health Organization 2020, Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations, World Health Organization, viewed 08/11/2020, <https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations#:~:text=According%20to%20current%20evidence%2C,transmission%20was%20not%20reported.>.

9.3. Reflective Journals

9.3.1 November

Over the last month my idea has been approved and I have had a couple of meetings with my supervisor, and he has given me some good tips on how to go about doing my project. I have also found what two datasets I am going to use for this study, and I have practiced using R Studio in the Data Application Development module. I feel a lot more confident using R Studio now for visualisations and to analyse the data.

Other than that, I have still not got a chance to get a good start on the main software project as the workload has been so intense, in the last month I have had 4 assignments due and an exam worth 50% of a module so this has not given me a real chance to start the main project as of yet. I have also been given another 2 assignments to do with 1 of them being due on the same day as the mid-point presentation for this module so I am worried as to how quickly I will be able to progress over the next month but hopefully I will be able to get enough work done at a good standard for both modules.

9.3.2 December

Over the last month the other assignments started to quieten down, and I finally had the time to get a proper start to my project. I was quite happy with the amount I submitted for my mid-point that was due on the 22nd of December, some of the things I got done were:

- Found the best datasets for the purpose of my project.

- Used excel to examine and do some basic cleaning of the datasets.

- Used R and RStudio to create some useful cluster plot graphs and choropleth maps and to also clean the datasets further.

- I played around with SPSS to gather some basic descriptive statistics.

- I created a couple of graphs in Tableau and familiarised myself with the creation of an interactive dashboard.

- I completed my mid-point word document and PowerPoint presentation video.

Next up I am going to start creating my interactive dashboard and start gathering some statistics that have better meaning than the ones I already have on SPSS. After the last month I am now happy and confident that I will be able to produce a good quality project and I feel that the progress I have made up to this point has been good.

9.3.3 January

For the month of January, I have now finalised the main COVID-19 dataset I am using, and the figures are now accurate. The submission during December consisted of temporary figures for the month of December as stated in my midpoint submission but now I have updated the data, so it is now accurate. Now that my data is finalised, I began produce some more meaningful statistics in both RStudio and SPSS. Over the next month or two I plan to gain a better understanding of SPSS as I now have a module that will run through exactly how to use it and will explain some of the interesting tests that can be carried out and analysed on SPSS. Next month I plan to create some more visualisations and further explore SPSS.

9.3.4 February

Over the last month I have begun to try and come up with some more relevant descriptive statistics. I had a very productive meeting with my supervisor, and he gave me some helpful tips on what tests I may want to carry out. Over the last month I have started to use SPSS more and have looked at tests such as the Holt winters test, the ARIMA test and K-means clustering. I have found some interesting results and I now plan to take my results and turn them into graphs and plots using Tableau and RStudio to make them easier to understand for the general viewer of the project.

Now that I have some more descriptive statistics, I plan to take a deeper look into the cluster analysis I had already carried out for the midpoint submission and try to gather some information that may lead to some predictions being made in the future. I have now also submitted most of the information required for the showcase and am just awaiting approval.

9.3.5 March

This month all of the data mining techniques that were carried out in RStudio which were K-means clustering, linear regression, and the interactive maps were finalised, and a Shapiro-Wilk normality test was also conducted to determine if the data was normally distributed or not. The reason this test was carried out was so it could be determined whether parametric or non-parametric tests had to be conducted. The Shapiro-Wilk test results indicated that the data was not normally distributed and therefore non-parametric tests had to be carried. The final report was also started, and the findings so far were added to the document and discussed.

Once it was identified that non-parametric tests were required, I had gathered some experience with non-parametric tests in the Advanced Business Data Analysis module and I was able to determine which tests would be suitable for this project and this data. The tests that were chosen were the Wilcoxon Rank test, the Mann-Whitney U test, and the Friedman test. Each of the tests were conducted and the results are ready for analysis.

Next month I plan to conduct a time series analysis in SPSS to try and predict the future number of cases, and I also plan to create an interactive dashboard in Tableau which contains various different graphs and plots.

9.3.6 April

This month the time series analysis was conducted in SPSS to try and predict future case numbers in three different states. All of this data was then added to the report and analysed. I also created the interactive dashboard in Tableau that contains graphs and plots to help viewer better understand the data. All that is left to do now is complete and finalise the document and create the showcase poster. Once both of these are done, I can then submit the final artefacts. Once the final artefacts are submitted the final thing I have to do is record my presentation.