Bike_Sharing_Analysis_Jul2020_Jun2021

Yan Houng

August 18, 2021 - October 5, 2021

Case Study: How Does a Bike-Share Navigate Speedy Success?

Scenario

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

Setting up my environment

First, we need to set the working directory to simplify calls to data.

Next, continue to set up the R environment by loading 'tidyverse', 'lubridate' and 'ggplot2' packages.

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

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

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

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

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

```
## The following object is masked from 'package:base':
##
##     date
```

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

STEP 1: COLLECT DATA

We need to upload Divvy datasets (csv files) so that the data can be used later.

```
jul_2020 <- read_csv("202007-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
aug_2020 <- read_csv("202008-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
sep_2020 <- read_csv("202009-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
oct_2020 <- read_csv("202010-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
nov_2020 <- read_csv("202011-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
dec_2020 <- read_csv("202012-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
jan_2021 <- read_csv("202101-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
feb_2021 <- read_csv("202102-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

```
mar_2021 <- read_csv("202103-divvy-tripdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

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

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

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

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

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

```
##
## -- Column specification --------------------------------------------------------
## 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()
## )
```

STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

After that, we need to compare the column names of each file. The column names need to be standardized before all the files are grouped together.

```
colnames(jul_2020)
```

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

```
colnames(aug_2020)
```

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

```
colnames(sep_2020)
```

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

```
colnames(oct_2020)
```

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

```
colnames(nov_2020)
```

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

```
colnames(dec_2020)
```

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

```
colnames(jan_2021)
```

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

```
colnames(feb_2021)
```

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

```
colnames(mar_2021)
```

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

```
colnames(apr_2021)
```

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

```
colnames(may_2021)
```

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

```
colnames(jun_2021)
```

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

All column names of each file are the same. So, we do not need to change the column names.

After that, we need to inspect the data frames and look for incongruencies.

```
str(jul_2020)
```

```
## spec_tbl_df [551,480 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:551480] "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77E
C1" "54AE594E20B35881" ...
##  $ rideable_type     : chr [1:551480] "docked_bike" "docked_bike" "docked_bike" "docked_bik
e" ...
##  $ started_at        : POSIXct[1:551480], format: "2020-07-09 15:22:02" "2020-07-24 23:56:3
0" ...
##  $ ended_at          : POSIXct[1:551480], format: "2020-07-09 15:25:52" "2020-07-25 00:20:1
7" ...
##  $ start_station_name: chr [1:551480] "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "Lak
e Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id  : num [1:551480] 180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : chr [1:551480] "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "Cl
ark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id    : num [1:551480] 291 461 156 94 301 289 140 31 191 142 ...
##  $ start_lat         : num [1:551480] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:551480] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:551480] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:551480] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:551480] "member" "member" "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()
##   .. )
```

```
str(aug_2020)
```

```
## spec_tbl_df [622,361 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:622361] "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B58
16" "C79FBBD412E578A7" ...
##  $ rideable_type     : chr [1:622361] "docked_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
##  $ started_at        : POSIXct[1:622361], format: "2020-08-20 18:08:14" "2020-08-27 18:46:0
4" ...
##  $ ended_at          : POSIXct[1:622361], format: "2020-08-20 18:17:51" "2020-08-27 19:54:5
1" ...
##  $ start_station_name: chr [1:622361] "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th
St" "Columbus Dr & Randolph St" "Daley Center Plaza" ...
##  $ 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" "St
ate St & Randolph St" "State St & Kinzie 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()
##   .. )
```

```
str(sep_2020)
```

```
## spec_tbl_df [532,958 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC77
8E" "57F6DC9A153DB98C" ...
##  $ rideable_type     : chr [1:532958] "electric_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-17 15:07:3
1" ...
##  $ ended_at          : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-17 15:07:4
5" ...
##  $ start_station_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway"
"W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...
##  $ 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 Oakdale Ave & N Broadway" "Montrose Harbor" ...
##  $ 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()
##   .. )
```

```
str(oct_2020)
```

```
## spec_tbl_df [388,653 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0
DF" "44A4AEE261B9E854" ...
##  $ rideable_type     : chr [1:388653] "electric_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:388653], format: "2020-10-31 19:39:43" "2020-10-31 23:50:0
8" ...
##  $ ended_at          : POSIXct[1:388653], format: "2020-10-31 19:57:12" "2020-11-01 00:04:1
6" ...
##  $ start_station_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Wave
land Ave" "Stony Island Ave & 67th St" "Clark St & Grace St" ...
##  $ 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"
"University Ave & 57th St" "Broadway & Sheridan Rd" ...
##  $ 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()
##   .. )
```

```
str(nov_2020)
```

```
## spec_tbl_df [259,716 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BD
C5" "E533E89C32080B9E" ...
##  $ rideable_type     : chr [1:259716] "electric_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:259716], format: "2020-11-01 13:36:00" "2020-11-01 10:03:2
6" ...
##  $ ended_at          : POSIXct[1:259716], format: "2020-11-01 13:45:40" "2020-11-01 10:14:4
5" ...
##  $ start_station_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illinois St"
"Lake Shore Dr & Monroe St" "Leavitt St & Chicago Ave" ...
##  $ 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 St & Polk St" "Stave St & Armitage Ave" ...
##  $ 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()
##   .. )
```

```
str(dec_2020)
```

```
## spec_tbl_df [131,573 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F
9A" "BE119628E44F871E" ...
##  $ rideable_type     : chr [1:131573] "classic_bike" "electric_bike" "electric_bike" "elect
ric_bike" ...
##  $ started_at        : POSIXct[1:131573], format: "2020-12-27 12:44:29" "2020-12-18 17:37:1
5" ...
##  $ ended_at          : POSIXct[1:131573], format: "2020-12-27 12:55:06" "2020-12-18 17:44:1
9" ...
##  $ 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()
##   .. )
```

```
str(jan_2021)
```

```
## spec_tbl_df [96,834 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F2
7" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr [1:96834] "electric_bike" "electric_bike" "electric_bike" "elect
ric_bike" ...
##  $ started_at        : POSIXct[1:96834], format: "2021-01-23 16:14:19" "2021-01-27 18:43:0
8" ...
##  $ ended_at          : POSIXct[1:96834], format: "2021-01-23 16:24:44" "2021-01-27 18:47:1
2" ...
##  $ start_station_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez
St" "California Ave & Cortez St" "California Ave & Cortez St" ...
##  $ 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()
##   .. )
```

```
str(feb_2021)
```

```
## spec_tbl_df [49,622 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB9
1" "B32D3199F1C2E75B" ...
##  $ 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 17:52:3
8" ...
##  $ ended_at          : POSIXct[1:49622], format: "2021-02-12 16:21:43" "2021-02-14 18:12:0
9" ...
##  $ start_station_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave"
"Clark St & Lake St" "Wood St & Chicago Ave" ...
##  $ start_station_id  : chr [1:49622] "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard S
t" "State St & Randolph St" "Honore St & Division St" ...
##  $ 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()
##   .. )
```

```
str(mar_2021)
```

```
## spec_tbl_df [228,496 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A2
84" "994D05AA75A168F2" ...
##  $ 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 01:26:2
8" ...
##  $ ended_at          : POSIXct[1:228496], format: "2021-03-16 08:36:34" "2021-03-28 01:36:5
5" ...
##  $ start_station_name: chr [1:228496] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armit
age Ave" "Shields Ave & 28th Pl" "Winthrop Ave & Lawrence Ave" ...
##  $ start_station_id  : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomin
gdale Ave" "Halsted St & 35th St" "Broadway & Sheridan Rd" ...
##  $ 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()
##   .. )
```

```
str(apr_2021)
```

```
## spec_tbl_df [337,230 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:337230] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80B
AD" "1887262AD101C604" ...
##  $ rideable_type     : chr [1:337230] "classic_bike" "docked_bike" "docked_bike" "classic_b
ike" ...
##  $ started_at        : POSIXct[1:337230], format: "2021-04-12 18:25:36" "2021-04-27 17:27:1
1" ...
##  $ ended_at          : POSIXct[1:337230], format: "2021-04-12 18:56:55" "2021-04-27 18:31:2
9" ...
##  $ start_station_name: chr [1:337230] "State St & Pearson St" "Dorchester Ave & 49th St" "L
oomis Blvd & 84th St" "Honore St & Division St" ...
##  $ start_station_id  : chr [1:337230] "TA1307000061" "KA1503000069" "20121" "TA1305000034"
...
##  $ end_station_name  : chr [1:337230] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th
St" "Loomis Blvd & 84th St" "Southport Ave & Waveland Ave" ...
##  $ 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()
##   .. )
```

```
str(may_2021)
```

```
## spec_tbl_df [531,633 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EF
C2" "7881AC6D39110C60" ...
##  $ rideable_type     : chr [1:531633] "electric_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:531633], format: "2021-05-30 11:58:15" "2021-05-30 11:29:1
4" ...
##  $ ended_at          : POSIXct[1:531633], format: "2021-05-30 12:10:39" "2021-05-30 12:14:0
9" ...
##  $ 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()
##   .. )
```

```
str(jun_2021)
```

```
## spec_tbl_df [729,595 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:729595] "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045D
F2" "B03C0FE48C412214" ...
##  $ rideable_type     : chr [1:729595] "electric_bike" "electric_bike" "electric_bike" "elec
tric_bike" ...
##  $ started_at        : POSIXct[1:729595], format: "2021-06-13 14:31:28" "2021-06-04 11:18:0
2" ...
##  $ ended_at          : POSIXct[1:729595], format: "2021-06-13 14:34:11" "2021-06-04 11:24:1
9" ...
##  $ 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()
##   .. )
```

The start_station_id an end_station_id columns in jul_2020, aug_2020, sep_2020, oct_2020, nov_2020 tables consist of data stored as numerical data type instead of character data type.

So we need to convert start_station_id and end_station_id to character data type so that they can stack correctly.

```
jul_2020 <-  mutate(jul_2020, start_station_id = as.character(start_station_id)
                    ,end_station_id = as.character(end_station_id))
aug_2020 <-  mutate(aug_2020, start_station_id = as.character(start_station_id)
                    ,end_station_id = as.character(end_station_id))
sep_2020 <-  mutate(sep_2020, start_station_id = as.character(start_station_id)
                    ,end_station_id = as.character(end_station_id))
oct_2020 <-  mutate(oct_2020, start_station_id = as.character(start_station_id)
                    ,end_station_id = as.character(end_station_id))
nov_2020 <-  mutate(nov_2020, start_station_id = as.character(start_station_id)
                    ,end_station_id = as.character(end_station_id))
```

Lastly, we can stack individual month's data frames into one big data frame.

```
all_trips <- bind_rows(jul_2020, aug_2020, sep_2020, oct_2020, nov_2020, dec_2020, jan_2021, f
eb_2021, mar_2021, apr_2021, may_2021, jun_2021)
```

STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

Then, we will inspect the new table “all_trips” that has been created.

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

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

```
nrow(all_trips)  #How many rows are in data frame?
```

```
## [1] 4460151
```

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

```
## [1] 4460151      13
```

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

```
## # A tibble: 6 x 13
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 762198~ docked_bike   2020-07-09 15:22:02 2020-07-09 15:25:52 Ritchie Ct & Ba~
## 2 BEC9C9~ docked_bike   2020-07-24 23:56:30 2020-07-25 00:20:17 Halsted St & Ro~
## 3 D2FD8E~ docked_bike   2020-07-08 19:49:07 2020-07-08 19:56:22 Lake Shore Dr &~
## 4 54AE59~ docked_bike   2020-07-17 19:06:42 2020-07-17 19:27:38 LaSalle St & Il~
## 5 54025F~ docked_bike   2020-07-04 10:39:57 2020-07-04 10:45:05 Lake Shore Dr &~
## 6 65636B~ docked_bike   2020-07-28 16:33:03 2020-07-28 16:49:10 Fairbanks St & ~
## # ... with 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>
```

```
str(all_trips)  #See list of columns and data types (numeric, character, etc)
```

```
## spec_tbl_df [4,460,151 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4460151] "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77
EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr [1:4460151] "docked_bike" "docked_bike" "docked_bike" "docked_bi
ke" ...
##  $ started_at        : POSIXct[1:4460151], format: "2020-07-09 15:22:02" "2020-07-24 23:56:
30" ...
##  $ ended_at          : POSIXct[1:4460151], format: "2020-07-09 15:25:52" "2020-07-25 00:20:
17" ...
##  $ start_station_name: chr [1:4460151] "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "La
ke Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id  : chr [1:4460151] "180" "299" "329" "181" ...
##  $ end_station_name  : chr [1:4460151] "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "C
lark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id    : chr [1:4460151] "291" "461" "156" "94" ...
##  $ start_lat         : num [1:4460151] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:4460151] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:4460151] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:4460151] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:4460151] "member" "member" "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()
##   .. )
```

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

```
##    ride_id          rideable_type        started_at
##  Length:4460151     Length:4460151     Min.   :2020-07-01 00:00:14
##  Class :character   Class :character   1st Qu.:2020-08-29 14:06:10
##  Mode  :character   Mode  :character   Median :2020-11-11 11:51:43
##                                        Mean   :2020-12-25 05:10:09
##                                        3rd Qu.:2021-05-11 20:43:11
##                                        Max.   :2021-06-30 23:59:59
##
##     ended_at                    start_station_name start_station_id
##  Min.   :2020-07-01 00:03:01   Length:4460151     Length:4460151
##  1st Qu.:2020-08-29 14:35:14   Class :character   Class :character
##  Median :2020-11-11 12:08:21   Mode  :character   Mode  :character
##  Mean   :2020-12-25 05:33:58
##  3rd Qu.:2021-05-11 21:00:43
##  Max.   :2021-07-13 22:51:35
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:4460151     Length:4460151     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:4460151
##  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   :5286    NA's   :5286
```

There are a few problems we will need to fix: (1) The data can only be aggregated at the ride-level, which is too granular. We will want to add some additional columns of data – such as day, month, year – that provide additional opportunities to aggregate the data. (2) We will want to add a calculated field for length of ride as "ride_length" to the entire data frame for consistency. (3) There are some rides where trip duration shows up as negative, including several hundred rides where Divvy took bikes out of circulation for Quality Control reasons. We will want to delete these rides. (4) We will add some calculated fields for the distance between start station and end station to the entire data frame in order to find out if the riders return the shared bike on the same station as the start station.

We add in columns that list the date, month, day, and year of each ride. This will allow us to aggregate ride data for each month, day, or year … before completing these operations we could only aggregate at the ride level https://www.statmethods.net/input/dates.html (https://www.statmethods.net/input/dates.html) more on date formats in R found at that link

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

Next, we will add a "ride_length" calculation to all_trips (in seconds). https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html (https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html)

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

After that, we inspect the structure of the columns.

```
str(all_trips)
```

```
## spec_tbl_df [4,460,151 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4460151] "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77
EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr [1:4460151] "docked_bike" "docked_bike" "docked_bike" "docked_bi
ke" ...
##  $ started_at        : POSIXct[1:4460151], format: "2020-07-09 15:22:02" "2020-07-24 23:56:
30" ...
##  $ ended_at          : POSIXct[1:4460151], format: "2020-07-09 15:25:52" "2020-07-25 00:20:
17" ...
##  $ start_station_name: chr [1:4460151] "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "La
ke Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id  : chr [1:4460151] "180" "299" "329" "181" ...
##  $ end_station_name  : chr [1:4460151] "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "C
lark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id    : chr [1:4460151] "291" "461" "156" "94" ...
##  $ start_lat         : num [1:4460151] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:4460151] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:4460151] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:4460151] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:4460151] "member" "member" "casual" "casual" ...
##  $ date              : Date[1:4460151], format: "2020-07-09" "2020-07-24" ...
##  $ month             : chr [1:4460151] "07" "07" "07" "07" ...
##  $ day               : chr [1:4460151] "09" "24" "08" "17" ...
##  $ year              : chr [1:4460151] "2020" "2020" "2020" "2020" ...
##  $ day_of_week       : chr [1:4460151] "Thursday" "Friday" "Wednesday" "Friday" ...
##  $ hour              : chr [1:4460151] "15" "23" "19" "19" ...
##  $ ride_length       : 'difftime' num [1:4460151] 230 1427 435 1256 ...
##   ..- attr(*, "units")= chr "secs"
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_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()
##   .. )
```

Then, we will convert “ride_length” from Factor to numeric so we can run calculations on the data.

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

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

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

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

Next, we will also add in all latitude, longitude and distance differences.

```
all_trips$lat_diff <- all_trips$end_lat - all_trips$start_lat
all_trips$lng_diff <- all_trips$end_lng - all_trips$start_lng
all_trips$lat_diff_m <- all_trips$lat_diff/0.00001*1.11
all_trips$lng_diff_m <- all_trips$lng_diff/0.00001*1.11
all_trips$distance_diff_m <- sqrt(all_trips$lat_diff_m^2 + all_trips$lng_diff_m^2)
all_trips$start_end_same_station <- ifelse(all_trips$start_station_name == all_trips$end_stati
on_name | all_trips$distance_diff_m <= 50, 1, 0)
```

I have decided that if the distance difference of a start position and an ending position is more than 50 m, then it is not considered as same location or station except the case of the start station is the same name as the end station.

Then, we will check for all the bad data first.

```
all_trips_negative_ride_length <- all_trips[(all_trips$ride_length<0),]
```

There are 9872 observations for negative ride length. This is equivalent to 0.221% of total observations.

```
all_trips_na_lat_lng <- all_trips[(all_trips$start_lat == "" | is.na(all_trips$start_lat) | al
l_trips$start_lng == "" | is.na(all_trips$start_lng) | all_trips$end_lat == "" | is.na(all_tri
ps$end_lat) | all_trips$end_lng == "" | is.na(all_trips$end_lng)),]
```

There are 5286 observations for lat and lng column is na. This is equivalent to 0.119% of total observations.

```
all_trips_na_station <- all_trips[(all_trips$start_station_name == "" | is.na(all_trips$start_
station_name) | all_trips$end_station_name == "" | is.na(all_trips$end_station_name)),]
```

There are 433776 observations for station name is na. This is equivalent to 9.726% of total observations. These observations are too much to be removed from the data. It is more than 1 month average data.

Remove “bad” data

The data frame includes a few thousand entries where ride_length was negative and missing value for start position and ending position. We will create a new version of the data frame (v2) since data is being removed. https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/ (https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/)

```
all_trips_v2 <- all_trips[!(all_trips$ride_length<0),] #remove the rows where the ride length
 is negative.
all_trips_v2 <- all_trips_v2[!(all_trips_v2$start_lat == "" | is.na(all_trips_v2$start_lat) |
 all_trips_v2$start_lng == "" | is.na(all_trips_v2$start_lng) | all_trips_v2$end_lat == "" | i
s.na(all_trips_v2$end_lat) | all_trips_v2$end_lng == "" | is.na(all_trips_v2$end_lng)),]
all_trips_v2 <- all_trips_v2[!((all_trips_v2$start_station_name == "WATSON TESTING - DIVVY" &
 !is.na(all_trips_v2$start_station_name)) | (all_trips_v2$start_station_name == "HUBBARD ST BI
KE CHECKING (LBS-WH-TEST)" & !is.na(all_trips_v2$start_station_name))),]
```

Stations “WATSON TESTING - DIVVY” & “HUBBARD ST BIKE CHECKING (LBS-WH-TEST)” are for checking and testing. Thus they are being removed from the data.

```
str(all_trips_v2)
```

```
## tibble [4,441,929 x 26] (S3: tbl_df/tbl/data.frame)
##  $ ride_id               : chr [1:4441929] "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA43
2C77EC1" "54AE594E20B35881" ...
##  $ rideable_type         : chr [1:4441929] "docked_bike" "docked_bike" "docked_bike" "docke
d_bike" ...
##  $ started_at            : POSIXct[1:4441929], format: "2020-07-09 15:22:02" "2020-07-24 2
3:56:30" ...
##  $ ended_at              : POSIXct[1:4441929], format: "2020-07-09 15:25:52" "2020-07-25 0
0:20:17" ...
##  $ start_station_name    : chr [1:4441929] "Ritchie Ct & Banks St" "Halsted St & Roscoe St"
"Lake Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id      : chr [1:4441929] "180" "299" "329" "181" ...
##  $ end_station_name      : chr [1:4441929] "Wells St & Evergreen Ave" "Broadway & Ridge Av
e" "Clark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id        : chr [1:4441929] "291" "461" "156" "94" ...
##  $ start_lat             : num [1:4441929] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng             : num [1:4441929] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat               : num [1:4441929] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng               : num [1:4441929] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual         : chr [1:4441929] "member" "member" "casual" "casual" ...
##  $ date                  : Date[1:4441929], format: "2020-07-09" "2020-07-24" ...
##  $ month                 : chr [1:4441929] "07" "07" "07" "07" ...
##  $ day                   : chr [1:4441929] "09" "24" "08" "17" ...
##  $ year                  : chr [1:4441929] "2020" "2020" "2020" "2020" ...
##  $ day_of_week           : chr [1:4441929] "Thursday" "Friday" "Wednesday" "Friday" ...
##  $ hour                  : chr [1:4441929] "15" "23" "19" "19" ...
##  $ ride_length           : num [1:4441929] 230 1427 435 1256 308 ...
##  $ lat_diff              : num [1:4441929] -0.000142 0.040375 0.003909 0.027544 -0.003729
...
##  $ lng_diff              : num [1:4441929] -0.00861 -0.01132 -0.01111 -0.00458 -0.0047 ...
##  $ lat_diff_m            : num [1:4441929] -15.8 4481.6 433.9 3057.4 -413.9 ...
##  $ lng_diff_m            : num [1:4441929] -956 -1257 -1233 -509 -521 ...
##  $ distance_diff_m       : num [1:4441929] 956 4655 1307 3099 666 ...
##  $ start_end_same_station: num [1:4441929] 0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(all_trips_v2)
```

```
##    ride_id          rideable_type        started_at
##  Length:4441929     Length:4441929     Min.   :2020-07-01 00:00:14
##  Class :character   Class :character   1st Qu.:2020-08-29 15:15:40
##  Mode  :character   Mode  :character   Median :2020-11-11 23:13:33
##                                        Mean   :2020-12-25 13:15:19
##                                        3rd Qu.:2021-05-12 10:45:07
##                                        Max.   :2021-06-30 23:59:59
##
##     ended_at                   start_station_name start_station_id
##  Min.   :2020-07-01 00:03:01   Length:4441929     Length:4441929
##  1st Qu.:2020-08-29 15:46:51   Class :character   Class :character
##  Median :2020-11-11 23:36:56   Mode  :character   Mode  :character
##  Mean   :2020-12-25 13:41:02
##  3rd Qu.:2021-05-12 11:03:37
##  Max.   :2021-07-13 22:51:35
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:4441929     Length:4441929     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           date
##  Min.   :41.51   Min.   :-88.07   Length:4441929     Min.   :2020-07-01
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character   1st Qu.:2020-08-29
##  Median :41.90   Median :-87.64   Mode  :character   Median :2020-11-11
##  Mean   :41.90   Mean   :-87.64                      Mean   :2020-12-24
##  3rd Qu.:41.93   3rd Qu.:-87.63                      3rd Qu.:2021-05-12
##  Max.   :42.16   Max.   :-87.44                      Max.   :2021-06-30
##
##     month               day                year           day_of_week
##  Length:4441929     Length:4441929     Length:4441929     Length:4441929
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      hour            ride_length         lat_diff             lng_diff
##  Length:4441929     Min.   :      0   Min.   :-0.4163532   Min.   :-0.3532472
##  Class :character   1st Qu.:    453   1st Qu.:-0.0097240   1st Qu.:-0.0096250
##  Mode  :character   Median :    820   Median : 0.0000000   Median : 0.0000000
##                     Mean   :   1543   Mean   : 0.0003019   Mean   :-0.0002616
##                     3rd Qu.:   1509   3rd Qu.: 0.0100000   3rd Qu.: 0.0091030
##                     Max.   :3356649   Max.   : 0.2739608   Max.   : 0.1821418
##
##    lat_diff_m          lng_diff_m        distance_diff_m start_end_same_station
##  Min.   :-46215.21   Min.   :-39210.44   Min.   :    0   Min.   :0.0
##  1st Qu.: -1079.36   1st Qu.: -1068.38   1st Qu.: 1024   1st Qu.:0.0
##  Median :     0.00   Median :     0.00   Median : 1928   Median :0.0
##  Mean   :    33.51   Mean   :   -29.04   Mean   : 2491   Mean   :0.1
##  3rd Qu.:  1110.00   3rd Qu.:  1010.43   3rd Qu.: 3391   3rd Qu.:0.0
##  Max.   : 30409.65   Max.   : 20217.74   Max.   :49886   Max.   :1.0
##                                                          NA's   :388008
```

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

So, we can conduct the descriptive analysis on ride_length (all figures in seconds).

```
mean(all_trips_v2$ride_length) #straight average (total ride length / rides)
```

```
## [1] 1543.071
```

```
median(all_trips_v2$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 820
```

```
max(all_trips_v2$ride_length) #Longest ride
```

```
## [1] 3356649
```

```
min(all_trips_v2$ride_length) #shortest ride
```

```
## [1] 0
```

We can condense the four lines above to one line using summary() on the specific attribute.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     453     820    1543    1509 3356649
```

Then, lets compare members and casual users.

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

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                2396.6310
## 2                     member                 892.9672
```

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

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                     1155
## 2                     member                      648
```

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

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                  3356649
## 2                     member                  2005282
```

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

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

We can see the average ride time by each day for members vs casual users

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

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                  Friday                2293.9896
## 2                      member                  Friday                 880.1806
## 3                      casual                  Monday                2315.4466
## 4                      member                  Monday                 859.7733
## 5                      casual                Saturday                2552.2271
## 6                      member                Saturday                 981.4182
## 7                      casual                  Sunday                2755.2208
## 8                      member                  Sunday                1009.7902
## 9                      casual                Thursday                2168.3222
## 10                     member                Thursday                 841.3316
## 11                     casual                 Tuesday                2119.2009
## 12                     member                 Tuesday                 843.2790
## 13                     casual               Wednesday                2168.1132
## 14                     member               Wednesday                 846.2170
```

The days of the week have been out of order. We can fix that with the following commands.

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

Now, let's compare the average ride time by each day for members vs casual users.

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

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                  Sunday                2755.2208
## 2                      member                  Sunday                1009.7902
## 3                      casual                  Monday                2315.4466
## 4                      member                  Monday                 859.7733
## 5                      casual                 Tuesday                2119.2009
## 6                      member                 Tuesday                 843.2790
## 7                      casual               Wednesday                2168.1132
## 8                      member               Wednesday                 846.2170
## 9                      casual                Thursday                2168.3222
## 10                     member                Thursday                 841.3316
## 11                     casual                  Friday                2293.9896
## 12                     member                  Friday                 880.1806
## 13                     casual                Saturday                2552.2271
## 14                     member                Saturday                 981.4182
```

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

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                  1920453
## 2                     member                  2521476
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$rideable_type + all_trips_v2$member_casual,
FUN = length)
```

```
##   all_trips_v2$rideable_type all_trips_v2$member_casual
## 1               classic_bike                     casual
## 2                docked_bike                     casual
## 3              electric_bike                     casual
## 4               classic_bike                     member
## 5                docked_bike                     member
## 6              electric_bike                     member
##   all_trips_v2$ride_length
## 1                   452496
## 2                   966589
## 3                   501368
## 4                   824102
## 5                  1070654
## 6                   626720
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FU
N = length)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Sunday                   365379
## 2                      member                   Sunday                   322230
## 3                      casual                   Monday                   208267
## 4                      member                   Monday                   336495
## 5                      casual                  Tuesday                   203041
## 6                      member                  Tuesday                   363235
## 7                      casual                Wednesday                   214494
## 8                      member                Wednesday                   387882
## 9                      casual                 Thursday                   211345
## 10                     member                 Thursday                   362671
## 11                     casual                   Friday                   279675
## 12                     member                   Friday                   373498
## 13                     casual                 Saturday                   438252
## 14                     member                 Saturday                   375465
```

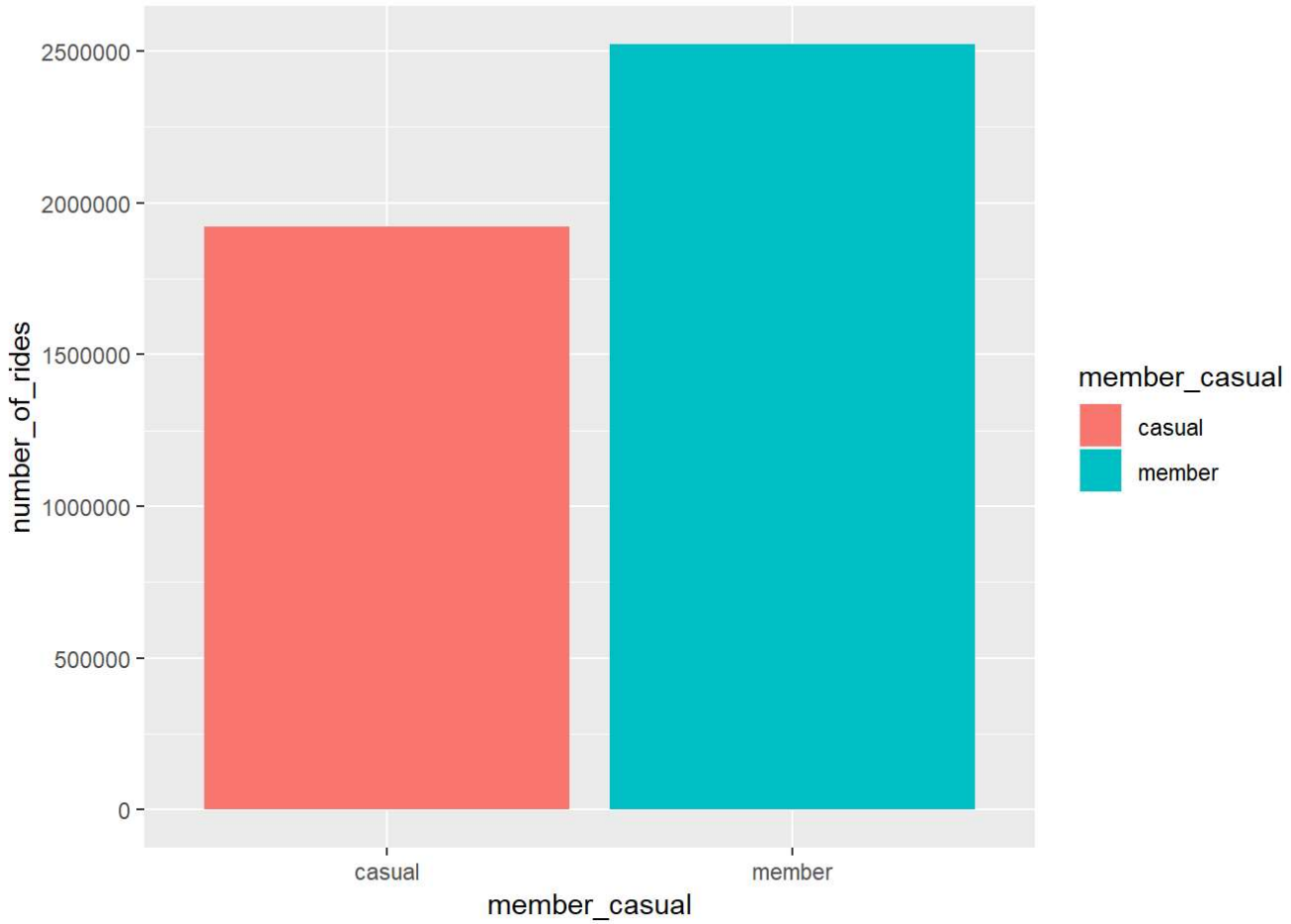
```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$start_end_same_
station, FUN = length)
```

```
##   all_trips_v2$member_casual all_trips_v2$start_end_same_station
## 1                     casual                                   0
## 2                     member                                   0
## 3                     casual                                   1
## 4                     member                                   1
##   all_trips_v2$ride_length
## 1                  1450163
## 2                  2172391
## 3                   290881
## 4                   140486
```

Data Visualization

Let us see the number of rides by rider type.

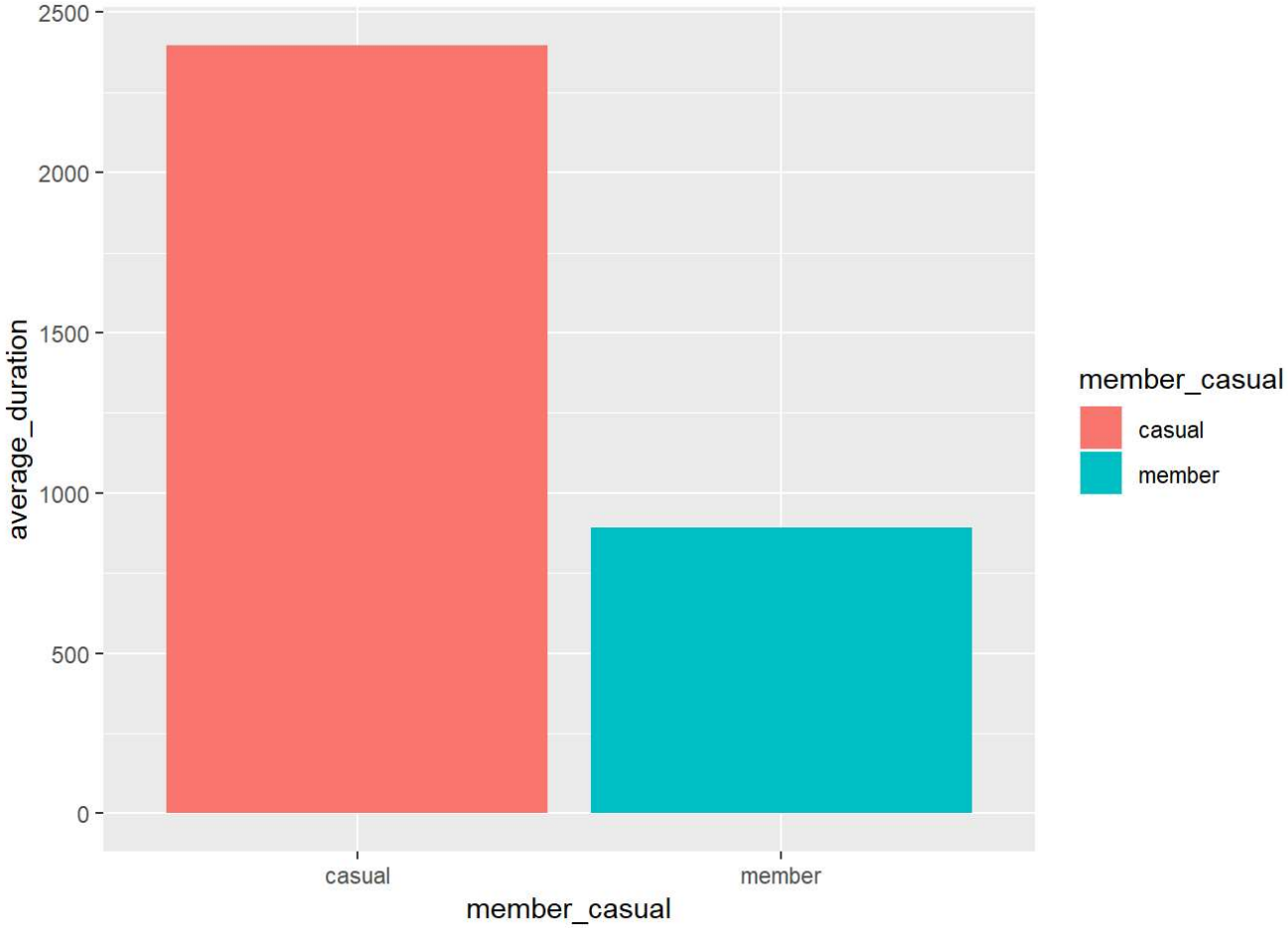
```
all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual)  %>%
  ggplot(aes(x = member_casual, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```



Based on the number of rides, the number of annual member is hgiher than the number of casual riders. There are total 2521476 rides for annual member while for the casual rider, there are total 1920453 rides.

Let us see the average riding duration by rider type.

```
all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual)  %>%
  ggplot(aes(x = member_casual, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

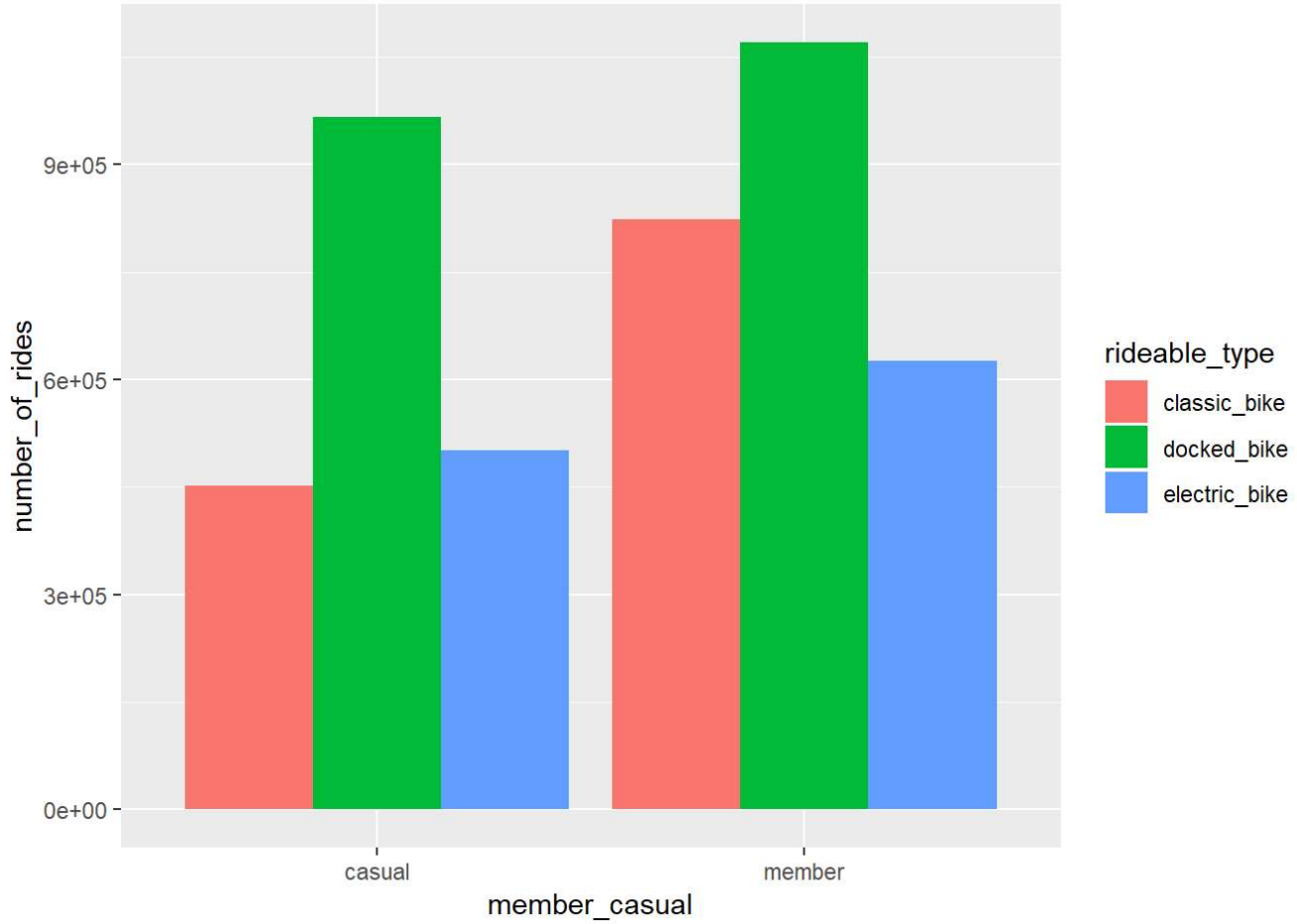


We can see that the average riding duration of casual riders is higher than annual member's. The average riding duration of casual riders is about 2.5 times of the annual member's average riding time. This mean that casual riders will usually rent bike to cycle around for quite a long distance while annual members only rent bicycle to travel to a specific location, for example to workplace for working.

Let us see the number of rides by bicycle type (classic bike, docked bike or electric bike).

```
all_trips_v2 %>%
  group_by(member_casual, rideable_type) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual)  %>%
  ggplot(aes(x = member_casual, y = number_of_rides, fill = rideable_type)) +
  geom_col(position = "dodge")
```

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

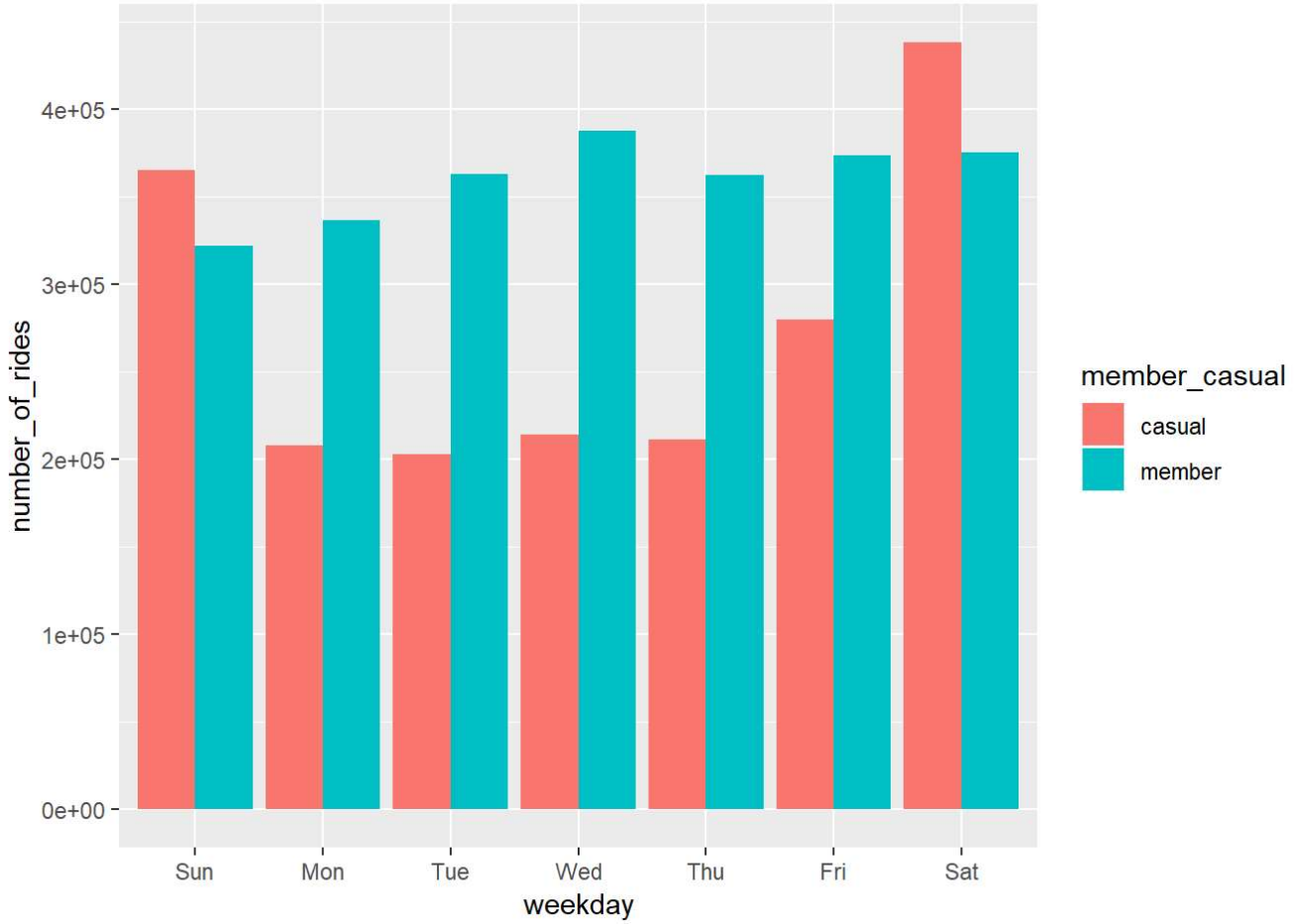


Docked bike is the most popular bicycle type among the 3 bicycle types.

Let us visualize the number of rides by rider type for different days of the week.

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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

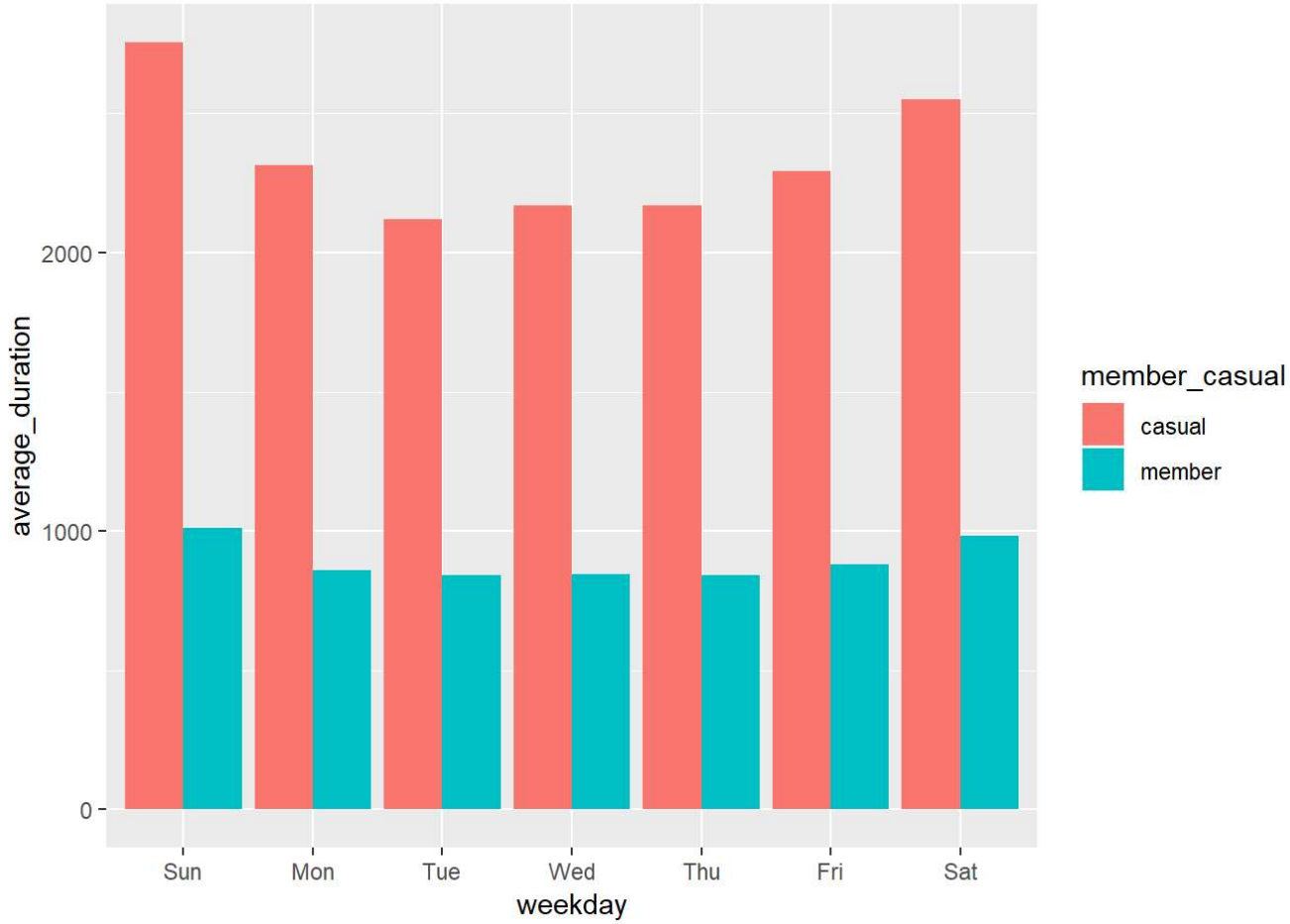


We can see that the number of casual riders is lesser than the number of annual members on the weekdays while it is the opposite for the weekends.

But if we visualize for average duration on different days of the week, we will see something different.

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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

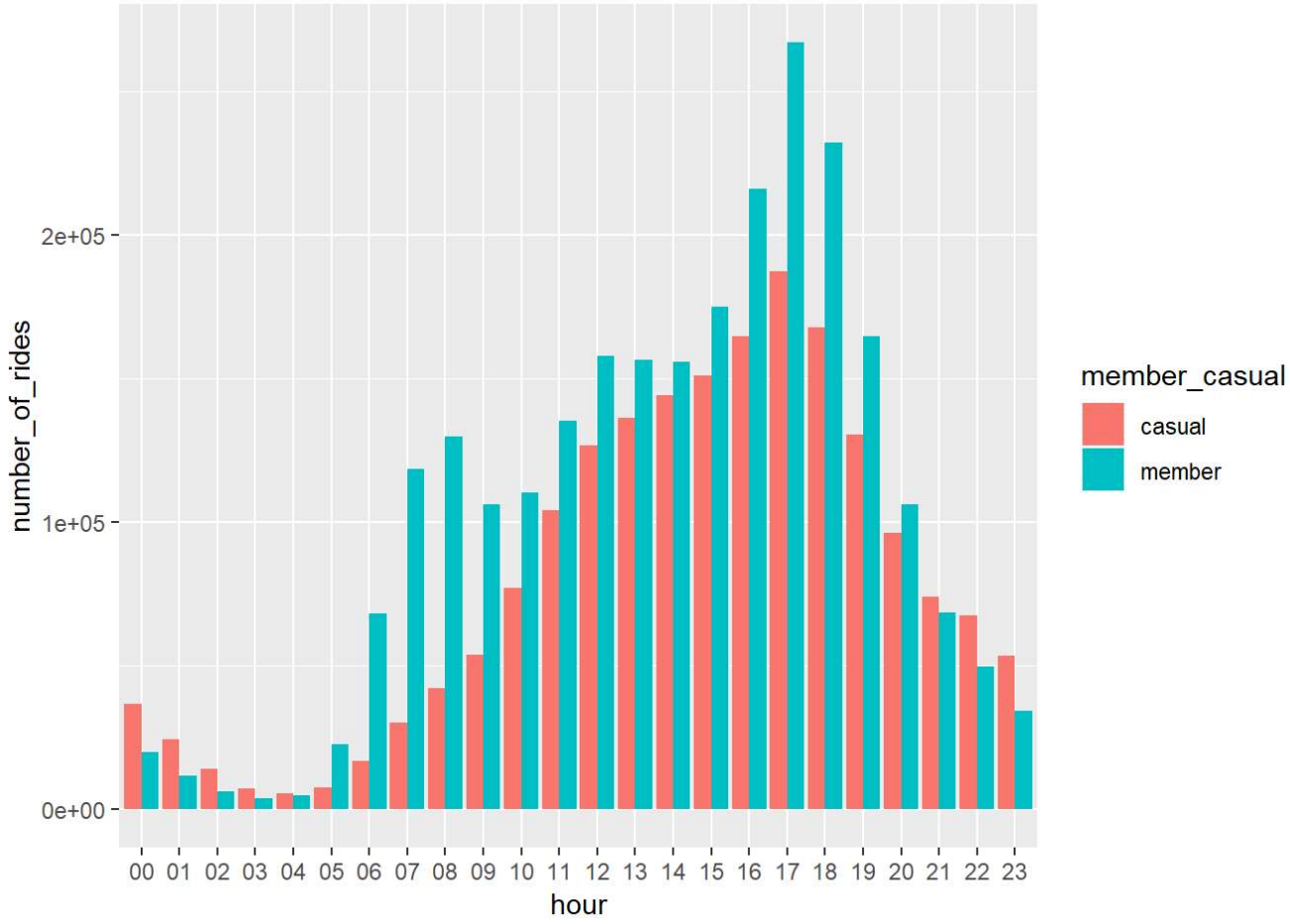


The average riding duration of casual rider is higher than annual member on everyday. The average riding duration is the highest on weekends.

Next, let's create a visualization for the number of rides by different hour in 1 day. This is the total number of rides sum up for everyday.

```
all_trips_v2 %>%
  group_by(member_casual, hour) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, hour)  %>%
  ggplot(aes(x = hour, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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

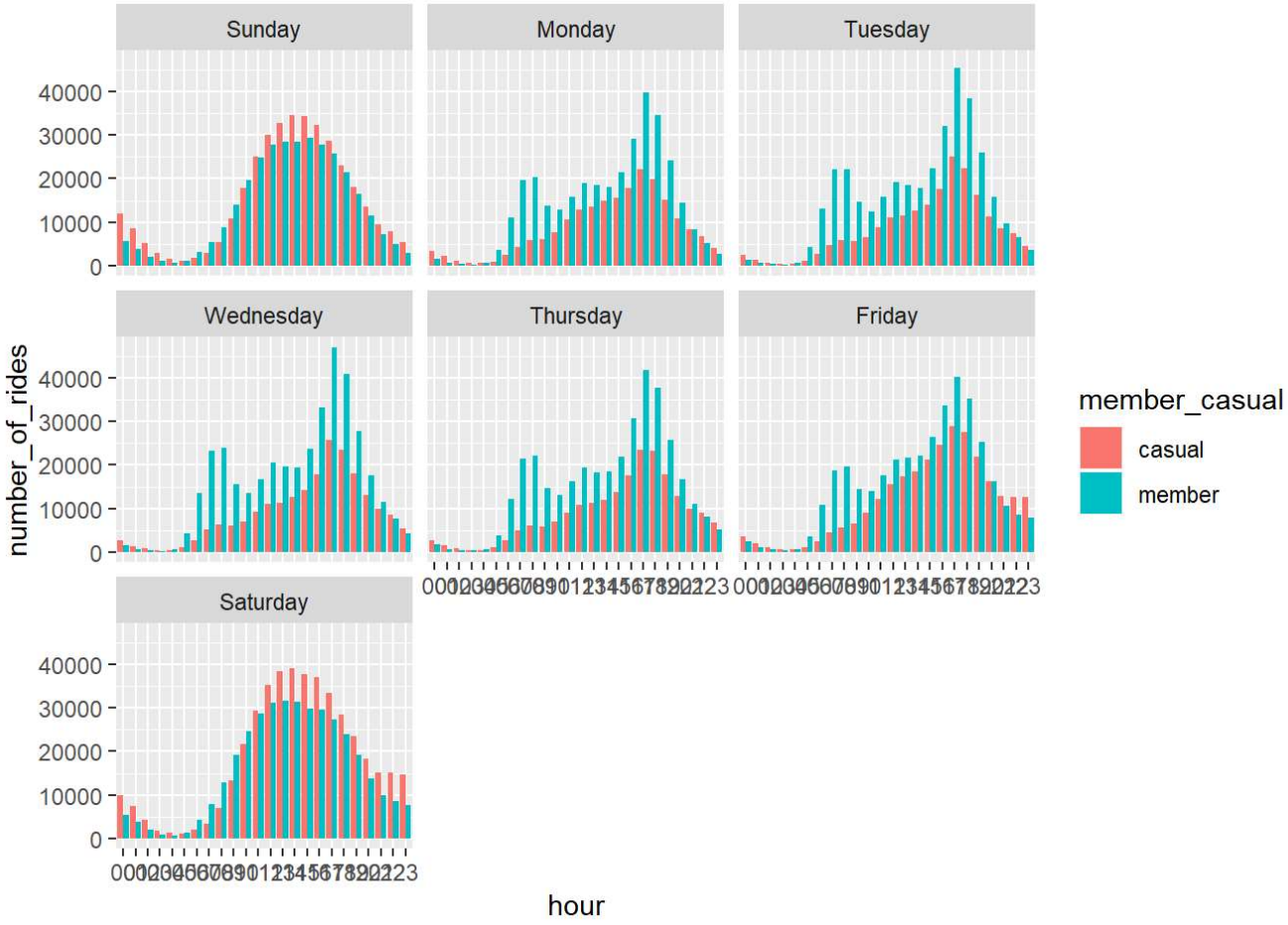


From the bar chart, we can see that most of the riders ride bicycle during evening time from 4pm - 6pm. It is the same for both casual riders and annual members. There is another peak hour for annual members which is from 7am - 9am. You may realised that there are quite a number of riders ride bicycle during midnight. But the number gets lesser and lesser when the time close to 4am in the morning.

Now, let us see the visualization for the number of rides by different hour in different days of week.

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week, hour) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, hour)  %>%
  ggplot(aes(x = hour, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  facet_wrap(~day_of_week)
```

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



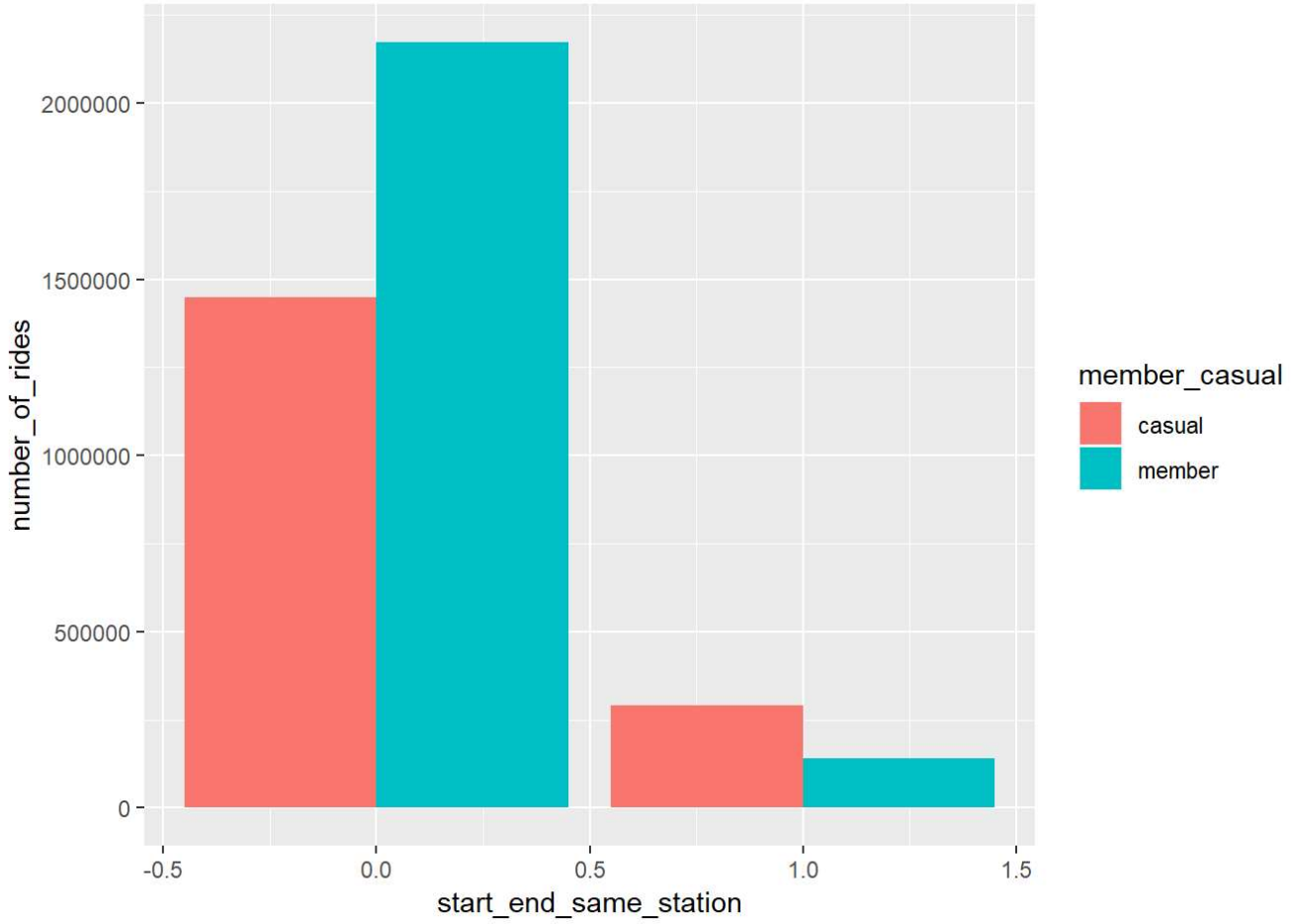
The trend is almost the same as the total number of rides as what we saw previously for the weekday, Monday to Friday. A different trend is observed for weekends, Saturday and Sunday. The number of rides get higher and higher from 5am in the morning and reach the peak at 2pm in the afternoon. There are more riders (both casual riders and annual members) ride bicycle to tour around during the weekend. While for weekday, there are more annual members who rent the bicycle to travel to and fro the workplace and their house.

Lastly, we create a visualization for the number of rides vs return to same station.

```
all_trips_v2 %>%
  group_by(member_casual, start_end_same_station) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, start_end_same_station)  %>%
  ggplot(aes(x = start_end_same_station, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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

```
## Warning: Removed 2 rows containing missing values (geom_col).
```



We can see that there are only a few riders will return the bike on the same station where they rented the bike. Among these riders, the number of casual riders is about 2 times of annual members. This mean that casual riders will rent bicycle to tour around an area then return the bicycles at the same station.

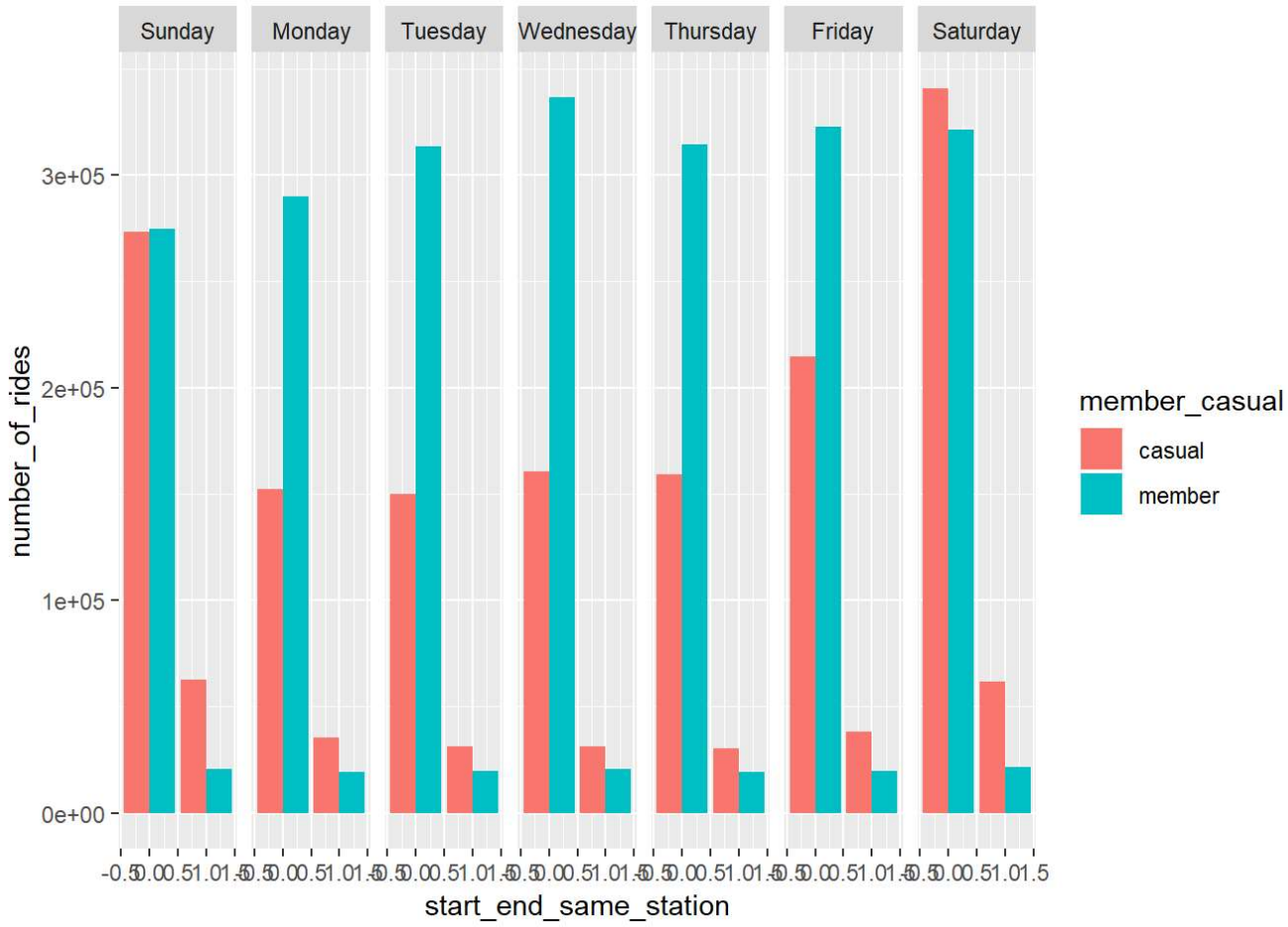
The total number of riders (both casualriders and annual members) who cycle and return bicyle at different station is 8 times of the number of riders who return the bicycle at the same station. Thus, we can understand that the riders usually will ride the bicycle to another place and return the bicycle there.

Then, we create visualization for number of rides vs return to same station for different days of week.

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week, start_end_same_station) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, start_end_same_station)  %>%
  ggplot(aes(x = start_end_same_station, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  facet_grid(cols = vars(day_of_week))
```

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

```
## Warning: Removed 14 rows containing missing values (geom_col).
```



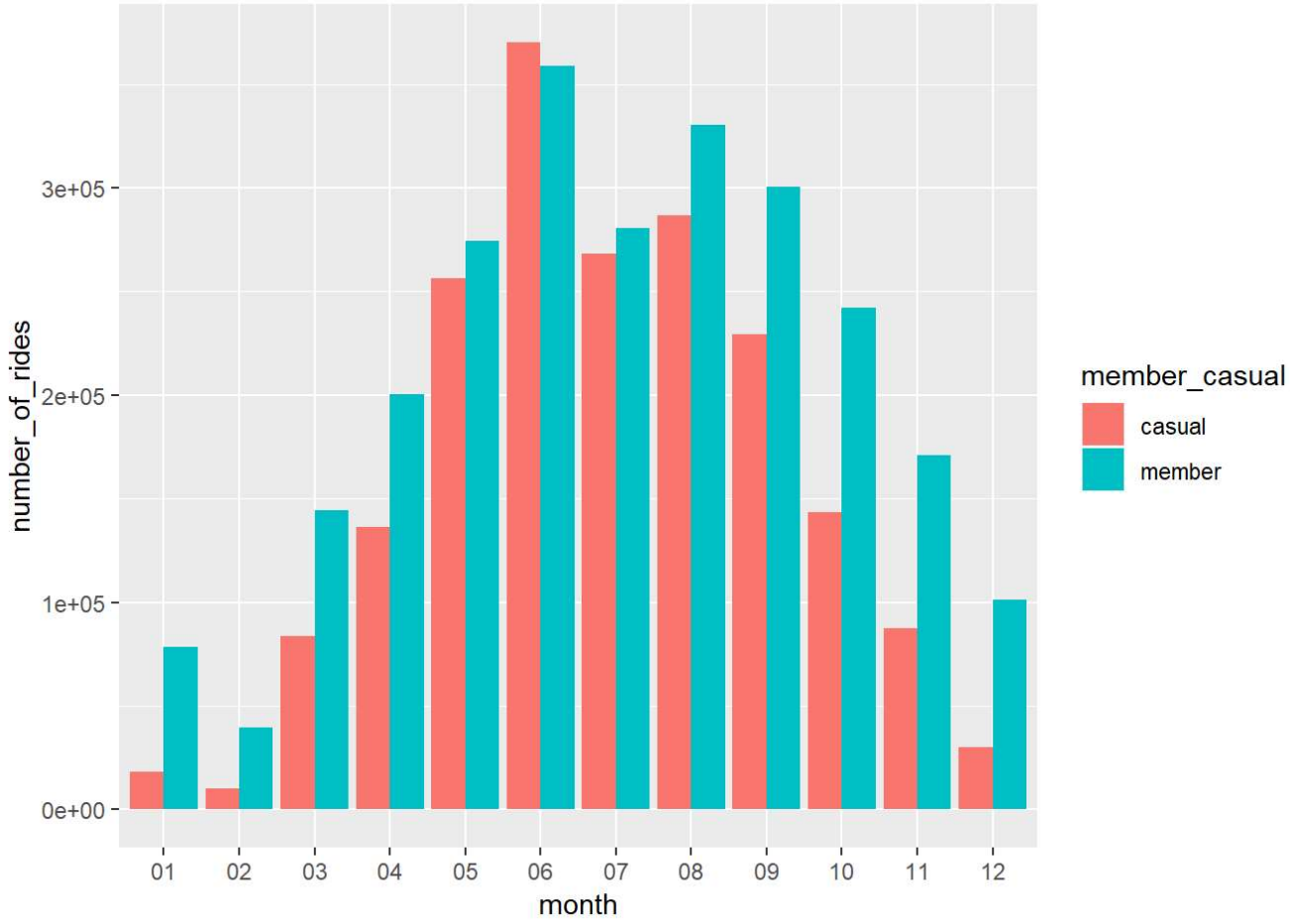
From the above bar charts, we can see that there are more annual members who rent and ride bicycle to another place and return the bicycle at another station on weekdays as compared to casual riders. The annual members who need to work on weekdays will usually rent bicycle to cycle to the workplace and return bicycle at the station nearby the workplace.

While for weekends, the number of casual riders and annual members who rent bicycle are almost the same for returning bicycle at different station.

Lastly, let's create a visualization for the number of rides by different months.

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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

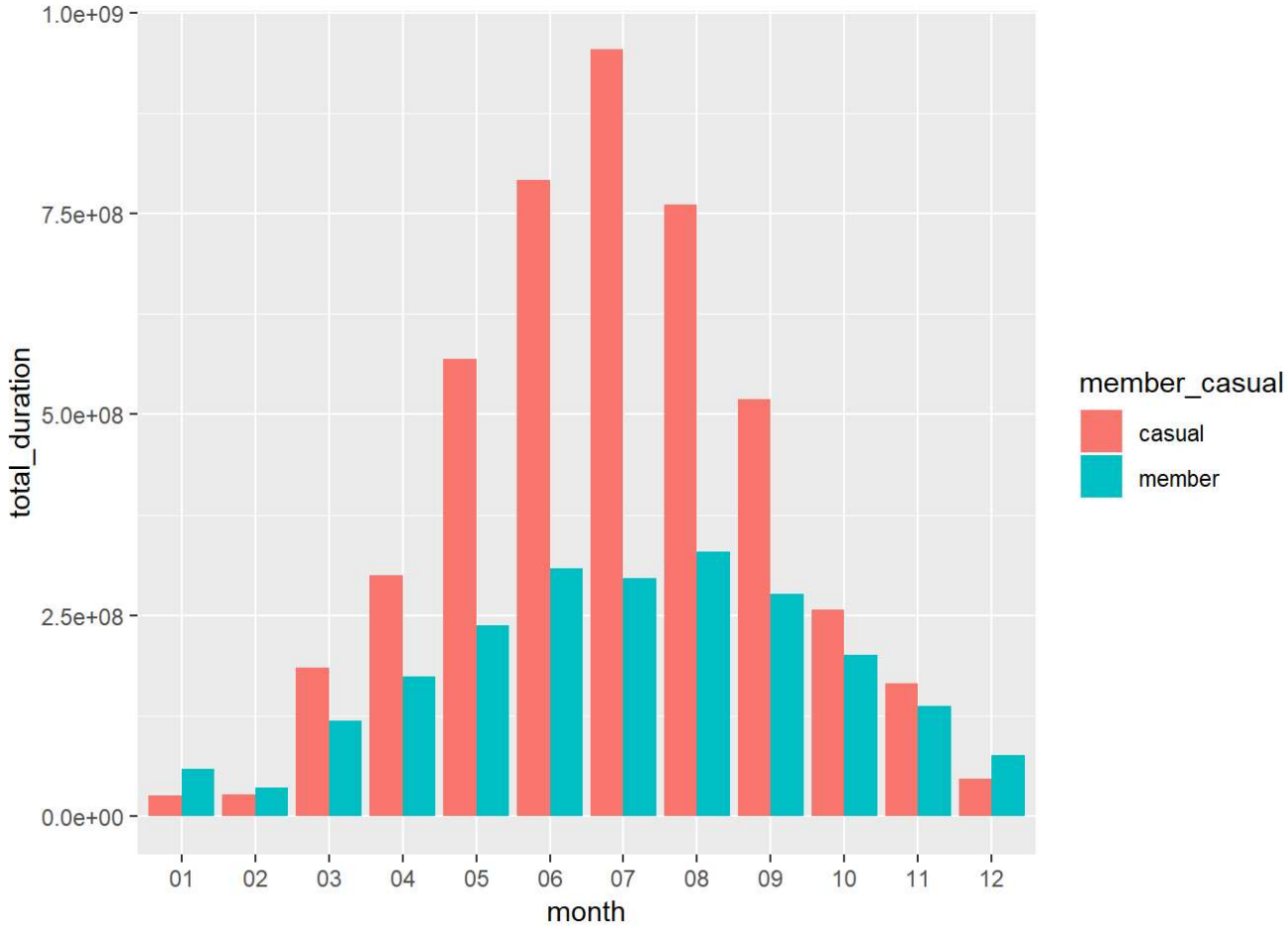


The above chart shown the total number of rides for different months. The number of rides starts to increase from February and reach to the highest number on June then decrease again till February.

Lastly, let's create a visualization for the total riding duration by different months.

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()
            ,total_duration = sum(ride_length)) %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = total_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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



The last chart shown the total riding duration of riders in different months. It have similar trend with the previous chart where the chart show a normal distribution. The chart shows that there is minimal ride duration on January and February then the riding duration pick up and reach the peak at July. Then the total duration drops again till December. From these 2 charts, we can see that there are very few people rent bicycle and ride in the winter season from December to February. Then, the number of rides and riding duration increase when the temperature increase in Spring season in March to May. The number of rides and riding duration reach maximum in the Summer (June to August). After Summer, the number of rides and riding duration decrease in Autumn season and reach the minimum in Winter.