Assignment 1

Harnoor Rangi

2017-07-09

CMTH 642: Advanced Methods Assignment 1

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

setwd("~/Downloads")  
library(readr)

## Warning: package 'readr' was built under R version 3.3.2

dat1<-read\_csv("USDA\_Micronutrients.csv")

## Parsed with column specification:  
## cols(  
## ID = col\_integer(),  
## Sodium = col\_number(),  
## Cholesterol = col\_integer(),  
## Sugar = col\_double(),  
## Calcium = col\_integer(),  
## Iron = col\_double(),  
## Potassium = col\_number(),  
## VitaminC = col\_double(),  
## VitaminE = col\_double(),  
## VitaminD = col\_double()  
## )

dat2<-read\_csv("USDA\_Macronutrients.csv")

## Parsed with column specification:  
## cols(  
## ID = col\_integer(),  
## Description = col\_character(),  
## Calories = col\_integer(),  
## Protein = col\_double(),  
## TotalFat = col\_double(),  
## Carbohydrate = col\_double()  
## )

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

USDA <- merge(dat2,dat1,by="ID")

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

summary(USDA)

## ID Description Calories Protein   
## Min. : 1001 Length:7057 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 8387 Class :character 1st Qu.: 85.0 1st Qu.: 2.29   
## Median :13293 Mode :character Median :181.0 Median : 8.20   
## Mean :14258 Mean :219.7 Mean :11.71   
## 3rd Qu.:18336 3rd Qu.:331.0 3rd Qu.:20.43   
## Max. :93600 Max. :902.0 Max. :88.32   
##   
## TotalFat Carbohydrate Sodium Cholesterol   
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 0.72 1st Qu.: 0.00 1st Qu.: 37.0 1st Qu.: 0.00   
## Median : 4.37 Median : 7.13 Median : 79.0 Median : 3.00   
## Mean : 10.32 Mean : 20.70 Mean : 322.1 Mean : 41.55   
## 3rd Qu.: 12.70 3rd Qu.: 28.17 3rd Qu.: 386.0 3rd Qu.: 69.00   
## Max. :100.00 Max. :100.00 Max. :38758.0 Max. :3100.00   
## NA's :83 NA's :287   
## Sugar Calcium Iron Potassium   
## Min. : 0.000 Min. : 0.00 Min. : 0.000 Min. : 0.0   
## 1st Qu.: 0.000 1st Qu.: 9.00 1st Qu.: 0.520 1st Qu.: 135.0   
## Median : 1.395 Median : 19.00 Median : 1.330 Median : 250.0   
## Mean : 8.257 Mean : 73.53 Mean : 2.828 Mean : 301.4   
## 3rd Qu.: 7.875 3rd Qu.: 56.00 3rd Qu.: 2.620 3rd Qu.: 348.0   
## Max. :99.800 Max. :7364.00 Max. :123.600 Max. :16500.0   
## NA's :1909 NA's :135 NA's :122 NA's :408   
## VitaminC VitaminE VitaminD   
## Min. : 0.000 Min. : 0.000 Min. : 0.0000   
## 1st Qu.: 0.000 1st Qu.: 0.120 1st Qu.: 0.0000   
## Median : 0.000 Median : 0.270 Median : 0.0000   
## Mean : 9.436 Mean : 1.488 Mean : 0.5769   
## 3rd Qu.: 3.100 3rd Qu.: 0.710 3rd Qu.: 0.1000   
## Max. :2400.000 Max. :149.400 Max. :250.0000   
## NA's :331 NA's :2719 NA's :2833

sapply(USDA,class)

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

str(USDA)

## 'data.frame': 7057 obs. of 15 variables:  
## $ ID : int 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...  
## $ Description : chr "BUTTER,WITH SALT" "BUTTER,WHIPPED,WITH SALT" "BUTTER OIL,ANHYDROUS" "CHEESE,BLUE" ...  
## $ 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 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 : 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 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 ...

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

# Removing duplicate values

desp1 <- as.data.frame(table(USDA$Description))  
duplicates <- subset(desp1,desp1$Freq >1)  
USDA\_clean<- aggregate(USDA[,-2],list(USDA$Description), mean, na.rm = TRUE)  
colnames(USDA\_clean)[1] <- "Description"  
summary(USDA\_clean)

## Description ID Calories Protein   
## Length:7053 Min. : 1001 Min. : 0.0 Min. : 0.00   
## Class :character 1st Qu.: 8387 1st Qu.: 85.0 1st Qu.: 2.29   
## Mode :character Median :13285 Median :181.0 Median : 8.20   
## Mean :14257 Mean :219.6 Mean :11.71   
## 3rd Qu.:18335 3rd Qu.:331.0 3rd Qu.:20.43   
## Max. :93600 Max. :902.0 Max. :88.32   
##   
## TotalFat Carbohydrate Sodium Cholesterol   
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 0.72 1st Qu.: 0.00 1st Qu.: 37.0 1st Qu.: 0.00   
## Median : 4.36 Median : 7.13 Median : 79.0 Median : 3.00   
## Mean : 10.31 Mean : 20.70 Mean : 322.1 Mean : 41.56   
## 3rd Qu.: 12.70 3rd Qu.: 28.17 3rd Qu.: 386.0 3rd Qu.: 69.00   
## Max. :100.00 Max. :100.00 Max. :38758.0 Max. :3100.00   
## NA's :83 NA's :287   
## Sugar Calcium Iron Potassium   
## Min. : 0.000 Min. : 0.00 Min. : 0.000 Min. : 0.0   
## 1st Qu.: 0.000 1st Qu.: 9.00 1st Qu.: 0.520 1st Qu.: 135.0   
## Median : 1.400 Median : 19.00 Median : 1.330 Median : 250.0   
## Mean : 8.263 Mean : 73.57 Mean : 2.829 Mean : 301.4   
## 3rd Qu.: 7.893 3rd Qu.: 56.00 3rd Qu.: 2.615 3rd Qu.: 349.0   
## Max. :99.800 Max. :7364.00 Max. :123.600 Max. :16500.0   
## NA's :1909 NA's :135 NA's :122 NA's :408   
## VitaminC VitaminE VitaminD   
## Min. : 0.000 Min. : 0.000 Min. : 0.0000   
## 1st Qu.: 0.000 1st Qu.: 0.120 1st Qu.: 0.0000   
## Median : 0.000 Median : 0.270 Median : 0.0000   
## Mean : 9.442 Mean : 1.488 Mean : 0.5772   
## 3rd Qu.: 3.100 3rd Qu.: 0.710 3rd Qu.: 0.1000   
## Max. :2400.000 Max. :149.400 Max. :250.0000   
## NA's :331 NA's :2718 NA's :2831

nrow(USDA)

## [1] 7057

nrow(USDA\_clean)

## [1] 7053

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

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

# 5. How many records remain in the data frame?

dim(USDA)

## [1] 7057 15

dim(USDAclean)

## [1] 6734 15

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

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

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

USDAclean1 <- na.omit(USDAclean)

# 8. How many records remain in the data frame?

dim(USDAclean1)

## [1] 6310 15

# 9. Which food has the highest sodium level?

USDAclean1[USDAclean1$Sodium == max(USDAclean1$Sodium), ]

## ID Description Calories Protein TotalFat Carbohydrate Sodium  
## 265 2047 SALT,TABLE 0 0 0 0 38758  
## Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD  
## 265 0 0 24 0.33 8 0 0 0

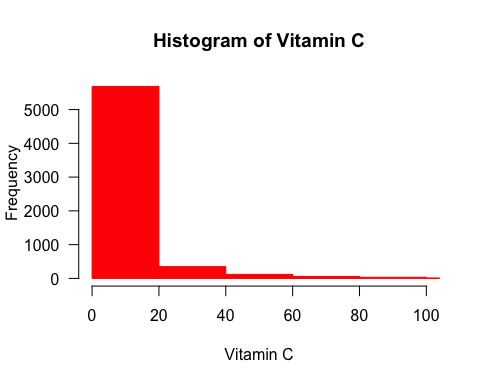
# 10. Create a scatter plot using Protein and Fat, with the plot title "Fat vs Protein", labeling the axes "Fat" and "Protein", and making the data points red.

plot(USDAclean1$Protein, USDAclean1$TotalFat, main=" Plot of Fat vs Protein", xlab="Protein", ylab="Fat", col="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( USDAclean1$VitaminC,   
 main="Histogram of Vitamin C",   
 xlab="Vitamin C",   
 border="red",   
 col="red",  
 xlim=c(0,100),  
 las=1,   
 breaks=100)



# 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.

USDAclean1["HighSodium"] <- USDAclean1$Sodium   
USDAclean1["HighSodium"] <- ifelse(USDAclean1["HighSodium"] > 648.9257, 1, 0)

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

USDAclean1["HighCalories"] <- USDAclean1$Calories  
USDAclean1["HighCalories"] <- ifelse(USDAclean1["HighCalories"] > 219.7, 1, 0)   
  
USDAclean1["HighProtein"] <- USDAclean1$Protein  
USDAclean1["HighProtein"] <- ifelse(USDAclean1["HighProtein"] > 11.71, 1, 0)   
  
USDAclean1["HighSugar"] <- USDAclean1$Sugar  
USDAclean1["HighSugar"] <- ifelse(USDAclean1["HighSugar"] > 8.257, 1, 0)   
  
USDAclean1["HighFat"] <- USDAclean1$TotalFat  
USDAclean1["HighFat"] <- ifelse(USDAclean1["HighFat"] > 10.32, 1, 0)

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

highsod.fat <- USDAclean1$Description[USDAclean1$HighSodium == 1 & USDAclean1$HighFat == 1]  
length(highsod.fat)

## [1] 400

# 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).

aggregate(USDAclean1$Iron~USDAclean1$HighProtein, USDAclean1,mean)

## HighProtein USDAclean1$Iron  
## 1 0 2.671081  
## 2 1 3.093716