Preprocessing and Exploratory Data Analysis of Large-Scale Taxi GPS Traces

(1) How many unique taxis are there in this dataset, and how many trips are recorded?

The number of unique taxis in this dataset is 13,385.

The number of trips recorded is 147,800,095.

(2) What is the distribution of the number of trips per taxi? Who are the top performers?

The distribution of the number of trips per taxi basically follows a normal distribution (Fig 1). During 2011, most taxis have trip numbers around 8,000 to 14,000. And the trip number over 21000 or less than 500 is rare. The most common trip number is 13558. The top performer is the taxi with ID 4816 who has 22842 trips in total.

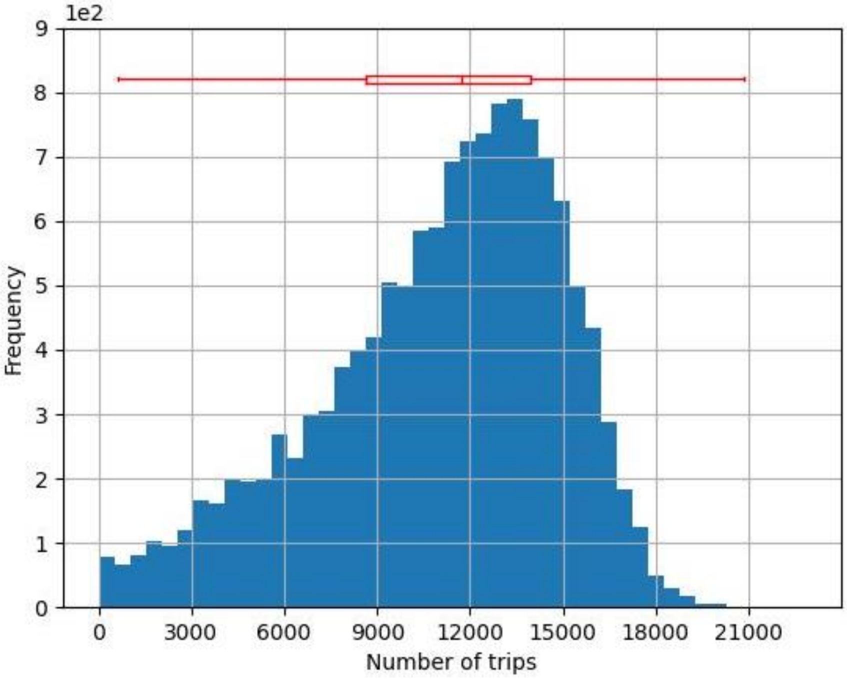

Fig 1. Histogram of the Number of Trips Per Taxi

(3) How does the daily trip count (i.e., number of trips per day) change throughout the year? Any rhythm or seasonality?

Because some trips ended on the second day, the time of every trip will be counted based on the pick-up time. According to the record, the average number of trips per day is 404,932. As shown in the Fig 2, over 213 days' total number of trips exceed the average. And the coefficient of variation of the number of trips per day is 14.90% less than 25%, which means the daily trip counts does not change much on the scale of day. It can be found from the Fig 2 that mid-month is usually the time when there are the most people who prefer to take taxi.

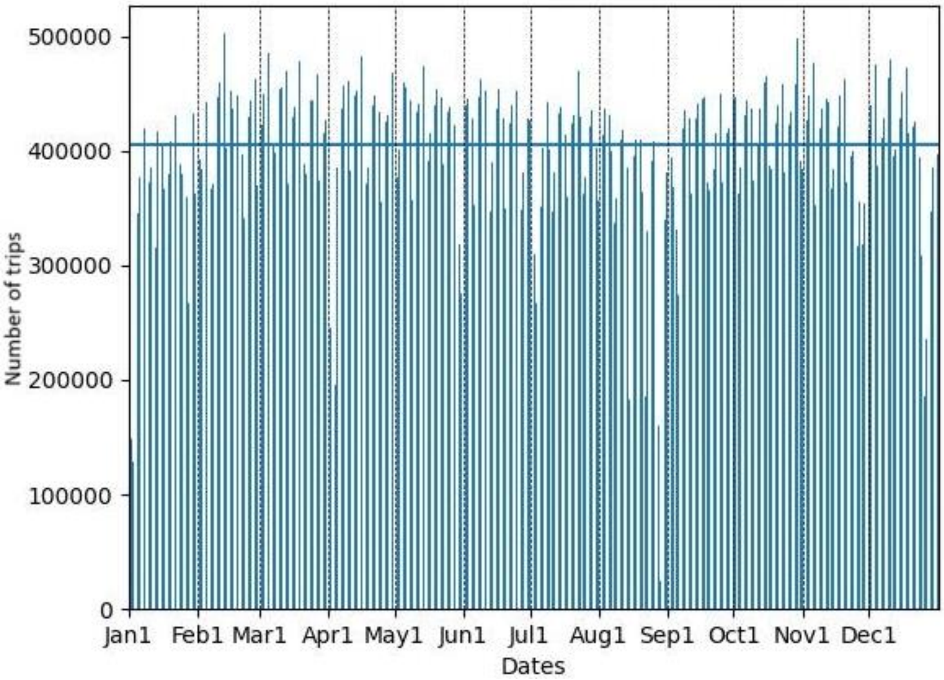

Fig 2. Number of Trips Per Day

In the scale of month, the average number of trips per month is 12,316,675. As shown in the Fig 3, over 7 months' total number of trips exceed the average. And the coefficient of variation of the number of trips per month is 5.54% less than 25%, which means the monthly trip counts does not change much on the scale of month. It can be found from the Fig 3 that the trend of number of trips is first increasing, then decreasing, and then increasing again all the year. The peak is March and October.

The probable reason for the rhythm may be due to the temperature change of the weather. In generally, there a less trip number in winter and summer, it may be too hot or cold to have an outside trip.

And the number may also be affected by the long-term school holiday. According to DiNapoli. TP (2024), New York will enroll around one million students from 2009, they will be an enormous potential customer for taxis. For example, the duration of Spring Break is about one to two weeks and usually arranged from mid-March to early April, and there are lots of non-local students who need to take taxis to go home or have a trip. So, the number of trips in March has a notable rise than January and February (Fig 3). This phenomenon can be also detected from the number of daily trips, which has a notable raise in the mid-March and min-April because of the outbound and inbound passengers flow, and the early April has few trips number because of the students has already left New York. The Summer Break starts on late May or early Juny and ends on late August or early September, it is obvious that the number of trips have a significant decrease in the term from Juny to August but have increase after this period. It has a similar rhythm to Spring Break. This rhythm can also be seen during the Winter Break, whose duration is from mid-December to middle January.

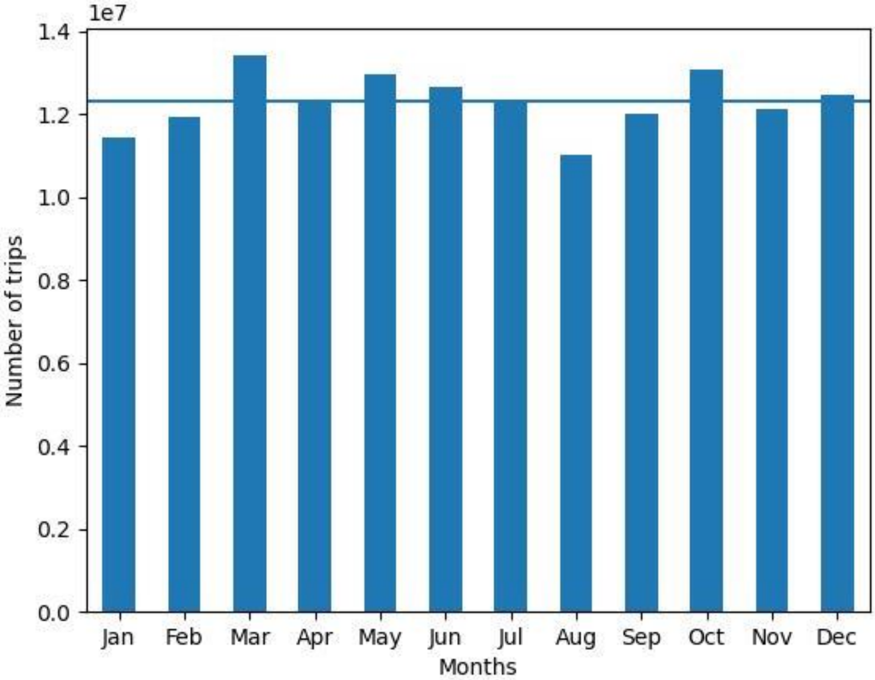

Fig 3. Number of Trips Per Month

(4) What is the distribution of the number of departure trips at different locations (i.e., intersections)? What about the distribution of arrival trips? What will you conclude from these two distributions?

We can get an intersection point list with number of departure trips and arrival trips by taxi_id.csv using Python. Then this list can be imported into ArcGIS Pro and symbolize the distribution of the number of departure trips and arrival trips at various locations (Fig 4).

It is notable that the inside of Manhattan has more trip numbers than the outside whether it is the number of departure trips or the number of arrival trips. It can be concluded that most people take taxis to the inside of Manhattan rather than the outside. And the north part of Manhattan has less trip number than the south and central parts whether it is the number of departure trips or the number of arrival trips. This may be because the south and central parts have much more scenic spots like Central Park and Empire Stata Building than the north parts, and which will attract substantial number of tourists who do not have private car. The north part is residential areas, and most American residents have private car, they basically do not need taxi, so that it has less trip number than the south and central.

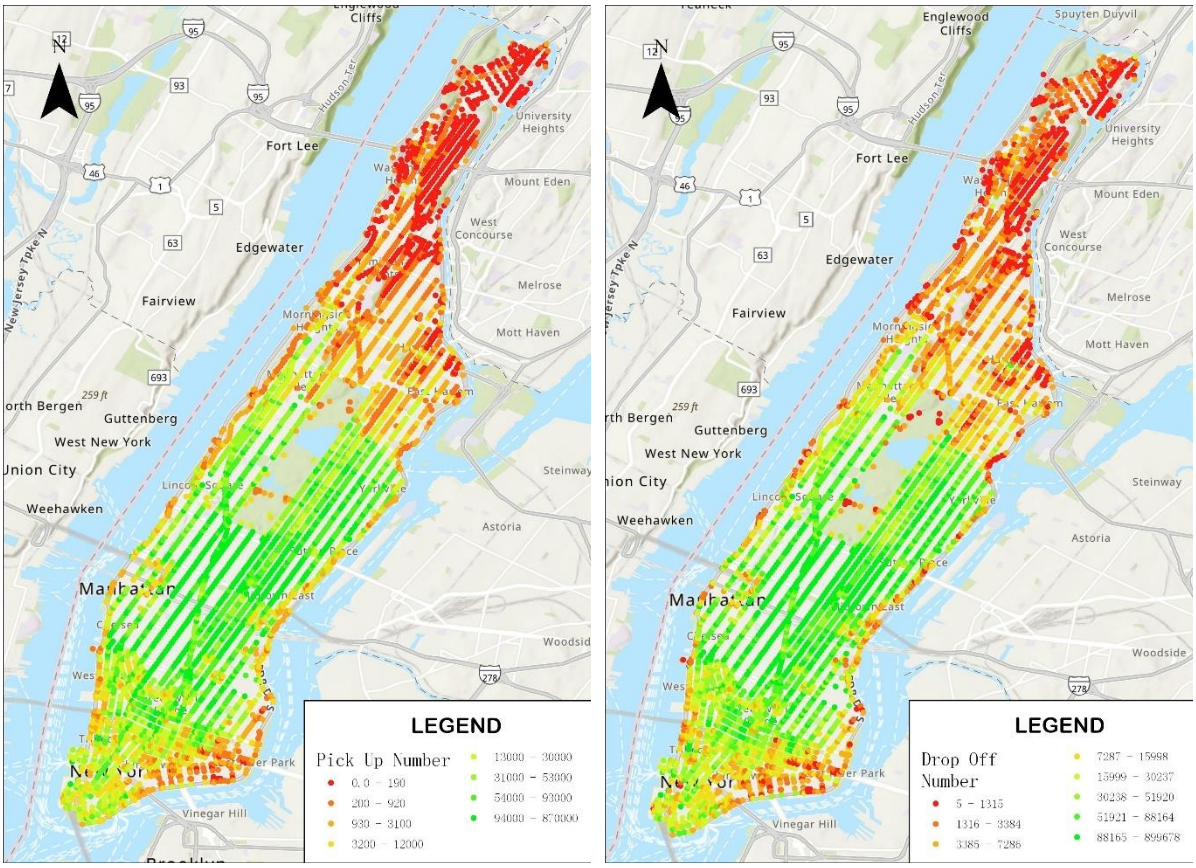

Fig 4. Number of Departure Trips (Left) and Arrival Trips (Right)

Using field calculator in ArcGIS Pro to subtract the number of arrivals from the number of departures, we can get a figure of the difference of departing and arriving situation of each intersection (Fig 5). It is notable that the north parts of the Manhattan have almost the same number of departures than arrivals, this mainly because that the north parts are residential areas, some people who live there would take taxi away for some reasons and finally take taxi back again and arrive at the same point.

But the south and central parts have an interesting phenomenon. It seems that the people, mainly tourists, in the south and central parts prefer to drop off at specific places (green points) to start their tour and take a taxi to leave at other specific places (red points). The intersections with more picking up and the intersections with more dropping off are distributed alternately with some space intervals. And the intersections with same attributes can always be linked by one road, this may because that people prefer to take taxi along the road rather than a single point.

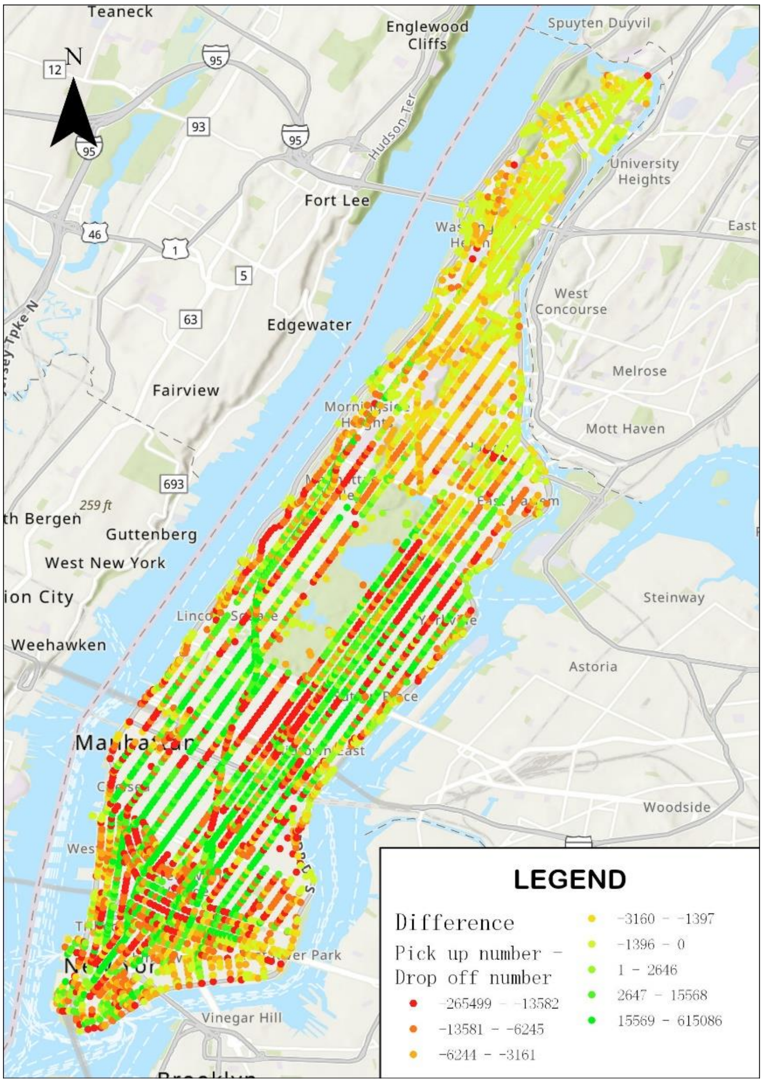

Fig 5. Difference Between Departures and Arrivals

(5) How does the number of trips change over time in a day? (You will be given three dates randomly selected from the dataset, and then plot the hourly variation of trips from the perspective of local time).

The given dates are 19th Aug 2011, 11th Nov 2011, and 16th Dec 2011. They are all Friday and have almost the same hourly variation of trips trends.

19th August 2011

During the 19th of Aug 2011, the peak hour of picking up is 20 o'clock, and the low peak hour of picking up is 4 o'clock. The peak hour of dropping off is 19 o'clock, and the low peak hour of dropping off is 5 o'clock. It can be known from the Fig 6 that the hourly variation of trips trends is descending during 0 o'clock to 5 o'clock, and the trips number increases quickly until 9 o'clock, where is the morning peak, then has a fluctuation descending until 16 o'clock when the evening peak starts. The trips number increases after that until 19 o'clock when the evening peak ends, then decreases until the end of the day. It is also obvious that the picking up trend has no significant difference between the dropping off trend, they all have almost the same variation except for some lag.

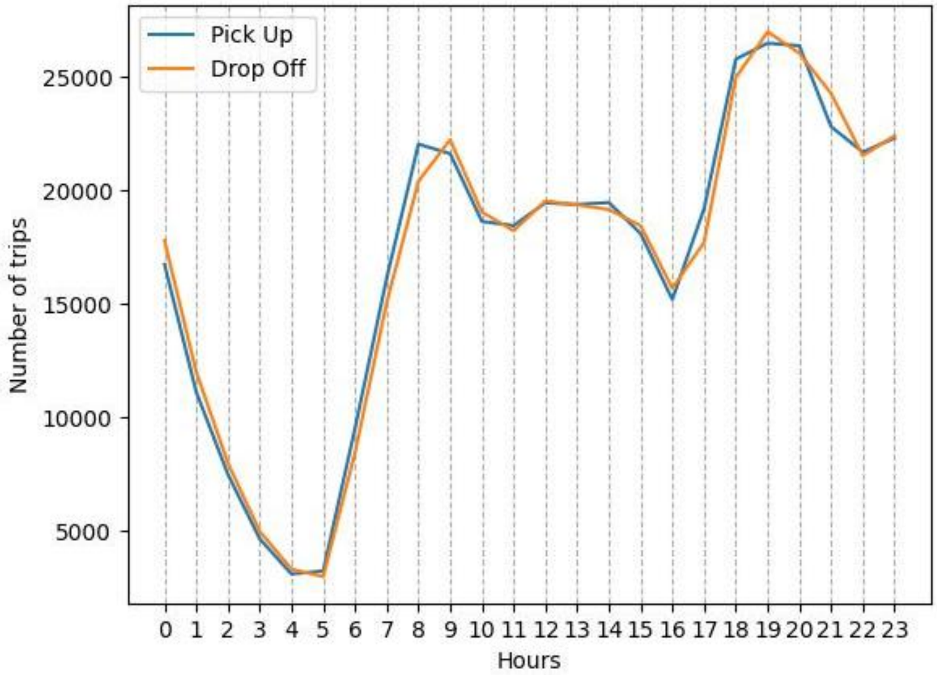

Fig 6. Number of Trips Per Hour During 19th Aug 2011

Subtracting the number of dropping off from the number of picking up, we can get a line chart of the difference of hourly variation of trips (Fig 7). It is notable that morning time from 0 o'clock to 4 o'clock has a greater number of dropping off because it is time to go home.

And there is a greater number of picking up from 5 o'clock to 8 o'clock because it is time to go to work. The no significant difference between the number of picking up and dropping off during 9 o'clock to 16 o'clock. And after 16 o'clock the number of picking up overflow the dropping off because it is time to get off work. At 21 o’clock the number of dropping of overflow the picking up because most people who were picked up before has arrived at their destination.

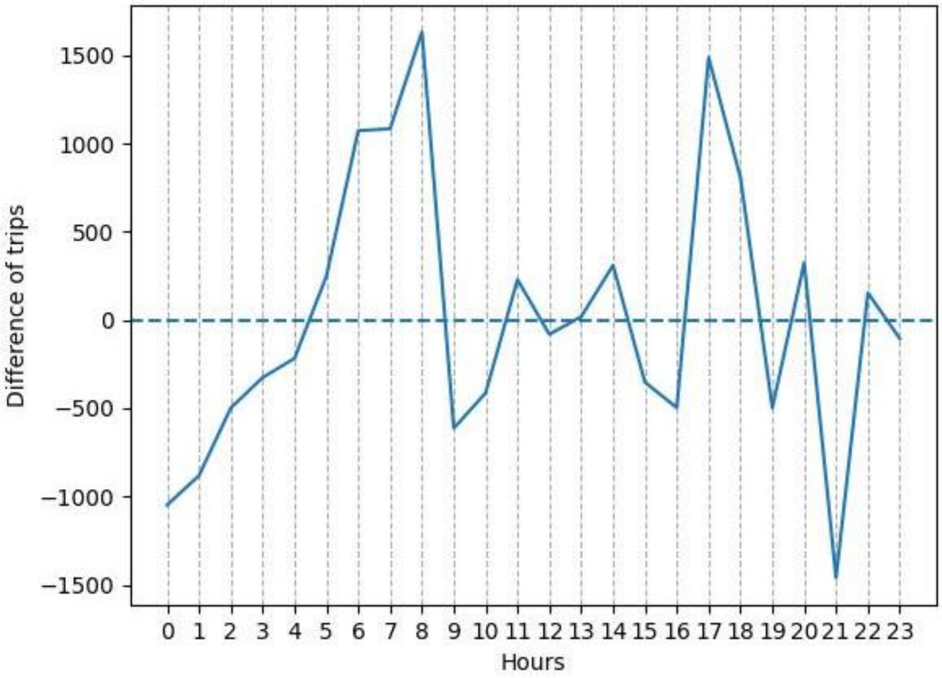

Fig 7. Difference of the Number of Trips Per Hour During 19th Aug 2011

11th November 2011

During the 11th of Nov 2011, the peak hour of picking up and dropping off are both 22 o’clock, and the low peak hour of picking up and dropping off are both 5 o'clock. It can be known from the Fig 8 that the hourly variation of trips trends is almost same with 19th Aug (Fig 6) but only has a slightly difference after 19 o’clock, the number of trips on 11th Nov keep flat until the end of the day rather than decrease. This may be because 11th Nov is Veterans Day and people have more free time to enjoy their night life. It is also obvious that the picking up trend has no significant difference between the dropping off trend, they all have almost the same variation except for some lag.

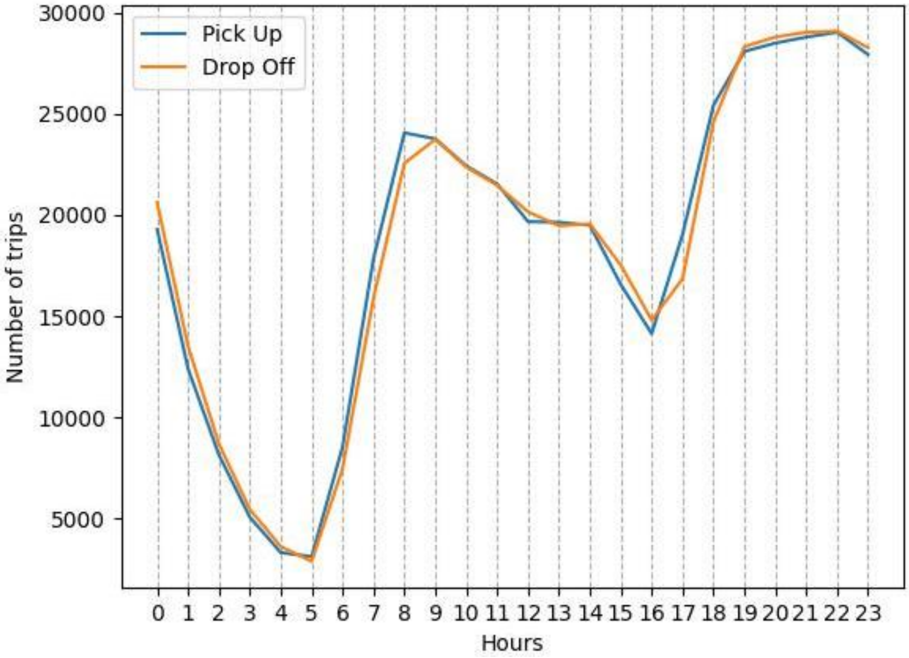

Fig 8. Number of Trips Per Hour During 11th Nov 2011

Subtracting the number of dropping off from the number of picking up, it can be founded that the trend of difference of hourly variation of trips on 11th Nov 2011 (Fig 9) is almost same with 19th Aug (Fig 7).

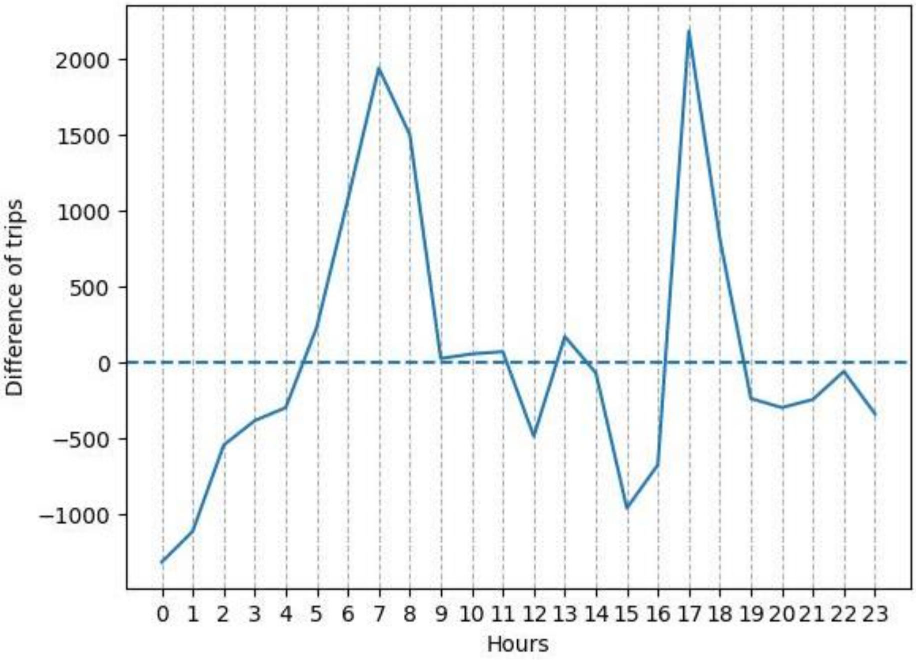

Fig 9. Difference of the Number of Trips Per Hour During 11th Nov 2011

16th December 2011

During the 16th of Dec 2011, the peak hour of picking up and dropping off are both 19 o’clock, and the low peak hour of picking up and dropping off are both 5 o'clock. It can be known from the Fig 10 that the hourly variation of trips trends is almost same with 19th Aug (Fig 6) and 11th Nov (Fig 8). It is also obvious that the picking up trend has no significant difference between the dropping off trend, they all have almost the same variation except slight lag.

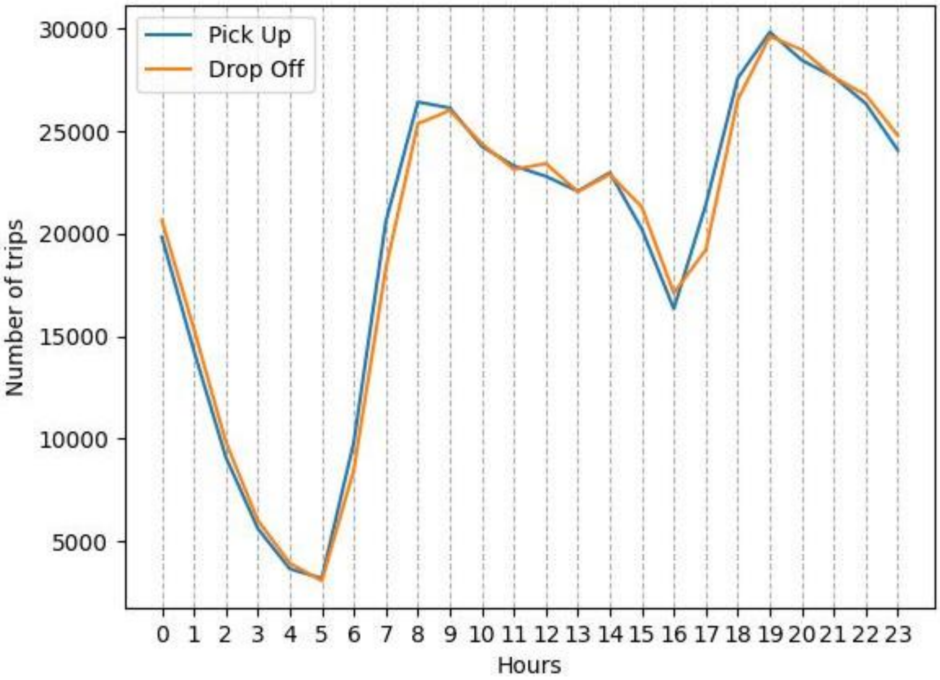

Fig 10. Number of Trips Per Hour During 16th Dec 2011

Subtracting the number of dropping off from the number of picking up, it can be founded that the trend of difference of hourly variation of trips on 16th Dec 2011 (Fig 11) is almost same with 19th Aug (Fig 7) and 11th Nov (Fig 9).

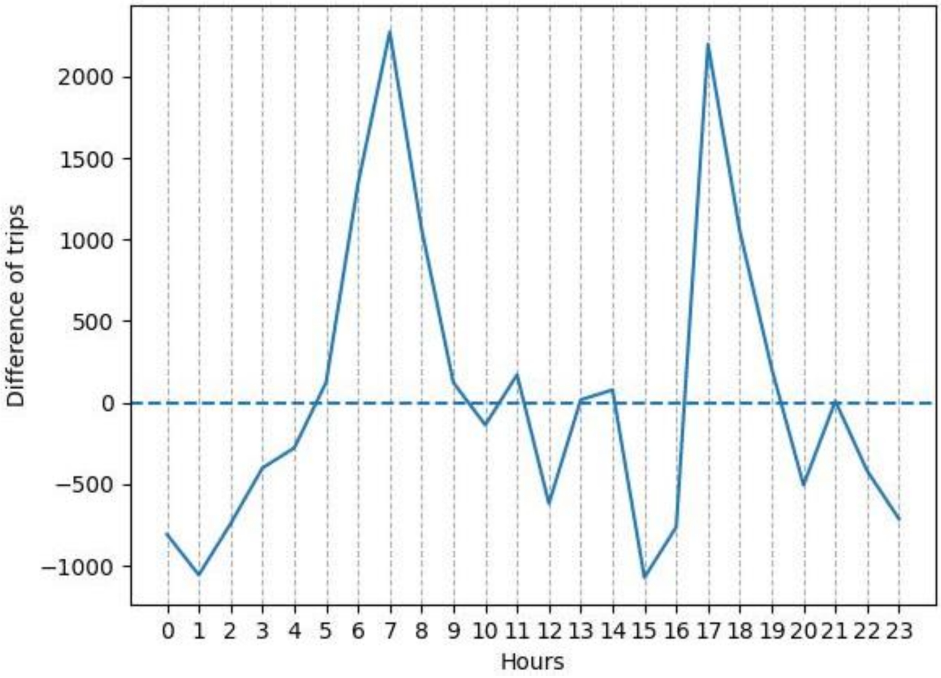

Fig 11. Difference of the Number of Trips Per Hour During 16th Dec 2011

(6) What is the probability distribution of the trip distance (measured as straight-line distance)? How about travel time (i.e., trip duration)? What will you conclude from these two distributions?

Some trips have the same departure and arrival intersection, this may be because some trips may have a circle route. Therefore, the records who have the same departure and arrival intersection will be considered as valid data.

The distribution of the trip distance is a skew distribution (Fig 12) and its skewness and kurtosis both greater than zero. During 2011, most taxis have a trip distance less than 3.5 KM and the travel distance over 6 KM is rare. And the most common travel distance is around 1.1 KM to 2.5 KM. The longest trip in distance is the taxi with ID 3265 who drove 20.89 KM in one trip, almost the length of diagonal of Manhattan. The average distance pre trips is 2.47 KM. This means that most taxi trips in Manhattan are short trips. There may be two reasons leading to this distribution. First is that the area of Manhattan is not large, which is only about 21.5 KM long and 3.7 KM width, it limits the upper limit of the travel distance. The second reason is that the price of taxis in Manhattan is expensive, so people in Manhattan may choose other mass transit like metros or buses for long trips.

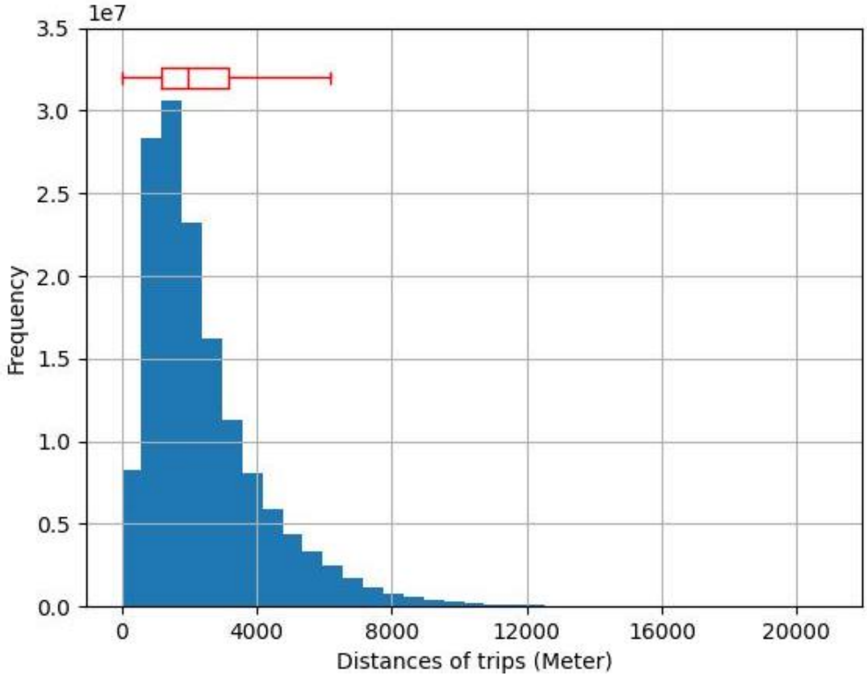

Fig 12. Histogram of the Trip Distance

However, the data between pick-up time and drop-off time have some problems. First, some data's pick-up time is after drop-off time, and these data will be considered invalid and be deleted directly. Second, some data's travel time are larger than 24 hours which is impossible in the real world, so the data with travel time which exceed the mean plus or minus three times of the standard deviation after the deleting of invalid data are considered as outliers and will be deleted.

After deleting of dirty data, the distribution of the trip time is a skew distribution (Fig 13Fig 12) and its skewness and kurtosis both greater than zero. During 2011, most taxis have a trip time which less than 15 minutes (900 seconds) and the travel time which over 25 minutes (1500 seconds) is rare. The common travel time is around 6.3 minutes (380 seconds) to 13.8

minutes (830 seconds). The longest trip in time is 30 mins and 57 seconds in one trip. The average time pre trips is 10 mins and 20 seconds.

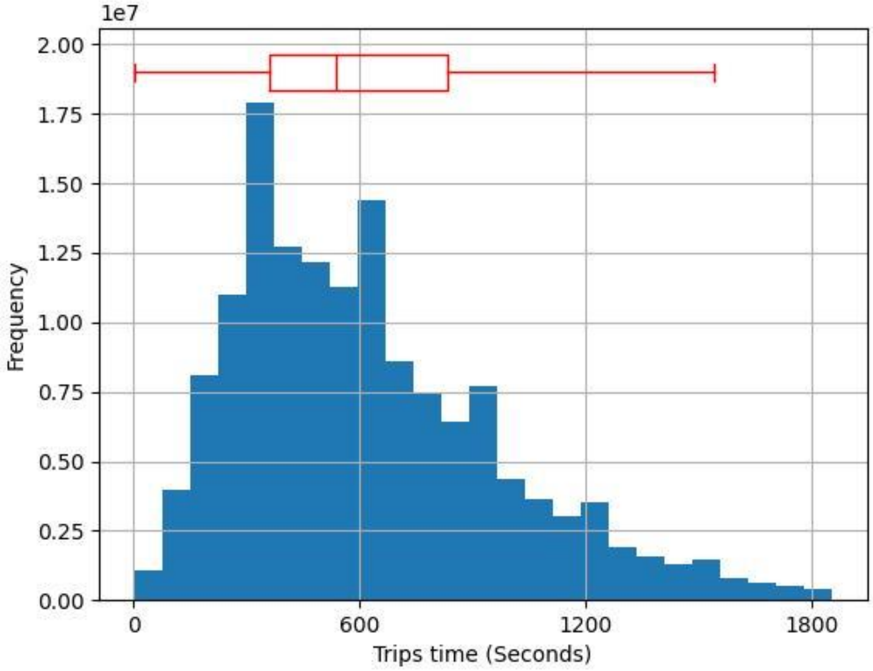

Fig 13. Histogram of the Trip Time

In summary, most taxi trips in Manhattan are short trips with travel time between 5 minutes to 15 minutes and travel distance less than 3.5 KM. This time consumptions are rational and in line with people's daily habit of taking taxis, not too short or too long. And the distribution patterns of travel time resonate with the distribution patterns of trip distance. On the other hand, the average speed in Manhattan is not fast, about 15.06 KM/h, which can indicate that the traffic condition in Manhattan is not very well.

Reference

DiNapoli. TP (Ed.) (2023). Higher education in New York: Evaluating competitiveness and ldentifying challenges. Office of the New York State Comptroller. https://www.osc.ny.gov/files/reports/pdf/higher-education-nys.pdf