Cyclist Bike Analysis

Vaibhav Kumar

2023-04-21

Problem Statement

Identify how annual members and casual riders use Cyclistic bikes differently.

Data Source

- Past 3 months cyclistic bike trip data for 2023 that has been made available by Motivate International Inc.
- This is public data that you can use to explore how different customer types are using Cyclistic bikes.

Data Processing

(Data processing has been done with the help of R programming.)

- importing csv files

```
df1 <- read.csv("C:/Users/Vaibhav/OneDrive/Desktop/Google Data
Analytics/Portfolio/Bike-Share Navigate Speedy Success/Working_Data/202301-
divvy-tripdata.csv", header = TRUE, sep = ",")
df2 <- read.csv("C:/Users/Vaibhav/OneDrive/Desktop/Google Data
Analytics/Portfolio/Bike-Share Navigate Speedy Success/Working_Data/202302-
divvy-tripdata.csv", header = TRUE, sep = ",")
df3 <- read.csv("C:/Users/Vaibhav/OneDrive/Desktop/Google Data
Analytics/Portfolio/Bike-Share Navigate Speedy Success/Working_Data/202303-
divvy-tripdata.csv", header = TRUE, sep = ",")
```

- Checking if data is correctly loaded

```
head(df1)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 F96D5A74A3E41399 electric_bike 2023-01-21 20:05:42 2023-01-21 20:16:33
## 2 13CB7EB698CEDB88  classic_bike 2023-01-10 15:37:36 2023-01-10 15:46:05
## 3 BD88A2E670661CE5 electric_bike 2023-01-02 07:51:57 2023-01-02 08:05:11
## 4 C90792D034FED968  classic_bike 2023-01-22 10:52:58 2023-01-22 11:01:44
## 5 3397017529188E8A  classic_bike 2023-01-12 13:58:01 2023-01-12 14:13:20
## 6 58E68156DAE3E311 electric_bike 2023-01-31 07:18:03 2023-01-31 07:21:16
```

```
##              start_station_name start_station_id
end_station_name
## 1   Lincoln Ave & Fullerton Ave     TA1309000058      Hampden Ct &
Diversey Ave
## 2         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave &
47th St
## 3        Western Ave & Lunt Ave           RP-005 Valli Produce - Evanston
Plaza
## 4         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave &
47th St
## 5         Kimbark Ave & 53rd St     TA1309000037        Greenwood Ave &
47th St
## 6 Lakeview Ave & Fullerton Pkwy     TA1309000019      Hampden Ct &
Diversey Ave
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1       202480.0  41.92407 -87.64628 41.93000 -87.64000        member
## 2   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 3            599  42.00857 -87.69048 42.03974 -87.69941        casual
## 4   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 5   TA1308000002  41.79957 -87.59475 41.80983 -87.59938        member
## 6       202480.0  41.92607 -87.63886 41.93000 -87.64000        member

colnames(df1)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

(There is a problem of formatting in column “start_station_id” and “end_station_id” and for the analysis purpose, I don’t need it. It will lead to error in combining the data frames in one, if not removed.)

- Dropping column 6 and 8 i.e. “start_station_id” and “end_station_id”.

```
df1 <- df1[, -c(6,8)]
df2 <- df2[, -c(6,8)]
df3 <- df3[, -c(6,8)]
```

- Checking data again to confirm if the columns are dropped or not.

```
head(df3)

##            ride_id rideable_type          started_at            ended_at
## 1 6842AA605EE9FBB3 electric_bike 2023-03-16 08:20:34 2023-03-16 08:22:52
```

```
## 2 F984267A75B99A8C electric_bike 2023-03-04 14:07:06 2023-03-04 14:15:31
## 3 FF7CF57CFE026D02  classic_bike 2023-03-31 12:28:09 2023-03-31 12:38:47
## 4 6B61B916032CB6D6  classic_bike 2023-03-22 14:09:08 2023-03-22 14:24:51
## 5 E55E61A5F1260040 electric_bike 2023-03-09 07:15:00 2023-03-09 07:26:00
## 6 123AAD676850F53C  classic_bike 2023-03-22 17:47:02 2023-03-22 18:01:29
##                     start_station_name             end_station_name
start_lat
## 1              Clark St & Armitage Ave    Larrabee St & Webster Ave
41.91841
## 2 Public Rack - Kedzie Ave & Argyle St
41.97000
## 3 Orleans St & Chestnut St (NEXT Apts)       Clark St & Randolph St
41.89820
## 4            Desplaines St & Kinzie St Sheffield Ave & Kingsbury St
41.88872
## 5                          Walsh Park        Sangamon St & Lake St
41.91448
## 6 Orleans St & Chestnut St (NEXT Apts)  Halsted St & Wrightwood Ave
41.89820
##   start_lng  end_lat   end_lng member_casual
## 1 -87.63645 41.92182 -87.64414        member
## 2 -87.71000 41.95000 -87.71000        member
## 3 -87.63754 41.88458 -87.63189        member
## 4 -87.64445 41.91052 -87.65311        member
## 5 -87.66801 41.88578 -87.65102        member
## 6 -87.63754 41.92914 -87.64908        member

colnames(df3)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "end_station_name"
##  [7] "start_lat"          "start_lng"          "end_lat"
## [10] "end_lng"            "member_casual"
```

For performing further analysis I've combined all the three data frames into one and then exported as a csv file to perform some calculations.

- Combining all data frames in one

```
df_final <- rbind(df1, df2, df3)
```

- Viewing the df_final dataset

```
View(df_final)
summary(df_final)

##    ride_id          rideable_type       started_at          ended_at
##  Length:639424      Length:639424      Length:639424      Length:639424
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##  start_station_name end_station_name     start_lat       start_lng
##  Length:639424      Length:639424      Min.   :41.65   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.89   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.53
##
##     end_lat         end_lng       member_casual
##  Min.   :41.63   Min.   :-87.90   Length:639424
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.89   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.08   Max.   :-87.52
##  NA's   :426     NA's   :426
```

- Dropping the rows with NA or Null values, as it might create some errors in results.

```
df_final <- na.omit(df_final)
```

- Exporting the df_final as csv file

```
write.csv(df_final, "exported_data.csv", row.names = FALSE)
```

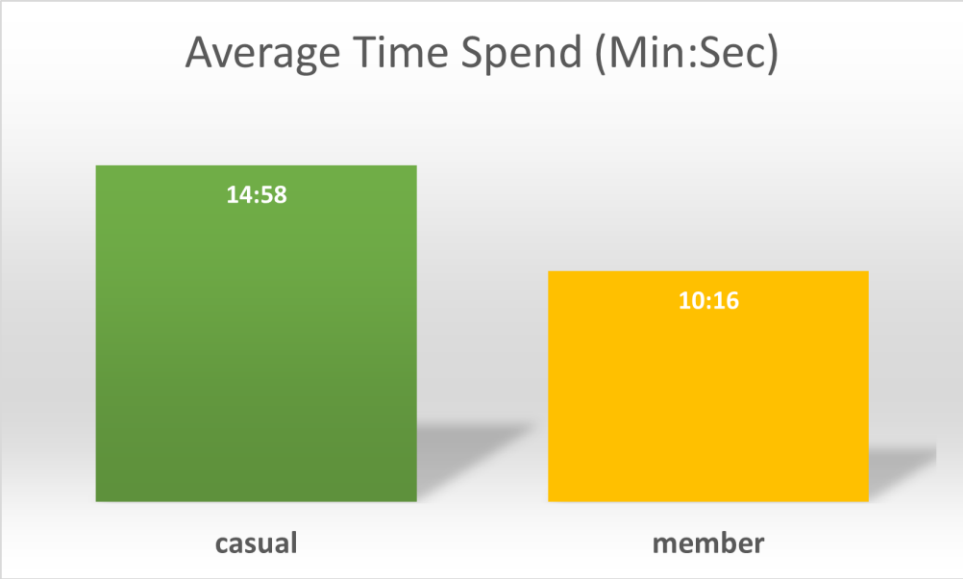
Operations performed in Excel

- A new column “time_length_min” was create using the formula “=D’row no.’-C’ row no.’” (D is ‘ended_at’ column , C is ‘started_at’ column).

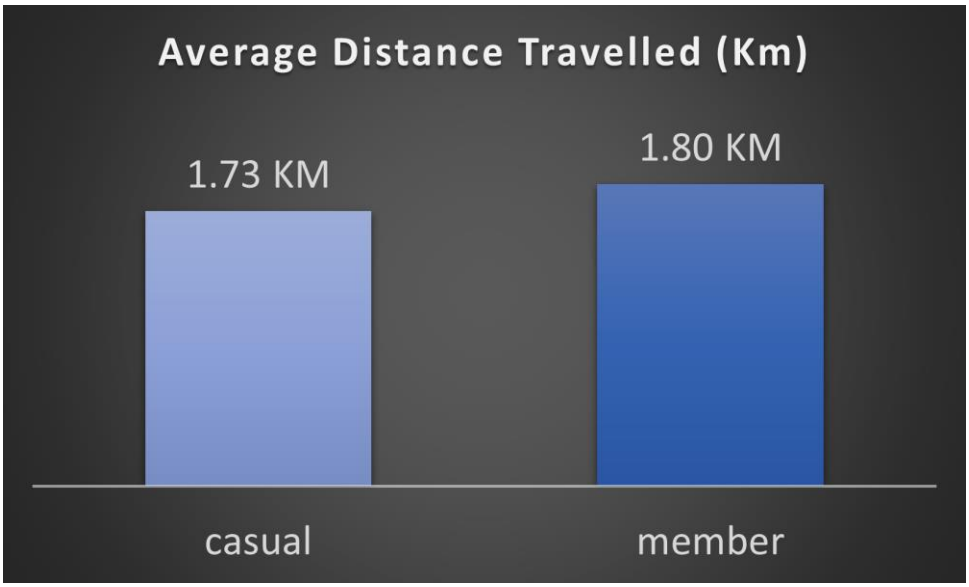
- Another new column “distance_travelled_km” was created using the formula “=ACOS(COS(RADIANS(90-G’row no.’))*COS(RADIANS(90-I’row no.’))+SIN(RADIANS(90-G’row no.’))*SIN(RADIANS(90-I’row no.’))*COS(RADIANS(H’row no.’-J’row no.’)))*6371” (G is ‘start_lat’ column, I is ‘end_lat’ column, H is ‘start_lng’ column, J is ‘end_lng’ column).

- Another new column “day_of_week” was created using the formula “=WEEKDAY(C’row no.’,1)” (C is ‘started_at’ column, noting that 1 = Sunday and 7 = Saturday)

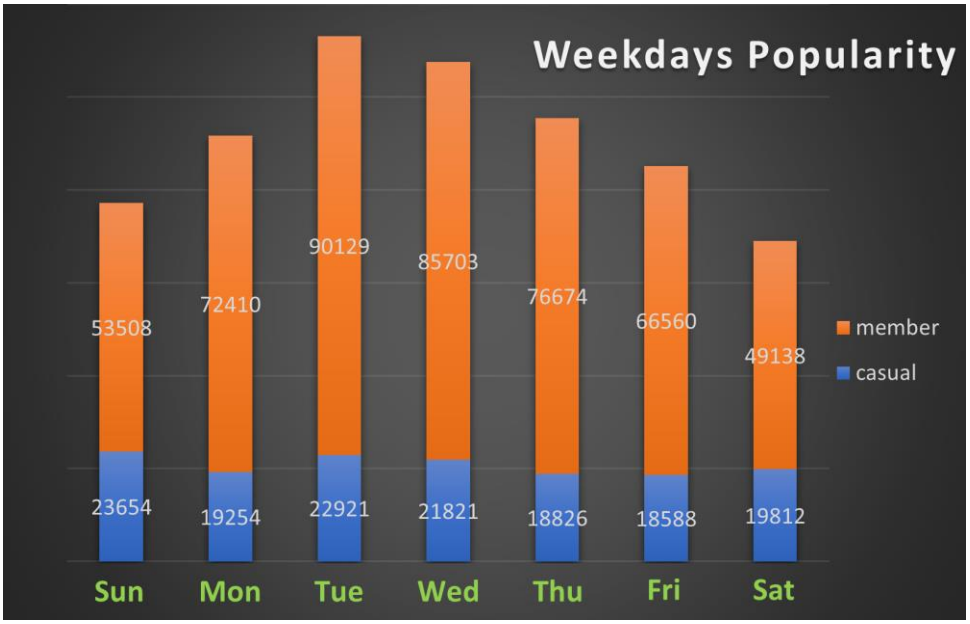
Pivot table operations was then conducted in order to get charts that are,



Average Time Spent on Bike



Average Distance Travelled on Bike



Popular Weekdays to ride a bike

Importing the dataset again to Rstudio in order to visualize

```
df_new <- read.csv("C:/Users/Vaibhav/OneDrive/Desktop/Google Data
Analytics/Portfolio/Bike-Share Navigate Speedy
Success/Working_Data/exported_data.csv", header = TRUE, sep = ",")
```

- Checking if data is correctly loaded

```
head(df_new)
```

```
##            ride_id rideable_type      started_at        ended_at
## 1 F96D5A74A3E41399 electric_bike 1/21/2023 20:05 1/21/2023 20:16
## 2 13CB7EB698CEDB88  classic_bike 1/10/2023 15:37 1/10/2023 15:46
## 3 BD88A2E670661CE5 electric_bike   1/2/2023 7:51   1/2/2023 8:05
## 4 C90792D034FED968  classic_bike 1/22/2023 10:52 1/22/2023 11:01
## 5 3397017529188E8A  classic_bike 1/12/2023 13:58 1/12/2023 14:13
## 6 58E68156DAE3E311 electric_bike  1/31/2023 7:18  1/31/2023 7:21
##              start_station_name              end_station_name start_lat
## 1   Lincoln Ave & Fullerton Ave      Hampden Ct & Diversey Ave  41.92407
## 2         Kimbark Ave & 53rd St        Greenwood Ave & 47th St  41.79957
## 3        Western Ave & Lunt Ave Valli Produce - Evanston Plaza  42.00857
## 4         Kimbark Ave & 53rd St        Greenwood Ave & 47th St  41.79957
## 5         Kimbark Ave & 53rd St        Greenwood Ave & 47th St  41.79957
## 6 Lakeview Ave & Fullerton Pkwy      Hampden Ct & Diversey Ave  41.92607
```

```
##   start_lng  end_lat   end_lng member_casual time_length_in_min
## 1 -87.64628 41.93000 -87.64000        member              10:51
## 2 -87.59475 41.80983 -87.59938        member              08:29
## 3 -87.69048 42.03974 -87.69941        casual              13:14
## 4 -87.59475 41.80983 -87.59938        member              08:46
## 5 -87.59475 41.80983 -87.59938        member              15:19
## 6 -87.63886 41.93000 -87.64000        member              03:13
##   distance_travelled_km day_of_week
## 1                  0.84           7
## 2                  1.20           3
## 3                  3.54           2
## 4                  1.20           1
## 5                  1.20           5
## 6                  0.45           3

colnames(df_new)

##  [1] "ride_id"               "rideable_type"         "started_at"
##  [4] "ended_at"              "start_station_name"    "end_station_name"
##  [7] "start_lat"             "start_lng"             "end_lat"
## [10] "end_lng"               "member_casual"         "time_length_in_min"
## [13] "distance_travelled_km" "day_of_week"
```

- Performing descriptive statistics on the df_new

```
max_result <- max(df_new$distance_travelled_km)
mode_result <- as.numeric(names(sort(table(df_new$day_of_week), decreasing =
TRUE)[1]))
cat("Mode value of days_of_week: ", mode_result, "\n")

## Mode value of days_of_week:  3

cat("Max value of distance_travelled_km: ", max_result, "\n")

## Max value of distance_travelled_km:  24.26
```

- Visualizing the number of rides per hour in order to find the rush hours.

```
options(repos = c(CRAN = "https://cran.rstudio.com"))

install.packages("dplyr")

## Installing package into 'C:/Users/Vaibhav/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Vaibhav\AppData\Local\R\win-
library\4.2\00LOCK\dplyr\libs\x64\dplyr.dll
## to C:\Users\Vaibhav\AppData\Local\R\win-
library\4.2\dplyr\libs\x64\dplyr.dll:
## Permission denied

## Warning: restored 'dplyr'

##
## The downloaded binary packages are in
##  C:\Users\Vaibhav\AppData\Local\Temp\Rtmp0wPcuf\downloaded_packages

install.packages("ggplot2")

## Installing package into 'C:/Users/Vaibhav/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Vaibhav\AppData\Local\Temp\Rtmp0wPcuf\downloaded_packages

install.packages("lubridate")

## Installing package into 'C:/Users/Vaibhav/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

## package 'lubridate' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'lubridate'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Vaibhav\AppData\Local\R\win-
library\4.2\00LOCK\lubridate\libs\x64\lubridate.dll
## to
## C:\Users\Vaibhav\AppData\Local\R\win-
library\4.2\lubridate\libs\x64\lubridate.dll:
## Permission denied

## Warning: restored 'lubridate'

##
## The downloaded binary packages are in
##  C:\Users\Vaibhav\AppData\Local\Temp\Rtmp0wPcuf\downloaded_packages

library(lubridate)

##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Converting the type of started_at
df_final$started_at <- lubridate::ymd_hms(df_final$started_at)

#New column with the hour at which the ride started
df_final$start_hour <- lubridate::hour(df_final$started_at)



# Count the number of rides per hour
rides_per_hour <- df_final %>%
  group_by(start_hour, member_casual) %>%
  summarise(num_rides = n())
```

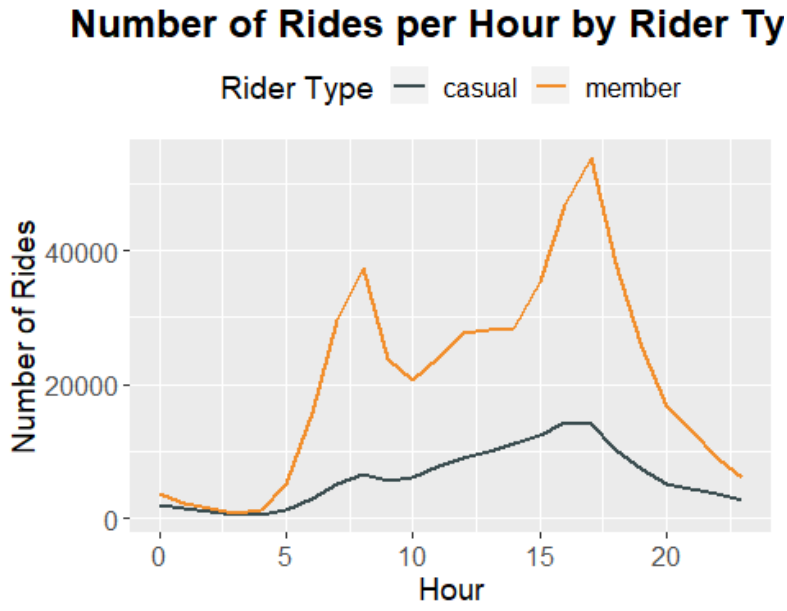
```
## `summarise()` has grouped output by 'start_hour'. You can override using
the
## `.groups` argument.
```

```
type_of_rider <- c(df_final$member_casual)

# Define color scheme
colors <- c("#3A4C4F", "#F28E2B")



# Count the number of rides per hour and rider type
rides_per_hour <- df_final %>%
  group_by(start_hour, member_casual) %>%
  summarise(num_rides = n())
```

```
## `summarise()` has grouped output by 'start_hour'. You can override using
the
## `.groups` argument.
```

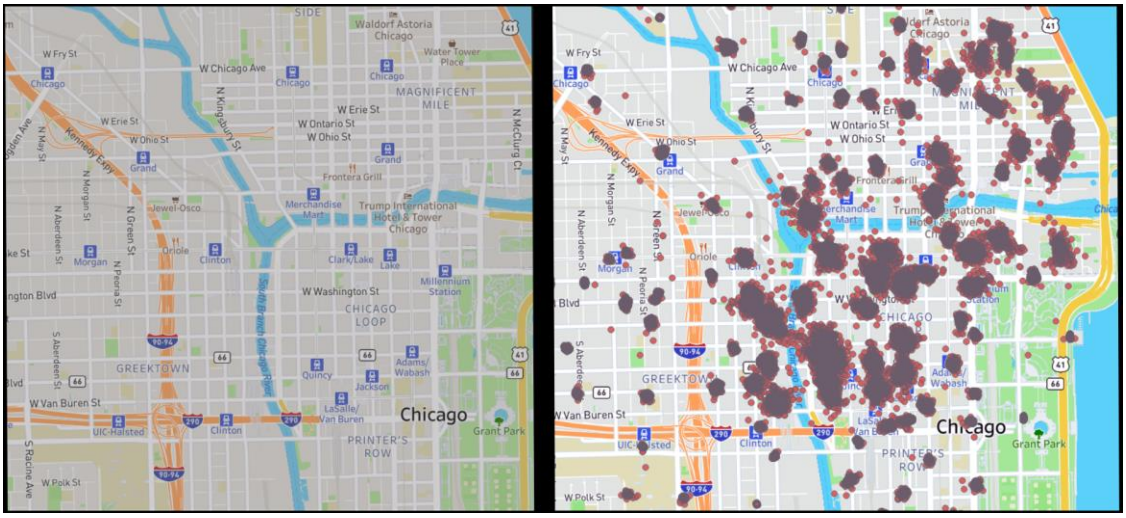
```r
# Create the line chart with custom x-axis labels and a legend based on rider
type
ggplot(data = rides_per_hour, aes(x = start_hour, y = num_rides, group =
member_casual, color = member_casual)) +
  geom_line(size = 1) +  # Increase line width

  scale_color_manual(values = colors) +  # Use custom color scheme
  labs(title = "Number of Rides per Hour by Rider Type", x = "Hour", y =
"Number of Rides", color = "Rider Type") +
  theme(plot.title = element_text(size = 18, face = "bold", hjust = 0.5),  #
Increase title font size and center it
        axis.title.x = element_text(size = 14),  # Increase x-axis label font
size
        axis.title.y = element_text(size = 14),  # Increase y-axis label font
size
        axis.text.x = element_text(size = 12),  # Increase x-axis tick label
font size
        axis.text.y = element_text(size = 12),  # Increase y-axis tick label
font size
        legend.position = "top",  # Move legend to top
        legend.title = element_text(size = 14),  # Increase legend title font
size
        legend.text = element_text(size = 12))  # Increase legend text font
size
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Importing dataset to Tableau

- This chart shows the distribution of the riders starting point.



Starting Point of Riders

- This chart shows the frequent riders starting point. It has been calculated in tableau using the most frequent occurrence of the starting station. Then the filter has been applied to just get the data for more than 3000 count (5906 was the most count occurred).

Most Popular Starting Point

- This chart shows the frequent starting points of the casual rider. It has been calculated in tableau using the most frequent occurrence of the starting station. Then the filter has been applied to just get the data for more than 3000 count (5906 was the most count occurred) and just for casual riders.

Most Popular Starting Point of Casual Riders

[All these analysis is then summarized and put up in the presentation to give recommendations to solve the particular problem.]

Connect with me on Linkedin - Click Here