August 10, 2022

Lefhoko Thapelo

Springboard apprentice

SPRINGBOARD

A study of covid 19 Death statistics

Understanding the Predictors of Covid 19 Deaths Using Data

Contents

[INTRODUCTION 1](#_Toc110776744)

[PROBLEM STATEMENT 1](#_Toc110776745)

[DATA WRANGLING 2](#_Toc110776746)

[Exploratory data analysis 3](#_Toc110776747)

# INTRODUCTION

The Covid 19 pandemic is one of the most significant things to have happened this century. In one way or another, it has affected every person on this planet. A lot of people died and others fell very ill but recovered. There are some who tested positive, but never had any symptoms and others who only had mild symptoms. Hospitals were getting filled with lots of patients and resources were limited.

According to the World Health Organization, Coronavirus disease 2019 (Covid 19) is an infectious disease caused by the SARS-CoV-2 virus. They go on to say, “Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention.” It goes on to mention that older people and those with underlying medical conditions are more likely to develop serious illness.

The primary concern for governments and health institutions around the world was trying to reduce the Covid 19 death rate. In order to do this, a whole lot of measures were taken. People were forced to wear masks in public, public gatherings were restricted by lockdowns and curfews, social distancing was enforced, and movement across borders was restricted.

The aim of this study was to find the major factors that lead to covid 19 deaths using a dataset collected from our word in data. Finding this out would help to understand what actions were most effective in preventing Covid 19 deaths, and may inform key decision-makers on what the best courses of action may be in the even of another similar pandemic. The study will also help to understand what could have been done better in order to reduce Covid 19 deaths by modelling the death rate statistics and predicting death rate based on key factors or features. The dataset that was used can be found at the following github link: <https://github.com/owid/covid-19-data>

# PROBLEM STATEMENT

This study seeks to identify the most significant associations and causes of covid 19 deaths based on group level data sourced from ourworldindata.com. From the study, one should be able to understand what actions increase or decrease the death rate of a country and to what degree they are expected to affect the death rate. Based on this, one should be able to make a prediction on the death rate based on various key attributes using a trained machine learning algorithm.

The importance of this problem cannot be overstated because Covid 19 has had a very big impact on the lives of people around the world. In an attempt to try reduce death rates, governments around the world have implemented various policies that have had a huge impact on the economies of their countries. For example, we need to understand whether stringent restrictions such as curfews and lockdowns really did have an impact on death rate because they had such an adverse impact on a lot of businesses. A lot of these businesses had to shut down because their access to the market was blocked. Understanding what could have been done better may inform leaders around the world on how to deal with the next pandemic that comes along.

The scope of this project is limited to only the United States and using only group level data since individual level data was not available. Individual data such as the actual age of the deceased or whether that person had underlying health problems is not available in the data. What is available in the data is summary data such as number of deaths each day, number of hospital admissions each day, number of vaccinations each day, number of booster vaccinations each day, and stringency index which is just a scale of the tightness of restrictions imposed by the government.

# DATA WRANGLING

As already mentioned, the data for this study was sourced from ourworldindata.com in a GitHub repository. It was quite a large dataset with covid data for many countries around the world for the period of February 2020 to April 2022 when this project started. There were 179218 records with 67 features. The project was done using python notebooks and can be found at the following link: <https://github.com/Thaps/DataScienceGuidedCapstone> . The main python package that was used for the project was Pandas. This is a great tool for manipulating data using an object known as a DataFrame.

The first step of the data cleaning process was to limit the data to only the United States data. Limiting the scope of the geographical location in the data will help control a lot of variability in the study like whether some countries may not have values for certain fields, or possibly different interpretation of fields in different countries due to cultural differences. After limiting the records to only United States data, there were 815 records left. This is still a fairly large number of records. Along with limiting the data to only United States, there were also location specific fields that were removed from the dataset such as iso code and continent. The number of features left was 61.

The next step was then to address missing values in the data. From this United States data, there was a total of 9388 missing values. Out of these, there were columns that had absolutely no values at all. It was a simple decision to delete all of these since they really had no value. Other features were also removed based on having no variability. This had to be because those variables also varied with countries.

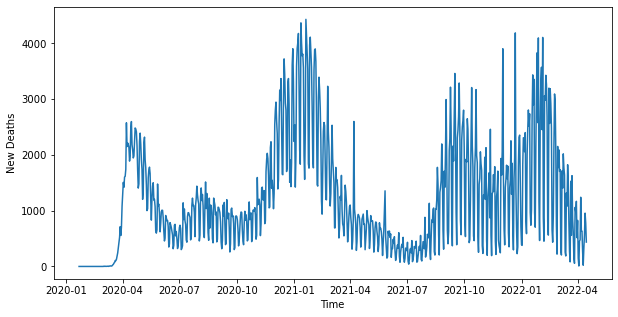
The last step in the data cleaning process was to check for duplicate features. There were no duplicate features found in the data.

# Exploratory data analysis

After taking a good look at the dataset, it was clear that it could be well modelled using a time series. To explore the data as a time series, the date column was set as the index of the d ataset. Find below the trend in new Covid cases and the trends for new deaths as well.

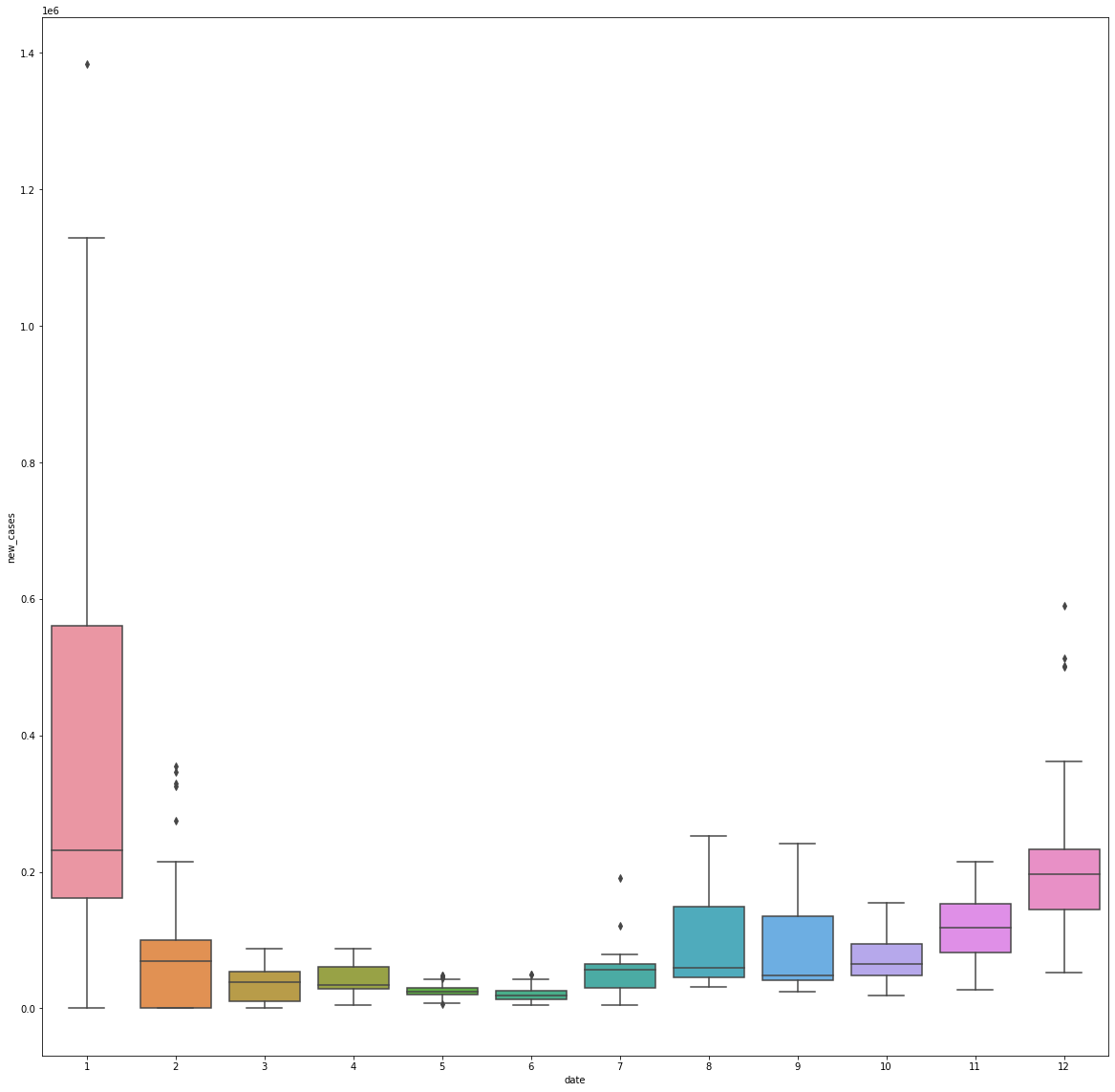
Chart, histogram

Description automatically generated



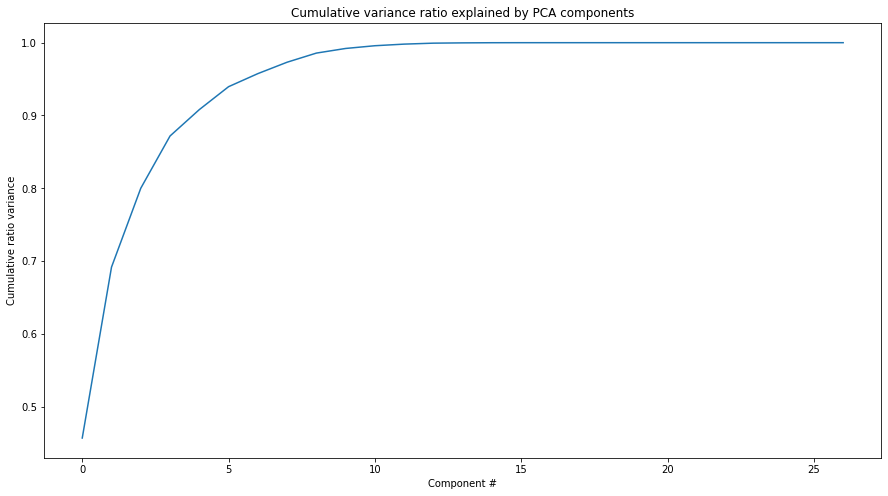
As expected, the two graphs are similar in shape suggesting that the more the covid cases, the more the new deaths and vice versa. It also suggests that a lot of the deaths are indeed related to Covid 19, which is important.

According to the data, the peak of covid deaths was reached around January 2021 with about 4431 deaths on January 20, 2021. The data also shows a peak in January 2022 which suggests that people were more likely to contract and die from the virus after Christmas holidays. This may have been because people move around a lot more at this time of year. Number of people hospitalized also shows a very similar pattern.



Above is a boxplot showing the distribution of new cases across months for the entire period. It is very clear that January has the highest distribution of new cases is in January, followed by December.

After the initial observation of the data, principal components analysis (PCA) was performed on the data in order to reduce its dimensionality and make better visualizations. The data was scaled around the mean before transformation.



As can be seen in the figure above, the first 2 components account for about 69% of the total variance in the data.

The most significant features taken from the PCA were stringency index, hospital patients per million, total tests, and people vaccinated per hundred.

Chart

Description automatically generated

From the figure above, it looks like PC1 and PC2 are just inverses of each other. But it cannot be forgotten that together they account for 69% of the variance in the data.

# FEATURE ENGINEERING

After the EDA, it was now time to start preparing the data so that it could be used to train our machine learning model. One of the most crucial steps to do this is to scale the data. The data was scaled using sklearn’s StandardScaler which basically removes the mean and divides by standard deviation.

# MODELLING

During the EDA, it was found that the relationships between the key features and death rate were quite linear. Based on this, a decision was made to use sklearn’s LinearRegression model. To do this, the data was split into a train set and test set at 75% and 25% respectively. A Grid Search object was trained using the training data. The estimator was the LinearRegression object, and the parameter grid included all hyperparameters for Linear Regression.

Since the EDA indicated what the most significant features were, it was decided that a few other less significant columns would be dropped before training the model. It was also decided that the feature for ICU patients is too related to death rate and would represent a false impression of the model’s ability to predict death rates.

The results of predicting the test data can be seen below:

Mean Squared Error: 0.383

R2 Score: 0.646