Homework 1

Jun Ryu, UID: 605574052

2023-04-07

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

Question 1

```
taxi <- read.csv("chicagotaxiraw.csv") # read in the data

# locate the column numbers of the columns that we want to extract (1:6, 11:17)
clean_taxi <- as_tibble(taxi[,c(1:6, 11:17)])
clean_taxi
```

```
## # A tibble: 100,000 x 13
##    Trip.ID      Taxi.ID Trip.~1 Trip.~2 Trip.~3 Trip.~4 Fare  Tips  Tolls Extras
##    <chr>        <chr>   <chr>   <chr>     <int>   <dbl> <chr> <chr> <chr> <chr>
##  1 2f946572a1f~ 626cdd~ 2014-1~ 2014-1~     240     0   $4.65 $0.00 $0.00 $0.00
##  2 340c309a437~ da1882~ 2014-0~ 2014-0~     480     2.2 $7.65 $1.50 $0.00 $0.00
##  3 3a1a1f33626~ db337a~ 2016-0~ 2016-0~     480     2.3 $9.50 $0.00 $0.00 $1.00
##  4 39018c6704e~ ef4143~ 2014-1~ 2014-1~     780     2.8 $10.~ $2.00 $0.00 $1.00
##  5 40d518fdeed~ dcb626~ 2014-1~ 2014-1~    1140     5.7 $15.~ $0.00 $0.00 $0.00
##  6 3825c467c1b~ af3b4b~ 2013-0~ 2013-0~    1380    18.2 $36.~ $0.00 $0.00 $3.00
##  7 3df87a82d46~ 386ace~ 2015-1~ 2015-1~    1200     8.6 $20.~ $0.00 $0.00 $0.00
##  8 3270aa9ee4a~ 94cb43~ 2014-0~ 2014-0~     720     5.7 $14.~ $0.00 $0.00 $0.00
##  9 2a61add9211~ 589abe~ 2014-1~ 2014-1~     420     1.5 $6.65 $0.00 $0.00 $1.00
## 10 29d3a51c542~ 429edc~ 2015-1~ 2015-1~    1020     6   $15.~ $3.05 $0.00 $0.00
## # ... with 99,990 more rows, 3 more variables: Trip.Total <chr>,
## #   Payment.Type <chr>, Company <chr>, and abbreviated variable names
## #   1: Trip.Start.Timestamp, 2: Trip.End.Timestamp, 3: Trip.Seconds,
## #   4: Trip.Miles
```

```
# now, we use tolower() for lowercasing
names(clean_taxi) <- tolower(names(clean_taxi))
```

```
clean_taxi
```

```
## # A tibble: 100,000 x 13
##    trip.id      taxi.id trip.~1 trip.~2 trip.~3 trip.~4 fare  tips  tolls extras
##    <chr>        <chr>   <chr>   <chr>      <int>   <dbl> <chr> <chr> <chr> <chr>
##  1 2f946572a1f~ 626cdd~ 2014-1~ 2014-1~     240     0   $4.65 $0.00 $0.00 $0.00
##  2 340c309a437~ da1882~ 2014-0~ 2014-0~     480     2.2 $7.65 $1.50 $0.00 $0.00
##  3 3a1a1f33626~ db337a~ 2016-0~ 2016-0~     480     2.3 $9.50 $0.00 $0.00 $1.00
##  4 39018c6704e~ ef4143~ 2014-1~ 2014-1~     780     2.8 $10.~ $2.00 $0.00 $1.00
##  5 40d518fdeed~ dcb626~ 2014-1~ 2014-1~    1140     5.7 $15.~ $0.00 $0.00 $0.00
##  6 3825c467c1b~ af3b4b~ 2013-0~ 2013-0~    1380    18.2 $36.~ $0.00 $0.00 $3.00
##  7 3df87a82d46~ 386ace~ 2015-1~ 2015-1~    1200     8.6 $20.~ $0.00 $0.00 $0.00
##  8 3270aa9ee4a~ 94cb43~ 2014-0~ 2014-0~     720     5.7 $14.~ $0.00 $0.00 $0.00
##  9 2a61add9211~ 589abe~ 2014-1~ 2014-1~     420     1.5 $6.65 $0.00 $0.00 $1.00
## 10 29d3a51c542~ 429edc~ 2015-1~ 2015-1~    1020     6   $15.~ $3.05 $0.00 $0.00
## # ... with 99,990 more rows, 3 more variables: trip.total <chr>,
## #   payment.type <chr>, company <chr>, and abbreviated variable names
## #   1: trip.start.timestamp, 2: trip.end.timestamp, 3: trip.seconds,
## #   4: trip.miles
```

Question 2

```
# use as.Date to extract the dates
range(as.Date(clean_taxi$trip.start.timestamp))
```

```
## [1] "2013-01-01" "2017-05-31"
```

```
range(as.Date(clean_taxi$trip.end.timestamp), na.rm = T)
```

```
## [1] "2013-01-01" "2017-05-31"
```

As we can see by the results above, the date of the first pickup is 2013/1/1 and the date of the last dropoff is 2017/5/31 in this dataset.

Question 3

```
# use table() to get a summary table based on the day of the week
table(weekdays(as.Date(clean_taxi$trip.start.timestamp)))
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##     16767     12406     15859     13003     14873     13196     13896
```

Based on the above table, we observe that Friday had the greatest number of trips begin with 16767 trips and Monday had the least with 12406 trips in this dataset.

Question 4

```
# use str_replace() to drop the dollar sign and as.numeric() for numeric conversion
clean_taxi$tips <- as.numeric(str_replace(clean_taxi$tips, "[$]", ""))

range(clean_taxi$tips)
```

```
## [1]  0 97
```

```
mean(clean_taxi$tips)
```

```
## [1] 1.154516
```

```
median(clean_taxi$tips)
```

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

```
table(clean_taxi$tips == 0)
```

```
##
## FALSE  TRUE
## 33207 66793
```

For a typical amount for a tip, we notice that going off of the median value of 0 is a better indicator than the mean. This is due to the fact that certain upper values like our maximum value of 97 will likely skew the mean. Indeed, looking at our table, we can verify that the majority of people actually do not tip for a ride, causing our median to be 0.