CMTH 642 Data Analytics: Advanced Methods Assignment 1

Paul Ycay

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

```
Macro_Data<-read.csv(file="C:\\Users\\Paul\\Desktop\\USDA_Macronutrients.csv",header=T,sep=",")
Micro_Data<-read.csv(file="C:\\Users\\Paul\\Desktop\\USDA_Micronutrients.csv",header=T,sep=",")
head(Macro_Data)
```

```
##     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(Micro_Data)
```

```
##      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(Macro_Data,Micro_Data,by="ID")
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
```

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)

#{r, eval=F, echo=T} #use this piece of code to only run the code and not output

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

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

```
USDA$Sodium<-gsub(',','',USDA$Sodium)
USDA$Potassium<-gsub(',','',USDA$Potassium)
USDA$Sodium<-as.numeric(USDA$Sodium)
USDA$Potassium<-as.numeric(USDA$Potassium)
sapply(USDA,class)
```

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

```
missingvalues=(rowSums(is.na(USDA)))
USDA=USDA[!missingvalues > 4,];
sprintf("These are the number of records remaining: %i ",nrow(USDA))
```

```
## [1] "These are the number of records remaining: 6887 "
```

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)

```
USDA$Sugar[is.na((USDA$Sugar))]<-mean(USDA$Sugar,na.rm = TRUE)
USDA$VitaminE[is.na((USDA$VitaminE))]<-mean(USDA$VitaminE,na.rm = TRUE)
USDA$VitaminD[is.na((USDA$VitaminD))]<-mean(USDA$VitaminD,na.rm = TRUE)



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   1395          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
```

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=USDA[complete.cases(USDA),]
str(USDAclean)
```

```
## 'data.frame':    6310 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      : num  714 827 2 1395 560 ...
##  $ Cholesterol : int  215 219 256 75 94 100 72 93 105 103 ...
##  $ Sugar       : num  0.06 0.06 0 0.5 0.51 ...
##  $ 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  24 26 5 256 136 152 187 93 98 95 ...
##  $ VitaminC    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VitaminE    : num  2.32 2.32 2.8 0.25 0.26 ...
##  $ VitaminD    : num  1.5 1.5 1.8 0.5 0.5 ...
```

```r
sprintf("Number of records remaining: %i", nrow(USDAclean))
```

```
## [1] "Number of records remaining: 6310"
```

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

```r
which.max(USDAclean$Sodium)
```

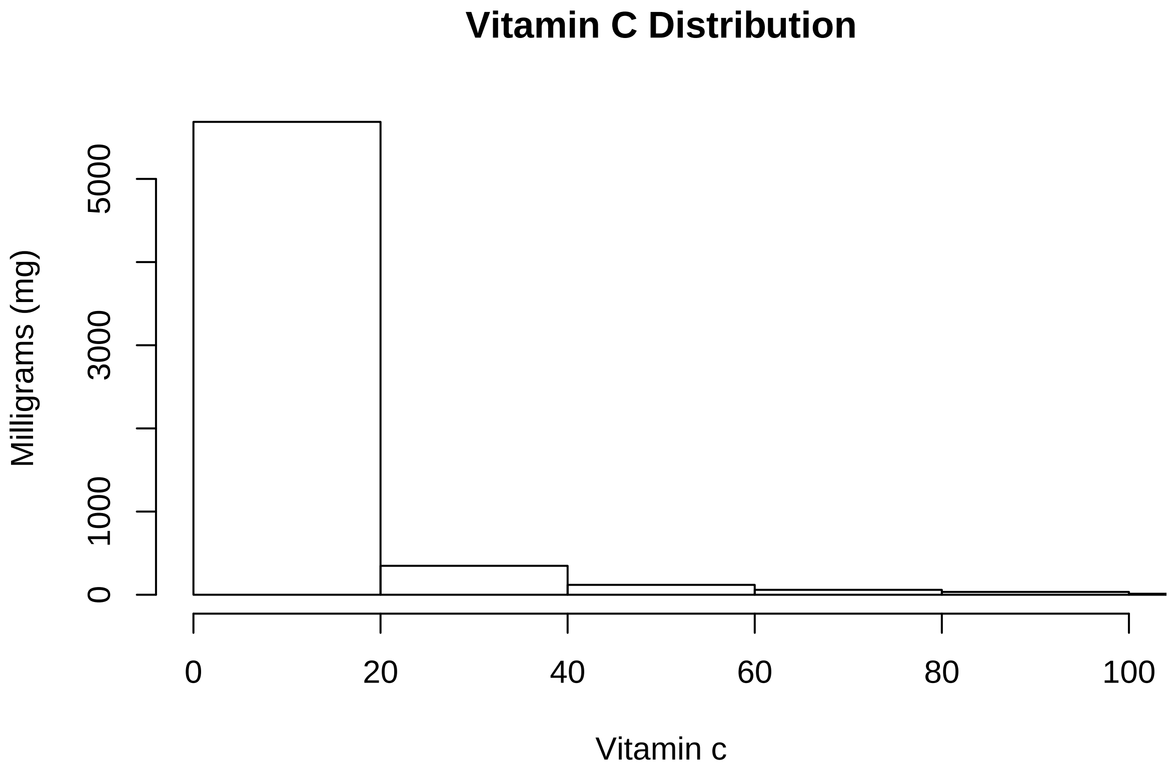
```
## [1] 262
```

```r
USDAclean$Description[265]
```

```
## [1] VANILLA EXTRACT
## 7053 Levels: ABALONE,MIXED SPECIES,RAW ... ZWIEBACK
```

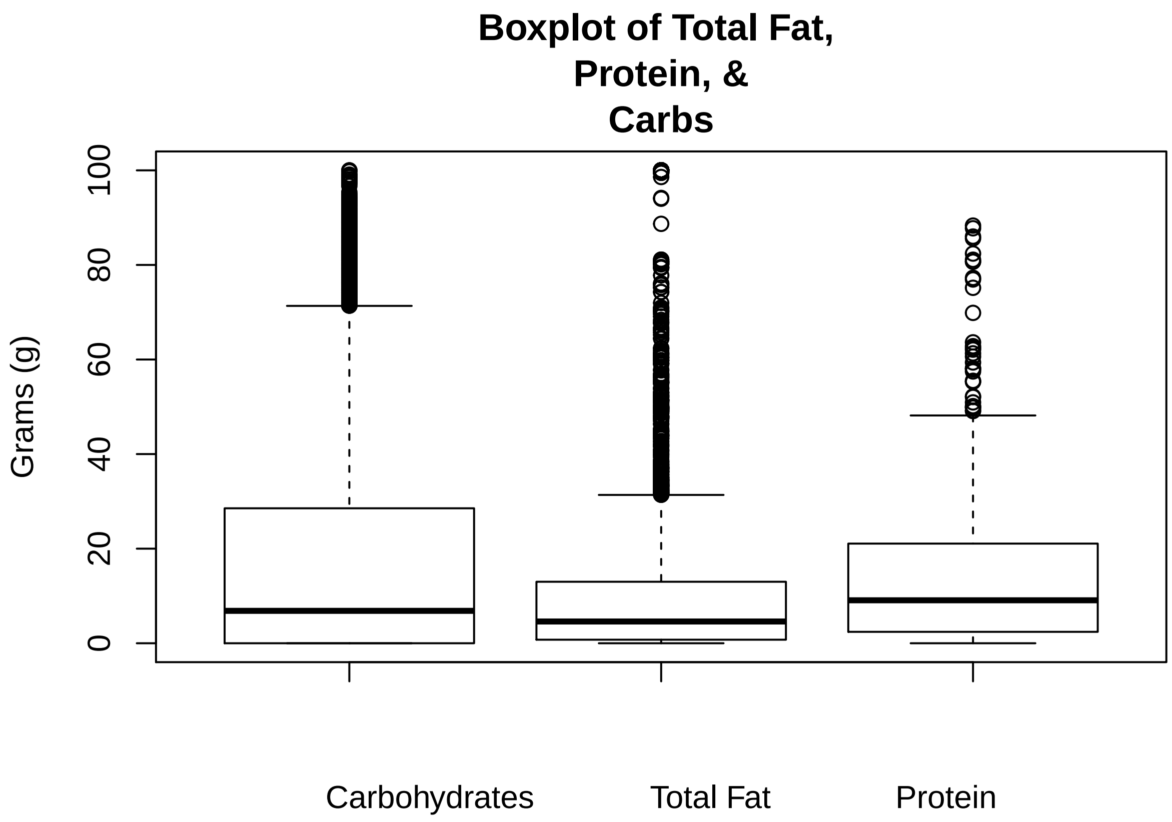
8. Create a histogram of Vitamin C distribution in foods, with a limit of 0 to 100 on the x-axis and breaks of 100. (6 points)

```r
hist(USDAclean$VitaminC, xlab="Vitamin c", ylab="Milligrams (mg)",
main= "Vitamin C Distribution", xlim=c(0,100), breaks=100)
```



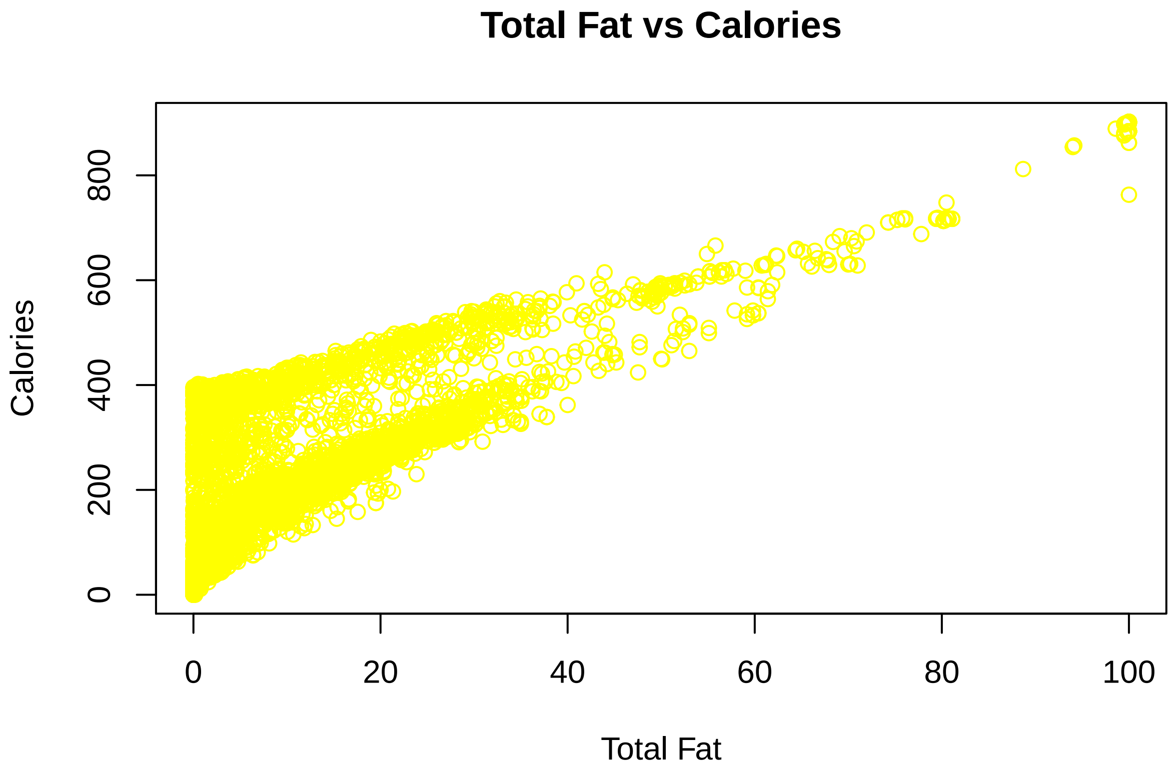
9. Create a boxplot to illustrate the distribution of values for TotalFat, Protein and Carbohydrate. (6 points)

```
boxplot(USDAclean$Carbohydrate, USDAclean$TotalFat, USDAclean$Protein,
main="Boxplot of Total Fat,
Protein, &
Carbs", ylab="Grams (g)",
xlab=("Carbohydrates             Total Fat              Protein"))
```



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,
xlab="Total Fat", ylab = "Calories",
main = "Total Fat vs Calories", col = "yellow")
```



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)

```
HighSodium = as.numeric(USDAclean$Sodium > mean(USDAclean$Sodium, na.rm=TRUE))
str(HighSodium)
```

```
## num [1:6310] 1 1 0 1 1 1 1 1 1 1 ...
```

```
HighCalories=as.numeric(USDAclean$Calories > mean(USDAclean$Calories,na.rm=TRUE))
str(HighCalories)
```

```
## num [1:6310] 1 1 1 1 1 1 1 1 1 1 ...
```

```
HighProtein = as.numeric(USDAclean$Protein > mean(USDAclean$Protein,na.rm=TRUE))
str(HighProtein)
```

```
## num [1:6310] 0 0 0 1 1 1 1 1 1 1 ...
```

```r
HighSugar = as.numeric(USDAclean$Sugar > mean(USDAclean$Sugar, na.rm=TRUE))
str(HighSugar)
```

```
##  num [1:6310] 0 0 0 0 0 0 0 1 0 1 ...
```

```r
HighFat = as.numeric(USDAclean$TotalFat > mean(USDAclean$TotalFat, na.rm=TRUE))

a<-table(HighSodium, HighFat);
highfs<-a[2,2];
paste0("Number of foods with high sodium and high fat: ", highfs)
```

```
## [1] "Number of foods with high sodium and high fat: 644"
```

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

```r
tapply(USDAclean$Iron, HighProtein, mean, na.rm=TRUE)
```

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

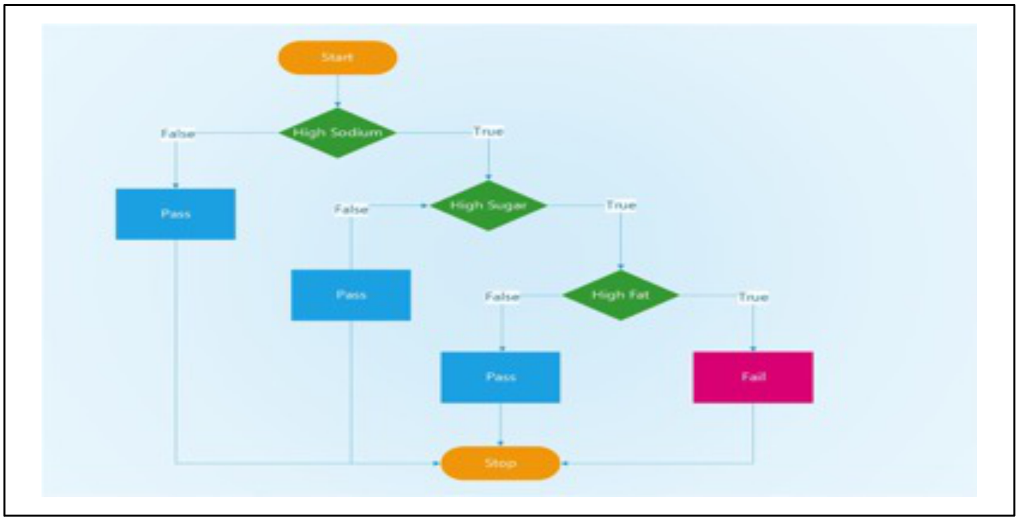
13. Create a script for a "HealthCheck" program to detect unhealthy foods. Use the algorithm flowchart below as a basis for this script. (8 points)

```r
require(jpeg)
```

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

```
## Warning: package 'jpeg' was built under R version 3.5.2
```

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



```r
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)

```r
USDAclean["HealthCheck"]<-healthcheck(HighSodium,HighSugar,HighFat)
head(USDAclean)
```

```
##     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   1395          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 HealthCheck
## 1      1.5        Pass
## 2      1.5        Pass
## 3      1.8        Pass
## 4      0.5        Pass
## 5      0.5        Pass
## 6      0.5        Pass
```

```r
tail(USDAclean)
```

```
##         ID                Description Calories Protein TotalFat
## 7052 48052         VITAL WHEAT GLUTEN      370   75.16     1.85
## 7053 80200              FROG LEGS,RAW       73   16.40     0.30
## 7054 83110            MACKEREL,SALTED      305   18.50    25.10
## 7055 90240 SCALLOP,(BAY&SEA),CKD,STMD      111   20.54     0.84
## 7056 90560                  SNAIL,RAW       90   16.10     1.40
## 7057 93600           TURTLE,GREEN,RAW       89   19.80     0.50
##      Carbohydrate Sodium Cholesterol Sugar Calcium Iron Potassium VitaminC
## 7052        13.79     29           0     0     142 5.20       100        0
## 7053         0.00     58          50     0      18 1.50       285        0
## 7054         0.00   4450          95     0      66 1.40       520        0
## 7055         5.41    667          41     0      10 0.58       314        0
## 7056         2.00     70          50     0      10 3.50       382        0
## 7057         0.00     68          50     0     118 1.40       230        0
##      VitaminE VitaminD HealthCheck
## 7052     0.00      0.0        Pass
## 7053     1.00      0.2        Pass
## 7054     2.38     25.2        Pass
## 7055     0.00      0.0        Pass
## 7056     5.00      0.0        Pass
## 7057     0.50      0.0        Pass
```

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

```r
nasty_foods<-sum(USDAclean$HealthCheck=="Fail", na.rm = TRUE)
paste0("Number of foods that fail the HealthCheck: ",nasty_foods)
```

```
## [1] "Number of foods that fail the HealthCheck: 237"
```

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

```r
USDAclean_Ycay<-USDAclean
head(USDAclean_Ycay)
```

```
##     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   1395          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 HealthCheck
## 1      1.5        Pass
## 2      1.5        Pass
## 3      1.8        Pass
## 4      0.5        Pass
## 5      0.5        Pass
## 6      0.5        Pass
```

```
tail(USDAclean_Ycay)
```

```
##         ID                Description Calories Protein TotalFat
## 7052 48052         VITAL WHEAT GLUTEN      370   75.16     1.85
## 7053 80200              FROG LEGS,RAW       73   16.40     0.30
## 7054 83110            MACKEREL,SALTED      305   18.50    25.10
## 7055 90240 SCALLOP,(BAY&SEA),CKD,STMD      111   20.54     0.84
## 7056 90560                  SNAIL,RAW       90   16.10     1.40
## 7057 93600           TURTLE,GREEN,RAW       89   19.80     0.50
##      Carbohydrate Sodium Cholesterol Sugar Calcium Iron Potassium VitaminC
## 7052        13.79     29           0     0     142 5.20       100        0
## 7053         0.00     58          50     0      18 1.50       285        0
## 7054         0.00   4450          95     0      66 1.40       520        0
## 7055         5.41    667          41     0      10 0.58       314        0
## 7056         2.00     70          50     0      10 3.50       382        0
## 7057         0.00     68          50     0     118 1.40       230        0
##      VitaminE VitaminD HealthCheck
## 7052     0.00      0.0        Pass
## 7053     1.00      0.2        Pass
## 7054     2.38     25.2        Pass
## 7055     0.00      0.0        Pass
## 7056     5.00      0.0        Pass
## 7057     0.50      0.0        Pass
```