AI-19 Advisor: Fighting COVID-19 With Predictive Analytics

Team Members:

Adwaith Hariharan, Shagun Khare, Nandana Kumar, Bhavini Pandey, Ananth Prabhu, Harini Rajadeva, Dhruv Vaidya, & Kavya Venkatesan

Advisors:
Dr. Madiha Jafri & Dr. Ghulam Rasool

June 2020-August 2020

Organization Affiliation: Agraj Seva Kendra
Program Coordinator : Sruthi Suresh

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of Dr. Madiha Jafri, Dr. Bruno Leao, Rafael Griffo, Dr. Ghulam Rasool, and Dr. Sindhu Suresh for their technical support in accomplishing our project objective. We would like to thank our TA Pranay Chinthaparthi for providing all logistic support and finally Sruthi Suresh the program coordinator for giving us the opportunity to explore our research interest in the field of Data science.

ABSTRACT

The word “Corona” or “COVID- 19” has been engraved into the minds of people around the world. The sudden outbreak of this impactful disease has led to thousands of deaths nationwide, economic shutdowns, and other various tragedies. This pandemic is so newly born, and there is so much that the world has yet to discover. As of now, everyone in the world is trying to take safety precautions and doing everything in their power to prevent the spread of this contagious disease. In these unprecedented times, it is imperative that people understand how susceptible they are to the virus. This would allow individuals to limit their actions and exposure to various situations involving COVID-19.

Since there is so much uncertainty regarding the transmission and symptoms of the virus, we as a team wanted to develop a COVID-19 risk calculator that would help people understand their mortality risk due to COVID-19 based on various factors. To accomplish this, we broke down the different underlying factors that may have an influence on humans dying from COVID-19. These included race, age, gender, location, and Non-Pharmaceutical Interventions (NPIs). Race played a role due to the fact that it guaranteed that all ethnicities and races had an equal chance of being affected by the pandemic. There are towns and neighborhoods that have a majority of one race and may not have access to healthcare supplies. Age was chosen as a factor, since people of older age seemed to be far more vulnerable to dying from the disease due to the fragility of the body's health. Gender was a factor that was included to differentiate whether males or females had a higher chance of getting the disease. Population density or the compliance of people in a certain area may cause the spike of COVID-19 cases in one state compared to another, which is why the location factor was included in our risk calculator and research. The NPIs were then considered as the final major factor. It was very important to account for the levels of precaution various states were taking. This included how well states practiced social distancing, whether people were wearing masks, etc. With these differentiating factors, a COVID-19 risk calculator was created to prevent and inform people on how likely they were to die from this newborn virus.

INTRODUCTION

Background: COVID-19 and Artificial Intelligence

On December 9th, 2019, a pneumonia-like outbreak was detected in Wuhan, China. This led to the identification of a novel virus that was later named the Coronavirus (COVID-19)[1]. Since its discovery, COVID-19 has affected countless lives and caused massive socioeconomic and health crises Worldwide. Problems associated with COVID-19 include the fact that the virus is known to mainly spread through respiratory droplets that can easily enter the mouth, nose, or lungs when people come in close contact[1]. Another problem is that COVID-19 symptoms may appear only 14 days after the person is infected[1]. Therefore, victims of COVID-19 can spread the virus unknowingly due to asymptomatic carriers. These serious issues are some of the reasons why COVID-19 is still continuing to affect our society and world.

The global population is now 8 months into this ongoing pandemic, which has changed the normal lifestyles of people. The “new normal” includes staying home, wearing masks when going outside, practicing social distancing, and remote learning and working. Despite this, it is imperative that people do what they can to help flatten the curve by taking the necessary preventative measures to protect themselves and others. To bring more awareness to the public and understanding about the virus, Artificial Intelligence (AI) is being harnessed in a variety of ways during the pandemic by policymakers, health care professionals, and companies.

AI involves designing and programming systems that can simulate human intelligence and solve complex problems. This powerful tool enables the world to uncover new trends and compute unimaginable possibilities. One application of AI during the pandemic is automated patient care. Especially during the COVID-19 pandemic, hospitals are overflowing with patients, and there is a shortage in healthcare workers. This means that important questions related to COVID-19 such as whether a person is showing symptoms of the virus may go unaddressed. Governments are implementing remote chatbots or virtual assistants that can automate patient care and provide “information about the outbreak, symptoms, precautionary measures, etc.”[2]. Patient care is also being automated by systems that are monitoring the health conditions of patients and collecting observational data. The data is currently being used to understand how

effective certain treatments are on COVID-19 patients. This information not only ensures resources can be allocated better, but also opens up the possibility for doctors being able to monitor COVID-19 patients' health remotely.

AI is also being used during the pandemic to aid with COVID-19 patient prioritization in Intensive Care Units (ICUs). In the past, it has been suggested to implement "algorithmic risk assessments of diseases such as cancer, diabetes, and cardiac-related diseases with Artificial Neural Networks (ANNs)"[2]. Similarly, ANNs are being used in algorithms that factors take into account like age, gender, and health conditions, to assess a COVID-19 patient's mortality risk. By knowing which patients are at most risk, doctors can get the treatment they need to them immediately and ensure more lives are saved. Furthermore, since a variety of COVID-19 symptoms have been reported, these AI-driven algorithms can make quick and accurate assessments that alleviate uncertainty in hospitals.

Finally, AI is being used around the world during this crisis for models and simulations of viral spread. These models "can capture the effects of interventions (e.g., social distancing) and differences among populations (e.g., herd immunity) to predict what might happen in different circumstances in a single region" [2]. By combining many of these models and simulations, a more accurate picture of the spread among several regions across the world can be developed, allowing better educated planning and decision making from the world in stopping the virus from spreading. Not only that, but simulations in particular could also help predict how much medical equipment might be needed for the future by allowing analysts to determine how much will be consumed based on previous models and real-time data. This process of analyzing historical and current data to make predictions about future outcomes is known as predictive analytics. Predictive analytics is a well-known area of AI that involves techniques like data mining, statistical analysis, and predictive modeling.

Machine Learning (ML) is also closely related to predictive analytics. ML is where a model learns from its previous experience with the training data to make observations or predictions. Over time, the model's learning curve and performance improves, and it is able to provide more accurate results when presented with new data. There are a variety of ML algorithms that can be applied to data. An algorithm is a set of instructions that is executed by a

machine. Examples of ML algorithms include linear regression, decision trees, and K-means. ML algorithms can be used for a variety of purposes such as predicting outcomes or even clustering data points based on the features.

Project Visualization

As the world is experiencing unprecedented times, it is imperative that people are aware of their COVID-19 risk levels and take the necessary precautions to lower that. To help create this awareness, we were motivated to design a solution that uses AI, predictive analytics, and ML to assess a person's vulnerability towards COVID-19. The goal was to design a reliable AI-driven system that predicts how vulnerable a person is to COVID-19 based on the following factors: race, age, gender, location, and NPIs. To achieve this goal, a project vision was established on how the model would function.

First, the user inputs personal information related to the 5 major input information into the Python-based web interface. The information is then processed by the AI algorithm for each factor. A percentage is derived for each factor based on the user's input and data we fed into the algorithms. These percentages are then averaged. The output of this calculation is the risk level of one dying from COVID-19, which appears as a percentage on the User Interface (UI).

METHODOLOGY

We decided to incorporate AI, predictive analytics, and ML in our project to effectively analyze the available, public data on COVID-19 death counts. This way, we can better understand the trends of who is most vulnerable to COVID-19 and inform the public about our critical findings. We divided the data for the five factors (race, age, gender, location, NPIs) among ourselves to work on, and each factor incorporated the predictive analytics in a unique way. However, the procedure of how the predictive AI models were implemented was the same for all the five factors overall .

First, the data for each of the five factors was mined and cleaned. This process involved using Microsoft Excel and other programs to sift through the data for any trends and remove outliers (values within a data set that deviate from the main trend). Finding key patterns and relationships in the data was important because they can help assess the predictive algorithms' performances. After refining the data, a statistical analysis was performed on the data. Values like median, mean, variation, and standard deviation were calculated. This helped us evaluate the quality of our data. Data visualizations such as charts, tables, and graphs were also used to model the correlations and trends of the data points, which would later help us in deciding the AI algorithm that was going to be implemented.

Next, research was conducted on the different AI algorithms that could possibly be used in the project. It was later decided that 5 AI algorithms such as linear and polynomial regression would be utilized because each factor's data sets contained unique trends that needed to be treated in different ways. We also researched problems that have occurred in the past with predictive models and identified strategies to mitigate them. Some of these problems that we came across included overfitting and underfitting. Overfitting occurs when the model learns the training data too well and starts to memorize all the specific details including the noise. This hinders the model from finding a general trend that it can use when faced with new data. Therefore, the model may perform well on the training data but not on the test data. To ensure overfitting did not occur with our models, the outliers were removed to improve the quality of

the data. Linear models were also used as much as possible for the project to address this possibility.

Underfitting was another problem where the algorithms did not model the data well and capture the general trend. To prevent underfitting from happening with our data, we researched different performance metrics that could be used to measure the predictive power of our models. These metrics included R squared (R^2) and mean squared error (MSE), which represent how close the data points were to the fitted regression line. All of this information was helpful for finalizing the algorithms and determining how we would train our models to make accurate predictions related to COVID-19.

The final step involved programming the selected predictive algorithm and validating it with test data. To achieve this, we used the numpy library, matplotlib (plotting library), and regression models that were imported from the sckit-learn software. We were able to plot the data points for each factor and fit them with a regression line (linear or nonlinear) that could be used to predict the death counts of a certain group. This predicted value was then used to calculate the probability of dying from COVID-19. The formula for this calculation was not the same for all the factors. Then, the models were connected with a program that took the output from each of the factor's algorithms and averaged the values to obtain one final probability. This final probability was displayed on the UI as the person's mortality risk from COVID-19.

RACE

Due to the COVID-19 pandemic, over 26 million people have fallen sick, and of those sick, over 870,000 people have died. Data on the disease is being recorded throughout the world, which explores how COVID-19 has impacted various groups of people. This includes studies on how the different races have been impacted by the virus. We were aware that some races were affected more than others, and that not every race was equally vulnerable to the virus. The purpose of this study was to observe the effect of COVID-19 on the separate races living in the United States of America. This information was then used to create a program, which would utilize AI algorithms to predict how individual races would be affected in the future, as well as any single person's likelihood of contracting COVID-19.

Procedure

The first step of the project that was taken was the collection of raw data. The data was collected from several sources, including the Centers for Disease Control and Prevention (CDC) and other COVID-19 tracking sites for the project[3]. Based on the numbers alone, it appeared that the white population had the most COVID-19 cases and deaths in the United States. However, this did not necessarily mean that the white population was the most vulnerable to the virus. The white cases and deaths were significantly higher than the other groups, because their population in the United States was also significantly higher. Therefore, the next step to find out which race was truly the most vulnerable to COVID-19 was to take the population of each race into account as well. Before we did this, we made a hypothesis as to which race was the most vulnerable to the virus. Before normalization, the white race had the most cases out of all of the races, so our initial thought was that they were the most vulnerable to COVID-19. Later on in the project, however, this would prove to be wrong due to normalization.

Analysis

Finding which race was truly vulnerable to COVID-19 was achieved through normalizing the data. In order to normalize the data, we divided the deaths for a single race by the population

of that race in the United States. Normalization helped ensure that the results were reliable and truly represented the intended information. The normalization of the raw data as well as some of the visualizations and graphs for the project were created in Microsoft Excel and Google Sheets. The Seaborn program was also used for some of the visualizations, which included heatmaps, scatter plots, and the final graph.

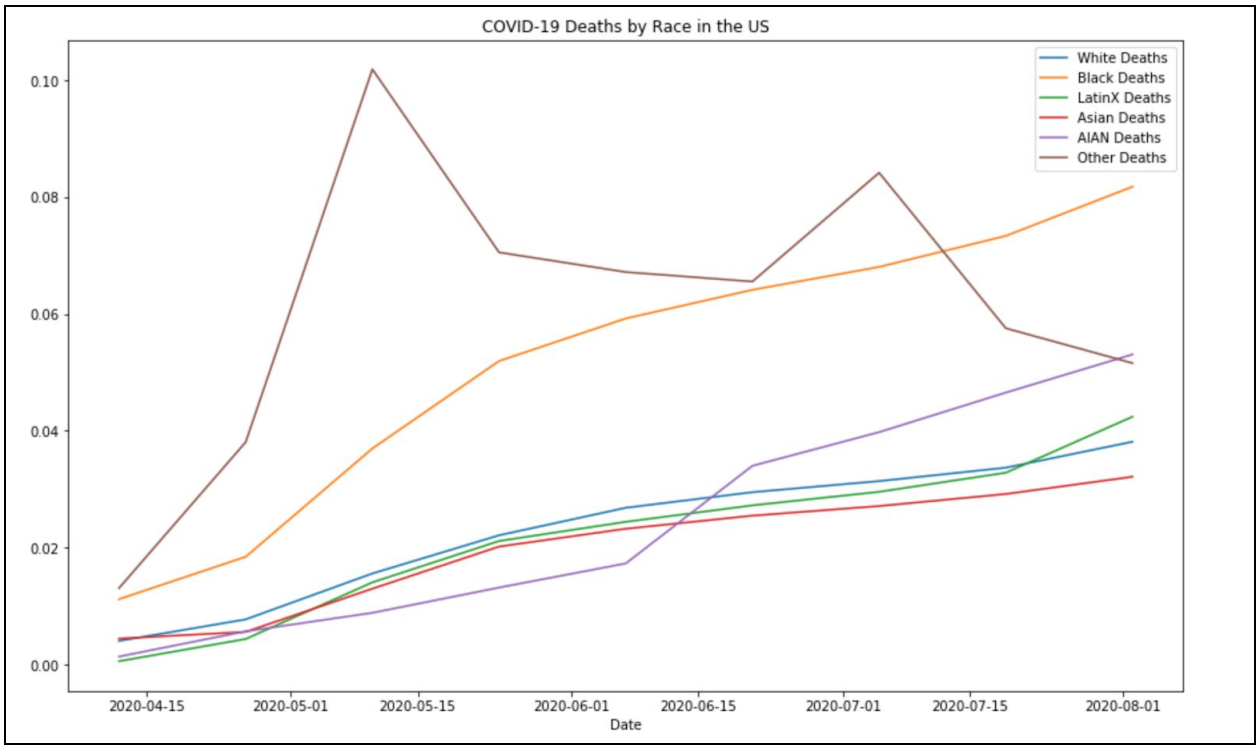


Figure 1: Relationship between time (Starting from April 15th, and ending around August 1st), and normalized percentages, representing the COVID-19 death rate among the races listed in the key above.

After the data was normalized and the graphs were created, linear regression models were developed for each race. We used the JupyterLab notebook to both write and run our code, which was in Python. Throughout the development of the final model, we referenced code provided by our project leads. The final model correctly predicted a person's susceptibility to COVID-19 based on race by allowing a person to input a value (representing their race). Also, based on the final model, we were able to conclude which race was most impacted by the virus. After careful analysis of the data, we figured out that the hypothesis we had made was rejected, and it was in fact the Black race that was the most vulnerable to the virus. The reason why it appeared that the

White population was the most vulnerable to the virus was because they had the most cases. However, after factoring in how many people were actually in each race, the numbers clearly showed that the Black population was the most affected (Figure 1). After making our final conclusions, we used our final model to output a percentage, which displayed the person's chances of dying from the disease solely based on race. In the final program, multiple different factors such as race, age, gender, location, NPIs were averaged to find the final probability of dying from COVID-19.

Results

Based on our model (Figure 1), we had made many conclusions on the consistency of each race group over time. The other category's consistency was extremely interesting, since they were the least consistent group. Originally, the line started off with a steep inclination throughout the first few weeks. Later on, we then observed a downfall. After this, the line briefly increased again before finally decreasing. The consistency of the Black population also certainly stood out, and this is because the Black population consistently increased by large amounts. It can be seen that the line seems almost exponential.

After completing our research and data analysis specifically for the race factor, it was found that Asian-Americans have a relatively low chance of dying from COVID-19, whereas the Black population has the highest chance of dying from COVID-19. If an Asian-American was to input his or her race into our model, the chances of dying from the disease would be a mere 4.15% based on the most recent data. On the other hand, if a Black person inputted his or her race into our model, their chance of dying from the disease would be 10.38%, which is significantly higher. The American Indian and Alaska Native (AIAN) population also showed a noteworthy increase throughout the months. Their population started off as one of the least vulnerable populations but eventually grew to be the second most vulnerable to COVID-19 behind the Black population. In addition, our model calculates that the percentage for a White person would be 4.9%, putting them in the lesser vulnerable group, which is interesting considering how the White population had the most COVID-19 deaths.

After spending about two months working with the race data, we have come to the conclusion that the Black Americans and groups of color are the most vulnerable to COVID-19. Black Americans are surely the most at risk from dying from the disease with a clear exponential growth in their COVID-19 cases. Several studies done by medical professionals at centers such as Johns Hopkins Medicine have come to similar conclusions. According to a Johns Hopkins article titled “Coronavirus in African Americans and Other People of Color”, “People of color, particularly African Americans, are experiencing more serious illness and death due to COVID-19 than white people”[4]. Other sources including ABC News have also suggested that the Latino community is quite vulnerable. According to an ABC News article discussing the Latino community, in the pandemic, “Many Latino communities have been disproportionately impacted by the coronavirus pandemic”[5], and the article continues to describe how the Latino community was extremely vulnerable in the pandemic due to racial disparities. However, based on the data that we collected, the Latino community is only slightly more vulnerable than the White population, which still puts them in the less vulnerable half.

In spite of the differences between the various studies done on COVID-19, there is no guarantee about the vulnerability of any given person in the United States, which is why our team, statisticians, medical professionals, and many more will continue to gather data and analyze it in order to learn more about the virus and keep the world safe.

AGE

The impact of age on the clinical characteristics and deaths related to COVID-19 patients has been closely tracked along with gender and race. Age is closely associated with certain changes in pulmonary physiology, pathology and function, during the period of lung infection[6]. We hypothesized that these age-related differences will have an impact on the health and physiology of COVID-19 patients, leading to increased deaths in elderly individuals.

Procedure

We identified, collected, and mined raw age-related data from the Centers for Disease Control (CDC) and Prevention Data Tracker[7]. The data set contained the weekly COVID-19 deaths from February 1st to August 1st for the following age groups: Under 1 year, 1-4 years, 5-14 years, 15-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 year, 65-74 years , 75-84 years, and 85 years and above. Initial analysis of the data indicated that the age groups of 85 years and above had the highest death counts. However, these numbers were not conclusive as the percentage of individuals that make up each group could be different. Therefore, the data was normalized, and the trends were inspected to achieve accuracy. Normalized data indicated that individuals who were 85 years or older made up 66% of total COVID-19 deaths. We then started analyzing the data by breaking it down into several different sets to examine trends. Given the trends we saw, age-related data was checked against three different regression models - linear regression, piecewise linear regression, and exponential regression to arrive at the best fit. Data analysis indicated that the Piecewise Linear model was the best fit for our data.

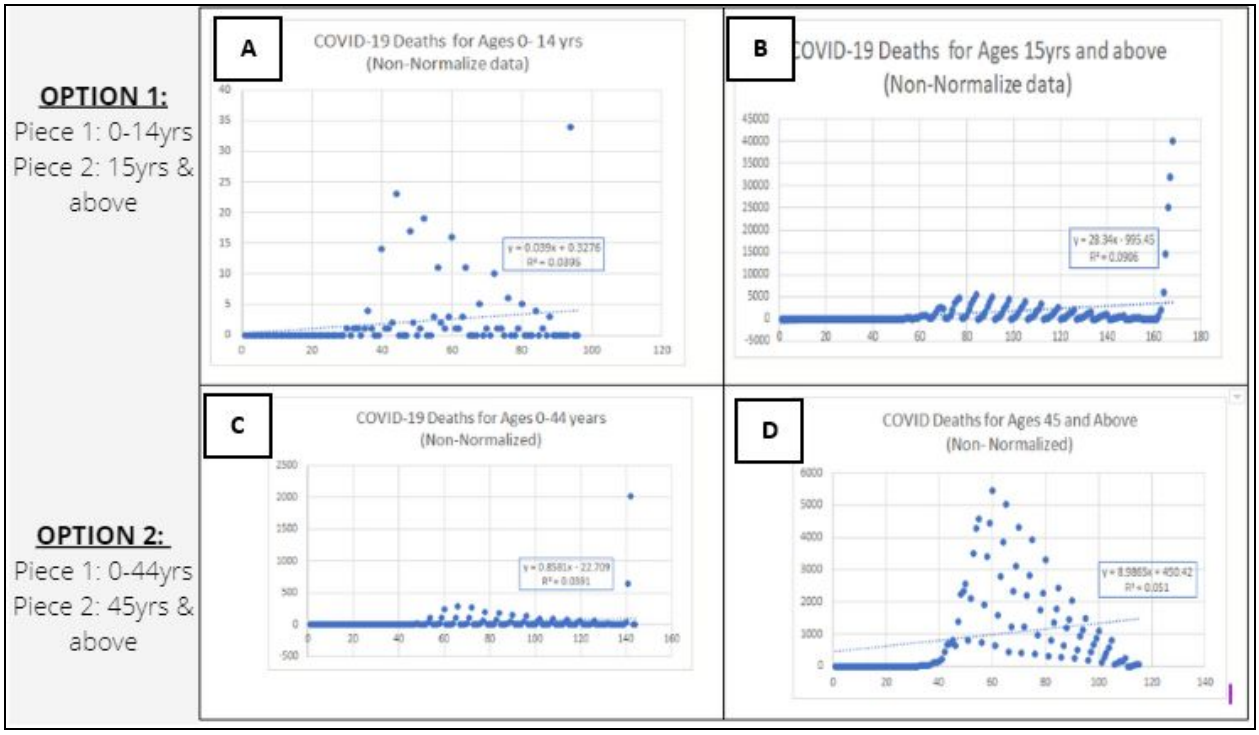


Figure 2: Visualization of Piecewise linear regression model for Age data. Visualizations in A & B and C & D shows Piecewise regressions for different data sets to find the line of best fit.

Analysis:

Despite the fact that the piecewise linear regression model seemed to be a good fit, we realized that using the data for the cumulative deaths during the summer months was ineffective for making predictions. A week-by-week breakdown for each age group would tell a more detailed story and yield better predictions for our model.

This led us to programming 11 models, one for each age group. Before deciding which type of regression algorithm to apply, the data points were plotted on graphs using Google Collaboratory. Unlike before, the x-axis was set to weeks (starting from 6/6/2020 and ending at 8/1/2020), and the y-axis was set to COVID-19 deaths. This data visualization helped us understand how the death counts of each age group changed over the time span of 9 weeks. One of the major trends we noticed was that from week 1 to week 3 (June), most of the age groups' death counts decreased. The death counts for all of the age groups increased from week 4 to 6 (early July). From week 6 and onward (mid-July to early August), there was a sharp decrease in

the COVID-19 death counts for all of the age groups(Figure 3). This trend may be attributed to the different restrictions that were placed and lifted at various times. Although the trend was interesting, there was a lot of fluctuation, which meant that the data could not be fed into a simple predictive algorithm like linear regression or piecewise linear regression.

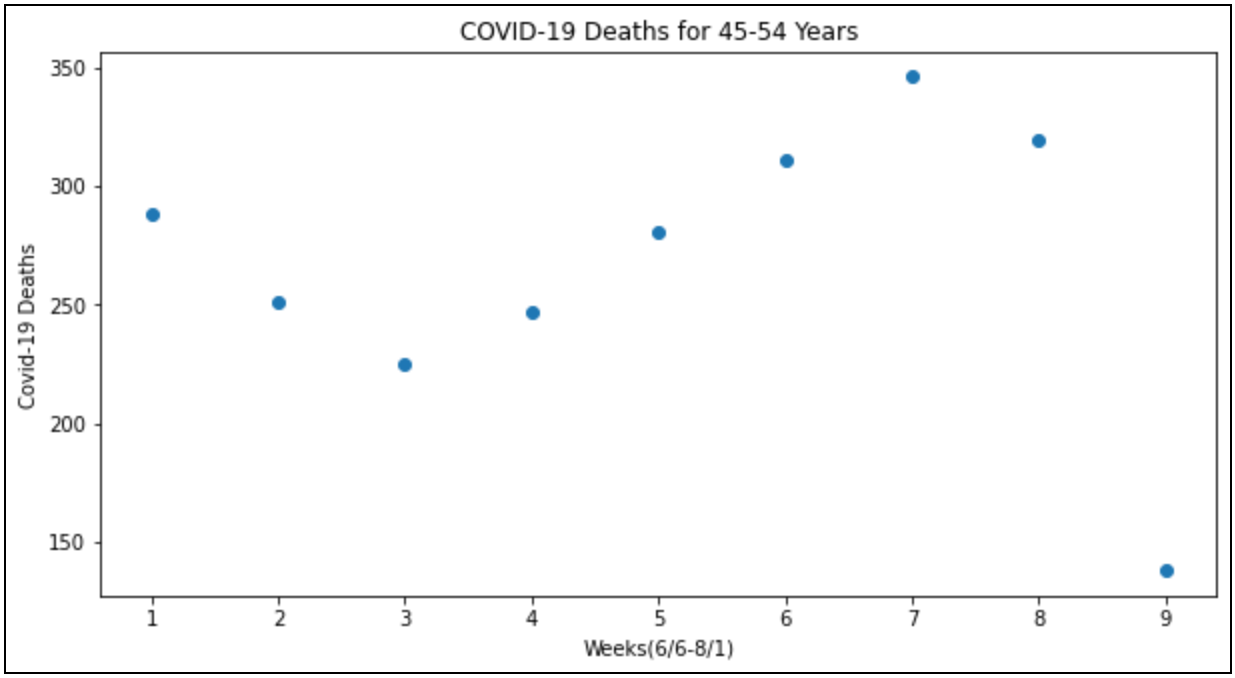


Figure 3: Trend of the change in death counts over the 9-week time span for 45-54 Years Age group.

We decided to apply the polynomial regression models to the data. Some of the age groups such as Under 1 Year, 1-4 Years, 5-14 Years, and 15-24 years were consolidated, bringing down the total number of models from 11 to 7. Then, the program for the polynomial regression was written using online tutorials and feedback from the advisors. Different robust linear estimators were also experimented with throughout the process of programming the model. The best robust linear estimator was chosen by calculating the mean squared error value. Mean squared error value is a performance metric that measures the average square difference between the estimated and actual values. The lower the mean squared error value, the better the model's predictive power is. The Linear Regression Estimator had the lowest mean squared error values, so we came to the conclusion that it was the most accurate predictor for all the age groups. The Linear Regression Estimator was then applied in the programs for all the 7 models, and the data was fitted with polynomial regression lines. From the graphs, it was evident that the

regression lines seemed to accommodate for the fluctuations in the data very well (Figure 4). This meant that these lines were reliable enough to be used for predicting COVID-19 deaths.

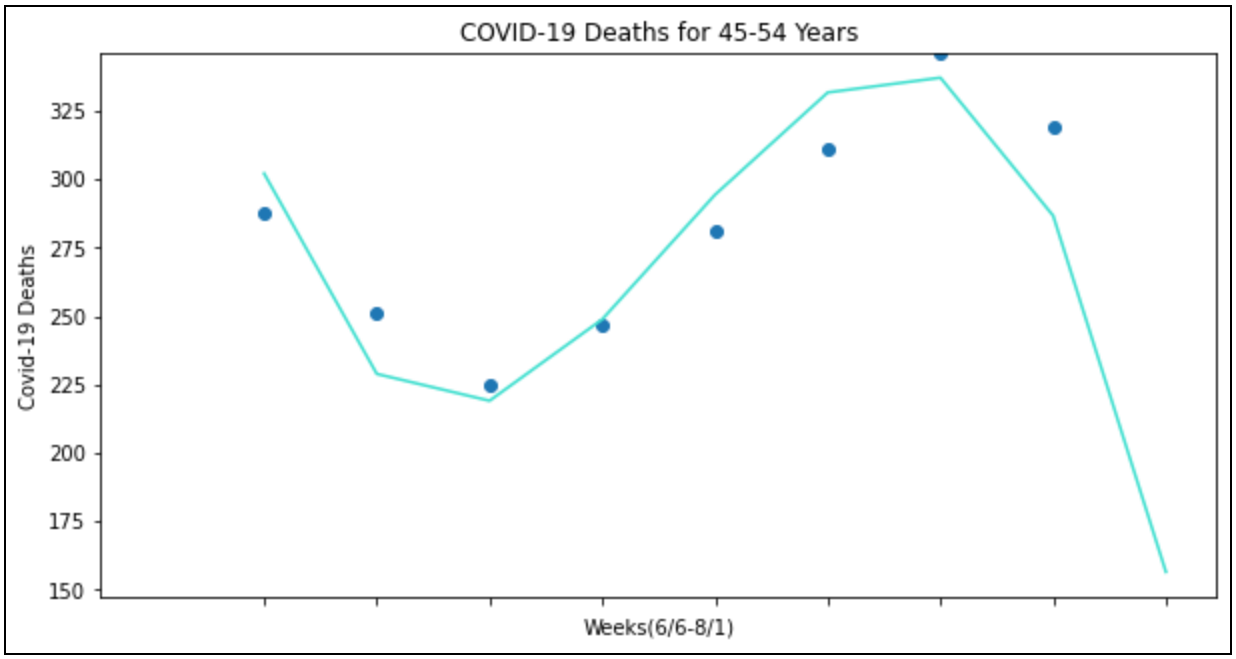


Figure 4: COVID-19 death counts over the 9-week time span for 45-54 Years Age Group. The data points in the scatter plot are fitted with a polynomial regression line.

Finally, the calculation was written for the probability of one dying from COVID-19 solely based on his or her age. We decided that the algorithms would be calculating the probability for the current time, which is the rest of August. The formula for the probability involved taking the predicted value of COVID-19 deaths and dividing it by the total deaths (including COVID-19). The total deaths was an additional part of our data set, and we were able to hard-code those values into our program for the probability calculations.

Results

The overall trend was that as age increased, the probability of dying from COVID-19 also increased. For instance, 0-24 years age group had a 0.6% probability of dying from COVID-19. Meanwhile, the 75-84 years age group had a 9.04% probability of dying from COVID-19. These

results were consistent with the literature survey, which showed that the older age groups are more vulnerable .

Surprisingly, the 85 and Above years age group had a 7.6% probability, which was lower than the percentage of the 75-84 years age group. Although 85 years and above had the greatest number of COVID-19 deaths, it seems that the virus affected the overall death counts of the 75-84 years more than the 85 years and above, which is why the probability was greater. In summation, we can conclude that the older age groups are the most vulnerable to dying from COVID-19, thus proving our hypothesis to be correct.

GENDER

There are many underlying factors that impact the overall cause of death when relating to COVID-19. This new found disease has struck the world with zero precaution leaving no room or time for humanity to prepare. One of the factors that we decided to research included gender. We wanted to see if a specific gender had a greater chance of dying of COVID-19 than the other. Before conducting our research and process, we hypothesized that females would be more prone to catching COVID-19 over males. This hypothesis was formed based on prior knowledge relating to the fact that there are more female health care workers than male. We assumed that since more female nurses are evident in hospitals and other places where COVID-19 is prevalent[8], they would have a higher death rate. Within our journey to either prove or disprove our hypothesis, we took many steps which included: cleaning up data, creating various visualizations to model trends, and creating a code to apply our algorithm.

Procedure

To start our research, we began to scour the internet for reliable data sources. We took the data that was provided on the CDC website (Centers for Disease Control and Prevention)[7]. We extracted the dates from February 1st through August 1st. Additionally, it was decided to only use the set biological genders that each person was assigned at birth (male and female).

After cleaning our data, we created various visualizations to observe the different trends in the data. We made a pivot table derived from the cleansed data. From our raw data, the pivot table was created to model the total deaths per week for each gender. This made the data easier to interpret, which showed that males had more weekly deaths. We also created different line graphs. From the line graph, the trend for both genders showed a steady, linear increase until April, when it then started to decrease. However, males had a slight increase of deaths over females. This discrepancy had rejected our hypothesis, as we had predicted that females would have a higher death rate.

After seeing that males had a higher death rate due to COVID-19, we brought down all the values and variables to the same range, through data normalization.

Normalizing the data changed the values of the numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. In order to normalize it, we divided each weekly value for both genders from our pivot table by the total population of either gender in the US population (which was then divided by the US male population for the male deaths and the US female population for the female deaths). After, we created another pivot table with the normalized data. We created a graph with the normalized data, resulting in a bell-curved graph. These results proved to be the same as the graphs and results we had created prior to normalizing our raw data.

Analysis

Throughout the data visualization process of the gender data we analyzed it in order to conclude which algorithm best fits our data. The algorithm that we deemed most effective and useful was the Linear Regression algorithm. We decided on this algorithm because we wanted the simplest, yet most effective algorithm that could model the linear trend seen earlier. We saw linear relationships between time and the cumulative female and male COVID deaths in our data visualizations, which meant that the data is best suited for linear regression.The end goal was to calculate the probability of dying from COVID-19 based off of the biological gender a person was assigned at birth. Google Collaboratory was used to apply our python programming. Throughout the coding process, we referenced online tutorials and code from the advisors. Although our data contained death counts from February 1st through August 1st, we wanted to use data from June 6th through August 1st. We wanted to use the most recent data that was

available to us as well as stay consistent with the other factors.

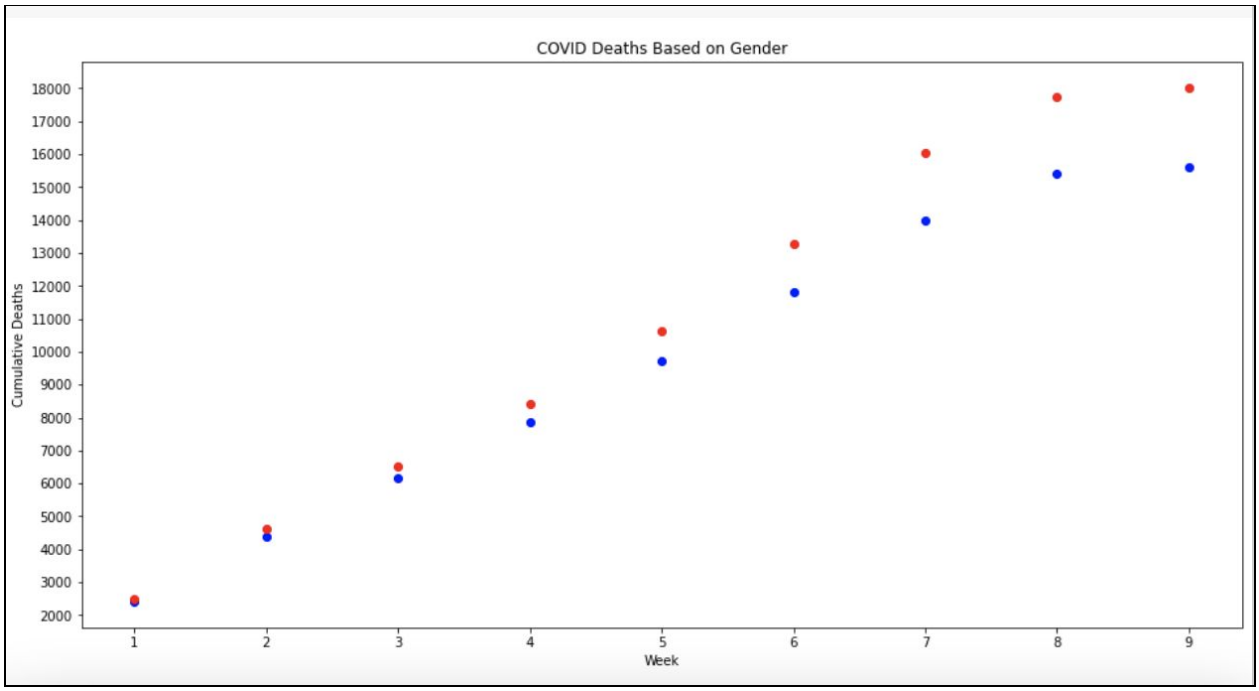


Figure 5: Linear Regression Model of the relationship between time and cumulative male and female COVID-19 Deaths (Male=Red, Female=blue)

Results:

Our scatter plot showed male and female cumulative deaths increasing every week. Males were more affected than females each week. We also calculated our linear regression equations and the R^2 values. Our R^2 values were extremely high, which meant that the relationship between the week and the cumulative deaths for males and females was very strong. Finally, we calculated the probability of dying from COVID-19. For our particular factor , we calculated the percent increase for both genders. We concentrated on the data from August since we wanted to use the most current trends. Each gender was assigned a number. Male was assigned 1 and Female was 2. Then, we added the calculation for finding the probabilities of dying from COVID-19 for each gender with its respective linear regression equations, which we derived from the algorithm. When we calculated the final probability for males, it came to 12.19%, and for females, it came to 11.73%. The values seemed to be accurate because our trends show that males were more impacted than females. It also makes sense that the

probabilities for males and females were very close because the data points that represent the male and female Cumulative Deaths in the scatter plot were close for each week. Overall, the gender of a person doesn't seem to play a large role in a person’s risk of dying from COVID-19.

LOCATION

In addition to examining factors, such as age, race, and gender, a large sum of the rapid growth in COVID-19 deaths can largely be attested towards the location of a person in the US. As each state had its own individual impact on the spread of the virus, there were several factors [9] that needed to be taken into account when examining the death rates of each state. A few of these attributes include each state's population density, non-COVID death causes, and the implementation of social-distancing guidelines and other non-pharmaceutical interventions. When analyzing the demographics, it was important to take these factors into consideration in order to account for the varying trends over the span of several months for each specific state. Therefore, from prior knowledge on the cumulative death counts on a state-by-state basis, our hypothesis was that people who lived near states such as New York and New Jersey would have a higher chance of dying from COVID-19 [10] as compared to lesser populated states such as Hawaii and Vermont.

Procedure

To start with analyzing the location data, the first measure we took was the collection of raw data from the Centers for Disease Control and Prevention Data Tracker. In the initial data sets, the location data were broken apart into two categories: states and dates. The state by state breakdown analyzed the cumulative cases and deaths for each respective state whereas the temporal data focused on the cumulative cases in the United States on a daily basis. After breaking down several different sets of data and examining similar trends, we realized that breaking apart the data into these categories would not be useful in comparing the increase in cases or deaths per state throughout the course of the months. Therefore, we had to mine the data again in order to find an accurate representation of data that would underlay the cases and deaths for each state by each day. By accomplishing this, it would be more accessible to compare and examine the trends each state had independently rather than on a cumulative basis. Thus, through extensive research, data was extracted to be fed into the algorithm[11].

This data was first exported through Excel as a Comma-Separated-Values file and then imported to Google Sheets for its easy use and shareable settings. As this location data set modeled information such as increase in deaths and rise in positive COVID-19 cases for each of the states on a daily basis, we chose to push through with examining our new data. After using features from Google Sheets to graph and analyze the data for total deaths and deaths increased daily, the trends allowed us to choose the algorithms that would best model the location data.

Our next steps were to proceed with Google Collaboratory, a python notebook, that would allow us to import our data and program the algorithm. Our plan was to fit the data with the regression lines and use them to predict the probability of someone dying from COVID-19 based on the state he or she resides in. To successfully import the location data for each state, data was directly imported from The COVID Tracking Project as a url-based source. With the data imported, the next step was to code the Linear and Polynomial Regression models using Python libraries, such as numpy and pandas. The data points were then fed into the models, producing scatter plots and simple line graphs. After that, we were able to retrieve input from the user based on their state and output the corresponding visual representation. Our final step in modeling the location data was to calculate the probability as percentage increase of deaths from a state by state comparison. After having inaccurate percentages by using the state populations, our final method in calculating this probability was dividing the COVID 19 deaths by the sum of the COVID-19 and non-COVID deaths for each state, and comparing this percentage to the top three states that are currently having the greatest rise in deaths during the allotted time period.

Analysis

Throughout the data visualization process, the location data was analyzed and used to make conclusions in order to accurately model its progress and results in comparison to the other factors. When comparing and contrasting the graphs from our initial data mining process, we arrived at the conclusion of using Polynomial Regression in order to better accommodate curvature. Regardless of the numbers each state outlined, the states all followed a format in which the cases rose rapidly from March to May but started leveling off after June. However, it was still necessary to account for the occasional spikes each state showcased independently.

Therefore, the location model took an extra step in combining the data with the Non-pharmaceutical Interventions implemented in each state in order to find a reason as to why these outlets existed in the models. After aligning the beginning and ending dates each NPI was implemented in with each respective state, we conducted extensive research on the compliance of these certain NPIs. Our results portrayed that out of all the states, Nevada, Pennsylvania, and Hawaii had the highest compliance to these NPIs. However, after overlaying these specific states with the corresponding NPI, there was no correlation between the trends of the location data and the implementation of these specific NPIs. Thus, we chose to model these sets of information separately in our model as they were independent from each other.

Thus, in the final model, the location data took into account the non-COVID deaths each state had in comparison in order to categorize which states were more at risk of contracting COVID deaths as to other states. As each individual state modeled its own percentage increase, these values were compared side-by-side with Georgia, Florida, and Texas because these states have shown the most active rise in deaths. All in all, the analysis of the data visualization helped us form conclusions to better construct the model of a person being able to compare the chances of being susceptible to COVID-19 deaths in comparison to other states.

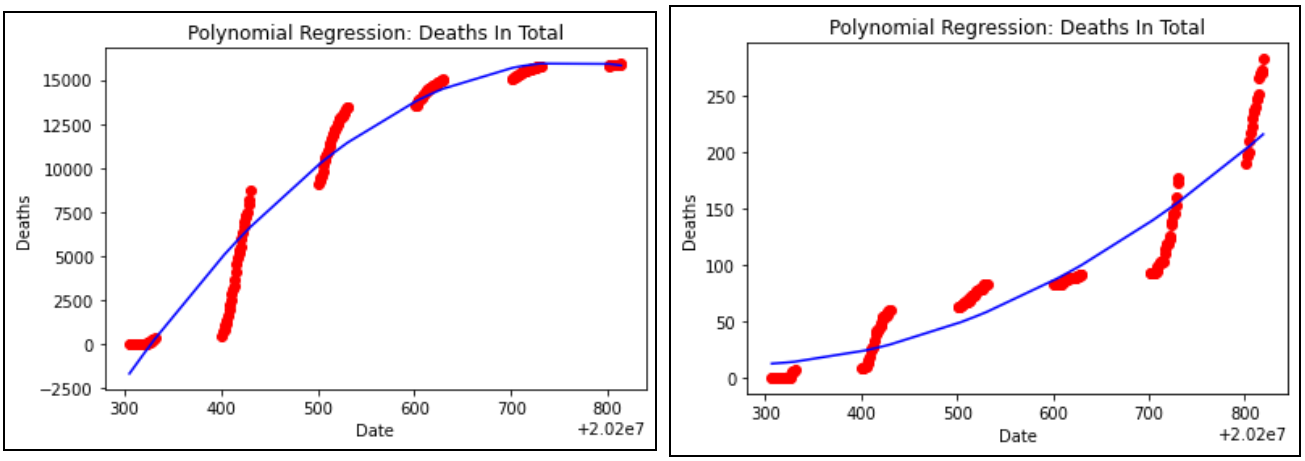


Figure 6: Polynomial Regression models of New Jersey (left) and Vermont (right) and compares the trends between months and cumulative COVID-19 deaths.

Results

In order to stay consistent with the other factors of this model and fulfill the project's purpose of predicting a person's likelihood of dying from the virus, the location data took user inputs of the state the person resides in and the current date. Using this information, the model extracts the updated data sheet for the corresponding state and imports the current date to find the corresponding number of deaths through the predicted curve of best fit. To find the most accurate model of best fit, it was necessary to compare the Root Mean Square Error (RMSE) values of both Linear and Polynomial Regression. After arriving at these values, the RMSE value for Linear Regression was more than double of the RMSE value modeling for Polynomial Regression. Therefore, it came to our understanding that we disregard the Linear Regression model and continue with Polynomial Regression in order to output a more accurate percentage. Then, the model undergoes a calculator hardcoded by the program to present the user with the probability of contracting COVID-19 deaths in comparison to the other states. In order to arrive at this probability, we had to compare and contrast several different results in order to select the most applicable percentage in comparison to the other factors. Initially, the location model predicted a percent increase of merely 1.29% as a rise in deaths for New Jersey. However, after a detailed discussion, the percentage rose to 34.45% by assessing New Jersey's total death count in comparison to its COVID-19 deaths.

NON-PHARMACEUTICAL INTERVENTION (NPI)

Non-pharmaceutical Interventions are community mitigation strategies implemented to stop the spread of various infectious diseases. NPIs, as defined by the Center for Disease Control and Prevention (CDC), are "actions, apart from medical interventions like getting vaccinated or taking medicine, that people and communities can take to help slow the spread of communicable diseases like pandemic influenza (flu)"[12].

Due to the absence of effective pharmaceutical interventions, such as a vaccine or antiviral to combat the spread of COVID-19, several NPIs have been implemented across the world to varying degrees to effectively control the spread of this pandemic. NPIs applied so far have ranged from extremely restrictive complete lockdowns to less severe strategies. As part of this research, we wanted to look at how NPIs like social distancing, school closures, closure of public venues, gathering size restrictions, non-essential business closures, and quarantine interventions factor into the transmission of this deadly virus. Based on observations and data from past pandemics like the Spanish flu and influenza[13-14], we hypothesized that NPIs should significantly contribute as a factor towards preventing the transmission of COVID-19. Therefore, our hypothesis that NPIs significantly contribute as a factor towards preventing the transmission of COVID-19 was accepted

Procedure

We first proceeded to identify valid and complete publicly available data sets. Reliable raw data sets from Keystone Strategies[15] were identified, sorted, grouped and cleaned up to meet the needs of this project. Mathematical analysis was performed and various visualizations for the 7 distinct NPIs available in the Keystone Strategies dataset were created to identify patterns and trends.

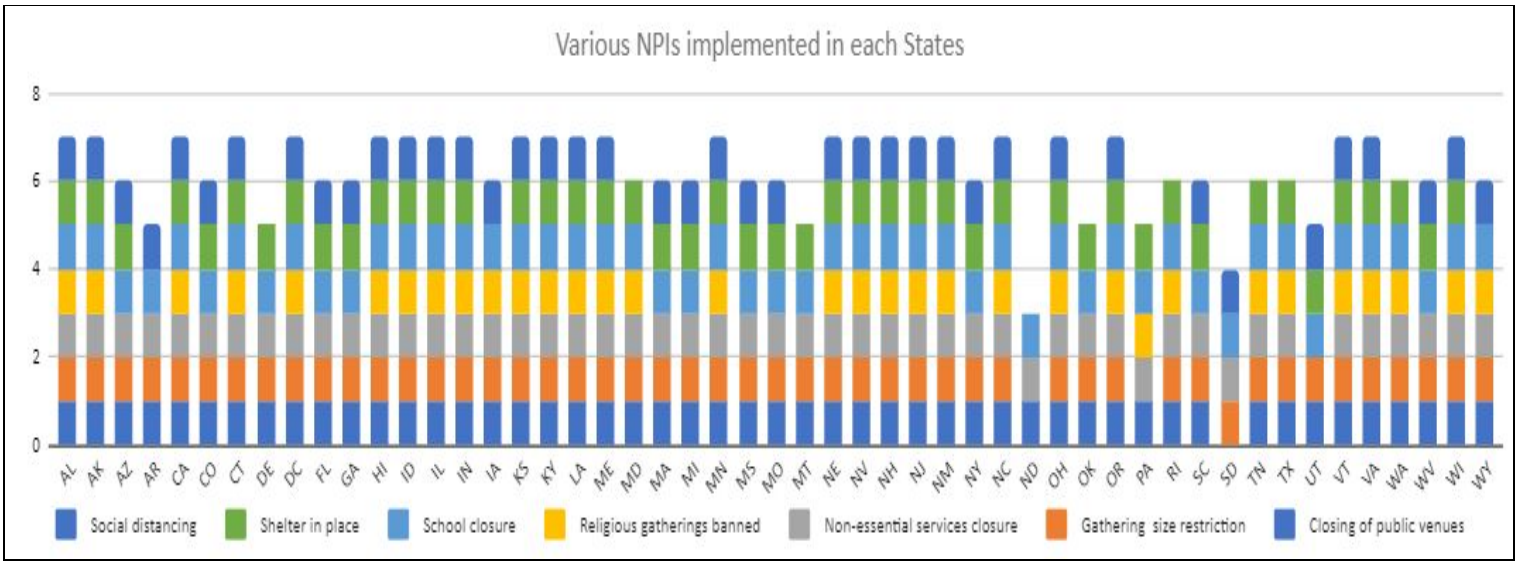


Figure 7: Implementation of 7 NPIs across all 50 US States and District of Columbia

Data indicated that by July 2020, North Dakota had the least number of NPI's implemented - only 3 which included School closures, non-essential services closures and closing of public venues, whereas 25 states – that is 50% of all the states in the US had implemented all 7 NPIs. Additionally, it was observed that 100% of the 50 US states and District of Columbia implemented School closures, while only 64% implemented banning of religious gatherings. Once we had this data, we quickly figured out that we cannot look at the NPI data in isolation. Since NPIs are tied to people and people are tied to location, we started looking at NPI data and Number of deaths by location together.

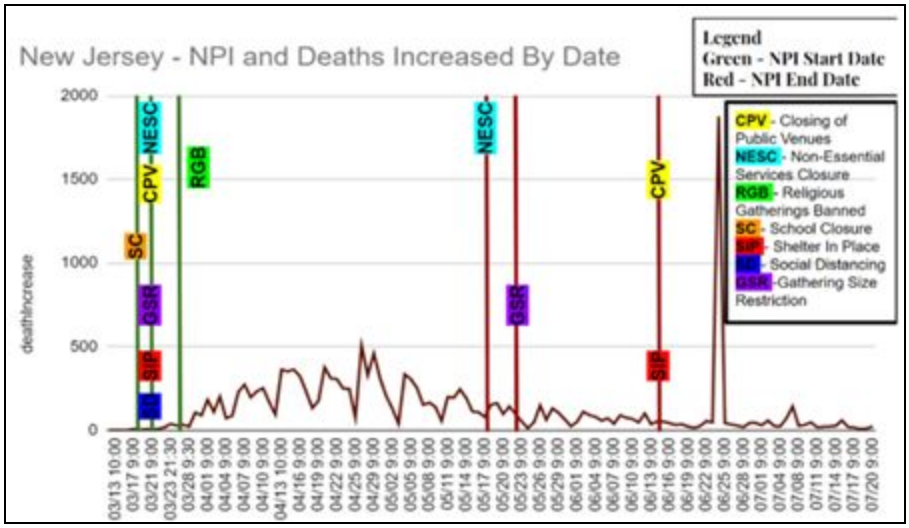


Figure 8: NPI Data overlaid on Location data for the state of New Jersey

As a next step, NPI data was applied to a particular location to see if there was an impact on the total number of deaths where the NPIs were applied. When NPI data was overlaid on top of Location data for the state of New Jersey, a downward trend in the number of deaths between Mar-Jul, 2020 indicated that NPIs could potentially have contributed to the drop in the number of deaths in New Jersey. Similar trends were observed for Nevada, Pennsylvania, and Ohio. Therefore, our hypothesis that NPIs significantly contribute as a factor towards preventing the transmission of COVID-19 was accepted.

Analysis

Since most interventions were implemented in quick succession in all states, it was difficult to clearly identify the individual levels of effect of each intervention. Further literature review of an article published in medRxiv showed that there is a considerable difference in the ways that different states in the US have applied various NPIs[16], which makes it very difficult to predict the separate benefit of each NPI by itself. A study published in *Nature* indicated that there is an 82% reduction in the transmission of the virus in lockdown conditions when compared to the pre-NPI intervention values[17].

Results

Given the complexity of the NPI data and its implementation, we decided to apply the existing literature value of 82%[17], if all NPIs (Social distancing, wearing mask in public places, school closures, mass gathering restrictions, non-essential business closures, Stay at home orders, with exceptions and measures to isolate symptomatic individuals and their contacts) were in place. Therefore, we arrived at the conclusion that once the probability of dying due to COVID-19 was calculated based on age, gender, race and location, the probability would be reduced by 82% of the final calculation if all NPIs were followed. It should be noted that even if any one of the NPIs was not followed, the NPI factor would not contribute towards the final calculation.

USER INTERFACE DEVELOPMENT PROCESS

The User Interface (UI) is a very important part of any application. It brings a view to the application and must be coded in a way that captures the inputs that the user selects and send them to the respective algorithms for the calculations to take place. For the user interface, a sidebar was added, which contained all of the values in the following order: Gender, Age, Race, State (location), and NPIs. A submit button was placed at the bottom of the sidebar, so when clicked, it would load the interactive gauge that displayed your mortality risk level. On the main screen, we added our title for the project: "AI-19 Advisor, Fighting COVID-19 with Predictive Analytics". A caption was added, which informed the user that it was a COVID-19 mortality risk level predictor. Following that, an interactive table was displayed, which contained all of the user inputs. When the user changes his/her inputs, the values would automatically change in the table, making the application more user friendly. At the bottom of the user interface, an interactive gauge was coded, which appeared only after the user submitted his or her inputs. The gauge displayed the mortality risk level as a percentage and increased in color based on severity. As the percentage increased, the gauge's color rose from yellow to dark orange to red.

Procedure

Our project visualization consisted of multiple parts. Simply put, the first part would be where the user selected his inputs from a set of dropdowns/checkboxes. Those inputs would be sent from the front end to the back end where the calculations would take place. After the calculations, the back end would send the percentage back to the front end for presentation.

The app was created through Streamlit, a free platform used to build data apps and is also compatible with Python, the programming language used throughout the project. We used a variety of tutorials to understand how to download Streamlit and use the platform [18-19]. The code was written using the platform Atom and uploaded on Github. At the start, we ran the

application on localhost using terminal. Afterwards, we deployed the application on Heroku, a free website for app development that supports many languages including Python. The software's compatibility and user-friendly developer tools were the reasons why we selected this platform. All that was required was to connect to the repository on GitHub, which contained the uploaded code for the app.

As previously stated, the code was written using Atom and then ran on localhost. After the first few trial runs were working, the code was placed inside of a GitHub repository. We added the Procfile, requirements.txt, and setup.sh files to the repository. We then created an account on Heroku, where we connected our repository on GitHub to the website and then deployed our application on the web. Since it was our first run, a test url was used, since only a unique url could be used every time. After we finished coding, we repeated this same process for the final product while adding a few touches.

Analysis

The sidebar on the user interface contains the following factors : Gender, Age, Race, Location, Non Pharmaceutical Interventions (NPIs) and a Submit button. It was designed to contain all of the factors in either a dropdown or checkbox format. For all of our factors except for NPIs, dropdowns were used for capturing the inputs. Checkboxes were only used for NPIs. Gender contains a dropdown for two values, male and female. Age contains a dropdown for 8 values, which contain the age groups. For Race, 6 values were displayed, which included: White, Black, Asian, LatinX, American Indian/Alaskan Native and Others. The State dropdown contained a list of all 50 states in alphabetical order. For NPIs, seven checkboxes are used where the algorithm reduces your mortality risk level by 82% if you select all the checkboxes. This percentage was derived from a study of the effects of using NPIs and the impacts they had. When the user selects all the desired inputs and clicks the Submit button, the inputs are sent to the respective algorithms. There, the calculations are performed, and results in a percentage, which is then displayed on the user interface, in the form of an interactive gauge. We originally had the idea of displaying the percentage in a bullet graph, but we decided to use an interactive gauge instead because it was more aesthetically pleasing.

Results

To connect the back end and front end, we ran each of the separate algorithms individually, before implementing them into the front end code. An equation was created, which summed the values of the Gender, Race, Age and Location and formed one percentage. The sum of this value was divided by four to obtain a weighted average. Then the NPI calculation, which reduced the percentage by 82% if all the checkboxes were checked, was applied. It is important to note that even if one checkbox was not checked, the 82% reduction would not be applied. Following this, the percentage was displayed to the user and some feedback was given, telling the user to continue social distancing and wearing a mask in public.

The end product consisted of the sidebar, logo, interactive table and gauge, which displayed the final percentage. The sidebar was where the user selected his inputs, which were sent to the algorithms for calculations. The image was our team logo and a table was displayed underneath it, which showed the users' inputs. They changed automatically as soon as the user selected a different value from the dropdown. Finally, after clicking the Submit button on the sidebar, the interactive gauge appeared, which showed the user what their mortality rate was in the form of a percentage.

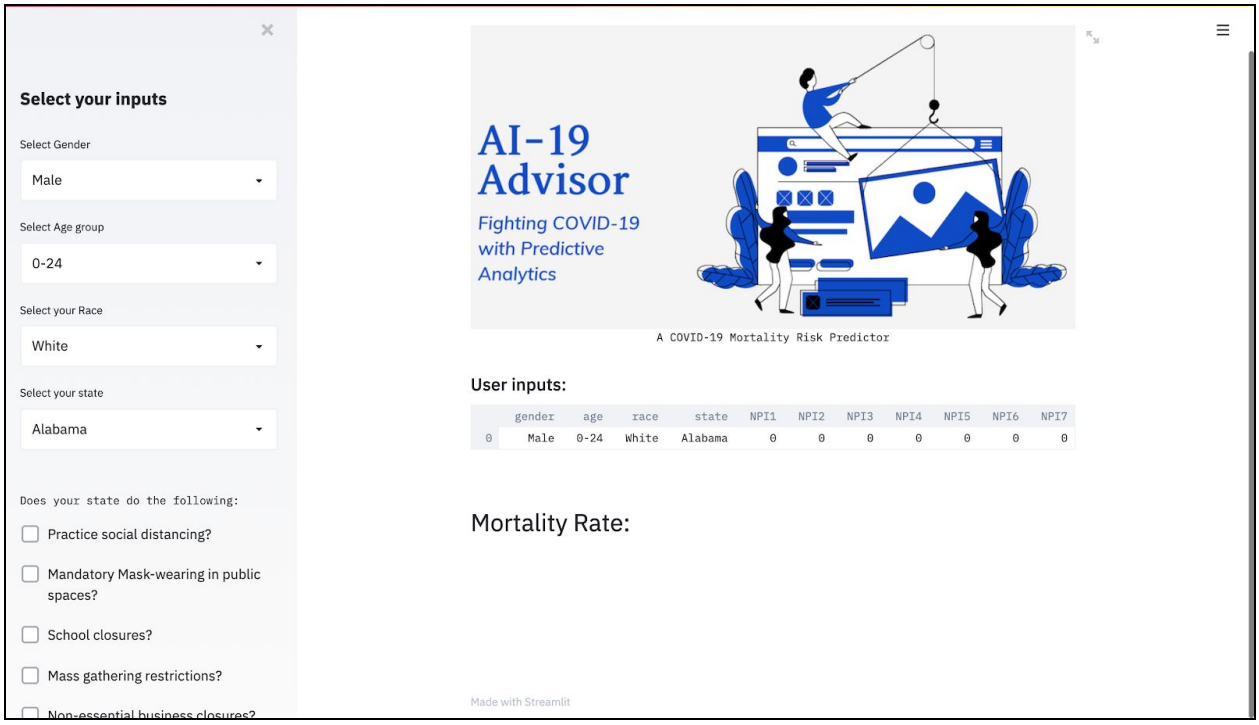


Figure 9: Picture of UI when the user selects his/her inputs

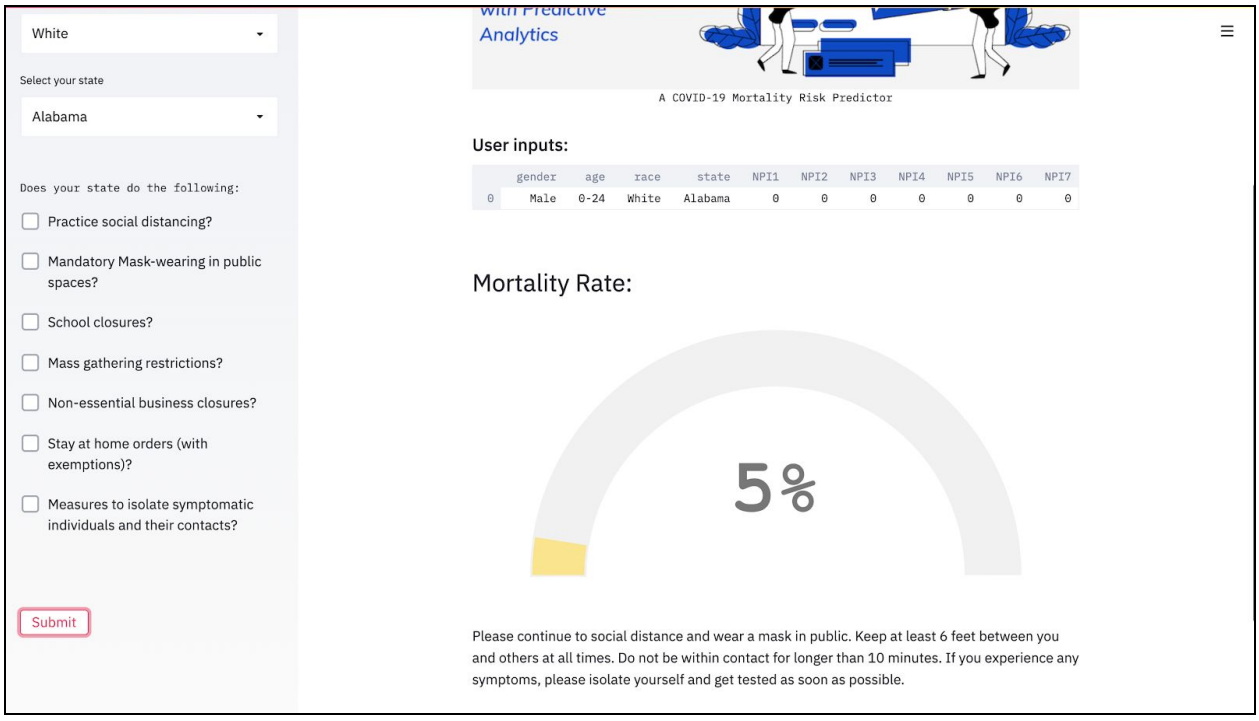


Figure 10: Picture of UI after user clicks Submit

CONCLUSION

Finishing the app, and reviewing the data once more has allowed the team to come to some conclusions about COVID-19 and its effect on groups and the world in general. First, the effects of the virus may last up to the end of this decade, as a result of the temporary worldwide shutdown of the economy. Secondly, the most vulnerable groups of people include Black people, people over the age of 55, people living in densely populated states (such as New Jersey and New York), and of course, those who refuse to follow social distancing guidelines. People who fall under one or more of these categories should be aware of the risk presented to them by the virus.

Our future work includes converting our web interface into a user-friendly and interactive app. We also want to conduct further analysis by overlaying the NPI and location data to see whether the implementation of interventions in certain areas caused increase or decrease in COVID-19 deaths. Additionally, we want to write a program for automatic data extraction, so the data is automatically fed into the algorithms as it is updated. Finally, we would like to implement more advanced data analytics solutions such as deep learning models and Natural Language Processing (NLP) for text-mining. This may enable our model to yield more accurate and reliable predictions. All in all, we hope to keep enhancing our solution and spread awareness about the mortality risk during the pandemic through our app and paper.

REFERENCES

[1] "Coronavirus (COVID-19) frequently asked questions," *Centers for Disease Control and Prevention*. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/faq.html. [Accessed: 20-Aug-2020].

[2]S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, and J. Crowcroft1, "(PDF) Leveraging Data Science To Combat COVID-19: A Comprehensive Review," *ResearchGate*, Apr-2020. [Online]. Available: https://www.researchgate.net/publication/340687152_Leveraging_Data_Science_To_Combat_COVID-19_A_Comprehensive_Review. [Accessed: 20-Aug-2020].

[3] "The COVID Racial Data Tracker," *The COVID Tracking Project*, 22-Jan-2020. [Online]. Available: https://covidtracking.com/race. [Accessed: 20-Aug-2020].

[4]S. H. Golden, "Coronavirus in African Americans and Other People of Color," *Coronavirus in African Americans and Other People of Color | Johns Hopkins Medicine*, 20-Apr-2020. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/covid19-racial-disparities. [Accessed: 20-Aug-2020].

[5]A. Yang, V. Moll-Ramirez, C. Fallon, A. Cruz, and C. Burton, *ABC News*, 03-Jul-2020. [Online]. Available: https://abcnews.go.com/US/covid-19-affecting-latino-community/story?id=71478786. [Accessed: 20-Aug-2020].

[6]E. J. Miller and H. M. Linge, "Age-Related Changes in Immunological and Physiological Responses Following Pulmonary Challenge," *International journal of molecular sciences*, 17-Jun-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5486115/. [Accessed: 20-Aug-2020].

[7]"Provisional COVID-19 Death Counts by Sex, Age, and Week," *Centers for Disease Control and Prevention*, 15-May-2020. [Online]. Available: https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-W/vsak-wrfu/data. [Accessed: 20-Aug-2020].

[8]J. Drees, "Gender roles and the coronavirus: Why women may have higher risk of catching COVID-19. Some health experts are worried that the coronavirus pandemic may present a higher risk for women due to the gender roles they play in society, according to The New York Times.," *Becker's Hospital Review*, 13-Mar-2020. [Online]. Available: https://www.beckershospitalreview.com/care-coordination/gender-roles-and-the-coronavirus-why-women-may-have-higher-risk-of-catching-covid-19.html. [Accessed: 20-Aug-2020].

[9]Kate Bradford, Tahra Johnson; Alise Garcia. State Action on Coronavirus (COVID-19), 19 Aug. 2020, www.ncsl.org/research/health/state-action-on-coronavirus-covid-19.aspx.

[10] McCaughey, Betsy. "COVID-19 Death Rates Reveal the States That Failed the Test: New York and New Jersey." New York Post, New York Post, 13 Aug. 2020, nypost.com/2020/08/12/covid-19-death-rates-reveal-ny-and-nj-are-states-that-failed-the-test/.

[11] J. Drees, "Gender roles and the coronavirus: Why women may have higher risk of catching COVID-19. Some health experts are worried that the coronavirus pandemic may present a higher risk for women due to the gender roles they play in society, according to The New York Times.," *Becker's Hospital Review*, 13-Mar-2020. [Online]. Available: https://www.beckershospitalreview.com/care-coordination/gender-roles-and-the-coronavirus-why-women-may-have-higher-risk-of-catching-covid-19.html. [Accessed: 20-Aug-2020].

[12]"Nonpharmaceutical Interventions (NPIs)," *Centers for Disease Control and Prevention*, 27-Apr-2020. [Online]. Available: https://www.cdc.gov/nonpharmaceutical-interventions/index.html. [Accessed: 20-Aug-2020].

[13]R. J. Hatchett, C. E. Mecher, and M. Lipsitch, "Public health interventions and epidemic intensity during the 1918 influenza pandemic," *PNAS*, 01-May-2007. [Online]. Available: https://www.pnas.org/content/104/18/7582. [Accessed: 20-Aug-2020].

[14]M. D. Howard Markel, "Nonpharmaceutical Interventions Implemented by US Cities During the 1918-1919 Influenza Pandemic," *JAMA*, 08-Aug-2007. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/208354. [Accessed: 20-Aug-2020].

[15] Keystone-Strategy, "Keystone-Strategy/covid19-intervention-data," *GitHub*. [Online]. Available: https://github.com/Keystone-Strategy/covid19-intervention-data/tree/master/archive. [Accessed: 20-Aug-2020].

[16]H. M. Korevaar, A. D. Becker, I. F. Miller, B. T. Grenfell, J. E. Metcalf, and M. J. Mina, "Quantifying the impact of US state non-pharmaceutical interventions on COVID-19 transmission," *medRxiv*, 01-Jul-2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.06.30.20142877v1. [Accessed: 20-Aug-2020].

[17]S. Flaxman, S. Mishra, A. Gandy, J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, and S. Bhatt, "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe," Nature News, 08-Jun-2020. [Online]. Available: https://www.nature.com/articles/s41586-020-2405-7. [Accessed: 20-Aug-2020].

[18]"Get started¶," *Get started - Streamlit 0.65.2 documentation*. [Online]. Available: https://docs.streamlit.io/en/stable/getting_started.html. [Accessed: 20-Aug-2020].

[19]“Tutorial: Create a data explorer app¶,” *Tutorial: Create a data explorer app - Streamlit 0.65.2 documentation*. [Online]. Available: https://docs.streamlit.io/en/stable/tutorial/create_a_data_explorer_app.html. [Accessed: 20-Aug-2020].

APPENDIX A

Code Implemented To Run The Algorithms For Each Respective Factor

[RACE]

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

def race_data(a):
  if (a == 1):
    Y = 0.0042137 + (0.0021101*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
    return(final_race)
  if (a == 2):
    Y = 0.012551 + (0.004345*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
    return(final_race)
  if (a == 3):
    Y = 9.1616e-05 + (0.0024151*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
    return(final_race)
  if (a == 4):
    Y = 0.0038881 + (0.0017926*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
```

```
    return(final_race)
  if (a == 5):
    Y = -0.0065058 + (0.0034341*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
    return(final_race)
  if (a == 6):
    Y = 0.048139 + (0.0014339*21)
    final_race = []
    ans = Y*100
    final_race.append(ans)
    return(final_race)

race = input("Enter your race: \n")
race_date(race)
```

[AGE]

```
def age_data(a,b):
  from matplotlib import pyplot as plt
  import numpy as np

  from sklearn.linear_model import (
    LinearRegression)
  from sklearn.metrics import mean_squared_error
  from sklearn.preprocessing import PolynomialFeatures
  from sklearn.pipeline import make_pipeline

  week = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
  deaths = np.array([1625, 1268, 1178, 1081, 1205, 1324, 1611, 1518, 841]).reshape(-1,1)

  model = make_pipeline(PolynomialFeatures(3), LinearRegression())
  fittingmodel = model.fit(week, deaths) #polynomial regression model
  y_plot = model.predict(week[:, np.newaxis].reshape(-1,1))
```

```
week2 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths2 = np.array([1250, 1043, 931, 910, 1028, 1341, 1547, 1484, 818]).reshape(-1,1)

model2 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel2 = model2.fit(week2, deaths2)
y_plot2 = model2.predict(week2[:, np.newaxis].reshape(-1,1))

week3 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths3 = np.array([1043, 850, 739, 763, 896, 1137, 1325, 1296, 617]).reshape(-1,1)

model3 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel3 = model3.fit(week3, deaths3)
y_plot3 = model3.predict(week3[:, np.newaxis].reshape(-1,1))

week4 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths4 = np.array([571, 564, 493, 512, 568, 728, 835, 754, 391]).reshape(-1,1)

model4 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel4 = model4.fit(week4, deaths4)
y_plot4 = model4.predict(week4[:, np.newaxis].reshape(-1,1))

week5 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths5 = np.array([288, 251, 225, 247, 281, 311, 346, 319, 138]).reshape(-1,1)

model5 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel5 = model5.fit(week5, deaths5)
y_plot5 = model5.predict(week5[:, np.newaxis].reshape(-1,1))

week6 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths6 = np.array([100, 96, 100, 83, 132, 161, 152, 127, 56]).reshape(-1,1)

model6 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel6 = model6.fit(week6, deaths6)
y_plot6 = model6.predict(week6[:, np.newaxis].reshape(-1,1))
```

```
week7 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths7 = np.array([40, 49, 32, 30, 56, 61, 65, 46, 12]).reshape(-1,1)

model7 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel7 = model7.fit(week7, deaths7)
y_plot7 = model7.predict(week7[:, np.newaxis].reshape(-1,1))

week8 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9]).reshape(-1,1)
deaths8 = np.array([8, 8, 10, 13, 21, 24, 19, 15, 4]).reshape(-1,1)

model8 = make_pipeline(PolynomialFeatures(3), LinearRegression())
fittingmodel8 = model8.fit(week8, deaths8)
y_plot8 = model8.predict(week8[:, np.newaxis].reshape(-1,1))

if(b == "06"):
  x = 0 #AUGUST IS REPRESENTED BY WEEK 1
  j8 = 1293 # total deaths(June)for 0-24 years age group
  j7 = 1508 # total deaths(June)for 25-34 years age group
  j6 = 1991 # total deaths(June)for 35-44 years age group
  j5 = 3669 # total deaths(June)for 45-54 years age group
  j4 = 7733 # total deaths(June)for 55-64 years age group
  j3 = 11398 # total deaths(June)for 65-74 years age group
  j2 = 13621 # total deaths(June)for 75-84 years age group
  j = 16768 # total deaths(June)for 85 and Above years age group
elif(b== "07"):
  x = 4 # JULY IS REPRESENTED BY WEEK 5
  j8 = 1167 # total deaths(July) for 0-24 years age group
  j7 = 1352 # total deaths(July)for 25-34 years age group
  j6 = 1871 # total deaths(July)for 35-44 years age group
  j5 = 3382 # total deaths(July)for 45-54 years age group
  j4 = 7571 # total deaths(July)for 55-64 years age group
  j3 = 11033 # total deaths(July)for 65-74 years age group
  j2 = 13168 # total deaths(July)for 75-84 years age group
  j  = 16191 # total deaths(July)for 85 and Above years age group
elif(b == "08"):
  x = 8 # AUGUST IS REPRESENTED BY WEEK 9
```

```
    j8 = 560 # total deaths(August) for 0-24 years age group
    j7 = 786 # total deaths(August)for 25-34 years age group
    j6 = 1156 # total deaths(August)for 35-44 years age group
    j5 = 2081 # total deaths(August)for 45-54 years age group
    j4 = 4902 # total deaths(August)for 55-64 years age group
    j3 = 7702 # total deaths(August)for 65-74 years age group
    j2 = 10023 # total deaths(August)for 75-84 years age group
    j = 12249 # total deaths(August)for 85 and Above years age group

  if(a>=0 and a <=24):
    z8 = model8.predict(week8[x].reshape(-1,1)) # predicting COVID-19 death counts for the corresponding
month
    probability8 = (z8/j8)*100 #total deaths - (total deaths - predicted COVID-19 death counts )
    print("Age Probability of Dying from COVID-19:", probability8, "%")
    return(probability8)
  elif(a>=25 and  a<=34):
    z7 = model7.predict(week7[x].reshape(-1,1))
    probability7 = (z7/j7)*100
    print("Age Probability of Dying from COVID-19:", probability7, "%")
    return(probability7)
  elif(a>=35 and  a<=44):
    z6 = model6.predict(week6[x].reshape(-1,1))
    probability6 = (z6/j6)*100
    print("Age Probability of Dying from COVID-19:", probability6, "%")
    return(probability6)
  elif(a>=45 and a<=54):
    z5 = model5.predict(week5[x].reshape(-1,1))
    probability5 = (z5/j5)*100
    print("Age Probability of Dying from COVID-19:", probability5, "%")
    return(probability5)
  elif(a>=55 and a<=64):
    z4 = model4.predict(week4[x].reshape(-1,1))
    probability4 = (z4/j4)*100
    print("Age Probability of Dying from COVID-19:", probability4, "%")
    return(probability4)
  elif(a>=65 and a<=74):
```

```
    z3 = model3.predict(week3[x].reshape(-1,1))
    probability3 = (z3/j3)*100
    print("Age Probability of Dying from COVID-19:", probability3, "%")
    return(probability3)
  elif(a>=75 and a<=84):
    z2 = model2.predict(week2[x].reshape(-1,1))
    probability2 = (z2/j2)*100
    print("Age Probability of Dying from COVID-19:", probability2, "%")
    return(probability2)
  elif(a>=85):
    z = model.predict(week[x].reshape(-1,1))
    probability = (z/j)*100
    print("Age Probability of Dying from COVID-19:", probability, "%")
    return(probability)

month = input("Enter the month in the format: 08\n")
age = int(input("Enter your age: \n"))

a = age_data(age, month)
```

[GENDER]

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

def gender_data(a):

 if (a == 1):
    val1 = 419.58 + 2088.8*(9)
    val2 = 419.58 + 2088.8*(8)
    val3 = ((val1-val2)/val2)*100
    final_gender = []
    final_gender.append(val3)
```

```
    print(final_gender)
    return(final_gender)
 if (a == 2):
    val1 =  917.36 + 1757.2*(9)
    val2  =  917.36 + 1757.2*(8)
    val3 = ((val1-val2)/val2)*100
    final_gender = []
    final_gender.append(val3)
    print(final_gender)
    return(final_gender)

gender = input("Enter your gender: \n")
g = gender_data(gender)
```

[LOCATION]

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

def location_data(b):

   ask_state = input("What state do you live in?\n")

   state_abbrev = {"Alabama" : "al", "Alaska" : "ak", "Arizona": "az", "Arkansas": "ar", "California":
"ca","Colorado": "co","Connecticut": "ct", "Delaware": "de","Florida": "fl","Georgia":"ga","Hawaii":
"hi","Idaho":"id","Illinois": "il", "Indiana": "in","Iowa":"ia", "Kansas": "ks","Kentucky": "ky","Louisiana":
"la","Maine": "me", "Maryland": "md", "Massachusetts": "ma", "Michigan": "mi", "Minnesota": "mn",
"Mississippi":"ms","Missouri": "mo", "Montana": "mt","Nebraska": "ne","Nevada": "nv","New
Hampshire": "nh","New Jersey": "nj","New Mexico": "nm","New York": "ny","North Carolina":
"nc","North Dakota": "nd","Ohio": "oh","Oklahoma":"ok","Oregon": "or","Pennsylvania": "pa" ,"Rhode
Island": "ri","South Carolina": "sc","South Dakota": "sd", "Tennessee": "tn", "Texas": "tx","Utah":
"ut","Vermont": "vt","Virginia": "va", "Washington": "wa","West Virginia": "wv", "Wisconsin": "wi",
"Wyoming": "wy"}

   state = state_abbrev.get(ask_state)
```

```
    #state = input("Please enter the abbreviated form of your state in a lowercase format: \n")
    csv_url = "https://covidtracking.com/api/v1/states/"+state+"/daily.csv"

    totaldeaths = {"al": 29611, "ak": 2068, "az": 38945, "ar": 17293, "ca": 153910 , "co": 23447, "ct": 18372,
"de": 5377, "fl": 121463, "ga": 47625, "hi": 6074, "id": 7719, "il": 65073, "in": 37404, "ia": 16424, "ks":
14288, "ky": 25005, "la": 27401, "me": 7952, "md": 31280, "ma": 38972, "mi": 58931, "mn": 25157, "ms":
18973, "mo": 34319, "mt": 5319, "ne": 8935, "nv": 14682, "nh": 7046, "nj": 55887, "nm": 10282, "ny":
66126, "nc": 42875, "nd": 3780 , "oh": 65811, "ok": 20335, "or" : 19284, "pa" :75308 , "ri": 5900, "sc":
29412, "sd": 4293, "tn": 41191, "tx": 117602, "ut": 10735, "vt": 3201 , "va": 39602, "wa": 31261, "wv":
9353, "wi": 29580, "wy": 2499} #Total Number Of Deaths For Each State

    df = pd.read_csv(csv_url)

    df = df.apply(pd.to_numeric,errors='coerce') # Making sure all NaN values are set to 0
    df = df.replace(np.nan,0)

    rawx = df.get("date")
    resX = rawx[::-1]
    XX = np.array(resX)
    x = XX.reshape(-1,1)

    rawy = df.get("death")
    Y = rawy[::-1]
    y = np.array(Y)

    date_pred = int(b)

    clf = LinearRegression()
    clf.fit(x,y)

    val1 = clf.predict([[date_pred]])
    a = (totaldeaths.get(state))
    final_location = val1/a*100
    print("\nLocation percentage is " + str(final_location) + "%")
```

```
    return(final_location)

month = input("Enter the month in the format: 08\n")
date = input("Enter the day of the month:\n")
year = input("Enter the year: \n")

strval = year+month+date
location_data(strval)
```

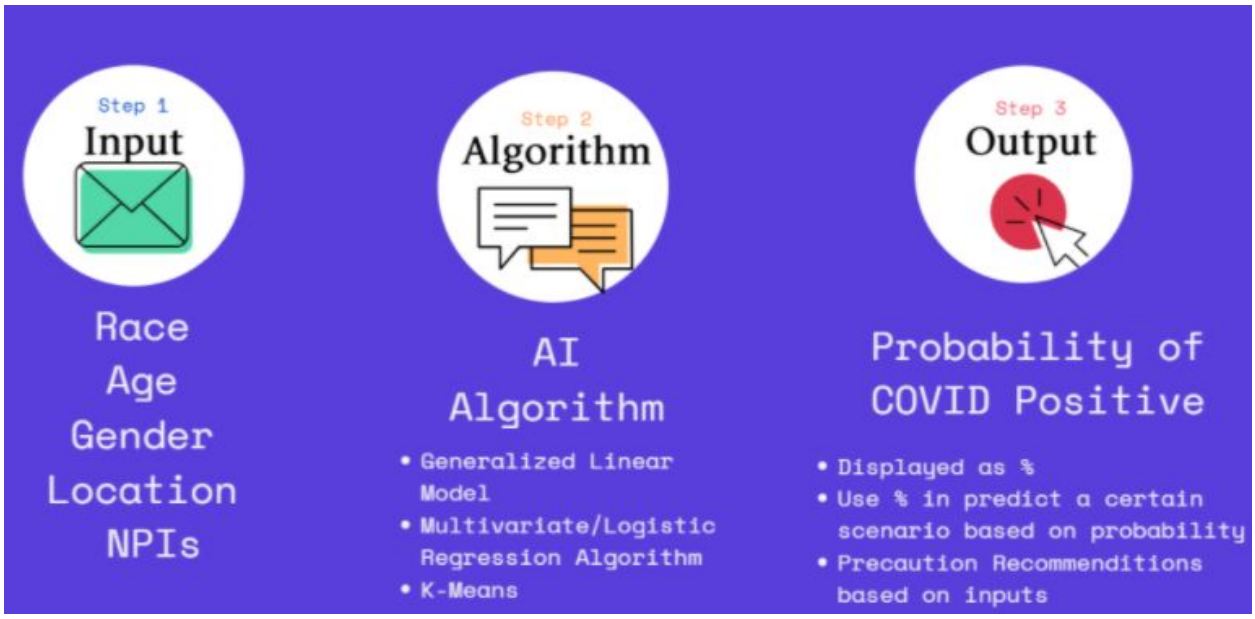
[NPI]

No code was developed, as we decided to use existing literature values to apply on top of our final product.

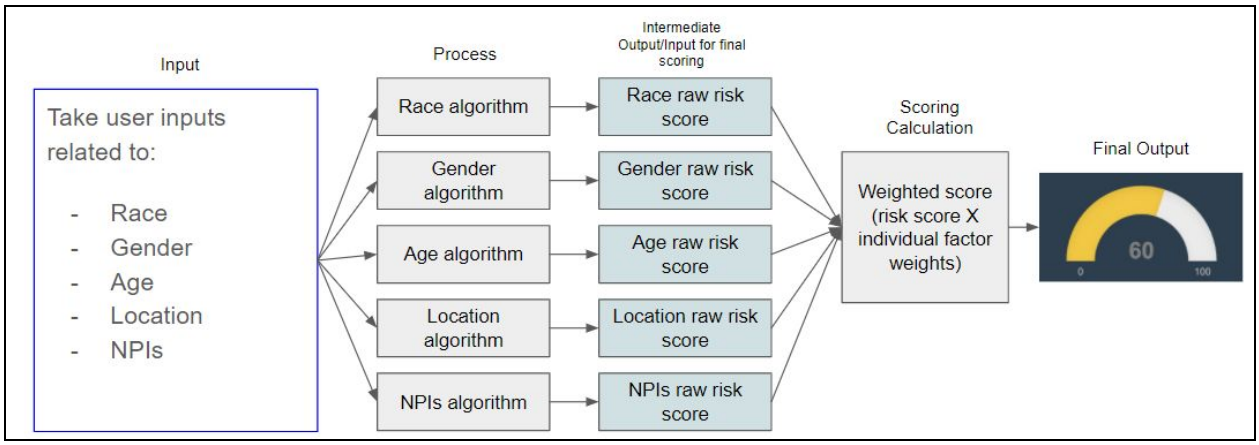
Appendix B

Project Visualizations: Overall, Front-end, and Back-end

Overall Project Visualization



Front-end Visualization



Back-end Processing