R Notebook

1. Import the data file Cereals.dat in the R environment.

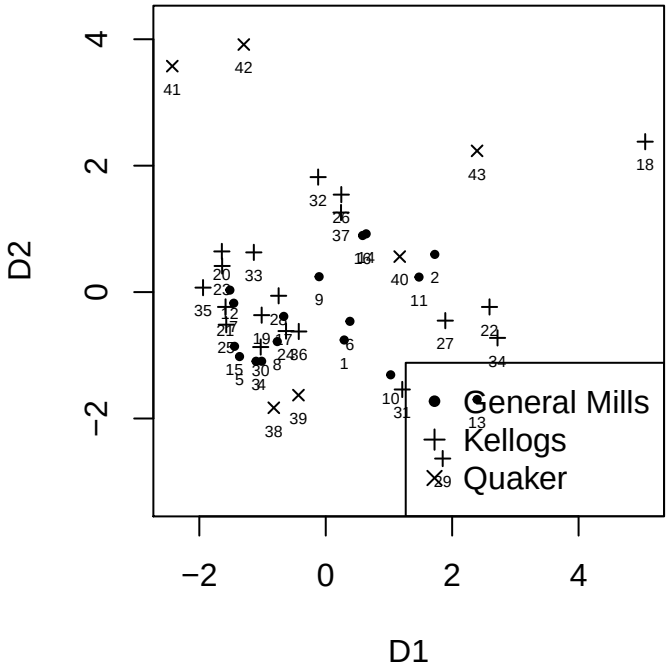
```
cereal_data <- as.data.frame(read.table("Cereals.dat", header = TRUE))
head(cereal_data, 5)
```

```
##           Brand Manufacturer Calories Protein Fat Sodium Fiber
## 1    ACCheerios            G      110       2   2    180   1.5
## 2      Cheerios            G      110       6   2    290   2.0
## 3    CocoaPuffs            G      110       1   1    180   0.0
## 4  CountChocula            G      110       1   1    180   0.0
## 5 GoldenGrahams            G      110       1   1    280   0.0
##   Carbohydrates Sugar Potassium
## 1          10.5    10        70
## 2          17.0     1       105
## 3          12.0    13        55
## 4          12.0    13        65
## 5          15.0     9        45
```

2. Compute the Euclidean distance matrix for the cereals, using the information on calories and all seven cereal components, and standardizing the variables prior to the calculation of the distance matrix. You can use the R functions scale and dist for this purpose. Paste the distance matrix of the first 5 specimens into your report.

```
##               ACCheerios Cheerios CocoaPuffs CountChocula GoldenGrahams
## ACCheerios      0.000000 4.389923  1.8800970    1.8678873      2.413378
## Cheerios        4.389923 0.000000  5.5151628    5.4964619      4.869285
## CocoaPuffs      1.880097 5.515163  0.0000000    0.1512638      1.700201
## CountChocula    1.867887 5.496462  0.1512638    0.0000000      1.720269
## GoldenGrahams   2.413378 4.869285  1.7002007    1.7202688      0.000000
```

3. Perform a metric MDS of the data, using the cmdscale program. Plot the two-dimensional solution, and label the cereals with a number or abbreviated name. Use a different colour or symbol to label each cereal according to its manufacturer.



The cereal brands are labeled 1 to 43 as they appear in the matrix of data.

4. Which pair of cereals is, according to the two-dimensional solution of the analysis, the most similar?

```
approx_dist <- as.matrix(dist(multidim_scale$points, upper = T))
approx_dist_non_matrix <- dist(multidim_scale$points)
which(approx_dist == min(approx_dist_non_matrix), arr.ind = T)
```

```
##                 row col
## Cheaties         16  14
## TotalWholeGrain  14  16
```

```
cheaties_manufacturer <- cereal_data$Manufacturer[[14]]
total_whole_grain_manufacturer <- cereal_data$Manufacturer[[16]]
cheaties_manufacturer
```

```
## [1] G
## Levels: G K Q
```

```
total_whole_grain_manufacturer
```

```
## [1] G
## Levels: G K Q
```

The most similar pair is **Cheaties** and **TotalWholeGrain**. It is worth noting that both are from the same manufacturer, **General Mills**. Just by looking at the MDS plot it can be seen that most of the manufacturers tend to cover a wide variety of the cereal spectrum, probably in order to appeal to a wider range of the

population. It's interesting to notice that **General Mills** follows a pattern of cereals characterized by dispersed clusters of two closely related cereal brands. It's not true for all of them, but a majority of them form a cluster with a clearly distinguished neighbor. In some cases, like **Cheaties** and **TotalWholeGrain** it seems it's due to a regular brand and a **dietetic** brand. It's not followed in all the close neighbor cases, so we cannot affirm certainly that this is the main motive for their existance.

5. Which pair of cereals would you, according to the two-dimensional solution of the analysis, classify as most distinct?

```
which(approx_dist == max(approx_dist), arr.ind = T)
```

```
##            row col
## PuffedRice  41  18
## AllBran     18  41
```

```
puffed_rice_manufacturer <- cereal_data$Manufacturer[[41]]
all_bran_manufacturer <- cereal_data$Manufacturer[[18]]
puffed_rice_manufacturer
```

```
## [1] Q
## Levels: G K Q
```

```
all_bran_manufacturer
```

```
## [1] K
## Levels: G K Q
```

The most distinct pair of cereals is **PuffedRice** and **AllBran**, manufactured by **Quakers** and **Kellogs** respectively. It seems reasonable that the highest differences are between different manufacturers, as it is more likely that they use different source materials and have different scopes and strategies to develop their brands.

6. Is it possible to find a configuration of the 43 cereals in k dimensions that will represent the original distance matrix exactly? Why or why not? If so, how many dimensions would be needed to obtain this exact representation?

Yes, as will be seen in the exercise **8.**, the matrix of scalar products is euclidean (semidefinite positive), so it admits exact representation. Furthermore, there are only 8 non-zero eigenvalues, so with 8 dimensions we get a full representation of the data. It makes sense that there are only 8 different dimensions, because there are only 8 different numerical variables that affect the distance. If the distances can be fully represented (the matrix of scalar products is euclidean), it makes sense that the trivial representation in 8D where each variable represents a dimension will be, in fact, a perfect representation.

7. Report the eigenvalues of the solution, and calculate the goodness-of-fit of the two-dimensional solution.

```
##  [1] 106.99792  77.90000  74.29086  36.46790  20.88663  15.01200   2.54114
##  [8]   1.90356   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
## [15]   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
## [22]   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
## [29]   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
## [36]   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
## [43]   0.00000
```

```
## [1] 0.5502914
```

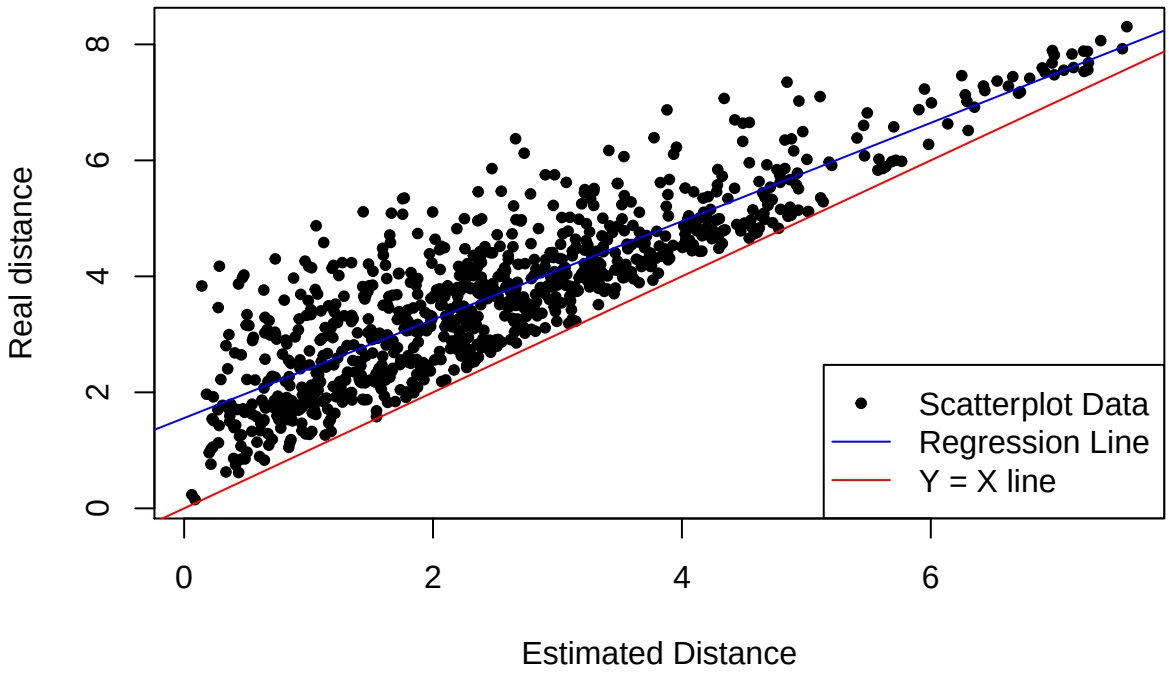
Above are the eigenvalus of the solution and the goodness-of-fit for two diensions. As can be seen the fit is quite poor, only 55%.

8. Are there any zero eigenvalues? Can you explain these?

Yes. The actual values obtained are not exactly 0, but when taking up until 5 decimal digits, they are approximated by 0. Furthermore, they are close to 10^{-15}, which means that they are highly likely 0 eigenvalues with some truncating errors. As said before, this is consistent with the fact that there are only 8 distinct numerical variables.

9. Compute the fitted distances according to the two-dimensional MDS solution. Graph fitted and observed distances and assess the goodness of fit by regression. What do you observe? Report the coefficient of determination of this regression.

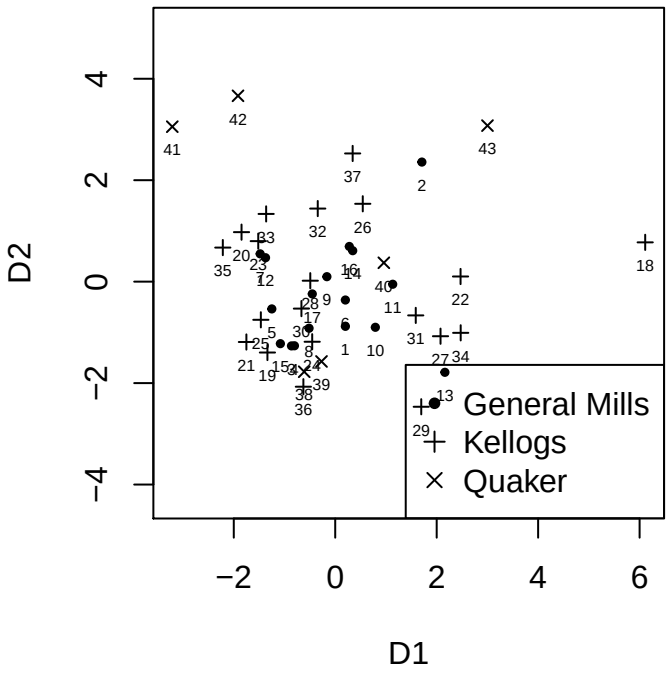
```
##
## Call:
## lm(formula = eucl_dist_no_matrix ~ approx_dist_no_matrix, data = helper_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4803 -0.5241 -0.1287  0.3820  2.5568
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.55673    0.04697   33.14   <2e-16 ***
## approx_dist_no_matrix  0.84830    0.01583   53.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7293 on 901 degrees of freedom
## Multiple R-squared:  0.7612, Adjusted R-squared:  0.7609
## F-statistic:  2871 on 1 and 901 DF,  p-value: < 2.2e-16
```



The coefficient of dtermination is 0.7609, which is slightly higher than the GOF obtained from exercise **7.** It can also be seen that the approximation tends to underpredict all the distance values. This underprediction is more severe in smaller distance values than higher ones, as more spread can be observe on the left part of the scatterplot.

10. Try now non-metric MDS with the isoMDS program. Plot the two-dimensional solution, labelling the points again with the name or number of the brand, and using different symbols for different manufacturers.

```
## initial  value 23.424101
## iter   5 value 17.264581
## final  value 17.035316
## converged
```



11. Which pair of cereals is, according to the two-dimensional solution of the non-metric analysis, most similar?

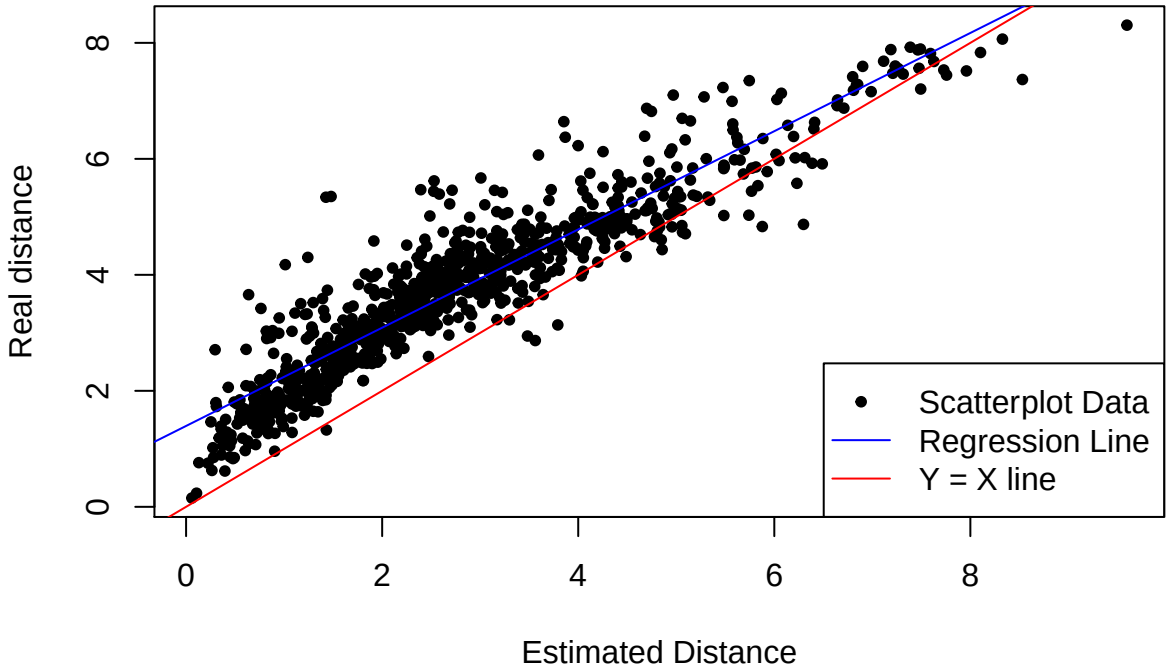
```
##              row col
## CountChocula   4   3
## CocoaPuffs     3   4
```

In this case, the closest products are **CountChocula** and **CocoaPuffs**, both from **General Mills** We would expect this result to correspond with the one obtained in the metric MDS. Non-metric MDS distances are not reliable absolute values by themselves, but they preserve the rank between distances. This discrepance is probably due to the bad representation of the metric MDS, specially in lower distance values (as seen in exercise **9.**). Again the two closest values are from **General Mills** and follow the pair cluster pattern.

12. Compute the fitted distances according to the two-dimensional non-metric MDS solution. Graph fitted and observed distances and assess the goodness of fit by regression. What do you observe? Report the coefficient of determination of this regression.

```
##
## Call:
## lm(formula = eucl_dist_no_matrix ~ approx_dist_non_metric_non_matrix)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85934 -0.35182 -0.05675  0.28835  2.73578
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.39442    0.03655   38.15   <2e-16 ***
## approx_dist_non_metric_non_matrix  0.84704    0.01148   73.80   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5622 on 901 degrees of freedom
## Multiple R-squared:  0.8581, Adjusted R-squared:  0.8579
## F-statistic:  5447 on 1 and 901 DF,  p-value: < 2.2e-16
```



The coefficient of determination is 0.8579, which is, surprisingly, higher than the one in metrid MDS. The lower distance values tend to be underpredicted, while the higher values tend to be overpredicted.

13. Compute the stress for a 1, 2, 3, 4 and 5 dimensional solution. How many dimensions do you think are necessary to obtain a "good fit"?

```
## initial  value 41.385754
## iter   5 value 30.785230
## final  value 30.602413
## converged
```

```
## initial  value 23.424101
## iter   5 value 17.264581
## final  value 17.035316
## converged
```

```
## initial  value 11.182343
## final  value 7.735158
## converged
```

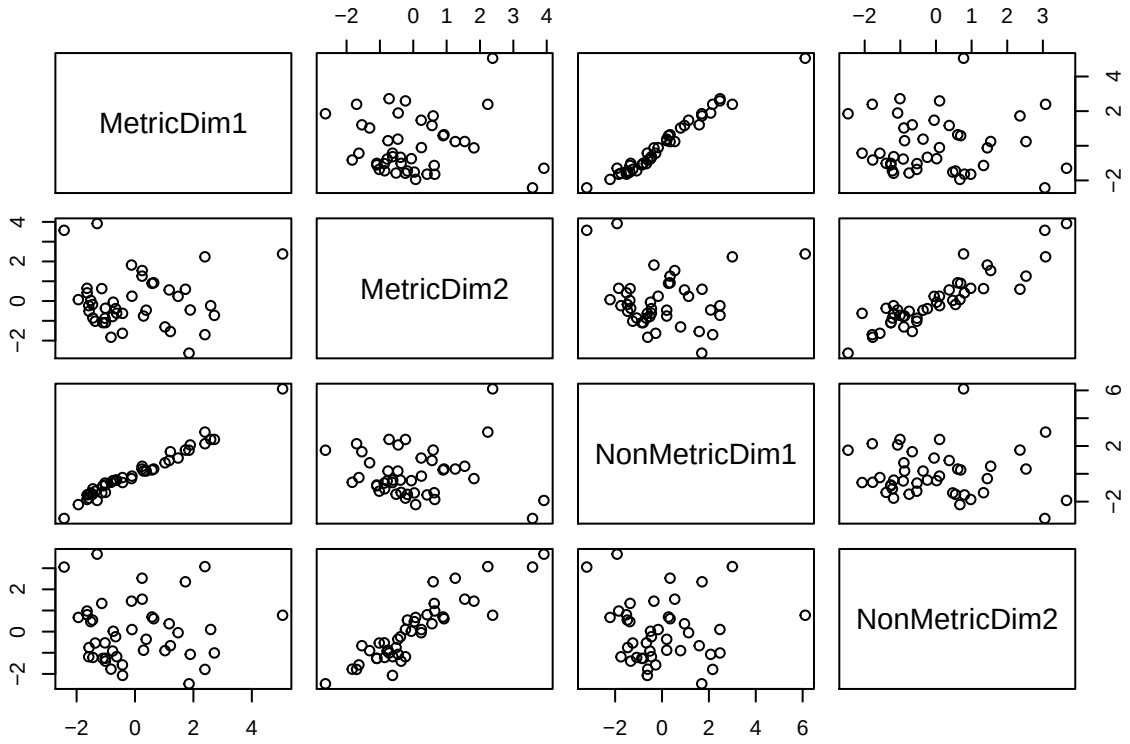
```
## initial  value 6.504622
## iter   5 value 3.880344
## final  value 3.812695
## converged
```

```
## initial  value 3.189552
## iter   5 value 2.361677
```

```
## iter  10 value 2.306782
## final  value 2.302776
## converged
```

Above are the results of running non-metric MDS until convergence for 1, 2, 3, 4, and 5 dimensions. We usually consider good fit under 5% stress, so we need 4 dimensions to obtain a good fit.

14. Make a scatterplot matrix of the first two dimensions of the metric MDS solution and the non-metric MDS solution (use k = 2). Calculate the correlation matrix of these four variables and comment on your results.




```
##                MetricDim1 MetricDim2 NonMetricDim1 NonMetricDim2
## MetricDim1          1.000      0.000         0.983        -0.028
## MetricDim2          0.000      1.000        -0.035         0.899
## NonMetricDim1       0.983     -0.035         1.000        -0.056
## NonMetricDim2      -0.028      0.899        -0.056         1.000
```

It can be seen that the first dimension of the Metric MDS highly corresponds to the first dimension of the non-metric MDS, and the same behavior is seen between the second dimensions of both analysis. Furthermore there's almost complete correlation between the first dimensions, so it can be inferred that most of the improvement of the non-metric analysis with respect to the metric one is due to a better approximation/choice of the second dimension.

15. We have used the Euclidean distance as a metric in this exercise, on the standardized variables. We could also have calculated the Euclidean distance matrix without prior standardization of the variables. Which approach do you think is preferable? Argue your answer.

The different numerical variables don't have comparable units. Variables as **Calories** and **Sodium** have values orders of magnitude higher than **Protein** or **Fat**, and have higher standard deviations too. This means that most of the contribution to the distance would come from these bigger variables. Thus smaller variables would probably be more misrepresented than bigger ones.

So, it's probably preferable to standardize the variables prior to the analysis.