CMTH642 - Assignment 1

Mohammed Amir

October 16, 2016

1. Read the csv files in the folder

```
macro <- read.csv(file="D:/Big Data/CMTH642 - DATA ANALYTICS ADVANCED METHODS/ASSIGNMENT 1/USDA_Macronu
micro <- read.csv(file="D:/Big Data/CMTH642 - DATA ANALYTICS ADVANCED METHODS/ASSIGNMENT 1/USDA_Micronu
```

2. Merge the data frames using the variable "ID". Name the Merged Data Frame "USDA"

```
USDA <- merge (macro, micro, by="ID")
```

3. Prepare the dataset for analysis

```
# ----- Check data set structure
str(USDA)
```

```
## 'data.frame':    7057 obs. of  15 variables:
##  $ ID          : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ Description : Factor w/ 7053 levels "ABALONE,MIXED SPECIES,RAW",..: 1302 1301 1297 2302 2303 2304
##  $ Calories    : int  717 717 876 353 371 334 300 376 403 387 ...
##  $ Protein     : num  0.85 0.85 0.28 21.4 23.24 ...
##  $ TotalFat    : num  81.1 81.1 99.5 28.7 29.7 ...
##  $ Carbohydrate: num  0.06 0.06 0 2.34 2.79 0.45 0.46 3.06 1.28 4.78 ...
##  $ Sodium      : Factor w/ 1197 levels "","0","1","1,000",..: 972 1069 371 194 819 889 1084 946 882 9
##  $ Cholesterol : int  215 219 256 75 94 100 72 93 105 103 ...
##  $ Sugar       : num  0.06 0.06 0 0.5 0.51 0.45 0.46 NA 0.52 NA ...
##  $ Calcium     : int  24 24 4 528 674 184 388 673 721 643 ...
##  $ Iron        : num  0.02 0.16 0 0.31 0.43 0.5 0.33 0.64 0.68 0.21 ...
##  $ Potassium   : Factor w/ 886 levels "","0","1","1,000",..: 313 335 608 331 168 186 225 863 876 868
##  $ VitaminC    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VitaminE    : num  2.32 2.32 2.8 0.25 0.26 0.24 0.21 NA 0.29 NA ...
##  $ VitaminD    : num  1.5 1.5 1.8 0.5 0.5 0.5 0.4 NA 0.6 NA ...
```

```
# ----- Check for missing data
# is.na(USDA)

# ----- Check head
head(USDA)
```

```
##     ID               Description Calories Protein TotalFat Carbohydrate
## 1 1001          BUTTER,WITH SALT      717    0.85    81.11         0.06
## 2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06
## 3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00
## 4 1004               CHEESE,BLUE      353   21.40    28.74         2.34
## 5 1005              CHEESE,BRICK      371   23.24    29.68         2.79
## 6 1006               CHEESE,BRIE      334   20.75    27.68         0.45
##   Sodium Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE
## 1    714         215  0.06      24 0.02        24        0     2.32
## 2    827         219  0.06      24 0.16        26        0     2.32
## 3      2         256  0.00       4 0.00         5        0     2.80
## 4  1,395          75  0.50     528 0.31       256        0     0.25
## 5    560          94  0.51     674 0.43       136        0     0.26
## 6    629         100  0.45     184 0.50       152        0     0.24
##   VitaminD
## 1      1.5
## 2      1.5
## 3      1.8
## 4      0.5
## 5      0.5
## 6      0.5
```

```
# ----- Check column name
colnames(USDA)
```

```
##  [1] "ID"           "Description"  "Calories"     "Protein"
##  [5] "TotalFat"     "Carbohydrate" "Sodium"       "Cholesterol"
##  [9] "Sugar"        "Calcium"      "Iron"         "Potassium"
## [13] "VitaminC"     "VitaminE"     "VitaminD"
```

```
# ----- Check data set summary
summary(USDA)
```

```
##        ID
##  Min.   : 1001
##  1st Qu.: 8387
##  Median :13293
##  Mean   :14258
##  3rd Qu.:18336
##  Max.   :93600
##
##                                                        Description
##  BEEF,CHUCK,UNDER BLADE CNTR STEAK,BNLESS,DENVER CUT,LN,0" FA:   2
##  CAMPBELL,CAMPBELL'S SEL MICROWAVEABLE BOWLS,HEA             :   2
##  OIL,INDUSTRIAL,PALM KERNEL (HYDROGENATED),CONFECTION FAT    :   2
##  POPCORN,OIL-POPPED,LOFAT                                    :   2
##  ABALONE,MIXED SPECIES,RAW                                   :   1
##  ABALONE,MXD SP,CKD,FRIED                                    :   1
##  (Other)                                                     :7047
##     Calories        Protein         TotalFat       Carbohydrate
##  Min.   :  0.0   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 85.0   1st Qu.: 2.29   1st Qu.:  0.72   1st Qu.:  0.00
##  Median :181.0   Median : 8.20   Median :  4.37   Median :  7.13
##  Mean   :219.7   Mean   :11.71   Mean   : 10.32   Mean   : 20.70
##  3rd Qu.:331.0   3rd Qu.:20.43   3rd Qu.: 12.70   3rd Qu.: 28.17
##  Max.   :902.0   Max.   :88.32   Max.   :100.00   Max.   :100.00
##
##      Sodium      Cholesterol          Sugar           Calcium
##  2      : 174   Min.   :   0.00   Min.   : 0.000   Min.   :   0.00
##  0      : 148   1st Qu.:   0.00   1st Qu.: 0.000   1st Qu.:   9.00
##  1      : 144   Median :   3.00   Median : 1.395   Median :  19.00
##  4      : 144   Mean   :  41.55   Mean   : 8.257   Mean   :  73.53
##  3      : 131   3rd Qu.:  69.00   3rd Qu.: 7.875   3rd Qu.:  56.00
##  5      : 117   Max.   :3100.00   Max.   :99.800   Max.   :7364.00
##  (Other):6199   NA's   :287       NA's   :1909     NA's   :135
##       Iron           Potassium       VitaminC           VitaminE
##  Min.   :  0.000          : 408   Min.   :   0.000   Min.   : 0.000
##  1st Qu.:  0.520   0      : 127   1st Qu.:   0.000   1st Qu.: 0.120
##  Median :  1.330   340    :  29   Median :   0.000   Median : 0.270
##  Mean   :  2.828   237    :  28   Mean   :   9.436   Mean   : 1.488
##  3rd Qu.:  2.620   262     :  28   3rd Qu.:   3.100   3rd Qu.:  0.710
##  Max.   :123.600   284     :  27   Max.   :2400.000   Max.   :149.400
##  NA's   :122       (Other):6410   NA's   :331        NA's   :2719
##     VitaminD
##  Min.   :  0.0000
##  1st Qu.:  0.0000
##  Median :  0.0000
##  Mean   :  0.5769
##  3rd Qu.:  0.1000
##  Max.   :250.0000
##  NA's   :2833
```

```
# ----- Change Sodium & Potassium from factor to numeric
USDA$Sodium <- as.numeric(USDA$Sodium)
USDA$Potassium <- as.numeric(USDA$Potassium)
```

4. Remove records with missing values in 4 or more vectors

```
USDA <- USDA[rowSums(is.na(USDA)) < 4, ]
```

5. How many records remain in the data frame?

```
rowCount <- nrow(USDA)
rowCount
```

```
## [1] 6757
```

6. For records with missing values for Sugar, Vitamin E and Vitamin D, replace missing values with mean value for the respective vector

```
Sugar    <- replace(USDA$Sugar,which(is.na(USDA$Sugar)),mean(USDA$Sugar, na.rm = TRUE))
VitaminC <- replace(USDA$VitaminC,which(is.na(USDA$VitaminC)),mean(USDA$VitaminC, na.rm = TRUE))
VitaminD <- replace(USDA$VitaminD,which(is.na(USDA$VitaminD)),mean(USDA$VitaminD, na.rm = TRUE))
VitaminE <- replace(USDA$VitaminE,which(is.na(USDA$VitaminE)),mean(USDA$VitaminE, na.rm = TRUE))

USDA <- data.frame(ID=USDA$ID,Description=USDA$Description,Calories=USDA$Calories, Protein=USDA$Protein
```

7. With a single line of code, remove all remaining records with missing values. Name the new Data Frame "USDAclean"

```
USDAClean <- na.omit(USDA)
```

8. How many records remain in the data frame?

```
USDAClean_Count <- nrow(USDAClean)
USDAClean_Count
```

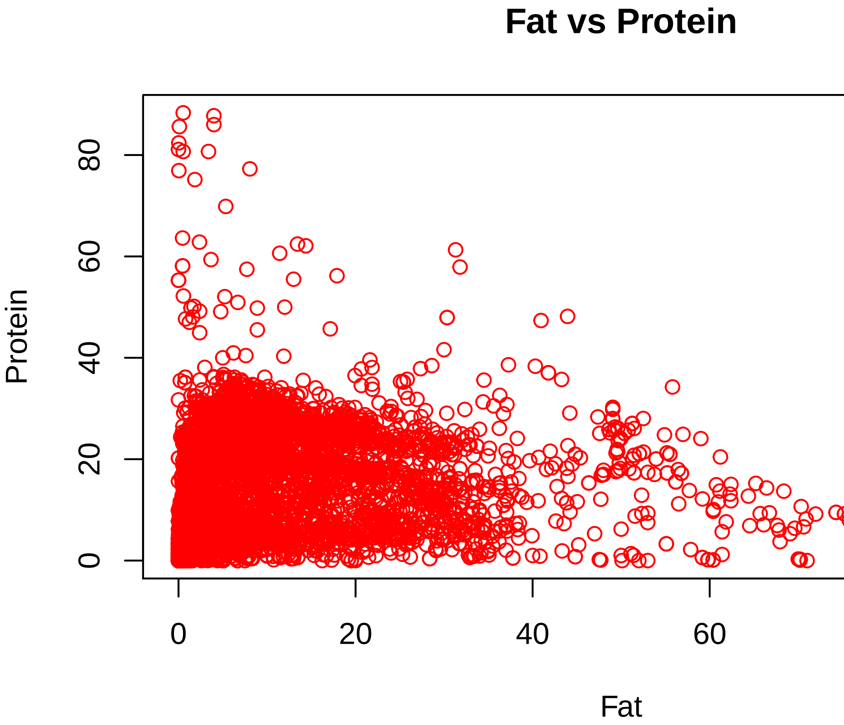
```
## [1] 6613
```

9. Which food has the highest sodium level?

```
Highest_Sodium <- USDAClean[which.max(USDAClean$Sodium),]
Highest_Sodium
```

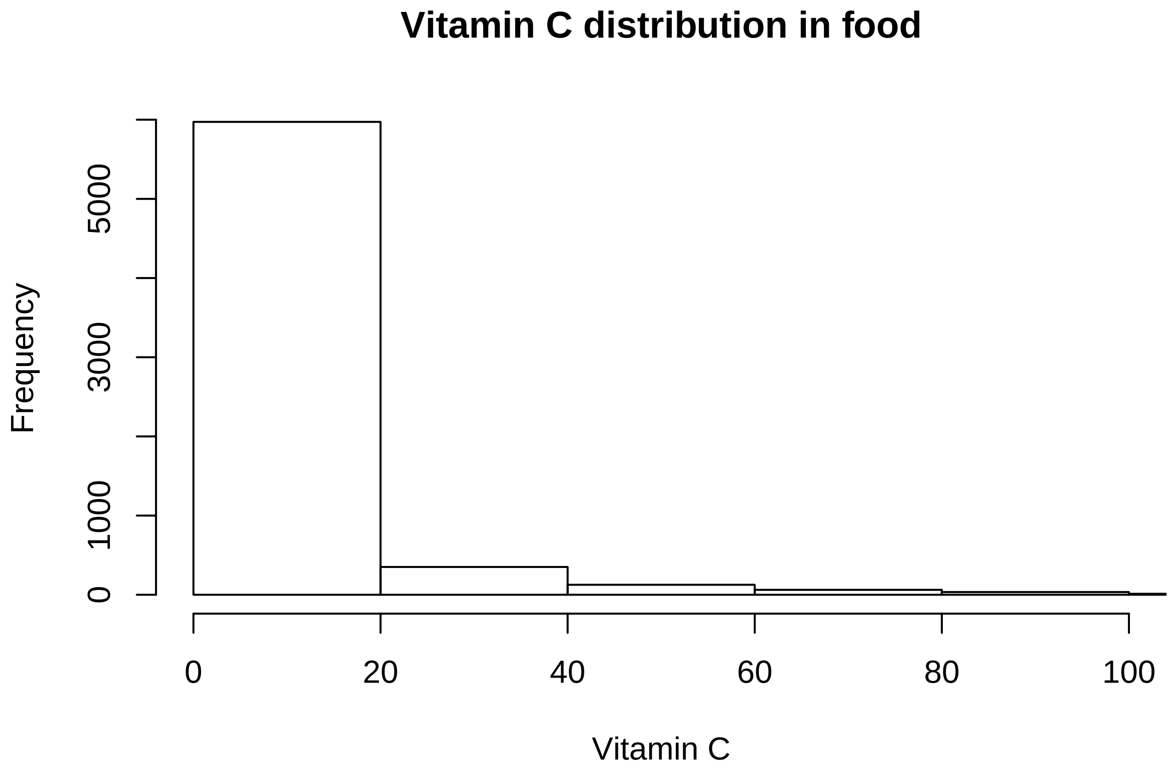
```
##         ID                                   Description Calories
## 4933 18014 BISCUIT, PLN OR BUTMLK, REFRI DOUGH, HIGHER FAT      322
##      Protein TotalFat Carbohydrate Sodium Cholesterol Calcium Iron
## 4933    6.66    13.63        43.27   1197           1      51 2.48
##      Potassium Sugar VitaminC  VitaminD VitaminE
## 4933       198   7.4        0 0.5771909     0.69
```

10. Create a scatter plot using Protein and Fat, with the plot title “Fat vs Protein”, labeling the axes “Fat” and “Protein”, and making the data points red



11. Create a histogram of Vitamin C distribution in foods, with a limit of 0 to 100 on the x-axis and breaks of 100

```
hist(USDAClean$VitaminC, breaks = 100, xlim=c(0,100), main="Vitamin C distribution in food", xlab="Vita
```



12. Add a new variable to the data frame that takes value 1 if the food has higher sodium than average, 0 otherwise. Call this variable HighSodium

```
#   ------ High Sodium
USDAClean$HighSodium <- ifelse(USDAClean$Sodium > mean(USDAClean$Sodium),1,0)
str(USDAClean)
```

```
## 'data.frame':    6613 obs. of  16 variables:
##  $ ID          : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ Description : Factor w/ 7053 levels "ABALONE,MIXED SPECIES,RAW",..: 1302 1301 1297 2302 2303 2304
##  $ Calories    : int  717 717 876 353 371 334 300 376 403 387 ...
##  $ Protein     : num  0.85 0.85 0.28 21.4 23.24 ...
##  $ TotalFat    : num  81.1 81.1 99.5 28.7 29.7 ...
##  $ Carbohydrate: num  0.06 0.06 0 2.34 2.79 0.45 0.46 3.06 1.28 4.78 ...
##  $ Sodium      : num  972 1069 371 194 819 ...
##  $ Cholesterol : int  215 219 256 75 94 100 72 93 105 103 ...
##  $ Calcium     : int  24 24 4 528 674 184 388 673 721 643 ...
##  $ Iron        : num  0.02 0.16 0 0.31 0.43 0.5 0.33 0.64 0.68 0.21 ...
##  $ Potassium   : num  313 335 608 331 168 186 225 863 876 868 ...
##  $ Sugar       : num  0.06 0.06 0 0.5 0.51 ...
##  $ VitaminC    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VitaminD    : num  1.5 1.5 1.8 0.5 0.5 ...
##  $ VitaminE    : num  2.32 2.32 2.8 0.25 0.26 ...
##  $ HighSodium  : num  1 1 0 0 1 1 1 1 1 1 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:144] 278 279 280 353 443 916 979 980 1021 1023 ..
##   .. ..- attr(*, "names")= chr [1:144] "278" "279" "280" "353" ...
```

13. Do the same for HighCalories, HighProtein, HighSugar, and HighFat

```r
#   ------ High Calories
USDAClean$HighCalories <- ifelse(USDAClean$Calories > mean(USDAClean$Calories),1,0)

#   ------ High Protein
USDAClean$HighProtein <- ifelse(USDAClean$Protein > mean(USDAClean$Protein),1,0)

#   ------ High Sugar
USDAClean$HighSugar <- ifelse(USDAClean$Sugar > mean(USDAClean$Sugar),1,0)

#   ------ High Fat
USDAClean$HighTotalFat <- ifelse(USDAClean$TotalFat > mean(USDAClean$TotalFat),1,0)
```

14. How many foods have both high sodium and high fat?

```r
High_Sodium_TotalFat <- USDAClean[USDAClean$HighSodium == 1,]
High_Sodium_TotalFat  <- High_Sodium_TotalFat[High_Sodium_TotalFat$HighTotalFat == 1,]
# High_Sodium_TotalFat
```

15. Calculate the average amount of iron by high and low protein (i.e. average amount of iron in foods with high protein and average amount of iron in foods with low protein)

```r
#    -- Average Iron for High protein
AverageIron_HighProtein <- USDAClean[USDAClean$HighProtein == 1,]
AverageIron_HighProtein <- mean(AverageIron_HighProtein$Iron)
AverageIron_HighProtein
```

```
## [1] 3.087864
```

```r
#    -- Average Iron for low protein
AverageIron_LowProtein <- USDAClean[USDAClean$HighProtein == 0,]
AverageIron_LowProtein <- mean(AverageIron_LowProtein$Iron)
AverageIron_LowProtein
```

```
## [1] 2.572456
```