Report_Case_Study1

Ulises Jimenez

8/27/2022

Introduction

The following report is my interpretation of the Google Data Analytics Capstone - Case Study 1. The case study involves a bike-sharing company called Cyclistic, and their customer's trip details over a 12 month period (from August 2021 to July 2022).The goal is to solve key business questions using the data analysis process: ask, prepare, process, analyze, share, and act.

Scenario

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.

Characters and Teams

- **Cyclistic**: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistics sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.
- **Lilly Moreno**: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
- **Cyclistic marketing analytics team**: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic mission and business goals - as well as how you, as a junior data analyst, can help Cyclistic achieve them.
- **Cyclistic executive team**: The notoriously detail-oriented team will decide whether to approve the recommended marketing program.

Objective

The objective of this case study is to find the differences between casual and annual members. The insights discovered will be used to promote annual membership among casual riders.

Load Libraries

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(scales)
library(data.table)
```

Load Data

All 12 csv files will be concatenated into one date frame.

```
bike_data <- list.files(pattern='tripdata',full.names=TRUE,recursive=TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
```

Inspect the data frame.

```
glimpse(bike_data)
```

```
## Rows: 5,901,463
## Columns: 13
## $ ride_id            <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-08-10 17:15:49, 2021-08-10 17:23:14, 2021-08-~
## $ ended_at           <dttm> 2021-08-10 17:22:44, 2021-08-10 17:39:24, 2021-08-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, "Clark St & Grace St", ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, "TA1307000127", NA, NA,~
## $ start_lat          <dbl> 41.77000, 41.77000, 41.95000, 41.97000, 41.79000, 4~
## $ start_lng          <dbl> -87.68000, -87.68000, -87.65000, -87.67000, -87.600~
## $ end_lat            <dbl> 41.77000, 41.77000, 41.97000, 41.95000, 41.77000, 4~
## $ end_lng            <dbl> -87.68000, -87.63000, -87.66000, -87.65000, -87.620~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

Clean Data

Unnecessary data will be removed from the data frame such as duplicate data, test data, and data that will not be used for the analysis.

Remove Data

Removed latitude and longitude data since it will not be part of the analysis.

```
bike_data <- bike_data %>%
  select(-c(start_lat:end_lng))

colnames(bike_data)
```

```
## [1] "ride_id"            "rideable_type"      "started_at"
## [4] "ended_at"           "start_station_name" "start_station_id"
## [7] "end_station_name"   "end_station_id"     "member_casual"
```

Remove Duplicate Data

Remove any duplicate ride id since the id is unique for every ride.

```
bike_data <- bike_data[!duplicated(bike_data$ride_id), ]

nrow(bike_data)
```

```
## [1] 5901463
```

No duplicate data was found.

Remove Test Data

The Cyclistic company conducts tests on random bikes and labels the start station name as "Test". Test data will be removed before analysis because test rides are conducted by the company and not actual riders.

```
#Number of test rides
nrow(subset(bike_data, tolower(start_station_name) %like% "test"))
```

```
## [1] 1
```

```
#Removing test rides
bike_data <- bike_data[!(tolower(bike_data$start_station_name) %like% "test"), ]
nrow(bike_data)
```

```
## [1] 5901462
```

Found 1 test ride and removed it from the data frame.

Data Manipulation

New columns will be created to help with analysis and visualizations.

Ride Length

The new column "ride_length" will represent the duration of a bike ride in seconds. The "ride_length" column will be calculated by subtracting the time a rider checks out a bike from the time a rider returns the bike. The "ride_length" column will be measured in seconds to maintain accuracy.

```
bike_data <- bike_data %>%
  mutate(ride_length = as.numeric(difftime(bike_data$ended_at, bike_data$started_at, units = "secs")))

glimpse(bike_data)
```

```
## Rows: 5,901,462
## Columns: 10
## $ ride_id            <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-08-10 17:15:49, 2021-08-10 17:23:14, 2021-08-~
## $ ended_at           <dttm> 2021-08-10 17:22:44, 2021-08-10 17:39:24, 2021-08-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, "Clark St & Grace St", ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, "TA1307000127", NA, NA,~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
## $ ride_length        <dbl> 415, 970, 973, 918, 522, 395, 941, 341, 18, 1047, 3~
```

Check for any discrepancies with the new ride_length column. Since time cannot be negative, any negative value in "ride_length" will be removed.

```
#Check for negative values
nrow(bike_data[bike_data$ride_length < 0, ])
```

```
## [1] 149
```

```
#Remove negative values
bike_data <- bike_data %>%
  filter(ride_length >= 0)
glimpse(bike_data)
```

```
## Rows: 5,901,313
## Columns: 10
## $ ride_id            <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-08-10 17:15:49, 2021-08-10 17:23:14, 2021-08-~
## $ ended_at           <dttm> 2021-08-10 17:22:44, 2021-08-10 17:39:24, 2021-08-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, "Clark St & Grace St", ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, "TA1307000127", NA, NA,~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
## $ ride_length        <dbl> 415, 970, 973, 918, 522, 395, 941, 341, 18, 1047, 3~
```

149 negative time values were found and removed.

Create More Columns

Create 3 more columns to show weekday name, hour of the day, and month year.

```
bike_data_clean <- bike_data %>%
  mutate(month_yr = strftime(started_at, "%b-%Y"),
         weekday_name = weekdays(as.Date(started_at)),
         start_hr = strftime(started_at, "%H"))

glimpse(bike_data_clean)
```

```
## Rows: 5,901,313
## Columns: 13
## $ ride_id            <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-08-10 17:15:49, 2021-08-10 17:23:14, 2021-08-~
## $ ended_at           <dttm> 2021-08-10 17:22:44, 2021-08-10 17:39:24, 2021-08-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, "Clark St & Grace St", ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, "TA1307000127", NA, NA,~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
## $ ride_length        <dbl> 415, 970, 973, 918, 522, 395, 941, 341, 18, 1047, 3~
## $ month_yr           <chr> "Aug-2021", "Aug-2021", "Aug-2021", "Aug-2021", "Au~
## $ weekday_name       <chr> "Tuesday", "Tuesday", "Saturday", "Saturday", "Thur~
## $ start_hr           <chr> "10", "10", "19", "23", "04", "05", "05", "07", "11~
```

Saving Results

Saved the results in a csv file.

```
bike_data_clean %>%
  write.csv('cyclists_clean.csv')
```

Analyze

Load Clean Data

```
cyclists <- read_csv('cyclists_clean.csv', col_names = TRUE)
```

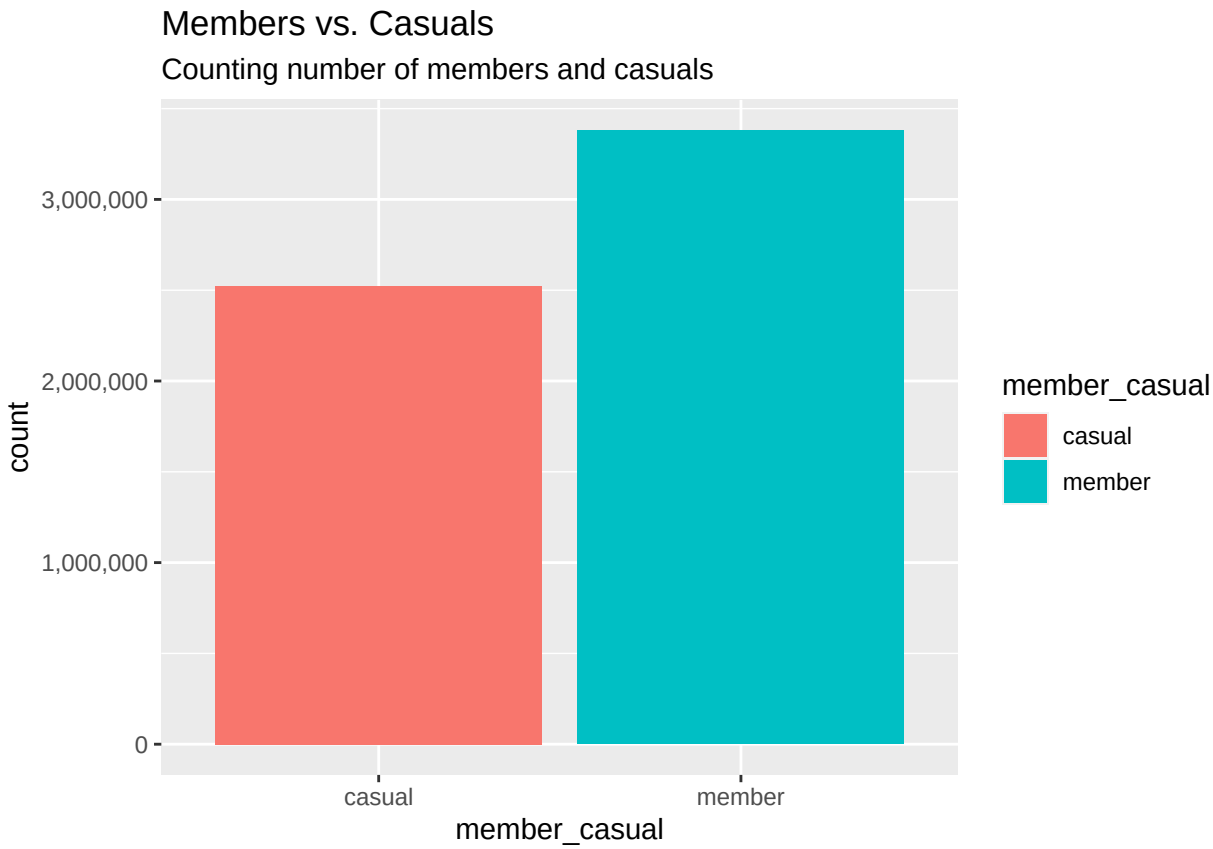
Customer Status

Number and percentage of annual members and casual riders.

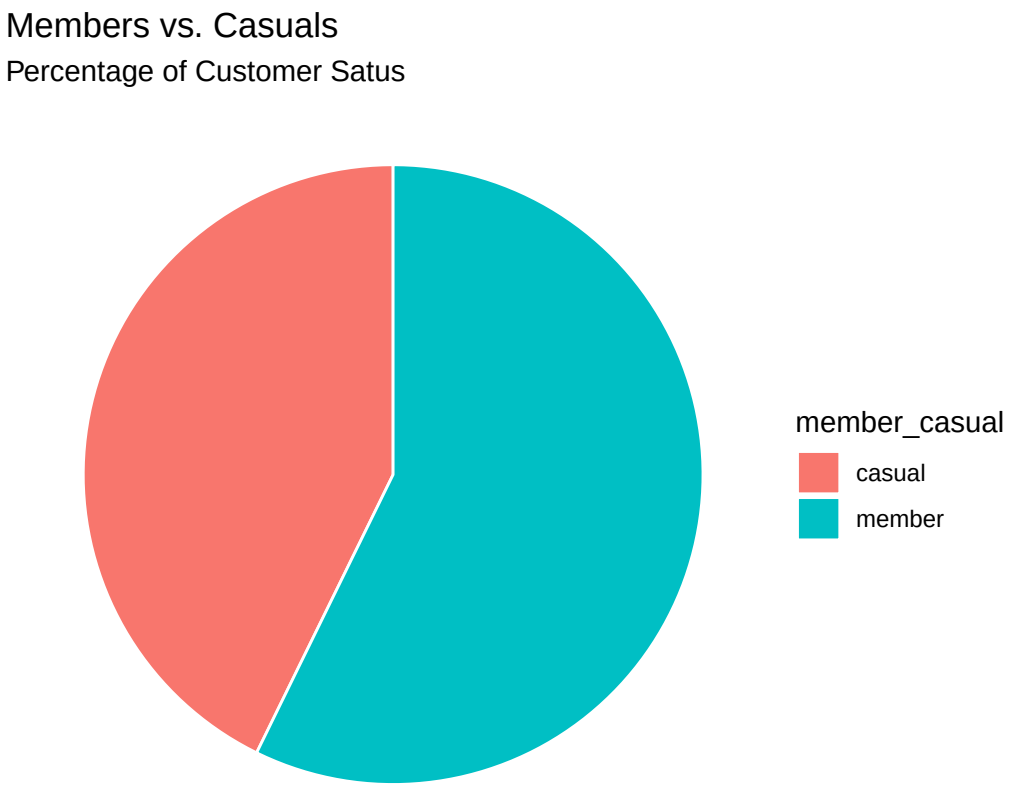
```
cyclists %>%
  group_by(member_casual) %>%
  summarise(count = n(),
            percentage = (count/nrow(cyclists))*100)
```

```
## # A tibble: 2 x 3
##   member_casual   count percentage
##   <chr>           <int>      <dbl>
## 1 casual        2522159       42.7
## 2 member        3379154       57.3
```

```
cyclists %>%
  ggplot(aes(x = member_casual, fill = member_casual))+
  geom_bar()+
  labs(title = "Members vs. Casuals", subtitle = "Counting number of members and casuals")+
  scale_y_continuous(labels = comma)
```




```
#Pie chart
df_pie <- cyclists %>%
  group_by(member_casual) %>%
  summarise(count = n())

df_pie %>%
  ggplot(aes(x = "", y = count, fill = member_casual))+
  geom_bar(stat = "identity", width = 1, color = "white")+
  coord_polar("y", start = 0)+
  labs(title = "Members vs. Casuals", subtitle = "Percentage of Customer Satus")+
  theme_void()
```



Based on the Members vs. Casuals bar chart there are more annual members than casual members. Furthermore the Members vs. Casuals pie chart shows there are roughly 15% more annual members than casual members. Thus indicating that the majority of the company's revenue comes from annual members.

Month

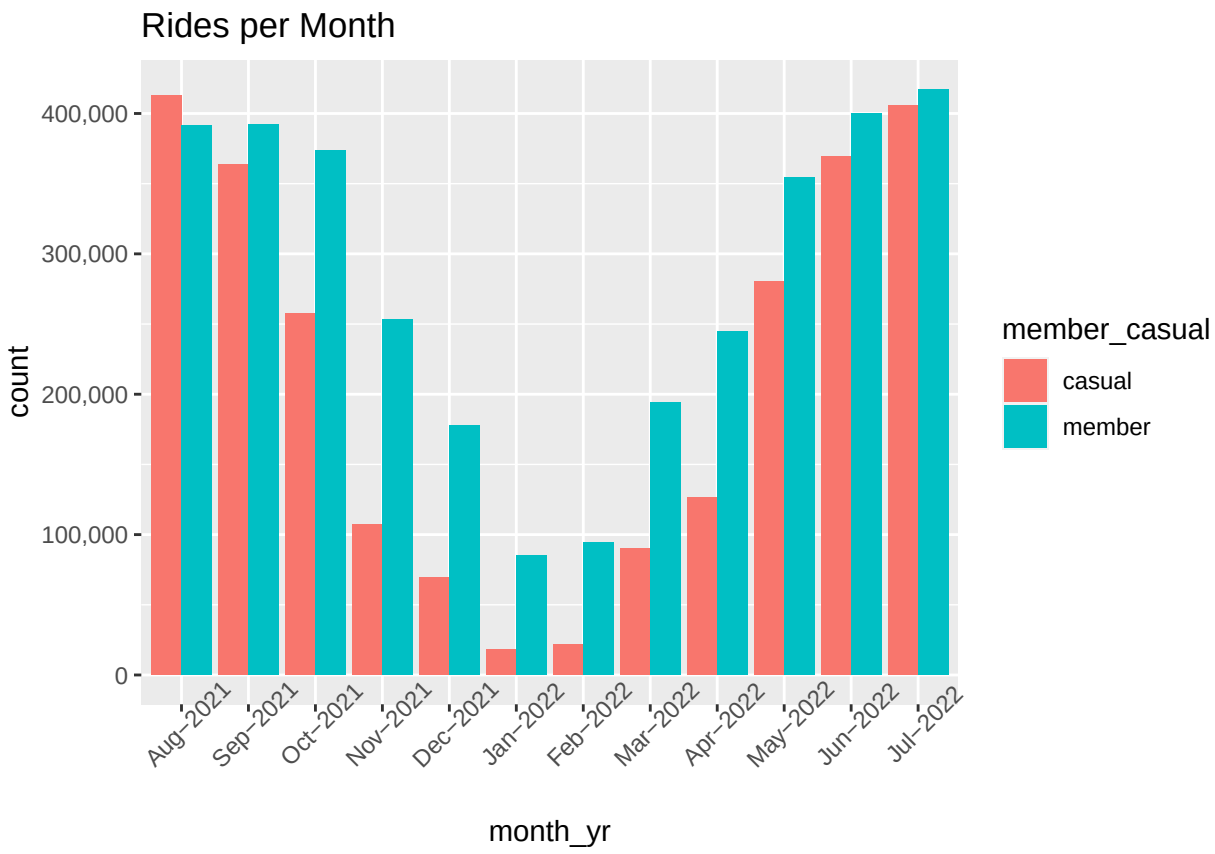
Number of rides per month for annual members and casual riders.

```
x1 <- cyclists %>%
  group_by(member_casual, month_yr) %>%
  summarise(number_of_rides = n(), avg_ride_time = mean(ride_length)) %>%
  arrange(member_casual, desc(number_of_rides))

print(tibble(x1), n = 24)
```

```
## # A tibble: 24 x 4
##    member_casual month_yr number_of_rides avg_ride_time
##    <chr>         <chr>              <int>         <dbl>
##  1 casual        Aug-2021          412662         1727.
##  2 casual        Jul-2022          406046         1757.
##  3 casual        Jun-2022          369044         1926.
##  4 casual        Sep-2021          363883         1669.
##  5 casual        May-2022          280414         1852.
##  6 casual        Oct-2021          257242         1720.
##  7 casual        Apr-2022          126417         1772.
```

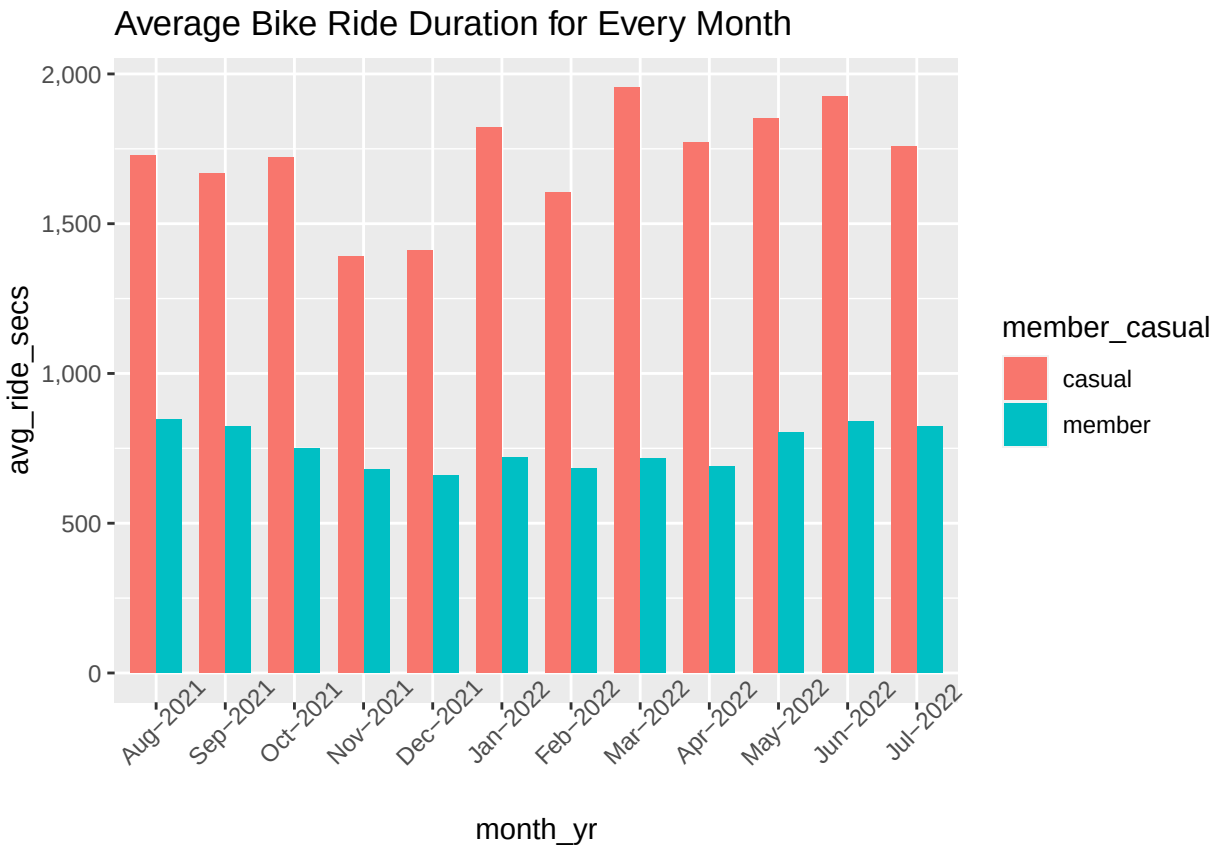
```
##  8 casual        Nov-2021          106898         1391.
##  9 casual        Mar-2022           89880         1956.
## 10 casual        Dec-2021           69738         1410.
## 11 casual        Feb-2022           21416         1603.
## 12 casual        Jan-2022           18519         1823.
## 13 member        Jul-2022          417426          823.
## 14 member        Jun-2022          400148          840.
## 15 member        Sep-2021          392228          824.
## 16 member        Aug-2021          391661          846.
## 17 member        Oct-2021          373984          750.
## 18 member        May-2022          354443          802.
## 19 member        Nov-2021          253027          679.
## 20 member        Apr-2022          244832          690.
## 21 member        Mar-2022          194160          717.
## 22 member        Dec-2021          177802          660.
## 23 member        Feb-2022           94193          684.
## 24 member        Jan-2022           85250          719.
```

```r
#Arrange month year
cyclists$month_yr <- ordered(cyclists$month_yr,
                             levels = c("Aug-2021","Sep-2021","Oct-2021","Nov-2021","Dec-2021",
                                        "Jan-2022","Feb-2022","Mar-2022","Apr-2022","May-2022",
                                        "Jun-2022","Jul-2022"))


cyclists %>%
  group_by(member_casual) %>%
  ggplot(aes(x = month_yr, fill = member_casual))+
  geom_bar(position = "dodge")+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "Rides per Month")+
  scale_y_continuous(labels = comma)
```



The chart above shows that both annual and casual members have high usage of shared-bikes during the months of May to October. While both types of members have extremely low usage during the end and beginning of the year (from November to April). Notice how annual members have more rides than casual riders in all months except for August 2021.

```
cyclists %>%
  group_by(member_casual, month_yr) %>%
  summarise(avg_ride_secs = mean(ride_length)) %>%
  ggplot(aes(x = month_yr, y = avg_ride_secs, fill = member_casual))+
  geom_col(width = 0.75, position = "dodge")+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "Average Bike Ride Duration for Every Month")+
  scale_y_continuous(labels = comma)
```

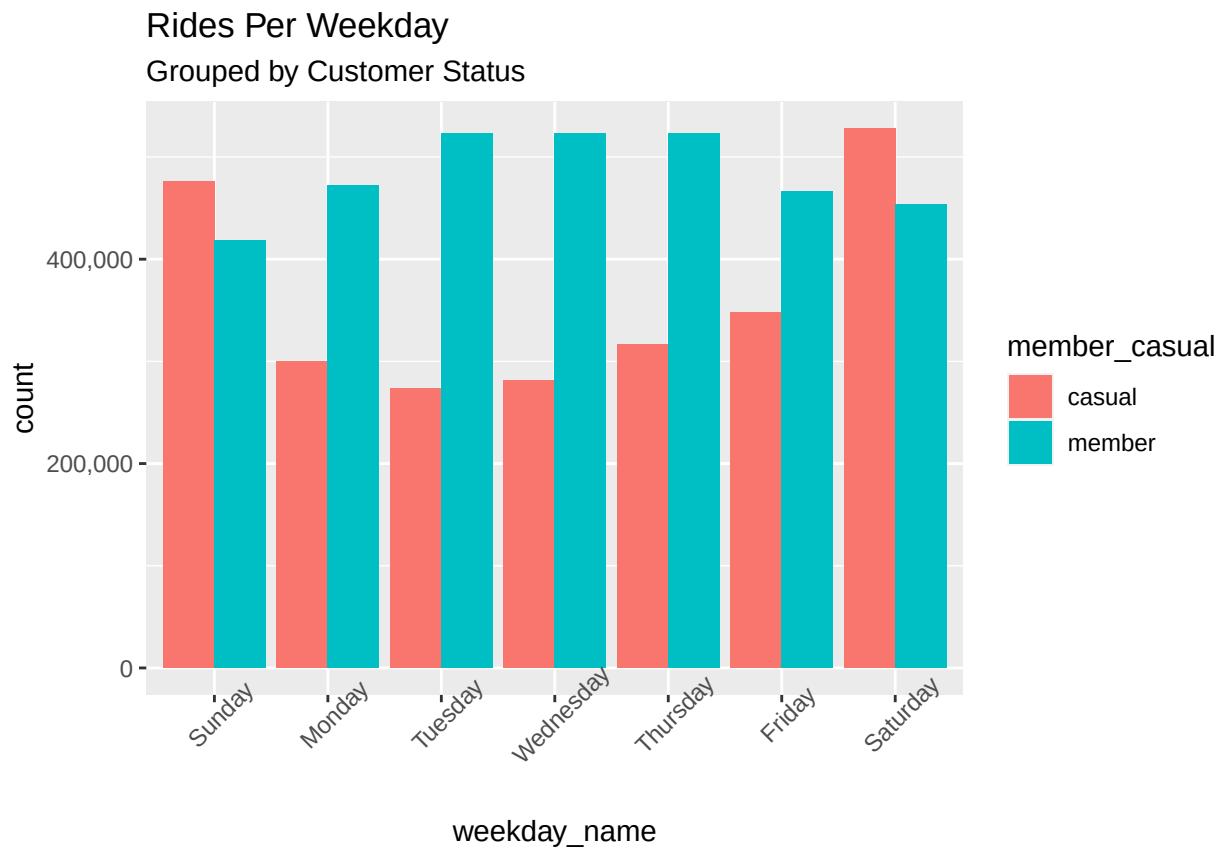


For annual members, the average bike ride duration for every month seems to be fairly consistent compared to more fluctuation with the casual riders. The consistency of annual members suggests that annual riders have routines that require shared-bike services. In addition, casual riders on average have twice as long duration on bike rides compared to annual members.

Weekday

Number of Rides Per day of the week for annual and casual riders.

```
cyclists$weekday_name <- ordered(cyclists$weekday_name,
                                 levels = c('Sunday', 'Monday', 'Tuesday', 'Wednesday',
                                            'Thursday', 'Friday', 'Saturday'))

cyclists %>%
  group_by(member_casual, weekday_name) %>%
  ggplot(aes(x = weekday_name, fill = member_casual))+
  geom_bar(position = 'dodge')+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "Rides Per Weekday", subtitle = "Grouped by Customer Status")+
  scale_y_continuous(labels = comma)
```



The graph above shows that annual members are more active during the middle of the week (Tuesday to Thursday) as opposed to casual members who are more active during the weekend (Saturday and Sunday).

```
#Find the mode
getmode <- function(x) {
  u = unique(x)
  u[which.max(tabulate(match(x,u)))]
}

aggregate(cyclists$weekday_name ~ cyclists$member_casual, FUN = getmode)
```

```
##   cyclists$member_casual cyclists$weekday_name
## 1                 casual              Saturday
## 2                 member               Tuesday
```

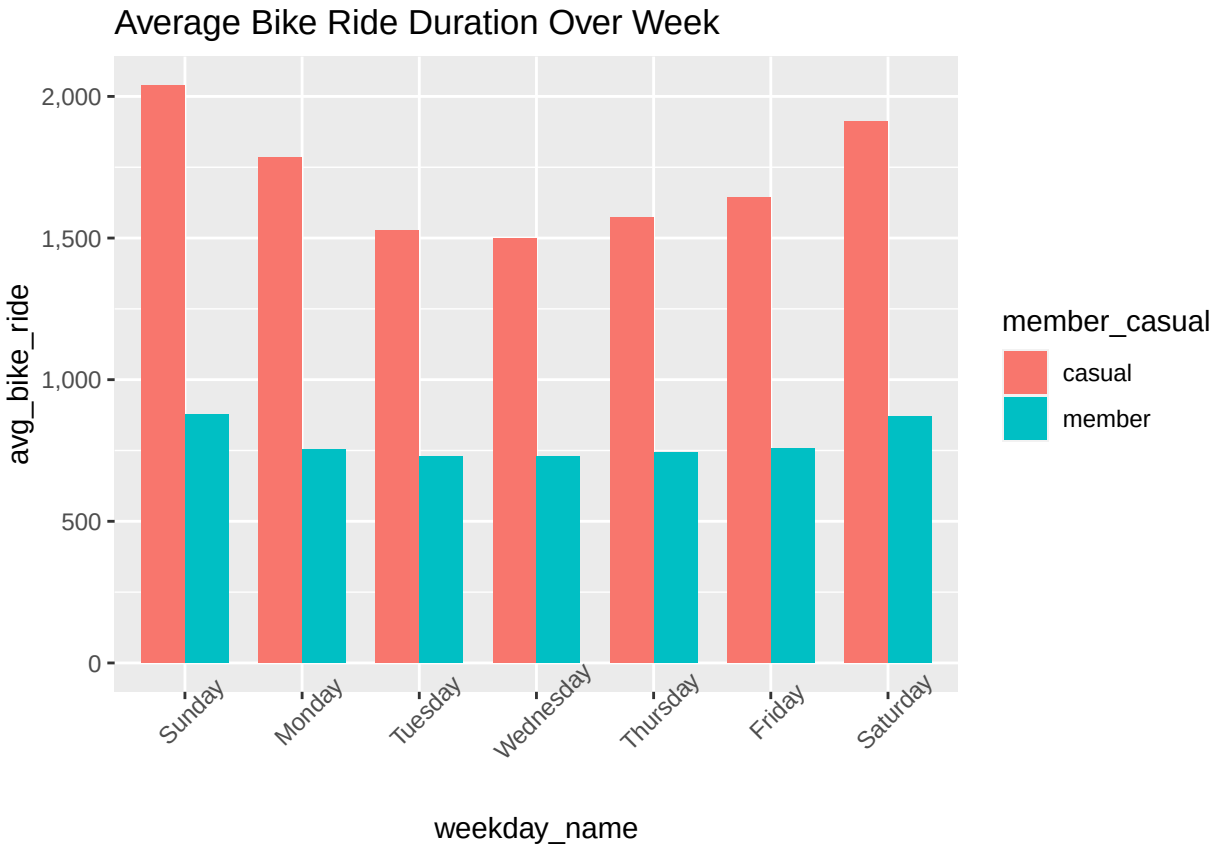
The busiest day for annual members is Tuesday. While the busiest day for casual riders is Saturday.

```
cyclists %>%
  group_by(member_casual, weekday_name) %>%
  summarise(number_of_rides = n(), avg_ride_time_secs = mean(ride_length)) %>%
  arrange(member_casual, desc(number_of_rides))
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday_name number_of_rides avg_ride_time_secs
```

```
##    <chr>        <ord>              <int>          <dbl>
##  1 casual       Saturday          527568         1910.
##  2 casual       Sunday            475591         2039.
##  3 casual       Friday            347636         1644.
##  4 casual       Thursday          316118         1572.
##  5 casual       Monday            299653         1783.
##  6 casual       Wednesday         281783         1500.
##  7 casual       Tuesday           273810         1527.
##  8 member       Tuesday           523377          729.
##  9 member       Thursday          522658          744.
## 10 member       Wednesday         522617          730.
## 11 member       Monday            472387          754.
## 12 member       Friday            466676          756.
## 13 member       Saturday          453486          868.
## 14 member       Sunday            417953          878.
```

```r
#Average ride length per day of the week
cyclists %>%
  group_by(member_casual, weekday_name) %>%
  summarise(avg_bike_ride = mean(ride_length)) %>%
  ggplot(aes(x = weekday_name, y = avg_bike_ride, fill = member_casual))+
  geom_col(width = 0.75, position = "dodge")+
  labs(title = "Average Bike Ride Duration Over Week")+
  theme(axis.text.x = element_text(angle = 45))+
  scale_y_continuous(labels = comma)
```

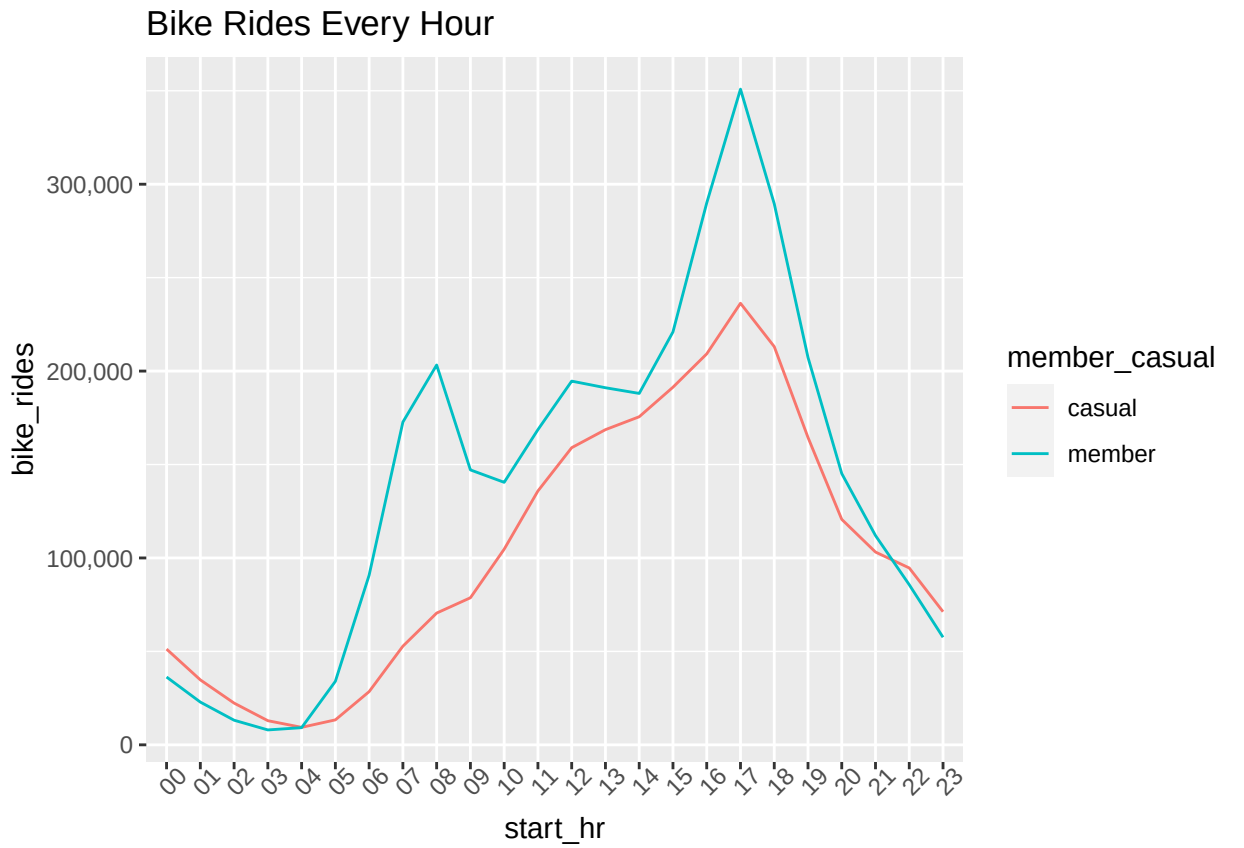


The average bike ride duration for annual members seems to be consistent throughout the week at around 700 seconds. The graph above also shows that casual members have about double the average bike ride duration to annual members for the entire week. Note this may suggest that casual riders travel longer distances or travel to multiple locations before returning the bike to a designated station. One consistency between both types of members is that they have longer bike rides during the weekend compared to the rest of the week.

Hour

Visualization of bike rides within a 24 Hour period.

```
cyclists %>%
  group_by(member_casual,start_hr) %>%
  summarise(bike_rides = n()) %>%
  ggplot(aes(x = start_hr, y = bike_rides, color = member_casual, group = member_casual))+
  geom_line()+
  labs(title = "Bike Rides Every Hour")+
  theme(axis.text.x = element_text(angle = 45))+
  scale_y_continuous(labels = comma)
```



Notice that annual members have three peaks at 8am, 12pm, and 5pm. These are common times for individuals to have breakfast, lunch, and dinner. The casual member is more active during 12pm to 5pm where bike rides dramatically fall after 5pm.

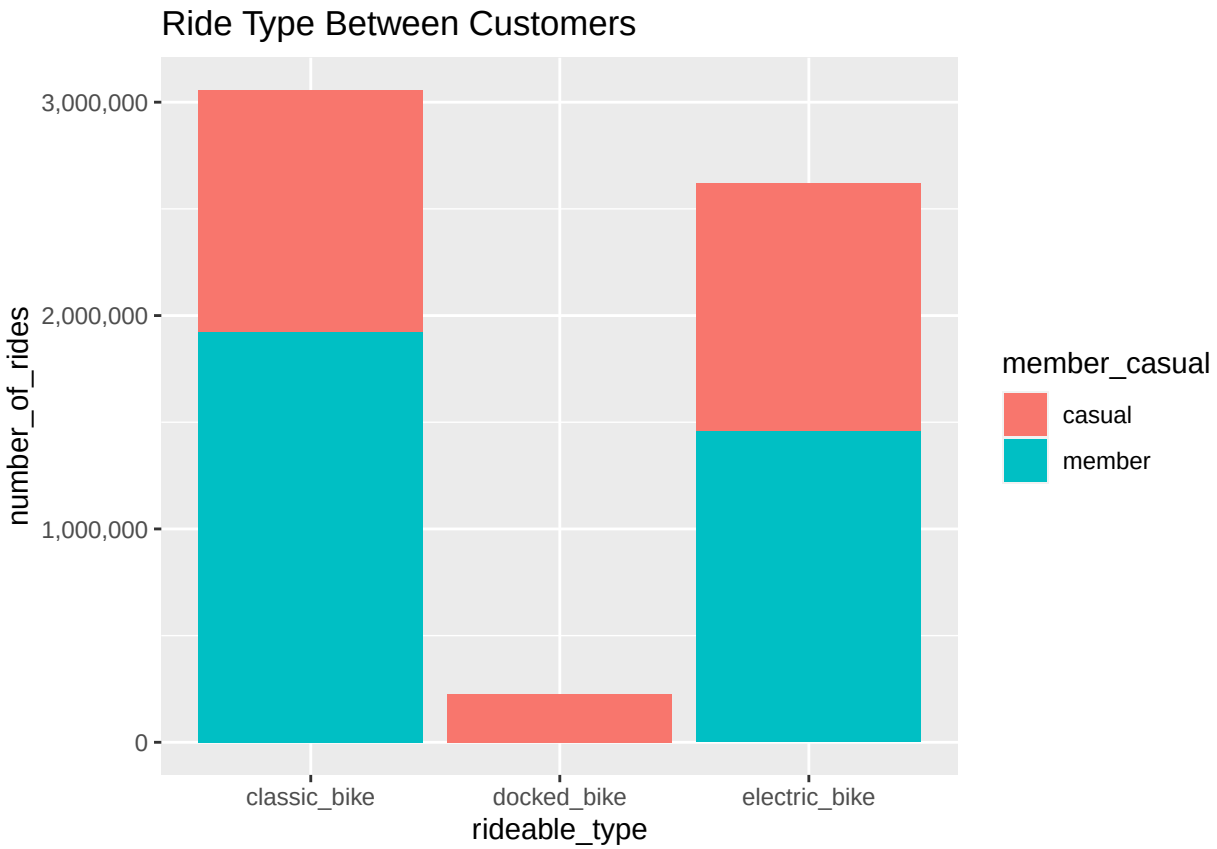
Type of Bike

Types of bikes customers use.

```
cyclists %>%
  group_by(member_casual, rideable_type) %>%
  summarise(bike_type_count = n()) %>%
  arrange(member_casual, desc(bike_type_count))
```

```
## # A tibble: 5 x 3
## # Groups:   member_casual [2]
##   member_casual rideable_type bike_type_count
##   <chr>         <chr>                  <int>
## 1 casual        electric_bike        1162568
## 2 casual        classic_bike         1132868
## 3 casual        docked_bike           226723
## 4 member        classic_bike         1922698
## 5 member        electric_bike        1456456
```

```
cyclists %>%
  group_by(member_casual, rideable_type) %>%
  summarise(number_of_rides = n()) %>%
  ggplot(aes(x = rideable_type, y = number_of_rides, fill = member_casual))+
  geom_col()+
  labs(title = "Ride Type Between Customers")+
  scale_y_continuous(labels = comma)
```



Classic bikes are primarily used by annual members. Electric bikes are evenly distributed among all riders. Docked bikes are only used by casual riders.

Summary of Ride Length

The following is a summary of statistics of the column "ride_length" for all riders.

```
summary(cyclists$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     370     657    1194    1189 2497750
```

Summary of Ride_length vs member_casual

The following is a summary of statistics of the column "ride_length" between casuals and annual members.

```
aggregate(cyclists$ride_length ~ cyclists$member_casual, FUN = mean)
```

```
##   cyclists$member_casual cyclists$ride_length
## 1                 casual            1752.8518
## 2                 member             776.0014
```

```
aggregate(cyclists$ride_length ~ cyclists$member_casual, FUN = median)
```

```
##   cyclists$member_casual cyclists$ride_length
## 1                 casual                  864
## 2                 member                  541
```

```
aggregate(cyclists$ride_length ~ cyclists$member_casual, FUN = max)
```

```
##   cyclists$member_casual cyclists$ride_length
## 1                 casual              2497750
## 2                 member                89998
```

```
aggregate(cyclists$ride_length ~ cyclists$member_casual, FUN = min)
```

```
##   cyclists$member_casual cyclists$ride_length
## 1                 casual                    0
## 2                 member                    0
```

Casual riders on average have longer duration on bike rides compared to annual members.

Findings

- Casual riders use bike-share services more during the weekend compared to annual members who use them at a consistent rate for the whole week.
- Annual members have more total rides per month with the exception of August 2021.
- Casual members have twice as long ride duration compared to annual members for all days of the week.
- During the start and end of the year bike-share services are at an all time low among both types of members.
- Casual Riders preferred docked bikes while annual members primarily used classic bikes.

Recommendations

- Promote discounts during the week day so more casual riders start using bike-share services consistently throughout the week. This consistent usage of bike-share services will entice casual members to upgrade to annual memberships to further indulge in the savings.
- Lower the price of annual membership renewal in order to retain current annual members while convincing casuals to convert to annual members.
- Offer discounts during non busy hours to increase rides throughout the day. Since casual members have less rides per hour, discounts will increase casual customer rides thus making casuals into returning customers then ultimately annual members.
- Increase the fleet of docked bikes since casual riders dominate the use of them.
- Create commercials about the use of bike-share services in the daily routine of various professions. The Commercial promotes the further use of bike-share services on a consistent basis which will lead to more returning customers. Returning customers have a higher probability of becoming annual members because annual memberships provide more savings compared to purchasing multiple daily passes throughout the year.

Further Research

- Data about Age and Gender might show what non-member population the company should target based on the current customers.
- Location data - Based on this data, the company can add more bikes to more popular stations. This will not only increase availability for current customers but also attract new customers as more bikes might encourage membership.
- Price between hour passes, daily passes, and annual memberships.
- Data about weather to see if there is any correlation between ride duration and weather or ride frequency and weather.