Cyclistic Project Log

Project Overview

Company: Cyclistic

- They are based in Chicago, since 2016
- Their bike share program features 692 docking stations and 5824 bikes with Geo-tracked feature
  - Usage of traditional bikes is 92%, with the assistive options: 8%
    - Reclining bikes, Hand tricycles, Two-wheeled bike
- The service usage: 70% to commute to work each day, 30% for leisure

Current Marketing Strategy

- Building general awareness to broad consumer segment by appealing flexibility of pricing plan
  - Casual riders: Single ride passes, full-day passes
  - Cyclistic members: annual membership

Stakeholders

- Executive team: They are detail oriented and will approve recommended plan
- Director of Marketing: Lily Moreno
  - She is responsible for development of campaigns and initiatives to promote the bike-share program that may include email, social media, and other channels
  - She believes that maximizing the number of annual members will be key to future growth because casual riders are already aware of the cyclistic program and have chosen Cyclistic for their mobility needs
- Analytics Team: Responsible for collecting, analyzing, and reporting data that helps guide the marketing strategy
- Finance Analysts: Concluded that annual members are much more profitable than casual riders

Goal of This Project

- Design the marketing strategies aimed to convert casual riders into annual members
  - Key guiding questions:
    1. How do annual members and casual riders use Cyclistic bikes differently?
    2. Why would casual riders buy Cyclistic annual memberships?
    3. How can Cyclistic use digital media to influence casual riders to become members?
- This Analysis is the a part of the 3 analysis series which will help answer the first key guiding questions

Analysis Process

1. A Clear Statement of the Business Task

- Identify key difference of trend in service usage between the casual riders and the annual members in order to design the marketing campaign that will encourage casual riders to get the annual membership

2. A Description of All Data Sources Used

Download data and store it appropriately

- The data was located on the company's cloud storage (Amazon Web Services)
  - https://divvy-tripdata.s3.amazonaws.com/index.html
  - The data was downloaded and stored in the locally safe location ### Identify how it's organized

```
##########################################
# Install required packages
# tidyverse for data import and wrangling
# lubridate for date functions
##########################################

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

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

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

```
library(ggplot2)  # ggplot2 helps visualize data

# getwd() displays your working directory
# setwd()  sets working directory to simplify calls to data
setwd("C:/Workstations/Cyclistics/tripdata csv")
```

```r
#=====================================
# STEP 1: COLLECT DATA
#=====================================
#Read 12 month datasets (csv file) here
tripdata_202008 <-read_csv("202008-divvy-tripdata.csv")
```

```
## Rows: 622361 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```r
tripdata_202009 <-read_csv("202009-divvy-tripdata.csv")
```

```
## Rows: 532958 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```r
tripdata_202010 <-read_csv("202010-divvy-tripdata.csv")
```

```
## Rows: 388653 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```r
tripdata_202011 <-read_csv("202011-divvy-tripdata.csv")
```

```
## Rows: 259716 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```
tripdata_202012 <-read_csv("202012-divvy-tripdata.csv")
```

```
## Rows: 131573 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.
```

```
tripdata_202101 <-read_csv("202101-divvy-tripdata.csv")
```

```
## Rows: 96834 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.
```

```
tripdata_202102 <-read_csv("202102-divvy-tripdata.csv")
```

```
## Rows: 49622 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.
```

```
tripdata_202103 <-read_csv("202103-divvy-tripdata.csv")
```

```
## Rows: 228496 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.
```

```
tripdata_202104 <-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.
```

```
tripdata_202105 <-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.
```

```
tripdata_202106 <-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.
```

```
tripdata_202107 <-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.
```

- The dataset consists of 12 '.csv' files representing each month from August 2020 to July 2021
- Each of those trip data files consist of 13 columns with various data types indicating the aspects of each ride activity
  - Names of columns are:
    1. ride_id: unique id representing each trip record

2. rideable_type: indicates which bike type used for the trip
3. started_at: indicates trip start time
4. ended_at: indicates trip end time
5. start_station_name: indicates trip start station name
6. start_station_id: indicates trip start station id
7. end_station_name: indicates trip end station name
8. end_station_id: indicates trip end station id
9. start_lat: indicates trip start latitude
10. start_lng: indicates trip start longitude
11. end_lat: indicates trip end latitude
12. end_lng: indicates trip end longitude

13. member_casual: indicates the trip was made by an annual member or a casual user

- All columns of 12 trip dataset have the matching column name
  - No need to rename columns

```
# Inspect the dataframes and look for incongruencies
str(tripdata_202008)
```

```
## spec_tbl_df [622,361 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:622361] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79F
##  $ rideable_type     : chr [1:622361] "docked_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:622361], format: "2020-08-20 18:08:14" "2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:622361], format: "2020-08-20 18:17:51" "2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:622361] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Colu
##  $ start_station_id  : num [1:622361] 329 168 195 81 658 658 196 67 153 177 ...
##  $ end_station_name  : chr [1:622361] "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St & 
##  $ end_station_id    : num [1:622361] 141 168 44 47 658 658 49 229 225 305 ...
##  $ start_lat         : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:622361] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(tripdata_202009)
```

```
## spec_tbl_df [532,958 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6
##  $ rideable_type     : chr [1:532958] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...
##  $ ended_at          : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...
##  $ start_station_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ start_station_id  : num [1:532958] 52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : chr [1:532958] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ end_station_id    : num [1:532958] 112 NA NA 249 24 NA 256 NA NA NA ...
##  $ start_lat         : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:532958] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:532958] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:532958] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(tripdata_202010)
```

```
## spec_tbl_df [388,653 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4
##  $ rideable_type     : chr [1:388653] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:388653], format: "2020-10-31 19:39:43" "2020-10-31 23:50:08" ...
##  $ ended_at          : POSIXct[1:388653], format: "2020-10-31 19:57:12" "2020-11-01 00:04:16" ...
##  $ start_station_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave"
##  $ start_station_id  : num [1:388653] 313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : chr [1:388653] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Universit
##  $ end_station_id    : num [1:388653] 125 260 423 256 185 53 125 313 199 635 ...
##  $ start_lat         : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:388653] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(tripdata_202011)
```

```
## spec_tbl_df [259,716 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533
##  $ rideable_type     : chr [1:259716] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:259716], format: "2020-11-01 13:36:00" "2020-11-01 10:03:26" ...
##  $ ended_at          : POSIXct[1:259716], format: "2020-11-01 13:45:40" "2020-11-01 10:14:45" ...
##  $ start_station_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shor
##  $ start_station_id  : num [1:259716] 110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : chr [1:259716] "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal S
##  $ end_station_id    : num [1:259716] 211 29 41 185 2 76 72 NA 288 273 ...
##  $ start_lat         : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:259716] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:259716] -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:259716] "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_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(tripdata_202012)
```

```
## spec_tbl_df [131,573 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
```

```
##  $ ride_id           : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE11
##  $ rideable_type     : chr [1:131573] "classic_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:131573], format: "2020-12-27 12:44:29" "2020-12-18 17:37:15" ...
##  $ ended_at          : POSIXct[1:131573], format: "2020-12-27 12:55:06" "2020-12-18 17:44:19" ...
##  $ start_station_name: chr [1:131573] "Aberdeen St & Jackson Blvd" NA NA NA ...
##  $ start_station_id  : chr [1:131573] "13157" NA NA NA ...
##  $ end_station_name  : chr [1:131573] "Desplaines St & Kinzie St" NA NA NA ...
##  $ end_station_id    : chr [1:131573] "TA1306000003" NA NA NA ...
##  $ start_lat         : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:131573] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:131573] -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:131573] "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(tripdata_202101)
```

```
## spec_tbl_df [96,834 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453
##  $ rideable_type     : chr [1:96834] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:96834], format: "2021-01-23 16:14:19" "2021-01-27 18:43:08" ...
##  $ ended_at          : POSIXct[1:96834], format: "2021-01-23 16:24:44" "2021-01-27 18:47:12" ...
##  $ start_station_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez St" "Cali
##  $ start_station_id  : chr [1:96834] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:96834] NA NA NA NA ...
##  $ end_station_id    : chr [1:96834] NA NA NA NA ...
##  $ start_lat         : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:96834] "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(tripdata_202102)
```

```
## spec_tbl_df [49,622 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3
##  $ rideable_type     : chr [1:49622] "classic_bike" "classic_bike" "electric_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:49622], format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct[1:49622], format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St
##  $ start_station_id  : chr [1:49622] "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State S
##  $ end_station_id    : chr [1:49622] "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num [1:49622] 42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:49622] 42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:49622] "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(tripdata_202103)
```

```
## spec_tbl_df [228,496 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D
##  $ rideable_type     : chr [1:228496] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:228496], format: "2021-03-16 08:32:30" "2021-03-28 01:26:28" ...
##  $ ended_at          : POSIXct[1:228496], format: "2021-03-16 08:36:34" "2021-03-28 01:36:55" ...
##  $ start_station_name: chr [1:228496] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave"
```

```
##  $ start_station_id  : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave
##  $ end_station_id    : chr [1:228496] "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num [1:228496] 41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:228496] 41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num [1:228496] -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:228496] "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(tripdata_202104)
```

```
## 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(tripdata_202105)
```

```
## spec_tbl_df [531,633 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881A
##  $ 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(tripdata_202106)
```

```
## 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(tripdata_202107)
```

```
## 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>
```

Determine the credibility of the data (ROCCC)

- The data follows ROCCC approach
- Bias: There seem to be no noticeable bias issues in this data
- Credibility: The data has been made available by Motivate International Inc. under the license by Divvy
  - https://www.divvybikes.com/data-license-agreement
- However, the dataset has a limitation since there are non personally identifiable information
  - Therefore, we won't be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes
  - Also, it will be hard to verify or track how many times a single user uses the service or if such one is an annual member or a casual user
  - There are some data with null value or inconsistency in the data format

Sort and filter the data

- start_station_id and end_station_id for tripdata_202008, tripdata_202009, tripdata_202010, tripdata_202011 are double() but others are character()

```
# Convert start_station_id and end_station_id to character
# so that datasets can be stacked correctly

tripdata_202008 <- mutate(tripdata_202008, start_station_id = as.character(start_station_id),
                          end_station_id = as.character(end_station_id))
tripdata_202009 <- mutate(tripdata_202009, start_station_id = as.character(start_station_id),
                          end_station_id = as.character(end_station_id))
tripdata_202010 <- mutate(tripdata_202010, start_station_id = as.character(start_station_id),
                          end_station_id = as.character(end_station_id))
tripdata_202011 <- mutate(tripdata_202011, start_station_id = as.character(start_station_id),
                          end_station_id = as.character(end_station_id))

# Confirm the data type for start_station_id, end_station_id changed to char
str(tripdata_202008)
```

```
## tibble [622,361 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:622361] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79F
##  $ rideable_type     : chr [1:622361] "docked_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:622361], format: "2020-08-20 18:08:14" "2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:622361], format: "2020-08-20 18:17:51" "2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:622361] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Colu
##  $ start_station_id  : chr [1:622361] "329" "168" "195" "81" ...
##  $ end_station_name  : chr [1:622361] "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St & 
##  $ end_station_id    : chr [1:622361] "141" "168" "44" "47" ...
##  $ start_lat         : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:622361] "member" "casual" "casual" "casual" ...
```

```
str(tripdata_202009)
```

```
## tibble [532,958 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6
##  $ rideable_type     : chr [1:532958] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...
##  $ ended_at          : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...
##  $ start_station_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ start_station_id  : chr [1:532958] "52" NA NA "246" ...
##  $ end_station_name  : chr [1:532958] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakda
##  $ end_station_id    : chr [1:532958] "112" NA NA "249" ...
##  $ start_lat         : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:532958] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:532958] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:532958] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:532958] "casual" "casual" "casual" "casual" ...
```

```
str(tripdata_202010)
```

```
## tibble [388,653 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4
##  $ rideable_type     : chr [1:388653] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:388653], format: "2020-10-31 19:39:43" "2020-10-31 23:50:08" ...
##  $ ended_at          : POSIXct[1:388653], format: "2020-10-31 19:57:12" "2020-11-01 00:04:16" ...
##  $ start_station_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave"
##  $ start_station_id  : chr [1:388653] "313" "227" "102" "165" ...
##  $ end_station_name  : chr [1:388653] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Universit
##  $ end_station_id    : chr [1:388653] "125" "260" "423" "256" ...
##  $ start_lat         : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:388653] "casual" "casual" "casual" "casual" ...
```

```
str(tripdata_202011)
```

```
## tibble [259,716 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533
##  $ rideable_type     : chr [1:259716] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:259716], format: "2020-11-01 13:36:00" "2020-11-01 10:03:26" ...
##  $ ended_at          : POSIXct[1:259716], format: "2020-11-01 13:45:40" "2020-11-01 10:14:45" ...
##  $ start_station_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shor
##  $ start_station_id  : chr [1:259716] "110" "672" "76" "659" ...
##  $ end_station_name  : chr [1:259716] "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal St
##  $ end_station_id    : chr [1:259716] "211" "29" "41" "185" ...
##  $ start_lat         : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:259716] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:259716] -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:259716] "casual" "casual" "casual" "casual" ...
```

```
# Stack individual data frames into one big data frame
all_trips <- bind_rows(tripdata_202008, tripdata_202009, tripdata_202010,
                       tripdata_202011, tripdata_202012, tripdata_202101,
                       tripdata_202102, tripdata_202103, tripdata_202104,
                       tripdata_202105, tripdata_202106, tripdata_202107)

# Inspect the new table that has been created

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

```
## [1] 4731081      13
```

```
# See the first 6 rows of data frame.
head(all_trips)
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 322BD2~ docked_bike   2020-08-20 18:08:14 2020-08-20 18:17:51 Lake Shore Dr &~
## 2 2A3AEF~ electric_bike 2020-08-27 18:46:04 2020-08-27 19:54:51 Michigan Ave & ~
## 3 67DC1D~ electric_bike 2020-08-26 19:44:14 2020-08-26 21:53:07 Columbus Dr & R~
## 4 C79FBB~ electric_bike 2020-08-27 12:05:41 2020-08-27 12:53:45 Daley Center Pl~
## 5 13814D~ electric_bike 2020-08-27 16:49:02 2020-08-27 16:59:49 Leavitt St & Di~
## 6 56349A~ electric_bike 2020-08-27 17:26:23 2020-08-27 18:07:50 Leavitt St & Di~
## # ... 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>
```

```
# See the last 6 rows of the data
tail(all_trips)
```

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 7B47CA~ electric_bike 2021-07-04 05:34:53 2021-07-04 05:36:46 <NA>
## 2 1E660B~ electric_bike 2021-07-04 10:40:41 2021-07-04 11:30:13 <NA>
## 3 A2448B~ electric_bike 2021-07-04 12:47:41 2021-07-04 12:54:46 <NA>
## 4 2D612B~ electric_bike 2021-07-03 21:41:58 2021-07-03 21:57:14 <NA>
## 5 6D615D~ electric_bike 2021-07-03 22:10:31 2021-07-03 22:11:39 <NA>
## 6 0F31D3~ electric_bike 2021-07-04 07:03:50 2021-07-04 07:32:38 <NA>
## # ... 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>
```

```
# See list of columns and data types
str(all_trips)
```

```
## tibble [4,731,081 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4731081] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79
##  $ rideable_type     : chr [1:4731081] "docked_bike" "electric_bike" "electric_bike" "electric_bike"
```

```
##  $ started_at        : POSIXct[1:4731081], format: "2020-08-20 18:08:14" "2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:4731081], format: "2020-08-20 18:17:51" "2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:4731081] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Colu
##  $ start_station_id  : chr [1:4731081] "329" "168" "195" "81" ...
##  $ end_station_name  : chr [1:4731081] "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St &
##  $ end_station_id    : chr [1:4731081] "141" "168" "44" "47" ...
##  $ start_lat         : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:4731081] "member" "casual" "casual" "casual" ...
```

```r
# Statistical summary of data. Mainly for numerics
summary(all_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:4731081     Length:4731081     Min.   :2020-08-01 00:00:01
##  Class :character   Class :character   1st Qu.:2020-10-03 08:51:57
##  Mode  :character   Mode  :character   Median :2021-04-05 13:41:29
##                                        Mean   :2021-02-17 10:22:09
##                                        3rd Qu.:2021-06-15 05:47:53
##                                        Max.   :2021-07-31 23:59:58
##
##     ended_at                   start_station_name start_station_id
##  Min.   :2020-08-01 00:04:41   Length:4731081     Length:4731081
##  1st Qu.:2020-10-03 09:13:58   Class :character   Class :character
##  Median :2021-04-05 14:03:51   Mode  :character   Mode  :character
##  Mean   :2021-02-17 10:44:21
##  3rd Qu.:2021-06-15 06:16:14
##  Max.   :2021-08-12 17:45:41
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:4731081     Length:4731081     Min.   :41.64   Min.   :-87.87
##  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   :-87.64
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.08   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   :41.51   Min.   :-88.07   Length:4731081
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.64
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.16   Max.   :-87.44
##  NA's   :5247    NA's   :5247
```

3. Documentation of Any Cleaning or Manipulation of Data

1) Add some additional columns of data, such as day, month year from started_at

```
# all_trips$date <- as.Date(all_trips$started_at)
all_trips$date <- as.Date(all_trips$started_at) #The default format is yyyy-mm-dd
all_trips$hour <- format(all_trips$started_at, "%H")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year_month <- format(as.Date(all_trips$date), "%Y_%m")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")

# Verify newly added columns
str(all_trips)
```

```
## tibble [4,731,081 x 18] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4731081] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79
##  $ rideable_type     : chr [1:4731081] "docked_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:4731081], format: "2020-08-20 18:08:14" "2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:4731081], format: "2020-08-20 18:17:51" "2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:4731081] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Colu
##  $ start_station_id  : chr [1:4731081] "329" "168" "195" "81" ...
##  $ end_station_name  : chr [1:4731081] "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St &
##  $ end_station_id    : chr [1:4731081] "141" "168" "44" "47" ...
##  $ start_lat         : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:4731081] "member" "casual" "casual" "casual" ...
##  $ date              : Date[1:4731081], format: "2020-08-20" "2020-08-27" ...
##  $ hour              : chr [1:4731081] "18" "18" "19" "12" ...
##  $ day               : chr [1:4731081] "20" "27" "26" "27" ...
##  $ year_month        : chr [1:4731081] "2020_08" "2020_08" "2020_08" "2020_08" ...
##  $ day_of_week       : chr [1:4731081] "Thursday" "Thursday" "Wednesday" "Thursday" ...
```

2) calculate ride_length from started_at and ended_at in min

```
all_trips$ride_length_min <- round(as.numeric(difftime(all_trips$ended_at,
                                                       all_trips$started_at,
                                                       units = "mins")), digits = 2)

# Verify newly added columns
str(all_trips)
```

```
## tibble [4,731,081 x 19] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4731081] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79
##  $ rideable_type     : chr [1:4731081] "docked_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : POSIXct[1:4731081], format: "2020-08-20 18:08:14" "2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:4731081], format: "2020-08-20 18:17:51" "2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:4731081] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Colu
##  $ start_station_id  : chr [1:4731081] "329" "168" "195" "81" ...
```

```
##  $ end_station_name  : chr [1:4731081] "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St &
##  $ end_station_id    : chr [1:4731081] "141" "168" "44" "47" ...
##  $ start_lat         : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:4731081] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:4731081] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:4731081] "member" "casual" "casual" "casual" ...
##  $ date              : Date[1:4731081], format: "2020-08-20" "2020-08-27" ...
##  $ hour              : chr [1:4731081] "18" "18" "19" "12" ...
##  $ day               : chr [1:4731081] "20" "27" "26" "27" ...
##  $ year_month        : chr [1:4731081] "2020_08" "2020_08" "2020_08" "2020_08" ...
##  $ day_of_week       : chr [1:4731081] "Thursday" "Thursday" "Wednesday" "Thursday" ...
##  $ ride_length_min   : num [1:4731081] 9.62 68.78 128.88 48.07 10.78 ...
```

3) Remove all duplicate ride_id

```
glimpse(all_trips)
```

```
## Rows: 4,731,081
## Columns: 19
## $ ride_id            <chr> "322BD23D287743ED", "2A3AEF1AB9054D8B", "67DC1D133E~
## $ rideable_type      <chr> "docked_bike", "electric_bike", "electric_bike", "e~
## $ started_at         <dttm> 2020-08-20 18:08:14, 2020-08-27 18:46:04, 2020-08-~
## $ ended_at           <dttm> 2020-08-20 18:17:51, 2020-08-27 19:54:51, 2020-08-~
## $ start_station_name <chr> "Lake Shore Dr & Diversey Pkwy", "Michigan Ave & 14~
## $ start_station_id   <chr> "329", "168", "195", "81", "658", "658", "196", "67~
## $ end_station_name   <chr> "Clark St & Lincoln Ave", "Michigan Ave & 14th St",~
## $ end_station_id     <chr> "141", "168", "44", "47", "658", "658", "49", "229"~
## $ start_lat          <dbl> 41.93259, 41.86438, 41.88464, 41.88409, 41.90299, 4~
## $ start_lng          <dbl> -87.63643, -87.62368, -87.61955, -87.62964, -87.683~
## $ end_lat            <dbl> 41.91569, 41.86422, 41.88497, 41.88958, 41.90300, 4~
## $ end_lng            <dbl> -87.63460, -87.62344, -87.62757, -87.62754, -87.683~
## $ member_casual      <chr> "member", "casual", "casual", "casual", "casual", "~
## $ date               <date> 2020-08-20, 2020-08-27, 2020-08-26, 2020-08-27, 20~
## $ hour               <chr> "18", "18", "19", "12", "16", "17", "20", "21", "19~
## $ day                <chr> "20", "27", "26", "27", "27", "27", "26", "26", "26~
## $ year_month         <chr> "2020_08", "2020_08", "2020_08", "2020_08", "2020_0~
## $ day_of_week        <chr> "Thursday", "Thursday", "Wednesday", "Thursday", "T~
## $ ride_length_min    <dbl> 9.62, 68.78, 128.88, 48.07, 10.78, 41.45, 19.97, 12~
```

```
all_trips <- distinct(all_trips, ride_id, .keep_all = TRUE)
glimpse(all_trips)
```

```
## Rows: 4,730,872
## Columns: 19
## $ ride_id            <chr> "322BD23D287743ED", "2A3AEF1AB9054D8B", "67DC1D133E~
## $ rideable_type      <chr> "docked_bike", "electric_bike", "electric_bike", "e~
## $ started_at         <dttm> 2020-08-20 18:08:14, 2020-08-27 18:46:04, 2020-08-~
## $ ended_at           <dttm> 2020-08-20 18:17:51, 2020-08-27 19:54:51, 2020-08-~
## $ start_station_name <chr> "Lake Shore Dr & Diversey Pkwy", "Michigan Ave & 14~
## $ start_station_id   <chr> "329", "168", "195", "81", "658", "658", "196", "67~
```

```
## $ end_station_name   <chr> "Clark St & Lincoln Ave", "Michigan Ave & 14th St",~
## $ end_station_id     <chr> "141", "168", "44", "47", "658", "658", "49", "229"~
## $ start_lat          <dbl> 41.93259, 41.86438, 41.88464, 41.88409, 41.90299, 4~
## $ start_lng          <dbl> -87.63643, -87.62368, -87.61955, -87.62964, -87.683~
## $ end_lat            <dbl> 41.91569, 41.86422, 41.88497, 41.88958, 41.90300, 4~
## $ end_lng            <dbl> -87.63460, -87.62344, -87.62757, -87.62754, -87.683~
## $ member_casual      <chr> "member", "casual", "casual", "casual", "casual", "~
## $ date               <date> 2020-08-20, 2020-08-27, 2020-08-26, 2020-08-27, 20~
## $ hour               <chr> "18", "18", "19", "12", "16", "17", "20", "21", "19~
## $ day                <chr> "20", "27", "26", "27", "27", "27", "26", "26", "26~
## $ year_month         <chr> "2020_08", "2020_08", "2020_08", "2020_08", "2020_0~
## $ day_of_week        <chr> "Thursday", "Thursday", "Wednesday", "Thursday", "T~
## $ ride_length_min    <dbl> 9.62, 68.78, 128.88, 48.07, 10.78, 41.45, 19.97, 12~
```

- Removed 209. Remaining rows: 4,730,872

4) Remove unnecessary columns

```
#  - end_station_id, end_station_name, end_lat, end_lng, start_station_id, start_station_name, start_la
all_trips <- all_trips %>%
  select(-c(end_station_id, end_station_name, end_lat, end_lng,
            start_station_id, start_station_name, start_lat, start_lng,
            ride_id,started_at, ended_at))

# Verify data set after the removal
str(all_trips)
```

```
## tibble [4,730,872 x 8] (S3: tbl_df/tbl/data.frame)
##  $ rideable_type  : chr [1:4730872] "docked_bike" "electric_bike" "electric_bike" "electric_bike" ..
##  $ member_casual  : chr [1:4730872] "member" "casual" "casual" "casual" ...
##  $ date           : Date[1:4730872], format: "2020-08-20" "2020-08-27" ...
##  $ hour           : chr [1:4730872] "18" "18" "19" "12" ...
##  $ day            : chr [1:4730872] "20" "27" "26" "27" ...
##  $ year_month     : chr [1:4730872] "2020_08" "2020_08" "2020_08" "2020_08" ...
##  $ day_of_week    : chr [1:4730872] "Thursday" "Thursday" "Wednesday" "Thursday" ...
##  $ ride_length_min: num [1:4730872] 9.62 68.78 128.88 48.07 10.78 ...
```

5) Make sure to remove any NA rows

```
all_trips_clean <- drop_na(all_trips)

nrow(all_trips)
```

```
## [1] 4730872
```

- There were no NA rows left to be removed

6) Remove negative ride length or over 1440min (24 hours)

```
all_trips_v2 <- all_trips_clean[!(all_trips_clean$ride_length_min<0 |
                                  all_trips_clean$ride_length_min>1440),]
nrow(all_trips_v2)
```

```
## [1] 4719477
```

- Removed 11,395. Dataset left with 4,719,477 rows

4. Conduct Analysis and Supporting Visualizations

```
# First, perform general summary of the ride_length_min

#mean(all_trips_v2$ride_length_min)     #straight avg (total ride length / rides)
#median(all_trips_v2$ride_length_min)   #midpoint number in the ascending array of ride lengths
#max(all_trips_v2$ride_length_min)      #longest ride - about 24 hours
#min(all_trips_v2$ride_length_min)      #shortest ride  - 0
summary(all_trips_v2$ride_length_min)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.35   13.18   21.55   24.07 1439.90
```

Service Usage - Overall Proportion

```
# Calculate usage proportion for the casual riders' trip duration in %
agg_sum_total <- sum(all_trips_v2$ride_length_min)
agg_sum <- aggregate(all_trips_v2$ride_length_min~all_trips_v2$member_casual,
          FUN = sum)
agg_sum_casual <- round(agg_sum[1,2] / agg_sum_total, 4) * 100
agg_sum_casual
```

```
## [1] 62.83
```

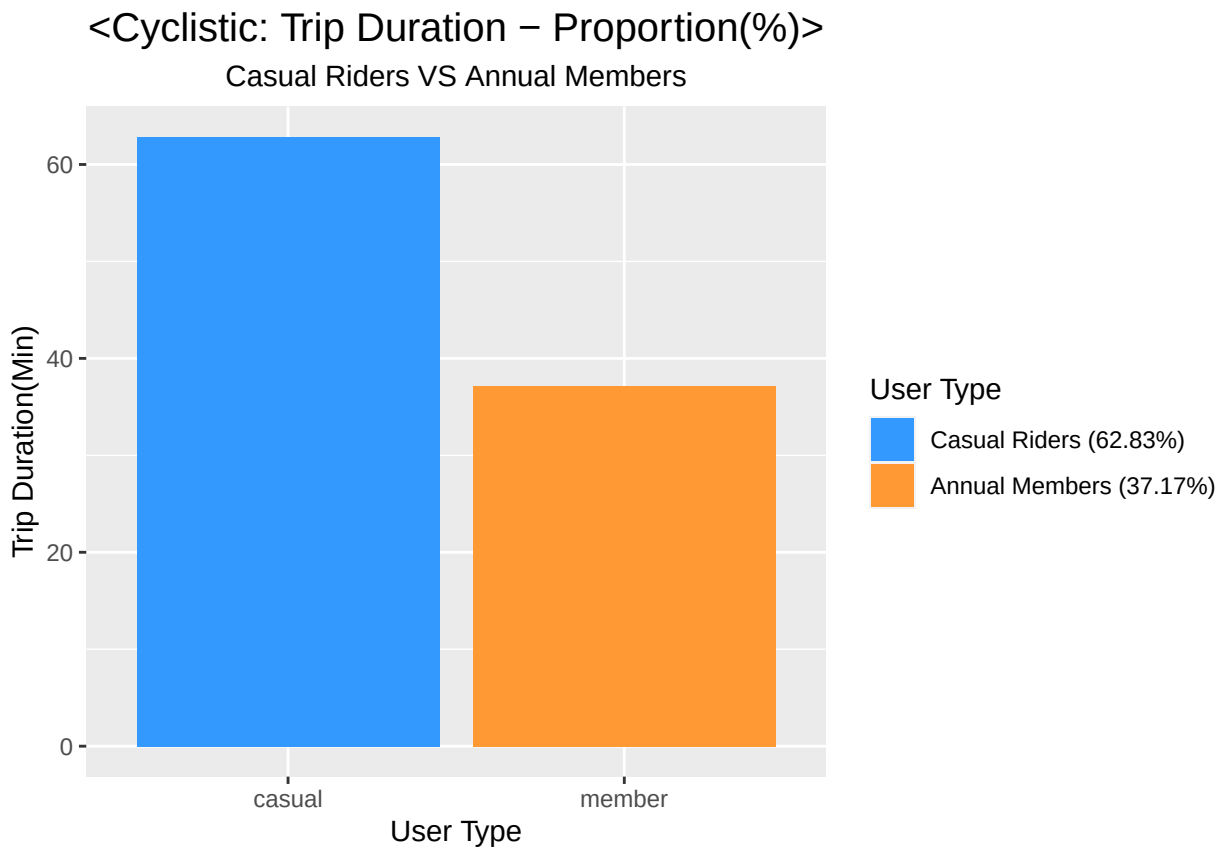
- Casual riders make up 62.83% of the sum of total trip duration

```
# Calculate usage proportion for the Annual members' trip duration in %
agg_sum_member <- round(agg_sum[2,2] / agg_sum_total, 4) * 100
agg_sum_member
```

```
## [1] 37.17
```

- Annual members make up 32.17% of the sum of total trip duration

```r
# setup plot
plot1_df <- data.frame(x_tmp=c("casual", "member"),
                       y_tmp=c(agg_sum_casual, agg_sum_member))

plot1 <- ggplot(plot1_df, aes(x= x_tmp, y= y_tmp, fill=x_tmp)) +
  geom_col(position = "dodge")

# format title and subtitle
plot1 <-  plot1 + labs(title = "<Cyclistic: Trip Duration - Proportion(%)>",
       subtitle = "Casual Riders VS Annual Members",
       x = "User Type", y = "Trip Duration(Min)")
plot1 <- plot1 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5))


# format legend
plot1 <- plot1 + scale_fill_manual(name="User Type",
                                   breaks = c("casual", "member"),
                                   labels= c("Casual Riders (62.83%)", "Annual Members (37.17%)" ),
                                   values=c("#3399FF", "#FF9933" ))

print(plot1)
```



- **Analysis: Trip duration for casual riders is *almost twice more* than annual members.**

```
agg_min <- aggregate(all_trips_v2$ride_length_min~all_trips_v2$member_casual, FUN = min)
agg_min
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length_min
## 1                     casual                            0
## 2                     member                            0
```

- agg_min does not hold any meaningful data

```
agg_max <- aggregate(all_trips_v2$ride_length_min~all_trips_v2$member_casual, FUN = max)
agg_max
```

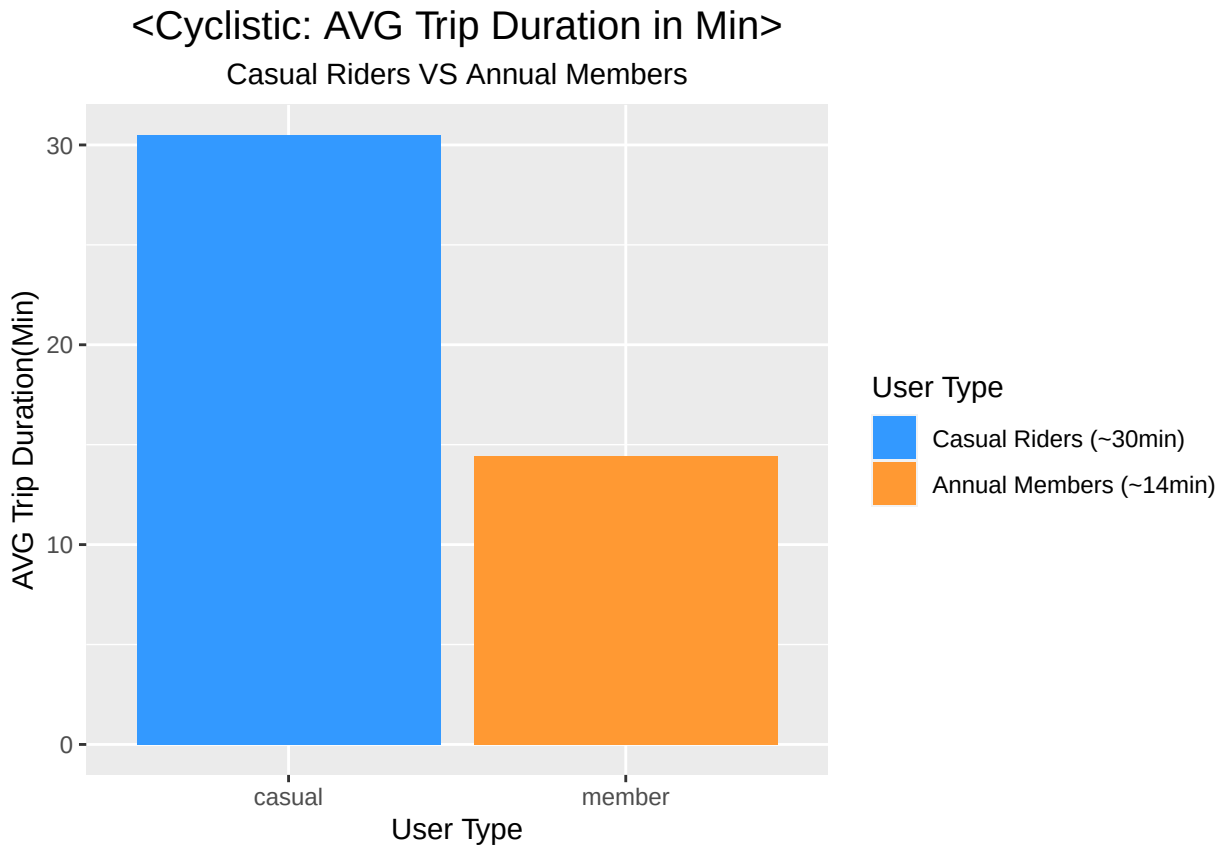
```
##   all_trips_v2$member_casual all_trips_v2$ride_length_min
## 1                     casual                      1439.90
## 2                     member                      1439.72
```

- agg_max also does not hold any meaningful data

```
agg_avg <- aggregate(all_trips_v2$ride_length_min~all_trips_v2$member_casual, FUN = mean)
agg_avg
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length_min
## 1                     casual                     30.48165
## 2                     member                     14.41568
```

```
plot2 <- all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(average_duration = mean(ride_length_min)) %>%
  arrange(member_casual) %>%
  ggplot(aes(x = member_casual, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")

# format title and subtitle
plot2 <-  plot2 + labs(title = "<Cyclistic: AVG Trip Duration in Min>",
       subtitle = "Casual Riders VS Annual Members",
       x = "User Type", y = "AVG Trip Duration(Min)")
plot2 <- plot2 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5))


# format legend
plot2 <- plot2 + scale_fill_manual(name="User Type",
                                   breaks = c("casual","member"),
                                   labels= c("Casual Riders (~30min)","Annual Members (~14min)"),
                                   values=c("#3399FF","#FF9933"))

print(plot2)
```



- **Analysis: Average trip duration for casual riders is also about *twice more* than annual members**

```
agg_count <- aggregate(all_trips_v2$ride_length_min~all_trips_v2$member_casual, FUN = length)
agg_count
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length_min
## 1                     casual                       2096608
## 2                     member                       2622869
```

```
# Calculate % of the trip count for casual riders
agg_trip_count_casual <- agg_count[1,2] / (agg_count[1,2] + agg_count[2,2])
agg_trip_count_casual <- round(agg_trip_count_casual * 100, 2)
agg_trip_count_casual
```

```
## [1] 44.42
```

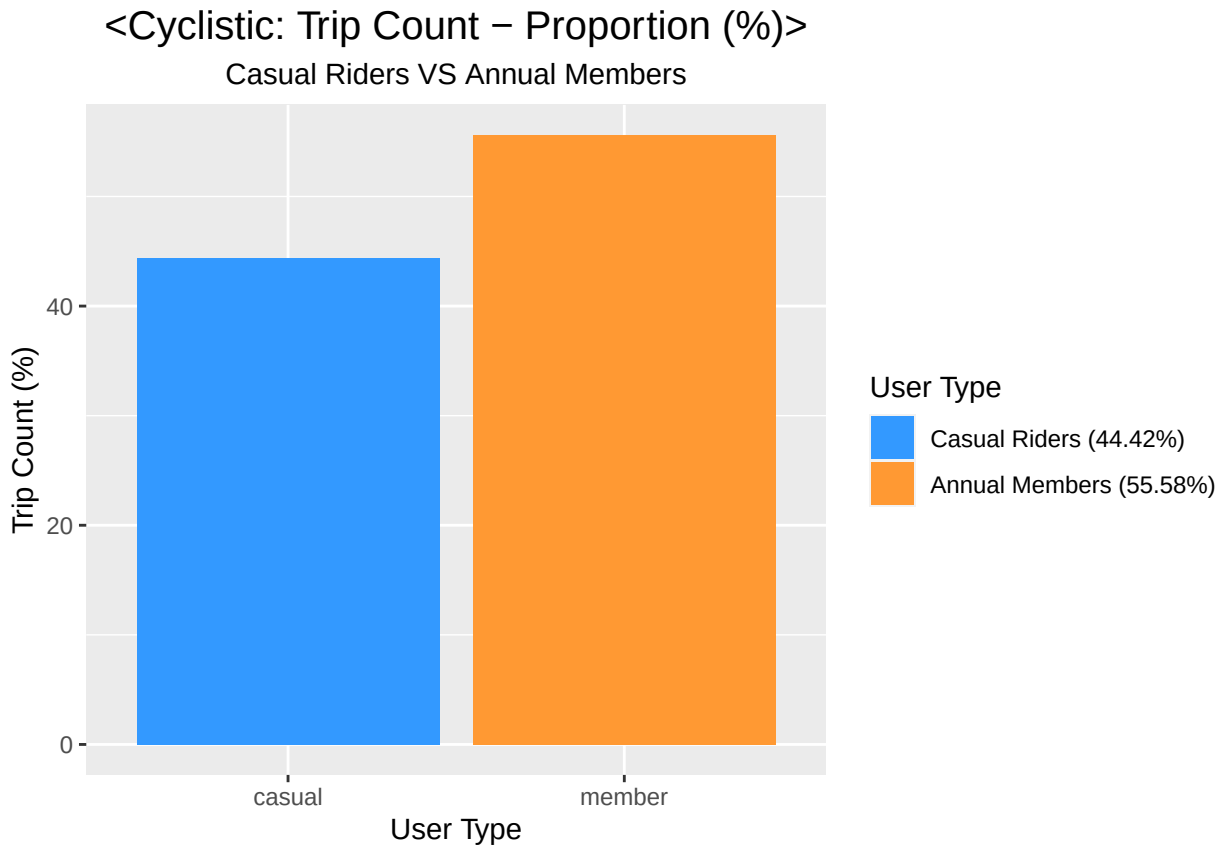
- casual riders make up 44.42% of the total trip count

```
# Calculate % of the trip count for annual members
agg_trip_count_member <- agg_count[2,2] / (agg_count[1,2] + agg_count[2,2])
agg_trip_count_member <- round(agg_trip_count_member * 100, 2)
agg_trip_count_member
```

```
## [1] 55.58
```

- Annual members make up 55.58% of the total trip count

```
# Visualize

plot3_df <- data.frame(x_tmp_cnt=c("casual", "member"),
                       y_tmp_cnt=c(agg_trip_count_casual, agg_trip_count_member))

plot3 <- ggplot(plot3_df, aes(x= x_tmp_cnt, y= y_tmp_cnt, fill=x_tmp_cnt)) +
  geom_col(position = "dodge")


# format title and subtitle
plot3 <-  plot3 + labs(title = "<Cyclistic: Trip Count - Proportion (%)>",
       subtitle = "Casual Riders VS Annual Members",
       x = "User Type", y = "Trip Count (%)")
plot3 <- plot3 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5))


# format legend
plot3 <- plot3 + scale_fill_manual(name="User Type",
                                   breaks = c("casual","member"),
                                   labels= c("Casual Riders (44.42%)","Annual Members (55.58%)"),
                                   values=c("#3399FF","#FF9933"))

print(plot3)
```



- **Analysis: Although trip duration for casual riders are *more* than annual members, but the actual trip count is *less*!**

Service Usage - Monthly Trend

```
plot4_df <- all_trips_v2 %>%
  group_by(member_casual, year_month) %>%
  summarise(trip_duration_sum = sum(ride_length_min))
```

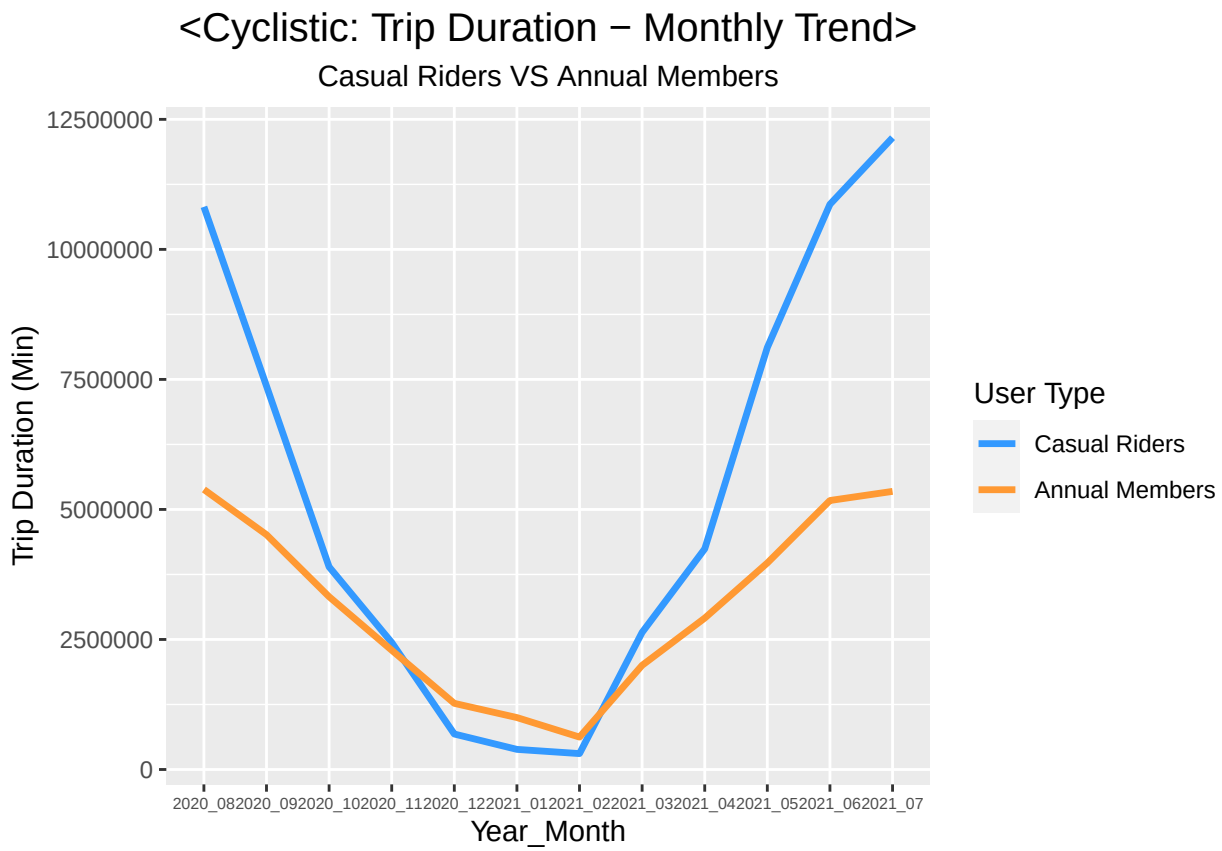
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot4_df
```

```
## # A tibble: 24 x 3
## # Groups:   member_casual [2]
##    member_casual year_month trip_duration_sum
##    <chr>         <chr>                  <dbl>
##  1 casual        2020_08            10818275.
##  2 casual        2020_09             7372470.
##  3 casual        2020_10             3895454.
##  4 casual        2020_11             2442368.
##  5 casual        2020_12              681975.
```

```
##  6 casual        2021_01              386917.
##  7 casual        2021_02              305542.
##  8 casual        2021_03             2635884.
##  9 casual        2021_04             4241672.
## 10 casual        2021_05             8113162.
## # ... with 14 more rows
```

```r
# Visualize
plot4 <- ggplot(plot4_df, aes(x=year_month, y=trip_duration_sum,
                              group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot4 <-  plot4 + labs(title = "<Cyclistic: Trip Duration - Monthly Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Year_Month", y = "Trip Duration (Min)")
plot4 <- plot4 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot4 <- plot4 + scale_color_manual(name="User Type",
                                    breaks = c("casual","member"),
                                    labels= c("Casual Riders","Annual Members"),
                                    values=c("#3399FF","#FF9933"))

plot4
```



- **Analysis: Trip duration start to decline drastically in Oct and hit lowest in Dec, start to pick up again in Feb. As expected, bike usage in Chicago during *the winter season is the lowest. the summer season is the highest*!**

```
# Calculate monthly AVG trip duration
plot5_df <- all_trips_v2 %>%
  group_by(member_casual, year_month) %>%
  summarise(trip_duration_avg = mean(ride_length_min))
```

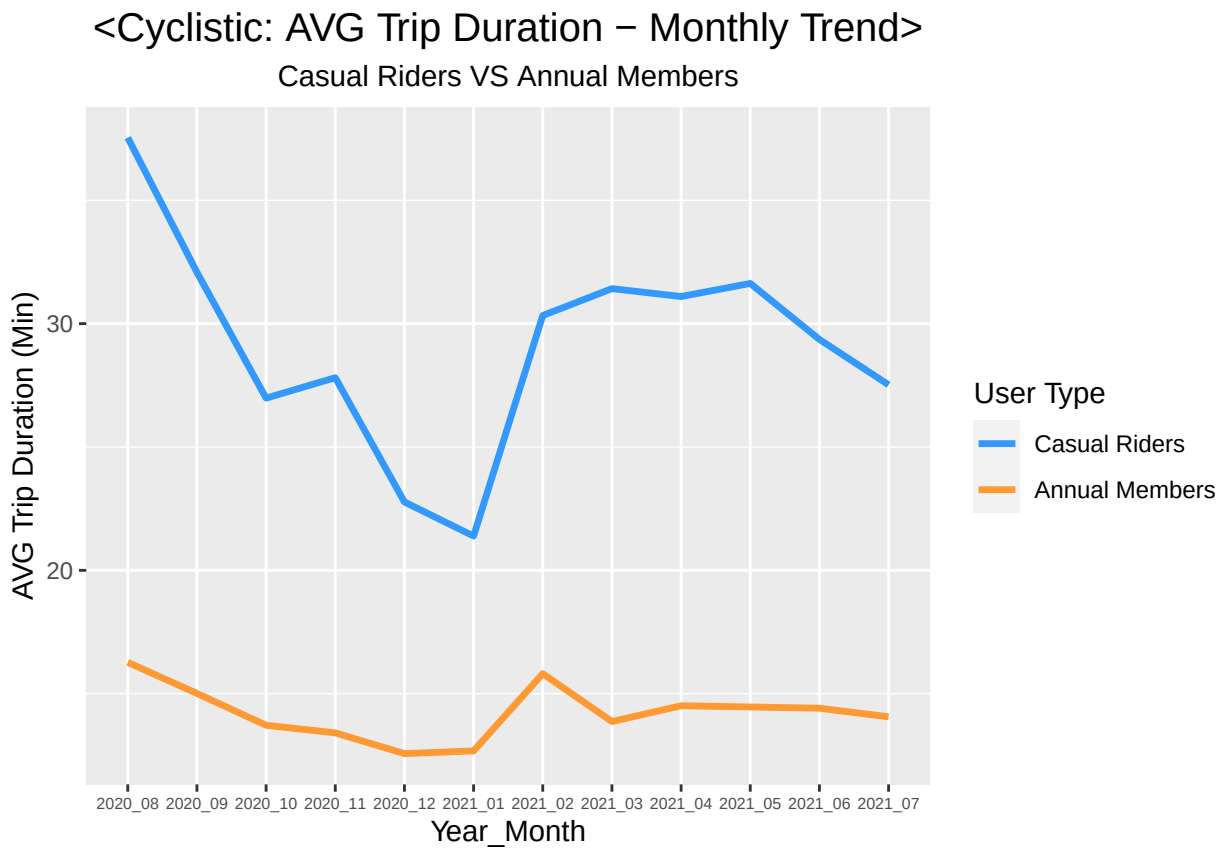
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot5_df[3] <- round(plot5_df[3],2)
plot5_df
```

```
## # A tibble: 24 x 3
## # Groups:   member_casual [2]
##    member_casual year_month trip_duration_avg
##    <chr>         <chr>                  <dbl>
##  1 casual        2020_08                 37.5
##  2 casual        2020_09                 32.1
##  3 casual        2020_10                 27.0
##  4 casual        2020_11                 27.8
##  5 casual        2020_12                 22.8
```

```
##  6 casual        2021_01                21.4
##  7 casual        2021_02                30.3
##  8 casual        2021_03                31.4
##  9 casual        2021_04                31.1
## 10 casual        2021_05                31.6
## # ... with 14 more rows
```

```r
# Visualize
plot5 <- ggplot(plot5_df, aes(x=year_month, y=trip_duration_avg,
                              group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot5 <-  plot5 + labs(title = "<Cyclistic: AVG Trip Duration - Monthly Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Year_Month", y = "AVG Trip Duration (Min)")
plot5 <- plot5 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot5 <- plot5 + scale_color_manual(name="User Type",
                                    breaks = c("casual","member"),
                                    labels= c("Casual Riders","Annual Members"),
                                    values=c("#3399FF","#FF9933"))

plot5
```



- **Analysis: AVG trip duration for casual riders *constantly changes* from 22min to 37min. However, AVG trip duration for annual members is *relatively stable* around 14min**

```
# Calculate monthly trip count
plot6_df <- all_trips_v2 %>%
  group_by(member_casual, year_month) %>%
  summarise(trip_duration_count = length(ride_length_min))
```

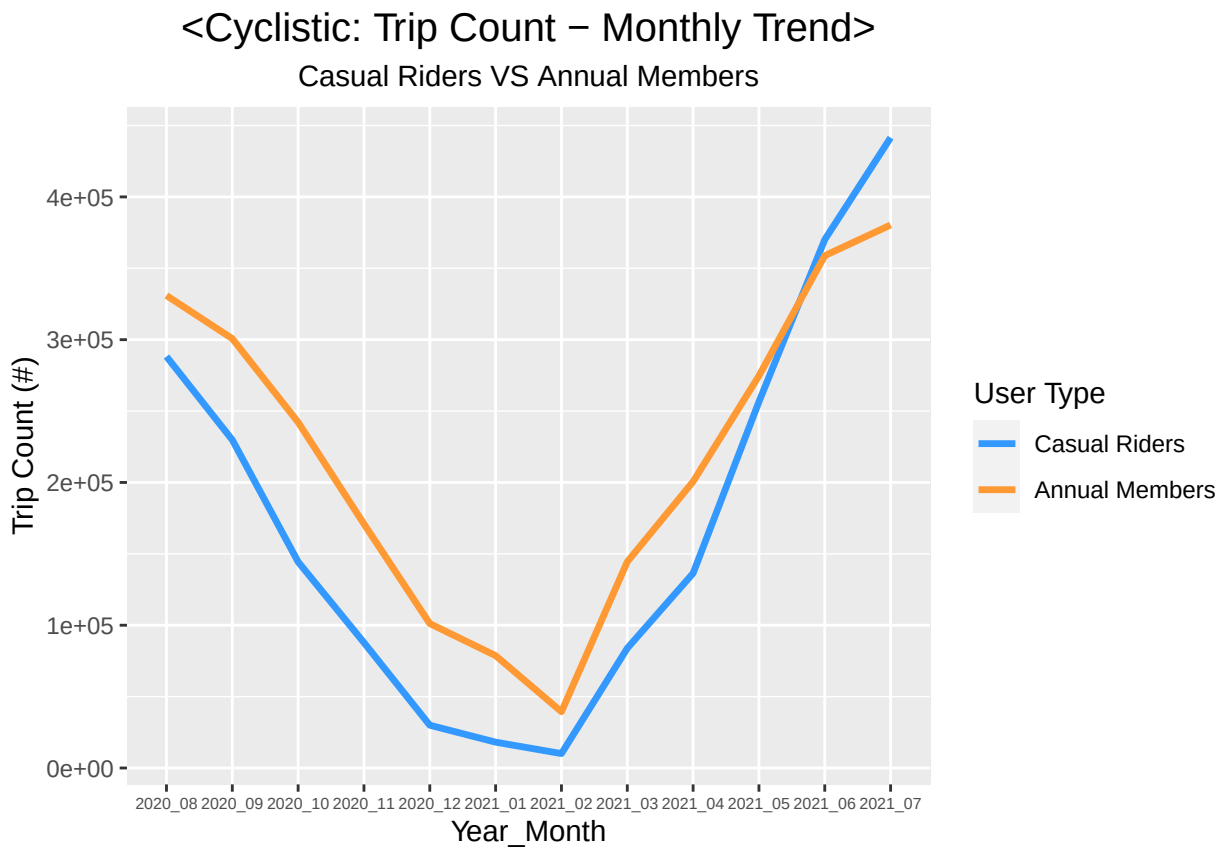
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot6_df
```

```
## # A tibble: 24 x 3
## # Groups:   member_casual [2]
##    member_casual year_month trip_duration_count
##    <chr>         <chr>                    <int>
##  1 casual        2020_08                 288183
##  2 casual        2020_09                 229800
##  3 casual        2020_10                 144368
##  4 casual        2020_11                  87820
##  5 casual        2020_12                  29956
##  6 casual        2021_01                  18090
##  7 casual        2021_02                  10073
```

```
##  8 casual        2021_03                 83880
##  9 casual        2021_04                136370
## 10 casual        2021_05                256508
## # ... with 14 more rows
```

```
# Visualize
plot6 <- ggplot(plot6_df, aes(x=year_month, y=trip_duration_count,
                              group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot6 <-  plot6 + labs(title = "<Cyclistic: Trip Count - Monthly Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Year_Month", y = "Trip Count (#)")
plot6 <- plot6 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot6 <- plot6 + scale_color_manual(name="User Type",
                                    breaks = c("casual","member"),
                                    labels= c("Casual Riders","Annual Members"),
                                    values=c("#3399FF","#FF9933"))

plot6
```



- **Analysis: Trip count for annual members is *almost always higher* than casual riders. This indicate, *annual members uses the service more frequently* even if trip duration for each ride may be only around 14min avg.**

Service Usage - Monthly Trend by Bike Type

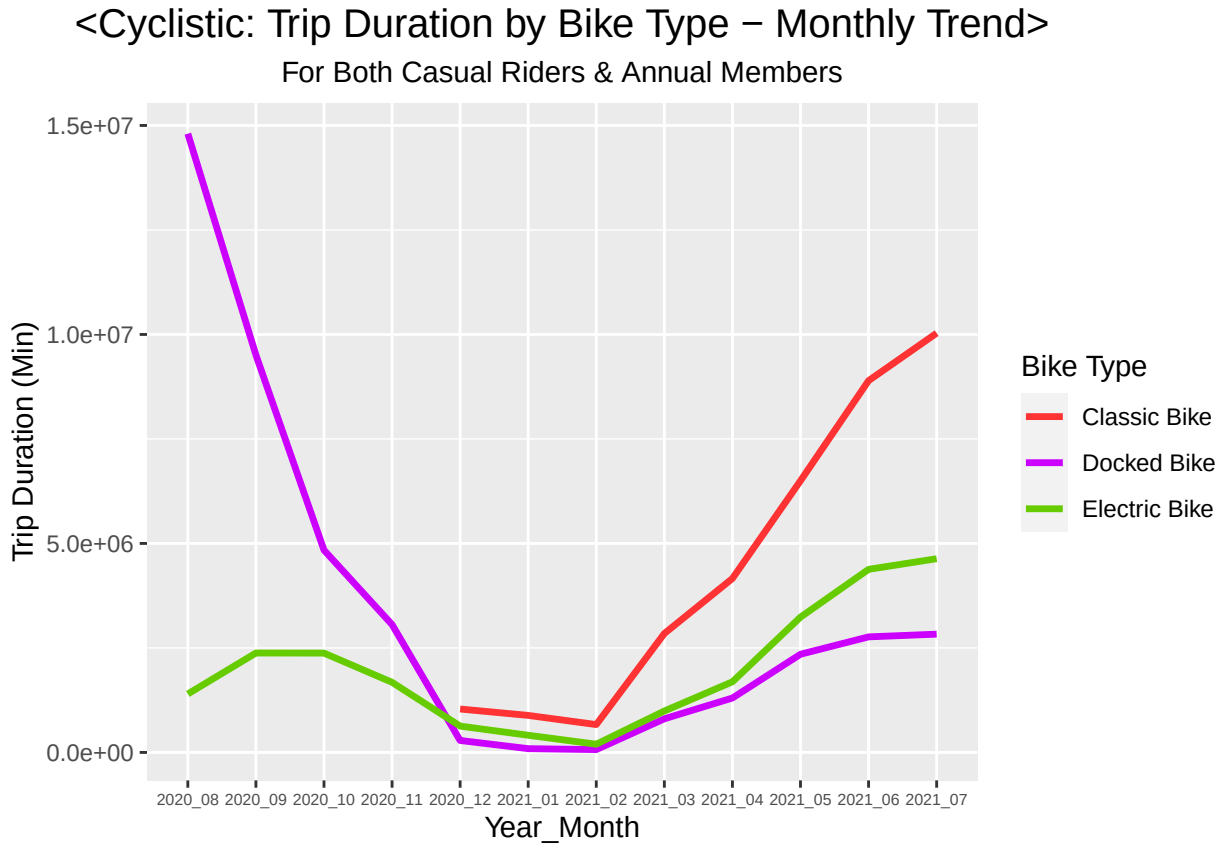
```
plot7_df <- all_trips_v2 %>%
  group_by(rideable_type, year_month) %>%
  summarise(trip_duration_sum_biketype = sum(ride_length_min))
```

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

```
plot7_df
```

```
## # A tibble: 32 x 3
## # Groups:   rideable_type [3]
##    rideable_type year_month trip_duration_sum_biketype
##    <chr>         <chr>                          <dbl>
##  1 classic_bike  2020_12                      1038375.
##  2 classic_bike  2021_01                       884974.
##  3 classic_bike  2021_02                       666107.
##  4 classic_bike  2021_03                      2844320.
##  5 classic_bike  2021_04                      4161664.
##  6 classic_bike  2021_05                      6499162.
##  7 classic_bike  2021_06                      8894596.
##  8 classic_bike  2021_07                     10031437.
##  9 docked_bike   2020_08                     14803881
## 10 docked_bike   2020_09                      9509037.
## # ... with 22 more rows
```

```
# Visualize
plot7 <- ggplot(plot7_df, aes(x=year_month, y=trip_duration_sum_biketype,
                              group=rideable_type)) +
  geom_line(aes(color=rideable_type), size=1.2)

# format title and subtitle
plot7 <-  plot7 + labs(title = "<Cyclistic: Trip Duration by Bike Type - Monthly Trend>",
       subtitle = "For Both Casual Riders & Annual Members",
       x = "Year_Month", y = "Trip Duration (Min)")
plot7 <- plot7 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot7 <- plot7 + scale_color_manual(name="Bike Type",
                                    breaks = c("classic_bike","docked_bike", "electric_bike"),
                                    labels= c("Classic Bike","Docked Bike", "Electric Bike"),
                                    values=c("#FF3333","#CC00FF", "#66CC00"))

plot7
```



- **Analysis:** ***Classic bike is the most favored*** **bike type!**

```
# See if casual riders also favors classic bike

plot8_df <- all_trips_v2 %>%
  group_by(rideable_type, year_month, member_casual) %>%
  filter(member_casual == "casual") %>%
  summarise(trip_duration_sum_biketype = sum(ride_length_min))
```

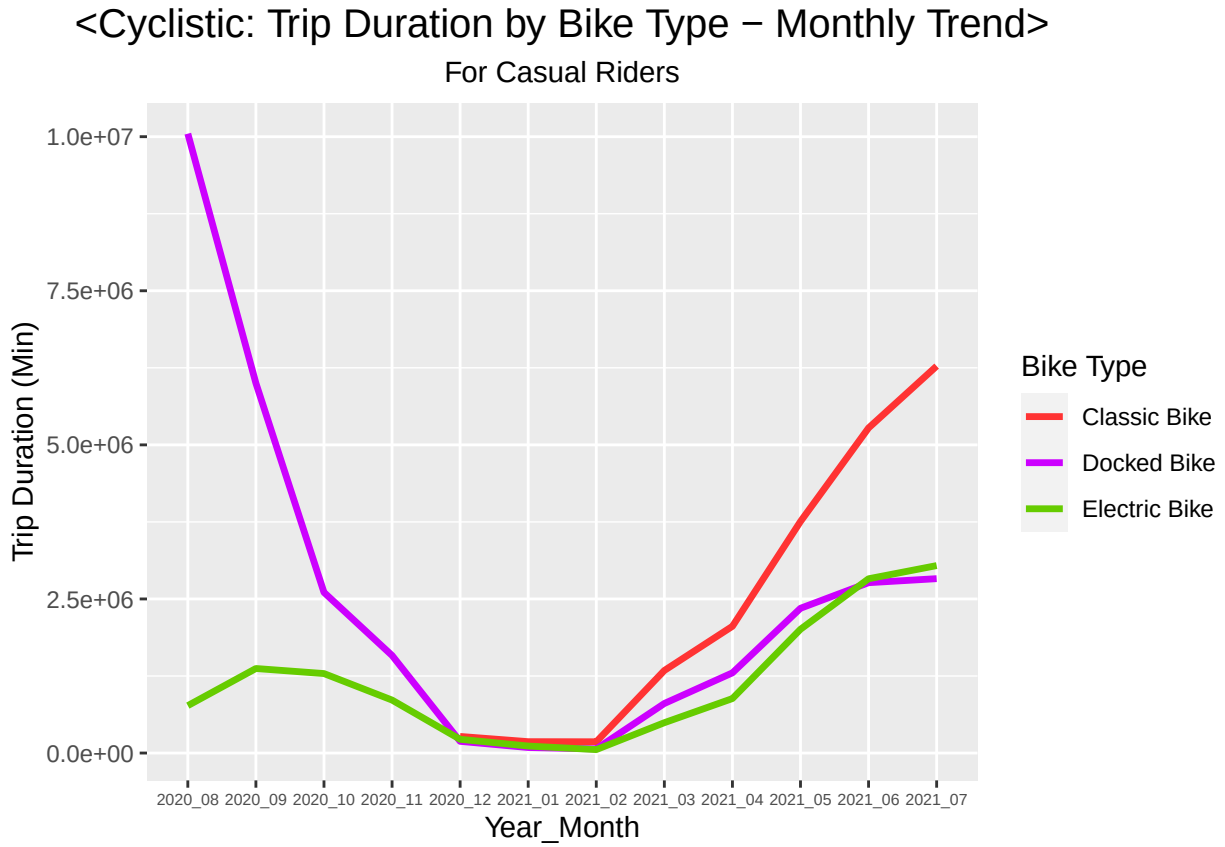
```
## `summarise()` has grouped output by 'rideable_type', 'year_month'. You can
## override using the `.groups` argument.
```

```
plot8_df
```

```
## # A tibble: 32 x 4
## # Groups:   rideable_type, year_month [32]
##    rideable_type year_month member_casual trip_duration_sum_biketype
##    <chr>         <chr>      <chr>                              <dbl>
##  1 classic_bike  2020_12    casual                           269172.
##  2 classic_bike  2021_01    casual                           184207.
##  3 classic_bike  2021_02    casual                           182772.
##  4 classic_bike  2021_03    casual                          1339420.
##  5 classic_bike  2021_04    casual                          2056571.
##  6 classic_bike  2021_05    casual                          3756683.
```

```
##  7 classic_bike  2021_06    casual                          5274439.
##  8 classic_bike  2021_07    casual                          6277790.
##  9 docked_bike   2020_08    casual                         10049070.
## 10 docked_bike   2020_09    casual                          6001070.
## # ... with 22 more rows
```

```
# Visualize
plot8 <- ggplot(plot8_df, aes(x=year_month, y=trip_duration_sum_biketype,
                              group=rideable_type)) +
  geom_line(aes(color=rideable_type), size=1.2)

# format title and subtitle
plot8 <-  plot8 + labs(title = "<Cyclistic: Trip Duration by Bike Type - Monthly Trend>",
       subtitle = "For Casual Riders",
       x = "Year_Month", y = "Trip Duration (Min)")
plot8 <- plot8 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot8 <- plot8 + scale_color_manual(name="Bike Type",
                                    breaks = c("classic_bike","docked_bike", "electric_bike"),
                                    labels= c("Classic Bike","Docked Bike", "Electric Bike"),
                                    values=c("#FF3333","#CC00FF", "#66CC00"))

plot8
```



- **Analysis: Yes!** ***Casual riders also favor Classic bike*** **type!**

```
# See if annual riders also favors classic bike

plot9_df <- all_trips_v2 %>%
  group_by(rideable_type, year_month, member_casual) %>%
  filter(member_casual == "member") %>%
  summarise(trip_duration_sum_biketype = sum(ride_length_min))
```

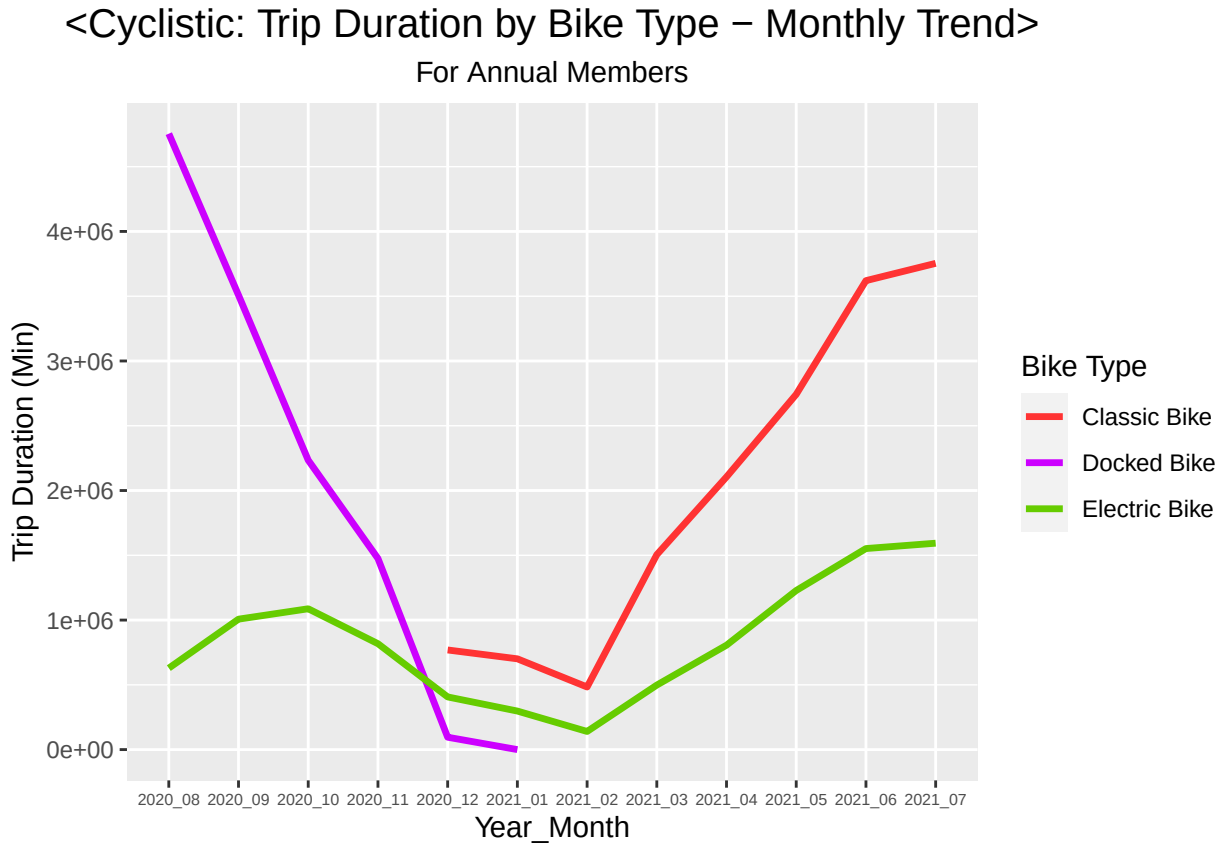
```
## `summarise()` has grouped output by 'rideable_type', 'year_month'. You can
## override using the `.groups` argument.
```

```
plot9_df
```

```
## # A tibble: 26 x 4
## # Groups:   rideable_type, year_month [26]
##    rideable_type year_month member_casual trip_duration_sum_biketype
##    <chr>         <chr>      <chr>                              <dbl>
##  1 classic_bike  2020_12    member                           769203.
##  2 classic_bike  2021_01    member                           700767.
##  3 classic_bike  2021_02    member                           483335.
##  4 classic_bike  2021_03    member                          1504900.
##  5 classic_bike  2021_04    member                          2105093.
##  6 classic_bike  2021_05    member                          2742478.
```

```
##  7 classic_bike  2021_06    member                          3620157.
##  8 classic_bike  2021_07    member                          3753647.
##  9 docked_bike   2020_08    member                          4754811.
## 10 docked_bike   2020_09    member                          3507967.
## # ... with 16 more rows
```

```r
# Visualize
plot9 <- ggplot(plot9_df, aes(x=year_month, y=trip_duration_sum_biketype,
                              group=rideable_type)) +
  geom_line(aes(color=rideable_type), size=1.2)

# format title and subtitle
plot9 <-  plot9 + labs(title = "<Cyclistic: Trip Duration by Bike Type - Monthly Trend>",
       subtitle = "For Annual Members",
       x = "Year_Month", y = "Trip Duration (Min)")
plot9 <- plot9 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                       plot.subtitle = element_text(hjust = 0.5),
                       axis.text.x = element_text(size=6))


# format legend
plot9 <- plot9 + scale_color_manual(name="Bike Type",
                                    breaks = c("classic_bike","docked_bike", "electric_bike"),
                                    labels= c("Classic Bike","Docked Bike", "Electric Bike"),
                                    values=c("#FF3333","#CC00FF", "#66CC00"))

plot9
```



- **Analysis:** ***Annual members clearly favor Classic-bike* type! and *completely stopped using Docked-bike*!**

```
# is the classic bike still the favorite from the perspective of trip-count?
plot10_df <- all_trips_v2 %>%
  group_by(rideable_type, year_month) %>%
  summarise(trip_duration_count_biketype = length(ride_length_min))
```

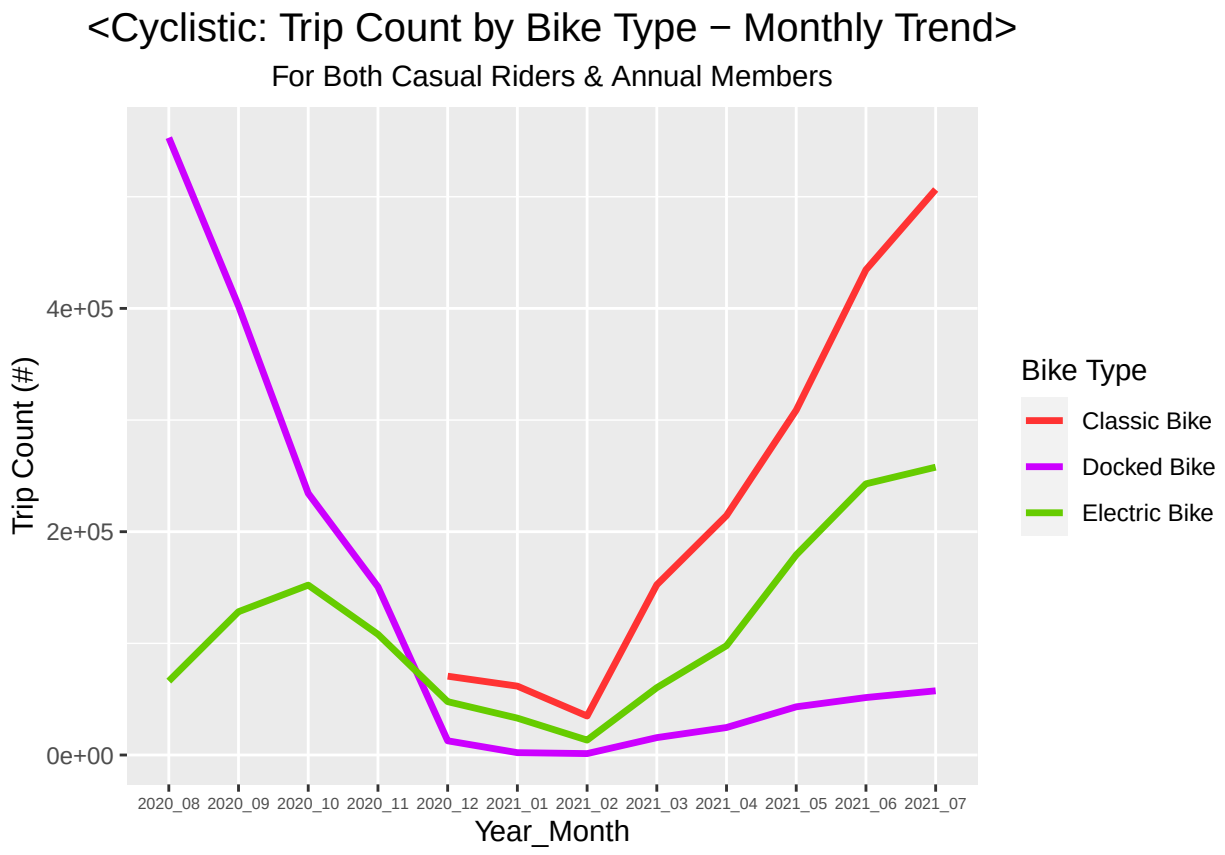
```
## `summarise()` has grouped output by 'rideable_type'. You can override using the
## `.groups` argument.
```

```
plot10_df
```

```
## # A tibble: 32 x 3
## # Groups:   rideable_type [3]
##    rideable_type year_month trip_duration_count_biketype
##    <chr>         <chr>                             <int>
##  1 classic_bike  2020_12                           70587
##  2 classic_bike  2021_01                           61675
##  3 classic_bike  2021_02                           34916
##  4 classic_bike  2021_03                          152470
##  5 classic_bike  2021_04                          214466
##  6 classic_bike  2021_05                          308860
##  7 classic_bike  2021_06                          434603
```

```
##  8 classic_bike  2021_07                          506473
##  9 docked_bike   2020_08                          552908
## 10 docked_bike   2020_09                          402168
## # ... with 22 more rows
```

```r
# Visualize
plot10 <- ggplot(plot10_df, aes(x=year_month, y=trip_duration_count_biketype,
                                group=rideable_type)) +
  geom_line(aes(color=rideable_type), size=1.2)

# format title and subtitle
plot10 <-  plot10 + labs(title = "<Cyclistic: Trip Count by Bike Type - Monthly Trend>",
       subtitle = "For Both Casual Riders & Annual Members",
       x = "Year_Month", y = "Trip Count (#)")
plot10 <- plot10 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                         plot.subtitle = element_text(hjust = 0.5),
                         axis.text.x = element_text(size=6))


# format legend
plot10 <- plot10 + scale_color_manual(name="Bike Type",
                                      breaks = c("classic_bike","docked_bike", "electric_bike"),
                                      labels= c("Classic Bike","Docked Bike", "Electric Bike"),
                                      values=c("#FF3333","#CC00FF", "#66CC00"))

plot10
```



- **Analysis: Yes! The *Classic-bike is the favorite bike type* for both user types!**

Service Usage - Days of Week Trend

```
#Rearrange the days of the week in order.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week,
                                    levels=c("Sunday", "Monday", "Tuesday",
                                             "Wednesday", "Thursday", "Friday",
                                             "Saturday"))

plot11_df <- all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(trip_duration_sum = sum(ride_length_min))
```

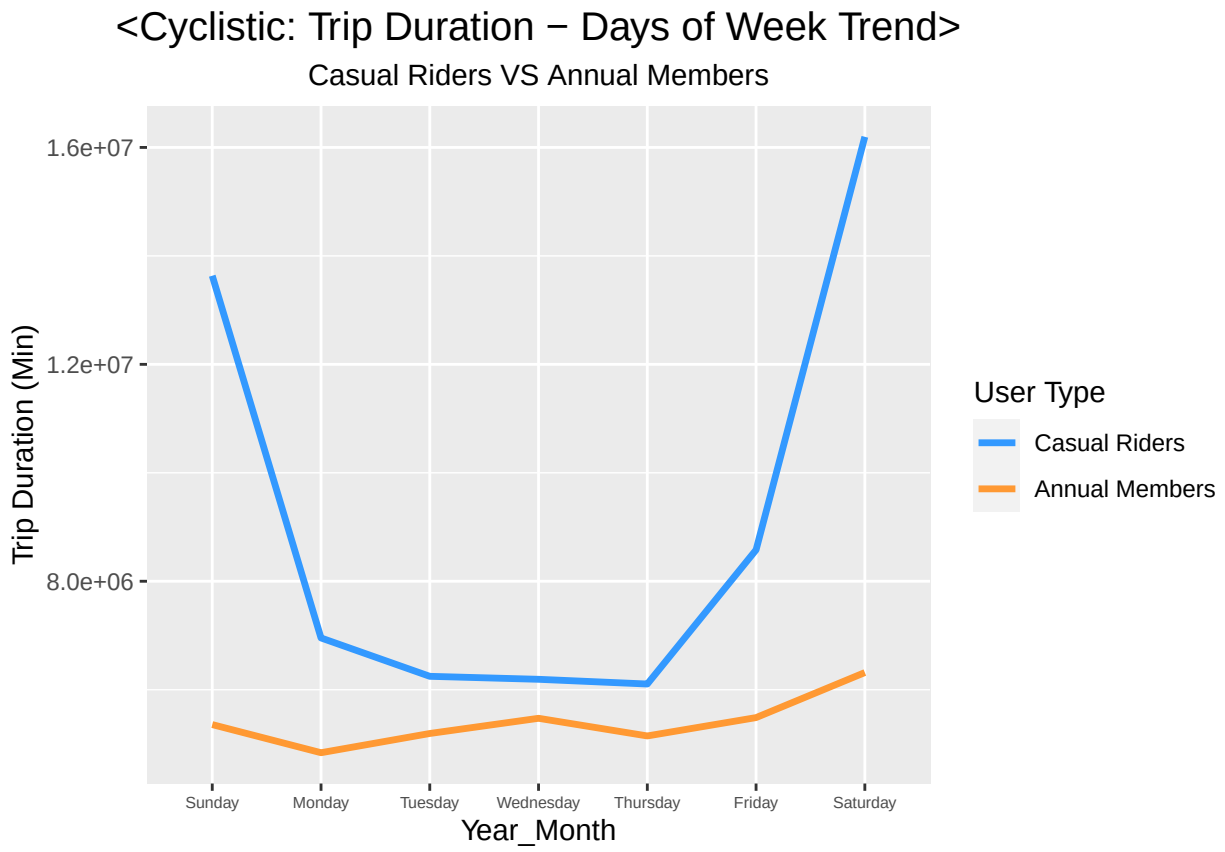
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot11_df
```

```
## # A tibble: 14 x 3
## # Groups:   member_casual [2]
##    member_casual day_of_week trip_duration_sum
##    <chr>         <ord>                   <dbl>
##  1 casual        Sunday              13634201.
##  2 casual        Monday               6954542.
##  3 casual        Tuesday              6247249.
##  4 casual        Wednesday            6192255.
##  5 casual        Thursday             6104964.
##  6 casual        Friday               8578532.
##  7 casual        Saturday            16196327.
##  8 member        Sunday               5357026.
##  9 member        Monday               4838078.
## 10 member        Tuesday              5192677.
## 11 member        Wednesday            5472913.
## 12 member        Thursday             5147419.
## 13 member        Friday               5485908.
## 14 member        Saturday             6316426.
```

```
# Visualize
plot11 <- ggplot(plot11_df, aes(x=day_of_week, y=trip_duration_sum,
                                group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot11 <-  plot11 + labs(title = "<Cyclistic: Trip Duration - Days of Week Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Year_Month", y = "Trip Duration (Min)")
plot11 <- plot11 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                         plot.subtitle = element_text(hjust = 0.5),
                         axis.text.x = element_text(size=6))
```

```
# format legend
plot11 <- plot11 + scale_color_manual(name="User Type",
                                      breaks = c("casual","member"),
                                      labels= c("Casual Riders","Annual Members"),
                                      values=c("#3399FF","#FF9933"))

plot11
```



- **Analysis: Casual riders clearly use bike on the *weekends*. This may indicate, *casual riders use bike for leisure***

```
plot12_df <- all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(trip_duration_count = length(ride_length_min))
```

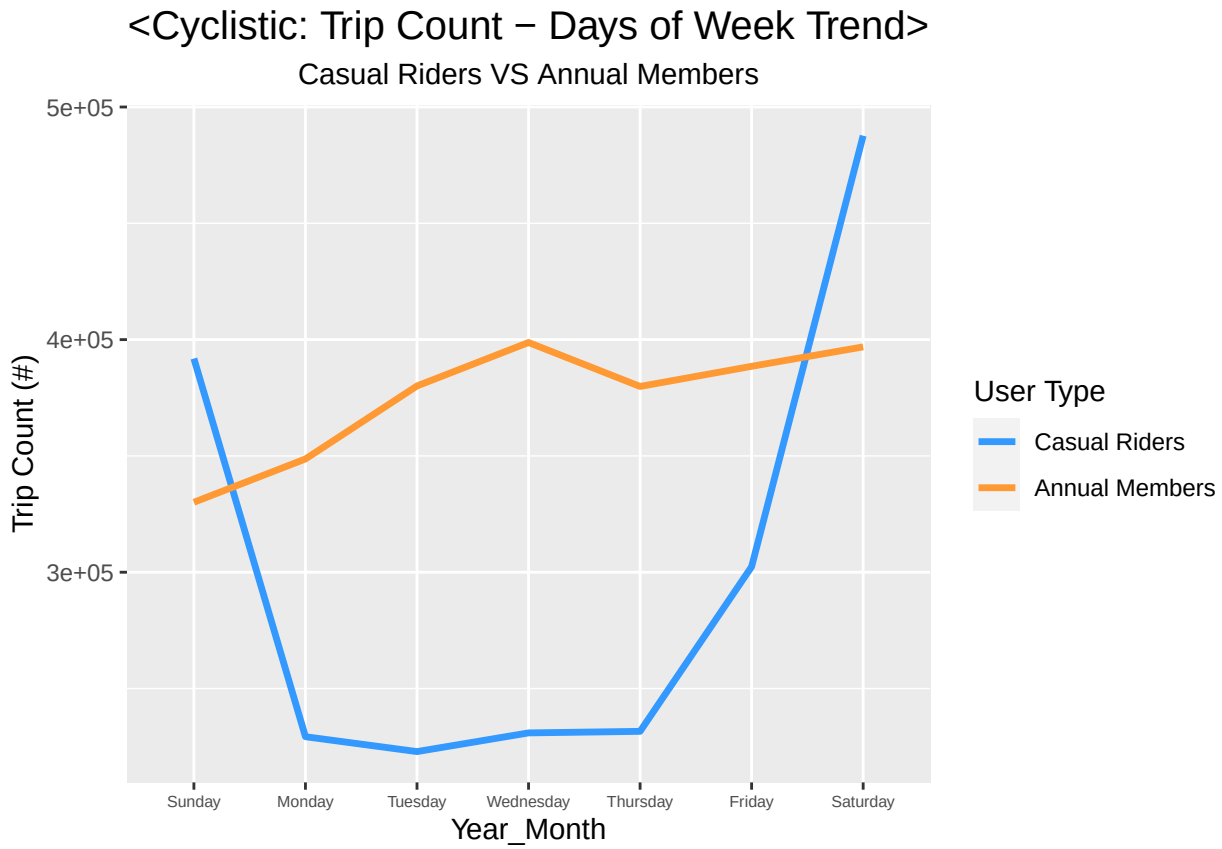
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot12_df
```

```
## # A tibble: 14 x 3
## # Groups:   member_casual [2]
##    member_casual day_of_week trip_duration_count
```

```
##    <chr>         <ord>                 <int>
##  1 casual        Sunday                391861
##  2 casual        Monday                229288
##  3 casual        Tuesday               222882
##  4 casual        Wednesday             230924
##  5 casual        Thursday              231593
##  6 casual        Friday                302378
##  7 casual        Saturday              487682
##  8 member        Sunday                330073
##  9 member        Monday                348698
## 10 member        Tuesday               380071
## 11 member        Wednesday             398756
## 12 member        Thursday              379866
## 13 member        Friday                388523
## 14 member        Saturday              396882
```

```r
# Visualize
plot12 <- ggplot(plot12_df, aes(x=day_of_week, y=trip_duration_count,
                                group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot12 <-  plot12 + labs(title = "<Cyclistic: Trip Count - Days of Week Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Year_Month", y = "Trip Count (#)")
plot12 <- plot12 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                         plot.subtitle = element_text(hjust = 0.5),
                         axis.text.x = element_text(size=6))


# format legend
plot12 <- plot12 + scale_color_manual(name="User Type",
                                      breaks = c("casual","member"),
                                      labels= c("Casual Riders","Annual Members"),
                                      values=c("#3399FF","#FF9933"))

plot12
```



- **Analysis: Annual Members clearly uses bikes during the *weekdays*. This may indicate, annual members *uses the service to commute* to and from work or school!**

Service Usage - Hourly Trend

```
plot13_df <- all_trips_v2 %>%
  group_by(member_casual, hour) %>%
  summarise(trip_duration_count = length(ride_length_min))
```

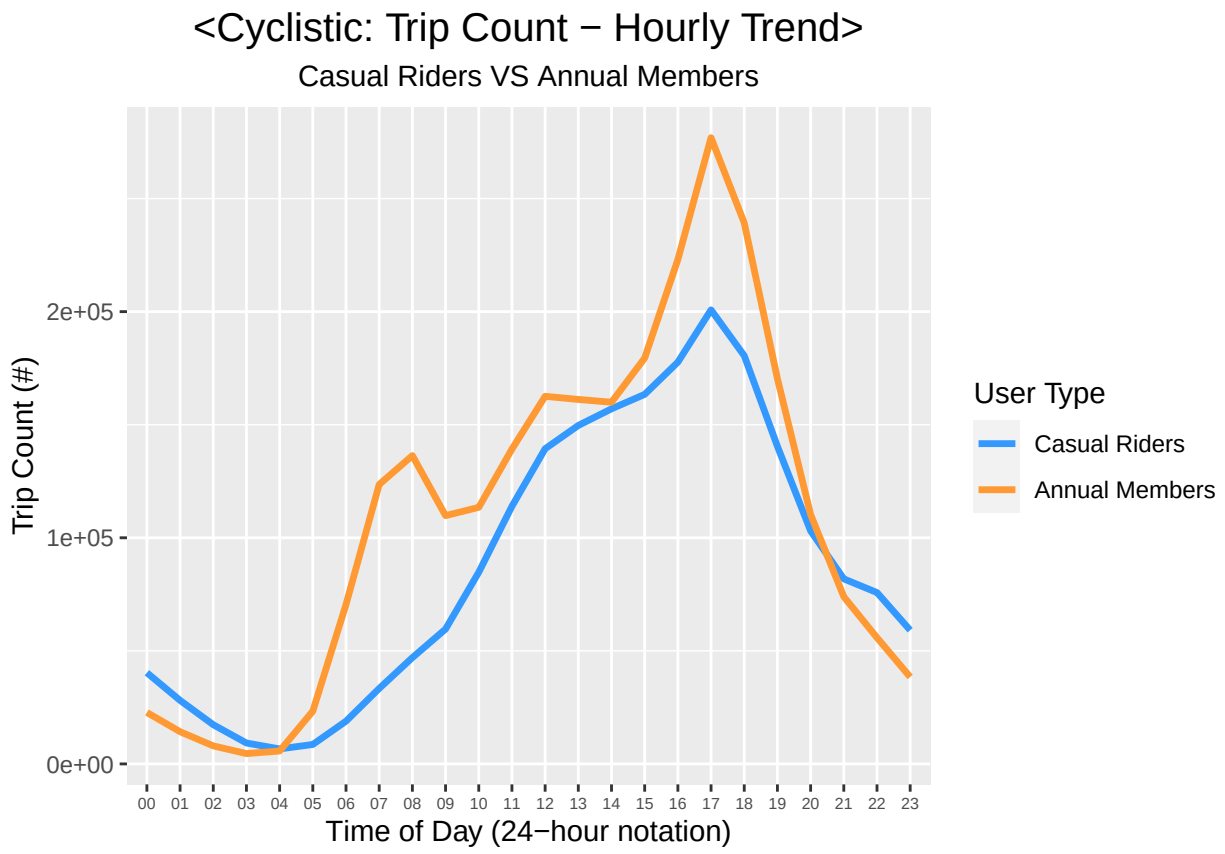
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
plot13_df
```

```
## # A tibble: 48 x 3
## # Groups:   member_casual [2]
##    member_casual hour  trip_duration_count
##    <chr>         <chr>               <int>
##  1 casual        00                  40331
##  2 casual        01                  28107
##  3 casual        02                  17339
##  4 casual        03                   9246
##  5 casual        04                   6627
```

```
##  6 casual        05                  8598
##  7 casual        06                 18902
##  8 casual        07                 33365
##  9 casual        08                 47006
## 10 casual        09                 59643
## # ... with 38 more rows
```

```
# Visualize
plot13 <- ggplot(plot13_df, aes(x=hour, y=trip_duration_count,
                                group=member_casual)) +
  geom_line(aes(color=member_casual), size=1.2)

# format title and subtitle
plot13 <-  plot13 + labs(title = "<Cyclistic: Trip Count - Hourly Trend>",
       subtitle = "Casual Riders VS Annual Members",
       x = "Time of Day (24-hour notation)", y = "Trip Count (#)")
plot13 <- plot13 + theme(plot.title = element_text(size = 15, hjust = 0.5),
                         plot.subtitle = element_text(hjust = 0.5),
                         axis.text.x = element_text(size=6))


# format legend
plot13 <- plot13 + scale_color_manual(name="User Type",
                                      breaks = c("casual","member"),
                                      labels= c("Casual Riders","Annual Members"),
                                      values=c("#3399FF","#FF9933"))

plot13
```



- **Analysis:** ***Annual members make three hills from 5am to 9am, from 11am to 2pm, and from 3pm to 6pm.*** **Afternoon hill is bigger than the one in the morning. This may indicate that** ***annual members use bike more to commute from work or school to home.*** **However, casual riders only make one big hill in the afternoon which may indicate that** ***casual riders prefer to use bike in the afternoon for leisure.***

5. Summary of the Analysis

On-Season and Off-Season

- The **preseason** starts in *February*
  - The preseaon is the *best time* to begin the marketing campaign
- The **on-season** starts in *April*
- The **off-season** starts in *October*

Analysis on Casual Riders

- **Casual riders** make up *62.83%* of the total trip duration. Avg trip duration is about *30 min* per ride. However, they *only make up 44.42%* of the total trip count. The **service usage declines in Autumn and hits the lowest during the winter season**, which is expected in the city of Chicago. The service usage **picks up in Feb and hit the highest in July**. Ever since the introduction of the **classic bike**, casual riders tend to use it than docked bike. The usage of the electric bike is the second favored bike type consistently. Also, casual riders tend to use the service on the **weekends** which may indicate they ride bike **for leisure in the afternoon**!

Analysis on Annual Members

- **Annual members** make up only *37.17%* of the total trip duration. Avg trip duration is about *14 min* per ride. However, they make up *55.58%* of the total trip count, which indicate that annual members **use the service more frequently than casual riders**. The **service usage declines in Autumn and hits the lowest during the winder season**, which is expected in the city of Chicago. The service usage **picks up in Feb and hit the highest in June**. Even for annual members, they prefer to use the bike during the **summer season** as well. Ever sine the introduction of the **classic bike**, annual members favored it over other bike types. They completely stopped using the docked-bike. Contrast to casual riders, service usage of the annual members is consistent throughout the week but **tend to have higher usage during the week**. This may indicate annual members tend to use the service **to commute from work or from school where the distance is within 15min by bike**. They tend to use bike *from 5am to 9am, from 11am to 2pm, and from 3pm to 6pm.* They seem to **use the service more in the afternoon to commute** after work or school*.

Conclusion of the Analysis

Top 3 Recommendations Based on the Analysis

Recommendation 1: Targeting AVG Trip Duration of Casual Riders

- Casual riders make up *62.83%* of the total trip duration and ride for about **30min** per ride.
  - Therefore, if there is an **annual membership have merits for *leisure riders* that will benefit those who *ride over 30 minutes per ride* will appeal many of the casual riders to get the membership**.

Recommendation 2: Targeting Weekend Casual Riders

- Casual riders tend to ride bike in the afternoon **on the weekend**.
  - Therefore, if there is an **annual membership have merits for *leisure riders* that will benefit *the weekend riders* will appeal many of the casual riders to get the membership**.

Recommendation 3: Targeting Favorite Bike Type of Casual Riders

- Casual riders favor classic bike.
  - Therefore, if there is an **annual *membership have merits for leisure riders* that will benefit the riders who use *classic bike* appeal to many of the casual riders to get the membership**.