Cyclistic Data Analysis

Edgar Elias

2022-05-31

Contents

Introduction

This case study was conducted as my capstone project for the 8th and final course of the **Google Data Analytics Professional Certificate** program to help students demonstrate the knowledge and skills gained during the whole course. I applied Google's procedures (Ask, Prepare, Process, Analyze, Share, and Act) to my own findings and analysis which are displayed below along with visualizations.

Scenario

The purpose of this script is to consolidate downloaded bike data and conduct analysis for Cyclistic, a bike-share company based in Chicago, Illinois. The director of marketing believes the company's primary path to success depends on maximizing the number of annual memberships amongs it's customers. Cyclistic's finance analysts have found that annual member riders are much more profitable than casual riders. Therefore the

marketing team is aimed at converting casual riders to annual memebers, but in order to do so, the marketing team needs better understanding on how annual members and casual riders differ and what can be done to maximize the conversion of casual riders into a membership.

Business Task

As my business task, I was given the key question to answer: "How do member and casual riders use Cyclistic bikes differently?"

Data Sources

The data has been provided by Motivate International Inc. under this *license*. The data is a collection of 12 months of bike trip data starting from April 2021 to March 2022. The procedure for my subsequent analysis was conducted in R Studio since this program could efficiently help in data importing, combining, cleaning, filtering, analyzing and visualizing.

Data Cleaning and Manipulations

Setting up environment

1. Load following libraries to inspect, wrangle, clean and manipulate the data

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
##
## Attaching package: 'lubridate'

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

```
library(ggplot2)  #helps visualize data
library(dplyr) #helps manipulate data
library(knitr) #helps filter out data
getwd() #displays your working directory
```

```
## [1] "/Users/EdgarElias/Desktop/CaseStudy/Tripdata"
```

```
setwd("/Users/EdgarElias/Desktop/CaseStudy/Tripdata") #sets working directory to simplify calls to data
```

2. Collect Data & Upload tripdata sets (csv files) of last 12 months.

```
apr_2021 <- read_csv("202104-divvy-tripdata.csv")
```

```
## Rows: 337230 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.
```

```
may_2021 <- read_csv("202105-divvy-tripdata.csv")
```

```
## Rows: 531633 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.
```

```
jun_2021 <- read_csv("202106-divvy-tripdata.csv")
```

```
## Rows: 729595 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.
```

```
jul_2021 <- read_csv("202107-divvy-tripdata.csv")
```

```
## Rows: 822410 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.
```

```
aug_2021 <- read_csv("202108-divvy-tripdata.csv")
```

```
## Rows: 804352 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.
```

```
sep_2021 <- read_csv("202109-divvy-tripdata.csv")
```

```
## Rows: 756147 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.
```

```
oct_2021 <- read_csv("202110-divvy-tripdata.csv")
```

```
## Rows: 631226 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.
```

```
nov_2021 <- read_csv("202111-divvy-tripdata.csv")
```

```
## Rows: 359978 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.
```

```
dec_2021 <- read_csv("202112-divvy-tripdata.csv")
```

```
## Rows: 247540 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.
```

```
jan_2022 <- read_csv("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.
```

```
feb_2022 <- read_csv("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.
```

```
mar_2022 <- read_csv("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.
```

3. Wrangling and combining data into a single file (Make sure column names match)

```
colnames(apr_2021)
```

```
##  [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(may_2021)
```

```
##  [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(jun_2021)
```

```
##  [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(jul_2021)
```

```
##  [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(aug_2021)
```

```
##  [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(sep_2021)
```

```
##  [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(oct_2021)
```

```
##  [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(nov_2021)
```

```
##  [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(dec_2021)
```

```
##  [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(jan_2022)
```

```
##  [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(feb_2022)
```

```
##  [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(mar_2022)
```

```
##  [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"
```

4. Inspect data frames for incongruences

```
str(apr_2021)
```

```
## spec_tbl_df [337,230 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:337230] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887
##  $ rideable_type     : chr [1:337230] "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:337230], format: "2021-04-12 18:25:36" "2021-04-27 17:27:11" ...
##  $ ended_at          : POSIXct[1:337230], format: "2021-04-12 18:56:55" "2021-04-27 18:31:29" ...
##  $ start_station_name: chr [1:337230] "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blv
##  $ start_station_id  : chr [1:337230] "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr [1:337230] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loo
##  $ end_station_id    : chr [1:337230] "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num [1:337230] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num [1:337230] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:337230] "member" "casual" "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(may_2021)
```

```
## spec_tbl_df [531,633 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881
##  $ rideable_type     : chr [1:531633] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:531633], format: "2021-05-30 11:58:15" "2021-05-30 11:29:14" ...
##  $ ended_at          : POSIXct[1:531633], format: "2021-05-30 12:10:39" "2021-05-30 12:14:09" ...
##  $ start_station_name: chr [1:531633] NA NA NA NA ...
##  $ start_station_id  : chr [1:531633] NA NA NA NA ...
##  $ end_station_name  : chr [1:531633] NA NA NA NA ...
##  $ end_station_id    : chr [1:531633] NA NA NA NA ...
##  $ start_lat         : num [1:531633] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:531633] 41.9 41.8 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:531633] "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(jun_2021)
```

```
## spec_tbl_df [729,595 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:729595] "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C
##  $ rideable_type     : chr [1:729595] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:729595], format: "2021-06-13 14:31:28" "2021-06-04 11:18:02" ...
##  $ ended_at          : POSIXct[1:729595], format: "2021-06-13 14:34:11" "2021-06-04 11:24:19" ...
##  $ start_station_name: chr [1:729595] NA NA NA NA ...
##  $ start_station_id  : chr [1:729595] NA NA NA NA ...
##  $ end_station_name  : chr [1:729595] NA NA NA NA ...
##  $ end_station_id    : chr [1:729595] NA NA NA NA ...
##  $ start_lat         : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
##  $ start_lng         : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
##  $ end_lng           : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:729595] "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(jul_2021)
```

```
## spec_tbl_df [822,410 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
```

```
##  $ ride_id           : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B5
##  $ rideable_type     : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:822410], format: "2021-07-02 14:44:36" "2021-07-07 16:57:42" ...
##  $ ended_at          : POSIXct[1:822410], format: "2021-07-02 15:19:58" "2021-07-07 17:16:09" ...
##  $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez St" "Wa
##  $ start_station_id  : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St
##  $ end_station_id    : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:822410] "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(aug_2021)
```

```
## spec_tbl_df [804,352 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834
##  $ rideable_type     : chr [1:804352] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:804352], format: "2021-08-10 17:15:49" "2021-08-10 17:23:14" ...
##  $ ended_at          : POSIXct[1:804352], format: "2021-08-10 17:22:44" "2021-08-10 17:39:24" ...
##  $ start_station_name: chr [1:804352] NA NA NA NA ...
##  $ start_station_id  : chr [1:804352] NA NA NA NA ...
##  $ end_station_name  : chr [1:804352] NA NA NA NA ...
##  $ end_station_id    : chr [1:804352] NA NA NA NA ...
##  $ start_lat         : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num [1:804352] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num [1:804352] -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:804352] "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(sep_2021)
```

```
## spec_tbl_df [756,147 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1
##  $ rideable_type     : chr [1:756147] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:756147], format: "2021-09-28 16:07:10" "2021-09-28 14:24:51" ...
##  $ ended_at          : POSIXct[1:756147], format: "2021-09-28 16:09:54" "2021-09-28 14:40:05" ...
##  $ start_station_name: chr [1:756147] NA NA NA NA ...
##  $ start_station_id  : chr [1:756147] NA NA NA NA ...
##  $ end_station_name  : chr [1:756147] NA NA NA NA ...
##  $ end_station_id    : chr [1:756147] NA NA NA NA ...
##  $ start_lat         : num [1:756147] 41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num [1:756147] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:756147] 41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num [1:756147] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:756147] "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(oct_2021)
```

```
## spec_tbl_df [631,226 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:631226] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "3629
##  $ rideable_type     : chr [1:631226] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:631226], format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct[1:631226], format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr [1:631226] "Kingsbury St & Kinzie St" NA NA NA ...
```

```
##  $ start_station_id  : chr [1:631226] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:631226] NA NA NA NA ...
##  $ end_station_id    : chr [1:631226] NA NA NA NA ...
##  $ start_lat         : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:631226] "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(nov_2021)
```

```
## spec_tbl_df [359,978 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:359978] "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3B
##  $ rideable_type     : chr [1:359978] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:359978], format: "2021-11-27 13:27:38" "2021-11-27 13:38:25" ...
##  $ ended_at          : POSIXct[1:359978], format: "2021-11-27 13:46:38" "2021-11-27 13:56:10" ...
##  $ start_station_name: chr [1:359978] NA NA NA NA ...
##  $ start_station_id  : chr [1:359978] NA NA NA NA ...
##  $ end_station_name  : chr [1:359978] NA NA NA NA ...
##  $ end_station_id    : chr [1:359978] NA NA NA NA ...
##  $ start_lat         : num [1:359978] 41.9 42 42 41.9 41.9 ...
##  $ start_lng         : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ end_lat           : num [1:359978] 42 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ member_casual     : chr [1:359978] "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(dec_2021)
```

```
## spec_tbl_df [247,540 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:247540] "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278
##  $ rideable_type     : chr [1:247540] "electric_bike" "electric_bike" "electric_bike" "classic_bike"
##  $ started_at        : POSIXct[1:247540], format: "2021-12-07 15:06:07" "2021-12-11 03:43:29" ...
##  $ ended_at          : POSIXct[1:247540], format: "2021-12-07 15:13:42" "2021-12-11 04:10:23" ...
##  $ start_station_name: chr [1:247540] "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted St
##  $ start_station_id  : chr [1:247540] "13307" "KP1705001026" "KA1504000117" "KA1504000117" ...
##  $ end_station_name  : chr [1:247540] "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Broadway &
##  $ end_station_id    : chr [1:247540] "TA1307000130" "TA1307000119" "13137" "KP1705001026" ...
##  $ start_lat         : num [1:247540] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:247540] -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:247540] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:247540] -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:247540] "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(jan_2022)
```

```
## spec_tbl_df [103,770 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:103770] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB8
##  $ rideable_type     : chr [1:103770] "electric_bike" "electric_bike" "classic_bike" "classic_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 Ave" "Sheffiel
##  $ start_station_id  : chr [1:103770] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:103770] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview 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(feb_2022)
```

```
## spec_tbl_df [115,609 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:115609] "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A178
##  $ rideable_type     : chr [1:115609] "classic_bike" "classic_bike" "classic_bike" "classic_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 Ave" "State S
##  $ start_station_id  : chr [1:115609] "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...
##  $ end_station_name  : chr [1:115609] "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Cana
##  $ 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(mar_2022)
```

```
## spec_tbl_df [284,042 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:284042] "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F44
##  $ rideable_type     : chr [1:284042] "classic_bike" "electric_bike" "classic_bike" "classic_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" "Broadway & Be
##  $ 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
##  $ 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>
```

Combine data

1. Individual trip data frames are combined into one large data frame with name "all_trip_data"

```
all_trip_data <- bind_rows(apr_2021, may_2021, jun_2021, jul_2021, aug_2021, sep_2021, oct_2021, nov_20
```

2. Inspect "all_trip_data"

```
colnames(all_trip_data) #List of column names
```

```
##  [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(all_trip_data) #How many rows are in data frame?
```

```
## [1] 5723532
```

```
dim(all_trip_data) #Dimensions of the data frame?
```

```
## [1] 5723532      13
```

```
head(all_trip_data) #See the first 10 rows of data frame.
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 6C992B~ classic_bike  2021-04-12 18:25:36 2021-04-12 18:56:55 State St & Pear~
## 2 1E0145~ docked_bike   2021-04-27 17:27:11 2021-04-27 18:31:29 Dorchester Ave ~
## 3 E498E1~ docked_bike   2021-04-03 12:42:45 2021-04-07 11:40:24 Loomis Blvd & 8~
## 4 188726~ classic_bike  2021-04-17 09:17:42 2021-04-17 09:42:48 Honore St & Div~
## 5 C12354~ docked_bike   2021-04-03 12:42:25 2021-04-03 14:13:42 Loomis Blvd & 8~
## 6 097E76~ classic_bike  2021-04-25 18:43:18 2021-04-25 18:43:59 Clinton St & Po~
## # ... with 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

```
tail(all_trip_data) #See the last 10 rows of data frame.
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 A2A6F0~ electric_bike 2022-03-09 20:29:48 2022-03-09 21:01:30 Sheridan Rd & I~
## 2 E23BE3~ docked_bike   2022-03-13 16:31:03 2022-03-13 16:39:32 Michigan Ave & ~
## 3 15AF71~ docked_bike   2022-03-09 06:56:02 2022-03-09 07:42:14 Broadway & Barr~
## 4 9C4CE6~ electric_bike 2022-03-09 15:55:26 2022-03-09 16:08:54 <NA>
## 5 F4E136~ electric_bike 2022-03-21 16:12:44 2022-03-21 16:18:24 <NA>
## 6 5AEC5F~ classic_bike  2022-03-03 18:13:40 2022-03-03 18:23:39 Clark St & Rand~
## # ... with 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

```
str(all_trip_data) #See list of columns and data types EX: numeric, character, etc.
```

```
## spec_tbl_df [5,723,532 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5723532] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "188
##  $ rideable_type     : chr [1:5723532] "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:5723532], format: "2021-04-12 18:25:36" "2021-04-27 17:27:11" ...
##  $ ended_at          : POSIXct[1:5723532], format: "2021-04-12 18:56:55" "2021-04-27 18:31:29" ...
##  $ start_station_name: chr [1:5723532] "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Bl
##  $ start_station_id  : chr [1:5723532] "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr [1:5723532] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Lo
##  $ end_station_id    : chr [1:5723532] "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num [1:5723532] 41.9 41.8 41.7 41.9 41.7 ...
```

```
##  $ start_lng         : num [1:5723532] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5723532] 41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num [1:5723532] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5723532] "member" "casual" "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>
```

```
summary(all_trip_data) #Statistical summary of data for numeric info
```

```
##    ride_id          rideable_type        started_at
##  Length:5723532     Length:5723532     Min.   :2021-04-01 00:03:18
##  Class :character   Class :character   1st Qu.:2021-06-22 15:20:26
##  Mode  :character   Mode  :character   Median :2021-08-17 18:25:49
##                                        Mean   :2021-08-26 22:25:18
##                                        3rd Qu.:2021-10-14 19:48:10
##                                        Max.   :2022-03-31 23:59:47
##
##     ended_at                    start_station_name start_station_id
##  Min.   :2021-04-01 00:14:29   Length:5723532     Length:5723532
##  1st Qu.:2021-06-22 15:47:37   Class :character   Class :character
##  Median :2021-08-17 18:44:32   Mode  :character   Mode  :character
##  Mean   :2021-08-26 22:46:50
##  3rd Qu.:2021-10-14 20:03:28
##  Max.   :2022-04-01 22:10:12
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5723532     Length:5723532     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.   :41.39   Min.   :-88.97   Length:5723532
##  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.   :42.17   Max.   :-87.49
##  NA's   :4716    NA's   :4716
```

Clean and Organize Data

1. Create copy (all_trip_data_v2) of combined data frame (all_trip_data) for cleaning and transforming it, starting with removing duplicate rides by checking ride_id

```
all_trip_data_v2 <- all_trip_data[!duplicated(all_trip_data$ride_id), ]
```

2. Remove unused columns (latitudes, longitudes, start and end station id's)

```
all_trip_data_v2 = subset(all_trip_data_v2, select = -c(start_station_id, end_station_id, start_lat, en
```

3. Add columns from started_at observation for matching month, day, year, day_of_the_week of each trip/ride

```
all_trip_data_v2$month <- format(as.Date(all_trip_data_v2$started_at), "%b") #adds month column
all_trip_data_v2$day <- format(as.Date(all_trip_data_v2$started_at), "%d") #adds day column
all_trip_data_v2$year <- format(as.Date(all_trip_data_v2$started_at), "%y") #adds year
all_trip_data_v2$day_of_week <- format(as.Date(all_trip_data_v2$started_at), "%a") #adds character day
```

4. Create a column called "ride_length." Calculating the length of each ride by subtracting the column "started_at" from the column "ended_at".

```
all_trip_data_v2$ride_length <- difftime(all_trip_data_v2$ended_at, all_trip_data_v2$started_at, unit =
```

5. Remove column started_at, ended_at since we have a split started column consisting of month, day and year and not concerned of ended_at since it was already used to calculated ride_length

```
all_trip_data_v2 = subset(all_trip_data_v2, select = -c(started_at, ended_at))
```

6. Inspect structure of new all_trip_data_v2 columns

```
str(all_trip_data_v2)
```

```
## tibble [5,723,532 x 10] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5723532] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "188
##  $ rideable_type     : chr [1:5723532] "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ start_station_name: chr [1:5723532] "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Bl
##  $ end_station_name  : chr [1:5723532] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Lo
##  $ member_casual     : chr [1:5723532] "member" "casual" "casual" "member" ...
##  $ month             : chr [1:5723532] "Apr" "Apr" "Apr" "Apr" ...
##  $ day               : chr [1:5723532] "12" "27" "03" "17" ...
##  $ year              : chr [1:5723532] "21" "21" "21" "21" ...
##  $ day_of_week       : chr [1:5723532] "Mon" "Tue" "Sat" "Sat" ...
##  $ ride_length       : 'difftime' num [1:5723532] 31.3166666666667 64.3 5697.65 25.1 ...
##   ..- attr(*, "units")= chr "mins"
```

7. Change ride_length to a numeric data type

```
all_trip_data_v2$ride_length <- as.numeric(all_trip_data_v2$ride_length)
```

8. Ensure ride_length is a numeric type

```
is.numeric(all_trip_data_v2$ride_length)
```

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

9. Remove trips with no trip duration and false station names like HQ (ride length <= 0 min)

```
all_trip_data_v2 <- all_trip_data_v2[!(all_trip_data_v2$start_station_name == "HQ QR" | all_trip_data_v
```

10. Remove NA rows from "member_casual" column

```
all_trip_data_v2 <- all_trip_data_v2[!is.na(all_trip_data_v2$member_casual), ]
```

11. Remove outlier trips with trip duration longer than 48hrs (ride length > 2880 min)

```
all_trip_data_v2 <- all_trip_data_v2[!(all_trip_data_v2$ride_length > 2880), ]
```

Prepare for Analysis

1. Statistical view of all_trip_data_v2

```
summary(all_trip_data_v2)
```

```
##    ride_id          rideable_type      start_station_name end_station_name
##  Length:4976727     Length:4976727     Length:4976727     Length:4976727
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  member_casual         month               day                year
##  Length:4976727     Length:4976727     Length:4976727     Length:4976727
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  day_of_week         ride_length
##  Length:4976727     Min.   :   0.0167
##  Class :character   1st Qu.:   6.7667
##  Mode  :character   Median :  11.9333
##                     Mean   :  20.6974
##                     3rd Qu.:  21.6833
##                     Max.   :2868.4500
```

```
str(all_trip_data_v2)
```

```
## tibble [4,976,727 x 10] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4976727] "6C992BD37A98A63F" "1E0145613A209000" "1887262AD101C604" "C123
##  $ rideable_type     : chr [1:4976727] "classic_bike" "docked_bike" "classic_bike" "docked_bike" ...
##  $ start_station_name: chr [1:4976727] "State St & Pearson St" "Dorchester Ave & 49th St" "Honore St
##  $ end_station_name  : chr [1:4976727] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Sou
##  $ member_casual     : chr [1:4976727] "member" "casual" "member" "casual" ...
##  $ month             : chr [1:4976727] "Apr" "Apr" "Apr" "Apr" ...
##  $ day               : chr [1:4976727] "12" "27" "17" "03" ...
##  $ year              : chr [1:4976727] "21" "21" "21" "21" ...
##  $ day_of_week       : chr [1:4976727] "Mon" "Tue" "Sat" "Sat" ...
##  $ ride_length       : num [1:4976727] 31.317 64.3 25.1 91.283 0.683 ...
```

```
sum(str_count(all_trip_data_v2$member_casual, "casual")) #total sum of casual riders
```

```
## [1] 2213538
```

```
sum(str_count(all_trip_data_v2$member_casual, "member")) #total sum of member riders
```

```
## [1] 2763189
```

```
summary(all_trip_data_v2$ride_length) #summary of ride_length
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0167    6.7667   11.9333   20.6974   21.6833 2868.4500
```

2. Retrieve the top 5 starting station names to see where most rides start from April 2021 - March 2022

After exploring the number of rides from all riders, Streeter Dr. & Grand Ave, Michigan Ave & Oak St, Wells St & Concord Ln, Millennium Park, and Clark St & Elm St were the most frequent starting stations from April 2021 - March 2022.

```
top_5_start_station_names <- sort(table(all_trip_data_v2$start_station_name), decreasing=TRUE)[1:5]
knitr::kable(top_5_start_station_names,
             col.names = c("Starting Station Name", "Number of rides"),
             caption = "Top 5 Starting Stations (April 2021 - March 2022)")
```

Table 1: Top 5 Starting Stations (April 2021 - March 2022)

Starting Station Name	Number of rides
Streeter Dr & Grand Ave	83973
Michigan Ave & Oak St	44248
Wells St & Concord Ln	44023
Millennium Park	42242
Clark St & Elm St	41201

3. Retrieve the top 5 station names by user type

```
members_only <- all_trip_data_v2[!(all_trip_data_v2$member_casual == "casual"),]
casuals_only <- all_trip_data_v2[!(all_trip_data_v2$member_casual == "member"),]
top_5_members_start <- sort(table(members_only$start_station_name), decreasing=TRUE)[1:5]
top_5_members_end <- sort(table(members_only$end_station_name), decreasing=TRUE)[1:5]
top_5_casuals_start <- sort(table(casuals_only$start_station_name), decreasing=TRUE)[1:5]
top_5_casuals_end <- sort(table(casuals_only$end_station_name), decreasing=TRUE)[1:5]
```

After exploring the number of rides from Annual Members, Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, Wells St & Elm St, and Dearborn St & Erie St were the most frequent starting stations from April 2021 - March 2022.

```
knitr::kable(top_5_members_start,
             col.names = c("Starting Station Name", "Number of Rides"),
             caption = "Annual Members, Top 5 Starting Stations (April 2021 - March 2022)")
```

Table 2: Annual Members, Top 5 Starting Stations (April 2021 - March 2022)

Starting Station Name	Number of Rides
Kingsbury St & Kinzie St	25152
Clark St & Elm St	24888
Wells St & Concord Ln	24107
Wells St & Elm St	21269
Dearborn St & Erie St	19419

After exploring the number of rides from Annual Members, Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, Wells St & Elm St, and Dearborn St & Erie St were the most frequent ending stations from April 2021 - March 2022.

```
knitr::kable(top_5_members_end,
             col.names = c("Ending Station Name", "Number of Rides"),
             caption = "Annual Members, Top 5 Ending Stations (April 2021 - March 2022)")
```

Table 3: Annual Members, Top 5 Ending Stations (April 2021 - March 2022)

Ending Station Name	Number of Rides
Kingsbury St & Kinzie St	24176
Clark St & Elm St	23982
Wells St & Concord Ln	23686
Wells St & Elm St	21008
Dearborn St & Erie St	19041

After exploring the number of rides from Casual Riders, Streeter Dr & Grand Ave , Millennium Park, Michigan Ave & Oak St, Shedd Aquarium, and Theater on the Lake were the most frequent starting stations from April 2021 - March 2022.

```
knitr::kable(top_5_casuals_start,
             col.names = c("Starting Station Name", "Number of Rides"),
             caption = "Casual Users, Top 5 Starting Stations (April 2021 - March 2022)")
```

Table 4: Casual Users, Top 5 Starting Stations (April 2021 - March 2022)

Starting Station Name	Number of Rides
Streeter Dr & Grand Ave	67371
Millennium Park	33273
Michigan Ave & Oak St	29670
Shedd Aquarium	23195
Theater on the Lake	21106

After exploring the number of rides from Casual Riders, Streeter Dr & Grand Ave , Millennium Park, Michigan Ave & Oak St, Theater on the Lake, and Shedd Aquarium on the Lake were the most frequent ending stations from April 2021 - March 2022.

```
knitr::kable(top_5_casuals_end,
             col.names = c("Ending Station Name", "Number of Rides"),
             caption = "Casual Users, Top 5 Ending Stations (April 2021 - March 2022)")
```

Table 5: Casual Users, Top 5 Ending Stations (April 2021 - March 2022)

Ending Station Name	Number of Rides
Streeter Dr & Grand Ave	68484
Millennium Park	33382
Michigan Ave & Oak St	30154
Theater on the Lake	21954
Shedd Aquarium	21157

4. Analyze bike type by rider type

Casual users overwhelmingly preferred docked bikes compared to member users which do not seem to use them. Casual riders used classic bikes more than electric bikes by almost a double margin while annual members used classic bikes more than electric bikes by an alarming margin.

```
member_bike_type <- table(members_only$rideable_type)
casual_bike_type <- table(casuals_only$rideable_type)
knitr::kable(casual_bike_type,
             col.names = c("Bike Type", "Number of Rides"),
             caption = "Casual Users, Total Rides by Bike Type (April 2021 - March 2022)")
```

Table 6: Casual Users, Total Rides by Bike Type (April 2021 - March 2022)

Bike Type	Number of Rides
classic_bike	1257512
docked_bike	303185
electric_bike	652841

```
knitr::kable(member_bike_type,
             col.names = c("Bike Type", "Number of Rides"),
             caption = "Annual Members, Total Rides by Bike Type (April 2021 - March 2022)")
```

Table 7: Annual Members, Total Rides by Bike Type (April 2021 - March 2022)

Bike Type	Number of Rides
classic_bike	1992903
electric_bike	770286

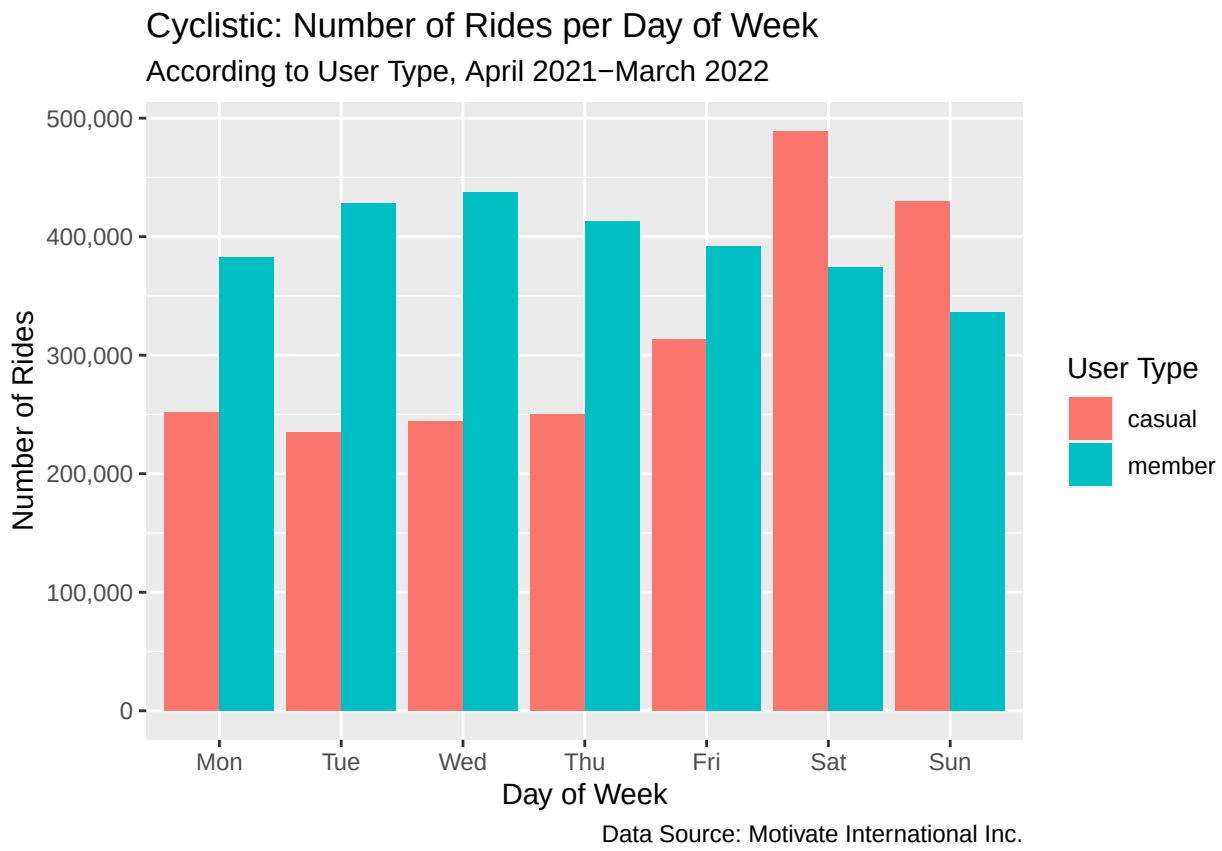
Visualizations

1. Visualization showing the number of rides per day of week by type of user, ordered day_of_week Mon-Sun.

Graph depicts member riders taking a similar average number of rides throughout the week while casual riders seem to prefer taking more rides during the weekend, Saturday and Sunday.

```
all_trip_data_v2$day_of_week <- ordered(all_trip_data_v2$day_of_week, levels=c("Mon", "Tue", "Wed", "Thu
all_trip_data_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n()) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(mapping = aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Cyclistic: Number of Rides per Day of Week", subtitle = "According to User Type, April 1
  scale_y_continuous(labels = scales::comma)
```

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

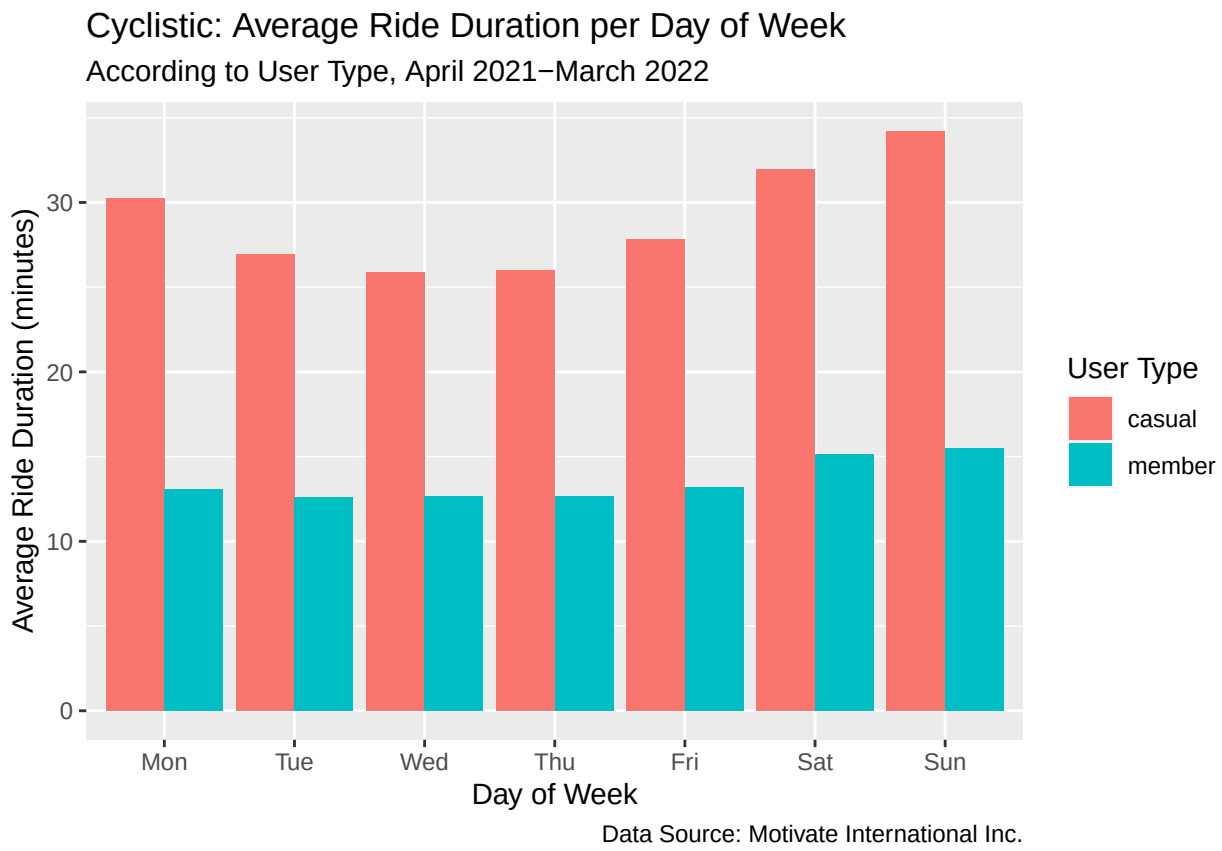


2. Visualization showing the average ride duration per day of week by type of user

Graph depicts casual users have significantly longer rides than annual members every day of the week.

```
all_trip_data_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Cyclistic: Average Ride Duration per Day of Week", subtitle = "According to User Type, A
```

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

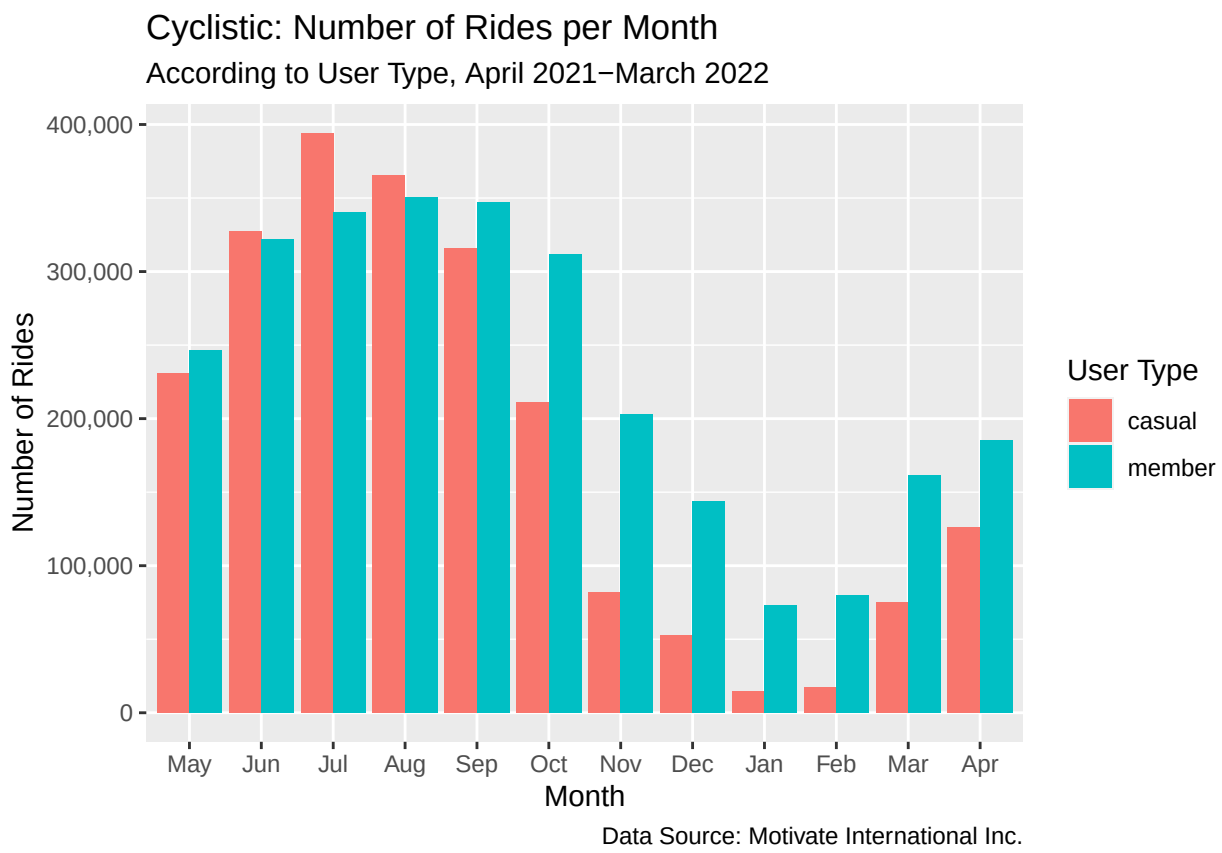


3. Visualization showing the number of rides per month by user type while reordering months starting with May through April

Graph depicts most casual users rent bikes in the months June through September drastically dropping off bike usage in the winter months. Annual members follow a similar patter to bike rentals as to the casual users peaking in the spring/summer months while falling off in the fall/winter seasons.

```
all_trip_data_v2$month <- ordered(all_trip_data_v2$month, levels=c("May", "Jun", "Jul", "Aug", "Sep", "
all_trip_data_v2 %>%
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n()) %>%
  arrange(member_casual, month)  %>%
  ggplot(mapping = aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Cyclistic: Number of Rides per Month", subtitle = "According to User Type, April 2021-Ma
```

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

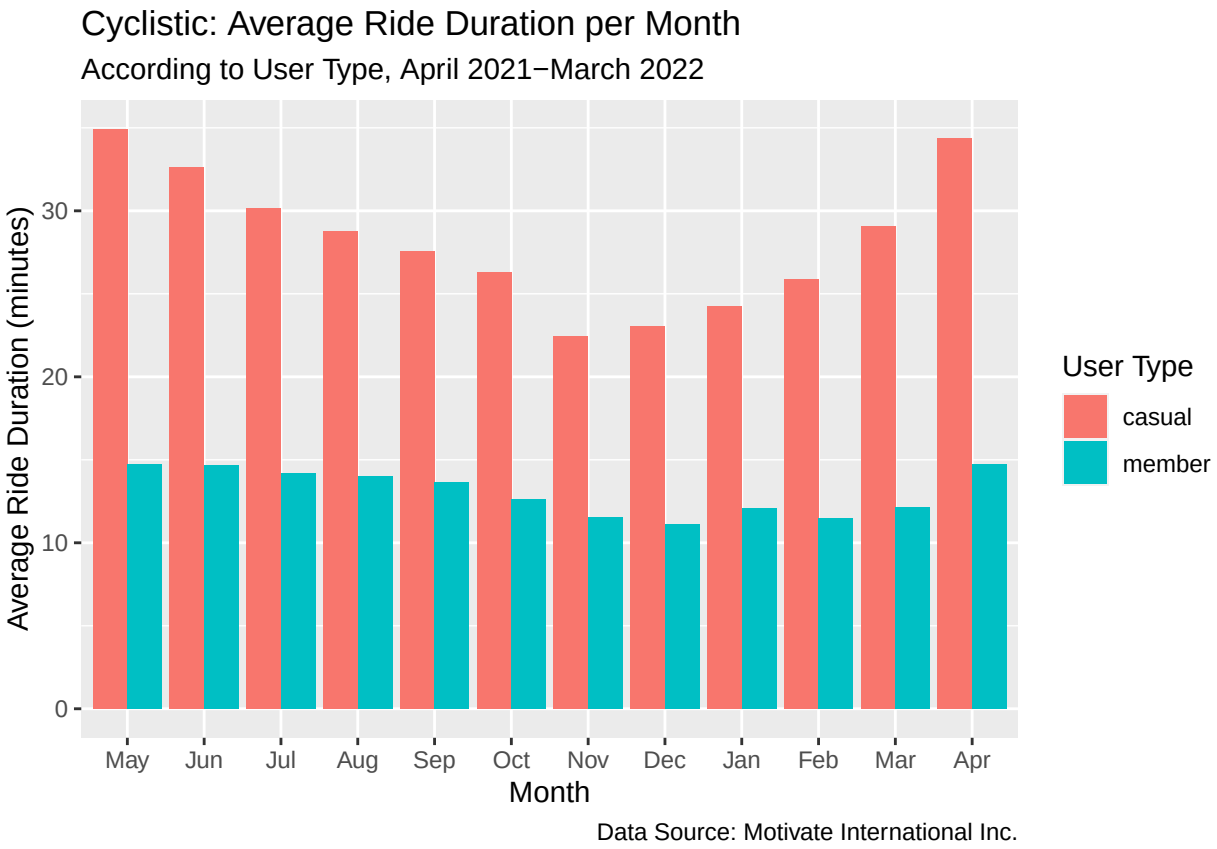


4. Visualization showing the average ride duration per month by user type

The graph below depicts casual users ride longer in the spring/summer seasons while annual members' ride lengths are fairly consistent all year-round.

```
all_trip_data_v2 %>%
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Cyclistic: Average Ride Duration per Month", subtitle = "According to User Type, April
```

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



Recap of Analysis and Visualizations

- I have determined that Streeter Dr. & Grand Ave, Michigan Ave & Oak St, Wells St & Concord Ln, Millennium Park, and Clark St & Elm St were the most frequent starting stations from April 2021 - March 2022 among all riders.
- Among annual members, annual members most frequently used the Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, Wells St & Elm St, and Dearborn St & Erie St as their start stations. Casual users most frequently used the Streeter Dr. & Grand Ave, Millennium Park, Michigan Ave & Oak St, Shedd Aquarium, and Theater on the Lake as their start stations.
- Among annual member, annual members most frequently stopped at the Kingsbury St & Kinzie St, Clark St & Elm St, Wells St & Concord Ln, Wells St & Elm St, and Dearborn St & Erie St end stations. Casual riders preffered the Streeter Dr & Grand Ave, Millennium Park, Michigan Ave & Oak St, Theater on the Lake, and Shedd Aquarium end stations.
- Both annual members and casual users overwhelmingly preferred classic bikes. Casual riders used classic bikes more than electric bikes by almost a double margin. Annual members used classic bikes more than electric bikes by an alarming margin while using no docked bikes.
- Member riders take a similar average number of rides throughout the week while casual riders seem to prefer taking more rides during the weekend: Saturday and Sunday. Casual riders have significantly longer rides than annual members every day of the week.
- Casual users rent more bikes in the months of June through September drastically dropping off bike usage in the winter months. Annual members follow a similar pattern to bike rentals as to the casual

users, peaking in the spring/summer months while falling off in the fall/winter seasons. Casual users ride longer in the summer/spring seasons while annual members' ride lengths are fairly consistent all year-round.

Application of Insights

Cyclistic can set up promotions or advertisements around the most frequently used starting and ending stations. These promotions could be ran during the spring summer season to maximize the amount of riders that would encounter the promotions as these months are the peak months for bike rentals. Cyclistic can adjust prices on bike types to encourage more/less usage of a certain type of bike. Dropping the price on electric bikes or running a weekend special price will encourage more electric bike rentals especially within the casual community since casuals go on more rides during the weekend as oppose to members. Cyclistic can offer summer pass membership or weekend pass memberships to convert casual riders to members. This type of pass can be used as a bridge to convert casuals to full year members. Offer a tier system of casual, summer/weekend pass and full year pass with each having its own perks based on tier of membership.

Top 3 Recommendations based on Analysis

1. Cyclistic can set up promotions or advertisements around the most frequently used starting and ending stations. These promotions could be run during the spring summer season to maximize the amount of riders that would encounter the promotions as these months are the peak months for bike rentals.
2. Cyclistic can adjust prices on bike types to encourage more/less usage of a certain type of bike. Dropping the price on electric bikes or running a weekend special price will encourage more electric bike rentals especially within the casual community since casuals go on more rides during the weekend as oppose to members. Cyclistic could also offer docked bikes or electric bikes rental rates at a discount to those holding a membership.
3. Since casuals take longer rides and seem to prefer riding on the weekends than Cyclist could introduce a weekend pass allowing unlimited 45 minutes rides on Friday, Saturday and Sunday only that is priced just bellow the annual price.