DSC630 Final Project

Carla Harrell

Date: 05/30/20

**Executive Summary**

COVID-19 is a newly discovered infectious disease that can cause mild to severe respiratory illness. It appears those with underlying conditions such as diabetes, hypertension, cardiovascular disease, and other serious illness are more at risk. COVID-19 is an RNA virus which belongs to the family coronavirus. It was initially emerged in China around December 2019. The wide spread of this virus has become a pandemic.

The COVID-19 pandemic causing severe respiratory syndrome has many scrambling all over the world trying to contain the spread of the virus. Many people are requested to stay at home if they do not have severe symptoms while others are walking around asymptomatic who can spread the virus. There are many challenges presented due to lack of testing and reporting of results.

The goal of this project is to analyze the daily cases of corona virus based on confirmed cases, deaths and other parameters and predict for future days number of confirmed cases and spread factor. Spread factor is like a growth measure of this pandemic. It is calculated based on the positive increase of cases each day compared to the previous day. If the spread factor goes down, the curve of this pandemic will flatten and go down and we will witness less spread everyday of this virus.

Machine learning and statistical methods like linear regression and polynomial regression were used to build our algorithms to predict for future days. There was an attempt to test a basic hypothesis using second algorithm (Model 2) – Whether population, median age and density of a state or region in United States affects the confirmed cases. Even though there were no strong correlations, population does matter to some extent in the spread of the virus.

**Problem Statement:**

**Scope:**

This scope of this project will predict the importance to flattening the curve by reducing the spread.

**METHODS:**

**Technical Approach:**

My goal is to discover when the curve will be flattened. In order to proceed with this project, I will use two models to make predictions. The first model will use polynomial regression to predict the spread factor which flattens the curve and the second model is based on the population, median age, and deaths in region/province/country to predict the number of positive cases.

**Data Sources:**

* **Population\_by\_country\_2020.csv**

<https://www.kaggle.com/tanuprabhu/population-by-country-2020>

* **states\_code.csv**

<https://aruljohn.com/usstates/>

* **us\_states\_median\_.csv**

<https://worldpopulationreview.com/states/median-age-by-state/>

* **us\_states\_population\_.csv**

<https://www.worldometers.info/world-population/us-population/>

* **us\_covid19\_daily.csv**

<https://www.kaggle.com/sudalairajkumar/covid19-in-usa>

* **us\_states\_covid19\_daily.csv**

<https://www.kaggle.com/sudalairajkumar/covid19-in-usa>

* <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series>

**Exploratory Data Analysis**

Exploratory data analysis was used to analyze the current situation in United States based upon positive cases, deaths, spread factor, negative cases, number of increased deaths and other factors.

Below are the graphs and description about those graphs:

A close up of a logo

Description automatically generated

The number of positive cases by the number of days increases with testing.

A close up of a mans face

Description automatically generated

The number of negative cases by the number of days increases with testing.

A close up of a mans face

Description automatically generated

As more people got infected, there is an increase in death due to this virus in US since the day of outbreak.

A close up of a device

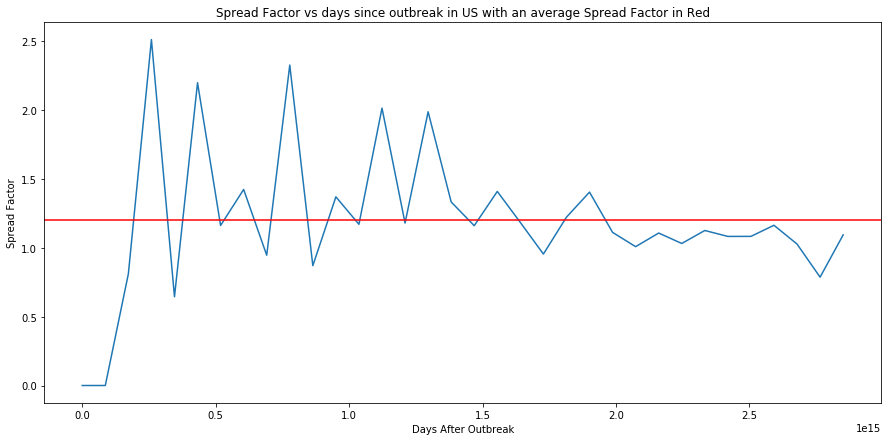
Description automatically generated

This above graph is an important one as it compares death increase from previous day since the outbreak. Although the death increased per day is getting higher, there were times and currently that it is flattening a bit. It will slow down in future as we will demonstrate using our first model.

**Calculation of Spread Factor**

Death Increase rate does not provide a proper indication of the spread of this outbreak. The death may depend upon other relating factors such as heart conditions, diabetes, etc. Those datasets were not available at the time of this project. To understand the spread of outbreak, the growth factor or spread factor of Covid-19 in United States must be calculated. The Spread Factor based on positive cases Increase each day compared to previous day. A trigger alert was added in the graph figure below to see if the spread is greater than a number to see if Covid-19 is spreading. The following formulas are used to calculate the spread factor and the average spread factor in the US.

* Spread Factor = Positive Cases for the Current Day / Positive Cases for the Previous Day
* Average Spread Factor for US¶= np.mean(us\_data.Spread\_factor)



The spread factor graph shows the spread spike initially. After more testing and isolation, it shows a gradual decrease with less spikes. Future days was added to Model 1 to predict the spread using polynomial regression.

**Analyzing State Data**

The state data for the US is analyzed to see if there are any interesting insights and patterns across states.

Us\_states\_covid19\_daily.csv was used. The file date was corrected to format yymmdd and a new column no\_days which was the days counted for each record since the first outbreak.

A screenshot of a cell phone

Description automatically generated

The above graph gives a comparative study of few major states based on number of confirmed cases since outbreak. New York has shown exponential increases in the cases compared to any other state. That’s why for model 1 with polynomial regression, we will be using New York State data to fit our model due to availability of more data.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

This above graph is a comparison of the states based on number of the deaths. New York has an exponential rise in deaths.

A view of a city

Description automatically generated

Above graph gives a bar chart of all the confirmed cases across the states in US.

**RESULTS:**

There are two models that were created. In the first model, polynomial regression including linear regression on New York was used.

**Model 1**

Model 1 is a Polynomial Regression model the following algorithm steps are used:

1. Selected the entire data for NY state
2. Calculate the spread factor for NY
3. Handle the missing values and NANs in NY data
4. Import Polynomial Regression and related libraries in Python
5. Create a regression model based on Spread Factor and Confirmed Cases along with days since outbreak factor.
6. Based on user input for degrees, the model can be transformed from linear regression to polynomial regression
7. User input decides the number of future days for which the prediction must be made.

Below is the output graph and the result of the model for next 5 future days and with degrees 2 for New York State.

Regression Degrees: 2

Spread\_factor Pred\_positiveIncrease Pred\_Confirmed

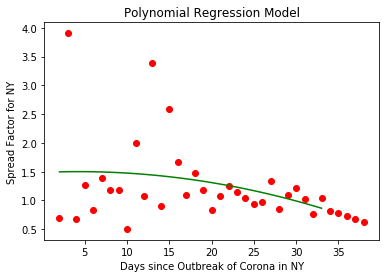
0 0.820594 7104 137793

1 0.774053 5498 143291

2 0.725973 3991 147282

3 0.676354 2699 149981

4 0.625196 1687 151668



Regression Degrees: 3

Spread\_factor Pred\_positiveIncrease Pred\_Confirmed

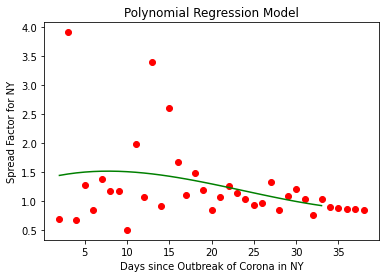
0 0.901578 7805 138494

1 0.884485 6903 145397

2 0.870184 6006 151403

3 0.858923 5158 156561

4 0.850947 4389 160950



Regression Degrees: 4

Spread\_factor Pred\_positiveIncrease Pred\_Confirmed

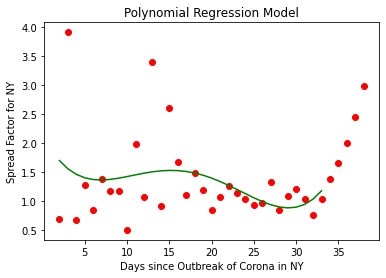
0 1.379318 11942 142631

1 1.651766 19725 162356

2 2.003647 39521 201877

3 2.445430 96645 298522

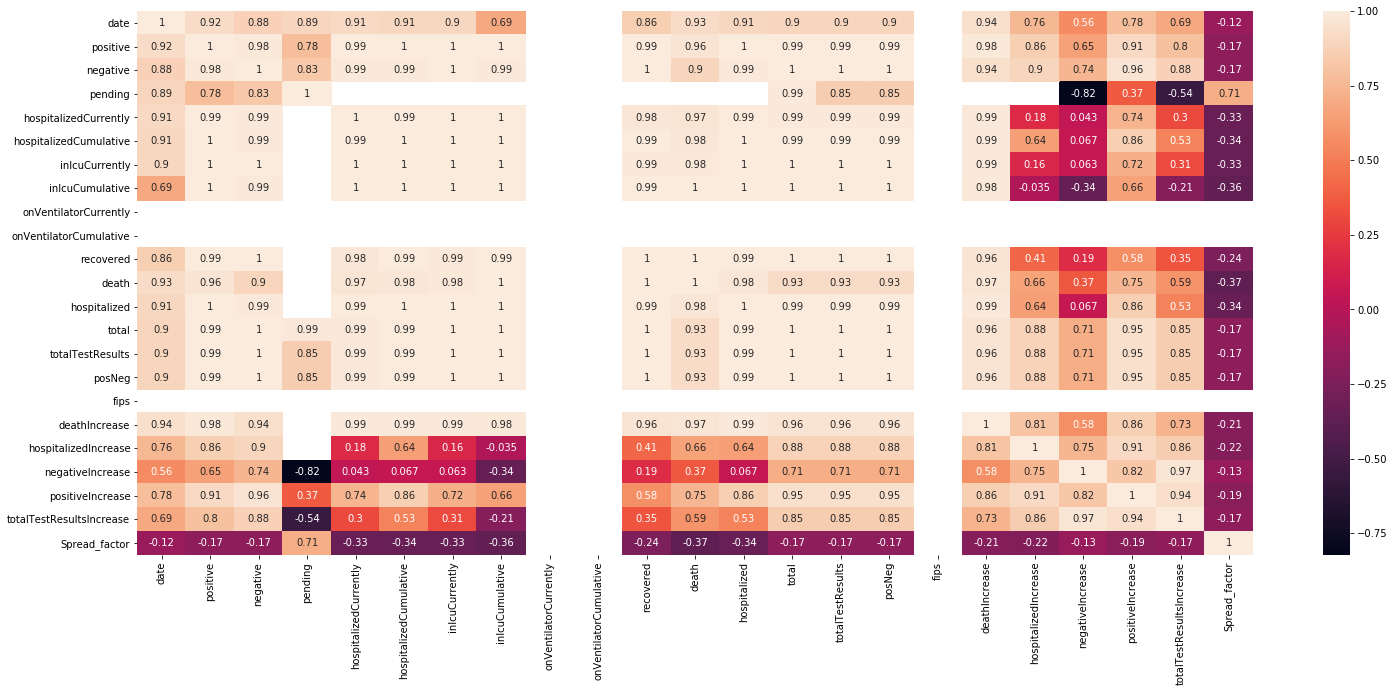
4 2.988148 288789 587311



**Model 2**

Before looking at the model 2,a correlation matrix is used to analyze the data based upon the correlation coefficient to see the relations between different attributes.

Below is the correlation matrix of the New York Data:

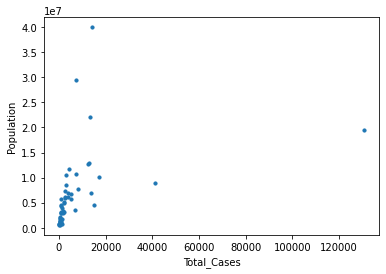


The above matrix gives an idea of how different attributes are correlated to each other.

## **Linear Regression Model 2**

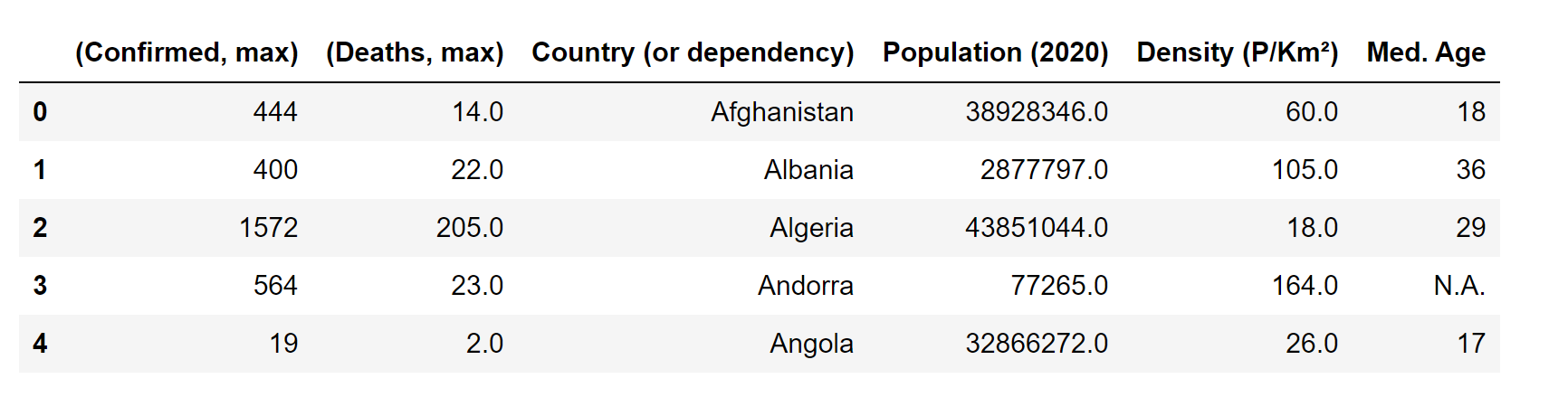
The Model 2 is kind of a hypothesis testing to verify whether population and median age of a region/state or province matters in the spread of this virus. As per the results of this regression analysis, it does not matter to a significant extent. I collected world’s stats for countries, their population, median age, and number of positive cases and used this to train a linear regression model. And I collected similar stats for states in US to use them as my validation data. The second model use linear regression and is based upon on the following assumptions:

1. Higher the population and population density of a country or region or state, higher the chances of spread and positive cases due to community spreading as can be seen in the image below.
2. Age is related to the spread of virus as old age people are at the risk of getting the virus compared to young age due to immune system.
3. Deaths are directly related to number of positive cases for each region or a country.



**Testing and Evaluation**

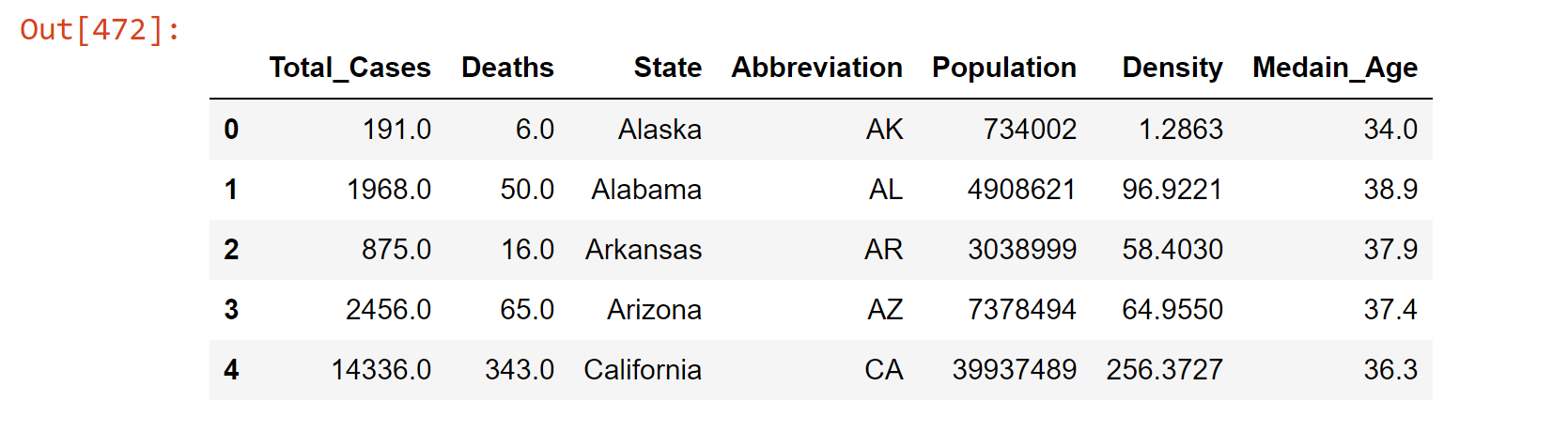
Different data files (mentioned in sources and references) to get a train data in the following shape.



Prediction will be performed based upon on all the other attributes in the above table. The model will be trained based upon on the data mentioned above for around 170 countries and tested our linear model on data for 51 states in USA. Multiple merging

and joins as each column mentioned above was fetched from different data sources.

The test data looked like below:



**CONCLUSION:**

Population is slightly correlated to total cases. However, this virus spreads irrespective of the population density and median age of the region. It is now in community spread phase and therefore, these factors are not able to quantify the spread or growth. This was the whole purpose of Model 2.

Through Model 1, I am already predicting the spread factor for next 5 (or any number) of future days and predicting the number of cases for these future days. (this can be seen in the graphs and results for Model 1)

**RMSE**

RMSE increases as you move from degrees 2 to degrees 3 and 4

1. 2 Degrees

* Root Mean Squared Value = 0.3638
* Mean Absolute Percentage Error= 32.45%
* Accuracy = 67.55%

1. 3 Degrees

* Root Mean Squared Value = 0.6328
* Mean Absolute Percentage Error= 59.61%
* Accuracy = 67.55%

1. 4 Degree

* Root Mean Squared Value = 0.3725
* Mean Absolute Percentage Error= 31.01%
* Accuracy = 68.99%

Implementation of a linear regression model to predict the total cases based on the given factors. However, these are not the only major factors which estimate the total cases in a region and the model will not produce accurate results. The second model is based upon the population, median age, and deaths in regions/province/country to predict positive cases. The model was trained and tested for linear regression and evaluated using RMSE values.

States Included:

print(metrics.mean\_absolute\_error(y\_test,y\_pred))

6406.811598857804

print(np.sqrt(metrics.mean\_squared\_error(y\_test,y\_pred)))

99596484.08658408

States Removed:

print(metrics.mean\_absolute\_error(y\_test,y\_pred))

12817.768496186587

print(np.sqrt(metrics.mean\_squared\_error(y\_test,y\_pred)))

1362885423.6272116

**Acknowledgements:**

The raw data pulled and arranged by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from the following resources:

Reference:

* <https://www.kaggle.com/tanuprabhu/population-by-country-2020>
* <https://aruljohn.com/usstates/>

* <https://worldpopulationreview.com/states/median-age-by-state/>
* <https://www.worldometers.info/world-population/us-population/>
* <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>
* <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>
* <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series>
* World Health Organization (WHO): <https://www.who.int/>
* DXY.cn. Pneumonia. 2020. <http://3g.dxy.cn/newh5/view/pneumonia>.
* BNO News: <https://bnonews.com/index.php/2020/02/the-latest-coronavirus-cases/>
* National Health Commission of the People’s Republic of China (NHC): http:://[www.nhc.gov.cn/xcs/yqtb/list\_gzbd.shtml](http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml)
* China CDC (CCDC): http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm
* Hong Kong Department of Health: <https://www.chp.gov.hk/en/features/102465.html>
* Macau Government: <https://www.ssm.gov.mo/portal/>
* Taiwan CDC: <https://sites.google.com/cdc.gov.tw/2019ncov/taiwan?authuser=0>
* US CDC: <https://www.cdc.gov/coronavirus/2019-ncov/index.html>
* Government of Canada: <https://www.canada.ca/en/public-health/services/diseases/coronavirus.html>
* Australia Government Department of Health: <https://www.health.gov.au/news/coronavirus-update-at-a-glance>
* European Centre for Disease Prevention and Control (ECDC): <https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases>