Early Data Exploration for COVID - 19

Studying correlation of Population Density, Life Expectancy and Cardiovascular predisposition to SARS-CoV-2 cases

James Casey
Yishu Qu
Philip Melanchthon

March 17th 2021

MSDS: 430 Python for Data Science

Introduction

The COVID-19 pandemic has been one of the most significant events in recent history, which can be seen by its social and economic impacts worldwide. We are interested in determining whether any relationships or correlations can be recognized from COVID-19 datasets sourced for this project. In particular, we would like to investigate specific factors that affect the total number of cases and deaths of COVID-19 with the goal of aiming to find some potential strategies to help mitigate the pandemic.

The dataset includes not only COVID epidemiology data such as confirmed cases of COVID-19, deaths attributed to COVID-19 and real-time estimate of the effective reproduction rate of COVID-19. It also includes plenty of other public health data involving life expectancy, population density, death rate from cardiovascular disease, hospital beds per 1,000 people, diabetes prevalence etc. It is an abundant resource for us to test different dependency hypotheses. The data is one single csv file and was retrieved from Our World in Data. It is updated daily and includes data on confirmed cases, deaths, hospitalizations, and testing, as well as other variables of potential interest across different countries worldwide. Some of the primary sources that were used to collect variables of interest are the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), European Centre for Disease Prevention and Control (ECDC), United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.

With this data, we wish to examine COVID-19 and recognize various trends associated with the number of cases and deaths from the pandemic. We hope to learn which factors affect the amount of COVID death/cases worldwide. Specifically, the factors we are interested in include life expectancy, population density, and cardiovascular related deaths. We believe that the data will reveal that:

- Life Expectancy in a given region is negatively correlated to COVID Deaths.
- Population density is positively correlated with total COVID cases.
- Total COVID cases are positively correlated with cardiovascular related deaths.

By preparing the data and performing analysis on the dataset, we will attempt to determine whether these correlations exist.

Data Preparation and Analysis

Before analysis could take place, data preparation and cleaning needed to be completed. For this to occur, a complete understanding of the data and what is necessary for our analysis needs to be present. Therefore, it is important to identify the variables of interest, which include: Date, Location, Life Expectancy, Total Deaths, Population Density, Total Cases, and Cardiovascular related deaths. The following provides an in depth description of each variable:

Variable	Description	Var. Name	Value Count	Object Type	Purpose
Date	Date of observation	date	63572 Values	object string	will be used to for observing progression of cases/deaths
Location	Country where COVID case(s) are observed	location	63572 Values	object string	will be used to compare COVID cases/deaths amongst different countries
Total Cases	Total confirmed cases of COVID-19	total_cases	62978 Values	float64	will be used in correlations
Cardiovascular	Death rate from	cardiovas	61134	float64	will be used in correlation with total

Related Deaths	cardiovascular disease (deaths per 100,000 people)	c_death_rate	Values		COVID cases
Population Density	Number of people divided by land area, measured in square kilometers, most recent year available	population_density	61875	float64	will be used in correlation with new COVID cases
Life Expectancy	Life expectancy at birth in 2019	life_excpectancy	62886	float64	will be used in correlation with Total COVID deaths
Total Deaths	Total deaths attributed to COVID-19	total_deaths	54354	float64	will be used in correlation with Life Expectancy

Table 1: Data Dictionary / Details for EDA Project.

Since these variables were the only ones necessary for the analysis, all other columns were cut. Next, rows with missing data were identified and eliminated. A ‘World’ country was also identified. Since this will not be useful in the analysis, these rows were also removed. After this, the dataset was updated to only include the last date of collection, February 2, 2021. This would ensure that the values for each country are the final totals for the date range of interest. Lastly, any potential outliers identified and assessed via boxplots that can be found in the Appendix. The following are the decisions made on the potential outliers:

- Cardiovascular Rate boxplot shows two outliers. These values are static for each country. If dropped, we will lose data for an entire country. Therefore, we will not drop these.
- We recognized three potential outliers for the total_cases column. These will not be dropped since they are the United States, India, and Brazil, and will likely have a high amount of cases.
- There is one potential outlier Population Density. Once again, since this number is consistent for the country, we do believe this is correct and do not want to remove an entire country's data.
- There are no potential outliers for Life Expectancy.
- We recognized two potential outliers for Total Deaths. Similar to total cases, these were also the United States and Brazil and will not be dropped. Since it is similar to total cases, this gives us confirmation that both should be kept.

With the data cleaned, the dataset for analysis included 171 countries in total. Each country had one observation for each variable. The descriptive characteristics of the data can be found in a table in the Appendix.

To start analysis, pairplots, which can be found in the Appendix, were made to explore the correlations between each variable. We found that all of our selected variables are not normally distributed, especially, the total cases, population density, and total deaths, which are highly skewed in distribution. We also observed patterns that did not agree with our hypotheses.

Correlational heatmaps were then created to observe whether or not there are any correlations between the relationships of interest:

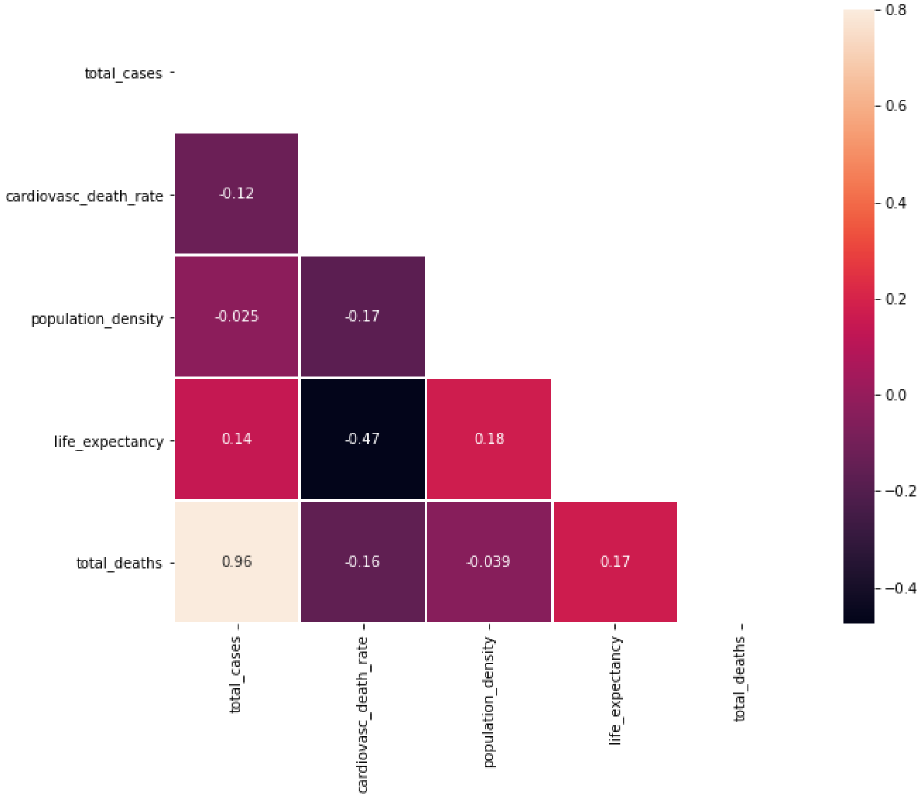
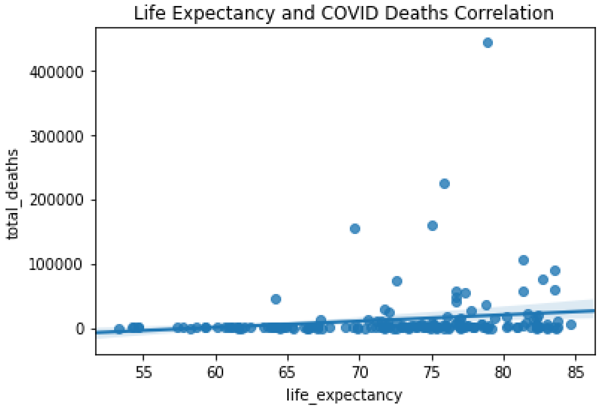


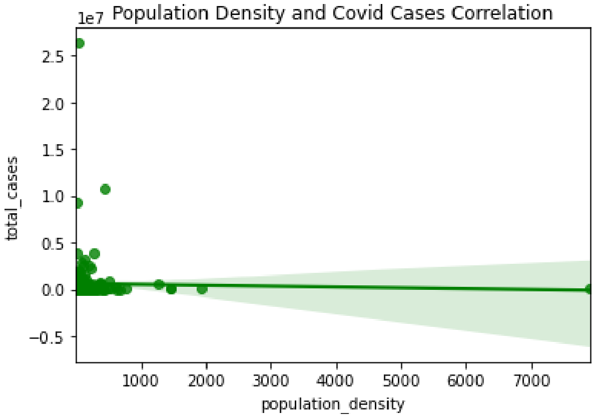
Fig 1: Correlations of Numeric Variables

Looking at relationships we are interested in, we can see that:

- There is a weak positive correlation between life expectancy and total COVID deaths (0.17). However, it was expected that a higher life expectancy would be negatively associated with total COVID deaths.
- There is almost no correlation between population density and total COVID cases (-0.025).
- There is low negative correlation between total COVID cases and cardiovascular related deaths (-0.12). It was expected that total COVID cases are positively correlated with cardiovascular related deaths.

Next, linear relationships between variables of interest were represented by scatterplots and regression models.





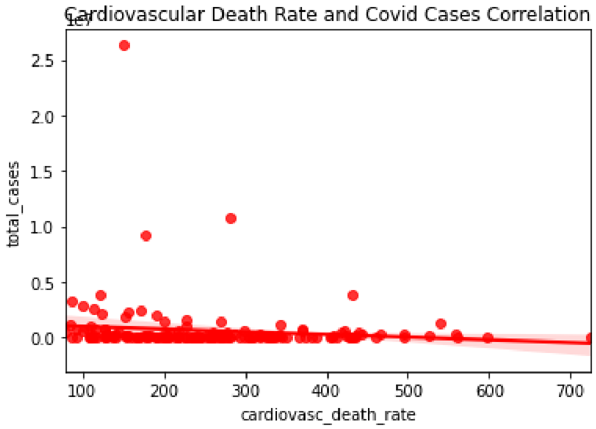


Fig 2: Scatterplots for Relationships Being Investigated

Relationship	**Slope**	**P-value**	**R^2**
Life Expectancy and COVID Deaths	976.93	0.03	0.03
Population Density and Total Covid Cases	-91.74	0.74	0.0006
Cardiovascular Death Rate and Total Covid Cases	-2427.19	0.11	0.02

Table 3: Slope, P-value, and R^2 for Linear Regressions

The slopes of univariate linear regression and the scatterplots suggest:

- As life expectancy of a country increases, COVID deaths increase.
- As population density increases, total cases decrease.
- As cardiovascular death rate increases, total cases decrease.

However, after considering the low P-values and R-squared values of linear regression models we can conclude that the only significant linear relationship could occur by chance and these relationships are not correlated. In order to further investigate this, a sensitivity analysis was performed by excluding potential outliers identified in the data preparation step. We didn't observe improvements in regards to R-squared values and p-values. Also, some of the slopes changed from positive to negative and vice versa. Therefore, there is not a correlation between the variables we are interested in.

Lastly, we categorized countries by population density, ranking in order to see if there is any apparent separation between different classes and if the correlation between population density and total COVID cases differs in each category. We split up data into tertile groups (low, medium, high) based on population density. We observe that the three groups have different mean total cases (low: 875760, medium: 327799, high: 563454). The medium population density group has the lowest number of total cases while the low group has the largest number of total cases.

We also created pairplots colored by population density categories:

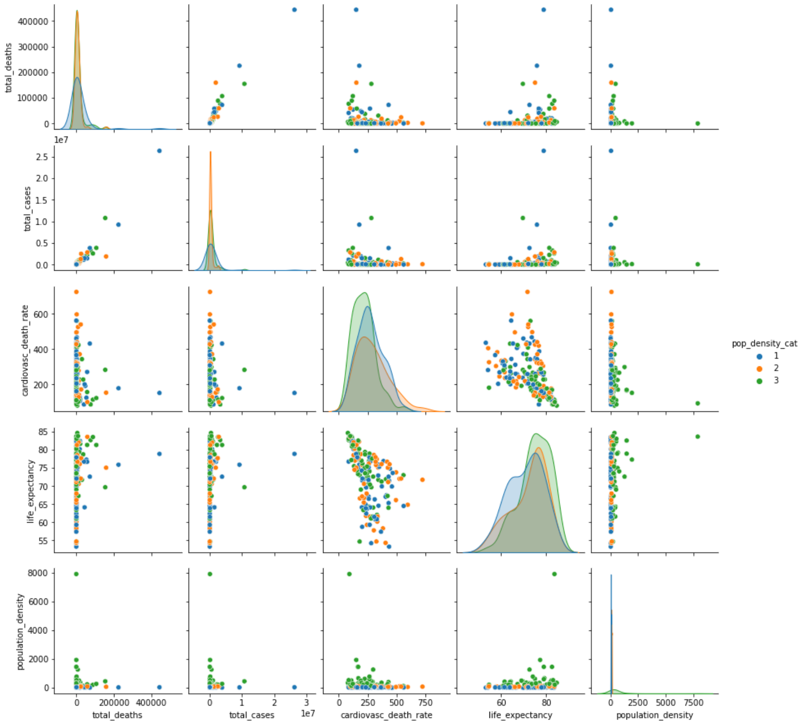


Fig 3: Seaborn Pairplot for Variables Coloring By Population Density Category

There are no clear segmentations between three population density groups for the relationships we hypothesized, especially for population density and total COVID cases. One can conclude for each population density category, no correlations or trends were observed for any of the relationships of interest.

Conclusion

Our main difficulty in the EDA project is the problem of outliers and highly skewed data. We included all the countries in the analysis for integrity. However, the methods may not be appropriate for data with non-normal distribution.

We did not identify any factors that are correlated with COVID cases/deaths. It turns out we cannot prove the initial hypotheses based on current evidence and come to the following conclusion:

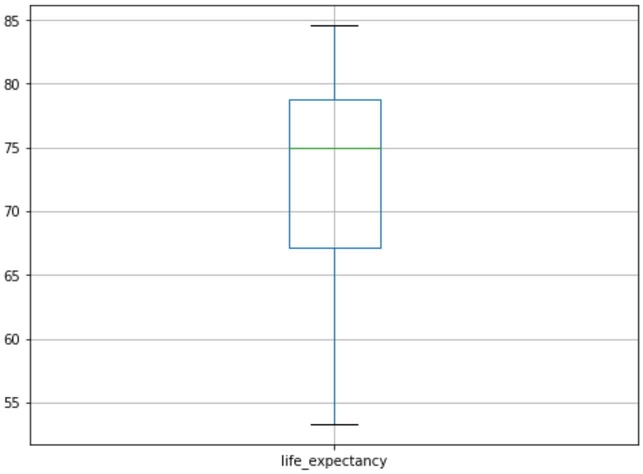
- Life Expectancy in a given region is not correlated to COVID Deaths.
- Population density is not correlated with total COVID cases.
- Total COVID cases are not correlated with cardiovascular related deaths.

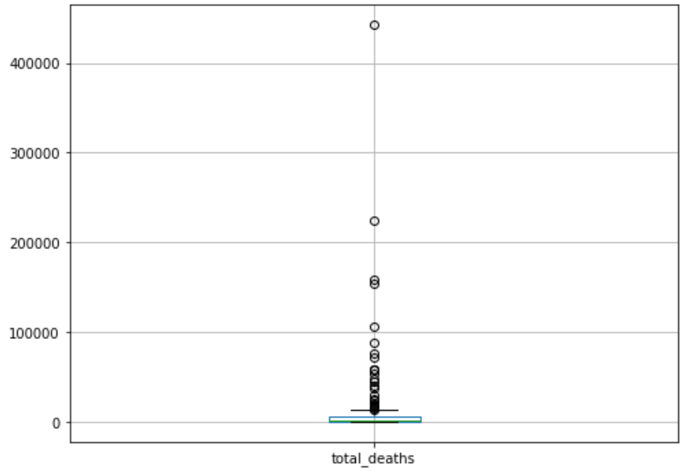
Possible next investigative steps includes:

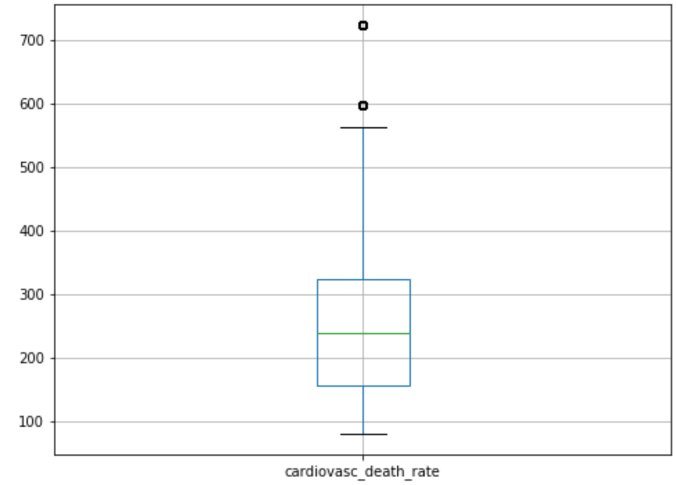
1) Consider other variables that could contribute to COVID cases/deaths (e.g. number of hospital beds).
2) Consider multivariable linear regressions between COVID cases/deaths and public health measures
3) Consider whether the interactions between these variables could affect COVID cases/deaths.

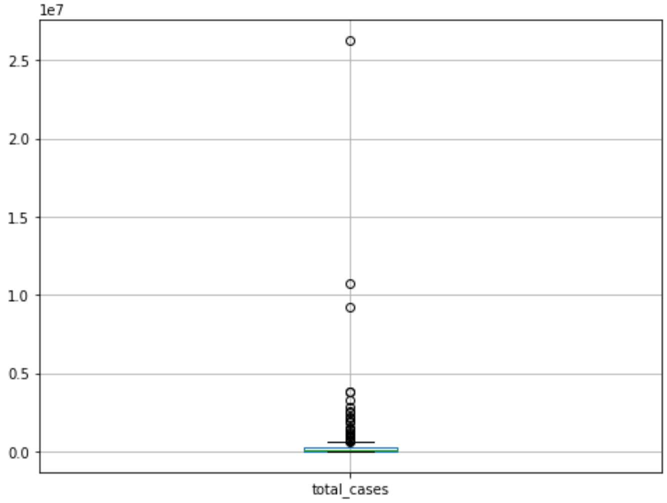
Appendix

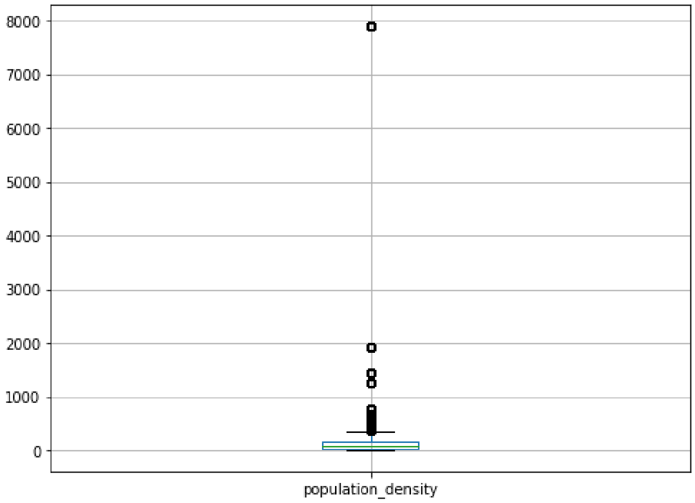
Variable Boxplots







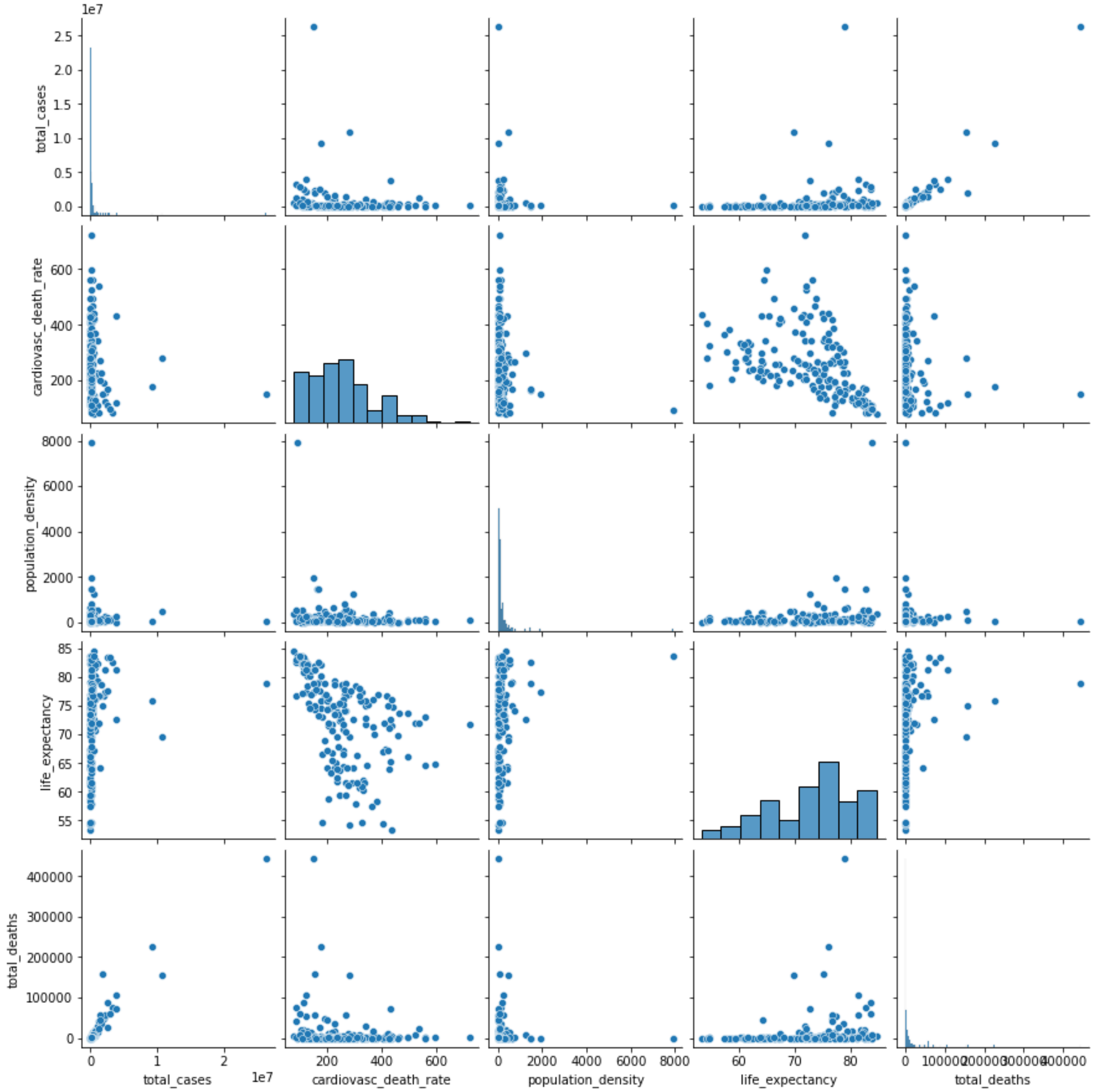




Descriptive Statistics of Variables

	Total Cases	Cardiovascular Death Rate	Population Density	Life Expectancy	Total Deaths
Count	171.00	171.00	171.00	171.00	171.00
Mean	600926.00	257.71	206.53	72.65	13005.19
Std	2338308.00	118.29	646.53	7.61	43706.52
Min	55.00	79.37	1.98	53.28	1.00
25%	10860.00	169.38	36.07	66.87	152.00
50%	78755.00	242.65	81.72	74.30	959.00
75%	264734.50	324.20	205.14	77.94	5484.50
Max	26321120.00	724.42	7915.73	84.63	443355.00

Pairplots of Variables


1e7
total_cases
cardiovasc_death_rate
population_density
life_expectancy
total_deaths
total_cases
1e7
cardiovasc_death_rate
population_density
life_expectancy
total_deaths