R-Assignment2

Payton Kim

9/19/2019

1. Download the c2015 dataset to your computer. Use function getwd() to check the current working directory. Use setwd() to change the current directory to the c2015 file

```
getwd()
```

```
## [1] "C:/Users/student/Documents"
```

```
setwd("C:/Users/student/Downloads")
```

2. 2. We need to install a package to read the xlsx file. (Let's not change the xlsx to csv here) There are a few packages for this. I recommend to use the readxl package. This package is contained in the tidyverse package so if you already installed tidyverse, you should have it already. If not, install and load the readxl package by

```
library(readxl)
```

3. Use read_excel() to read the c2015 dataset. Use function class() to check the type of data you just read in. You will notice that the data now is not just a data frame, it is also a tibble. A tibble is a generalization of a data frame, so you can still use all the functions and syntax for data frame with tibble.

```
c <- read_excel("C:/Users/student/Downloads/c2015.xlsx")
class(c)
```

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

4. Use dim function to check the dimension of the data. Since this data is quite big, a common practice is to randomly subset the data to analyze. Use sample function to create a new dataset that has a random 1000 observations from the original data. Use set.seed(2019) before using the sample function to set the seed for the randomness so that everyone in class is working with the same random subset of the data.

```
dim(c)
```

```
## [1] 80587    28
```

```
set.seed(2019)
samplec <- c[sample(nrow(c),1000),]
```

5. Use summary function to have a quick look at the data. You will notice there is one variable is actually a constant. Remove that variable from the data.

```r
summary(samplec)
```

```
##     STATE              ST_CASE           VEH_NO           PER_NO
##  Length:1000        Min.   : 10020   Min.   : 0.000   Min.   : 1.000
##  Class :character   1st Qu.:122408   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median :270249   Median : 1.000   Median : 1.000
##                     Mean   :276444   Mean   : 1.385   Mean   : 1.697
##                     3rd Qu.:420726   3rd Qu.: 2.000   3rd Qu.: 2.000
##                     Max.   :560071   Max.   :13.000   Max.   :48.000
##
##      COUNTY            DAY           MONTH                HOUR
##  Min.   :  1.00   Min.   : 1.00   Length:1000        Min.   : 0.00
##  1st Qu.: 32.50   1st Qu.: 8.00   Class :character   1st Qu.: 8.00
##  Median : 71.00   Median :16.00   Mode  :character   Median :16.00
##  Mean   : 93.05   Mean   :15.89                      Mean   :14.26
##  3rd Qu.:117.00   3rd Qu.:24.00                      3rd Qu.:20.00
##  Max.   :810.00   Max.   :31.00                      Max.   :99.00
##
##      MINUTE          AGE                SEX                PER_TYP
##  Min.   : 0.00   Length:1000        Length:1000        Length:1000
##  1st Qu.:14.00   Class :character   Class :character   Class :character
##  Median :27.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :27.76
##  3rd Qu.:43.00
##  Max.   :59.00
##  NA's   :5
##    INJ_SEV            SEAT_POS           DRINKING              YEAR
##  Length:1000        Length:1000        Length:1000        Min.   :2015
##  Class :character   Class :character   Class :character   1st Qu.:2015
##  Mode  :character   Mode  :character   Mode  :character   Median :2015
##                                                           Mean   :2015
##                                                           3rd Qu.:2015
##                                                           Max.   :2015
##
##    MAN_COLL            OWNER             MOD_YEAR
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    TRAV_SP           DEFORMED           DAY_WEEK
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     ROUTE              LATITUDE        LONGITUD         HARM_EV
##  Length:1000        Min.   :21.30   Min.   :-160.34   Length:1000
##  Class :character   1st Qu.:33.48   1st Qu.: -97.59   Class :character
##  Mode  :character   Median :36.42   Median : -87.43   Mode  :character
##                     Mean   :36.72   Mean    : -91.83
##                     3rd Qu.:40.40   3rd Qu.: -81.41
##                     Max.   :61.54   Max.    : -67.72
##                     NA's   :7       NA's   :7
##    LGT_COND           WEATHER
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
c2015 = select(samplec, -"YEAR")
summary(c2015)
```

```
##     STATE              ST_CASE           VEH_NO           PER_NO
##  Length:1000        Min.   : 10020   Min.   : 0.000   Min.   : 1.000
##  Class :character   1st Qu.:122408   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median :270249   Median : 1.000   Median : 1.000
##                     Mean   :276444   Mean   : 1.385   Mean   : 1.697
##                     3rd Qu.:420726   3rd Qu.: 2.000   3rd Qu.: 2.000
##                     Max.   :560071   Max.   :13.000   Max.   :48.000
##
##      COUNTY             DAY           MONTH                HOUR
##  Min.   :  1.00   Min.   : 1.00   Length:1000        Min.   : 0.00
##  1st Qu.: 32.50   1st Qu.: 8.00   Class :character   1st Qu.: 8.00
##  Median : 71.00   Median :16.00   Mode  :character   Median :16.00
##  Mean   : 93.05   Mean   :15.89                      Mean   :14.26
##  3rd Qu.:117.00   3rd Qu.:24.00                      3rd Qu.:20.00
##  Max.   :810.00   Max.   :31.00                      Max.   :99.00
##
##      MINUTE           AGE                SEX                PER_TYP
##  Min.   : 0.00   Length:1000        Length:1000        Length:1000
##  1st Qu.:14.00   Class :character   Class :character   Class :character
##  Median :27.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :27.76
##  3rd Qu.:43.00
##  Max.   :59.00
##  NA's   :5
##    INJ_SEV           SEAT_POS          DRINKING
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    MAN_COLL            OWNER             MOD_YEAR
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    TRAV_SP           DEFORMED          DAY_WEEK
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     ROUTE              LATITUDE        LONGITUD         HARM_EV
##  Length:1000        Min.   :21.30   Min.   :-160.34   Length:1000
##  Class :character   1st Qu.:33.48   1st Qu.: -97.59   Class :character
##  Mode  :character   Median :36.42   Median : -87.43   Mode  :character
##                     Mean   :36.72   Mean   : -91.83
##                     3rd Qu.:40.40   3rd Qu.: -81.41
##                     Max.   :61.54   Max.   : -67.72
##                     NA's   :7       NA's   :7
##    LGT_COND          WEATHER
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

6. Check the number of missing values (NA) in each column.

```
sum(is.na(c2015))
```

```
## [1] 494
```

```
sum(is.na(c2015[,1]))
```

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

```
sum(is.na(c2015[,2]))
```

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

```
sum(is.na(c2015[,3]))
```

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

```
sum(is.na(c2015[,4]))
```

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

```
sum(is.na(c2015[,5]))
```

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

```
sum(is.na(c2015[,6]))
```

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

```
sum(is.na(c2015[,7]))
```

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

```
sum(is.na(c2015[,8]))
```

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

```
sum(is.na(c2015[,9]))
```

```
## [1] 5
```

```
sum(is.na(c2015[,10]))
```

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

```
sum(is.na(c2015[,11]))
```

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

```
sum(is.na(c2015[,12]))
```

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

```
sum(is.na(c2015[,13]))
```

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

```
sum(is.na(c2015[,14]))
```

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

```
sum(is.na(c2015[,15]))
```

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

```
sum(is.na(c2015[,16]))
```

```
## [1] 95
```

```
sum(is.na(c2015[,17]))
```

```
## [1] 95
```

```
sum(is.na(c2015[,18]))
```

```
## [1] 95
```

```
sum(is.na(c2015[,19]))
```

```
## [1] 95
```

```
sum(is.na(c2015[,20]))
```

```
## [1] 95
```

```
sum(is.na(c2015[,21]))
```

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

```
sum(is.na(c2015[,22]))
```

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

```
sum(is.na(c2015[,23]))
```

```
## [1] 7
```

```r
sum(is.na(c2015[,24]))
```

```
## [1] 7
```

```r
sum(is.na(c2015[,25]))
```

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

```r
sum(is.na(c2015[,26]))
```

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

```r
sum(is.na(c2015[,27]))
```

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

7. There are missing values in this data that are not NAs. Identify the form of these missing values. Check the number of these missing values in each column. Notice that you may want to use na.rm = TRUE when counting these missing values.

```r
colSums(c2015 == "Unknown")
```

```
##    STATE  ST_CASE   VEH_NO   PER_NO   COUNTY      DAY    MONTH     HOUR 
##        0        0        0        0        0        0        0        0 
##   MINUTE      AGE      SEX  PER_TYP  INJ_SEV SEAT_POS DRINKING MAN_COLL 
##       NA       16        9        0        8       10        0       NA 
##    OWNER MOD_YEAR  TRAV_SP DEFORMED DAY_WEEK    ROUTE LATITUDE LONGITUD 
##       NA       NA       NA       NA        0       36       NA       NA 
##  HARM_EV LGT_COND  WEATHER 
##        0        5        0
```

8. Change the missing values in SEX variable to “Female”

```r
c2015$SEX <- ifelse(c2015$SEX == "Unknown","Female",c2015$SEX)

colSums(c2015 == "Unknown")
```

```
##    STATE  ST_CASE   VEH_NO   PER_NO   COUNTY      DAY    MONTH     HOUR 
##        0        0        0        0        0        0        0        0 
##   MINUTE      AGE      SEX  PER_TYP  INJ_SEV SEAT_POS DRINKING MAN_COLL 
##       NA       16        0        0        8       10        0       NA 
##    OWNER MOD_YEAR  TRAV_SP DEFORMED DAY_WEEK    ROUTE LATITUDE LONGITUD 
##       NA       NA       NA       NA        0       36       NA       NA 
##  HARM_EV LGT_COND  WEATHER 
##        0        5        0
```

9. Fix the AGE variable so that it is in the right form and has no missing values. Hint: • Change the value Less than 1 to 0 (string 0, not a number 0) • Change the type of the variable to numeric using as.numeric function • Change the missing values to the average of the age.

```r
c2015$AGE <- ifelse(c2015$AGE == "Less than 1","0", c2015$AGE)
c2015$AGE <- as.numeric(c2015$AGE)
```

```
## Warning: NAs introduced by coercion
```

```r
mean <- mean(c2015$AGE,na.rm = TRUE)
c2015$AGE <- ifelse(is.na(c2015$AGE),mean, c2015$AGE)
```

10. Put the TRAV_SP(Travel Speed) variable in the right form (type) and remove all missing values. Calculate the average speed. You can use a non-base R function for this question. Hint: check out the function str_replace

```r
library(stringr)
noMPH <- str_replace(c2015$TRAV_SP, "MPH","")
noMPHnumeric <- as.numeric(noMPH)
```

```
## Warning: NAs introduced by coercion
```

```r
c2015$TRAV_SP <- noMPHnumeric
new2015 <- na.omit(c2015)
travsp <- na.omit(noMPHnumeric)
mean(travsp)
```

```
## [1] 50.77188
```

11. Compare the average speed of those who had “No Apprent Injury” and the rest. What do you observe?

```r
mean(new2015$TRAV_SP[new2015$INJ_SEV == "No Apparent Injury (O)"])
```

```
## [1] 44.51724
```

```r
mean(new2015$TRAV_SP[new2015$INJ_SEV != "No Apparent Injury (O)"])
```

```
## [1] 53.09914
```

```r
###Those with no apparent injury were driving, on average, slower than those with injuries
```

12. Use the SEAT_POS variable to filter the data so that there is only drivers in the dataset. Compare the average speed of man drivers and woman drivers. Comment on the results.

```r
question12 <- new2015%>%
  filter(new2015$SEAT_POS == "Front Seat, Left Side")
man <- mean(question12$TRAV_SP[question12$SEX == "Male"])
woman <- mean(question12$TRAV_SP[question12$SEX == "Female"])
man
```

```
## [1] 51.63087
```

```
woman
```

```
## [1] 45.57895
```

```
###Women were, on average, driving about 6 MPH slower than men were if they got into a car accident
```

13. Compare the average speed of drivers who drink and those who do not. Comment on the results. Hint: This calculation can be done manually or by using the aggregate function or by function in base R.

```
drink <- mean(question12$TRAV_SP[question12$DRINKING == "Yes (Alcohol Involved)"])
nodrink <- mean(question12$TRAV_SP[question12$DRINKING != "Yes (Alcohol Involved)"])

drink
```

```
## [1] 68.25
```

```
nodrink
```

```
## [1] 47.07865
```

```
###The drivers in which alcohol was inolver drove, on average, over 20MPH more than drivers who did not
```

14. Hypothesize about the age range of drivers who may drive more aggressively. Test your hypothesis by comparing the average speed of those in this age range and the rest. Comment on the results.

```
###I would hypothesize that drivers less than age 30 drive more aggressively/faster than drivers who ar
lessthan30 <- mean(question12$TRAV_SP[question12$AGE < 30])
thirtyandup <- mean(question12$TRAV_SP[question12$AGE >= 30])

lessthan30
```

```
## [1] 54.32787
```

```
thirtyandup
```

```
## [1] 48.11724
```

```
###Drivers who were less than the age of thirty were driving on average 54MPH which is about 6MPH more
```

15. If the data did not confirm your hypothesis in 14. Could you identify an age group of drivers who may drive more aggressively?

```
###The data in my hypothesis from question 14 was correct.
```