Cyclist Case Study Analysis

Malay Raj

2023-03-02

## Divvy\_Exercise\_Full\_Year\_Analysis

This analysis is based on the Divvy case study “‘Sophisticated, Clear, and Polished’: Divvy and Data Visualization”.

The purpose of this script is to consolidate downloaded Divvy data into a single dataframe and then conduct simple analysis to help answer the key question: “In what ways do members and casual riders use Divvy bikes differently?”

## Installing required packages

tidyverse for data import and wrangling lubridate for date functions ggplot for visualization

library(tidyverse) #helps wrangle data

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(lubridate) #helps wrangle date attributes  
library(ggplot2) #helps visualize data  
getwd() #displays your working directory

## [1] "C:/Users/KIIT/OneDrive/Documents"

## Setting up working directory

setwd("C:/Users/KIIT/OneDrive/Documents")

## STEP 1: COLLECT DATA

Uploading Divvy datasets (csv files) here:

q2\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202202-divvy-tripdata.csv")

## Rows: 190301 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.

q3\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202203-divvy-tripdata.csv")

## Rows: 284042 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.

q4\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202204-divvy-tripdata.csv")

## Rows: 371249 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.

q5\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202205-divvy-tripdata.csv")

## Rows: 634858 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.

q6\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202206-divvy-tripdata.csv")

## Rows: 769204 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.

q7\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202207-divvy-tripdata.csv")

## Rows: 823488 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.

q8\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202208-divvy-tripdata.csv")

## Rows: 785932 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.

q9\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202209-divvy-publictripdata.csv")

## Rows: 701339 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.

q10\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202210-divvy-tripdata.csv")

## Rows: 558685 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.

q11\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202211-divvy-tripdata.csv")

## Rows: 337735 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.

q12\_2022 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202212-divvy-tripdata.csv")

## Rows: 181806 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.

q01\_2023 <- read\_csv("C:/Users/KIIT/OneDrive/Documents/Datasets/202301-divvy-tripdata.csv")

## Rows: 190301 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: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

Comparing column names each of the files. While the names don’t have to be in the same order, they DO need to match perfectly before we can use a command to join them into one file.

colnames(q2\_2022)

## [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(q3\_2022)

## [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(q4\_2022)

## [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(q5\_2022)

## [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(q6\_2022)

## [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(q7\_2022)

## [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(q8\_2022)

## [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(q9\_2022)

## [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(q10\_2022)

## [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(q11\_2022)

## [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(q12\_2022)

## [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(q01\_2023)

## [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"

##Inspecting the dataframes and look for incongruencies

str(q2\_2022)

## spc\_tbl\_ [190,301 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:190301] "8A18EBD4A724A351" "6FAF1E235BBDA252" "D915F47D01416C18" "573590CDF9DE476D" ...  
## $ rideable\_type : chr [1:190301] "electric\_bike" "electric\_bike" "classic\_bike" "electric\_bike" ...  
## $ started\_at : chr [1:190301] "31-01-2023 23:56" "31-01-2023 23:53" "31-01-2023 23:53" "31-01-2023 23:52" ...  
## $ ended\_at : chr [1:190301] "01-02-2023 00:03" "31-01-2023 23:57" "31-01-2023 23:57" "31-01-2023 23:58" ...  
## $ start\_station\_name: chr [1:190301] NA NA "Paulina St & 18th St" NA ...  
## $ start\_station\_id : chr [1:190301] NA NA "TA1307000159" NA ...  
## $ end\_station\_name : chr [1:190301] NA "Western Ave & Winnebago Ave" "Racine Ave & 18th St" "Western Ave & Berteau Ave" ...  
## $ end\_station\_id : chr [1:190301] NA "13068" "13164" "438" ...  
## $ start\_lat : num [1:190301] 42 41.9 41.9 42 41.9 ...  
## $ start\_lng : num [1:190301] -87.7 -87.7 -87.7 -87.7 -87.6 ...  
## $ end\_lat : num [1:190301] 42 41.9 41.9 42 41.9 ...  
## $ end\_lng : num [1:190301] -87.7 -87.7 -87.7 -87.7 -87.6 ...  
## $ member\_casual : chr [1:190301] "member" "member" "member" "member" ...  
## - 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(q3\_2022)

## spc\_tbl\_ [284,042 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:284042] "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F446FD9DEE3F389" ...  
## $ rideable\_type : chr [1:284042] "classic\_bike" "electric\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:284042], format: "2022-03-21 13:45:01" "2022-03-16 09:37:16" ...  
## $ ended\_at : POSIXct[1:284042], format: "2022-03-21 13:51:18" "2022-03-16 09:43:34" ...  
## $ start\_station\_name: chr [1:284042] "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & Berwyn Ave" "Wabash Ave & Wacker Pl" ...  
## $ start\_station\_id : chr [1:284042] "TA1307000131" "13042" "13109" "TA1307000131" ...  
## $ end\_station\_name : chr [1:284042] "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apts)" "Broadway & Ridge Ave" "Franklin St & Jackson Blvd" ...  
## $ end\_station\_id : chr [1:284042] "KA1503000043" "620" "15578" "TA1305000025" ...  
## $ start\_lat : num [1:284042] 41.9 41.9 42 41.9 41.9 ...  
## $ start\_lng : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:284042] 41.9 41.9 42 41.9 41.9 ...  
## $ end\_lng : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.7 ...  
## $ member\_casual : chr [1:284042] "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(q4\_2022)

## spc\_tbl\_ [371,249 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:371249] "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F07FF" "84D4751AEB31888D" ...  
## $ rideable\_type : chr [1:371249] "electric\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:371249], format: "2022-04-06 17:42:48" "2022-04-24 19:23:07" ...  
## $ ended\_at : POSIXct[1:371249], format: "2022-04-06 17:54:36" "2022-04-24 19:43:17" ...  
## $ start\_station\_name: chr [1:371249] "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted St & Polk St" "Wentworth Ave & Cermak Rd" ...  
## $ start\_station\_id : chr [1:371249] "515" "13075" "TA1307000121" "13075" ...  
## $ end\_station\_name : chr [1:371249] "University Library (NU)" "Green St & Madison St" "Green St & Madison St" "Delano Ct & Roosevelt Rd" ...  
## $ end\_station\_id : chr [1:371249] "605" "TA1307000120" "TA1307000120" "KA1706005007" ...  
## $ start\_lat : num [1:371249] 42 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:371249] 42.1 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:371249] "member" "member" "member" "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(q5\_2022)

## spc\_tbl\_ [634,858 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF59852924528F8" ...  
## $ rideable\_type : chr [1:634858] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...  
## $ ended\_at : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...  
## $ start\_station\_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St" "Clinton St & Madison St" "Clinton St & Madison St" ...  
## $ start\_station\_id : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...  
## $ end\_station\_name : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood St & Milwaukee Ave" "Clark St & Randolph St" ...  
## $ end\_station\_id : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...  
## $ start\_lat : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...  
## $ member\_casual : chr [1:634858] "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(q6\_2022)

## spc\_tbl\_ [769,204 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:769204] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD771D2" "C9C320375DE1D5C6" ...  
## $ rideable\_type : chr [1:769204] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:769204], format: "2022-06-30 17:27:53" "2022-06-30 18:39:52" ...  
## $ ended\_at : POSIXct[1:769204], format: "2022-06-30 17:35:15" "2022-06-30 18:47:28" ...  
## $ start\_station\_name: chr [1:769204] NA NA NA NA ...  
## $ start\_station\_id : chr [1:769204] NA NA NA NA ...  
## $ end\_station\_name : chr [1:769204] NA NA NA NA ...  
## $ end\_station\_id : chr [1:769204] NA NA NA NA ...  
## $ start\_lat : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...  
## $ start\_lng : num [1:769204] -87.6 -87.6 -87.7 -87.7 -87.6 ...  
## $ end\_lat : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...  
## $ end\_lng : num [1:769204] -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ member\_casual : chr [1:769204] "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(q7\_2022)

## spc\_tbl\_ [823,488 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:823488] "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6BE44314590E6" ...  
## $ rideable\_type : chr [1:823488] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:823488], format: "2022-07-05 08:12:47" "2022-07-26 12:53:38" ...  
## $ ended\_at : POSIXct[1:823488], format: "2022-07-05 08:24:32" "2022-07-26 12:55:31" ...  
## $ start\_station\_name: chr [1:823488] "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)" ...  
## $ start\_station\_id : chr [1:823488] "13224" "15541" "15541" "15541" ...  
## $ end\_station\_name : chr [1:823488] "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan Ave & 8th St" "Woodlawn Ave & 55th St" ...  
## $ end\_station\_id : chr [1:823488] "KA1503000043" "623" "623" "TA1307000164" ...  
## $ start\_lat : num [1:823488] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:823488] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:823488] 41.9 41.9 41.9 41.8 41.9 ...  
## $ end\_lng : num [1:823488] -87.6 -87.6 -87.6 -87.6 -87.7 ...  
## $ member\_casual : chr [1:823488] "member" "casual" "casual" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_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(q8\_2022)

## spc\_tbl\_ [785,932 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:785932] "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB7FD" "F597830181C2E13C" ...  
## $ rideable\_type : chr [1:785932] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:785932], format: "2022-08-07 21:34:15" "2022-08-08 14:39:21" ...  
## $ ended\_at : POSIXct[1:785932], format: "2022-08-07 21:41:46" "2022-08-08 14:53:23" ...  
## $ start\_station\_name: chr [1:785932] NA NA NA NA ...  
## $ start\_station\_id : chr [1:785932] NA NA NA NA ...  
## $ end\_station\_name : chr [1:785932] NA NA NA NA ...  
## $ end\_station\_id : chr [1:785932] NA NA NA NA ...  
## $ start\_lat : num [1:785932] 41.9 41.9 42 41.9 41.9 ...  
## $ start\_lng : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...  
## $ end\_lat : num [1:785932] 41.9 41.9 42 42 41.8 ...  
## $ end\_lng : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...  
## $ member\_casual : chr [1:785932] "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(q9\_2022)

## spc\_tbl\_ [701,339 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:701339] "5156990AC19CA285" "E12D4A16BF51C274" "A02B53CD7DB72DD7" "C82E05FEE872DF11" ...  
## $ rideable\_type : chr [1:701339] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:701339], format: "2022-09-01 08:36:22" "2022-09-01 17:11:29" ...  
## $ ended\_at : POSIXct[1:701339], format: "2022-09-01 08:39:05" "2022-09-01 17:14:45" ...  
## $ start\_station\_name: chr [1:701339] NA NA NA NA ...  
## $ start\_station\_id : chr [1:701339] NA NA NA NA ...  
## $ end\_station\_name : chr [1:701339] "California Ave & Milwaukee Ave" NA NA NA ...  
## $ end\_station\_id : chr [1:701339] "13084" NA NA NA ...  
## $ start\_lat : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:701339] "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(q10\_2022)

## spc\_tbl\_ [558,685 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:558685] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E653FDC2D99236" ...  
## $ rideable\_type : chr [1:558685] "classic\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:558685], format: "2022-10-14 17:13:30" "2022-10-01 16:29:26" ...  
## $ ended\_at : POSIXct[1:558685], format: "2022-10-14 17:19:39" "2022-10-01 16:49:06" ...  
## $ start\_station\_name: chr [1:558685] "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne Ave & Balmoral Ave" "Rush St & Cedar St" ...  
## $ start\_station\_id : chr [1:558685] "13290" "13288" "655" "KA1504000133" ...  
## $ end\_station\_name : chr [1:558685] "Larrabee St & Division St" "Damen Ave & Cullerton St" "Western Ave & Leland Ave" "Orleans St & Chestnut St (NEXT Apts)" ...  
## $ end\_station\_id : chr [1:558685] "KA1504000079" "13089" "TA1307000140" "620" ...  
## $ start\_lat : num [1:558685] 41.9 41.9 42 41.9 41.9 ...  
## $ start\_lng : num [1:558685] -87.7 -87.7 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:558685] 41.9 41.9 42 41.9 41.9 ...  
## $ end\_lng : num [1:558685] -87.6 -87.7 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:558685] "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(q11\_2022)

## spc\_tbl\_ [337,735 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:337735] "BCC66FC6FAB27CC7" "772AB67E902C180F" "585EAD07FDEC0152" "91C4E7ED3C262FF9" ...  
## $ rideable\_type : chr [1:337735] "electric\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:337735], format: "2022-11-10 06:21:55" "2022-11-04 07:31:55" ...  
## $ ended\_at : POSIXct[1:337735], format: "2022-11-10 06:31:27" "2022-11-04 07:46:25" ...  
## $ start\_station\_name: chr [1:337735] "Canal St & Adams St" "Canal St & Adams St" "Indiana Ave & Roosevelt Rd" "Indiana Ave & Roosevelt Rd" ...  
## $ start\_station\_id : chr [1:337735] "13011" "13011" "SL-005" "SL-005" ...  
## $ end\_station\_name : chr [1:337735] "St. Clair St & Erie St" "St. Clair St & Erie St" "St. Clair St & Erie St" "St. Clair St & Erie St" ...  
## $ end\_station\_id : chr [1:337735] "13016" "13016" "13016" "13016" ...  
## $ start\_lat : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:337735] "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(q12\_2022)

## spc\_tbl\_ [181,806 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:181806] "65DBD2F447EC51C2" "0C201AA7EA0EA1AD" "E0B148CCB358A49D" "54C5775D2B7C9188" ...  
## $ rideable\_type : chr [1:181806] "electric\_bike" "classic\_bike" "electric\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:181806], format: "2022-12-05 10:47:18" "2022-12-18 06:42:33" ...  
## $ ended\_at : POSIXct[1:181806], format: "2022-12-05 10:56:34" "2022-12-18 07:08:44" ...  
## $ start\_station\_name: chr [1:181806] "Clifton Ave & Armitage Ave" "Broadway & Belmont Ave" "Sangamon St & Lake St" "Shields Ave & 31st St" ...  
## $ start\_station\_id : chr [1:181806] "TA1307000163" "13277" "TA1306000015" "KA1503000038" ...  
## $ end\_station\_name : chr [1:181806] "Sedgwick St & Webster Ave" "Sedgwick St & Webster Ave" "St. Clair St & Erie St" "Damen Ave & Madison St" ...  
## $ end\_station\_id : chr [1:181806] "13191" "13191" "13016" "13134" ...  
## $ start\_lat : num [1:181806] 41.9 41.9 41.9 41.8 41.9 ...  
## $ start\_lng : num [1:181806] -87.7 -87.6 -87.7 -87.6 -87.7 ...  
## $ end\_lat : num [1:181806] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:181806] -87.6 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:181806] "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(q01\_2023)

## spc\_tbl\_ [190,301 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:190301] "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C90792D034FED968" ...  
## $ rideable\_type : chr [1:190301] "electric\_bike" "classic\_bike" "electric\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:190301], format: "2023-01-21 20:05:42" "2023-01-10 15:37:36" ...  
## $ ended\_at : POSIXct[1:190301], format: "2023-01-21 20:16:33" "2023-01-10 15:46:05" ...  
## $ start\_station\_name: chr [1:190301] "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Western Ave & Lunt Ave" "Kimbark Ave & 53rd St" ...  
## $ start\_station\_id : chr [1:190301] "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...  
## $ end\_station\_name : chr [1:190301] "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli Produce - Evanston Plaza" "Greenwood Ave & 47th St" ...  
## $ end\_station\_id : chr [1:190301] "202480.0" "TA1308000002" "599" "TA1308000002" ...  
## $ start\_lat : num [1:190301] 41.9 41.8 42 41.8 41.8 ...  
## $ start\_lng : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:190301] 41.9 41.8 42 41.8 41.8 ...  
## $ end\_lng : num [1:190301] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:190301] "member" "member" "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>

##Converting ride\_id and rideable\_type to character so that they can stack correctly:

q2\_2022 <- mutate(q2\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q3\_2022 <- mutate(q3\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q4\_2022 <- mutate(q4\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))  
q5\_2022 <- mutate(q5\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q6\_2022 <- mutate(q6\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q7\_2022 <- mutate(q7\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))  
q8\_2022 <- mutate(q8\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q9\_2022 <- mutate(q9\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q10\_2022 <- mutate(q10\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))  
q11\_2022 <- mutate(q11\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q12\_2022 <- mutate(q12\_2022, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))   
q01\_2023 <- mutate(q01\_2023, ride\_id = as.character(ride\_id)  
 ,rideable\_type = as.character(rideable\_type))

##Converting started\_at and ended\_at of q2\_2022 to datetime so that they can stack correctly

q2\_2022 <- mutate(q2\_2022, started\_at = as\_datetime(started\_at)  
 ,ended\_at = as\_datetime(ended\_at))

##Stacking the individual quarter’s data frames into one big data frame.

all\_trips <- bind\_rows(q2\_2022, q3\_2022, q4\_2022, q5\_2022, q6\_2022, q7\_2022, q8\_2022, q9\_2022, q10\_2022, q11\_2022, q12\_2022, q01\_2023)

##Removing few columns such as:- start\_lat start\_lng end\_lat end\_lng because we don’t need them in our 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 Inspecting 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) #How many rows are in data frame?

## [1] 5828940

dim(all\_trips) #Dimensions of the data frame?

## [1] 5828940 9

head(all\_trips) #See the first 6 rows of data frame

## # A tibble: 6 × 9  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 8A18EBD4A724A… electr… 2031-01-20 23:23:56 2001-02-20 23:00:03 <NA> <NA>   
## 2 6FAF1E235BBDA… electr… 2031-01-20 23:23:53 2031-01-20 23:23:57 <NA> <NA>   
## 3 D915F47D01416… classi… 2031-01-20 23:23:53 2031-01-20 23:23:57 Paulin… TA1307…  
## 4 573590CDF9DE4… electr… 2031-01-20 23:23:52 2031-01-20 23:23:58 <NA> <NA>   
## 5 56FF83E0D18B8… electr… 2031-01-20 23:23:51 2031-01-20 23:23:59 Wells … TA1307…  
## 6 03C7B7A14C694… classi… 2031-01-20 23:23:51 2031-01-20 23:23:55 Univer… KA1503…  
## # … with 3 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # member\_casual <chr>, and abbreviated variable names ¹​rideable\_type,  
## # ²​start\_station\_name, ³​start\_station\_id

tail(all\_trips) #See the last 6 rows of data frame

## # A tibble: 6 × 9  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 A3DC3E8358DB1… electr… 2023-01-17 18:36:00 2023-01-17 19:00:26 Clark … TA1307…  
## 2 A303816F2E8A3… electr… 2023-01-11 17:46:23 2023-01-11 17:57:31 Clark … TA1307…  
## 3 BCDBB142CC610… classi… 2023-01-30 15:08:10 2023-01-30 15:33:26 Wester… TA1307…  
## 4 7D1C7CA805171… classi… 2023-01-06 19:34:50 2023-01-06 19:50:01 Clark … TA1307…  
## 5 1A4EB636346DF… classi… 2023-01-13 18:59:24 2023-01-13 19:14:44 Clark … TA1307…  
## 6 069971675AC7D… electr… 2023-01-02 13:48:29 2023-01-02 13:59:29 Clark … TA1307…  
## # … with 3 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # member\_casual <chr>, and abbreviated variable names ¹​rideable\_type,  
## # ²​start\_station\_name, ³​start\_station\_id

str(all\_trips) #See list of columns and data types (numeric, character, etc)

## tibble [5,828,940 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5828940] "8A18EBD4A724A351" "6FAF1E235BBDA252" "D915F47D01416C18" "573590CDF9DE476D" ...  
## $ rideable\_type : chr [1:5828940] "electric\_bike" "electric\_bike" "classic\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:5828940], format: "2031-01-20 23:23:56" "2031-01-20 23:23:53" ...  
## $ ended\_at : POSIXct[1:5828940], format: "2001-02-20 23:00:03" "2031-01-20 23:23:57" ...  
## $ start\_station\_name: chr [1:5828940] NA NA "Paulina St & 18th St" NA ...  
## $ start\_station\_id : chr [1:5828940] NA NA "TA1307000159" NA ...  
## $ end\_station\_name : chr [1:5828940] NA "Western Ave & Winnebago Ave" "Racine Ave & 18th St" "Western Ave & Berteau Ave" ...  
## $ end\_station\_id : chr [1:5828940] NA "13068" "13164" "438" ...  
## $ member\_casual : chr [1:5828940] "member" "member" "member" "member" ...

summary(all\_trips) #Statistical summary of data. Mainly for numerics.

## ride\_id rideable\_type started\_at   
## Length:5828940 Length:5828940 Min. :2001-01-20 23:00:01.00   
## Class :character Class :character 1st Qu.:2022-06-01 19:30:31.50   
## Mode :character Mode :character Median :2022-07-28 08:28:12.50   
## Mean :2022-05-01 17:09:44.88   
## 3rd Qu.:2022-09-23 22:21:31.75   
## Max. :2031-01-20 23:23:56.00   
## ended\_at start\_station\_name start\_station\_id   
## Min. :2001-01-20 23:00:02.00 Length:5828940 Length:5828940   
## 1st Qu.:2022-06-01 19:48:40.75 Class :character Class :character   
## Median :2022-07-28 08:41:15.50 Mode :character Mode :character   
## Mean :2022-05-01 18:09:33.58   
## 3rd Qu.:2022-09-23 22:39:04.50   
## Max. :2031-01-20 23:23:59.00   
## end\_station\_name end\_station\_id member\_casual   
## Length:5828940 Length:5828940 Length:5828940   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

## Checking number of member and casual riders.

table(all\_trips$member\_casual)

##   
## casual member   
## 2362112 3466828

##There are a few problems we will need to fix:

1. Add columns that list the date, month, day, and year of each ride This will allow us to aggregate ride data for each month, day, or year … before completing these operations we could only aggregate at the ride level.
2. We will want to add a calculated field for length of ride.We will add “ride\_length” to the entire dataframe.

all\_trips$date <- as.Date(all\_trips$started\_at) #The default format is yyyy-mm-dd  
all\_trips$month <- format(as.Date(all\_trips$date), "%m")  
all\_trips$day <- format(as.Date(all\_trips$date), "%d")  
all\_trips$year <- format(as.Date(all\_trips$date), "%Y")  
all\_trips$day\_of\_week <- format(as.Date(all\_trips$date), "%A")  
  
all\_trips$ride\_length <- difftime(all\_trips$ended\_at,all\_trips$started\_at)

## Inspecting the structure of the columns:

str(all\_trips)

## tibble [5,828,940 × 15] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5828940] "8A18EBD4A724A351" "6FAF1E235BBDA252" "D915F47D01416C18" "573590CDF9DE476D" ...  
## $ rideable\_type : chr [1:5828940] "electric\_bike" "electric\_bike" "classic\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:5828940], format: "2031-01-20 23:23:56" "2031-01-20 23:23:53" ...  
## $ ended\_at : POSIXct[1:5828940], format: "2001-02-20 23:00:03" "2031-01-20 23:23:57" ...  
## $ start\_station\_name: chr [1:5828940] NA NA "Paulina St & 18th St" NA ...  
## $ start\_station\_id : chr [1:5828940] NA NA "TA1307000159" NA ...  
## $ end\_station\_name : chr [1:5828940] NA "Western Ave & Winnebago Ave" "Racine Ave & 18th St" "Western Ave & Berteau Ave" ...  
## $ end\_station\_id : chr [1:5828940] NA "13068" "13164" "438" ...  
## $ member\_casual : chr [1:5828940] "member" "member" "member" "member" ...  
## $ date : Date[1:5828940], format: "2031-01-20" "2031-01-20" ...  
## $ month : chr [1:5828940] "01" "01" "01" "01" ...  
## $ day : chr [1:5828940] "20" "20" "20" "20" ...  
## $ year : chr [1:5828940] "2031" "2031" "2031" "2031" ...  
## $ day\_of\_week : chr [1:5828940] "Monday" "Monday" "Monday" "Monday" ...  
## $ ride\_length : 'difftime' num [1:5828940] -944007833 4 4 6 ...  
## ..- attr(\*, "units")= chr "secs"

## Converting “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

## REMOVING BAD DATA

## There are some rides where tripduration shows up as negative, including several hundred rides where Divvy took bikes out of circulation for Quality Control reasons. We will want to delete these rides.

We will create a new version of the dataframe (v2) since data is being removed.

all\_trips\_v2 <- all\_trips[!(all\_trips$start\_station\_name == "HQ QR" | all\_trips$ride\_length<0),]

## STEP 4: CONDUCTING DESCRIPTIVE ANALYSIS

Descriptive analysis on ride\_length (all figures in seconds).

mean(all\_trips\_v2$ride\_length) #straight average (total ride length / rides)

## [1] NA

median(all\_trips\_v2$ride\_length) #midpoint number in the ascending array of ride lengths

## [1] NA

max(all\_trips\_v2$ride\_length) #longest ride

## [1] NA

min(all\_trips\_v2$ride\_length) #shortest ride

## [1] NA

## Condensing the four lines above to one line using summary() on the specific attribute

summary(all\_trips\_v2$ride\_length)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0 334 607 5273 1105 725846884 851641

## Comparing 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 8191.029  
## 2 member 3296.375

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 795  
## 2 member 511

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 725846884  
## 2 member 31622460

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

## 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 4114.127  
## 2 member Friday 3372.714  
## 3 casual Monday 8433.234  
## 4 member Monday 3647.459  
## 5 casual Saturday 7899.775  
## 6 member Saturday 3577.492  
## 7 casual Sunday 11851.658  
## 8 member Sunday 4948.374  
## 9 casual Thursday 8053.091  
## 10 member Thursday 2530.686  
## 11 casual Tuesday 8940.811  
## 12 member Tuesday 2594.457  
## 13 casual Wednesday 7568.468  
## 14 member Wednesday 2936.700

## As we can 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", "Wednesday", "Thursday", "Friday", "Saturday"))

## Now, we’ll 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 Sunday 11851.658  
## 2 member Sunday 4948.374  
## 3 casual Monday 8433.234  
## 4 member Monday 3647.459  
## 5 casual Tuesday 8940.811  
## 6 member Tuesday 2594.457  
## 7 casual Wednesday 7568.468  
## 8 member Wednesday 2936.700  
## 9 casual Thursday 8053.091  
## 10 member Thursday 2530.686  
## 11 casual Friday 4114.127  
## 12 member Friday 3372.714  
## 13 casual Saturday 7899.775  
## 14 member Saturday 3577.492

## Analyzing ridership data by type and weekday

all\_trips\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 dplyr::summarise(number\_of\_rides = dplyr::n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, weekday)

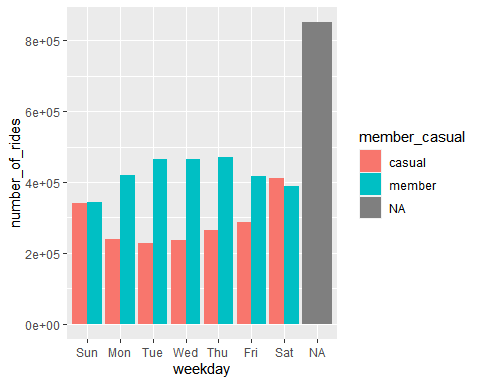
## `summarise()` has grouped output by 'member\_casual'. You can override using the  
## `.groups` argument.

## # A tibble: 15 × 4  
## # Groups: member\_casual [3]  
## member\_casual weekday number\_of\_rides average\_duration  
## <chr> <ord> <int> <dbl>  
## 1 casual Sun 339501 11852.  
## 2 casual Mon 240212 8433.  
## 3 casual Tue 229108 8941.  
## 4 casual Wed 236691 7568.  
## 5 casual Thu 265230 8053.  
## 6 casual Fri 286686 4114.  
## 7 casual Sat 412075 7900.  
## 8 member Sun 342710 4948.  
## 9 member Mon 418941 3647.  
## 10 member Tue 466187 2594.  
## 11 member Wed 466092 2937.  
## 12 member Thu 469343 2531.  
## 13 member Fri 415756 3373.  
## 14 member Sat 388658 3577.  
## 15 <NA> <NA> 851641 NA

## 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) %>%   
 dplyr::summarise(number\_of\_rides = dplyr::n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, weekday) %>%   
 ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge")

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

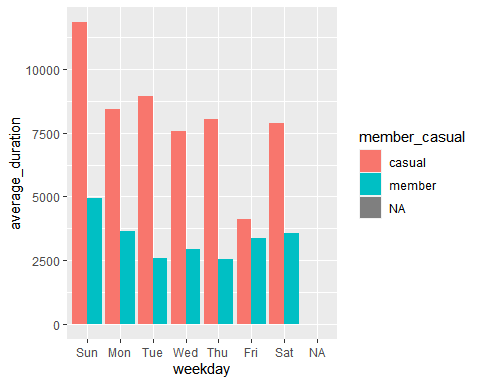


## Let’s create a visualization for average duration

all\_trips\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 dplyr::summarise(number\_of\_rides = dplyr::n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, weekday) %>%   
 ggplot(aes(x = weekday, y = average\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge")

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

## Warning: Removed 1 rows containing missing values (`geom\_col()`).



## STEP 5: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

Creating a csv file that we will visualize for further analysis.

counts <- aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$day\_of\_week, FUN = mean)  
  
write.csv(counts, file = 'C:/Users/KIIT/OneDrive/Documents/avg\_ride\_length.csv')