Exploratory Data Analysis

COVID-19

Abdullah Bushnag

Database

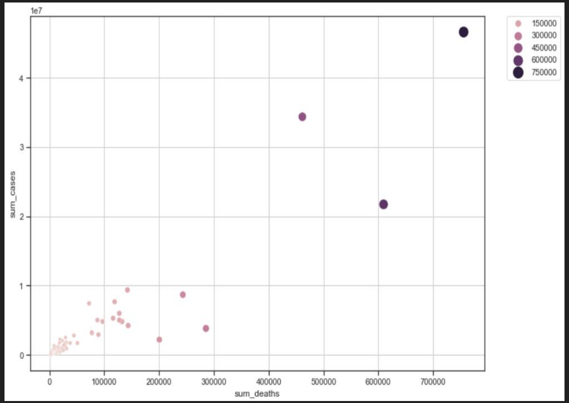
- CSV into the SQL SERVER caused a few problems: Wrong data-types causing issues with queries, Nulls values also affecting queries and finally string characters of nominal type.
- The SQL columns data-types were converted from nvchar to float64, Nulls were replaced with Zeros and finally a mapping from the nominal-string type to nominal-int type.
- The Nulls were converted to zero because they are not saying there is no data at this point but it rather that there was no cases at this point.
- The steps were done firstly on a test database then on the actual data. Then pushed to the SQL server where the EDA processes was possible now that the database is fixed.
- The data had one large table so self-join was applied.
- The data contains 140k+ points increasing weekly with 65 features going back to the beginning of the pandemic.

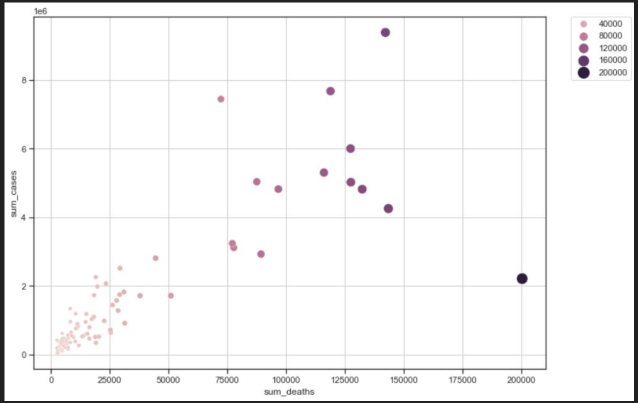
Dataframes

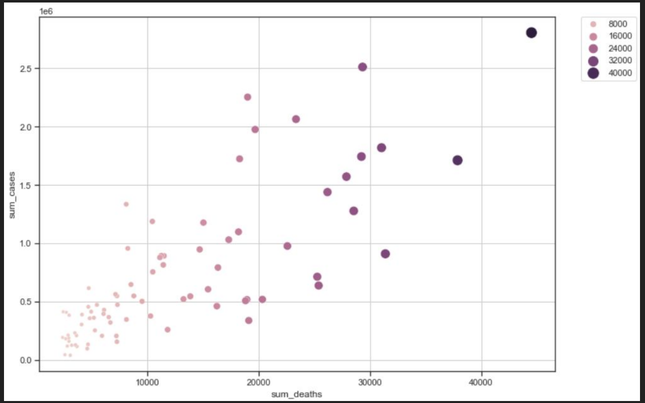
- The pandas dataframe has a to_SQL function which once connected to the database can be use to run queries normally.
- First view of the data showed that the dataframe has unwanted columns due to the fixing of the database in the previous step. So the first step was to remove that column.
- When checking for data-types and null values no issues were faced as they were resolved in the previous step
- Then the number of unique data-types for each columns checked so are the values. Which showed that some had a finite continuous-like values.
- Some countries in the location column were not countries such as: Europe,Africa..etc and they to be removed.
- Based on the questions decided on. The two extreme groups were picked:Top(x) and Bottom(x)

Top 100 countries sorted by number of deaths.

1. USA is the highest in number of deaths followed by Brazil and India.
2. The scatter plot shows a trend that the increase in cases, generally increases the number of deaths.
3. The scatter plot show that some countries with the same number of cases has less deaths.
4. The outliers are not affecting the data generalization.

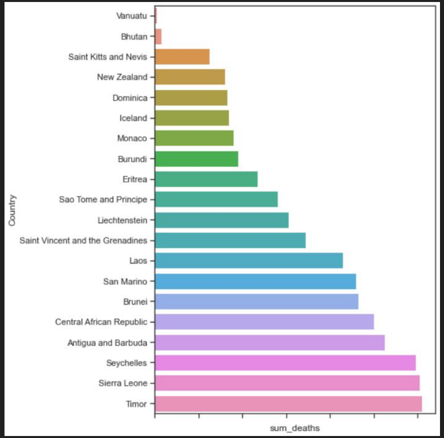


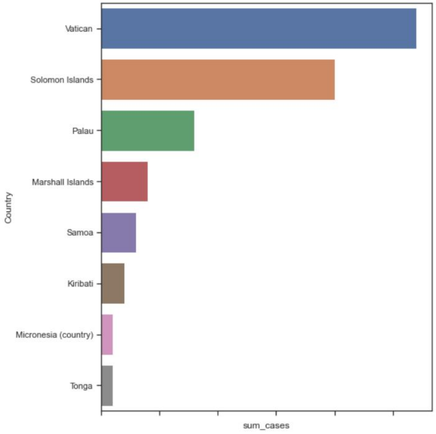


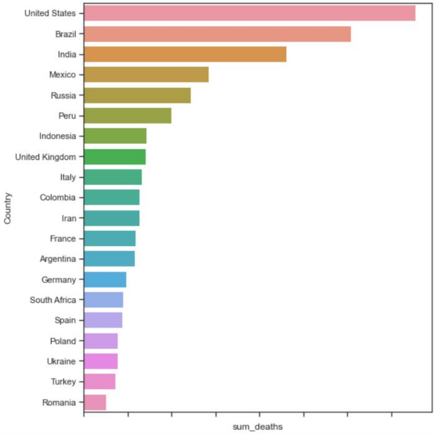


Which countries has the most and least Deaths?

1. The country with most cases is India
2. The country with the most deaths is USA
3. The country with NO cases is Macao, Jersey, Bermuda, Bonaire Sint Eustatius and Saba...etc
4. The country with the MOST cases and NO deaths is Vatican.
5. The country with the LEAST cases and NO deaths is Tonga
6. The country with the LEAST cases and has deaths is Vanuatu
7. The country with the highest death percentage/ratio to the population is Montenegro

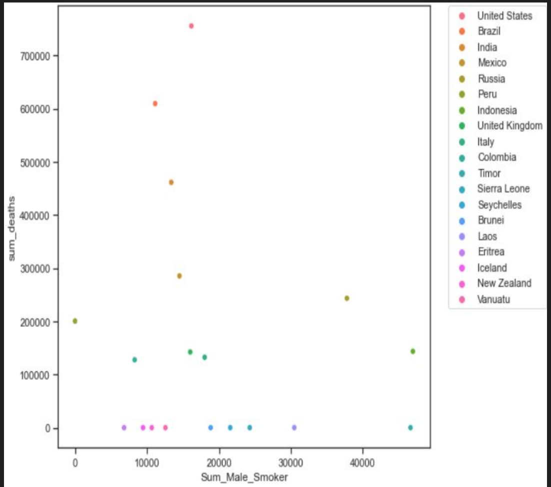


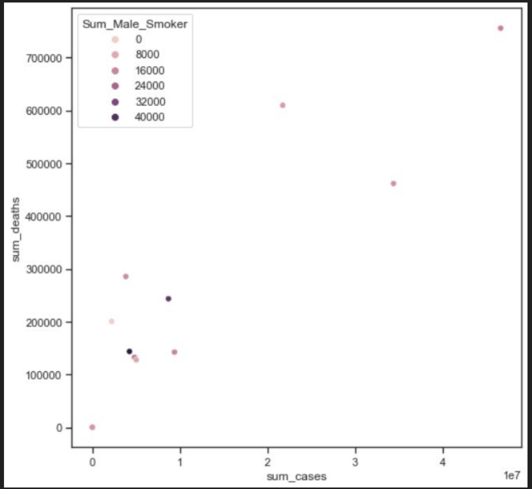


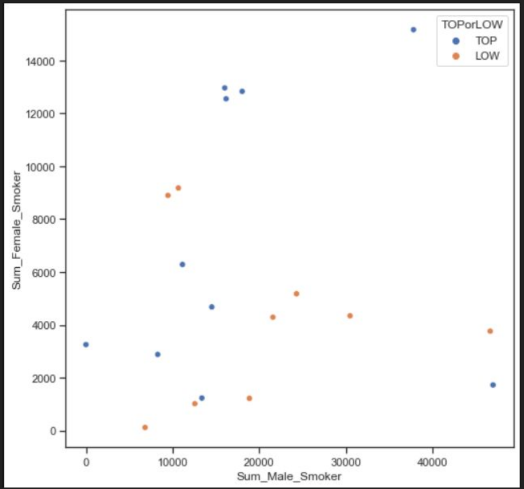


How does being a smoker affect the death numbers?

1. Some countries has cases but none are smokers.
2. The lowest country in case number with the most deaths and smokers is Timor.
3. The lowest country in case number with the least deaths and smokers is Vanuatu.
4. The rich, more populated, countries tends to have more cases with female smokers
5. The poor, less populated, countries tends to have more cases with male smokers
6. The numbers of deaths are higher when the country has smokers amongst the cases.

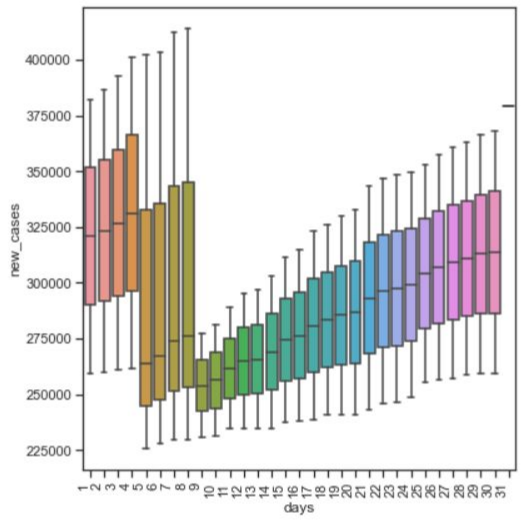


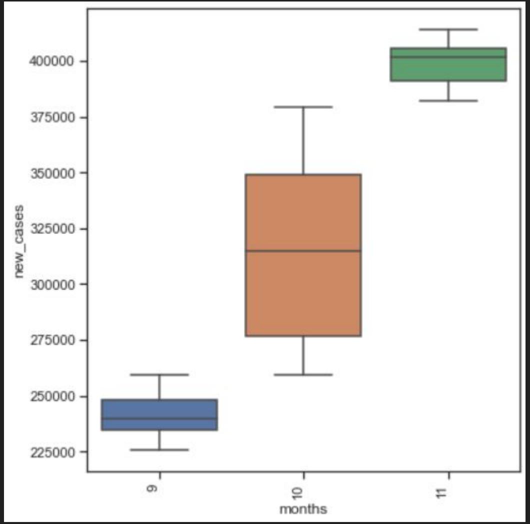




The new cases by months/days for the top 1/3 countries = 65

1. The cases increase by the day. The mean, min, max, Q1, Q2 and Q3 show that.
2. Even though the cases were increase daily the increases per-month were different as some months had higher number of cases.
3. Some months has spikes compared to other months.





The End