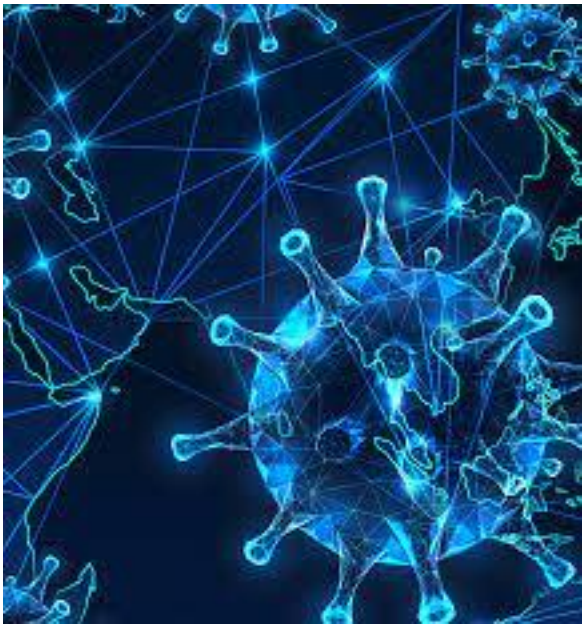


COVID-19 CONFIRMED CASES PREDICTION

using machine learning algorithms and time series forecasting models

Student Name: Thi Ngoc Diem Nguyen
Student ID: 0778061

DAB 402 - CAPSTONE PROJECT
Instructor: Dr. SAVITA SEHARAWAT

Table of Contents

Overview

In early December 2019, an outbreak of coronavirus illness 2019 (COVID-19) was reported in Wuhan City, Hubei Province, China, caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The World Health Organization labeled the outbreak a Public Health Emergency of International Concern on January 30, 2020. Globally, 49,053 laboratory-confirmed cases and 1,381 deaths had been reported as of February 14, 2020. Many countries have implemented a range of control measures in response to the perceived risk of contracting sickness.

People have suffered from this pandemic in many aspects of life. It cost people's time, money, and even human lives. Fact that many people died because of this virus, the world's economy is severely impacted by its consequences. There is no doubt to say it is one of the most catastrophic pandemics in human history. Until now, there are around 495 million people infected Covid-19 and approximately 6.17 million who died because of this virus. The mortality rate is 1.25% which means the risk of dying because of Covid 19 when infected is 1.25%. According to Statista, the fatality rate of major virus outbreaks worldwide as of 2020 is as below:

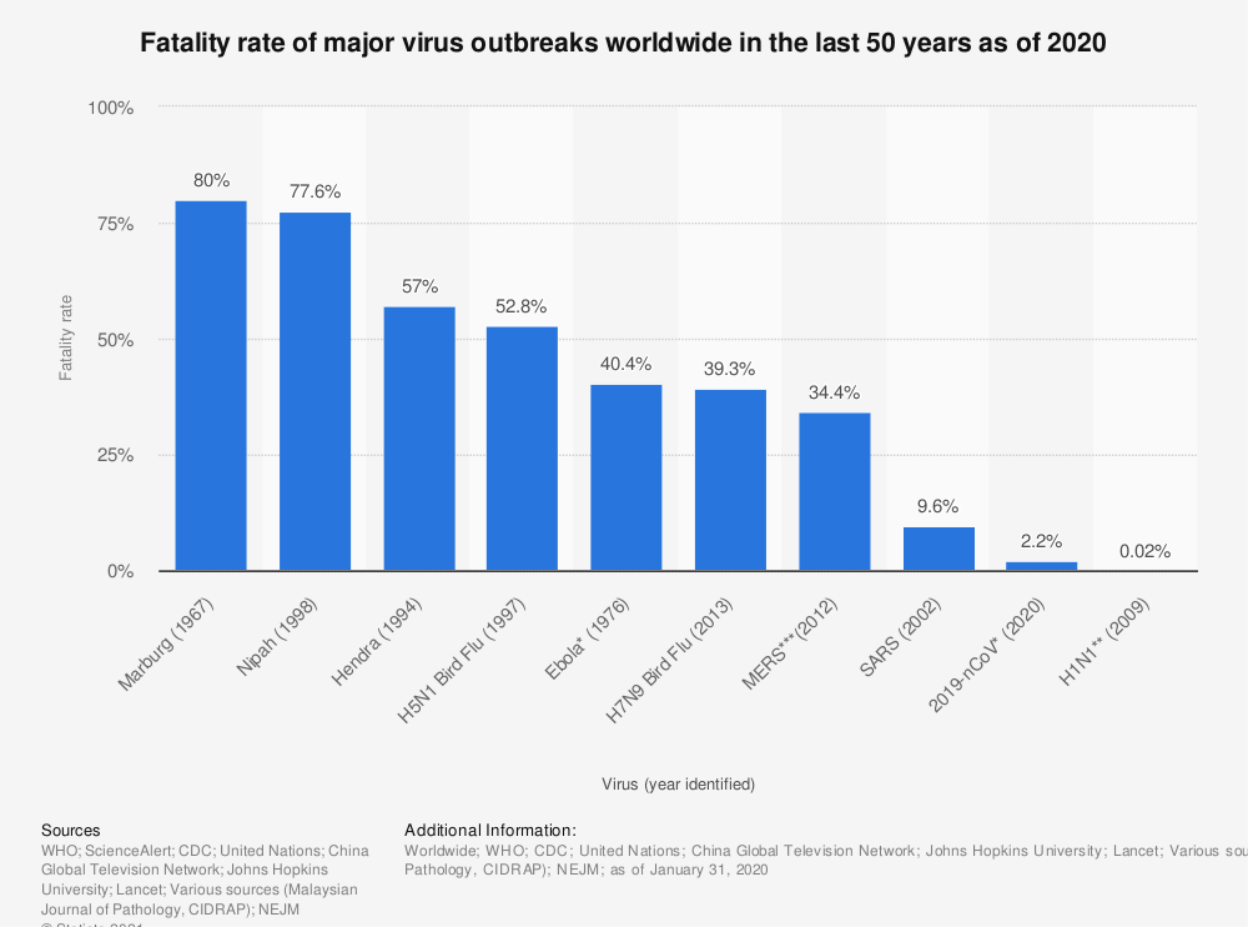


Figure 1: Fatality comparison as of 2020

We can see that compared to other viruses, Covid-19 is less severe than another virus, but it spreads at lightning speed from the nose and mouth through respiratory droplets at close range and through virus particles that float through the air and can stay suspended for quite a while especially in places with poor ventilation. Because of this, Covid-19 is an inevitable disease all over the world in recent 2 years.

Abstract

Thanks to the development of Covid-19 vaccines in late 2020, the number of people dying slowed down by vaccine protection. This concern has become a topic of discussion recently. I am not an exception; I have a great interest in exploring data related to Covid-19 to have an updated look at this matter and how efficient vaccination campaigns bring to us. Hence, I decided to do a little research to understand how Covid-19 status changed in the last 2 years and where Canada is right now in the fight against this kind of virus. I picked my dataset from the [Our World in Data] (https://ourworldindata.org/) website. It comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The reason I chose this data set is that it is updated and clearly shows the vaccination progress for every country, including Canada of course, on daily basis. I downloaded this data set on Jan 22nd, 2022. Therefore, my original dataset includes 156,546 records and 67 attributes which were reported from Feb 15th, 2020, to Jan 22nd, 2022. This data set includes many missing values and outliers. Therefore, I will select 10 attributes among 67 attributes that are related to the purpose of this project for further analysis. Then data will be cleaned, normalized, and pre-processed before moving to the modeling section. This data set includes unlabeled attributes and data points ordered in time so this project will use machine learning algorithms from unsupervised learning and time series analysis for future prediction.

The key outcome of this project is to visualize the correlation of vaccination to several new cases and several new deaths and to use data science capabilities, such as machine learning and time series analysis to predict future values. It's feasible to build a forecasting model that provides sufficiently accurate predictions.

Keywords:

Covid-19 spread, new confirmed cases prediction analysis, death ratio, vaccination progress, Clustering, Polynomial Regression, Random Forests, ARIMA, Time series analysis

Research questions

This project aims to solve these questions:

1. What is the likelihood of dying if people are infected Coviwith d-19 in their countries?

2. How does vaccination impact the number of people dying because of Covid-19?

3. Build a model to predict new confirmed cases in Canada in the next 14 days by time series forecasting.

Tools:

Data extract from Excel, formulation, and data visualization in Python and Power BI.

Github account:

https://github.com/EmilyDiemNguyen/Capstone-Project

Introduction

Literature review

The China Health Authority notified the World Health Organization (WHO) on December 31, 2019, about multiple cases of pneumonia with an unknown cause in Wuhan, Hubei Province, central China. Many patients worked at or resided near the nearby Huanan Seafood Wholesale Market when the cases were first reported on December 8, 2019, however other early cases had no connection to this market [1]. On January 7, a novel coronavirus, designated 2019-to by the World Health Organization, was discovered in a patient's throat swab sample [2]. The Coronavirus Study Group renamed this pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the condition coronavirus disease 2019 (COVID-19) by the WHO. As of the 10th of April, there are around 495 million people infected Covid-19 and approximately 6.17 million who died because of this virus. To be updated on the number of cases, deaths, and vaccine doses, we can visit: *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* (Johns Hopkins University, accessed 7 April 2021) [3]. Virus transmission and mortality have been reduced since the beginning of the pandemic through a variety of measures, including individual precautions such as social distancing, wearing facemasks, hand hygiene, and limiting interpersonal contact to outdoor settings; widespread testing to identify individuals infected with the virus; and non-pharmaceutical policy responses from governments, such as school and workplace closures, bans on public gatherings, and travel restrictions. Governments are now looking to vaccination as a critical answer to the epidemic, following the successful creation, evaluation, and production of various vaccinations.

We need timely, comparable data across countries to assess the scope and rate of vaccination implementation. Our World in Numbers The COVID-19 dataset is a publicly available global aggregated dataset. It covers the entire time beginning on Feb 15, 2020 and has been updated regularly since then. As additional nations release official data on their national reporting and immunization efforts, the COVID-19 dataset continues to grow. The dataset covers 238 nations as of Feb 22, 2022. Official sources, such as health ministries, government reports, and official social media accounts, were used to compile this information.

The scientific community has widely utilized our dataset in a variety of fields. It has been used to draw attention to worldwide discrepancies in vaccine access, prompting strong demands for action to speed up financial and policy responses to overcome the gaps that exist [4].

Researchers have used it to identify nations with exceptionally good vaccination rollouts, allowing for investigations of how this was accomplished [5].

Kaggle is a data scientist and machine learning practitioner community. Users can use Kaggle to search and publish datasets, explore, and construct models in a web-based data-science environment, collaborate with other data scientists and machine learning experts, and participate in data-science competitions. In the case of the Covid-19 pandemic, the portal posts a new challenge every week for people to work on Covid-19 data.

It is worth mentioning that there are some limitations and challenges in the data set.

- Time-varying nature: The outbreak's needs necessitate a quick response, which means gathering the most up-to-date information. This poses some significant difficulties. Government policies, for example, are always changing. Information is frequently out of date by the time it is discovered. Every day, the number of countries enacting or changing measures grows [6].

When working with numerous data sources at the same time, data availability daily can be a problem.

- The number of confirmed cases is not a reliable metric: A confirmed case is defined as a person with laboratory confirmation of Covid-19 infection, regardless of clinical signs and symptoms, according to the WHO global [7]. At the time of the pandemic's onset, access to mass testing was extremely limited, and only a small percentage of hospitalized cases were tested in a laboratory setting. As a result, the complex and constrained testing method filters out the vast majority of illness reports. In addition, only a few datasets include statistics on the number of suspected cases. Even if everyone with modest symptoms is tested, the results will only provide an approximation of the disease's symptomatic cases. The study of the fraction of asymptomatic cases is an important subject of research, not only because it is one key to estimating the total number of infected individuals, but also because it is critical to the virus's transmission [8].

- The mortality rate is difficult to estimate: During the most severe stages of a country's virus transmission, the number of mortality cases recorded by the administration often ranges significantly from the actual number. Because only deaths with prior test proof of the condition are included, this is the case.

- Government transparency: There are significant discrepancies in how the countries are reporting Covid-19 statistics. Furthermore, there are some questions about the countries' transparency in terms of the data they offer.

The goal of this project is to describe current knowledge on COVID-19 and emphasize the effectiveness of vaccines in reducing the risk of ICU (intensive care unit) and deaths. Besides that, I also applied some machine learning capabilities to build multiple regression models (Polynomial

Regression, Random Forest Regression) and time series analysis models (Double Exponential Smoothing, Triple Exponential Smoothing, ARMIA) to forecast the number of the new confirmed case in Canada in the next 14 days.

Data analysis to determine the current stage of the pandemic and construct forecasting models: different models can be created utilizing the variables to estimate the current condition of the pandemic and predict the reaction to Covid-19 spread. Estimation of the infected population is an example of an estimation and forecasting analysis.

- Economic impact estimation.
- The number of infected will be used to forecast the impact on the health system.
- Trying to figure out how much of an impact there is in terms of mortality.
- Seasonal behavior is examined.

Decision-making: The data-driven models' ultimate goal is to provide useful tools that will aid governments and institutions in anticipating and evaluating their responses to Covid-19 dissemination. Among these, the following are the most important:

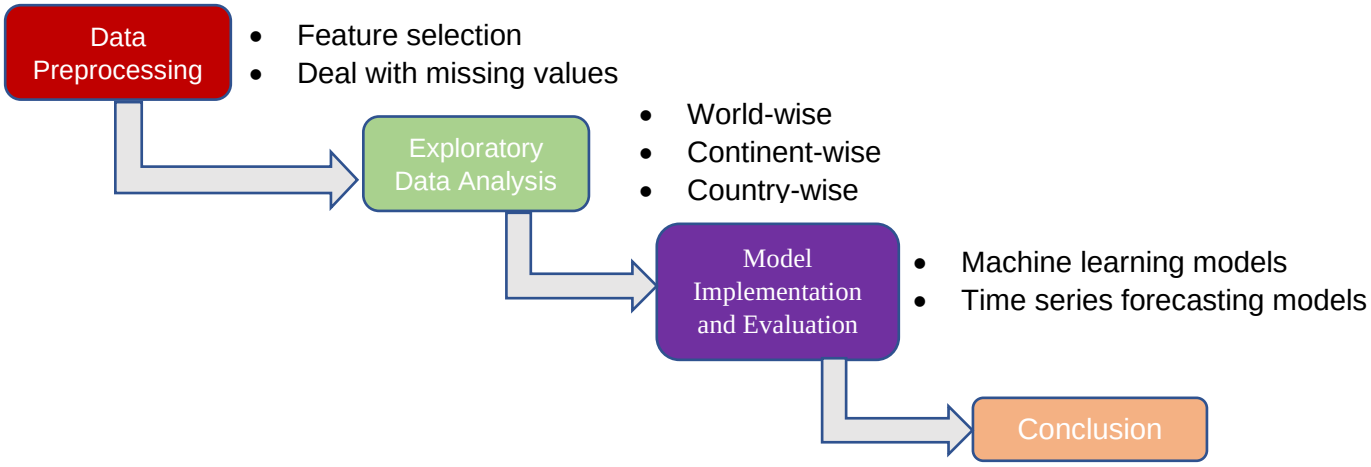
- The effectiveness of the measures is being evaluated.
- Government actions are being planned ahead of time.

No one can deny that the current COVID-19 pandemic is an international public health problem. There are many types of research on this data set by scientists or even students.
In this project, we evaluate key open-data sources to better comprehend Covid-19's global dissemination. We list the variables that must be included in epidemiological and forecasting models. We concentrate on the Covid-19 confirmed case, deaths, people fully vaccinated, and time series forecasting.

The current state of the available Covid-19 open data is examined. Unfortunately, a variety of challenges such as data inaccuracy, shifting criteria, a huge diversity of sources, non-comparable measures between countries, delays, and so on make it far from perfect. Despite the challenges, the open-data resources on Covid-19 and related variables present several potentials to various communities. Epidemiologists, data-driven researchers, health-care specialists, the machine-learning community, data scientists, and others.

Methodology:

The workflow to build models can be summarised as follows:



Data Preprocessing

Features selection:

The original data set includes 156,546 records and 67 attributes. In this step, I only picked 10 attributes among 67 attributes from the original data set which are most related to the purpose of this project.

The first step is to check the data types of each attribute and then check statistics for numeric attributes.

They are continent, location, date, total cases, new cases, total deaths, new deaths, ICU patients, people fully vaccinated, and population.

Detailed Data Dictionary:

In this step, I took each attribute in clean data set to do the analysis. First, I assigned the correct data type of data attribute. For numeric attributes, I checked five mean number summaries, standard deviation, minimum, 1^{st,} and 3rd quartile, and maximum. Next, I will check the value count of each categorical attribute. Then I checked the missing values in the data set.

Table 1: Summary of Categorical Attributee

Attribute	Data Type	Value count	Description
continent	object	6	Continent of the geographical location

Attribute	Data Type	Value count	Description
location	object	224	Geographical location
date	object	730	Date of observation

Table 2: Summary of Numeric Attributes

Attribute	Description	mean	std	min	25%	50%	75%	max
total_cases	Total confirmed cases of COVID-19. Counts can include probable cases, where reported.	2222006	13200489	1	1674	21998	266867	3.491346e+08
new_cases	New confirmed cases of COVID-19. Counts can include probable cases, where reported.	9576	66857	-74347	1	73	964	4.232499e+06
total_deaths	Total deaths attributed to COVID-19. Counts can include probable deaths, where reported.	136289	283933	1	71	705	6681	5.591704e+06
new_deaths	New deaths attributed to COVID-19. Counts can include probable deaths, where reported.	170	829	-1918	0	2	19	1.806100e+04
icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day	901	2700	0	26	140	582	2.889100e+04

Attribute	Description	mean	std	min	25%	50%	75%	max
people_fully_vaccinated	Total number of people who received all doses prescribed by the vaccination protocol. If a person receives the first dose of a 2-dose vaccine, this metric stays the same. If they receive the second dose, the metric goes up by 1	58281582	28334392 9	1	233092	1831694	10690639	4.091460e+09
population	Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv for full list of sources	148100306	70715343 0	47	1172369	8715494	33933611	7.874966e+09

- **Confirmed cases and deaths:** our data comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The cases & deaths dataset is updated daily. Note: the number of cases or deaths reported by any institution—including JHU, the WHO, the ECDC, and others—on a given day does not necessarily represent the actual number on that date. This is because of the long reporting chain that exists between a new case/death and its inclusion in statistics. This also means that negative values in cases and deaths can sometimes appear when a country corrects historical data because it had previously overestimated the number of cases/deaths. Alternatively, large changes can sometimes (although rarely) be made to a country's entire time series if JHU decides (and has access to the necessary data) to correct values retrospectively.

Table 3: Summary of Date Attribute

Attribute	Start Date	End Date
date	2020-02-15	2022-01-22

Missing values:

Table 4: Missing values

Attribute Name	No of Missing Value
continent	9410
total_cases	2832
new_cases	2896
total_deaths	20257
new_deaths	20095
icu_patients	135254
people_fully_vaccinated	120338
population	1033

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Having clean data will enhance overall productivity and help you to make decisions based on the best quality information available. These are steps that I did for the data cleaning part.

Change data type

The date is the only data type I need to change from object to date time for the visualization part later.

Deal with null values

- Continent attribute

When the continent is null, the location attribute showed not only the country name but also some other groups, eg: European Union, High income, International, Upper middle income, World. In this project, I analyzed several metrics based on a country level, so I decided to remove any records having null values at the continent attribute. After removing these records, the location attribute showed 225 country names, so I changed that attribute's name from location to country.

- Population attribute

Next, I want to check the population attribute because, in the latter part, I will count how many percent of the population infected with Covid-19 or died because of this virus. Therefore, population data is vital for my project. I will check which country has null data in population then

I will remove them from the data set. There was only one country has no value in population Northern Cyprus. When I remove this country, my country attribute included 224 countries.

- For cumulative data: total cases, total deaths, ICU patients, people fully vaccinated

Because this data set is time base data in which the next values depend on the previous values, so I apply the ffill() function to fill null values. By applying this function, null values will be replaced by their previous value. In case, there is no value before, the null values were filled with 0.

- For non-cumulative data: new cases, new deaths

As we already filled total cases and total deaths attributes, then from this, we can calculate new cases and new deaths on each day by taking the difference between the next day to the current day.

Check duplicate values

The last step is to check whether there were any duplicate values or not. Our data set didn't include any duplicates. Now our data was cleaned and ready for the next parts.

Exploratory Data Analysis (EDA)

Geographical analysis

In this project, new cases and new deaths are the primary attributes I want to focus on to study the spread and fatality of Covid-19 to human beings' health. Therefore, I will plot these two attributes to see the correlation between infection and death.

We can see the number of cases and deaths are not on the same scale. The mean of new cases is 9,576 and the means of new death is 170. Therefore, to plot them on the same chart to see the correlation, we need to normalize data first by using the the the the Min-Max Scaler function. This function transforms features by scaling each feature to a given range. By doing so, all features will be transformed into the range [0,1] meaning that the minimum and maximum value of a feature/variable are going to be 0 and 1, respectively.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

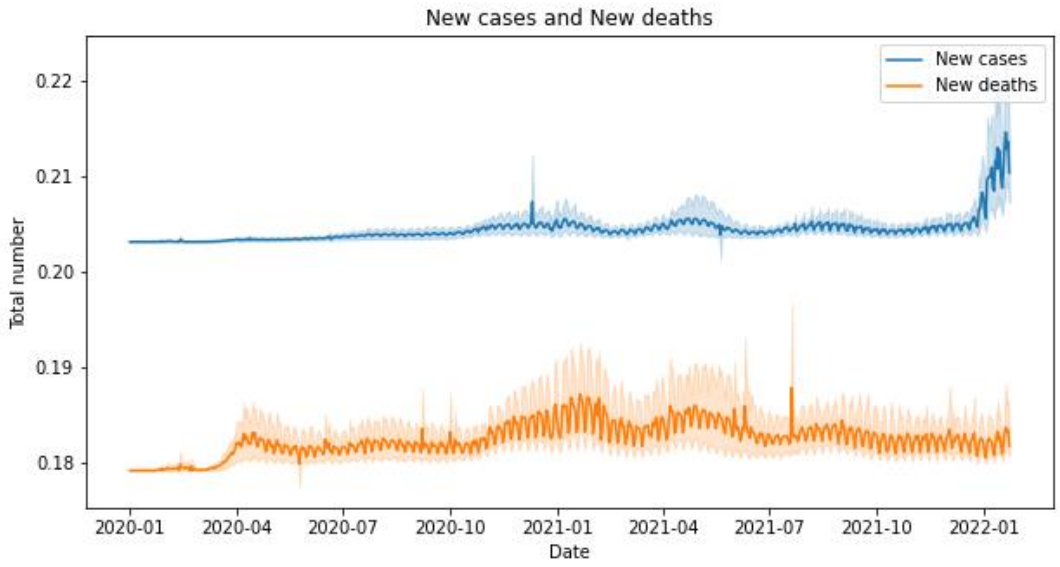


Figure 2: New cases and new deaths over the time

When new cases and new deaths are plotted on the same scale, we can see:

The number of cases is much higher than the number of deaths.

- There are 4 waves of new cases that affected the world which correspondingly match with 4 variants of Covid-19: the original variant, beta, delta, and omicron. But the latest variant (Omicron) has rapidly surged past other variants to become the dominant SARS-CoV-2 strain. The truth is Omicron is the most super spreading variant till now.
- There is a similarity between waves of new cases and waves of new deaths. When the new cases increased, new deaths also increase too. But after vaccinations (end of 2020), new deaths are likely to decrease. While the trend of new cases is going up, the trend of new deaths went down and remained unchanged. This can be explained by 2 reasons: the effectiveness of vaccines and the Omicron variant is less severe than other previous ones.

Next, we will look at the data set on the continent level. How many per the center of the population were infected with Covid-19 and how many percent died because of this virus?

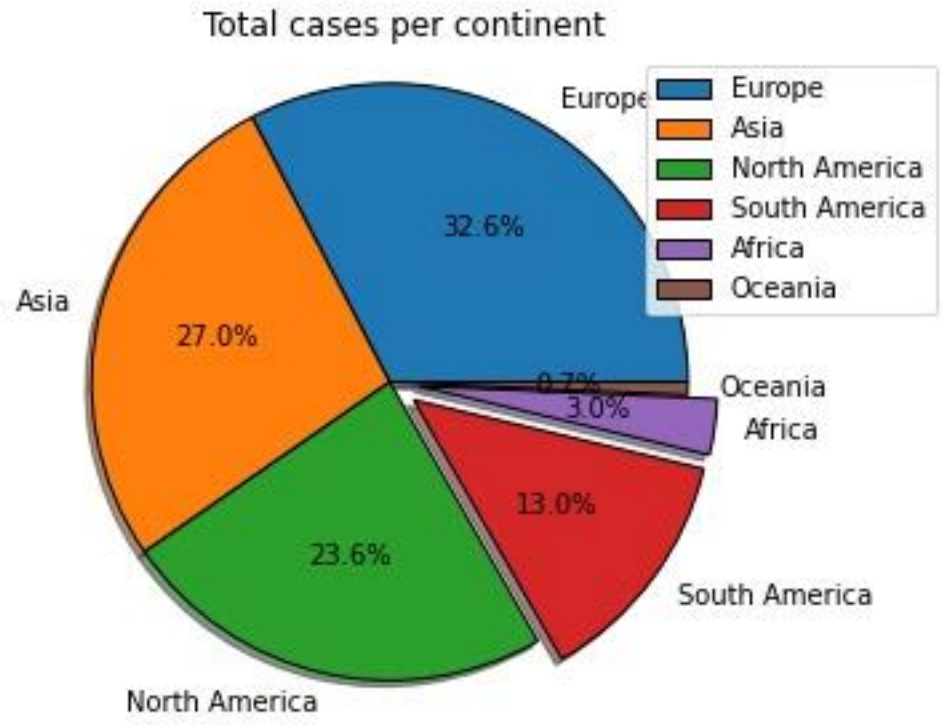


Figure 3: Total cases per continent

There are 6 continents which have been shown here: Europe, Asia, North America, South America, Africa, and Oceania. Europe is the leading with a 32.6% number of confirmed cases globally. Next is Asia with 27%, North America with 23.6%, South America with 13%. The minor part belongs to Africa 3% and Oceania 0.7%. Through this chart, I could say that the pandemic mainly spread in Europe, Asia, and North America. We will find out which countries have the highest number of cases in the next part.

Similarly, we want to know the percentage of deaths cases globally at the continent level.

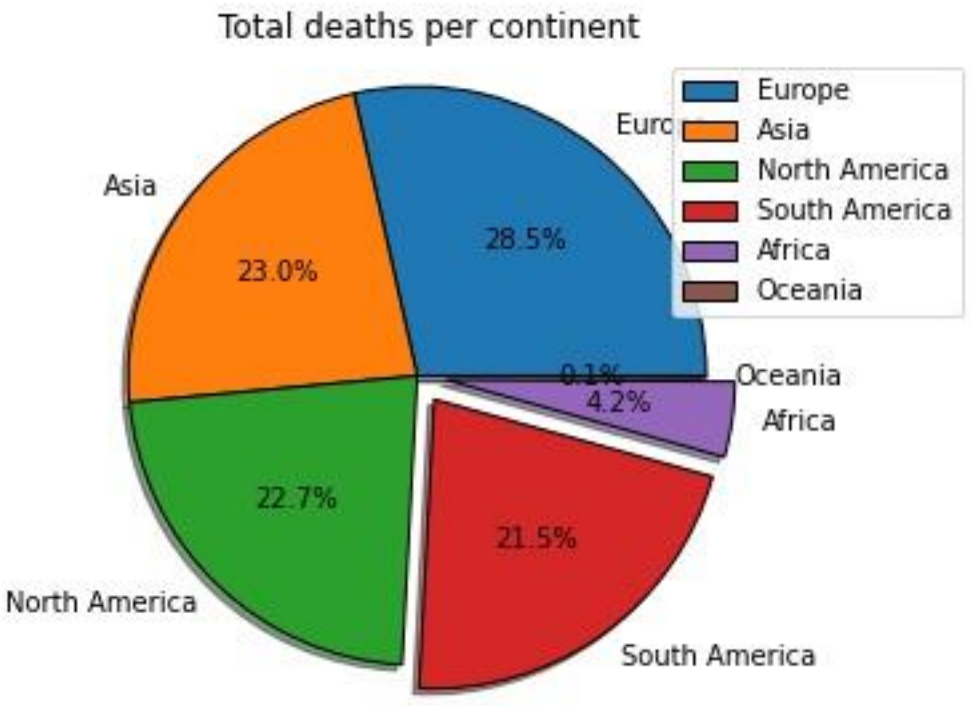


Figure 4: Total deaths per continent

Europe was leading in the ranking of death cases around the world with 28.5%. The number of deaths in Asia, North America, and South America is nearly equal.

Through these charts, we can see in most cases, the percentage of new cases is less than the percentage of deaths. But this is not true in South America and Africa. We can say that the death ratio (total deaths/ total cases) in South America and Africa is higher than in other continents. It means that people in South America and Africa have a higher risk of death when they are infected with Covid-19.

Finally, we will look at the data set at the country level. In this section, I created 3 new attributes which were the infected ratio, death ratio, and the vaccination ratio. They will be defined by:

- Infected ratio: total cases/ population

- Death ratio: total deaths/ total cases
- Vaccination ratio: people fully vaccinated/ population

After that, I only considered those countries which have a population of more than 10million to plot them.

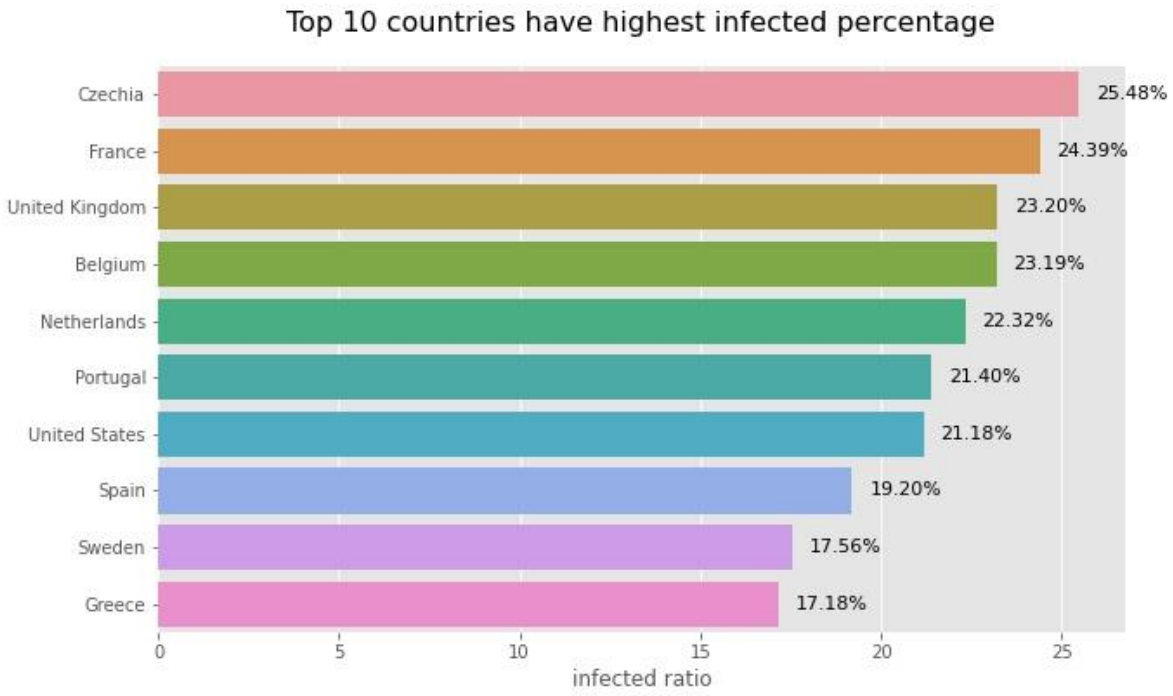


Figure 5: Top 10 countries with the highest infected ratio

Here are countries that have the highest confirmed ratio. According to this data set, 25% of people in Czechia are positive for Covid-19. This is an extremely high number. Most of these countries belong to Europe and America. Canada was not on this list.

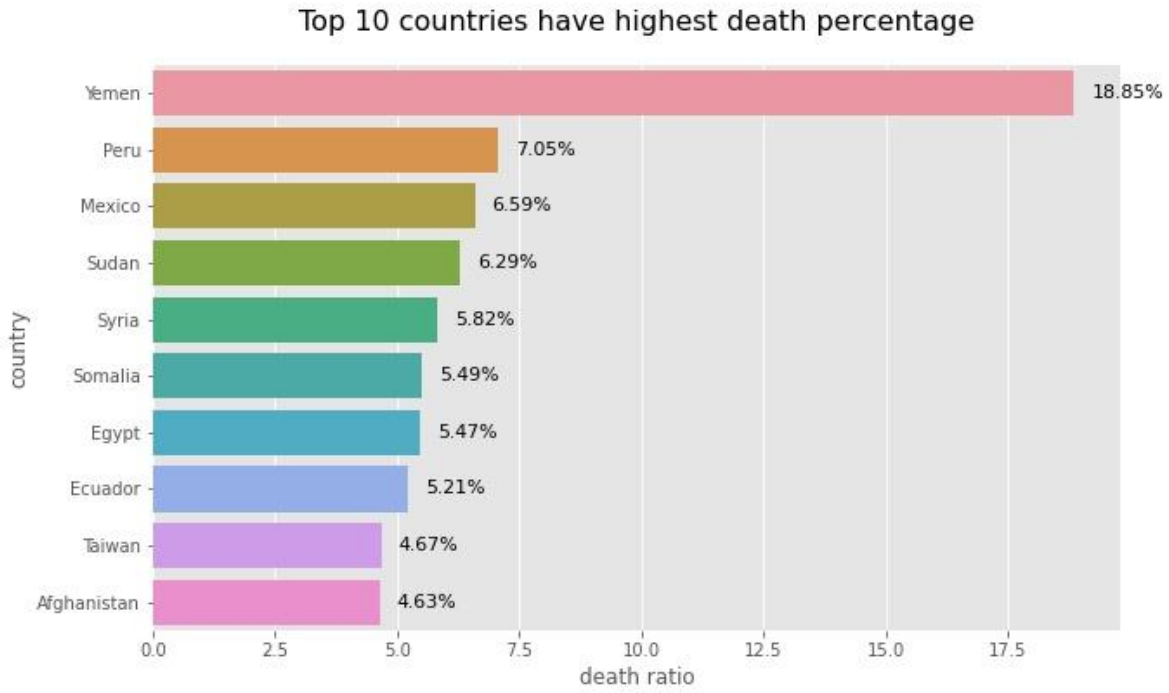


Figure 6: Top 10 countries with the highest death ratio

However, most death cases belong to very different countries, not the countries in the highest confirmed cases list. The top 10 countries with have highest death ratio belong to Asia, South America, and Africa. By logical thinking that countries have high confirmed ratios should have high death ratios accordingly. Why these lists are very different?

Then vaccination ratio comes into the picture. Countries have a high confirmed ratio, but they rolled out vaccine campaigns successfully and can help reduce the risk of death. Especially, Portugal and Spain are typical examples.

Canada is ranked 27th in the number of people deaths, the possibility of being infected with Covid 19 in Canada is 7.58% of people are infected with Covid-19, the likely hood of dying because of

Covid-19 is 1.12%. And Canada is doing a great effort in vaccination progress when 78.2% of Canadians are fully vaccinated.

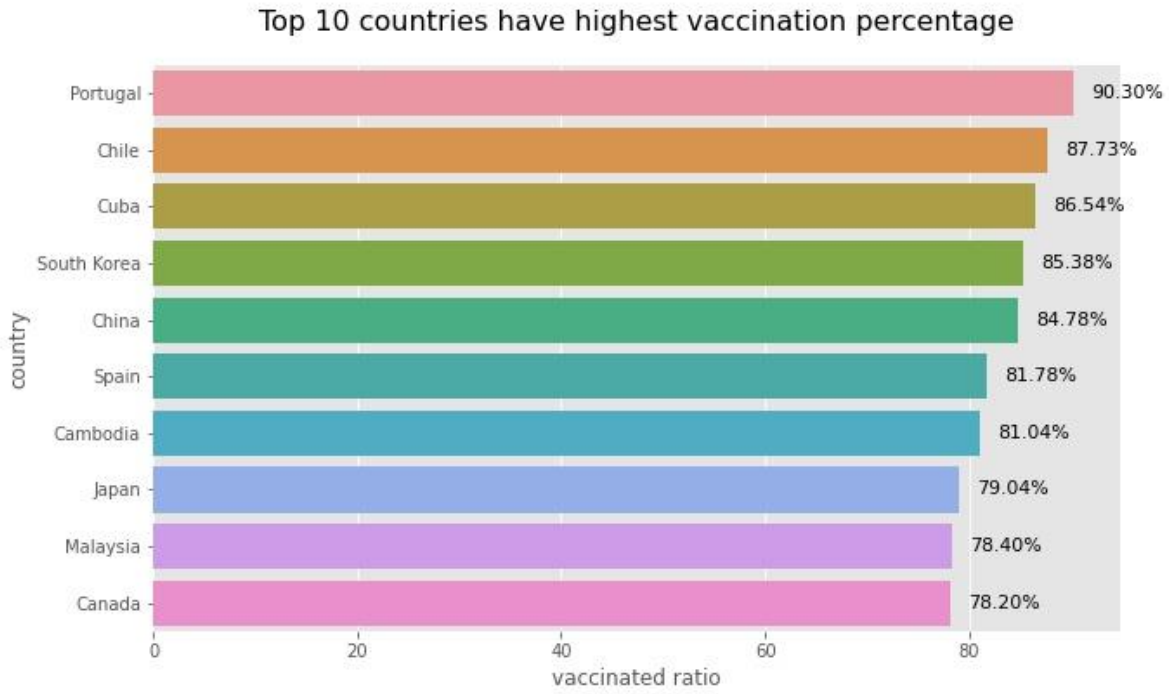


Figure 7: Top 10 countries with the highest vaccination percentage

Countries with a high ratio of vaccination do not exist in the list of high death ratios. We will study the relationship between vaccination and death ratio in the next part of this project.

Clustering countries by death ratio

The death ratio of these countries varied in a wide range from 0.88% to 18.85%. So before coming to clustering, I normalized data by using Standard Scaling since K-means clustering is a distance-based algorithm. To identify the number of optimal k, I used the hierarchical clustering technique and elbow technique. Hierarchical cluster analysis (HCA), often known as HCA, is an unsupervised clustering approach that includes constructing groups with dominant ordering from

top to bottom. The program divides objects into clusters based on their similarity. The endpoint is a collection of clusters or groups, each of which is distinct from the others yet the items inside each cluster are broadly similar. And the elbow approach is a heuristic used in cluster analysis to determine the number of clusters in a data set. Plotting the explained variation as a function of the number of clusters and selecting the elbow of the curve as the number of clusters to utilize is the method. Elbow is one of the most well-known approaches for determining the optimal value of k and improving model performance. It selects a set of values and selects the best among them. It calculates the average distance and the sum of the squares of the spots.

Both methods gave the same result of k, the optimal value is 3. It means that there are 3 groups of countries with different death ratios. Look at the scatter plot below:

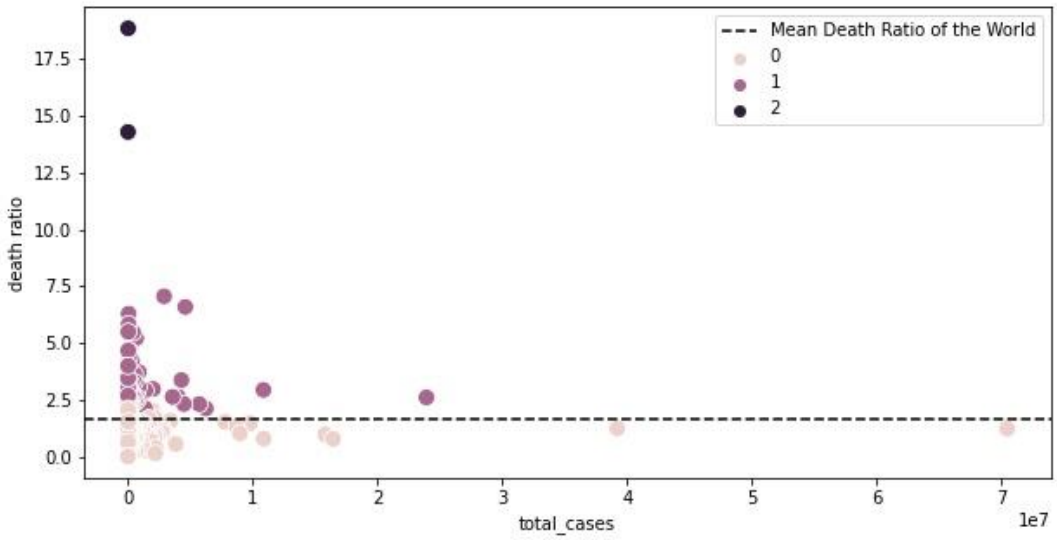


Figure 8: Clustering by death ratio

The dashed line is the average mortality rate of the whole world. It was estimated at around 2%. Cluster 0 are countries below the average line which means they are below the risk of average

death and those above the line are above the risk. Cluster 0 which have a high number of total cases but a low death ratio while cluster 2 lower number of confirmed cases but a higher death ratio. Countries belonging to Cluster 2 are alerted in the fight against Covid-19. Most of them are poor countries that have low standards of health care systems.

Few Countries belonging to Cluster 0: ['United States', 'India', 'United Kingdom', 'Italy', 'France', 'Argentina', 'Germany', 'Spain', 'Turkey', 'Philippines']

Few Countries belonging to Cluster 1: ['Brazil', 'Russia', 'Mexico', 'Peru', 'Indonesia', 'Iran', 'Colombia', 'Ukraine', 'Poland', 'South Africa']

Few Countries belonging to Cluster 2: ['Yemen', 'Vanuatu']

US, India, UK, Italy, France, Germany, Spain, and Turkey are countries in the top 10 that have the highest Covid-19 cases in the world; but they are belonging to Cluster 0 which has a low death ratio. One of the reasons is that they got vaccines in the early stage.

How does vaccination impact the number of people dying because of Covid-19?

In this part, we will explore the effectiveness of Covid-19 vaccines in 3 countries: Canada, the US, and the UK. As mentioned in the previous part of this report, I am more concerned about the status of Canada on the journey of fighting Covid-19, so in this part, I will choose Canada to analyze the correlation between vaccines and confirmed cases, death cases, or ICU cases.

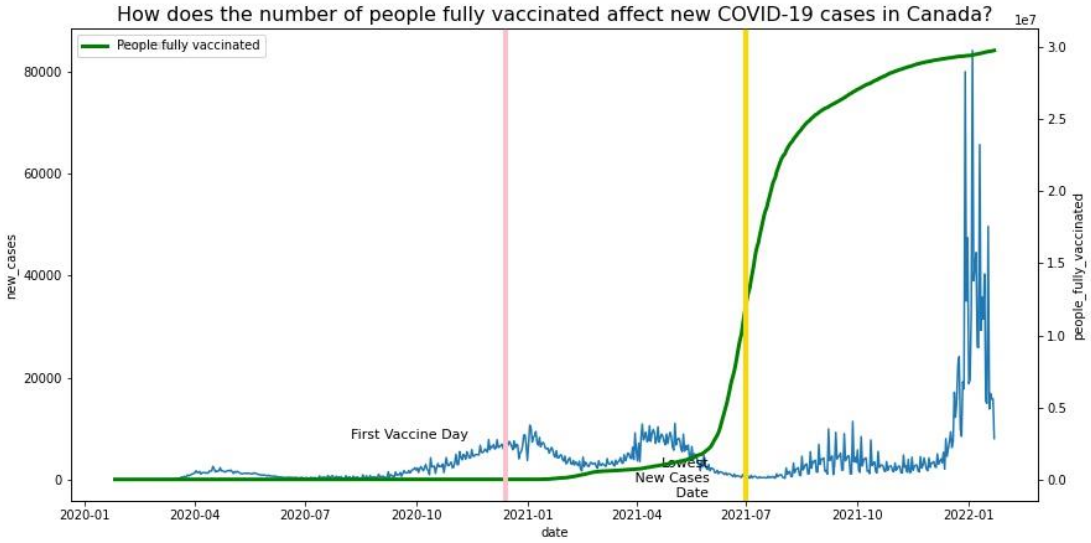


Figure 9: The impact of vaccines on confirmed new cases in Canada

Canada rolled out the 1st vaccine on December 14, 2020, and on July 1st, 2021, the number of new Covid cases generally reached the bottom when many people were fully vaccinated. But despite the number of fully vaccinated Canadian surging significantly, the new cases also reached a peak in Jan 2022. This was because of a new variant - Omicron. Therefore, we could say the current vaccines can't stop spreading the virus, especially the new variant.

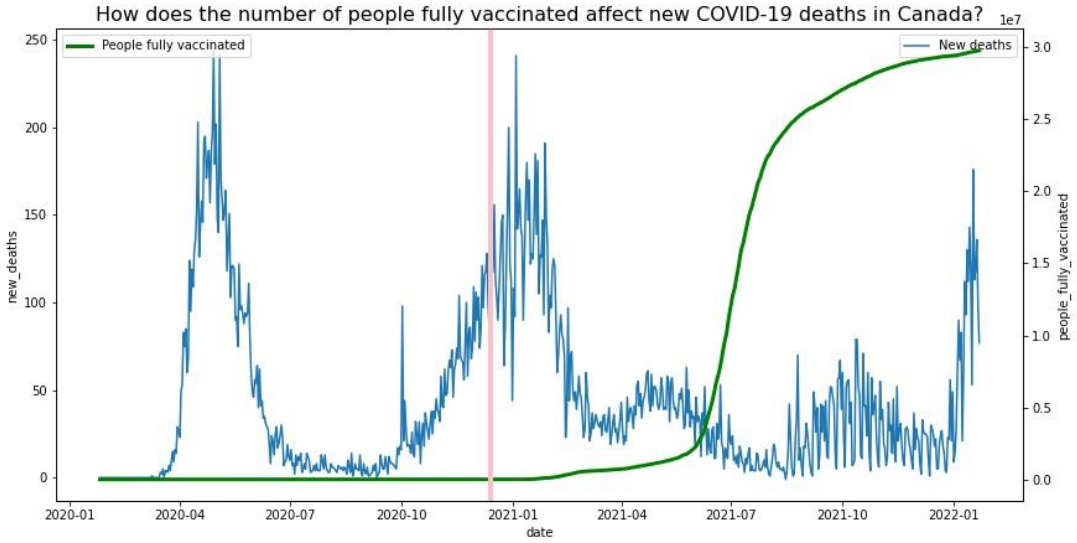


Figure 10: The impact of vaccines on the number of deaths in Canada

From the chart above, I noted that there is the opposite trend of deaths and people fully vaccinated. This proves the effectiveness of vaccines to reduce the risk of dying because of Covid19.

To make it clearer about the effectiveness of vaccines, I compared the infected ratio, death ratio, and ICU ratio before and after vaccination in three countries: Canada, the US, and the UK.

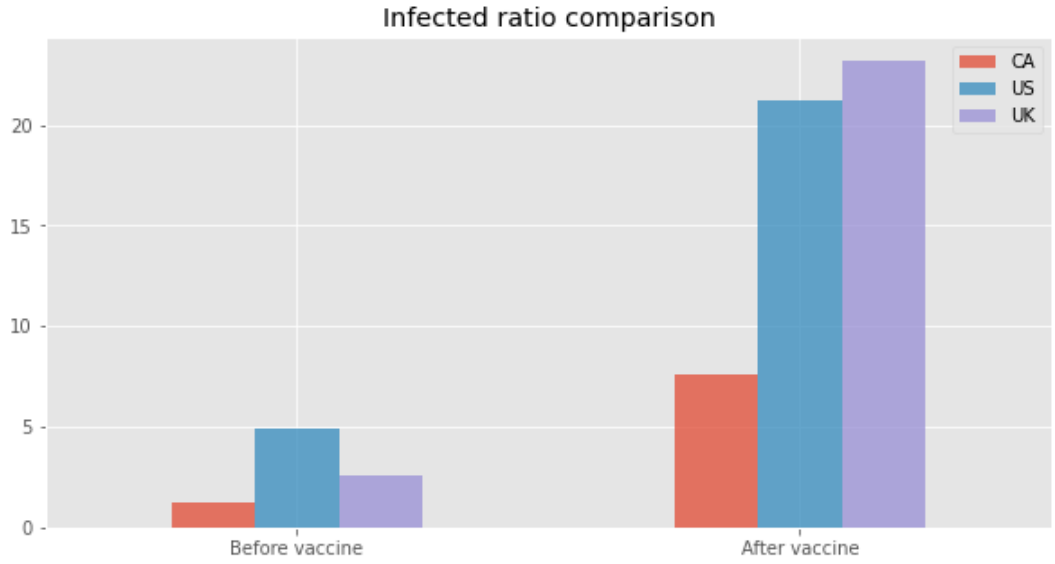


Figure 11: Infected ratio comparison

The bar plot above shows that the percentage of the population infected with Covid-19 increased significantly over time despite there being more people who are fully vaccinated. One more time, we can say vaccines can't stop spreading the virus among communities.

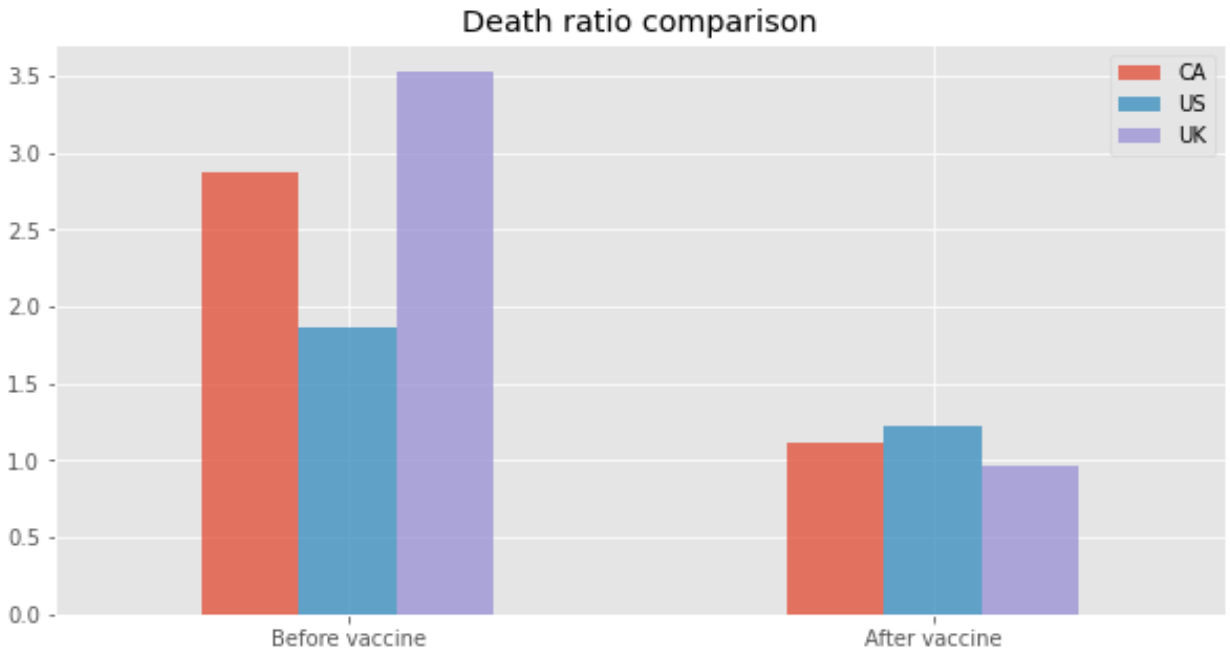


Figure 12: Death ratio comparison

From the above chart, we can conclude that vaccines can help reduce the risk of dying when they are positive for Covid 19. After being vaccinated, the fatality rate significantly dropped. Canada, the US, and the UK are typical examples.

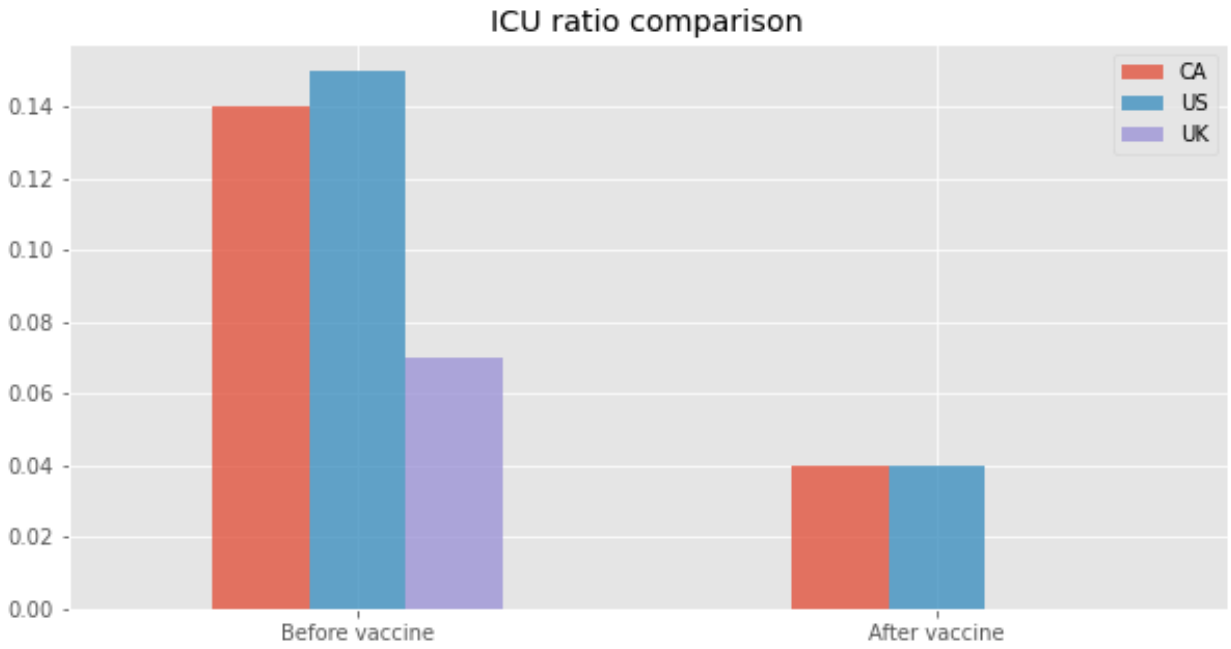


Figure 13: IUC ratio comparison

This chart proves more clearly the effectiveness of the vaccine. It reduces the percentage of ICU patients after vaccination. Or in the other way, we can say, the vaccine can lessen the severity of Covid 19 when they are infected.

Build models to predict new cases in Canada in the next 14 days.

Prediction using machine learning models

In this model, we assigned dates for X and new cases for y. This is the scatter plot that shows the number of cases over time.

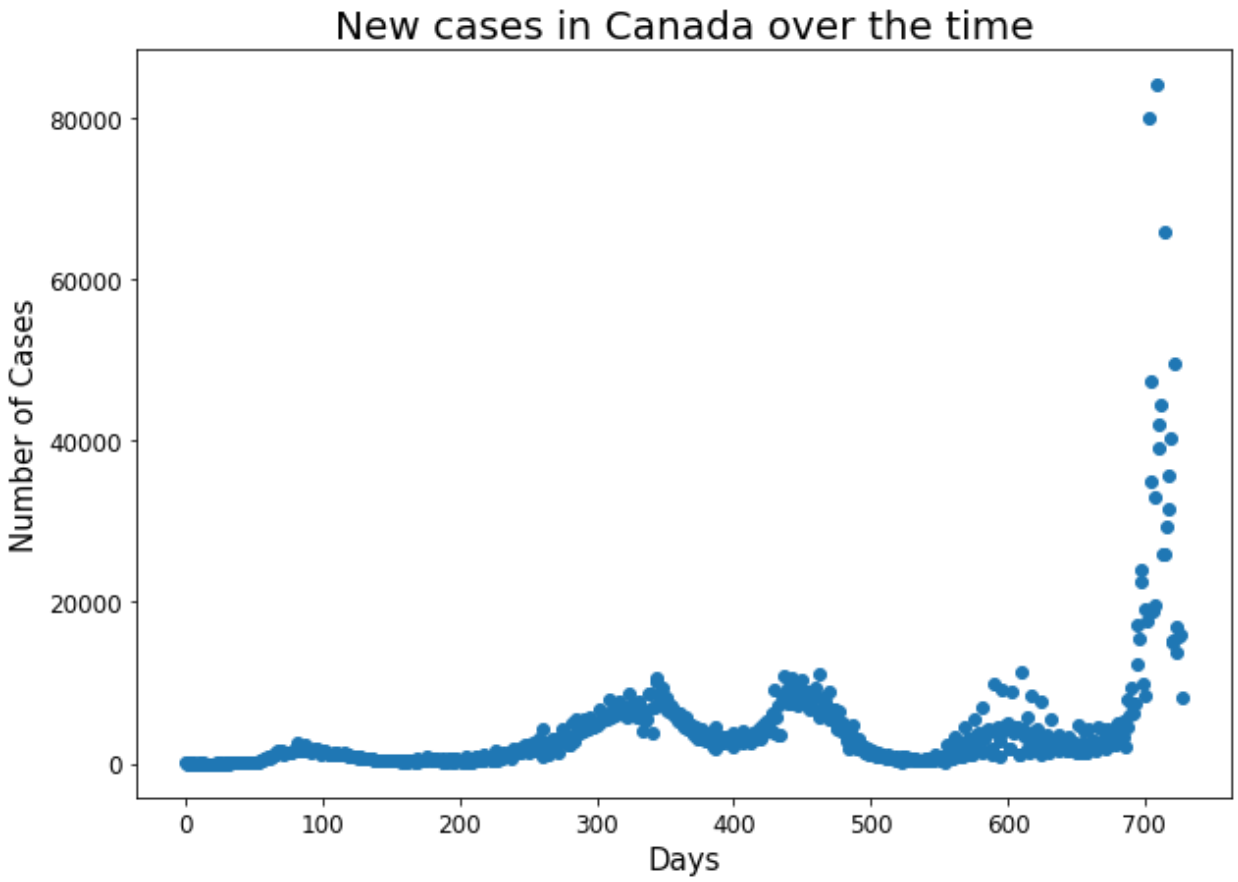


Figure 14: New cases in Canada over the time

From this scatter plot, we can see that X and y have a non-linear relationship.

Now we will apply some machine learning models and time series forecasting models to our prepared data and then compare RMSE to find the optimal model.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals. In other words, it tells us how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

Before coming to the modeling part, we used the train test split function to divide the data set into 2 parts, training, and testing. The training set includes 90% of values and the testing set includes 10%.

Polynomial Regression Model

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y.

To find the best model of polynomial regression, we evaluated the RMSE of each model with a degree from 1 to 100. As a result, the best model which provides the lowest RMSE has a degree is 5 and RMSE is 15,566.39.

This is the plot of the real cases and the prediction by Polynomial regression.

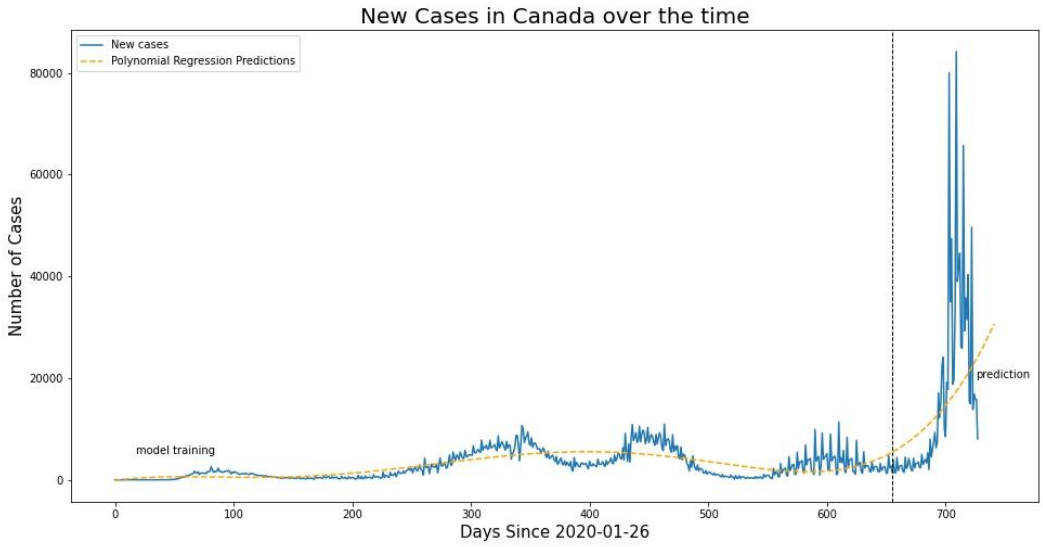


Figure 15: New cases in Canada with Polynomial Regression Predictions

This dashed line is quite similar to the original data, but the trend of prediction is going up while the original data is going downwards.

Random Forest Regression

The second model we used is Random Forest Regression. Random forests or random decision forests is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. In this project, I will use the Regression task. Similar to the previous model, to find the optimal parameters we will evaluate models by comparing

RMSE. Then the lowest RMSE of the Random Forest Regression model is 21,910. Then we will plot the original data with the prediction to see how fits this model.

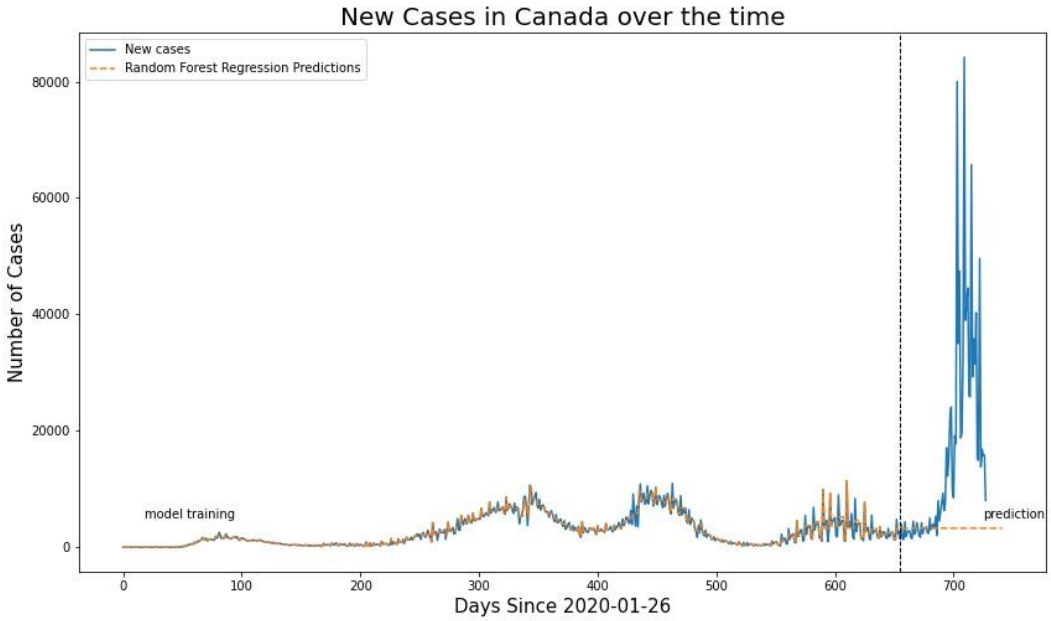


Figure 16: New cases in Canada with Random Forest Regression predictions

The prediction fitted the training set but did not fit the testing set from this plot. Therefore, this model is not suitable to predict the number of new cases.

Time Series Forecasting Models

In time series analysis, the train set and test set are divided by the order of time. Because the Canada data set has 728 records, I will assign a train set from 0 to 715 and the rest 14 records are for the test set.

Double Exponential Smoothing Model

Double Exponential Smoothing is an extension to Exponential Smoothing that explicitly adds support for trends in the univariate time series.

In addition to the alpha parameter for controlling the smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in a trend called beta (b).

The method supports trends that change in different ways: an additive and a multiplicative, depending on whether the trend is linear or exponential respectively.

Double Exponential Smoothing with an additive trend is classically referred to as Holt's linear trend model, named for the developer of the method Charles Holt.

Additive Trend: Double Exponential Smoothing with a linear trend.

Multiplicative Trend: Double Exponential Smoothing with an exponential trend.

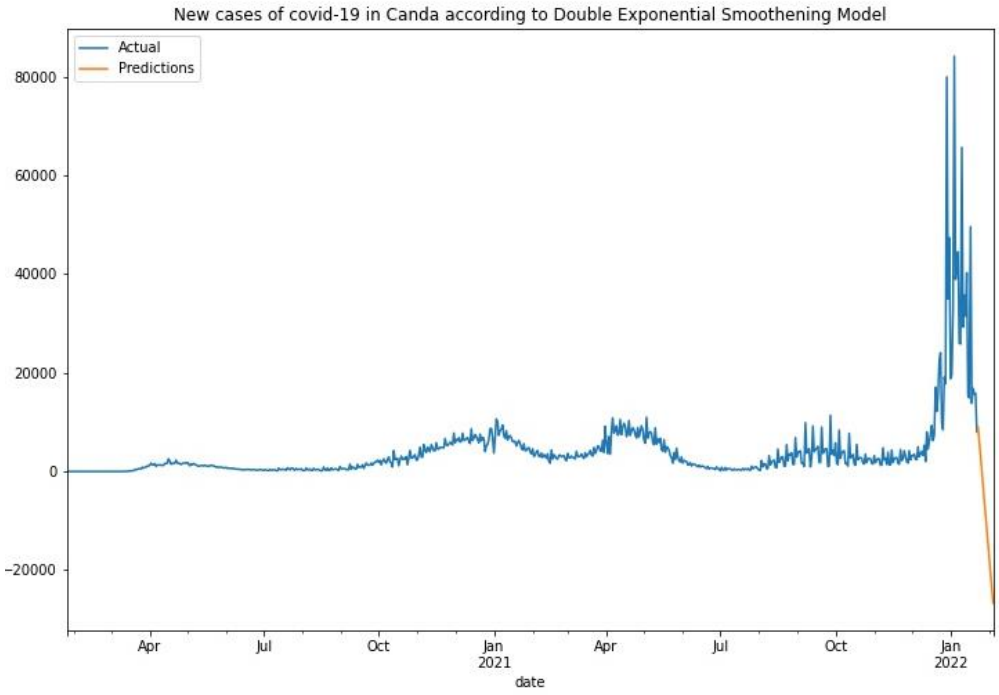


Figure 17: New cases in Canada with the Double Exponential Smoothing model

According to this chart, the new cases in the next 14 days will significantly drop to a negative value. This is not an accurate prediction. The RMSE of this model is 31,833.

Triple Exponential Smoothing Model

Triple exponential smoothing is used to handle the time series data containing a seasonal component. This method is based on three smoothing equations: stationary component, trend, and seasonal. Both seasonal and trends can be additive or multiplicative. Triple exponential smoothing is also known as Holt-Winters Exponential Smoothing.

Triple exponential smoothing is the most advanced variation of exponential smoothing and through configuration, it can also develop double and single exponential smoothing models.

Being an adaptive method, Holt-Winter's exponential smoothing allows the level, trend, and seasonality patterns to change over time.

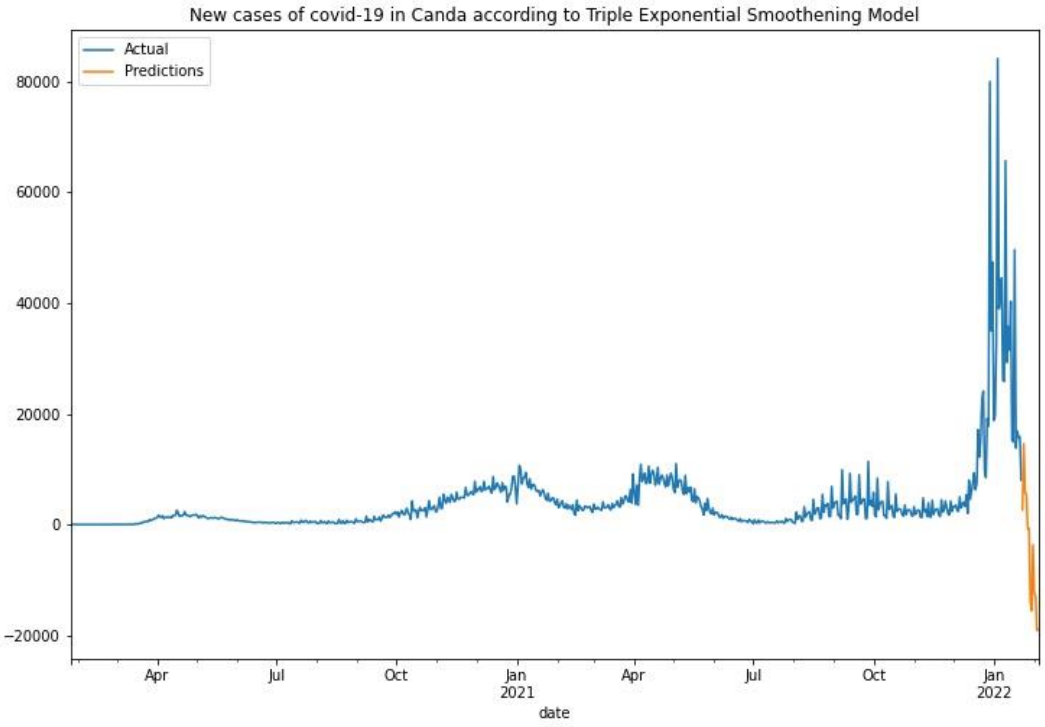


Figure 18: New cases in Canada with the Triple Exponential Smoothing model

Again, this method did not provide accurate predictions when predicting new cases in Canada for the next 14 days. Its RMSE is 34,780 which is even higher than the Double Exponential Smoothing.

ARIMA models Autoregressive Integrated Moving Average

A time series is a sequence where a metric is recorded over regular time intervals

Forecasting is the next step where you want to predict the future values the series is going to take. ARIMA, short for 'Auto-Regressive Integrated Moving Average' is a class of models that

explains a given time series based on its past values, that is, its lags and the lagged forecast errors, so that equation can be used to forecast future values.

An ARIMA model is characterized by 3 terms: p, d, q where:

p is the order of the AR term

q is the order of the MA term

d is the number of differences required to make the time series stationary

First, we need to check data is stationary or not. And for Canada date set, by one time differencing, it was stationary. Then we decomposed the data set to see the new case attribute's trend, seasonal, and residuals.

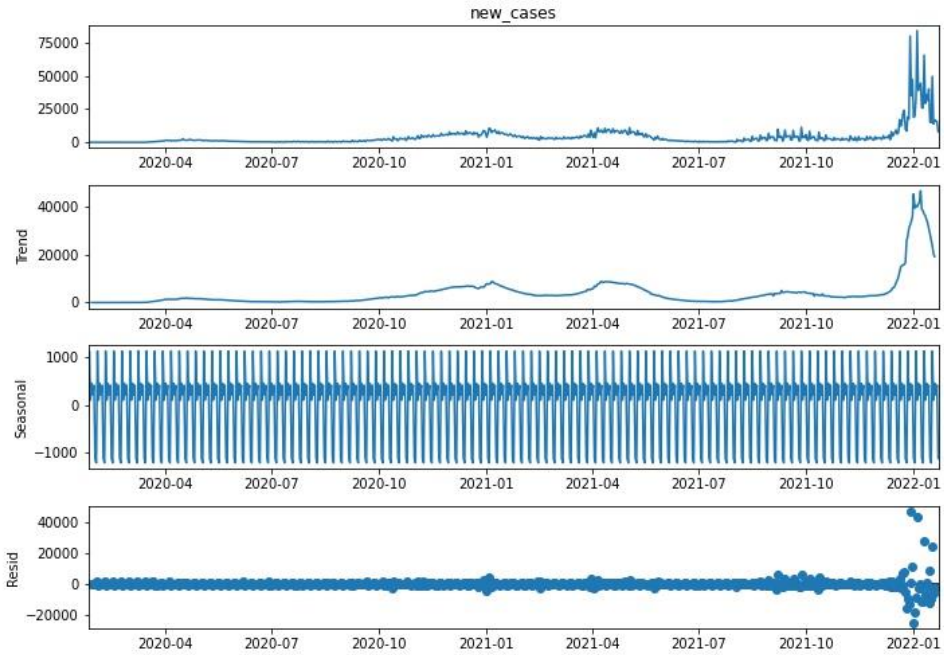


Figure 19: New cases decomposition

Here we can see that trend is significantly going up in Jan 2022, then dropping after that. The increase in the number of cases can be attributed to some of the severely affected by the new variant Omicron.

The seasonality shows us a sinusoidal trend which can be attributed to the continuous increasing trend in the number of confirmed cases. We can see some noise components in the later months of Dec 2021 and Jan 2022 which can be attributed to being poorly affected by the new variant Omicron.

The next step, the very important part is to find the optimal p, q for the ARIMA model. By applying auto_arima() function, I found the best model is ARIMA(3,1,2)(0,0,0)[0]

Then we fitted this model to our data set, the RMSE is 19,007 and the prediction was shown below:

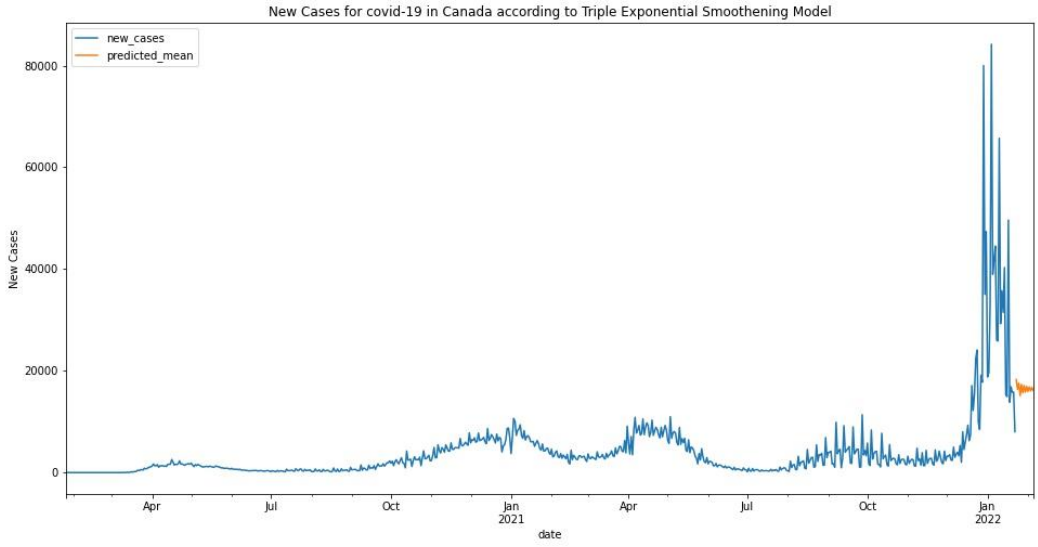


Figure 20: New cases in Canada with ARIMA model

Conclusion

After building forecasting models, we compared our models by the model scores:

Table 4: Model scores

	Model	Root Mean Squared Error
0	Polynomial Regression	15566.393572
4	ARIMA	19007.680418
1	Random Forest Regression	21910.874692
2	Double Exponential Smoothing	31833.705266
3	Triple Exponential Smoothing	34780.509307

The lower RMSE, the better model. As RMSE for ARIthe MA model is less than that of other models (except Polynomial Regression). However, considering this dataset is a time-series dataset, I will proceed ahead with the ARIMA model to predict.

By applying the ARIMA model, I can predict the number of new cases in Canada from Jan 24, 2022, to Feb 06, 2022, which was predicted to fluctuate from around 15,000 to 18,000 cases per day.

Fighting Covid-19 is a long journey. Although most people around the world have been fully vaccinated, this virus is still spreading all over the world but with less severity. Building forecasting models is essential to support the Government in planning and preparing resources to cope with the consequences of this virus on people's health and the world's economies.

References

[1] Hale, T. et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).

[2] Hasell, J. et al. A cross-country database of COVID-19 testing. *Sci. Data.* **7**, 345 (2020).

[3] *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* (Johns Hopkins University, accessed 7 April 2021) https://arcg.is/0fHmTX

[4] Sachs, J. et al. Priorities for the COVID-19 pandemic at the start of 2021: statement of the *Lancet* COVID-19 Commission. *Lancet* **397**, 947–950 (2021).

[5] Rosen, B., Waitzberg, R. & Israeli, A. Israel's rapid rollout of vaccinations for COVID-19. Is. *J. Health Policy Res.* **10**, 6 (2021).

[6] ACAPS. Report on COVID19 Government Measures Updates. 2020. Available online: **https://www.acaps.org/special-report/covid-19-government-measures-update** (accessed on 12 April 2020).

[7] (Covid-19 surveillance document https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov))

[8] Ferretti, L.; Wyman, C.; Kendall, M.; Zhao, L.; Murray, A.; Abeler-Dörner, L.; Parker, M.; Bonsall, D.; Fraser, C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **2020**, eabb6936.