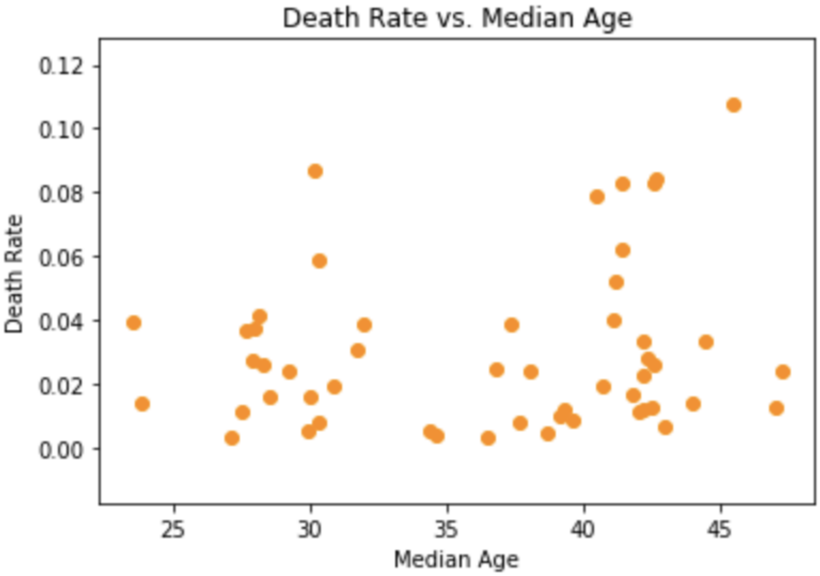
Math189Z: HW1 Deliverable

Rachael Soh

9 April 2020

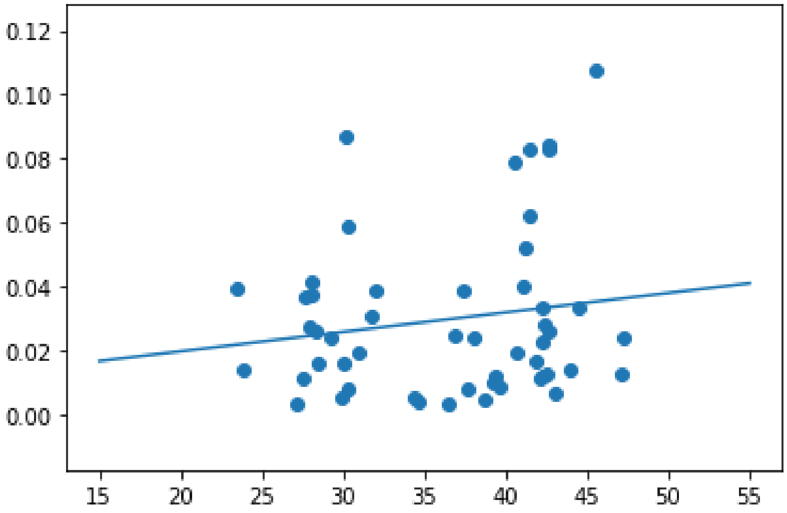
1 Task 1:

Filter on Sample Size and Re-Run Regression
Here, I filtered by the sample size or number of confirmed cases by only taking data with more than 1000 samples. I then repeated the steps used in the example where we calculated death rates of each country and removed any null values. We then plotted the death rate against the median age to get:



Once again, there is no obvious relationship between the 2 variables so we run a regression test on it to check if the results were statistically significant. Unfortunately, we find again that median age had no statistically significant effect on the death rate. Although we see a slightly more positive coefficient here than previously. I think the problem may be that the older infected population were more likely to die but the median age of everyone in the country is not quite representative of the median age of the infected population. To test this more accurately, maybe we could use the infected population's median age per country.

```
p-values: 0.2626107312969402
R^2: 0.02503189765307188
Slope: 0.0006084849030462419
```

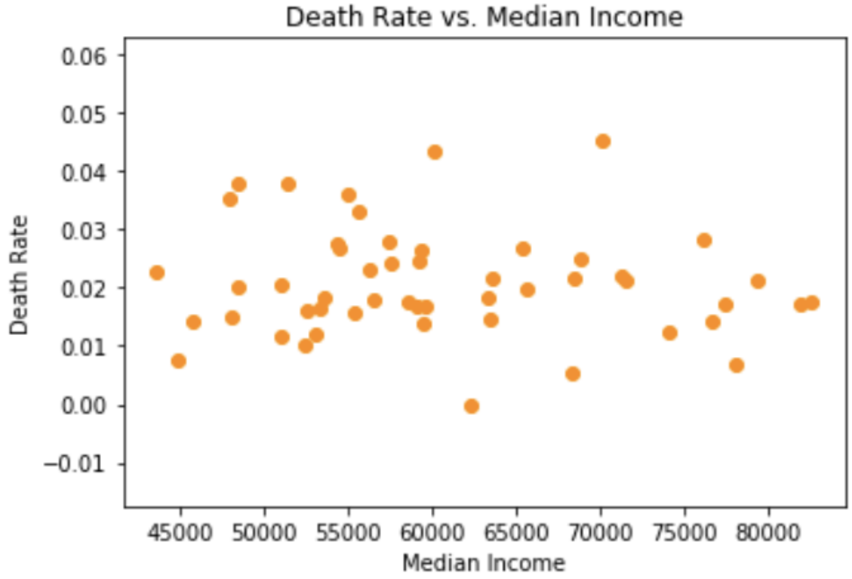


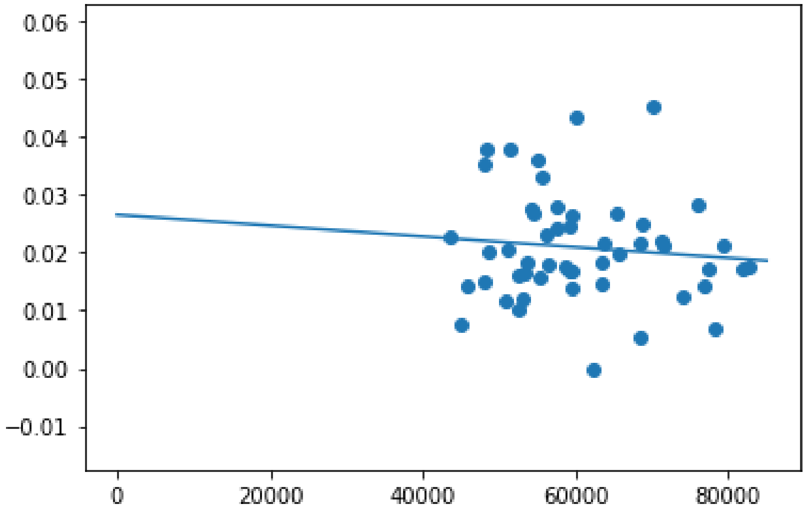
2 Task 2:

Find Your Own Data Here, I would like to further investigate causal relationships and correlations between the U.S. COVID-19 data and possible population statistics that could affect death rates. I decided to test death rate against median household income. I was wondering if those living in poorer areas had less access to healthcare, food or testing and maybe those in higher income areas could easily afford the healthcare if necessary. So let's explore this relationship. To find this information, I downloaded a dataset from here which gave the population and median household information of every state in the US. I also downloaded this data which gave the counts of confirmed cases and deaths in each state and county from here.

Now, we want to calculate the death rate at each state and merge the death rate with the median income data frame. We then remove any states lost in the merge, in this case the median income dataset did not include the U.S. territories and so we want to remove them.

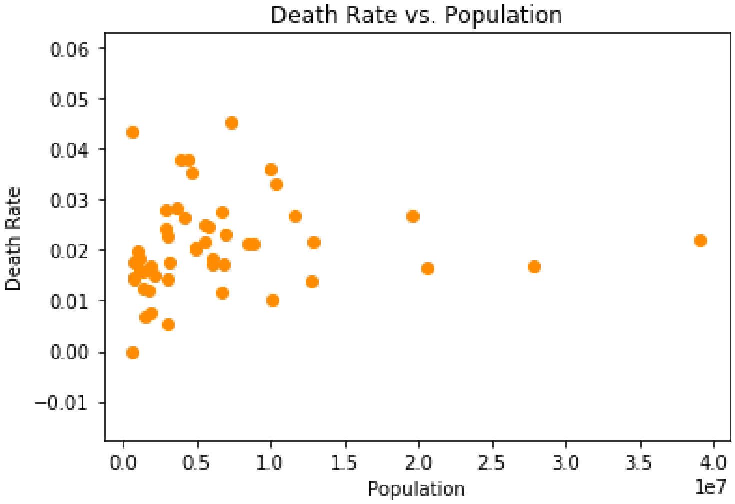
Now, we can plot our death rates against median income. It seems that there isn't an obvious relationship between the two variables and when we run a linear regression between the two, it becomes evident that there is no statistically significant relationship between the two variables. I think one reason it failed is because each state has such a large disparity and so I wanted to repeat this with the counties data, however, I could not find a dataset for median household income by county for 2020.

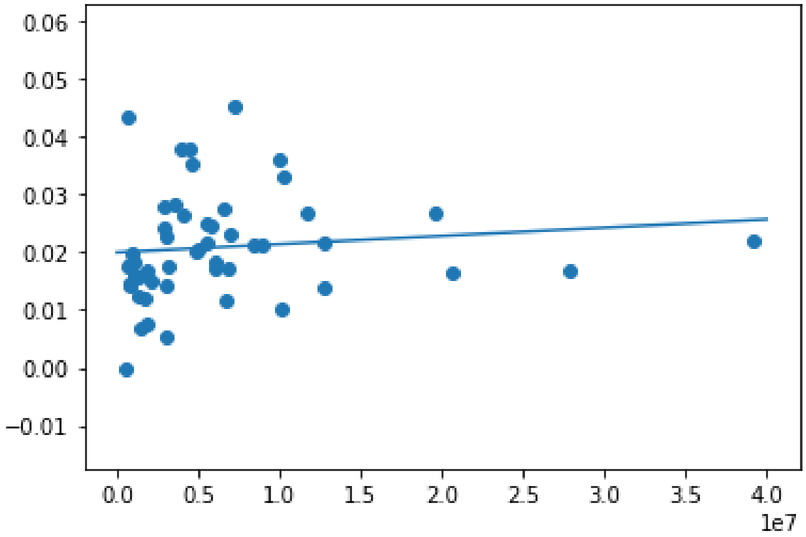




Now let's try finding a relationship between the population size and the death rates. I wanted to see if having a denser and more populated state would have an effect on death rates as a larger population would mean a higher chance of contact, making transmission easier. Having a larger population could also overflow healthcare services in the event of a pandemic. So we repeat the steps

we did previously.
Once again we see no statistically significant relationship between the two variables. As mentioned previously, I think that a big part of why we don't see a significant relationship may be because the state data is not informative enough and that using county data may give us more information. Ex. New York city is extremely dense and populated but New York also has states like Ithaca with a much smaller and less dense population. Another variable I would try to use in the future would also be population density. Ex. Singapore would have a small population but an extremely large population density considering its land size.





3 Task 3:

A few sentences describing why you decided to take this course/what you were hoping to get from it. I was hoping to learn a bit about data analysis. I've taken stats classes before so I feel that I have learned a some basic statistical concepts but I don't feel that I've done enough data wrangling and applications of these concepts. I'm also interested in healthcare and this pandemic situation we are in so I think this is a good opportunity for me to pursue my interests while applying what I've learned in the past.

How long it took you to complete this assignment It took me around 4 hours because this is my first time using Jupyter notebook and pandas so reading documentation took a bit of time. Although most of my time was trying to find a dataset I could use for my research question. Originally I wanted to explore the relationship between weather and deaths but I could not get good weather data and once I got it, I didn't know how to make it usable so I had to give up on it as it was taking too much time.