Week8.3_FinalProject1_KoppulaVeera

Veera Koppula

July 30 2021

##Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?

We live in interesting times with a pandemic that affects the way of life of every living being on the globe. Every government and public agency has tried to either slow down spread of Covid-19 by travel restrictions/lockdowns/masking & social distancing policies or mitigate its impacts by vaccination campaign. I would like to explore impact of factors that could impact the outcome of Covid infection. (as sex,age, comorbities etc)

##Draft 5-10 Research questions that focus on the problem statement/topic.

What are the possible data sources to explore, showing the Covid-19 case history/Trend? What are the various conditions that impact the spread of the Covid-19 virus? How is the data collected by identified sources? Incase of any missing elements/gaps in data, how are they mutated? Which geographical region is a best representation for sample analysis? are there any comorbities for the patient? what is the age/sec/Race that defines the patient?

##Provide a concise explanation of how you plan to address this problem statement.

I would like to validate different data sources,to collect the information on Covid-19 infection/case rate(Concentrating on USA), Collect different mitigation actions taken to slow the spread of infections and Collect vaccination rate for the same period/regions. Once this data is collected, I would like to look at the geographical region that has best representation by completeness and cohesive data. Once the sample is identified, would like to preform calculations on correlating factors and as well try to identify a model that would should best possible predictors for infection spread rate. This calculated model should identify the best possible predictors that would determine the outcome of virus infection.

##Discuss how your proposed approach will address (fully or partially) this problem.

as described above, the calculated model should give us the best predictors to calculate the possible outcome of virus infection. This should help understanding the correlation of seriousness of virus outcomes, this would help health agencies to take specific actions to address or prioritize any specific demographic that could be impacted by the virus more than others.

##Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets) #Original source where the data was obtained is cited and, if possible, hyperlinked.

Source1 - CDC covid data - https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf Source2 - JohnHopkins Covid Data - https://coronavirus.jhu.edu/us-map Source3 - Worldometer- https://www.worldometers.info/coronavirus/country/us/

#Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.). CDC Covid Data - CDC is the main health policy maker and premier disease research institute in USA. CDC collected the Covid-19 data reported by different

county and state health departments. CDC collects that data to track/understand/predict and issue guidances for various infectious diseases spread across US. CDC has collected various factors of data including demographics, any exposure history, disease severity indicators and outcomes, presence of any underlying medical conditions and risk behaviors. Most of this data is collected from various county/city/state reporting agencies that reported information to CDC. Incase of missing data CDC made notes and provided explanation on gaps in data.

JohnHopkins Data - JohnHopkins is a leading philonthropic organization that provides suggestions and inputs on critical public health policies. JohnHopkin collects its data from various public reported sources of county/city/state data sources. They collate different sources, analyzes and provides trends as well on Covid-19 positivity rate, death rate,vaccination, hospitalization including the civic mitigating actions taken by states (https://coronavirus.jhu.edu/data/state-timeline/new-confirmed-cases/alaska)

Worldometer- Worldometer is a independent run website that collects information from different public sources and manually compiled to validate. The data contains the number of cases deaths, active cases, recovered, total tests and test/1M, population and sources. Website also indicates projectons on how cases could have an outcome based on few factor. This website also cleary indicates the gap in data or facotrs impacting the data collection due to any change in policies(https://www.worldometers.info/coronavirus/about/)

##Identify the packages that are needed for your project.

Following are some of the packages that I would need for this analysis: knitr ggplot2 readxl readr purrr RcmdrMisc MASS lattice graphics grDevices dplyr datasets car carData olsrr

##What types of plots and tables will help you to illustrate the findings to your research questions?

Following plots: histogram (with Kernal Density plot) boxplot scatterplot (with Matrices) line chart

also would build tables using data frame, while laoding data into data frame using read csv and xl. also use str and head to analyze the content of data.

##What do you not know how to do right now that you need to learn to answer your research questions?

I am able to dip my toes into the modeling and test of the models built. I need to learn more on differnt regression modelling to identify correlation and confounders that could affect (using different types of variables)