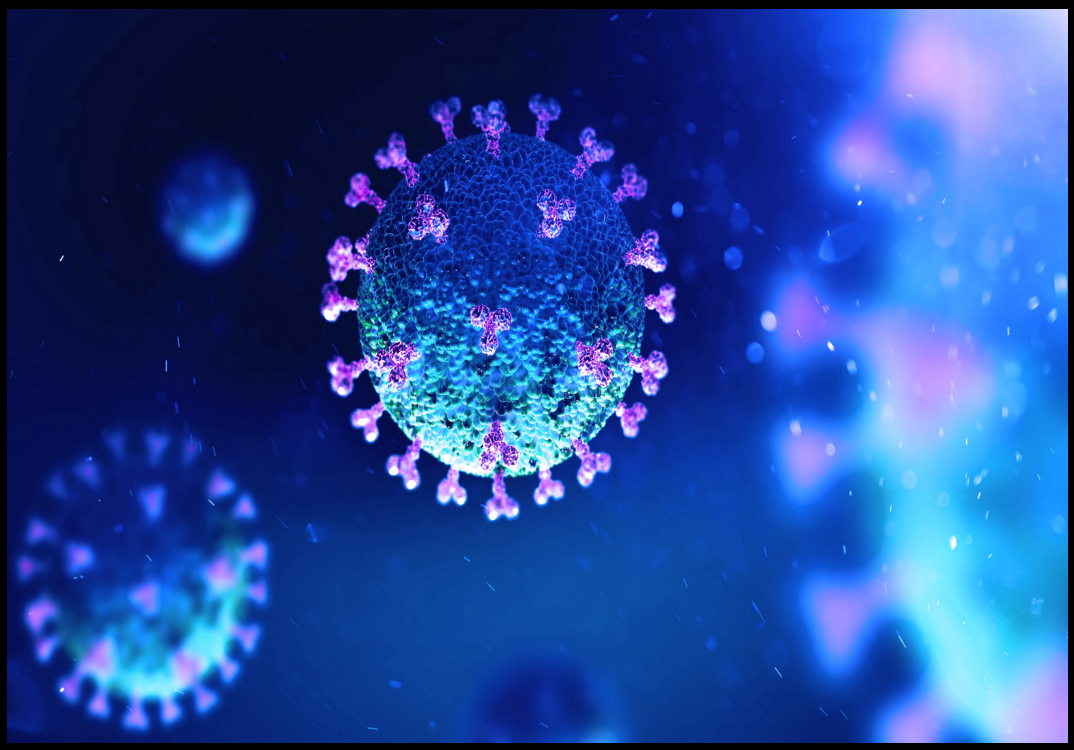

GLOBLAL COVID-19 ANALYSIS

Presented By:
Mesaye Bahiru

Table of content

Introduction

Covid-19 also known as (SARS-CoV-2 or Coronavirus) is viral respiratory disease first discovered in Wuhan, China in December 2019. Since its discovery, it has spread across the globe, leading to the ongoing pandemic. The virus causes illness ranging from mild flu-like symptoms to severe pneumonia. and so far, there is over 273 million confirmed cases worldwide. The virus can be transmitted via droplets from infected person. The common prevention methods include wearing a mask and maintaining social distance in social gathering.

The goal of this project is to build a classification and unsupervised learning model that examines the recovery rate of covid-19 and how it is impact by factors like confirmed cases, death, and country/region. Furthermore, linear discriminate analysis (LDA) was used for classification model, regression tree was utilized for tree-based analysis and finally hierarchical clustering using dendrogram supported in understanding clusters formations.

The project analysis is designed to be presented for healthcare directors working in World Health Organization (WHO) and policy makers who are interested in acquiring knowledge on covid-19 recovery rate and working to avert next pandemic.

Data Description and Visualization

Data Description

The dataset in this project was collected from Kaggle, which is popular platform that provides a reliable data repository for data scientist. There were 8 variables in total. The table below contains description of the data.

Table-1: Data Dictionary

Column Name	Description
SNo	Unique identifier
ObservationDate	The date the observation recorded
Province/State	The Provinces or States name of the country
Country/Region	The Country name
Last Update	The data and time observation were updated
Confirmed	The number of confirmed cases
Deaths	The number of deaths recorded
Recovered	The number of recoveries recorded

Data visualization

Critical visuals that highlighted the global covid-19 impacts were presented as part of data visualization. Figure 1 plot showed recovered cases per country, it was evident India had the highest recovery cases followed by Brazil and Russia.

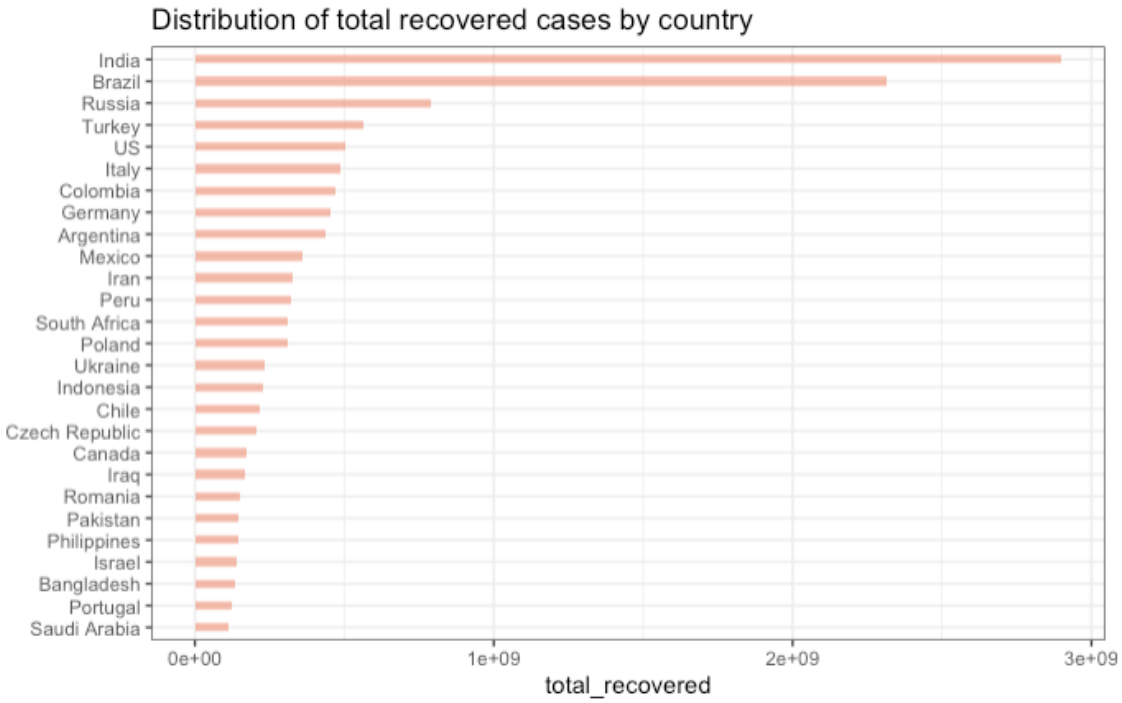

Figure 1: Recovered cases by country

Total death across different countries was checked to examine which countries experiencing highest mortality rate. The United States had the highest mortality as shown in figure 2 followed by Brazil and India.

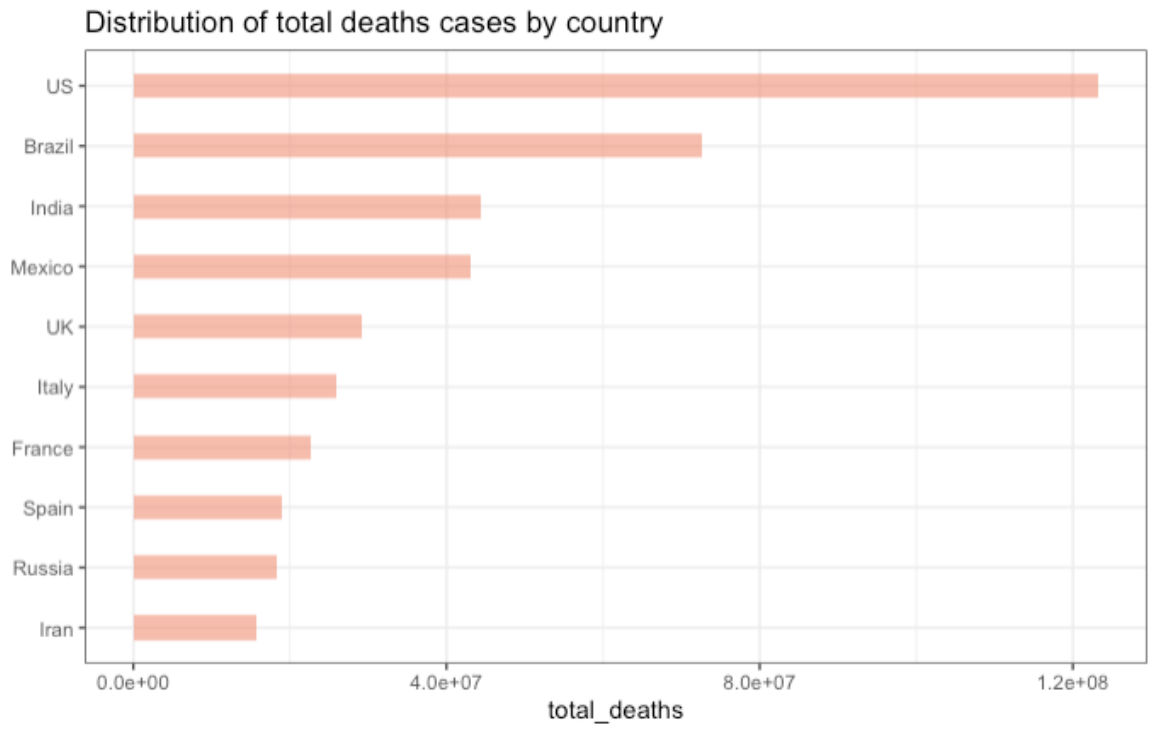

Figure 2: Mortality by country

Furthermore, USA States confirmed cases was studied and the results showed California, Texas and Florida had the highest confirmed cases.

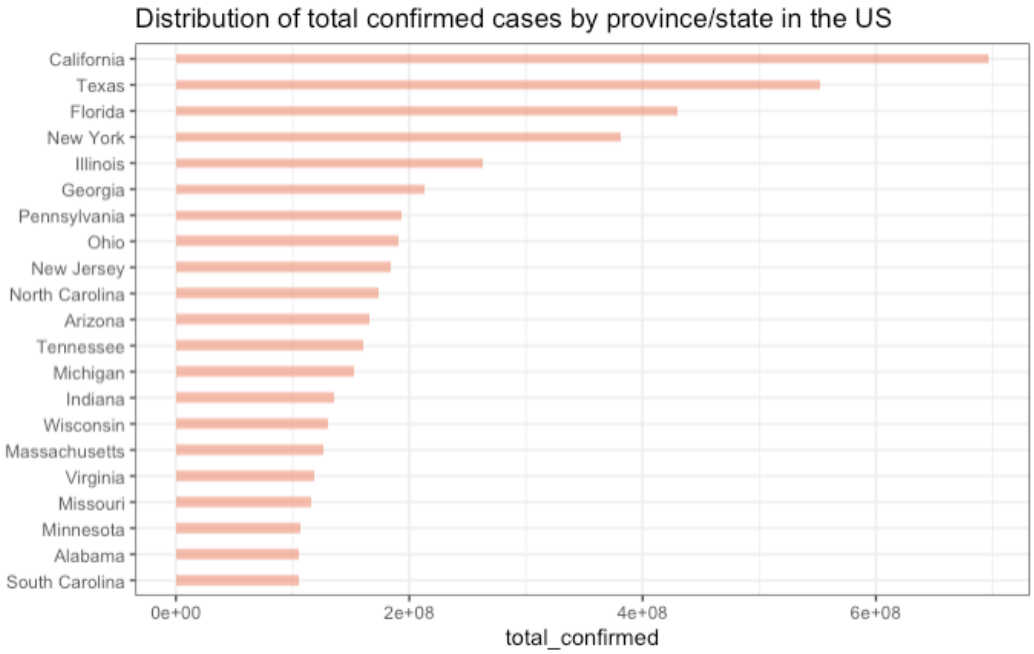

Figure 3: Total confirmed cases by US States

Additionally, Canada provinces confirmed cases were studied to compare how well Canada managed compared to USA. Ontario and Quebec were two provinces with the most cases in Canada.

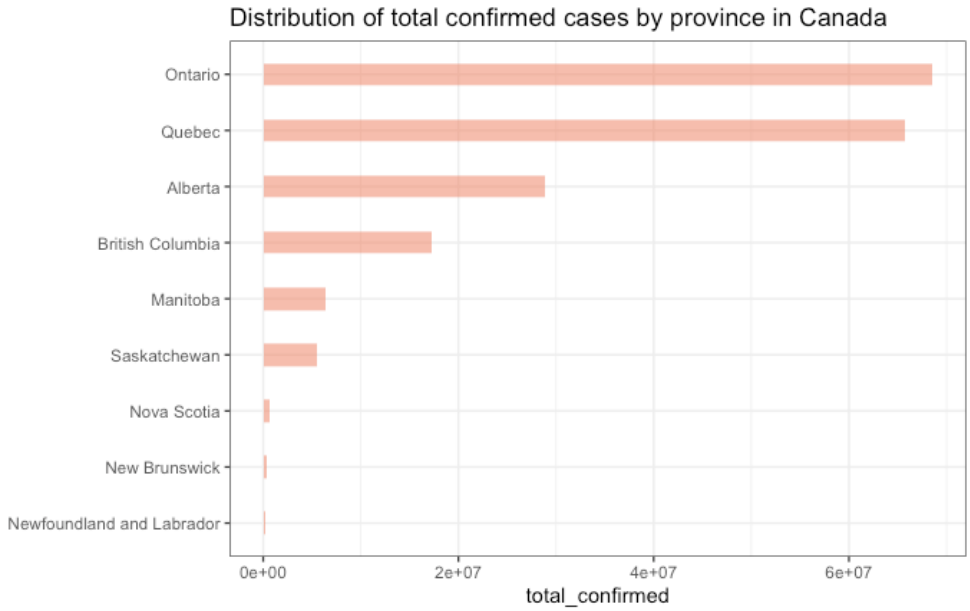

Figure 4: Canadian Provinces Confirmed Cases

Finally, using world map global confirmed cases as of May 29th ,2020 was presented in Figure 5. Brazil and India had the highest confirmed cases at the time. For more supporting visuals refer to appendix and the R-code.

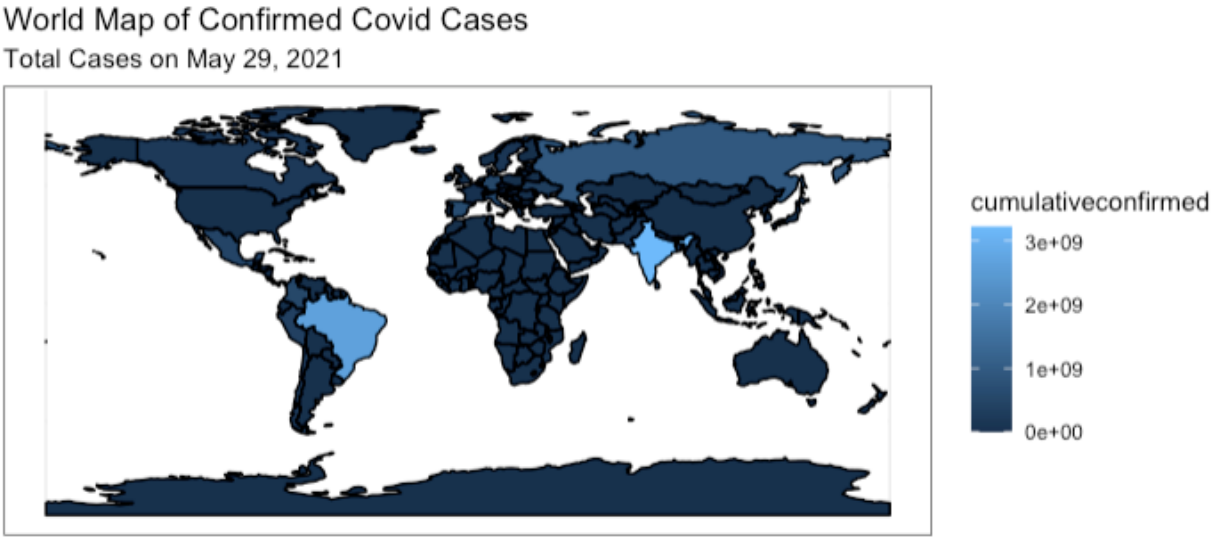

Figure 5: World Map with Cumulative Confirmed Cases

Nowadays, it is common to hear health officials in press conferences talking about flatten the curve and using logscale the world confirmed cases was plotted to see if health guidelines had impact on leveling the curve.

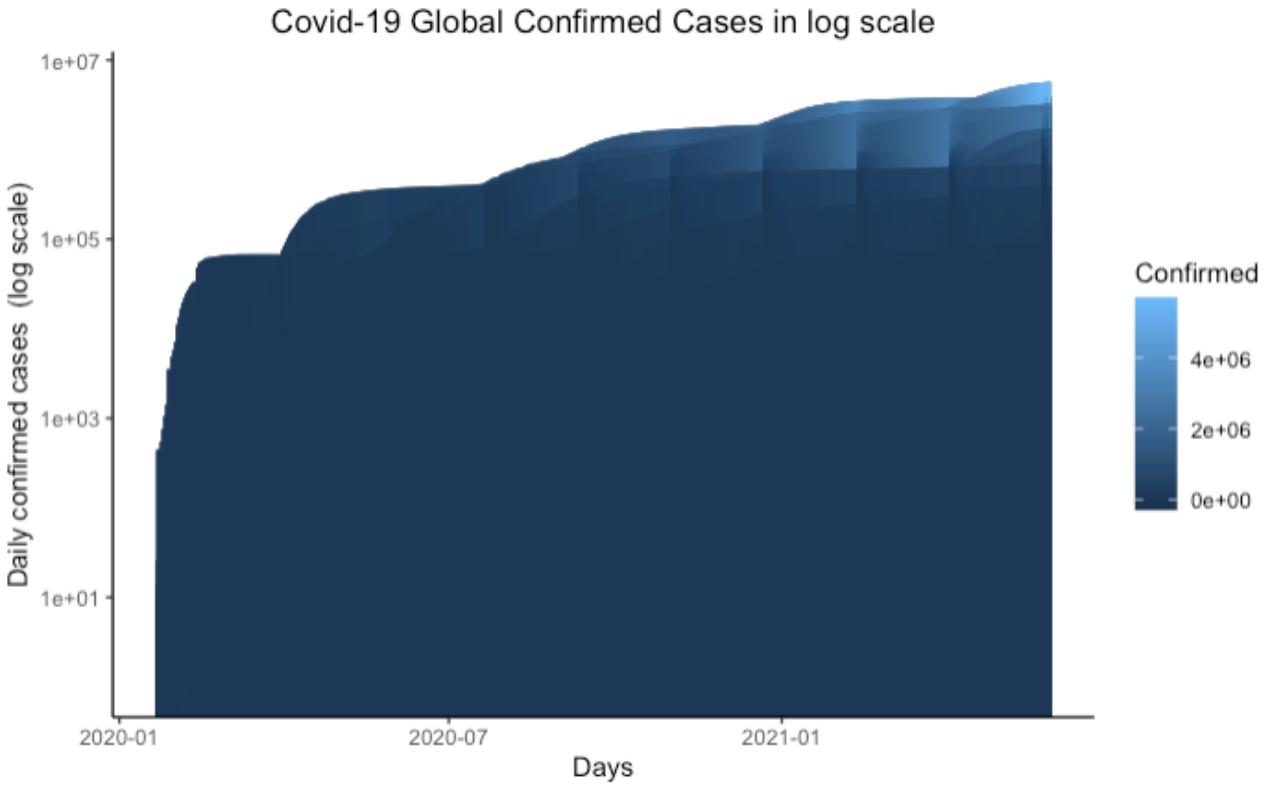

Figure 6: World confirmed cases in Logscale

Model Selection and Methodology

For the covid-19 analysis, three statistical models were built. These models were linear discriminate analysis, regression tree, and hierarchical clustering.

Data Preprocessing

Data preprocessing needs to be performed before placing the raw data in the machine learning model. In the process of data preparation, all missing values and the negative values were eliminated from the dataset. In addition, a new variable (Recovery Rate) was created. This was created to show two classes (**0= low recovery** rate and **1=high recovery rate**). To create this variable, all recovery values above the mean were considered as high while all values below the mean were considered as low. Once, data preprocessing was completed, then the next step is to start modelling process.

Model 1: Linear Discriminate Analysis

Linear Discriminate Analysis (LDA) is classification algorithms utilized when dealing with two or more categorical target variables. Linear discriminate analysis provides great insight over logistics regression for this covid-19 analysis for two main reasons. First, linear discriminate analysis is stable when working with response variables well separated. Second, there is not many classifications analysis done on covid-19 using linear discriminate analysis and this provides a challenge and great learning opportunity. Linear discriminate analysis relies on Bayesian theorem as shown the formula (1) below to perform classification task.

The goal of this model is to predict Pr (Y= Y^k |X=x), which denotes the probability the recovery rate (Y=Y^k) given the recovery, death, confirmed variables (X= x). Pr (X=x | Y= Y= Y^k) is known as probability density function of X for an observation that comes from the Kth class. $\prod_k$ represents the prior probability of Kth class.

$$Pr(Y = Y^K | X = x) = \frac{Pr(X = x | Y = Y^k).\Pi_k}{P(X = x)} \quad (1)$$

Estimating Pr (X=x | Y= Y^k) can be computationally difficult, this can be resolved by simplifying into $f_k(x)$, which assumes the distribution of all K categories are normally distributed with mean value of (μ_k) and standard deviation (σ_k).

The $f_k(x)$ can be write as

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right) \quad (2)$$

The posterior probability is given as

$$P(Y = k|X = x) = \frac{\pi_k f_k(X)}{\sum_{l=1}^{K} \pi_l f_l(X)} \quad (3)$$

$$Pr(Y = Y^k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}. \quad (4)$$

MASS package in RStudio was used to conduct linear discriminate analysis. The dataset was split into 80 % train and 20 % test set.

Then, build LDA model using recovery rate as the target variable and recovery, confirmed, death as predictors. The prior probabilities of response class for an observation were checked. 79.9% of all recoveries fall under the low recovery rate and 20.1% of all recoveries fall under high recovery rate. The mean value (k) for response classes for a given X=x is checked. When different features fit into a particular response class, this reflected their mean values. For their recovery rate class, it was noticed a clear difference between the proportion of recovered cases vs deaths (1508.30 vs 669.82). The greater the gap between the mean, the easier it is to classify observations. Linear discriminant coefficients were the coefficients of the

linear equation that were used to classify the response classes is defined here. Because there are only two response classes in this model, there will only be one set of coefficients (LD1). Afterwards, partition plot was generated to visualize how well the model classified the response class based on recovered, death and confirmed variables.

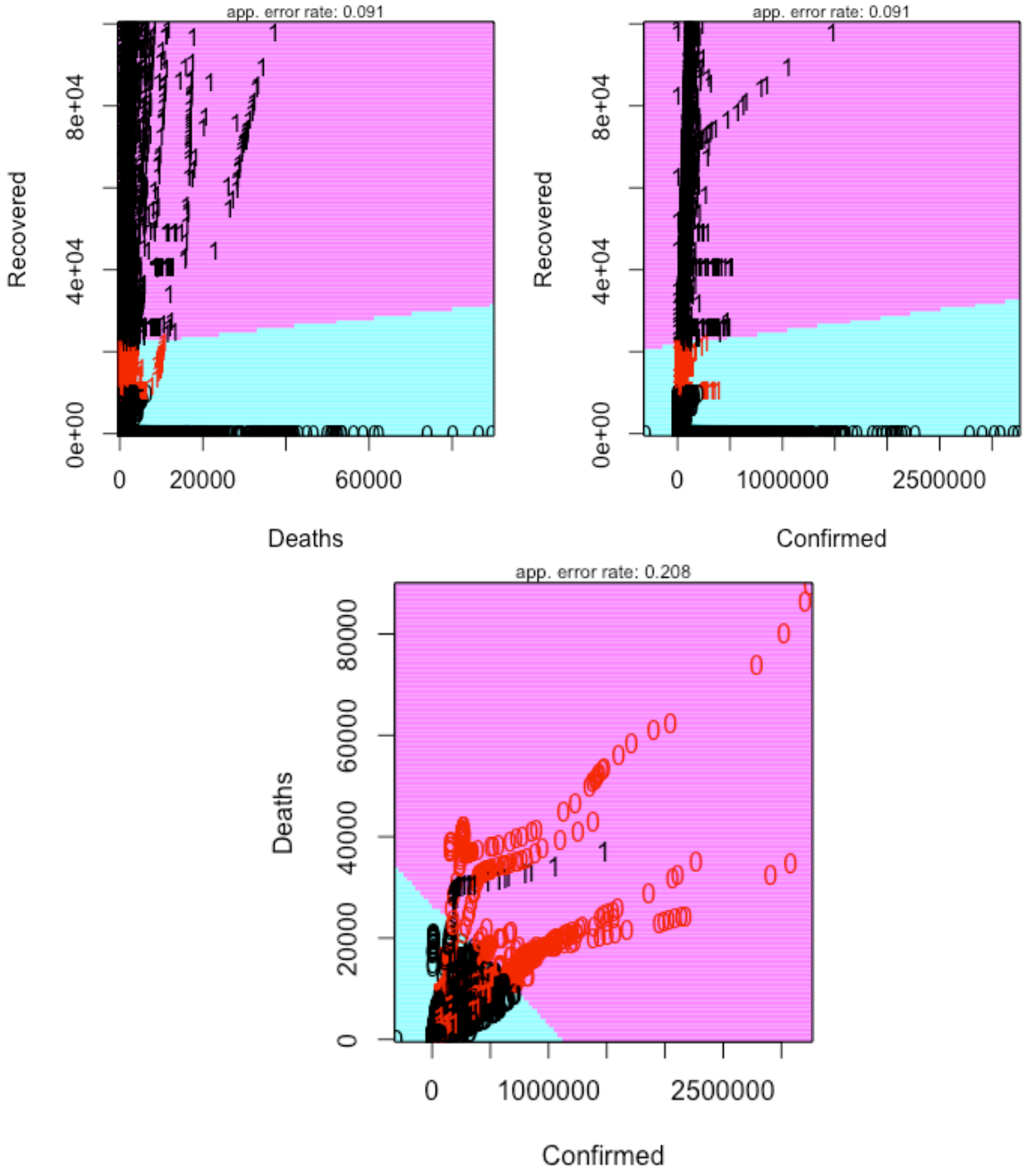

Figure 7: Partition Plots

The model was fit on the training dataset and prediction was obtained on the testing dataset. The accuracy of the LDA model was 74%. The table 2 confusion matrix summarizes the model prediction to test data observations

Table-2 Confusion matrix

	Recovery Rate	
Predicted	0	1
0	31942	3666
1	0	4391

LDA made incorrect prediction for 3666 observations as high recovery rate when it should be classified as low recovery rate, and there was no incorrect prediction for high recovery rate. In addition, ROC (Receiver Operating Characteristics) curve was used to show a graphical representation how well the model separates the classes and ROC curve conveys the overall performance of classifier, summarized over all possible thresholds was given by the area under the curve (AUC). Linear discriminate analysis AUC was shown to be 100%, and this indicated the model was perfectly capable of classifying response classes accordingly.

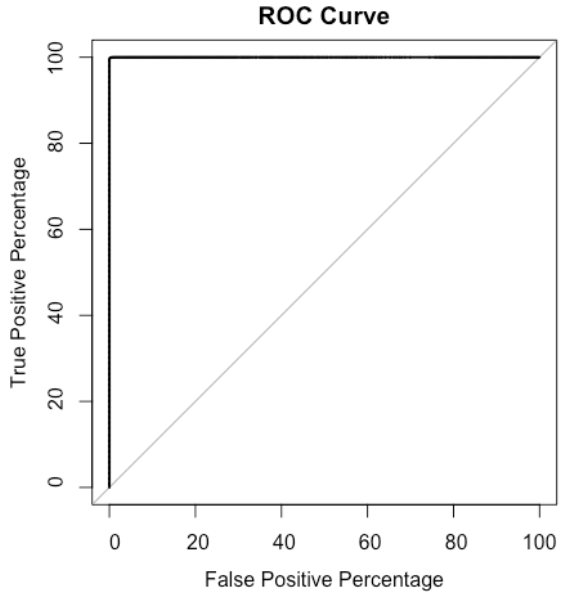

Figure 8: ROC Curve

Model 2: Regression Tree Model

The principle behind growing a tree involves dividing the predictor space into J regions using certain criteria and for every observation in region R_j, we find the mean response. Regression tree applies recursive binary splitting to minimise residual sum of squares. Moreover, complex parameter cp a number is between 0 to 1 is selected, and the goal is to reduce RSS by more than 100*cp%. Tree should stop trying to split a branch when splitting improves RSS less than 1 %. *rpart* and *rpart.plot* are two main libraries used for the regression tree analysis.

The predictors in this case where *Death*, *Country.Region*, *Recovery_rate* and the response variable was *Recovered*. Build regression tree model and cp was selected to be 0.08. Then, the regression tree was visualized using rpart.plot.

The regression tree showed 97% percent of the observation was placed under recovery rate (yes) node and country/region is next important variable. Regression tree illustrated individuals with high recovery rate is likely make of the most of recovered individuals. Next, individuals with confirmed cases in European countries like Belgium had a better chance of recovery.

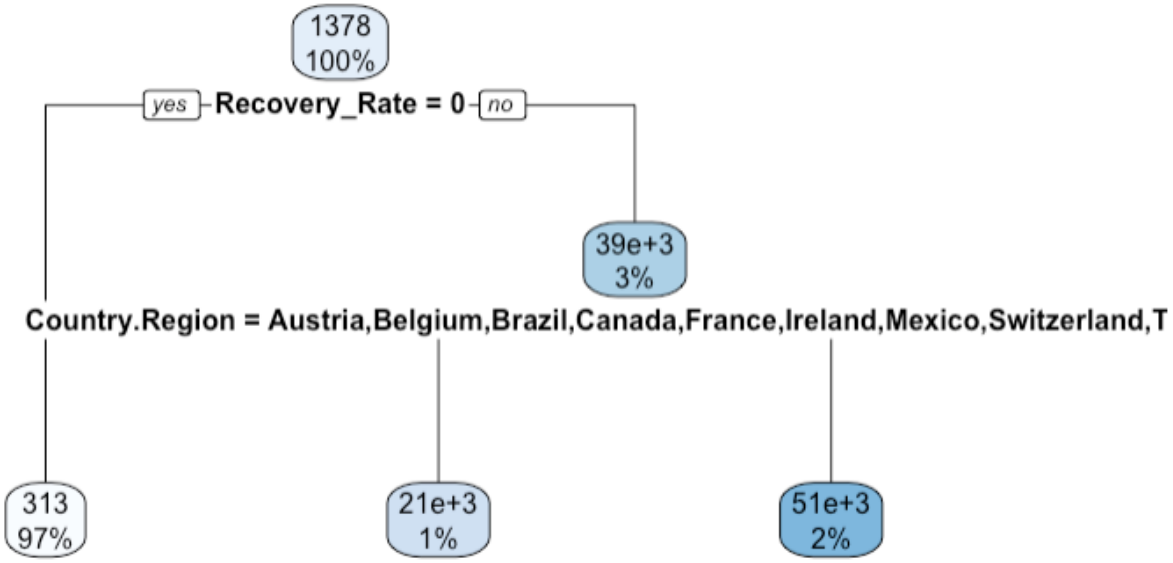

Figure 9: Regression Tree

Pruning regression trees was required to determine the optimal cp value. To find the optimal cp value, another regression model was built with smaller cp value of 0. 008.The goal was finding the cp value that minimizes error and it had better out of sample performance. Then, based on the best cp value, new regression tree was built the regression tree.

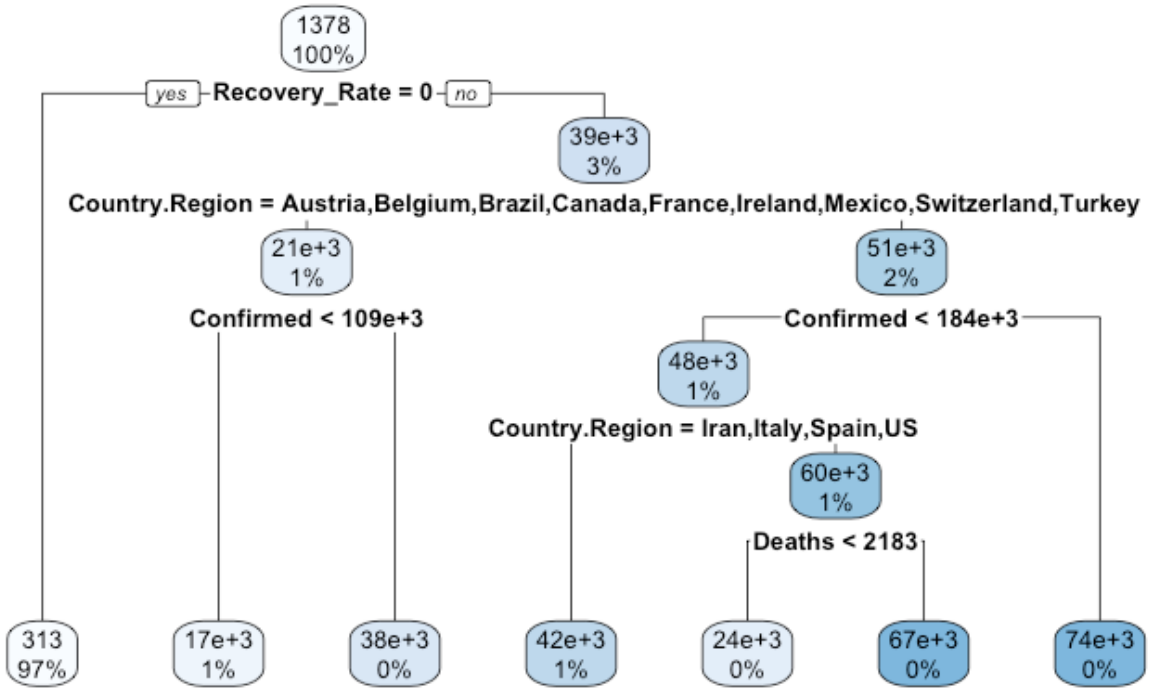

Figure 10: Best regression tree

The best regression tree showed 97% percent of the observation was placed under recovery rate (yes) node and country/region is next important variables followed by confirmed cases. The regression tree illustrates individuals with high recovery rate is likely make of the most of recovered individuals. Next, individuals with confirmed cases in Austria, Belgium, Brazil, and Canada had better chance of recovery and in addition individuals in Iran, Italy also had better chance compared with Spain and US.

Model 3: Hierarchical clustering

Hierarchical clustering is an unsupervised learning technique which groups are produced in a hierarchical order (or a pre-determined ordering). Objects are grouped into a hierarchy comparable to a tree-shaped structure in hierarchical clustering, which is used to analyze hierarchical clustering models.

The hierarchical clustering process is as follows:

i) It starts by calculating the distance between every pair of observation points and store it in a distance matrix.

ii) It then puts every point in its own cluster.

iii) Then it starts merging the closest pairs of points based on the distances from the distance matrix and as a result the number of clusters goes down by 1.

iv) Then it recomputes the distance between the new cluster and the old ones and stores them in a new distance matrix.

v) Lastly it repeats steps 2 and 3 until all the clusters are merged into one single cluster.

To plot the dendrogram, smaller dataset that contains 40 observation was selected and null values were removed. The euclidean distance between each observation was calculated.

In the model, the cluster method used was average linkage. The average linkage computed mean intercluster dissimilarity. Distance was euclidean and no. of objects are 40. Distances were transformed into heights in a dendrogram, which is a hierarchy of clusters. It divides n units or objects into smaller groups, each with a p property. A horizontal line connects units in the same cluster. Individual units are represented by the leaves at the bottom. It showed clusters as a visual representation.

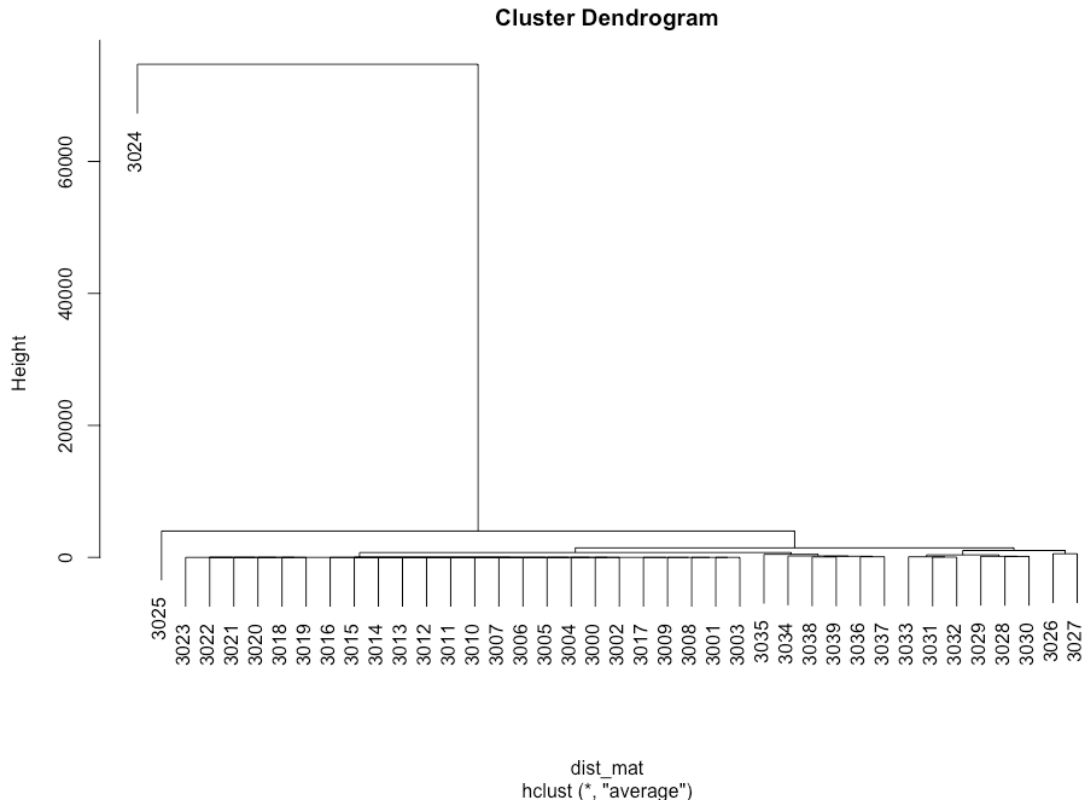

Figure 11: Cluster Dendrogram

From the above dendrogram, there were two main clusters formed, the first cluster (3024 -> country. region= China) contained high number of values compared to rest of clusters formed.

Let's examine the dendrogram, by cutting the tree into two clusters and the leaves were color coded. The result was shown below

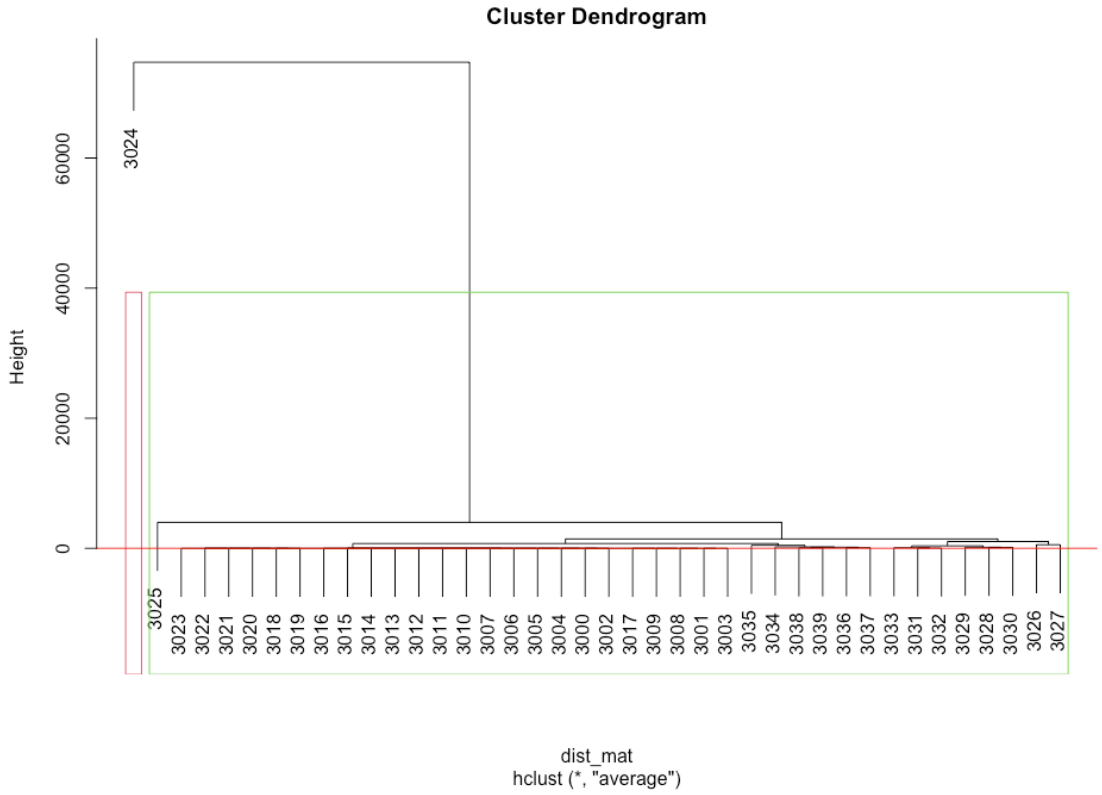

Figure 12: Dendrogram with two clusters

In unsupervised learning, the goal is not to make a prediction, likewise, there was no best approach to measure the dendrogram performance. For this analysis, confusion matrix was generated to see how well the model clustered recovery rate compared with actual recovery rate. The results in the table 3 showed hierarchical clustering with 100% accuracy placed the clusters based on recovery rate.

Table 3: Confusion Matrix for Hierarchical clustering

	Predicted	
Actual	**1**	**2**
0	39	0
1	0	1

Healthcare Managerial implications & Conclusion

The covid-19 analysis shines the light on the recovery rate from covid-19 virus. The Analysis provides insight for healthcare manager or policy makers working World Health Organization (WHO) who want to understand covid-19 recovery rate and the factors associated with recovery process. The insight provided in this analysis can be developed into recommendation. Moreover, using this analysis WHO Officials can allocate resources to countries struggling low recovery rates.

The data visualization presented will help policy makers better understand how the virus impacted the world and provide urgencies on how to avert another pandemic in the future.

The three statistical models consist of linear discriminate analysis, regression tree and hierarchical clustering was built to understand and visualize the intricate relationship between recovery rate to confirmed cases, country/region, recovered cases and death. The linear discriminate analysis models were capable of classifying dataset into low and high recovery rate. with 74% accuracy. The regression tree highlighted recovery rate and country/region was highly indicative of recovery state. Unsupervised learning with hierarchical clustering of 40 observations showed the formation of dendrogram in the analysis.

Appendix

More Exploratory Data analysis

Picked 4 countries from different continents to examine the daily death from 2020-2021.

Daily Death in USA

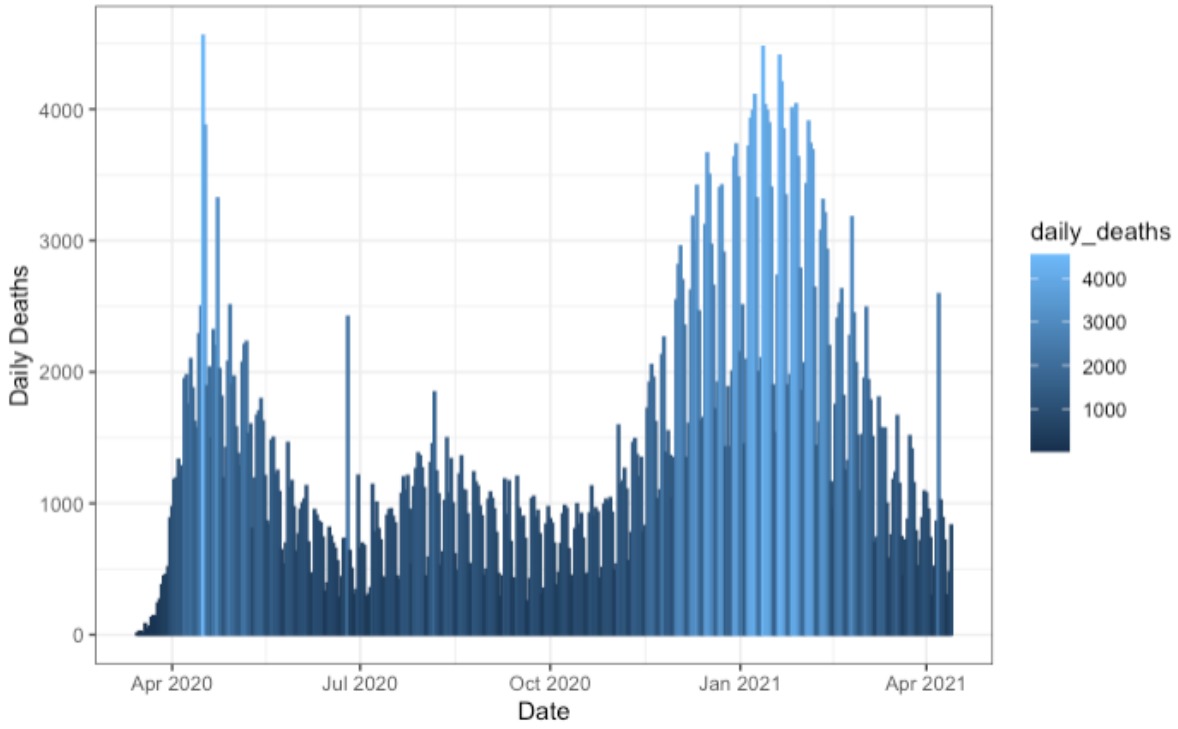

Figure 13: daily death from 2020-2021 in USA

Daily Death in Brazil

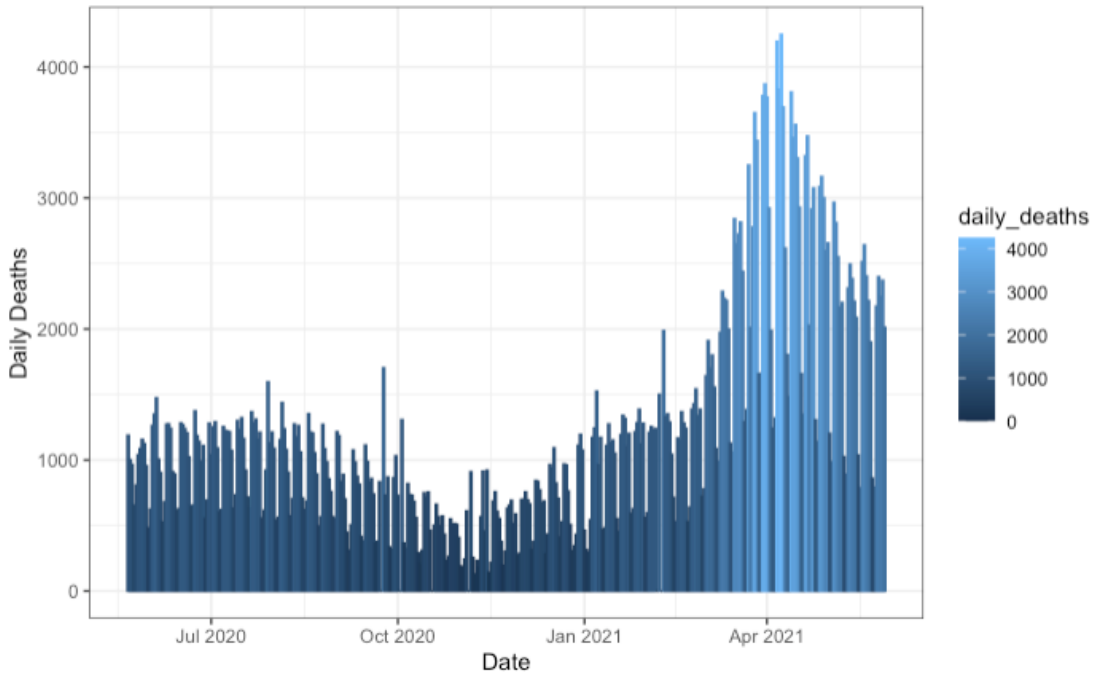

Figure 14: daily death from 2020-2021 in Brazil

Daily Death in China

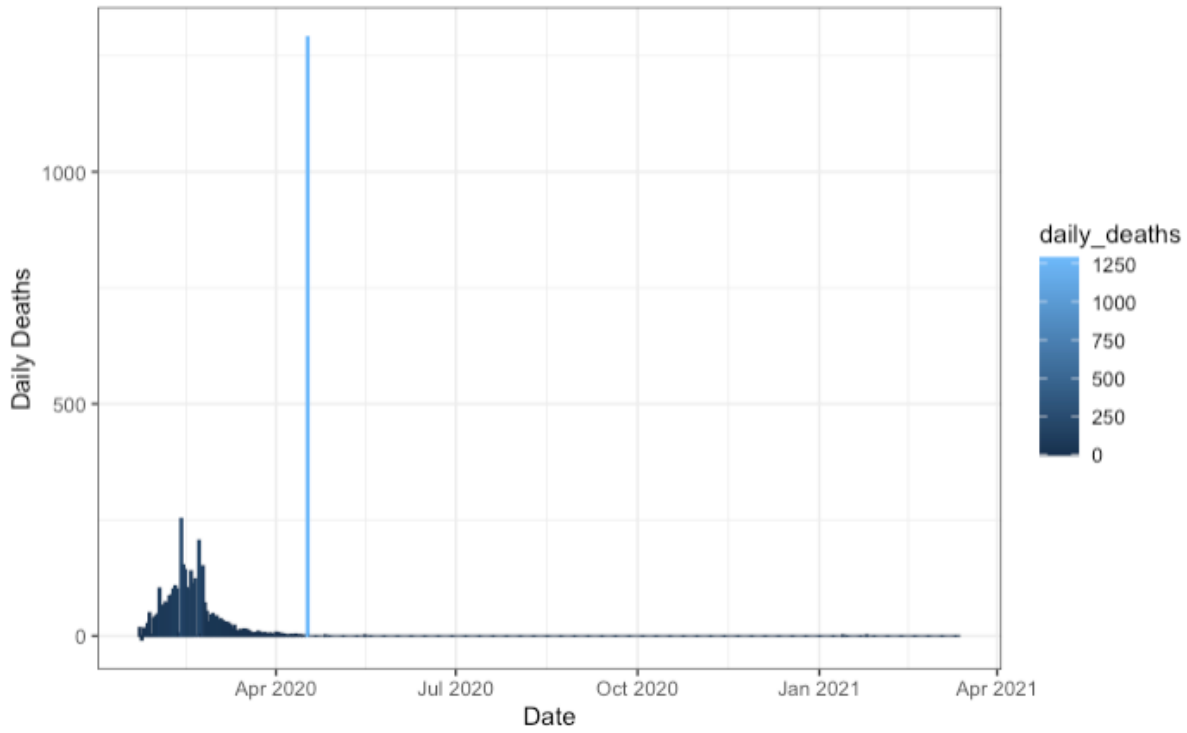

Figure 15: daily death from 2020-2021 in China

Daily Death in Australia

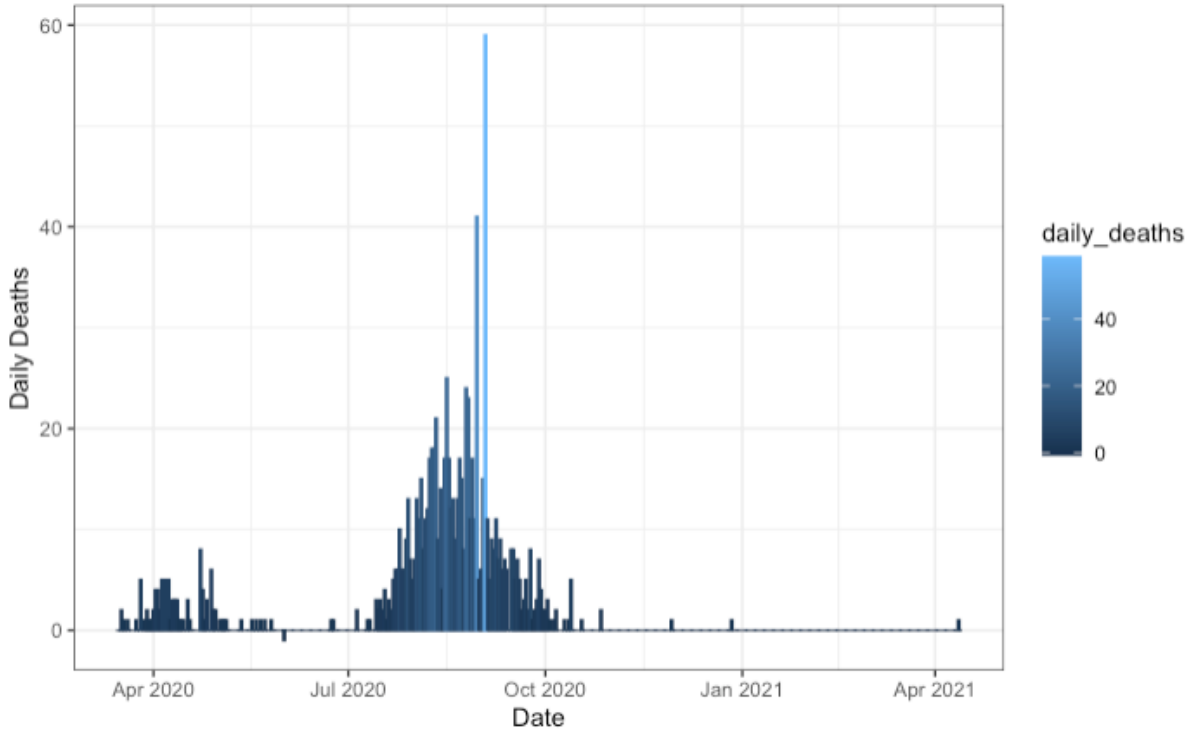

Figure 16: daily death from 2020-2021 in Australia