Cyclistic Project Report

Diego Sanoja

2022-05-28

Google Certificate Case Study

Objective of the project

Determine how casual riders and annual members used Cyclistic bikes differently during a one year period to develop a marketing strategy whose goal is to turn casual riders into annual members.

Data Used

For this project, 12 csv files which contained data collected from the Cyclistic customers from the beginning of May, 2021 to the end of April, 2022 were downloaded from a public dataset whose URL direction is https://divvy-tripdata.s3.amazonaws.com/index.html.

Each csv file contains information of the use of a Cyclistic bike like the bike id, the type of bike rented, the starting and ending time of the lease, the station where the bike was rented and where it was returned, and finally the customer type of a particular month.

The data has some limitations: it only provided information from the last 12 months, it didn't provide the prices to rent the bikes for each type of customer nor season discounts, only the observations where the start and end stations names are given were used since some of the trips where there were no stations mentioned were only tests to evaluate the quality of the bikes.

Changelog

Version 1.0.0 (05-17-2022)

The following applies to the 12 csv files used in this project.

New

- Created a folder for the project files, this folder is called Capstone_Project.
- Saved the original file in a sub folder called original_files.
- Added a new column called ride_length that contains the amount of time that the bike was used before being returned in a time format that shows the hours, minutes and seconds used.
- Added a new column called day_of_week whose value represents the day of the week that the bike was rented using the WEEKDAY function (the values go from 1 to 7 where 1 presents the Sunday and the 7 represents the Saturday).
- Saved the file with the cleaned data in a sub folder called cleaned_files.

Changes

- Removed the rows where the start_station_name and end_station_name were both blank.
- Sorted the rows by the start_station_time column in ascending order.
- Replaced the original columns ride_id, start_station_name and end_station_name by the values generated using the TRIM function in those columns.
- Changed the alignment of the header row to the left and the rest of rows to the right.

Fixes

- Fixed the format of the started_at and ended_at columns by adding the second when the bike was rented and returned.

Version 1.1.0 (05-18-2022)

The following applies to the 12 excel files created in the previous version.

New

- Added a new sheet called Pivot_table containing 6 pivots tables. The tables are the following:

a. Table 1 contains values of the type_of_customer variable in the rows and 3 columns which show the average value of the ride_length variable, the maximum value of the ride_length variable and number of subjects in each group of the type_of_customer variable.

b. Table 2 contains values of the type_of_customer variable in the rows and the columns are the values of the day_of_week variable. The values in the cells represent the average value of the ride_length each day of the week.

c. Table 3 contains values of the type_of_customer variable in the rows and the columns are the values of the day_of_week variable. The values in the cells represent the number of persons from each group of the type_of_customer that rented a bike in that particular day during the month.

d. Table 4 contains values of the type_of_customer variable in the rows and the columns are the values of the rideable_type variable. The values in the cells represent the number of persons from each group of the type_of_customer that rented 1 of the 3 types of bikes in the rideable_type. Note: if the cell in this table is empty it means that the bike was not rented.

e. Table 5 contains values of the start_station_name variable in the rows and the columns are the values of the type_of_customer variable. The values in the cells represent the number of persons from each group of the type_of_customer that rented a bike in that station. Note: if the cell in this table is empty it means that nobody in the subgroup rented a bike in that station.

f. Table 6 contains values of the end_station_name variable in the rows and the columns are the values of the type_of_customer variable. The values in the cells represent the number of persons from each group of the type_of_customer that returned a bike in that station. Note: if the cell in this table is empty it means that nobody in the subgroup returned a bike in that station.

- Added a sheet called Summary which has a table containing answers to the following questions for each value in the type_of_customer variable and other important information: Number of people in the group, Mean of ride_length, Max ride_length, Day with highest rent, Day with lowest rent, Most used bike, How many times the bike above was used?, Was there a bike that wasn't used? If yes which?, Most common start station, How many times the start station name above appears?, Most common end station and How many times the end station above appears?
- Saved the file with the same as it originally was in the same subfolder.

Changes

- Changed the name of column O from member_casual to type_of_customer.
- Changed the format of the ride_length column from Time to Custom to display the number of hours of each trip. The same was done to the cells in the Pivot_table sheet containing values of time to see the maximum values of the ride_length column for the Summary sheet.
- Added a border to the table in the Summary sheet.

Version 1.2.0 (05-19-2022)

New

- Created an RStudio document where to proceed the analysis of the 12 excel files. The name of this file is Project_analysis.

Changes

- Added 2 columns to each of the excel files, the first column is called ride_length(minutes) which contains the values of the ride_length column in minutes. The second column is called month which states the month where each trip was done.
- Switched the tool of analysis from Excel to R to combine the 12 excel files and study them all together.

Fixes

- Fixed the name of the column containing the function WEEKDAY() in the excel files of June and August by changing their name from date_of_week to day_of_week.
- Removed the rows of each excel file that had a negative value in the column ride_length(minutes).

Version 1.3.0 (05-20-2022)

New

- Created a data frame containing all the data from the 12 excel file that contained the cleaned data. The name of this data frame is full_year_frame.
- Created a final data frame for analysis called full_year_frame_v5.
- Created 4 sub data frames from full_year_frame_v5. The names of those data frames are: max_frame, mean_frame, mean_frame_short, and long_trips

Changes

- Modified the full_year_frame by adding to it 5 new columns and removing 3 columns.
- The columns removed from the frame are ride_length(minutes), rideable_type and type_of_customer.
- The name of the first 2 columns added are Year and Day which contain the year and day each customer used a bike. The other 3 columns are trip_length_in_minutes, type_of_bike_used and type_of_customer which are modified versions of the columns ride_length(minutes), rideable_type and type_of_customer. The name of the data frame with those modifications is full_year_frame_v5.

Fixes

- Changed the strings in the type_of_customer column to 'Member' and 'Casual'.
- Changed the strings in the rideable_type column to 'Classic', 'Electric' and 'Docked'.

- Removed the rows where the start_station_name column had NA as a value. Same for the rows where the end_station_name column had NA as a value.

Analysis summary

For the analysis phase, the R programming language tool was used. The reason to use R for the analysis phase was due to its ability to process larger amounts of data faster than excel and the had packages with functions and codes that can create effective data visuals.

After cleaning the data of the 12 excel files, they were combined in a single data frame to study all the cleaned data together. The name of this data frame was full_year_frame.

By making some modifications to this data frame, questions regarding how the casual riders and customers with annual memberships use Cyclistic bikes differently could be answered.

The first modification to this frame was adding 2 columns to the data frame using vectors created by "for loops" using the columns day_of_week and Month, the first column was named Day which contains the day of the week when the bike was rented, the second one was named Year which contains the year when the rent was made. The modified data frame was called full_year_frame_v2.

Next, the entries where either the name of the station where the bike was rented or the name of the station where the bike was returned was not given were removed. The data frame created by this change is called full_year_frame_v3.

Additionally, the entries of the Day column were altered so that when they were plotted in a graph, the first value was Monday, then Tuesday, etc. A similar alteration was done to the entries of the column Month so that the first value was May, then June and so on.

Then, the column type_of_customer was modified by changing the first letter of each entry to a capital letter and changed the name of the column containing the duration of the rent from ride_length(minutes) to trip_length_in_minutes. The result of this was the data frame called full_year_frame_v4.

Finally, the column rideable_type was altered by changing the name of the column from rideable_type to type_of_bike_used and deleting the _bike piece from each entry and making the first letter capital. The final result of this was full_year_frame_v5.

The full_year_frame_v5 was used to gather valuable insights from the data. This data frame contained 710699 observations/rows and 19 columns.

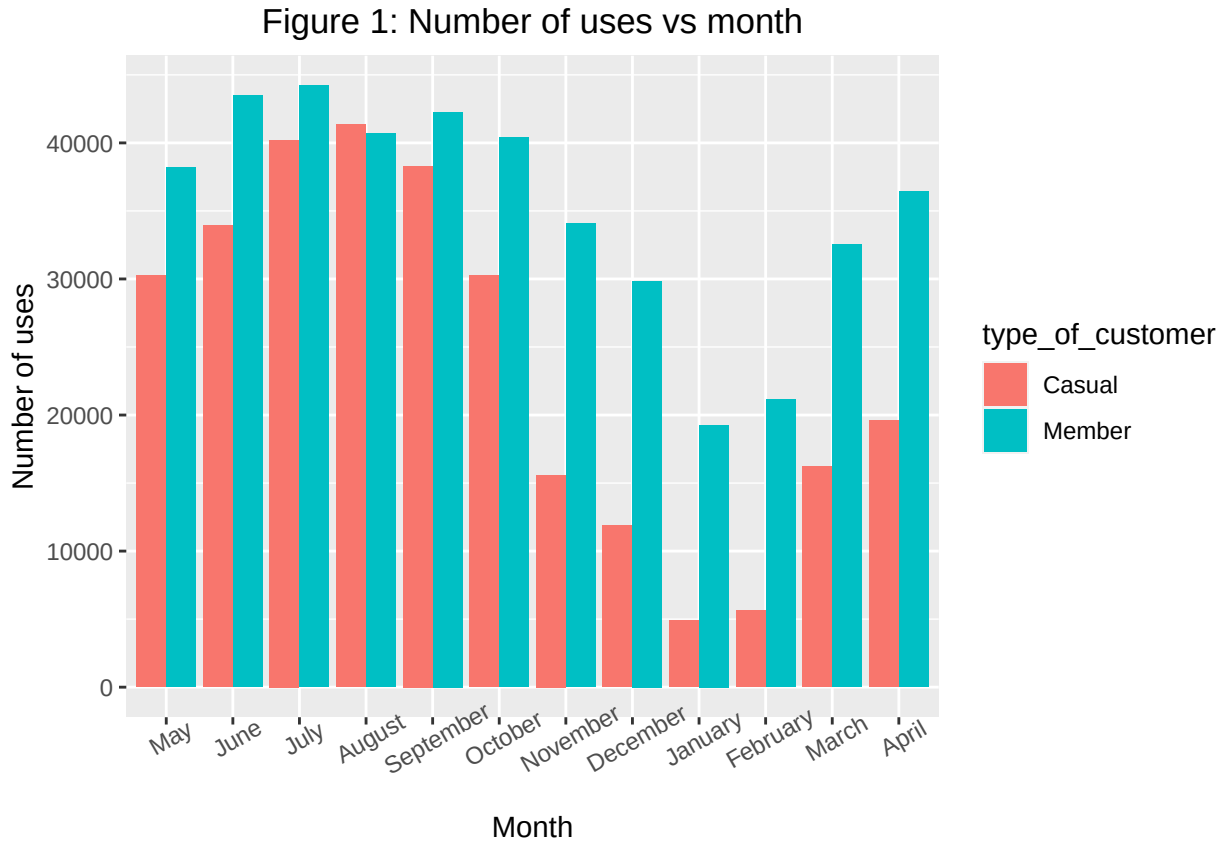


Figure 1 is a bar chart containing the number of bikes used each month on the y-axis and the month on the x-axis. To compare how many times casual riders and annual members used bikes each month, the number of entries of each month was separated by the type of customer. Red bars represent casual riders and the blue ones represent annual members.

By looking at this graph, the following insights were found:

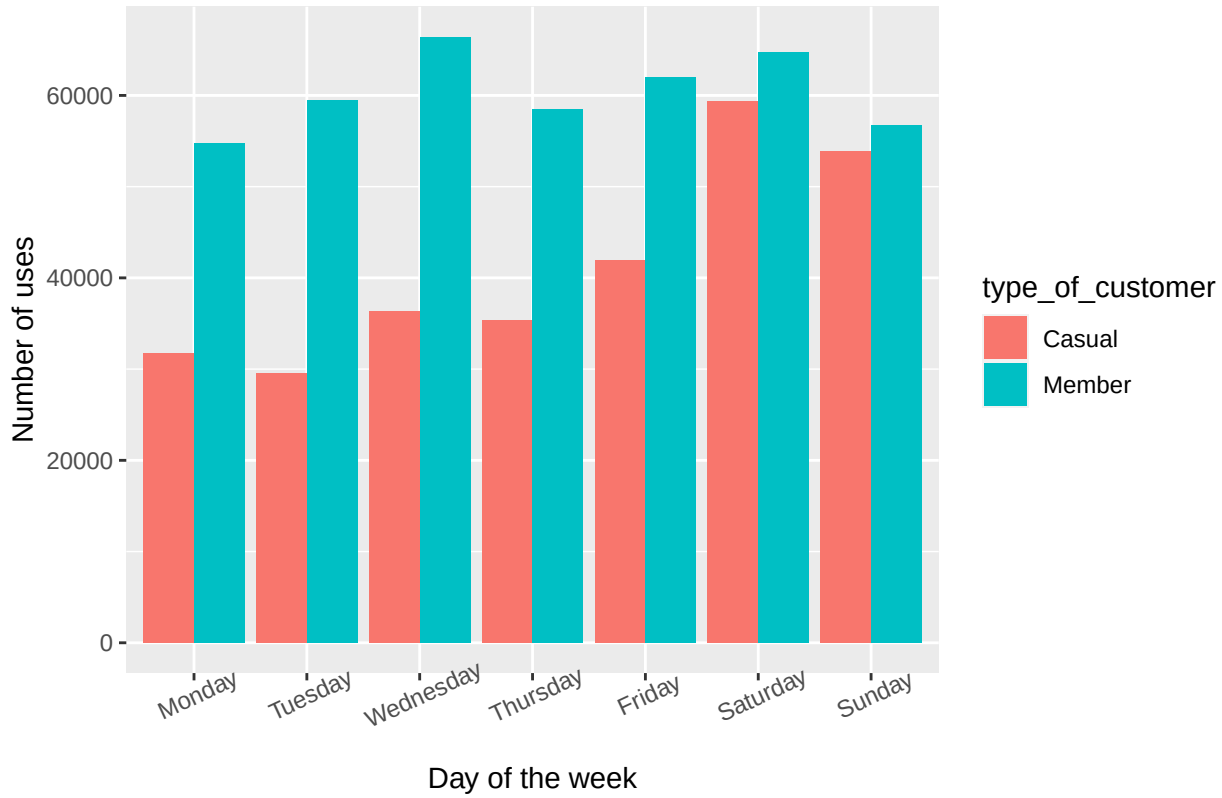
a. The months where bikes were used the most were June, July, August and September which correspond to the months of the summer season.

b. The months where bikes were used the least were December, January and February which correspond to the months of the winter season.

c. Through most of the year, customers with annual memberships used more bikes.

The next will be to change the values of the x-axis from months to days. This results in Figure 2 shown below.

Figure 2: Number of uses vs day of the week



By using Figure 2, the number of bikes used were measured by the day of the week. By splitting the observations by the type of customer like in Figure 1 the following conclusions were obtained:

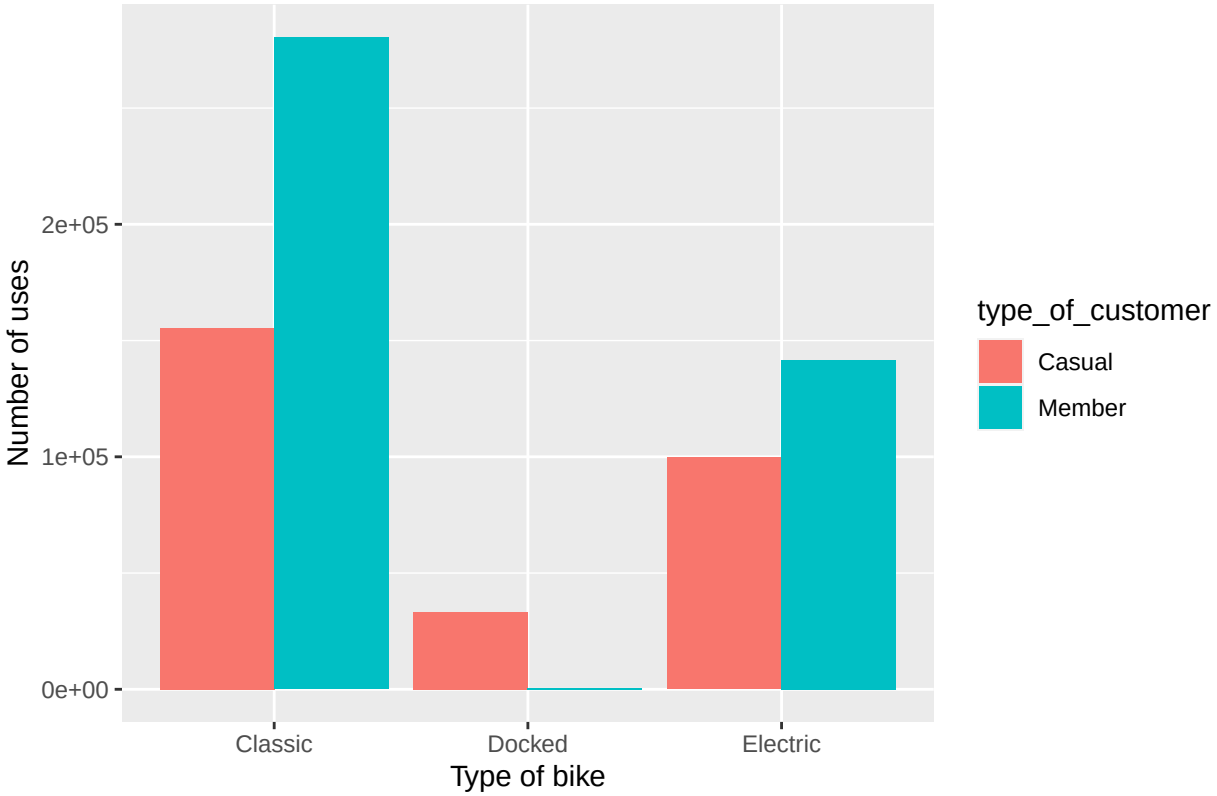
a. Each day of the week, customers with annual memberships used more bikes than casual riders.

b. Casual riders used bikes mainly during the weekend.

c. Customers with annual memberships used bikes mostly on Wednesdays and Saturdays.

To continue, the values of the x-axis are changed again. Now, the x-axis represents the type of bike used.

Figure 3: Number of uses vs type of bike used



With Figure 3 which shows the numbers of bikes used against the type of bike used by the customers, the analysis revealed the following:

a. Both types of customers used classic bikes mostly through the year.

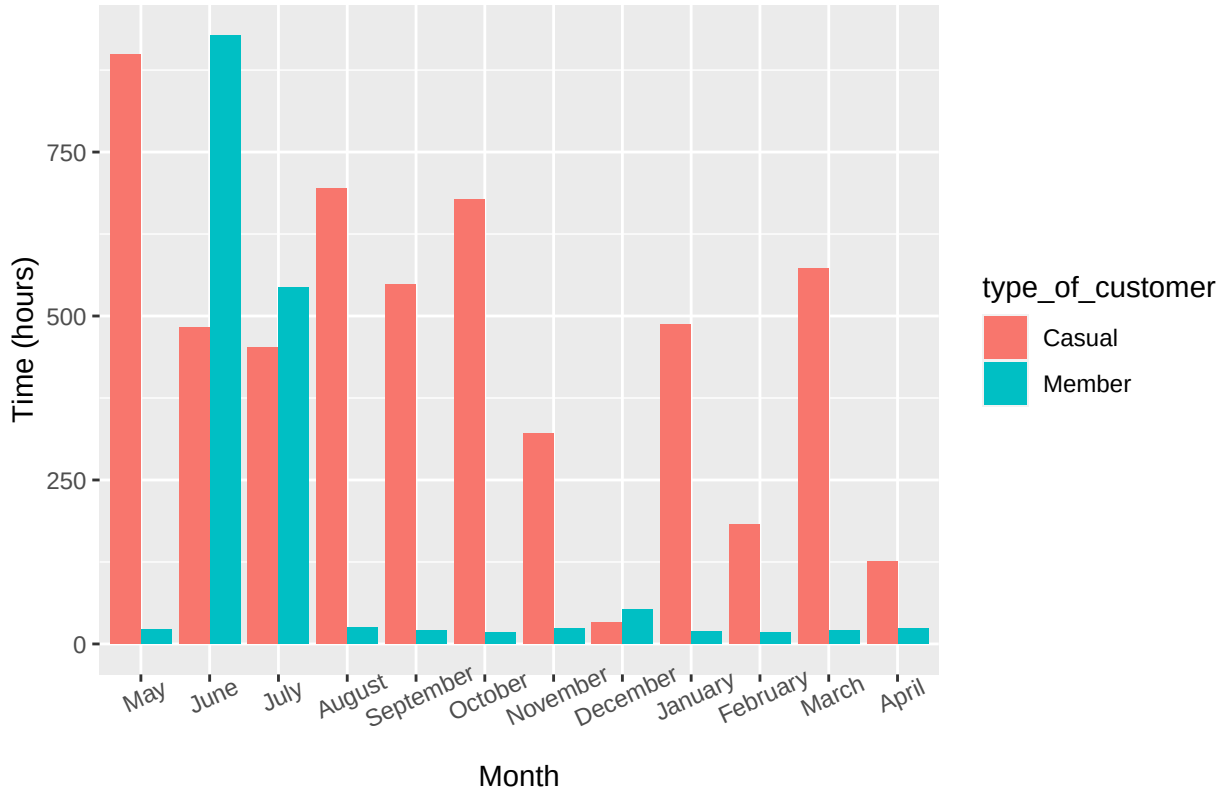
b. From the 3 types of bikes, the one that was used the least was the docked bike (especially for customers with memberships). There was no annual member who used a docked bike in the one year period.

c. Casual riders used more docked bikes than customers with annual memberships. This is the only of the 3 types of bikes where casual riders used more bikes than annual members.

To continue with the analysis process, the full_year_frame_v5 was used to create 4 smaller data frames, each one containing important information.

The first data frame was named max_frame which has 24 observations and 3 columns. Each observation indicated the longest time in hours that a bike was rented by a customer of each group during each month.

Figure 4: Longest bike rent per month

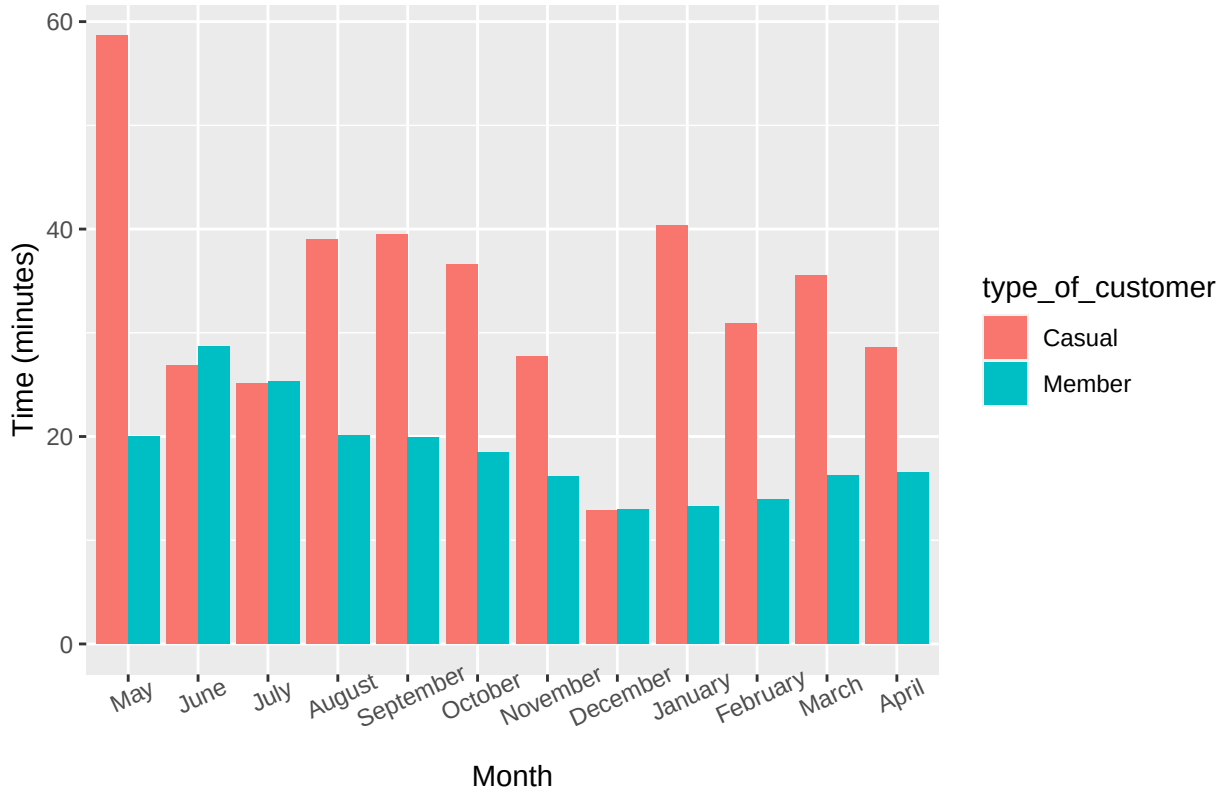


By using the values of the max_frame to create figure 4, it was shown that:

The longest bike rent of almost every month (except for June, July and December) was made by casual riders.

The second of the 4 data frames was called mean_frame which has 24 observations and 4 columns. This data frame contains the means of the time the bikes were used each month by type of customer. The mean is the average value of the time the bikes were used in the data frame.

Figure 5: Average bike use duration per month



The values of the mean_frame were used to create Figure 5 which displays the average bike use in minutes by each customer group each month. The most relevant insights given by the graph were:

a. In average, casual riders used the bikes for longer periods of time through most of the year (except during June, July and December). The longest average was the one of May which is almost 60 minutes and the lowest one was on December which is slightly higher than 10 minutes.

b. Customers with an annual memberships rented a bike for periods of time higher than 10 minutes and lower than 30 minutes in average through the year.

Since the values of figure 4 are measured in hours and more than half of them were above 100 hours, the results of the figure 5 could be skewed by those values and other which are close to them. For this reason the next and last 2 data frames were created.

The third of the four was mean_frame_short. It has the same amount of columns and observations than mean_frame but its values were calculated using the observations of full_year_frame_v5 where the values of the trip_length_in_minutes were higher than 0 minutes and less than or equal to 60. To compute the value of this frame, 681193 observations were used which is equivalent to 95.84% of the data from the original frame.

Figure 6: Average time of short bike trips per month

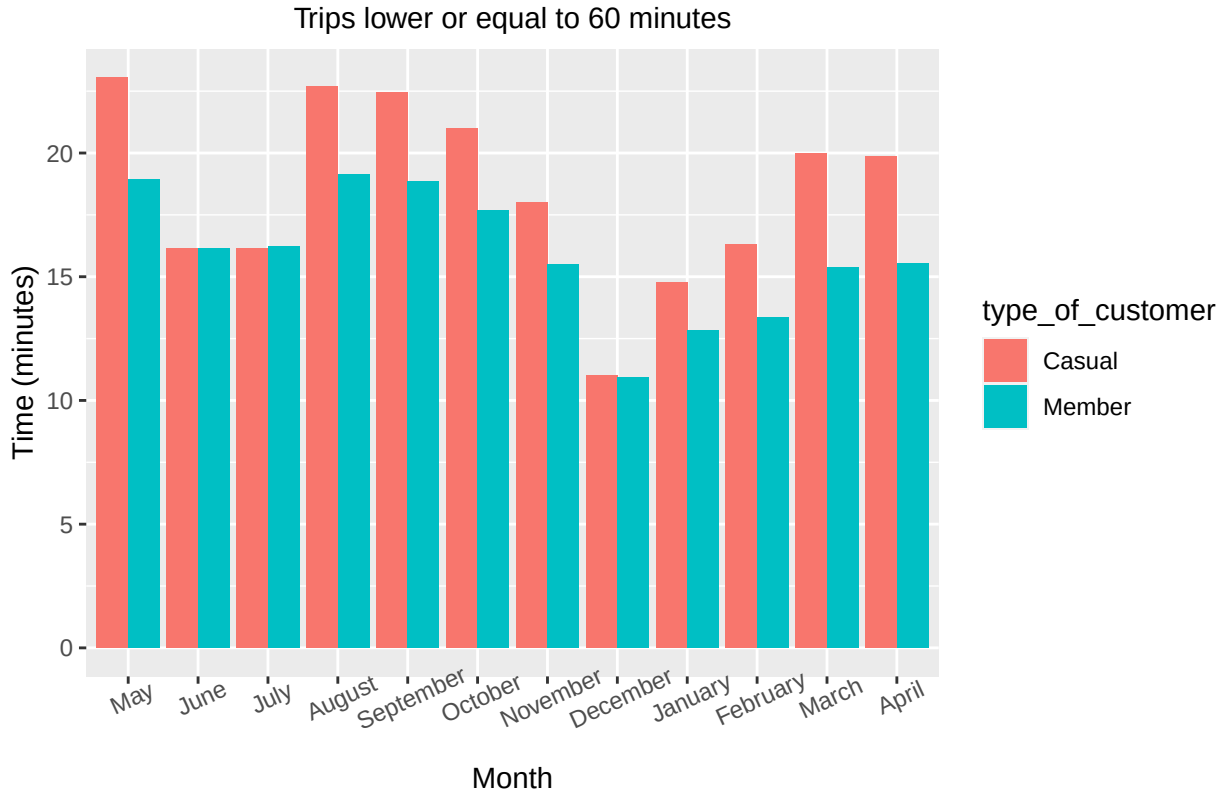


Figure 6 displays the means of the short bike trips by customer group each month. By comparing Figures 5 and 6, it was demonstrated that:

a. The average of each group of customers reduced each month (especially for casual riders).

b. Each average was between 10 and 30 minutes.

c. Casual riders still made longer trips than customers with annual memberships in average through most of the year.

The above shows that by removing observations where customers didn't make short trips, the averages for the bikes trips of each customers were reduced significantly.

The last data frame was called long_trips and was used to study the amount of customers who used bikes for periods of time longer than 60 minutes. This frame is a subset of the full_year_frame_v5 where the observations have values higher than 60 in the trip_length_in_minutes column. It contains 29491 observations which is equivalent to 4.15% of the data in the original frame.

Figure 7: Long bike trips per month

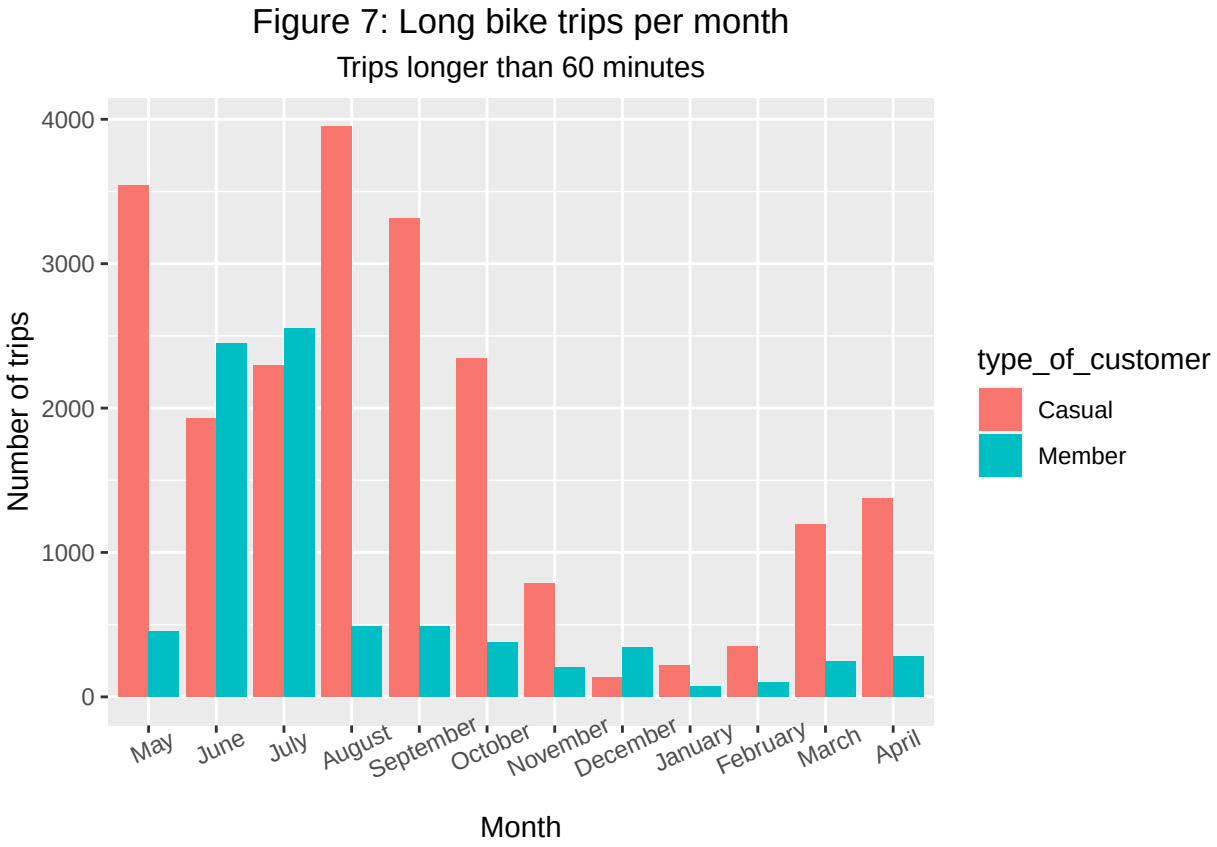


Figure 7 shows how many customers of each group did long bike trips each month of the 12 month period covered by the analysis. The insights shown from this graph were:

a. The number of casual riders who rented bikes for long time periods is higher than the one annual members through most of the year (except on June and July).

b. Those customers are the main reason of why the average trip duration of all the data are so different between the 2 groups of customers in some months.

Conclusions

- The months where bikes were used the most are the ones of the summer season (from June to September), where annual members used more bikes than casual riders.
- Casual riders used bikes mostly on weekends while customers with annual memberships used them mostly on Wednesdays and Saturdays.
- Classic and docked bikes were the bikes used the most and least respectively by each customer.
- Casual riders rented bikes by longer periods of time than annual members.

Recommendations to turn casual riders into annual members

a. Send an email to casual riders which explains them the benefits of becoming annual members.

b. Propose a discount or a special offer during the months of summer to turn casual riders into annual members.

c. Repeat this analysis after 1 year to review the effectiveness of the campaign to improve and/or create new ideas to increase the number of annual members.