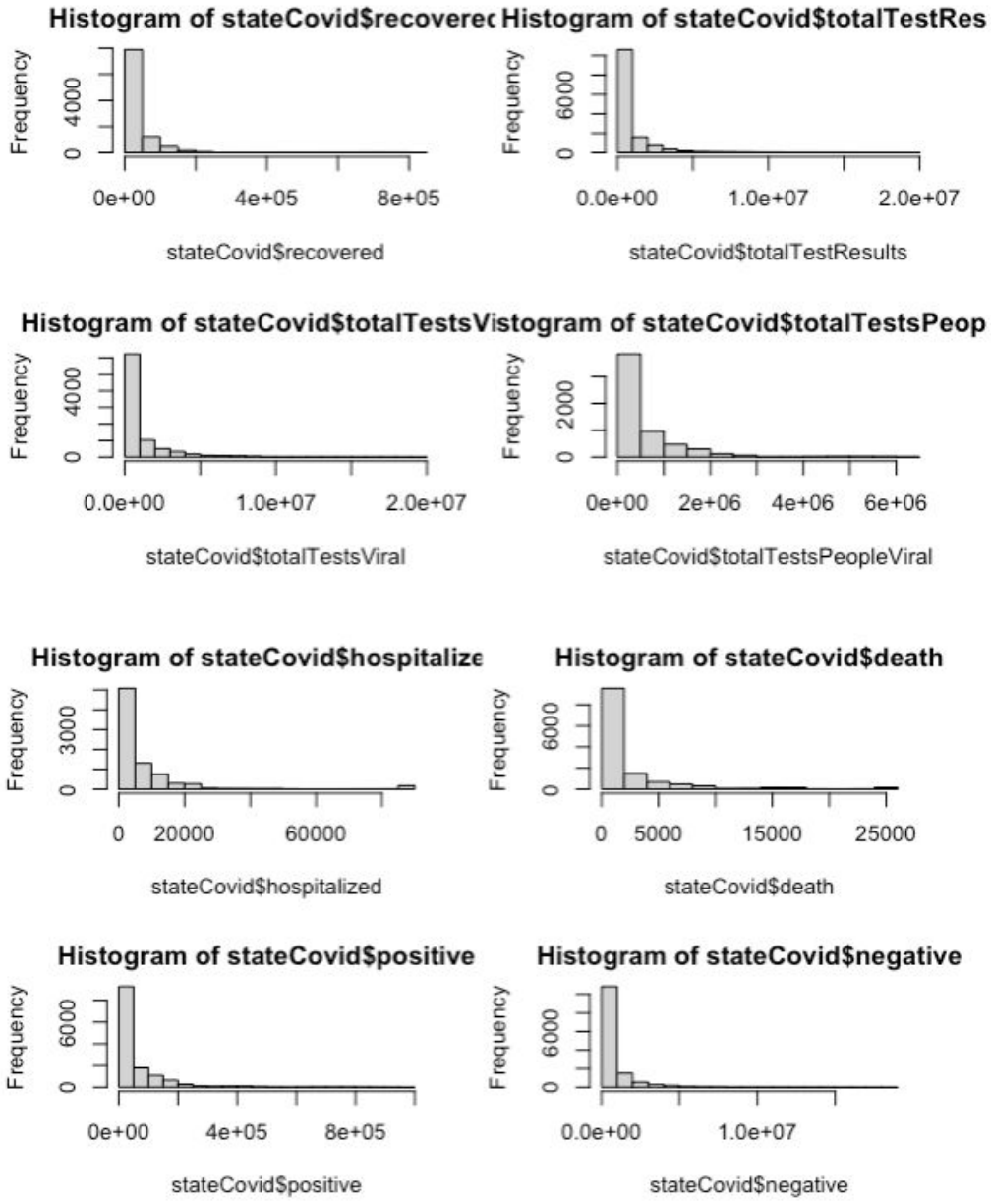
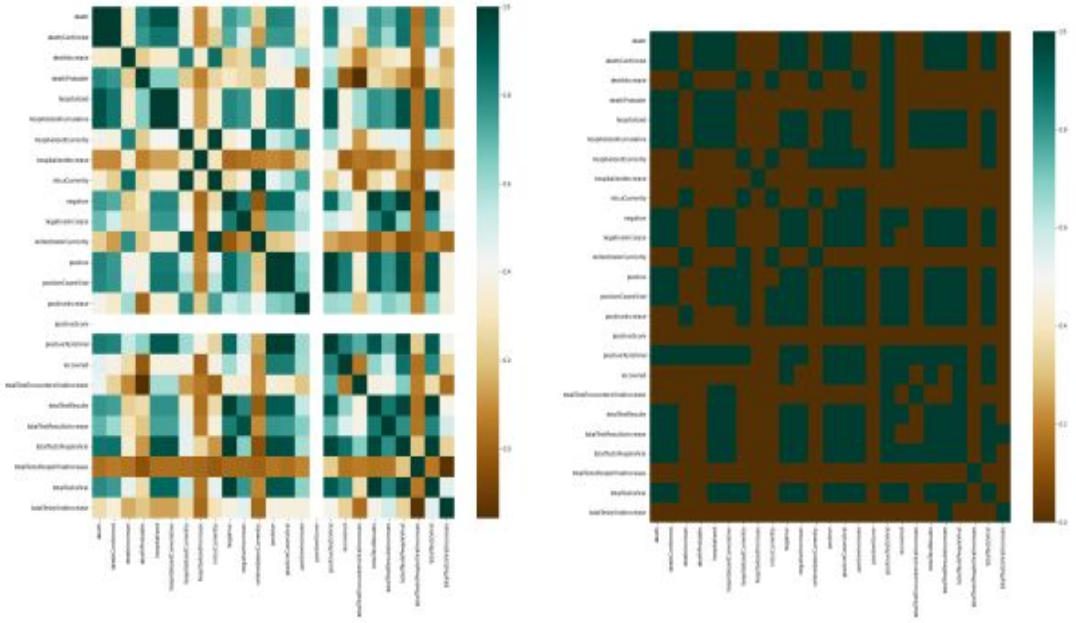
# Preliminary Data

Miles Tweed, Steven Spielman, Julian Palazzo

1) Preliminary descriptive analysis on the data:



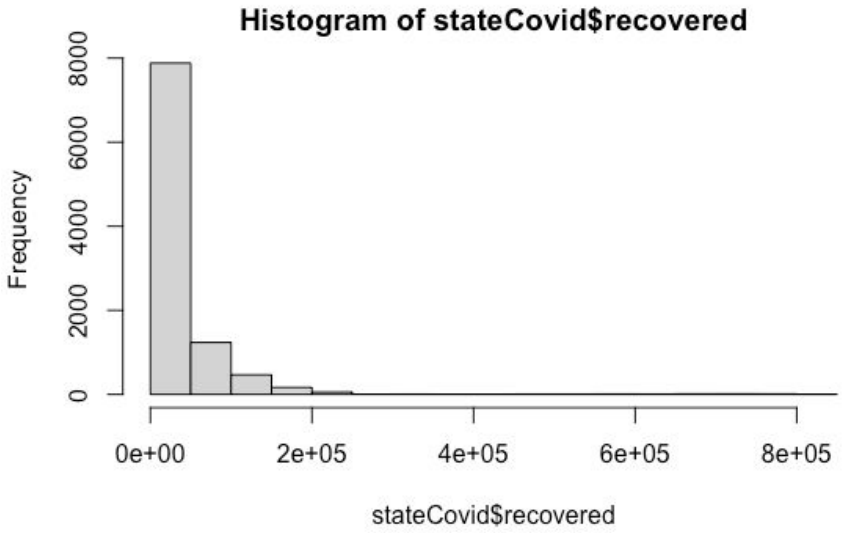
To look at the variable relations to one another and gauge their strength, we created this visual. It looks at correlation between variables with teal representing a strong relationship and brown representing a weak relationship.



We decided to look at a correlation of 0.5 as a hard limit to find relationships between variables, which can be seen in the changes above.

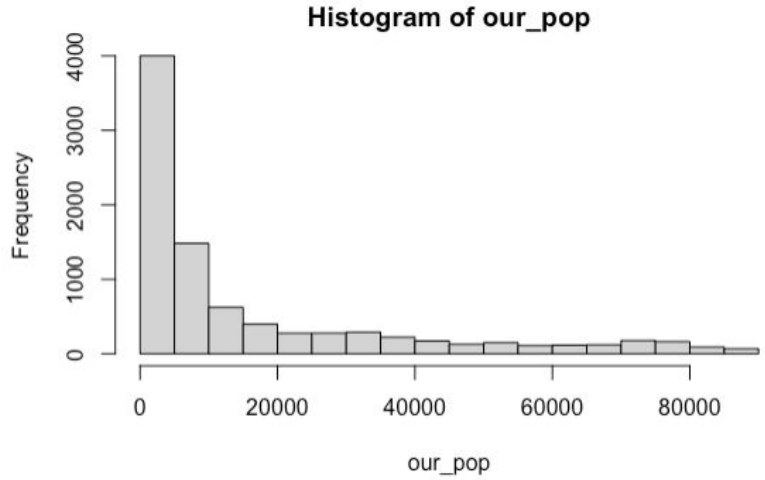
2) We decided to look at the recovered variable as our response. We conducted a single sample statistical inference testing on the response. We did some exploration into the summary statistics and performed a t-test with a selected mean value to test as well as computed the effect size.

```
[1] 32625.7
[1] 7863.5
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      2    1725    7864   32626   36379  811330    4089

        One Sample t-test

data:  stateCovid$recovered
t = 3.712, df = 9909, p-value = 0.0002067
alternative hypothesis: true mean is not equal to 30000
95 percent confidence interval:
 31239.16 34012.25
sample estimates:
mean of x
  32625.7

[1] 0.03728854
```




After looking at the distribution and summary of the recovered data, the distribution is not normal and contains extreme outliers. To see if there was a change in the distribution or overall data summary without the outliers, we performed another test.

```
[1] 16423.99
[1] 6116.5
[1] 1024
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      2    1432    6116   16424   22619   88780    4089

        One Sample t-test

data:  our_pop
t = 6.057, df = 8885, p-value = 1.443e-09
alternative hypothesis: true mean is not equal to 15000
95 percent confidence interval:
 15963.15 16884.84
sample estimates:
mean of x
 16423.99

[1] 0.06425492
```




The distribution is still highly skewed, making t-test results still suspect. We have decided to use non parametric hypothesis testing via bootstrap. Pending results.

3) We ran a t-test to examine whether or not the number of daily COVID-related deaths in florida were significantly different as compared to the average nationwide COVID-related death rate. We multiplied the nationwide death rate by the population of Florida to get the expected daily death rate for Florida and used this value as our population average. We ran the one-sample two-tailed t-test on the observed daily increase in COVID-related deaths compared to the expected daily death rate based on the national average to produce the following result:

```
        One Sample t-test

data:  florida$deathIncrease
t = -110811, df = 283, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 409223.9
95 percent confidence interval:
 53.68608 68.22238
sample estimates:
mean of x
 60.95423
```

We also ran a one-sample t-test to compare the daily increase in COVID-related death counts in Florida to the national daily increase in COVID-related deaths.

```
        One Sample t-test

data:  florida$deathIncrease
t = 12.073, df = 283, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 16.37531
95 percent confidence interval:
 53.68608 68.22238
sample estimates:
mean of x
 60.95423
```

4) In order to compare regional COVID statistics, we subsetted our original COVID dataset into four regions: northeast, midwest, south, and west; according to their respective designations by the US Census Bureau. We then performed a Welch's two-sided t-test to compare the average daily COVID-related death counts between regions. First, we compared the average daily death counts between southern and western states to generate the following result:

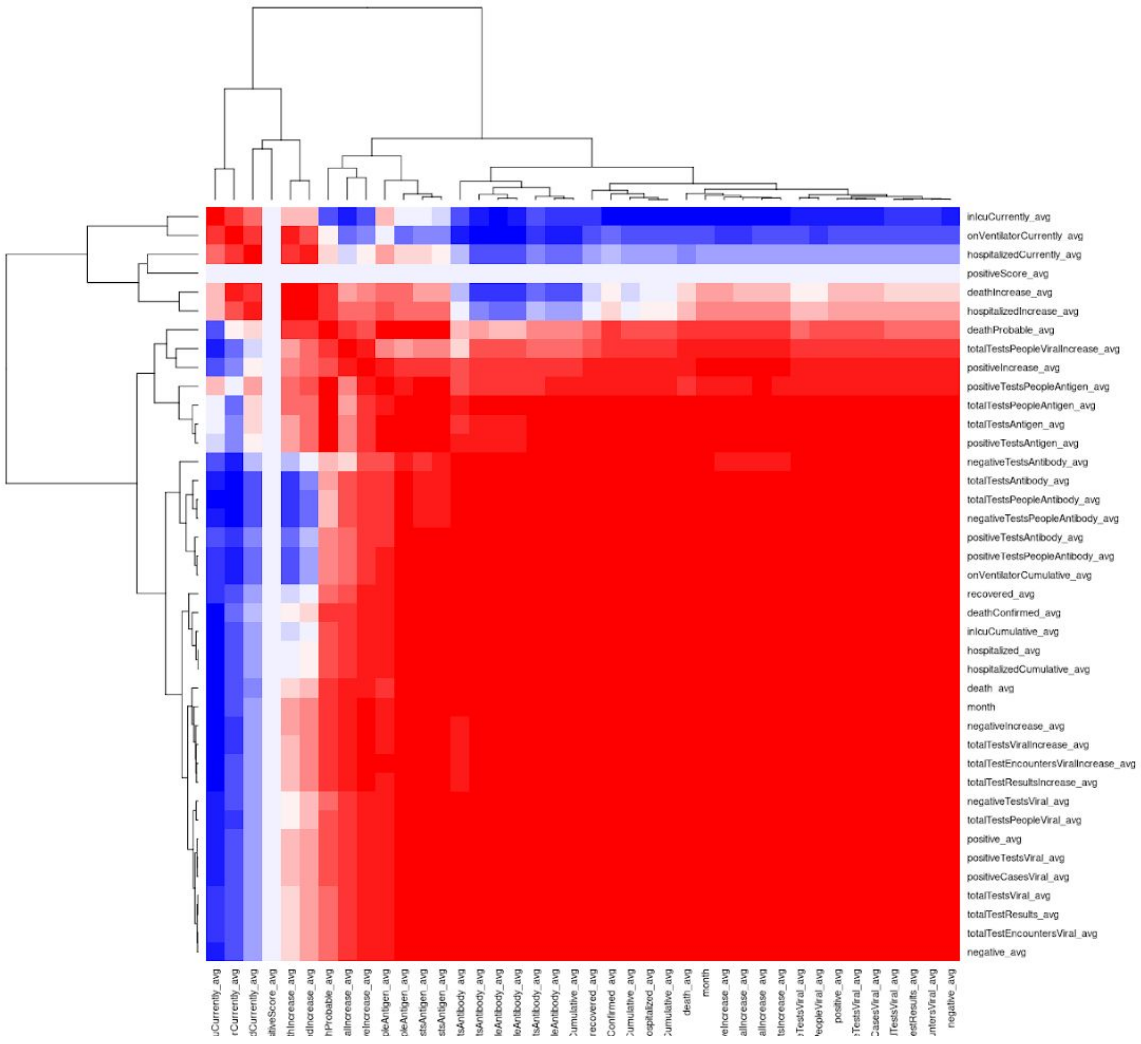
```
        Welch Two Sample t-test

data:  south$deathIncrease and west$deathIncrease
t = 12.994, df = 7466.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.829215 10.611171
sample estimates:
mean of x mean of y
 19.81510  10.59491
```

Using another CDC dataset, we compared the counts of flu-related deaths in Florida in 2018 (the most recent data) to the counts of COVID-related deaths in Florida by month using a paired t-test. We also calculated the Cohen's D value which indicated a high level of practical significance, expectedly.

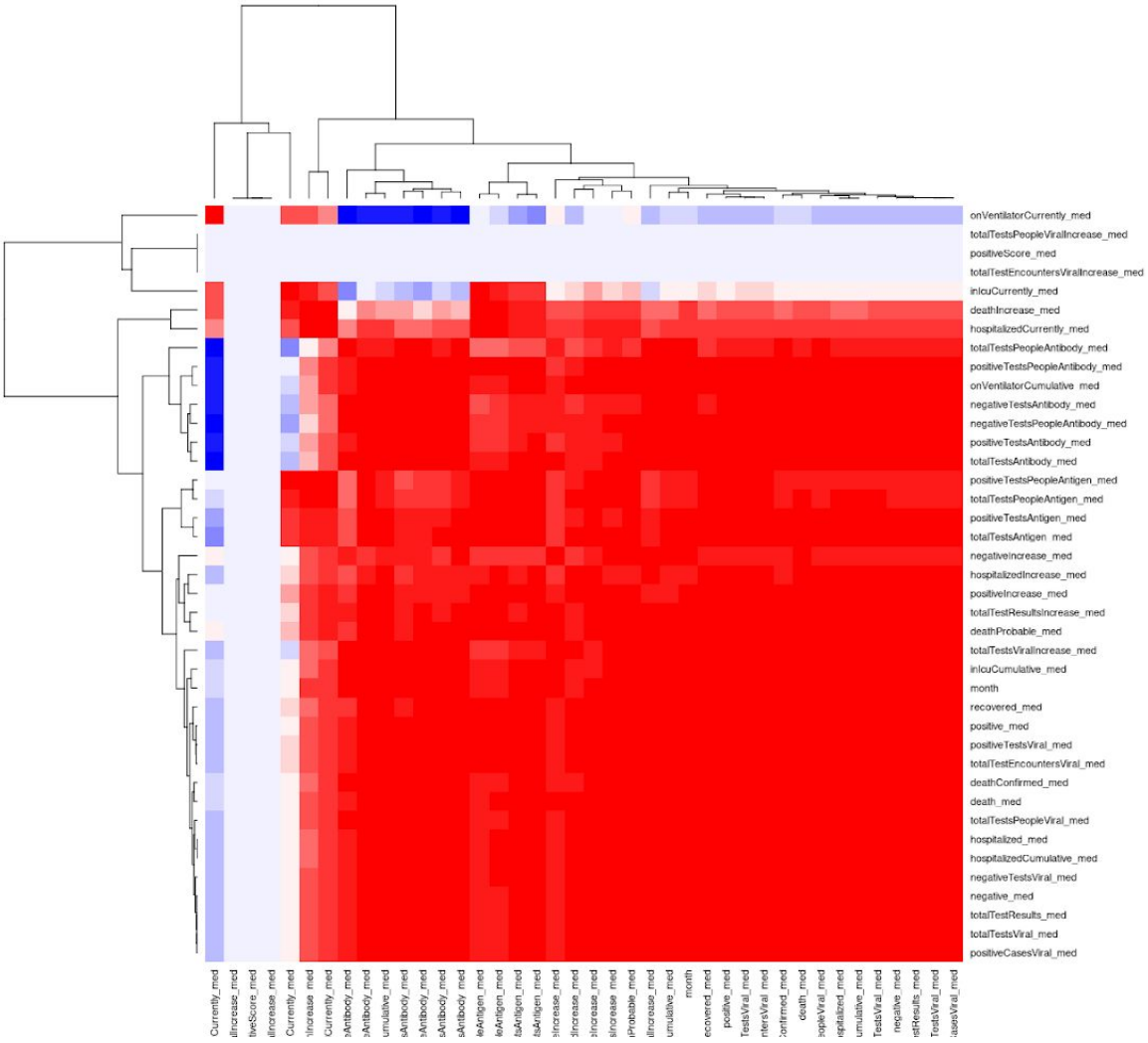
```
        Paired t-test

data:  fl.mo$totalDeaths and flu.fl1$Deaths
t = 3.4165, df = 8, p-value = 0.009135
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  557.0448 2870.5108
sample estimates:
mean of the differences
               1713.778

[1] "Cohen's D:  1.64"
```

5) We decided to perform multiple regression on the data with ‘recovered’ as our response variable. We analyzed the average, median, and total values by month for the sample population. We first inspected the correlation between each variable and our response variable.

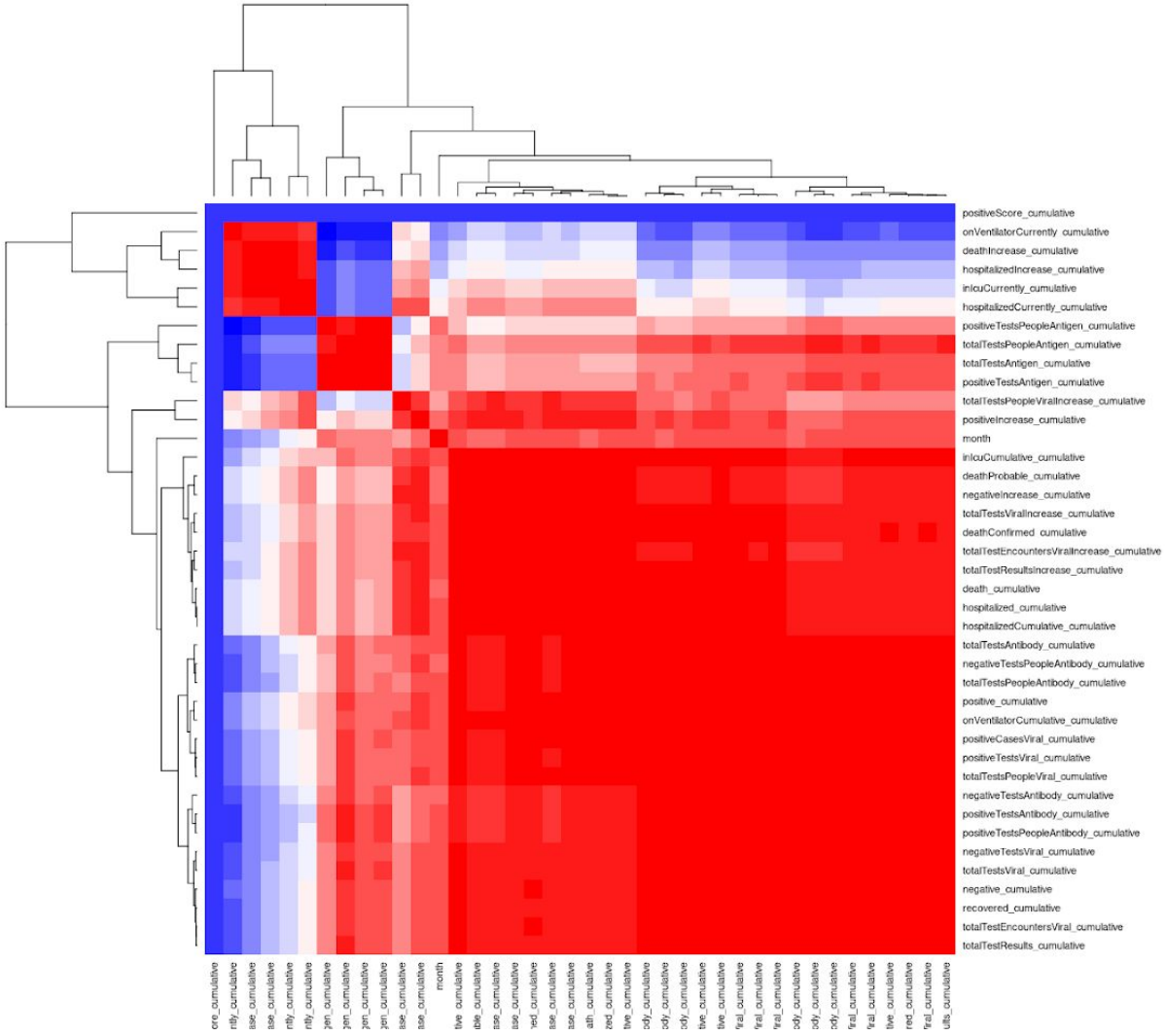
**Correlation of mean values**



## Correlation of median values

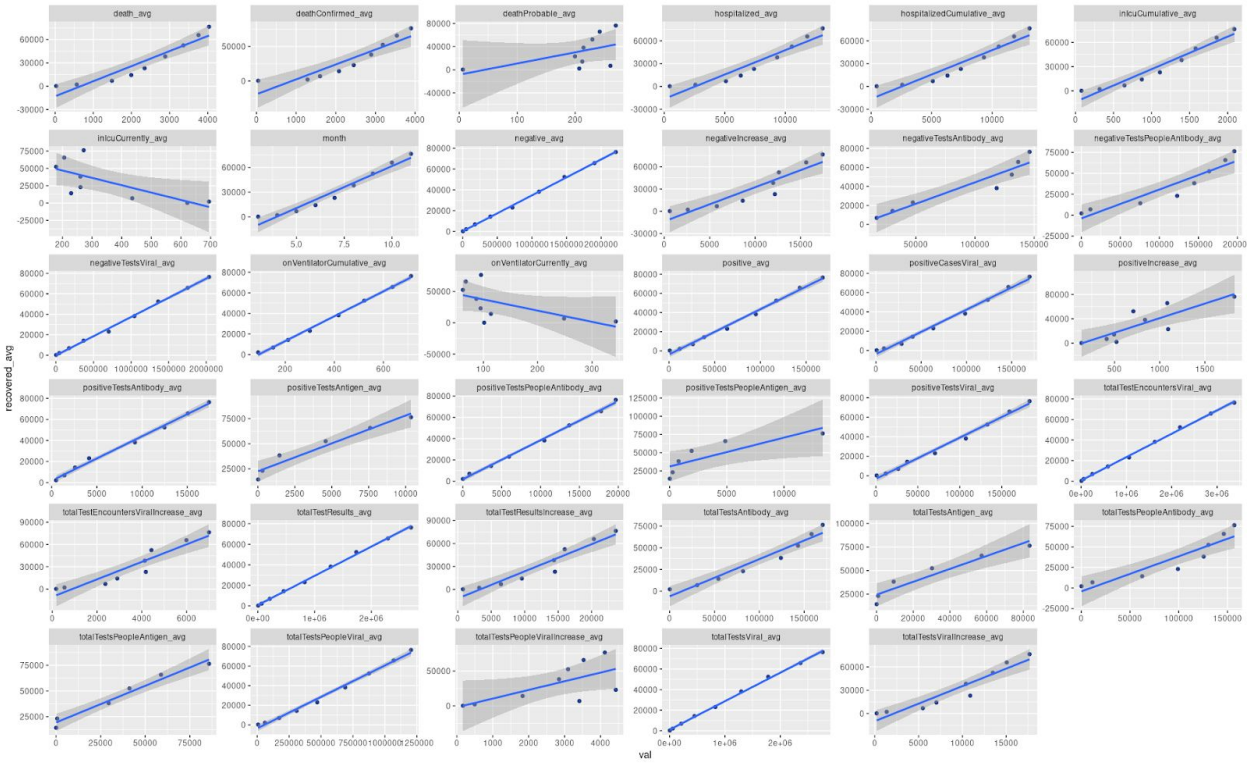


**Correlation of cumulative values**

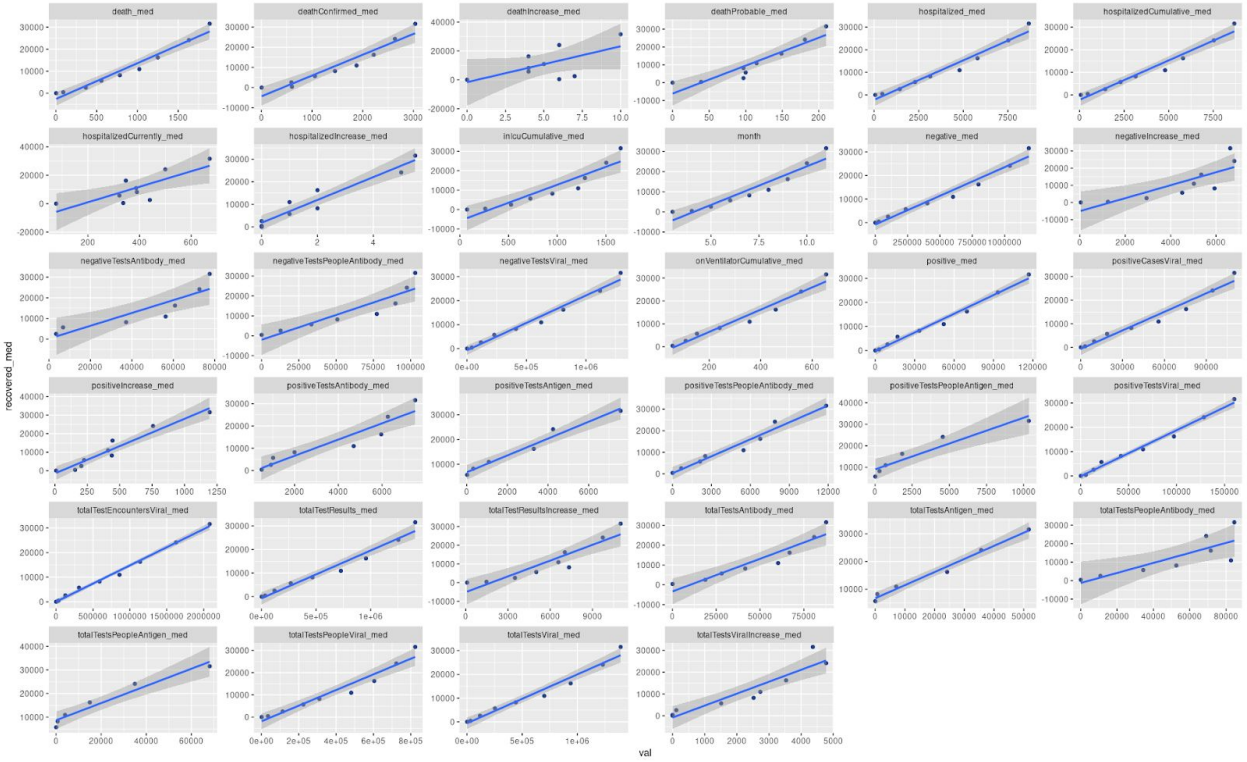



We selected variables that had a correlation greater than 0.5 or less than -0.5. We then plotted each of these against our response variable.

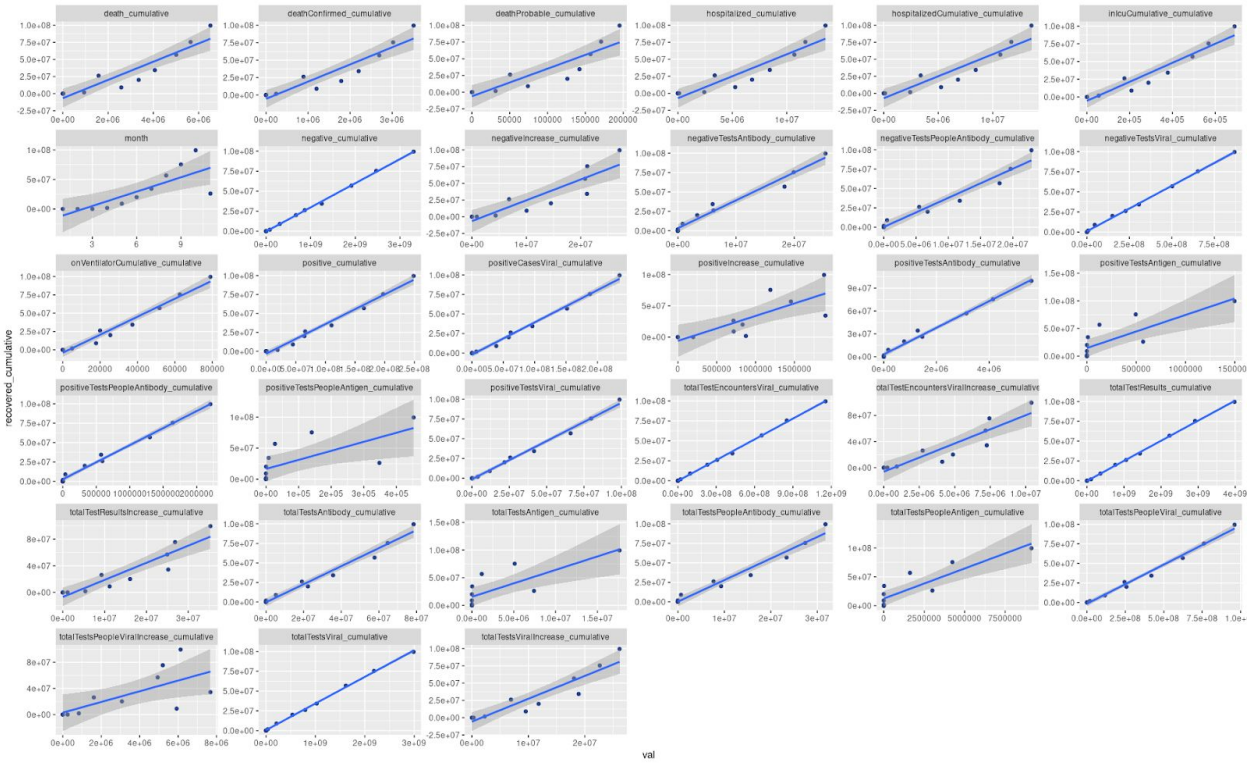
## Regression plots of avg values




## Regression plots of median values



**Regression plots of cumulative values**




The next step is to choose the variables that are the best predictor of the ‘recovered’ variable and discover if any of them are good candidates for multiple regression. This means finding two or more explanatory variables that are not correlated with each other but are correlated with ‘recovered’. Once the analysis is complete we can select the most interesting variable relationships to present.

# Questions:

For hypothesis testing, are we performing one test with a single mean value we choose or performing multiple tests on samples drawn from the dataset as the population and comparing to the mean of the dataset?