Data Analysis Cyclistic: Bike-Share Program

ASK:

Clear statement of the business task:

How do annual members and casual riders use Cyclistic bikes differently?

Help the executive team to create marketing strategies aimed at converting casual riders into annual members.

After analyzing the datasets supplied, we are going to provide relevant and qualified information in order to help Lily Moreno and her marketing team with suggestions of channels and campaign ideas and the executive team to make a data-driven-decision.

The goal is to make business more profitable, engaging more casual into the membership annual plan.

PREPARE:

Description of all data sources used:

The Dataset has been made available in this address by Motivate International Inc:
https://divvy-tripdata.s3.amazonaws.com/index.html

There is a 'Data License Agreement' published online in the address:
https://www.divvybikes.com/data-license-agreement

These have 10 sections that we are aware of when using this data for this analysis.

The privacy agreement stands that none information can be used to personally identify the customers. So this means we are not able to cross information about payment methods in order to find out if the casual riders live in the Cyclistic service area or if they have purchased multiple passes.

The last 2 years of the dataset are standardized by monthly files in CSV format.

After downloading it was stored in my computer for cleaning, processing and analyzing, it is located in my home, and the computer is locked with a password. The integrity of data was checked by opening each dataset, checking the shape of the tables, missing arguments, sorting and filtering to understand and get familiar with.

Data follows the rules for a good dataset because it is reliable, original, comprehensive, current and cited. It was provided by the company so is first hand delivered, this makes it reliable, original and cited.

The information inside the dataset is comprehensive because there is just a few missing information about started or ended stations. But it did not lead to error because it is possible to cross information using the geolocation of the bike when started and ended the ride if necessary.

Data is up to date. Although the company provided historical data from 2013, we will use only the last 23 months (from January 2021 to november 2022), and this part of the dataset follows all the requirements we need to have a solid analysis.

The data is composed by 13 variables, one for each column as follows:
‘ride_id',
'rideable_type',
'started_at',
'ended_at',
‘start_station_name',
’start_station_id',
‘end_station_name',
‘End_station_id',
'start_lat',
'start_lng',
'end_lat',
'end_lng',
'user_type'.

Each row stands for one ride, the rows from each file are not sorted chronologically.

PROCESS
Documentation of any cleaning or manipulation of data

The programming language used to manipulate the dataset was Python, with the Pandas library. The methods and syntaxes of this language are simple, and Jupyter Notebook is

a IDE to work in data analysis with a cell processing workflow that makes it clear and easy to build and visualize the commands flow.

All 23 files were opened, checked if all column names are the same, using automated process coded.

The columns ‘started_at’ and ‘ended_at’ were used to create a ‘trip_duration’ column, in order to check the average of time in which the bikes were being used.

There are 247 observations where the end time is before start time, these data rows were removed from the dataset and stored in a separate file for further investigation later. Despite this there were no problems with the data, using the information available and manipulating it we will be able to answer the questions and get some insights on the data.

There were 9255 observations that exceed 24h ride_length, this data was removed from the main dataset and separated for further investigation, because it might mean some maintenance procedure and will affect the statistics.

The ‘started_at’ data was used to create 4 more columns (day, month, year and weekday) this way we can aggregate data and determine which weekday the service was most used for or find some trend in the service use during the month or the year.

The data was filtered with the objective to compare ride_length and number of rides for each kind of customer in different time frames.

Start stations and End Stations were grouped by casual users and rides registered in each one were counted in order to check concentration of this kind of user in the top 20 list stations.

About the rideable_types of bike, it was aggregated in a dataframe considering the type of user, in order to understand which kind of bike is most used for both kinds of users.

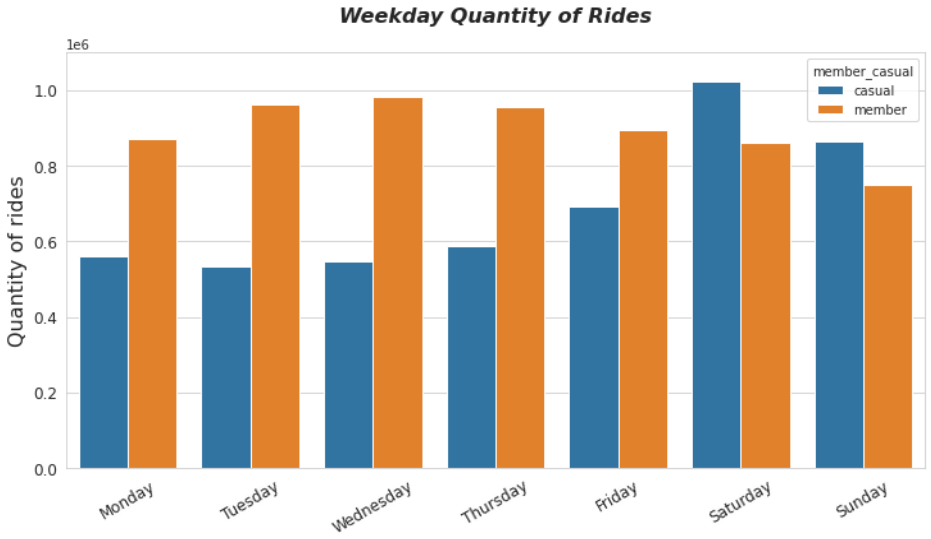
ANALYSIS

Summary of Analysis

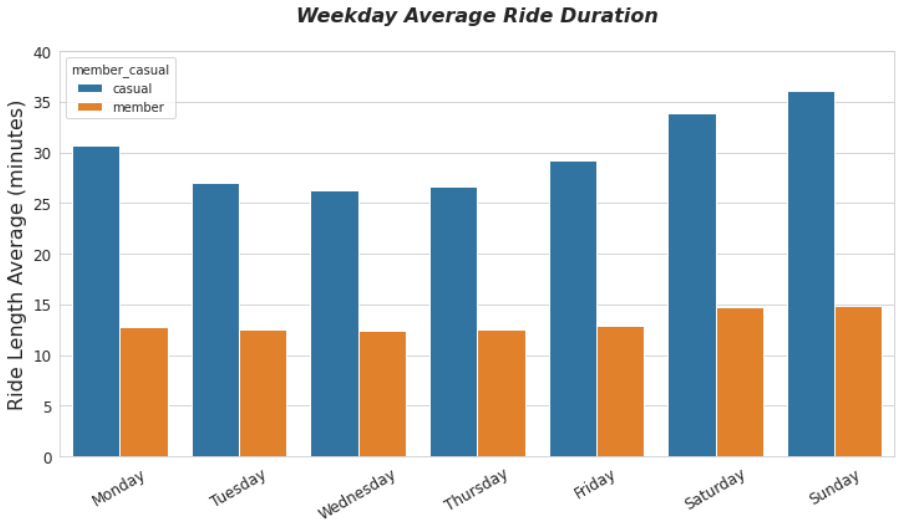
After cleaning and processing the dataset, organizing and filtering data with the objective of answer the business task: in which way members and casual riders differ using the service?, we found some interesting evidences in the data:

We have analyzed the data from the operation of this company in the last 23 months. It has 11.080.974 rides. And it shows consumer behavior categorized by multiple variables.

- Casual users have 4,806,143 rides (43.4%) and members have 6,274,831 rides. (56.6%).
- The number of daily rides for members is almost the same for each day of the week. With Wednesday slightly higher being the day in which the service is most used on the week by these users.
- The number of daily rides for casual riders are steady from Monday to Thursday, increasing on Friday, on Saturday it almost doubles the number of rides than in the four first days of the week. On Sunday it is 25% less than on Saturday but still is the second day with the most casual rides in the week.



- The overall average ride length is 20.8 minutes. The average ride length for casual users is 30.71 minutes. The average ride length for members is 13.21 minutes. The average duration of the rides for casual users are 2.26x higher than members.
- The average of ride_length by day of week for members is more stable and smaller than casual users. For casual riders, we have the shortest ride length average on wednesday, and following the week we find the longest average ride on Sunday. It is important to make clear that this length is based on rent time not distance. So we can infer that casual riders stay longer with the bike rented than members.



- When we organize stations by the information in which the ride starts and in which the ride ends, we do not find any extra pattern, just found the most used stations, that is probably because they are located in the city center with more nearby points of interest.

Streeter Dr & Grand Ave	124040
Millennium Park	58809
Michigan Ave & Oak St	54925
DuSable Lake Shore Dr & Monroe St	47792
Shedd Aquarium	43067
Theater on the Lake	39751
DuSable Lake Shore Dr & North Blvd	38401
Wells St & Concord Ln	35877
Clark St & Lincoln Ave	30299
Indiana Ave & Roosevelt Rd	30108

ACT

After analyze the data, we achieve some conclusions:

Casual riders locate bikes less frequently and use them for more time. Members locate more often and use it for a short amount of time.

Member riders locate bikes in stable quantity, with a small standard deviation during the week, in the meanwhile, members have a peak from Friday to Sunday.

People that work 9 to 5, are not able to use the service in this period and usually are able to use it on the weekends.

The idea we came as insights for the marketing team are:

1 - Create a different membership for these casual users, that is cheaper than the regular membership, it would be another contract, because they will have some constraints about quantity of rides, and even days of week.

2 - Post a campaign encouraging people to use the system more often for their health, not just on the weekends, and convince them that having a membership is almost as if you have your own bike, you do not rent, you just take it and put it back, with no headaches about maintenance.

3 - For the riders that have used the service in the last 3 months and can prove it by receipts, the company might offer free membership for 3 months. This can be used to show to these customers how convenient it is to have the right to use the service without having to manage daily tickets.