Stat 650 sec 1 Midterm —-Ford Go Bike

\*\* Liu, Jiayi\*\* In this notebook I download and unzip the [Ford Go Bike](https://www.fordgobike.com/) [data](https://www.fordgobike.com/system-data).

library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.8  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.2.1 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)  
library(kableExtra)

##   
## Attaching package: 'kableExtra'

## The following object is masked from 'package:skimr':  
##   
## kable

Create a directory /data in your directory. Download the files. First one is not zipped, the remaining are zipped.

URL <- "https://s3.amazonaws.com/fordgobike-data/2017-fordgobike-tripdata.csv"  
download.file(URL, destfile = "./data/2017-fordgobike-tripdata.csv", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201801-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201801-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201802-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201802-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201803-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201803-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201804-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201804-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201805-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201805-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201806-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201806-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201807-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201807-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201808-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201808-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201809-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201809-fordgobike-tripdata.csv.zip", method="libcurl")  
URL <- "https://s3.amazonaws.com/fordgobike-data/201810-fordgobike-tripdata.csv.zip"  
download.file(URL, destfile = "./data/201810-fordgobike-tripdata.csv.zip", method="libcurl")

Loop over the one value in the url and filename that changes.

URL <- "https://s3.amazonaws.com/fordgobike-data/2017-fordgobike-tripdata.csv"  
download.file(URL, destfile = "./data/2017-fordgobike-tripdata.csv", method="curl")  
  
for (i in 1:10) {  
URL <- paste0("https://s3.amazonaws.com/fordgobike-data/20180",i,"-fordgobike-tripdata.csv.zip")  
download.file(URL, destfile = paste0("./data/20180",i,"-fordgobike-tripdata.csv.zip"), method="curl")  
}

Unzip downloaded files.

unzip("./data/201801-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201802-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201803-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201804-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201805-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201806-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201807-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201808-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201809-fordgobike-tripdata.csv.zip",exdir="./data")  
unzip("./data/201810-fordgobike-tripdata.csv.zip",exdir="./data")

Clean up data directory.

fn <- "./data/201801-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201802-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201803-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201804-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201805-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201806-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201807-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201808-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201809-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)  
fn <- "./data/201810-fordgobike-tripdata.csv.zip"  
if (file.exists(fn)) file.remove(fn)

Read the.csv files

fordgobike2017 <- read\_csv(file="./data/2017-fordgobike-tripdata.csv")  
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")  
fordgobike201809 <- read\_csv(file="./data/201809-fordgobike-tripdata.csv")  
fordgobike201810 <- read\_csv(file="./data/201810-fordgobike-tripdata.csv")

#glimpse(fordgobike201805)  
#glimpse(fordgobike201806)  
#glimpse(fordgobike201807)  
#glimpse(fordgobike201808)  
glimpse(fordgobike201808)

## Observations: 192,162  
## Variables: 16  
## $ duration\_sec <dbl> 69465, 60644, 57922, 82573, 57677, 769...  
## $ start\_time <dttm> 2018-08-31 19:18:08, 2018-08-31 21:40...  
## $ end\_time <dttm> 2018-09-01 14:35:54, 2018-09-01 14:30...  
## $ start\_station\_id <chr> "3", "50", "17", "19", "3", "19", "157...  
## $ start\_station\_name <chr> "Powell St BART Station (Market St at ...  
## $ start\_station\_latitude <dbl> 37.78638, 37.78053, 37.79225, 37.78898...  
## $ start\_station\_longitude <dbl> -122.4049, -122.3903, -122.3971, -122....  
## $ end\_station\_id <chr> "3", "50", "30", "34", "3", "285", "15...  
## $ end\_station\_name <chr> "Powell St BART Station (Market St at ...  
## $ end\_station\_latitude <dbl> 37.78638, 37.78053, 37.77660, 37.78399...  
## $ end\_station\_longitude <dbl> -122.4049, -122.3903, -122.3953, -122....  
## $ bike\_id <dbl> 813, 1964, 535, 3080, 132, 3056, 1123,...  
## $ user\_type <chr> "Customer", "Subscriber", "Customer", ...  
## $ member\_birth\_year <dbl> NA, 1990, 1997, 2000, NA, 2000, 1992, ...  
## $ member\_gender <chr> NA, "Male", "Male", "Female", NA, "Fem...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "N...

glimpse(fordgobike201809)

## Observations: 186,217  
## Variables: 16  
## $ duration\_sec <dbl> 76491, 50832, 74498, 79396, 60287, 707...  
## $ start\_time <dttm> 2018-09-30 18:37:01, 2018-09-30 19:49...  
## $ end\_time <dttm> 2018-10-01 15:51:53, 2018-10-01 09:56...  
## $ start\_station\_id <chr> "120", "NULL", "55", "11", "15", "188"...  
## $ start\_station\_name <chr> "Mission Dolores Park", "NULL", "Webst...  
## $ start\_station\_latitude <dbl> 37.76142, 37.41000, 37.77705, 37.79728...  
## $ start\_station\_longitude <dbl> -122.4264, -121.9400, -122.4296, -122....  
## $ end\_station\_id <chr> "26", "NULL", "55", "55", "15", "190",...  
## $ end\_station\_name <chr> "1st St at Folsom St", "NULL", "Webste...  
## $ end\_station\_latitude <dbl> 37.78729, 37.41000, 37.77705, 37.77705...  
## $ end\_station\_longitude <dbl> -122.3944, -121.9400, -122.4296, -122....  
## $ bike\_id <dbl> 903, 4154, 2696, 2458, 2497, 1483, 270...  
## $ user\_type <chr> "Customer", "Customer", "Customer", "C...  
## $ member\_birth\_year <dbl> 1998, 1992, NA, 1990, NA, 1985, 1996, ...  
## $ member\_gender <chr> "Female", "Male", NA, "Male", NA, "Mal...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "N...

glimpse(fordgobike201810)

## Observations: 201,458  
## Variables: 16  
## $ duration\_sec <dbl> 80066, 56428, 44825, 55461, 47724, 328...  
## $ start\_time <dttm> 2018-10-31 20:49:03, 2018-10-31 20:18...  
## $ end\_time <dttm> 2018-11-01 19:03:30, 2018-11-01 11:58...  
## $ start\_station\_id <chr> "138", "75", "139", "67", "259", "345"...  
## $ start\_station\_name <chr> "Jersey St at Church St", "Market St a...  
## $ start\_station\_latitude <dbl> 37.75090, 37.77379, 37.75102, 37.77664...  
## $ start\_station\_longitude <dbl> -122.4274, -122.4212, -122.4119, -122....  
## $ end\_station\_id <chr> "147", "86", "122", "16", "259", "109"...  
## $ end\_station\_name <chr> "29th St at Tiffany Ave", "Market St a...  
## $ end\_station\_latitude <dbl> 37.74407, 37.76931, 37.76030, 37.79413...  
## $ end\_station\_longitude <dbl> -122.4215, -122.4268, -122.4189, -122....  
## $ bike\_id <dbl> 3458, 1798, 1575, 2978, 2392, 1466, 44...  
## $ user\_type <chr> "Customer", "Customer", "Customer", "S...  
## $ member\_birth\_year <dbl> 1987, NA, 1983, 1992, NA, 1975, 1995, ...  
## $ member\_gender <chr> "Female", NA, "Female", "Female", NA, ...  
## $ bike\_share\_for\_all\_trip <chr> "No", "No", "No", "No", "No", "No", "N...

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

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

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

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

fordgobike201808 <- fordgobike201808 %>%  
 mutate(start\_station\_id = as.integer(start\_station\_id),  
 end\_station\_id= as.integer(start\_station\_id) )

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

fordgobike201809 <- fordgobike201809 %>%  
 mutate(start\_station\_id = as.integer(start\_station\_id),  
 end\_station\_id= as.integer(start\_station\_id) )

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

fordgobike201810 <- fordgobike201810 %>%  
 mutate(start\_station\_id = as.integer(start\_station\_id),  
 end\_station\_id= as.integer(start\_station\_id) )

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

fordgobike2018 <- bind\_rows(fordgobike201801, fordgobike201802, fordgobike201803, fordgobike201804,  
 fordgobike201805, fordgobike201806, fordgobike201807, fordgobike201808,  
 fordgobike201809, fordgobike201810)  
  
#glimpse(fordgobike2018)  
#View(fordgobike2018)

fordgobike2018 %>% select(start\_station\_id,start\_station\_name, start\_station\_latitude,start\_station\_longitude) %>%  
 arrange(start\_station\_id) %>%  
 distinct() %>%  
 head() %>%  
 kable(, format = "rst")

================ ============================================ ====================== ======================= start\_station\_id start\_station\_name start\_station\_latitude start\_station\_longitude ================ ============================================ ====================== ======================= 3 Powell St BART Station (Market St at 4th St) 37.78638 -122.4049 4 Cyril Magnin St at Ellis St 37.78588 -122.4089 5 Powell St BART Station (Market St at 5th St) 37.78390 -122.4084 6 The Embarcadero at Sansome St 37.80477 -122.4032 7 Frank H Ogawa Plaza 37.80456 -122.2717 8 The Embarcadero at Vallejo St 37.79995 -122.3985 ================ ============================================ ====================== =======================

## 1. Explain what the GBFS is?

General Bikeshare Feed Specification, known as GBFS, is the open data standard for bikeshare. GBFS will make real-time data feeds publicly available online in a uniform format so that map and transportation based apps can easily incorporate this data into their platforms. Also, the gbfs package supplies a set of functions to interface with General Bikeshare Feed Specification .json feeds in R, allowing users to save and accumulate tidy .rds datasets for specified cities/bikeshare programs. ##2. Explain any difficulties you encountered getting the code to work. Since I don’t want just delete unnecessary code, I tried to use some code into the chunk to hide some big result. I always failed but to add “#” to undo the code. ##3. The analysis is to work with the 2017 and 2018 data. How many bike rentals were there in 2017? 519700 How many bike rentals were there in 2018? 1210548 How many bike rentals have there been since the beginning of Ford GoBikes?1730248 ##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. line 175, 176 etc. From glimpse()function, we can find that it is int for start\_station\_id and end\_station\_id in 2018/05, while it is chr after 2018/06. When we try to combine the dataset together, we need unionform the type of data for each column.

dim(fordgobike2017)

## [1] 519700 15

fordgobike2017 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 519700

#nrow(fordgobike201801) + nrow(fordgobike201802) + nrow(fordgobike201803) + nrow(fordgobike201804)  
  
dim(fordgobike2018)

## [1] 1598223 16

fordgobike2018 %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 1598223

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

## [1] 2117923 16

fordgobike %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 2117923

dim(fordgobike)

## [1] 2117923 16

# add age, year, month, day and week\_day  
fordgobike <- fordgobike %>% mutate(age = 2018 - member\_birth\_year)  
  
fordgobike %>% count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 2117923

dim(fordgobike)

## [1] 2117923 17

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

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 2117923

dim(fordgobike)

## [1] 2117923 20

fordgobike <- fordgobike %>% mutate(week\_day = wday(start\_time) )  
  
  
level <- c("M","T","W","TH","F","SAT","SUN")  
  
fordgobike$week\_day <- factor(fordgobike$week\_day)

library(tidyverse)  
fordgobike$week\_day <- weekdays(as.Date(fordgobike$start\_time))  
level <- c("M","T","W","TH","F","SAT","SUN")  
fordgobike$week\_day <- factor(fordgobike$week\_day)  
dim(fordgobike)

## [1] 2117923 21

#View(fordgobike)  
fordgobike\_select<- fordgobike %>%  
 select(year,month,user\_type)  
write.csv(fordgobike\_select, "fordgobike\_select.csv")  
#View(fordgobike)

#library(lubridate)  
#library(data.table)  
#fordgobike <- fordgobike %>% mutate(week\_day = wday(as.POSIXlt(start\_time)))  
#fordgobike %>% count()   
#dim(fordgobike)  
#View(fordgobike)  
# week\_day seems not work very well. can't change level