A REPORT ON NYC YELLOW TAXI'S:
FACTORS THAT IMPACT TRIP DURATION AND PROFITABILITY

Akira Takihara Wang
School of Computing and Information Systems
The University of Melbourne
Student ID: 913391

August 22, 2019

ABSTRACT

In New York City (NYC), the iconic Yellow Taxi rides illustrate the life of New Yorkers - a busy city buzzing with citizens by day with a vivid nightlife to match it. With a total of 1.2 million rides driven by NYC's Yellow Taxis in 2018 and accurate weather observations from 3 central weather stations, this report aims to see if we can determine or find potential relationships that can affect a passengers' trip duration. The investigations of the report come to a plausible conclusion that factors such as weather, time, and roadworks can affect the duration and profitability of taxi rides.

1 A look at a Sampled TLC Yellow Taxi Data

The Yellow Taxi data used in the report is a subset of TLC's trip data [1], exceeding well past one billion trips since its initial release in 2009. Although the data used for the report is exclusive to 2018, a total of 1.2 million instances with 17 attributes is still computationally expensive even on a powerful computer. As such, additional tools outside of Python (the base tool of this report) were considered. These included some optimizations techniques such as serializing the dataframe and the use of external languages such as Apache Arrow. Hence, the full utilization of the data sets became possible, and the population sample should be representative of other years. Therefore, the findings from the report should hold for any other year except for significantly unexpected events. (Figure 1) illustrates the availability of the data by zones.

For the weather data used in the report, special permission was obtained from Mr Daryl Herzmann and the recently deceased Dr Ray Arrit from Iowa State University. The observations in the data are automated through minute-by-minute timeframes via an Automated Surface Observing System (ASOS), which are then labelled according to METAR codes [2].

1.1 Data Types and Integrity

Since the Yellow Taxi data post-2015 July switched to taxi zones over the more precise longitude/latitude due to privacy concerns [1], the raw data set for 2018 suggests that it has been cleaned before release. As such, the data types were consistent with attributes and provided shapefiles apart from the DateTime attribute, which can be easily converted to a timestamp format.

Also, a Data Dictionary was provided [3] to help verify the integrity of data. A trivial function to return unique values implied that nearly all the data was correct apart from a few instances. These were dropped without question and have no impact on the analysis.

1.2 Trip Duration

A "trip_duration" (in minutes) attribute was created by taking the difference from "DOdatetime" (dropoff) to "PUdatetime" (pickup). As shown in (Figure 2), the main distribution of points lies around the 50-minute mark, with several points lying in the high 1000's - the equivalent of 20+ hours of trip duration.

Absurd as it is, there is a possibility that customers hired a Yellow Taxi for a whole day going back and forth to locations (i.e tourists or tourism agencies), or the more likely scenario that the taxi driver had simply forgotten to turn the meter off during the

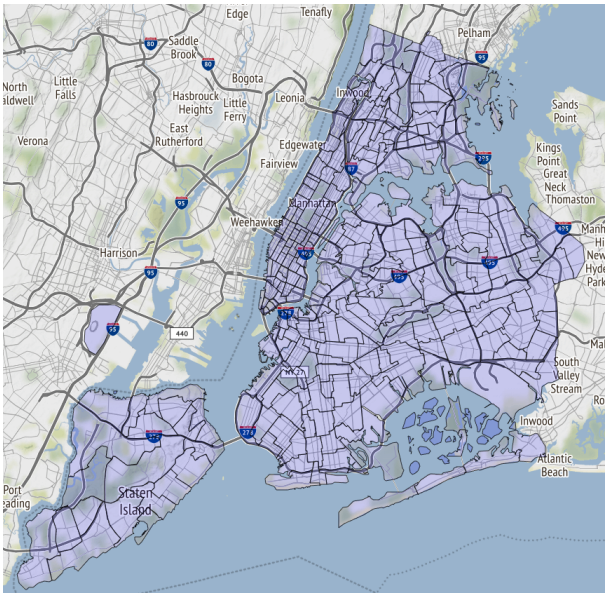

Figure 1: Zone Availability from the 2018 data set.

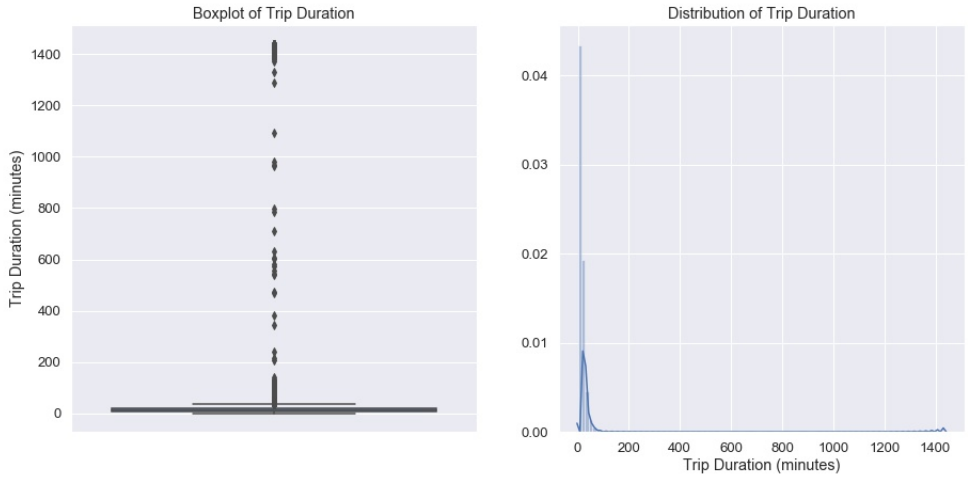

Figure 2: Initial Trip Duration (minutes) and Distribution from Sampled Data ($n = 10,000$).

end of the shift. Since the number of points that fit this description are few and could be valid, there should not be any effect on the visualization and will not be removed.

payment_type	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime
2	5	5	5
3	13	13	13
4	14	14	14

Figure 3: Negative Values grouped by their Payment Type from Sampled Data (2015).

1.3 Negative Values in Fare Amount

The "fare_amount" column also held negative values, which initially may seem invalid since money can only be positive. However, (Figure 3) illustrates the payment types of these instances (2 = Cash, 3 = No Charge, 4 = Dispute) which may be of relevance.

Cash payments maybe have been the result of incorrect balances before it was discovered to be wrong, whilst the negative balances in No Charge or Dispute could be the outcome of the driver losing money from the exchange. Since these are valid instances according to the Data Dictionary, they will not be removed for analytical purposes.

1.4 Zero Values in Distance

Some instances in "trip_distance" were zero, which should not be a valid value. Although a minority can be explained with a Rate Code ID of 5 (Negotiated Fare: Trip can be negotiated or prepaid for a fixed distance), the number of instances is disproportionate to the remaining zero distance trips. After consideration, they will be removed during preprocessing.

1.5 Result

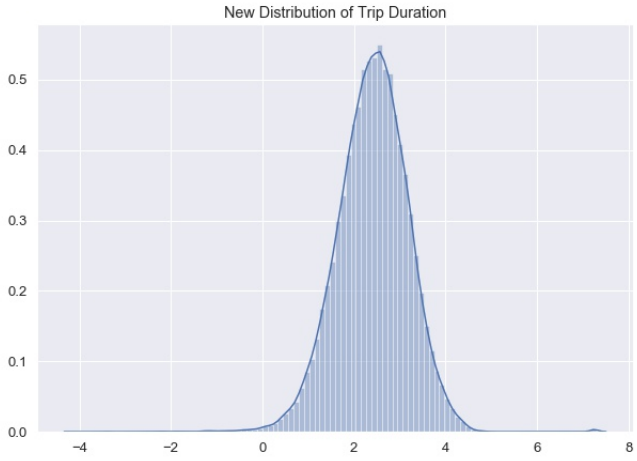

Figure 4: New Trip Duration (minutes) Distribution from Sampled Data (2015).

After taking the above into account, the resulting "trip_duration" follows a much nicer distribution albeit left-skew as shown in (Figure 4).

2 ETL (Extract, Transform, Load) and Preprocessing

2.1 ASOS Weather Data (txt)

1. Missing values and Trace Amount (denoted "M", "T" respectively) were replaced with NaN's and later imputed with 0's. This is because the readings are only active during snow conditions and off otherwise (according to Daryl Herzmann).
2. The "wxcodes" attribute corresponded to weather conditions [2], where normal conditions were denoted as "M" (or NaNs). Hence, all normal conditions were converted to code "NW" for "Nice Weather".
3. With the "ice_accretion" attribute, several instances (due to snow seasons) were invalid or left blank, and can, therefore, be assumed to be 0. However, once imputations were complete, it became apparent that the majority of instances were 0, and under the recommendation of Daryl Herzmann (manager of the data set), the attributes were dropped.
4. Attributes of particular interest were generated, notably "heavy rain/snow", "rain/snow". These were computed using the "wxcodes" flag, where a "+" or "-" indicated intense or mild intense weather conditions respectively. Other flags such as hail, thunderstorms, or windy were not used since it was observed to be a subset of snow and rain.
5. The "valid" (standing for validated timestamp) column was converted to standard DateTime format.

6. A group-by over "valid" using mean aggregation was taken to smooth out the readings from the three stations.
7. The "valid" attribute was renamed to be "DateTime", and rounded to the closest hour. This is due to the precision of minutes being unnecessary in the context of the analysis, and memory considerations for adding both minutes and seconds were at play.
8. The processed file was then outputted to a simple ".csv" since there were less than 500 instances.

2.2 NYC TLC Yellow Taxi's Data (csv)

(*For each monthly dataset:*)

1. The "tpep_pickup_datetime" and "tpep_dropoff_datetime" attributes were converted into timestamps.
2. A checking function was applied to some attributes to ensure all values were well defined from the provided "Data Dictionary" and the considerations explained above, and all invalid rows were dropped.
3. A "trip_duration" attribute was created by subtracting "PUdatetime" from "DOdatetime" and rounded to the nearest minute due to memory implications of any further accuracy.
4. Entries whose distances or time were 0 (possibly due to a "RatecodeID" of 5) were too few to be included and were added to the indices to be dropped.
5. Once invalid instances were dropped, attributes made redundant were also removed from the data set (Notably "RatecodeID" was taken into account by the "mta_tax" according to [1]). After dropping them, we are left with these attributes: ["PUdatetime","DOdatetime","PULocationID","trip_duration","trip_distance"].
6. A subset using a groupby on "payment_type == 1" is also created with attributes ["PUdatetime","DOdatetime", "PULocationID","fare_amount","tip_amount"]. This is because only credit card tips (payment type 1) are consistently recorded, hence any analysis relating to costs will use this subset.

After each month was processed, they were merged to make an *annual* dataset with a total of 1.15 million instances. Due to time and space implications of such a large dataset (12GB as a ".csv"), the data was serialized into either feather formats [4] or pickle formats [5] and read in partitions during analysis.

2.3 NYC TLC Taxi Zone Shapefile (shp)

1. Since the polygon shapes are a collection of coordinates in an arbitrary space, they were converted to standard longitude/latitude using WGS84 latitude-longitude projection.
2. Redundant attributes such as: ["OBJECTID", "Shape_Leng", Shape_Area", "zone", "borough"] were dropped due to the "taxi+_zone_lookup.csv" containing the same attributes already.
3. The full availability of zones is provided in (Figure 1) as a reference.

3 EDA (Exploratory Data Analysis)

3.1 An Overview of NYC trip durations

(Figure 5) projects trip duration on a logarithmic scale to a map of NYC. From initial observations, central NYC and Manhattan share the shortest trip duration, with the surrounding suburbs becoming longer in duration.

A proposed explanation may be that suburban citizens travel via taxi from the surrounding suburbs into Manhattan for work, whilst trips within Manhattan are mainly taken for short trips between workplaces. Exceptions to this include the airports LaGuardia, Newark and JFK, as well as Arden Heights which seems to have the longest trip duration for some reason that cannot be explained by the current data.

The common "holiday" or "peak hour" theories should also hold in NYC, where the frequency and duration of trips are expected to increase over the peak seasons and hours respectively. Likewise, with the extreme weather conditions in New York, we also expect to see some taxi trips being affected during the winter season. However, the question remains as to why Arden Heights has more than double the duration compared to anywhere else has still yet to be explained by the data.

Arden Heights has an expressway (New York State Route 440) connecting itself to New Jersey and Central NYC, so trip durations should not be any longer than its surrounding neighbours. A look at roadwork construction data for 2018 (provided by DOT) reveals the cause. For the year 2018, West Shore Expressway experienced two major construction works during May [6] and November [7] separately. Several sections of the expressway were closed off to make way for new auxiliary lanes, resulting in detours and delays for commuters. This means that pickups based in Alden Heights would experience some form of delays, and may serve as a possible cause to the significantly higher trip duration time.

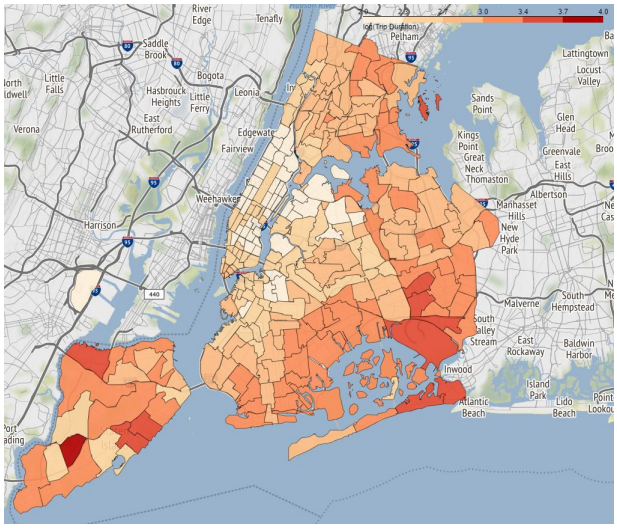

Figure 5: Choropleth of NYC with respect to trip duration (log minutes).

3.2 Time as a Factor for Trip Duration

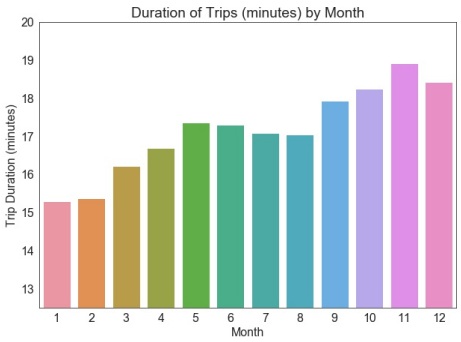

Figure 6: Expected Monthly Trip Durations (minutes).

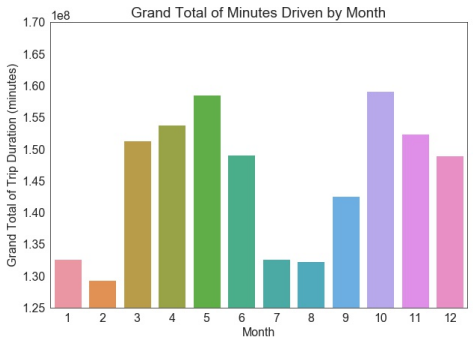

Figure 7: Monthly Grand Total Trip Durations (minutes).

From (Figure 6), we can examine a positive trend over months concerning expected trip duration. One question that arises from this is if it is a gradual trend building upon previous years, or if it is independent within years. Since the data in the report is exclusive to 2018, it is unsure which case it falls under. Complimenting the plot is (Figure 7), which displays the cumulative total of minutes drive by months - computed using:

$$\text{Grand Total} = E[\text{Monthly Trip Duration}] \times \text{Total Trips in Month}. \tag{1}$$

From this figure, we identify the months between March and May to have a large increase in total minutes driven, with a sharp decline in the months leading up to October. This means that there must be a factor affecting the frequency of trips, and will be explored later on in the report.

Exploring (Figure 6) in more detail are (Figures 8 and 9), which display the number of trips driven by Yellow Taxi's in May and November due to their noticeable local peaks in the plot. Notable national events in May were the Armed Forces Day and Memorial Day, whilst November had State Elections, Veterans Day, Thanksgiving Day and Black Friday (notoriously known for Black Friday Shopping) [8]. Although it is unknown whether or not they have a significant impact on the trip duration and profitability for taxi drivers, the national events certainly bring more citizens into Manhattan as observed in (Figures 8 and 9) and serve as one plausible explanation.

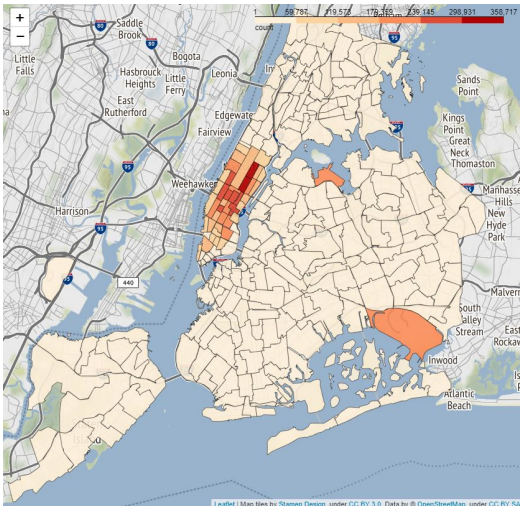

Figure 8: Choropleth Map of May's Total Trip Counts.

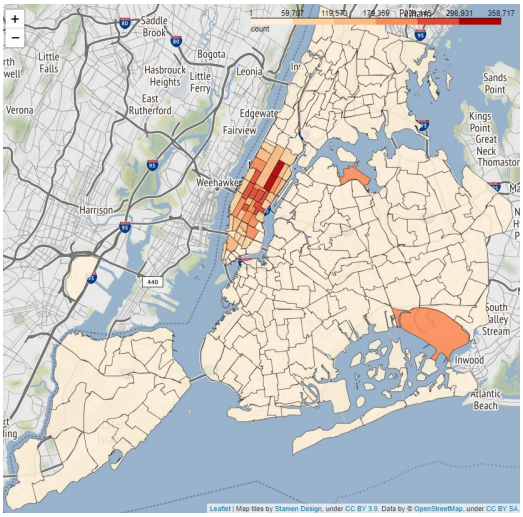

Figure 9: Choropleth Map of September's Total Trip Counts.

As for the holiday theory, (Figures 6 and 7) illustrates peaks during the holiday months (May - July, October - December) with a much larger spike in the latter holiday months. Although there was no statistical test conducted to sufficiently prove this, the holiday theory seems to generally hold according to the visualizations.

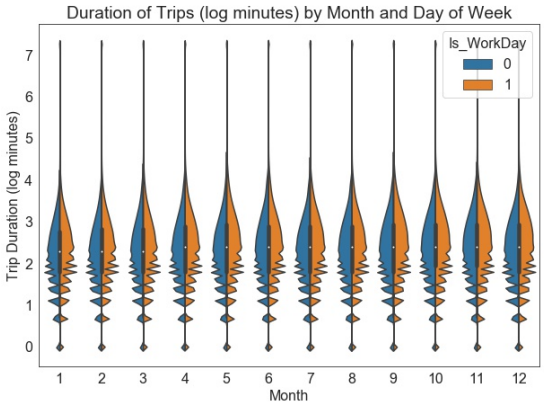

Figure 10: A violinplot of log(Trip Duration) with respect to Workdays vs Weekends over Months.

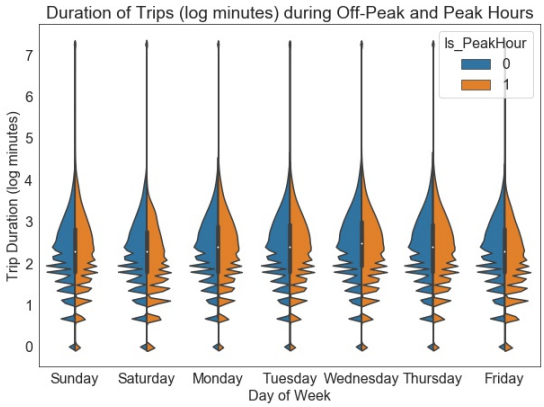

Figure 11: A violinplot of log(Trip Duration) with respect to Peak vs Off-Peak Hours over Days of Week.

From (Figure 10), workdays (Orange: Mon - Fri) consistently show an average of 3.7 minutes longer trip durations than weekends (Blue: Sat-Sun), which seem to take the majority of the upper outliers. Interestingly, (Figure 11) contradicts the peak-hour theory [9] and we observe overall trip durations to drop by 1.2 minutes on average, despite the number of trips remaining equivalent to off-peak hours. This may be due to the additional $1 peak hour fee [10] that is active between 4 pm - 8 pm weekdays, where people may be reluctant to catch a taxi during these hours.

3.3 Pickup Zones as a Factor for Trip Duration and Profits

Firstly, we define our profit measurements in this report by two separate rates - minutes (t) or distance (d). This is because taxies follow a strict pricing structure as explained in [10], but there is no accurate information regarding the exact fare breakdown in the data. Hence, an additional independent profit measurement that avoids these factors had to be created.

1. Time is a useful measure since there may be drivers whom *prefer* short trips at a higher frequency.

$$\text{Zone Profit}_t = \log\left(E\left[\frac{\text{Fare Amount} + \text{Tip Amount}}{\text{Trip Duration (minutes)}}\right] \times \frac{\text{Frequency of Trips in Zone}}{\text{Total Number of Trips}}\right) \tag{2}$$

2. In contrast, Distance is a useful measure since there may be drivers who enjoy longer trips at a lower frequency.

$$\text{Zone Profit}_d = \log\left(E\left[\frac{\text{Fare Amount} + \text{Tip Amount}}{\text{Trip Distance (Miles)}}\right] \times \frac{\text{Frequency of Trips in Zone}}{\text{Total Number of Trips}}\right) \tag{3}$$

Note: *The logarithm is taken it is the relationship concerning its unit (time or distance) that is important, and the pickup zones with frequencies lower than 30 per month (1 per day) are ignored due to the high variances it may cause.*

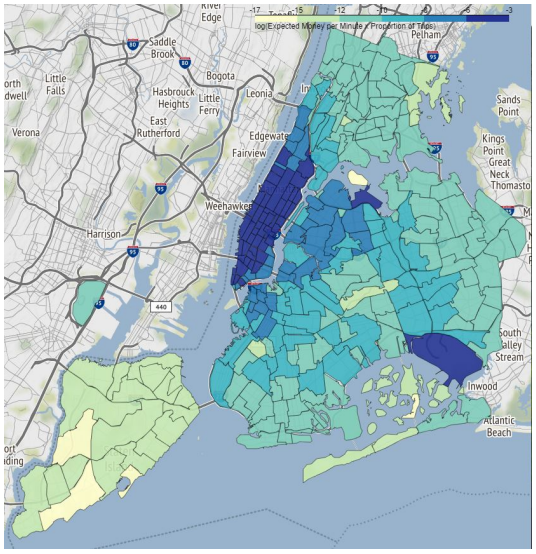

Figure 12: Estimated Zone Profit using Time (minutes).

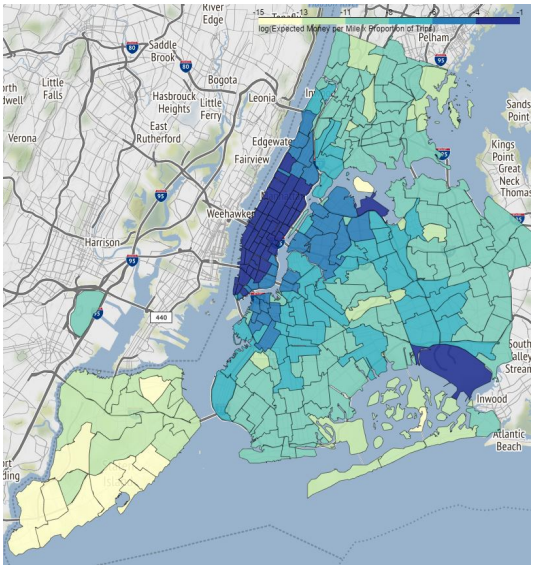

Figure 13: Estimated Zone Profit using Distance (miles).

Regardless of the zone profit calculations, (Figures 12 and 13) appear to contrast (Figure 5). One might expect longer trips to be more profitable, but by taking zone frequency into account we now see that Central NYC / Manhattan, LaGuardia Airport and JFK Airport are the most profitable. For taxi drivers looking to make several profitable trips in a day, Central NYC / Manhattan seem to be good pickup zones (with drop-offs remaining within Central NYC / Manhattan). However, both LaGuardia Airport and JFK Airport have profit to be made but the downsides are numerous. Notably, the trips are much longer and there is a fiercer competition, resulting in a less profitable zone contrary to the points made above.

The largest competitors for JFK Airport to Manhattan according to [11]:

- Uber / Lyft has an average cost of $42.50 to get into Manhattan with a long wait time.
- Airport Shuttles are even cheaper at a flat rate of $19 with the downside of stopping at locations.
- The Airtrain is the cheapest way to travel to Manhattan at a one-way cost of $7.50 but requires the traveller to carry their luggage.

When compared to the minimum fare of $52 for a taxi, it is obvious that hailing a Yellow Taxi at an airport will not be the most popular option for the average economical traveller. However, the Yellow Taxi does operate 24/7 and according to some local travelling guides, it may be an excellent choice for travellers in groups of four. Appropriately, we can conclude that Manhattan may be better for taxi drivers during the graveyard hours (with the added surcharge fees) but not as good during the daytime hours.

	trip_duration	fare_amount
trip_duration	1.000000	0.889485
fare_amount	0.889485	1.000000

Figure 14: Fare Amount shares a strong positive correlation of 0.8895 with Trip Durations.

	fare_amount	tip_amount
fare_amount	1.000000	0.756705
tip_amount	0.756705	1.000000

Figure 15: Tip Amount shares a strong positive correlation of 0.7567 with Fare Amount.

From the bigger picture, we can devise some potential strategies for Yellow Taxi drivers:

1. Although LaGuardia Airport and JFK Airport are profitable, it does not take into account the competition and number of taxi's waiting for customers. But, the taxis have one competitive advantage - they are allowed to operate 24/7. So, travellers who arrive during the graveyard hours will still need transportation but will have more limited options (the trains and shuttle buses do not run after certain hours). Therefore, it is recommended for drivers who can or are content with night shifts to make pickups from the airports.
2. By custom, tips for NYC Taxis are 10%-20% [12] of the fare rate and can be shown to have a strong positive correlation (0.7567) with larger fare rates. (Figure 14 and 15) show the correlation tables with the assumption that longer trip durations will have a larger fare, hence the tipping amount will also be larger. Consequently, it is recommended that off-peak hour trips from the surrounding suburbs are a good choice for accumulating large sums of tips if the driver is happy to wage their bets on it.
3. During weekday peak-hours, taxi drivers should aim to target workers or tourists who are looking to head into the surrounding suburbs passing through Manhattan South. In addition to the peak-hour surcharge, Yellow Taxi drivers may also collect an additional $2.50 Congestion Surcharge and any toll fees along the way [10]. This means there is much more money to be made, and more trips can be accomplished if the suburbs are close by.

3.4 Weather Conditions (Snow) as a Factor for Trip Duration

The roads of NYC spend at least 3 months of the year under the snow, with March being the most extreme. From (Figure 16), March has the most number of warnings for heavy snow, but unexpectedly boasts the most number of trips despite the observations made from (Figure 7), which had suggested that March was not a busy month.

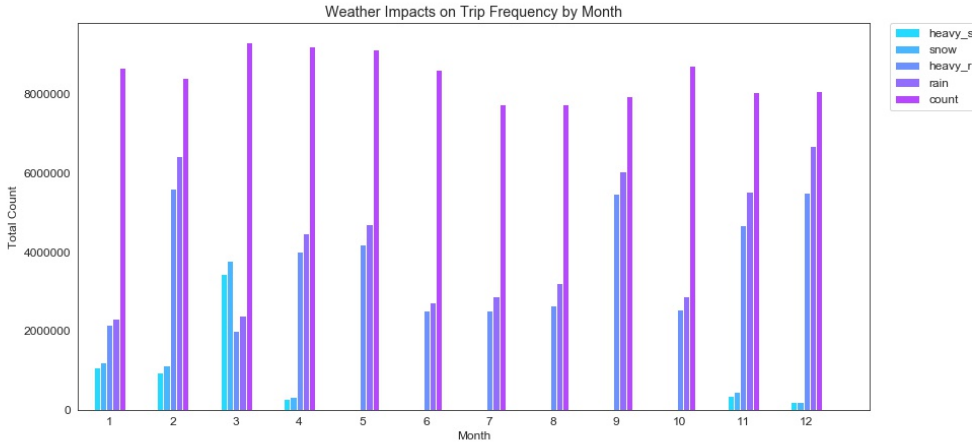

Figure 16: Total Number of Trips and Weather Warnings per Month.

One such justification of this observed behaviour could be due to the intense snowfall. Citizens will want to avoid walking outside in these conditions and may wish to opt-in for a short taxi ride which would otherwise be walkable under normal weather conditions. (Figure 17) gives an example as to what an average March evening may look like, and how snow buildup on the road is a major issue. The official response to such concerns is led by clear roads [13] and DOT.

Figure 17: *"There's actually a lot of snow in New York"* - Business Insider, March 2018.

According to the DOT's "Snow and Ice Control" procedures [14];

> *The Department's goal is to provide highways and central roads that are passable and reasonably safe for vehicular traffic as much of the time as possible within the limitations imposed by weather conditions and the availability of equipment, material, and personnel.*

In practice, only zones close to Central NYC / Manhattan will remain in full operation during high levels of snowfall and since resources are finite, several roads may be closed off creating delays in travel. For the report, we will generously define a pickup zone as "operable" under heavy snow conditions if there has been a bare minimum of at least 2 trips completed in one day.

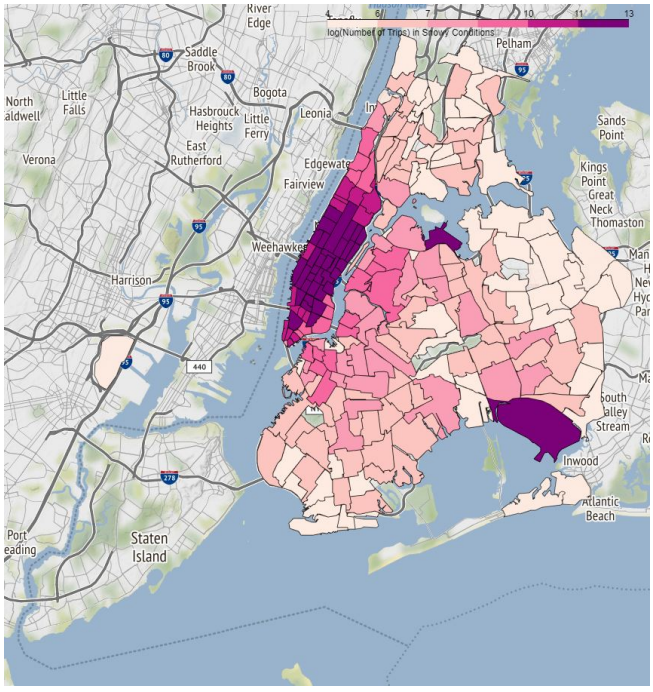

Figure 18: log(Number of trips) made in "Operable" zones in heavy snow conditions.

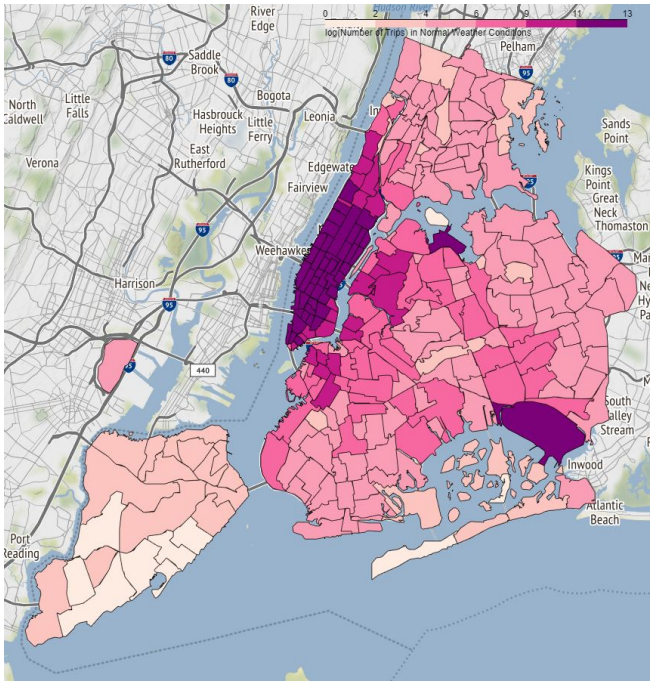

Figure 19: log(Number of trips) made in any zones during normal weather conditions.

The availability of "operable" zones in heavy snow conditions can be viewed in (Figure 18), with a complimenting comparison map in (Figure 19) for normal weather conditions. Instantly, an observation can be made that the number of active zones during heavy snow conditions is far less, but according to our justification above, we should expect to see a higher density of trips concentrated within central NYC / Manhattan.

From (Figure 20), "trip_duration" shares a strong negative correlation (-0.5) with both "heavy_snow" and "snow", where "heavy_snow" and "snow" have a -0.6 and 0.5 correlation with "trip_distance" and "count" (number of trips) respectively. To visualize this correlation, bar plots concerning heavy snow conditions can be seen through (Figures 21 and 22). There is a lower trip duration/trip distance respectively, and (Figure 21) shows a higher number of trips completed during snow conditions despite an enforced minimum zone frequency of two per day.

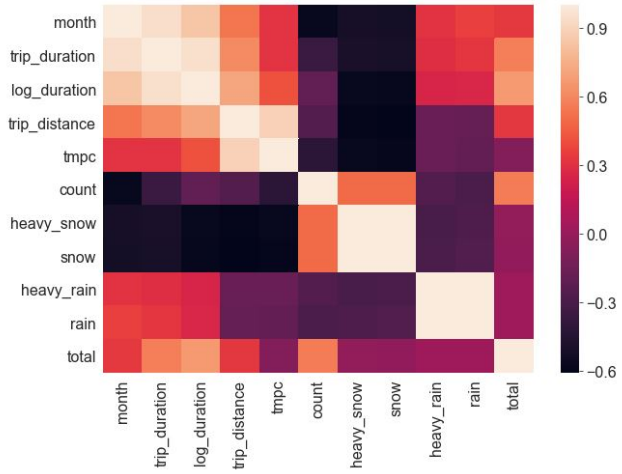

Figure 20: Correlation HeatMap of Weather against other Attributes.

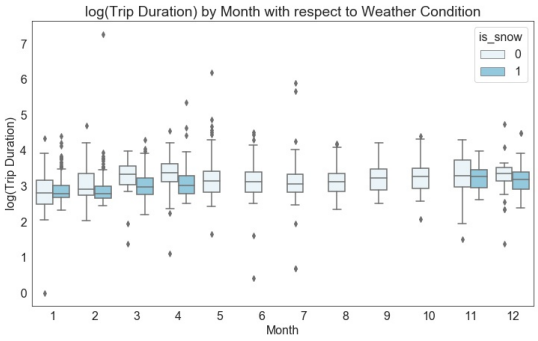

Figure 21: BarPlot of log(Trip Duration) by Month with respect to Snowy Conditions.

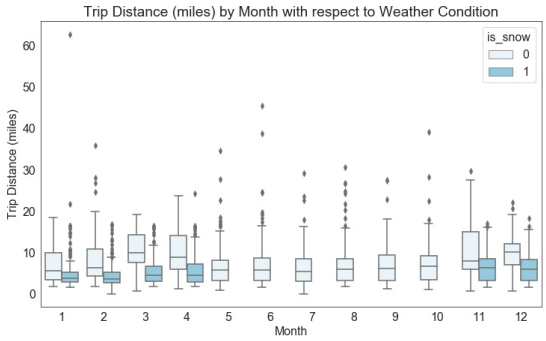

Figure 22: BarPlot of Trip Distance (miles) by Month with respect to Snowy Conditions.

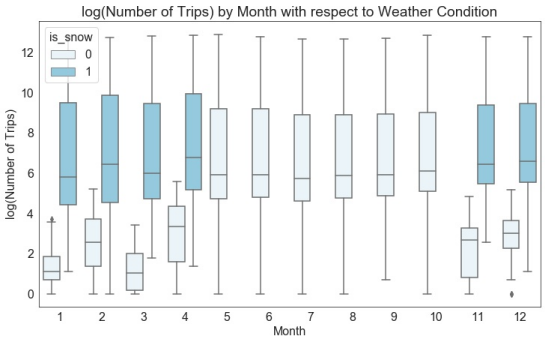

Figure 23: BarPlot of log(Number of Trips) by Month with respect to Snowy Conditions.

As such, a conclusion remaining consistent with the findings from (Figures 6 and 16) can be formed from the above, which contends that taxi trips during snow conditions will be more frequent, albeit shorter in duration and distance. This also aligns with the initial justification with the observed behaviour.

For the brave taxi drivers who wish to be profitable during heavy snow periods, some new strategies can be formed:

1. Consistent with the first proposed strategy from before, LaGuardia and JFK Airport seem to be unaffected by the snow conditions and will remain busy. Taxi drivers should aim to find travellers heading into Manhattan rather than the suburbs, as they are more likely to find a return trip to the Airports.
2. Taxi drivers should cluster within Central NYC / Manhattan and attempt to make as many short trips as possible during their shift. By doing so, they can make the most out of the initial $2.50 hailing fee, as well as the additional surcharges during the peak hour times.
3. Lastly, taxi drivers may also gamble on the tips made on longer trips into the outer suburbs with the added toll fees if lucky. By travelling back into Central NYC / Manhattan avoiding the toll roads through normal roads, they can also make a decent sum of profit.

4 Conclusion

It was found after analysis that external factors such as weather, time, roadworks and pricing can affect the duration and profitability of Yellow Taxi rides. Although the assumption that each factor was independent of each other was made and additional measurements such as profitability or operability were devised, the results of the analysis suggest evidence that weather conditions and events play one major role, notably when NYC enters the snow season. An important factor which could be considered in a future attempt may be tourist data or event data, which were not taken into account for this report due to data availability.

Interesting findings to note included; a decrease in trip duration with no increase in trip frequency during weekday peak hours, an increase in trip frequency with a decrease in trip duration and distance during heavy snow conditions, and passengers were happy to pay larger tips for longer distances despite the larger fare overall.

Since no rigorous statistical testing was carried outside of the visual plots and covariance matrices, future considerations include conducting tests to see if they are statistically relevant. These will serve as a stronger basis for conclusions and yield more effective strategies. Also, a look at the distribution based around the median may convey a different perceptive since it is less affected by outliers in the data. Finally, it may be used possible to predict whether or not a pickup zone is profitable based on the accessibility, weather condition and time of year.

References

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] https://en.wikipedia.org/wiki/METAR#Example_METAR_codes

[3] https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[4] https://pypi.org/project/feather-format/

[5] https://docs.python.org/3/library/pickle.html

[6] https://www.dot.ny.gov/news/press-releases/2018/2018-05-14

[7] https://www.dot.ny.gov/news/press-releases/2018/2018-11-021

[8] https://www.tripsavvy.com/november-weather-in-the-united-states-3301219

[9] https://en.wikipedia.org/wiki/Rush_hour#United_States

[10] https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

[11] https://www.nycbynatives.com/visitors_center/jfk_to_manhattan.php

[12] https://www.tripadvisor.com.au/Travel-g60763-c71510/New-York-City:New-York:Tipping.In.Nyc.html

[13] https://clearroads.org/

[14] https://www.dot.ny.gov/divisions/operating/oom/transportation-maintenance/repository/NYS_SI_Manual_Apr2006_RevJan2012.pdf