Divvy Bike data challenge

Divvy Bike data challenge with R Version 1.0

Student: Hung Cao 984401
Professor: Dr. Anil Maheshwari

Author: Hung Cao

Divvy Bike data challenge

Author: Hung Cao

Introduction

Divvy is a bicycle sharing system located in the City of Chicago operated by Motivate for the Chicago Department of Transportation. It operates with 4760 bicycles at 476 stations in an area bounded by 75th Street on the south, Touhy Avenue on the north, Lake Michigan on the east, and Pulaski Road on the west.

Summary

Data schema

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
stationpair*	Station Source and Destination
latitude_from*	Latitude of from station
longtitude_from*	Longitude of from station
latitude_to*	Latitude of to station
longtitude_to*	Longitude of to station

* Newly added during analysis.

Number of trips

2454634

Total duration of trips

2515853900

Average duration of trips

~1025

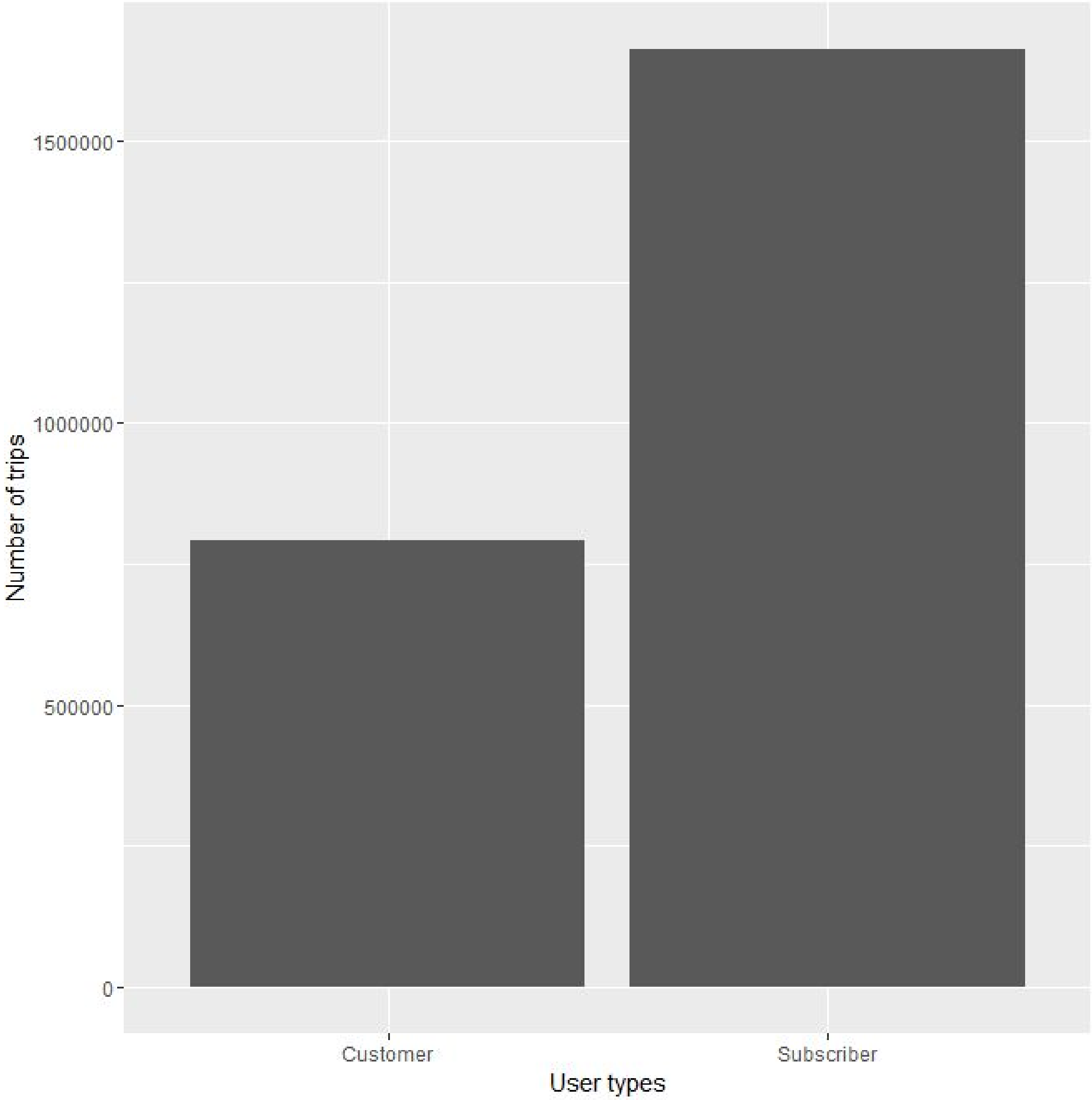
Number of users

Subscriber

1663394

Customers

791240



Largest duration

86392

Stations

At launch of the Divvy service, there were 68 stations online with a maximum capacity of 1,352 bikes total. Over the next 6 months, 232 more stations came online, for a total of 300 stations and capacity of 5040 bikes.

Top 10 station pairs

Name	Number of trips
Lake Shore Dr & Monroe St && Streeter Dr & Illinois St	8019
Lake Shore Dr & Monroe St && Lake Shore Dr & Monroe St	5445
Streeter Dr & Illinois St && Lake Shore Dr & Monroe St	5166
Theater on the Lake && Streeter Dr & Illinois St	4898
Streeter Dr & Illinois St && Streeter Dr & Illinois St	4483
Michigan Ave & Oak St && Michigan Ave & Oak St	4353
Streeter Dr & Illinois St && Theater on the Lake	4089
Streeter Dr & Illinois St && Millennium Park	3665
Lake Shore Dr & North Blvd && Streeter Dr & Illinois St	3509

Top 10 start station

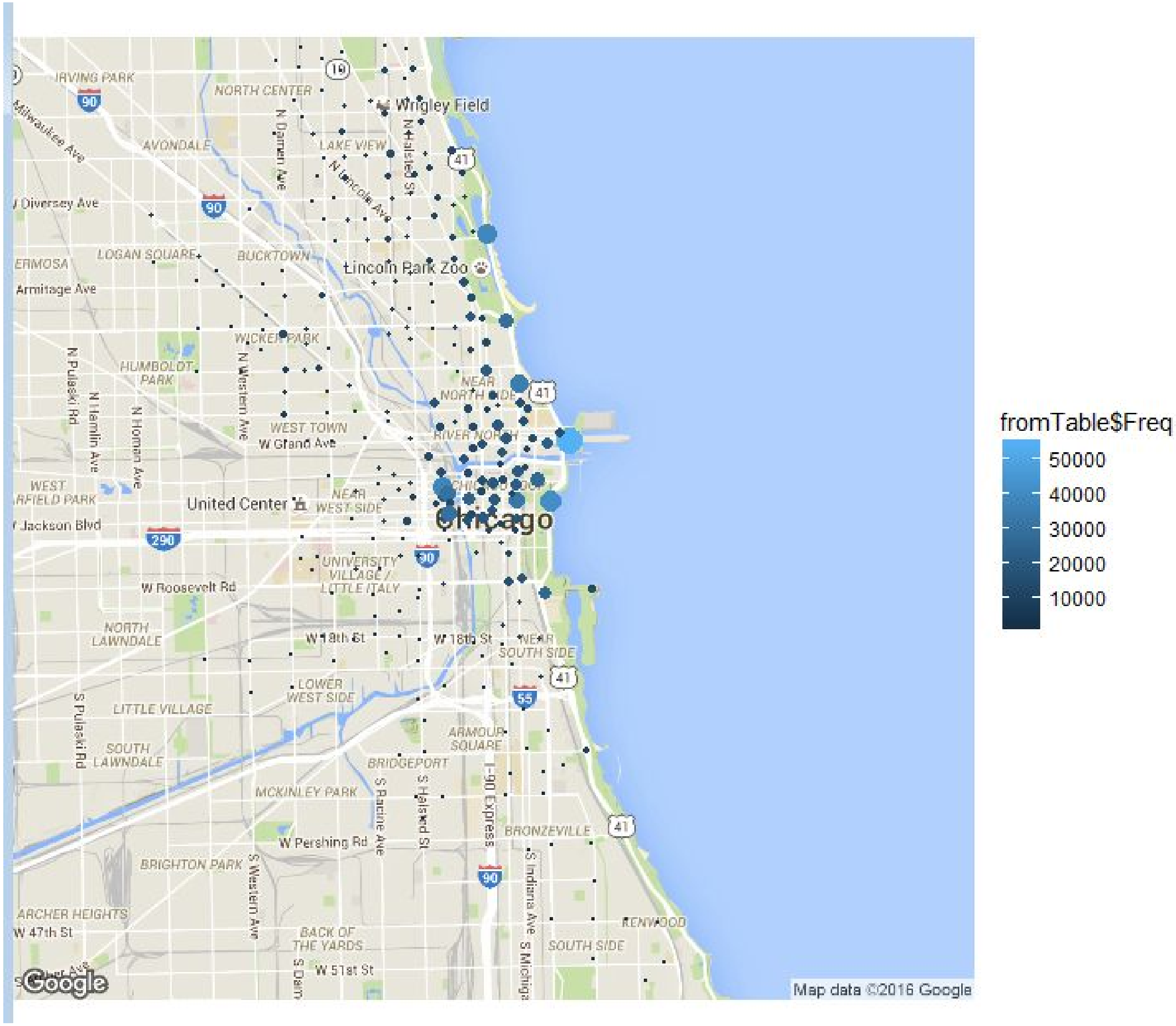
Name	Number of trips
Streeter Dr & Illinois St	54214
Lake Shore Dr & Monroe St	41326
Theater on the Lake	38667
Clinton St & Washington Blvd	37755
Michigan Ave & Oak St	34668
Millennium Park	32075
Canal St & Madison St	30277
Canal St & Adams St	30165
Lake Shore Dr & North Blvd	29208
Columbus Dr & Randolph St	26797

Top 10 end station

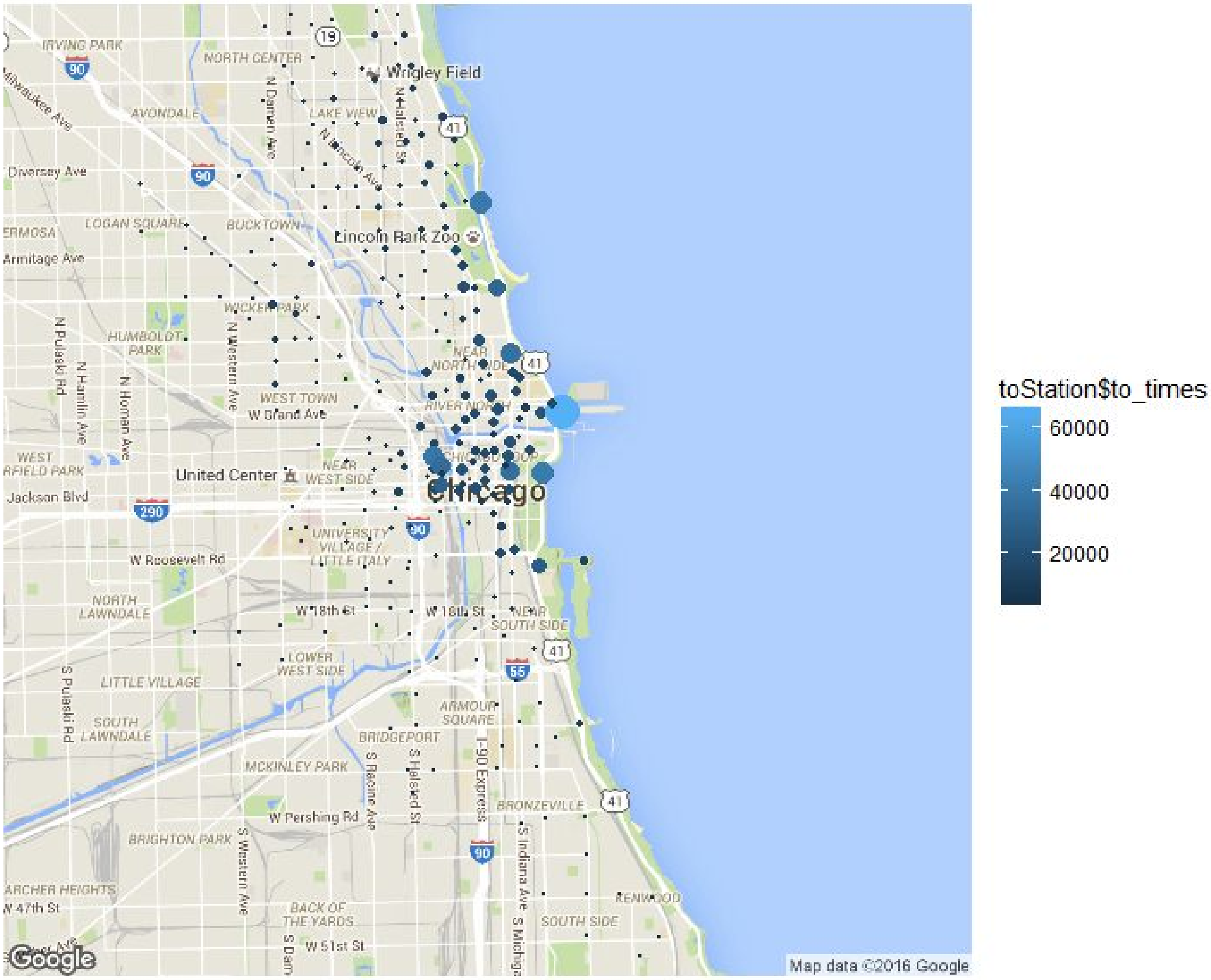
Name	Number of trips
Streeter Dr & Illinois St	67048
Lake Shore Dr & Monroe St	42060
Theater on the Lake	41297
Clinton St & Washington Blvd	39517
Michigan Ave & Oak St	37422
Millennium Park	35481
Canal St & Madison St	34650
Lake Shore Dr & North Blvd	32613
Canal St & Adams St	29082
Museum Campus	26982

Of course, most of these stations are located around tourist attractions: Navy Pier, the Mag Mile, and Millenium Park, harbour. So in the future we may want to expand our stations in those places. To be more clear, take a look at below 2 images.

Map of start stations with their trips

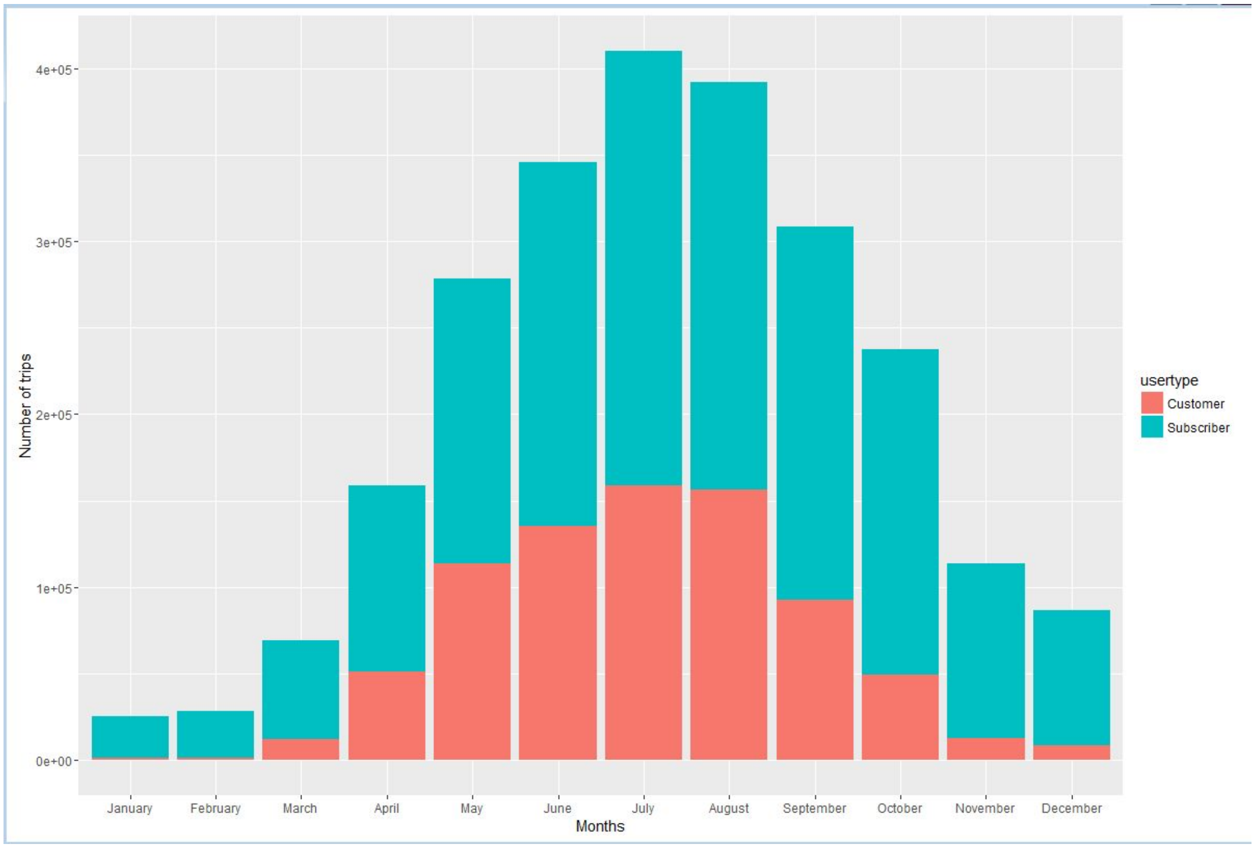


Map of end stations with their trips

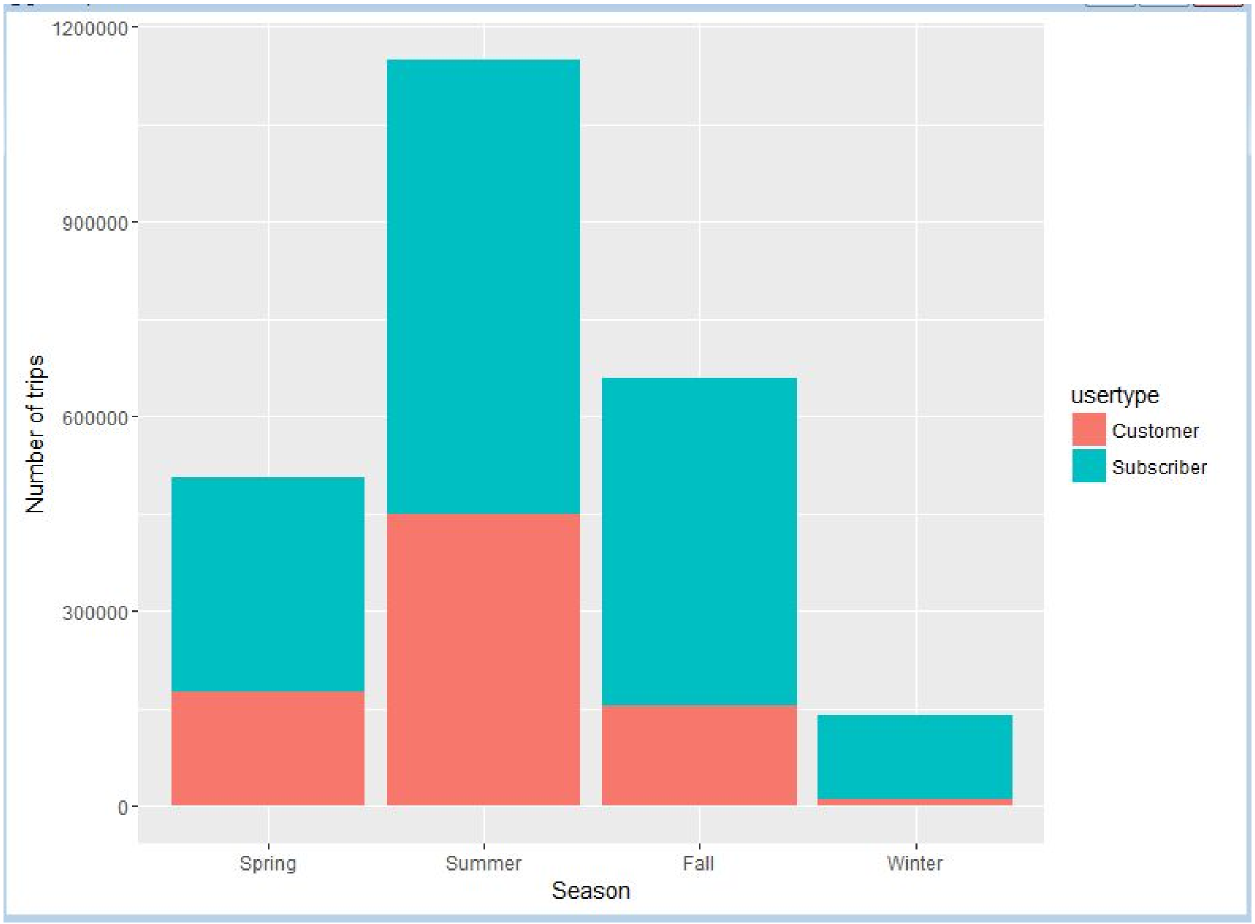

toStation$to_times
60000
40000
20000
Wrigley Field
Lincoln Park Zoo
United Center
Chicago
IRVING PARK
NORTH CENTER
LAKE VIEW
AVONDALE
LOGAN SQUARE
BUCKTOWN
WICKER PARK
HUMBOLDT PARK
WEST TOWN
NEAR NORTH SIDE
RIVER NORTH
NEAR WEST SIDE
UNIVERSITY VILLAGE / LITTLE ITALY
NORTH LAWNDALE
LOWER WEST SIDE
LITTLE VILLAGE
SOUTH LAWNDALE
ARMOUR SQUARE
BRIDGEPORT
MCKINLEY PARK
BRONZEVILLE
BRIGHTON PARK
ARCHER HEIGHTS
BACK OF THE YARDS
KENWOOD
SOUTH SIDE
NEAR SOUTH SIDE
Milwaukee Ave
Diversey Ave
Armitage Ave
Jackson Blvd
W Roosevelt Rd
W 18th St
W Grand Ave
W Pershing Rd
W 47th St
W 51st St
N Pulaski Rd
N Hamlin Ave
N Homan Ave
N Western Ave
N Damen Ave
N Halsted St
N Lincoln Ave
S Pulaski Rd
S Western Ave
S Racine Ave
S Halsted St
I-90 Express
S Indiana Ave
S Michigan
19
90
41
290
55
Google
Map data ©2016 Google

Rides

Number of trips and months

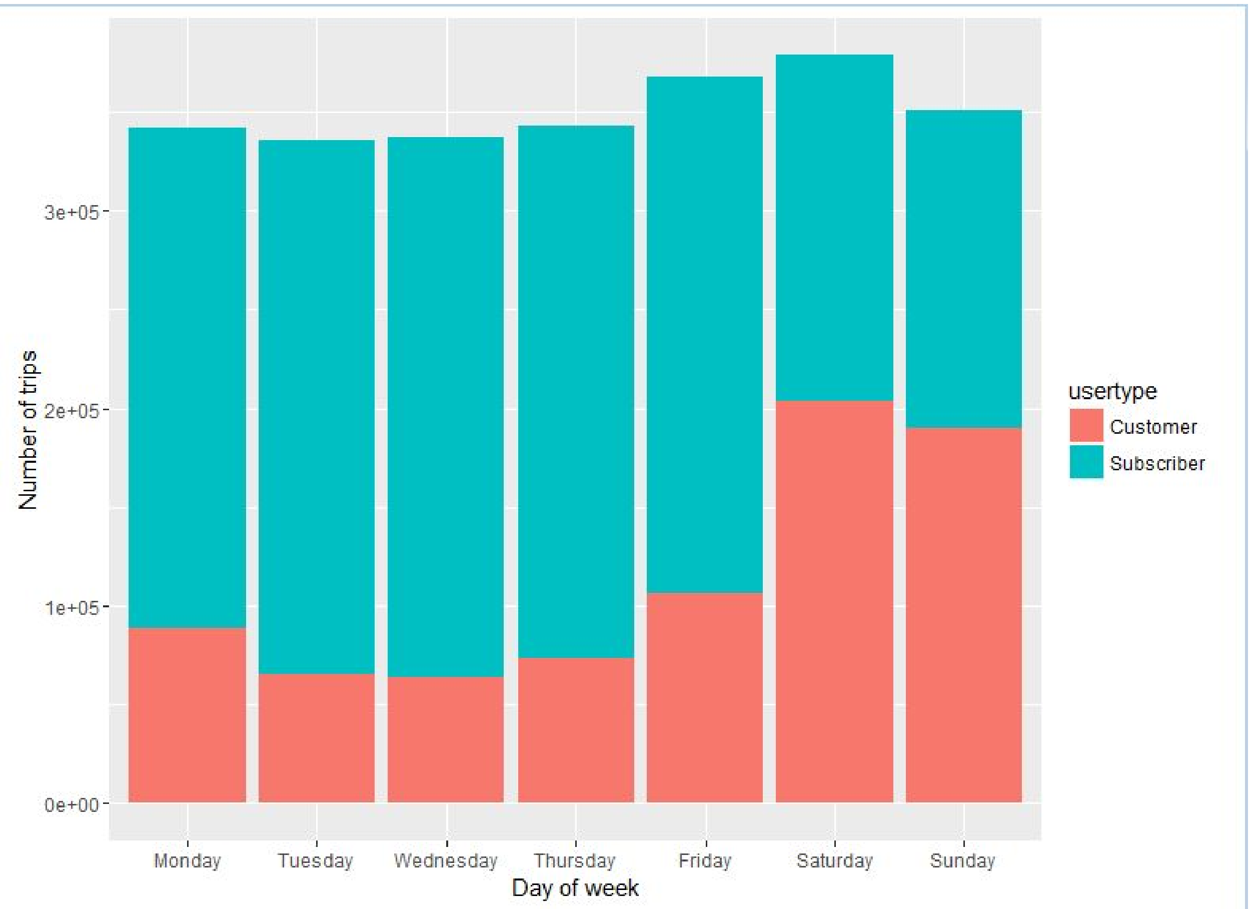


Number of trips and seasons



With this map, we can understand that the number of trips was increasing significantly in Summer (June to August) for both customer and subscriber.
Subscriber always tends to use our services more than customer, especially in Fall and Winter.

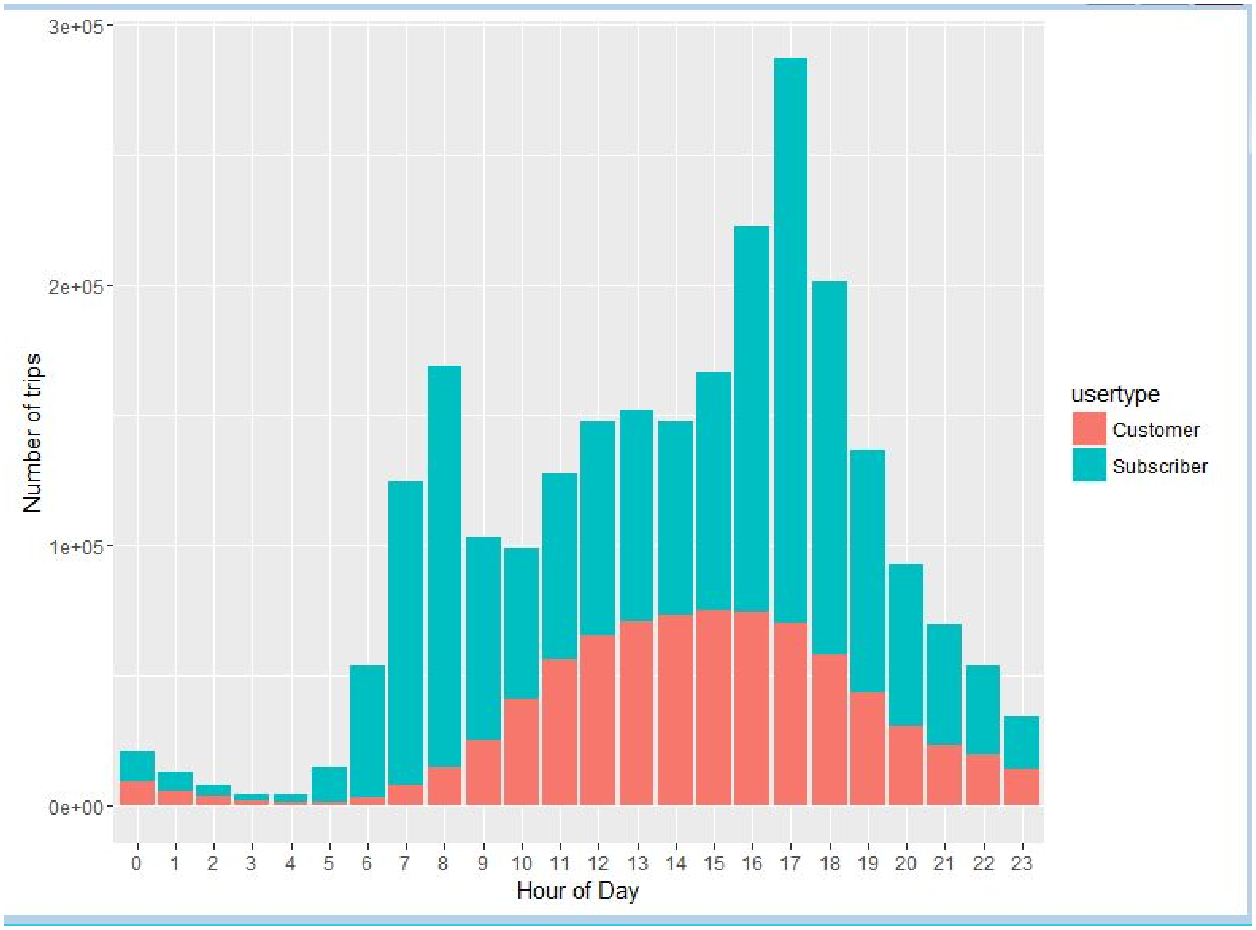
Number of trips and day of week



It's clear that in work day (Monday to Friday) subscriber user tends to use our services much more than normal customer. They are using it for daily commuting as well as other activities. In Saturday and Sunday, the number of customer was increasing significantly because it's weekend so the needs is high.

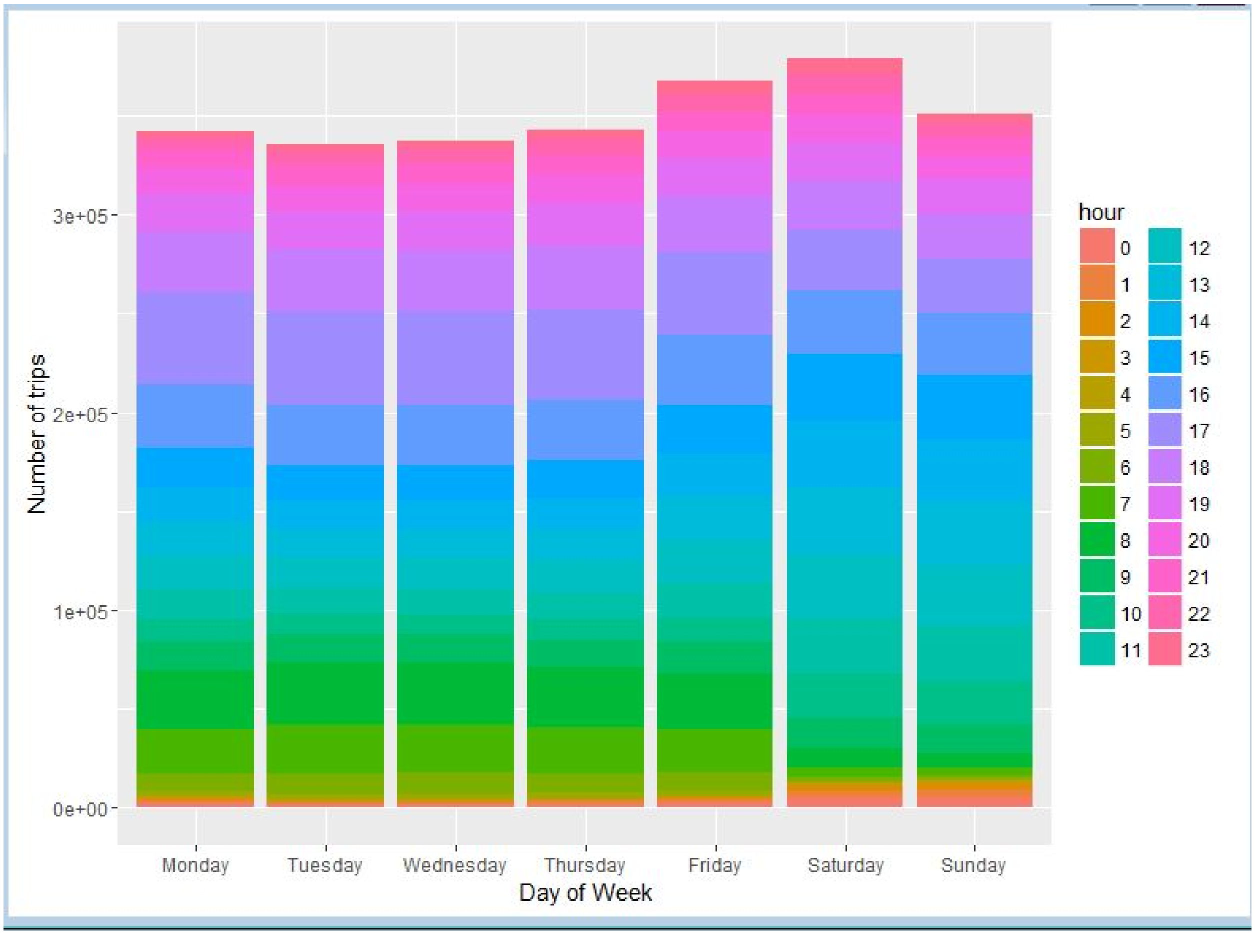
The reason the number of subscriber was decreasing on weekends because a lot subscribers use our service as daily commute to work.

Number of trips and hour



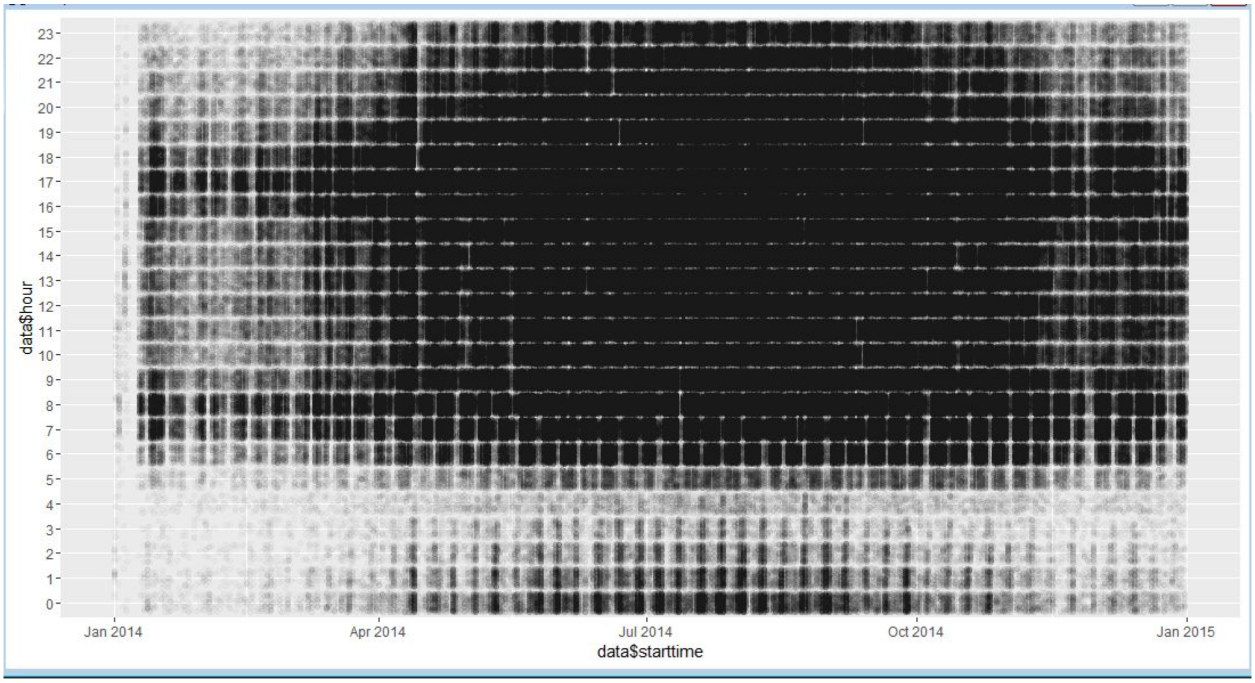
This chart's message is straightforward to say that the user's need is highest at 16-18 every days because they are ideal hours for people to go outside at the weekend or coming back in working days.

Number of trips in everyday filled by hour



This chart show us again there many people who use our service as daily commuting to work. They often start using in the morning and coming back in the afternoon.

Busy hour in year



Bikes

Top 10 most used bikes

Bike ID	Freq
1768	1232
30	1187
1975	1174
1080	1160
2837	1160
2915	1155
513	1147

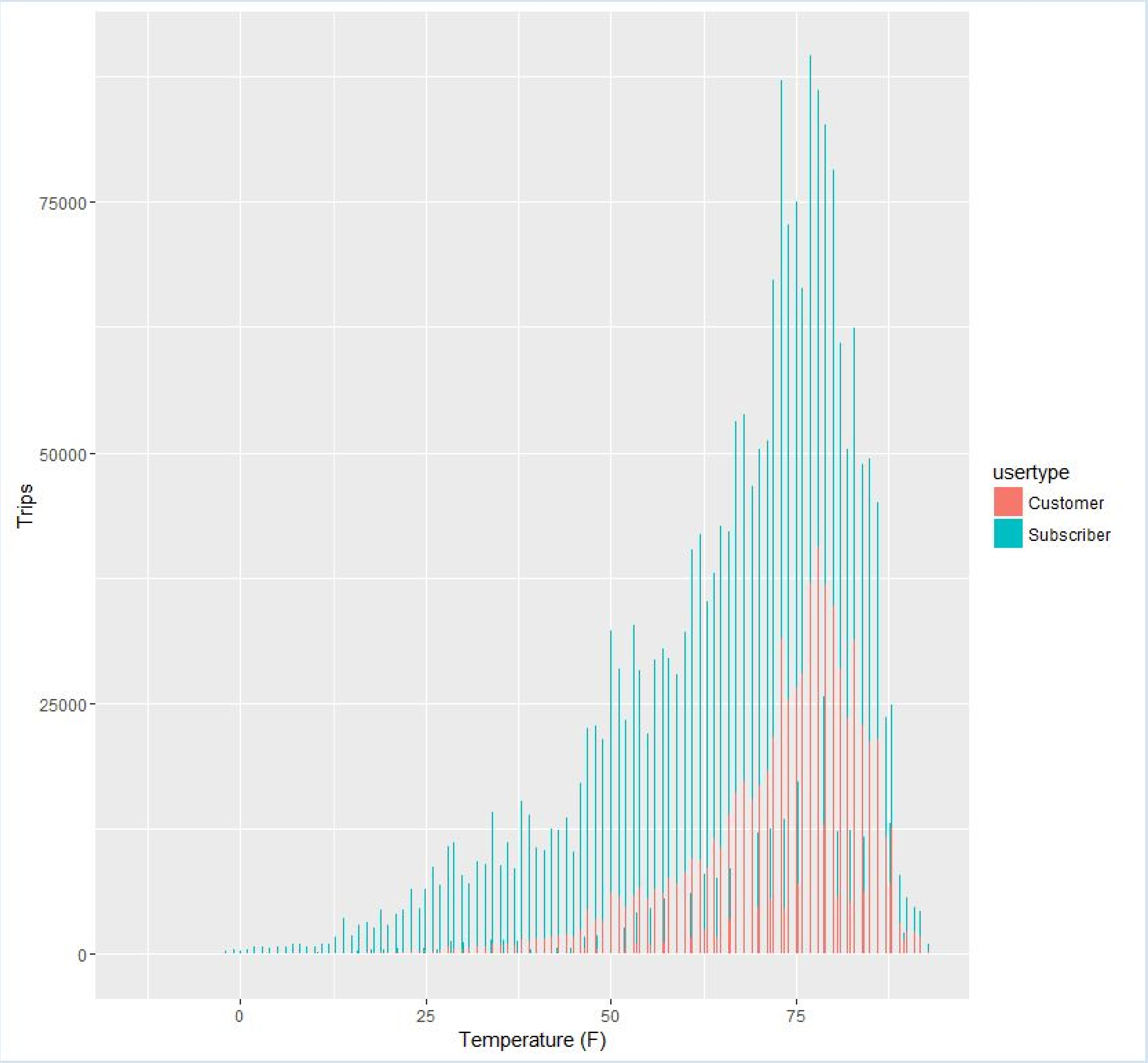
2176	1146
2651	1145
1979	1140

Top 10 less used bikes

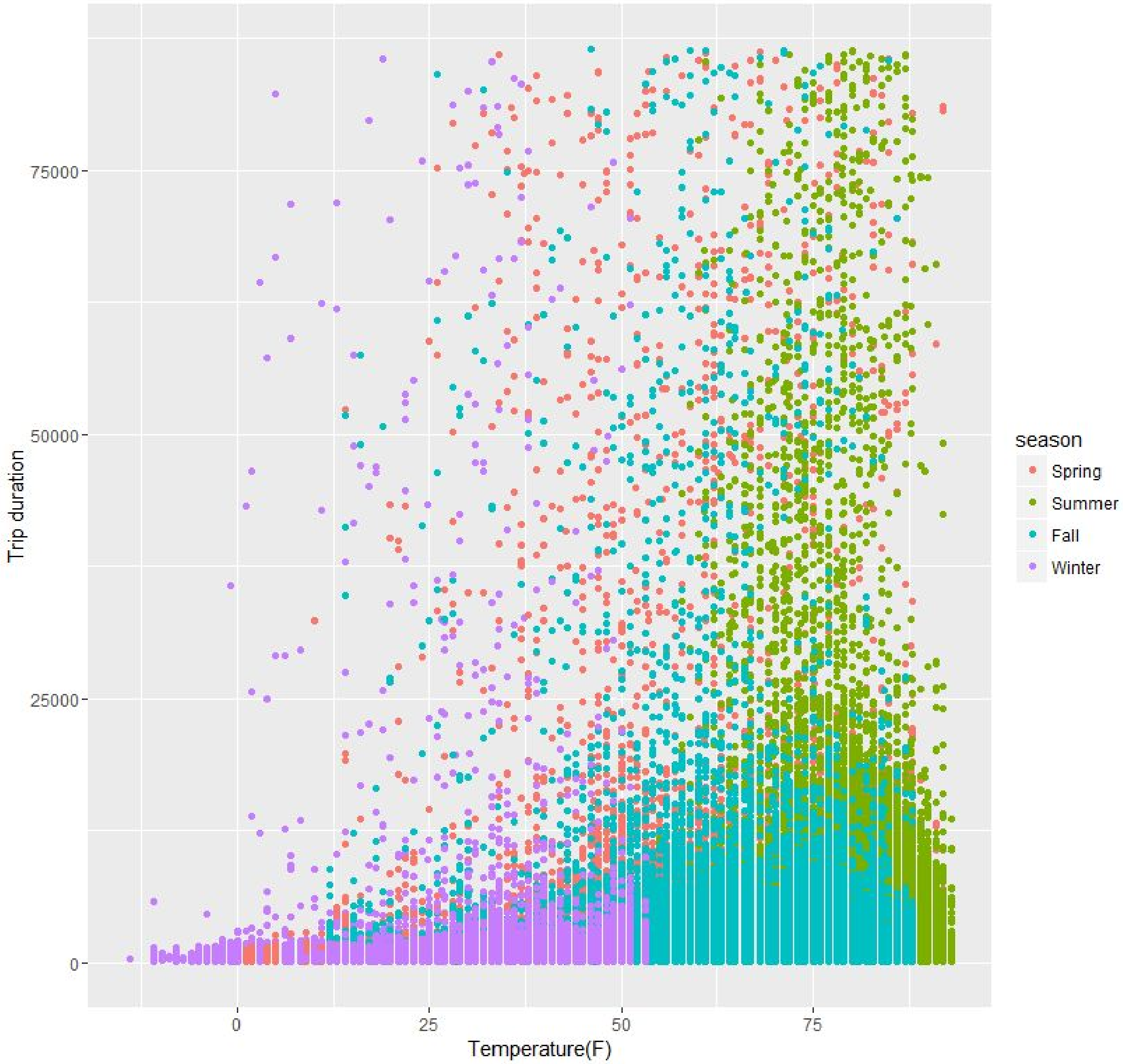
BikeID	Freq
2351	69
1508	68
1963	63
312	58
2969	57
2250	56
424	47
37	35
358	32
39	31

Weathers

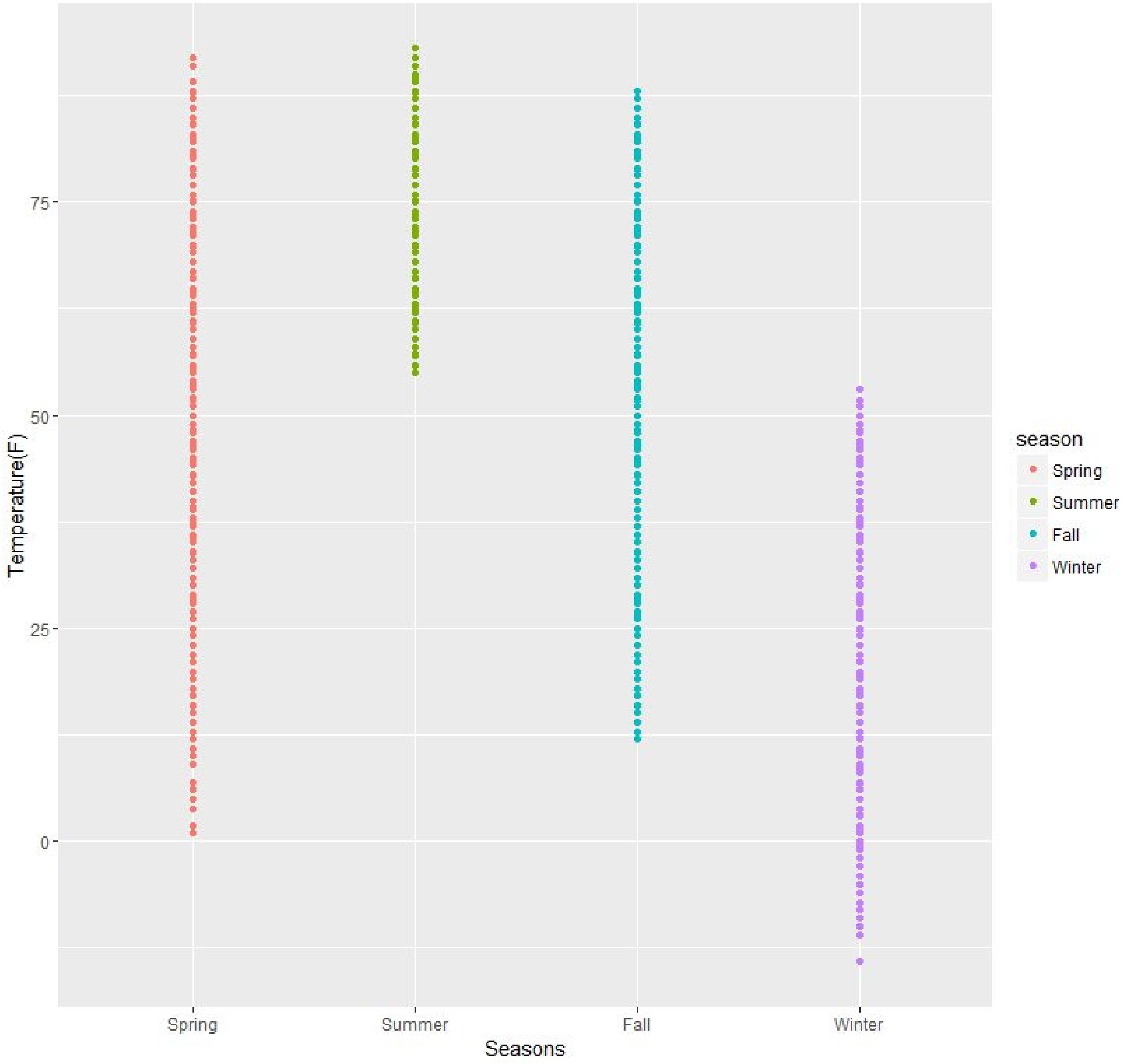
Number of trips in different temperatures



Trip duration in different temperatures by season



Temperature in different seasons



With these above maps, we can see the trend of users in a very clear way. People loves to go outside when the temperature is from 65-75 degrees especially in Summer and Spring. In Winter, the number of trips were decreased due to the cold but when the weather got warmer they still went out.

And there is a huge number of users who use our services every day in all weather conditions.

Appendix: Used R scripts

Prepare data

```
setwd("/Users/hungqcao/Desktop/R/Ass/")
> data1<-read.csv("Divvy_Trips_2014_Q1Q2.csv")
> data2<-read.csv("Divvy_Trips_2014-Q3-07.csv")
> data3<-read.csv("Divvy_Trips_2014-Q3-0809.csv")
> data4<-read.csv("Divvy_Trips_2014-Q4.csv")

> data<-rbind(data1, data2, data3, data4)
data<-read.csv
install.packages("ggplot2")
install.packages("gplots")
library("ggplot2")
data$starttime <- strptime(data$starttime,"%m/%d/%Y %H:%M")
data$stoptime <- strptime(data$stoptime,"%m/%d/%Y %H:%M")
data$weekday <- weekdays(data$starttime)
data$month <- months(data$starttime)
#order data by correct order
data$month<-factor(data$month,
levels=c("January","February","March","April","May","June","July","August","September","October","November","December"))
data$season[data$month=="January"] <- "Winter"
data$season[data$month=="February"] <- "Winter"
data$season[data$month=="March"] <- "Spring"
data$season[data$month=="April"] <- "Spring"
data$season[data$month=="May"] <- "Spring"
data$season[data$month=="June"] <- "Summer"
data$season[data$month=="July"] <- "Summer"
data$season[data$month=="August"] <- "Summer"
data$season[data$month=="September"] <- "Fall"
data$season[data$month=="October"] <- "Fall"
data$season[data$month=="November"] <- "Fall"
data$season[data$month=="December"] <- "Winter"
data$season <- as.factor(data$season)
> data$season<-factor(data$season,c("Spring","Summer","Fall","Winter"))
data$month <- as.factor(data$month)
data$weekday <- as.factor(data$weekday)
data$hour <- format(data$starttime,"%H")
data$stationpair <- paste(data$from_station_name,"&&",data$to_station_name)
data$season <- as.factor(data$season)
data$month <- as.factor(data$month)
```

Author: Hung Cao

```
data$weekday <- as.factor(data$weekday)
data$stationpair <- as.factor(data$stationpair)
#createdata function will return a list of training and testing sets
createdata <- 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 <- createdata(data, seed=20000)
lapply(splits,nrow)
testset<- splits$trainset
testset$hour <- hour(testset$startTime)

data$hour<-factor(hour(data$starttime),c("0","1","2","3","4","5","6","7","8","9","10","11","12","13"
,"14","15","16","17","18","19","20","21","22","23","24"))

testset$stationpair <- paste(testset$from_station_name,"--",testset$to_station_name)
percent <- round(summary(testset$weekday) * 100 / nrow(testset))
labels <- sprintf("%s (%d%%)", levels(testset$weekday), percent)

stations<-read.csv("Divvy_Stations_2014.csv")
save.image("C:\\Users\\hcao\\Documents\\R\\workspace")

stations$from_station_id<-stations$id

newdata<-merge(data, stations, by="from_station_id", all.x=TRUE)
install.packages(“ggmap”)
library("ggmap")
tmp<-data.frame(lat=c(40.725095,40.725116,40.724652,40.723371),
               lon=c(-73.999115,-73.999775,-73.995937,-73.996085),
               name=c("Apple Store", "Kidrobot", "Puck Fair", "McNally Jackson Books"))

install.package("foreach")
timezone <- "America/Chicago"
getWeatherDataForChicagoIn2014 <- function(){
          # create folder to contains the data
          data_folder <- "weather_data"
          ifelse(!dir.exists(data_folder), dir.create(data_folder), FALSE)
          # loop through all days of year 2014, and get the weather data
```

```
    start_date <- as.Date("2014-01-1")
    end_date <- as.Date("2014-12-31")
    days <- seq(start_date, end_date, by = "day")
    weather <- NA
    foreach(day = days) %do% {
        dateString <- format(day,"%Y%m%d")
        dateStringFileName <- format(day,"%Y-%m-%d")
        filename <- paste0(file.path(data_folder, dateStringFileName), ".csv")
        ifelse(!file.exists(filename),
download.file(sprintf("https://www.wunderground.com/history/airport/KMDW/%s/DailyHistor
y.html?req_city=Chicago&req_state=IL&req_statename=Illinois&reqdb.zip=60290&reqdb.magi
c=1&reqdb.wmo=99999&format=1", format(day,"%Y/%m/%d")), filename)
, FALSE)
processedData <- read.csv(filename)[, 1:2]
        colnames(processedData) <- c("TimeCST","TemperatureF")
processedData$TimeCST <- paste0(dateStringFileName," ",processedData$TimeCST)
        processedData$TimeCST <- format(strptime(processedData$TimeCST,
"%Y-%m-%d %I:%M %p", timezone),"%Y-%m-%d %H")
        processedData = processedData[!duplicated(processedData[,1]),]

ifelse(is.na(weather), weather <- processedData, weather <- rbind(weather, processedData))
        rm(processedData)
    }
weather
}

weather <- getWeatherDataForChicagoIn2014()
write.csv(weather, "weather_chicago_2014.csv")

data$TimeCST<-format(strptime(data$starttime ,format="%m/%d/%Y %H:%M"), "%Y-%m-%d
%H")
data <- merge(data, weather, by="TimeCST", all.x = TRUE)

processNATemperature<- function(col){
condition <- (!is.na(col) & col!=-9999)
idx <- c(0, which(condition))[cumsum(condition) + 1]
return(col[idx])
}
data$TemperatureF <- processNATemperature(data$TemperatureF)
```

Some commons map create functions

```
qmap("Prince St & Mercer St, New York City", zoom = 16, maptype="hybrid")+
  geom_point(aes(x=lon, y=lat, color=name), data=tmp,  size=5)+
  theme(legend.title=element_blank()) # turn off legend title


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

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

barplot(table(data$usertype, data$season),  beside=T, col=heat.colors(2),
xlab="Customer/Subscriber - Season ", ylab="Number of Trips", legend=T, ylim=c(0,800000))

qplot(testset$usertype,  data=testset, geom="bar")

qplot(data$hour,  data=data, geom="bar", ylab="Number of trips", fill=usertype, xlab="Hour of
Day")

qmap("Chicago", zoom = 12)+
  geom_point(aes(x=fromTable$longitude,
y=fromTable$latitude,colour=fromTable$Freq,fill=fromTable$Freq), data=fromTable,
size=fromTable$Freq/10000)

qplot(newdata$season,newdata$TemperatureF,  data=newdata, geom = "point",
colour=season, xlab="Temperature(F)", ylab="Trip duration")
```