Exploratory Data Analysis of New York Yellow Taxi: Investigating New York Yellow Taxi Trip Behaviour During Winter

Nathanael Luira Yoewono
Bachelor of Science, Data Science
University of Melbourne
Student ID: 1000582

November 1, 2020

Abstract

A detailed analysis of New York yellow taxi trip behaviour was done by examining the historical trip frequency and trip distance trend, both hourly and daily. A graphical method, such as choropleth was used to further visualize the trend in each area around New York. The result showed that trip frequency would fall during Christmas and heavy snow fall rate. Average trip distance would also decrease when snow thickness reach beyond five inches. The trip frequency and distance can be further described by adding another dimension, that is hour and day. Identifying this behaviour may be important for taxi driver, as may be able to estimate during specific day and hour, which area may have the most demand for taxi ride.

1 Section 1 - Introduction Data and Attribute Selection

To begin with, the appearance of high frequency for hire vehicle, such as Uber, may have dominated New York taxi daily trips (Guse, 2020). This may be due to its innovative business model that disrupt the transportation market, especially taxi. However, one specific uniqueness that a taxi has is the ability for a passenger to hail the taxi without any booking appointment. Hence, it may be better if the taxi driver would be able to notice the time, place and behaviour in which people may need rides. As a result, this analysis will focus on dismantling the past behaviour of New York taxi trips, and it will be restricted to a particular season, which is winter. The period will be limited to December-February from 2016-2019, as the element of interest for winter here is snow fall rate and depth, which usually peak around January-February in New York (The climate of NYC, 2020).

The first step of this analysis is to chose the appropriate data. It is decided that Yellow taxi TLC data set is chosen, as it contains a considerable amount of real time historical taxi trip data with various attributes (TLC Trip Record Data, 2020). Additionally, the reason why the period of year is restricted to 2016-2019 is due to the change of data format from 2016 below, whereas beyond 2019, the Yellow Taxi performance may be affected by Corona virus pandemic. Consequently, the snow data will also be restricted to this period to match the taxi data, and it was taken from New York National Weather Forecast website (National Weather Service Climate, 2020). The taxi data consist of 17 attributes and an average of around 7 million trips recorded in each month. Most of the data types are continuous; however, there are also some string and discrete values, such as pickup date time and location id. The snow data consist of 270 daily predictions with five attributes: max temperature, min temperature, average temperature, snow fall rate and snow depth. The temperature are all stored in Fahrenheit with float data type, while the depth and rate are in inches with string data type. Moreover, both TLC taxi and snow data has the wide rectangular data format, which is very convenient for data processing.

In addition, after researching on attributes that may describe taxi trip behaviour, a decent starting point is to investigate the historical trip frequency for each period, that is 2016-2017, 2017-2018 and 2018-2019 from December to February. This part of the analysis might give an overview of the trip frequency trend, and how the trend may change progressively through out each period. One may also see how certain event may disrupt the trend behaviour of the trip frequency. The next selected attribute is hour and day. This attribute will add a dimension in the analysis, as they would expand the granularity of the behaviour analysis, resulting in hourly and daily behaviour of the trips. Certain cycle might be expected for the trip frequency in hourly and daily basis, and they might also be affected when snow fall in New York city. The last attribute that will be discussed is trip distance. This attribute may reveal whether people tend to take taxi trips for short or long distance trips, and whether specific snow condition could affect New York people's preference for their distance travelled.

Ultimately, the expectation of this analysis is that by using the TLC New York yellow taxi data and the New York National Weather Forecast snow data, one would be able to examine the trip behaviour of New York yellow taxi trips from their past trip frequency and distance preference, both daily and hourly, and how snow fall might caused changes to these behaviour.

2 Data Preprocessing

Due to the limit of Ram, the preprocessing was divided into five layers. There are nine data frames that went through each of these layers. None of these data frame has any missing values; hence, no impute was done to any of its attribute column values. In each layer, the data frame was saved and reopened in the next layer. Theses data frame consist of all New York Yellow Taxi historical rides from 2016-2019, within December - February. To start with, the first step of the preprocessing was to eliminate rows based on certain attribute values. There are eight guards that were implemented for slicing the data frame. Firstly, passenger attribute indicates the number of passenger that is in the taxi. However, some of the passenger values are 0; hence, the data frame will be sliced based on passenger that is bigger than 0. Second, the data scope was limited to only cash and credit payment type, as dispute and no charge payment type has negative values in its fare amount, and this may distort the analysis. Third, there were some undefined rate code id, such as 99; thus, these rows were be excluded. Next, all trip distance should be greater than 0, and the minimum fare is $2.5 (Taxi Fare - TLC, 2020). Lastly, speed and time taken, a feature engineered attributes, were extracted from trip distance over time taken in hour, and pickup date time minus drop off date time. They were cleaned by taking both time and speed greater than 0. This may aid eliminating errors in recording the date time pickup and drop off data, resulting in negative time travelled for the taxi trips.

In addition, the second layer of the preprocessing step focused more on the fare rate and other miscellaneous attributes, such as tolls and extra. After inspecting the data further, it appears that there were some fare amount values that are below 0. Those indexes with negative values were sliced from the data frame, as these instances are not relevant to the aim of this analysis. Moreover, huge values of fare amount were also identified. For example, there are instances where a taxi was paid for $135 dollar for just one minute trip time and 0.3 miles of trip distance. To handle these data, a ratio of fare and speed was made to assist in finding these anomalies. By inspection, data within 0.025-0.975 quantile show a decent ratio; henceforth, the dataframe was sliced based on these quantiles of the ratio attribute. Not only that, the speed attribute was limited to below 65 mph, as it is the maximum speed in New York state (Weiss, 2012). There were some occurrences in the data where the speed was beyond 1000 mph. Additionally, some predefined rules were listed in TLC website regarding to the fare rate, especially for JFK and Newark, rate code id 2 and 3. For JFK, the rate is fixed, with $52 as its cost, whereas for Newark, the rate was set to a minimum of $17.5 for each ride. Ergo, additional slicing guard was implemented to ensure consistency in the data quality. Lastly, miscellaneous attributes such as Tolls amount and extra has some anomalies in it, for instance, negative values in extra, and a very expensive cost in tolls amount. Therefore, further slicing were done to ensure that there are no negative values in extra, and limit the tolls cost by $100 (Tolls On Popular Bridges and Tunnels, 2020).

For the third layer, it turns out that some anomalies in the fare amount and trip distance still exist in the data. These anomalies can be shown in the trip distance and fare amount ratio. For example, with only 0.01 miles, the passenger was charged with $52. In this case, it might be that this trip was cancelled, but was recorded mistakenly. Thus, the proposed solution was to create a ratio of fare amount over trip distance attribute to locate indexes that contain these anomaly. There were two approaches that was been conducted to slice these indexes. One is using the outlier whisker detection by setting a range for the data to be within q1-1.5*iqr and q3-1.5*iqr, iqr = inter quantile. However, the result was not satisfying, as some anomalies were still detected. The other method is to examined it manually. As a result, by inspection, those indexes that has a fare/trip ratio beyond 200 were be excluded. Moreover, in this layer, hour and day attribute were extracted from the pick up date time attributes. It was extracted in this layer as now most of the data frame size has reduced by a quarter approximately; hence, less RAM is used to compute the feature engineering process.

Furthermore, in the fourth layer, each data frame went through five different groupby method. The first groupby was grouped based on pickup and drop off location id with its date and rate code id. In this way, the data will consist of each date trip destination with its rate code id distribution. Second, the dataframe was also grouped by pickup, drop off, day and hour. This may help show the daily and hourly trend of each attribute that may affect trip frequency and distance. Lastly, the last groupby is on rate code id, pickup location, day and hour. This groupby mainly focus on the trend of the rate code id.

Finally, the last layer's function was to append all of these seperate data frames into one big data frame. The analyst managed to reduce a total of 4.5 GB data into 2 GB, cleaned and extracted. The files were saved in feather format to enable fast reading of data frame object.

Aside from the TLC data, snow data also need to be preprocessed. The missing values in the snow fall and snow depth are replaced with 'T' and 'M'; hence, this value are replaced with its previous 3 days average, as the assumption here is that

today's weather may be affected by the previous weather condition. Overall, the data has no other missing values, and minor type changes were done to the snow fall and depth attributes.

3 Analysis and Visualisation

Firstly, before examining the proposed attribute, a set of hypothesis questions ought to be listed in order to guide the flow and aim of this analysis. Firstly, it is hypothesised that there are less trip frequency during snow and low temperature. Second, national event or holiday may disrupt the trip frequency as well as the cycle of the trip behaviour. Third, there may exist a cycle in hourly and daily behaviour in trip frequency, and that these cycle may also be disrupted by the heavy snow fall rate. Also, it is hypothesised that each area in New York may have different hourly trend in its trip frequency. Fifth, the average trip distance of a taxi may decrease when the snow thickness is higher than usual. Finally, it is also hypothesised that a decrease in average trip distance may correlate with the decrease in trip frequency as well. To examine all of these hypothesis statement, one may need to consider this visualisation section.

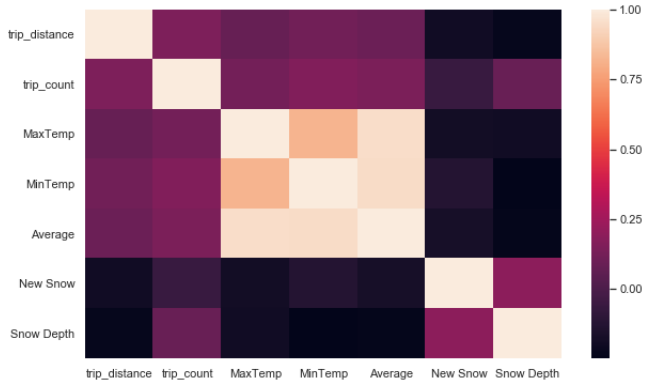


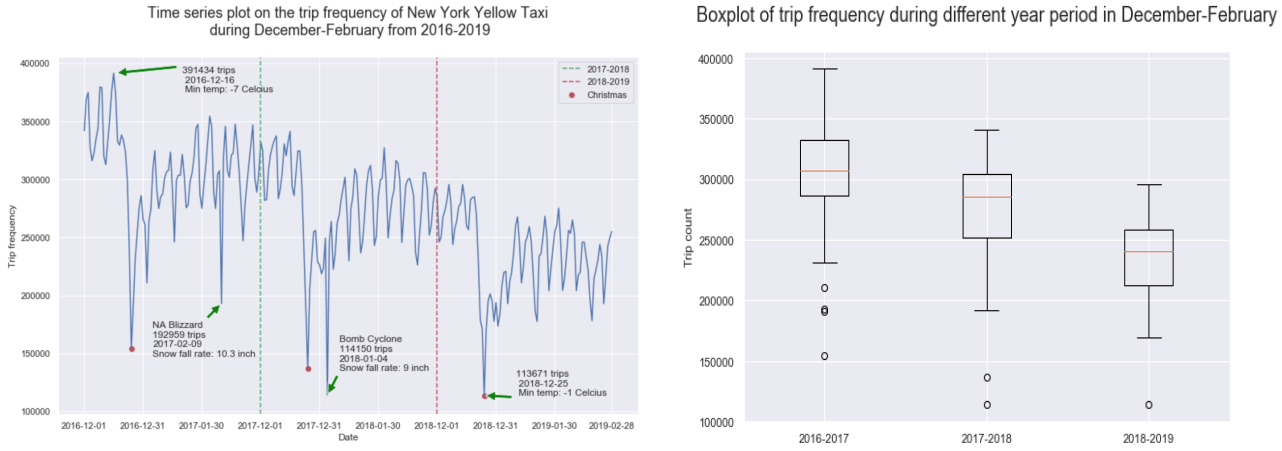
Figure 1: Correlation plot between selected attributes in TLC data

To begin with, an introduction correlation plot may help identify certain relationships in the selected attribute of this analysis. Figure one shows that the highest correlation of trip count or frequency has is with minimum temperature, which is around 0.15, followed by trip distance, that is 0.14. It turns out that snow fall and depth has a low correlation, with -0.05 and 0.08 respectively. However, this correlation plot is only a raw plot, as there may still be hidden pattern in the trip behaviour that may not yet be explained using this heat map. Henceforth, the visualisation analysis below will be split into three main section, that is trip frequency, day and hour, and trip distance.

3.1 Trip Frequency

Firstly, time series is a decent way to find pattern and trend in time series data. Since TLC data is a time series data, it may be helpful to plot a time series figure on the historical trip trend to visualize changes in trip demand from each period. Based on the plot below, there may exist a downward trend from 2016-2019. This might indicate that the demand for taxi trip has decreased yearly. The mean trip frequency in the 2016-2017 is 305,970 trips with a standard deviation of 41,946 trips. Over the year, both the mean and standard deviation of the trips declined, that is 234,296 and 34,863 taxi trips for mean and standard deviation respectively. None of these periods came from a normal distribution, as tested using Wilks test method; hence, it is decided not to proceed a Welch significance testing for mean differences. A lesser standard deviation might means that the market segmentation between Yellow taxi and high frequency for hire vehicle has become more mature, but there can be various reasons why this can also happened, and it is beyond the scope of this analysis. The maximum demand from 2016-2019 was in 16th of December 2016 where there was 391,343 trips recorded. The minimum temperature forecast at that time was around -7°C. The lowest demand was achieved during Christmas in 2018, with only 113,671 recorded trips, and the minimum temperature is at -1°C. Moreover, there are two interesting events that can be seen from this graph. First, the Christmas drop. These drops were found by using the whisker outlier detection after gazing at the boxplot in figure 2b. The data shows that in each year, the demand for taxi trips

during Christmas dropped drastically around 0.3-0.5% from previous day. The worst was experienced during 2016, with 0.35% dropped approximately. The second interesting dropped was the blizzard in New York, which is on 2nd of February 2019 and 4th of January 2018. The famous one is the "Bomb Cyclone" blizzard, 4th of January 2018, as the demand fell by 0.54%, that is around 61,594 less trips, greatest through out these three period. It may be the case that New York Yellow Taxi took a huge toll in 2017-2018 period, due to annual Christmas drop, and blizzard. The indication of a blizzard is the prediction of snow fall rate, as both has 10.3 and 9 inch prediction. Overall, one may infer that national holidays and heavy snow fall rate may caused an abrupt decrease in trip demand.



(a) Time series plot of trip frequency

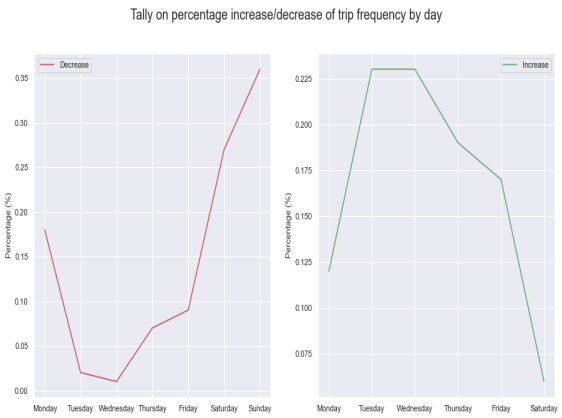
(b) Trip frequency box plot for each time period

Figure 2

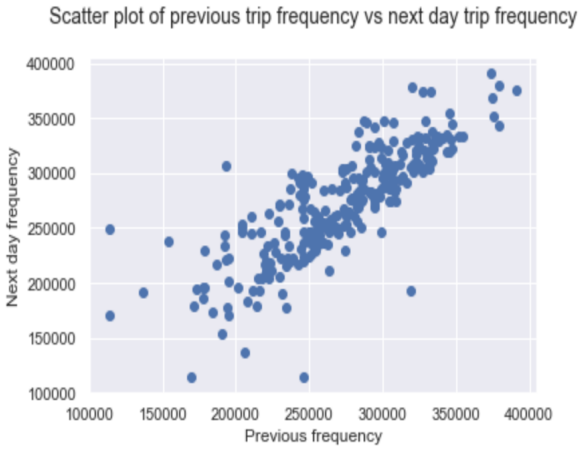
Aside from these outliers, the time series graph also depicts an oscillation motion in each year. This may imply that there may be pattern on which day people are more desirable to take taxi trips. Figure 3a below visualize the pattern on which day trip frequency may increase or decrease. Based on the graph, it can be seen that most of the time, trip decreased on Monday, Saturday and Sunday. The mode on the most increased trip is on Tuesday and Wednesday. However, this graph does not imply which day has the most trip on average, as it only breaks down the oscillation motion on figure one. More than 35% of the time trips decreases from Saturday to Sunday, and there is no data showing any increase of trips within these days; hence, trips demand on Saturday were always superior in comparison to Sunday. Friday and Wednesday were also superior most of the time compared to Thursday and Saturday, respectively. This oscillation pattern leads to another investigation, that is to test the correlation between previous day trip frequency with its next day, as depicted from figure 3b. The result shows a strong relationship, that is 0.81. This may indicate that an increase in yesterday's trip frequency linearly correlated to today's frequency. However, one may need to test further whether previous day trip frequency can be used to predict the next day frequency.

In addition, lower frequency of trip may not generally be correlated to lower temperature. A time series plot of the historical minimum temperature in figure 4a reveals that the maximum of the minimum temperature has a fairly higher trips than the lowest temperature by 86,061 trips. But, this statement can be argued, as that minimum frequency was also caused due to a blizzard. Additionally, figure 4b scatter plot also depicts that the correlation between trip frequency and minimum temperature is quite low, with only 0.16. By looking closely to the scatter plot below, is can be seen that there is one outlier in -5°C that pull out the deviation for points in that temperature. Furthermore, there is a wide range for trips in 0°C, that is a range of 264,846 trips, which shows the inconsistency on the trip demand. A more tight range are located at -10°below and 4°above, but there are less data points located in these areas as well. One could also see that the time the mean temperature may not decrease yearly as significant as the trip frequency; henceforth, this analysis shows that temperature may not have a significant impact in trip frequency, but further research still need to be done.

Previously, it is also found that heavy snow fall rate may caused a reduction in daily average of trip frequency. This finding may be supported by the graph on figure 5at, as the demand frequency dropped drastically when the snow fall rate is higher than 6 inch. However, the summary statistics shows that there is only two samples, which means there is only two days recorded with snow fall rate higher than 6 inch; thus, one may need more sample in order to support this result further. Nonetheless, even after considering its deviation from its mean, their trips are still far below the mean of other group. The most consistent group is achieved by the 2-4 inch with an interquartile, standard deviation and range of only 9,722, 9,689 and 24,375 trips, smallest compared to others. Moreover, the upward trend in the first three bar may not be accurate, as 0-2 inch statistics reveal that it may

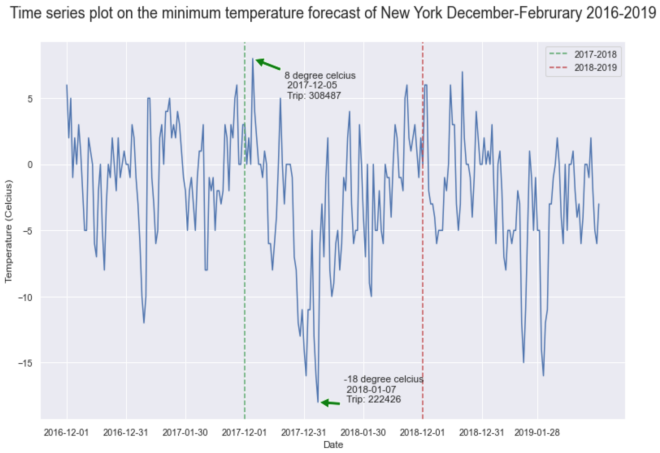


(a) Percentage increase/decrease of daily trip frequency

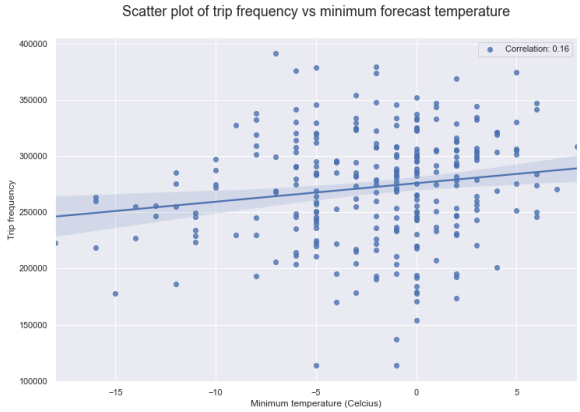


(b) Scatter plot of previous day trip frequency vs the following day

Figure 3



(a) Time series plot of minimum temperature

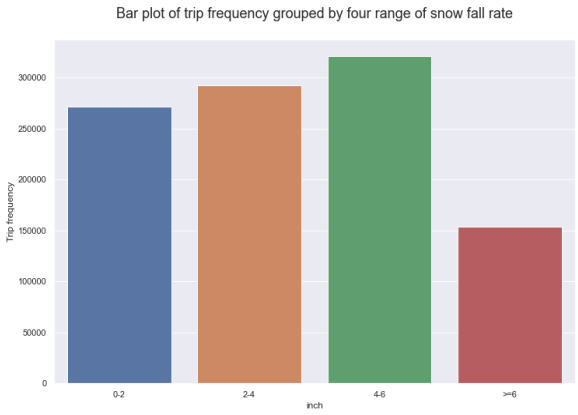


(b) Scatter plot of daily trip frequency vs minimum temperature

Figure 4

contain some outlier, as the there is a great distance between its mean and min value, and the min value is even smaller than the minimum value for the 6 inch. Overall, there might be some indication that heavy snow fall effect the trip demand, but the analysis may need more sample data to support this statement.

Lastly, one may need to investigate impacted trip demand areas around New York due to its past blizzard, and Christmas. From The choropleth in figure 6, it is visible that the trip demand in most area of the New York state decreased compared to the usual normal day. The data shows that the overall median trip decrease in each location is at 46.3%. Brooklyn suffered a median decrease trip by 50%, followed by Manhattan with 46.7%. However, the latter value may have bigger impact, as 91.5% of the yellow taxi cap demand on average are from Manhattan. Furthermore, the LaGuardia airport also experienced a huge fall, with 5,463 reduced in trip demand. This is quite detrimental for New York Yellow taxi, as pick up from this location ranked 15th on sorted trip frequency from highest to lowest during normal day. Despite this lost, there were several locations that experienced a higher demand rate than its usual, such as Cypress Hills, Roosevelt Island, and West concourse, and each with a total daily trip distance of around 60-100 miles. Interestingly, there is no access from Roosevelt Island to Manhattan using subway or other public transport, as the only way to get to the island is by walking, or driving. It may be the case that due to the extreme weather condition, people living in the island may prefer driving instead of walking to travel; thus, those that used to commute by walking may prefer taking a taxi cap instead. The trip pick up in this island increased by 70%.



(a) Trip frequency bar plot based on snow fall rate

	0-2 inch	2-4 inch	4-6 inch	>=6 inch
count	258.000000	5.000000	5.000000	2.000000
mean	271922.879845	292554.000000	321409.200000	153554.500000
std	48162.278837	9689.001445	36027.534771	55726.378319
min	113671.000000	283298.000000	285893.000000	114150.000000
25%	242323.500000	285089.000000	291094.000000	133852.250000
50%	274967.000000	291899.000000	318808.000000	153554.500000
75%	305422.750000	294811.000000	337415.000000	173256.750000
max	391434.000000	307673.000000	373836.000000	192959.000000

(b) Statistics of the bar plot on trip frequency in each snow fall rate group

Figure 5

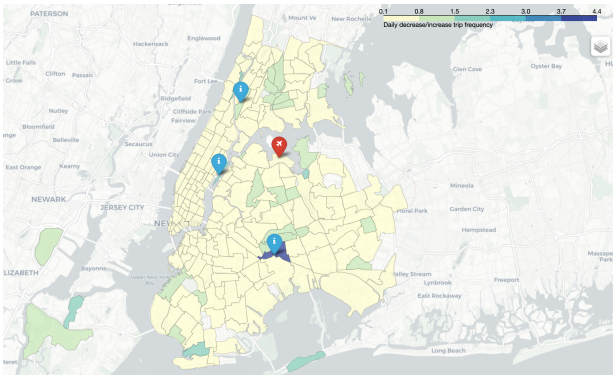


Figure 6: Percentage on daily increase/decrease of trips during blizzard

The trip demand behaviour during Christmas is quite different than blizzard. For this part, it is advisable to open the source code html file in the timeslider section for full experience of the analysis. The trip period that is used for the slider is during the 24th-25th of December 2018, and the focus point in this graph is at Manhattan, JFK and LaGuardia airport. On the 24th of December, most trips were happening around Manhattan, especially in Midtown, Upper East and West.Nonetheless, there is increase in trip frequency on both JFK and LaGuardia by 10.3% and 57% respectively. However, Manhattan was filled with trips again on Boxing day, 26thth of December, Madison Square dominated the pick up area in Manhattan, and pickup from JFK and LaGuardia surged again by 54.7% and 88.6%.

All in all, this section has demonstrated that trip frequency of New York Yellow taxi may be affected due to extreme weather condition, such as blizzard and national holidays. However, Snow and cold temperature in general may not give a significant impact in the frequency. The next section of the analysis will discuss more on the hourly and daily behaviour of the yellow taxi trip frequency.

3.2 Day and Hour

To begin with, it is hypothesised that trip frequency around New York ought to be affected by time and day, as time may be a good indication on when is a rush hour and when is not. Day may also reveals whether there might be different behaviour in trip frequency during different hours.

3.2.1 Day

To start of examining the daily behaviour, one may need to get the overall view on which day has the most taxi trip based on the data. The violin plot below demonstrated that there are some outliers in Thursday and Tuesday. This can be seen from the long tail of each violin. Henceforth, one may need to look at the median, and not mean. The median, the white dot in each violin, for Thursday and Friday is quite similar, with values around 300,000 trips. The lowest median is achieved by both Sunday and Monday, that is around 17% less trips than the median of Friday. Moreover, based on these violins, one can see that there is an upward trend from Sunday to Friday, followed by a downward trend from Friday and back to Sunday. In addition, Thursday, Friday, Tuesday and Wednesday has a wider violin compared to the other. This means the trip frequency located in the widest point of the violin have higher probability compared to the rest of the frequency in that violin. For example, historically, there is higher probability that in Friday, there exist around 310,000 trips than 250,000 trips. To sump up, the data shows that highest on average daily trip frequency is on Friday, while the lowest is on Sunday or Monday.

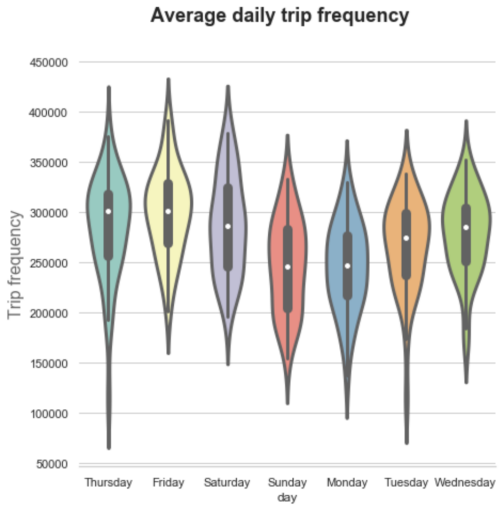


Figure 7: Violin plot of average number of trip frequency in each day

3.2.2 Hour

Aside from day, hour may also affect taxi trip frequency. Not only that, one may need to investigate whether snow may disrupt the usual behaviour of this taxi frequency. Thus, the figure below reveals that the trend between snow and no snow day is fairly similar. There might be a slight deviance at around 8-9 am, but this may not be significant. However, there is a clear difference in the hourly trend, as there were more trip demand on 6pm compared to those at 5am, with differences for more than 130,000 trips on average approximately.

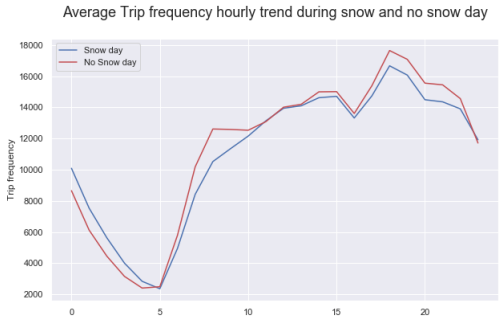


Figure 8: Hourly plot of trip frequency based on snow and no snow day

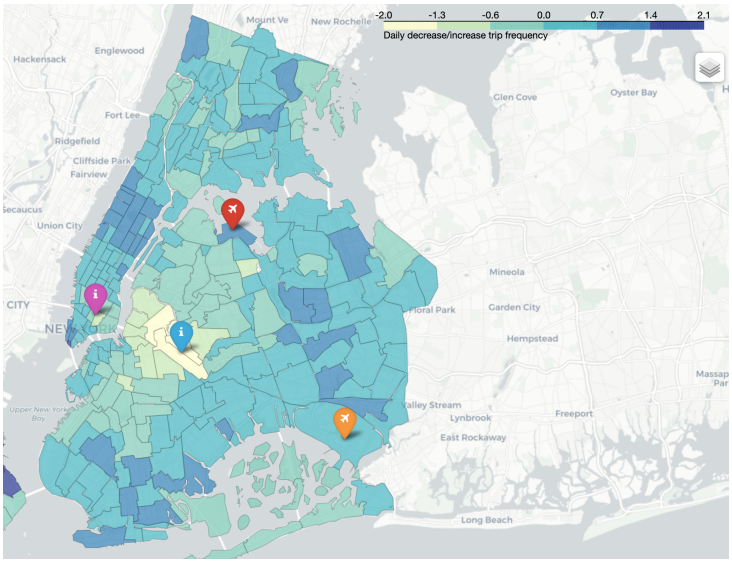
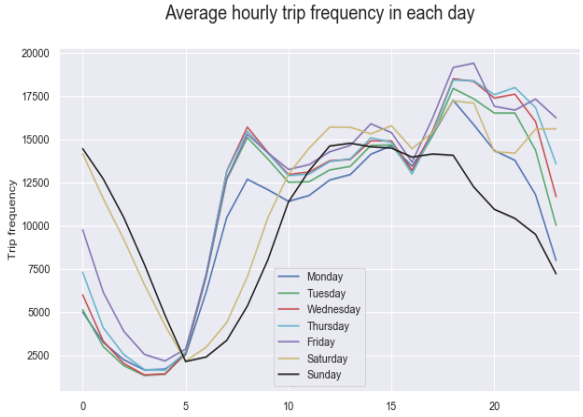


Figure 9: an increase/decrease on trip frequency choropleth based on working hour and non working hour. Red: LaGuardia, Purple: Lower East Side, Blue: Bushwick, Orange: JFK

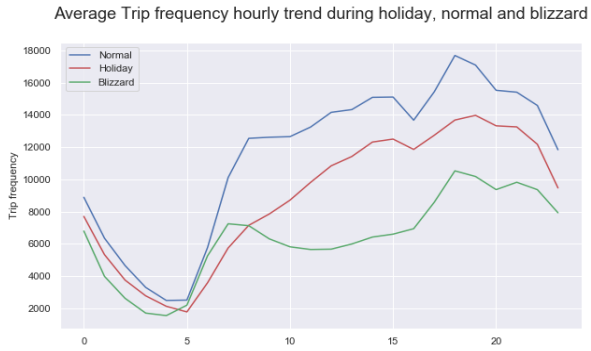
Furthermore, a choropleth of trip frequency rate of increase/decrease between working hour, 8am-19pm, and non working hour, 20pm-7am, is plotted to visualize how each location trip frequency is impacted by hour. However, before taking the difference of the trip frequency, the values are normalized using log transformation, as some areas are superior in its trip frequency compared to other areas in general. The plot shows that there are some specific areas that has more trips after working hour, colored in light green and bright yellow, such as Lower East Village, Greenpoint, Bushwick and Williamsburg. In contrast, New York Financial District and areas around the central park are colored dark blue. This is expected, as these are New York's famous business district area. In addition, it turns out that LaGuardia airport was has more trips during business hour compared to non business hour, while JFK is more neutral. It might be that business traveler may prefer to land on LaGuardia, as compared to JFK, since LaGuardia is closer to Manhattan by 10 miles. Ergo, from these visual aids, hourly trip frequency in each location may exist.

3.2.3 Combination of hour and day

Last but not least, one may also need to examine the hourly trip frequency behaviour of each day, and to investigate the effect of certain events that may disrupt the hourly trend. To answer this question, first, one may need to look at figure 10a. This plot display the overall hourly behaviour in different days. From the plot, it can be seen that there are three types behaviour here, weekdays, Saturday and Sunday. On weekdays, there are two spikes happening within the day, that is 8-9 am and 18-19 pm. However, on Saturday and Sunday, the spike moves from 8 am to 12 pm. Another spike is happening on 18-19 pm too on Saturday, but not on Sunday, as it has no more spike based on the historical data. But, both Saturday and Sunday has higher trips at 0-4 am, and this may indicate New York night life behaviour. The maximum average trip in this chart is from Friday at 7 pm with 19,431 trips recorded. The minimum frequency is at 4-5 am for every day with less than 2,500 trips. Moreover, the figure on the right demonstrate that trip frequency in different events indeed showed major differences, especially in the range of 9am-20pm; Nonetheless, it appears that mostly, they have similar hourly pattern, except for 7am to 16pm during blizzard, as it forms a curve instead of a linearly straight line. The graph also shows that there is an intersection at 4am between holiday and blizzard, which indicate that there are more trips on 4am-9am during blizzard compared to holiday, with average trips of 4,668 and 4,069 trips, respectively. All in all, this analysis has demonstrated that frequency trip behaviour may be affected by both hour and day, as they may be highly correlated with the daily cycle of people in New York.



(a) Daily and hourly trend of trip frequency

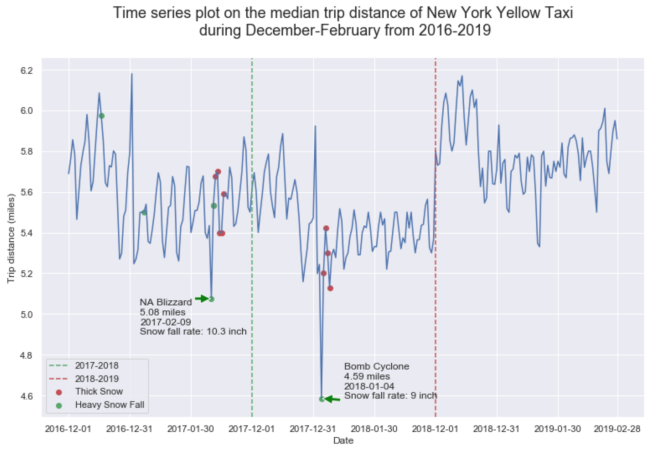


(b) Trip frequency hourly trend based on specific event

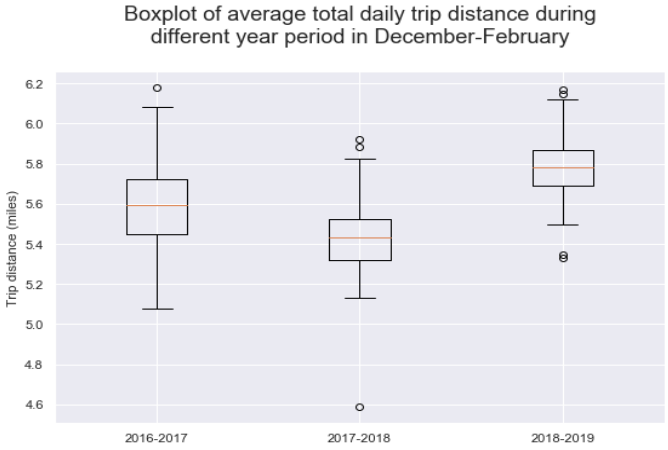
Figure 10

3.3 Trip distance

Finally, the last attribute that is going to be analysed is trip distance. Visually, from the figure below, it is plausible to say that average trip distance of taxi trip may increase in 2018-2019 period. This notion is supported by the boxplot figure on the right, as the median of 2018-2019 trip distance is at 5.8 miles, compared to the first two period, that is 5.6 and 5.4 miles respectively. Since the boxplot indicate that the first and last period may have a decent bell curve shape; thus, a normality test using wilk test was conducted. The result of the test revealed that both may have a normal distribution. As a result, a welch t-test was done with different standard deviation. The result has a low p-value, that is around 1.41e-11; ergo, there may be enough evidence to reject the null hypothesis, which means these period may have different mean value. The time series plot also reveals parts in which there was a high snow fall rate and thick snow, and it can be seen that there may be some pattern in the plot. Average trip distance did drop when there is a heavy snow fall rate, defined as bigger than 5 inch, and both of these were caused by a blizzard. A few days after the blizzard, snow depth increased; however, these snow depth may not impacted much on the median trip distance of the taxi trip, as shown by the sideways trend. After the blizzard incident, the trip distance is slowly recovering, but still in a oscillation motion. Moreover, the outliers in the boxplot also help in detecting an interesting pattern, as every year, around the 12^{th}-16^{th} of December, the median trip distance has its all time high during December-February.



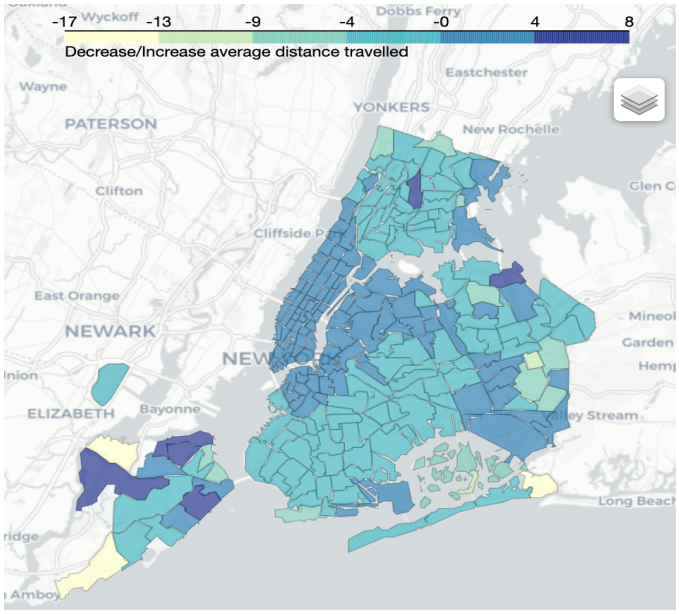
(a) Median trip distance time series plot



(b) Boxplot on trip distance in each period

Figure 11

Furthermore, one may also need to check the behaviour of the median trip distance and trip frequency when there is a thick snow, above 5 inches. The plot figure below visualize the number of daily trips and its median trip distance during thick and less snow, after some outliers have been sliced. On this plot, most of the snow points are located at the higher part of the plot. Even though one may not see a clear separation between these points, this may still indicate that during thick snow, there might be more short distance trip. The correlation result on the trip frequency and median trip distance travelled is 0.38 during snow, but only 0.05 when there is no snow. This may indicate that the relationship between average trip distance and trip frequency becomes more apparent during high thick snow. To further support this statement, the choropleth figure below shows that when the snow depth is bigger than 5 inch, that is few days after blizzard, areas around Southern part of Brooklyn and Bronx experienced a decrease in median average distance of each taxi. This may indicate that people may chose to do short distance travel instead of long distance, as they might still be recovering from the storm. In fact, the data shows that all five borough had a decreased in average median trip distance in each trip, with the worst impacted is Bronx with around minus 1.7 miles on average. This correlate with our previous findings in trip frequency, that during heavy snow fall rate, trip frequency also decreased significantly. Moreover, most of the second dark blue colored areas in the plot are business district of New York, such as Manhattan and Downtown Brooklyn. However, not all of these areas experienced an increase in the trip distance median, as some has values of -0. decimals. There is also an increased trip distance travelled in airports, that is JFK and LaGuardia, but it only increased by 0.04-0.06 miles. Also, areas near the business district were affected as well. Interestingly, there is a single outliers in Bronx, that is the Bronx park, with an increase of median trip distance by around more than 4.5 miles; however, this may not be accurate, as the frequency is really low, with only six trips recorded during the past three period, compared to less snow depth with 200 trips. This may also indicate that Bronx Park may not be a popular place for a taxi pickup. This issue also happened in Staten Island, as generally, there is not much of a taxi frequency pickups happening in the island.

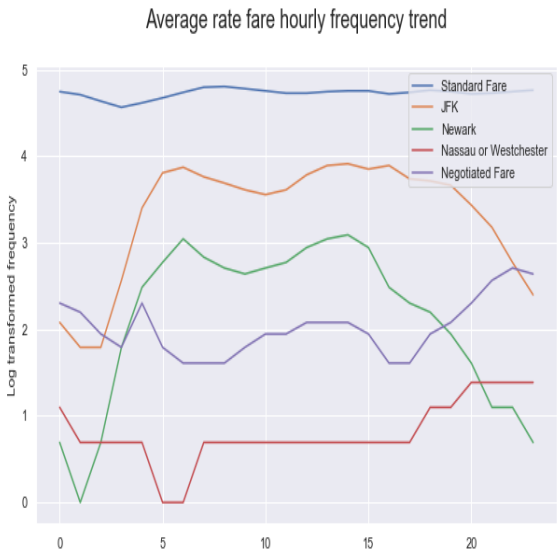


(a) Average trip distance decrease/increase during thick snow

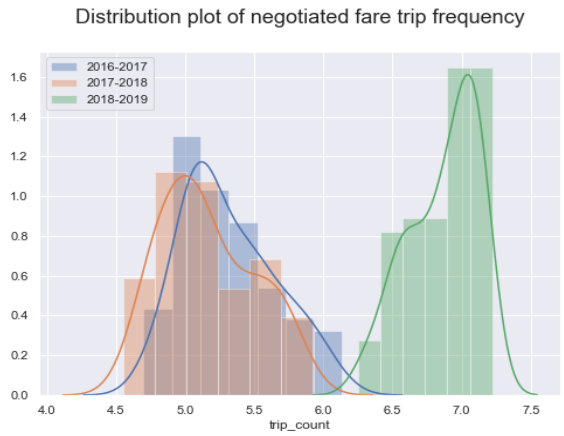
(b) Scatter plot of trip frequency vs average trip distance during normal and thick snow

Figure 12

Unfortunately, this finding may contradict on the time series plot, as there may be evidence that average daily trip distance increase in the last period, but there was actually less snow fall rate in 2018-2019. Thus, to investigate this matter, one may need to analyse the rate code trend. After much investigation, the data shows that rate code id 5, negotiated fare, sky rocketed by more than 450% from the previous period. The distribution plot below also indicate a major deviation of negotiated fare trip frequency in period three from its peers. This rate code id applies to passenger that travels outside the city of New York. The median trip distance for rate code id 5 is around 10 miles, compared to the standard meter fare, rate code id 1, with median of 2.63 miles. Thus, this abrupt surge in negotiated fare demand might alleviate the median daily trip distance of the taxi trip in the last period.



(a) Hourly trend of each taxi rate type



(b) Distribution plot of negotiated fare daily log normalized trip frequency in each period

Figure 13

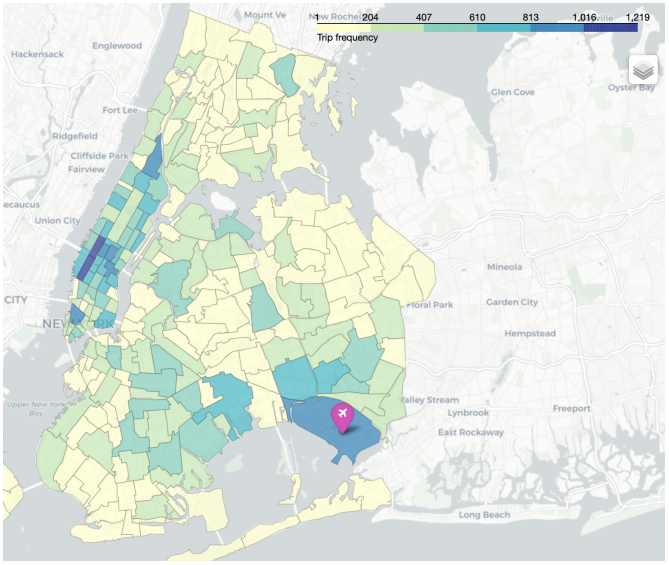


Figure 14: Choropleth of negotiated fare trip frequency in each location id

Plot 14 visualize last period's historical trip frequency demand for negotiated fare, and most of them are located in Clinton East and East Chelsea. JFK is actually one of the negotiated fare source, as it is placed near the east border of New York state. Interestingly, there is also a huge demand in Central Harlem compared to the area around this location. Also, for the northern border, the record was hold by Co-op city in Bronx. Moreover, this rate trip distribution may also be affected by hour, as seen from figure 14. Here, one can see that travelling to JFK or from JFK dropped from 16pm on wards, and plunge back up at 2 am in the morning. This trend is followed by Newark as well, but with less frequency. Furthermore, negotiated fare shows a sideways pattern throughout the day, but it has a sturdy upward trend from 18pm above. Nonetheless, the daily rate trend for all rate code id is fairly constant; hence, the matter will not be investigated further.

Ultimately, this part of the visualisation section has depicted a change of trend in the median travel distance of the New York Yellow Taxi trip in the last period. The analysis also showed that during higher thick snow, the correlation of average trip

distance and trip frequency are more significant, and that these caused a decrease in the average trip distance travelled for the taxi across New York. Moreover, there may be a new market opportunities for yellow taxi driver to take rate 5, which is the negotiated fare, as more passengers in New York were eager to take this rate.

4 Section 4 - Discussion

This discussion section ought to review the hypothesis that was stated in the previous section. First, by looking at the trip frequency visualisation, there may be evidence historically that heavy snow fall rate may reduced the trip frequency of New York yellow taxi. This was actually caused by two blizzard that occurred in New York on the 9th of February, 2017 and the 4th of January, 2018. The blizzard, especially Bomb Cyclone, did wreak havoc all over New York, as seen from figure 6, causing most schools to be closed and cancelled flights in JFK (Steinbuch, 2018). Hence, people may prefer to stay at home, as the weather were quite extreme, which is -18°C. Nonetheless, in general, mild snow fall and low temperature may not decrease the trip frequency, as shown from the histogram table in figure 5a. The trip would decrease when the snow fall rate is higher than or equal to 6 inches, which is quite rare from these three chosen period. The scatter plot in figure 4b also indicates that the trip frequency may vary when the temperature is not extreme. This might be due to New York as being one of tourist's favorite place to visit, especially during winter to see the snow and various attractions. However, every year during Christmas, the frequency dropped drastically. This is expected, as during Christmas eve, Upper East, West and Midtown are filled with fancy restaurants and various attractions. Not only that, people may be doing their grocery shopping too, as most restaurant, business and stores are usually closed during Christmas in New York (Stack and Victor, 2016). 26th of December is Boxing day in America, and it is customary for stores to declare huge sales on their items; thus, this may explain why the trips surged again. In addition, there are some interesting additional findings that may not be related to the hypothesis, that is the correlation between previous day's trip frequency and the following day. One main reason why this might happen is due to the cycle that trend that is found in the day analysis, that is trip frequency increase linearly from Monday to Friday, followed by a linear decrease afterwards. Ultimately, the appearance of snow fall and low temperature may not result in the declining taxi trip frequency, but a severe snow fall, temperature and Christmas did make a significant impact in the changes in trip frequency behaviour.

In addition, there may be evidence historically to show that there exist a trend and cycle of trip frequency behaviour in each day and hour. From the second section of the visualisation, one can see that two days before entering weekend are the busiest time for a taxi driver, and not weekend. Not only that, figure 7 is also a small part of the cycle found in the time series plot from trip frequency section, and this prove the existence of the weekly oscillation motion; hence forming a cycle for the trip frequency in each period. Furthermore, hourly trend of the trip can be seen in figure 8, and the trend did match the general normal working hour. There may not be enough evidence to say that the trend change during regular snow fall. Nevertheless, there may be different trends during holiday and blizzard. By looking at figure 10b, the holiday line shows that people in New York tend to flatten their 8-9 am activities, and during blizzard, less people are taking trips from 7 am on wards, as most places were forced to cease activities due to the weather condition. Overall, taxi driver could still expect a peak at around 6-7 pm, regardless of the condition, as this marks the end of normal working hour. Not only that, by looking at figure 10a, weekdays and weekend tend to have different hourly trip frequency pattern. It is interesting to see that most night life in New York usually peaks at Saturday and Sunday, and not Friday. Hence, taxi driver could try to find extra income during midnight at Saturday and Sunday in areas around Lower East Village, Greenpoint, Bushwick, and Williamsbug, where there are a considerable number of bar and night life venues (Hickey, 2020). Moreover, Bushwick and Greendpoint are also known for their hippie culture and night music festival (New York City's Hippest Neighborhoods Right Now, 2020). To sum up, different cycle of trip frequency behaviour exist in each location id, both hourly and daily, and regular snow fall and depth may not disrupt this cycle.

Furthermore, the trip distance visualisation part demonstrate that the average trip distance did decrease when the thickness of the snow is beyond 5 inches. The choropleth plot in trip distance visualize the impacted parts of New York, and it can be seen that mostly, the suburb part of Brooklyn and Bronx has lower average trip distance compared to other areas. Evidently, the thick snow in Brooklyn reach to 7-12 inches, higher compared to Manhattan, with 4-7 inches (Tata, 2018). As a result, it may be dangerous for car to travel in Brooklyn; hence, this may explain the decrease in average trip distance. Finally, for the last hypothesis, average trip distance may have higher correlation to daily trip frequency during thick snow, but not when the thickness is below 5 inches. Aside from these hypothesis, the data also indicate that there is an increase in trip distance in the last period, and this has been tested using welch's significant test. One reason that may explain this phenomenon is due to the rise of negotiated fare demand, that is trips that go beyond the city of New York (Taxi Fare - TLC, 2020). Detecting the rise of this fare may be of interest for taxi driver, as on average, the rate per miles for negotiated fare are higher by $1.96 compared to the standard fare. Another forte for this fare might be the ability for the driver to negotiate the fare; thus, the amount received by the passenger also depends on the negotiation ability for the driver. These demand may be found in areas around Clinton East

and East Chelsea, as they are located near the border of another state of New York and Newark New Jersey, and it is estimated that based on past data, the demand has its maximum height at 9 pm above. All in all, the average trip distance travelled might decrease when the thickness of the snow is beyond 5 inches, and trip frequency may not have a strong correlation to average trip distance, unless the snow thickness reach above 5 inches.

Identifying these behaviours from the historical data are quite significant, as the trend for the trip frequency has become for apparent, and taxi driver may be able to use this to increase their chances in getting more passenger. Driver may also be able to plan ahead on the trip destination ahead that may maximise their earnings daily, as the analysis has identified the hourly demand trend for certain locations. Last but not least, in times of blizzard or Christmas, driver may know now which part of New York that still has steady demand for a ride. The result from this analysis may not be true for conditions that are out of the scope from this analysis, as this result is based on winter period in New York, specifically December - February.

Ultimately, the next step from this exploratory data analysis is to try predict next day's trip frequency. This might be done by using decision tree regressor, or linear regression. Linear regression might be applicable in this case, as the daily frequency trend is linear based on the data, and the analysis has shown that previous day trip frequency do have a strong linear relationship to the following day's trip frequency. The reason why decision tree regressor might be better, as several discrete attributes, such as day and hour, may be better learned using the tree structure, as tree are known for learning categorical and discrete values well. One might also use day and hour in linear regression, but this will turn the model into a less than full rank model.

5 Conclusion

In conclusion, the analysis has shown New York yellow taxi trip behaviour during winter through historical trip frequency and distance data, both hourly and daily trend. The result indicate that heavy snow fall rate, snow thickness beyond five inches, and Christmas may disrupt the usual trend for usual trip behaviour. However, low temperature may not decrease trip frequency. Moreover, daily and hourly cycle exist in trip frequency, and that each location in New York may have specific hourly trend for trip demand.

References

[1] Guse, C., 2020. Driving NYC Taxis Out Of Business: How Uber And Lyft Doomed The Once-Solid Yellow Cab Industry. [online] Nydailynews.com. Available at: <https://www.nydailynews.com/new-york/ny-medallion-foreclosures-taxi-bailout-plan-uber-lyft-20200130-s2mjkhjubzgptdxasoxddwdote-story.html> [Accessed 3 September 2020].

[2] Weather Atlas. 2020. The Climate Of NYC. [online] Available at: <https://www.weather-us.com/en/new-york-usa/new-york-climate#:~:text=When%20does%20it%20snow%20in,December%20are%20months%20with%20snowfall.> [Accessed 3 September 2020].

[3] Www1.nyc.gov. 2020. TLC Trip Record Data. [online] Available at: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page> [Accessed 3 September 2020].

[4] W2.weather.gov. 2020. National Weather Service Climate. [online] Available at: <https://w2.weather.gov/climate/xmacis.php?wfo=okx> [Accessed 3 September 2020].

[5] Www1.nyc.gov. 2020. Taxi Fare - TLC. [online] Available at: <https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page> [Accessed 3 September 2020].

[6] Weiss, M., 2012. Fastest Road In America Maximum Speed Limits In New York. [online] Weiss Associates, P.C. Available at: <https://nytrafficticket.com/fastest-road-in-america-and-maximum-speed-limits-in-new-york/> [Accessed 29 August 2020].

[7] Beta.costtodrive.com. 2020. Tolls On Popular Bridges And Tunnels. [online] Available at: <https://beta.costtodrive.com/tolls-on-popular-bridges-and-tunnels/> [Accessed 29 August 2020].

[8] Stack, L. and Victor, D., 2016. Christmas Holiday Hours For Some Major Stores. [online] Nytimes.com. Available at: <https://www.nytimes.com/2016/12/23/business/christmas-holiday-store-hours.html> [Accessed 3 September 2020].

[9] Hickey, W., 2020. The Best Bars In New York City Are In These Neighborhoods. [online] Business Insider Australia. Available at: <https://www.businessinsider.com.au/map-the-best-bars-in-new-york-city-are-in-these-neighborhoods-2013-10?r=US&IR=T> [Accessed 29 August 2020].

[10] Beyond Times Square. 2020. New York City's Hippest Neighborhoods Right Now. [online] Available at: <https://btsq.com/new-york-citys-hippest-neighborhoods-right-now/> [Accessed 29 August 2020].

[11] Steinbuch, Y., 2018. Temperatures Plummet Day After 'Bomb Cyclone'. [online] Nypost.com. Available at: <https://nypost.com/2018/01/05/temperatures-plummet-day-after-bomb-cyclone/> [Accessed 29 August 2020].

[12] Tata, S., 2018. 'Bomb Cyclone' Snowstorm Could Dump 12+ Inches On New York Region, Whip Up Blizzard Conditions. [online] WPIX. Available at: <https://www.pix11.com/2018/01/03/bomb-cyclone-monster-snowstorm-could-dump-12-inches-on-parts-of-tri-state-area/> [Accessed 29 August 2020].