Big Data Analytics Final Project Report

New York University

Tandon School of Engineering

Group Members

Fenil Tailor (N18730085)

Sahil Shah (N12706992)

Ajinkya Avinash Shukla (N17644394)

GitHub URL of Project

https://github.com/sahil2211/BigDataproject

Project Proposal

With the introduction of green taxis, Ubers and CitiBike, an interesting question is how these new modes of transportation have impacted urban movement in New York City.

We attempt to answer the following questions

1. Are previously underserved regions (e.g., Harlem) better served now by the green taxis?
2. Has Uber negatively affected the number of trips or is uber attracting different set of customers entirely?
3. Is there a particular hour of the day where each of these services thrive?
4. Does the weather of the new york city affect the total revenue and the total number of trips per day of the green and yellow taxi and uber?
5. Does public holidays has any effect on the total revenue per day of the yellow taxi ?

We will use the taxi geo data to track the regions where nightlife of New York is active.
We will also try to find out how does uber perform in extreme weather and does the surge pricing play a negative role in its performance.

Index

1. Introduction

We are going to analyze the following questions with the help of this project.

1. How does Uber and green taxi arrival impact the Yellow Taxi in New York?
    a. Progression of rise of uber and green taxi in all 5 boroughs of New York City over the period of April 2014 to June 2015.
    b. Is there a particular area which is better served with the arrival of Uber and green taxi?
    c. Do people prefer yellow taxi, green taxi and uber during a particular time of the day? Is there a clear favorite in any of the borough on different hours during the day.

2. a. How is the total revenue per day of the yellow and green taxi affected by the weather in NYC
   b. How is the total number of trips per day of yellow and green taxi and uber affected by the weather in NYC?
   c. How is the total number of trips per day of yellow and green taxi affected on the public holidays ?

3. Which are popular nightlife locations in New York city. At what location do people of new york typically hangout between 9 pm and 2 am in both weekends and weekdays.

   We are going to use Hadoop Streaming to make sense of the massive datasets into a csv containing useful few rows .
   We carried out our tasks in AWS EMR cluster.

4. Are there any points in the city where citi bike takes as much time as yellow taxi? How many such locations are there?
   To visualize the data we have used R and QGIS for spatial data.

2. Understanding the Data

Dataset Links

Taxi data (yellow and green)
- http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Uber Data
- https://github.com/fivethirtyeight/uber-tlc-foil-response

Weather Data
- http://www.ncdc.noaa.gov/qclcd/QCLCD?prior=N

Timeline

Date	Deliverables
04/19/2016	Cleaning and transforming the Data
04/24/2016	Question 1 analysis using Hadoop(Map+Reduce)
04/29/2016	Question 2,3 analysis using Hadoop(Map+Reduce)
05/04/2016	Question 4 and Visualization
05/09/2016	Data Visualization,Final report and Presentation

Data Issues

i) For Uber We only have the data between April 2014 to September 2014 and January 2015 to June 2015. We only have access to pickup date and pickup location.
ii) We have some dirty geo data. Some pickup coordinates are absent.
iii) We don't have longitude latitudes for the 2015 uber data. We have location ids instead which are mapped in a separate lookup table with their boroughs and zone.
iv) The weather data consist of missing values for weather parameters.

V) The yellow and green taxi data

We employed the following methods to clean the data.

1)Cleaning the data

If there are any null values or outliers there,we first try to analyze and clean the data using of various techniques(ignoring null values, replacing it with approximate values) For some coordinates in spatial data we were getting boroughs as 'Unknown' so we neglected those coordinates.

2)Data transformation

Transform the PickUpDate, PickUpTime,DropOffDate,DropOffTime
into DateTime EST format so that we'll be able to take that format for analysis.

3)**Data Reduction**

All the data attributes which we think will not helpful to our analysis ,we will eliminate giving way only to the useful data for further analysis.

3. Issues Encountered

1. In uber data the date Format is “mm/dd/yyyy” and the date format in Taxi data is in “yyyy-mm-dd”. This can be resolved in our respective mapper code by using strftime method in Python.
2. For the year 2015 data we are not provided with the latitude and longitude of pickup location. We are only provided with the location ids. We are given corresponding zones and boroughs for that locationid.
3. In the Yellow and Green Taxi data we have some Outliers that lie outside NYC some are 0,0 coordinates .We are trying to deal with that tuple either by removing it from our consideration or by just putting approximate value by seeing the PickUp Location and DropOff Location.
4. One major issue we encountered for question 1 and question3 was how do we determine the borough from a given latitude and longitude. We planned to use geopy module which is python’s geocoding modules. But we decided against that as it took a lot of time to execute. We decided to use rtree indexing on zillowshapefile to identify the boroughs. This approach reduced the execution and query time significantly. Although we did use geopy module for generating corresponding latitude longitude pair from the lookup table.
5. In the shapefile record the LGA and JFK zone was missing so we used polygon contains method of the spatial lab to solve that problem. We verified the boundary given in the lab and it turned out to be correct.
6. In the weather dataset for question 2, we have considered the missing values as M or blank as it is in the output because there is no operation to be performed over the values in the reducer.

4. Algorithms

We have partitioned our Project into 3 questions as listed in our introduction. Here we briefly describe how we used MapReduce to achieve our goals.

Question 1

Task-1a

We first import datetime module of python. With the help of date time we extract the year and month from the datetime string.

Mapper

For every record in uber,green taxi and yellow taxi file we take our key as Pickup month, Borough,tag. The tag can either be 'y', 'g' or 'u' for yellow taxi, green taxi or uber respectively. And our value will be 1. We employed rtree indexing on the shapefile to get respective borough by indexing pickup geolocation to RTree of New York shapefile. We queried the index with the help of the pickup longitude and latitude.

Reducer

We simply counted the amount of keys in our reducer side. So our output will be pickup_month,borough,tag,count. The tag identifies whether it is yellow or green taxi or uber and count is the amount of trips for that particular taxi in a particular borough for a particular month.

Task-1b

For this we faced a couple of issues.

We wanted to see the distribution of uber, green taxi and yellow taxi on NYC map. But the precision level of longitudes and latitudes varied across the 3 datasets. We decided to use the precision level of upto 3 decimal places across all the datasets. This allowed us to approximate some coordinates.

The other issue that we faced was for the uber Jan-June 2015 data we were not given the pickup longitude and latitudes. We were given location ids and a lookup table to find out the zone and boroughs of respective location id. We decided to use geopy module of python to generate the longitude latitude pair from zone and boroughs. This is ill advised if the data set is too large but our lookup table had only 265 rows so we could complete finding longitude latitude pair in a reasonable amount of time. We generated 2 hash tables from lookup table one to determine the borough from the location id and other to determine the longitude latitude pair from location id. We stored the hash tables in our mapper file.

Mapper

For each record we found the borough from rtree indexing of shapefile. We extracted the pickup month from pickup_datetime field. For every record we output pickup_month,borough,pickup_longitude,Pickup_latitude,tag as key and 1 as value.

Reducer

We simply add the values of each keys. As a result we get the count of pickups from a particular location for each month.

Task-1c

This was pretty much same as ques1a but instead of extracting the month we extracted the hour from the datetime. We also used the datetime module to determine if a given date is a weekday or a weekend.

For **mapper** we emit (pickup_hour,Borough,tag,'weekday' or 'weekend') as key and 1 as value

For **reducer** we count the keys.

Question 2

Task-2a (Impact of weather on total revenue per day)

Mapper

The mapper emits the key as day and values are total amount for yellow and green taxi and weather parameters namely max temperature, minimum temperature, average temperature, weather type, depth, water level, snowfall, precipitation and average speed.

Reducer

The reducer sums up the total amount for yellow and the green taxi and emits key as the day and values as the total revenue per day of the yellow and the green taxi along with the weather parameters for that day.

Task-2b (Impact of weather on total total number of trips per day)

Mapper

The mapper emits the key as day and values are trip count as1 for yellow and green taxi and uber as well as the weather parameters namely max temperature, minimum temperature, average temperature, weather type, depth, water level, snowfall, precipitation and average speed.

Reducer

The reducer sums up the trips for each day and emits key as day and value as the total trip count per day for the yellow and the green taxi along with the weather parameters of the day.

Task-2c (Total revenue per month)

Mapper

The mapper emits key as the year and month and value as the total amount for the green and the yellow taxi.

Reducer

The reducer sums up the total revenue for each month and emits key as the year-month and values as the total revenue for the green and the yellow taxi.

Question 3

Mapper

We like the previous question determine the borough of the long lat pair and filter the the time that fall between 9 pm to 2 am for both weekdays and weekends. We round down the long lat pair to 3 decimal places. This was verified in Google maps that is rounding down to 3 decimal places doesn't lead to major loss in accuracy.

Reducer

We simply count the keys in the reducer and output the csv file.

Question 4

Mapper

We determined the pickup and dropoff neighborhoods within boroughs using the shape file and took the pickup neighborhood and dropoff neighborhood as key. The value was the citibike or yellow taxi tag and the trip duration

Reducer

We take the average of the trip duration and output them in a csv.

5. AWS Implementation

We performed Hadoop streaming in AWS EMR cluster. We used 1 Master and 4 core nodes for ques 1a and ques 1c and 1 master and 7 core nodes for ques 1b and ques 3, 1 Master and 2 core nodes for question 2.

We needed to install rtree, libspatialindex and pyshp across all nodes. We did it by writing a shell script and installing it across all nodes in the bootstrapping stage.
Example
We will show the implementation commands used for ques 1a here for reference.

* Termination protection: Yes

* Logging: Enabled (remember to input your S3 bucket to store log file)

* Hadoop distribution: Amazon 2.7.2

* Bootstrap action: This is a very important step because the sample scripts

make use of python rtree library, but Amazon AMI 2.7.2 does not have rtree installed.

Click 'Add bootstrap action' -> Custom action -> Configure and add ->

Put the following in 'S3 location': s3://safprojectbigdata/rtree.sh

* Don't add any step at this point

* Cluster Auto-terminate: No

Then we add the neighborhoods files that we use in our mapper to the S3 bucket.

neighborhoods: s3://safprojectbigdata/neighborhoods

Finally to generate output csv file

Replace safprojectbigdatawith your bucket name, except in Input

* Mapper: s3://safprojectbigdata/scripts/ques1a/map.py

* Reducer: s3://safprojectbigdata/scripts/ques1a/reduce.py

* Input: s3://safprojectbigdata/input/ques1/

* Output: s3://safprojectbigdata/ques1aoutputfinal

Arguments: -D mapred.reduce.tasks=1 -files s3://safprojectbigdata/scripts/ques1a/map.py,s3://safprojectbigdata/scripts/ques1a/reduce.py,s3://safprojectbigdata/neighborhoods/ZillowNeighborhoods-NY.shp,s3://safprojectbigdata/neighborhoods/ZillowN

eighborhoods-NY.prj,s3://safprojectbigdata/neighborhoods/ZillowNeighborhoods-NY.shp.xml,s3://safprojectbigdata/neighborhoods/ZillowNeighborhoods-NY.shx,s3://safprojectbigdata/neighborhoods/ZillowNeighborhoods-NY.dbf -files s3://safprojectbigdata/scripts/ques1a/map.py,s3://safprojectbigdata/scripts/ques1a/reduce.py -mapper map.py

Likewise we did for every questions. For question 2 we did not require any external library and thus skipped the bootstrapping shell script stage of rtree.sh. We used 2 reducers for every tasks. However this can be accomplished using any amount of reducers as the order is not important in the output file.

The output and script links will be provided in further sections

Output links

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Night_Life_Analysis/part-00000_yellow_green

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Monthly_Total_Revenue/part-00000

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_hour_Analysis/part-00000_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_hour_Analysis/part-00001_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_hour_Analysis/part-00000_yellow_green

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis/part-00000_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis/part-00001_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis/part-00000_yellow_green

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis_with_long_lat/part-00000_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis_with_long_lat/part-00001_uber

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Per_month_Analysis_with_long_lat/part-00000_yellow_green

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Weather_Impact_On_Total_Revenue/2015/part-00000

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Weather_Impact_On_Total_Trips/2015/part-00000

https://s3-us-west-2.amazonaws.com/fenik/Big_Data_Projects_outputs/Weather_Impact_On_Total_Trips/2014-2015/part-00000

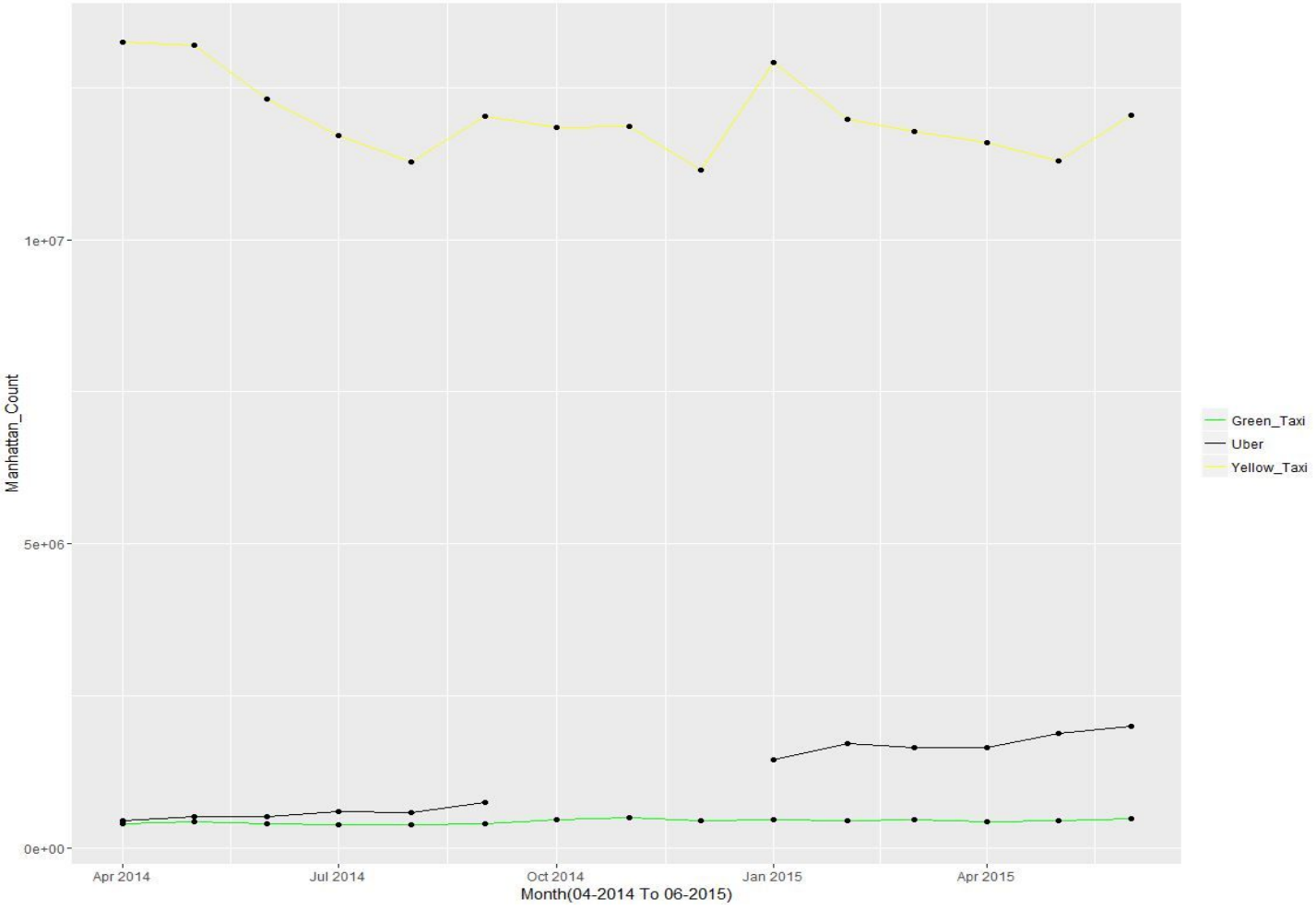
Script links can be found on github repository
The link for github repository is

https://github.com/sahil2211/BigDataproject

6. Visualization

1. Visualization for Trips Per Month Per Borough Analysis

For Manhattan



As we can see from the above graph that a large fraction of the pickups are done by the yellow taxis only.From the counts we can also deduce that Manhattan has the largest number of pickups than any other boroughs.

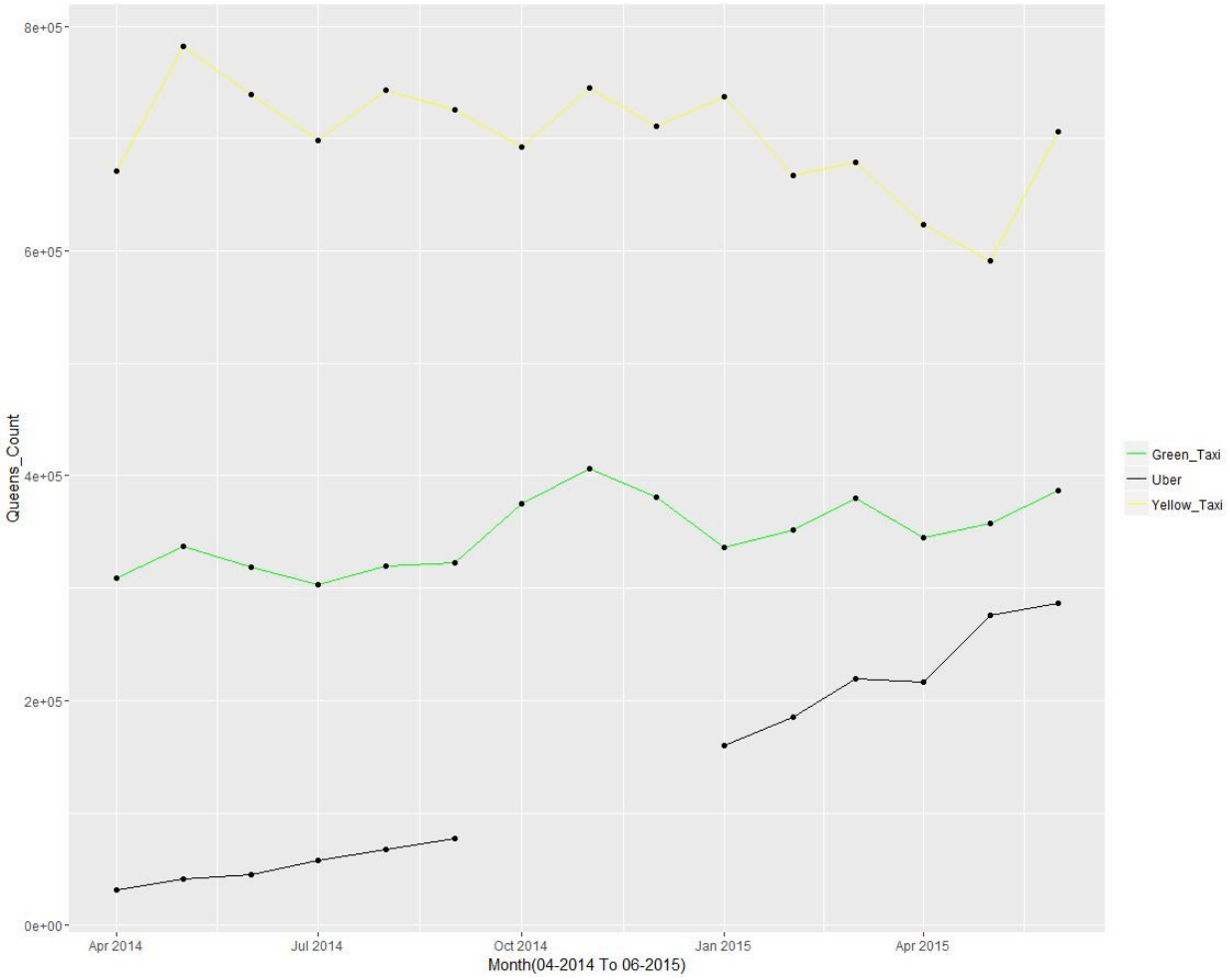
By seeing the graph we can also deduce that Uber trips count overtook the count for Green taxi in all of the 12 months starting from April-2014 to September-2014 and from January-2015 to June-2015

We can say that Uber Trips count are increasing drastically(Approx 500%) in manhattan by looking at the values for Apr 2014 and Apr 2015
Apr 2014 =3.1*10^7(Approx)
Apr-2015 =18*10^7(Approx)

There is also interesting to see that yellow taxi count remains almost constant. So Assuming the revenue correlates heavily with the number of trips we can assume as far as Manhattan is concerned that the yellow taxi business can be expected to hold its own against the threat of uber. This also can lead us to assume that uber is attracting the customers who don't usually travel by yellow taxi. So is uber eating into Subway and citibike revenue?

As far as green taxis are concerned the pickup rate being so low can be explained by the fact that green taxis are not allowed pickups below East 96th Street and West 110th street which pretty much nullifies the entire Manhattan region

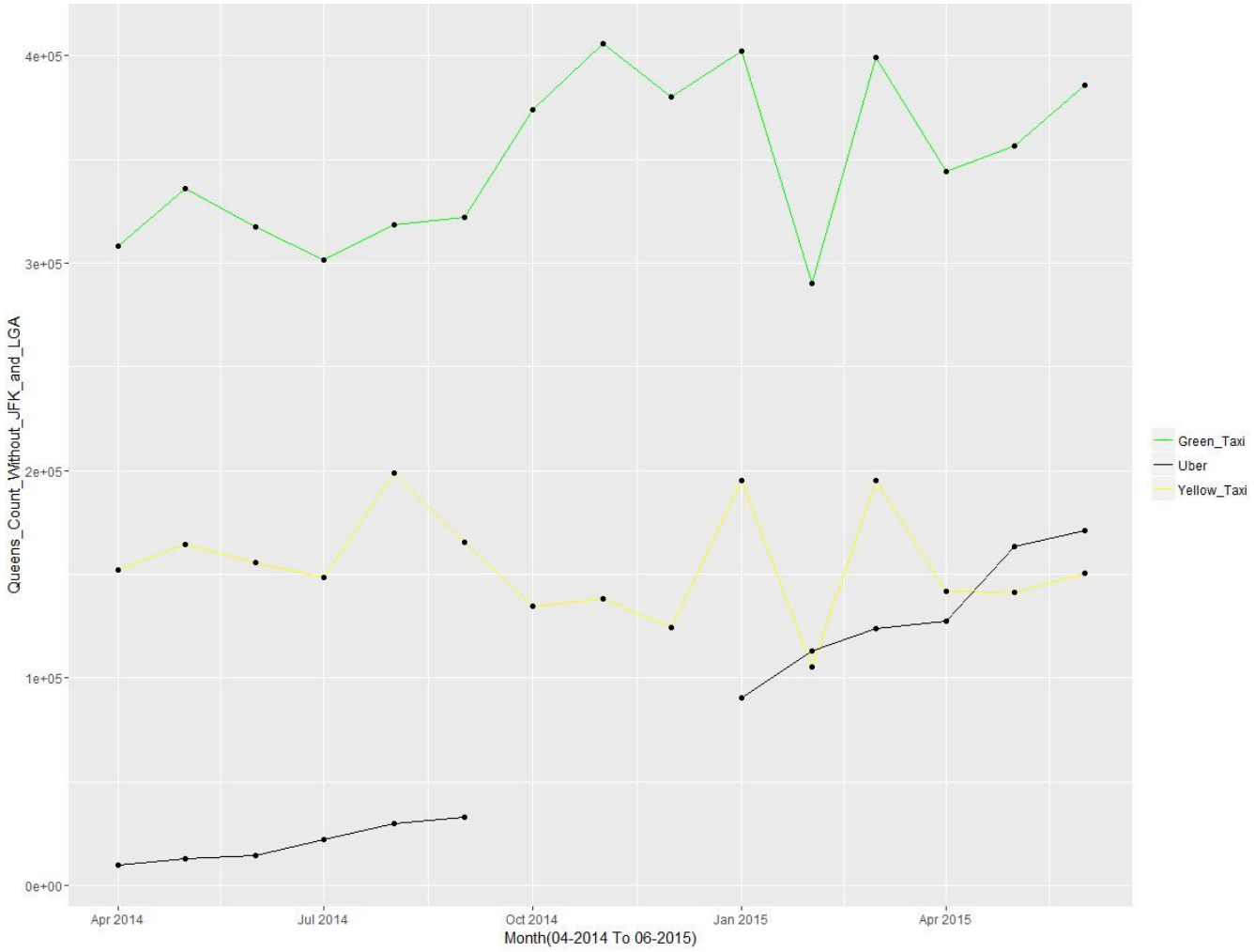
For Queens



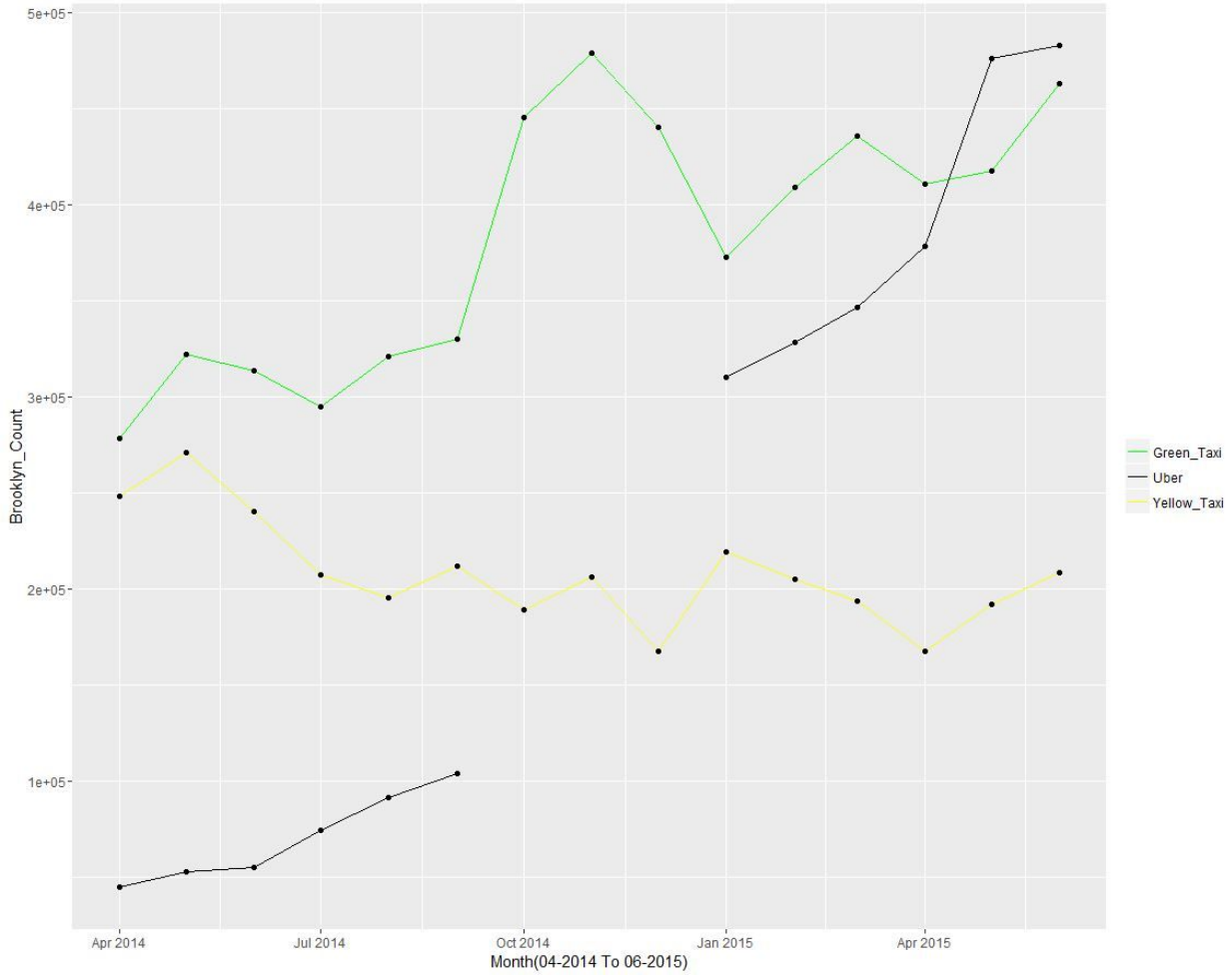
We can conclude from the above graph that Yellow taxi dominates the market in Queens by large fraction of the trip count.

Though Uber Count is increasing,it is still struggling in Queens to cope-up with the Green Taxis count

As long as we restrict our analysis to Green and Yellow taxis,the reason we see green taxis underperforming in Queens is because they are not allowed their pickups in the airports JFK and LGA. This is a huge market. If we exclude the airport then green taxis easily outperforms yellow which can be shown in the below graph.


4e+05
3e+05
2e+05
1e+05
0e+00
Queens_Count_Without_JFK_and_LGA
Apr 2014
Jul 2014
Oct 2014
Jan 2015
Apr 2015
Month(04-2014 To 06-2015)
Green_Taxi
Uber
Yellow_Taxi

For Brooklyn

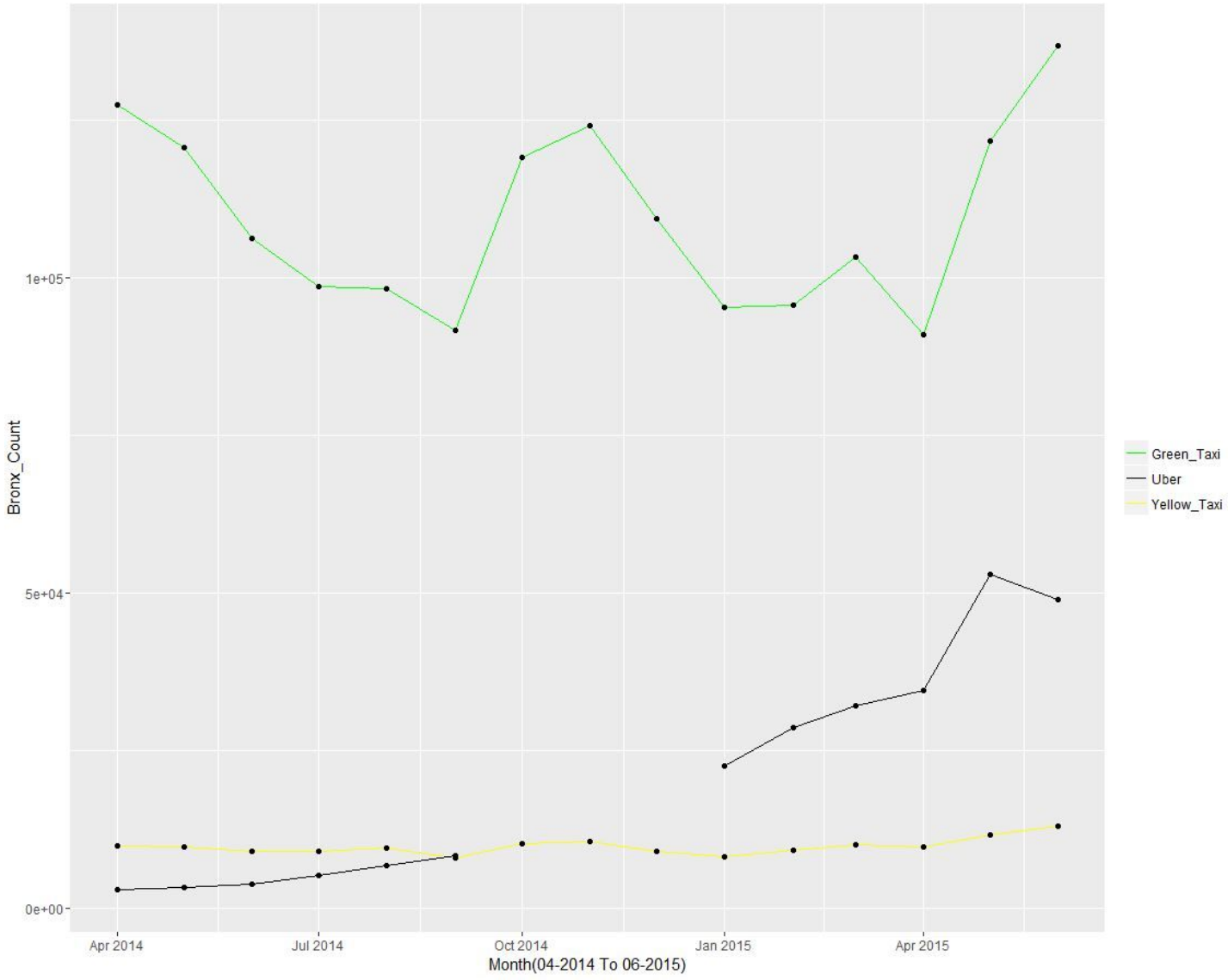


In Brooklyn,till April 2015 green taxis dominates the market but after that Uber just overtook Green Taxi and became a very strong player in this borough.

Yellow Taxi is a average player in Brooklyn which was overtaken by Uber somewhere in between October and January

Yellow taxi is clearly declining in Brooklyn. However the decline is slow whereas the rise of uber is drastic that means uber is cultivating new customers.

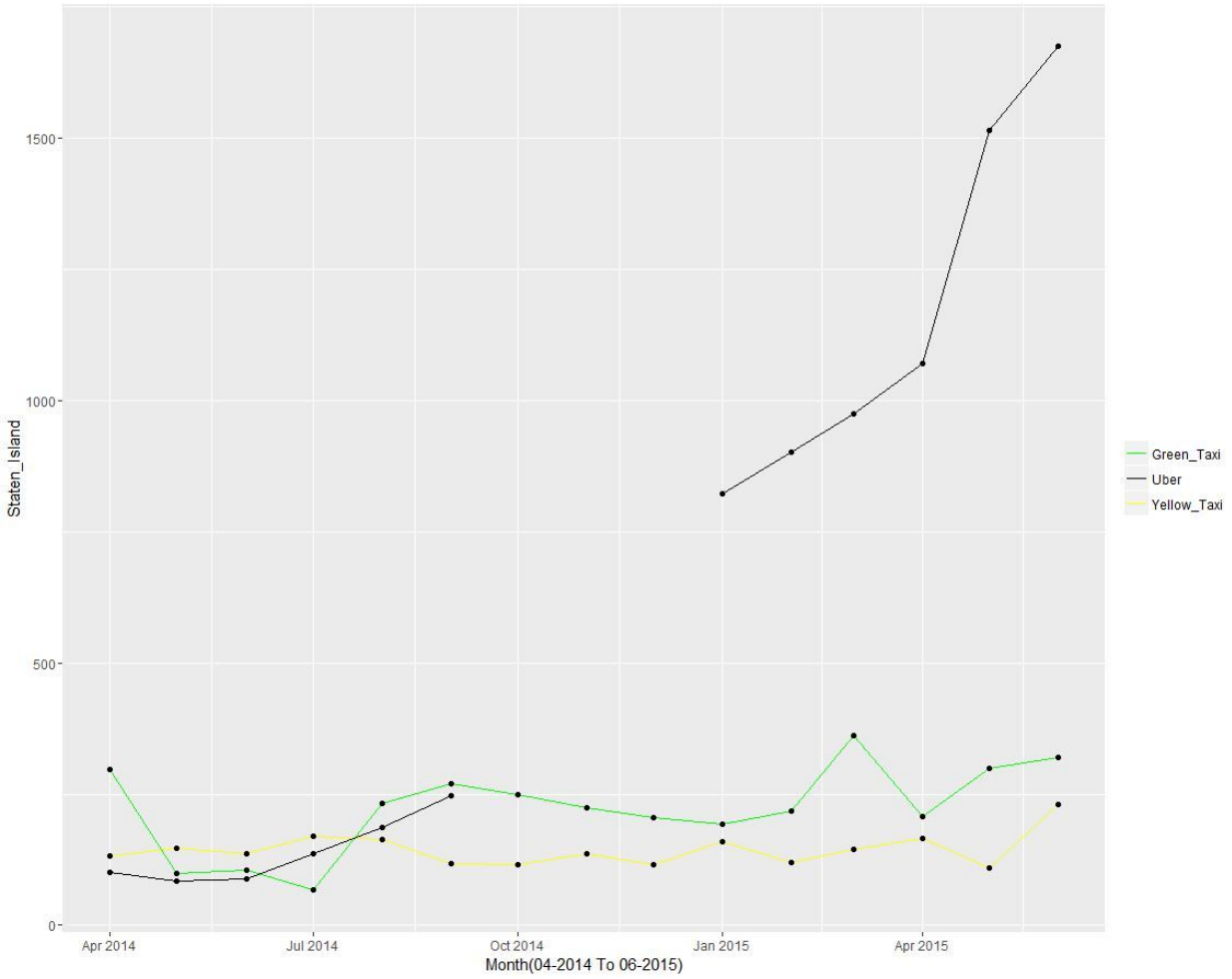
For Bronx



For Bronx,We can say that Green Taxi takes the game by significant margin.As usual like other boroughs Uber shows an upward trend and it took over Yellow Taxi in January 2015.After January 2015 Uber was given higher Preference than Yellow Taxis by people commuting from Bronx.

Thus green taxis are serving their purpose and the sheer number by which they have beaten yellow and uber is because of the fact that the demands of New Yorkers were not being met by the supply of yellow taxi earlier and thus green taxis have infact replaced yellow in Bronx.

For Staten-Island

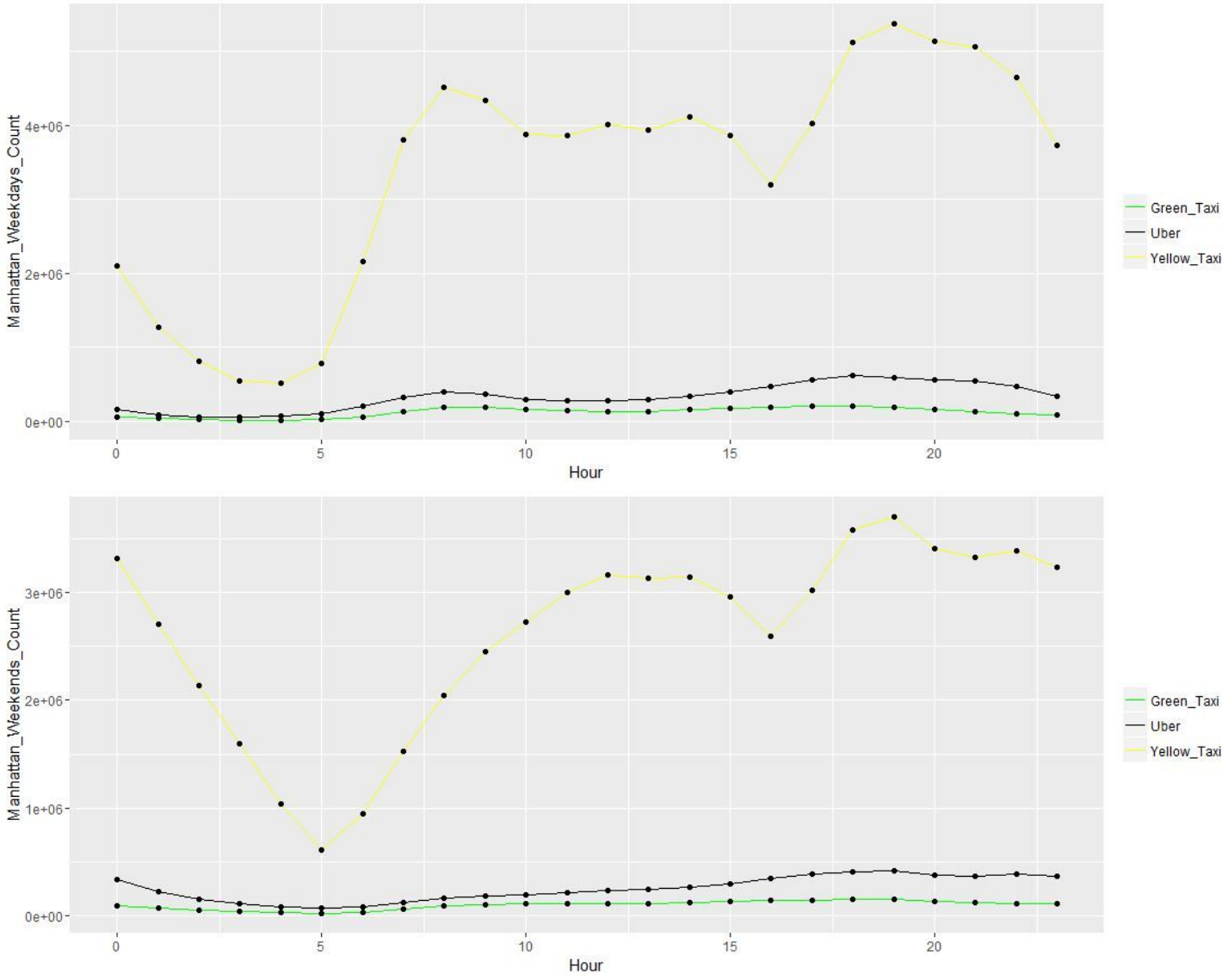


As depicted from the above graph that in Staten Island we have minimum number of pickups than any other boroughs.

The numbers are insufficient to declare uber the winner although the graph says otherwise because the maximum the trips go is around 1700 which is drastically low as compared to other boroughs. So this means the people of Staten Island don't prefer this mode of transportation.

1)Visualization for Trips Per hour Per Borough Analysis

For Manhattan



As we can see from the above graph that number of yellow taxis trips in the weekends between hour 00:00 and 04:00 is greater than the number of yellow taxis trips in the weekdays during these hours.So we can conclude that there is actually high night activitity during weekends than weekdays

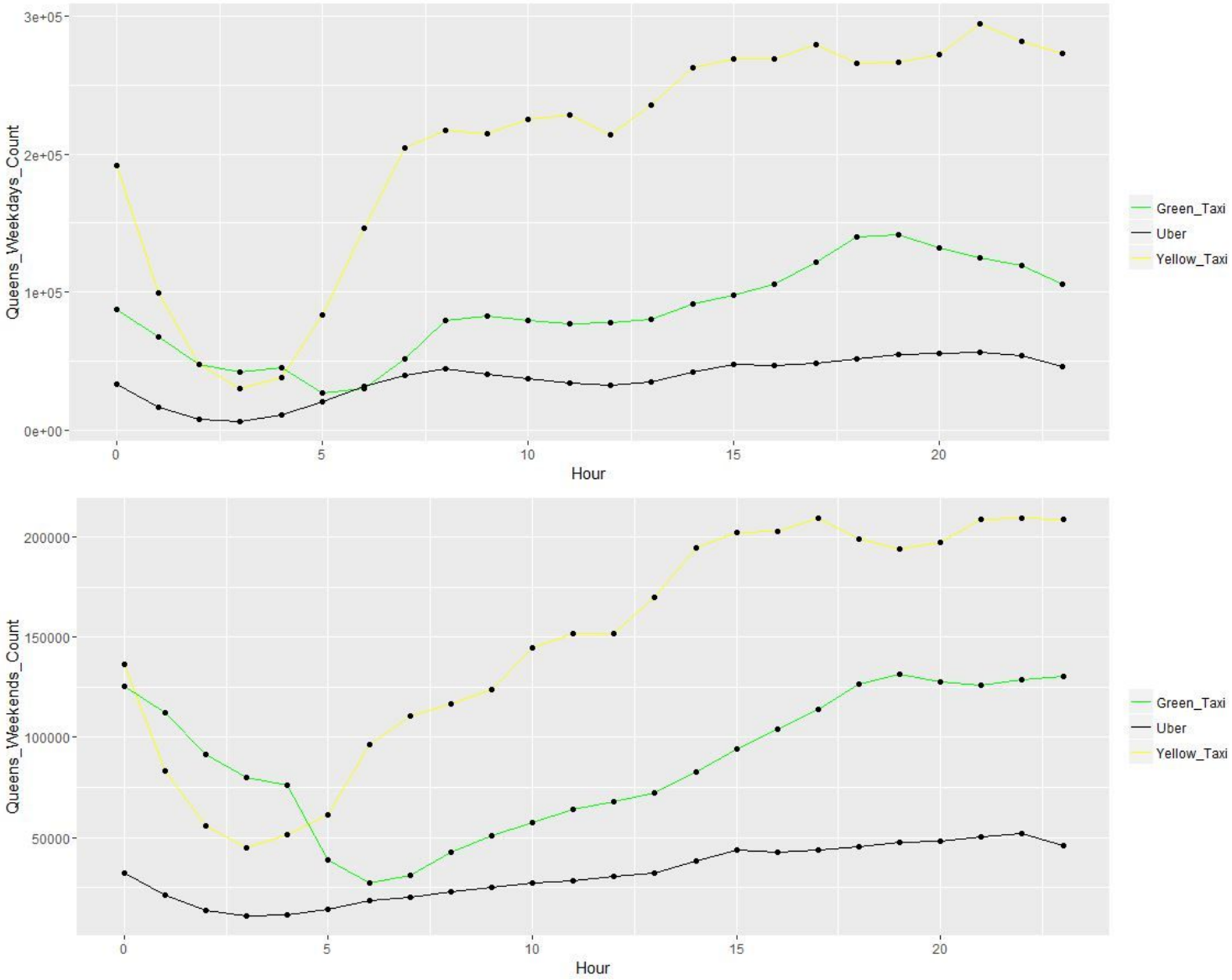
Also we can say that the count of Uber trips overtook the count of Green taxis trips for all hours.In Manhattan, not surprisingly around 90% of the pickups for the entire day are done by the yellow taxis

The trend for both of the weekdays and weekends graph are more or less same except that for weekdays we have the local maximum at Hour 8:00 for the yellow taxis trips but for weekends the graph is steadily increases from Hour 5:00 to Hour 12:00 .

Can we deduce what are the rush hours in manhattan from this analysis?

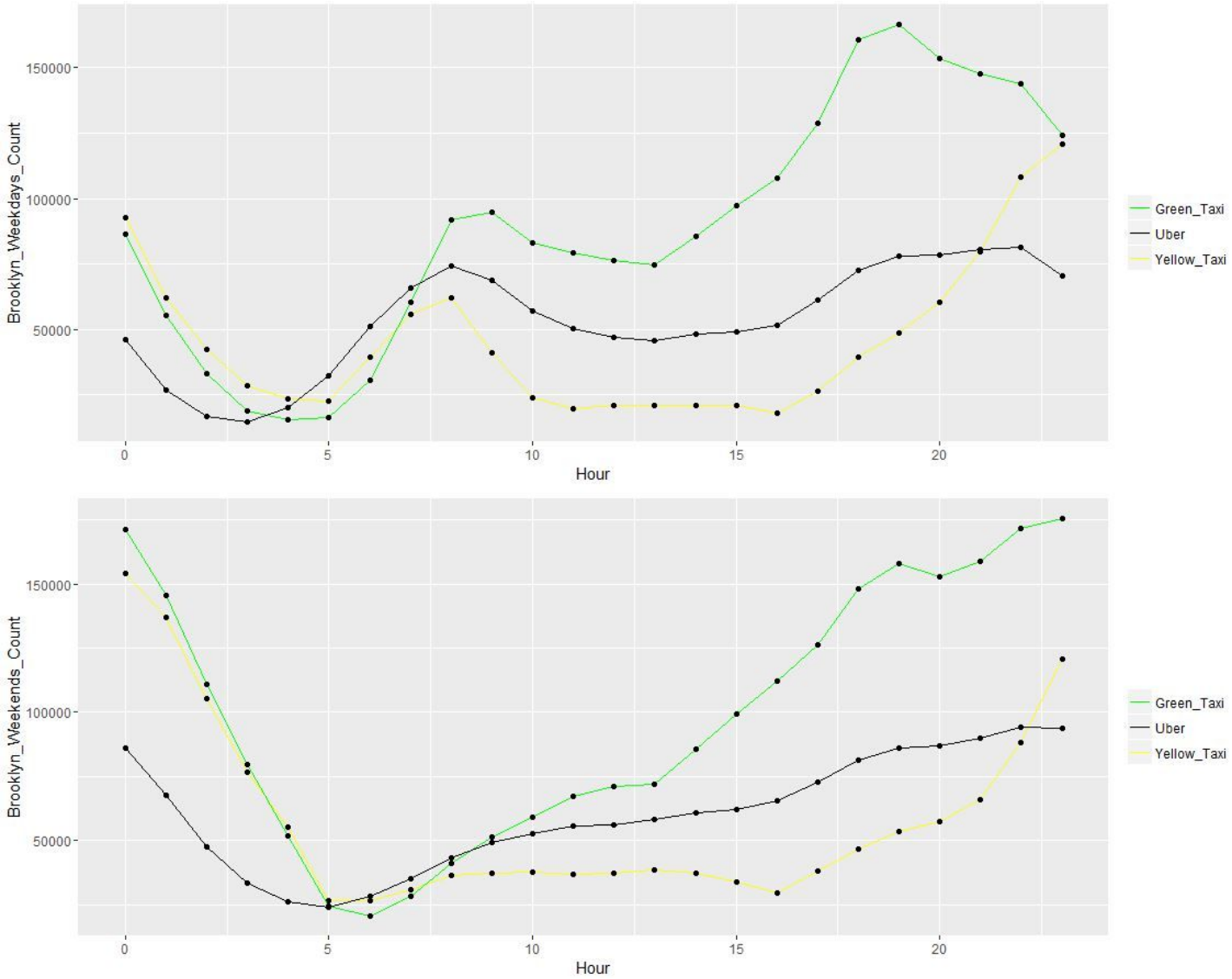
Yes,We can. people living in manhattan generally heading towards their offices from their homes at around 8:00 and 9:00 during weekdays and people leaving for their homes from work generally around at 19:00. These hours are rush hours for manhattan.

For Queens



The Yellow Taxi Queens data is heavily influenced by the LGA and JFK airports for hours in the range 14:00 to 23:00 .

For Brooklyn



We can see that in weekends people of Brooklyn generally prefer green or yellow taxis than the Uber between hour 00:00 and 05:00.After then people prefer to commute via green taxis much more than the Uber and yellow taxis.

Also,There is a local maximum in the weekdays graph at 8:00 for all the three taxi types but weekends graph shows the increasing trend after 6:00 and there is no local maximum in the range between 05:00 to 10:00.Further,there is much more activity in the evening than in the morning for weekdays.

Moreover we can also say that count for Uber overtook the count for Yellow taxis for both weekends and weekdays in hour range between and including 06:00 to 21:00

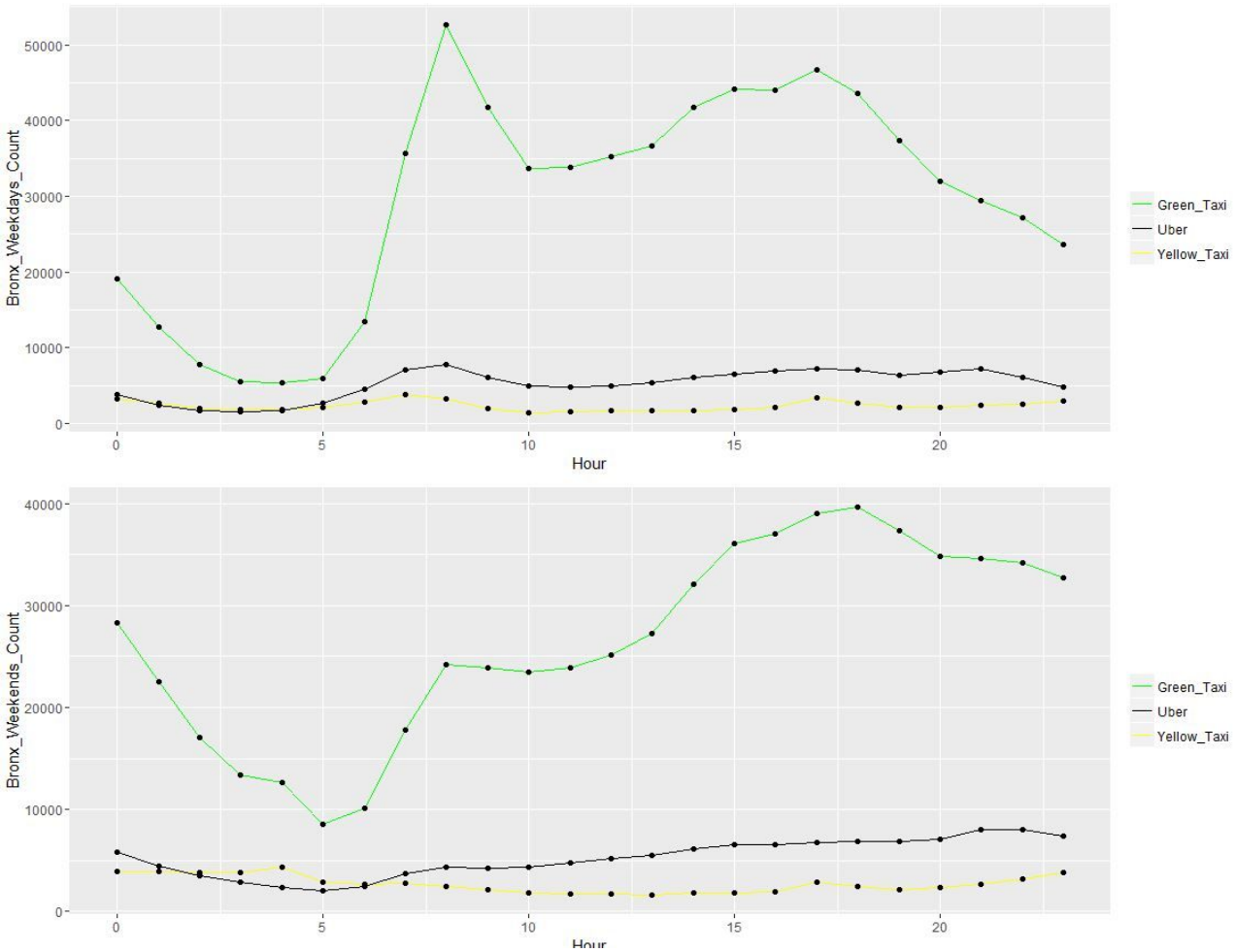
Still for going to late night parties and hangouts people of Brooklyn prefer yellow or green taxis over uber.Why is that happening?

It's just that generally Uber hike their prices for the weekends during late night hours.

So,What time of the weekday you should prefer if you are heading somewhere and it's not urgent?And At What time of the day you can get the Cab very easily?

By looking at the graph,we can say that you should avoid hours 08:00 to 10:00 in the morning and 18:00 to 19:00 in the evening if it's not that urgent. Also among both those peak hours people still favor taxis in evening. That means more people are willing to take subway or walk in the morning than evening.

For Bronx



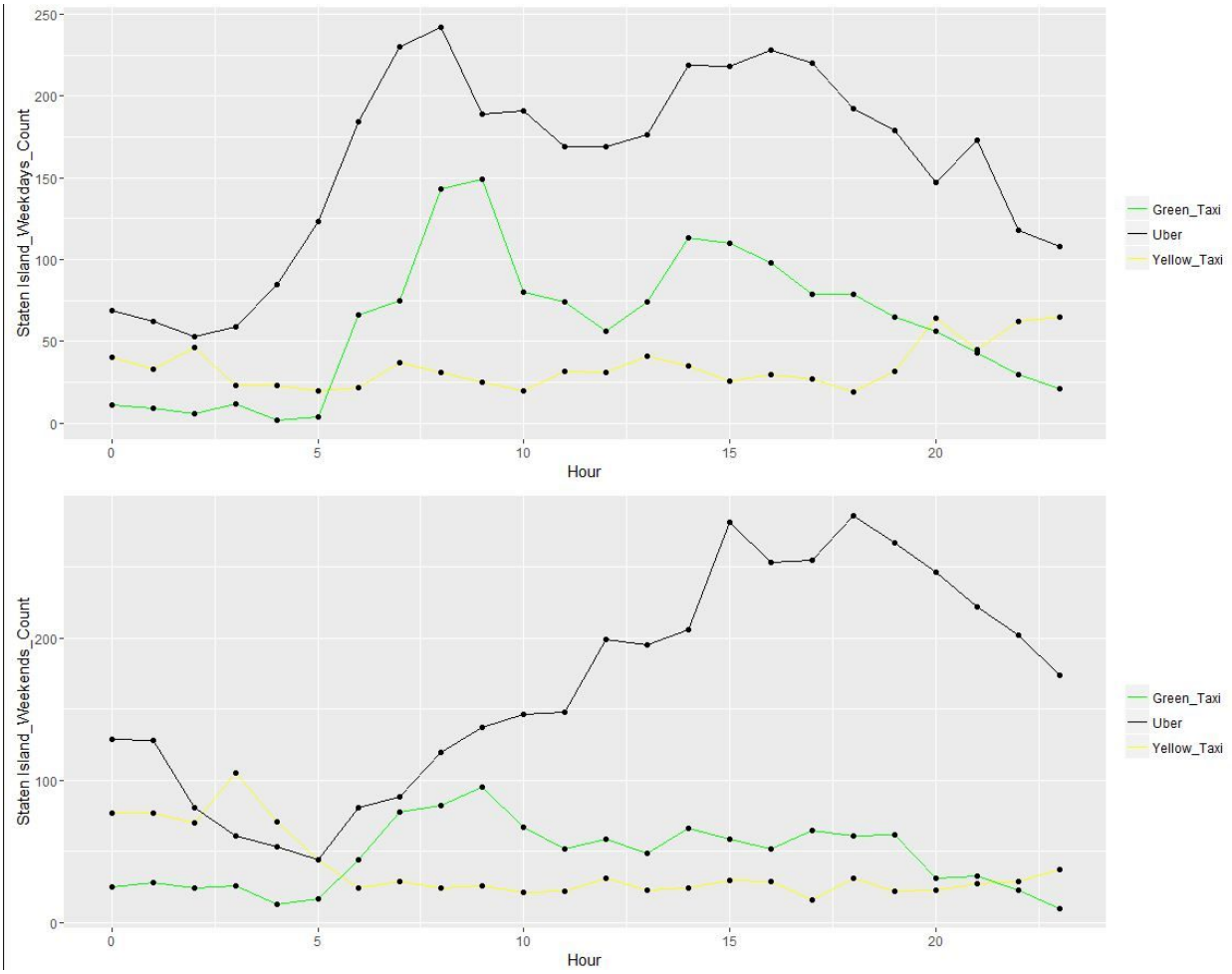
We can see that Green Taxi is dominant throughout the 24 hours.As usual like the other boroughs,the graph for Taxi counts in Bronx has a global maximum at 8:00 which is rush hour.

There is one strange behaviour that we can notice by analyzing the weekends graph.Unlike other boroughs,people living in Bronx prefer to take Uber than the yellow taxi in the hour ranges between 00:00 to 2:00

So,What time of the weekday you should prefer if you are heading somewhere and it's not urgent?And At What time of the day you can get the Cab very easily?What is the case for weekend?

By looking at the graph,we can say that you should avoid hours 08:00 to 10:00 in the morning and 17:00 to 18:00 in the evening if it's not that urgent

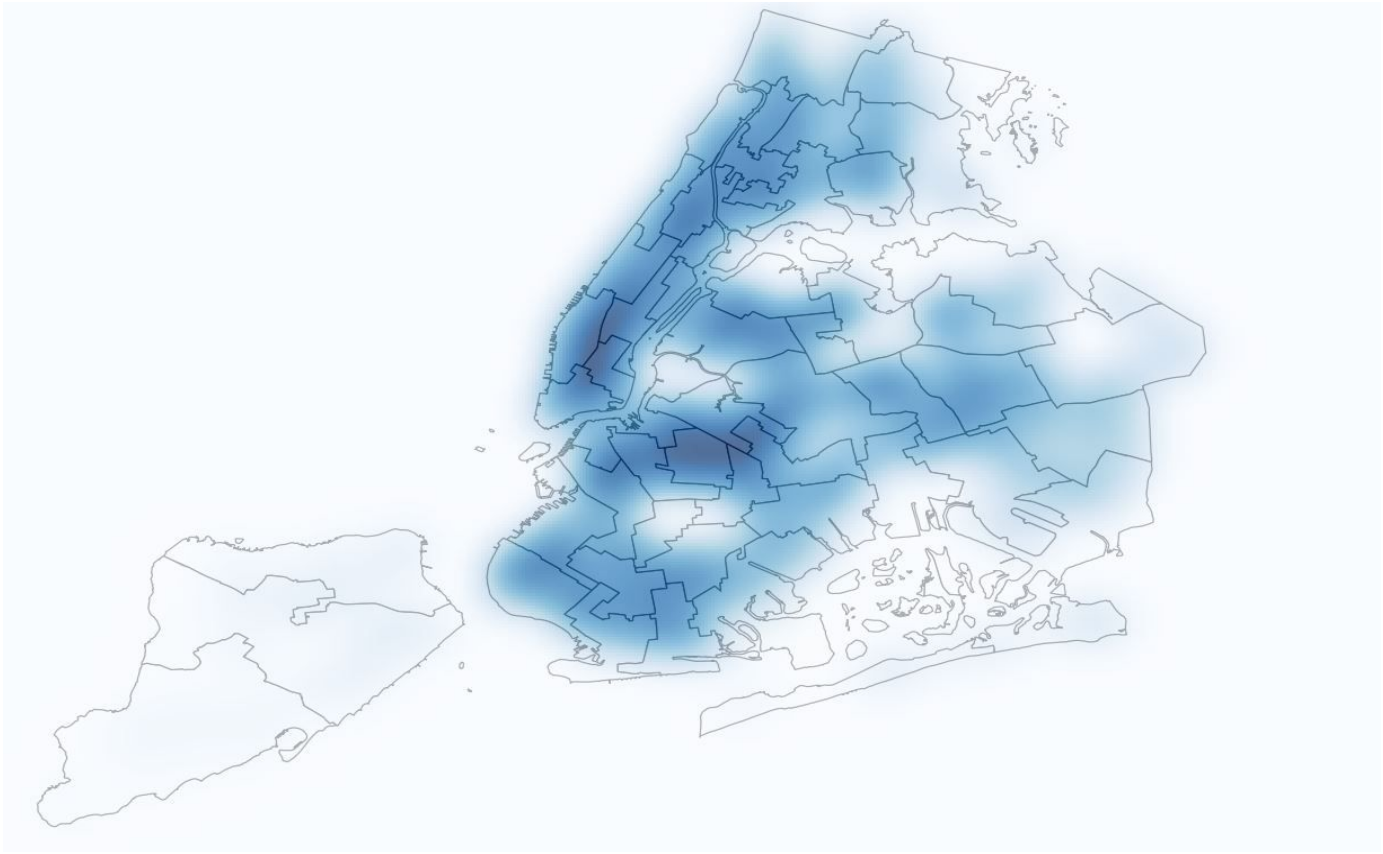
For Staten Island



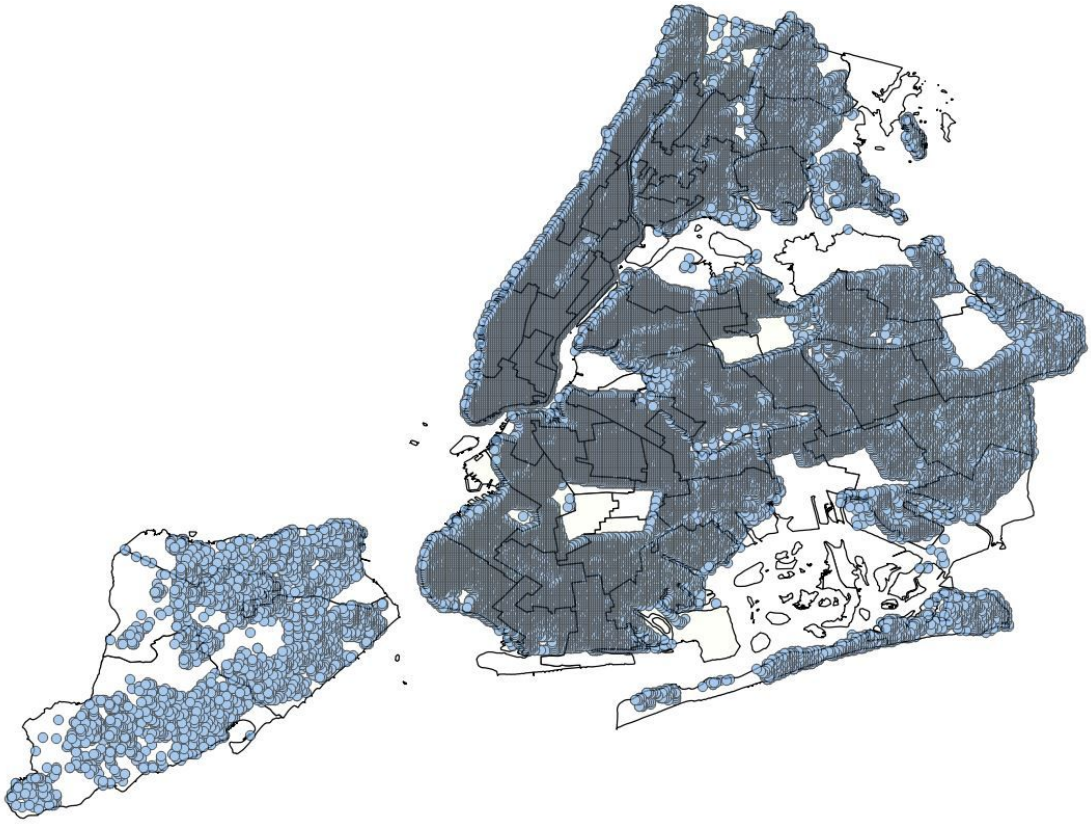
We can see that Uber is dominant throughout the 24 hours.As usual like the other boroughs,the graph for Taxi counts in Staten Island has a global maximum at 8:00 for the weekdays and global maximum at 18:00 for weekends which are rush hour.

Visualization for Night Life Analysis

1)Heat Map



2)Points Plotted into the MAP



For doing Night Life Analysis,We are utilizing the taxi information to draw a few derivations about what parts of the city are well known for going out late around evening time by taking count of pickups that happen between 10 PM and 2 AM.

From Above two graphs we can clearly deduce that Lower East Manhattan and Upper West Brooklyn are the most favourite among the people that loves night life in NYC

The darker shade in Heat map suggests that number of the dropoffs at that location is higher.The Neighbourhoods into which the heat map is dark are listed below from where we can infer their names where there is most lively night life

1) Bedford-Stuyvesant in Brooklyn

2)Williamsburg in Brooklyn

3)Bushwick in Brooklyn

4)Harlem in Manhattan

5)Greenwich Village in Manhattan

6)Lower East Manhattan(Specifically Delancey Street,little italy)

On the other hand if we want to find the place where there is significantly less activity at night, then we can also deduce that from the above two graphs

From the points graph and the heat map we can say that following are some of the regions where there is not such lively night life

1)Long Island City which creates boundary between Brooklyn and Queens

2)Mcguinness Boulevard in Brooklyn

3)Upper West side of Manhattan

4)Lower south of the Brooklyn

5)West side of Queens

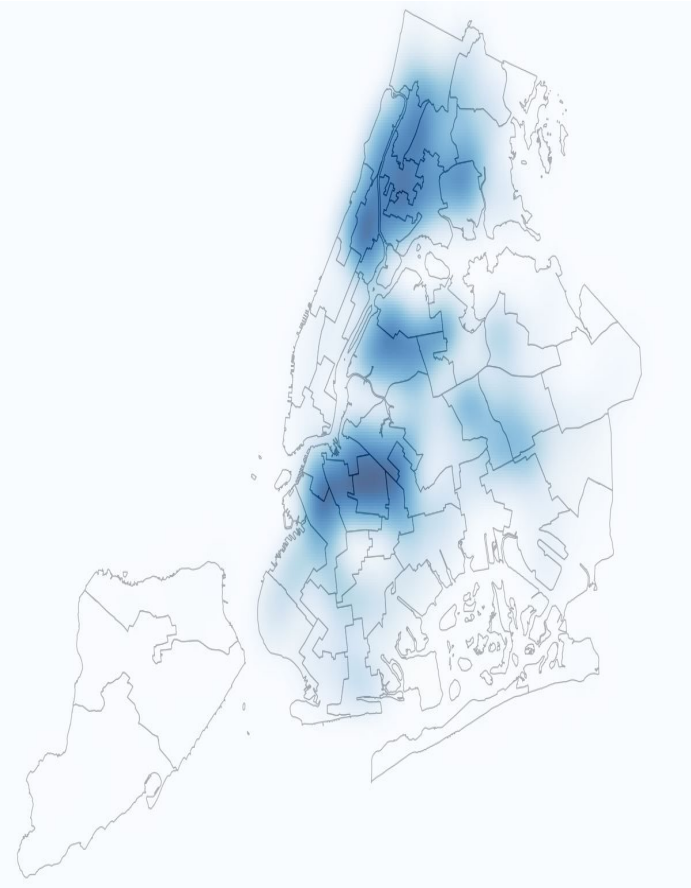
6)Upper north side of Bronx

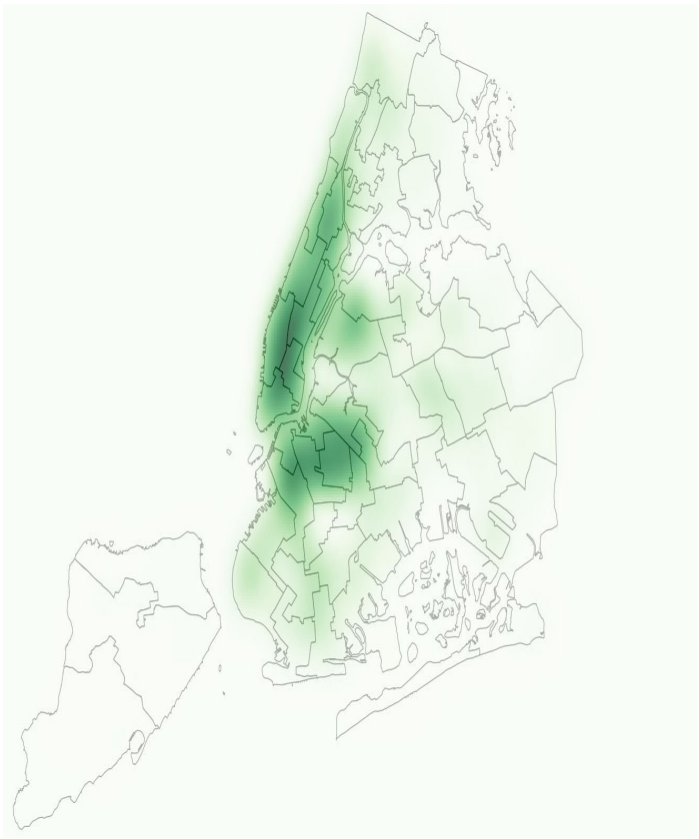
Visualization for Trips Per Borough With Longitude and Latitude

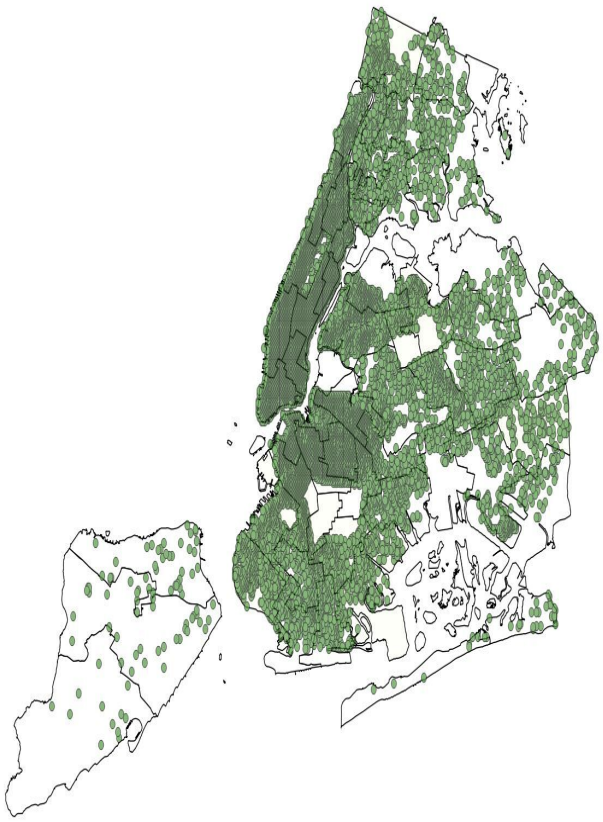
For all of the visualizations below, Blue heatmap is for Green Taxi,Green Heatmap is for Uber and Orange Heatmap is for Yellow Taxi Counts as per their respective Longitude and Latitude.

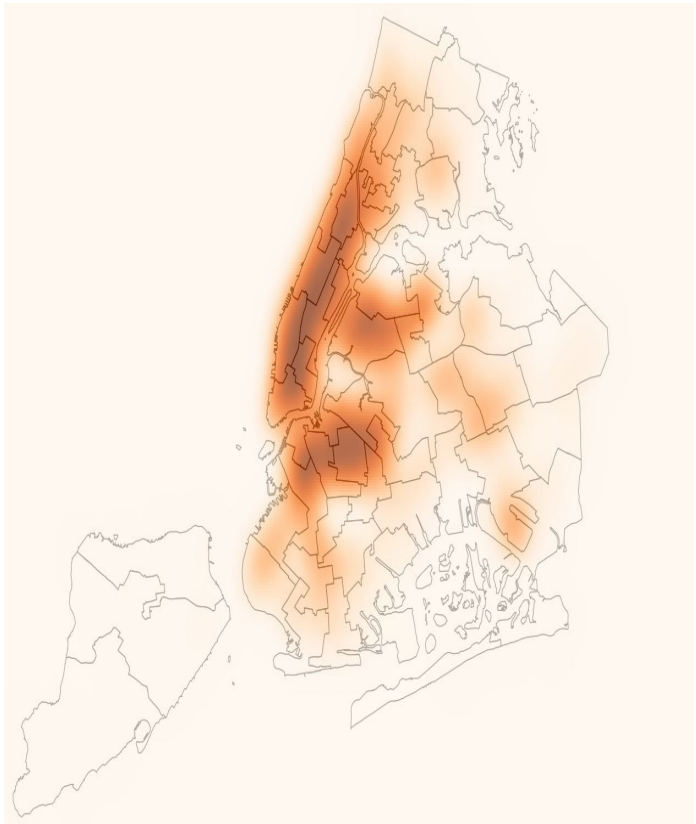
For April 2014 to September 2014

Green Taxi









As we can see from above graphs that Green taxis were able to pick significant number of passengers into east side of Manhattan and in the north east part of the Brooklyn.

Wondered why there is no significant Green taxis trips in the east side of Manhattan?

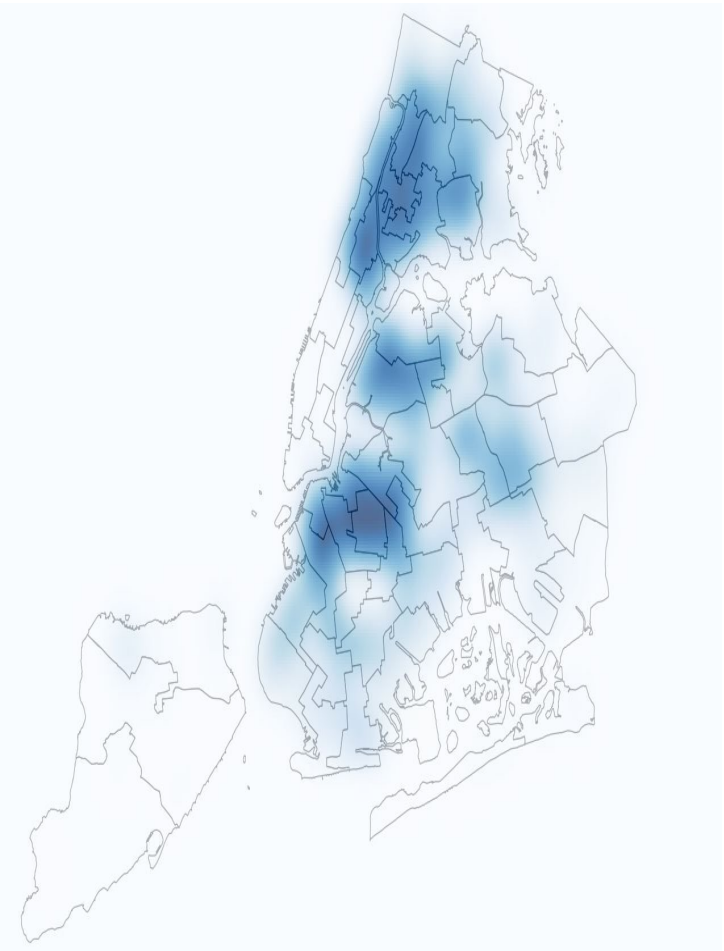
As Green Taxis are not allowed to operate in the east side of the manhattan,Uber and Yellow taxis took the game for east side of the Manhattan

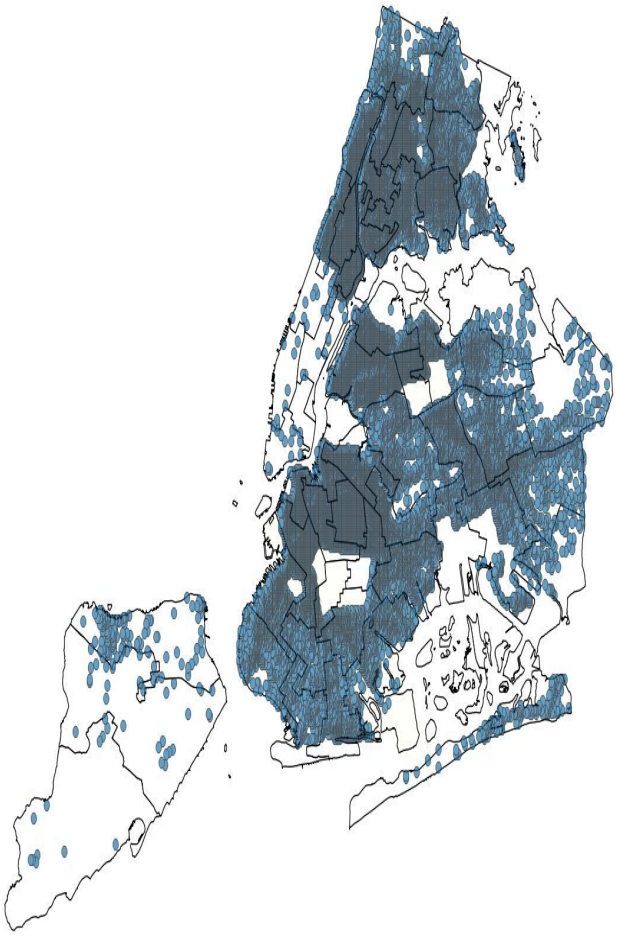
Green Taxi dominant areas are outer west part of the manhattan, east part of the Bronx and Upper north side of the Brooklyn

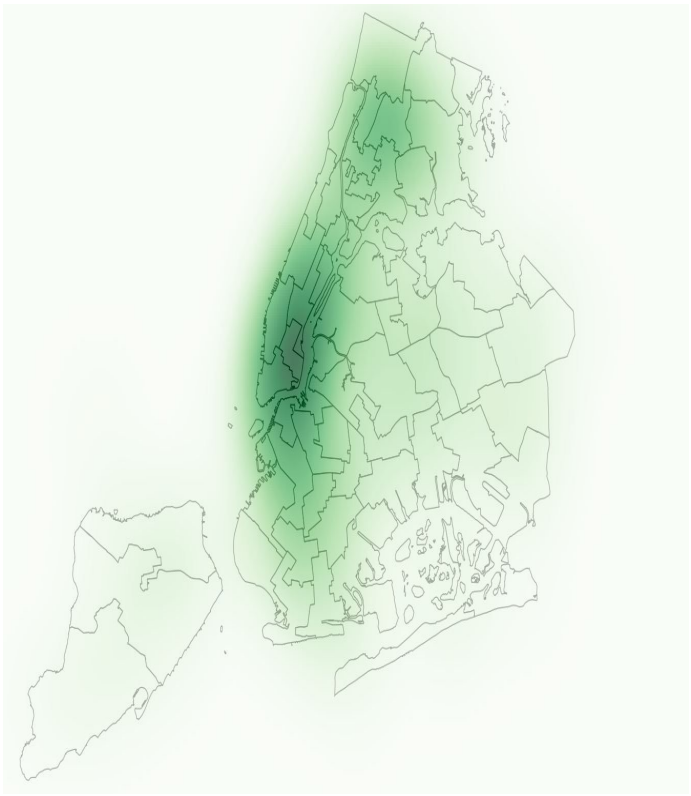
Uber taxis were able to catch countable number of passengers from east side of Manhattan,north east side and some of the west side neighbourhoods of Brooklyn and from neighbourhoods situated in upper side of the Queens.As count for Uber trips was much lesser than the count for Yellow and Green taxis,Uber wasn't dominant in any of the neighbourhoods during April 2014

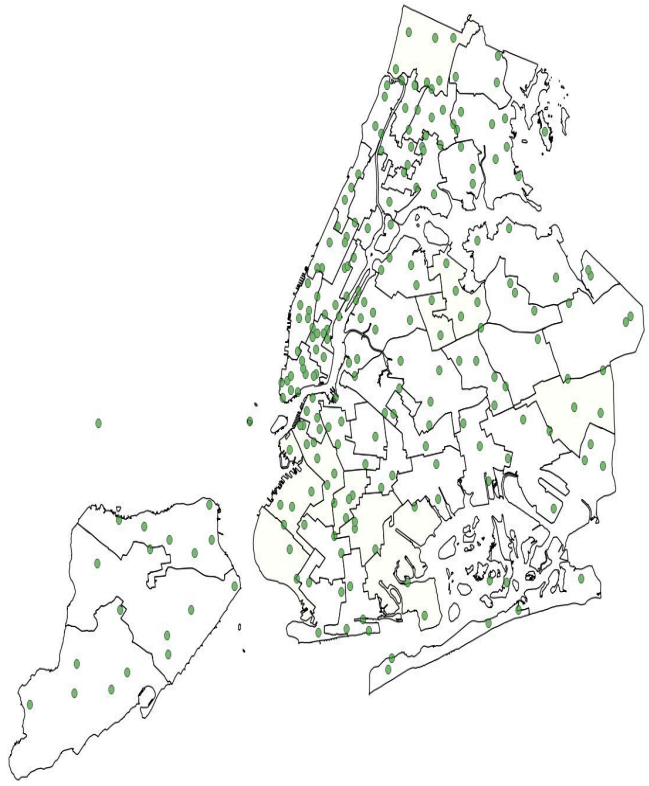
Yellow taxis able to get most of their passengers(around 80% of the trips) from Manhattan.As we can see from the heat map that yellow taxis are dominant in the Manhattan,nighbourhoods in the down west side of the Brooklyn and in the neighbourhodds of the upper side of the Queens

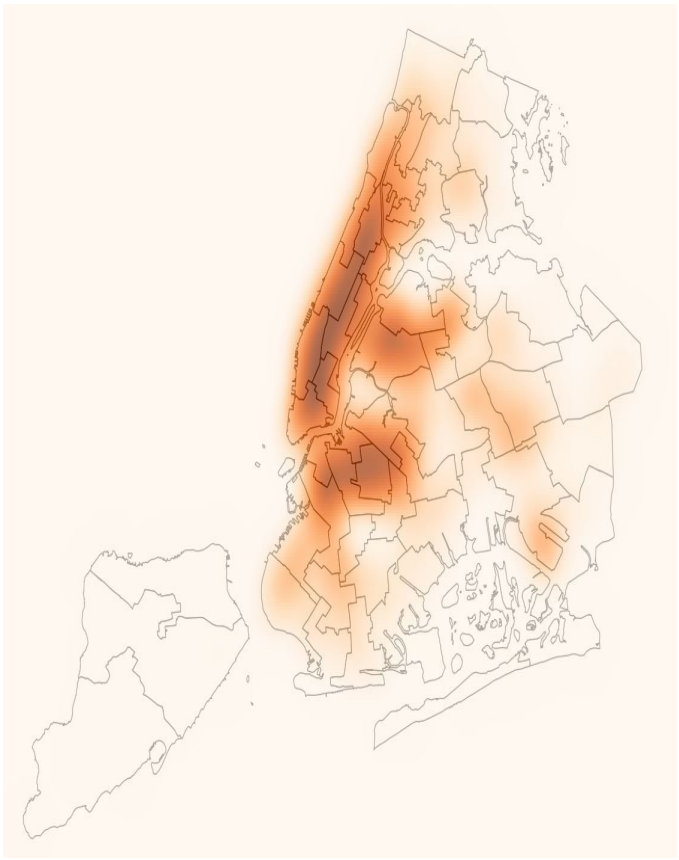
For January 2015 to June-2015











Green and Yellow taxi shows the same behaviour as that was for April 2014 to September 2014.But we can see an interesting change for the Uber Counts trend.

As you can see from the points graph for the uber that the number of the location from which Uber tends to pick up their customers has decreased drastically still some of the regions are became much darker in the heatmap for these months.

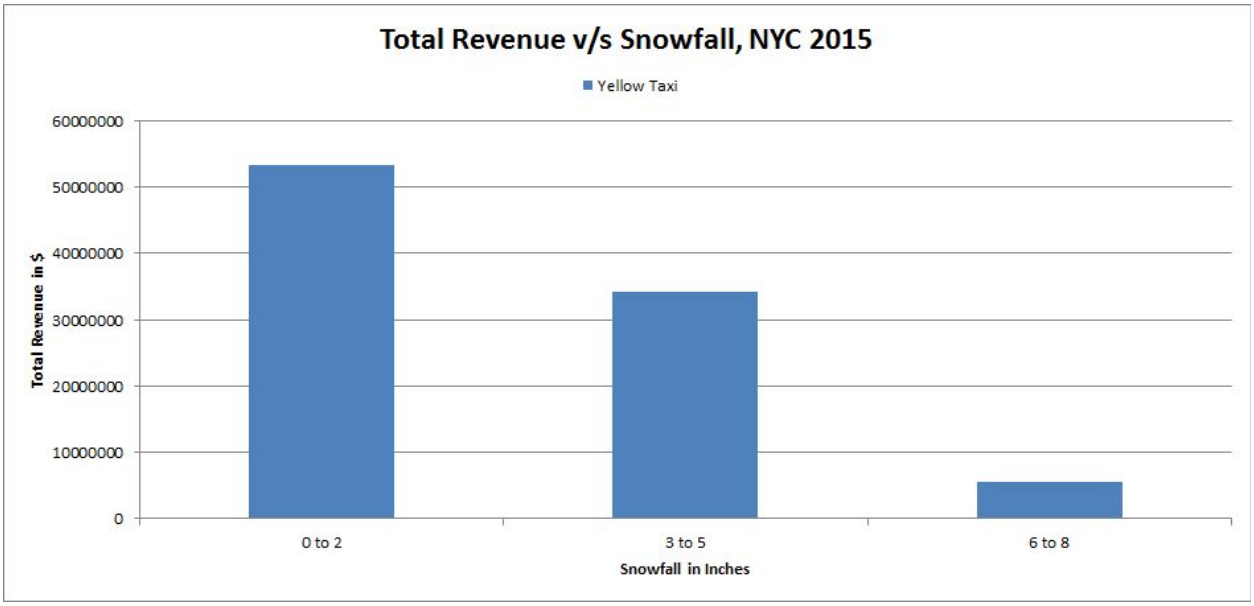
The reason for this phenomenon is that though number of locations(longitude,latitude) from which Uber pick up their customers decreases,the count of the number of trips from some locations(longitude,latitude) has tremendously increased for the month January 2016 to June 2016

Question 2

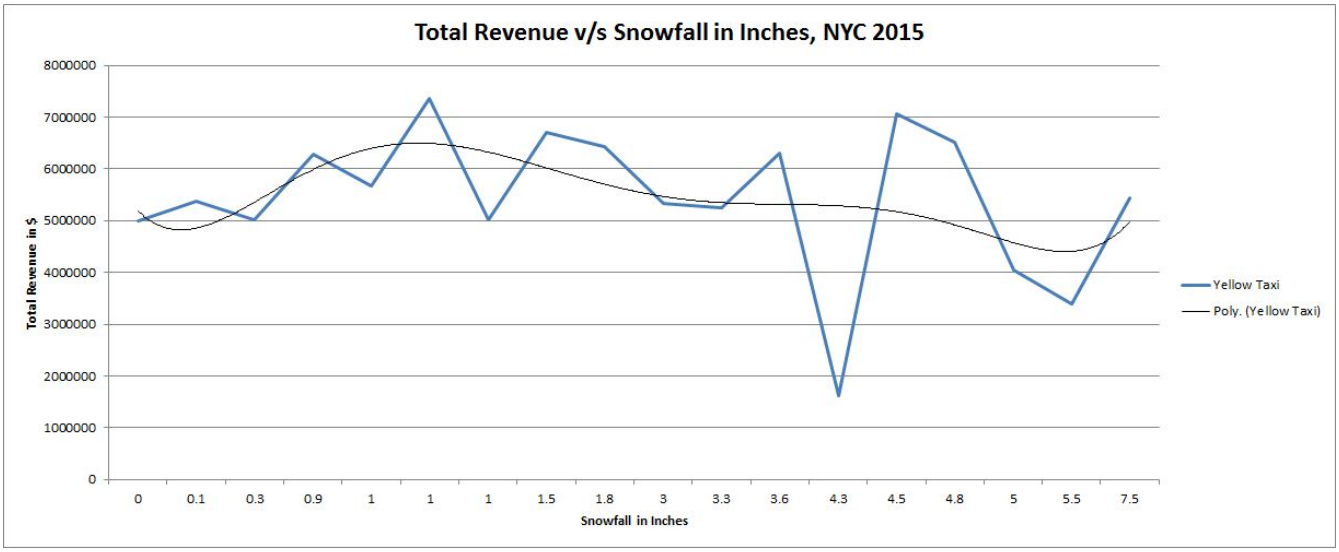
Task-2a

Yellow Taxi

Impact of Snowfall on Total Revenue per day for 2015

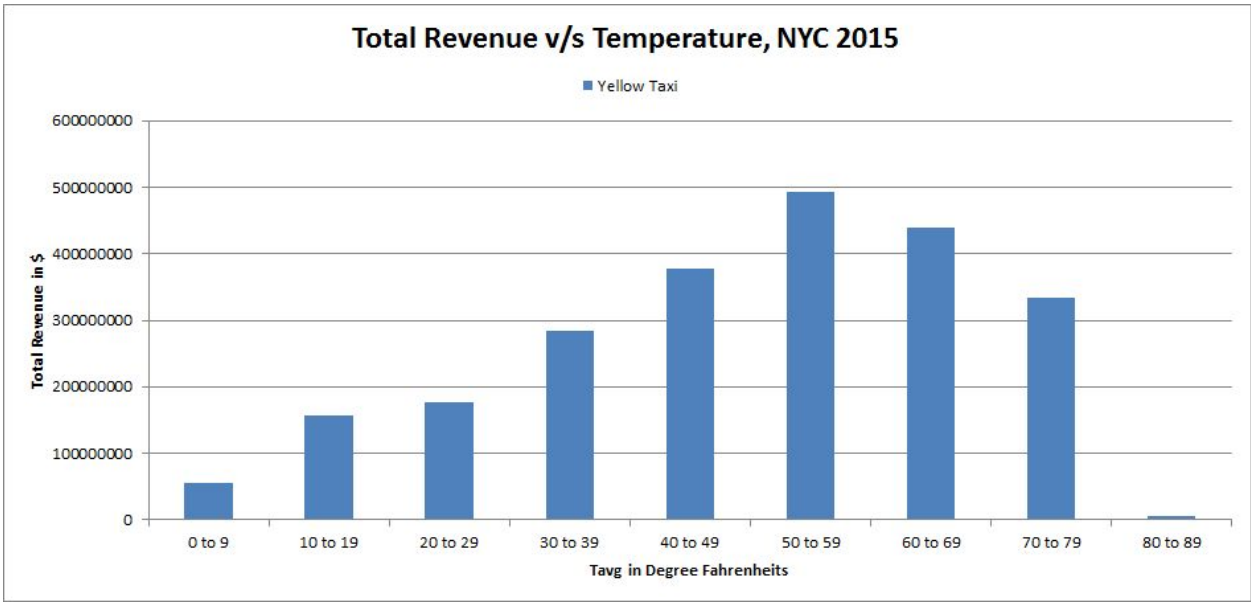


The figure indicates the total revenue in $ for the yellow taxi on y-axis and snowfall in inches on the x-axis. It can be observed from the bar graph that as the snowfall increases from left to right, the total revenue decreases. Thus,as the snowfall increased, the total revenue decreased. However, increase in snowfall does not necessarily lead to decline in revenue, since the number of days with heavy snowfall is less than number of days with light snowfall.

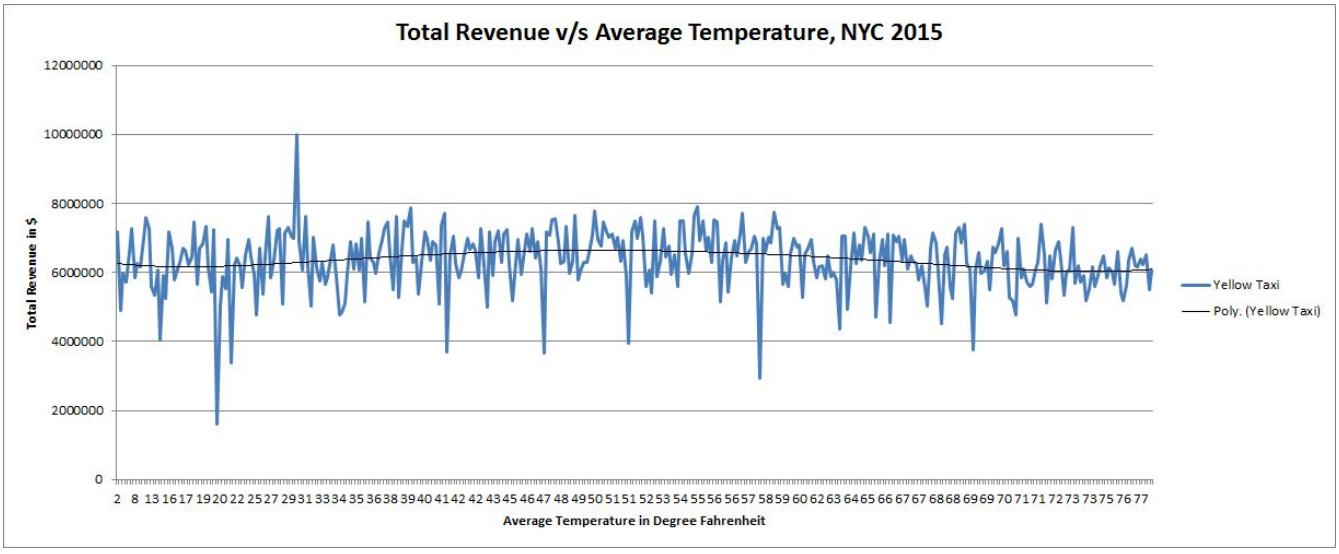


The black trendline shows that the total revenue decreases with the increase in snowfall. However the decrease is not significant.

Impact of Temperature on Total Revenue per day for 2015

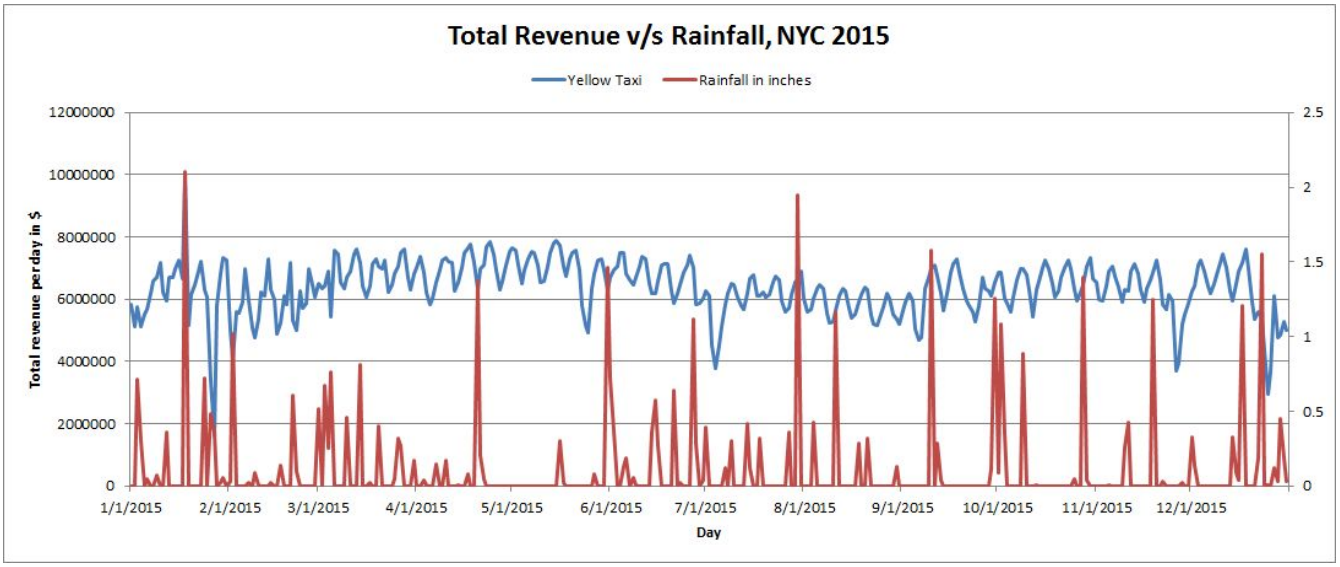


The figure indicates the total revenue in $ for the yellow taxi on y-axis and the temperature in degree fahrenheits on the x-axis. It can be observed from the bar graph that as the temperature increases from left to right, the total revenue increases and then decreases. The revenue is low when the temperature is very low or very high because the number of days with very low and very high temperature is less as compared to the number of days with optimal temperature.



The black trendline shows that the total revenue per day does not significantly vary with the average temperature per day.

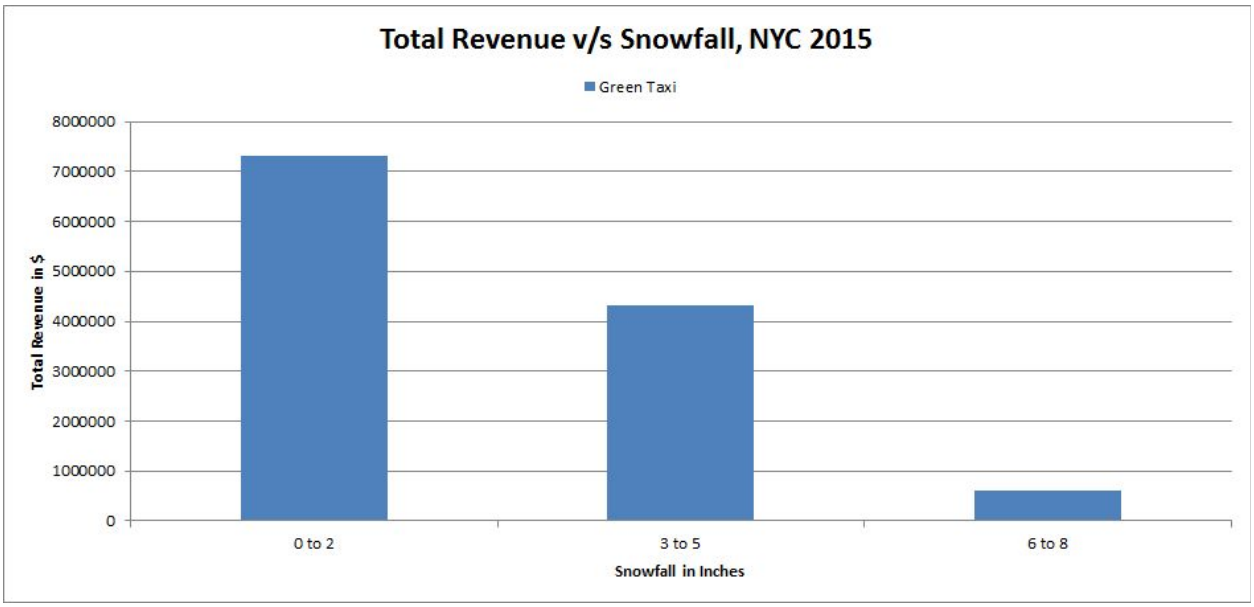
Impact of rainfall on total revenue per day for 2015



The figure shows that rainfall has a very little effect on the total revenue per day because when the rainfall is very high or very low, there is no significant change in the revenue.

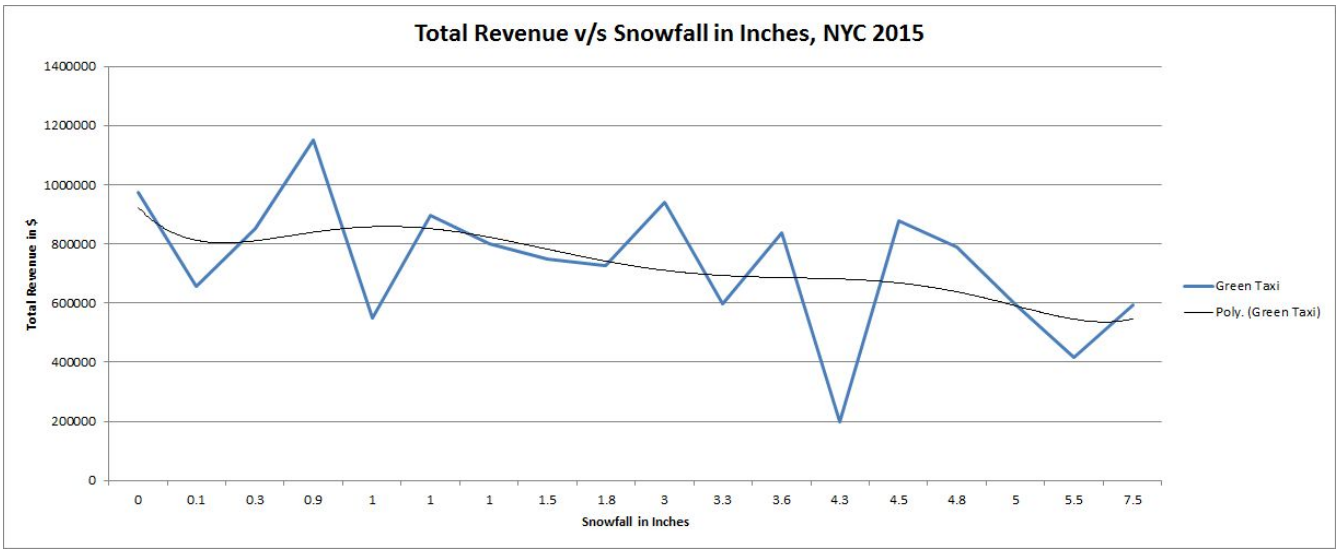
Green Taxi

Impact of snowfall on total revenue per day for 2015



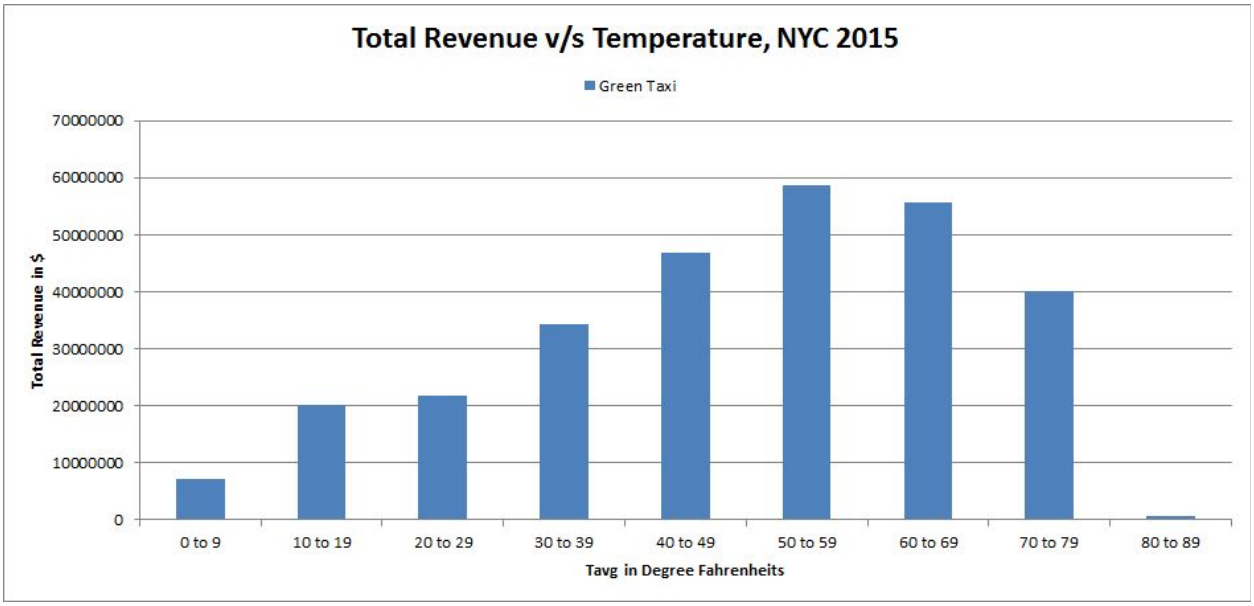
The figure indicates the total revenue in $ for the green taxi on y-axis and snowfall in inches on the x-axis. It can be observed from the bar graph that as the snowfall increases from left to right, the total revenue decreases. Thus,as the snowfall increased, the total revenue decreased. However, increase in

snowfall does not necessarily lead to decline in revenue, since the number of days with heavy snowfall is less than number of days with light snowfall. This trend is same as for the yellow taxi.



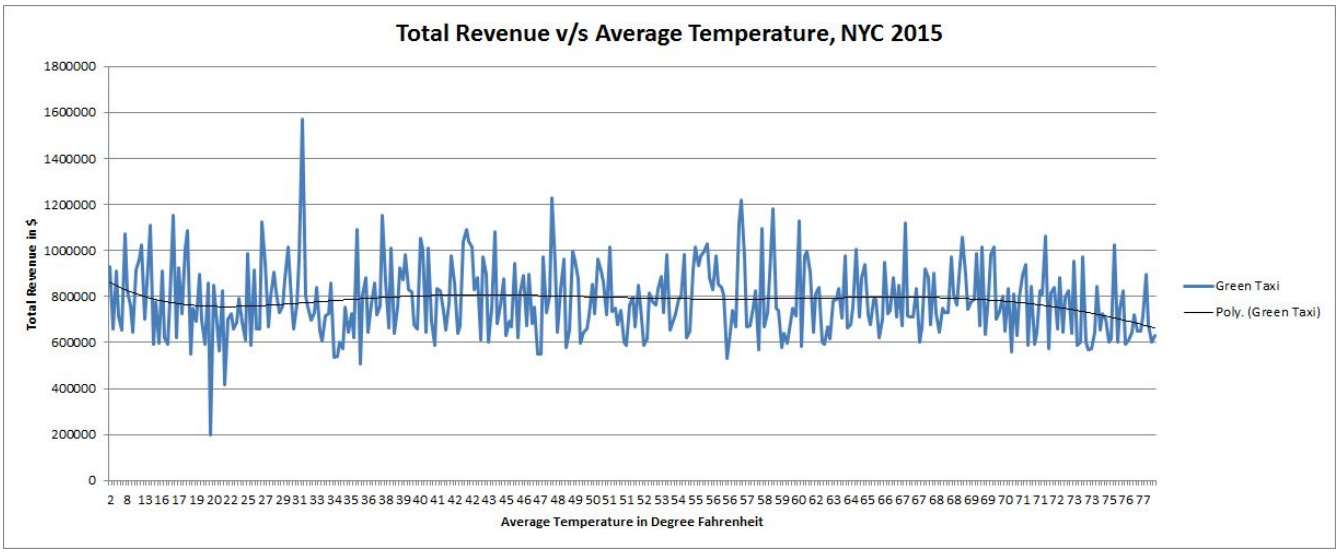
The black trendline shows that the total revenue decreases with the increase in snowfall. However the decrease is not significant.

Impact of temperature on total revenue for 2015



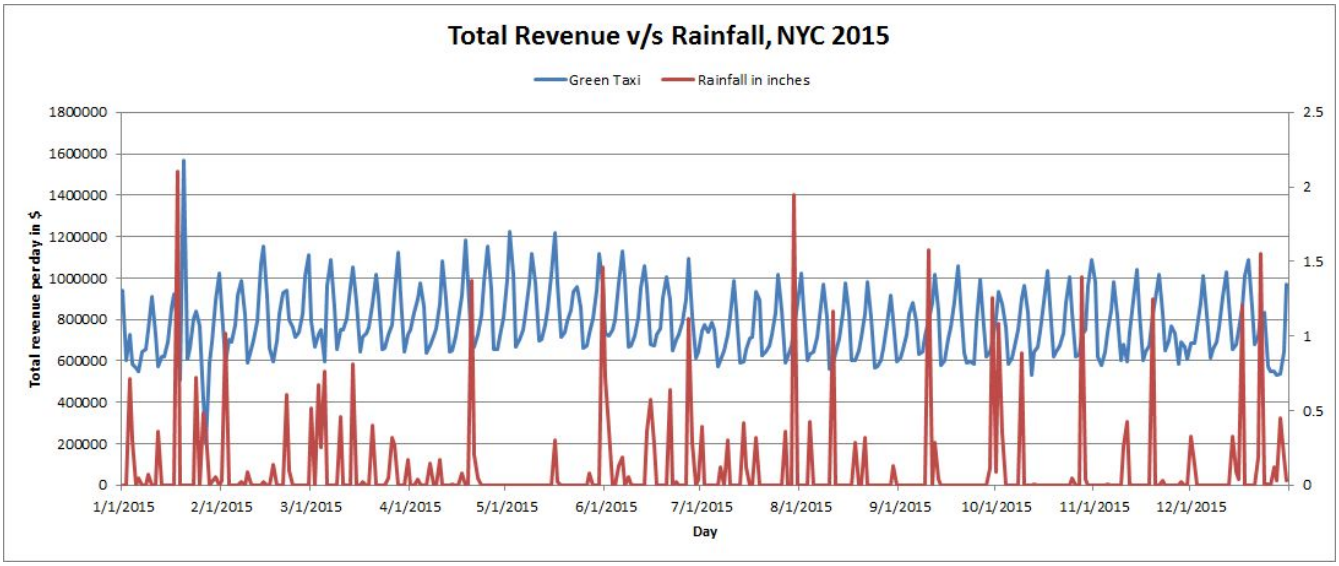
The figure indicates the total revenue in $ for the green taxi on y-axis and the temperature in degree fahrenheits on the x-axis. It can be observed from the bar graph that as the temperature increases from left to right, the total revenue increases and then decreases. The revenue is low when the temperature is

very low or very high because the number of days with very low and very high temperature is less as compared to the number of days with optimal temperature. This trend is same as for the yellow taxi.



The black trendline shows that the total revenue per day does not significantly vary with the average temperature per day.

Impact of Rainfall on total revenue for 2015



The figure shows that rainfall has a very little effect on the total revenue per day because when the rainfall is very high or very low, there is no significant change in the revenue.

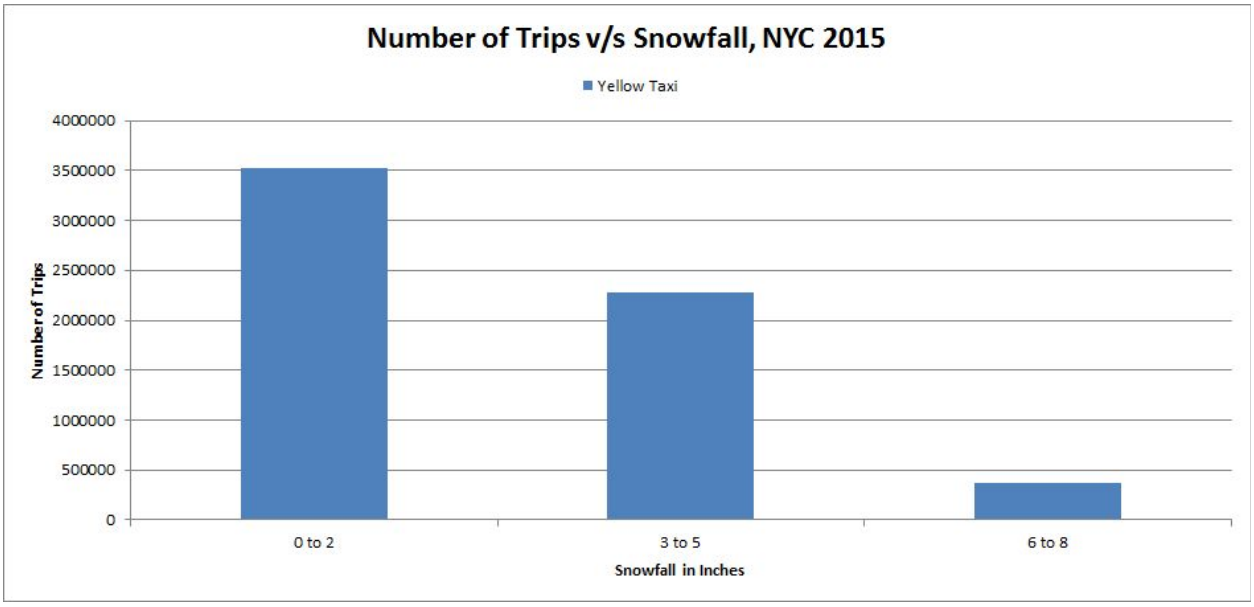
Thus we conclude that while snowfall has a little effect on the total revenue per day, temperature and rainfall does not significantly impact the total revenue per day. This is because the weather hardly remains constant throughout the day. It rains or snows for some duration of the day which does not affect the total revenue per day.

Task -2b

Number of Trips

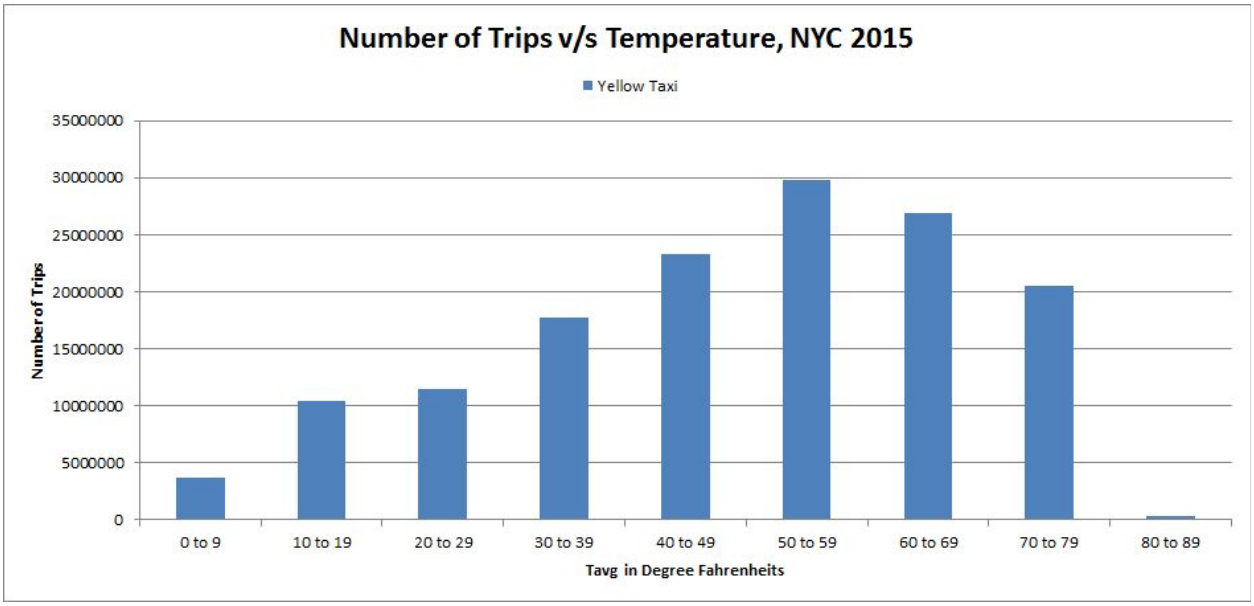
Yellow Taxi

Impact of snowfall on number of trips for 2015



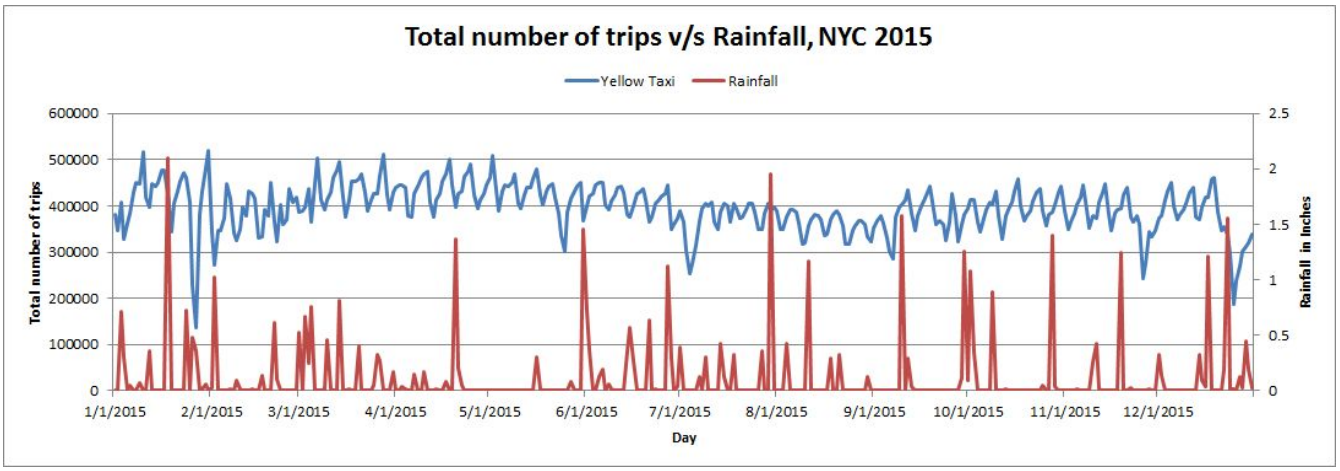
The figure indicates the number of trips for the yellow taxi on y-axis and snowfall in inches on the x-axis. It can be observed from the bar graph that as the snowfall increases from left to right, the total number of trips decreases.Thus,as the snowfall increased, the total number of trips decreased. However, increase in snowfall does not necessarily lead to decline in number of trips, since the number of days with heavy snowfall is less than number of days with light snowfall.

Impact of temperature on number of trips for 2015



The figure indicates the total number of trips for the yellow taxi on y-axis and the temperature in degree fahrenheits on the x-axis. It can be observed from the bar graph that as the temperature increases from left to right, the total number of trips increases and then decreases. The number of trips is low when the temperature is very low or very high because the number of days with very low and very high temperature is less as compared to the number of days with optimal temperature.

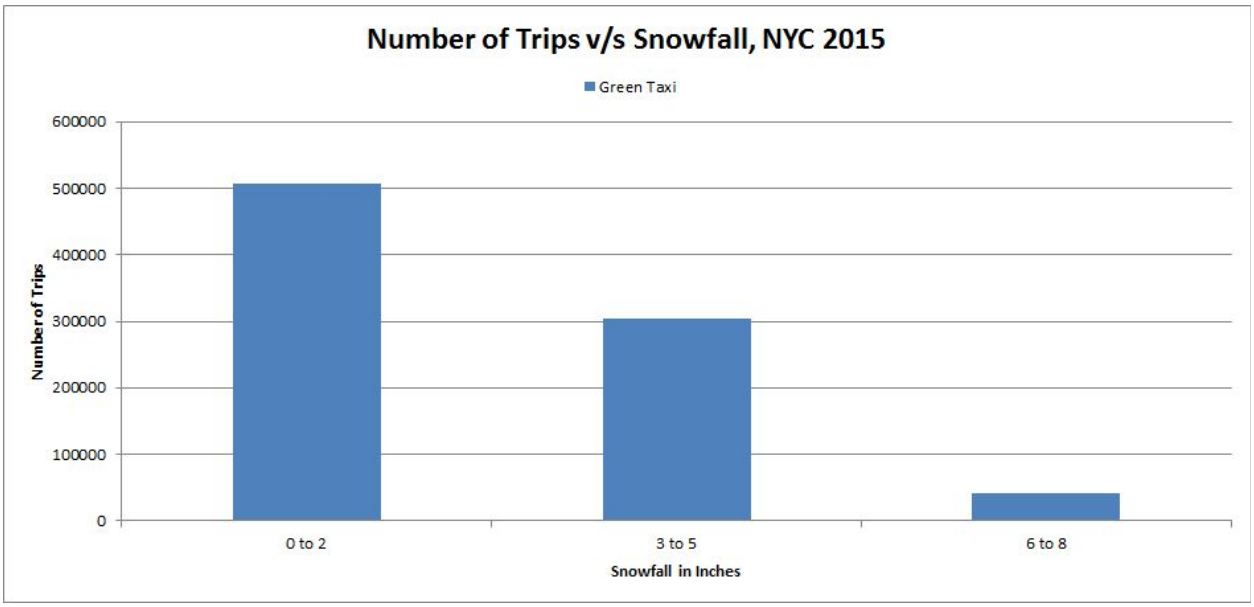
Impact of rainfall on number of trips for 2015



The figure shows that rainfall has no impact on the total number of trips of yellow taxi because the rainfall occurs only for a small duration in a day and does not affect the total revenue per day.

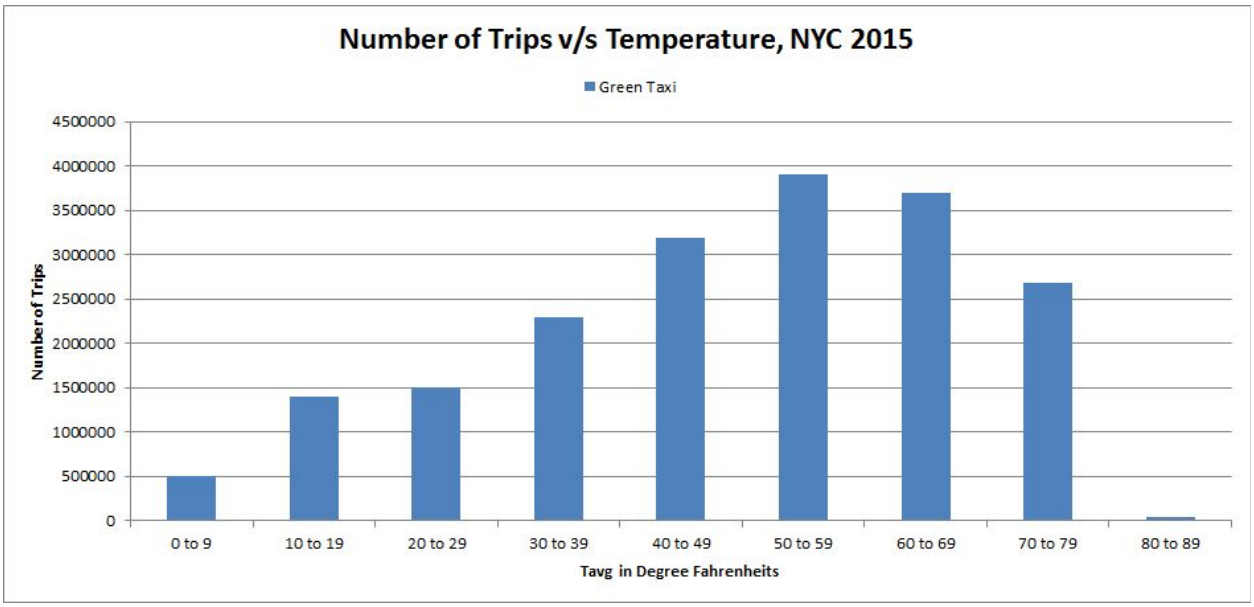
Green Taxi

Impact of snowfall on the number of trips for 2015



The figure indicates the number of trips in $ for the green taxi on y-axis and snowfall in inches on the x-axis. It can be observed from the bar graph that as the snowfall increases from left to right, the total number of trips decreases. Thus,as the snowfall increased, the total number of trips decreased. However, increase in snowfall does not necessarily lead to decline in number of trips, since the number of days with heavy snowfall is less than number of days with light snowfall.

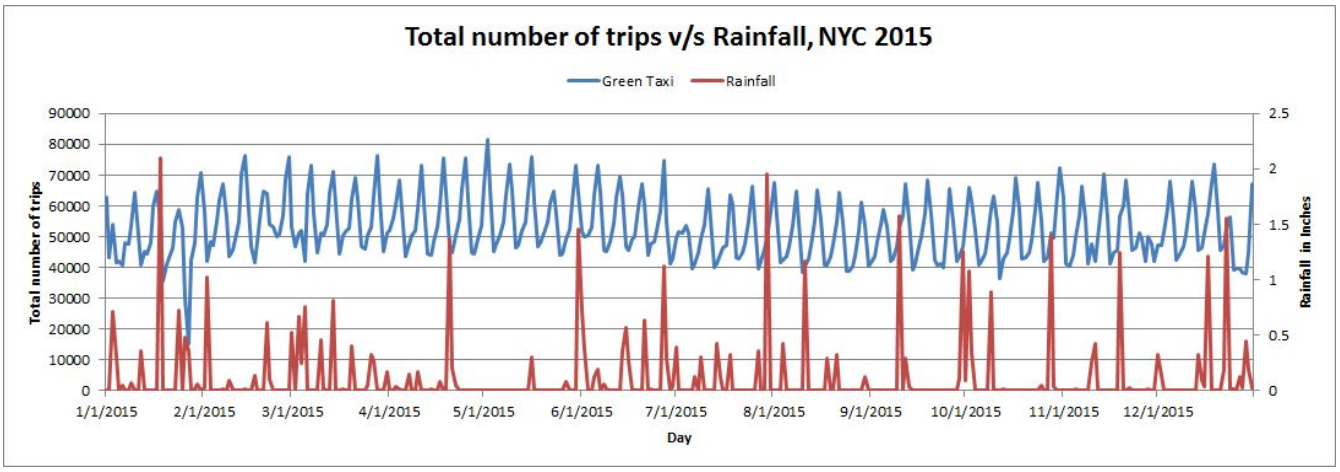
Impact of temperature on the number of trips for 2015



The figure indicates the total number of trips for the green taxi on y-axis and the temperature in degree fahrenheits on the x-axis. It can be observed from the bar graph that as the temperature increases from

left to right, the total number of trips increases and then decreases. The number of trips is low when the temperature is very low or very high because the number of days with very low and very high temperature is less as compared to the number of days with optimal temperature.

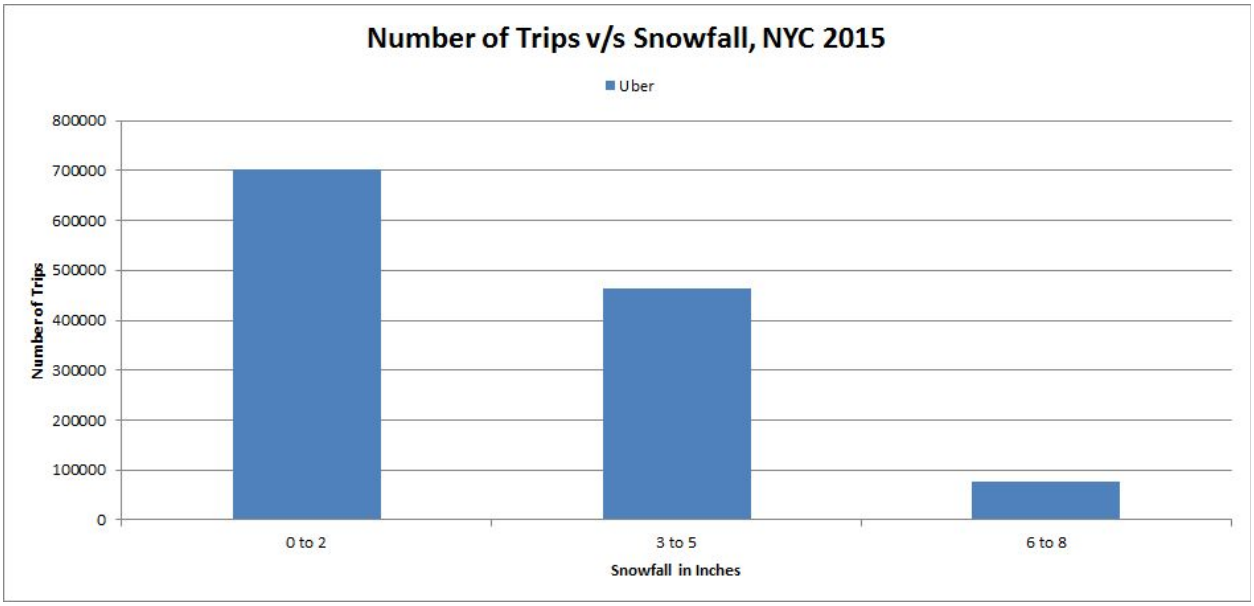
Impact of rainfall on green taxi for 2015



The figure shows that rainfall has no impact on the total number of trips of green taxi because the rainfall occurs only for a small duration in a day and does not affect the total revenue per day.

Uber

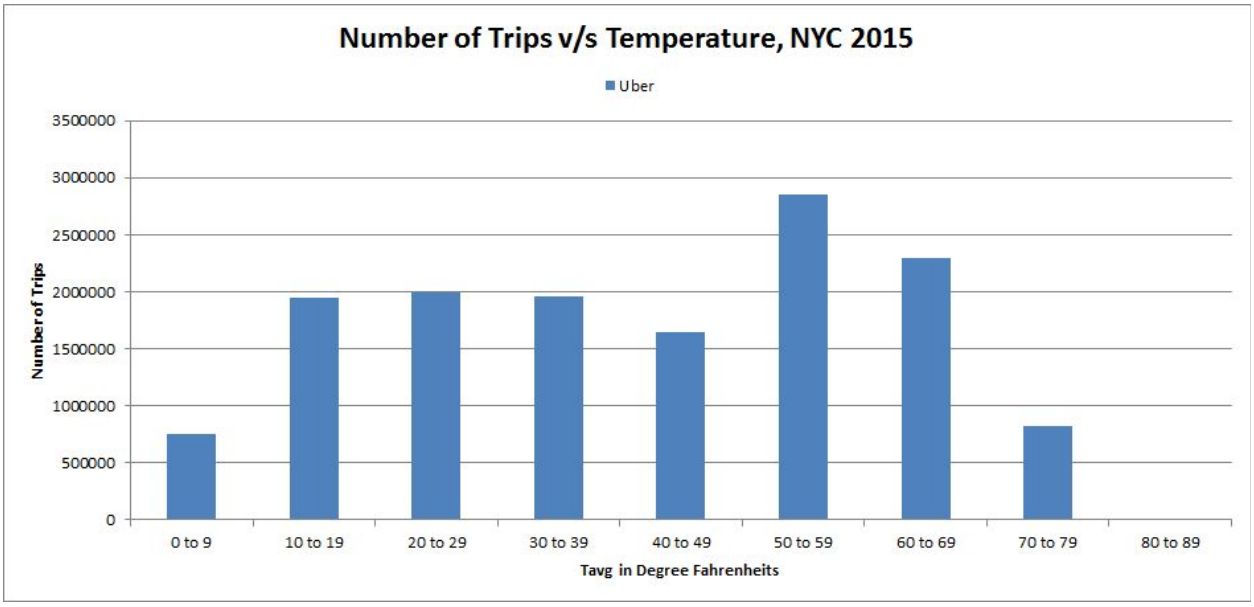
Impact of snowfall on the number of trips for 2015



The figure indicates the number of trips in $ for the uber on y-axis and snowfall in inches on the x-axis. It can be observed from the bar graph that as the snowfall increases from left to right, the total number of trips decreases. Thus,as the snowfall increases, the total number of trips decreased. However,

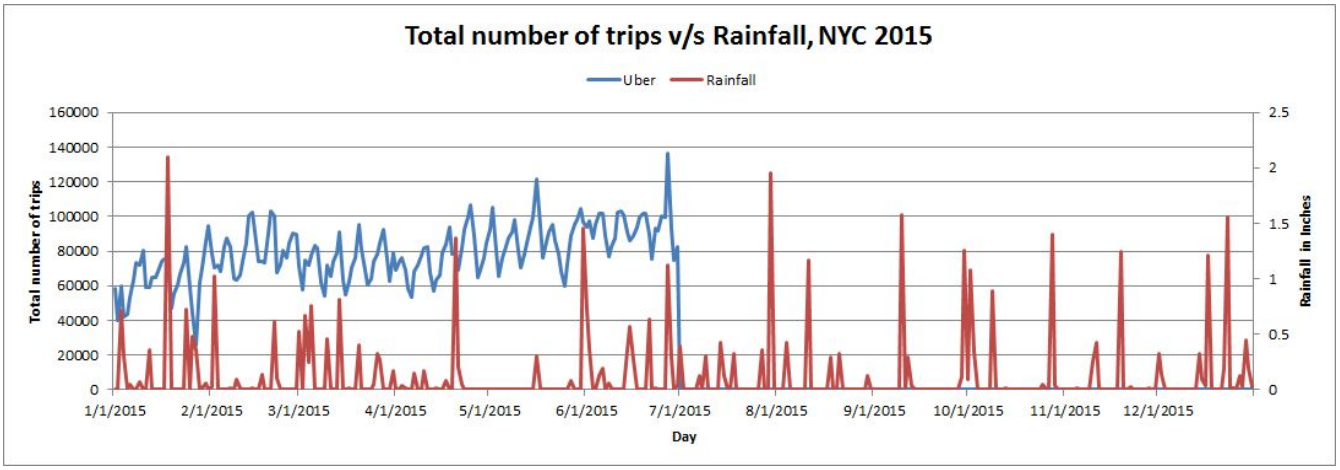
increase in snowfall does not necessarily lead to decline in number of trips, since the number of days with heavy snowfall is less than number of days with light snowfall.

Impact of temperature on number of trips per day for 2015



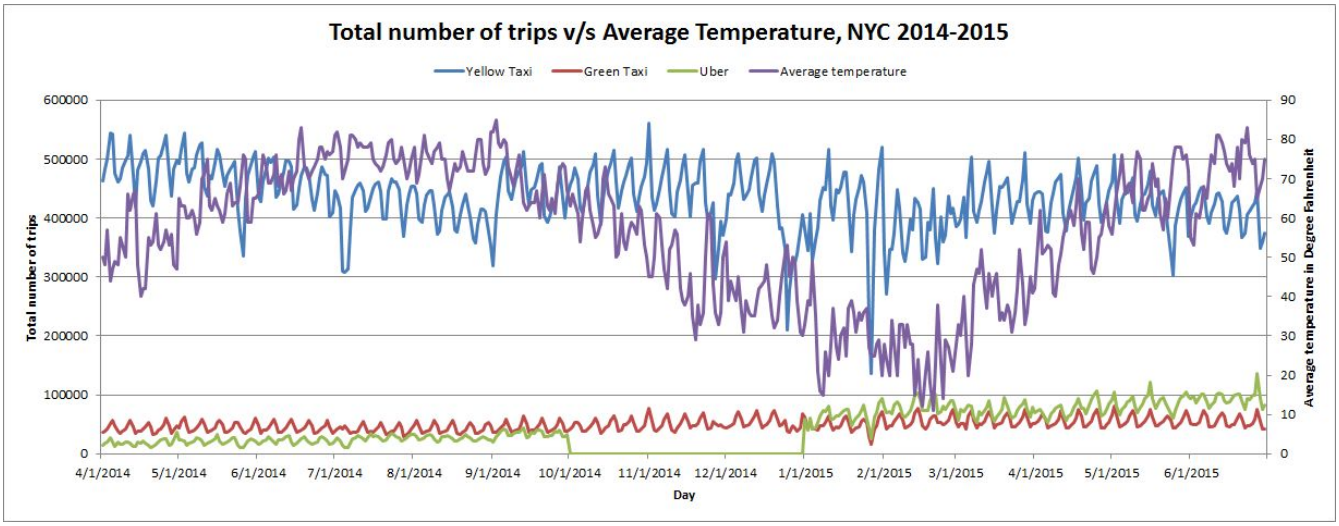
The figure indicates the total number of trips for the uber on y-axis and the temperature in degree fahrenheits on the x-axis. It can be observed from the bar graph that as the temperature increases from left to right, the total number of trips varies randomly. Thus temperature has very little effect on the number of trips.

Impact of rainfall on the number of trips per day for 2015



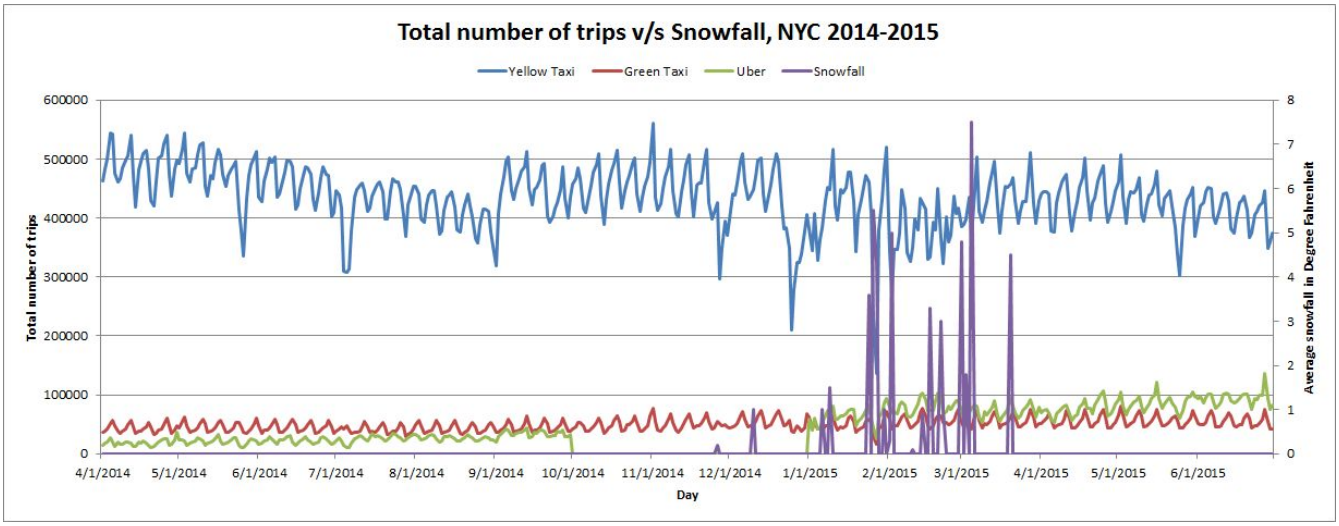
The figure shows that rainfall has no impact on the total number of trips of uber because the rainfall does not occur continuously during the day and does not affect the total revenue per day.

Impact of temperature on the total number of trips per day from April 2014 to June 2015



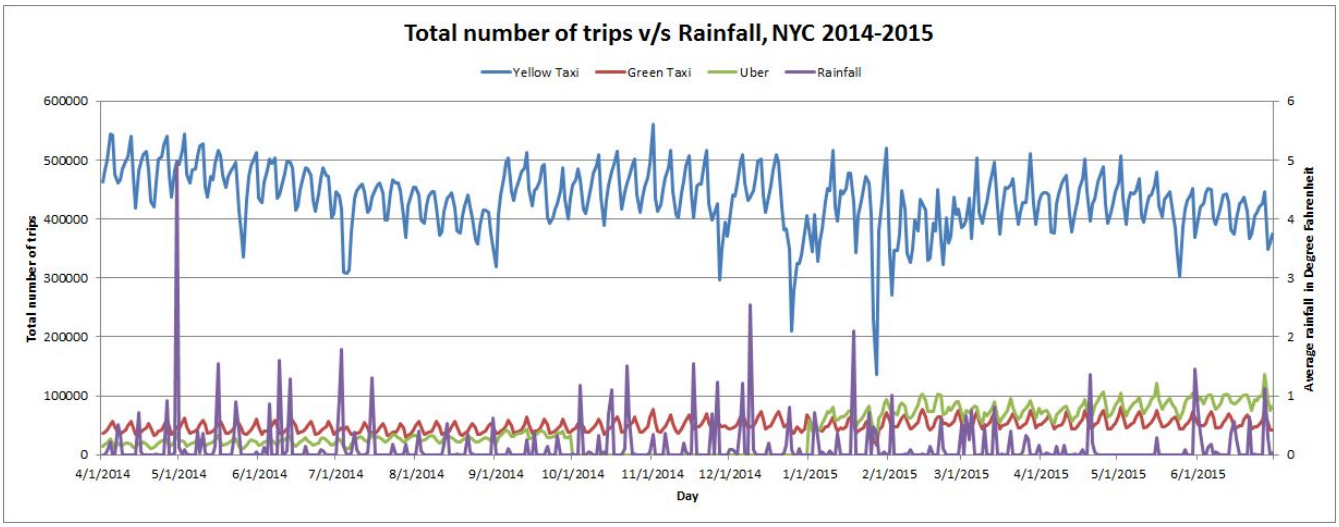
The figure indicates the total number of trips per day on the left y-axis and the average temperature on right y-axis. The number of trips increases to some extent when the average temperature per day rises but it is not significant impact on the total number of trips of green taxi, yellow taxi and uber per day. This is indicated in the above graph when the temperature is low around Jan to March 2015, the total number of trips per day follows similar trend as before Jan 2015 or after March 2015.

Impact of snowfall on the total number of trips per day from April 2014 to June 2015



The figure indicates the total number of trips per day on the left y-axis and the average snowfall on the right y-axis. The average snowfall per day has very little significant impact on the total number of trips per day. This is because the snowfall does not occur continuously throughout the day and thus the total number of trips per day is affected only for the duration of the snow.

Impact of rainfall on the total number of trips per day from April 2014 to June 2015

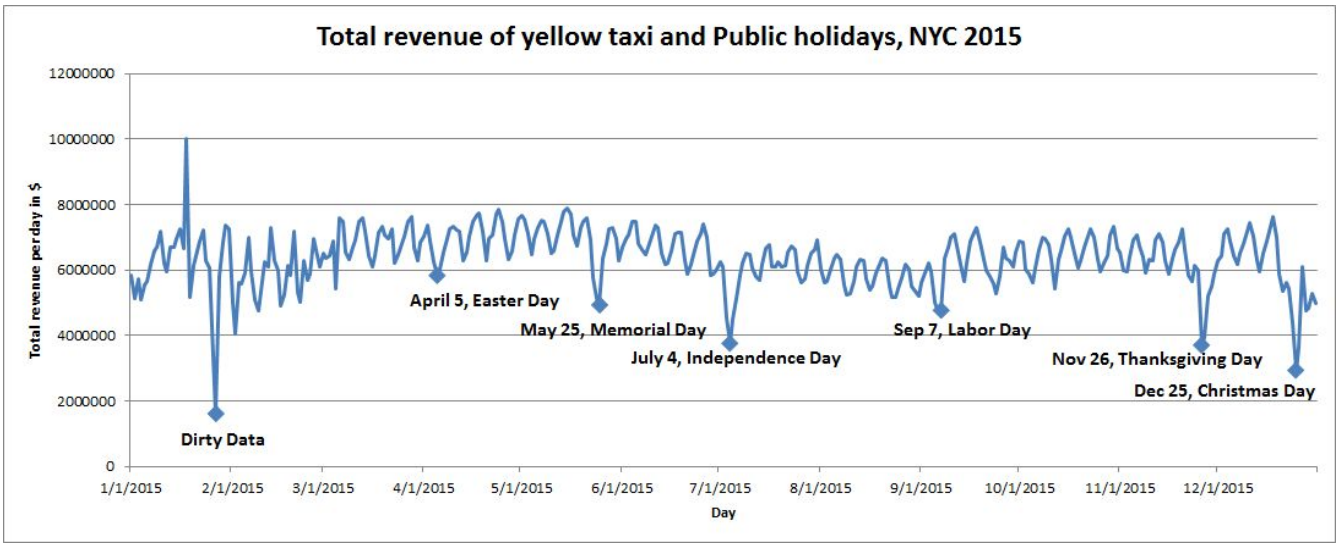


The figure shows that the rainfall has no effect on the total revenue per day for yellow taxi, green taxi and uber. This is because the rainfall occurs intermittently in a day and this does not affect the total revenue per day.

Thus we conclude that while snowfall had an impact on the total number of trips per day but not significant and rainfall and temperature did not have an significant impact on the total number of trips

per day. This is because the snowfall, rainfall and temperature are not constant throughout the day and occur intermittently in a day, it does not affect the total number of trips per day significantly.

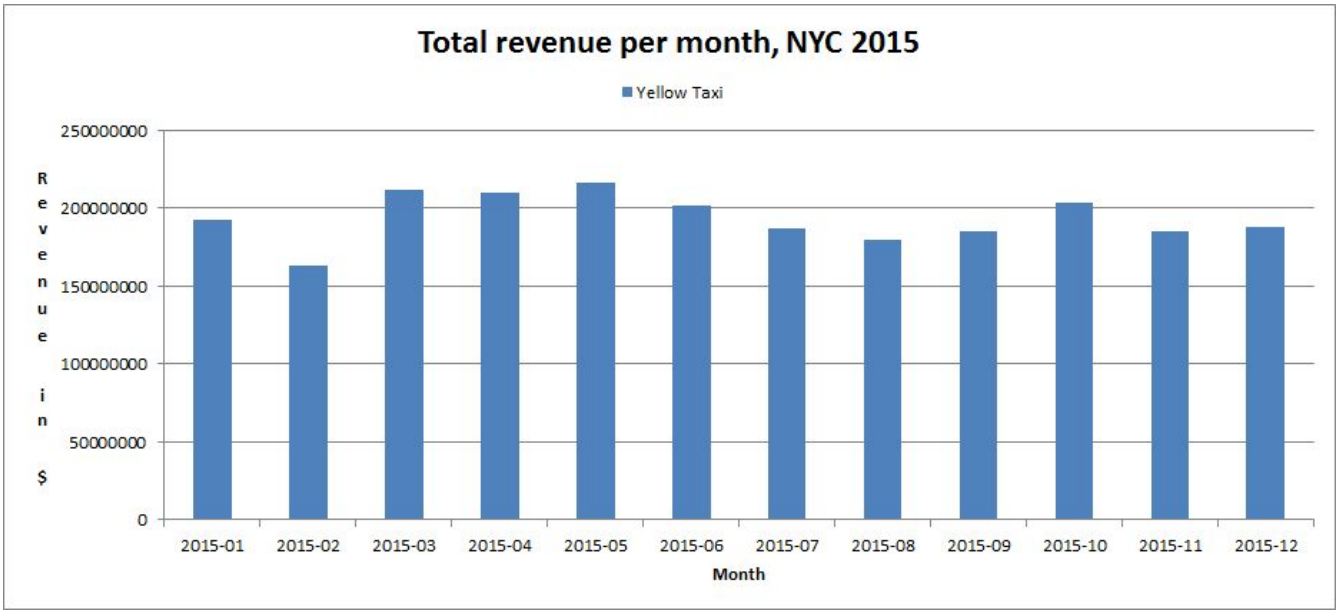
Impact of public holidays on the total revenue per day



The figure shows that the total revenue per day decreases significantly on Easter day, Memorial day, Independence day, Labor day, Thanksgiving day and christmas day. Thus we conclude that public holidays has a significant impact on the total revenue per day because on public holidays not too many people tend to leave their house and thus the number of trips per day decreases significantly. The dirty data is an outlier on January 27, 2015 since there was no public holiday or any event.

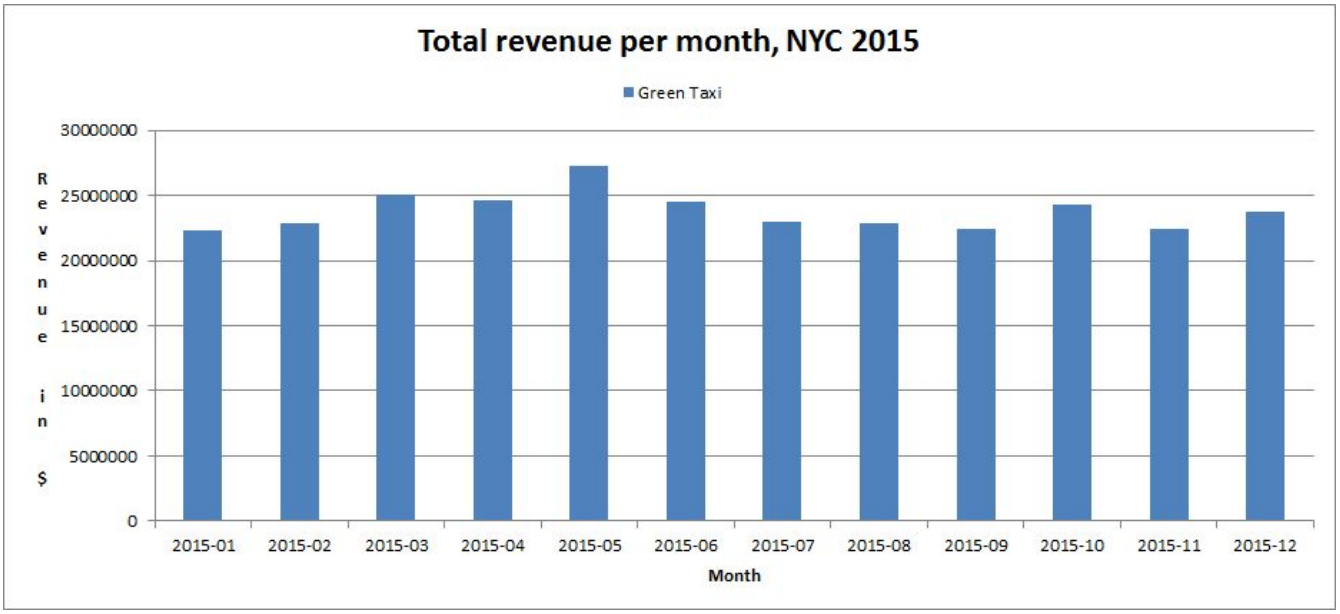
Task-2c

Total revenue per month of yellow taxi for the year 2015



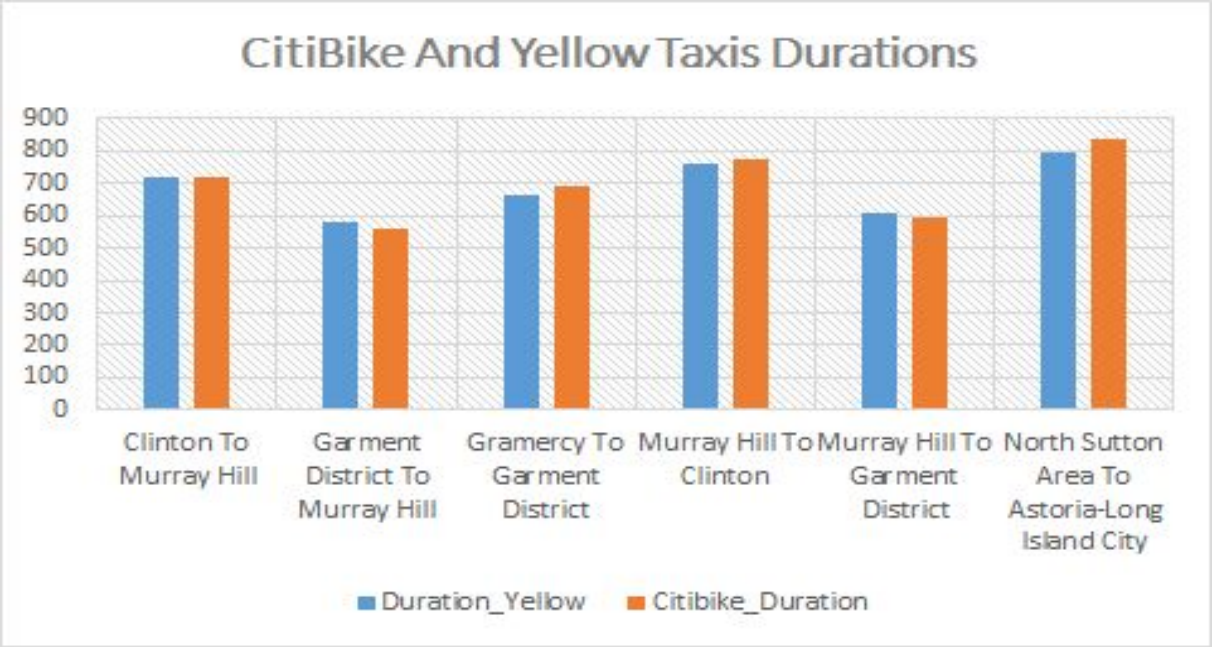
The total revenue is maximum for the month of may and more during the last months of spring and beginning of the summer.

Total revenue per month of green taxi for the year 2015



The trend is similar to the yellow taxi. The total revenue is maximum during the month of may and slightly more during the end of spring and beginning of summer.

For comparing CitiBike Trip Duration and Yellow Taxi Durations between two locations



As we can see that travelling to Gramercy to Garment District,Murray Hill to Clinton and North Sutton Area to Astoria-Long Island City Citibike takes less average time than Yellow taxis.

8.Reproducibility

All the information for Reproducing the output are listed into the github repository and experimental setup steps are included into Section 5 in the document

9.Conclusion

We have taken four data set in our project to do the meaningful analysis

1)Yellow Taxi data

2)Uber data

3)Green Taxi data

4)Weather data

We successfully able to get the relationship between weather data set and trip counts for the 15 months(April 2014 -May 2015)

We answered all the questions which we set out to answer at the beginning of the project:

1) **How does Uber and green taxi arrival impact the Yellow Taxi in New York?**

Yellow taxis are clear winners in Manhattan. The number of trips have remain steady in Manhattan whereas uber is showing sharp rise which means somehow uber is finding passengers who don't usually take the taxi. Uber is the most successful in Brooklyn as compared to its counterparts. Green is close second. Yellow is a leader in Queens thanks to both the airports as green is not allowed to operate near those airports. Bronx and Harlem belongs to Green taxi.

In Manhattan uber enjoys significant customers from East Manhattan area whereas yellow taxi is localised more to the central manhattan area. Green is more or less absent in Manhattan which law requires it too but it has taken over the Manhattan region. Yellow is leading the market when it comes to both airports it has almost four times the passengers from the airports than the remaining part of the Queens. Uber also has a huge amount passengers from the airports but it is significantly less than Yellow Taxis.

People tend to avoid uber between 00 and 0400 Hours. Also during weekdays 0800 to 0900 and 1800 to 1900 hours is time when people generally take taxis and in weekends traffic increases gradually through the night till 00 hours.

2. a. How is the total revenue per day of the yellow and green taxi affected by the weather in NYC ?

NYC is a metropolitan and a busy city. Weather has some impact impact on the total revenue per day. People tend to travel in taxi during snow or rain rather than walking. With increase in temperature, the revenue decreases to some extent because people prefer walking. However because of the culture and lifestyle of the city, it is ever busy and thus weather has little impact on the total revenue per day unless there is a heavy snowfall or any other severe weather condition.

b. How is the total number of trips per day of yellow and green taxi and uber affected by the weather in NYC?

Total number of trips per day and the total revenue per day are directly correlated. As stated in the part a above for the total revenue, the total number of trips is related to weather to some extent but the impact is not significant except for some severe weather condition or snowfall.

c. How is the total number of trips per day of yellow taxi affected on the public holidays ?

Public holidays have significant impact on the total number of trips per day. Most of the people tend to stay at home on public holidays and as a result the total number of trips per day drops significantly. Decrease in the number of trips also leads to decrease in the total revenue per day.

3. Which are popular nightlife locations in New York city ? At what location do people of new york typically hangout between 9 pm and 2 am in both weekends and weekdays?

We were successfully able to triangulate the active/party zones in NYC using dropoff data from Yellow and green taxis. To name the few :- Williamsburg in Brooklyn,Bushwick in Brooklyn,Harlem in Manhattan and Lower east side of Manhattan.

10. References

https://github.com/ViDA-NYU/TaxiVis

http://fivethirtyeight.com/tag/uber/

11. Contribution

Sahil Shah (N12706992)	Question 1, 3 and 4 Per month analysis with long lat Per hour analysis Per month analysis Citibike vs yellow duration
Fenil Tailor (N18730085)	Question 1,3 and 4 Nightlife analysis Per hour analysis Per month analysis Citibike vs yellow duration
Ajinkya Avinash Shukla (N17644394)	Question 2 Weather impact on total revenue per day Weather impact on total trips per day Impact of public holidays on total revenue