The Report of New York City Taxi Trip Data

Yicheng Wang 875174

The university of Melbourne

yichengw1@student.unimelb.edu.au

1 Introduction

Taxi service is an essential part of the New York City urban transportation network. The trips data of the 2009-2020 NYC yellow, green taxi, and FHV are provided by the NYC Taxi and Limousine Commission (TLC) (NYC Taxi and Limousine Commission, n.d.). The primary purpose of this report is to gain preliminary insight into the 2018 New York Yellow Taxi data trip and to explore the in-depth information reflected in the data through data analysis and visualization.

2 Data

2.1 NYC Taxi & Limousine Commission (TLC) Service Yellow Taxi Trip Record Data (.csv)

This database records a tremendous amount of trip data for all registered New York taxis since 2009. The investigation of this database will be the main research direction of this report. The database records around valid 100 million trips data across the whole New York with a number of significative features including pick-up and drop-off locations, operating hours, fare amount, tip amount, etc.

Since 2016 July, TLC converted the format of location from longitude/latitude to a more obscure way, area code. The investigation quality may be affected by this in the microscopic sense. However, it becomes more conducive to practical analysis based on the region, giving more general guidance on specific information.

The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

2.2 Taxi Zone Shapefile (.shp)

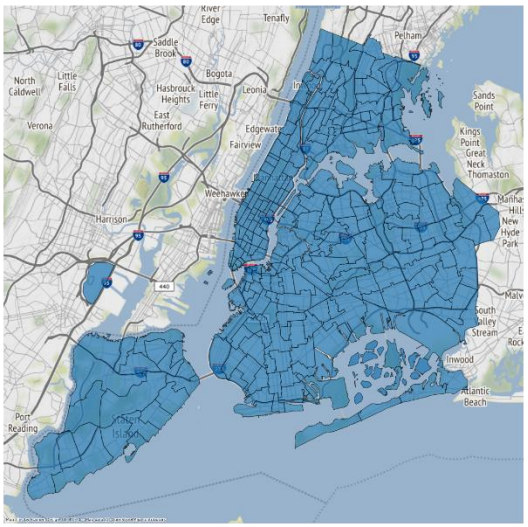

Figure 1: choropleth of NYC Regions Sample

A shapefile (NYC Taxi and Limousine Commission, n.d.) contains New York City polygons (areas) geometric information that divides it into 263 regions with a unique area code. Figure 1 illustrates the spatial area bound via choropleth plot.

2.3 NOAA weather Data (.csv)

National Oceanic and Atmospheric Administration provide free access to global historical weather and climate data in addition to station history information (National Oceanic and Atmospheric Administration, n.d.). These data include daily measurements of temperature, precipitation, snow, etc.

There are several monitoring stations within New York City. In this repot, weather data was collected from one of these stations located at central park (STATIONID: GHCND: USW00094728). This station data is assumed to represent the weather in the whole New York, which may cause the analysis unrepresentative and elimination of regional characteristics.

New York often encounters extreme weather, such as snowstorms, hurricanes, and heavy rains due to geographical location. Combined with taxi trip data, the influence of weather on travel time and profitability for a taxi would be explored.

2.4 Taxi Zone Lookup Table (.csv)

A CSV file works as a dictionary, to indicate area code and corresponding area name. (NYC Taxi and Limousine Commission, n.d.)

3 Period Selection

Limited by insufficient computing power and different research purposes, it is inappropriate to investigate historical data for several years. In this report, trip data and corresponding weather resources for the whole year of 2018 is selected for the following reasons:

1) 2018 encountered many extreme weathers, it is worth exploring further.
2) The location-based part of this report focuses on the analysis of regional characteristics, rather than detailed geographic location.
3) There are around 100 million rows of data in 2018. Analysing the data for a whole year provides seasonal, periodic, and other dimensions for discovery and analysis.

4 Data Processing and cleaning

4.1 NYC Yellow Taxi trip data

The 12-month data are processed separately and output as a file in *.feather format. And merge 12 months of data into one file.

1. (“tpep_dropoff_datetime”, “tpep_pickup_datetime”) were converted to Timestamp format. An new attribute (“duration”) was calculated by the difference of two previous attributes and unit was set by mins.
2. Several attributes having time property were created, including (“month”,”day”,”weekday”,”dayofyear”,”hour”) which is prepared for grouping data in various time degrees.
3. New attribute (“period”) distinguish daytime(6am-10pm) and late nighttime(10pm-6am).
4. Redundant columns removed. (“VendorID”, “datetime”, “tpep_dropoff_datetime”, “RatecodeID”, “store_and_fwd_flag”, “extra”, “improvement_surcharge”)
5. Some technical errors may happen on the taxi meter, these errors in data represent as outlier. some constraints were applied to remove these outliers and errors.
    i. Using boxplot/describe() function to detect outlier, such as Unreliable large/small fare amount or tips, abnormal duration.
    ii. Removed error data, such as wrong time dates or periods, outside New York pick-up/drop-off, negative value.

4.2 NOAA Weather data

1. (“DATE”) was converted to Timestamp format.

2. New Attributes (“RAIN_F”,”SNOW_F”) were created to determine the extreme weather condition.
3. Redundant columns removed. (“STATION”,”NAME”)

5 A General Look on Data

The attributes in the data are roughly divided into several research directions, namely timeline, location based, and fare related. This part forms some basic introduction and analysis of trip data from the perspective of time and space.

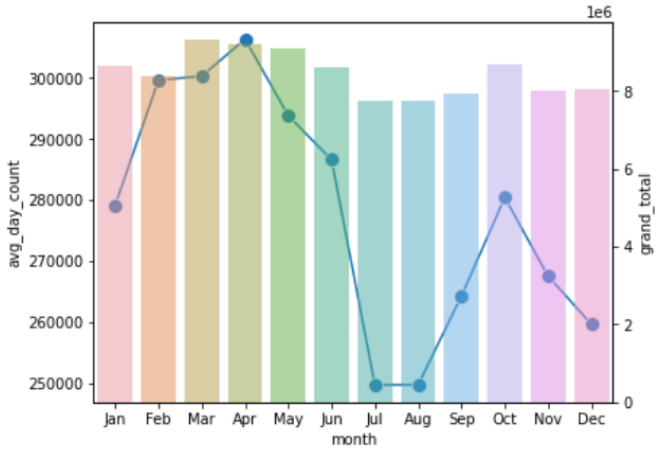

Figure 2: monthly total trip count and daily average count per month

Figure 2 shows the average number of daily trips and the total number of monthly trips for each month in New York in 2018. Looking at the trend, the number of trips is strongly related to seasonal changes. It can be seen from the figure that the average daily trip number fell to the bottom in the summer of July and August but reached its peak in March and April.

Historically speaking, the climate in April is becoming comfortable. People would start to take more economical public transportation as the weather condition improves. Abnormally in April 2018, New York suffered the biggest snowstorm in 15 years. Heavy snow will greatly restrict travel method, so taxis could become a hot choice. Extreme weather may be a key factor for people to increase their choice of taxi travel. Therefore, the impact of weather on the trips will be discussed and analyzed.

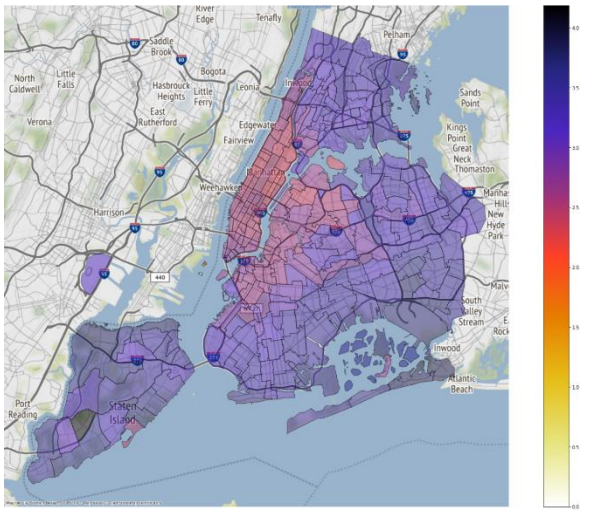

Figure 3: choropleth of log mean duration of trips based on pick-up location in mins

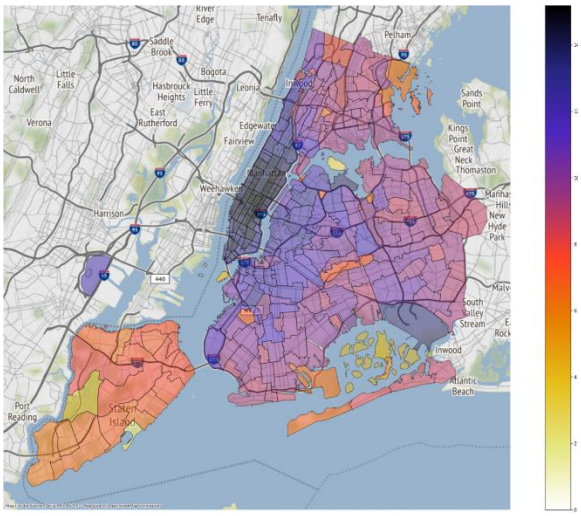

Figure 4: choropleth of log mean trip count base on drop-off location

The different pick-up locations affect the difference in trips time. From figure 3, the shade of the color represents the length of the trips from the pick-up location, radiating from Manhattan as the center. The farther from Manhattan, the darker the color. Therefore, we can have enough guesses that a large part of the trips is from all over New York to Manhattan.

Figure 5 also proves the above point by counting the count of drop off everywhere. As the core area of New York, Manhattan has also become a destination for many taxi trips. So, Manhattan has become the most congested area.

In addition to Manhattan, Figure 5 also suggests that the two airports LGA (LaGuardia Airport) and JFK (John F. Kennedy International Airport) have also become popular drop-off destinations. There is fewer drop-off located at EWR (Newark Liberty International Airport), another larger airport around New York. EWR is 16 miles from Manhattan. The distance is not the reason why this airport is left out of the Manhattan residents. I speculate that it is because the location of EWR is like a transportation hub. Going to Manhattan or other cities such as Boston is more convenient than living in Manhattan. Compared to the expensive and congested Manhattan, more tourists will choose to land at this airport and live near this airport or in New Jersey.

6 Timeline Based Analysis

6.1 Trips quantity related to weekday and weekend

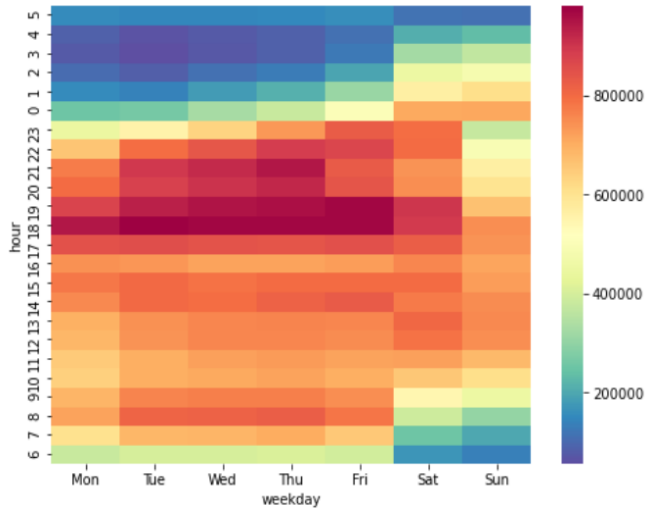

Figure 4: Heat map of weekdays and weekend hourly trip count

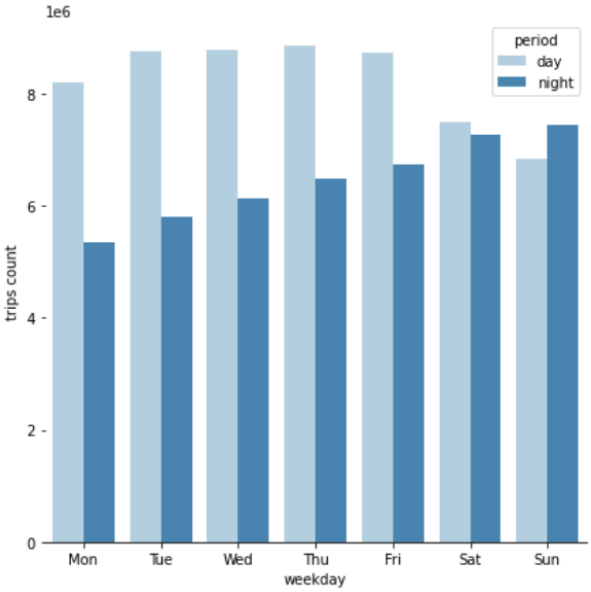

Figure 3 bar chart of weekly daytime and nighttime trip count

The utilization rate of taxis could reflect human travel habits. Figure 5 & 6 show the difference in the behavior of people using taxis during weekdays and weekends. At 6 pm on the weekday, the trip count of taxis is the highest at any time. It is easy to guess that the rush hour is ushered in this period. After 0 am on weekdays, taxis usage has dropped significantly, which is in line with people living habits of resting at home at night during the weekdays. Another rush hour shows likewise at 8am on weekdays. It is inferred that taxi is a popular choice for white-collar worker although taxi fare is not cheap.

Conversely, usage on weekends night is much higher than the same period on weekdays especially from 0 am to 4 am. It can be discovered that New York citizens prefer to go out on weekend nights and choose taxis as their transportation method. This is also related to the suspension or reduction of the most public transportations at late night. Family group self-driving trips and the setting of days off may be the reasons for the fewer trips during the day on weekends

In general, a more effective strategy for taxis driver is to be more active during the daytime and before 10pm on weekdays. On weekends, drivers can choose to work more at night to avoid idling and congestion.

6.2 late night pickup suggestion

We can use the proportion of taxi data at night to speculate on areas with high-quality nightlife in New York, to provide information to those who want to have fun at night. The range of late-night is defined as 10 pm-5 am. Some areas always look calm during the day, but at night it is a different scene. As shown in Figure 7, Brooklyn is one of the highest late-night trip proportion boroughs among the whole New York.

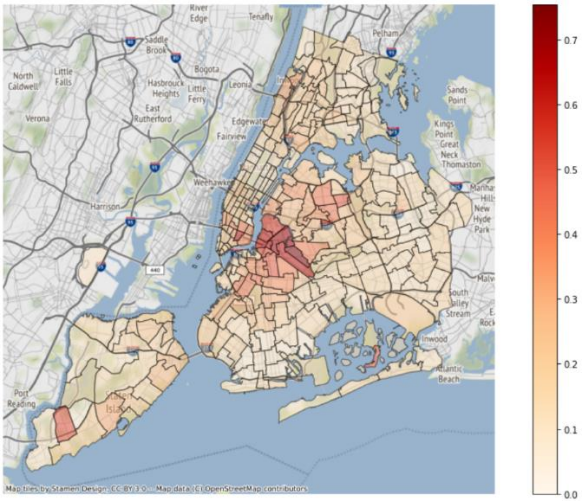

Figure 5: choropleth of late-night trips count based on pick-up location

Bushwick and Williamsburg are two typical representatives. The number of late night (10pm-5am) taxi trips exceeds 65% of the total day. Besides, nights in downtown Manhattan is also lively, and if you want to get out of there, it's a good idea to go to Midtown or Central Park to escape the crowds and increase the probability of getting a taxi.

6.3 Trip analysis throughout the day

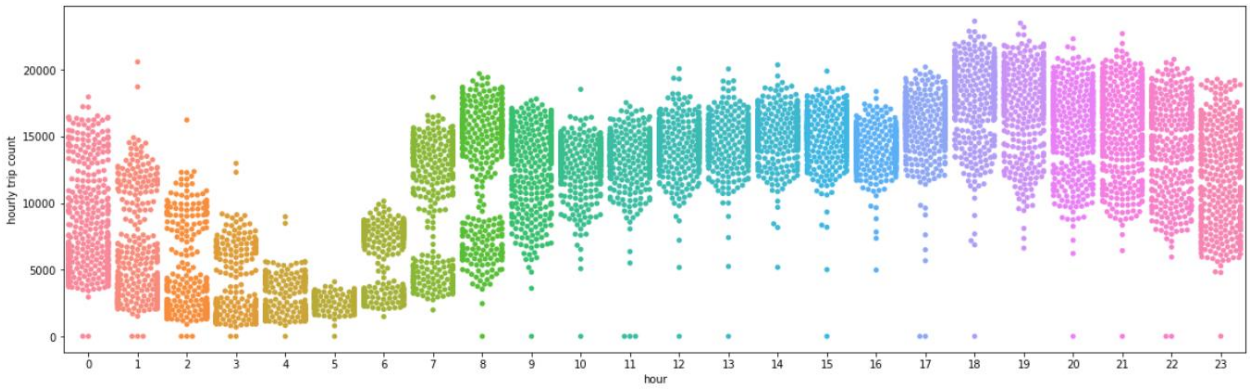

Figure 6: swarm plot of hourly trip count

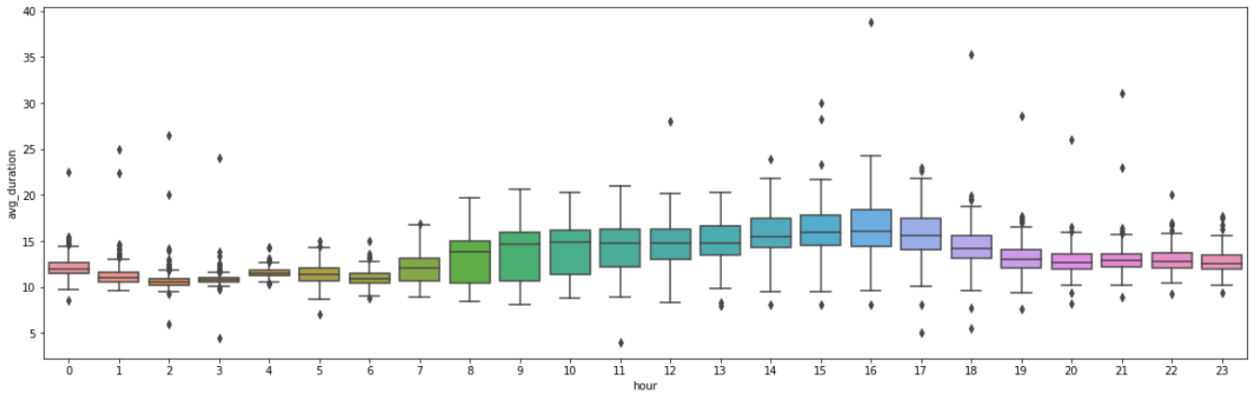

Figure 7: boxplot of hourly average trip duration in mins

Figure 8 tells a lot about the whole day trend of hourly trip count. From 1 am to 5 am, the number of orders dropped sharply, matches human sleep behavior. Orders rebounded from the bottom at 6 am, and after that, reaching an early peak from 7 am to 9 am. The public travel pattern of the commute is shared by public transport, private cars, etc. and people do not favor taxi travel. Assuming an hour commute, it is possible to determine the period between 8 am and 10 am for a small concentration of working hours. Orders grew significantly from 5 pm, peaking at about 7 pm, and the average number of trips remained high until 23 pm. The reason for this might be that people prefer taxis when they are out at night than during the day. It is possible people leave work gradually.

After modifying a few outliers, Figure 9 indicates that there has a much higher chance for taxi drivers receiving a long-distance order before 7am, since traffic congestion would not be considered involve.

7 Airport

figure5 illustrates that the airport is another place to pay attention by facilitating tourism and trade. Taxis are an important connection tool between the airport and the urban area. Researching and understanding the data of taxis to and from the airport can help practitioners improve work efficiency and income. It can also provide transportation information and

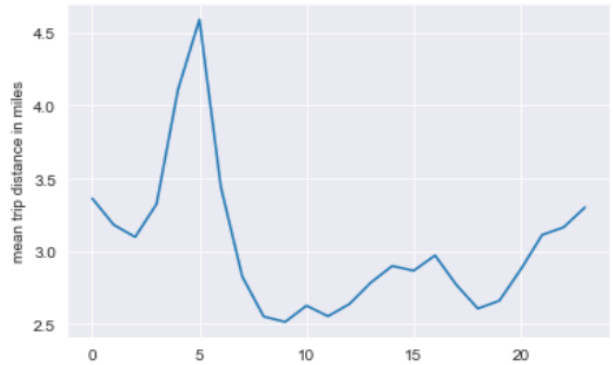

Figure 10: line chart of hourly mean trip distance in miles

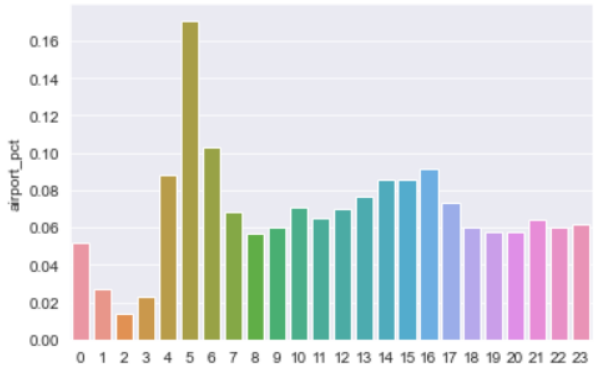

Figure 11: bar chart of hourly airport trip proportion

suggestions to people who travel to and from the airport, including tourists and citizens, helping them to plan their time more reasonably.

7.1 when do people go Airport?

There are three airports around New York, namely JFK, LGA and EWR. Most long-period, long-distance trips are related to airports and are defined as airport trips. Figure 10 points out that the mean trip distance of 5am exceeds 4.5 miles, which is much higher than other times. Since 5am cannot be defined as an early rush hour, we have reason to believe that most long-distance trips during this period are airport trips. According to speculation, there are the following reasons:

1) Choosing a flight departing early in the morning avoids traffic congestion and makes time to go to the airport more abundant, especially for residents living in Manhattan.

2) Early morning flights are mostly economical, and this behavior is also in line with human working habits. For some people who work out, they will not waste precious time during the day.

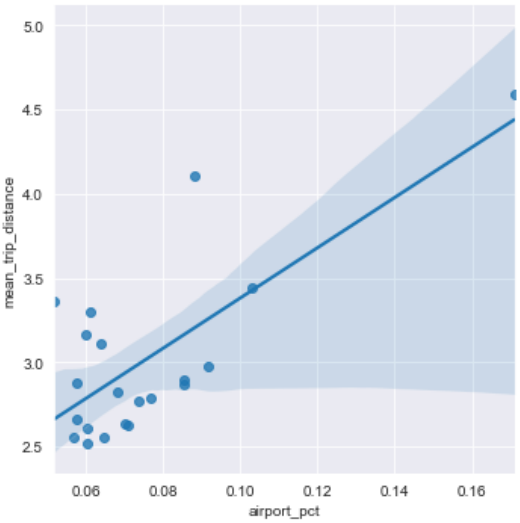

Figure 12: scatter plot of fitted line of airport trip count and mean trip distance

	airport_pct	mean_trip_distance
airport_pct	1.000000	0.729626
mean_trip_distance	0.729626	1.000000

Figure 13: correlation of airport trip proportion and mean trip distance

Combined with figure 11, the proportion of airport trips at 5am exceeds other periods, which also validates the previous analysis. On the whole, the distribution of airport trips is proportional to the trip distance. After eliminating the 2am where the airport trip count accounted for less than 2%, figure 12 showed a strong positive relationship between airport trip percentage and long-distance trip with a 0.73 Pearson correlation score, as figure 13 indicates. There is sufficient evidence that most long trips are airport trips. If drivers prefer to pick up long-distance trips, then the airport is a reliable pick-up location.

7.2 Manhattan to airport

Figure 14: choropleth of airport trip frequency area

The areas where people choose to take taxis to the airport are polarized. Almost all pick-up locations are concentrated in midtown and downtown Manhattan (figure 14. This may be because most of the people living outside Manhattan are locals, and they tend to drive to the airport or public transportation. Downtown gathers Wall Street, Broadway, and a large number of star-rated hotels and the Midtown district where Fifth Avenue is located is a typical wealthy area. People who live there would choose more private transportation. This explains well the phenomenon of pick-up concentrated in Manhattan.

7.3 hail a cab in advanced

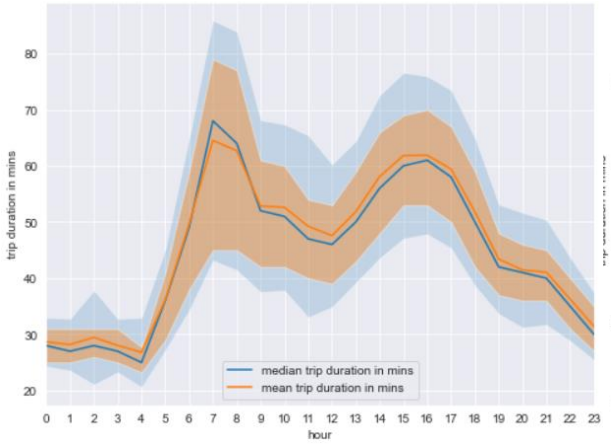

Figure 15: line chart of midtown to JFK mean and median duration in mins

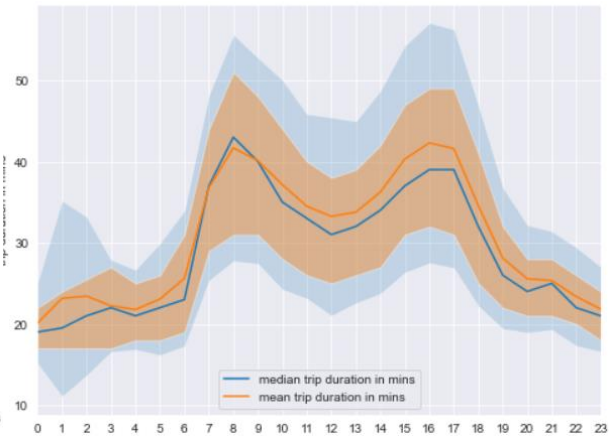

Figure 16: line chart of midtown to LGA mean and median duration in mins

Airport trip is a form of long-distance traffic with strict timeliness requirements. People always encounter missed flights due to unexpected situations such as heavy traffic jams and road closures. This situation is particularly obvious in Manhattan congested traffic environment. I

counted the median and mean of each time period from midtown Manhattan to JFK and LGA. (figure 15 & figure 16)

No matter which airport is destination, there will be two time periods that will take longer, morning and afternoon rush hour. For example, if you depart from Manhattan at 7 am, it will take about 70 minutes on average to reach JFK Airport. It takes twice as long as in non-congested periods.

Therefore, the habit of people always choosing to go to the airport at 5am is in line with the general understanding of traffic conditions, effectively avoiding traffic congestion.

8 Weather conditions affect taxi industry

The weather condition will affect the people travelling way. On snowy days, for example, low visibility and snow on the road prevent people walking. Bad weather conditions can make it easier to subconsciously to choose private transportation because unknown circumstances create uncertainty, that most people do not like that. This part will discuss the impact of weather changes on taxi trips.

8.1 Different weather condition different taxi usage

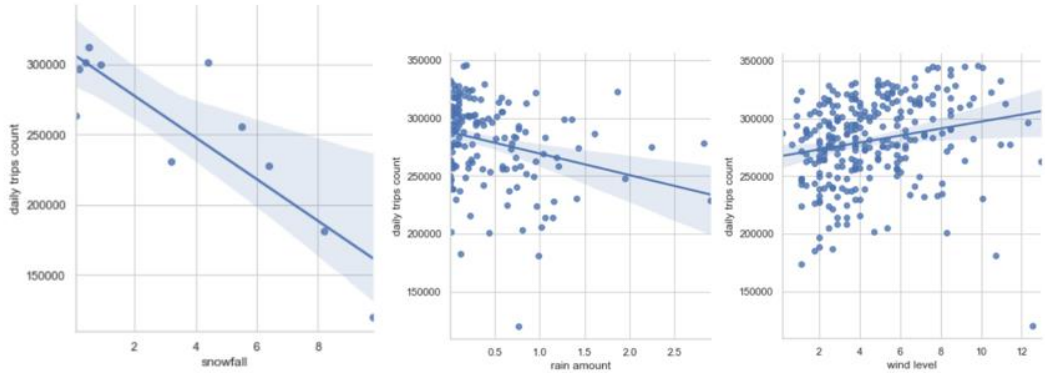

Figure 17: scatter plot of fitted line of different weather conditions and trips count

From the figure 17, three different weather conditions are listed. Snowy days have the greatest impact on taxi usage among these three conditions. Snowfall and taxi usage show a strong negative relationship. As the amount of snow increases, daily taxi trip counts become smaller and lower than the annual average. Contrary to the original assumption, people would choose to reduce unnecessary travel due to heavy snow, as well as heavy snow on the roads, traffic conditions would also be a factor affecting taxi. One suggestion for taxi drivers is to reduce the driving time in snowy days, which can effectively reduce the loss of profits caused by no-load.

However, for windy and rainy days, the above picture does not show an explicit relationship with daily taxi usage. In other words, the figure shows a weak relationship between these two factors and trip count. I speculate that these two types of weather are not enough to change

the way of travelling, but only increase the difficulty of travel. For example, heavy rain only requires people to hold umbrellas, while heavy snow truly affects the road conditions .

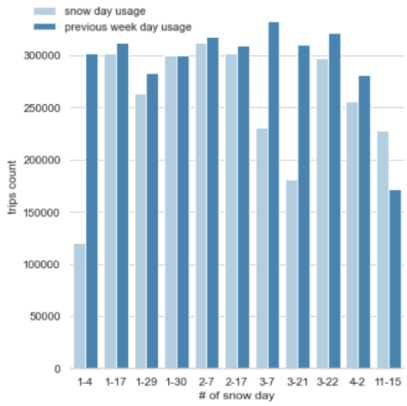

Figure 18: bar chart of snow day vs. previous week daily trips count

A total of 9 days of snow in 2018 were detected by the central park station. However, there are more than 120 weather monitoring stations located in different areas in New York. It is not convictive and general to select the data of one weather station as the whole weather in New York. This is also the shortcoming of this analysis report, and the part that can be improved.

As figure 18 shown, among all the snowy days of the year, the trip count of 10 days(almost all) is lower than the count of the previous week (or the following week). Even on January 4 and March 7, there was almost a double gap. Thus, we can conclude that snow is an important factor affecting taxi trips or the decision of whether to take the trip.

9 Conclusion

This report gives a preliminary investigation for mining the logical relationship behind the NYC taxi trip data. For this purpose, we used spatial-temporal association and multi-level directed graphs, which have given several interesting findings and analysis on multiple aspects including: different time-based pickup frequency, airport peak hour with road condition, and how extreme weather influence taxi industry. Through the above findings.

Taxi trips may be affected by many factors, including time, people's living habits and the appearance of extreme weather. But there are other unexplored factors that need to be considered. Only by knowing the entire scope of the problem, the society can discover in other areas by understanding taxi data.

10 Reference

National Oceanic and Atmospheric Administration. (n.d.). *Daily Summaries Station Details*. Retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

NYC Taxi and Limousine Commission. (n.d.). *Taxi Zone Lookup Table*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

NYC Taxi and Limousine Commission. (n.d.). *Taxi Zone Shapefile*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

NYC Taxi and Limousine Commission. (n.d.). *TLC Trip Record Data*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page