Project Milestone 5

DSC 540

Abed Tabbalat

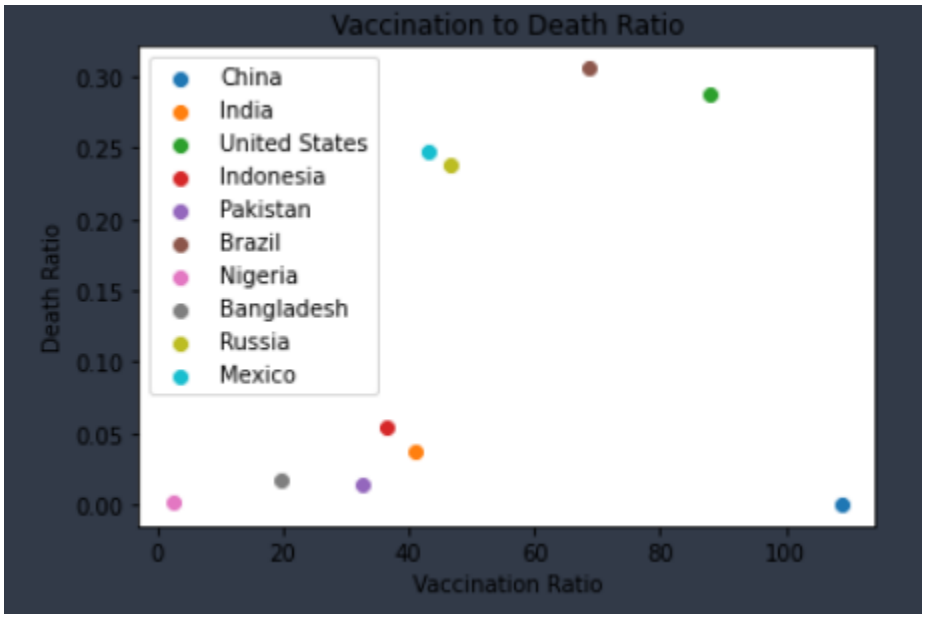
The project has been successfully completed. During the time spent on taking different sources, combining them, and merging them together in database. I have learned plenty of new skills that has reshaped my views on how to wrangle data.

While wrangling the data, I have come across a few obstacles that got me scratching my head into how to make this work. The idea was to grab all variables related to covid-19 and try to see what the ratios look like to the population of each country. I have determined that my primary key is the country. However, there were situations where each dataset was using different labels to certain countries. For example, 2 of the datasets used 'United States' and one used 'US'. In the merge process, I had to go back and do a validation check on all the countries.

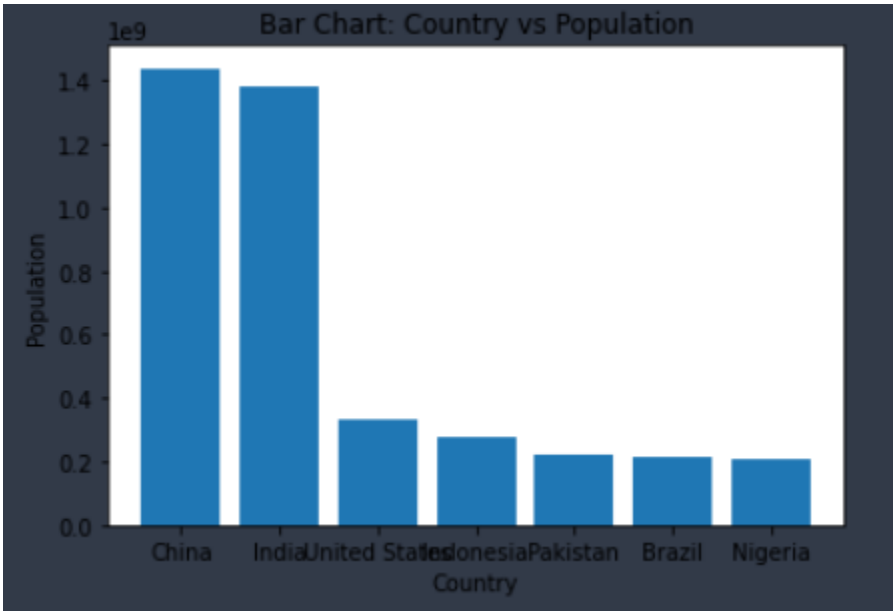
Another complication was the data provided from each source had more countries than others, which lead to the decision any field that has missing data, the entire country will be dropped out of the dataset. I believe having full data from the source is the most accurate way to view data, up to a certain threshold. If there are more missing data than existent data, the entire project should be shutdown. In my situation, that wasn't the case.

I have computed 5 visualizations that I will give a summary of what they show.

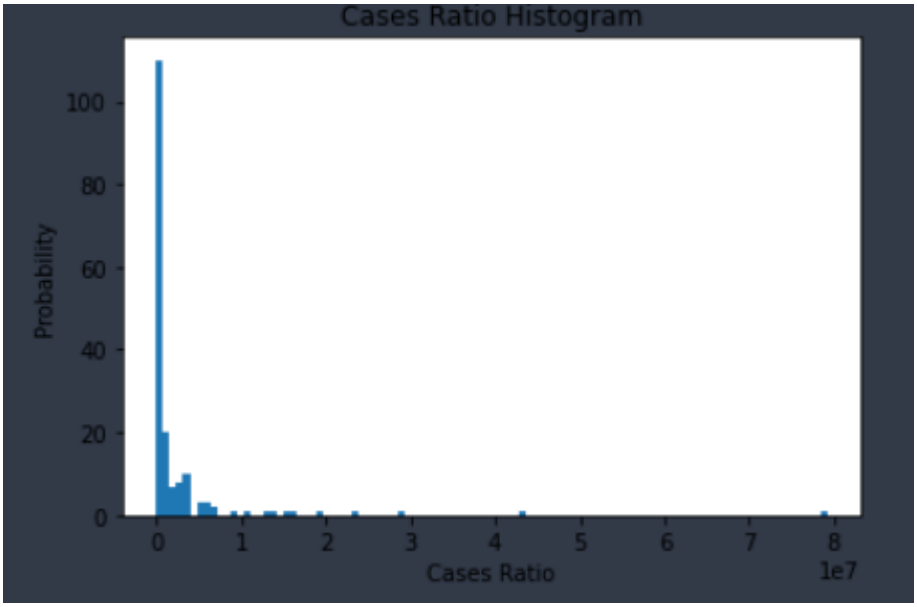
Below shows the first visualization:



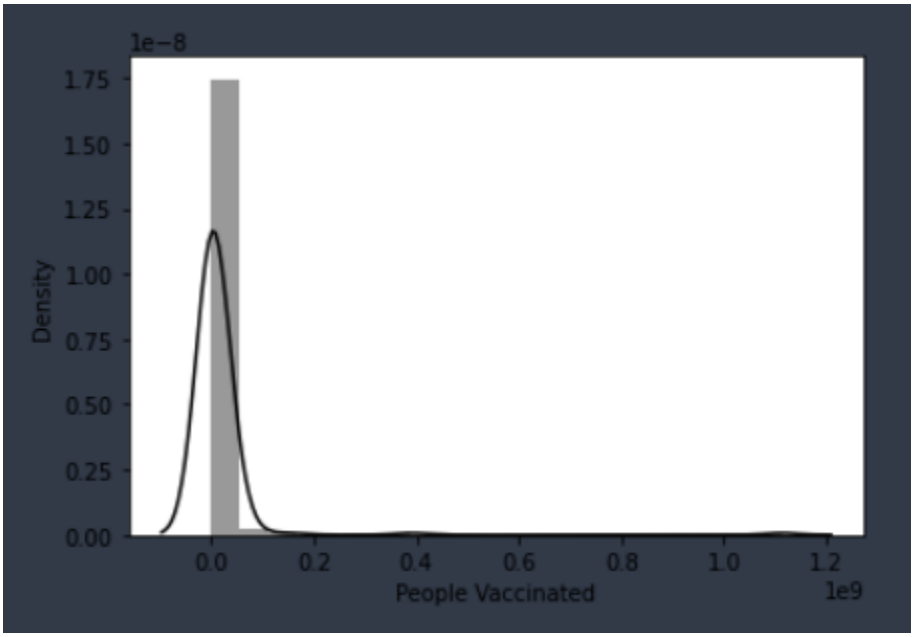
Below shows the second visualization:



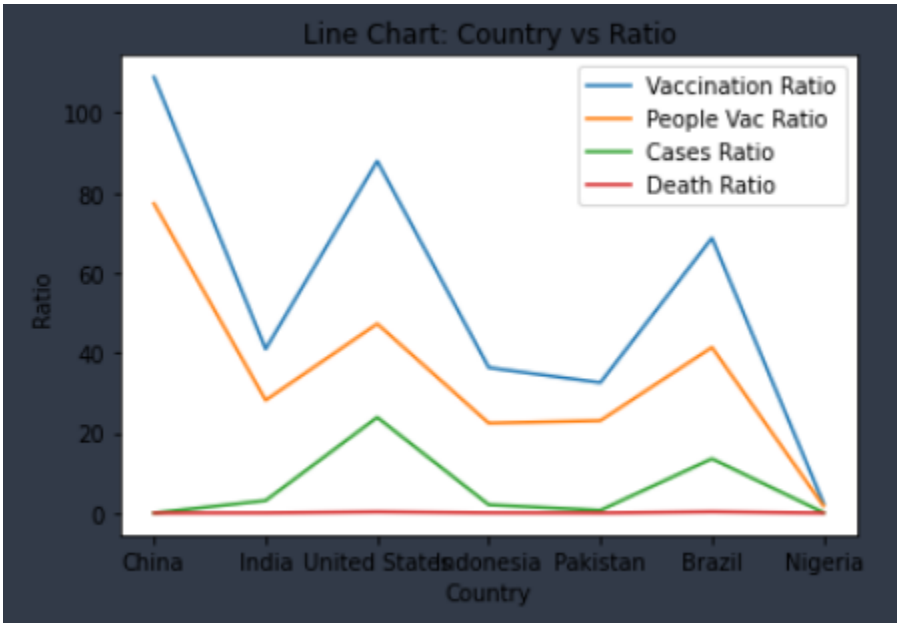
Below shows the third visualization



Below shows the fourth visualization:



Below shows the fifth visualization:



I would like to steer the focus on the last visualization since it combines all the ratios calculated for every variable that has been done for this exercise and gives a clearer picture of what is going on with the covid-19 situation.

We can see the number of Vaccines and people vaccinated are lower in high populated 3rd world countries and this makes sense as their technological advancements would be behind developed countries. In addition, we need to factor in the data collection for this exercise coming from certain countries could be inaccurate, like Nigeria for example. Everything points to a low ratio in Nigeria which doesn't make any sense, so we can count it as an outlier for minimal data has been provided.

The death ratio seems to be flat in all regions, meaning the covid-19 cause of death impact now is much better than when the pandemic started, as we have advanced research going behind that for the past 2 years that got us to this point.