COVID-19 Data Analysis

Md Reazul Islam
Department of Computer Science
Georgia State University
mislam20@student.gsu.edu

Ricardo Villarreal
Department of Computer Science
Georgia State University
Rvillarreal2@student.gsu.edu

Tomy Tran
Department of Computer Science
Georgia State University
ttran134@student.gsu.edu

***Abstract*— We are living in the "information age"is a popular saying; however, we are actu-ally living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business society, science and engineer-ing, medicine, and almost every other aspect of daily life. This explosive growth of avail-able data volume is a result of the comput-erization of our society and the fast devel-opment of powerful data collection and storage tools. Businesses worldwide generate gi-gantic data sets, including sales transactions, stock trading records, product descriptions,sales promotions, company profiles and per-formance, and customer feedback. For example, large stores, such as Walmart, handle hundreds of millions of transactions per week at thousands of branches around the world.Scientific and engineering practices generate high orders of petabytes of data in a con-tinuous manner, from remote sensing, pro-cess measuring, scientific experiments, system performance, engineering observations,and environment surveillance. Data mining also gives the way into utilizing the machine learning models and can be extended to deep learning methods such as CNN, AE [1] and in real life scenarios such as twitter analysis[2].**

Keywords—regression, clustering, k-means, data

I. Introduction

The recent global pandemic known as COVID-19 has been causing mass hysteria, city wide shutdowns and a high death toll. One of the scariest things about the virus is that there is little known about how to prevent and treat it. In this paper we hope to use linear regression and clustering to help identify the optimal conditions for the spread of COVID-19. We believe that Northern states, more specifically highly populated ones, will suffer the most outbreaks in winter. This belief primarily stems from knowing that the cold weakens the human immune system. We would also like to see if there is evidence that the virus can survive longer in colder environments. One good counterpoint to our hypothesis is that people tend to be less physically active and social in the winter time. One of the best safety practices in a pandemic is to quarantine yourself, which more people tend to naturally do in the winter time. We are curious if this will affect our hypothesis. By providing this information we hope to help solve the mysteries of COVID-19, create awareness and promote proper safety measures.

II. Related Work

A. *India Covid-19 Cases*

While our efforts and goals were focused on the states of the US. Others have been researching into other countries, In one particular case, their priorities were focused on the rise of cases in India. Their Research and work can be found on https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/ . How their work relates to ours, is that they were able to use linear regression to produce some visualization in the rise of affected cases. In their first case, they were able to produce their results by splitting the data into a 80% and 20% training set. Doing this allowed the model to produce predictions of active cases based on the daily active cases.

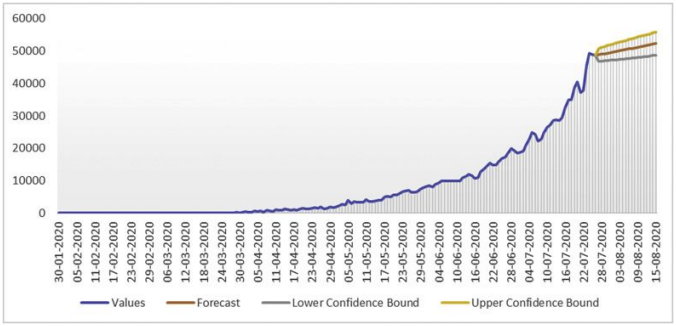
Table 4

Values obtained after training with multiple linear regression prediction model.

Data set	Intercept	Coefficient ($\beta_0, \beta_1, \beta_2$)	Score (R^2)	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Odisha	−21.972463	[0.78035157, −0.45427124, 13.57489689]	**0.9985**	45.5734759e-06	3826.5455326712213e-06	61.85907801342679e-06
India	−3.2596290111541 75e-09	[-1,1,1]	**1.0**	2.6540334374658414e-09	7.7358572080180 85e-18	2.781340901079564 7e-09

Bold indicates R-squared or Coefficient of Multiple Determination.

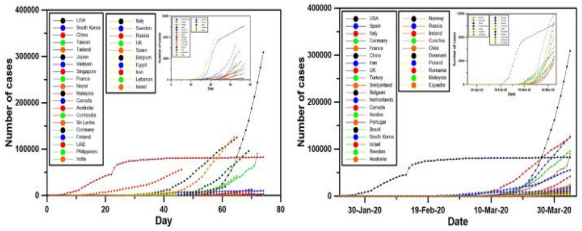
In the second case, instead of using simple linear regression for the first case, they use multiple linear aggression and just like the last case, they split the data into a training and test set. The inputs are daily positive cases, recovered and deceased cases to predict the daily number of active cases. They were able to obtain a relation between the above variables using the correlation factor from their correlation table of India in daily Covid-19 cases.



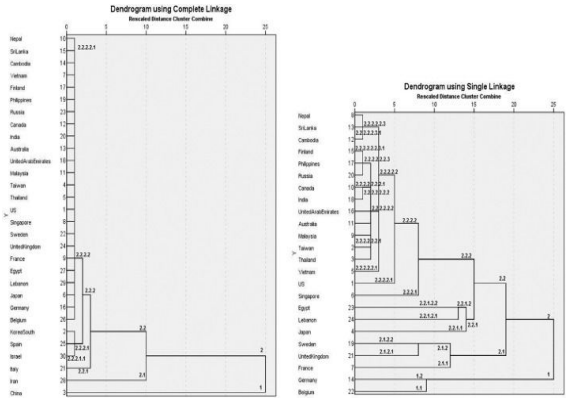
B. *Hierarchical analysis*

In this data analysis, this group was able to apply clustering to their Covid-19 data . Just like our goal of applying clustering to determine the events that lead to

outbreaks in specific areas in countries, this group applies clustering in hopes of finding possible causes of the Covid-19 epidemic in many countries, along with their many other goals. Their analysis and research can be found on this website https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7258836/ .



This figure shows us the evolution of cases on a daily basis. The left graph tells us the number of cases rising in the thirty countries with the oldest cases, starting from the first day, and the right graph tells us the thirty countries with the most cases.



In the image above, the image shows us the advancements of cases and the clustering of countries on a daily basis. The Clustering of the second image is obtained using the daily number of Covid-19 cases in the first image.

III. Dataset

The dataset is publicly available at https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36. The dataset contains 15,800 records and 15 attributes. We chose this dataset because of how much we can relate to the current global pandemic that is covid-19. The dataset contains dates, number of cases, states in which the outbreak was confirmed, the number of deaths, etc, which contains both numeric and categorical variables and is useful for expressing quantitative values and information.

The dataset contains a few empty values. Since there is a huge number of records, we will remove the records that contain null values as null values can negatively affect the results of the model. The dataset should then be able to be analyzed and prepared for data mining.

IV. Algorithms

A. Linear Regression

There are several ways linear regression can be useful in this dataset. We plan to perform linear regression of time with infections and time with deaths for all 50 states individually. Thus, we can predict how the infections will spread in the future and why the death rate may rise. We have dates associated with each record, so we can perform linear regression on the data based on different seasons such as fall, winter, summer, etc. This may also help in clustering.

B. Clustering

Clustering can help us determine where and when the outbreaks occurred through data analysis and pattern recognition. Using clustering we can also sort the states regionally, which may help us understand if there were any events that may have led to a mass outbreak in one area of the country. Through the use of organizing data and processing it, we should be able to classify when and where the covid outbreaks occurred. Thus the method should give us a more detailed report on what increases occurrences of the outbreak from a spatiotemporal perspective.

C. Principal Component Analysis(PCA)

Since we have a lot of variables, it will be difficult to visualize them or perform clustering algorithms. To make visualization easier, we can use principal component analysis to reduce the number of attributes to 2 while retaining most of the information. We can use the elbow method to ensure that we will not lose a lot of information. Since most of the variables are of similar scale we will probably get good results without scaling the variables. However, we will implement the algorithms with both standardization and normalization.

V. Cleaning and Standardizing the Data

A. Importing the DataSet

corona=pd.read_csv('United_States_COVID-19_Cases_and_Deaths_by_State_over_Time.csv'). Here the data is in csv format and then imported into the python for us to analyze the data.

B. Missing Values

The first problem of our approach to the dataset was filling in missing values. Our first task when looking at the dataset was looking to see if we would see any missing values, which will be represented with NaN or None. To our surprise the data contained NaN values and the question that we wanted to ask was if the values were missing because it wasn't recorded or because it didn't exist? Most of the data was filled with NaN values and the solution we took to fix this problem was filling in the NaN values with 0's.

C. Splitting the dataset

We imported libraries from sklearn for our train_test split. Our line of code X_train, X_test, y_train, y_test = train_test_split(dataframe[['Cases']], dataframe[['Deaths']], test_size = 0.25), split the data in 75:25 ratio.. 75% of the data

would be used for training the model while 25% will be used for testing the model that is constructed out of it.We have defined and passed the following parameters which is x and y that we had previously defined test_size: This is set 0.25 thus defining the test size will be 25% of the dataset. The Concerns that were left with us were either having parameters having greater variance if we have less training data or performance statistics would have greater variance.

D.. Standardizing

Our goal now at this point, was to carry through the preprocessing process. One of these objectives was to change the raw features of the data into a more suitable rendering depiction and making sure that the mean of each feature is 0. The idea is simple, but the problems we would be facing is that each variable would be measured at different scales and would not be uniform in terms of when contributing to how the machine learning model will generalize data, thus creating a bias. There were many approaches we could take such as removing the mean value of each feature and dividing the non-constant features with the standard deviation to scale it.

VI. Process

A. Applying Linear Regression

We used two different sets in the application of linear regression. The first one compared the total deaths to the total cases. We believe this is important because it will show the variance of the mortality rate for all people in the United States. We assume that there will be a low level of variance between the increase of the total deaths and total cases. We then plotted a more specific graph to help in the analysis of each State's death to case ratio over time. This is important because it helps us determine the factors at a specific period in time, that lead to a high rise in deaths in a specific area and gives us the month with the highest rate of death. The coefficient of determination will also help determine if the data is best fit for linear regression or mean analysis. The last set we analyzed is the new deaths vs new cases. This set is important because it shows us when outbreaks occur. This can be used to help link past events as a factor that plays a role in the increased infection and death rate of COVID-19.

B. Applying Clustering

We intend to implement clustering in an attempt to visualize any obvious separation between data points. Prior to implementing clustering, we must determine an appropriate amount of centroids. The elbow method will be used to determine the number of clusters.

C. Applying Polynomial Regression

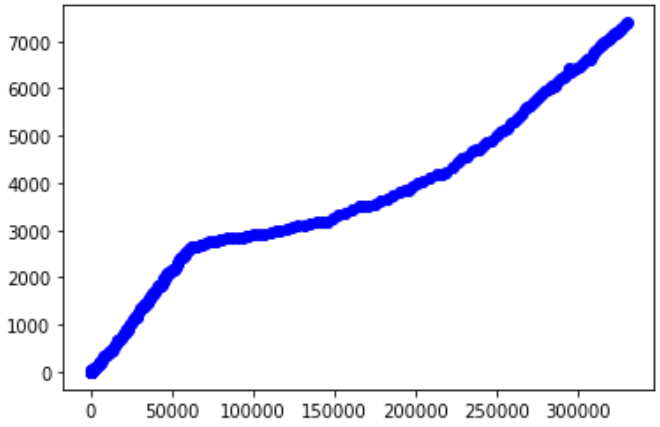
Polynomial Regression is important with a dataset like this because each state has different times when they are affected the worse hy COVID-19. This causes a lot of variance when attempting to fit a regression line. Polynomial regression will give us a higher R^2 score which will help with predicting test data.

D. Applying Principal Component Analysis

Principal Component Analysis provides us with weights to help determine which variables are the most useful. We can reduce the complexity of our analysis by focus on the features that carry the most weight. We assume that the date, total cases and total deaths will be weighted the highest.

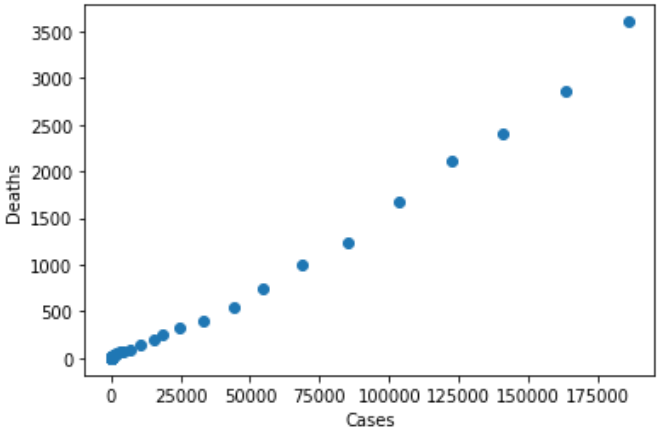
VII. Visualizations

In these Visualizations, we use Georgia as an example state out of all the many states we could've chosen from, for our visualization.

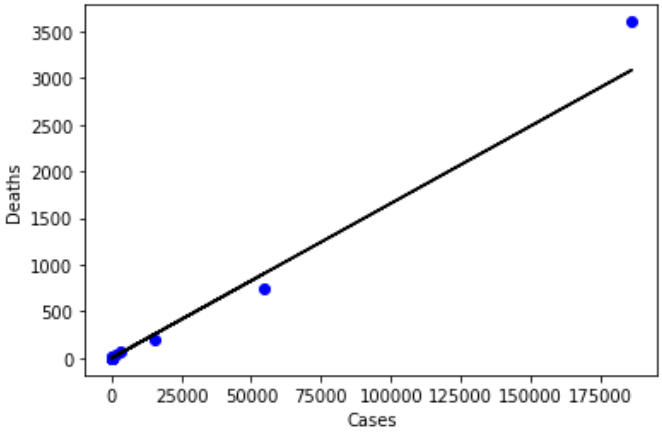


"Georgia"

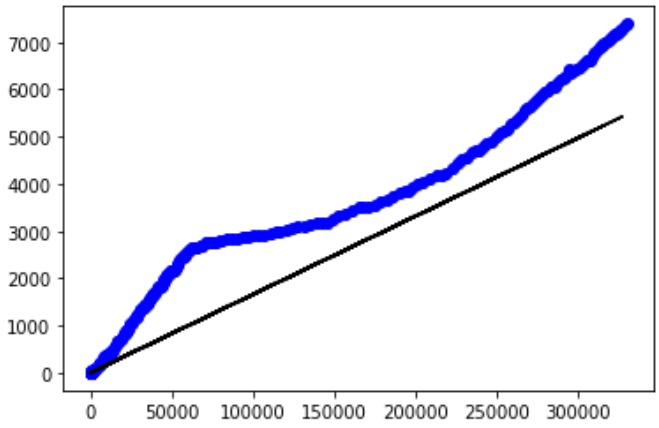
This figure shows us the Total Cases vs Total Deaths in the state of Georgia



"Scatter plot of total number of cases and Deaths"

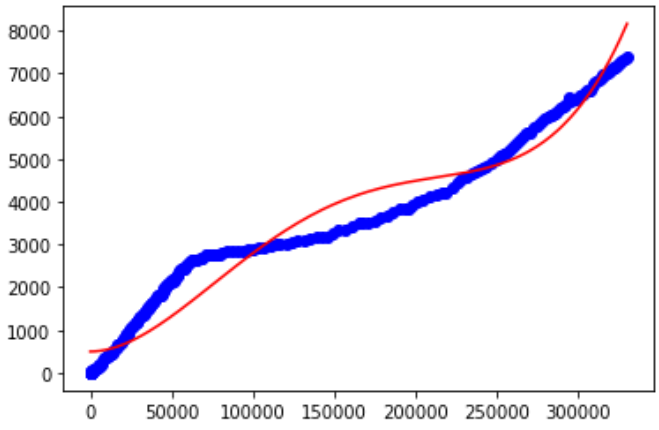


"Georgia"



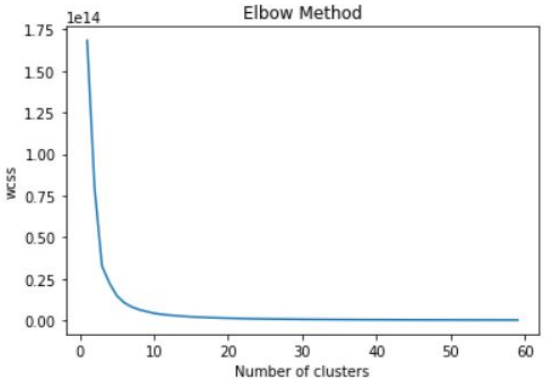
"Georgia"

We used the state_wise Linear Regression for this model and calculated the R_squared score of the state of Georgia . As you can see, Georgia wasn't able to fit the data well on a straight line compared to other states.

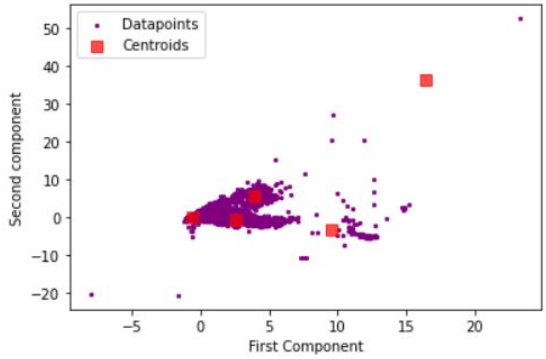


"Georgia"

In this figure, we use polynomial Regression, which provided a better accuracy for the state of Georgia but was unable to be much better of a result compared to other states.



In this figure we chose to implement the elbow method to help select a number of clusters with a range of values of K. This method allows us to see the number of clusters and find any observable separations.



In this result of the scatter plot , we were unable to conclude any information from this due to the relationship of data points being in clusters not being evident.

VIII. Results

A. *Linear Regression*

We first plotted the total deaths vs the total cases. This resulted in a regression score of .94. This means that 94 percent of the variance in the dependent variable is due to the independent variable. The remaining 6 percent of variance is due to unknown factors. This high score shows a strong correlation between variables, which is what we expected. When we repeated this for each state the data surprised us. While most states have a high regression score, there were a few outliers. One State that stood out is CO. Colorado had a comparatively low amount of total cases and an extremely low regression score. While perplexing at first, we have determined some reasons for the why. Our first assumption is that since Colorado is not densely populated, the rate of infection took longer to pick up. Another reason is that Colorado is not a major player in the American transportation/shipping game. There are no significantly large airports in Colorado and there is not a high amount of goods being exchanged and shipped. The final reason is the time of year and climate of Colorado. The Colorado climate is generally cold and calm. The average temperature in the summertime is between 70 and 87 degrees fahrenheit. While

the exact temperature at which COVID-19 thrives is still unknown, most other viruses and bacteria thrive in these temperate conditions. Another state with a similar climate and population density is Maine. Both states share similar reg scores and graphs.

In contrast we can observe New York and New York City. New York has a regression score of negative .14. Not only is this extremely low, but it is negative, which means that there is a large amount of variance due to unknown variables. At first we believed this is due to the fact that NYC also resides within New York. We then became aware that the dataset separates New York State and New York City. The regression score for New York City was even more confounding. The score is negative 1.42, which is extremely high and invalidates any logical meaning. We believe this is due to the extremely high population density and cold climate of NYC.

While our data backs our belief that cold climates and high population density increase the infection rate of COVID-19, there are still other states that would be considered outliers. One obvious outlier is the state of Florida. Florida is a large land mass with a subtropical to tropical climate. On paper it should be hard to catch the virus while in Florida, but the data shows that Florida has had a steadily increasing rate of infection since January. We believe this has less to do with climate and population density, and everything to do with public opinion and enforced safety regulations. Florida is considered a republican state and a high traffic coastal trading/tourism destination. The amount of imported and exported goods can definitely play a role in the consistent spread of the virus, but we believe that takes a back to seat to how government leadership has affected the state. Florida governor Ron DeSantis was one of the first government officials to retract all state wide COVID-19 business restrictions. We believe his disregard for safety is the primary reason for such a consistent increase in cases. Florida's neighboring state Georgia also shows similar data, was a red state and has extremely high transportation traffic. Georgia's governor, Ron Kemp, also completely disregarded the enforcement of COVID-19 safety regulations for sometime, even going as far as to open a lawsuit against Atlanta mayor Keisha Bottoms for giving businesses the right to deny service to anyone without a mask. Based on this information we believe NYC managed to deter the possibility of a much higher mortality rate through proper education and safety practices. Based on the data, we also foresee a spike in cases in any state that refuses to enforce safety precautions in the coming winter. Next we plotted the death ratio against the submission dates for each state for mean analysis. Our intention is to analyze states that have the highest percent of infected patients that died and to compare the month it happened. The month with the highest death rate for 2020 is the month of February. Washington State holds the record for highest death ratio in a month at a whopping 33 percent. This data can be deceptive though as there were only 3 diagnosed patients in the month of February, which implies only one person died. Through further research it was found out that the majority of deaths in Washington due to COVID-19 were elderly or retired citizens. Based on this information, we could make the assumption that states with a higher youthful population will have a lower death rate than those with a high amount of elderly folks. In many cases the months of February to June have the highest percentage of deaths. A few of the midwestern states have managed to keep an extremely low percentage. Unlike many republican states, Kansas managed to effectively contain the spread of COVID-19. The death rate by month in Kansas is consistently under 3 percent (with the exception of February).Kansas received two more cases of COVID-19 from men traveling from Florida on the 12th of March, which is the same day that the Kansas governor declared a state of emergency. The governor's quick reaction and pro safety stance is definitely a large part of the reason Kansas was able to avoid a deadly outbreak. Kansas and a few other midwestern States disprove any correlation between a states political identity and its rate of infection. Turns out choosing to care about your own health and the health of others has nothing to do with politics!

B. Clustering

We try to perform k-means clustering on the dataset to see if there exists any visible clusters or obvious separation between data points. We use the elbow method to determine the number of clusters. From the elbow curve it seems that 5 clusters is a good choice. We plot the diagram in a scatter plot using the two components calculated using PCA and also plot the centroids on the same plot. It seems like in higher dimensions the data points might be in clusters but the relationship is not obvious in 2 dimensions, so we could not conclude any useful information from clustering results.

C. Polynomial Regression

In the case of polynomial Regression , we use this method to take advantage of providing the best approximation of the relationship between the dependent variable, which is the rising number of cases and the independent variable, which is the number of deaths. One of the problems that we could face using this method was the data being affected by outliers. When applying polynomial regression we were able to improve the accuracy of test predictions significantly compared to linear regression with not much error. There are many factors that can contribute to this such as the increased amount of positive values based on the predict() function and changing the model's complexity. When doing Visual inspection of variables comparing polynomial regression and linear regression, it is clear that polynomial regression is a better straightforward way to model curves.

D. Principal Component analysis

Before applying principal component analysis we must scale our data. After implementation we were given weights of the most useful targets in our dataset. The heaviest weights included submission date, tot_cases and tot_deatb. These variables are primarily used throughout all of our testing.

IX. Future Work

In our comprehensive study of the new deadly outbreak of Covid-19, we were able to provide valuable insights and provide suggestions on how the pandemic can be tackled better and how we can shift our resources in a more efficient manner. However, during the course of our study, we felt there is a lot more to be done in order to further extend this study and provide even better insights.

Our study was limited in a sense due to lack of proper data. An obvious example is the lack of demographic data, such as latitude and longitude. This data is important to study the spreading patterns of different areas. Although we performed clustering on our dataset, we could not provide much useful insight from the clustering results because of the lack of this data. Hence, we believe that collecting demographic/geographic data will help us analyze and dig deeper on finding patterns in different areas.

Although we had a lot of features, but a good portion of them were not much useful to us because they represented the same thing but in different forms, such as confirmed cases, probable cases, new cases and pnew cases all mean the almost same thing. Hence, collecting variables that are different and more diverse will help us to find more clear and accurate patterns from the data.

X. Conclusion

Our analysis of the COVID-19 dataset has revealed some interesting trends and factors. The dataset supports our belief that safety regulations and education are the best forms of proactive prevention. Kansas state is an example of responding appropriately to COVID-19. The data also support the idea that colder climates affect the rate of infection, but not as much as safety. Sheer population density also has a larger impact than the time of the year. When coupled with a lack of safety regulations, areas with large population densities are a COVID-19 breeding grounds. This is supported by multiple states in the north east that have a very high population densities and death rates. A low death rate is also an indicator of an effective medical system. While we were unable to determine any helpful information on the use of medicine, we were still able to pull out a lot of information from this data set. We also did not find an undoubtable correlation between a state's political party and how well they handle COVID-19. This again goes to reiterate that the best thing a state can do to prevent COVID-19 is to enforce safety regulations and educate the population. We can also conclude that multiple factors, outside of this dataset, play a role in every reported case of COVID-19. By analyzing the Covid-19 dataset, we hope to gain valuable insights about how the virus is spreading in different states and how different factors are affecting the outbreak. We believe that we did achieve this.. In the end, we hope that our models are able to provide help predicting the rate of infection and the factors that affect this rate. We achieved this by evaluating clean data and performing methods such as linear regression and polynomial regression. It was our hope that our work would reveal a solid trend in the data, which it did, showing that there is a strong correlation between infection rate and time.

XI. References

[1] World Health Organization. Naming the coronavirus disease (COVID-19) and the virus that causes it. The ICTV's page is here: International Committee on Taxonomy of Viruses (ICTV). 2020. https://talk.ictvonline.org/.

[2] World Health Organization. WHO statement regarding cluster of pneumonia cases in Wuhan, Chinam Jan 9, 2020. 2020. https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china. Accessed 15 Feb 2020.

[3] Ren LL, Wang YM, Wu ZQ, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study [published online ahead of print, 2020 Feb 11]. Chin Med J (Engl). 2020. 10.1097/CM9.0000000000000722.

[4] Centers for Disease Control and Prevention. 2019 Novel Coronavirus (2019-nCoV), Wuhan, China. 2019. https://www.cdc.gov/coronavirus/2019-nCoV/summary.html.

[5] Muthusami R, Bharathi A, Saritha K. COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the World. J GED Organ. 2020;33(2):534–549.

[6] Santosh, KC . AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. J Med Syst. 2020;44:93.

[7] Lauren G. Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE January 23, 2020. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html

[8] Clustering analysis of countries using the COVID-19 cases dataset. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7258836/

[9] United States COVID-19 Cases and Deaths by State over Time. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36