Cyclistic Case Study Using R and RStudio

1. Scenario and Business Task:

A company called Cyclistic, a bike-share company in Chicago is planning to design a new marketing strategy to convert casual riders (who purchase single-ride or full-day passes) into annual members (annual membership) as annual members are much more profitable than casual riders.

You have just joined the Cyclistic Marketing Analytics Team as a Junior Data Analyst and are tasked to design a new marketing strategy to convert casual riders to annual members. In order to do that, however, the Marketing Analyst Team needs to better understand how different customers (annual and casual) are using Cyclistic bikes, why casual riders would buy a membership, and how digital media could affect their marketing tactics.

2. Data Sources Used:

The Cyclistic data is organized year wise and quarterly zip files. For the purpose of this case study, I used and analyzed one year worth of data from Cyclistic. More specifically, the trip data from Q2 2021-Q1 2022. The zip files are downloaded and unzipped locally and then uploaded to the Integrated Development Environment (IDE). For the purpose of this case study I attempted to upload and analyze the data in two different environments - RStudio and Jupyter Notebook. I will explain how I performed the Case Study using R in this document.

3. Data Cleaning:

This is the prepare phase of data analysis where the parameters like completeness, accuracy, relevance and uniqueness of the data are checked. The following are the steps I took to ensure organized data, maintain data integrity and credibility.

- Set up the working directory (to be the one where I stored all the datasets to be analyzed) and created a R file (source code)
- The necessary packages were installed (i.e., Tidyverse, Lubridate, ggplot2)
- Collect the data from csv files to data frames (Q2_2022, Q3_2022, Q4_2022, Q1_2023) using read.csv function
- Used str, colnames to examine the data frames, columns, rows and the column data types
- Checked ***col_names*** of cols and renamed it to be the same ‘name’ using the **rename** function so its easier to combine these 4 datasets into one large dataframe
- Checked ***datatypes*** of cols and change the datatype using the **mutate** function to be the same ‘datatype’ so its easier to combine these 4 datasets into one large dataframe
- Combined the four smaller datasets into one large dataframe (all_trips)
- Deleted columns in the large data frame(all_trips) that were not consistent in all datasets

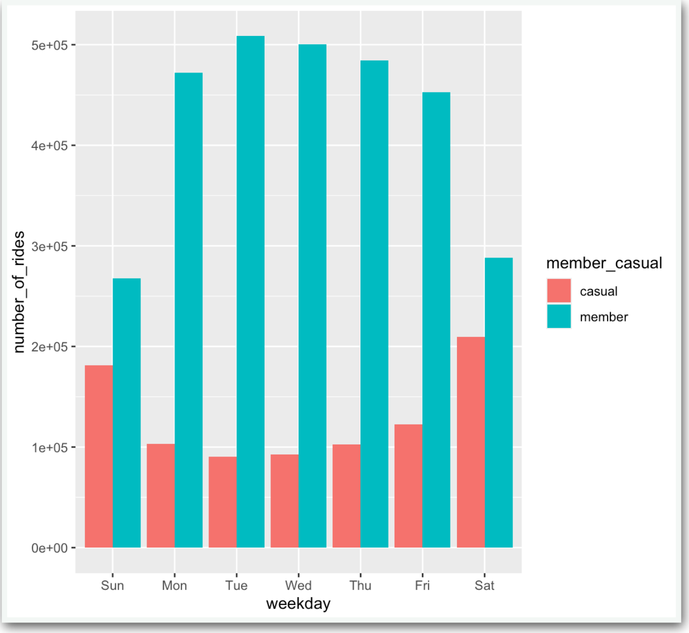
4. Data Manipulation (to prepare for analysis)

- There are four notations for customer type in ‘member_casual’ column. Subscriber and Member means the same and Customer and casual means the same. Hence using recode function, the column is recoded to convert all values that are subscriber to member and customer to casual leaving the column with only two variables: ‘casual’ and ‘member’
- Converting the datatype of ‘started at’ column that contains time at which the rider started the ride to a ‘Date’ datatype to extract useful information for analysis
- Added three new columns month, day, year, day_of_week using the format function
- Add a another column ‘ride_length’ that gives the length of the ride from subtracting end time from the start time and also made sure the ride_length is of type numeric to run calculations on the data
- Noticed some entries in the dataframe where ‘ride_length’ is negative values and the column ‘started_at’ contains “HQ_QR” which means the bikes were taken out of docks and checked for quality by the company

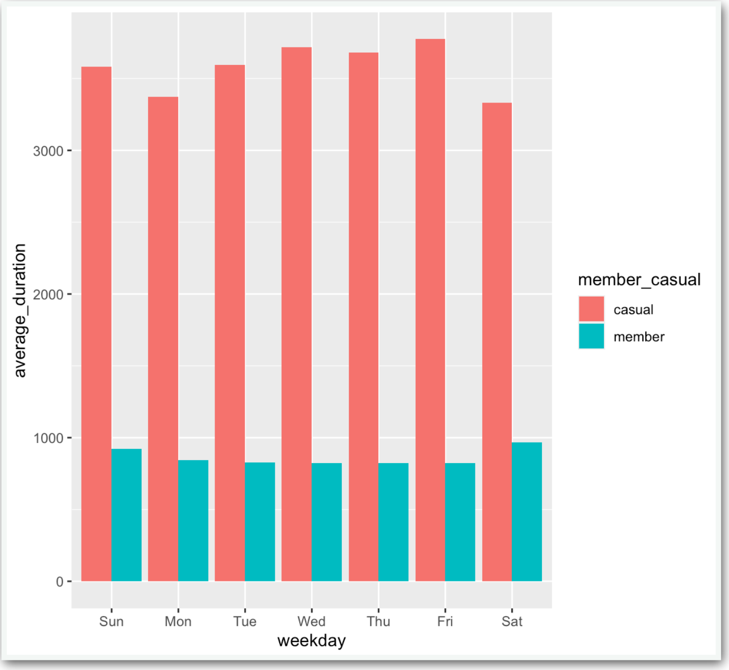
5. Analysis

- Performed descriptive analysis on ride_length (in seconds)
- Was able to analyze data to gain different insights like: average ride length, maximum, minimum and mode ride length, trends in riding time v/s hours, days, months and seasons.
- Calculated mean average ride_length and grouped them by rider type (member or casual)
- Further, calculated the mean average ride_length and grouped them by both rider type and day of the week.
- Used ‘ordered’ function to order the ‘day_of_week’ column for analysis purposes to print the week in an orderly way from Sunday to Saturday
- Putting all the above together, analyzed ridership data by rider type and weekday and calculated two important data columns- ‘number of rides’ and ‘average duration’ for each category
- Visualized the data using ggplot2 and geom_col( ) and plotted two graphs: number_of_rides v/s rider type and average_duration v/s rider type categorized into days of the week.

6. Visualization



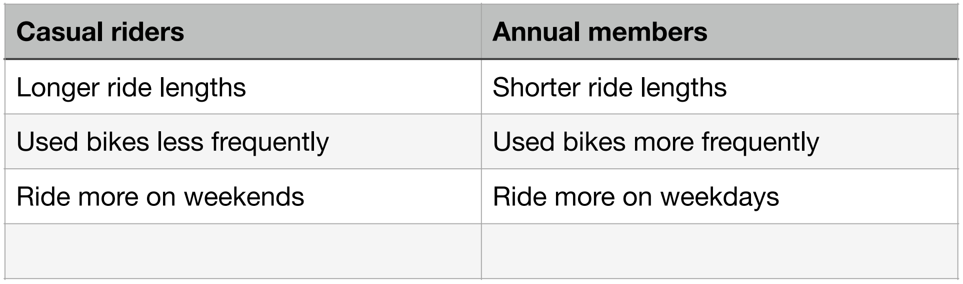
Number of rides for each ride type



Average Duration for each rider type

7. Summary

- Analyzed historical bike data from April 2021-March 2022 to identify trends in how annual members and casual riders use Cylistic bikes differently.
From the bar graphs above, we can summarize how casual and annual members use Cyclistic bikes differently and thus conclude the following things:



Casual riders	Annual members
Longer ride lengths	Shorter ride lengths
Used bikes less frequently	Used bikes more frequently
Ride more on weekends	Ride more on weekdays

8. Top three recommendations

- To offer more flexibility, include weekend passes
- Further analysis can be done and trends can be identified with respect to seasons, determining which season is the busiest in terms of bike usage. Special offers/ promotion campaigns/ membership discounts can be run right before the the busiest season.
- Casual riders being the customers that use the service for longer duration, we can visually highlight the value and benefits of becoming a member.