Divvy Internal Report

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

I made sure all the files use the same naming convention to make it easier when writing the code to load them.

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, so no corrections needed.

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

- Reviewed the combined dataframe to verify column names and data types.
- Checked for duplicate rows. There are none.
- Dropped all rows with N/A values from the dataframe.
- Added a column with a calculation for ride length in minutes.
- Created a box plot to detect outliers for ride length. There are a great many rows with extraordinarily long ride times that are outliers.
- Calculated which outliers to remove using the IQR.
  - Removed values over 36 minutes.
  - Removed values under 5 minutes. The lower limit is -12.0, so I manually set the lower limit to 5 to eliminate negative ride times, any problems resulting in the bike not being used for a trip and any possible internal testing. There are no specific identifiers for internal testing.
  - Created another box plot that shows a cleaner dataset with fewer and less extreme outliers for ride length that will not skew the analysis.
- Added columns for date (original column is combined date and time), day, year, month, day of week and hour.
- Added columns for season(Spring, Summer, Fall, Winter) and time of day (Morning, Afternoon, Evening, Night).

Initial combined file has 5.7 million records. Analysis and cleaning reduced the record count to just under 796,000.

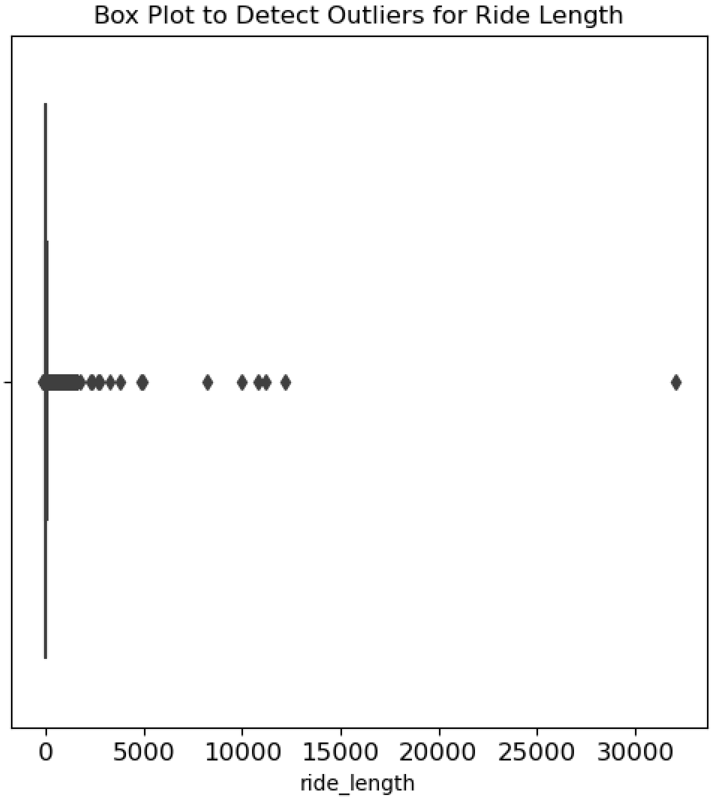
Verifying The Cleaned Data

Basic statistics for number of rides and ride times appear reasonable. Number of rides and ride times broken out by day also appear reasonable. Preliminary plots are on page 2.

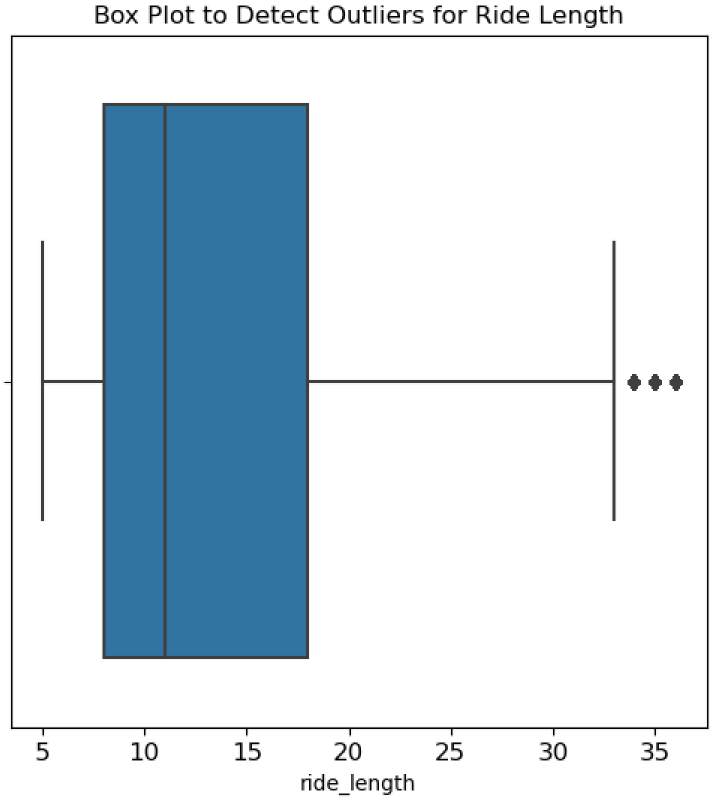
I saved the clean data to a CSV file for further analysis and visualization in Tableau.

Last Updated October 7, 2023

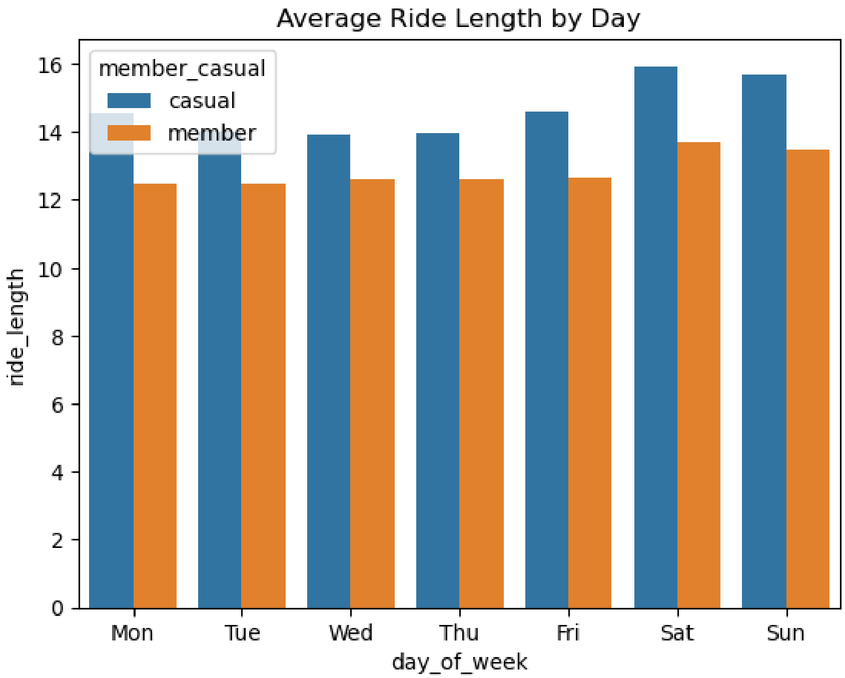
Initial Box Plot for Ride Length



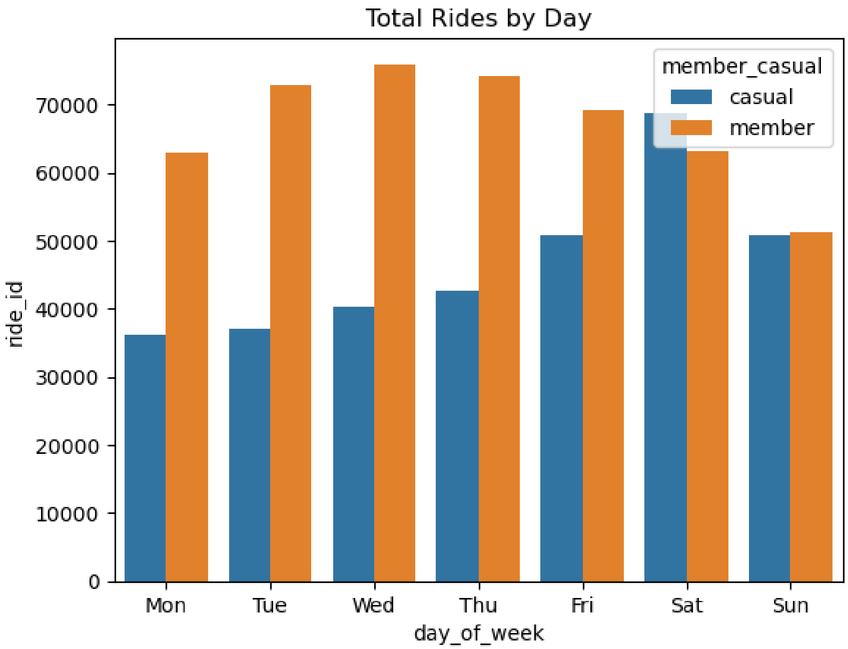
Box Plot for Ride Length after Cleaning



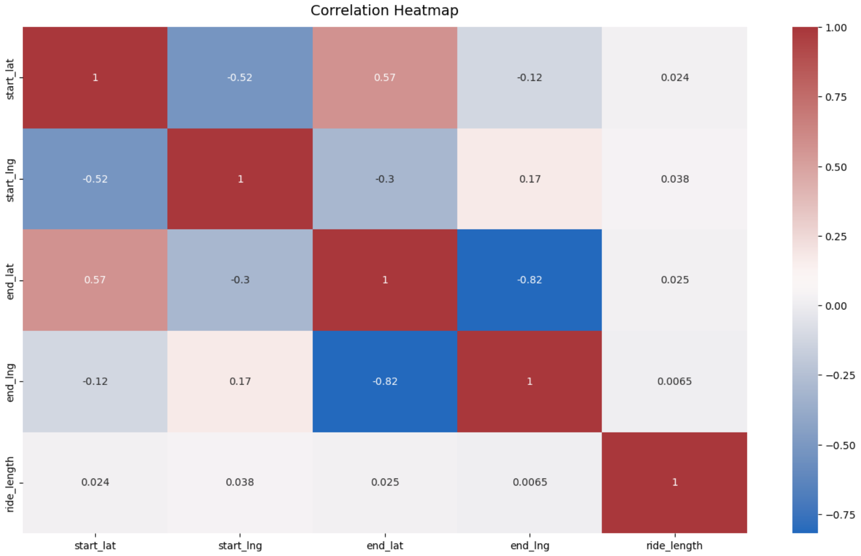
Average Ride Length in Minutes by Day



Number of Rides by Day



Correlation Heatmap



Latitude and longitude correlate with each other. Ride length doesn't correlate with any other numeric, so new new avenues of exploration.

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