# Week10_FinalProject3_KoppulaVeera

Veera Koppula

August 10 2021

##Overall, write a coherent narrative that tells a story with the data as you complete this section. #Introduction. Analyzing data and drawing useful information has always been a passion of mine. With the interesting times we are living in, with a global pandemic, I have been trying to follow various publicly collected data sources to see if any interesting conclusions can be drawn by analyzing the massive amounts of data being collected by various public sector and private sector sources.

#Summarize the problem statement you addressed.
For this Assignment, I have picked a specific problem statement related to covid infections. I wanted to identify if there is any relationship to serious infection outcomes as hospitalization/ICU admittance or Death, in relation to patient's age, race/ethnicity or underlying medical conditions.

#Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented). For addressing the above problem statement, I have looked at three public data sources from CDC, Johnhopkins and worldometer. After reviewing through the data, I have picked Covid infections data collected by CDC as most coherent to address my problem statement. This data set had various fields including the date patient was identified as first infection, and other aspects as age group, ethnicity/race and sex of the patient, along with the serious outcomes as if the patient was hospitalized, admitted into ICU or had passed away. This dat set consisted of 27.9M rows, after scrubbing the data removing all the rows that have missing elements or Unknown entries from all this dataset, I was left with 481,381 rows. I have calculated a new field that is set as Yes/No based on any serious outcome of the infection faced by patient. I have performed various graphs & scatter plots to see individual attributes as Age or Ethnicity/Race or Sex or Underlying conditions impact on the resultant serious outcomes. Based on the analysis and data type, I suggest a multiple regression of serious outcome prediction, using Age, Ethnicity/Race,Sex and underlying conditions as predicting variables.

#Summarize the interesting insights that your analysis provided. Based on the data split, graphs and analysis here are the insights I see from the data, Infected people who are either 0-9 yrs or 70+yrs old have higher probability of serious outcomes due to infection, which is in general expected as kids or advanced age patients have weakened immune system - that results in a serious outcome due to most of the infections Most of the infected population with serious outcomes due to infection appeared to be white/Caucasian. Infected population having serious outcome seemed to be equally distributed among Male and female. There seems to be very strong correlation to serious outcomes with underlying co-morbidities.

when data was validated, it looked like as the pandemic progressed from January 2020 to recent Jul 2021, looks like the serious outcomes due to infection have reduced. Also, post scrubbing most of the data seemed to be centered largely around White/Caucasian ethnicity.

#Summarize the implications to the consumer (target audience) of your analysis. One of the key stakeholders consumers for my analysis would be public health policy makers. Government entities or organizations that propose health policies or influence public health decisions. Based on the above analysis some of the key implications are, Covid is an equal opportunity offender, with no distinction of sex of the person. there is a higher probability of serious outcomes for the people with weakened immune system, such as infants/toddlers or advanced age. People with underlying conditions have highest probability of serious outcomes.

Also, based on the analysis that most of the recent data has less serious outcomes, suggests that some of the actions that public organizations have taken has been helping in reducing the serious outcomes.

#Discuss the limitations of your analysis and how you, or someone else, could improve or build on it. Here are some of the limitations that I see with my analysis and how it could be improved.

My scrubbing resulted in close to 99% of data being removed. A data wrangling analysis to add missing elements, would have helped in having higher data coverage. Based on the data scrubbing, most of the data left out is centered around White/Caucasian ethnicity, this could be a resultant of most coherent data being collected only in most of white populated areas. A data cleanup/correction exercise to increase the inclusion of other race/ethnicity would help for a better distribution of data to have better predictions. Dataset considered for analysis could be improved by overlaying another data set that shows the mitigation actions taken by public officials to reduce the spread of Covid infection and vaccination efforts that might have helped in reducing the seriousness of infections. Joining My current analysis has been looking at single prediction variable of serious outcome, using a single predictor. A multiple variable prediction could give probably better indication on the factors that could influence seious outcomes. This would need building regression model and predicting the outcome using combination of multiple variables as predictor.

#Concluding Remarks Based on the analysis, most of the insights support current public health policy communications. Covid infections have serious implications on people who are either immuno-compromised or with weakend Immunity system. Covid infections serious outcomes have a strong correlation with underlying health conditions. There are various opportunities to improve the analysis done in addresing the problem statement, to get more in depth insights and possible actions based on those insights. Some of these opportunities are outlined in limitations section with potential corrective actions.