

# National College of Ireland

Technology Management BSHTM4

Data Analysis

2020/2021

Ciara Quigley

x17408654

[x17408654@student.ncirl.ie](mailto:x17408654@student.ncirl.ie)

Covid 19

Technical Report

## Contents

Executive Summary .....	2
1.0 Introduction .....	2
1.1. Background .....	2
1.2. Aims.....	3
1.3. Technology.....	4
1.4. Structure .....	5
2.0 Data .....	6
3.0 Methodology.....	11
3.1 Data Cleaning .....	12
3.2 Data Integration .....	12
3.3 Data Selection .....	13
3.4 Data Transformation .....	13
3.5 Data Mining.....	13
3.6 Knowledge Presentation.....	13
4.0 Analysis .....	13
5.0 Results .....	15
6.0 Conclusions .....	25
7.0 Further Development or Research .....	26
8.0 References .....	28
9.0 Appendices.....	29
9.1. Project Plan .....	29
9.2. Reflective Journals .....	31
9.3. Project Proposal .....	36
9.3.1 Objectives.....	37
9.3.2 Background .....	37
9.3.3 Technical Approach.....	38
9.3.4 Project Plan .....	40
9.3.5 Technical Details .....	42
9.3.6 Bibliography .....	42

## Executive Summary

I decided to conduct this analysis based on Covid 19 data because it is a relevant topic nationally and globally. Covid 19 has brought the entire world to a standstill. There have been millions of cases reported worldwide and as of December 20th, 2020, 1.7 million people have lost their lives to this awful disease. I intend to compare the distribution of Covid 19 by continent and look at data from Asia, Africa, Europe, South America, North America, and Oceania. In this report I will analyse the distribution of virus to identify which continents were most affected. I will compare the instance of coronavirus in 215 countries, specifically looking at number of patients who have recovered. I will compare the trends in the spread of Covid 19 Ireland in Ireland and New Zealand in relation to the closing of borders and school opening.

### 1.0 Introduction

#### 1.1. Background

Coronavirus is a newly detected disease. It is thought by experts that the SARS-CoV-2 disease originates in bats. It is believed that this disease made its way to humans in Wuhan in China through an open air market otherwise known as a wet market. These markets are where locals can purchase meat and fish often times the animals being killed on the spot. Retailers in these markets often sell wild animals for consumption which includes bats. Now that the virus has infected human beings, it is no longer solely transmitted through animals, but humans themselves. It is an air born disease meaning physical contact is not required in order to obtain the virus. It travels through bodily fluids such as coughs or sneezes.

The virus was first detected in December 2019 in Wuhan, China and has since infected the entire world resulting in mass shutting down of all non essential business such as barbers, construction, and retailers. 160,000,000 people all over the world have been infected by this disease. (CSO.ie, 2021) In Ireland, 92% of all deaths have occurred in persons over the age of 65. (CSO.ie, 2021) People who are infected by Covid 19 experience flu like symptoms. They can range from mild to severe respiratory problems. The majority of people infected will experience mild to moderate symptoms likened to a cold or the flu however in some cases, the disease can be deadly and mostly affects people with underlying medical problems or the elderly. (who.int, 2021)

I decided to conduct this analysis on Covid 19 data because it is a relevant topic not only nationally but globally too. Covid 19 has brought the entire world to a standstill. There have been millions of cases reported worldwide and as of December 20<sup>th</sup>, 2020, 1.7 million people

have lost their lives to this awful disease. (Worldometer.info, December 2020) As this is such a new and contagious disease, there is so much we do not know about it. I am performing this analysis to hopefully gain some new insights into the disease which has caused this global pandemic. Nobody could have foreseen the impact this would have on the global economy and on people's lives, I want to investigate into how the initial shutting down of the country which initially was supposed to last for two weeks has now in some capacity, has lasted nine, going on ten months and beyond.

Over a year on, we are still in the same boat. We still have many cases, and they are not going down at the rate we would like them to. However, there has been mass improvement within Ireland over the last 6 months. In Ireland, we saw a huge spike in cases in January after the Christmas period. We saw our cases peak on January 8<sup>th</sup>, 2021 with 8,227 cases reported in just one day, with a 7 day average of 6,363. (hse.ie, 2021)

## 1.2. Aims

I have three main aims for this project:

1. Determine the relationship between Covid cases rising in Ireland and schools reopening
2. Compare the response to Covid 19 between Ireland and New Zealand, two countries with similar populations but a huge difference in cases
3. Investigate the global distribution of the virus, which areas were most impacted, where did the most deaths occur etc.

I also hope to find some interesting information about the virus which may not necessarily be recognised a lot in the media but has also got scientific confirmation that the information is true and accurate. I aim to conduct an investigation into the vaccination and look into how they were able to develop this life saving injection so quickly. When more people have been administered the vaccination, I hope to analyse the impact it has had, mainly within the elderly and vulnerable population.

I aim to compare countries and how they have coped with the virus. A country I will pay particular attention to is New Zealand. Their reaction to the virus has been incredible. In a country of 4.8 million people, since the outbreak of Covid 19, have only had 2,121 reported cases and 25 deaths. (Worldometer.info, December 2020) Comparing that to Ireland with a similar population of 4.9 million people, we have had 79,542 reported cases and 2,158 Covid related deaths as of December 2020. (Worldometer.info, December 2020)

Another aim of mine is to keep my report clear and precise. I do not want to include anything for the sake of the word count. If it is not relevant to my project, I will not include it. This will help to assist in the easy reading of my report. I aim to have it well structured and

easy to follow. Every section I discuss will be labelled and a direct link to each section can be found in the table of contents. This structure will ensure that the report is coherent, and anybody can pick it up and instantly understand it. As Covid 19 is such a relevant topic to everybody, I anticipate other people to read it, not only NCI staff. I know that my friends and family and interested is seeing the information that I have found out. The majority of my friends and family do not come from a technical background so I would like to produce a report which all levels can understand.

### 1.3. Technology

I technologies I will be using to achieve my goals set out are, Microsoft Excel, SPSS, R Studio, and Jupyter Notebook. During my work placement in third year, I became familiar with Power Bi so if I get the chance to, I would also like to revisit this technology.

#### **Microsoft Excel**

Microsoft Excel is a spreadsheet program. It is used as a tool to express visualisations to display your analysis. For this project, I have used excel as a tool to load my raw data sets into. In excel I can quickly see how many rows and cells are in the data set and I can gain an idea of what information I want to use for my analysis and set out a plan for my analysis. Here I can also use it to do some data cleaning meaning I can remove any data which I don't believe to be useful to my study.

#### **Microsoft Excel - Problems**

Over my four years in college and also outside in the industry, I have been lucky enough to have learned a lot about excel and its many functions. As a result of this, I have not found many problems when conducting this project in terms of excel.

#### **SPSS**

SPSS stands for Statistical Package for the Social Sciences. SPSS is a widely used technology in statistical analysis. It has also been very successful in aiding people in other areas such as researching (both market and human), surveying companies, government organisations, educational researchers, data miners, and many others. For the purpose of my project, I have used SPSS to aid me in my statistical analysis. I have used skills gained from other modules to help me with this.

#### **SPSS - Problems**

For the most part, I have found SPSS to be pretty straight forward compared to other technologies. The hardest part I found was initially when I began using the program. This was difficult for me as although I have used it in many of my modules, I was never taught the basis

which resulted in some confusion at first but resulted in teaching myself using online resources such as YouTube and of course a lot of trial and error.

### **R Studio**

R Studio is an integrated development environment for R. R is a programming language mainly used for statistical computing. R Studio is a desktop application. This is the technology I felt most comfortable with aside from Microsoft Excel. I used this for data cleaning and mainly for creating graphs. I used it to create pie charts, bar charts, histograms, etc.

### **R Studio - Problems**

The largest problem I faced whilst using R Studio was all of the different packages. In the beginning, I would often forget to install the packages I needed and was left perplexed as to why my code was not working. These problems were quickly solved by a quick Google search.

### **Jupyter Notebook**

Jupyter Notebook is a program which is primarily used to support the Python programming language. I used this technology to add to my technologies already mentioned. It was used to create more graphs to show diversity in my analysis skills.

### **Jupyter Notebook - Problems**

Jupyter Notebook was a technology which I was no familiar with at all a couple of months ago. During another module we were instructed to use Python and I did not know which program to use. As this was a group project, a team member of mine found Jupyter Notebook and we all like the look of it. It was straight forward to follow and ran in Google Chrome which made it versatile and different from other technologies I had used in the past. The biggest problem I faced was learning how to use the technology as I went along. Being so new to me, it was difficult to keep up with it sometimes.

## [1.4. Structure](#)

This document will adhere to the following structure:

1. Data – In this section I will discuss where I sourced the data sets used in this analysis along with why I chose them.
2. Methodology – I will outline the methodologies I used in this project providing step by step explanation on each.
3. Analysis – Here is where I will discuss the different approaches I took with the data and justify my answers.
4. Results – In this section I will describe in detail the results I have uncovered providing supporting graphs, tables, and figures.
5. Conclusion – I will provide a conclusion to the document highlighting my key findings

6. Future Works – I will discuss my future ambitions I may have with this project and highlight any changes I would make if I had more time.
7. References – Other resources/ persons work that I used to aid in my analysis.
8. Appendices – Includes project plan, project proposal, reflective journals, and other materials I used.

## 2.0 Data

In order to achieve my desired results, I have used many data sets in for this analysis. As Covid is ever evolving, the data following suits and also evolves. During this academic year it has been my duty to stay up to date with the Covid data and keep on top of the new variants of the disease in terms of data selection. As a result of this, I have used various data sets through out the year, trying my best to keep with the most up to date and relevant data. There have been milestones which have occurred recently which, if it had occurred earlier, I would have focused a lot of my analysis on. These milestones include the serge in cases throughout India and the lack of oxygen supply in the country. If I had more time on this project or this tragedy had occurred earlier, I would have looked greater into this however, I have touched over it in my results section.

Owid-covis-data.xlsx (Ourworldindata.org, December 2020) consists of 55,324 rows and 48 columns. I was quickly about to find this information out using the nrow and ncol functions in R studio.

Below I will describe the data set in detail:

Iso\_ code: These are the unique codes which are allocated to each country in the world. The ISO or The International Organisation for Standardisation

Continent: One of the six continents used in this data set, Asia, North America, South America, Europe, Africa, and Oceania

Location: The country in which the case is located

Date: The date the cases were reported

Total\_ cases: Total number of reported cases

New\_ cases: Total number of new cases

Total\_ deaths: Total number of reported deaths

New\_ deaths: Total number of new deaths

Total\_ cases\_ per\_ million: Total number of reported cases per million

New\_ cases\_ per\_ million: Total number of new cases per million

Total\_deaths\_per\_million: Total number of reported deaths per million

New\_deaths\_per\_million: Total number of new deaths per million

Reproduction\_rate: How fast the virus is spreading between humans

Icu\_patients: Total number of patients in the intensive care unit in hospital due to the virus

Icu\_patients\_per\_million: Total number of patients in the intensive care unit in hospital due to the virus per million

Hosp\_patients: Total number of people in hospital due to the virus

Hosp\_patients\_per\_million: Total number of people in hospital due to the virus per million

Weekly\_icu\_admissions: The weekly number of people being admitted into the intensive care unit due to the virus

Weekly\_icu\_admissions\_per\_millions: The weekly number of people being admitted into the intensive care unit due to the virus per million

Weekly\_hosp\_admissions: The weekly number of people being admitted into hospital due to the virus

Weekly\_hosp\_admissions\_per\_million: The weekly number of people being admitted into hospital due to the virus per million

New\_tests: How many new Covid tests that have been carried out

Total\_tests: Total number of Covid tests that have been carried out

Total\_tests\_per\_thousand: Total number of Covid tests that have been carried out per thousand

New\_tests\_per\_thousand: How many new Covid tests that have been carried out per thousand

New\_tests\_smoothed: Total number of new tests carried out based on a 7 day period

New\_tests\_smoothed\_per\_thousand: Total number of new tests carried out based on a 7 day period per thousand

Positive\_rate: The rate of positive Covid cases over a 7 day period

Tests\_per\_case: The number of tests carried out per confirmed case

Tests\_units: The units used based on which location is reporting

Total\_vaccinations: Total number of vaccinations administered

Total\_vaccinations\_per\_hundred: Total number of vaccinations administered per hundred

Stringency\_index: The governments response to the virus



Population: The population of each location at the time

Population\_density: The total number of people in a given areas divided by square kilometres

Median\_age: The median age of the population

Aged\_65\_older: Total number of people aged 65 or older

Aged\_70\_older: Total number of people aged 70 or older

Gdp\_per\_capita: Gross domestic product at purchasing power parity

Extreme\_poverty: The percentage of the country living in extreme poverty

Female\_smokers: The percentage of woman who smoke

Male\_smokers: The percentage of men who smoke

Handwashing\_facilities: The percentage of the population with basic handwashing facilities

Hospital\_beds\_per\_thousand: Hospital beds available per thousand

Life\_expectancy: A persons expected life span

Human\_development\_index: An index used to measure the basic dimensions of human development

```
> nrow(Covid)
[1] 55324
> ncol(Covid)
[1] 48
```

Figure 1: Owid-covis-data nrow and ncol

I found this data set on ourworldindata.org. It gives a very detailed breakdown of the cases reported daily from each country. The columns include the continent, the location, the date, total cases, number of new cases, total deaths, number of new deaths etc. It also displays how many patients are in hospital at any given time including ICU. It lists how many vaccines there are per country. The population, population density, median age and how many people are aged 65 and older and 70 and older. The life expectancy per each location is also given. There is so much detail in this data set a lot of analysis can be conducted.

The second data set I have chosen to analysis is covid-19-geographic-distribution-worldwide.xlsx (Europa.eu, December 2020).

Below I will describe the data set in detail:

Date Rep: The date the case was reported

Day: The day of the month it was reported

Month: The month it was reported

Year: The year it was reported

Cases: The total number of cases reported

Countries and Territories: The country and or territory the case was reported

Geo Id: These are the unique codes which are allocated to each country in the world. The ISO or The International Organisation for Standardisation

Country Territory Code: Similar to Geo Id- a unique code for the country

Pop Data 2019: The population of the country based on 2019 data

Continent: One of the six continents used in this data set, Asia, North America, South America, Europe, Africa, and Oceania

Cumulative\_ number\_ for\_ 14\_ days\_ of\_ COVID-19\_ cases\_ per\_ 100000: Total number of cases reported over a 14 day period per 100,000

```
covid19 )  
> nrow(Covid2)  
[1] 61900  
> ncol(Covid2)  
[1] 12  
>
```

Figure 2: covid-19-geographic-distribution-worldwide nrow and ncol

There are 61,900 rows and 12 columns in this data set. I chose this set as I thought that it would be easier to break down the data into months, displaying the number of cases.

The third data set I will definitely be using is 2019\_nCoV\_data.csv (Kaggle.com, December 2020). I have not explored this set much yet however I believe it will be very useful as there are few columns. This data set will enable me to get a high level view, perfect for discovering quick answers.

Below I will describe the data set in detail:

Sno: The number correlating to each row

Date: The date the cases were reported

Province/ State: The province or state the cases are located

Country: The country the cases are located

Last Updated: The most recent date the data was updated

Confirmed: The total number of confirmed cases

Deaths: The total number of confirmed deaths

Recovered: The total number of confirmed recovered cases

```
> nrow(Covid3)
[1] 24709
> ncol(Covid3)
[1] 8
> |
```

Figure 3: 2019\_nCoV\_data nrow and ncol

There are 24,709 rows in this data set with 8 columns. As you can see this is a much easier data set to manage and analyse thus giving more time to analyse the larger data sets.

Below I will describe the data set in detail:

Date: The date the cases were reported

Country: The country the cases are located

Confirmed: The total number of confirmed cases

Recovered: The total number of confirmed recovered cases

Deaths: The total number of confirmed deaths

The next data set that I have used in my analysis is countries-aggregated.csv which I obtained from Kaggle.com once again. (Kaggle.com, April 2021)

```
> nrow(Covid19)
[1] 91776
> ncol(Covid19)
[1] 5
>
```

Figure 4: countries-aggregated nrow and ncol

Figure 4 shows that countries-aggregated data set has 91,776 rows and 5 columns. This data set is on the larger side in terms of my analysis. Although there are fewer columns, there are many rows. This data set gives a day by day breakdown of how many reported cases, deaths and most importantly, the number recoveries from when the Coronavirus outbreak first occurred up until the most recent date of May 13<sup>th</sup> 2021. This data set is useful when visualising the long term impact on the countries around the world in terms of their confirmed cases, confirmed deaths, and confirmed recoveries.

### 3.0 Methodology

The analysis of my data sets followed the KDD methodology. KDD stands for Knowledge Discovery in Databases. It is a repetitive process where evaluation measures can be improved, data mining can be polished, new data can be integrated and transformed in order to obtain the appropriate results. (Geeksforgeeks.org, 2019). This methodology is summarised in seven steps:

1. Data cleaning – the removal of any inconsistent or data outliers.
2. Data integration – the combination of multiple data sources.
3. Data selection – the selection and retrieval of relevant data to the analysis.
4. Data transformation – the transformation and consolidation of data into appropriate forms for mining by performing summary operations.
5. Data mining – the discovery of patterns within large sets of data.
6. Pattern evaluation – patterns in the data sets are identified.
7. Knowledge presentation – the development of visualization tools and knowledge techniques to display your data discoveries to the user, in a user friendly manner.

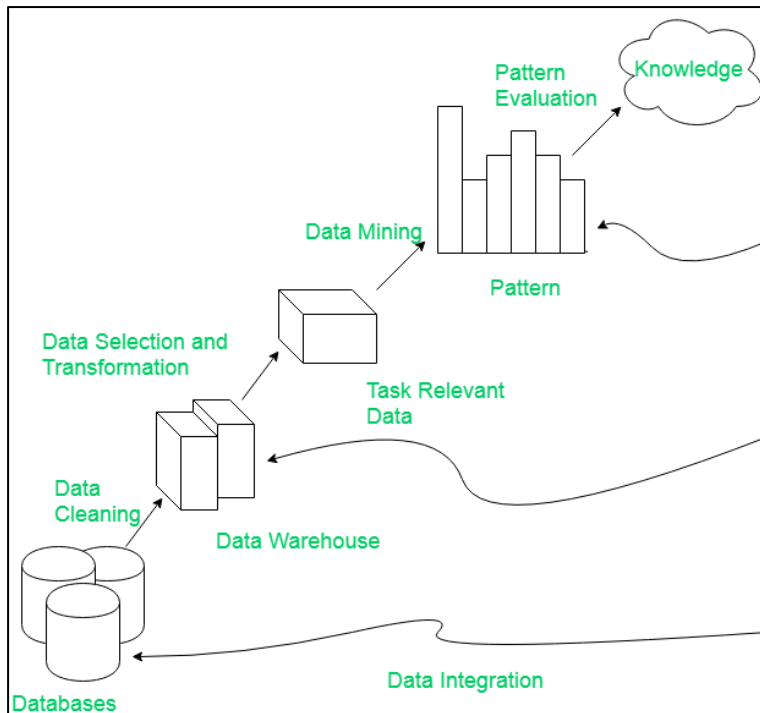


Figure 5 KDD Process

### 3.1 Data Cleaning

The first step of the KDD methodology was data cleaning. The analysis began with the removal of any noisy data. Noisy data is any data which has been deemed meaningless and not fit for use in the analysis. Data corruption was also removed during this step, any/ all corrupt data which would hinder the analysis.

The data set which I have chosen has multiple columns I deleted when conducting my preliminary analysis using Excel. Columns such as **new\_cases\_smoothed\_per\_million** and **new\_tests\_smoothed** were deleted as I found them to be not relevant to the data which I was analysing. There were five of these columns in total I decided not to use for my analysis.

### 3.2 Data Integration

To produce the best results possible for this analysis, I will be combining multiple data sets and comparing them with each other. At the moment, I have three definite data sets which will be used in my analysis. Owid-covis-data.xlsx (Ourworldindata.org, December 2020), covid-19-geographic-distribution-worldwide.xlsx (Europa.eu, December 2020), countries-aggregated.csv (Kaggle.com, April 2021), and 2019\_nCoV\_data.csv (Kaggle.com, December 2020).

### 3.3 Data Selection

The selection of the data deemed relevant to the analysis was conducted during this stage of the KDD methodology. The columns that were selected were used in the development of the visualization tools such as the pie charts and bar plots. This data was retrieved from the database in order for the selection to occur.

### 3.4 Data Transformation

Once the data cleaning, data integration and data selection were completed, it was time to perform a data transformation. Data transformation is the transformation and consolidation of data into appropriate forms for mining by performing summary and aggregation operations. By using the summary function, it should us all of the variables within the console in R. It displays which variable type is in each column.

After these operations were conducted it was found that there were no errors in relation to the viable types for all pf the columns across the three data sets. The only transformation which was needed was when the data was in a worded for, it was tasked to change this or transform it into a different data type, a function, to be able to perform the preliminary analysis on it.

### 3.5 Data Mining

After preforming the data transformation, it made the data mining step much more obtainable. The changing of data types ie: changing worded data from a character to a factor, it allowed for the implementation of the data mining process. Data mining is the discovery of patterns within a large data group.

### 3.6 Knowledge Presentation

The final step in the KDD methodology is knowledge presentation. This is the stage where you get to be creative and create your visualizations. It is now, after all of the previous stages have been completed, you can truly analysis your work and be able to visualize the patterns which have emerged from the data selected.

## 4.0 Analysis

I used the KDD methodology approach when conducting my analysis. This is described in detail in section 3.0.

I will now discuss my analytics approach for each of the technologies I have chosen to use in this project.

## **SPSS**

SPSS was the technology I used most throughout this project. I used it to locate the statistical information about the data sets. I also used it to create graphs for this analysis. SPSS is a great tool for any data analyst as it is easy to get used to and provides a great variety of internal tools to choose from

## **Microsoft Excel**

Microsoft Excel was used to obtain the raw data. I used the filters in excel, saved a copy to a workbook, and used the filtered data for many of my graphs. I found this the most efficient and effective way to carry out my analysis. I also made some pivot tables to aid with the analysis.

## **R Studio**

R Studio was used to develop easier graphs and to produce basic information. I decided to do that as I have become quite familiar with R Studio over the past year, and I want to challenge myself by replicating the output of data information and graphs using a different technology and methods to what I was used to. R Studio still does feature quite a bit in this project, however.

## **Jupyter Notebook**

Jupyter Notebook was the newest technology to me for this analysis. I decided to use this as I have only used it a hand full of times and wished to challenge myself. I spent a lot of time figuring out the language of Python to further my analysis on Covid 19.

One approach I wish I had the chance to use in this analysis was to create dashboards in Power BI. If I had sufficient time, I would have developed a dashboard similar to the dashboard in the "Covid Tracker Ireland" application on IOS and Android developed by the government. It would have taken up too much time to develop this as well as conducting the analysis itself. It would have been an extremely useful tool especially for a Covid project as it would have refreshed the data automatically saving me looking for the most up to date data.

## 5.0 Results

Below is the first graph I quickly pulled together. It is a pie chart which represents where the Coronavirus cases are distributed in terms of continents. As seen in the below image, Asia, Africa, and Europe are pretty even from a visual point of few with each contributing to in and around a quarter of the total cases. The final quarter is divided into North America, South America, and Oceania. Oceania consists of places such as New Zealand and Australia.

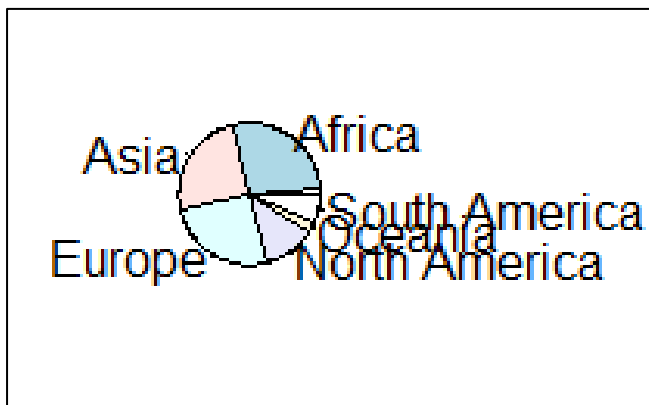


Figure 6: Pie chart of continents

```
pie(table(Covid$continent))
```

Figure 7: Pie chart of continents code

This graph was made using the Owid-covis-data.xlsx (Ourworldindata.org, December 2020) data set.

The following graph was created used the 2019\_nCoV\_data.csv (Kaggle.com, December 2020) data set. It represents the correlation between the total number of confirmed cases, versus the total number of recoveries.



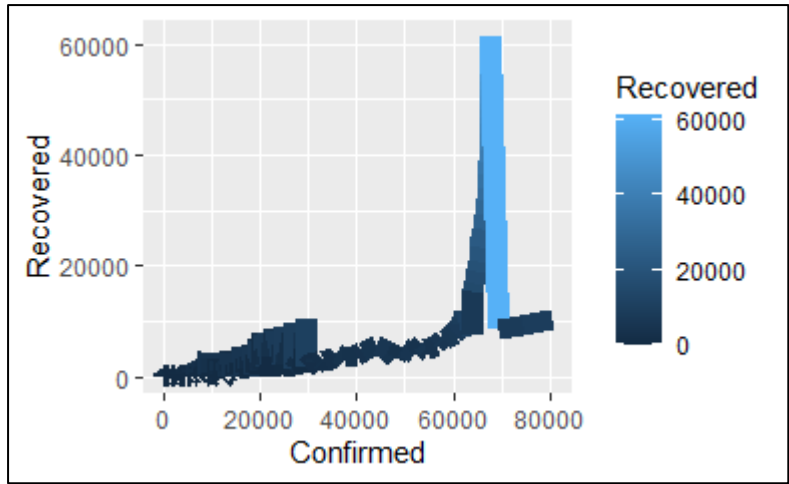


Figure 8: Line chart no cases vs recovered

```
qplot(data=Covid3,x=Confirmed,y=Recovered, geom = "line", colour=Recovered, size=I(4))
```

Figure 9: Line chart no cases vs recovered code

Using excel, I was able to create a pivot table displaying all the reported cases in each country for the month of December. There are 215 countries listed so I will not include all of them into this report, but for an example please see below:

	A	B
1	month	12
2		
3	<b>Row Labels</b>	<b>Sum of cases</b>
4	Afghanistan	3429
5	Albania	10905
6	Algeria	9881
7	Andorra	626
8	Angola	1085
9	Anguilla	6
10	Antigua_and_Barbuda	7
11	Argentina	79366
12	Armenia	13558
13	Aruba	211
14	Australia	138
15	Austria	41854
16	Azerbaijan	57679
17	Bahamas	178
18	Bahrain	2356
19	Bangladesh	28126
20	Barbados	17

Figure 10: Excel pivot table

The model description gives an overview of the data in the data set countries-aggregated. As there are not many rows in this data set the overview is minimal. See figure 11.

### PPlot

Model Description		
Model Name		MOD_2
Series or Sequence	1	Recovered
	2	Confirmed
	3	Deaths
	4	Date
Transformation		None
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period		No periodicity
Standardization		Not applied
Distribution	Type	Normal
	Location	estimated
	Scale	estimated
Fractional Rank Estimation Method		Blom's
Rank Assigned to Ties		Mean rank of tied values
Applying the model specifications from MOD_2		

Figure 11: SPSS Model Description

Below is the descriptive statistics for all numerical data in the countries-aggregated data set. This table was produced in SPSS and it includes information such as:

N = the number of values

Minimum = the lowest figure

Maximum = the highest figure

Mean = the average

Standard Deviation = how the figure differs from the mean figure

From this we can see the first date in the data set the Covid cases were detected, January 22<sup>nd</sup>, 2020. We also have the most up to date data included in this data set which is May 13<sup>th</sup>, 2021. We can also identify the maximum number of confirmed cases, deaths, and recoveries in the data set. This gives us a quick and easy representation of the data.

→ **Descriptives**

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Date	91776	22-JAN-20	13-MAY-21	16-SEP-20	137 23:41:3...
Confirmed	91776	.000000000	32852871.00	257084.4048	1464225.101
Deaths	91776	0	584487	6199.20	29705.505
Recovered	91776	.000000000	19734823.00	150397.6020	765409.6372
Valid N (listwise)	91776				

Figure 12: SPSS Descriptive Statistics

Figure 13 shows a Q-Q Plot of all recovered cases in the countries-aggregated data set. This graph represents all of the observed values of recovered cases within the data set and compares it to the expected values based on the given values. We can see that in the beginning the observed values is follows the expected values linear line however, as time goes by the number of recovered cases does not meet the expected number. We can conclude that this is the case due to the number of reported cases decreasing thus it results in a lower number of recoveries. Another reason for this could be the new variants of the disease being stronger meaning it is taking people to recover from it.

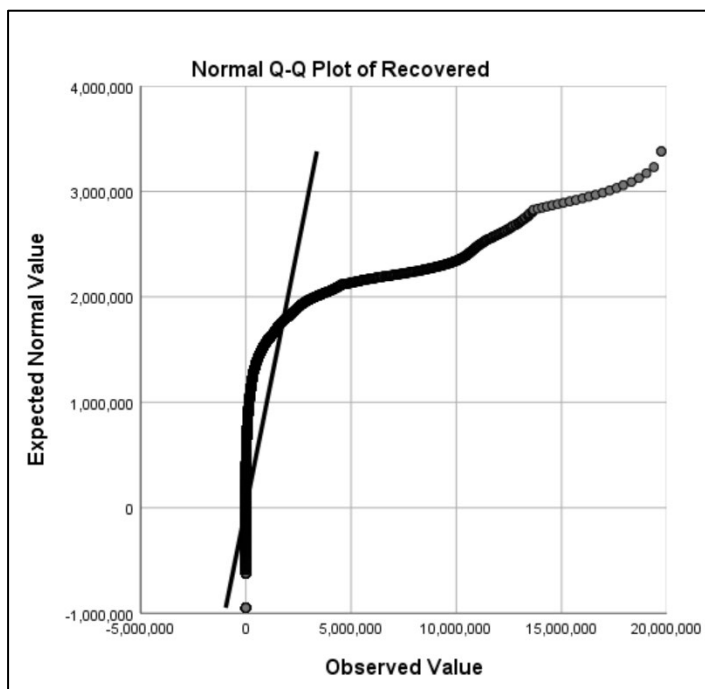


Figure 13: SPSS QQPlot Recovered

Below is the Q-Q Plot for all confirmed cases of Covid 19. Again, much like the recovered cases, the observed values and the expected values complement each other however, as time goes by, there are less cases reported than expected. This can be the result due to many reasons. The introduction of mandatory mask wearing, and social distancing helped to fight the disease resulting in fewer cases but also the establishment of the vaccination has played a huge roll in the control of the virus. The vaccination has reduced the number of cases dramatically worldwide. However, as we can see towards the end of the observe value line, there is a spike in confirmed cases. I have concluded that this is due to the mass influx in cases in India. India accounts for 23,703,665 cases and 258,351 deaths as of May 13<sup>th</sup>, 2021 according to worldometers.info. (worldometers.info,2 021)

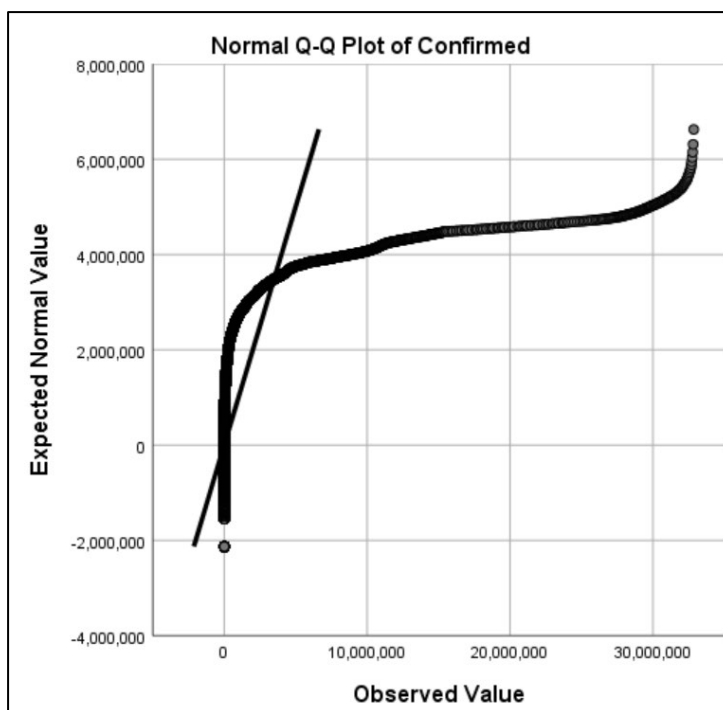


Figure 14: SPSS QQPlot Confirmed

The Q-Q Plot for the number of deaths mirrors nearly exactly the same as the confirmed cases graph. See figure 15 below.

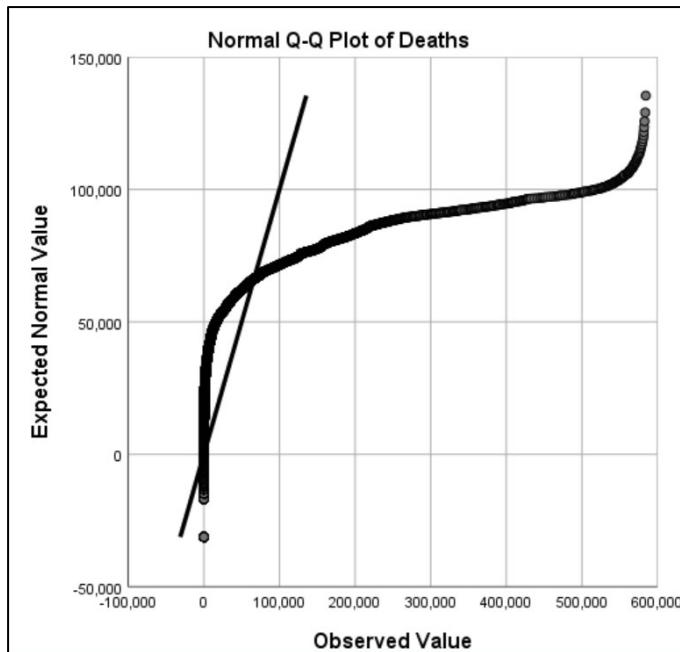


Figure 15: SPSS QQPlot Deaths

Below in figure 16 we have a bar chart made in SPSS displaying the difference in the number of confirmed coronavirus cases between Ireland and New Zealand. I chose to compare Ireland to New Zealand as their Covid prevention programme was/ is first class. At the first hint of the virus entering the country, their Prime Minister Jacinda Kate Laurell Ardern completely shut down the country and close the borders. No body was allowed in or out of the country for a period of time. From January 8<sup>th</sup>, 2021 to January 20<sup>th</sup>, 2021 New Zealand achieved zero covid meaning there were zero covid cases reported during this time. Whereas in Ireland as mentioned earlier, this was our peak time in cases with over 8,000 reported daily. (worldmeters.org, 2021)

The number of Covid cases in New Zealand spiked on April 8<sup>th</sup>, 2021 with just 24 cases. This is an astonishing figure when we compare that to Ireland. (worldmeters.org, 2021) New Zealand has a similar population to Ireland with 4.9 million inhabitants reported in 2018 where as Ireland has 4.9 million inhabitants as of 2019. Having two countries with roughly 100,000 persons in the difference having such a vast difference in the number of reported cases is shocking. (worldmeters.org, 2021)

According to Worldmeters.org, Ireland has had a total of roughly 255,000 cases to date with 5,000 people sadly losing their lives whereas New Zealand's total number of cases is just 2,645 and the total number of deaths is 26. Almost double the amount of people in Ireland have lost their lives compared to the total number of cases reported in New Zealand. See figure 16 to see the comparison in the number of deaths between the countries.

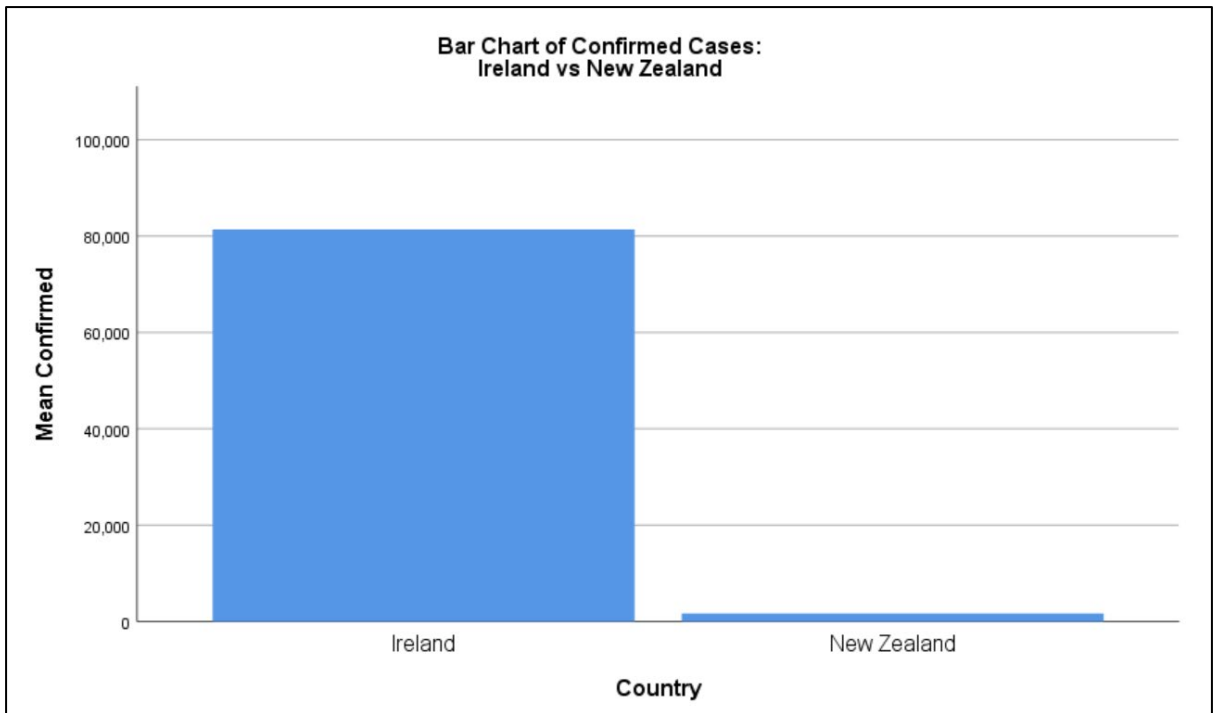


Figure 16: SPSS Bar Chart Ireland vs New Zealand

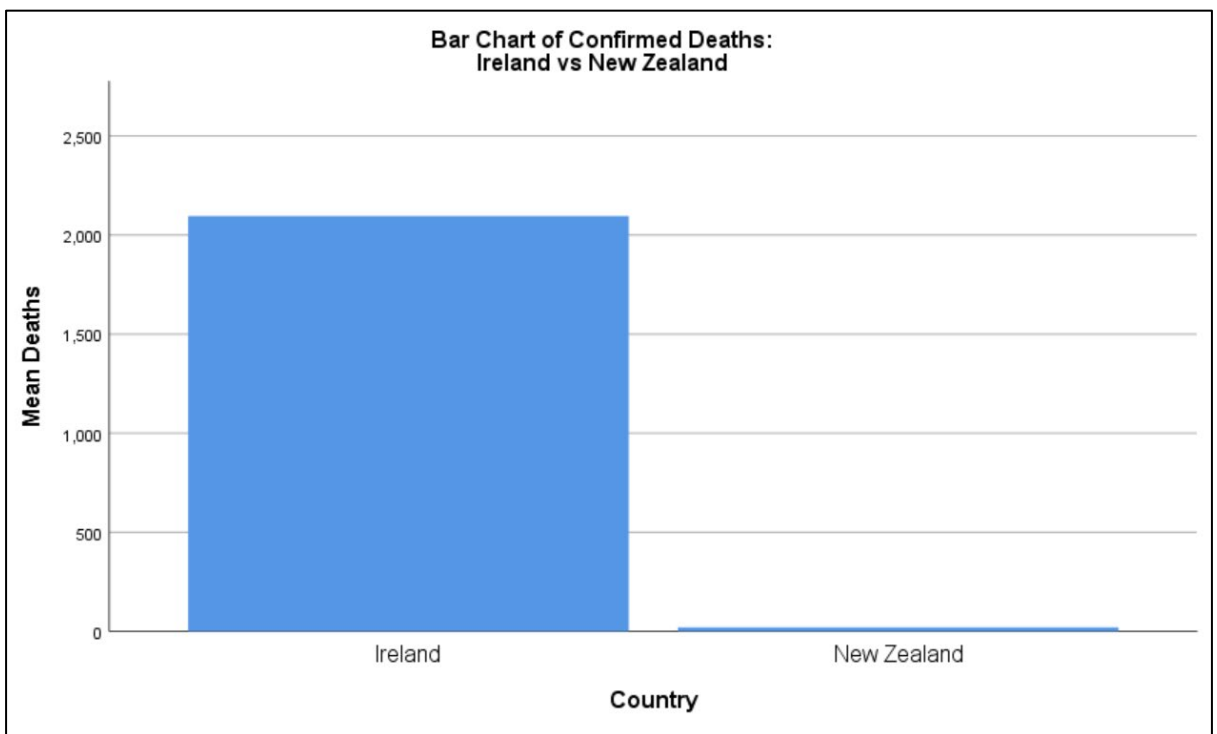


Figure 17: SPSS Bar Chart Ireland vs New Zealand

There is simply no comparison between the two countries, the difference is astonishing.

Figure 18 was made in SPSS and is a more advanced version of the pie chart found under figure 5. We can clearly see the distribution of the virus throughout the continents. Asia has the greatest number of confirmed cases with 27.33% of the cases worldwide. This is perhaps due to the surge of cases in India. North America is a close second with 26.38%. Next is Europe is 22.09% followed by South America, 20.21%, Africa, 3.92%, and Oceania, 0.08%.

My hypothesis as to the reason the African reported cases are so low is due to the lack of funding for tests. I believe that there are far more cases than this, they simply go undocumented.

Figure 19 displays the distribution of deaths between the continents. It mirrors the reported cases' pie chart with some percentages varying slightly.

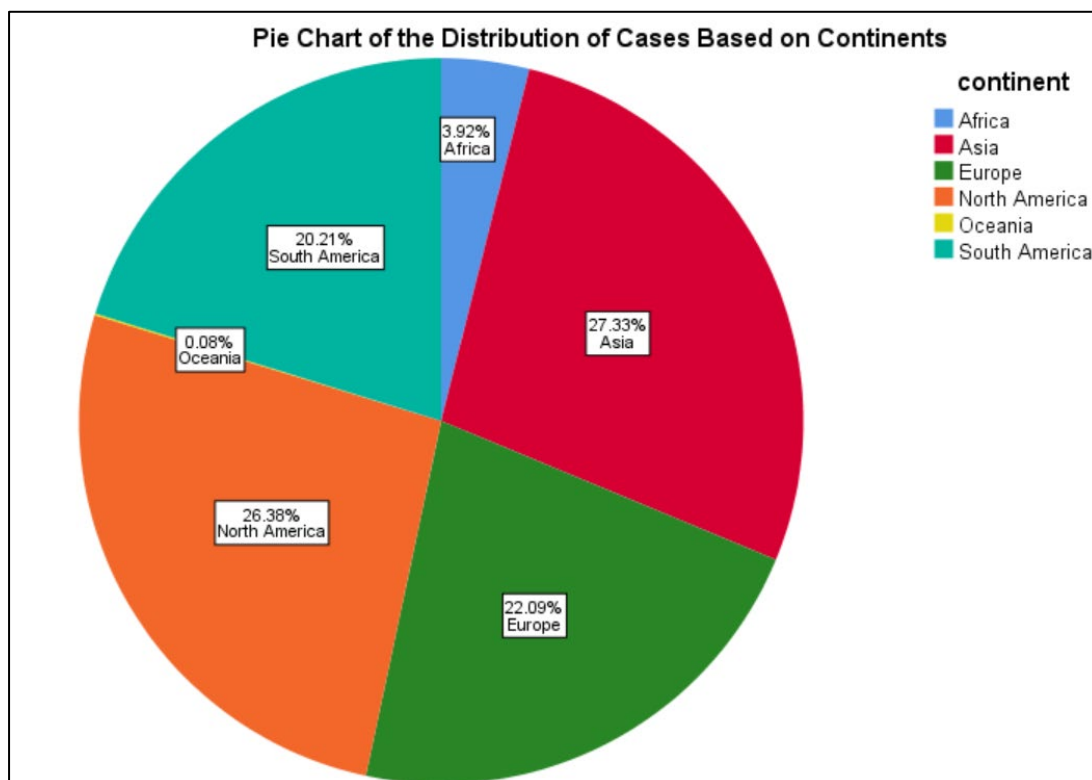


Figure 18: SPSS Pie Chart Continents

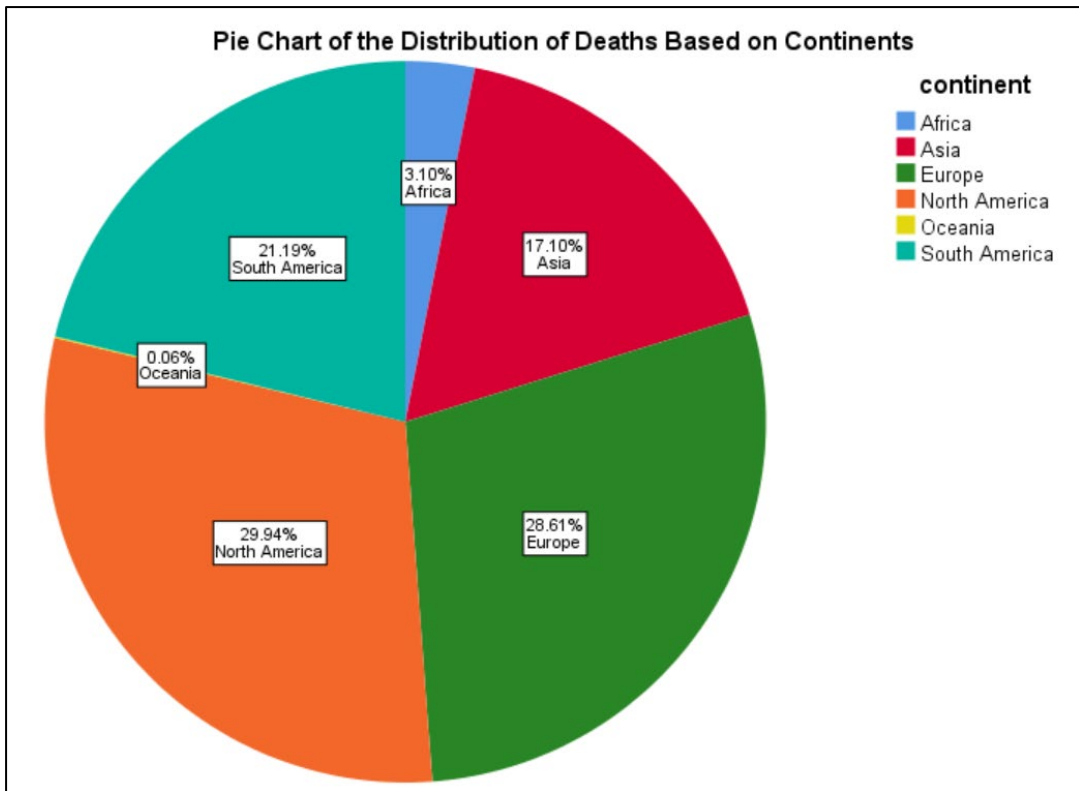


Figure 19: SPSS Pie Chart Continents

In figure 20, I investigated whether there was a correlation between the number of Covid cases and children going back to school for the first time since mid March 2020. The graph represents data from Ireland between the months of August to October 2020 (8 = October, 9 = September, 10 = October) The results are represented on a line graph. I was not surprised with the results. Cases began to fall towards the end of August and as soon as September began, there was a huge spike in cases. This surge could have been caused by parents and children going back to school shopping to collect their school supplies. Then, in October, there is another increase in the number of cases which would be the results of the children going back to school in September. This space in time is the time it takes for a person to show their symptoms and get tested resulting in a positive result.



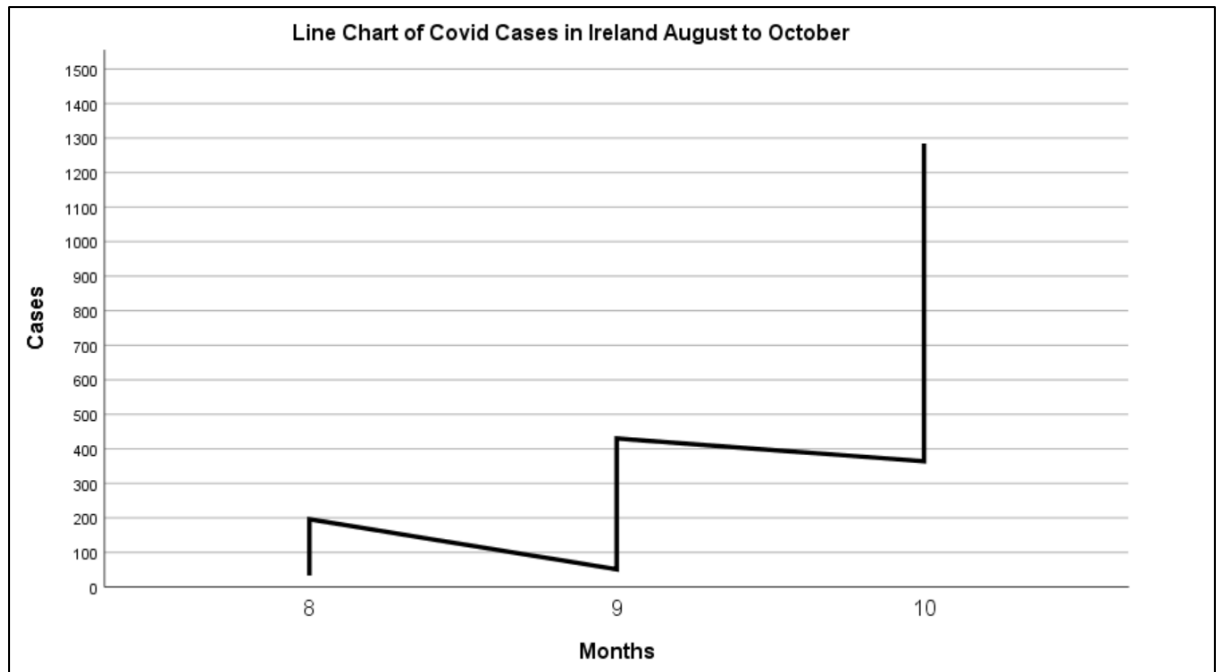


Figure 20: SPSS Line Chart Back to School

To challenge myself, I decided to create this line chart in Python using Jupyter Notebook. This was a challenge for me as I have only used Python once or twice before. It was fresh in my mind as I have just completed another project using this technology. This graph represents daily confirmed cases, daily recovered cases, daily deaths, and the total number of cases. This graph is a collect of the world's Covid data all pulled together into one graph.

A positive which I have found from this graph is that the number of deaths line (in green) is almost parallel with the x axis meaning compared to the total number of cases, the death rate is very low. Another positive that I have noticed is that the recovered line (in yellow) is constantly growing meaning people are always on the road to recovery.

A negative I am noticing whilst reading this graph is that the cases are rising, not falling. Ideally, we would like to see the graph skewing the opposite way. It is currently having a left skewed distribution, meaning the cases are rising, we want to see a right skewed distribution meaning the number of cases is falling. From the graph we can see that there was a small decline in cases around the beginning of December, however, after this, it surged once again.

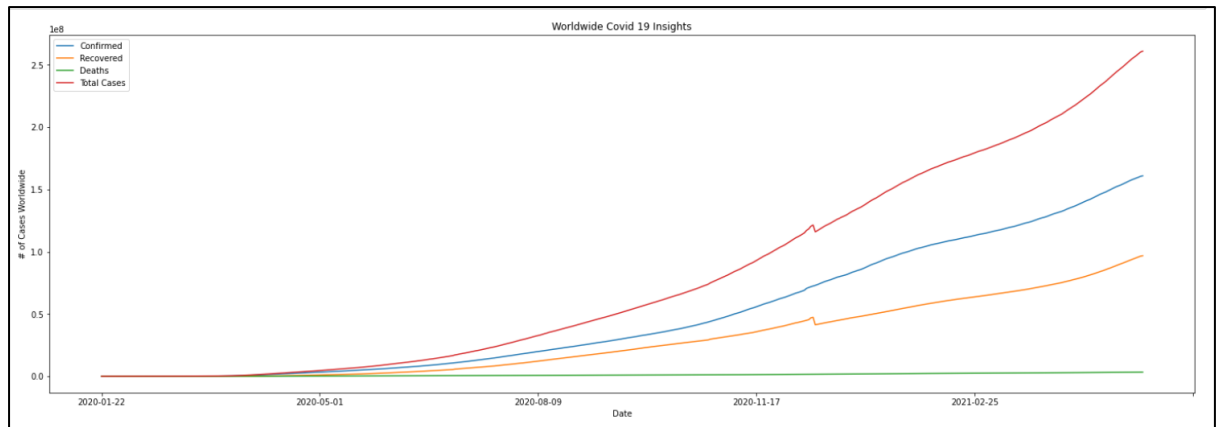


Figure 21: Python line graph worldwide data

## 6.0 Conclusions

Revising my aims I set at the beginning of this project:

1. Determine the relationship between Covid cases rising in Ireland and schools reopening
2. Compare the response to Covid 19 between Ireland and New Zealand, two countries with similar populations but a huge difference in cases
3. Investigate the global distribution of the virus, which areas were most impacted, where did the most deaths occur etc.

I believe that I achieved all of the goals I had set out at the beginning of this report. I discovered that there was a relationship between children going back to school and Covid cases rising. I compared and identified the differences between the Irish and New Zealanders (Kiwis) initial response to the virus. I identified the global distribution of the virus and identified the most affected areas in the world.

I discovered that Covid cases in Ireland began to fall towards the end of August and as soon as September began, there was a huge spike in cases. This surge could have been caused by parents and children going back to school shopping to collect their school supplies. Then, in October, there is another increase in the number of cases which would be the results of the children going back to school in September. This space in time is the time it takes for a person to show their symptoms and get tested resulting in a positive result. See figure 20 above for results.

I conducted a detailed comparison between Ireland and New Zealand in terms of each country's initial response to the virus. A very sad reality for Ireland I discovered was that according to Worldmeters.org, Ireland has had a total of roughly 255,000 cases to date with 5,000 people sadly losing their lives whereas New Zealand's total number of cases is just 2,645

and the total number of deaths is 26. Almost double the amount of people in Ireland have lost their lives compared to the total number of cases reported in New Zealand. Although now, over a year on, we have the same measures in place as New Zealand, it took losing thousand of Irish lives for our government to realise the severity of the disease and to shut down the country entirely. Too little, too late? See figures 16 and 17 for reference.

I investigated the distribution of the virus throughout the world based on six continents, Asia, North America, South America, Europe, Africa, and Oceania. Figure 21 displays the full spread of the virus throughout the world including total cases, confirmed cases, deaths and recovered cases. Figures 18 and 19 are pie chart representations of the distribution of the total cases and total deaths due to the virus across the world based on continents. Asia has the greatest number of confirmed cases with 27.33% of the cases worldwide. This is perhaps due to the surge of cases in India. North America is a close second with 26.38%. Next is Europe is 22.09% followed by South America, 20.21%, Africa, 3.92%, and Oceania, 0.08%.

A major advantage of conducting my analysis on Covid 19 and the current global pandemic is that it is a global issue and very current. Because of this, there is an abundant amount of public data available to analyse. Although there are great benefits from having many data sources to choose from, it can also be seen as a disadvantage as sometimes there is too much choose. Also, with having many data sets to choose from, this leaves a lot of room for false data or 'dirty data' to be presented. I have had to be extremely careful when selecting my data sets to ensure they are coming from a reliable source such as Kaggle.

A disadvantage is the fact that the data is constantly change, each day new data is being released in every country. This has caused problems for me when creating visualizations as I always have to ensure I have the latest version of the data set.

## 7.0 Further Development or Research

If a situation arose where I was granted extra time, there are many things I would do differently or would like to add to the project. I believe that even if I had all the time in the world, it would still not be enough. There is always more information to investigate especially with a topic such as Covid 19 that changes day by day.

I had hoped to have looked into using Power BI to further my analysis by creating dashboards to automatically refresh the data I used without manually having to do it. However, I was unable to achieve this goal as I had an insufficient amount of time. I had visited this technology during my third year work placement at Irish Water where I was tasked to create dashboards for the company. If I were granted more time, this would be on the top of my to do list.

My original project idea was to compare Covid data from public sources vs private sources. I had wanted to compare the testing data that the Irish government has lead vs Ashford Studios in County Wicklow. Ashford Studio have maintained full production of their tv show 'The Vikings' throughout the entirety of the Covid period and through lockdowns. They have been able to achieve this by hosting their own testing centre. They test their staff multiple times a week with next day results being produced. In the case where somebody tests positive, that section of the production would be stopped and everybody in that bubble would isolate for 14 days. The cast and crew who were on set are tested three times a week and all other staff is tested twice a week. The tests are then taken to Belfast each day to get tested in order to receive the results the next day.

I would have loved to have completed this project however, at the beginning my first supervisor suggested that this would not be a good idea as it would have been too difficult to gain approval to access the private data for an undergraduate degree project saying that the access would be granted at masters' level. This was disappointing to me however, I followed his advice and continued on the project with solely the public data. Given the chance to progress with this current project I would love to gain access to the private data now that I have completed a full analysis on the public data. I would like to compare their testing techniques ie: how fast they can turn around test results vs the government, the positivity rate; public vs private and measure taken to prevent the spread.

For the past number of months, countries all over the world have been involved in the vaccine rollout. Given extra time and resources I would have liked to look further into this and investigate the reasons as to why it is taking Ireland so much longer to vaccinate such a small population compared to other countries such as Italy. I would also like to investigate the distribution of the vaccine and visualise how skewed it really is. I would have like to have developed a model which displays the even distribution of the vaccination throughout the world will a focus on third world countries or regions which may be struggling more than other with the rapid spread of the virus.

This brings me onto the final item I would add to this report given the opportunity. I would like to explore India further. I had mentioned it earlier in the report, however there is so much heart break happening over there with the spread of the virus. They are battling to fight the spread; however, it is very much an uphill battle for them with the supply of their oxygen rapidly diminishing. I believe this investigation would be an eye opening exploration and would hopefully discover some alternate measure to help in supporting them.

## 8.0 References

Who.int. (2021). *Coronavirus*. [online] Available at: [Coronavirus \(who.int\)](https://www.who.int/coronavirus) Accessed April 2021

Worldometers.info. (2020). *Coronavirus Update Live*. [online] Available at: [Coronavirus Update \(Live\): 77,397,288 Cases and 1,703,368 Deaths from COVID-19 Virus Pandemic - Worldometer \(worldometers.info\)](https://www.worldometers.info/coronavirus/) Accessed December 2020

Geeksforgeeks.org. (2019). *KDD Process in Data Mining*. [online] Available at: <https://www.geeksforgeeks.org/kdd-process-in-data-mining/> [Accessed March 2020]

Worldometers.info. (2020). *New Zealand Coronavirus*. [online] Available at: [New Zealand Coronavirus: 2,121 Cases and 25 Deaths - Worldometer \(worldometers.info\)](https://www.worldometers.info/coronavirus/new-zealand/) Accessed December 2020

Worldometers.info. (2020). *Ireland Coronavirus*. [online] Available at: [Ireland Coronavirus: 79,542 Cases and 2,158 Deaths - Worldometer \(worldometers.info\)](https://www.worldometers.info/coronavirus/ireland/) Accessed December 2020

Europa.eu (2020) *Coronavirus Datasets*. [online] Available at: [COVID-19 Coronavirus data - Datasets \(europa.eu\)](https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&plugin=1) Accessed December 2020

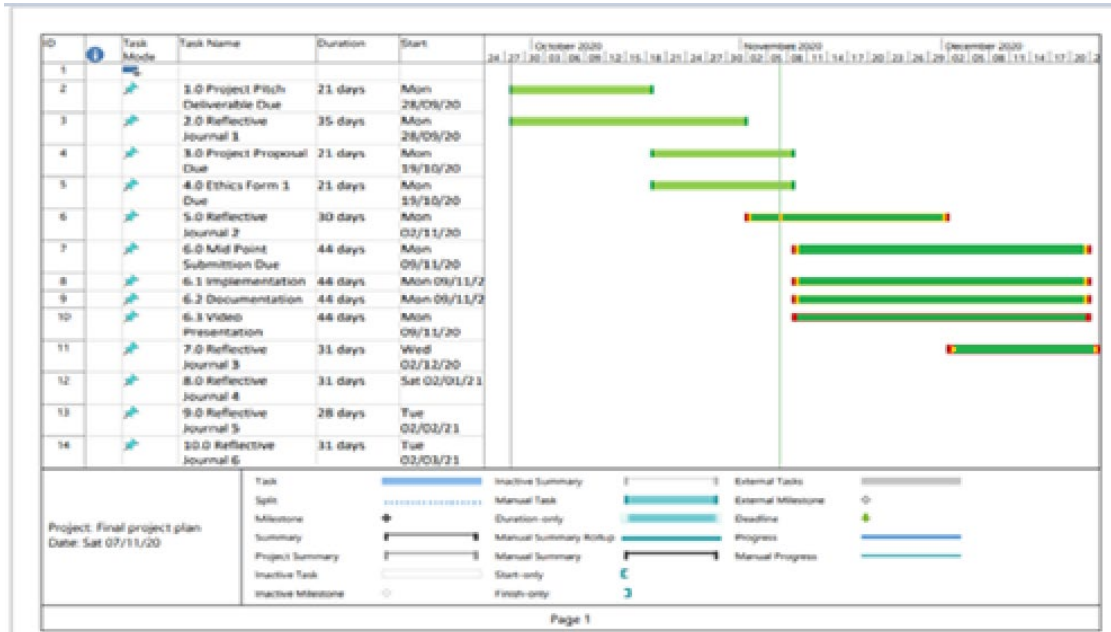
Ourworlddata.org. (2020) *New Zealand Coronavirus*. [online] Available at: [New Zealand: Coronavirus Pandemic Country Profile - Our World in Data](https://ourworldindata.org/coronavirus-zealand) Accessed December 2020

Kaggle.com. (2020) *Coronavirus dataset*. [online] Available at: [COVID-19 Coronavirus Dataset | Kaggle](https://www.kaggle.com/coronavirus) Accessed December 2020

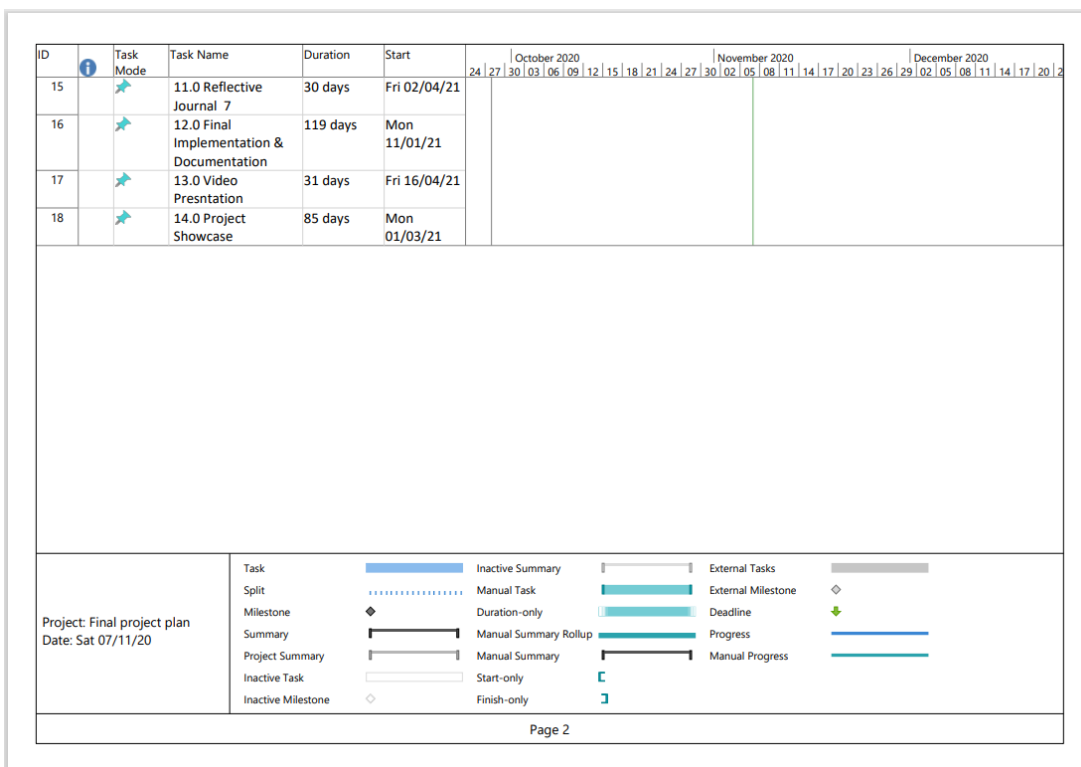
Kaggle.com. (2021) *Coronavirus Data*. [online] Available at: [Real-time Covid 19 Data | Kaggle](https://www.kaggle.com/coronavirus) Accessed April 2021

## 9.0 Appendices

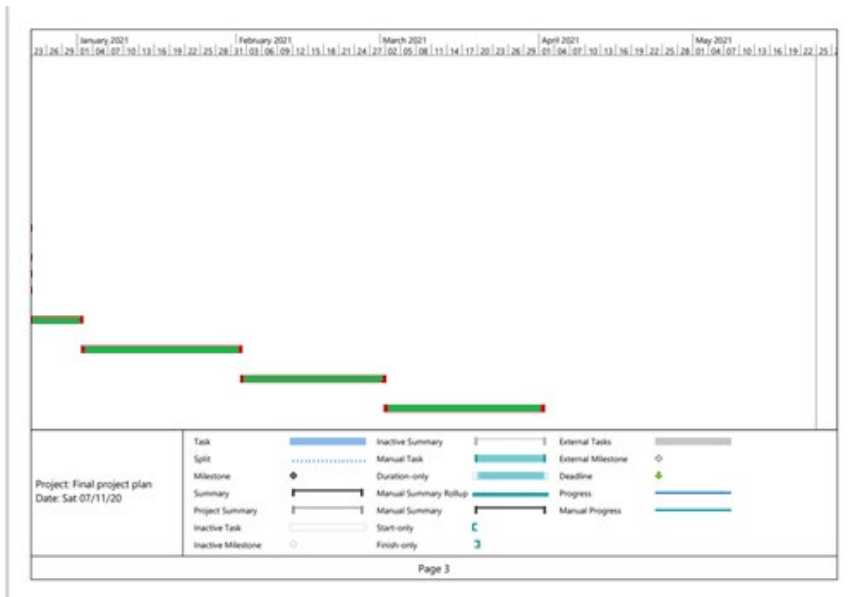
### 9.1. Project Plan



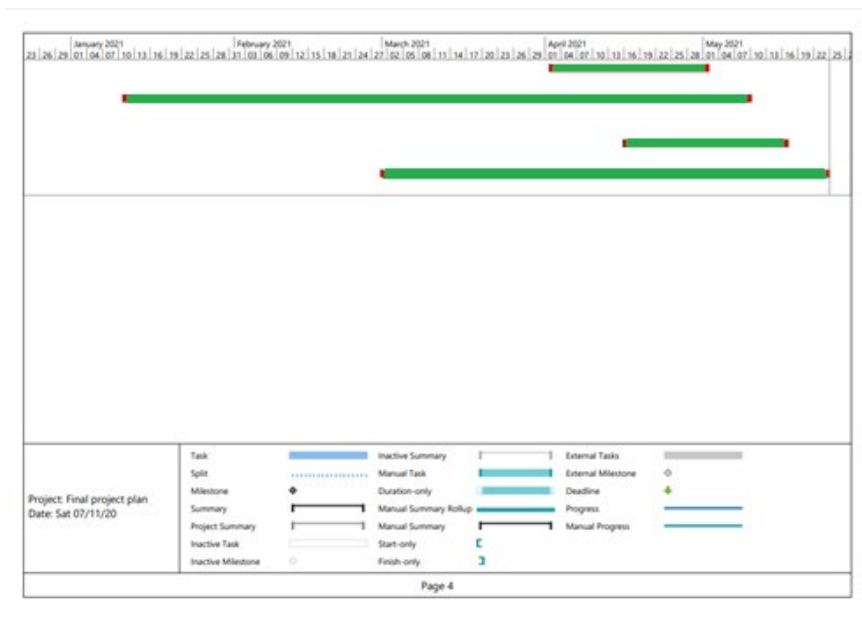
Gantt Chart Page 1






Gantt Chart Page 2



Gantt Chart Page 3



Gantt Chart Page 4

- Key:**
-  Completed Task
  -  Pending Task
  -  Not Yet Started Task

## 9.2. Reflective Journals

### Reflective Journal 1 October 2020

Ciara Quigley x17408654 BSHTM4

This month started on Monday September 28<sup>th</sup>, 2020, our first Final Year Project class. We began the class at 9am with Frances Sheridan where she gave us an introduction to the module and an outline of how the next year will pan out. She went over deadlines, objectives and other vital components of our final year project. Later, at 10am, we had an Ethics seminar with Cristina Muntean. During this seminar Cristina explained to us under what circumstances you would need to fill in an Ethics form and how to go about doing so. It all depends on what your project idea is and if it involves people or not. Our first ethics form submission is on Sunday November 8<sup>th</sup>, 2020. As our project ideas change and/ or develop, we may need to resubmit our forms with updated information. These dates are as follows, December 13<sup>th</sup>, 2020, February 7<sup>th</sup>, 2020, March 7<sup>th</sup>, 2020, and April 12<sup>th</sup>, 2020.

During week two Frances introduced our project pitches. She reminded us to keep thinking about an idea for the project and gave us a detailed explanation on what is required in the project pitch which was due Sunday October 18<sup>th</sup>, 2020. Frances also told us about these reflective journals which we are required to submit on the 1<sup>st</sup> of every month to track our progress with our projects.

Week three began with a Q&A session with Frances to give us students a chance to ask any question we had relating to our video project pitches. This session lasted for about an hour. After this is had a one-hour seminar with Pat Delaney all about project time management. He gave us some great tips and tricks to keep on top of our work load for the next year and reminded us to spread it all out, so we are not left with a mountain of work to do last minute. On Sunday October 18<sup>th</sup>, we submitted our project pitches in the form of a video. My project idea is to analyse data from Ashford Studios in relation to the filming of the TV Show the Vikings and how they are managing to cope with Covid 19. I also want to take this data and compare it to the governments data regarding Covid testing as The Vikings have their own testing team and doctors on set.

At 9am on Monday of week four, we had our project proposal session with Frances once again. During this session we looked at what is required in the written project proposal and she also gave us a template. Later at 10am we had a Commercialisation seminar with Anu Sahni, and she gave us an overview of the process of commercialising our projects if we so wish to do so.

We had no class in week five due to a bank holiday however during the week we did get told who our project supervisor is for the next year. My project supervisor for this year is Sachin



Sharma. I have yet to make contact with him so if I do not hear from him by the evening of Monday November 2<sup>nd</sup>, I will get in contact with him on Tuesday morning to see if my project idea is a viable one for this year.

## Reflective Journal 2 November 2020

Ciara Quigley x17408654 BSHTM4

On November 2<sup>nd</sup> we had a seminar at 9am with Michael Bradford. In this seminar he discussed the topic of Finding Good Data and Preliminary Analysis. This seminar was particularly beneficial to us students in the data analysis stream. Later, at 10am, Paul Stynes discussed Gathering and Documenting Requirements. This seminar will also come in use to us.

Week 7 was reading week. During reading week, I met with my project supervisor for the first time. This is where I confirmed my project idea and it set me off on the right foot.

Week 8 at 9am we had a Report Writing and Referencing Session with Keith Brittle. This was a good refresher seminar for not only our final year project documentations, but also it helps with our other modules.

Week 9 found us in a seminar with Emer Thornbury the new Computing Support lecturer. She took us through GitHub once again. This was a nice refresher course and it allowed us to ask questions as we needed to.

Finally, this week, week 10 was our final class with Frances. During this final class she took us through the mid point deliverables and the grading rubrics. She gave us tips and tricks to aid us with our time management as the stress begins to pick up.

Regarding my final year project, not a whole lot of work happened. Outside of the project I had four CAs, a presentation, and an exam worth 50% all during the month of November. I did not get a lot of spare time to work on the deliverables for the mid point presentation for the project. I did however get a chance to meet with my supervisor. We decided to modify my project idea a little bit so instead of comparing public vs private data relating to Covid 19, we decided to do a more in depth analysis of the public data as we found that there would be too many ethical problems if I included private data.

## Reflective Journal 3 December 2020

Ciara Quigley x17408654 BSHTM4

This month was spent doing work towards the mid point submission. Due to other modules having a lot of deadlines before December, they took priority leaving this month for the mid point work. For the mid point submission, we were to submit our mid point implementation, documentation, and video presentation. Due to Covid19 and the restrictions which come along with it, our mid point presentation had to be a recording for the internal and external faculty to grade. This was all due on Tuesday December 22<sup>nd</sup>.

To start off my work on the submission, I took another look into my project proposal as the first submission was not graded. I made some changes and completed sections which I was not able to complete prior to this point. Then I took a look at R Studio and played around with the three data sets I had chosen. I ran some code to cleanse the data and discard any dirty data. I was then able to create some graphs which I was able to show in my mid point report. I completed the remainder of the report with the relevant information required. As this was only the mid point submission, there were some sections of the report which I could not yet contribute to or complete in full. As we have not yet learned all of the skills required to complete the project, a lot of the report included tasks which I plan on doing, these tasks were supported with a Gantt chart.

Finally, I recorded a video presentation. This presentation consisted of three videos uploaded to YouTube under a private listing meaning you need the link to the video in order to view it. The first video was of me, it was an introduction of myself and of the project. The video which followed this was a screen recording of my PowerPoint presentation with a voice over of me describing and discussing the slides. The third and final video was again a screen record with a voice over but this time it was a look into my code in R Studio and the graphs which I had created.

## Reflective Journal 4 January 2021

Ciara Quigley x17408654 BSHTM4

January was a quite month in terms of working on my final year project deliverables. For the first two weeks of the month, I was completing my TABA, terminal based assessments, for my other five modules. This was a very stressful time for me as there were so many heavily weighting assignments due within a two week period. The week following this was a week off, I decided to take this week for myself to regroup and prepare myself for my semester in college. A lot of this time was spent organising my at home work space and also some relaxing.

Classes for semester two, my final semester, commenced on Monday 25<sup>th</sup> January 2021. We had a class with Frances Sheridan on this day which was a recap of what we have completed and also an overview of the work to come. On Wednesday 27<sup>th</sup> January, I was informed by Frances Sheridan that my project supervisor was changed from Sachin Sharma to Enda

Stafford as a result of a change within the faculty. I was in contact with Enda, and we plan to meet this week to discuss my project and inform him of my progress thus far and what I plan on doing in the next few months. We will also discuss the results of my mid point submission once those results are released.

### Reflective Journal 5 February 2021

Ciara Quigley x17408654 BSHTM4

On February 1<sup>st</sup>, 2021, we received our mid point submission results. I achieved a 2:1 grade which I was happy with as this is grade; I hope to achieve overall. The first of February there was no class instead we were instructed to work independently on our projects. On Monday 8<sup>th</sup> of February, we had a seminar on Frameworks with Michael Bradford, on February 15<sup>th</sup>, 2021 there was a seminar on Cybersecurity with Vikas Sahni and on the 22<sup>nd</sup> of this month, Enda Stafford held a seminar on unit testing.

The more I work on my project the more I realise how fast my data becomes outdated. As my project is based on Covid 19 data, it changes every single day. After some thought, I have decided to work on the parts of my report which do not require data and also focus on my other modules. I plan to choose a cut off date which both gives me plenty of time to complete the project and also gives the me most up to date data. By doing this, I will improve upon my skills using R Studio and SPSS and thus will result in an improvement in my project in the long run. I have also been working on other elements of my project such as the project profile for the website for the end of year showcase.

### Reflective Journal 6 March 2021

Ciara Quigley x17408654 BSHTM4

On March 1<sup>st</sup>, 2021 we had class with Frances Sheridan. During this class, Frances gave us an update on our project showcase profiles. We were given an update about the layout and how to edit our profiles. We were reminded to include up to three images to the profile to be set as cover photos. We also were reminded to include a clear professional profile picture. Later on in the class, we had a talk with revenue about their graduate programme. This was very interesting.

On March 8<sup>th</sup>, 2021 John Bohan came to give us a lecture about data visualisation and techniques we could use to improve our projects. This lecture was extremely beneficial to me and I will review it once again closer to the time of our final submission. During our reading

week, I really focused on my project as well as doing my other assignments for other modules. I did a lot of research into learning new techniques John Bohan had previously mentioned to us. I made some graphs during this time to go towards my final submission.

Week 9 was extremely helpful to all data analysis students. Rejwanul Haque come on and gave us a lecture about the documentation specifically for us data analysis students. We all found this session very useful as we learned how to create a h1 level final report. We were also given the lecture slides which will also come in handy closer to the time of submission.

Our final ever project lecture class was on March 29<sup>th</sup> and it was about the Final Presentation and Demonstration for preparing for the presentation. Was class was given to us by Lisa Murphy and again was extremely useful to us all. She also talked through what we should include in our recordings and the PowerPoint.

## Reflective Journal 7 April 2021

Ciara Quigley x17408654 BSHTM4

The beginning of April was very difficult not only for us as students but also the entire staff in NCI. We were the victims of a ransomware attack which resulted in major IT obstruction. We had no access to Moodle or any college services. The only college system I had access to was Microsoft Teams. Moodle even to this day can be quite temperamental and extremely slow unfortunately. Due to this disturbance, it was very difficult to advance with my project as I relied on the virtual desktop for some aspects of my project. I had to make different arrangements for myself which I am happy to say that since then I have been progressing greatly with my project. As a result of the 8 day disturbance on all college systems, we were all granted a week extension on all submission across the college. For this project it means it is now due on Sunday May 16<sup>th</sup> instead of Sunday May 9<sup>th</sup>.

Mid to late April I solely focused on my TABAs for my other modules. For us, this semester was only 10 weeks unlike the usual 12 as it allowed us time to complete our TABAs and give us the sufficient time to complete our projects after. My final TABA is due on May 4<sup>th</sup> and from then on, my sole focus will be on completing this project and uploading it on time. I will then create my PowerPoint and film my project presentation and submit that for May 23<sup>rd</sup>.



# National College of Ireland

## Project Proposal

<Covid 19>

<7/11/20>

<Technology Management>

<Data Analytics>

<2020/2021>

<Ciara Quigley>

<x17408654>

<x17408654@student.ncirl.ie>

## Contents

1.0	Objectives.....	37
2.0	Background .....	37
3.0	Technical Approach.....	38
4.0	Project Plan .....	40
5.0	Technical Details .....	42
6.0	Evaluation .....	<b>Error! Bookmark not defined.</b>
7.0	Bibliography .....	42

### 9.3.1 Objectives

The objectives for my final year project are:

1. To analyse Covid 19 public data
2. To compare the Irish response to the virus vs other countries such as the United States of America and Italy
3. To identify trends in the data ie: when people stay at home case numbers drop, when children went back to school in September, case numbers rose again.
4. To visualise these trends and comparisons using visualisation tools such as Power BI
5. Identify problems some countries may be facing and suggest solutions, based off other countries who are succeeding with maintaining low Covid cases ie: New Zealand.

### 9.3.2 Background

Covid 19 or the Novel Coronavirus has taken over the world over the past nine to twelve months. Covid 19 is a new virus which originates from Wuhan, China. It is an illness whereby the virus attacks your lungs and airways. Symptoms of the virus range from mild to severe and in 2.5% of cases they result in death. (covid19.who.int, accessed November 2020) Often, people who catch the virus do not realise they have it. This is known as being asymptomatic. As of November 7<sup>th</sup>, 2020, according to the WHO (World Health Organisation) there has been 49,106,931 confirmed cases worldwide and 1,239,157 recorded deaths. (covid19.who.int, accessed November 2020)

On March 11<sup>th</sup>, 2020, the Novel Coronavirus was declared a global pandemic. (covid19.who.int, accessed November 2020) This virus has halted the global economy resulting in millions of job losses and business closers. Since March 2020, the entire world has had restrictions placed upon them in some capacity. Depending on the country, every fortnight or so, new restrictions are put in place or the current ones and extended. For example, in Ireland, we are currently at the level 5 stage in living with Covid. This is the highest possible stage which means the country is in lockdown for the second time this year. This level 5 mean that we can no longer travel outside of the 5km radius of your house unless it is absolutely necessary. Everybody has been asked to reduce their movements and work from home where possible. People who are over the age of 70 are being asked to stay at home and ask others for assistance with regards to getting food and medicines. This is known as cocooning.

### 9.3.3 Technical Approach

For my analysis I will be using the KDD approach. KDD stands for Knowledge Discovery in Databases.

Figure 1 illustrates the steps involved in the process of KDD:

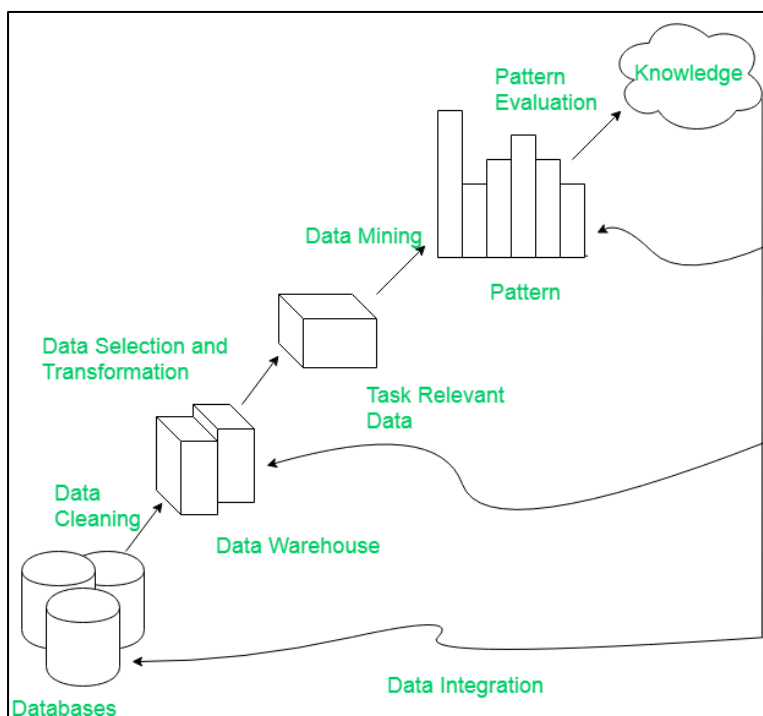


Figure 1; (geeksforgeeks.org, accessed November 2020)

Step 1: Data cleaning:

This is the process of cleaning up the data by removing any error or variances within the data set

**Step 2: Data Integration:**

This is the process whereby you combine common data from multiple data sources. I.e: using the ETL process (Extract Load Transformation)

**Step 3: Data Selection:**

This is the stage where you identify what data is relevant to your analysis and remove the remaining data from your data collection.

**Step 4: Data Transformation:**

This is where you transform your data into the appropriate data structure you are working with. There are two steps within this step. Data Mapping and Code Generation.

**Step 5: Data Mining:**

This is a technique used to pull/ extract data which has similar patterns. It characterises the data which has been extracted.

**Step 6: Pattern Evaluation:**

Pattern evaluation is identifying patterns which are noticeably increasing based on given reasonings/ measures. This is where you will source interesting patterns to create visually pleasing visualisations.

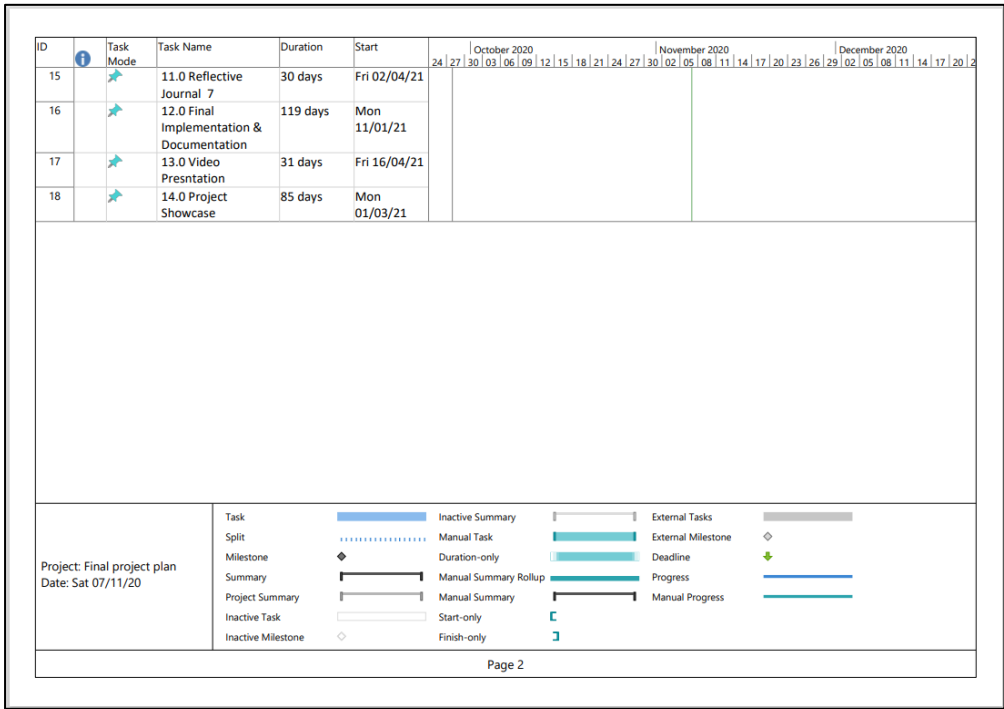
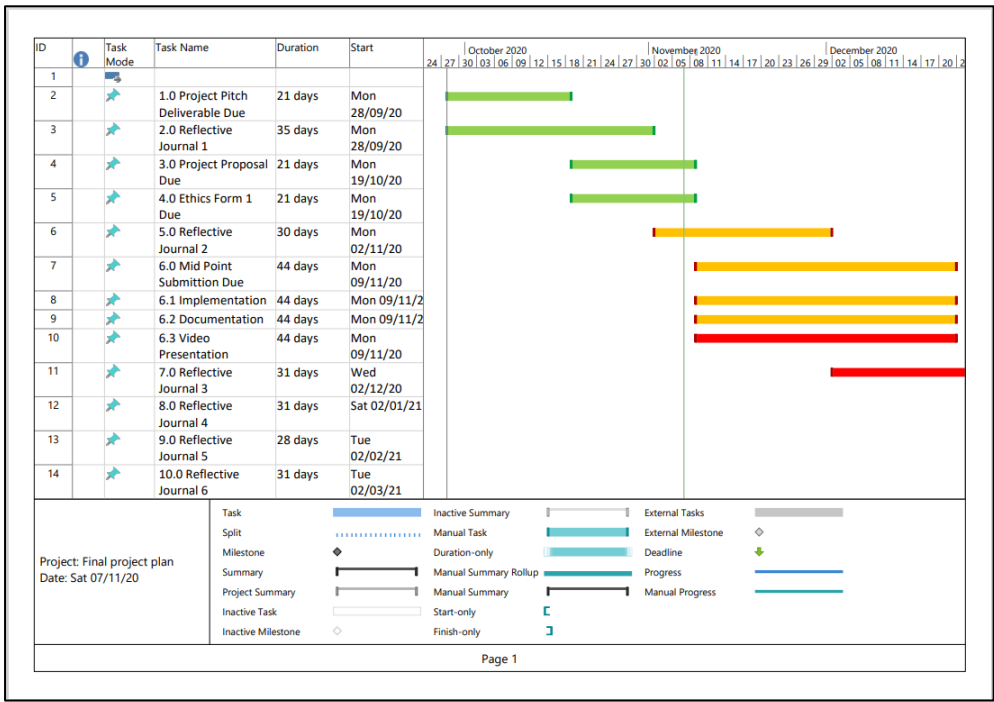
**Step 7: Knowledge Representation:**

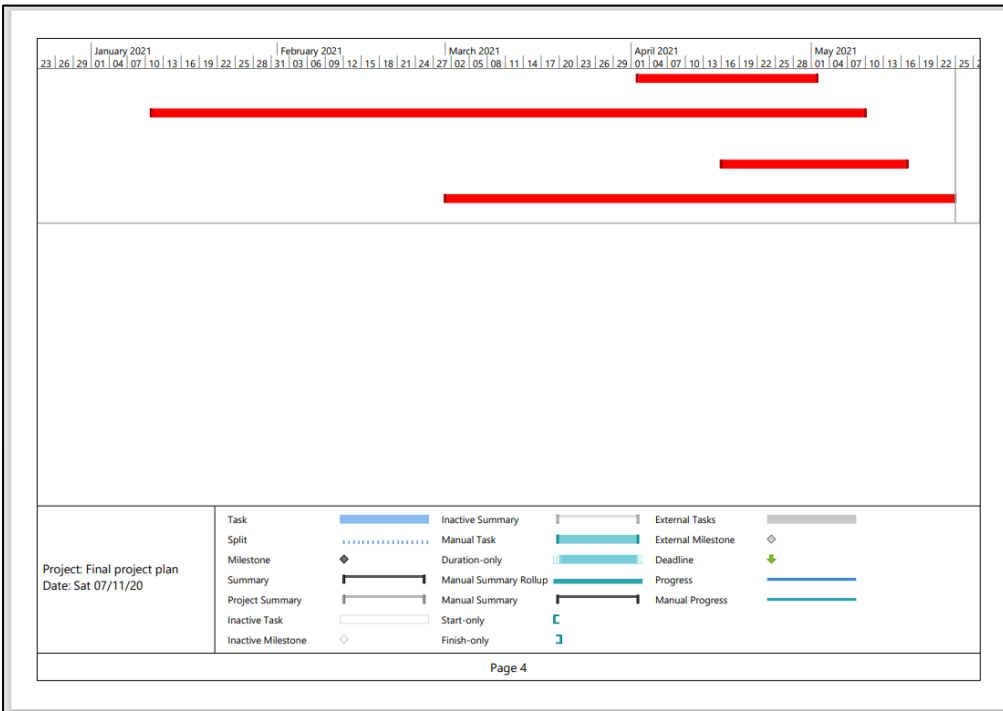
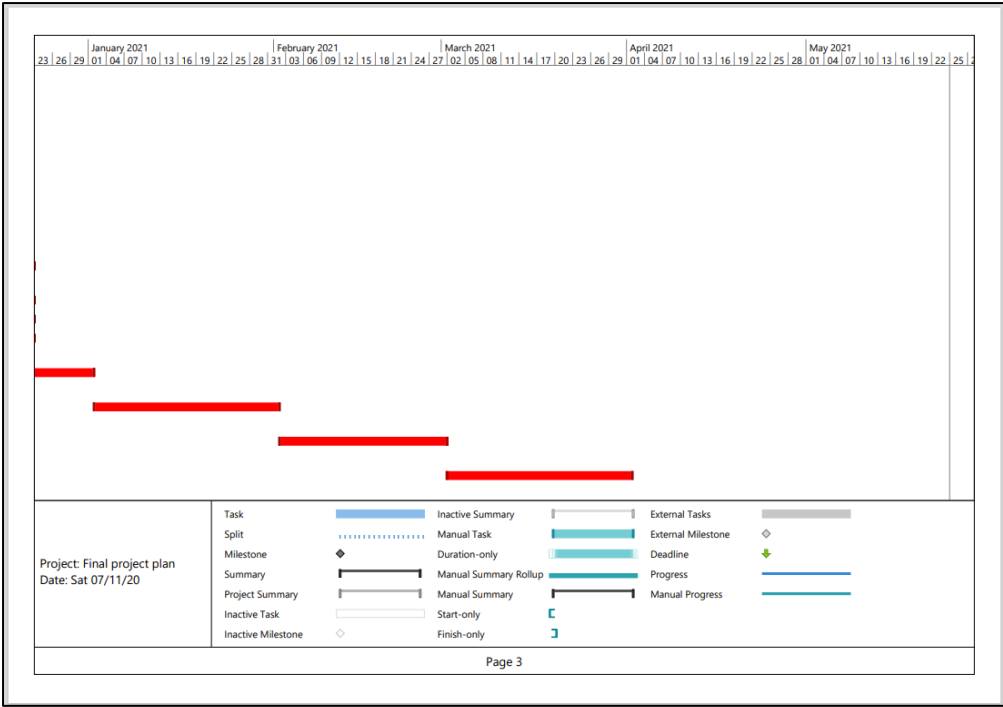
Knowledge representation uses visualization tools and represents the results you gained from data mining. This is where you generate graphs, reports, tables etc.

Other machine learning methods which I will decide upon in the second semester are: SVM Support Vector Machine, Naïve Bayes method for machine learning, random decision tree, linear regression multiple regression and deep learning method.






### 9.3.4 Project Plan





**Key:**

-  Completed Task
-  Pending Task
-  Not Yet Started Task

### 9.3.5 Technical Details

I technologies I will be using to achieve my goals set out are, Microsoft Excel, SPSS, R Studio, and Jupyter Notebook. During my work placement in third year, I became familiar with Power Bi so if I get the chance to, I would also like to revisit this technology.

### 9.3.6 Bibliography

Covid.who.int. (2020). Coronavirus. [online] Available at: <https://covid19.who.int/> [Accessed November 2020]

Geeksforgeeks.org. (2019). KDD Process In Data Mining. [online] Available at: <https://www.geeksforgeeks.org/kdd-process-in-data-mining/> [Accessed November 2020]