Ask Phase:

Business Objective:

1. The company cyclist has two types of consumers which are casual riders and subscribed members. Recently, executives at the company have identified that subscribed members generate more revenue for the company than casual riders and want to find the best incentive to convert casual riders into subscribed members. They acknowledge that the best way to approach this problem is by first identifying the differences between casual riders and subscribed members using product usage data and then use these differences to generate member conversion incentives.

Primary stakeholders:

1. Director of Marketing
2. My manager Lily Moreno

Deliverable:

1. To analyze data from several months of bicycle use from both members and casual riders throughout the year and take note of their differences.
2. Identify what these differences are and how they can be used to create an effective marketing regime geared towards convincing casual riders to sign up for an annual membership

Prepare Data:

Credibility

The dataset is from a public source so its credibility can be validated. The dataset is also reliable as it provides information about members and casual riders from making it feasible to explore trends likely revealed by changes in seasons. Finally the data is organized and divided in the form of multiple CSV files making it easy to download and analyze using programs like sql in cloud platforms like bigquery.

Datasources:

1. August divvy tripdata 2020
2. March divvy tripdata 2020
3. February divvy tripdata 2020
4. September divvy tripdata 2020

Process (in SQL):

1. Clean the data and prepare the data for analysis

# check if there are any duplicate ride\_id's

SELECT

ride\_id,

Count(\*),

FROM `temp-424017.Cyclistic\_Bike\_Share\_Analysis.Cyclistic Bicycle Data `

GROUP BY

ride\_id

HAVING

Count(\*)>1;

# check if there are any null values in any of the columns which could potentially render the analysis inaccurate

SELECT

start\_lat IS NULL OR ended\_at IS NULL OR start\_lng IS NULL OR end\_lat IS NULL OR end\_lng IS NULL

FROM `temp-424017.Cyclistic\_Bike\_Share\_Analysis.Cyclistic Bicycle Data `;

#comparing the minimum, and maximum started\_at times with the minimum and maximum ended\_at times and checking whether they are on the same or separate days. The results of the minimum and maximum starting at points and ending at points will be retrieved using the sql database and the results will be compared manually.

SELECT

min(started\_at), max(started\_at),

min(ended\_at), max(ended\_at),

FROM `temp-424017.Cyclistic\_Bike\_Share\_Analysis.Cyclistic Bicycle Data `;

Process(in R):

# install package and import file with columns added and modified with the help of SQL

install.packages('tidyverse')

library(tidyverse)

library(readr)

Bicycle\_Trip\_Data<-read.csv("Bicycle Trip Data 2020-Divvy\_Trips\_2020\_Q`.csv")

# View the first several rows of data, column names, and number of distinct rows to make sure that all of the data is intact and included in the imported file

View(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

> head(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

#View the column names associated with the table

> colnames(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

[1] "ride\_id" "rideable\_type" "started\_at" "ended\_at" "start\_station\_name"

#View the number of distinct rows associated with the table to make sure that all the data is included

> n\_distinct(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1$ride\_id)

[1] 426887

> nrow(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

[1] 426887

**Analyze:**

1. **Identify trends and relationships:**
   1. Now that the data has been cleaned and processed, the differences in trends exhibited by members and casual riders are about to be analyzed

# create a new bike frame with two columns indicating the difference in times between starting and ending locations and differences in distance between starting and ending locations

SELECT

member\_casual, AVG(TIMESTAMP\_DIFF(ended\_at, started\_at,SECOND)) AS avg\_duration\_seconds, AVG(SQRT(POW((end\_lat-start\_lat),2)+POW((end\_lng-start\_lng), 2)))\*111 AS avg\_distance\_km

FROM `temp-424017.Cyclistic\_Bike\_Share\_Analysis.Cyclistic Bicycle Data `

GROUP BY

member\_casual;

# Find out the number of rows there are dedicated to started\_at and ended\_at positions to determine the frequency of cycle usage between members and casual riders

SELECT

member\_casual,

COUNT(started\_at),

COUNT(ended\_at)

FROM `temp-424017.Cyclistic\_Bike\_Share\_Analysis.Cyclistic Bicycle Data `

GROUP BY

member\_casual;

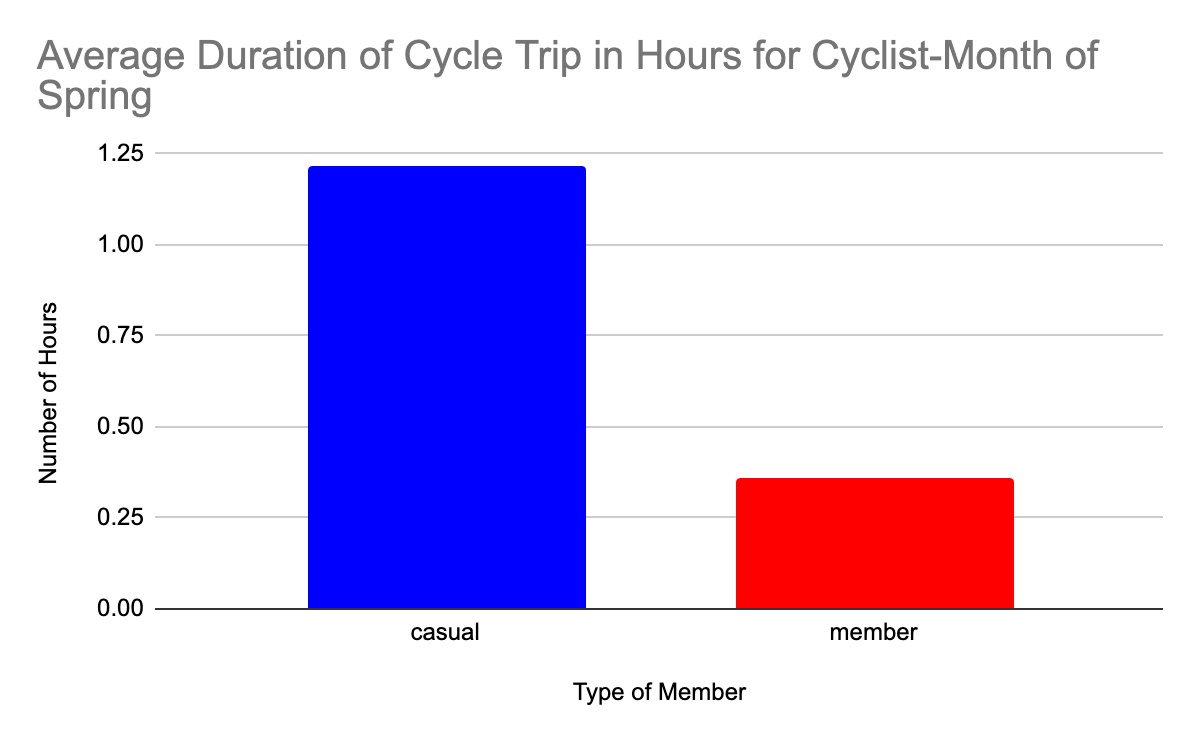
Results:

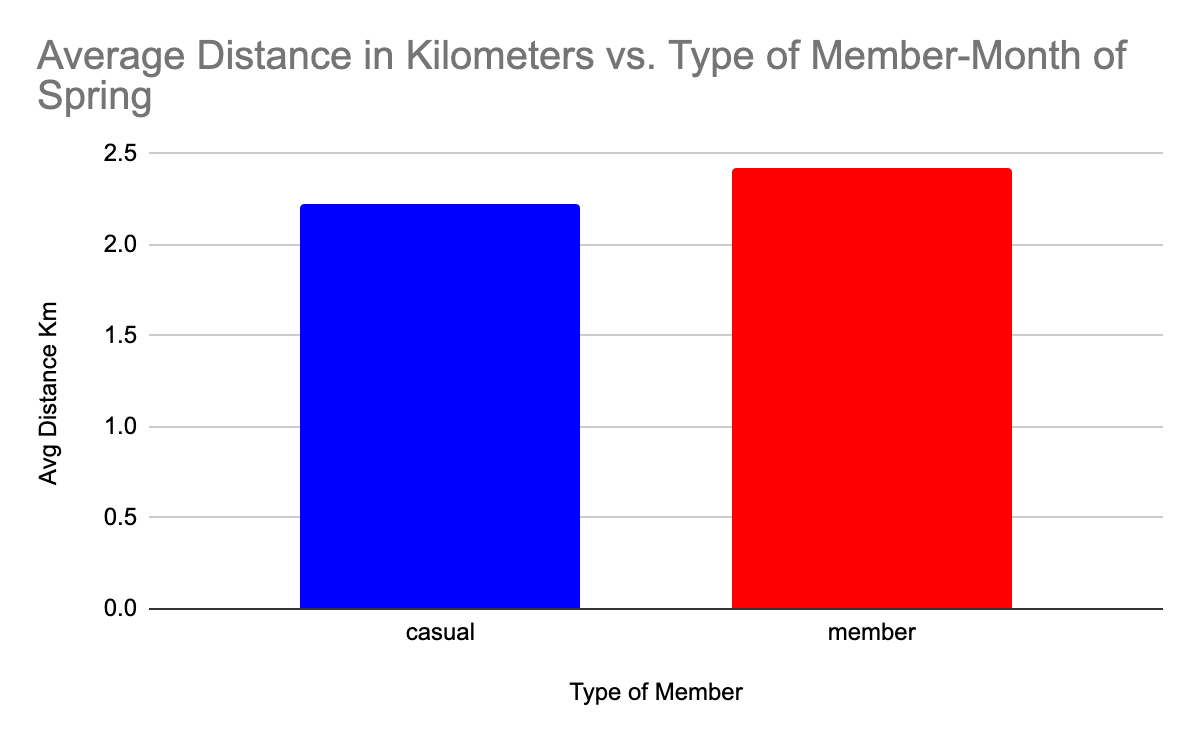
1. Data that is retrieved from SQL database and converted into tabular form using Microsoft Excel
2. Data presented in the table was then converted into graphs using Microsoft Excel

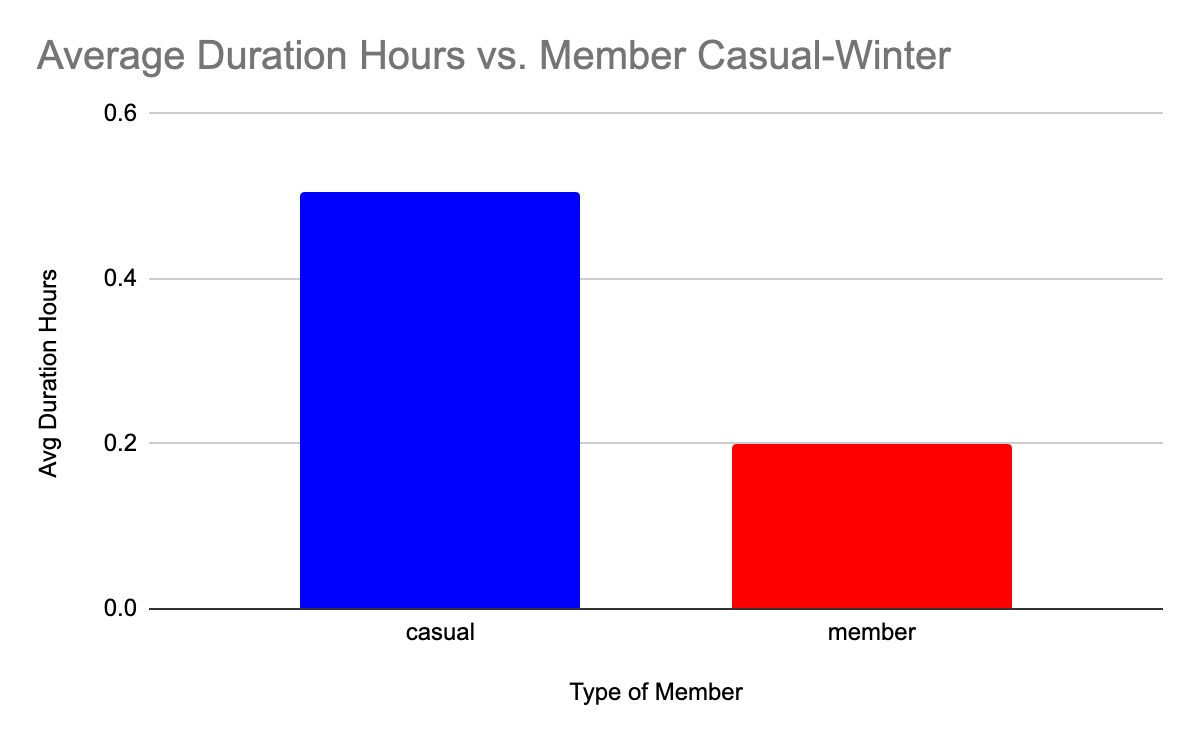
Data:

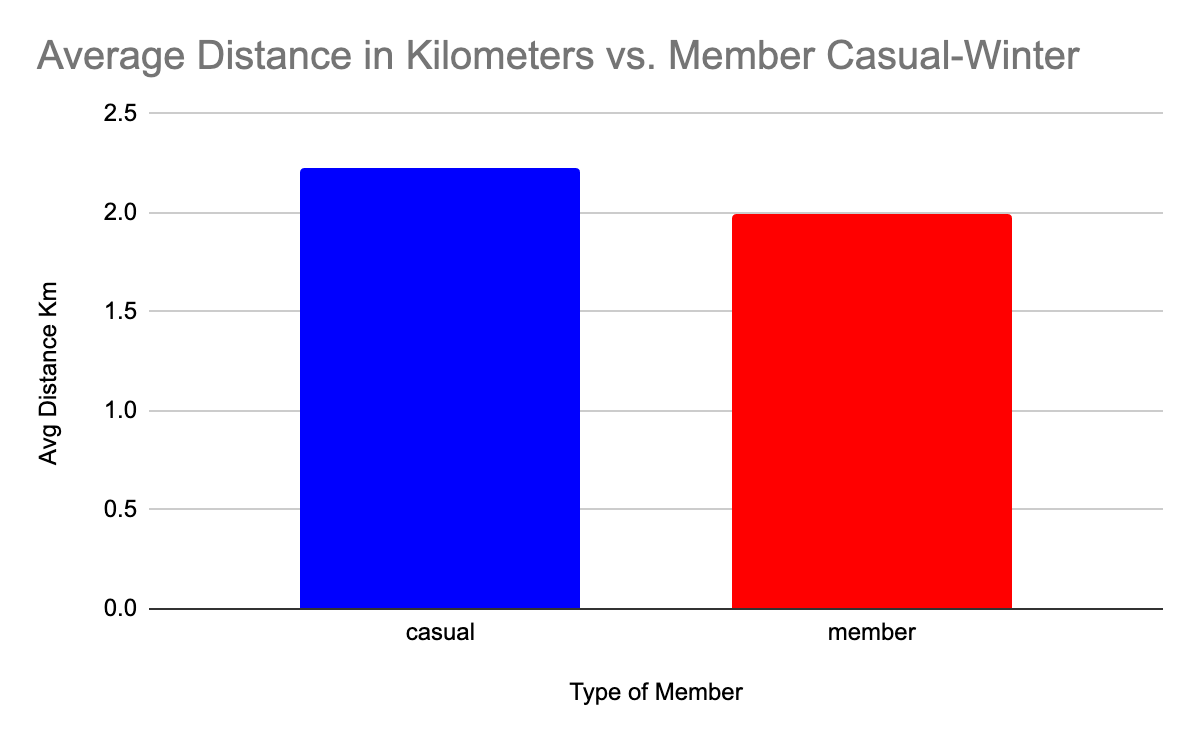
| **Comparison of Time and Distance Between Casual Users and Members For Each Season of the Year** |  |  |
| --- | --- | --- |
| **Fall** | **Avg Duration (Hours)** | **Avg Distance (km)** |
| casual | 2286.969626 | 2.566188356 |
| member | 927.6530572 | 2.609218363 |
| **Winter** | **Avg Duration (Hours)** | **Avg Distance (km)** |
| casual | 0.5062909167 | 2.228543666 |
| member | 0.1996918312 | 1.997499535 |
| **Spring** | **Avg Duration (Hours)** | **Avg Distance (km)** |
| casual | 1.2178531 | 2.219233635 |
| member | 0.3577916739 | 2.425823865 |
| **Summer** | **Avg Duration (Hours)** | **Avg Distance (km)** |
| casual | 0.9968628149 | 2.295270497 |
| member | 0.2948236592 | 2.67993026 |

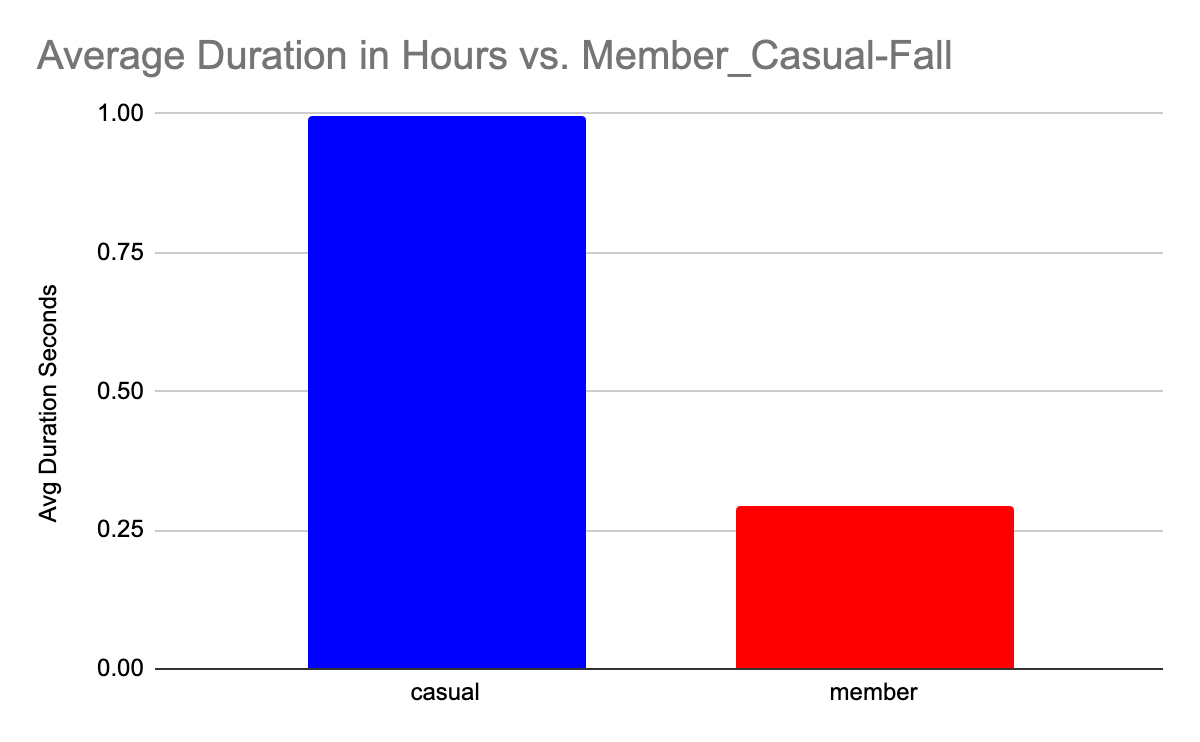
| **Comparison of Cycle Use Frequency Over the Various Months** |  |  |
| --- | --- | --- |
| **Fall** | **Cycle Use Frequency** | **Difference in Use Between Member and Casual** |
| member | 230692 | 71574 |
| member | 302266 |  |
| **Winter** | **Cycle Use Frequency** | **Difference in Use Between Member and Casual** |
| casual | 18520 | 66730 |
| member | 85250 |  |
| **Spring** | **Cycle Use Frequency** | **Difference in Use Between Member and Casual** |
| casual | 23628 | 37520 |
| member | 61148 |  |
| **Summer** | **Cycle Use Frequency** | **Difference in Use Between Member and Casual** |
| casual | 269296 | 12888 |
| member | 282184 |  |

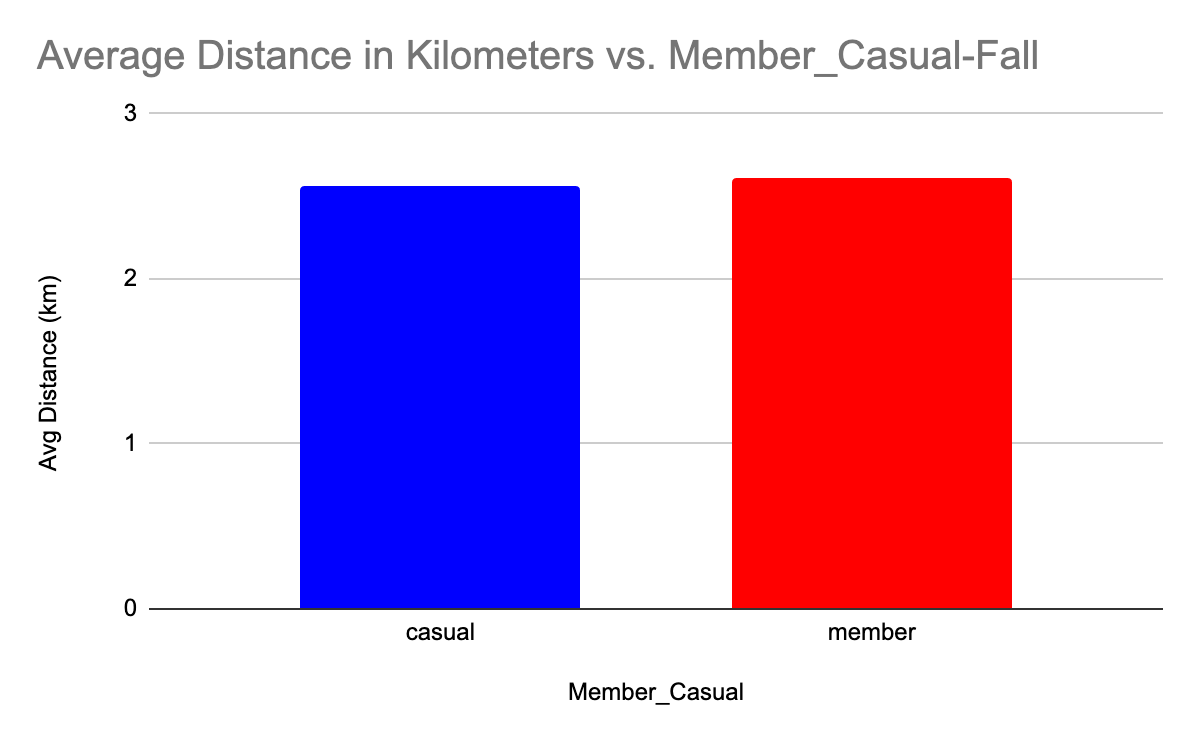


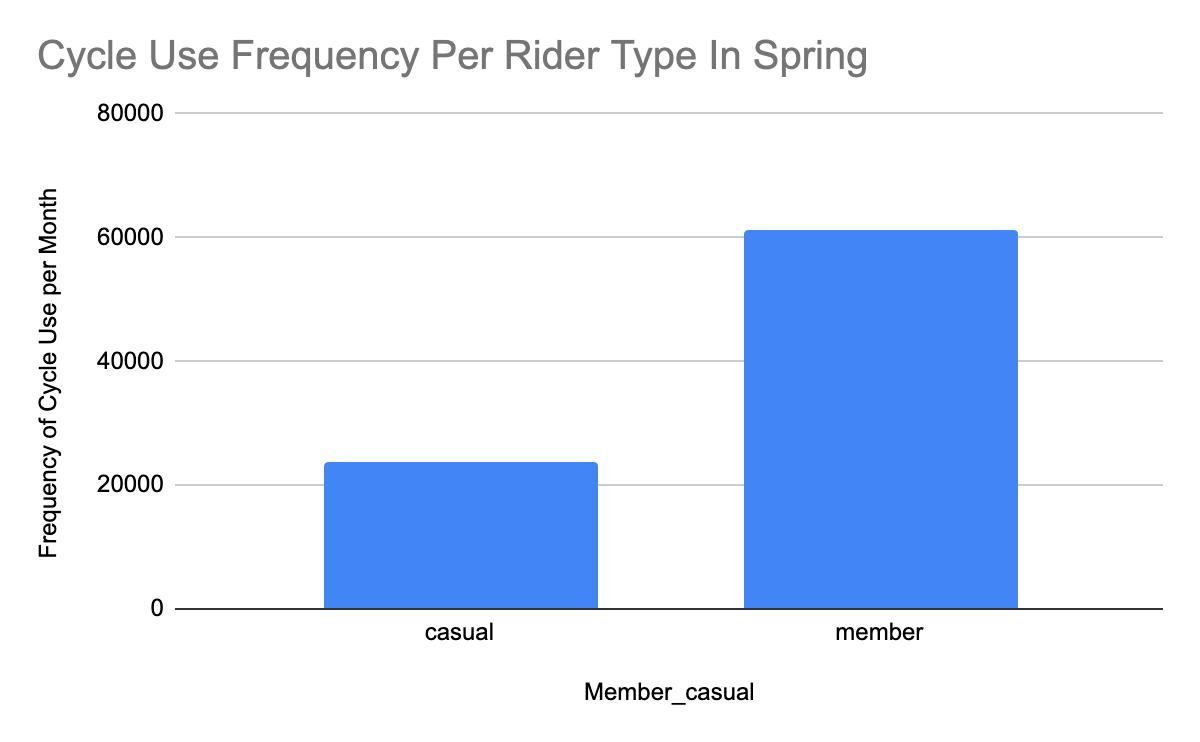


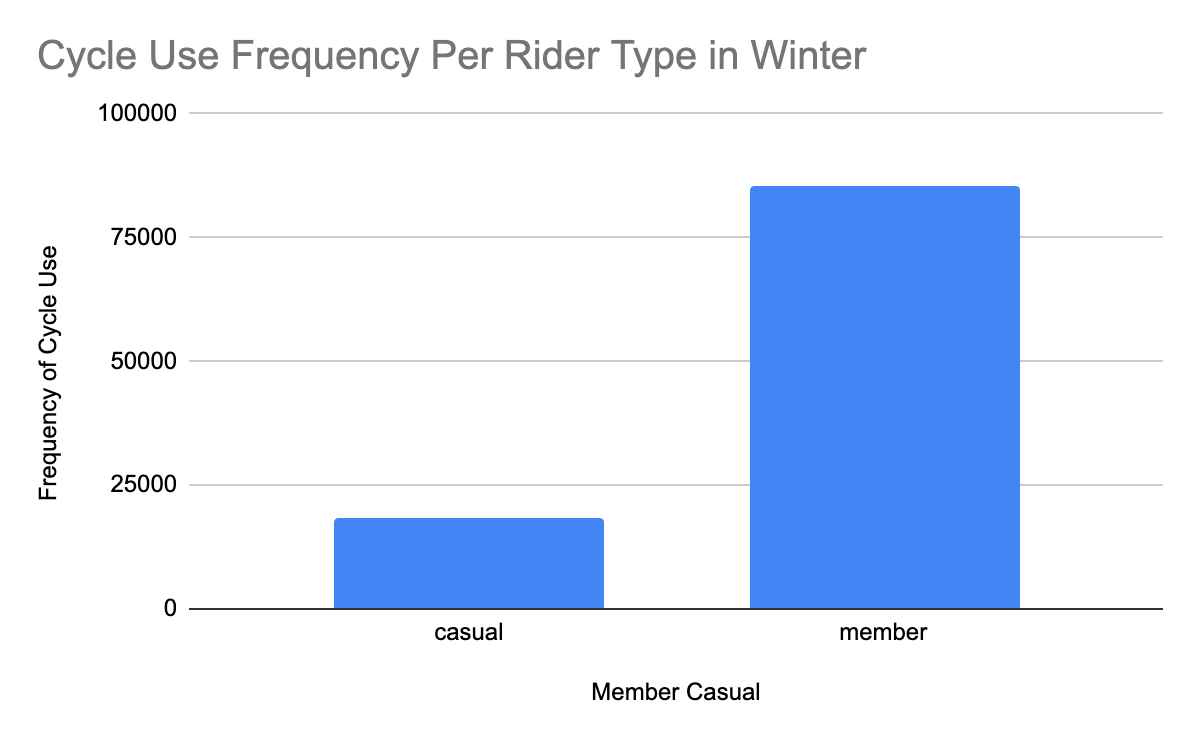


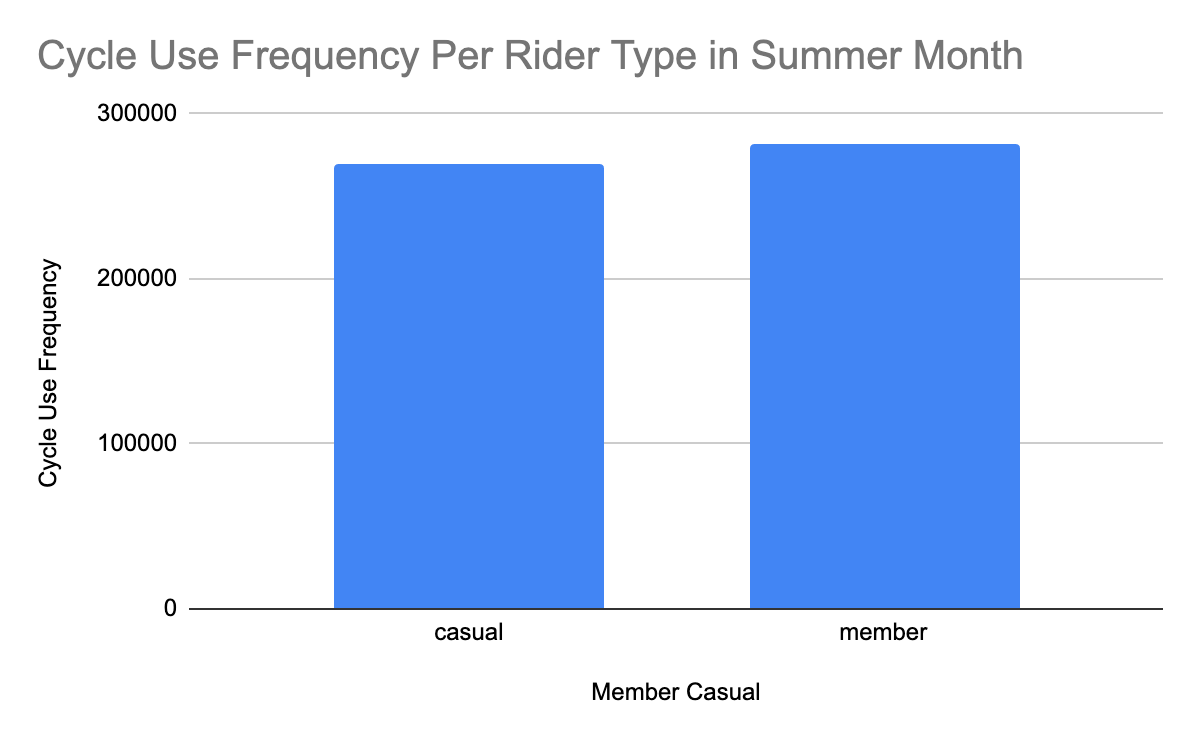


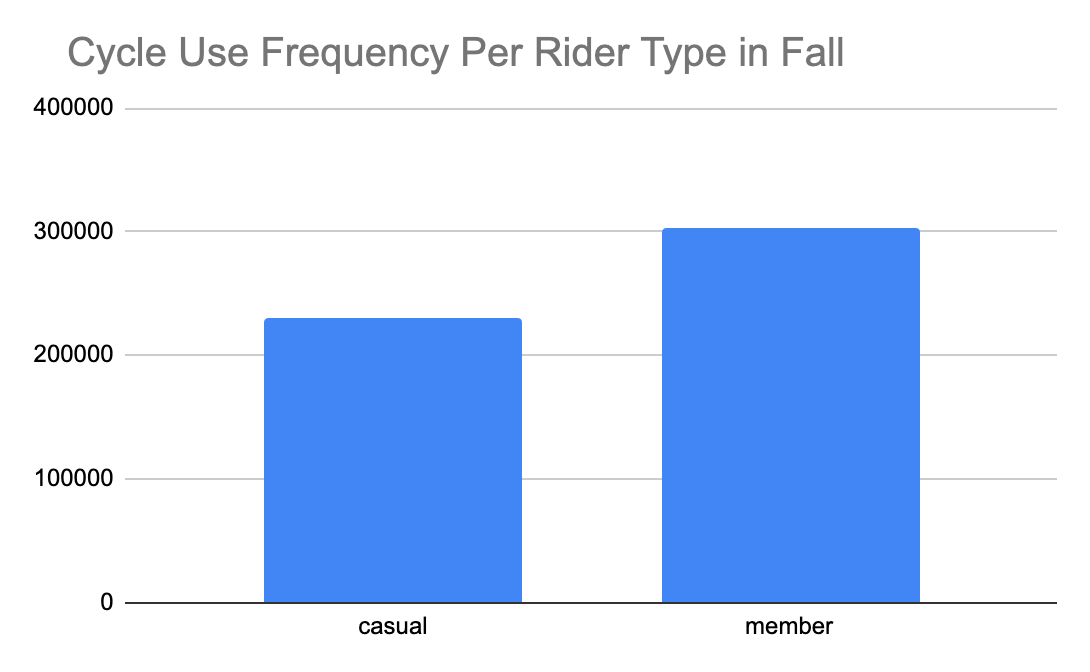












Analyze(In R):

# Select data that indicates the summary statistics of ride length to get a good understanding of how longest, shortest and longest amount of times cyclists have spent riding. This way you can get a good understanding of what is considered to be a good ride length, a below average ride length and a long ride length.

> Bicycle\_Data %>%

+ select(member\_casual,ride\_length) %>%

+ summary()

member\_casual ride\_length

Length:426887 Min. : -0.00639

Class :character 1st Qu.: 0.00381

Mode :character Median : 0.00637

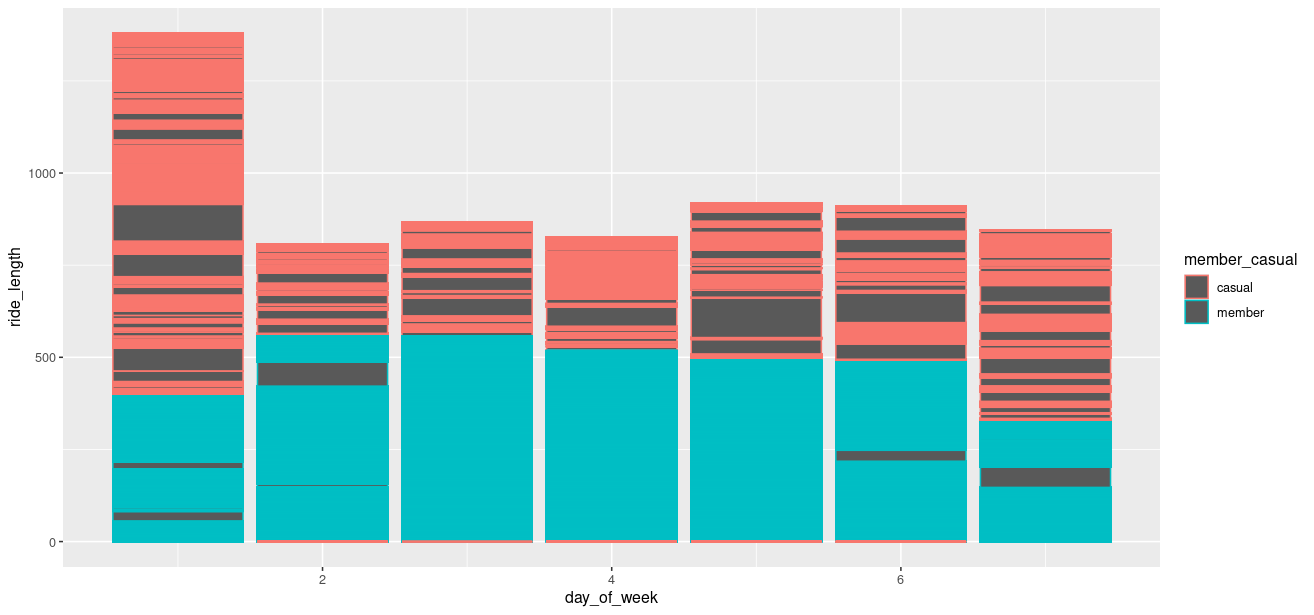
Mean : 0.01536

3rd Qu.: 0.01098

Max. :108.64611

# Creates a visualization using R to illustrate the relationship between type of member, the amount of time each ride takes and the days of the week they spend time riding

ggplot(data=Bicycle\_Data, aes(x=day\_of\_week, y=ride\_length, color=member\_casual)) + geom\_col()



The amount of distance traveled by the casual rider is higher than that of a member while the amount of time spent by the casual rider over members is higher. Casual riders seem to use their bicycles less frequently than regular members and spend significantly more of their time riding on mondays than members.

1. One way to address this marketing trend is to offer discounts for riders who spend less than two days a week riding and offer different payment options deviating based on the frequency of days a rider chooses to cycle. For example, the lowest payment offer could be offered to riders who ride less only once a week, medium to riders who ride 2-3 days and the highest payment option to riders who ride more than three days.

2. Offer a discount to new members that varies per season. The highest in the winter(the greatest difference in cycle use between members and the lowest in the summer(the highest difference in cycle use between members).

3. Offer an additional discount to new members that ride their bicycle for more than 30 minutes at a time, an activity that casual members have preference towards. This will push greater incentives towards joining as a new member as casual members have a tendency to ride for longer periods of time than traditional members.

Processing of Data:

View(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

> head(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

# A tibble: 6 × 15

ride\_id rideable\_type started\_at ended\_at start\_station\_name start\_station\_id end\_station\_name end\_station\_id start\_lat

<chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <dbl>

1 EACB19130B0… docked\_bike 2020-01-2… 2020-01… Western Ave & Lel… 239 Clark St & Lela… 326 42.0

2 8FED874C809… docked\_bike 2020-01-3… 2020-01… Clark St & Montro… 234 Southport Ave &… 318 42.0

3 789F3C21E47… docked\_bike 2020-01-0… 2020-01… Broadway & Belmon… 296 Wilton Ave & Be… 117 41.9

4 C9A388DAC6A… docked\_bike 2020-01-0… 2020-01… Clark St & Randol… 51 Fairbanks Ct & … 24 41.9

5 943BC3CBECC… docked\_bike 2020-01-3… 2020-01… Clinton St & Lake… 66 Wells St & Hubb… 212 41.9

6 6D9C8A69381… docked\_bike 2020-01-1… 2020-01… Wells St & Hubbar… 212 Desplaines St &… 96 41.9

> colnames(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

[1] "ride\_id" "rideable\_type" "started\_at" "ended\_at" "start\_station\_name"

[6] "start\_station\_id" "end\_station\_name" "end\_station\_id" "start\_lat" "start\_lng"

[11] "end\_lat" "end\_lng" "member\_casual" "ride\_length" "day\_of\_week"

> n\_distinct(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1$Id)

[1] 0

> n\_distinct(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1$ride\_id)

[1] 426887

> nrow(Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1)

[1] 426887

> Bicycle\_Data <- Bicycle\_Trip\_Data\_2020\_Divvy\_Trips\_2020\_Q1

> Bicycle\_Data %>%

+ select(member\_casual,ride\_length)

# A tibble: 426,887 × 2

member\_casual ride\_length

<chr> <dbl>

1 member 0.00522

2 member 0.00258

3 member 0.00198

4 member 0.00612

5 member 0.00384

6 member 0.0033[24m4

7 member 0.00334

8 member 0.00344

9 member 0.00341

10 member 0.00235

> Bicycle\_Data %>%

+ select(member\_casual,ride\_length) %>%

+ summary()

member\_casual ride\_length

Length:426887 Min. : -0.00639

Class :character 1st Qu.: 0.00381

Mode :character Median : 0.00637

Mean : 0.01536

3rd Qu.: 0.01098

Max. :108.64611

ggplot(data=Bicycle\_Data, aes(x=member\_casual, y=ride\_length)) + geom\_col()

Ride length per day of the week:

