STAT 3355 Team Report

October 31st, 2019

Group Name: The Tandem Bikers

Group Members:

Laila Kamel, Jayden Cellers, Anya Sagues

Semester: Fall 2019

Professor: Dr. Qiwei Li

Table of Contents

[**Introduction**](#_29fz0r3xax5n) **3**

[**Questions and Findings**](#_207efvzxnu2) **4**

[**Plots**](#_fyp7itic5ku0) **6**

[**Conclusion**](#_214dlidl2qnu) **14**

[**References**](#_ba6lysuca6sh) **15**

[**Appendix**](#_2ybtngmczd8l) **16**

# 

# **Introduction**

The dataset used in this project was the Bike Sharing Dataset from the UCI Machine Learning Repository. This dataset contains the hourly and daily count of rental bikes between the years 2011 and 2012 in Washington D.C., USA with the corresponding weather and seasonal information. The variables in this dataset included: the record index, the date, the season, the year, month, hour, whether the day was a holiday or not, the day of the week, if it’s a workday, the weather, the actual temperature, the realfeel temperature, humidity, wind speed, the amount of casual users (logging in to the service as ‘guest; typically tourists), registered users (typically locals) and total users (a combination of registered and casual users).

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, the user is able to easily rent a bike from a particular location and leave it at another location. Currently, there are over 500 bike-sharing programs around the world, which is composed of over 500,000 bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns the bike sharing system into a network that can be used for accurately detecting mobility in the city.

Especially with similar services like *Lime* and *Bird* in the metropolitan areas of Dallas and Austin, we were curious to see what information this dataset held. We chose to explore which different environmental aspects affected the amount of users at a certain moment. We looked at factors like weather, time of day, time of year, and day of the week to infer what exactly drove increases and decreases in rental bike usage. With the conclusions from the graphs, we were hoping to discover what other attributes of the data we could further explore and what could be done to further increase the use of these bike sharing systems.

# 

# **Questions and Findings**

We began by cleaning the data set in order to create some great plots and come to some interesting findings and conclusions. Many of the time variables were initially given as numeric values to represent different years or seasons, so we made all of these types of variables ordered factor variables. Additionally, many of the temperature variables were given as normalized values, so we adjusted these values to be numeric digits that are easier to understand and analyze. We also made a variable called ‘time\_of\_day’ to group the hours of the day into four time chunks: “Early” (Midnight - 6:00 AM), “AM” (6:00 AM - Noon), “PM” (Noon - 6:00 PM), “Late” (6:00 PM - Midnight), and thus make it easier to observe trends in bike usage across different times in a day.

The questions we used to drive our investigation are as follows:

1. Are registered users using bikes more frequently during the weekdays (to commute, potentially)?
2. Do casual users use bikes more frequently on weekends?
3. How does bike usage compare between casual and registered users?
4. Do bikes get used more during the afternoon (warmest time of day)?
5. Which day of the week has the heaviest bike usage?
6. Which time period of the day has the heaviest bike usage?
7. What could be some possible explanations for specific days/times of the day having increased usage?
8. Does usage go down during inclement weather?
9. Is there a significant difference in usage between clear days and light precipitation days?
10. How does casual user usage compare to itself on sunny days versus rainy days?

The first interesting finding comes from **Figure 1**, the Dumbbell Chart. We found that there was a significant increase in the total number of users from 2011 to 2012 across the entire dataset. This Figure plots the average number of total users in an hour in each month in 2011 (orange) and in 2012 (green) for each month. Every single month saw an increase in the average of at least 50 users per hour.

The next interesting finding comes from **Figure 2**, the Calendar Heat Map. Using the new ‘time\_of\_day’ variable, the days of the week, and the seasons, we observed that weekdays in the Summer during the afternoon had the highest bike usage. This is logical because this time everyone is getting off of work, school is out of session, and the weather is warm.

Our third interesting finding comes from **Figure 3**, the Population Pyramid. We compared and contrasted the usage of Registered and Casual Users on different days of the week. Within the realm of Registered Users, there is significantly more usage on weekdays than on weekends. This suggests that registered users with this bike sharing service could be using the shared bikes for commuting to and from work. Within the realm of Casual Users, there is much heavier usage on the weekends. Thus, users that function as ‘guests’ of this service may be using these bikes in a social context. And overall, there is much heavier usage by Registered Users than by Casual Users.

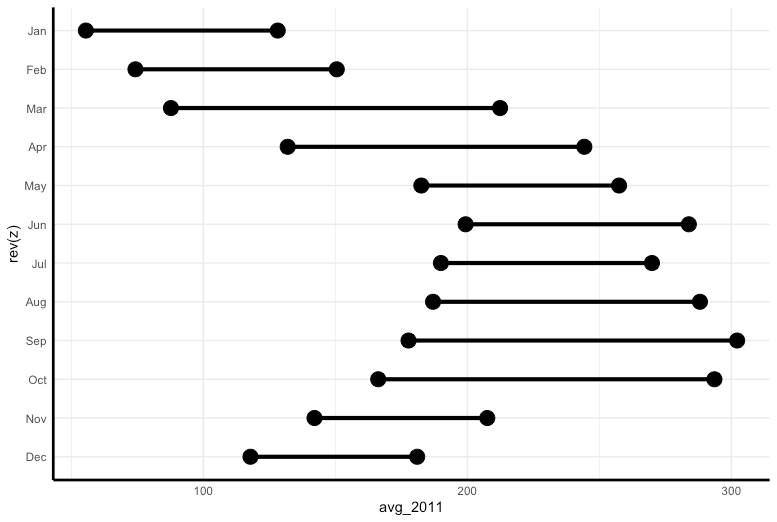
The final interesting finding for this report comes from **Figure 4**, the Box Plot. This plot confirms one of our previous findings, that between noon and 6:00PM is the period of time in the day with the heaviest usage across the entire 2 years that this data-set spans. By taking a look at casual users' bike usage across different types of weather, it is clear that ‘Clear - Partly Cloudy’ weather accounts for the heaviest usage, followed closely by ‘Cloudy mist’. However, once the weather becomes ‘Light Precipitation - Thunderstorm’, casual users' bike usage decreases drastically. We chose to omit ‘Heavy Precipitation - Foggy Snow’ from this plot, due to there being almost no Casual Users using the bikes during this weather.

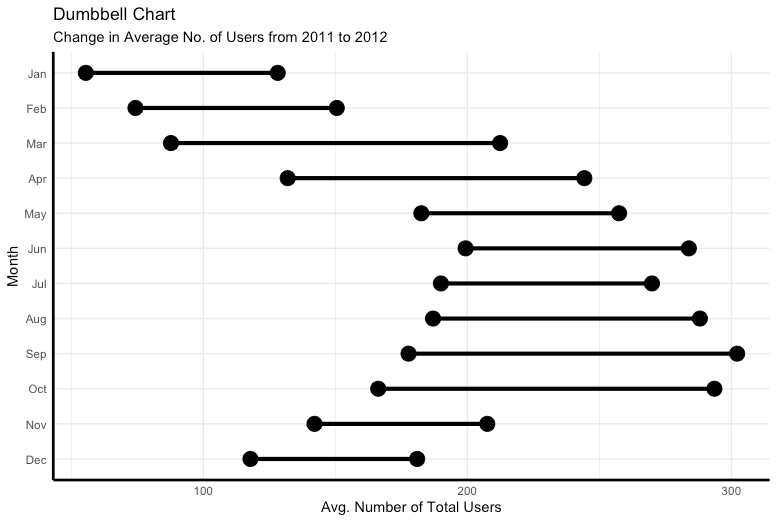
# 

# **Plots**

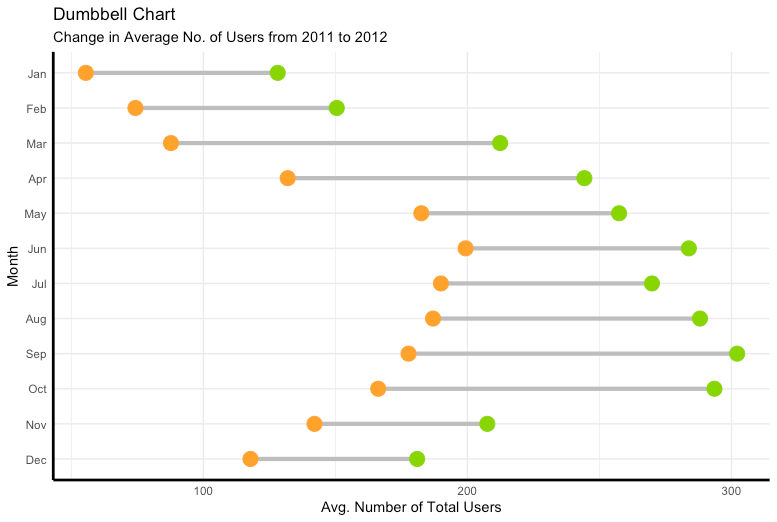
Iterations for each of the four plots in our report and presentation.

*Dumbbell plot iterations:*

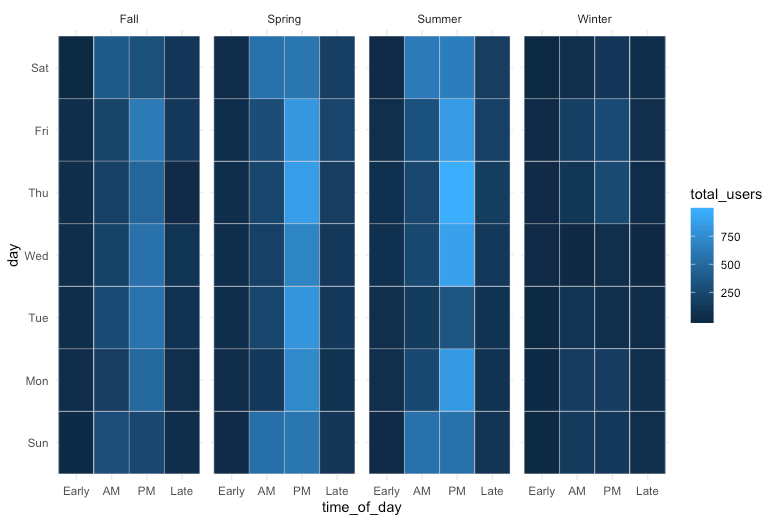


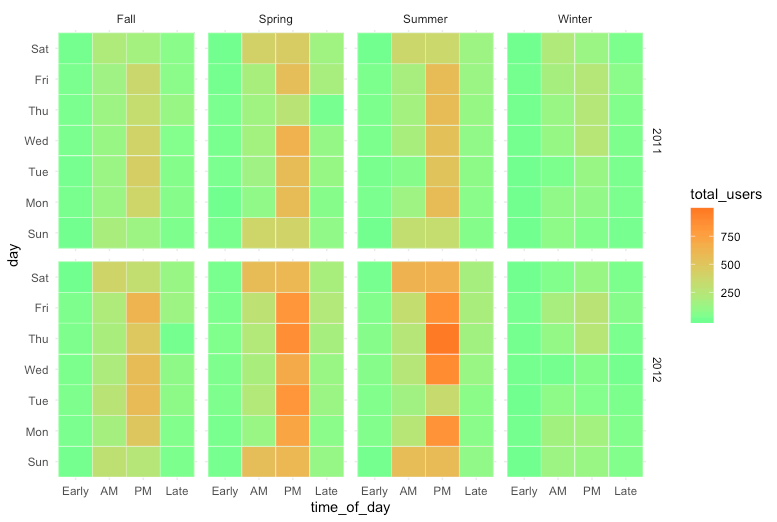


**Figure 1: Final Dumbbell Chart**

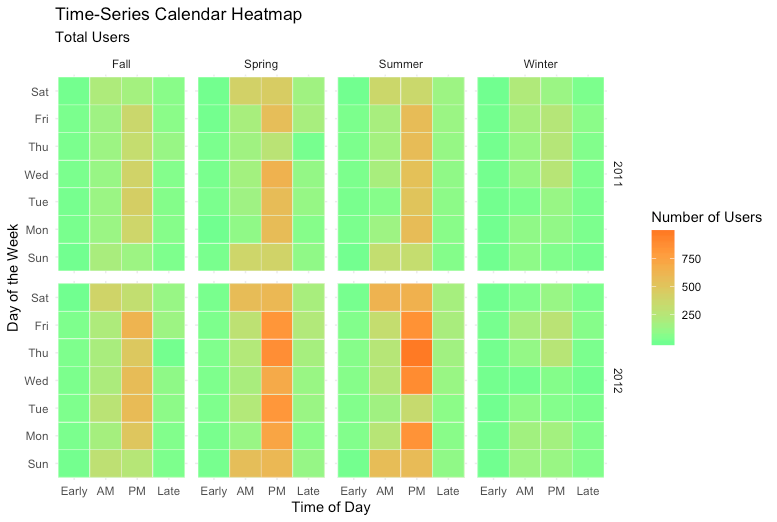


*Heat Map Iterations*

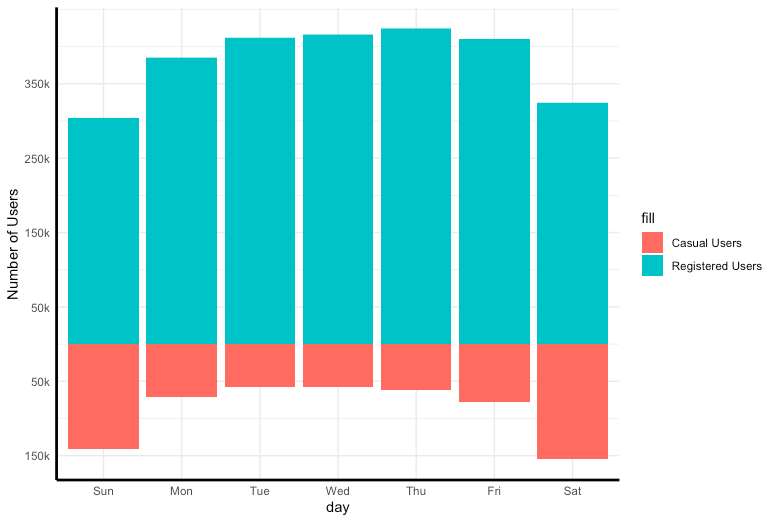


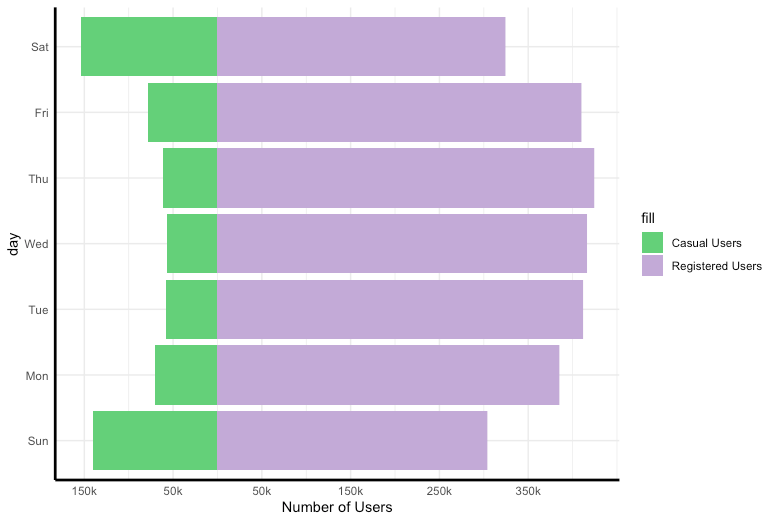


**Figure 2: Final Time-Series Calendar Heatmap**

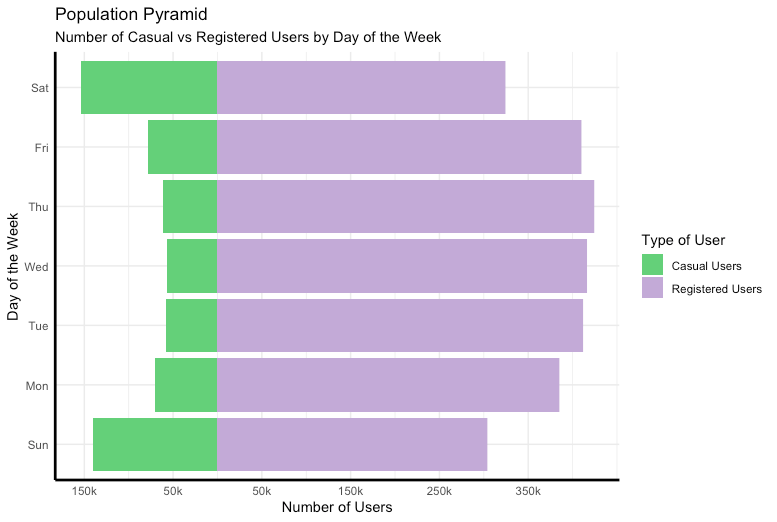


*Population Pyramid Iterations*

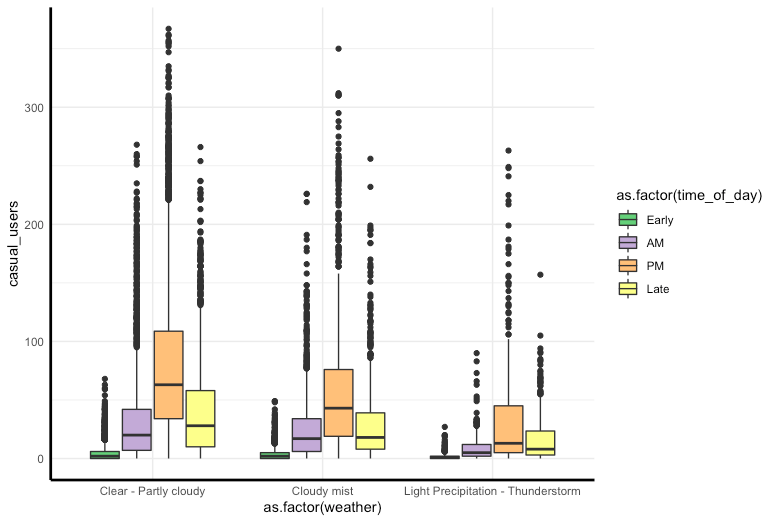
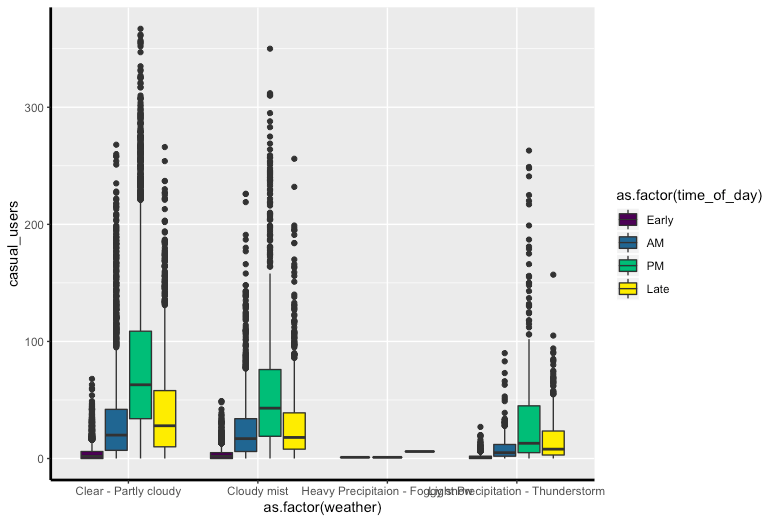




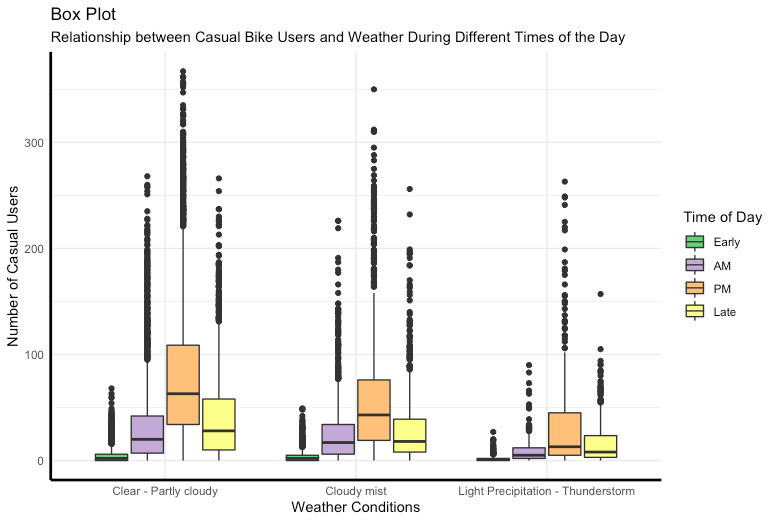
**Figure 3: Final Population Pyramid**



*Boxplot iterations*



**Figure 4: Final Box Plot**



# **Conclusion**

In analyzing the bike sharing dataset we found that the fairly intuitive hypotheses we had were confirmed. And we even made inferences as to why the results presented that way.

Throughout all of the graphs we saw that there was a very obvious increase in users from the first year to the second year, which we attributed to the growing popularity of such systems especially with advertising mediums like social media. Continuing with the trends among users: we saw that in terms of regular versus casual users, there were more registered users. Among registered users, there was an increase in use during the workweek which we assume is due to use for transport to work and school. And conversely, we saw that for casual users, more of them used bikes on the weekend which we assumed was due to casual users typically being visitors or tourists.

Specifying the timing of these upticks in use, we saw that warmer weather with relatively clear skies in the hours of noon to 6 pm (especially Thursdays in 2012) saw the most use. These seemed like very specific qualifiers but it made plenty of sense since during the spring and summer vacations optimal weather conditions are more likely to occur with the least precipitation, if you are going to be outside. In addition to this, PM hours are when most people get out of school and work, but it isn’t so late that it’s too dark and visibility is poor.

These conclusions that we made only raised more questions. As we saw the highest usage in the PM hours, we figured that if we broke the hours up into smaller intervals, we might be able to see clearer patterns in commuting in the area and pinpoint exactly when the usage increases. Even seeing what exactly goes on Thursdays in this town, or if there is one Thursday in particular that sees a sharp increase in usage.

Moreover, we would like to see if the patterns observed in this one town hold in other towns in other nations all across the globe. That would further inform whether patterns are observed in relation to culture specific reasons, or if the relation is purely based on each variable and nothing outside the data. Answers to these questions would help us to better understand the bike sharing market and predict its future.

# 

# **References**

“Be Awesome in ggplot2: A Practical Guide to Be Highly Effective - R Software and Data Visualization.” *STHDA*, http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-visualization.

“Bike Sharing Dataset.” *UCI Machine Learning Repository: Bike Sharing Dataset Data Set*, National Science Foundation, https://archive.ics.uci.edu/ml/datasets/Bike Sharing Dataset.

Prabhakaran, Selva. “Top 50 ggplot2 Visualizations.” *r-Statistics.co*, http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html.

# **Appendix**

# Installation of necessary packages for displaying graphs

install.packages("digest")

install.packages("reshape2")

install.packages("devtools")

devtools::install\_github("hrbrmstr/ggalt")

install.packages("dplyr")

install.packages("purrr")

library("dplyr")

library(ggalt)

library("reshape2")

library("UsingR")

library(ggplot2)

# Loading in the data set file

hour <- read.csv("hour.csv")

hour <- as.data.frame(hour)

# Data Cleaning

# Clean up names

names(hour) <- c("instant", "date", "season", "year", "month", "hour", "holiday",

"day", "workday", "weather", "temperature", "feels\_like",

"humidity", "windspeed", "casual\_users", "registered\_users",

"total\_users")

# Name the seasons

index\_spring <- which(hour$season == 2)

index\_summer <- which(hour$season == 3)

index\_fall <- which(hour$season == 4)

index\_winter <- which(hour$season == 1)

hour$season[index\_spring] <- "Spring"

hour$season[index\_summer] <- "Summer"

hour$season[index\_fall] <- "Fall"

hour$season[index\_winter] <- "Winter"

# List the actual years

index\_2011 <- which(hour$year == 0)

index\_2012 <- which(hour$year == 1)

hour$year[index\_2011] <- 2011

hour$year[index\_2012] <- 2012

# List the actual Months

hour$month <- factor(hour$month,

ordered = TRUE,

levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),

labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",

"Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

# Holiday or not

hour$holiday <- as.logical(hour$holiday)

# Days of the week

hour$day <- factor(hour$day,

ordered = TRUE,

levels = c(0, 1, 2, 3, 4, 5, 6),

labels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))

# Workday or not

hour$workday <- as.logical(hour$workday)

# Weather codes

hour$weather[which(hour$weather == 1)] <- "Clear - Partly cloudy"

hour$weather[which(hour$weather == 2)] <- "Cloudy mist"

hour$weather[which(hour$weather == 3)] <- "Light Precipitation - Thunderstorm"

hour$weather[which(hour$weather == 4)] <- "Heavy Precipitation - Foggy snow"

# Converting Temp from relative C to actual F rounding to one decimal point

hour$temperature <- round(1.8 \* (hour$temperature \* 47 - 8) + 32, 1)

hour$feels\_like <- round(1.8 \* (hour$feels\_like \* 66 - 16) + 32, 1)

# Turning humidity into percent

hour$humidity <- hour$humidity \* 100

# Converting normalized wind speed to mph rounding to one decimal point

hour$windspeed <- round(hour$windspeed \* 67 / 1.609, 1)

# Graphs-------------------------------------------------------------------------------------------------------------------

## Dumbbell Plot

# Creating new data frames, one for each year and exclusion of leap day

hour\_new <- filter(hour, year == "2011")

hour\_new2 <- filter(hour, year == "2012")

hour\_new2 <- filter(hour\_new2, date != "2012-02-29")

hour\_new2$instant <- c(1:8711)

hour\_new$total\_users\_2011 <- hour\_new$total\_users

hour\_new2$total\_users\_2012 <- hour\_new2$total\_users

# Joining the data for the use of a dumbbell plot

hour\_dumbbell <- inner\_join(hour\_new, hour\_new2, by = "instant")

# Identifying the indexes for each month in the data

jan\_index <- which(hour\_dumbbell$month.x == "Jan")

feb\_index <- which(hour\_dumbbell$month.x == "Feb")

mar\_index <- which(hour\_dumbbell$month.x == "Mar")

apr\_index <- which(hour\_dumbbell$month.x == "Apr")

may\_index <- which(hour\_dumbbell$month.x == "May")

jun\_index <- which(hour\_dumbbell$month.x == "Jun")

jul\_index <- which(hour\_dumbbell$month.x == "Jul")

aug\_index <- which(hour\_dumbbell$month.x == "Aug")

sep\_index <- which(hour\_dumbbell$month.x == "Sep")

oct\_index <- which(hour\_dumbbell$month.x == "Oct")

nov\_index <- which(hour\_dumbbell$month.x == "Nov")

dec\_index <- which(hour\_dumbbell$month.x == "Dec")

# Averaging the total amount of users per month for both years

jan2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[jan\_index])

jan2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[jan\_index])

feb2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[feb\_index])

feb2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[feb\_index])

mar2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[mar\_index])

mar2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[mar\_index])

apr2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[apr\_index])

apr2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[apr\_index])

may2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[may\_index])

may2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[may\_index])

jun2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[jun\_index])

jun2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[jun\_index])

jul2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[jul\_index])

jul2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[jul\_index])

aug2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[aug\_index])

aug2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[aug\_index])

sep2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[sep\_index])

sep2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[sep\_index])

oct2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[oct\_index])

oct2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[oct\_index])

nov2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[nov\_index])

nov2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[nov\_index])

dec2011\_avg <- mean(hour\_dumbbell$total\_users\_2011[dec\_index])

dec2012\_avg <- mean(hour\_dumbbell$total\_users\_2012[dec\_index])

# Turning those averages into vectors

avg\_2011 <- c(jan2011\_avg, feb2011\_avg, mar2011\_avg, apr2011\_avg, may2011\_avg,

jun2011\_avg, jul2011\_avg, aug2011\_avg, sep2011\_avg, oct2011\_avg,

nov2011\_avg, dec2011\_avg)

avg\_2012 <- c(jan2012\_avg, feb2012\_avg, mar2012\_avg, apr2012\_avg, may2012\_avg,

jun2012\_avg, jul2012\_avg, aug2012\_avg, sep2012\_avg, oct2012\_avg,

nov2012\_avg, dec2012\_avg)

# Creating labels for the months

z <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun",

"Jul", "Aug", "Sep", "Oct", "Nov", "Dec")

z <- factor(z, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",

"Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

# Creating a data frame from the labels and the averages for the two years

df <- data.frame(z, avg\_2011, avg\_2012)

# Plotting the data on a dumbbell plot

ggplot(df, mapping = aes(x = avg\_2011, xend = avg\_2012, y = rev(z))) +

geom\_dumbbell(size = 1.5,

size\_x = 5,

size\_xend = 5,

colour = "grey",

colour\_x = "tan1",

colour\_xend = "yellowgreen") +

theme\_minimal() +

theme(axis.line = element\_line(colour = "black",

size = 1, linetype = "solid")) +

scale\_y\_discrete(labels = rev(z)) +

labs(title = "Dumbbell Chart",

subtitle = "Change in Average No. of Users from 2011 to 2012",

x = "Avg. Number of Total Users",

y = "Month")

## Calendar Heat Map

# Creating a new categorical variable "time of day" breaking down the 24 hour

# day to 4 larger time periods of 6 hours each.

hour$time\_of\_day <- hour$hour

index\_early\_am <- which(hour$time\_of\_day <= 5)

index\_am <- which(hour$time\_of\_day > 5 & hour$time\_of\_day <= 11)

index\_pm <- which(hour$time\_of\_day > 11 & hour$time\_of\_day <= 17)

index\_late\_pm <- which(hour$time\_of\_day > 17 & hour$time\_of\_day <= 23)

hour$time\_of\_day[index\_early\_am] <- "Early"

hour$time\_of\_day[index\_am] <- "AM"

hour$time\_of\_day[index\_pm] <- "PM"

hour$time\_of\_day[index\_late\_pm] <- "Late"

hour$time\_of\_day <- factor(hour$time\_of\_day, ordered = TRUE,

levels = c("Early", "AM", "PM", "Late"))

#The creation of the heat map with the previous alteration in the data

ggplot(hour, mapping = aes(x = time\_of\_day, y = day, fill = total\_users)) +

geom\_tile(colour = "white") +

facet\_grid(year ~ season) +

scale\_fill\_gradient(low = "palegreen", high = "sienna1") +

theme\_minimal() +

labs(x = "Time of Day",

y = "Day of the Week",

title = "Time-Series Calendar Heatmap",

subtitle = "Total Users",

fill = "Number of Users")

## Population Pyramid

# Creating the population pyramid by manipulation of the bar plot

ggplot(data = hour) +

geom\_bar(mapping = aes(x = day, y = -casual\_users, fill = "Casual Users"),

stat = "identity") +

geom\_bar(mapping = aes(x = day, y = registered\_users,

fill = "Registered Users"), stat = "identity") +

coord\_flip() +

labs(title = "Population Pyramid",

subtitle = "Number of Casual vs Registered Users by Day of the Week",

fill = "Type of User") +

xlab("Day of the Week") +

theme(legend.title = element\_blank()) +

theme\_minimal() +

theme(axis.line = element\_line(colour = "black",

size = 1, linetype = "solid")) +

scale\_fill\_brewer(palette = "Accent") +

ylab("Number of Users") +

scale\_y\_continuous(breaks = seq(-150000,400000,100000),

labels = paste0(as.character(c(150, 50, 50, 150, 250, 350)), "k"))

## Boxplot

# Creation of the boxplot

ggplot(data = subset(hour, weather != "Heavy Precipitaion - Foggy snow")) +

geom\_boxplot(mapping = aes(y = casual\_users, x = as.factor(weather),

fill = as.factor(time\_of\_day))) +

labs(fill = "Time of Day", x = "Weather Conditions",

y = "Number of Casual Users",

subtitle = "Relationship between Casual Bike Users and Weather During Different Times of the Day",

title = "Box Plot") +

theme\_minimal() + scale\_fill\_brewer(palette = "Accent") +

theme(axis.line = element\_line(colour = "black",

size = 1, linetype = "solid"))