Assignment2

Harnoor Rangi

2017-07-18

## R Markdown

# 1. Create a visualization to illustrate the distribution of values for Total Fat, Protein and Carbohydrate. (12p)

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

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

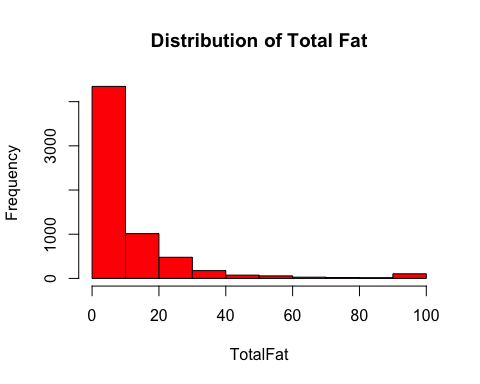
USDAclean<-read\_csv("MyData.csv")

## Warning: Missing column names filled in: 'X1' [1]

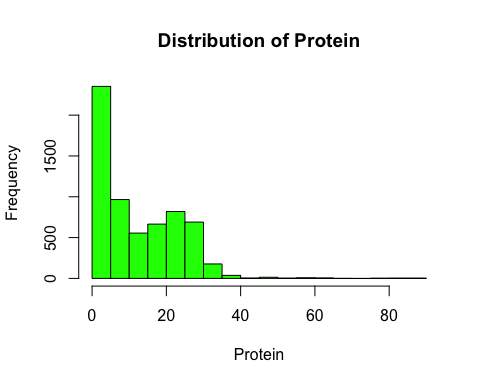
## Parsed with column specification:  
## cols(  
## .default = col\_integer(),  
## Description = col\_character(),  
## Protein = col\_double(),  
## TotalFat = col\_double(),  
## Carbohydrate = col\_double(),  
## Sugar = col\_double(),  
## Iron = col\_double(),  
## VitaminC = col\_double(),  
## VitaminE = col\_double(),  
## VitaminD = col\_double()  
## )

## See spec(...) for full column specifications.

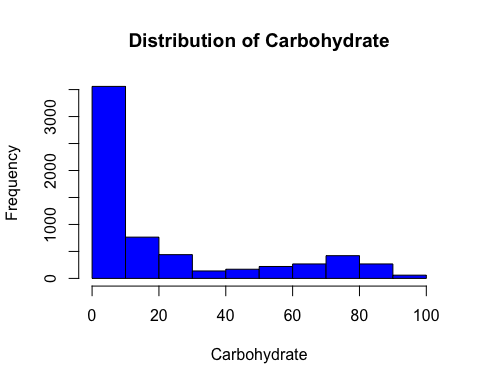
attach(USDAclean)  
hist(TotalFat, main = "Distribution of Total Fat", col = "red")



hist(Protein, main = "Distribution of Protein", col = "green")

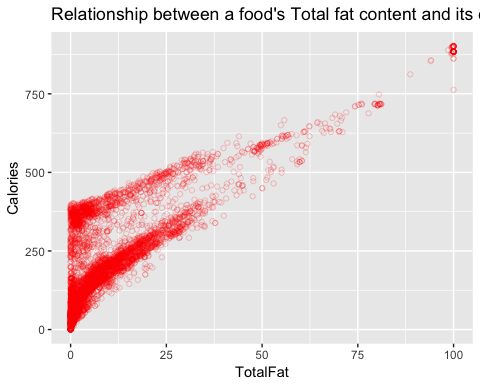


hist(Carbohydrate, main = "Distribution of Carbohydrate", col = "blue")



# 2.Create a visualization to illustrate the relationship between a food's Total Fat content and its calorie content. (12p)

library(ggplot2)  
g<-ggplot(USDAclean)  
g+geom\_point(aes(x=TotalFat,y=Calories),shape = 1,col="red",alpha=0.2)+labs(title="Relationship between a food's Total fat content and its calorie content")



# 3. Create a logistic regression model, using High Calories as the dependent variable, and Carbohydrate, Protein, Total Fat and Sodiumas independent variables. (16p)

logmodel1<-glm(HighCals ~ Carbohydrate+Protein+TotalFat+Sodium, data = USDAclean, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(logmodel1)

##   
## Call:  
## glm(formula = HighCals ~ Carbohydrate + Protein + TotalFat +   
## Sodium, family = "binomial", data = USDAclean)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.9191 -0.0457 -0.0033 0.0122 5.3606   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.437e+01 6.432e-01 -22.338 <2e-16 \*\*\*  
## Carbohydrate 2.388e-01 1.084e-02 22.029 <2e-16 \*\*\*  
## Protein 2.251e-01 1.447e-02 15.549 <2e-16 \*\*\*  
## TotalFat 6.465e-01 2.750e-02 23.506 <2e-16 \*\*\*  
## Sodium -6.051e-05 6.902e-05 -0.877 0.381   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8554.8 on 6309 degrees of freedom  
## Residual deviance: 1188.8 on 6305 degrees of freedom  
## AIC: 1198.8  
##   
## Number of Fisher Scoring iterations: 9

# 4. Which independent variable is not significant? (10 p)

# Answer: Sodium

# 5. Which independent variable has the strongest positive predictive power in the model?(10p)

# Answer: Carbohydrate, Protein, Total Fat

# 6. Create a script for a HealthCheck function to detect unhealthy foods. Foods that are high in salt, sugar and fat fail the HealthCheck, while all other foods pass. Foods that pass the HealthCheck should be assigned a 1, while foods that fail should be assigned a 0. Use the algorithm flowchart below as a basis for this script.(20p)

HealthCheck <- function(y){  
 x<-as.vector(numeric(6310))  
 for(i in 1:nrow(y))  
 {  
 if(y$HighSodium[i]==1 & y$HighSugar[i]==1 & y$HighFat[i] ==1)  
 x[i] <- 0  
 else if(y$HighSodium[i] ==0 | y$HighSugar[i] ==0 | y$HighFat[i] ==0 )  
 x[i] <- 1  
 }  
 return(x)  
}

# 7. Add a new column called HealthCheck to the USDAclean data frame using the output of the function. (10p)

He<-HealthCheck(USDAclean)  
USDAclean$HealthCheck <- He

# 8. How many foods in the USDAclean data frame fail the HealthCheck?(10p)

length(which(USDAclean$HealthCheck == 0))

## [1] 237