Analysis of COVID-19 and Transportation in the United States

Christian Miyares

***Abstract*—COVID-19 first appeared in China in early 2020. Soon after the virus spread throughout the whole world in what has been the worst pandemic in a century. Many countries made measures to combat the virus, some good others bad. A good example was China, the lockdown measures were strict but kept the virus from spreading. Meanwhile in the United States politics took precedence over public health and became the country with the most COVID cases in the world. This analysis looks to find correlation between transportation and the spread of COVID. We used data from John Hopkins University to calculate Confirmed and Fatal Rates. We also used data from The Bureau of Transportation Statistics. Using this data, we calculated Analysis of Variance on Low Medium and High groups of Confirmed and Fatal Rates. We also did Multiple and Linear Regression using a combination of variables from the data set in order to see if there was correlation between travel in the United States and confirmed rates of COVID as well as fatalities.**

I. INTRODUCTION

COVID 19 is easily spread thus the reason for the current global pandemic. In this analysis of COVID-19 we performed analysis of variance, as well as multiple and linear regression on data provided by John Hopkins University and The Bureau of Transportation Statistics. In this analysis we discovered that as transportation rises so do confirmed and fatal rates.

II. BODY

A. Data

COVID-19 has been an ultimate test of data collecting for governments and institutions. Since the virus is so widespread, a method of containing the virus is offering mass testing. Mass testing also means mass communication between hospitals pop up testing sites, and back too organizations to condense the data. Therefore, the data from John Hopkins University could contain some error. The United States is also a vast country and that makes it so that different states and politics may sometimes play into how the data is given. On the other hand, I believe that the data from the Bureau of Transportation Statistics (BTS) is accurate. The BTS was created in 1992 and therefore is used to handling the data. In fact, there are less people driving due to lockdowns in different states which further eases the load of data. In order to perform ANOVA, Multiple Regression, and Logistic Regression, we must make assumptions about the data. We will assume that the data is normally distributed. However, doing the Kolmogorov-Smirnov test we see that the data is not normally distributed for the variables Confirmed, Deaths, and People Tested. We will have to later perform non-parametric versions of ANOVA such as Kruskal-Wallis in order to justify our finding.

B. Methods

The first method that we did in this analysis was to calculate confirmed people over people tested in order to get a confirmed rate per state. The same was done for fatalities where it was number of deaths over people confirmed. The fifty states were then sorted and divided into groups of low, medium and high. I divided my groups using quartiles for both the confirmed and fatal rates. At this point having the groups we were able to perform analysis of variance. After performing regular analysis of variance, I performed the Kruskal Wallis Test as well in order to verify that our results were accurate do to the data not being normal. Then due to the results we performed a post-hoc Tukey HSD in order to see which groups were significantly different. All these methods were done for both groups. At this point we move on to the transportation data. For the transportation data we performed many logistic and multiple regression models. We compared the confirmed rates as well as the fatal rates to the transportation variables where the confirmed and fatal rates were divided into low, medium and high groups. After initially running models we

removed certain variables in order to have a more accurate model.

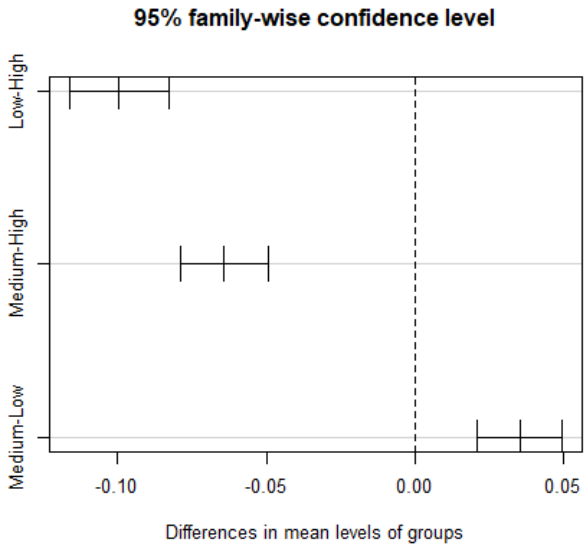
C. Analysis/Results

Beginning with the Confirmed and Fatal Rates. Here is an image as to how these were grouped by state.

	State	confirmedRates	fatalRate
1	Alabama	0.14282977	1.5430221
2	Alaska	0.02706854	0.5013757
3	Arizona	0.13883073	2.4310214
4	Arkansas	0.08329052	1.7158392
5	California	0.05010897	1.8944518
6	Colorado	0.05465870	2.1290076

Moving onto the Analysis of Variance for the confirmed Rates. Due to a very small p-value of almost 0 we can conclude that there exist significant differences between the paired of the low medium and high groups. To go further I also performed a Kruskal-Wallis test and got the same result. This result is trivial. The United States is very broad and as mentioned earlier in the *data* section, perhaps this data is incomplete making the differences in low medium and high different.

Performing a post-hoc test we can tell which groups are the ones that are significantly different from each other. In the case of confirmed rates that is all the groups.



The Tukey HSD test shows us that all of three groups are significantly different. The graphic of the confidence intervals above further prove this point do to the fact that none of the intervals span zero.

Next, we look at the Analysis of Variance of the Fatal Rates. The results are very similar to those of the Confirmed Rates. The p-value for both ANOVA and Kruskal-Wallis are essentially zero. Therefore, we can conclude that there exist significant differences between the through group combinations low medium and high.

```
Analysis of Variance Table

Response: treatments2
           Df    Sum Sq   Mean Sq F value             Pr(>F)
groups2     2 0.0080855 0.0040428  54.345 0.0000000000005972
Residuals 47 0.0034963 0.0000744
```

Furthermore, performing the Tukey-HSD post hoc test we get the same result that all three groups are significantly different from their pairs. Performing a confidence interval further proves the point with none of the intervals containing zero.

Onto the Logistic Regression. For Logistic regression we compared confirmed rate groups of low, medium and high to Active fatality rates and miles traveled. We also did the same for the fatal rates by comparing them to the confirmed rates as well as miles traveled. Beginning with high confirmed rates. We find no significance with fatal rates however; we do find several trips in the ranges below.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.57220  -0.64764  -0.34417  -0.01355   1.91467

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.232e+00  7.541e-01  -2.959  0.00308 **
Number.of.Trips..1       7.820e-08  4.421e-08   1.769  0.07694 .
Number.of.Trips.5.10    -1.008e-07  5.983e-08  -1.684  0.09214 .
Number.of.Trips.50.100  -6.375e-07  4.032e-07  -1.581  0.11389
Number.of.Trips.100.250  2.370e-06  9.831e-07   2.411  0.01590 *
Number.of.Trips...500   -1.648e-05  1.052e-05  -1.565  0.11749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.108  on 49  degrees of freedom
Residual deviance: 40.101  on 44  degrees of freedom
AIC: 52.101

Number of Fisher Scoring iterations: 6
```

As we can see the only significant variable is Number of Trips 100-250 miles. Therefore, states with high confirmed rates have a correlation of driving those miles and spreading the disease. This makes sense because at a small distance a person might take more precaution like at a grocery store. However, long distance a person may be going to an event or to see other family and may spread the disease and places such as those. Moving onto the medium group. We find no significance in any of the variables. However, the fatal rate variable is the closest at a p-value of 0.06. This is saying that states with medium confirmed rates tend to have a small increase in log odds of fatalities. Finally, with the low group I was not very surprised with the results. It was the group with the most significant variables. The low group had a connection between confirmed rates and Number of Trips between 250 and 500 miles. This is somewhat contradictory to the high rate group results but perhaps

due to state lines if someone traveled such a long distance, they may have infected someone in another state and therefore this anomaly exists.

Moving onto Logistic Regression comparing fatal rate groups of low, medium and high to confirmed rates as well as transportation data. Starting with the high group. The states with the highest fatalities also had a log odd increase in following variables below.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.137e-01  8.141e-01  -0.631  0.52804
Number.of.Trips.3.5  -9.365e-07  4.608e-07  -2.032  0.04213 *
Number.of.Trips.5.10  1.171e-06  5.285e-07   2.215  0.02675 *
Number.of.Trips.10.25 -5.091e-07  2.671e-07  -1.906  0.05668 .
Number.of.Trips.25.50  1.016e-06  4.074e-07   2.495  0.01261 *
Number.of.Trips.100.250 -1.709e-06  1.268e-06  -1.348  0.17769
Number.of.Trips.250.500 -1.614e-05  5.741e-06  -2.811  0.00494 **
Number.of.Trips...500   -2.191e-05  1.183e-05  -1.852  0.06398 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Other than the variable 100-250 miles every other variable is close or significant. That makes sense given that COVID is spread by contact and the more people are traveling the more at-risk groups encounter the disease and die. Moving onto the medium group the variables stayed plenty, but the miles moved down. This time mostly shorter trips led to medium fatalities. Finally, for the low group a combination of low trips and one long trip (250-500 miles) are significant. For more logistic regression model please see the appendix.

Now I will discuss multiple regression for the confirmed and fatal rates as well as the transportation variables. Beginning with Confirmed rates versus Fatal rates and transportation variables. For our first model we are comparing high confirmed rates to number of trips. Here are the results.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8941 -0.1530  0.1152  0.2949  0.6026

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.517e-01  8.355e-02  10.194 2.84e-13 ***
Number.of.Trips.1.3  -9.931e-09  7.024e-09  -1.414  0.16432
Number.of.Trips.5.10  1.847e-08  1.097e-08   1.684  0.09906 .
Number.of.Trips.50.100  6.939e-08  3.046e-08   2.278  0.02751 *
Number.of.Trips.100.250 -2.438e-07  7.560e-08  -3.225  0.00235 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4023 on 45 degrees of freedom
Multiple R-squared:  0.2015,    Adjusted R-squared:  0.1305
F-statistic: 2.839 on 4 and 45 DF,  p-value: 0.03502
```

As you can see the model is significant with a p-value of 0.04. More specifically there exists a correlation between states with high confirmed rates and traveling 50-100, and 100-250 miles. These are good ranges since essentially the ranges merge. As discussed earlier in the logistic regression section this result makes sense. Since most places are on lockdown people only go out to buy stuff from the grocery store or get gas. However, a person traveling longer distances spread the virus more.

The medium group is an odd one as there was no correlation found between any of the variables and the model was not significant. Finally, the low group, surprisingly there was a connection with states with low confirmed rates and fatal rates. This result must be due to perhaps states with a certain demographic of people that regardless of confirmed rate they are susceptible to the virus.

The last models I ran were doing multiple regression for the fatal rates versus the confirmed rates and transportation variables. Beginning with the high fatal group.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.961e-01  7.615e-02   7.828 9.74e-10 ***
Number.of.Trips.3.5   7.265e-08  2.675e-08   2.716  0.00955 **
Number.of.Trips.5.10 -7.895e-08  2.699e-08  -2.925  0.00554 **
Number.of.Trips.10.25 2.300e-08  1.381e-08   1.666  0.10325
Number.of.Trips.25.50 -5.967e-08  1.693e-08  -3.524  0.00104 **
Number.of.Trips.100.250  1.720e-07  7.847e-08   2.192  0.03397 *
Number.of.Trips.250.500  8.571e-07  2.846e-07   3.012  0.00438 **
Number.of.Trips...500    1.240e-06  7.082e-07   1.752  0.08714 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see there is a strong correlation between states that have a high fatal rate and travelers at almost all miles low and high. States with high fatal rates tend to be metropolitan due to higher population. In these states people tend to travel more whether it be long or short distance. However, due to the population density this probably made it so that more people die at all levels. For the medium group we get a mixed bag of results. There exists correlation at low miles traveled as well as high. This group being as big as it was probably making for less accurate result. Finally, the low fatal rates group. In this model we have significant variables at less miles traveled. This result fits in with the other models. The less miles traveled the less fatalities. Please see the appendix for more models.

III. Conclusion

COVID-19 has been a test for the entire world. At times like this, science is the most important tool to mankind. With analysis such as these we can draw different conclusions in order to learn more and help ourselves in getting better and saving lives. This is one of many analyses that we can do. However, an analysis of this type on transportation can help government leaders to make choices on whether to restrict transportation in order to stop the further spread or fatalities of its citizens. In this analysis I have concluded that generally the more one travels, the more confirmed rates there are. Both the logistic and multiple regression gave similar results. On the other hand, states with lower levels of confirmed and fatal rates have a correlation with less

distance traveled. Looking into the future, we could make hundred more models taking variables in and out as we see fit in order to see if we get different significance. As of the time of writing this the COVID-19 virus is very much alive and getting worse by the day in the United States which is our place of study. In the future it would be good to look back at all the travel data as well as the cumulative confirmed and fatal rates in order to truly see what difference it made. We could even go more in-depth state by state and add factors such as governors easing lockdowns and other protection measures. All in all, COVID-19 is a deadly and easily transmissible disease and one should take precautions to be safe.

IV. APPENDIX

A. ANOVA Graphics

```
Analysis of Variance Table

Response: treatments
          Df    Sum Sq  Mean Sq F value    Pr(>F)
groups     2 0.063645 0.031823  104.73 < 2.2e-16 ***
Residuals 47 0.014281 0.000304
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Kruskal-Wallis rank sum test

data:  treatments by groups
Kruskal-Wallis chi-squared = 41.353, df = 2, p-value = 1.048e-09
```

Analysis of Variance and Kruskal-Wallis between confirmed rates for low medium and high groups.

```
Analysis of Variance Table

Response: treatments2
          Df     Sum Sq   Mean Sq F value    Pr(>F)
groups2    2 0.0080855 0.0040428  54.345 5.972e-13 ***
Residuals 47 0.0034963 0.0000744
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Kruskal-Wallis rank sum test

data:  treatments by groups
Kruskal-Wallis chi-squared = 41.353, df = 2, p-value = 1.048e-09
```

Analysis of Variance and Kruskal-Wallis between fatal rates for low medium and high groups.

B. Logistic Regression

High Confirmed

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.57220  -0.64764  -0.34417  -0.01355   1.91467

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.232e+00  7.541e-01  -2.959  0.00308 **
Number.of.Trips..1       7.820e-08  4.421e-08   1.769  0.07694 .
Number.of.Trips.5.10    -1.008e-07  5.983e-08  -1.684  0.09214 .
Number.of.Trips.50.100  -6.375e-07  4.032e-07  -1.581  0.11389
Number.of.Trips.100.250  2.370e-06  9.831e-07   2.411  0.01590 *
Number.of.Trips...500   -1.648e-05  1.052e-05  -1.565  0.11749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.108  on 49  degrees of freedom
Residual deviance: 40.101  on 44  degrees of freedom
AIC: 52.101
```

Medium Confirmed

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9147  -1.0691  -0.5709   1.1308   2.0616

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.877e-01  5.789e-01   0.670   0.5031
fatalRate      -4.712e-01  2.536e-01  -1.858   0.0631 .
Number.of.Trips..1  4.020e-09  2.708e-09   1.485   0.1376
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 63.555  on 47  degrees of freedom
AIC: 69.555
```

Low Confirmed

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9177  -0.4886  -0.2008   0.1324   2.4695

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.259e-01  1.019e+00  -0.712   0.4762
fatalRate                8.522e-01  4.375e-01   1.948   0.0514 .
Number.of.Trips.1.3     -2.458e-07  1.333e-07  -1.844   0.0652 .
Number.of.Trips.3.5      4.688e-07  2.519e-07   1.861   0.0627 .
Number.of.Trips.250.500 -6.300e-06  3.184e-06  -1.979   0.0479 *
Number.of.Trips...500    1.295e-05  8.216e-06   1.576   0.1151
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.306  on 49  degrees of freedom
Residual deviance: 32.558  on 44  degrees of freedom
AIC: 44.558
```

High fatal

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.80551  -0.26530  -0.07668   0.09430   1.95426

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.137e-01  8.141e-01  -0.631  0.52804
Number.of.Trips.3.5     -9.365e-07  4.608e-07  -2.032  0.04213 *
Number.of.Trips.5.10     1.171e-06  5.285e-07   2.215  0.02675 *
Number.of.Trips.10.25   -5.091e-07  2.671e-07  -1.906  0.05668 .
Number.of.Trips.25.50    1.016e-06  4.074e-07   2.495  0.01261 *
Number.of.Trips.100.250 -1.709e-06  1.268e-06  -1.348  0.17769
Number.of.Trips.250.500 -1.614e-05  5.741e-06  -2.811  0.00494 **
Number.of.Trips...500   -2.191e-05  1.183e-05  -1.852  0.06398 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Medium fatal

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6067  -0.8041  -0.5568   0.8550   1.8630

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.493e+00  5.421e-01  -2.754  0.00589 **
Number.of.Trips.5.10    -1.305e-07  6.835e-08  -1.910  0.05616 .
Number.of.Trips.10.25    1.891e-07  9.561e-08   1.978  0.04788 *
Number.of.Trips.25.50   -2.184e-07  1.045e-07  -2.090  0.03660 *
Number.of.Trips.250.500  3.834e-06  1.708e-06   2.245  0.02476 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 53.762  on 45  degrees of freedom
AIC: 63.762
```

Low fatal

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.58904  -0.45921  -0.04052   0.35873   1.63054

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -6.388e-01  9.405e-01  -0.679   0.4970
Number.of.Trips.3.5      6.285e-07  3.648e-07   1.723   0.0849 .
Number.of.Trips.5.10    -6.811e-07  3.032e-07  -2.246   0.0247 *
Number.of.Trips.100.250  1.491e-06  8.878e-07   1.679   0.0931 .
Number.of.Trips.250.500  6.776e-06  3.518e-06   1.926   0.0541 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.295  on 49  degrees of freedom
Residual deviance: 30.312  on 45  degrees of freedom
AIC: 40.312
```

C. Multiple Regression

High Confirmed

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8941 -0.1530  0.1152  0.2949  0.6026

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.517e-01  8.355e-02  10.194 2.84e-13 ***
Number.of.Trips.1.3     -9.931e-09  7.024e-09  -1.414  0.16432
Number.of.Trips.5.10     1.847e-08  1.097e-08   1.684  0.09906 .
Number.of.Trips.50.100   6.939e-08  3.046e-08   2.278  0.02751 *
Number.of.Trips.100.250 -2.438e-07  7.560e-08  -3.225  0.00235 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Medium Confirmed

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.9117 -0.4737  0.1535  0.4490  0.8454

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.346e-01  1.328e-01   3.271  0.00201 **
fatalRate                9.226e-02  4.690e-02   1.967  0.05508 .
Number.of.Trips.3.5 -1.568e-09  9.914e-10  -1.581  0.12052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Low Confirmed

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8140 -0.1312  0.1159  0.2009  0.9495

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.787e-01  1.281e-01   6.081 2.19e-07 ***
fatalRate               -9.301e-02  4.173e-02  -2.229   0.0307 *
Number.of.Trips.25.50   -1.510e-08  8.450e-09  -1.787   0.0805 .
Number.of.Trips.100.250  1.327e-07  5.872e-08   2.260   0.0286 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

High Fatal

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.66447 -0.18729  0.07253  0.22805  0.60210

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.961e-01  7.615e-02   7.828 9.74e-10 ***
Number.of.Trips.3.5      7.265e-08  2.675e-08   2.716  0.00955 **
Number.of.Trips.5.10    -7.895e-08  2.699e-08  -2.925  0.00554 **
Number.of.Trips.10.25    2.300e-08  1.381e-08   1.666  0.10325
Number.of.Trips.25.50   -5.967e-08  1.693e-08  -3.524  0.00104 **
Number.of.Trips.100.250  1.720e-07  7.847e-08   2.192  0.03397 *
Number.of.Trips.250.500  8.571e-07  2.846e-07   3.012  0.00438 **
Number.of.Trips...500    1.240e-06  7.082e-07   1.752  0.08714 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Medium Fatal

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8364 -0.3432  0.1748  0.2842  1.0602

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.541e-01  9.353e-02   8.062 4.58e-10 ***
Number.of.Trips.1.3      2.443e-08  1.362e-08   1.794   0.0800 .
Number.of.Trips.3.5     -7.748e-08  5.285e-08  -1.466   0.1501
Number.of.Trips.5.10     5.905e-08  3.893e-08   1.517   0.1368
Number.of.Trips.10.25   -4.185e-08  1.921e-08  -2.178   0.0350 *
Number.of.Trips.25.50    3.369e-08  2.010e-08   1.676   0.1012
Number.of.Trips.250.500 -5.386e-07  3.049e-07  -1.766   0.0846 .
Number.of.Trips...500   -1.443e-06  1.042e-06  -1.385   0.1734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Low Fatal

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.89173 -0.22705  0.05975  0.28788  0.60469

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.834e-01  1.247e-01   6.281  1.2e-07 ***
confirmedRates          -3.015e+00  1.527e+00  -1.975  0.05445 .
Number.of.Trips..1      -7.264e-09  3.106e-09  -2.339  0.02385 *
Number.of.Trips.10.25    1.888e-08  6.770e-09   2.788  0.00774 **
Number.of.Trips.100.250 -1.146e-07  5.553e-08  -2.063  0.04491 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

D. Other

Testing Normality

```
        One-sample Kolmogorov-Smirnov test

data:  Confirmed
D = 1, p-value = 8.882e-16
alternative hypothesis: two-sided

> ks.test(People_Tested,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  People_Tested
D = 1, p-value = 8.882e-16
alternative hypothesis: two-sided

> ks.test(Deaths,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  Deaths
D = 1, p-value = 8.882e-16
alternative hypothesis: two-sided
```

E. R-Code

```
# Covid Project
options(scipen=0)

# loading in data
library(ggplot2)
library(mosaic)
library(readxl)
library(dplyr)
library(tidyverse)
library(car)
library(MASS)
library(olsrr)

coviddata =
read_excel("C:/Users/chris/Desktop/COVIDPROJ/coviddata.xlsx")
View(coviddata)
attach(coviddata)
names(coviddata)

#calculates the confirmed rates based off of Confirmed cases to people tested per each state
confirmedRates = (Confirmed/People_Tested)
confirmedRates


#making low medium and high groups


summary(confirmedRates)

#low: 0.005339 - 0.042911
#medium : 0.042912 - 0.083103
#high:0.083104 - 0.177211

#sorting data

sort(confirmedRates, decreasing= FALSE)

lowC = c(0.00533858, 0.01068466, 0.01797617, 0.02598260,
0.02706854, 0.02865046
      ,0.02923934, 0.03111400, 0.03141964,0.03493618,
0.04001250, 0.04019186
      ,0.04244679 )
mediumC = c(0.04430501, 0.04496721, 0.04823623,
0.05010897, 0.05109119,
         0.05211561, 0.05277674, 0.05282349, 0.05285332,
0.05465870,
         0.05656046, 0.06162199, 0.06574467, 0.06595213,
0.06772489,
         0.06851954, 0.06899711, 0.07099350, 0.07387818,
0.07539736,
         0.07590111, 0.08041048, 0.08042079, 0.08254203,
0.08329052)
highC = c(0.08435180, 0.09800616, 0.10126921, 0.10185355,
0.11432847, 0.12878861, 0.12998364,
       0.13883073, 0.14282977, 0.14432320, 0.16828769,
0.17721129)

#combination of low medium and high rates
treatments = c(lowC,mediumC,highC)
groups =
factor(c(rep("Low",13),rep("Medium",25),rep("High",12)))

x = data.frame(treatments,groups)
x
#anova
fit = lm(treatments~groups)
model1 = anova(fit)
model1

#anova model to fit tukey function
aovModel1 = aov(fit)

kruskal.test(treatments~groups)

#tukey
TukeyHSD(aovModel1)
plot(TukeyHSD(aovModel1))

# there is no statistically significant difference between the means


# fatality rates

fatalRate = Deaths/Confirmed

fatalRate

summary(fatalRate)


sort(fatalRate, decreasing= FALSE)
```

```
lowF =
c(0.005013757,0.005267932,0.006542337,0.009011887,0.009
217893,0.009240738,0.009735636,

0.010890992,0.011432578,0.011931870,0.012381769,0.0128
62908,0.013353046,0.013850157)
mediumF =
c(0.014531219,0.015338038,0.015430221,0.015941158,0.016
328659,0.016912280,0.017158392,

0.017596094,0.018765331,0.018944518,0.019806685,0.0201
66676,0.020884758,0.021290076,

0.021897182,0.022009098,0.022065446,0.022115358,0.0222
80479,0.024152812,0.024310214,
          0.024456256,0.024576142)
highF =
c(0.026617715,0.027746338,0.028375616,0.028544682,0.032
473803

,0.036533431,0.039000841,0.041355133,0.043576326,0.0630
04490,0.064825087,0.066025933,
         0.0687304)

#combination of low medium and high rates
treatments2 = c(lowF,mediumF,highF)
groups2 =
factor(c(rep("Low",14),rep("Medium",23),rep("High",13)))


x = data.frame(treatments2,groups2)
x
#anova
fit2 = lm(treatments2~groups2)
model2 = anova(fit2)
model2

#anova model to fit tukey function
aovModel2 = aov(fit2)

kruskal.test(treatments~groups)

#tukey
TukeyHSD(aovModel2)
plot(TukeyHSD(aovModel2))


# expected confirmed rate per 100 people

eCRate = lm(treatments*100~groups)
eCRate
anova(eCRate)
aov(eCRate)
TukeyHSD(aov(eCRate))
plot(TukeyHSD(aov(eCRate)))
#expected fatal rate per 100 people

fRate = lm(treatments2*100~groups2)
fRate
anova(fRate)
aov(fRate)
TukeyHSD(aov(fRate))
plot(TukeyHSD(aov(fRate)))
########################################

#testing normality of data


ks.test(Confirmed,"pnorm")
ks.test(People_Tested,"pnorm")
ks.test(Deaths,"pnorm")


################################################
####

data_confirmed <-
data.frame(State=coviddata$Province_State,
confirmedRates=confirmedRates, fatalRate = Mortality_Rate)
data_confirmed


#multiple regression
setwd("C:/Users/chris/Desktop/COVIDPROJ/")

trip<-read.csv("Trips_by_Distance.csv")
trip

# delete the county observations.
tripdata<-trip[which(trip$Level=="State"),]
tail(tripdata)
head(tripdata)

################################################
###############
# 1.pick up the trip data of Oct
# 2.aggregate the transportation data by month
require('tidyverse')
trip_1<-as_tibble(tripdata)
trip_a<-trip_1 %>% filter(str_detect(tolower(Date), pattern =
"2020/10")) %>% group_by(State.Postal.Code) %>%

summarise_at(vars(Number.of.Trips..1:Number.of.Trips...500)
,~sum(.))
trip_a


names(trip_a)[1] <- "State"

###### set State name of trip_a

# Get state names of the data
State <- trip_a$State
State
# No "DC" in state.abb
state.abb=="DC"
# Delete the row of DC.
data_name<-trip_a[-8,]
data_name
```

```
state_f <- character(50)
for (i in 1:50) {
  state_f[i]<-
state.name[which(state.abb==data_name$State[i])]
}
state_f

data_2<-data.frame(State=state_f,data_name[2:11])
head(data_2)


data<- merge(data_confirmed,data_2,by="State")
head(data)
str(data)


sort(data$confirmedRates)

hist(data$confirmedRates)

#logistic regression confirmed rates
################################

#high group
highRate = ifelse(
  data$confirmedRates>0.084,0,1)
highRate

groupHigh = data.frame(data,highRate)


log<- glm(highRate == 0 ~
fatalRate+Number.of.Trips..1+Number.of.Trips.1.3 +
          Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
          Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
          Number.of.Trips.250.500+Number.of.Trips...500,
data=groupHigh,family = "binomial")
summary(log)

step = stepAIC(log,direction = "both")
summary(step)



#medium group
medRate = ifelse(

(data$confirmedRates>=0.04430501&data$confirmedRates<=
0.08329052),0,1)
medRate

groupMed = data.frame(data,medRate)

log2<- glm(medRate == 0 ~
fatalRate+Number.of.Trips..1+Number.of.Trips.1.3 +
          Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
          Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
          Number.of.Trips.250.500+Number.of.Trips...500,
data=groupMed,family = "binomial")
summary(log2)

step = stepAIC(log2,direction = "both")
summary(step)

#lowgroup

lowRate = ifelse(
  (data$confirmedRates>0 & data$confirmedRates<0.044),0,1)
lowRate

groupLow = data.frame(data,lowRate)

log3<- glm(lowRate == 0 ~
fatalRate+Number.of.Trips..1+Number.of.Trips.1.3 +
          Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
          Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
          Number.of.Trips.250.500+Number.of.Trips...500,
data=groupLow,family = "binomial")
summary(log3)

step = stepAIC(log3,direction = "both")
summary(step)



############################
# log regression fatal rates

#high fatal Rate
sort(data$fatalRate)

highFRate = ifelse(
  data$fatalRate>2.6,0,1)
highFRate

groupFHigh = data.frame(data,highFRate)


log<- glm(highFRate == 0 ~
confirmedRates+Number.of.Trips..1+Number.of.Trips.1.3 +
          Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
          Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
          Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFHigh,family = "binomial")
summary(log)

step = stepAIC(log,direction = "both")
summary(step)

#med fatal rate

medFRate = ifelse(
  (data$fatalRate>1.4&data$fatalRate<2.5),0,1)
```

```
medFRate

groupFmed = data.frame(data,medFRate)


log2<- glm(medFRate == 0 ~
confirmedRates+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFmed,family = "binomial")
summary(log2)

step = stepAIC(log2,direction = "both")
summary(step)

#low fatal rate

lowFRate = ifelse(
  (data$fatalRate>0&data$fatalRate<1.4),0,1)
lowFRate

groupFlow = data.frame(data,lowFRate)


log3<- glm(lowFRate == 0 ~
confirmedRates+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFlow,family = "binomial")
summary(log3)

step = stepAIC(log3,direction = "both")
summary(step)

##############################
#multiple regression confirmed rates

lm1 = lm(highRate~
fatalRate+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupHigh)
summary(lm1)

steplm1 = stepAIC(lm1,direction="both")
summary(steplm1)



lm2 = lm(medRate ~
fatalRate+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupMed)
summary(lm2)

steplm2 = stepAIC(lm2,direction="both")
summary(steplm2)

lm3 = lm(lowRate ~ fatalRate +
Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupLow)
summary(lm3)

steplm3 = stepAIC(lm3,direction = "both")
summary(steplm3)



#multiple regression fatal rates

lmf1 = lm(highFRate~
confirmedRates+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFHigh)
summary(lmf1)

steplmf1 = stepAIC(lmf1,direction="both")
summary(steplmf1)



lmf2 = lm(medFRate ~
confirmedRates+Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFmed)
summary(lmf2)

steplmf2 = stepAIC(lmf2,direction="both")
summary(steplmf2)

lmf3 = lm(lowFRate ~ confirmedRates +
Number.of.Trips..1+Number.of.Trips.1.3 +
        Number.of.Trips.3.5+
Number.of.Trips.5.10+Number.of.Trips.10.25 +
```

```
        Number.of.Trips.25.50 +
Number.of.Trips.50.100+Number.of.Trips.100.250+
        Number.of.Trips.250.500+Number.of.Trips...500,
data=groupFlow)
summary(lmf3)

steplmf3 = stepAIC(lmf3,direction = "both")
summary(steplmf3)
```