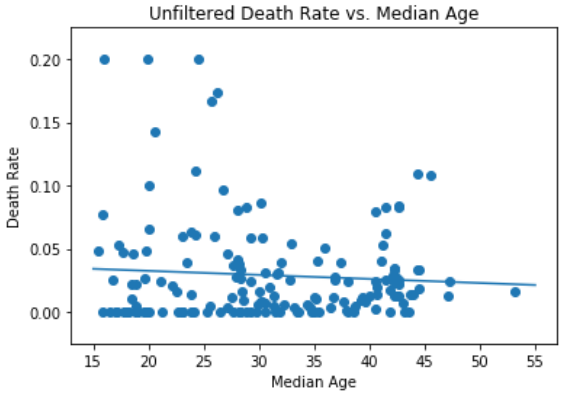
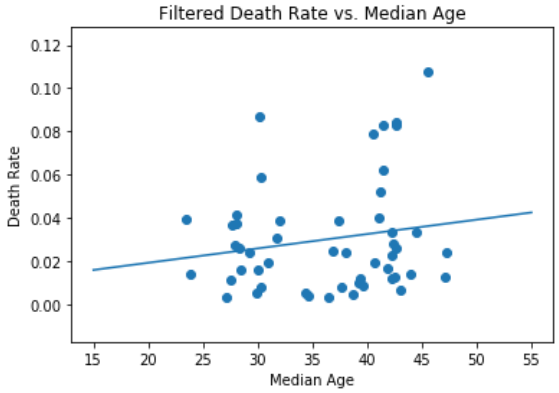
Victoria Lloyd
Homework 1 Deliverable
COVID-19: Data Analytics and Machine Learning
Due Thursday, April 9, 2020

For Task 1, I decided to filter out countries with fewer than 1000 cases as of April 2. After creating this filter I re-ran the linear regression, which yielded the following scatter plot:



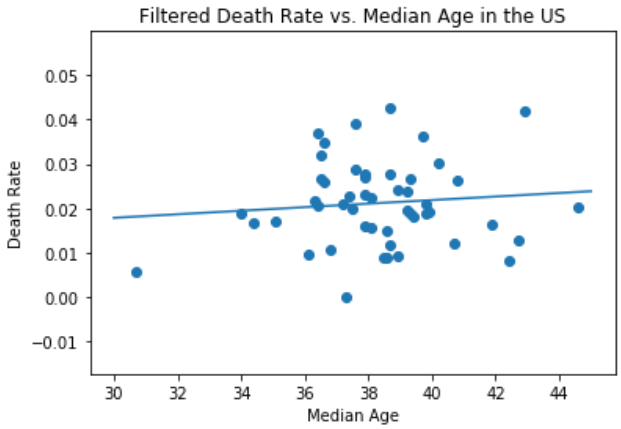


p-values: 0.35339370626947875
R^2: 0.005285847299216592
Slope: -0.00032013215700298314

p-values: 0.22587212168821744
R^2: 0.0297855879051554
Slope: 0.000665030949548829

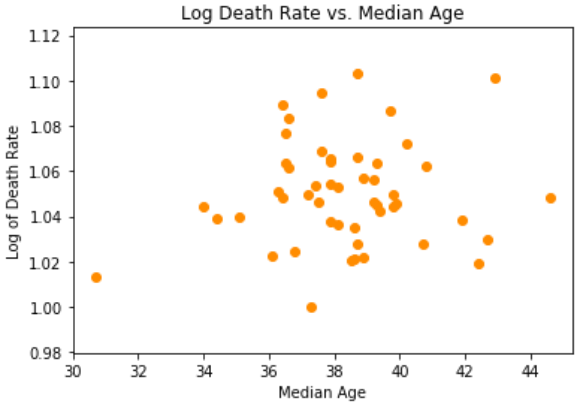
We see that this filter did not improve the variation by much, and even after I used more selective cutoff values the variation I was not able to obtain a more statistically significant linear regression. However, I did notice that by removing the countries with fewer than 1000 recent cases many of the countries with a very small median age were filtered out of the dataset. While this is most likely a result of countries with greater access to testing having the medical resources to sustain an older population, it is possible that this might indicate that older populations are contracting the virus in greater numbers. Additionally, while the linear regression fit does not have a large enough statistical significance to draw definite conclusions from, I noticed that the death rate seems to increase somewhat as median age increases. While this is in line with what I know of the virus, from this dataset alone I am not able to claim that there is a correlation between the two variables, only that the general trend suggests a possible connection.

For Task 2, I was interested in performing a similar analysis for states in the US using the States Current Values dataset from The COVID Tracking Project which lists the number of current positive and negative COVID test results as well as the number of deaths and test accuracy scores as of April 3, 2020 and the Median Age By State 2020 dataset from the World Population Review which lists the total median age and the median ages by male and female. After obtaining the data and cleaning it (changing the Median Age dataset state labels to match the abbreviated labels used in the COVID dataset and changing from a numerical index to a state index), I designed a filter imposing a cutoff point for the same reasons as in part I, which in this case was 600 cases by April 3. This yielded the following result:



p-values: 0.4852143566258089
R^2: 0.010202121464731838
Slope: 0.00039947002310030294

I saw again that the error of this fit is too high for these results to be statistically significant, but in this case the variation was lower than the distribution of countries, with the death rate ranging from 0.00 to 0.05 rather than 0.00 to 0.11. In this case, while our results are more statistically significant since our p-value increases between the two filtered models, the model is still not statistically significant enough to draw conclusions and the model for the US and our new model is able to explain less of our variation than in the worldwide dataset. Overall, I was not able to get more logical or statistically significant results for the statewide case compared to the worldwide case. Even after I tried taking a logarithmic plot of the death rate to see if there was some other regression type I could use, I found no distinctive pattern to follow.



I decided to take this course because I have always enjoyed working on research projects and contributing positively to my community. I have worked with machine learning and big data analytics before, but all of it has been self-taught over the course of my different research projects. I am excited to have the opportunity to more fully understand these topics in a low-stress structured environment and to make some positive contribution to all the chaos of the world right now using my research skill set. I have no real expectations from this course, other than having the opportunity to meaningfully interact with the body of COVID data out there and potentially contribute to some of the scientific efforts to understand it. This problem set (including setting up Jupyter notebook on my laptop) took about an hour and a half to complete.