Shanni Lam

Math 189Z HW1

9 April 2020

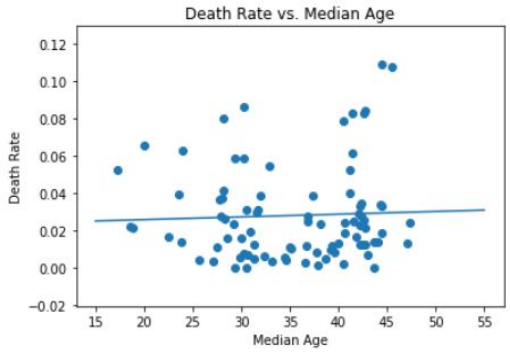
Task 1:

I cleaned the dataset so that we only consider countries with at least 230 confirmed cases; I chose the number 230 because it is near the median number of confirmed cases of 233 cases.

We see that there is lots of spread in the data. There are several points that show that older median ages tend to have higher death rates, which our hypothesis stated.

However, in the end, we may have to rethink our hypothesis or wait for a bigger sample size. We can think of p-values as the likelihood that the results of a study are due to random choice; here, the p-values are way above 0.05, meaning there is likely little relation between both variables. Additionally, the slope and the R^2 value, which is the square of the correlation, are minimal, implying very little of the data is explained by the relation between median age and death rate, which goes against our hypothesis. We also see that there is almost no correlation.

```
p-values: 0.6936657921384319
R^2: 0.0018128900106691415
Slope: 0.00014590492713919978
```



Correlation: `0.04257804611145445`

Task 2:

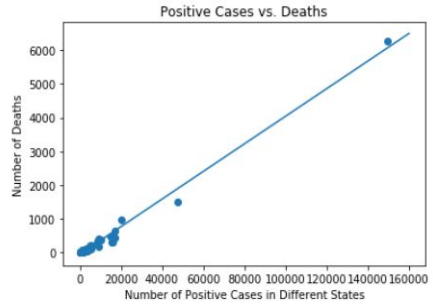
I will be using this dataset to observe the correlation between the number of positive COVID-19 cases and deaths for each state in the United States: https://docs.google.com/spreadsheets/u/2/d/e/2PACX-1vRwAqp96T9sYYq2-i7Tj0pvTf6XVHjDSMIKBdZHXiCGGdNC0ypEU9NbngS8mxea55JuCFuua1MUeOj5/pubhtml#.

Initially, I thought that there may not be enough cases in each state to warrant a strong correlation, but of those with a sizeable number, it may be interesting to notice a pattern among states that have high or low death rates. Regardless, in the results below, we still found a very strong correlation. We can conjecture the varying death rates may be due to access to healthcare, unemployment checks, political leaning, etc.

I converted this online spreadsheet to a CSV, took out empty rows, all columns except "Positive" and "Deaths", and replaced empty cells with the number 0.

According to the statistics below, there is almost a correlation of 1, an R^2 which states that about 98.5% of the variability in the data can be explained, and the p-values are almost 0, meaning it is extremely likely we got these results by random chance. The slope is small, but this may just mean that there are very few deaths per positive case; this makes sense since from the news, we know that the death rate is low, about 2% for young people but higher for old people.

```
p-values: 7.364424046176145e-48
R^2: 0.9857912879280191
Slope: 0.04093971543099365
```



correlation: 0.9928702271334453

Task 3:

This homework took me about 4 hours. I am taking this class because I want to learn more big data/data analytics/machine learning techniques, get research experience, fight the coronavirus, and be prepared for future internships and jobs.