Covid-19 Dataset Analysis

Abstract

Covid-19 had and has been affecting our society in many aspects that would bring the importance of analyzing and understanding its components. In this study, we aimed to study covid-19 dataset with the help of exploratory data analysis. Different methods were implemented such as data cleaning, missing value handling, outlier handling, and visualization.

Introduction

By scrutinizing diverse variables such as total cases per million, total deaths per million, vaccination rates, and socio-economic indicators across different global locations and timeframes, our analysis seeks to provide a comprehensive understanding of the pandemic's progression and the factors influencing its trajectory. This process is divided into several parts:

1- data loading
2-missing value handling
3- outlier handling
4- visualization of continents
5- effect of smoking in genders
6- explore life_expectancy
7- Compare China as the source of COVID-19 vs the rest of the world
8- comprehensive analysis countrywide

Methods

1-data loading

In order to understand the basic components of our dataset, we first need to get to know its shape, features, see some samples and etc.
By getting its shape we can see the shape of the data is (355449, 67) meaning we have 355449 samples and 67 features.

Now let’s take a look at our features.

```
Index(['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'],
      dtype='object')
```

Looks like the data has full cover over the necessary features relative to covid-19 analysis.
Only raised question is about the _smoothed label at the end of some features. Which is answered by referring to the documentation of the data.
A feature being 7-days smoothed is a method to reduce the effect of outliers by averaging (in this case 7 most recent days) instead of the raw average of the day the feature is recorded.

Now let's take a look at feature values for the first 5 samples:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed
0	AFG	Asia	Afghanistan	2020-01-03	NaN	0.0	NaN	NaN	0.0	NaN
1	AFG	Asia	Afghanistan	2020-01-04	NaN	0.0	NaN	NaN	0.0	NaN
2	AFG	Asia	Afghanistan	2020-01-05	NaN	0.0	NaN	NaN	0.0	NaN
3	AFG	Asia	Afghanistan	2020-01-06	NaN	0.0	NaN	NaN	0.0	NaN
4	AFG	Asia	Afghanistan	2020-01-07	NaN	0.0	NaN	NaN	0.0	NaN

weekly_icu_admissions	weekly_icu_admissions_per_million	weekly_hosp_admissions	weekly_hosp_admissions_per_million	total_tests	new_tests
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN

2-Missing Value Handling

It is clear that a large portion of the data is missing. Let's get the percentage of missing values for each column:

iso_code 0.000000
continent 4.752299
location 0.000000
date 0.000000
total_cases 10.688735
new_cases 2.699121
new_cases_smoothed 3.053321
total_deaths 16.799034

new_deaths 2.684492
new_deaths_smoothed 3.030533
total_cases_per_million 10.688735
new_cases_per_million 2.699121
new_cases_smoothed_per_million 3.053321
total_deaths_per_million 16.799034
new_deaths_per_million 2.684492
new_deaths_smoothed_per_million 3.030533
reproduction_rate 48.004636
icu_patients 89.381880
icu_patients_per_million 89.381880
hosp_patients 89.002642
hosp_patients_per_million 89.002642
weekly_icu_admissions 97.111822
weekly_icu_admissions_per_million 97.111822
weekly_hosp_admissions 93.421278
weekly_hosp_admissions_per_million 93.421278
total_tests 77.665713
new_tests 78.786549
total_tests_per_thousand 77.665713
new_tests_per_thousand 78.786549
new_tests_smoothed 70.751078
new_tests_smoothed_per_thousand 70.751078
positive_rate 73.012443
tests_per_case 73.456670
tests_units 69.956871
total_vaccinations 77.574279
people_vaccinated 78.536161
people_fully_vaccinated 79.470191
total_boosters 86.512552
new_vaccinations 81.513522
new_vaccinations_smoothed 48.617383
total_vaccinations_per_hundred 77.574279
people_vaccinated_per_hundred 78.536161
people_fully_vaccinated_per_hundred 79.470191
total_boosters_per_hundred 86.512552
new_vaccinations_smoothed_per_million 48.617383
new_people_vaccinated_smoothed 48.682933
new_people_vaccinated_smoothed_per_hundred 48.682933
stringency_index 44.393992

```
population_density                        15.091335
median_age                                21.045213
aged_65_older                             23.799757
aged_70_older                             21.836888
gdp_per_capita                            22.612245
extreme_poverty                           50.121677
cardiovasc_death_rate                     22.429097
diabetes_prevalence                       18.481695
female_smokers                            41.808529
male_smokers                              42.600204
handwashing_facilities                    61.998486
hospital_beds_per_thousand                31.516758
life_expectancy                           7.982580
human_development_index                   24.815093
population                                0.000000
excess_mortality_cumulative_absolute      96.564627
excess_mortality_cumulative               96.564627
excess_mortality                          96.564627
excess_mortality_cumulative_per_million   96.564627
dtype: float64
```

Many of our features hava a very large percentage of missing values that would make them rather unuseful. So we decided to delete the columns with more than 60% of its data being missing.
By doing so the and again obtaining the shape of the dataset (355449, 36) we can see the number of our features is reduced to 36. Still there is left some columns with missing value which we are going to fill through ***imputation methods***.

First let's see what percentage each remaining column have missing value.

```
iso_code                                  0.000000
continent                                 4.752299
location                                  0.000000
date                                      0.000000
total_cases                               10.688735
new_cases                                 2.699121
new_cases_smoothed                        3.053321
total_deaths                              16.799034
new_deaths                                2.684492
new_deaths_smoothed                       3.030533
total_cases_per_million                   10.688735
```

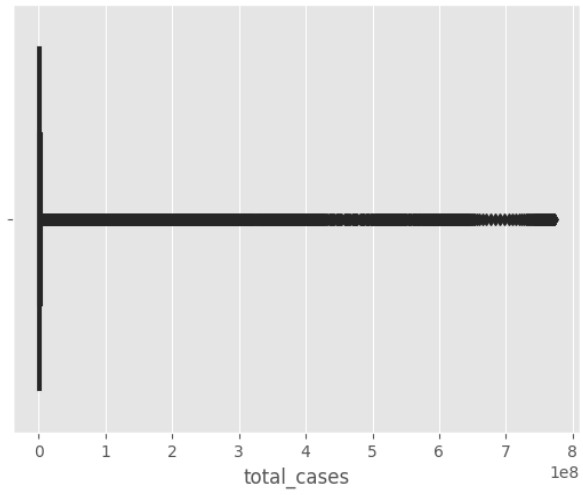
```
new_cases_per_million                        2.699121
new_cases_smoothed_per_million               3.053321
total_deaths_per_million                    16.799034
new_deaths_per_million                       2.684492
new_deaths_smoothed_per_million              3.030533
reproduction_rate                           48.004636
new_vaccinations_smoothed                   48.617383
new_vaccinations_smoothed_per_million       48.617383
new_people_vaccinated_smoothed              48.682933
new_people_vaccinated_smoothed_per_hundred  48.682933
stringency_index                            44.393992
population_density                          15.091335
median_age                                  21.045213
aged_65_older                               23.799757
aged_70_older                               21.836888
gdp_per_capita                              22.612245
extreme_poverty                             50.121677
cardiovasc_death_rate                       22.429097
diabetes_prevalence                         18.481695
female_smokers                              41.808529
male_smokers                                42.600204
hospital_beds_per_thousand                  31.516758
life_expectancy                             7.982580
human_development_index                     24.815093
population                                  0.000000
dtype: float64
```

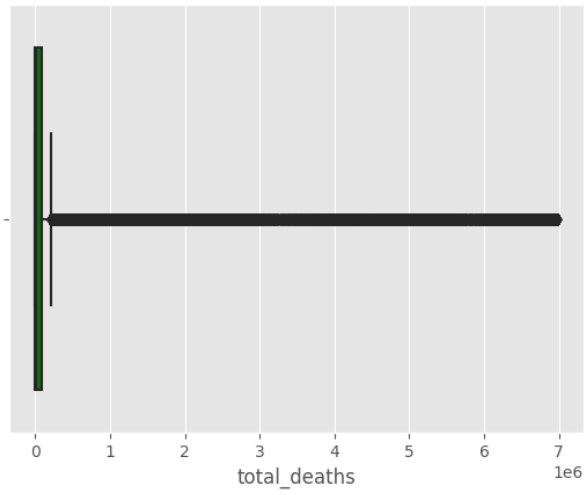
As you can see remaining columns have different data types. Some are numberical and some are categorical. For numbercials we choose to impute mean of each column and for categorical data we impute the most frequent value of each column.
Now that we have no missing values left in our dataset. It's time to go to the next step of this EDA process which is outlier handling.

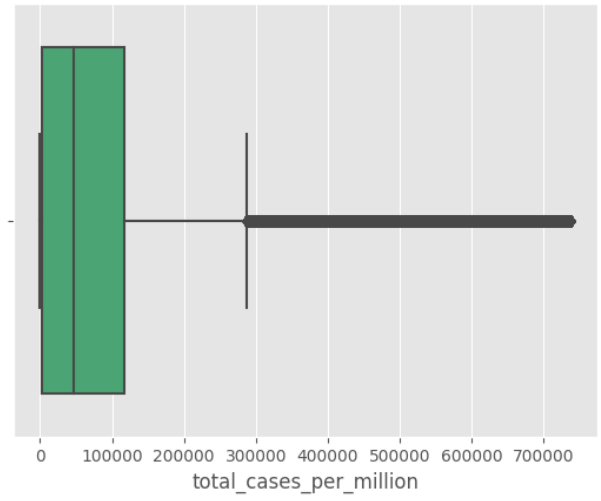
3- Outlier Handling

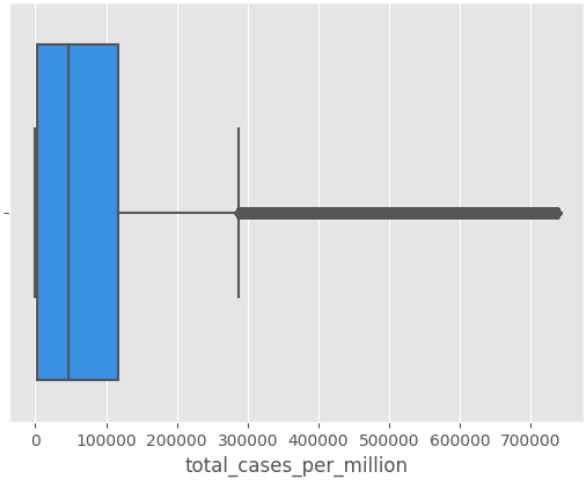
We use ***Interquartile Range (IQR) Method*** which is a way of dividing the data into four parts The upper quartile (Q3) is the value greater than or equal to 75% of the other values in the dataset, and the lower quartile (Q1) is the value greater than or equal to 25% of the other values. The IQR is calculated by subtracting Q1 from Q3. We remove any data that is fallen out of this upper and lower bound.

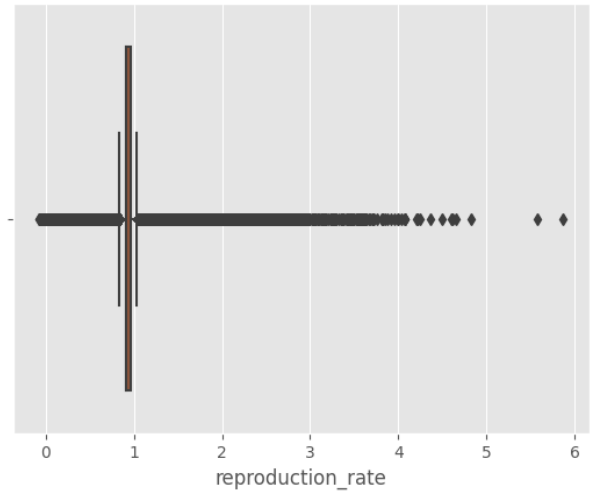
We can first visualize the distribution of some of the features which are going to be studied using boxplots.

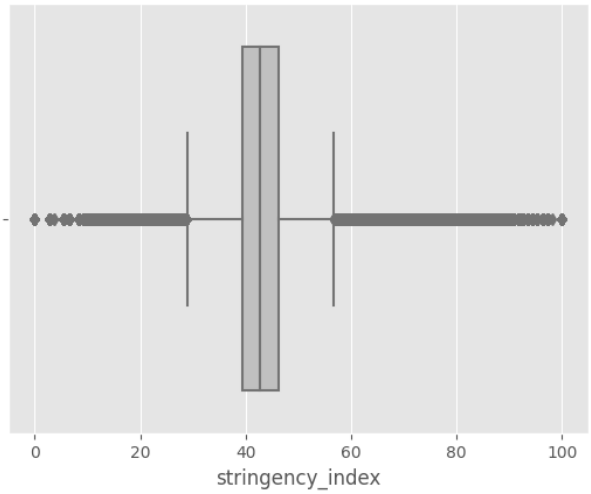


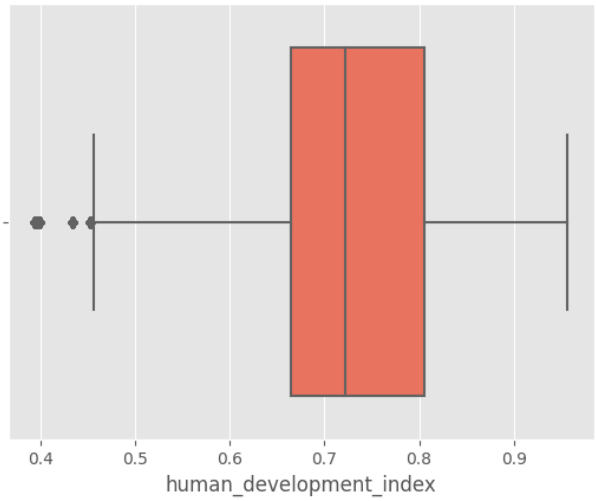


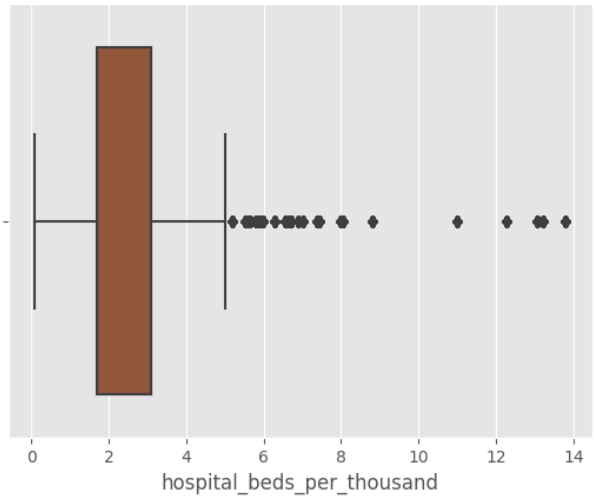


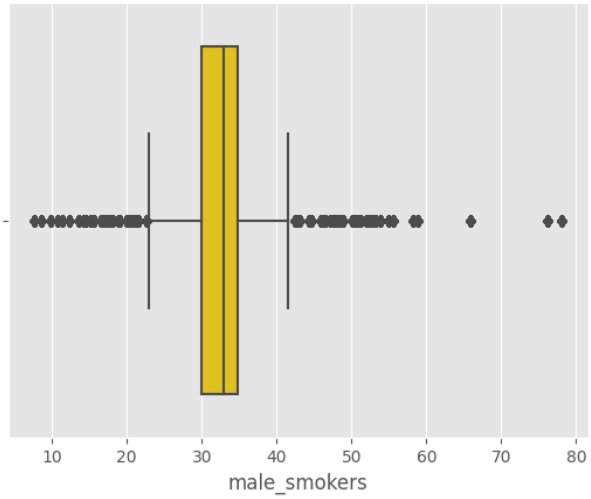


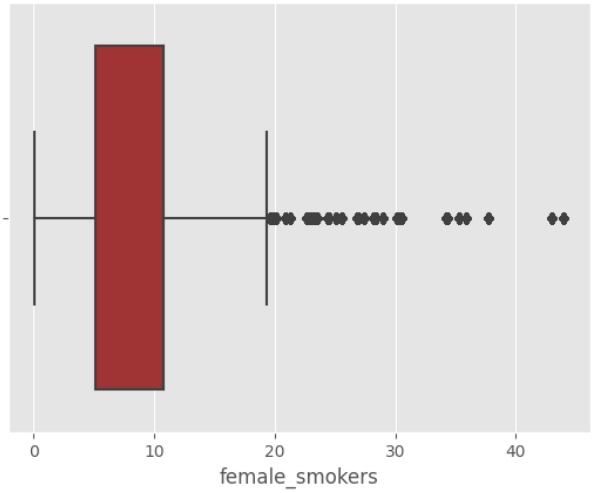


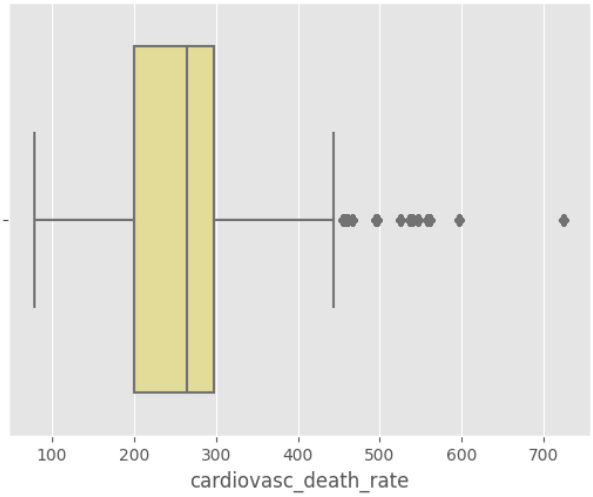


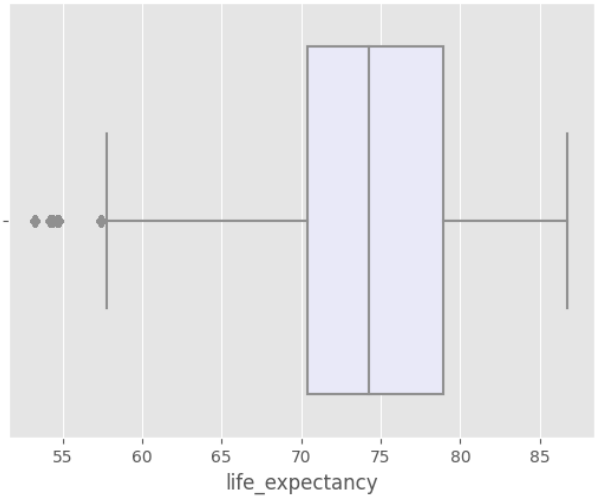

0.4
0.5
0.6
0.7
0.8
0.9
human_development_index


0
2
4
6
8
10
12
14
hospital_beds_per_thousand


10
20
30
40
50
60
70
80
male_smokers

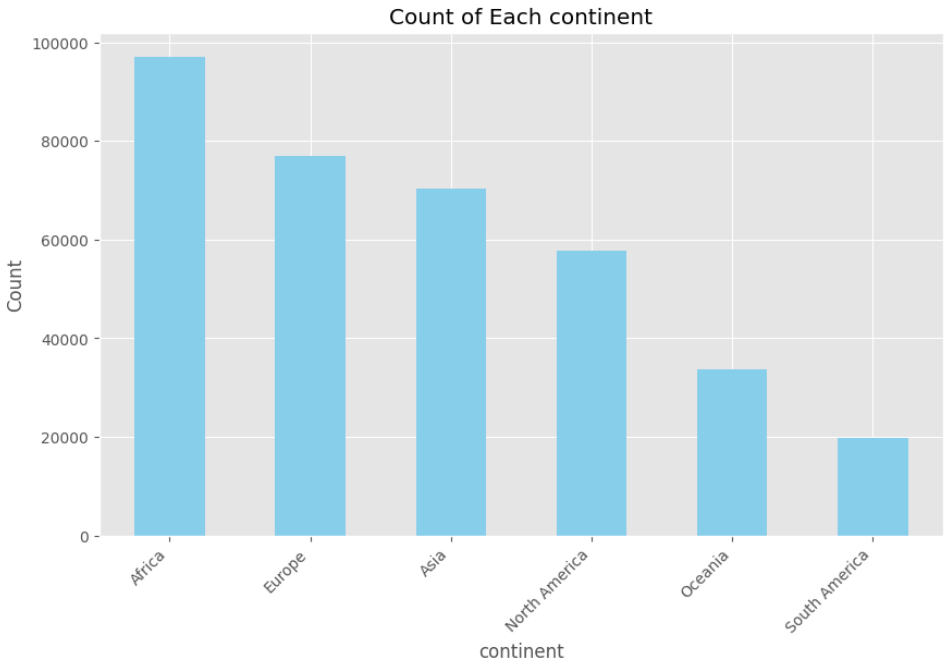

0
10
20
30
40
female_smokers


100
200
300
400
500
600
700
cardiovasc_death_rate

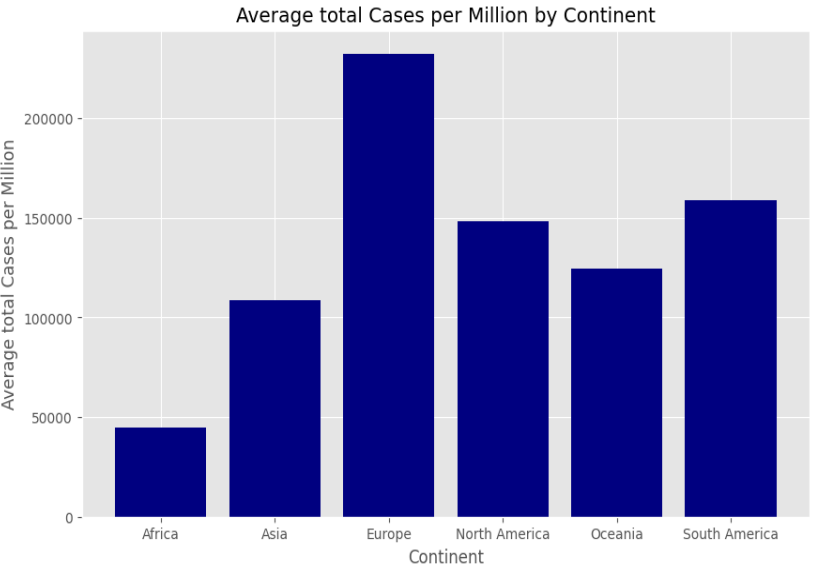

55
60
65
70
75
80
85
life_expectancy

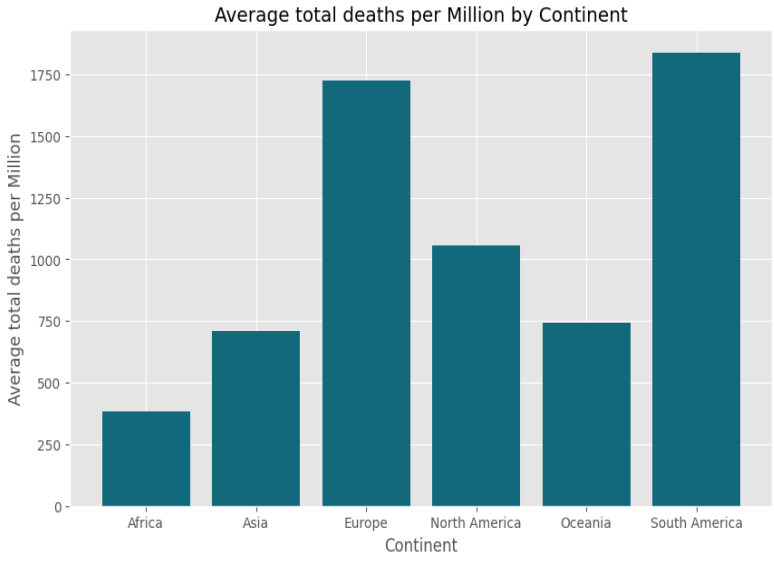
3-Visualization

The first feature that needs to be analyzed is the continent. Because of its general division of the world, it provides a platform for analyzing many features, worldwide. First, let's see what portion of our data is related to each continent.



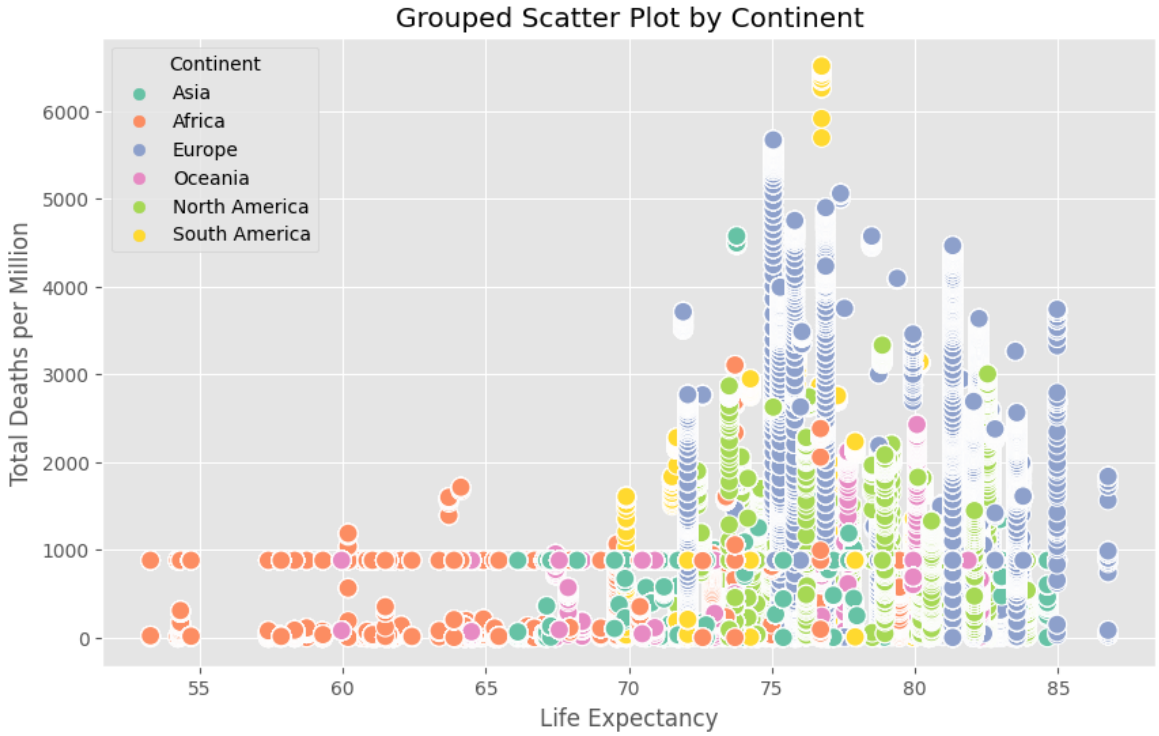
Most of our data is related to Africa. This means either Africa was the country most invaded by corona virus or simply most of the data recorded happens to be from this continent. In order to examine this question we can make use of new_cases_per_million and total_deaths_per_million features by plotting them continent-wise.





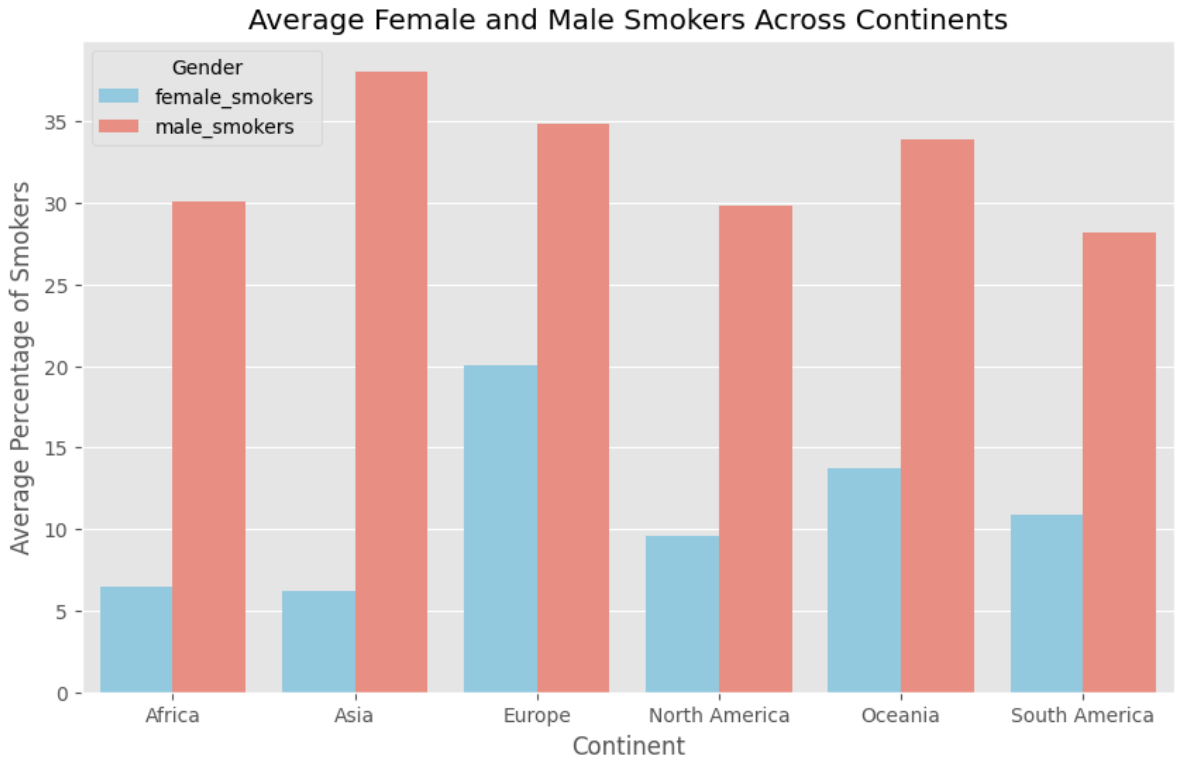
As these two plots indicate, Africa only had the most data samples from continents. the highest rate of people getting infected with coronavirus is from Europe and the highest rate of people dying from it, is from South America.

Since we have another feature *life_expectancy.* We suspect that there is a positive correlation between *life_expectancy* and *total_deaths_per_million.* We examine our suspicion with the help of a grouped scatterplot.



As shown above, it seems the dataset distribution is against our expectations. Since continents with higher life expectancy have more deaths per million.
Now let's examine the mentioned features plus *male_smokers* and *female_smokers* in order to examine the effect of coronavirus on smokers and nonsmokers males and females.

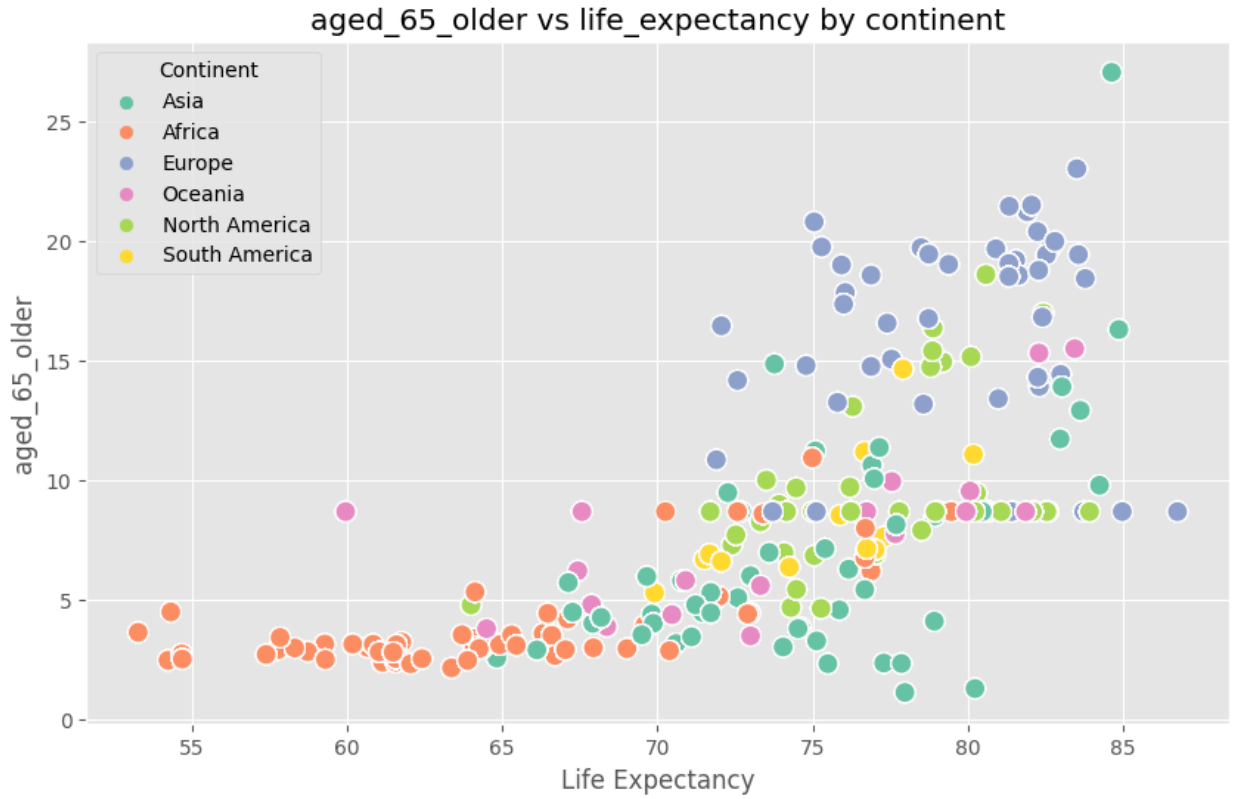
This heatmap implies a positive correlation between smoking and getting infected with coronavirus however it seems like its correlation for females is 3 times higher than for males. However this might be a false correlation since the majority of female smokers tend to be from more modern continents than more traditional ones. Let's examine this matter.



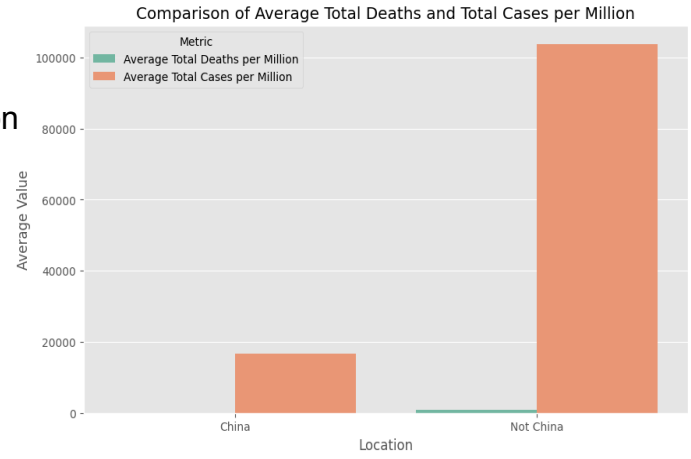
This plot relatively confirms our suspicion of a false correlation between gender and the effect of smoking on getting or dying from coronavirus.
To see if there is truly a correlation we can check their correlation across one continent. To obtain such a goal we create separate filtered data frames and a new heatmap for them. We chose two continents with the highest and lowest differences between male and female smokers portion.

As we suspected correlation between female smokers and total cases and total deaths is a lot less in Europe and it is also less than before in Asia. pay attention that in Europe the effect of female smokers is approximately twice that of men and in Asia it 4 times. This study confirms our previous guess however, it still seems like the relationship between smoking and getting infected with COVID-19 or dying from it is indeed higher in women than men.

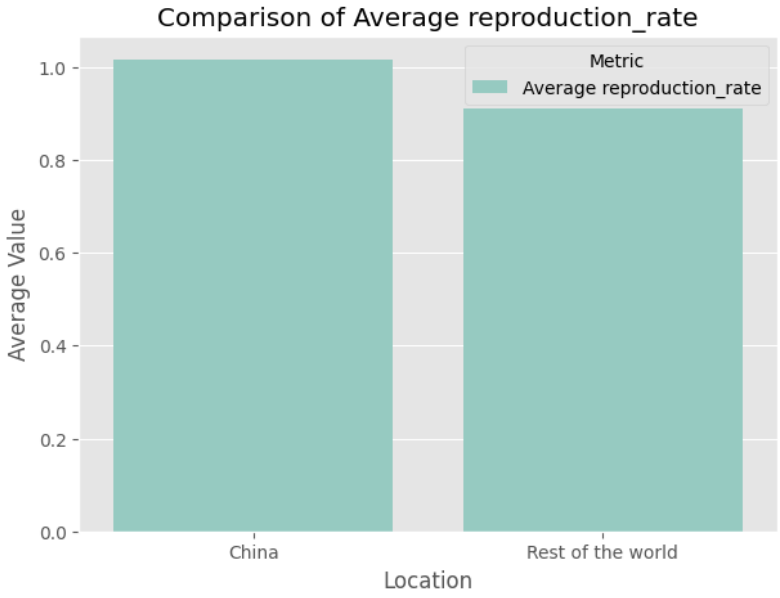
Now let's focus on another aspect of this data, as you can see in the general heatmap, there is positive total_cases_per_million and total_deaths_per_million which is in contrast with common knowledge. However, this might be because of the fact that the countries with higher life expectancy majority, are old people, which are the first target of getting infected and dying from a disease. The exponential relationship below confirms this matter.



Now let's have a better inspection on specific locations. We chose China as the source of COVID-19 as the first location to inspect. We created two separate data frames whose location is China and not to compare this country's interaction with COVID-19 to the rest of the world.

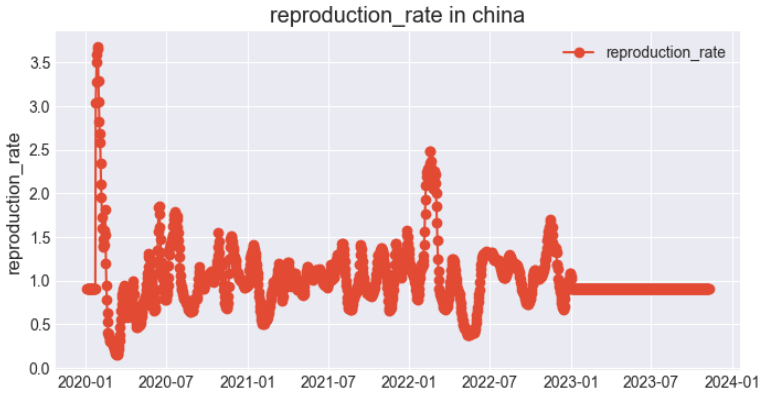


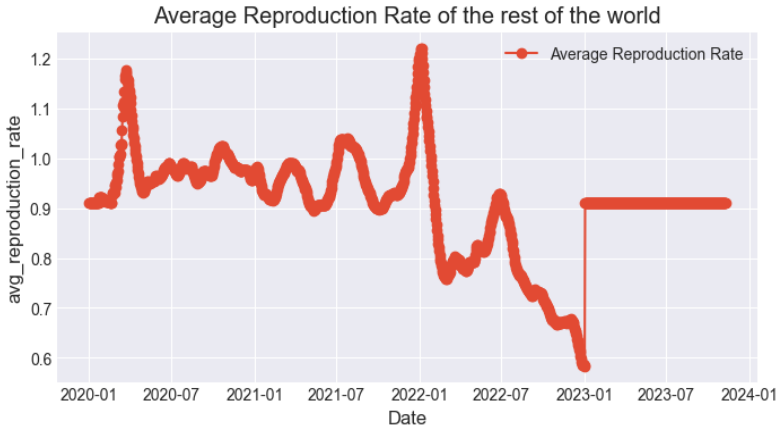
Since china was the source of spreading the Corona virus. What is the rate of reproduction in china vs rest of the world?



This plot confirms the common knowledge of China spreading the virus more compared to the rest of the world. However, it didn't continue to have more deaths and infections.

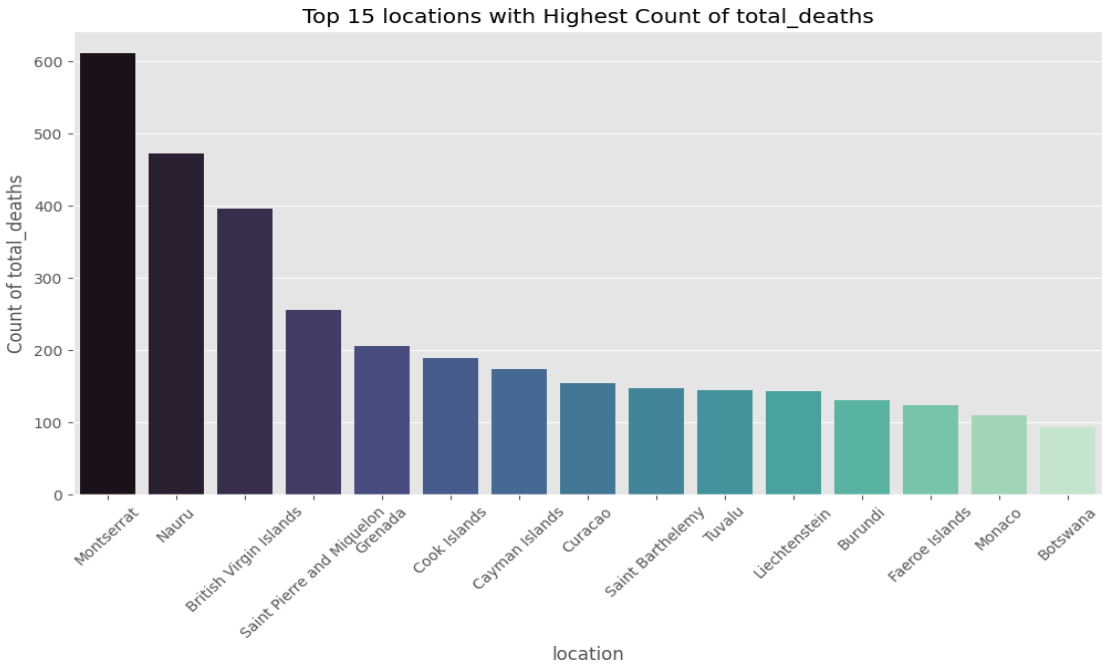
Below we can see a time series plot of reproduction_rate per month for china vs the rest of the world.

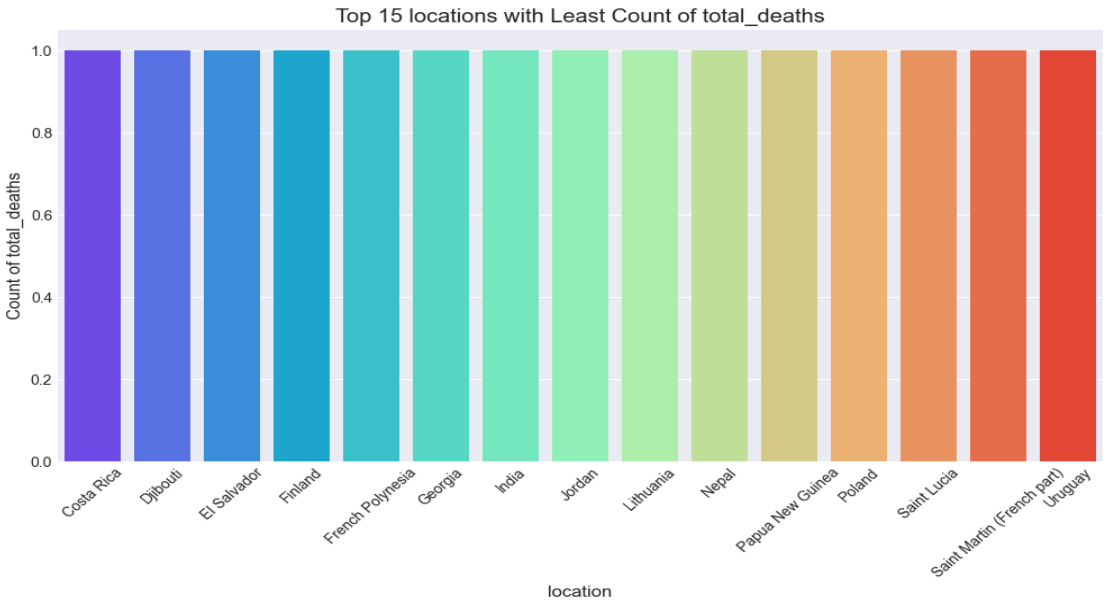




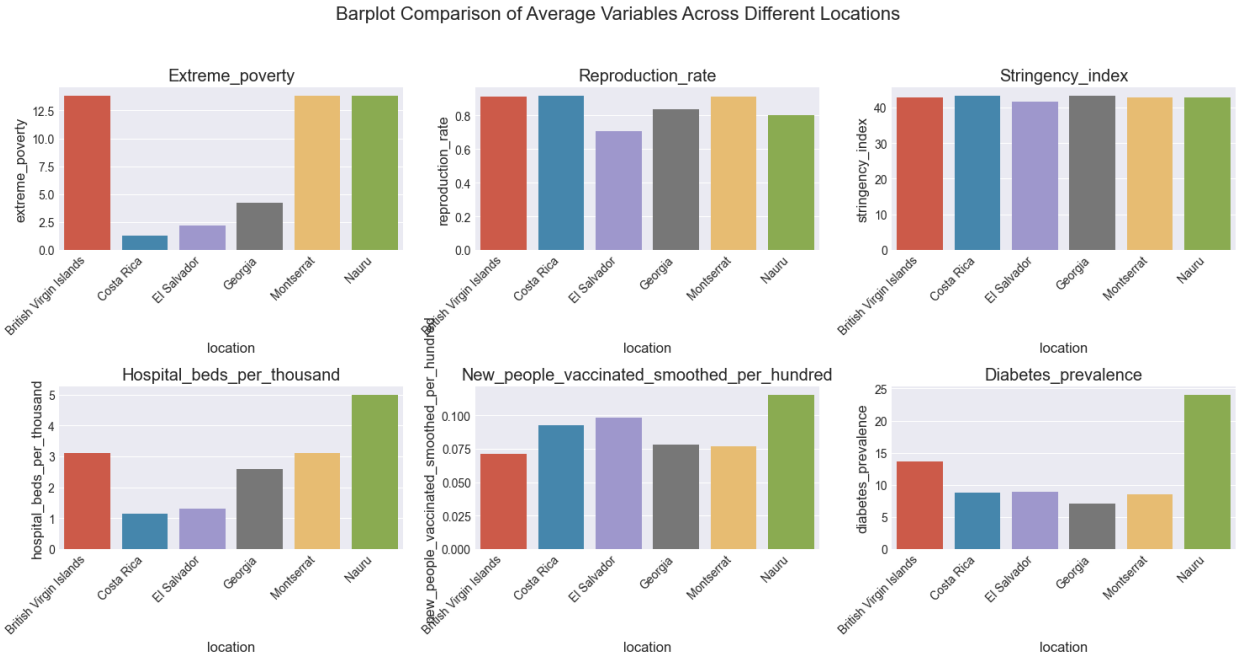
Since we have a daily log of reproduction_rate per day in many locations. We introduced a new feature the average reproduction rate of each day for the rest of the world.
What the comparison of these two plots tells us is that China started infecting the rest of the world at a very high rate it eventually reduced and stayed at approximately the same rate. While the rest of the world's reproduction rate reduced to a much smaller amount than China.

Now moving from China, let's examine other locations. Which are the top 15 locations with the highest amount of total deaths in a day recorded? Which ones has the least? Two plots below is our answer to these questions.





It looks like Montserrat in North America had the highest deaths and the top 15 least deaths all had equal amounts of 1 death recorded.
Let's more specifically compare the top 3 locations from each set (the least deaths are selected randomly).



What we can understand from these plots is that extreme poverty, hospital beds per thousand, and diabetes prevalence demonstrate a significant difference between these two groups.

Diabetic people are at more risk of dying from coronavirus, and countries with extreme poverty are also at risk, however, countries with fewer hospitals per thousand are the ones that are responsible for more deaths. But this can be justified by the fact that these per thousand isn't per patient. And the people infected with coronavirus as expected are fewer in these countries so they wouldn't need more hospital beds.

Conclusion:

Critical facts we concluded via this dataset:

1- Europe had the most amount of total cases recorded.

2- South America had the most amount of total deaths recorded.

3- smoking has a positive correlation with getting infected with COVID-19 and this correlation is larger among women.

4- countries with higher life expectancy had higher amounts of deaths but this is because these countries are populated with more old people and as you know old people are much more at risk.

5- china was responsible for reproducing COVID-19 a lot more than other countries however it had fewer deaths and infection records.

6- diabetes and extreme poverty have a significant impact on dying from COVID-19.