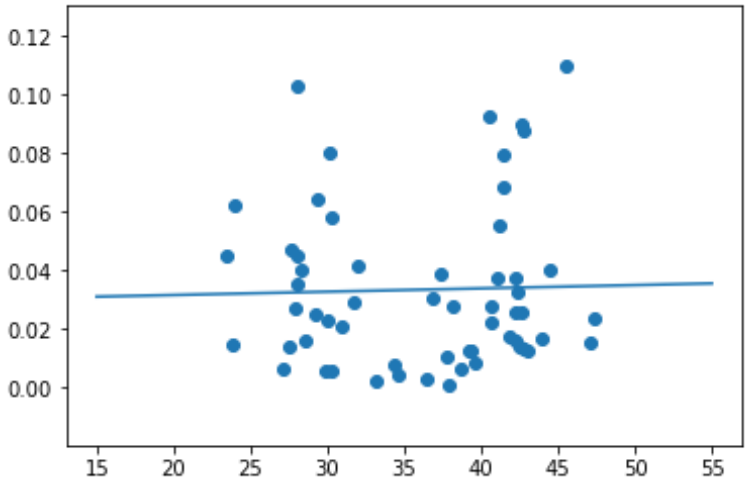
189Z Homework 1: Ethan Greenberg

Task 1: New Linear Regression

Our first task was to filter data and remove entries with insufficient samples in an attempt to improve our model for death rate as a function of median age. I only included countries with over 1000 confirmed tests as of April 4 in my model. You can see the results below:

```
p-values: 0.8378739413309912
R^2: 0.0007540407730892922
Slope: 0.00011301076019544737
Y-label: Death Rate
X-label: Median Age
```

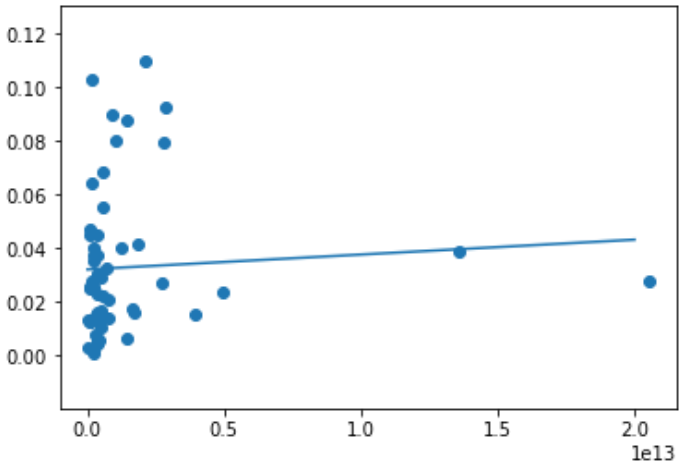


As you can see, the results of this regression are not very good. The R^2 value is incredibly low (0.000754) which suggests the obvious: that our regression does not fit the data well. This suggests that median age is not a good indicator of death rate, and/or there is not enough data to show the relationship.

Task 2: Research Questions

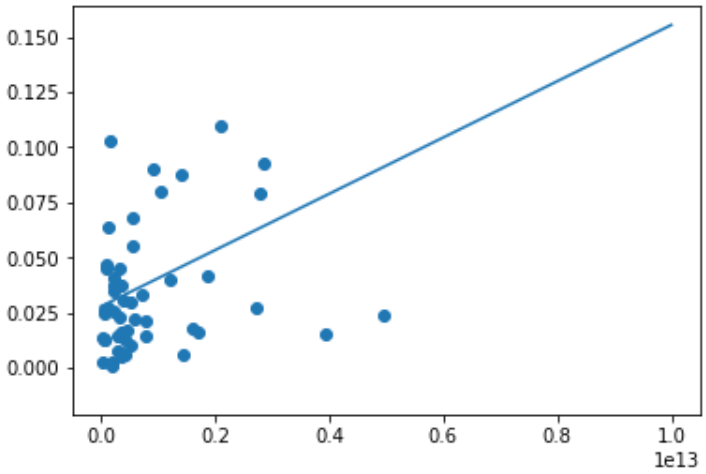
For my own exploration, I was interested in seeing whether there was a relationship between death rate and indicators of wealth. For my first regression, I pulled a dataset from the World Bank (https://data.worldbank.org/indicator/NY.GDP.MKTP.CD) with the GDP of countries in USD from 2018. I imported this into my workspace, and merged it with my filtered death rate dataframe, excluding countries as appropriate. I then ran a linear regression on the data. Here are the results:

```
p-values: 0.6296230337649651
R^2: 0.0044199353012690425
Slope: 5.543194760730156e-16
Y-label: Death Rate
X-label: GDP
```



As you can see, again, the regression does not fit the data well. I decided to also run a regression on the dataset after removing countries with GDP's above .6e13:

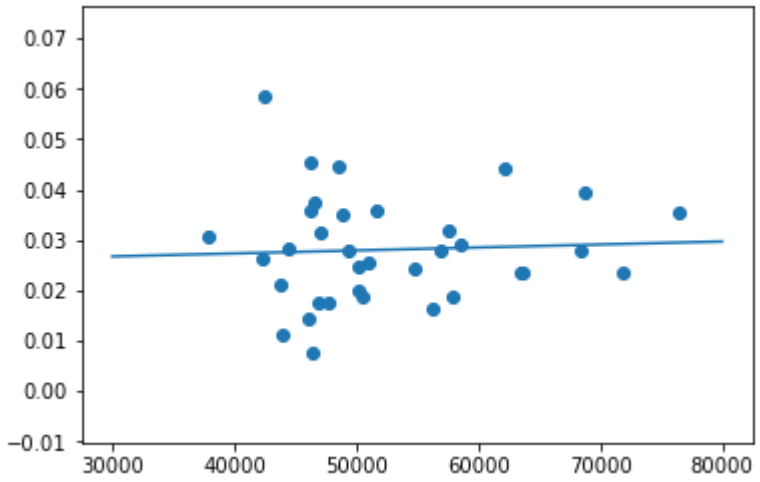
```
p-values: 0.09040250839944104
R^2: 0.0551892407197655
Slope: 6.39432066060142e-15
Y-label: Death Rate
X-label: GDP
```



Again, the fit is not very good. This suggests that GDP is not a good indicator for death rate in a country, or that there is not enough data to suggest a connection.

For my second research question, I explored the stat-specific Covid19 dataset. I wanted to see if average income for adults in a state was an indicator for death rate. I used 2018 data from the U.S. Bureau of Economic Analysis (BEA) for average state income and imported it into my workspace. Next, I joined this dataset with the state date for Covid19 using an intermediary array to translate between state abbreviations and full names. I then computed a linear regression from the dataset:

```
p-values: 0.7701995119522468
R^2: 0.002621793165324843
Slope: 5.927995561801271e-08
Y-label: Death Rate
X-label: Average Income
```



Yet again, we see that the linear regression does not seem to fit the data well. This suggests that average income is not a good indicator for death rate or at least there is not enough data to show a correlation.

Task 3: Me

I took this class because, first, it seemed like an interesting and motivating topic overall and online classes have generally been anything but motivating for me. Data analytics, ML, etc. are interesting topics to me, and framed by Covid19, I think this is a really great opportunity to learn a bit more about all of that stuff. Finally, I am really excited by the project component; I am excited to explore something about Covid19 that could be helpful to people while also improving my skills/learning.

This assignment took me 4 hours.