Kristofer Siimar

Covid-19 Health Effects Project Report

Motivation

Risk factors of covid-19 cause deaths is a widely discussed topic in today's world. Some of the factors doctors say to increase the probability of a person dying to covid-19 are older age, smoking, and obesity. Correlations between those factors and covid-19 caused deaths are already well analyzed; therefore, I decided to do my project on analyzing those factors and covid-19 caused deaths on countries' level - finding if countries with higher older population proportion have more on covid-caused deaths.

My research questions were:

Do countries with a higher proportion of older people have more covid-19 caused deaths?
Do countries with higher smoking rates have covid-19 deaths?
Do countries with higher obesity rates have more covid-19 deaths?

Data Sources

Dataset 1

Name: oldpop.csv

Size (in records and/or bytes): 246 rows, 65 columns

Location (URL or other method): https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS

Format: csv

Access Method: downloading csv

Variables: Country Name, 2018

Time Period: 2018

Dataset 2

Name: covid-19.csv

Size (in records and/or bytes): 276 rows, 6 columns

Location (URL or other method): https://documenter.getpostman.com/view/10808728/SzS8rjbc?version=latest#00030720-fae3-4c72-8aea-ad01ba17adf8

Format: json

Access Method: API

Variables: Country, TotalConfirmed, TotalDeaths

Time Period: 2020

Dataset 3

Name: smoking.csv

Size (in records and/or bytes): 1174 rows, 6 columns

Location (URL or other method): https://worldpopulationreview.com/countries/smoking-rates-by-country/

Format: csv

Access Method: downloading csv

Variables: name, totalSmokingRate

Time Period: 2020

Dataset 4

Name: obesity.csv

Size (in records and/or bytes): 1200 rows, 6 columns

Location (URL or other method): https://ourworldindata.org/obesity

Format: csv

Access Method: downloading csv

Variables: Entity, Year, Total adults who are obese (%)

Time Period: 2017

Dataset 5

Name: population.csv

Size (in records and/or bytes): 263 rows, 59 columns

Location (URL or other method): https://www.worldometers.info/world-population/population-by-country/

Format: csv

Access Method: downloading csv

Variables: Country, Year_2016

Time Period: 2016

Dataset 6

Name: testing.csv

Size (in records and/or bytes): 128 rows, 8 columns

Location (URL or other method): https://www.worldometers.info/world-population/population-by-country/

Format: csv

Access Method: downloading csv

Variables: Country or region, Date, Tests

Time Period: 2019 - 2020

Data Processing

Before any data processing, I wrote a code 'covid_data.ipynb' that took a json file and read it into a csv file to better understand the dataset before any processing.

The first data processing step I took was extracting the most recent data by using the loc method. After that, I changed the headers to follow the same format (WordWord) to keep the header style consistent throughout all the data sets. Additionally, I had to change data types and format on the testing.csv 'Tests' column - from strings that included commas to float. Furthermore, I had to create data corresponding to 100k people to eliminate the size difference bias and to better visualize the data.

Problems:

- We need to have similar keys to merge data.
    - I used the 'rename' methode to change columns that had country names to 'Country'.
- To better understand the datasets, we need to eliminate the unuseful data.
    - I used the drop function to drop columns that I didn't need for the analysis.
- Some countries are in the covid-19 dataframe but not in others.
    - I used inner joins for each merging to have a complete data set.
- We need to have testing data in testing.csv float type to calculate and visualize testing and population size correlation.
    - I defined a function to eliminate commas ('replace' function) from the data and then turned it into a float. After defining the function, I used the 'apply' function to change the data into float type.
- We should have understandable headers that are formatted consistently.
    - I used the 'rename' method to rename columns and to format headers consistently.
- We should have covid deaths and testing data per 100, 0000 people in order to eliminate the population size difference bias from the results.
    - I defined functions that took in testing, deaths and population country for each country and produced tests taken and deaths per 100,000 people for each country. Then I applied the function to the merged data frame and created new columns consisting of the produced data.

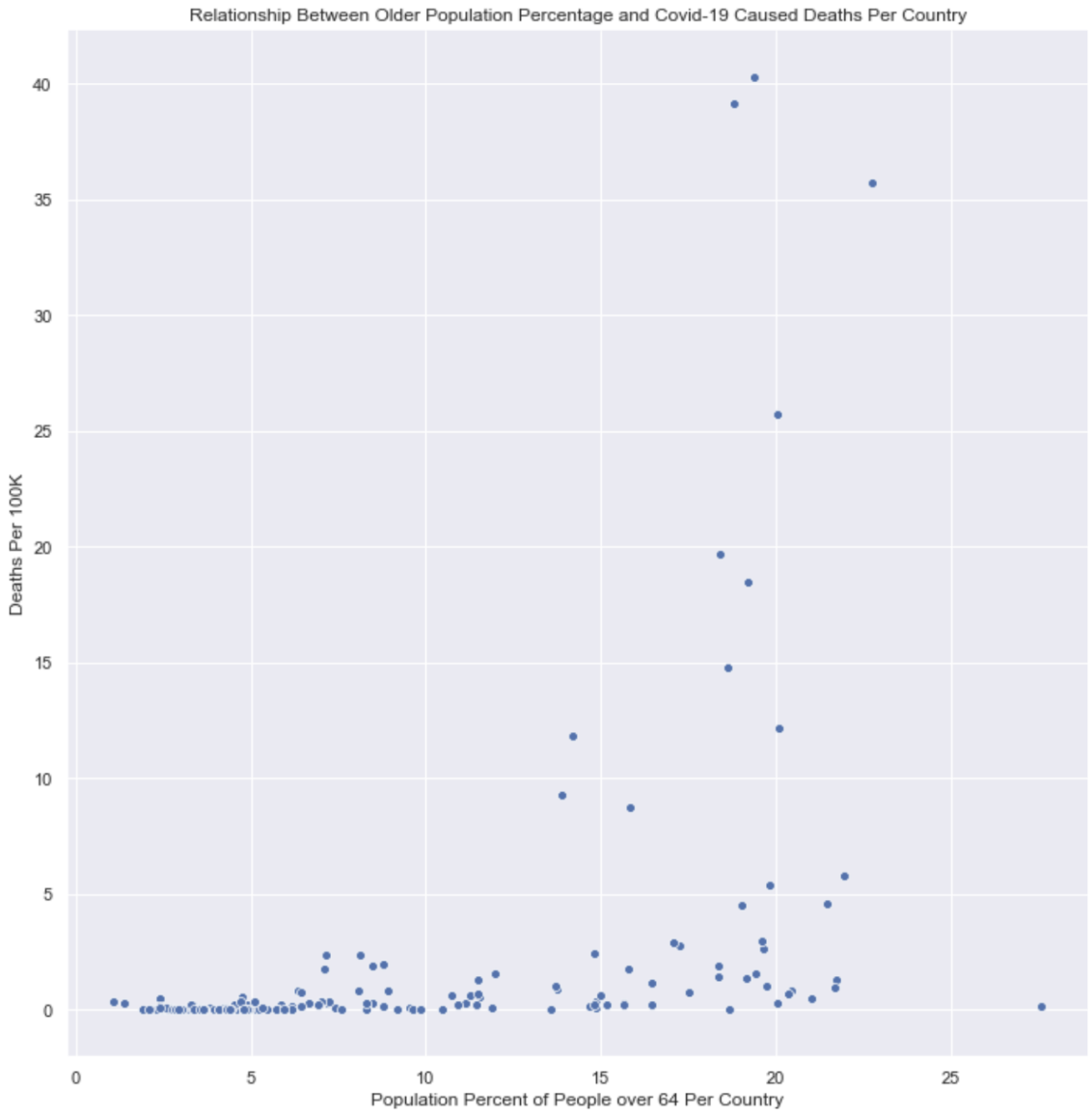
Analysis and Visualization

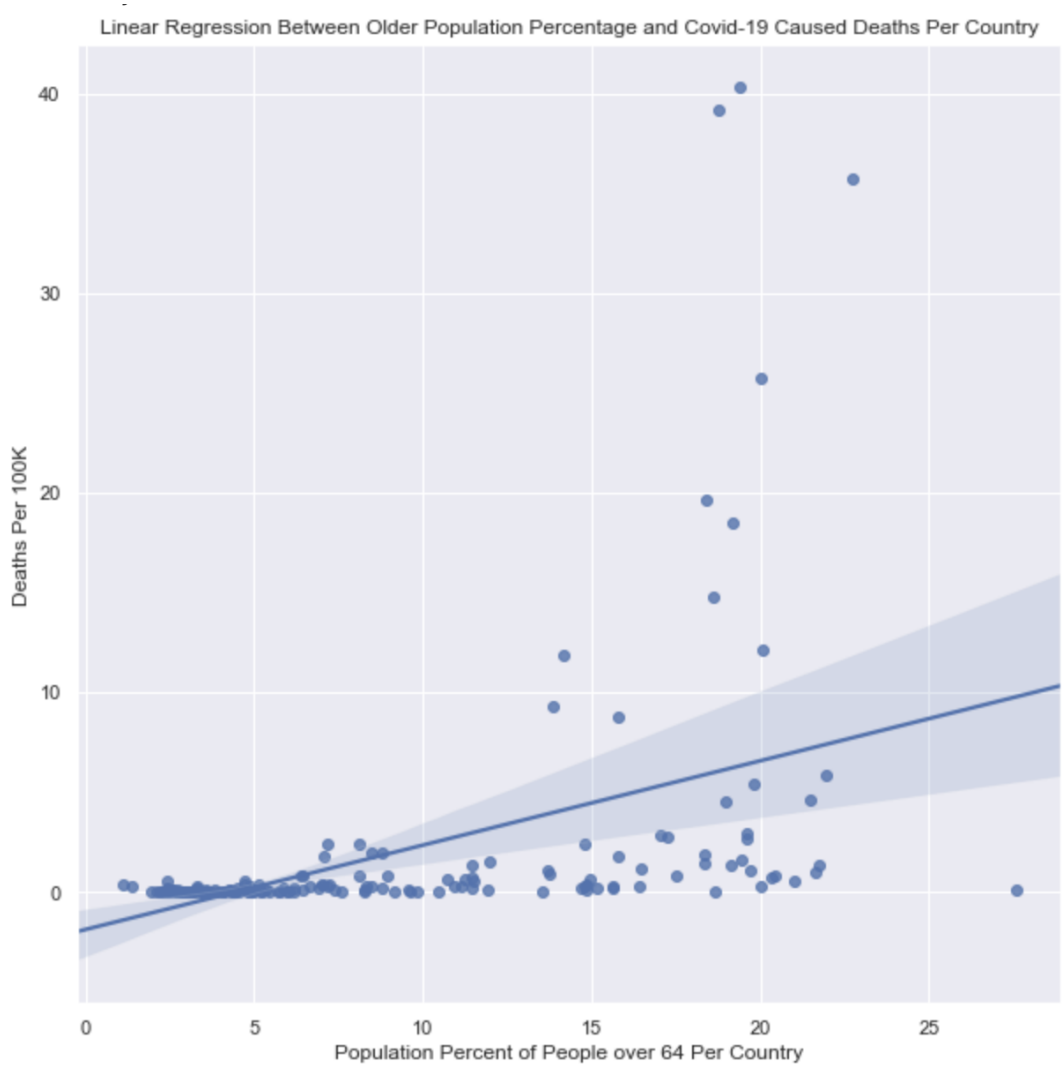
a) First I analyzed the correlation between the proportion of older population (over 64 years) and covid-19 caused deaths by country.

   Note: the analysis is performed with a code named ‘deaths_age.ipynb’.

   First of all, I read in all the necessary data frames (covid-19.csv, oldpop.csv, population.csv, testing.csv). Then, I dropped the unnecessary columns and renamed the remaining ones. Furthermore, I changed the ‘Tests’ column’s data from string to float in testing.csv. Then I merged all the described data frames on the ‘Country’ column. Besides that, I generated tests and deaths for 100k people columns to eliminate the population size difference bias.

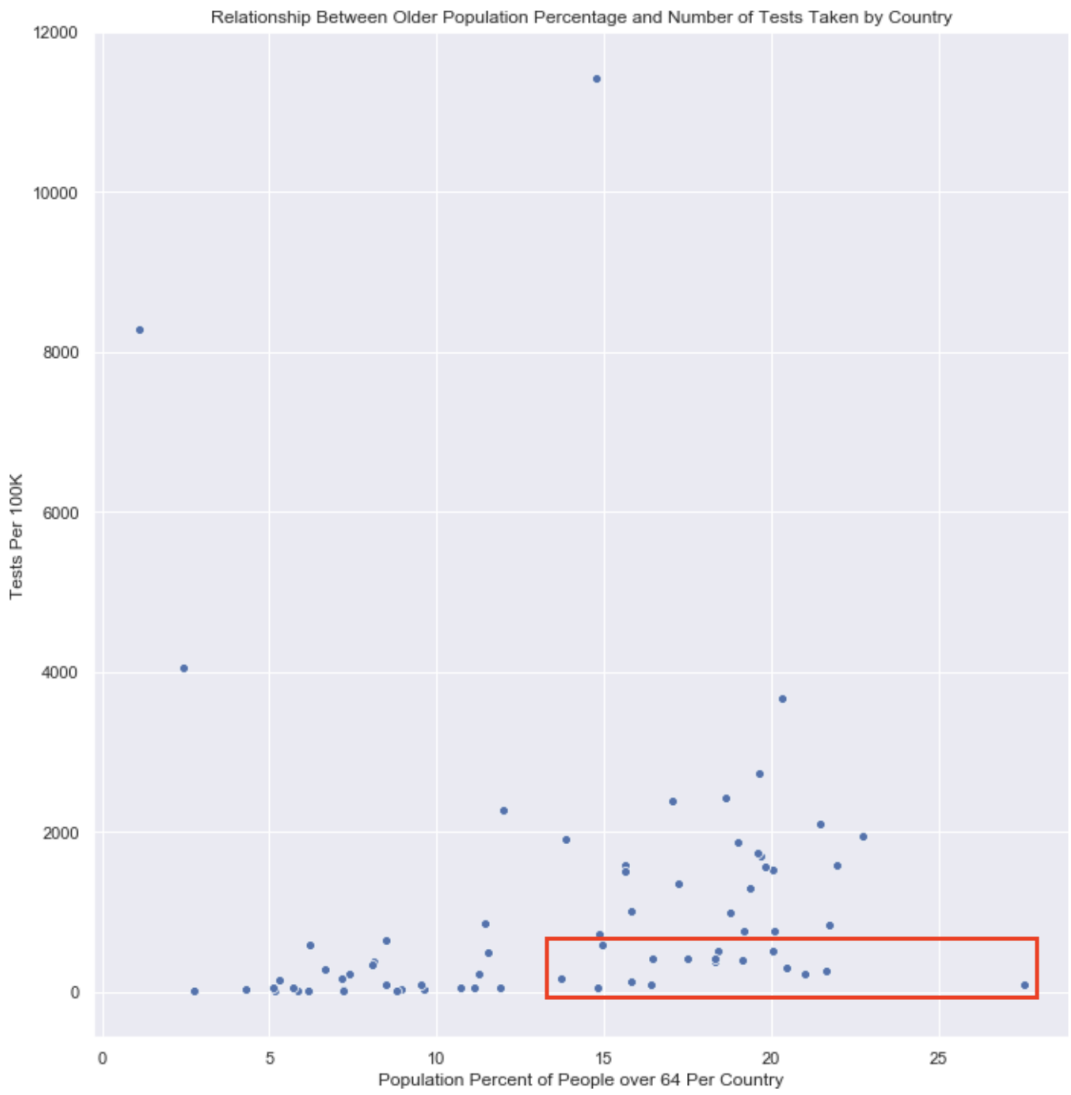
   After all the data was ready for analysis, I created a scatterplot and a regression plot of the relationship between older population percentage and covid-19 caused deaths per 100k people by country.


Relationship Between Older Population Percentage and Covid-19 Caused Deaths Per Country
Deaths Per 100K
0
5
10
15
20
25
30
35
40
0
5
10
15
20
25
Population Percent of People over 64 Per Country



My hypothesis was that countries with a greater proportion of older people (over 64 years) have a greater number of covid-19 caused deaths because older people are at higher risk of losing life to the virus.

From the graphs, we see a moderate correlation that countries with a higher proportion of older people (over 64 years) have a greater number of covid-19 caused deaths. However, there are still some countries that have a relatively high proportion of older people but have a low amount of deaths. One reason for that might be the number of tests countries perform - countries with low capacity of testing do not know if and how many people have actually died from covid-19.



From the graph, we see that some countries that have a higher percentage of older people haven't performed many tests compared to other countries. Another reason for that could be that some countries with a greater proportion of older people have a better healthcare system which enables them to save many of their older people that are at risk of dying.

Overall, based on the analysis, we can say that countries with a greater proportion of older people have more covid-19 deaths, relative to their population size.

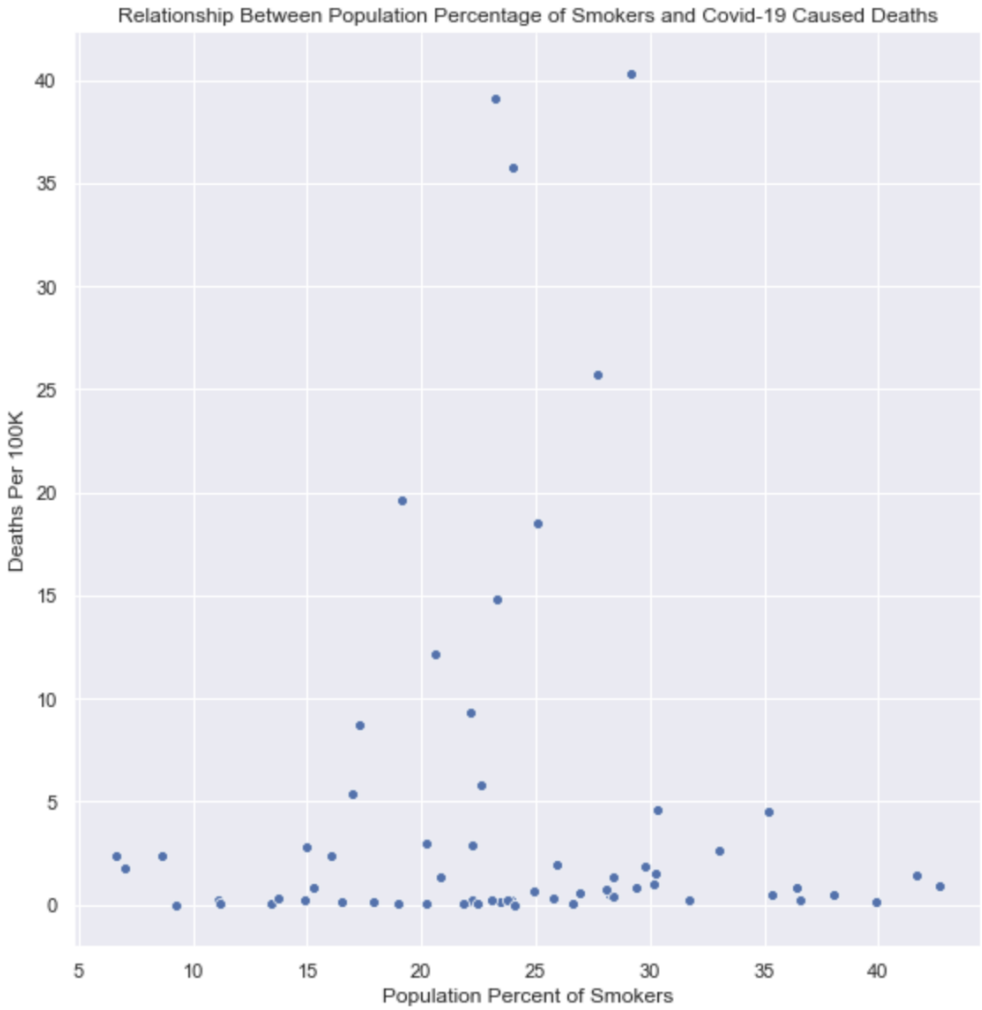
b) Secondly, I analyzed the correlation between the proportion of people that smoke and covid-19 caused deaths by country.

Note: the analysis is performed with a code named 'deaths_smoking_obesity.ipynb'.

My hypothesis was that countries with higher smoking rates have more covid-19 caused deaths, relative to their population size.

I started by reading in necessary data frames (covid-19.csv, population.csv, smoking.csv, testing.csv). Then I renamed each data frame's column headers to follow a consistent style and renamed columns that contained country names to 'Country' to be able to join the data frames. Then I joined the data frames using inner joins. Besides that, I generated tests and deaths for 100k people columns to eliminate the population size difference bias.

After the data was ready for analysis, I created a scatter plot visualizing the relationship between the population percentage of smokers and covid-19 caused deaths by country.



The graph does not show a significant relationship between the population percentage of smokers and covid-19 caused deaths. We see that some of the countries that have smoking rates between 15 - 20 percent have a relatively high amount of deaths, however, there are many countries with similar smoking rates but a smaller amount of deaths.

Based on the graph, my hypothesis is not true. However, if it is still true that smokers are at more risk, then my analysis could have been wronged by the testing performance data.

	Country	PopulationSize	SmokingRate	TestsPer100K
25	Greece	10746740.0	42.65	265.978334
51	Serbia	7057412.0	41.65	372.346123
29	Indonesia	261115456.0	39.90	12.638471
15	Chile	17909754.0	38.00	490.202155
36	Latvia	1960424.0	36.60	1524.976230
18	Croatia	4170600.0	36.45	295.449096
13	Bulgaria	7127822.0	35.30	223.055514
4	Austria	8747358.0	35.15	1861.316297
21	Estonia	1316481.0	33.05	2736.385865
60	Ukraine	45004645.0	31.70	95.152400

From the table, we can see that some of the countries that have high smoking rates have not tested many people. Therefore, for those countries, covid-19 caused deaths data might also be significantly more inaccurate than other countries'.

Overall, based on the analysis, we cannot say that countries with greater smoking rates have more covid-19 deaths, relative to their population size.

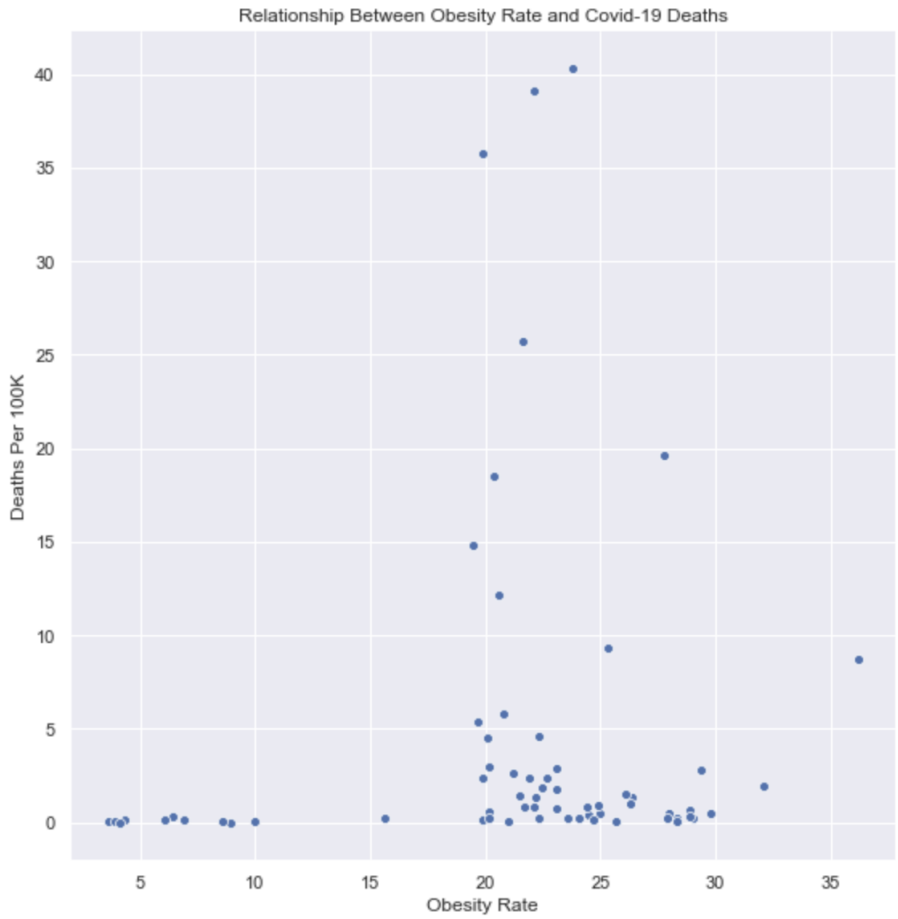
c) Secondly, I analyzed the correlation between obesity rate and covid-19 caused deaths by country.

Note: the analysis is performed with a code named 'deaths_smoking_obesity.ipynb'.

My hypothesis was that countries with higher obesity rates have more from covid-19 caused deaths because obese people tend to be at higher risk, according to scientists.

I started the analysis by reading in the obesity.csv file and filtered out the most recent obesity data. Then I renamed the columns and joined it to the dataframe I used for analyzing the smoking and death relationship. After that, I renamed the merged dataframe and dropped all the unnecessary columns ('SmokingRate', 'Temperature').

I did my analysis process by creating a scatterplot visualizing the relationship between obesity rate adn covid-19 caused deaths by country.



From the graph, we can see that there is no correlation between obesity rates and covid-19 caused deaths. Again, this could be due to the diverse testing capability and healthcare performance by countries. Countries with lower testing capability don't find infected and dead people which biases the data. Similarly, countries with lower healthcare capabilities cannot help all of the covid infected people and therefore they don't get an accurate overview of covid-19 caused damage to their population.

Overall, based on the analysis, we cannot say that countries with higher obesity rates have more covid-19 deaths relative to their population size.