Cyclistic Bike Share Data Analysis

Nayan Kadhre

2023-12-24

Introduction

Hello everyone, this analysis is part of the Google Data Analysis Professional Certificate Course 8 Case Study. As part of this case study, I undertake the responsibilities of a real-world junior data analyst within the marketing analyst team at Cyclistic, a fictional company.

Business Problem and Objecive

To design a new Marketing Strategy to convert Casual Riders to Cyclistic Member based on insights, trends and patterns found about how Casual Riders differ from Cyclistic Members from analyzing historical Cyclistic trip data.

Environment

- RStudio Version: 2023.12.0+369
- R version: 4.3.2
- tidyverse version 2.0.0
- lubridate version 1.9.3

About Data

Cyclistic's historical trip data is utilized for this analysis. You can download the data from here (https://divvy-tripdata.s3.amazonaws.com/index.html). The data used in this analysis pertains to the year 2023, excluding the data for the month of December 2023, as it is not available right now. The data for December 2023 will possibly be updated and made available next month, in January 2024. The data is stored in multiple zip files, with each zip file representing the data for a specific month. The data is provided in CSV file format. The original, unclean data set comprises over 5.49 million rows, necessitating cleaning and modification before conducting the analysis.

Code

Import the data

```
q1_2023_01 <- read_csv("2023-cyclist-bike-share-data/202301-divvy-tripdata.csv")
```

```
## Rows: 190301 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q1_2023_02 <- read_csv("2023-cyclist-bike-share-data/202302-divvy-tripdata.csv")
```

```
## Rows: 190445 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q1_2023_03 <- read_csv("2023-cyclist-bike-share-data/202303-divvy-tripdata.csv")
```

```
## Rows: 258678 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q2_2023_04 <- read_csv("2023-cyclist-bike-share-data/202304-divvy-tripdata.csv")
```

```
## Rows: 426590 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q2_2023_05 <- read_csv("2023-cyclist-bike-share-data/202305-divvy-tripdata.csv")
```

```
## Rows: 604827 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q2_2023_06 <- read_csv("2023-cyclist-bike-share-data/202306-divvy-tripdata.csv")
```

```
## Rows: 719618 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q3_2023_07 <- read_csv("2023-cyclist-bike-share-data/202307-divvy-tripdata.csv")
```

```
## Rows: 767650 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q3_2023_08 <- read_csv("2023-cyclist-bike-share-data/202308-divvy-tripdata.csv")
```

```
## Rows: 771693 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q3_2023_09 <- read_csv("2023-cyclist-bike-share-data/202309-divvy-tripdata.csv")
```

```
## Rows: 666371 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q4_2023_10 <- read_csv("2023-cyclist-bike-share-data/202310-divvy-tripdata.csv")
```

```
## Rows: 537113 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q4_2023_11 <- read_csv("2023-cyclist-bike-share-data/202311-divvy-tripdata.csv")
```

```
## Rows: 362518 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Check column names for each of the files for any inconsistencies that need to be solved

```
colnames(q1_2023_01)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q1_2023_02)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q1_2023_03)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q2_2023_04)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q2_2023_05)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q2_2023_06)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q3_2023_07)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q3_2023_08)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q3_2023_09)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q4_2023_10)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(q4_2023_11)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

View the structure of the data

```
str(q1_2023_01)
```

```
## spc_tbl_ [190,301 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED9
68" ...
##  $ rideable_type     : chr [1:190301] "electric_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" ...
##  $ ended_at          : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" ...
##  $ start_station_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western Ave & Lun
t Ave" "Kimbark Ave & 53rd St" ...
##  $ start_station_id  : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
##  $ end_station_name  : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli Produce - E
vanston Plaza" "Greenwood Ave & 47th St" ...
##  $ end_station_id    : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...
##  $ start_lat         : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
##  $ start_lng         : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:190301] 41.9 41.8 42 41.8 41.8 ...
##  $ end_lng           : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:190301] "member" "member" "casual" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q1_2023_02)
```

```
## spc_tbl_ [190,445 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:190445] "CBCD0D7777F0E45F" "F3EC5FCE5FF39DE9" "E54C1F27FA9354FF" "3D561E04F739CC
45" ...
##  $ rideable_type     : chr [1:190445] "classic_bike" "electric_bike" "classic_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:190445], format: "2023-02-14 11:59:42" "2023-02-15 13:53:48" ...
##  $ ended_at          : POSIXct[1:190445], format: "2023-02-14 12:13:38" "2023-02-15 13:59:08" ...
##  $ start_station_name: chr [1:190445] "Southport Ave & Clybourn Ave" "Clarendon Ave & Gordon Ter" "Southport A
ve & Clybourn Ave" "Southport Ave & Clybourn Ave" ...
##  $ start_station_id  : chr [1:190445] "TA1309000030" "13379" "TA1309000030" "TA1309000030" ...
##  $ end_station_name  : chr [1:190445] "Clark St & Schiller St" "Sheridan Rd & Lawrence Ave" "Aberdeen St & Mon
roe St" "Franklin St & Adams St (Temp)" ...
##  $ end_station_id    : chr [1:190445] "TA1309000024" "TA1309000041" "13156" "TA1309000008" ...
##  $ start_lat         : num [1:190445] 41.9 42 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:190445] -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:190445] 41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:190445] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:190445] "casual" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q1_2023_03)
```

```
## spc_tbl_ [258,678 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:258678] "6842AA605EE9FBB3" "F984267A75B99A8C" "FF7CF57CFE026D02" "6B61B916032CB6
D6" ...
##  $ rideable_type     : chr [1:258678] "electric_bike" "electric_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:258678], format: "2023-03-16 08:20:34" "2023-03-04 14:07:06" ...
##  $ ended_at          : POSIXct[1:258678], format: "2023-03-16 08:22:52" "2023-03-04 14:15:31" ...
##  $ start_station_name: chr [1:258678] "Clark St & Armitage Ave" "Public Rack - Kedzie Ave & Argyle St" "Orlean
s St & Chestnut St (NEXT Apts)" "Desplaines St & Kinzie St" ...
##  $ start_station_id  : chr [1:258678] "13146" "491" "620" "TA1306000003" ...
##  $ end_station_name  : chr [1:258678] "Larrabee St & Webster Ave" NA "Clark St & Randolph St" "Sheffield Ave &
Kingsbury St" ...
##  $ end_station_id    : chr [1:258678] "13193" NA "TA1305000030" "13154" ...
##  $ start_lat         : num [1:258678] 41.9 42 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:258678] -87.6 -87.7 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:258678] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:258678] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:258678] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2023_04)
```

```
## spc_tbl_ [426,590 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:426590] "8FE8F7D9C10E88C7" "34E4ED3ADF1D821B" "5296BF07A2F77CB5" "40759916B76D5D
52" ...
##  $ rideable_type     : chr [1:426590] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:426590], format: "2023-04-02 08:37:28" "2023-04-19 11:29:02" ...
##  $ ended_at          : POSIXct[1:426590], format: "2023-04-02 08:41:37" "2023-04-19 11:52:12" ...
##  $ start_station_name: chr [1:426590] NA NA NA NA ...
##  $ start_station_id  : chr [1:426590] NA NA NA NA ...
##  $ end_station_name  : chr [1:426590] NA NA NA NA ...
##  $ end_station_id    : chr [1:426590] NA NA NA NA ...
##  $ start_lat         : num [1:426590] 41.8 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:426590] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:426590] 41.8 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:426590] -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:426590] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2023_05)
```

```
## spc_tbl_ [604,827 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:604827] "0D9FA920C3062031" "92485E5FB5888ACD" "FB144B3FC8300187" "DDEB93BC2CE9AA
77" ...
##  $ rideable_type     : chr [1:604827] "electric_bike" "electric_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:604827], format: "2023-05-07 19:53:48" "2023-05-06 18:54:08" ...
##  $ ended_at          : POSIXct[1:604827], format: "2023-05-07 19:58:32" "2023-05-06 19:03:35" ...
##  $ start_station_name: chr [1:604827] "Southport Ave & Belmont Ave" "Southport Ave & Belmont Ave" "Halsted St
& 21st St" "Carpenter St & Huron St" ...
##  $ start_station_id  : chr [1:604827] "13229" "13229" "13162" "13196" ...
##  $ end_station_name  : chr [1:604827] NA NA NA "Damen Ave & Cortland St" ...
##  $ end_station_id    : chr [1:604827] NA NA NA "13133" ...
##  $ start_lat         : num [1:604827] 41.9 41.9 41.9 41.9 42 ...
##  $ start_lng         : num [1:604827] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:604827] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:604827] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:604827] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q2_2023_06)
```

```
## spc_tbl_ [719,618 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:719618] "6F1682AC40EB6F71" "622A1686D64948EB" "3C88859D926253B4" "EAD8A5E0259DEC
88" ...
##  $ rideable_type     : chr [1:719618] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:719618], format: "2023-06-05 13:34:12" "2023-06-05 01:30:22" ...
##  $ ended_at          : POSIXct[1:719618], format: "2023-06-05 14:31:56" "2023-06-05 01:33:06" ...
##  $ start_station_name: chr [1:719618] NA NA NA NA ...
##  $ start_station_id  : chr [1:719618] NA NA NA NA ...
##  $ end_station_name  : chr [1:719618] NA NA NA NA ...
##  $ end_station_id    : chr [1:719618] NA NA NA NA ...
##  $ start_lat         : num [1:719618] 41.9 41.9 42 42 42 ...
##  $ start_lng         : num [1:719618] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:719618] 41.9 41.9 41.9 42 42 ...
##  $ end_lng           : num [1:719618] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:719618] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2023_07)
```

```
## spc_tbl_ [767,650 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:767650] "9340B064F0AEE130" "D1460EE3CE0D8AF8" "DF41BE31B895A25E" "9624A293749EF7
03" ...
##  $ rideable_type     : chr [1:767650] "electric_bike" "classic_bike" "classic_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:767650], format: "2023-07-23 20:06:14" "2023-07-23 17:05:07" ...
##  $ ended_at          : POSIXct[1:767650], format: "2023-07-23 20:22:44" "2023-07-23 17:18:37" ...
##  $ start_station_name: chr [1:767650] "Kedzie Ave & 110th St" "Western Ave & Walton St" "Western Ave & Walton
St" "Racine Ave & Randolph St" ...
##  $ start_station_id  : chr [1:767650] "20204" "KA1504000103" "KA1504000103" "13155" ...
##  $ end_station_name  : chr [1:767650] "Public Rack - Racine Ave & 109th Pl" "Milwaukee Ave & Grand Ave" "Damen
Ave & Pierce Ave" "Clinton St & Madison St" ...
##  $ end_station_id    : chr [1:767650] "877" "13033" "TA1305000041" "TA1305000032" ...
##  $ start_lat         : num [1:767650] 41.7 41.9 41.9 41.9 42 ...
##  $ start_lng         : num [1:767650] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:767650] 41.7 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:767650] -87.7 -87.6 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:767650] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2023_08)
```

```
## spc_tbl_ [771,693 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:771693] "903C30C2D810A53B" "F2FB18A98E110A2B" "D0DEC7C94E4663DA" "E0DDDC5F84747E
D9" ...
##  $ rideable_type     : chr [1:771693] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:771693], format: "2023-08-19 15:41:53" "2023-08-18 15:30:18" ...
##  $ ended_at          : POSIXct[1:771693], format: "2023-08-19 15:53:36" "2023-08-18 15:45:25" ...
##  $ start_station_name: chr [1:771693] "LaSalle St & Illinois St" "Clark St & Randolph St" "Clark St & Randolph
St" "Wells St & Elm St" ...
##  $ start_station_id  : chr [1:771693] "13430" "TA1305000030" "TA1305000030" "KA1504000135" ...
##  $ end_station_name  : chr [1:771693] "Clark St & Elm St" NA NA NA ...
##  $ end_station_id    : chr [1:771693] "TA1307000039" NA NA NA ...
##  $ start_lat         : num [1:771693] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:771693] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:771693] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:771693] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:771693] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q3_2023_09)
```

```
## spc_tbl_ [666,371 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:666371] "011C1903BF4E2E28" "87DB80E048A1BF9F" "7C2EB7AF669066E3" "57D197B010269C
E3" ...
##  $ rideable_type     : chr [1:666371] "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:666371], format: "2023-09-23 00:27:50" "2023-09-02 09:26:43" ...
##  $ ended_at          : POSIXct[1:666371], format: "2023-09-23 00:33:27" "2023-09-02 09:38:19" ...
##  $ start_station_name: chr [1:666371] "Halsted St & Wrightwood Ave" "Clark St & Drummond Pl" "Financial Pl & I
da B Wells Dr" "Clark St & Drummond Pl" ...
##  $ start_station_id  : chr [1:666371] "TA1309000061" "TA1307000142" "SL-010" "TA1307000142" ...
##  $ end_station_name  : chr [1:666371] "Sheffield Ave & Wellington Ave" "Racine Ave & Fullerton Ave" "Racine Av
e & 15th St" "Racine Ave & Belmont Ave" ...
##  $ end_station_id    : chr [1:666371] "TA1307000052" "TA1306000026" "13304" "TA1308000019" ...
##  $ start_lat         : num [1:666371] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:666371] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:666371] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:666371] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:666371] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q4_2023_10)
```

```
## spc_tbl_ [537,113 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:537113] "4449097279F8BBE7" "9CF060543CA7B439" "667F21F4D6BDE69C" "F92714CC6B019B
96" ...
##  $ rideable_type     : chr [1:537113] "classic_bike" "electric_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:537113], format: "2023-10-08 10:36:26" "2023-10-11 17:23:59" ...
##  $ ended_at          : POSIXct[1:537113], format: "2023-10-08 10:49:19" "2023-10-11 17:36:08" ...
##  $ start_station_name: chr [1:537113] "Orleans St & Chestnut St (NEXT Apts)" "Desplaines St & Kinzie St" "Orle
ans St & Chestnut St (NEXT Apts)" "Desplaines St & Kinzie St" ...
##  $ start_station_id  : chr [1:537113] "620" "TA1306000003" "620" "TA1306000003" ...
##  $ end_station_name  : chr [1:537113] "Sheffield Ave & Webster Ave" "Sheffield Ave & Webster Ave" "Franklin St
& Lake St" "Franklin St & Lake St" ...
##  $ end_station_id    : chr [1:537113] "TA1309000033" "TA1309000033" "TA1307000111" "TA1307000111" ...
##  $ start_lat         : num [1:537113] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:537113] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:537113] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:537113] -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:537113] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(q4_2023_11)
```

```
## spc_tbl_ [362,518 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:362518] "4EAD8F1AD547356B" "6322270563BF5470" "B37BDE091ECA38E0" "CF0CA5DD26E4F9
0E" ...
##  $ rideable_type     : chr [1:362518] "electric_bike" "electric_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:362518], format: "2023-11-30 21:50:05" "2023-11-03 09:44:02" ...
##  $ ended_at          : POSIXct[1:362518], format: "2023-11-30 22:13:27" "2023-11-03 10:17:15" ...
##  $ start_station_name: chr [1:362518] "Millennium Park" "Broadway & Sheridan Rd" "State St & Pearson St" "Thea
ter on the Lake" ...
##  $ start_station_id  : chr [1:362518] "13008" "13323" "TA1307000061" "TA1308000001" ...
##  $ end_station_name  : chr [1:362518] "Pine Grove Ave & Waveland Ave" "Broadway & Sheridan Rd" "State St & Pea
rson St" "Theater on the Lake" ...
##  $ end_station_id    : chr [1:362518] "TA1307000150" "13323" "TA1307000061" "TA1308000001" ...
##  $ start_lat         : num [1:362518] 41.9 42 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:362518] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:362518] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:362518] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:362518] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Combine the data into single data frame all_trips

```
all_trips <- bind_rows(q1_2023_01, q1_2023_02, q1_2023_03, q2_2023_04, q2_2023_05, q2_2023_06, q3_2023_07, q3_202
3_08, q3_2023_09, q4_2023_10, q4_2023_11)
```

Remove columns not necessary for analysis

```
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng, start_station_name, start_station_id, end_station_name, end_s
tation_id))
```

Columns present in the data

```
colnames(all_trips)
```

```
## [1] "ride_id"       "rideable_type" "started_at"    "ended_at"
## [5] "member_casual"
```

Number of Rows present in the data

```
nrow(all_trips)
```

```
## [1] 5495804
```

Dimensions of the Data Frame (rows x columns)

```
dim(all_trips)
```

```
## [1] 5495804       5
```

Display first 6 observations

```
head(all_trips)
```

```
## # A tibble: 6 × 5
##   ride_id    rideable_type started_at          ended_at            member_casual
##   <chr>      <chr>         <dttm>              <dttm>              <chr>
## 1 F96D5A74A… electric_bike 2023-01-21 20:05:42 2023-01-21 20:16:33 member
## 2 13CB7EB69… classic_bike  2023-01-10 15:37:36 2023-01-10 15:46:05 member
## 3 BD88A2E67… electric_bike 2023-01-02 07:51:57 2023-01-02 08:05:11 casual
## 4 C90792D03… classic_bike  2023-01-22 10:52:58 2023-01-22 11:01:44 member
## 5 339701752… classic_bike  2023-01-12 13:58:01 2023-01-12 14:13:20 member
## 6 58E68156D… electric_bike 2023-01-31 07:18:03 2023-01-31 07:21:16 member
```

Display last 6 observations

```
tail(all_trips)
```

```
## # A tibble: 6 × 5
##   ride_id    rideable_type started_at          ended_at            member_casual
##   <chr>      <chr>         <dttm>              <dttm>              <chr>
## 1 B80D69303… classic_bike  2023-11-03 17:04:45 2023-11-03 17:07:50 member
## 2 30B44BD4C… classic_bike  2023-11-24 08:39:27 2023-11-24 08:47:03 member
## 3 094A79892… classic_bike  2023-11-06 09:07:20 2023-11-06 09:10:00 member
## 4 F0A7DF8A4… electric_bike 2023-11-10 19:35:30 2023-11-10 19:44:28 member
## 5 4D5E3685B… classic_bike  2023-11-27 09:11:23 2023-11-27 09:13:23 member
## 6 1FA95C375… electric_bike 2023-11-20 16:16:03 2023-11-20 16:17:43 member
```

View the structure of the data

```
str(all_trips)
```

```
## tibble [5,495,804 × 5] (S3: tbl_df/tbl/data.frame)
##  $ ride_id      : chr [1:5495804] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED968"
...
##  $ rideable_type: chr [1:5495804] "electric_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at   : POSIXct[1:5495804], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" ...
##  $ ended_at     : POSIXct[1:5495804], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" ...
##  $ member_casual: chr [1:5495804] "member" "member" "casual" "member" ...
```

View basic statistical summary of the data

```
summary(all_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5495804     Length:5495804     Min.   :2023-01-01 00:01:58.00
##  Class :character   Class :character   1st Qu.:2023-05-18 18:15:14.50
##  Mode  :character   Mode  :character   Median :2023-07-16 13:09:35.50
##                                        Mean   :2023-07-10 06:47:12.70
##                                        3rd Qu.:2023-09-09 12:49:02.75
##                                        Max.   :2023-11-30 23:59:14.00
##     ended_at                      member_casual
##  Min.   :2023-01-01 00:02:41.00   Length:5495804
##  1st Qu.:2023-05-18 18:32:05.75   Class :character
##  Median :2023-07-16 13:30:46.00   Mode  :character
##  Mean   :2023-07-10 07:05:34.51
##  3rd Qu.:2023-09-09 13:10:32.00
##  Max.   :2023-12-01 20:42:31.00
```

Find the unique categories present in the ‘members_casual’ column

```
unique(all_trips[,"member_casual"])
```

```
## # A tibble: 2 × 1
##   member_casual
##   <chr>
## 1 member
## 2 casual
```

Find the count of each of the unique categories present in the ‘members_casual’ column

```
table(all_trips$member_casual)
```

```
##
##  casual  member
## 2007507 3488297
```

View the structure of the data (similar to str() function)

```
glimpse(all_trips)
```

```
## Rows: 5,495,804
## Columns: 5
## $ ride_id       <chr> "F96D5A74A3E41399", "13CB7EB698CEDB88", "BD88A2E670661CE…
## $ rideable_type <chr> "electric_bike", "classic_bike", "electric_bike", "class…
## $ started_at    <dttm> 2023-01-21 20:05:42, 2023-01-10 15:37:36, 2023-01-02 07…
## $ ended_at      <dttm> 2023-01-21 20:16:33, 2023-01-10 15:46:05, 2023-01-02 08…
## $ member_casual <chr> "member", "member", "casual", "member", "member", "membe…
```

####Add ‘date’ column representing the start date of the ride

```
all_trips$date <- as.Date(all_trips$started_at)
```

Add ‘month’ column representing the month the ride started

```
all_trips$month <- format(as.Date(all_trips$date), "%m")
```

Add ‘day’ column representing the day the ride started

```
all_trips$day <- format(as.Date(all_trips$date), "%d")
```

Add ‘year’ column representing the year the ride started

```
all_trips$year <- format(as.Date(all_trips$date), "%Y")
```

Add ‘day_of_week’ column representing the day of the week the ride started

```
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
```

Add ‘ride_length’ column representing time take for the ride (in seconds)

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
```

Convert month to month names

```
all_trips$month_names <- format(all_trips$started_at, "%B")
```

Create a new column with abbreviated month names

```
all_trips$month_abbrev <- format(all_trips$started_at, "%b")
```

Convert “ride_length” from Factor to numeric so we can run calculations on the data

```
is.factor(all_trips$ride_length)
```

```
## [1] FALSE
```

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

```
## [1] TRUE
```

Once again view the structure of the data

```
glimpse(all_trips)
```

```
## Rows: 5,495,804
## Columns: 13
## $ ride_id       <chr> "F96D5A74A3E41399", "13CB7EB698CEDB88", "BD88A2E670661CE…
## $ rideable_type <chr> "electric_bike", "classic_bike", "electric_bike", "class…
## $ started_at    <dttm> 2023-01-21 20:05:42, 2023-01-10 15:37:36, 2023-01-02 07…
## $ ended_at      <dttm> 2023-01-21 20:16:33, 2023-01-10 15:46:05, 2023-01-02 08…
## $ member_casual <chr> "member", "member", "casual", "member", "member", "membe…
## $ date          <date> 2023-01-21, 2023-01-10, 2023-01-02, 2023-01-22, 2023-01…
## $ month         <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01", "0…
## $ day           <chr> "21", "10", "02", "22", "12", "31", "15", "25", "25", "0…
## $ year          <chr> "2023", "2023", "2023", "2023", "2023", "2023", "2023", …
## $ day_of_week   <chr> "Saturday", "Tuesday", "Monday", "Sunday", "Thursday", "…
## $ ride_length   <dbl> 651, 509, 794, 526, 919, 193, 840, 561, 747, 753, 589, 5…
## $ month_names   <chr> "January", "January", "January", "January", "January", "…
## $ month_abbrev  <chr> "Jan", "Jan", "Jan", "Jan", "Jan", "Jan", "Jan", "Jan", …
```

Check number of null values in each column

```
colSums(is.na(all_trips))
```

```
##       ride_id rideable_type    started_at      ended_at member_casual
##             0             0             0             0             0
##          date         month           day          year   day_of_week
##             0             0             0             0             0
##   ride_length   month_names  month_abbrev
##             0             0             0
```

Remove duplicates and inspect the modification

```
all_trips_v2 <- distinct(all_trips)
nrow(all_trips)
```

```
## [1] 5495804
```

```
nrow(all_trips_v2)
```

```
## [1] 5495804
```

```
colSums(is.na(all_trips))
```

```
##       ride_id rideable_type    started_at      ended_at member_casual
##             0             0             0             0             0
##          date         month           day          year   day_of_week
##             0             0             0             0             0
##   ride_length   month_names  month_abbrev
##             0             0             0
```

```
colSums(is.na(all_trips_v2))
```

```
##       ride_id rideable_type    started_at      ended_at member_casual
##             0             0             0             0             0
##          date         month           day          year   day_of_week
##             0             0             0             0             0
##   ride_length   month_names  month_abbrev
##             0             0             0
```

Remove rows with ride_length equal to 0 or negative

```
all_trips_v2 <- all_trips_v2[!(all_trips_v2$ride_length <= 0),]
# Number of rows removed = 1214
```

Sub-set data based on casual riders and member riders

```
casual_riders <- all_trips_v2[all_trips_v2$member_casual == "casual",]
nrow(casual_riders)
```

```
## [1] 2006959
```

```
colSums(is.na(casual_riders))
```

```
##       ride_id rideable_type    started_at      ended_at member_casual
##             0             0             0             0             0
##          date         month           day          year   day_of_week
##             0             0             0             0             0
##   ride_length   month_names  month_abbrev
##             0             0             0
```

```
member_riders <- all_trips_v2[all_trips_v2$member_casual == "member",]
nrow(member_riders)
```

```
## [1] 3487631
```

```
colSums(is.na(member_riders))
```

```
##       ride_id rideable_type    started_at      ended_at member_casual
##             0             0             0             0             0
##          date         month           day          year   day_of_week
##             0             0             0             0             0
##   ride_length   month_names  month_abbrev
##             0             0             0
```

Descriptive analysis on ride_length

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     328     577    1103    1026 5909344
```

Compare members and casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                1708.0781
## 2                     member                 754.8659
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                      718
## 2                     member                      515
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                  5909344
## 2                     member                    93580
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                        1
## 2                     member                        1
```

See the average ride time by each day for members vs casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                  Friday                1654.0087
## 2                      member                  Friday                 752.2389
## 3                      casual                  Monday                1674.8897
## 4                      member                  Monday                 715.2766
## 5                      casual                Saturday                1942.9229
## 6                      member                Saturday                 843.2250
## 7                      casual                  Sunday                1979.0545
## 8                      member                  Sunday                 843.7574
## 9                      casual                Thursday                1497.4745
## 10                     member                Thursday                 722.7094
## 11                     casual                 Tuesday                1516.8259
## 12                     member                 Tuesday                 724.2108
## 13                     casual               Wednesday                1469.8835
## 14                     member               Wednesday                 720.1731
```

Fix the days of the week that are out of order.

```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday"
, "Thursday", "Friday", "Saturday"))
```

View the average ride time by each day for members vs casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                  Sunday                1979.0545
## 2                      member                  Sunday                 843.7574
## 3                      casual                  Monday                1674.8897
## 4                      member                  Monday                 715.2766
## 5                      casual                 Tuesday                1516.8259
## 6                      member                 Tuesday                 724.2108
## 7                      casual               Wednesday                1469.8835
## 8                      member               Wednesday                 720.1731
## 9                      casual                Thursday                1497.4745
## 10                     member                Thursday                 722.7094
## 11                     casual                  Friday                1654.0087
## 12                     member                  Friday                 752.2389
## 13                     casual                Saturday                1942.9229
## 14                     member                Saturday                 843.2250
```

Analyze ridership data by type and weekday

```
all_trips_v2 %>%
  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)
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

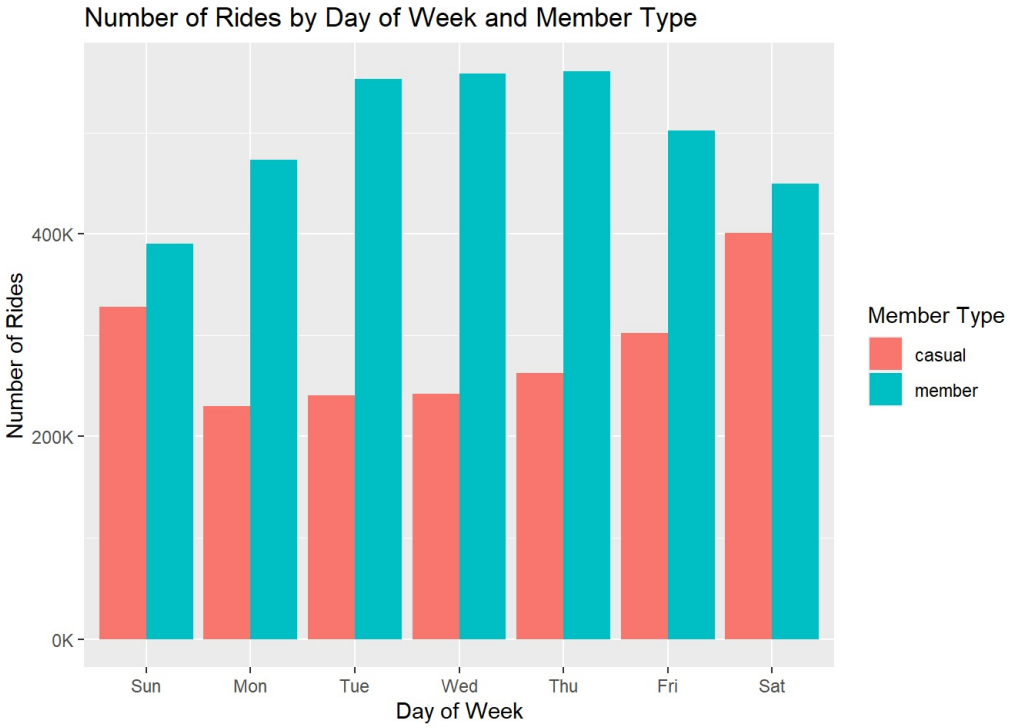
```
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              328144            1979.
##  2 casual        Mon              229721            1675.
##  3 casual        Tue              240843            1517.
##  4 casual        Wed              242114            1470.
##  5 casual        Thu              262903            1497.
##  6 casual        Fri              302324            1654.
##  7 casual        Sat              400910            1943.
##  8 member        Sun              390401             844.
##  9 member        Mon              472888             715.
## 10 member        Tue              553175             724.
## 11 member        Wed              558591             720.
## 12 member        Thu              560952             723.
## 13 member        Fri              502184             752.
## 14 member        Sat              449440             843.
```

Data Visualizations

a. visualize the number of rides by rider type and day of week

```
all_trips_v2 %>%
  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 = "Number of Rides by Day of Week and Member Type",
       x = "Day of Week",
       y = "Number of Rides",
       fill = "Member Type") + scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



b. visualization for average ride duration

```
all_trips_v2 %>%
  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 = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Average ride length by Day of Week and Member Type",
       x = "Day of Week",
       y = "Average Ride Length",
       fill = "Member Type")
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

Preferred Bike type in general and based on casual and member riders

```
table(all_trips_v2$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##       2591250         78287       2825053
```

```
table(casual_riders$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##        856342         78287       1072330
```

```
table(member_riders$rideable_type)
```

```
##
##  classic_bike electric_bike
##       1734908       1752723
```

Distribution of data in general and based on casual and member riders by days of week

```
table(all_trips_v2$day_of_week)
```

```
##
##    Sunday    Monday   Tuesday Wednesday  Thursday    Friday  Saturday
##    718545    702609    794018    800705    823855    804508    850350
```

```
table(casual_riders$day_of_week)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##    302324    229721    400910    328144    262903    240843    242114
```

```
table(member_riders$day_of_week)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##    502184    472888    449440    390401    560952    553175    558591
```

Create subsets for each quarter using the “month” column

```
q1 <- all_trips_v2[all_trips_v2$month %in% c("01", "02", "03"), ]
q2 <- all_trips_v2[all_trips_v2$month %in% c("04", "05", "06"), ]
q3 <- all_trips_v2[all_trips_v2$month %in% c("07", "08", "09"), ]
q4 <- all_trips_v2[all_trips_v2$month %in% c("10", "11", "12"), ]
```

Display the number of rides in each quarter

```
cat("Number of rides in Q1:", nrow(q1), "\n")
```

```
## Number of rides in Q1: 639388
```

```
cat("Number of rides in Q2:", nrow(q2), "\n")
```

```
## Number of rides in Q2: 1750855
```

```
cat("Number of rides in Q3:", nrow(q3), "\n")
```

```
## Number of rides in Q3: 2205024
```

```
cat("Number of rides in Q4:", nrow(q4), "\n")
```

```
## Number of rides in Q4: 899323
```

Create a data frame for visualizing quarterly data

```
rides_data <- data.frame(Quarter = c("Q1", "Q2", "Q3", "Q4"), Rides = c(nrow(q1), nrow(q2), nrow(q3), nrow(q4)))
```

Calculate quarter for each ride

```
all_trips_v2$Quarter <- cut(all_trips_v2$started_at, breaks = "quarters", labels = c("Q1", "Q2", "Q3", "Q4"))
```

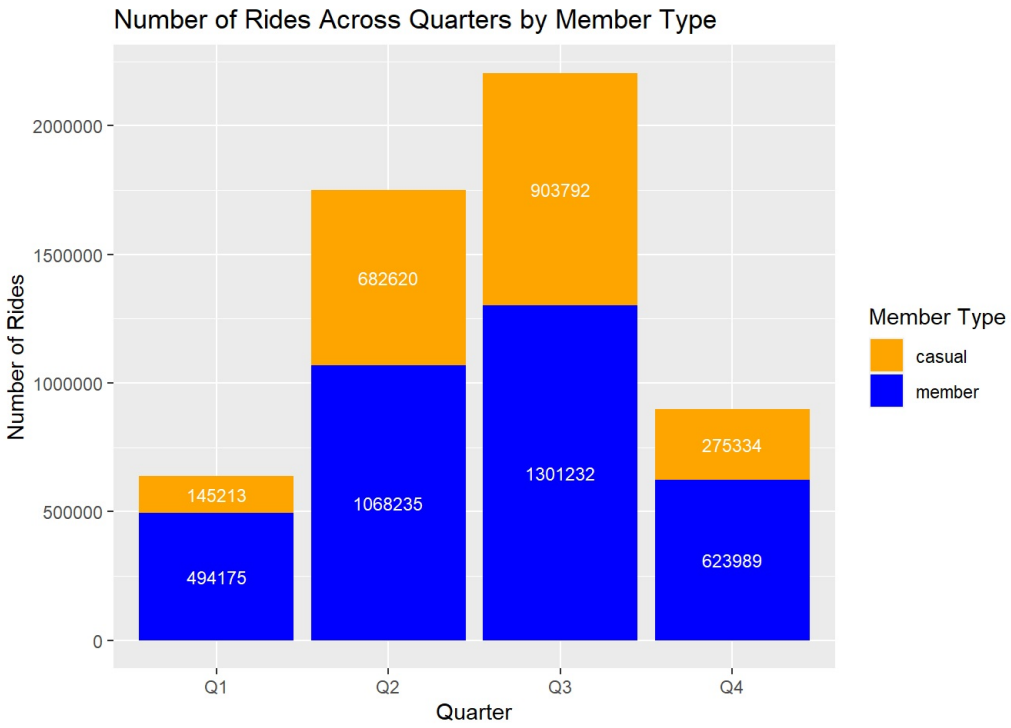
Group by member_casual and Quarter

```
rides_data_quarters <- all_trips_v2 %>%
  group_by(member_casual, Quarter) %>%
  summarise(number_of_rides = n())
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

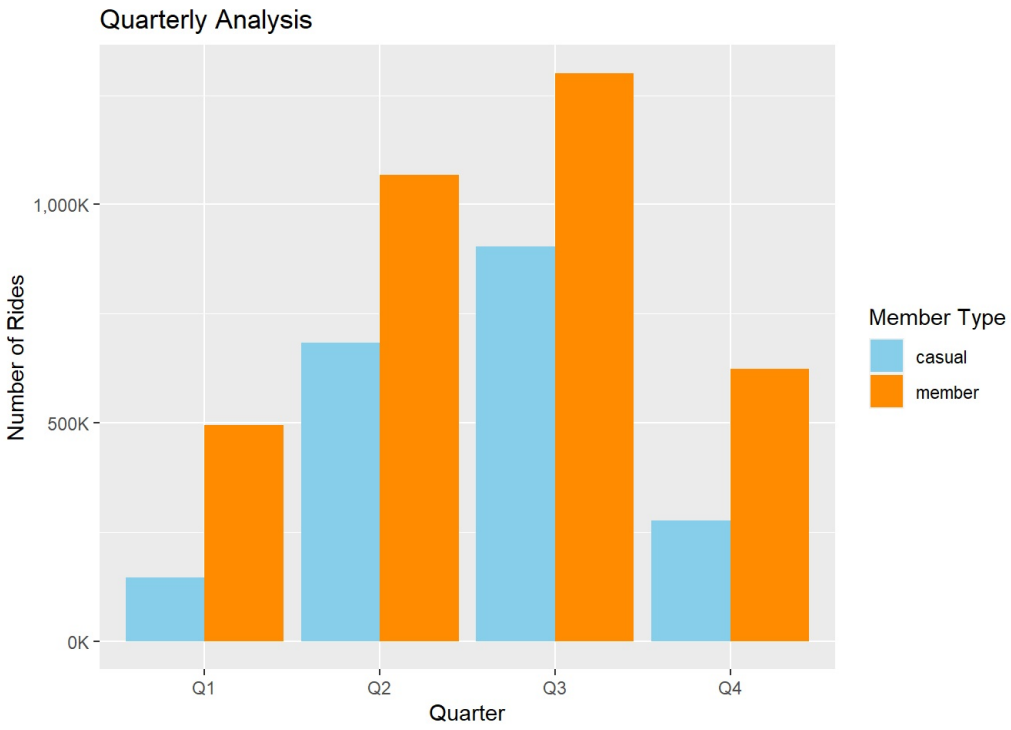
c1. Quarterly Analysis (stacked)

```
quarterly_trend_plot_stacked <- ggplot(rides_data_quarters, aes(x = Quarter, y = number_of_rides, fill = member_c
asual)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = number_of_rides), position = position_stack(vjust = 0.5), size = 3, color="white") +
  labs(title = "Number of Rides Across Quarters by Member Type",x = "Quarter",y = "Number of Rides",fill = "Membe
r Type") +
  scale_fill_manual(values = c("casual" = "orange", "member" = "blue"))

print(quarterly_trend_plot_stacked)
```



c2. Quarterly Analysis

```
quarterly_trend_plot <- ggplot(all_trips_v2, aes(x = Quarter, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Quarterly Analysis",
       x = "Quarter",
       y = "Number of Rides",
       fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "skyblue", "member" = "darkorange")) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))

print(quarterly_trend_plot)
```



Group by member_casual and Quarter

```
rides_data_month <- all_trips_v2 %>%
  group_by(member_casual, month_abbrev) %>%
  summarise(number_of_rides = n())
```

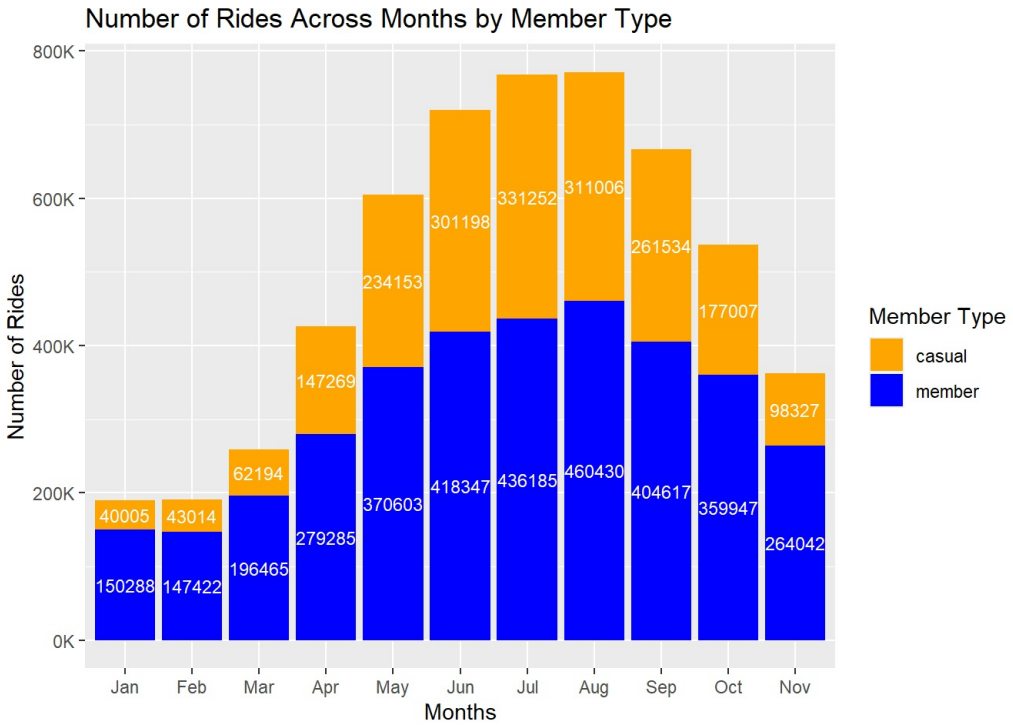
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

Set the order of months

```
rides_data_month$month_abbrev <- factor(rides_data_month$month_abbrev, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
```

d1. Monthly Ride Distribution (stacked)

```
monthly_ride_distribution <- ggplot(rides_data_month, aes(x = month_abbrev , y = number_of_rides, fill = member_casual)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = number_of_rides), position = position_stack(vjust = 0.5), size = 3, color = "white") +
  labs(title = "Number of Rides Across Months by Member Type", x = "Months", y = "Number of Rides", fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "orange", "member" = "blue")) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))

print(monthly_ride_distribution)
```

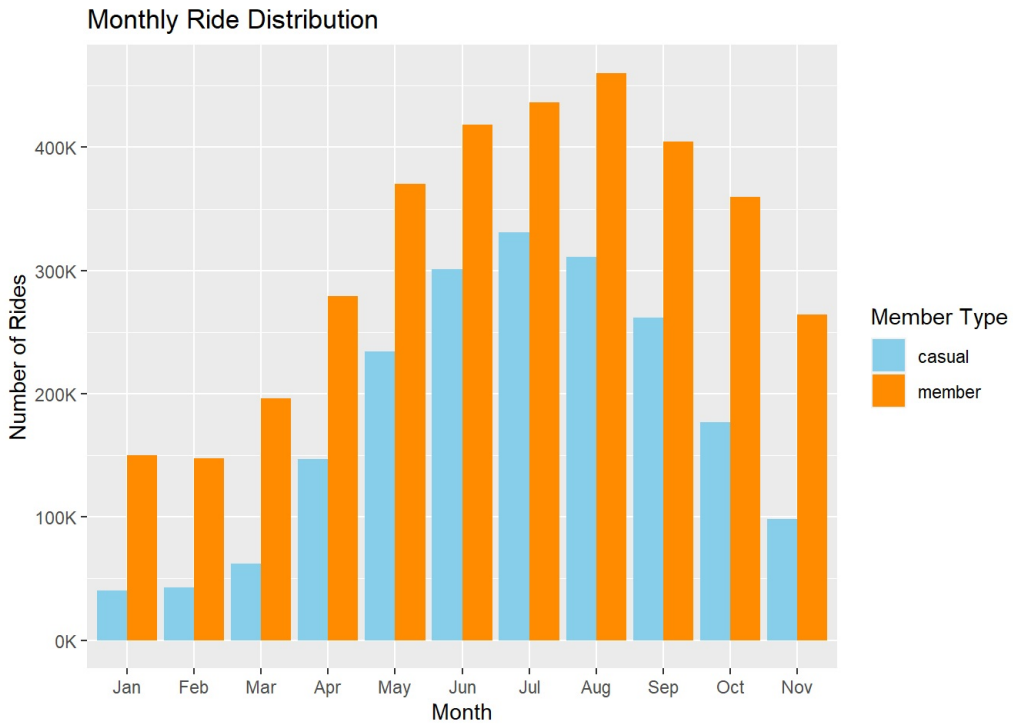


Set the order of months

```
all_trips_v2$month_abbrev <- factor(all_trips_v2$month_abbrev, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
```

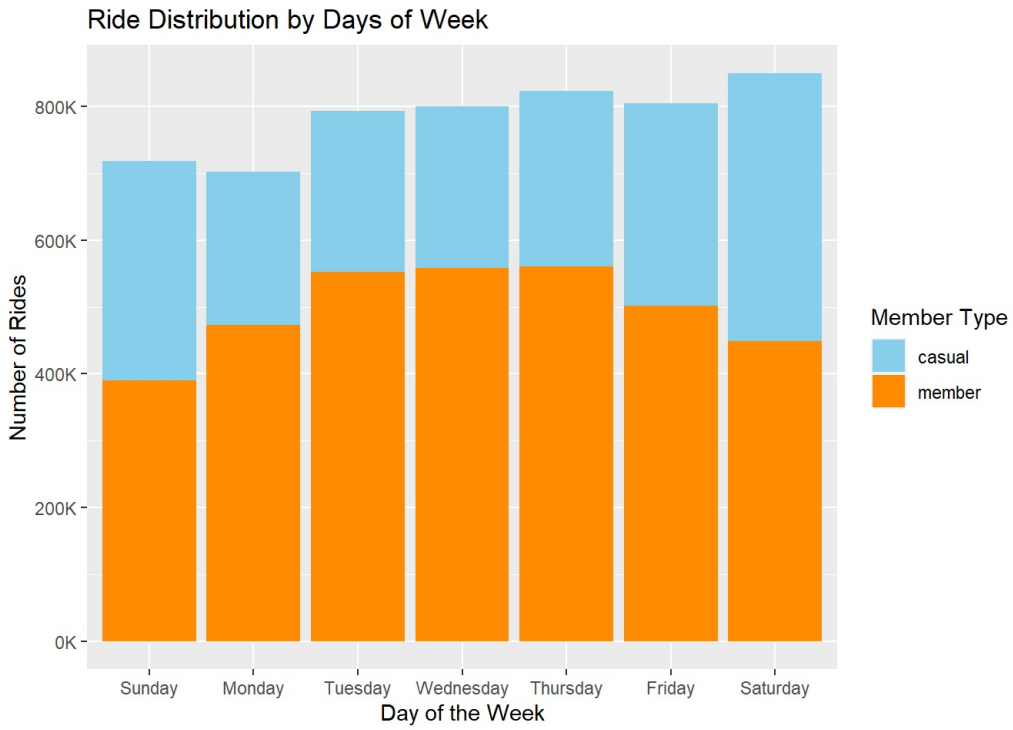
d2. Monthly Ride Distribution

```
monthly_dist_plot <- ggplot(all_trips_v2, aes(x = month_abbrev, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Monthly Ride Distribution",
       x = "Month",
       y = "Number of Rides",
       fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "skyblue", "member" = "darkorange")) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))

print(monthly_dist_plot)
```



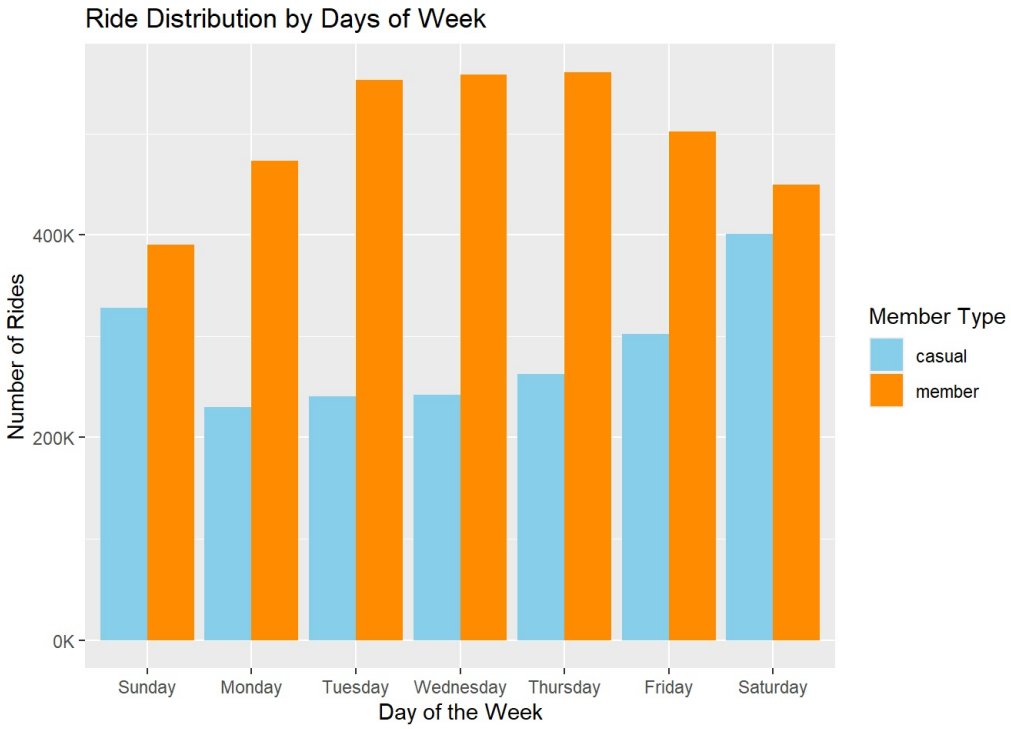
e1. Ride Distribution by Days of Week (stacked)

```
day_of_week_plot_stacked <- ggplot(all_trips_v2, aes(x = day_of_week, fill = member_casual)) +
  geom_bar(position = "stack", stat = "count") +  # Change position to "stack"
  labs(title = "Ride Distribution by Days of Week",
       x = "Day of the Week",
       y = "Number of Rides",
       fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "skyblue", "member" = "darkorange")) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))

print(day_of_week_plot_stacked)
```



e2. Ride Distribution by Days of Week

```
day_of_week_plot <- ggplot(all_trips_v2, aes(x = day_of_week, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Ride Distribution by Days of Week",
       x = "Day of the Week",
       y = "Number of Rides",
       fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "skyblue", "member" = "darkorange")) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "K"))

print(day_of_week_plot)
```



f. Rideable Type Distribution

```
rideable_type_dist_plot <- ggplot(all_trips_v2, aes(x = rideable_type, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Rideable Type Distribution",
       x = "Rideable Type",
       y = "Number of Rides",
       fill = "Member Type") +
  scale_fill_manual(values = c("casual" = "skyblue", "member" = "darkorange"))

print(rideable_type_dist_plot)
```