

Case Studys

Wilton Ngo

2022-02-23

Ask Phase

To analyze how annual members different from Causal Riders that use Cyclistic Bikes. How does casual riders use their bikes compare to annual members. So we can create a marketing strategy for casual members to convert to annual members.

Prepare Phase

The data was download from Divvy Bikes Website. <https://ride.divvybikes.com/system-data>
The data is organized in Quartely CVS Zip Files. The data is orginial since its from their own company dataset. I use Google Sheets to open the dataset, when I opened I saw their are some missing entries for gender and birthyear. This personal information isn't necessary for purchasing bikes so some people don't write it down When I clean my data I will use R program because there are about 3.8 millon entries it will be easier to use R because it will be able to handle large datasets

Process Phase

##Install Required Packages and load Them #Install Require Packages #Tidyverse for data importing #lubridate for data function #ggplot for visualization

```
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Collect Data

#Upload Divvy Datasets(csv file here)

```
q2_2019 <- read_csv("Divvy_Trips_2019_Q2.csv")
q3_2019 <- read_csv("Divvy_Trips_2019_Q3.csv")
q4_2019 <- read_csv("Divvy_Trips_2019_Q4.csv")
q1_2020 <- read_csv("Divvy_Trips_2020_Q1.csv")
```

Combine Data into Single File

##Rename the column names so they are consistent with q1_2020

```
(q4_2019 <- rename(q4_2019,
                  ride_id = trip_id,
                  rideable_type = bikeid,
```

```

        started_at = start_time,
        ended_at = end_time,
        start_station_name = from_station_name,
        start_station_id = from_station_id,
        end_station_name = to_station_name,
        end_station_id = to_station_id,
        member_casual = usertype))

(q3_2019 <- rename(q3_2019,
        ride_id = trip_id,
        rideable_type = bikeid,
        started_at = start_time,
        ended_at = end_time,
        start_station_name = from_station_name,
        start_station_id = from_station_id,
        end_station_name = to_station_name,
        end_station_id = to_station_id,
        member_casual = usertype))

(q2_2019 <- rename(q2_2019
        ,ride_id = "01 - Rental Details Rental ID"
        ,rideable_type = "01 - Rental Details Bike ID"
        ,started_at = "01 - Rental Details Local Start Time"
        ,ended_at = "01 - Rental Details Local End Time"
        ,start_station_name = "03 - Rental Start Station Name"
        ,start_station_id = "03 - Rental Start Station ID"
        ,end_station_name = "02 - Rental End Station Name"
        ,end_station_id = "02 - Rental End Station ID"
        ,member_casual = "User Type"))

# convert variable to character type
q4_2019 <- mutate(q4_2019, ride_id = as.character(ride_id)
        ,rideable_type = as.character(rideable_type))
q3_2019 <- mutate(q3_2019, ride_id = as.character(ride_id)
        ,rideable_type = as.character(rideable_type))
q2_2019 <- mutate(q2_2019, ride_id = as.character(ride_id)
        ,rideable_type = as.character(rideable_type))
#Stack the individuals quater data fame into a single big data frame
all_trips <- bind_rows(q2_2019,q3_2019,q4_2019,q1_2020)

#remove fields of data
all_trips <- all_trips %>%
        select(-c(start_lat,start_lng,end_lat,birthyear,gender,
        "01 - Rental Details Duration In Seconds Uncapped",
        "05 - Member Details Member Birthday Year","Member Gender",
        "tripduration"))

##Clean Up And Add Data

```

#Reassign the desired values that will go with current 2020 labels

```
all_trips <- all_trips %>%  
  mutate(member_casual = recode(member_casual  
    , "Subscriber" = "member"  
    , "Customer" = "casual"))
```

#Add columns to list the Date, month, day, and year of each ride

```
all_trips$date <- as.Date(all_trips$started_at)  
all_trips$month <- format(as.Date(all_trips$date), "%m")  
all_trips$day <- format(as.Date(all_trips$date), "%d")  
all_trips$year <- format(as.Date(all_trips$date), "%y")  
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
```

#Add a ride length calculation to calculate the trip length in seconds

```
all_trips$ride_length <- difftime(all_trips$ended_at, all_trips$started_at)
```

#Convert ride length from factor to numeric so we can perform calculations

```
is.factor(all_trips$ride_length)
```

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

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))  
is.numeric(all_trips$ride_length)
```

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

#We will create a new dataframe(v2) since data is being removed

```
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" |  
  all_trips$ride_length < 0),]
```

Analyze Phase ## Descriptive Analysis

#Summary of ride length (in seconds)

```
summary(all_trips_v2$ride_length)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##         1      412      712   1479   1289 9387024
```

#Compare members and casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length  
## 1                          casual          3552.7502  
## 2                          member          850.0662
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length  
## 1                          casual             1546  
## 2                          member              589
```

```
aggregate(all_trips_v2$ride_length~ all_trips_v2$member_casual, FUN = max)
```

```
## all_trips_v2$member_casual all_trips_v2$ride_length
## 1 casual 9387024
## 2 member 9056634
```

```
aggregate(all_trips_v2$ride_length~ all_trips_v2$member_casual, FUN = min)
```

```
## all_trips_v2$member_casual all_trips_v2$ride_length
## 1 casual 2
## 2 member 1
```

#See the average time by each day for members vs casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$day_of_week, FUN = mean)
```

```
## all_trips_v2$day_of_week all_trips_v2$ride_length
## 1 Friday 1452.156
## 2 Monday 1296.635
## 3 Saturday 1964.200
## 4 Sunday 1993.965
## 5 Thursday 1324.161
## 6 Tuesday 1244.759
## 7 Wednesday 1275.481
```

#Fixing the days out of order

```
all_trips_v2$day_of_week <-
ordered(all_trips_v2$day_of_week, levels=c("Sunday",
"Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
```

#Run the average time by each day member vs casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual +
all_trips_v2$day_of_week, FUN = mean)
```

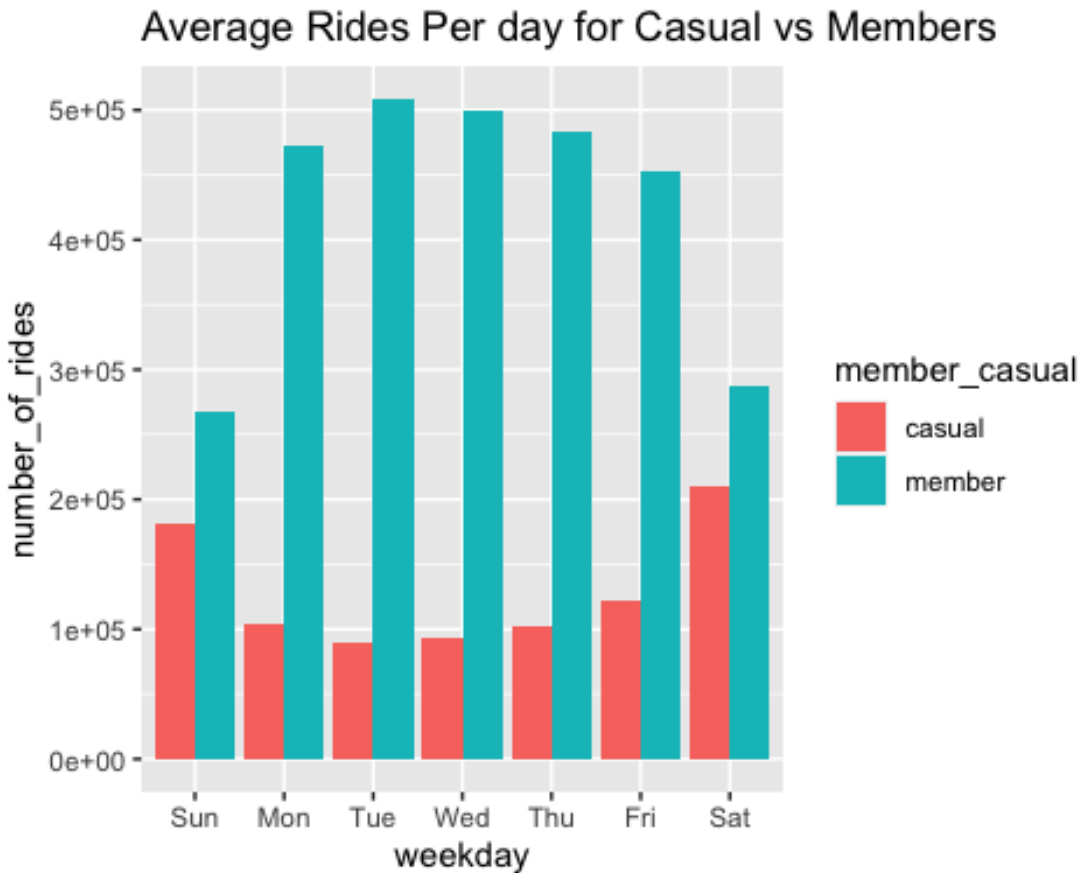
```
## all_trips_v2$member_casual all_trips_v2$day_of_week
all_trips_v2$ride_length
## 1 casual Sunday
3581.4054
## 2 member Sunday
919.9746
## 3 casual Monday
3372.2869
## 4 member Monday
842.5726
## 5 casual Tuesday
3596.3599
## 6 member Tuesday
826.1427
## 7 casual Wednesday
3718.6619
```

## 8	member	Wednesday
823.9996		
## 9	casual	Thursday
3682.9847		
## 10	member	Thursday
823.9278		
## 11	casual	Friday
3773.8351		
## 12	member	Friday
824.5305		
## 13	casual	Saturday
3331.9138		
## 14	member	Saturday
968.9337		

#Show the number riders per day

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label= TRUE)) %>%
  group_by(member_casual, weekday)%>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length))%>%
  arrange(member_casual, weekday)%>%
  ggplot(aes(x=weekday, y=number_of_rides, fill=member_casual))+
  geom_col(position="dodge")+
  labs(title = "Average Rides Per day for Casual vs Members")
```

`summarise()` has grouped output by 'member_casual'. You can override using the
`.groups` argument.

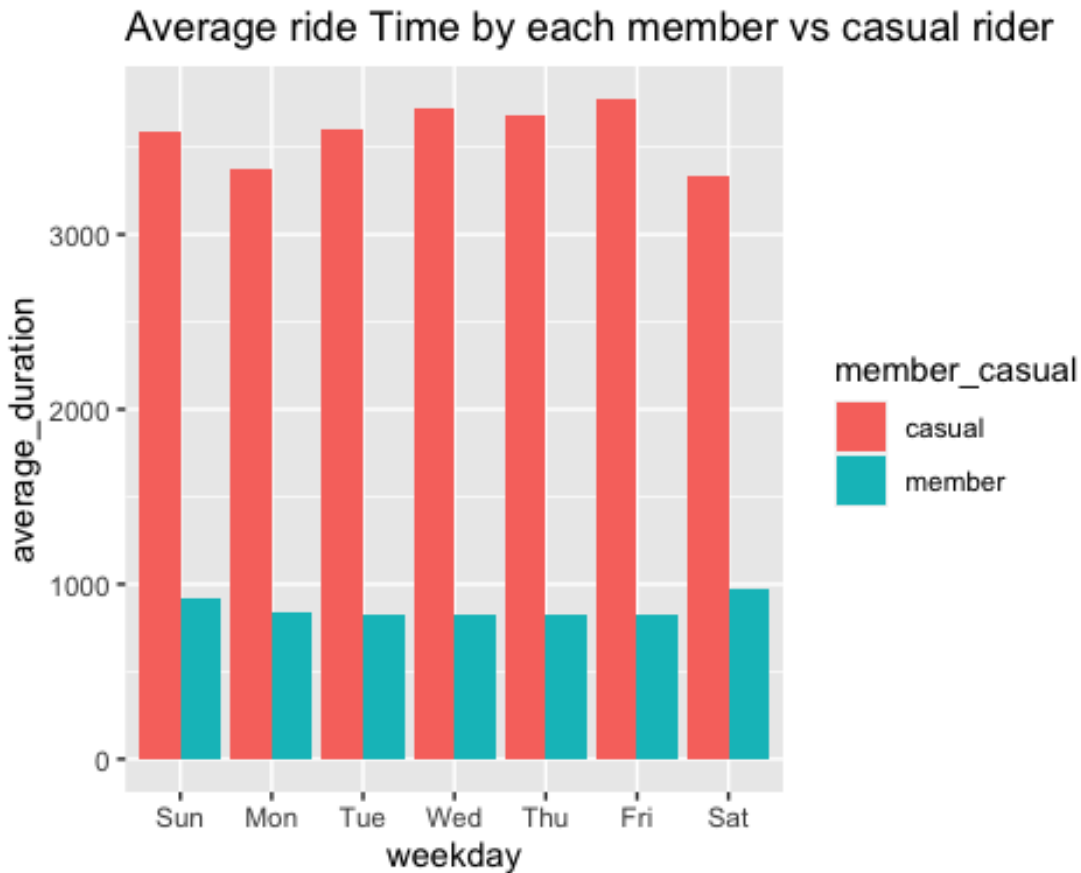


#Share

Phase

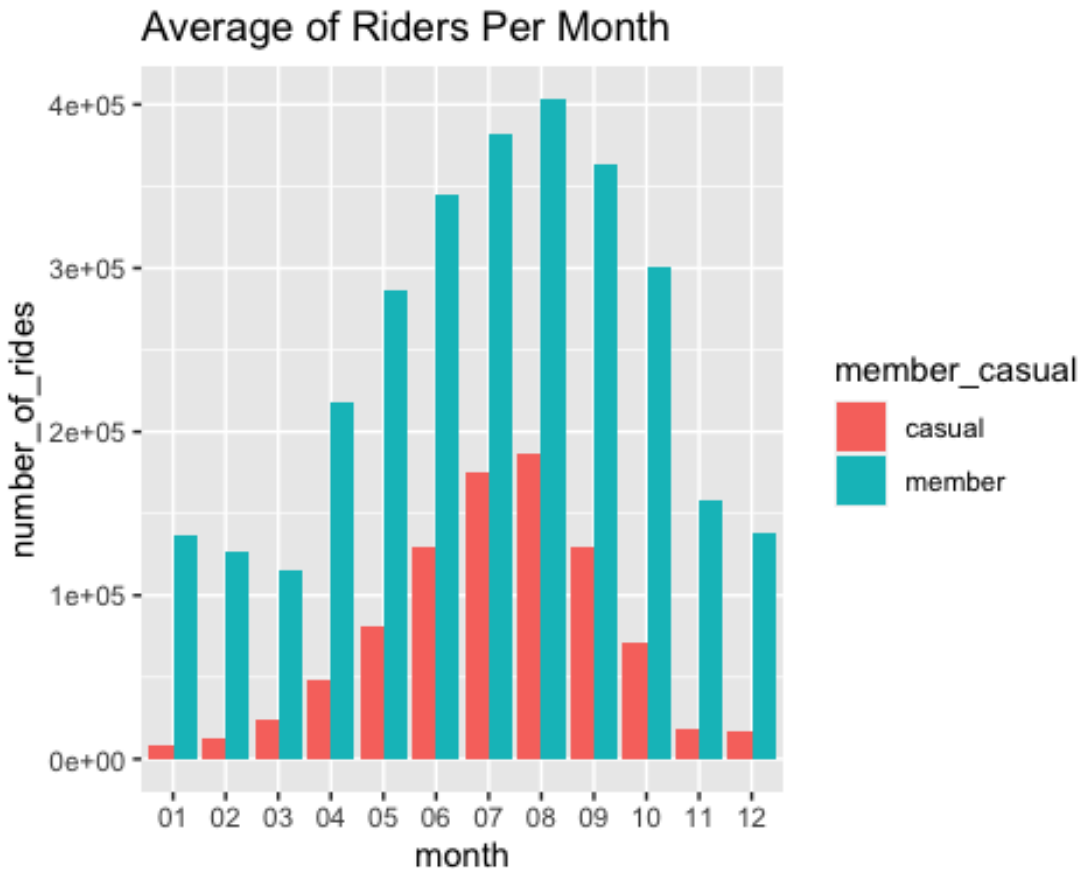
```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label= TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday) %>%
  ggplot(aes(x=weekday, y=average_duration, fill=member_casual))+
  geom_col(position="dodge")+
  labs(title = "Average ride Time by each member vs casual rider")
```

`summarise()` has grouped output by 'member_casual'. You can override using the
`.groups` argument.



```
all_trips_v2 %>%  
  mutate(weekday = month(started_at, label= TRUE)) %>%  
  group_by(member_casual, month)%>%  
  summarise(number_of_rides = n(),  
            average_duration = mean(ride_length))%>%  
  arrange(member_casual, month)%>%  
  ggplot(aes(x=month, y=number_of_rides, fill=member_casual))+  
  geom_col(position="dodge")+  
  labs(title = "Average of Riders Per Month")
```

`summarise()` has grouped output by 'member_casual'. You can override using the
`.groups` argument.



Conclusion

- My conclusion from my analysis of Cyclistic Bikes is that members ride more often, usually to work or to buy something without any cars or transportation. They ride short distances compared to casual riders who ride long distances for a hike or out with family and want to ride out and have fun.
- Members need to ride the bikes for transportation
- Members are more professional riders that do to prepare for a race or an event so they have to practice more.
- Casual riders only ride bikes mostly on the weekend because it is usually when they don't have work or school and can have fun.
- Casual riders usually ride more during the summer because that's when their kids are off, so they ride with family
- Some next steps that I would take with my stakeholders based on my findings is find a way for making the membership cheaper so people would buy the membership. We would collaborate to find the best solution to increase benefits to the membership that correlates for casual riders that fit their needs.

My top three Recommendations

- Reduce Membership Cost

- Increase Benefits for the membership Change
- How you charge, for example the distance, so people will get the membership just to be able to bike farther.