CMTH642 - Assignment # 2

By - Mohammed Amir

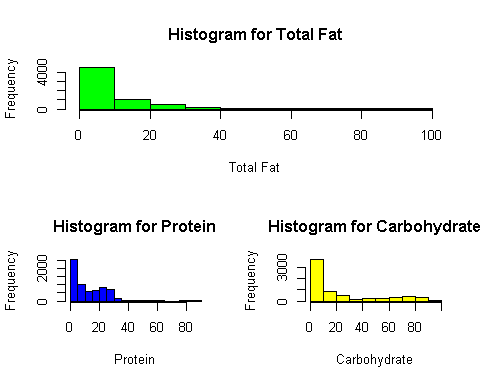
November 26, 2016

# Read the csv files in the folder

USDAClean <- read.csv(file="D:/Big Data/CMTH642 - DATA ANALYTICS ADVANCED METHODS/ASSIGNMENT 2/USDAClean\_MohammedAmir.csv",head=TRUE,sep=",")

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

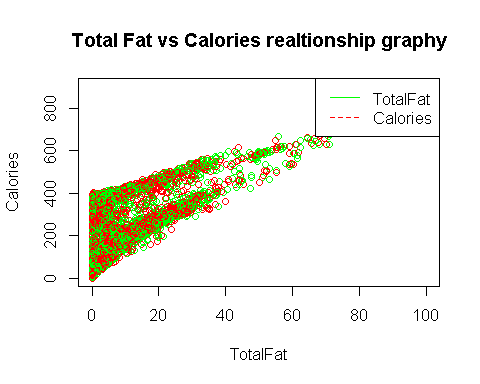
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))  
hist(USDAClean$TotalFat, main="Histogram for Total Fat", xlab="Total Fat", col="Green" )  
hist(USDAClean$Protein, main="Histogram for Protein", xlab="Protein", col="Blue")  
hist(USDAClean$Carbohydrate, main="Histogram for Carbohydrate", xlab="Carbohydrate" , col="Yellow")



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

Calories <- USDAClean$Calories  
TotalFat <- USDAClean$TotalFat  
plot (TotalFat, Calories, col = c("green", "red"), xlab="TotalFat", ylab="Calories", main="Total Fat vs Calories realtionship graphy")  
legend("topright", legend=c("TotalFat", "Calories"), col=c("green", "red"), lty=1:2, cex=1, box.lty=1)

# Based on the graphy below the relationship between Calories and Total Fat shows that, with lower total fat, the calories value concentration is high and as the total fat value increases the calories concentration starts to get less.



# 3. Create a logistic regression model, using High Calories as the dependent variable, and Carbohydrate, Protein, Total Fat and Sodium as independent variables.

Carbohydrate <- USDAClean$Carbohydrate  
Protein <- USDAClean$Protein  
TotalFat <- USDAClean$TotalFat  
Sodium <- USDAClean$Sodium  
HighCalories <- USDAClean$HighCalories  
  
logreg\_model <- glm (HighCalories ~ Carbohydrate + Protein + TotalFat + Sodium, data=USDAClean, family = binomial, )

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

summary(logreg\_model)

##   
## Call:  
## glm(formula = HighCalories ~ Carbohydrate + Protein + TotalFat +   
## Sodium, family = binomial, data = USDAClean)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.0892 -0.0401 -0.0030 0.0088 5.4038   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.450e+01 6.689e-01 -21.67 <2e-16 \*\*\*  
## Carbohydrate 2.475e-01 1.135e-02 21.82 <2e-16 \*\*\*  
## Protein 2.383e-01 1.541e-02 15.46 <2e-16 \*\*\*  
## TotalFat 6.722e-01 2.857e-02 23.53 <2e-16 \*\*\*  
## Sodium -2.848e-04 3.031e-04 -0.94 0.347   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8971.1 on 6612 degrees of freedom  
## Residual deviance: 1196.0 on 6608 degrees of freedom  
## AIC: 1206  
##   
## Number of Fisher Scoring iterations: 9

# 4. Which independent variable is not significant?

# As per the logistic regression value Sodium is not significant because its p value of 0.347 is greater than 0.05 and has no star rating associated to it.

# 5. Which independent variable has the strongest positive predictive power in the model?

# Comparing the coefficient number of all dependent variables, TotalFat has the highest coefficient value making it the highest positive predictive variable in the model

# 6. Create a script for a HealthCheck function to detect unhealthy foods.

HealthCheck <- function(sodium, sugar, fat)  
{  
 if(sodium == 0)  
 {  
 print ("1")  
 }  
 else  
 {  
 if(sugar == 0 && sodium == 1)  
 {  
 print ("1")   
 }  
 else  
 {  
 if(fat == 0 && sugar == 1 && sodium == 1)  
 {  
 print ("1")   
 }  
 else  
 {  
 print ("0")  
 }  
   
 }  
 }  
 }

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

USDAClean$HealthCheck <- mapply(HealthCheck,USDAClean$HighSodium,USDAClean$HighSugar,USDAClean$TotalFat)

# 8. How many foods in the USDAcle an data frame fail the HealthCheck?

table(USDAClean$HealthCheck)

##   
## 0 1   
## 512 6101

# 512 foods have failed the health check test were as 6101 foods have passed the test.