## Divvy Bikeshare Full Year Analysis

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

#### Loading libraries

library(tidyverse)

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

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 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)

#### Data Collection

reading .csv files in to dataframes

tripdata\_2020\_09 <- read\_csv("202009-divvy-tripdata.csv")

## Rows: 532958 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, memb...  
## dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, e...  
## dttm (2): started\_at, ended\_at

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

tripdata\_2020\_10 <- read\_csv("202010-divvy-tripdata.csv")

## Rows: 388653 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, memb...  
## dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, e...  
## dttm (2): started\_at, ended\_at

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

tripdata\_2020\_11 <- read\_csv("202011-divvy-tripdata.csv")

## Rows: 259716 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, memb...  
## dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, e...  
## dttm (2): started\_at, ended\_at

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

tripdata\_2020\_12 <- read\_csv("202012-divvy-tripdata.csv")

## Rows: 131573 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_01 <- read\_csv("202101-divvy-tripdata.csv")

## Rows: 96834 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_02 <- read\_csv("202102-divvy-tripdata.csv")

## Rows: 49622 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_03 <- read\_csv("202103-divvy-tripdata.csv")

## Rows: 228496 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_04 <- read\_csv("202104-divvy-tripdata.csv")

## Rows: 337230 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_05 <- read\_csv("202105-divvy-tripdata.csv")

## Rows: 531633 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_06 <- read\_csv("202106-divvy-tripdata.csv")

## Rows: 729595 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_07 <- read\_csv("202107-divvy-tripdata.csv")

## Rows: 822410 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

tripdata\_2021\_08 <- read\_csv("202108-divvy-tripdata.csv")

## Rows: 804352 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

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

#### Data consolidation

First a look at the column names of each dataframe

colnames(tripdata\_2020\_09)

## [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(tripdata\_2020\_10)

## [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(tripdata\_2020\_11)

## [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(tripdata\_2020\_12)

## [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(tripdata\_2021\_01)

## [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(tripdata\_2021\_02)

## [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(tripdata\_2021\_03)

## [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(tripdata\_2021\_04)

## [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(tripdata\_2021\_05)

## [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(tripdata\_2021\_06)

## [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(tripdata\_2021\_07)

## [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(tripdata\_2021\_08)

## [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"

Now to check dataframe structures for incongruencies

str(tripdata\_2020\_09)

## spec\_tbl\_df [532,958 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:532958] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...  
## $ rideable\_type : chr [1:532958] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:532958], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...  
## $ ended\_at : POSIXct[1:532958], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...  
## $ start\_station\_name: chr [1:532958] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...  
## $ start\_station\_id : num [1:532958] 52 NA NA 246 24 94 291 NA NA NA ...  
## $ end\_station\_name : chr [1:532958] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Montrose Harbor" ...  
## $ end\_station\_id : num [1:532958] 112 NA NA 249 24 NA 256 NA NA NA ...  
## $ start\_lat : num [1:532958] 41.9 41.9 41.9 42 41.9 ...  
## $ start\_lng : num [1:532958] -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num [1:532958] 41.9 41.9 41.9 42 41.9 ...  
## $ end\_lng : num [1:532958] -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:532958] "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\_double(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_double(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(tripdata\_2020\_10)

## spec\_tbl\_df [388,653 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE261B9E854" ...  
## $ rideable\_type : chr [1:388653] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:388653], format: "2020-10-31 19:39:43" "2020-10-31 23:50:08" ...  
## $ ended\_at : POSIXct[1:388653], format: "2020-10-31 19:57:12" "2020-11-01 00:04:16" ...  
## $ start\_station\_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave" "Stony Island Ave & 67th St" "Clark St & Grace St" ...  
## $ start\_station\_id : num [1:388653] 313 227 102 165 190 359 313 125 NA 174 ...  
## $ end\_station\_name : chr [1:388653] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "University Ave & 57th St" "Broadway & Sheridan Rd" ...  
## $ end\_station\_id : num [1:388653] 125 260 423 256 185 53 125 313 199 635 ...  
## $ start\_lat : num [1:388653] 41.9 41.9 41.8 42 41.9 ...  
## $ start\_lng : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num [1:388653] 41.9 41.9 41.8 42 41.9 ...  
## $ end\_lng : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:388653] "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\_double(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_double(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(tripdata\_2020\_11)

## spec\_tbl\_df [259,716 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533E89C32080B9E" ...  
## $ rideable\_type : chr [1:259716] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:259716], format: "2020-11-01 13:36:00" "2020-11-01 10:03:26" ...  
## $ ended\_at : POSIXct[1:259716], format: "2020-11-01 13:45:40" "2020-11-01 10:14:45" ...  
## $ start\_station\_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shore Dr & Monroe St" "Leavitt St & Chicago Ave" ...  
## $ start\_station\_id : num [1:259716] 110 672 76 659 2 72 76 NA 58 394 ...  
## $ end\_station\_name : chr [1:259716] "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal St & Polk St" "Stave St & Armitage Ave" ...  
## $ end\_station\_id : num [1:259716] 211 29 41 185 2 76 72 NA 288 273 ...  
## $ start\_lat : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:259716] -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:259716] -87.6 -87.7 -87.6 -87.7 -87.6 ...  
## $ member\_casual : chr [1:259716] "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\_double(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_double(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(tripdata\_2020\_12)

## spec\_tbl\_df [131,573 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE119628E44F871E" ...  
## $ rideable\_type : chr [1:131573] "classic\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:131573], format: "2020-12-27 12:44:29" "2020-12-18 17:37:15" ...  
## $ ended\_at : POSIXct[1:131573], format: "2020-12-27 12:55:06" "2020-12-18 17:44:19" ...  
## $ start\_station\_name: chr [1:131573] "Aberdeen St & Jackson Blvd" NA NA NA ...  
## $ start\_station\_id : chr [1:131573] "13157" NA NA NA ...  
## $ end\_station\_name : chr [1:131573] "Desplaines St & Kinzie St" NA NA NA ...  
## $ end\_station\_id : chr [1:131573] "TA1306000003" NA NA NA ...  
## $ start\_lat : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...  
## $ start\_lng : num [1:131573] -87.7 -87.7 -87.7 -87.7 -87.6 ...  
## $ end\_lat : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...  
## $ end\_lng : num [1:131573] -87.6 -87.7 -87.7 -87.7 -87.6 ...  
## $ member\_casual : chr [1:131573] "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(tripdata\_2021\_01)

## spec\_tbl\_df [96,834 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...  
## $ rideable\_type : chr [1:96834] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:96834], format: "2021-01-23 16:14:19" "2021-01-27 18:43:08" ...  
## $ ended\_at : POSIXct[1:96834], format: "2021-01-23 16:24:44" "2021-01-27 18:47:12" ...  
## $ start\_station\_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" ...  
## $ start\_station\_id : chr [1:96834] "17660" "17660" "17660" "17660" ...  
## $ end\_station\_name : chr [1:96834] NA NA NA NA ...  
## $ end\_station\_id : chr [1:96834] NA NA NA NA ...  
## $ start\_lat : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...  
## $ end\_lat : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...  
## $ member\_casual : chr [1:96834] "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(tripdata\_2021\_02)

## spec\_tbl\_df [49,622 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75B" ...  
## $ rideable\_type : chr [1:49622] "classic\_bike" "classic\_bike" "electric\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:49622], format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" ...  
## $ ended\_at : POSIXct[1:49622], format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" ...  
## $ start\_station\_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St" "Wood St & Chicago Ave" ...  
## $ start\_station\_id : chr [1:49622] "525" "525" "KA1503000012" "637" ...  
## $ end\_station\_name : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Randolph St" "Honore St & Division St" ...  
## $ end\_station\_id : chr [1:49622] "660" "16806" "TA1305000029" "TA1305000034" ...  
## $ start\_lat : num [1:49622] 42 42 41.9 41.9 41.8 ...  
## $ start\_lng : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num [1:49622] 42 42 41.9 41.9 41.8 ...  
## $ end\_lng : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...  
## $ member\_casual : chr [1:49622] "member" "casual" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(tripdata\_2021\_03)

## spec\_tbl\_df [228,496 x 13] (S3: spec\_tbl\_df/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" ...  
## - 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(tripdata\_2021\_04)

## spec\_tbl\_df [337,230 x 13] (S3: spec\_tbl\_df/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" ...  
## - 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(tripdata\_2021\_05)

## spec\_tbl\_df [531,633 x 13] (S3: spec\_tbl\_df/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" ...  
## - 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(tripdata\_2021\_06)

## spec\_tbl\_df [729,595 x 13] (S3: spec\_tbl\_df/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" ...  
## - 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(tripdata\_2021\_07)

## spec\_tbl\_df [822,410 x 13] (S3: spec\_tbl\_df/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" ...  
## - 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(tripdata\_2021\_08)

## spec\_tbl\_df [804,352 x 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>

Converting column data types to match for consolidation

tripdata\_2020\_09 <- mutate(tripdata\_2020\_09, start\_station\_id = as.character(start\_station\_id), end\_station\_id = as.character(end\_station\_id))  
tripdata\_2020\_10 <- mutate(tripdata\_2020\_10, start\_station\_id = as.character(start\_station\_id), end\_station\_id = as.character(end\_station\_id))  
tripdata\_2020\_11 <- mutate(tripdata\_2020\_11, start\_station\_id = as.character(start\_station\_id), end\_station\_id = as.character(end\_station\_id))

Combine dataframes into one

all\_data <- bind\_rows(tripdata\_2020\_09, tripdata\_2020\_10, tripdata\_2020\_11, tripdata\_2020\_12, tripdata\_2021\_01, tripdata\_2021\_02, tripdata\_2021\_03, tripdata\_2021\_04, tripdata\_2021\_05, tripdata\_2021\_06, tripdata\_2021\_07, tripdata\_2021\_08)

Removing columns unnecessary for analysis

all\_data <- all\_data %>%   
 select(-c(start\_lat, start\_lng, end\_lat, end\_lng))

#### Data cleaning and prep

first a brief overall look at the data and it’s properties

colnames(all\_data) #List of column names

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

nrow(all\_data) #How many rows are in data frame?

## [1] 4913072

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

## [1] 4913072 9

head(all\_data) #See the first 6 rows of data frame. Also tail(all\_trips)

## # A tibble: 6 x 9  
## ride\_id rideable\_type started\_at ended\_at start\_station\_n~  
## <chr> <chr> <dttm> <dttm> <chr>   
## 1 2B22BD5F95FB2629 electric\_bike 2020-09-17 14:27:11 2020-09-17 14:44:24 Michigan Ave & ~  
## 2 A7FB70B4AFC6CAF2 electric\_bike 2020-09-17 15:07:31 2020-09-17 15:07:45 W Oakdale Ave &~  
## 3 86057FA01BAC778E electric\_bike 2020-09-17 15:09:04 2020-09-17 15:09:35 W Oakdale Ave &~  
## 4 57F6DC9A153DB98C electric\_bike 2020-09-17 18:10:46 2020-09-17 18:35:49 Ashland Ave & B~  
## 5 B9C4712F78C1AE68 electric\_bike 2020-09-17 15:16:13 2020-09-17 15:52:55 Fairbanks Ct & ~  
## 6 378BBCE1E444EB80 electric\_bike 2020-09-17 18:37:04 2020-09-17 19:23:28 Clark St & Armi~  
## # ... with 4 more variables: start\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_id <chr>, member\_casual <chr>

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

## tibble [4,913,072 x 9] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:4913072] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...  
## $ rideable\_type : chr [1:4913072] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:4913072], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...  
## $ ended\_at : POSIXct[1:4913072], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...  
## $ start\_station\_name: chr [1:4913072] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...  
## $ start\_station\_id : chr [1:4913072] "52" NA NA "246" ...  
## $ end\_station\_name : chr [1:4913072] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Montrose Harbor" ...  
## $ end\_station\_id : chr [1:4913072] "112" NA NA "249" ...  
## $ member\_casual : chr [1:4913072] "casual" "casual" "casual" "casual" ...

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

## ride\_id rideable\_type started\_at   
## Length:4913072 Length:4913072 Min. :2020-09-01 00:00:07   
## Class :character Class :character 1st Qu.:2020-12-09 18:57:27   
## Mode :character Mode :character Median :2021-05-26 16:51:05   
## Mean :2021-04-11 06:27:48   
## 3rd Qu.:2021-07-17 03:36:04   
## Max. :2021-08-31 23:59:35   
## ended\_at start\_station\_name start\_station\_id   
## Min. :2020-09-01 00:04:43 Length:4913072 Length:4913072   
## 1st Qu.:2020-12-09 18:29:54 Class :character Class :character   
## Median :2021-05-26 17:12:45 Mode :character Mode :character   
## Mean :2021-04-11 06:48:57   
## 3rd Qu.:2021-07-17 04:16:25   
## Max. :2021-09-01 17:37:35   
## end\_station\_name end\_station\_id member\_casual   
## Length:4913072 Length:4913072 Length:4913072   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

There are a few issues that needs to be fixed:

1. In the “member\_casual” column, there are two names for members (“member” and “Subscriber”) and two names for casual riders (“Customer” and “casual”). These will be consolidated from four to two labels
2. The data can only be aggregated at the ride-level, which is too granular. Additional columns of data – such as day, month, year – will be added to provide additional opportunities to aggregate the data.
3. A calculation will be needed to determine the length of each trip, this will add another column which will be called “ride\_length” to the data frame.

Reassigning values to the correct labels:

all\_data <- all\_data %>%   
 mutate(member\_casual = recode(member\_casual  
 ,"Subscriber" = "member"  
 ,"Customer" = "casual"))

Adding additional columns for detailed analysis

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

Adding a function to calculate the ride length

all\_data$ride\_length <- difftime(all\_data$ended\_at,all\_data$started\_at)

inspecting the structure of the dataframe

str(all\_data)

## tibble [4,913,072 x 15] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:4913072] "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...  
## $ rideable\_type : chr [1:4913072] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:4913072], format: "2020-09-17 14:27:11" "2020-09-17 15:07:31" ...  
## $ ended\_at : POSIXct[1:4913072], format: "2020-09-17 14:44:24" "2020-09-17 15:07:45" ...  
## $ start\_station\_name: chr [1:4913072] "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...  
## $ start\_station\_id : chr [1:4913072] "52" NA NA "246" ...  
## $ end\_station\_name : chr [1:4913072] "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Montrose Harbor" ...  
## $ end\_station\_id : chr [1:4913072] "112" NA NA "249" ...  
## $ member\_casual : chr [1:4913072] "casual" "casual" "casual" "casual" ...  
## $ date : Date[1:4913072], format: "2020-09-17" "2020-09-17" ...  
## $ month : chr [1:4913072] "09" "09" "09" "09" ...  
## $ day : chr [1:4913072] "17" "17" "17" "17" ...  
## $ year : chr [1:4913072] "2020" "2020" "2020" "2020" ...  
## $ day\_of\_week : chr [1:4913072] "Thursday" "Thursday" "Thursday" "Thursday" ...  
## $ ride\_length : 'difftime' num [1:4913072] 1033 14 31 1503 ...  
## ..- attr(\*, "units")= chr "secs"

Converting ride\_length to numeric data type to perform calculations

is.factor(all\_data$ride\_length)

## [1] FALSE

all\_data$ride\_length <- as.numeric(as.character(all\_data$ride\_length))  
is.numeric(all\_data$ride\_length)

## [1] TRUE

Remove bad data: A few hundred times bikes where checked out for service and repairs by “HQ QR” while some rides show a negative ride time. A new database called “all\_data\_v2” will be created by filtering out such entries

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

#### Descriptive Analysis

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

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual, FUN = mean)

## all\_data\_v2$member\_casual all\_data\_v2$ride\_length  
## 1 casual 2144.9725  
## 2 member 865.3867

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual, FUN = median)

## all\_data\_v2$member\_casual all\_data\_v2$ride\_length  
## 1 casual 1055  
## 2 member 618

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual, FUN = max)

## all\_data\_v2$member\_casual all\_data\_v2$ride\_length  
## 1 casual 3356649  
## 2 member 1870176

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual, FUN = min)

## all\_data\_v2$member\_casual all\_data\_v2$ride\_length  
## 1 casual 0  
## 2 member 0

Now a look at the average ride time by day for members vs casual users

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual + all\_data\_v2$day\_of\_week, FUN = mean)

## all\_data\_v2$member\_casual all\_data\_v2$day\_of\_week all\_data\_v2$ride\_length  
## 1 casual Friday 2057.1686  
## 2 member Friday 853.1022  
## 3 casual Monday 2122.2700  
## 4 member Monday 832.1529  
## 5 casual Saturday 2303.5785  
## 6 member Saturday 962.2649  
## 7 casual Sunday 2465.3786  
## 8 member Sunday 990.0270  
## 9 casual Thursday 1856.7460  
## 10 member Thursday 809.1883  
## 11 casual Tuesday 1902.8772  
## 12 member Tuesday 813.4891  
## 13 casual Wednesday 1913.1304  
## 14 member Wednesday 817.1874

Fixing order of days in dataframe

all\_data\_v2$day\_of\_week <- ordered(all\_data\_v2$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

Now recalculating average ride time by day for member vs casual users

aggregate(all\_data\_v2$ride\_length ~ all\_data\_v2$member\_casual + all\_data\_v2$day\_of\_week, FUN = mean)

## all\_data\_v2$member\_casual all\_data\_v2$day\_of\_week all\_data\_v2$ride\_length  
## 1 casual Sunday 2465.3786  
## 2 member Sunday 990.0270  
## 3 casual Monday 2122.2700  
## 4 member Monday 832.1529  
## 5 casual Tuesday 1902.8772  
## 6 member Tuesday 813.4891  
## 7 casual Wednesday 1913.1304  
## 8 member Wednesday 817.1874  
## 9 casual Thursday 1856.7460  
## 10 member Thursday 809.1883  
## 11 casual Friday 2057.1686  
## 12 member Friday 853.1022  
## 13 casual Saturday 2303.5785  
## 14 member Saturday 962.2649

Analyze rider data by user type and week day

all\_data\_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)

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

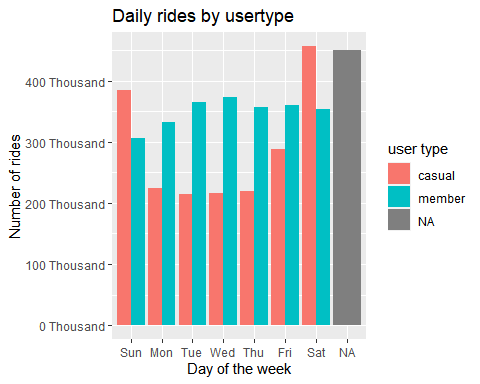
## # A tibble: 15 x 4  
## # Groups: member\_casual [3]  
## member\_casual weekday number\_of\_rides average\_duration  
## <chr> <ord> <int> <dbl>  
## 1 casual Sun 385495 2465.  
## 2 casual Mon 224316 2122.  
## 3 casual Tue 215069 1903.  
## 4 casual Wed 215627 1913.  
## 5 casual Thu 220439 1857.  
## 6 casual Fri 288555 2057.  
## 7 casual Sat 457163 2304.  
## 8 member Sun 306804 990.  
## 9 member Mon 333219 832.  
## 10 member Tue 366045 813.  
## 11 member Wed 373211 817.  
## 12 member Thu 357171 809.  
## 13 member Fri 360326 853.  
## 14 member Sat 354241 962.  
## 15 <NA> <NA> 449991 NA

#### Visualizations

Visualization of number rides by user type

all\_data\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 summarise(number\_of\_rides = n(), average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, weekday) %>%   
 ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge")+  
 scale\_y\_continuous(labels = function(x) paste0(x/1e3, " Thousand"))+  
 labs(title = "Daily rides by usertype", x = "Day of the week", y = "Number of rides", fill = "user type")

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



Visualization of average ride duration(seconds)

all\_data\_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(title = "Average daily ride duration by user type", x ="Day of the week", y = "Average ride duration", fill = "User type")

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

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

