Is There a Direct Relationship Between How Far the COVID-19 Pandemic Progresses and Taxi Demand in NYC?

Project 1 for Applied Data Science

Andre Media
Student ID: 980155

August 18, 2021

1 Introduction

The following report will test the significance of the duration of the pandemic on the taxi counts in NYC to see whether there has been an increase or decrease to taxi counts for the same daily data but for different weeks since the COVID-19 pandemic started. This report is targeted towards current and potential taxi drivers as to whether, as the pandemic increases even further, it would be beneficial to find a different job or start getting back in the taxi due to an expected increase or decrease in demand for the months to come.

1.1 Abstract and Timeline

The datasets will be cleaned by prepossessing algorithms and then converted to a form similar to what is mentioned below. Correlation methods will pick out a few significant statistics and transformations. Finally a linear statistical model will be created which can model the taxi count over each day of the year (DOY) (for 2020) for a specific:

- Part of the day (POD)
- Borough (one of the five boroughs in NYC)
- For either taxis (green or yellow NYC taxi cabs) or 'For-Higher Vehicle' (FHV) (examples include Uber and Lyft)

Given the following data related to the DOY

- The day of the week (DOW) (Monday - Sunday)
- The taxi count of the respective day in 2019
- A few COVID-19 related statistics
- The week of the year (WOY)

The WOY parameter will then be tested for significance using ANOVA.

1.2 Datasets

1.2.1 NYC TLC dataset

This first dataset comes from the New York City Taxi & Limousine Commission (NYC TLC) [2] and consists of all the trip records for any taxi or FHV over the last decade. Each trip record has the time and place the trip commenced and terminate. There are also some other attributes included per record; for example the fare amount and trip distance for the taxi specific records. 22 months of data will be used, with ∼30 million trips per month (More of the size is mentioned later).

The pickup time will be used for categorization, with the drop-off time being ignored as it is very close to the pickup-time. This pickup time will be binned into groups including the DOW, WOY, and POD. The different POD groups will be the following:

- Night: 12am - 5:59am
- Morning: 6am - 11:59am
- Afternoon: 12pm - 5:59pm
- Evening: 6pm - 11:59pm

The pickup and drop-off locations will be grouped based on borough by use of a 'Taxi Zone Lookup Table' also provided by the NYC TLC [3].

1.2.2 COVID-19 Case Counts

This dataset contains data from the COVID-19 pandemic in NYC [5]. It consists of a range of different counts such as the daily new cases and new hospitalizations during the pandemic. Along with these counts is the 7 day averages for each. Furthermore, for each of these, the counts are repeated for each of the boroughs and the combined total. Resulting in ∼50 different counts over ∼500 days

1.2.3 Pandemic Restrictions in NYC

This is the final dataset used and was generated by looking at several news articles ([6] - [11]) relating to the pandemic to create a running list of which restrictions were occurring on which days of the pandemic. There are ∼20 different categories which each have their own categorical variable responding to whether the activity was restricted or not for each of the ∼500 days. These categories include restrictions on different levels of NYC school (E.g. elementary) to outdoor dining.

1.3 Other Key Assumptions

Major assumptions being made by this model are:

- The taxi count can somewhat be modeled by the other parameters
- There are no other confounding factors unaccounted for (E.g. seasonal weather patterns are accommodated by the 2019 data)
- The only factor that will give the model a sense of time is the WOY parameter, therefore without this parameter, the model will be predicting the taxi count without consideration of how long the pandemic has been occurring for, and only based on the current day's data.

2 Preprocessing

2.1 NYC TLC Dataset

2.1.1 Initial visualization

The plot shown in figure 1 contains plots of a sub sample of the raw data, investigating the trip distance and fare amount of June taxi counts in 2019. In both histograms, there is a small bump on the right tail, this likely comes from the airport trips which are around that distance and fare amount respectively. With the scatter plot there is a clear main cluster with outliers in every direction of it. Clearly some general filtering must occur.

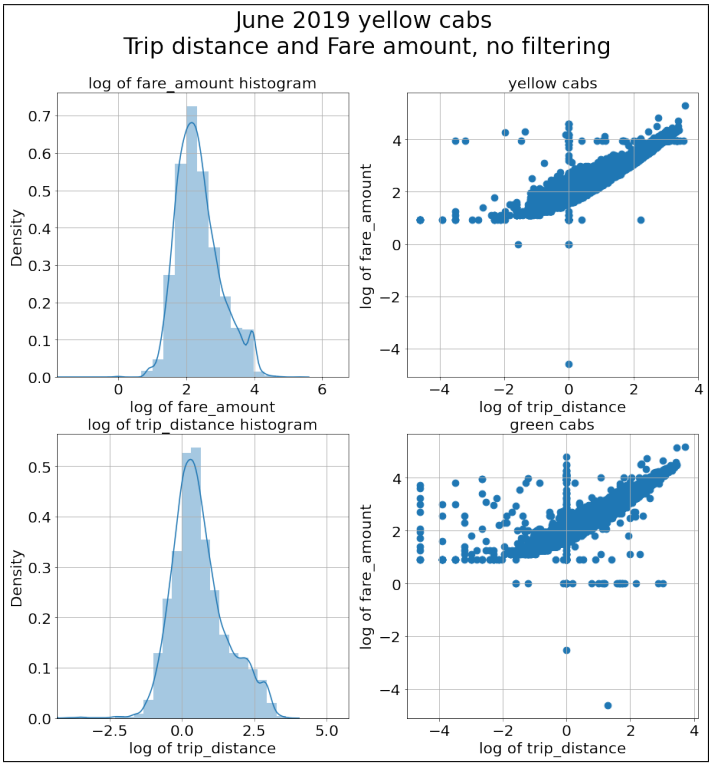

Figure 1: Trip distance and fare amount plots, no filtering

2.1.2 General cleanup

The goal with this clean up is to remove most of the outliers and only be left with 'standard' taxi rides, this is to remove any systemic difference between the 2019 data and 2020 data, such as share rides being banned in 2020. Starting with general cleaning for all datasets, as each trip was grouped by borough, any trip without a borough was removed. This means removing any trip which went to or came from a locations which;

1. Did not have an associated borough (E.g. N/A locations)
2. Went to an airport. As counting them would skew borough counts towards boroughs containing airports.

Specific cleaning for FHV included removing any shared trips for the aforementioned reason.

Specific cleaning for the taxi cab datasets included removing any record where the following was true:

1. Less than 1 person traveled; as an empty taxi is not a standard ride
2. The rate-code was not the standard rate; we are looking at standard taxi rides only (also excluded airports for the aforementioned reason)
3. Payment type was disputes, no charge etc.. (only kept cash and card as standard payments)

After this the data still had outliers, Therefore only the central 99% of trip distance and fare amount were kept. As well as making sure the fare amount was above the minimum for a taxi ride (2.5 USD [1]) and the distance was above 0 miles.

Figure 2 is data after filtering. As expected, the data is a lot more uniform.

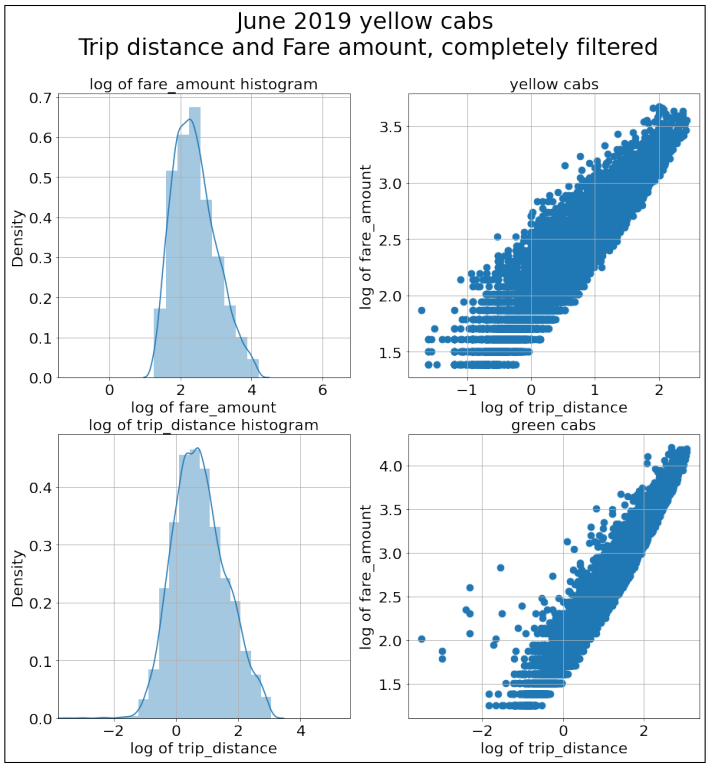

Figure 2: Trip distance and fare amount plots, full filtering

2.2 Change in Size of Datasets from filtering

For 2019 taxis, the datasets went from ∼8 million trips per month (Mtpm) to ∼5 Mtpm. As for FHV, that went from ∼22 Mtpm to ∼14 Mtpm. Both cases, around 40% of the data was removed. Looking at where this removal occurred; Around 20% was removed during the location filtering and the rest was non standard trips. In 2020, around the same percentage was removed, however taxi counts were down ∼90% of the same time the previous year and FHV were down ∼65%. The removal rate is high, However the justification is sound, therefore this smaller dataset will be continued with.

2.3 Tallying data

These taxi trips were counted per bin (mentioned in below) and combined with the restrictions and case counts on a per day of the year (DOY) basis. This dataset had 3400 rows and 55 columns. The rows were made of 10 rows (for each borough and taxi/ FHV combination) repeated for each DOY COVID-19 was relevant in 2020 (around 340 days). As for the columns, they are as follows:

- 5 columns for the binning of the row: DOY, DOW, WOY, borough and FHV flag (1 for FHV, 0 for taxi)
- 8 columns for the taxi counts for each POD in 2019 and 2020
- 20 columns for NYC total and borough specific COVID-19 and case counts of that day
- 20 columns for the restrictions of that day

3 Preliminary analysis

3.1 Correlation functions used

Before moving onto a statistical model, the number of attributes must be decreased sharply. Therefore the correlation needs to be checked between some variables and a linear relationships must be discovered. To test for correlation, three functions will be used.

1. Cramér's V for two categorical variables.
2. Pearson's correlation for two continuous variables.
3. Correlation ratio for one categorical and one continuous variable.

3.2 Testing Correlation and Transformations

3.2.1 Best Average Correlation

Several correlation matrices were created and it was found that:

1. Borough specific COVID-19 counts were the only necessary counts due to high correlation to total counts, and hospitalization had the best representation overall. (View figure 3)
2. Afternoon and night counts for taxis and FHV in Queens were a good representation for most other taxi counts. (View figures 5 and 4)

3.2.2 Log-Log Transformation

Plotting the best averagely correlated taxi counts vs COVID-19 counts gives figure 6, it is clear that a transformation is needed. A log-log transformation gives the data a clear linear relationship seen in figure 7.

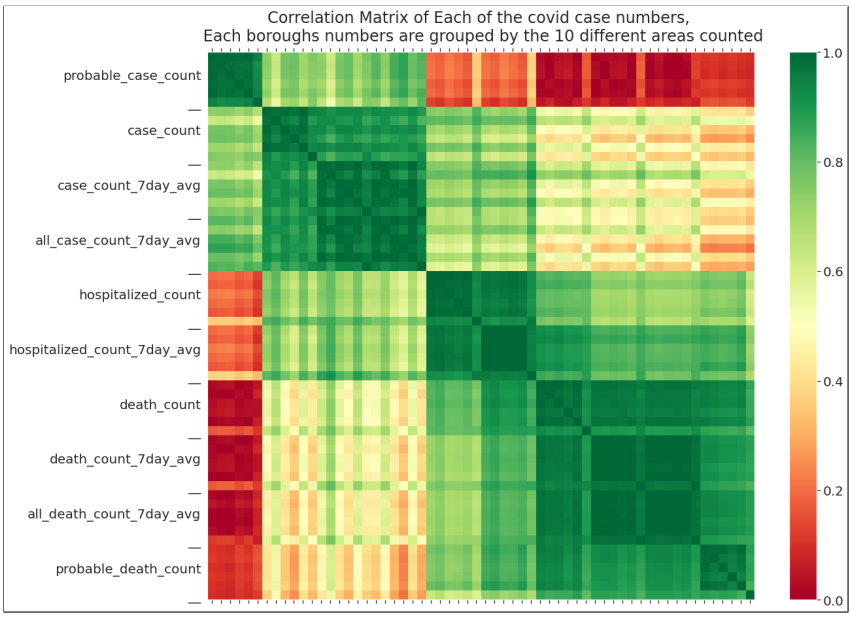

Figure 3: Correlation matrix of different COVID-19 case counts

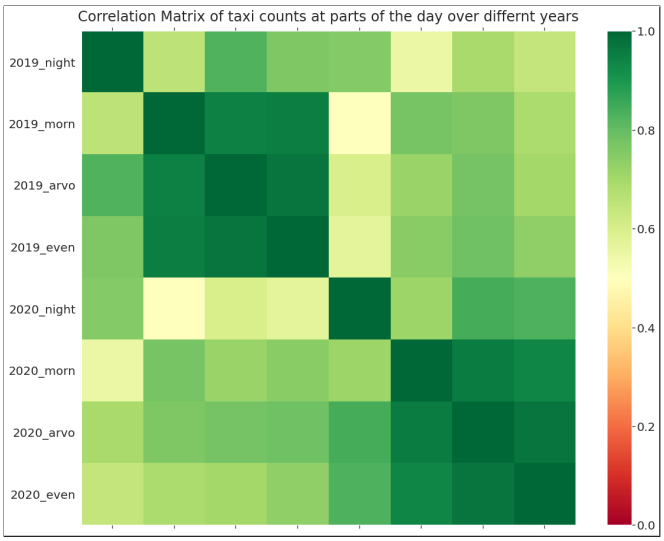

Figure 4: Correlation matrix of different POD

3.2.3 Time Delta of COVID-19 counts

Plotting the COVID-19 counts next to the taxi counts shows evidence of a 'lag' between the cases rising and taxi counts dropping. This is expected as it takes time for news to spread. Testing the correlation of different lag amounts showed an optimal amount of 5 days. This was then rounded to a week as there were weekly patterns in all taxi and COVID-19 counts.

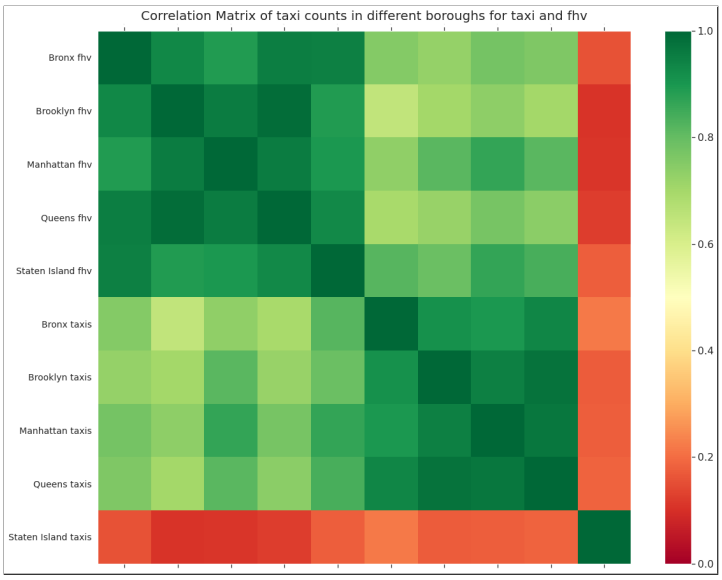

Figure 5: Correlation matrix of different borough and FHV combinations

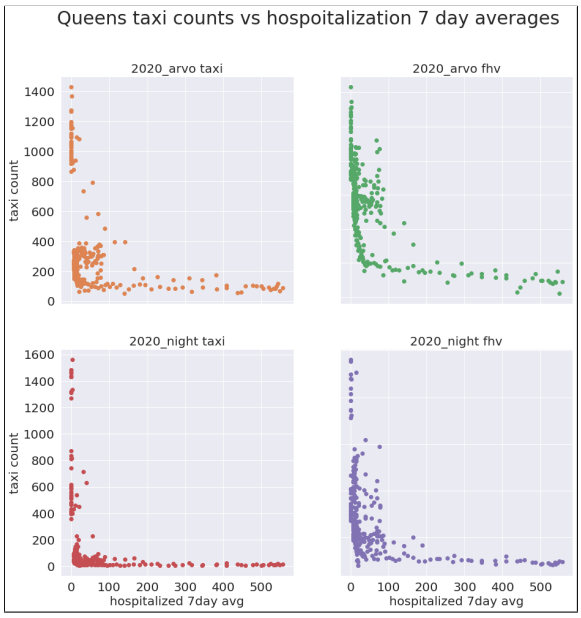

Figure 6: COVID-19 case counts vs taxi counts

3.2.4 Which COVID-19 Restrictions to Use

A correlation matrix of different restrictions to each other (figure 8) shows several restrictions with high correlation. Furthermore looking at the correlation between restrictions and taxi counts, there are no standouts overall. Surprisingly the curfew correlates poorly. A few highly cross correlated restrictions chosen for further analysis were stadium capacity, restaurants, high schools, indoor religious and phase 1.

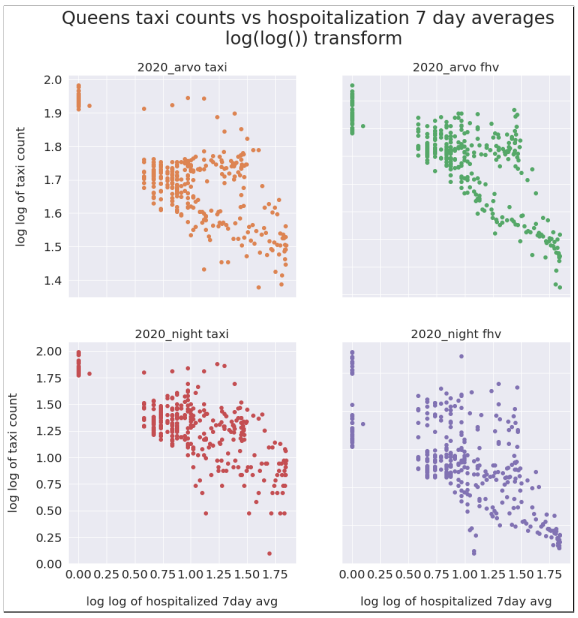

Figure 7: log-log of COVID-19 case counts vs log-log of taxi counts

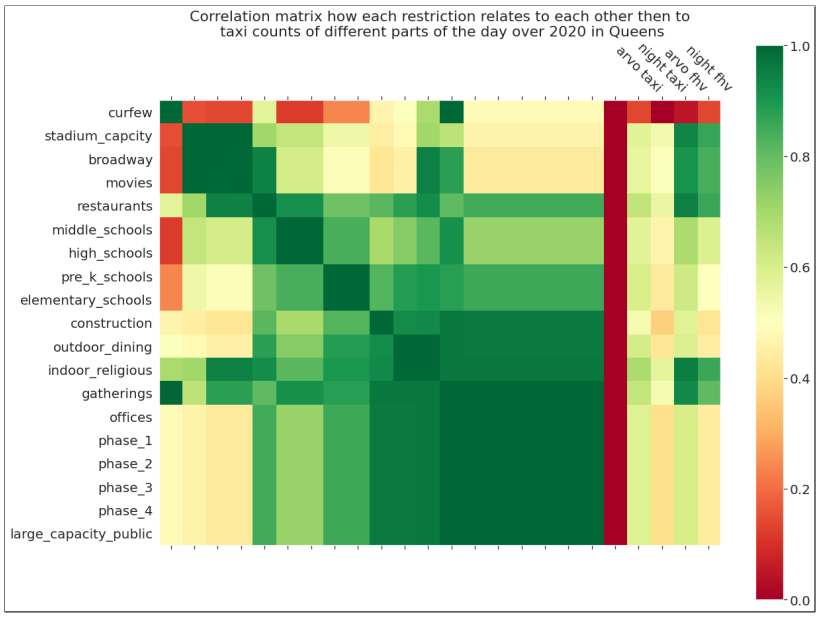

Figure 8: Correlation matrix of restrictions to taxi counts

3.2.5 Which COVID-19 counts to use

Now the time delta and log-log transformations have occurred, we can view correlation between case counts and taxi counts (figure 9). A count plus its 7 day average were taken from each of the 3 large groups with good correlation for further analysis.

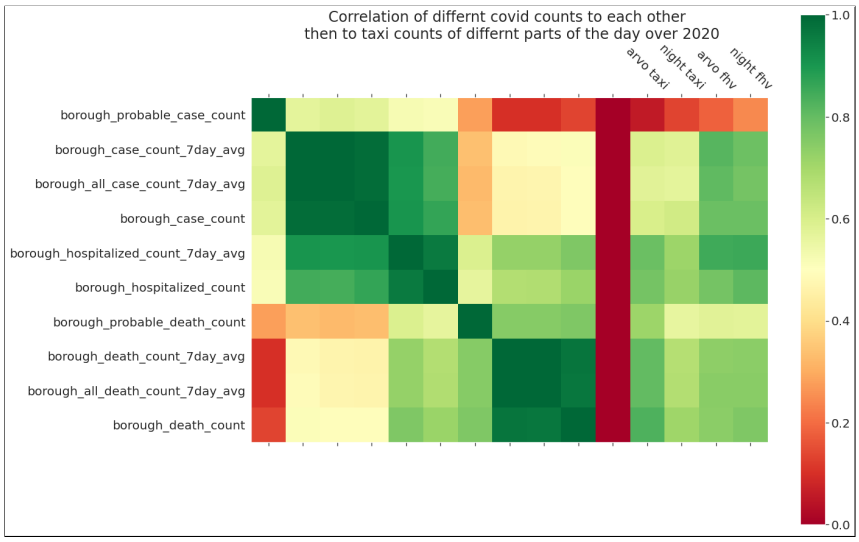

Figure 9: Correlation matrix of adjusted case counts to taxi counts

4 Statistical Modelling

4.1 Initial model with AIC

The amount of features have decreased from ∼50 to ∼15, a significant improvement, however more must be removed to prevent over-fitting. Four models were produced (for each POD) and step-wise AIC was conducted to simplify the models. Via ANOVA, the p-value of each remaining COVID-19 statistics was calculated. For the COVID-19 counts, only two were significant (at 1%) in 2 of the 4 models. The same observations occurred for restrictions. However, these 4 variables were significant enough overall to move forward.

4.2 Final model

Each of the four final linear models (for each POD) will have the following variables:

- Categorical variables with no interaction will be DOW, WOY, and borough
- Categorical variables with interaction to the continuous variables will be the FHV flag, stadium capacity and indoor religious restrictions.
- Continuous variables with interaction are the log-log of each:
  - Corresponding 2019 taxi count
  - 7 day average of the borough death count and borough case counts with a 7 day lag
- The response variable was the log-log of the taxi count for that POD

After fitting the model, a plot was created for the 7 day average count of each model (figure 10). Each model was plotted with and without the WOY variable (referenced as week).

4.3 Results and Discussion

All of the models had an adjusted R^2 value of above 0.97 which indicates a high level of correlation, However the difference in adjusted R^2 between the models contrasting the WOY was only 0.0005 at its worst. This combined with viewing the plots shows there is no major visible difference between the models with and without WOY.

This being said, the P-value of an ANOVA test was less then 10^{-6} in all models which shows high levels of significance. The exponential of the averaged 95% CI for the parameter of WOY was (1.0020, 1.0031). Therefore in general WOY gives a slight exponential increase. Together this means there was an extremely significant and very slight increase to the taxi count as the pandemic progresses.

Figure 10: Final four models plotted over each day of 2020

5 Recommendations

As the evidence points towards there not being a noticeable change to taxi counts as the pandemic continues, the demand will therefore also not increase. This suggest that new taxi drivers will struggle to receive any customers due to the unchanging demand. However, there is also strong evidence that there is a very slight increase in demand with time, therefore it is unlikely that the taxi counts will

drop further as the pandemic continues. Therefore if you are currently a driver and are making a sustainable wage, do not look for a new job as demand will continue to be sustainable.

6 Conclusion

As the NYC TLC data for 2021 had not be released at the time of writing this report, the results are only valid for the start of 2021 which is already out-of-date. Furthermore, vaccines had not been introduced so they were not taken into account in the model. The introduction of vaccines would change the model significantly. Constructing this same model a year or two in the future, after the pandemic has ended, would produce more realistic results.

References

[1] "Taxi Fare." Taxi Fare - TLC. Accessed August 1, 2021.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[2] "Taxi Datasets." Trip Record Data - TLC. Accessed August 1, 2021
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[3] "Taxi lookup table." Taxi Lookup Table - TLC . Accessed August 1, 2021
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[4] "Categorical correlation" Categorical correlation - Towards Data science. Accessed August 11, 2021
https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

[5] "COVID-19 Dataset." COVID-19 case counts - DOHMH. Accessed August 2, 2021
https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3

[6] "COVID-19 Timeline" NYC COVID-19 timeline - Investopedia. Accessed August 2, 2021
https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986

[7] "NYC phase 1-4 restrictions" New York phase reopening - New York times. Accessed August 2, 2021
https://www.nytimes.com/article/new-york-phase-reopening.html

[8] "COVID-19 restrictions" State pause executive order - Governor of NY. Accessed August 2, 2021
https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order

[9] "COVID-19 restaurants restrictions" Restaurants and bars timeline - NY eater . Accessed August 2, 2021
https://ny.eater.com/2020/12/30/22203053/nyc-coronavirus-timeline-restaurants-bars-2020

[10] "COVID-19 new restrictions" New York Shutdown - The guardian. Accessed August 2, 2021
https://www.theguardian.com/us-news/2020/nov/13/new-york-coronavirus-shutdown-cases-deaths

[11] Overleaf Word Count: exactly 2000 words