MAST30034 Assignment 1 - Report

New York City Yellow Taxi Data Analysis

Andrew Tjen, 939206

Introduction

With the ongoing Covid-19 pandemic causing travel restrictions, it becomes ever so difficult for the taxi industry, specifically its drivers, to generate revenue. This report aims to analyze the factors which are likely affecting the profitability of yellow taxi drivers of the state of New York in this pandemic season from *New York Taxi and Limousine Trip Record Dataset (TLC),* which is openly provided by the New York City government. We will discuss the methodology and data preparation, along with analysis of the visualization, and finally, conclude our analysis with the effect of factors and what drivers can expect with it.

Period and Attribute Selection

To start, we will set the case in the period of COVID-19 outbreaks. The first reported COVID-19 case in New York City was on March 1 of 2020. Therefore, we will set our analysis case period from January of 2020 where there is no effect of COVID-19, to the end of June of 2020 where COVID-19 has become common to the society. We set this period with the purpose to observe the effect of COVID-19 on the attributes we want to analyze on.

The dataset we are working with provides us with several attributes that can be used to represent profitability. In this report, we will be using the rate of tips (in minutes). This attribute will be generated using the tip amount and duration of each trip, will be explained deeper in the data cleaning and pre-processing section. Justification of this attribute would be due to the specific stakeholder we are targeting, which is taxi drivers. After some research, contract agreements such as percentage profit, tolls, fuel, etc between drivers and companies are relative. And also due to complication of some extra charges are not included in the dataset, attributes relating to fare amount is quite vague. Since America has a strong sense of gratuity and tips are fully owned by the drivers, it is a valid choice to analyze the rate of tips for our analysis.

We will analyze the rate of tips with other attributes, mainly with the area of pickup locations, the number of COVID-19 cases, and different weather conditions (between rain/snow and normal weather). 2016 data onwards has a change in data collection as they do not provide specific longitude and latitude to protect privacy. Because of this, we are limited to taxi zone id to visualize the area that yields a higher tip rate. We will use the number of COVID-19 and precipitation levels to see how external conditions affect our rate of tips. We hypothesize that COVID-19 will impact negatively to our dataset due to travel restrictions being in place. While we hypothesize that precipitation level will increase our yield of tips as people tend to appreciate taxi rides more in rain or snow weather. We will also briefly analyze trip counts and duration per area to see which areas are more popular to tailor recommendations for taxi drivers.

Data Cleansing and Preprocessing

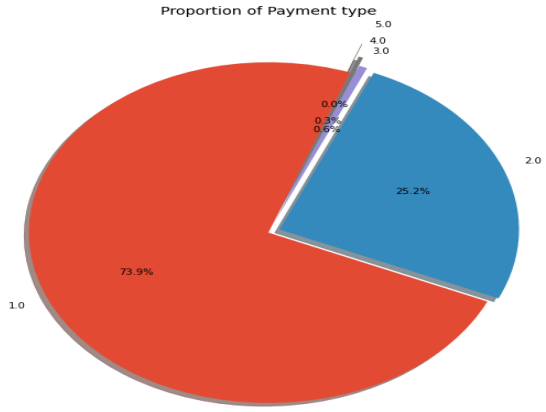
All data sources and references will be included in the reference section.

The first stage is to prepare the taxi dataset. We will download data for each month from January of 2020 to June of 2020 specific to Yellow Taxi from New York TLC dataset. In total, we are dealing with 16,499,407 of data. The sheer size of the data makes it prone to outliers, input errors, and may bias in our result. Thus, we need to set a pre-defined rule to validate our data.

The first step is removing that the data doesn't have a positive duration, meaning that the data must have a pick-up time smaller than the drop off time. We first generate a new attribute duration where it is the difference in the drop off time and pick up time and convert it into seconds using DateTime conversion in python and then multiplying it by 60 to be in minutes measurement. Then we query the data frame so that only a positive duration is what is left from the dataset.

The next step is to generate our profitability attribute. For every record in the data, we generate a new attribute called "tip_per_minute" where the tip amount attributes from the dataset is divided with the duration attribute. The justification of this new feature is so that we will have more normalized distributed data to analyze instead of having an infinite range of possible outlier values with the only tip_amount. An example would be having a tip of $20,000 on a certain pickup area would skew the data.

Since the dataset records only tip amount from credit card payments, we ensured that when analyzing the rate of tips, we took only data that is paid with a credit card. This is to prevent outliers from having many 0 tip rates since the TLC dataset only records tips that are paid with credit cards.



1.0 = Credit Card Payment
2.0 = Cash
3.0 = No Charge
4.0 = Dispute
5.0 = Unknown
6.0 = Voided Trips

Figure 1 – Proportion of Payment Type in 2020

Finally, we ensure that the pickup time is from January to June of 2020 and concatenate all data into one data frame.

Then, we need to prepare the dataset for the weather. We will use precipitation levels from the dataset of wunderground.com. The precipitation level from wunderground.com represents how deep (in inches) excess of precipitation that day. Meaning that it can be snow or rain weather. Due to the limits in the knowledge of weather, we will analyze taxi tip rates in rain/snow or normal condition. We first copy the date and precipitation level from January to June of 2020 to our excel. We then convert it into CSV format with ";" separation and load it into our Virtual Machine

framework. Since excel has a different DateTime format, we need to convert the DateTime format to fit python's format. We will consider rain/snow weather if precipitation depth is above 0.1 inches.

Finally, we need to prepare the dataset of COVID-19 cases. The dataset link in the reference below. The dataset starts on March 1 of 2020. Thus, we need to self-sample data to have a range of data starting from January 1 of 2020. We will sample data from January 1 until February 29 with 0 COVID-19 and death cases. We will then generate a new attribute where we have the number of new COVID-19 cases per day to represent COVID-19. This is a better measurement because daily changes of COVID-19 are more volatile and we can observe how daily changes can affect the profitability of drivers. It's also due to social bias where we judge the state of COVID-19 by the number of new cases. We generate this attribute by subtracting the number of total cases today with the previous day. Then we ensure that the dataset is within the first 6 months of 2020.

Finally, we implement a Left Join method where we merge the weather and COVID-19 dataset to the taxi dataset by the pickup date. We will then implement a group by methods and data frame querying to produce the necessary plot we want to achieve.

‘

Visualizations

This section will discuss on how we achieve the visualizations and analysis on it. We will first have a general look at our data.

Figure 2 – Trip Counts with and precipitation level and Covid-19 daily across the first half of 2020

The daily trip count of taxis seems to be consistent until March where it significantly dropped. We see the correlation with the COVID-19 case as we see the case spiked in the same month. We do not see the correlation clearly with precipitation level. To be sure, we will plot the correlation table.

	Trip count	Percipitation level	Covid Daily Change
Trip count	1.000000	-0.053157	-0.595026
Percipitation level	-0.053157	1.000000	0.099795
Covid Daily Change	-0.595026	0.099795	1.000000

Figure 3 – Correlation Table

We can see a huge negative correlation of -0.592026 between daily change in COVID-19 and count of trips. This shows how the increase in COVID-19 cases negatively impacts the taxi industry. From this table, we also can see a small negative correlation with precipitation level indicating how we should see different weather conditions to weather conditions to trip counts.

With this in mind, to reduce bias in our geospatial analysis. We need to divide our data into two groups. Period of pre-COVID and period of after COVID. The date dividing this will be March 1, the first reported case date in New York. This can also be useful as New York is entering the phase of reopening. The taxi driver can use this information and can expect what will be the conditions.

Trip Counts and Duration Per Area Relationship

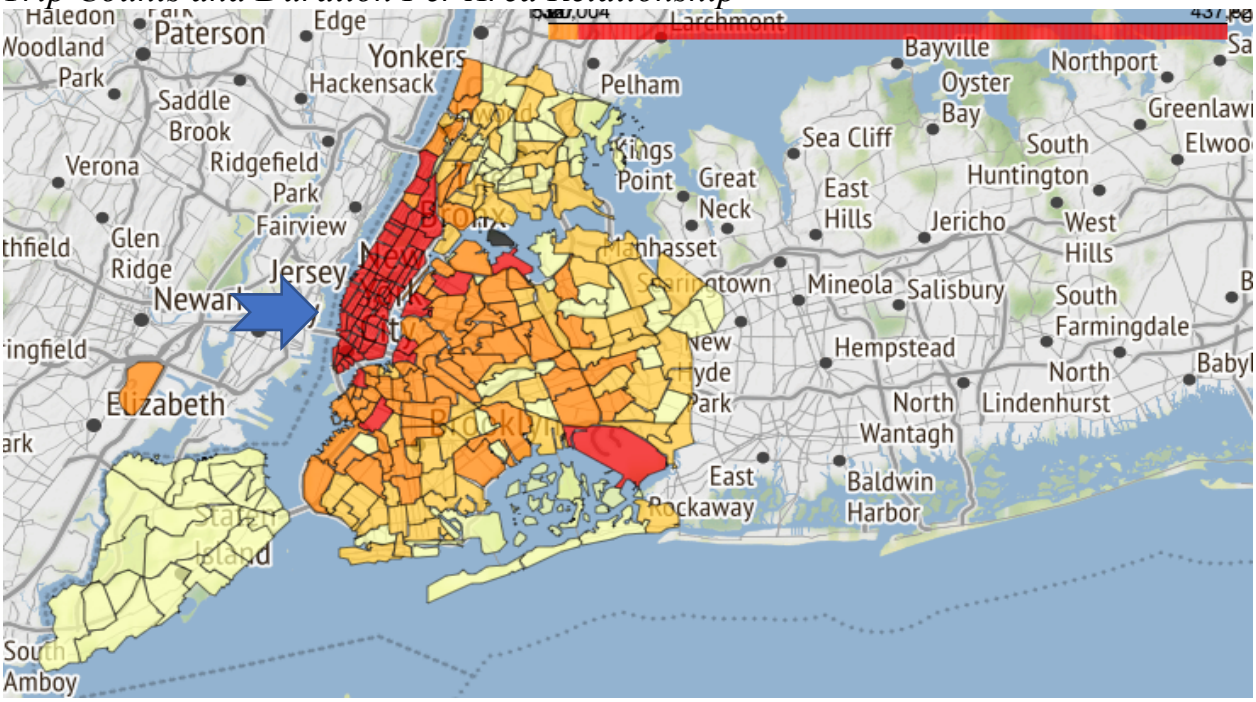


Figure 4 – Drop off Count Per Area (darker indicate higher count)

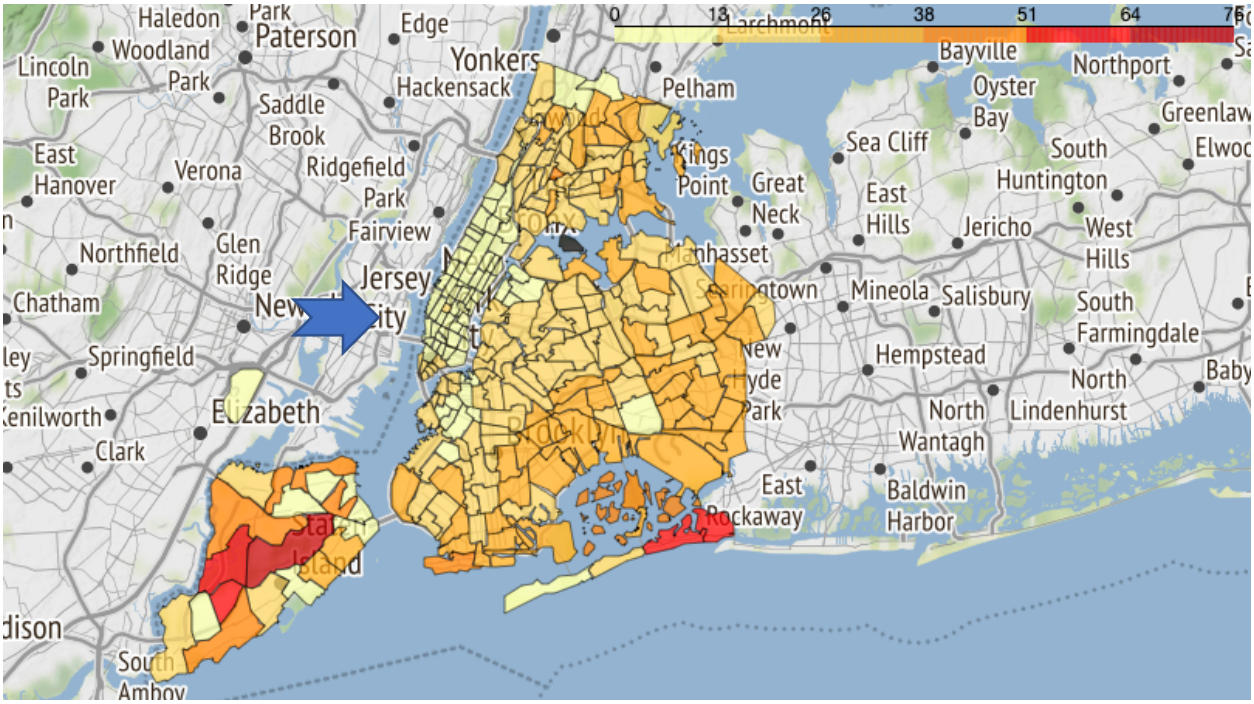


Figure 5 – Average Trip Duration Per Area (darker indicate higher duration)

From Figure 4 and Figure 5, we want to see the relationship between the median trip duration and drop off counts per area. Let us focus on the area of Manhattan, which is pointed by the arrow. It has a low median of duration but has a high drop off count. While the southern area of New York has low drop-offs but high duration. This may indicate the hotspots of the state of New York where the majority of people work, find entertainment, etc. We can assume median people from the south of New York state travel to Manhattan for the reasons stated above, which causes the south area to yield a higher median duration. This is an interesting finding that we should keep in mind when analyzing the rate of tips.

Rate of Tips with Different Weather Conditions and Covid-19 Condition for Different Areas New York City

In this section, we will discuss the rate of tips from different areas on different conditions and different COVID-19 situations. The aim of this plot to give general knowledge on which areas are yielding a higher rate of tips.

To visualize this, we first query the data to pre and post COVID. We also query the data frame based on the condition weather conditions as we assume precipitation level greater than 0.1 inches is considered rain condition. We then group the data with median aggregation with the reason stated later in the report.

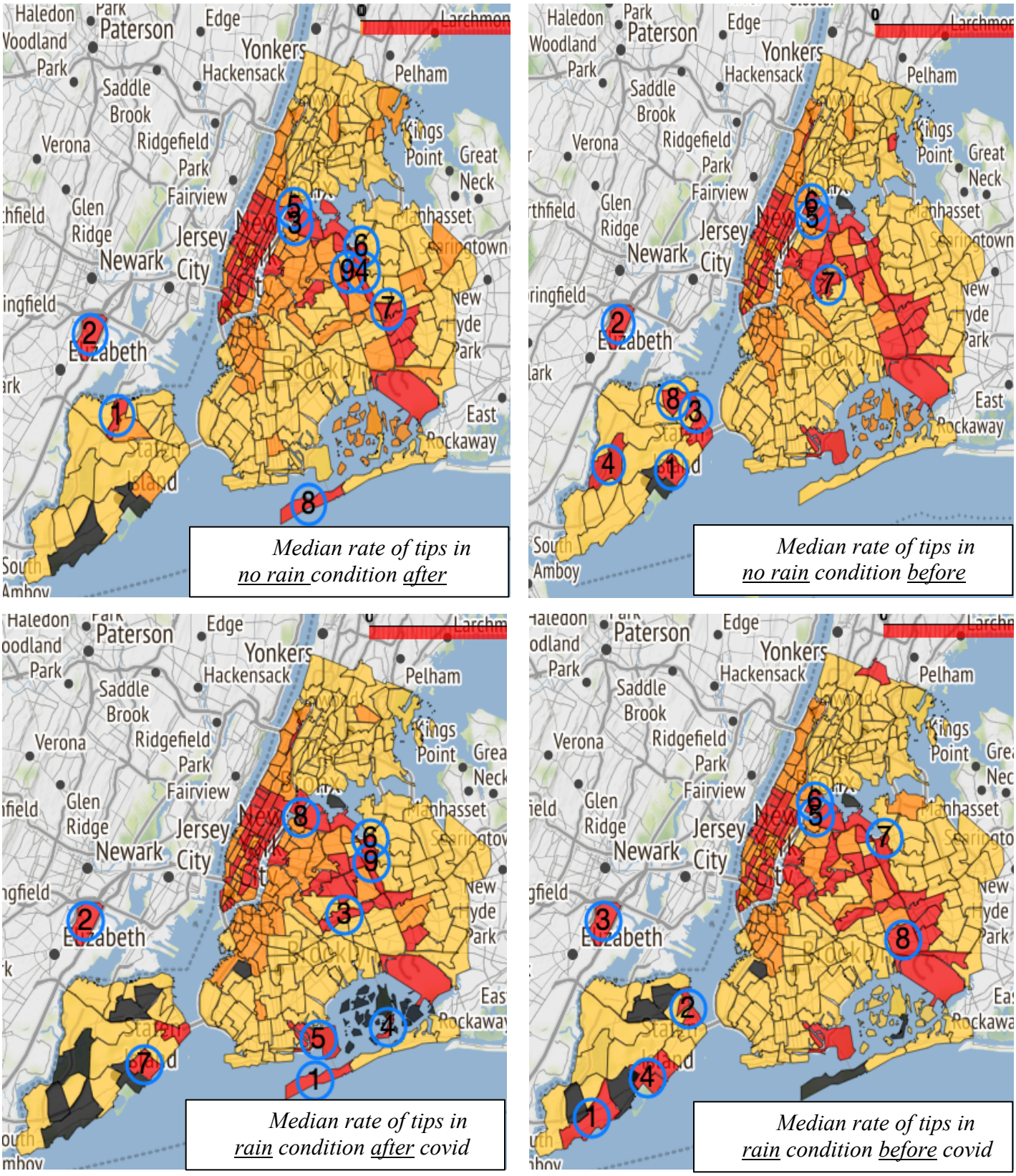


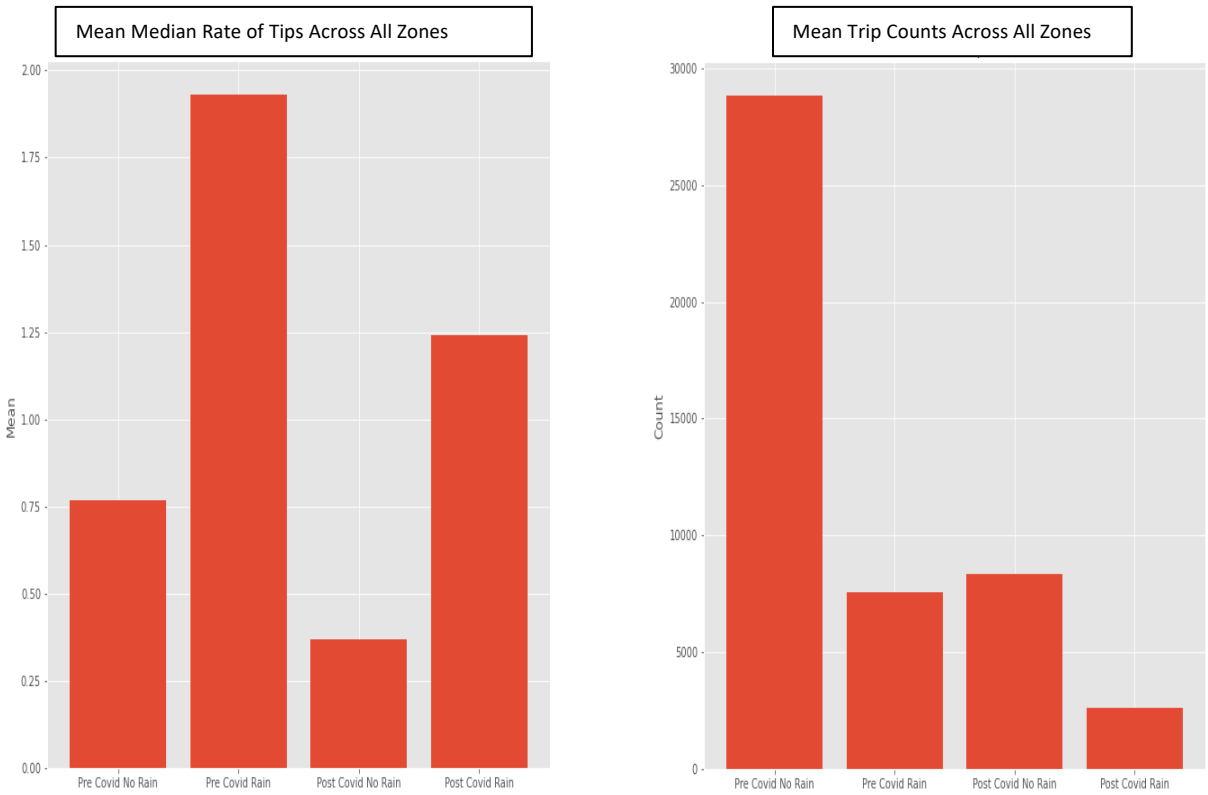
Figure 6 – Rate of tips per area (Darker indicates a higher yield of tip rates)

Here we observe different areas of New York city having different rates of trips with different conditions. Black areas are areas that mean that there is no valid data to be represented in that area after the result of our query. The circle markers with numbers show the rank of areas with the top 10 highest rates of tips. Each graph is tailored to bins of 5 quantile levels of their data to show contrast. We use the median for the rate of tips since it will be a more valid option as the mean is very prone to outliers. And also since we are dealing with a large amount of data.

First, we can see slightly more areas yielding a high rate of tips before COVID-19 conditions. One thing we can see is that the Manhattan area tends to have a higher scale in the rate of tips for every condition. This can be due to the reason stated in the previous section. However, it was never the top 10 areas yielding the highest rate of tips area. This may be caused by the Manhattan pick-ups has more data than the rest of the area which causes a consistent median estimate.

Figure 6 main purpose is to guide and generalize high performing areas of New York city to drivers factoring different conditions. We will summarize, how different weather and COVID-19 condition affects areas of New York.

Rate of Tips and Trip Counts Summary Analysis



.

Figure 7 – Mean of the rate of tips and the number of trips across all zones of New York

An interesting finding from our first graph (left), we found is that rain condition tends to have a higher yield of tip rates. We can understand this result through a psychological standpoint where passengers are more sentiment and grateful towards drivers operating in those weather conditions. This remains the same in the era of COVID-19. We can observe the effect of COVID-19, how it negatively affects the rate of tips as we observe zones before and after COVID-19. However, a higher rate of tips does not conclude to higher profitability. Although rain/snow weather condition affects the rate of tips positively, we can observe a negative correlation with the count of trips from our second graph (right). The impact of COVID-19 remains negative to trip counts. This means that despite rain conditions increases the rate of tips, the frequency of trips will less than average. Although the rate of tips seems to be a good estimate for the profitability of a driver, it is still dependent on the frequency of trips.

.

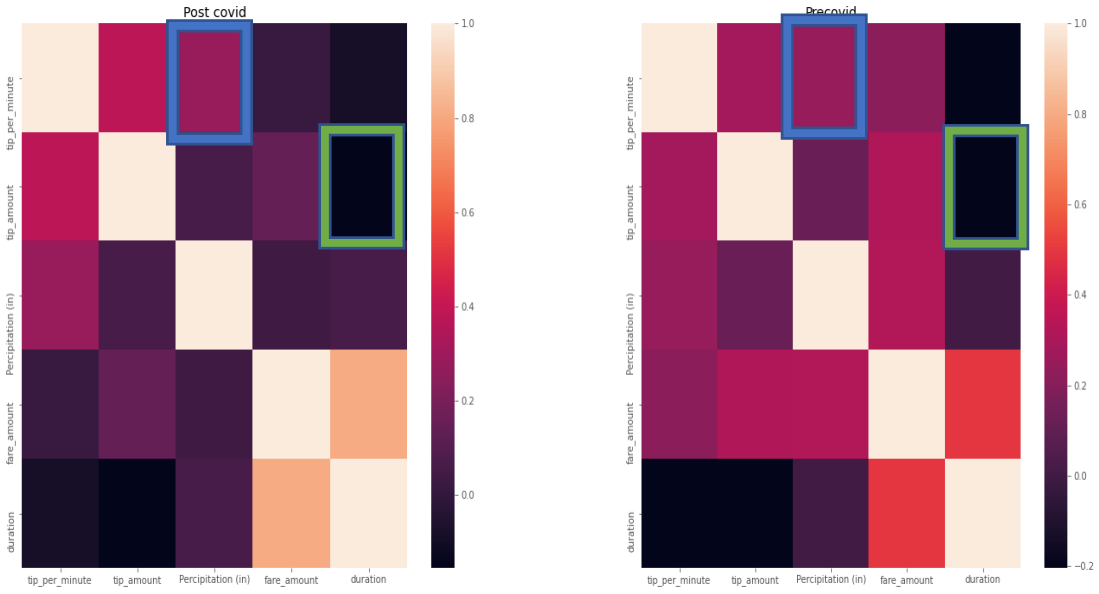


Figure 8– Correlation heatmap of attributes in Post COVID and Pre COVID period

From figure 7, we want to focus on a specific attribute. The highlighted blue box shows the correlation between the rate of tips (tip_per_minute) with weather conditions (Precipitation Level). We can solidify the previous argument where rain conditions increase the rate of tips as figure 7 shows a positive correlation.

The box highlighted green also shows an interesting finding where the total amount of tip yields a negative correlation with the duration of trips. We can observe this from a psychological standpoint as well where if a taxi driver completed trips early, passengers will be more satisfied and tip a larger amount. However, this contradicts the gratuity culture of America where 10% of tips are usually enforced across all jobs. Longer trips should yield a higher average of tip amount. This anomaly may be caused by the result of using a group by where you lose information on some features.

In general, we have darker color in condition post COVID, showing a weaker correlation between attributes. This may be caused due to the difference in the number of datasets on cases of COVID as it loses information. Another may be caused due to the bias of COVID, which shows a stronger feature to correlate to.

Conclusion

Now we will summarize our findings.

We found that COVID-19 has a strong negative correlation with our profitability attribute rate of tips. It also has a significant negative impact of frequency, which we should take account when we observe how weather condition increases the rate of tips. The area of Manhattan seems to have a more consistent frequency of trips and rate of tips since it can be considered the hotspots of New York. Focusing on areas, some areas are yielding higher tips than others. This might be due to the design or the construction of specific areas (parks, the number of public transports) which we require more knowledge on. The Southwest area of New York City can also be an option as it tends to have at least one of the top 10 highest rates of tips area there.

In this pandemic time, it is very important to maximize profit as drivers. Figure 6 is made with purpose, so drivers have an idea which zones yield a higher rate of tips. As New York entering the reopening phase, drivers can relate to the pre-COVID condition of New York. And in case sudden second wave lockdown, drivers can relate to the post-COVID condition. Utilizing weather is one-factor drivers can focus on as it has a positive correlation with tip rates. Although the driver should take into account how both conditions will affect the frequency of tips. The profitability of taxi drivers also comes from the amount of fare, however we need more knowledge on contracts with respective companies and hidden unrecorded fees.

What can be improved from this research and analysis is more knowledge of geographic area features such as public transports, housing, tourist spots, etc. Analyzing the total fare amount as at least 30% revenue of taxi drivers revenue comes from. Using time of day deciding rush hour times. And the comparison with ride-hailing app drivers to see which areas and condition taxi weaker or stronger.

References

Dataset:
https://data.humdata.org/dataset/nyt-covid-19-data?#metadata-0
https://www.wunderground.com/history/monthly/us/nj/newark/KEWR/date/2020-6
Research:
Shala M, June 2018, How Much of Fare Do Taxi Drivers Keep?
https://work.chron.com/much-fare-taxi-drivers-keep-22871.html#:~:text=The%20U.S.%20Bureau%20of%20Labor,the%20other%20half%20earned%20less.
Sharara, R. (2020). The North American Tipping Culture - Why Do We Tip So Much? - The Bull & Bear. Retrieved 3 September 2020, from http://bullandbearmcgill.com/the-north-american-tipping-culture-why-do-we-tip-so-much/