Divvy case study

Contents

[Install required packages 2](#_Toc121293803)

[STEP 1: COLLECT DATA 3](#_Toc121293804)

[STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE 4](#_Toc121293805)

[Compare column names each of the files 4](#_Toc121293806)

[Inspect the data-frames and look for incongruencies 6](#_Toc121293807)

[Stack individual month’s data into one big data frame 17](#_Toc121293808)

[STEP 3: DATA CLEANING 18](#_Toc121293809)

[Making ride\_id unique without checking which rows are affected. USING distinct() 18](#_Toc121293810)

[What are the types of bikes available for the riders to use? 18](#_Toc121293811)

[Drop rows with missing values 18](#_Toc121293812)

[Creating a ride\_duration column 18](#_Toc121293813)

[What data does member\_casual column have? 21](#_Toc121293814)

[Are there any anomalies with start\_station\_name ? 21](#_Toc121293815)

[Are there any anomalies with start\_station\_id ? 22](#_Toc121293816)

[finding 1: 23](#_Toc121293817)

[finding 2: 23](#_Toc121293818)

[Finding 3: 24](#_Toc121293819)

[Are there any anomalies with ‘end\_station\_name’ ? 26](#_Toc121293820)

[Are there any anomalies with end\_station\_id ? 27](#_Toc121293821)

[finding 1: 27](#_Toc121293822)

[Finding 2: 28](#_Toc121293823)

[Are there any anomalies with the latitude, longitude columns? 30](#_Toc121293824)

[Cleanup 30](#_Toc121293825)

[STEP 4: ANALYZE 31](#_Toc121293826)

[Ride Duration Analysis 31](#_Toc121293827)

[Organize data into two datasets one for casual riders and the other for member riders. 31](#_Toc121293828)

[Creating a subset to analyse ride\_duration. 31](#_Toc121293829)

[Histogram for ride duration based on member type. 32](#_Toc121293830)

[Ride Type Analysis 36](#_Toc121293831)

[Type of membership to start location. 38](#_Toc121293832)

Arunakchutha

2022-12-07

# Install required packages

tidyverse for data import and wrangling lubridate for date functions ggplot for visualization dplyr for dataframe manipulation stringr for regex

r = getOption("repos")  
r["CRAN"] = "http://cran.us.r-project.org"  
options(repos = r)  
  
install.packages("dplyr")

## Installing package into 'C:/Users/aruna/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## C:\Users\aruna\AppData\Local\R\win-library\4.2\00LOCK\dplyr\libs\x64\dplyr.dll  
## to C:\Users\aruna\AppData\Local\R\win-library\4.2\dplyr\libs\x64\dplyr.dll:  
## Permission denied

## Warning: restored 'dplyr'

##   
## The downloaded binary packages are in  
## C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\downloaded\_packages

install.packages("tidyverse")

## Installing package into 'C:/Users/aruna/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\downloaded\_packages

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

library(tidyverse) #helps wrangle data

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.4.0 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ stringr 1.5.0.9000  
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2   
## ✔ readr 2.1.3   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(lubridate) #helps wrangle date attributes

## Loading required package: timechange  
##   
## Attaching package: 'lubridate'  
##   
## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

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

# STEP 1: COLLECT DATA

parsing the datasets(csv) files here

aug\_2021 <- read\_csv("202108-divvy-tripdata.csv", show\_col\_types = FALSE)  
sep\_2021 <- read\_csv("202109-divvy-tripdata.csv", show\_col\_types = FALSE)  
oct\_2021 <- read\_csv("202110-divvy-tripdata.csv", show\_col\_types = FALSE)  
nov\_2021 <- read\_csv("202111-divvy-tripdata.csv", show\_col\_types = FALSE)  
dec\_2021 <- read\_csv("202112-divvy-tripdata.csv", show\_col\_types = FALSE)  
jan\_2022 <- read\_csv("202201-divvy-tripdata.csv", show\_col\_types = FALSE)  
feb\_2022 <- read\_csv("202202-divvy-tripdata.csv", show\_col\_types = FALSE)  
mar\_2022 <- read\_csv("202203-divvy-tripdata.csv", show\_col\_types = FALSE)  
apr\_2022 <- read\_csv("202204-divvy-tripdata.csv", show\_col\_types = FALSE)  
may\_2022 <- read\_csv("202205-divvy-tripdata.csv", show\_col\_types = FALSE)  
jun\_2022 <- read\_csv("202206-divvy-tripdata.csv", show\_col\_types = FALSE)  
jul\_2022 <- read\_csv("202207-divvy-tripdata.csv", show\_col\_types = FALSE)  
aug\_2022 <- read\_csv("202208-divvy-tripdata.csv", show\_col\_types = FALSE)  
sep\_2022 <- read\_csv("202209-divvy-tripdata.csv", show\_col\_types = FALSE)

# STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

## Compare 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(aug\_2021)

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

colnames(sep\_2021)

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

colnames(oct\_2021)

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

colnames(nov\_2021)

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

colnames(dec\_2021)

## [1] "ride\_id" "rideable\_type" "started\_at"   
## [4] "ended\_at" "start\_station\_name" "start\_station\_id"   
## [7] "end\_station\_name" "end\_station\_id" "start\_lat"   
## [10] "start\_lng" "end\_lat" "end\_lng"   
## [13] "member\_casual"

colnames(jan\_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(feb\_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(mar\_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(apr\_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(may\_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(jun\_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(jul\_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(aug\_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(sep\_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"

No renaming of columns is necessary since the column names are consistant across all datasets.

## Inspect the data-frames and look for incongruencies

str(aug\_2021)

## spc\_tbl\_ [804,352 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(sep\_2021)

## spc\_tbl\_ [756,147 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(oct\_2021)

## spc\_tbl\_ [631,226 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(nov\_2021)

## spc\_tbl\_ [359,978 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(dec\_2021)

## spc\_tbl\_ [247,540 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(jan\_2022)

## spc\_tbl\_ [103,770 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(feb\_2022)

## spc\_tbl\_ [115,609 × 13] (S3: spec\_tbl\_df/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" ...  
## - 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(mar\_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(apr\_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(may\_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(jun\_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(jul\_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(aug\_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(sep\_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>

## Stack individual month’s data into one big data frame

all\_trips <- bind\_rows(aug\_2021,sep\_2021,oct\_2021,nov\_2021,dec\_2021,jan\_2022,feb\_2022,mar\_2022,apr\_2022,may\_2022,jun\_2022,jul\_2022,aug\_2022,sep\_2022)  
head(all\_trips)

## # A tibble: 6 × 13  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 99103BB87CC6C… electr… 2021-08-10 17:15:49 2021-08-10 17:22:44 <NA> <NA>   
## 2 EAFCCCFB0A3FC… electr… 2021-08-10 17:23:14 2021-08-10 17:39:24 <NA> <NA>   
## 3 9EF4F46C57AD2… electr… 2021-08-21 02:34:23 2021-08-21 02:50:36 <NA> <NA>   
## 4 5834D3208BFAF… electr… 2021-08-21 06:52:55 2021-08-21 07:08:13 <NA> <NA>   
## 5 CD825CB87ED1D… electr… 2021-08-19 11:55:29 2021-08-19 12:04:11 <NA> <NA>   
## 6 612F12C94A964… electr… 2021-08-19 12:41:12 2021-08-19 12:47:47 <NA> <NA>   
## # … with 7 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, and abbreviated variable names ¹​rideable\_type,  
## # ²​start\_station\_name, ³​start\_station\_id

# STEP 3: DATA CLEANING

## Making ride\_id unique without checking which rows are affected. USING distinct()

Creating a new dataset ‘all\_trips\_v2’ since we might drop some rows while trying to make ride\_id distinct

all\_trips\_v2 <- distinct(all\_trips, ride\_id, .keep\_all = TRUE)

conclusion: all\_trips\_v2 contains unique ride\_id

## What are the types of bikes available for the riders to use?

unique(all\_trips\_v2$rideable\_type)

## [1] "electric\_bike" "classic\_bike" "docked\_bike"

## Drop rows with missing values

nrow(all\_trips\_v2)

## [1] 7388734

all\_trips\_v2 <- na.omit(all\_trips\_v2)  
nrow(all\_trips\_v2) # to see if any rows were dropped

## [1] 5769700

conclusion: We have removed all NA values from the dataset and 21.91% of rows were omitted. Still the dataset is large enough to proceed with the analysis

## Creating a ride\_duration column

This is to observe if there are any anomalies with the started\_at and ended\_at data. If there exists any such anomalies we might try to identify the nature of the cause and omit those rows.

all\_trips\_v2$ride\_duration <- difftime(all\_trips\_v2$ended\_at,all\_trips\_v2$started\_at)  
str(all\_trips\_v2)

## tibble [5,769,700 × 14] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5769700] "DD06751C6019D865" "79973DC3B232048F" "0249AD4B258806AD" "F41EB054E44ACFDA" ...  
## $ rideable\_type : chr [1:5769700] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:5769700], format: "2021-08-08 17:21:26" "2021-08-27 08:53:52" ...  
## $ ended\_at : POSIXct[1:5769700], format: "2021-08-08 17:25:37" "2021-08-27 09:18:29" ...  
## $ start\_station\_name: chr [1:5769700] "Desplaines St & Kinzie St" "Larrabee St & Armitage Ave" "Aberdeen St & Jackson Blvd" "Michigan Ave & Oak St" ...  
## $ start\_station\_id : chr [1:5769700] "TA1306000003" "TA1309000006" "13157" "13042" ...  
## $ end\_station\_name : chr [1:5769700] "Kingsbury St & Kinzie St" "Michigan Ave & Oak St" "Aberdeen St & Jackson Blvd" "Michigan Ave & Oak St" ...  
## $ end\_station\_id : chr [1:5769700] "KA1503000043" "13042" "13157" "13042" ...  
## $ start\_lat : num [1:5769700] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:5769700] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:5769700] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:5769700] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:5769700] "member" "member" "member" "casual" ...  
## $ ride\_duration : 'difftime' num [1:5769700] 251 1477 37 282 ...  
## ..- attr(\*, "units")= chr "secs"  
## - attr(\*, "na.action")= 'omit' Named int [1:1619034] 1 2 3 4 5 6 7 8 9 10 ...  
## ..- attr(\*, "names")= chr [1:1619034] "1" "2" "3" "4" ...

head(all\_trips\_v2)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 DD06751C6019D… classi… 2021-08-08 17:21:26 2021-08-08 17:25:37 Despla… TA1306…  
## 2 79973DC3B2320… classi… 2021-08-27 08:53:52 2021-08-27 09:18:29 Larrab… TA1309…  
## 3 0249AD4B25880… classi… 2021-08-08 12:59:18 2021-08-08 12:59:55 Aberde… 13157   
## 4 F41EB054E44AC… classi… 2021-08-12 16:52:09 2021-08-12 16:56:51 Michig… 13042   
## 5 B149E6C71A1C3… classi… 2021-08-23 15:33:04 2021-08-23 16:09:00 Michig… 13042   
## 6 C41829CD6CC5A… classi… 2021-08-23 10:11:09 2021-08-23 10:51:11 Aberde… 13157   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <drtn>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

Since ride\_duration is ‘difftime’ and not ‘numeric’ we have to change that inorder to perform calculations on that data.

all\_trips\_v2$ride\_duration <- as.numeric(as.character(all\_trips\_v2$ride\_duration))  
str(all\_trips\_v2)

## tibble [5,769,700 × 14] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5769700] "DD06751C6019D865" "79973DC3B232048F" "0249AD4B258806AD" "F41EB054E44ACFDA" ...  
## $ rideable\_type : chr [1:5769700] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:5769700], format: "2021-08-08 17:21:26" "2021-08-27 08:53:52" ...  
## $ ended\_at : POSIXct[1:5769700], format: "2021-08-08 17:25:37" "2021-08-27 09:18:29" ...  
## $ start\_station\_name: chr [1:5769700] "Desplaines St & Kinzie St" "Larrabee St & Armitage Ave" "Aberdeen St & Jackson Blvd" "Michigan Ave & Oak St" ...  
## $ start\_station\_id : chr [1:5769700] "TA1306000003" "TA1309000006" "13157" "13042" ...  
## $ end\_station\_name : chr [1:5769700] "Kingsbury St & Kinzie St" "Michigan Ave & Oak St" "Aberdeen St & Jackson Blvd" "Michigan Ave & Oak St" ...  
## $ end\_station\_id : chr [1:5769700] "KA1503000043" "13042" "13157" "13042" ...  
## $ start\_lat : num [1:5769700] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:5769700] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:5769700] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:5769700] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:5769700] "member" "member" "member" "casual" ...  
## $ ride\_duration : num [1:5769700] 251 1477 37 282 2156 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:1619034] 1 2 3 4 5 6 7 8 9 10 ...  
## ..- attr(\*, "names")= chr [1:1619034] "1" "2" "3" "4" ...

head(all\_trips\_v2)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 DD06751C6019D… classi… 2021-08-08 17:21:26 2021-08-08 17:25:37 Despla… TA1306…  
## 2 79973DC3B2320… classi… 2021-08-27 08:53:52 2021-08-27 09:18:29 Larrab… TA1309…  
## 3 0249AD4B25880… classi… 2021-08-08 12:59:18 2021-08-08 12:59:55 Aberde… 13157   
## 4 F41EB054E44AC… classi… 2021-08-12 16:52:09 2021-08-12 16:56:51 Michig… 13042   
## 5 B149E6C71A1C3… classi… 2021-08-23 15:33:04 2021-08-23 16:09:00 Michig… 13042   
## 6 C41829CD6CC5A… classi… 2021-08-23 10:11:09 2021-08-23 10:51:11 Aberde… 13157   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -7745 383 669 1093 1199 2497750

Omitting the rows that fit the following conditions: Negative ride durations in this dataset (as seen in the summary). Ride duration that were below 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it was secure). Ride durations that are greater than 24 hours (86400s) because, after 24 hrs from starting a trip, if these bikes are not returned, these cases in dataset might be error, loss/theft of bike and are out of scope for the analysis at the moment.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 filter(ride\_duration >= 60 & ride\_duration <= 86400)  
summary(all\_trips\_v2$ride\_duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60 394 680 1086 1212 86362

## What data does member\_casual column have?

count() is a convenient way to get a sense of the distribution of values in a dataset

all\_trips\_v2 %>% count(member\_casual)

## # A tibble: 2 × 2  
## member\_casual n  
## <chr> <int>  
## 1 casual 2408020  
## 2 member 3271217

conclusion: member\_casual column is clean

## Are there any anomalies with start\_station\_name ?

sorting the dataset by ascending order of the str\_length(start\_station\_name) to identify abbreviations or incomplete street names.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 arrange(str\_length(start\_station\_name), start\_station\_name)  
head(all\_trips\_v2)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 5E181D51F7C39… electr… 2021-08-04 16:31:29 2021-08-04 16:48:33 351 351   
## 2 7A492BA13DC98… electr… 2022-06-03 21:12:47 2022-06-03 21:54:42 WestChi DIVVY …  
## 3 7325C8783518F… electr… 2022-06-03 19:58:33 2022-06-03 20:00:46 WestChi DIVVY …  
## 4 053289A38DEB1… electr… 2022-07-20 08:10:53 2022-07-20 08:16:17 WestChi DIVVY …  
## 5 A80F66E86B532… electr… 2022-08-12 10:45:23 2022-08-12 11:18:03 WestChi DIVVY …  
## 6 6CD2E6D1D3F39… classi… 2021-08-19 17:42:34 2021-08-19 18:24:18 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

The start\_station\_name and start\_station\_id for row 1 is the same (351), indicating that the station ID was misinterpreted into station name. We can find out which station has an ID of 351 and rectify the data.

head(all\_trips\_v2 %>%   
 filter(start\_station\_id == "351") %>%  
 select(start\_station\_id, start\_station\_name))

## # A tibble: 6 × 2  
## start\_station\_id start\_station\_name   
## <chr> <chr>   
## 1 351 351   
## 2 351 Mulligan Ave & Wellington Ave  
## 3 351 Mulligan Ave & Wellington Ave  
## 4 351 Mulligan Ave & Wellington Ave  
## 5 351 Mulligan Ave & Wellington Ave  
## 6 351 Mulligan Ave & Wellington Ave

‘Mulligan Ave & Wellington Ave’ is the station name for station id 351.

all\_trips\_v2$start\_station\_name[all\_trips\_v2$start\_station\_name == "351"] <-"Mulligan Ave & Wellington Ave"  
head(all\_trips\_v2 %>%   
 filter(start\_station\_id == "351") %>%  
 select(start\_station\_id, start\_station\_name))

## # A tibble: 6 × 2  
## start\_station\_id start\_station\_name   
## <chr> <chr>   
## 1 351 Mulligan Ave & Wellington Ave  
## 2 351 Mulligan Ave & Wellington Ave  
## 3 351 Mulligan Ave & Wellington Ave  
## 4 351 Mulligan Ave & Wellington Ave  
## 5 351 Mulligan Ave & Wellington Ave  
## 6 351 Mulligan Ave & Wellington Ave

performing a frequency table check to look for misspelled or irrelevant characters in start\_station\_names

temp\_data <- all\_trips\_v2 %>% count(start\_station\_name)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## start\_station\_name n  
## <chr> <int>  
## 1 10101 S Stony Island Ave 1  
## 2 111th St - Morgan Park Metra 6  
## 3 2112 W Peterson Ave 1017  
## 4 532 E 43rd St 17  
## 5 63rd & Western Ave - north corner 5  
## 6 63rd & Western Ave - south corner 7

after manually observing all the data in this row, it is evident that the entries in the ‘start\_station\_name’ column are clean.

CONCLUSION: start\_station\_name is clean

## Are there any anomalies with start\_station\_id ?

performing a frequency table check to look for misspelled or irrelevant characters in start\_station\_id

temp\_data <- all\_trips\_v2 %>% count(start\_station\_id)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## start\_station\_id n  
## <chr> <int>  
## 1 021320 202  
## 2 1011 21  
## 3 1012 2  
## 4 1013 6  
## 5 1014 2  
## 6 1015 1

### finding 1:

Some of the IDs have a trailing decimal point on them (row: 309, start\_station\_id: 20247.0). The whole column is of character data type.Have to replace those ‘.0’ in the dataset to maintain consistancy.

all\_trips\_v2$start\_station\_id = str\_replace\_all(all\_trips\_v2$start\_station\_id, "\\.0$", "")  
View(all\_trips\_v2)  
  
temp\_data <- all\_trips\_v2 %>% count(start\_station\_id)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## start\_station\_id n  
## <chr> <int>  
## 1 021320 202  
## 2 1011 21  
## 3 1012 2  
## 4 1013 6  
## 5 1014 2  
## 6 1015 1

All the “.0” entries have been corrected to no decimal.

### finding 2:

Some starting station IDs are ‘LP-’ and ‘RN-’ maybe its missing a suffix?

all\_trips\_v2 %>%  
 filter(start\_station\_id == "LP-" |start\_station\_id == "RN-" ) %>%  
 select(start\_station\_name, start\_station\_id) %>%  
 count(start\_station\_name, start\_station\_id)

## # A tibble: 2 × 3  
## start\_station\_name start\_station\_id n  
## <chr> <chr> <int>  
## 1 Franklin St & Illinois St RN- 19115  
## 2 Lincoln Park Conservatory LP- 16571

‘RN-’ and ‘LP-’ refers to ‘Franklin St & Illinois St’ and ‘Lincoln Park Conservatory’ respectively.

all\_trips\_v2 %>%  
 filter(end\_station\_id == "LP-" |end\_station\_id == "RN-" ) %>%  
 select(end\_station\_name, end\_station\_id) %>%  
 count(end\_station\_name, end\_station\_id)

## # A tibble: 2 × 3  
## end\_station\_name end\_station\_id n  
## <chr> <chr> <int>  
## 1 Franklin St & Illinois St RN- 18626  
## 2 Lincoln Park Conservatory LP- 16058

as long as ‘RN-’ and ‘LP-’ rows from end\_station\_id refers to the same names in end\_station\_name, we dont have to drop these rows. This has to me mentioned in the documentaion.

### Finding 3:

some entries in the start\_station\_id has station names instead of ID. performing an orderby length of the data in start\_station\_id column reveals this.

temp\_data <- all\_trips\_v2 %>%   
 arrange(desc(str\_length(all\_trips\_v2$start\_station\_id))) %>%  
 select(rideable\_type, start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, ride\_duration)   
 View(temp\_data)  
 head(temp\_data)

## # A tibble: 6 × 6  
## rideable\_type start\_station\_name start…¹ end\_s…² end\_s…³ ride\_…⁴  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 electric\_bike Pawel Bialowas - Test- PBSC cha… Pawel … Pawel … Pawel … 155  
## 2 electric\_bike Wilton Ave & Diversey Pkwy - Ch… Wilton… Wilton… Wilton… 278  
## 3 classic\_bike Wilton Ave & Diversey Pkwy - Ch… Wilton… Larrab… 13193 524  
## 4 classic\_bike Wilton Ave & Diversey Pkwy - Ch… Wilton… Wilton… TA1307… 525  
## 5 electric\_bike Wilton Ave & Diversey Pkwy - Ch… Wilton… Greenv… TA1307… 865  
## 6 classic\_bike Wilton Ave & Diversey Pkwy - Ch… Wilton… Sheffi… TA1306… 483  
## # … with abbreviated variable names ¹​start\_station\_id, ²​end\_station\_name,  
## # ³​end\_station\_id, ⁴​ride\_duration

creating an ordered frequency table(descending) based on the length of the start\_station\_id.

temp\_data <- all\_trips\_v2 %>%   
 count(start\_station\_id) %>%  
 arrange(desc(str\_length(start\_station\_id)))  
 View(temp\_data)  
 head(temp\_data, n=20)

## # A tibble: 20 × 2  
## start\_station\_id n  
## <chr> <int>  
## 1 Pawel Bialowas - Test- PBSC charging station 1  
## 2 Wilton Ave & Diversey Pkwy - Charging 17  
## 3 Bissell St & Armitage Ave - Charging 16  
## 4 Hubbard Bike-checking (LBS-WH-TEST) 1108  
## 5 DIVVY 001 - Warehouse test station 7  
## 6 2059 Hastings Warehouse Station 43  
## 7 Divvy Valet - Oakwood Beach 9  
## 8 Hastings WH 2 1  
## 9 chargingstx0 1589  
## 10 chargingstx1 8299  
## 11 chargingstx3 7823  
## 12 chargingstx4 3478  
## 13 chargingstx5 5126  
## 14 KA1503000001 147  
## 15 KA1503000002 1704  
## 16 KA1503000003 300  
## 17 KA1503000004 3093  
## 18 KA1503000005 2442  
## 19 KA1503000007 9519  
## 20 KA1503000009 669

collecting more details on chargingstx0-5

head(all\_trips\_v2 %>%  
 filter(start\_station\_id == "chargingstx0") %>%  
 select(start\_station\_id, start\_station\_name))

## # A tibble: 6 × 2  
## start\_station\_id start\_station\_name   
## <chr> <chr>   
## 1 chargingstx0 Wilton Ave & Diversey Pkwy\*  
## 2 chargingstx0 Wilton Ave & Diversey Pkwy\*  
## 3 chargingstx0 Wilton Ave & Diversey Pkwy\*  
## 4 chargingstx0 Wilton Ave & Diversey Pkwy\*  
## 5 chargingstx0 Wilton Ave & Diversey Pkwy\*  
## 6 chargingstx0 Wilton Ave & Diversey Pkwy\*

since the station ID and the station name are not exactly the same, we can move onto other potential issues. It can be observed that from the first 8 rows of the frequency table shown above, the start\_station\_id s that have issues are clearly 13 characters in length or more. Since our frequency table is ordered by descending order of the length of the IDs, we can drop the rows containing those 8 IDs.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 filter(str\_length(start\_station\_id) <= 12)

temp\_data <- all\_trips\_v2 %>%   
 count(start\_station\_id) %>%  
 arrange(desc(str\_length(start\_station\_id)))  
 View(temp\_data)  
 head(temp\_data, n=20)

## # A tibble: 20 × 2  
## start\_station\_id n  
## <chr> <int>  
## 1 chargingstx0 1589  
## 2 chargingstx1 8299  
## 3 chargingstx3 7823  
## 4 chargingstx4 3478  
## 5 chargingstx5 5126  
## 6 KA1503000001 147  
## 7 KA1503000002 1704  
## 8 KA1503000003 300  
## 9 KA1503000004 3093  
## 10 KA1503000005 2442  
## 11 KA1503000007 9519  
## 12 KA1503000009 669  
## 13 KA1503000010 307  
## 14 KA1503000011 129  
## 15 KA1503000012 19091  
## 16 KA1503000013 467  
## 17 KA1503000014 24677  
## 18 KA1503000015 12014  
## 19 KA1503000018 395  
## 20 KA1503000019 357

Conclusion: The start\_station\_id column is clean.

## Are there any anomalies with ‘end\_station\_name’ ?

sorting the dataset by ascending order of the str\_length(end\_station\_name) to identify abbreviations or incomplete street names.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 arrange(str\_length(end\_station\_name), end\_station\_name)  
head(all\_trips\_v2)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 58B441E68B86A… classi… 2021-08-14 19:01:32 2021-08-14 19:17:42 Smith … 643   
## 2 E1E6509A0BCE9… classi… 2021-08-07 11:06:34 2021-08-07 12:04:56 Smith … 643   
## 3 B8E690C8B0A87… classi… 2021-08-14 19:00:53 2021-08-14 19:19:39 Smith … 643   
## 4 24933988DF96F… electr… 2021-08-03 20:51:40 2021-08-03 21:18:59 Smith … 643   
## 5 E936E896092FC… classi… 2021-08-30 08:56:40 2021-08-30 09:06:07 Smith … 643   
## 6 61ABBABD7456B… classi… 2021-08-16 15:09:38 2021-08-16 15:13:05 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

performing a frequency table check to look for misspelled or irrelevant characters in end\_station\_names

temp\_data <- all\_trips\_v2 %>% count(end\_station\_name)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## end\_station\_name n  
## <chr> <int>  
## 1 10101 S Stony Island Ave 4  
## 2 111th St - Morgan Park Metra 2  
## 3 2112 W Peterson Ave 1200  
## 4 532 E 43rd St 14  
## 5 63rd & Western Ave - north corner 11  
## 6 63rd & Western Ave - south corner 5

after manually observing all the data in this row, it is evident that the entries in the ‘end\_station\_name’ column are clean.

## Are there any anomalies with end\_station\_id ?

performing a frequency table check to look for misspelled or irrelevant characters in end\_station\_id

temp\_data <- all\_trips\_v2 %>% count(end\_station\_id)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## end\_station\_id n  
## <chr> <int>  
## 1 021320 184  
## 2 1011 33  
## 3 1012 3  
## 4 1013 8  
## 5 1015 2  
## 6 1016 3

### finding 1:

Some of the IDs have a trailing decimal point on them (row: 234, end\_station\_id: 20.0). The whole column is of character data type.Have to replace those ‘.0’ in the dataset to maintain consistancy.

all\_trips\_v2$end\_station\_id = str\_replace\_all(all\_trips\_v2$end\_station\_id, "\\.0$", "")  
View(all\_trips\_v2)  
  
temp\_data <- all\_trips\_v2 %>% count(end\_station\_id)  
View(temp\_data)  
head(temp\_data)

## # A tibble: 6 × 2  
## end\_station\_id n  
## <chr> <int>  
## 1 021320 184  
## 2 1011 33  
## 3 1012 3  
## 4 1013 8  
## 5 1015 2  
## 6 1016 3

All the “.0” entries have been corrected to no decimal.

### Finding 2:

some entries in the end\_station\_id has station names instead of ID. performing an orderby length of the data in end\_station\_id column reveals this.

temp\_data <- all\_trips\_v2 %>%   
 arrange(desc(str\_length(all\_trips\_v2$end\_station\_id))) %>%  
 select(rideable\_type, start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, ride\_duration)   
 View(temp\_data)  
 head(temp\_data)

## # A tibble: 6 × 6  
## rideable\_type start\_station\_name start\_statio…¹ end\_s…² end\_s…³ ride\_…⁴  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 electric\_bike Wells St & Concord Ln TA1308000050 Wilton… Wilton… 803  
## 2 electric\_bike Clark St & Armitage Ave 13146 Wilton… Wilton… 395  
## 3 classic\_bike Lincoln Ave & Roscoe St TA1307000138 Wilton… Wilton… 506  
## 4 classic\_bike Sheridan Rd & Buena Ave TA1309000027 Wilton… Wilton… 828  
## 5 classic\_bike Bissell St & Armitage Ave 13059 Wilton… Wilton… 407  
## 6 classic\_bike Clark St & Wellington Ave TA1307000136 Wilton… Wilton… 325  
## # … with abbreviated variable names ¹​start\_station\_id, ²​end\_station\_name,  
## # ³​end\_station\_id, ⁴​ride\_duration

creating an ordered frequency table(descending) based on the length of the end\_station\_id.

temp\_data <- all\_trips\_v2 %>%   
 count(end\_station\_id) %>%  
 arrange(desc(str\_length(end\_station\_id)))  
 View(temp\_data)  
 head(temp\_data, n=20)

## # A tibble: 20 × 2  
## end\_station\_id n  
## <chr> <int>  
## 1 Wilton Ave & Diversey Pkwy - Charging 15  
## 2 Bissell St & Armitage Ave - Charging 17  
## 3 DIVVY CASSETTE REPAIR MOBILE STATION 6  
## 4 Hubbard Bike-checking (LBS-WH-TEST) 323  
## 5 2059 Hastings Warehouse Station 2  
## 6 Divvy Valet - Oakwood Beach 15  
## 7 chargingstx0 1326  
## 8 chargingstx1 8880  
## 9 chargingstx3 8207  
## 10 chargingstx4 3638  
## 11 chargingstx5 5906  
## 12 KA1503000001 222  
## 13 KA1503000002 1756  
## 14 KA1503000003 246  
## 15 KA1503000004 3193  
## 16 KA1503000005 2145  
## 17 KA1503000007 8903  
## 18 KA1503000009 530  
## 19 KA1503000010 352  
## 20 KA1503000011 112

It can be observed that from the first 6 rows of the frequency table shown above, the end\_station\_id s that have issues are clearly more than 13 characters in length . Since our frequency table is ordered by descending order of the length of the IDs, we can drop the rows containing those 6 IDs.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 filter(str\_length(end\_station\_id) <= 12)

temp\_data <- all\_trips\_v2 %>%   
 count(end\_station\_id) %>%  
 arrange(desc(str\_length(end\_station\_id)))  
 View(temp\_data)  
 head(temp\_data, n=20)

## # A tibble: 20 × 2  
## end\_station\_id n  
## <chr> <int>  
## 1 chargingstx0 1326  
## 2 chargingstx1 8880  
## 3 chargingstx3 8207  
## 4 chargingstx4 3638  
## 5 chargingstx5 5906  
## 6 KA1503000001 222  
## 7 KA1503000002 1756  
## 8 KA1503000003 246  
## 9 KA1503000004 3193  
## 10 KA1503000005 2145  
## 11 KA1503000007 8903  
## 12 KA1503000009 530  
## 13 KA1503000010 352  
## 14 KA1503000011 112  
## 15 KA1503000012 19043  
## 16 KA1503000013 382  
## 17 KA1503000014 24495  
## 18 KA1503000015 12208  
## 19 KA1503000018 518  
## 20 KA1503000019 347

Conclusion: The end\_station\_id column is clean.

## Are there any anomalies with the latitude, longitude columns?

summary(all\_trips\_v2$start\_lat)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 41.65 41.88 41.90 41.90 41.93 42.06

summary(all\_trips\_v2$start\_lng)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -87.83 -87.66 -87.64 -87.64 -87.63 -87.53

summary(all\_trips\_v2$end\_lat)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 41.65 41.88 41.90 41.90 41.93 42.17

summary(all\_trips\_v2$end\_lng)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -87.83 -87.66 -87.64 -87.64 -87.63 -87.53

Conclusion: The latitude and longitude data are clean.

## Cleanup

rm(aug\_2021)  
rm(sep\_2021)  
rm(oct\_2021)  
rm(nov\_2021)  
rm(dec\_2021)  
rm(jan\_2022)  
rm(feb\_2022)  
rm(mar\_2022)  
rm(apr\_2022)  
rm(may\_2022)  
rm(jun\_2022)  
rm(jul\_2022)  
rm(temp\_data)

# STEP 4: ANALYZE

## Ride Duration Analysis

### Organize data into two datasets one for casual riders and the other for member riders.

all\_trips\_v3\_casual <- all\_trips\_v2 %>% filter(  
 member\_casual == "casual"  
)  
all\_trips\_v3\_member <- all\_trips\_v2 %>% filter(  
 member\_casual == "member"  
)  
head(all\_trips\_v3\_casual)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 58B441E68B86A… classi… 2021-08-14 19:01:32 2021-08-14 19:17:42 Smith … 643   
## 2 E1E6509A0BCE9… classi… 2021-08-07 11:06:34 2021-08-07 12:04:56 Smith … 643   
## 3 B8E690C8B0A87… classi… 2021-08-14 19:00:53 2021-08-14 19:19:39 Smith … 643   
## 4 61ABBABD7456B… classi… 2021-08-16 15:09:38 2021-08-16 15:13:05 Smith … 643   
## 5 1C28307A849B1… classi… 2021-08-30 18:51:30 2021-08-30 19:28:11 Smith … 643   
## 6 E4938454847D6… electr… 2021-08-06 17:15:21 2021-08-06 18:09:47 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

View(all\_trips\_v3\_casual)  
head(all\_trips\_v3\_member)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 24933988DF96F… electr… 2021-08-03 20:51:40 2021-08-03 21:18:59 Smith … 643   
## 2 E936E896092FC… classi… 2021-08-30 08:56:40 2021-08-30 09:06:07 Smith … 643   
## 3 0E136E35E531A… classi… 2021-08-20 10:41:28 2021-08-20 11:58:50 Smith … 643   
## 4 758BAE9B6D5AB… electr… 2021-08-08 12:46:29 2021-08-08 13:03:26 Smith … 643   
## 5 3C8949185DBFD… classi… 2021-08-06 19:51:19 2021-08-06 20:08:15 Smith … 643   
## 6 9E21B303C1D77… classi… 2021-08-01 15:39:47 2021-08-01 15:41:09 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

View(all\_trips\_v3\_member)

### Creating a subset to analyse ride\_duration.

all\_trips\_v3\_ride\_duration <- subset(all\_trips\_v2, select = c(member\_casual, ride\_duration)  
)  
all\_trips\_v3\_ride\_duration <- all\_trips\_v3\_ride\_duration %>%   
 mutate(ride\_duration = ride\_duration / 60)  
View(all\_trips\_v3\_ride\_duration)  
head(all\_trips\_v3\_ride\_duration)

## # A tibble: 6 × 2  
## member\_casual ride\_duration  
## <chr> <dbl>  
## 1 casual 16.2   
## 2 casual 58.4   
## 3 casual 18.8   
## 4 member 27.3   
## 5 member 9.45  
## 6 casual 3.45

### Histogram for ride duration based on member type.

if(!require(devtools)) install.packages("devtools")

## Loading required package: devtools

## Loading required package: usethis

devtools::install\_github("kassambara/ggpubr")

## Downloading GitHub repo kassambara/ggpubr@HEAD

## tidyr (1.2.1 -> 21818c67a...) [GitHub]  
## xfun (0.34 -> 0.35 ) [CRAN]  
## stringr (2d2f81b42... -> 4220eda13...) [GitHub]

## Installing 1 packages: xfun

## Installing package into 'C:/Users/aruna/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)

## package 'xfun' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'xfun'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## C:\Users\aruna\AppData\Local\R\win-library\4.2\00LOCK\xfun\libs\x64\xfun.dll  
## to C:\Users\aruna\AppData\Local\R\win-library\4.2\xfun\libs\x64\xfun.dll:  
## Permission denied

## Warning: restored 'xfun'

##   
## The downloaded binary packages are in  
## C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\downloaded\_packages

## Downloading GitHub repo tidyverse/tidyr@HEAD

## stringr (2d2f81b42... -> 4220eda13...) [GitHub]

## Downloading GitHub repo tidyverse/stringr@HEAD

##   
## ── R CMD build ─────────────────────────────────────────────────────────────────  
## checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a58a9f79c1\tidyverse-stringr-4220eda/DESCRIPTION' ... checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a58a9f79c1\tidyverse-stringr-4220eda/DESCRIPTION' ... ✔ checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a58a9f79c1\tidyverse-stringr-4220eda/DESCRIPTION' (715ms)  
## ─ preparing 'stringr': (999ms)  
## checking DESCRIPTION meta-information ... checking DESCRIPTION meta-information ... ✔ checking DESCRIPTION meta-information  
## ─ checking for LF line-endings in source and make files and shell scripts (587ms)  
## ─ checking for empty or unneeded directories  
## ─ building 'stringr\_1.5.0.9000.tar.gz'  
##   
##

## Warning: package 'stringr' is in use and will not be installed

## ── R CMD build ─────────────────────────────────────────────────────────────────  
## checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a586e082d41\tidyverse-tidyr-21818c6/DESCRIPTION' ... checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a586e082d41\tidyverse-tidyr-21818c6/DESCRIPTION' ... ✔ checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a586e082d41\tidyverse-tidyr-21818c6/DESCRIPTION' (599ms)  
## ─ preparing 'tidyr': (1.8s)  
## checking DESCRIPTION meta-information ... checking DESCRIPTION meta-information ... ✔ checking DESCRIPTION meta-information  
## ─ cleaning src  
## ─ checking for LF line-endings in source and make files and shell scripts (711ms)  
## ─ checking for empty or unneeded directories  
## ─ building 'tidyr\_1.2.1.9001.tar.gz'  
##   
##

## Warning: package 'tidyr' is in use and will not be installed

## Downloading GitHub repo tidyverse/stringr@HEAD

## Skipping stringr, it is already being installed

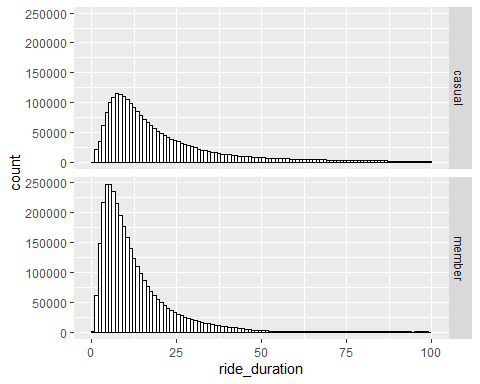
## ── R CMD build ─────────────────────────────────────────────────────────────────  
## checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a583316ee5\kassambara-ggpubr-e4277dc/DESCRIPTION' ... checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a583316ee5\kassambara-ggpubr-e4277dc/DESCRIPTION' ... ✔ checking for file 'C:\Users\aruna\AppData\Local\Temp\RtmpIjgQUd\remotes1a583316ee5\kassambara-ggpubr-e4277dc/DESCRIPTION' (601ms)  
## ─ preparing 'ggpubr': (3.2s)  
## checking DESCRIPTION meta-information ... checking DESCRIPTION meta-information ... ✔ checking DESCRIPTION meta-information  
## ─ checking for LF line-endings in source and make files and shell scripts (1.7s)  
## ─ checking for empty or unneeded directories  
## ─ building 'ggpubr\_0.5.0.999.tar.gz'  
##   
##

## Installing package into 'C:/Users/aruna/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)

## Warning in i.p(...): installation of package  
## 'C:/Users/aruna/AppData/Local/Temp/RtmpIjgQUd/file1a587266346/ggpubr\_0.5.0.999.tar.gz'  
## had non-zero exit status

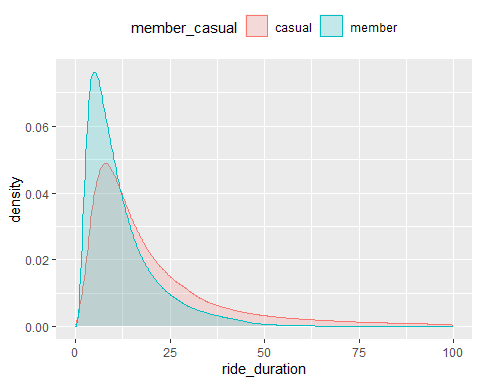
library("ggpubr")

all\_trips\_v3\_ride\_duration %>%   
ggplot(aes(x=ride\_duration)) + geom\_histogram(color="black",fill="white", breaks = c(seq(from = 0, to = 100, by = 1))) + facet\_grid(member\_casual ~ .)



Another way to visualize the ride duration,

ggplot(all\_trips\_v3\_ride\_duration , aes(x=ride\_duration, fill = member\_casual, colour = member\_casual)) +  
 geom\_density(alpha = 0.2, na.rm = TRUE) + xlim(0, 100) +  
 theme(legend.position = "top")



# Converting Ride Duration from seconds to minutes for casual members data set  
all\_trips\_v3\_casual <- all\_trips\_v3\_casual %>%   
 mutate(ride\_duration = ride\_duration / 60)  
View(all\_trips\_v3\_casual)  
head(all\_trips\_v3\_casual)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 58B441E68B86A… classi… 2021-08-14 19:01:32 2021-08-14 19:17:42 Smith … 643   
## 2 E1E6509A0BCE9… classi… 2021-08-07 11:06:34 2021-08-07 12:04:56 Smith … 643   
## 3 B8E690C8B0A87… classi… 2021-08-14 19:00:53 2021-08-14 19:19:39 Smith … 643   
## 4 61ABBABD7456B… classi… 2021-08-16 15:09:38 2021-08-16 15:13:05 Smith … 643   
## 5 1C28307A849B1… classi… 2021-08-30 18:51:30 2021-08-30 19:28:11 Smith … 643   
## 6 E4938454847D6… electr… 2021-08-06 17:15:21 2021-08-06 18:09:47 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

# Converting Ride Duration from seconds to minutes for annual members data set  
all\_trips\_v3\_member <- all\_trips\_v3\_member %>%   
 mutate(ride\_duration = ride\_duration / 60)  
View(all\_trips\_v3\_member)  
head(all\_trips\_v3\_member)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 24933988DF96F… electr… 2021-08-03 20:51:40 2021-08-03 21:18:59 Smith … 643   
## 2 E936E896092FC… classi… 2021-08-30 08:56:40 2021-08-30 09:06:07 Smith … 643   
## 3 0E136E35E531A… classi… 2021-08-20 10:41:28 2021-08-20 11:58:50 Smith … 643   
## 4 758BAE9B6D5AB… electr… 2021-08-08 12:46:29 2021-08-08 13:03:26 Smith … 643   
## 5 3C8949185DBFD… classi… 2021-08-06 19:51:19 2021-08-06 20:08:15 Smith … 643   
## 6 9E21B303C1D77… classi… 2021-08-01 15:39:47 2021-08-01 15:41:09 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

# Descriptive Statistics on both member types  
summary(all\_trips\_v3\_casual$ride\_duration)

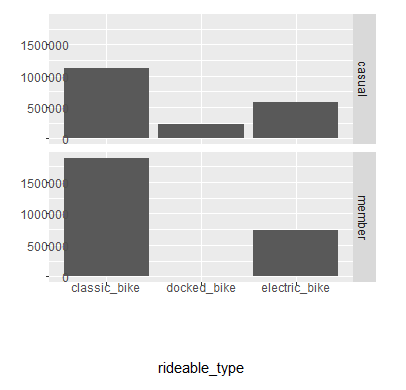
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 8.667 14.933 25.195 27.433 1439.367

summary(all\_trips\_v3\_member$ride\_duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.583 9.383 12.859 15.900 1435.467

## Ride Type Analysis

all\_trips\_v2 %>% ggplot(aes(x=rideable\_type)) + geom\_bar() + facet\_grid(member\_casual ~ .)

On executing this, the resulting viz needs further investigation. 

library("dplyr")  
all\_trips\_v4\_ride\_type <- all\_trips\_v2 %>% filter(  
 rideable\_type == "docked\_bike")  
View(all\_trips\_v4\_ride\_type)

From this visualization, It can be assumed that the movement of the docked\_bikes could refer to the divvy personnel redistributing the bikes to and fro different stations to optimize demand. This assumption is being made after checking the divvy website and realizing that they make two types of bikes available. One is the classic bike and the other is the e-bike.

So going forward, we are going to exclude the docked\_bike’s data from our analysis. by removing those data from the all\_trips\_v2 table. Since ride duration analysis used the data from all\_trips\_v2, we recommend re-running the code chunks 37 through 42 after this deletion.

all\_trips\_v2 <- all\_trips\_v2 %>%   
 filter(rideable\_type != "docked\_bike")  
View(all\_trips\_v2)  
head(all\_trips\_v2)

## # A tibble: 6 × 14  
## ride\_id ridea…¹ started\_at ended\_at start…² start…³  
## <chr> <chr> <dttm> <dttm> <chr> <chr>   
## 1 58B441E68B86A… classi… 2021-08-14 19:01:32 2021-08-14 19:17:42 Smith … 643   
## 2 E1E6509A0BCE9… classi… 2021-08-07 11:06:34 2021-08-07 12:04:56 Smith … 643   
## 3 B8E690C8B0A87… classi… 2021-08-14 19:00:53 2021-08-14 19:19:39 Smith … 643   
## 4 24933988DF96F… electr… 2021-08-03 20:51:40 2021-08-03 21:18:59 Smith … 643   
## 5 E936E896092FC… classi… 2021-08-30 08:56:40 2021-08-30 09:06:07 Smith … 643   
## 6 61ABBABD7456B… classi… 2021-08-16 15:09:38 2021-08-16 15:13:05 Smith … 643   
## # … with 8 more variables: end\_station\_name <chr>, end\_station\_id <chr>,  
## # start\_lat <dbl>, start\_lng <dbl>, end\_lat <dbl>, end\_lng <dbl>,  
## # member\_casual <chr>, ride\_duration <dbl>, and abbreviated variable names  
## # ¹​rideable\_type, ²​start\_station\_name, ³​start\_station\_id

A NOTE ON REDOING THE RIDE DURATION ANALYSIS

before excluding the docked\_bike data, the descriptive stats were as follows:

Min. 1st Qu. Median Mean 3rd Qu. Max.

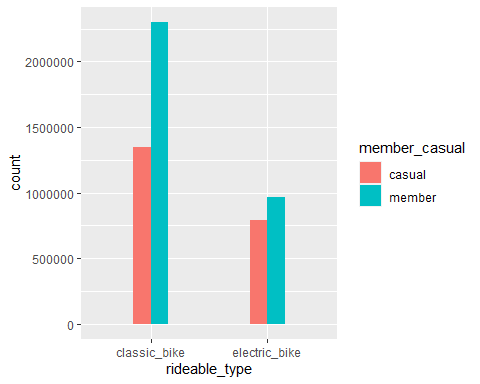
1.000 8.867 15.317 25.763 28.067 1439.367 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 5.550 9.333 12.797 15.817 1435.467

After removing the docked\_bike data the descriptive stats are as follows:

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 8.433 14.233 22.788 25.467 1438.583 Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 5.550 9.333 12.797 15.817 1435.467

# running the ride type analysis once more,  
  
all\_trips\_v2 %>%   
ggplot(aes(x=rideable\_type, fill = member\_casual)) + geom\_bar(width = 0.3, position='dodge')



## Type of membership to start location.

(Do I see a cycle near my home)

start\_location\_membership <- all\_trips\_v2 %>% group\_by(start\_station\_name) %>% count(member\_casual)   
View(start\_location\_membership)

# Splitting casual and annual members data in different data sets for comparison purposes.  
start\_location\_casual <- start\_location\_membership %>%   
 filter(member\_casual != "member")  
start\_location\_casual <- start\_location\_casual[order(-start\_location\_casual$n),]  
View(start\_location\_casual)  
head(start\_location\_casual)

## # A tibble: 6 × 3  
## # Groups: start\_station\_name [6]  
## start\_station\_name member\_casual n  
## <chr> <chr> <int>  
## 1 Streeter Dr & Grand Ave casual 57869  
## 2 DuSable Lake Shore Dr & Monroe St casual 29328  
## 3 DuSable Lake Shore Dr & North Blvd casual 27113  
## 4 Michigan Ave & Oak St casual 26919  
## 5 Millennium Park casual 25421  
## 6 Wells St & Concord Ln casual 20427

start\_station\_lat\_lng <- unique(all\_trips\_v2[,c('start\_station\_name','start\_lat','start\_lng')])  
View(start\_station\_lat\_lng)