EDA on COVID Dataset

Sarah Hosseini Feshtami - 400222026

Shahid Beheshti University

1. Data Collection

The COVID data (github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv) is maintained by Our World in Data (OWID) and aims to provide reliable and comprehensive COVID-19-related information from around the world. The dataset includes a wide range of features related to COVID-19, such as daily cases, deaths, hospitalizations, testing, vaccination progress, and various demographic and socio-economic indicators.

The CSV file follows a format of 1 row per location and date.

The dataset contains the following attributes:

```
['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',
    'new_cases_smoothed', 'total_deaths', 'new_deaths',
    'new_deaths_smoothed', 'total_cases_per_million',
    'new_cases_per_million', 'new_cases_smoothed_per_million',
    'total_deaths_per_million', 'new_deaths_per_million',
    'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',
    'icu_patients_per_million', 'hosp_patients',
    'hosp_patients_per_million', 'weekly_icu_admissions',
    'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',
    'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests',
    'total_tests_per_thousand', 'new_tests_per_thousand',
    'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
    'positive_rate', 'tests_per_case', 'tests_units', 'total_vaccinations',
    'people_vaccinated', 'people_fully_vaccinated', 'total_boosters',
```

'new_vaccinations', 'new_vaccinations_smoothed', 'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred', 'total_boosters_per_hundred', 'new_vaccinations_smoothed_per_million', 'new_people_vaccinated_smoothed', 'new_people_vaccinated_smoothed_per_hundred', 'stringency_index', 'population_density', 'median_age', 'aged_65_older', 'aged_70_older', 'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers', 'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand', 'life_expectancy', 'human_development_index', 'population', 'excess_mortality_cumulative_absolute', 'excess_mortality_cumulative', 'excess_mortality', 'excess_mortality_cumulative_per_million']

Out of these attributes, only ‘continent’ and ‘test_units’ are categorical (less than or equal to 20 unique values) and the rest are numerical.

2. Data Cleaning: We checked for duplicate records with the ‘`drop_duplicates`’ function. Then, we looked at the content of some of the columns. For example, the ‘iso_code’ column looks like this:

```
0         AFG
1         AFG
2         AFG
         ...
357230    ZWE
357231    ZWE
357232    ZWE
```

The ‘location’ feature has the names of countries as its values:

```
0         Afghanistan
1         Afghanistan
2         Afghanistan
             ...
357230       Zimbabwe
357231       Zimbabwe
357232       Zimbabwe
```

The ‘tests_units’ column consists of these values:

```
array([nan, 'tests performed', 'units unclear', 'samples tested',

       'people tested'], dtype=object)
```

After that, we replaced missing values in our data frame with NaN.

Additionally, we determined the unnecessary columns that need to be dropped. These columns are calculated on two criteria: first, the columns that are null with a percentage of more than 50. Second, the columns that have only one value. These columns are from the first group:

['icu_patients', 'icu_patients_per_million', 'hosp_patients', 'hosp_patients_per_million', 'weekly_icu_admissions', 'weekly_icu_admissions_per_million', 'weekly_hosp_admissions', 'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests', 'total_tests_per_thousand', 'new_tests_per_thousand', 'new_tests_smoothed', 'new_tests_smoothed_per_thousand', 'positive_rate', 'tests_per_case', 'tests_units', 'total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'total_boosters', 'new_vaccinations', 'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred', 'total_boosters_per_hundred', 'extreme_poverty', 'handwashing_facilities', 'excess_mortality_cumulative_absolute', 'excess_mortality_cumulative', 'excess_mortality', 'excess_mortality_cumulative_per_million']

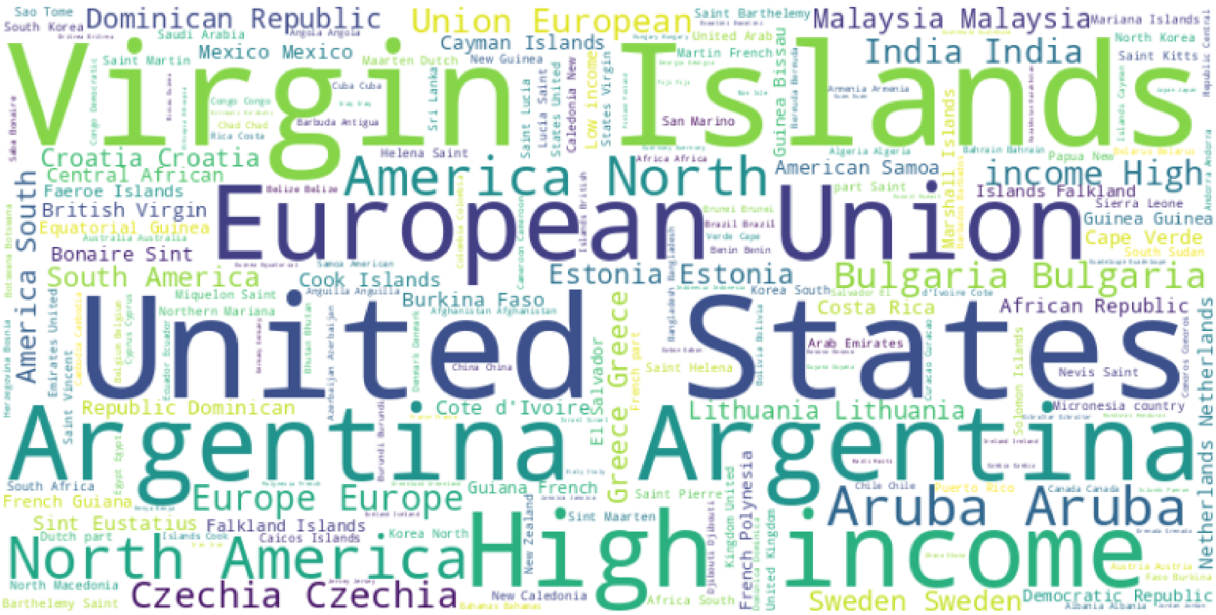
And there were no identical columns.

After these steps, the shape of our dataframe changes from (357233, 67) to (357233, 35).

Then, we use a z-score of 3 to detect and remove outliers and the shape shrinks to (55442, 35). This means that only 15 percent of our data has remained and 85 percent were detected as outliers. For example, the whole data from Afghanistan was removed. So, this method of outlier detection was not appropriate and we dismissed this method.

We then used another method that calculates the z-score of each numeric column using the 'stats' library from Scipy. The shape of the dataframe reduces to (351561, 35) which indicates a reasonable amount of outliers removed. So, we choose to work with this cleaned data.

3. Univariate Analysis: One of the interesting columns is the 'location' column. We used a wordcloud to see the frequency of its values:

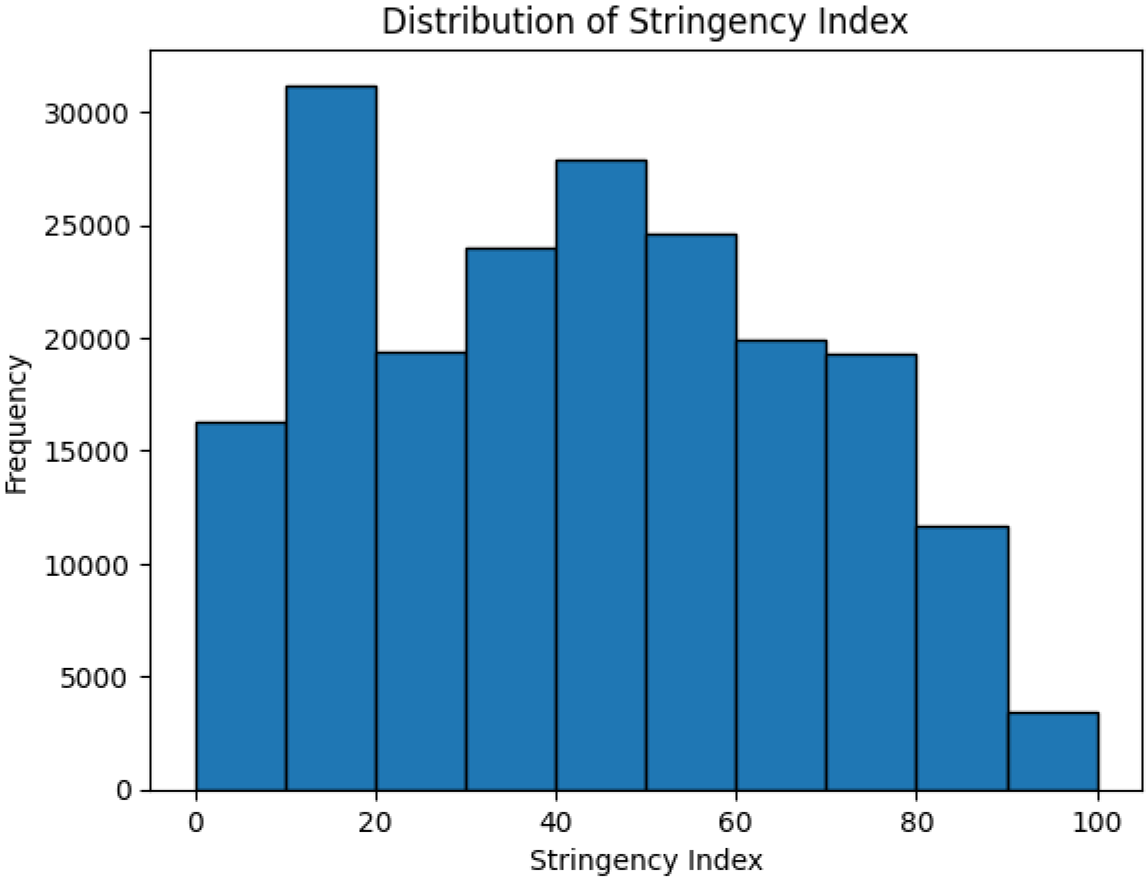


One thing that captured our attention is the appearance of the word 'high income'. We furthermore see that all of the records with a 'low income' location have an iso_code of 'OWID_LIC' and a 'high income' location an iso_code of 'OWID_HIC'.

Another notable column is ‘stringency_index’. A higher score of stringency indicates a stricter response from the government to the virus, implying that the government has implemented more extensive measures to control the spread of the virus.

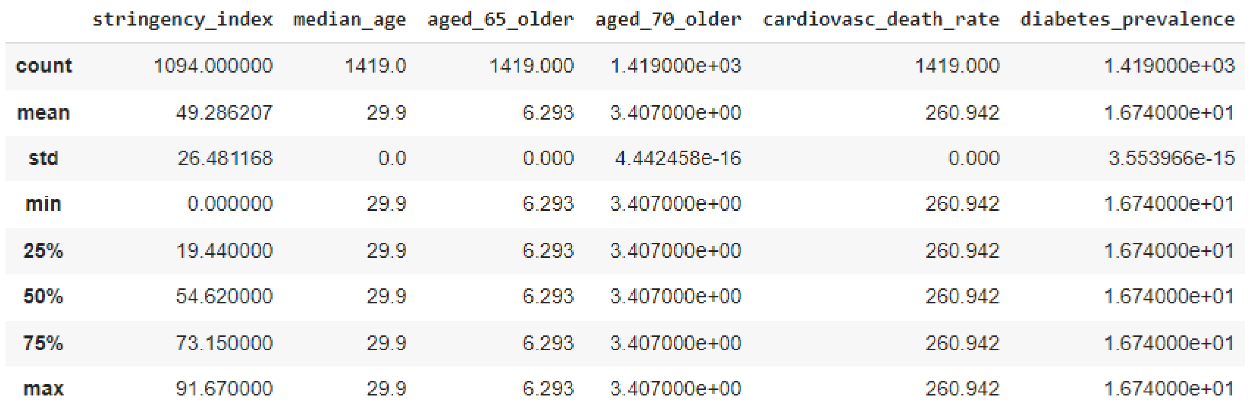
This column has:

```
count    197651.000000
mean         42.714021
std          24.911007
min           0.000000
25%          22.220000
50%          42.590000
75%          62.040000
max         100.000000
```

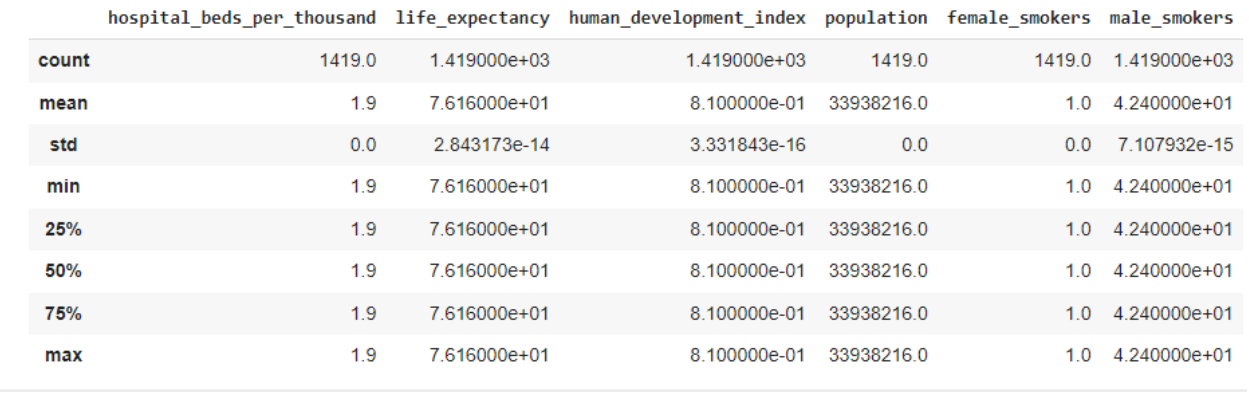


This means that more locations in more dates tended to be have a less strict response.

We also used ‘.describe()’ on the rows with Malaysia location to get an insight on the variables. After analyzing the data from the following table, we see that stringency index is not a consistent value for a specific location and it is changing. But ‘median_age’ (Median age of the population), ‘aged_65_older’ (Share of the population that is 65 years and older, most recent year available), 'aged_70_older', 'cardiovasc_death_rate' (Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)), 'diabetes_prevalence’(Diabetes prevalence (% of population aged 20 to 79) in 2017), 'hospital_beds_per_thousand', 'life_expectancy', 'human_development_index', 'population', 'female_smokers' (share of women who smoke), and 'male_smokers' (share of male smokers) are all constant valued for this location.



	stringency_index	median_age	aged_65_older	aged_70_older	cardiovasc_death_rate	diabetes_prevalence
count	1094.000000	1419.0	1419.000	1.419000e+03	1419.000	1.419000e+03
mean	49.286207	29.9	6.293	3.407000e+00	260.942	1.674000e+01
std	26.481168	0.0	0.000	4.442458e-16	0.000	3.553966e-15
min	0.000000	29.9	6.293	3.407000e+00	260.942	1.674000e+01
25%	19.440000	29.9	6.293	3.407000e+00	260.942	1.674000e+01
50%	54.620000	29.9	6.293	3.407000e+00	260.942	1.674000e+01
75%	73.150000	29.9	6.293	3.407000e+00	260.942	1.674000e+01
max	91.670000	29.9	6.293	3.407000e+00	260.942	1.674000e+01

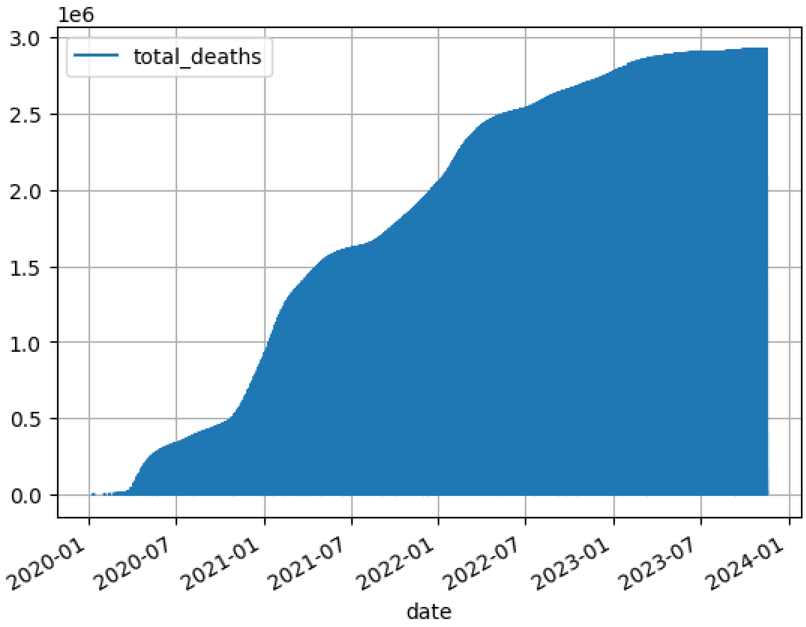


	hospital_beds_per_thousand	life_expectancy	human_development_index	population	female_smokers	male_smokers
count	1419.0	1.419000e+03	1.419000e+03	1419.0	1419.0	1.419000e+03
mean	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01
std	0.0	2.843173e-14	3.331843e-16	0.0	0.0	7.107932e-15
min	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01
25%	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01
50%	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01
75%	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01
max	1.9	7.616000e+01	8.100000e-01	33938216.0	1.0	4.240000e+01

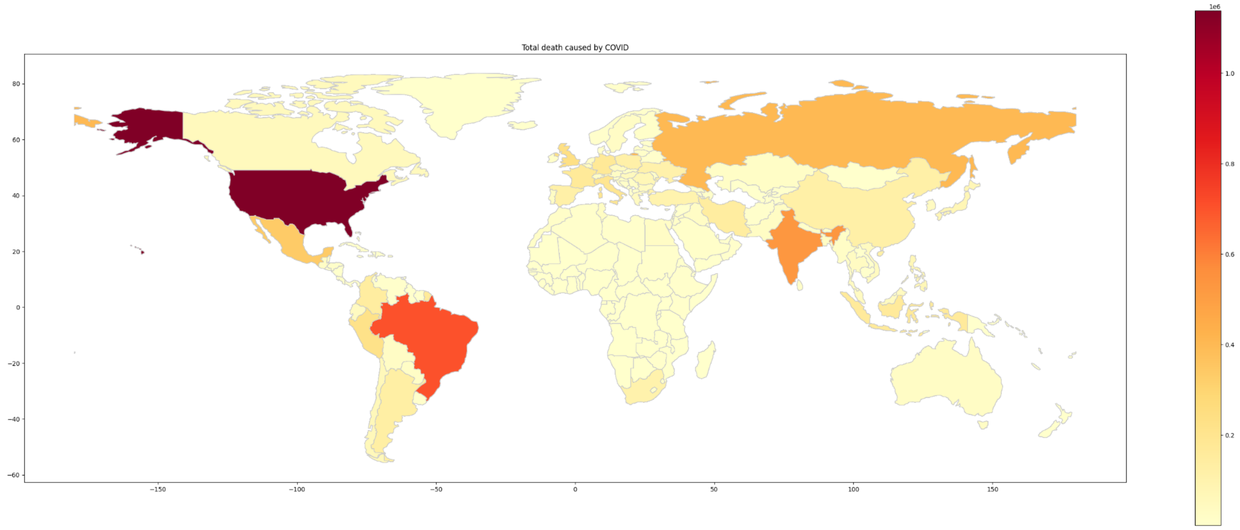
However, upon further investigating these columns in other locations, we realize that the values are not always constant and the unvariablity of these columns here is probably just lack of updated data from Malaysia.

4. Bivariate Analysis:

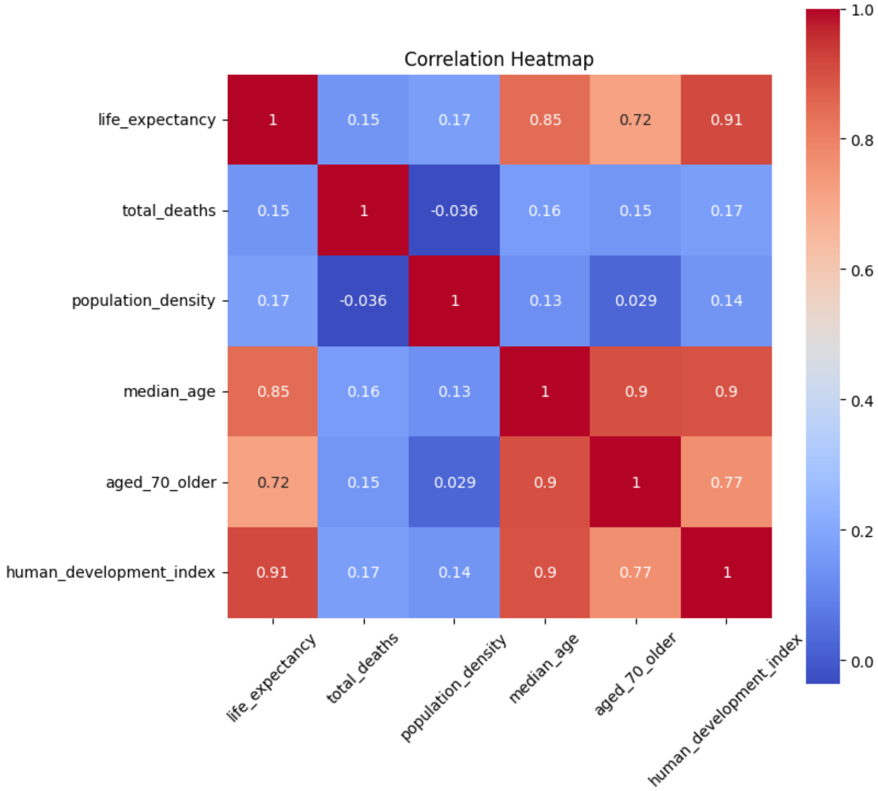
We plotted total deaths' growth based on time which shows an increasing rate:



Then, we plotted the total deaths based on locations:

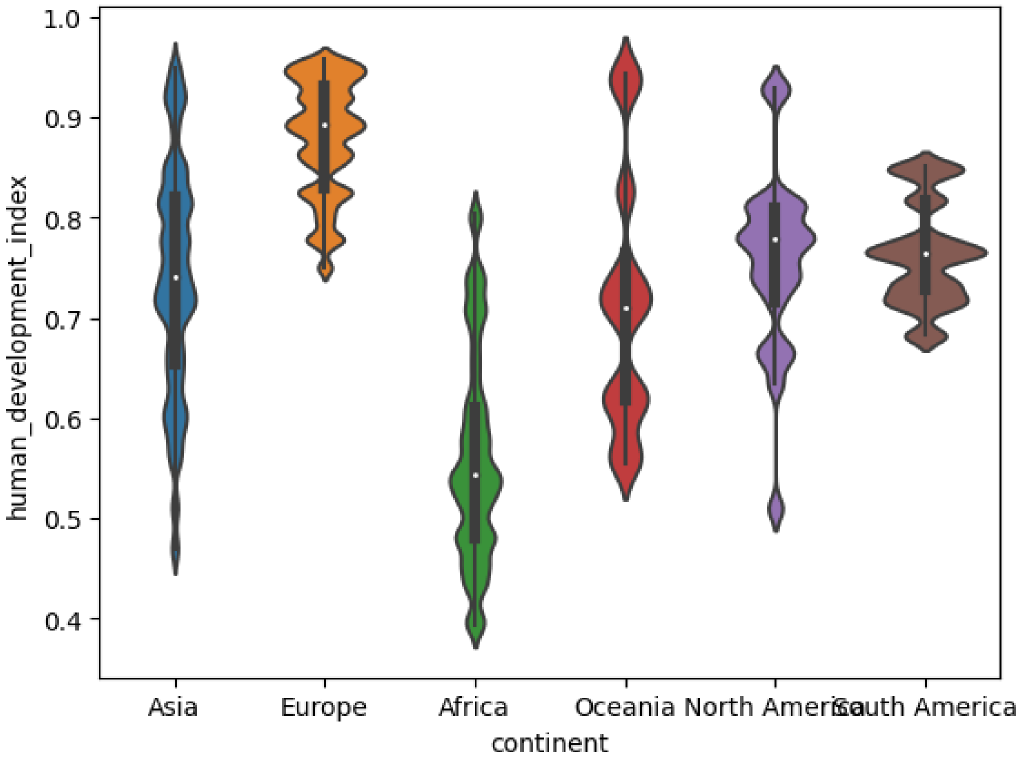


We then plotted the correlation heatmap for a few columns. This plot shows that life expectancy and median age are highly related. But total deaths and population density have a negative relation. This means that the more the density, the higher the number of total deaths.



We can also recognize that human development index has a positive relation to median age and life expectancy.

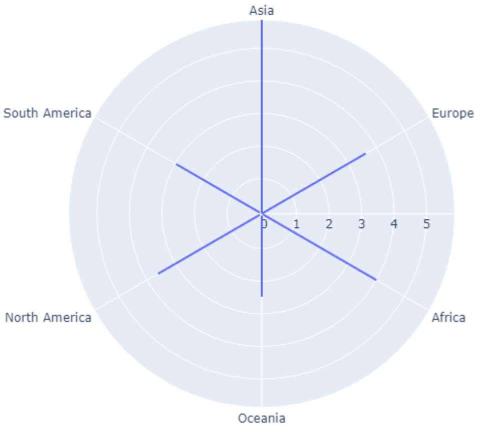
This plot shows the human development index for different continents:

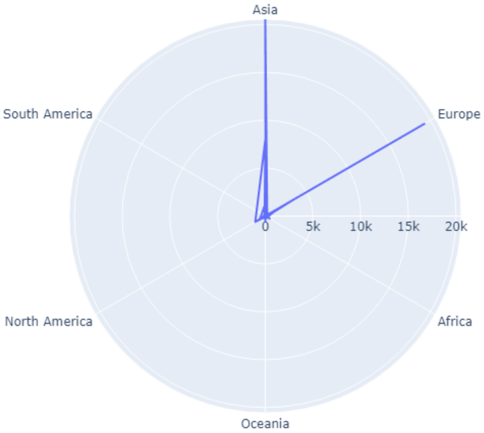


Africa has lowest and Europe the highest values.

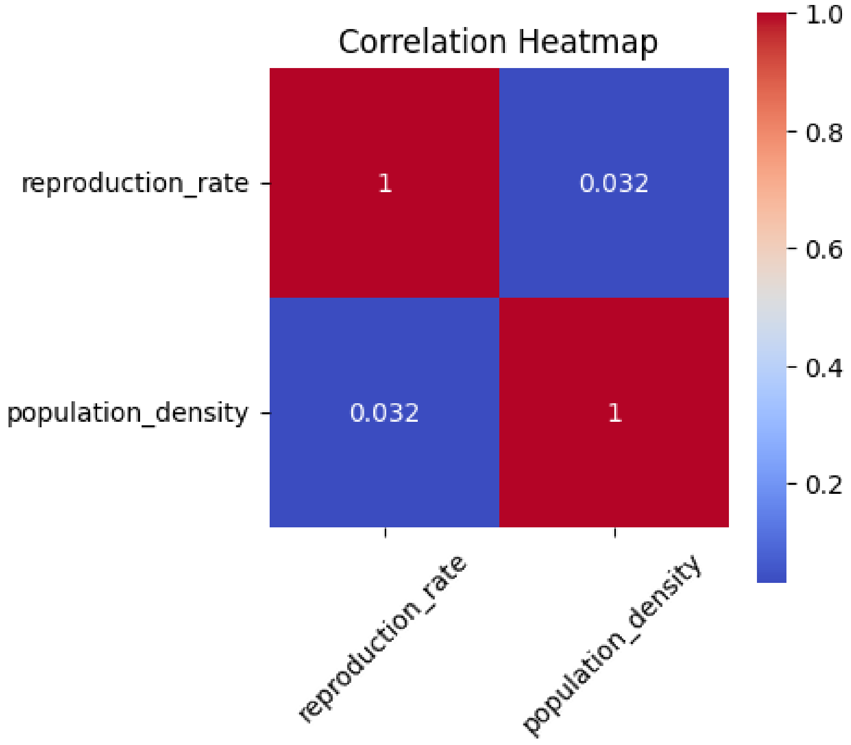
We tried to see which countries have the most reproduction rate of the virus:

```
South Korea      5.87
Iran             4.82
Turkey           4.25
Philippines      4.22
Japan            4.08
Eswatini         4.00
China            3.68
Malta            3.62
United States    3.61
Italy            3.54
```

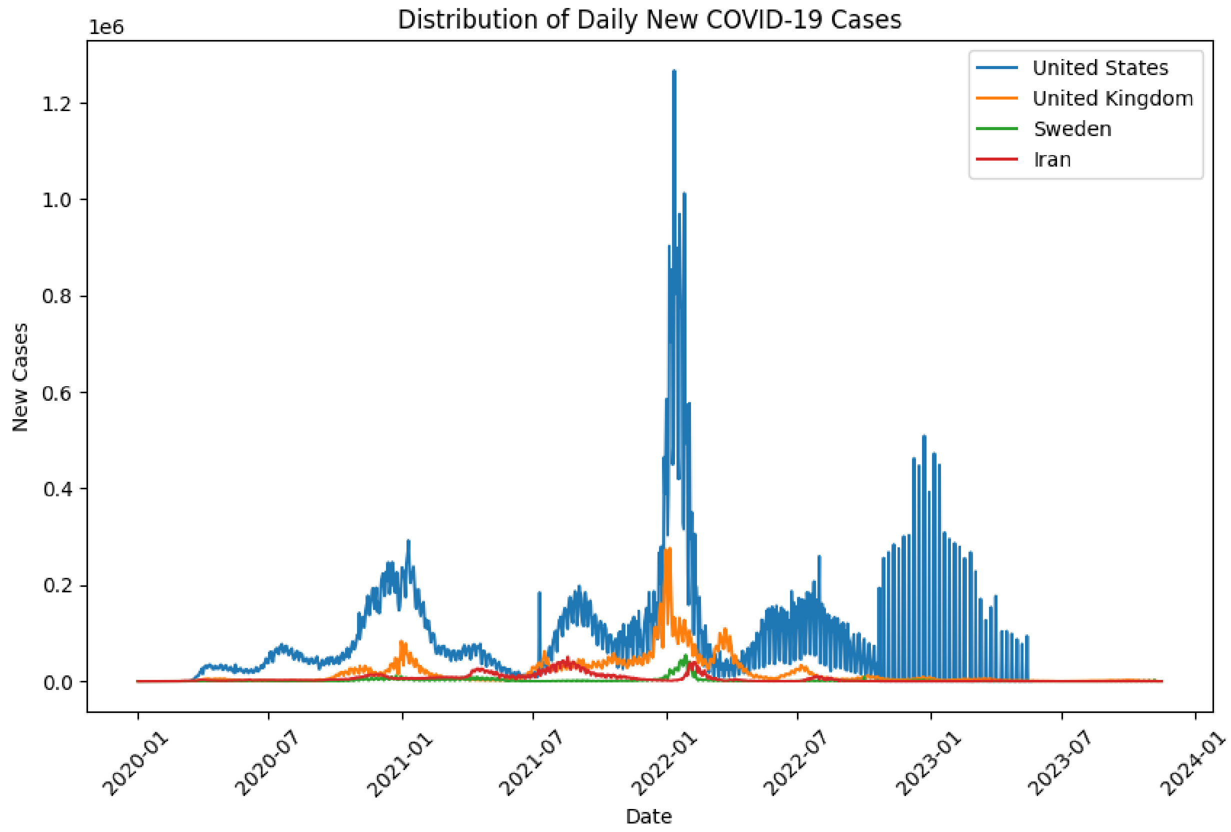




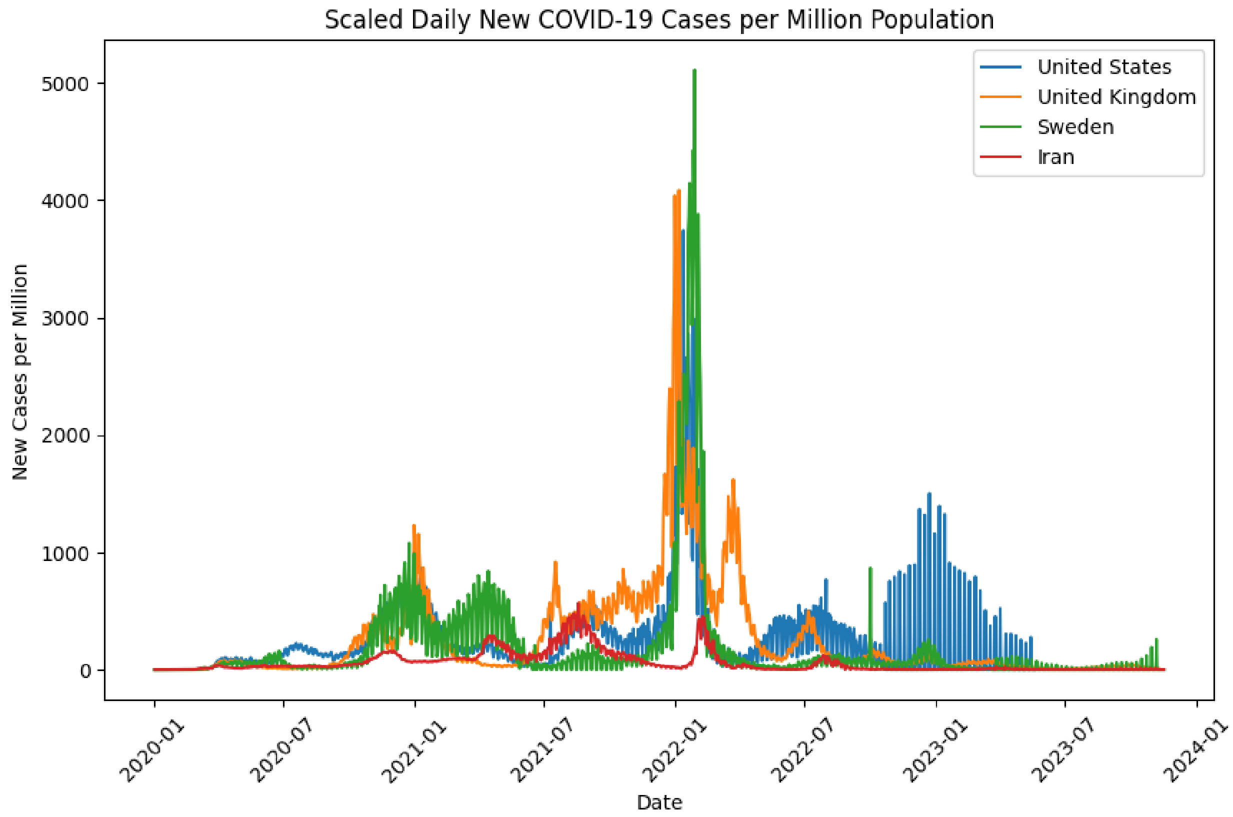
However, since the correlation metric between these two is 0.032, we can not infer that locations with a higher population density have higher reproduction rate.



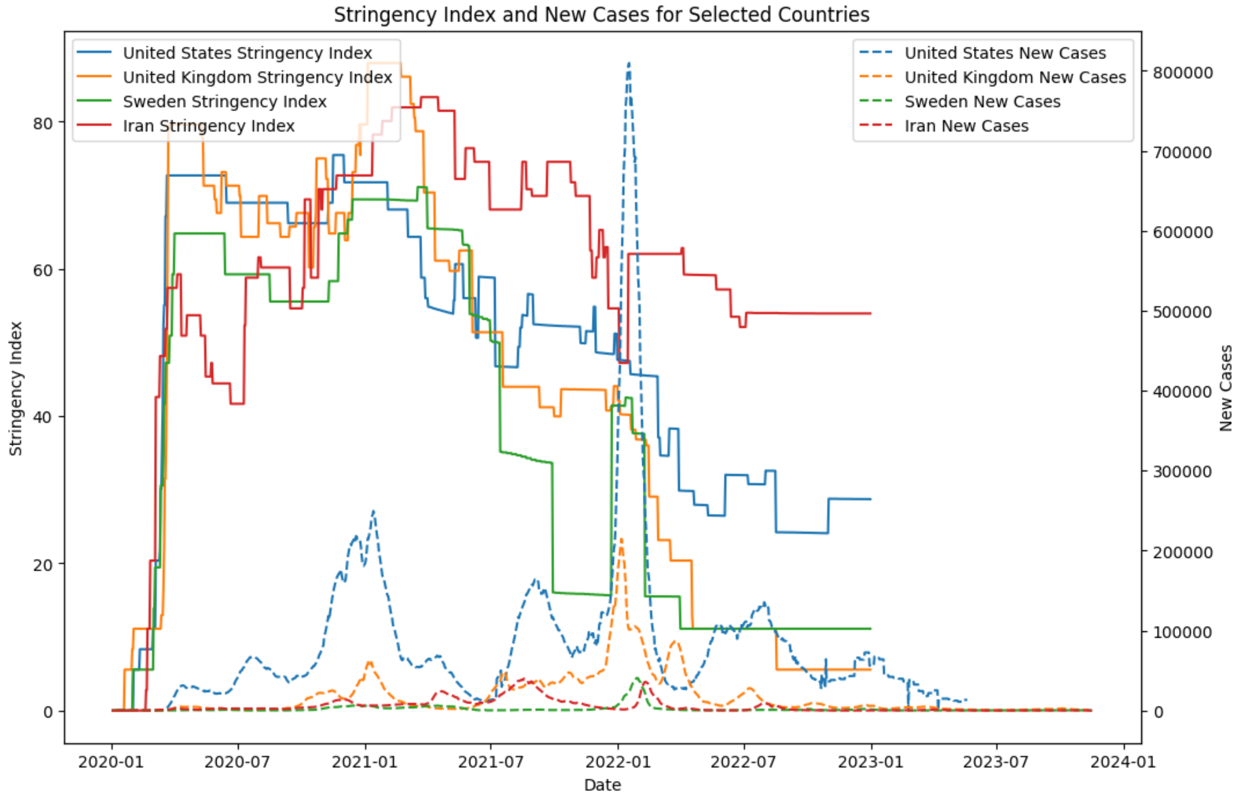
We then plotted the distribution of daily new cases. This plot shows that there was a peak in 2022-01 that happened in both US and UK.



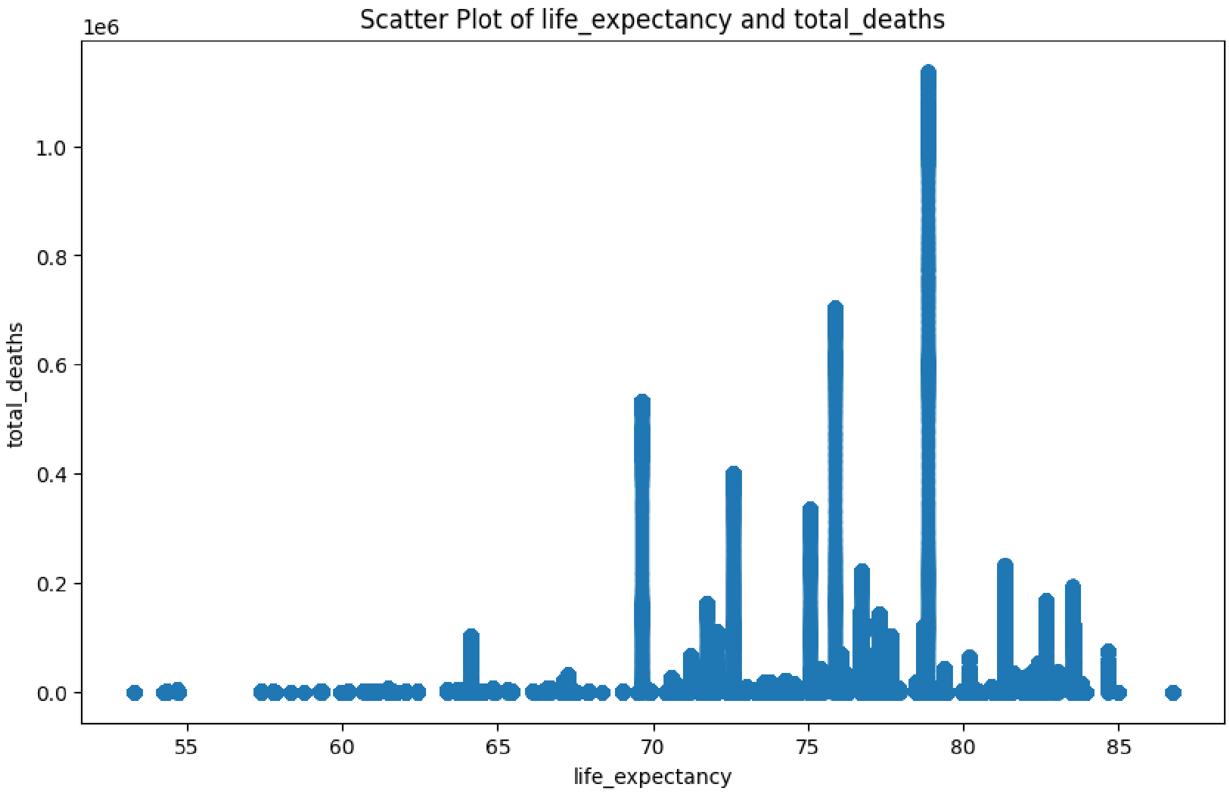
Here is a scaled version of this plot:



Then we tried to visually understand if there is any relation between stringncy index and new cases. There appears to be no significant relation but it s notable that just before the peak, the stringency was decreasing:

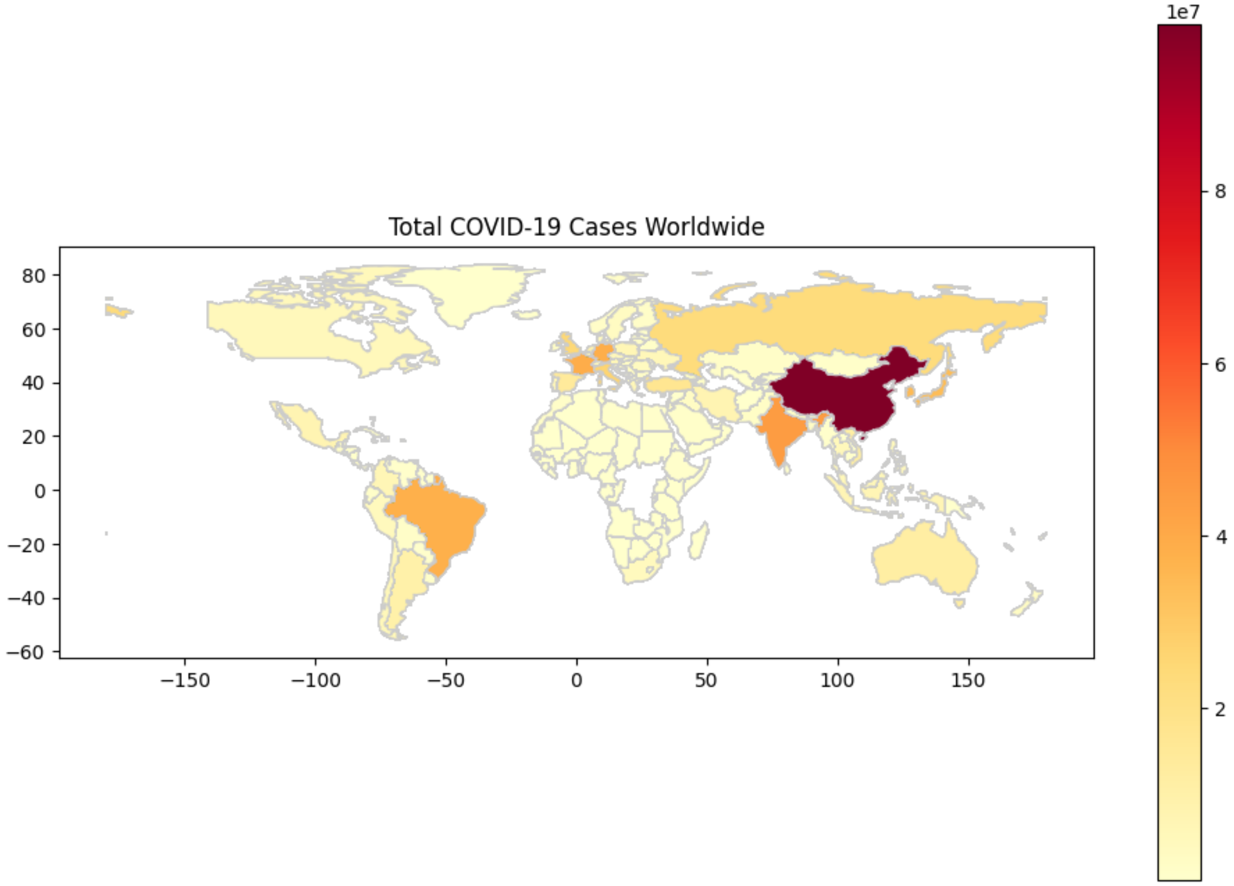

Stringency Index and New Cases for Selected Countries
United States Stringency Index
United Kingdom Stringency Index
Sweden Stringency Index
Iran Stringency Index
United States New Cases
United Kingdom New Cases
Sweden New Cases
Iran New Cases
Stringency Index
0
20
40
60
80
New Cases
0
100000
200000
300000
400000
500000
600000
700000
800000
Date
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01

This is the scatterplot of total deaths and life expectancy:



```
                total_deaths  life_expectancy
total_deaths        1.000000         0.096632
life_expectancy     0.096632         1.000000
```

The correlation is also calculated to be 0.09 which means they are probably not strongly related.



This map displays the geographical distribution of COVID-19 cases worldwide.

Countries are color-coded based on the total number of COVID-19 cases, with darker shades indicating higher case counts. Regions with a higher concentration of COVID-19 cases can be identified through the color gradient on the map.

Hotspots are represented by countries with darker colors, indicating a significant number of cases. The visualization allows for a quick comparison between countries in terms of the severity of the COVID-19 outbreak. Countries with a higher number of cases can be easily distinguished from those with fewer cases.

5. Conclusion: The exploratory data analysis (EDA) conducted in this report has provided valuable insights into the dataset under investigation. Through careful examination and visualization of the data, we have gained a deeper understanding of its characteristics and underlying patterns. The EDA process has allowed us to identify potential trends, outliers, and relationships within the data.