CMTH 642 Data Analytics: Advanced Methods Assignment 1 - Derick Tung

1. Read the csv files in the folder. (4 points)

```
USDA_Macro = read.csv("C:\\Users\\Derick\\Downloads\\USDA_Macronutrients.csv")
USDA_Micro = read.csv("C:\\Users\\Derick\\Downloads\\USDA_Micronutrients.csv")
head(USDA_Macro)
```

```
##     ID                                        Description Calories Protein
## 1 2047                                         SALT,TABLE        0       0
## 2 2048                                    VINEGAR,CIDER       21       0
## 3 2053                                VINEGAR,DISTILLED       18       0
## 4 2073          CAMPBELL SOUP CO,PACE,DRY TACO SEAS MIX      188       0
## 5 6597 CAMPBELL SOUP COMPANY,PACE,CHIPOTLE CHUNKY SALSA       25       0
## 6 6598 CAMPBELL SOUP COMPANY,PACE,CILANTRO CHUNKY SALSA       25       0
##   TotalFat Carbohydrate
## 1        0         0.00
## 2        0         0.93
## 3        0         0.04
## 4        0        56.29
## 5        0         6.25
## 6        0         6.25
```

```
head(USDA_Micro)
```

```
##      ID Sodium Cholesterol Sugar Calcium  Iron Potassium VitaminC VitaminE
## 1  4038      0           0  0.00       0  0.00         0      0.0   149.40
## 2  8504    813          NA 17.17      45 67.67       630    239.7    80.46
## 3 25021    386           0 16.90     886 14.20       412     68.0    64.25
## 4  8590    242           0 14.30      47  8.70       296     89.0    58.96
## 5  4532      0           0  0.00       0  0.00         0      0.0    47.20
## 6  8568    251           0 28.00     233  4.20       721     70.0    46.90
##   VitaminD
## 1      0.0
## 2       NA
## 3      3.1
## 4      0.0
## 5       NA
## 6       NA
```

2. Merge the data frames using the variable "ID". Name the Merged Data Frame "USDA". (4 points)

```
USDA = merge(USDA_Macro, USDA_Micro, by = "ID")
head(USDA)
```

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

3. Check the datatypes of the attributes. Delete the commas in the Sodium and Potasium records. Assign Sodium and Potasium as numeric data types. (6 points)

```
sapply(USDA, class)
```

```
##          ID  Description     Calories      Protein     TotalFat Carbohydrate
##    "integer"     "factor"    "integer"    "numeric"    "numeric"    "numeric"
##       Sodium  Cholesterol        Sugar      Calcium         Iron    Potassium
##     "factor"    "integer"    "numeric"    "integer"    "numeric"     "factor"
##     VitaminC     VitaminE     VitaminD
##    "numeric"    "numeric"    "numeric"
```

```
USDA$Potassium = gsub(",", "", USDA$Potassium)
USDA$Sodium = gsub(",", "", USDA$Sodium)
USDA$Potassium = as.numeric(USDA$Potassium)
USDA$Sodium = as.numeric(USDA$Sodium)

sapply(USDA, class)
```

```
##          ID  Description     Calories      Protein     TotalFat Carbohydrate
##    "integer"     "factor"    "integer"    "numeric"    "numeric"    "numeric"
##       Sodium  Cholesterol        Sugar      Calcium         Iron    Potassium
##    "numeric"    "integer"    "numeric"    "integer"    "numeric"    "numeric"
##     VitaminC     VitaminE     VitaminD
##    "numeric"    "numeric"    "numeric"
```

4. Remove records (rows) with missing values in more than 4 attributes (columns). How many records remain in the data frame? (6 points)

```
USDA2 = USDA[rowSums(is.na(USDA)) < 5, ]

head(USDA2)
```

```
##     ID              Description Calories Protein TotalFat Carbohydrate Sodium
## 1 1001         BUTTER,WITH SALT      717    0.85    81.11         0.06    714
## 2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06    827
## 3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00      2
## 4 1004             CHEESE,BLUE      353   21.40    28.74         2.34   1395
## 5 1005            CHEESE,BRICK      371   23.24    29.68         2.79    560
## 6 1006             CHEESE,BRIE      334   20.75    27.68         0.45    629
##   Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD
## 1         215  0.06      24 0.02        24        0     2.32      1.5
## 2         219  0.06      24 0.16        26        0     2.32      1.5
## 3         256  0.00       4 0.00         5        0     2.80      1.8
## 4          75  0.50     528 0.31       256        0     0.25      0.5
## 5          94  0.51     674 0.43       136        0     0.26      0.5
## 6         100  0.45     184 0.50       152        0     0.24      0.5
```

```
"6887 records remain in the data frame."
```

```
## [1] "6887 records remain in the data frame."
```

5. For records with missing values for Sugar, Vitamin E and Vitamin D, replace missing values with mean value for the respective variable. (6 points)

```
sugarmean = mean(USDA2$Sugar, na.rm = TRUE)
USDA2$Sugar[is.na(USDA2$Sugar)] = sugarmean

vitemean = mean(USDA2$VitaminE, na.rm = TRUE)
USDA2$VitaminE[is.na(USDA2$VitaminE)] = vitemean

vitdmean = mean(USDA2$VitaminD, na.rm = TRUE)
USDA2$VitaminD[is.na(USDA2$VitaminD)] = vitdmean

head(USDA2)
```

```
##     ID              Description Calories Protein TotalFat Carbohydrate Sodium
## 1 1001         BUTTER,WITH SALT      717    0.85    81.11         0.06    714
## 2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06    827
## 3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00      2
## 4 1004             CHEESE,BLUE      353   21.40    28.74         2.34   1395
## 5 1005            CHEESE,BRICK      371   23.24    29.68         2.79    560
## 6 1006             CHEESE,BRIE      334   20.75    27.68         0.45    629
##   Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD
## 1         215  0.06      24 0.02        24        0     2.32      1.5
## 2         219  0.06      24 0.16        26        0     2.32      1.5
```

```
## 3         256  0.00       4 0.00         5        0     2.80      1.8
## 4          75  0.50     528 0.31       256        0     0.25      0.5
## 5          94  0.51     674 0.43       136        0     0.26      0.5
## 6         100  0.45     184 0.50       152        0     0.24      0.5
```

6. With a single line of code, remove all remaining records with missing values. Name the new Data Frame "USDAclean". How many records remain in the data frame? (6 points)

```
USDAclean = USDA2[complete.cases(USDA2), ]
"6310 records remain in USDAclean."
```

```
## [1] "6310 records remain in USDAclean."
```

```
head(USDAclean)
```

```
##     ID              Description Calories Protein TotalFat Carbohydrate Sodium
## 1 1001         BUTTER,WITH SALT      717    0.85    81.11         0.06    714
## 2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06    827
## 3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00      2
## 4 1004             CHEESE,BLUE      353   21.40    28.74         2.34   1395
## 5 1005            CHEESE,BRICK      371   23.24    29.68         2.79    560
## 6 1006             CHEESE,BRIE      334   20.75    27.68         0.45    629
##   Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD
## 1         215  0.06      24 0.02        24        0     2.32      1.5
## 2         219  0.06      24 0.16        26        0     2.32      1.5
## 3         256  0.00       4 0.00         5        0     2.80      1.8
## 4          75  0.50     528 0.31       256        0     0.25      0.5
## 5          94  0.51     674 0.43       136        0     0.26      0.5
## 6         100  0.45     184 0.50       152        0     0.24      0.5
```

7. Which food has the highest sodium level? (6 points)

```
USDAclean$Description[which.max(USDAclean$Sodium)]
```

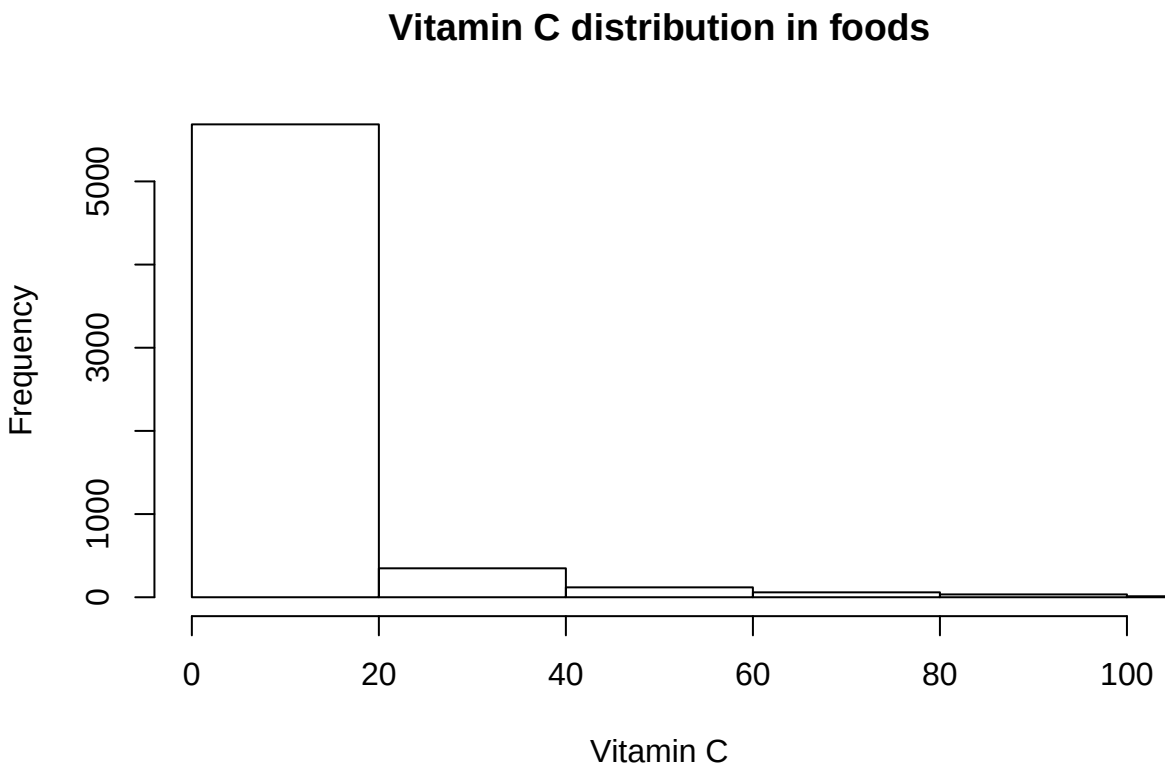
```
## [1] SALT,TABLE
## 7053 Levels: ABALONE,MIXED SPECIES,RAW ABALONE,MXD SP,CKD,FRIED ... ZWIEBACK
```

```
"Table Salt has the highest sodium level"
```

```
## [1] "Table Salt has the highest sodium level"
```

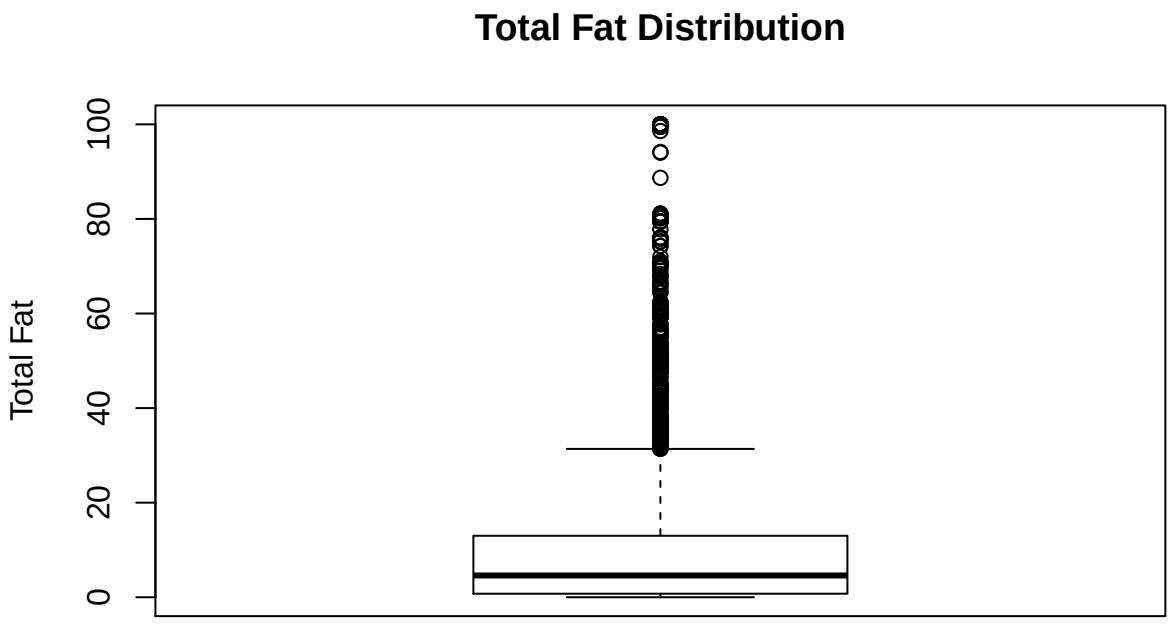
8. Create a histogram of Vitamin C distribution in foods, with a limit of 0 to 100 on the x-axis and breaks of 100. (6 points)

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

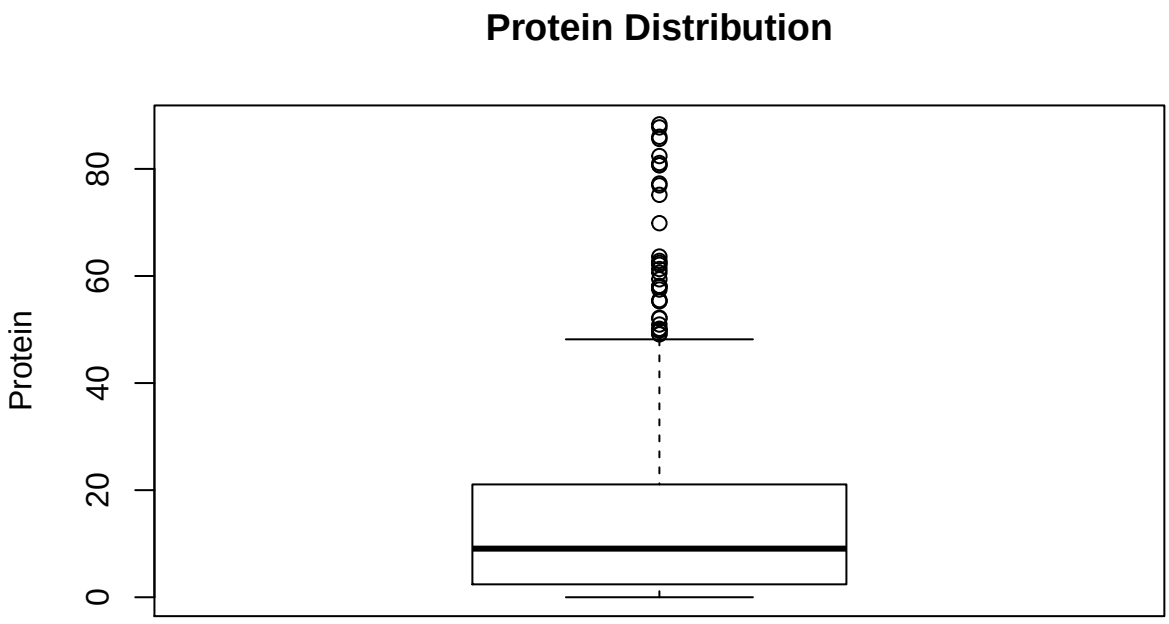


9. Create a boxplot to illustrate the distribution of values for TotalFat, Protein and Carbohydrate. (6 points)

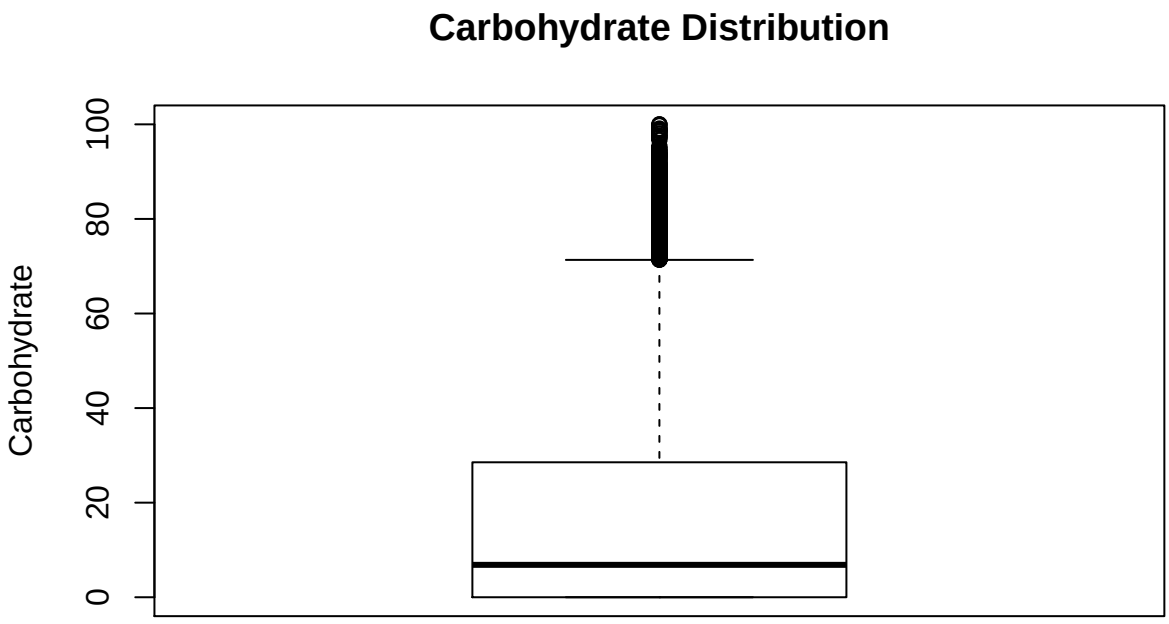
```
boxplot(USDAclean$TotalFat, main = "Total Fat Distribution", ylab = "Total Fat")
```




```
boxplot(USDAclean$Protein, main = "Protein Distribution", ylab = "Protein")
```

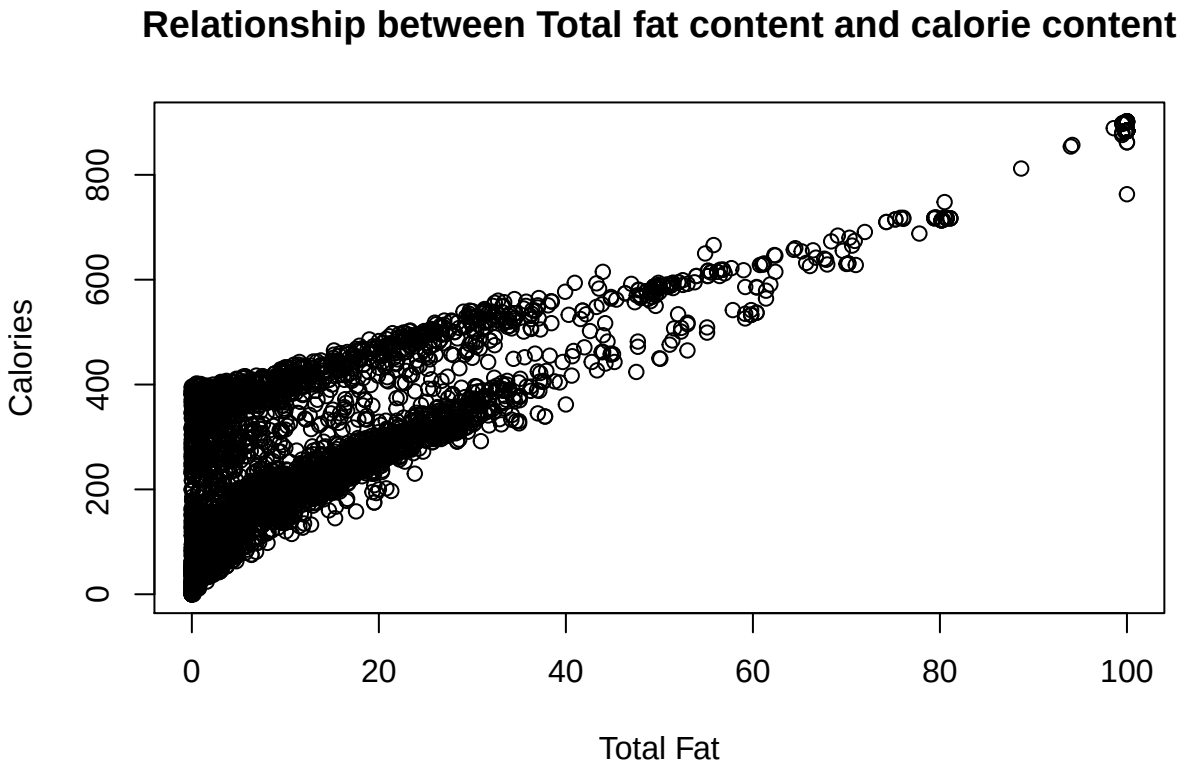



```
boxplot(USDAclean$Carbohydrate, main = "Carbohydrate Distribution", ylab = "Carbohydrate")
```



10. Create a scatterplot to illustrate the relationship between a food's TotalFat content and its calorie content. (6 points)

```
plot(USDAclean$TotalFat, USDAclean$Calories, main = "Relationship between Total fat content and calorie
```



11. Add a variable to the data frame that takes value 1 if the food has higher sodium than average, 0 otherwise.Call this variable HighSodium. Do the same for High Calories, High Protein, High Sugar, and High Fat. How many foods have both high sodium and high fat? (8 points)

```
USDAclean["HighSodium"] = ifelse(mean(USDAclean$Sodium) < USDAclean$Sodium, 1, 0)
USDAclean["HighCalories"] = ifelse(mean(USDAclean$Calories) < USDAclean$Calories, 1, 0)
USDAclean["HighProtein"] = ifelse(mean(USDAclean$Protein) < USDAclean$Protein, 1, 0)
USDAclean["HighSugar"] = ifelse(mean(USDAclean$Sugar) < USDAclean$Sugar, 1, 0)
USDAclean["HighFat"] = ifelse(mean(USDAclean$TotalFat) < USDAclean$TotalFat, 1, 0)

head(USDAclean)
```

```
##     ID              Description Calories Protein TotalFat Carbohydrate Sodium
## 1 1001         BUTTER,WITH SALT      717    0.85    81.11         0.06    714
## 2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06    827
## 3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00      2
## 4 1004             CHEESE,BLUE      353   21.40    28.74         2.34   1395
## 5 1005            CHEESE,BRICK      371   23.24    29.68         2.79    560
## 6 1006             CHEESE,BRIE      334   20.75    27.68         0.45    629
##   Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD
## 1         215  0.06      24 0.02        24        0     2.32      1.5
## 2         219  0.06      24 0.16        26        0     2.32      1.5
## 3         256  0.00       4 0.00         5        0     2.80      1.8
```

```
## 4          75  0.50     528 0.31       256        0     0.25      0.5
## 5          94  0.51     674 0.43       136        0     0.26      0.5
## 6         100  0.45     184 0.50       152        0     0.24      0.5
##   HighSodium HighCalories HighProtein HighSugar HighFat
## 1          1            1           0         0       1
## 2          1            1           0         0       1
## 3          0            1           0         0       1
## 4          1            1           1         0       1
## 5          1            1           1         0       1
## 6          1            1           1         0       1
```

```
table(USDAclean$HighSodium, USDAclean$HighFat)
```

```
##
##        0    1
##   0 3233 1250
##   1 1183  644
```

```
"The number of foods with high sodium and high fat is 644."
```

```
## [1] "The number of foods with high sodium and high fat is 644."
```

12. Calculate the average amount of iron, sorted by high and low protein. (8 points)

```
tapply(USDAclean$Iron, USDAclean$HighProtein, mean)
```

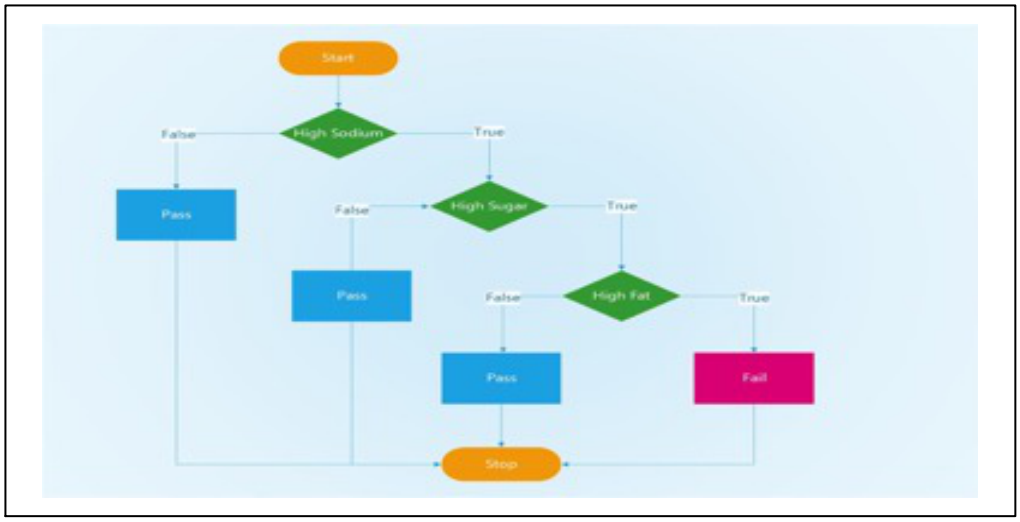
```
##        0        1
## 2.696634 3.069541
```

13. Create a script for a "HealthCheck" program to detect unhealthy foods. Use the algorithm flowchart below as a basis for this script. (8 points)

```
require(jpeg)
```

```
## Loading required package: jpeg
```

```
img <- readJPEG("HealthCheck.jpg")
plot(1:4, ty = 'n', ann = F, xaxt = 'n', yaxt = 'n')
rasterImage(img,1,1,4,4)
```




```
HealthCheck = function(x,y,z){
  ifelse(x==1,
         ifelse(y==1,
                ifelse(z==1, "Fail", "Pass"), "Pass"), "Pass")
}
```

14. Add a new variable called HealthCheck to the data frame using the output of the function. (8 points)

```
USDAclean["HealthCheck"] = HealthCheck(USDAclean$HighSodium, USDAclean$HighSugar, USDAclean$HighFat)
```

15. How many foods in the USDAclean data frame fail the HealthCheck? (8 points)

```
x = sum(USDAclean$HealthCheck == "Fail")
x
```

```
## [1] 237
```

16. Save your final data frame as "USDAclean_ [your last name]" (4 points)

```r
write.csv(USDAclean, "USDAclean_Tung")
```