~~~
---
  title: "Project Report Draft 5"
author: "Vivian Zheng, Tyler Braito, Calvin Aberg, Mattia Saladini, and Ndidi
Nwosu"
date: "2022-11-10"
output:
  html_document:
  df_print: paged
fig_height: 5
fig_width: 5
editor_options:
  chunk_output_type: console
---

  ```{r, include=FALSE, warning=FALSE}
knitr::opts_chunk$set(echo = TRUE)
#Loading data:
#setwd("~/Documents/Sem 1/STAT 405/City Bike Data")
bike_dta <- read.csv("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/202208-
capitalbikeshare-tripdata/202208-capitalbikeshare-tripdata.csv", header =
TRUE, sep = ",")
library(ggplot2)
library(dplyr)
library(ggthemes)
library(RSQLite)
library(stringr)
library(lubridate)
library(ggeffects)
library(dotwhisker)

# Import bikeshare and weather csv as SQL
dcon <- dbConnect(SQLite(), dbname = "mydba.sqlite")
table <- read.csv(paste0("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/data
project/stat405-project/Washington,DC,USA weather 2022-08-01 to
2022-08-31.csv"))
dbWriteTable(conn = dcon, name = "weather",
             table, append = TRUE, row.names = FALSE)
dbListTables(dcon)
table2 <- read.csv(paste0("C:/Users/Ndidi/Desktop/2022-23/fall/STAT
405/202208-capitalbikeshare-tripdata/202208-capitalbikeshare-tripdata.csv"))
dbWriteTable(conn = dcon, name = "bikes",
             table2, append = TRUE, row.names = FALSE)
dbListTables(dcon)
```
~~~

## Plots on City Bike Data in Washington DC

Our data measures all city bike rentals in the city of Washington DC throughout the month of August 2022. Every bike rental is tracked along with information such as bike type, rental start and stop time, start and stop station names, and member type. There are over a million entries, each one recording the information corresponding with one bike ride in the Washington DC area.

We also have data that has the temperature and weather information throughout the month of August 2022 for Washington DC.
As we analyze the data, we seek to discover if there are any correlations between the different variables for each ride. Is there a pattern in what bike locations are most popular, or a correlation between the type of bike share member, the duration of the ride or the type of bike chosen? How does the weather affect the number of people using city bikes? Knowing if such a pattern exists would be beneficial to the bike share company, seeing as they could use the information to maximize the amount of customers they get by varying the number of bikes of each type, or at each station.
These various data analyses will hopefully reveal such patterns.

1. Scatterplot of Ride Duration by Date

~~~
```{r, echo=FALSE, warning=FALSE}
# library(ggplot2)
# install.packages("tidyverse")
start_time <- bike_dta$started_at
end_time <- bike_dta$ended_at
bike_dta$start_time <- as.POSIXct(start_time,tz="EST")
bike_dta$end_time <- as.POSIXct(end_time,tz="EST")
bike_dta$duration <- as.numeric((bike_dta$end_time - bike_dta$start_time) /
60)

# Identify top station (Lincoln Memorial)
# tab <- table(bike_dta$start_station_name)
# sort(tab)
linc_mem <- subset(bike_dta, bike_dta$start_station_name == "Lincoln Memorial"
&
                     bike_dta$end_station_name != "Lincoln Memorial")
linc_mem <- na.omit(linc_mem)

ggplot(data = linc_mem, aes(x = start_time, y = duration)) +
  geom_point(color = "skyblue") +
  ggtitle("Duration by Date for Lincoln Memorial Station, August 2022") +
  xlab("Date") + ylab("Rental Duration (min)")
```
~~~

This scatter plot shows the rental duration differences between dates throughout the month of August. Most of the rentals are clustered under 60 minutes, with a few outliers above 1000 minutes. These outliers may reflect errors in the tracking system of the bike, or instances when users forgot to return their bikes to another station. There seems to be a pretty uniform distribution of rentals by date throughout the month.

2. Daily Usage Rank of the Top 10 Most Common Starting Stations

~~~
```{r, echo=FALSE, warning=FALSE}
# library(ggplot2)
# library(dplyr)
bike_dta$date <- format(bike_dta$start_time, '%Y-%m-%d')
bike_data_sub <- subset(bike_dta, select = c("start_station_name",
"end_station_name", "date"))
bike_data_sub <- bike_data_sub %>%
~~~

```
  add_count(start_station_name, name = 'start_occurance')
bike_data_sub <- bike_data_sub %>%
  add_count(end_station_name, name = 'end_occurance')
# View(bike_data_sub)

#Only use 10 most common stations
bike_data_sub <- subset(bike_data_sub, end_station_name %in%
names(sort(table(bike_data_sub$start_station_name),decreasing=TRUE)[2:11]))
bike_data_sub <- subset(bike_data_sub, start_station_name %in%
bike_data_sub$end_station_name)

length(unique(bike_data_sub$start_station_name))
length(unique(bike_data_sub$end_station_name))


bike_sub2 <- bike_data_sub %>%
  group_by(date) %>%
  add_count(start_station_name, name = 'occurance_by_day') %>%
  ungroup()

bike_sub2 <- bike_sub2 %>% distinct(start_station_name, date, .keep_all=TRUE)

bike_sub3 <- bike_sub2 %>%
  group_by(date) %>%
  arrange(date, desc(occurance_by_day), start_station_name) %>%
  mutate(rank=row_number())%>%
  ungroup()

bike_sub3$date <- as.Date(bike_sub3$date)


ggplot(bike_sub3, aes(x=date, y = rank, group = start_station_name))+
  geom_line(aes(color = start_station_name), size = 2, alpha = .75)+
  geom_point(aes(color = start_station_name), size = 4, alpha = .75)+
  scale_y_reverse(breaks = 1:10)+
  scale_color_brewer(palette = "Paired", name = "Station Name")+
  scale_x_date(date_breaks = "6 days")+
  theme(legend.position = "bottom") +
  guides(color = guide_legend(nrow = 4)) +
  labs(title = "Daily Usage Rank of the Top 10 Most Common Starting Stations",
x = "Date", y = "Rank")
```

This plot is visualizing the usage ranks of the 10 most common starting of August for each day of the month. From the plot, we can see that generally the top 4 stations at the very start of the month remained in the top five ranks for most of the rest of the month. Conversely, it was rather rare for a station that initially ranked below the top 5 to reach a top five rank. The red line that was exactly ranked #5 at the beginning exhibits the widest range of ranks for the rest of the month. From this we can conclude that the very most popular stations will consistently draw the most demand from day-to-day, even if their individual ranks within these top few spots will fluctuate.

3. Density Plot of Riding Distance by Bike Type

~~~
```{r, echo=FALSE, output=TRUE, warning=FALSE}
#Calculating duration:
# start_time <- as.POSIXct(bike_dta$started_at,tz="EST")
# end_time <- as.POSIXct(bike_dta$ended_at,tz="EST")
bike_dta$duration_sec <- bike_dta$end_time - bike_dta$start_time
bike_dta$duration_min <- bike_dta$duration_sec/60



#Densities:
bike_dta <- na.omit(bike_dta)
bike_dta$distance <-sqrt((69*(bike_dta$end_lat - bike_dta$start_lat))**2 +
55*((bike_dta$end_lat - bike_dta$start_lat))**2)
bike_dta2 <- bike_dta[bike_dta$distance<=3, ]

ggplot(bike_dta2, aes(x=distance)) + geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(distance)), color="blue", linetype="dashed",
size=1) + labs(x="Distance (miles)", y = "Density") + facet_grid(. ~
rideable_type)
```
~~~

In this plot, we are visualizing the differences in riding distance by type of rideable bike. To do this, we can see three separate density plots of each rideable type of bike that Capital Bikes offers. This visualization gives us a great deal of information of the tendencies of the users of each bike type. With the docked bike, there is a very large concentration of riders riding very short distances and stopping. However, with the electric bikes, there is a much larger skew of riders riding much longer distances. These tendencies can give us insight into which bike types are preferable in different locations, and how users are likely to use different bikes.

4. Barplots of Distance Ridden by Type of Bike

~~~
```{r, echo=FALSE, output=TRUE, warning=FALSE}

dist <- sqrt((69*(bike_dta$end_lat - bike_dta$start_lat))**2 +
55*((bike_dta$end_lat - bike_dta$start_lat))**2)

bike_dta$dist <- dist

brks <- seq(-250000, 250000, 50000)
lbls = paste0(as.character(c(seq(250, 0, -50), seq(50, 250, 50))), "k")

#bike_dta$rideable_type

# Plot
ggplot(bike_dta, aes(x = rideable_type, y = dist, fill = member_casual)) +
  geom_bar(stat = "identity", width = .6) +
  coord_flip() +
  theme_tufte() +
  scale_fill_brewer(palette = "Dark2") +
  theme(plot.title = element_text(hjust = .5), axis.ticks = element_blank()) +
  labs(title = "Distance Ridden By Type of Bike") +
~~~

~~~
  xlab("Type of Bike") +
  ylab("Distance Ridden in August") +
  scale_y_continuous(breaks = brks, labels = lbls) +
  theme(plot.title = element_text(hjust = .5, size = 25), axis.ticks =
element_blank()) +
  labs(fill = "Type of Member")

```
~~~

This plot illustrates the distance ridden on different types of bikes colored by whether the rider was a member of Capital Bikes or just a casual rider. From a business perspective, this plot is very helpful in illustrating the tendencies of different customers and how they use Capital Bikes' products. With this, we can see that the bulk of the usage of the bikes are centered on classic bikes with the majority of member distance being on the classic bikes. However, the docked bike has seen very little usage by member customers. The reasons for this disparity could be a question to be examined by those with a stake in the business' success.

5. Boxplots of Bike Types and Rental Durations

~~~
```{r, echo = FALSE, output = TRUE, warning=FALSE}
#Calculating duration:
# start_time <- as.POSIXct(bike_dta$started_at,tz="EST")
# end_time <- as.POSIXct(bike_dta$ended_at,tz="EST")
bike_dta$duration_sec <- bike_dta$end_time - bike_dta$start_time
bike_dta$duration_min <- bike_dta$duration_sec/60

#Boxplots:
par(mfrow = c(1, 3))
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "classic_bike"), main="Classic Bikes Boxplot",
        ylab="Duration (min)", col="red")
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "electric_bike"), main="Electric Bikes Boxplot",
col="green")
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "docked_bike"), main="Docked Bikes Boxplot",
col="blue")

```
~~~

The first dataset describes different characteristics (duration, location, etc.) of each bike rental ride in the DC Area. The three box plots each display the minimum, first quartile, median, third quartile, and maximum bike rental durations for Classic, Electric, and Docked bikes separately.

6. Most Popular Stations by Start and End Location

~~~
```{r, echo = FALSE, output = TRUE}
counts_start <- head(sort(table(bike_dta$start_station_name), decreasing =
TRUE), 6)[-1]
~~~

~~~
counts_end <- head(sort(table(bike_dta$end_station_name), decreasing = TRUE),
6)[-1]

stations <- rbind(counts_start,counts_end)

par(mfrow = c(1, 1))
barplot(stations, col = c('darkblue', 'red'),
        names.arg = c("Lincoln",'Jefferson', "T St.", '15th', 'Union St'),
        main = "Most Popular Stations by Start and End Location",
        legend = c('End Location', 'Start Location'),
        args.legend = list(x = "topright",
                           inset = c(0, -0.1)),
        xlab = 'Stations',
        ylab = "Count of Riders")
```
~~~

This plot describes the most popular starting stations and ending stations for Capital Bike riders in August. The total height of the bar is the sum of trips started and ended at a given station, with the blue being the number of trips ended and red being the number of trips started. These five stations were the most commonly visited stations in the Washington D.C. area. We can see that there isn’t a majority station within the data, and, within the top five stations, there is a fairly balanced flow in and out of each station.

7. Histogram of Usage Frequency by Days of the Week and Date

~~~
```{r, echo = FALSE, output = TRUE}
par(mfrow = c(1, 2))
# bike_dta$started_at <- as.POSIXct(bike_dta$started_at, tz = "EST")
# bike_dta$ended_at <-as.POSIXct(bike_dta$ended_at, tz = "EST")
# bike_dta$duration <-  as.numeric(bike_dta$end_time - bike_dta$start_time)
bike_dta$weekday <-as.Date(bike_dta$start_time)
bike_dta$weekday <- format(bike_dta$weekday, "%a")
# bike_dta$weekdayn <- as.numeric(format(bike_dta$start_time, "%w"))
# hist(bike_dta$weekdayn, breaks = -.5+0:7, labels =
unique(bike_dta$weekday[order(bike_dta$weekdayn)]), xlab = "", xaxt = "n",
main ="Ride Frequency Per Day of Week", col = "deepskyblue")
barplot(table(bike_dta$weekday), main = "Ride Frequency Per Day of Week", xlab
= "Day of the Week", col = "deepskyblue")
date_formatted <- as.Date(bike_dta$start_time, format = "%Y-%m-%d")
ride_start_days <- format(date_formatted, "%d")
barplot(table(ride_start_days), main = "Rides By Day in August", xlab = "Day",
ylab = "Number of Rides", col = "yellow")

```
~~~

The plot on the right, "Ride Frequency Per Day of Week" zooms in on the trend observed in the “Rides by Day in August” plot by giving an account of which days of the week exhibit a peak in bike usage activity. Based on the plot, there seems to be both a midweek peak and a weekend peak on Saturday.
The plot on the right, "Rides By Day in August", shows the frequency of bike rides using the Capital Bike bikes throughout the month of August by day. Every value on the x-axis represents a day of August (1st to the 31st) and the height of each bar represents the number of bike rides that day.

The minimum number of rides occurs on August 10th, with 7,771 rides, and the maximum number of rides occurs on August 13th, with 14,860 rides. There is no drastic peak or valley, and the amount of bike rides is relatively uniformly distributed besides minor peaks around August 13th and August 27th. On average, there were 12,182.48 rides per day.

8. Temperature in DC in August 2022 with Average Temp based on Weather Condition

~~~
```{r, echo = FALSE, output = TRUE, warning=FALSE}
# Create query for average based on condition
res <- dbSendQuery(conn = dcon, "
SELECT conditions, avg(temp)
FROM weather
GROUP BY conditions
ORDER BY avg(temp) DESC;
")

# store results in df
conditions <- dbFetch(res, -1)
dbClearResult(res)
# conditions

# Plot graph
plot(pull(table, var = "temp"), type = "l", lwd = "3",
     xlab = "Day of month in August 2022",
     ylab = "Temperature (°F)",
     main = "Temperature in DC in August 2022 with
     Average Temp based on Weather Condition")

# add lines for average temp of conditions
abline(h = conditions[1:4,2], lwd = "2", col = c("gray", "orange",
"red","blue"))

# add legend
legend("topright", c("Partially Cloudy", "Rain, Partially Cloudy",
                     "Clear", "Rain, Overcast"),
       col = c("gray", "orange", "red", "blue"),
       cex = 0.7, lty = 1)
```
~~~

This plot use from the weather data gathered from August 2022 in Washington DC, and shows the temperature per day in the month, with lines to display where the average temperature was for each type of weather: partially cloudy, rainy and partially cloudy, clear and rainy and overcast. This graph shows that temperature spiked in the August 5th and August 25th times of year, and hit an all time low around August 15th, when it was typically raining.

9. Bike Rides vs Average Temperature of That Day

~~~
```{r, echo = FALSE, output = TRUE}
weather <- read.csv("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/data
project/stat405-project/Washington,DC,USA weather 2022-08-01 to
2022-08-31.csv")
~~~

~~~
# Create query for average based on condition
# res <- dbSendQuery(conn = dcon, "
#               SELECT w.datetime, w.temp, b.started_at
#               FROM bikes b, weather w;")
#
# # store results in df
# rides <- dbFetch(res, -1)
# dbClearResult(res)

rides_per_day <- data.frame(weather$datetime, weather$temp,
table(ride_start_days))

ggplot(rides_per_day)  +
  geom_bar(aes(x=ride_start_days, y=Freq),stat="identity",
fill="lightgreen",color="darkgreen")+
  geom_line(aes(x=ride_start_days, y=weather.temp *150, group =
1),stat="identity",color="red",size=2)+
  labs(title= "Bike Rides vs Temperature (in Fahrenheit)",
       x="day",y="Number of Bike Rides")+
  scale_y_continuous(sec.axis=sec_axis(
    ~./ 150,name="Temperature"))

```
~~~

This graph directly compares the temperature of the day (represented by the red line) to the number of bike rides that were taken on city bikes that day. As the graph shows, there seems to be some correlation. When the temperature dipped around the 12th of the month, there was an increase in bike rides, likely due to the fact that people may prefer to bike in slightly cooler weather. But on the 15th, the coolest day of the month, there was a sudden decrease in bike rides, which makes sense because as seen in graph number 8, that was a rainy day. Bike rides seem to relatively correlate with temperature, but precipitation also seems to play a role.

10. Concentration of Trip Starts

~~~
```{r, echo = FALSE, output = TRUE}
# Create query for average based on condition
res <- dbSendQuery(conn = dcon, "
              SELECT start_station_name, end_station_name, COUNT(*) as
num_trips
              FROM bikes b
              GROUP BY start_station_name, end_station_name
              ORDER BY (-1 * num_trips)
              LIMIT 10;
                   ")

# store results in df
station_travel <- dbFetch(res, -1)
dbClearResult(res)
# station_travel

ggplot(data = station_travel[-1,], aes(x = start_station_name, y = num_trips))
+
  geom_bar(stat = 'identity') +
~~~

~~~
  coord_flip() +
  ggtitle("Concentration of Trip Starts") +
  labs(x = "Start Station Name", y = "Number of Trips") +
  theme_minimal()
```
~~~

This graph shows the number of trips that start from a select number of stations. As the graph shows, there seems to be a high concentration of bike rides that start from Jefferson Dr & 14th St SW, which this station having the most bike rides by a large margin. Meanwhile, the least popular starting station seems to be Roosevelt Island. Jefferson Dr & 14th St SW may be a popular station where there is a higher concentration of bike riders, so investing in more bikes at this station, or creating more bike stations near this area may be a profitable move for this Bike Share company, while investing less in places near Roosevelt Island may not be.

~~~
=======
  ---
  title: "Project Report Draft 3"
date: '2022-11-03'
author: "Vivian Zheng, Tyler Braito, Calvin Aberg, Mattia Saladini, and Ndidi
Nwosu"
output: pdf_document
fig_height: 5
fig_width: 5
editor_options:
  chunk_output_type: console
---

  ```{r setup, include=FALSE, warning=FALSE}
knitr::opts_chunk$set(echo = TRUE)
#Loading data:
#setwd("~/Documents/Sem 1/STAT 405/City Bike Data")
bike_dta <- read.csv("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/202208-
capitalbikeshare-tripdata/202208-capitalbikeshare-tripdata.csv", header =
TRUE, sep = ",")
library(ggplot2)
library(dplyr)
library(ggthemes)
library(RSQLite)

# Import bikeshare and weather csv as SQL
dcon <- dbConnect(SQLite(), dbname = "mydba.sqlite")
table <- read.csv(paste0("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/data
project/stat405-project/Washington,DC,USA weather 2022-08-01 to
2022-08-31.csv"))
dbWriteTable(conn = dcon, name = "weather",
             table, append = TRUE, row.names = FALSE)
dbListTables(dcon)
table2 <- read.csv(paste0("C:/Users/Ndidi/Desktop/2022-23/fall/STAT
405/202208-capitalbikeshare-tripdata/202208-capitalbikeshare-tripdata.csv"))
dbWriteTable(conn = dcon, name = "bikes",
             table2, append = TRUE, row.names = FALSE)
dbListTables(dcon)
~~~

~~~
```
~~~

## Plots on City Bike Data in Washington DC

Our data measures all city bike rentals in the city of Washington DC throughout the month of August 2022. Every bike rental is tracked along with information such as bike type, rental start and stop time, start and stop station names, and member type. There are over a million entries, each one recording the information corresponding with one bike ride in the Washington DC area.
We also have data that has the temperature and weather information throughout the month of August 2022 for Washington DC.
As we analyze the data, we seek to discover if there are any correlations between the different variables for each ride. Is there a pattern in what bike locations are most popular, or a correlation between the type of bike share member, the duration of the ride or the type of bike chosen? How does the weather affect the number of people using city bikes? Knowing if such a pattern exists would be beneficial to the bike share company, seeing as they could use the information to maximize the amount of customers they get by varying the number of bikes of each type, or at each station.
These various data analyses will hopefully reveal such patterns.

1. Scatterplot of Ride Duration by Date

~~~
```{r, echo=FALSE, warning=FALSE}
# library(ggplot2)
# install.packages("tidyverse")
start_time <- bike_dta$started_at
end_time <- bike_dta$ended_at
bike_dta$start_time <- as.POSIXct(start_time,tz="EST")
bike_dta$end_time <- as.POSIXct(end_time,tz="EST")
bike_dta$duration <- as.numeric((bike_dta$end_time - bike_dta$start_time) /
60)

# Identify top station (Lincoln Memorial)
# tab <- table(bike_dta$start_station_name)
# sort(tab)
linc_mem <- subset(bike_dta, bike_dta$start_station_name == "Lincoln Memorial"
&
                     bike_dta$end_station_name != "Lincoln Memorial")
linc_mem <- na.omit(linc_mem)

ggplot(data = linc_mem, aes(x = start_time, y = duration)) +
  geom_point(color = "skyblue") +
  ggtitle("Duration by Date for Lincoln Memorial Station, August 2022") +
  xlab("Date") + ylab("Rental Duration (min)")
```
~~~

This scatter plot shows the rental duration differences between dates throughout the month of August. Most of the rentals are clustered under 60 minutes, with a few outliers above 1000 minutes. These outliers may reflect errors in the tracking system of the bike, or instances when users forgot to return their bikes to another station. There seems to be a pretty uniform distribution of rentals by date throughout the month.

## 2. Daily Usage Rank of the Top 10 Most Common Starting Stations

~~~
```{r, echo=FALSE, warning=FALSE}
# library(ggplot2)
# library(dplyr)
bike_dta$date <- format(bike_dta$start_time, '%Y-%m-%d')
bike_data_sub <- subset(bike_dta, select = c("start_station_name",
"end_station_name", "date"))
bike_data_sub <- bike_data_sub %>%
  add_count(start_station_name, name = 'start_occurance')
bike_data_sub <- bike_data_sub %>%
  add_count(end_station_name, name = 'end_occurance')
# View(bike_data_sub)

#Only use 10 most common stations
bike_data_sub <- subset(bike_data_sub, end_station_name %in%
names(sort(table(bike_data_sub$start_station_name),decreasing=TRUE)[2:11]))
bike_data_sub <- subset(bike_data_sub, start_station_name %in%
bike_data_sub$end_station_name)

length(unique(bike_data_sub$start_station_name))
length(unique(bike_data_sub$end_station_name))


bike_sub2 <- bike_data_sub %>%
  group_by(date) %>%
  add_count(start_station_name, name = 'occurance_by_day') %>%
  ungroup()

bike_sub2 <- bike_sub2 %>% distinct(start_station_name, date, .keep_all=TRUE)

bike_sub3 <- bike_sub2 %>%
  group_by(date) %>%
  arrange(date, desc(occurance_by_day), start_station_name) %>%
  mutate(rank=row_number())%>%
  ungroup()

bike_sub3$date <- as.Date(bike_sub3$date)


ggplot(bike_sub3, aes(x=date, y = rank, group = start_station_name))+
  geom_line(aes(color = start_station_name), size = 2, alpha = .75)+
  geom_point(aes(color = start_station_name), size = 4, alpha = .75)+
  scale_y_reverse(breaks = 1:10)+
  scale_color_brewer(palette = "Paired", name = "Station Name")+
  scale_x_date(date_breaks = "6 days")+
  theme(legend.position = "bottom") +
  guides(color = guide_legend(nrow = 4)) +
  labs(title = "Daily Usage Rank of the Top 10 Most Common Starting Stations",
x = "Date", y = "Rank")
```
~~~

This plot is visualizing the usage ranks of the 10 most common starting of August for each day of the month. From the plot, we can see that generally the top 4 stations at the very start of the month remained in the top five ranks for most of the rest of the month. Conversely, it was rather rare for a station that initially ranked below the top 5 to reach a top five rank. The red line that was exactly ranked #5 at the beginning exhibits the widest range of ranks for the rest of the month. From this we can conclude that the very most popular stations will consistently draw the most demand from day-to-day, even if their individual ranks within these top few spots will fluctuate.

3. Density Plot of Riding Distance by Bike Type

~~~
```{r, echo=FALSE, output=TRUE, warning=FALSE}
#Calculating duration:
# start_time <- as.POSIXct(bike_dta$started_at,tz="EST")
# end_time <- as.POSIXct(bike_dta$ended_at,tz="EST")
bike_dta$duration_sec <- bike_dta$end_time - bike_dta$start_time
bike_dta$duration_min <- bike_dta$duration_sec/60



#Densities:
bike_dta <- na.omit(bike_dta)
bike_dta$distance <-sqrt((69*(bike_dta$end_lat - bike_dta$start_lat))**2 +
55*((bike_dta$end_lat - bike_dta$start_lat))**2)
bike_dta2 <- bike_dta[bike_dta$distance<=3, ]

ggplot(bike_dta2, aes(x=distance)) + geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(distance)), color="blue", linetype="dashed",
size=1) + labs(x="Distance (miles)", y = "Density") + facet_grid(. ~
rideable_type)
```
~~~

In this plot, we are visualizing the differences in riding distance by type of rideable bike. To do this, we can see three separate density plots of each rideable type of bike that Capital Bikes offers. This visualization gives us a great deal of information of the tendencies of the users of each bike type. With the docked bike, there is a very large concentration of riders riding very short distances and stopping. However, with the electric bikes, there is a much larger skew of riders riding much longer distances. These tendencies can give us insight into which bike types are preferable in different locations, and how users are likely to use different bikes.

4. Barplots of Distance Ridden by Type of Bike

~~~
```{r, echo=FALSE, output=TRUE, warning=FALSE}

dist <- sqrt((69*(bike_dta$end_lat - bike_dta$start_lat))**2 +
55*((bike_dta$end_lat - bike_dta$start_lat))**2)

bike_dta$dist <- dist

brks <- seq(-250000, 250000, 50000)
lbls = paste0(as.character(c(seq(250, 0, -50), seq(50, 250, 50))), "k")
~~~

~~~~
#bike_dta$rideable_type

# Plot
ggplot(bike_dta, aes(x = rideable_type, y = dist, fill = member_casual)) +
  geom_bar(stat = "identity", width = .6) +
  coord_flip() +
  theme_tufte() +
  scale_fill_brewer(palette = "Dark2") +
  theme(plot.title = element_text(hjust = .5), axis.ticks = element_blank()) +
  labs(title = "Distance Ridden By Type of Bike") +
  xlab("Type of Bike") +
  ylab("Distance Ridden in August") +
  scale_y_continuous(breaks = brks, labels = lbls) +
  theme(plot.title = element_text(hjust = .5, size = 25), axis.ticks =
element_blank()) +
  labs(fill = "Type of Member")

```
~~~~

This plot illustrates the distance ridden on different types of bikes colored by whether the rider was a member of Capital Bikes or just a casual rider. From a business perspective, this plot is very helpful in illustrating the tendencies of different customers and how they use Capital Bikes’ products. With this, we can see that the bulk of the usage of the bikes are centered on classic bikes with the majority of member distance being on the classic bikes. However, the docked bike has seen very little usage by member customers. The reasons for this disparity could be a question to be examined by those with a stake in the business’ success.

5. Boxplots of Bike Types and Rental Durations

~~~~
```{r, echo = FALSE, output = TRUE, warning=FALSE}
#Calculating duration:
# start_time <- as.POSIXct(bike_dta$started_at,tz="EST")
# end_time <- as.POSIXct(bike_dta$ended_at,tz="EST")
bike_dta$duration_sec <- bike_dta$end_time - bike_dta$start_time
bike_dta$duration_min <- bike_dta$duration_sec/60

#Boxplots:
par(mfrow = c(1, 3))
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "classic_bike"), main="Classic Bikes Boxplot",
        ylab="Duration (min)", col="red")
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "electric_bike"), main="Electric Bikes Boxplot",
col="green")
boxplot(subset(bike_dta$duration_min, bike_dta$duration_min<=60 &
bike_dta$rideable_type == "docked_bike"), main="Docked Bikes Boxplot",
col="blue")

```
~~~~

The first dataset describes different characteristics (duration, location, etc.) of each bike rental ride in the DC Area. The three box plots each display the minimum, first quartile, median, third quartile, and maximum bike rental durations for Classic, Electric, and Docked bikes separately.

6. Most Popular Stations by Start and End Location

~~~
```{r, echo = FALSE, output = TRUE}
counts_start <- head(sort(table(bike_dta$start_station_name), decreasing =
TRUE), 6)[-1]

counts_end <- head(sort(table(bike_dta$end_station_name), decreasing = TRUE),
6)[-1]

stations <- rbind(counts_start,counts_end)

par(mfrow = c(1, 1))
barplot(stations, col = c('darkblue', 'red'),
        names.arg = c("Lincoln",'Jefferson', "T St.", '15th', 'Union St'),
        main = "Most Popular Stations by Start and End Location",
        legend = c('End Location', 'Start Location'),
        args.legend = list(x = "topright",
                           inset = c(0, -0.1)),
        xlab = 'Stations',
        ylab = "Count of Riders")
```
~~~

This plot describes the most popular starting stations and ending stations for Capital Bike riders in August. The total height of the bar is the sum of trips started and ended at a given station, with the blue being the number of trips ended and red being the number of trips started. These five stations were the most commonly visited stations in the Washington D.C. area. We can see that there isn’t a majority station within the data, and, within the top five stations, there is a fairly balanced flow in and out of each station.

7. Histogram of Usage Frequency by Days of the Week and Date

~~~
```{r, echo = FALSE, output = TRUE}
par(mfrow = c(1, 2))
# bike_dta$started_at <- as.POSIXct(bike_dta$started_at, tz = "EST")
# bike_dta$ended_at <-as.POSIXct(bike_dta$ended_at, tz = "EST")
# bike_dta$duration <-  as.numeric(bike_dta$end_time - bike_dta$start_time)
bike_dta$weekday <-as.Date(bike_dta$start_time)
bike_dta$weekday <- format(bike_dta$weekday, "%a")
# bike_dta$weekdayn <- as.numeric(format(bike_dta$start_time, "%w"))
# hist(bike_dta$weekdayn, breaks = -.5+0:7, labels =
unique(bike_dta$weekday[order(bike_dta$weekdayn)]), xlab = "", xaxt = "n",
main ="Ride Frequency Per Day of Week", col = "deepskyblue")
barplot(table(bike_dta$weekday), main = "Ride Frequency Per Day of Week", xlab
= "Day of the Week", col = "deepskyblue")
date_formatted <- as.Date(bike_dta$start_time, format = "%Y-%m-%d")
ride_start_days <- format(date_formatted, "%d")
barplot(table(ride_start_days), main = "Rides By Day in August", xlab = "Day",
ylab = "Number of Rides", col = "yellow")
~~~

~~~
```
~~~

The plot on the right, "Ride Frequency Per Day of Week" zooms in on the trend observed in the “Rides by Day in August” plot by giving an account of which days of the week exhibit a peak in bike usage activity. Based on the plot, there seems to be both a midweek peak and a weekend peak on Saturday.
The plot on the right, "Rides By Day in August", shows the frequency of bike rides using the Capital Bike bikes throughout the month of August by day. Every value on the x-axis represents a day of August (1st to the 31st) and the height of each bar represents the number of bike rides that day.
The minimum number of rides occurs on August 10th, with 7,771 rides, and the maximum number of rides occurs on August 13th, with 14,860 rides. There is no drastic peak or valley, and the amount of bike rides is relatively uniformly distributed besides minor peaks around August 13th and August 27th. On average, there were 12,182.48 rides per day.

8. Temperature in DC in August 2022 with Average Temp based on Weather Condition

~~~
```{r, echo = FALSE, output = TRUE, warning=FALSE}
# Create query for average based on condition
res <- dbSendQuery(conn = dcon, "
SELECT conditions, avg(temp)
FROM weather
GROUP BY conditions
ORDER BY avg(temp) DESC;
")

# store results in df
conditions <- dbFetch(res, -1)
dbClearResult(res)
# conditions

# Plot graph
plot(pull(table, var = "temp"), type = "l", lwd = "3",
     xlab = "Day of month in August 2022",
     ylab = "Temperature (°F)",
     main = "Temperature in DC in August 2022 with
     Average Temp based on Weather Condition")

# add lines for average temp of conditions
abline(h = conditions[1:4,2], lwd = "2", col = c("gray", "orange",
"red","blue"))

# add legend
legend("topright", c("Partially Cloudy", "Rain, Partially Cloudy",
                     "Clear", "Rain, Overcast"),
       col = c("gray", "orange", "red", "blue"),
       cex = 0.7, lty = 1)
```
~~~

This plot use from the weather data gathered from August 2022 in Washington DC, and shows the temperature per day in the month, with lines to display where the average temperature was for each type of weather: partially cloudy, rainy and partially cloudy, clear and rainy and overcast. This graph shows that temperature spiked in the August 5th and August 25th times of year, and hit an all time low around August 15th, when it was typically raining.

9. Bike Rides vs Average Temperature of That Day

~~~
```{r, echo = FALSE, output = TRUE}
weather <- read.csv("C:/Users/Ndidi/Desktop/2022-23/fall/STAT 405/data
project/stat405-project/Washington,DC,USA weather 2022-08-01 to
2022-08-31.csv")
# Create query for average based on condition
# res <- dbSendQuery(conn = dcon, "
#                SELECT w.datetime, w.temp, b.started_at
#                FROM bikes b, weather w;")
#
# # store results in df
# rides <- dbFetch(res, -1)
# dbClearResult(res)

rides_per_day <- data.frame(weather$datetime, weather$temp,
table(ride_start_days))

ggplot(rides_per_day)  +
  geom_bar(aes(x=ride_start_days, y=Freq),stat="identity",
fill="lightgreen",color="darkgreen")+
  geom_line(aes(x=ride_start_days, y=weather.temp *150, group =
1),stat="identity",color="red",size=2)+
  labs(title= "Bike Rides vs Temperature (in Fahrenheit)",
       x="day",y="Number of Bike Rides")+
  scale_y_continuous(sec.axis=sec_axis(
    ~./ 150,name="Temperature"))

```
~~~

This graph directly compares the temperature of the day (represented by the red line) to the number of bike rides that were taken on city bikes that day. As the graph shows, there seems to be some correlation. When the temperature dipped around the 12th of the month, there was an increase in bike rides, likely due to the fact that people may prefer to bike in slightly cooler weather. But on the 15th, the coolest day of the month, there was a sudden decrease in bike rides, which makes sense because as seen in graph number 8, that was a rainy day. Bike rides seem to relatively correlate with temperature, but precipitation also seems to play a role.

10. Concentration of Trip Starts

~~~
```{r, echo = FALSE, output = TRUE}
# Create query for average based on condition
res <- dbSendQuery(conn = dcon, "
              SELECT start_station_name, end_station_name, COUNT(*) as
num_trips
              FROM bikes b
~~~

~~~
              GROUP BY start_station_name, end_station_name
              ORDER BY (-1 * num_trips)
              LIMIT 10;
                   ")

# store results in df
station_travel <- dbFetch(res, -1)
dbClearResult(res)
# station_travel

ggplot(data = station_travel[-1,], aes(x = start_station_name, y = num_trips))
+
  geom_bar(stat = 'identity') +
  coord_flip() +
  ggtitle("Concentration of Trip Starts") +
  labs(x = "Start Station Name", y = "Number of Trips") +
  theme_minimal()
```
~~~

This graph shows the number of trips that start from a select number of stations. As the graph shows, there seems to be a high concentration of bike rides that start from Jefferson Dr & 14th St SW, which this station having the most bike rides by a large margin. Meanwhile, the least popular starting station seems to be Roosevelt Island. Jefferson Dr & 14th St SW may be a popular station where there is a higher concentration of bike riders, so investing in more bikes at this station, or creating more bike stations near this area may be a profitable move for this Bike Share company, while investing less in places near Roosevelt Island may not be.

11. Effects of Afternoon Rain on Bike Usage

~~~
```{r, echo = FALSE, output = TRUE}
## Join two dataframes based on start date
res <- dbSendQuery(conn = dcon, "
SELECT *
FROM (SELECT *, date(started_at) as date FROM bikes) as a, weather
WHERE a.date = weather.datetime
ORDER BY a.date")
mydf <- dbFetch(res,-1)
dbClearResult(res)

## Search Description for presence of After Noon Rain
mydf$after_noon_rain <- str_detect(mydf$description, "afternoon rain")
mydf$after_noon_rain <-  factor(mydf$after_noon_rain, labels = c("No Afternoon
Rain", "Afternoon Rain"))

#Extract Hour From Time Stamp
mydf$hour <- hour(mydf$started_at)

#Plot it
ggplot(data = mydf, aes(x = hour))+
  geom_histogram(stat = "count", fill = "deepskyblue3")+
  facet_wrap(~after_noon_rain, scales = "free")+
  labs(title = "Effects of Afternoon Rain on Bike Usage", y = "Trip Count", x
= "Hour of Day",
~~~

```
       facet = c("No Rain", "Rain"))+
  theme_bw()
```

This graph shows how the presence of afternoon rain will affect the distribution of bike usage throughout the day. When there is no Afternoon rain, there is usually a sharp spike in usage around 5-6pm. When there is rain, however, the hours from 7am-6pm see more even usage levels. The shape of the beginning or end of the day is still not effected very much.

12. Effects of Weather Description on Bike Rides

```{r, echo = FALSE, output = TRUE}
weather$cloudy <- as.factor(as.numeric(str_detect(weather$description,
"cloud")))
weather$clear <- as.factor(as.numeric(str_detect(weather$description,
"Clear")))
weather$rain <- as.factor(as.numeric(str_detect(weather$description, "rain")))
## Join two dataframes based on start date
res <- dbSendQuery(conn = dcon, "
SELECT *
FROM (SELECT *, date(started_at) as date FROM bikes) as a, weather
WHERE a.date = weather.datetime
ORDER BY a.date")
mydf <- dbFetch(res,-1)
dbClearResult(res)

table2$date <- str_sub(table2$started_at, 1, 10)
bikes <- table2

# aggregating by rides per day
by_date <- bikes %>%
  group_by(date) %>%
  summarise(total_rides = n())
weather$date <- as.Date(weather$datetime)
by_date$date <- by_date$date
dates <- by_date$date
by_date$date <- as.POSIXct(dates,tz=Sys.timezone())
# joining weather data with bike data
ride_weather <- inner_join(by_date, weather)

mod <- lm(total_rides ~ rain+clear+cloudy, data = ride_weather)
summary(mod)
# creating a substantive effect plot
dwplot(mod, vline = geom_vline(xintercept = 0, colour = "grey60", linetype =
2))%>%
  relabel_predictors(c(rain1= "Rain", clear1 = "Clear",cloudy1 = "Cloudy")) +
  ggtitle("Effects of Weather Description Factors on Rides")

```

These two graphs further expand on the effects of weather on the number of bike rides. This plot displays the average amount of bike rides based on the weather description for that day. AS the graph shows, when the day is described as having "rain" or "rainy", the average number of bike rides is significantly lower than on "cloudy" or clear" days. The averages for "clear" and "cloudy" days, however, are almost identical, with the range of the "clear" days being slightly larger, implying that there is more variation of bike rides on "clear" days. This would make sense, seeing as some days can be both "cloudy" and "rainy", which would decreased the number of rides on "cloudy" days.

As seen from the various plots, there are various patterns that can be observed within this bike share data, such as the popularity of classic bikes, and the more frequent usage by member as opposed to causual riders. This information could be crucial to the Washington DC city bike share company, as they could maximze profits by focusing advertising on their membership plan, or investing more in classic bikes the most riders seem to enjoy riding more.