Case Study- Cyclistic bike-share

2023-02-12

Introduction

This case study is my Google Data Analytics Certificate course final project. In this project I will be analyzing public dataset provided by course using R programming language.

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Lily Moreno (the director of marketing and my manager) has assigned you the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

Step 01 - ASK

Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ. We will analyze the Cyclistic historical bike trip data to identify trends. Find the differences between the casual riders and annual members.

Step 02 - PREPARE

I will be using Cyclistic's historical trip data. The data has been made available by Motivate International Inc. under this license (https://ride.divvybikes.com/data-license-agreement). Datasets are available here (https://divvy-tripdata.s3.amazonaws.com/index.html). I will be using data for the last 12 months - Jan 2022 to Dec 2022.

Install required packages

```
library(tidyverse) #helps wrangle data
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.1     ✓ purrr   1.0.1
## ✓ tibble  3.1.8     ✓ dplyr   1.1.0
## ✓ tidyr   1.3.0     ✓ stringr 1.5.0
## ✓ readr   2.1.3     ✓ forcats 1.0.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)   #helps visualize data
library(janitor)   #helps examining and cleaning data
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate) #helps wrangle date attributes
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(plyr)      #helps data cleaning and transformation
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following object is masked from 'package:purrr':
##
##     compact
```

Collect Data

Import data -12 csv files, each representing 1 of the 12 months of trip data.

```
y22_jan <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202201-divvy-tripdata.csv")
```

```
## Rows: 103770 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.
```

```
y22_feb <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202202-divvy-tripdata.csv")
```

```
## Rows: 115609 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.
```

```
y22_mar <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202203-divvy-tripdata.csv")
```

```
## Rows: 284042 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.
```

```
y22_apr <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202204-divvy-tripdata.csv")
```

```
## Rows: 371249 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.
```

```
y22_may <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202205-divvy-tripdata.csv")
```

```
## Rows: 634858 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.
```

```
y22_jun <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202206-divvy-tripdata.csv")
```

```
## Rows: 769204 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.
```

```
y22_jul <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202207-divvy-tripdata.csv")
```

```
## Rows: 823488 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.
```

```
y22_aug <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202208-divvy-tripdata.csv")
```

```
## Rows: 785932 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.
```

```
y22_sep <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202209-divvy-tripdata.csv")
```

```
## Rows: 701339 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.
```

```
y22_oct <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202210-divvy-tripdata.csv")
```

```
## Rows: 558685 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.
```

```
y22_nov <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202211-divvy-tripdata.csv")
```

```
## Rows: 337735 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.
```

```
y22_dec <- read_csv("C:\\Users\\sathu\\OneDrive\\Desktop\\Case Study\\Data\\202212-divvy-tripdata.csv")
```

```
## Rows: 181806 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.
```

Inspecting Data

```
str(y22_jan)
```

```
## spc_tbl_ [103,770 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:103770] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C
66D" "CBB80ED419105406" ...
##  $ rideable_type     : chr [1:103770] "electric_bike" "electric_bike" "classic_bike" "clas
sic_bike" ...
##  $ started_at        : POSIXct[1:103770], format: "2022-01-13 11:59:47" "2022-01-10 08:41:
56" ...
##  $ ended_at          : POSIXct[1:103770], format: "2022-01-13 12:02:44" "2022-01-10 08:46:
17" ...
##  $ start_station_name: chr [1:103770] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Av
e" "Sheffield Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...
##  $ start_station_id  : chr [1:103770] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:103770] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Green
view Ave & Fullerton Ave" "Paulina St & Montrose Ave" ...
##  $ end_station_id    : chr [1:103770] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num [1:103770] 42 42 41.9 42 41.9 ...
##  $ start_lng         : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:103770] 42 42 41.9 42 41.9 ...
##  $ end_lng           : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:103770] "casual" "casual" "member" "casual" ...
##  - 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(y22_feb)
```

```
## spc_tbl_ [115,609 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:115609] "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C
4AF" "A1789BDF844412BE" ...
##  $ rideable_type     : chr [1:115609] "classic_bike" "classic_bike" "classic_bike" "classi
c_bike" ...
##  $ started_at        : POSIXct[1:115609], format: "2022-02-19 18:08:41" "2022-02-20 17:41:
30" ...
##  $ ended_at          : POSIXct[1:115609], format: "2022-02-19 18:23:56" "2022-02-20 17:45:
56" ...
##  $ start_station_name: chr [1:115609] "State St & Randolph St" "Halsted St & Wrightwood Av
e" "State St & Randolph St" "Southport Ave & Waveland Ave" ...
##  $ start_station_id  : chr [1:115609] "TA1305000029" "TA1309000061" "TA1305000029" "13235"
...
##  $ end_station_name  : chr [1:115609] "Clark St & Lincoln Ave" "Southport Ave & Wrightwood
Ave" "Canal St & Adams St" "Broadway & Sheridan Rd" ...
##  $ end_station_id    : chr [1:115609] "13179" "TA1307000113" "13011" "13323" ...
##  $ start_lat         : num [1:115609] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:115609] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:115609] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:115609] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:115609] "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(y22_mar)
```

```
## spc_tbl_ [284,042 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:284042] "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66
109" "9F446FD9DEE3F389" ...
##  $ rideable_type     : chr [1:284042] "classic_bike" "electric_bike" "classic_bike" "class
ic_bike" ...
##  $ started_at        : POSIXct[1:284042], format: "2022-03-21 13:45:01" "2022-03-16 09:37:
16" ...
##  $ ended_at          : POSIXct[1:284042], format: "2022-03-21 13:51:18" "2022-03-16 09:43:
34" ...
##  $ start_station_name: chr [1:284042] "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Br
oadway & Berwyn Ave" "Wabash Ave & Wacker Pl" ...
##  $ start_station_id  : chr [1:284042] "TA1307000131" "13042" "13109" "TA1307000131" ...
##  $ end_station_name  : chr [1:284042] "Kingsbury St & Kinzie St" "Orleans St & Chestnut St
(NEXT Apts)" "Broadway & Ridge Ave" "Franklin St & Jackson Blvd" ...
##  $ end_station_id    : chr [1:284042] "KA1503000043" "620" "15578" "TA1305000025" ...
##  $ start_lat         : num [1:284042] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:284042] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:284042] "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(y22_apr)
```

```
## spc_tbl_ [371,249 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:371249] "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F0
7FF" "84D4751AEB31888D" ...
##  $ rideable_type     : chr [1:371249] "electric_bike" "classic_bike" "classic_bike" "class
ic_bike" ...
##  $ started_at        : POSIXct[1:371249], format: "2022-04-06 17:42:48" "2022-04-24 19:23:
07" ...
##  $ ended_at          : POSIXct[1:371249], format: "2022-04-06 17:54:36" "2022-04-24 19:43:
17" ...
##  $ start_station_name: chr [1:371249] "Paulina St & Howard St" "Wentworth Ave & Cermak Rd"
"Halsted St & Polk St" "Wentworth Ave & Cermak Rd" ...
##  $ start_station_id  : chr [1:371249] "515" "13075" "TA1307000121" "13075" ...
##  $ end_station_name  : chr [1:371249] "University Library (NU)" "Green St & Madison St" "G
reen St & Madison St" "Delano Ct & Roosevelt Rd" ...
##  $ end_station_id    : chr [1:371249] "605" "TA1307000120" "TA1307000120" "KA1706005007"
...
##  $ start_lat         : num [1:371249] 42 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:371249] 42.1 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:371249] "member" "member" "member" "casual" ...
##  - 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(y22_may)
```

```
## spc_tbl_ [634,858 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF59852924528F8" ...
##  $ rideable_type     : chr [1:634858] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...
##  $ ended_at          : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...
##  $ start_station_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St" "Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
##  $ end_station_name  : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood St & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:634858] "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(y22_jun)
```

```
## spc_tbl_ [769,204 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:769204] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD77
1D2" "C9C320375DE1D5C6" ...
##  $ rideable_type     : chr [1:769204] "electric_bike" "electric_bike" "electric_bike" "ele
ctric_bike" ...
##  $ started_at        : POSIXct[1:769204], format: "2022-06-30 17:27:53" "2022-06-30 18:39:
52" ...
##  $ ended_at          : POSIXct[1:769204], format: "2022-06-30 17:35:15" "2022-06-30 18:47:
28" ...
##  $ start_station_name: chr [1:769204] NA NA NA NA ...
##  $ start_station_id  : chr [1:769204] NA NA NA NA ...
##  $ end_station_name  : chr [1:769204] NA NA NA NA ...
##  $ end_station_id    : chr [1:769204] NA NA NA NA ...
##  $ start_lat         : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
##  $ start_lng         : num [1:769204] -87.6 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num [1:769204] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:769204] "casual" "casual" "casual" "casual" ...
##  - 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(y22_jul)
```

```
## spc_tbl_ [823,488 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:823488] "954144C2F67B1932" "292E027607D218B6" "57765852588AD
6E0" "B5B6BE44314590E6" ...
##  $ rideable_type     : chr [1:823488] "classic_bike" "classic_bike" "classic_bike" "classi
c_bike" ...
##  $ started_at        : POSIXct[1:823488], format: "2022-07-05 08:12:47" "2022-07-26 12:53:
38" ...
##  $ ended_at          : POSIXct[1:823488], format: "2022-07-05 08:24:32" "2022-07-26 12:55:
31" ...
##  $ start_station_name: chr [1:823488] "Ashland Ave & Blackhawk St" "Buckingham Fountain (T
emp)" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)" ...
##  $ start_station_id  : chr [1:823488] "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr [1:823488] "Kingsbury St & Kinzie St" "Michigan Ave & 8th St"
"Michigan Ave & 8th St" "Woodlawn Ave & 55th St" ...
##  $ end_station_id    : chr [1:823488] "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num [1:823488] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:823488] -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:823488] 41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num [1:823488] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:823488] "member" "casual" "casual" "casual" ...
##  - 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(y22_aug)
```

```
## spc_tbl_ [785,932 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:785932] "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB
7FD" "F597830181C2E13C" ...
##  $ rideable_type     : chr [1:785932] "electric_bike" "electric_bike" "electric_bike" "ele
ctric_bike" ...
##  $ started_at        : POSIXct[1:785932], format: "2022-08-07 21:34:15" "2022-08-08 14:39:
21" ...
##  $ ended_at          : POSIXct[1:785932], format: "2022-08-07 21:41:46" "2022-08-08 14:53:
23" ...
##  $ start_station_name: chr [1:785932] NA NA NA NA ...
##  $ start_station_id  : chr [1:785932] NA NA NA NA ...
##  $ end_station_name  : chr [1:785932] NA NA NA NA ...
##  $ end_station_id    : chr [1:785932] NA NA NA NA ...
##  $ start_lat         : num [1:785932] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:785932] 41.9 41.9 42 42 41.8 ...
##  $ end_lng           : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:785932] "casual" "casual" "casual" "casual" ...
##  - 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(y22_sep)
```

```
## spc_tbl_ [701,339 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:701339] "5156990AC19CA285" "E12D4A16BF51C274" "A02B53CD7DB72
DD7" "C82E05FEE872DF11" ...
##  $ rideable_type     : chr [1:701339] "electric_bike" "electric_bike" "electric_bike" "ele
ctric_bike" ...
##  $ started_at        : POSIXct[1:701339], format: "2022-09-01 08:36:22" "2022-09-01 17:11:
29" ...
##  $ ended_at          : POSIXct[1:701339], format: "2022-09-01 08:39:05" "2022-09-01 17:14:
45" ...
##  $ start_station_name: chr [1:701339] NA NA NA NA ...
##  $ start_station_id  : chr [1:701339] NA NA NA NA ...
##  $ end_station_name  : chr [1:701339] "California Ave & Milwaukee Ave" NA NA NA ...
##  $ end_station_id    : chr [1:701339] "13084" NA NA NA ...
##  $ start_lat         : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:701339] "casual" "casual" "casual" "casual" ...
##  - 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(y22_oct)
```

```
## spc_tbl_ [558,685 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:558685] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F
873" "47E653FDC2D99236" ...
##  $ rideable_type     : chr [1:558685] "classic_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:558685], format: "2022-10-14 17:13:30" "2022-10-01 16:29:
26" ...
##  $ ended_at          : POSIXct[1:558685], format: "2022-10-14 17:19:39" "2022-10-01 16:49:
06" ...
##  $ start_station_name: chr [1:558685] "Noble St & Milwaukee Ave" "Damen Ave & Charleston S
t" "Hoyne Ave & Balmoral Ave" "Rush St & Cedar St" ...
##  $ start_station_id  : chr [1:558685] "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr [1:558685] "Larrabee St & Division St" "Damen Ave & Cullerton S
t" "Western Ave & Leland Ave" "Orleans St & Chestnut St (NEXT Apts)" ...
##  $ end_station_id    : chr [1:558685] "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat         : num [1:558685] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:558685] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:558685] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:558685] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:558685] "member" "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(y22_nov)
```

```
## spc_tbl_ [337,735 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:337735] "BCC66FC6FAB27CC7" "772AB67E902C180F" "585EAD07FDEC0
152" "91C4E7ED3C262FF9" ...
##  $ rideable_type     : chr [1:337735] "electric_bike" "classic_bike" "classic_bike" "class
ic_bike" ...
##  $ started_at        : POSIXct[1:337735], format: "2022-11-10 06:21:55" "2022-11-04 07:31:
55" ...
##  $ ended_at          : POSIXct[1:337735], format: "2022-11-10 06:31:27" "2022-11-04 07:46:
25" ...
##  $ start_station_name: chr [1:337735] "Canal St & Adams St" "Canal St & Adams St" "Indiana
Ave & Roosevelt Rd" "Indiana Ave & Roosevelt Rd" ...
##  $ start_station_id  : chr [1:337735] "13011" "13011" "SL-005" "SL-005" ...
##  $ end_station_name  : chr [1:337735] "St. Clair St & Erie St" "St. Clair St & Erie St" "S
t. Clair St & Erie St" "St. Clair St & Erie St" ...
##  $ end_station_id    : chr [1:337735] "13016" "13016" "13016" "13016" ...
##  $ start_lat         : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:337735] "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(y22_dec)
```

```
## spc_tbl_ [181,806 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:181806] "65DBD2F447EC51C2" "0C201AA7EA0EA1AD" "E0B148CCB358A
49D" "54C5775D2B7C9188" ...
##  $ rideable_type     : chr [1:181806] "electric_bike" "classic_bike" "electric_bike" "clas
sic_bike" ...
##  $ started_at        : POSIXct[1:181806], format: "2022-12-05 10:47:18" "2022-12-18 06:42:
33" ...
##  $ ended_at          : POSIXct[1:181806], format: "2022-12-05 10:56:34" "2022-12-18 07:08:
44" ...
##  $ start_station_name: chr [1:181806] "Clifton Ave & Armitage Ave" "Broadway & Belmont Av
e" "Sangamon St & Lake St" "Shields Ave & 31st St" ...
##  $ start_station_id  : chr [1:181806] "TA1307000163" "13277" "TA1306000015" "KA1503000038"
...
##  $ end_station_name  : chr [1:181806] "Sedgwick St & Webster Ave" "Sedgwick St & Webster A
ve" "St. Clair St & Erie St" "Damen Ave & Madison St" ...
##  $ end_station_id    : chr [1:181806] "13191" "13191" "13016" "13134" ...
##  $ start_lat         : num [1:181806] 41.9 41.9 41.9 41.8 41.9 ...
##  $ start_lng         : num [1:181806] -87.7 -87.6 -87.7 -87.6 -87.7 ...
##  $ end_lat           : num [1:181806] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:181806] -87.6 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:181806] "member" "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>
```

Merge Data

```
y22_merged <- bind_rows(y22_jan, y22_feb, y22_mar, y22_apr, y22_may, y22_jun, y22_jul, y22_au
g, y22_sep, y22_oct, y22_nov, y22_dec)
```

Inspect the new table that has been created

```
colnames(y22_merged)
```

```
##  [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"
```

```
nrow(y22_merged)
```

```
## [1] 5667717
```

```
dim(y22_merged)
```

```
## [1] 5667717      13
```

```
head(y22_merged)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ started_at          ended_at            start…² start…³
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 C2F7DD78E82EC… electr… 2022-01-13 11:59:47 2022-01-13 12:02:44 Glenwo… 525
## 2 A6CF8980A652D… electr… 2022-01-10 08:41:56 2022-01-10 08:46:17 Glenwo… 525
## 3 BD0F91DFF741C… classi… 2022-01-25 04:53:40 2022-01-25 04:58:01 Sheffi… TA1306…
## 4 CBB80ED419105… classi… 2022-01-04 00:18:04 2022-01-04 00:33:00 Clark … KA1504…
## 5 DDC963BFDDA51… classi… 2022-01-20 01:31:10 2022-01-20 01:37:12 Michig… TA1309…
## 6 A39C6F6CC0586… classi… 2022-01-11 18:48:09 2022-01-11 18:51:31 Wood S… 637
## # … with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²start_station_name, ³start_station_id
```

```
tail(y22_merged)
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ started_at          ended_at            start…² start…³
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 7BDEDE9860418… classi… 2022-12-07 06:52:45 2022-12-07 06:56:36 Sangam… 13409
## 2 43ABEE85B6E15… classi… 2022-12-05 06:51:04 2022-12-05 06:54:48 Sangam… 13409
## 3 F041C89A3D1F0… electr… 2022-12-14 17:06:28 2022-12-14 17:19:27 Bernar… 18016
## 4 A2BECB88430BE… classi… 2022-12-08 16:27:47 2022-12-08 16:32:20 Wacker… KA1503…
## 5 37B392960E566… classi… 2022-12-28 09:37:38 2022-12-28 09:41:34 Sangam… 13409
## 6 2DD1587210BA4… classi… 2022-12-09 00:27:25 2022-12-09 00:35:28 Southp… 13235
## # … with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²start_station_name, ³start_station_id
```

```
str(y22_merged)
```

```
## spc_tbl_ [5,667,717 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5667717] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741
C66D" "CBB80ED419105406" ...
##  $ rideable_type     : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "cla
ssic_bike" ...
##  $ started_at        : POSIXct[1:5667717], format: "2022-01-13 11:59:47" "2022-01-10 08:4
1:56" ...
##  $ ended_at          : POSIXct[1:5667717], format: "2022-01-13 12:02:44" "2022-01-10 08:4
6:17" ...
##  $ start_station_name: chr [1:5667717] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Av
e" "Sheffield Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...
##  $ start_station_id  : chr [1:5667717] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:5667717] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Gree
nview Ave & Fullerton Ave" "Paulina St & Montrose Ave" ...
##  $ end_station_id    : chr [1:5667717] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ start_lng         : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ end_lng           : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:5667717] "casual" "casual" "member" "casual" ...
##  - 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>
```

```
summary(y22_merged)
```

```
##    ride_id          rideable_type        started_at
##  Length:5667717     Length:5667717     Min.   :2022-01-01 00:00:05.00
##  Class :character   Class :character   1st Qu.:2022-05-28 19:21:05.00
##  Mode  :character   Mode  :character   Median :2022-07-22 15:03:59.00
##                                        Mean   :2022-07-20 07:21:18.74
##                                        3rd Qu.:2022-09-16 07:21:29.00
##                                        Max.   :2022-12-31 23:59:26.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-01-01 00:01:48.00   Length:5667717     Length:5667717
##  1st Qu.:2022-05-28 19:43:07.00   Class :character   Class :character
##  Median :2022-07-22 15:24:44.00   Mode  :character   Mode  :character
##  Mean   :2022-07-20 07:40:45.33
##  3rd Qu.:2022-09-16 07:39:03.00
##  Max.   :2023-01-02 04:56:45.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5667717     Length:5667717     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :45.64   Max.   :-73.80
##
##     end_lat         end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5667717
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.37   Max.   :  0.00
##  NA's   :5858    NA's   :5858
```

Step 03 - PROCESS

Documentation of any cleaning or manipulation of data. Transforming the data so you can work with it effectively.

Add columns that list day of the week and month.This will allow us to aggregate ride data for each day and each month. We will add “day_of_week” and “month”. More on date formats are found here (https://www.statmethods.net/input/dates.html).

```
y22_merged$month <- format(as.Date(y22_merged$started_at), "%b")

y22_merged$day_of_week <- format(as.Date(y22_merged$started_at), "%A")
```

We will want to add a calculated field for length of ride since the data did not have the “trip duration” column. We will add “ride_length”(in seconds) to the entire dataframe for consistency.

```
y22_merged$ride_length <- difftime(y22_merged$ended_at,
                                   y22_merged$started_at)
```

View and inspect the structure of columns

```
str(y22_merged)
```

```
## spc_tbl_ [5,667,717 × 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5667717] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741
C66D" "CBB80ED419105406" ...
##  $ rideable_type     : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "cla
ssic_bike" ...
##  $ started_at        : POSIXct[1:5667717], format: "2022-01-13 11:59:47" "2022-01-10 08:4
1:56" ...
##  $ ended_at          : POSIXct[1:5667717], format: "2022-01-13 12:02:44" "2022-01-10 08:4
6:17" ...
##  $ start_station_name: chr [1:5667717] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Av
e" "Sheffield Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...
##  $ start_station_id  : chr [1:5667717] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:5667717] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Gree
nview Ave & Fullerton Ave" "Paulina St & Montrose Ave" ...
##  $ end_station_id    : chr [1:5667717] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ start_lng         : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ end_lng           : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:5667717] "casual" "casual" "member" "casual" ...
##  $ month             : chr [1:5667717] "Jan" "Jan" "Jan" "Jan" ...
##  $ day_of_week       : chr [1:5667717] "Thursday" "Monday" "Tuesday" "Tuesday" ...
##  $ ride_length       : 'difftime' num [1:5667717] 177 261 261 896 ...
##   ..- attr(*, "units")= chr "secs"
##  - 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>
```

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

```
y22_merged$ride_length <- as.numeric(as.character(y22_merged$ride_length))
is.numeric(y22_merged$ride_length)
```

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

Remove “Bad” Data

The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or ride_length was zero or negative.

We will create a new dataframe since data is being removed. Delete or Drop rows conditions (https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/).

```
y22_merged_clean<- y22_merged[!(y22_merged$ride_length <= 0),]
```

Step 04 - ANALYZE

Perform calculations.Identify trends and relationships.

Conducting descriptive analysis on “ride_length”.

```
summary(y22_merged_clean$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     349     617    1167    1108 2483235
```

Compare members and casual users.

```
aggregate(y22_merged_clean$ride_length ~ y22_merged_clean$member_casual, FUN = mean)
```

```
##   y22_merged_clean$member_casual y22_merged_clean$ride_length
## 1                         casual                    1748.9000
## 2                         member                     762.8915
```

```
aggregate(y22_merged_clean$ride_length ~ y22_merged_clean$member_casual, FUN = median)
```

```
##   y22_merged_clean$member_casual y22_merged_clean$ride_length
## 1                         casual                          780
## 2                         member                          530
```

```
aggregate(y22_merged_clean$ride_length ~ y22_merged_clean$member_casual, FUN = max)
```

```
##   y22_merged_clean$member_casual y22_merged_clean$ride_length
## 1                         casual                       2483235
## 2                         member                         93594
```

```
aggregate(y22_merged_clean$ride_length ~ y22_merged_clean$member_casual, FUN = min)
```

```
##   y22_merged_clean$member_casual y22_merged_clean$ride_length
## 1                         casual                            1
## 2                         member                            1
```

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

```
y22_merged_clean$day_of_week <- ordered(y22_merged_clean$day_of_week,
                                        levels=c("Sunday", "Monday", "Tuesday", "Wednesday",
"Thursday", "Friday", "Saturday"))
  aggregate(y22_merged_clean$ride_length ~ y22_merged_clean$member_casual +
            y22_merged_clean$day_of_week, FUN = mean )
```

```
##    y22_merged_clean$member_casual y22_merged_clean$day_of_week
## 1                          casual                       Sunday
## 2                          member                       Sunday
## 3                          casual                       Monday
## 4                          member                       Monday
## 5                          casual                      Tuesday
## 6                          member                      Tuesday
## 7                          casual                    Wednesday
## 8                          member                    Wednesday
## 9                          casual                     Thursday
## 10                         member                     Thursday
## 11                         casual                       Friday
## 12                         member                       Friday
## 13                         casual                     Saturday
## 14                         member                     Saturday
##    y22_merged_clean$ride_length
## 1                     2043.6343
## 2                      841.9355
## 3                     1751.3805
## 4                      736.2531
## 5                     1549.5189
## 6                      727.8171
## 7                     1485.1319
## 8                      726.3364
## 9                     1533.0238
## 10                     737.6191
## 11                    1682.8110
## 12                     751.8978
## 13                    1957.0725
## 14                     848.4573
```

analyze ridership data by type and weekday

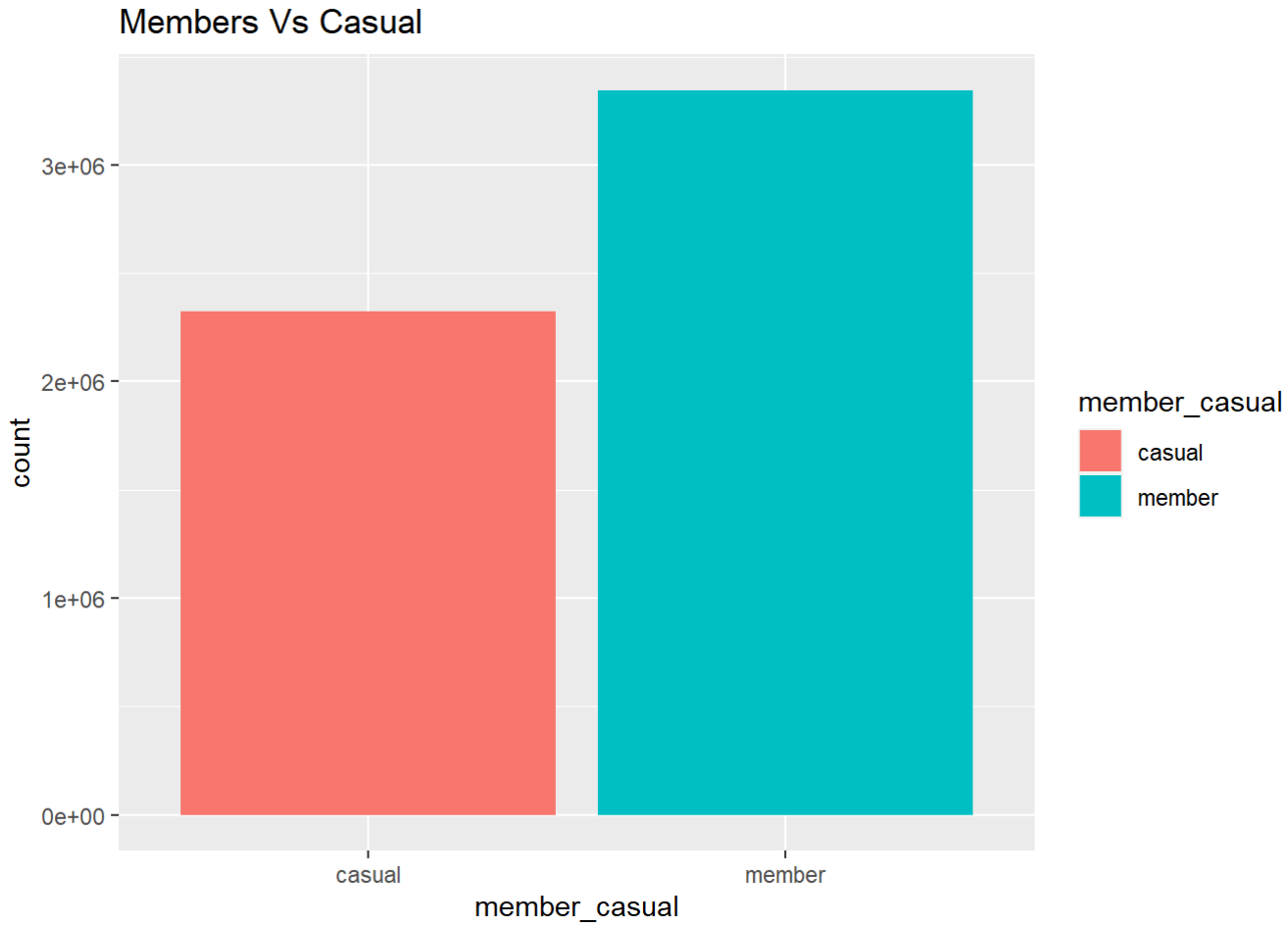
```
y22_merged_clean %>%
  mutate(weekday = wday(started_at, label= TRUE)) %>%
  group_by(member_casual,weekday) %>%
  dplyr::summarise(number_of_rides= n(), .groups="drop",
                   average_duration= mean(ride_length)) %>%
  arrange(member_casual,weekday)
```

```
## # A tibble: 14 × 4
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              388981            2044.
##  2 casual        Mon              277649            1751.
##  3 casual        Tue              263706            1550.
##  4 casual        Wed              274339            1485.
##  5 casual        Thu              309297            1533.
##  6 casual        Fri              334667            1683.
##  7 casual        Sat              473130            1957.
##  8 member        Sun              387180             842.
##  9 member        Mon              473305             736.
## 10 member        Tue              518584             728.
## 11 member        Wed              523836             726.
## 12 member        Thu              532215             738.
## 13 member        Fri              467051             752.
## 14 member        Sat              443246             848.
```

Visualizing data

Let's visualize members and casuals by the total ride taken

```
ggplot(data = y22_merged_clean)+
  geom_bar(mapping= aes(x= member_casual, fill= member_casual)) +
  labs(title = "Members Vs Casual")
```

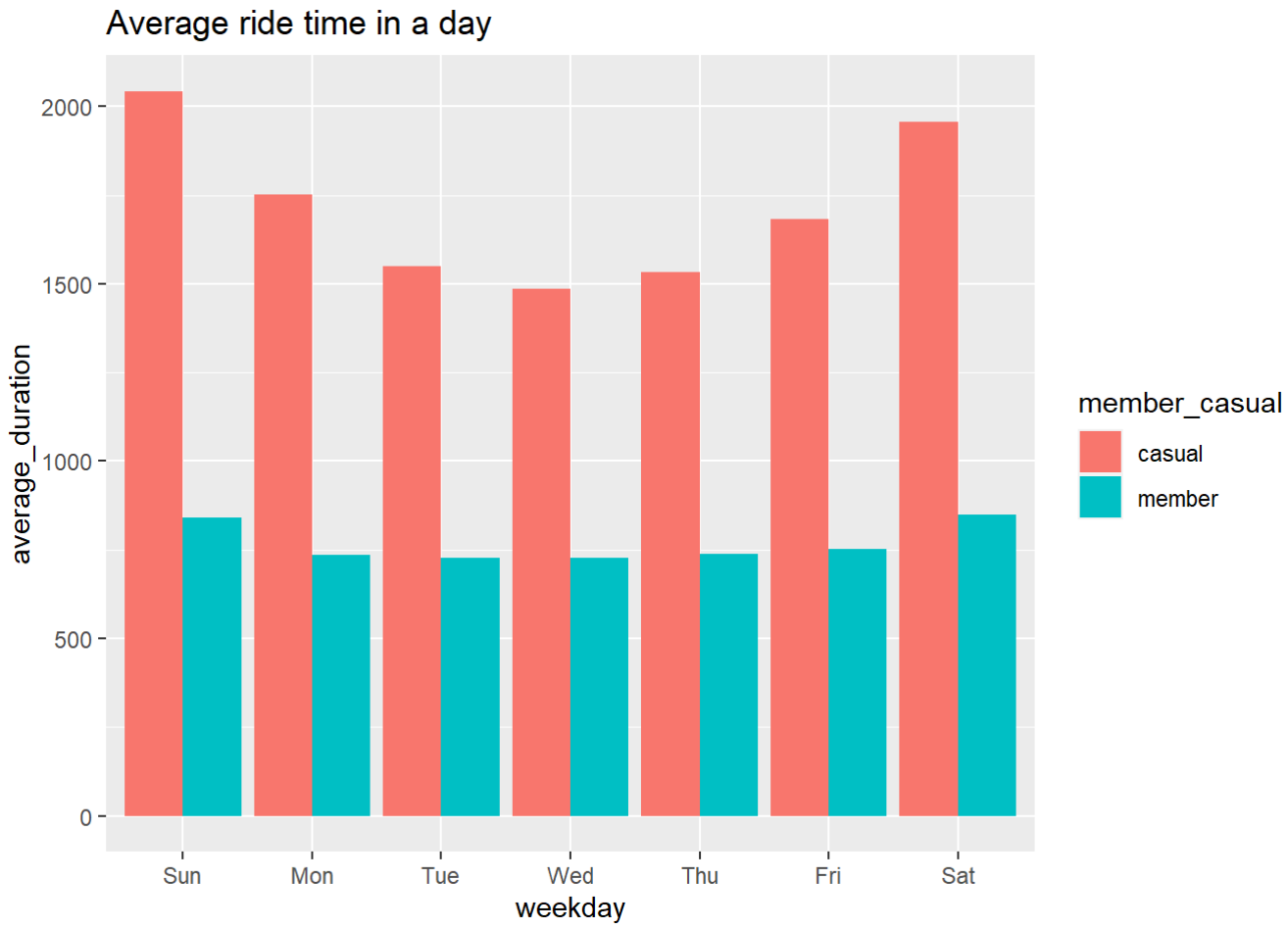


Let's Visualize the number of rides by rider type

```
y22_merged_clean %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  dplyr::summarise(number_of_rides = n(), .groups="drop",
                   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 = "Total rides in a day")
```

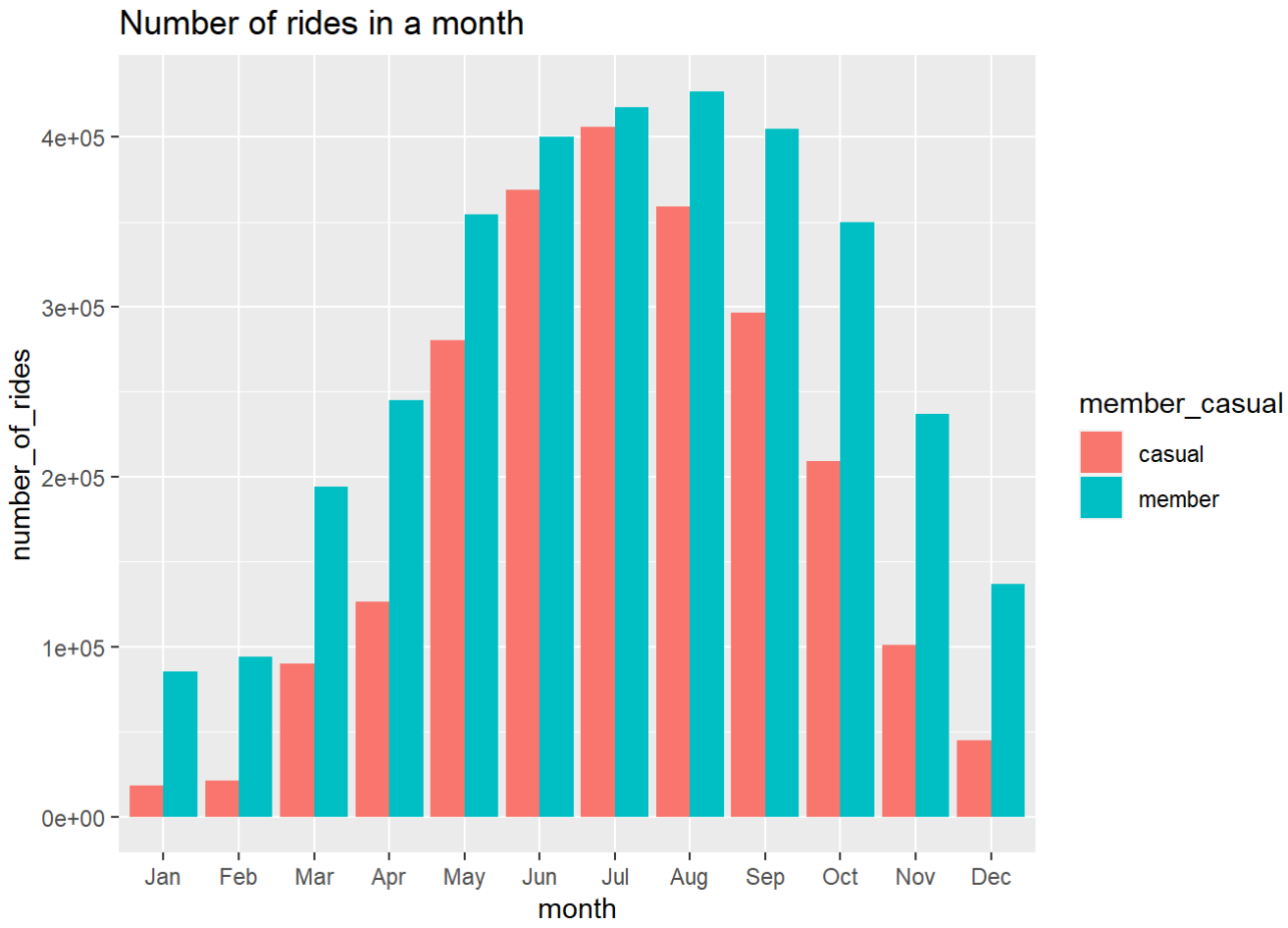
Let's create a visualization for average duration

```
y22_merged_clean %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  dplyr::summarise(number_of_rides = n(), .groups="drop",
                   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 time in a day")
```



Let's visualize the total rides taken by members and casuals by month

```
  y22_merged_clean$month <- ordered(y22_merged_clean$month,
                                    levels=c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "J
ul", "Aug", "Sep", "Oct", "Nov", "Dec"))
y22_merged_clean %>%
  group_by(member_casual, month) %>%
  dplyr::summarise(number_of_rides = n(), .groups="drop") %>%
  arrange(member_casual,month) %>%
  ggplot(aes(x= month, y= number_of_rides, fill= member_casual))+
  geom_col(position = "dodge") +
  labs(title = "Number of rides in a month")
```



Step 05 - SHARE

Present your findings. Ensure your work is accessible. You can view my work here (https://github.com/Vinay-Sathupati/Case-Study--Cyclistic-bike-share.git).

Findings:

- membership people are more compared with casual members.
- members renting bikes show consistency throughout the week. Whereas, casual are lowest during week and higher than members during weekend.
- This data also shows a seasonal pattern as Chicago's climate is typically continental with cold winters, warm summers. Lowest usage of bikes is in winters and Highest during summer.

Step 06 - ACT

My recommendations:

- Provide weekend and/or seasonal memberships.
- provide discounts during summer, since bike usage is peak during this season.
- Host campaigns, events to attract more members to get membership.