Assignment 2

Rachel Cardarelli

9/19/2019

Question 1

```
getwd()
```

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

```
setwd("C:/Users/student/Documents/R Programming")
```

Question 2

```
#install.packages('readxl')
library(readxl)
```

Question 3

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

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

Question 4

```
dim(c2015)
```

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

```
set.seed(2019); index <- sample(1:nrow(c2015), 1000)
c2015 <- c2015[index,]
```

Question 5

```
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              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$YEAR <- NULL
```

Question 6

```
colSums(is.na.data.frame(c2015), na.rm = FALSE)
```

```
##    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
```

Question 7

```
sum(c2015 == "Unknown", na.rm = TRUE)
```

```
## [1] 220
```

```
sum(c2015 == "Unkno", na.rm = TRUE)
```

```
## [1] 12
```

```
sum(c2015 == "Not Rep", na.rm = TRUE)
```

```
## [1] 461
```

Question 8

```
sum(c2015$SEX=="Unknown")
```

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

```r
sum(c2015$SEX=="Unkno")
```

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

```r
sum(c2015$SEX=="Not Rep")
```

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

```r
c2015$SEX[c2015$SEX == "Unknown"] <- "Female"
c2015$SEX[c2015$SEX == "Not Rep"] <- "Female"
```

Question 9

```r
c2015$AGE[c2015$AGE == "Less than 1"] <- "0"
c2015$AGE[c2015$AGE == "Unknown"] <- "NA"
c2015$AGE <- as.numeric(c2015$AGE)
```

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

```r
class(c2015$AGE)
```

```
## [1] "numeric"
```

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

```
## # A tibble: 6 x 27
##   STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE   AGE SEX
##   <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <dbl> <chr>
## 1 New ~  340336      1      1     27    19 Sept~     3     17  39.3 Fema~
## 2 Ariz~   40327      1      1     13     7 May      22     15  47   Fema~
## 3 Tenn~  470789      1      1    163     2 Dece~     8     26  23   Male
## 4 Minn~  270119      2      4     59    16 May      21     59  15   Fema~
## 5 Miss~  290576      1      1    201     2 Octo~    15     38  55   Male
## 6 Cali~   62865      1      1     19     6 June     15     20  56   Male
## # ... with 16 more variables: PER_TYP <chr>, INJ_SEV <chr>,
## #   SEAT_POS <chr>, DRINKING <chr>, 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>
```

Question 10

```r
#install.packages("stringr", repos='http://cran.us.r-project.org')
library("stringr")

c2015$TRAV_SP <- str_replace(c2015$TRAV_SP," MPH","")

c2015$TRAV_SP[c2015$TRAV_SP == "Unknown"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Not Rep"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Greater"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Stopped"] <- "0"

c2015$TRAV_SP <- as.numeric(c2015$TRAV_SP)
```

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

```r
mean(c2015$TRAV_SP,na.rm=TRUE)
```

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

Question 11

```r
#Those with no apparent injuries had lower travel speeds on average
mean(c2015$TRAV_SP[c2015$INJ_SEV=="No Apparent Injury (0)"],na.rm=TRUE)
```

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

```r
mean(c2015$TRAV_SP,na.rm=TRUE)
```

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

Question 12

```r
drivers <- c2015[c2015$SEAT_POS == "Front Seat, Left Side",]
head(drivers,50)
```

```
## # A tibble: 50 x 27
##    STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE   AGE SEX
##    <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <dbl> <chr>
##  1 New ~  340336      1      1     27    19 Sept~     3     17  39.3 Fema~
##  2 Ariz~   40327      1      1     13     7 May      22     15  47   Fema~
##  3 Tenn~  470789      1      1    163     2 Dece~     8     26  23   Male
##  4 Miss~  290576      1      1    201     2 Octo~    15     38  55   Male
##  5 Cali~   62865      1      1     19     6 June     15     20  56   Male
##  6 Alab~   10286      5      1    115    30 May      14     36  32   Male
##  7 Sout~  450153      1      1     29    19 March    14     15  54   Male
##  8 Indi~  180596      2      1    141    12 Octo~    17     50  44   Male
##  9 Texas  481831      1      1    347     9 Augu~     4     56  39   Fema~
## 10 Kent~  210262      2      1    151    17 June     13     25  44   Male
```

```
## # ... with 40 more rows, and 16 more variables: PER_TYP <chr>,
## #   INJ_SEV <chr>, SEAT_POS <chr>, DRINKING <chr>, MAN_COLL <chr>,
## #   OWNER <chr>, MOD_YEAR <chr>, TRAV_SP <dbl>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

```
#the mean speeds of female and male drivers are nearly identical
mean(drivers$TRAV_SP[drivers$SEX =="Female"],na.rm = TRUE)
```

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

```
mean(drivers$TRAV_SP[drivers$SEX =="Male"],na.rm = TRUE)
```

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

Question 13

```
mean(c2015$TRAV_SP[c2015$DRINKING == "Yes (Alcohol Involved)"],na.rm=TRUE)
```

```
## [1] 66.3871
```

```
mean(c2015$TRAV_SP[c2015$DRINKING == "No (Alcohol Not Involved)"],na.rm=TRUE)
```

```
## [1] 37.22222
```

Question 14

```
#Hypothesis: younger drivers drive more aggresively
mean(subset(c2015, SEAT_POS == "Front Seat, Left Side" & AGE <= 21)$TRAV_SP,na.rm=TRUE)
```

```
## [1] 43.77778
```

```
mean(subset(c2015, SEAT_POS == "Front Seat, Left Side")$TRAV_SP, na.rm = TRUE)
```

```
## [1] 43.10833
```

```
#my hypothesis was incorrect. the speeds are almost the same
```

Question 15

```
mean(subset(c2015, SEAT_POS == "Front Seat, Left Side" & AGE > 21 & AGE <= 32)$TRAV_SP,na.rm=TRUE)
```

```
## [1] 46.34483
```

```
#21-31 year-olds drive more aggresively on average
```