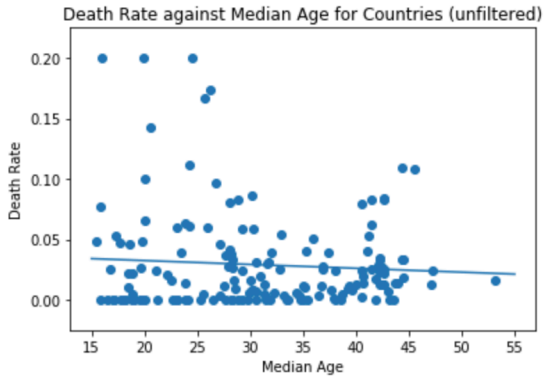
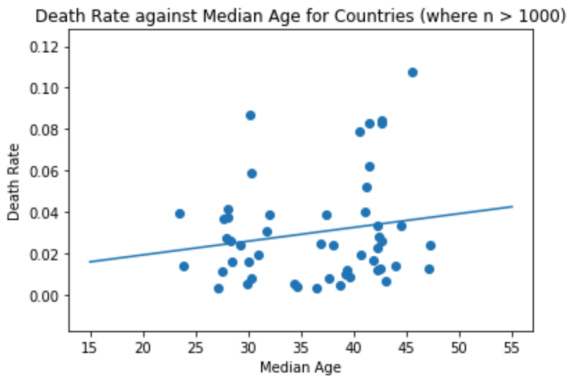
Daphne Poon

MATH198: PSET 1

Task 1:





For the unfiltered data,

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

For the filtered data,

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

If we compare the two sets of data, we see that the R^2 value for the filtered data is about 6 times larger (though still very far from perfect). The slope is also now slightly positive, as opposed to slightly negative. The p-values are also smaller for the filtered data, though not small enough to confidently say there is a relationship between the median age and death rate of a country. It also assumes that the COVID-19 cases are evenly distributed around the median age.

There is a lot of noise in the data since each country has different standards for reporting, testing and healthcare. A better way to determine if there's a relationship between age and death rate is by comparing the death rates of each age group in a country.

Task 2:

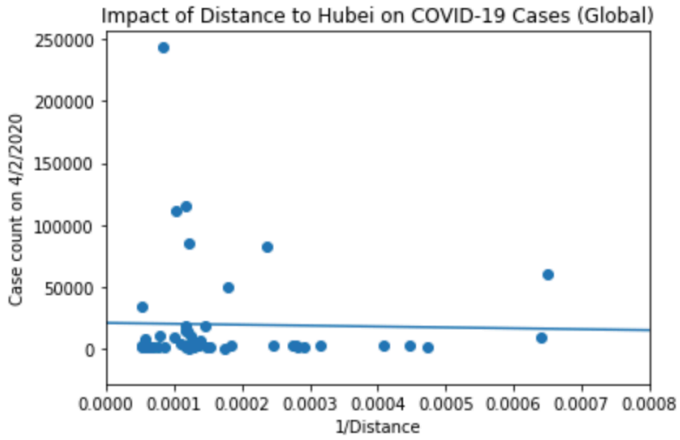
Research question: Is there a negative relationship between a country's (alternatively: province's) distance from Hubei (the epicenter of the epidemic) and the number of coronavirus cases it has?

I used the same Johns Hopkins data that was provided in the starter code. I wanted to explore the relationship between distance and case count, but I also guessed in advanced that there would be no relation, considering how travel between China and the rest of the world is not really limited by distance, and more so by which airports are located where / how popular those destinations are with tourists.

Daphne Poon

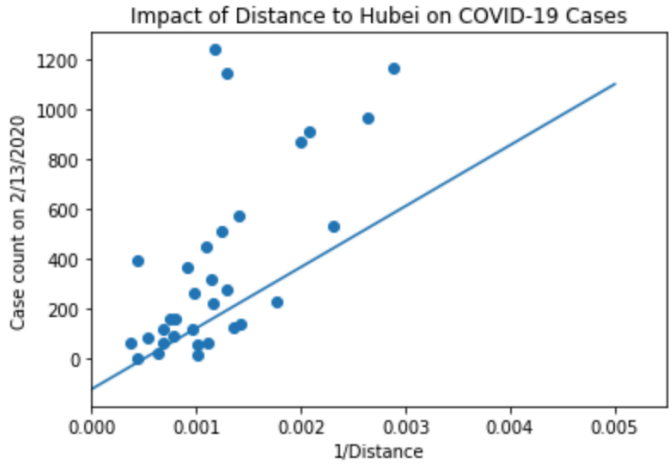
I combined the data for separate counties/provinces and filtered the countries such that only countries with over 1000 cases on 4/2/2020 were included. I used the original latitude / longitude coordinates that was dropped in the confirmed table from Task 1, and calculated the inverse of a country's distance (in km^{-1}) from Hubei as a new column. Plotting that information, I got the graph below. There isn't a real correlation between the two, and that's something we can see from the p-value and the R^2 value.

```
p-values: 0.7808837937434487
R^2: 0.0015618719639924515
Slope: -12103549.47331528
```



Seeing this, I wondered if this pattern (or lack of) would be the case on a localized level, specifically whether provinces in China would be affected by their distance to Hubei. Since China has effectively plateaued in terms of COVID-19 case growth, I wanted to compare data points from an earlier point in time. I decided to use data from 2/13/2020, around 3 weeks after Hubei was shut down and around 3 weeks after Chinese New Year (when many people were travelling). My hypothesis for this was since moving within China for work was a pretty common thing, many people from Hubei would probably be visiting family in nearby provinces, increasing the relative number of cases in those provinces. The graph shows the results I got from this plot:

```
p-values: 1.830234624009114e-05
R^2: 0.4629593916295409
Slope: 408547.62692306656
```



Daphne Poon

For this graph, $R^2 = 0.46$ and even though that's not that close to 1, it suggests there may be a correlation between the distance from Hubei and the total number of COVID-19 cases detected in the provinces of China.

I can't confidently say there is a relationship between the variables, mainly because there are many other factors that we haven't considered; certain parts of China are much more densely populated than others, and a few of provinces that were included by Johns Hopkins (e.g. Mongolia / Tibet / Xinjiang) are very sparsely populated and very far from the rest of China, meaning they have been relatively unaffected by the virus. There's definitely more to do because a conclusion can be made.

Task 3:
I'm hoping that this course will teach me the fundamentals of machine learning and data analysis, so that I'll have the basic knowledge necessary to talk about it in the future with other engineers. I think knowing how to analyze data is also just a good skill to have. I also think it'd be cool (wrong word, maybe) to be using data that is so relevant to our lives right now.

This assignment took me about 4-5 hours.