WHY YELLOW TAXI

Big Data Analysis On New York City Yellow Cab Trips Along Subway Lines

Abstract

Yuan Shi, Sichen Tang, Chunqing Xu, Dongjie Fan, Huy T Vo (Mentor), NYU CUSP

New York City has a sophisticated public transportation system, the subway is the cheapest and most necessary mean of transportation, but in some situation, people prefer the yellow cab, why and which line or lines have most business 'stolen' by yellow cabs. Our paper will mainly aim at answer these questions as well as some other related study. We processed the data set in Hadoop use pyspark and conducted our statistical analysis using ipython notebook. Our result showed that temperature affected people's choice the most. People use the yellow cab for creation and subway for working commute.

1.Background

As one of the largest cities in the world, New York City has a complex public transportation system. Taxi and Subway are two main choices for new yorkers. Sometimes, people could take a subway train from one place to another without any transference, but they choose to take a taxi instead. We are curious about which one among the subway lines is the one that people are most unwilling to take. To be specific, we would find out the subway line with the most trips by taxi which could have been taken that line. Information of location and time are included to detect people's choice pattern.
This research will help deliver an answer for why people would prefer a taxi than subway and then help both the government and transportation services to improve the efficiency of public transportation.

2. Data Source

1. TLC Trip Record data
   TLC Trip Record data is from the website of New York City Taxi & Limousine Commission (TLC). TLC Provides record data for both the yellow and green taxi cabs, capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Only the yellow taxi data is included in our research.

Datasets for yellow taxi cabs of each month is about 1.6 GB; the total is about 19 GB for July 2015 to June 2016.

2. 2016 (May) New York City Subway Station Entrances
   Data on New York City Subway station entrances is accessible from GIS Lab of Baruch College, the City University of New York. In the dataset, subway station entrance locations for the MTA NYC Transit Subway were plotted using the coordinates and saved as a spatial layer in the local state plane coordinate system.
   This dataset id of 11.1 MB, presenting 1868 subway station entrances of New York City (excluding Staten Island Railway stations).

3. 2016 (May) New York City Subway Routes
   Subway routes data of New York City is also from GIS Lab of Baruch, CUNY. Subway routes were plotted and saved as a spatial layer in the local state plane coordinate reference system. The routes dataset is of 274 KB showing 22 subway lines.

3.Data Preparation

1. First of all, for each subway entrance, build a buffer area (Figure 3.1) with 100 us-ft radius (around 30 meters), which is the acceptable distance for people to be willing to take the subway.

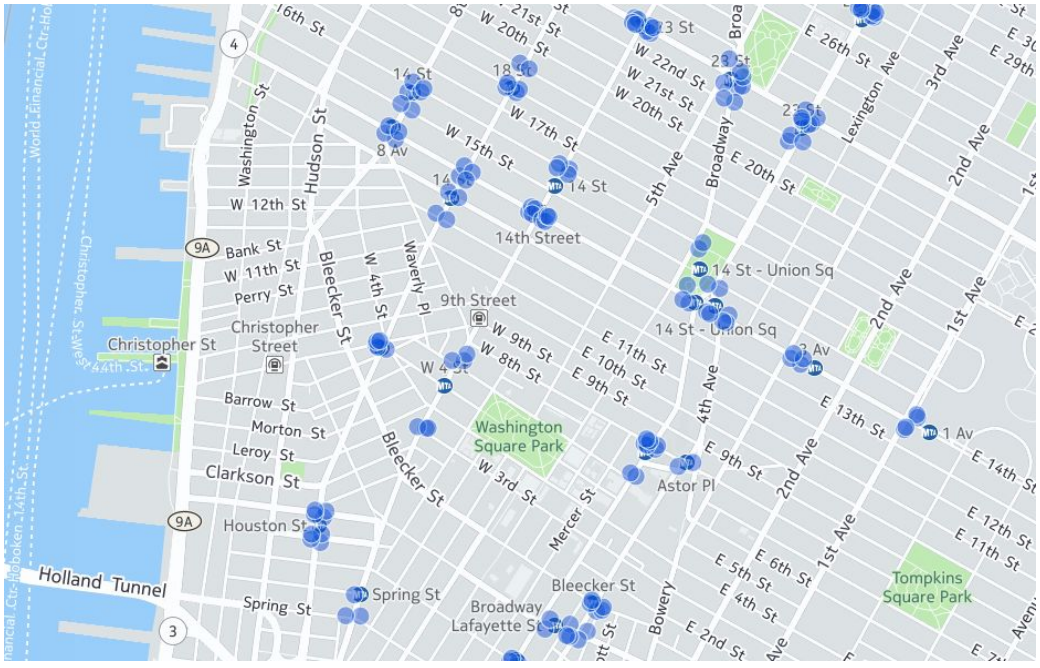

Figure 3.1

2. Next, for each yellow taxi record, extract its pickup and drop-off geo-location (longitude and latitude). Do spatial join with buffer areas. If both pickup and drop-off locations are within buffer zones, identify its two corresponding entrances and the stations. (Here we use rtree to run spatial join faster)

3. For each filtered record, find out all subway lines which have stops at both stations mentioned above.
4. Return the counts of each subway line.
5. We also consider the weekend and late night service time changes (*Figure 3.2*). Beside we ignore the record which pick-up and drop-off locations are within one entrance buffer.

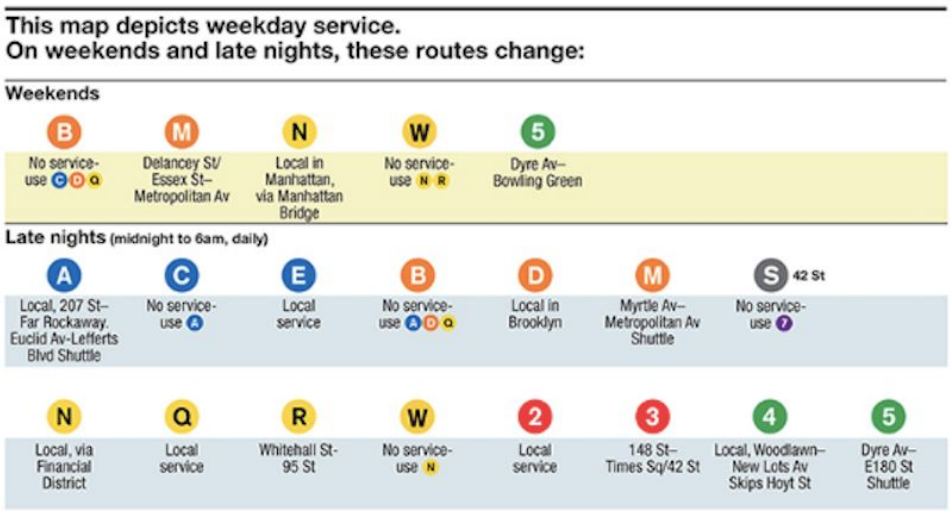

Figure 3.2

For example, the pick-up *(Figure 3.3)* location is around West 4th St NW, and the drop-off *(Figure 3.4)* location is around 59 St - Columbus Circle West.

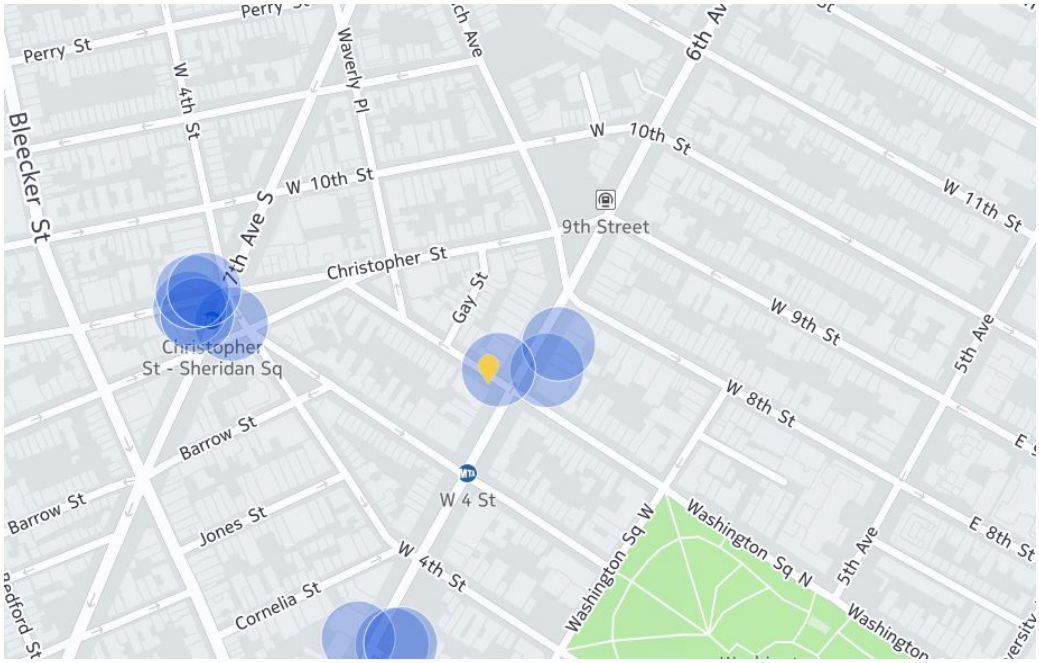

Figure 3.3

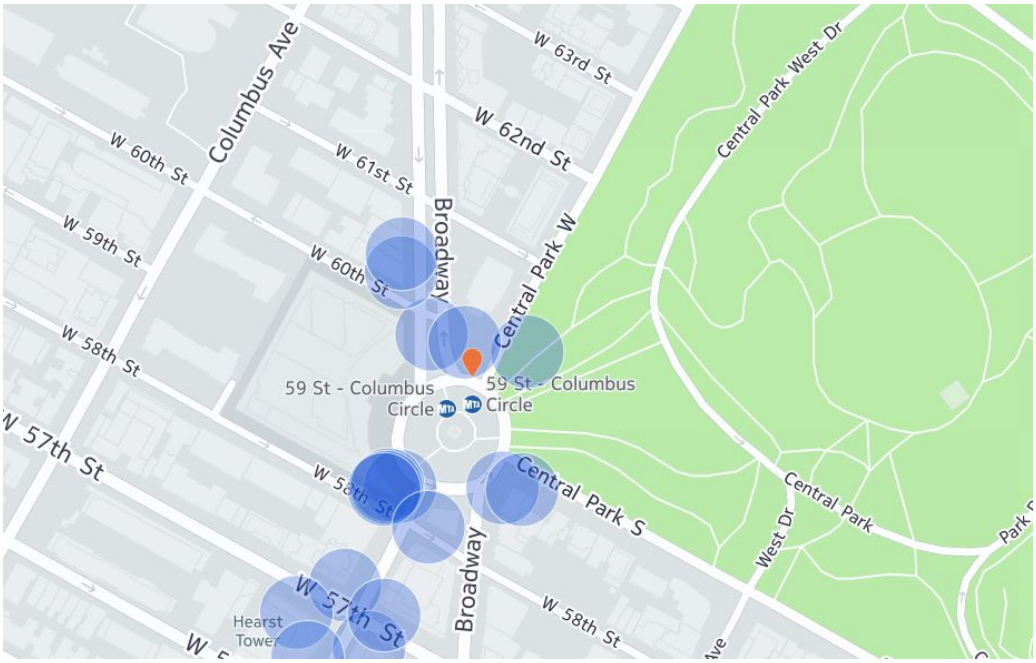

Figure 3.4

Station	**Route**
West 4th St NW	A, B, C, D, E, F & M
59 St - Columbus Circle West	A, B, C, D & 1

Two stations are connected by A, B, C and D Train, which means people can take these subway lines to their destination directly without any transfer. Thus, add one count for A, B, C and D Train.

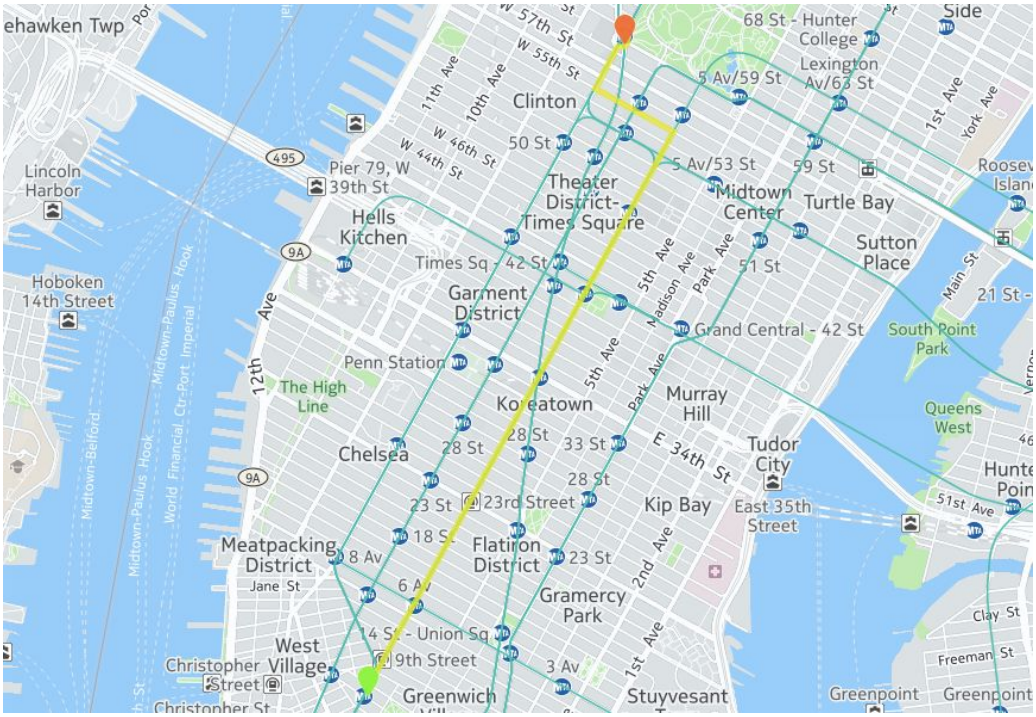

Figure 3.5

4.Analysis

4.1. anomaly day

By aggregating all yellow cab trips around all the subway stations by day, then from Figure 4.1.1 the time series from July 1, 2015 to June 30, 2016 shows there are two significant drops. One is Dec 25, 2016. When we zoom in, Figure 4.1.2 shows the other drop is Jan 23, 2016. Christmas had a drop is understandable, as more drivers and passengers were out of the city. But why Jan 23 is an anomaly day?

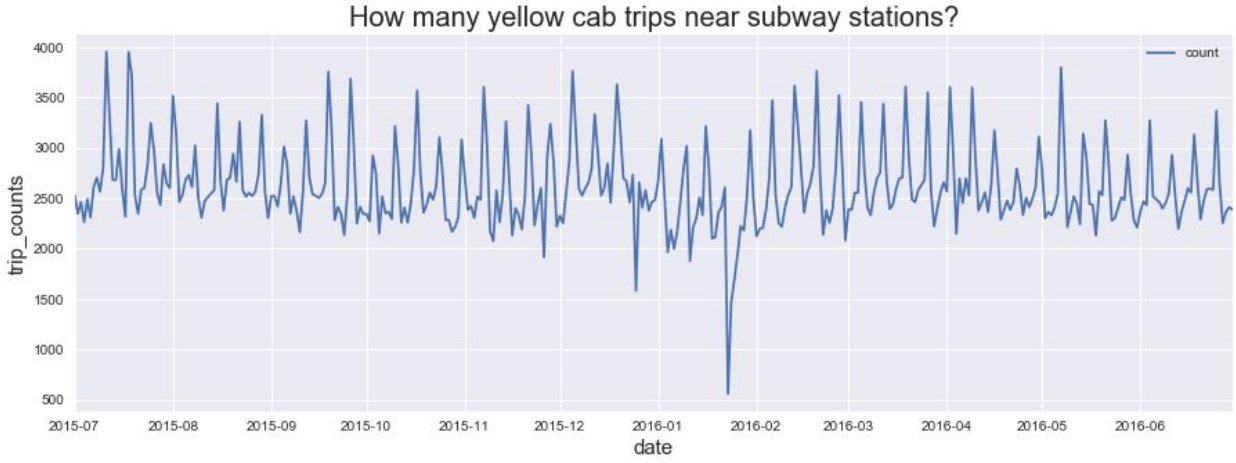

Figure 4.1.1

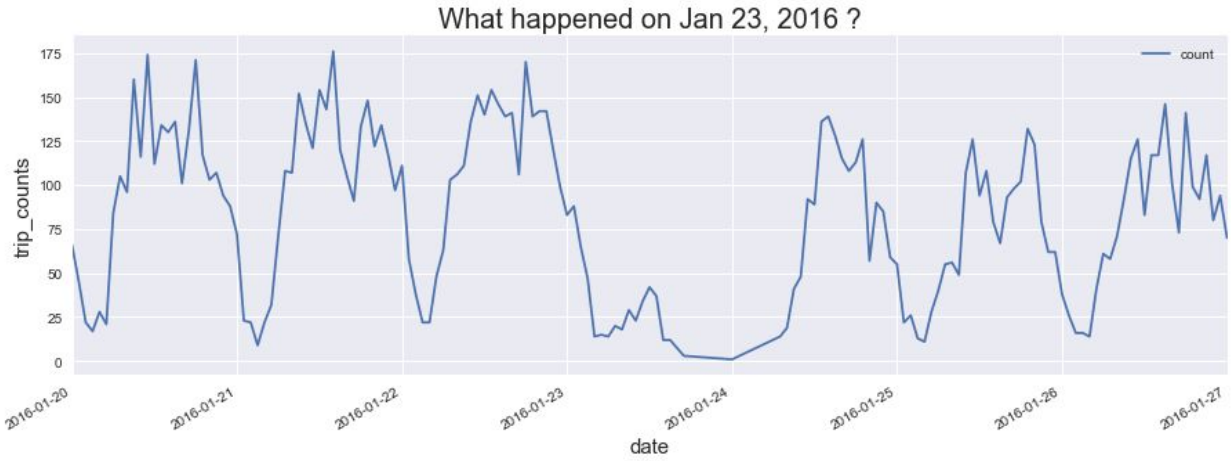

Figure 4.1.2

What happened on Jan 23, 2016? When we searched the news, we noticed on that day there was snow storm --“Winter Storm Jonas Was New York City's Heaviest Snowstorm on Record” So from this, we learned that in extreme weather, even if people would like to take a cab to avoid walking, there may not be enough drivers in the street to provide service.

After looking at the significant drop, how about the peaks? From Figure 4.1.1 the highest two peaks were in June 2015, almost reaching 4000 trips a day. Which days are they?

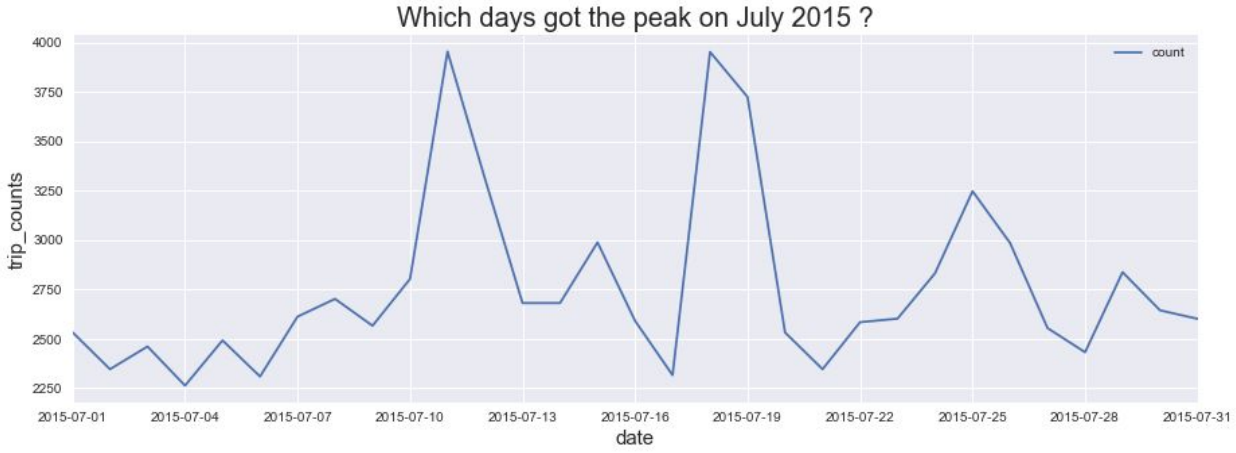

Figure 4.1.3

From Figure 4.1.3 we find out it was June 11, and June 18, 19. June 11 and June 18, 2015 were Saturday, and June 19 was Sunday, s we could say that people who used yellow cab instead of no-transfer subway trips were not going for work. Instead, yellow cab was serving people going out for recreation. From Figure 4.1.4, we notice that after midnight of June 18, the trip count is still quite high, which confirms that yellow cab is more attractive to people hanging out late at night. Based on the highest peaks were during the summer holiday, we could also guess tourists are the primary contributors to these trips.

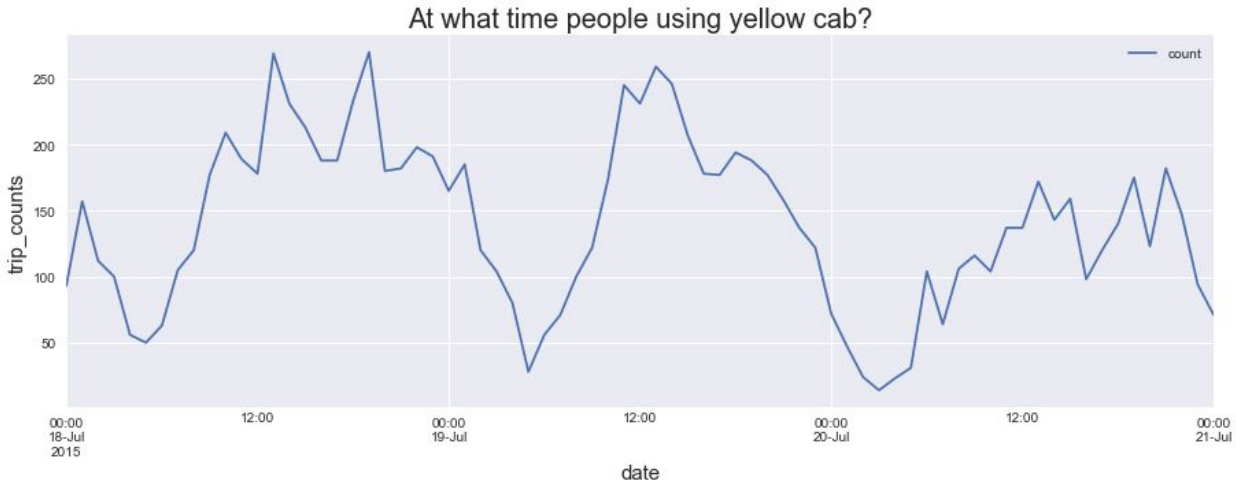

Figure 4.1.4

But are these peaks statistically significant anomaly? As there are a lot of similar peaks like these in Figure 4.1.1, we use a statistical test to identify whether they are an anomaly. Firstly, take a look at the distribution of trips, as Figure 4.1.5, and it seems not a normal distribution.

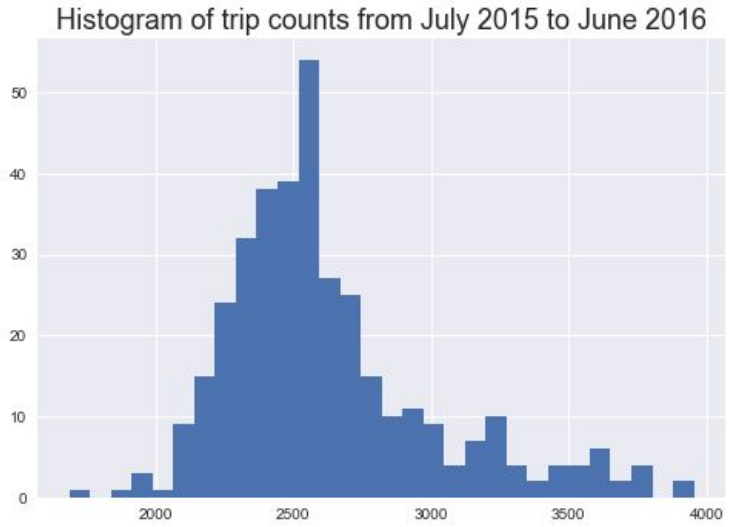

Figure 4.1.5

Set alpha as 5%, and run D'Agostino's K2 test with a null hypothesis that trip counts are normally distributed. It returns p-value 1.38e-15. So we have to reject null hypothesis. As the data is not normally distributed, we cannot use t-test, z-test, so we decide to use boxplot to find the outliers.

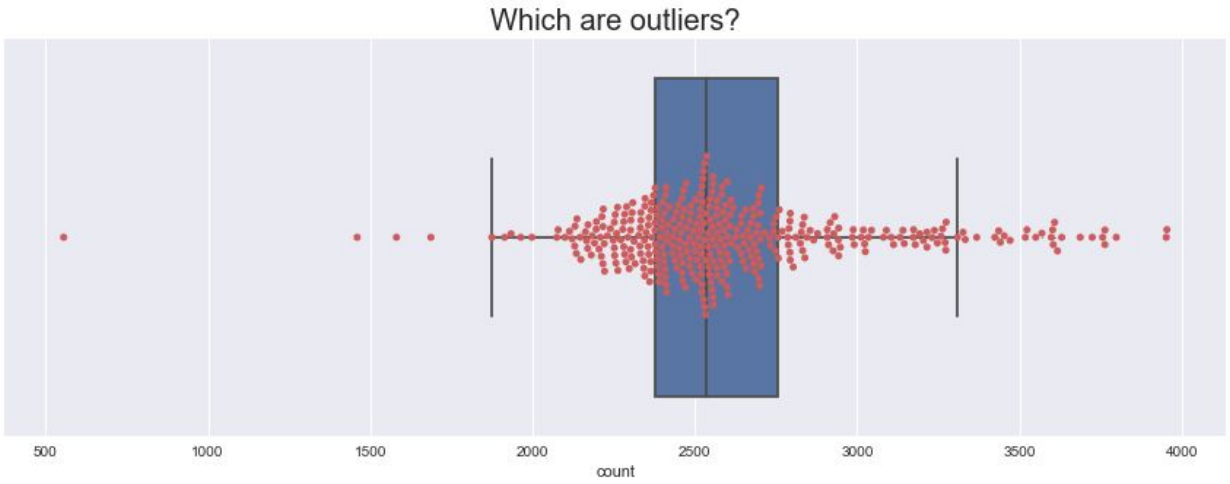

Figure 4.1.6

The boxplot returns outliers. The upper bound and lower bound shown in time series as in Figure 4.1.7. A lot of peaks are picked as an anomaly , and all of them is on weekends, especially Saturday.

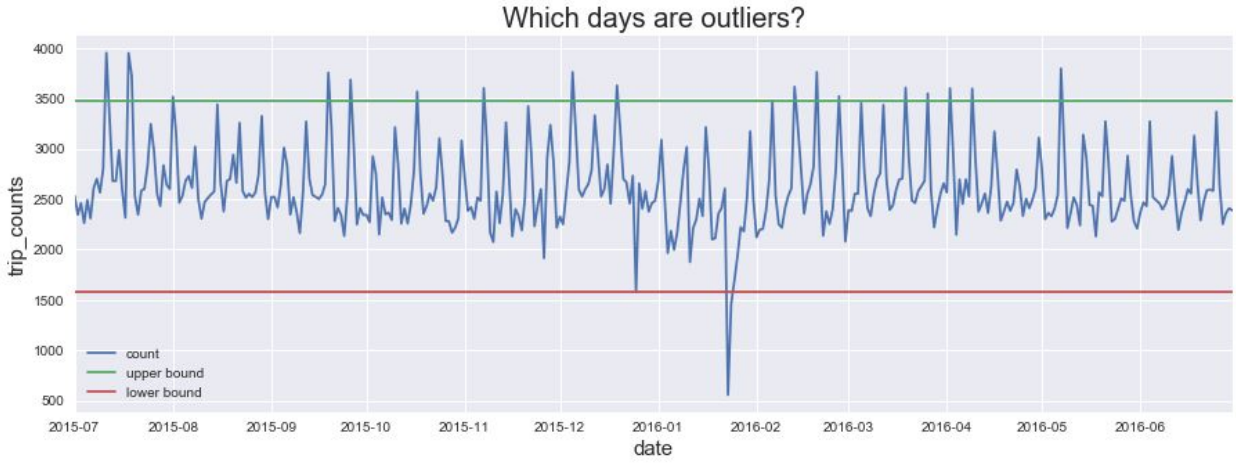

Figure 4.1.7

4.2 Anomaly Hour

After separating trips by lines, we notice C line has an anomaly hour. It was Jan 10, 2016, 11 am. Why has that time a significant peak?

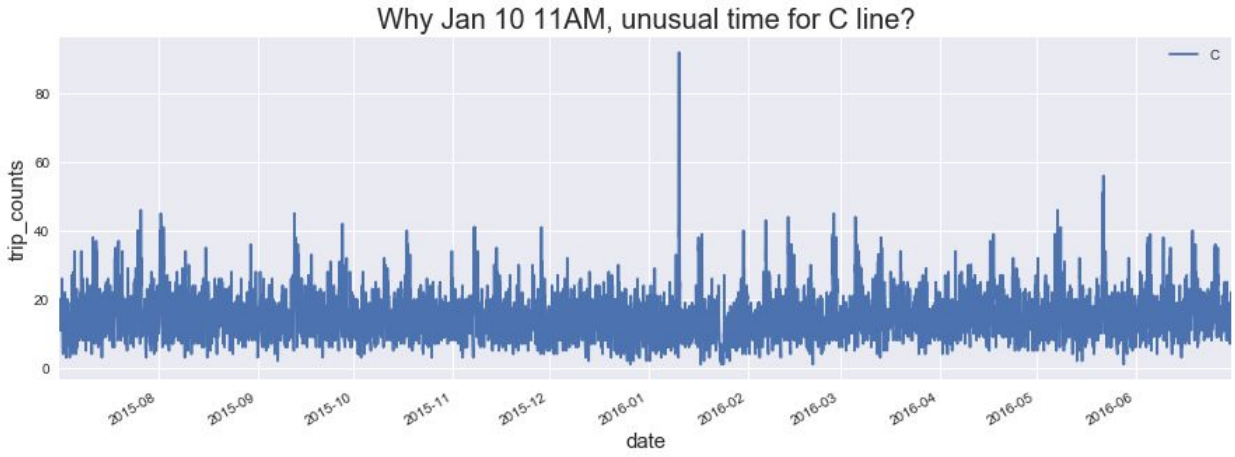

Figure 4.2.1

As there were no significant events or breaking news happened that day we find could cause such a peak, then we looked up weather. Then we find out that it was Sunday, and from Figure 4.4.2 we can see the weather that day got suddenly warm since 10 AM compared to the weather before and after Jan 10.

SUN 1/3	MON 1/4	TUE 1/5	WED 1/6	THU 1/7	FRI 1/8	SAT 1/9
Actual Temp 45°/35°	Actual Temp 36°/14°	Actual Temp 29°/11°	Actual Temp 41°/25°	Actual Temp 46°/31°	Actual Temp 46°/31°	Actual Temp 47°/40°
Hist. Avg. 39°/28°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°

SUN 1/10	MON 1/11	TUE 1/12	WED 1/13	THU 1/14	FRI 1/15	SAT 1/16
Actual Temp 59°/40°	Actual Temp 40°/26°	Actual Temp 44°/25°	Actual Temp 30°/22°	Actual Temp 38°/22°	Actual Temp 51°/34°	Actual Temp 52°/42°
Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°	Hist. Avg. 38°/27°

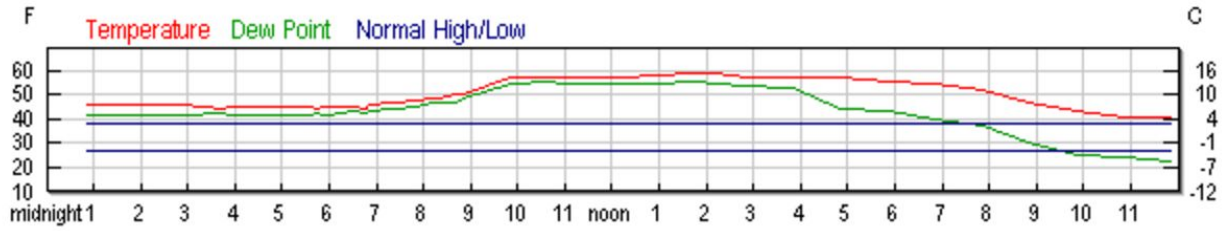

Figure 4.2.2 Source: NWS Daily Summary

From this, we learned that good weather is more attractive for people going out and taking a cab instead of the subway.

4.3. Monthly Pattern

The figure 4.2.1 show the monthly aggregate mean count of all pick-up for 12 months. From this figure, we can find that the mean of monthly pick-up on January was significantly smaller than other months. As we have shown in the previous progress, there was a massive snow storm in January; thus there were fewer pickups in January due to less taxi available as well as fewer people will go out during that period. To have a comprehensive understanding of this phenomenon, we have to check other characteristics of the monthly pattern. The figure 4.2.2 shows the monthly sum count of the pick-up during 12 months. We can see that the pick-up number has slightly dropped every two months, basically because some month has less day than others, especially February.

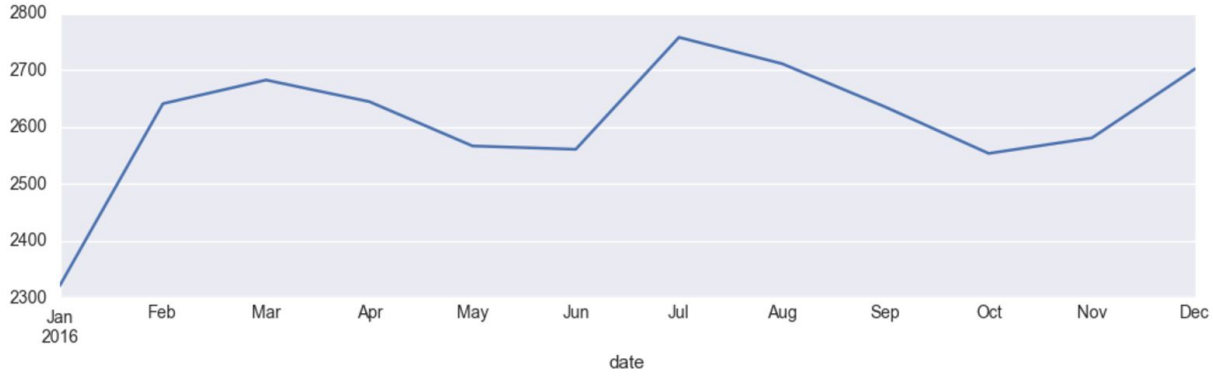

Figure 4.3.1

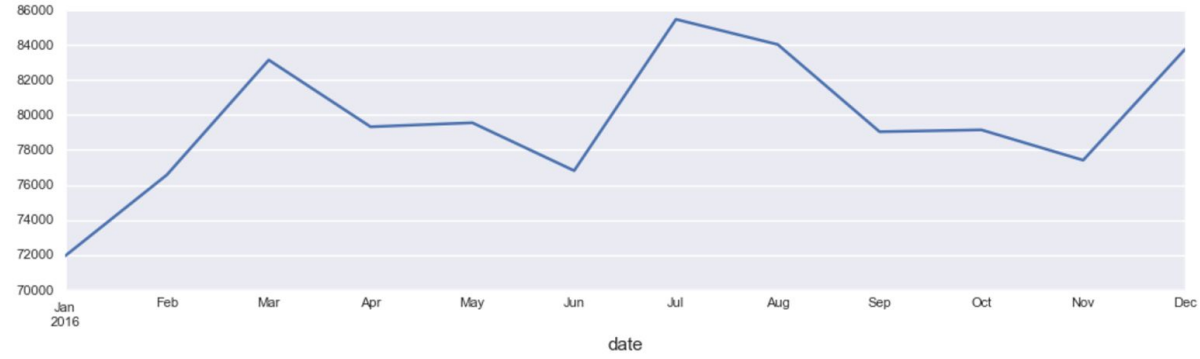

Figure 4.3.2

4.4 Seasonal pick-up pattern for lines

After we have detected the timely pattern of pick-ups, we also want to check if there are any timely patter differ by lines. So we divided the time into seasons and checked the count of each line for season(Figure 4.3.1). From the output figure, we can see that there was no significant difference for each line by seasons.

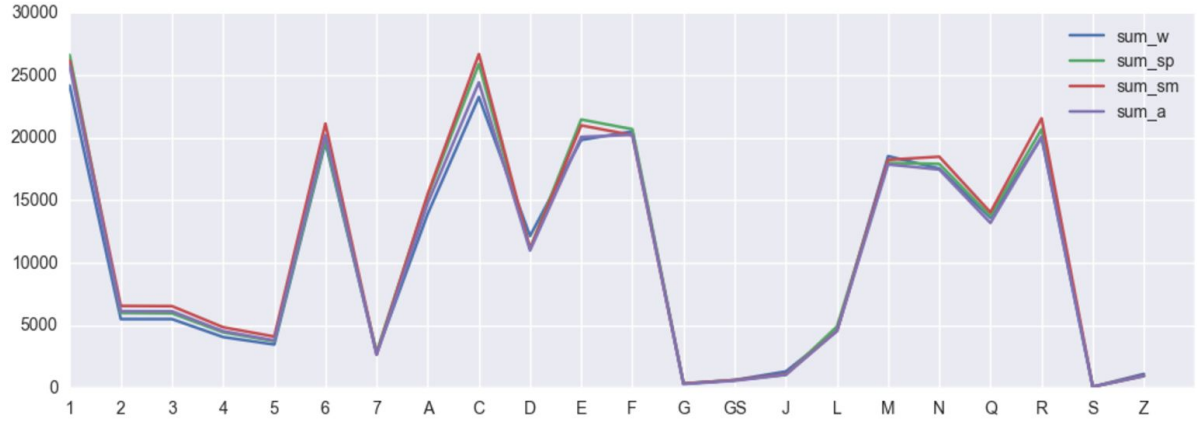

Figure 4.4.1

Figure 4.4.2

4.5. Which line is most undesired?

Seen from Figure 4.5.1 (the legend ranks by the trip counts decreasingly), through a year, the total yellow cab trips along subway line 1 ranks first, and line C ranks second, line R ranks third. As both line 1 and line C has more than 100,000 trips, are their difference statistically significant? From Figure 4.5.2, we notice the distribution of yellow trips along line C and line 1 are not skewed, and it is approximately normally distributed, so we use two sample t-test to verify.

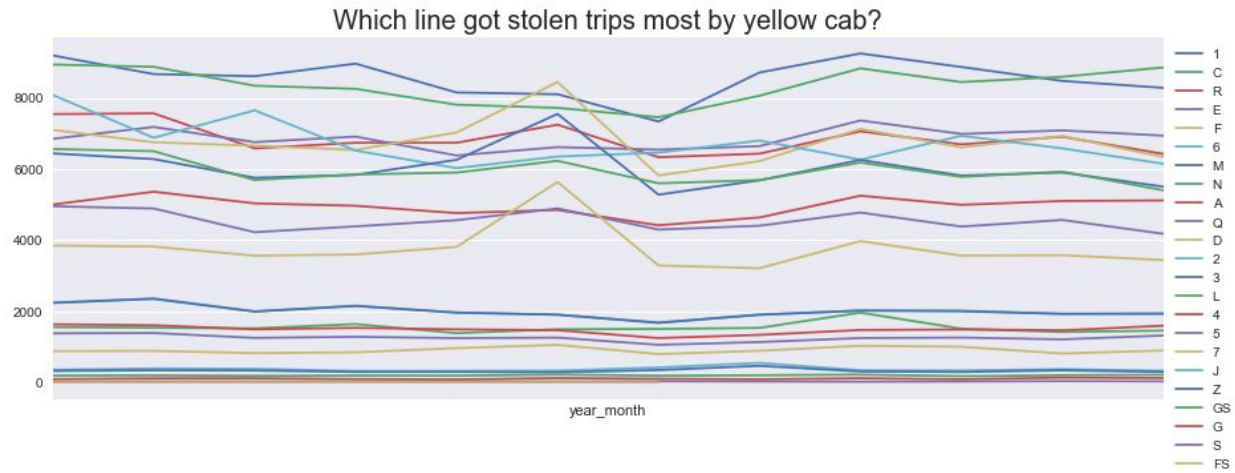

Figure 4.5.1

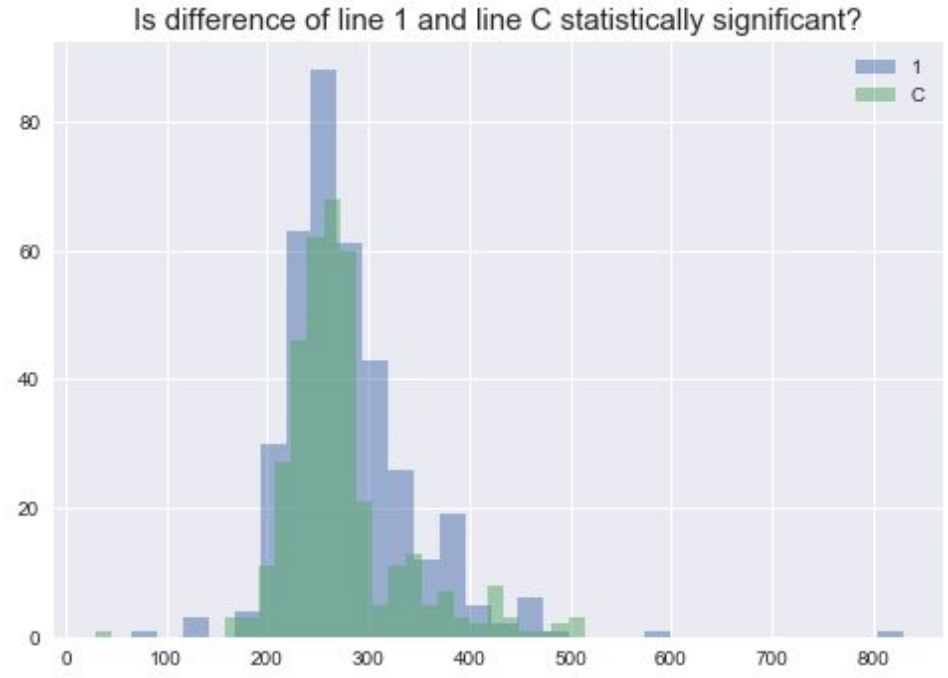

Figure 4.5.2

Two sample t-test has p-value 0.17, which means we cannot reject the null hypothesis that this two sample has equal mean. So line 1 and line C don't have a significant difference. Then we use two sample t-test test line 1 with line R , and line C with line R again. Both P-value is less than 1e-32, so the differences of line 1, C with R are significant. Then we would say line 1 and line C got 'stolen' business most by the yellow cab.

4.6 Which two lines are most similar ?

Using Pearson's correlation we compare each pair of lines, we filter out pairs with the p-value less than 0.05, and Pearson's correlation coefficient more than 0.9. The pairs are: (2,3), (J,Z),

(N,R), (N,Q), (Q,R), (5,4), (F,M). The pairs are listed in the order of Pearson correlation coefficient decreasingly. For example, line 2 and line 3 are highly alike, with p-value near 0, and Pearson correlation coefficient as 0.9948. Also, line J and line Z are highly alike as well, with Pearson correlation coefficient as 0.9804. But as line 2 and 3 are almost on the same track, so are line J and line Z, it's not surprising to find out this result. We can see it from Figure 4.6.1

Figure 4.6.1

All these pairs share tracks to some extent. From these pairs, only (F,M) are not on the same tracks for long. They share the track from Broadway/Lafayette St to 47-50st Rockefeller Center, only seven stops altogether. However, they still get 0.9305 Pearson correlation coefficient. So we can infer that the area between Broadway/Lafayette st to 47-50st Rockefeller Center is the yellow cab densely used area.

5.Conclusion

1. People use yellow cab as a transport mode when go for recreation, not for working commute.

2. When weather is pleasant, more people go out with yellow cab; when snow storm comes, the trips drop, maybe because drivers are not working as well.
3. Saturday has significant peak on yellow cab trips.
4. People will choose yellow taxi other than subway in Winter and Summer.
5. Broadway/Lafayette st to 47-50st Rockefeller Center is the active area people using yellow cab a lot, even if subway is really accessible as well, maybe due to tourists.
6. Line 1 and C got 'stolen' business most by yellow cab.

6.Future

1. Since the population density at the location of each subway station varies, some results of our study should be modified in the future works. We would make a normalization to minimize the effect of different population density in our research. The dataset will be used to represent the population density at each subway station in the future is 2015 New York City Subway Complexes and Ridership. It could be accessed at the website of GIS Lab of Baruch, CUNY. Annual, average weekday and average weekend ridership is provided for 2007 to 2014 (excluding Staten Island Railway stations).
2. The length of each subway line also varies. There may be much more subway stations in a longer line. People may tend to take the subway to save their time because many stations before destination would cost extra time. A normalization on different subway length will also be made to minimize the effect of subway line length on our research.

7.Reference

1. Jianting Zhang. 2012. The City University of New York. Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC.
   (Jianting Zhang 2012)
2. Joya A. Deri, José M. F. Moura. 2015. Carnegie Mellon University. Taxi data in New York city: A network perspective.
   (Joya and José 2015)
3. Kenneth Stuart, Marc Mednick, Johanna Bockman. 2014. New York University. Structural Equation Model of Customer Satisfaction for the New York City Subway System.
   (Kenneth, Marc and Johanna 2014)
4. Ricardo Lagos. 2003. An Analysis of the Market for Taxicab Rides in New York City.
   (Ricardo 2003)