Divvy Bikes Dataset Analysis using R

Data Analysis Report

Author: Janardhan Bonu

Date: May-13-2015 .

Submitted to Professor

Anil K. Maheshwari

**Maharishi University of Management**

Table of Contents

Introduction: 3

Summary of Divvy data 3

Number of Trips: 3

Trip duration Max : 3

Number of Users: 3

Stations Information:- 4

Top 5 stations Start trip 4

Top 5 stations End trip 4

Reports: 5

Trip per hour of a day 5

Customer Vs Subscribers 6

Trips per week day 7

Trips per month 8

Trips per season: 9

Divvy Bikes data analysis 10

Busy hour of Customers and Subscribers 10

Hour x dayofweek 11

User type x day of week 12

User type x season 12

# Introduction:

Divvy bikes is a bike share program in Chicago, USA. The public can either subscribe for year or use with day pass. There are many Divvy bike stations in Chicago for the commuters. There is transaction log generated by the Divvy bike application. The divvy bikes data analysis is done with R language

The analysis is summarized in 3 different forms.

* Summary
* Reports
* Analysis

# Summary of Divvy data

|  |  |
| --- | --- |
| Column Name | Description |
| trip\_id | Trip Id – unique |
| starttime | Trip start time |
| stoptime | Trip end time |
| bikeid | Bike Id |
| Tripduration | Time taken complete single trip |
| from\_station\_id | From station Id |
| from\_station\_name | From station name |
| to\_station\_id | Destination station Id |
| to\_station\_name | Destination station name |
| usertype | Customer or Subscriber |
| gender | Gender of Subscriber |
| birthyear | Birth Year of Subscriber |
| Weekday\* | Week day of the trip |
| Month\* | Month of the trip |
| Season\* | Season of the trip |
| Hour\* | Hour of the trip |
| Station pair\* | Station Source and Destination |

\* The green marked columns are newly added at the time of data analysis.

## Number of Trips:

2,454,634 (Nearly 2.5 million trips)

## Trip duration Max :

86392

## Number of Users:

Customers: 791240

Subscribers: 1663394

## Stations Information:-

### Top 5 stations Start trip

|  |  |
| --- | --- |
| **From station name** | **No. of trips** |
| Streeter Dr & Illinois St | 54214 |
| Lake Shore Dr & Monroe St | 41326 |
| Theater on the Lake | 38667 |
| Clinton St & Washington Blvd | 37755 |
| Michigan Ave & Oak St | 34668 |
| Millennium Park | 32075 |

### Top 5 stations End trip

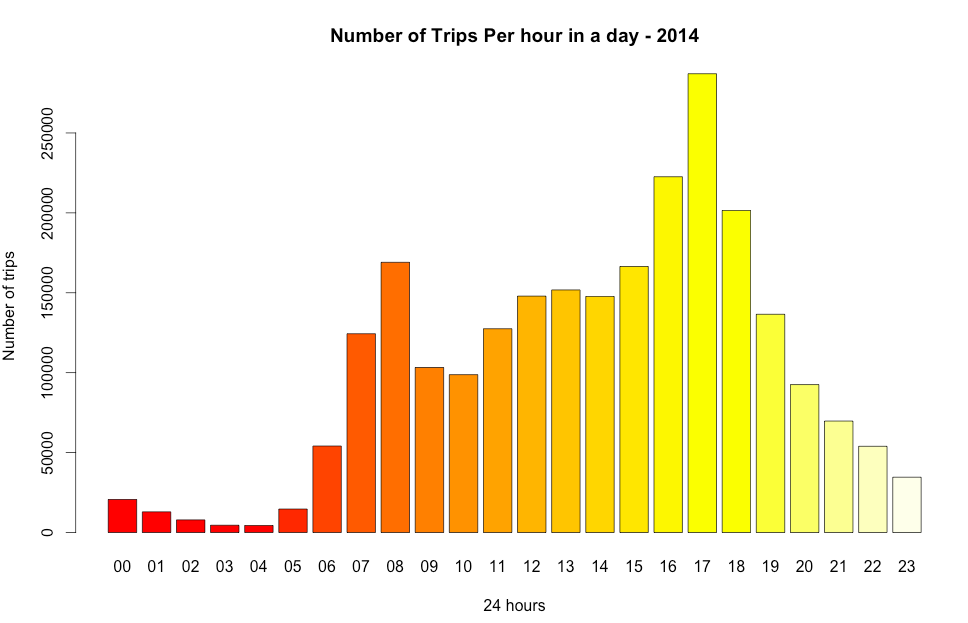
|  |  |
| --- | --- |
| **To station name** | **No. of trips** |
| Streeter Dr & Illinois St | 67048 |
| Lake Shore Dr & Monroe St | 42060 |
| Theater on the Lake | 41297 |
| Clinton St & Washington Blvd | 39517 |
| Michigan Ave & Oak St | 37422 |
| Millennium Park | 35481 |

# Reports:

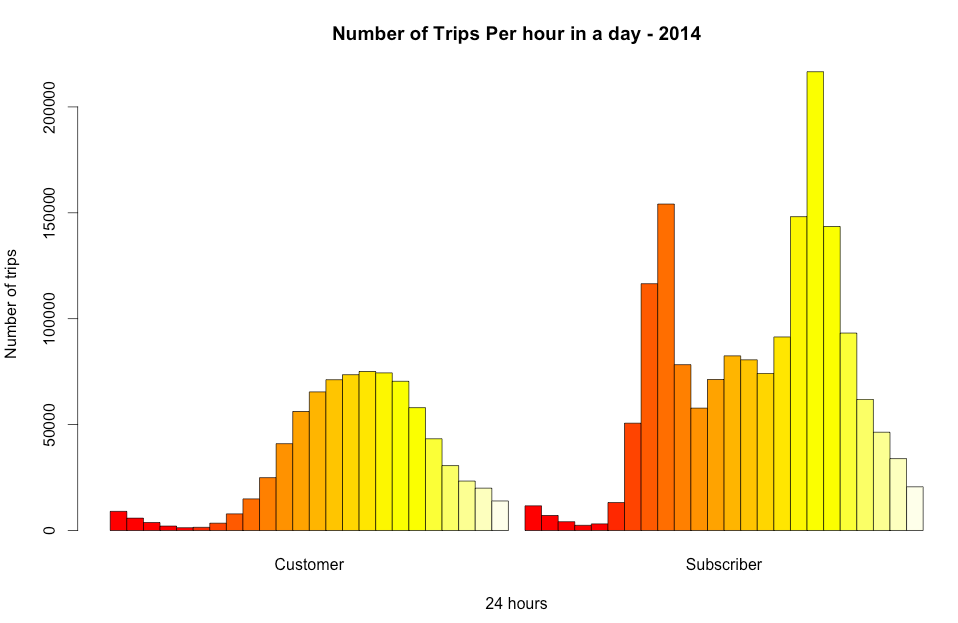
## Trip per hour of a day

1. The below graph shows the number of trips per hour in a day for the complete 2014 year.

Here we can the peak hour is **17hr.**

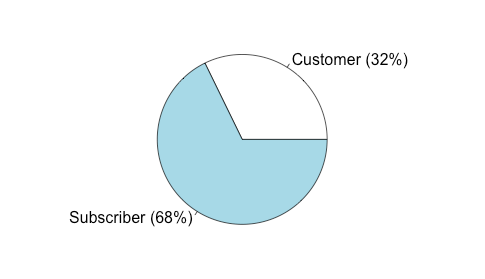


2. The Customer and Subscribers uses the bikes at different timings as shown below



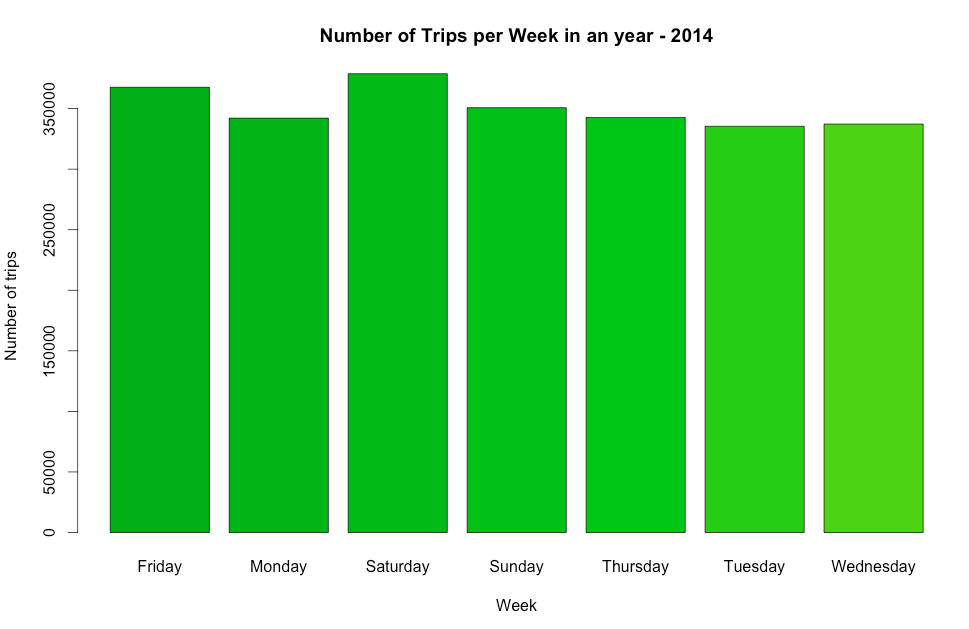
## Customer Vs Subscribers

This pie chart provide the percentage of customers and subscribers.

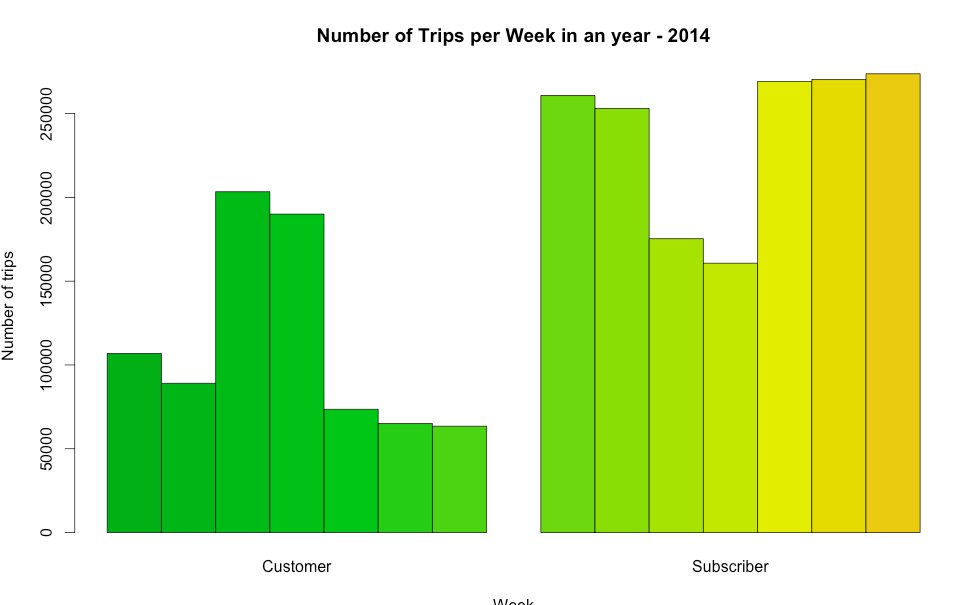


## Trips per week day

The number of trips per day in a week

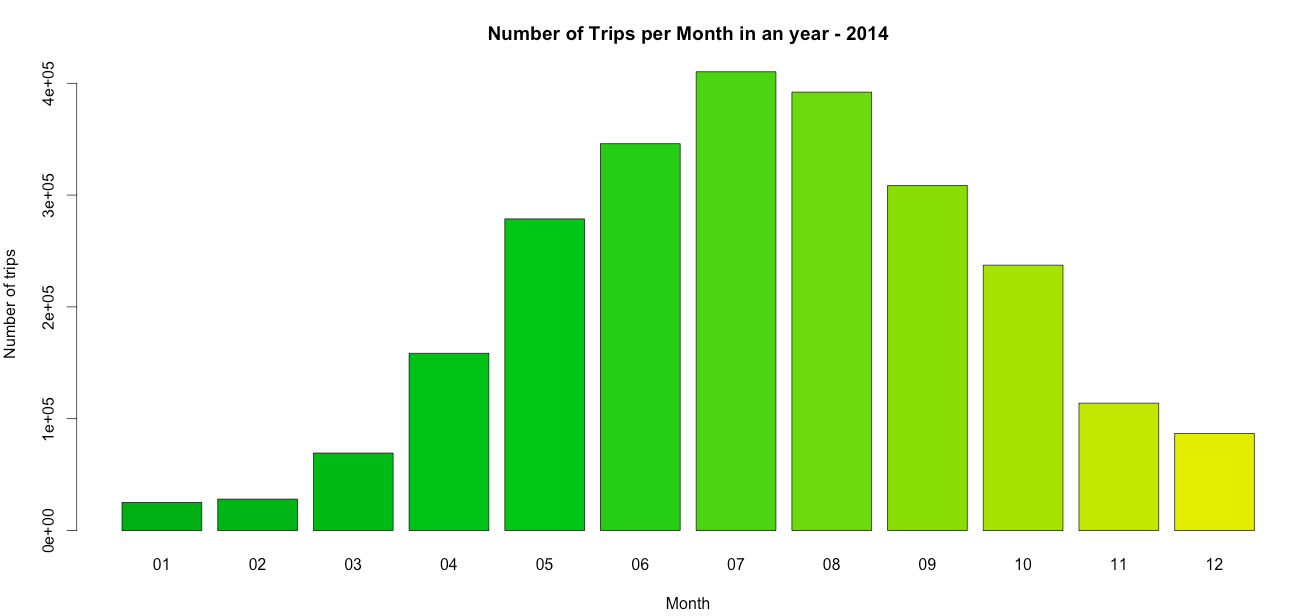


The number of trips per day in a week – Customer and Subscriber

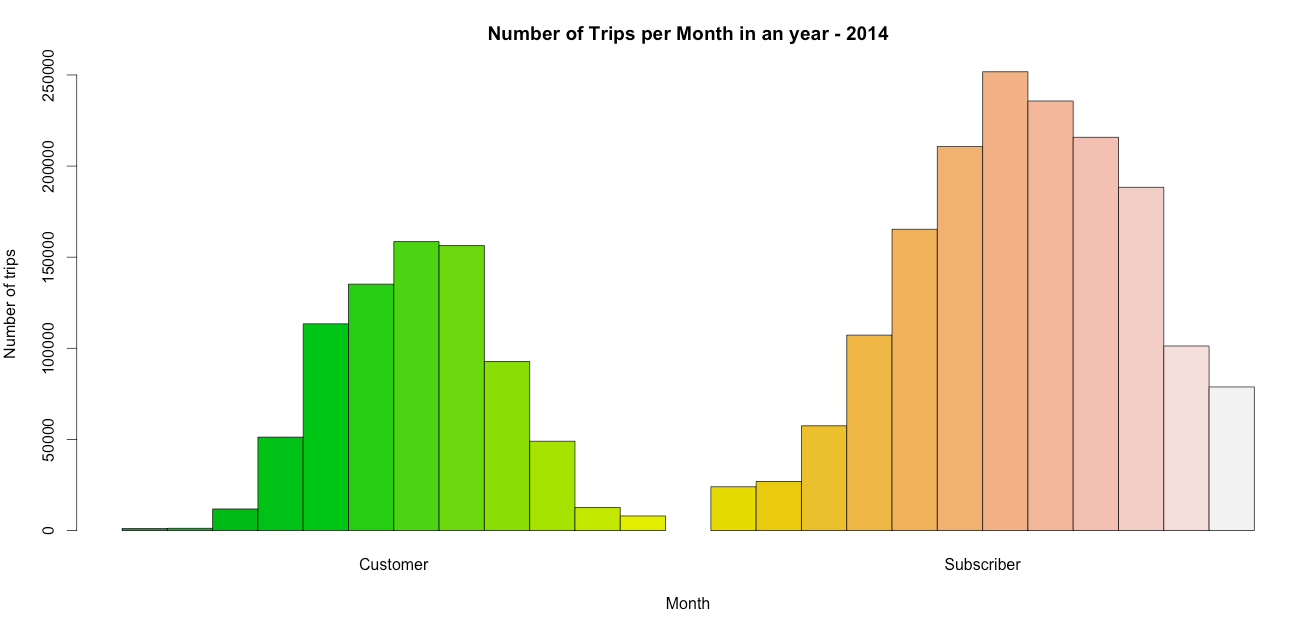


## Trips per month

The number of trips per month

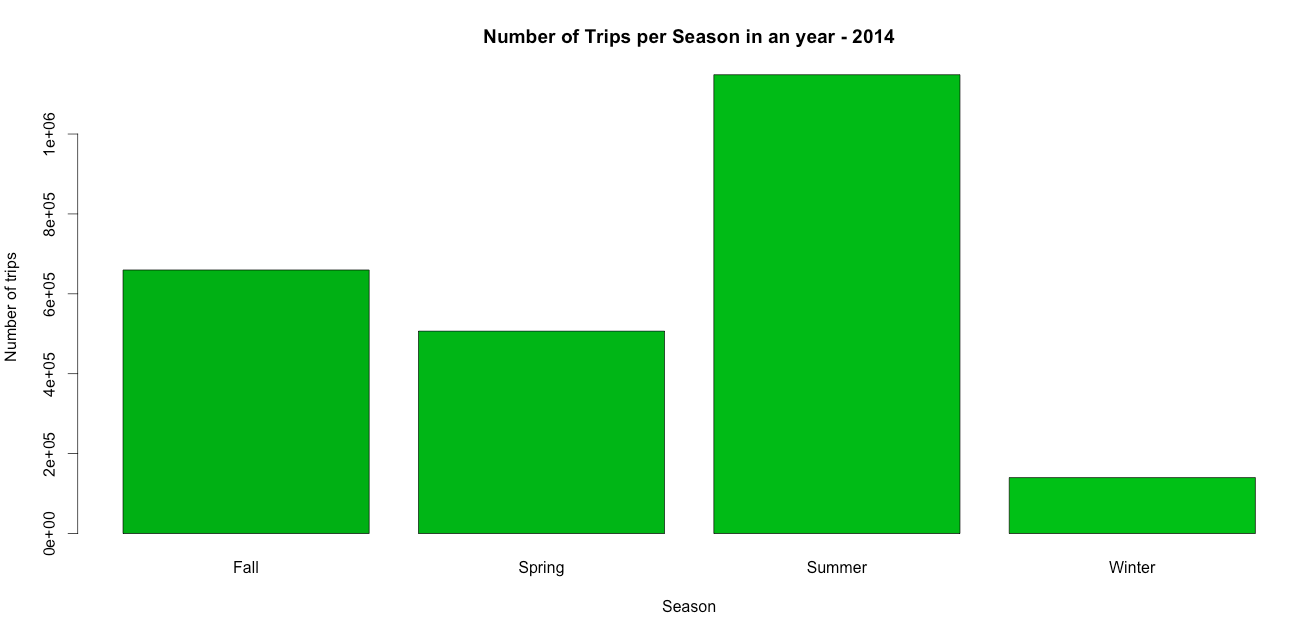


The number of trips per month – Customer and Subscriber

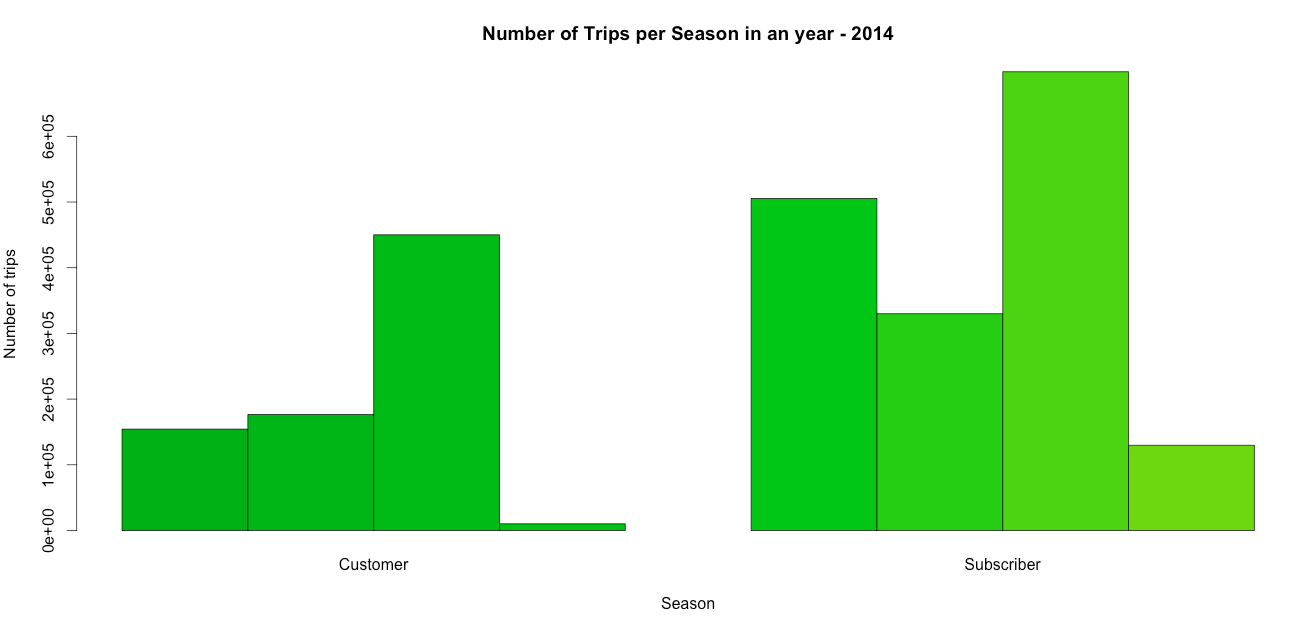


## Trips per season:

Number of Trips in a season

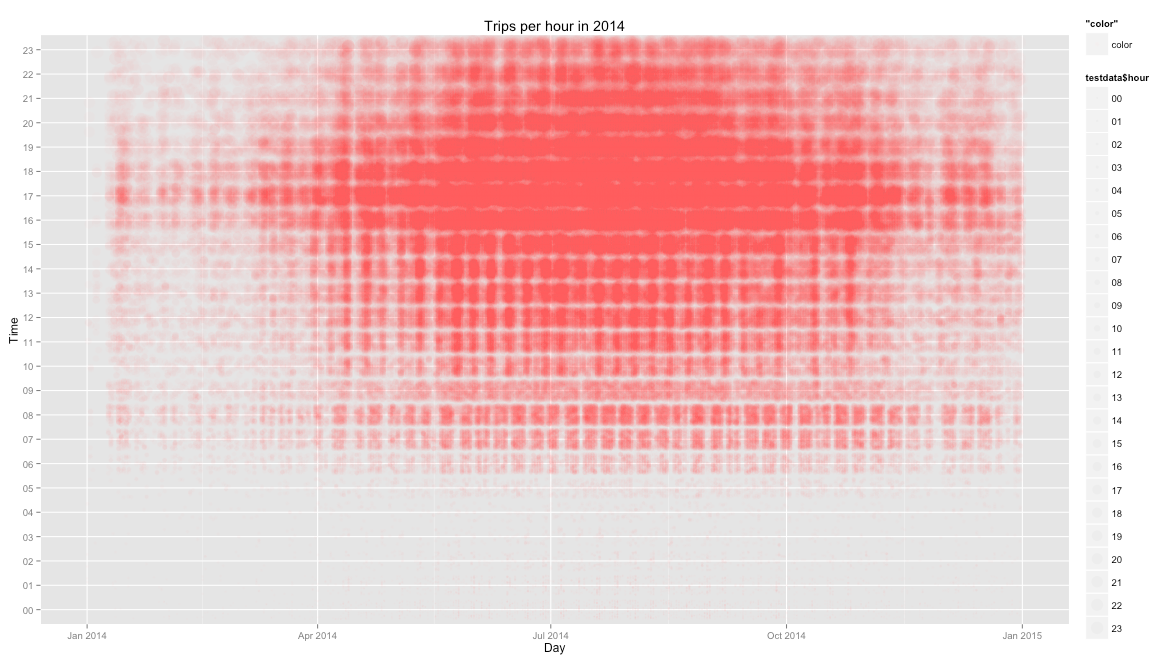


Number of Trips in a season – Customer and Subscriber

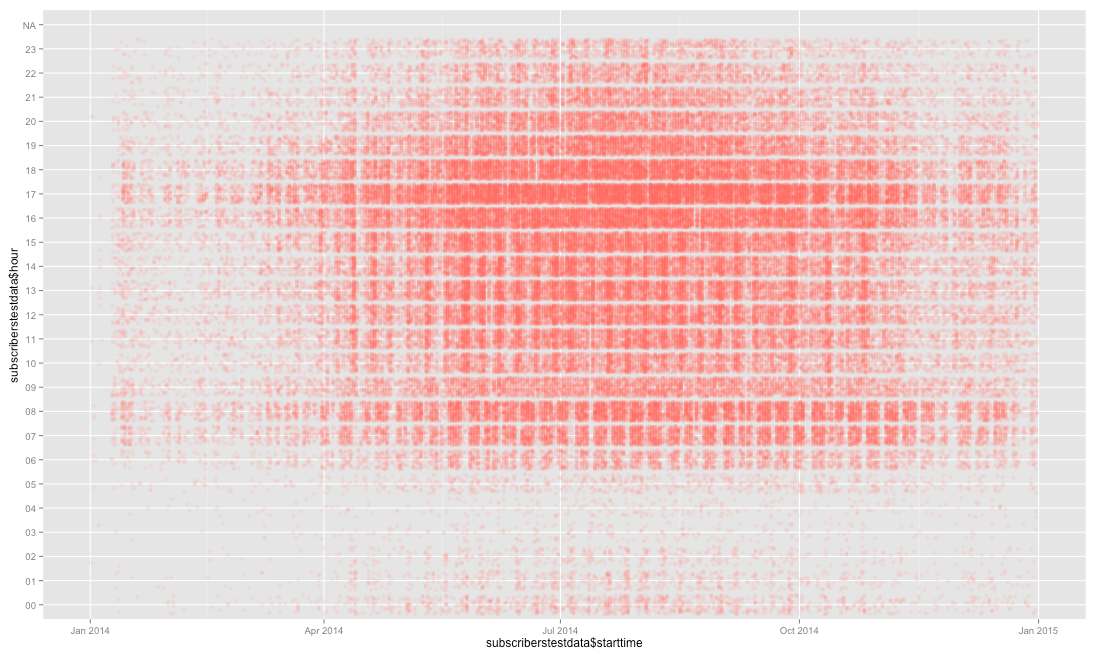


# Divvy Bikes data analysis

## Busy hour of Customers and Subscribers

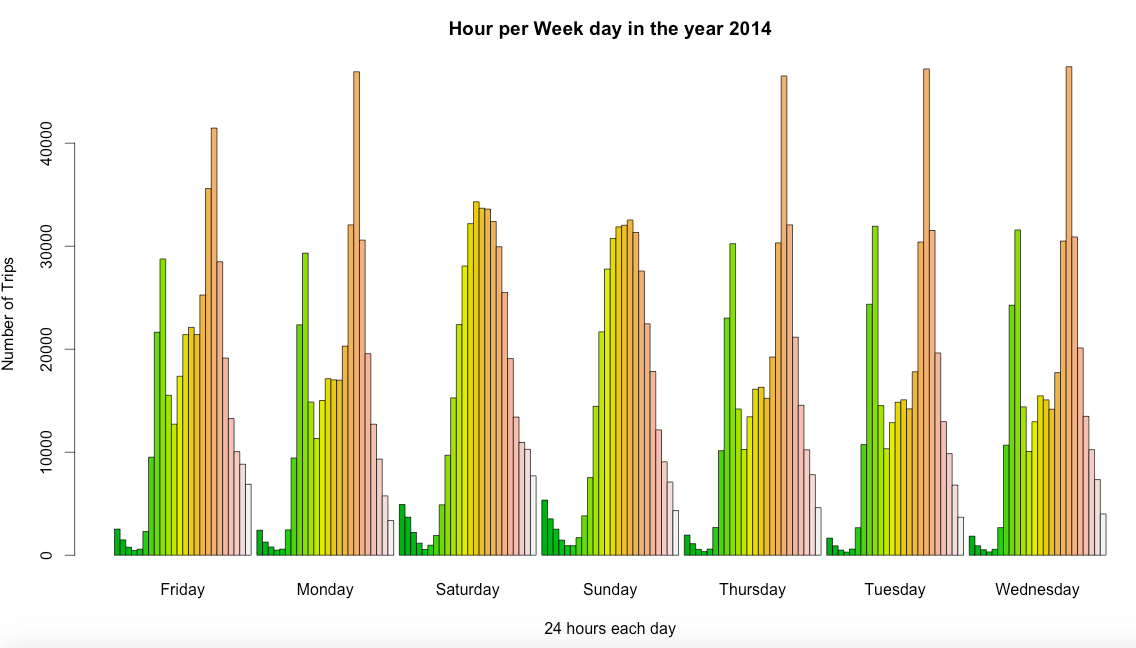




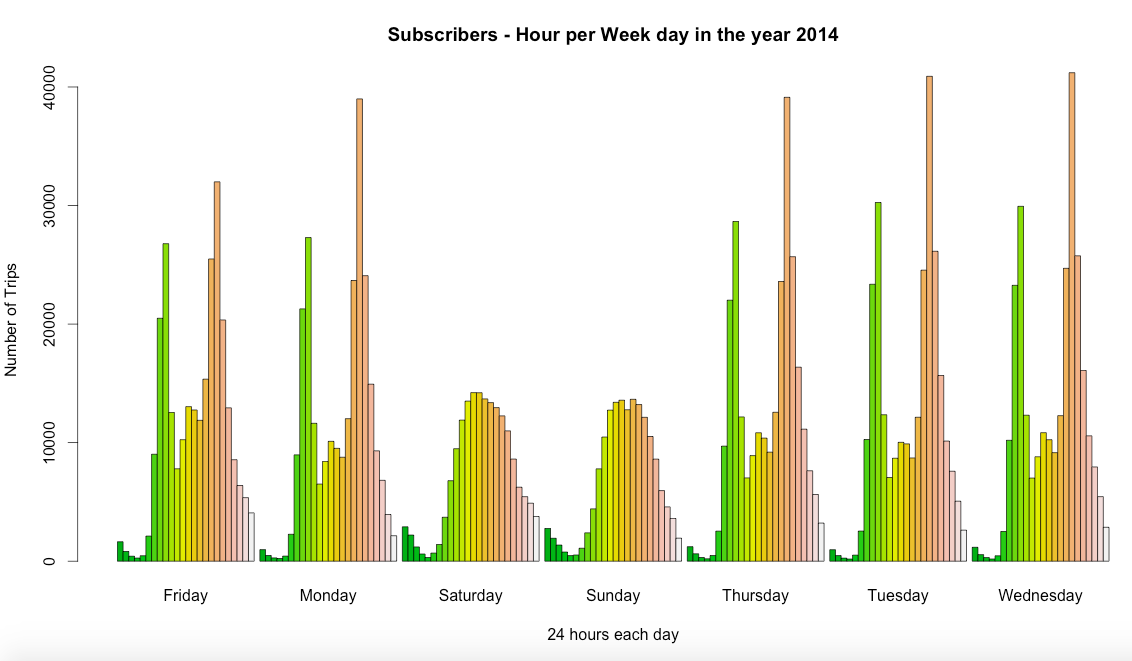


## Hour x dayofweek

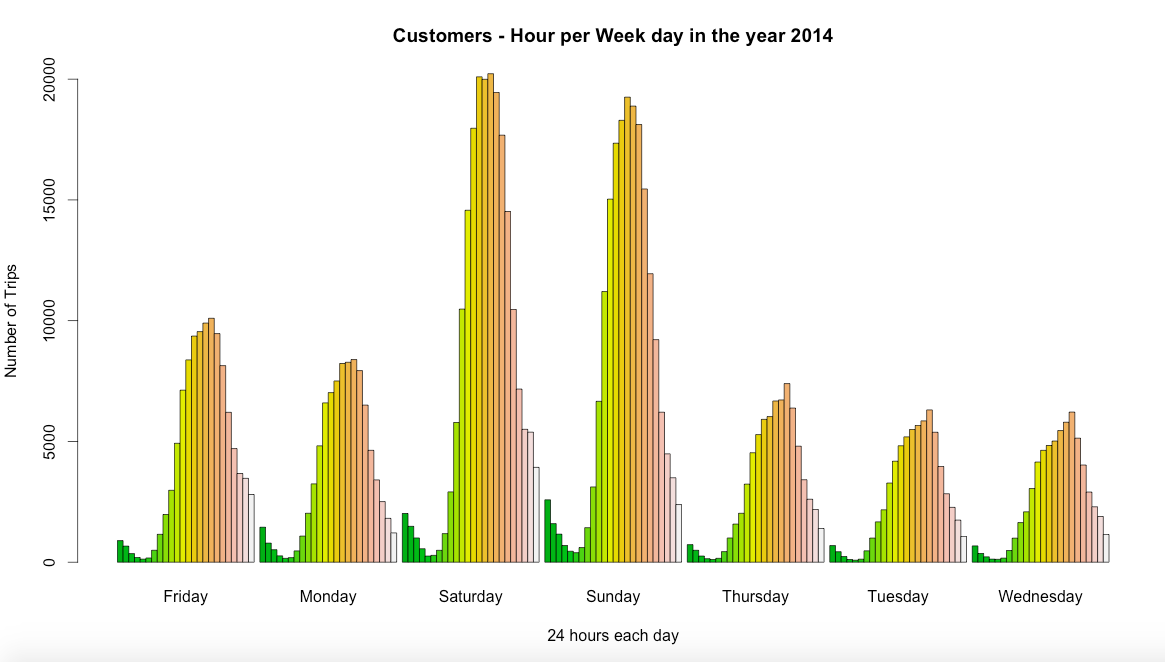
Trips per hour in a weekday



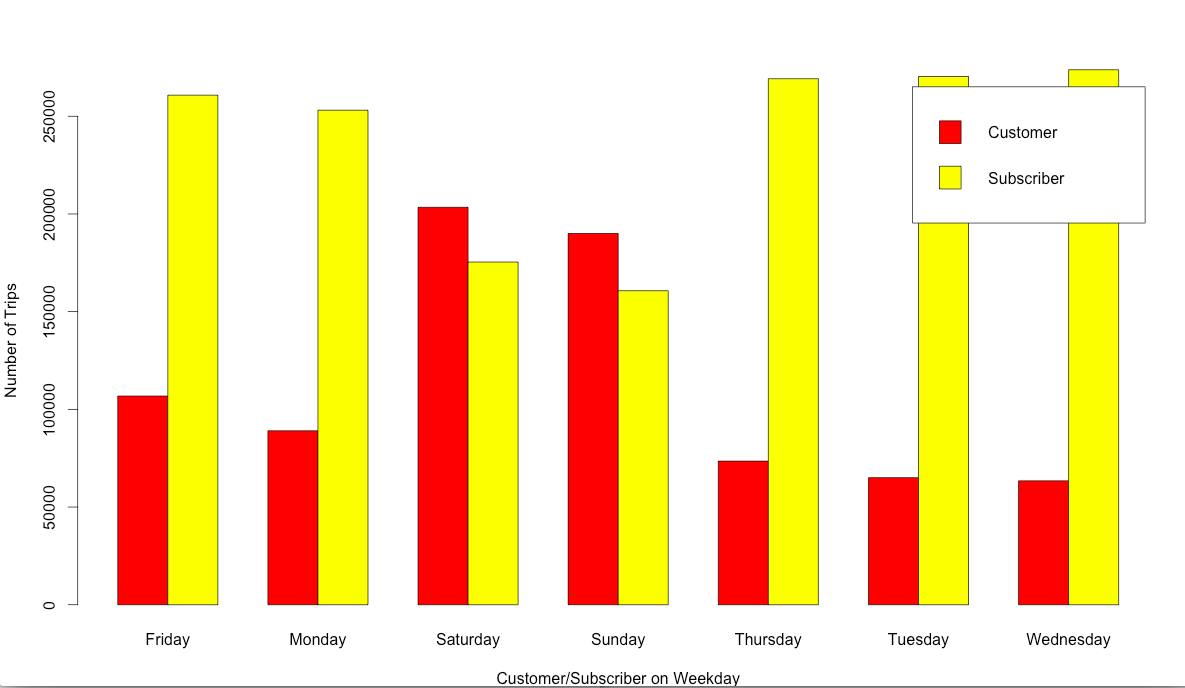
Trips per hour in a weekday - Subscriber



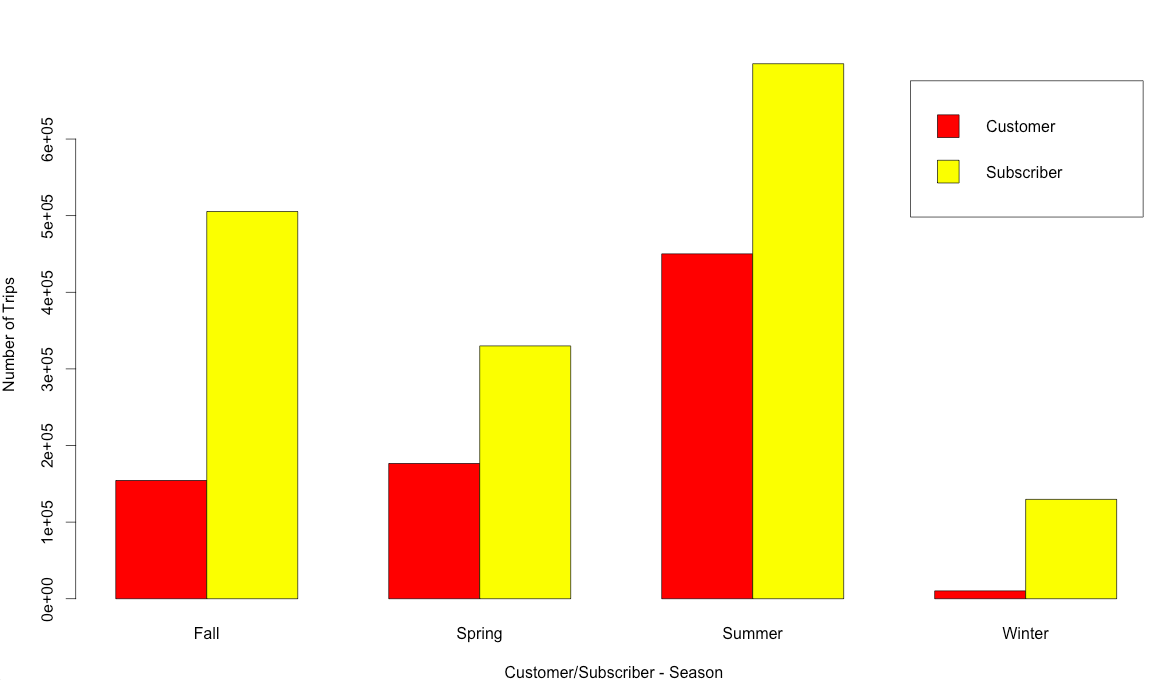
Trips per hour in a weekday - Customer



## User type x day of week



## User type x season



Divvy R-script:

library("ggplot2", lib.loc="/Library/Frameworks/R.framework/Versions/3.2/Resources/library")

library("gplots", lib.loc="/Library/Frameworks/R.framework/Versions/3.2/Resources/library")

divvydata <- read.csv("/Users/janardhanbonu/OneDrive/MUM/datamining/rawdata/Divvy\_Stations\_Trips\_2014-Q1Q2/Divvy\_Trips\_2014\_Q1Q2.csv")

names(divvydata)

nrows(divvydata)

# splitdf function will return a list of training and testing sets

splitdf <- function(dataframe, seed=NULL) {

if (!is.null(seed)) set.seed(seed)

index <- 1:nrow(dataframe)

trainindex <- sample(index, trunc(length(index)/10))

trainset <- dataframe[trainindex, ]

testset <- dataframe[-trainindex, ]

list(trainset=trainset,testset=testset)

}

splits <- splitdf(divvydata, seed=20000)

lapply(splits,nrow)

testdata <- splits$trainset

if (!is.null(seed)) set.seed(seed)

index <- 1:nrow(dataframe)

divvydata$starttime <- strptime(divvydata$starttime,"%m/%d/%Y %H:%M")

divvydata$stoptime <- strptime(divvydata$stoptime,"%m/%d/%Y %H:%M")

divvydata$weekday <- weekdays(divvydata$starttime)

divvydata$month <- months(divvydata$starttime)

divvydata$season[divvydata$month=="January"] <- "Winter"

divvydata$season[divvydata$month=="February"] <- "Winter"

divvydata$season[divvydata$month=="March"] <- "Spring"

divvydata$season[divvydata$month=="April"] <- "Spring"

divvydata$season[divvydata$month=="May"] <- "Spring"

divvydata$season[divvydata$month=="June"] <- "Summer"

divvydata$season[divvydata$month=="July"] <- "Summer"

divvydata$season[divvydata$month=="August"] <- "Summer"

divvydata$season[divvydata$month=="September"] <- "Fall"

divvydata$season[divvydata$month=="October"] <- "Fall"

divvydata$season[divvydata$month=="November"] <- "Fall"

divvydata$season[divvydata$month=="December"] <- "Winter"

divvydata$season <- as.factor(divvydata$season)

divvydata$month <- as.factor(divvydata$month)

divvydata$weekday <- as.factor(divvydata$weekday)

divvydata$hour <- format(divvydata$starttime,"%H")

divvydata$stationpair <- paste(divvydata$from\_station\_name,"&&",divvydata$to\_station\_name)

divvydata$season <- as.factor(divvydata$season)

divvydata$month <- as.factor(divvydata$month)

divvydata$weekday <- as.factor(divvydata$weekday)

divvydata$stationpair <- as.factor(divvydata$stationpair)

qplot(table(divvydata@stationpair))

barplot(table(divvydata$usertype, divvydata$weekday), beside=T, col=heat.colors(2), xlab="Customer/Subscriber on Weekday", ylab="Number of Trips", legend=T)

barplot(table(divvydata$usertype, divvydata$stationpair), beside=T, col=heat.colors(2), xlab="Customer/Subscriber - Station ", ylab="Number of Trips", legend=T)

barplot(table( divvydata$stationpair, divvydata$usertype), beside=T, col=heat.colors(2), xlab="Customer/Subscriber - Station ", ylab="Number of Trips", legend=T)

barplot(table(divvydata$usertype, divvydata$stationpair), beside=T, col=heat.colors(2), xlab="Customer/Subscriber - Station ", ylab="Number of Trips", legend=T)

barplot(table(divvydata$usertype, divvydata$season), beside=T, col=heat.colors(2), xlab="Customer/Subscriber - Station ", ylab="Number of Trips", legend=T)

barplot(table(divvydata$hour, divvydata$stationpair), beside=T, col=heat.colors(2), xlab="Customer/Subscriber - Season ", ylab="Number of Trips", legend=T)

testdata$hour <- format(testdata$starttime,"%H")

testdata$stationpair <- paste(testdata$from\_station\_name,"&&",testdata$to\_station\_name)

percent <- round(summary(testdata$weekday) \* 100 / nrow(testdata))

labels <- sprintf("%s (%d%%)", levels(testdata$weekday), percent)

pie(summary(testdata$weekday), lab=labels)

pie(table(usertype))

percent <- round(summary(testdata$usertype) \* 100 / nrow(testdata))

labels <- sprintf("%s (%d%%)", levels(testdata$usertype), percent)

pie(summary(testdata$usertype), lab=labels)

barplot(table(hour))

barplot(table(month))

barplot(table(season))

barplot(table(testdata$season))

barplot(table(testdata$weekday))

testdata$stationpair <- as.factor(testdata$stationpair)

summary(testdata$stationpair)

qplot(testdata$starttime, testdata$trip\_id, geom = "jitter", alpha = I(1/50), size=testdata$from\_station\_name)