**Manhattan's Yellow Taxis and COVID-19**

Dongsoo Kim

Through the analyses and testing conducted in this project, it can be concluded that the differences in means of both passengers and distance traveled for Manhattan's Yellow Taxis between days with a high amount of COVID-19 cases and days with a low amount of COVID-19 cases are statistically statistic. The same could be said for days with a high amount of COVID-19 deaths and days with a low amount of COVID-19 deaths.

COVID-19 has affected a large part of nearly everyone's daily lifestyle. The citizens and businesses in New York City are no different, and the city has become an interesting case to study during these times due to its high population density and ease of transportation. When thinking of Manhattan's transportation methods, the iconic Yellow Taxis are what would pop into anyone's mind, and so this project will analyze the relationship between the Yellow Taxis in Manhattan and the severity of COVID-19 in Manhattan. Originally, the hypothesis for this analysis was that a high number of passengers and distance traveled by taxis would result in a high number of COVID-19 cases and deaths. However, this analysis will suggest otherwise, as we delve into Yellow Taxi data and COVID-19 data for Manhattan for the time period of March 22, 2020 to June 30, 2020.

The reason for this specific time period is to focus in on the time period that COVID-19 had a large impact on society in Manhattan. March 22, 2020 is the date that the governor's office issued a stay-at-home order to Manhattan, and June 30, 2020 is the last available date for Yellow Taxi data. All of the data is provided by https://www1.nyc.gov/ which is the official website of New York City. In order to clean the data for any errors inputted in the data and to focus on the relevant time period we are looking at, each month's dataframe's null values and data outside the relevant time period were taken out. The same was done to the COVID-19 data. The data for each month's taxi rides includes the time/date and locations of pickup and drop-off, the number of passengers, the distance of the ride (in miles), all the charges applied to the customer, and miscellaneous ID's that we won't be using. For the COVID-19 data, the number of daily cases, hospitalizations, deaths, and the 7-day averages of those categories are listed for the whole city and each of the boroughs. Since we are looking at Yellow Taxis which are only present in Manhattan, we will only be using the daily Manhattan COVID-19 data which consists of the number of daily cases, hospitalizations, and deaths.

In order to understand what the Yellow Taxi data can tell us, a network analysis of each month's Yellow Taxi rides is conducted. This was set up by using the pickup and drop-off locations of each taxi ride with a trip distance greater than 10 miles. This is done to filter out only the significant taxi rides and avoid data with taxi rides starting and ending in the same location, as it wouldn't be helpful in a network analysis. Directed graphs for each month are created and populated by the nodes (each zone in Manhattan) and the edges (each taxi ride going from one zone/node to another). With this, we can find the density of pickup/drop-off locations, create degree histograms, draw out the network of taxi rides, and list the betweenness centralities. The first item on the dashboard included in this project is the densities, which show what proportion of possible pickup and drop-off combinations are present in the data. As we progress in months, we can see that more combinations of pickup and drop-off locations are ridden, meaning taxi rides are likely becoming more frequent and more spread out. Next on the dashboard is the selected month's degree histogram, which shows how many zones have an x amount of unique destinations. Similar to the densities, the histograms show data moving more towards the right as we progress in months. This means that more pickup locations have a higher number of destinations and taxi travel is picking up. Right below the histogram is the network drawings for the selected month. Although the drawing may seem very cluttered and unreadable, what we want to focus on is not the inner area of the drawings, but the outer areas, which are fortunately more readable. Zones that are on the outside are zones that generally have less taxi rides to and from other zones, while those on the inside experience a greater number of taxi rides to and from different zones. As we move on from month to month, the number of zones on the outside seem to be decreasing, meaning more zones are experiencing more traffic from other zones. Finally for the network analysis, the betweenness centralities for each month are displayed. Here we see which zones are most frequently traveled by proportion. Looking at the top 5 for each month, JFK airport is consistently at the top for each month, showing the most frequency in travel by Yellow Taxis. Something to point out is that each of the months also contain a "nan" zone. These are zones which locations are not within the standard locations that these Yellow Taxis operate in, and can be seen as a miscellaneous collection of locations. As we move on in months, the proportion of rides involving JFK Airport dwindles while other locations see more taxi rides, meaning there is an increase in Yellow Taxi travel within Manhattan.

This network analysis gave quite a few insights into the data we are looking at and what we should expect for the next analysis we do. As the months go on, travel by Yellow Taxis seems to become more widespread despite the ongoing outbreaks in New York City. What might be the reason for this? In order to find out more, we will analyze the correlations of Manhattan's Yellow Taxi data and Manhattan's COVID-19 data.

In order to create a dataframe to conduct this analysis, we will concatenate and merge all of the data that we have into one dataframe. First, we concatenate, or append the rows of, each month's Yellow Taxi rides to each other, and group them by the pickup date, summing all the values so that the data consists of each day's total passengers, trip distance, fare amount, etc. Next, we will merge this concatenated dataframe with the COVID-19 data by the same dates. This way, the total COVID-19 cases, hospitalized, and deaths correspond to the correct day for the taxi data. This new merged dataframe is available as output.csv for further analysis.

With the merging complete, we can now analyze the data properly. A heatmap can be made, displayed below the network analysis, showing the correlation coefficients for each variable against another. These numbers range from 1 to -1, where numbers closer to 1 mean the variables have a stronger positive correlation, while numbers closer to -1 mean the variables have a stronger negative correlation. Surprisingly, we can see that all of the taxi data is negatively correlated to the COVID-19 data. This means that, for example, days with a high number of passengers have a low number of COVID-19 cases, and vice versa. As the variables for taxi data are all very strongly, positively correlated to each other, we will just choose variables that make more sense to analyze by our own discretion. The same goes for the COVID-19 data. For this analysis, we will be using the following variables: passenger\_count, trip\_distance, MN\_CASE\_COUNT, and MN\_DEATH\_COUNT, as these variables were involved in the original hypothesis.

By plotting scatter plots (below the heatmap) for each combination of the four variables, we can see that the shape of the data points is fairly distinguishable and that there is a clear pattern to this data. In order to test if the relationship between these variables is statistically significant, we will conduct a few hypothesis tests. For our null hypotheses, our original hypothesis may not work out well. Instead, it seems like a lower amount of cases or deaths may cause people to feel safer and travel more. It may also be that government actions to discourage travel while cases or deaths are high causes people to travel less by taxi. Either way, the null hypothesis will be that the population means of the number of [passengers/miles traveled] between days with [>=100 cases/>=10 deaths] and days with [<100 cases/<10 deaths] are the same. The alternative hypothesis would then be that the population means are not the same and that there is statistically significant evidence to conclude this. The alpha level will be 0.05, and as long as the calculated p-value is smaller than this alpha level, we can reject the null hypothesis and conclude that the difference in population means is statistically significant.

After carrying out these hypothesis tests, all four rejected the null hypothesis, and therefore we can conclude that the differences in means of both passengers and distance traveled for Manhattan's Yellow Taxis between days with a high amount of COVID-19 cases and days with a low amount of COVID-19 cases are statistically statistic. It turns out that the new hypothesis made in the earlier paragraph may be true, and that these correlations exist because people either fear the virus less when there are less cases/deaths, or that the government intervenes enough to restrict or discourage travel when there are higher cases/deaths. Thus, a high amount of taxi rides does not correlate positively with the number of cases/deaths from COVID-19, as there is instead a negative correlation. Therefore, we may not need to worry as much about the Yellow Taxis being a dangerous avenue for COVID-19 to spread, as the reasons above likely caused people to travel by Yellow Taxis in Manhattan less when there were more cases from March 22, 2020 to June 30, 2020.

Link to the Repository: <https://github.com/DongsooDKim/nyc-taxis-covid-19>

Link to the Dashboard: <https://ma346-dkim-taxis1.herokuapp.com/>