Google Data Analytics Case Study (Bike-Share)

Raddad Numan

2022-03-20

Introduction

This is the final capstone for the Google data analytics program, this case study will let me apply many skills I have gained to a real world data set, to drive business dicisions.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Cyclistic:

A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can’t use a standard two-wheeled bike.

Ask Phase

Identifying the business task.

- The differences between casual riders and the annual members, and how to consider them or other factors to encourage casual riders to buy annual memberships.

- digital media marketing strategy to increase the annual members.

Prepare Phase

Data source

> Data has been downloaded from Motivate International Inc (https://divvy-tripdata.s3.amazonaws.com/index.html). under the License (https://www.divvybikes.com/data-license-agreement).
>
> The data was properly stored locally with copies have been stored securely on Google Drive.the data is organized by month, with (.CSV) format. for the twelve months of 2021.
>
> We will assume that the data is credible since it’s public and provided by the google data analytics program. .

Using Rstudio to begin cleaning and combining the data

Installing the required packages

```
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

```
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

```
install.packages("lubridate", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

```
install.packages("janitor", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

```
install.packages("skimr", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

```
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q7/lcx3p6nx60s220lqcvyl0wvr0000gn/T//Rtmp9r9Osy/downloaded_packages
```

Calling the installed packages

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

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

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

```
library("lubridate")
```

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

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

```
library("skimr")
library("janitor")
```

```
##
## Attaching package: 'janitor'
```

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

```
library("ggplot2")
```

Importing the Data

```
divvy202101 <- 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.
```

```
divvy202102 <- 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.
```

```
divvy202103 <- 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.
```

```
divvy202104 <- 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.
```

```
divvy202105 <- 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.
```

```
divvy202106 <- 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.
```

```
divvy202107 <- 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.
```

```
divvy202108 <- 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.
```

```
divvy202109 <- 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.
```

```
divvy202110 <- 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.
```

```
divvy202111 <- 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.
```

```
divvy202112 <- 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.
```

Comparing the individual data frames structure and making them ready to combine them into one large data set for further analysis.

To view the first six rows of the first data frame

```
head(divvy202101)
```

```
## # A tibble: 6 × 13
##   ride_id rideable_type started_at          ended_at            start_station_n…
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 E19E6F… electric_bike 2021-01-23 16:14:19 2021-01-23 16:24:44 California Ave …
## 2 DC88F2… electric_bike 2021-01-27 18:43:08 2021-01-27 18:47:12 California Ave …
## 3 EC45C9… electric_bike 2021-01-21 22:35:54 2021-01-21 22:37:14 California Ave …
## 4 4FA453… electric_bike 2021-01-07 13:31:13 2021-01-07 13:42:55 California Ave …
## 5 BE5E8E… electric_bike 2021-01-23 02:24:02 2021-01-23 02:24:45 California Ave …
## 6 5D8969… electric_bike 2021-01-09 14:24:07 2021-01-09 15:17:54 California Ave …
## # … 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>
```

To view quick summary of all and comparing

```
summary(divvy202101)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:96834       Length:96834       Min.   :2021-01-01 00:02:05  
##  Class :character   Class :character   1st Qu.:2021-01-08 20:55:02  
##  Mode  :character   Mode  :character   Median :2021-01-15 06:05:04  
##                                        Mean   :2021-01-15 17:57:29  
##                                        3rd Qu.:2021-01-22 09:28:48  
##                                        Max.   :2021-01-31 23:57:00  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-01-01 00:08:39   Length:96834       Length:96834      
##  1st Qu.:2021-01-08 21:14:23   Class :character   Class :character  
##  Median :2021-01-15 06:19:58   Mode  :character   Mode  :character  
##  Mean   :2021-01-15 18:12:46                                        
##  3rd Qu.:2021-01-22 09:41:18                                        
##  Max.   :2021-02-01 15:33:15                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:96834       Length:96834       Min.   :41.64   Min.   :-87.78  
##  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.   :42.06   Max.   :-87.53  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.64   Min.   :-87.81   Length:96834      
##  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.   :42.07   Max.   :-87.51                     
##  NA's   :103     NA's   :103
```

```
summary(divvy202102)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:49622       Length:49622       Min.   :2021-02-01 00:55:44  
##  Class :character   Class :character   1st Qu.:2021-02-09 08:20:56  
##  Mode  :character   Mode  :character   Median :2021-02-22 13:17:53  
##                                        Mean   :2021-02-18 01:16:52  
##                                        3rd Qu.:2021-02-26 16:02:13  
##                                        Max.   :2021-02-28 23:59:41  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-02-01 01:22:48   Length:49622       Length:49622      
##  1st Qu.:2021-02-09 08:36:02   Class :character   Class :character  
##  Median :2021-02-22 13:39:20   Mode  :character   Mode  :character  
##  Mean   :2021-02-18 01:41:18                                        
##  3rd Qu.:2021-02-26 16:19:32                                        
##  Max.   :2021-03-05 15:11:45                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:49622       Length:49622       Min.   :41.65   Min.   :-87.77  
##  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.   :42.06   Max.   :-87.53  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.54   Min.   :-87.77   Length:49622      
##  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.   :42.07   Max.   :-87.53                     
##  NA's   :214     NA's   :214
```

```
summary(divvy202103)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:228496      Length:228496      Min.   :2021-03-01 00:01:09  
##  Class :character   Class :character   1st Qu.:2021-03-10 10:45:36  
##  Mode  :character   Mode  :character   Median :2021-03-19 17:37:20  
##                                        Mean   :2021-03-17 23:22:08  
##                                        3rd Qu.:2021-03-25 08:39:23  
##                                        Max.   :2021-03-31 23:59:08  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-03-01 00:06:28   Length:228496      Length:228496     
##  1st Qu.:2021-03-10 11:04:40   Class :character   Class :character  
##  Median :2021-03-19 17:55:05   Mode  :character   Mode  :character  
##  Mean   :2021-03-17 23:45:00                                        
##  3rd Qu.:2021-03-25 08:54:12                                        
##  Max.   :2021-04-06 11:00:11                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:228496      Length:228496      Min.   :41.65   Min.   :-87.78  
##  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.   :42.07   Max.   :-87.53  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.64   Min.   :-88.07   Length:228496     
##  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.   :42.08   Max.   :-87.53                     
##  NA's   :167     NA's   :167
```

```
summary(divvy202104)
```

```
##    ride_id          rideable_type        started_at
##  Length:337230      Length:337230      Min.   :2021-04-01 00:03:18
##  Class :character   Class :character   1st Qu.:2021-04-07 12:07:56
##  Mode  :character   Mode  :character   Median :2021-04-15 22:37:04
##                                        Mean   :2021-04-15 22:47:10
##                                        3rd Qu.:2021-04-24 08:31:49
##                                        Max.   :2021-04-30 23:59:53
##
##     ended_at                   start_station_name start_station_id
##  Min.   :2021-04-01 00:14:29   Length:337230      Length:337230
##  1st Qu.:2021-04-07 12:31:51   Class :character   Class :character
##  Median :2021-04-15 23:00:10   Mode  :character   Mode  :character
##  Mean   :2021-04-15 23:11:18
##  3rd Qu.:2021-04-24 08:52:47
##  Max.   :2021-05-05 22:14:39
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:337230      Length:337230      Min.   :41.64   Min.   :-87.78
##  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.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   :41.59   Min.   :-87.85   Length:337230
##  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.   :42.15   Max.   :-87.52
##  NA's   :267     NA's   :267
```

```
summary(divvy202105)
```

```
##    ride_id           rideable_type        started_at                 
##  Length:531633      Length:531633      Min.   :2021-05-01 00:00:11  
##  Class :character   Class :character   1st Qu.:2021-05-10 17:40:50  
##  Mode  :character   Mode  :character   Median :2021-05-19 07:44:31  
##                                        Mean   :2021-05-17 19:52:32  
##                                        3rd Qu.:2021-05-24 19:32:22  
##                                        Max.   :2021-05-31 23:59:16  
##                                                                     
##     ended_at                  start_station_name start_station_id  
##  Min.   :2021-05-01 00:03:26   Length:531633      Length:531633     
##  1st Qu.:2021-05-10 17:57:59   Class :character   Class :character  
##  Median :2021-05-19 07:59:43   Mode  :character   Mode  :character  
##  Mean   :2021-05-17 20:18:34                                        
##  3rd Qu.:2021-05-24 19:57:20                                        
##  Max.   :2021-06-10 22:17:11                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:531633      Length:531633      Min.   :41.65   Min.   :-87.78  
##  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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.56   Min.   :-87.85   Length:531633     
##  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.   :42.09   Max.   :-87.52                     
##  NA's   :452     NA's   :452
```

```
summary(divvy202106)
```

```
##    ride_id           rideable_type        started_at                 
##  Length:729595      Length:729595      Min.   :2021-06-01 00:00:38  
##  Class :character   Class :character   1st Qu.:2021-06-08 16:03:57  
##  Mode  :character   Mode  :character   Median :2021-06-14 19:46:47  
##                                        Mean   :2021-06-15 09:48:47  
##                                        3rd Qu.:2021-06-21 19:10:47  
##                                        Max.   :2021-06-30 23:59:59  
##                                                                     
##     ended_at                    start_station_name start_station_id  
##  Min.   :2021-06-01 00:06:22   Length:729595      Length:729595     
##  1st Qu.:2021-06-08 16:23:54   Class :character   Class :character  
##  Median :2021-06-14 20:13:55   Mode  :character   Mode  :character  
##  Mean   :2021-06-15 10:14:52                                        
##  3rd Qu.:2021-06-21 19:31:59                                        
##  Max.   :2021-07-13 22:51:35                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:729595      Length:729595      Min.   :41.64   Min.   :-87.78  
##  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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.51   Min.   :-87.86   Length:729595     
##  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.   :42.08   Max.   :-87.49                     
##  NA's   :717     NA's   :717
```

```
summary(divvy202107)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:822410      Length:822410      Min.   :2021-07-01 00:00:22  
##  Class :character   Class :character   1st Qu.:2021-07-08 17:44:35  
##  Mode  :character   Mode  :character   Median :2021-07-17 13:58:37  
##                                        Mean   :2021-07-16 22:23:15  
##                                        3rd Qu.:2021-07-24 18:23:39  
##                                        Max.   :2021-07-31 23:59:58  
##                                                                     
##     ended_at                    start_station_name start_station_id  
##  Min.   :2021-07-01 00:04:51   Length:822410      Length:822410     
##  1st Qu.:2021-07-08 18:02:01   Class :character   Class :character  
##  Median :2021-07-17 14:28:04   Mode  :character   Mode  :character  
##  Mean   :2021-07-16 22:47:28                                        
##  3rd Qu.:2021-07-24 18:46:20                                        
##  Max.   :2021-08-12 17:45:41                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:822410      Length:822410      Min.   :41.65   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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.63   Min.   :-87.85   Length:822410     
##  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.   :42.15   Max.   :-87.49                     
##  NA's   :731     NA's   :731
```

```
summary(divvy202108)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:804352      Length:804352      Min.   :2021-08-01 00:00:04  
##  Class :character   Class :character   1st Qu.:2021-08-08 12:06:10  
##  Mode  :character   Mode  :character   Median :2021-08-16 07:57:11  
##                                        Mean   :2021-08-16 10:44:36  
##                                        3rd Qu.:2021-08-23 17:33:34  
##                                        Max.   :2021-08-31 23:59:35  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-08-01 00:03:11   Length:804352      Length:804352     
##  1st Qu.:2021-08-08 12:30:18   Class :character   Class :character  
##  Median :2021-08-16 08:12:14   Mode  :character   Mode  :character  
##  Mean   :2021-08-16 11:06:14                                        
##  3rd Qu.:2021-08-23 17:52:03                                        
##  Max.   :2021-09-01 17:37:35                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:804352      Length:804352      Min.   :41.65   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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.58   Min.   :-87.85   Length:804352     
##  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.   :42.15   Max.   :-87.51                     
##  NA's   :706     NA's   :706
```

```
summary(divvy202109)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:756147      Length:756147      Min.   :2021-09-01 00:00:06  
##  Class :character   Class :character   1st Qu.:2021-09-08 11:14:14  
##  Mode  :character   Mode  :character   Median :2021-09-15 16:43:37  
##                                        Mean   :2021-09-15 18:19:01  
##                                        3rd Qu.:2021-09-23 12:29:54  
##                                        Max.   :2021-09-30 23:59:48  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-09-01 00:00:41   Length:756147      Length:756147     
##  1st Qu.:2021-09-08 11:33:01   Class :character   Class :character  
##  Median :2021-09-15 17:01:16   Mode  :character   Mode  :character  
##  Mean   :2021-09-15 18:39:32                                        
##  3rd Qu.:2021-09-23 12:44:08                                        
##  Max.   :2021-10-01 22:55:35                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:756147      Length:756147      Min.   :41.65   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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.57   Min.   :-87.87   Length:756147     
##  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.   :-87.50                     
##  NA's   :595     NA's   :595
```

```
summary(divvy202110)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:631226      Length:631226      Min.   :2021-10-01 00:00:09  
##  Class :character   Class :character   1st Qu.:2021-10-08 12:25:58  
##  Mode  :character   Mode  :character   Median :2021-10-15 05:31:57  
##                                        Mean   :2021-10-15 08:38:27  
##                                        3rd Qu.:2021-10-21 19:25:00  
##                                        Max.   :2021-10-31 23:59:49  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-10-01 00:03:11   Length:631226      Length:631226     
##  1st Qu.:2021-10-08 12:46:34   Class :character   Class :character  
##  Median :2021-10-15 05:56:26   Mode  :character   Mode  :character  
##  Mean   :2021-10-15 08:57:32                                        
##  3rd Qu.:2021-10-21 19:37:25                                        
##  Max.   :2021-11-03 21:45:48                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:631226      Length:631226      Min.   :41.65   Min.   :-87.83  
##  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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.60   Min.   :-87.96   Length:631226     
##  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.   :42.13   Max.   :-87.52                     
##  NA's   :484     NA's   :484
```

```
summary(divvy202111)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:359978      Length:359978      Min.   :2021-11-01 00:00:14  
##  Class :character   Class :character   1st Qu.:2021-11-06 17:34:18  
##  Mode  :character   Mode  :character   Median :2021-11-12 08:32:12  
##                                        Mean   :2021-11-13 21:27:31  
##                                        3rd Qu.:2021-11-20 13:39:34  
##                                        Max.   :2021-11-30 23:59:56  
##                                                                     
##     ended_at                   start_station_name start_station_id  
##  Min.   :2021-11-01 00:04:06   Length:359978      Length:359978     
##  1st Qu.:2021-11-06 17:53:19   Class :character   Class :character  
##  Median :2021-11-12 08:46:55   Mode  :character   Mode  :character  
##  Mean   :2021-11-13 21:42:19                                        
##  3rd Qu.:2021-11-20 13:57:54                                        
##  Max.   :2021-12-02 06:41:33                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:359978      Length:359978      Min.   :41.65   Min.   :-87.84  
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66  
##  Mode  :character   Mode  :character   Median :41.89   Median :-87.64  
##                                        Mean   :41.89   Mean   :-87.65  
##                                        3rd Qu.:41.93   3rd Qu.:-87.63  
##                                        Max.   :42.07   Max.   :-87.53  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.39   Min.   :-88.97   Length:359978     
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character  
##  Median :41.89   Median :-87.64   Mode  :character  
##  Mean   :41.89   Mean   :-87.65                     
##  3rd Qu.:41.93   3rd Qu.:-87.63                     
##  Max.   :42.12   Max.   :-87.53                     
##  NA's   :191     NA's   :191
```

```
summary(divvy202112)
```

```
##    ride_id          rideable_type        started_at                 
##  Length:247540      Length:247540      Min.   :2021-12-01 00:00:01  
##  Class :character   Class :character   1st Qu.:2021-12-06 12:51:05  
##  Mode  :character   Mode  :character   Median :2021-12-13 13:04:54  
##                                        Mean   :2021-12-13 23:39:29  
##                                        3rd Qu.:2021-12-20 10:14:01  
##                                        Max.   :2021-12-31 23:59:48  
##                                                                     
##     ended_at                    start_station_name start_station_id  
##  Min.   :2021-12-01 00:02:40   Length:247540      Length:247540     
##  1st Qu.:2021-12-06 13:02:03   Class :character   Class :character  
##  Median :2021-12-13 13:18:39   Mode  :character   Mode  :character  
##  Mean   :2021-12-13 23:54:00                                        
##  3rd Qu.:2021-12-20 10:24:38                                        
##  Max.   :2022-01-03 17:32:18                                        
##                                                                     
##  end_station_name   end_station_id       start_lat       start_lng     
##  Length:247540      Length:247540      Min.   :41.64   Min.   :-87.84  
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.67  
##  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.   :42.07   Max.   :-87.52  
##                                                                        
##     end_lat         end_lng       member_casual     
##  Min.   :41.48   Min.   :-87.85   Length:247540     
##  1st Qu.:41.88   1st Qu.:-87.67   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.   :42.07   Max.   :-87.52                     
##  NA's   :144     NA's   :144
```

The data seems to be consistent and ready to combine

For further certainty we run the ‘compare_df_cols’ function

```
compare_df_cols(
  divvy202101,
  divvy202102,
  divvy202103,
  divvy202104,
  divvy202105,
  divvy202106,
  divvy202107,
  divvy202108,
  divvy202109,
  divvy202110,
  divvy202111,
  divvy202112, return = "mismatch")
```

```
##  [1] column_name divvy202101 divvy202102 divvy202103 divvy202104 divvy202105
##  [7] divvy202106 divvy202107 divvy202108 divvy202109 divvy202110 divvy202111
## [13] divvy202112
## <0 rows> (or 0-length row.names)
```

The data is ready to combine!

Combining the data sets into one This code chunk will combine the 12 individual data frames into one large data frame for analysis.

```
divvy2021 <- bind_rows(
                       divvy202101,
                       divvy202102,
                       divvy202103,
                       divvy202104,
                       divvy202105,
                       divvy202106,
                       divvy202107,
                       divvy202108,
                       divvy202109,
                       divvy202110,
                       divvy202111,
                       divvy202112 )
```

To view the data as a table we run the view function on the first month data avoiding the run problem for the larger data set.

```
view(divvy202101)
```

To view the structure of the larger data set.

```
str(divvy2021)
```

```
## spec_tbl_df [5,595,063 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5595063] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94
683FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr [1:5595063] "electric_bike" "electric_bike" "electric_bik
e" "electric_bike" ...
##  $ started_at        : POSIXct[1:5595063], format: "2021-01-23 16:14:19" "2021-01-27
18:43:08" ...
##  $ ended_at          : POSIXct[1:5595063], format: "2021-01-23 16:24:44" "2021-01-27
18:47:12" ...
##  $ start_station_name: chr [1:5595063] "California Ave & Cortez St" "California Ave &
Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" ...
##  $ start_station_id  : chr [1:5595063] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5595063] NA NA NA NA ...
##  $ end_station_id    : chr [1:5595063] NA NA NA NA ...
##  $ start_lat         : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5595063] "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>
```

```
dim(divvy202101)
```

```
## [1] 96834    13
```

For quick skim to check if there is misleading or missing data

```
skim(divvy2021)
```

Data summary

Name	divvy2021
Number of rows	5595063

Number of columns	13
_______________________	
Column type frequency:	
character	7
numeric	4
POSIXct	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	16	16	0	5595063	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	690809	0.88	3	53	0	847	0
start_station_id	690806	0.88	3	36	0	834	0
end_station_name	739170	0.87	10	53	0	844	0
end_station_id	739170	0.87	3	36	0	832	0
member_casual	0	1.00	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
start_lat	0	1	41.90	0.05	41.64	41.88	41.90	41.93	42.07	▁▁▇▇▁
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-87.52	▁▁▆▇▁
end_lat	4771	1	41.90	0.05	41.39	41.88	41.90	41.93	42.17	▁▁▁▇▁
end_lng	4771	1	-87.65	0.03	-88.97	-87.66	-87.64	-87.63	-87.49	▁▁▁▁▇

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2021-01-01 00:02:05	2021-12-31 23:59:48	2021-08-01 01:52:11	4677998
ended_at	0	1	2021-01-01 00:08:39	2022-01-03 17:32:18	2021-08-01 02:21:55	4671372

There is no need this four columns in our analysis so we exclude them

```
divvy2021 <- divvy2021 %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

Adding the required columns for analysis the date, month, day, and the day of the week to see the trends

```
divvy2021$date <- as.Date(divvy2021$started_at)
divvy2021$month <- format(as.Date(divvy2021$date), "%m")
divvy2021$day <- format(as.Date(divvy2021$started_at), "%d")
divvy2021$day_in_week <- format(as.Date(divvy2021$started_at), "%A")
```

Add the ride time column

```
divvy2021$ride_time_length <- difftime(divvy2021$ended_at, divvy2021$started_at)
```

Check for misleading time

```
skim(divvy2021$ride_time_length)
```

Data summary

Name	divvy2021$ride_time_lengt…
Number of rows	5595063
Number of columns	1
________________________	
Column type frequency:	
difftime	1
________________________	
Group variables	None

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
data	0	1	-3482 secs	3356649 secs	720 secs	25645

Delete the misleading or the false data (the negative times)

```
divvy2021 <- divvy2021[!(divvy2021$ride_time_length < 0),]
```

Checking again

```
skim(divvy2021$ride_time_length)
```

Data summary

Name	divvy2021$ride_time_lengt…
Number of rows	5594916
Number of columns	1
_______________________	
Column type frequency:	
difftime	1
________________________	
Group variables	None

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
data	0	1	0 secs	3356649 secs	720 secs	25540

check if there is other type than member and casual

```
table(divvy2021$member_casual)
```

```
##
##  casual  member
## 2528946 3065970
```

the mean or the average of the ride time for member and casual

```
aggregate(divvy2021$ride_time_length ~ divvy2021$member_casual, FUN = median)
```

```
##   divvy2021$member_casual divvy2021$ride_time_length
## 1                  casual                   958 secs
## 2                  member                   576 secs
```

the mean time of every day of the week with ride type

```
aggregate(divvy2021$ride_time_length ~ divvy2021$member_casual + divvy2021$day_in_week,
FUN = mean)
```

```
##    divvy2021$member_casual divvy2021$day_in_week divvy2021$ride_time_length
## 1                   casual                Friday             1820.9160 secs
## 2                   member                Friday              799.4950 secs
## 3                   casual                Monday             1912.5269 secs
## 4                   member                Monday              794.8517 secs
## 5                   casual              Saturday             2082.3740 secs
## 6                   member              Saturday              915.8742 secs
## 7                   casual                Sunday             2253.9949 secs
## 8                   member                Sunday              939.4763 secs
## 9                   casual              Thursday             1662.1955 secs
## 10                  member              Thursday              766.5710 secs
## 11                  casual               Tuesday             1678.3396 secs
## 12                  member               Tuesday              767.2874 secs
## 13                  casual             Wednesday             1659.4383 secs
## 14                  member             Wednesday              769.1496 secs
```

Finally Exporting the data with the summery for further analysis with tableau or other tool

```
write.csv(divvy2021, file = "/Volumes/Hard/R projects/Data.csv")
```

```
counts <- aggregate(divvy2021$ride_time_length ~ divvy2021$member_casual + divvy2021$day
_in_week, FUN = mean)
write.csv(counts, file = "/Volumes/Hard/R projects/counts.csv")
```

Exporting the time average for the ride type

```
type_median <- aggregate(divvy2021$ride_time_length ~ divvy2021$member_casual, FUN = med
ian)
write.csv(type_median, file = "/Volumes/Hard/R projects/member_casual_time_ave.csv")
```

by using group by we summerize the number of rides within each day grouped by the ride type

```
divvy2021 %>%
  group_by(member_casual, day_in_week) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n() ,average_duration = mean(ride_time_length)) %>%    #ca
lculates the number of rides and average duration
  arrange(member_casual, day_in_week)
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual day_in_week number_of_rides average_duration
##    <chr>         <chr>                 <int> <drtn>
##  1 casual        Friday               364075 1820.9160 secs
##  2 casual        Monday               286373 1912.5269 secs
##  3 casual        Saturday             557994 2082.3740 secs
##  4 casual        Sunday               481104 2253.9949 secs
##  5 casual        Thursday             286064 1662.1955 secs
##  6 casual        Tuesday              274388 1678.3396 secs
##  7 casual        Wednesday            278948 1659.4383 secs
##  8 member        Friday               446423  799.4950 secs
##  9 member        Monday               416204  794.8517 secs
## 10 member        Saturday             433041  915.8742 secs
## 11 member        Sunday               376117  939.4763 secs
## 12 member        Thursday             451520  766.5710 secs
## 13 member        Tuesday              465509  767.2874 secs
## 14 member        Wednesday            477156  769.1496 secs
```

to export

```
rides_number  <- divvy2021 %>%
  group_by(member_casual, day_in_week) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n() ,average_duration = mean(ride_time_length)) %>%
  arrange(member_casual, day_in_week)
```

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

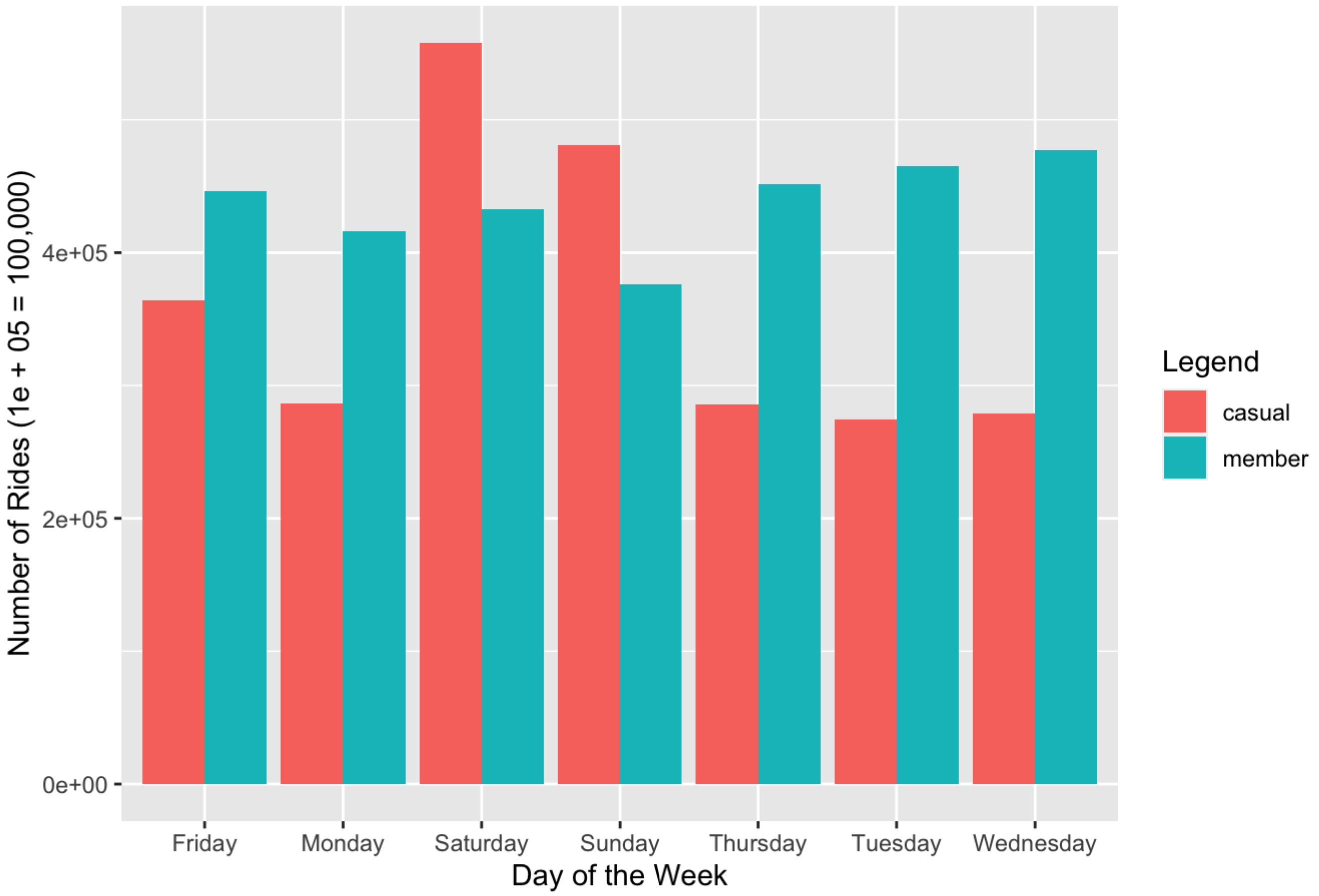
```
write.csv(rides_number, file = "/Volumes/Hard/R projects/rides_number_by_type.csv")
```

Analysis using R

The first graph to do with the R commands for the relation the kind or the Rider type and the day of the week

```
ggplot(data = divvy2021) + stat_count(mapping = aes(x= day_in_week, fill = member_casua
l), position = "dodge") +
  ggtitle("Figure 2: Number of Rides by Rider Type and Day of the Week") + ylab("Number
 of Rides (1e + 05 = 100,000)") +
  xlab("Day of the Week") + labs(fill = "Legend")
```

Figure 2: Number of Rides by Rider Type and Day of the Week



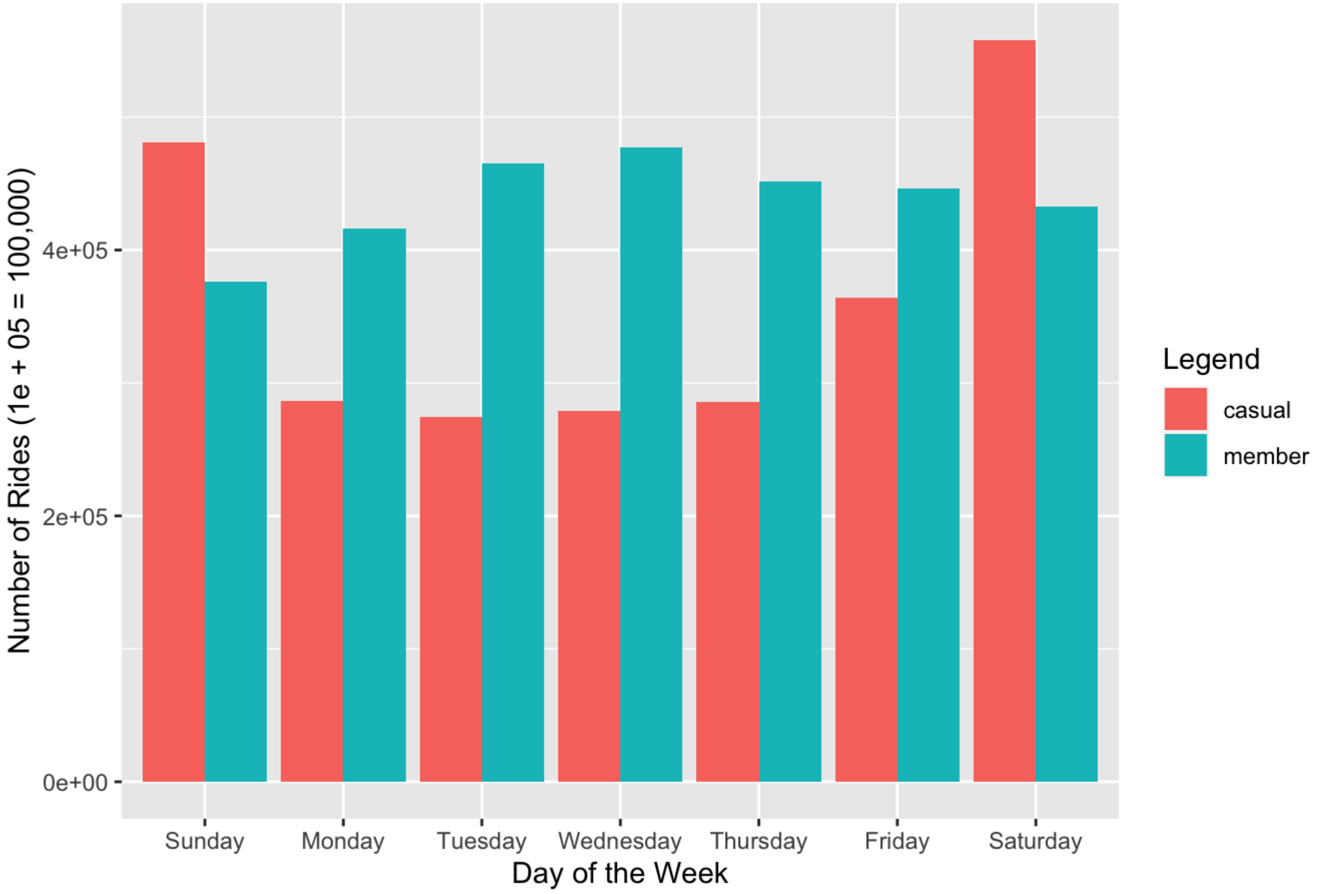
Noticing that the days are not in the right order

```
divvy2021$day_in_week <- ordered(divvy2021$day_in_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
```

we run the plot code again

```
ggplot(data = divvy2021) + stat_count(mapping = aes(x= day_in_week, fill = member_casual), position = "dodge") +
  ggtitle("Figure 2: Number of Rides by Rider Type and Day of the Week") + ylab("Number of Rides (1e + 05 = 100,000)") +
  xlab("Day of the Week") + labs(fill = "Legend")
```

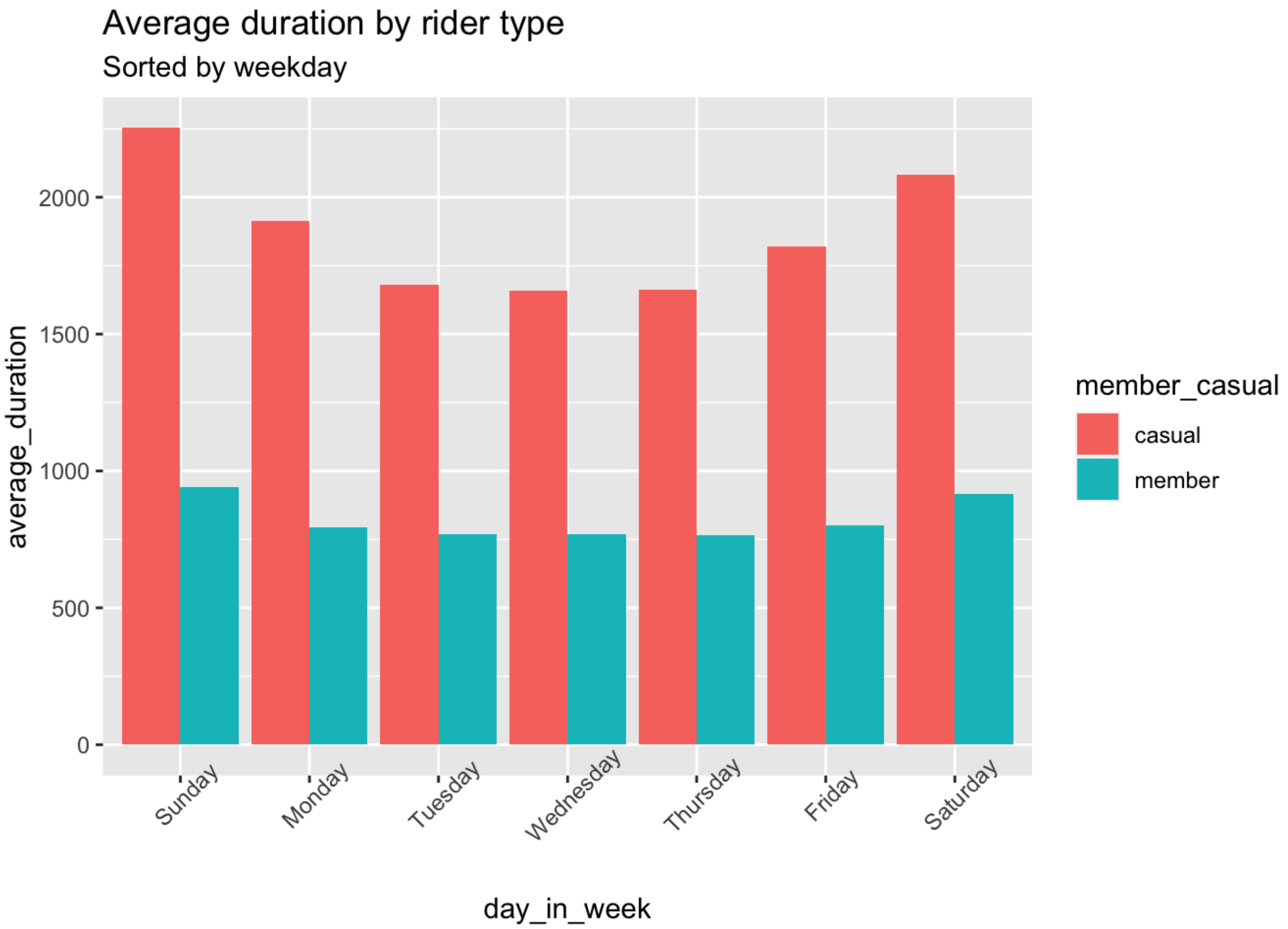
Figure 2: Number of Rides by Rider Type and Day of the Week



by number of rides

```
divvy2021 %>%
  group_by(member_casual, day_in_week) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n() ,average_duration = mean(ride_time_length)) %>%
  arrange(member_casual, day_in_week) %>%
  ggplot(aes(x = day_in_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Average duration by rider type", subtitle = "Sorted by weekday") +
  theme(axis.text.x = element_text(angle = 45))
```

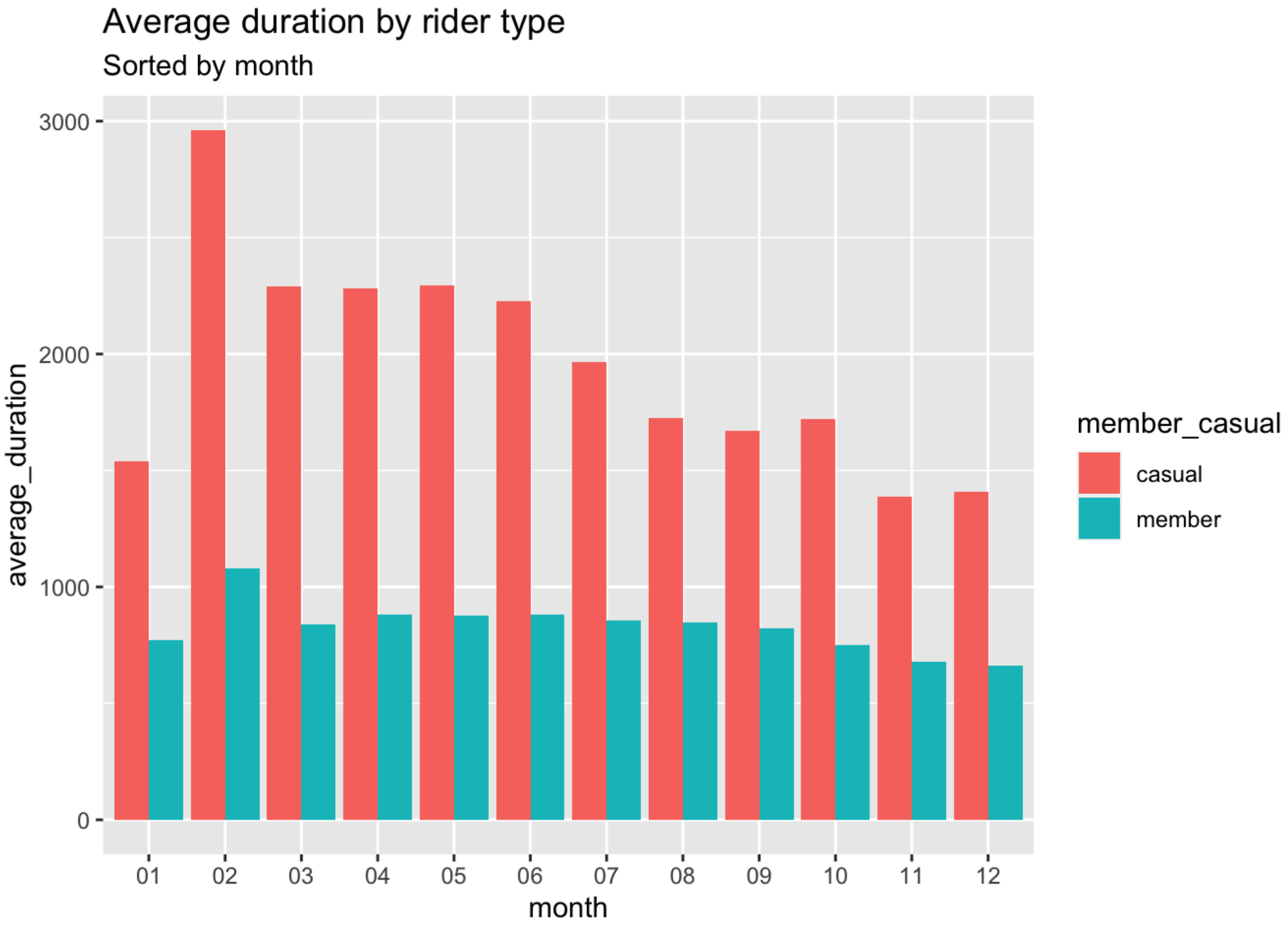
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
## Don't know how to automatically pick scale for object of type difftime.
## Defaulting to continuous.
```



By month

```
divvy2021 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n() ,average_duration = mean(ride_time_length)) %>%
  arrange(member_casual, month) %>%
ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Average duration by rider type", subtitle = "Sorted by month")
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
## Don't know how to automatically pick scale for object of type difftime.
## Defaulting to continuous.
```



Average duration by rider type

Sorted by month