Covid-19 Impact Analysis

Kesav Adithya Venkidusamy

Statistical Hypothesis/Question:

Coronavirus disease or COVID-19 is a global pandemic infectious disease caused by virus called sars-cov-2. Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. **Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are most likely to develop serious complications from COVID-19 illness.**

The center for disease control and prevention (CDC) is the national public health agency of the United States. The agency's main goal is the protection of public health and safety through control and prevention of disease, injury, and disability in US and worldwide. CDC plays an essential role in the response to COVID-19. The agency collects the data on regular basis and provide for public use. Among numerous datasets available in CDC, below are the ones considered for analysis

1. Provisional COVID-19 deaths by week, sex and age
2. Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21

Outcome of your EDA

The outcome of my EDA resulted in rejection of the null hypothesis that the high covid death count for older people having age greater than 55 compared to young people whose age is less than 55 was due to chance. It also resulted in rejection of null hypothesis that the high covid death count for people with underlying condition compared to healthy people was due to chance.

In addition, I could see that the correlation between old aged people and Covid death is high compared to the correlation between young aged people. Interestingly, I also noticed that the correlation between the people without underlying condition and Covid death is high compared to those having underlying condition.

What do you feel was missing during your analysis?

I really wish that I would have looked for more numerical information to include my analysis. However, covid deaths is the only numerical variable available in the dataset. Moreover, the age has been given as range or group instead of individual value. If it was given in as individual value, I would have been able to create better regression model.

Were there any variables you felt could have helped the analysis?

Vaccine information is missing in the dataset. If vaccine information was provided, it would have been helped to perform analysis on vaccine impact to covid deaths and would have been used to find the correlation between two. I also would like to compare the Covid-19 impact by state.

Were there any assumptions made you felt were incorrect?

I assumed that correlation between the people with underlying condition and covid deaths would be high compared to healthy people and covid deaths. However, this is not correct as correlation between healthy people and covid death is high compared to other.

What challenges did you face, what did you not fully understand?

The data contains duplicates across group, state and age group. It was bit challenging at the initial stage to clean and filter the data that are required for analysis. Moreover, the format of the variable or column “Covid Death” was incorrect. So, required to change that too before considering for analysis. Additionally, the number of deaths reported by CDC may not be accurate. Only the deaths occurred in hospital certified by doctors are reported in the data set. I personally feel it does not provide complete or thorough accounting for all adverse events.

Other thoughts?

I really enjoyed working on this project applying all the techniques that I have learnt throughout this course. It was great experience working with different sets of large datasets. Overall, based on the analysis of CDC dataset used, I would say covid-19 impact was high to old people and people with underlying compared to young and heathy people. However, it is worthy to extended the analysis/investigation with more reliable dataset.