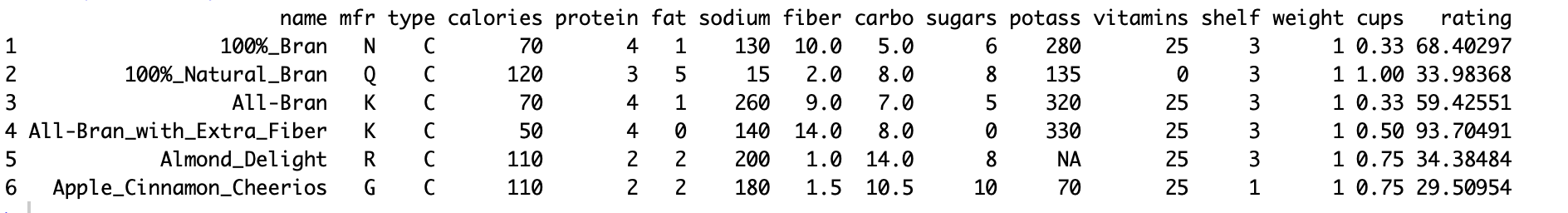
# **Assignment -1**

1. Import the data to R. Copy the R code used below. (0.25 points)

cereals.df <- read.csv(file="cereals.csv", header=TRUE, sep=",")



1. Which variables are numerical? Which variables are categorical? (0.25 points)

“Name is the primary key which uniquely identifies each row in the data frame”

"Manufacturer, shelf and type are **categorical variables** "

"calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, weight, cups, rating are all **numerical variables**"

(c) Compute the mean, median, min, max, and standard deviation for each of the numeric variables.

Copy the code used and the resulting table below. (Tip: this can be done through R’s sapply() function.)   
 (0.5 point)

Numeric\_cereals.df <- cereals.df[,-c(1:3,13)]

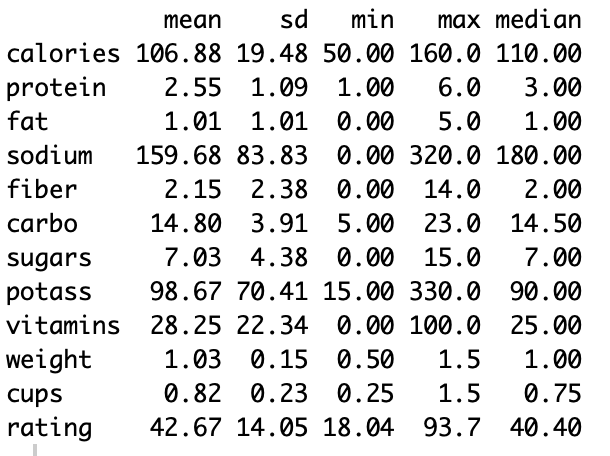
round(data.frame(mean=sapply(Numeric\_cereals.df, mean, na.rm=TRUE),

sd=sapply(Numeric\_cereals.df, sd, na.rm=TRUE),

min=sapply(Numeric\_cereals.df, min, na.rm=TRUE),

max=sapply(Numeric\_cereals.df, max, na.rm=TRUE),

median=sapply(Numeric\_cereals.df, median, na.rm=TRUE)),2)



(d) Plot a histogram for each of the numeric variable. Copy the code used and the generated charts

below. Based on the histograms and summary statistics, which variables have the largest variability?

Which variables seem skewed? (1.25 points = 0.5+0.25+0.25)

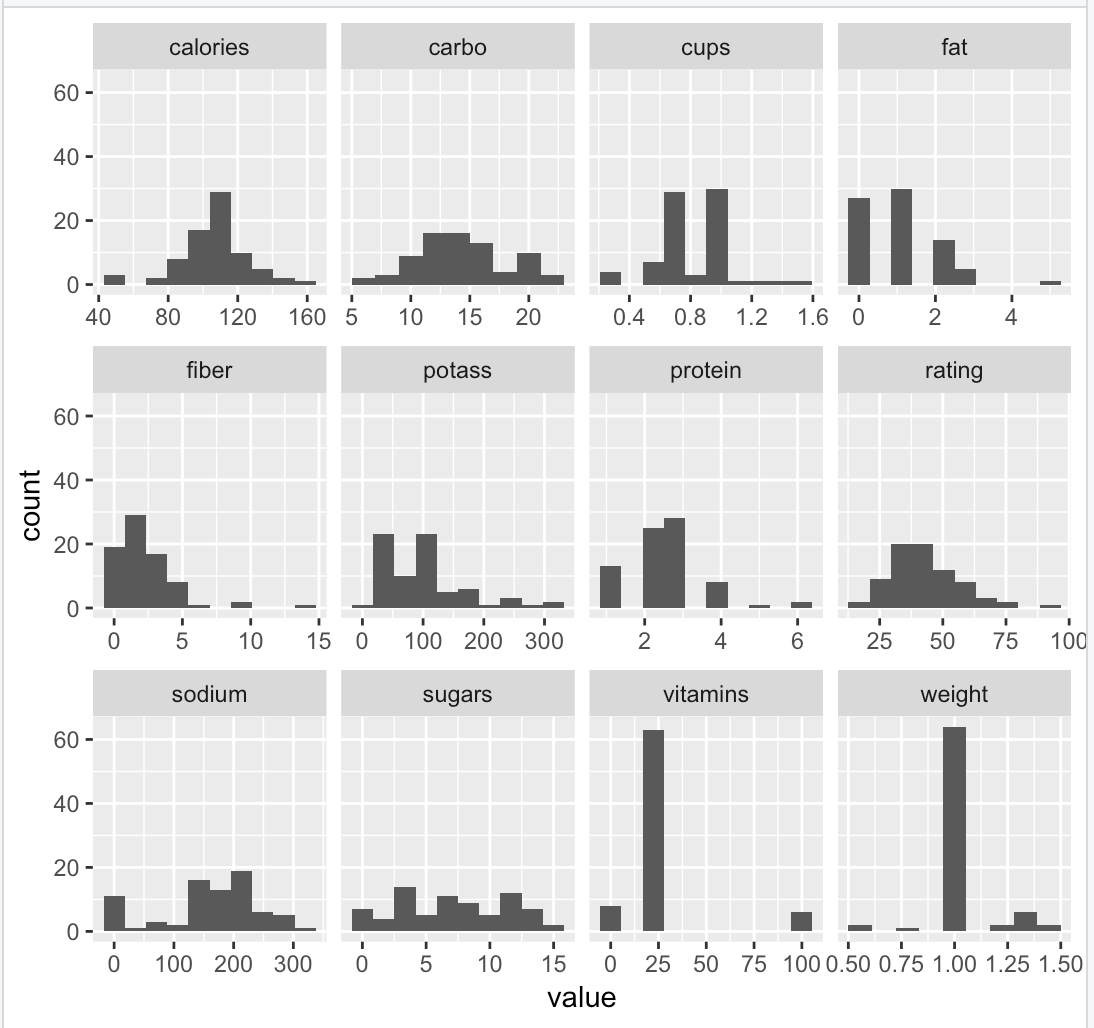
library(tidyr)

library(ggplot2)

ggplot(gather(Numeric\_cereals.df, na.rm=TRUE ), aes(value)) +

geom\_histogram(bins = 10) +

facet\_wrap(~key, scales = 'free\_x')



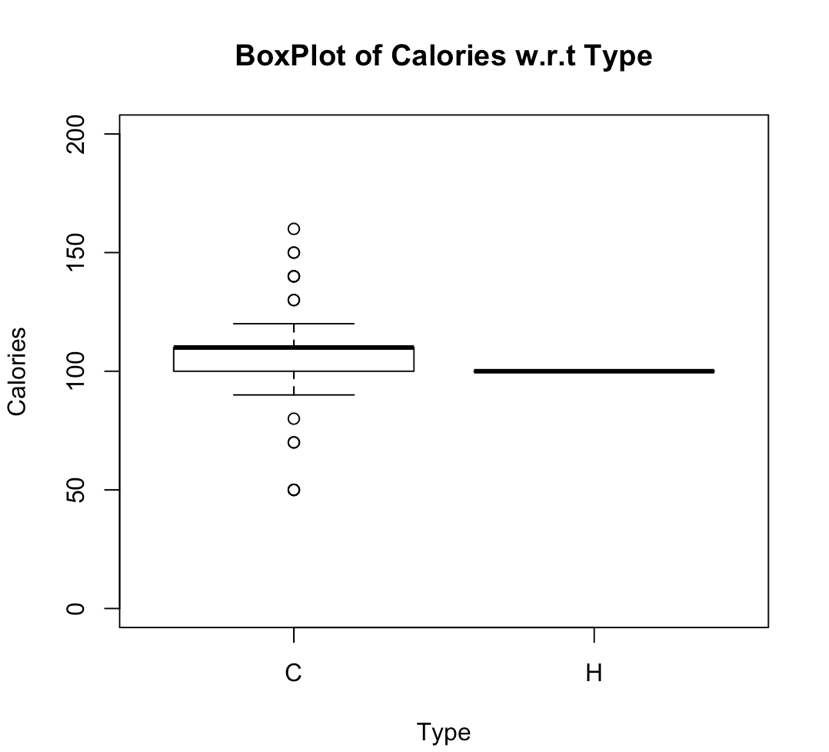
We observe almost all variables have significant variability, based on the histogram plots and summary statistics such as sd we observe sugars, carbo and sodium variables have the highest variability.

Based on the histogram plots, we observe protein, fat, fiber, potass variables are right skewed."

(e) Plot a side-by-side boxplot comparing the calories in hot vs. cold cereals. Copy the code used and

the generated chart below. What does this plot show us? (1 point = 0.5+0.5)

boxplot (cereals.df$calories ~ cereals.df$type, xlab = "Type", ylab = "Calories", ylim= c(0,200), main = "BoxPlot of Calories w.r.t Type")



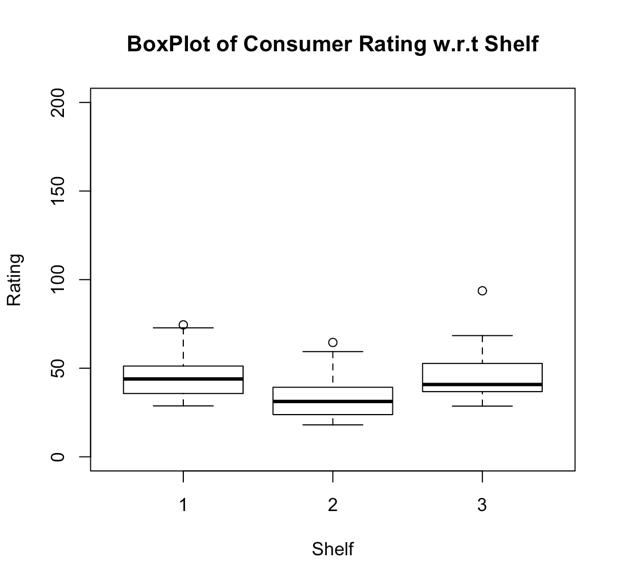
" From the boxplot above, we observe that the cold breakfast cereals on an average have higher calories compared to hot ones"

(f) Plot a side-by-side boxplot of consumer rating as a function of the shelf height. Copy the code

used and the generated chart below. If we were to predict consumer rating from shelf height, does it

appear that we need to keep all three categories of shelf height? Why? (1.5 points = 1+0.5)

boxplot (cereals.df$rating ~ cereals.df$shelf, xlab = "Shelf", ylab = "Rating", ylim=c(0,200), main = "BoxPlot of Consumer Rating w.r.t Shelf")



"From the boxplot we observe that the shelf height 1 has highest average consumer ratings followed by shelf height3 and 2 "

“Not necessary to have all 3 categories of shelf height to predict the consumer rating as we can reduce by combining categories that are close to each other. From the above boxplot we observe shelf1 and shelf 3 are close to each other, we can include either one to predict the average consumer rating.

(g) Compute the correlation table for the numeric variables (function cor()). In addition, generate a

matrix plot for these variables (function plot(data)). Copy the code used and the results below. Which

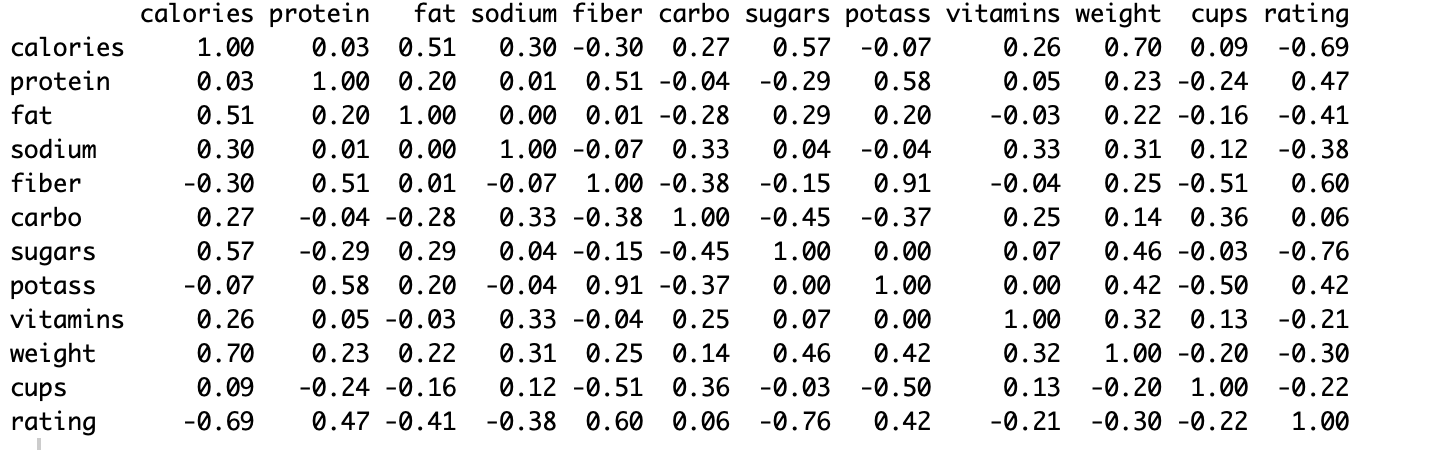
pair of variables is most strongly corrected? How can we reduce the number of variables based on

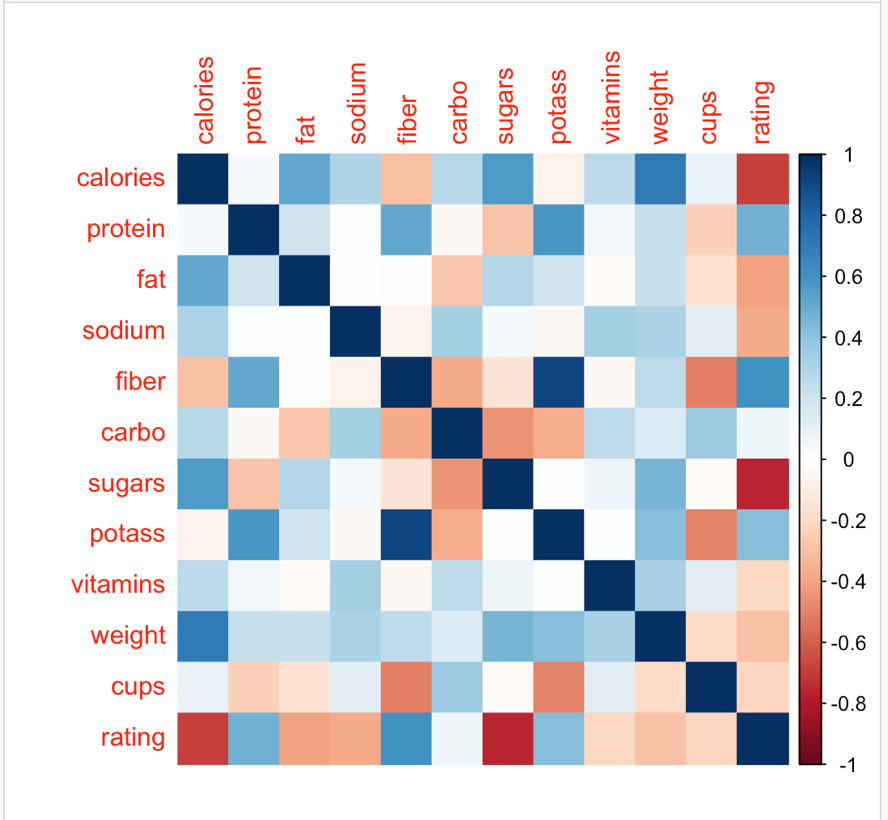
these correlations? How would the correlations change if we normalized the data first? (1.75 points = 0.5+0.5+0.25+0.25+0.25)

library(corrplot)

Correlation\_matrix <- round(cor(Numeric\_cereals.df,use="complete.obs"),2)

corrplot(Correlation\_matrix, method="color")





" From the plot of correlation matrix we observe that rating and calories, rating and sugars, potass and fiber are strongly correlated"

“We can remove highly correlated variables based on the correlation matrix to reduce the set of numeric variables

#Correlation plot of normalized data

Normalized\_Numeric\_cereals.df <- scale(Numeric\_cereals.df)

Normalized\_Correlation\_matrix <- round(cor(Normalized\_Numeric\_cereals.df,use="complete.obs"),2)

corrplot(Normalized\_Correlation\_matrix, method="color")

"We observe that the correlations does not change with the normalization of data"

" Since the formula for calculating the correlation coefficient standardizes the variables, changes in scale or units of measurement will not affect its value."

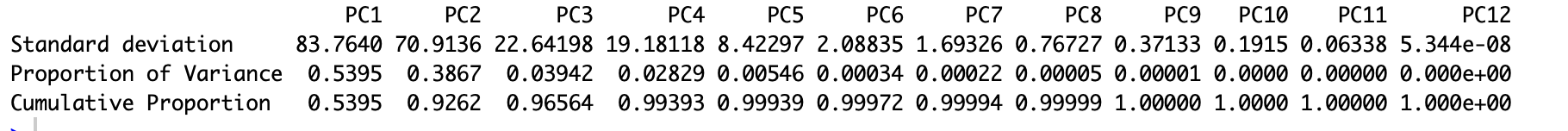
(h) As that on page 25 of class slides used for Week 3, conduct PCA (without normalization).

Consider the first PC of the analysis. Describe briefly what this PC represent. Furthermore, what is

the score of this PC for the first record? Also, show us how you can manually derive this value. (1.5 points = 0.25+0.5+0.25+0.5)

pcs <- prcomp(na.omit(cereals.df[,-c(1:3,13)]))

summary(pcs)



" In the first PC of the analysis we observe that, PC1 explains 53.95% of the total variance."

pcs$rotation[1,1]

[1] 0.07798472

The score of this PC for the first record is 0.07798472.

Also, show us how you can manually derive this value.