Cyclistic Bike Share Preliminary Report

by Jeff Smith, Data Analyst
Last Updated July 23, 2023

Background

Pulled public data from https://divvy-tripdata.s3.amazonaws.com/index.html
for the 12 months starting July 1, 2022 and ending June 30, 2023.

Divvy (now Lyft Bikes and Scooters, LLC) provides anonymized data for non-commercial use under this license https://ride.divvybikes.com/data-license-agreement

This is first-party data known to be reliable, current, comprehensive and properly cited.

Initial Preparation

The files for the most recent 12 months all follow the same naming convention
yyyymm-divvy-tripdata.csv (ex. 202207-divvy-tripdata.csv)

except for the September 2022 file which is named
202209-divvy-publictripdata.csv

so I changed it to
202209-divvy-tripdata.csv

to match the other 11 files and make it easier to load and manipulate all 12 files.

Loading, Validating and Prepping the Data

Loaded the 12 monthly data files (July 2022 through June 2023) and reviewed the column names. They're the same in all 12 files. If they weren't, I'd have to correct the differing names before combining everything into one file. I also made sure the data types are the same for every column in every file.

The files combined successfully, so I moved ahead with data cleanup.

- Removed columns for latitude and longitude.
- Inspected the new table to verify column names and data types.
- Added columns for date (original column is combined date and time), month, day, year, day of week and hour.
- Added ride length calculation in minutes for all trips.
- Added columns for season(Spring, Summer, Fall, Winter) and time of day (Morning, Afternoon, Evening, Night).
- Removed rows for internal quality checks and NA values (no ride length)
- Removed duplicate rows.
- Also removed any rides less then 10 minutes or longer than one day (1,440 minutes)
    * If the ride is less than 10 minutes, the bike might not have actually gone anywhere.
    * If the ride is longer than one day, it may be due the bike being abandoned or not being returned in a timely manner.

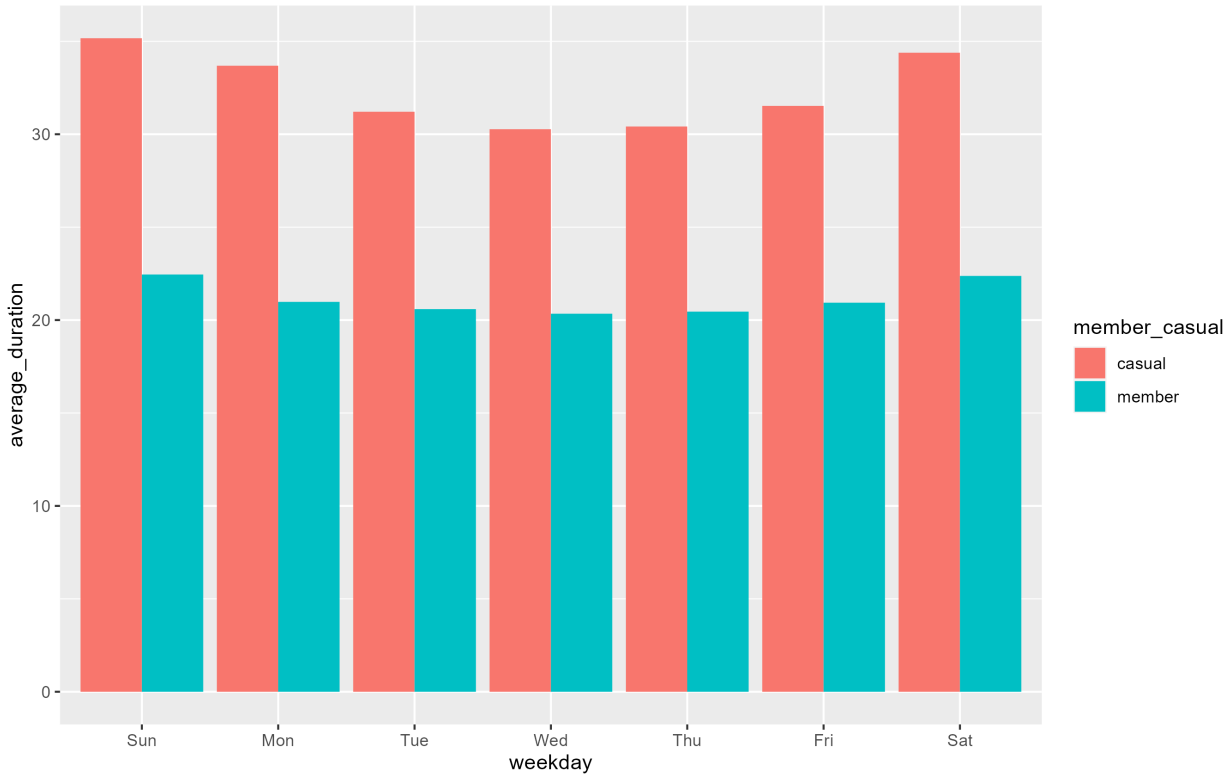
Initial combined files has 5.78 million records. The first round of cleanup reduced that to 2.8 million records. The second round of cleanup reduced that to 1.84 million records.

Verifying The Cleaned Data

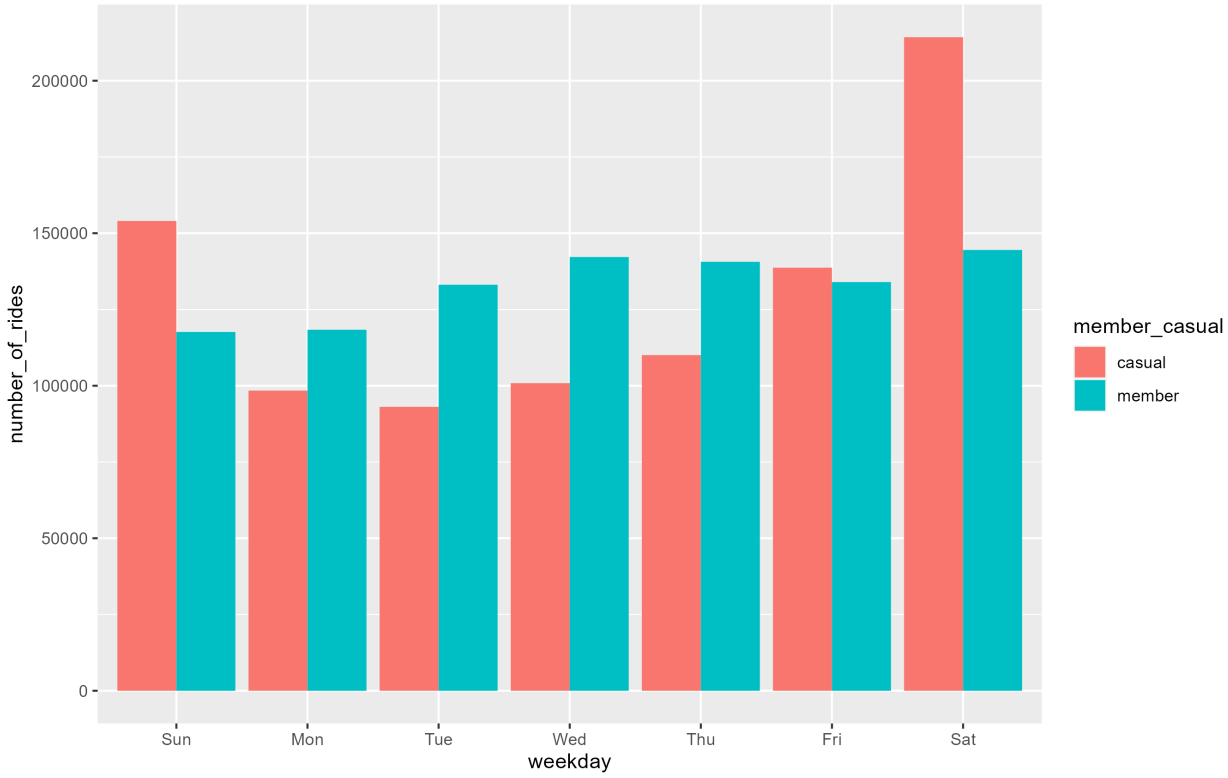
The min, median, mean and max values for member and non-member (casual) ride time look good, overall and broken out by day of week.

I generated a final clean data file and linked it to Tableau for further analysis and visualization. Preliminary plots are on page 2. Further analysis, visualizations and recommedations are in the Cyclistic presentation.

Ride Duration in Minutes by Weekday



Number of Rides by Weekday



Primary Sources

Data Source: Divvy (Lyft Bikes and Scooters, LLC)
https://divvy-tripdata.s3.amazonaws.com/index.html

Data License
https://ride.divvybikes.com/data-license-agreement

Other Sources

Sophisticated, Clear, and Polished': Divvy and Data Visualization (by Kevin Hartman)
https://artscience.blog/home/divvy-dataviz-case-study

Date Formats in R
https://www.statmethods.net/input/dates.html

Time Intervals/Differences in R
https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html

Deleting Rows That Meet Certain Conditions
https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/

Exporting Data from R to CSV Files
https://datatofish.com/export-dataframe-to-csv-in-r/