Cyclistic Bikes Analysis

Ramya Srinivasan

2024-01-18

Introduction

Welcome to the Cyclistic bike-share analysis case study! As a junior data analyst at Cyclistic, my mission is to understand how casual riders and annual members utilize Cyclistic bikes differently. This analysis is crucial for designing a new marketing strategy aimed at converting casual riders into annual members. Let's embark on this journey to uncover the distinctive patterns and behaviors of Cyclistic's user base by following the 5 stage process of Data Analysis i.e., Ask, Prepare, Process, Share and Act.

Dataset for the analysis is taken from here https://divvy-tripdata.s3.amazonaws.com/index.html.

Ask

The guiding question for this analysis is: How do annual members and casual riders use Cyclistic bikes differently?

Prepare

For this analysis, Cyclistic's historical trip data from the past 12 months has been obtained, provided by Motivate International Inc. While the dataset has a different name due to Cyclistic being a fictional company, it is suitable for answering the business questions at hand. This public data allows exploration of how different customer types utilize Cyclistic bikes. It's crucial to note that data-privacy regulations prohibit the use of personally identifiable information, preventing the connection of pass purchases to credit card numbers for determining the residence of casual riders or their pass-purchasing patterns. This limitation ensures compliance with privacy standards while still allowing for a comprehensive analysis of bike usage trends among Cyclistic's diverse user base.

Process

The tool chosen to conduct data analysis is primarily RStudio Desktop and its associated packages, including tidyverse, lubridate, hms, data.table, and DescTools. R is a powerful and versatile programming language for data analysis. Furthermore, it can handle processing large datasets with ease and this dataset containts approx. 5 million rows in total.

Steps Taken to Ensure Data Cleanliness:

- **Merging Data Frames**:

Monthly data files in csv format were merged into a single dataframe (data_2023) to consolidate information and streamline analysis.

```
data_2023 <- rbind(data_2023_01, data_2023_02, data_2023_03, data_2023_04, data_2023_05, data_2023_06, data_2023_
07, data_2023_08, data_2023_09, data_2023_10, data_2023_11, data_2023_12)
```

- **Removing Duplicate Rows**:

Duplicate rows were identified and removed from the merged dataframe (data_2023_2) using the distinct function.

```
data_2023_2 <- distinct(data_2023_2)
```

- **Handling Missing Values**:

Rows with null values were removed from the dataframe using na.omit to ensure that the analysis is based on complete and accurate data.

```
data_2023_2 <- na.omit(data_2023_2)
```

- **Column Selection**:

Unneeded columns, such as ride_id, start_station_id, end_station_id, start_lat, start_lng, end_lat, and end_lng, were removed to focus on relevant variables.

```
data_2023_2 <- data_2023_2 %>%
  select(-c(ride_id, start_station_id, end_station_id, start_lat, start_lng, end_lat, end_lng))
```

Analyze

To perform effective analysis, the data has been organized and formatted in a way that facilitates exploration and insights. Here are the key steps taken:

1. Organizing Data:

- **Day of Week and Month Columns**:

Columns for day_of_week and month were created to categorize trips by the day of the week and month, providing insights into usage patterns.

```
data_2023_2$day_of_week <- wday(data_2023_2$started_at)
data_2023_2$month <- format(as.Date(data_2023_2$started_at), "%m")
```

- **Season Classification**:

A column named season was introduced to classify rides into seasons based on their corresponding months.

```
data_2023_2 <- data_2023_2 %>% mutate(season =
                                        case_when(month == "01" ~ "Winter",
                                                  month == "02" ~ "Winter",
                                                  month == "03" ~ "Spring",
                                                  month == "04" ~ "Spring",
                                                  month == "05" ~ "Spring",
                                                  month == "06" ~ "Summer",
                                                  month == "07" ~ "Summer",
                                                  month == "08" ~ "Summer",
                                                  month == "09" ~ "Fall",
                                                  month == "10" ~ "Fall",
                                                  month == "11" ~ "Fall",
                                                  month == "12" ~ "Winter")
)
```

- **Ride Length Calculation**: A new column, ride_length, was created by calculating the time difference between the started_at and ended_at columns in minutes. Rides with a length less than or equal to zero were removed to ensure data accuracy.

```
data_2023_2 <- data_2023_2 %>%
  mutate(ride_length = as.numeric(difftime(ended_at, started_at, units = "mins")))
```

2. Key Calculations:

- **Member Type Count**:

The count of rides for each member type (member_casual) was obtained to understand the distribution of riders.

```
data_2023_2 %>% count(member_casual)
```

- **Rideable Type Count**:

The total number of rides was analyzed based on the type of rideable bike to identify preferences among users.

```
data_2023_2 %>%
  group_by(rideable_type) %>%
  count(rideable_type)
```

- **Ride Length Summary**:

Summary statistics, including minimum, maximum, median, and mean of ride lengths, were calculated to gain insights into the duration of rides.

```
summary(data_2023_2$ride_length)
```

- **Mode of Day of Week**:

The mode of the day_of_week was calculated using the Mode function, providing information on the most common day for rides.

```
mode_day_of_week <- Mode(data_2023_2$day_of_week)
```

- **Average Ride Length**:

Pivot tables were created to calculate the average ride length for members and casual riders, as well as for users on different days of the week.

```
pivot_table_1 <- data_2023_2 %>%
  group_by(member_casual) %>%
  summarize(Average_ride_length = mean(ride_length, na.rm = TRUE))

pivot_table_2 <- data_2023_2 %>%
  group_by(day_of_week) %>%
  summarize(Average_ride_length = mean(ride_length, na.rm = TRUE))
```

Trends and Relationships:

- **Member Type Distribution**:

The count of member types revealed that 36% of the riders were of the 'Casual' type and 64% of them were of the 'Member' type.

- **Rideable Type Preferences**:

The examination of rideable types revealed that Electric Bikes were the preferred choice, with 52% of users opting for this type. Following closely, 48% of users favored Classic Bikes, while Docked Bikes were the least popular, accounting for only 1% of users.

- **Ride Length Insights**:

The analysis of ride lengths revealed that, on average, casual riders had longer rides with an average ride length of 20.7 minutes, while annual members had slightly shorter rides with an average ride length of 12.1 minutes.

- **Day of Week Trends**:

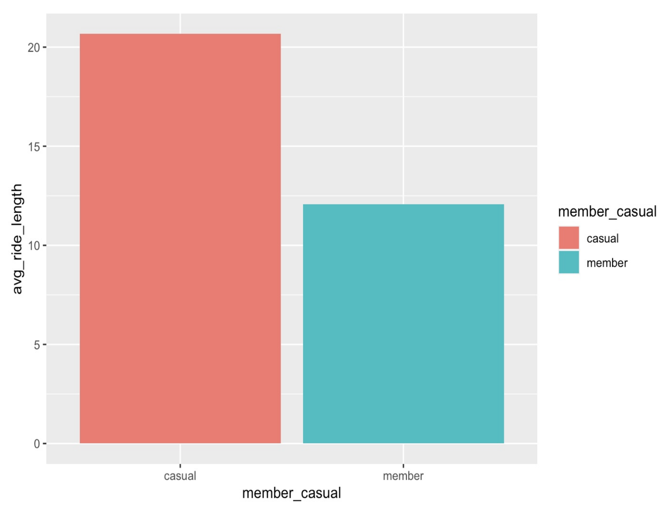
The analysis of ride lengths based on the day of the week showed variations in user behavior. On weekdays (Monday to Friday), the average ride lengths ranged from 13.4 to 18.2 minutes, with the highest average on Monday and the lowest on Thursday.

During the weekends, specifically on Saturday and Sunday, the average ride lengths were relatively higher, with peaks on Sunday (18.1 minutes) and Saturday (15.0 minutes)

Share

Comparison of Average Ride Length

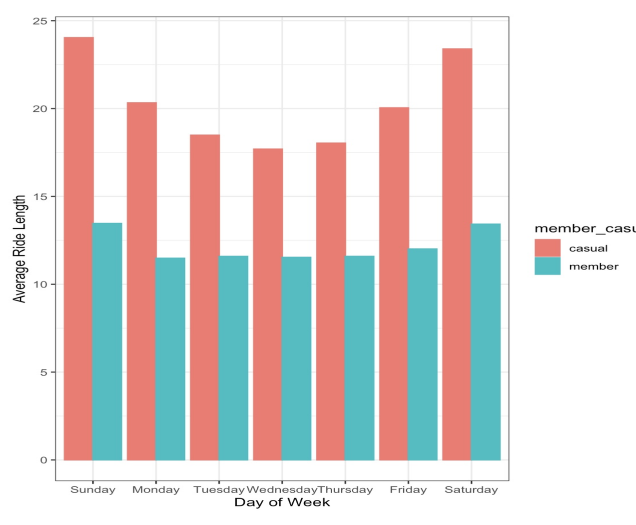
Casual riders have longer average ride lengths (20.7) compared to Member riders (12.1).



Average Ride Length by Day of the Week

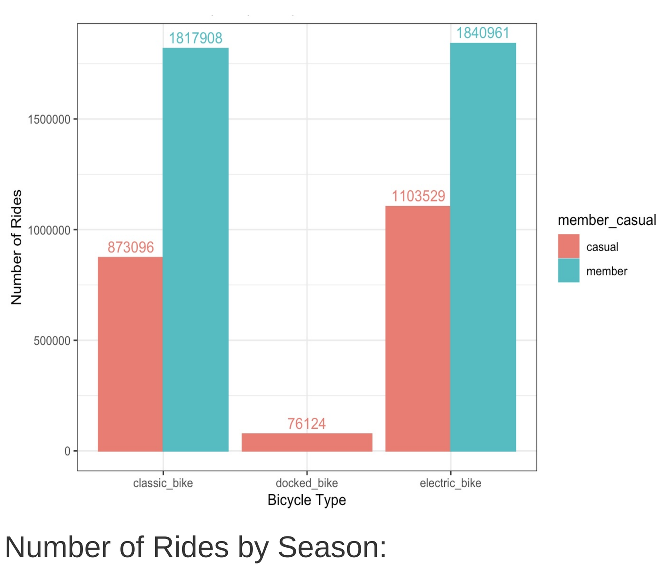
As observed in the data, annual members tend to display a consistent average ride length across the week. The weekdays, from Monday to Friday, show relatively stable patterns with minimal fluctuations in ride lengths.

Interestingly, both annual members and casual riders exhibit an increase in average ride length during the weekends, particularly on Saturdays and Sundays.



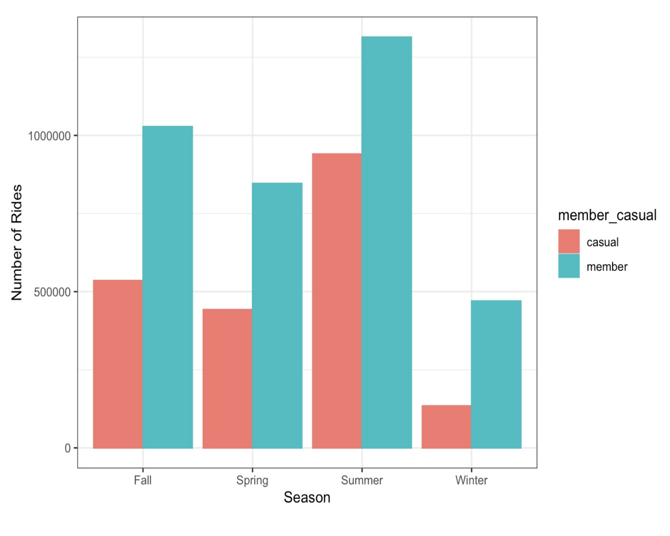
Number of Trips by Bicycle Type:

Electric bikes emerge as the preferred choice for both member types, with classic bikes following closely. Notably, there is a absence of member types utilizing docked bikes, and the usage of docked bikes by casual riders is significantly limited.



Number of Rides by Season:

Seasonal trends in bike usage reveal a clear preference, with the Summer season being the most popular, followed by Fall and Spring. Winter witnesses the lowest bike usage among the seasons. The statistics are consisent among the member types as well.



Act

Refined Recommendations for Cyclistic Marketing Strategy

1. **Targeted Membership Promotions**:

Leverage the insight that 36% of riders are "Casual" and 64% are "Member" to design promotions specifically for casual riders. Offer time-sensitive discounts, exclusive perks, or bundled services to entice casual riders to upgrade to annual memberships. Customize promotions based on individual usage patterns, emphasizing the financial benefits and convenience of annual membership.

2. **Enhanced Weekend Offerings**:

Recognize the trend of higher average ride lengths on weekends. Consider special promotions, events, or partnerships during weekends to capitalize on increased user activity and engagement.

3. **Rideable Type Management**:

Given that Electric Bikes are the most popular, allocate resources to maintain and expand the electric bike fleet. Evaluate the feasibility of introducing new models or incentives to further boost their usage and phase out the provision of Docked Bikes.