Data Analysis on the Divy Bike Dataset

ASK

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

PREPARE

Data is available via csv file. I have decided to use Rstudio as my development environment for processing and analyzing the data. This environment is great for adding libraries for helping clean and sort the data. This environment also enables me to step through the data as it is being processed, and visualize in real time. The data has been made available by Motivate International Inc. under this license.

Data Storage:

1. Data is stored locally on computer
2. Data is stored in Google Drive
3. Data is duplicated and redundant
4. Analysis will be on duplicated data, original data will be saved incase of emergency

PROCESS

After downloading the data and storing it locally, and in duplicate, we are able to start processing the data. The data was in groups by fiscal quarter, the second, third, and fourth quarters of 2019, as well as the first quarter of 2020. Divy changed up their column name conventions between 2019 and 2020, so we will have to convert older column names into the respective 2020 column names, so the data frame can be stacked on top of each other nicely. Using the str() function we can get structures of the quarterly reports, and using the mutate function we can change the column names to match. In order to stack the quarters on top of one another, we will use the bind_rows function to combine the csv files into one data frame.
We also have to drop some data that is no longer being recorded, so the data frames match each other even more uniformly. I dropped the columns start_lat, start_lng, end_lat, end_lng, birthyear, and gender.

ANALYZE

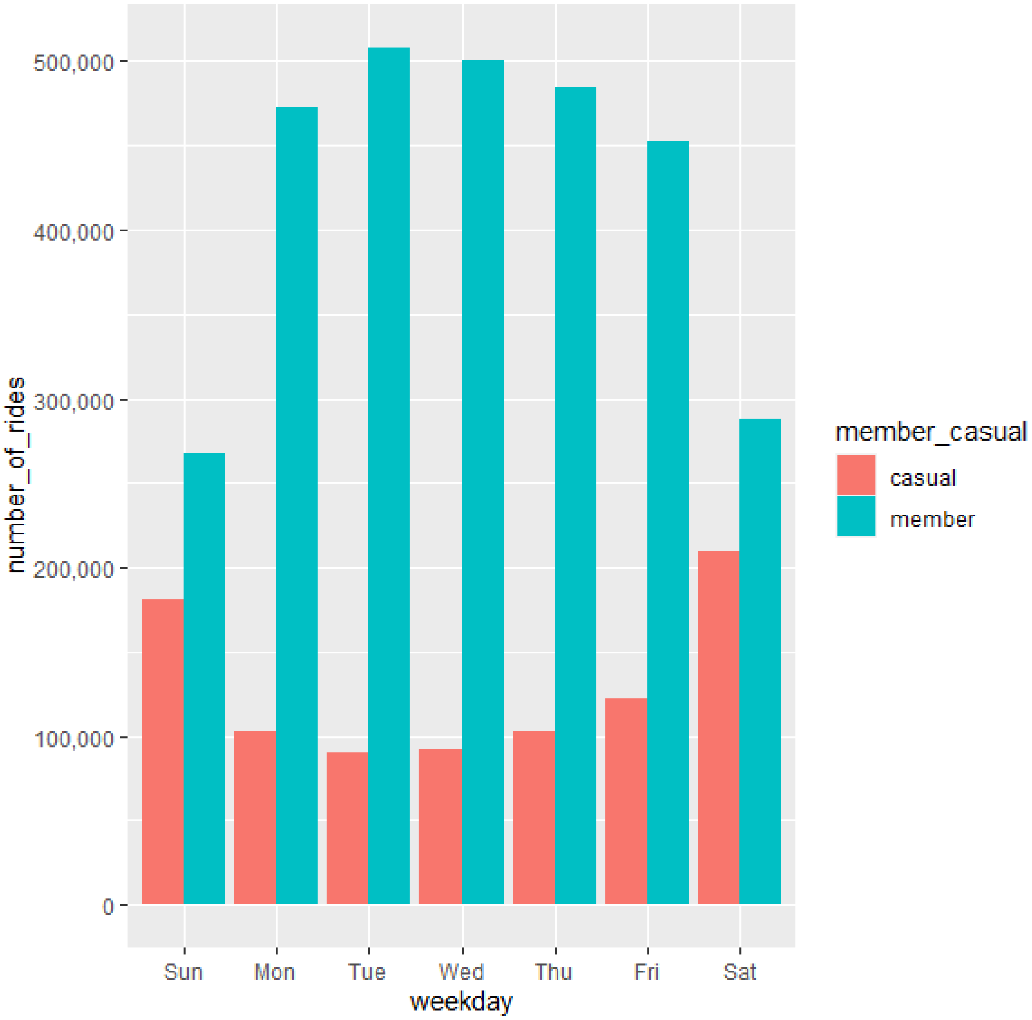
There are a few more issues I needed to deal with before the data was ready to be fully analyzed.

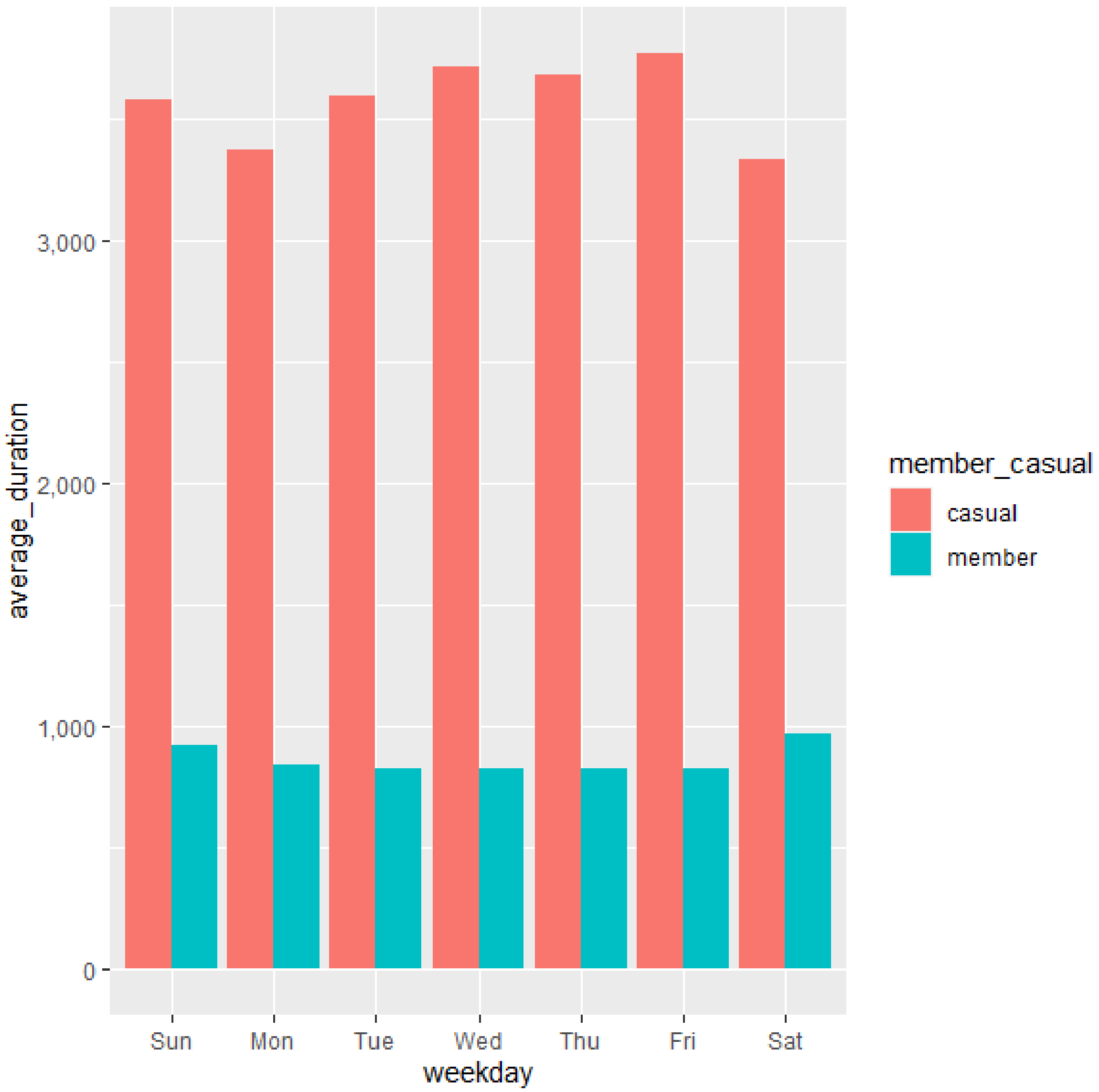
1. In the “member_casual” column, there are two names for members, as well as two names for casual riders. I will consolidate these four labels into two labels, casual and member.
2. The data also can only be aggregated at the ride-level. This is too granular and specific for our needs. I want some additional columns of data, such as day, month, and year, that will provide additional values in which to group and analyze data.
3. I will add a ride_length column to the entire dataframe
4. Want to delete any rows in which Divy has taken bikes out of rotation, which have led to negative values in ride_length.

Using the mutate function and recode function, I was able to rename all values in the member_causual column to be uniform.
Using the as.date function, as well as format, I was able to split the value in the started_at column into separate columns each corresponding to date, day, month, year.
Using the lubridate package, and the contained difftime function. Was able to calculate the ride duration by getting the difference of the started_at and ended_at columns. Stored those independent values in their own respective column.
Anytime it comes to removing data from a data frame, I admittedly get nervous. So instead of overwriting the data frame with the new (albeit smaller) dataframe. I stored the new dataframe (with dropped data) into a new data frame name. This final data frame is the data frame I will be using for analysis.

Running the data frame through ggplot, and using geom_col, to create charts to compare the usage rates of Divy bikes, I found some interesting relationships between usage and memberships.

<u>SHARE</u>





A few key points are of interest to me.
General usage rates:

1. Member’s use the bikes much more frequently.
2. Casual’s have much higher duration of use.

Weekday usage rates:

1. Members use the bikes more during the weekdays, than on the weekends. Tuesday being the busiest day for members to use the bikes.
2. Casual’s use the bikes more during the weekends than during the week.
3. Trend lines of duration use are all pretty flat. The day of the week does not seem to have a correlation to the duration of trips.

ACT

Due to the low duration, but high frequency of use on weekdays It would appear to me that the members of divy are using the bikes as transportation to and from work. While the casual riders are using the bikes as day trips during the weekend. Each of these types of customers are equally important customer bases, and help keep the bikes earning money each day.

With further information, such as rates for members vs. casual, one could figure out which is more profitable. Longer duration, but less frequent, or more frequent use but with lower durations.

Also, there could be analysis run on the pick up and drop off locations. To determine which of the stations are used most frequently. Using the latitude and longitude values, one could determine if there's a general trend in cardinal direction as well. This could create opportunities to track user activity based on the direction they are going, and could help choose locations for additional drop off or pick up locations. These questions are outside of the scope of the original business task of this analysis.