Assignment 2

Hyungue Lim

9/16/2019

1, 2

```
library(readxl)



c2015 <- read_excel("~/MATH 421/c2015.xlsx")
```

3

```
class(c2015)
```

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

4

```
dim(c2015)
```

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

```
set.seed(2019)

c2015_sample <- c2015[sample(nrow(c2015),1000),]
```

5

```
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
##
```

```
##
##
##
```

```
c2015_sample <- c2015_sample[,-16]
```

6

```
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
```

```
'summary function shows NAs in data'
```

```
## [1] "summary function shows NAs in data"
```

```
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
##  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
##
##
##
##
```

7

```
colSums(c2015_sample == "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

```
a <- c2015_sample$SEX == "Unknown"
c2015_sample$SEX[a] <- "Female"
```

9

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

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

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

10

```
library("stringr")
c2015_sample$TRAV_SP <- str_replace(c2015_sample$TRAV_SP, " MPH", "")
c2015_sample$TRAV_SP <- str_replace(c2015_sample$TRAV_SP, "Stopped", "0")

c2015_sample$TRAV_SP <- as.numeric(c2015_sample$TRAV_SP)
```

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

```
mean(c2015_sample$TRAV_SP, na.rm=TRUE)
```

```
## [1] 43.79245
```

```
c2015_sample1 = c2015_sample[!is.na(c2015_sample$TRAV_SP),]
```

11

```
mean(c2015_sample1$TRAV_SP[c2015_sample1$INJ_SEV == "No Apparent Injury (O)"])
```

```
## [1] 33.57265
```

```r
mean(c2015_sample1$TRAV_SP[c2015_sample1$INJ_SEV != "No Apparent Injury (O)"])
```

```
## [1] 48.5
```

```r
#People with no apparent injury were driving slower on average than people with some observed injuries.
```

12

```r
c2015_sample2 <- subset(c2015_sample1, SEAT_POS == "Front Seat, Left Side")

mean(c2015_sample2$TRAV_SP[c2015_sample2$SEX == "Male"])
```

```
## [1] 45.57647
```

```r
mean(c2015_sample2$TRAV_SP[c2015_sample2$SEX == "Female"])
```

```
## [1] 37.11429
```

```r
#Men were driving faster than women on average
```

13

```r
aggregate(c2015_sample2$TRAV_SP, list(c2015_sample2$DRINKING), mean)[c(1,4),]
```

```
##                     Group.1        x
## 1 No (Alcohol Not Involved) 37.22086
## 4    Yes (Alcohol Involved) 65.89655
```

```r
#Drivers who consumed alcohol were driving faster than those who did not.
```

14

```r
#My hypothesis is that drivers under the age of 25 drive more aggresively than those who are 25 and abo

mean(c2015_sample2$TRAV_SP[c2015_sample2$AGE < 25])
```

```
## [1] 46.68085
```

```r
mean(c2015_sample2$TRAV_SP[c2015_sample2$AGE >= 25])
```

```
## [1] 42.23834
```

```
#The average driving speed for those under 25 is higher than those who are 25 and above. This might ind
```

15

```
#Data confirmed my hypothesis in #14
```