Project 2

VADYM DUDARENKO

PCA - Calories and Macronutrients Per Capita per Day US 1909-2010

The goal of this project is to use PCA (Principal Component Analysis) algorithm for dimension reduction on the calories from food group in the USA from 1909 -2010 Problem of high dimensional data occurs when the dimension of the dataset (each numeric variable is a dimension) is large in comparison to number of observations. The goal of dimension reduction is to decrease the size of the dataset preserving as much information as possible.

Dataset

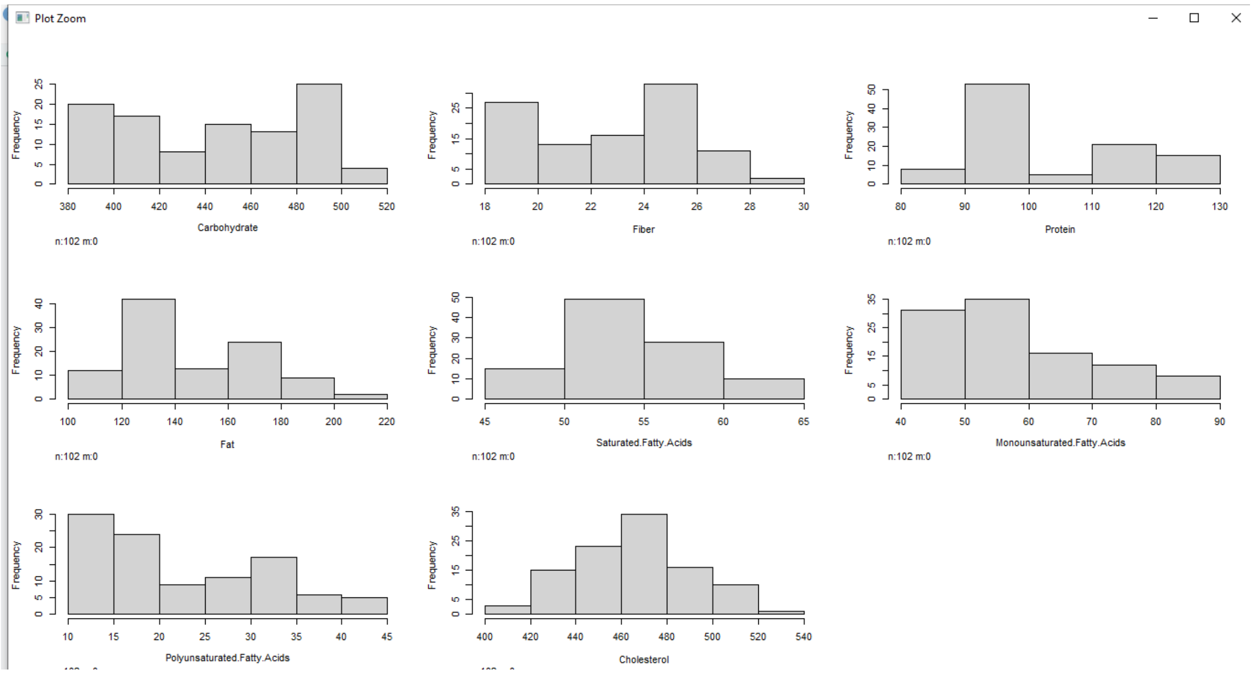
Dataset used in this project contains nutritional values from food group in the USA from 1909 -2010. Whole dataset can be found on data.world website (https://data.world/garyhoov/us-calories-from-food-groups). Each year is described by 8 variables: Carbohydrate, Fiber, Protein, Fat, Saturated.Fatty.Acids, Monounsaturated.Fatty.Acids, Polyunsaturated.Fatty.Acids., Cholesterol

Descriptive statistics:

```
summary(df)
Carbohydrate      Fiber        Protein          Fat
 Min.   :384.0   Min.   :18.0   Min.   : 86.0   Min.   :113.0
 1st Qu.:405.2   1st Qu.:20.0   1st Qu.: 93.0   1st Qu.:129.0
 Median :445.5   Median :24.0   Median : 96.5   Median :140.0
 Mean   :444.8   Mean   :23.4   Mean   :102.9   Mean   :147.7
 3rd Qu.:483.0   3rd Qu.:26.0   3rd Qu.:117.0   3rd Qu.:166.0
 Max.   :506.0   Max.   :29.0   Max.   :125.0   Max.   :202.0
 Saturated.Fatty.Acids Monounsaturated.Fatty.Acids Polyunsaturated.Fatty.Acids
 Min.   :46.00         Min.   :42.00               Min.   :12.00
 1st Qu.:53.00      1st Qu.:49.00         1st Qu.:15.00
 Median :54.50       Median :54.00          Median :20.00
 Mean   :54.72       Mean   :58.46         Mean   :23.22
 3rd Qu.:56.00       3rd Qu.:68.00         3rd Qu.:31.00
 Max.   :64.00      Max.   :88.00         Max.   :45.00
  Cholesterol
 Min.   :410.0
 1st Qu.:460.0
 Median :470.0
 Mean   :469.8
 3rd Qu.:490.0
 Max.   :530.0
```

Histograms:

```
library(Hmisc)
hist.data.frame(df)
```



Dimensions of the dataset:

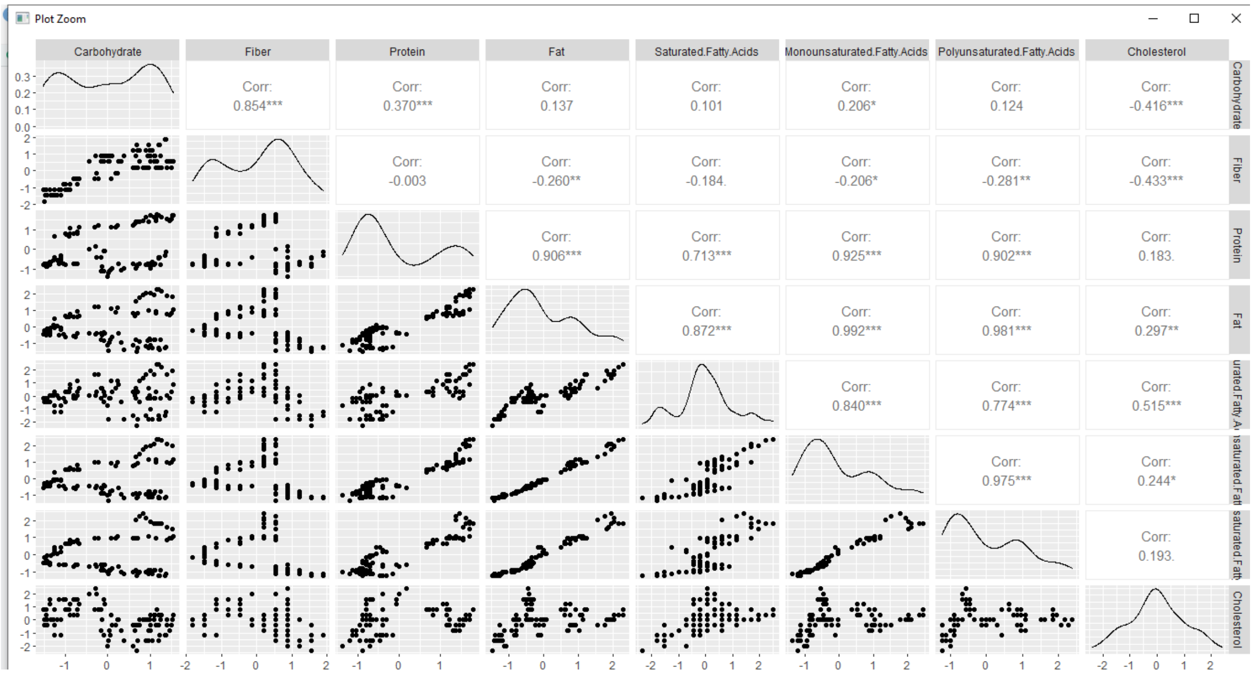
```
dim(df)
[1] 102   8
```

Before further analysis the data has been normalized:

```
install.packages("AppliedPredictiveModeling")
library(AppliedPredictiveModeling)
library(caret)
preproc <- preProcess(df, method=c("center", "scale"))
df_norm <- predict(preproc, df)
```

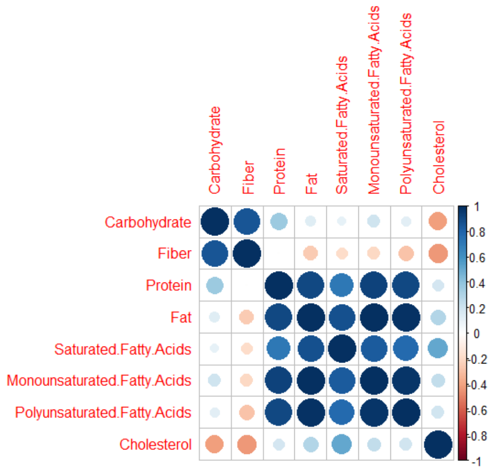
Matrix of plots:

```
library(GGally)
ggpairs(df_norm)
```



Correlation matrix:

```
install.packages("corrplot")
install.packages("xlsx")
library(corrplot)
library(xlsx)
cor<-cor(df_norm, method="pearson")
corrplot(cor)
```



On the correlation matrix it’s visible that some variables are positively correlated with saturated fat. Also the correlation between protein and carbohydrates can be easily spotted.

PCA:

```
pca <- prcomp(df_norm, center=FALSE, scale=FALSE)
```

PCA projections has been calculated using prcomp function which uses the singular value decomposition.

```
pca$rotation
                                    PC1         PC2         PC3          PC4
Carbohydrate                -0.06337142  0.64735444 -0.20362001 -0.045735436
Fiber                        0.11543745  0.60517306 -0.38158920  0.020413784
Protein                     -0.42084336  0.18154532  0.05035632 -0.572891432
Fat                         -0.45619584  0.01209118  0.10034472  0.092209461
Saturated.Fatty.Acids       -0.40961788 -0.04492674 -0.34206792  0.721616207
Monounsaturated.Fatty.Acids -0.45234960  0.05952808  0.12345255  0.008430391
Polyunsaturated.Fatty.Acids -0.44229852  0.02311302  0.26804226 -0.072772180
Cholesterol                 -0.17146289 -0.41892324 -0.77215679 -0.366998036
                                    PC5         PC6           PC7           PC8
Carbohydrate                 0.71171272  0.06014061  0.1502083175 -0.0254760671
Fiber                       -0.59366737 -0.32517182 -0.1274168173 -0.0001615947
```

```
Protein                  -0.30094235  0.59298595  0.1272019753  0.0285101431
Fat                      -0.01267066 -0.19350913 -0.0876588351 -0.8531571433
Saturated.Fatty.Acids       -0.11030524  0.34515802  0.1884345395  0.1601420658
Monounsaturated.Fatty.Acids         0.13003458     -0.12552491     -0.7814389302
0.3649826607
Polyunsaturated.Fatty.Acids      -0.04236913      -0.56737617           0.5396136852
0.3343674267
Cholesterol               0.13937660 -0.21157031 -0.0008340112  0.0012667458
```

Choosing number of components:

There are 3 most common methods used to select the number of components:

1)Kaiser rule

Kaiser rule focuses on component’s eigenvalues. An eigenvalue is an index that indicates how good a component is as a summary of the data (if an eigenvalue equals to 1, it means that the component contains the same amount of information as a single variable). This approach suggests that only components with eigenvalues higher than 1 should be chosen.

```
df_norm.cov<-cov(df_norm)
df_norm.eigen<-eigen(df_norm.cov)
df_norm.eigen$values
```

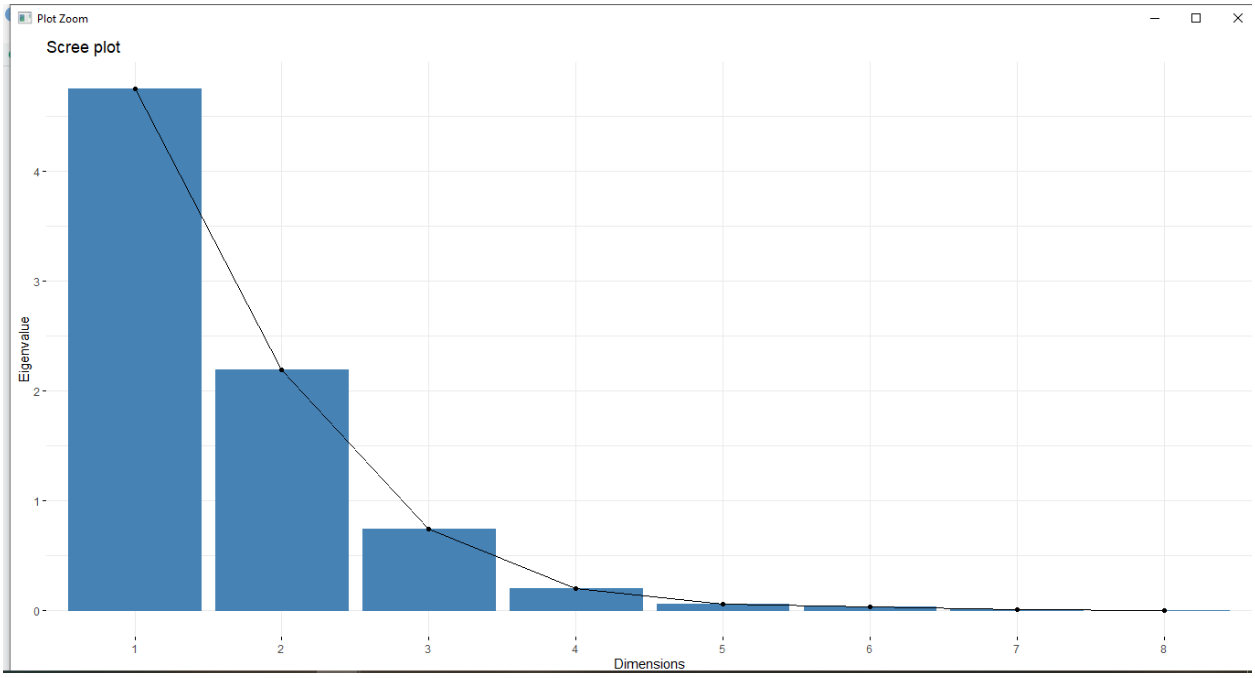
```
[1]  5.5275818362  2.3985857399  0.7482610837  0.2040697277  0.0677550609
0.0348692153
[7] 0.0119704484 0.0061688051 0.0007380829
```

Eigenvalues displayed above indicate that only 2 components should be chosen.

2)Scree plot

The second approach relies on the scree plot. This plot visualizes the eigenvalues of the components in the ascending order. Scree plot approach suggests that the appropriate number of components is the number of bars preceding the bend of the line connecting eigenvalues.

```
library("factoextra")
fviz_eig(pca, choice='eigenvalue')
```



This approach, as well as the Kaiser rule, indicates that the right number of components is 2.

3)Proportion of variance explained

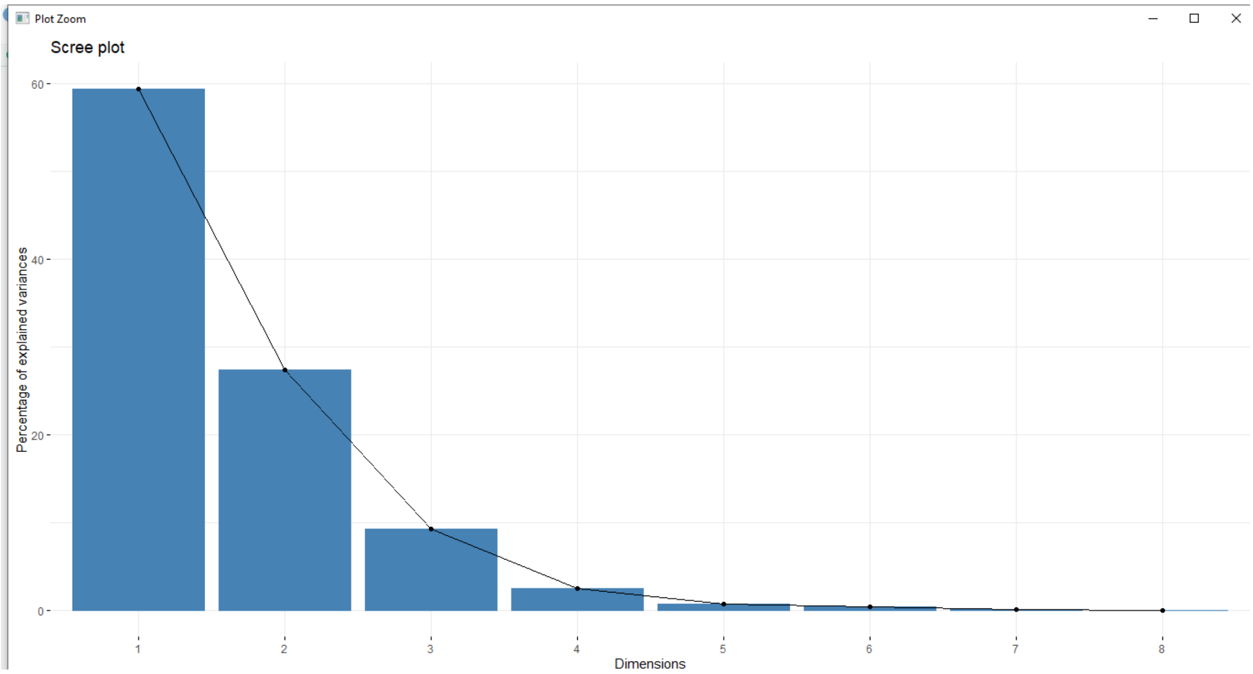
The last approach suggests that chosen components should explain over 2/3 of the variance.

```
summary(pca)

Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8
```

```
Standard deviation     2.1793 1.4798 0.86483 0.4508 0.24910 0.18668 0.10929 0.02847
Proportion of Variance 0.5937 0.2737 0.09349 0.0254 0.00776 0.00436 0.00149 0.00010
Cumulative Proportion  0.5937 0.8674 0.96089 0.9863 0.99405 0.99841 0.99990 1.00000
```

```
fviz_eig(pca)
```

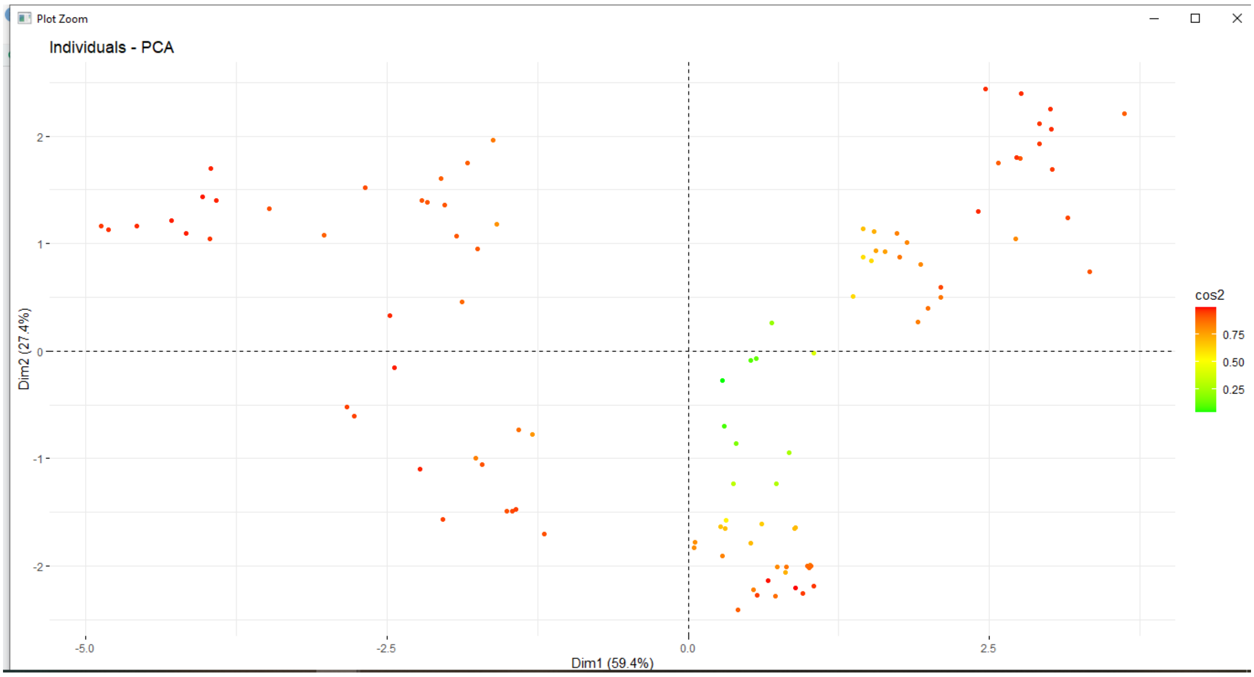


Cumulative proportion of explained variance displayed above indicates that 4 components are able to explain over 95% of the variance. It means that this proportion of information can be preserved after reducing number of variables by half. First two components are able to explain over 85% of the variance so this number of components is enough. It means that results given by all three methods are the same.

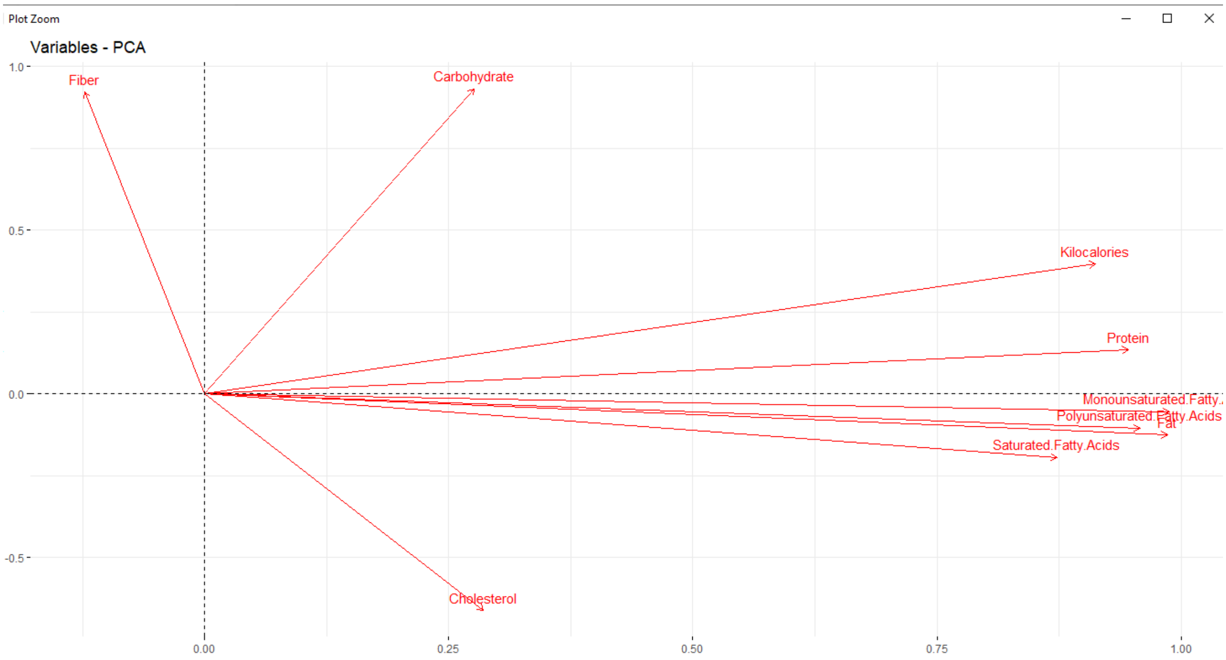
Components analysis:

The “cloud of points” graph shows individual observations quality of representation.

```
fviz_pca_ind(pca, col.ind="cos2", geom = "point", gradient.cols = c("green", "yellow", "red" ))
```

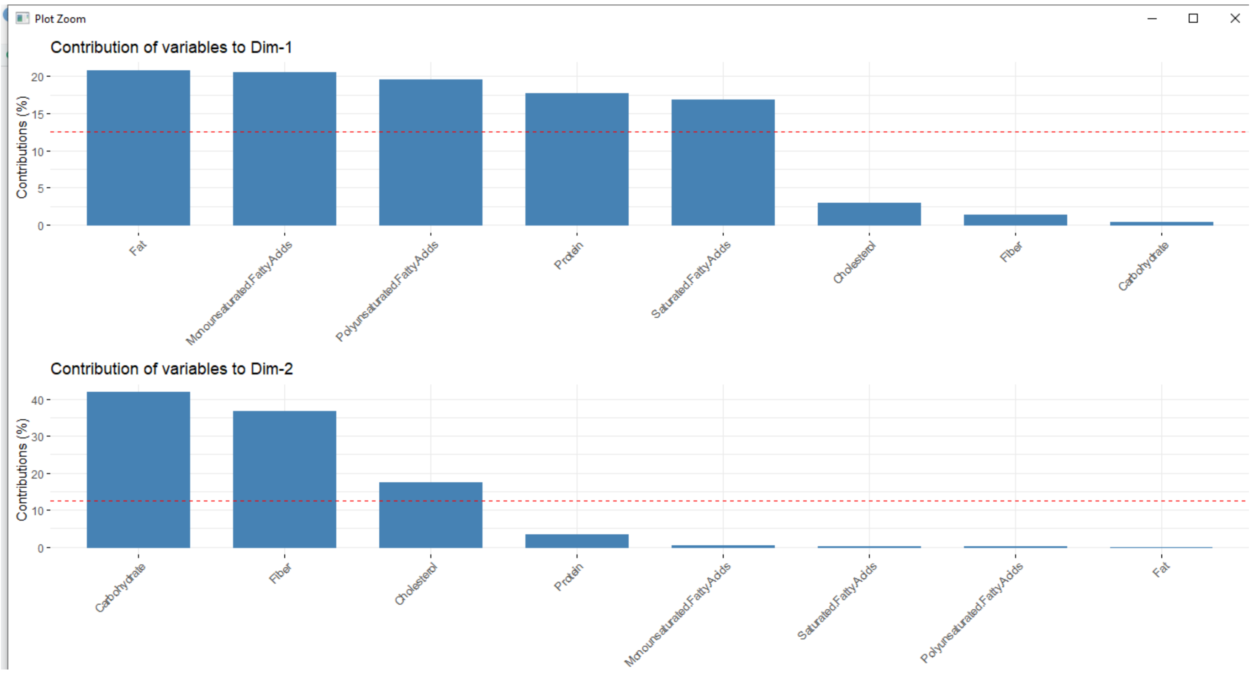


```
fviz_pca_var(pca, col.var = "red")
```



The plot displayed above shows relations between variables as well as the "quality" of all factors. Variables correlated positively are close to each other whereas those correlated negatively are on the opposite sites of the plot. "Quality" of the variable is presented by the distance from the center - “the best” variables. Just by looking on this graph, it’s hard to clearly distinguish the components.

```
library(gridExtra)
PC1 <- fviz_contrib(pca, choice = "var", axes = 1)
PC2 <- fviz_contrib(pca, choice = "var", axes = 2)
grid.arrange(PC1, PC2)
```



On the first plot it’s visible that the first component consists of Monounsaturated.Fatty.Acids, Fat, Protein, Saturated.Fatty.Acids. The second one consists of Carbohydrate, Fiber, Cholesterol.

Conclusions

Dimension reduction simply refers to the process of reducing the number of dimensions in a dataset. The aim of this process is to preserve as much information as possible by reducing the number of features. Conducted research shows that over 95% of the variance can be explained by only a half of the variables and 2 variables out of 9 are able to keep over 85% of the information included in the original dataset. Dimension reduction techniques are very powerful when it comes to analysis and storage of huge datasets.