Data Visualization: Assignment 3 Report

Goutham U R, Adithya Nagaraja Somasalle and Pannaga Bhat

CONTENTS

I. STRINGENCY INDEX AND WAVES

A. Preliminary Visualization

The dataset [1] provides a field showing new cases per country per day as well as a 'Stringency Index' which is a "composite measure based on 9 response indicators including school closures, workplace closures, and travel bans", which ranges from 1-100. We could use these 2 columns to show the progression of the pandemic and the dynamics of government responses in different continents.

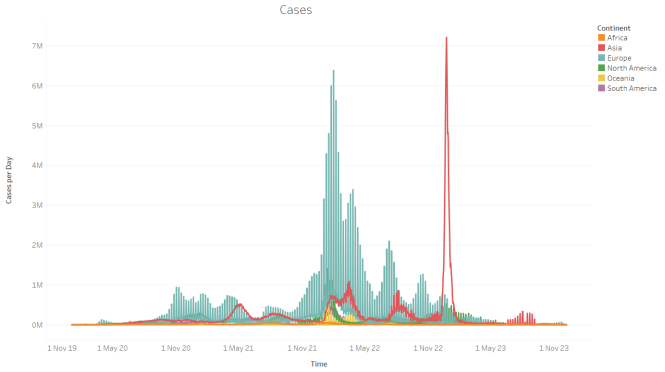

Figure 1: Line graph of new cases daily per day in different continents over time

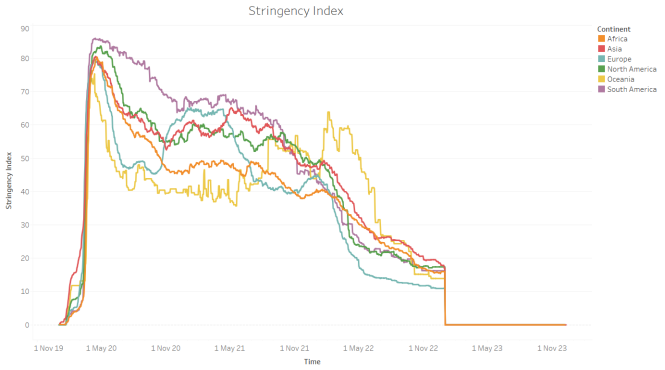

Figure 2: Stringency Index in different continents

Figs. 1 and 2 show the values of these columns over the course of the pandemic. Since data is given per country, a sum of all cases for every country in a continent is taken. We see that the lines are very jagged and no data is provided for Stringency Index after December 2022. Because some continents have a large number of cases, the data for the others is not seen clearly. Furthermore, since Stringency Index is given per country, an average is taken as the aggregate measure to calculate the value for a continent. This may be inaccurate as countries with smaller populations may have a disproportionately large effect and bigger countries may have a disproportionately small effect.

B. Data Transformation

Filtering: The data after December 2022 is hidden so that the NULL values of Stringency Index can be removed.

Weighted Stringency: A new field is calculated by multiplying the Stringency Index for a particular country by its own population and then dividing by the population of the continent the country is part of. The sum of this value for all countries in a continent gives a weighted average that also lies in between 1-100.

Note that since [1] uses a constant population for every country and continent for every date, this may be slightly inaccurate.

Cases per Capita: A new field is calculated by dividing the new daily cases for every country by the population of the corresponding continent. By adding up the value for all the different countries, we will calculate the new daily cases per capita for the entire continent. This will provide a more accurate measure of the severity of the pandemic across different continents.

Smoothing: By implementing a 7-day rolling average for both the measures, we can reduce the jagged nature of the plot and make it more presentable.

C. Refined Visualization

Figs. 3 and 4 show the line graphs with the above transformations applied. We see that it now much easier to draw conclusions from this data.

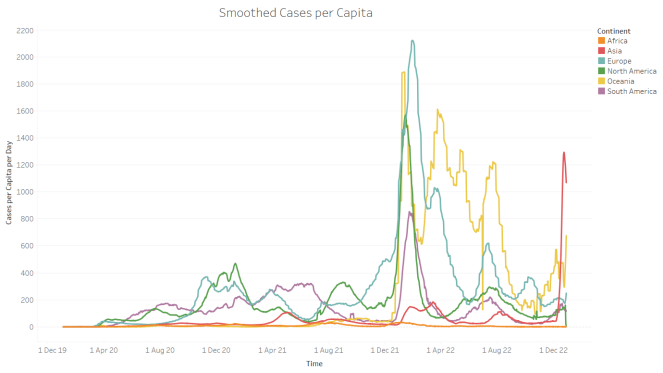

Figure 3: New cases per capita per day in different continents

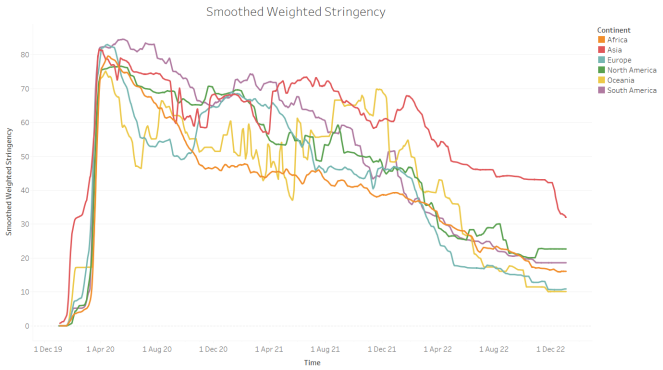

Figure 4: Weighted Stringency for different continents

D. Conclusions

1. At the onset of the pandemic, governments worldwide implemented significantly stringent restrictions. As the pandemic progressed and a better understanding emerged, these restrictions tended to ease.
2. A large uptick in cases denotes a wave of the pandemic, and this is accompanied by a corresponding tightening of restrictions by governments.
3. Applying smoothing to time-series data will limit the effect of random fluctuations and give a better understanding of the information. A small interval may be heavily influenced by random noise, but conversely a large window size may erase short term trends in the data.

II. Deaths and Vaccinations

For exploration of this task I have used both Python and Tableau.

A. Pre-processing

Removed duplicates and converted date column to pandas recognizable date form.

B. Data Analysis and Statistical Analysis

First among the given 30 features, 16 features were chosen for plotting correlation matrix. These 16 features were chosen as these columns are more relevant with respect to upcoming exploration and other columns had large amount of NULL values.

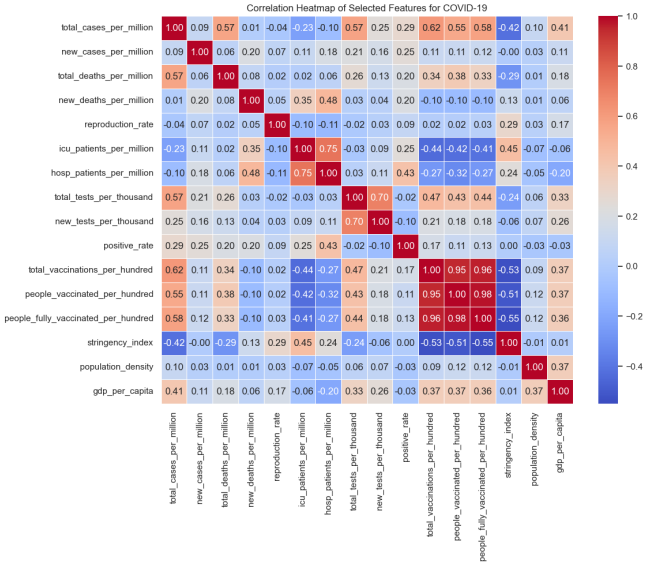

Figure 5: Correlation Matrix

High Positive Correlations

- **total_cases_per_million** and **total_deaths_per_million** Correlation: 0.57
- **total_cases_per_million** and **total_tests_per_thousand** Correlation: 0.57
- **total_cases_per_million** and **total_vaccinations_per_hundred** Correlation: 0.62
- **hosp_patients** and **icu_patients** Correlation: 0.75

- **total_vaccinations_per_hundred** and **people_fully_vaccinated_per_hundred**
  Correlation: 0.96

HIGH NEGATIVE CORRELATIONS

- **icu_patients_per_million** and **stringency_index**
  Correlation: -0.450929
- **icu_patients_per_million** and **people_vaccinated_per_hundred**
  Correlation: -0.435668
- **icu_patients_per_million** and **people_fully_vaccinated_per_hundred**
  Correlation: -0.413003

From above we can see that hospital_patients and icu_patients have good amount of correlation, to analyse it further we can plot a bar graph side by side for different continents.

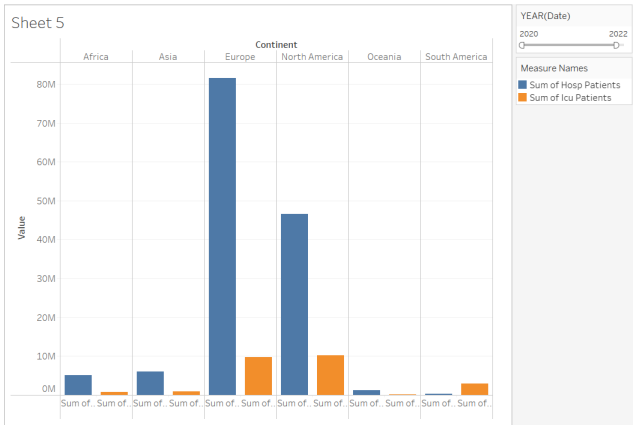

Figure 6: Bar chart of Hospital Patients and ICU patients for different continents

From Fig. 6 we see that for Europe and North America the number of hospital patients are way higher than ICU patients, so we can infer that in these continents the disease is not severe.

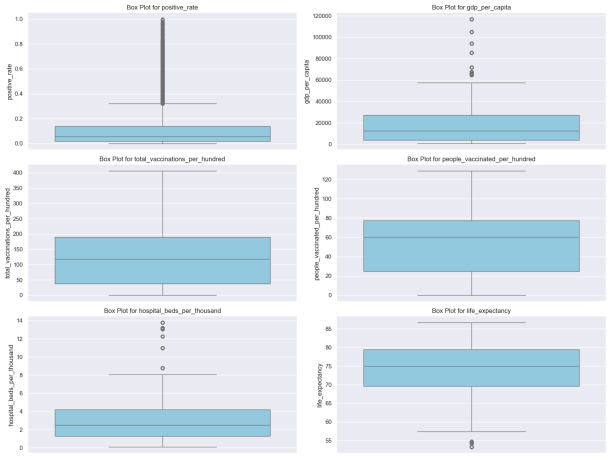

Figure 7: Box-Plots of various columns

BoxPlots: These Box plots gives us idea of the distribution of each of the column is varying.

C. *Forecast of Total Deaths with Date*

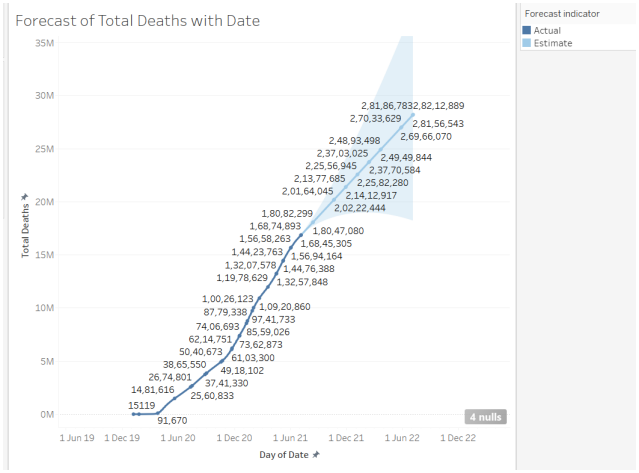

Figure 8: Forecast of Total Deaths

For forecasting of Total Deaths, Tableau feature forecast of data analytics has been used. This gives us an idea of how the total deaths may vary in future and help us predict an estimate of it.

Following diagram is used for both the workflows.

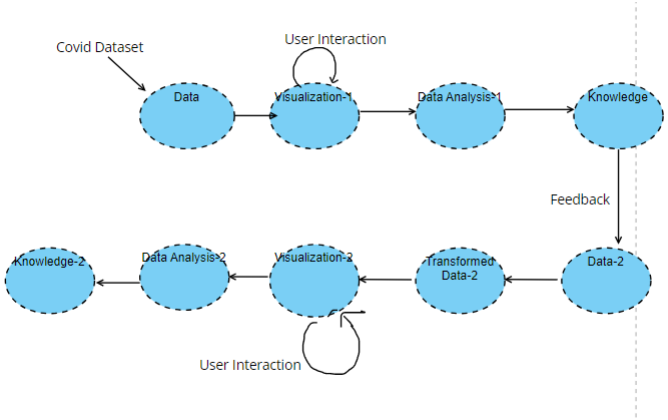

Figure 9: WorkFlow Diagram

D. *Workflow 1*

Workflow 1 explores the aspects of mortality during a specific period of time.

Data: In this we take the input as covid dataset and then we choose columns required for visual exploration. In this case they are date,location, total_deaths, total_cases and population.

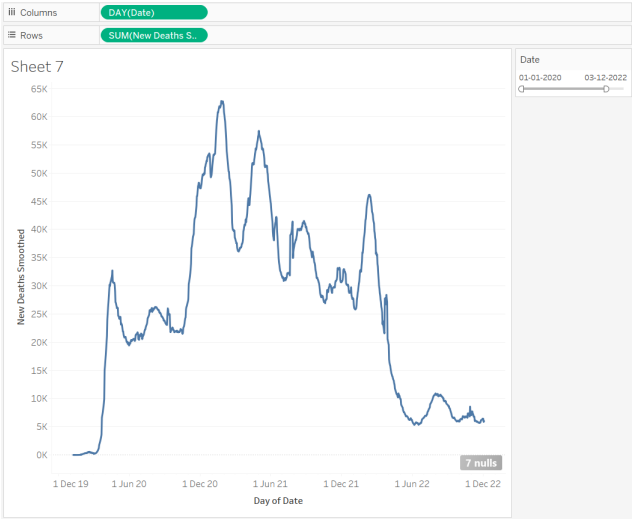

Figure 10: New Deaths smoothed vs Date

Visualization-1*:* Fig. 10 shows us plot of New Deaths smoothes vs Date. We can also include visualization of forecast of total deaths Fig. 8 here.

Data Analysis-1*:* Fig. 10 was explored in A1. We see peaks in New Deaths Smoothed at August 2020-January 2021,March 2021-June 2021 and Jan 2022-March 2022, these peaks correspond to Covid wave-1, wave-2 and wave-3 respectively.
For Fig. 8 we get the Root Mean Square Error to be 11,744 and Mean Absolute Error to be 7,275.

Knowledge-1*:* New Deaths Smoothed vs Date may be a great way for studying temporal trends but it doesn't take into consideration, the population change and specific disease for reason of death.

Data - 2*:* Clearly we need additional metric to evaluate mortality considering factors gained by 'Knowledge' state. So we introduce two metrics called Crude Mortality Rate(CMR) and Case Fatality Rate(CFR) [2]. Crude Mortality Rate answers the question "if someone is infected, how likely they are to die?". Case Fatality Rate tells us how serious the case is.

Tranformed Data -2 *:* We add 2 more columns Crude Mortality Rate and Case Fatality Rate to our data, which is calculated as

$$\text{CFR} = \left(\frac{\text{Number of Deaths}}{\text{Number of Cases}}\right) \times 100$$

$$\text{CMR} = \left(\frac{\text{Number of Deaths}}{\text{Total Population}}\right) \times 100000$$

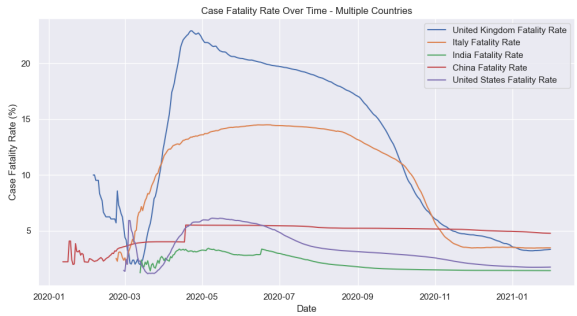

Figure 11: Case Fatality Rate vs Date For Various Countries

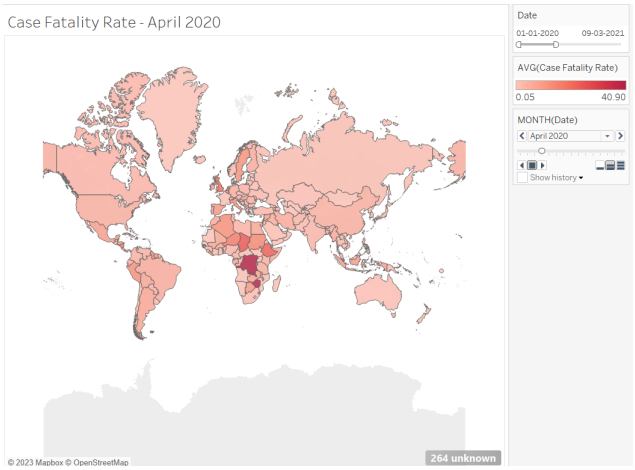

Figure 12: Case Fatality Rate over World Map

Visualization-2*:* Fig. 11 shows Cases Fatality Rate for United Kingdom, India, China, Italy and United States from Jan 2020 to Jan 2021. We can see that the peak is attained by United Kingdom first and then by Italy.

Fig. 12 shows the Case Fatality Rate for the world using a chloropeth map for April 2020. The areas which are darker in colour represents higher CFR values. We can see that Zimbabwe and Republic of Congo are dark red. Zimbabwe has CFR of 42.39 and Republic of Congo has CFR of 59.14.

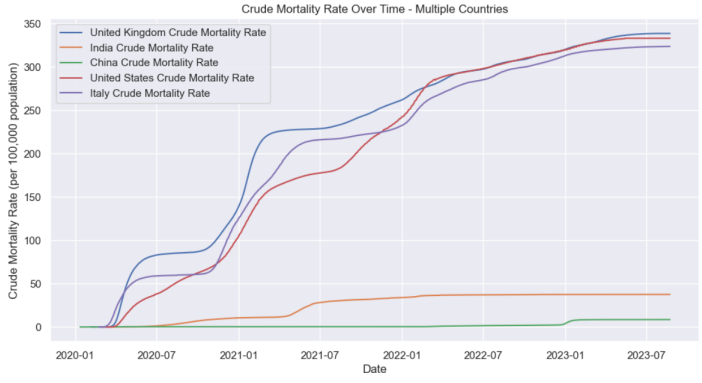

Figure 13: Crude Mortality Rate vs Date For Various Countries

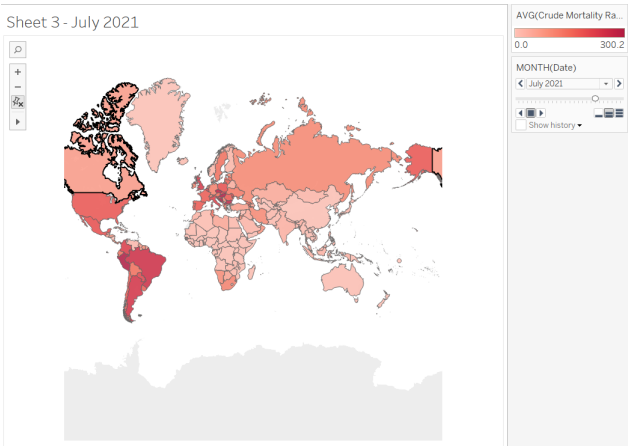

Figure 14: Crude Mortality Rate vs Date For Various Countries

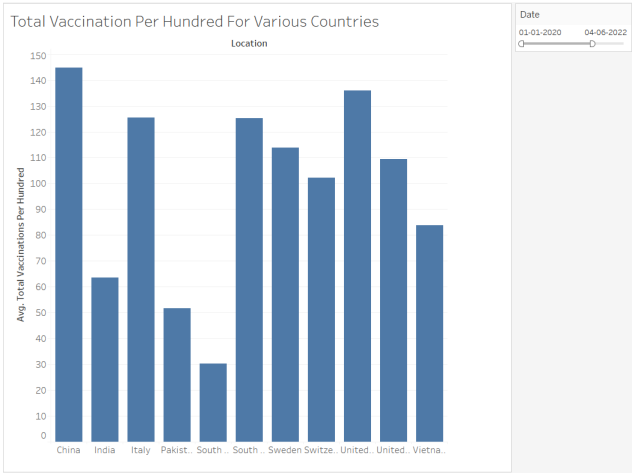

Figure 15: Total Vaccination per hundred For Various Countries

From Fig. 13 we can see that the CMR of the countries United Kingdom, United States and Italy rises faster than of China and India.
From Fig. 12 we can see that Brazil and Argentina have high Crude Mortality Rate as of July 2021.

Data Analysis-2*:* The mean of Case Fatality Rate is 2.024876 and that of Crude Mortality Rate is 0.8523312. They have maximum values of 94.945248 and 6.501224.

Knowledge*:* In Fig. 11, we can infer from the peaks that the large portion of people diagonised with covid have died in United Kingdom and Italy, this may be due to bad healtcare system in these areas. They may have limited resources, insufficent medical infrastructure to handle corona virus. It can also be seen that India, China and United States were successful in keeping the deaths stable by having CFR near 5%.
In Fig. 12, we can see that Republic of Congo has CFR of 59.14, we can infer that the severity of disease was high in this region and medical infrastructure was poor.

From Fig. 13 we can infer that India and China have relatively stable CMR, which means that they have done relatively better in handling number of deaths for given population. In Fig. 14, high CMR values of Brazil and Argentina can be due to problem of Vaccination coverage for the large population, not following public health measures like social distancing and unable to treat patients who were tested positive due to lack of ventilators and hospital beds for large population.

E. Workflow-2

This workflow explores aspects of Vaccination in different countries and economic groups.

Data*:* : In this we take the input as covid dataset and then we choose columns required for visual exploration. In this case they are location , date, total_vaccinations_per_hundred,population , gdp_per_capita. In this we choose columns of interest in our dataframe and filter date until 2022, as after that year vaccination for covid isn't much relevant.

Fig. 15 gives us the total vaccination per hundred for countries China,India,Italy, Pakistan, South Africa, South Korea, Sweden, Switzerland, United States, United Kingdom and Vietnam.

Data Analysis-1*:* United States has value of 109.4, China has value of 144.8, Italy has value of 125.5, South Africa being least of them has 30.2.

Knowledge-1*:* We can see China has highest value followed by United Kingdokm. We can infer that these high values for China and United Kingdom must be because of well organized vaccination camps, vaccine availability and health care infrastructures. However, here we are missing the fact that the purchasing power of these countries are not same.

Data-2*:* The previous data doesn't tell us given their purchasing power how well the country has done in terms of vaccinating people. That's why we need to transform our data such that it includes income of the country.

Transformed Data-2*:* Since in our input dataset we don't have any of the attributes or features related to income of the country, we use gdp_per_capita as the criteria to divide the countries as high-income, upper-middle income, lower-middle income and low income countries.
We set the following thresholds on gdp_per_capita gdp_per_capita

$$threshold\_high = 40000$$

$$threshold\_upper\_middle = 15000$$

$$threshold\_low = 5000$$

Then we applying Weighted Average of the Total Vaccination per hundred in each income group.

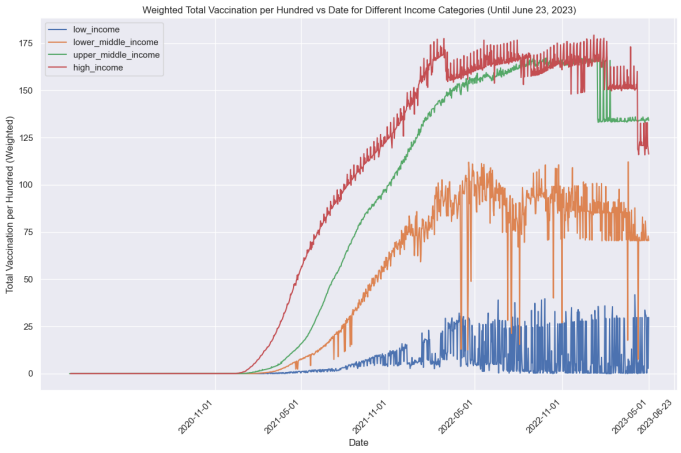

Figure 16: Weighted total Vaccination per hundred for 4 levels of income groups

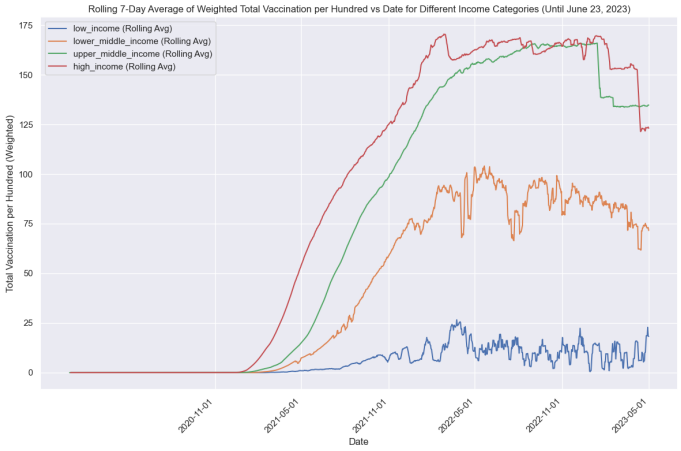

Figure 17: Smoothed Weighted total Vaccination per hundred for 4 levels of income groups

Visualization: Fig. 16 shows the Weighted total Vaccination per hundred for 4 levels of income groups.

Fig. 16 shows the Smoothed Weighted total Vaccination per hundred for 4 levels of income groups. It uses 7-day rolling average.

Data Analysis-2: The mean is of gdp_per_capita is 18993.540929. The mean value of Weighted Vaccination Per Hundred of low income country is 5.638232, lower-middle income countries is 42.569209, upper middle income country is 76.003587 and for high income countries it is 87.81583. Low income countries deviate the least with 9.163746 and high income countries deviate max with 71.695801.

Knowledge-2: From Fig. 17 we can see that the value of total vaccination per hundred increases faster in high income groups and upper-middle income countries faster than lower-middle income and low income countries. This is because the have higher purchasing power and these income groups can afford vaccination for higher prices compared to other income groups. We can infer that these high income countries like United States, Germany and upper-middle income countries like China, Brazil have greater financial resources and access to diverse vaccine portfolio compared to countries belonging to other income groups.

III. Analyzing the number of covid cases and its density

These visualizations were generated using python and machine learning methods like principal component analysis (PCA) was used.

A. Visualization 1

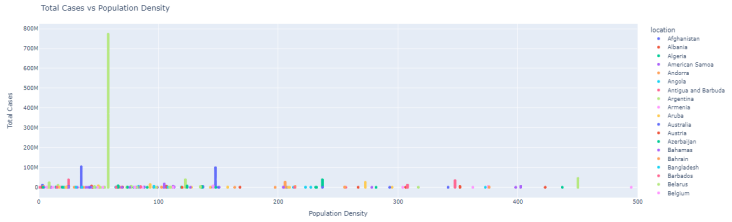

Figure 18: Total number of cases plotted against population density for every country and the world

1) Inference:

1. As expected, the total number of cases recorded is the highest for the world because it includes the number of cases recorded in every country (the longest green coloured line whose x co-ordinate is between 0 and 100)
2. Next in line, we have China, USA and India and this observation from the dataset is in par with what actually happened during covid times.

1) Feedback Loop: Because of the presence of a very large outlier, the data (total number of cases in this case) for other countries cannot be interpreted/understood easily from the visualization. The solution will be to remove the outlier and get a more clear visualization with only the countries.

B. Visualization 2

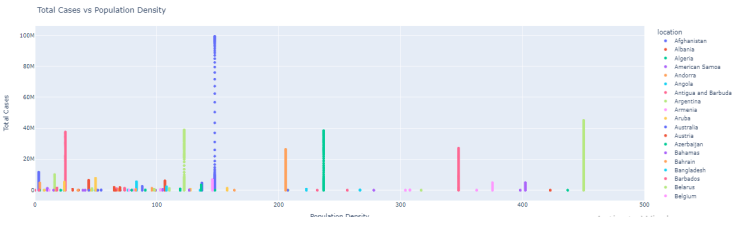

Figure 19: Total number of cases plotted against population density for every country

2) Inference: The total number of cases for every country can now be interpreted clearly using Figure. 19

2) Feedback Loop: Since we are considering total number of cases, it is hard to predict the intensity of covid in a particular

country, as it obvious that countries with larger populations will be having more number of cases (like China, USA, India) than the countries with smaller populations. The solution will be to use total cases per million instead of total cases which is being done in the next visualization.

C. Visualization 3

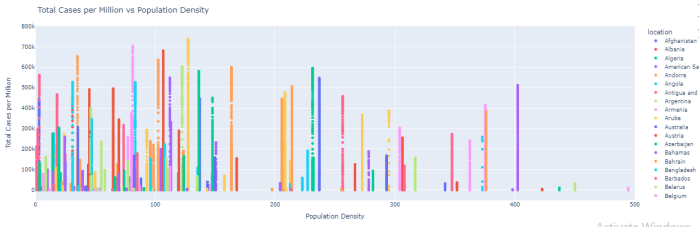

Figure 20: Total cases per million plotted against population density for every country

3) Inference:

1. Since we are using total cases per million instead of total number of cases, even countries will smaller populations like Cyrus, Austria and Brunei have high values compared to other countries.

2. On the other hand, countries which had very high number of cases have very low case density values.

3. It's reasonable to infer that the longer the bar representing a country, the more significant the impact of COVID-19 on that specific nation.

3) Feedback Loop: Population density cannot be the only factor which determined the case density (as we don't see a regular trend). There might be several other factors which might have affected the spread of covid. The solution will be to introduce a third variable and use a bubble chart. The third variable will be features like life expectancy, hospital beds per thousand, life expectancy and human development index.

D. Visualization 4

The next four visualizations are bubble plots where the x-axis represents the population density and the y axis represents one of the features among life expectancy, hospital beds per thousand, life expectancy and human development index. The size of the bubbles are determined by case density (total cases per million) and hovering over each bubble gives us the country name, population density, value of the feature being considered and total cases per million.

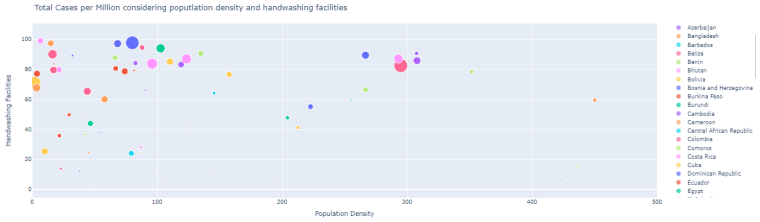

Figure 21: Total cases per million considering population density and handwashing facilities

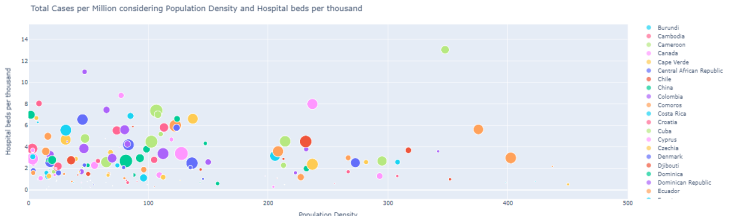

Figure 22: Total cases per million considering population density and hospital beds per thousand

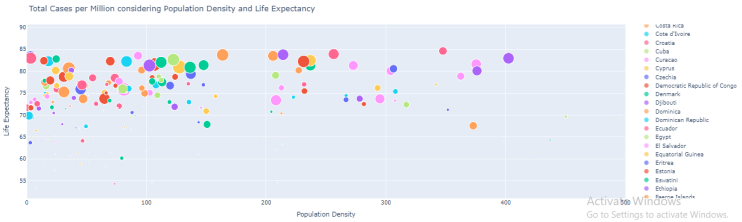

Figure 23: Total cases per million considering population density and life expectancy

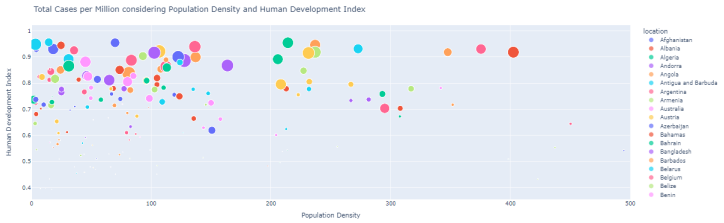

Figure 24: Total cases per million considering population density and human development index

4) Inference:

1. From fig.21, in general, countries with more advanced handwashing facilities tend to have lower case density compared to those with less developed handwashing infrastructure. Although, few countries like Serbia and Marshall islands are few exceptions as they had a higher case density in spite of having one of the best handwashing facilities.

2. From fig.22, the general trend is that countries with higher number of hospital beds per thousand have had lesser case density than those with lower number of hospital beds per thousand. Few exceptions to this rule are countries like Bolivia and Guyana, these countries had a low case density

despite having lower number of hospital beds (per thousand).

3. From fig.23, There seems to be no clear correlation between life expectancy and case density. Namibia with a low life expectancy of 63, has a lower case density than Iceland which has a life expectancy of 83. On the other hand, Sri Lanka with life expectancy of 76 has a lower case density than Georgia which has a life expectancy of 74.

4. From fig.25, when we consider countries with lower population densities, those with higher HDI have higher case density than those with lower HDI. In regions with high population densities, there is little disparity in case densities, irrespective of variations in their Human Development Index (HDI) rankings.

4) Feedback Loop: Although we have considered a lot of features, the visualizations are generated only with population densities. What if multiple factors are influencing case densities? In that we will have to interpret a lot of visualizations to arrive at a conclusion. The solution will be to use Principal Component Analysis or PCA in order to find the 2 most dominant factos which can be used to determine case densities.

E. Visualization 5

In this step, I have used PCA in order to find the 2 most dominant factors which have influenced the case densities in various countries.

```
Names of the Two Dominant Factors:
For Principal Component 1: Index(['population_density', 'hospital_beds_per_thousand'], dtype='object')
```

Figure 25: Results obtained after applying PCA algorithm

5) Inference: The features population density and hospital beds per thousand are the best 2 features which can be used to decide/determine case density.

F. Conclusion

1. Only considering total number of cases will not help us interpret the impact of covid on a particular country. Total cases per million or case density is a better measure.

2. Population density cannot be the sole factor which can determine case density. We need to consider features like life expectancy, handwashing facilities, HDI and hospital beds per thousand.

3. Interpreting a visualization with 5 features will be very difficult. Hence, it is important for us to use PCA.

4. Using PCA, we have found out that the features population density and hospital beds per thousand affect case density the most.

IV. Author Contributions

- Stringency Index and Waves: Pannaga Bhat, IMT2021080
- Deaths and Vaccinations : Adithya Nagaraja Somasalle, IMT2021054
- Analyzing the number of covid cases and its density: Goutham U R, IMT2021045

References

[1] Our World in Data - github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

[2] https://ourworldindata.org/mortality-risk-covid

[3] Link to A1 report
https://github.com/Bambo0st/DV_V1/blob/main/Report.pdf