**Introduction**

Welcome to my Cyclistic bike-share analysis project. In this endeavor, I'll be taking on the role of a junior data analyst, working for the fictional company Cyclistic. Throughout this project, I'll have the opportunity to tackle real-world data analysis tasks and interact with various characters and team members.

My main objective is to address critical business questions by following the data analysis process. This process involves asking insightful questions, preparing and processing data, analyzing findings, sharing valuable insights, and ultimately taking action.

By the end of this project, I'll have a case study that demonstrates my analytical skills and knowledge. I'm excited to dive into the world of Cyclistic bike-share data analysis and can't wait to share my findings with you as we progress through this fascinating journey!

**Scenario**

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. 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. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Shareholders: Director of Marketing (Lily Moreno), the Marketing Analytics team, and the Executive team.

**Ask**

Three questions will guide the future marketing program:

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

2. Why would casual riders buy Cyclistic annual memberships?

3. How can Cyclistic use digital media to influence casual riders to become members?

Lily Moreno has assigned you the first question to answer: **How do annual members and casual riders use Cyclistic bikes differently?**

Deliverables:

1. A clear statement of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top three recommendations based on your analysis

Business Task: Analyze the Cyclistic trip data from the last six months to identify trends in how each customer type uses Cyclistic bikes.

**Prepare**

For this project, I will be conducting an analysis of Cyclistic's historical trip data to identify trends and extract insights. The dataset has been obtained and is available for cleaning and analysis.

I have chosen to complete an analysis on the first two quarters of 2022 and adhere to strict data privacy regulations while working with this publicly available dataset. My primary focus is on discerning overarching trends and patterns that can inform Cyclistic's strategic decisions while safeguarding user privacy.

Guiding questions

* The data has been made available by Motivate International Inc. under this [link](https://divvy-tripdata.s3.amazonaws.com/index.html).
* It is organized in CSV (comma-separated values) format, comprising a total of 13 columns.
* However, it's important to acknowledge potential issues related to bias, credibility, and data consistency. Notably, there are inconsistencies in column names, where some sheets refer to customers as "Customer" and "Subscriber," while others use "Member" and "Casual."
* Additionally, there are discrepancies in data standards across various sheets. Some sheets contain more rows than others, which may impact data completeness.
* To ensure data integrity, I have taken several measures, including removing duplicates, implementing routine backups, and controlling access rights to the data. These steps aim to maintain data accuracy and consistency throughout the analysis process.

Key tasks

1. Download data and store it appropriately.

2. Identify how it’s organized.

3. Sort and filter the data.

4. Determine the credibility of the data.

Deliverable: All the data sources utilized for this analysis are internal and have been provided by Motivate International Inc. These sources consist of data available through links, stored in CSV format and excel sheets.

**Process**

Guiding questions

* I've employed a combination of tools for data analysis, specifically Excel for preliminary cleaning and R for in-depth analysis. The choice of Excel initially allows for efficient data cleaning and organization, while R is utilized for more complex analysis tasks due to its robust statistical capabilities and flexibility in handling large datasets.
* Data integrity has been a paramount concern throughout the project, and I've taken measures to ensure that it adheres to the ROCCC principles, meaning it is Reliable, Original, Comprehensive, Current, and Cited. This ensures that the data can be trusted for accurate and valid analysis, and that any findings or insights derived from it are based on a solid foundation of reliable information.

Key tasks

1. Check the data for errors.

2. Choose your tools.

3. Transform the data so you can work with it effectively.

4. Document the cleaning process.

Deliverable: Documentation of any cleaning or manipulation of data. See steps below

Cleaning Process

I began the process in Excel by…

* Organizing the data into cleaner columns
* Removing any duplicates in the data
* Renaming the columns so they were in a consistent structure
* Made sure the member\_type column only listed “Member” and “Casual” as types of riders
* Including columns for date, month, year, and week
* Included a column for ride length and removed the columns that had a negative value

I then moved my data into RStudio for further analysis…

* I installed and loaded the necessary packages

**install.packages("tidyverse")**

**install.packages("lubridate")**

* I imported the data

**jan\_2022 <- read\_csv("202201-divvy-tripdata.csv")**

**feb\_2022 <- read\_csv("202202-divvy-tripdata.csv")**

**mar\_2022 <- read\_csv("202203-divvy-tripdata.csv")**

**apr\_2022 <- read\_csv("202204-divvy-tripdata.csv")**

**may\_2022 <- read\_csv("202205-divvy-tripdata.csv")**

**jun\_2022 <- read\_csv("202206-divvy-tripdata.csv")**

* I then merged the specific months together in two datasets titled Q1\_Data and Q2\_Data. This allowed me to work with smaller datasets at a time and avoid any unnecessary software crashes

**Q1\_Data <- bind\_rows(jan\_2022, feb\_2022, mar\_2022)**

**Q2\_Data <-bind\_rows(apr\_2022, may\_2022, jun\_2022)**

**Analyze**

Guiding questions

* To prepare the data for analysis, I followed a structured approach. Firstly, I bound the rows to consolidate the information and ensure that all relevant data was accounted for.
* Next, I removed any unnecessary columns, keeping only those essential to the analysis, and omitted rows with missing or NA values for data completeness.
* Additionally, I ensured that the data was properly formatted before commencing the analysis, which is crucial for accurate and efficient data processing.
* These meticulous steps in data preparation lay the foundation for deriving meaningful insights. The insights obtained from this analysis will play a pivotal role in addressing the primary business question: "How do annual members and casual riders use Cyclistic bikes differently?"

Key tasks

1. Aggregate your data so it’s useful and accessible.

2. Organize and format your data.

3. Perform calculations.

4. Identify trends and relationships.

Deliverable: A summary of your analysis. See Analysis Process below.

Analysis in R

* Started by calculating the mean, median, maximum, and minimum ride\_length for each quarter

**mean(Q1\_Clean.Data$ride\_length[Q1\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**mean(Q2\_Clean.Data$ride\_length[Q2\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**median(Q1\_Clean.Data$ride\_length[Q1\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**median(Q2\_Clean.Data$ride\_length[Q2\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**max(Q1\_Clean.Data$ride\_length[Q1\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**max(Q2\_Clean.Data$ride\_length[Q2\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**min(Q1\_Clean.Data$ride\_length[Q1\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**min(Q2\_Clean.Data$ride\_length[Q2\_Clean.Data$ride\_length > 0], na.rm = TRUE)**

**Q1\_Clean.Data %>% filter(!is.na(ride\_length) & ride\_length > 0) %>% group\_by(member\_casual) %>% summarise( average\_ride\_length = mean(ride\_length), median\_length = median(ride\_length), max\_ride\_length = max(ride\_length), min\_ride\_length = min(ride\_length) )**

A tibble: 2 × 5

member\_casual avg\_ride\_len, med\_length, max\_ride\_length, min\_ride\_length

*<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 casual 2082.34 807 2061244 1

2 member 717.12 473.98 93594 1

**Q2\_Clean.Data %>% filter(!is.na(ride\_length) & ride\_length > 0) %>% group\_by(member\_casual) %>% summarise( average\_ride\_length = mean(ride\_length), median\_length = median(ride\_length), max\_ride\_length = max(ride\_length), min\_ride\_length = min(ride\_length) )**

A tibble: 2 x 5

| member\_casual | avg\_ride\_len, | median\_length | max\_ride\_length | min\_ride\_length |
| --- | --- | --- | --- | --- |
| casual | 2018.77 | 911 | 2175468 | 1 |
| member | 802.7559456 | 566 | 89998 | 1 |

* Grouped the members and casual riders into two groups to visualize the total rides taken for each type

**result <- Q1\_Clean.Data %>%**

**drop\_na(member\_casual) %>% # Remove rows with NA values in 'member\_casual'**

**group\_by(member\_casual) %>%**

**summarise(ride\_count = n())**

member\_casual ride\_count

casual 107355

member 313978

**result <- Q2\_Clean.Data %>%**

**drop\_na(member\_casual) %>% # Remove rows with NA values in 'member\_casual'**

**group\_by(member\_casual) %>%**

**summarise(ride\_count = n())**

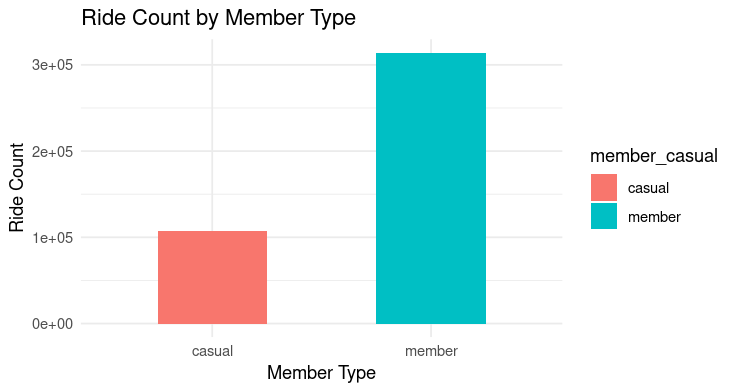
member\_casual ride\_count

casual 668784

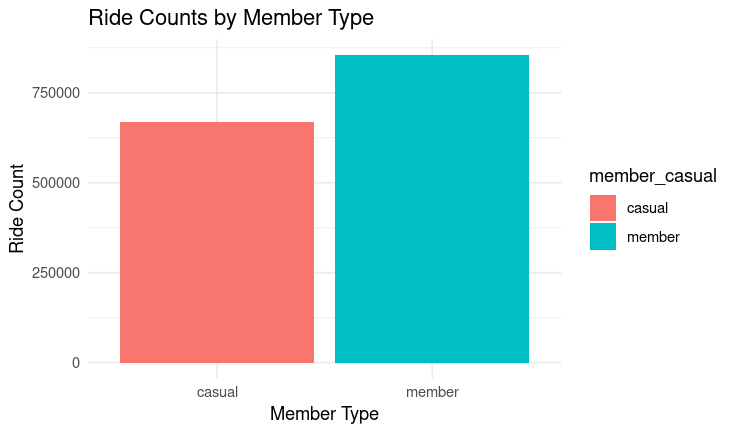
member 855988

**ggplot(result, aes(x = member\_casual, y = ride\_count, fill = member\_casual)) + geom\_bar(stat = "identity", position = "dodge", width = 0.5) + labs( title = "Ride Count by Member Type", x = "Member Type", y = "Ride Count" ) + theme\_minimal()**

Q1\_2022



Q2\_2022

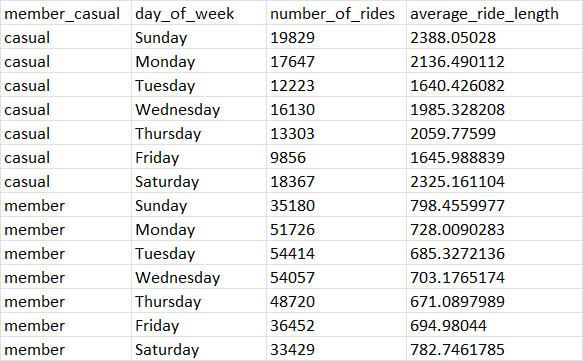


* Made sure to aggregate the data in each quarter

**aggregate(Q1\_Clean.Data$ride\_length ~ Q1\_Clean.Data$member\_casual + Q1\_Clean.Data$day\_of\_week, FUN = mean)**

**Q1\_Clean.Data$day\_of\_week <- ordered(Q1\_Clean.Data$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))**

**aggregate(Q1\_Clean.Data$ride\_length ~ Q1\_Clean.Data$member\_casual + Q1\_Clean.Data$day\_of\_week, FUN = mean)**



**aggregate(Q2\_Clean.Data$ride\_length ~ Q2\_Clean.Data$member\_casual + Q2\_Clean.Data$day\_of\_week, FUN = mean)**

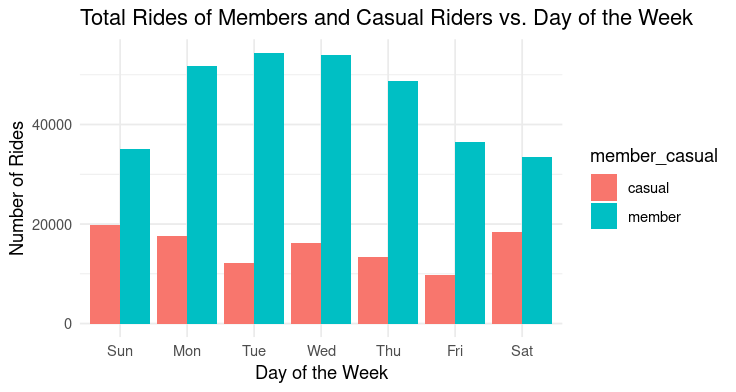
**Q2\_Clean.Data$day\_of\_week <- ordered(Q2\_Clean.Data$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))**

**aggregate(Q2\_Clean.Data$ride\_length ~ Q2\_Clean.Data$member\_casual + Q2\_Clean.Data$day\_of\_week, FUN = mean)**

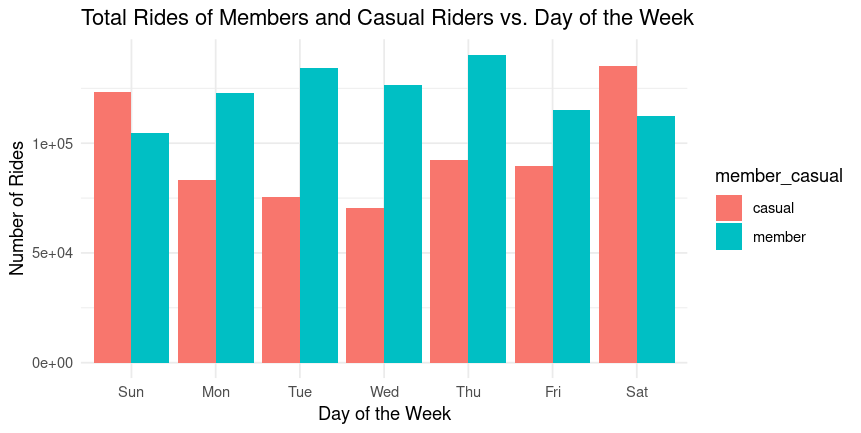
* Analyzed the total rides and average ride time by each day for members vs casual riders for each quarter

**Q1\_Clean.Data %>% filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length' mutate(weekday = wday(started\_at, label = TRUE)) %>% group\_by(member\_casual, weekday) %>% summarise( number\_of\_rides = n(), average\_duration = mean(ride\_length) ) %>% arrange(member\_casual, weekday) %>% ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) + geom\_col(position = "dodge") + labs( title = "Total Rides of Members and Casual Riders vs. Day of the Week", x = "Day of the Week", y = "Number of Rides" ) + theme\_minimal()**

Q1\_2022



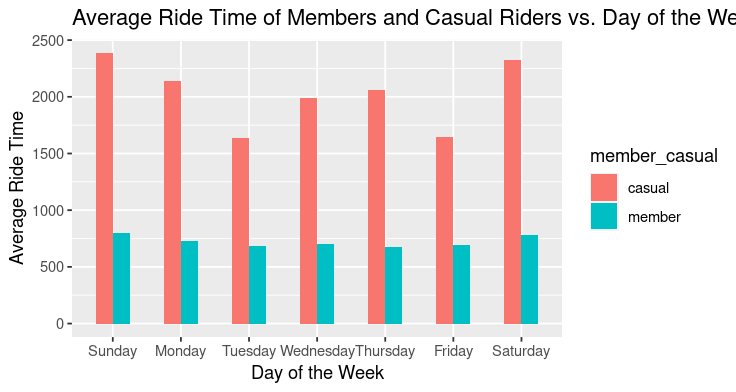
**Q2\_Clean.Data %>% filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length' mutate(weekday = wday(started\_at, label = TRUE)) %>% group\_by(member\_casual, weekday) %>% summarise( number\_of\_rides = n(), average\_duration = mean(ride\_length) ) %>% arrange(member\_casual, weekday) %>% ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) + geom\_col(position = "dodge") + labs( title = "Total Rides of Members and Casual Riders vs. Day of the Week", x = "Day of the Week", y = "Number of Rides" ) + theme\_minimal()**

Q2\_2022

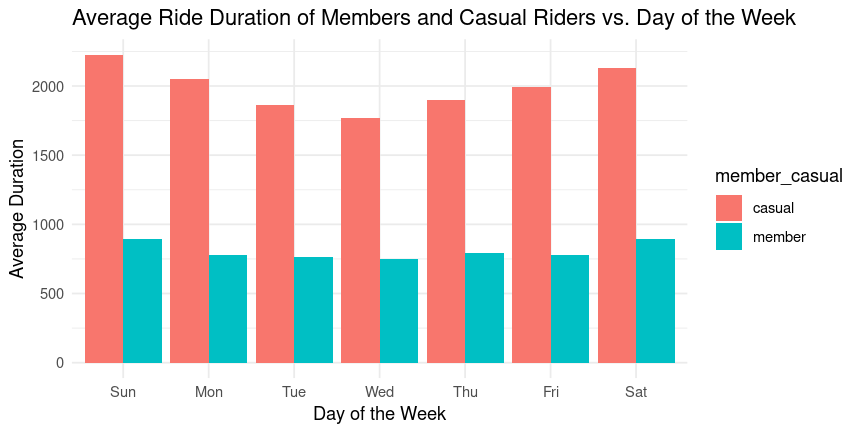
* Analyzed the average ride by day of the week for each quarter

**Q1\_Clean.Data %>% filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length' group\_by(member\_casual, day\_of\_week) %>% summarise( average\_ride\_length = mean(ride\_length), .groups = "drop" ) %>% ggplot(aes(x = day\_of\_week, y = average\_ride\_length, fill = member\_casual)) + geom\_col(width = 0.5, position = position\_dodge(width = 0.5)) + labs( title = "Average Ride Time of Members and Casual Riders vs. Day of the Week (Excluding NA Values)", x = "Day of the Week", y = "Average Ride Time" )**

Q1\_2022



**Q2\_Clean.Data %>% filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length' group\_by(member\_casual, day\_of\_week) %>% summarise( average\_ride\_length = mean(ride\_length), .groups = "drop" ) %>% ggplot(aes(x = day\_of\_week, y = average\_ride\_length, fill = member\_casual)) + geom\_col(width = 0.5, position = position\_dodge(width = 0.5)) + labs( title = "Average Ride Time of Members and Casual Riders vs. Day of the Week (Excluding NA Values)", x = "Day of the Week", y = "Average Ride Time" )**

Q2\_2022

* Analyzed the total rides taken by members and casuals by month

**Q1\_Clean.Data %>%**

**filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length'**

**mutate(weekday = wday(started\_at, label = TRUE)) %>%**

**group\_by(member\_casual, weekday) %>%**

**summarise(**

**number\_of\_rides = n(),**

**average\_duration = mean(ride\_length)**

**) %>%**

**arrange(member\_casual, weekday) %>%**

**ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +**

**geom\_col(position = "dodge") +**

**labs(**

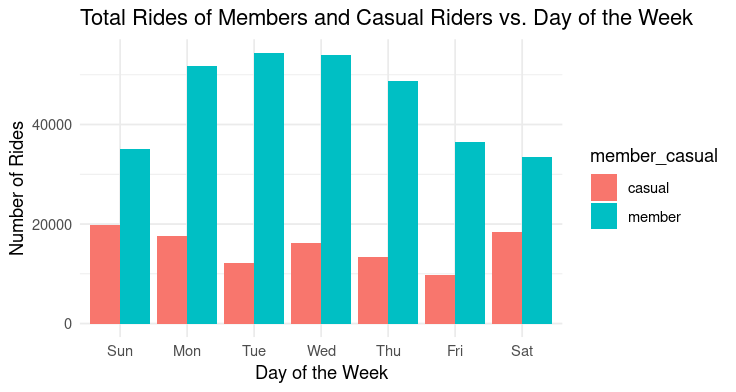
**title = "Total Rides of Members and Casual Riders vs. Day of the Week",**

**x = "Day of the Week",**

**y = "Number of Rides"**

**) +**

**theme\_minimal()**

Q1\_2022

**Q2\_Clean.Data %>%**

**filter(!is.na(ride\_length)) %>% # Remove rows with NA values in 'ride\_length'**

**mutate(weekday = wday(started\_at, label = TRUE)) %>%**

**group\_by(member\_casual, weekday) %>%**

**summarise(**

**number\_of\_rides = n(),**

**average\_duration = mean(ride\_length)**

**) %>%**

**arrange(member\_casual, weekday) %>%**

**ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +**

**geom\_col(position = "dodge") +**

**labs(**

**title = "Total Rides of Members and Casual Riders vs. Day of the Week",**

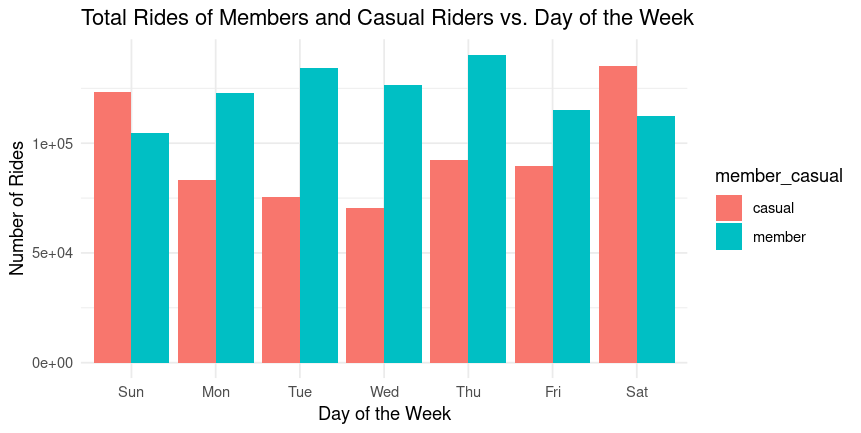
**x = "Day of the Week",**

**y = "Number of Rides"**

**) +**

**theme\_minimal()**

Q2\_2022



**Share**

Indeed, I was able to answer the question regarding how annual members and casual riders utilize Cyclistic bikes differently.

The data reveals distinct patterns:

* members tend to take more rides
* casual riders embark on longer journeys in terms of both distance and duration
* members ride less frequently on weekends.

These findings directly align with the original inquiry into the differing usage behaviors of these two customer segments.

In terms of audience, my stakeholders include the Director of Marketing, Lily Moreno, the Marketing Analytics team, and the Executive team. To effectively communicate these insights, I plan to deliver a professional PowerPoint presentation. This format allows for concise yet comprehensive sharing of the analysis, ensuring that the findings are accessible and easily understood.

Data visualization will certainly play a pivotal role in conveying the results. It serves as a powerful means to visually represent the data trends and patterns, enhancing the clarity and impact of the presentation.

Furthermore, the presentation is designed to be accessible to the audience, walking them through each step of the analysis, including the methodologies, code used, and the visuals employed to arrive at the findings. This comprehensive approach ensures that the audience can follow and engage with the analysis process, making it an effective way to share the results and insights derived from the data analysis.

**Act**

The "Act" phase involves taking action based on the insights and trends uncovered through data analysis, specifically addressing the guiding question. During this phase, I've formulated three recommendations aimed at assisting the marketing team in their efforts to convert casual riders into registered members.

**Weekend Membership Promotion:** Launch a targeted marketing campaign that offers special weekend membership promotions. Highlight the benefits of becoming a member, such as unlimited weekend rides at a discounted rate.

**Ride Rewards Program:** Implement a loyalty program specifically designed for casual riders. Offer rewards or discounts for reaching milestones in terms of ride duration or frequency. This encourages casual riders to consistently use the service and gradually transition into becoming members to unlock more substantial benefits.

**Personalized Recommendations:** Utilize data analytics to provide personalized recommendations to casual riders based on their riding behavior. Send targeted emails or notifications suggesting the benefits of membership tailored to their preferences.