The impact weather conditions on trip distance, tip amount and spatial distribution of destinations

Ze Pang
Faculty of Data Science
University of Melbourne
Student ID: 955698

September 12 , 2020

ABSTRACT

This paper focuses on the impact of weather in terms of temperature and precipitation on the taxi demand, trip distance, tip amount and spatial distribution of trip destination in New York City. The main idea is to help to analysing passenger behaviour and make decisions for stakeholders in different weather conditions. In this paper, the dataset of New York City Taxi and Limousine Service Trip Record Data (TLC dataset) has been used for analysis and visualised using a set of graphs to demonstrate whether different temperature and precipitation conditions can influence the user behaviour. To the end, a result of temperature can affect total taxi demand and passenger decision on the length of trip and tip amount, whereas the precipitation does not seem to have a solid correlation with total taxi demand and other passenger behaviour. Moreover, both of those factors are not likely to have an effect on destination locations. As a result, further decisions and models can be built base on the temperature to determine the passenger habits and thus improves efficiency and profitability for the taxi industry.

Section01 - Describe Data

Since the main objective of this paper is to investigate the relationship between weather conditions and taxi trip features, the weather dataset from the OpenWeather website has been used apart from the TLC Dataset. This dataset contains 25 attributes in total describing different atmospheric conditions and some other quantitive measures like sea level for each hour from 2016 to 2019 in New York City. It has two temperature-related columns - max_temp and min_temp which stand for maximum and minimum temperature at that moment and several precipitation-related columns like volume of rain and snow for past one and three hours. For this analysis, the granularity of time has been set to one hour, as such two columns have been chosen for analysing the precipitation condition are rain_1h and snow_1h.

For the TLC data, the yellow taxi data for all months in 2016 and 2019 have been chosen. The reason for choosing 2019 in the analysis is it contains taxi data for all months and also the most recent. Similarly, 2016 contains data for all months and has a 2-year gap with 2019, which makes it more comprehensive and less likely to suffer from biases. The three attributes of interest are trip distance, tip amount and spatial distribution of destination. The motivation for picking these three attributes is that these are likely to be affected by weather conditions. It is more likely to make these assumptions. The first one is the worse weather conditions may affect the length of trip, or more passengers tend to go to the further destination by taxi instead of public transport. The next one is taxi driver may receive more tips under bad weather conditions, and the last one is the destination maybe some certain places under certain weather conditions, like going to parks or beaches during good weather.

Section02 - ETL

For the weather dataset, in order to get an estimate of temperature in each hour, a new attribute namely avg_temp which stands for an average of both maximum and minimum temperature has been derived using those two columns. Also, a new attribute tot_prec, which means the total amount of precipitation has been derived by using rain_1h and snow_1h columns. After investigating the sliced dataset with only three attribute - datetime, avg_temp and tot_prec, there is a huge amount of missing values have been found. A valid assumption is all the missing values mean there is no recorded raining or snowing during that specific time due to there is no 0 value in these columns; therefore all the missing values are replaced by 0.

For the TLC data, since all datasets in 2019 for each months have been used throughout the whole analysis, the selected attributes are tpep_pickup_datetime, trip_distance, DOLocationID and tip_amount. The datasets in 2016 are not going to analyse the spatial distribution of drop off locations; thus the features selected are the same as in 2019 except the DOLocationID. All the rows with missing values in TLC dataset have been deleted because each of those contains a huge amount of data and will not be affected.

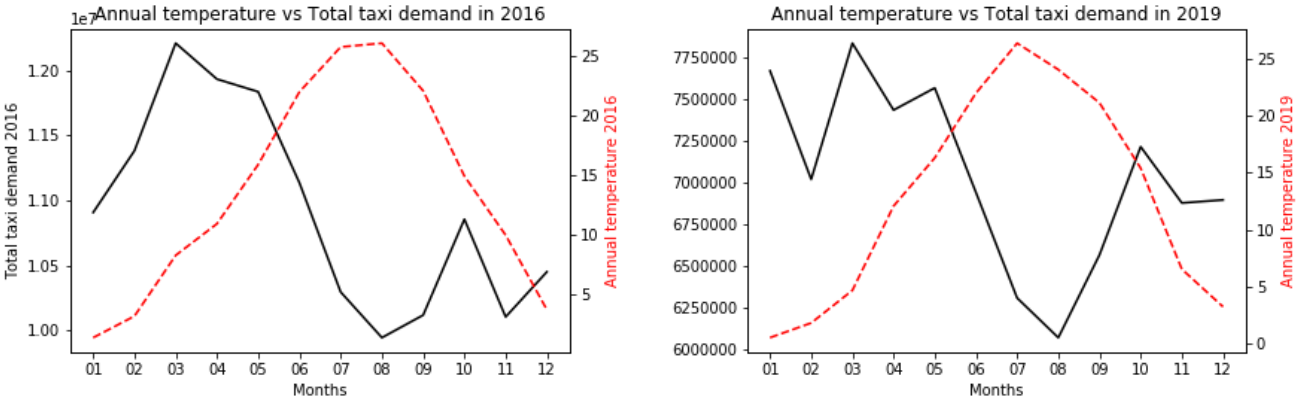
The dataset for weather data is relatively clean compared with the TLC dataset. There is no need to drop any outliers or solve any anomalies apart from select rows with right time. However, there are many problems in the TLC data. Taking yellow taxi data in 2019 as an example, it contains a large number of records which did not happen in this year, instead, it has a large proportion of records in 2018 and surprisingly two records in 2088. Thus all the incorrect data has to be sorted out, and the same procedure also needed to be carried out in 2016 taxi dataset.

There are some anomalies in a subset of 2016 yellow taxi trip dataset. All the datasets after June in 2016 have a different format with those in the first six months. The columns names in those incorrect datasets are shifted backwards two positions, i.e. the first two columns become the index, and the last two columns contain only nan value. This needed to be solved by firstly keeping the original column names, resetting the index, dropping last two empty columns and finally reassigning the column names.

Moreover, all the TLC trip datasets have to be transformed into feather type due to the csv file is low weight and fast serialisation for a large dataset (lab01). However, the weather dataset is much smaller than trip data and no need to be transformed into feather type.

Section03 - EDA

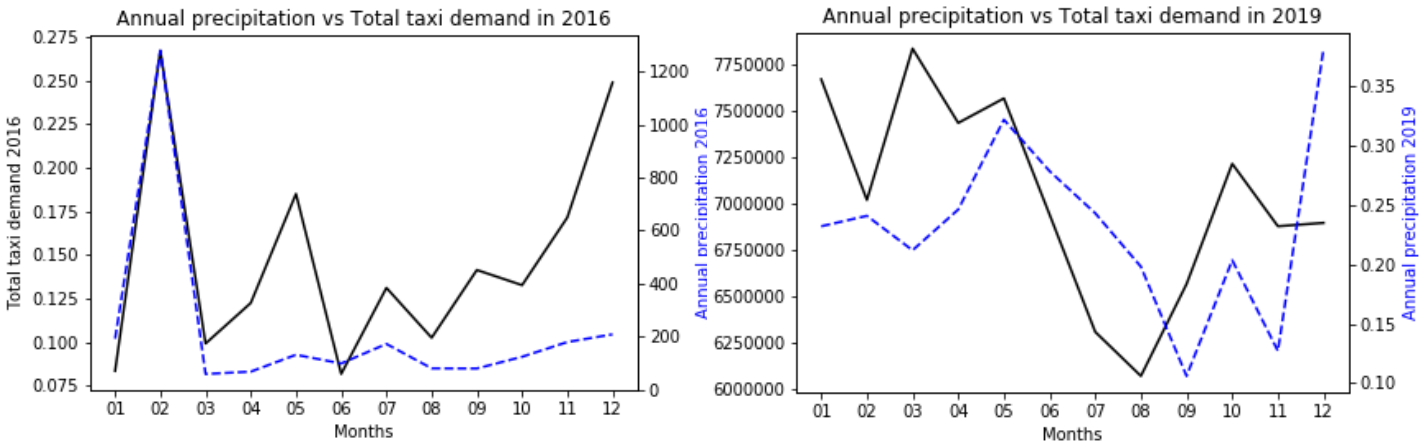
Before analysing how do weather conditions can affect these three attributes, a reasonable question can be raised is that whether temperature and precipitation can affect total taxi demand or is it logical to consider whether is actually a predictor for passenger behaviour? Therefore it is necessary to first investigate whether the temperature and precipitation can be useful for making decisions in taxi industries.



(Figure 01 - Two line plots showing the relationship between annual temperature and total taxi demand for each month in 2016 and 2019.)

From the line plots shown in figure1, it is worth noting that the total taxi demand seems to have an inverse trend with annual temperature in both 2016 and 2019. The total taxi demand has peaked in March and started to drop in summer, which is expected due to high temperature will impel fewer travellers or commuters. After

summer the trend starts to rebound and reaches a local maximum in autumn. It finally falls down in November.

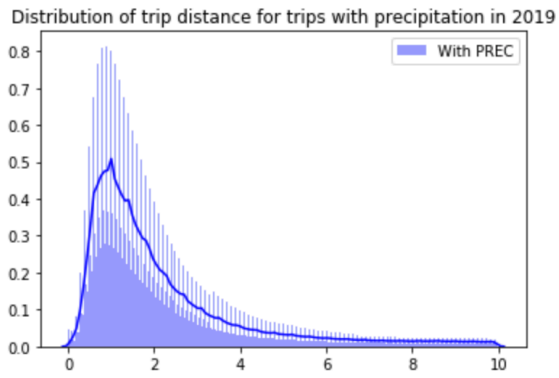


(Figure 02 - Two line plots showing the relationship between annual precipitation and ratio between number of orders during rainy/snowy and total taxi demand for each month in 2016 and 2019.)

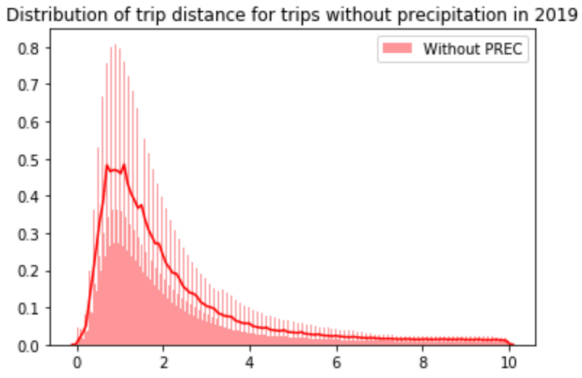
After having an idea of temperature may have an inverse impact on the taxi demand, figure2 shows how does precipitation influence the taxi orders. In order to reduce the influence caused by the difference for the total number of taxi demand in each month, the total taxi demand has been replaced by the ratio between a number of taxi demand during precipitation conditions and total demand in all weather conditions for that month. For these two plots, there is some trend that matches, like the first three months in 2016, the ratio has the same pattern as precipitation. However, there are some differences which could not be ignored; for instance, the ratio has reached its two maximum in May and December whereas the amount of precipitation reaches its maximum in March and October. As a result, precipitation condition may affect the total taxi demand, but taxi demand is more likely to have other factors other than precipitation.

After being aware of two weather predictors can possibly influence the taxi demand and not totally unrelated, it is reasonable to investigate do the weather conditions can have an impact on the trip distance.

(Figure03 - Two distribution plots of trip distance with different weather conditions in 2019.)

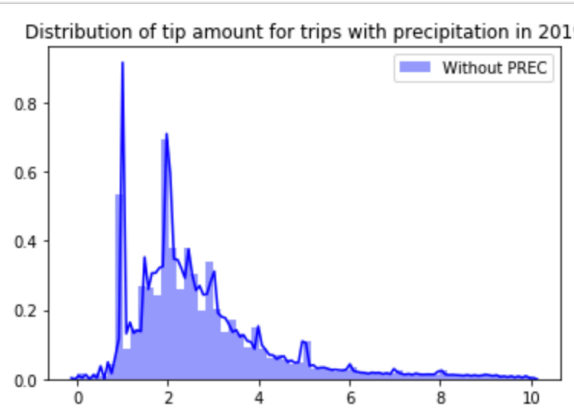



```
mean of trip distance with PREC: 2.17
variance of trip distance with PREC: 3.58
```

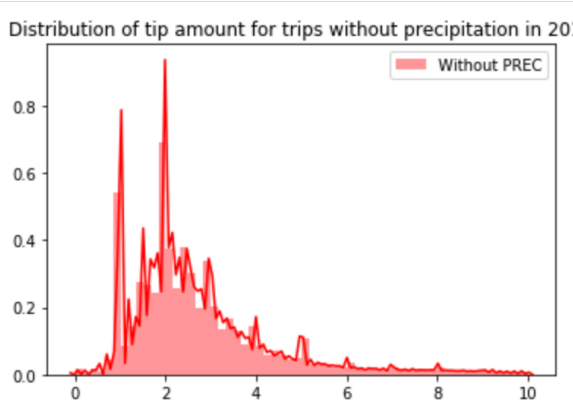



```
mean of trip distance without PREC: 2.18
variance of trip distance without PREC: 3.56
```

All the plots and analysis below only use yellow trip data in 2019 due to the results are similar in both years and the conclusion is representative because of the dataset is already vast. Figure 3 has indicated that the distribution of trip distances for all months in 2019 looks similar, all of them are skewed to the right and looks like normal distribution on the left. The statistics given also looks very similar, the mean of all trip distance under rainy or snowy conditions is 0.01 lower than another, and the variance is 0.02 higher than those in sunny days. Therefore, the precipitation could hardly affect the trip distance overall.




```
mean of tip amount with PREC: 2.77
variance of tip amount with PREC: 2.67
```

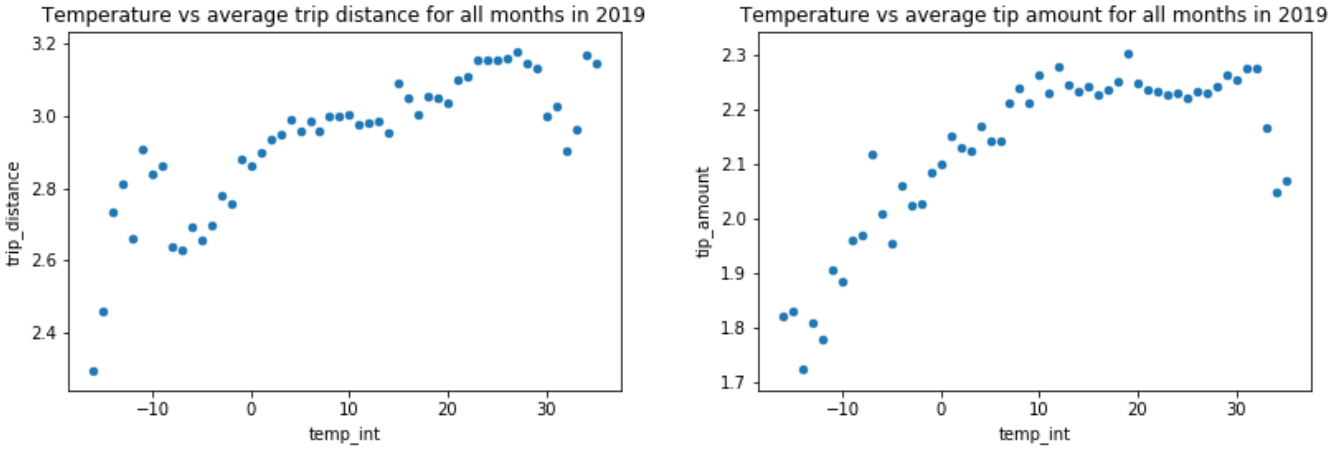



```
mean of tip amount without PREC: 2.76
variance of tip amount without PREC: 2.66
```

(Figure 04 -Two distribution plots of tip amount for each trip with different weather conditions in 2019.)

Apart from there is a not strong relationship between precipitation and distribution of trip distance, the tip amount also can be hardly affected by the precipitation conditions. From the figure 04, the distribution of tip amount in different weather conditions seems to be similar, both with two peaks around 1 dollar and 2 dollars except the trip in sunny days have more tips around 1 dollar than another one and fewer tips around 2 dollars. However, both means and variances are very close to each other with only 0.01 difference.

On the other hand, the temperature seems to have a better correlation with trip distance and tip amount. From figure 5, some interesting observations can be found in two scatter plots.



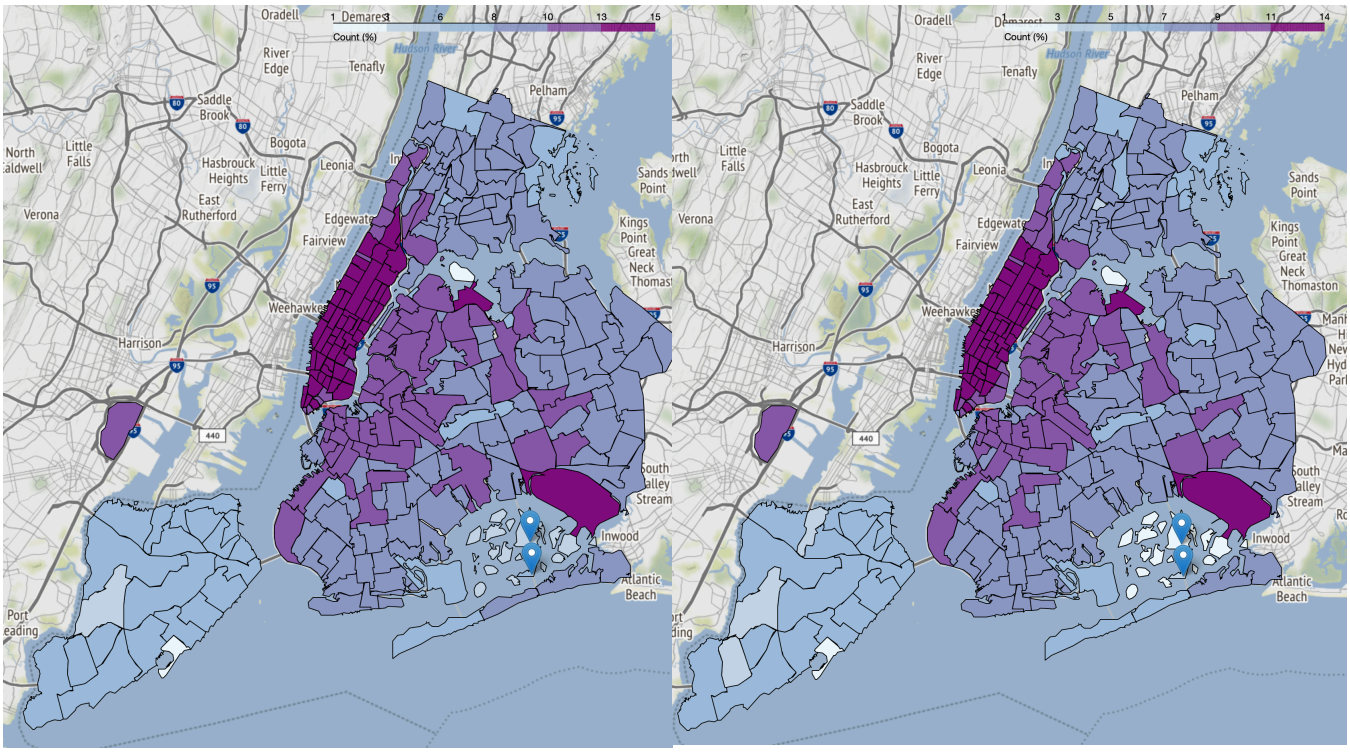
(Figure 05 - Two scatter plots showing how do average trip distance and tip amount change with the temperature.)

Both trip distance and tip amount is likely to have a linear relationship when the temperature. Although there are some points around -10 Celsius degree and greater than 30 Celsius degree have a different pattern than the other, the overall trend is linear, which means it may conclude that the warmer the temperature, the distance of trip will increase. For another one, the temperature has a clear linear relationship before 10 Celsius degree. After that, the trend becomes flat and have dropped after 30 Celsius degree.

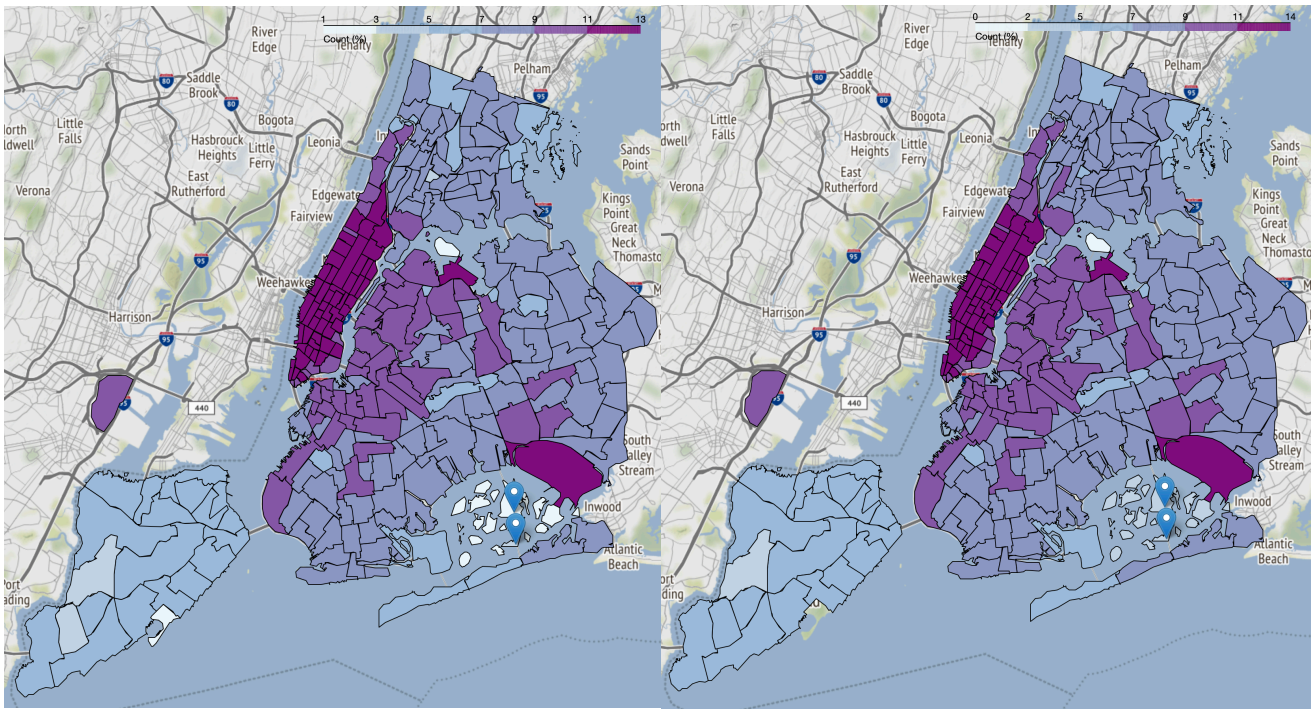
Finally, the last objective of this paper is to analysis the geospatial distribution for drop-off locations in different weather conditions. For the purpose of revealing geospatial distribution for those places without too much data and do not skew to some regions with too much data recorded, like JFK airport, a log scale

transformation has been used. Figure 6 has shown that most of the regions have similar drop-off data recorded. Except for some islands marked in the plot.

(Figure 6 - Two geospatial plots showing the distribution of drop-off locations in NYC with left shows the trip in sunny days and right in rainy/snowy days)



(Figure 6 - Two geospatial plots showing the distribution of drop-off locations in NYC with left shows the trip with high temperature and right with low temperature)



For analysing the different temperature could affect the geospatial distribution of trips, the temperature for each hour for all months in 2019 have been sorted and divided into four parts, then choosing the top 25 per cent data and bottom 25 per cent data joining with yellow trip 2019 datasets to plot the graph. Similar to the precipitation, figure 7 indicates that temperature also seems to have a tiny changes to the distribution of drop-off locations. The same location marked in the plot as the previous one has been found that it has different trip data when the temperature varies. The name of this place is Jamaica Bay Wildlife Refuge after investigation. A reasonable assumption is this place may be home for some wildlife, and those wildlife may come to those place in certain weather conditions.

Section04 - Discussions

As indicated above, the precipitation condition seems to have limited influence on the trip distance, tip amount and geospatial distribution of drop-off locations. Temperature as a predictor may be helpful for predicting the trip distance and tip amount, but same as precipitation can hardly influence the destination locations. Moreover, since the temperature is likely to have a linear relationship with the trip distance and tip amount, a generalised linear model can be built in the further analysis in order to make predictions on average of trip length, and the average tip can drivers receive during the weather with a specific temperature. This may be extremely useful for improves efficiency and profitability for those taxi companies, for instance, adjust the current price if this year is more likely to have a higher temperature than the previous year. In addition, this analysis can also be helpful for other cities or other countries with similar weather conditions; the model built using New York City TLC data may also have a satisfying performance over those places.

Section05 - Conclusions

This paper has analysed whether different weather conditions can affect passenger behaviour in three aspects: trip distance, tip amount and destinations locations. The temperature tends to have a relationship with both trip distance and tip amount, whereas the precipitation cannot really affect those two aspects. However, for both temperature and precipitation have no obvious influence on the distribution of drop-off locations. In further analysis, a generalised linear model may be useful for predicting both trip distance and tip amount using the temperature. This could be helpful for decision making and policy adjustment for taxi companies.

Section04 - Reference

1. New York City Taxi and Limousine Commission (2009-2020) *TLC Trip Record Data.* Retrieved from *https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page*