Data cleaning process

Tanatip Pairod

2021-05-24

Contents

Setup

```
library(tidyverse)

### import data with multiple csv files into 1 data frame and export for next time usage
library(data.table)
files <- list.files(pattern = ".csv")
temp <- lapply(files, fread, sep = ",")
data <- rbindlist(temp)
cyclist_2020 <- data
```

```
## check data type for each column
str(data)
```

```
## Classes 'data.table' and 'data.frame':   3541683 obs. of  21 variables:
##  $ V1                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA7
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct, format: "2020-04-26 17:45:00" "2020-04-17 17:08:00" ...
##  $ start_date        : IDate, format: "2020-04-26" "2020-04-17" ...
##  $ start_time        : chr  "17:45:14" "17:08:54" "17:54:13" "12:50:19" ...
##  $ ended_at          : POSIXct, format: "2020-04-26 18:12:00" "2020-04-17 17:17:00" ...
##  $ ended_date        : IDate, format: "2020-04-26" "2020-04-17" ...
##  $ ended_time        : chr  "18:12:03" "17:17:03" "18:08:36" "13:02:31" ...
##  $ ride_length       : chr  "0:26:49" "0:08:09" "0:14:23" "0:12:12" ...
##  $ ride_length_min   : num  26.82 8.15 14.38 12.2 52.92 ...
##  $ day_of_week       : chr  "Sun" "Fri" "Wed" "Tue" ...
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "Calif
##  $ start_station_id  : chr  "86" "503" "142" "216" ...
##  $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt
##  $ end_station_id    : chr  "152" "499" "255" "657" ...
```

```
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
cyclist_2020 <- data
```

```
## change cyclist_2020 to data.frame
cyclist_2020 <- as.data.frame(cyclist_2020)
```

```
# drop start_date, started_time, ended_date, ended_time, and ride_length column
drop <- c("V1","start_date","start_time","ended_date", "ended_time", "ride_length")
cyclist_2020 <- cyclist_2020[ , !(names(cyclist_2020) %in% drop)]

## check data type for each column again
str(cyclist_2020)
```

```
## 'data.frame':    3541683 obs. of  15 variables:
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA7
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct, format: "2020-04-26 17:45:00" "2020-04-17 17:08:00" ...
##  $ ended_at          : POSIXct, format: "2020-04-26 18:12:00" "2020-04-17 17:17:00" ...
##  $ ride_length_min   : num  26.82 8.15 14.38 12.2 52.92 ...
##  $ day_of_week       : chr  "Sun" "Fri" "Wed" "Tue" ...
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "Calif
##  $ start_station_id  : chr  "86" "503" "142" "216" ...
##  $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt
##  $ end_station_id    : chr  "152" "499" "255" "657" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

Cleaning Data

Step 1: Quick checks

- Check class of data (must be data frame)
- Check the number of rows and columns the data frame has
- Quick check for summary statistic

```
class(cyclist_2020) # "data.frame"
```

```
## [1] "data.frame"
```

```
dim(cyclist_2020) # 3541683 15
```

```
## [1] 3541683      15
```

```
summary(cyclist_2020)
```

```
##    ride_id          rideable_type        started_at
##  Length:3541683     Length:3541683     Min.   :2020-01-01 00:04:44
##  Class :character   Class :character   1st Qu.:2020-06-18 19:10:40
##  Mode  :character   Mode  :character   Median :2020-08-09 11:42:17
##                                        Mean   :2020-07-28 22:19:56
##                                        3rd Qu.:2020-09-25 10:54:24
##                                        Max.   :2020-12-31 23:59:59
##
##     ended_at                    ride_length_min    day_of_week
##  Min.   :2020-01-01 00:10:54   Min.   :-1439.03   Length:3541683
##  1st Qu.:2020-06-18 19:41:00   1st Qu.:    7.48   Class :character
##  Median :2020-08-09 12:14:22   Median :   13.85   Mode  :character
##  Mean   :2020-07-28 22:44:46   Mean   :   12.93
##  3rd Qu.:2020-09-25 11:18:17   3rd Qu.:   25.52
##  Max.   :2021-01-03 08:54:11   Max.   : 1376.52
##                                NA's   :1119
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:3541683     Length:3541683     Length:3541683     Length:3541683
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.64   Min.   :-87.87   Min.   :41.54   Min.   :-87.89
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.64
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.08   Max.   :-87.52   Max.   :42.16   Max.   :-87.44
##                                   NA's   :4255    NA's   :4255
##  member_casual
##  Length:3541683
##  Class :character
##  Mode  :character
##
##
##
##
```

Step 2: Correcting the error!

```r
# change all the text to lowercase in a particular column
cyclist_2020$rideable_type <- tolower(cyclist_2020$rideable_type)
cyclist_2020$start_station_name <- tolower(cyclist_2020$start_station_name)
cyclist_2020$end_station_name <- tolower(cyclist_2020$end_station_name)

# trimming whitespaces
library(stringr)
cyclist_2020$rideable_type <- str_trim(cyclist_2020$rideable_type)
cyclist_2020$start_station_name <- str_trim(cyclist_2020$start_station_name)
cyclist_2020$end_station_name <- str_trim(cyclist_2020$end_station_name)
```

```r
# checking for missing values in the entire data frame
any(is.na(cyclist_2020))
```

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

```r
# checking for the total number of missing values in a particular column
sum(is.na(cyclist_2020$ride_id))            # 0
sum(is.na(cyclist_2020$rideable_type))      # 0
sum(is.na(cyclist_2020$started_at))         # 0
sum(is.na(cyclist_2020$ended_at))           # 0
sum(is.na(cyclist_2020$ride_length_mins))   # 0
sum(is.na(cyclist_2020$day_of_week))        # 0
sum(is.na(cyclist_2020$start_station_name))# 0
sum(is.na(cyclist_2020$start_station_id))   # 83583
sum(is.na(cyclist_2020$end_station_name))   # 0
sum(is.na(cyclist_2020$end_station_id))     # 98105
sum(is.na(cyclist_2020$start_lat))          # 0
sum(is.na(cyclist_2020$start_lng))          # 0
sum(is.na(cyclist_2020$end_lat))            # 4255
sum(is.na(cyclist_2020$end_lng))            # 4255
sum(is.na(cyclist_2020$member_casual))      # 0
## Total of missing values: 190198
```

```r
# check for duplicate row
sum(duplicated(cyclist_2020))
```

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

```r
# eliminating missing values completely from the entire dataframe
cyclist_2020 <- na.omit(cyclist_2020)               # 135869 rows affected
```

```r
# fix and remove row with white space in selected column
cyclist_2020 <- cyclist_2020[!(is.na(cyclist_2020$start_station_id)|
                               cyclist_2020$start_station_id==""), ] # 11,699 rows affected
cyclist_2020 <- cyclist_2020[!(is.na(cyclist_2020$end_station_id)|
                               cyclist_2020$end_station_id==""), ] # 5,852 rows affected
cyclist_2020 <- cyclist_2020[!(is.na(cyclist_2020$end_lat)|
                               cyclist_2020$end_lat==""), ] # 0 row affected
cyclist_2020 <- cyclist_2020[!(is.na(cyclist_2020$end_lng)|
                               cyclist_2020$end_lng==""), ] # 0 row affected
```

```r
# checking for number of the white space value for particular column
sum(cyclist_2020$ride_id=="")            # 0
sum(cyclist_2020$rideable_type=="")      # 0
sum(cyclist_2020$ride_length_mins=="")   # 0
sum(cyclist_2020$day_of_week=="")        # 0
sum(cyclist_2020$start_station_name=="")# 0
sum(cyclist_2020$start_station_id=="")   # 0
sum(cyclist_2020$end_station_name=="")   # 0
sum(cyclist_2020$end_station_id=="")     # 0
sum(cyclist_2020$start_lat=="")          # 0
sum(cyclist_2020$start_lng=="")          # 0
sum(cyclist_2020$end_lat=="")            # 0
sum(cyclist_2020$end_lng=="")            # 0
sum(cyclist_2020$member_casual=="")      # 0
```

```r
# 2nd checking for the total number of missing values in a particular column
missing <- cyclist_2020 %>%
  summarize_all(funs(mean(is.na(.))))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```r
glimpse(missing[1:15])
```

```
## Rows: 1
## Columns: 15
## $ ride_id            <dbl> 0
## $ rideable_type      <dbl> 0
## $ started_at         <dbl> 0
## $ ended_at           <dbl> 0
## $ ride_length_min    <dbl> 0
## $ day_of_week        <dbl> 0
## $ start_station_name <dbl> 0
## $ start_station_id   <dbl> 0
## $ end_station_name   <dbl> 0
## $ end_station_id     <dbl> 0
## $ start_lat          <dbl> 0
## $ start_lng          <dbl> 0
## $ end_lat            <dbl> 0
## $ end_lng            <dbl> 0
## $ member_casual      <dbl> 0
```

```
## remove rows that negative values in ride_length_min column
cyclist_2020 <- cyclist_2020 %>%
  filter(ride_length_min>0) # 35,736 rows affected
```

```
summary(cyclist_2020)
```

```
##    ride_id          rideable_type        started_at
##  Length:3352527     Length:3352527     Min.   :2020-01-01 00:04:44
##  Class :character   Class :character   1st Qu.:2020-06-15 12:51:53
##  Mode  :character   Mode  :character   Median :2020-08-06 11:09:49
##                                        Mean   :2020-07-25 04:27:16
##                                        3rd Qu.:2020-09-20 18:32:17
##                                        Max.   :2020-12-31 23:54:39
##     ended_at                   ride_length_min   day_of_week
##  Min.   :2020-01-01 00:10:54   Min.   :   0.02   Length:3352527
##  1st Qu.:2020-06-15 13:18:45   1st Qu.:   7.73   Class :character
##  Median :2020-08-06 11:33:24   Median :  14.12   Mode  :character
##  Mean   :2020-07-25 04:49:35   Mean   :  22.36
##  3rd Qu.:2020-09-20 18:58:34   3rd Qu.:  25.82
##  Max.   :2021-01-02 22:03:22   Max.   :1376.52
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:3352527     Length:3352527     Length:3352527     Length:3352527
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.65   Min.   :-87.77   Min.   :41.65   Min.   :-87.77
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.64
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.06   Max.   :-87.53   Max.   :42.07   Max.   :-87.53
##  member_casual
##  Length:3352527
##  Class :character
##  Mode  :character
##
##
##
```

Total removed data: 189,156 rows

Step 3: Export

```
## export data for making data visualization in Tableau
write_csv(cyclist_2020, "clean_cyclist_2020.csv")
```