**Script:**

LAFinal <- read.csv(file.choose())

DCFinal <- read.csv(file.choose())

CGFinal <- read.csv(file.choose())

CGFinal <- read.csv(file.choose())

DCFinal <- read.csv(file.choose())

LAFinal <- read.csv(file.choose())

View(LAFinal)

#Taking count of all values present in the LA dataset

nrow(LAFinal) #93378 values

#Since we have a primary key of trip id, we try and see if we have any duplicate values and dedupe the #dataset by using the dedupe() function and then calculating the final count of all values present

Deduped <- unique(LAFinal)

nrow(Deduped) #93105 values left

#We have to take date from the date-time stamp separately for hourly and daily analysis. For this we have used as.Date and lubridate packages

LAFinal$Date<-as.Date(LAFinal$Start.date, format = "%m/%d/%Y")

CGFinal$Date<-as.Date(CGFinal$Start.date, format = "%m/%d/%Y")

DCFinal$Date<-as.Date(DCFinal$Start.date, format = "%m/%d/%Y")

LAFinal$WDAY<-wday(LAFinal$Date, abbr = TRUE, label = TRUE)

DCFinal$WDAY<-wday(DCFinal$Date, abbr = TRUE, label = TRUE)

CGFinal$WDAY<-wday(CGFinal$Date, abbr = TRUE, label = TRUE)

LAFinal$WDAY<-Hour(LAFinal$Start.date)

DCFinal$WDAY<-Hour(DCFinal$Start.date)

CGFinal$WDAY<-Hour(CGFinal$Start.date)

#The following plot function uses ggplot to derive a relation between the hours of the day the bikes are #operated and considered months to check if there is any difference of usage in these months

ggplot(LAFinal, aes(LAFinal$Month, LAFinal$Hour))+ geom\_count(col="maroon",show.legend = F) + labs(x="Month", y='Hours of Day', title = "Monthly Bike Usage - Hourly Distribution")

#Our first research question for the project is :

#Q1. How can Seattle create the best bike share system based on trends from other cities?

#For the above question, we try to find relationships between various factors such as trip duration, plan #holder types, busiest days of week and busiest stations.The following plot function uses ggplot to derive #a relation between the hours of the day the bikes are operated and days of the week to see a relation #between them and try to find a pattern or relation between the trip duration and the route trip to #understand the which users have round trips and how many have one way trips w.r.t the trip duration #undertaken

ggplot(LAFinal,aes(LAFinal$Trip.route,LAFinal$Duration))+geom\_jitter(col="#008081",show.legend = F) + labs(x="Route Type", y='Trip Duration’', title = "Trip Duration and Route Category relation")

ggplot(LAFinal,aes(LAFinal$Trip.route,LAFinal$Duration))+geom\_jitter(col="#008081",show.legend = F) + labs(x="Route Type", y='Trip Duration’', title = "Trip Duration and Membership Relationship")

#we try and analyze if there is any relation between busy hours of the day and busy days of the week and #to understand if any of the relationships are consistent or exponential in nature.

ggplot(LAFinal, aes(LAFinal$WDAY, LAFinal$Hour))+ geom\_count(col="maroon",show.legend = F) + labs(x="Day of Week", y='Hour of Day', title = "Weekly Bike Usage - Hourly Distribution")

#Q2. Should Seattle be proposed dockless or docked bicycles?

#Next we try to decipher a pattern between the trip route category and passholder type. This would help #us understand if flexible, weekly or daily passholders use the bicycles and how often are these trips #made. Are they round trips or one way trips? If one way trips are more, we would propose docked #systems.

ggplot(LAFinal,aes(LAFinal$trip\_route\_category,LAFinal$passholder\_type))+geom\_count(col="salmon",show.legend = F) + labs(x="Route Category", y="Pass Holder", title = "Trip Duration Based on Passholder Type")

#Q3. How do busy hours affect the Bicycle patterns?

#For this question and the sheer vastness of the probably factors, we have at first taken the summary() #function to find the busiest start and end stations for LA to determine the stations which need to be most #populated with bicycles.

summary(CGFinal)

#Following are additional plots for contextual research and analysis of the system. Overall, the following plots are responsible for relations between

#member type, total bike usage, route type, gender distribution and age distribution along with relation between ridership and temperature as well

ggplot() + geom\_col(data=LAFinal, aes(x=LAFinal$Date, y = LAFinal$Duration/360), color="darkcyan") + geom\_line(data=LAWeather, aes(x=LAWeather$Date, y = LAWeather$Avg.Temp),size =2, color="red") + labs(x="Month", y="Duration of Trips/ Avg. Temperature", title = "LA - Total Bike Usage vs Temperature Distribution over Quarter 2")

ggplot() + geom\_col(data=DCFinal, aes(x=DCFinal$Date, y = DCFinal$Duration/1800), color="darkcyan") + geom\_line(data=DCWeather, aes(x=DCWeather$Date, y = DCWeather$Avg.Temp),size =2, color="red") + labs(x="Month", y="Duration of Trips/ Avg. Temperature", title = "DC - Total Bike Usage vs Temperature Distribution over Quarter 2")

ggplot() + geom\_col(data=CGFinal, aes(x=CGFinal$Date, y = CGFinal$Duration/1800), color="darkcyan") + geom\_line(data=CGWeather, aes(x=CGWeather$Date, y = CGWeather$Avg.Temp),size =2, color="red") + labs(x="Month", y="Duration of Trips/ Avg. Temperature", title = "CG - Total Bike Usage vs Temperature Distribution over Quarter 2")

ggplot(LAFinal, aes(LAFinal$Member.type)) + geom\_bar(fill = "#008081") +labs(title = "LA - Bike Usage by Customer Type", x="Customer Type", y="Usage Count")

ggplot(DCFinal, aes(DCFinal$Member.type)) + geom\_bar(fill = "#008081") +labs(title = "DC - Bike Usage by Customer Type", x="Customer Type", y="Usage Count")

ggplot(CGFinal, aes(CGFinal$Member.type)) + geom\_bar(fill = "#008081") +labs(title = "CG - Bike Usage by Customer Type", x="Customer Type", y="Usage Count")

ggplot(CGFinal, aes(CGFinal$Gender)) + geom\_bar(fill = "#008081") +labs(title = "CG - Bike Usage by Customer Gender", x="Gender", y="Usage Count")

ggplot(LAFinal,aes(LAFinal$Trip.route,LAFinal$Member.type))+geom\_jitter(col="#008081",show.legend = F, size = 0.5) + labs(x="Route Category", y="Pass Holder", title = "LA - Membership Distribution based on Route")

ggplot(LAFinal, aes(LAFinal$Month, LAFinal$Hour))+ geom\_count(col="maroon",show.legend = F) + labs(x="Month", y='Hours of Day', title = "Monthly Bike Usage - Hourly Distribution")

ggplot(DCFinal, aes(DCFinal$WDAY, DCFinal$Hour))+ geom\_count(col="#008081",show.legend = F) + labs(x="Day of Week", y='Hour of Day', title = "DC - Weekly Bike Usage - Hourly Distribution")

ggplot(LAFinal, aes(LAFinal$WDAY, LAFinal$Hour))+ geom\_count(col="#008081",show.legend = F) + labs(x="Day of Week", y='Hour of Day', title = "LA - Weekly Bike Usage - Hourly Distribution")

ggplot(CGFinal, aes(CGFinal$WDAY, CGFinal$Hour))+ geom\_count(col="#008081",show.legend = F) + labs(x="Day of Week", y='Hour of Day', title = "CG - Weekly Bike Usage - Hourly Distribution")

DCBikesAvailable<-length(unique(DCFinal$Bike.number))

LABikesAvailable<-length(unique(LAFinal$Bike.number))

CGBikesAvailable<-length(unique(CGFinal$Bike.number))

#Summary is used to find most popular spots in the cities and google maps are used to plot these spots for visualization

summary(LAFinal)

#Top 5 popular start stations :

#Ocean Front Walk & Navy

#7th & Flower

#Ocean Front Walk & North Venice

#Union Station West Portal

#Main & 1st

#Top 5 popular end stations :

#Downtown Santa Monica Expo Line Station

#Ocean Front Walk & Navy

#7th & Flower

#Union Station West Portal

#Ocean Front Walk & North Venice

#Similarly we did this for CG and DC as well

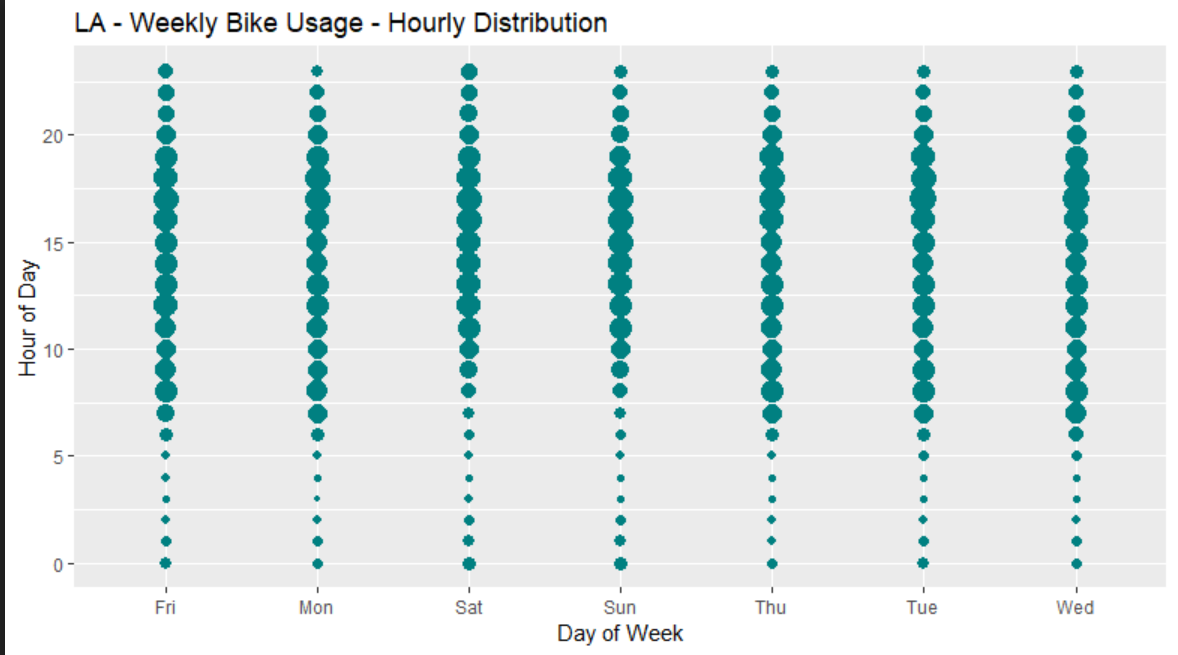
#Following link has the plots that we have generated using these functions and scripts:

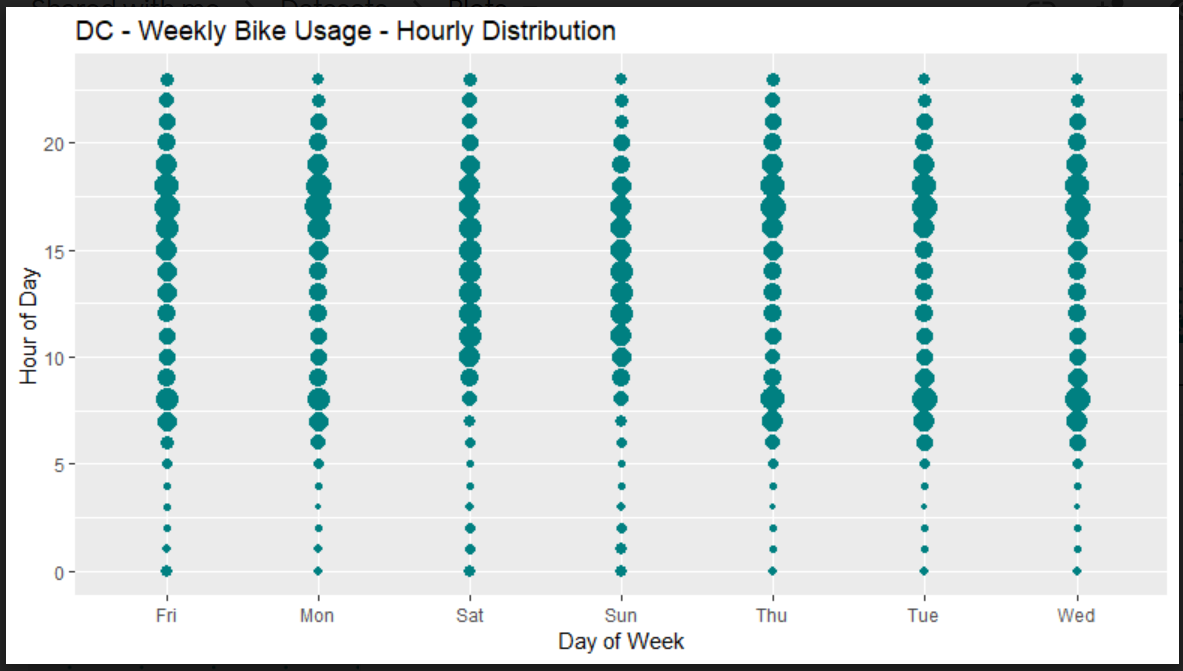
# https://drive.google.com/open?id=174rK-AyEufKip4\_zndvrjifEuc1t5P4n

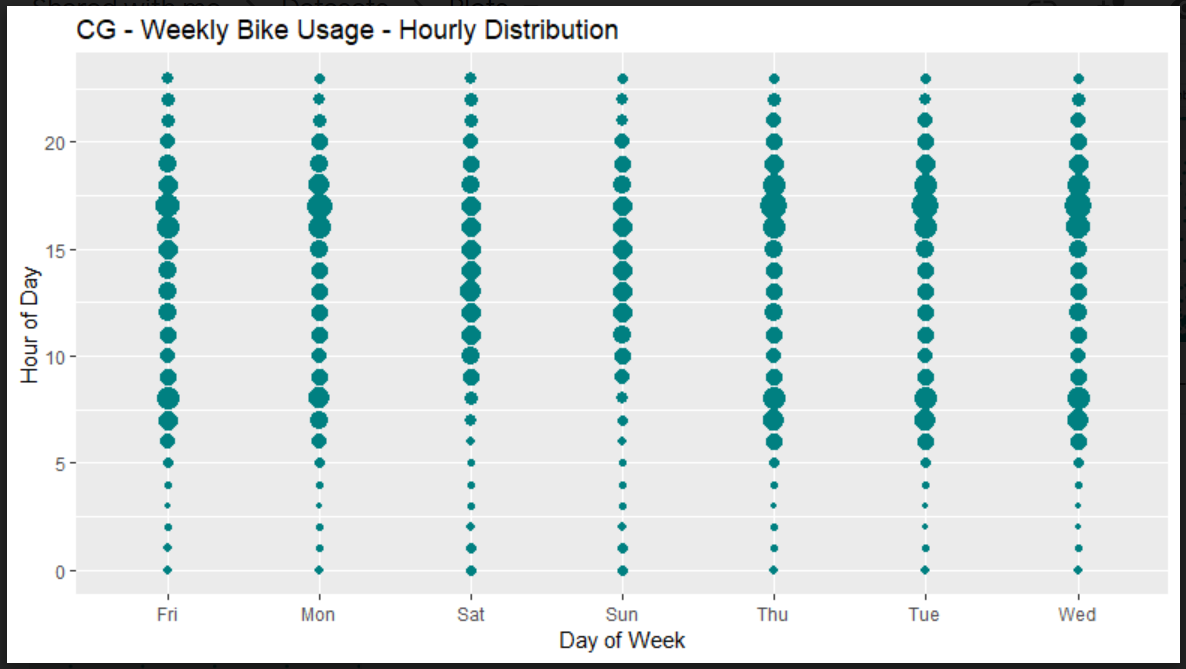
#This link is accessible by UMD IDs only

**Plots and Interpretations:**

#Following are the plots generated for Bike Usage comparing the hour of the day and Day of the week for LA, Chicago and DC.

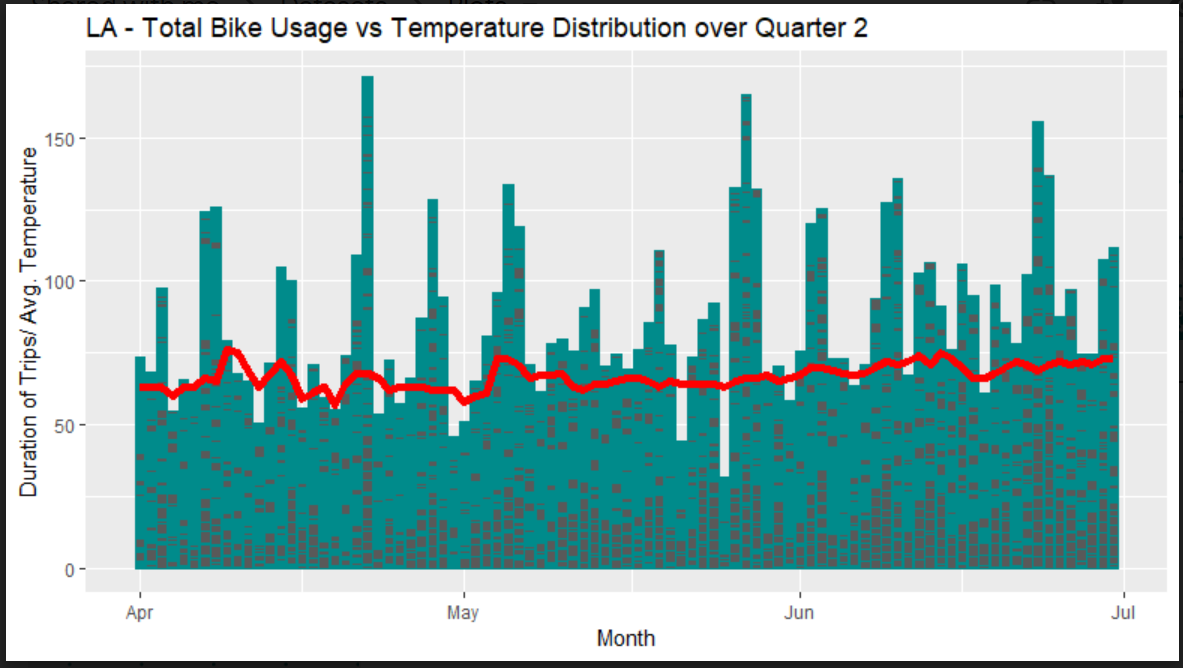


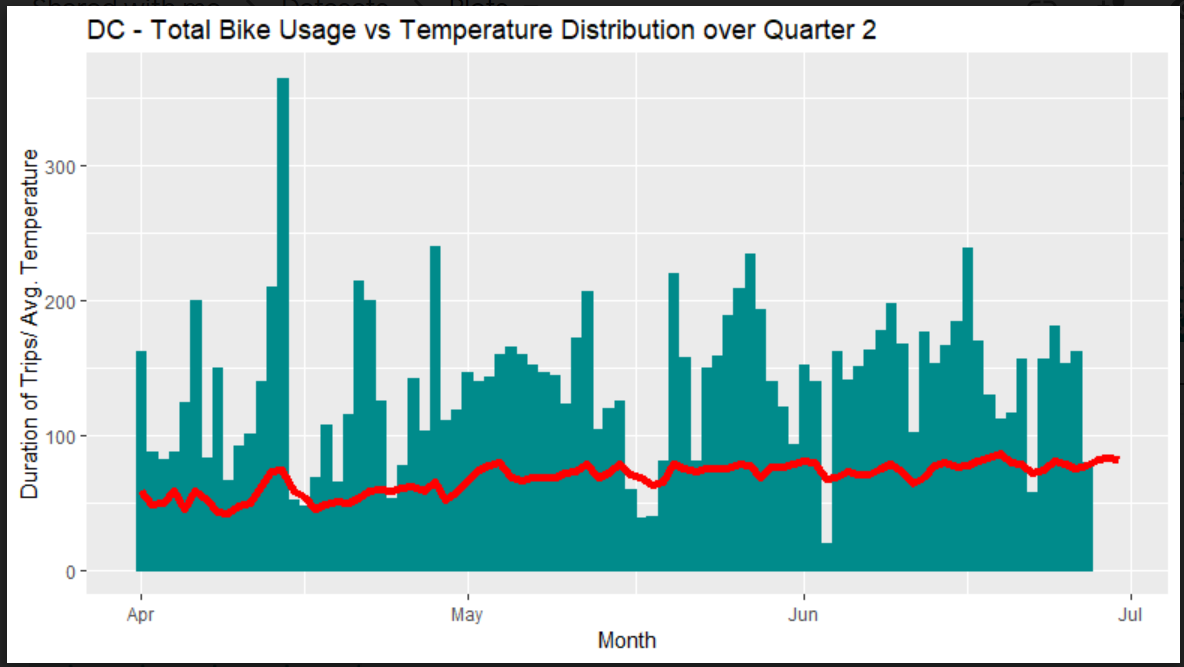


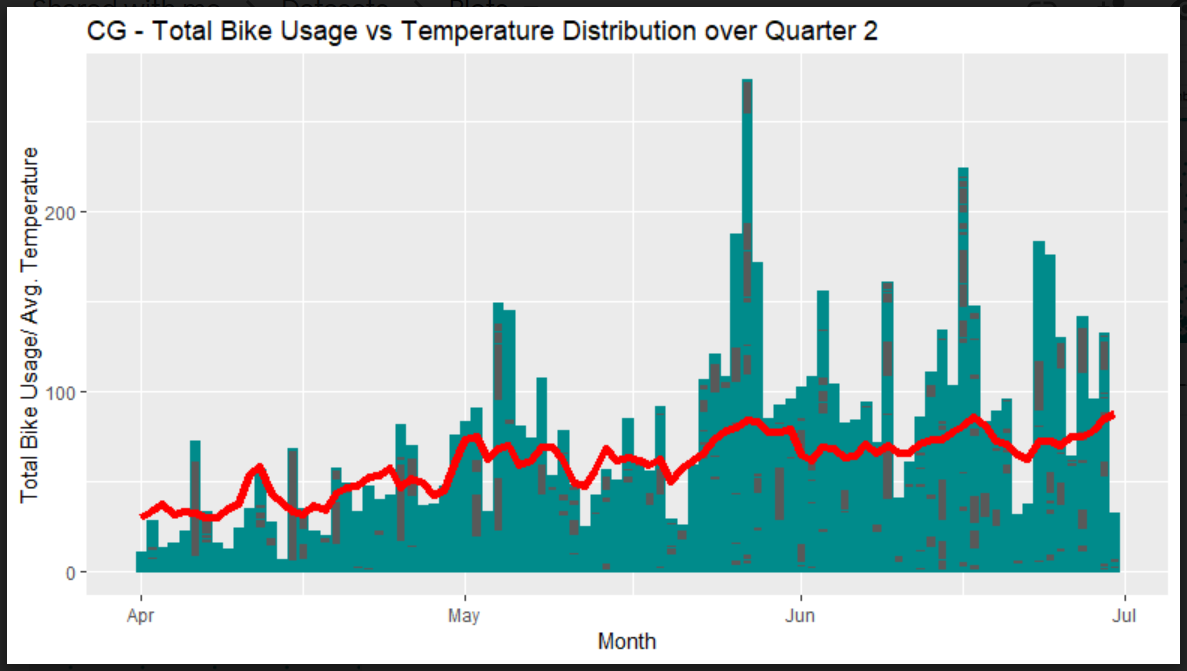


We observe here that during weekdays (Mon – Fri), the bike usage is at peak during 6am-10am and again 3pm-8pm. These are office start and end times, by which we can safely say working class people use these bikes. On weekends, the rush hours are from 9am-5pm, which might suggest people are using these bikes to visit or roam the city hence bikes are also used by tourists.

#Plots for Bike Usage and Temperature Distribution

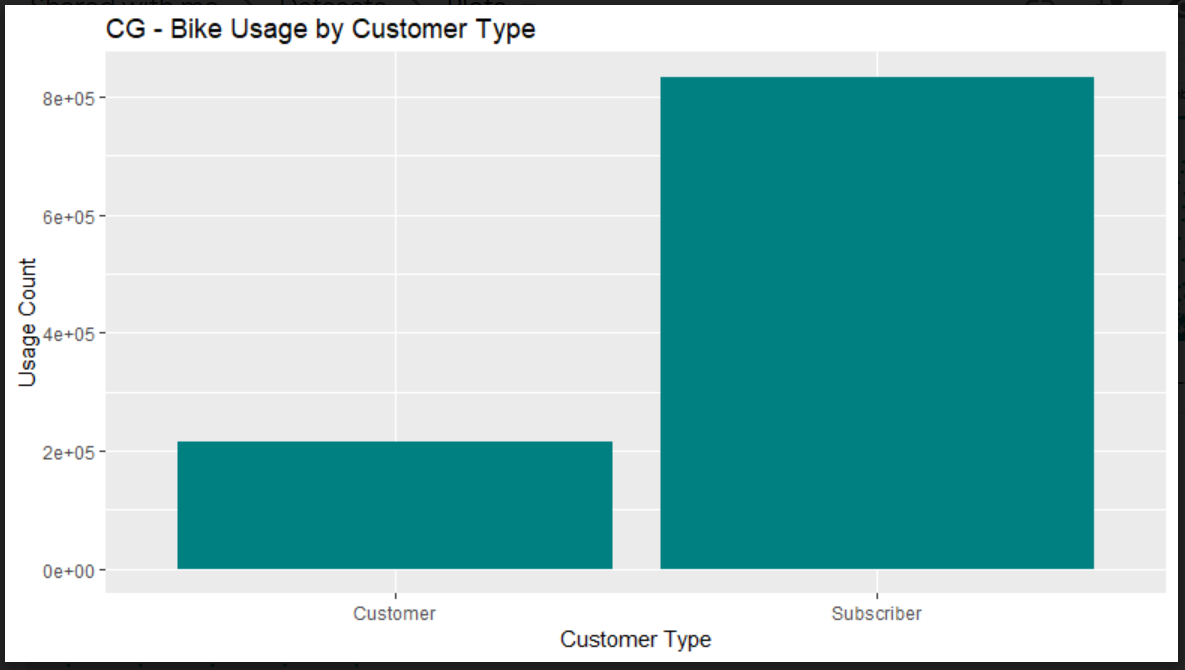


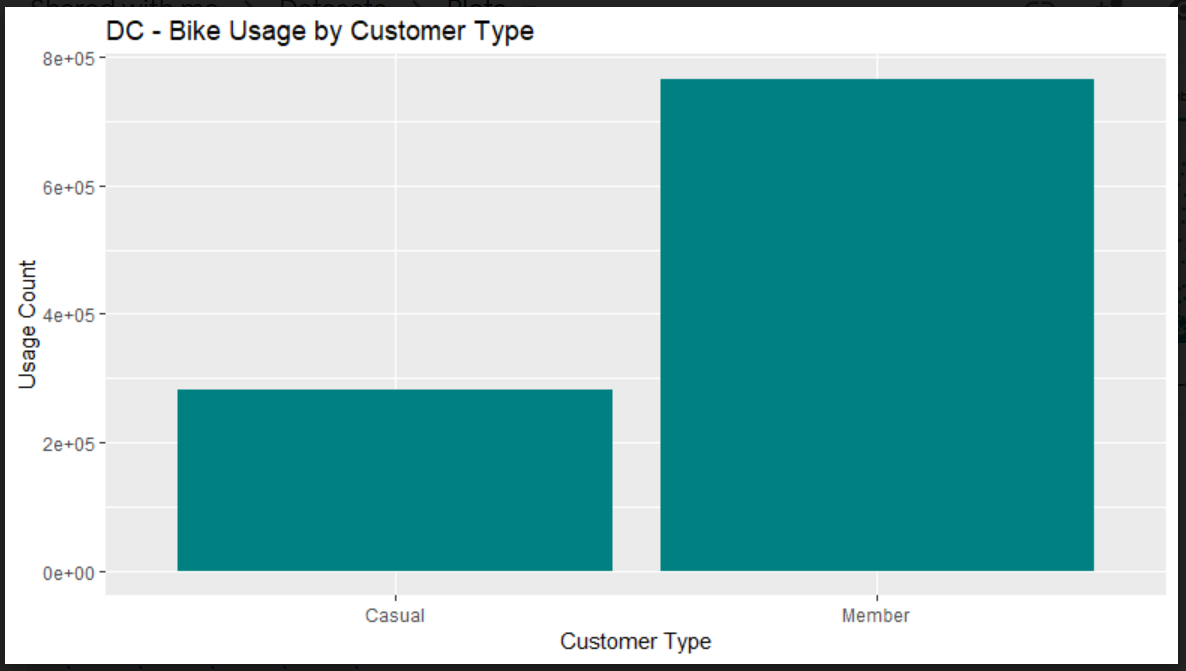


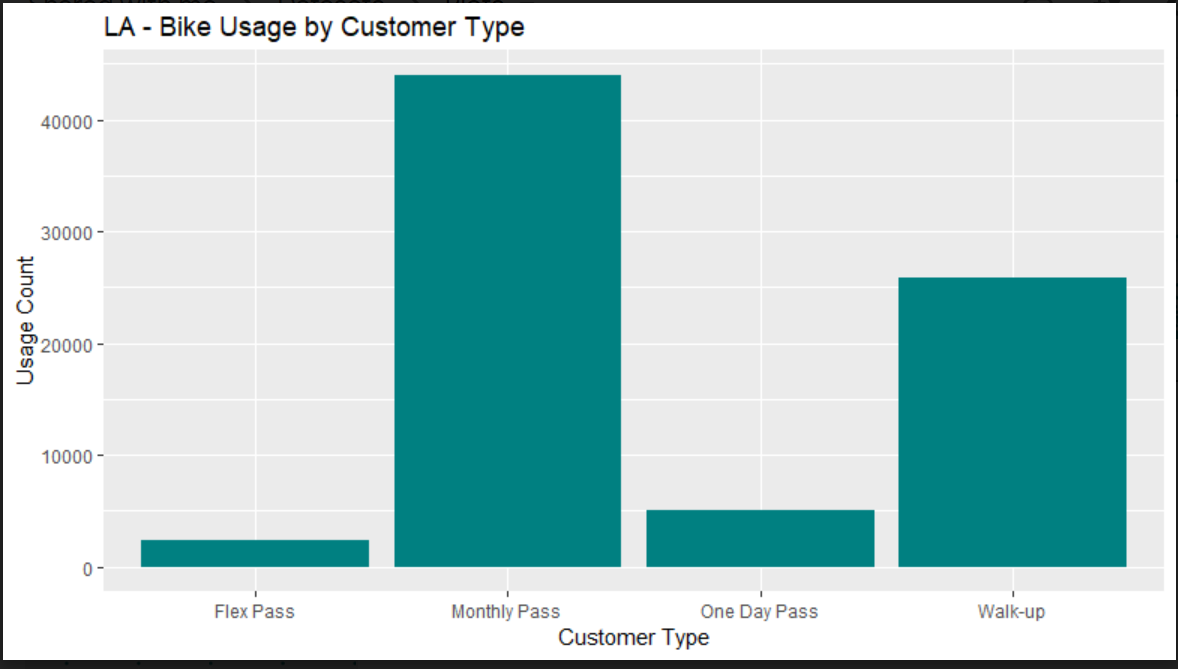


We observe here that in every city, as the temperature dips (below 50 degrees mostly), the bike usage drops too.

#Plots for Bike Usage by Customer Type

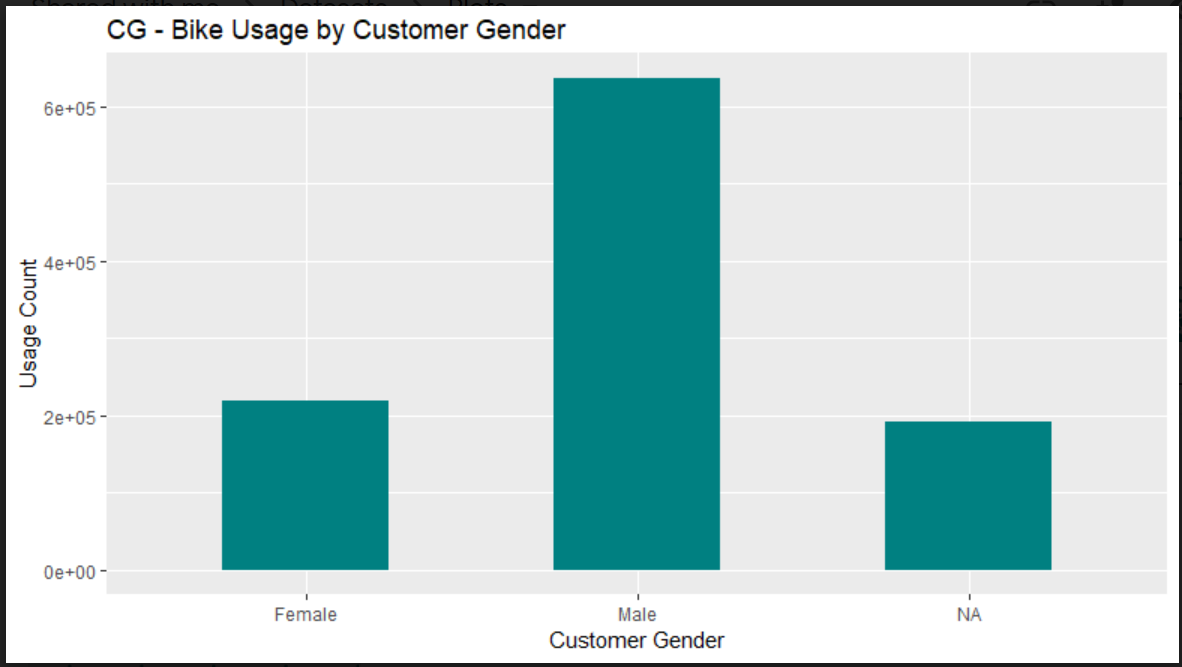






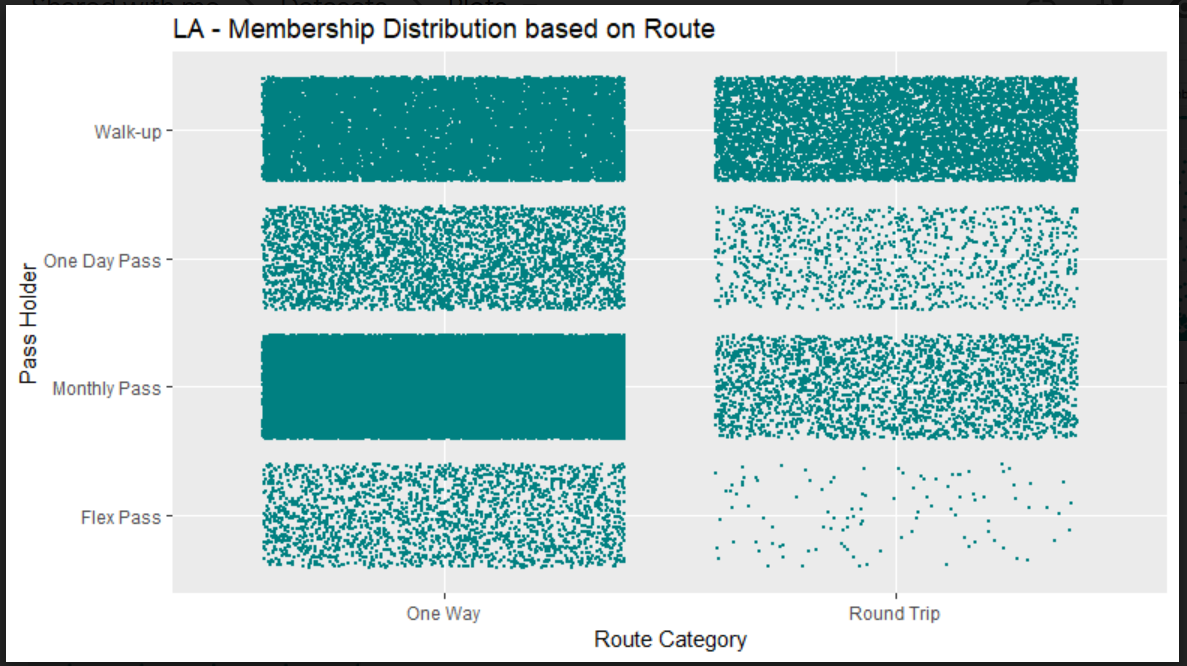
We can interpret from these graphs that users who are members do more frequent trips than users who are not members.

#Plots for Customer Gender and Bike Usage for Chicago since other Datasets didn’t have genders.

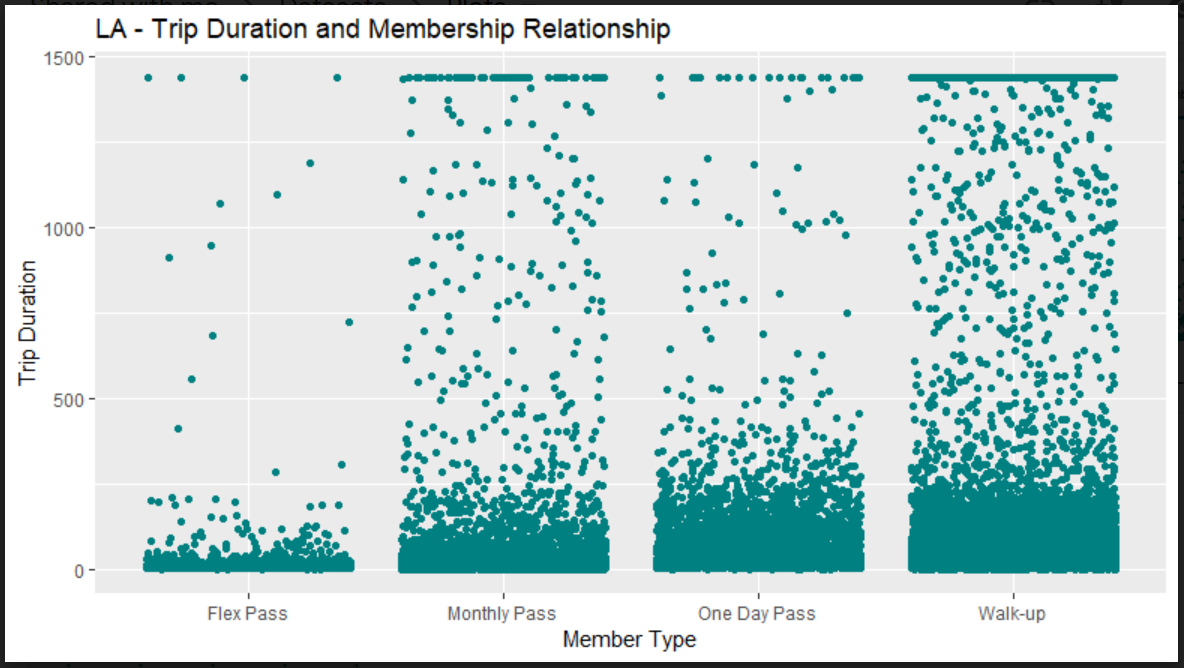


#Here we see that most users are male who are doing the longest trips, followed by Females

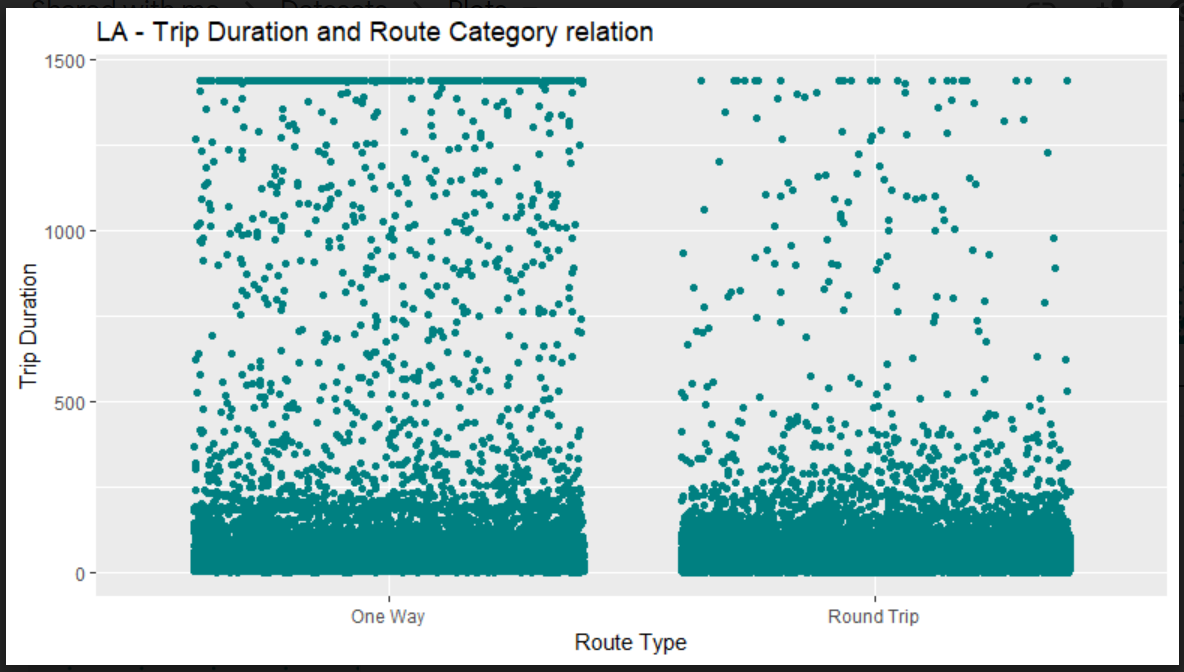
#Plots for LA (since it had most of the data required) for interpretations:



We see here that most trips made are one-way trips instead of round trips and they are mostly made by Monthly Pass Holders., followed by Walk-ups which are our second leading users and the users who do make a round trip unlike most other users.



We interpret here that the longest trip duration is taken up walk ups, which as we determined above also do round trips, hence the coagulation of trips at the top. This is followed by Monthly pass users, who do round trips but rarely, followed by One day passes and Flex passes.



Here we interpret that most trips are One Way trips and they might be of higher duration than round trips, implying the users taking these trips are travelling longer distances, increasing usage.