“Analysis of Factors That Related To Covid-19 Pandemic Spread Out”

Prepared for

Dr. Shobhana Stoyanov
University of California, Berkeley

Prepared by

Winston Mai, Caixia Zhang, Jiahui Huang,
Mary Guo, Yaoying Cai, Yupeng Zhang

Introduction

To prevent COVID-19, the government had issued some policies in the past months, such as “Large Gathering Ban,” “Restaurant Limits,” and “Face Covering Requirement.” We wanted to use these variables to analyze the relationship between the confirmed cases and the different government policies, and we compared the difference between the proportion of the confirmed cases in October 2020 and the proportion of the confirmed cases in June 2020. First, we will try to generate a graphic visualization to see the proportion difference in June 2020 and in Oct 2020 for all the states under different large gathering ban policies and different restaurants’ limits policies. Secondly, we were interested in the relationship between air quality and confirmed cases. Research shows that increased rates of COVID-19 in areas with high levels of air pollution. Thus, we would examine the correlation between air quality and confirmed cases by using regression. Lastly, under the current case, face covering is an essential way to slow down the virus. We are going to use a hypothesis test to check if the face mask requirement does make some changes.

Variable Description:

Large Gathering Ban: This variable is about different states’ Gathering Ban statuses. It’s a categorical variable.

Restaurant Limits: This variable is about different states’ Restaurant Limits statuses. It’s a categorical variable.

June 2020 Cases: This variable is the confirmed cases in different states in June. It’s a quantitative variable.

October 2020 Cases: This variable is the confirmed cases in different states in October. It’s a quantitative variable.

AQI 2020: This variable is the average air quality index in different states. It’s a quantitative variable.

Face Covering Requirement: This variable is about different states’ face covering orders. It’s a categorical variable.

States: This variable is about the fifty states in the U.S. It’ s a categorical variable.

Exploratory Data Analysis

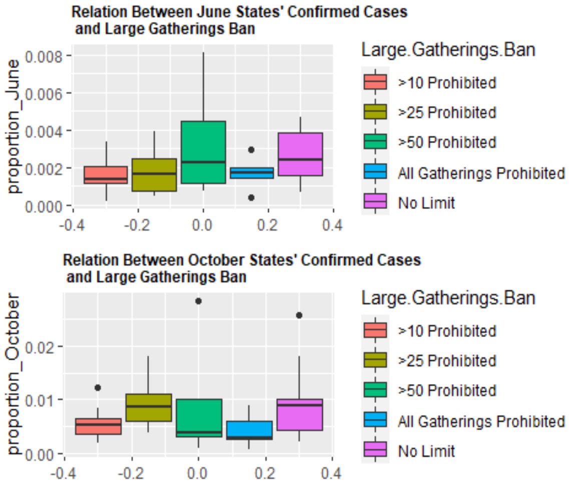

Figure 1 By Yupeng Zhang

The first boxplot shows the number of states' confirmed cases in June with different Large Gathering Ban policies. The second one shows the number of states' confirmed cases in October with different Large Gathering Ban policies.

Additionally, we notice that the categories with >10 prohibited, >50 prohibited, and no limit gatherings ban has extremely high outliers in October compared with no outlier under the same categories in June. The states with all gatherings prohibited gathering policy remains relatively low numbers of confirmed cases from June to October compared with the states with other gathering ban policies. The boxplot of no limit gathering ban policy has a relatively high median in October. In contrast, the medians in the boxplot of June are relatively the same. Moreover, the boxplots of >50 prohibited, and all gatherings prohibited gathering policies in October are skewed to the right. The boxplot of no limit gathering policy in October is skewed to the left.

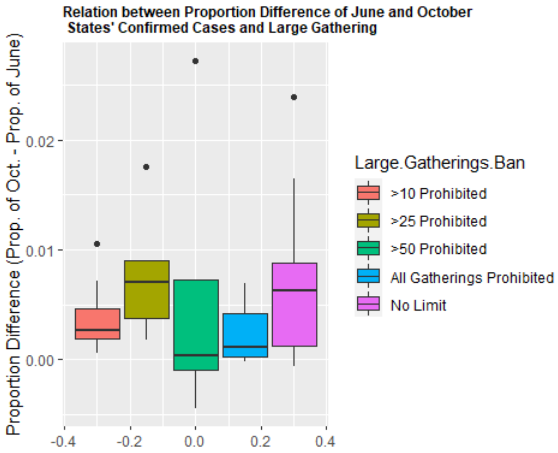

Figure 2 By Yaoying Cai

Summary Statistics

	Median	Mean	SD.
>10 Prohibited	0.002669	0.003661	0.002476
>25 Prohibited	0.007082	0.007430	0.004884
>50 Prohibited	0.000436	0.005913	0.012465
All Gathering Prohibited	0.001121	0.002488	0.002703
No Limit	0.006312	0.006475	0.006154

In this part, the boxplot shows the difference between confirmed cases proportion between June and October 2020, according to different Large Gatherings Ban Policy. Although some states using >50 prohibited gathering ban policy have reduced confirmed cases, the other states under the same category still have an increased number of confirmed cases, especially the outlier in this boxplot.

Moreover, the graph shows that the states with >10 prohibited gathering ban policy, >50 restricted gathering ban policy, and all gatherings prohibited ban policy have relatively lower medians. In contrast, the other two categories have relatively higher medians. Furthermore, the categories of >50 gatherings policy and all gatherings Prohibited policy skew to the right.

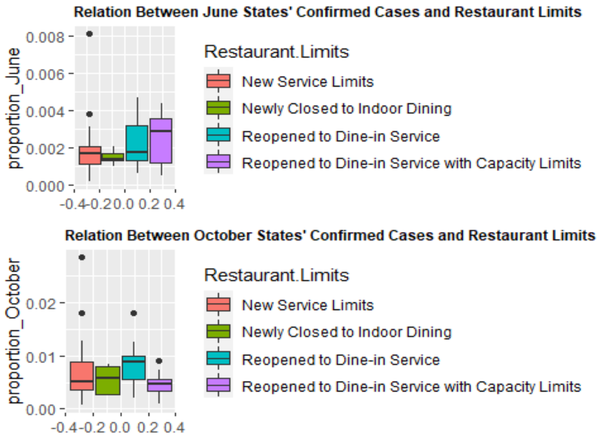

Figure 3 By Mary Guo

The first boxplot shows the June confirmed cases in states with different Restaurant

Limits policies. The second one shows the October confirmed cases in states with varying limits of restaurant policies. We can see that the number of confirmed cases across the U.S. increases in general compared with these two boxplots. Additionally, we notice that only the new service limits category in June has outliers. However, Reopened to Dine-in Service and Reopened to Dine-in Service with Capacity limits categories also have outliers in October. Moreover, the Reopened to Dine-in Service category in June has a similar median like New Service Limits, and Newly Closed to Indoor Dining. In contrast, it has a higher median in October compared with these two. Besides, the distance between the median and 75th percentile of Reopened to Dine-in Service is considerable in June, which means the data skewed to the right. However, the distance between the median and 25th percentile of Reopened to Dine-in Service is considerable in October, which means the data skewed to the left. The shift on the reopened to Dine-in Service is more perceptible than other categories. We know that Dine-in Service is easier for people to get contacted, so the change in the number of confirmed cases might relate to the reopened to Dine-in Service policy.

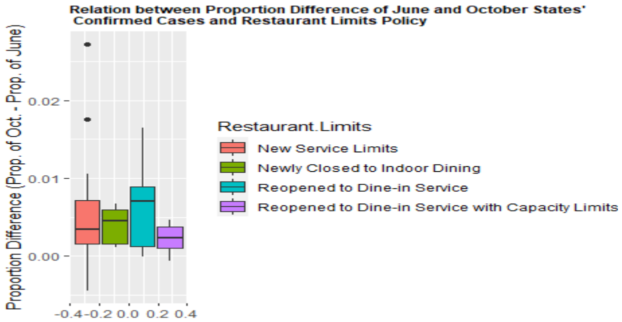

Figure 4 By Winston Mai

Summary Statistics

	Median	Mean	SD
New Service Limits	0.003443	0.005510	0.001448
Newly Closed to Indoor Dining	0.004527	0.003980	0.001010
Reopened to Dine-in Service	0.007082	0.006122	0.001274
Reopened to Dine-in Service with Capacity Limits	0.002318	0.002272	0.000540

This graph shows the proportion difference of confirmed cases across the U.S between June and October according to different state restaurants' limits policy.

The graph and the summary statistic show that the Reopened to Dine-in Service policy category has the highest median of proportion difference of confirmed cases and the largest IQR range. We can also see that all of the 25th percentiles in four Restaurant Limits are above zero. These all imply that most states with Reopened to Dine-in Service had more confirmed covid cases between June and October than other states.

Besides these, only the states with New Service Limits Policy have the outliers in the boxplot. Additionally, the New Service Limits category has the lowest minimum point and the highest standard deviation compared with other restaurant limits policies. So, with New Service Limits, some states have many more people infected by covid -19 while some states have fewer people infected between June and October.

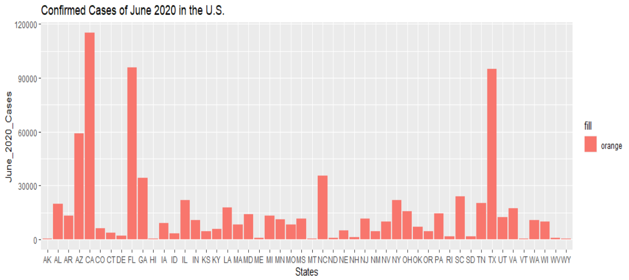

Figure 5 By Caixia Zhang

The above graph shows the confirmed cases of June 2020 in the U.S. One variable labeled "states" is categorical. There are 50 bars in the graph, and each bar stands for one state. Another variable labeled "June_2020_cases" is quantitative, which measures the total amount of confirmed cases of June in all states. As can be seen, the bar of different lengths represents different amounts of confirmed cases. The highest bar in the histogram represents California, which had the largest confirmed cases by almost 120,000, followed by Texas, Florida, and others.

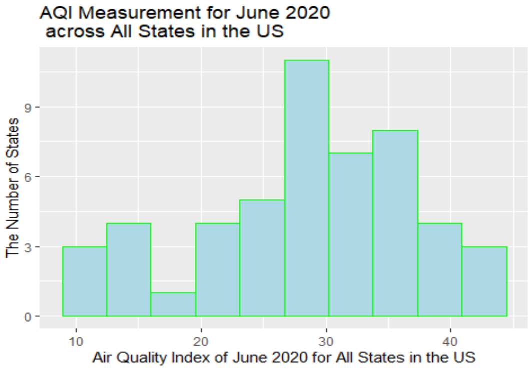

Figure 6 By Jiahui Huang

This histogram shows the distribution of the air quality index of all U.S. states in 2020. They can see that most of the states have an air quality index between 23 and 37. There are 11 states with air quality index between 27 and 31. On the contrary, there is only one state with air quality index between 16 and 19.

Data Analysis and Inference

We tried to see the relationship between different government policies and the increased rates of the virus in the exploratory data analysis. However, when talking about the spread of the virus, one thing that we must investigate is the air quality. "Studies are coming out that are finding increased rates of COVID-19 in areas of high pollution exposure." Here is a quote from Stanford researcher Mary Prunicki in an article published by Stanford Medicine. Research finds that there is a correlation between air pollution and covid-19. Another study published by Harvard University found that people living in high-particulate pollution areas are 8% more likely to die from covid than those living in an area with just one small unit less pollution. From here, we will examine the correlation between air quality and confirmed cases by using regression.

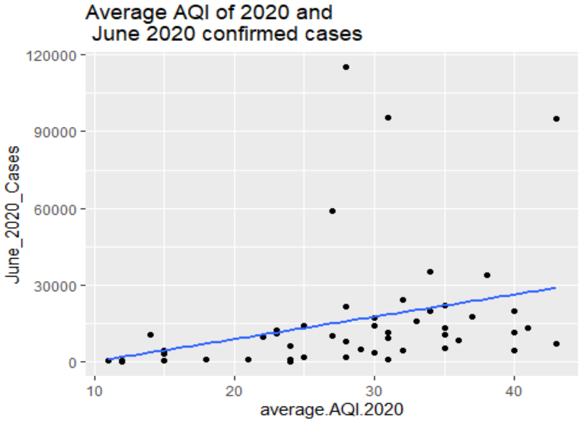

Figure 7 By Winston Mai

The graph above shows the correlation between the average air quality index of 2020 and the total confirmed cases in June for all 50 states in the U.S. Since the average air quality index of 2020 was measured in June, we only compare it to the total confirmed cases in June to get a more accurate correlation. However, as we can see, there are a couple of outliers in the graph, which will heavily affect the regression line by pulling towards them. Since our main data are clustered at the bottom of the graph, we want to omit those outliers and construct a new regression graph.

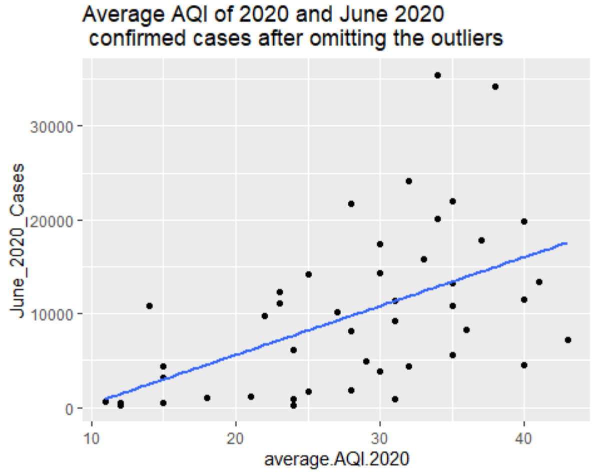

Figure 8 By Winston Mai

This is the regression graph after omitting the outliers. Our initial assumption of the relationship between these two variables is that on average as the air quality index gets bigger, the confirmed cases will be increasing. The greater the air quality index, the worse the air quality in that area. By looking at the graph and the summary statistics, there is a positive slope regression line. The p value for the slope is 0.000282, which means that we reject the hypothesis of the true slope equal to 0. Our estimated slope is 521.5 with standard error of 132.1. The estimated intercept is -4838.8 with a standard error of 3878.9. However, we cannot reject that the true intercept is 0 since the p value for the hypothesis test is 0.2188, which is greater than a normal significance level of 0.05. The correlation coefficient is 0.2615. The value is closer to 0. It indicates that there is a weak association between the air quality and confirmed cases.

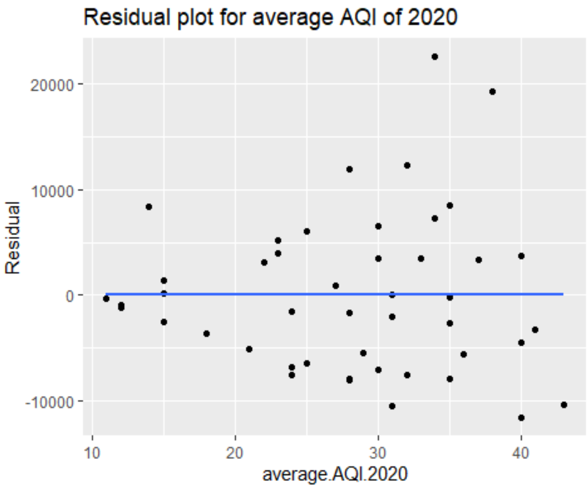

Figure 9 By Winston Mai

The scatterplot does not come out to be “nice” and one of the reasons is probably because we have a small sample size of 50. Therefore, we decided to construct a residual plot to ensure that it is appropriate to use the regression test. The residuals show prediction errors, and we plot the graph with vertical residual for each average air quality index. As the above graph and statistics shown, the regression line for residuals is horizontal at 0 and the residuals average out to 0. The result comes out to be it is valid to use regression to do analysis since the residual plot is not non-linear.

Hypothesis Test

Even though there is only a weak association between air quality and the confirmed cases, it is always good for us to wear a mask and prevent breathing in the pollution and viral components in the air. The most efficient way to stop the virus from spreading is minimizing social activities by staying at home. However, people are not going to follow it simply because you told them to. We learned that from our real-life experience in the past 9 months. People went out of their houses for freedom, to protest and so on. Even with the strict stay home order, it did not implement for a long enough time period to make the virus disappear. In this case, face covering has become the most essential way to slow down the virus. We will use a hypothesis test to check if the face mask requirement is making some changes or not. Our null hypothesis and alternative hypothesis are as follows:

Ho: There is no real difference in the mean increased proportion of the confirmed cases (Oct-Jun) between states that required face covering and those with no action on face covering

H1: The mean increased proportion of confirmed cases of the states with required face covering is lower than the no action states.

Two sample test

We will do a two-sample hypothesis test for the states that implement a face covering requirement with those that did not. Our initiation is that the average increment difference of no action states is higher than the average increment difference of states with requirement. We will assume that we have simple random samples here.

Average difference of proportion of no action states = 0.0129

SD of the difference of proportion of no action states = 0.00776

Average difference of proportion of face covering required states = 0.00478

SD of the difference of proportion of face covering required states = 0.00533

SE of difference = 0.00455

z = ((0.00478-0.0129)-0)/0.00455 = -1.79

p = pnorm(-1.79) = 0.0366

Since our P value is less than the standard significance level of 0.05, we reject the null hypothesis. This indicates that the face covering requirement is making some difference. By looking at the raw data, we can see that there are some states with face covering requirements that have a greater increment compared to all the states with no action on face covering. There are also states with required face covering that have decreasing proportions, which we do not see in the no action states data. However, there might be some inaccuracies since we are working with limited data and a small sample size (only 3 states with no action on face covering requirement). There are many reasons for the increase of confirmed cases, such as policies on restaurant limits, gathering ban, reopening status and so on.

Conclusion

To analyze the relations between different ways of avoiding pandemic spread and the number of confirmed cases in the U.S, we have used the variables, such as large gathering bans orders, restaurant limits policies, and face-covering requirements from the Kaiser Family Foundation. We compared the number of confirmed cases in the states with different gathering bans orders, restaurant limits, and face-covering requirements. In addition, we utilized the Air Quality Index in June 2020 to see if air quality related to Covid -19 spread out. We visually represented these statistics using histograms, bar graphs, boxplots, and scatterplots and saw some interesting exploration ideas to expand and narrow down our study.

First, we decided to draw a histogram to see the number of confirmed cases in different states. Then, we would like to compare states' confirmed cases under different government policies. We took the variables, such as large gathering bans orders and restaurant limits policies with the June and October confirmed cases, to compare the distribution of confirmed cases proportion under different policy categories by drawing two boxplots. The boxplots show that the states with more than ten people Prohibited and All gatherings Prohibited have relatively lower growths of confirmed cases than the states with other gathering bans. For the restaurant limits policies, we found that the states with Reopened to Dine-in Service have relatively higher growth of confirmed cases than other states.

Additionally, we decided to compare the air quality levels and the number of confirmed cases for all the states. We took the Air Quality Index in June 2020 as our explanatory variable and the number of confirmed cases in June as our response variable. By doing a regression test, we saw a positive correlation between air quality and the number of confirmed cases. The positive correlation indicates that as one increases, the other variable increases, on average. However, as the correlation coefficient is small (r = 0.2615), we concluded that this is a weak correlation, not a causation.

Furthermore, we decided to do hypothesis testing on the face-covering policies and the number of confirmed cases to investigate whether there is a real difference between the difference proportion of the confirmed cases (Oct-June) in states that required face covering and those without requirements. To do this, we carried out a one-sided two-sample z-test with the assumption that we have two random sampling due to the limited data. We then concluded that the difference between the difference in proportion of the confirmed cases (Oct-June) in states that required face covering and those without requirements is not due to chance (p-value = 0.0366).

Although there might be other confounding variables that influence this result, it's important to notice that wearing masks is a good way to protect people from the COVID-19, as well as other government policies like restaurant limits and larger gathering bans and natural factors like good air quality.

Works Cited

Air pollution linked with higher COVID-19 death rates. (2020, May 05). Retrieved December 12, 2020, from https://www.hsph.harvard.edu/news/hsph-in-the-news/air-pollution-linked-with-higher-covid-19-death-rates/

Costello, A., Huber, A., MacCormick, A., & Benzkofer, A. (2020, July 22). Why air pollution is linked to severe cases of COVID-19. Retrieved December 12, 2020, from https://scopeblog.stanford.edu/2020/07/17/why-air-pollution-is-linked-to-severe-cases-of-covid-19/

(n.d.). Retrieved December 12, 2020, from https://worldpopulationreview.com/states/state-abbreviations