# Project: COVID death rate and World Health Statistics

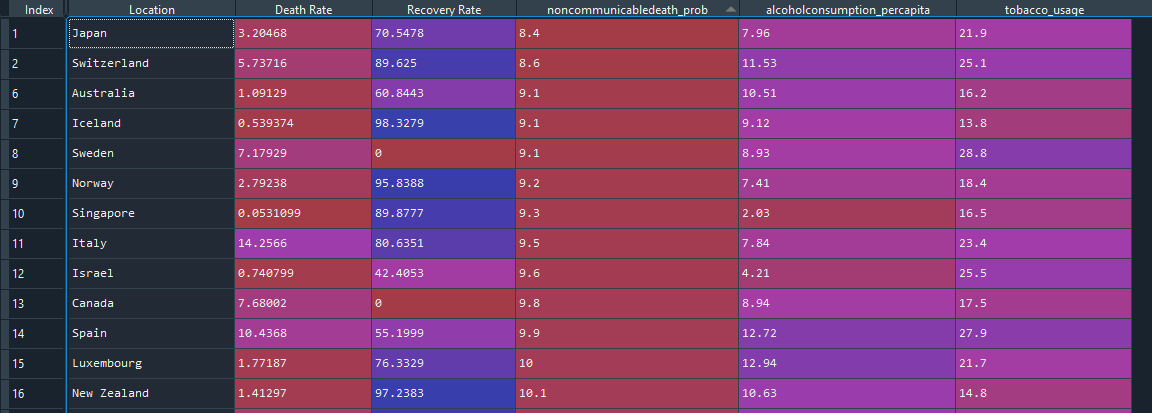
# by Jyothi Samudrala

# Outcome of my EDA:

To get my dataset ready for detailed analysis, I did the following clean up:

1. In World Health Statistics datasets, I selected only certain years to remove “location” duplicates (having multiple years for the same country). The assumption is that the statistics may not have changed in a span of 2 to 3 years. Example: Probability of Death due to non-communicable disease like cancer in 2016 and 2018 are similar.
2. In covid-19 dataset, created a death rate column; this will be derived by dividing total covid deaths by total covid cases.
3. In covid-19 dataset, created a recovery rate column; this will be derived by dividing total recoveries by total covid cases.
4. Checked for outliers – thankfully, the dataset had no outliers based on descriptive statistics and histogram visuals.
5. Merged all four datasets with Country as the Primary Key, and included only the necessary columns/variables.
6. There were no null values or duplicates to remove in the final dataset.

The final data set looked like so:



# What do you feel was missed during the analysis?

Post project and analysis, I realized:

1. The dataset is NOT large enough.
2. The data may be suffering from measurement scale issues – We are assuming that all the variables came to us normalized between values of 1 and 100.
3. I think I could have created multiple data frames instead of just one, to analyze relation between each world health statistic and covid death rate separately. The results would have been more precise and accurate.
4. Doing a similar analysis on recovery rate could have shed more light on the world health statistics in general.

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

Due to only choosing certain variables (with a preconceived notion that certain health statistics affect covid death rate more), the R-Squared value was low. Having other variables could have helped increase R-Squared, giving a better understanding of covid death rate. So, my initial questions could have been framed better with more domain knowledge/research.

# Were there any assumptions made you felt were incorrect?

For T test and Z test, I assumed data normality. I am not so certain of this. Also, covariance and correlation numbers are based on a very small dataset (we only have so many countries), so I am not sure the numbers are entirely reliable.

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

I do not fully understand the following:

1. Was I unable to use PMF because of plotting a wrong scenario? Was it accurate of me to not rely on PMF because my data is continuous?
2. Plotting PDF and CDF in the context was a little challenging.
3. Choosing a test statistic for Hypothesis testing was a little challenging as well because I was not sure my assumptions (mentioned in the power point) are accurate.