Case1_analysis

Joe

2024-11-04

Cyclistic bike-share analysis (case study)

Stakeholders: Marketing analytics team and Executive team, Director of Marketing Lily Moreno

Reporting to: Manager Lily Moreno

Inspect the Data to understand it (dimensions, shape. . .)

```
##  [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
##  [5] "start_station_name" "start_station_id"   "end_station_name"   "end_station_id"
##  [9] "start_lat"          "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
## tibble [6,471,332 × 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:6471332] "903C30C2D810A53B" "F2FB18A98E110A2B" "D0DEC7C94E4663DA" "E0DE
##  $ rideable_type     : chr [1:6471332] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:6471332] "2023-08-19 15:41:53" "2023-08-18 15:30:18" "2023-08-30 16:15
##  $ ended_at          : chr [1:6471332] "2023-08-19 15:53:36" "2023-08-18 15:45:25" "2023-08-30 16:27
##  $ start_station_name: chr [1:6471332] "LaSalle St & Illinois St" "Clark St & Randolph St" "Clark St
##  $ start_station_id  : chr [1:6471332] "13430" "TA1305000030" "TA1305000030" "KA1504000135" ...
##  $ end_station_name  : chr [1:6471332] "Clark St & Elm St" "" "" "" ...
##  $ end_station_id    : chr [1:6471332] "TA1307000039" "" "" "" ...
##  $ start_lat         : num [1:6471332] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:6471332] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:6471332] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:6471332] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:6471332] "member" "member" "member" "member" ...
## Rows: 6,471,332
## Columns: 13
## $ ride_id           [3m[38;5;246m<chr>[39m[23m "903C30C2D810A53B", "F2FB18A98E110A2B", "D0DEC7C94E4
## $ rideable_type     [3m[38;5;246m<chr>[39m[23m "electric_bike", "electric_bike", "electric_bike", "
## $ started_at        [3m[38;5;246m<chr>[39m[23m "2023-08-19 15:41:53", "2023-08-18 15:30:18", "2023-
## $ ended_at          [3m[38;5;246m<chr>[39m[23m "2023-08-19 15:53:36", "2023-08-18 15:45:25", "2023-
## $ start_station_name [3m[38;5;246m<chr>[39m[23m "LaSalle St & Illinois St", "Clark St & Randolph St"
## $ start_station_id  [3m[38;5;246m<chr>[39m[23m "13430", "TA1305000030", "TA1305000030", "KA150400013
## $ end_station_name  [3m[38;5;246m<chr>[39m[23m "Clark St & Elm St", "", "", "", "", "", "", "", "",
## $ end_station_id    [3m[38;5;246m<chr>[39m[23m "TA1307000039", "", "", "", "", "", "", "", "", "
## $ start_lat         [3m[38;5;246m<dbl>[39m[23m 41.89072, 41.88451, 41.88498, 41.90310, 41.88555, 41
## $ start_lng         [3m[38;5;246m<dbl>[39m[23m -87.63148, -87.63155, -87.63079, -87.63467, -87.6320
## $ end_lat           [3m[38;5;246m<dbl>[39m[23m 41.90297, 41.93000, 41.91000, 41.90000, 41.89000, 41
## $ end_lng           [3m[38;5;246m<dbl>[39m[23m -87.63128, -87.64000, -87.63000, -87.62000, -87.6800
## $ member_casual     [3m[38;5;246m<chr>[39m[23m "member", "member", "member", "member", "member", "m
##    ride_id          rideable_type       started_at          ended_at
##  Length:6471332     Length:6471332     Length:6471332     Length:6471332
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:6471332     Length:6471332     Length:6471332     Length:6471332
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    start_lat       start_lng         end_lat         end_lng        member_casual
##  Min.   :41.63   Min.   :-87.94   Min.   : 0.00   Min.   :-144.05   Length:6471332
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.: -87.66   Class :character
##  Median :41.90   Median :-87.64   Median :41.90   Median : -87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65   Mean   :41.90   Mean   : -87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.: -87.63
##  Max.   :42.07   Max.   :-87.46   Max.   :87.96   Max.   :   0.00
##                                   NA's   :8783    NA's   :8783
## [1] 6471332
## [1] 6471332      13
```

```
## # A tibble: 6 × 13
##   ride_id          rideable_type started_at     ended_at start_station_name start_station_id
##   <chr>            <chr>         <chr>          <chr>    <chr>              <chr>
## 1 903C30C2D810A53B electric_bike 2023-08-19 15... 2023-08... LaSalle St & Illi... 13430
## 2 F2FB18A98E110A2B electric_bike 2023-08-18 15... 2023-08... Clark St & Randol... TA1305000030
## 3 D0DEC7C94E4663DA electric_bike 2023-08-30 16... 2023-08... Clark St & Randol... TA1305000030
## 4 E0DDDC5F84747ED9 electric_bike 2023-08-30 16... 2023-08... Wells St & Elm St  KA1504000135
## 5 7797A4874BA260CA electric_bike 2023-08-22 15... 2023-08... Clark St & Randol... TA1305000030
## 6 DF4DE734EBC4DF66 electric_bike 2023-08-24 12... 2023-08... Milwaukee Ave & F... 428
## #  7 more variables: end_station_name <chr>, end_station_id <chr>, start_lat <dbl>,
## #   start_lng <dbl>, end_lat <dbl>, end_lng <dbl>, member_casual <chr>
```

Data Cleaning and Feature Engineering

```
## Columns: 12
## $ ride_id            <chr> "903C30C2D810A53B", "F2FB18A98E110A2B", "D0DEC7C94E4663DA", "E0...
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", "electric_bi...
## $ started_at         <dttm> 2023-08-19 15:41:53, 2023-08-18 15:30:18, 2023-08-30 16:15:08,...
## $ ended_at           <dttm> 2023-08-19 15:53:36, 2023-08-18 15:45:25, 2023-08-30 16:27:37,...
## $ start_station_name <chr> "LaSalle St & Illinois St", "Clark St & Randolph St", "Clark St...
## $ start_station_id   <chr> "13430", "TA1305000030", "TA1305000030", "KA1504000135", "TA130...
## $ end_station_name   <chr> "Clark St & Elm St", "", "", "", "", "", "", "", "", "", "", ""...
## $ end_station_id     <chr> "TA1307000039", "", "", "", "", "", "", "", "", "", "", "", "",...
## $ member_casual      <chr> "member", "member", "member", "member", "member", "member", "me...
## $ trip_duration      <Duration> 703s (~11.72 minutes), 907s (~15.12 minutes), 749s (~12.48...
## $ day_of_week        <ord> Saturday, Friday, Wednesday, Wednesday, Tuesday, Thursday, Thur...
## $ year_month         <date> 2023-08-01, 2023-08-01, 2023-08-01, 2023-08-01, 2023-08-01, 20...
##  Data Summary
##                           Values
## Name                      aggregated_tibble_clean
```

```
## Number of rows             6471332
## Number of columns          12
## _________________________
## Column type frequency:
##   character                7
##   Date                     1
##   factor                   1
##   POSIXct                  2
##   Timespan                 1
## _________________________
## Group variables            None
##
##  Variable type: character
##   skim_variable      n_missing complete_rate min max   empty n_unique whitespace
## 1 ride_id                    0             1  16  16       0  6471121          0
## 2 rideable_type              0             1  11  13       0        3          0
## 3 start_station_name         0             1   0  64 1087616     1739          0
## 4 start_station_id           0             1   0  14 1087616     1702          0
## 5 end_station_name           0             1   0  64 1131701     1749          0
## 6 end_station_id             0             1   0  36 1131701     1711          0
## 7 member_casual              0             1   6   6       0        2          0
##
##  Variable type: Date
##   skim_variable n_missing complete_rate min        max        median     n_unique
## 1 year_month            0             1 2023-08-01 2024-08-01 2024-04-01       13
##
##  Variable type: factor
##   skim_variable n_missing complete_rate ordered n_unique
## 1 day_of_week           0             1 TRUE           7
##   top_counts
## 1 Sat: 1031405, Wed: 981243, Thu: 945681, Fri: 942802
##
##  Variable type: POSIXct
##   skim_variable n_missing complete_rate min                 max
## 1 started_at            0             1 2023-08-01 00:00:06 2024-08-31 23:58:30
## 2 ended_at              0             1 2023-08-01 00:01:03 2024-08-31 23:59:53
##   median              n_unique
## 1 2024-04-01 14:43:21  5855514
## 2 2024-04-01 14:54:15  5862609
##
##  Variable type: Timespan
##   skim_variable n_missing complete_rate     min     max median n_unique
## 1 trip_duration         0             1 -999391 5909344   591.  1659213
```

Trip duration stats are questionable (presence of < 0 min durations and unreasonably high max duration)

```
## # A tibble: 2 × 2
##   row_name           high_durations
##   <chr>                       <int>
## 1 high_durations              16980
## 2 negative_durations            404
```

We will remove all negative durations and keep the unreasonably high trip durations (need more context to decide on what durations to keep)

Limitation: There are trips with unreasonably high trip durations ($>$ ~9 weeks). We Will need more context

to understand why and whether to remove some of the unreasonably high trip durations.

Now, let's find out how many empty strings are there in the start and end station ids and names

```
## 1 trip_duration         0             1 -999391 5909344   591.  1659213
## # A tibble: 4 × 2
##   row_name          empty_strings
##   <chr>                     <int>
## 1 empty_start_ids         1087616
## 2 empty_start_names       1087616
## 3 empty_end_id            1131701
## 4 empty_end_names         1131701
```

Limitation: there are many observations(rows) with empty ids and names for start stations and end stations. This would make it difficult for further analysis into stations that member or casual riders frequently use.

Analysis

For The analysis, we will be investigating a variety of questions:

1. How many members and casual riders are there for each month? *Limitation:* Cannot Get this info since personal identifying info is not provided

Year (08/2023 - 08/2024)

2. What is the total number of trips for each user category for the whole year?
3. What is the average duration of trips taken by each user category?
4. What rideable type do each user type prefer?

Month

5. What is the total monthly number of trips for each user category?
6. What is the average monthly trip duration for each user category?
7. What rideable type do each user type prefer?

Day

8. What is the total number of trips for each user category?
9. What is the average daily trip duration for each user category?
10. What rideable type do each user type prefer?

Let's perform some aggregate descriptive analysis on trip durations for each user type

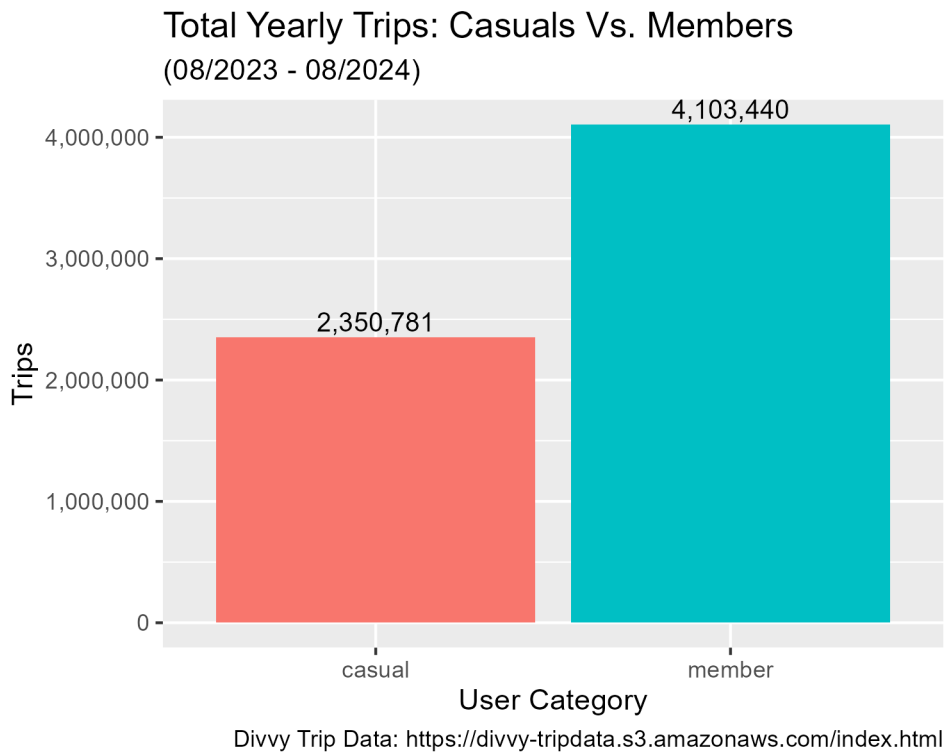
```
##   aggregated_tibble_clean$member_casual aggregated_tibble_clean$trip_duration
## 1                               casual                              1625.3893
## 2                               member                               784.5545
##   aggregated_tibble_clean$member_casual aggregated_tibble_clean$trip_duration
## 1                               casual                                740.244
## 2                               member                                529.000
##   aggregated_tibble_clean$member_casual aggregated_tibble_clean$trip_duration
## 1                               casual                                5909344
## 2                               member                                  93588
##   aggregated_tibble_clean$member_casual aggregated_tibble_clean$trip_duration
## 1                               casual                                     5
## 2                               member                                     5
##   aggregated_tibble_clean$day_of_week aggregated_tibble_clean$trip_duration
## 1                              Sunday                              1350.4556
```

```
## 2                                Monday                                1004.4757
## 3                               Tuesday                                 941.7203
## 4                             Wednesday                                 969.9080
## 5                              Thursday                                 966.9606
## 6                                Friday                                1083.8616
## 7                              Saturday                                1310.3171
##    aggregated_tibble_clean$member_casual aggregated_tibble_clean$day_of_week
## 1                                 casual                               Sunday
## 3                                 casual                               Monday
## 5                                 casual                              Tuesday
## 7                                 casual                            Wednesday
## 9                                 casual                             Thursday
## 11                                casual                               Friday
## 13                                casual                             Saturday
## 2                                 member                               Sunday
## 4                                 member                               Monday
## 6                                 member                              Tuesday
## 8                                 member                            Wednesday
## 10                                member                             Thursday
## 12                                member                               Friday
## 14                                member                             Saturday
##    aggregated_tibble_clean$trip_duration
## 1                              1901.7695
## 3                              1567.9483
## 5                              1403.1028
## 7                              1435.2602
## 9                              1440.6378
## 11                             1605.0010
## 13                             1795.9262
## 2                               877.7829
## 4                               745.9535
## 6                               751.6747
## 8                               764.0645
## 10                              750.3630
## 12                              770.9644
## 14                              866.2321
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                 <int>            <dbl>
##  1 casual        Sunday               392071            1902.
##  2 casual        Monday               255169            1568.
##  3 casual        Tuesday              263385            1403.
##  4 casual        Wednesday            300128            1435.
##  5 casual        Thursday             295966            1441.
##  6 casual        Friday               352793            1605.
##  7 casual        Saturday             491269            1796.
##  8 member        Sunday               457302             878.
##  9 member        Monday               556164             746.
## 10 member        Tuesday              639432             752.
## 11 member        Wednesday            678502             764.
## 12 member        Thursday             647248             750.
## 13 member        Friday               587587             771.
## 14 member        Saturday             537205             866.
```

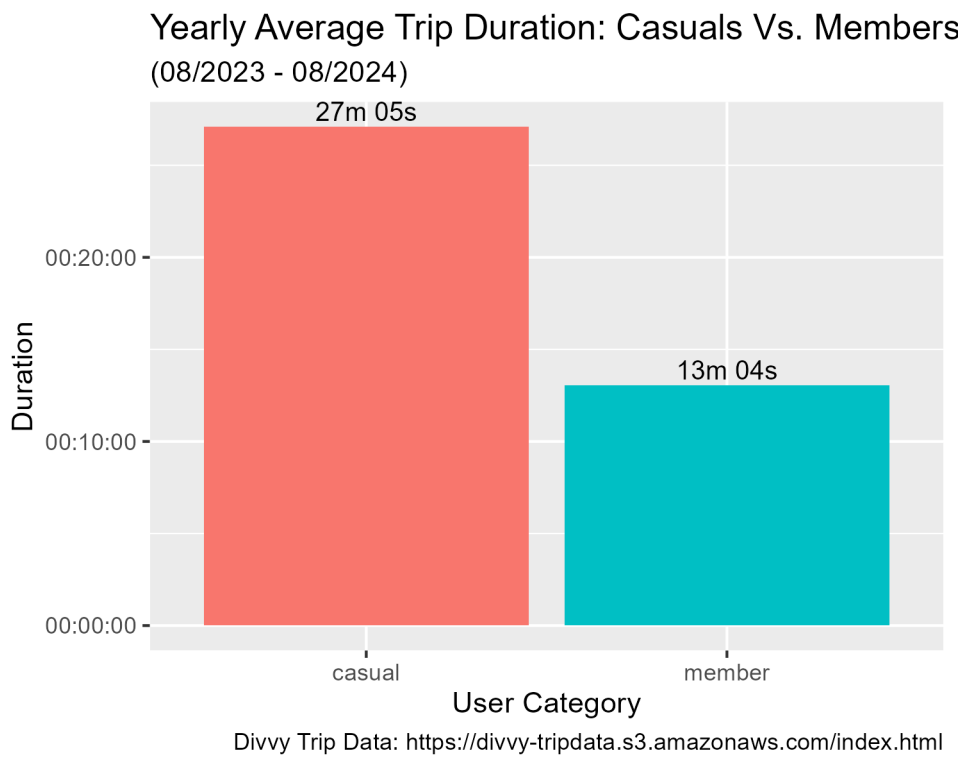
Plotting

Year

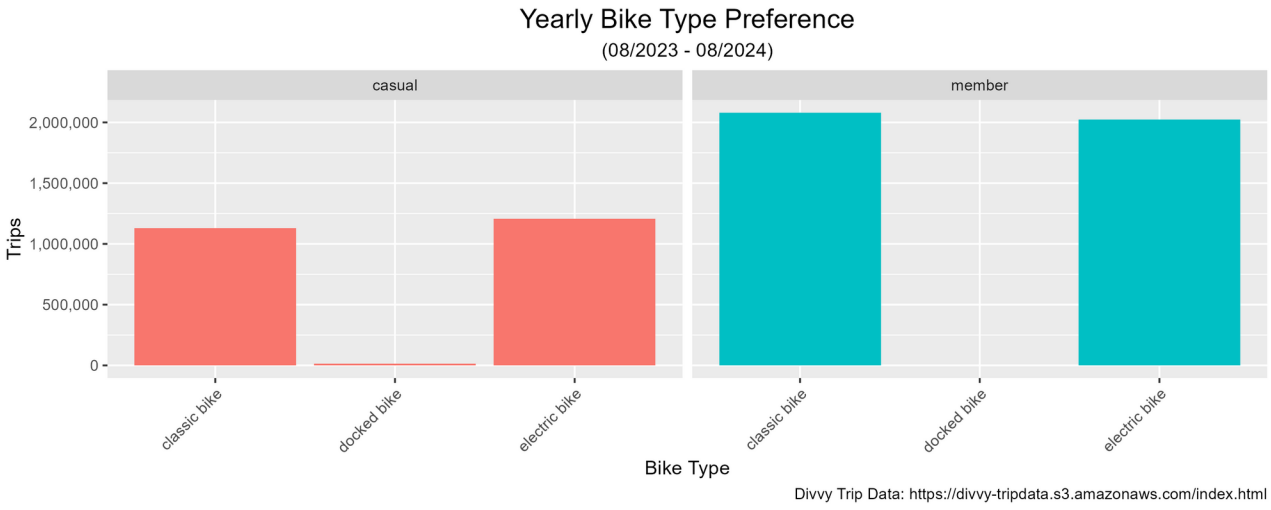
Total trips taken by each user category



Total Average Duration of trips taken by each user category

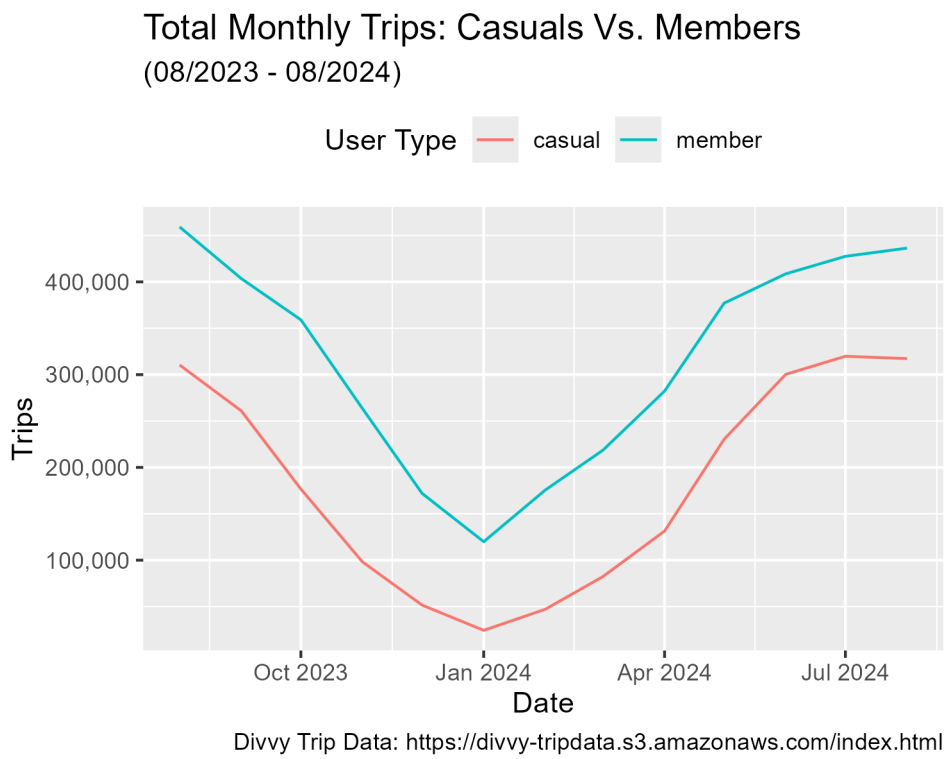


Yearly rideable type preference

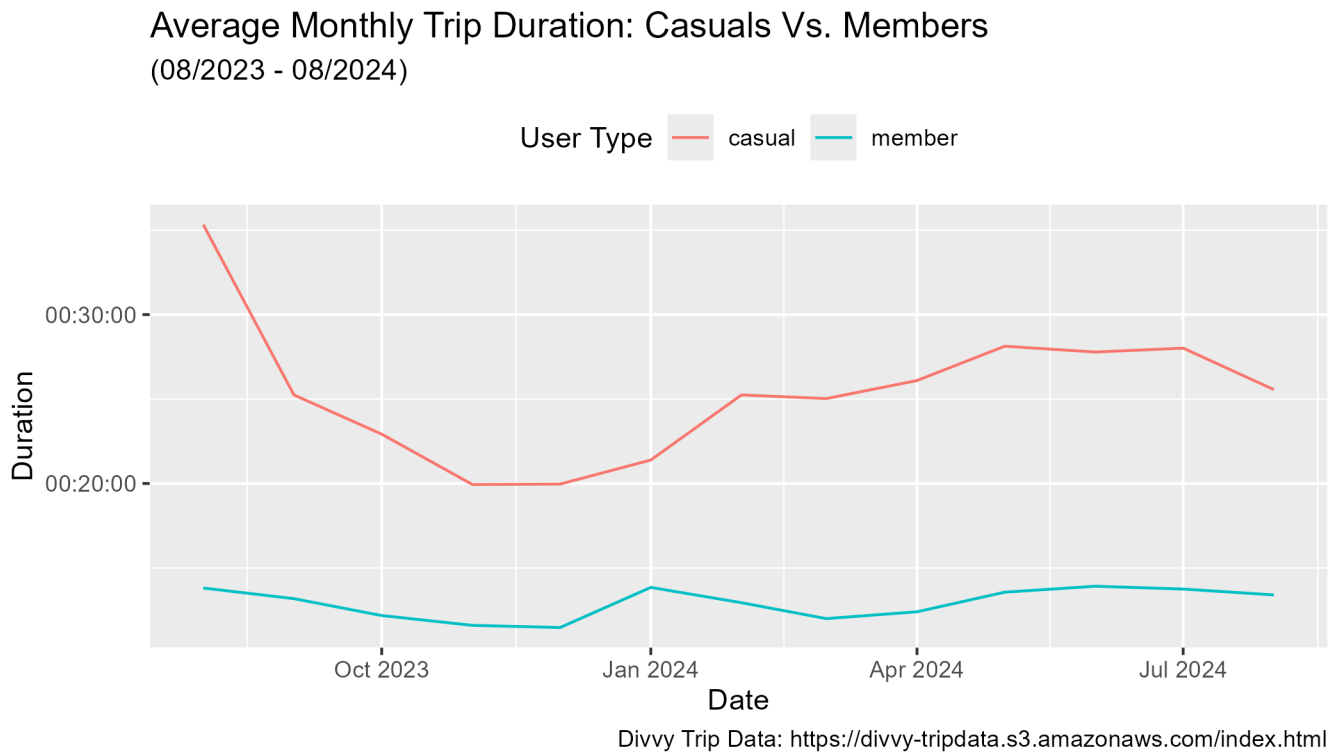


Month

Monthly total trips

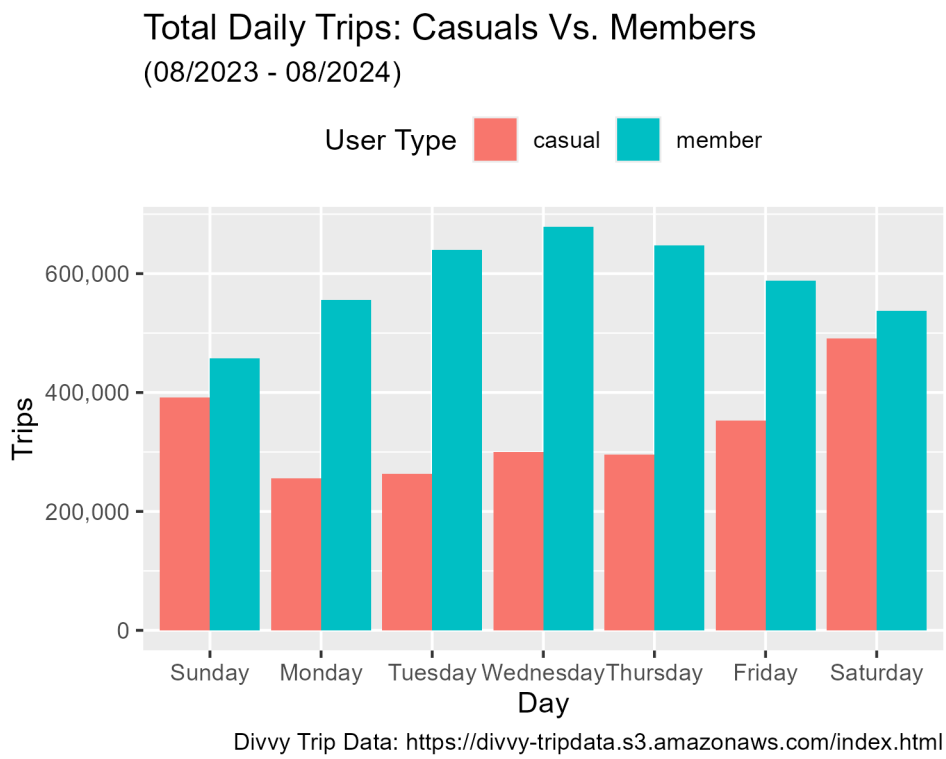


Average monthly trip duration

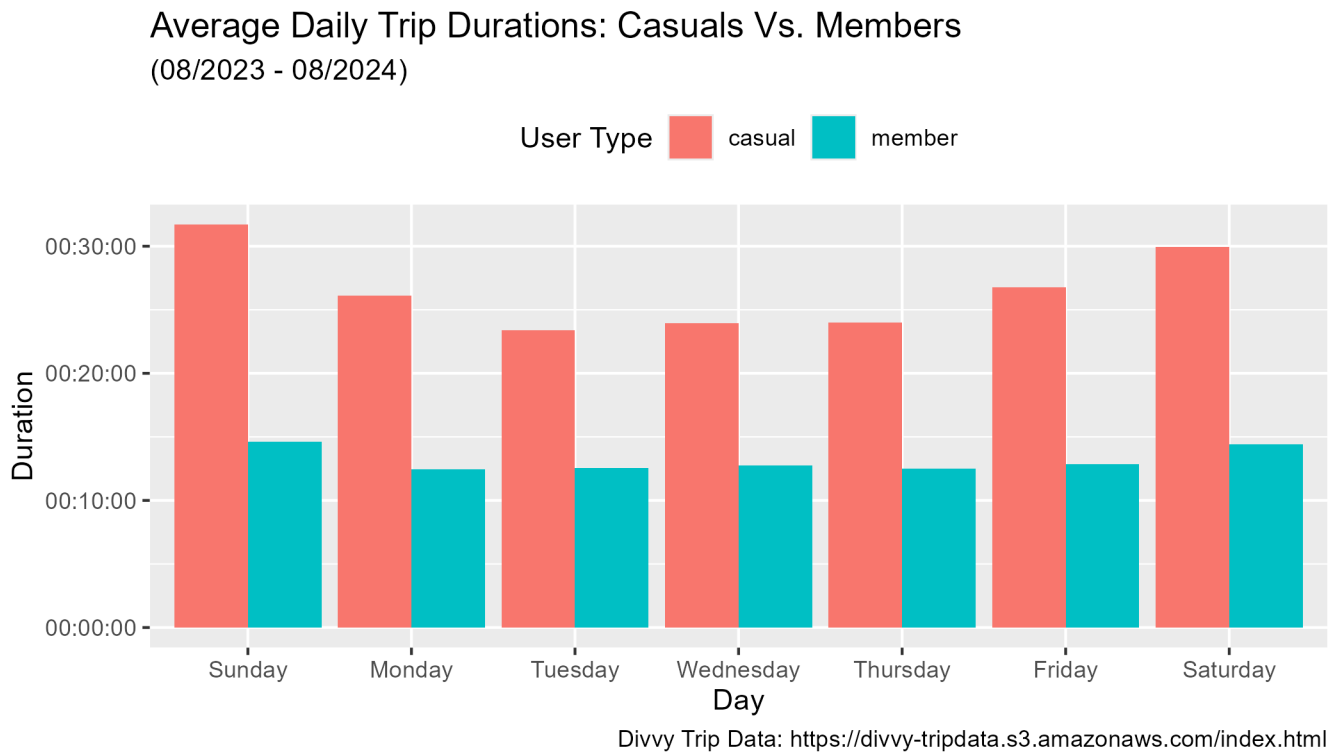


Monthly rideable type preference

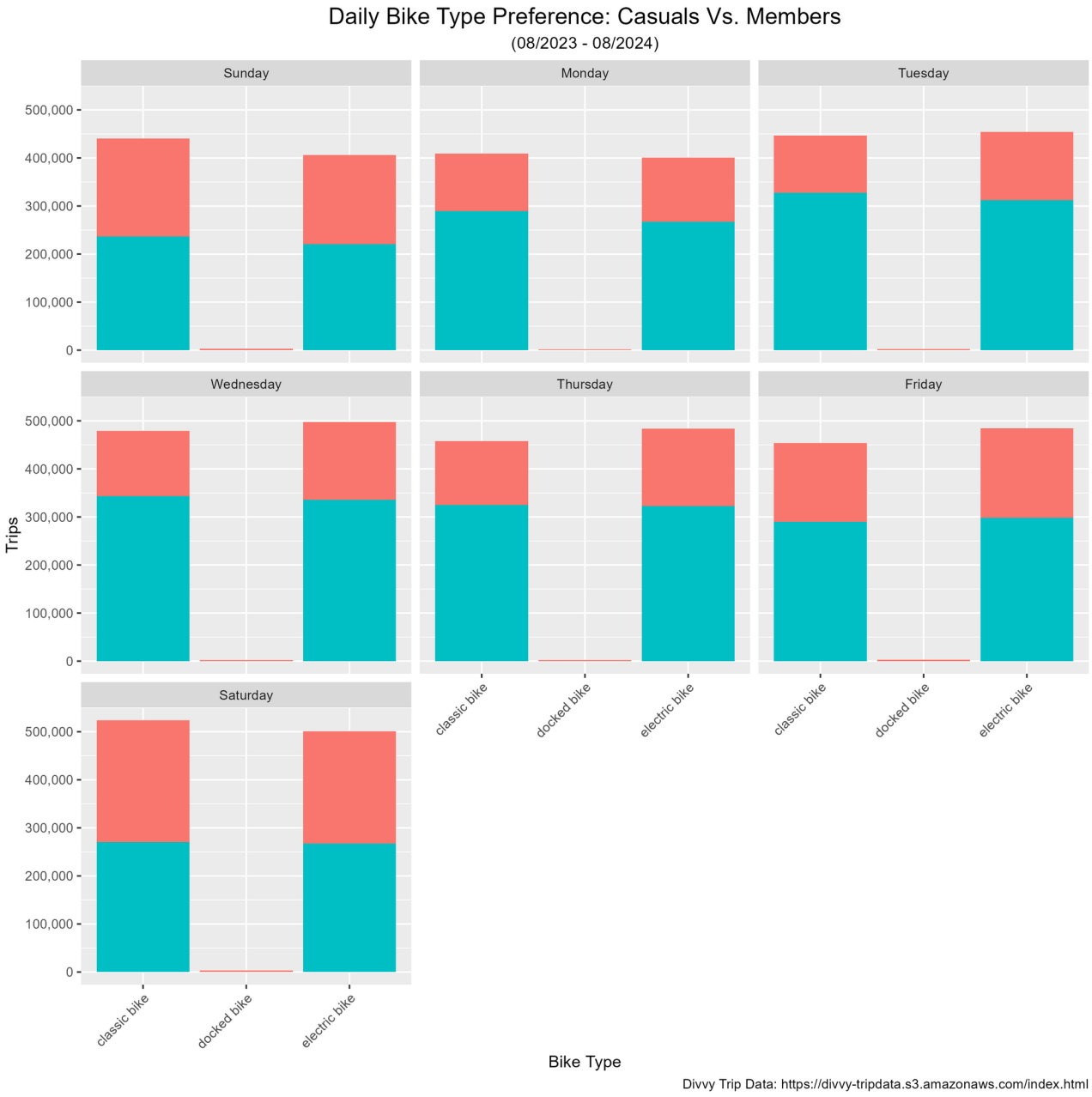
Day Total number of daily trips



Average daily trip durations



Daily Rideable type preference



Findings

On the **Yearly, Monthly and Daily basis**.

- Member riders take many more trips than the casual riders on the yearly, monthly and daily basis. (Year: members trips ~ 2x casual trips).
- Casual riders, on the other hand, typically ride the bikes for longer trips.

On a **Monthly Basis**

- On a monthly basis, casual riders have a bigger variation in their average trip durations compared to the members.
- Members as well as casual riders do not use the bikes as much during the winter season (Nov 2023 -

March 2024) with the lowest ride counts for both user types in the month of January 2024

Generally

- Casual riders slighlty preferred electric bikes over classic bikes while members preferred the opposite (based on the yearly trips done with both rideable type bikes for each user type).