Case study – Bike ride Data Analysis (Cyclistic)

Ask stage

Defining how casual riders and annual members use the bike share service differently in order to make recommendations to the marketing team.

Stakeholders are the exec team, marketing director and marketing analyst team

Prepare stage

The data has been downloaded from an online repository of a trustworthy source so data integrity is maintained. The files from the most recent 12 months from August 2022 to July 2023 were downloaded so the data is current. The files were named based on month that the data was collected and are organized based on the date in the file name. The dataset also has thousands of rows giving us a large enough sample size.

Thus we have good reliable data available but there are shortcomings such as we have no customer information available or how many trips have been done for each customer or the pricing data for the trips for analysis.

Process stage

Moving all files to a specific folder named 2022-2023 cyclistic data

The first file opened was the 202208-divvy-tripdata.

The rows that had any blank cells were filtered and then removed completely from the dataset in order to have a complete dataset.

The dataset was checked for duplicate rows and none were found

The dataset was also checked for blank spaces in all the columns but none were found.

The day of week column was computed using the started\_at column which tells the specific weekday on which the ride the started.

The ride length column was computed using started\_at and ended\_at columns which tells the amount of time taken for the ride to be completed

The cells in the ride length column with time as 0 min or with time in negative were highlighted and then removed from the dataset because they were not relevant to the analysis since the ride time cannot be so zero or in negative

This completes all the operations to be performed on the file as part of the process stage

Similarly all the operations mentioned above were performed for all the files in the 2022-2023 cyclistic data from 202208-divvy-tripdata to 2023-07-divvy-tripdata

Thus finally all the files relevant to the data analysis completed the process stage

Analysis stage

All the files were sorted based on the started\_at column in ascending order for analysis purpose.

The data formats of all the columns are appropriate in all files.

R is being used to perform analysis on the data so the software being used is RStudio

The data is first imported to RStudio using the read\_csv function. First 202208-divvy-tripdata is imported and checked to ensure that the import is happening correctly. Then all the files are imported one-by-one into RStudio.

Using bind\_rows function from the dplyr package,all the datasets were binded together in order to create one large dataset.

Since the data was already sorted based on started\_at column, it just had to be binded together based on month and year to create a fully sorted dataset based on the started\_at column.

The View() function can be used to view the dataset.

The summary() function can be used to analyse the dataset and find out various statistical measures such as minimum, maximum, first quartile, mean, median, third quartile and also the NA values for the columns with numeric column types

The character column types provide the length of the column(no of rows), class and mode for each of the columns.

The no of rows in the dataset were 4287202 after the binding of all files

The skim\_without\_charts() function is also used to analyse the dataset and provide various statistical measures.

It provides the total no of rows and columns and also the frequency of each datatype in columns

For the variable type columns, it provides the no of missing values, complete rate(how complete each column is), min, max, empty values, unique values and the no of whitespaces as well.

There are a total of 9 columns with variable type data. The dataset does not have any empty values or whitespaces and the columns are complete but there is a missing value in all columns.

The ride\_length column is the only column of difftime datatype and there 4 missing values but it is complete.

For difftime datatype we have the missing values, complete rate, min, max, median, unique values as the statistical measures available.

There are 5 columns which are of type numeric and again there is a missing value for all of them and are complete.

For numeric datatype, the measures available are missing values, complete rate, mean, standard deviation, p0(min), p25(first quartile), p50(median), p75(third quartile), p100(max).

The ride\_length column is converted to numeric for analysis purpose and all the times are basically shown in seconds.

The years, month, day, hour, minute columns are computed using the started\_at column for analysis later on.

The ride count for casuals and members

# A tibble: 2 × 2

member\_casual ride\_count

*<chr>* *<int>*

1 casual 1593202

2 member 2693995

The ride count for members is substantially higher than for casuals over the 12 month period.

The ride count for members and casuals across months

# A tibble: 24 × 3

# Groups: member\_casual [2]

member\_casual month ride\_count

*<chr>* *<chr>* *<int>*

1 casual 01 29277

2 casual 02 32408

3 casual 03 46160

4 casual 04 109157

5 casual 05 175087

6 casual 06 217587

7 casual 07 242909

8 casual 08 267648

9 casual 09 218986

10 casual 10 149914

11 casual 11 72857

12 casual 12 31212

13 member 01 116723

14 member 02 114897

15 member 03 150612

16 member 04 209412

17 member 05 281939

18 member 06 311312

19 member 07 324735

20 member 08 331382

21 member 09 310639

22 member 10 259720

23 member 11 180108

24 member 12 102516

The months of April, May, June, July, August, September, October witness much higher ride counts compared to the rest of the months. Also the overall increase is higher for casuals compared to members but the overall numbers for members are higher compared to casuals.

The ride count for casuals and members across the seven days of a week

# A tibble: 14 × 3

# Groups: member\_casual [2]

member\_casual day\_of\_week ride\_count

*<chr>* *<dbl>* *<int>*

1 casual 1 248433

2 casual 2 189983

3 casual 3 186631

4 casual 4 191029

5 casual 5 210305

6 casual 6 241870

7 casual 7 324951

8 member 1 288175

9 member 2 387977

10 member 3 426518

11 member 4 433721

12 member 5 433531

13 member 6 385940

14 member 7 338133

For casuals, the ride count is increasing on the weekend and then falls as the week starts. But for members, the ride count is low on Monday and then increases till Friday and then falls again on the weekend.

The ride count for members and casuals across the hours of a day

# A tibble: 48 × 3

# Groups: member\_casual [2]

member\_casual hours ride\_count

*<chr>* *<int>* *<int>*

1 casual 0 28220

2 casual 1 17894

3 casual 2 10440

4 casual 3 5511

5 casual 4 3792

6 casual 5 8242

7 casual 6 23165

8 casual 7 39304

9 casual 8 53619

10 casual 9 53642

11 casual 10 67773

12 casual 11 86686

13 casual 12 102196

14 casual 13 105635

15 casual 14 111867

16 casual 15 124156

17 casual 16 142414

18 casual 17 156727

19 casual 18 133864

20 casual 19 99101

21 casual 20 70255

22 casual 21 59058

23 casual 22 51634

24 casual 23 38007

25 member 0 24062

26 member 1 14663

27 member 2 7994

28 member 3 4951

29 member 4 5836

30 member 5 25503

31 member 6 81293

32 member 7 150714

33 member 8 183381

34 member 9 119988

35 member 10 108492

36 member 11 127928

37 member 12 146576

38 member 13 144990

39 member 14 145900

40 member 15 180471

41 member 16 245794

42 member 17 291680

43 member 18 227602

44 member 19 161042

45 member 20 110173

46 member 21 84546

47 member 22 61623

48 member 23 38793

For casuals, the hours 11 to 19 have a higher ride count in a day. But for members, the hours 7 to 20 have higher ride counts.

The no of rides across different types of bikes for casual and member riders

# A tibble: 5 × 3

# Groups: member\_casual [2]

member\_casual rideable\_type ride\_count

*<chr>* *<chr>* *<int>*

1 casual classic\_bike 782165

2 casual docked\_bike 125336

3 casual electric\_bike 685701

4 member classic\_bike 1676113

5 member electric\_bike 1017882

The casuals like the classic and electric bikes equally and the docked bikes are used to a lesser extent by them. The members like the classic more than the electric ones but they don’t seem to use the docked bikes.

The no of rides across different bikes across the seven days of a week

# A tibble: 35 × 4

# Groups: member\_casual, rideable\_type [5]

member\_casual rideable\_type day\_of\_week ride\_count

*<chr>* *<chr>* *<dbl>* *<int>*

1 casual classic\_bike 1 128681

2 casual classic\_bike 2 90815

3 casual classic\_bike 3 88201

4 casual classic\_bike 4 88109

5 casual classic\_bike 5 99091

6 casual classic\_bike 6 116555

7 casual classic\_bike 7 170713

8 casual docked\_bike 1 22524

9 casual docked\_bike 2 15155

10 casual docked\_bike 3 13351

11 casual docked\_bike 4 12919

12 casual docked\_bike 5 14377

13 casual docked\_bike 6 19007

14 casual docked\_bike 7 28003

15 casual electric\_bike 1 97228

16 casual electric\_bike 2 84013

17 casual electric\_bike 3 85079

18 casual electric\_bike 4 90001

19 casual electric\_bike 5 96837

20 casual electric\_bike 6 106308

21 casual electric\_bike 7 126235

22 member classic\_bike 1 184952

23 member classic\_bike 2 243037

24 member classic\_bike 3 264483

25 member classic\_bike 4 265060

26 member classic\_bike 5 264807

27 member classic\_bike 6 237472

28 member classic\_bike 7 216302

29 member electric\_bike 1 103223

30 member electric\_bike 2 144940

31 member electric\_bike 3 162035

32 member electric\_bike 4 168661

33 member electric\_bike 5 168724

34 member electric\_bike 6 148468

35 member electric\_bike 7 121831

For casuals, the ride count is increasing on the weekends consistently for all electric, docked and classic and then decreasing on weekdays.

For members, the ride count is increasing and staying consistent on weekdays for both electric and classic bikes and then decreases on the weekend.

The top 10 starting stations according to ride count

| **Start\_station** | | **Frequency** | |
| --- | --- | --- | --- |
|  |  | |  |
| **1** | Streeter Dr & Grand Ave | | 64402 |
| **2** | DuSable Lake Shore Dr & North Blvd | | 36514 |
| **3** | Michigan Ave & Oak St | | 35870 |
| **4** | DuSable Lake Shore Dr & Monroe St | | 35776 |
| **5** | Wells St & Concord Ln | | 32411 |
| **6** | Clark St & Elm St | | 31877 |
| **7** | Millennium Park | | 30870 |
| **8** | Kingsbury St & Kinzie St | | 30591 |
| **9** | Theater on the Lake | | 29226 |
| **10** | Wells St & Elm St | | 27407 |

The top 10 ending stations according to ride count

| **End\_station** | | **Frequency** | |
| --- | --- | --- | --- |
|  |  | |  |
| **1** | Streeter Dr & Grand Ave | | 62589 |
| **2** | DuSable Lake Shore Dr & Monroe St | | 36984 |
| **3** | Michigan Ave & Oak St | | 34836 |
| **4** | DuSable Lake Shore Dr & North Blvd | | 33553 |
| **5** | Clark St & Elm St | | 32458 |
| **6** | Wells St & Concord Ln | | 31752 |
| **7** | Kingsbury St & Kinzie St | | 31219 |
| **8** | Millennium Park | | 29708 |
| **9** | Theater on the Lake | | 28481 |
| **10** | Wells St & Elm St | | 27088 |

The top 10 stations for both starting and ending stations based on ride count are the same with just some ranking positions being different.

The ride count across different bikes across months for members and casuals

# A tibble: 60 × 4

# Groups: member\_casual, rideable\_type [5]

member\_casual rideable\_type month ride\_count

*<chr>* *<chr>* *<chr>* *<int>*

1 casual classic\_bike 01 13772

2 casual classic\_bike 02 15348

3 casual classic\_bike 03 19236

4 casual classic\_bike 04 48326

5 casual classic\_bike 05 91503

6 casual classic\_bike 06 115213

7 casual classic\_bike 07 141176

8 casual classic\_bike 08 127236

9 casual classic\_bike 09 104231

10 casual classic\_bike 10 60899

11 casual classic\_bike 11 32710

12 casual classic\_bike 12 12515

13 casual docked\_bike 01 1675

14 casual docked\_bike 02 2139

15 casual docked\_bike 03 2932

16 casual docked\_bike 04 8608

17 casual docked\_bike 05 12738

18 casual docked\_bike 06 14471

19 casual docked\_bike 07 17735

20 casual docked\_bike 08 25735

21 casual docked\_bike 09 19354

22 casual docked\_bike 10 12316

23 casual docked\_bike 11 5769

24 casual docked\_bike 12 1864

25 casual electric\_bike 01 13830

26 casual electric\_bike 02 14921

27 casual electric\_bike 03 23992

28 casual electric\_bike 04 52223

29 casual electric\_bike 05 70846

30 casual electric\_bike 06 87903

31 casual electric\_bike 07 83998

32 casual electric\_bike 08 114677

33 casual electric\_bike 09 95401

34 casual electric\_bike 10 76699

35 casual electric\_bike 11 34378

36 casual electric\_bike 12 16833

37 member classic\_bike 01 75645

38 member classic\_bike 02 73552

39 member classic\_bike 03 86839

40 member classic\_bike 04 119629

41 member classic\_bike 05 175436

42 member classic\_bike 06 194725

43 member classic\_bike 07 216803

44 member classic\_bike 08 213342

45 member classic\_bike 09 198896

46 member classic\_bike 10 150560

47 member classic\_bike 11 110565

48 member classic\_bike 12 60121

49 member electric\_bike 01 41078

50 member electric\_bike 02 41345

51 member electric\_bike 03 63773

52 member electric\_bike 04 89783

53 member electric\_bike 05 106503

54 member electric\_bike 06 116587

55 member electric\_bike 07 107932

56 member electric\_bike 08 118040

57 member electric\_bike 09 111743

58 member electric\_bike 10 109160

59 member electric\_bike 11 69543

60 member electric\_bike 12 42395

For casuals, the classic, docked and electric are all used at a higher rate from April to October and then ride count dips from November to March.

The members also have a higher rate from April to October for classic and electric bikes and then it dips from November to March.

The total ride time for casual and members across seven days of a week

# A tibble: 14 × 3

# Groups: member\_casual [2]

member\_casual day\_of\_week ride\_time

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

1 casual 1 385756560

2 casual 2 255146100

3 casual 3 229412580

4 casual 4 221644620

5 casual 5 251898660

6 casual 6 320215200

7 casual 7 497649180

8 member 1 235103880

9 member 2 273156240

10 member 3 300592440

11 member 4 304234140

12 member 5 306279780

13 member 6 280962660

14 member 7 279454080

The ride time for casuals is higher as the weekend arrives and then starts dropping as the weekdays begin. But the ride time is overall around the same across the week for members.

The total ride time across months for members and casuals

# A tibble: 24 × 3

# Groups: member\_casual [2]

member\_casual month ride\_time

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

1 casual 01 26445840

2 casual 02 34748220

3 casual 03 46943160

4 casual 04 150026100

5 casual 05 259776840

6 casual 06 316805040

7 casual 07 371131500

8 casual 08 377322420

9 casual 09 288998940

10 casual 10 185362920

11 casual 11 76102800

12 casual 12 28059120

13 member 01 71229180

14 member 02 73035240

15 member 03 93774840

16 member 04 148139700

17 member 05 217953300

18 member 06 244635240

19 member 07 263262180

20 member 08 263610900

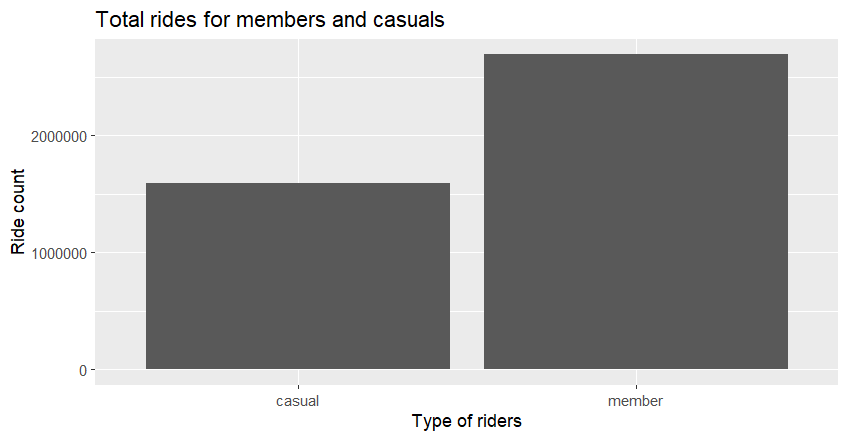
21 member 09 237991620

22 member 10 184176900

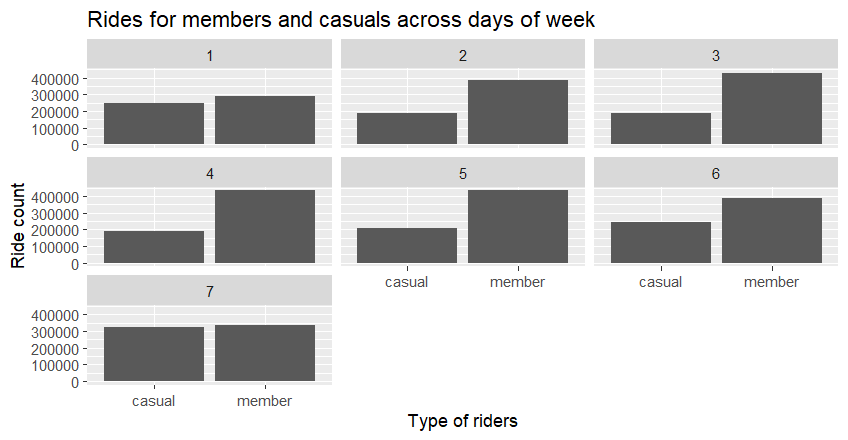
23 member 11 118388580

24 member 12 63585540

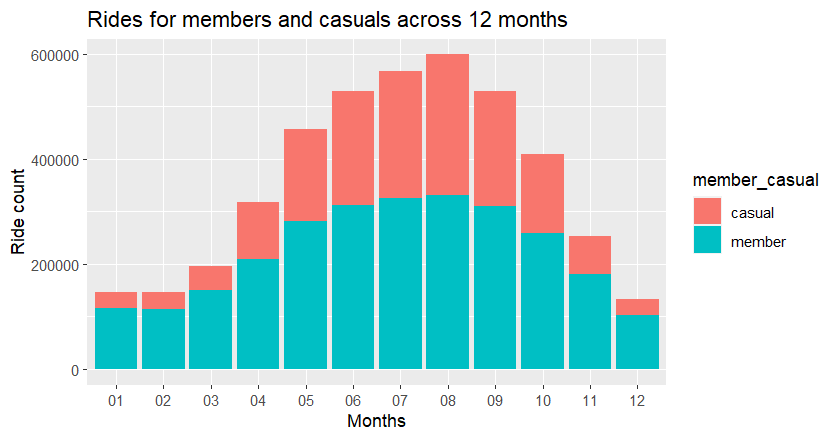
The ride time for both casuals and members is overall higher in April, May, June, July, August, September and October compared to the other months.

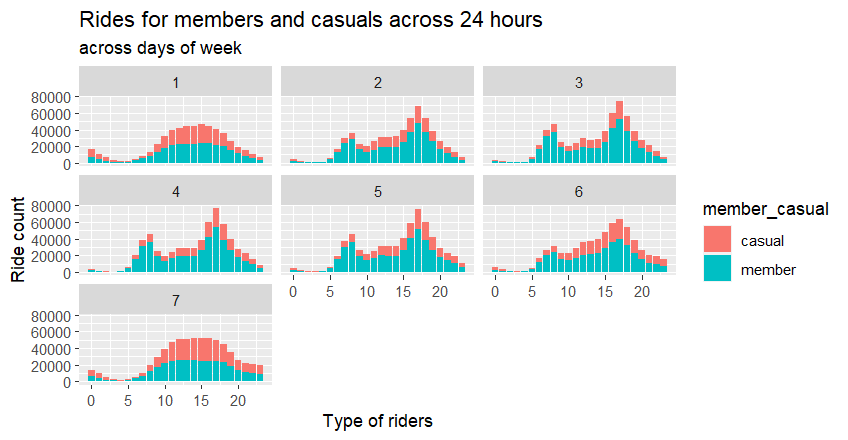
Share stage

The total rides members are much higher than casuals over the last 12 months. Thus members are using the rides more often than casuals. Now lets look more deeply into how the ride count differs over days of a week and months in a year.

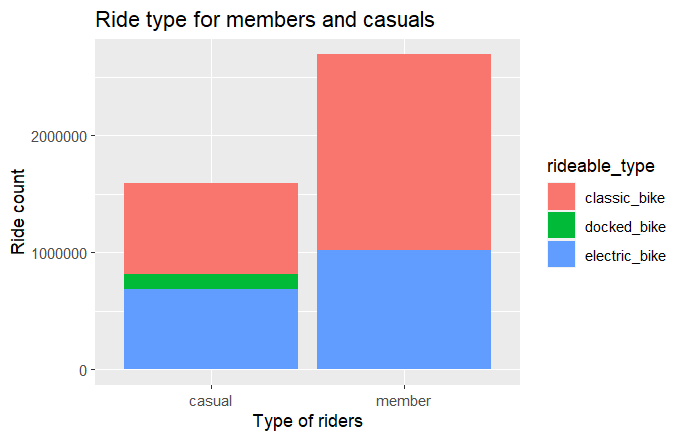


The trend that is clearly visible is that casuals prefer to take rides more on the weekend while members prefer to take rides more on the weekdays even if the number of casuals is lesser than members.

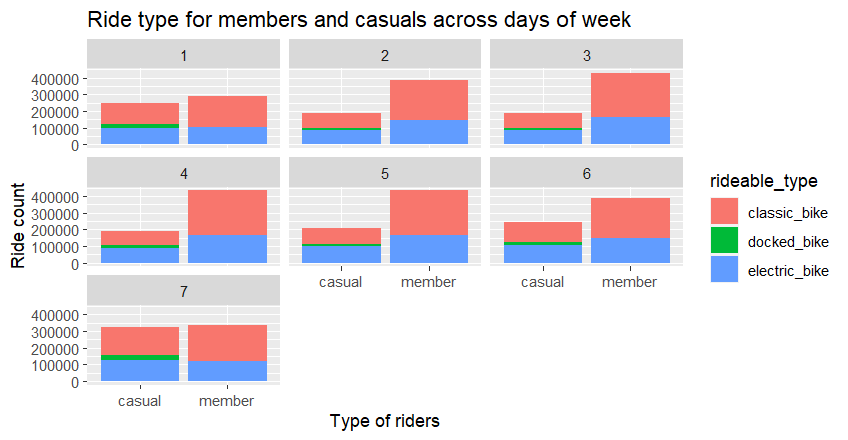
The ride counts increase significantly from April to October (Months 4 to 10) for both casuals and members and decreases for both in the remaining months. But the increase in casuals is comparatively higher than members.



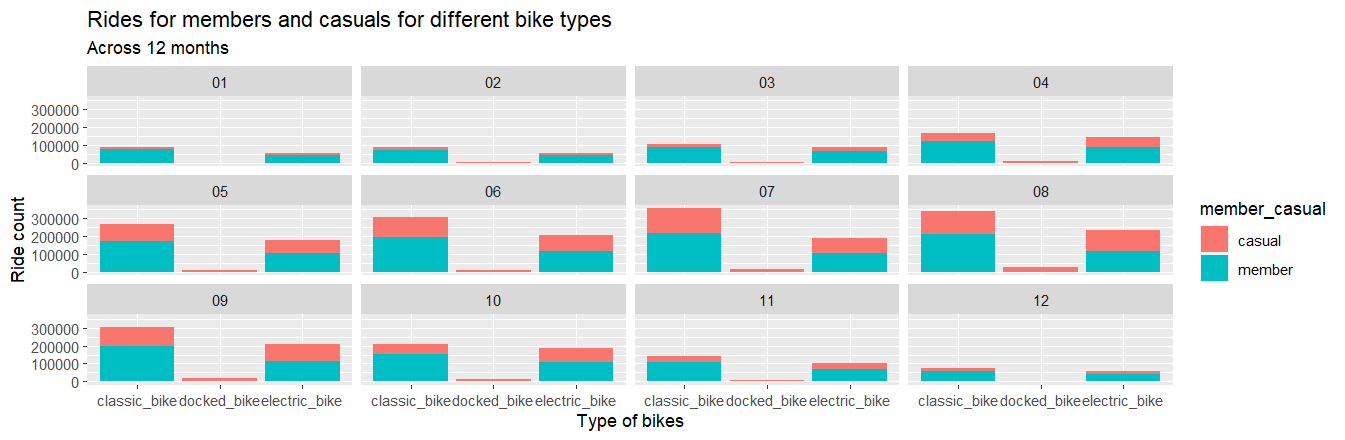
Casuals mostly use the rides from hours 11 to 19 and mostly on weekends while for members, they use it from hours 7 to 20 and mostly on weekdays.

Now lets see how the different types of bikes are used by members and casuals

Casuals prefer electric and classic bikes equally and a small percentage prefer the docked ones. The members prefer classic more than electric bikes overall but they don’t use the docked ones.



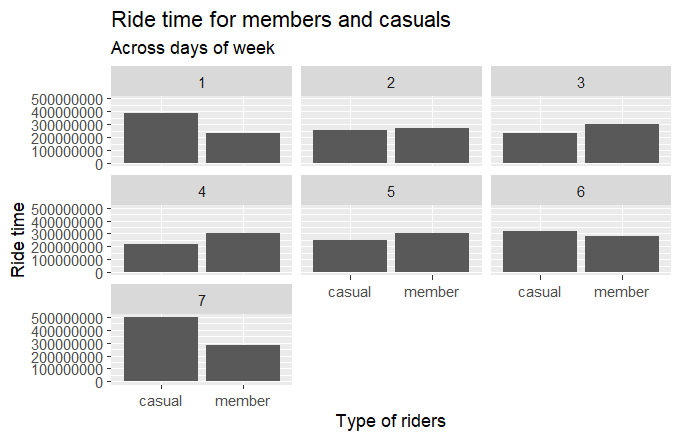
The casuals again have a similar usage for both classic and electric bikes which increases as the weekend comes. The members again have a higher usage for classics than electric but the difference is less on the weekend.



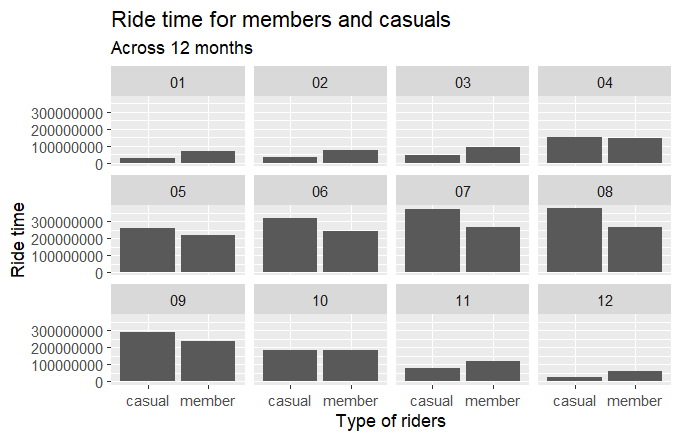
For casuals, the months of May to October have a higher ride usage on all 3 types of bikes.

For members, the months of May to October also have a higher ride usage on classic and electric bikes.

But for the remaining months of November to April, members have a higher overall usage than casuals on classic and electric bikes. The docked bikes are only used by casual riders.

Now lets see how the ride time differs for members and casuals

The ride time is significantly higher over the weekend for casuals compared to the weekdays. For the members, the ride time is generally consistent across all days but slightly higher on the weekdays.



The ride time for both casuals and members is significantly high from May to October but is more for casuals compared to members. But in the remaining months, even though the ride time is low for both, it is actually more for members compared to casuals.

Act stage

Now based on the recommendations being done based on the analysis, the executive team and the marketing team will take appropriate steps to solve the business task.

Recommendations

1. The top 10 stations with the highest ride count can have advertisements in their vicinity showing the benefits of becoming a member.

2. The months of April to October generally have a higher ride count so the marketing strategies can be focused on those months to maximize the impact

3. Casuals usually ride bikes on weekends so they most probably use it for leisure activites while members generally ride on weekdays so they use it while commuting to work or college. Thus marketing aimed at casual riders on the weekend can influence them to opt for bike rides on weekdays as well when they commute to work or college.

4. Casuals use the ride service mostly from 11 AM to 7 PM, so that time period should be targeted for marketing initiatives.