R Assignment 2

Jeff Moise

9/16/2019

Questions

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.

```
work_dir <- "C://Users//student//Documents//RStudio"
getwd()
```

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

```
setwd(work_dir)
```

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

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

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`.

```
path <- "C:/Users/student/Documents/RStudio/c2015.xlsx"
library(readxl)
data_excel=read_excel(path)
class(data_excel)
```

```
## [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(data_excel)
```

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

```r
set.seed(2019)
c2015sample<-data_excel[sample(1:80587,1000),]
dim(c2015sample)
```

```
## [1] 1000   28
```

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(c2015sample)
```

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

```
data_excel2 = subset(c2015sample, select = -c(YEAR) )
summary(data_excel2)
```

```
##     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.

```
colSums(is.na(data_excel2))
```

```
##    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 LATITUD LONGITUD
##       95       95       95       95        0        0        7        7
##  HARM_EV LGT_COND  WEATHER
##        0        0        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
table(data_excel2$TRAV_SP)
```

```
##
## 004 MPH 005 MPH 006 MPH 008 MPH 009 MPH 010 MPH 014 MPH 015 MPH 018 MPH
##       1       9       1       1       1      14       1       7       1
## 020 MPH 025 MPH 026 MPH 030 MPH 033 MPH 034 MPH 035 MPH 038 MPH 040 MPH
##       4       8       1       7       1       1      19       1      22
## 041 MPH 043 MPH 045 MPH 048 MPH 049 MPH 050 MPH 053 MPH 055 MPH 057 MPH
##       1       1      28       2       3      15       3      37       1
## 058 MPH 059 MPH 060 MPH 062 MPH 063 MPH 064 MPH 065 MPH 066 MPH 067 MPH
##       3       1      19       2       1       3      26       1       1
## 068 MPH 070 MPH 073 MPH 075 MPH 076 MPH 077 MPH 079 MPH 080 MPH 082 MPH
##       3      25       3      12       1       1       1      10       1
## 083 MPH 085 MPH 089 MPH 090 MPH 100 MPH 107 MPH 113 MPH Not Rep Stopped
##       1       4       1       5       2       1       1     459      51
## Unknown
##      75
```

```r
table(data_excel2$OWNER)
```

```
##
## Driver (in this crash) Not Registered Owner (Other Private Owner Listed)
##                                                                     293
##                               Driver (in this crash) was  Registered Owner
##                                                                     469
##                         Driverless/Motor Vehicle Parked/Stopped Off Roadway
##                                                                       1
##                                    Not Applicable, Vehicle Not Registered
##                                                                      16
##                                                                 Unknown
##                                                                      23
##                  Vehicle Registered as Business/Company/Government Vehicle
##                                                                      89
##                                      Vehicle Registered as Rental Vehicle
##                                                                      11
##                                   Vehicle was Stolen (reported by police)
##                                                                       3
```

```r
table(data_excel2$DEFORMED)
```

```
##
##  Disabling Damage Functional Damage      Minor Damage         No Damage
##               660                87                78                16
##      Not Reported           Unknown
##                44                20
```

8. Change the missing values in `SEX` variable to "Female"

```r
data_excel2$SEX[data_excel2$SEX=='Not Rep']= 'Female'
data_excel2$SEX[data_excel2$SEX=='Unknown']= 'Female'
table(data_excel2$SEX)
```

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

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.

```
data_excel2$SEX[data_excel2$AGE=='Less than 1']= '0'

data_excel2$AGE <- as.numeric(data_excel2$AGE)
```

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

```
avgAge<-mean(data_excel2$AGE[!is.na(data_excel2$AGE)])

data_excel2$AGE[is.na(data_excel2$AGE)]<- avgAge

avgAge
```

```
## [1] 39.48315
```

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`

```
library(stringr)
data_excel2$TRAV_SP[data_excel2$TRAV_SP=='Stopped'] <- '0'
data_excel2$TRAV_SP[data_excel2$TRAV_SP=='Not Rep' | data_excel2$TRAV_SP=='Unknown'] <- NA
data_excel2$TRAV_SP<- stringr::str_replace(data_excel2$TRAV_SP," MPH", "")
data_excel2$TRAV_SP <- as.numeric(data_excel2$TRAV_SP)
mean(data_excel2$TRAV_SP, na.rm = TRUE)
```

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

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

```
data_excel2<- subset(data_excel2,data_excel2$INJ_SEV=='No Apparent Injury (O)')
mean(data_excel2$TRAV_SP, na.rm=TRUE)
```

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

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.

```
table(data_excel2$SEAT_POS)
```

```
##
##       Front Seat, Left Side          Front Seat, Middle
##                         175                           1
##      Front Seat, Right Side                Not Reported
##                          54                           1
## Other Passenger in enclosed      Second Seat, Left Side
##                           1                          16
##         Second Seat, Middle     Second Seat, Right Side
##                           3                          10
##        Second Seat, Unknown       Third Seat, Left Side
##                           1                           1
##                     Unknown
##                           2
```

```
data_excel2 <-subset(data_excel2,data_excel2$SEAT_POS=='Front Seat, Left Side')
table(data_excel2$SEAT_POS)
```

```
##
## Front Seat, Left Side
##                   175
```

```
MaleSet<-(subset(data_excel2,data_excel2$SEX=='Male'))
FemaleSet<-(subset(data_excel2,data_excel2$SEX=='Female'))
mean(MaleSet$TRAV_SP,na.rm = TRUE)
```

```
## [1] 36.33333
```

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

```
## [1] 34.05263
```

```
# Males tend to drive faster slightly in this data set
```

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. For example:

```
aggregate(data_excel2$TRAV_SP, by=list(data_excel2$DRINKING),FUN=mean, na.rm=TRUE)
```

```
##                     Group.1        x
## 1 No (Alcohol Not Involved) 34.32857
## 2              Not Reported 28.33333
## 3 Unknown (Police Reported) 60.00000
## 4    Yes (Alcohol Involved) 54.75000
```

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.

```r
# I hypothesize that drivers under 30 would drive faster
data_excel2$AGE[data_excel2$AGE=='Unknown'] <- NA
over30 = subset(data_excel2$TRAV_SP,data_excel2$AGE>30)
under30 = subset(data_excel2$TRAV_SP,data_excel2$AGE<=30)
over30 <- as.numeric(over30)
under30 <- as.numeric(under30)
mean(over30,na.rm=TRUE)
```

```
## [1] 35.3
```

```r
mean(under30,na.rm=TRUE)
```

```
## [1] 37.31579
```

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

```r
# I was unable to find a hige difference in driving speeds, it seems to random across the ages
aggregate(data_excel2$TRAV_SP, by=list(data_excel2$AGE),FUN=mean, na.rm=TRUE)
```

```
##     Group.1        x
## 1  17.00000 49.00000
## 2  18.00000      NaN
## 3  19.00000 40.00000
## 4  20.00000 40.00000
## 5  21.00000 32.50000
## 6  22.00000 33.00000
## 7  23.00000 40.00000
## 8  24.00000      NaN
## 9  25.00000      NaN
## 10 26.00000 70.00000
## 11 27.00000 45.00000
## 12 28.00000 50.00000
## 13 29.00000 18.33333
## 14 30.00000      NaN
## 15 31.00000 15.00000
## 16 32.00000 55.50000
## 17 33.00000 62.50000
## 18 34.00000 65.00000
## 19 35.00000 37.60000
## 20 36.00000 17.50000
## 21 37.00000 40.00000
## 22 38.00000      NaN
## 23 39.00000 55.00000
## 24 39.48315 40.00000
## 25 40.00000      NaN
## 26 41.00000 12.50000
## 27 43.00000 17.50000
## 28 44.00000 65.00000
## 29 45.00000      NaN
## 30 46.00000      NaN
## 31 47.00000 24.00000
```

```
## 32 48.00000 32.50000
## 33 49.00000  0.00000
## 34 50.00000  0.00000
## 35 51.00000 21.66667
## 36 52.00000      NaN
## 37 53.00000 57.50000
## 38 54.00000 30.00000
## 39 55.00000 60.00000
## 40 56.00000 58.00000
## 41 57.00000 32.50000
## 42 58.00000 65.00000
## 43 59.00000  0.00000
## 44 60.00000 10.00000
## 45 61.00000 35.00000
## 46 62.00000 33.33333
## 47 63.00000 37.50000
## 48 64.00000 40.00000
## 49 68.00000 20.00000
## 50 70.00000      NaN
## 51 71.00000 70.00000
## 52 73.00000 40.00000
## 53 74.00000      NaN
## 54 75.00000      NaN
## 55 76.00000 65.00000
## 56 78.00000 36.50000
## 57 89.00000 25.00000
```