Task 1:

For this task, I filtered the data based on a minimum size of 3000 because it provided a larger sample size (21 countries), while ignoring some of the outliers from considering the entire range. Below is the linear regression:

A close up of a map

Description automatically generated

The corresponding statistics are:

p-values: 0.1635283921118379

R^2: 0.07615340256810098

Slope: 0.00150537969525609

Because the R^2 value is incredibly small and the p-value is much greater than 0.05, we cannot conclusively determine anything. In other words, the correlation between median age of a country and its death rate cannot conclusively show that older individuals are at a higher risk of death due to COVID-19. The inconclusive nature of our investigation makes sense, as just looking at the data without the regression, we can see that it is pretty spread out in general, so we wouldn’t expect a high R^2 value anyways for a linear regression. Furthermore, the median ages considered are well below the age of 80 and does not take into account comorbidities, so it would make sense that median age is not a meaningful predictor.

Task 2:

Question 1: How does population density relate to the fatality of COVID-19?

Data Source: I found population density by country data from <https://datahub.io/world-bank/en.pop.dnst/r/data.csv> and used the JHU COVID-19 data. I also used data from <https://datahub.io/JohnSnowLabs/country-and-continent-codes-list> to relate country codes to their continent!

Methods: First I cleaned up the data so that I only had relevant information: Country name, death rate which we had calculated before, density values, and continent. When I did a quick plot of the whole world, I ended up getting:

A screenshot of a cell phone

Description automatically generated

As a result, I decided to narrow my scope to just Europe, and setting the maximum density value to 600 (which excluded the outlier Monaco with a population density of 20000).

Results:

A close up of a map

Description automatically generated

The corresponding statistics are:

p-values: 1.1528086620238059e-05

R^2: 0.3779848989954087

Slope: 0.00014865099786682878

Discussion: Since the p-value is very small (p < 0.05) and the R^2 value is nominally better than what we discovered in the first task, we can say that there is, to some extent, a relationship between population density and death rate. While this intuitively makes sense given what we know about COVID-19, there are many confounding variables we are overlooking in this model. For one, despite China having a high population density and even at times high infection rates, their death rate was significantly lower than Italy’s. This suggests that it might be less about the density of a country, but rather how draconian their measures are in combatting the spread of COVID-19 that ultimately lead to lower death-rates.

Question 2: In a [NYTimes article](https://www.nytimes.com/2020/04/07/climate/air-pollution-coronavirus-covid.html?fbclid=IwAR35f8BaZg2Nt0IJRYwHnDZ8YxOIOn5zg9vtWb10Pkt5MX-bgacNN9djhPo), they stated that pollution levels linked to higher COVID-19 death rates. I wanted to verify this

Data Sources: I used NYTimes data (<https://github.com/nytimes/covid-19-data>), which broke down cumulative cases and deaths by US county by day, and air pollution data (PM2.5 Wtd AM) by US county from the EPA (<https://www.epa.gov/air-trends/air-quality-cities-and-counties>).

Methods: First I cleaned up the data so that I only had relevant information: total deaths by April 8, the county, and that county’s PM10 level (the number of inhalable particles, with diameters that are generally 10 micrometers and smaller per cubic meter). Since the data from the NYTimes was cumulative, I used deaths on April 8 as the determinant of the plot’s y-axis, ‘Deaths’, which I plotted against ‘Pollution levels’ per county. This yielded:

A screenshot of a cell phone

Description automatically generated

According to the EPA, any pollution level above 15 μg/m^3 is considered moderate air quality. Thus, I filtered the data according to exclude any PM2.5 data below 50. I then created a linear regression for the data. The results were terrible. I tried again with a different measure of air quality (ozone), and also got poor results, even when filtering out good air quality levels (above 0.065 ppm).

Results:

A screenshot of a cell phone

Description automatically generated

PM2.5 Statistics:

p-values: 0.5765677161245109

R^2: 0.0016795320902042794

Slope: -0.11708493056977806

A screenshot of a cell phone

Description automatically generated

O3 Statistics:

p-values: 0.02061945073694575

R^2: 0.03768433049801812

Slope: 1489.5330194897717

Discussion: In the first investigation using PM2.5 data, the p-value was large (p > 0.05), and the R^2 value was very small, indicating a poor linear fit and no conclusive correlation. In my second investigation, the p-values were very small (p < 0.05) as was the R^2 value, which means there is nothing conclusive we can draw. This makes sense because ozone levels fluctuate, increasing when high temperatures are accompanied by weak winds, so there might be some hidden biases. For one, we know that the outbreak began in Washington state. It turns out that Washington has the third best air quality index in the US, so the data is already skewed as result. Moreover, since the model just considers cumulative deaths, it doesn’t take into account when the outbreaks began in respective states and counties. Lastly, in the NYTimes article, the study looked at pollution exposure globally and its correlation to pulmonary-disease-related deaths, whereas this regression *only* looked at COVID-19 deaths. Therefore, it might be conclusive that PM2.5 exposure has drastic effects on lung health, which in turn could weaken the lungs of those who have contracted COVID-19, but death and pollutant exposure is not conclusively link.

Task 3:

I chose to take this class because I am interested in the intersection of mathematics, health, and computer science. It encompasses a large part of what I envision for my future career, and the ML coding skills (regression tests, integration tests, Markov Models, etc.) I hope to learn will translate well in my (remote?) internship this summer!

This assignment took me about 20 hours to complete. The first task took me about 3hrs as it took me a lot of time and energy chasing down Google wormholes trying to figure out what was going on in the code and how to actually use the tools offered. Task 2 took a long time as well, because it took me a while to decide *what* research questions I could look into using data I had access to, and how to format/clean that data in a way that was useful to me.