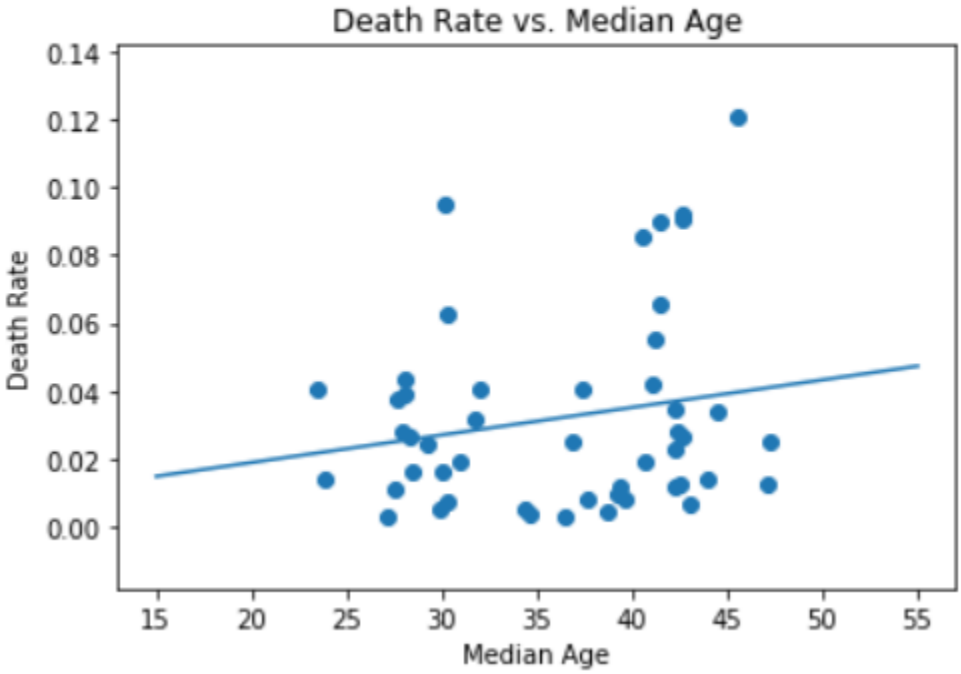
Homework 1: Analyzing COVID-19 Data with Regression

Task 1: Filter on Sample Size and Re-Run Regression

Graph and Important Statistics:

```
p-values: 0.19129095079483016
R^2: 0.036054248451831604
Slope: 0.0008102589143363457
```



Results, Interpretation, and Discussion:

We have a p-value of 0.1913, an R^2 value of 0.0361, and a slope of 0.0008. This shows that there is a more positive correlation between age and death rate because of the positive slope (albeit a very small correlation). However, we cannot reject the null hypothesis (our p-value is so high), the data may be statistically insignificant. It should be noted that our R^2 value is very low, which suggests that this data is not fitted properly. Perhaps a sample size of larger than 1000 cases would improve the regression model and provide a more conclusive result or interpretation. It does make sense to me that the data would be better (in terms of fit) with the additional constraint of a sample size.

Task 2: Find Your Own Data

Research Question

How do death toll and the number of confirmed cases correlate with total hospitalizations across the US? If states have a high number of hospitalizations, then their death toll and the number of confirmed cases will be higher. The null hypothesis would be there is no correlation (or statistical significance) between death toll and the number of confirmed cases with total hospitalizations in the US.

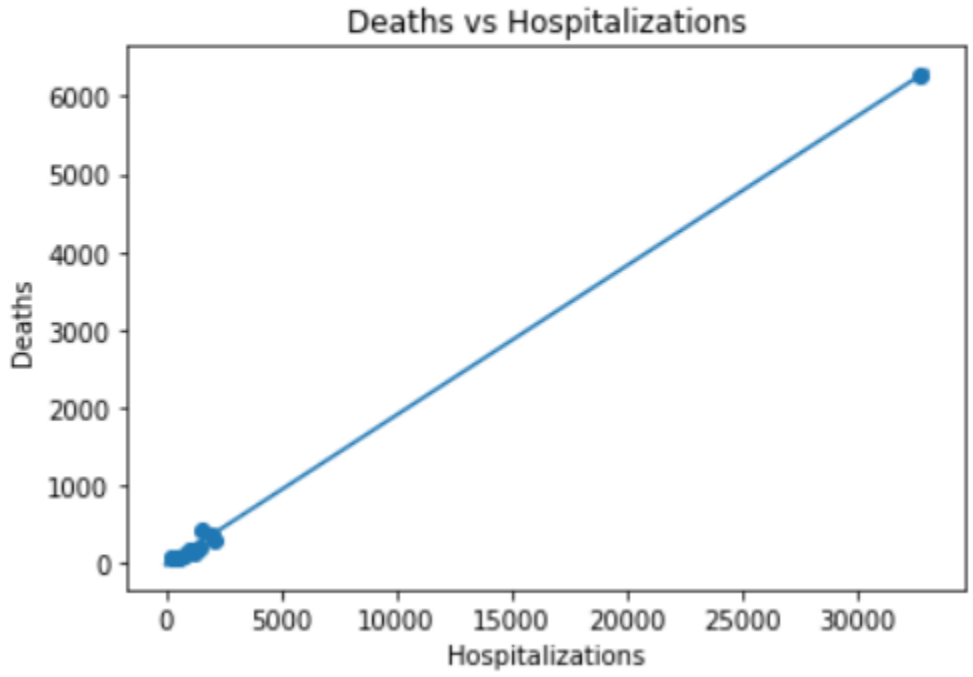
Data Source and Description

The data came from The COVID Tracking Project's GitHub repository. This data shows a variety of statistics across states in the US, notably positive cases and deaths resulting from COVID19.

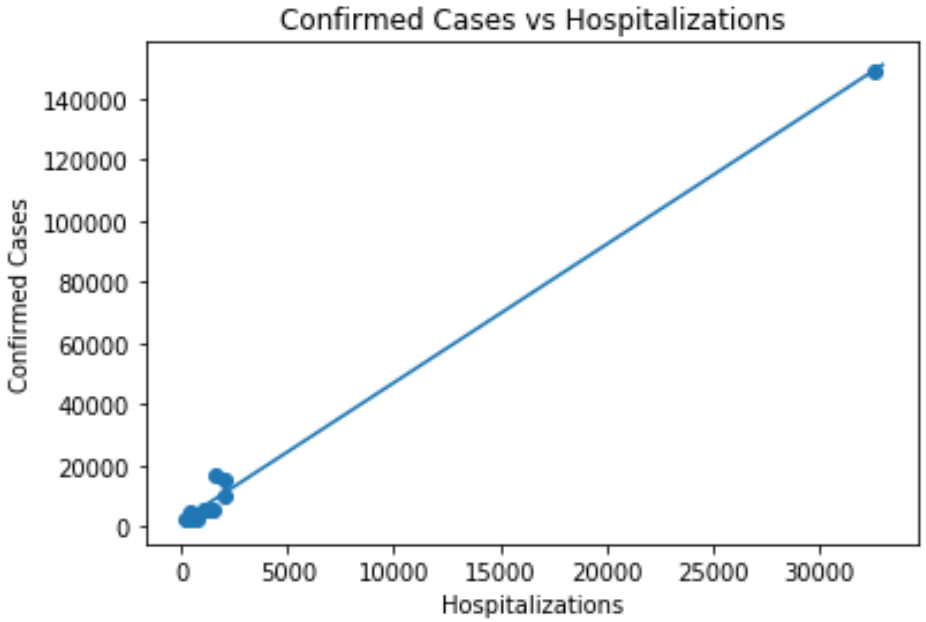
Graph, Important Statistic, and Expectation:

I expect there to be a positive relationship between deaths and hospitalizations across states in the US. Additionally, I expect for there to be a positive relationship between confirmed cases and hospitalizations across the US. I expect both graphs, and the corresponding statistics, to show us that the data is statistically significant.

```
p-values: 2.7223842959063155e-17
R^2: 0.9987340294380046
Slope: 0.19256218057629665
```



```
p-values: 1.0421694881315281e-13
R^2: 0.9943281008640964
Slope: 4.527164661196226
```



Methods:

I followed a similar process to the tutorial in terms of filtering out data. I decided to explore Deaths vs. Hospitalizations and Confirmed Cases vs. Hospitalizations for states within in the US. In previous task we added a minimum sample size for the data; similarly, I made sure to add a minimum sample size to the new data. My choice to increase it to 2000 for this task was because of how increasing sample sized improved the fit on the data in task 1. I also graphed deaths instead of death rate, as otherwise the graph would be even more scrunched into the bottom left corner. I contemplated removing the point that is very far away from the graph, but since I did not do any calculations for outliers and the graph seems to have a good fit, I decided to leave it. I left the potential outlier, for the same reason, in the Confirmed Cases vs. Hospitalization graph.

Results:

From this exploration we can see that there is a more positive trend for death toll and hospitalizations compared to the variables we observed in task one. There is an even more positive trend for confirmed cases and hospitalizations. Additionally, the regression fit is excellent – for both graphs – as we have an R^2 value of almost one. Also, we can reject the null hypothesis due to the size of the p-value (approximately 2.72e-17 for graph one and approximately 1.04e-13 for graph two). Therefore, it seems like the data for total deaths versus hospitalizations is statistically significant. Additionally, the data for confirmed cases versus hospitalizations is statistically significant, which confirms our initial hypothesis. However, I would continue to follow up on this as the situation progress, if I were to continue to explore this

question. I would be curious to see if these results hold, and for how long they do hold during this time.

Credibility:

From the standpoint of where the data comes from, the source seems to be credible (a lot of their data matches up with CDC numbers). However, I am not sure about the collection of the data. There may be biases that deserve some contemplation in terms of how these statistics are recorded (i.e. are these deaths recorded from COVID19 being the main cause? Could there be other deaths not attributed to COVID19, but COVID19 was the cause for hospitalization? Could there be hospitalizations caused by something else, but death was resulted from COVID19?). However, for an exploratory search I believe this data should suffice to gain some insight into what kind of impact COVID19 is currently having. As for the results themselves, they do seem to be pretty credible with how low the p-value was, the R^2 value being so close to one, and the positive relationship between the variables (for both graphs). However, this was from only one pool of data, and it was just within the US. To add more credibility to what the results suggest, we would need a larger sample size and a more random sample.

Task 3: Reflection

I decided to take this course to learn more about big data, and how to use statistical tools to understand and use big data. Furthermore, it seems like this course could offer me a taste of machine learning before I take the class at Mudd, which I think would be quite beneficial. Additionally, I think the topic of the class is relevant and interests me, so I am quite curious and excited to be working with the data.

This assignment took me approximately 8 hours.