Assignment 2: Data Wrangling with Base R

Louis Dion

September 12, 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.

```
getwd()
```

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

```
setwd("C:/Users/student/Documents/MATH421/data")
#Will not allow me to set the directory to the exact file
#additionally it only sets the directory for this chunk and resets for other chunks
```

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

```
library(readxl)
```

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.

```
c2015<-read_excel('MATH421/data/c2015.xlsx')
#using directory that was reset to because directory did not carry over to this chunk
class(c2015)
```

```
## [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(c2015)
```

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

```
set.seed(2019)
sample2015<-c2015[sample(nrow(c2015),1000),] #########
sample2015
```

```
## # A tibble: 1,000 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 New ~  340336      1      1     27    19 Sept~     3     17 Unkn~ Unkn~
##  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
##  7 New ~  330095      0      1     15     3 Dece~    14     32 26    Male
##  8 Iowa   190173      0      1    127    30 Augu~    20     20 63    Male
##  9 Cali~   62263      2      4     13    17 Dece~     7     41 6     Male
## 10 Alab~   10286      5      1    115    30 May      14     36 32    Male
## # ... with 990 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>
```

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.

```
summary(sample2015)
```

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

##

```
sample2015<-subset(sample2015,select= -YEAR)
```

6. Check the number of missing values (NA) in each column.

```
x<-is.na(sample2015)
sum(x)
```

```
## [1] 494
```

```
colSums(x)
```

```
##    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
```

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.

```
sample2015
```

```
## # A tibble: 1,000 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> <chr> <chr>
##  1 New ~  340336      1      1     27    19 Sept~     3     17 Unkn~ Unkn~
##  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
##  7 New ~  330095      0      1     15     3 Dece~    14     32 26    Male
##  8 Iowa   190173      0      1    127    30 Augu~    20     20 63    Male
##  9 Cali~   62263      2      4     13    17 Dece~     7     41 6     Male
## 10 Alab~   10286      5      1    115    30 May      14     36 32    Male
## # ... with 990 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 <chr>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

```
#By looking at the sample dataset, some variable have values "Unknown" "Not Rep" that are missing value

sample2015<-replace(sample2015,sample2015=="Unknown"|sample2015=="Not Rep",NA)
sample2015
```

```
## # A tibble: 1,000 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> <chr> <chr>
##  1 New ~  340336      1      1     27    19 Sept~     3     17 <NA>  <NA>
##  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
##  7 New ~  330095      0      1     15     3 Dece~    14     32 26    Male
##  8 Iowa   190173      0      1    127    30 Augu~    20     20 63    Male
##  9 Cali~   62263      2      4     13    17 Dece~     7     41 6     Male
## 10 Alab~   10286      5      1    115    30 May      14     36 32    Male
## # ... with 990 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 <chr>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

```
x<-is.na(sample2015)
sum(x)
```

```
## [1] 1175
```

```
colSums(x)
```

```
##    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       16       11        0        8       10        0       97
##    OWNER MOD_YEAR  TRAV_SP DEFORMED DAY_WEEK    ROUTE LATITUDE LONGITUD
##      118      111      629      115        0       36        7        7
##  HARM_EV LGT_COND  WEATHER
##        0        5        0
```

8. Change the missing values in SEX variable to “Female”

```
sample2015['SEX'][is.na(sample2015['SEX'])]<-"Female"
sample2015
```

```
## # A tibble: 1,000 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> <chr> <chr>
##  1 New ~  340336      1      1     27    19 Sept~     3     17 <NA>  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
##  7 New ~  330095      0      1     15     3 Dece~    14     32 26    Male
##  8 Iowa   190173      0      1    127    30 Augu~    20     20 63    Male
##  9 Cali~   62263      2      4     13    17 Dece~     7     41 6     Male
## 10 Alab~   10286      5      1    115    30 May      14     36 32    Male
## # ... with 990 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 <chr>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

9. Fix the AGE variable so that it is in the right form and has no missing values.

```
#step 1
sample2015['AGE'][sample2015['AGE']=='Less than 1']<-'0'
#step 2
sample2015$AGE<-as.numeric(sample2015$AGE)
#step 3
sample2015$AGE[is.na(sample2015$AGE)]<-mean(sample2015$AGE,na.rm=TRUE)
sample2015
```

```
## # A tibble: 1,000 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
##  7 New ~  330095      0      1     15     3 Dece~    14     32  26   Male
##  8 Iowa   190173      0      1    127    30 Augu~    20     20  63   Male
##  9 Cali~   62263      2      4     13    17 Dece~     7     41   6   Male
## 10 Alab~   10286      5      1    115    30 May      14     36  32   Male
## # ... with 990 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 <chr>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

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.

```
sample2015$TRAV_SP<-substr(sample2015$TRAV_SP, 1, nchar(sample2015$TRAV_SP)-4)
sample2015$TRAV_SP<-as.numeric(as.character(sample2015$TRAV_SP))
```

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

```
nomissing<-sample2015[!is.na(sample2015$TRAV_SP),]
mean(nomissing$TRAV_SP)
```

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

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

```
noinjury<-nomissing[nomissing$INJ_SEV=='No Apparent Injury (O)',]
injury<-nomissing[nomissing$INJ_SEV!='No Apparent Injury (O)',]
mean(noinjury$TRAV_SP,na.rm=TRUE)
```

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

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

```
## [1] 53.25652
```

```
# Travel speed with injury is higher than travel speed without injury
```

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.

```
driver<-nomissing[nomissing$SEAT_POS=="Front Seat, Left Side",]
maledriver<-driver[driver$SEX=='Male',]
femaledriver<-driver[driver$SEX=='Female',]
mean(maledriver$TRAV_SP,na.rm=TRUE)
```

```
## [1] 51.65333
```

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

```
## [1] 45.57895
```

```
#Males in accidents tend to drive faster than females in accidents
```

13.Compare the average speed of drivers who drink and those who do not. Comment on the results.

```r
drink<-driver[driver$DRINKING=='Yes (Alcohol Involved)',]
nodrink<-driver[driver$DRINKING=='No (Alcohol Not Involved)',]
mean(drink$TRAV_SP,na.rm=TRUE)
```

```
## [1] 68.25
```

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

```
## [1] 44.94074
```

```r
#People who drink tended to drive faster than people who did not drink
```

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 am hypothesizing that drivers under the age of 30 drive more aggressively, meaning they drive faster
young<-driver[driver$AGE<30,]
old<-driver[driver$AGE>=30,]
mean(young$TRAV_SP,na.rm=TRUE)
```

```
## [1] 54.32787
```

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

```
## [1] 48.16438
```

```r
#These results show that people aged under 30 tended to drive faster than people aged 30 or older
```

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
# Number 14 did confirm my hypothesis, so I do not need to complete this step.
```