Cyclist Bike Share Analysis

Details of the Case Study

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analysing the Cyclistic historical bike trip data to identify trends.

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members.

But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Moreno has assigned you the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

Business Task

Aim: To analyse rider's usage patterns for marketing membership conversion programs.

Tools**:** R for data cleaning and data visualization.

Dataset**:** [Cyclistic's historical trip data from December 2021 to November 2022]

Approach

1. Data Cleaning

```
#Load libraries
library(tidyverse) #calculations
library(lubridate) #dates
library(hms) #time
library(data.table) #exporting data frame
library(readxl)

#Load original .csv files, a years worth of data from Dec 2021 to Nov 2022
dec <- read.csv("R:/Rutuja/Projects/R Project/Data/202112-divvy-tripdata.csv")
jan <- read.csv("R:/Rutuja/Projects/R Project/Data/202201-divvy-tripdata.csv")
feb <- read.csv("R:/Rutuja/Projects/R Project/Data/202202-divvy-
tripdata.csv")
mar <- read.csv("R:/Rutuja/Projects/R Project/Data/202203-divvy-tripdata.csv")
apr <- read.csv("R:/Rutuja/Projects/R Project/Data/202204-divvy-tripdata.csv")
may <- read.csv("R:/Rutuja/Projects/R Project/Data/202205-divvy-tripdata.csv")
jun <- read.csv("R:/Rutuja/Projects/R Project/Data/202206-divvy-tripdata.csv")
jul <- read.csv("R:/Rutuja/Projects/R Project/Data/202207-divvy-tripdata.csv")
aug <- read.csv("R:/Rutuja/Projects/R Project/Data/202208-divvy-tripdata.csv")
sep <- read.csv("R:/Rutuja/Projects/R Project/Data/202209-divvy-
publictripdata.csv")
oct <- read.csv("R:/Rutuja/Projects/R Project/Data/202210-divvy-tripdata.csv")
nov <- read.csv("R:/Rutuja/Projects/R Project/Data/202211-divvy-tripdata.csv")

#Merge all of the data frames into one year view
cyclistic_df <- rbind (dec, jan, feb, mar, apr, may, jun, jul, aug, sep, oct,
nov)

#Remove individual month data frames to clear up space in the environment
remove(dec, jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov)

#Firstly remove all the irrelevant columns that won't be used for analysis
```

```r
cyclistic_df <- cyclistic_df %>%
  select(-c(start_lat, start_lng, end_lat, end_lng,
start_station_id,end_station_id, end_station_name))

#Review of the data and its parameters.
colnames(cyclistic_df)  #List of column names
nrow(cyclistic_df)  #Number of rows are in data frame
dim(cyclistic_df)  #Dimensions of the data frame
head(cyclistic_df, 6)  #See the first 6 rows of data frame.
str(cyclistic_df)  #See list of columns and data types
summary(cyclistic_df) #Inspect the date and its dimensions before moving onto
cleaning


#Additional columns must be created for date and time.
#The default format is yyyy-mm-dd
cyclistic_df$date <- as.Date(cyclistic_df$started_at)
cyclistic_df$month <- format(as.Date(cyclistic_df$date), "%m")
cyclistic_df$day <- format(as.Date(cyclistic_df$date), "%d")
cyclistic_df$year <- format(as.Date(cyclistic_df$date), "%Y")
cyclistic_df$day_of_week <- format(as.Date(cyclistic_df$date), "%A")
cyclistic_df$time <- format(cyclistic_df$started_at, format= "%H:%M")
cyclistic_df$time <- as.POSIXct(cyclistic_df$time, format= "%H:%M")

#create calculated field to isolate time spent on every ride.
cyclistic_df$ride_length <- (as.double(difftime(cyclistic_df$ended_at,
cyclistic_df$started_at))) / 60

#calculate ride length by subtracting ended_at time from started_at time and
converted it to minutes
cyclistic_date$ride_length <- difftime(cyclistic_df$ended_at,
cyclistic_df$started_at, units = "mins")

#Alter data type for time
cyclistic_df$ride_length <- as.numeric(as.character(cyclistic_df$ride_length))
#change datatype to numeric for further analysis

#Remove all blank entries from the dataset
cyclistic_df<- cyclistic_df[!(cyclistic_df$start_station_name == "HQ QR" |
cyclistic_df$ride_length<0),]

#Summarize the data before proceeding to analyze
summary(cyclistic_df$ride_length)
```

```
> summary(cyclistic_df$ride_length)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
    0.00     5.83    10.30    19.42    18.50 41387.25 
```

Fig: Summary of data for column 'ride_length'

2. Data Analysis

```
#Analyze data
#Calculating the mean, median, max, min - figures to determine statistical spread of membership type
aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = mean)
aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = median)
aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = max)
aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = min)

#Order day's of week within new dataset for future use
cyclistic_df$day_of_week <- ordered(cyclistic_df$day_of_week,
levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

#Create a weekday field as well as view column specifics
cyclistic_df %>%
  mutate(day_of_week = wday(started_at)) %>%  #creates weekday field using wday()
  group_by(member_casual, day_of_week ) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n())
```

```
> summary(cyclistic_df$ride_length)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     5.83    10.30    19.42    18.50 41387.25
> aggregate(trips20fill$ride_length ~ trips20fill$member_casual, FUN = mean)
Error in eval(predvars, data, env) : object 'trips20fill' not found
> aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = mean)
  cyclistic_df$member_casual cyclistic_df$ride_length
1                     casual                 29.10894
2                     member                 12.70896
> aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = median)
  cyclistic_df$member_casual cyclistic_df$ride_length
1                     casual                13.050000
2                     member                 8.833333
> aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = max)
  cyclistic_df$member_casual cyclistic_df$ride_length
1                     casual                 41387.25
2                     member                  1559.90
> aggregate(cyclistic_df$ride_length ~ cyclistic_df$member_casual, FUN = min)
  cyclistic_df$member_casual cyclistic_df$ride_length
1                     casual                        0
2                     member                        0
```

Data Visualization

a. Rides per day of the week.

```
#Total rides broken down by weekday
cyclistic_df$day_of_week  <- format(as.Date(cyclistic_df$date), "%A")
cyclistic_df %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, day_of_week) %>%
```

```
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
geom_col(position = "dodge") +
  labs(x='Day of Week', y='Total Rides', title='Rides per Weekday', fill =
'Membership type') +
  scale_y_continuous(breaks = c(250000, 400000, 550000), labels = c("250K",
"400K", "550K"))
```

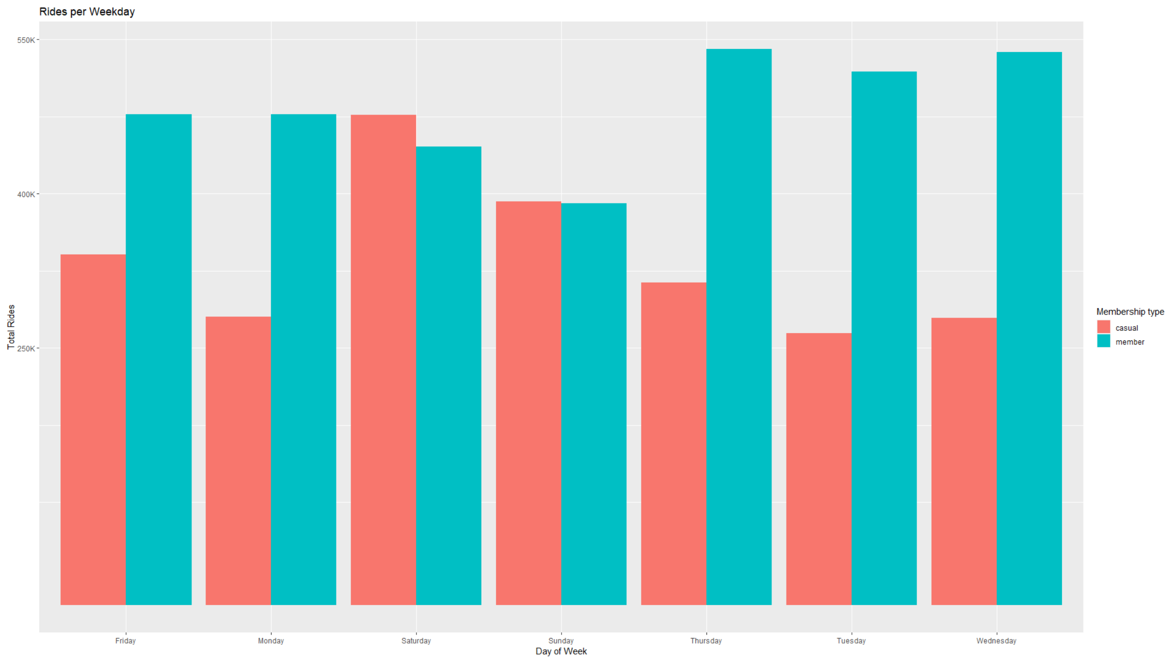

Fig: Plot for Rides per day of the week.

b. <u>Rides per Month</u>

```
#Total rides broken down by month
cyclistic_df %>%
  group_by(member_casual, month) %>%
  summarise(total_rides = n(),`average_duration_(mins)` = mean(ride_length))
%>%
  arrange(member_casual) %>%
  ggplot(aes(x=month, y=total_rides, fill = member_casual)) +
geom_col(position = "dodge") +
  labs(x= "Month", y= "Total Rides", title = "Rides per Month", fill =
"Membership type") +
  scale_y_continuous(breaks = c(100000, 200000, 300000, 400000), labels =
c("100K", "200K", "300K", "400K")) + theme(axis.text.x = element_text(angle =
45))
```

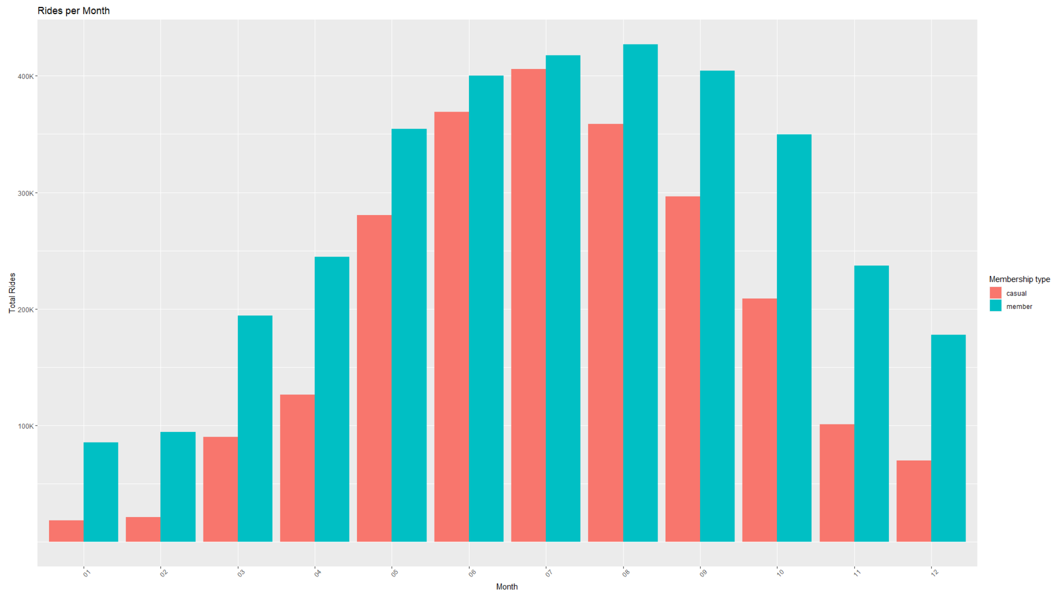

Fig: Plot for rides per month

c. <u>Most used bike types</u>

```
#Looking at breakdown of bike types rented
cyclistic_df %>%
  ggplot(aes(x = rideable_type, fill = member_casual)) + geom_bar(position =
"dodge") +
  labs(x= 'Type of Bike', y='Number of Rentals', title='Most used bike type',
fill = 'Membership type') +
  scale_y_continuous(breaks = c(500000, 1000000, 1500000), labels = c("500K",
"1Mil", "1.5Mil"))
```

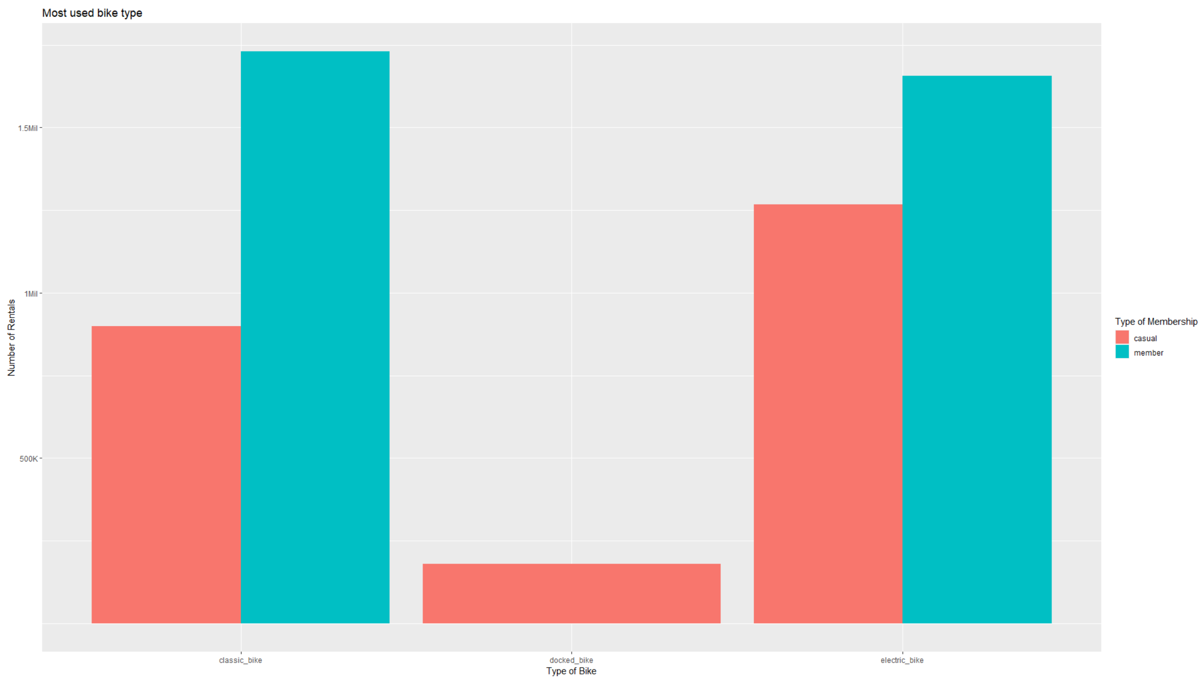

Fig: Plot for most used bike type

d. Average ride time per week

```
#Find the average time spent riding by each membership type per individual day
cyclistic_df %>%
  mutate(day_of_week = wday(started_at)) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + labs(x='Days of the week', y='Average
duration - Hrs', title='Average ride time per week', fill='Membership type')
```

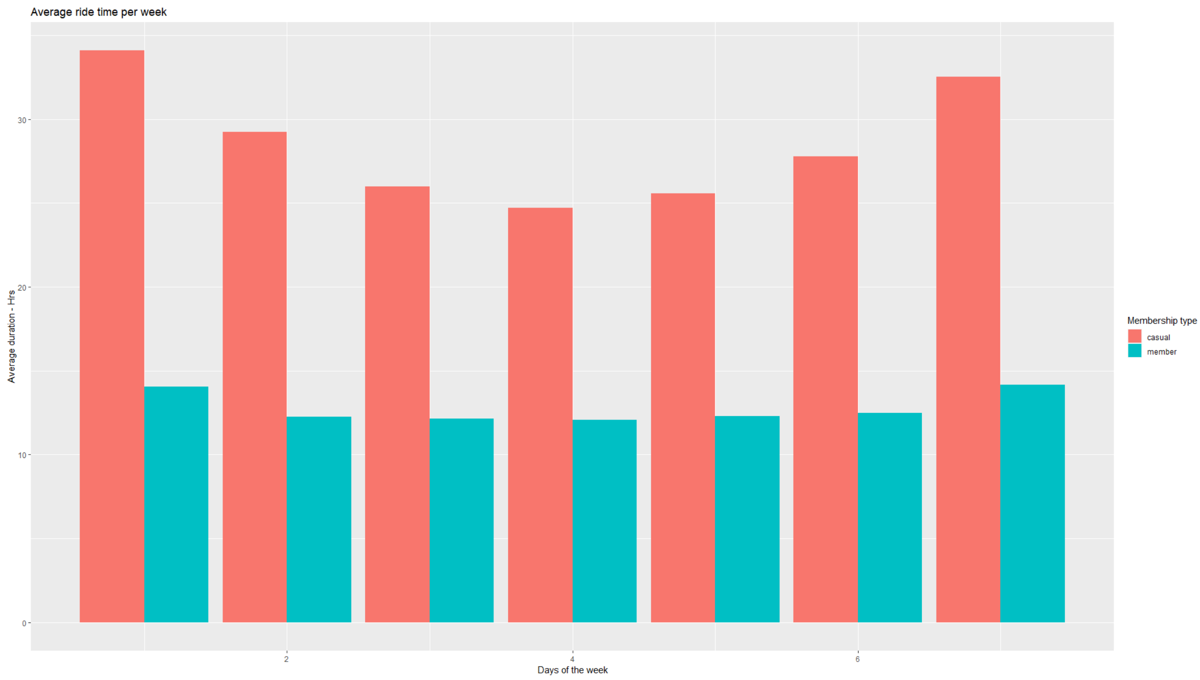

Fig: Plot showing average ride time per week

Results and Interpretations

- Casual users tended to ride more so in the warmer months of Chicago, namely June- August. Their participation exceeded that of the long-term members.
- To further that the Casual demographic spent on average a lot longer time per ride than their long-term counter-parts.
- The days of the week also further shows that causal riders prefer to use the service during the weekends as their usage peaked then. The long-term members conversely utilised the service more-so throughout the typical work week i.e. (Monday- Friday)
- Long term riders tended to stick more so to classic bikes as opposed to the docked or electric bikes.

Conclusion

- Introducing plans that may be more appealing to casuals for the summer months.
- The casual users might be more interested in a membership option that allows for per-use balance card. Alternatively, the existing payment structure may be altered in order to make single-use more costly to the casual riders as well as lowering the long-term membership rate.
- Membership rates specifically for the warmer months as well as for those who only ride on the weekends would assist in targeting the casual riders more specifically