LIS4805 – Week 3: Breakfast Cereals Lab

Gabriel Myles

## Set working directory (optional)

# Example ONLY (edit then uncomment if you need it):  
# setwd("~/Desktop/LIS4805")

## 1) Load the data

# Make sure 'Cereals.csv' is in the SAME folder as this .Rmd  
cereals.df <- read.csv("Cereals.csv", stringsAsFactors = TRUE)  
  
# Quick checks  
str(cereals.df)

## 'data.frame': 77 obs. of 16 variables:  
## $ name : Factor w/ 77 levels "100%\_Bran","100%\_Natural\_Bran",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ mfr : Factor w/ 7 levels "A","G","K","N",..: 4 6 3 3 7 2 3 2 7 5 ...  
## $ type : Factor w/ 2 levels "C","H": 1 1 1 1 1 1 1 1 1 1 ...  
## $ calories: int 70 120 70 50 110 110 110 130 90 90 ...  
## $ protein : int 4 3 4 4 2 2 2 3 2 3 ...  
## $ fat : int 1 5 1 0 2 2 0 2 1 0 ...  
## $ sodium : int 130 15 260 140 200 180 125 210 200 210 ...  
## $ fiber : num 10 2 9 14 1 1.5 1 2 4 5 ...  
## $ carbo : num 5 8 7 8 14 10.5 11 18 15 13 ...  
## $ sugars : int 6 8 5 0 8 10 14 8 6 5 ...  
## $ potass : int 280 135 320 330 NA 70 30 100 125 190 ...  
## $ vitamins: int 25 0 25 25 25 25 25 25 25 25 ...  
## $ shelf : int 3 3 3 3 3 1 2 3 1 3 ...  
## $ weight : num 1 1 1 1 1 1 1 1.33 1 1 ...  
## $ cups : num 0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...  
## $ rating : num 68.4 34 59.4 93.7 34.4 ...

head(cereals.df)

## name mfr type calories protein fat sodium fiber carbo  
## 1 100%\_Bran N C 70 4 1 130 10.0 5.0  
## 2 100%\_Natural\_Bran Q C 120 3 5 15 2.0 8.0  
## 3 All-Bran K C 70 4 1 260 9.0 7.0  
## 4 All-Bran\_with\_Extra\_Fiber K C 50 4 0 140 14.0 8.0  
## 5 Almond\_Delight R C 110 2 2 200 1.0 14.0  
## 6 Apple\_Cinnamon\_Cheerios G C 110 2 2 180 1.5 10.5  
## sugars potass vitamins shelf weight cups rating  
## 1 6 280 25 3 1 0.33 68.40297  
## 2 8 135 0 3 1 1.00 33.98368  
## 3 5 320 25 3 1 0.33 59.42551  
## 4 0 330 25 3 1 0.50 93.70491  
## 5 8 NA 25 3 1 0.75 34.38484  
## 6 10 70 25 1 1 0.75 29.50954

## 2) Identify variable types

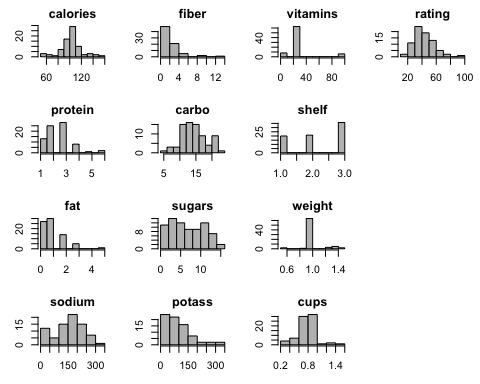
**Nominal (IDs/labels):** name, mfr, type  
**Ordinal:** shelf (1 = low, 2 = middle, 3 = high)  
**Quantitative/Numerical:** calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, weight, cups, rating

## 3) Summary stats for quantitative variables

quant.cols <- c("calories","protein","fat","sodium","fiber","carbo","sugars",  
 "potass","vitamins","shelf","weight","cups","rating")  
  
summary.stats <- sapply(cereals.df[quant.cols], function(x) c(  
 mean = mean(x, na.rm=TRUE),  
 median = median(x, na.rm=TRUE),  
 min = min(x, na.rm=TRUE),  
 max = max(x, na.rm=TRUE),  
 sd = sd(x, na.rm=TRUE)  
))  
round(summary.stats, 2)

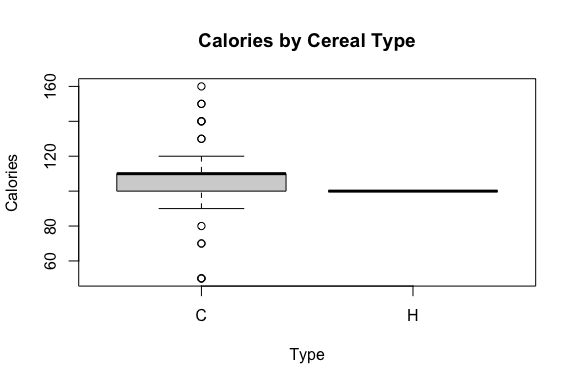
## calories protein fat sodium fiber carbo sugars potass vitamins shelf  
## mean 106.88 2.55 1.01 159.68 2.15 14.80 7.03 98.67 28.25 2.21  
## median 110.00 3.00 1.00 180.00 2.00 14.50 7.00 90.00 25.00 2.00  
## min 50.00 1.00 0.00 0.00 0.00 5.00 0.00 15.00 0.00 1.00  
## max 160.00 6.00 5.00 320.00 14.00 23.00 15.00 330.00 100.00 3.00  
## sd 19.48 1.09 1.01 83.83 2.38 3.91 4.38 70.41 22.34 0.83  
## weight cups rating  
## mean 1.03 0.82 42.67  
## median 1.00 0.75 40.40  
## min 0.50 0.25 18.04  
## max 1.50 1.50 93.70  
## sd 0.15 0.23 14.05

par(mfcol=c(4,4), mar=c(3,3,2,1))  
for (nm in quant.cols) {  
 hist(cereals.df[[nm]], main=nm, xlab="", col="gray")  
}  
par(mfcol=c(1,1))

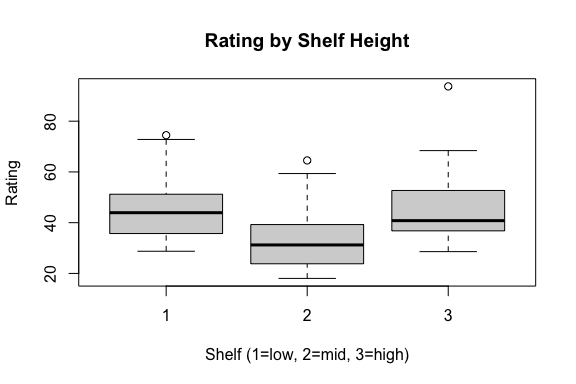


## 4) Boxplots

boxplot(calories ~ type, data=cereals.df,  
 main="Calories by Cereal Type", xlab="Type", ylab="Calories")



boxplot(rating ~ shelf, data=cereals.df,  
 main="Rating by Shelf Height", xlab="Shelf (1=low, 2=mid, 3=high)", ylab="Rating")

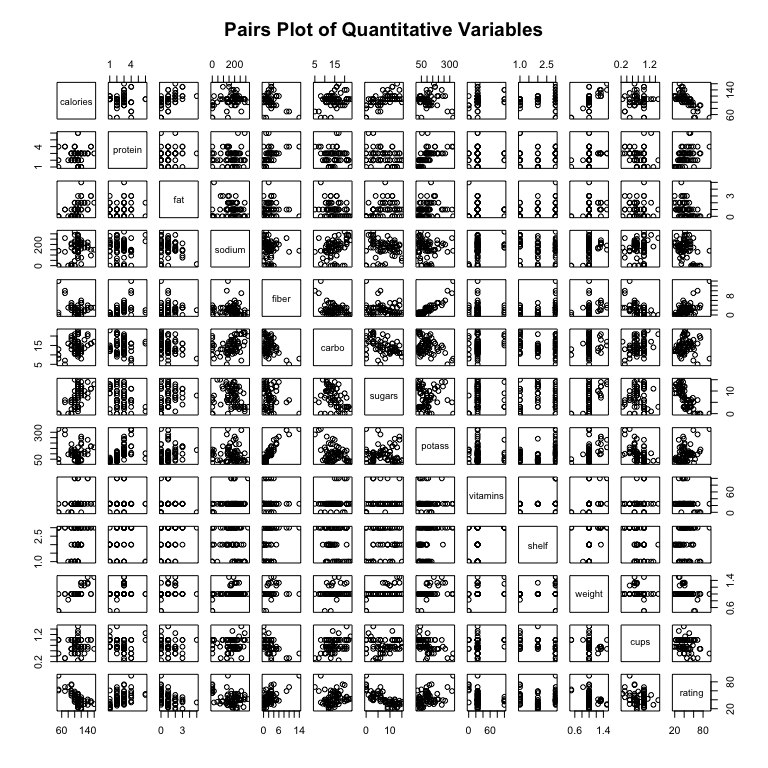


## 5) Correlation matrix + normalization

options(digits=2)  
corr\_raw <- cor(na.omit(cereals.df[quant.cols]))  
corr\_raw

## calories protein fat sodium fiber carbo sugars potass  
## calories 1.000 0.034 0.50737 0.29625 -0.295 0.271 0.5691 -0.0714  
## protein 0.034 1.000 0.20235 0.01156 0.514 -0.037 -0.2866 0.5787  
## fat 0.507 0.202 1.00000 0.00082 0.014 -0.285 0.2872 0.1996  
## sodium 0.296 0.012 0.00082 1.00000 -0.071 0.328 0.0371 -0.0394  
## fiber -0.295 0.514 0.01404 -0.07073 1.000 -0.379 -0.1509 0.9115  
## carbo 0.271 -0.037 -0.28493 0.32841 -0.379 1.000 -0.4521 -0.3650  
## sugars 0.569 -0.287 0.28715 0.03706 -0.151 -0.452 1.0000 0.0014  
## potass -0.071 0.579 0.19964 -0.03944 0.912 -0.365 0.0014 1.0000  
## vitamins 0.260 0.055 -0.03051 0.33158 -0.039 0.254 0.0730 -0.0026  
## shelf 0.089 0.196 0.27798 -0.12190 0.314 -0.189 0.0614 0.3946  
## weight 0.696 0.231 0.22171 0.31253 0.246 0.145 0.4605 0.4206  
## cups 0.089 -0.242 -0.15758 0.11958 -0.514 0.358 -0.0324 -0.5017  
## rating -0.694 0.467 -0.40505 -0.38301 0.603 0.056 -0.7560 0.4158  
## vitamins shelf weight cups rating  
## calories 0.2598 0.089 0.70 0.089 -0.694  
## protein 0.0548 0.196 0.23 -0.242 0.467  
## fat -0.0305 0.278 0.22 -0.158 -0.405  
## sodium 0.3316 -0.122 0.31 0.120 -0.383  
## fiber -0.0387 0.314 0.25 -0.514 0.603  
## carbo 0.2536 -0.189 0.14 0.358 0.056  
## sugars 0.0730 0.061 0.46 -0.032 -0.756  
## potass -0.0026 0.395 0.42 -0.502 0.416  
## vitamins 1.0000 0.284 0.32 0.134 -0.214  
## shelf 0.2844 1.000 0.19 -0.351 0.051  
## weight 0.3204 0.193 1.00 -0.202 -0.300  
## cups 0.1336 -0.351 -0.20 1.000 -0.223  
## rating -0.2145 0.051 -0.30 -0.223 1.000

pairs(na.omit(cereals.df[quant.cols]), main="Pairs Plot of Quantitative Variables")



scaled <- scale(cereals.df[quant.cols])  
corr\_scaled <- cor(na.omit(scaled))  
all.equal(corr\_raw, corr\_scaled) # Pearson r is scale-invariant

## [1] TRUE

corr\_scaled

## calories protein fat sodium fiber carbo sugars potass  
## calories 1.000 0.034 0.50737 0.29625 -0.295 0.271 0.5691 -0.0714  
## protein 0.034 1.000 0.20235 0.01156 0.514 -0.037 -0.2866 0.5787  
## fat 0.507 0.202 1.00000 0.00082 0.014 -0.285 0.2872 0.1996  
## sodium 0.296 0.012 0.00082 1.00000 -0.071 0.328 0.0371 -0.0394  
## fiber -0.295 0.514 0.01404 -0.07073 1.000 -0.379 -0.1509 0.9115  
## carbo 0.271 -0.037 -0.28493 0.32841 -0.379 1.000 -0.4521 -0.3650  
## sugars 0.569 -0.287 0.28715 0.03706 -0.151 -0.452 1.0000 0.0014  
## potass -0.071 0.579 0.19964 -0.03944 0.912 -0.365 0.0014 1.0000  
## vitamins 0.260 0.055 -0.03051 0.33158 -0.039 0.254 0.0730 -0.0026  
## shelf 0.089 0.196 0.27798 -0.12190 0.314 -0.189 0.0614 0.3946  
## weight 0.696 0.231 0.22171 0.31253 0.246 0.145 0.4605 0.4206  
## cups 0.089 -0.242 -0.15758 0.11958 -0.514 0.358 -0.0324 -0.5017  
## rating -0.694 0.467 -0.40505 -0.38301 0.603 0.056 -0.7560 0.4158  
## vitamins shelf weight cups rating  
## calories 0.2598 0.089 0.70 0.089 -0.694  
## protein 0.0548 0.196 0.23 -0.242 0.467  
## fat -0.0305 0.278 0.22 -0.158 -0.405  
## sodium 0.3316 -0.122 0.31 0.120 -0.383  
## fiber -0.0387 0.314 0.25 -0.514 0.603  
## carbo 0.2536 -0.189 0.14 0.358 0.056  
## sugars 0.0730 0.061 0.46 -0.032 -0.756  
## potass -0.0026 0.395 0.42 -0.502 0.416  
## vitamins 1.0000 0.284 0.32 0.134 -0.214  
## shelf 0.2844 1.000 0.19 -0.351 0.051  
## weight 0.3204 0.193 1.00 -0.202 -0.300  
## cups 0.1336 -0.351 -0.20 1.000 -0.223  
## rating -0.2145 0.051 -0.30 -0.223 1.000