Citibike Data Analysis

Data 621: Final Project - Group 4

Michael Yampol
Ann Liu-Ferrara
Sachid Deshmukh
Vishal Arora

12/14/2019

Contents

Abstract

- Short summary of the research problem and its importance, what you do and what you find
- A person reading your abstract should get a good sense of what problem you addressed and how you addressed it without having to look at the rest of the paper

Key words

Bikeshare, Weather, Cycling

Introduction

- What is the general area? What is the exact problem you are addressing?
- Why is it important? (why should I be interested as a reader?)
- What are the objectives of the research? What are your hypotheses?
- How is the paper structured?

Literature review

Looking at these papers/articles:

- SCHMIDT, Charles Active Travel for All? The Surge in Public Bike-Sharing Programs
- ZHOU, Xiaolu Understanding Spatiotemporal Patterns of Biking Behavior by Analyzing Massive Bike Sharing Data in Chicago.
- JIA, Yingnan, ... Effects of new dock-less bicycle-sharing programs on cycling: a retrospective study in Shanghai
- JIA, Yingnan, ... Association between innovative dockless bicycle sharing programs and adopting cycling in commuting and non-commuting trips
- HOSFORD,Kate ... Who is in the near market for bicycle sharing? Identifying current, potential, and unlikely users of a public bicycle share program in Vancouver, Canada
- HOSFORD,Kate ... Evaluation of the impact of a public bicycle share program on population bicycling in Vancouver, BC
- WESTLAND, James ... Demand cycles and market segmentation in bicycle sharing
- DELL'AMICO,... The bike sharing rebalancing problem: Mathematical formulations and benchmark instances

- DELL'AMICO,... The Bike sharing Rebalancing Problem with Stochastic Demands
- WANG, Shuai BRAVO: Improving the Rebalancing Operation in Bike Sharing with Rebalancing Range Prediction
- VOGEL, Patrick, ... "Strategic and Operational Planning of Bike-Sharing Systems by Data Mining – A Case Study"
- FULLER, Daniel,... Impact of a public transit strike on public bicycle share use: An interrupted time series natural experiment study
- FULLER, Daniel,... Impact evaluation of a public bicycle share program on cycling: a caseexample of BIXI in Montreal, Quebec
- FAGHIH-IMANI, Ahmadreza A finite mixture modeling approach to examine New York City bicycle sharing system (CitiBike) users' destination preferences
- AN, Ran, ... Weather and cycling in New York: The case of Citibike
- HEANEY, Alexandra, ... Climate Change and Physical Activity: Estimated Impacts of Ambient Temperatures on Bikeshare Usage in New York City

(I think the ones near the bottom of the list may be most promising...)

Methodology

- Define data collection method
- Accurate representation of the sample population and coverage issue from the target population
- Description of Data
- A complete description of the desired output
- Data Analysis
- Describe the instrumentation
- Describe the analysis plan
- Describe the scope and limitations of the methodology

- The data-set is currently composed of XXXXX records and VVV variables.
- Comment on missing values
- Comment on cleanup

in order to obtain the maximum information possible, we had to discard the use of many variables and put our focus into the following variables:

- `var1`
- `var2`
- `var3`
- `var4`

Experimentation and Results

The Results Section

- Needs to systematically and clearly articulate the study findings. If the results are unclear, the reviewer must decide whether the analysis of the data was poorly executed or whether the Results section is poorly organized.

From the above, we decided to ...

The Discussion Section

– Should state whether their hypotheses were verified or proven untrue or, if no hypotheses were given, whether their research questions were answered. The authors should also comment on their results in light of previous studies and explain what differences (if any) exist between their findings and those reported by others and attempt to provide an explanation for the discrepancies.

The Figures and Graphs

- Should illustrate the important features of the methods and results.
- Should allow the reader to understand the figure or graph without having to refer back to the text of the manuscript.
- Common mistakes made by inexperienced authors are failing to include figures that best depict their findings, writing unclear figure legends, and making poor use of arrows.

Tables

- Should summarize the data, make the data more easily understandable, and point out important comparisons.
- Description of the data in the text, if possible, is preferable to the use of a space-consuming table.

Conclusions and Summary

- Recap briefly what you do in the paper
- Evaluate the effectiveness of your research and provide recommendations (if applicable)
- Make sure that all of the questions raised in the introduction and the literature review have been addressed
- Compare the final results against the original aims and objectives
- Identify any shortcomings and future research

References

Appendix

```
knitr::opts_chunk$set(echo = TRUE, fig.pos = 'h')
mydir = "C:/Users/Michael/Dropbox/priv/CUNY/MSDS/201909-Fall/DATA621_Nasrin/20201214_FinalProject/"
setwd(mydir)
knitr::opts_knit$set(root.dir = mydir)
options(digits=7,scipen=999,width=120)
datadir = paste0(mydir,"/Data/")


### This contains all the data -- total 90.4M rows, 17GB size, 77 monthly files
rawdatadir = "C:/temp/CitibikeData/"
### This contains 1/1000 of the rows from each of the data files -- total 90.4K rows, 17MB size
slimdatadir = "C:/temp/CitibikeDataSlim/"
```

Load libraries

```
library(tidyverse)
```

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

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(sp)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(forcats)
```

1. Data Exploration

Load data

Weather data

```
# Weather data is obtained from the  NCDC (National Climatic Data Center) via https://www.ncdc.noaa.gov/cdo-web/
# click on search tool  https://www.ncdc.noaa.gov/cdo-web/search
# select "daily summaries"
# select Search for Stations
# Enter Search Term "USW00094728" for Central Park Station:
# https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail
# "add to cart"


weatherfilenames=list.files(path="./",pattern = '.csv$', full.names = T)    # ending with .csv ; not .zip
weatherfilenames
```

```
## [1] "./NYC_Weather_Data_2013-2019.csv" "./NYC_Weather_Data_2019.csv"
```

```
weatherfile <- "NYC_Weather_Data_2013-2019.csv"

## Perhaps we should rename the columns to more clearly reflect their meaning?
weatherspec <- cols(
  STATION = col_character(),
  NAME = col_character(),
  LATITUDE = col_double(),
  LONGITUDE = col_double(),
  ELEVATION = col_double(),
  DATE = col_date(format = "%F"),          #  readr::parse_datetime() :    "%F" = "%Y-%m-%d"
  AWND = col_double(),                     # Average Daily Wind Speed
  AWND_ATTRIBUTES = col_character(),
  PGTM = col_double(),                    # Peak Wind-Gust Time
  PGTM_ATTRIBUTES = col_character(),
  PRCP = col_double(),                    # Amount of Precipitation
  PRCP_ATTRIBUTES = col_character(),
  SNOW = col_double(),                    # Amount of Snowfall
  SNOW_ATTRIBUTES = col_character(),
```

```
  SNWD = col_double(),                    # Depth of snow on the ground
  SNWD_ATTRIBUTES = col_character(),
  TAVG = col_double(),                    # Average Temperature (not populated)
  TAVG_ATTRIBUTES = col_character(),
  TMAX = col_double(),                    # Maximum temperature for the day
  TMAX_ATTRIBUTES = col_character(),
  TMIN = col_double(),                    # Minimum temperature for the day
  TMIN_ATTRIBUTES = col_character(),
  TSUN = col_double(),                    # Daily Total Sunshine (not populated)
  TSUN_ATTRIBUTES = col_character(),
  WDF2 = col_double(),                    # Direction of fastest 2-minute wind
  WDF2_ATTRIBUTES = col_character(),
  WDF5 = col_double(),                    # Direction of fastest 5-second wind
  WDF5_ATTRIBUTES = col_character(),
  WSF2 = col_double(),                    # Fastest 2-minute wind speed
  WSF2_ATTRIBUTES = col_character(),
  WSF5 = col_double(),                    # fastest 5-second wind speed
  WSF5_ATTRIBUTES = col_character(),
  WT01 = col_double(),                    # Fog
  WT01_ATTRIBUTES = col_character(),
  WT02 = col_double(),                    # Heavy Fog
  WT02_ATTRIBUTES = col_character(),
  WT03 = col_double(),                    # Thunder
  WT03_ATTRIBUTES = col_character(),
  WT04 = col_double(),                    # Sleet
  WT04_ATTRIBUTES = col_character(),
  WT06 = col_double(),                    # Glaze
  WT06_ATTRIBUTES = col_character(),
  WT08 = col_double(),                    # Smoke or haze
  WT08_ATTRIBUTES = col_character(),
  WT13 = col_double(),                    # Mist
  WT13_ATTRIBUTES = col_character(),
  WT14 = col_double(),                    # Drizzle
  WT14_ATTRIBUTES = col_character(),
  WT16 = col_double(),                    # Rain
  WT16_ATTRIBUTES = col_character(),
  WT18 = col_double(),                    # Snow
  WT18_ATTRIBUTES = col_character(),
```

```
  WT19 = col_double(),                    # Unknown source of precipitation
  WT19_ATTRIBUTES = col_character(),
  WT22 = col_double(),                    # Ice fog
  WT22_ATTRIBUTES = col_character()
)



# load all the daily weather data
weather <- read_csv(weatherfile,col_types = weatherspec)

# extract just 2019
weather2019 <- weather[(weather$DATE>="2019-01-01" & weather$DATE<="2019-12-31"),]


# extract just one month
weather201906 <- weather[(weather$DATE>="2019-06-01" & weather$DATE<="2019-06-30"),]
```

Function to load up a citibike datafile

```
read_CB_data_file = function(f){
  Startloadtime = Sys.time()
  print(paste("reading data file   ", f, " at ", Startloadtime))

  ### Extract the year and month from the datafile.  Needed below for inconsistent date/time formats by month.
  YYYYMM <- sub("^.*/","",f) %>% sub("-citibike-tripdata.csv","",.)
  print(paste("YYYYMM = ", YYYYMM))


  ### Read the datafile according to the format specifications
  datafile = read_csv(f,skip = 1,
#The column names have slight format differences across months.  So, replace all column names with these:
                  col_names=c("trip_duration",            # in seconds
                              "s_time",                   # start date/time
                              "e_time",                   # end date/time
                              "s_station_id",             # station ID for beginning of trip
                              "s_station_name",
                              "s_lat",                    # start station latitude
```

```
                              "s_long",                    # start station longitude
                              "e_station_id",              # station ID for end of trip
                              "e_station_name",
                              "e_lat",                     # latitude
                              "e_long",                    # longitude
                              "bike_id",                   # every bike has a 5-digit ID number
                              "user_type",                 # Annual Subscriber or Daily Customer
                              "birth_year",                # Can infer age from this
                              "gender")                    # 1=Male,2=Female,0=unknown
#                    ,col_types = "dTTffddffddifif"    # d=decimal; T=datetime; f=factor; i=integer
### specify the data type for each of the above columns
### Note: because of changes in the format across months, we will  have to read the date/time as char for now
### also we will have to read the birth_year as char for now because of missing data (either "\\N" or "NULL")
                   ,col_types = "dccffddffddifcf"    # d=decimal; c=character; f=factor; i=integer
                   )
  Endloadtime = Sys.time()
  print(paste("done reading data file ",  f, " at ", Endloadtime))
  Totalloadtime = round(Endloadtime - Startloadtime, 2)
  print(paste("Totaltime = ", Totalloadtime))

## Fix format changes on time and birth_year variables
  s_time <- pull(.data=datafile, var = "s_time")
  e_time <- pull(.data=datafile, var = "e_time")

  ### Early and recent files use format "%Y-%m-%d %H:%M:%OS"
  if (YYYYMM < "201409" | YYYYMM > "201609") timeformat="%Y-%m-%d %H:%M:%OS"

  ### time between the months uses format "%m/%d/%Y %H:%M:%OS"
  if (YYYYMM >= "201409" & YYYYMM <= "201609") timeformat="%m/%d/%Y %H:%M:%OS"
  ### except for the first 3 months of 2015, time is only HH:MM -- no seconds!
  if (YYYYMM >= "201501" & YYYYMM <= "201503") timeformat="%m/%d/%Y %H:%M"
  ### Same for June 2015, time is only HH:MM -- no seconds!
  if (YYYYMM == "201506") timeformat="%m/%d/%Y %H:%M"

  datafile[,"s_time"] <- as.POSIXct(s_time, format=timeformat)
  datafile[,"e_time"] <- as.POSIXct(e_time, format=timeformat)

#### note:  on the first Sunday of November, clocks move back 1 hour.
```

```
#### This means that the hour 1am-2am EDT is followed by the hour 1am-2am EST.
#### If a bicycle was rented during this hour "EDT",
#### but returned during the subsequent hour "EST",
#### then the trip duration could appear negative.
#### This is because the default loader will assume all times on this date are EST.
#### In this case, the below will force such start-times back an hour:

iii = which(datafile$s_time>datafile$e_time)
if(length(iii)>0) {
  print("***DAYLIGHT SAVINGS PROBLEM***")
  print(datafile[iii,])
  print("**Start times:")
  print(pull(datafile[iii,2]))
  print(pull(datafile[iii,2]) %>% as.numeric())
  print(unclass(datafile[iii,2])$s_time)
  print("**End times:")
  print(pull(datafile[iii,3]))
  print(pull(datafile[iii,3]) %>% as.numeric())
  print(unclass(datafile[iii,3])$e_time)


  print("***CHANGING s_time backward***")
  new_s_time <- ifelse(datafile$s_time>datafile$e_time,
                       datafile$s_time-60*60,  # pushes back 1 hour from EST to EDT
                       datafile$s_time) %>% as.POSIXct(., origin= "1970-01-01")
  print("***CHANGING e_time forward***")
  new_e_time <- ifelse(datafile$s_time>datafile$e_time,
                       datafile$e_time+60*60,  # pushes forward 1 hour from EDT to EST
                       datafile$e_time) %>% as.POSIXct(., origin= "1970-01-01")
  before_diff <-  datafile[iii,3] - datafile[iii,2]
  print(paste("BEFORE difference: ", before_diff))

  datafile[,"s_time"] <- new_s_time
  datafile[,"e_time"] <- new_e_time

print("**AFTER CHANGE**")
  print(datafile[iii,])
  print("**Start times**")
```

```
  print(pull(datafile[iii,2]))
  print(pull(datafile[iii,2]) %>% as.numeric())
  print(unclass(datafile[iii,2])$s_time)
  print("**End times**")
  print(pull(datafile[iii,3]))
  print(pull(datafile[iii,3]) %>% as.numeric())
  print(unclass(datafile[iii,3])$e_time)

  after_diff <-  datafile[iii,3] - datafile[iii,2]
  print(paste("AFTER difference: ", after_diff))


  }

##
## set missing birth years to NA
  birth_year <- pull(.data=datafile, var = "birth_year")
## Fix missing birth year on early data (occurs when YYYYMM < "201409")
  birth_year[birth_year=='\\N']<-NA
## Fix missing birth year on 2017 (occurs when "201704" YYYYMM < "201712")
  birth_year[birth_year=='NULL']<-NA
## Convert the available birth_years to their integer equivalents (while retaining above NAs)
  datafile[,"birth_year"] <- as.integer(birth_year)

## There are numerous cases between 201610 and 201703 where the usertype is not specified.
## (It should be "Subscriber" or "Customer", but in such cases it is blank.)
## We will set it to "UNKNOWN"

#library(forcats)   # loaded above
  datafile$user_type<-fct_explicit_na(datafile$user_type, "UNKNOWN")

## There was a trial of DOCKLESS BIKES in the Bronx starting from August 2018:
## https://nyc.streetsblog.org/2018/08/16/a-hit-and-miss-debut-for-dockless-citi-bikes-in-the-bronx/
## https://d21xlh2maitm24.cloudfront.net/nyc/bronx-service-area-map.png?mtime=20180809110452
## https://webcache.googleusercontent.com/search?q=cache:9Xz02WSdeOYJ:https://www.citibikenyc.com/how-it-works/dockless-faqs+

## For these trips, the latitute and longitude of the bike start and stop is given, but
## the start and end station ID and station name are set to "NULL" in the input datafiles.
```

```
## For clarity, we will change such values to "DOCKLESS" :
  levels(datafile$s_station_id)[levels(datafile$s_station_id)=="NULL"] <- "DOCKLESS"
  levels(datafile$s_station_name)[levels(datafile$s_station_name)=="NULL"] <- "DOCKLESS"
  levels(datafile$e_station_id)[levels(datafile$e_station_id)=="NULL"] <- "DOCKLESS"
  levels(datafile$e_station_name)[levels(datafile$e_station_name)=="NULL"] <- "DOCKLESS"


## for certain months, the datafile is not sorted on s_time (instead it is sorted on s_station_id then s_time)
## ensure that this month's data is sorted on s_time
  datafile <- datafile[order(datafile$s_time),]
  print("----------------------------------------------------------------")
  return(datafile)
}


### November 1, 2015
#read_CB_data_file(filenames[29])

### November 6, 2016
#read_CB_data_file(filenames[41])
```

List the names of available citibike data files

```
filenames=list.files(path=slimdatadir,pattern = '.csv$', full.names = T)     # ending with .csv ; not .zip
length(filenames)
```

```
## [1] 77
```

```
t(t(filenames))
```

```
##       [,1]
##  [1,] "C:/temp/CitibikeDataSlim/201307-citibike-tripdata.csv"
##  [2,] "C:/temp/CitibikeDataSlim/201308-citibike-tripdata.csv"
##  [3,] "C:/temp/CitibikeDataSlim/201309-citibike-tripdata.csv"
##  [4,] "C:/temp/CitibikeDataSlim/201310-citibike-tripdata.csv"
##  [5,] "C:/temp/CitibikeDataSlim/201311-citibike-tripdata.csv"
```

```
##  [6,] "C:/temp/CitibikeDataSlim/201312-citibike-tripdata.csv"
##  [7,] "C:/temp/CitibikeDataSlim/201401-citibike-tripdata.csv"
##  [8,] "C:/temp/CitibikeDataSlim/201402-citibike-tripdata.csv"
##  [9,] "C:/temp/CitibikeDataSlim/201403-citibike-tripdata.csv"
## [10,] "C:/temp/CitibikeDataSlim/201404-citibike-tripdata.csv"
## [11,] "C:/temp/CitibikeDataSlim/201405-citibike-tripdata.csv"
## [12,] "C:/temp/CitibikeDataSlim/201406-citibike-tripdata.csv"
## [13,] "C:/temp/CitibikeDataSlim/201407-citibike-tripdata.csv"
## [14,] "C:/temp/CitibikeDataSlim/201408-citibike-tripdata.csv"
## [15,] "C:/temp/CitibikeDataSlim/201409-citibike-tripdata.csv"
## [16,] "C:/temp/CitibikeDataSlim/201410-citibike-tripdata.csv"
## [17,] "C:/temp/CitibikeDataSlim/201411-citibike-tripdata.csv"
## [18,] "C:/temp/CitibikeDataSlim/201412-citibike-tripdata.csv"
## [19,] "C:/temp/CitibikeDataSlim/201501-citibike-tripdata.csv"
## [20,] "C:/temp/CitibikeDataSlim/201502-citibike-tripdata.csv"
## [21,] "C:/temp/CitibikeDataSlim/201503-citibike-tripdata.csv"
## [22,] "C:/temp/CitibikeDataSlim/201504-citibike-tripdata.csv"
## [23,] "C:/temp/CitibikeDataSlim/201505-citibike-tripdata.csv"
## [24,] "C:/temp/CitibikeDataSlim/201506-citibike-tripdata.csv"
## [25,] "C:/temp/CitibikeDataSlim/201507-citibike-tripdata.csv"
## [26,] "C:/temp/CitibikeDataSlim/201508-citibike-tripdata.csv"
## [27,] "C:/temp/CitibikeDataSlim/201509-citibike-tripdata.csv"
## [28,] "C:/temp/CitibikeDataSlim/201510-citibike-tripdata.csv"
## [29,] "C:/temp/CitibikeDataSlim/201511-citibike-tripdata.csv"
## [30,] "C:/temp/CitibikeDataSlim/201512-citibike-tripdata.csv"
## [31,] "C:/temp/CitibikeDataSlim/201601-citibike-tripdata.csv"
## [32,] "C:/temp/CitibikeDataSlim/201602-citibike-tripdata.csv"
## [33,] "C:/temp/CitibikeDataSlim/201603-citibike-tripdata.csv"
## [34,] "C:/temp/CitibikeDataSlim/201604-citibike-tripdata.csv"
## [35,] "C:/temp/CitibikeDataSlim/201605-citibike-tripdata.csv"
## [36,] "C:/temp/CitibikeDataSlim/201606-citibike-tripdata.csv"
## [37,] "C:/temp/CitibikeDataSlim/201607-citibike-tripdata.csv"
## [38,] "C:/temp/CitibikeDataSlim/201608-citibike-tripdata.csv"
## [39,] "C:/temp/CitibikeDataSlim/201609-citibike-tripdata.csv"
## [40,] "C:/temp/CitibikeDataSlim/201610-citibike-tripdata.csv"
## [41,] "C:/temp/CitibikeDataSlim/201611-citibike-tripdata.csv"
## [42,] "C:/temp/CitibikeDataSlim/201612-citibike-tripdata.csv"
## [43,] "C:/temp/CitibikeDataSlim/201701-citibike-tripdata.csv"
## [44,] "C:/temp/CitibikeDataSlim/201702-citibike-tripdata.csv"
```

```
## [45,] "C:/temp/CitibikeDataSlim/201703-citibike-tripdata.csv"
## [46,] "C:/temp/CitibikeDataSlim/201704-citibike-tripdata.csv"
## [47,] "C:/temp/CitibikeDataSlim/201705-citibike-tripdata.csv"
## [48,] "C:/temp/CitibikeDataSlim/201706-citibike-tripdata.csv"
## [49,] "C:/temp/CitibikeDataSlim/201707-citibike-tripdata.csv"
## [50,] "C:/temp/CitibikeDataSlim/201708-citibike-tripdata.csv"
## [51,] "C:/temp/CitibikeDataSlim/201709-citibike-tripdata.csv"
## [52,] "C:/temp/CitibikeDataSlim/201710-citibike-tripdata.csv"
## [53,] "C:/temp/CitibikeDataSlim/201711-citibike-tripdata.csv"
## [54,] "C:/temp/CitibikeDataSlim/201712-citibike-tripdata.csv"
## [55,] "C:/temp/CitibikeDataSlim/201801-citibike-tripdata.csv"
## [56,] "C:/temp/CitibikeDataSlim/201802-citibike-tripdata.csv"
## [57,] "C:/temp/CitibikeDataSlim/201803-citibike-tripdata.csv"
## [58,] "C:/temp/CitibikeDataSlim/201804-citibike-tripdata.csv"
## [59,] "C:/temp/CitibikeDataSlim/201805-citibike-tripdata.csv"
## [60,] "C:/temp/CitibikeDataSlim/201806-citibike-tripdata.csv"
## [61,] "C:/temp/CitibikeDataSlim/201807-citibike-tripdata.csv"
## [62,] "C:/temp/CitibikeDataSlim/201808-citibike-tripdata.csv"
## [63,] "C:/temp/CitibikeDataSlim/201809-citibike-tripdata.csv"
## [64,] "C:/temp/CitibikeDataSlim/201810-citibike-tripdata.csv"
## [65,] "C:/temp/CitibikeDataSlim/201811-citibike-tripdata.csv"
## [66,] "C:/temp/CitibikeDataSlim/201812-citibike-tripdata.csv"
## [67,] "C:/temp/CitibikeDataSlim/201901-citibike-tripdata.csv"
## [68,] "C:/temp/CitibikeDataSlim/201902-citibike-tripdata.csv"
## [69,] "C:/temp/CitibikeDataSlim/201903-citibike-tripdata.csv"
## [70,] "C:/temp/CitibikeDataSlim/201904-citibike-tripdata.csv"
## [71,] "C:/temp/CitibikeDataSlim/201905-citibike-tripdata.csv"
## [72,] "C:/temp/CitibikeDataSlim/201906-citibike-tripdata.csv"
## [73,] "C:/temp/CitibikeDataSlim/201907-citibike-tripdata.csv"
## [74,] "C:/temp/CitibikeDataSlim/201908-citibike-tripdata.csv"
## [75,] "C:/temp/CitibikeDataSlim/201909-citibike-tripdata.csv"
## [76,] "C:/temp/CitibikeDataSlim/201910-citibike-tripdata.csv"
## [77,] "C:/temp/CitibikeDataSlim/201911-citibike-tripdata.csv"
```

Load up data file or files

```
#### Load all the data files, noting how much time it takes to load them
```

```r
Starttime = Sys.time()
print(paste("Start time: ", Starttime))
```

```
## [1] "Start time:  2019-12-24 00:04:11"
```

```r
### loads up the data for all files -- problem is too much data for my computer to handle if all are loaded
print("About to load multiple datafiles:")
```

```
## [1] "About to load multiple datafiles:"
```

```r
#suppress listing
#print(filenames)


# call read_CB_data_files to load the files specified
CB <- do.call(rbind,lapply(filenames,read_CB_data_file))
```

```
## [1] "reading data file    C:/temp/CitibikeDataSlim/201307-citibike-tripdata.csv  at  2019-12-24 00:04:11"
## [1] "YYYYMM =  201307"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201307-citibike-tripdata.csv  at  2019-12-24 00:04:11"
## [1] "Totaltime =  0.06"
## [1] "-----------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201308-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201308"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201308-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.04"
## [1] "-----------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201309-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201309"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201309-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.04"
## [1] "-----------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201310-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201310"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201310-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.07"
## [1] "-----------------------------------------------------------------"
```

```
## [1] "reading data file    C:/temp/CitibikeDataSlim/201311-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201311"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201311-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.03"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201312-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201312"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201312-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.02"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201401-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201401"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201401-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.02"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201402-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201402"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201402-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.02"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201403-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201403"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201403-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.04"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201404-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201404"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201404-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "Totaltime =  0.03"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201405-citibike-tripdata.csv  at  2019-12-24 00:04:12"
## [1] "YYYYMM =  201405"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201405-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.06"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201406-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201406"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201406-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.04"
```

```
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201407-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201407"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201407-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201408-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201408"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201408-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.06"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201409-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201409"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201409-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201410-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201410"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201410-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201411-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201411"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201411-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.02"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201412-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201412"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201412-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.01"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201501-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201501"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201501-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "Totaltime =  0.01"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201502-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201502"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201502-citibike-tripdata.csv  at  2019-12-24 00:04:13"
```

```
## [1] "Totaltime =  0.01"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201503-citibike-tripdata.csv  at  2019-12-24 00:04:13"
## [1] "YYYYMM =  201503"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201503-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.02"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201504-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201504"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201504-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201505-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201505"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201505-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201506-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201506"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201506-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201507-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201507"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201507-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201508-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201508"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201508-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201509-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201509"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201509-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201510-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201510"
```

```
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201510-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201511-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "YYYYMM =  201511"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201511-citibike-tripdata.csv  at  2019-12-24 00:04:14"
## [1] "Totaltime =  0.04"
## [1] "***DAYLIGHT SAVINGS PROBLEM***"
## # A tibble: 1 x 15
##   trip_duration s_time              e_time              s_station_id s_station_name s_lat s_long e_station_id
##           <dbl> <dttm>              <dttm>              <fct>        <fct>          <dbl>  <dbl> <fct>
## 1           935 2015-11-01 01:55:34 2015-11-01 01:11:10 463          9 Ave & W 16 ~  40.7  -74.0 253
## # ... with 7 more variables: e_station_name <fct>, e_lat <dbl>, e_long <dbl>, bike_id <int>, user_type <fct>,
## #   birth_year <chr>, gender <fct>
## [1] "**Start times:"
## [1] "2015-11-01 01:55:34 EST"
## [1] 1446360934
## [1] "2015-11-01 01:55:34 EST"
## [1] "**End times:"
## [1] "2015-11-01 01:11:10 EDT"
## [1] 1446354670
## [1] "2015-11-01 01:11:10 EDT"
## [1] "***CHANGING s_time backward***"
## [1] "***CHANGING e_time forward***"
## [1] "BEFORE difference:  -1.74"
## [1] "**AFTER CHANGE**"
## # A tibble: 1 x 15
##   trip_duration s_time              e_time              s_station_id s_station_name s_lat s_long e_station_id
##           <dbl> <dttm>              <dttm>              <fct>        <fct>          <dbl>  <dbl> <fct>
## 1           935 2015-11-01 01:55:34 2015-11-01 01:11:10 463          9 Ave & W 16 ~  40.7  -74.0 253
## # ... with 7 more variables: e_station_name <fct>, e_lat <dbl>, e_long <dbl>, bike_id <int>, user_type <fct>,
## #   birth_year <chr>, gender <fct>
## [1] "**Start times**"
## [1] "2015-11-01 01:55:34 EDT"
## [1] 1446357334
## [1] "2015-11-01 01:55:34 EDT"
## [1] "**End times**"
## [1] "2015-11-01 01:11:10 EST"
## [1] 1446358270
```

```
## [1] "2015-11-01 01:11:10 EST"
## [1] "AFTER difference:  15.6"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201512-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201512"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201512-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201601-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201601"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201601-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201602-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201602"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201602-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201603-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201603"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201603-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201604-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201604"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201604-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201605-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201605"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201605-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201606-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201606"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201606-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201607-citibike-tripdata.csv  at  2019-12-24 00:04:15"
```

```
## [1] "YYYYMM =  201607"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201607-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "Totaltime =  0.04"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201608-citibike-tripdata.csv  at  2019-12-24 00:04:15"
## [1] "YYYYMM =  201608"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201608-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.05"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201609-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201609"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201609-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.06"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201610-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201610"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201610-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.08"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201611-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201611"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201611-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.04"
## [1] "***DAYLIGHT SAVINGS PROBLEM***"
## # A tibble: 1 x 15
##   trip_duration s_time              e_time              s_station_id s_station_name s_lat s_long e_station_id
##           <dbl> <dttm>              <dttm>              <fct>        <fct>          <dbl>  <dbl> <fct>
## 1           748 2016-11-06 01:53:03 2016-11-06 01:05:32 461          E 20 St & 2 A~  40.7  -74.0 312
## # ... with 7 more variables: e_station_name <fct>, e_lat <dbl>, e_long <dbl>, bike_id <int>, user_type <fct>,
## #   birth_year <chr>, gender <fct>
## [1] "**Start times:"
## [1] "2016-11-06 01:53:03 EST"
## [1] 1478415183
## [1] "2016-11-06 01:53:03 EST"
## [1] "**End times:"
## [1] "2016-11-06 01:05:32 EDT"
## [1] 1478408732
## [1] "2016-11-06 01:05:32 EDT"
## [1] "***CHANGING s_time backward***"
```

```
## [1] "***CHANGING e_time forward***"
## [1] "BEFORE difference:  -1.79194444444444"
## [1] "**AFTER CHANGE**"
## # A tibble: 1 x 15
##   trip_duration s_time              e_time              s_station_id s_station_name s_lat s_long e_station_id
##           <dbl> <dttm>              <dttm>              <fct>        <fct>          <dbl>  <dbl> <fct>
## 1           748 2016-11-06 01:53:03 2016-11-06 01:05:32 461          E 20 St & 2 A~  40.7  -74.0 312
## # ... with 7 more variables: e_station_name <fct>, e_lat <dbl>, e_long <dbl>, bike_id <int>, user_type <fct>,
## #   birth_year <chr>, gender <fct>
## [1] "**Start times**"
## [1] "2016-11-06 01:53:03 EDT"
## [1] 1478411583
## [1] "2016-11-06 01:53:03 EDT"
## [1] "**End times**"
## [1] "2016-11-06 01:05:32 EST"
## [1] 1478412332
## [1] "2016-11-06 01:05:32 EST"
## [1] "AFTER difference:  12.4833333333333"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201612-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201612"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201612-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201701-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201701"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201701-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.02"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201702-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201702"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201702-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201703-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201703"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201703-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "Totaltime =  0.03"
## [1] "------------------------------------------------------------"
```

```
## [1] "reading data file    C:/temp/CitibikeDataSlim/201704-citibike-tripdata.csv  at  2019-12-24 00:04:16"
## [1] "YYYYMM =  201704"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201704-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.05"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201705-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201705"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201705-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.06"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201706-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201706"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201706-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.06"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201707-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201707"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201707-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.06"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201708-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201708"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201708-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.07"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201709-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201709"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201709-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.08"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201710-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "YYYYMM =  201710"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201710-citibike-tripdata.csv  at  2019-12-24 00:04:17"
## [1] "Totaltime =  0.07"
## [1] "-------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201711-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201711"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201711-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.06"
```

```
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201712-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201712"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201712-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201801-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201801"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201801-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201802-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201802"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201802-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201803-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201803"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201803-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201804-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201804"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201804-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.06"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201805-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201805"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201805-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.09"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201806-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "YYYYMM =  201806"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201806-citibike-tripdata.csv  at  2019-12-24 00:04:18"
## [1] "Totaltime =  0.09"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201807-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201807"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201807-citibike-tripdata.csv  at  2019-12-24 00:04:19"
```

```
## [1] "Totaltime =  0.09"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201808-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201808"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201808-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "Totaltime =  0.1"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201809-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201809"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201809-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "Totaltime =  0.09"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201810-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201810"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201810-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "Totaltime =  0.08"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201811-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201811"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201811-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "Totaltime =  0.06"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201812-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "YYYYMM =  201812"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201812-citibike-tripdata.csv  at  2019-12-24 00:04:19"
## [1] "Totaltime =  0.04"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201901-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201901"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201901-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201902-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201902"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201902-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.05"
## [1] "------------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201903-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201903"
```

```
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201903-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.06"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201904-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201904"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201904-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.08"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201905-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201905"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201905-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.09"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201906-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201906"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201906-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "Totaltime =  0.1"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201907-citibike-tripdata.csv  at  2019-12-24 00:04:20"
## [1] "YYYYMM =  201907"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201907-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "Totaltime =  0.11"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201908-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "YYYYMM =  201908"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201908-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "Totaltime =  0.1"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201909-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "YYYYMM =  201909"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201909-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "Totaltime =  0.09"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201910-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "YYYYMM =  201910"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201910-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "Totaltime =  0.09"
## [1] "-----------------------------------------------------------"
## [1] "reading data file    C:/temp/CitibikeDataSlim/201911-citibike-tripdata.csv  at  2019-12-24 00:04:21"
```

```
## [1] "YYYYMM =  201911"
## [1] "done reading data file  C:/temp/CitibikeDataSlim/201911-citibike-tripdata.csv  at  2019-12-24 00:04:21"
## [1] "Totaltime =  0.07"
## [1] "------------------------------------------------------------"
```

```
### Problem:  loading up multiple files is too much for my computer to handle !!!!



### load up just a single month of data ("filename")
#### Look at just June 2019
##filename = filenames[6]
##print(paste("Loading data for ", filename))
##CB <- read_CB_data_file(filename)

Endtime = Sys.time()
print(paste("End time: ", Endtime))
```

```
## [1] "End time:  2019-12-24 00:04:22"
```

```
Totaltime = round(Endtime - Starttime,2)
print(paste("Totaltime for loading above file(s) = ", Totaltime))
```

```
## [1] "Totaltime for loading above file(s) =  11.46"
```

```
## Save a copy of the loaded data, in case we need it during manipulations below
save_CB <- CB
```

glimpse the dataset

```
glimpse(CB)
```

```
## Observations: 90,370
## Variables: 15
## $ trip_duration  <dbl> 634, 437, 1398, 1124, 1199, 221, 861, 403, 965, 384, 940, 577, 783, 526, 381, 405, 275, 1036...
## $ s_time         <dttm> 2013-07-01 00:00:00, 2013-07-01 06:54:02, 2013-07-01 08:03:38, 2013-07-01 08:37:40, 2013-07...
```

```
## $ e_time         <dttm> 2013-07-01 00:10:34, 2013-07-01 07:01:19, 2013-07-01 08:26:56, 2013-07-01 08:56:24, 2013-07...
## $ s_station_id   <fct> 164, 479, 157, 496, 432, 475, 458, 466, 2004, 507, 153, 414, 507, 412, 485, 147, 161, 334, 3...
## $ s_station_name <fct> E 47 St & 2 Ave, 9 Ave & W 45 St, Henry St & Atlantic Ave, E 16 St & 5 Ave, E 7 St & Avenue ...
## $ s_lat          <dbl> 40.75323, 40.76019, 40.69089, 40.73726, 40.72622, 40.73524, 40.75140, 40.74395, 40.72440, 40...
## $ s_long         <dbl> -73.97033, -73.99126, -73.99612, -73.99239, -73.98380, -73.98759, -74.00523, -73.99145, -74....
## $ e_station_id   <fct> 504, 243, 375, 500, 466, 537, 465, 212, 116, 379, 488, 310, 482, 368, 408, 534, 347, 477, 49...
## $ e_station_name <fct> 1 Ave & E 15 St, Fulton St & Rockwell Pl, Mercer St & Bleecker St, Broadway & W 51 St, W 25 ...
## $ e_lat          <dbl> 40.73222, 40.68798, 40.72679, 40.76229, 40.74395, 40.74026, 40.75514, 40.74335, 40.74178, 40...
## $ e_long         <dbl> -73.98166, -73.97847, -73.99695, -73.98336, -73.99145, -73.98409, -73.98658, -74.00682, -74....
## $ bike_id        <int> 16950, 16151, 15997, 17750, 17671, 16490, 16067, 16025, 19485, 15641, 19184, 19895, 17044, 1...
## $ user_type      <fct> Customer, Subscriber, Subscriber, Subscriber, Subscriber, Subscriber, Subscriber, Subscriber...
## $ birth_year     <int> NA, 1987, 1987, 1959, 1983, 1956, 1982, 1970, 1954, 1989, 1988, NA, 1974, 1961, 1987, 1975, ...
## $ gender         <fct> 0, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 0, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1,...
```

str - structure of the dataset

```
str(CB)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    90370 obs. of  15 variables:
##  $ trip_duration : num  634 437 1398 1124 1199 ...
##  $ s_time        : POSIXct, format: "2013-07-01 00:00:00" "2013-07-01 06:54:02" "2013-07-01 08:03:38" "2013-07-01 08:37:40" ...
##  $ e_time        : POSIXct, format: "2013-07-01 00:10:34" "2013-07-01 07:01:19" "2013-07-01 08:26:56" "2013-07-01 08:56:24" ...
##  $ s_station_id  : Factor w/ 945 levels "164","479","157",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ s_station_name: Factor w/ 962 levels "E 47 St & 2 Ave",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ s_lat         : num  40.8 40.8 40.7 40.7 40.7 ...
##  $ s_long        : num  -74 -74 -74 -74 -74 ...
##  $ e_station_id  : Factor w/ 947 levels "504","243","375",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ e_station_name: Factor w/ 966 levels "1 Ave & E 15 St",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ e_lat         : num  40.7 40.7 40.7 40.8 40.7 ...
##  $ e_long        : num  -74 -74 -74 -74 -74 ...
##  $ bike_id       : int  16950 16151 15997 17750 17671 16490 16067 16025 19485 15641 ...
##  $ user_type     : Factor w/ 3 levels "Customer","Subscriber",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ birth_year    : int  NA 1987 1987 1959 1983 1956 1982 1970 1954 1989 ...
##  $ gender        : Factor w/ 3 levels "0","1","2": 1 2 2 3 3 2 3 3 2 3 ...
```

Summary of the dataset

```r
summary(CB)
```

```
##  trip_duration           s_time                         e_time                       s_station_id
##  Min.   :     61.0   Min.   :2013-07-01 00:00:00   Min.   :2013-07-01 00:10:34   519    :  867
##  1st Qu.:    375.0   1st Qu.:2015-12-12 13:18:44   1st Qu.:2015-12-12 13:32:47   497    :  651
##  Median :    621.0   Median :2017-07-27 22:56:59   Median :2017-07-27 23:14:00   435    :  583
##  Mean   :    911.5   Mean   :2017-04-25 17:56:54   Mean   :2017-04-25 18:12:06   402    :  571
##  3rd Qu.:   1064.0   3rd Qu.:2018-10-18 18:30:21   3rd Qu.:2018-10-18 18:40:07   426    :  569
##  Max.   :1688083.0   Max.   :2019-11-30 23:59:28   Max.   :2019-12-01 00:06:58   285    :  560
##                                                                                  (Other):86569
##                s_station_name      s_lat           s_long        e_station_id                e_station_name
##  Pershing Square North:  766   Min.   :40.65   Min.   :-74.03   519    :  799   Pershing Square North:  713
##  E 17 St & Broadway   :  651   1st Qu.:40.72   1st Qu.:-74.00   497    :  701   E 17 St & Broadway   :  701
##  W 21 St & 6 Ave      :  583   Median :40.74   Median :-73.99   435    :  653   W 21 St & 6 Ave      :  653
##  Broadway & E 22 St   :  571   Mean   :40.74   Mean   :-73.99   426    :  643   West St & Chambers St:  643
##  West St & Chambers St:  569   3rd Qu.:40.75   3rd Qu.:-73.98   402    :  596   Broadway & E 22 St   :  596
##  8 Ave & W 31 St      :  563   Max.   :40.86   Max.   :-73.89   293    :  579   Lafayette St & E 8 St:  579
##  (Other)              :86667                                    (Other):86399   (Other)              :86485
##      e_lat           e_long          bike_id           user_type       birth_year   gender
##  Min.   :40.65   Min.   :-74.05   Min.   :14529   Customer  :10634   Min.   :1885   0: 8989
##  1st Qu.:40.72   1st Qu.:-74.00   1st Qu.:18065   Subscriber:79687   1st Qu.:1969   1:61095
##  Median :40.74   Median :-73.99   Median :22059   UNKNOWN   :   49   Median :1981   2:20286
##  Mean   :40.74   Mean   :-73.99   Mean   :23798                      Mean   :1978
##  3rd Qu.:40.75   3rd Qu.:-73.98   3rd Qu.:28852                      3rd Qu.:1988
##  Max.   :40.86   Max.   :-73.89   Max.   :42044                      Max.   :2003
##                                                                      NA's   :5881
```

2. Data Preparation

Check for missing values

Let's check whether any variables have missing values, i.e., values which are NULL or NA.

```
## [1] "Number of columns with missing values =  1"
```

```
## [1] "Names of columns with missing values =  birth_year"
```

We know that there are missing birth_year values.

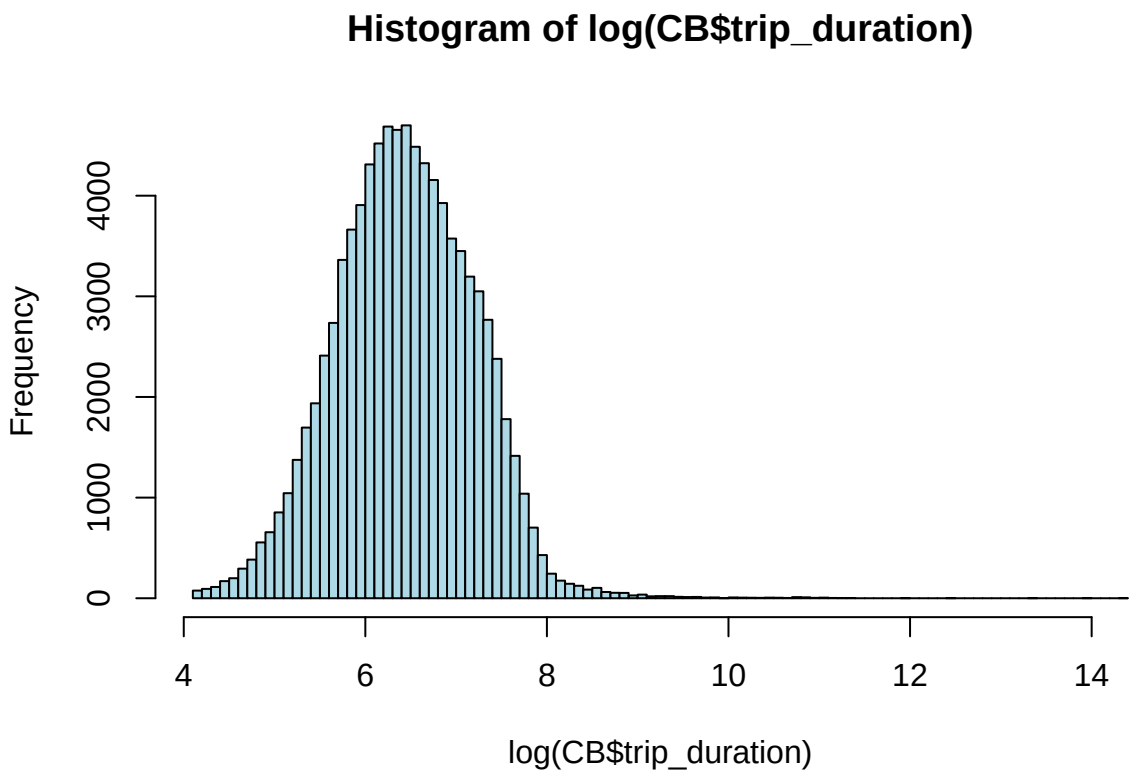
Look for unusual values / outliers

Examine trip_duration

The trip_duration is specified in seconds, but there seem to be some outliers which may be incorrect, as the value for Max is quite high: 1688083 seconds, or 19.5379977 days. We can assume that this data is bad, as nobody would willingly rent a bicycle for this period of time, given the fees that would be charged.

Histogram of log(trip_duration)

```
hist(log(CB$trip_duration),col='lightblue', breaks=100)
```



It may be easier to think of trip duration in other units (i.e., minutes, hours, or days) rather than in seconds, so lets create such variables. Also, let's confirm that the value shown (in seconds) is consistent with the difference between the start time and the end time:

Summary of trip durations before censoring/truncation:

```
#express trip duration in seconds, minutes, hours, days
# note: we needed to fix the November daylight savings problem to eliminate negative trip times

#### Seconds
CB$trip_duration_s = as.numeric(CB$e_time - CB$s_time,"secs")
print("Seconds:")
```

```
## [1] "Seconds:"
```

```
summary(CB$trip_duration_s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      60     375     621     912    1065 1688083
```

```
#### Minutes
CB$trip_duration_m = as.numeric(CB$e_time - CB$s_time,"mins")
print("Minutes")
```

```
## [1] "Minutes"
```

```
summary(CB$trip_duration_m)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.00     6.25    10.35    15.20    17.75 28134.72
```

```
#### Hours
CB$trip_duration_h = as.numeric(CB$e_time - CB$s_time,"hours")
print("Hours")
```

```
## [1] "Hours"
```

```
summary(CB$trip_duration_h)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0167   0.1042   0.1725   0.2533   0.2958 468.9119
```

```
#### Days
CB$trip_duration_d = as.numeric(CB$e_time - CB$s_time,"days")
summary(CB$trip_duration_d)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  0.000694  0.004340  0.007188  0.010556  0.012326 19.537998
```

```
print("Days")
```

```
## [1] "Days"
```

```
sum(CB$trip_duration_h>6)
```

```
## [1] 79
```

Let's assume that nobody would rent a bicycle for more than a specified timelimit, and drop the records which exceed this:

```
total_rows=dim(CB)[1]
print(paste("Initial number of trips: ", total_rows))
```

```
## [1] "Initial number of trips:  90370"
```

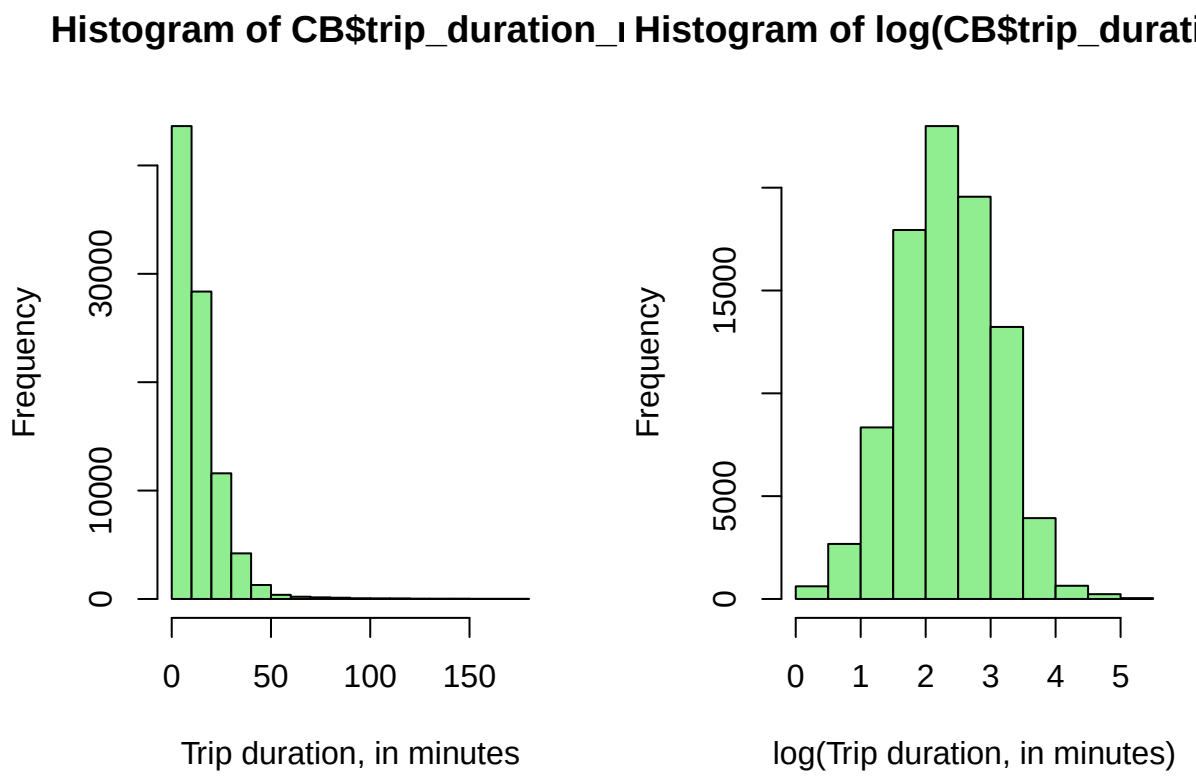
```
# choose only trips that were at most 3 hrs, as longer trips may reflect an error
# remove long trips from the data set -- something may be wrong (e.g., the system failed to properly record the return of a bike)
longtripthreshold_s = 60 * 60 *3  # 10800 seconds = 180 minutes = 2 hours
longtripthreshold_m = longtripthreshold_s / 60
longtripthreshold_h = longtripthreshold_m / 60

long_trips <- CB %>% filter(trip_duration_s > longtripthreshold_s)
num_long_trips_removed = dim(long_trips)[1]
pct_long_trips_removed = round(100*num_long_trips_removed / total_rows, 3)

CB <- CB %>% filter(trip_duration <= longtripthreshold_s)
reduced_rows = dim(CB)[1]

print(paste0("Removed ", num_long_trips_removed, " trips (", pct_long_trips_removed, "%) of longer than ", longtripthreshold_h, " hours.")
```

```
## [1] "Removed 157 trips (0.174%) of longer than 3 hours."
```

```r
print(paste0("Remaining number of trips: ", reduced_rows))
```

```
## [1] "Remaining number of trips: 90213"
```

```r
par(mfrow=c(1,2))
hist(CB$trip_duration_m, col="lightgreen",  xlab="Trip duration, in minutes")
hist(log(CB$trip_duration_m), col="lightgreen",  xlab="log(Trip duration, in minutes)")
```



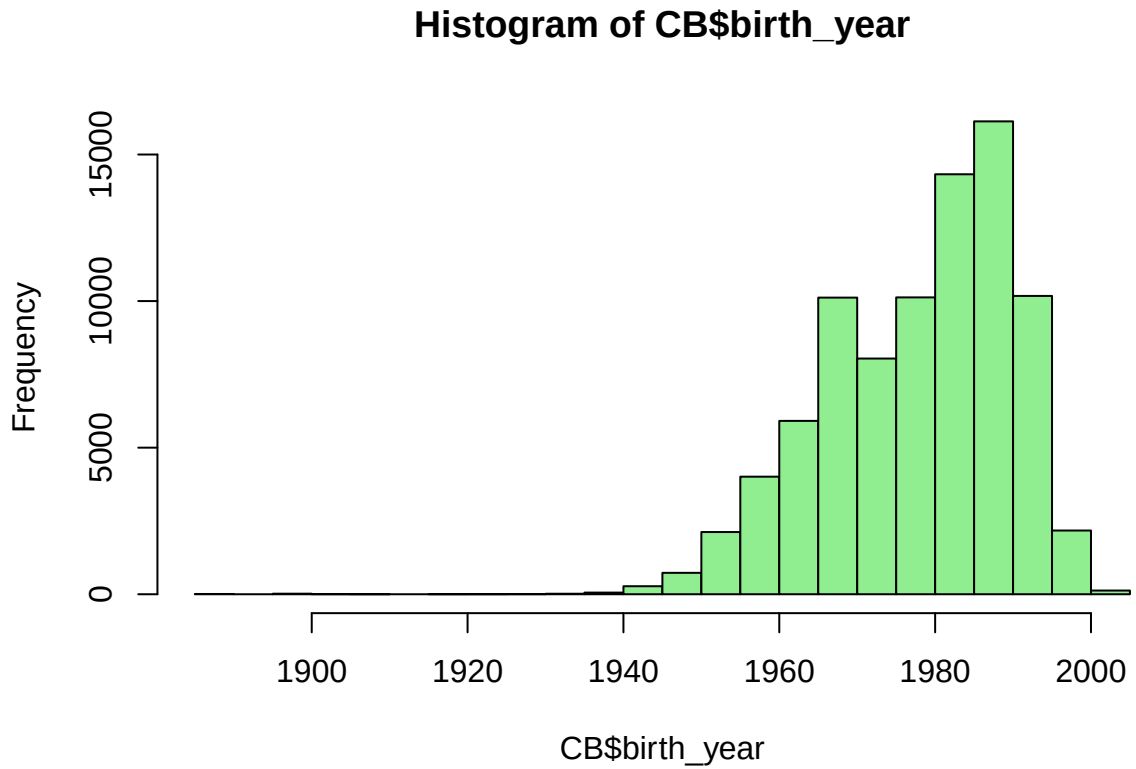
Examine birth year

The birth year for some users is as old as 1885, which is not possible.

```
summary(CB$birth_year)
```

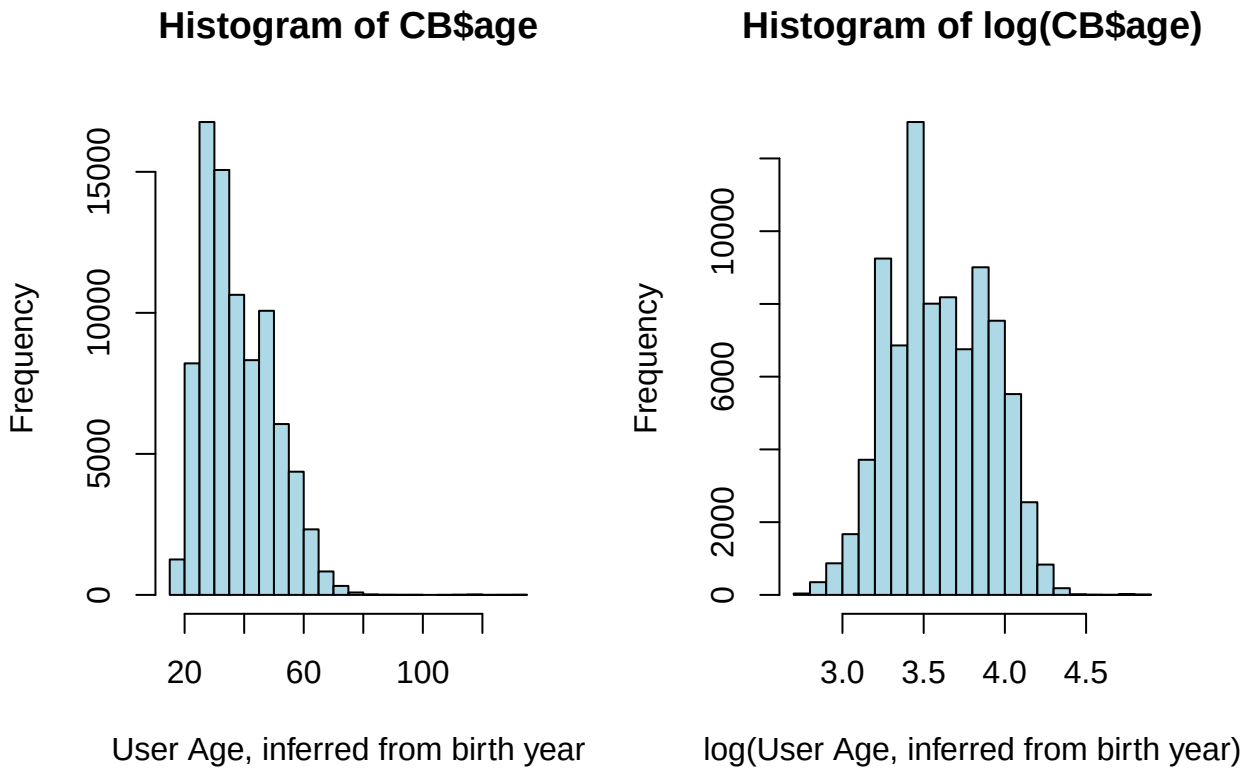
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1885    1969    1981    1978    1988    2003    5828
```

```
hist(CB$birth_year, col="lightgreen")
```



Remove trips associated with very old users

```
# Deduce age from trip date and birth year
#library(lubridate) #loaded above
CB$age <- year(CB$s_time) - CB$birth_year

par(mfrow=c(1,2))
hist(CB$age, col="lightblue",  xlab="User Age, inferred from birth year")
hist(log(CB$age), col="lightblue",  xlab="log(User Age, inferred from birth year)")
```



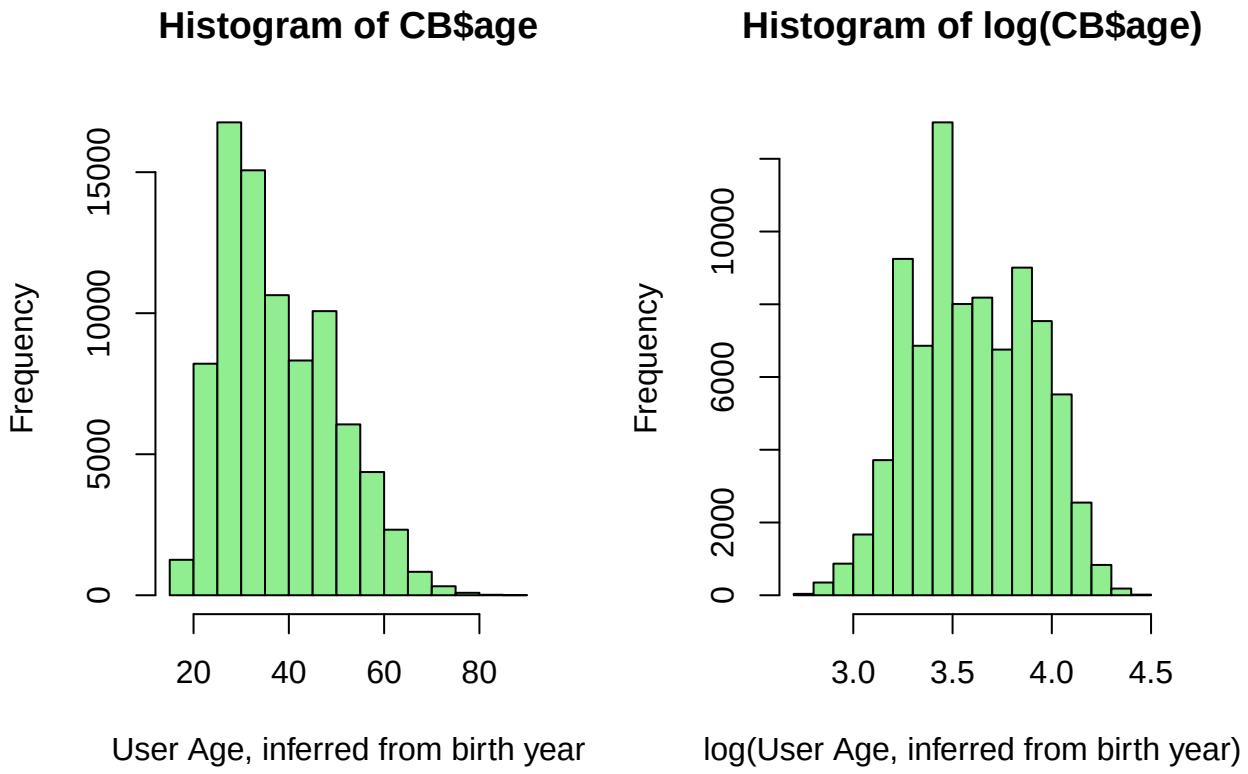
```r
# choose only trips where the user was born after a certain year,  as older users may reflect an error
age_threshhold = 90


aged_trips <- CB %>% filter(age > age_threshhold)
num_aged_trips_removed = dim(aged_trips)[1]
pct_aged_trips_removed = round(100*num_aged_trips_removed / total_rows, 3)

CB <- CB %>% filter(age <= age_threshhold)
reduced_rows = dim(CB)[1]

print(paste0("Removed ", num_aged_trips_removed, " trips (", pct_aged_trips_removed, "%) of users older than ", age_threshhold, " years.")
```

```
## [1] "Removed 40 trips (0.044%) of users older than 90 years."
```

```r
print(paste0("Remaining number of trips: ", reduced_rows))
```

```
## [1] "Remaining number of trips: 84345"
```

```r
par(mfrow=c(1,2))
hist(CB$age, col="lightgreen",  xlab="User Age, inferred from birth year")
hist(log(CB$age), col="lightgreen",  xlab="log(User Age, inferred from birth year)")
```



Compute distance between start and end stations

This is straight-line distance – it doesn't incorporate an actual route.
There are services (e.g., from Google) which can compute and measure a recommended bicycle route between points, but use of such services requires a subscription and incurs a cost.

```
# Compute the distance between start and end stations
s_lat_long <- CB %>% select(c(s_lat,s_long)) %>%  as.matrix
e_lat_long <- CB %>% select(c(e_lat,e_long)) %>%  as.matrix
#library(sp) # loaded above
CB$distance_km <- spDists(s_lat_long, e_lat_long, longlat=T, diagonal = TRUE)
```

```r
# There is a time-based usage fee for rides longer than an initial period.
# For user_type=Subscriber, the fee is $2.50 per 15 minutes following an initial free 45 minutes.
# For user_type=Customer, the fee is $4.00 per 15 minutes following an initial free 30 minutes.
# There are some cases where the user type is not specified (we have relabeled as "UNKNOWN")


CB$trip_fee[CB$user_type=="Subscriber"] <- 2.50 * (ceiling(
  CB$trip_duration_m[CB$user_type=="Subscriber"]  / 15)-3)  # first 45 minutes are free
```

```
## Warning: Unknown or uninitialised column: 'trip_fee'.
```

```r
CB$trip_fee[CB$user_type=="Customer"]   <- 4.00 * (ceiling(
  CB$trip_duration_m[CB$user_type=="Customer"]  / 15)-2)  # first 30 minutes are free
CB$trip_fee[CB$trip_fee<0] <- 0   # fee is non-negative
```

Summary of trip durations AFTER censoring/truncation:

```r
#express trip duration in seconds, minutes, hours, days
# note: we needed to fix the November daylight savings problem to eliminate negative trip times

#### Seconds
CB$trip_duration_s = as.numeric(CB$e_time - CB$s_time,"secs")
print("Seconds:")
```

```
## [1] "Seconds:"
```

```r
summary(CB$trip_duration_s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   364.0   594.5   778.2   997.8 10617.5
```

```r
#### Minutes
CB$trip_duration_m = as.numeric(CB$e_time - CB$s_time,"mins")
print("Minutes")
```

```
## [1] "Minutes"
```

```
summary(CB$trip_duration_m)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.067   9.908  12.970  16.630 176.959
```

```
#### Hours
CB$trip_duration_h = as.numeric(CB$e_time - CB$s_time,"hours")
print("Hours")
```

```
## [1] "Hours"
```

```
summary(CB$trip_duration_h)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01667 0.10111 0.16514 0.21616 0.27716 2.94931
```

```
#### Days
CB$trip_duration_d = as.numeric(CB$e_time - CB$s_time,"days")
summary(CB$trip_duration_d)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0006944 0.0042130 0.0068807 0.0090068 0.0115485 0.1228879
```

```
print("Days")
```

```
## [1] "Days"
```

```
sum(CB$trip_duration_h>6)
```

```
## [1] 0
```

Make a smaller dataset, numeric, without multicollinearities, for correlation calculations

```r
# extract selected fields
CBlite  <- select(CB, c(trip_duration, trip_fee, distance_km,
                        s_station_id, s_lat, s_long,
                        e_station_id, e_lat, e_long,
                        user_type, gender, age))

#make numeric variables
CBlite$user_type <- as.integer(CBlite$user_type)
CBlite$gender <- as.integer(CBlite$gender)

# function to revert factor back to its numeric levels
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}

CBlite$s_station_id <- as.numeric.factor(CBlite$s_station_id)
```

```
## Warning in as.numeric.factor(CBlite$s_station_id): NAs introduced by coercion
```

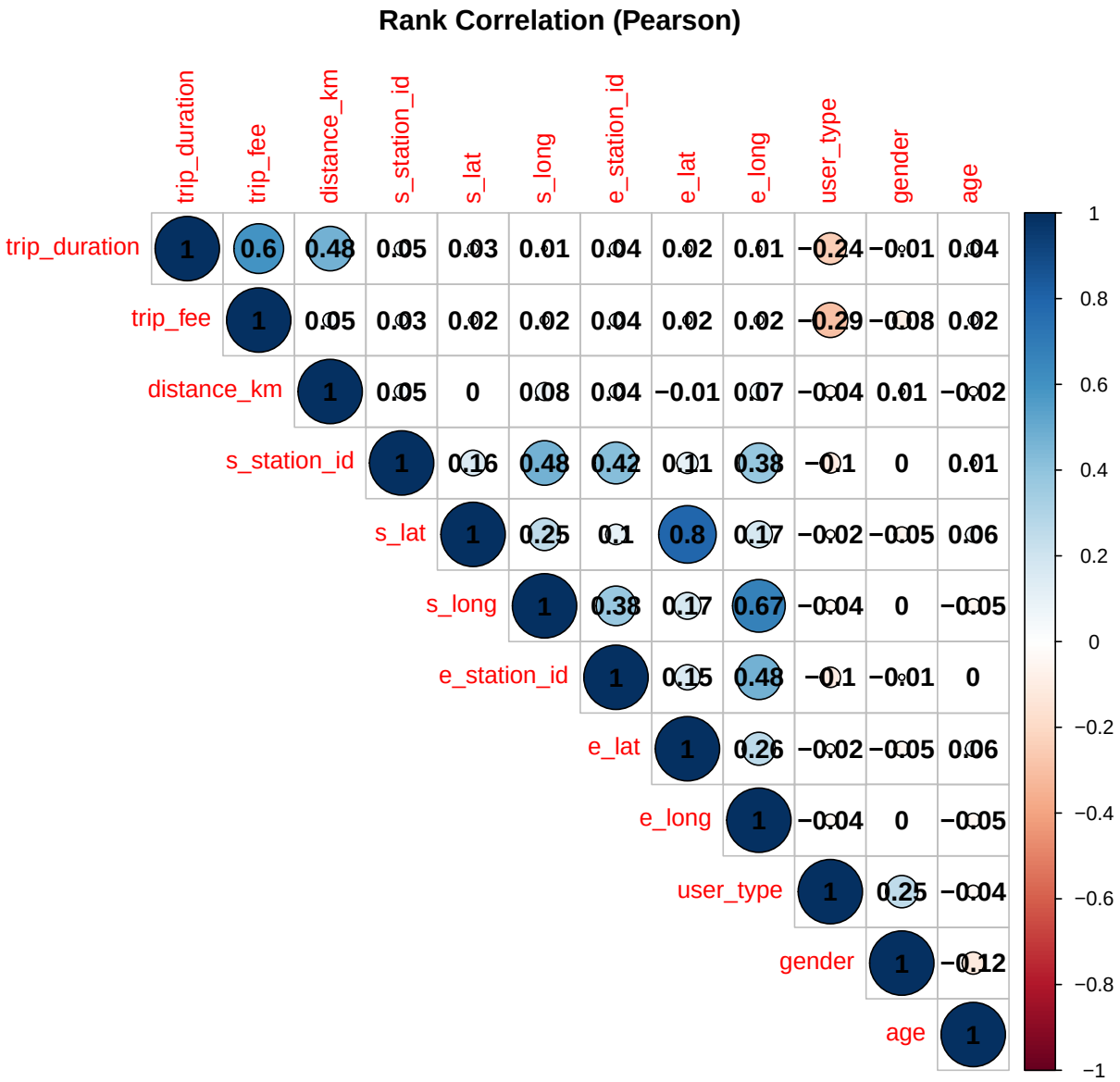
```r
CBlite$e_station_id <- as.numeric.factor(CBlite$e_station_id)
```

```
## Warning in as.numeric.factor(CBlite$e_station_id): NAs introduced by coercion
```

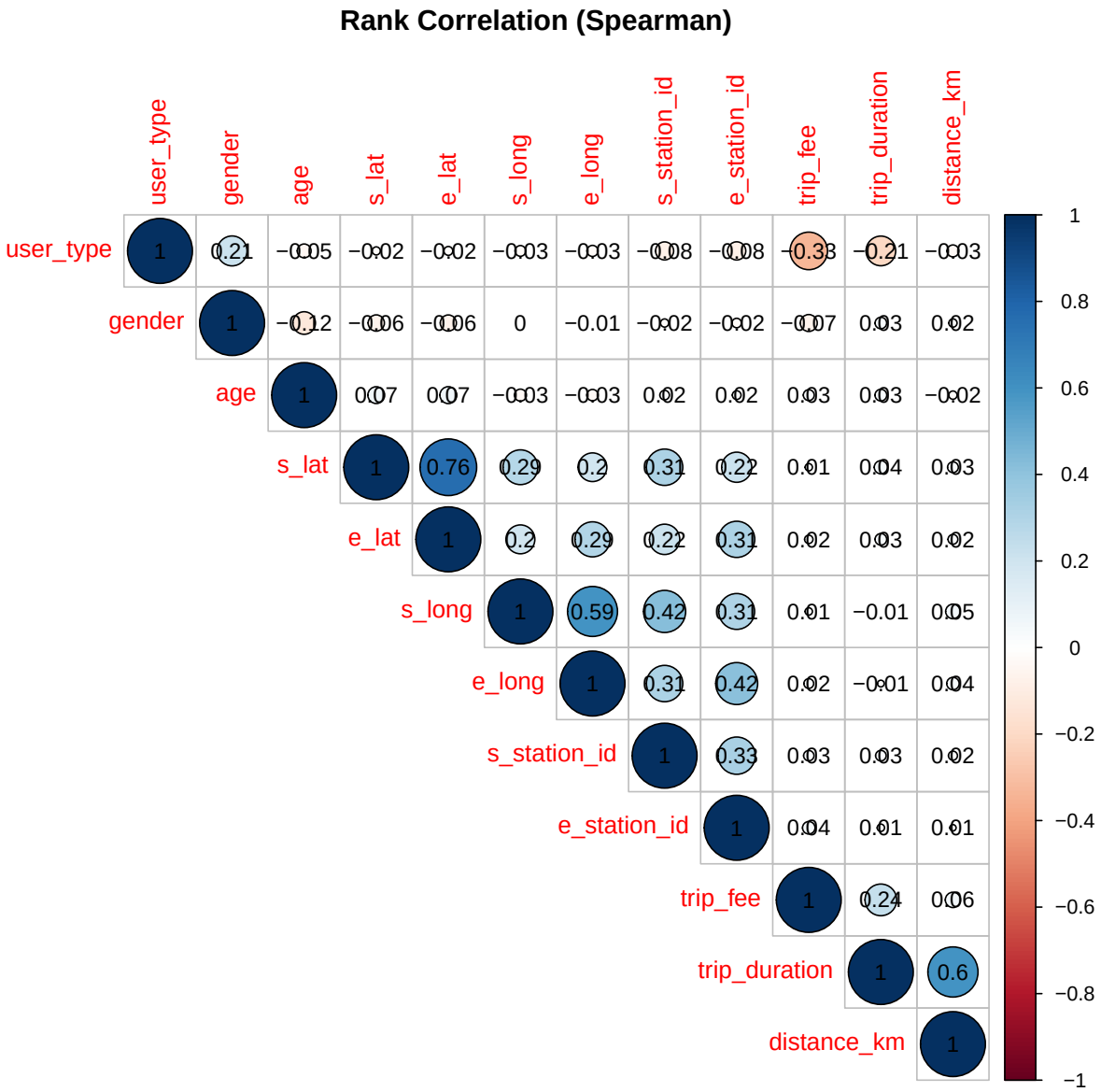
compute correlations

```r
#library(Hmisc) #loaded above
res2<-rcorr(as.matrix(CBlite))
respearson=rcorr(as.matrix(CBlite),type = "pearson")
resspearman=rcorr(as.matrix(CBlite),type = "spearman")
res3 <- cor(as.matrix(CBlite))
```

Pearson rank correlation


Rank Correlation (Pearson)
trip_duration
trip_fee
distance_km
s_station_id
s_lat
s_long
e_station_id
e_lat
e_long
user_type
gender
age
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Spearman rank correlation


Rank Correlation (Spearman)
user_type
gender
age
s_lat
e_lat
s_long
e_long
s_station_id
e_station_id
trip_fee
trip_duration
distance_km
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

3. Build Models

4. Select Models