Cyclistic_study

Hasan Mustafa

2025-08-06

Contents

1 Cyclistic Bike-Share Case Study

1.1 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.

1.2 Setup

1.3 CLEANING THE DATA

1.3.1 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)
```

Table 1: 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

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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)
  )
```

1.4 remove rides <1 min or >1 day

```r
bike_data <- bike_data %>%
  filter(ride_length > 1, ride_length < 1440)
```

1.5 Save Cleaned Data

```r
saveRDS(bike_data, "cleaned_data/bike_data.rds")
```

1.6 Analysis

1.7 Structure of the Dataset

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

```r
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~
```

```r
head(bike_data)
```

```
## # A tibble: 6 x 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>
```

1.8 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 x 2
##   member_casual      n
##   <chr>          <int>
## 1 casual        588261
## 2 member        801417
```

1.9 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 x 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.
```

1.10 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 x 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
```

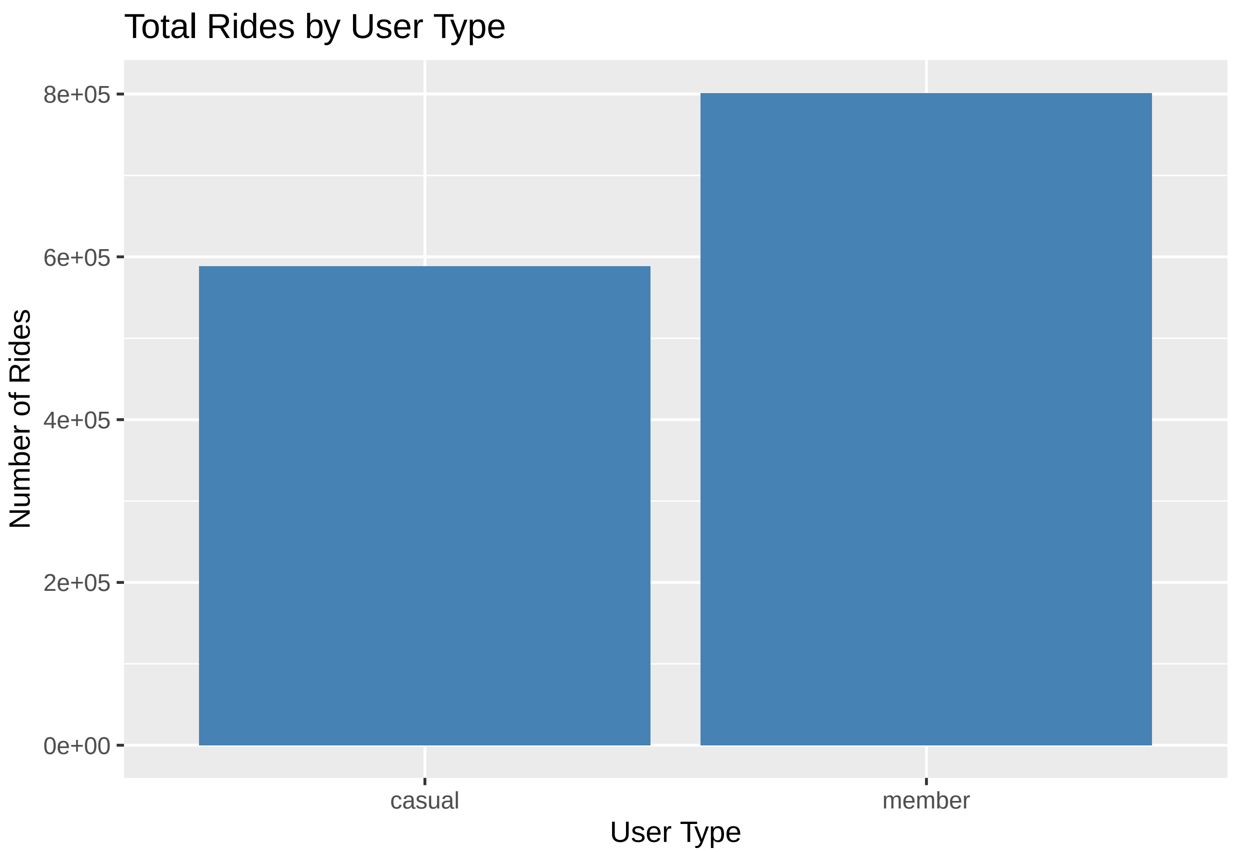
1.11 Visualize

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

1.12 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.

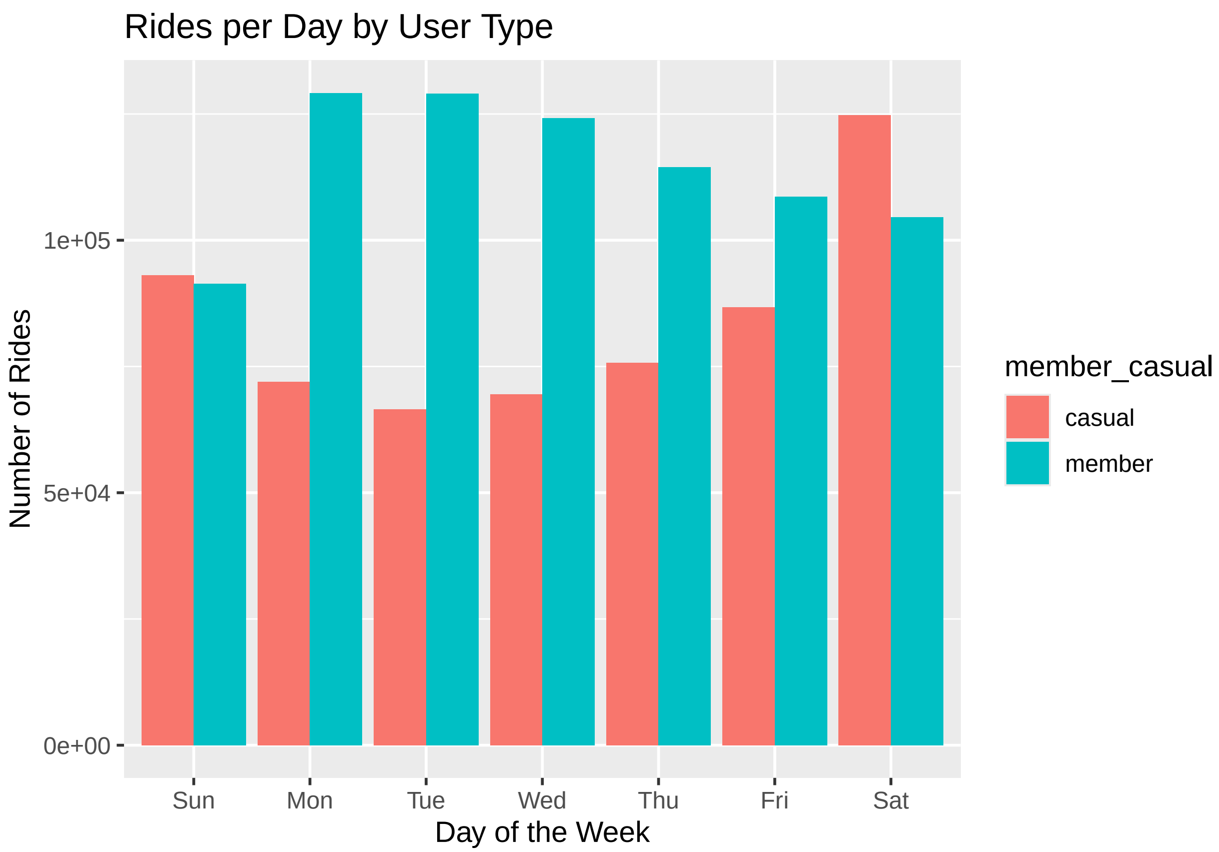
```r
library(ggplot2)

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")
```



1.13 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.

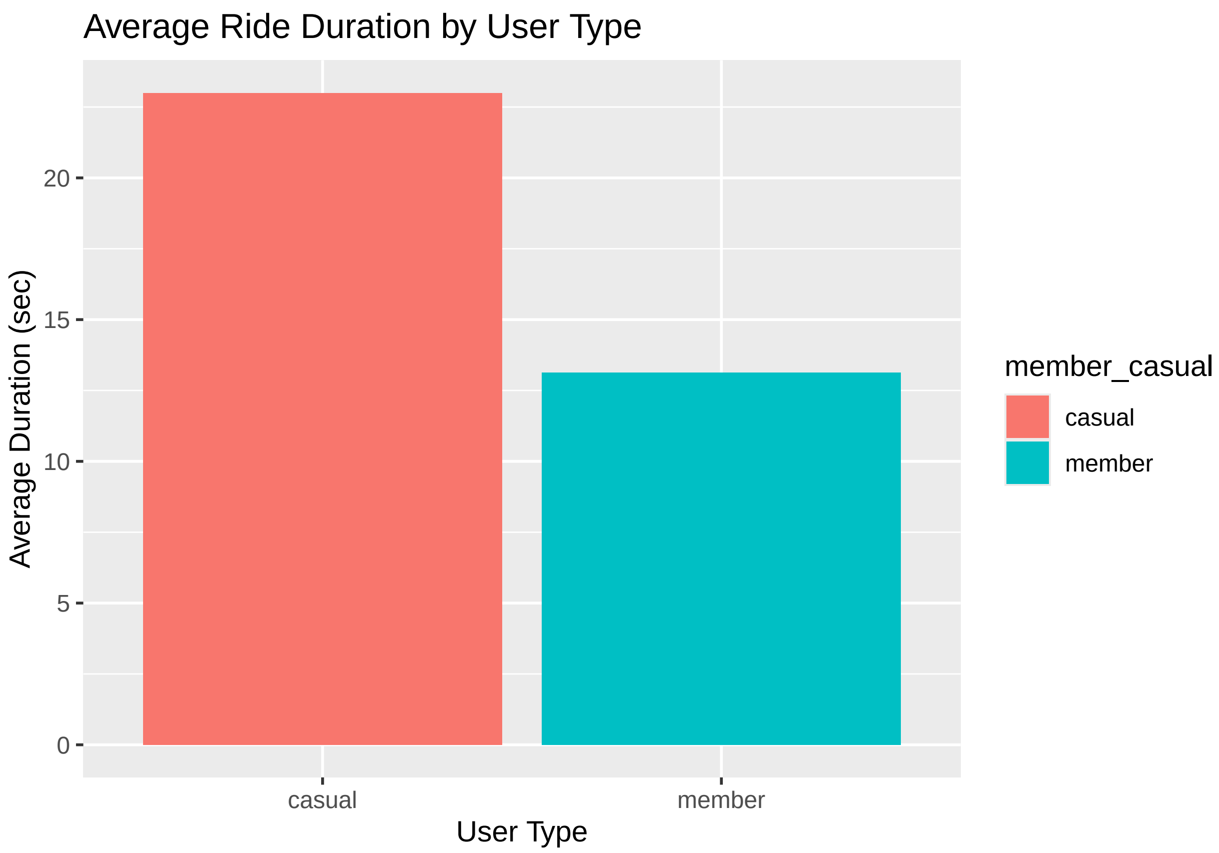
```r
bike_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(num_rides = n(), .groups = "drop") %>%
  ggplot(aes(x = day_of_week, y = num_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Rides per Day by User Type", x = "Day of the Week", y = "Number of Rides")
```



1.14 3.Average Ride Duration by User Type

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

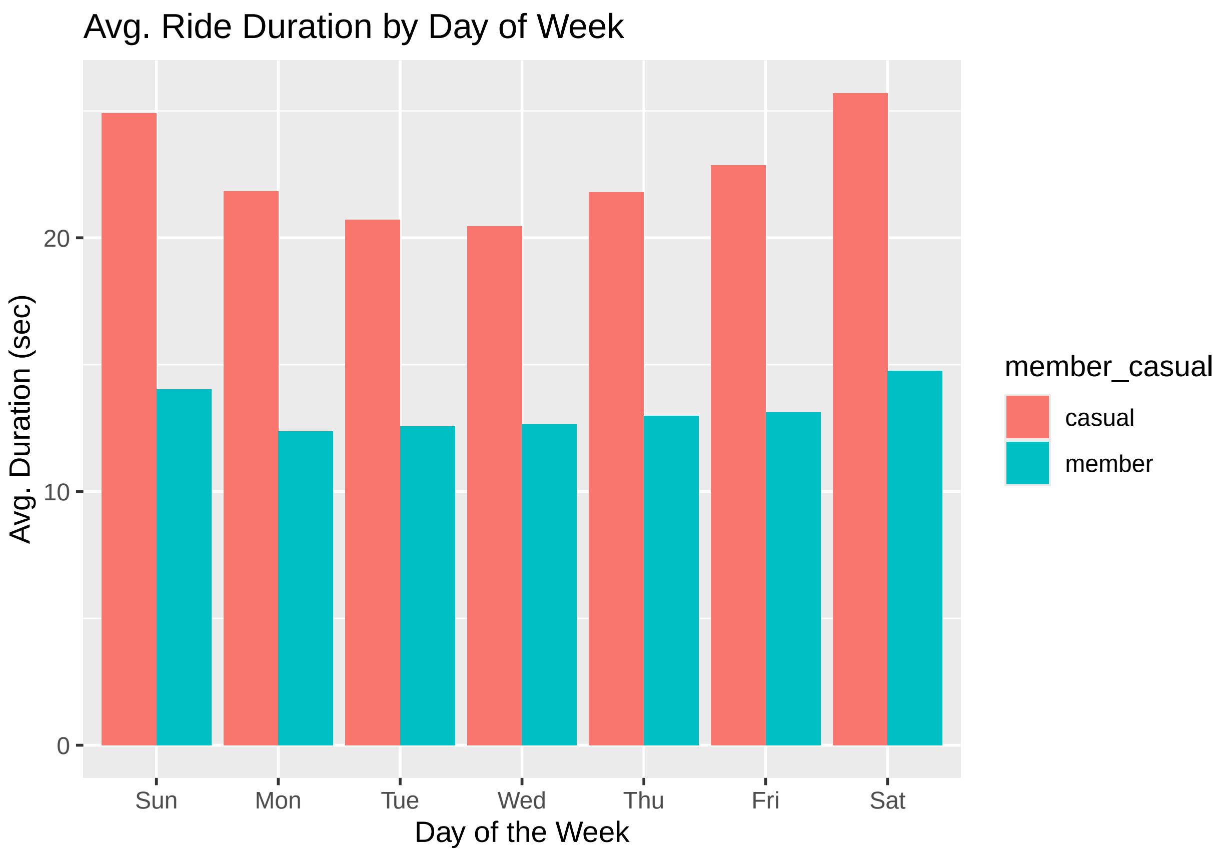
```
bike_data %>%
  group_by(member_casual) %>%
  summarise(avg_duration = mean(ride_length)) %>%
  ggplot(aes(x = member_casual, y = avg_duration, fill = member_casual)) +
  geom_col() +
  labs(title = "Average Ride Duration by User Type", x = "User Type", y = "Average Duration (sec)")
```



1.15 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)")
```



1.16 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**.

1.17 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.

1.18 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.
