Assignment 2

McKenzie Kozma

9/16/2019

1 - 5

```
getwd()
```

```
## [1] "C:/Users/student/Documents/GitHub/A2"
```

```
library(readxl)
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   0.8.3     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()
```

```
c2015 <- read_xlsx("c2015.xlsx")
class(c2015)
```

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

```
dim(c2015)
```

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

```
set.seed(2019)
c2015_sample <- c2015[sample(nrow(c2015), 1000),]
summary(c2015_sample)
```

```
##     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
#year is a constant. this variable will be removed from the dataset.
c2015_sample$YEAR <- NULL
```

6 - 10

```r
colSums(is.na(c2015_sample))
```

```
##    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
##        5        0        0        0        0        0        0       95
##    OWNER MOD_YEAR  TRAV_SP DEFORMED DAY_WEEK    ROUTE LATITUDE LONGITUD
##       95       95       95       95        0        0        7        7
##  HARM_EV LGT_COND  WEATHER
##        0        0        0
```

```r
colSums(c2015_sample == "Unknown", na.rm = TRUE)
```

```
##    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
##        0       16        9        0        8       10        0        2
##    OWNER MOD_YEAR  TRAV_SP DEFORMED DAY_WEEK    ROUTE LATITUDE LONGITUD
##       23       16       75       20        0       36        0        0
##  HARM_EV LGT_COND  WEATHER
##        0        5        0
```

```r
c2015_sample$SEX[c2015_sample$SEX == "Unknown"] <- "Female"
c2015_sample$AGE[c2015_sample$AGE == "Less than 1"] <- "0"
c2015_sample$AGE <- as.numeric(c2015_sample$AGE)
```

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

```r
c2015_sample$AGE[is.na(c2015_sample$AGE)] <- mean(c2015_sample$AGE, na.rm = TRUE)

c2015_sample$TRAV_SP <- as.numeric(str_remove(c2015_sample$TRAV_SP, "MPH"))
```

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

```r
c2015_sample2 <- c2015_sample[!(is.na(c2015_sample$TRAV_SP)), ]
```

11 - 15

```r
mean(c2015_sample2$TRAV_SP[c2015_sample2$INJ_SEV == "No Apparent Injury (O)"], na.rm = TRUE)
```

```
## [1] 44.63636
```

```r
mean(c2015_sample2$TRAV_SP[c2015_sample2$INJ_SEV != "No Apparent Injury (O)"], na.rm = TRUE)
```

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

```r
#those who have no apparent injury were traveling, on average, at a lower speed

c2015_sample3 <- c2015_sample2[c2015_sample2$SEAT_POS == "Front Seat, Left Side", ]

by(c2015_sample3$TRAV_SP, c2015_sample3$SEX, FUN = mean)
```

```
## c2015_sample3$SEX: Female
## [1] 45.57895
## ------------------------------------------------------------
## c2015_sample3$SEX: Male
## [1] 51.65333
```

```r
#males drive faster on average in comparison to females

by(c2015_sample3$TRAV_SP, c2015_sample3$DRINKING, FUN = mean)
```

```
## c2015_sample3$DRINKING: No (Alcohol Not Involved)
## [1] 44.94074
## ------------------------------------------------------------
## c2015_sample3$DRINKING: Not Reported
## [1] 52.7
## ------------------------------------------------------------
## c2015_sample3$DRINKING: Unknown (Police Reported)
## [1] 54.14706
## ------------------------------------------------------------
## c2015_sample3$DRINKING: Yes (Alcohol Involved)
## [1] 68.25
```

```r
#those who were drinking were driving faster, on average, than those who were not drinking


#i hypothesized that those under the age of 25 would drive faster/more aggressively
by(c2015_sample3$TRAV_SP, c2015_sample3$AGE < 25, FUN = mean)
```

```
## c2015_sample3$AGE < 25: FALSE
## [1] 48.52381
## ------------------------------------------------------------
## c2015_sample3$AGE < 25: TRUE
## [1] 56.25641
```

```r
#my hypothesis appears to be true, those who were under the age of 25 were driving faster on average
```