**DATA SCIENCE TASKS TO CONSIDER**

1. Clearly state the goals for the project as well as the hypothesis you hope to prove (or disprove).

The coronavirus COVID-19 pandemic is the defining global health crisis of our time and the greatest challenge we have faced. I collected other respiratory disease death rates to see the impact COVID-19 is threatening our lives by comparing with other disease death rates and chose the States with high COVID-19, Pneumonia, and Influenza mortality rate, which are California, Massachusetts,New Jersey, New York, New York City, Pennsylvania.

I would like to compare death rate of these disease death rate by age, race, and sex and want to find any interest in a relationship between those respiratory diseases according to the provisional respiratory death counts datasets.

1. Data access: download the data set into your R environment (do not use the Data Import feature in RStudio, but instead write the R code for data access).

Read csv files of provisional death count by state, race, and age

Merge those two files to see the plots by both ract and age.

1. Use the data repository to get a definition of all the variables in the dataset. Perform feature engineering to select variables that support your hypothesis.

State

Age

Race

Sex: White, Black, Hispanic, Asian, and More race

Covid: Death count by COVID-19

Tot\_death: Total death count

Pneumonia: Death count by Pneumonia

Pneum\_covid: Death count by Pneumonia and COVID-19

Influenza: Death count by Influenza

Pneu\_flu\_covid: Death count by Pneumonia, Influenza, and COVID-19

1. Perform any data transformations you feel are necessary to achieve the desired goals.

Convert Class of Variables

* + 1. as.factor(character columns)
    2. as.numeric(integer columns)

Merge by States

* + 1. Provisional Death Counts by State, Age, and Sex
    2. Deaths involving Coronavirus disease by Race and Hispanic origin and Age, by State

Remove Outliers & NAs

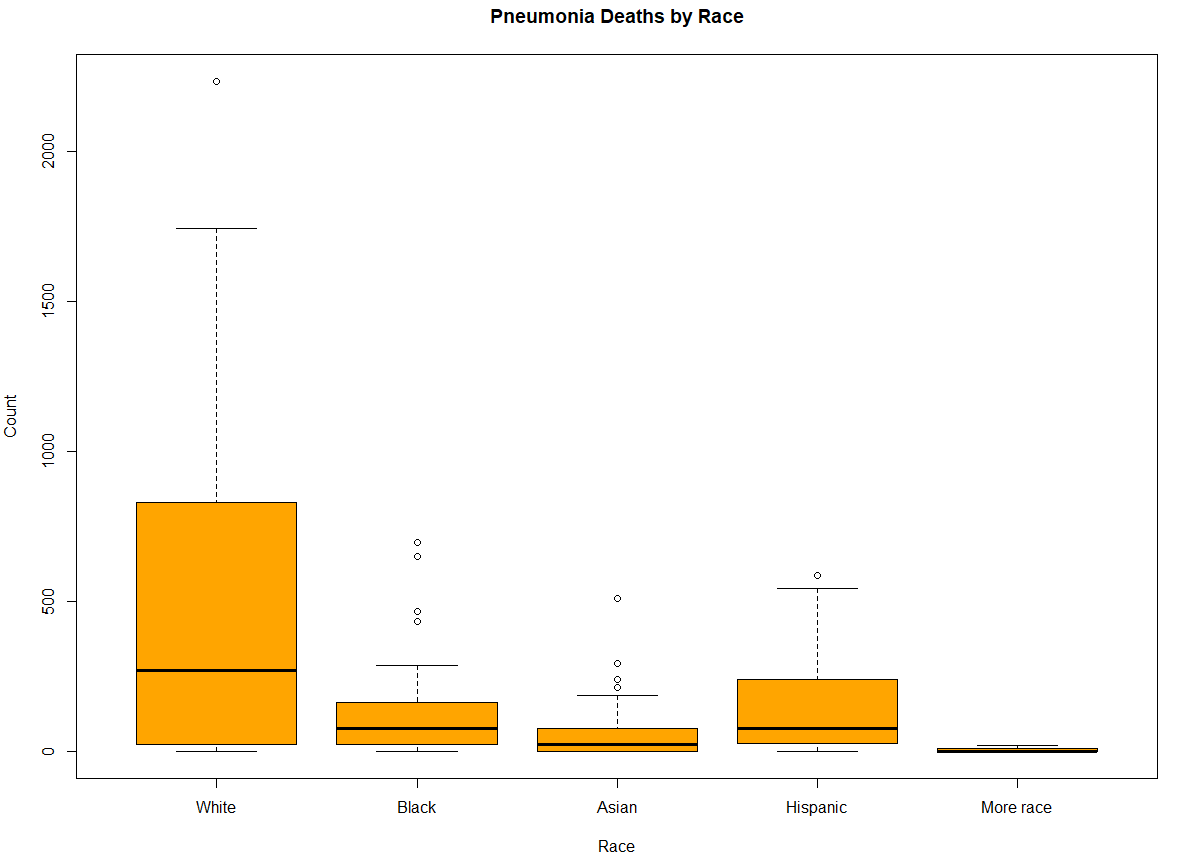
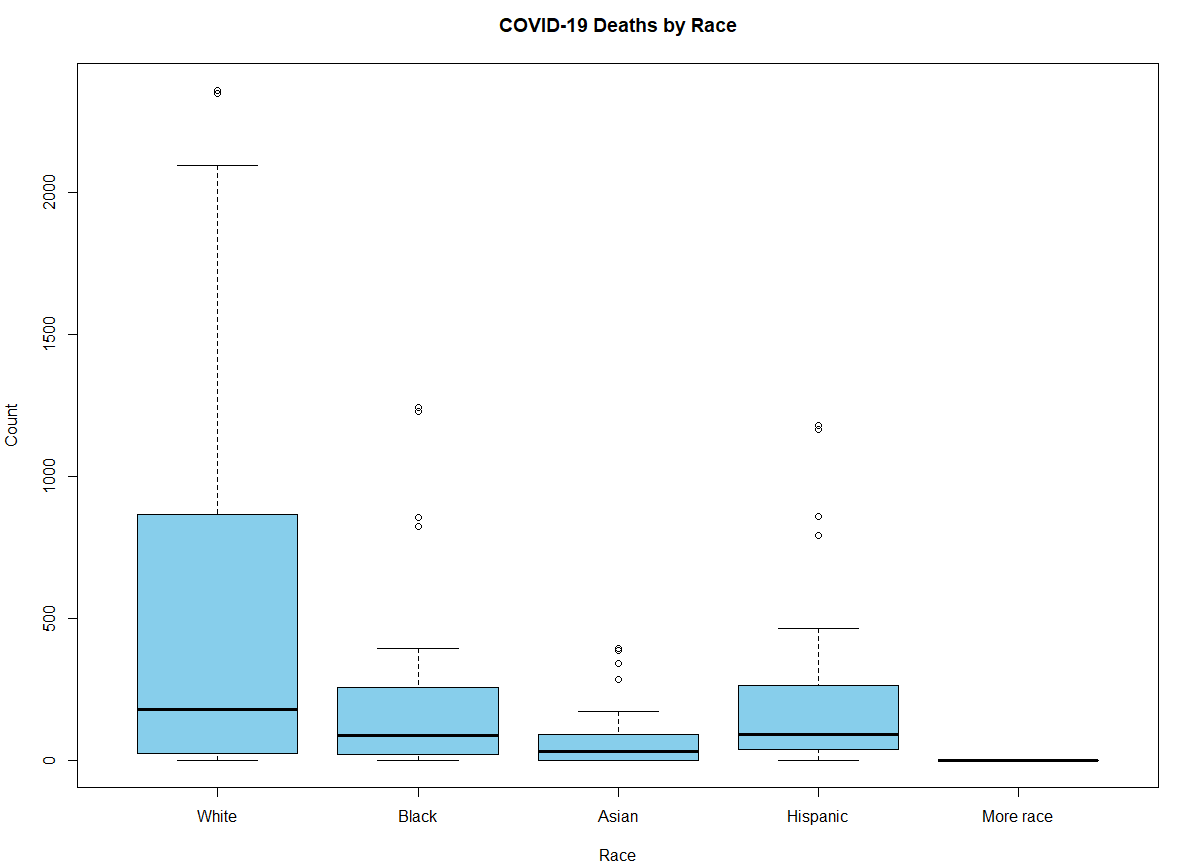
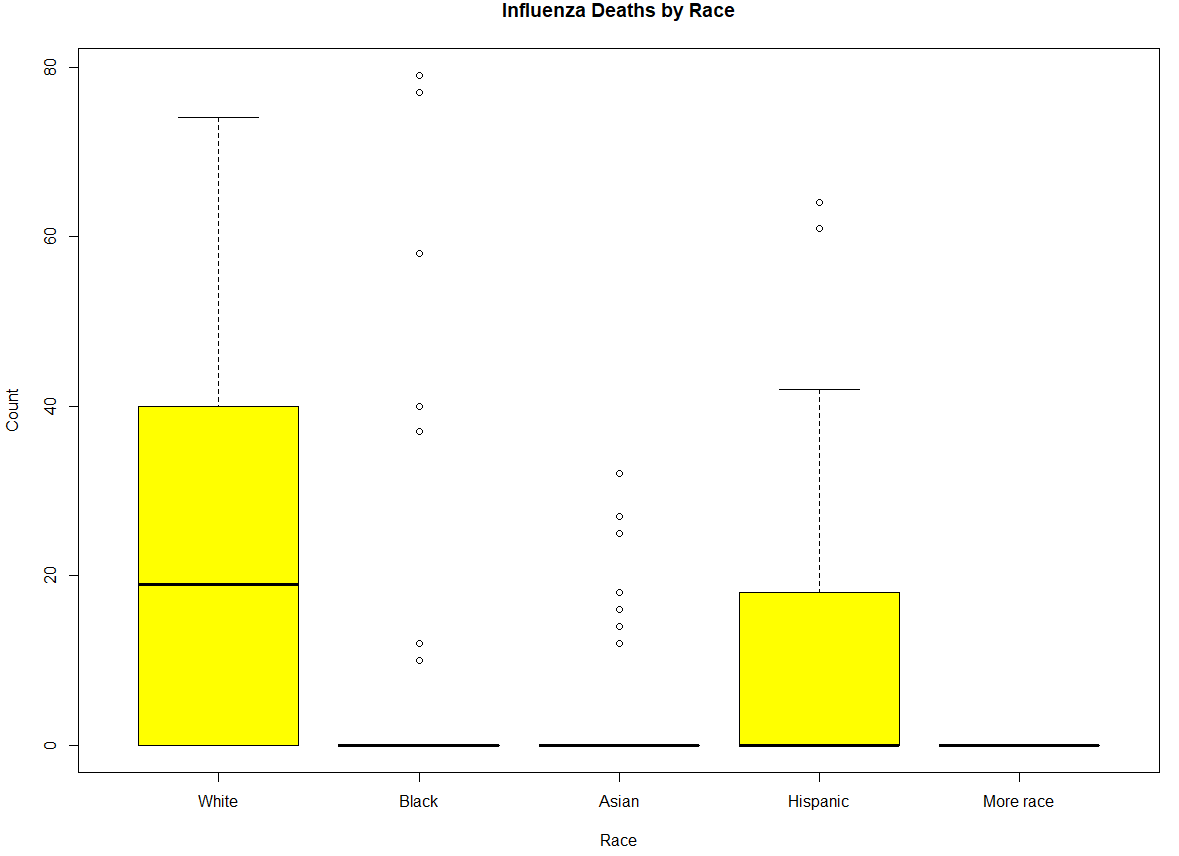
* + 1. Age (1 - 24 years)
    2. States

Re-order & Visualize

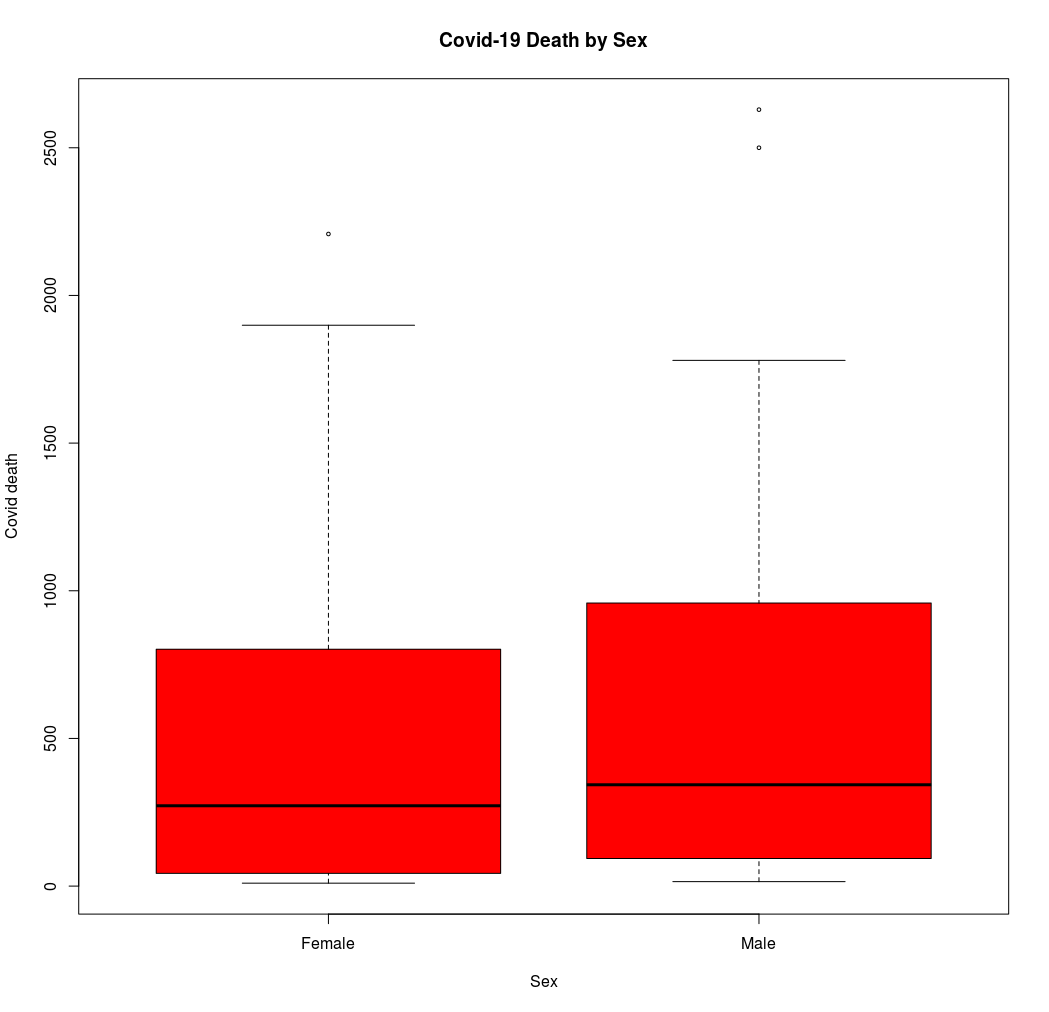
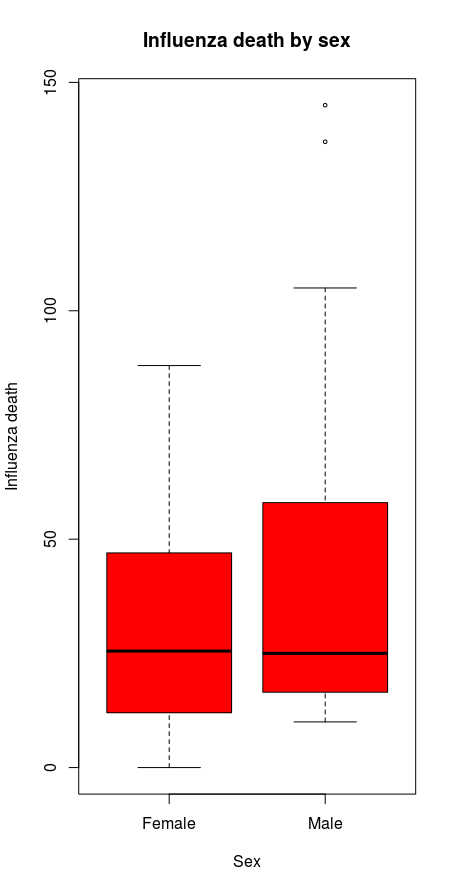
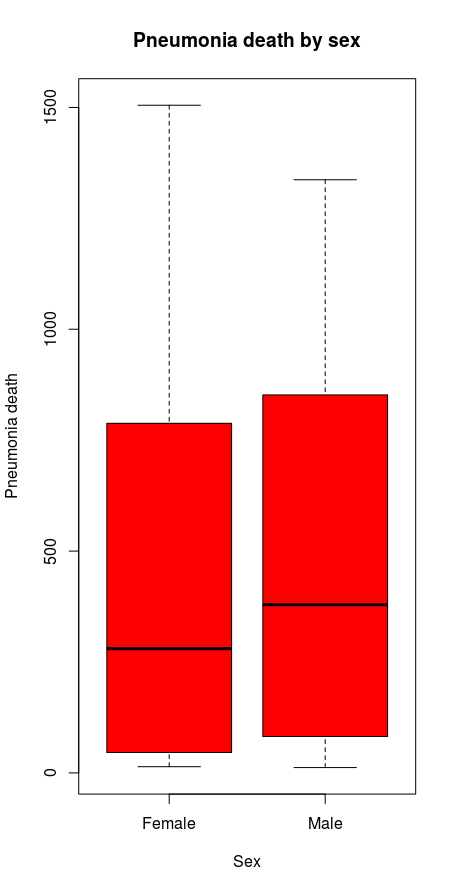
* + 1. Race Column
    2. Ordered by Age

1. Use various EDA and simple statistical analysis techniques to gain a deep understanding for the data.
2. Use R’s plotting features to produce both exploratory and expository data visualizations.

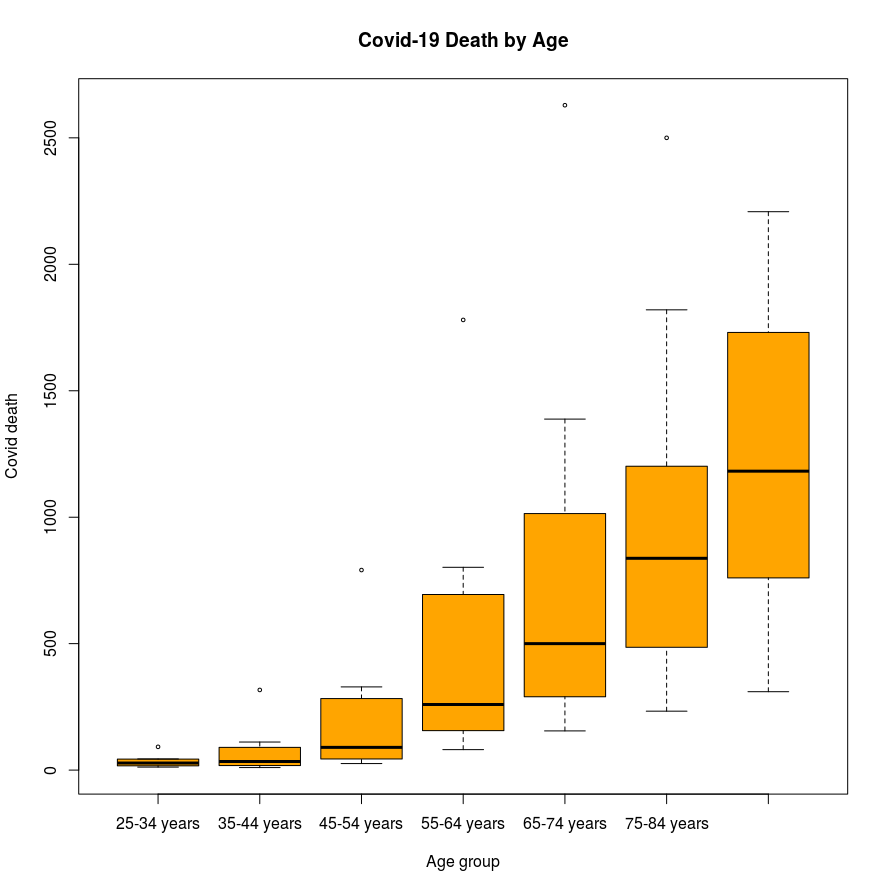
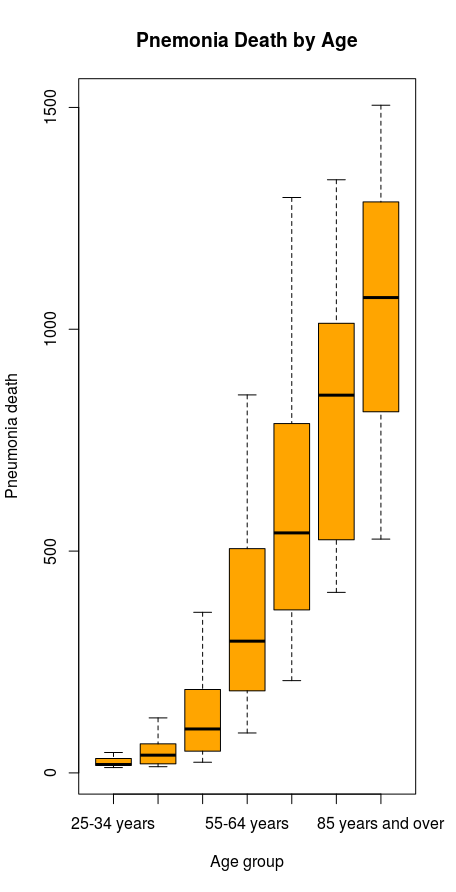
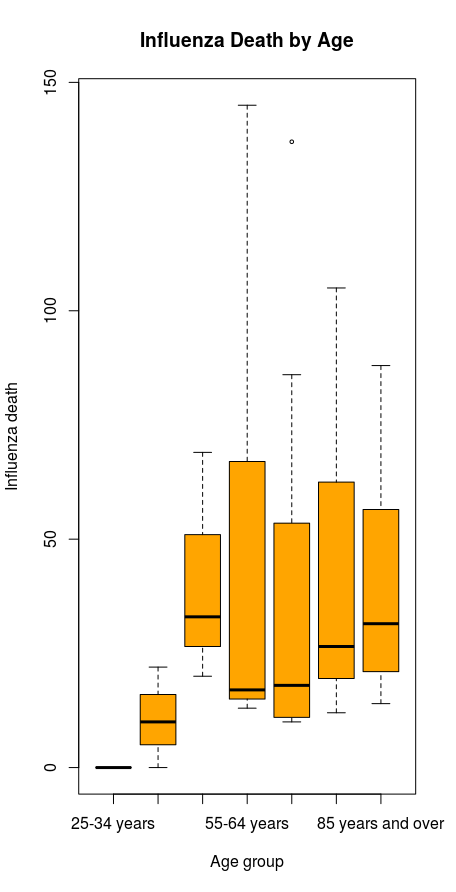
Box Plot (Deaths Counts by Race)



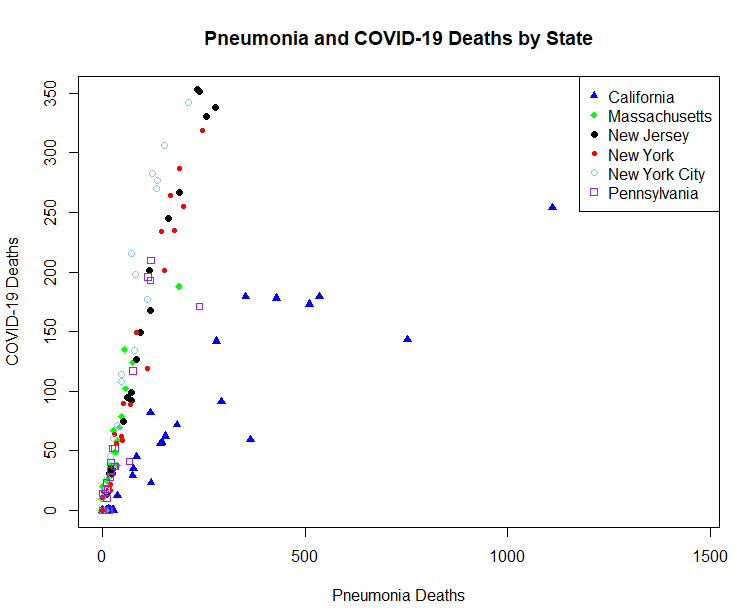
Box Plot (Deaths Counts by Sex)



Box Plot (Deaths Counts by Age Group)



Scatter Plot (Relationship betweem Pneumonia & COVID-19 Deaths by State)



1. Select one or more of R’s statistical learning algorithms to make predictions, and/or discoveries.
2. In the case of predictions, use the trained algorithm on new data and make a case for the algorithm’s accuracy.
3. Prepare a report using techniques of “data storytelling” to present the results to a management-level audience – state the goals of the project, the data sets used, EDA results, data visualization, overview of how you used machine learning algorithms, and final conclusions.
   1. Pneumonia death has a relationship with Covid-19 death. That is, we can deduce that pneumonia has an impact on Covid-19. However, influenza death does not have any relationship with Covid-19 death. The projection is based on if the current preventions (such as, at least 6 feet distance, face covering, and hand washing) are done. The numbers look good for all the states; however, Pennsylvania needs to be careful with the various high and low points on their plot.
4. Create your own GitHub repository and publish the results of the project:

project description, R script, final report, and all data sets used.