Cyclistic Analysis

Key Question: “How do annual members and casual riders use Cyclistic bikes differently?”

**Ask**

* Business Task: To find trends that show how annual members and casual riders us Cyclistic differently. Use these trends to give recommendations for the marketing team to turn casual riders into members.
* Shareholders: Marketing team and Executive team

**Prepare**

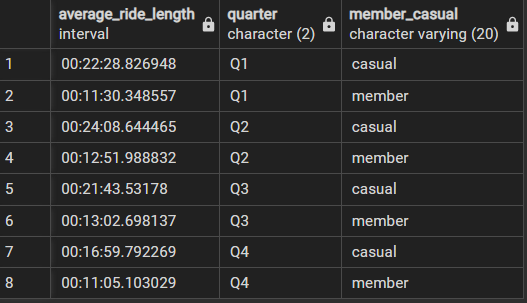
* Data is stored in .csv files
* Pros:
  + Data is from the company itself so it should be quite reliable.
* Issues:
  + Data is too big to fit in an Excel spreadsheet (~1,600,000 rows per sheet)
  + Column names are inconsistent
  + Data have inconsistent standards. Some sheets use “Customer” and “Subscriber” while some use “Member” and “Casual”
  + Some have more columns than others
  + A decent amount of missing data, bike rides with no station starting name, or with no trip duration
* Steps Moving Forward:
  + Bring Data into an SQL database and into R to make it fit
  + Make the column names consistent
  + Make all member types use “Member” and “Casual”
  + Use only the most essential columns when loading them into a database
  + Make note of empty data but proceed on as there is still ample data available. Also use numbers.Abs() to convert any negative trip durations to positive
  + Transform data in excel to get down to the essential columns and data for SQL

**Process**

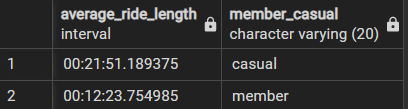
* Tools:
  + Microsoft Power Query for preliminary analysis and data cleaning. Take the cleaned data from here and use it in SQL and R. Use pivot tables to find some trends among the data.
  + SQL for main analysis, can handle all of the data properly. Create a PostgreSQL table to hold all of the data for the year. Upload all data and list its quarter.
  + R for some extra analysis. See if there are any trends that went unnoticed with R.
* Cleaning Process:
  + EXCEL
    - Used Power Query to load in and transform all 5+ million rows
    - Most Important Columns: Ride\_id, Start\_Time, End\_time, Start\_Station\_ID, Start\_Station\_Name, End\_Station\_ID, End\_Station\_Name, Member\_Type
    - Add Rows: For Ride\_Duration, Day\_of\_Week
    - Delete any other columns
    - Make all the dates into date types
    - Delete rows whose end time is before their start time
    - Make sure all data in Member\_Type is either “casual” or “member”
  + SQL
    - Add the quarter of the year to each row

**Analyze**

* SQL
  + Average of ride length by quarter and user type w/ time < 24 hours

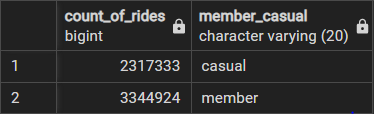


Average of ride length by user type w/ time < 24 hours

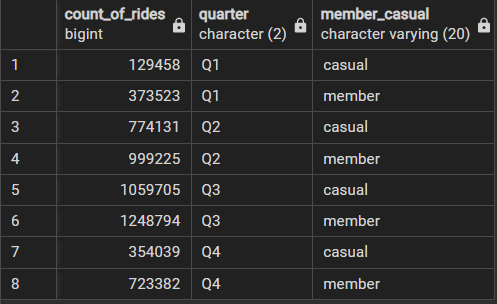


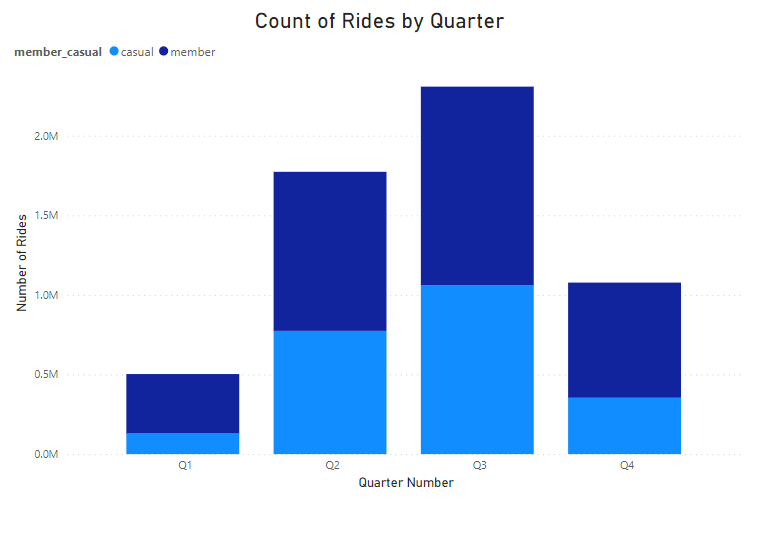
* + Find counts of rides

Member and Casual

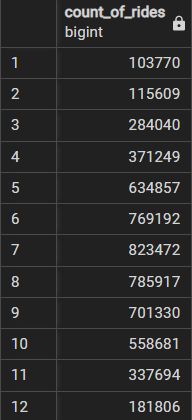


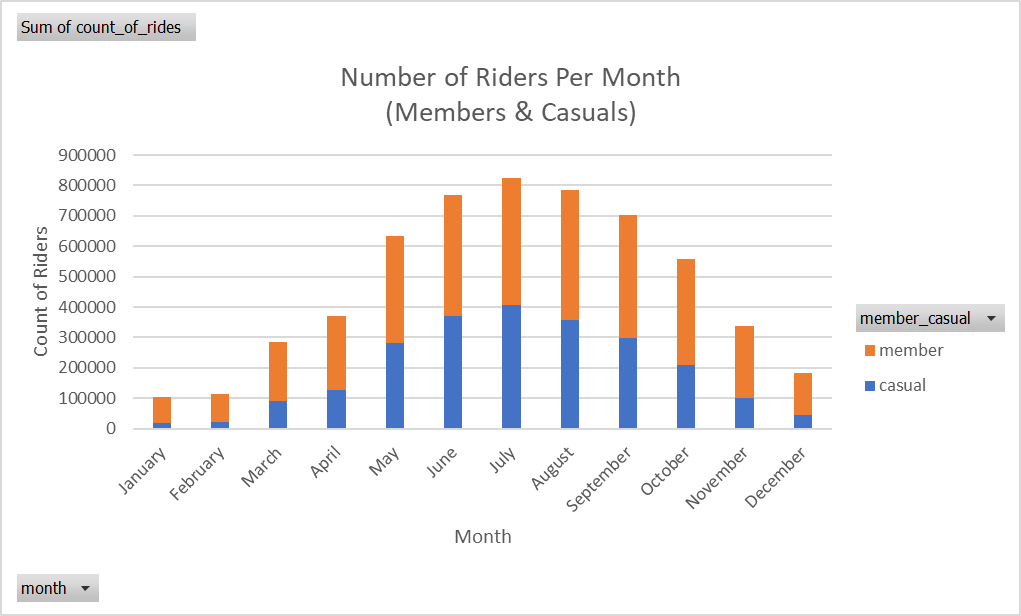
Member and Casual per quarter (quarter 3 is when casual riders are at peak, probably want to throw out the most marketing efforts in quarters 2 & 3)

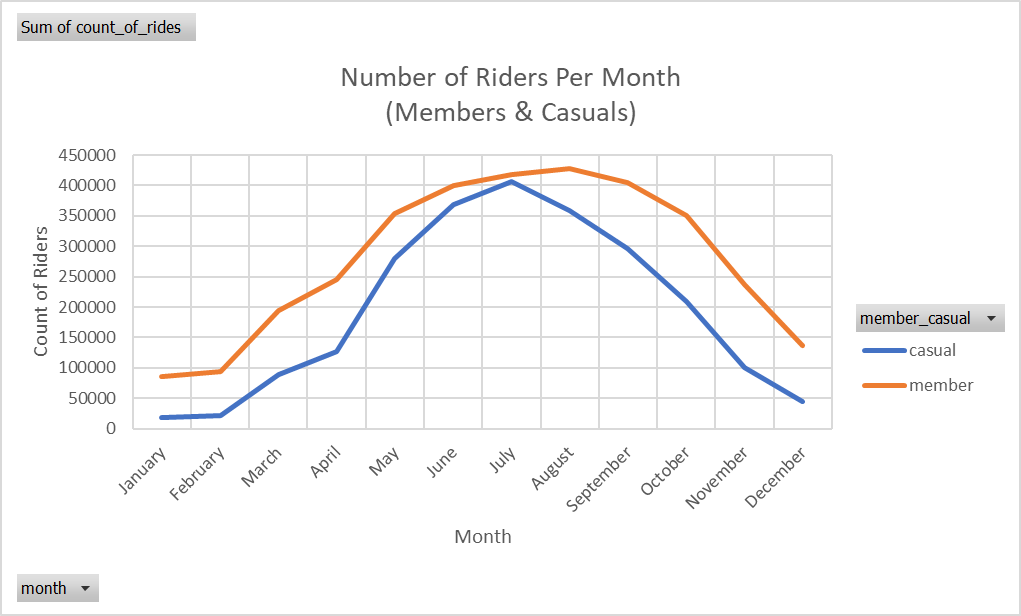




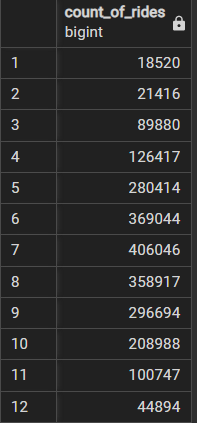
All members, highest number in July





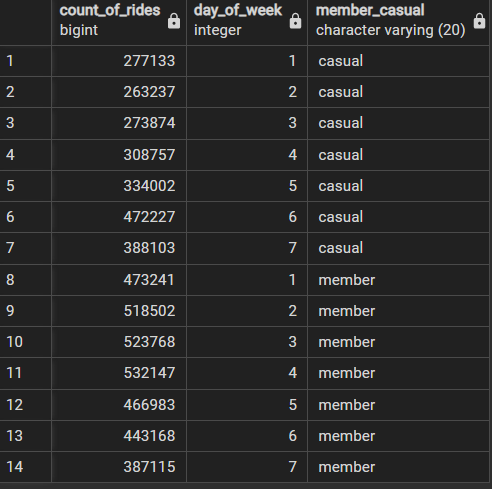


Only Casual members, highest number still in July

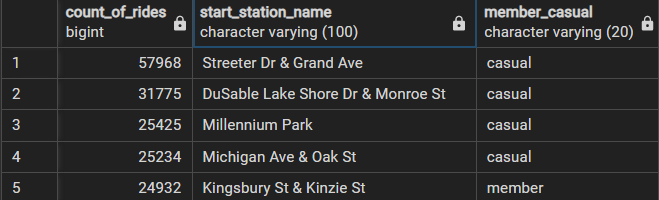


* + Find count of rides per day

Count of rides grouped by day and user type (casual riders ride the most on 6 & 7 (Saturday & Sunday) while members ride most during the weekday

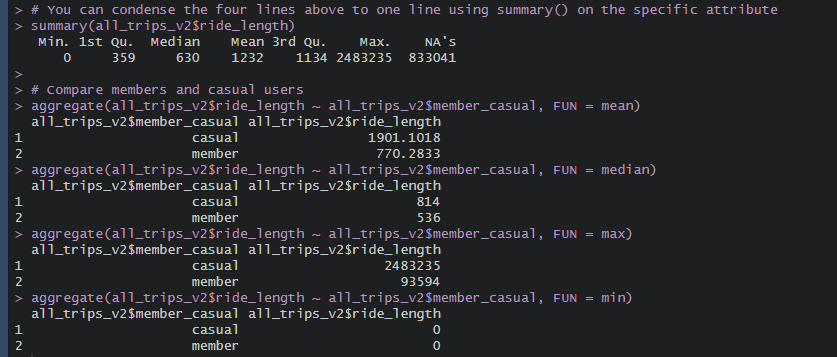


* + Most popular locations grouped by user type (4 of the top 5 locations grouped by user type are casual riders, should try to target these locations)

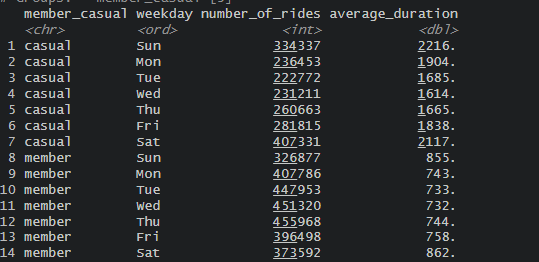


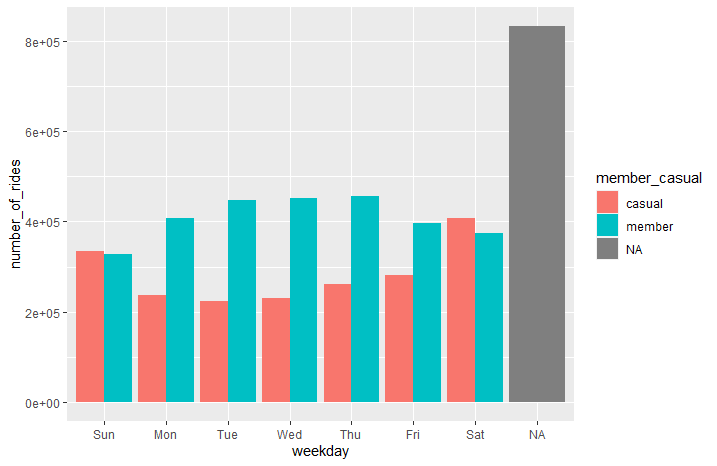
* R
  + Create a column in the data for what day of the week a ride is (easier to do in R than in SQL)

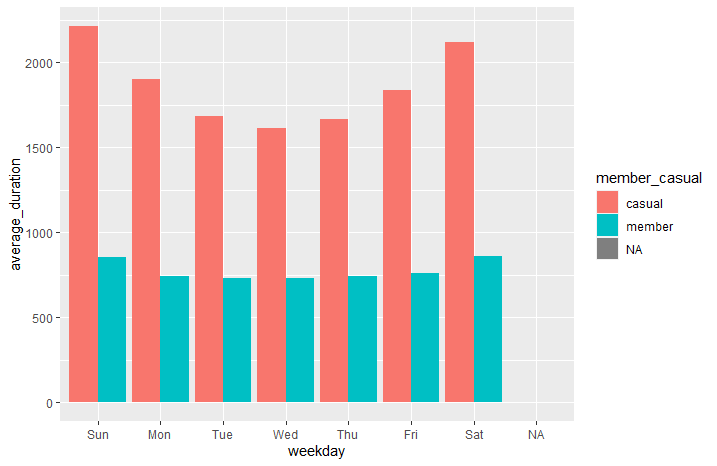
Show mean, median, max, and min for casual and members. Casual members take longer rides in all regards



Shows the number of rides and average duration of rides per day of week grouped by user type



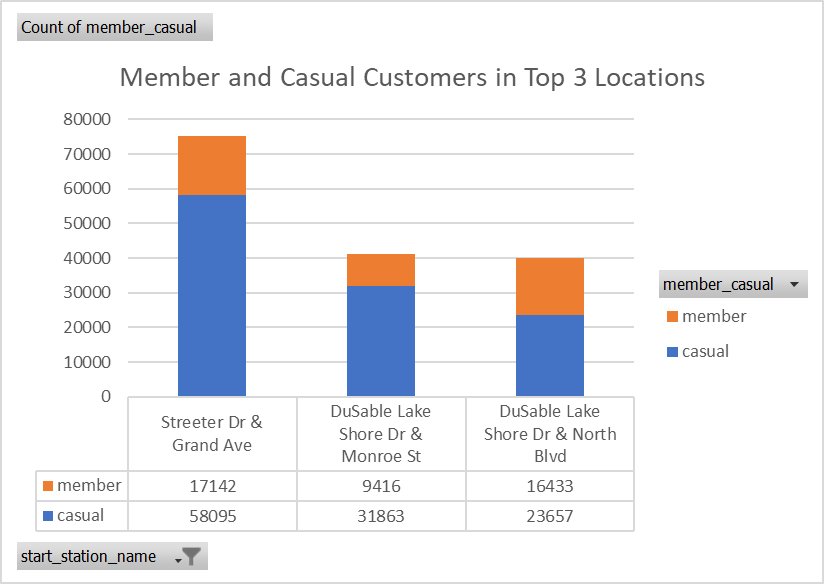




* Excel
  + Files are too big to use Excel but get a feel for the data in EXCEL

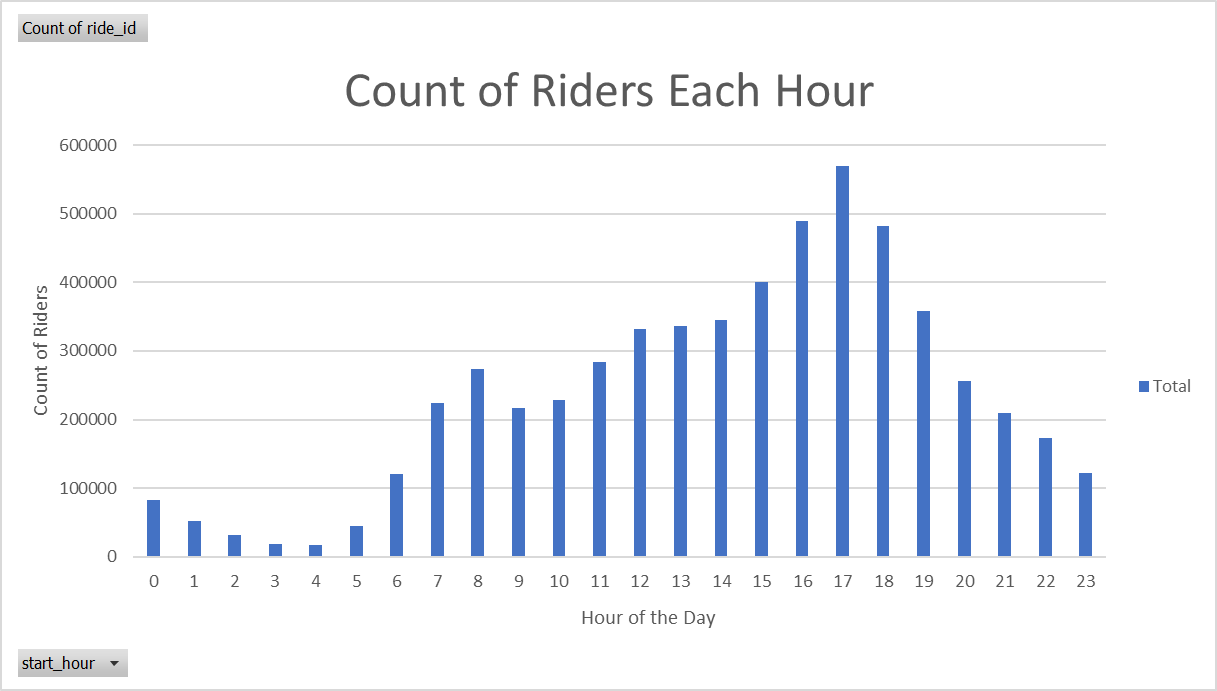
**FINDINGS**

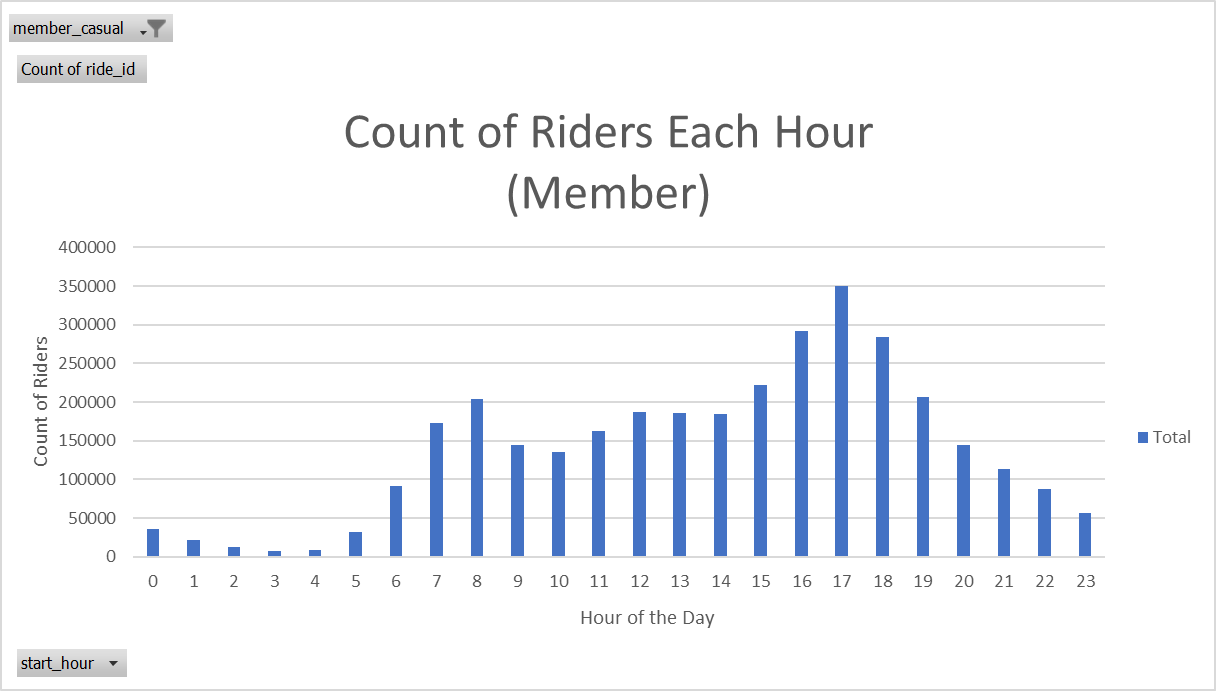
* Streeter Dr & Grand Ave is the most popular starting station overall (75k) and among casuals (58k) but only 11th for members. Advertise to this location in particular. DuSable Lake Shore Dr & Monroe St is another one to target. The most popular locations are more started by casuals than members. In the future try to keep data more consistent for station names since many were blank and analysis could be better w/o NULL values

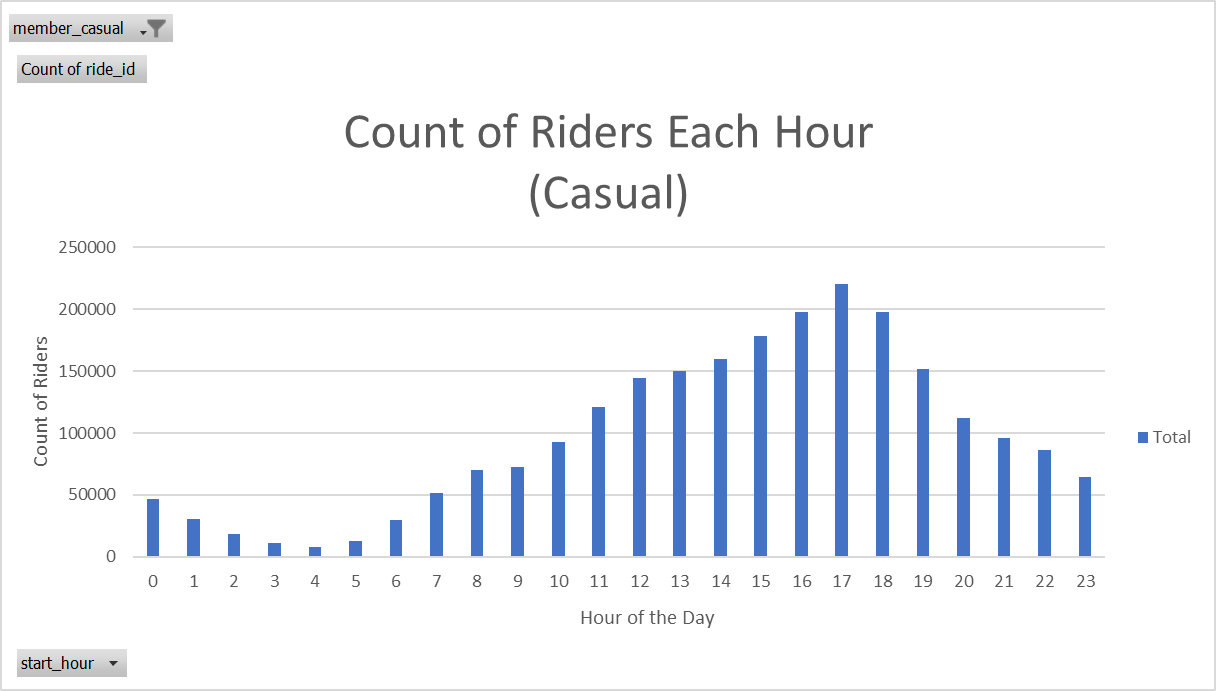


**Share**

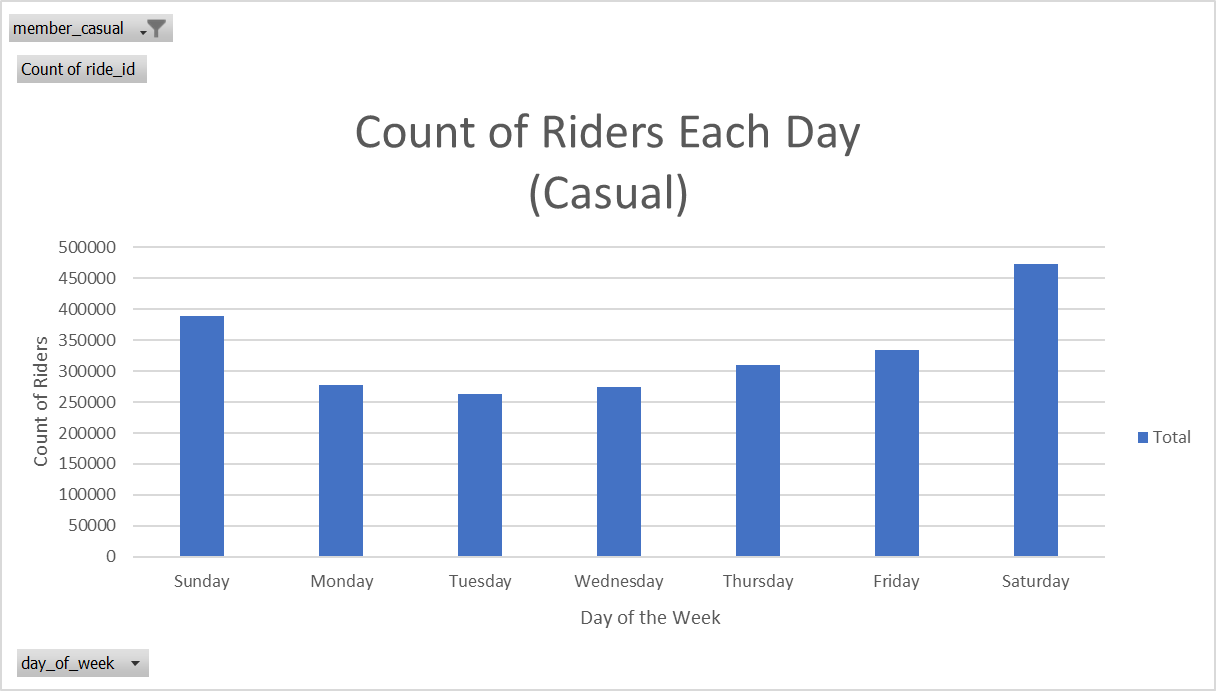
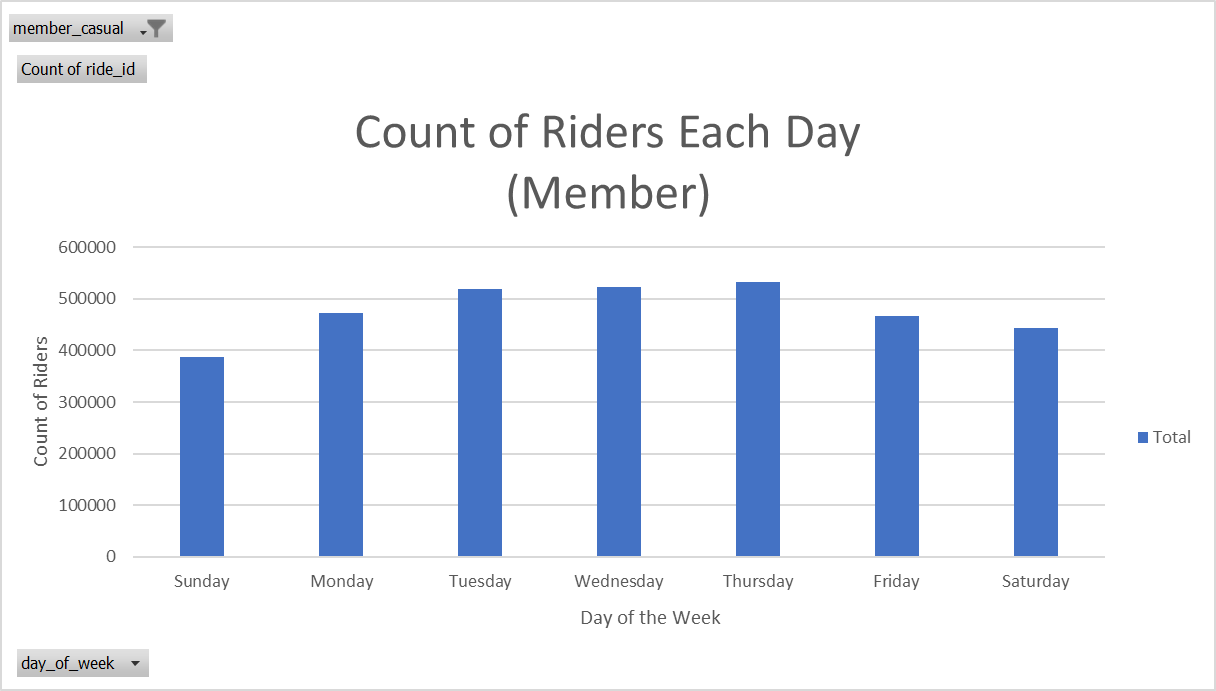
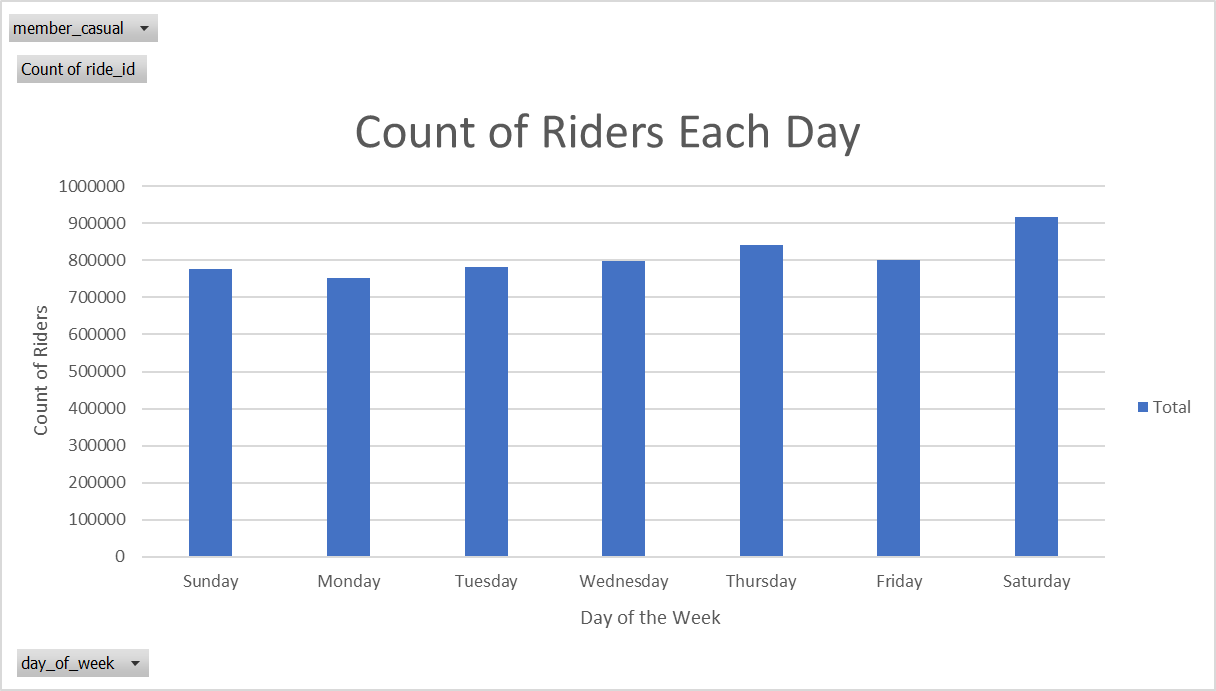
Rides at each hour (members use them more at commuting times than casuals who use them more in the afternoon, market to casuals that they can commute)

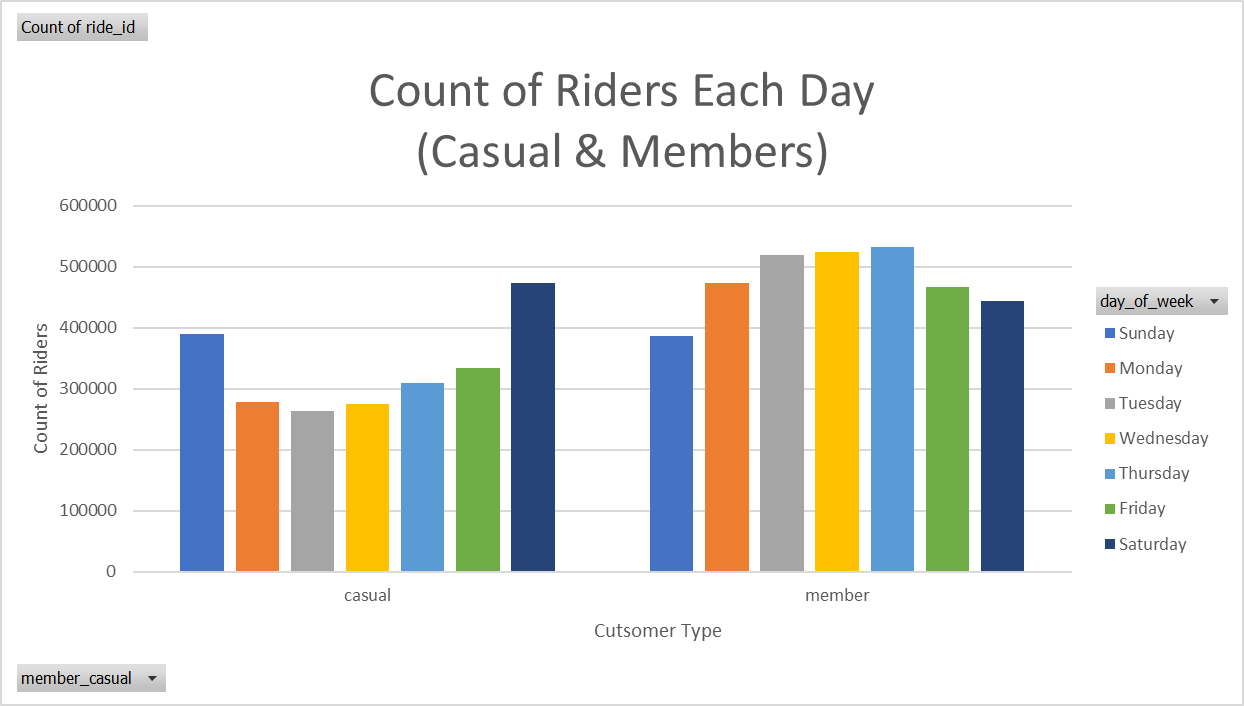




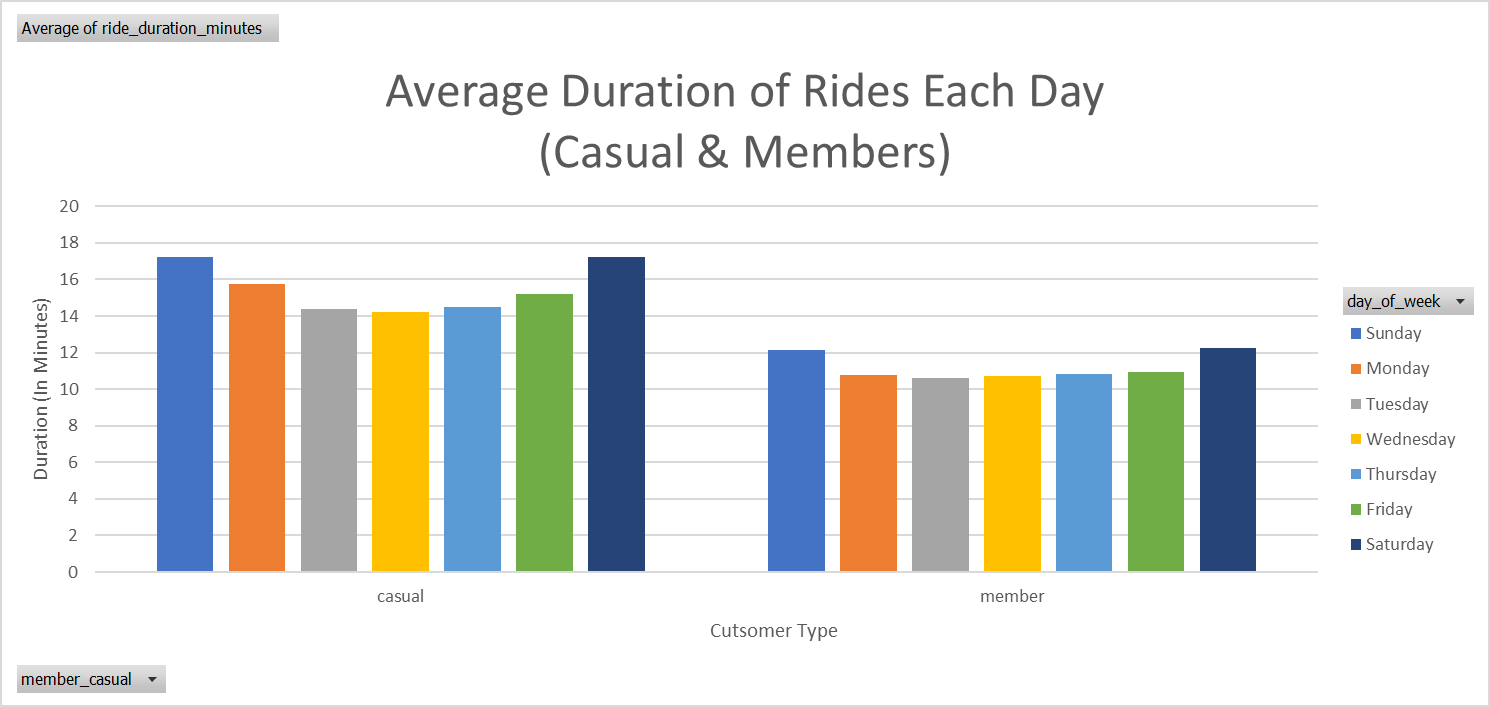


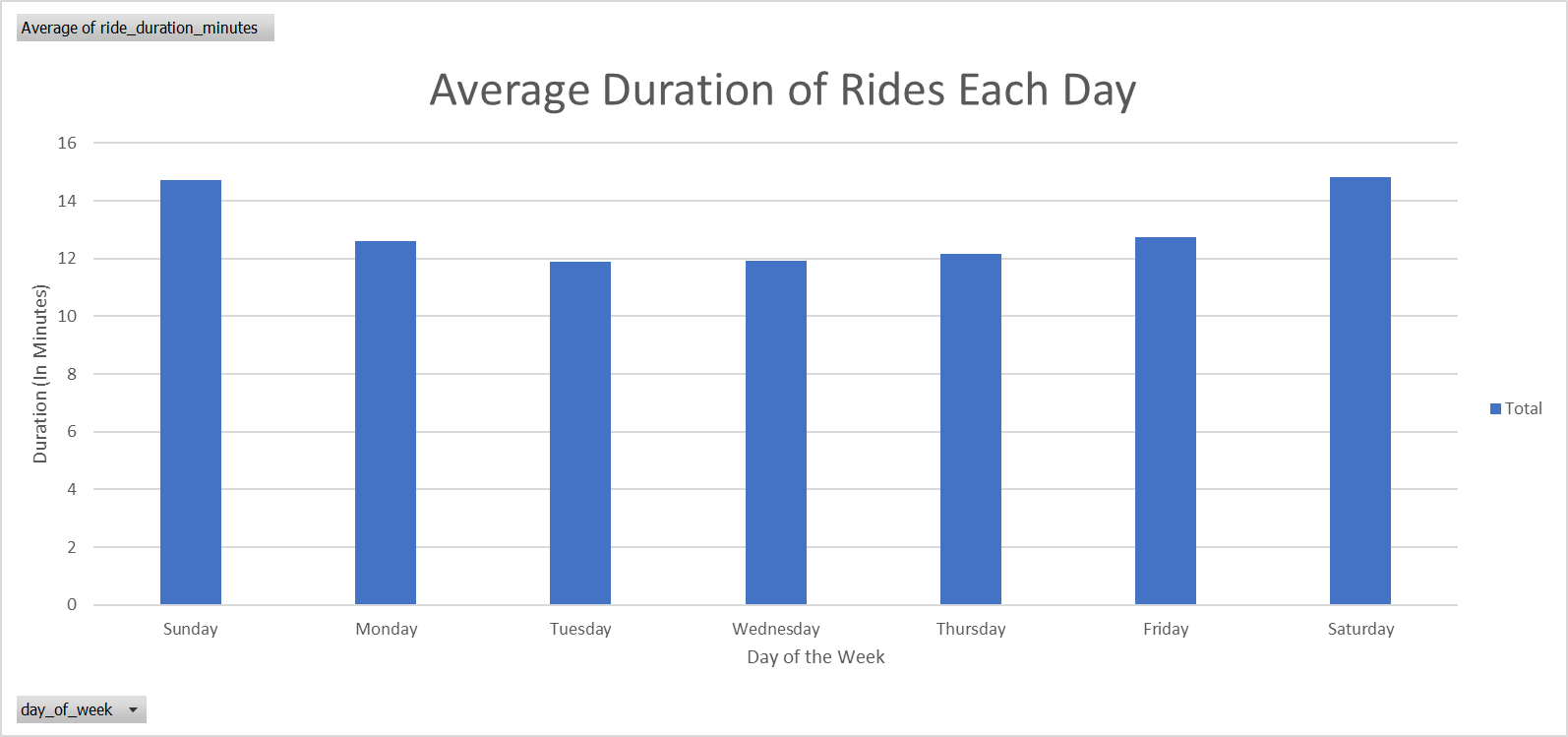
Rides on each day (members ride more during the week, possibly commuting, casuals ride more on the weekend. Another reason to market to casuals to start commuting)



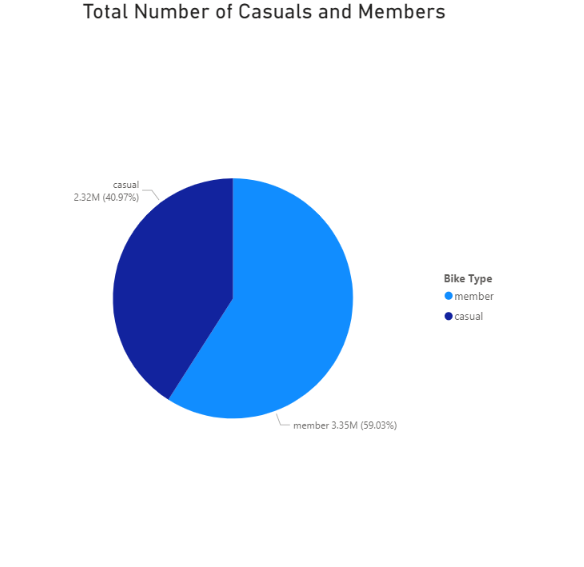


Average Ride duration by day of week (members take longer rides on the weekend likely for recreation, market to take weekend rides)

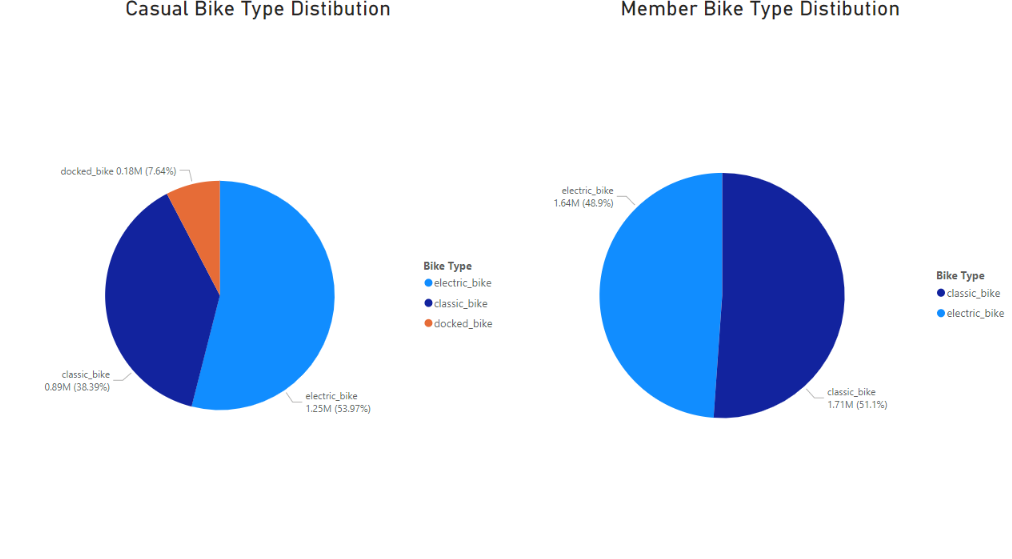




Casual vs Member



Bike Type Distribution



**Act**

Recommendations

1. Make advertisements to show casual riders that they can use Cyclistic to commute during the week and not just for recreation.
2. Target ads during quarter 3 of the year. Especially in July.
3. Target ads in the locations where casual members ride the most

**Extras**

Elevator pitch: “I used data from a bike-sharing service to find out how to turn casual customers into repeat members. I developed an action plan based on those findings to bring in more annual members.”