Pritam shrestha DSC\_630

August 3, 2020 Prof. Fadi Alsaleem

**Executive Summary**

According to the CDC, Coronavirus (Covid\_19) is an illness caused by a virus that can spread from person to person. The virus that causes COVID-19 is a new coronavirus that has spread throughout the world. Its symptoms can range from mild to severe illness. Coronavirus disease (COVID-19) is an inflammation disease from a new virus. Basically, it affects the respiratory system of the human, and its known symptoms are fever, common cold, and cough. But it could cause severe damage to the respiratory system and breathing problems. We can quickly become infected by coming into close contact with a person who has COVID-19 positive. It can be transferred through anything which has a virus. So COVID-19 has created a global health crisis and impacting the global economy and our daily lives. Lots of people are dying each day, but some people are still thinking it is just hoax and ignoring social distancing, face mask, and other preventing activities.

There are currently no vaccines to protect us. The best way of protecting us is to avoid being exposed to the virus that causes COVID-19. So I have chosen this project to predict the death of the people of the USA. As a dataset, I downloaded data from one trustable source and implemented machine learning algorithms to predict death by COVID-19. For model implementation, I have used two models one is a linear regression model, and another one is the Holt forecasting model. After implementing this model, I have predicted the future death of the people in the USA. The primary purpose of my project is predicting death using machine learning algorithms and encouraging the public to follow social distancing and other preventive activities because people are not taking it seriously.

**Abstract**

The outbreak of coronavirus disease 2019 (COVID-19) has created a global health crisis and impacting the global economy and our daily lives. Due to the fast spread of the coronavirus, social distancing, use of hand sanitizer, frequent hand wash have come as major prevention activities. Coronavirus disease (COVID-19) is an inflammation disease from a new virus. Basically, it affects the respiratory system of the human, and its known symptoms are fever, common cold, and cough. But it could cause severe damage to the respiratory system and breathing problems. Most of the countries of the world are facing this virus because, until now, there is no known medicine to control it. Vaccines are being trialed worldwide, and hopefully, we will get it very soon.

Here, I am presenting a model that could be useful to predict the death of COVID-2019. For that, I will perform a regression model based and Holt forecasting model based on the available dataset.

**Intro/Background of the problem**

Since the first known case of COVID-19 in Hubei china, 9451515 have tested positive, 482301 have died, and 5111629 have recovered until today (6/24/2020). Gradually mortality rate is decreasing as of today due to the several actions have taken by countries. Every nation implemented different types of models to stop the spread of covid\_19, such as lockdown, social distancing, frequent hand washing, quarantine, and isolation, etc. After a few months, most of the countries are reopened based on the WHO, CDC guidelines. But it is impacting to stop the spread of COVID-19. I just watched, governor Gavin Newsome reviled that after reopening California, positive test case has increased by 69% which is warning for future and outside human activities. So COVID-19 is impacting every sector of our life, such as the economy, health, social activities, sports, schools, and so on. Due to the loneliness, growing fear, bad economic status, a lot of people is being depressed, which could lead to the mental health crisis in the world.

Hence, it is clear that it is affecting everywhere, so it is called a pandemic. Here, I am planning to use a couple of datasets related to COVID-19 and will extract some insights based on the available data. First of all, I will work on my datasets to predict death. More precisely, I will merge, join, and drop the variables. Besides that, I will apply the best fit model to answer my hypothetical questions and to solve the problems.

**Used resources:**

Resources are something that is used to complete the project. Here I have used the following resources to complete this project.

1. Data source:
2. Other technical resources:

Anaconda, Jupyter Notebook, python libraries, etc.

**Data Source**

Data is distinct pieces of information, usually formatted in a special way. It can be used to analyze and visualize implementing different machine learning algorithms. Data is the most valuable thing in data science because everything relies on it. Bad data could lead to bad output and project failure. So data finding is a major milestone of the project. Finally, I found data and downloaded data as a CSV file format from the following source. The source is given below.

**Source:** [**https://github.com/datasets/covid-19/tree/master/data**](https://github.com/datasets/covid-19/tree/master/data)

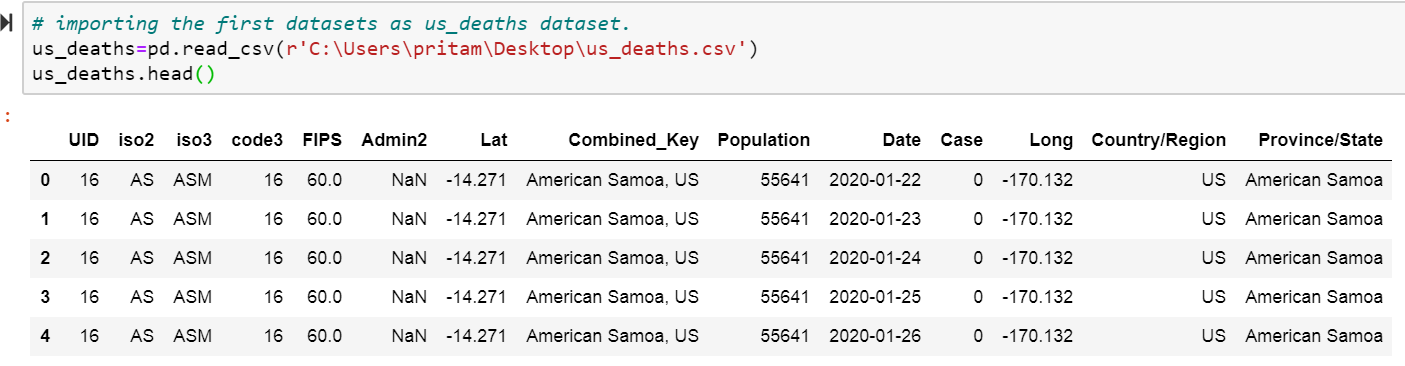
**Data understanding:**

Understanding the problem statement is the best way of finding a solution. To understand the problem, we need to find the problem first. Problem finding is not easy because it could be related to anything. So first of all, we need to categories the type of problem, whether it is solvable or not. If it is solvable, then they need to pay attention to it. Likewise, I figured out the problem statement and started to find possible factors being responsible for prediction.

The structure of the data is provided as an image. Basically, I have downloaded the data from GitHub and saved it in a CSV file. Later I read that data using the pandas library of python for further analysis.

**Structure of the data:**

Before implementing the data for modeling, we have to go and analyze data for better understanding and perform exploratory data analysis. Raw data is not always ready for model implementation, so data preparation is another essential part of the data science project. Here is the structure of the downloaded data.



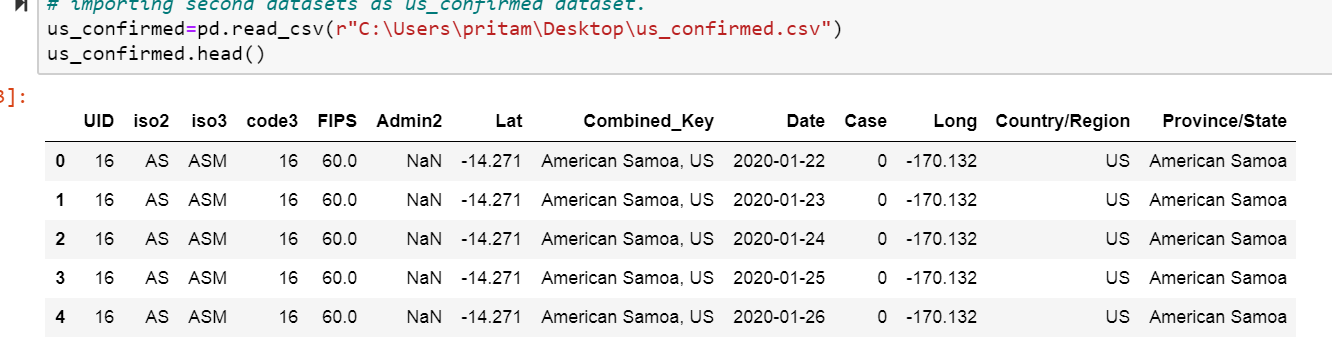


Fig: Structure of the datasets

Where I have used two datasets, each of them have 570675 records with 13 and 14 columns. This data are not ready for the modeling, so I will perform exploratory data analysis for data preparation.

**Exploratory data analysis:**

Exploratory data analysis is an approach to analyzing the dataset by summarizing its main characteristics implementing visual methods. The primary purpose of EDA is to see what the data can tell us beyond the formal modeling or hypothesis testing task. Data is not always ready for further analysis, so I did some exploratory data analysis using python.

1. Replaced header: I replaced the header of the downloaded data to make it more readable and understandable.
2. Fixed casing and inconsistent data: When I checked the structure of the data, then I have found some missing values in multiple columns so filled with 'Zero' value.

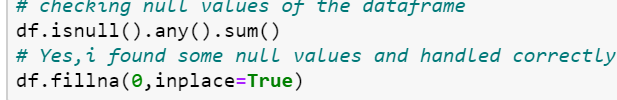
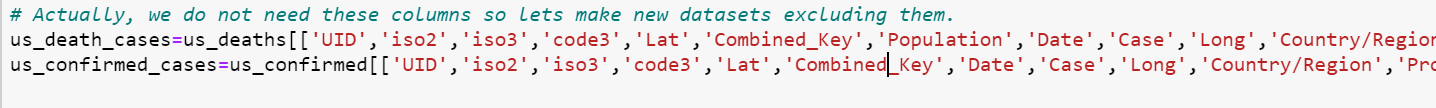
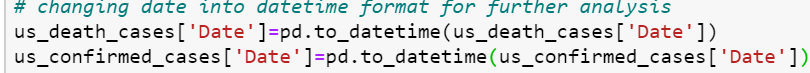


Fig: handling missing data.

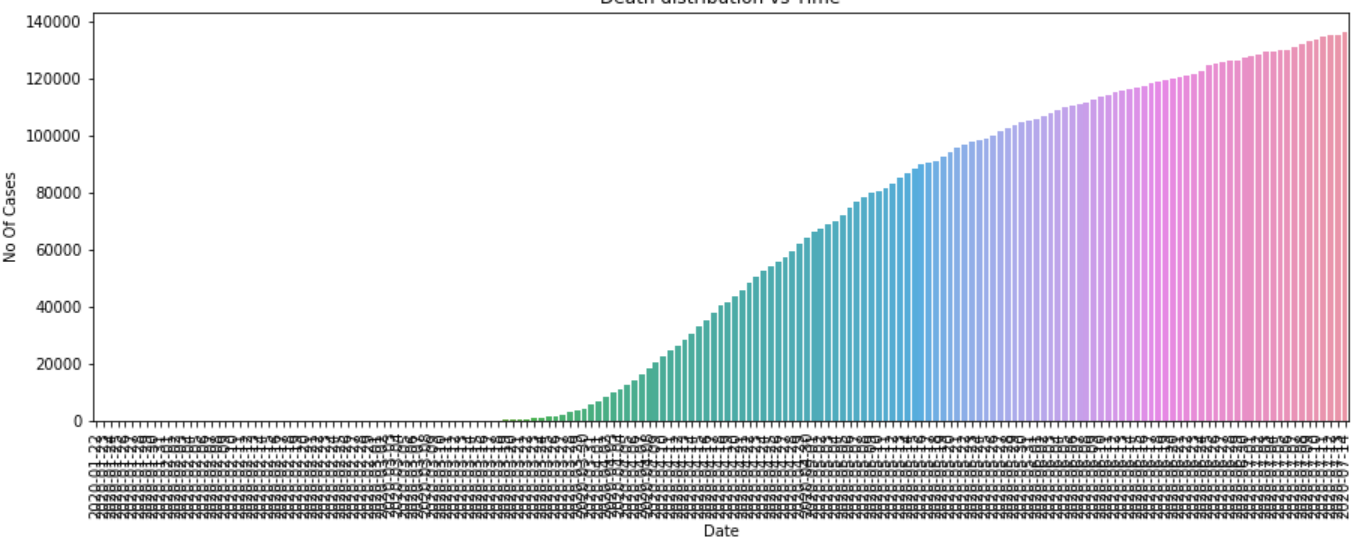
1. Checked outliers and bad data: Sometimes we get outlier. Outlier means extreme values that fall a long way outside of the other observations. Finding outliers and proper handling of the outlier is very important; otherwise, it might lead the wrong output or result might bias to the higher value. But in my dataset, I didn't find any outliers.
2. Data cleansing or cleaning: It is a process of finding and correcting corrupt or inaccurate records from the dataset. I also dropped unnecessary data and made a more cleaned data frame.

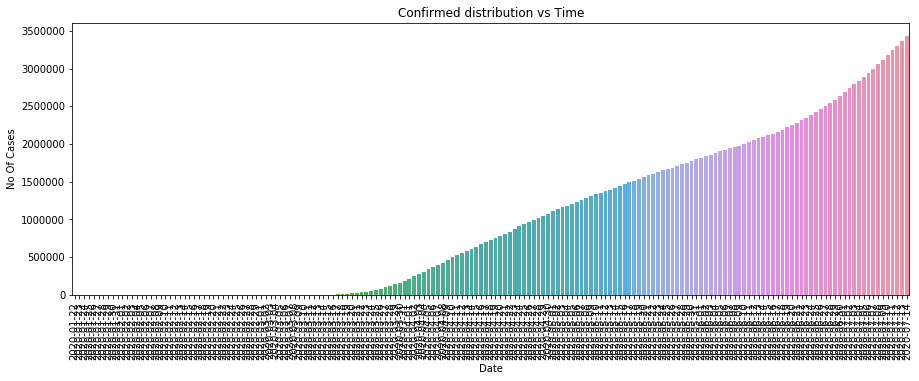


1. Data transformation: It is a process of converting data from one format to another format that makes data more readable and understandable. For this project, I have converted all my raw data into the required data format.

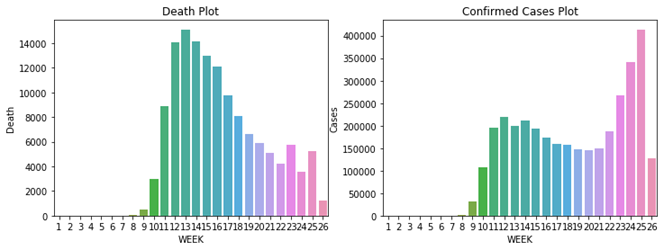


1. Visualization: It is a pictorial representation of data. It helps us to analyze the data, so I plotted a couple of plots using matplotlib package.





Weekly trend:



From the above figure, I have discovered some insights about the type of data and its structure. It means I have seen a trend of the confirmed cases and deaths of the people based on time period.

**Feature Selection:**

Feature selection is the process of selecting a subset of relevant features or variables for modeling or use in the model construction. It is also called variable selection. In my case, I have 14 variables or features; some of them are not correlated to the target variable, so I did not include them as my input for modeling. I only chose the following variables as my input variables and target variables.

1. Input features: date, city, confirmed cases, etc.
2. Target variable: death

In the process of feature selection, I did not include some of them because they are not correlated to the target variable.

My dataset considered as less dimensional dataset because I have only 14 features. If I had high dimensional data, I would have used dimension reduction techniques using either SVD (singular value decomposition) or PCA (Principle component analysis).

**Data Modeling:**

Data modeling means training machine learning algorithms to predict the target from the features. For the implementation of the model, we have to split out data into two data sets, such as training dataset and testing dataset, so I have split my input and target dataset into two datasets.

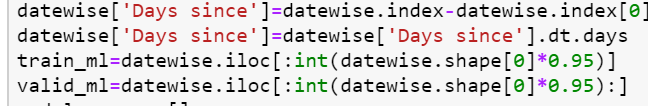


Fig: data splitting

1. Training dataset: It contains 95% data for training purposes.
2. Testing dataset. It contains 5% data for testing purposes.

**Model selection:**

Now input and target data are ready to train the model and test the model, so I have selected two models for data modeling because of the linear relationship between input and output variables.

1. Linear regression:

Linear regression is a supervised learning algorithm used to predict the target variable. This is one of the most widely known modeling techniques in data science or statistics. In simple linear regression, a single independent variable is used to predict the value of the independent variable. The equation of the simple linear regression is given below.

Y=mx+c

Where Y= dependent variable

M=slope

X=independent variable.

C=intercept

It is easy to calculation and finds the relationship between the variable because it uses only one dependent variable.

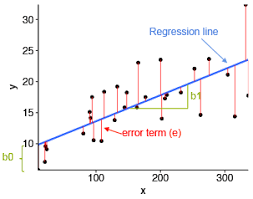
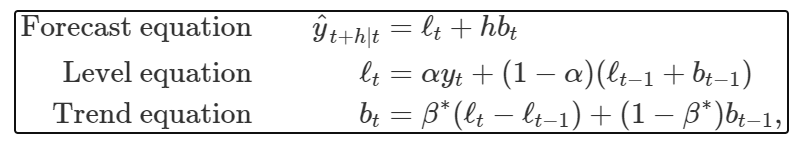


Fig: Linear regression

Source: <https://nextjournal.com/intelrefinery/simple-linear-regression>

1. Holt forecasting model:

It is another model for the forecast. Holt's two-parameter model, it is also called as linear, exponential smoothing is a popular smoothing model for forecasting data with the trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for the previous period's trend. The second equation updates the trend over time. The third equation is used to generate the final forecast.

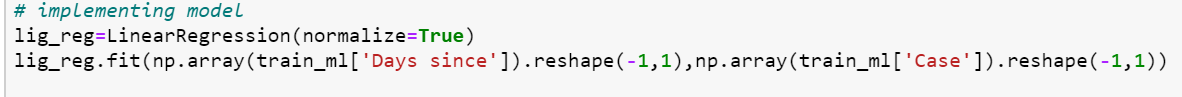


Where ℓt denotes an estimate of the level of the series at time t, bt denotes an estimate of the trend (slope), and alpha is the smoothing parameter for level 0<= alpha<=1 and beta is another smoothing parameter for 0<=beta<=1.

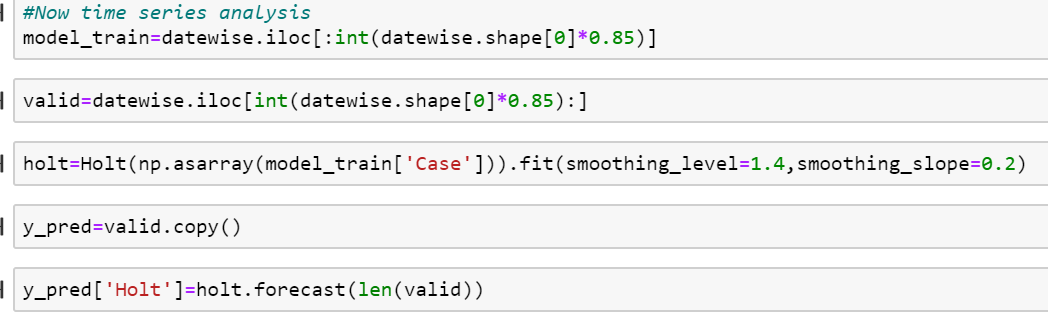
Model implementation:

I have chosen two models and implemented them to predict and forecast the result.

Linear regression model implementation



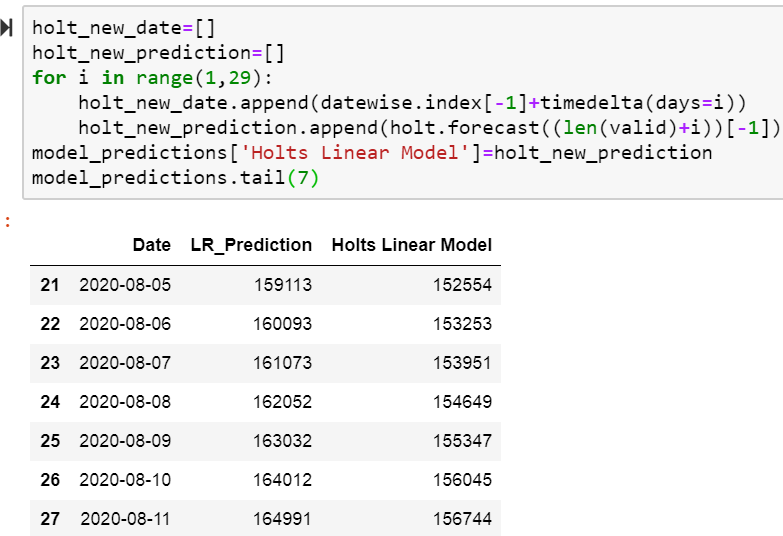
Holt's model implementation



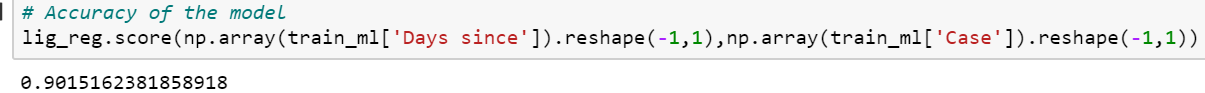
So to predict the death of COVID-19 patients in the USA, I have used these two above mentioned models.

**Result:**

Finally, I have implemented the two models and predicted and forecasted the death of the coming one week. The output of the model is given below.



Accuracy is the way of measuring how much model is predicting the accurate value of the target variable. In my regression model, I am getting 90% accuracy. If I use another model, I might get more accuracy.



Here I am displaying only one week to make it easy to read and understand, but we can predict as much as we want. According to the result, today's actual death is 160,998, and my prediction is 159113. It means a linear model giving a better result. But most of the cases, Holt model forecasts better.

**Discussion/Conclusion**

Coronavirus disease (COVID-19) is an inflammation disease from a new virus and spreading all over the world, contacting its surrounding. It means it can spread through close contact of an infected person, food, and other various materials used by an infected person. So, until now, there are no vaccines and medicine have invented. Due to this bitter truth, we have only one method to stop it is social distancing. There are so many biomedical companies are working rapidly to invent vaccines such as Gillard, a foster city-based company's vaccine Remdesivir is being used for trial and several vaccines are also being used for the third phase of the trial. Besides that so, many data scientists are also working to discover the insights about it. Recently, I have seen my professor Fadi Alsaleem has also invented a new concept to minimize the coronavirus spread using kinsa thermometers.

To complete this project was very challenging. Because I have used two different datasets, and I have completed each step of exploratory data analysis, such as data cleaning, transformation, duplicate finding, data conversions, and best fit model implementation. Besides that, I had to use two different programming languages as minimum requirements.

So, I have used exploratory data analysis and data visualization using R and completed modeling and plotting using python programming languages. While working in this project, I have faced a couple of problems but solved under the supervision of my professor Fadi Alsaleem. Indeed, I have completed this project with expected output.

**Acknowledgment:**

I can't express enough thanks to my professor Fadi Alsaleem for his continuous support and encouragement. The completion of this project could not have been accomplished without the support of my classmates and beautiful daughter, Prisumsa.

Thanks to my wife Suchan as well. The countless times you kept the children during our hectic schedule will not be forgotten.

**References:**

[1] <https://link.springer.com/article/10.1007/s00477-020-01827-8>

[2] <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv>

[3] <https://www.frontiersin.org/research-topics/13638/coronavirus-disease-covid-19-the-impact-and-role-of-mass-media-during-the-pandemic>

[4] <https://otexts.com/fpp2/holt.html>

[5] <https://www.worldometers.info/coronavirus/country/us/>

[6] <https://machinelearningmastery.com/linear-regression-for-machine-learning/>