STA 104: Take Home Project
Looking into States that are being affected from Covid-19

Aman Singh & Peter Mueller
amasin@ucdavis.edu
pfmueller@ucdavis.edu
UC Davis
Instructor: Amy T. Kim

2/28/2021

(I) Introduction

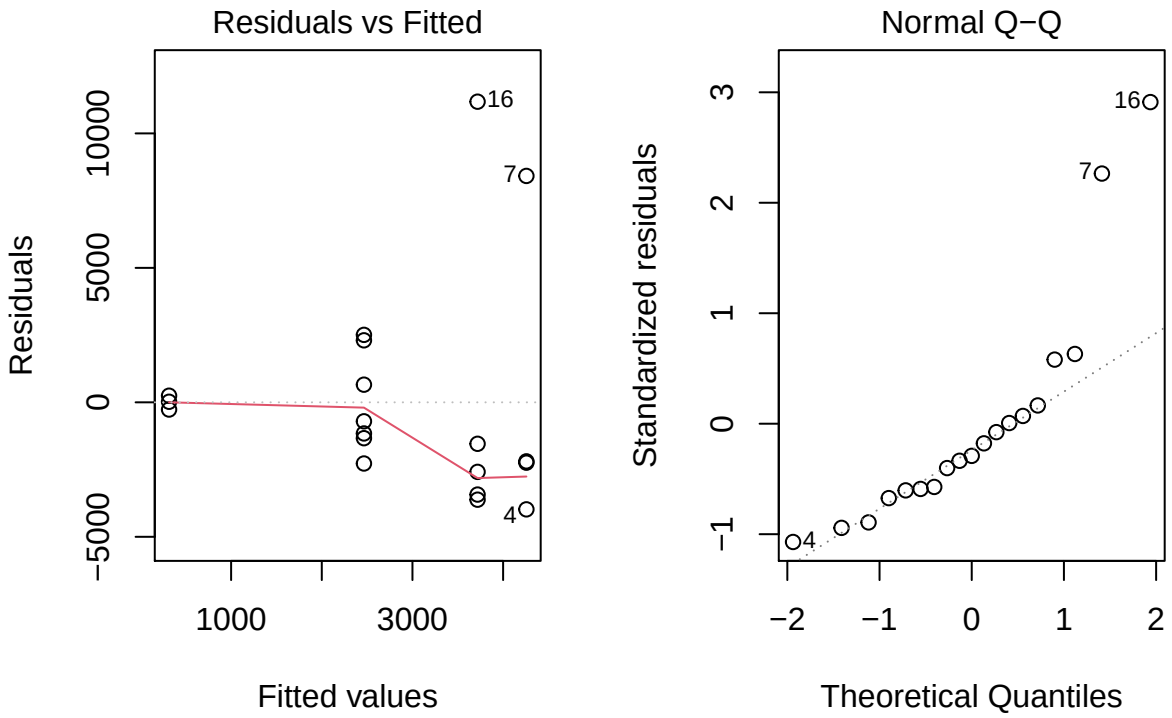
Covid-19 has vastly changed the way of how the world works. Millions of people in America are all unable to do their work properly and have been forced to work at home. We are currently analyzing an data set called CovidA. which lists all the months of states as will as there monthly deaths. We chose to instead subset the dataset and work with 4 states. We chose California, New York City, Florida, and Alabama. We chose California because that is where my partner and I are located and we would like to know more about California compared to other states. We chose New York City because NYC is a typically very popular spot for visitors and we were curious on how the analysis will work with an heavily populated city with Corona. We chose Alabama because they are an republican state are are widely known for not believeing in the Coronavirus. We also chose Florida largely because it is one of the only states where nobody believed in the Coronavirus. All the bars, indoor dining and tourism is rampant over there with the state not caring if anyone wears a mask. We are curious on how Florida and Alabama is doing in terms of the virus compared to states that are very progressive and strict on guidelines such as California and NYC.

(II) Summary of Data

Table 1: Summary Statistics

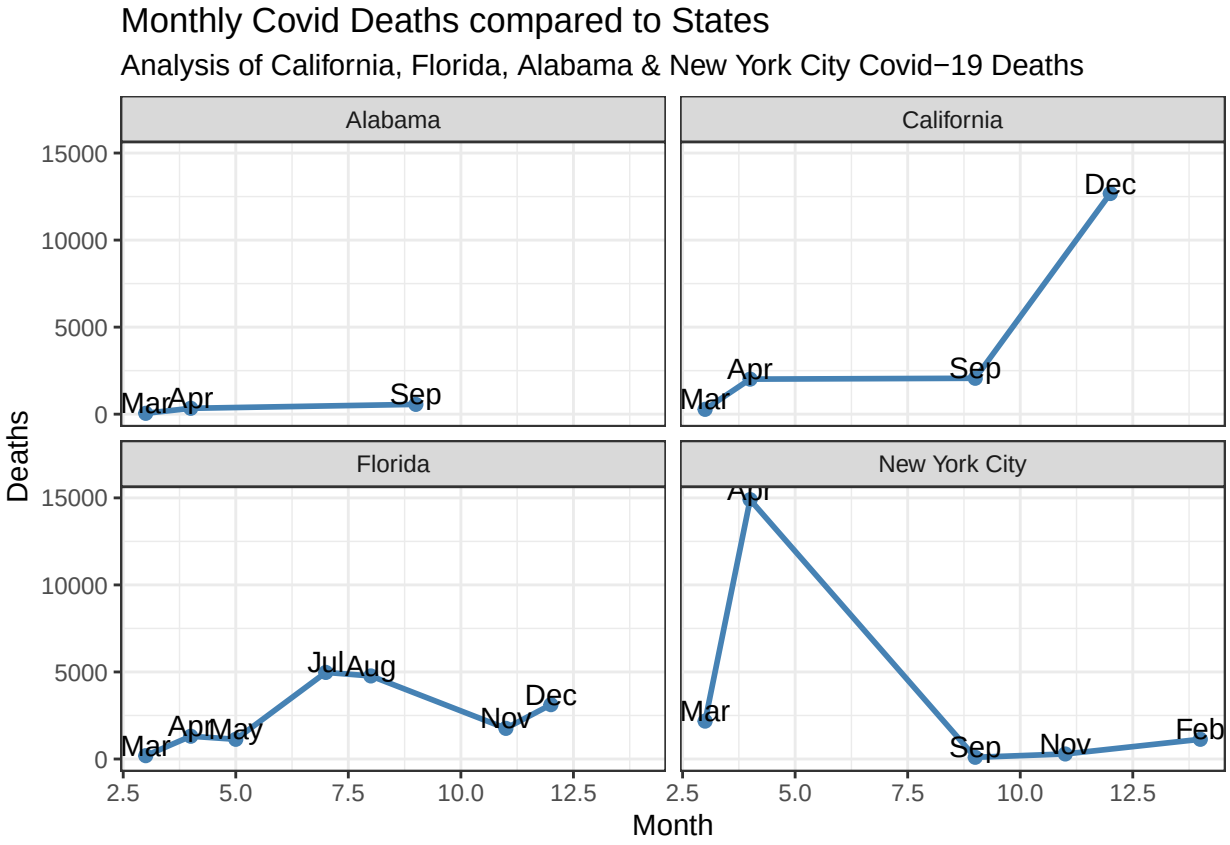
	California	Florida	Alabama	New York City
Group Mean	4256.750	2465.00000	315.666667	3718.200
Group SD	5673.232	1863.53750	257.663217	6302.866
Rank Mean	11.750	11.42857	4.666667	9.800
Sample Size	4.000	7.00000	3.00000	5.000

%



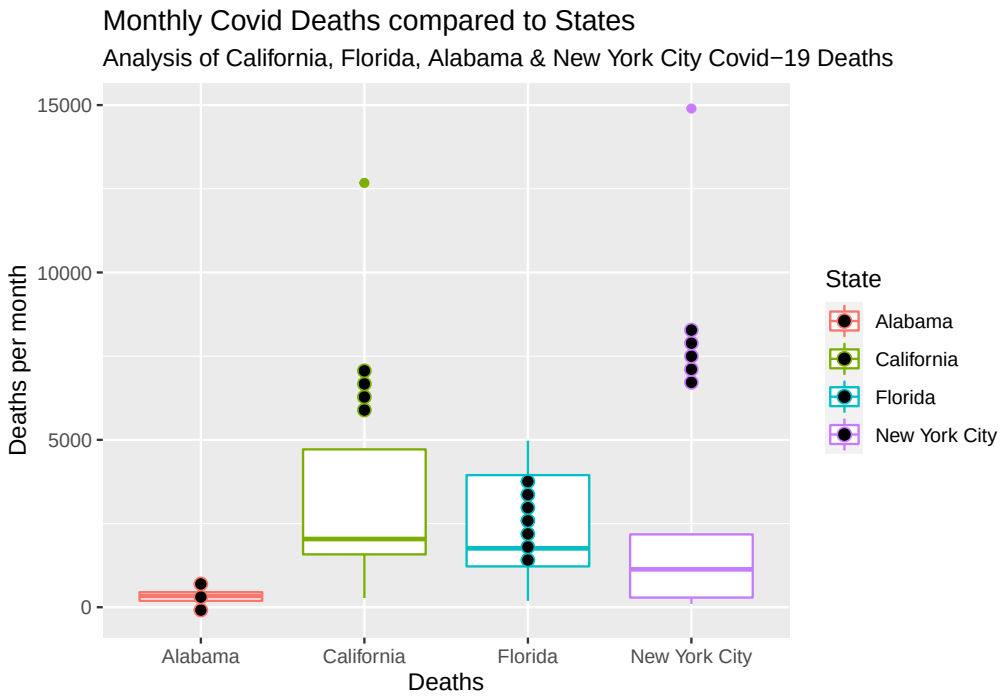
We see from the Shapiro- Wilks test based on the hypothesis that the variable is normally distributed in the dataset. We see that since the p-value is really low, we reject the null hypothesis and confirm that that the variable is not normally distributed. We also use the Levene test to test if K samples have equal variances and find that we fail to reject the null hypothesis thus concluding that no violations of any assumptions have happened.

In this forthcoming graph, I purposely ended up changing the dataset where I made an replica of the subsetted graph that we had from the CovidA dataset. I than changed the months to where month 1 was January of 2020 and month 14 was February of 2021. This helped in viewing the graphs as there was not data for every month and when I first did the graph, specifically for NYC, there was no data of February of 202 but there was of February 2021. However, the reason why I changed it was because that data of 2021 was showing up earlier in the 2020 panel which was messing up the graph thus the reason as to why I changed the index. I also added another column in the dataset which has the abbreviation names for the months for aesthetically pleasing graphs using GGplot.



We see here that surprisingly, there were a lot of deaths in California, despite it being a much more Democratic and very progressive state. While they had it mostly low, the sudden increase in deaths was due to the fact that it was December, and people were probably visiting their friends and loved ones regardless which led to an increase in deaths. Florida, in this case is very surprising how they have managed to keep their deaths down despite the fact that everything is open there and nobody believes in the virus. Alabama while is not too surprising because its not as a popular spot to travel to and is not know for tourism. However, it is a shock that not too many deaths happened there as it is a republican state and indoor dining has been going on regardless of the pandemic. NYC it seems like had the most death earlier on in 202, which makes sense as a lot of people still weren't the processing the fact of how serious the pandemic was and so people were still traveling and doing work as well as flights being very active which shows in the graph a to why NYC had too many deaths in April of 2020.

Lastly we also graph the distribution of deaths in order to analyze which test we would be using, the F-test or the Kruskall-Wallis Test.



Looking at this dataset, we see that there are outliers in multiple groups as well as skewed distributions as well. This leads us to believe to use the Kruskal - Wallis Test in order to understand the results a bit more. We also use Kruskal - Wallis as it has more power than a permutation test when exposed to outliers, or an skewed distribution in 1 or more of the groups. We will be using the regular Kruskal - Wallis Test as the Large Sample Approximation requires the sample size to be $n \geq 30$ which in this case is only 19.

(III) Analysis

(IV) Interpretation

(V) Conclusion

VI

Code Appendix

```
# cuttingoffcode
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)
# importing dataset
library(readr)
CovidA <- read_csv("CovidA.csv")
# subsetting data
CovidA_sub = subset(CovidA, CovidA$State == "California" | CovidA$State ==
    "New York City" | CovidA$State == "Florida" | CovidA$State == "Alabama")
CovidA_sub$Rank = rank(CovidA_sub$Death, ties = "average")
Group.order = aggregate(Death ~ State, data = CovidA_sub, mean)$State
Xi = aggregate(Death ~ State, data = CovidA_sub, mean)$Death
si = aggregate(Death ~ State, data = CovidA_sub, sd)$Death
Ri = aggregate(Rank ~ State, data = CovidA_sub, mean)$Rank
ni = aggregate(Death ~ State, data = CovidA_sub, length)$Death
results = rbind(Xi, si, Ri, ni)
rownames(results) = c("Group Mean", "Group SD", "Rank Mean", "Sample Size")
colnames(results) = as.character(Group.order)
SR.2 = var(CovidA_sub$Rank)
Ano = aov(Death ~ State, data = CovidA_sub)
par(mfrow = c(1, 2))
plot(Ano, 1)
plot(Ano, 2)
shapiro.test(CovidA_sub$Death)
library(lawstat)
levene.test(CovidA_sub$Death, as.factor(CovidA_sub$State))
library(ggplot2)
library(dplyr)
# converting months as to not get confused by year so month 1 is
# January 2020 and month 14 is February 2021
CovidA_sub_graph = CovidA_sub
CovidA_sub_graph$Month[19] = 14
CovidA_sub_graph$month.name = month.abb[CovidA_sub$Month]
# ggplot graph with facetwrap
ggplot(data = CovidA_sub_graph, aes(Month, Death)) + geom_line(color = "steelblue",
    size = 1) + geom_point(color = "steelblue", size = 2) + labs(title = "Monthly Covid Deaths compared
    subtitle = "Analysis of California, Florida, Alabama & New York City Covid-19 Deaths",
    y = "Deaths", x = "Month") + facet_wrap(~State) + theme_bw() + geom_text(aes(label = month.name),
    hjust = 0.5, vjust = 0)
ggplot(CovidA_sub, aes(x = Death, y = State, color = State)) + geom_boxplot() +
    coord_flip() + geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.8) +
    labs(title = "Monthly Covid Deaths compared to States", subtitle = "Analysis of California, Florida
        y = "Deaths", x = "Deaths per month")
```