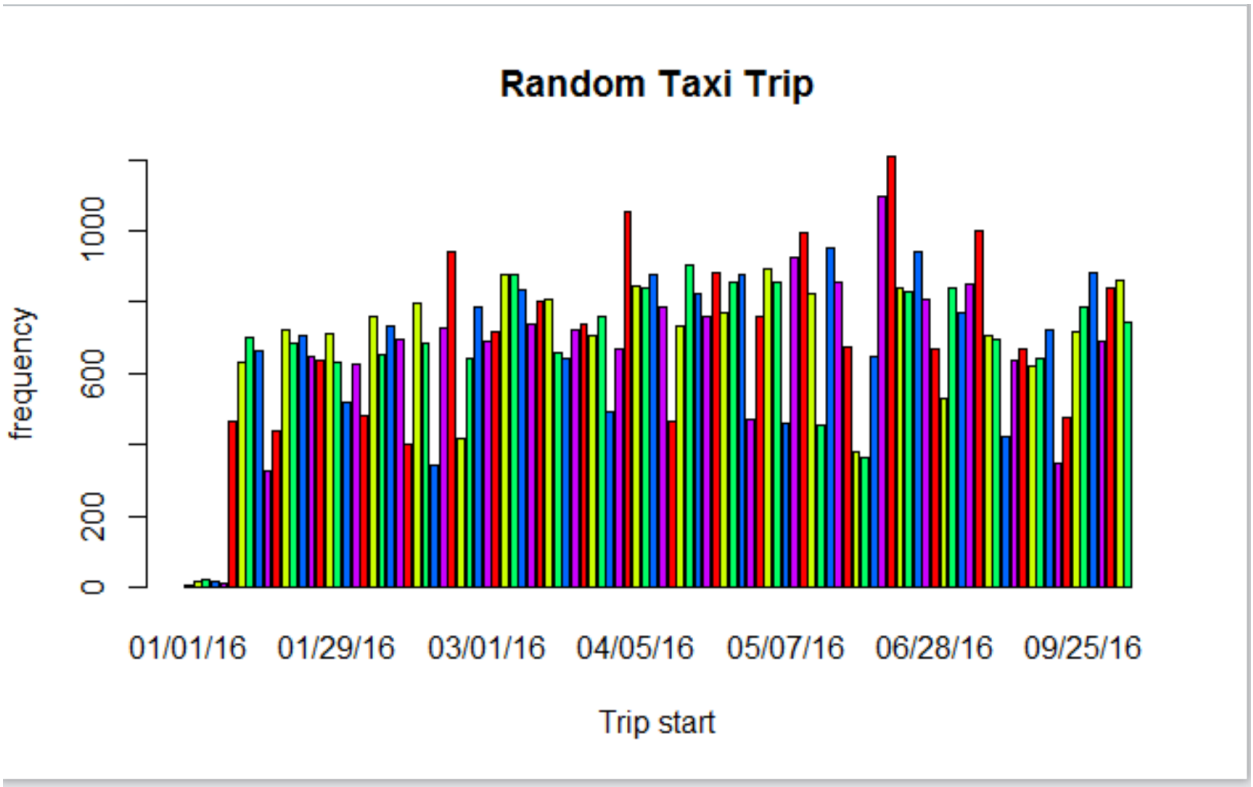
Abinaya Janakan
A20376287

Homework 2

- A study in human behavior
  - Use the Chicago Data Portal's Taxi Trips dataset as a means of studying behavior
- Hypotheses:
  - Cubs fans use taxis more than Sox fans
    - Examine Wrigleyville vs. Bridgeport trips on game days
  - People use taxis more when the weather is bad
    - Examine a sampling of rainy, snowy, very cold, very hot days
  - "Kids" are using taxis less than before
    - Examine Lollapalooza days over the years
- Analysis
  - Study "random" trips to build a sense of "normal"
  - Compare to relevant trips (e.g., date, neighborhood, etc.) to try to find patterns of "abnormal"
  - Assess each hypothesis as well as the credibility of this analysis approach (2 pts)

Abinaya Janakan
A20376287

Studying Random trips:



The above chart shows the random taxi trips

Code:

```
install.packages('strngr')
install.packages('dplyr')
install.packages('sqldf')

library(stringr)
library(dplyr)
library(sqldf)

install.packages("stringr", dependencies=TRUE)
```

Abinaya Janakan
A20376287

```
library(stringr)
require(stringr)



#reading random trips:
randomtrips <-read_csv("C:/Users/abinaya/Desktop/random.csv")
randomplot <- randomtrips %>%
  count(TripStart)
View(randomplot)

randomtrips$TripStart<-strptime(as.character(randomtrips$TripStart),'%m/%d/%y')
plot( randomplot$TripStart , randomplot$n,  main = 'random plot' ,
     xlab = "Start", ylab = "Frequency" ,
     xlim=c(as.POSIXct('2013-04-01', format="%Y-%m-%d"),
          as.POSIXct('2013-09-29', format="%Y-%m-%d")),
     col=rainbow(4) )

counts <- table(randomtrips$TripStart)
barplot(counts, main="Random Taxi Trip",
       xlab="Trip start",ylab="frequency", col=rainbow(5))
```

Abinaya Janakan
A20376287

Hypotheses:1 Cubs fans use taxis more than Sox fans

Examining Wrigleyville vs. Bridgeport trips on game days

Data taken:

Cubs and Sox game days from http://www.baseball-reference.com

Chicago taxi trip dataset for the year 2013 and 2014 for the community areas 60(Bridgeport) and 6(Wrigleyville)

```
install.packages('strngr')

install.packages('dplyr')

install.packages('sqldf')


library(stringr)

library(dplyr)

library(sqldf)


install.packages("stringr", dependencies=TRUE)

library(stringr)

require(stringr)


Taxi60final <- read_csv("C:/Users/abinaya/Desktop/Taxi60final.csv")

TaxiTrip60<- read_csv("C:/Users/Abinaya/Desktop/Taxi60final.csv")

TaxiTrips60

View(TaxiTrips60)


SOX2013 <- read_csv("C:/Users/Abinaya/Desktop/CHW2013Game.csv")

SOX2013

View(SOX2013)


SOX2014 <- read_csv("C:/Users/Abinaya/Desktop/CHW2014Game.csv")
```

```
SOX2014

View(SOX2014)


CUB2013 <- read_csv("C:/Users/Abinaya/Downloads/CHC2013Game.csv")

CUB2013

View(CUB2013)


CUB2014 <- read_csv("C:/Users/Abinaya/Desktop/CHC2014Game.csv")

CUB2014

View(CUB2014)




SOX2013games = sqldf("SELECT  TripID,TripStart, TripSeconds, TripMiles  FROM TaxiTrip60JOIN SOX2013

            ON SOX2013.Date2013 = Taxi_Trips_60r.TripStart")

SOX2013games

View(SOX2013games)


SOX2014games = sqldf("SELECT  TripID,TripStart, TripSeconds, TripMiles  FROM TaxiTrip60JOIN SOX2014

            ON SOX2014.Date2013 = TaxiTrip60.TripStart")

SOX2014games

View(SOX2014games)


CUB2013games = sqldf("SELECT TripID, TripStart, TripSeconds, TripMiles  FROM TaxiTrip60JOIN
CUB2013

            ON CUB2013.Date2013 = TaxiTrips60.TripStart")

CUB2013games

View(CUB2013games)
```

```
CUB2014games = sqldf("SELECT TripID, TripStart, TripSeconds, TripMiles  FROM TaxiTrip60JOIN
CUB2014
                ON CUB2014.Date2013 = TaxiTrip60.TripStart")
CUB2014games
View(CUB2014games)


SOXplot <- SOX2013games %>%
  count(TripStart)
View(SOXplot)


SOXplot_1 <- SOX2014games %>%
  count(TripStart)
View(SOXplot)



CUBplot <- CUB2013games %>%
  count(TripStart)
View(CUBplot)


CUBplot_1 <- CUB2014games %>%
  count(TripStart)
View(CUBplot)


SOXplot$TripStart<-strptime(as.character(SOXplot$TripStart),'%m/%d/%y')
plot( SOXplot$TripStart , SOXplot$n,  main = 'SOX2013 plot' ,
     xlab = "Start", ylab = "Frequency" ,
     xlim=c(as.POSIXct('2013-04-01', format="%Y-%m-%d"),
```

```
        as.POSIXct('2013-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )
counts <- table(SOX2013games$TripStart)
barplot(counts, main="SOX game 2013 Taxi Trip",
      xlab="Trip start", col=rainbow(4))


SOXplot_1$TripStart<-strptime(as.character(SOXplot_1$TripStart),'%m/%d/%y')
plot( SOXplot_1$TripStart , SOXplot_1$n,  main = 'SOX2014 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2014-04-01', format="%Y-%m-%d"),
        as.POSIXct('2014-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )
counts <- table(SOX2014games$TripStart)
barplot(counts, main="SOX game 2014 Taxi Trip",
      xlab="Trip start", col=rainbow(4))


CUBplot$TripStart<-strptime(as.character(CUBplot$TripStart),'%m/%d/%y')
plot( CUBplot$TripStart , CUBplot$n,  main = 'CUB2013 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2013-04-09', format="%Y-%m-%d"),
        as.POSIXct('2013-09-25', format="%Y-%m-%d")),
    col=rainbow(4) )
counts <- table(CUB2013games$TripStart)
barplot(counts, main=" CUB game 2013 Taxi Trips",
      xlab="Trip start", col=rainbow(4))
```

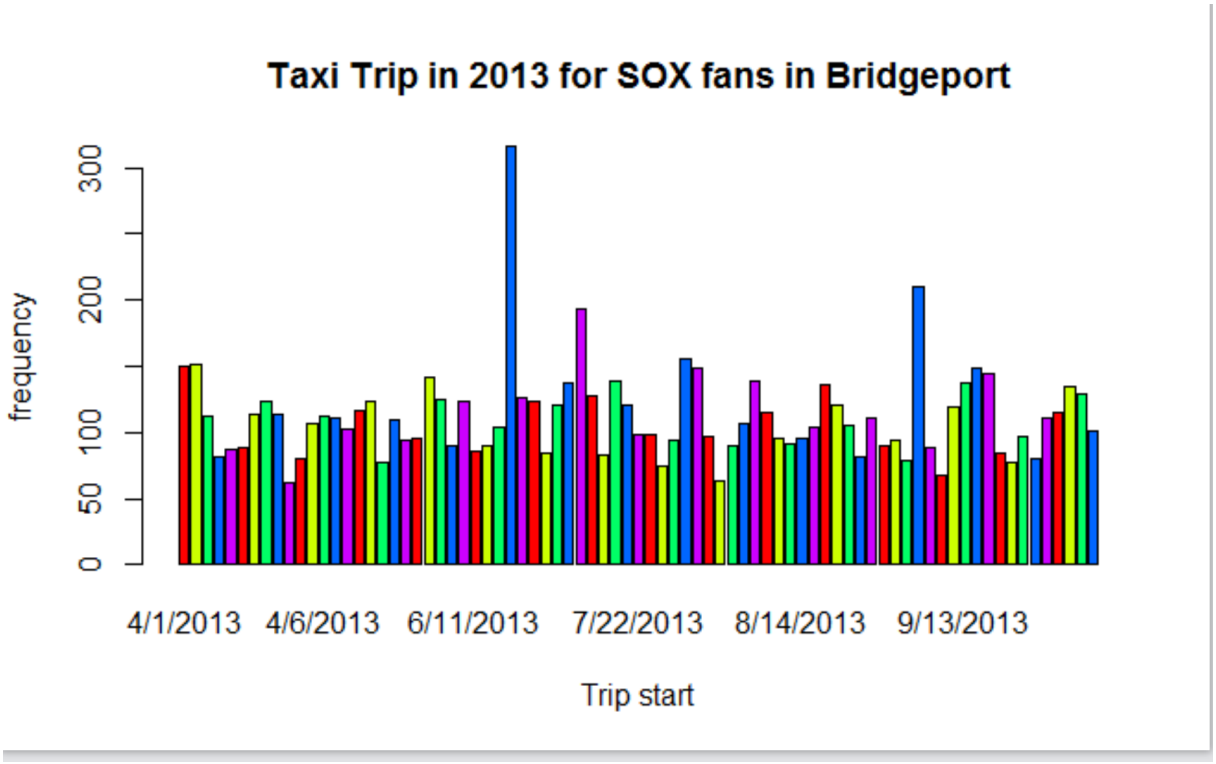
```
CUBplot_1$TripStart<-strptime(as.character(CUBplot_1$TripStart),'%m/%d/%y')
plot( CUBplot_1$TripStart , CUBplot_1$n,  main = 'CUB2014 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2014-04-09', format="%Y-%m-%d"),
        as.POSIXct('2014-09-25', format="%Y-%m-%d")),
    col=rainbow(4) )
counts <- table(CUB2014games$TripStart)
barplot(counts, main=" CUB game 2014 Taxi Trips",
     xlab="Trip start", col=rainbow(4))


Taxi6final <- read_csv("C:/Users/abinaya/Desktop/Taxi6final.csv")
Taxi_Trips_6r <- read_csv("C:/Users/Abinaya/Desktop/Taxi6final.csv")
Taxi_Trips_6r
View(Taxi_Trips_6r)

SOX2013games_6 = sqldf("SELECT  TripID,TripStart, TripSeconds, TripMiles  FROM Taxi_Trips_6r JOIN
SOX2013
            ON SOX2013.Date2013 = Taxi_Trips_6r.TripStart")
SOX2014games_6 = sqldf("SELECT  TripID,TripStart, TripSeconds, TripMiles  FROM Taxi_Trips_6r JOIN
SOX2014
            ON SOX2014.Date2013 = Taxi_Trips_60r.TripStart")
SOX2014games
View(SOX2014games)

SOX2013game_6
View(SOX2013games_6)
```

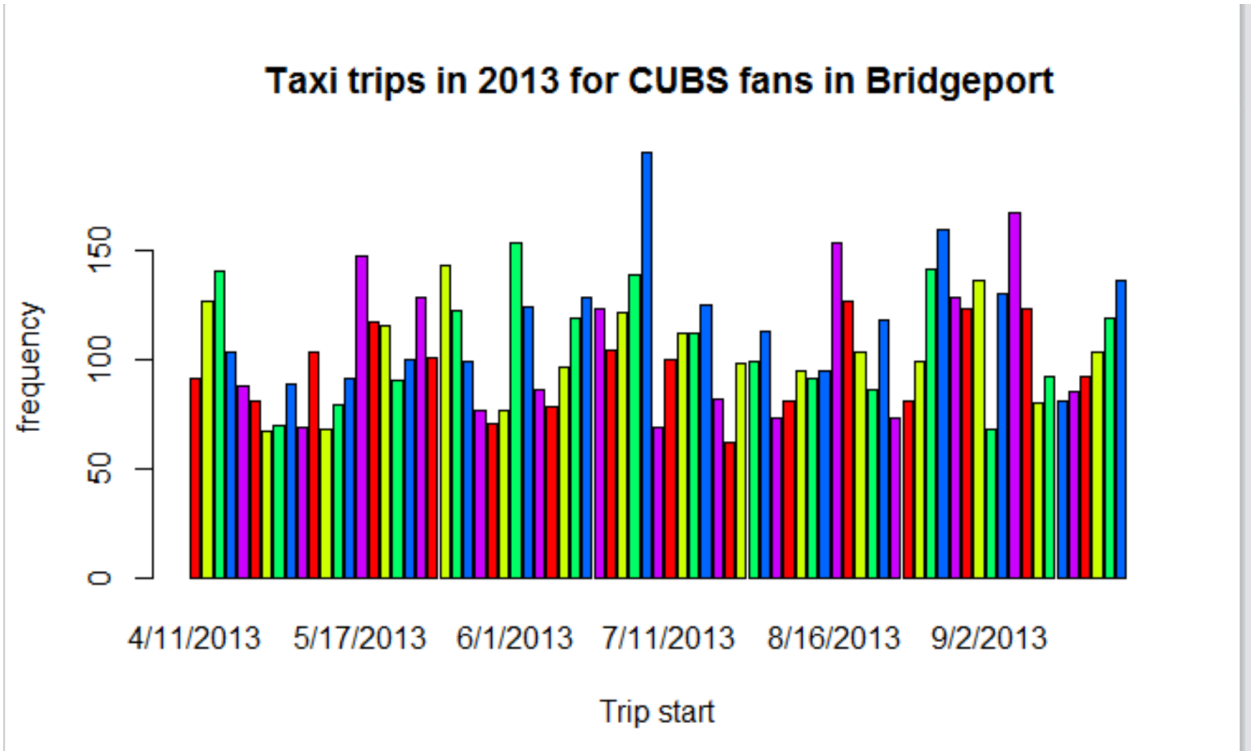
```
CUB2013games_6 = sqldf("SELECT TripID, TripStart, TripSeconds, TripMiles  FROM Taxi_Trips_6r JOIN CUB2013
                       ON CUB2013.Date2013 = Taxi_Trips_6r.TripStart")
CUB2013games_6
View(CUB2013games_6)

CUB2014games_6 = sqldf("SELECT TripID, TripStart, TripSeconds, TripMiles  FROM Taxi_Trips_6r JOIN CUB2014
                       ON CUB2014.Date2013 = Taxi_Trips_6r.TripStart")
CUB2014games
View(CUB2014games)



SOXplot_6 <- SOX2013games_6 %>%
  count(TripStart)
View(SOXplot_6)

SOXplot_1_6 <- SOX2014games_6 %>%
  count(TripStart)
View(SOXplot)

CUBplot_6 <- CUB2013games_6 %>%
  count(TripStart)
View(CUBplot_6)

CUBplot_1_6 <- CUB2014games_6 %>%
  count(TripStart)
View(CUBplot_6)
```

```
SOXplot_6$TripStart<-strptime(as.character(SOXplot_6$TripStart),'%m/%d/%y')
plot( SOXplot_6$TripStart , SOXplot_6$n,  main = 'SOX2013 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2013-04-01', format="%Y-%m-%d"),
        as.POSIXct('2013-09-29', format="%Y-%m-%d")),
    col=rainbow(5) )
counts <- table(SOX2013games_6$TripStart)
barplot(counts, main="SOX game 2013 Taxi Trip",
      xlab="Trip start", col=rainbow(5))

SOXplot_1_6$TripStart<-strptime(as.character(SOXplot_6$TripStart),'%m/%d/%y')
plot( SOXplot_1_6$TripStart , SOXplot_1_6$n,  main = 'SOX2014 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2014-04-01', format="%Y-%m-%d"),
        as.POSIXct('2014-09-29', format="%Y-%m-%d")),
    col=rainbow(5) )
counts <- table(SOX2014games_6$TripStart)
barplot(counts, main="SOX game 2014 Taxi Trip",
      xlab="Trip start", col=rainbow(5))


CUBplot_6$TripStart<-strptime(as.character(CUBplot$TripStart),'%m/%d/%y')
plot( CUBplot_6$TripStart , CUBplot$n,  main = 'CUB2013 plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2013-04-09', format="%Y-%m-%d"),
        as.POSIXct('2013-09-25', format="%Y-%m-%d")),
    col=rainbow(4) )
```

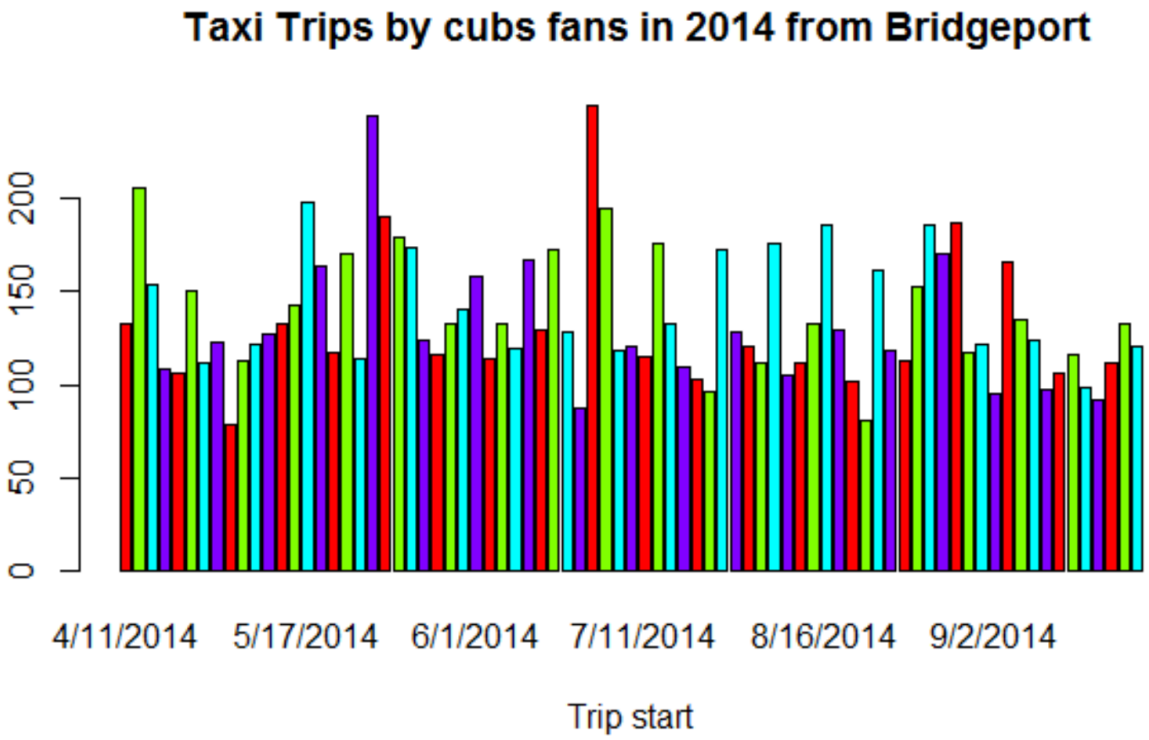
```
counts <- table(CUB2013games_6$TripStart)
barplot(counts, main=" CUB game 2013 Taxi Trips",
       xlab="Trip start")


CUBplot_1_6$TripStart<-strptime(as.character(CUBplot_1_6$TripStart),'%m/%d/%y')
plot( CUBplot_1_6$TripStart , CUBplot_1_6$n,  main = 'CUB2014 plot' ,
     xlab = "Start", ylab = "Frequency" ,
     xlim=c(as.POSIXct('2014-04-09', format="%Y-%m-%d"),
          as.POSIXct('2014-09-25', format="%Y-%m-%d")),
     col=rainbow(5) )
counts <- table(CUB2014games_6$TripStart)
barplot(counts, main=" CUB game 2014 Taxi Trips",
       xlab="Trip start", col=rainbow(5))
```

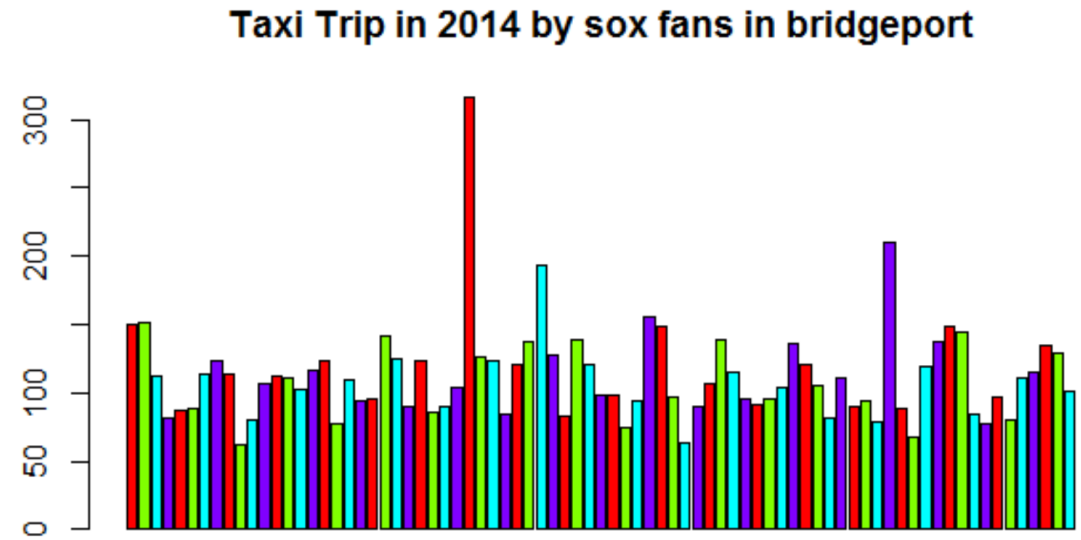
Abinaya Janakan
A20376287



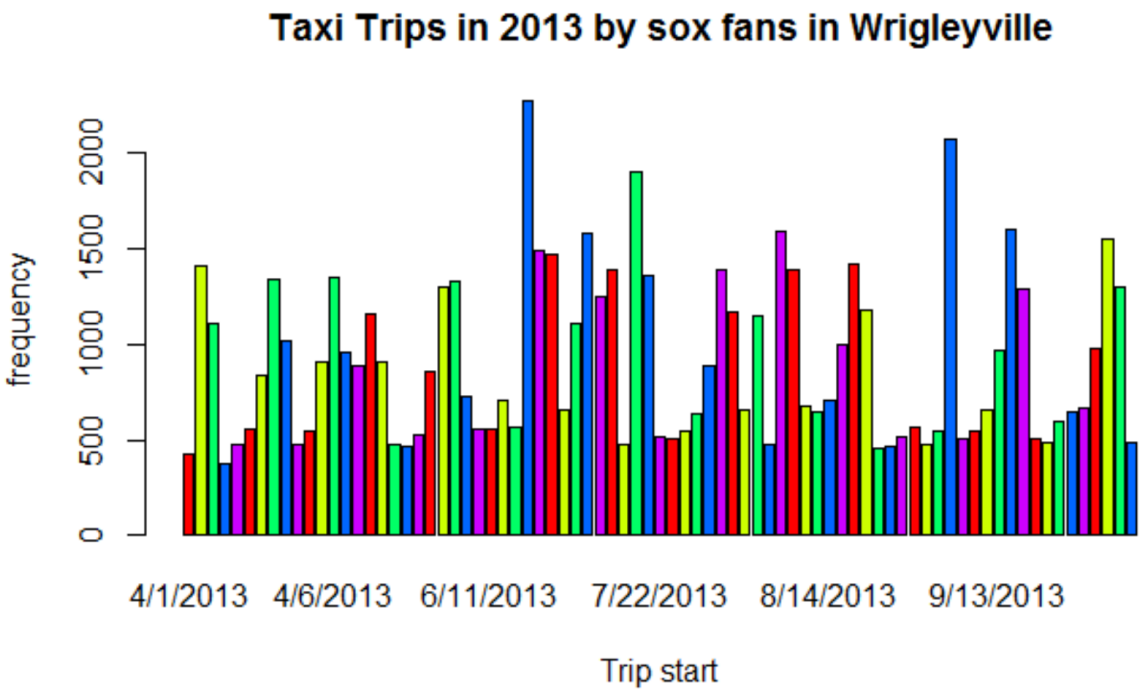


Abinaya Janakan
A20376287

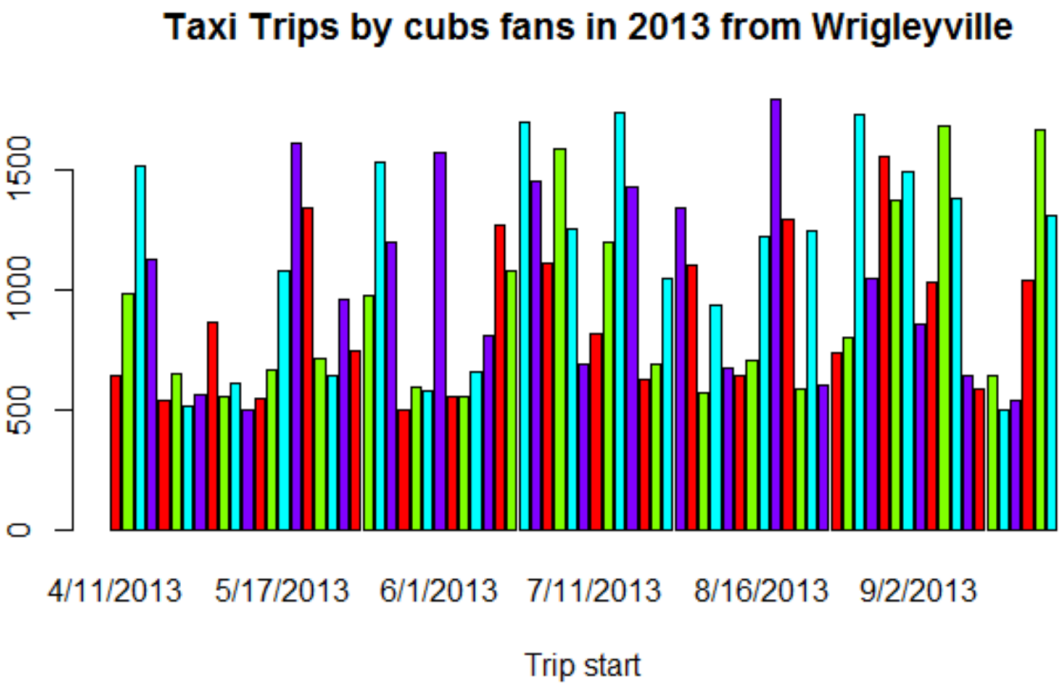


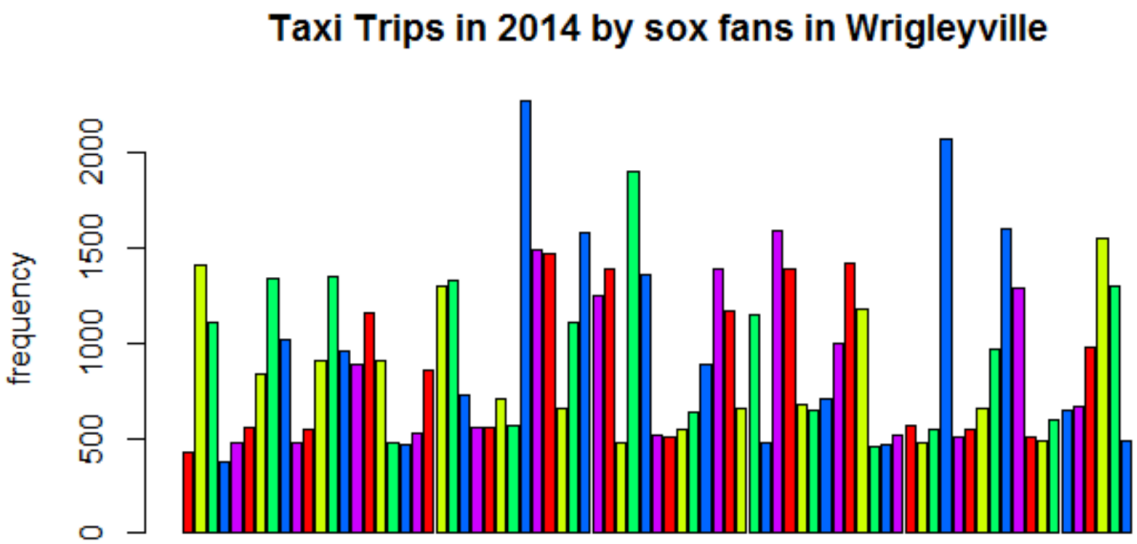


Abinaya Janakan
A20376287

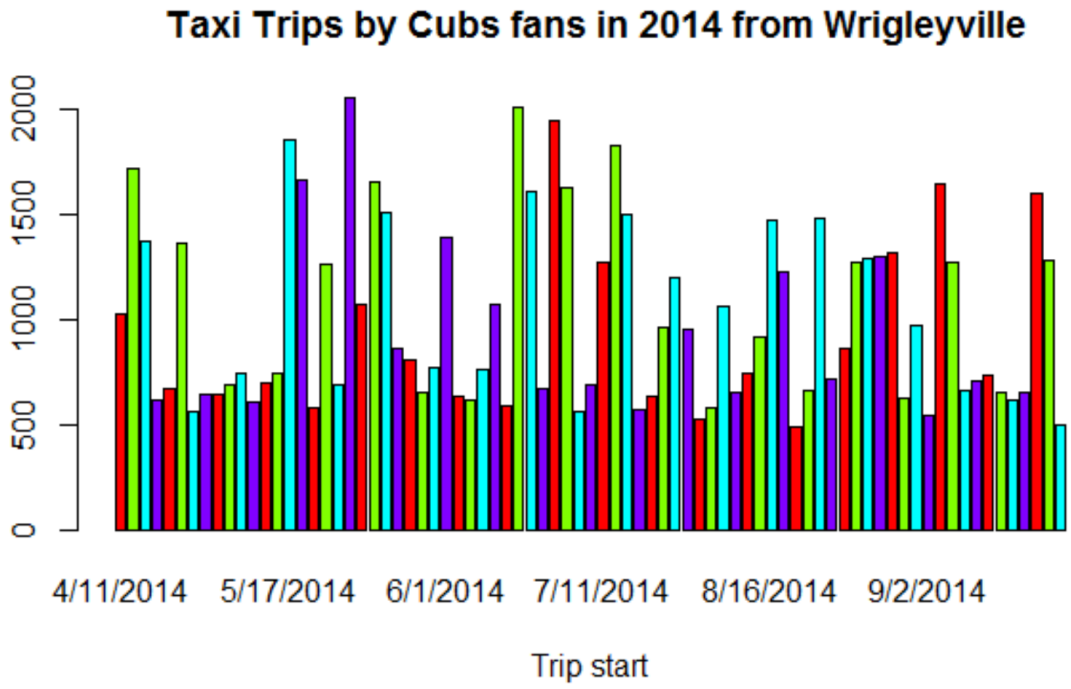


Abinaya Janakan
A20376287





Abinaya Janakan
A20376287



Analysis:

Cubs fans Vs Sox fans:

After analyzing the cubs and Sox fans usage of Taxi for two years we could find that the cubs fans indeed use Taxi more than the Sox fans.Though the number of Taxi trips taken by the Cubs fans is comparatively higher we can't conclude that Cubs fans use Taxi more than Sox fans because Cubs is a much popular team than Sox and hence the number of people who watches the Cubs matches are comparatively higher , which ultimately increases the taxi trips on the Cubs game days.

Moreover we are considering the taxi trips of Bridgeport and Wrigleyville on game days . Wrigley Field a baseball stadium which is also the home of Chicago Cubs team is located nearer to Wrigleyville.So the number of people using Taxi's will ultimately be higher in Cubs game days.

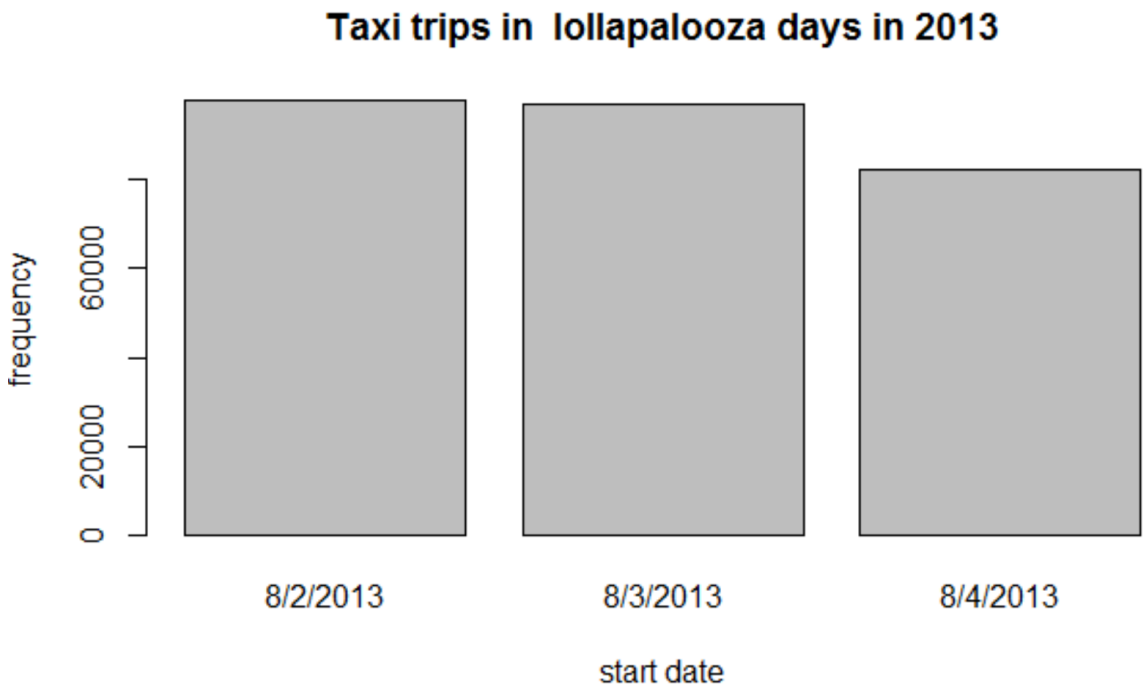
Abinaya Janakan
A20376287

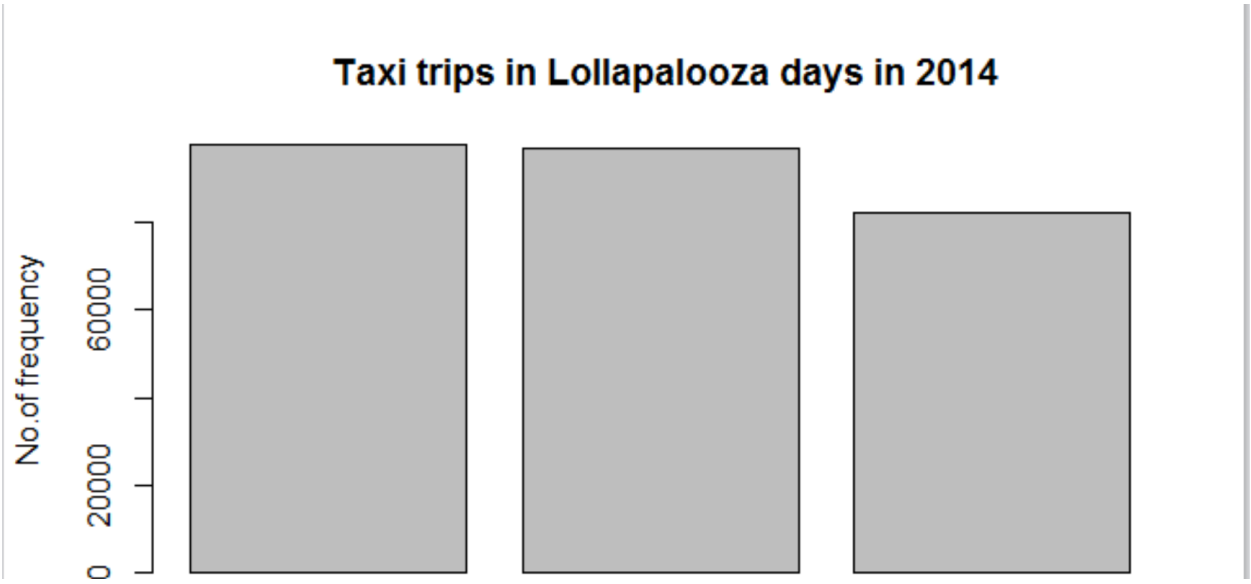
Hypothesis 2:

```
Taxi_Trips_2013Ldays<- read_csv("C:/Users/Abinaya/Desktop/Taxi_ldays.csv")

Taxi_Trips_2013Ldays

View(Taxi_Trips_2013Ldays)

#lollapalooza  days in 2013

ldays2013 <- read_csv("C:/Users/Abinaya/Desktop/lollapalooza2013.csv")

ldays2013

View(ldays2013)

# number of trips per day in 2013 lallapalooza

ldays2013tripsPlot <- ldays2013trips %>%

  count(TripStart_l)

View(ldays2013tripsPlot)

counts <- table(ldays2013trips$TripStart)

barplot(counts, main=" Lollapalooza  days 2013 trips",

        xlab="Trip start",ylab=”frequency” col=rainbow(4))



l1 <- ggplot(ldays2013tripsPlot, aes(x=ldays2013tripsPlot$TripStart_l, y=ldays2013tripsPlot$n, xlab="Trip
start")) +

  geom_bar(stat="identity") +

  xlab("Lollapalooza 2013 ") +

  ylab("Frequency of the taxi trips (n)") +

  ggtitle("2013 Taxi Trips on Lollapalooza Days") +

  geom_text(aes(label=ldays2013tripsPlot$n), colour="white")


Taxi_Trips_2014Ldays<- read_csv("C:/Users/Abinaya/Desktop/Taxi60final.csv")
```

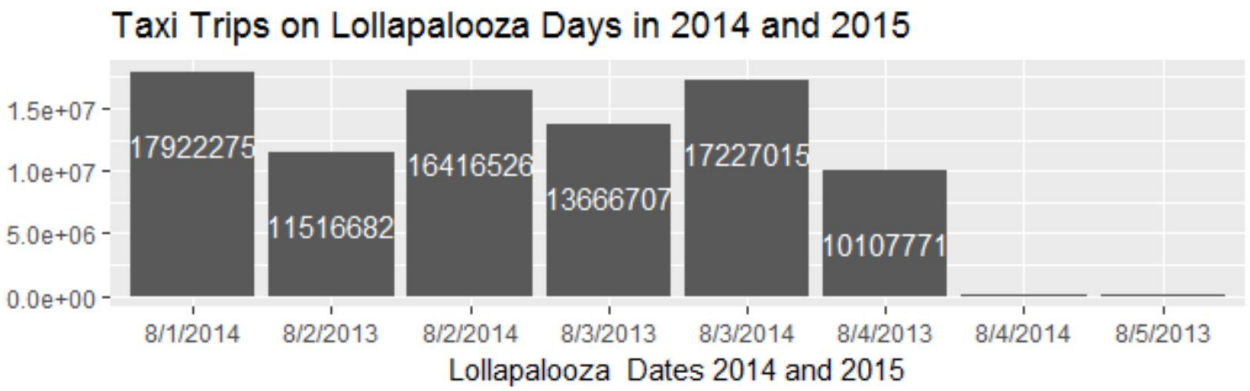
```
Taxi_Trips_2014Ldays

View(Taxi_Trips_2014Ldays)


ldays2014 <- read_csv("C:/Users/Abinaya/Desktop/Taxi_ldays.csv")

View(ldays2014)

#lollapalooza in 2014 with the taxi trips

ldays2014trips= sqldf("SELECT TripID, TripStart, TripSeconds, TripMiles

                      FROM Taxi_Trips_2014LD JOIN ldays2014

                      ON ldays2014.TripStart_l = Taxi_Trips_2014LD.TripStart")

ldays2014trips

View(ldays2014trips)

ldays2014tripsPlot <- ldays2014trips %>%

  count(TripStart)

View(ldays2014tripsPlot)

Taxi2014Lplot <- barplot(counts, main=" Lollapalooza 2014 trips",

                  xlab="Trip start", ylab="No.of Trips/day",legend = rownames(counts),  col=rainbow(4))


l2 <- ggplot(ldays2014tripsPlot, aes(x=ldays2014tripsPlot$TripStart, y=ldays2014tripsPlot$n, xlab="Trip
start")) +

  geom_bar(stat="identity") +

  xlab("Lollapalooza  Dates in 2014") +

  ylab("Frequency of the taxi trips (n)") +

  ggtitle("2014 Taxi Trips on Lollapalooza Days") +

  geom_text(aes(label=ldays2014tripsPlot$n), vjust=3.5, colour="white")


grid.arrange(l1, l2)
```

Abinaya Janakan
A20376287





Abinaya Janakan
A20376287



As seen from the bar graph above the number of people using taxi in lollapalooza days is higher in 2014 than 2013.As well as the number of taxi users also goes down after the lollapalooza days.So, the hypothesis that people use less taxi than before is not true.Since we are considering the taxi trips throught Chicago the analysis may not be accurate.However the analysis could be improved if we consider the taxi trips around the venue of the lollapalooza event

Weather data:

Hypothesis 3:People use taxis more when the weather is bad

Data:

Lets take data for the community area 76 and analyze the bad weather days in 2016.

Code:

```
library(readr)

weather <- read_csv("C:/Users/abinaya/Desktop/weather.csv")


weather <- data.frame(weather)

View(weather)


snow <- subset(weather, snow_depth>1)

snow_1 <- data.frame(snow)

View(snow_1)


snowfall <- subset(weather, snowfall>0.5)
```

```
snowfall <- data.frame(snowfall)
View(snowfall)

rainydays <- subset(weather, precipitation>0.5)
rainydays <- data.frame(rainydays)
View(rainydays)



hotdays <- subset(weather, max_temperature>60)
hotdays <- data.frame(hotdays)
View(hotdays)

colddays <- subset(weather, min_temperature<10)
colddays <- data.frame(colddays)
View(colddays)

windydays <- subset(weather, average_wind_speed>10)
windydays <- data.frame(windydays)
View(windydays)

weatherplot_rainydays$date<-strptime(as.character(weatherplot_rainydays$date),'%m/%d/%y')
plot( weatherplot_rainydays$date , weatherplot_rainydays$n,  main = 'trips on rainy days plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
         as.POSIXct('2016-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )

weathertrips_rainydays = merge(Taxiweather_1,rainydays,by="date",all=FALSE)
```

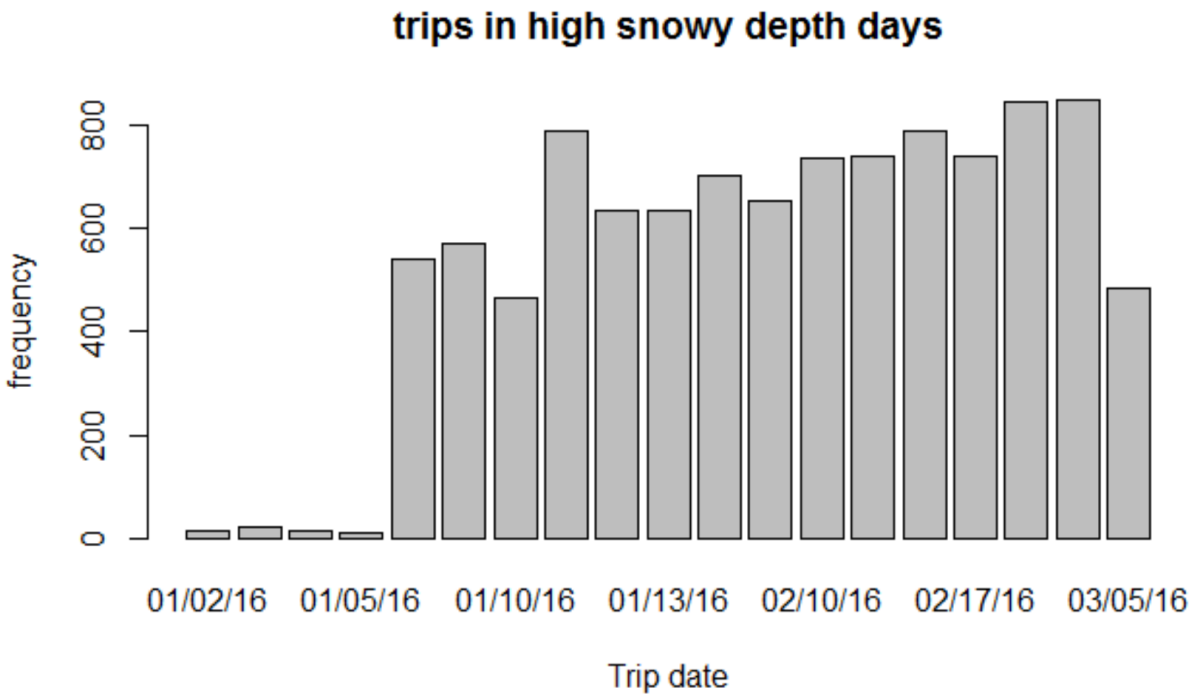
```
weathertrips_rainydays
View(weathertrips_rainydays)
head(weathertrips_rainydays)

counts_5 <- table(weathertrips_windydays$date)
barplot(counts_5, main="trips in rainy  days",
      xlab="Trip start",ylab="frequency")



library(readr)
weather_trips <- read_csv("C:/Users/abinaya/Desktop/weather_trips.csv")
View(weather_trips)

Taxiweather <- read_csv("C:/Users/abinaya/Desktop/weather_trips.csv")
Taxiweather_1 <- data.frame(Taxiweather)
Taxiweather_1
View(Taxiweather_1)



#high snow depth days m
weathertrips_snow = merge(Taxiweather_1,snow_1,by="date",all=FALSE)
weathertrips_snow
View(weathertrips_snow)
head(weathertrips_snow)



weathertrips_snowfall = merge(Taxiweather_1,snowfall,by="date",all=FALSE)
```

```
weathertrips_snowfall
View(weathertrips_snowfall)
head(weathertrips_snowfall)

weathertrips_hotdays = merge(Taxiweather_1,hotdays,by="date",all=FALSE)
weathertrips_hotdays
View(weathertrips_hotdays)
head(weathertrips_hotdays)

weathertrips_colddays = merge(Taxiweather_1,colddays,by="date",all=FALSE)
weathertrips_colddays
View(weathertrips_colddays)
head(weathertrips_colddays)

weathertrips_windydays = merge(Taxiweather_1,windydays,by="date",all=FALSE)
weathertrips_windydays
View(weathertrips_windydays)
head(weathertrips_windydays)

install.packages('plyr')
library(plyr)
weatherplot_snow <- count(weathertrips_snow,'date')
View(weatherplot_snow)

# weather plot snow normalising date

weatherplot_snow$date<-strptime(as.character(weatherplot_snow$date),'%m/%d/%y')
plot( weatherplot_snow$date , weatherplot_snow$n,  main = 'Snow depth plot' ,
```

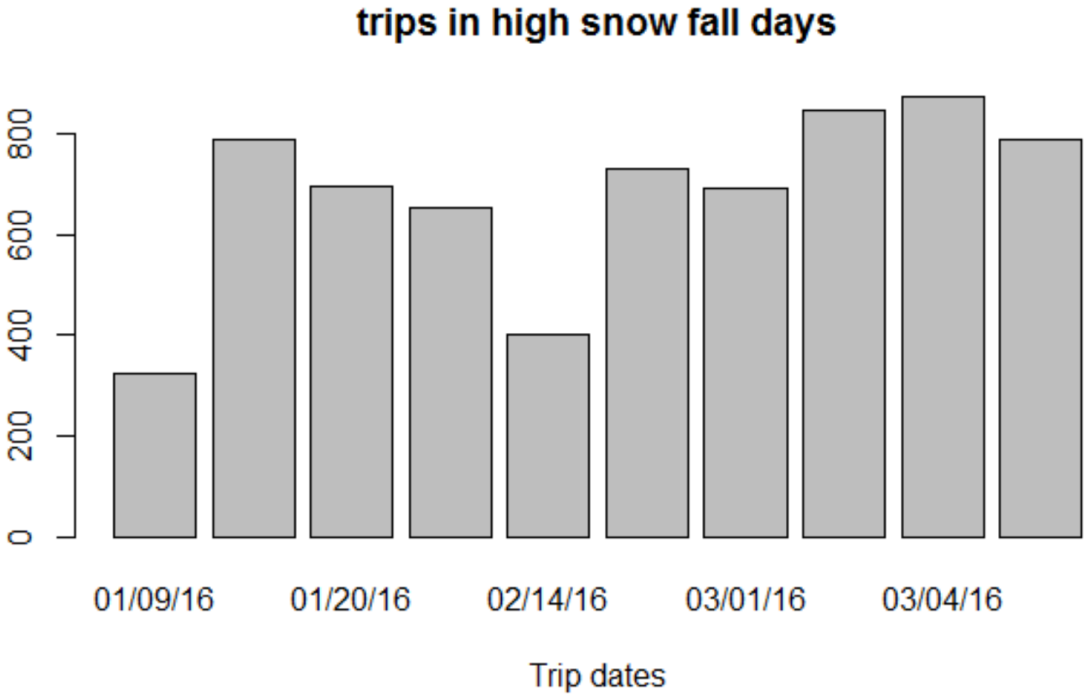
```
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
         as.POSIXct('2016-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )

counts <- table(weathertrips_snow$date)
barplot(counts, main="trips in high snowy depth days",
      xlab="Trip start", col=rainbow(4))



weatherplot_snowfall <- count(weathertrips_snowfall,'date')
View(weatherplot_snowfall)
#weather plot in snowfall days
weatherplot_snowfall$date<-strptime(as.character(weatherplot_snowfall$date),'%m/%d/%y')
plot( weatherplot_snowfall$date , weatherplot_snowfall$n,  main = 'Snow fall plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
         as.POSIXct('2016-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )

counts <- table(weathertrips_snowfall$date)
barplot(counts, main="trips in high snowy depth days",
      xlab="Trip start", col=rainbow(4))

# hot days
weatherplot_hot <- count(weathertrips_hotdays,'date')
View(weatherplot_hot)
```

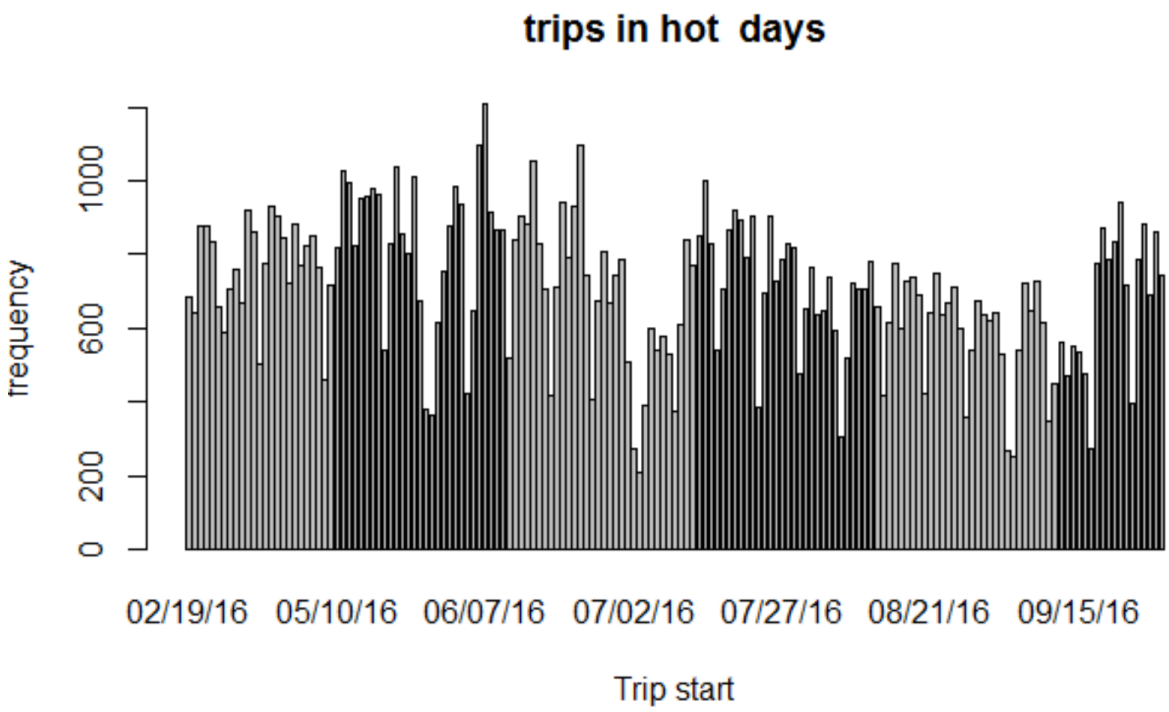
```
weatherplot_hot$date<-strptime(as.character(weatherplot_hot$date),'%m/%d/%y')
plot( weatherplot_hot$date , weatherplot_hot$n,  main = 'trips on hot days plot' ,
     xlab = "Start", ylab = "Frequency" ,
     xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
          as.POSIXct('2016-09-29', format="%Y-%m-%d")),
     col=rainbow(4) )

counts <- table(weathertrips_hotdays$date)
barplot(counts, main="trips in hot  days",
       xlab="Trip start", col=rainbow(4))



#cold days
weatherplot_cold <- count(weathertrips_colddays,'date')
View(weatherplot_cold)

weatherplot_cold$date<-strptime(as.character(weatherplot_cold$date),'%m/%d/%y')
plot( weatherplot_cold$date , weatherplot_cold$n,  main = 'trips on cold days plot' ,
     xlab = "Start", ylab = "Frequency" ,
     xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
          as.POSIXct('2016-09-29', format="%Y-%m-%d")),
     col=rainbow(4) )

counts <- table(weathertrips_colddays$date)
barplot(counts, main="trips in cold  days",
       xlab="Trip start", col=rainbow(4))
```

```
#windy days
weatherplot_windy <- count(weathertrips_windydays,'date')
View(weatherplot_windy)

weatherplot_windy$date<-strptime(as.character(weatherplot_windy$date),'%m/%d/%y')
plot( weatherplot_windy$date , weatherplot_windy$n,  main = 'trips on windy days plot' ,
    xlab = "Start", ylab = "Frequency" ,
    xlim=c(as.POSIXct('2016-04-01', format="%Y-%m-%d"),
         as.POSIXct('2016-09-29', format="%Y-%m-%d")),
    col=rainbow(4) )

counts <- table(weathertrips_windydays$date)
barplot(counts, main="trips in windy  days",
      xlab="Trip start", col=rainbow(4))
```

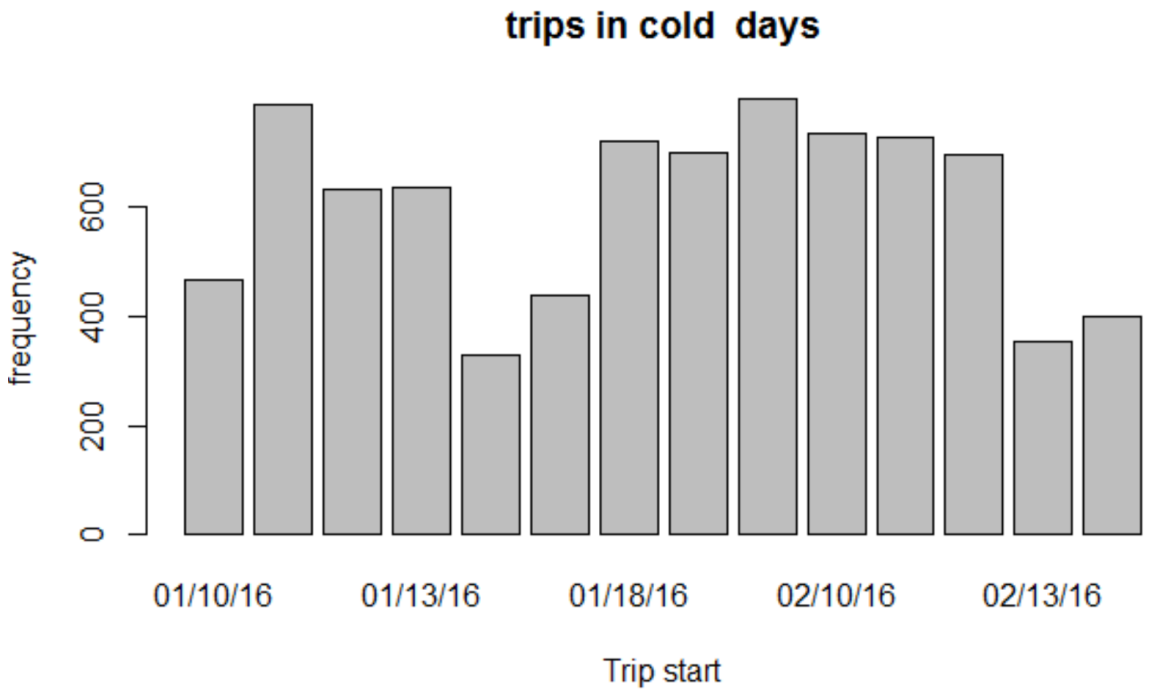


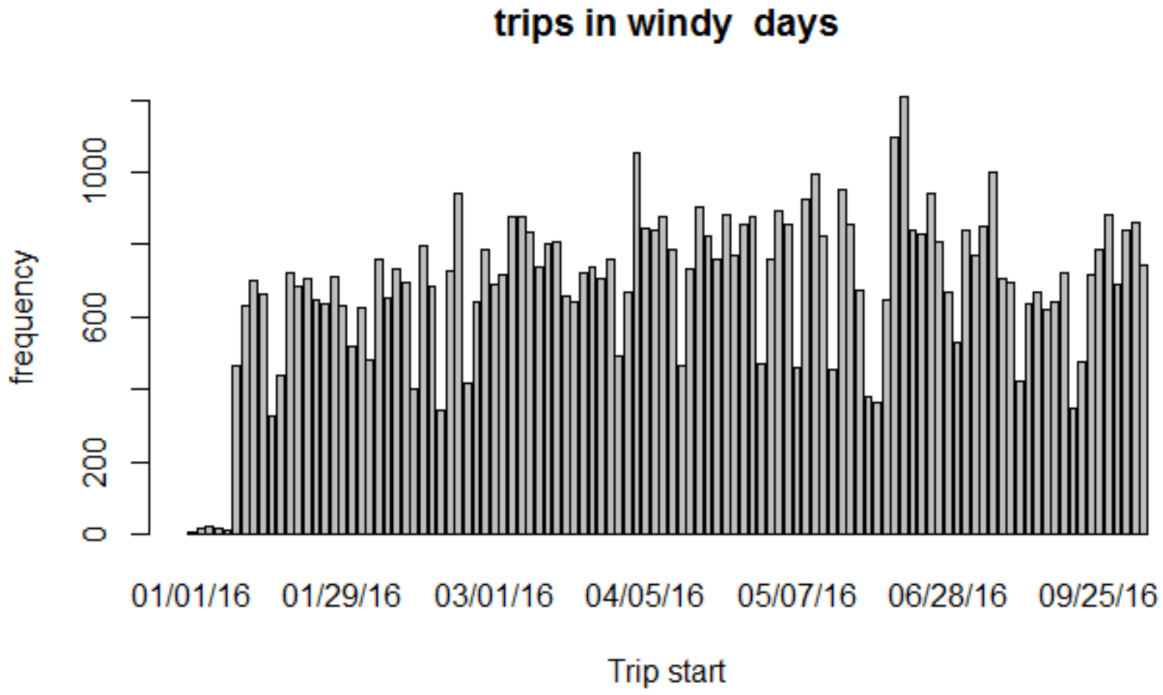
Abinaya Janakan
A20376287



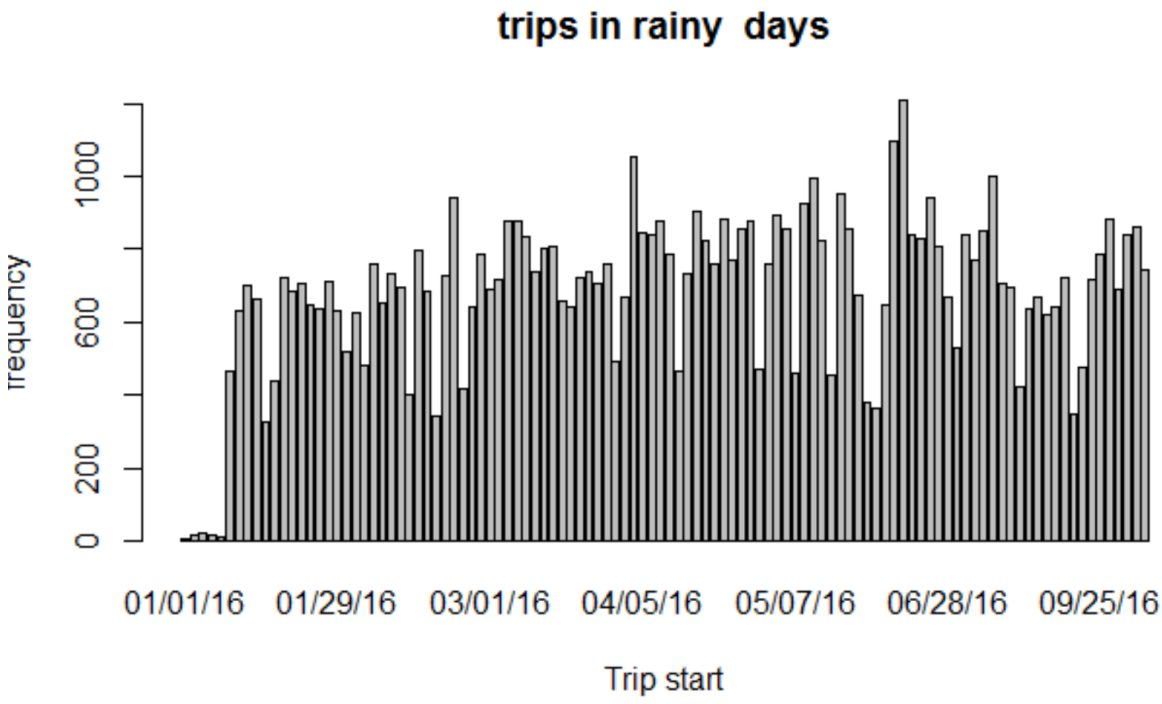


Abinaya Janakan
A20376287





Abinaya Janakan
A20376287



The taxi trips are not generally high in bad weather days.It is higher in hot days and high wind days and rainy days.So, the hypothesis that the taxi usage is higher in all bad weather days is not completely true.Since we have taken only a particular community area and analyzed its weather for only one year the result may not be accurate.