Cyclistic Bikeshare

Zee Setash

07-22-2023

Introduction

Cyclistic is a (fictional) bikeshare company based in Chicago. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.The marketing director has set the goal of converting existing casual users into annual members. The data for this analysis is provided here. My analysis is focusing on the time frame of July 2022 - June 2023. In order to guide business decisions, I am focusing on a singular question:

How do annual **members** and **casual** riders use Cyclistic bikes differently?

Preparing Data

Limitations

The current data records rides, not riders. I do not have information to compare how many times a specific rider uses the service. I also do not have any information about the cost of the rides. The best way to convince people to sign up for a membership is to convince them that it is the more cost effective choice. I do not currently know how the cost for the customes varies across the different passes.

Cleaning Data

Prior to importing the data into R, I made some cleaning adjustments. I verified that the column names were consistent across files. I added the column *ride_length* using the difference between the columns *ended_at* and *started_at* in order to later learn more about the differences in ride lengths between users. I also added a *day_of_week* column using the WEEKDAY function in order to later compare usage across days of the week.

Importing Packages and Data

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(lubridate)
library(hms)
library(chron)
```

```
july <- read_csv("C:/Users/Z/Desktop/Bikeshare/202207-divvy-tripdata.csv")
august <- read_csv("C:/Users/Z/Desktop/Bikeshare/202208-divvy-tripdata.csv")
september <- read_csv("C:/Users/Z/Desktop/Bikeshare/202209-divvy-tripdata.csv")
october <- read_csv("C:/Users/Z/Desktop/Bikeshare/202210-divvy-tripdata.csv")
november <- read_csv("C:/Users/Z/Desktop/Bikeshare/202211-divvy-tripdata.csv")
december <- read_csv("C:/Users/Z/Desktop/Bikeshare/202212-divvy-tripdata.csv")
january <- read_csv("C:/Users/Z/Desktop/Bikeshare/202301-divvy-tripdata.csv")
february <- read_csv("C:/Users/Z/Desktop/Bikeshare/202302-divvy-tripdata.csv")
march <- read_csv("C:/Users/Z/Desktop/Bikeshare/202303-divvy-tripdata.csv")
april <- read_csv("C:/Users/Z/Desktop/Bikeshare/202304-divvy-tripdata.csv")
may <- read_csv("C:/Users/Z/Desktop/Bikeshare/202305-divvy-tripdata.csv")
june <- read_csv("C:/Users/Z/Desktop/Bikeshare/202306-divvy-tripdata.csv")
```

I then combined the data into a single dataframe, to better see overall trends.

```
year <- rbind(january, february, march, april, may, june, july, august, september, october, november, d
```

I then removed any duplicate rows based on the ride ID.

```
year_no_dup <- year %>%
  distinct(ride_id, .keep_all = TRUE)
print(paste("Removed", nrow(year) - nrow(year_no_dup), "duplicated rows"))
```

```
## [1] "Removed 0 duplicated rows"
```

I then removed any rows that contained NA values.

```
year_clean <- drop_na(year_no_dup)
print(paste("Removed", nrow(year_no_dup) - nrow(year_clean), "rows with NA values"))
```

```
## [1] "Removed 1370355 rows with NA values"
```

Data Transformation

I used existing values to create several columns to streamline my workflow. First, I created a column that displayed the length of the bike rides in minutes by finding the difference between the starting and ending times.

```
year_clean <- year_clean %>%
  mutate(ride_time_minutes = as.numeric(year_clean$ended_at - year_clean$started_at) / 60)
print(summary(year_clean$ride_time_minutes))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  -168.70      5.72     10.00     16.05     17.85  32035.45
```

Through the summary, I discovered that the data set included impossible ride times. There were instances where the ride ended before it started. There were also times that were too short to be actual rides. I removed these values to focus the analysis on more authentic ride times.

```
year_clean_times <- year_clean %>%
  filter(ride_time_minutes > 2)
print(paste("Removed", nrow(year_clean) - nrow(year_clean_times), "rows with impossible or inauthentic v
```

```
## [1] "Removed 167665 rows with impossible or inauthentic values"
```

I then added columns that documented the day of the week of the ride start times. I also added a column that gave the start times a "time of day", splitting the times into morning, afternoon, evening, and night.

```
year_clean_times <- year_clean_times %>%
  mutate(day_of_week = wday(started_at, label = TRUE))

breaks <- hour(hm("00:00", "4:00", "12:00", "19:00", "23:00"))
labels <- c("Night", "Morning", "Afternoon", "Evening")

year_clean_times$time_of_day <- cut(x=hour(year_clean_times$started_at), breaks = breaks, labels = label
```

I then removed any times great than one day and created a column that group the lengths of the rides into more meaningful groups.

```
year_clean_times2 <- year_clean_times %>%
  filter(ride_time_minutes <= 1440)

print(paste("Removed", nrow(year_clean_times) - nrow(year_clean_times2), "rows with ride duration greate
```

```
## [1] "Removed 105 rows with ride duration greater than a day"
```

```
breaks2 <- c(2, 5, 10, 20, 60, 360, 720, 1440)
labels2 <- c("Less than 5 minutes","Between 5 and 10 minutes", "Between 10 and 20 minutes", "Between 20

year_clean_times2$ride_length <- cut(x=(year_clean_times2$ride_time_minutes), breaks = breaks2, labels =
```

While not truly necessary, I split the data into two separate data sets, one focuing on the members and the other focusing on the casual riders.

```
year_member <- year_clean_times2 %>%
  filter(member_casual == 'member')

year_casual <- year_clean_times2 %>%
  filter(member_casual == 'casual')
```

```
print("Members")
```

```
## [1] "Members"
```

```
glimpse(year_member)
```

```
## Rows: 2,613,888
## Columns: 17
```

```
## $ ride_id            <chr> "F96D5A74A3E41399", "13CB7EB698CEDB88", "C90792D034~
## $ rideable_type      <chr> "electric_bike", "classic_bike", "classic_bike", "c~
## $ started_at         <dttm> 2023-01-21 20:05:42, 2023-01-10 15:37:36, 2023-01-~
## $ ended_at           <dttm> 2023-01-21 20:16:33, 2023-01-10 15:46:05, 2023-01-~
## $ start_station_name <chr> "Lincoln Ave & Fullerton Ave", "Kimbark Ave & 53rd ~
## $ start_station_id   <chr> "TA1309000058", "TA1309000037", "TA1309000037", "TA~
## $ end_station_name   <chr> "Hampden Ct & Diversey Ave", "Greenwood Ave & 47th ~
## $ end_station_id     <chr> "202480.0", "TA1308000002", "TA1308000002", "TA1308~
## $ start_lat          <dbl> 41.92407, 41.79957, 41.79957, 41.79957, 41.92607, 4~
## $ start_lng          <dbl> -87.64628, -87.59475, -87.59475, -87.59475, -87.638~
## $ end_lat            <dbl> 41.93000, 41.80983, 41.80983, 41.80983, 41.93000, 4~
## $ end_lng            <dbl> -87.64000, -87.59938, -87.59938, -87.59938, -87.640~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
## $ ride_time_minutes  <dbl> 10.850000, 8.483333, 8.766667, 15.316667, 3.216667,~
## $ day_of_week        <ord> Sat, Tue, Sun, Thu, Tue, Sun, Wed, Wed, Fri, Thu, T~
## $ time_of_day        <fct> Evening, Afternoon, Morning, Afternoon, Morning, Ev~
## $ ride_length        <ord> Between 10 and 20 minutes, Between 5 and 10 minutes~
```

```
print("Casual Riders")
```

```
## [1] "Casual Riders"
```

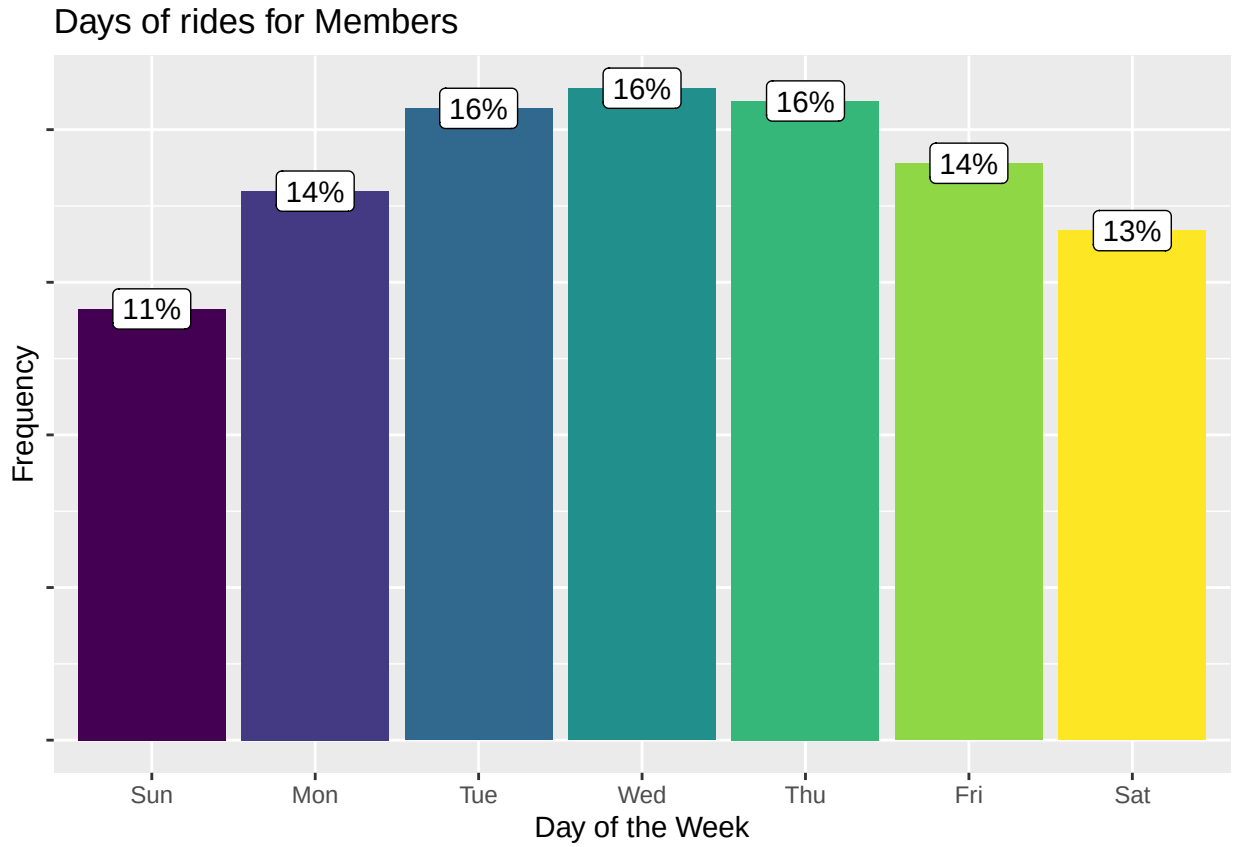
```
glimpse(year_casual)
```

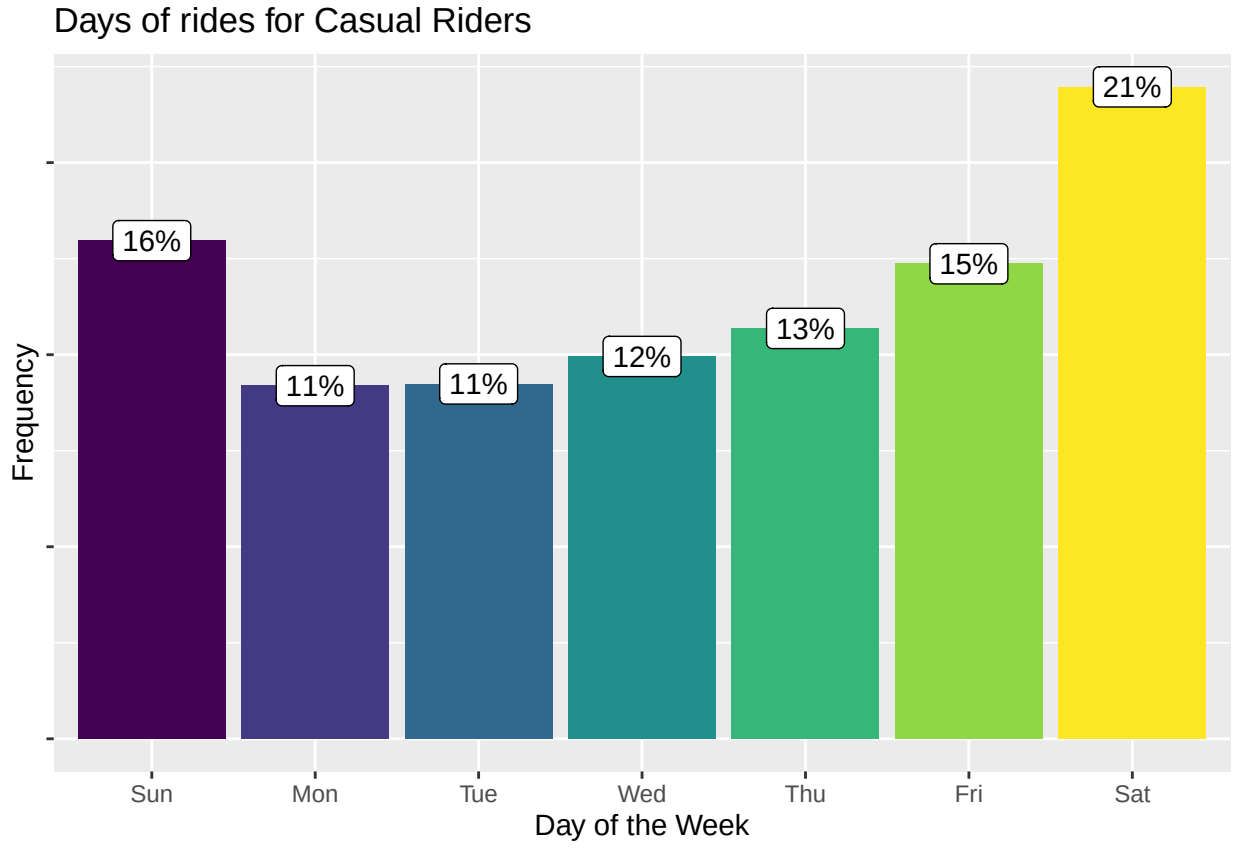
```
## Rows: 1,627,431
## Columns: 17
## $ ride_id            <chr> "BD88A2E670661CE5", "9DC70E5EE9D6A93F", "689D537E5D~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2023-01-02 07:51:57, 2023-01-03 20:25:53, 2023-01-~
## $ ended_at           <dttm> 2023-01-02 08:05:11, 2023-01-03 20:35:50, 2023-01-~
## $ start_station_name <chr> "Western Ave & Lunt Ave", "Broadway & Waveland Ave"~
## $ start_station_id   <chr> "RP-005", "13325", "RP-005", "KA1504000146", "15624~
## $ end_station_name   <chr> "Valli Produce - Evanston Plaza", "Hampden Ct & Div~
## $ end_station_id     <chr> "599", "202480.0", "599", "KA1504000148", "439", "K~
## $ start_lat          <dbl> 42.00857, 41.94911, 42.00861, 41.97803, 41.95318, 4~
## $ start_lng          <dbl> -87.69048, -87.64863, -87.69052, -87.66856, -87.731~
## $ end_lat            <dbl> 42.03974, 41.93000, 42.03974, 41.99086, 41.95000, 4~
## $ end_lng            <dbl> -87.69941, -87.64000, -87.69941, -87.66972, -87.690~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
## $ ride_time_minutes  <dbl> 13.233333, 9.950000, 15.266667, 7.300000, 10.150000~
## $ day_of_week        <ord> Mon, Tue, Thu, Mon, Wed, Tue, Fri, Sat, Fri, Sun, T~
## $ time_of_day        <fct> Morning, Evening, Afternoon, Morning, Afternoon, Af~
## $ ride_length        <ord> Between 10 and 20 minutes, Between 5 and 10 minutes~
```

Analyzing the Data

Throughout the course of my analysis, I focused on percentages of riders. My analyis included a million more rides from members than casual riders. Looking at raw numbers and comparing the values using the same y-axis for members and casual riders could be misleading. Since my goal was to determine the differences betweeen the types of riders, I honed in on the proportion of riders within a group, rather then the number of individual riders.

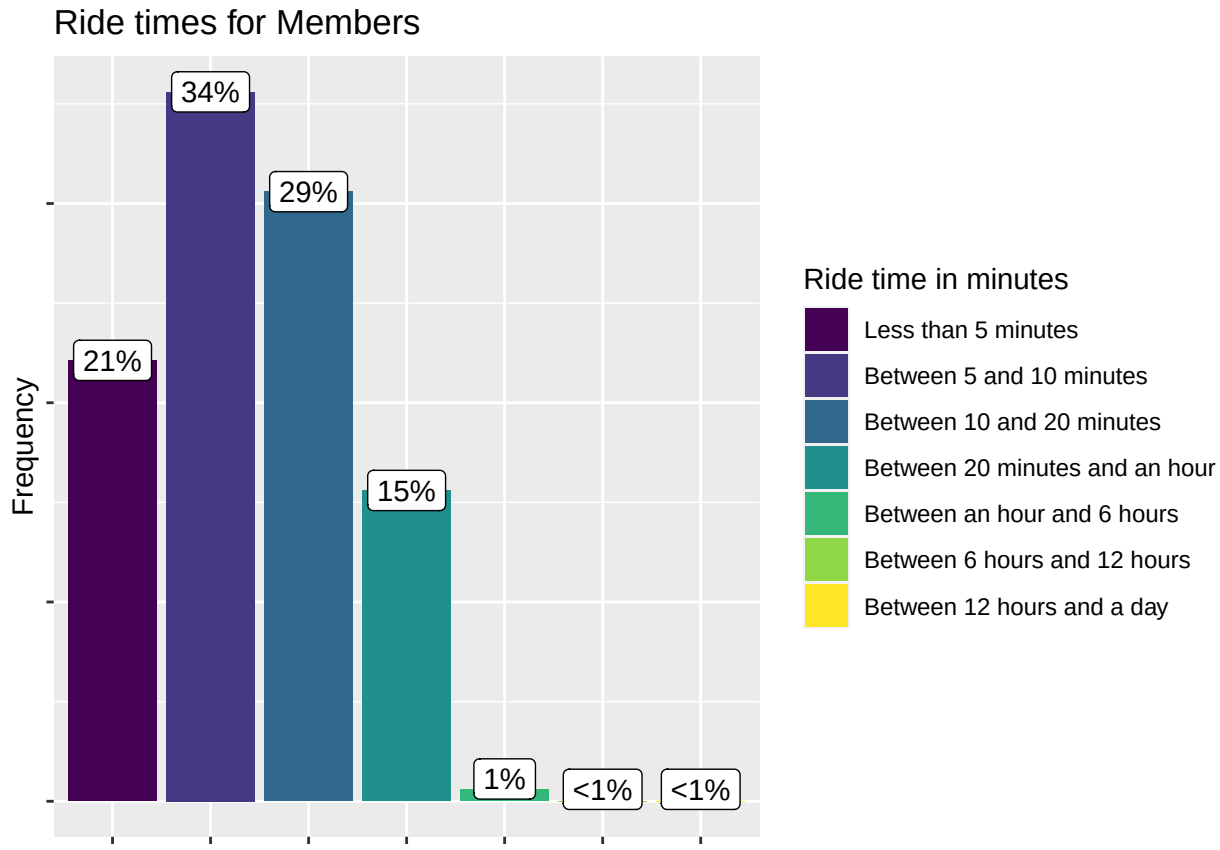
Day of the Week

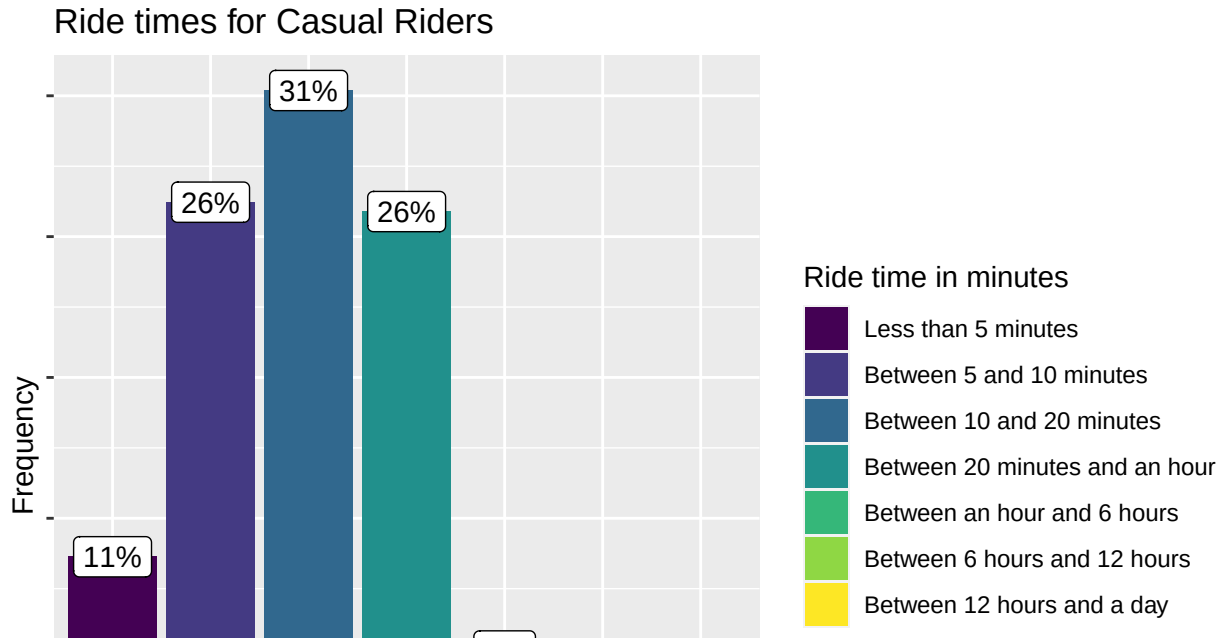




The most popular days of the week for members are Tuesday, Wedneday, and Thursday, making up a total of 48% of their rides. The most popular days of the week for causual riders are Friday, Saturday, and Sunday, making up a total of 52% of their rides. These differences indicate that members use the bikes for utility purposes while casual riders use the bikes for recreational purposes.

Duration of Rides

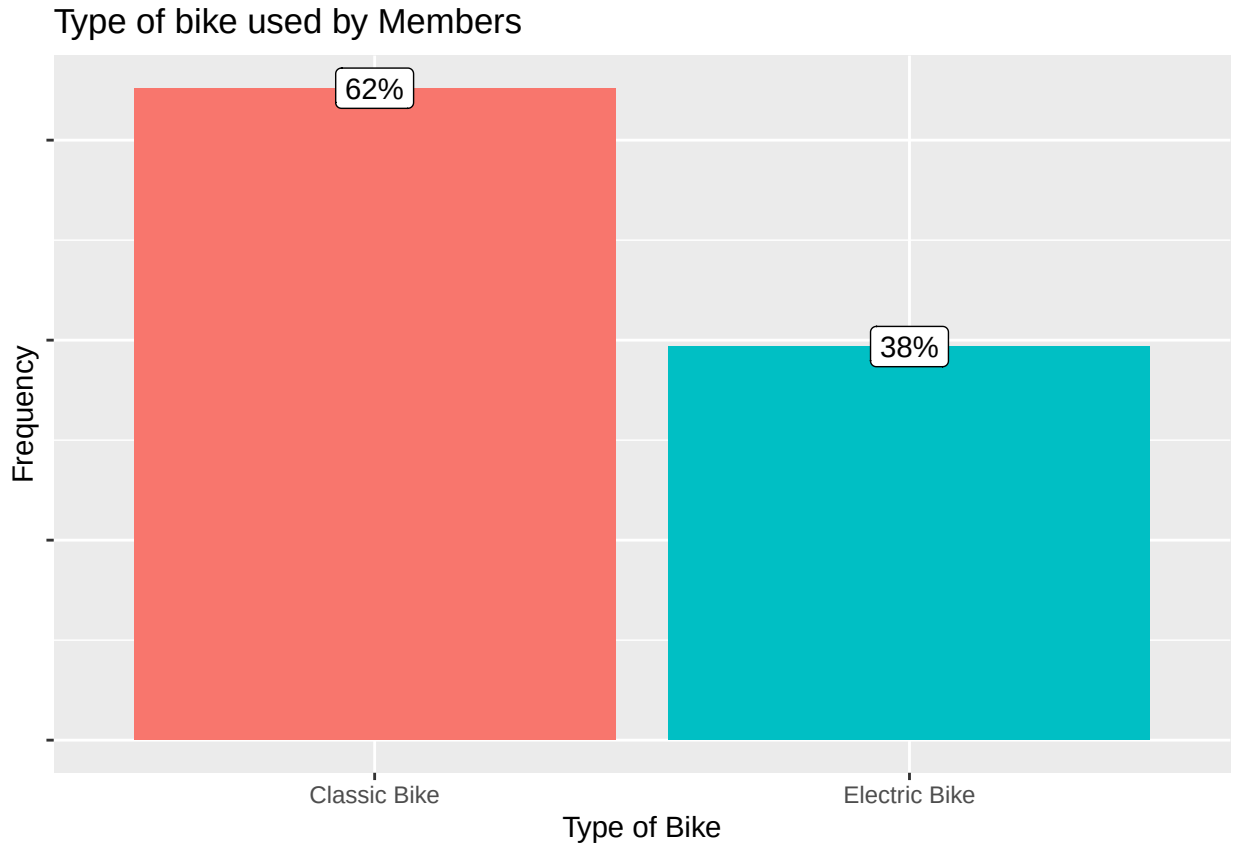


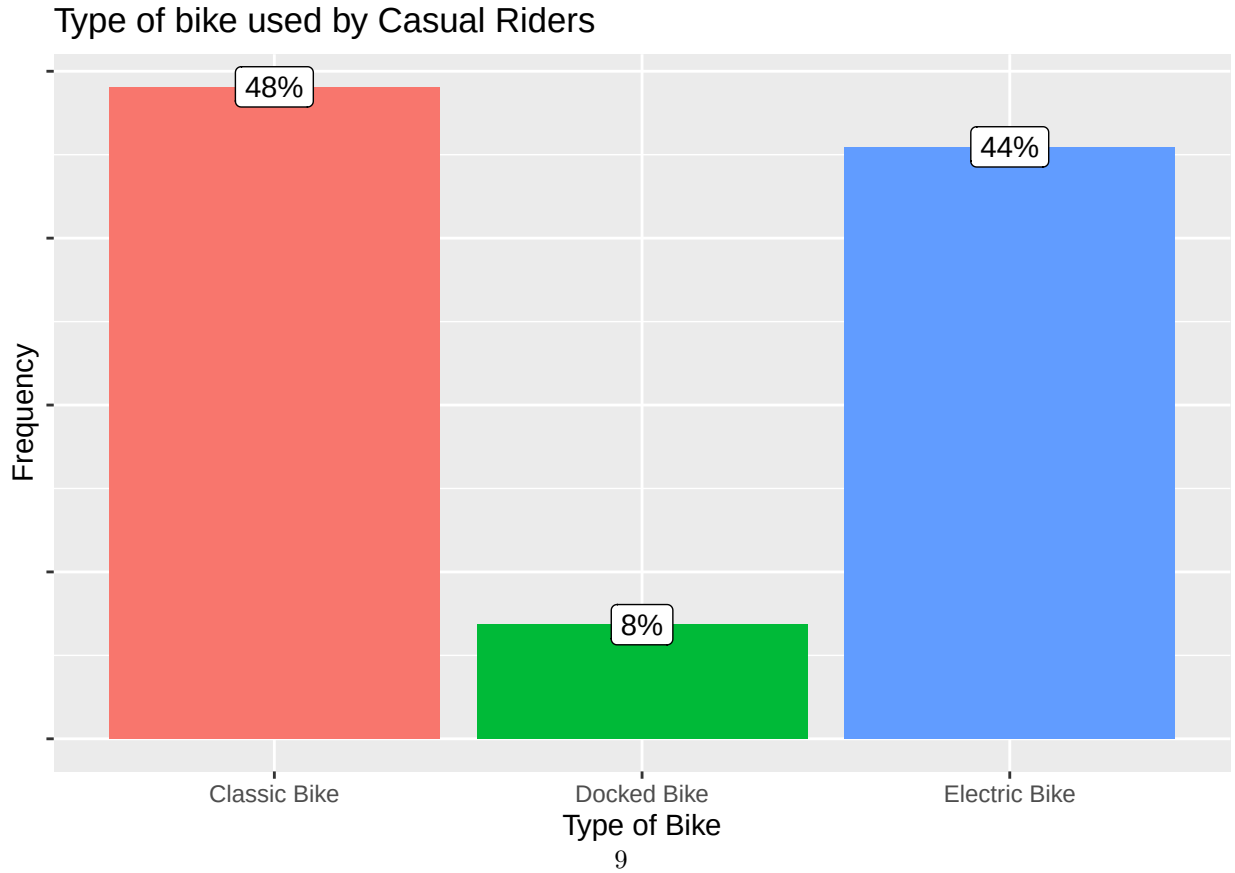


Members tend to take shorter rides than casual riders. 84% of rides by members were shorter than 20 minutes, whereas only 68% of rides by casual riders were under 20 minutes. Only 1% of rides by members took over an hour, whereas 7% of rides by casual riders took over an hour.

The most frequent ride duration for members was between 5 and 10 minutes while the most frequent duration for casual riders was between 10 and 20 minutes.

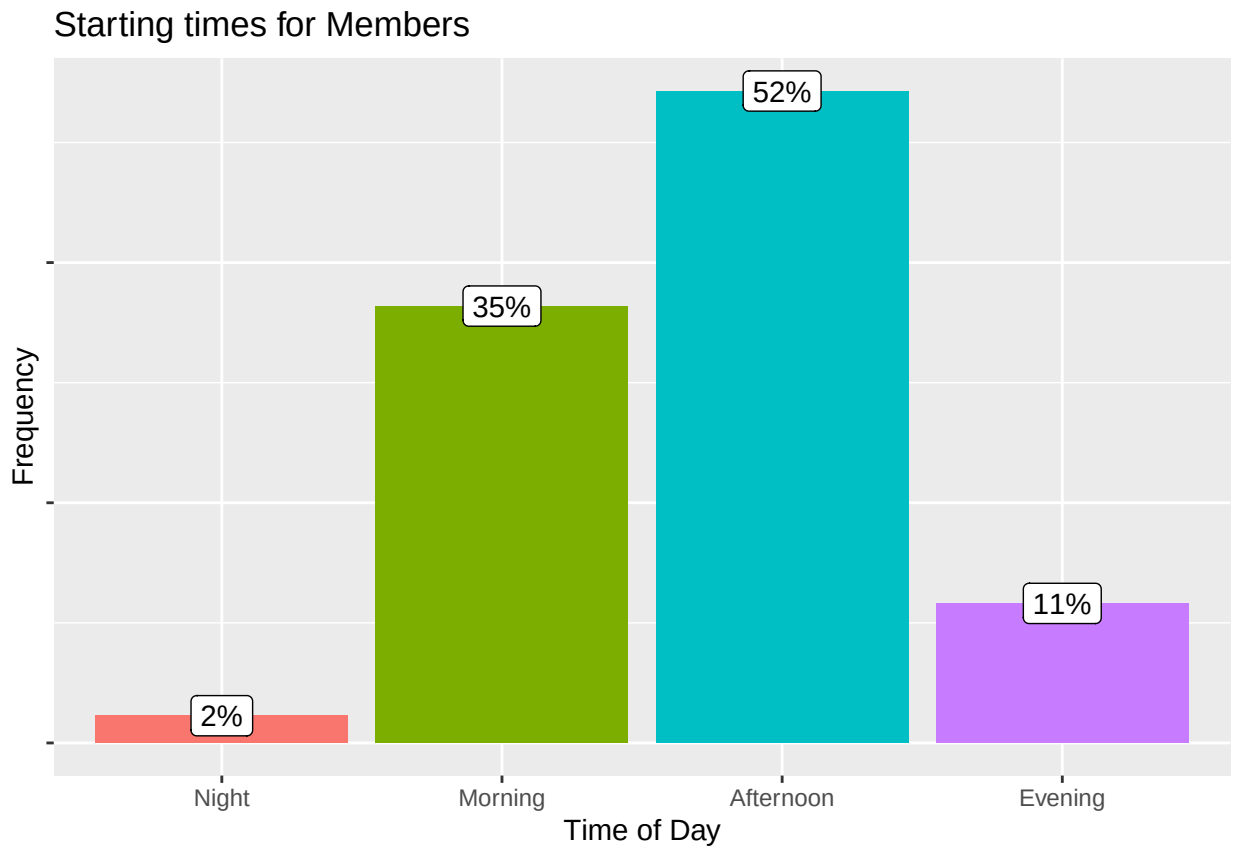
Bike Type

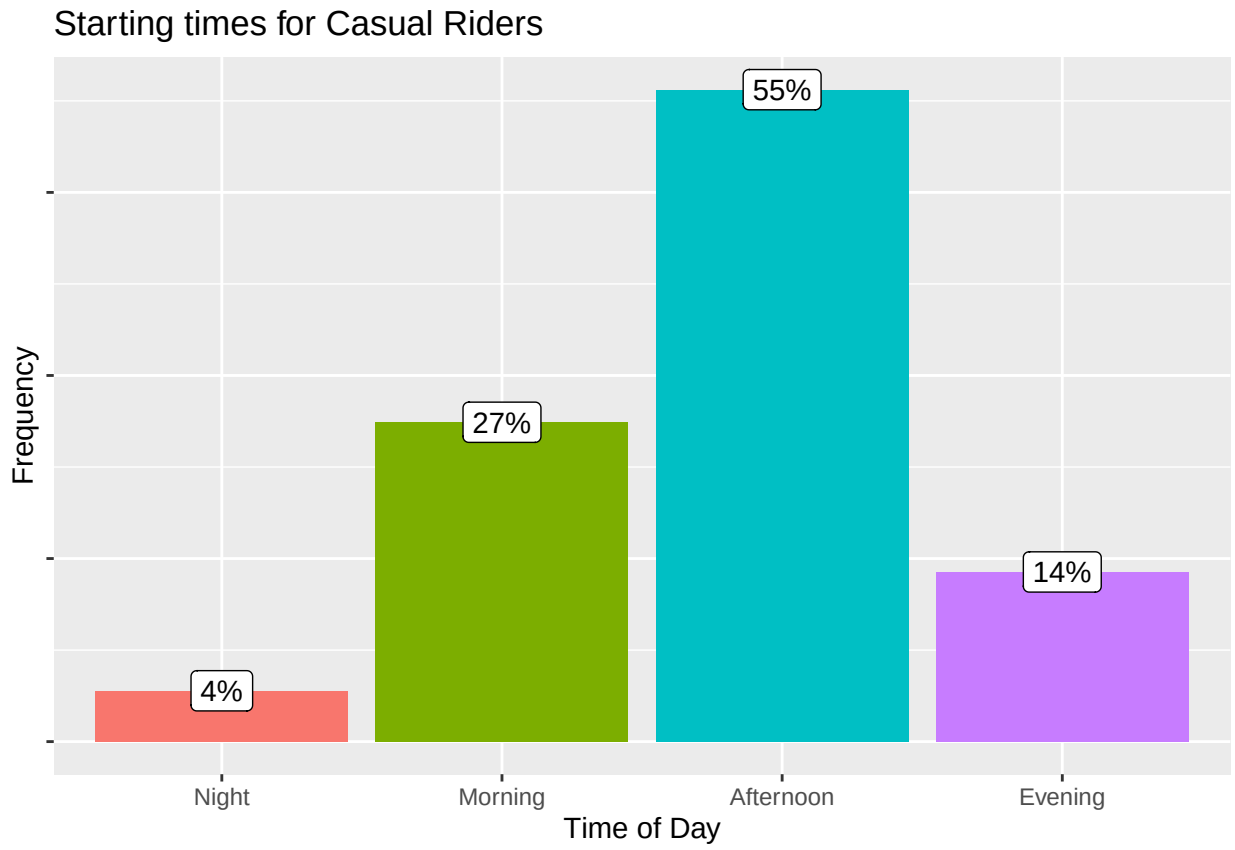

Type of bike used by Members
62%
38%
Frequency
Classic Bike
Electric Bike
Type of Bike


Type of bike used by Casual Riders
48%
8%
44%
Frequency
Classic Bike
Docked Bike
Electric Bike
Type of Bike

It is unclear whether the "docked" bikes are classic, electric, or a mix. Regardless, a greater proportion of casual riders select electric bikes than members do. This supports the idea that casual riders primarily ride for recreaction and more tourist activities.

Ride Start Time





Riders from both groups follow the same trends in riding start times. Over half of the rides in both groups started in the afternoon.

Conclusions

The current goal in the company in too encourage casual riders to transition into Cyclistic Members. Tourists are not likely to become members, as they will not get enough use out of the membership. Locals need to feel it is more cost effective to be a member rather than paying per ride or per day.

Recommendations

I recommend implementing a pricing structure that has prices that vary based on the day of the week and time of day. The cost for members would be uneffected, and proper implementation will make membership more desirable to repeat casual riders.

1. Weekends are popular ride times for casual riders, therefore I recommend making single ride or day passes more expensive on Fridays, Saturdays, and Sundays.
2. The afternoon is the most popular time to start a ride, therefore I recommend increasing the price for single rides if they start in the afternoon.
3. Casual riders are more likely to take longer rides, therefore I recommend making single rides more expensive if they last greater than an hour.

I recognize that recommendation #3 may already be addressed by the existence of day passes, so I have an additional recommendation if necessary.

4. Electric bikes are more commonly selected by casual riders, therefore I recommend having a number of electric bikes reserved for members. This may encourage casual riders to become members in order to gain access to electric bikes.

Thank You

Thank you for taking the time to read my report. Thank you to the Google Data Analytics course for providing me with the tools to accomplish this.