CMTH 642 Data Analytics: Advanced Methods
Assignment 1

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

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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)

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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)

```
# INSERT YOUR ANSWER HERE
```

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)

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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)

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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)

```
# INSERT YOUR ANSWER HERE
```

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

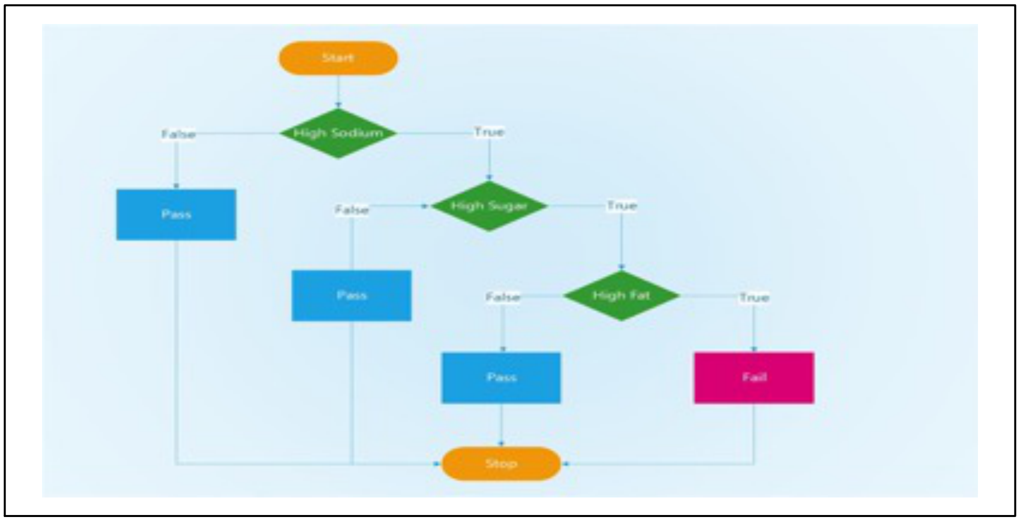
```
# INSERT YOUR ANSWER HERE
```

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)
```



```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```

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

```
# INSERT YOUR ANSWER HERE
```