Cyclistic_Bike_Share_Full_Year_Analysis:202102-202201

Yaxin Guan

2022/3/1

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

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

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

```
library(data.table) # help creates data table and import data
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(lubridate) # for date functions
```

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

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year

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

```r
library(ggplot2)  # for data visualization
# getwd() #displays your working directory
# setwd("/Users/usernames/Desktop/Divvy_Exercise/csv") #sets your working directory to simplify calls t
```

STEP 1: COLLECT DATA

```r
# Filepath <- "/Users/usernames/Desktop/Divvy_Exercise/"
trip_202201 <- read_csv(paste0(Filepath,"202201-divvy-tripdata/202201-divvy-tripdata.csv"))
```

```
## Rows: 103770 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
trip_202102 <- read_csv(paste0(Filepath,"202102-divvy-tripdata/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.
```

```r
trip_202103 <- read_csv(paste0(Filepath,"202103-divvy-tripdata/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.
```

```r
trip_202104 <- read_csv(paste0(Filepath,"202104-divvy-tripdata/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.
```

```r
trip_202105 <- read_csv(paste0(Filepath,"202105-divvy-tripdata/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.
```

```r
trip_202106 <- read_csv(paste0(Filepath,"202106-divvy-tripdata/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.
```

```r
trip_202107 <- read_csv(paste0(Filepath,"202107-divvy-tripdata/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.
```

```
trip_202108 <- read_csv(paste0(Filepath,"202108-divvy-tripdata/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.
```

```
trip_202109 <- read_csv(paste0(Filepath,"202109-divvy-tripdata/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.
```

```
trip_202110 <- read_csv(paste0(Filepath,"202110-divvy-tripdata/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.
```

```
trip_202111 <- read_csv(paste0(Filepath,"202111-divvy-tripdata/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.
```

```
trip_202112 <- read_csv(paste0(Filepath,"202112-divvy-tripdata/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.
```

STEP 2: CHECK DATA AND COMBINE INTO A SINGLE FILE

```
# Check to see if all the CSV files have the same column names.
colnames(trip_202201)
```

```
##  [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(trip_202102)
```

```
##  [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(trip_202103)
```

```
##  [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(trip_202104)
```

```
##  [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(trip_202105)
```

```
##  [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(trip_202106)
```

```
##  [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(trip_202107)
```

```
##  [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(trip_202108)
```

```
##  [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(trip_202109)
```

```
##  [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(trip_202110)
```

```
##  [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(trip_202111)
```

```
##  [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(trip_202112)
```

```
##  [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"
```

```
# Inspect the data frame and look for incongruencies
str(trip_202201)
```

```
## spec_tbl_df [103,770 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:103770] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB8
##  $ rideable_type     : chr [1:103770] "electric_bike" "electric_bike" "classic_bike" "classic_bike"
##  $ started_at        : chr [1:103770] "1/13/2022 11:59" "1/10/2022 8:41" "1/25/2022 4:53" "1/4/2022 
##  $ ended_at          : chr [1:103770] "1/13/2022 12:02" "1/10/2022 8:46" "1/25/2022 4:58" "1/4/2022 
##  $ start_station_name: chr [1:103770] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffiel
##  $ start_station_id  : chr [1:103770] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:103770] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave &
##  $ end_station_id    : chr [1:103770] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num [1:103770] 42 42 41.9 42 41.9 ...
##  $ start_lng         : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:103770] 42 42 41.9 42 41.9 ...
##  $ end_lng           : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:103770] "casual" "casual" "member" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(trip_202102)
```

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

```
str(trip_202103)
```

```
## spec_tbl_df [228,496 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D
##  $ rideable_type     : chr [1:228496] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : POSIXct[1:228496], format: "2021-03-16 08:32:30" "2021-03-28 01:26:28" ...
##  $ ended_at          : POSIXct[1:228496], format: "2021-03-16 08:36:34" "2021-03-28 01:36:55" ...
##  $ start_station_name: chr [1:228496] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave"
##  $ start_station_id  : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave
##  $ end_station_id    : chr [1:228496] "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num [1:228496] 41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:228496] 41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num [1:228496] -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:228496] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(trip_202104)
```

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

```
str(trip_202105)
```

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

```
str(trip_202106)
```

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

```
str(trip_202107)
```

```
## spec_tbl_df [822,410 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B5
##  $ rideable_type     : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct[1:822410], format: "2021-07-02 14:44:36" "2021-07-07 16:57:42" ...
##  $ ended_at          : POSIXct[1:822410], format: "2021-07-02 15:19:58" "2021-07-07 17:16:09" ...
##  $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez St" "Wa
##  $ start_station_id  : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St
##  $ end_station_id    : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:822410] "casual" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(trip_202108)
```

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

```
str(trip_202109)
```

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

```
str(trip_202110)
```

```
## spec_tbl_df [631,226 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:631226] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "36294
##  $ rideable_type     : chr [1:631226] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:631226], format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct[1:631226], format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr [1:631226] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:631226] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:631226] NA NA NA NA ...
##  $ end_station_id    : chr [1:631226] NA NA NA NA ...
##  $ start_lat         : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:631226] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(trip_202111)
```

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

```
str(trip_202112)
```

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

```
trip_202201 <-mutate(trip_202201, started_at = mdy_hm(started_at,tz = "UTC"),
ended_at = mdy_hm(ended_at, tz = "UTC"))
```

```
# Stack individual month's data frames into one big data frame
all_trips <- bind_rows(trip_202102, trip_202103, trip_202104, trip_202105, trip_202106, trip_202107,tri
```

```
# Filter out the data that will not be used in the analysis
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

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

```
# Inspect the new table 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"     "member_casual"
```

```
nrow(all_trips)  # rows in data frame
```

```
## [1] 5601999
```

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

```
## [1] 5601999       9
```

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

```
## # A tibble: 6 x 9
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 89E7AA~ classic_bike  2021-02-12 16:14:56 2021-02-12 16:21:43 Glenwood Ave & ~
## 2 0FEFDE~ classic_bike  2021-02-14 17:52:38 2021-02-14 18:12:09 Glenwood Ave & ~
## 3 E6159D~ electric_bike 2021-02-09 19:10:18 2021-02-09 19:19:10 Clark St & Lake~
## 4 B32D31~ classic_bike  2021-02-02 17:49:41 2021-02-02 17:54:06 Wood St & Chica~
## 5 83E463~ electric_bike 2021-02-23 15:07:23 2021-02-23 15:22:37 State St & 33rd~
## 6 BDAA7E~ electric_bike 2021-02-24 15:43:33 2021-02-24 15:49:05 Fairbanks St & ~
## # ... with 4 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, member_casual <chr>
```

```
tail(all_trips)
```

```
## # A tibble: 6 x 9
##   ride_id rideable_type started_at          ended_at            start_station_n~
##   <chr>   <chr>         <dttm>              <dttm>              <chr>
## 1 9C80CD~ electric_bike 2022-01-09 18:56:00 2022-01-09 19:02:00 Broadway & Wave~
## 2 8788DA~ electric_bike 2022-01-18 12:36:00 2022-01-18 12:46:00 Clinton St & Wa~
## 3 C6C3B6~ electric_bike 2022-01-27 11:00:00 2022-01-27 11:02:00 Racine Ave & Ra~
## 4 CA281A~ electric_bike 2022-01-10 16:14:00 2022-01-10 16:20:00 Broadway & Wave~
## 5 44E348~ electric_bike 2022-01-19 13:22:00 2022-01-19 13:24:00 Racine Ave & Ra~
## 6 E477C5~ electric_bike 2022-01-13 17:24:00 2022-01-13 17:28:00 Clinton St & Wa~
## # ... with 4 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, member_casual <chr>
```

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

```
## tibble [5,601,999 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5601999] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32
##  $ rideable_type     : chr [1:5601999] "classic_bike" "classic_bike" "electric_bike" "classic_bike"
##  $ started_at        : POSIXct[1:5601999], format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct[1:5601999], format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr [1:5601999] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark 
##  $ start_station_id  : chr [1:5601999] "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr [1:5601999] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "Stat
##  $ end_station_id    : chr [1:5601999] "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ member_casual     : chr [1:5601999] "member" "casual" "member" "member" ...
```

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

```
##    ride_id           rideable_type        started_at                 
##  Length:5601999     Length:5601999     Min.   :2021-02-01 00:55:44  
##  Class :character   Class :character   1st Qu.:2021-06-11 12:40:12  
##  Mode  :character   Mode  :character   Median :2021-08-04 22:01:30  
##                                        Mean   :2021-08-04 20:30:48  
##                                        3rd Qu.:2021-09-28 16:39:49  
##                                        Max.   :2022-01-31 23:58:00  
##     ended_at                      start_station_name start_station_id  
##  Min.   :2021-02-01 01:22:48   Length:5601999     Length:5601999    
##  1st Qu.:2021-06-11 13:03:36   Class :character   Class :character  
##  Median :2021-08-04 22:23:12   Mode  :character   Mode  :character  
##  Mean   :2021-08-04 20:52:44                                        
##  3rd Qu.:2021-09-28 16:55:21                                        
##  Max.   :2022-02-01 01:46:00                                        
##  end_station_name   end_station_id     member_casual     
##  Length:5601999     Length:5601999     Length:5601999    
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
## 
```

```
# Continue the inspection
table(all_trips$member_casual)
```

```
##
##  casual  member
## 2529408 3072591
```

```
# Add columns that list the date, month, day, and year of each ride
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")
```

```
# Add a "ride_length" calculation to all_trips (in seconds)
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
```

```
# Inspect the structure of the columns
str(all_trips)
```

```
## tibble [5,601,999 x 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5601999] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32
##  $ rideable_type     : chr [1:5601999] "classic_bike" "classic_bike" "electric_bike" "classic_bike"
##  $ started_at        : POSIXct[1:5601999], format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct[1:5601999], format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr [1:5601999] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark
##  $ start_station_id  : chr [1:5601999] "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr [1:5601999] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State
##  $ end_station_id    : chr [1:5601999] "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ member_casual     : chr [1:5601999] "member" "casual" "member" "member" ...
##  $ date              : Date[1:5601999], format: "2021-02-12" "2021-02-14" ...
##  $ month             : chr [1:5601999] "02" "02" "02" "02" ...
##  $ day               : chr [1:5601999] "12" "14" "09" "02" ...
##  $ year              : chr [1:5601999] "2021" "2021" "2021" "2021" ...
##  $ day_of_week       : chr [1:5601999] "Friday" "Sunday" "Tuesday" "Tuesday" ...
##  $ ride_length       : 'difftime' num [1:5601999] 407 1171 532 265 ...
##   ..- attr(*, "units")= chr "secs"
```

```
# 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
```

```
# The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quali
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length<0),]
```

```
# Remove duplicated data
all_trips_v2 <- all_trips_v2[!duplicated(all_trips_v2$ride_id), ]
dim(all_trips_v2)
```

```
## [1] 4903431      15
```

```
# Rmove any missing data
all_trips_v2 <-all_trips_v2[complete.cases(all_trips_v2),]
dim(all_trips_v2)
```

```
## [1] 4584805      15
```

```
#Check the data
summary(all_trips_v2)
```

```
##    ride_id          rideable_type        started_at
##  Length:4584805     Length:4584805     Min.   :2021-02-01 01:07:04
##  Class  :character  Class  :character  1st Qu.:2021-06-08 11:28:51
##  Mode   :character  Mode   :character  Median :2021-07-31 19:23:11
##                                        Mean   :2021-07-31 19:01:59
##                                        3rd Qu.:2021-09-23 07:21:28
##                                        Max.   :2022-01-31 23:58:00
##     ended_at                   start_station_name start_station_id
##  Min.   :2021-02-01 01:47:45   Length:4584805     Length:4584805
##  1st Qu.:2021-06-08 11:53:56   Class  :character  Class  :character
##  Median :2021-07-31 19:49:03   Mode   :character  Mode   :character
##  Mean   :2021-07-31 19:23:47
##  3rd Qu.:2021-09-23 07:34:06
##  Max.   :2022-02-01 00:12:00
##  end_station_name   end_station_id     member_casual           date
##  Length:4584805     Length:4584805     Length:4584805     Min.   :2021-02-01
##  Class  :character  Class  :character  Class  :character  1st Qu.:2021-06-08
##  Mode   :character  Mode   :character  Mode   :character  Median :2021-07-31
##                                                           Mean   :2021-07-31
##                                                           3rd Qu.:2021-09-23
##                                                           Max.   :2022-01-31
##     month               day                 year           day_of_week
##  Length:4584805     Length:4584805     Length:4584805     Length:4584805
##  Class  :character  Class  :character  Class  :character  Class  :character
##  Mode   :character  Mode   :character  Mode   :character  Mode   :character
##
##
##
##   ride_length
##  Min.   :      0
##  1st Qu.:    416
##  Median :    730
##  Mean   :   1307
##  3rd Qu.:   1323
##  Max.   :3356649
```

```
# Remove outliers
Q1 <- quantile(all_trips_v2$ride_length, prob = c(0.25, 0.75))[1]
Q3 <- quantile(all_trips_v2$ride_length, prob = c(0.25, 0.75))[2]
IQR_ride_length <- IQR(all_trips_v2$ride_length)

all_trips_v3 <- all_trips_v2[!(all_trips_v2$ride_length < Q1 - 1.5*IQR_ride_length |all_trips_v2$ride_l
dim(all_trips_v3)
```

```
## [1] 4235100      15
```

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

```
# Descriptive analysis on ride_length (all figures in seconds)
summary(all_trips_v3$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   395.0   673.0   832.5  1134.0  2683.0
```

```
# Compare members and casual users
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual, FUN = mean)
```

```
##   all_trips_v3$member_casual all_trips_v3$ride_length
## 1                     casual                 992.0368
## 2                     member                 721.2801
```

```
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual, FUN = median)
```

```
##   all_trips_v3$member_casual all_trips_v3$ride_length
## 1                     casual                      847
## 2                     member                      571
```

```
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual, FUN = max)
```

```
##   all_trips_v3$member_casual all_trips_v3$ride_length
## 1                     casual                     2683
## 2                     member                     2683
```

```
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual, FUN = min)
```

```
##   all_trips_v3$member_casual all_trips_v3$ride_length
## 1                     casual                        0
## 2                     member                        0
```

```
# See the average ride time by each day for members vs casual users
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual + all_trips_v3$day_of_week, FUN = mean)
```

```
##    all_trips_v3$member_casual all_trips_v3$day_of_week all_trips_v3$ride_length
## 1                      casual                   Friday                 962.0339
## 2                      member                   Friday                 706.8241
## 3                      casual                   Monday                 990.0699
## 4                      member                   Monday                 700.7933
## 5                      casual                 Saturday                1057.4062
## 6                      member                 Saturday                 788.8175
## 7                      casual                   Sunday                1072.0294
## 8                      member                   Sunday                 800.6298
## 9                      casual                 Thursday                 907.0809
## 10                     member                 Thursday                 687.4831
## 11                     casual                  Tuesday                 927.5026
## 12                     member                  Tuesday                 690.7724
## 13                     casual                Wednesday                 916.0496
## 14                     member                Wednesday                 693.6589
```

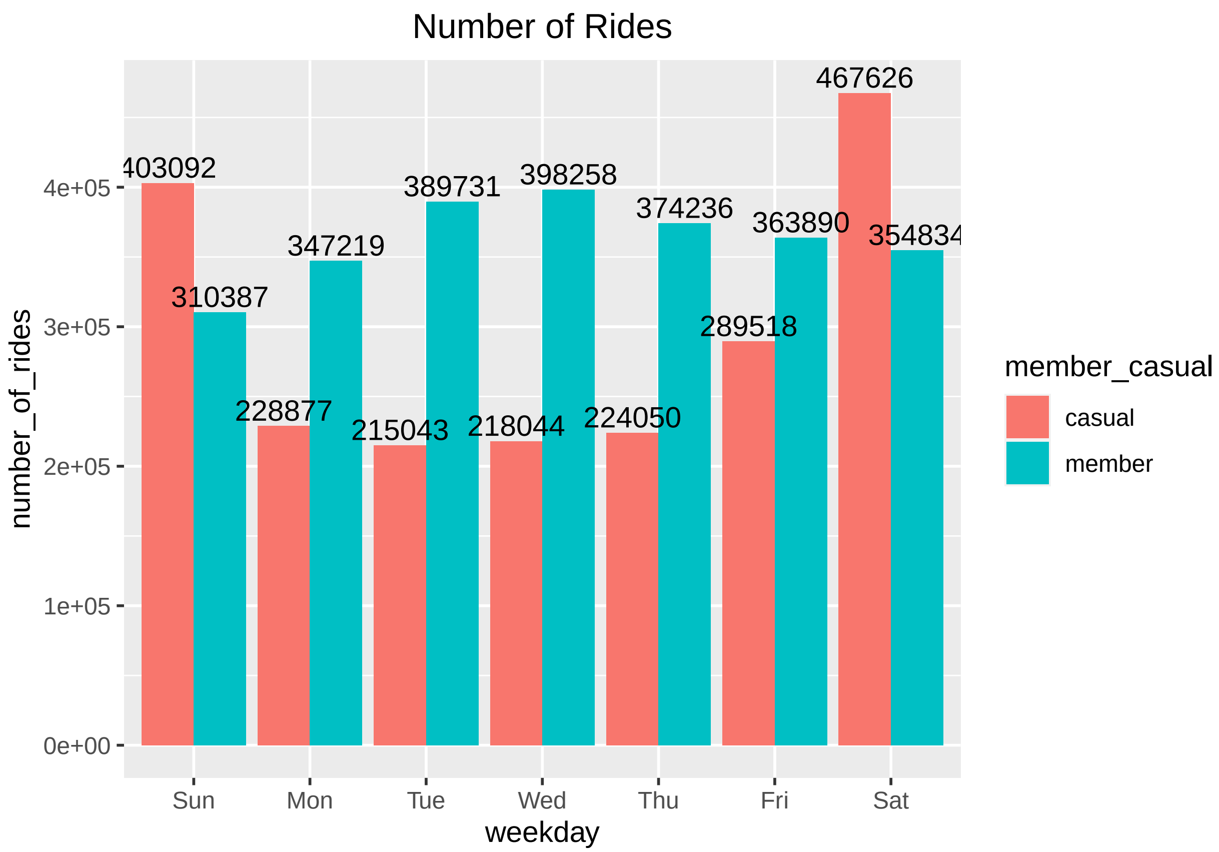
```
# Fix the order of the week's days for version3.
all_trips_v3$day_of_week <- ordered(all_trips_v3$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "W

# Fix the order of the week's days for version2 for the analysis later.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "W
# Run the average ride time by each day for members vs casual users
aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual + all_trips_v3$day_of_week, FUN = mean)
```

```
##    all_trips_v3$member_casual all_trips_v3$day_of_week all_trips_v3$ride_length
## 1                      casual                   Sunday                1072.0294
## 2                      member                   Sunday                 800.6298
## 3                      casual                   Monday                 990.0699
## 4                      member                   Monday                 700.7933
## 5                      casual                  Tuesday                 927.5026
## 6                      member                  Tuesday                 690.7724
## 7                      casual                Wednesday                 916.0496
## 8                      member                Wednesday                 693.6589
## 9                      casual                 Thursday                 907.0809
## 10                     member                 Thursday                 687.4831
## 11                     casual                   Friday                 962.0339
## 12                     member                   Friday                 706.8241
## 13                     casual                 Saturday                1057.4062
## 14                     member                 Saturday                 788.8175
```

```
# Visualize the number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") + labs(title = "Number of Rides")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label=number_of_rides),position=position_dodge(width=0.9),
            vjust=-0.25)
```

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

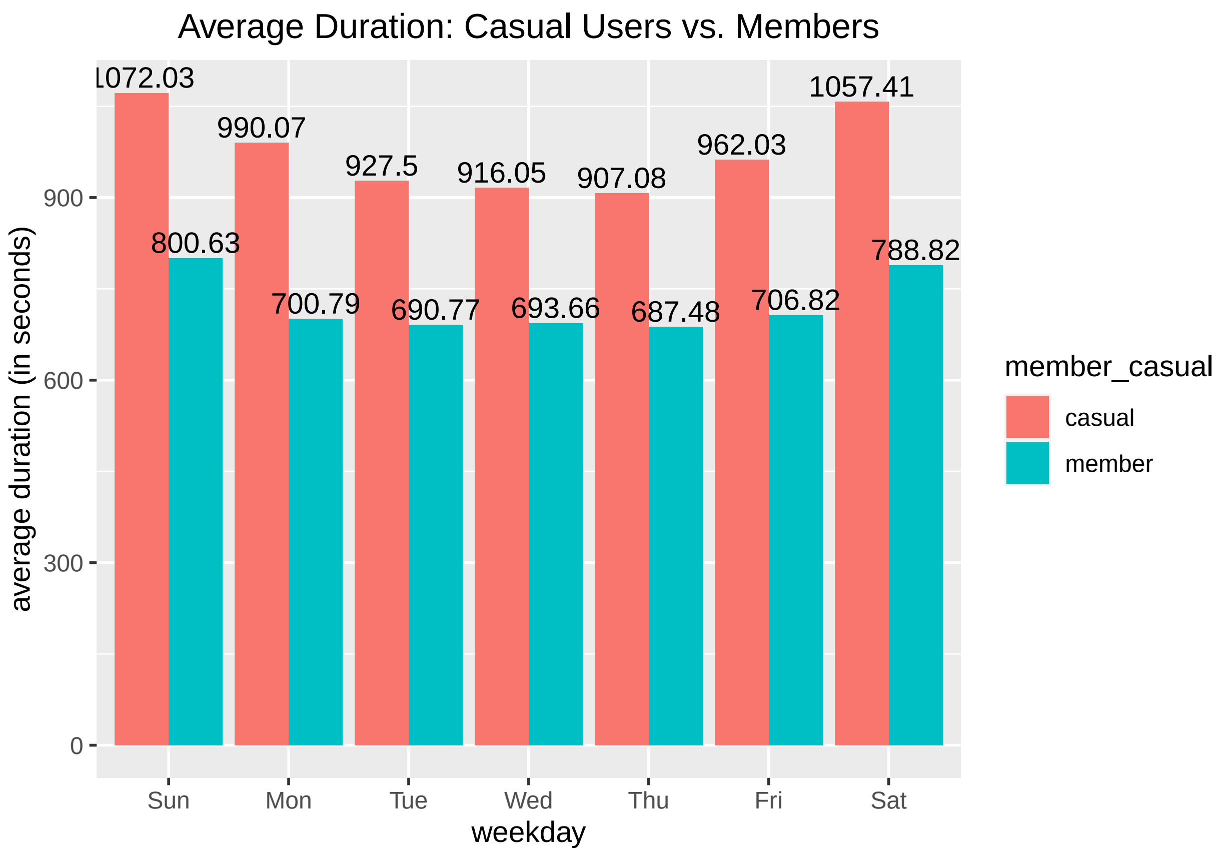


```
ggsave("Number_of_Rides.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

```
# Create a visualization for average duration
all_trips_v3 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") + labs(y = "average duration (in seconds)" ,title = "Average Duration: Ca
  theme(plot.title = element_text(hjust = 0.5)) + geom_text(aes(label=round(average_duration,2)),positio
            vjust=-0.25)
```

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



```
ggsave("Average_Ride_Length.jpg")
```

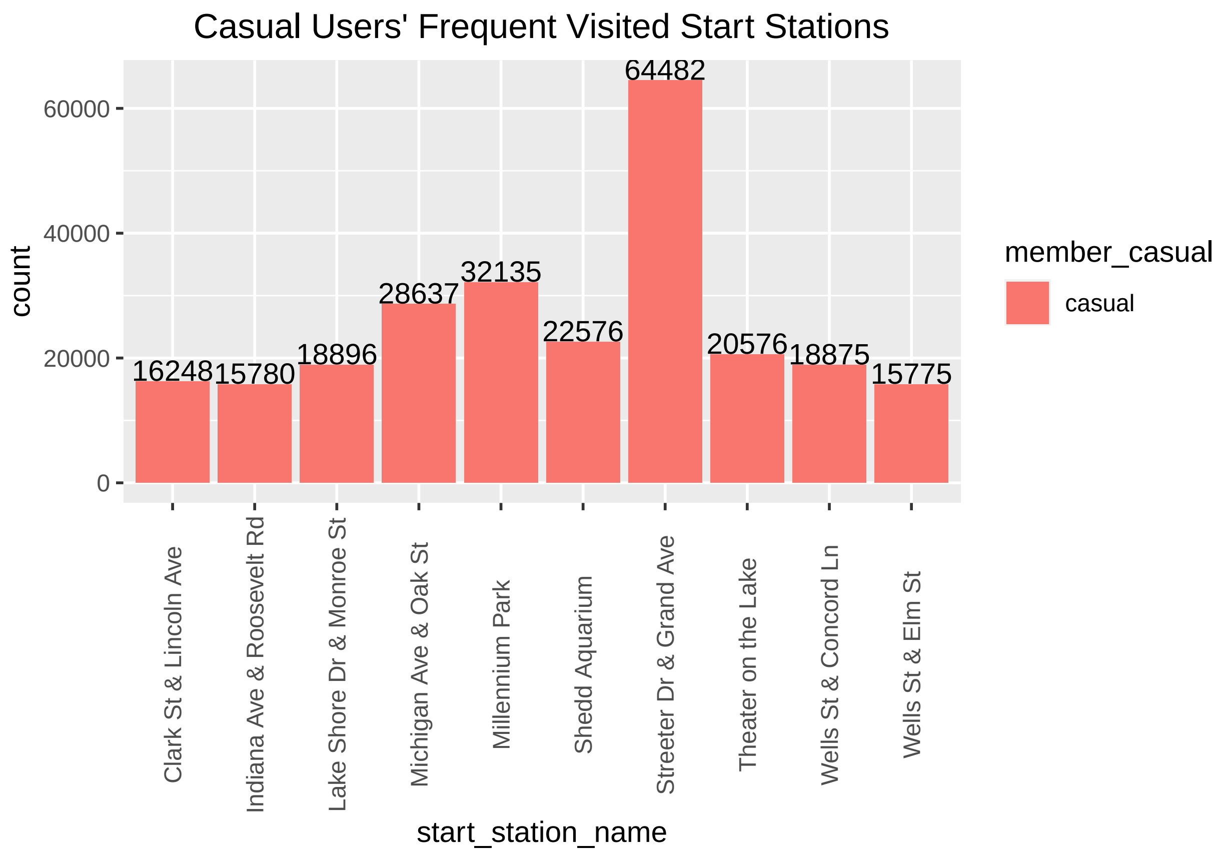
```
## Saving 6.5 x 4.5 in image
```

```
# Frequent Visited Start Station
Freq_start_station <- all_trips_v2 %>%
  group_by(member_casual,start_station_name) %>%
  summarise(count = n()) %>%
  arrange(-count)
```

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

```
Top_10_casual <- Freq_start_station[Freq_start_station$member_casual == "casual",
                                    ][1:10,]
Top_10_member <- Freq_start_station[Freq_start_station$member_casual == "member",
                                    ][1:10,]
```

```
# Casual Users' Frequent Visited Start Stations Graph
ggplot(Top_10_casual,aes(x = start_station_name, y = count,
                       fill = member_casual))+
  geom_bar(stat = "identity")+
  labs(title = "Casual Users' Frequent Visited Start Stations") +
    theme(axis.text.x =element_text(angle = 90, vjust = 0.5),
      plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label = count), vjust = -0.03)
```



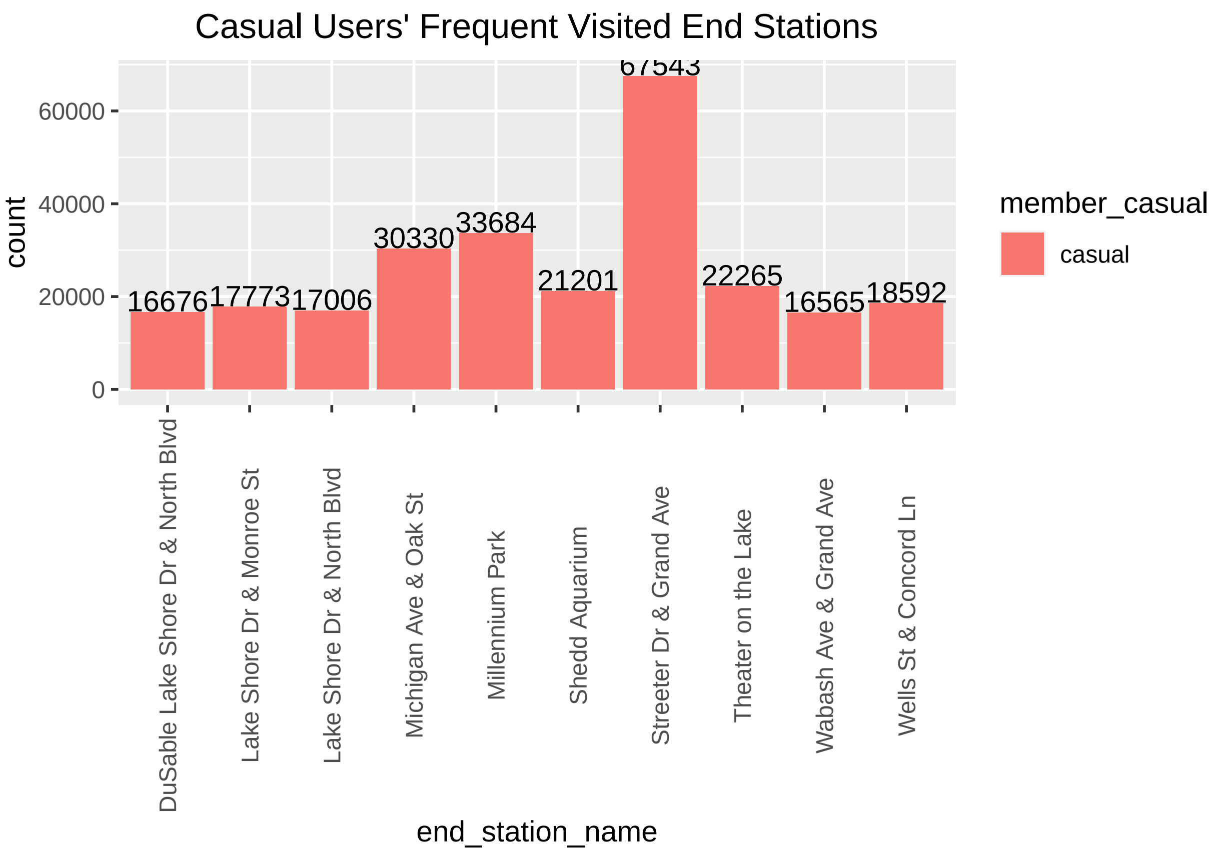
```
ggsave("Frequent Visited Start Startions.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

```
# Frequent Visited End Station
Freq_end_station <- all_trips_v2 %>%
  group_by(member_casual,end_station_name) %>%
  summarise(count = n()) %>%
  arrange(-count)
```

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

```
Top_10_casual_endstation <- Freq_end_station[Freq_end_station$member_casual == "casual",
                                    ][1:10,]
Top_10_member_endstation <- Freq_end_station[Freq_end_station$member_casual == "member",
                                    ][1:10,]
ggplot(Top_10_casual_endstation,aes(x = end_station_name, y = count,
                        fill = member_casual))+
  geom_bar(stat = "identity")+
  labs(title = "Casual Users' Frequent Visited End Stations") +
    theme(axis.text.x =element_text(angle = 90, vjust = 0.5),
      plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label = count), vjust = -0.03)
```



```
ggsave("Frequent Visited End Startions.jpg")
```

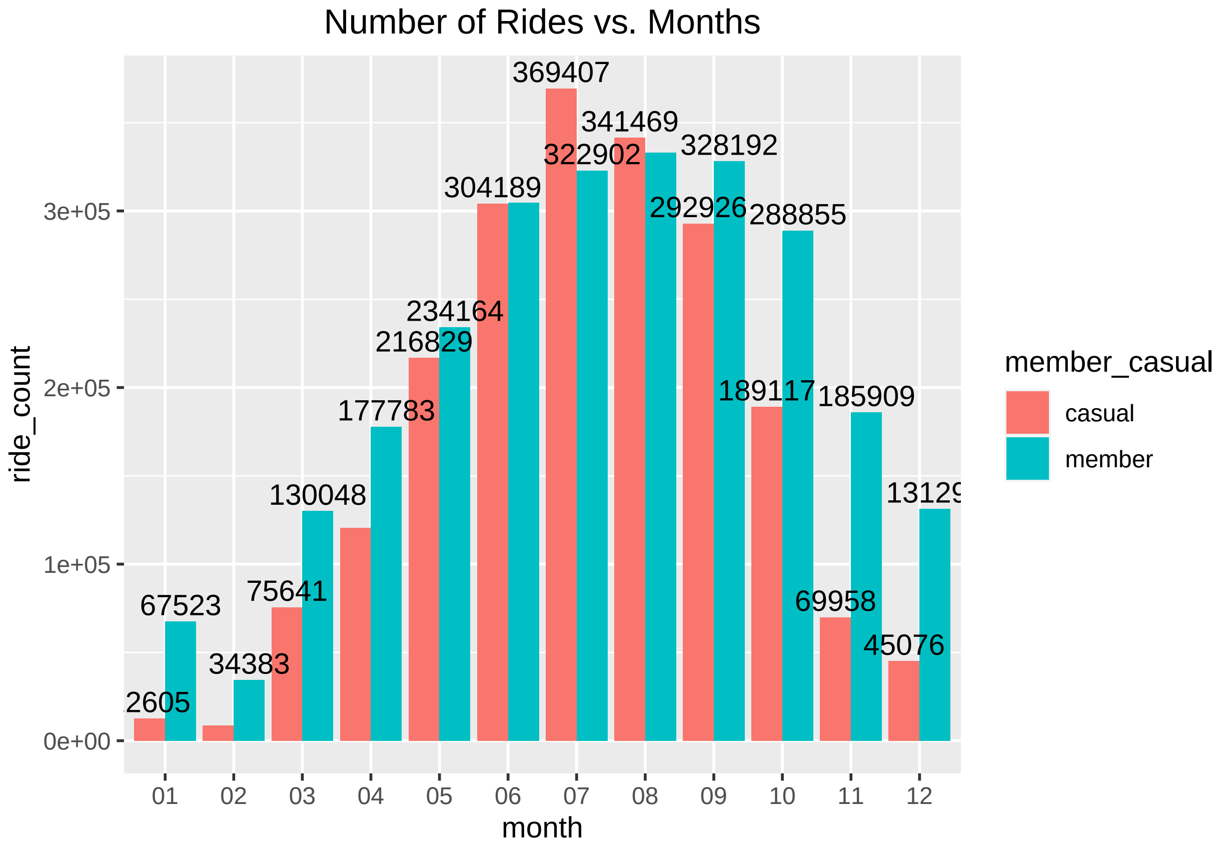
```
## Saving 6.5 x 4.5 in image
```

```
# Casual users: overlap of stations
overlap_stations <-Top_10_casual$start_station_name[Top_10_casual$start_station_name %in% Top_10_casual_
overlap_stations
```

```
## [1] "Streeter Dr & Grand Ave"   "Millennium Park"
## [3] "Michigan Ave & Oak St"     "Shedd Aquarium"
## [5] "Theater on the Lake"       "Lake Shore Dr & Monroe St"
## [7] "Wells St & Concord Ln"
```

```
# Causal users and members' number of rides by months
all_trips_v2 %>%
  group_by(month, member_casual) %>%
  summarise(ride_count = n()) %>%
  ggplot(aes(x = month, y = ride_count, fill = member_casual)) +
  geom_col(position = "dodge") + labs(title = "Number of Rides vs. Months")+
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=ride_count),position=position_dodge(width=0.9),
            vjust=-0.3, check_overlap = TRUE)
```

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

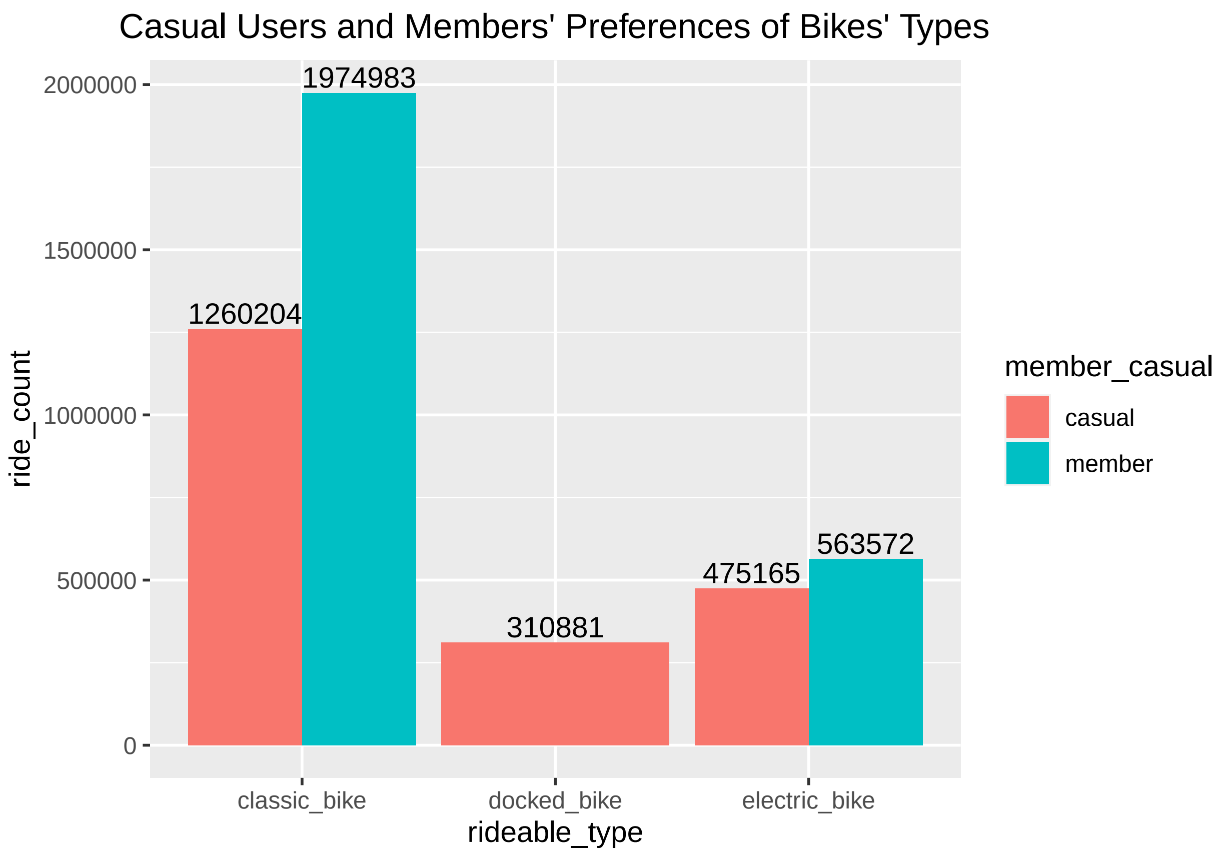


```
ggsave("Number of Rides vs. Months.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

```
# Casual users and members' preferences of types of bikes
all_trips_v2 %>%
  group_by(rideable_type, member_casual) %>%
  summarise(ride_count = n()) %>%
  ggplot(aes(x = rideable_type, y = ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Casual Users and Members' Preferences of Bikes' Types")+
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=ride_count),position=position_dodge(width=0.9),
            vjust=-0.25)
```

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



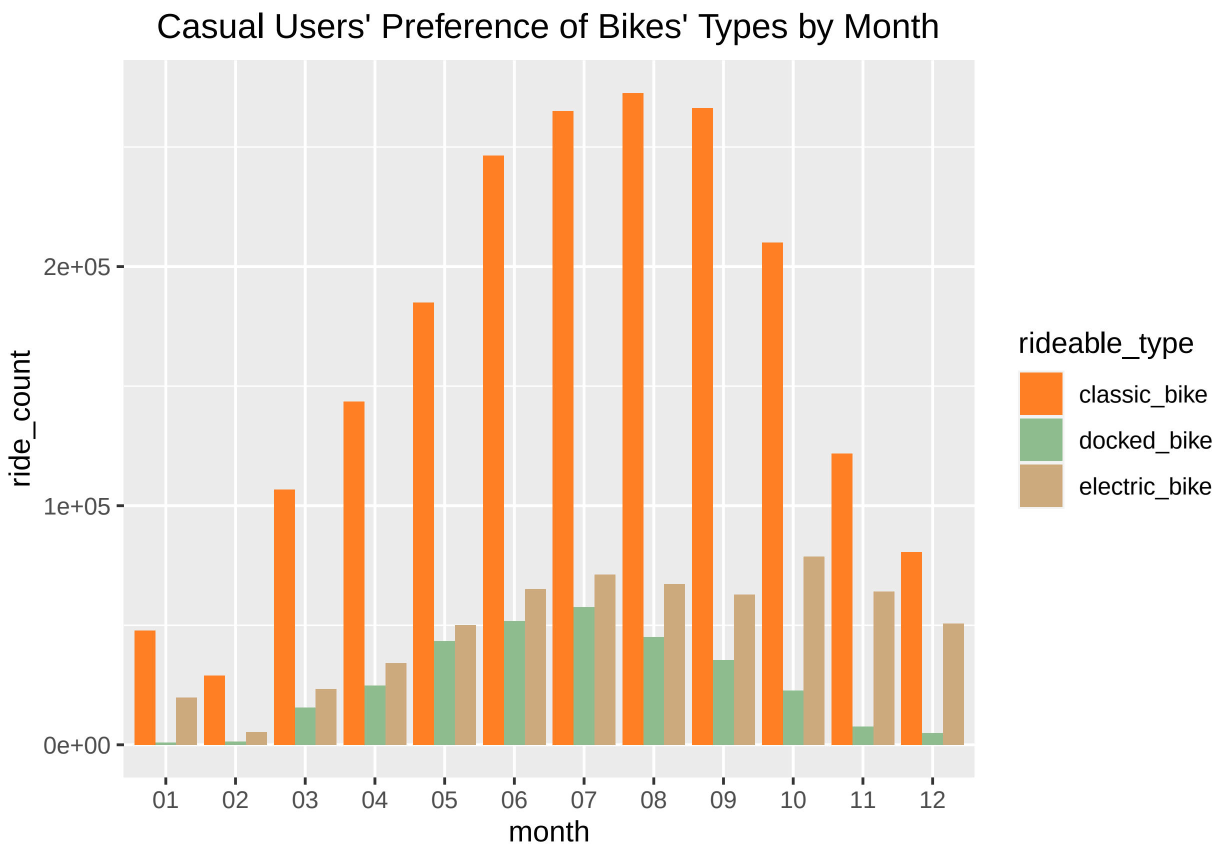
```
ggsave("Preferences of Bikes' Types.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

```
# Casual users' preference of bikes' types by month
preference_type_by_month <-all_trips_v2 %>%
  group_by(month, member_casual, rideable_type) %>%
  summarise(ride_count = n())
```

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

```
casual_preference_by_month <-preference_type_by_month[preference_type_by_month$member_casual == "casual
ggplot(data =preference_type_by_month ,aes(x = month, y = ride_count,
fill = rideable_type)) + geom_col(position = "dodge") +
  scale_fill_manual(values = c("classic_bike" = "chocolate1",
      "docked_bike" = "darkseagreen", "electric_bike" = "burlywood3"))+
  labs(title = "Casual Users' Preference of Bikes' Types by Month")+
  theme(plot.title = element_text(hjust = 0.5))
```



STEP 5: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

```
# Create a csv file
counts <- aggregate(all_trips_v3$ride_length ~ all_trips_v3$member_casual + all_trips_v3$day_of_week, FU

write.csv(counts, file = "D:/Career/Google Data Analytics Program/Case Study/Google Data Analytics Certi
# Choose the file path
```

```
# Set a data table to extract the needed data for Student T-Test
data1 <- setDT(all_trips_v3)[,.(average_duration = sum(ride_length)/length(ride_length)), by = .(member_

count_casual <- data1[member_casual == "casual" & order(day_of_week), average_duration]
count_member <- data1[member_casual == "member" & order(day_of_week), average_duration]

# Check if the variances are equal
var.test(count_casual,count_member)
```

```
##
##  F test to compare two variances
##
## data:  count_casual and count_member
## F = 1.8884, num df = 6, denom df = 6, p-value = 0.4586
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.3244747 10.9898183
## sample estimates:
## ratio of variances
##           1.888364
```

```
# The result shows that the variance of casual users is different from the variance of member.

#Student T-Test
t.test(count_casual,count_member, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  count_casual and count_member
## t = 8.042, df = 10.963, p-value = 3.174e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  195.6183      Inf
## sample estimates:
## mean of x mean of y
##  976.0246  724.1399
```

```
# significant greater

# Repeat the sames steps above for number of rides by day of weeks
data2 <- setDT(all_trips_v2)[,.(number_of_ride = .N),
        by = .(member_casual, day_of_week)][order(member_casual,day_of_week)]

ride_count_casual <- data2[member_casual == "casual"& order(day_of_week), number_of_ride]
ride_count_member <- data2[member_casual == "member"& order(day_of_week), number_of_ride]

# Check if the variances are equal
var.test(ride_count_casual,ride_count_member)
```

```
##
##  F test to compare two variances
##
## data:  ride_count_casual and ride_count_member
## F = 12.227, num df = 6, denom df = 6, p-value = 0.007692
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   2.100926 71.157439
## sample estimates:
## ratio of variances
##           12.22688
```

```
# different variances
t.test(ride_count_casual,ride_count_member, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  ride_count_casual and ride_count_member
## t = -1.7433, df = 6.9749, p-value = 0.9375
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -146801     Inf
## sample estimates:
## mean of x mean of y
##  292321.4  362650.7
```

```
#significant greater
```

```
# Location
location_counts<- all_trips_v2 %>%
  group_by(member_casual,start_station_name) %>%
  summarise(number_of_station = n())
```

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

```
write.csv(location_counts, file = "D:/Career/Google Data Analytics Program/Case Study/Google Data Analy
```