Stat 441
Assignment 2

Pengyu Xiong

1501982

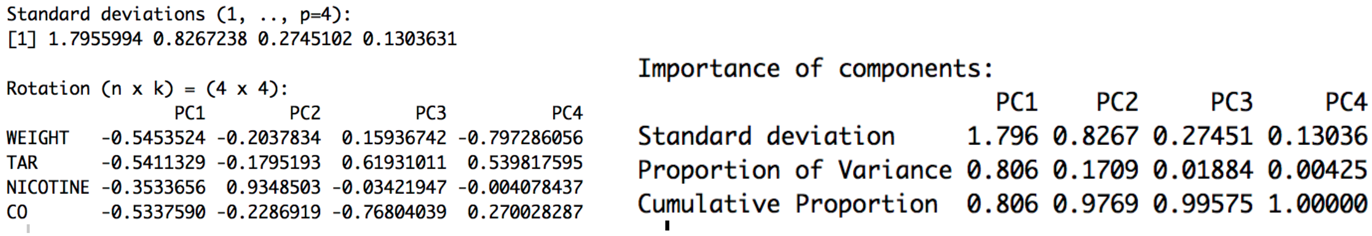
Q1

- **Data Processing**

The cigs data should only have four columns that are Weight, Tar, Nicotine, CO. The rest of the noise data should be deleted. After the clean up, the name of each column should be assigned correctly. The row names need to be assigned correctly. Extracted each brand name under that column and assign them as the name for each row. After all these procedures, the data will have the four columns correspond to 25 different brands.

- **Build Principle Components**

There are going to have four Principle Components correspond to the number of columns. Due to that the dimension will also be D4. The initial four components that was calculated from R can be found at figure 1.1 . From the output, the PC1 and PC2 combination could explain almost 98% of variance for the dataset. Apparently, PC4 has small coverage inside the dataset.



```
Standard deviations (1, .., p=4):
[1] 1.7955994 0.8267238 0.2745102 0.1303631

Rotation (n x k) = (4 x 4):
                PC1        PC2         PC3          PC4
WEIGHT   -0.5453524 -0.2037834  0.15936742 -0.797286056
TAR      -0.5411329 -0.1795193  0.61931011  0.539817595
NICOTINE -0.3533656  0.9348503 -0.03421947 -0.004078437
CO       -0.5337590 -0.2286919 -0.76804039  0.270028287
```

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.796 0.8267 0.27451 0.13036
Proportion of Variance 0.806 0.1709 0.01884 0.00425
Cumulative Proportion  0.806 0.9769 0.99575 1.00000
```

Figure 1.1 *The data description for the four PCs (principle component)*

The result from the above can be visualized with figure 1.2. It is pretty obvious that the PC4 has very slight effect on the dataset. The previous three PCs can explain enough Variance the PC4 can be eliminated.

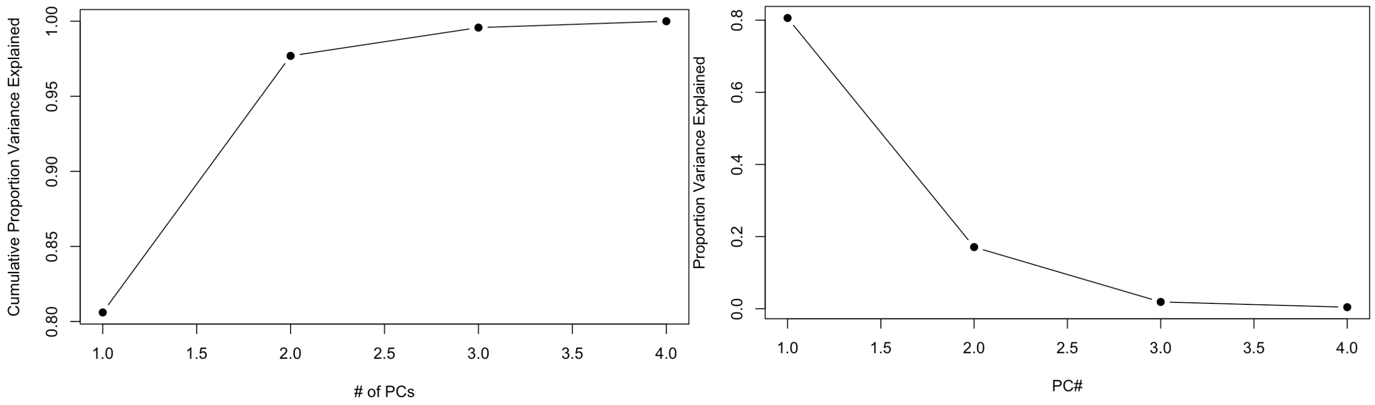


Figure 1.2 *The variance explained by number of pcs (left), the correspond explained variance for each pc (right)*

From figure 1.3, it can be seen that under two PCs the correlation between Weight, Tar, CO are much stronger than Nicotine. The biplot can explain the dimension of D2 constructed by PC1 and PC2. The data clusters have the trend that can be explained using the plots. Also, from the figure, it is clear that brand “Durham” is an outlier which can be considered to dumped away.

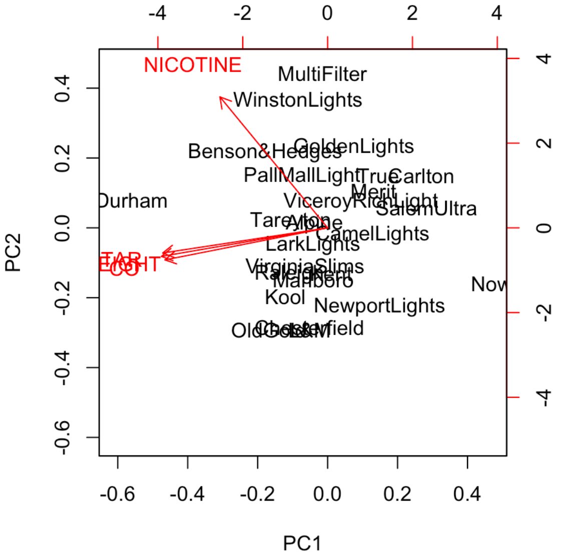


Figure 1.3 *Biplot for PC1 and PC2 explaining the dataset*

- **Summary**

In general, the dimension can be reduced from D4 to D3 with less than 1% loss of information. However, if more loss is acceptable; by adapting only CP1 and Cp2 the data can be reduced to D2 while having around 2.5% of loss. The first principle component has more than 80% of variance being explained which suggest this component or vector inside the vector space is much more closer to the actual data.

Q2

- **Data Processing**

The input data has 14 columns and 77 rows. Each row represents a cereal brand while each column indicates a ingredient associate to the cereal brands. The data was read in as data frame format in R.

However, when it comes to the later column analyzing regards cereal ingredient in columns; the first two columns need to be removed, because there is not as much can be learned. The matrix of the properties can then be used to compare the ingredients inside Cereal.

- **Distance Calculation**

The Euclidean distance between the cereal brands can show whether there is any similarities exist inside all the rows. The smaller distance shows more similarity between the brands and greater distance meant low similarity. The calculation follows the following formula:

$$d(A, B) = \sqrt{(A1 - B1)^2 + ... + (Ap - Bp)^2}$$

- **Multidimension Scaling**

After the Euclidean distance is being calculated, with the dissimilarities the scaling became possible. The following procedure would be reconstruct the high dimension spaces using the distance matrix. And finally select the dimension transformation that has lowest sum of squares.

However, like in Q1, the output dimensions is hard to decide. What has been done is output the first dimension that has stress lower than 1%. The stress seems rather low and same with sum of squares. The dimension that was selected is D4 which is fairly small compare to the initial 14 dimensions. The stress rate of 4 principle components dimension is 0.121879%. The stress formula is

$$\sum_{i \neq j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}$$

The resulting dimension is 4D which is hard to Visualize in R. Due to the fact that most of information is contained inside CP1 and CP2, it is sufficient to present relation using the first two PCs. The following are the first two Principle Components graphs Figure 2.1. It can be found out that there are several Cereal brands having special ingredients from the figure since it is far away from the most data. The similarity of the data can be detected according to Figure 2.1.

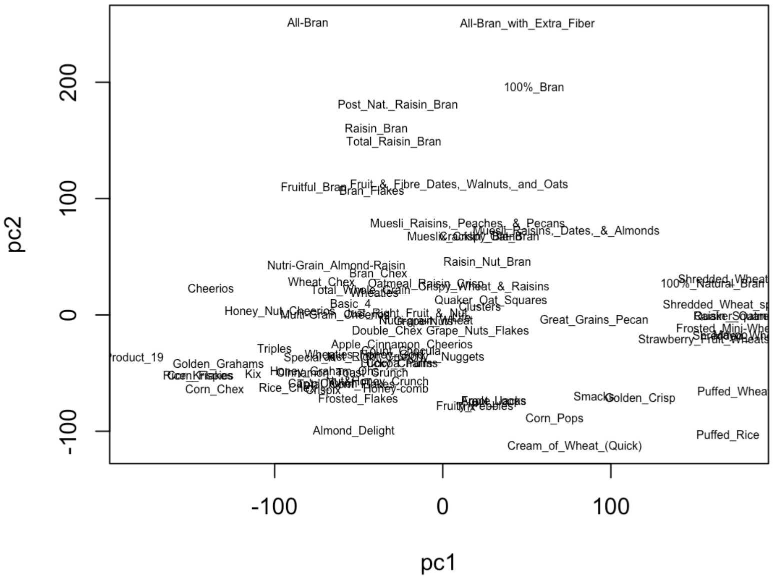


***Figure 2.1** The first 2 PC transformation for different brands (brand is represented in index on x axis).*

- **Analyzing Cereal Properties**

Using four principle components seems unable to be visualized. And there is limited information being kept inside component 3 and component 4. Thus, it is reasonable to analyze the data using the first two principle components when there is most information can be preserved under dimension of 2.

After the data processing, let the matrix with columns as the analyzing subject. The general trend of the ingredients of the Cereal can be found at figure 2.2 which is plotted from the matrix. This can be reported as the general trend of Cereal ingredients. There are more fat and Vitamins inside Cereal than Calories and fibre in general.

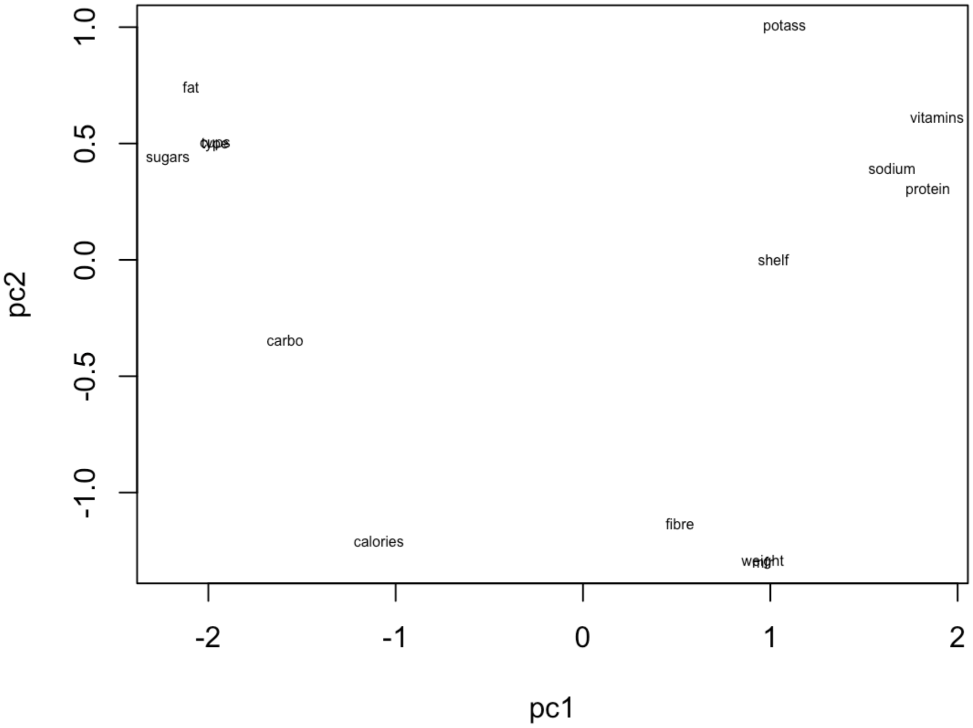


Figure 2.2 *The first 2 PC transformation for different ingredients of the brands*

- **Summary**

To summarize, it seems rather efficient to transform the original D14 dataset into D4 dimension. The 4 PCs are sufficient to illustrate the relations of the Cereal brands. The MDS is able to find a minimized dimension transformation under these conditions. When it comes to analyze the relation of the ingredients, it seems sufficient to have D2 dimension inside. With the previous work, it can be concluded that the cereal have more healthy ingredients inside.

Appendix

Q1.

```
# please make sure the data file is at the current directory

# read in data
cigs = read.table("cigs.data", header = TRUE, na.strings = "?", fill = TRUE, row.names = 2)
cigs[, 1] = NULL;

plot(cigs)

# check component
pcs <- prcomp(cigs, scale = T)
pcInf <- summary(pcs)
summary(pcs)
# plot component
biplot(pcs)

eqscplot()
plot(pcInf$importance[3, ], type="b", lty = 1, pch = 19, xlab = "# of PCs", ylab = "Cumulative Proportion Variance
Explained")
eqscplot()
plot(pcInf$importance[2, ], type = "b", pch = 19, xlab = "PC#", ylab = "Proportion Variance Explained")
eqscplot()
biplot(prcomp(cigs[-3], scale = T))
```

Q2.

```
Cereal.data = read.table("cerealbrands.data", header = TRUE, na.strings = "?", fill = TRUE)
Cereal <- as.matrix(Cereal.data)
library(MASS)

# convert the origin datasets into dist
cerealDist <- dist(Cereal)

# max dimension is 14
i = 1
while (i <= 14){
  output = isoMDS(cerealDist, k = i)
  # make sure there is less than 1% stress
  if (output$stress > 1){
    i <- i + 1
  } else{
    print(paste0("selected dimension: ", i))
    break
  }
}

# the selected dimension is 4d in this case

# plot cmd scaling
data <- cmdscale(dist(Cereal), k = i, eig = T)
```

```
plot(data$points, type = 'n', xlab = "pc1", ylab = "pc2")
text(data$points, labels = row.names(Cereal.data), cex = 0.5)

# transpose
# Pre processing
Cereal.slice = (Cereal.data[, 3:14])
Cereal.cor <- cor(Cereal.slice)

Cereal.t.dist <- dist(cor(Cereal.cor))
Cereal.t.MDS <- cmdscale(Cereal.t.dist, k = 2, eig = T)
plot(Cereal.t.MDS$points, type = 'n', xlab = "pc1", ylab = "pc2")
text(Cereal.t.MDS$points, labels = names(Cereal.data), cex = 0.5)

eqscplot()
plot(data$points[, 1], xlab = "x", ylab = "PC1 Scaling")
eqscplot()
plot(data$points[, 2], xlab = "x", ylab = "PC2 Scaling")
eqscplot()
plot(data$points[, 3], xlab = "x", ylab = "PC3 Scaling")
eqscplot()
plot(data$points[, 4], xlab = "x", ylab = "PC4 Scaling")

# MDS scaling
eqscplot()
plot(output$points[, 1], xlab = "x", ylab = "PC1 Scaling")
eqscplot()
plot(output$points[, 2], xlab = "x", ylab = "PC2 Scaling")
eqscplot()
plot(output$points[, 3], xlab = "x", ylab = "PC3 Scaling")
eqscplot()
plot(output$points[, 4], xlab = "x", ylab = "PC4 Scaling")

# biplot
eqscplot()
biplot(prcomp(output$points, scale = T))
```