Cyclistic\_study

Hasan Mustafa

2025-08-06

Table of Contents

# Cyclistic Bike-Share Case Study

## Introduction

The goal of this project is to analyze Cyclistic’s historical bike-share data and identify differences in usage patterns between **casual riders** and **annual members**. This insight will help guide marketing strategies aimed at converting casual riders into members.

## Setup

## CLEANING THE DATA

### renaming, ride\_length, filtering.

bike\_data <- clean\_names(bike\_data)  
glimpse(bike\_data)

## Rows: 1,389,678  
## Columns: 17  
## $ ride\_id <chr> "2658E319B13141F9", "B2176315168A47CE", "C2A9D33DF7…  
## $ rideable\_type <chr> "electric\_bike", "electric\_bike", "electric\_bike", …  
## $ started\_at <dttm> 2024-07-11 08:15:14, 2024-07-11 15:45:07, 2024-07-…  
## $ ended\_at <dttm> 2024-07-11 08:17:56, 2024-07-11 16:06:04, 2024-07-…  
## $ start\_station\_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, "California Ave & M…  
## $ start\_station\_id <chr> NA, NA, NA, NA, NA, NA, NA, NA, "13084", NA, NA, NA…  
## $ end\_station\_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, "California Ave & M…  
## $ end\_station\_id <chr> NA, NA, NA, NA, NA, NA, NA, NA, "13084", NA, NA, NA…  
## $ start\_lat <dbl> 41.80000, 41.79000, 41.79000, 41.88000, 41.95000, 4…  
## $ start\_lng <dbl> -87.59000, -87.60000, -87.59000, -87.64000, -87.640…  
## $ end\_lat <dbl> 41.79000, 41.80000, 41.79000, 41.90000, 41.91000, 4…  
## $ end\_lng <dbl> -87.59000, -87.59000, -87.60000, -87.67000, -87.620…  
## $ member\_casual <chr> "casual", "casual", "casual", "casual", "casual", "…  
## $ ride\_length <dbl> 2.692517, 20.939867, 3.284083, 28.080000, 12.061667…  
## $ day\_of\_week <ord> Thu, Thu, Thu, Thu, Thu, Thu, Thu, Thu, Thu, Wed, T…  
## $ month <ord> Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, J…  
## $ hour <int> 8, 15, 8, 8, 18, 16, 8, 14, 13, 20, 18, 2, 9, 15, 1…

skim(bike\_data)

Data summary

|  |  |
| --- | --- |
| Name | bike\_data |
| Number of rows | 1389678 |
| Number of columns | 17 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 7 |
| factor | 2 |
| numeric | 6 |
| POSIXct | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ride\_id | 0 | 1.00 | 16 | 16 | 0 | 1389678 | 0 |
| rideable\_type | 0 | 1.00 | 12 | 13 | 0 | 2 | 0 |
| start\_station\_name | 264878 | 0.81 | 10 | 64 | 0 | 1545 | 0 |
| start\_station\_id | 264878 | 0.81 | 3 | 14 | 0 | 2681 | 0 |
| end\_station\_name | 265575 | 0.81 | 10 | 64 | 0 | 1537 | 0 |
| end\_station\_id | 265575 | 0.81 | 3 | 35 | 0 | 2676 | 0 |
| member\_casual | 0 | 1.00 | 6 | 6 | 0 | 2 | 0 |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| day\_of\_week | 0 | 1 | TRUE | 7 | Sat: 229318, Mon: 201166, Tue: 195497, Fri: 195302 |
| month | 0 | 1 | TRUE | 3 | Jul: 731200, Jun: 658321, May: 157, Jan: 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start\_lat | 0 | 1 | 41.91 | 0.04 | 41.64 | 41.88 | 41.90 | 41.93 | 42.07 | ▁▁▇▇▁ |
| start\_lng | 0 | 1 | -87.65 | 0.03 | -87.89 | -87.66 | -87.64 | -87.63 | -87.52 | ▁▁▂▇▁ |
| end\_lat | 33 | 1 | 41.91 | 0.06 | 41.48 | 41.88 | 41.90 | 41.93 | 87.96 | ▇▁▁▁▁ |
| end\_lng | 33 | 1 | -87.65 | 0.06 | -144.05 | -87.66 | -87.64 | -87.63 | -87.42 | ▁▁▁▁▇ |
| ride\_length | 0 | 1 | 17.31 | 32.06 | 1.00 | 6.49 | 11.14 | 19.39 | 1438.70 | ▇▁▁▁▁ |
| hour | 0 | 1 | 14.30 | 4.96 | 0.00 | 11.00 | 15.00 | 18.00 | 23.00 | ▁▃▅▇▅ |

**Variable type: POSIXct**

| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| --- | --- | --- | --- | --- | --- | --- |
| started\_at | 0 | 1 | 2024-06-30 09:48:17 | 2025-06-30 23:57:15 | 2024-07-30 17:15:23 | 1389373 |
| ended\_at | 0 | 1 | 2024-07-01 00:00:15 | 2025-06-30 23:59:49 | 2024-07-30 17:29:56 | 1389201 |

bike\_data <- bike\_data %>%  
 mutate(  
 started\_at = ymd\_hms(started\_at),  
 ended\_at = ymd\_hms(ended\_at),  
 ride\_length = as.numeric(difftime(ended\_at, started\_at, units = "mins")),  
 day\_of\_week = wday(started\_at, label = TRUE),  
 month = month(started\_at, label = TRUE),  
 hour = hour(started\_at)  
 )

## remove rides <1 min or >1 day

bike\_data <- bike\_data %>%  
 filter(ride\_length > 1, ride\_length < 1440)

## Save Cleaned Data

saveRDS(bike\_data, "cleaned\_data/bike\_data.rds")

## Analysis

## Structure of the Dataset

We load the cleaned dataset and view its structure and sample rows to understand the contents and format.

bike\_data <- readRDS("cleaned\_data/bike\_data.rds")  
glimpse(bike\_data)

## Rows: 1,389,678  
## Columns: 17  
## $ ride\_id <chr> "2658E319B13141F9", "B2176315168A47CE", "C2A9D33DF7…  
## $ rideable\_type <chr> "electric\_bike", "electric\_bike", "electric\_bike", …  
## $ started\_at <dttm> 2024-07-11 08:15:14, 2024-07-11 15:45:07, 2024-07-…  
## $ ended\_at <dttm> 2024-07-11 08:17:56, 2024-07-11 16:06:04, 2024-07-…  
## $ start\_station\_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, "California Ave & M…  
## $ start\_station\_id <chr> NA, NA, NA, NA, NA, NA, NA, NA, "13084", NA, NA, NA…  
## $ end\_station\_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, "California Ave & M…  
## $ end\_station\_id <chr> NA, NA, NA, NA, NA, NA, NA, NA, "13084", NA, NA, NA…  
## $ start\_lat <dbl> 41.80000, 41.79000, 41.79000, 41.88000, 41.95000, 4…  
## $ start\_lng <dbl> -87.59000, -87.60000, -87.59000, -87.64000, -87.640…  
## $ end\_lat <dbl> 41.79000, 41.80000, 41.79000, 41.90000, 41.91000, 4…  
## $ end\_lng <dbl> -87.59000, -87.59000, -87.60000, -87.67000, -87.620…  
## $ member\_casual <chr> "casual", "casual", "casual", "casual", "casual", "…  
## $ ride\_length <dbl> 2.692517, 20.939867, 3.284083, 28.080000, 12.061667…  
## $ day\_of\_week <ord> Thu, Thu, Thu, Thu, Thu, Thu, Thu, Thu, Thu, Wed, T…  
## $ month <ord> Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, J…  
## $ hour <int> 8, 15, 8, 8, 18, 16, 8, 14, 13, 20, 18, 2, 9, 15, 1…

head(bike\_data)

## # A tibble: 6 × 17  
## ride\_id rideable\_type started\_at ended\_at   
## <chr> <chr> <dttm> <dttm>   
## 1 2658E319B13141F9 electric\_bike 2024-07-11 08:15:14 2024-07-11 08:17:56  
## 2 B2176315168A47CE electric\_bike 2024-07-11 15:45:07 2024-07-11 16:06:04  
## 3 C2A9D33DF7EBB422 electric\_bike 2024-07-11 08:24:48 2024-07-11 08:28:05  
## 4 8BFEA406DF01D8AD electric\_bike 2024-07-11 08:46:06 2024-07-11 09:14:11  
## 5 ECD3EF02E5EB73B6 electric\_bike 2024-07-11 18:18:16 2024-07-11 18:30:20  
## 6 A3C62391BBBAC107 electric\_bike 2024-07-11 16:03:59 2024-07-11 16:32:38  
## # ℹ 13 more variables: start\_station\_name <chr>, start\_station\_id <chr>,  
## # end\_station\_name <chr>, end\_station\_id <chr>, start\_lat <dbl>,  
## # start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>, member\_casual <chr>,  
## # ride\_length <dbl>, day\_of\_week <ord>, month <ord>, hour <int>

## Count of Rides by User Type

This tells us how many rides were made by casual users versus members.

bike\_data %>%  
 count(member\_casual)

## # A tibble: 2 × 2  
## member\_casual n  
## <chr> <int>  
## 1 casual 588261  
## 2 member 801417

## Ride Duration Summary

We analyze how long rides typically last for each type of user — this helps us understand usage patterns and preferences.

bike\_data %>%  
 group\_by(member\_casual) %>%  
 summarise(  
 average\_duration = mean(ride\_length),  
 median\_duration = median(ride\_length),  
 max\_duration = max(ride\_length)  
 )

## # A tibble: 2 × 4  
## member\_casual average\_duration median\_duration max\_duration  
## <chr> <dbl> <dbl> <dbl>  
## 1 casual 23.0 13.7 1439.  
## 2 member 13.1 9.67 1423.

## Rides by Day of Week

This shows us how riding behavior varies across the week for each user type.

bike\_data %>%  
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(num\_rides = n(), .groups = "drop") %>%  
 arrange(member\_casual, day\_of\_week)

## # A tibble: 14 × 3  
## member\_casual day\_of\_week num\_rides  
## <chr> <ord> <int>  
## 1 casual Sun 93074  
## 2 casual Mon 71988  
## 3 casual Tue 66484  
## 4 casual Wed 69500  
## 5 casual Thu 75751  
## 6 casual Fri 86700  
## 7 casual Sat 124764  
## 8 member Sun 91417  
## 9 member Mon 129178  
## 10 member Tue 129013  
## 11 member Wed 124156  
## 12 member Thu 114497  
## 13 member Fri 108602  
## 14 member Sat 104554

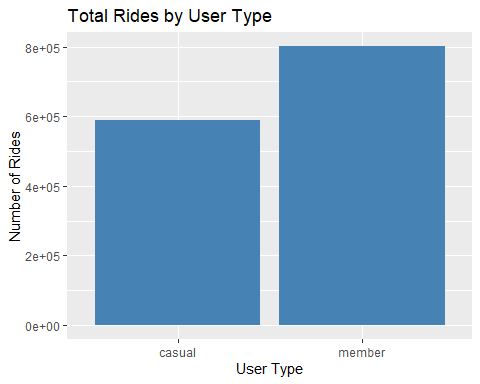
## Visualize

In this section, we create visualizations to compare ride behavior between **casual users** and **members**.

## 1. Total Rides by User Type

This bar chart shows the total number of rides taken by members vs. casual users. It gives a quick overview of who uses the service more.

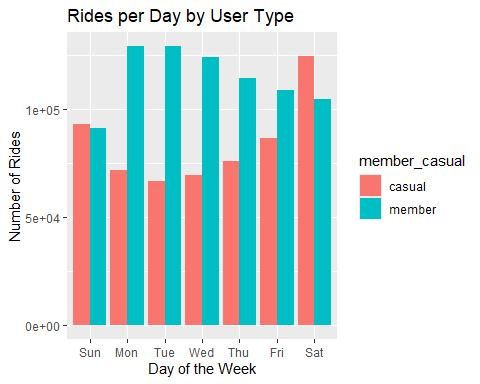
library(ggplot2)  
  
ggplot(bike\_data, aes(x = member\_casual)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Total Rides by User Type", x = "User Type", y = "Number of Rides")



## 2.Rides per Day of the Week by User Type

This chart breaks down the number of rides per weekday for both user types. It helps us understand on which days each group prefers to ride.

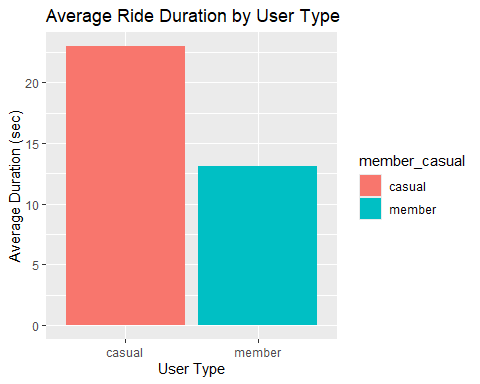
bike\_data %>%  
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(num\_rides = n(), .groups = "drop") %>%  
 ggplot(aes(x = day\_of\_week, y = num\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 labs(title = "Rides per Day by User Type", x = "Day of the Week", y = "Number of Rides")



## 3.Average Ride Duration by User Type

This bar chart compares the average ride duration between casual and member users.

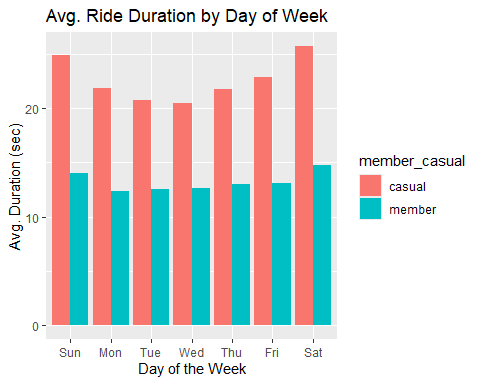
bike\_data %>%  
 group\_by(member\_casual) %>%  
 summarise(avg\_duration = mean(ride\_length)) %>%  
 ggplot(aes(x = member\_casual, y = avg\_duration, fill = member\_casual)) +  
 geom\_col() +  
 labs(title = "Average Ride Duration by User Type", x = "User Type", y = "Average Duration (sec)")



## 4.Average Ride Duration by Day of Week

This chart shows how the average duration of rides varies across the week for each user type. It may reveal patterns like longer weekend rides for casual users.

bike\_data %>%  
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(avg\_duration = mean(ride\_length), .groups = "drop") %>%  
 ggplot(aes(x = day\_of\_week, y = avg\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 labs(title = "Avg. Ride Duration by Day of Week", x = "Day of the Week", y = "Avg. Duration (sec)")



## Summary of Findings

Based on the analysis, we observed the following patterns:

* **Casual users** tend to take longer rides than **annual members**, especially on weekends.
* **Members** ride more frequently and consistently throughout the week.
* **Casual users** ride more often on weekends, while **members** have higher usage during weekdays.
* There are clear differences in behavior based on **day of the week** and **ride duration**.

## Business Recommendations

To convert more casual riders into annual members, Cyclistic could consider:

1. **Targeted Marketing Campaigns** on weekends when casual use is high.
2. **Promotions/discounts** to encourage frequent riders to become members.
3. **Feature improvements** (like app suggestions, ride history benefits) that cater to casual users.
4. Partnering with event organizers on weekends to increase exposure.

## Next Steps

* Analyze seasonal trends or weather impact on rides.
* Incorporate location-based data (start/end station) for deeper route analysis.
* Conduct user surveys to understand decision factors.