Big Data Project

Utku, Nikita & Deepti

[ue225, nn1119, dv697]

INTRODUCTION

The NYC Taxi and Limousine commission has made available an enormous, detailed dataset, covering over 600 million taxi trips. The staggering amount of information, spanning 7 years opens the opportunity to explore transport and traffic habits of one of the most populated cities in the United States. With the study we conducted on this data, we found many an insight into the transport habits and preferences of New Yorkers. Though taxi usage appears to be declining, is yellow taxis are still a strong pillar of the transport system in New York.

GOALS:

The taxi data is enormous, at around 25G. So our initial goal was to first parse through the all data and find any data quality issues. We report our findings on this front in Part I. We then wanted to study the taxi habits of New York residents and tourists. To do so, we studied a stream of hypotheses which are discussed in detail in Part II. Ultimately our goal was to understand how and when New Yorkers take taxis and whether the taxi institution is as strong today as it once was.

Part I: Data summary and data quality issues

Every file had a header and not all files have the same header. Header's are different for most of the file. Hence, there was no unified way to delete them. We handled this part in our reader function.

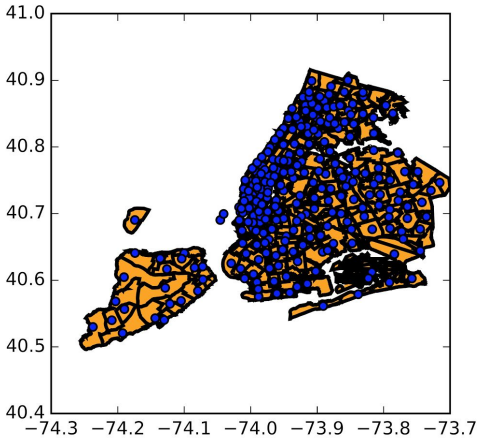
The schema was different every 2 years.

- 2013-2014

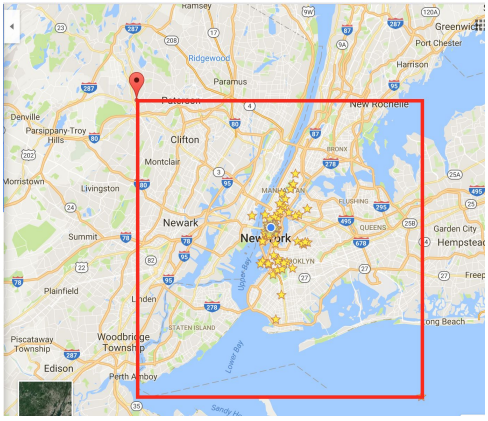
    - Used strings for vendor id and payment type.
    - Improvement surcharge was missing.
- 2015-2016 (till June) matched what was given in dictionary
- After 2016 June, location ids were used instead of gps coordinates.

While checking the number of fields in files we found that before 2015 the number of fields were just 18 instead of 19 as given in the dictionary file. The field that was missing was improvement surcharge. Improvement surcharge was levied in 2015. So a column was inserted with value 0.0 for files before 2015.

Apart from this we found that the last 6 months of data contained location id's instead of gps coordinates. The data after June 2016, doesn't have the four GPS pickup-dropoff fields anymore. Instead there are PULocationID and DOLocationID indicating TLC taxi-zones locations. There are 263 zones and additional 2 unknown zones. To merge these two data types we wrote a function to convert the gps coordinates into their corresponding zones by using the .shp file provided. We used pyshp library to load the .shp file and used Lambert Conformal Conic projection to project the points into longitude–latitude pairs. Than we used matplotlib to verify whether a point is in a polygon. We search all 263 polygons to find out the zone, for each point. It turned out to be quite slow for the big data we have. Therefore we decided to make the conversion other way around. We assigned a fixed GPS coordinate by calculating the mean of the polygon points for each zone. On the right you can see the plot of the zones and the mean points calculated.



There are two unknown zones enumerated 264,265 and we converted this values as NULL's and counted them as invalid. We've also checked whether the drop-off/pick-up locations are inside NYC or-not by checking the coordinates are inside the enclosing rectangle or not.



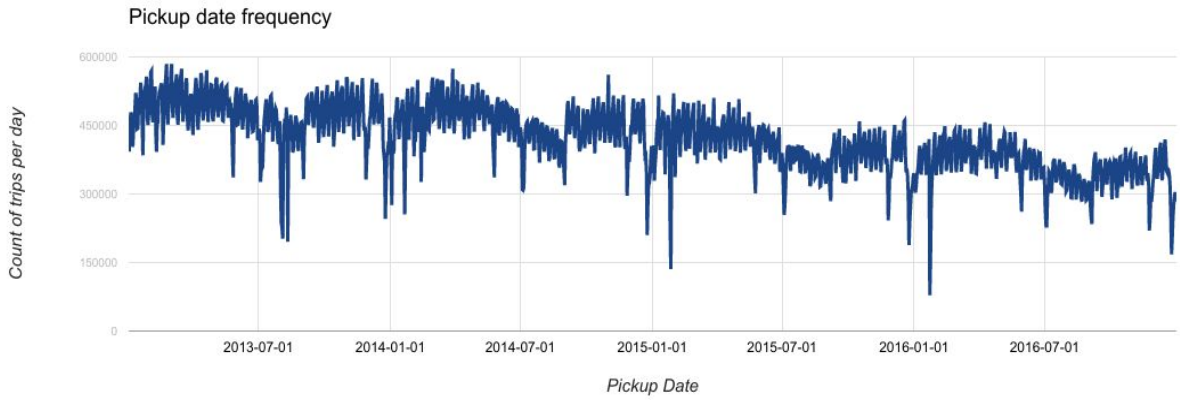
We found that very few drop off dates (<< 1 percent) had invalid year.

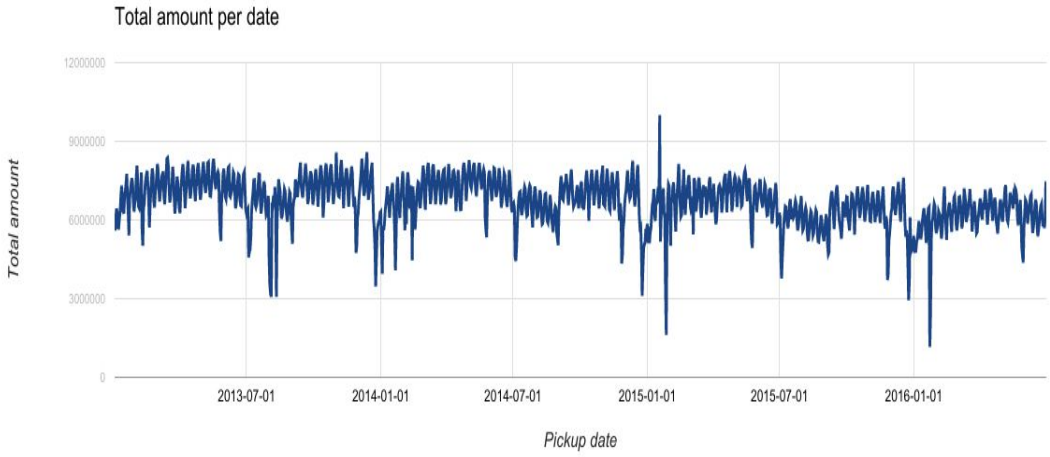
Fields like vendor id, payment type, pick up date did not have any invalids. But trip distances, amounts we found many negative values. There are also cases when the pick up and drop off coordinates match but trip distance is greater than 0 and vice versa.

Even though improvement surcharge was levied only 2015 we see a huge amounts of null which is surprising as it can be either 0 or 0.3. We are planning to investigate it further.

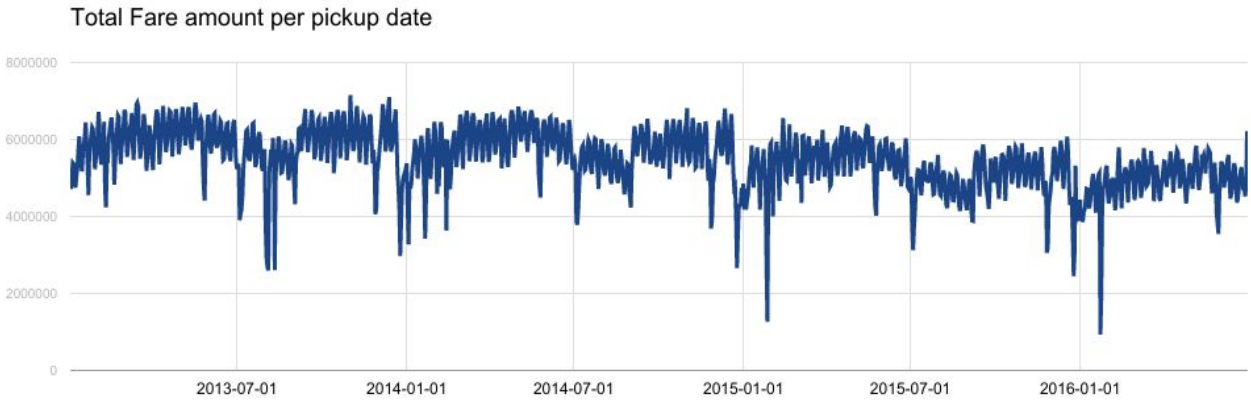
Data Summary:

- Generally the number of trips per day should coincide with total amount generated per date. But as seen from the plots below, there are some instances where they don't match. For example on date 2015-01-18 there was a peak in the total amount but the number of trips did not peak. Hence, it is an outlier. Like this we noticed about 647 outliers.

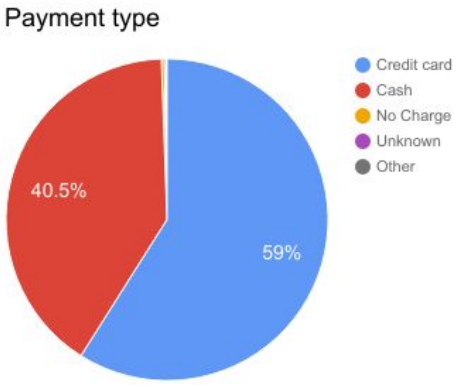




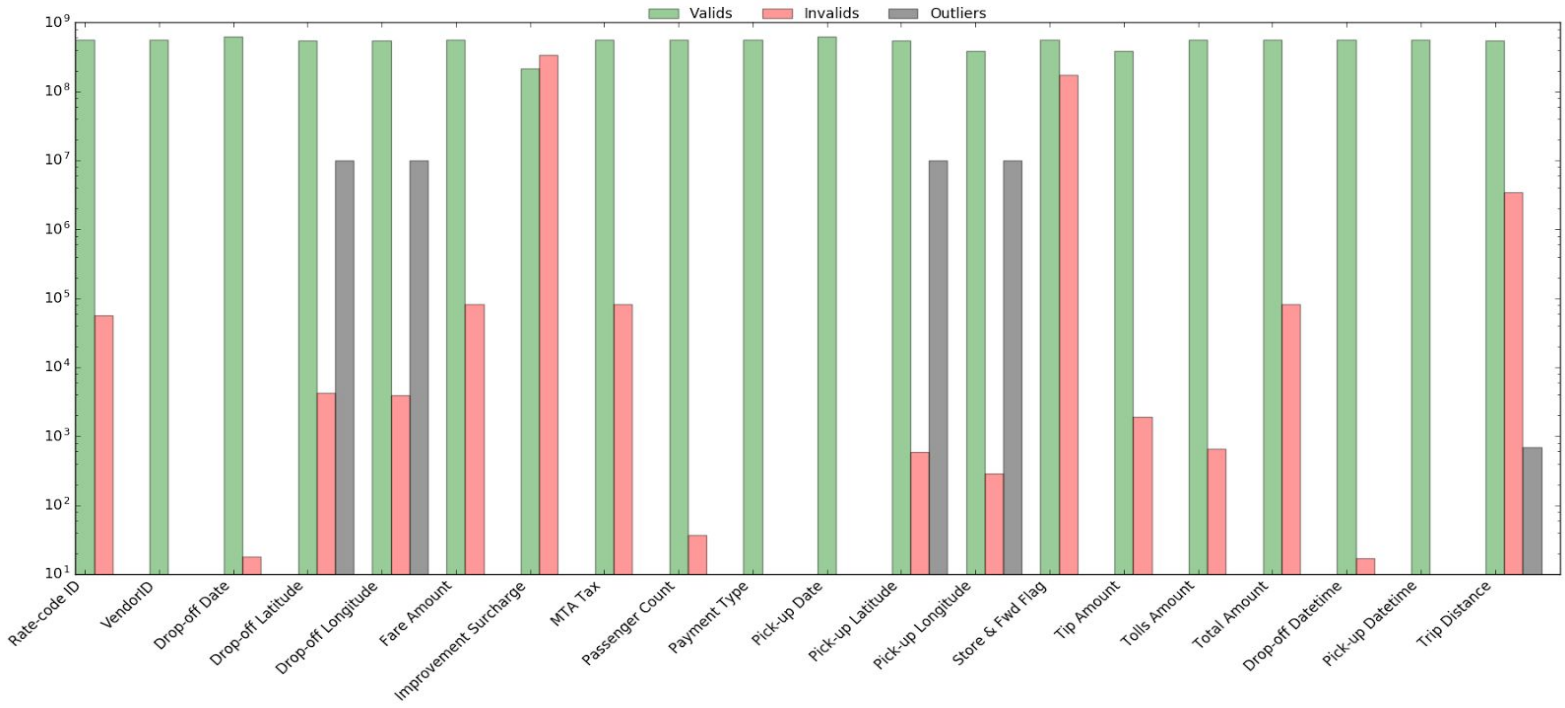
- We observed that the fare amount holds a better correlation with the number of trips per day. Clearly, we can conclude that the total amounts field has errors/faulty values.



- Looking at payment data it is interesting to see that many customers prefer using credit card over cash.



Finally, we present the following bar plot which shows the ratio of Valid, Invalid, and Outlier data points per column. The plot is presented on a log-scale due to the comparatively small percentage of invalid and outlier data for most columns.



- Issues per column:
    - Vendor Id: In general Vendor Id is pretty consistent with the dictionary after January 2015. Prior to 2015 we found that for vendor id, instead of integers 1 or 2 as mentioned in the dictionary and in data after 2015 Jan, strings “CMT” and “VTS” were used. Similarly, for payment type strings like “CRD”, “CSH” were used instead of integers. Hence, we used the following mapping to transform the data.

  Mapping table for vendor id that was used for data before 2015 for data cleaning is given below:

TPEP provider	Integer Code
CMT	1
VTS	2

    - Dropoff Date: We found 18 dates out of the whole column to be invalid. The rest of it were date format and the date between 2013-01-01 to 2017-01-01

- Pickup Date: No known issues were found with values in this attribute. But when both Pick up and drop off dates were compared, we found 1252247 records with pickup and drop off dates to be equal and 24233 records we found where drop off time was before pickup time.
- Store & Fwd Flag: This attribute has an inordinate number of invalid values (nearly 30 prct). But on further investigation, it was found that this primarily comes from 2013 and 2014 data. Starting Jan 2015 the field was populated with 'Y' or 'N'.
- Improvement Surcharge: There was a really high percentage of invalids in Improvement Surcharge which was really intriguing. A lot of the data prior to 2015 for the column were nulls.
- Trip Distance: 99.3% of the data we found was valid. And out of the valid we calculated the mean and the standard deviation and found 689 records as outliers(i.e. Anything greater than 4 standard deviations). Out of all the invalid trip distances we got 3,438,233 records with trip distance as 0 (0.6 % of the data) and 7 records with trip distance being negative.
- Tip amount: As seen from the plot 42.8% of the trips had zero tip amount. Negligible amount of trips had negative amount (1886 of total trips till June 2016). We also noticed about 400k trips which had tip amount greater than fare amount.
- Payment Type: For years 2013-2014 we found that they used strings instead of integers (as part of the dictionary) . Hence, we used the following mapping table to clean the data prior to Dec 2014.

  Payment type mapping for data before 2015:

Payment type	Integer code
CRD	1
CSH	2
NOC	3
DIS	4
UNK	5

CONCLUSION on DATA QUALITY

We found that there are some data issues with a few of the columns which have a high percentage of invalids. Additionally, there is a non-negligible percentage of outliers in the pickup and dropoff coordinate data, these outliers are coordinate values that lie outside New York city limits. Way may look into these further and do a more detailed filtering and include some areas within New Jersey (specifically due to Newark airport).

Ultimately though, this data is clean and cohesive enough to base further investigative work on.

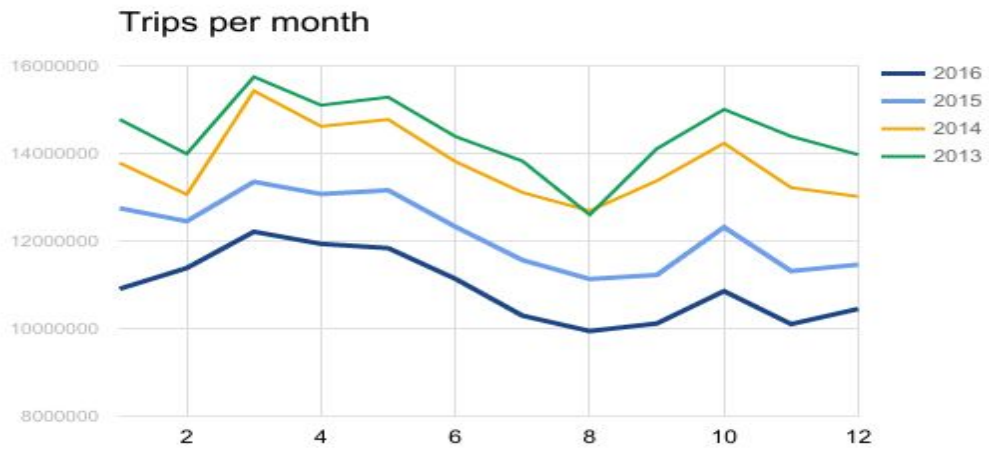
Part II

After getting a sense of what NYC taxi data looks like, and noting the anomalies we already observed, we built a list of questions and hypothesis we wanted to answer. We used these as a guiding mechanism to dig deeper into the data. Since one of our motivating objectives is to gain insight into New York residents and visitors by means of studying their taxi usage, we looked at a diversity of hypothesis that question how and when customers use taxis.

Hypothesis 1: Seasonality of Taxi Trips

New York is an extremely walkable city but suffers from harsh winters. Therefore, we intuitively assume that taxi trips will increase during harsher climates and fall when the weather is more walkable. So we expect the data to show an increase in trips during the winter season and a drop in the summer. Furthermore, during the fall, as families return from holidays and students return to school, we expect an increase in taxi usage.

The following is a plot showing taxi usage by month, over 4 years. We see a clear trend were taxi usage is greater in winter months than summer months. [In the plot we notice an October spike, which is a curiosity we looked into further. Unfortunately, we didn't find any clear source of this spike. The taxi trips in October have the same location distribution as months like August. We suspect that October might be high tourist season, so the distribution of taxi usage remains unchanged but there is simply increased usage and frequency due to an influx of tourists.]

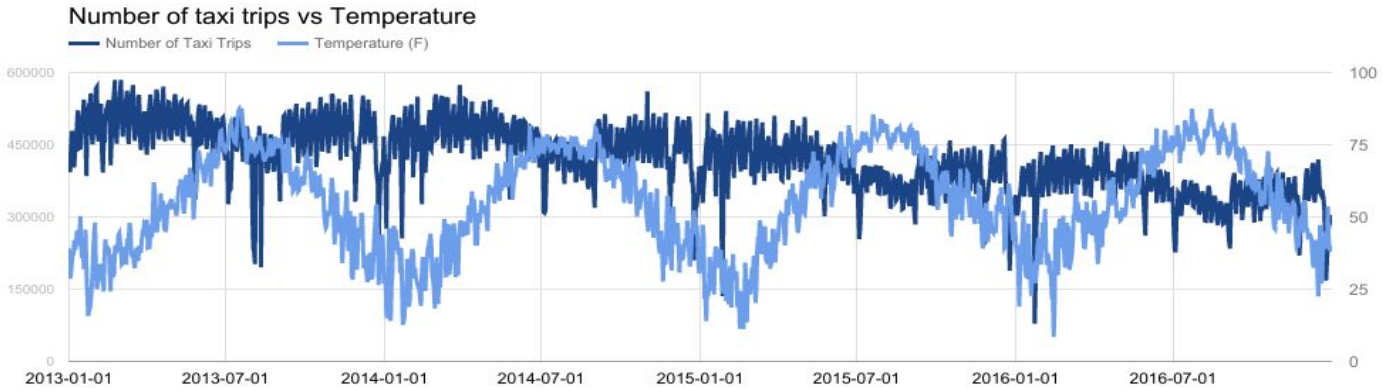


(Script: *run_trip_frequency_by_month.sh*)

The Pearson correlation coefficients we calculated are listed below,

- Number of trips with temperature = -0.2324659977505238
- Number of trips with wind-speed = 0.09900347208933871
- Number of trips with precipitation = -0.0044164110844505935
- Number of trips with snow depth = -0.11470904193132954
- Number of trips with temperature for winter months = -0.22551026620056025
- Number of trips with temperature for summer months = -0.26307551324797235
- Number of trips with temperature for fall months = -0.16402766528471788

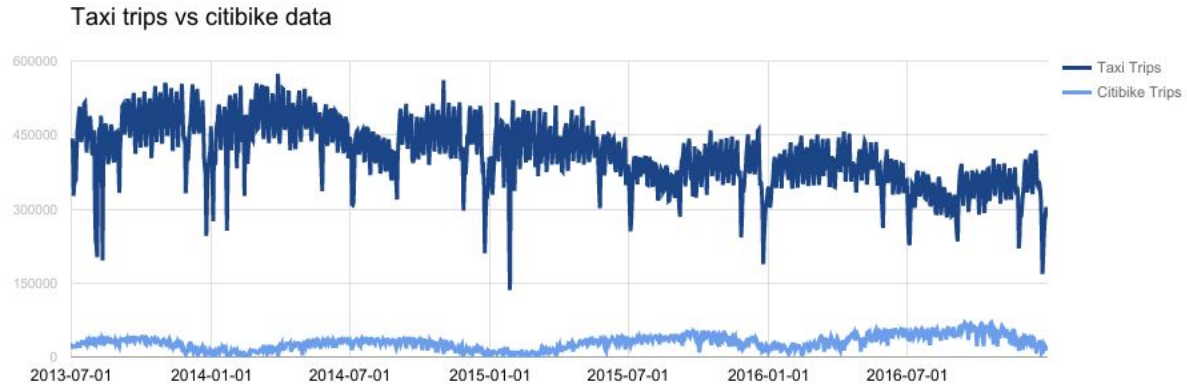
The following plot simply shows the cyclical weather patterns alongside taxi trip frequency.



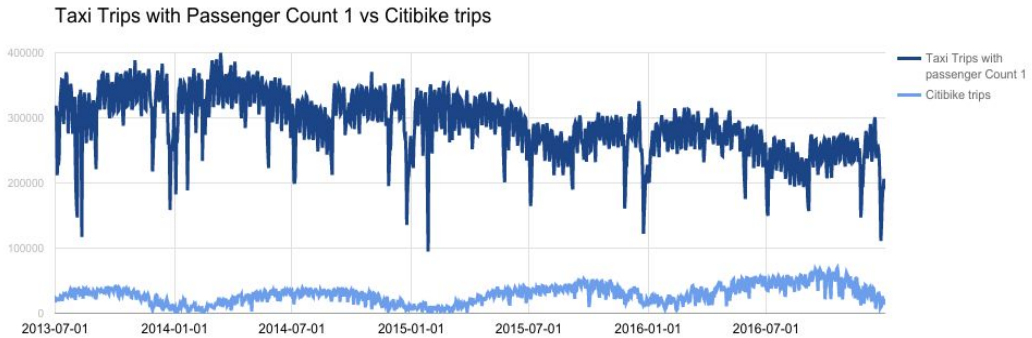
Hypothesis 2: Overarching Decline: Uber, Lyft, Citibikes

With New York being a tech savvy city, and the advent of car sharing apps like Uber and Lyft, we expected yellow taxi usage to decline. Unfortunately, Uber and Lyft data aren't publicly available. Citibike data however, is available. Even though Citibike usage is undoubtedly much lower than taxi usage, we do hypothesize that bike trips increase over the years as taxi trips decline. We expect this because the world is becoming more concerned with physical health, and New Yorkers specifically care about being healthy and drinking and eating only things that are tinged with green.

So, we compared citibike data with taxi data. We expect that summertime usage of citibikes to be much higher than the winter. We see a clear cyclical trend in the plot below. We also observe that the summertime usage of Citibikes rises year after year, helping to confirm our hypothesis.



The above plot consider all taxi trips. However, we thought it would make more sense to compare taxi trips with a single rider to Citibike trips. The following plot does just that,



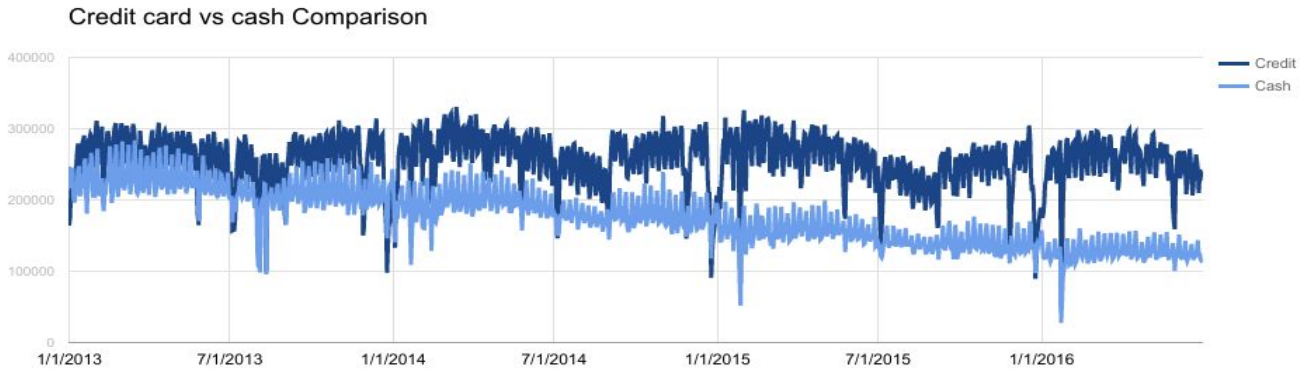
The above plot shows a clear inverse correlation in citibike usage and taxi usage. The Pearson correlation values we calculated (using scripts: *run_citibike_correlation_script.sh*) are as follows,

- Correlation value of number of trips with citibike number of trips = -0.31036
- Correlation value of number of trips with citibike number of trips with passenger count 1 = -0.24985
- Correlation value of number of trips with citibike number of trips in summer and Spring months = -0.45456
- Correlation value of number of trips with citibike number of trips in summer and Spring months with passenger count 1= -0.37567

Hypothesis 3: Credit Card Usage and the Death of Cash

Carrying cash is becoming an increasingly unnecessary chore when living in a city where everyone, from your corner bodega to the cool new coffee shop, accepts credit card. This report confirms the hypothesis of the advent of credit card payments. Therefore, we anticipate credit card payments with taxis to also increase over time.

The following plot shows a clear, steady, decline in cash usage over time. The co-occurrence of an upward trend of credit card payments is less evident but we expect the overarching decline in taxi usage offset any such trend.



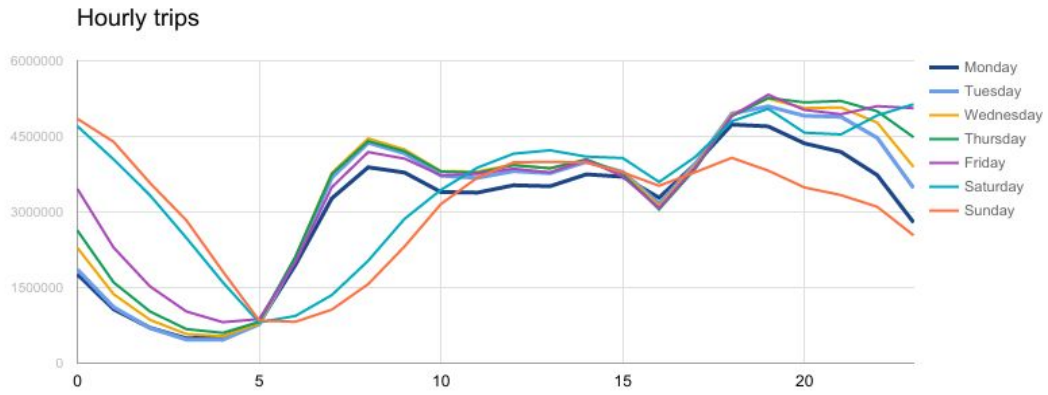
(frequency data from script: *credit_card_usage_frequency.py*)

Hypothesis 4: After Dinner and Brunch-time Cab Ride

As people head home overstuffed, exhausted, or inebriated, after dinner, we imagine they often

take taxis. We hypothesized that there would be an increase in taxi rides in the evening, post-dinner hours. Furthermore, with the popularity of brunch, we expect there to be a rise in taxi usage around brunch-time hours on the weekends.

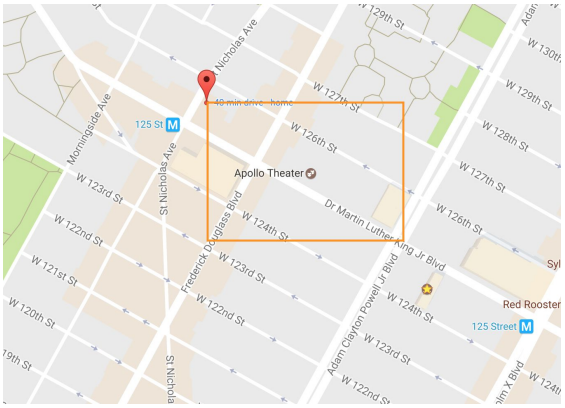
As we see in the plot below, there is a post dinner rise in taxi usage. During the weekdays, there is a sharp increase in taxi trips during the morning rush hours, from 7–9AM, as people head to work. On the weekends however, there is a steady upward slope in the morning spanning the lazy weekend brunch hours.



(script: *pickup_time_by_day_frequency.py*)

Hypothesis 5: NYC Theaters vs Taxi Data

New York is an extremely lively city, the home to many a theater and music venue. We wanted to look at taxi data specific to such venues. We expect to see different hourly characteristics around theatres in comparison to the mean.



We found that the number of taxi trips over a period of day shows different behaviour around theaters. To prove this hypothesis we used the NYC Theaters data[1] . This dataset has the coordinates of 117 theaters of NYC. Most of the theaters are around Midtown/Time Square and only 2 out of 117 theaters are in Queens.

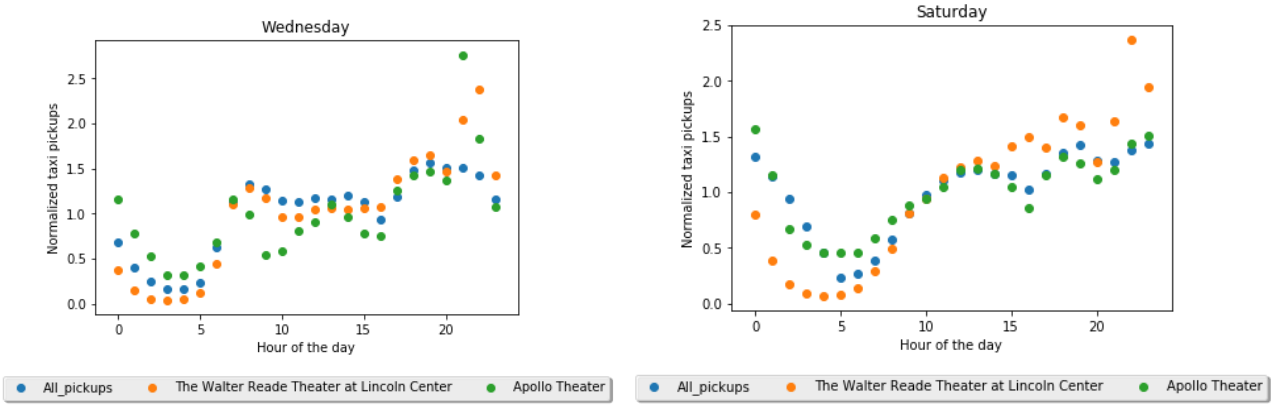
So we first we loaded the data to spark and defined a box (width=~300 meters, height=~200 meters) around

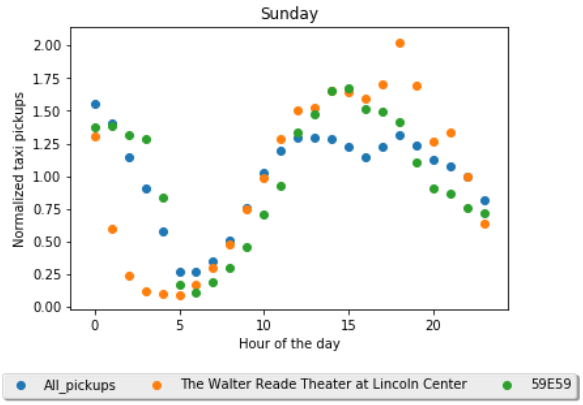
[1] https://data.cityofnewyork.us/Recreation/Theaters/kdu2-865w

the theater locations to filter the taxi trips. Since there are too many theaters in Midtown next to each other we picked some theaters out of that zone like Apollo Theater, Walter Reade Theater at Lincoln Center and 59E59. We first filtered trips for each location separately and assign a weekday/hour bin to count the total number of trips for each day of the week and hour. To be able to compare different count magnitudes and focus on the trend instead of the counts, we have normalized the counts with the mean value of each bin.

Comparison of individual theaters

Here we show that theaters have different characteristics in terms of their taxi pickup counts. We saw different trends for different theaters. For example Apollo Theater has a pretty obvious peaks around 21pm on Wednesdays, whereas it follows the mean on saturday. Walter Reade Theater has peaks both on Wednesday and Saturday. We saw that Apollo Theater hosts many shows on Wednesdays and as expected, we see different trends for different theaters, since their schedule change. The 59E59 theater has many shows on Sunday evenings and we can see the increase in the taxi-pickup counts. For the rest of the plots of every other day please refer to the jupyter notebook[2]. (script: <*theatersVsTaxi.py*>)

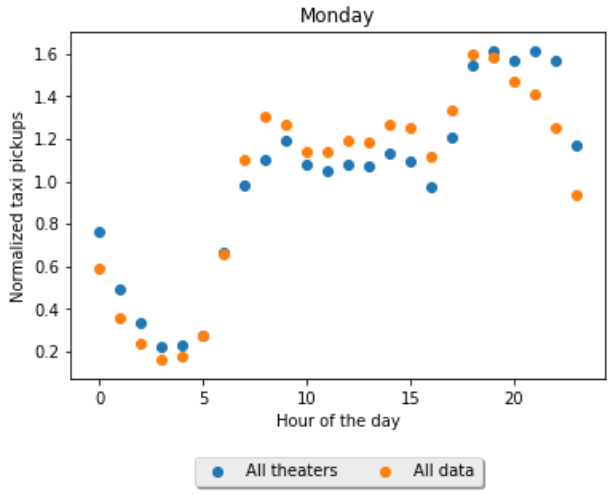


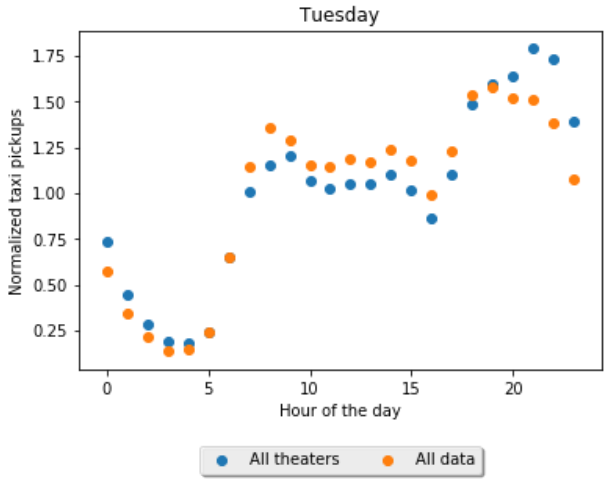


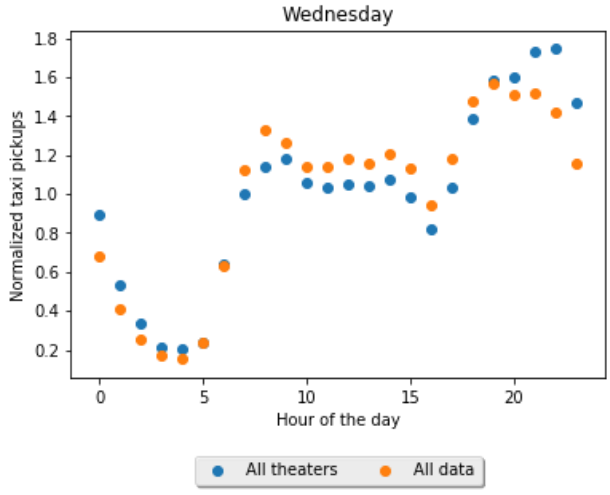
[2] https://github.com/anudeepti2004/big_data_project/blob/master/notebooks/plotTheatersData.ipynb

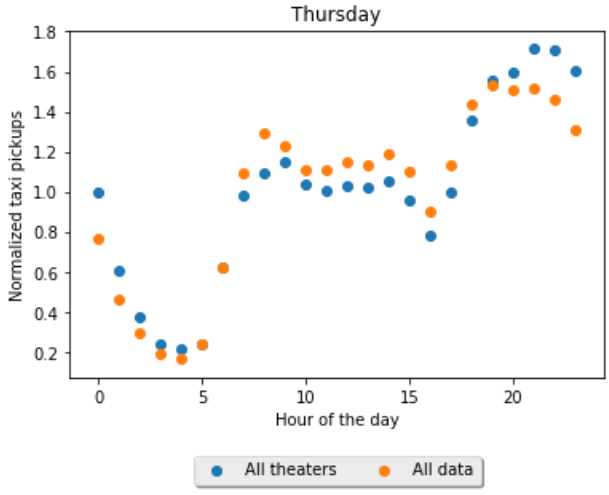
Comparison of taxi-pickup trajectory of theaters vs all

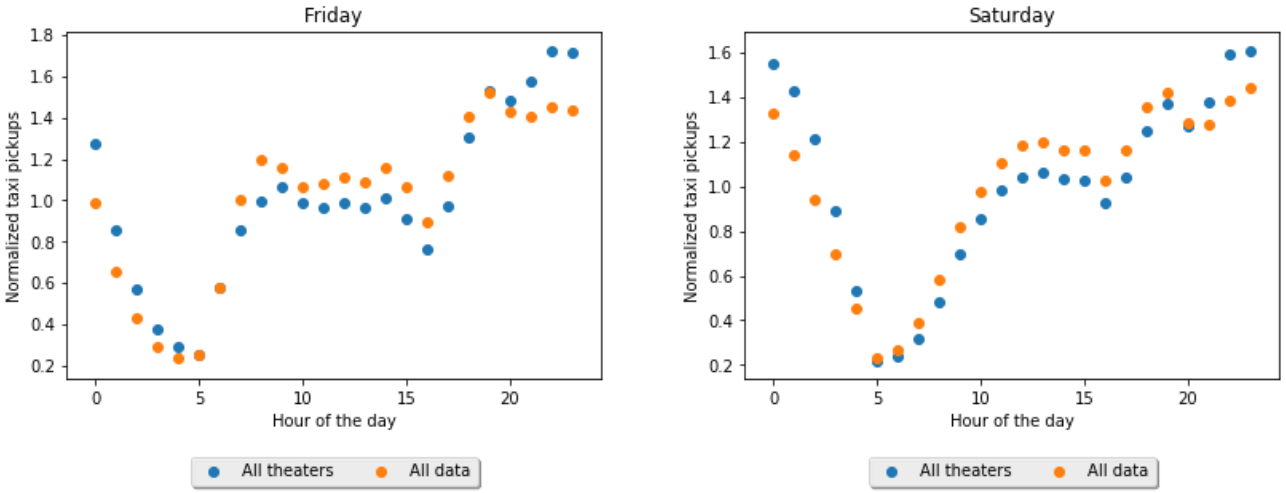
To have a bigger comparison we than filtered out all data using the all_theaters and the boxes around. We use the same box-dimensions we used for the first part and then we remove the data points which are outside of any of the boxes. Results for the seven days of the week plot below. One can clearly see that the two trajectories follow a similar trend on the day time, whereas at night, as expected, we see a significant increase in counts. This observation repeats almost in every day except on Sunday and as there are not many events on Sunday nights. But one can clearly see the increase during the Sunday afternoon.

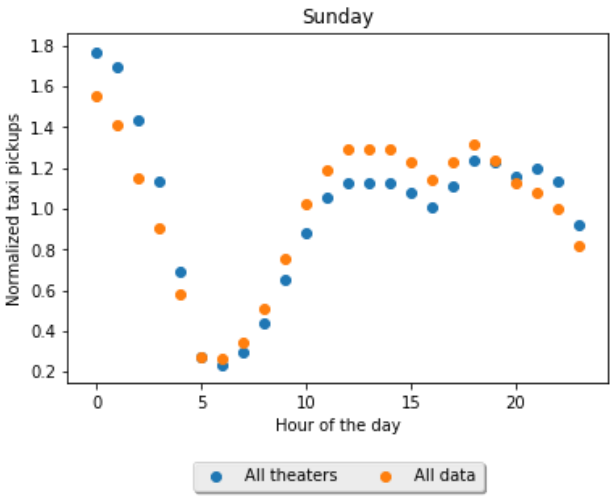






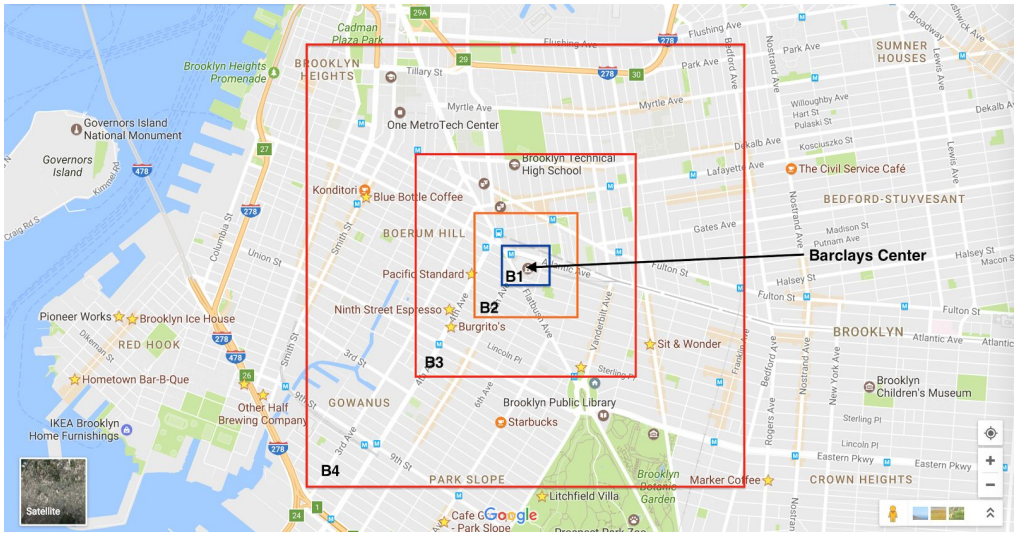




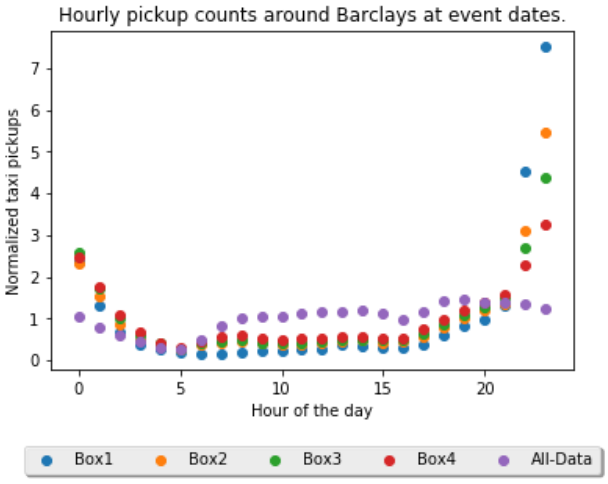


Hypothesis 6: Barclay Center vs Taxi Data

We expect a similar behaviour of taxi pickups around Barclay center. So, we investigated and did indeed find a higher activity around Barclay Center in the evening!



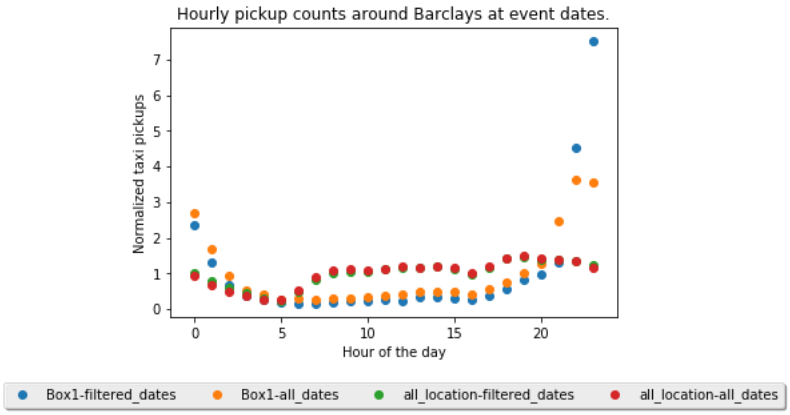
We parsed the Wikipedia page of Concerts at Barclays Center[3] and extracted the dates of the all events. There are 179 concerts since the Barclays Center is opened in 2012 with the Jay-Z concerts. We first started with the same box size that we used for the previous hypothesis and filtered out the trips according to that. Than we filtered out more trips that happened other than the event-dates. Then we counted trips per hour. We repeated this process 4 times for 4 different box sizes. As expected we observed that the behaviour of counts approaches to the mean as we increase the box size. This shows that Barclays center indeed has a special behaviour on the event dates. We plot below the mean-normalized[4] results. One can see that around 11pm on the event dates there are 7x more taxi pickups than the mean inside the Box-1. We can also observe that as we increase the box size the unusual behaviour fades out.



(script: *barclays.py*)

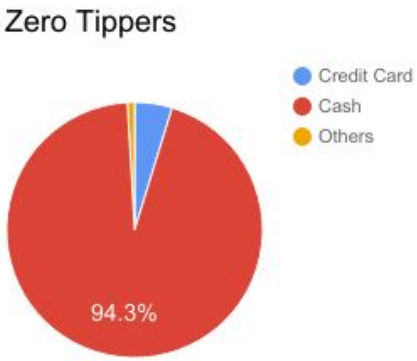
After observing this unusual behaviour we compared the event dates with with the all-data. On the plot below you can see the normalized hourly pickup counts of all the data (red). When we filter the data with the event dates then we get the green counts and it is almost same as the red curve. This shows that the event dates follow the general distribution. However when we repeat the same steps with the pickups happened inside the Box-1(around Barclays Center) we see that on the event dates the taxi-pickups are doubled compare to the normal trend.

[3] https://en.wikipedia.org/wiki/List_of_concerts_at_Barclays_Center
[4] Same normalization explained in the previous section



Hypothesis 7: Zero Tippers

While looking into the standard tipping practices of New Yorkers, we noticed that many trips had zero tips. These strikes as extremely unusual and we believe that this is likely because taxi drivers don't report cash tips. The yellow taxi dictionary file mentions that, "Cash tips are not included", leading us to conclude that this is a common and acknowledged practice.



(Script: *zero_tippers_by_paymenttype_frequency.py*)

Hypothesis 8: Tip Correlated with Fare Amount

Further in the direction of tipping practices, we hypothesized that there should be a strong correlation between tip and fare amount. We excluded zero-tip events for this study to avoid

skewing the analysis.

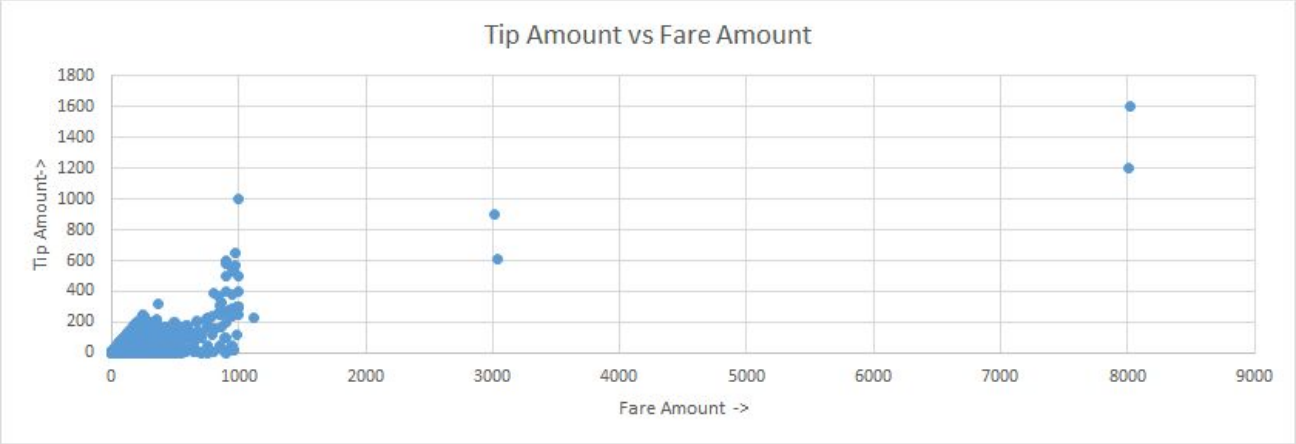
A curiosity we discovered is that there are 463,835 records where tip amount is greater than Fare amount. Some of these instances are extremely unusa, with a 3 million dollar tip for a $22 fare amount, we think it is a fare to assume this was an error. Due to these abnormalities, we filtered out data where the tip amount exceeds fare amount. In summary, these are the conditions we made for filtering the data:

- Fare amount and Tip amount are positive.
- Fare amount and tip amount are non-zero (to remove all the zero tippers)
- Fare amount is greater than tip amount

The correlation value we found between tip amount and fare amount was 0.871 thus proving our hypothesis. (Script: *run_correlation_between_fare_amount_tip_amount.sh*)

While working on this analysis we found that apart from the zero tippers, there were 1,019,285 trips with tip less than a $1. This made up 2.87% of valid trips.

The following is a scatter plot of tip amount vs fare amount:



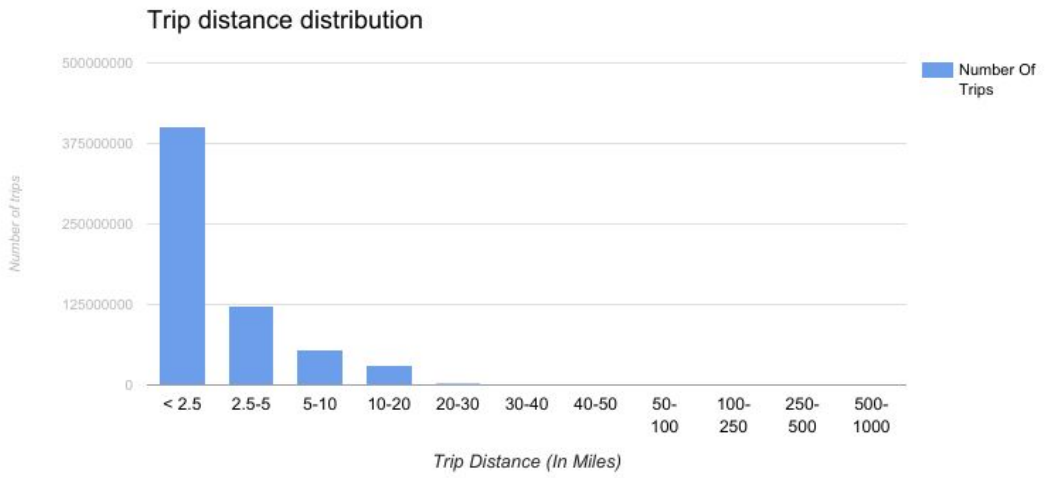
Even after removing records with tip amount greater than fare amount, we still retain records where the trip fare was $8k, and the customer tipped $1600! We imagine someone hired a taxi for a few days.

Hypothesis 9: Most Trips are Short

Seeing as New York has an extensive subway system, and taxis are expensive, we imagine most trips would be over short distances. To confirm this hypothesis, we looked at the distribution of

trip distances.

Before plotting the histogram, we filtered the data to be valid by dictating it must be non-zero, positive, and within 4 standard deviations of the mean.



As seen the plot that maximum trips were short distances. We found that nearly 86% of the trips were between 0-5 miles. Hence, proving our hypothesis (Script : *run_generate_trip_distance_histogram.sh*).

Hypothesis 10: Tip correlated with Trip Distance

Knowing already that tips are correlated with fare amount, it was only natural for us to hypothesize that tip amounts must be correlated with trip distance.

Before looking into this relation we filtered the data using the following declartions,

- Trip distance and Tip amount must be positive
- Tip amount is less than Fare amount (which is also positive)
- Trip distance value is not an outlier and tip amount is also not an outlier
- Restricting trip distances to 1000 miles.

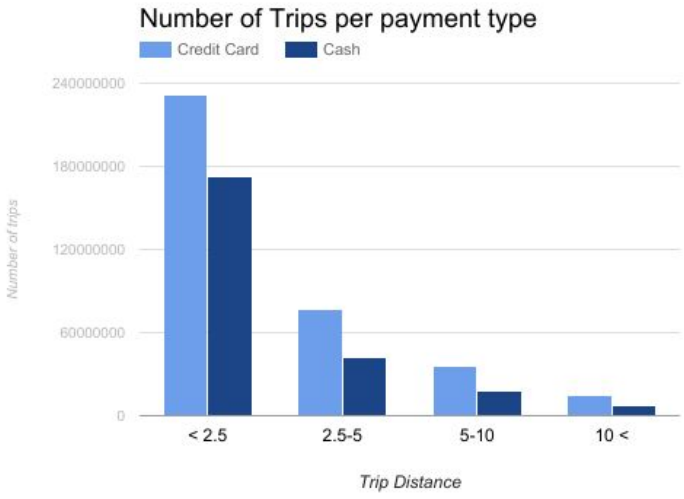
We found that the correlation value of tip amount and trip distance is 0.824 which is expected and this proves our hypothesis. It is slightly lower than the correlation value between Fare amount and tip amount. (Script: *run_correlation_between_trip_distance_tip_amount.sh*)

Hypothesis 11: Fare Amount correlated with Payment Type

As we have already stated, there is a decline in cash usage. Therefore, we would assume that with larger fare amount and trip distances, the tendency to pay with credit cards must surely increase.

The following table shows that with increasing trip distance, customers do indeed pay with credit cards more often.

Trip Distance	Percentage of Trips paid by Cash	Percentage of Trips paid by Credit Card
< 2.5	43.1214%	57.692%
2.5-5	34.217%	62.158%
5-10	31.9892%	65.843%



(script: *run_trip_distance_payment_type_relation.sh*)

Hypothesis 12: Daylight Savings Peculiarity

There is one abnormality in the data that we observed wherein, for one day a year there was a large spike in taxi usage at 1AM, and a spike at 2AM for another day. We realized that the first

spike happens in the Fall, and the second in the Spring, leading us to believe it could be a data abnormality due to daylight savings! As it turns out, the days with the unusual activity, are daylight saving days: the first sunday of November and the second sunday of March.

The following table lists the 10 hours with the trips during the chosen day (script: *run_trip_frequency_by_time.sh*):

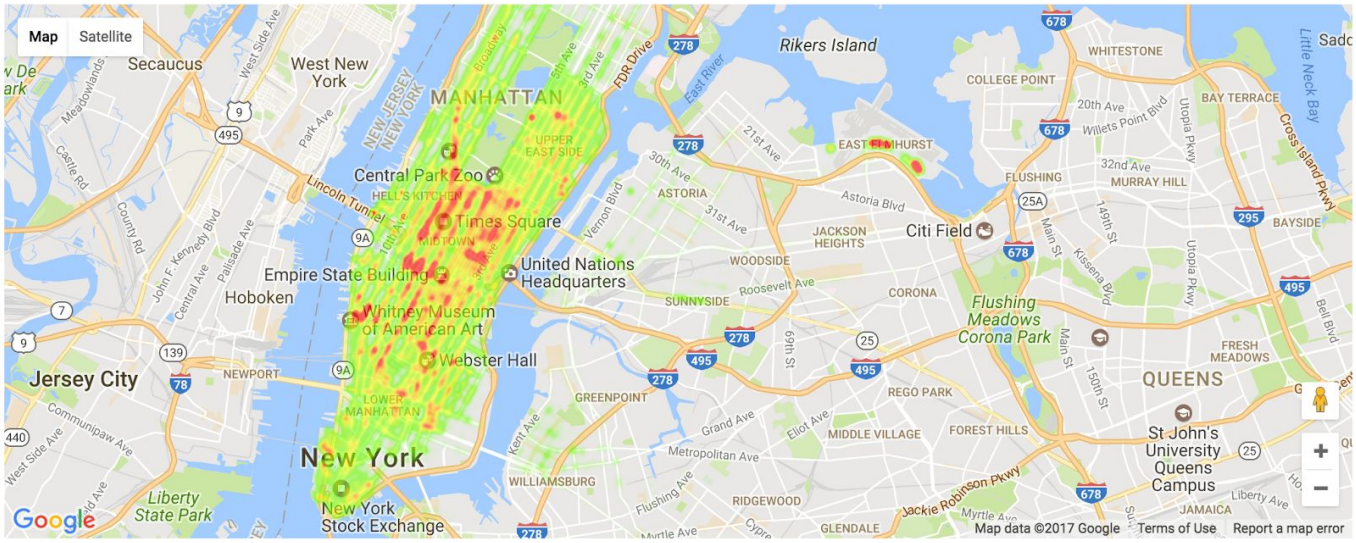
Date	Hour (0-23)	Number of Taxi trips
11/3/2013	1	53078
11/1/2015	1	45853
11/2/2014	1	41894
1/24/2013	19	37103
4/10/2013	19	37090
2/21/2013	19	36811
3/1/2013	19	36725
5/3/2013	19	36572
3/8/2013	19	36569
3/22/2013	19	36528

We see that the taxi trips per hour on the 1st sunday of November are unusually high, confirming our daylight saving hypothesis. Furthermore, in March, the taxi-counts at 2AM are zero, or very small, on daylight savings days.

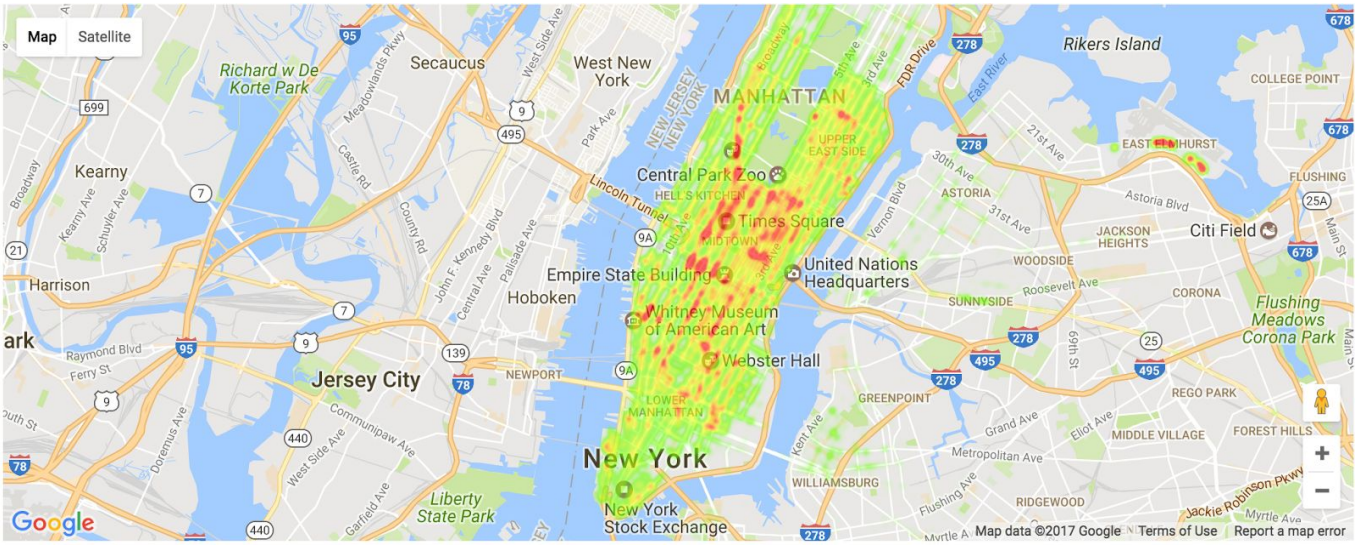
Date	Hour	Count
2013-03-10	2	0
2014-03-09	2	0
2015-03-08	2	157
2016-03-13	2	35

Hypothesis 13: Most Visited by Month

We wondered whether different parts of New York have seasonal popularity. We wanted to see what the trip distribution through the city looks like from month to month. To do so, we rendered heat maps on New York showing the frequency of taxi pickup locations.



Data from month of August



Data from month of January

The distributions looks rather similar but in the winter months, the taxi pickups are more spread

out through the city. (script: month_coordinates.py)

DISCUSSION and CONCLUSION

In our analysis we found that taxi usage is declining. This is likely due to an increase in car-sharing apps, and the popularity of Citibikes might also be taking away from taxi rides during the summer months. In spite this, taxis are clearly still a common form of transport, and one people often resort to in the evening hours when they simply want to get home and don't want to deal with sporadic public transit. We also found that theater life in New York is lively and kicking, with a higher frequency of taxi rides in the theater district.

Therefore, we seen that the taxi data provides some insight in the lives of New York residents and tourists, revealing some of their daily habits and routines.

ISSUES FACED

The issues faced while working on Part II are listed below:

- Weather data in general had no issues. We didn't have to clean the data. But as per the site we can see that there were many missing values. Missing values of precipitation, snow depth and wind. We replaced the missing values with 0 the reason being that most stations just report missing value if it had not rained that day.. We did not find any missing values with temperature.
- Citibiki data schema changed constantly. They had double quotes in the data, had to strip them to remove noise. There were three different formats of dates and we had to make them uniform in order to have the format so that we could join with our pickup dates in taxi data. Wanted pickup dates to match our taxi data to do a join merge (join function in RDD). Lot of time was spent on it we had sure everything matches.
- Our data had schema change in 2016
- There was no uber data available on the Open data site. After Googling we found some data on github. There were few issues around it
    - We didn't know if we could trust it.
    - It was discontinuous and also there wasn't much data. Only through june of 2014.
    - In order to get the data we had to submit the foil request which didn't come through in time.
- We found there were nearly (400k trips) where tip amount is greater than fare amount.

Technically, it is possible but very rare. Since it skewed our data and did not get the correlations that met our hypothesis. Thus, we filtered them out.

- There are few fare amount that were $8000 and $1600k tip which is 20% of the fare. This sounds fine and it might have been that the person hired the cab for days. This is why we didn't do outlier detection claiming that a cab ride can't be more than 3-4 hours.
- We did not get the correlation we expected in winter months. It would have been better to to separate the coldest days from the rest and then run correlation again.
- The last 6 months of 2016 had neighbourhood information instead of coordinate information. We first tried to get neighborhood information from the coordinates by using shapefile regions. To do this we wanted to find the coordinates that correspond to their respective neighbourhood boundaries. However, the task of detecting if a point is in a polygon is slow. Instead of doing the coordinate to neighbourhood conversion, we resorted to assigning mean value of coordinates to the neighbourhood data.