Case Study Cyclistic

unknownda

2022-07-17

Setting up the data frame

Installing and loading *readr*, *tidyr*, *lubridate* and *ggplot2* packages. Then import the data from the four quarters.

Cleaning Data

```
# Exploring the structure of all four datasets

str(Trips_Q1_temp)
str(Trips_Q2_temp)
str(Trips_Q3_temp)
str(Trips_Q4_temp)

# renaming column names of Trips_Q2_temp

col_names <- names(Trips_Q1_temp)
names(Trips_Q2_temp) <- col_names

# mutating to add a new column for Quarter

Trips_Q1_temp <- Trips_Q1_temp %>% mutate(quarter=1)
Trips_Q2_temp <- Trips_Q2_temp %>% mutate(quarter=2)
Trips_Q3_temp <- Trips_Q3_temp %>% mutate(quarter=3)
Trips_Q4_temp <- Trips_Q4_temp %>% mutate(quarter=4)

# combining datasets of all quarters

Trips <- rbind(Trips_Q1_temp,Trips_Q2_temp,Trips_Q3_temp,Trips_Q4_temp)

# changing column types of Trips

Trips$trip_id <- as.integer(Trips$trip_id)
Trips$bikeid <- as.integer(Trips$bikeid)
Trips$from_station_id <- as.integer(Trips$from_station_id)
Trips$to_station_id <- as.integer(Trips$to_station_id)
Trips$usertype <- as.factor(Trips$usertype)
Trips$gender <- as.factor(Trips$gender)
```

```r
# checking number of unique trip_id

nrow_Trips <- nrow(Trips)
no_unique_trip_id_Trips <- NROW(unique(Trips$trip_id))
if(nrow_Trips==no_unique_trip_id_Trips) print("True")

# checking number of NA's in each column

find.na <- function(x){
  sum(is.na(x))
}

nas <- apply(Trips,2,find.na)

# removing nas in gender column

gender_levels <- levels(Trips$gender)
gender_levels[length(gender_levels)+1] <- "None"
Trips$gender <- factor(Trips$gender, levels = gender_levels)
Trips$gender[is.na(Trips$gender)] <- "None"

nas <- apply(Trips,2,find.na)

# checking trip duration

td <- as.double(difftime(Trips$end_time, Trips$start_time, units = "sec"))
identical(Trips$tripduration, td, num.eq=TRUE)

# transform incorrect trip durations

idx <- which(Trips$tripduration!=td)
Trips$tripduration[idx] <- td[idx]
identical(Trips$tripduration, td, num.eq = TRUE)

# ignoring nas in birthyear

Trips_cleaned <- Trips

# mutate to include weekday

Trips_cleaned <- Trips_cleaned %>% mutate(day_of_week=wday(start_time, label=TRUE))
```

Analyzing Data

```r
stat_tripduration <- Trips_cleaned %>% summarize(mean=mean(tripduration), max=max(tripduration))

mode<-function(x){
  which.max(tabulate(x))
  }

stat_day_of_week_idx <- mode(Trips_cleaned$day_of_week)
```

```
stat_day_of_week <- levels(Trips_cleaned$day_of_week)[stat_day_of_week_idx]

stat_usertype <- Trips_cleaned %>%
  group_by(usertype, gender) %>%
  summarize(average_tripduration=mean(tripduration))
```

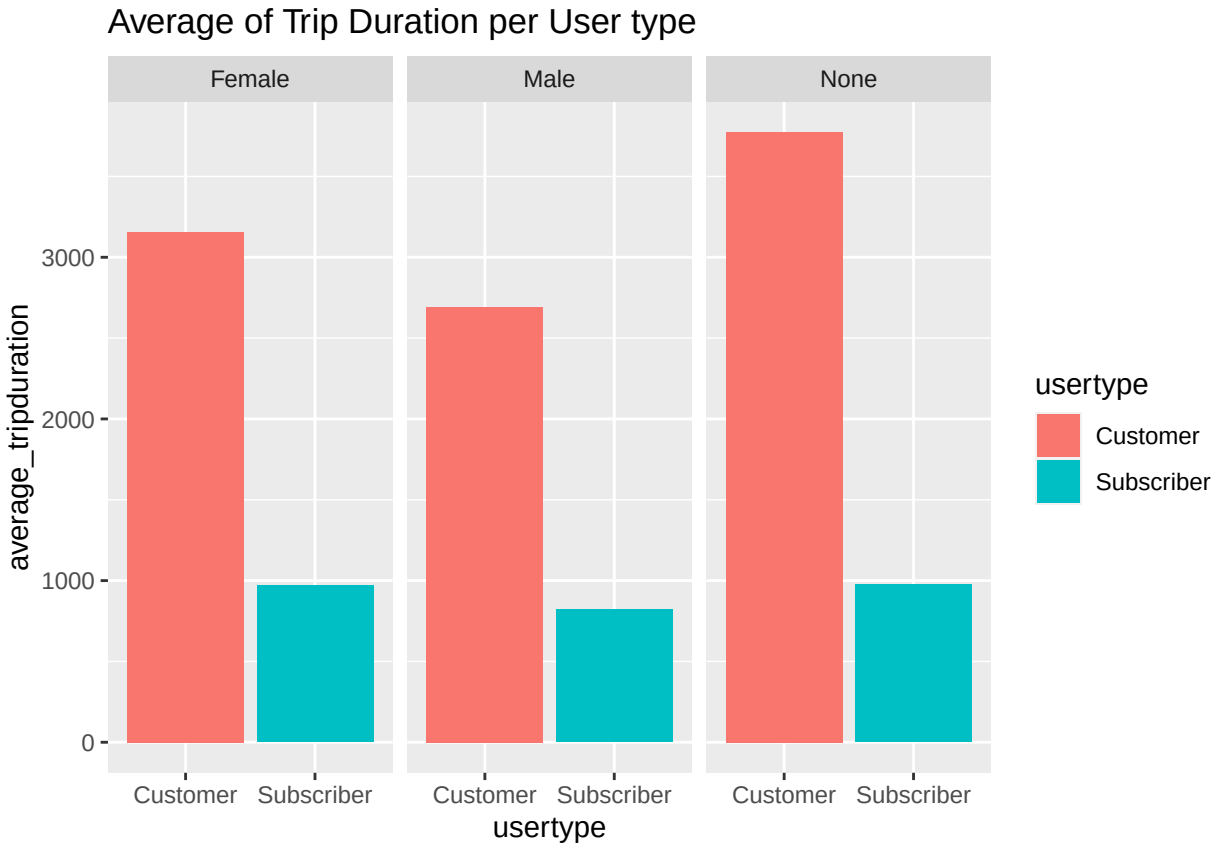
```
## `summarise()` has grouped output by 'usertype'. You can override using the
## `.groups` argument.
```

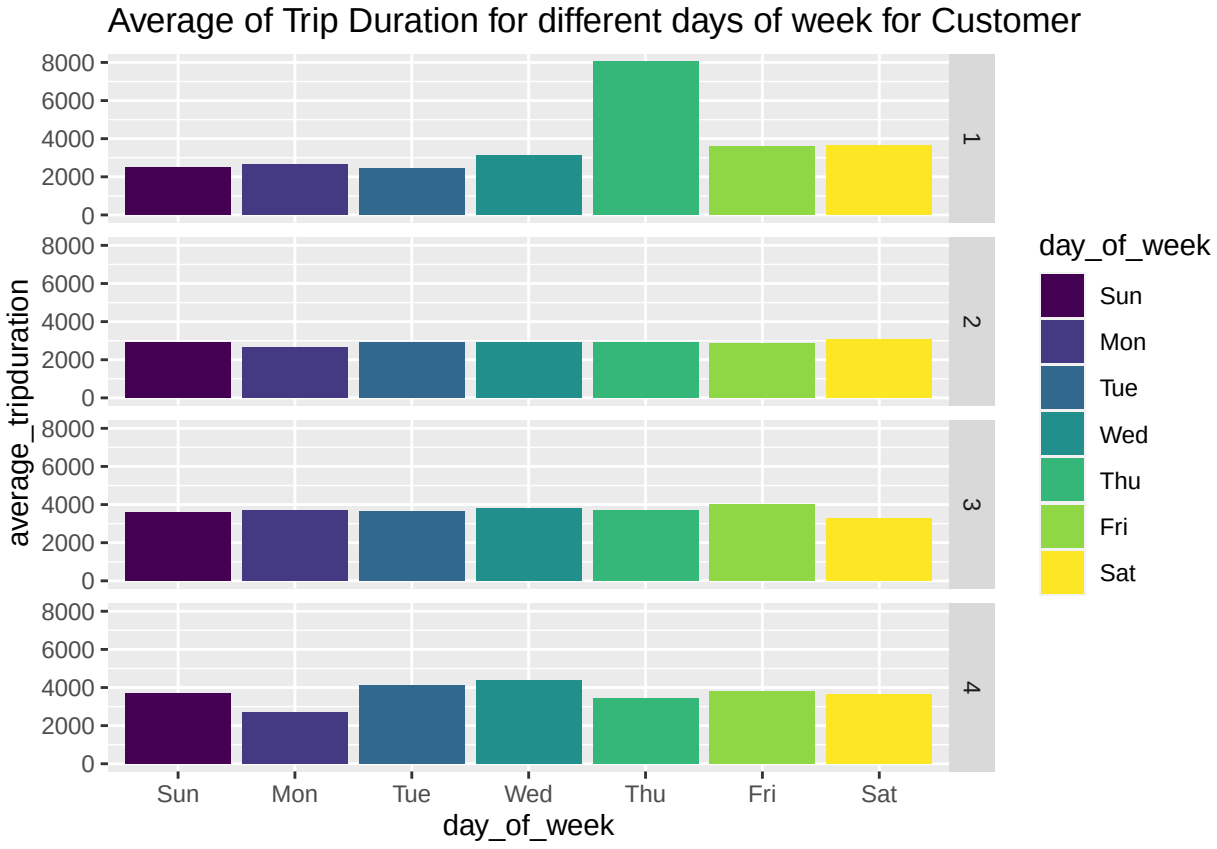
```
stat2_day_of_week<- Trips_cleaned %>%
  group_by(quarter, usertype, day_of_week) %>%
  summarize(average_tripduration=mean(tripduration), number_of_rides=n())
```

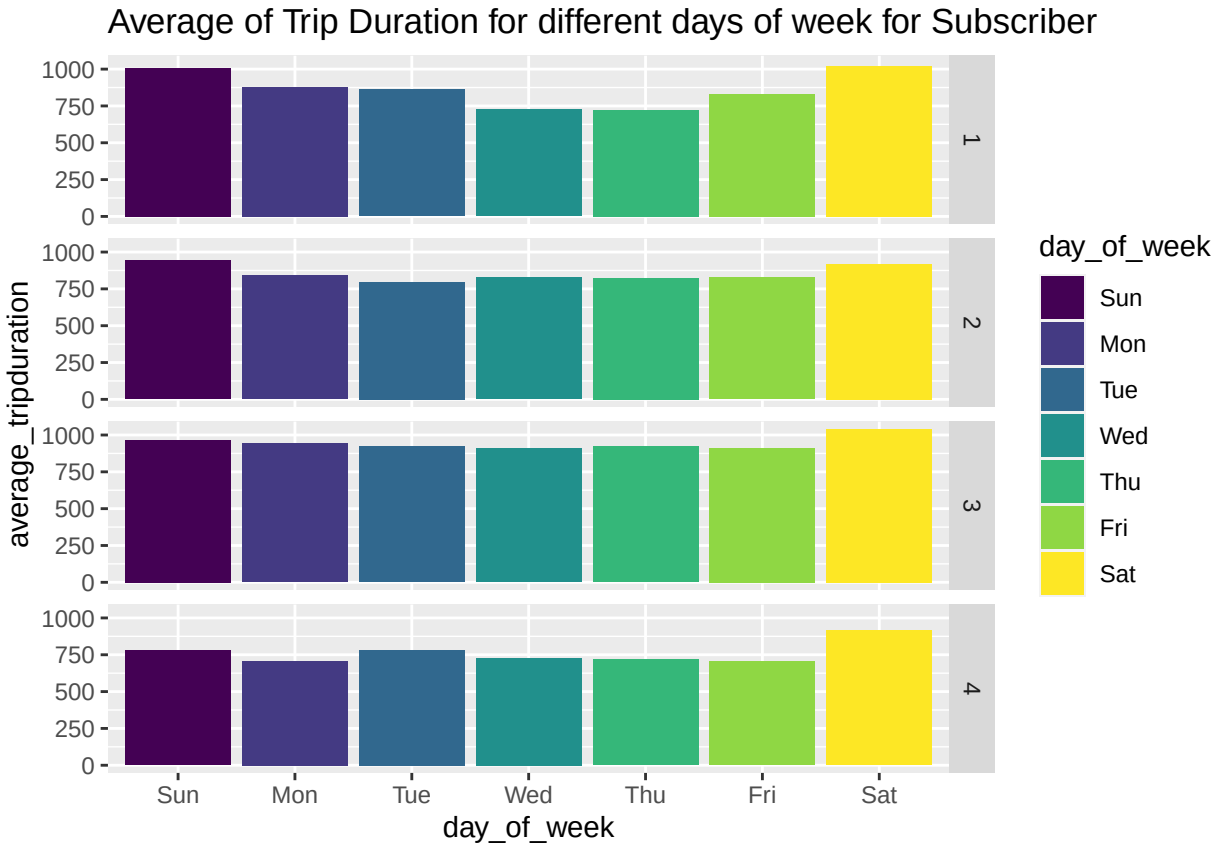
```
## `summarise()` has grouped output by 'quarter', 'usertype'. You can override
## using the `.groups` argument.
```

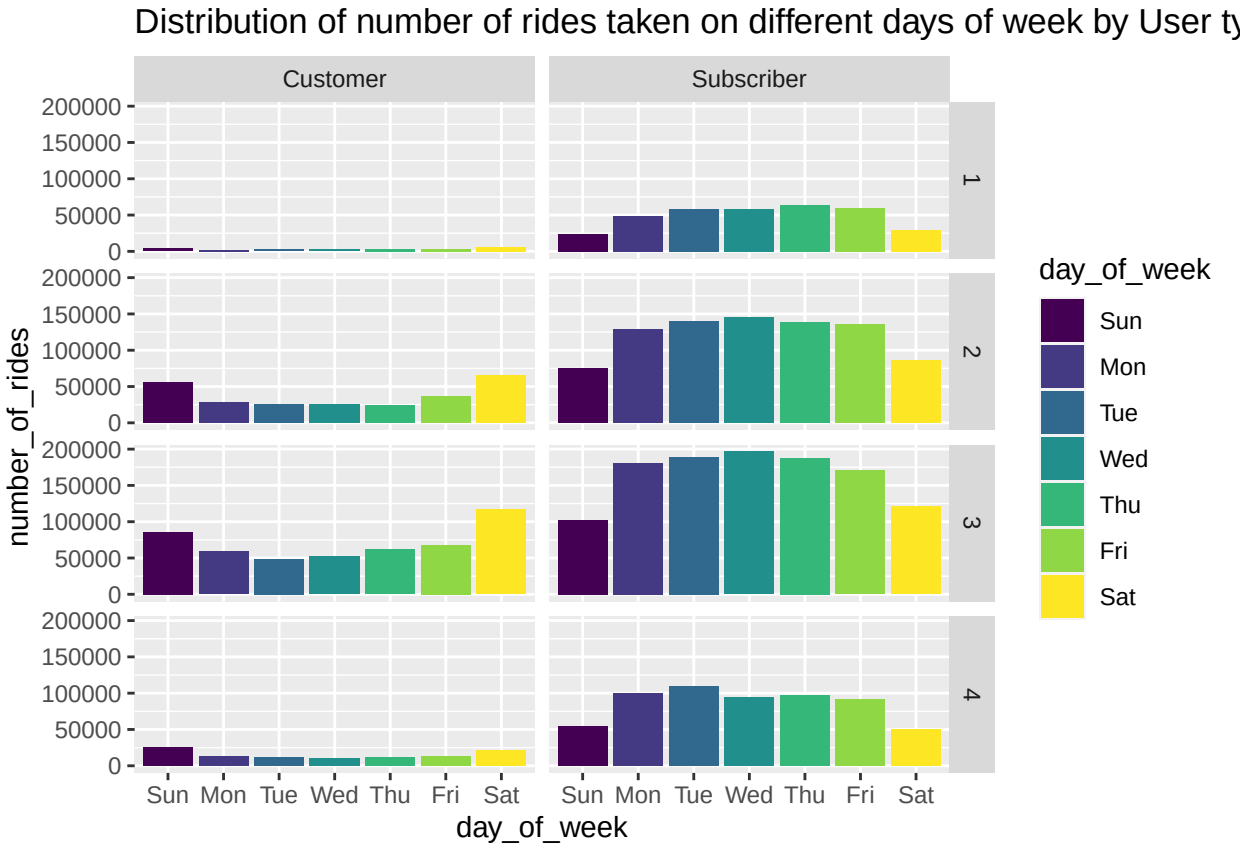
```
stat3_day_of_week <- Trips_cleaned %>%
  group_by(day_of_week) %>%
  count(name="number_of_rides")

stat_gender <- Trips_cleaned %>%
  group_by(usertype, gender) %>%
  count(name="number_of_rides")
```

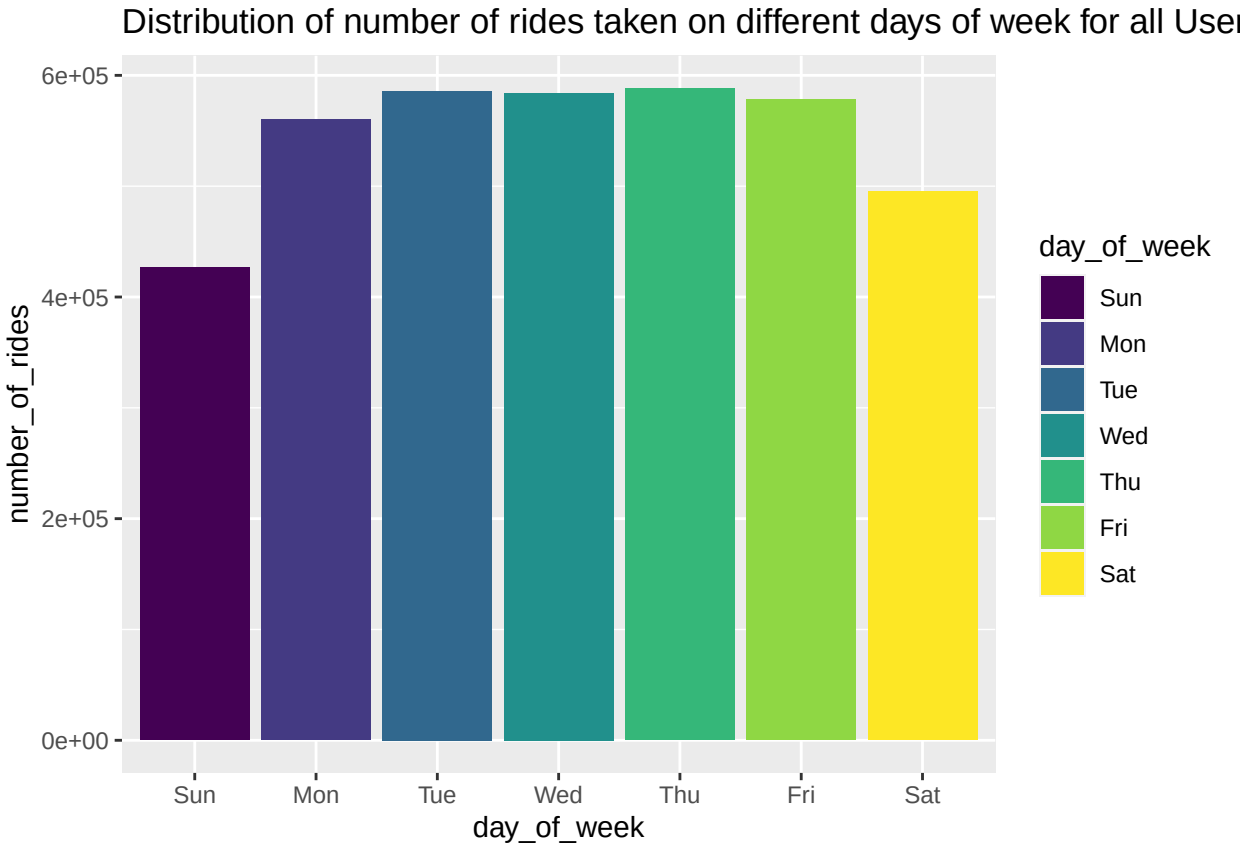
Visualizing Data

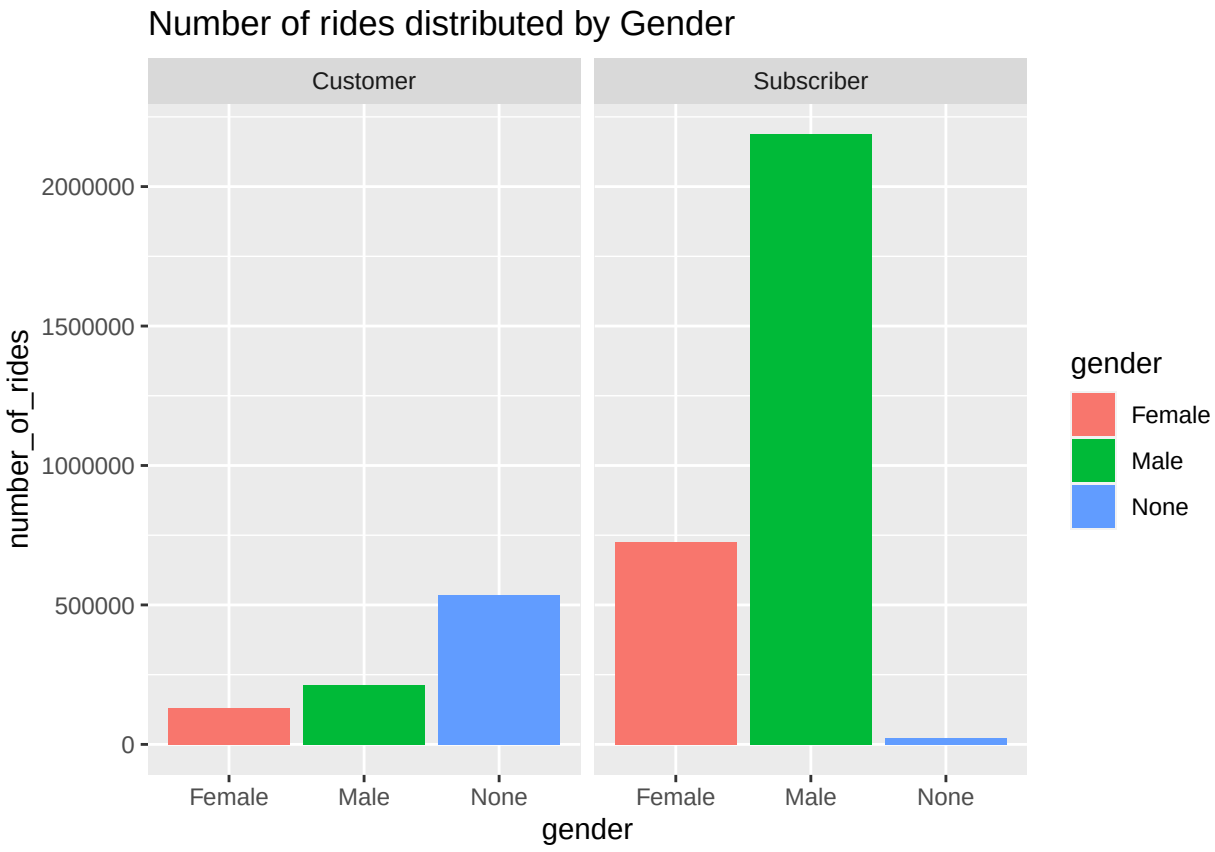





Average of Trip Duration for different days of week for Subscriber
average_tripduration
1000
750
500
250
0
1
2
3
4
Sun
Mon
Tue
Wed
Thu
Fri
Sat
day_of_week
day_of_week
Sun
Mon
Tue
Wed
Thu
Fri
Sat


Distribution of number of rides taken on different days of week by User ty
Customer
Subscriber
1
2
3
4
200000
150000
100000
50000
0
number_of_rides
Sun
Mon
Tue
Wed
Thu
Fri
Sat
day_of_week
day_of_week
Sun
Mon
Tue
Wed
Thu
Fri
Sat


Distribution of number of rides taken on different days of week for all User
6e+05
4e+05
2e+05
0e+00
number_of_rides
Sun
Mon
Tue
Wed
Thu
Fri
Sat
day_of_week
day_of_week
Sun
Mon
Tue
Wed
Thu
Fri
Sat



Conclusion

- Average trip duration is highest for customers having “None” gender.
- Average trip duration for customers is much larger than that of subscribers.
- Customers ride longest on Thursdays.
- Subscribers ride longest on weekends.
- Subscribers ride more times than customers in general.
- Peak subscriber traffic is in weekdays.
- Peak subscriber season is Q3.
- In totality, there more number of rides on weekdays.
- Most Customers are “None” and most Subscribers are “Males”.

Recommendations

- Give heavy discounts to Subscribers in the third quarter of the year.
- Provide incentives to Subscribers if they ride on weekends.
- Design advertisements keeping Male attributes in mind.