Capstone: Cyclistic Bike Share Performance

Allison Barton

12/5/2021

To see the presentation version of this project without the code, please click here.

Objective:

Identify the differences between annual and causal riders. Cyclistic is a bike-sharing company based in Chicago that has provided the historical data from the past 12 months (November 2020 - October 2021) of their ridership trends.

Data Preparation

The data was first cleaned in Excel to make importing easier. After ensuring strings were consistent and other values were formatted correctly, the first order of business was calculating the trip duration from the start and end times of each ride. Rows that contained a negative value, a zero, or whose starting station was "HQ QR" were then deleted, as this indicated maintenance from Cyclistic and was not a rider. Then, Excel's `=WEEKDAY()` function determined the day of week each trip occurred. Once the data was cleaned by month, it was imported into the R console for aggregation and analysis.

```
glimpse(all_trips)
```

```
## Rows: 5,336,079
## Columns: 9
## $ ride_id            <chr> "BD0A6FF6FFF9B921", "96A7A7A4BDE4F82D", "C61526D065~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "11/1/2020 13:36", "11/1/2020 10:03", "11/1/2020 0:~
## $ ended_at           <chr> "11/1/2020 13:45", "11/1/2020 10:14", "11/1/2020 1:~
## $ start_station_name <chr> "Dearborn St & Erie St", "Franklin St & Illinois St~
## $ end_station_name   <chr> "St. Clair St & Erie St", "Noble St & Milwaukee Ave~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
## $ ride_length        <time> 00:09:40, 00:11:19, 00:29:01, 00:09:15, 00:33:27, ~
## $ day_of_week        <dbl> 1, 1, 1, 1, 1, 7, 7, NA, 7, 7, 7, NA, 7, NA, 7, 7, ~
```

This data set contains a lot of useful information, but in a format that does not allow the R console to fully utilize it. In order to aggregate data by month or day, it is first necessary to parse that information from the `started_at` column. Since the dates in the `started_at` column are not in the standard YYYY-MM-DD format, a column changing the format was created in order to more easily retrieve the month and day. The `day_of_week` column was also reformatted to be more easily read.

```
all_trips$date <- as.Date(all_trips$started_at,"%m/%d/%Y")
all_trips$month <- format(as.Date(all_trips$date), "%b")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips <- all_trips %>%
  mutate(day_of_week = recode(day_of_week,
                              "1" = "Sunday",
                              "2" = "Monday",
                              "3" = "Tuesday",
                              "4" = "Wednesday",
                              "5" = "Thursday",
                              "6" = "Friday",
                              "7" = "Saturday"))
all_trips[1:5, c(3,9,10,11,12)]
```

```
## # A tibble: 5 x 5
##   started_at      day_of_week date       month day
##   <chr>           <chr>       <date>     <chr> <chr>
## 1 11/1/2020 13:36 Sunday      2020-11-01 Nov   01
## 2 11/1/2020 10:03 Sunday      2020-11-01 Nov   01
## 3 11/1/2020 0:34  Sunday      2020-11-01 Nov   01
## 4 11/1/2020 0:45  Sunday      2020-11-01 Nov   01
## 5 11/1/2020 15:43 Sunday      2020-11-01 Nov   01
```

Data Processing

Now that the data is clean and appropriately formatted, it's time to begin processing the data. Since the business objective seeks to identify how members and casual riders differ, a short summary of the ridership habits seemed like a good first step.

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

```
##   all_trips$member_casual all_trips$ride_length
## 1                  casual        1609.3261 secs
## 2                  member         807.3259 secs
```

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

```
##   all_trips$member_casual all_trips$ride_length
## 1                  casual            85482 secs
## 2                  member            74934 secs
```

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

```
##   all_trips$member_casual all_trips$ride_length
## 1                  casual                1 secs
## 2                  member                1 secs
```

```
all_trips %>% count(member_casual)
```

```
## # A tibble: 3 x 2
##   member_casual       n
##   <chr>           <int>
## 1 casual        2151861
## 2 member        2587613
## 3 <NA>           596605
```

From this quick glimpse, it appears that there are approximately the same number of casual riders and members. However, on average, a casual rider's trip is almost double the length than a member's trip. The minimum trip duration doesn't say very much, but the maximum duration does further emphasize that casual riders spend longer on the bicycles. That being said, these trends are a bit too broad. To evaluate trends from the past year, a comparison of the average ride time by month for both members and casual users was created, keeping in mind the data is from November 2020 to October 2021.

```
all_trips$month <- ordered(all_trips$month, levels = c("Nov",
                                                       "Dec",
                                                       "Jan",
                                                       "Feb",
                                                       "Mar",
                                                       "Apr",
                                                       "May",
                                                       "Jun",
                                                       "Jul",
                                                       "Aug",
                                                       "Sep",
                                                       "Oct"))
aggregate(all_trips$ride_length ~ all_trips$member_casual + all_trips$month, FUN = mean)
```

```
##    all_trips$member_casual all_trips$month all_trips$ride_length
## 1                   casual             Nov        1642.3993 secs
## 2                   member             Nov         802.4258 secs
## 3                   casual             Dec        1327.2852 secs
## 4                   member             Dec         739.5616 secs
## 5                   casual             Jan        1217.8992 secs
## 6                   member             Jan         735.3712 secs
## 7                   casual             Feb        1654.3488 secs
## 8                   member             Feb         880.4638 secs
## 9                   casual             Mar        1790.7750 secs
## 10                  member             Mar         817.6855 secs
## 11                  casual             Apr        1773.6907 secs
## 12                  member             Apr         850.0359 secs
## 13                  casual             May        1833.6370 secs
## 14                  member             May         855.8354 secs
## 15                  casual             Jun        1706.7201 secs
## 16                  member             Jun         845.3545 secs
## 17                  casual             Jul        1620.0237 secs
## 18                  member             Jul         825.8547 secs
## 19                  casual             Aug        1545.1770 secs
## 20                  member             Aug         813.8798 secs
## 21                  casual             Sep        1484.8757 secs
## 22                  member             Sep         789.3322 secs
## 23                  casual             Oct        1377.5858 secs
## 24                  member             Oct         723.2376 secs
```

Then, to seek an even more detailed trend, a similar comparison was ran by each day.

```
all_trips$day_of_week <- ordered(all_trips$day_of_week, levels = c("Sunday",
                                                                   "Monday",
                                                                   "Tuesday",
                                                                   "Wednesday",
                                                                   "Thursday",
                                                                   "Friday",
                                                                   "Saturday"))
aggregate(all_trips$ride_length ~ all_trips$member_casual + all_trips$day_of_week, FUN = mean)
```

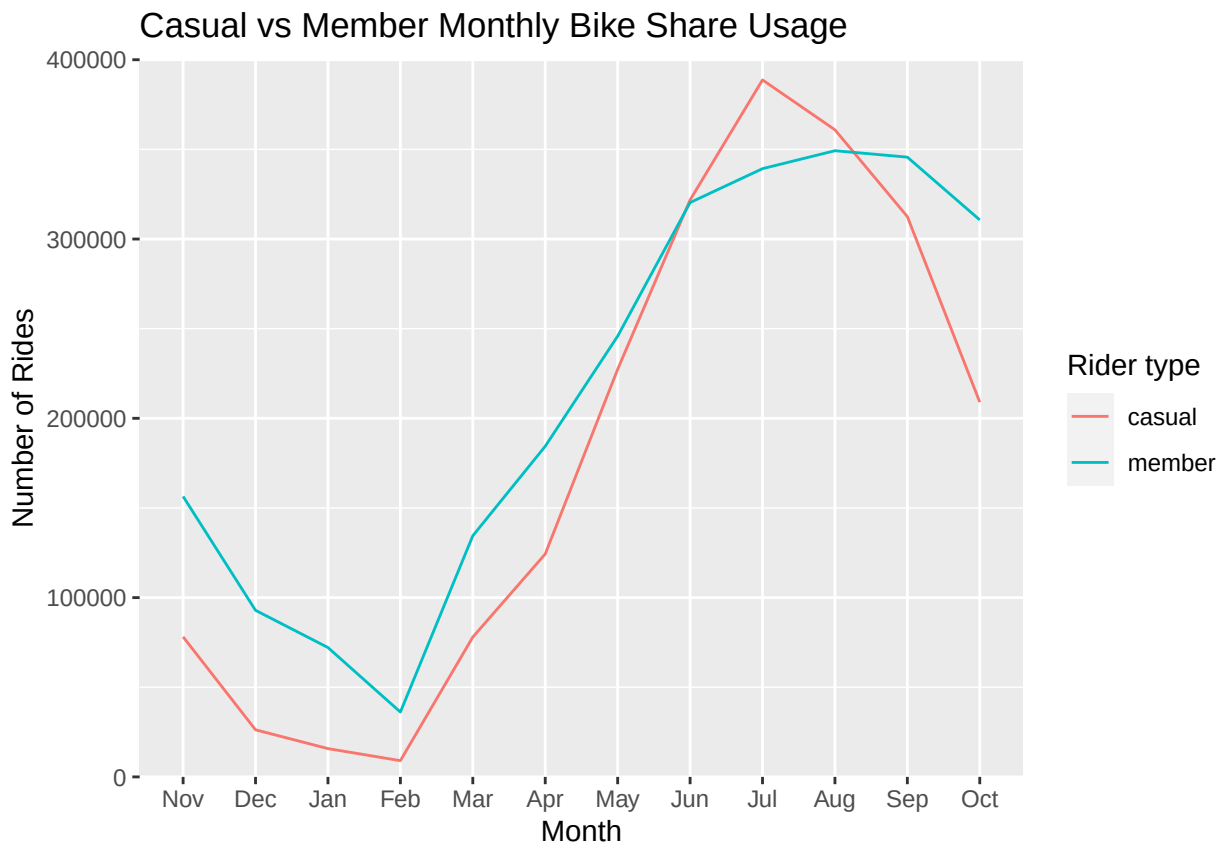
```
##    all_trips$member_casual all_trips$day_of_week all_trips$ride_length
## 1                  casual                 Sunday        1868.8242 secs
## 2                  member                 Sunday         922.4770 secs
## 3                  casual                 Monday        1642.8024 secs
## 4                  member                 Monday         780.6108 secs
## 5                  casual                Tuesday        1453.1181 secs
## 6                  member                Tuesday         760.1325 secs
## 7                  casual              Wednesday        1397.4431 secs
## 8                  member              Wednesday         763.5338 secs
## 9                  casual               Thursday        1369.9584 secs
## 10                 member               Thursday         757.7785 secs
## 11                 casual                 Friday        1475.1843 secs
## 12                 member                 Friday         783.7872 secs
## 13                 casual               Saturday        1739.1735 secs
## 14                 member               Saturday         901.1330 secs
```

These tables both seem to make sense given the initial query which suggested casual riders take longer trips than members. While these summaries are helpful, they are difficult to read in their current form. Creating visualizations from them will help identify patterns.

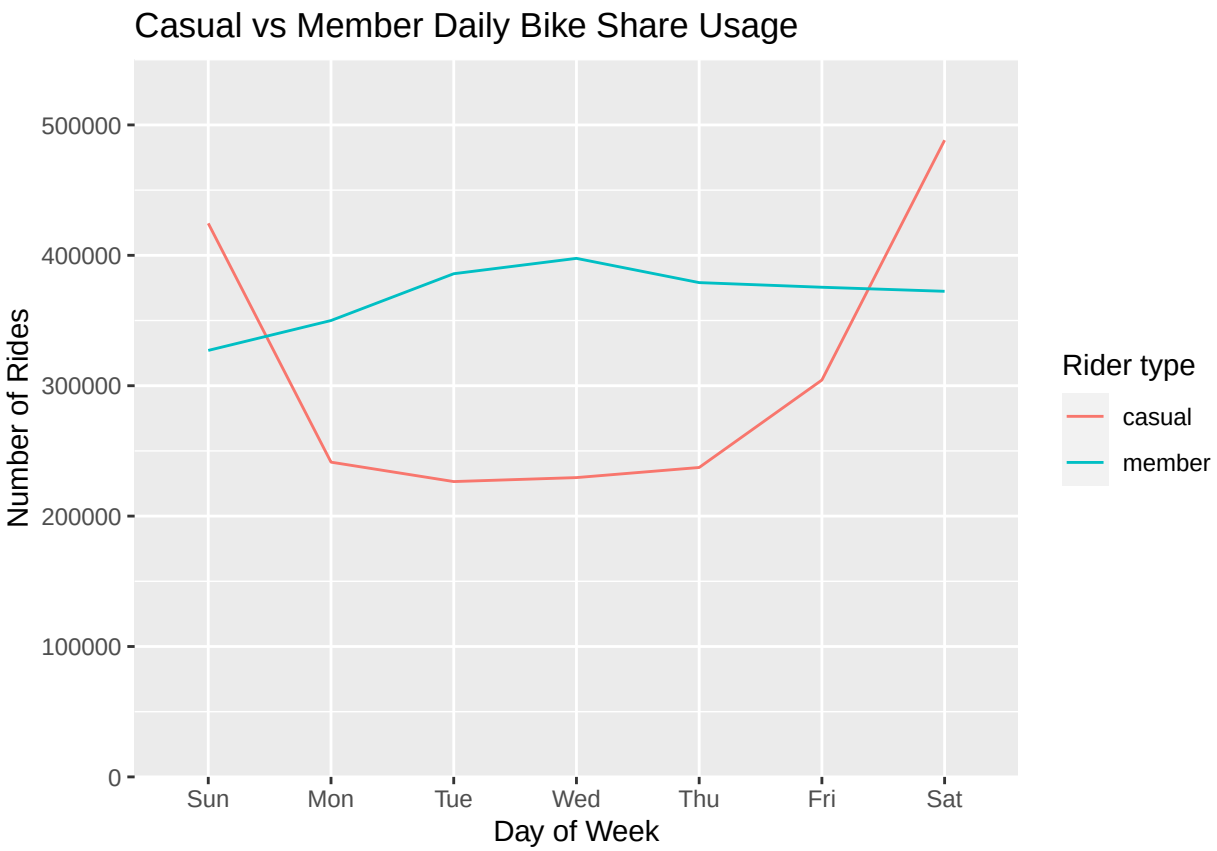
Data visualization

First, two queries to visualize the number of casual riders versus members by both monthly and daily trends were created to gain insights on their usage.

```
all_trips %>%
  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,
             group = member_casual,
             color = as.factor(member_casual))) +
  geom_line() +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 400000)) +
  labs(x = "Month", y = "Number of Rides",
       title = "Casual vs Member Monthly Bike Share Usage",
       color = "Rider type",
       options(scipen=10000))
```

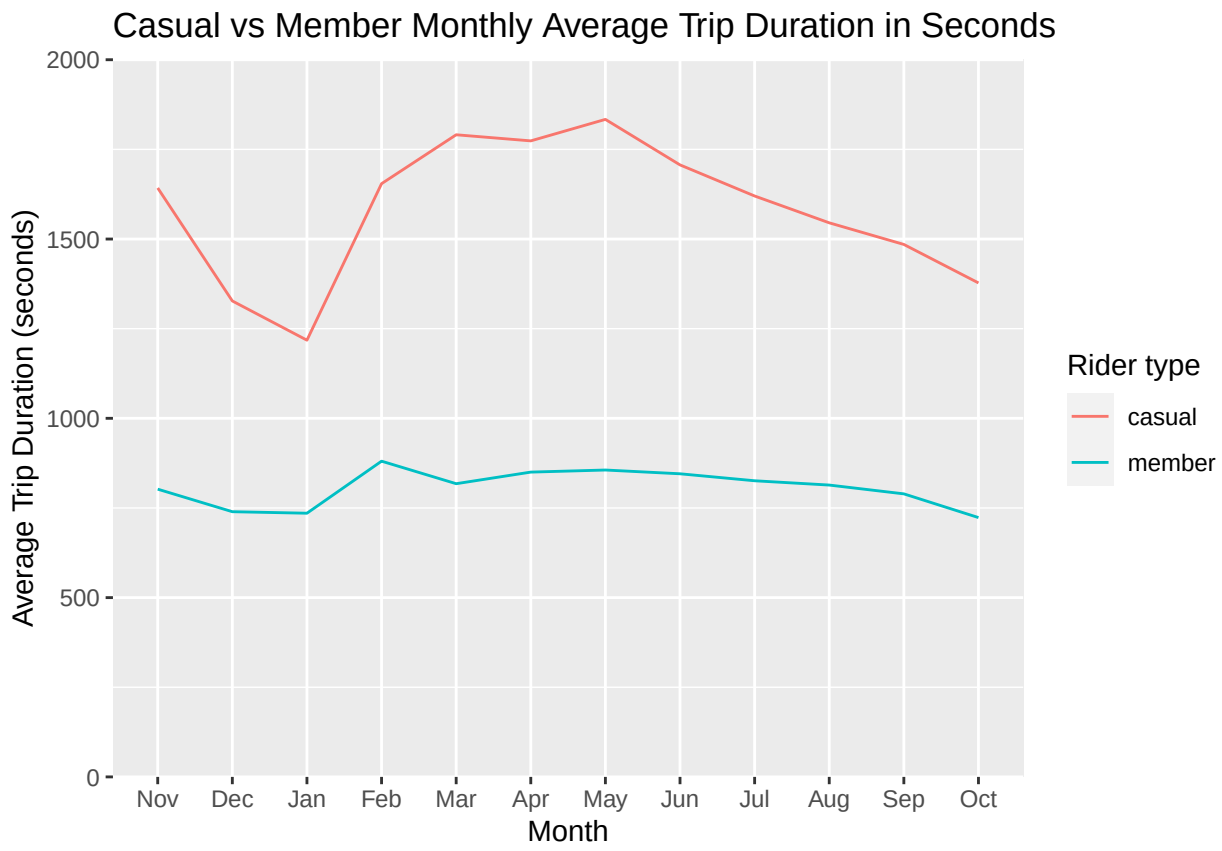



```
all_trips %>%
  mutate(day_of_week = wday(date,label = TRUE)) %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length)) %>%
  arrange(member_casual,day_of_week) %>%
  ggplot(aes(x=day_of_week, y = number_of_rides,
             group = member_casual,
             color = as.factor(member_casual))) +
    geom_line() +
    scale_y_continuous(expand = c(0, 0), limits = c(0, 550000)) +
    labs(x = "Day of Week", y = "Number of Rides",
         title = "Casual vs Member Daily Bike Share Usage",
         color = "Rider type",
         options(scipen=10000))
```

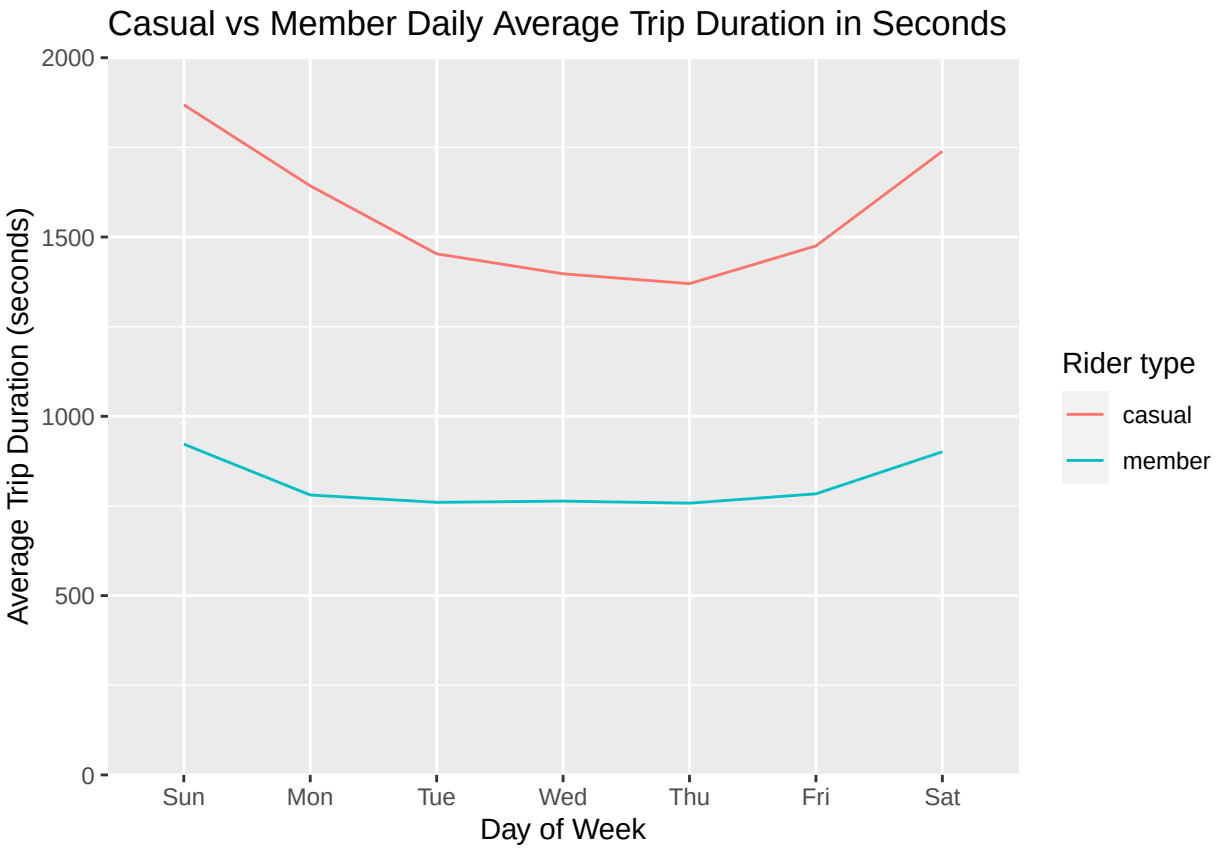


Then, two more queries to see the average time the riders spent with their bicycle per trip, again by monthly and daily trends.

```
all_trips %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length)) %>%
  arrange(member_casual,month) %>%
  ggplot(aes(x=month, y = average_duration,
             group = member_casual,
             color = as.factor(member_casual))) +
  geom_line() +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 2000)) +
  labs(x = "Month", y = "Average Trip Duration (seconds)",
       title = "Casual vs Member Monthly Average Trip Duration in Seconds",
       color = "Rider type")
```




```
all_trips %>%
  mutate(day_of_week = wday(date,label = TRUE)) %>%
  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,
             group = member_casual,
             color = as.factor(member_casual))) +
  geom_line() +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 2000)) +
  labs(x = "Day of Week", y = "Average Trip Duration (seconds)",
       title = "Casual vs Member Daily Average Trip Duration in Seconds",
       color = "Rider type")
```



Analysis

These visualizations help provide valuable insights to the ridership habits of annual membership riders and casual riders for Cyclistic bikes. The first graph suggests that both members and casual riders follow similar trends during throughout the year, but casual riders overtake members in the summer months. The second indicates that the number of members using the bicycles during the week is relatively even, but casual riders show a significant increase in usage during the weekends and very little usage during the week.

In addition, the third and fourth graphs indicate that casual riders as a whole take longer rides than members, whether it is examined yearly or weekly. A notable trend casual riders show is a considerable decrease in ride time during the cold months of December and January.

Conclusion and Further Research

Overall, this study found that casual users, as opposed to annual members, are more likely to ride during weekends and summer months, which makes those prime times to advertise memberships.

The study proposes further analysis on the purpose of bicycle usage between the groups, and hypothesizes that Cyclistic members may be riding to and from work or school while casual riders might be using the bicycles for leisure or exercise. Surveys could also be used to include demographic information such as age or gender of members to understand which groups are being missed in advertising.

Sources

This public data has been provided by Motivate International Inc.