How does a bike-share navigate speedy success?

Bashiru Mukaila

2024-04-19

1 Introduction

This capstone project is part of the Google Data Analytics program, where I am tasked with assuming the role of a junior data analyst recently joining Cyclistic's marketing analytics team in Chicago. Cyclistic is a bike-share company, and our team of data analysts focuses on collecting, analysing, and reporting data to inform the company's marketing strategy.

Under the guidance of Lily Moreno, the director of marketing and my manager, our team aims to enhance Cyclistic's future success by increasing the number of annual memberships. Financial analysis has shown that annual members contribute significantly more to profitability compared to casual riders. Consequently, our objective is to explore the differences in behaviour between casual riders and annual members to inform the development of a targeted marketing strategy aimed at converting casual riders into annual members.

The report is structured as follows: Section 2 outlines the specific business tasks, Section 3 provides details on the data sources used, Section 4 describes the data cleaning and manipulation process, Section 5 presents a summary of the analysis accompanied by supporting visualizations, section 6 highlight the key findings, while Section 7 offers recommendations based on the analysis and potential future work.

2 Clear Statement of the Business Task

The primary objective of this project is to investigate how annual members and casual riders use Cyclistic bikes differently.

The secondary objectives is to explore the following questions:

1. Why would casual riders buy Cyclistic annual memberships?
2. How can Cyclistic use digital media to influence casual riders to become members?

3 The Description of All Data Sources Used

The data for this project is the **last twelve months Cyclistic's historical trip data** between April 2023 and March 2024. This is public data that can be used to explore how different customer types are using Cyclistic bikes. But the data-privacy issues prohibit us from using riders' personally identifiable information. This means that we won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

The data source is reliable and verified to be original. It provides comprehensive insights into how various customer types use Cyclistic bikes. Additionally, the data is current and relevant, being released monthly. This data is provided by Motivate International, Inc. under this **license** and it can be accessed **here**.

The data was downloaded as individual monthly files in CSV format, thus requiring consolidation into a single file for analysis.

Each of the monthly files has the following columns:

1. **Ride id:** This shows the unique number giving to each trip.
2. **Rideable type:** This specify the type of bike used during trip which are either classic bike, docked bike or electric bike.
3. **Started at:** This represents the starting time for the trip.
4. **Ended at:** This shows the end time of the trip.
5. **Start station name:** This is where the trip started.
6. **Start station id:** This is the unique identification code giving to the start station.
7. **End station name:** This is where the trip ended.
8. **End station id:** This is the unique identification code giving to the end station.
9. **Start lat:** This is the latitude coordinate of where the trip started.
10. **Start lng:** This is the longitude coordinate of where the trip started.
11. **End lat:** This is the latitude coordinate of where the trip ended.
12. **End lng:** This is the longitude coordinate of where the trip ended.
13. **Member or casual:** This specify the customer type which can either be member or casual rider.

4 Data Cleaning and Manipulation Process

We have chosen R as our primary tool for this project for the following reasons:

1. R's robust capabilities make it ideal for handling large data sets like the last twelve months of Cyclistic's historical trip data.
2. Given the need for thorough data aggregation, cleaning, and manipulation, R offers a variety of powerful packages tailored for effective data manipulation and analysis.
3. To explore bike usage trends through compelling visualizations, R provides an extensive selection of packages for generating high-quality, customizable plots and graphs.
4. Leveraging R Markdown, we will effortlessly produce the required project report, ensuring it's easily reproducible for anyone interested.

The following steps were taken to ensure that the data is clean and ready for exploration:

4.1 Step 1: Collect Data

The historical trip data for Cyclistic from April 2023 to March 2024, spanning the last twelve months, was imported into R as individual monthly files.

4.2 Step 2: Wrangle Data and Combine into a Single File

We cross-checked the column names across all files. Although the order of column names may vary, they must match exactly before merging them into a single file. After this comparison, we confirmed that all files share identical column names, as outlined in section 3. Consequently, we merged the individual monthly data frames into one comprehensive data frame.

We identified formatting issues in the **ride id** and **rideable type** columns, which we resolved by converting them into character strings. To ensure clarity of name, we rename **member_casual** column to **user_type**. Additionally, we eliminated redundant latitude and longitude columns, as they are unnecessary variables for this project.

4.3 Step 3: Clean Up and Add Data to Prepare for Analysis

Upon completing step 2, we conducted an inspection of the merged data, revealing a total of 5,750,177 observations with a reduction in column names to 9. Subsequent checks were performed to address identified issues:

1. We validated the **User_type** column, which ideally should only contain two categories: member and casual. This validation aligned with the information provided by the business.

2. Recognizing that the data granularity was too limited to the ride-level, we opted to enhance it by incorporating additional columns such as day, month, and year. This augmentation offers broader opportunities for data aggregation.

3. To enrich the dataset, we introduced a calculated field denoting the length of each ride, labeled **ride_length**.

4. During our scrutiny, we observed instances of negative ride lengths, zero ride lengths and trips with missing information. To maintain data integrity, we eliminated these anomalies from the dataset.

5. We also set the maximum ride length to 5,400 seconds (1 hour and 30 minutes) which we consider a reasonable duration to ride bike either for work or for leisure. We have regarded ride length higher than this as an outlier.

Following the aforementioned steps, our final data set comprises 4,307,385 observations. This figure reflects the removal of 1,442,792 observations containing missing information, inaccurately reported times and outliers.

The relevant code for this task can be found within the code chunk of this document's RMD file.

5 Summary of Data Analysis with Supporting Visualizations

5.1 Intial Observations

5.1.1 Summary of overall rides by User type

The donut chart in figure 1 illustrates the distribution of rides taken by both members and casual riders over the past twelve months, starting from April 2023 to March 2024. The total number of rides within this period amounted to 4,307,385, with members accounting for 65% and casual riders for 35% of the total rides. Despite members taking a significantly higher number of rides than casual riders, the box plot in figure 1 reveals that, on average, casual riders tend to have longer ride duration compared to members.

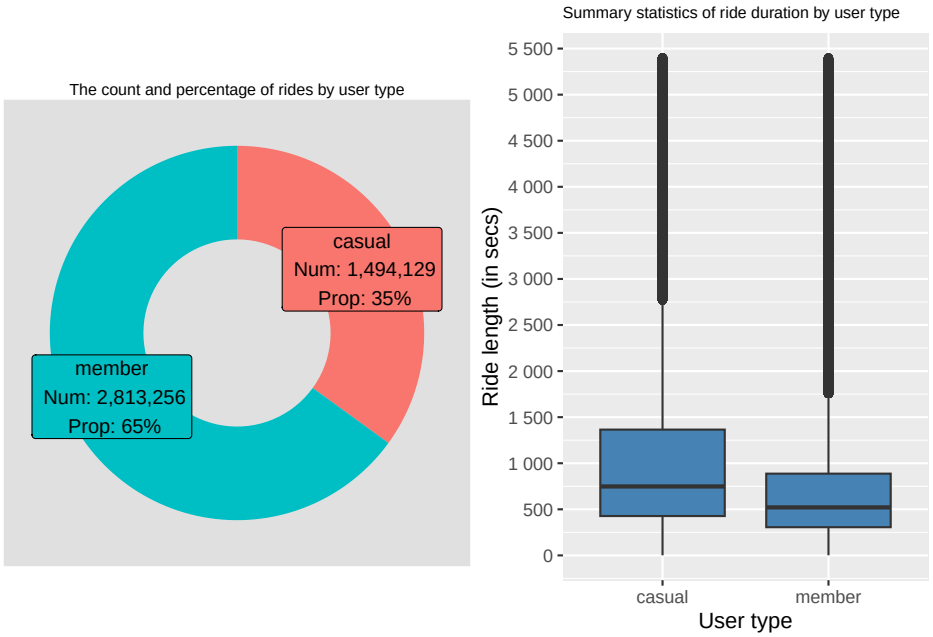

Figure 1: Summary of overall rides by User type.

5.1.2 Summary of rides by ride type for members

Of the 2,813,256 member rides recorded in the past twelve months, figure 2 shows that 66% were on classic bikes and 34% were on electric bikes. The average ride duration for both ride types among members is similar.

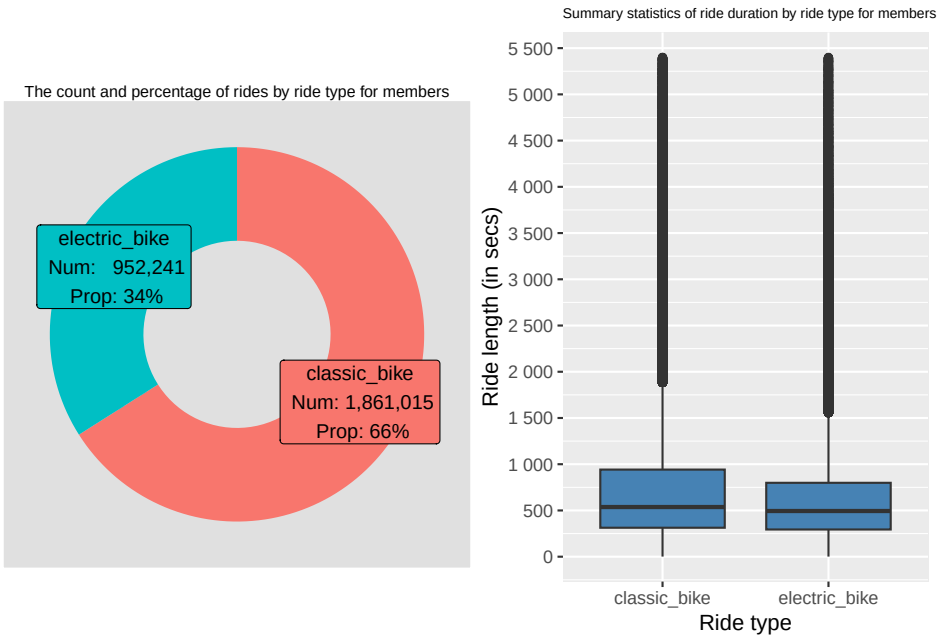

Figure 2: Summary of rides by ride type for members.

5.1.3 Summary of rides by ride type for casual riders

Out of the 1,494,129 rides recorded for casual riders in the past twelve months, 58% were on classic bikes, 38% were on electric bikes, and 4% were on docked bikes as shown in figure 3. The average ride duration for classic bikes, electric bikes, and docked bikes are 1,170.3, 850.825, and 1,809.43 seconds, respectively. After comparing the box plots in Figures 2 and 3, it is evident that casual riders, on average, have longer ride duration across all ride types compared to members.

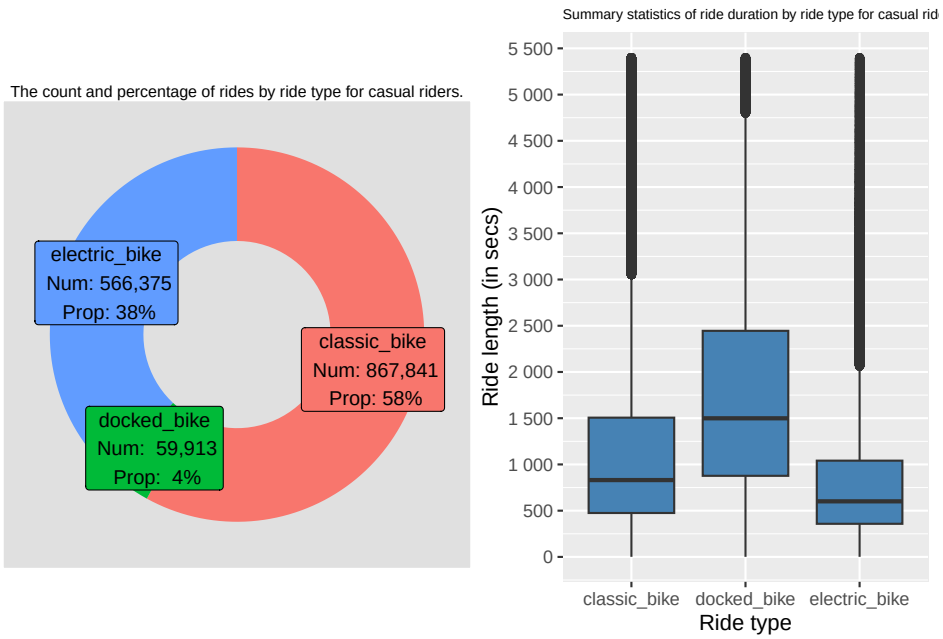

Figure 3: Summary of rides by ride type for casual riders.

5.2 Summary of rides by user type for each month

The lowest number of rides for both members and casual riders occurs in January, while the highest number of rides occurs in August for members and July for casual riders. Figure 4 demonstrates a general trend where the number of rides tends to peak during the summer months, likely due to the favorable weather conditions during this period.

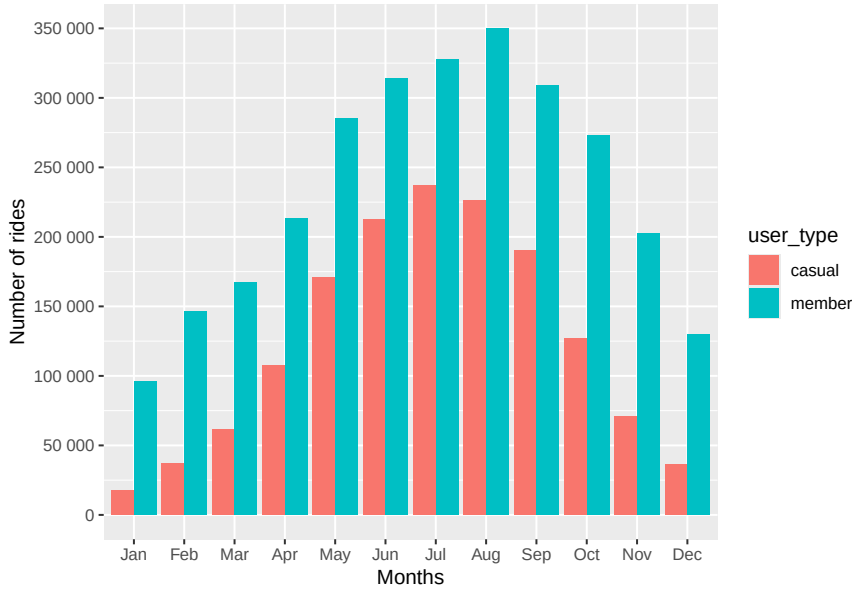

Figure 4: The number of rides by user type for each month.

The increased number of rides during the summer, as depicted in figure 4, also comes with longer period of rides during the summer as shown in figure 5 for both members and casual riders.

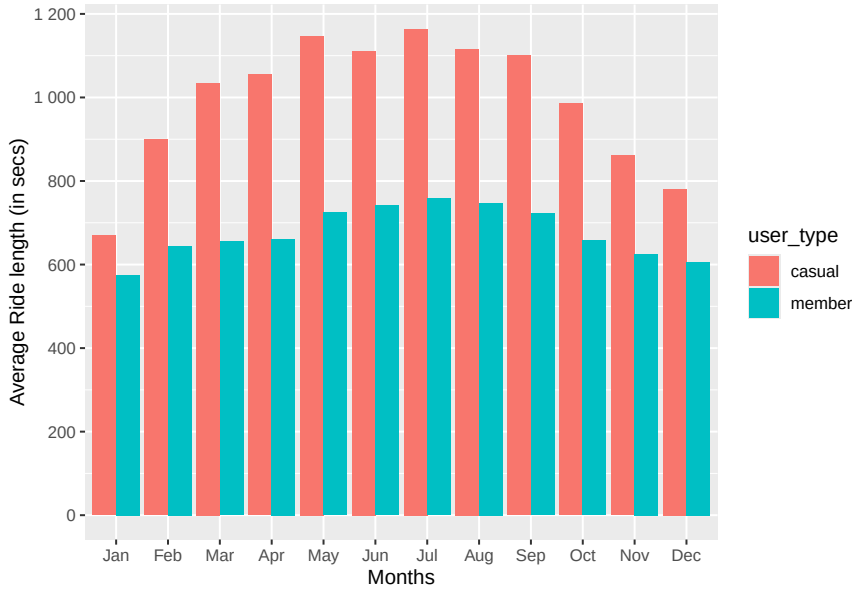

Figure 5: The average ride duration by user type for each month.

5.3 Summary of rides by user type for each day of the week

The number of rides between Monday and Friday in figure 6 is not significantly different for members and for casual riders. However, there is a notable increase in rides on Saturday and Sunday for casual riders, while the number of rides drops during the weekend for members. This suggests that casual riders may mostly use Cyclistic bikes for leisure purposes.

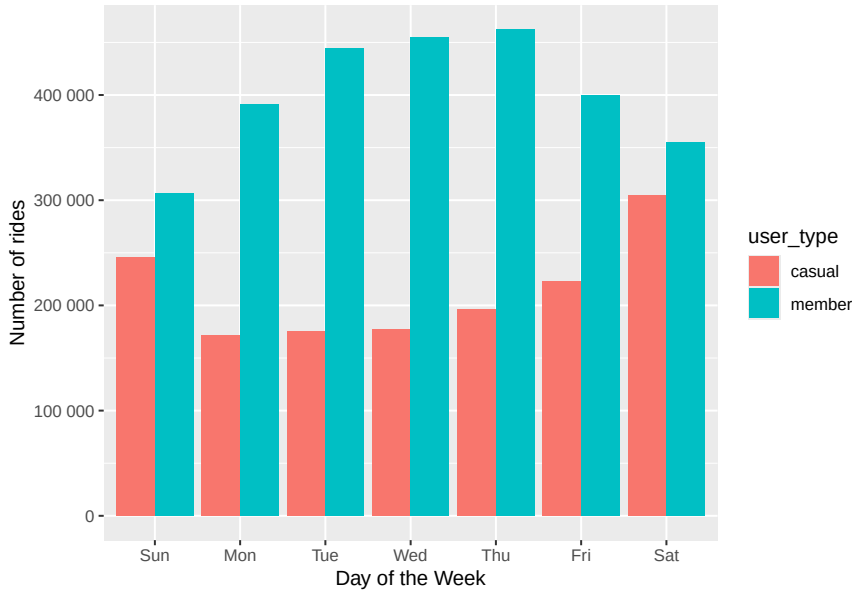

Figure 6: The number of rides by user type for each day of the week.

While the number of rides for members is higher than that of casual riders across the week, as shown in figure 6, figure 7 indicates that casual riders tend to use Cyclistic bikes for longer periods of time every day than the members.

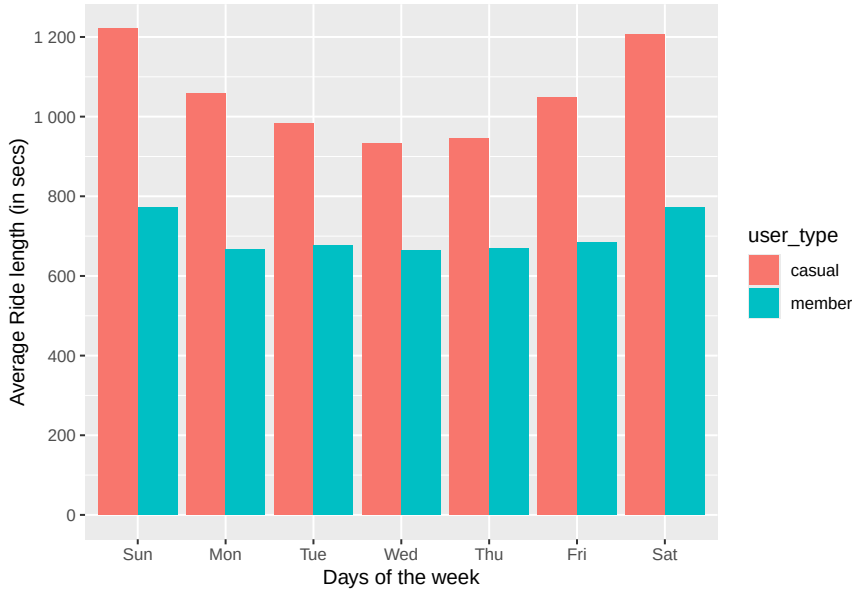

Figure 7: The average ride duration by user type for each day of the week.

5.4 Summary of rides by user type for each day of the month

Even though members and casual riders generally follow the same pattern of rides everyday, according to figure 8, we can see that the number of rides peaks for members on the 7th day of the month and for casual riders on the 4th day of the month. Both groups experience very low ride counts on the 31st day of the month.

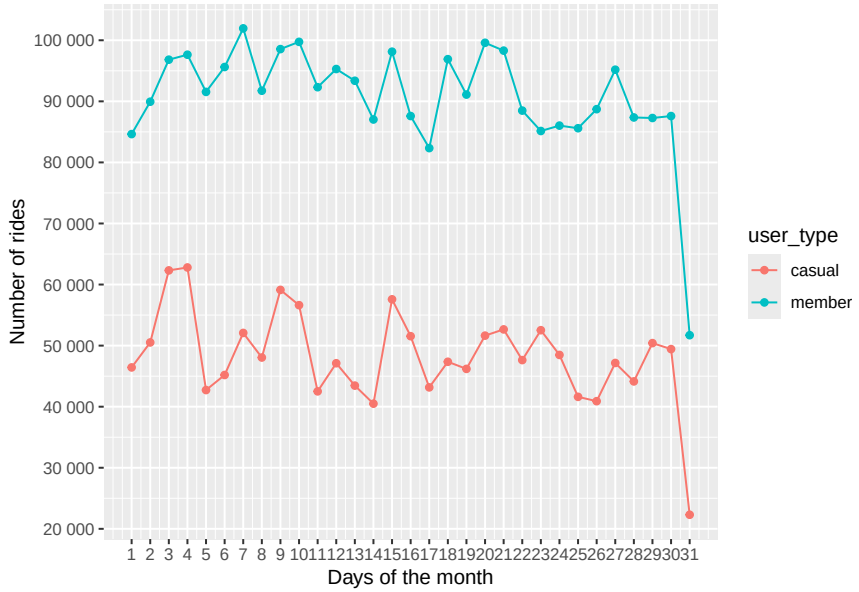

Figure 8: The number of rides by user type for each day of the month.

While member rides outnumber casual rider rides on a daily basis, on average, casual riders use Cyclistic bikes for longer duration on daily basis compared to members based on figure 9.

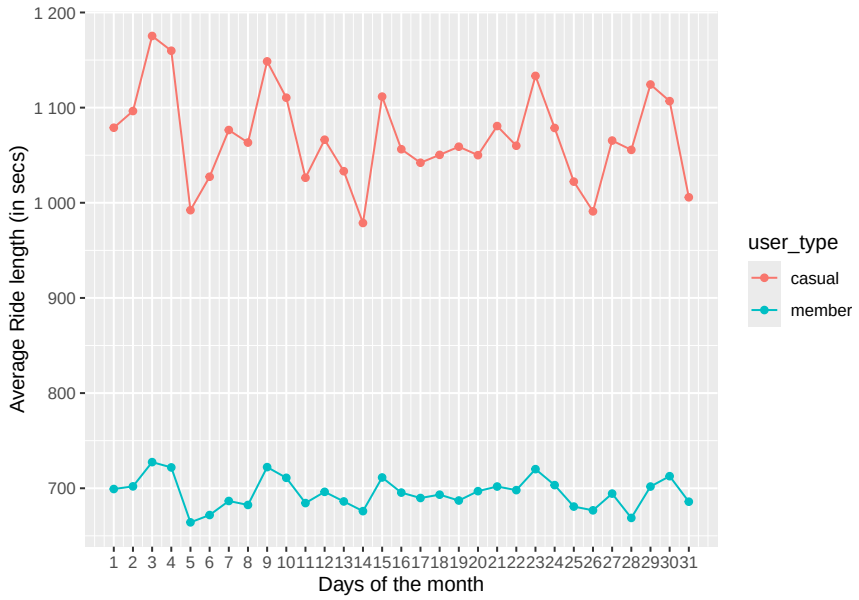

Figure 9: The average ride duration by user type for each day of the month.

5.5 Overview of rides by user type across different hours of day

Figure 10 indicates that both members and casual riders predominantly begin their trips at 5pm, suggesting that there is high usage of Cyclistic bikes after the close of business. Additionally, both groups experience their lowest activity levels between 1am and 5am, and this can be attributed to little or no activity during this periods.

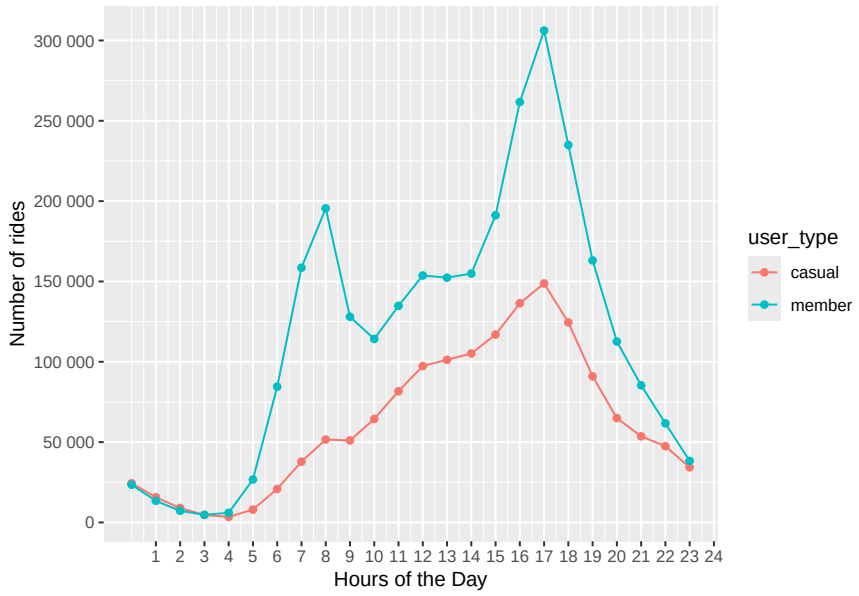

Figure 10: Number of rides by user type across different hours of the day.

For members, based on figure 11, the highest average ride duration occurs at 5pm, while for casual riders, it is between 11am and 2pm. This mean that in as much as there is high demand for Cyclistic's bike at close of business by members, there is also possibility that the bikes will be driven for longer time than usual.

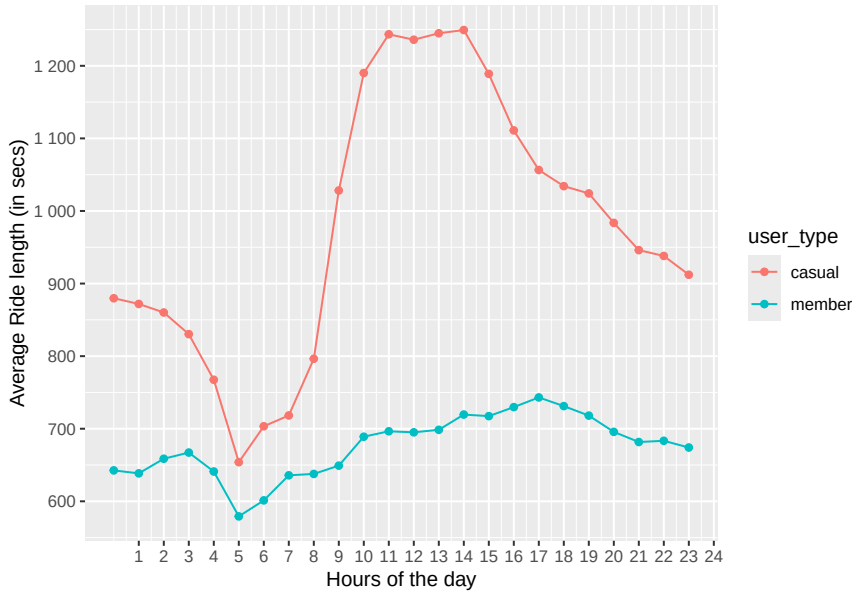

Figure 11: Average ride duration by user type across different hours of the day.

The relevant code for this task can be found within the code chunk of this document's RMD file.

6 Highlight of Key Findings

Below are the key findings:

1. Members constitute the largest proportion of Cyclistic bike riders across all levels.
2. Casual riders, regardless of ride type, month, week, day, or hour, typically use Cyclistic's bikes for longer duration on average than members. This underscores the importance for casual riders to invest in Cyclistic annual memberships, ensuring they receive maximum value for their money.
3. The classic bike is the preferred ride type for both members and casual riders.
4. There is an increase in rides during the summer months for both members and casual riders, with longer duration compared to other seasons.
5. The busiest day for Cyclistic bikes is Saturday for casual riders and Thursday for members.
6. Rush hour for Cyclistic bikes occurs at 5pm for both members and the casual riders.

7 Recommendations Based on the Analysis

Cyclistic can employ the following strategies to convert casual riders into members:

1. **Implementing a time limit per trip for casual packages:** Given that casual riders tend to ride for longer duration on average than members, Cyclistic can introduce time limits to casual packages to encourage registration as a member. For example, setting a time limit of 900 seconds (15 minutes) per trip for casual riders could prompt over 40% of them to opt for membership before their next ride.
2. **Introducing usage frequency limits for casual packages:** Cyclistic could restrict the number of times riders can use bikes through casual packages. For instance, if a casual rider uses Cyclistic bikes three times, with each ride duration being less than 900 seconds (15 minutes), they would be prompted to register as a member for their fourth ride. This approach would encourage prospective regular casual riders to become members.
3. **Offering a flexible payment plan exclusively for casual riders converting to members:** Alongside the above strategies, Cyclistic could introduce a flexible payment plan allowing casual riders converting to members to pay their annual membership fees in installments. This initiative would motivate casual riders who may find it challenging to afford the annual membership fee in one payment.

4. **Expanding the scope of advertising:** Cyclistic can intensify the usage of digital platforms to showcase benefits and testimonies from satisfied current annual members to attract more casual riders to become members.

Potential future work may involve:

1. Providing insights into the personal information of riders, including their full names, to determine whether casual riders have purchased multiple passes or only single passes. This analysis would enable us to quantify the percentage of casual riders who could potentially be converted by introducing limits on usage frequency.
2. Analyzing price insights based on user type to forecast the impact of introducing various discounts. This examination would help project the potential effects of different discount strategies.

8 Reference

https://www.statmethods.net/input/dates.html

https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html

https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/

https://datatofish.com/export-dataframe-to-csv-in-r/

https://www.optimonk.com/limited-time-offers/