Google Data Analytics Course Capstone Project

I just completed the Google Data Analytics Course from Coursera and would like to share my process and what I have learned from the program. For the project, I will be using the case study provided by the course. For data analysis, I will be using the R programming language.

Background
Cyclistic is a bike-share company in Chicago founded in 2016. The company has a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bicycles can be unlocked from one station and returned to any other station in the system anytime.
The company's customers are classified in two segments: casual riders and members. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Ask
The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members.

Prepare
For this project, I will be using the dataset of January to December of 2022 provided by Motivate International Inc. https://divvy-tripdata.s3.amazonaws.com/index.html. Due to data privacy issues, some riders' personally identifiable information such as credit card is not available. This will limit our ability to determine if casual riders live in the serviced area or if they have purchased multiple single passes.

Process
First I loaded all the 12 months data frames in RStudio and checked that all the columns are consistent. After that I combined the data frames of all 12 months into a single data frame for the complete year.
The data frame contain 13 columns:
1. ride_id: unique id of each ride trip
2. rideable_type: type of bicycle: classic, docked and electric
3. started_at: date and time of the start of the trip
4. ended_at: date and time of the end of the trip
5. start_station_name: starting station name
6. start_station_id: starting station id
7. end_station_name: ending station name
8. end_station_id: ending station id
9. start_lat: latitude of the starting location
10. start_lng: longitude of the starting location
11. end_lat: latitude of the ending location
12. end_lng: longitude of the ending location
13. member_casual: type of riders: casual or member

Analyze
To analyze the data, I first cleaned the data frame a little bit by removing some irrelevant columns, added extra columns by breaking down the data in the date column, created a column with the calculated length of each trip and checked that negative and incomplete data are removed.
I also calculated the mean, median, max and min from the data frame.

Share
After having all the data cleaned and organized, I created some data visualizations. First I group the data by type of riders for each day of the week. Then we added the number of rides for each day and calculated the average duration of the rides for each day of the week.

For the first graphic, we plug the weekday on the x-axis and the number of rides in the y-axis.
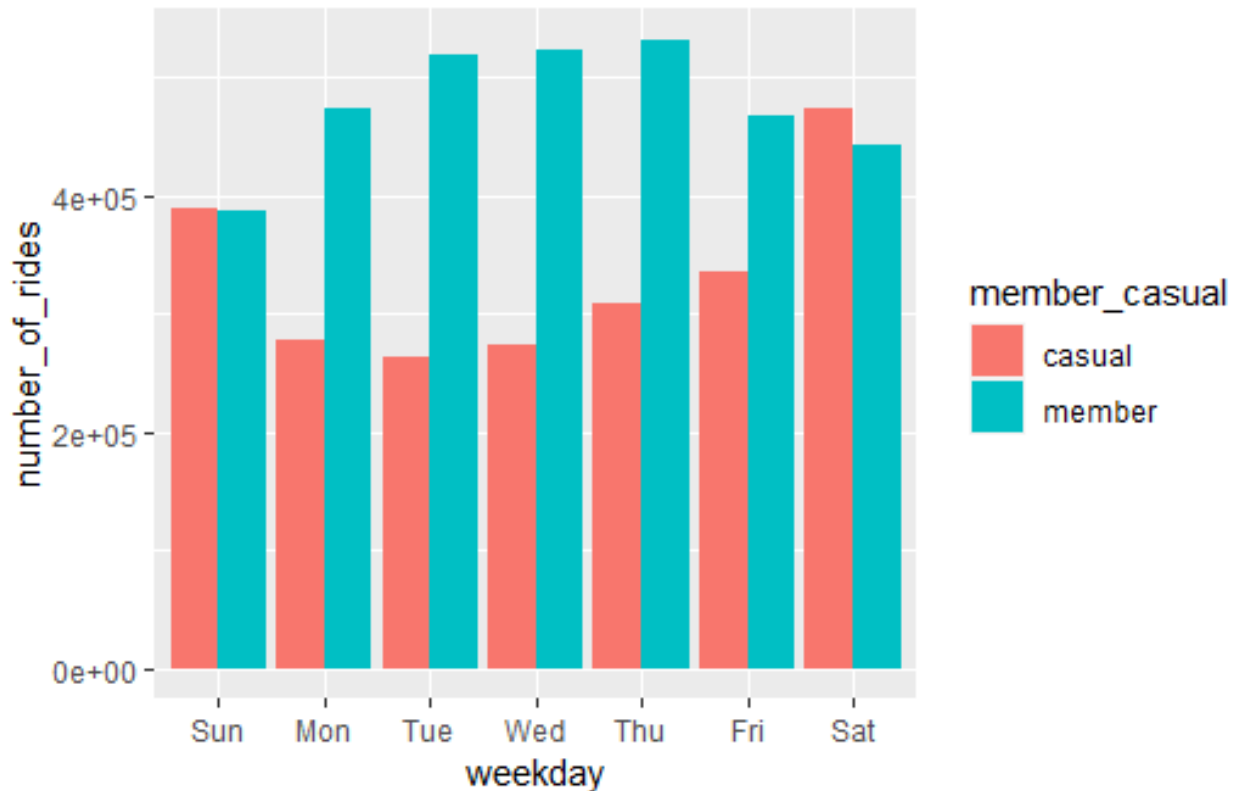
This graphic shows that members tend to ride more from Monday through Friday, while casual riders tend to ride more on Saturday and Sunday.

For the next graphic, we plug the weekday on the x-axis and the average duration of the rides in the y-axis.
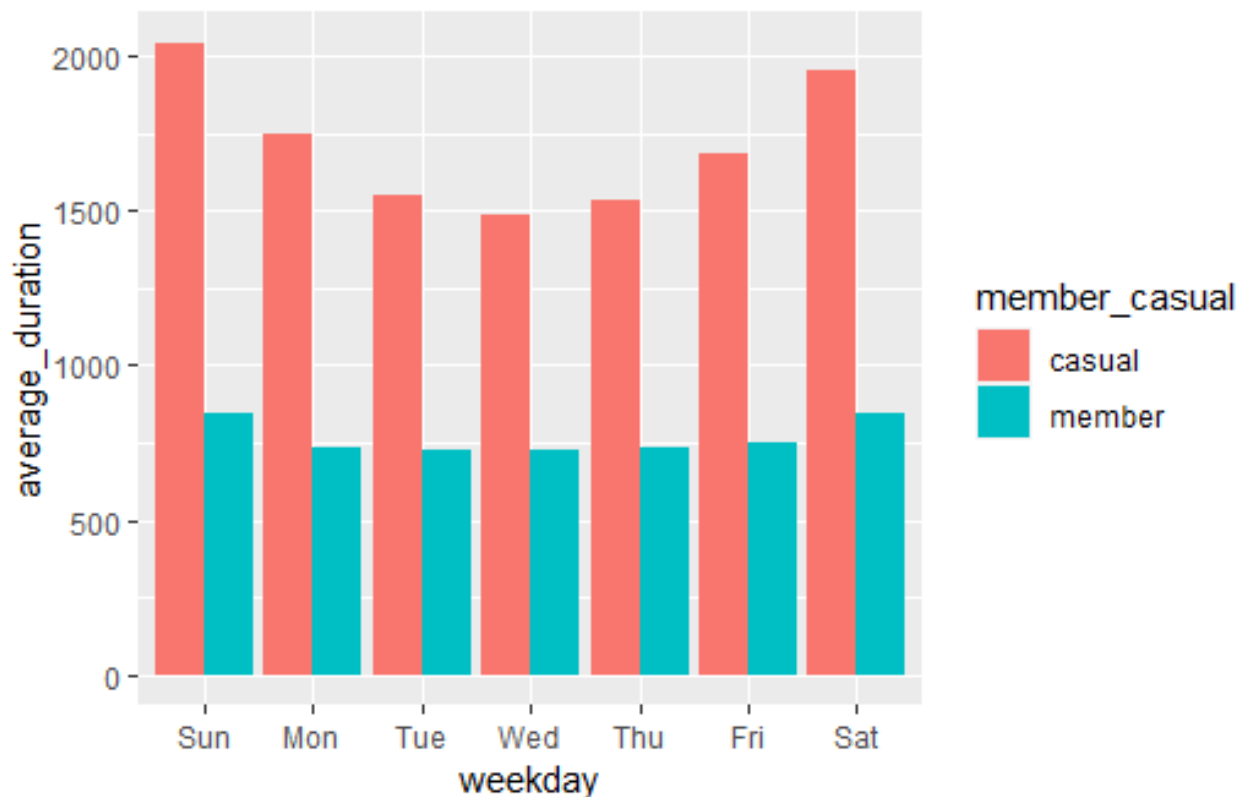
This graphic shows that casual riders' duration is about double the duration of the member's rider duration.

Act
Based on the data, we cannot give a recommendation that converting casual riders into members might be the company's right strategy. We need to get more information on the casual riders. Information such as: is the casual rider a local or a tourist. Or is the casual rider a first time user that is not familiar with the service or doesn't know how to return the bike. All of these will need further study to get a more precise recommendation.