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

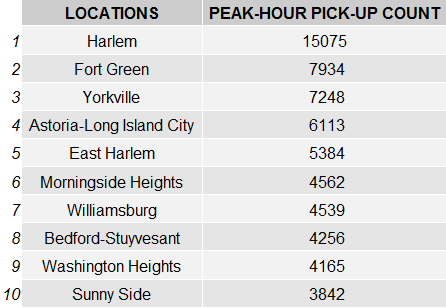
nightData<-subset(mainData,strftime(mainData$lpep\_pickup\_datetime, format="%H:%M:%S")>nightTime & strftime(mainData$lpep\_pickup\_datetime, format="%H:%M:%S")<=beforeMidNightTime)

#Breaking this data into MidNight Slot

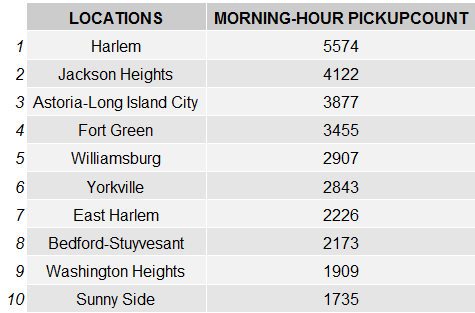
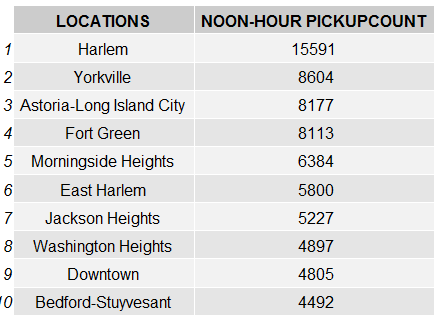
midnightData<-subset(mainData,strftime(mainData$lpep\_pickup\_datetime, format="%H:%M:%S")>=midnightTime & strftime(mainData$lpep\_pickup\_datetime, format="%H:%M:%S")<=morningTime)

# Analysis-

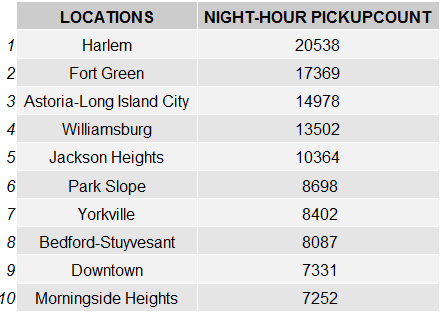
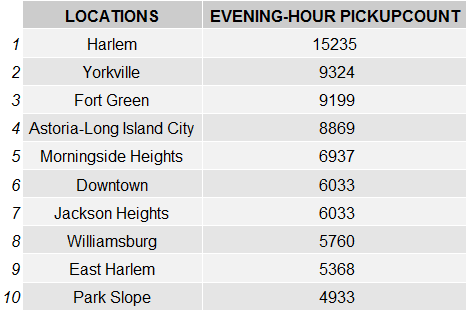
**PEAK-HOURS :**



**MORNING-HOURS: NOON-HOURS:**

**EVENING-HOURS: NIGHT-HOURS:**



**MIDNIGHT HOURS:**



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-

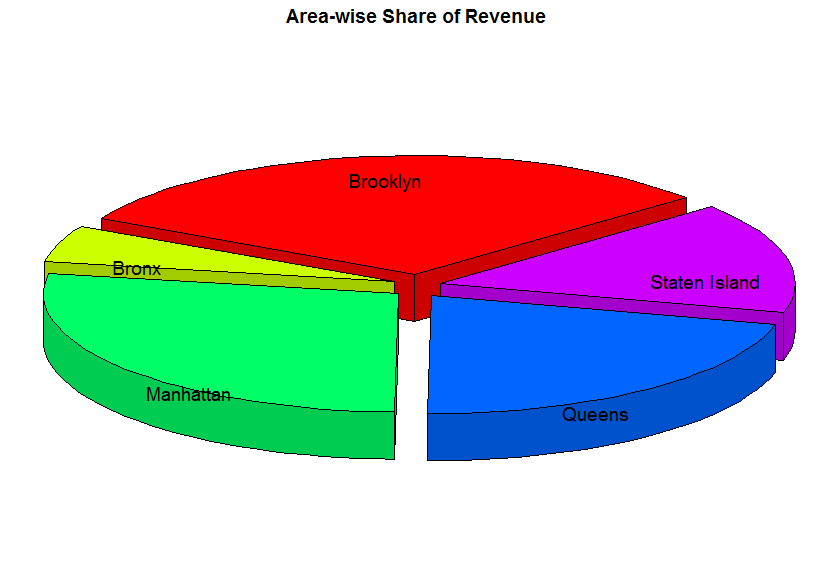
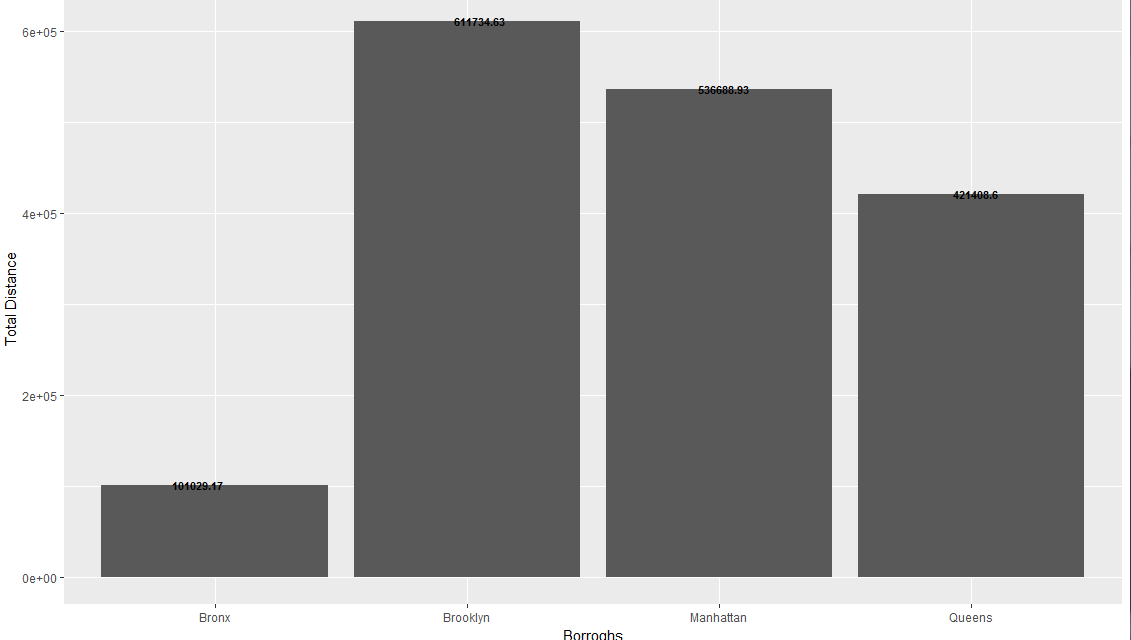
manhattanData<-subset(mainData,mainData$County=="New York")

bronxData<-subset(mainData,mainData$County=="Bronx")

brooklynData<-subset(mainData,mainData$County=="Kings")

queensData<-subset(mainData,mainData$County=="Queens")

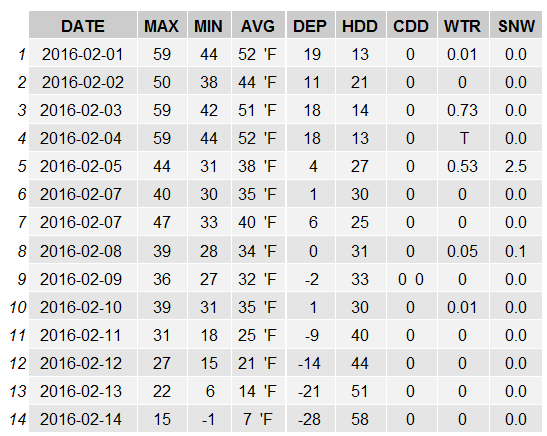
statenIsland<-subset(mainData,mainData$County=="Staten Island ")

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.



Code-

weather<-read.csv("WeatherData.csv",header=T,sep=",")weather$AVG<-paste(as.character(weather$AVG)," 'F")

weather<-weather[-c(15,16),]

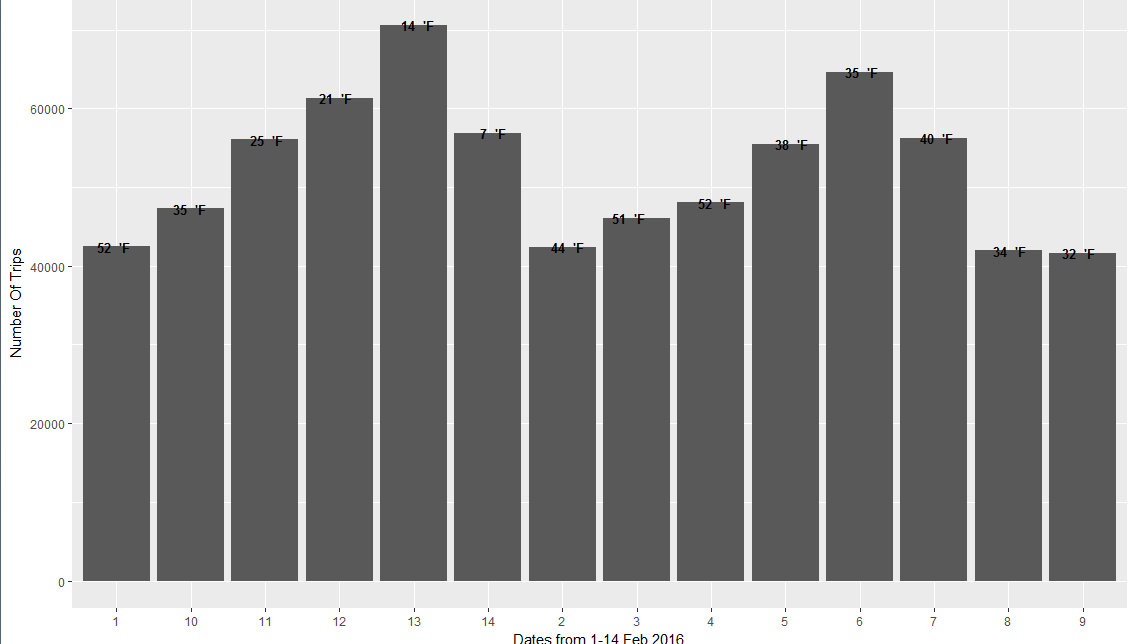
dateData<-table(unlist(as.Date(mainData$lpep\_pickup\_datetime,tz="EST")))

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.