Case Study: Bike Sharing

```
library(tidyverse)  #helps wrangle data

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(lubridate)  #helps wrangle date attributes

##
## Attaching package: 'lubridate'

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

library(skimr) #get summary data
library(janitor)

##
## Attaching package: 'janitor'

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

library(dplyr)
```

STEP 1: COLLECT DATA

```
#=========================================================
```

```
Trips_Apr20 <- read_csv('202004-divvy-tripdata.csv')

##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_May20 <- read_csv('202005-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_June20 <- read_csv('202006-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_July20 <- read_csv('202007-divvy-tripdata.csv')
```

```
##
## -- Column specification ---------------------------------------------------
------
## 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()
## )
```

```
Trips_Aug20 <- read_csv('202008-divvy-tripdata.csv')
```

```
##
## -- Column specification ---------------------------------------------------
------
## 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()
## )
```

```
Trips_Sep20 <- read_csv('202009-divvy-tripdata.csv')
```

```
##
## -- Column specification ---------------------------------------------------
```

```
------
## 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()
## )
```

```
Trips_Oct20 <- read_csv('202010-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Nov20 <- read_csv('202011-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Dec20 <- read_csv('202012-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Jan21 <- read_csv('202101-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Feb21 <- read_csv('202102-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Mar21 <- read_csv('202103-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

```
Trips_Apr21 <- read_csv('202004-divvy-tripdata.csv')
```

```
##
## -- Column specification --------------------------------------------------
------
## 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()
## )
```

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(Trips_Apr20)

##  [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(Trips_May20)

##  [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(Trips_June20)

##  [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(Trips_July20)

##  [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(Trips_Aug20)
```

```
##  [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(Trips_Sep20)
```

```
##  [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(Trips_Oct20)
```

```
##  [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(Trips_Nov20)
```

```
##  [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(Trips_Dec20)
```

```
##  [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(Trips_Jan21)
```

```
##  [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(Trips_Feb21)
```

```
##  [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(Trips_Mar21)

##  [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(Trips_Apr21)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

Inspect the dataframes and look for incongruencies

#===

```
str(Trips_Apr20)

## spec_tbl_df [84,776 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:84776] "A847FADBBC638E45" "5405B80E996FF60D"
"5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:84776] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:84776], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:84776], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:84776] "Eckhart Park" "Drake Ave & Fullerton
Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : num [1:84776] 86 503 142 216 125 173 35 434 627 377
...
##  $ end_station_name  : chr [1:84776] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : num [1:84776] 152 499 255 657 323 35 635 382 359
508 ...
##  $ start_lat         : num [1:84776] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:84776] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:84776] "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_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()
##   .. )
```

```
str(Trips_May20)
```

```
## spec_tbl_df [200,274 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:200274] "02668AD35674B983"
"7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA84" ...
##  $ rideable_type     : chr [1:200274] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:200274], format: "2020-05-27 10:03:52"
"2020-05-25 10:47:11" ...
##  $ ended_at          : POSIXct[1:200274], format: "2020-05-27 10:16:49"
"2020-05-25 11:05:40" ...
##  $ start_station_name: chr [1:200274] "Franklin St & Jackson Blvd" "Clark
St & Wrightwood Ave" "Kedzie Ave & Milwaukee Ave" "Clarendon Ave & Leland
Ave" ...
##  $ start_station_id  : num [1:200274] 36 340 260 251 261 206 261 180 331
219 ...
##  $ end_station_name  : chr [1:200274] "Wabash Ave & Grand Ave" "Clark St &
Leland Ave" "Kedzie Ave & Milwaukee Ave" "Lake Shore Dr & Wellington Ave" ...
##  $ end_station_id    : num [1:200274] 199 326 260 157 206 22 261 180 300
305 ...
##  $ start_lat         : num [1:200274] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:200274] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:200274] 41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:200274] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:200274] "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_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()
##   .. )
```

```
str(Trips_June20)
```

```
## spec_tbl_df [343,005 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:343005] "8CD5DE2C2B6C4CFC"
"9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85FA1" ...
##  $ rideable_type     : chr [1:343005] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:343005], format: "2020-06-13 23:24:48"
"2020-06-26 07:26:10" ...
##  $ ended_at          : POSIXct[1:343005], format: "2020-06-13 23:36:55"
"2020-06-26 07:31:58" ...
##  $ start_station_name: chr [1:343005] "Wilton Ave & Belmont Ave" "Federal
St & Polk St" "Daley Center Plaza" "Broadway & Cornelia Ave" ...
##  $ start_station_id  : num [1:343005] 117 41 81 303 327 327 41 115 338 84
...
##  $ end_station_name  : chr [1:343005] "Damen Ave & Clybourn Ave" "Daley
Center Plaza" "State St & Harrison St" "Broadway & Berwyn Ave" ...
##  $ end_station_id    : num [1:343005] 163 81 5 294 117 117 81 303 164 53
...
##  $ start_lat         : num [1:343005] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:343005] -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:343005] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:343005] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:343005] "casual" "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_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()
##   .. )
```

```
str(Trips_July20)
```

```
## spec_tbl_df [551,480 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:551480] "762198876D69004D"
"BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr [1:551480] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:551480], format: "2020-07-09 15:22:02"
"2020-07-24 23:56:30" ...
##  $ ended_at          : POSIXct[1:551480], format: "2020-07-09 15:25:52"
"2020-07-25 00:20:17" ...
##  $ start_station_name: chr [1:551480] "Ritchie Ct & Banks St" "Halsted St
& Roscoe St" "Lake Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...
##  $ start_station_id  : num [1:551480] 180 299 329 181 268 635 113 211 176
31 ...
##  $ end_station_name  : chr [1:551480] "Wells St & Evergreen Ave" "Broadway
& Ridge Ave" "Clark St & Wellington Ave" "Clark St & Armitage Ave" ...
##  $ end_station_id    : num [1:551480] 291 461 156 94 301 289 140 31 191
142 ...
##  $ start_lat         : num [1:551480] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:551480] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:551480] 41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:551480] -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:551480] "member" "member" "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()
##   .. )
```

```
str(Trips_Aug20)
```

```
## spec_tbl_df [622,361 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:622361] "322BD23D287743ED"
"2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79FBBD412E578A7" ...
##  $ rideable_type     : chr [1:622361] "docked_bike" "electric_bike"
"electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:622361], format: "2020-08-20 18:08:14"
```

```
"2020-08-27 18:46:04" ...
##  $ ended_at          : POSIXct[1:622361], format: "2020-08-20 18:17:51"
"2020-08-27 19:54:51" ...
##  $ start_station_name: chr [1:622361] "Lake Shore Dr & Diversey Pkwy"
"Michigan Ave & 14th St" "Columbus Dr & Randolph St" "Daley Center Plaza" ...
##  $ start_station_id  : num [1:622361] 329 168 195 81 658 658 196 67 153
177 ...
##  $ end_station_name  : chr [1:622361] "Clark St & Lincoln Ave" "Michigan
Ave & 14th St" "State St & Randolph St" "State St & Kinzie St" ...
##  $ end_station_id    : num [1:622361] 141 168 44 47 658 658 49 229 225 305
...
##  $ start_lat         : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:622361] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:622361] -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:622361] "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_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()
##   .. )
```

```
str(Trips_Sep20)
```

```
## 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()
##   .. )
```

```
str(Trips_Oct20)
```

```
## 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()
##   .. )
```

```
str(Trips_Nov20)
```

```
## 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()
##   .. )
```

```
str(Trips_Dec20)
```

```
## 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()
##   .. )
```

```
str(Trips_Jan21)
```

```
## 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()
##   .. )
```

```
str(Trips_Feb21)
```

```
## 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()
##   .. )
```

```
str(Trips_Mar21)
```

```
## 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()
##   .. )
```

```
str(Trips_Apr21)
```

```
## spec_tbl_df [84,776 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:84776] "A847FADBBC638E45" "5405B80E996FF60D"
"5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:84776] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:84776], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:84776], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:84776] "Eckhart Park" "Drake Ave & Fullerton
Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : num [1:84776] 86 503 142 216 125 173 35 434 627 377
...
##  $ end_station_name  : chr [1:84776] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : num [1:84776] 152 499 255 657 323 35 635 382 359
508 ...
##  $ start_lat         : num [1:84776] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:84776] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:84776] "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_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()
##   .. )
```

we can compare column datatype across all dataframe by using compare_df_cols when we have large dataset, that would be more easy

```
compare_df_cols(Trips_Apr20, Trips_May20, Trips_June20, Trips_July20,
Trips_Aug20, Trips_Sep20, Trips_Oct20, Trips_Nov20, Trips_Dec20,
Trips_Jan21, Trips_Feb21, Trips_Mar21, Trips_Apr21, return = "mismatch")

##        column_name Trips_Apr20 Trips_May20 Trips_June20 Trips_July20
## 1   end_station_id     numeric     numeric      numeric      numeric
## 2 start_station_id     numeric     numeric      numeric      numeric
##   Trips_Aug20 Trips_Sep20 Trips_Oct20 Trips_Nov20 Trips_Dec20 Trips_Jan21
## 1     numeric     numeric     numeric     numeric   character   character
## 2     numeric     numeric     numeric     numeric   character   character
##   Trips_Feb21 Trips_Mar21 Trips_Apr21
## 1   character   character     numeric
## 2   character   character     numeric
```

Convert end_station_id and start_station_id to character so that they can stack correctly

```
Trips_Apr20 <- mutate(Trips_Apr20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_May20 <- mutate(Trips_May20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_June20 <- mutate(Trips_June20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_July20 <- mutate(Trips_July20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_Aug20 <- mutate(Trips_Aug20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_Sep20 <- mutate(Trips_Sep20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_Oct20 <- mutate(Trips_Oct20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_Nov20 <- mutate(Trips_Nov20, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
Trips_Apr21 <- mutate(Trips_Apr21, end_station_id =
as.character(end_station_id), start_station_id =
as.character(start_station_id))
```

double check column datatype across all dataframe

```
compare_df_cols(Trips_Apr20, Trips_May20, Trips_June20, Trips_July20,
Trips_Aug20, Trips_Sep20, Trips_Oct20, Trips_Nov20, Trips_Dec20,
Trips_Jan21, Trips_Feb21, Trips_Mar21, Trips_Apr21, return = "mismatch")
```

```
##  [1] column_name  Trips_Apr20  Trips_May20  Trips_June20 Trips_July20
##  [6] Trips_Aug20  Trips_Sep20  Trips_Oct20  Trips_Nov20  Trips_Dec20
## [11] Trips_Jan21  Trips_Feb21  Trips_Mar21  Trips_Apr21
## <0 rows> (or 0-length row.names)
```

Stack individual data frames into one big data frame

```
all_trips <- bind_rows(Trips_Apr20, Trips_May20, Trips_June20, Trips_July20,
Trips_Aug20, Trips_Sep20, Trips_Oct20, Trips_Nov20, Trips_Dec20,
Trips_Jan21, Trips_Feb21, Trips_Mar21, Trips_Apr21)
```

Remove unused column

```
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

Rename Columns

```
all_trips <- all_trips %>%  rename(trip_id= ride_id ,ride_type =
rideable_type
                  ,start_time = started_at,end_time =ended_at
                  ,from_station_name = start_station_name
                  ,from_station_id = start_station_id
                  ,to_station_name = end_station_name
                  ,to_station_id = end_station_id
                  ,usertype = member_casual)
```

STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

#== # Inspect the new table that has been created

```
colnames(all_trips)  #List of column names

## [1] "trip_id"           "ride_type"         "start_time"
## [4] "end_time"          "from_station_name" "from_station_id"
## [7] "to_station_name"   "to_station_id"     "usertype"

dim(all_trips)  #Dimensions of the data frame?

## [1] 3574524       9

head(all_trips) #See the first 6 rows of data frame.

## # A tibble: 6 x 9
##   trip_id   ride_type start_time          end_time
from_station_name
##   <chr>     <chr>     <dttm>              <dttm>              <chr>
## 1 A847FADB~ docked_b~ 2020-04-26 17:45:14 2020-04-26 18:12:03 Eckhart Park
## 2 5405B80E~ docked_b~ 2020-04-17 17:08:54 2020-04-17 17:17:03 Drake Ave &
Fulle~
## 3 5DD24A79~ docked_b~ 2020-04-01 17:54:13 2020-04-01 18:08:36 McClurg Ct &
Erie~
## 4 2A59BBDF~ docked_b~ 2020-04-07 12:50:19 2020-04-07 13:02:31 California
Ave & ~
## 5 27AD306C~ docked_b~ 2020-04-18 10:22:59 2020-04-18 11:15:54 Rush St &
Hubbard~
## 6 356216E8~ docked_b~ 2020-04-30 17:55:47 2020-04-30 18:01:11 Mies van der
Rohe~
## # ... with 4 more variables: from_station_id <chr>, to_station_name <chr>,
## #   to_station_id <chr>, usertype <chr>

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

## tibble [3,574,524 x 9] (S3: tbl_df/tbl/data.frame)
##  $ trip_id          : chr [1:3574524] "A847FADBBC638E45"
"5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ ride_type        : chr [1:3574524] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ start_time       : POSIXct[1:3574524], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ end_time         : POSIXct[1:3574524], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ from_station_name: chr [1:3574524] "Eckhart Park" "Drake Ave &
Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ from_station_id  : chr [1:3574524] "86" "503" "142" "216" ...
##  $ to_station_name  : chr [1:3574524] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
```

```
##  $ to_station_id    : chr [1:3574524] "152" "499" "255" "657" ...
##  $ usertype         : chr [1:3574524] "member" "member" "member" "member"
...
```

```
summary(all_trips)  #Statistical summary of data. Mainly for numerics
```

```
##    trip_id           ride_type           start_time                 
##  Length:3574524     Length:3574524     Min.   :2020-04-01 00:00:30  
##  Class :character   Class :character   1st Qu.:2020-07-11 15:53:56  
##  Mode  :character   Mode  :character   Median :2020-08-27 15:44:17  
##                                        Mean   :2020-09-06 13:37:36  
##                                        3rd Qu.:2020-10-17 22:11:16  
##                                        Max.   :2021-03-31 23:59:08  
##     end_time                   from_station_name  from_station_id   
##  Min.   :2020-04-01 00:10:45   Length:3574524     Length:3574524    
##  1st Qu.:2020-07-11 16:27:48   Class :character   Class :character  
##  Median :2020-08-27 16:07:07   Mode  :character   Mode  :character  
##  Mean   :2020-09-06 14:02:38                                        
##  3rd Qu.:2020-10-17 22:36:28                                        
##  Max.   :2021-04-06 11:00:11                                        
##  to_station_name    to_station_id        usertype        
##  Length:3574524     Length:3574524     Length:3574524    
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
```

```
skim(all_trips) #get summary of data, check missing data
```

Data summary

Name	all_trips
Number of rows	3574524
Number of columns	9
_______________________	
Column type frequency:	
character	7
POSIXct	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace

trip_id	0	1.00	16	16	0	3489539	0
ride_type	0	1.00	11	13	0	3	0
from_station_name	122175	0.97	10	53	0	708	0
from_station_id	122801	0.97	1	35	0	1259	0
to_station_name	143341	0.96	10	53	0	706	0
to_station_id	143802	0.96	1	35	0	1259	0
usertype	0	1.00	6	6	0	2	0

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_time	0	1	2020-04-01 00:00:30	2021-03-31 23:59:08	2020-08-27 15:44:17	3040228
end_time	0	1	2020-04-01 00:10:45	2021-04-06 11:00:11	2020-08-27 16:07:07	3027775

Add columns that list the date, month, day, and year of each ride

This will allow us to aggregate ride data for each month, day, or year … before completing these operations we could only aggregate at the ride level

```
all_trips$date <- as.Date(all_trips$start_time) #The default format is yyyy-mm-dd
all_trips$month <- format(as.Date(all_trips$date), "%m")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year <- format(as.Date(all_trips$date), "%Y")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
```

Add a "ride_length" calculation to all_trips (in seconds)

```
all_trips$ride_length <- difftime(all_trips$end_time,all_trips$start_time)
```

Convert "ride_length" from Factor to numeric so we can run calculations on the data

```
is.factor(all_trips$ride_length)
```

```
## [1] FALSE
```

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

```
## [1] TRUE
```

Remove “bad” data

The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Divvy or ride_length was negative

```
skim(all_trips$ride_length)
```

Data summary

Name	all_trips$ride_length
Number of rows	3574524
Number of columns	1
_______________________	
Column type frequency:	
numeric	1
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	1501.77	23732.69	-1742998	474	873	1600	3523202	▁▇▁▁▁

```
all_trips_v2 <- all_trips[!(all_trips$ride_length<0),]
skim(all_trips_v2)
```

Data summary

Name	all_trips_v2
Number of rows	3563921
Number of columns	15
_______________________	

Column type frequency:	
character	11
Date	1
numeric	1
POSIXct	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
trip_id	0	1.00	16	16	0	3479196	0
ride_type	0	1.00	11	13	0	3	0
from_station_name	122128	0.97	10	53	0	708	0
from_station_id	122754	0.97	1	35	0	1259	0
to_station_name	143257	0.96	10	53	0	706	0
to_station_id	143718	0.96	1	35	0	1259	0
usertype	0	1.00	6	6	0	2	0
month	0	1.00	2	2	0	12	0
day	0	1.00	2	2	0	31	0
year	0	1.00	4	4	0	2	0
day_of_week	0	1.00	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2020-04-01	2021-03-31	2020-08-27	363

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ride_length	0	1	1688.35	15883.15	0	477	876	1603	3523202	▇▁▁▁▁

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_time	0	1	2020-04-01 00:00:30	2021-03-31 23:59:08	2020-08-27 14:56:13	3035417
end_time	0	1	2020-04-01 00:10:45	2021-04-06 11:00:11	2020-08-27 15:21:31	3020300

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

#===================================== # Descriptive analysis on ride_length (all figures in seconds)

```
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     477     876    1688    1603 3523202
```

Export to CSV file for further analysis

```
write.csv(all_trips_v2, "data.csv")
```