Capstone

Benjamin T Burke

2023-04-18

Background Scenario

Cyclistic is an imaginary bike-sharing company that operates in Chicago, Illinois. My assignment was to work as a data analyst on the marketing team. Cyclistic has two types of customers: annual members and casual riders. The director of marketing believes that the company's future success depends on maximizing the number of members. She asks the marketing team to design a new strategy to convert casual riders into annual members based on an analysis of how casual riders and annual members use Cyclistic bikes differently.

Setting up my Environment

```
install.packages ("tidyverse")
```

```
## Error in install.packages : Updating loaded packages
```

```
install.packages ("readr")
```

```
## Error in install.packages : Updating loaded packages
```

```
install.packages("sqldf")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(tidyverse)
library(ggplot2)
library(tidyr)
library(readr)
library(dplyr)
library(data.table)
library(lubridate)
library(sqldf)
library(hms)
```

Importing Data

This data is up to date in this fictional setting and has been given to us by a made up reliable source.

```
trip_data <- read.csv("tripdata.csv")
View(trip_data)
```

Cleaning Data

Creating a trimmed up version of the data set

```
trimmed_trip_data <- select(trip_data, rideable_type, started_at, ended_at, member_casual, ride_length,
View(trimmed_trip_data)
```

Create new data frame to contain new info

```
new_trip_data <- trimmed_trip_data
```

Create ride_length column

```
new_trip_data$ended_at <- as.POSIXct(new_trip_data$ended_at, format = "%m/%d/%Y %H:%M")
new_trip_data$started_at <- as.POSIXct(new_trip_data$started_at, format = "%m/%d/%Y %H:%M")
new_trip_data$ride_length <- difftime(new_trip_data$ended_at, new_trip_data$started_at, units = "mins")
new_trip_data$ride_length <- round(new_trip_data$ride_length, digits = 1)
```

Create columns for all date calculations

```
new_trip_data$date <- as.Date(new_trip_data$started_at)
new_trip_data$day_of_week <- weekdays(new_trip_data$started_at) #day of week calculation
new_trip_data$day_of_week <- format(as.Date(new_trip_data$date),"%A") #day of week column
new_trip_data$month <- format(as.Date(new_trip_data$date), "%m") #month column
new_trip_data$day <- format(as.Date(new_trip_data$date), "%d") #day column
new_trip_data$year <- format(as.Date(new_trip_data$date), "%Y") #year column
new_trip_data$time <- format(as.Date(new_trip_data$date), "%H:%M:%S") #time formatted HH:MM:SS
new_trip_data$time <- as_hms(new_trip_data$started_at) #time column
new_trip_data$hour <- hour(new_trip_data$time) #create new column for hour
```

Remove where ride_length is 0, negative, or duplicated

```
new_trip_data <- na.omit(new_trip_data) #remove rows with NA
new_trip_data <- distinct(new_trip_data) #remove duplicate rows
new_trip_data <- new_trip_data[!(new_trip_data$ride_length <=0),]
```

Analyzing Data

Counting Member types

```
count_member <- sqldf("SELECT COUNT(member_casual)
                      FROM new_trip_data
                      WHERE member_casual='member' ")
count_member
```

```
##   COUNT(member_casual)
## 1               171822
```

```
count_casual <- sqldf("SELECT COUNT(member_casual)
                      FROM new_trip_data
                      WHERE member_casual='casual' ")
count_casual
```

```
##   COUNT(member_casual)
## 1                57916
```

- Member Total: 171822
- Casual Total: 57916
- Average ride: 13.90983 mins
- Most Popular Day: Wednesday

Counting Bike Types

```
count_e <- sqldf("SELECT COUNT(rideable_type)
                        FROM new_trip_data
                        WHERE rideable_type='electric_bike' ")
count_e
```

```
##   COUNT(rideable_type)
## 1               129924
```

```
count_c <- sqldf("SELECT COUNT(rideable_type)
                        FROM new_trip_data
                        WHERE rideable_type='classic_bike' ")
count_c
```

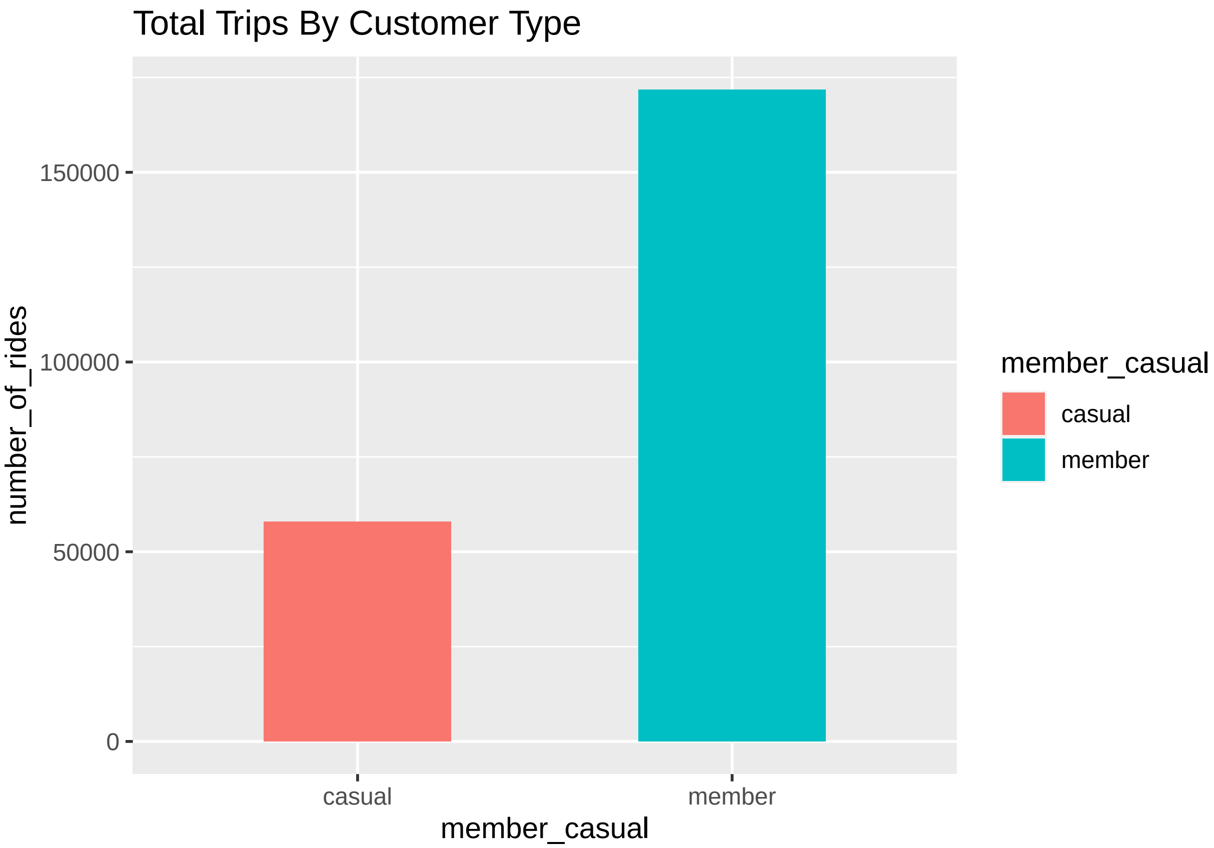
```
##   COUNT(rideable_type)
## 1                97099
```

- Electric Total: 129924
- Classic Total: 97099

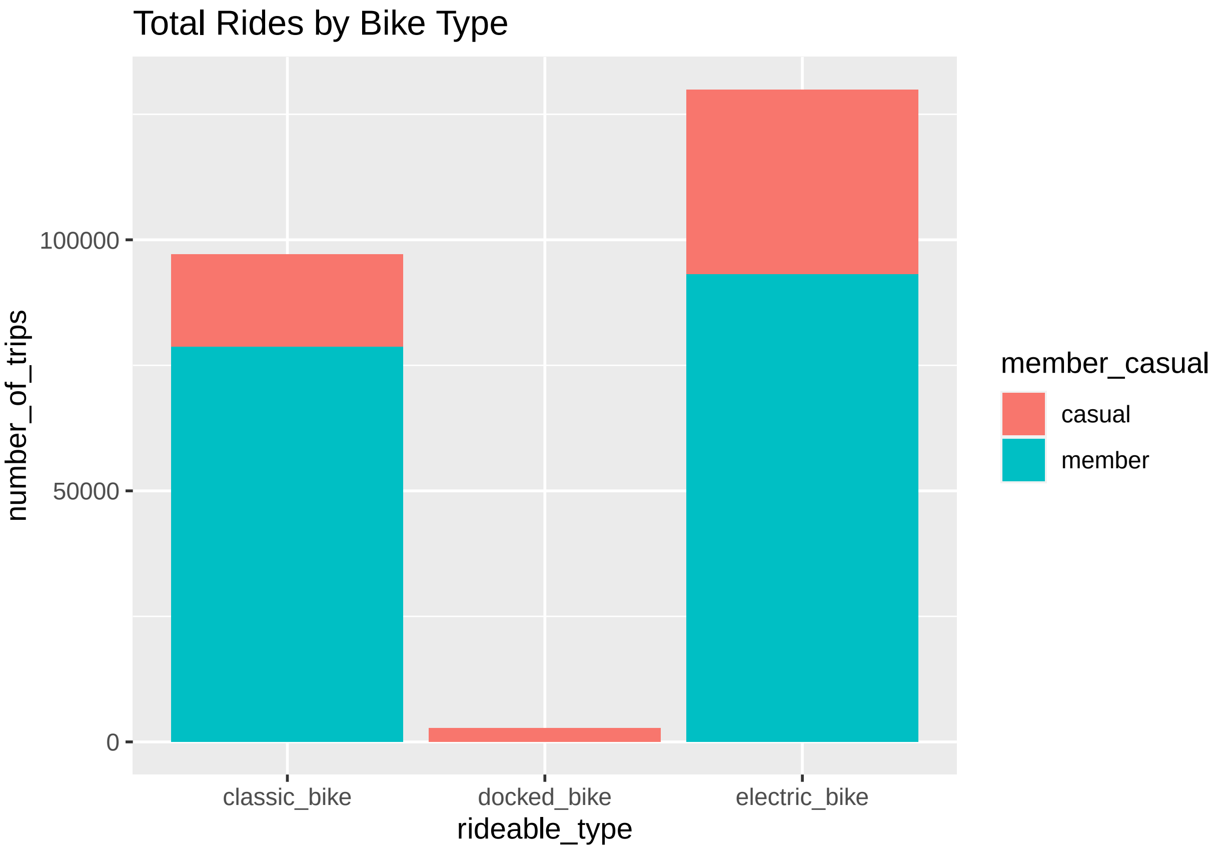
Most Popular Bike Via Member - Casual: Electric - Member: Electric

Visuals

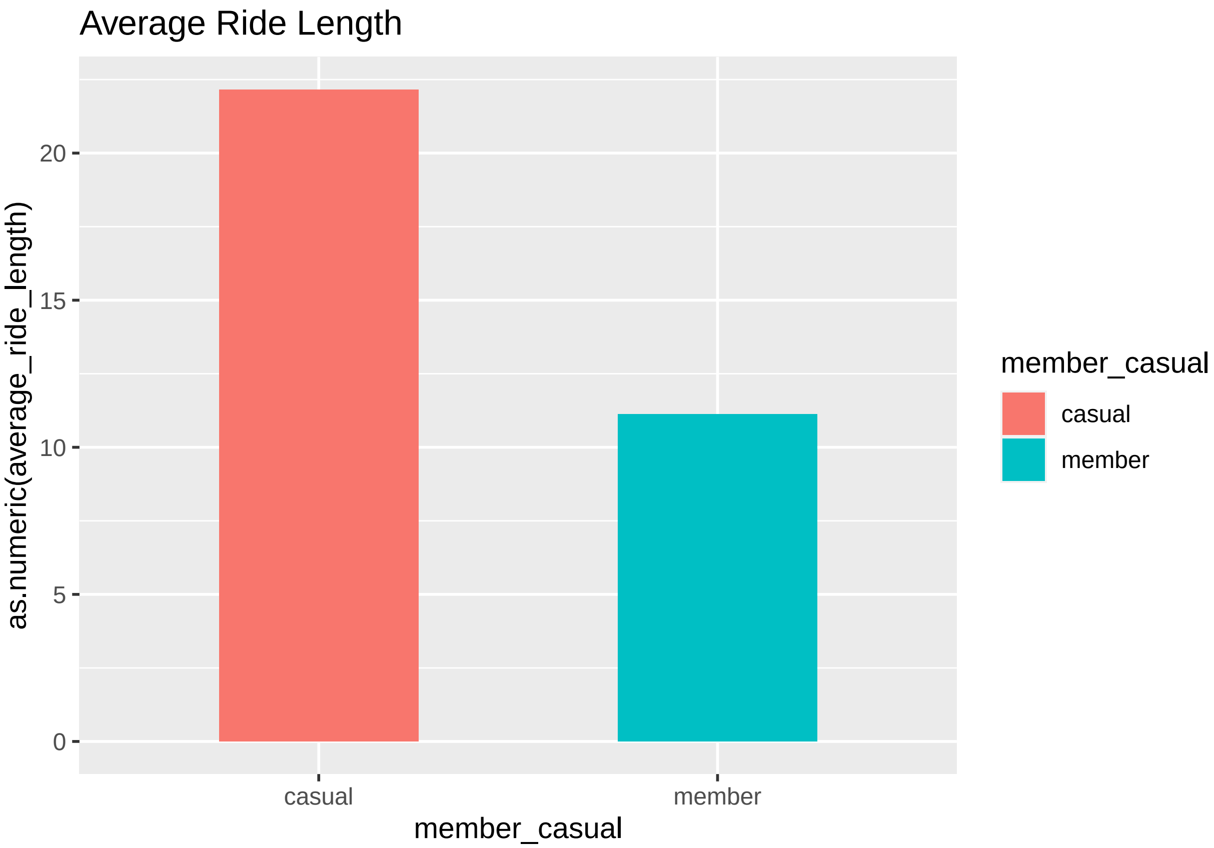
```
new_trip_data %>%
     group_by(member_casual)%>%
     summarize(number_of_rides = n())%>%
     arrange(member_casual)%>%
     ggplot(aes(x = member_casual, y = number_of_rides, fill = member_casual)) +
     labs(title = "Total Trips By Customer Type") +
     geom_col(width = 0.5, position = position_dodge(width = 0.5)) +
     scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



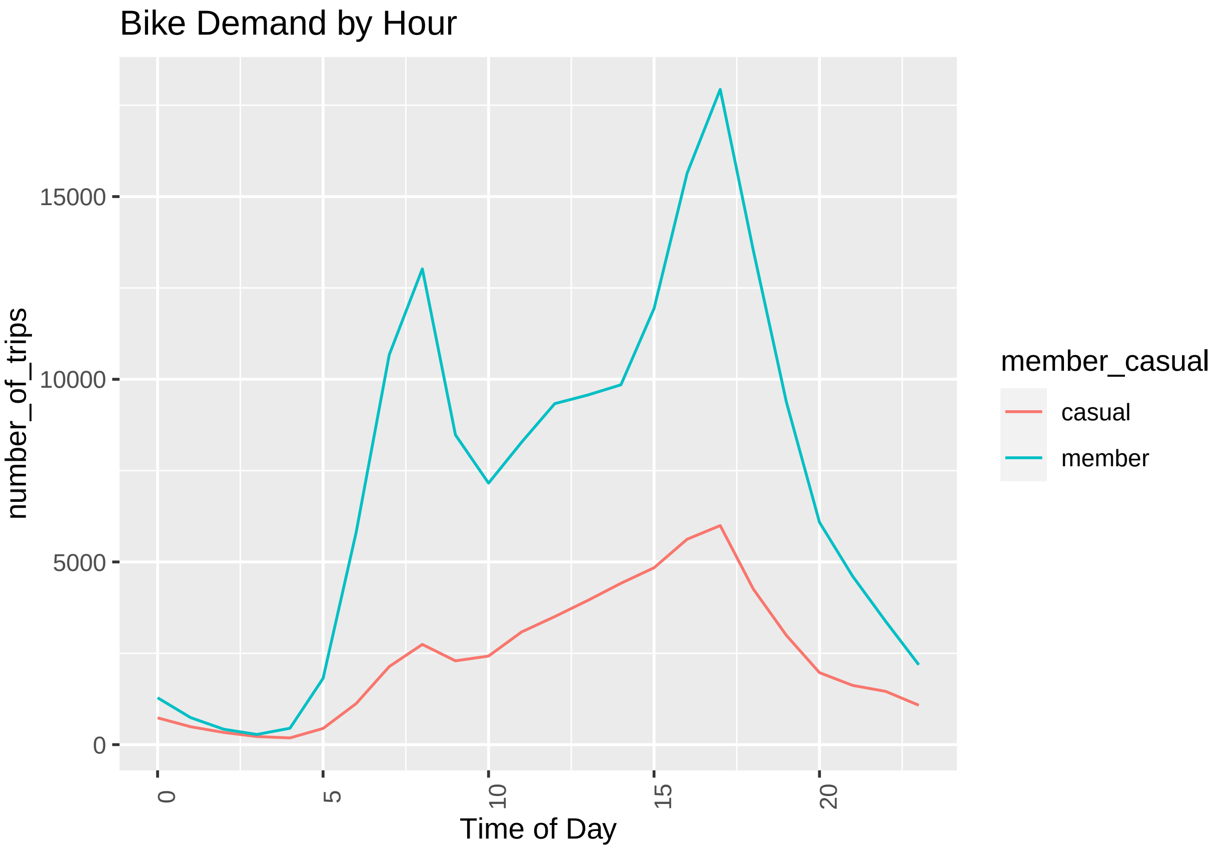
```
new_trip_data %>%
     group_by(rideable_type, member_casual)%>%
     summarize(number_of_trips = n(), .groups = "drop")%>%
     ggplot(aes(x = rideable_type, y = number_of_trips, fill = member_casual)) +
     geom_bar(stat = 'identity') +
     labs(title = "Total Rides by Bike Type") +
     scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```


Total Rides by Bike Type

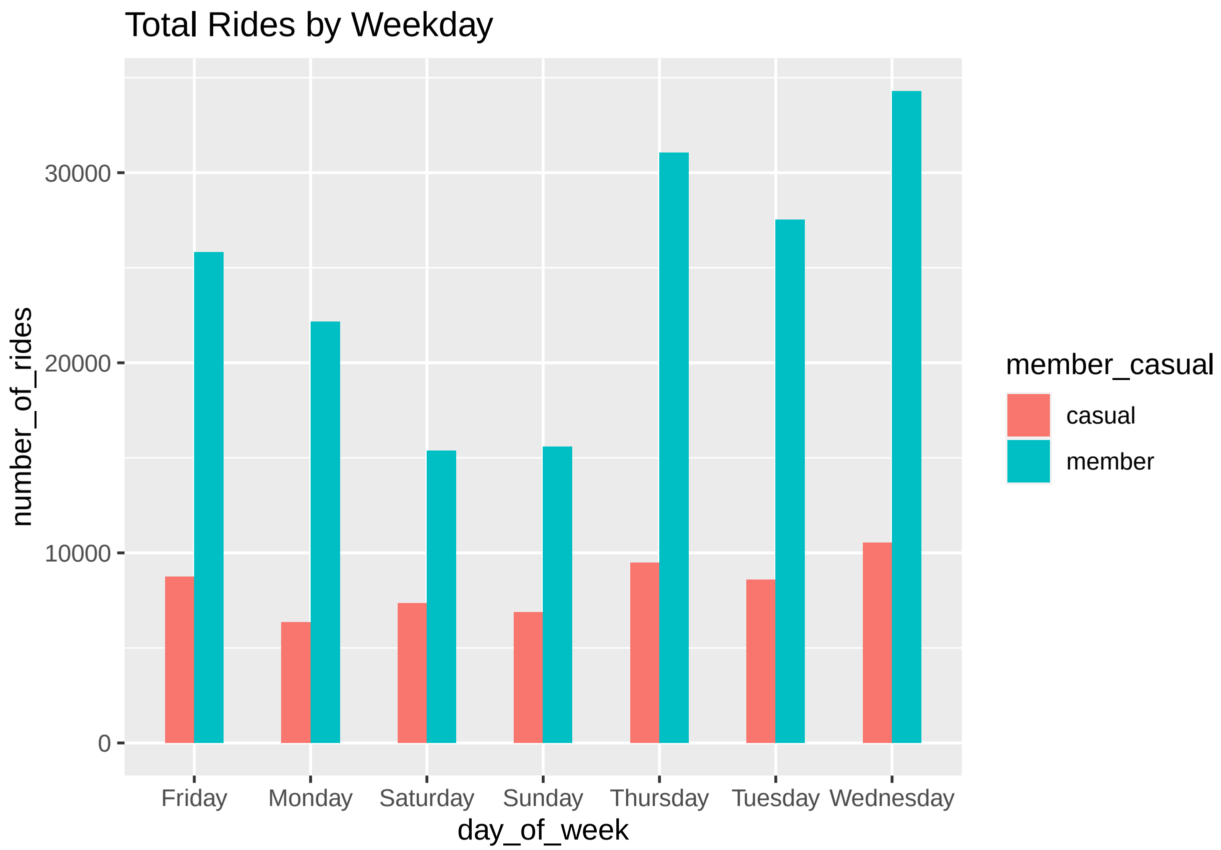
```
new_trip_data %>%
  group_by(member_casual)%>%
  summarize(average_ride_length = mean(ride_length))%>%
  ggplot(aes(x = member_casual, y = as.numeric(average_ride_length), fill = member_casual)) +
  labs(title = "Average Ride Length") +
  geom_col(width = 0.5, position = position_dodge(width = 0.5))
```



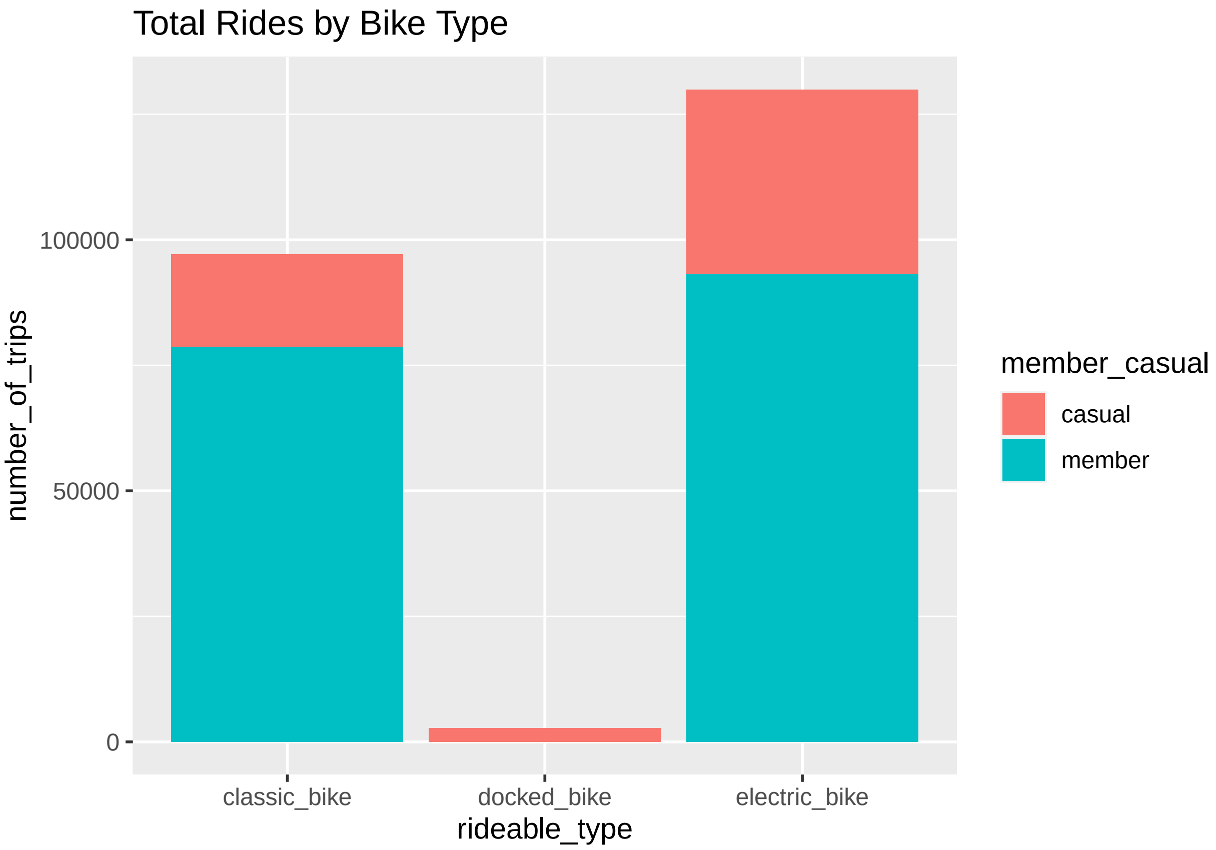
```
new_trip_data %>%
  group_by(member_casual, hour)%>%
  summarize(number_of_trips = n(), .groups = 'drop')%>%
  ggplot(aes(x = hour, y = number_of_trips, color = member_casual, group = member_casual)) +
  geom_line() +
  labs(title = "Bike Demand by Hour", x = "Time of Day") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



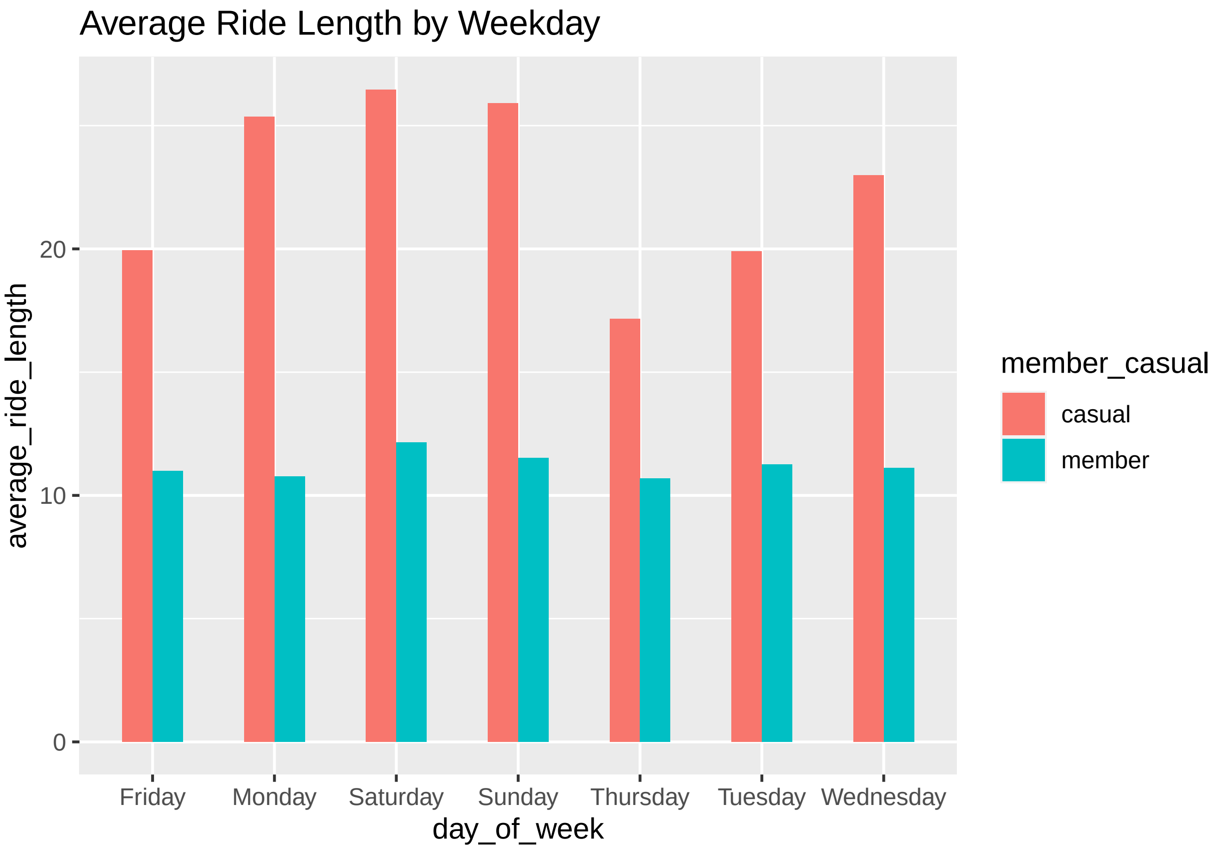
```
new_trip_data %>%
  group_by(member_casual, day_of_week)%>%
  summarize(number_of_rides = n(), .groups = 'drop')%>%
  arrange(member_casual, day_of_week)%>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  labs(title = "Total Rides by Weekday") +
  geom_col(width = 0.5, position = position_dodge(width = 0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



```
new_trip_data %>%
  group_by(rideable_type, member_casual)%>%
  summarize(number_of_trips = n(), .groups = 'drop')%>%
  ggplot(aes(x = rideable_type, y = number_of_trips, fill = member_casual)) +
  geom_bar(stat = 'identity') +
  labs(title = "Total Rides by Bike Type") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



```
new_trip_data %>%
  mutate(ride_length_seconds = as.numeric(ride_length)) %>%
  group_by(member_casual, day_of_week) %>%
  summarize(average_ride_length = mean(ride_length_seconds), .groups = 'drop') %>%
  ggplot(aes(x = day_of_week, y = average_ride_length, fill = member_casual)) +
  labs(title = "Average Ride Length by Weekday") +
  geom_col(width = 0.5, position = position_dodge(width = 0.5))
```



Findings

- We find that members compared to casual users are more consistent in the usage of the bikes
- Casual users are more likely to ride for more on the weekends, this suggest that casual user use this service more for recreational purposes then the member users who use it most likely for work commute
- The electric bikes are more popular to use for both member types
- Wednesday is the most popular day in total amount of rides
- Between 3pm and and 6pm is the most popular time for bike riding