NYC GREEN CABS DATA ANALYSIS

SUMMARY REPORT

Introduction:

This analysis was done based on the Nyc taxi and limousine commission data on green cabs trip information for a particular month(February 2016) for a particular span of time(1st to 14th February).

In the analysis, the data was suitably divided into time slots such as morning evening and night so that a detailed information about trip origins in all the boroughs and exact locations of the ride pickups would be known.That would give us a fair idea where most of the green-cabs pickup originate and how to would help the drivers if they tend to be in those regions in those specific time slot.The Data was also divided into the 5 boroughs so that some trends could be estimated from the data like which borough had more revenue share and where people tend to pay more tips in.

A weather data was also taken from w2.weather.gov for that time frame of 1st to 14th and the number of green cab rides were compared to see if weather changes affected people`s intention to take a cab.

Programming tool used was R.

The full code is here: https://github.com/soham287/NYC-Green-Cab-Recommendation

Pre-processing Data:

Basic task of pre-processing the data was to start with taking just the number of days that was asked to be put in consideration and remove the other dates.

Regarding location, the only data fields related to location in the csv files are the latitude and longitude. However I wanted to find which boroughs each record belongs to. One way I could do is through some geolocation API such as Google Maps API. The problem is that google charge for it and limit the frequency that we can query per day. 2500 per day is of course not suitable for this project. So I downloaded the NYC shape file from Zillow which sets the boundary for each boroughs.The Shape-file was not giving the correct estimation of Staten Island and somehow producing it to be NA,but I double checked the lats and longs of those entries just in case to understand those belonged to Staten Island.

CODE IN R –

```
data<-read.csv('https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2016-02.csv', header = T, sep = ',')
#Changing the pickup and drop-off time to POSIX
data$lpep_pickup_datetime<-as.POSIXct(strptime(data$lpep_pickup_datetime, "%Y-%m-%d %H:%M:%S"))
data$Lpep_dropoff_datetime<-as.POSIXct(strptime(data$Lpep_dropoff_datetime, "%Y-%m-%d %H:%M:%S"))
#Taking the data only between 1st and 14th of February.
mainData<-subset(data,data$Lpep_dropoff_datetime<"2016-02-14 23:59:59 EST")
```

```
# Added Burrough and area information to the data by reversegeocoding the latitude and longitude using Zillow Shapefile
nyc.shp <- readShapePoly("ZillowNeighborhoods-NY.shp")
proj4string(nyc.shp) <- CRS("+proj=longlat+ellps=WGS84")
coordinates<- data.frame(mainData$Pickup_longitude,mainData$Pickup_latitude)
coordinates.sp<-coordinates
coordinates.sp<-SpatialPoints(coordinates.sp,proj4string=CRS("+proj=longlat+ellps=WGS84"))
city.data<-over(coordinates.sp,nyc.shp)
County<-city.data$COUNTY
Location<-city.data$NAME
mainData<-cbind(mainData,data.frame(County,Location))
mainData$County<-as.character(mainData$County)
mainData$County[is.na(mainData$County)]<-"Staten Island"
}
```

Analysis based on Time Slots

Instead of trying to figure out for all the 14 days where a cab driver gets more pickup,it was assumed that finding in which time of the day, locations are more popular for green cab pickup was assumed.

Here is the code for that :

```
noonTime="12:00:00"
eveTime="16:00:00"
nightTime="19:00:00"
peakTimeStart="07:00:00"
peakTimeEnd="11:00:00"
beforeMidNightTime="23:59:59"
midnightTime="00:00:00"
morningTime="04:00:00"
#Breaking this data into Morning Slot

mornData<-subset(mainData,strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")>=morningTime & strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")<=noonTime)
mornData<-mornData[!(mornData$lpep_pickup_datetime %in%
peakmornData$lpep_pickup_datetime),]


#Breaking this data into Peak hours Morning Slot
peakmornData<-subset(mainData,strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")>peakTimeStart & strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")<=peakTimeEnd)

#Breaking this data into Noon Slot
noonData<-subset(mainData,strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")>noonTime & strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")<=eveTime)

#Breaking this data into Evening Slot
```

```
eveData<-subset(mainData,strftime(mainData$lpep_pickup_datetime, format="%H:%M:%S")>eveTime
& strftime(mainData$lpep_pickup_datetime, format="%H:%M:%S")<=nightTime)

#Breaking this data into Night Slot

nightData<-subset(mainData,strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")>nightTime & strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")<=beforeMidNightTime)

#Breaking this data into MidNight Slot

midnightData<-subset(mainData,strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")>=midnightTime & strftime(mainData$lpep_pickup_datetime,
format="%H:%M:%S")<=morningTime)
```

Analysis-

PEAK-HOURS :

	LOCATIONS	PEAK-HOUR PICK-UP COUNT
1	Harlem	15075
2	Fort Green	7934
3	Yorkville	7248
4	Astoria-Long Island City	6113
5	East Harlem	5384
6	Morningside Heights	4562
7	Williamsburg	4539
8	Bedford-Stuyvesant	4256
9	Washington Heights	4165
10	Sunny Side	3842

MORNING-HOURS:

	LOCATIONS	MORNING-HOUR PICKUPCOUNT
1	Harlem	5574
2	Jackson Heights	4122
3	Astoria-Long Island City	3877
4	Fort Green	3455
5	Williamsburg	2907
6	Yorkville	2843
7	East Harlem	2226
8	Bedford-Stuyvesant	2173
9	Washington Heights	1909
10	Sunny Side	1735

NOON-HOURS:

	LOCATIONS	NOON-HOUR PICKUPCOUNT
1	Harlem	15591
2	Yorkville	8604
3	Astoria-Long Island City	8177
4	Fort Green	8113
5	Morningside Heights	6384
6	East Harlem	5800
7	Jackson Heights	5227
8	Washington Heights	4897
9	Downtown	4805
10	Bedford-Stuyvesant	4492

EVENING-HOURS:

NIGHT-HOURS:

	LOCATIONS	EVENING-HOUR PICKUPCOUNT
1	Harlem	15235
2	Yorkville	9324
3	Fort Green	9199
4	Astoria-Long Island City	8869
5	Morningside Heights	6937
6	Downtown	6033
7	Jackson Heights	6033
8	Williamsburg	5760
9	East Harlem	5368
10	Park Slope	4933

	LOCATIONS	NIGHT-HOUR PICKUPCOUNT
1	Harlem	20538
2	Fort Green	17369
3	Astoria-Long Island City	14978
4	Williamsburg	13502
5	Jackson Heights	10364
6	Park Slope	8698
7	Yorkville	8402
8	Bedford-Stuyvesant	8087
9	Downtown	7331
10	Morningside Heights	7252

MIDNIGHT HOURS:

	LOCATIONS	MIDNIGHT PICKUPCOUNT
1	Williamsburg	8882
2	Jackson Heights	7346
3	Astoria-Long Island City	5839
4	Harlem	5597
5	Fort Green	4861
6	Bedford-Stuyvesant	3130
7	Forest Hills	2199
8	Woodside	1925
9	Park Slope	1912
10	Boerum Hill	1595

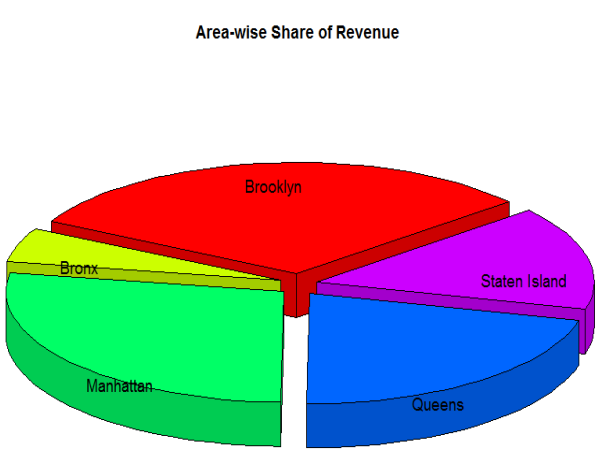
I found out that Harlem seems to be the most popular green-cab pickup destination and williamsburg at night.This list could be recommended to the drivers to understand where the potential cab ride lies in a particular time slot.

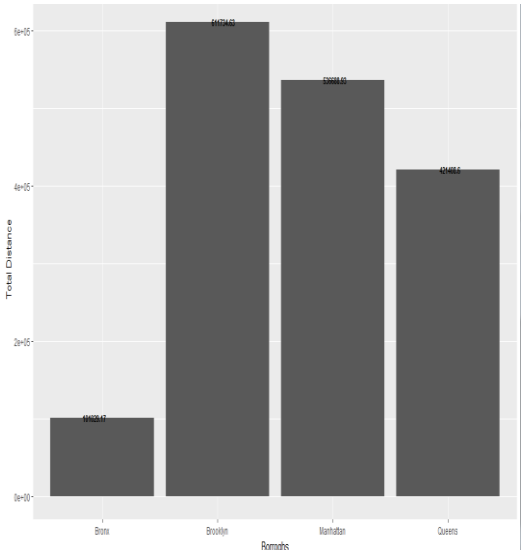
ANALYSIS BASED ON BOROUGHS :

Here I divided the whole data based on individual boroughs to understand which borough factored in more revenue, more number of rides or as simple as who paid more tips.

Code-

```
manhattanData<-subset(mainData,mainData$County=="New York")
bronxData<-subset(mainData,mainData$County=="Bronx")
brooklynData<-subset(mainData,mainData$County=="Kings")
queensData<-subset(mainData,mainData$County=="Queens")
statenIsland<-subset(mainData,mainData$County=="Staten Island ")
```





It seems Brooklyn has the highest revenue for these 14 days. Its interesting because the number of trips are more in Manhattan.But the total distance travelled is more in Brooklyn which increases its revenue.

WEATHER ANALYSIS W.R.T CAB PICKUPS :

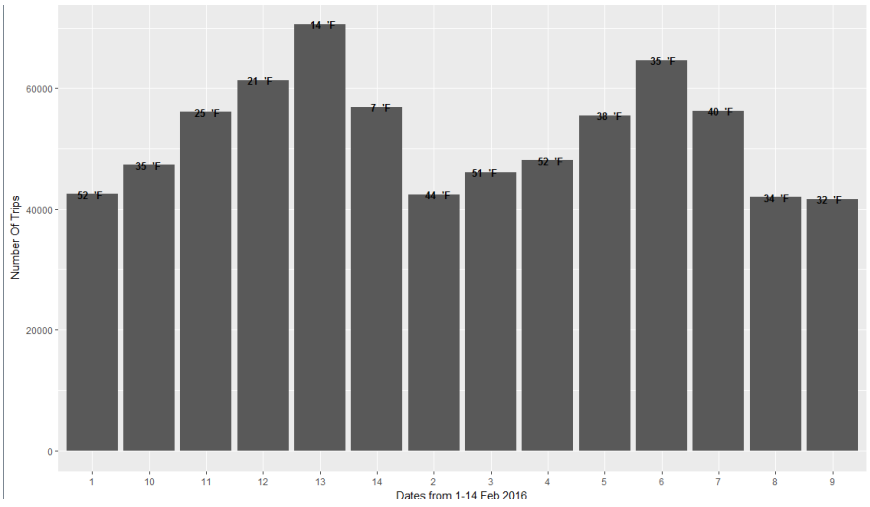
The weather data was taken from w2.weather.gov. I wanted to find out whether there is relation between the average temperature of a day and number of cab rides.

	DATE	MAX	MIN	AVG	DEP	HDD	CDD	WTR	SNW
1	2016-02-01	59	44	52 'F	19	13	0	0.01	0.0
2	2016-02-02	50	38	44 'F	11	21	0	0	0.0
3	2016-02-03	59	42	51 'F	18	14	0	0.73	0.0
4	2016-02-04	59	44	52 'F	18	13	0	T	0.0
5	2016-02-05	44	31	38 'F	4	27	0	0.53	2.5
6	2016-02-07	40	30	35 'F	1	30	0	0	0.0
7	2016-02-07	47	33	40 'F	6	25	0	0	0.0
8	2016-02-08	39	28	34 'F	0	31	0	0.05	0.1
9	2016-02-09	36	27	32 'F	-2	33	0 0	0	0.0
10	2016-02-10	39	31	35 'F	1	30	0	0.01	0.0
11	2016-02-11	31	18	25 'F	-9	40	0	0	0.0
12	2016-02-12	27	15	21 'F	-14	44	0	0	0.0
13	2016-02-13	22	6	14 'F	-21	51	0	0	0.0
14	2016-02-14	15	-1	7 'F	-28	58	0	0	0.0

Code-

```
weather<-read.csv("WeatherData.csv",header=T,sep=",")weather$AVG<-
paste(as.character(weather$AVG)," 'F")
weather<-weather[-c(15,16),]
dateData<-table(unlist(as.Date(mainData$lpep_pickup_datetime,tz="EST")))
dateDataFrame=data.frame(dateData,weather$AVG)
dateDataPlot<-ggplot(data=dateDataFrame, aes(x=as.character(1:14),
y=dateDataFrame$Freq,label=dateDataFrame$weather.AVG)) +
 geom_bar(stat="identity") +geom_text(fontface =
"bold",position=position_jitter(width=.1,height=1),size=3.5)+xlab("Dates from 1-14
Feb,2016")+ylab("Number Of Trips")
```

ANALYSIS :



As it shows weather doesn't affect the daily New-Yorker much if its not in the extreme range.February 13th had the highest number of pickup even though average temperatures were 14'F.