Assignment 2

Huiyi Huang

2019/9/17

#1-#5

```
setwd("C:/Users/student/Documents/Fall2019/")
getwd()
```

```
## [1] "C:/Users/student/Documents/Fall2019"
```

```
library(stringr)
library(readxl)
read_excel("C:/Users/student/Documents/Fall2019/c2015.xlsx")
```

```
## # A tibble: 80,587 x 28
##    STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE AGE   SEX
##    <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <chr> <chr>
##  1 Alab~   10001      1      1    127     1 Janu~     2     40 68    Male
##  2 Alab~   10002      1      1     83     1 Janu~    22     13 49    Male
##  3 Alab~   10003      1      1     11     1 Janu~     1     25 31    Male
##  4 Alab~   10003      1      2     11     1 Janu~     1     25 20    Fema~
##  5 Alab~   10004      1      1     45     4 Janu~     0     57 40    Male
##  6 Alab~   10005      1      1     45     7 Janu~     7      9 24    Male
##  7 Alab~   10005      2      1     45     7 Janu~     7      9 60    Male
##  8 Alab~   10006      1      1    111     8 Janu~     9     59 64    Male
##  9 Alab~   10006      1      2    111     8 Janu~     9     59 17    Male
## 10 Alab~   10007      1      1     89     8 Janu~    18     33 80    Male
## # ... with 80,577 more rows, and 17 more variables: PER_TYP <chr>,
## #   INJ_SEV <chr>, SEAT_POS <chr>, DRINKING <chr>, YEAR <dbl>,
## #   MAN_COLL <chr>, OWNER <chr>, MOD_YEAR <chr>, TRAV_SP <chr>,
## #   DEFORMED <chr>, DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>,
## #   LONGITUD <dbl>, HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

```
c2015=read_excel("C:/Users/student/Documents/Fall2019/c2015.xlsx")
class(c2015)
```

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

```
dim(c2015)
```

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

```
set.seed(2019)
x<- c2015[sample(1:dim(c2015)[1],1000,replace = FALSE),]
summary(x)
```

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

```
#remove constant
x$YEAR = NULL
x$LONGITUD = NULL
x$LATITUDE = NULL
x$DAY = NULL
x$MONTH = NULL
x$HOUR = NULL
x$VEH_NO = NULL
x$PER_NO = NULL
x$ROUTE = NULL
x$DAY_WEEK = NULL
x$MINUTE = NULL
x$MOD_YEAR = NULL
x$ST_CASE = NULL
```

#6

```
colSums(is.na(x))
```

```
##    STATE   COUNTY      AGE      SEX  PER_TYP  INJ_SEV SEAT_POS DRINKING
##        0        0        0        0        0        0        0        0
## MAN_COLL    OWNER  TRAV_SP DEFORMED  HARM_EV LGT_COND  WEATHER
##       95       95       95       95        0        0        0
```

#7

```
sum(x$STATE=='Unknown', na.rm = TRUE)
```

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

```
sum(x$COUNTY=='Unknown', na.rm = TRUE)
```

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

```
sum(x$SEX=='Unknown', na.rm = TRUE)
```

```
## [1] 9
```

```r
sum(x$AGE=='Unknown', na.rm = TRUE)
```

```
## [1] 16
```

```r
sum(x$PER_TYP=='Unknown', na.rm = TRUE)
```

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

```r
sum(x$INJ_SEV=='Unknown', na.rm = TRUE)
```

```
## [1] 8
```

```r
sum(x$SEAT_POS=='Unknown', na.rm = TRUE)
```

```
## [1] 10
```

```r
sum(x$DRINKING=='Unknown', na.rm = TRUE)
```

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

```r
sum(x$MAN_COLL=='Unknown', na.rm = TRUE)
```

```
## [1] 2
```

```r
sum(x$OWNER=='Unknown', na.rm = TRUE)
```

```
## [1] 23
```

```r
sum(x$TRAV_SP=='Unknown', na.rm = TRUE)
```

```
## [1] 75
```

```r
sum(x$DEFORMED=='Unknown', na.rm = TRUE)
```

```
## [1] 20
```

```r
sum(x$HARM_EV=='Unknown', na.rm = TRUE)
```

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

```r
sum(x$LGT_COND=='Unknown', na.rm = TRUE)
```

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

```r
sum(x$WEATHER=='Unknown', na.rm = TRUE)
```

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

#8

```r
table(x$SEX)
```

```
##
##  Female    Male Not Rep Unknown
##     336     653       2       9
```

```r
x[x$SEX=='Unknown','SEX']='Female'
x[x$SEX=='Not Rep','SEX']='Female'
table(x$SEX)
```

```
##
## Female   Male
##    347    653
```

```r
#Question 9 is at the end.
```

#10

```r
x$TRAV_SP= str_replace(x$TRAV_SP,' MPH','')
x$TRAV_SP<-as.numeric(as.character(x$TRAV_SP))
```

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

```r
x = x[!is.na(x$TRAV_SP),]
mean(x$TRAV_SP)
```

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

#11

```r
mean(x$TRAV_SP[x$INJ_SEV=='No Apparent Injury (O)'])
```

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

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

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

```r
#People who have no injuries are having lower speed comparing to others who has injuries.
```

#12

```r
x = x[x$SEAT_POS=='Front Seat, Left Side',]
mean(x$TRAV_SP[x$SEX=='Male'])>mean(x$TRAV_SP[x$SEX=='Female'])
```

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

```r
#Male drivers tend too drive faster than female drivers
```

#13

```r
mean(x$TRAV_SP[x$DRINKING=='Yes (Alcohol Involved)']) > mean(x$TRAV_SP[x$DRINKING=='No (Alcohol Not Inv
```

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

```r
#or
by(x$TRAV_SP, x$DRINKING, FUN = mean)
```

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

```r
#Drivers who drink alcohol tend to drive faster than drivers that don't drink alcohol.
```

#14

```r
#Hypothsis: Young drivers (16-25)tend to drive faster than adult drivers(25-35).
mean(x$TRAV_SP[x$AGE<25]) > mean(x$TRAV_SP[x$AGE<35 &x$AGE>25])
```

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

```r
# Hypothesis proven true. Yong drivers are more aggressive drivers comparing to adult drivers.
```

#9

```r
#change"less than 1" to 0

x[x$AGE=='Less than 1','AGE']=0
```

```r
#change data type
typeof(x$AGE)
```

```
## [1] "character"
```

```r
x$AGE<-as.numeric(as.character(x$AGE))
```

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

```r
typeof(x$AGE)
```

```
## [1] "double"
```

```r
#change missing value to avg

y=mean(x$AGE, na.rm = 1)
x[is.na(x$AGE),]=y
x[x$AGE=='Unknown','AGE']=y
```