Google Data Analytics Capstone Project

Audra Roberts

2022-05-28

#Install required packages  
#tidyverse for data import and wrangling  
#lubridate for date functions  
#ggplot2 for visualization  
#janior for data cleaning  
  
library(tidyverse)

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

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

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(ggplot2)  
library(janitor)

##   
## Attaching package: 'janitor'

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

#Display my working directory  
getwd()

## [1] "C:/Users/audra/Documents/Capstone/Cyclistic"

setwd("C:/Users/audra/Documents/Capstone/Cyclistic/dataset/Excel Workbook")  
  
#Step 1: Collect Data  
#Upload Divvy datasets (.xlsx files)  
  
mar\_2021 <- readxl::read\_xlsx("202103-divvy-tripdata.xlsx")  
apr\_2021 <- readxl::read\_xlsx("202104-divvy-tripdata.xlsx")  
may\_2021 <- readxl::read\_xlsx("202105-divvy-tripdata.xlsx")  
jun\_2021 <- readxl::read\_xlsx("202106-divvy-tripdata.xlsx")  
jul\_2021 <- readxl::read\_xlsx("202107-divvy-tripdata.xlsx")  
aug\_2021 <- readxl::read\_xlsx("202108-divvy-tripdata.xlsx")  
sept\_2021 <- readxl::read\_xlsx("202109-divvy-tripdata.xlsx")  
oct\_2021 <- readxl::read\_xlsx("202110-divvy-tripdata.xlsx")  
nov\_2021 <- readxl::read\_xlsx("202111-divvy-tripdata.xlsx")  
dec\_2021 <- readxl::read\_xlsx("202112-divvy-tripdata.xlsx")  
jan\_2022 <- readxl::read\_xlsx("202201-divvy-tripdata.xlsx")  
feb\_2022 <- readxl::read\_xlsx("202202-divvy-tripdata.xlsx")  
  
#Step 2: Wrangle Data and Combine into a Single File  
#Viewed each data frame, and all column names matched and were in order  
#Inspect the data frames and look for inconsistencies

str(mar\_2021)

## tibble [228,496 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168F2" ...  
## $ rideable\_type : chr [1:228496] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:228496], format: "2021-03-16 08:32:30" "2021-03-28 01:26:28" ...  
## $ 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" "Shields Ave & 28th Pl" "Winthrop Ave & Lawrence Ave" ...  
## $ start\_station\_id : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...  
## $ end\_station\_name : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted St & 35th St" "Broadway & Sheridan Rd" ...  
## $ end\_station\_id : chr [1:228496] "13266" "18017" "TA1308000043" "13323" ...  
## $ start\_lat : num [1:228496] 41.9 41.9 41.8 42 42 ...  
## $ start\_lng : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num [1:228496] 41.9 41.9 41.8 42 42.1 ...  
## $ end\_lng : num [1:228496] -87.7 -87.7 -87.6 -87.6 -87.7 ...  
## $ member\_casual : chr [1:228496] "casual" "casual" "casual" "casual" ...

str(apr\_2021)

## tibble [337,230 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:337230] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887262AD101C604" ...  
## $ 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 Blvd & 84th St" "Honore St & Division St" ...  
## $ start\_station\_id : chr [1:337230] "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...  
## $ end\_station\_name : chr [1:337230] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loomis Blvd & 84th St" "Southport Ave & Waveland Ave" ...  
## $ end\_station\_id : chr [1:337230] "13235" "KA1503000069" "20121" "13235" ...  
## $ start\_lat : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...  
## $ start\_lng : num [1:337230] -87.6 -87.6 -87.7 -87.7 -87.7 ...  
## $ end\_lat : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...  
## $ end\_lng : num [1:337230] -87.7 -87.6 -87.7 -87.7 -87.7 ...  
## $ member\_casual : chr [1:337230] "member" "casual" "casual" "member" ...

str(may\_2021)

## tibble [531,633 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881AC6D39110C60" ...  
## $ 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" ...

str(jun\_2021)

## tibble [729,595 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:729595] "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C0FE48C412214" ...  
## $ 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" ...

str(jul\_2021)

## tibble [822,410 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B58EAB20E8AA5" ...  
## $ 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" "Wabash Ave & 16th St" "California Ave & Cortez St" ...  
## $ 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 & Hubbard St" "Carpenter St & Huron 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" ...

str(aug\_2021)

## tibble [804,352 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834D3208BFAF1DA" ...  
## $ 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" ...

str(sept\_2021)

## tibble [756,147 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1DE133B3DBF55" ...  
## $ 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" ...

str(oct\_2021)

## tibble [631,226 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:631226] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "362947F0437E1514" ...  
## $ 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" ...

str(nov\_2021)

## tibble [359,978 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:359978] "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3BE33085BCFF02" ...  
## $ 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" ...

str(dec\_2021)

## tibble [247,540 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:247540] "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278BA87BF698339" ...  
## $ 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 & North Branch St" "Halsted St & North Branch 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 & Barry Ave" "LaSalle Dr & Huron St" ...  
## $ 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" ...

str(jan\_2022)

## tibble [103,770 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:103770] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB80ED419105406" ...  
## $ rideable\_type : chr [1:103770] "electric\_bike" "electric\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:103770], format: "2022-01-13 11:59:47" "2022-01-10 08:41:56" ...  
## $ ended\_at : POSIXct[1:103770], format: "2022-01-13 12:02:44" "2022-01-10 08:46:17" ...  
## $ start\_station\_name: chr [1:103770] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffield Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...  
## $ 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 & Fullerton Ave" "Paulina St & Montrose 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" ...

str(feb\_2022)

## tibble [115,609 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:115609] "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A1789BDF844412BE" ...  
## $ rideable\_type : chr [1:115609] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:115609], format: "2022-02-19 18:08:41" "2022-02-20 17:41:30" ...  
## $ ended\_at : POSIXct[1:115609], format: "2022-02-19 18:23:56" "2022-02-20 17:45:56" ...  
## $ start\_station\_name: chr [1:115609] "State St & Randolph St" "Halsted St & Wrightwood Ave" "State St & Randolph St" "Southport Ave & Waveland Ave" ...  
## $ start\_station\_id : chr [1:115609] "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...  
## $ end\_station\_name : chr [1:115609] "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Canal St & Adams St" "Broadway & Sheridan Rd" ...  
## $ end\_station\_id : chr [1:115609] "13179" "TA1307000113" "13011" "13323" ...  
## $ start\_lat : num [1:115609] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:115609] -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num [1:115609] 41.9 41.9 41.9 42 41.9 ...  
## $ end\_lng : num [1:115609] -87.6 -87.7 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:115609] "member" "member" "member" "member" ...

#Stack each data frame into one, huge data frame  
all\_trips <- rbind(mar\_2021, apr\_2021, may\_2021, jun\_2021, jul\_2021, aug\_2021, sept\_2021, oct\_2021, nov\_2021, dec\_2021, jan\_2022, feb\_2022)  
  
#Make a copy of the data in case I blunder  
all\_trips\_v2 <- all\_trips  
  
#Step 3: Clean Up and Add Data to Prepare for Analysis  
#Add relevant columns to aggregate data

#add a date column  
all\_trips\_v2$date <- as.Date(all\_trips\_v2$started\_at)  
  
#add a month column formatted as mm\_yyyy  
all\_trips\_v2$month <- format(as.Date(all\_trips\_v2$started\_at), "%b\_%y")  
  
#add a day column  
all\_trips\_v2$day <- format(as.Date(all\_trips\_v2$date), "%d")  
  
#add a year column  
all\_trips\_v2$year <- format(as.Date(all\_trips\_v2$date), "%Y")  
  
#add a day of week column  
all\_trips\_v2$weekday <- format(as.Date(all\_trips\_v2$date), "%A")  
  
#add a time started column  
all\_trips\_v2$time <- format(all\_trips\_v2$started\_at, format = "%H:%M")  
  
#change format for the time column for analysis later  
all\_trips\_v2$time <- as.POSIXct(all\_trips\_v2$time, format = "%H:%M")  
  
#calculate the ride length in minutes  
## https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html  
all\_trips\_v2$ride\_length <- as.double(difftime(all\_trips\_v2$ended\_at, all\_trips\_v2$started\_at, units = "mins"))  
   
#Inspect the structure of the columns  
str(all\_trips\_v2)

## tibble [5,667,986 x 20] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5667986] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168F2" ...  
## $ rideable\_type : chr [1:5667986] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:5667986], format: "2021-03-16 08:32:30" "2021-03-28 01:26:28" ...  
## $ ended\_at : POSIXct[1:5667986], format: "2021-03-16 08:36:34" "2021-03-28 01:36:55" ...  
## $ start\_station\_name: chr [1:5667986] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields Ave & 28th Pl" "Winthrop Ave & Lawrence Ave" ...  
## $ start\_station\_id : chr [1:5667986] "15651" "15651" "15443" "TA1308000021" ...  
## $ end\_station\_name : chr [1:5667986] "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted St & 35th St" "Broadway & Sheridan Rd" ...  
## $ end\_station\_id : chr [1:5667986] "13266" "18017" "TA1308000043" "13323" ...  
## $ start\_lat : num [1:5667986] 41.9 41.9 41.8 42 42 ...  
## $ start\_lng : num [1:5667986] -87.7 -87.7 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num [1:5667986] 41.9 41.9 41.8 42 42.1 ...  
## $ end\_lng : num [1:5667986] -87.7 -87.7 -87.6 -87.6 -87.7 ...  
## $ member\_casual : chr [1:5667986] "casual" "casual" "casual" "casual" ...  
## $ date : Date[1:5667986], format: "2021-03-16" "2021-03-28" ...  
## $ month : chr [1:5667986] "Mar\_21" "Mar\_21" "Mar\_21" "Mar\_21" ...  
## $ day : chr [1:5667986] "16" "28" "11" "11" ...  
## $ year : chr [1:5667986] "2021" "2021" "2021" "2021" ...  
## $ weekday : chr [1:5667986] "Tuesday" "Sunday" "Thursday" "Thursday" ...  
## $ time : POSIXct[1:5667986], format: "2022-05-28 08:32:00" "2022-05-28 01:26:00" ...  
## $ ride\_length : num [1:5667986] 4.07 10.45 16.4 28.98 17.93 ...

colnames(all\_trips\_v2)

## [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" "date" "month"   
## [16] "day" "year" "weekday"   
## [19] "time" "ride\_length"

nrow(all\_trips\_v2) #number of rows in data frame

## [1] 5667986

dim(all\_trips\_v2) #dimensions of the data frame

## [1] 5667986 20

head(all\_trips\_v2) #see first 6 rows of data frame

## # A tibble: 6 x 20  
## ride\_id rideable\_type started\_at ended\_at start\_station\_n~  
## <chr> <chr> <dttm> <dttm> <chr>   
## 1 CFA86D~ classic\_bike 2021-03-16 08:32:30 2021-03-16 08:36:34 Humboldt Blvd &~  
## 2 30D9DC~ classic\_bike 2021-03-28 01:26:28 2021-03-28 01:36:55 Humboldt Blvd &~  
## 3 846D87~ classic\_bike 2021-03-11 21:17:29 2021-03-11 21:33:53 Shields Ave & 2~  
## 4 994D05~ classic\_bike 2021-03-11 13:26:42 2021-03-11 13:55:41 Winthrop Ave & ~  
## 5 DF7464~ classic\_bike 2021-03-21 09:09:37 2021-03-21 09:27:33 Glenwood Ave & ~  
## 6 CEBA85~ classic\_bike 2021-03-20 11:08:47 2021-03-20 11:29:39 Glenwood Ave & ~  
## # ... with 15 more variables: start\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_id <chr>, start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>,  
## # end\_lng <dbl>, member\_casual <chr>, date <date>, month <chr>, day <chr>,  
## # year <chr>, weekday <chr>, time <dttm>, ride\_length <dbl>

tail(all\_trips\_v2) #see last 6 rows of data frame

## # A tibble: 6 x 20  
## ride\_id rideable\_type started\_at ended\_at start\_station\_n~  
## <chr> <chr> <dttm> <dttm> <chr>   
## 1 BDEB7A~ electric\_bike 2022-02-25 23:25:25 2022-02-25 23:29:26 <NA>   
## 2 211BE0~ electric\_bike 2022-02-23 17:47:49 2022-02-23 18:02:29 <NA>   
## 3 D4D53E~ electric\_bike 2022-02-04 10:43:47 2022-02-04 10:50:52 <NA>   
## 4 9E85F0~ electric\_bike 2022-02-28 09:16:33 2022-02-28 09:28:11 Wood St & Chica~  
## 5 B61B55~ electric\_bike 2022-02-10 16:55:16 2022-02-10 16:57:53 <NA>   
## 6 841C70~ electric\_bike 2022-02-21 16:35:20 2022-02-21 16:42:53 <NA>   
## # ... with 15 more variables: start\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_id <chr>, start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>,  
## # end\_lng <dbl>, member\_casual <chr>, date <date>, month <chr>, day <chr>,  
## # year <chr>, weekday <chr>, time <dttm>, ride\_length <dbl>

summary(all\_trips\_v2) #statistical summary of data

## ride\_id rideable\_type started\_at   
## Length:5667986 Length:5667986 Min. :2021-03-01 00:01:09   
## Class :character Class :character 1st Qu.:2021-06-13 11:41:58   
## Mode :character Mode :character Median :2021-08-07 19:12:50   
## Mean :2021-08-10 07:33:56   
## 3rd Qu.:2021-10-02 14:16:11   
## Max. :2022-02-28 23:58:44   
##   
## ended\_at start\_station\_name start\_station\_id   
## Min. :2021-03-01 00:06:28 Length:5667986 Length:5667986   
## 1st Qu.:2021-06-13 12:09:22 Class :character Class :character   
## Median :2021-08-07 19:35:11 Mode :character Mode :character   
## Mean :2021-08-10 07:55:41   
## 3rd Qu.:2021-10-02 14:38:59   
## Max. :2022-03-01 08:55:17   
##   
## end\_station\_name end\_station\_id start\_lat start\_lng   
## Length:5667986 Length:5667986 Min. :41.64 Min. :-87.84   
## Class :character Class :character 1st Qu.:41.88 1st Qu.:-87.66   
## Mode :character Mode :character Median :41.90 Median :-87.64   
## Mean :41.90 Mean :-87.65   
## 3rd Qu.:41.93 3rd Qu.:-87.63   
## Max. :45.64 Max. :-73.80   
##   
## end\_lat end\_lng member\_casual date   
## Min. :41.39 Min. :-88.97 Length:5667986 Min. :2021-03-01   
## 1st Qu.:41.88 1st Qu.:-87.66 Class :character 1st Qu.:2021-06-13   
## Median :41.90 Median :-87.64 Mode :character Median :2021-08-07   
## Mean :41.90 Mean :-87.65 Mean :2021-08-09   
## 3rd Qu.:41.93 3rd Qu.:-87.63 3rd Qu.:2021-10-02   
## Max. :42.17 Max. :-87.49 Max. :2022-02-28   
## NA's :4617 NA's :4617   
## month day year weekday   
## Length:5667986 Length:5667986 Length:5667986 Length:5667986   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## time ride\_length   
## Min. :2022-05-28 00:00:00 Min. : -58.03   
## 1st Qu.:2022-05-28 11:31:00 1st Qu.: 6.67   
## Median :2022-05-28 15:32:00 Median : 11.87   
## Mean :2022-05-28 14:43:29 Mean : 21.75   
## 3rd Qu.:2022-05-28 18:22:00 3rd Qu.: 21.57   
## Max. :2022-05-28 23:59:00 Max. :55944.15   
##

#remove duplicates  
#https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/  
all\_trips\_v2 <- distinct(all\_trips\_v2)  
  
#remove negative rides  
all\_trips\_v2 <- all\_trips\_v2[!all\_trips\_v2$ride\_length < 1,]  
  
#remove rides over 24 hours (1440 minutes)  
all\_trips\_v2 <- all\_trips\_v2[!all\_trips\_v2$ride\_length > 1440,]  
  
#remove rows that have missing data  
all\_trips\_v2 <- all\_trips\_v2[!(is.na(all\_trips\_v2$ride\_id) |   
 is.na(all\_trips\_v2$rideable\_type) |  
 is.na(all\_trips\_v2$started\_at) |  
 is.na(all\_trips\_v2$ended\_at) |  
 is.na(all\_trips\_v2$start\_station\_name) |  
 is.na(all\_trips\_v2$start\_station\_id) |  
 is.na(all\_trips\_v2$end\_station\_name) |  
 is.na(all\_trips\_v2$end\_station\_id) |  
 is.na(all\_trips\_v2$start\_lat) |  
 is.na(all\_trips\_v2$start\_lng) |  
 is.na(all\_trips\_v2$end\_lat) |  
 is.na(all\_trips\_v2$end\_lng)),]  
  
#remove rows that have start and end station names that are associated with quality control  
all\_trips\_v2 <- all\_trips\_v2[!(all\_trips\_v2$start\_station\_name == "DIVVY CASSETTE REPAIR MOBILE STATION" | all\_trips\_v2$end\_station\_name == "DIVVY CASSETTE REPAIR MOBILE STATION"),]  
  
#to keep our analysis in order based on weekday and month  
all\_trips\_v2$weekday <- ordered(all\_trips\_v2$weekday, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))  
all\_trips\_v2$month <- ordered(all\_trips\_v2$month, levels = c("Mar\_21", "Apr\_21", "May\_21", "Jun\_21", "Jul\_21","Aug\_21", "Sept\_21", "Oct\_21", "Nov\_21", "Dec\_21", "Jan\_22", "Feb\_22"))  
  
#change NA cells in Month column to Sept\_21  
all\_trips\_v2$month[is.na(all\_trips\_v2$month)] <- "Sept\_21"  
  
#Step 4: Conduct Descriptive Analysis  
#min, max, median, and average ride lengths  
summary(all\_trips\_v2$ride\_length)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 7.033 12.250 19.978 22.117 1439.367

#rides broken down as either member or casual  
table(all\_trips\_v2$member\_casual)

##   
## casual member   
## 2031039 2538407

#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 28.40690  
## 2 member 13.23362

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 16.76667  
## 2 member 9.75000

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 1439.367  
## 2 member 1426.567

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 1  
## 2 member 1

#average ride time by each day for members versus casual users  
aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$weekday, FUN = mean)

## all\_trips\_v2$member\_casual all\_trips\_v2$weekday all\_trips\_v2$ride\_length  
## 1 casual Sunday 32.66234  
## 2 member Sunday 15.29059  
## 3 casual Monday 28.94682  
## 4 member Monday 12.77175  
## 5 casual Tuesday 25.98493  
## 6 member Tuesday 12.40914  
## 7 casual Wednesday 24.54399  
## 8 member Wednesday 12.48410  
## 9 casual Thursday 24.28307  
## 10 member Thursday 12.40082  
## 11 casual Friday 26.35903  
## 12 member Friday 12.89285  
## 13 casual Saturday 30.64016  
## 14 member Saturday 14.88800

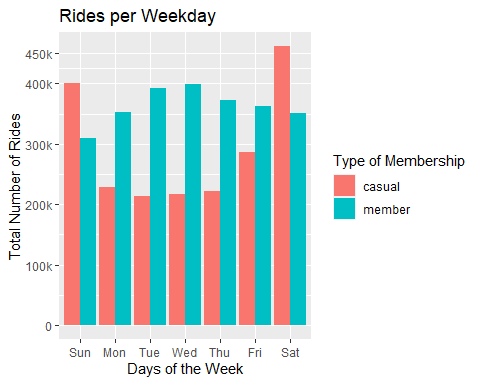
#analyze rideship data by type and weekday  
all\_trips\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>% #creates weekday field using wday()  
 group\_by(member\_casual, weekday) %>% #groups by usertype and weekday  
 summarise(number\_of\_rides = n() #calculates the number of rides and average duration  
 , average\_duration = mean(ride\_length)) %>% #calculates the average duration  
 arrange(member\_casual, weekday) #sorts

## `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 400541 32.7  
## 2 casual Mon 229094 28.9  
## 3 casual Tue 213847 26.0  
## 4 casual Wed 216633 24.5  
## 5 casual Thu 222197 24.3  
## 6 casual Fri 287122 26.4  
## 7 casual Sat 461605 30.6  
## 8 member Sun 310255 15.3  
## 9 member Mon 352920 12.8  
## 10 member Tue 391974 12.4  
## 11 member Wed 398067 12.5  
## 12 member Thu 372769 12.4  
## 13 member Fri 361786 12.9  
## 14 member Sat 350636 14.9

#Step 5: Data Visualization  
#number of rides by weekday  
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(x = 'Days of the Week', y = 'Total Number of Rides', title = 'Rides per Weekday', fill = 'Type of Membership') +  
 scale\_y\_continuous(breaks = c(0, 100000, 200000, 300000, 400000, 450000), labels = c("0", "100k", "200k", "300k", "400k", "450k"))

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



#number of rides by month  
all\_trips\_v2 %>%   
 group\_by(member\_casual, month) %>%   
 summarise(total\_rides = n()  
 , average\_duration = mean(ride\_length)) %>%  
 arrange(member\_casual, month) %>%  
 ggplot(aes(x = month, y = total\_rides, fill = member\_casual)) +geom\_col(position = "dodge") +  
 labs(x = "Month", y = "Total Number of Rides", title = "Rides per Month", fill = "Type of Membership") +  
 scale\_y\_continuous(breaks = c(0, 100000, 200000, 300000, 400000), labels = c("0", "100k", "200k", "300k", "400k"))

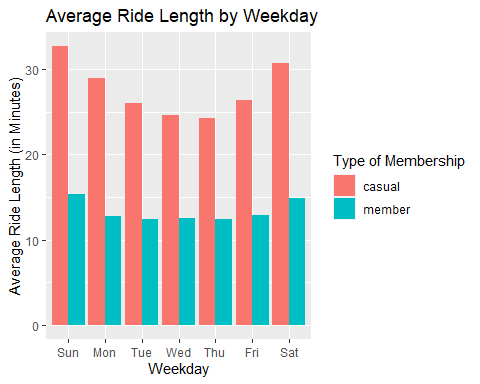
## `summarise()` has grouped output by 'member\_casual'. You can override using the  
## `.groups` argument.

Chart, bar chart

Description automatically generated

#average ride length by weekday  
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(x = "Weekday", y = "Average Ride Length (in Minutes)", title = "Average Ride Length by Weekday", fill = "Type of Membership")

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



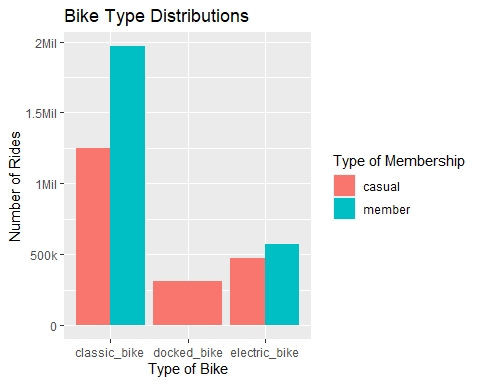
#average ride length by month  
all\_trips\_v2 %>%   
 group\_by(member\_casual, month) %>%   
 summarise(number\_of\_rides = n()  
 , average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, month) %>%   
 ggplot(aes(x = month, y = average\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge") +   
 labs(x = "Month", y = "Average Ride Length (in Minutes)", title = "Average Ride Length by Month", fill = "Type of Membership")

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

Chart, bar chart

Description automatically generated

#investigating the relationship between rideable type and member\_casual  
all\_trips\_v2 %>%   
 ggplot(aes(x = rideable\_type, fill = member\_casual)) +   
 geom\_bar(position = "dodge") +  
 labs(x = "Type of Bike", y = "Number of Rides", title = "Bike Type Distributions", fill = "Type of Membership") +  
 scale\_y\_continuous(breaks = c(0, 500000, 1000000, 1500000, 2000000), labels = c("0", "500k", "1Mil", "1.5Mil", "2Mil"))



#investigating the demand for bikes within 24 hours  
all\_trips\_v2 %>%   
 group\_by(member\_casual, time) %>%   
 summarise(total\_rides = n()) %>%   
 ggplot(aes(x = time, y = total\_rides, color = member\_casual, group = member\_casual)) +  
 geom\_smooth() + scale\_x\_datetime(date\_breaks = "1 hour",  
 date\_labels = "%H:%M", expand = c(0,0)) +  
 theme(axis.text.x = element\_text(angle = 45)) +  
 labs(title = "Ride Demands During the Day", x = "Time", y = "Number of Rides", fill = "Type of Membership")

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

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

Chart, line chart

Description automatically generated