The Impact of COVID-19 on New York Yellow Taxis in Each Borough

Darrell Chew
Student ID: 1081440

August 16, 2021

Word Count: 2000

1 Introduction

Taxicabs in New York have long been a vital organism in the heart of the city, connecting the city with the five boroughs. Yellow taxis are used to pick up passengers regardless of the borough, whilst green taxis are only allowed to pick up passengers in Brooklyn, Queens, Staten Island, the Bronx, and Upper Manhattan (Pickhardt, 2021).
In March 2020, COVID-19 was declared a global pandemic, impacting urban movement globally (Moore et al., 2020). The objective of this report is to help provide indicators by which the New York mayor office and the Taxi and Limousine Commission (TLC) may use to identify trends present in 2020 through exploratory data analysis. Such relationships can help the TLC plan how to increase taxi pickups in New York.

1.1 Hypothesis

This report aims to describe the impact that COVID-19 had on the demand of 13,000 yellow taxi licensees in each of the boroughs on a daily basis. The hypothesis is that COVID-19 influences the trip distance, passenger count and ultimately taxi pickup count per borough. This is because of the growing inherent risk the virus poses to the general public and hence may limit the amount of urban movement in New York.

1.2 Timeline

To evaluate the effect of COVID-19 on taxicabs, a timeline ranging from March 2020 to December 2020 was chosen as TLC data on 2021 cabs had not yet been released. Yellow taxis were selected as the primary taxi choice to analyse as they have unrestricted passage to all boroughs as compared to green taxis and can therefore portray the true change in ride counts due to changes in COVID-19 cases. Average trip distance and passenger count were used as attributes as a measure of social distancing and risk of exposure due to being in a confined space for long periods of time (Johanson, 2020). To compare the relationship between taxi pickups and COVID-19 cases, a dataset obtained from Statista was used in order to visualise the daily number of COVID-19 positive cases in New York from March 2020 to December 2020 (Elflein, 2021).

1.3 Assumptions

The investigation assumes that the majority of taxi passengers are aware of the COVID-19 situation, especially during the beginning few months of the virus. Furthermore, the hypothesis assumes that the taxi driver remembered to input passenger count and trip distance was calculated for each taxi ride.

2 Preprocessing

2.1 NYC TLC Dataset

As the COVID-19 external dataset only started recording from March 8 2020, the taxi dataset had to be sliced from March 8 onwards (Taxi and Limousine Commission, 2020). Concatenating all the months together, this resulted in a dataframe size of over 10 million instances. However, many instances had invalid passenger counts and trip distances and hence were then removed from the dataset. The passenger count and trip distance were also averaged per borough hence leading to continuous passenger counts, however it is inevitable when taking averages and hence should be considered by the mayor's office. Furthermore, as scanning through the dataset quickly shows that some pickup dates had the wrong month recorded in their respective dataframe (e.g. April was present as a pickup date in the March dataset), these instances were also removed to avoid any additional mistakes that may have been made when recording that specific instance. The same issue was found in the wrong recording of the year (e.g. 2021 was shown rather than 2020). The aforementioned steps were taken as it was unsure how TLC recorded their data. For example, if they added a March pickup in April, it may have been because they had unknowingly omitted that instance previously when publishing the March dataset. Therefore, ignoring potential incorrect instances ensures the integrity of the results.
In the taxi data, there were instances whereby the recordings did not comply with the data recording notes presented in the user guide. Such examples are given in Table 1 below. Invalid instances were then removed from the dataset.

Table 1: Dataset Cleaning and Their Reasons

Incorrect Recording	**Reason**
Negative Passenger Count	Assumed to be a mistake
Negative Trip Distance	May occur (refunds) but is irrelevant for the objective
Pickup LocationID outside of specified range (1-263)	Assumed to be strictly New York zones

2.2 COVID-19 Dataset

As the COVID-19 dataset used was being updated on a daily basis and the NYC TLC dataset had not released the recorded taxi data for 2021, the COVID-19 dataset had to be sliced into the days before 2021 to comply with the taxi dataset availability (Elflein, 2021).

3 Preliminary Analysis

As aforementioned, large outliers were found in trip distance and passenger count. Trip distance was shown to have a average distance of 4.347 miles with a standard deviation of 483 miles. This may have been skewed towards the higher values due to recording of a maximum distance of 350914.980

miles, suggesting that the trip went the same distance as travelling around the circumference of the Earth 1.5 times (which is highly doubtful). Such outliers were then removed in preprocessing.

3.1 Geospatial Visualisation

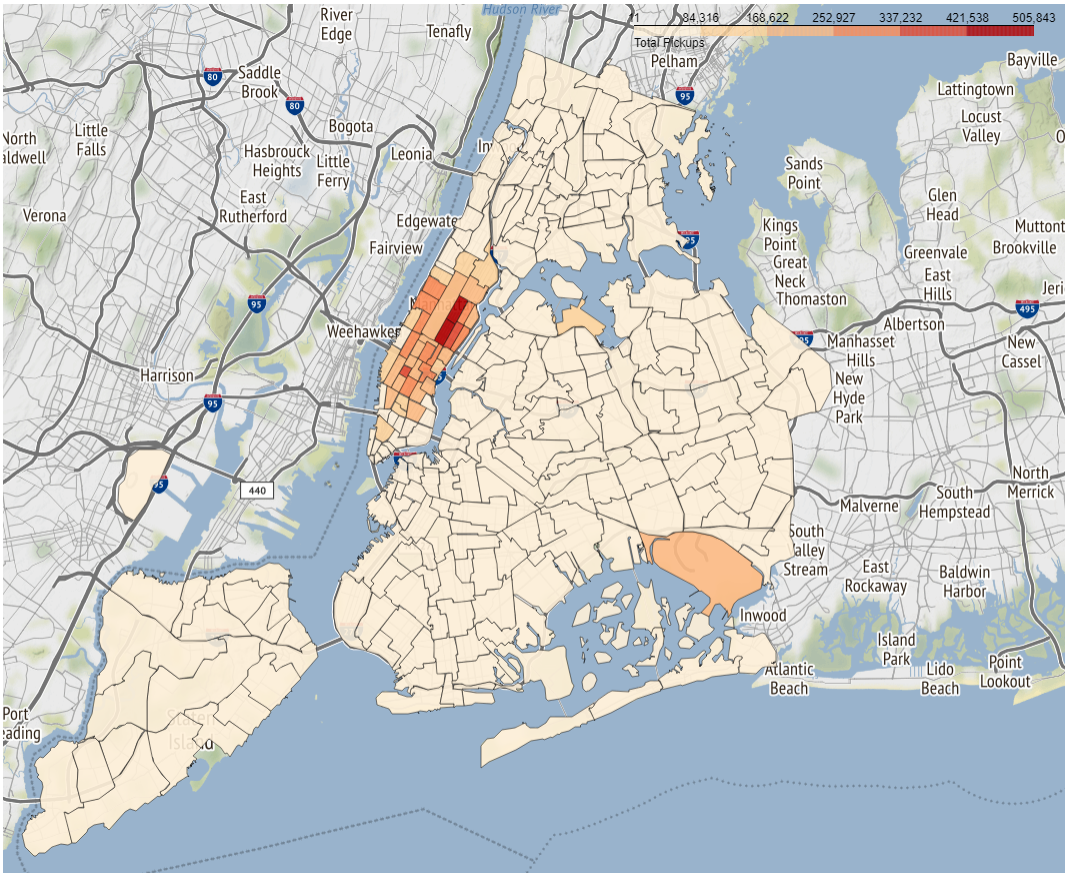

Figure 1: Total Pickups per Taxi Zone

As shown in Figure 1 above, whilst the choropleth map may not demonstrate a continuous change, it provides a rough understanding of the distribution of taxi pickups in New York during the COVID-19 pandemic. As demonstrated, there were a large number of pickups from the major airport zones, such as LaGuardia and JFK International Airport, with around 100 thousand pickups and 300 thousand pickups respectively. This may suggest why upon looking at the total rides per borough in Figure 2 below, the Queens borough is consistently second highest in number of daily pickups.

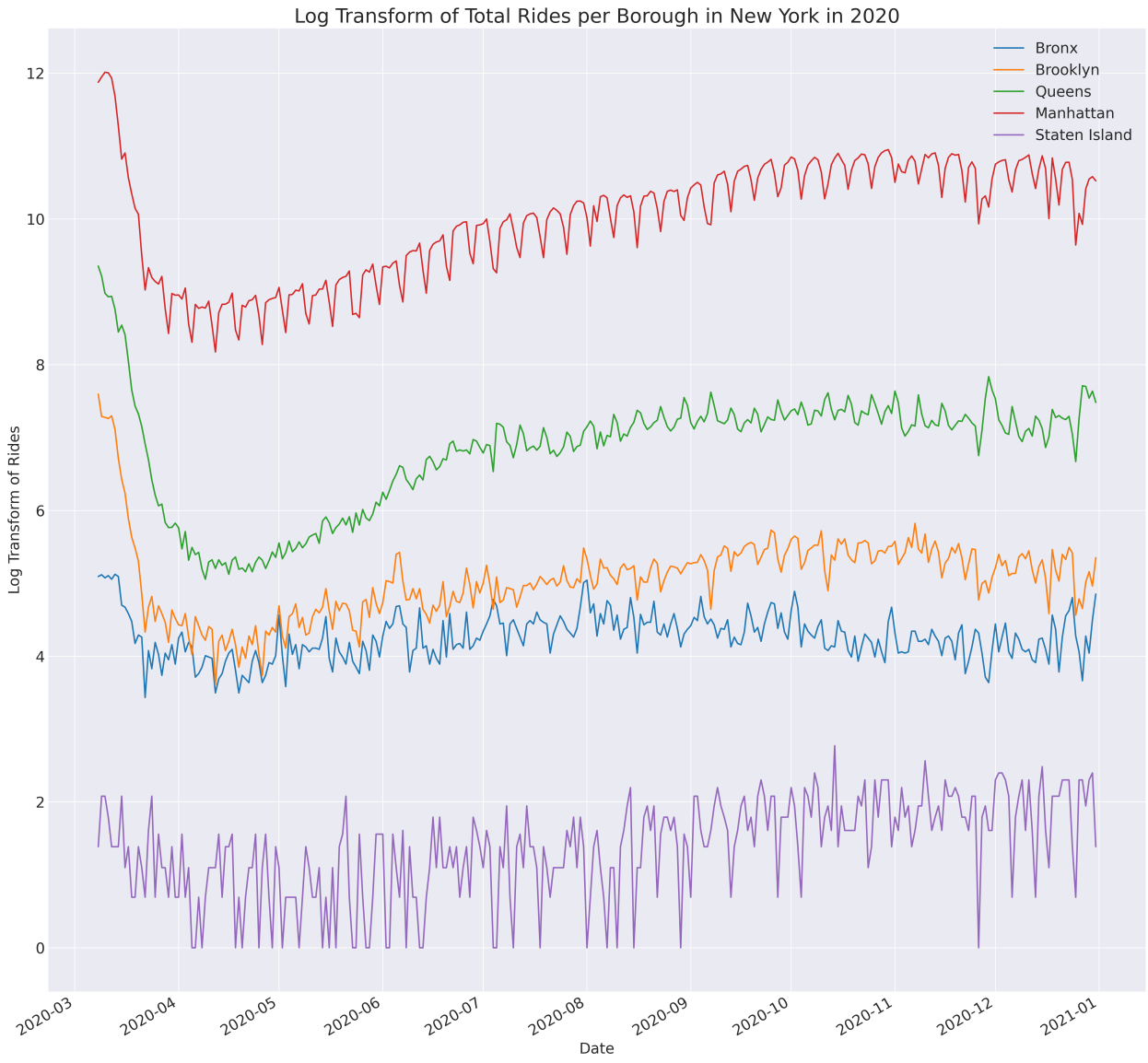

Figure 2: Log Transform of Total Rides per Borough

Figure 2 demonstrates the log transform of the total rides that were recorded for each borough. Log transformation was used for plotting as the Manhattan total pickups were consistently and significantly higher than the other boroughs as shown in Figure 1, hence transformation brings the data down into an easier scale to analyse. The trend observed with all the boroughs except Staten Island is that the total daily pickups per borough drops significantly at the end of March. Such a trend seems to correlate well with the hypothesis of rising COVID-19 cases causing a drop in taxi pickups that will be discussed further.

As aforementioned, Manhattan had the largest count of daily pickups. This may be attributed to the 1.5 million commuters who work in Manhattan during the day yet reside elsewhere during the night (Moss & Qing, 2012). With a significant working population in Manhattan, the borough holds the highest ratio of change between day and night populations (1.92), suggesting that during the day, the working population almost doubles the night population (Moss & Qing, 2012).

3.2 COVID-19 Cases Analysis

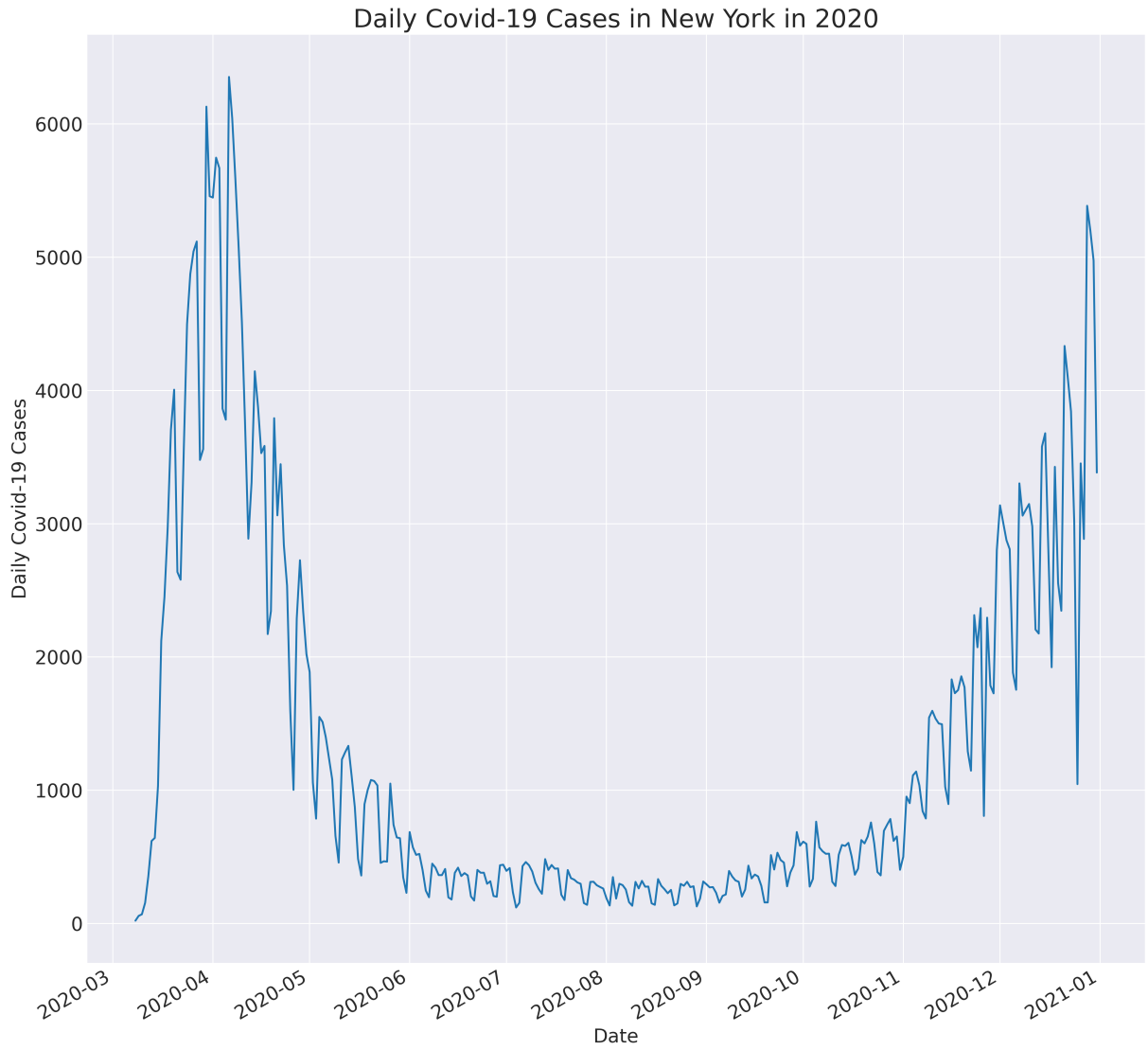

Figure 3: Daily COVID Cases in New York

In Figure 3, there was a large spike in COVID-19 cases between March and April 2020, rising from 50 cases to above 6000 daily cases. Due to the sudden increase, Governor Cuomo issued a stay-at-home order that explains the drop in cases after May to around 500 daily cases (Evelyn, 2020). In June, the order was removed, and its effect on taxis is shown most significantly in the Queens daily pickup rate in Figure 2, rising with a steeper slope than other boroughs. The cases then rose steeply in the later months of the year.

Whilst decreasing COVID-19 cases seemed to encourage urban movement in New York, Staten Island remains to witness the least daily taxi pickup count. This may be due to the already low numbers of recorded pickups in Staten Island, ranging from 1 to 16 daily pickups with a mean of 4. Such a small range would not lead to a noticeable change hence may require further analysis to why there are a low number of rides.

Focusing primarily on Manhattan, it seems COVID-19 may not impact the taxi rates as significantly as

the other boroughs. Whilst there was a steep drop in March, pickup counts regained by 50% suggesting that the Manhattan demographic tend to have more urban movement than other boroughs.

3.3 Trip Distance and Passenger Count

Though daily pickup counts seem to correlate with COVID-19 data, the average trip distance and passenger count per borough tend to be more stable. Staten Island had the highest average trip distance with Manhattan having the lowest. This seems to be in accordance to the aforementioned demographic of Manhattan, where more people seem to travel into Manhattan for work whilst living in other parts of the city, leading to longer trip distance as compared to other boroughs who are closer towards Manhattan.
Average daily passenger count was consistently lowest in Staten Island whilst the other 4 boroughs fluctuated between 1 and 2 passengers. However, Staten Island also had the largest average with passenger count with 6.

4 Statistical Modelling

To fit a model to determine the relationship between COVID-19 cases, total rides, passenger count and trip distance, the linear regression method was used to minimise the sum of square differences between predicted and observed values. Backwards elimination was used to identify which interaction effects were significant (p-value < 0) in predicting total ride count and hence led to the final model for each individual borough. Furthermore, each attribute was standardised in order to bring the large value attributes (COVID-19) into scale and avoid any unwanted significant influence that may arise. Hence the coefficients do not represent a true relationship between attributes.

4.1 Model

Bronx Rides Model:

$$BronxRides = -0.0247 - 0.3512Covid - 0.1726Distance + 0.1227Covid * Distance \tag{1}$$

Brooklyn Rides Model:

$$BrooklynRides = -6.462 \times 10^{-17} - 0.1876Covid \tag{2}$$

Queens Rides Model:

$$QueensRides = -0.0112 - 0.2185Covid + 0.1214Covid * Distance \tag{3}$$

Manhattan Rides Model:

$$ManhattanRides = 1.305 \times 10^{-16} + 0.4758Passengers \tag{4}$$

Staten Island Rides Model:

$$StatenIslandRides = 0.0046 + 0.1257Covid * Distance \tag{5}$$

4.2 Results

As shown in the final models, COVID-19 had a statistically significant impact in most boroughs in relation to determining total ride count. After backwards elimination, the interaction effects were found to be p-value statistically insignificant other than the COVID-19*Distance interaction term. In accordance to the aforementioned discussion with COVID-19's impact on Manhattan, it seemed as though Manhattan total rides were not dependent on COVID-19 cases and were more dependent on the average number of passengers. Bronx total rides were dependent on COVID-19 cases and average distance whilst Brooklyn was dependent solely on COVID-19 cases. Queens did not have a statistically significant effect from distance as a sole attribute however the interaction term between COVID-19 and distance was significant, suggesting that COVID-19 may affect the average distance taken for pickups in Queens.

4.3 Discussion

The exploratory data analysis taken in this project showed the diversity in New York's population. Whilst Manhattan demonstrates the necessity of a large working population, the Bronx shows its high vulnerability to COVID-19 cases. As a result, for each borough, different approaches to increase taxi pickup rates apply.
The Staten Island model relies only on the interaction term between COVID-19 cases and distance. This may have been influenced from the low pickup count range and sporadic passenger count which prevents attributes from determining the outcome. Therefore, the Staten Island office may not conclude much from the final model shown.
Whilst the models demonstrate a relationship between total rides and COVID-19 cases, it is worth noting that the R^2 values were often low, with a maximum of 0.226 in the Manhattan model. As R^2 explains the variability of the model on the response data, this may suggest that the sample size of 299 days may be too small.

5 Recommendations

As the primary aim of the report was to identify the relationship between COVID-19 cases and pickup count in each borough of New York, a possible extension of the analysis would be to obtain a dataset that specifically records the daily cases for each borough individually. This would allow for better understanding of how positive cases may affect each borough as the population demographic may differ significantly and hence result in a change in the number of people willing to take yellow taxis. The TLC can use such information to increase taxi pickups once COVID-19 cases start dropping by placing more taxis in the affected boroughs.
Further recommendations would be to investigate what factors result in low daily pickups on Staten Island, such as increasing sample size once more COVID-19 data is available. This would help confirm whether the interaction term in the final model is accurate or not.

6 Conclusion

The COVID-19 pandemic has affected many countries both socially and economically. This report aims to identify its impact on the economy by measuring the effect on urban movement in New York. The models appear to agree with the initial hypothesis of COVID-19 having a large impact on taxi rides. Whilst Manhattan proved the most resilient borough that has not been significantly impacted by the coronavirus, the other 4 boroughs were shown to react in line with rising cases. Such information

can help the mayor's office and TLC to analyse how taxi counts may rely on COVID-19 cases and lockdowns, increasing understanding of how economic recovery may be achieved.

References

[1] Anthes, E. (2021, January 17). How to (Literally) drive the coronavirus away. The New York Times - Breaking News, US News, World News and Videos. https://www.nytimes.com/2021/01/16/health/coronavirus-transmission-cars.html

[2] Elflein, J. (2021, August 4). COVID-19 cases in New York City by date 2021. Statista. https://www.statista.com/statistics/1109711/coronavirus-cases-by-date-new-york-city/

[3] Evelyn, K. (2020, March 21). Here's what a 'stay home' order means for New York. the Guardian. https://www.theguardian.com/us-news/2020/mar/20/new-york-90-day-stay-home-order-what-it-means

[4] Johanson, M. (2020, December 3). Why our reliance on cars could start booming. BBCpage. https://www.bbc.com/worklife/article/20201202-why-our-reliance-on-cars-could-start-booming

[5] Moore, S. A., Faulkner, G., Rhodes, R. E., Brussoni, M., Chulak-Bozzer, T., Ferguson, L. J., Mitra, R., O'Reilly, N., Spence, J. C., Vanderloo, L. M., & Tremblay, M. S. (2020). Impact of the COVID-19 virus outbreak on movement and play behaviours of Canadian children and youth: A national survey. International Journal of Behavioral Nutrition and Physical Activity. https://doi.org/10.21203/rs.3.rs-34730/v1

[6] Moss, M., & Qing, C. (2012, March 1). The dynamic population of Manhattan. NYU Wagner. https://wagner.nyu.edu/impact/research/publications/dynamic-population-manhattan

[7] Pickhardt, S. (2021, June 11). How to get a taxi in NYC. Free Tours by Foot. https://freetoursbyfoot.com/how-to-get-a-taxi-in-nyc/

[8] Taxi and Limousine Commission. (2020). TLC trip record data. Welcome to NYC.gov — City of New York. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page