Blue Bikes is a bicycle sharing system in the Boston, Massachusetts. The bikes sharing program started on 28 July 2011. This program aimed for individuals to use it for short-term basis for a price. The program allows individuals to borrow a bike from a dock station and retrun to another dock station after using it. According to 2017 [report](https://www.bostonglobe.com/business/2017/05/25/hubway-set-for-big-expansion-and-year-round-service/TIfhZ7eZmzU3OE4cDayNzL/story.html), there are more than 1600 bicycles in used.

The aims of the current project are to gain insight of the user who used this program such as gender. It also aimed to investigate the duration of the usage. The dataset used in this project can be found in this [link](https://www.bluebikes.com/system-data). For the purpose of this project, two datasets are used the station dataset as recorded in 16th August 2017 and the trip data on 8th August 2018.

library(data.table) # for faster loading on large dataset

library(tidyverse) # load ggplot, dplyr

library(leaflet) # interactive mapping

library(leaflet.extras) #extra mapping for leaflet

library(geosphere) #used to calculate distance

library(lubridate) #formatting date and time

library(gridExtra) # multiple plot arrange

library(grid) #multiple plot arrange

library(psych) #to get describe function

#load the data set

data <- as.data.frame(fread("201807-bluebikes-tripdata.csv"))

#load station data

station <- read.csv("Hubway\_Stations\_as\_of\_July\_2017.csv")

Before the analysis, let's visualize all the dock stations around Boston.

map1 <- station %>%

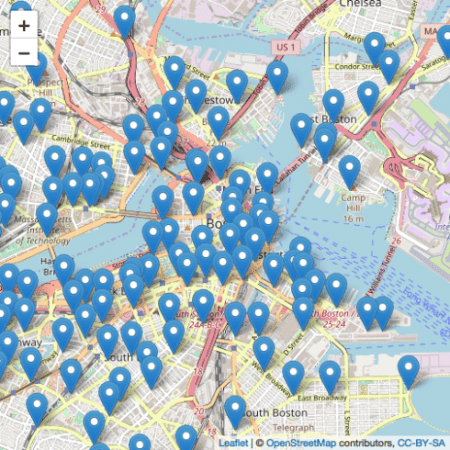
leaflet() %>%

setView(lng = -71.0589, lat = 42.3601, zoom = 13) %>%

addTiles() %>%

addMarkers(lng = station$Longitude, lat = station$Latitude, popup = station$Station)

map1

[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-2-1.png?ssl=1)

As you can see above, these are the distribution of dock station around Boston. You can run the code in RStudio and interact with the map.

**Data Cleaning**

glimpse(data)

*## Observations: 242,891*

*## Variables: 15*

*## $ tripduration 220, 6578, 1311, 831, 1318, 609, 523...*

*## $ starttime "2018-07-01 00:01:06.5090", "2018-07...*

*## $ stoptime "2018-07-01 00:04:46.9020", "2018-07...*

*## $ `start station id` 11, 26, 20, 116, 194, 75, 149, 75, 7...*

*## $ `start station name` "Longwood Ave at Binney St", "Washin...*

*## $ `start station latitude` 42.33863, 42.34152, 42.35991, 42.370...*

*## $ `start station longitude` -71.10650, -71.06892, -71.05143, -71...*

*## $ `end station id` 45, 26, 61, 117, 100, 179, 104, 179,...*

*## $ `end station name` "Jersey St. at Boylston St.", "Washi...*

*## $ `end station latitude` 42.34468, 42.34152, 42.34876, 42.366...*

*## $ `end station longitude` -71.09785, -71.06892, -71.08238, -71...*

*## $ bikeid 1188, 2020, 2875, 2334, 2813, 2152, ...*

*## $ usertype "Subscriber", "Subscriber", "Subscri...*

*## $ `birth year` 1969, 1969, 1961, 1990, 1990, 1993, ...*

*## $ gender 0, 0, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, ...*

A quick overview of the dataset shows that there is 242,891 observation of the dat and 15 variables. The date range from 1st July 2018 to 31st July 2018. In addition, the data is not as clean as it looks. First, the trip duration is caluculated in second which we need to convert into minutes. Next, only birth year gave which we need to perform a simple calculation to calculate users' age. Besides that, gender variable is dummy coded and the data dictionary does not explain what does each value mean. Therefore, I'm going to code them based on assumption (0 = Female, 1 = Male, 2 = Prefered not to say). Although the data dictionary explained that there are different user types such as Casual = 24-Hour or 72-Hour Pass user; Member = Annual or Monthly Member. But, all have been coded into one type which is a subscriber. This does not provide any use for the analysis. Lastly, I'm going to calculate the trip distances based on the start and end point of longitude and latitude using *haversine* formula. The formula is given below:

\(a = \sin^2(\frac{\bigtriangleup \varphi}{2} + \cos \varphi\_l \* \cos \varphi\_2 \* \sin^2(\frac{\bigtriangleup \lambda}{2}))\)

\(c = 2 \* a \tan2 (\sqrt{a}, \sqrt{(1 – a)}\)

\(d = R \* c\)

where \(\varphi\) is latitude, \(\lambda\) is longitude, R is earth's radius which is 6371. This formula calculates the distance into the meter. To convert it into miles changes the earth's radius to 3958.756. The details of the formula can be seen [here.](https://www.movable-type.co.uk/scripts/latlong.html). The distance calculated here might not be accurate because haversine formula is calculated based on shortest distance without considering the slope, hill and other factors.

#convert second to minutes and round it non decimal points

data$minutes <- round(data$tripduration/60, 0)

#calculate the age from birth year

data$age <- 2018 - data$`birth year`

#recoding gender variables

data$gender <- as.factor(recode(data$gender, "0" = "Female", "1" = "Male", "2" = "Prefer not to say"))

#convert the time and date variable to correct format

data$starttime <- ymd\_hms(data$starttime)

data$stoptime <- ymd\_hms(data$stoptime)

#separate date and time

data <- data %>%

mutate\_at(vars(starttime), ymd\_hms) %>%

mutate\_at(vars(starttime), funs("start\_date" = date(.)))

data <- data %>%

mutate\_at(vars(stoptime), ymd\_hms) %>%

mutate\_at(vars(stoptime), funs("stop\_date" = date(.)))

#extracting day, weekday and hour

data$day <- day(data$starttime)

data$weekday <- wday(data$starttime, label = TRUE)

data$hour <- hour(data$starttime)

# writing function for the distance formula

dist <- function(long1, lat1, long2, lat2) {

R <- 6371

radian <- pi / 180

a1 <- lat1 \* radian

a2 <- long1 \* radian

b1 <- lat2 \* radian

b2 <- long2 \* radian

diff\_long <- b2 - a2

diff\_lat <- b1 - a1

a <- (sin(diff\_lat/2))^2 + cos(a1) \* cos(b1) \* (sin(diff\_long/2))^2

c <- 2 \* atan2(sqrt(a), sqrt(1 - a))

d <- R \* c

return(d)

}

# return the calculation into 2 decimals and data frame

a <- as.data.frame(round(dist(data$`start station longitude`, data$`start station latitude`, data$`end station longitude`, data$`end station latitude`),2))

#binding two data frame

new\_data <- cbind(data, a)

#select relevant variable

bike\_data <- select(new\_data, "tripduration", "start\_date", "stop\_date", "start station latitude", "start station longitude", "end station latitude", "end station longitude", "gender", "minutes", "age", "round(dist(data$`start station longitude`, data$`start station latitude`, data$`end station longitude`, data$`end station latitude`), 2)", "day", "weekday", "hour")

#rename variabe

colnames(bike\_data) <- c("tripduration", "start\_date", "stop\_date", "start\_lat", "start\_long", "end\_lat", "end\_long", "gender", "minutes", "age", "distances", "day", "weekday", "hour")

# convert into factor

bike\_data$weekday <- as.factor(bike\_data$weekday)

glimpse(bike\_data)

*## Observations: 242,891*

*## Variables: 14*

*## $ tripduration 220, 6578, 1311, 831, 1318, 609, 523, 564, 1499, ...*

*## $ start\_date 2018-07-01, 2018-07-01, 2018-07-01, 2018-07-01, ...*

*## $ stop\_date 2018-07-01, 2018-07-01, 2018-07-01, 2018-07-01, ...*

*## $ start\_lat 42.33863, 42.34152, 42.35991, 42.37080, 42.38614,...*

*## $ start\_long -71.10650, -71.06892, -71.05143, -71.10441, -71.0...*

*## $ end\_lat 42.34468, 42.34152, 42.34876, 42.36610, 42.39697,...*

*## $ end\_long -71.09785, -71.06892, -71.08238, -71.08639, -71.1...*

*## $ gender Female, Female, Male, Male, Male, Male, Male, Pre...*

*## $ minutes 4, 110, 22, 14, 22, 10, 9, 9, 25, 22, 5, 23, 3, 1...*

*## $ age 49, 49, 57, 28, 28, 25, 45, 26, 24, 26, 31, 28, 2...*

*## $ distances 0.98, 0.00, 2.83, 1.57, 3.87, 0.92, 1.86, 0.92, 3...*

*## $ day 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...*

*## $ weekday Sun, Sun, Sun, Sun, Sun, Sun, Sun, Sun, Sun, Sun,...*

*## $ hour 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...*

describe(bike\_data$age)

*## vars n mean sd median trimmed mad min max range skew*

*## X1 1 242891 35.65 11.81 32 34.91 11.86 16 132 116 0.64*

*## kurtosis se*

*## X1 -0.07 0.02*

A quick examination shows that there are many impossible values from the age where the maximum age is 132. By guessing, it maybe because some of the users prefered not to identify their age and choose the default setting from the Blue Bikes system. Therefore, any age above 99 will be removed.

bike\_data <- bike\_data[bike\_data$age < 99,]

**Exploratory Data Analysis**

First we are going to examine the Blue Bikes usage in July, 2018.

group <- bike\_data %>%

group\_by(start\_date) %>%

summarise(total = n())

group %>% ggplot(aes(x = start\_date, y = total)) +

geom\_line(lty = 1) +

geom\_point() +

labs(title = "Total of Blue Bikes Usage in July, 2018",

x = "Date", y = "Total of Blue Bikes Used") +

theme\_bw() +

annotate(geom="text", x = as.Date("2018-07-03"), y = 5000, label = "

mean = 7833.87\n

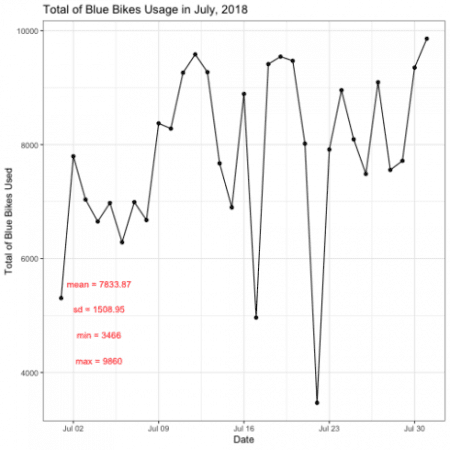
sd = 1508.95\n

min = 3466\n

max = 9860",

color = "red",

size = 3.5)

[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-7-1.png?ssl=1)

The plot above shows that total usage of the Blue Bikes on July 2018. It shows that there are two lowest usage of Blue Bikes during 17th and 22nd of July. Is there a reason why this is happening maybe due to weather, weekend or holiday?

Next, I'm going to examine which day have the most usage of the Blue Bikes.

days <- bike\_data %>%

group\_by(weekday, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-08"))

day2 <- bike\_data %>%

group\_by(weekday, start\_date) %>%

summarise(n = n()) %>%

filter(start\_date < as\_date("2018-07-15") & start\_date > as\_date("2018-07-07"))

day3 <- bike\_data %>%

group\_by(weekday, start\_date) %>%

summarise(n1 = n()) %>%

filter(start\_date < as\_date("2018-07-22") & start\_date > as\_date("2018-07-14"))

day4 <- bike\_data %>%

group\_by(weekday, start\_date) %>%

summarise(n2 = n()) %>%

filter(start\_date < as\_date("2018-07-29") & start\_date > as\_date("2018-07-21"))

p1 <- ggplot(days, aes(x = weekday, y = total)) +

geom\_col(fill = "red") +

geom\_label(aes(label = total), color = "black", hjust = 1) +

coord\_flip() +

labs(title = "From 1st July to 7th July",

x = "Weekday",

y = "Total") +

theme\_bw()

p2 <- ggplot(day2, aes(x = weekday, y = n)) +

geom\_col(fill = "red") +

geom\_label(aes(label = n), color = "black", hjust = 1) +

coord\_flip() +

labs(title = "From 8th July to 14th July",

x = "Weekday",

y = "Total") +

theme\_bw()

p3 <- ggplot(day3, aes(x = weekday, y = n1)) +

geom\_col(fill = "red") +

geom\_label(aes(label = n1), color = "black", hjust = 1) +

coord\_flip() +

labs(title = "From 15th July to 21st July",

x = "Weekday",

y = "Total") +

theme\_bw()

p4 <- ggplot(day4, aes(x = weekday, y = n2)) +

geom\_col(fill = "red") +

geom\_label(aes(label = n2), color = "black", hjust = 1) +

coord\_flip() +

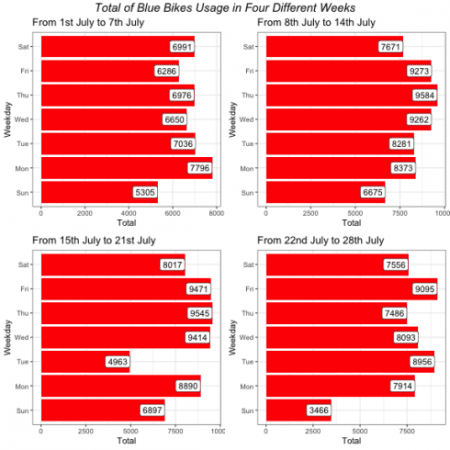
labs(title = "From 22nd July to 28th July",

x = "Weekday",

y = "Total") +

theme\_bw()

grid.arrange(p1,p2,p3,p4, top = textGrob("Total of Blue Bikes Usage in Four Different Weeks",gp = gpar(fontsize = 15, font = 3)))

[](https://i2.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-8-1.png?ssl=1)

As presented above, it shows that weekdays have the highest Blue Bikes usage while weekends have the lowest.

Next, I'm going to visulize hour that have most Blue Bikes usage.

Sun <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-02"))

Mon <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-03") & start\_date > as\_date("2018-07-01"))

Tue <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-04") & start\_date > as\_date("2018-07-02"))

Wed <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-05") & start\_date > as\_date("2018-07-03"))

Thu <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-06") & start\_date > as\_date("2018-07-04"))

Fri <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-07") & start\_date > as\_date("2018-07-05"))

Sat <- bike\_data %>%

group\_by(hour, start\_date) %>%

summarise(total = n()) %>%

filter(start\_date < as\_date("2018-07-08") & start\_date > as\_date("2018-07-06"))

m1 <- Sun %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Sunday", x = "Hour", y = "Total") +

theme\_bw()

m2 <- Mon %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Monday", x = "Hour", y = "Total") +

theme\_bw()

m3 <- Tue %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Tuesday", x = "Hour", y = "Total") +

theme\_bw()

m4 <- Wed %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Wednesday", x = "Hour", y = "Total") +

theme\_bw()

m5 <- Thu %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Thursday", x = "Hour", y = "Total") +

theme\_bw()

m6 <- Fri %>% ggplot(aes(x = hour, y = total)) +

geom\_line(lty = 2) +

geom\_point() +

labs(title = "Friday", x = "Hour", y = "Total") +

theme\_bw()

m7 <- Sat %>% ggplot(aes(x = hour, y = total)) +

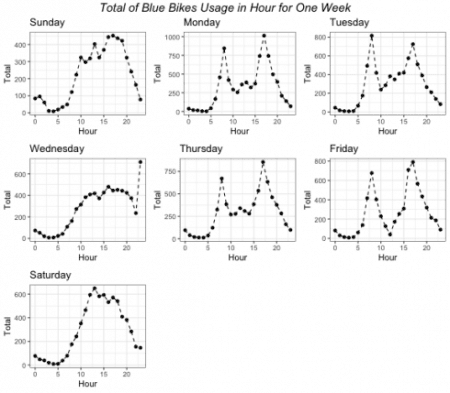
geom\_line(lty = 2) +

geom\_point() +

labs(title = "Saturday", x = "Hour", y = "Total") +

theme\_bw()

grid.arrange(m1,m2,m3,m4,m5,m6,m7, top = textGrob("Total of Blue Bikes Usage in Hour for One Week",gp = gpar(fontsize = 15,font = 3)))

[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-9-1.png?ssl=1)

As you can see in the plot, it shows that there is a pattern across different hour during weekdays and weekends. On weekends, Blue Bikes mostly likely to be used during the afternoon. However, during weekday Blue Bikes are most likely to be used on before and after working hour.

Next, I'm going to visualize the gender and age group most likely to use Blue Bikes.

gender <- bike\_data %>%

group\_by(gender) %>%

summarise(total = n()) %>%

filter(gender != "Prefer not to say")

gender %>% ggplot(aes(x = gender, y = total, fill = gender)) +

geom\_col() +

geom\_text(aes(label = total), vjust = -0.1) +

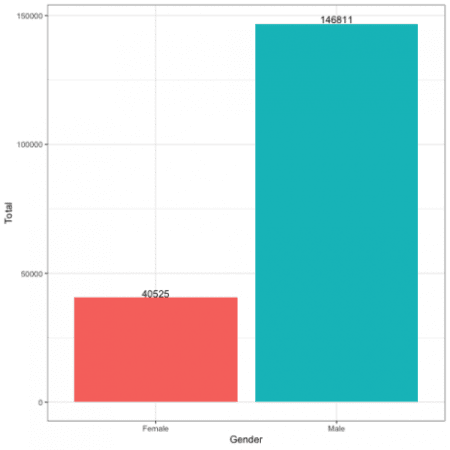
labs(main = "Gender on Blue Bikes Usage",

x = "Gender",

y = "Total") +

guides(fill = FALSE) +

theme\_bw()

[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-10-1.png?ssl=1)

Based on the observation, Male are more likely to used Blue Bikes compared to female in July 2018. However, this plot might not be true because we code the gender based on our assumption.

Let's test is there is a significant difference between gender and distance travel using Blue Bikes.

Before that, let's examine the distance.

ggplot(data = bike\_data, aes(x = distances)) +

geom\_density(alpha = 0.3, fill = "black") +

guides(fill = FALSE) +

labs(title = "Desntiy Plot of Distance",

x = "Distance",

y = "Density") +

annotate(geom ="text", x = 6, y = 0.3, label = "

mean = 1.97\n

sd = 1.37\n

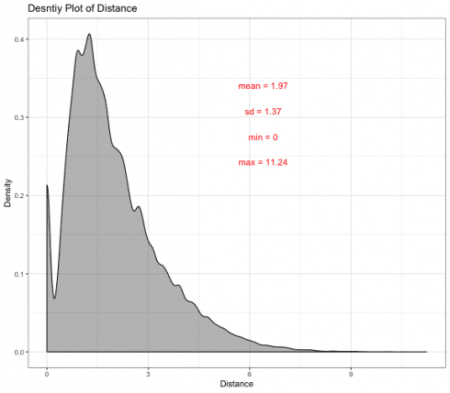
min = 0\n

max = 11.24",

color = "red",

size = 4) +

theme\_bw()

[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-11-1.png?ssl=1)

Based on the description analysis, it shows that the distance is not normally distributed and it is positively skewed. It was suggested that on average the travel distances of Blue Bikes are 1.97 meter and the maximum travel distance is 11.24 meter.

Let's examine the difference between male and female.

gender\_distance <- bike\_data %>%

group\_by(gender, distances) %>%

summarise(total = n()) %>%

filter(gender != "Prefer not to say")

gender\_distance %>% ggplot(aes(x = distances)) +

geom\_density(alpha = 0.3, fill = "black") +

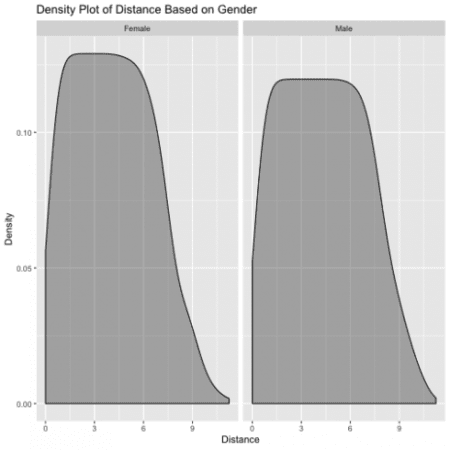
guides(fill = FALSE) +

labs(title = "Density Plot of Distance Based on Gender" ,

x = "Distance",

y = "Density") +

facet\_wrap(~ gender)

[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-12-1.png?ssl=1)

The plot shows that both male and female have a similar distribution, but this does not give much information about the differeces between gender on distance travel using Blue Bikes. I'm going to conduct a T-Test to test the hypothesis.

t.test(gender\_distance$distances ~ gender\_distance$gender)

*##*

*## Welch Two Sample t-test*

*##*

*## data: gender\_distance$distances by gender\_distance$gender*

*## t = -2.2447, df = 1610.1, p-value = 0.02492*

*## alternative hypothesis: true difference in means is not equal to 0*

*## 95 percent confidence interval:*

*## -0.52101431 -0.03509147*

*## sample estimates:*

*## mean in group Female mean in group Male*

*## 4.078028 4.356081*

The T-test shows that \(t\_{1610.1} = -2.2447\) with p = 0.025 which is lower than 0.05 suggested that there is a significant difference between male and female when using Blue Bikes. It was suggested that male are more likely to travel futther 0.28 meter compared to female.

gender\_distance %>%

ggplot(aes(x = gender, y = distances, fill = gender)) +

geom\_boxplot() +

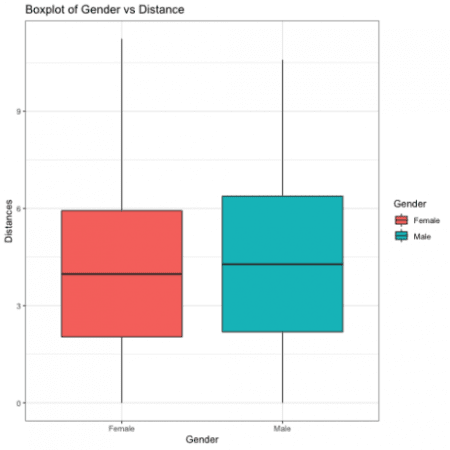
labs(title = "Boxplot of Gender vs Distance",

x = "Gender",

y = "Distances") +

guides(fill=guide\_legend(title = "Gender")) +

theme\_bw()

[](https://i2.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-14-1.png?ssl=1)

Next, I'm going to visualize the heatmap of starting location and ending location of using Blue Bikes.

#starting location

#mapping

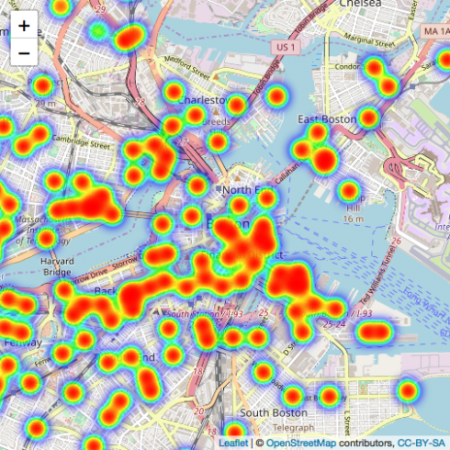
bike\_data %>%

leaflet() %>%

setView(lng = -71.0589, lat = 42.3601, zoom = 13) %>%

addTiles() %>%

addHeatmap(lng = bike\_data$start\_long, lat = bike\_data$start\_lat, max = 2, radius = 15)

[](https://i2.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-15-1.png?ssl=1)

#ending location

#mapping

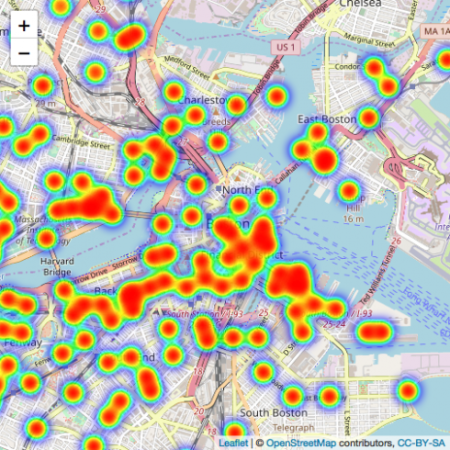
bike\_data %>%

leaflet() %>%

setView(lng = -71.0589, lat = 42.3601, zoom = 13) %>%

addTiles() %>%

addHeatmap(lng = bike\_data$end\_long, lat = bike\_data$end\_lat, max = 2, radius = 15)

[](https://i2.wp.com/datascienceplus.com/wp-content/uploads/2018/09/unnamed-chunk-16-1.png?ssl=1)

**Conclusion**

This analysis is just my personal interest and trying to mapping using leaflet.