STAT650\_Midterm\_Section 2\_fordgobike

**Prof. Eric A. Suess**

# Question1: Explain what the GBFS is?

Answer: GBFS ,which is the abbreviaion of General Bikeshare Feed Specification , supports publicly available station bike/dock availability information and does not require the use of an API token. Developers are encouraged to review the GBFS documentation available on Github as they determine how to incorporate BCycle system information into their applications and websites and it is the open data standard for bikeshare and also makes real-time data publicly available online in a uniform format so that maps and transportation based apps can easily incorporate this data into their platforms.

# Question 2: Explain any difficulties you encountered getting the code to work.

Answer: 1. When I try to download the each file , i need to add the “lib” before “curl” like typed as “libcurl”. 2. I am also having trouble with the setting my own directory , after this i learned how to set directory by my own every time i work with some files . 3.I found out that some date of birth were inaccurate and found out there were some outliers regarding to age variable .and also i need to remove these outliers. 4.Data type of age changed from int to char and had trouble with changing them while calculating the mean.

library(tidyverse)

## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

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

library(tictoc)  
library(ggmap)  
library(skimr)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(forcats)

Set working directory.

setwd("C:/Users/Patime/Desktop/midterm/data")

Download the files into the data directory. First one is not zipped, the remaining are zipped.

Read the.csv files

fordgobike2017 <- read\_csv(file="./data/2017-fordgobike-tripdata.csv")

## Warning: 2 parsing failures.  
## row # A tibble: 2 x 5 col row col expected actual file expected <int> <chr> <chr> <chr> <chr> actual 1 7717 start\_ti~ closing quote at end~ "" './data/2017-fordgobike-~ file 2 7717 <NA> 15 columns 2 colum~ './data/2017-fordgobike-~

fordgobike201801 <- read\_csv(file="./data/201801-fordgobike-tripdata.csv")  
fordgobike201802 <- read\_csv(file="./data/201802-fordgobike-tripdata.csv")  
fordgobike201803 <- read\_csv(file="./data/201803-fordgobike-tripdata.csv")  
fordgobike201804 <- read\_csv(file="./data/201804-fordgobike-tripdata.csv")  
fordgobike201805 <- read\_csv(file="./data/201805-fordgobike-tripdata.csv")  
fordgobike201806 <- read\_csv(file="./data/201806-fordgobike-tripdata.csv")  
fordgobike201807 <- read\_csv(file="./data/201807-fordgobike-tripdata.csv")  
fordgobike201808 <- read\_csv(file="./data/201808-fordgobike-tripdata.csv")

Check the head() and tail() of the data.frames that are loaded.

head(fordgobike2017)

## # A tibble: 6 x 15  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 80110 2017-12-31 16:57:39 2018-01-01 15:12:50 74  
## 2 78800 2017-12-31 15:56:34 2018-01-01 13:49:55 284  
## 3 45768 2017-12-31 22:45:48 2018-01-01 11:28:36 245  
## 4 62172 2017-12-31 17:31:10 2018-01-01 10:47:23 60  
## 5 43603 2017-12-31 14:23:14 2018-01-01 02:29:57 239  
## 6 9226 2017-12-31 22:51:00 2018-01-01 01:24:47 30  
## # ... with 11 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>

head(fordgobike201801)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 75284 2018-01-31 22:52:35 2018-02-01 19:47:19 120  
## 2 85422 2018-01-31 16:13:34 2018-02-01 15:57:17 15  
## 3 71576 2018-01-31 14:23:55 2018-02-01 10:16:52 304  
## 4 61076 2018-01-31 14:53:23 2018-02-01 07:51:20 75  
## 5 39966 2018-01-31 19:52:24 2018-02-01 06:58:31 74  
## 6 6477 2018-01-31 22:58:44 2018-02-01 00:46:41 236  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201802)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 598 2018-02-28 23:59:47 2018-03-01 00:09:45 284  
## 2 943 2018-02-28 23:21:16 2018-02-28 23:36:59 6  
## 3 18587 2018-02-28 18:20:55 2018-02-28 23:30:42 93  
## 4 18558 2018-02-28 18:20:53 2018-02-28 23:30:12 93  
## 5 885 2018-02-28 23:15:12 2018-02-28 23:29:58 308  
## 6 921 2018-02-28 23:14:19 2018-02-28 23:29:40 312  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201803)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 71766 2018-03-31 16:58:33 2018-04-01 12:54:39 4  
## 2 62569 2018-03-31 19:03:35 2018-04-01 12:26:25 78  
## 3 56221 2018-03-31 20:13:13 2018-04-01 11:50:14 258  
## 4 85844 2018-03-31 11:28:07 2018-04-01 11:18:52 186  
## 5 1566 2018-03-31 23:37:56 2018-04-01 00:04:02 193  
## 6 281 2018-03-31 23:58:07 2018-04-01 00:02:49 197  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201804)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 72393 2018-04-30 22:49:32 2018-05-01 18:56:06 4  
## 2 81034 2018-04-30 17:46:04 2018-05-01 16:16:39 122  
## 3 86142 2018-04-30 16:07:13 2018-05-01 16:02:56 41  
## 4 68839 2018-04-30 17:11:57 2018-05-01 12:19:16 284  
## 5 59091 2018-04-30 18:45:21 2018-05-01 11:10:13 196  
## 6 68093 2018-04-30 15:39:18 2018-05-01 10:34:12 21  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201805)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 56791 2018-05-31 21:41:51 2018-06-01 13:28:22 44  
## 2 52797 2018-05-31 18:39:53 2018-06-01 09:19:51 186  
## 3 43204 2018-05-31 21:09:48 2018-06-01 09:09:52 17  
## 4 67102 2018-05-31 14:09:54 2018-06-01 08:48:17 106  
## 5 58883 2018-05-31 16:07:23 2018-06-01 08:28:47 16  
## 6 22858 2018-05-31 23:06:40 2018-06-01 05:27:38 163  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201806)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 59088 2018-06-30 23:32:44 2018-07-01 15:57:33 76   
## 2 60358 2018-06-30 21:48:19 2018-07-01 14:34:18 248   
## 3 63654 2018-06-30 20:26:53 2018-07-01 14:07:47 23   
## 4 50508 2018-06-30 20:29:59 2018-07-01 10:31:48 58   
## 5 51697 2018-06-30 18:24:56 2018-07-01 08:46:33 196   
## 6 36708 2018-06-30 20:25:34 2018-07-01 06:37:22 8   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201807)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 59989 2018-07-31 18:20:32 2018-08-01 11:00:22 197   
## 2 60232 2018-07-31 17:24:26 2018-08-01 10:08:18 77   
## 3 43864 2018-07-31 21:03:26 2018-08-01 09:14:30 NULL   
## 4 51522 2018-07-31 18:54:23 2018-08-01 09:13:06 114   
## 5 83380 2018-07-31 09:22:29 2018-08-01 08:32:09 213   
## 6 49546 2018-07-31 18:44:11 2018-08-01 08:29:57 139   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

head(fordgobike201808)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 69465 2018-08-31 19:18:08 2018-09-01 14:35:54 3   
## 2 60644 2018-08-31 21:40:01 2018-09-01 14:30:46 50   
## 3 57922 2018-08-31 20:04:28 2018-09-01 12:09:50 17   
## 4 82573 2018-08-31 13:07:14 2018-09-01 12:03:28 19   
## 5 57677 2018-08-31 19:48:45 2018-09-01 11:50:03 3   
## 6 76960 2018-08-31 13:07:46 2018-09-01 10:30:26 19   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike2017)

## # A tibble: 6 x 15  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 1160 2017-12-27 12:17:07 2017-12-27 12:36:27 84  
## 2 1190 2017-12-27 12:16:26 2017-12-27 12:36:16 84  
## 3 1156 2017-12-27 12:16:48 2017-12-27 12:36:05 84  
## 4 1134 2017-12-27 12:17:04 2017-12-27 12:35:58 84  
## 5 1383 2017-12-27 12:11:57 2017-12-27 12:35:00 36  
## 6 246 2017-12-27 12:30:33 NA NA  
## # ... with 11 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>

tail(fordgobike201801)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 752 2018-01-01 00:19:13 2018-01-01 00:31:46 23  
## 2 695 2018-01-01 00:19:58 2018-01-01 00:31:33 23  
## 3 600 2018-01-01 00:19:48 2018-01-01 00:29:49 17  
## 4 1151 2018-01-01 00:09:31 2018-01-01 00:28:43 97  
## 5 714 2018-01-01 00:07:52 2018-01-01 00:19:47 74  
## 6 145 2018-01-01 00:07:41 2018-01-01 00:10:06 316  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201802)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 404 2018-02-01 00:16:47 2018-02-01 00:23:31 89  
## 2 136 2018-02-01 00:20:14 2018-02-01 00:22:30 182  
## 3 345 2018-02-01 00:08:39 2018-02-01 00:14:25 122  
## 4 439 2018-02-01 00:02:25 2018-02-01 00:09:45 284  
## 5 524 2018-02-01 00:00:05 2018-02-01 00:08:49 113  
## 6 319 2018-02-01 00:00:39 2018-02-01 00:05:59 72  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201803)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 594 2018-03-01 00:39:50 2018-03-01 00:49:45 66  
## 2 471 2018-03-01 00:28:29 2018-03-01 00:36:20 180  
## 3 285 2018-03-01 00:17:32 2018-03-01 00:22:18 183  
## 4 408 2018-03-01 00:13:37 2018-03-01 00:20:25 27  
## 5 368 2018-03-01 00:14:14 2018-03-01 00:20:22 27  
## 6 164 2018-03-01 00:13:31 2018-03-01 00:16:15 240  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201804)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 656 2018-04-01 00:06:53 2018-04-01 00:17:50 45  
## 2 887 2018-04-01 00:00:08 2018-04-01 00:14:55 194  
## 3 387 2018-04-01 00:08:06 2018-04-01 00:14:33 30  
## 4 480 2018-04-01 00:06:21 2018-04-01 00:14:21 44  
## 5 503 2018-04-01 00:04:36 2018-04-01 00:13:00 100  
## 6 192 2018-04-01 00:02:03 2018-04-01 00:05:16 176  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201805)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 392 2018-05-01 00:12:49 2018-05-01 00:19:22 336  
## 2 645 2018-05-01 00:07:24 2018-05-01 00:18:09 186  
## 3 135 2018-05-01 00:15:53 2018-05-01 00:18:09 244  
## 4 316 2018-05-01 00:10:04 2018-05-01 00:15:20 30  
## 5 183 2018-05-01 00:12:01 2018-05-01 00:15:05 243  
## 6 78 2018-05-01 00:02:01 2018-05-01 00:03:20 106  
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <int>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201806)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 119 2018-06-01 00:09:28 2018-06-01 00:11:28 105   
## 2 491 2018-06-01 00:03:12 2018-06-01 00:11:24 41   
## 3 507 2018-06-01 00:02:28 2018-06-01 00:10:55 118   
## 4 377 2018-06-01 00:04:10 2018-06-01 00:10:27 186   
## 5 391 2018-06-01 00:03:37 2018-06-01 00:10:08 30   
## 6 283 2018-06-01 00:02:37 2018-06-01 00:07:21 36   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201807)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 165 2018-07-01 00:08:12 2018-07-01 00:10:58 278   
## 2 130 2018-07-01 00:07:39 2018-07-01 00:09:49 253   
## 3 445 2018-07-01 00:02:16 2018-07-01 00:09:42 179   
## 4 490 2018-07-01 00:00:41 2018-07-01 00:08:51 307   
## 5 219 2018-07-01 00:04:32 2018-07-01 00:08:12 70   
## 6 368 2018-07-01 00:01:22 2018-07-01 00:07:31 4   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

tail(fordgobike201808)

## # A tibble: 6 x 16  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <chr>   
## 1 1132 2018-08-01 00:00:23 2018-08-01 00:19:16 343   
## 2 631 2018-08-01 00:04:38 2018-08-01 00:15:10 247   
## 3 325 2018-08-01 00:07:56 2018-08-01 00:13:21 100   
## 4 678 2018-08-01 00:00:01 2018-08-01 00:11:20 5   
## 5 405 2018-08-01 00:01:26 2018-08-01 00:08:12 100   
## 6 285 2018-08-01 00:02:35 2018-08-01 00:07:21 314   
## # ... with 12 more variables: start\_station\_name <chr>,  
## # start\_station\_latitude <dbl>, start\_station\_longitude <dbl>,  
## # end\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_latitude <dbl>, end\_station\_longitude <dbl>,  
## # bike\_id <int>, user\_type <chr>, member\_birth\_year <int>,  
## # member\_gender <chr>, bike\_share\_for\_all\_trip <chr>

by dim()function , we can get total number of row observations which is 519700 and total number column variables which is 15 in fordgobike2017 data frame.

dim(fordgobike2017)

## [1] 7717 15

fordgobike2017 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 7717

view total number of row observations and total number column variables which are from Jan of 2018 to Aug of 2018 in fordgobike2018 data frame.

fordgobike201801 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 94802

fordgobike201802 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 106718

fordgobike201803 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 111382

fordgobike201804 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 131169

fordgobike201805 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 179125

fordgobike201806 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 195968

fordgobike201807 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 199222

by glimpse()function , we can see that start\_station\_id data types for below months are chr .

glimpse(fordgobike201805)

## Observations: 179,125  
## Variables: 16  
## $ duration\_sec <int> 56791, 52797, 43204, 67102, 58883, 228...  
## $ start\_time <dttm> 2018-05-31 21:41:51, 2018-05-31 18:39...  
## $ end\_time <dttm> 2018-06-01 13:28:22, 2018-06-01 09:19...  
## $ start\_station\_id <int> 44, 186, 17, 106, 16, 163, 197, 61, 61...  
## $ start\_station\_name <chr> "Civic Center/UN Plaza BART Station (M...  
## $ start\_station\_latitude <dbl> 37.78107, 37.80132, 37.79225, 37.76324...  
## $ start\_station\_longitude <dbl> -122.4117, -122.2626, -122.3971, -122....  
## $ end\_station\_id <int> 78, 338, 93, 47, 30, 212, 197, 8, 8, 8...  
## $ end\_station\_name <chr> "Folsom St at 9th St", "13th St at Fra...  
## $ end\_station\_latitude <dbl> 37.77372, 37.80319, 37.77041, 37.78095...  
## $ end\_station\_longitude <dbl> -122.4116, -122.2706, -122.3912, -122....  
## $ bike\_id <int> 1230, 3414, 2677, 4224, 3392, 1235, 15...  
## $ user\_type <chr> "Customer", "Subscriber", "Customer", ...  
## $ member\_birth\_year <int> NA, 1983, NA, 1979, 1986, 1992, 1985, ...  
## $ member\_gender <chr> NA, "Male", NA, "Male", "Male", "Male"...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "Y...

glimpse(fordgobike201806)

## Observations: 195,968  
## Variables: 16  
## $ duration\_sec <int> 59088, 60358, 63654, 50508, 51697, 367...  
## $ start\_time <dttm> 2018-06-30 23:32:44, 2018-06-30 21:48...  
## $ end\_time <dttm> 2018-07-01 15:57:33, 2018-07-01 14:34...  
## $ start\_station\_id <chr> "76", "248", "23", "58", "196", "8", "...  
## $ start\_station\_name <chr> "McCoppin St at Valencia St", "Telegra...  
## $ start\_station\_latitude <dbl> 37.77166, 37.85596, 37.79146, 37.77662...  
## $ start\_station\_longitude <dbl> -122.4224, -122.2598, -122.3910, -122....  
## $ end\_station\_id <chr> "95", "239", "50", "88", "272", "4", "...  
## $ end\_station\_name <chr> "Sanchez St at 15th St", "Bancroft Way...  
## $ end\_station\_latitude <dbl> 37.76622, 37.86881, 37.78053, 37.77003...  
## $ end\_station\_longitude <dbl> -122.4311, -122.2588, -122.3903, -122....  
## $ bike\_id <int> 2100, 653, 3235, 3675, 3232, 577, 1764...  
## $ user\_type <chr> "Subscriber", "Customer", "Subscriber"...  
## $ member\_birth\_year <int> 1975, NA, 1962, 1992, 1989, NA, NA, 19...  
## $ member\_gender <chr> "Male", NA, "Female", "Male", "Female"...  
## $ bike\_share\_for\_all\_trip <chr> "Yes", "No", "No", "No", "No", "No", "...

glimpse(fordgobike201807)

## Observations: 199,222  
## Variables: 16  
## $ duration\_sec <int> 59989, 60232, 43864, 51522, 83380, 495...  
## $ start\_time <dttm> 2018-07-31 18:20:32, 2018-07-31 17:24...  
## $ end\_time <dttm> 2018-08-01 11:00:22, 2018-08-01 10:08...  
## $ start\_station\_id <chr> "197", "77", "NULL", "114", "213", "13...  
## $ start\_station\_name <chr> "El Embarcadero at Grand Ave", "11th S...  
## $ start\_station\_latitude <dbl> 37.80885, 37.77351, 37.41000, 37.76448...  
## $ start\_station\_longitude <dbl> -122.2497, -122.4160, -121.9400, -122....  
## $ end\_station\_id <chr> "181", "356", "NULL", "345", "198", "3...  
## $ end\_station\_name <chr> "Grand Ave at Webster St", "Valencia S...  
## $ end\_station\_latitude <dbl> 37.81138, 37.76919, 37.41000, 37.76647...  
## $ end\_station\_longitude <dbl> -122.2652, -122.4223, -121.9400, -122....  
## $ bike\_id <int> 1953, 3010, 4273, 1043, 1336, 697, 605...  
## $ user\_type <chr> "Customer", "Subscriber", "Subscriber"...  
## $ member\_birth\_year <int> 1995, 1994, 1998, 1990, 1982, 1991, 19...  
## $ member\_gender <chr> "Male", "Female", "Male", "Female", "M...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "N...

# Question 3: The analysis is to work with the 2017 and 2018 data. How many bike rentals were there in 2017? How many bike rentals were there in 2018? How many bike rentals have there been since the beginning of Ford GoBikes?

Answer: Number of Bike rentals in 2017 was 272 Number of Bike rentals in 2018 was 962 Number of Bike rentals since the beginning of Ford GoBikes = 272+962=1234

# Question 4: There is a part of the code that uses the as.integer() function for some reason. Explain what this function is being used for in the code.

Answer: Data type we used for Start\_Station\_id, end\_station\_id variables in the data frame, which are fordgobike201806, fordgobike201807 and fordgobike201806 data frame ,are character type.So we need to convert this kind of type to integer type by using as.integer()function. we use this function in order to match data type with other data frame. and by the way , if we use bind()function ,combine ()function to combine the data frames with different data type variables ,we will encounter an error . So before we use any combining function , we have to be sure our variable with the same data type .

Converting the start\_station\_id data type to integer for forgobike201806, forgobike201807, and forgobike201808 to match with other data. Combine all the multiple 2018 data set into single data set

fordgobike201806 <- fordgobike201806 %>%  
 mutate(start\_station\_id = as.integer(start\_station\_id),  
 end\_station\_id= as.integer(end\_station\_id) )

## Warning in evalq(as.integer(start\_station\_id), <environment>): NAs  
## introduced by coercion

## Warning in evalq(as.integer(end\_station\_id), <environment>): NAs introduced  
## by coercion

fordgobike201807 <- fordgobike201807 %>%  
 mutate(start\_station\_id = as.integer(start\_station\_id),  
 end\_station\_id= as.integer(end\_station\_id) )

## Warning in evalq(as.integer(start\_station\_id), <environment>): NAs  
## introduced by coercion

## Warning in evalq(as.integer(end\_station\_id), <environment>): NAs introduced  
## by coercion

fordgobike2018 <- bind\_rows(fordgobike201801, fordgobike201802, fordgobike201803, fordgobike201804,  
 fordgobike201805, fordgobike201806, fordgobike201807)  
  
glimpse(fordgobike2018)

## Observations: 1,018,386  
## Variables: 16  
## $ duration\_sec <int> 75284, 85422, 71576, 61076, 39966, 647...  
## $ start\_time <dttm> 2018-01-31 22:52:35, 2018-01-31 16:13...  
## $ end\_time <dttm> 2018-02-01 19:47:19, 2018-02-01 15:57...  
## $ start\_station\_id <int> 120, 15, 304, 75, 74, 236, 110, 81, 13...  
## $ start\_station\_name <chr> "Mission Dolores Park", "San Francisco...  
## $ start\_station\_latitude <dbl> 37.76142, 37.79539, 37.34876, 37.77379...  
## $ start\_station\_longitude <dbl> -122.4264, -122.3942, -121.8948, -122....  
## $ end\_station\_id <int> 285, 15, 296, 47, 19, 160, 134, 93, 4,...  
## $ end\_station\_name <chr> "Webster St at O'Farrell St", "San Fra...  
## $ end\_station\_latitude <dbl> 37.78352, 37.79539, 37.32600, 37.78095...  
## $ end\_station\_longitude <dbl> -122.4312, -122.3942, -121.8771, -122....  
## $ bike\_id <int> 2765, 2815, 3039, 321, 617, 1306, 3571...  
## $ user\_type <chr> "Subscriber", "Customer", "Customer", ...  
## $ member\_birth\_year <int> 1986, NA, 1996, NA, 1991, NA, 1988, 19...  
## $ member\_gender <chr> "Male", NA, "Male", NA, "Male", NA, "M...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "N...

we combined data frame fordgobike2017 and fordgobike2018 .

dim(fordgobike2017)

## [1] 7717 15

fordgobike2017 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 7717

nrow(fordgobike201801) + nrow(fordgobike201802) + nrow(fordgobike201803) + nrow(fordgobike201804)

## [1] 444071

dim(fordgobike2018)

## [1] 1018386 16

fordgobike2018 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1018386

fordgobike <- bind\_rows(fordgobike2017, fordgobike2018)  
  
dim(fordgobike)

## [1] 1026103 16

# Question 5: How is the Age variable created? Are there any outliers in the data? How many outliers have you removed? State what you think is a good cut off is to remove any outliers.

Answer: if we look at the fordgobike data frame,we can figure out that member\_ Birth\_year variable’s value is given. So before create the new Age variable to calculate the age, we need to use mutate()function to create a new variable which the age and also this variable is created by subtracting the current year from the given variable (member\_birth\_date) .

if we use geom\_histogram() function with filter()function on fordgobike data ,we can see the outliers. from histogram below, we can see that the data is showing a highly right skewed pattern. however we infer that there exests an outlier. Major portion of data lies between age range between 18 and 80 is majority proportion of data .except of this , there still showing some age group which are greater than above age 80.so we need to remove these outliers or replace them with NA value.

So I removed anything beyond 80 yrs which is considered as outliers and I replaced with NA values.

add another new variable Age by using mutate ()function

fordgobike %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1026103

dim(fordgobike)

## [1] 1026103 16

fordgobike <- fordgobike %>% mutate(age = 2018 - member\_birth\_year)  
  
fordgobike %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1026103

dim(fordgobike)

## [1] 1026103 17

fordgobike <- fordgobike %>% mutate(year=year(start\_time), month=month(start\_time), day=day(start\_time) )

fordgobike %>% group\_by( age ) %>% count()

## # A tibble: 83 x 2  
## # Groups: age [83]  
## age n  
## <dbl> <int>  
## 1 18 812  
## 2 19 5627  
## 3 20 10152  
## 4 21 13960  
## 5 22 14644  
## 6 23 23435  
## 7 24 28003  
## 8 25 39322  
## 9 26 35346  
## 10 27 40385  
## # ... with 73 more rows

fordgobike %>% group\_by( age ) %>% summary()

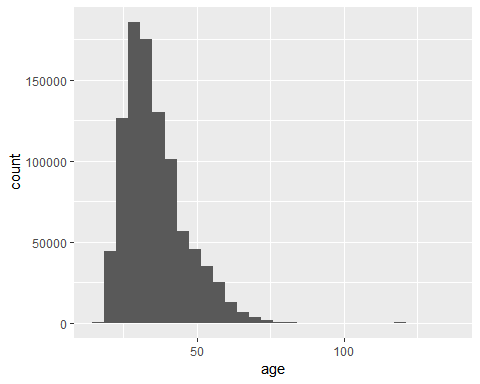
## duration\_sec start\_time   
## Min. : 61.0 Min. :2017-12-26 14:16:57   
## 1st Qu.: 351.0 1st Qu.:2018-03-14 19:57:38   
## Median : 557.0 Median :2018-05-11 08:13:15   
## Mean : 888.5 Mean :2018-05-02 13:28:52   
## 3rd Qu.: 879.0 3rd Qu.:2018-06-22 12:22:50   
## Max. :86366.0 Max. :2018-07-31 23:57:19   
##   
## end\_time start\_station\_id start\_station\_name  
## Min. :2017-12-27 12:35:00 Min. : 3.0 Length:1026103   
## 1st Qu.:2018-03-14 20:13:33 1st Qu.: 30.0 Class :character   
## Median :2018-05-11 08:23:44 Median : 84.0 Mode :character   
## Mean :2018-05-02 13:43:52 Mean :114.1   
## 3rd Qu.:2018-06-22 12:39:13 3rd Qu.:180.0   
## Max. :2018-08-01 11:00:22 Max. :357.0   
## NA's :1 NA's :5246   
## start\_station\_latitude start\_station\_longitude end\_station\_id   
## Min. :37.31 Min. :-122.44 Min. : 3.0   
## 1st Qu.:37.77 1st Qu.:-122.41 1st Qu.: 30.0   
## Median :37.78 Median :-122.40 Median : 81.0   
## Mean :37.77 Mean :-122.35 Mean :112.5   
## 3rd Qu.:37.80 3rd Qu.:-122.29 3rd Qu.:180.0   
## Max. :45.51 Max. : -73.57 Max. :357.0   
## NA's :1 NA's :1 NA's :5246   
## end\_station\_name end\_station\_latitude end\_station\_longitude  
## Length:1026103 Min. :37.28 Min. :-122.44   
## Class :character 1st Qu.:37.77 1st Qu.:-122.41   
## Mode :character Median :37.78 Median :-122.40   
## Mean :37.77 Mean :-122.35   
## 3rd Qu.:37.80 3rd Qu.:-122.29   
## Max. :45.51 Max. : -73.57   
## NA's :1 NA's :1   
## bike\_id user\_type member\_birth\_year member\_gender   
## Min. : 11 Length:1026103 Min. :1881 Length:1026103   
## 1st Qu.:1186 Class :character 1st Qu.:1977 Class :character   
## Median :2290 Mode :character Median :1985 Mode :character   
## Mean :2197 Mean :1982   
## 3rd Qu.:3224 3rd Qu.:1990   
## Max. :4307 Max. :2000   
## NA's :1 NA's :72788   
## bike\_share\_for\_all\_trip age year month   
## Length:1026103 Min. : 18.00 Min. :2017 Min. : 1.000   
## Class :character 1st Qu.: 28.00 1st Qu.:2018 1st Qu.: 3.000   
## Mode :character Median : 33.00 Median :2018 Median : 5.000   
## Mean : 35.54 Mean :2018 Mean : 4.605   
## 3rd Qu.: 41.00 3rd Qu.:2018 3rd Qu.: 6.000   
## Max. :137.00 Max. :2018 Max. :12.000   
## NA's :72788   
## day   
## Min. : 1.00   
## 1st Qu.: 9.00   
## Median :16.00   
## Mean :16.15   
## 3rd Qu.:24.00   
## Max. :31.00   
##

skim(fordgobike)

## Skim summary statistics  
## n obs: 1026103   
## n variables: 20   
##   
## -- Variable type:character ---------------------------------------------------------------------  
## variable missing complete n min max empty n\_unique  
## bike\_share\_for\_all\_trip 7717 1018386 1026103 2 3 0 2  
## end\_station\_name 1 1026102 1026103 4 63 0 311  
## member\_gender 72526 953577 1026103 4 6 0 3  
## start\_station\_name 1 1026102 1026103 4 63 0 311  
## user\_type 1 1026102 1026103 8 10 0 2  
##   
## -- Variable type:integer -----------------------------------------------------------------------  
## variable missing complete n mean sd p0 p25 p50  
## bike\_id 1 1026102 1026103 2197.07 1195.28 11 1186 2290  
## day 0 1026103 1026103 16.15 8.81 1 9 16  
## duration\_sec 0 1026103 1026103 888.55 2578.05 61 351 557  
## end\_station\_id 5246 1020857 1026103 112.48 95.51 3 30 81  
## member\_birth\_year 72788 953315 1026103 1982.46 10.53 1881 1977 1985  
## start\_station\_id 5246 1020857 1026103 114.15 95.63 3 30 84  
## p75 p100 hist  
## 3224 4307 <U+2586><U+2586><U+2586><U+2586><U+2587><U+2587><U+2587><U+2586>  
## 24 31 <U+2587><U+2587><U+2587><U+2587><U+2586><U+2587><U+2587><U+2587>  
## 879 86366 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>  
## 180 357 <U+2587><U+2585><U+2583><U+2582><U+2582><U+2582><U+2581><U+2581>  
## 1990 2000 <U+2581><U+2581><U+2581><U+2581><U+2581><U+2582><U+2587><U+2587>  
## 180 357 <U+2587><U+2586><U+2583><U+2582><U+2582><U+2582><U+2581><U+2581>  
##   
## -- Variable type:numeric -----------------------------------------------------------------------  
## variable missing complete n mean sd p0  
## age 72788 953315 1026103 35.54 10.53 18   
## end\_station\_latitude 1 1026102 1026103 37.77 0.1 37.28  
## end\_station\_longitude 1 1026102 1026103 -122.35 0.17 -122.44  
## month 0 1026103 1026103 4.61 2.04 1   
## start\_station\_latitude 1 1026102 1026103 37.77 0.1 37.31  
## start\_station\_longitude 1 1026102 1026103 -122.35 0.17 -122.44  
## year 0 1026103 1026103 2017.99 0.086 2017   
## p25 p50 p75 p100 hist  
## 28 33 41 137 <U+2587><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>  
## 37.77 37.78 37.8 45.51 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>  
## -122.41 -122.4 -122.29 -73.57 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>  
## 3 5 6 12 <U+2585><U+2583><U+2587><U+2585><U+2585><U+2581><U+2581><U+2581>  
## 37.77 37.78 37.8 45.51 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>  
## -122.41 -122.4 -122.29 -73.57 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>  
## 2018 2018 2018 2018 <U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>  
##   
## -- Variable type:POSIXct -----------------------------------------------------------------------  
## variable missing complete n min max median  
## end\_time 1 1026102 1026103 2017-12-27 2018-08-01 2018-05-11  
## start\_time 0 1026103 1026103 2017-12-26 2018-07-31 2018-05-11  
## n\_unique  
## 1026045  
## 1026040

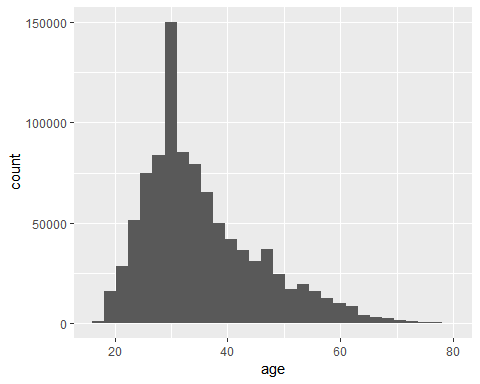
fordgobike %>% ggplot(aes(x=age)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



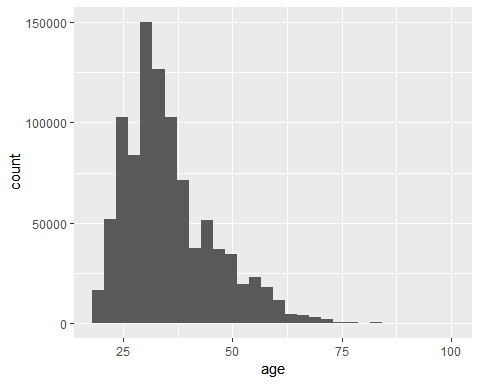
fordgobike %>% filter(age <= 80) %>% ggplot(aes(x=age)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



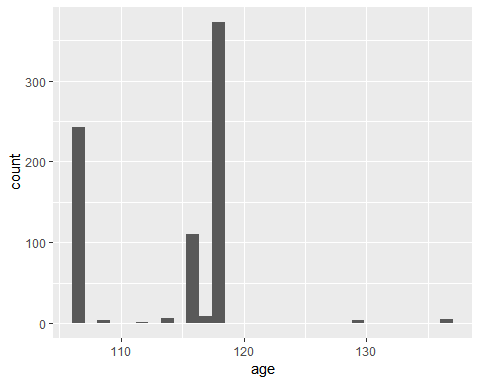
fordgobike %>% filter(age <= 100) %>% ggplot(aes(x=age)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



fordgobike %>% filter(age > 100) %>% ggplot(aes(x=age)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



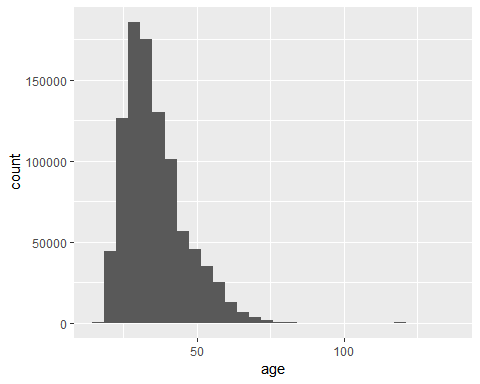
fordgobike %>% group\_by( member\_gender, age ) %>% count()

## # A tibble: 201 x 3  
## # Groups: member\_gender, age [201]  
## member\_gender age n  
## <chr> <dbl> <int>  
## 1 Female 18 268  
## 2 Female 19 1514  
## 3 Female 20 2412  
## 4 Female 21 4301  
## 5 Female 22 4617  
## 6 Female 23 7112  
## 7 Female 24 7902  
## 8 Female 25 9598  
## 9 Female 26 9915  
## 10 Female 27 12507  
## # ... with 191 more rows

fordgobike %>% ggplot(aes(x=age, class=member\_gender)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 72788 rows containing non-finite values (stat\_bin).



fordgobike %>% ggplot(aes(x=age, class=member\_gender)) + geom\_histogram(aes(y=..density..))

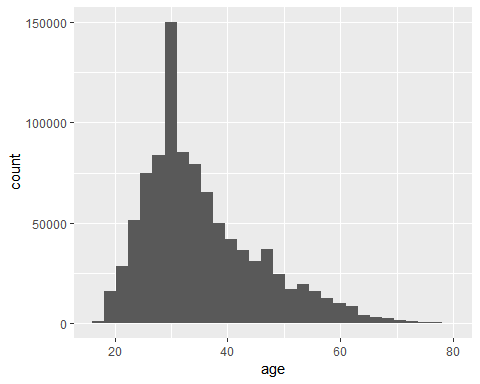
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 72788 rows containing non-finite values (stat\_bin).



fordgobike %>% filter(age <= 80) %>% ggplot(aes(x=age)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

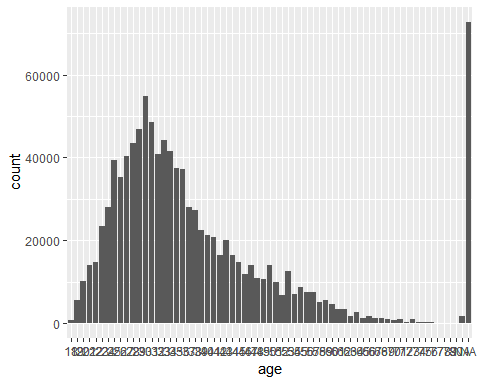
 Replacing the Age > 80 yrs with NA value.

Fordgobike\_outlier\_NA<-fordgobike %>% mutate(age=replace(age,age>80,"NA"))  
Fordgobike\_outlier\_NA\_Only<-Fordgobike\_outlier\_NA %>% filter(age=="NA") %>% count()  
glimpse(Fordgobike\_outlier\_NA)

## Observations: 1,026,103  
## Variables: 20  
## $ duration\_sec <int> 80110, 78800, 45768, 62172, 43603, 922...  
## $ start\_time <dttm> 2017-12-31 16:57:39, 2017-12-31 15:56...  
## $ end\_time <dttm> 2018-01-01 15:12:50, 2018-01-01 13:49...  
## $ start\_station\_id <int> 74, 284, 245, 60, 239, 30, 259, 284, 2...  
## $ start\_station\_name <chr> "Laguna St at Hayes St", "Yerba Buena ...  
## $ start\_station\_latitude <dbl> 37.77643, 37.78487, 37.87035, 37.77452...  
## $ start\_station\_longitude <dbl> -122.4262, -122.4009, -122.2678, -122....  
## $ end\_station\_id <int> 43, 96, 245, 5, 247, 30, 259, 284, 20,...  
## $ end\_station\_name <chr> "San Francisco Public Library (Grove S...  
## $ end\_station\_latitude <dbl> 37.77877, 37.76621, 37.87035, 37.78390...  
## $ end\_station\_longitude <dbl> -122.4159, -122.4266, -122.2678, -122....  
## $ bike\_id <int> 96, 88, 1094, 2831, 3167, 1487, 3539, ...  
## $ user\_type <chr> "Customer", "Customer", "Customer", "C...  
## $ member\_birth\_year <int> 1987, 1965, NA, NA, 1997, NA, 1991, NA...  
## $ member\_gender <chr> "Male", "Female", NA, NA, "Female", NA...  
## $ bike\_share\_for\_all\_trip <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ age <chr> "31", "53", NA, NA, "21", NA, "27", NA...  
## $ year <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 20...  
## $ month <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12...  
## $ day <int> 31, 31, 31, 31, 31, 31, 31, 31, 31, 31...

Fordgobike\_outlier\_NA %>% ggplot(aes(x=age)) + geom\_histogram(stat = "count",binwidth = 30)

## Warning: Ignoring unknown parameters: binwidth, bins, pad



**# Question 6: In 2018, what month had the highest number of riders? What month had the lowest number of riders? Interpret any seasonal patterns.**

Answer: from below results, July had the highest number of riders which is 199222 .and the January month had the lowest number of riders which is 94802.

if we interpret these two results with seasonal pattern , we can say because on July month, most of the people are willing to go outside and they will be more active ,so total rides would peak to the highest number and its number went down as the weather changes .In summer season like July ,June was first and second highest number .but in Winters ,there was showing less bike rides ,since winter is cold ,it is reasonabale to see less bike riders outside .therefore total bike rides follows to the seasonal pattern .it increases from winter to spring and reaches peak in July .

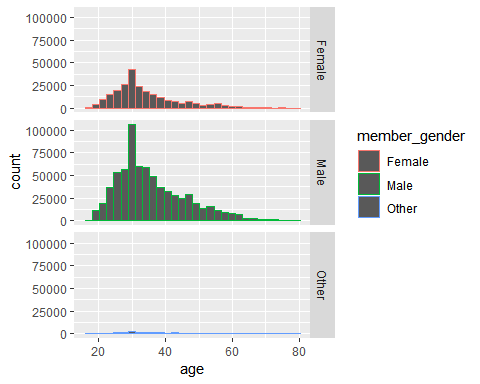
Filtering the 2018 data from the combined and outlier removed/NA data

fordgobike\_outlier\_NA\_2018 <- Fordgobike\_outlier\_NA %>% filter(year==2018)  
Riders<-fordgobike\_outlier\_NA\_2018 %>% group\_by(month) %>% summarise(total\_rides=n()) %>% arrange(desc(total\_rides))  
Riders

## # A tibble: 7 x 2  
## month total\_rides  
## <dbl> <int>  
## 1 7 199222  
## 2 6 195968  
## 3 5 179125  
## 4 4 131169  
## 5 3 111382  
## 6 2 106718  
## 7 1 94802

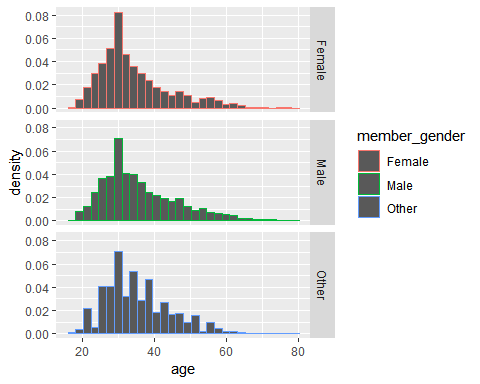
fordgobike %>% filter(age <= 80) %>% ggplot(aes(x=age, color=member\_gender)) +   
 geom\_histogram(position="identity") +   
 facet\_grid(member\_gender ~ .)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



fordgobike %>% filter(age <= 80) %>% ggplot(aes(x=age, color=member\_gender)) +   
 geom\_histogram(aes(y=..density..),position="identity") +   
 facet\_grid(member\_gender ~ .)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Question 7: What start station had the highest number of rides? That is, which start station was used most to start rides?

Answer: Start\_station\_ID 15 had highest number of rides. That is San Francisco Ferry Building was used most to start rides.

Group by station id and name

station<-Fordgobike\_outlier\_NA %>% group\_by(start\_station\_id,start\_station\_name) %>% summarise(A=n()) %>% arrange(desc(A))  
station

## # A tibble: 312 x 3  
## # Groups: start\_station\_id [309]  
## start\_station\_id start\_station\_name A  
## <int> <chr> <int>  
## 1 15 San Francisco Ferry Building (Harry Bridges Pla~ 22368  
## 2 30 San Francisco Caltrain (Townsend St at 4th St) 20427  
## 3 81 Berry St at 4th St 19778  
## 4 6 The Embarcadero at Sansome St 19509  
## 5 67 San Francisco Caltrain Station 2 (Townsend St ~ 19338  
## 6 58 Market St at 10th St 19089  
## 7 21 Montgomery St BART Station (Market St at 2nd St) 18217  
## 8 3 Powell St BART Station (Market St at 4th St) 17229  
## 9 16 Steuart St at Market St 15785  
## 10 22 Howard St at Beale St 14890  
## # ... with 302 more rows

# Question 8: What was the Age of the youngest rider? What was the Age of the oldest rider, after removing the outliers? What was the mean Age of the rider? What was the mean Age of the Female riders? What was the mean Age of the Male riders?

Answer: Age of the youngest rider ——18 yrs

Age of the oldest rider —— 80 yrs

mean age of rider —— 36.54

Mean age of Male rider - —–36.54

Mean age of Female rider - 34.69

Fordgobike\_outlier\_NA %>% group\_by(age) %>% count() %>% arrange(desc(age))

## # A tibble: 65 x 2  
## # Groups: age [65]  
## age n  
## <chr> <int>  
## 1 NA 1572  
## 2 80 31  
## 3 79 13  
## 4 78 35  
## 5 77 45  
## 6 76 184  
## 7 75 293  
## 8 74 220  
## 9 73 923  
## 10 72 205  
## # ... with 55 more rows

Fordgobike\_outlier\_NA$age<-as.integer(Fordgobike\_outlier\_NA$age)

## Warning: NAs introduced by coercion

Fordgobike\_outlier\_NA %>% na.omit(age)

## # A tibble: 940,636 x 20  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 75284 2018-01-31 22:52:35 2018-02-01 19:47:19 120  
## 2 71576 2018-01-31 14:23:55 2018-02-01 10:16:52 304  
## 3 39966 2018-01-31 19:52:24 2018-02-01 06:58:31 74  
## 4 453 2018-01-31 23:53:53 2018-02-01 00:01:26 110  
## 5 180 2018-01-31 23:52:09 2018-01-31 23:55:10 81  
## 6 996 2018-01-31 23:34:56 2018-01-31 23:51:32 134  
## 7 825 2018-01-31 23:34:14 2018-01-31 23:47:59 305  
## 8 432 2018-01-31 23:34:26 2018-01-31 23:41:39 89  
## 9 601 2018-01-31 23:29:46 2018-01-31 23:39:48 223  
## 10 887 2018-01-31 23:24:16 2018-01-31 23:39:04 308  
## # ... with 940,626 more rows, and 16 more variables:  
## # start\_station\_name <chr>, start\_station\_latitude <dbl>,  
## # start\_station\_longitude <dbl>, end\_station\_id <int>,  
## # end\_station\_name <chr>, end\_station\_latitude <dbl>,  
## # end\_station\_longitude <dbl>, bike\_id <int>, user\_type <chr>,  
## # member\_birth\_year <int>, member\_gender <chr>,  
## # bike\_share\_for\_all\_trip <chr>, age <int>, year <dbl>, month <dbl>,  
## # day <int>

glimpse(Fordgobike\_outlier\_NA)

## Observations: 1,026,103  
## Variables: 20  
## $ duration\_sec <int> 80110, 78800, 45768, 62172, 43603, 922...  
## $ start\_time <dttm> 2017-12-31 16:57:39, 2017-12-31 15:56...  
## $ end\_time <dttm> 2018-01-01 15:12:50, 2018-01-01 13:49...  
## $ start\_station\_id <int> 74, 284, 245, 60, 239, 30, 259, 284, 2...  
## $ start\_station\_name <chr> "Laguna St at Hayes St", "Yerba Buena ...  
## $ start\_station\_latitude <dbl> 37.77643, 37.78487, 37.87035, 37.77452...  
## $ start\_station\_longitude <dbl> -122.4262, -122.4009, -122.2678, -122....  
## $ end\_station\_id <int> 43, 96, 245, 5, 247, 30, 259, 284, 20,...  
## $ end\_station\_name <chr> "San Francisco Public Library (Grove S...  
## $ end\_station\_latitude <dbl> 37.77877, 37.76621, 37.87035, 37.78390...  
## $ end\_station\_longitude <dbl> -122.4159, -122.4266, -122.2678, -122....  
## $ bike\_id <int> 96, 88, 1094, 2831, 3167, 1487, 3539, ...  
## $ user\_type <chr> "Customer", "Customer", "Customer", "C...  
## $ member\_birth\_year <int> 1987, 1965, NA, NA, 1997, NA, 1991, NA...  
## $ member\_gender <chr> "Male", "Female", NA, NA, "Female", NA...  
## $ bike\_share\_for\_all\_trip <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ age <int> 31, 53, NA, NA, 21, NA, 27, NA, NA, NA...  
## $ year <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 20...  
## $ month <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12...  
## $ day <int> 31, 31, 31, 31, 31, 31, 31, 31, 31, 31...

Fordgobike\_outlier\_NA %>% group\_by(age) %>% summary()

## duration\_sec start\_time   
## Min. : 61.0 Min. :2017-12-26 14:16:57   
## 1st Qu.: 351.0 1st Qu.:2018-03-14 19:57:38   
## Median : 557.0 Median :2018-05-11 08:13:15   
## Mean : 888.5 Mean :2018-05-02 13:28:52   
## 3rd Qu.: 879.0 3rd Qu.:2018-06-22 12:22:50   
## Max. :86366.0 Max. :2018-07-31 23:57:19   
##   
## end\_time start\_station\_id start\_station\_name  
## Min. :2017-12-27 12:35:00 Min. : 3.0 Length:1026103   
## 1st Qu.:2018-03-14 20:13:33 1st Qu.: 30.0 Class :character   
## Median :2018-05-11 08:23:44 Median : 84.0 Mode :character   
## Mean :2018-05-02 13:43:52 Mean :114.1   
## 3rd Qu.:2018-06-22 12:39:13 3rd Qu.:180.0   
## Max. :2018-08-01 11:00:22 Max. :357.0   
## NA's :1 NA's :5246   
## start\_station\_latitude start\_station\_longitude end\_station\_id   
## Min. :37.31 Min. :-122.44 Min. : 3.0   
## 1st Qu.:37.77 1st Qu.:-122.41 1st Qu.: 30.0   
## Median :37.78 Median :-122.40 Median : 81.0   
## Mean :37.77 Mean :-122.35 Mean :112.5   
## 3rd Qu.:37.80 3rd Qu.:-122.29 3rd Qu.:180.0   
## Max. :45.51 Max. : -73.57 Max. :357.0   
## NA's :1 NA's :1 NA's :5246   
## end\_station\_name end\_station\_latitude end\_station\_longitude  
## Length:1026103 Min. :37.28 Min. :-122.44   
## Class :character 1st Qu.:37.77 1st Qu.:-122.41   
## Mode :character Median :37.78 Median :-122.40   
## Mean :37.77 Mean :-122.35   
## 3rd Qu.:37.80 3rd Qu.:-122.29   
## Max. :45.51 Max. : -73.57   
## NA's :1 NA's :1   
## bike\_id user\_type member\_birth\_year member\_gender   
## Min. : 11 Length:1026103 Min. :1881 Length:1026103   
## 1st Qu.:1186 Class :character 1st Qu.:1977 Class :character   
## Median :2290 Mode :character Median :1985 Mode :character   
## Mean :2197 Mean :1982   
## 3rd Qu.:3224 3rd Qu.:1990   
## Max. :4307 Max. :2000   
## NA's :1 NA's :72788   
## bike\_share\_for\_all\_trip age year month   
## Length:1026103 Min. :18.00 Min. :2017 Min. : 1.000   
## Class :character 1st Qu.:28.00 1st Qu.:2018 1st Qu.: 3.000   
## Mode :character Median :33.00 Median :2018 Median : 5.000   
## Mean :35.44 Mean :2018 Mean : 4.605   
## 3rd Qu.:41.00 3rd Qu.:2018 3rd Qu.: 6.000   
## Max. :80.00 Max. :2018 Max. :12.000   
## NA's :74360   
## day   
## Min. : 1.00   
## 1st Qu.: 9.00   
## Median :16.00   
## Mean :16.15   
## 3rd Qu.:24.00   
## Max. :31.00   
##

male<-Fordgobike\_outlier\_NA %>% filter(member\_gender=="Male") %>% group\_by(age) %>% summary()  
male

## duration\_sec start\_time   
## Min. : 61.0 Min. :2017-12-26 15:11:36   
## 1st Qu.: 329.0 1st Qu.:2018-03-14 17:57:03   
## Median : 517.0 Median :2018-05-11 06:40:09   
## Mean : 737.5 Mean :2018-05-02 13:24:10   
## 3rd Qu.: 793.0 3rd Qu.:2018-06-22 09:35:19   
## Max. :86142.0 Max. :2018-07-31 23:57:19   
##   
## end\_time start\_station\_id start\_station\_name  
## Min. :2017-12-27 12:35:00 Min. : 3.0 Length:699580   
## 1st Qu.:2018-03-14 18:06:57 1st Qu.: 30.0 Class :character   
## Median :2018-05-11 06:49:59 Median : 81.0 Mode :character   
## Mean :2018-05-02 13:36:28 Mean :112.6   
## 3rd Qu.:2018-06-22 09:44:25 3rd Qu.:176.0   
## Max. :2018-08-01 11:00:22 Max. :357.0   
## NA's :3535   
## start\_station\_latitude start\_station\_longitude end\_station\_id   
## Min. :37.31 Min. :-122.44 Min. : 3.0   
## 1st Qu.:37.77 1st Qu.:-122.41 1st Qu.: 30.0   
## Median :37.78 Median :-122.40 Median : 81.0   
## Mean :37.77 Mean :-122.35 Mean :111.1   
## 3rd Qu.:37.80 3rd Qu.:-122.30 3rd Qu.:176.0   
## Max. :45.51 Max. : -73.57 Max. :357.0   
## NA's :3535   
## end\_station\_name end\_station\_latitude end\_station\_longitude  
## Length:699580 Min. :37.31 Min. :-122.44   
## Class :character 1st Qu.:37.77 1st Qu.:-122.41   
## Mode :character Median :37.78 Median :-122.40   
## Mean :37.77 Mean :-122.35   
## 3rd Qu.:37.80 3rd Qu.:-122.30   
## Max. :45.51 Max. : -73.57   
##   
## bike\_id user\_type member\_birth\_year member\_gender   
## Min. : 11 Length:699580 Min. :1881 Length:699580   
## 1st Qu.:1204 Class :character 1st Qu.:1976 Class :character   
## Median :2317 Mode :character Median :1984 Mode :character   
## Mean :2220 Mean :1982   
## 3rd Qu.:3253 3rd Qu.:1990   
## Max. :4307 Max. :2000   
##   
## bike\_share\_for\_all\_trip age year month   
## Length:699580 Min. :18.00 Min. :2017 Min. : 1.000   
## Class :character 1st Qu.:28.00 1st Qu.:2018 1st Qu.: 3.000   
## Mode :character Median :34.00 Median :2018 Median : 5.000   
## Mean :35.88 Mean :2018 Mean : 4.591   
## 3rd Qu.:42.00 3rd Qu.:2018 3rd Qu.: 6.000   
## Max. :80.00 Max. :2018 Max. :12.000   
## NA's :984   
## day   
## Min. : 1.00   
## 1st Qu.: 9.00   
## Median :16.00   
## Mean :16.13   
## 3rd Qu.:24.00   
## Max. :31.00   
##

Female<-Fordgobike\_outlier\_NA %>% filter(member\_gender=="Female") %>% group\_by(age) %>% summary()  
Female

## duration\_sec start\_time   
## Min. : 61.0 Min. :2017-12-26 14:16:57   
## 1st Qu.: 392.0 1st Qu.:2018-03-19 09:57:18   
## Median : 619.0 Median :2018-05-14 07:36:18   
## Mean : 915.1 Mean :2018-05-04 13:42:17   
## 3rd Qu.: 960.0 3rd Qu.:2018-06-24 07:54:24   
## Max. :86208.0 Max. :2018-07-31 23:54:51   
##   
## end\_time start\_station\_id start\_station\_name  
## Min. :2017-12-27 12:42:13 Min. : 3.0 Length:239123   
## 1st Qu.:2018-03-19 10:06:12 1st Qu.: 37.0 Class :character   
## Median :2018-05-14 07:48:29 Median : 93.0 Mode :character   
## Mean :2018-05-04 13:57:33 Mean :120.4   
## 3rd Qu.:2018-06-24 08:04:32 3rd Qu.:188.0   
## Max. :2018-08-01 10:08:18 Max. :357.0   
## NA's :1472   
## start\_station\_latitude start\_station\_longitude end\_station\_id  
## Min. :37.31 Min. :-122.44 Min. : 3   
## 1st Qu.:37.77 1st Qu.:-122.41 1st Qu.: 30   
## Median :37.78 Median :-122.40 Median : 93   
## Mean :37.77 Mean :-122.35 Mean :118   
## 3rd Qu.:37.80 3rd Qu.:-122.27 3rd Qu.:186   
## Max. :44.95 Max. : -93.22 Max. :357   
## NA's :1472   
## end\_station\_name end\_station\_latitude end\_station\_longitude  
## Length:239123 Min. :37.31 Min. :-122.44   
## Class :character 1st Qu.:37.77 1st Qu.:-122.41   
## Mode :character Median :37.78 Median :-122.40   
## Mean :37.77 Mean :-122.35   
## 3rd Qu.:37.80 3rd Qu.:-122.27   
## Max. :44.95 Max. : -93.22   
##   
## bike\_id user\_type member\_birth\_year member\_gender   
## Min. : 11 Length:239123 Min. :1900 Length:239123   
## 1st Qu.:1147 Class :character 1st Qu.:1980 Class :character   
## Median :2215 Mode :character Median :1986 Mode :character   
## Mean :2147 Mean :1984   
## 3rd Qu.:3167 3rd Qu.:1991   
## Max. :4307 Max. :2000   
## NA's :262   
## bike\_share\_for\_all\_trip age year month   
## Length:239123 Min. :18.00 Min. :2017 Min. : 1.000   
## Class :character 1st Qu.:27.00 1st Qu.:2018 1st Qu.: 3.000   
## Mode :character Median :32.00 Median :2018 Median : 5.000   
## Mean :34.14 Mean :2018 Mean : 4.659   
## 3rd Qu.:38.00 3rd Qu.:2018 3rd Qu.: 6.000   
## Max. :79.00 Max. :2018 Max. :12.000   
## NA's :539   
## day   
## Min. : 1.00   
## 1st Qu.: 9.00   
## Median :16.00   
## Mean :16.14   
## 3rd Qu.:24.00   
## Max. :31.00   
##

# Question 9: Using the Amelia package and the missmap() function determine the rate of missing data in the month of June.

Answer: 1% of data is missing and 99% of data observed in the month of june.

To calculate missing observation in the month of June

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.5, built: 2018-05-07)  
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

fordgobike\_june<-Fordgobike\_outlier\_NA %>% filter(month==6)  
fordgobike\_june

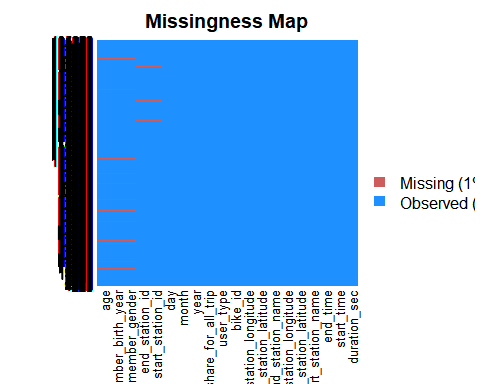
## # A tibble: 195,968 x 20  
## duration\_sec start\_time end\_time start\_station\_id  
## <int> <dttm> <dttm> <int>  
## 1 59088 2018-06-30 23:32:44 2018-07-01 15:57:33 76  
## 2 60358 2018-06-30 21:48:19 2018-07-01 14:34:18 248  
## 3 63654 2018-06-30 20:26:53 2018-07-01 14:07:47 23  
## 4 50508 2018-06-30 20:29:59 2018-07-01 10:31:48 58  
## 5 51697 2018-06-30 18:24:56 2018-07-01 08:46:33 196  
## 6 36708 2018-06-30 20:25:34 2018-07-01 06:37:22 8  
## 7 46380 2018-06-30 17:23:18 2018-07-01 06:16:18 237  
## 8 7224 2018-06-30 23:01:00 2018-07-01 01:01:24 284  
## 9 4294 2018-06-30 23:32:03 2018-07-01 00:43:38 240  
## 10 2209 2018-06-30 23:55:13 2018-07-01 00:32:02 133  
## # ... with 195,958 more rows, and 16 more variables:  
## # start\_station\_name <chr>, start\_station\_latitude <dbl>,  
## # start\_station\_longitude <dbl>, end\_station\_id <int>,  
## # end\_station\_name <chr>, end\_station\_latitude <dbl>,  
## # end\_station\_longitude <dbl>, bike\_id <int>, user\_type <chr>,  
## # member\_birth\_year <int>, member\_gender <chr>,  
## # bike\_share\_for\_all\_trip <chr>, age <int>, year <dbl>, month <dbl>,  
## # day <int>

missmap(fordgobike\_june)

## Warning in if (class(obj) == "amelia") {: the condition has length > 1 and  
## only the first element will be used

## Warning: Unknown or uninitialised column: 'arguments'.  
  
## Warning: Unknown or uninitialised column: 'arguments'.

## Warning: Unknown or uninitialised column: 'imputations'.



# Question 10: What Type of rider uses the Ford goBikes more? Subscribers or Customers?

Based on the below results, we can find that subscriber group are more than customers group.

To find out no of subscribers and customer user types

Fordgobike\_outlier\_NA %>% filter(user\_type=="Subscriber") %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 862842

Fordgobike\_outlier\_NA %>% filter(user\_type=="Customer") %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 163260