CMTH 642: Advance Methods

Assignment 1

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

2. Merge the data frames using the variable "ID". Name the Merged Data Frame "USDA". (6 points)

3. Prepare the dataset for analysis. (6 points)

4. Remove records with missing values in 4 or more vectors. (6 points)

5. How many records remain in the data frame? (6 points)

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

7. With a single line of code, remove all remaining records with missing values. Name the new Data Frame "USDAclean". (6 points)

8. How many records remain in the data frame? (6 points)

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

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. (8 points)

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. (8 points)

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. (8 points)

13. Do the same for HighCalories, HighProtein, HighSugar, and HighFat. (8 points)

14. How many foods have both high sodium and high fat? (8 points)

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). (8 points)