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
```

```
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

```
# Filepath <- "/Users/usernames/Desktop/Divvy_Exercise/"
trip_202201 <- fread(paste0(Filepath,"202201-divvy-tripdata/202201-divvy-tripdata.csv"))
trip_202102 <- fread(paste0(Filepath,"202102-divvy-tripdata/202102-divvy-tripdata.csv"))
trip_202103 <- fread(paste0(Filepath,"202103-divvy-tripdata/202103-divvy-tripdata.csv"))
trip_202104 <- fread(paste0(Filepath,"202104-divvy-tripdata/202104-divvy-tripdata.csv"))
trip_202105 <- fread(paste0(Filepath,"202105-divvy-tripdata/202105-divvy-tripdata.csv"))
trip_202106 <- fread(paste0(Filepath,"202106-divvy-tripdata/202106-divvy-tripdata.csv"))
trip_202107 <- fread(paste0(Filepath,"202107-divvy-tripdata/202107-divvy-tripdata.csv"))
trip_202108 <- fread(paste0(Filepath,"202108-divvy-tripdata/202108-divvy-tripdata.csv"))
trip_202109 <- fread(paste0(Filepath,"202109-divvy-tripdata/202109-divvy-tripdata.csv"))
trip_202110 <- fread(paste0(Filepath,"202110-divvy-tripdata/202110-divvy-tripdata.csv"))
trip_202111 <- fread(paste0(Filepath,"202111-divvy-tripdata/202111-divvy-tripdata.csv"))
trip_202112 <- fread(paste0(Filepath,"202112-divvy-tripdata/202112-divvy-tripdata.csv"))
```

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)
```

```
## Classes 'data.table' and 'data.frame':   103770 obs. of  13 variables:
##  $ ride_id           : chr  "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB80ED4191054
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "1/13/2022 11:59" "1/10/2022 8:41" "1/25/2022 4:53" "1/4/2022 0:18" ...
##  $ ended_at          : chr  "1/13/2022 12:02" "1/10/2022 8:46" "1/25/2022 4:58" "1/4/2022 0:33" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffield Ave & F
##  $ start_station_id  : chr  "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr  "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave & Fullerton
##  $ end_station_id    : chr  "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num  42 42 41.9 42 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 42 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "member" "casual" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202102)
```

```
## Classes 'data.table' and 'data.frame':   49622 obs. of  13 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E7
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct, format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake S
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Rand
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num  42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202103)
```

```
## Classes 'data.table' and 'data.frame':   228496 obs. of  13 variables:
##  $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-03-16 08:32:30" "2021-03-28 01:26:28" ...
##  $ ended_at          : POSIXct, format: "2021-03-16 08:36:34" "2021-03-28 01:36:55" ...
##  $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields A
##  $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted
##  $ end_station_id    : chr  "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num  41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202104)
```

```
## Classes 'data.table' and 'data.frame':   337230 obs. of  13 variables:
##  $ ride_id           : chr  "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887262AD101C6
##  $ rideable_type     : chr  "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-04-12 18:25:36" "2021-04-27 17:27:11" ...
##  $ ended_at          : POSIXct, format: "2021-04-12 18:56:55" "2021-04-27 18:31:29" ...
##  $ start_station_name: chr  "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blvd & 84th S
##  $ start_station_id  : chr  "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr  "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loomis Blvd &
##  $ end_station_id    : chr  "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202105)
```

```
## Classes 'data.table' and 'data.frame':   531633 obs. of  13 variables:
##  $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D39110C
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-05-30 11:58:15" "2021-05-30 11:29:14" ...
##  $ ended_at          : POSIXct, format: "2021-05-30 12:10:39" "2021-05-30 12:14:09" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202106)
```

```
## Classes 'data.table' and 'data.frame':   729595 obs. of  13 variables:
##  $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C0FE48C4122
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-06-13 14:31:28" "2021-06-04 11:18:02" ...
##  $ ended_at          : POSIXct, format: "2021-06-13 14:34:11" "2021-06-04 11:24:19" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.8 41.8 41.8 41.8 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.8 41.8 41.8 41.8 41.8 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202107)
```

```
## Classes 'data.table' and 'data.frame':   822410 obs. of  13 variables:
##  $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8A
##  $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-07-02 14:44:36" "2021-07-07 16:57:42" ...
##  $ ended_at          : POSIXct, format: "2021-07-02 15:19:58" "2021-07-07 17:16:09" ...
##  $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez St" "Wabash Ave &
##  $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St & Hubbard
##  $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202108)
```

```
## Classes 'data.table' and 'data.frame':   804352 obs. of  13 variables:
```

```
##  $ ride_id           : chr  "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFAF1
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-08-10 17:15:49" "2021-08-10 17:23:14" ...
##  $ ended_at          : POSIXct, format: "2021-08-10 17:22:44" "2021-08-10 17:39:24" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202109)
```

```
## Classes 'data.table' and 'data.frame':   756147 obs. of  13 variables:
##  $ ride_id           : chr  "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1DE133B3DBF
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-09-28 16:07:10" "2021-09-28 14:24:51" ...
##  $ ended_at          : POSIXct, format: "2021-09-28 16:09:54" "2021-09-28 14:40:05" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202110)
```

```
## Classes 'data.table' and 'data.frame':   631226 obs. of  13 variables:
##  $ ride_id           : chr  "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362947F0437E15
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct, format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr  "Kingsbury St & Kinzie St" "" "" "" ...
##  $ start_station_id  : chr  "KA1503000043" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202111)
```

```
## Classes 'data.table' and 'data.frame':   359978 obs. of  13 variables:
##  $ ride_id           : chr  "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3BE33085BCFF
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2021-11-27 13:27:38" "2021-11-27 13:38:25" ...
##  $ ended_at          : POSIXct, format: "2021-11-27 13:46:38" "2021-11-27 13:56:10" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 42 42 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ end_lat           : num  42 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(trip_202112)
```

```
## Classes 'data.table' and 'data.frame':   247540 obs. of  13 variables:
##  $ ride_id           : chr  "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278BA87BF6983
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-12-07 15:06:07" "2021-12-11 03:43:29" ...
##  $ ended_at          : POSIXct, format: "2021-12-07 15:13:42" "2021-12-11 04:10:23" ...
##  $ start_station_name: chr  "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted St & North B
##  $ start_station_id  : chr  "13307" "KP1705001026" "KA1504000117" "KA1504000117" ...
##  $ end_station_name  : chr  "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Broadway & Barry Ave
##  $ end_station_id    : chr  "TA1307000130" "TA1307000119" "13137" "KP1705001026" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
##  - attr(*, ".internal.selfref")=<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,trip
```

```
# 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.
```

```
##             ride_id rideable_type          started_at            ended_at
## 1: 89E7AA6C29227EFF  classic_bike 2021-02-12 16:14:56 2021-02-12 16:21:43
## 2: 0FEFDE2603568365  classic_bike 2021-02-14 17:52:38 2021-02-14 18:12:09
## 3: E6159D746B2DBB91 electric_bike 2021-02-09 19:10:18 2021-02-09 19:19:10
## 4: B32D3199F1C2E75B  classic_bike 2021-02-02 17:49:41 2021-02-02 17:54:06
## 5: 83E463F23575F4BF electric_bike 2021-02-23 15:07:23 2021-02-23 15:22:37
## 6: BDAA7E3494E8D545 electric_bike 2021-02-24 15:43:33 2021-02-24 15:49:05
##            start_station_name start_station_id           end_station_name
## 1:   Glenwood Ave & Touhy Ave              525 Sheridan Rd & Columbia Ave
## 2:   Glenwood Ave & Touhy Ave              525   Bosworth Ave & Howard St
## 3:         Clark St & Lake St     KA1503000012     State St & Randolph St
## 4:      Wood St & Chicago Ave              637    Honore St & Division St
## 5:         State St & 33rd St            13216      Emerald Ave & 31st St
## 6: Fairbanks St & Superior St            18003      LaSalle Dr & Huron St
##    end_station_id member_casual
## 1:            660        member
## 2:          16806        casual
## 3:   TA1305000029        member
## 4:   TA1305000034        member
## 5:   TA1309000055        member
## 6:   KP1705001026        casual
```

```
tail(all_trips)
```

```
##             ride_id rideable_type          started_at            ended_at
## 1: 9C80CD03B685B1B4 electric_bike 2022-01-09 18:56:00 2022-01-09 19:02:00
## 2: 8788DA3EDE8FD8AB electric_bike 2022-01-18 12:36:00 2022-01-18 12:46:00
## 3: C6C3B64FDC827D8C electric_bike 2022-01-27 11:00:00 2022-01-27 11:02:00
## 4: CA281AE7D8B06F5A electric_bike 2022-01-10 16:14:00 2022-01-10 16:20:00
## 5: 44E348991862319B electric_bike 2022-01-19 13:22:00 2022-01-19 13:24:00
## 6: E477C594A182AE58 electric_bike 2022-01-13 17:24:00 2022-01-13 17:28:00
##              start_station_name start_station_id           end_station_name
## 1:      Broadway & Waveland Ave            13325
## 2: Clinton St & Washington Blvd           WL-012
## 3:     Racine Ave & Randolph St            13155
## 4:      Broadway & Waveland Ave            13325       Clark St & Grace St
## 5:     Racine Ave & Randolph St            13155
## 6: Clinton St & Washington Blvd           WL-012 Desplaines St & Kinzie St
##    end_station_id member_casual
## 1:                       casual
## 2:                       casual
## 3:                       casual
## 4:   TA1307000127        casual
## 5:                       casual
## 6:   TA1306000003        casual
```

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

```
## Classes 'data.table' and 'data.frame':   5601999 obs. of  9 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct, format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Rando
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
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
# unique(is.na(all_trips)) # The results show no missing values in the data frame
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)
```

```
## Classes 'data.table' and 'data.frame':   5601999 obs. of  15 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : POSIXct, format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...
##  $ ended_at          : POSIXct, format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Rando
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
##  $ date              : Date, format: "2021-02-12" "2021-02-14" ...
##  $ month             : chr  "02" "02" "02" "02" ...
##  $ day               : chr  "12" "14" "09" "02" ...
##  $ year              : chr  "2021" "2021" "2021" "2021" ...
##  $ day_of_week       : chr  "Friday" "Sunday" "Tuesday" "Tuesday" ...
##  $ ride_length       : 'difftime' num  407 1171 532 265 ...
##   ..- attr(*, "units")= chr "secs"
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# 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
```

```r
# 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),]
```

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

```r
# Descriptive analysis on ride_length (all figures in seconds)
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     403     718    1316    1303 3356649
```

```r
# 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                1922.1317
## 2                     member                 816.4348
```

```r
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                      957
## 2                     member                      574
```

```r
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                    93596
```

```r
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
```

```r
# 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, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Friday                1822.0164
## 2                      member                   Friday                 799.0720
## 3                      casual                   Monday                1915.5927
## 4                      member                   Monday                 791.4776
## 5                      casual                 Saturday                2084.9814
## 6                      member                 Saturday                 914.4328
## 7                      casual                   Sunday                2253.5273
## 8                      member                   Sunday                 939.1134
## 9                      casual                 Thursday                1669.3037
## 10                     member                 Thursday                 765.2494
## 11                     casual                  Tuesday                1676.1755
## 12                     member                  Tuesday                 767.1518
## 13                     casual                Wednesday                1664.7192
## 14                     member                Wednesday                 766.3527
```

```
# Notice that the days of the week are out of order. Let's fix that.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "W
# Now, let's run 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, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Sunday                2253.5273
## 2                      member                   Sunday                 939.1134
## 3                      casual                   Monday                1915.5927
## 4                      member                   Monday                 791.4776
## 5                      casual                  Tuesday                1676.1755
## 6                      member                  Tuesday                 767.1518
## 7                      casual                Wednesday                1664.7192
## 8                      member                Wednesday                 766.3527
## 9                      casual                 Thursday                1669.3037
## 10                     member                 Thursday                 765.2494
## 11                     casual                   Friday                1822.0164
## 12                     member                   Friday                 799.0720
## 13                     casual                 Saturday                2084.9814
## 14                     member                 Saturday                 914.4328
```

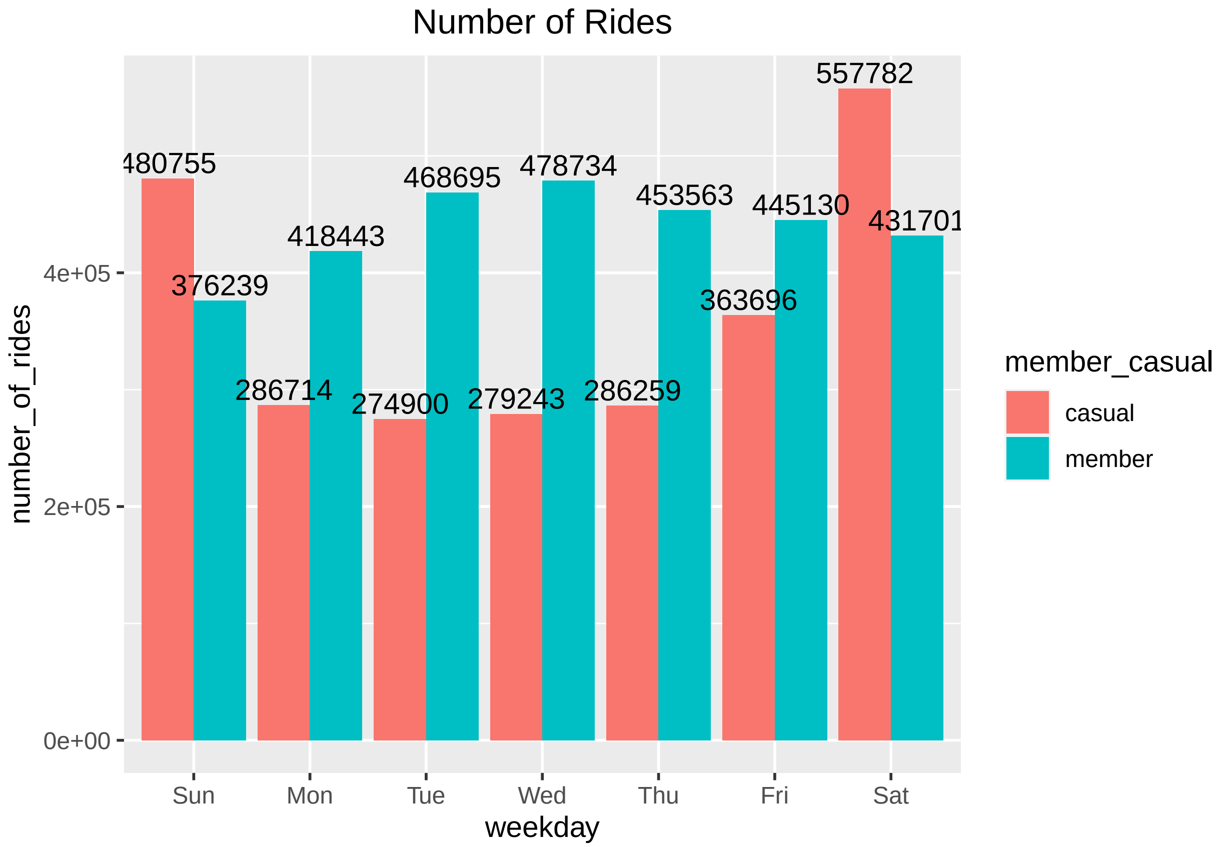
```
# analyze ridership data by type and weekday
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>% # create weekday field
  group_by(member_casual, weekday) %>%  # groups by usertype and weekday
  summarise(number_of_rides = n(), # calculates the number of rides and average duration
            average_duration = mean(ride_length)) %>% # calculates the average duration
    arrange(member_casual, weekday)
```

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

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              480755            2254.
##  2 casual        Mon              286714            1916.
##  3 casual        Tue              274900            1676.
##  4 casual        Wed              279243            1665.
##  5 casual        Thu              286259            1669.
##  6 casual        Fri              363696            1822.
##  7 casual        Sat              557782            2085.
##  8 member        Sun              376239             939.
##  9 member        Mon              418443             791.
## 10 member        Tue              468695            767.
## 11 member        Wed              478734            766.
## 12 member        Thu              453563            765.
## 13 member        Fri              445130            799.
## 14 member        Sat              431701            914.
```

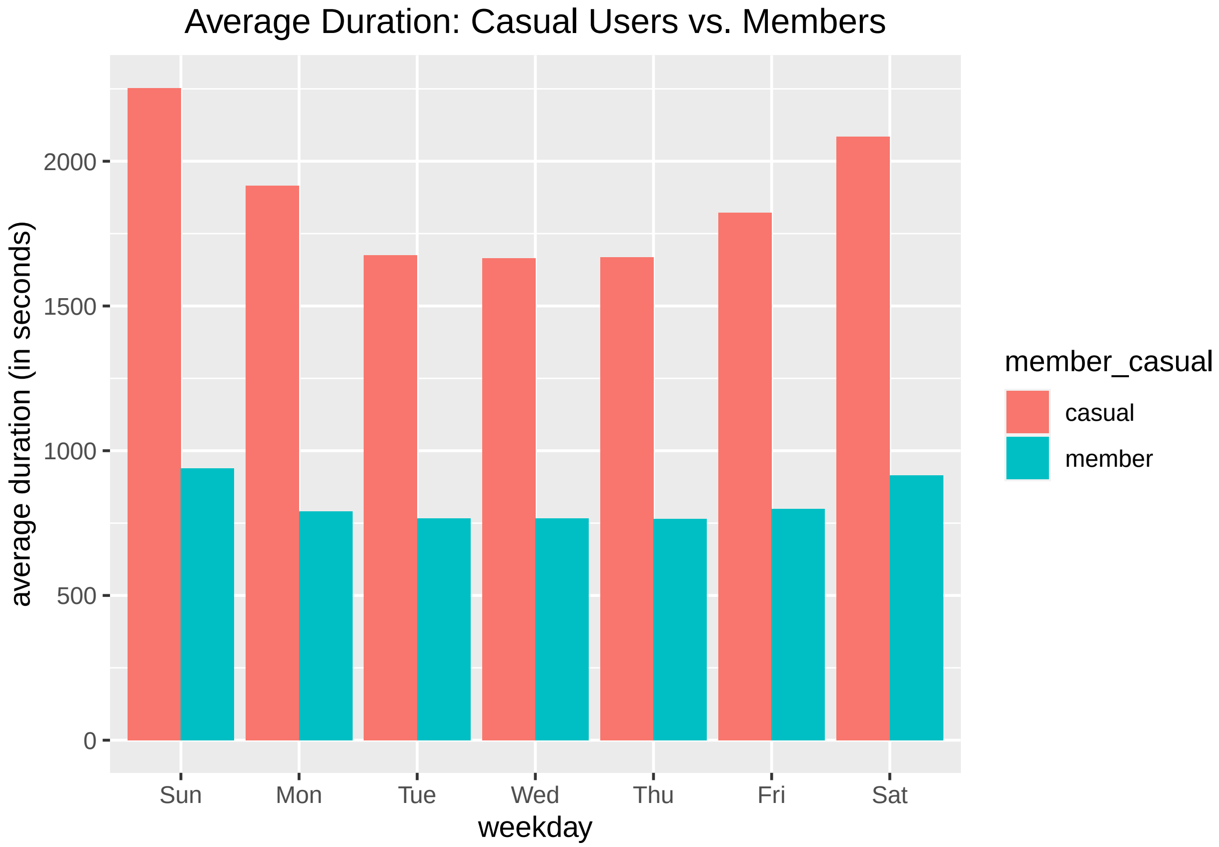
```
# Let's visualize the number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = 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")
# Let's create a visualization for average duration
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") + labs(y = "average duration (in seconds)" ,title = "Average Duration: Ca
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `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
```

STEP 5: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

```
# Create a csv file
counts <- aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, F
#write.csv(counts, file = "D:/Career/Google Data Analytics Program/Case Study/Google Data Analytics Cer
# Choose the file path
```

```
# Set a data table to extract the needed data for Student T-Test
data1 <- setDT(all_trips_v2)[,.(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 = 9.6194, num df = 6, denom df = 6, p-value = 0.01443
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.652891 55.982687
## sample estimates:
## ratio of variances
##           9.619421
```

```
# 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 = 11.48, df = 7.2341, p-value = 3.326e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  876.7823      Inf
## sample estimates:
## mean of x mean of y
## 1869.4737  820.4071
```

```
# 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.
```

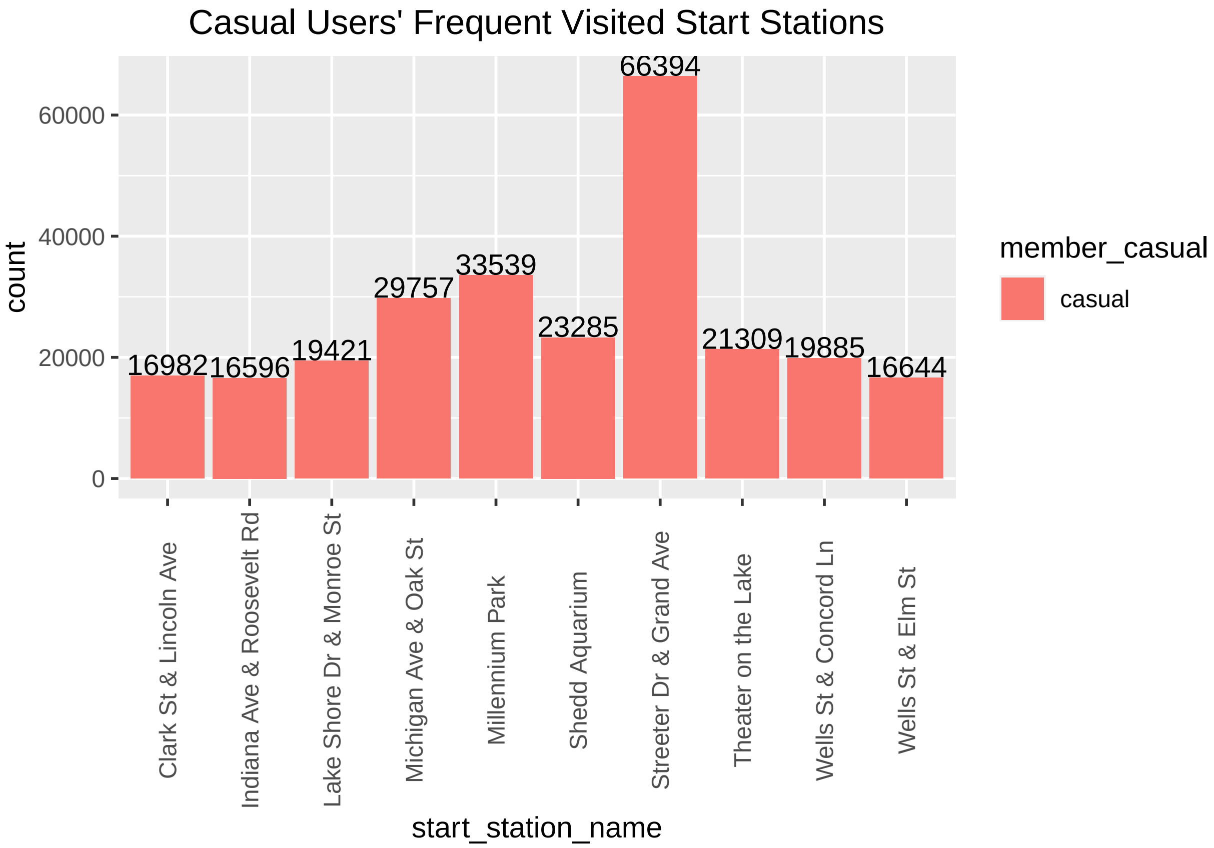
```r
#write.csv(location_counts, file = "D:/Career/Google Data Analytics Program/Case Study/Google Data Anal
```

```r
# 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.
```

```r
Top_10_casual <- Freq_start_station[Freq_start_station$member_casual == "casual",
                                   ][2:11,]
Top_10_member <- Freq_start_station[Freq_start_station$member_casual == "member",
                                   ][2:11,]
```

```r
# Summary Tables & 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")
```