**Practice Assignment I**

Business Analysts at a large global Asset Management firm (such as [Vanguard](https://investor.vanguard.com/home/), J P Morgan etc) have been asked by the Portfolio Management team to help build a risk model for the Dow Jones Industrial Average.  The Dow Jones Industrial Average is a basket of 30 large publicly traded stocks.  The Business Analysts would like get an idea of common risks between the stocks. Using the daily returns on the stocks in the years 2012 – 2015, develop a factor analysis model to determine if there exit any common factors.  Interpret and explain your results. The data is provided separately (DowJones.csv).

**Solution:**

1. All analyses must begin with exploration of the data. Typically univariate exploration includes computation of mean, median and standard deviations of the variables and bivariate analyses include computation of pairwise correlations. Graphical analyses include construction of histograms and scatterplots.

In this instance, there are 30 variables. Hence plots may be cumbersome. Detail outputs are given in Appendix along with R codes.

Important Observations:

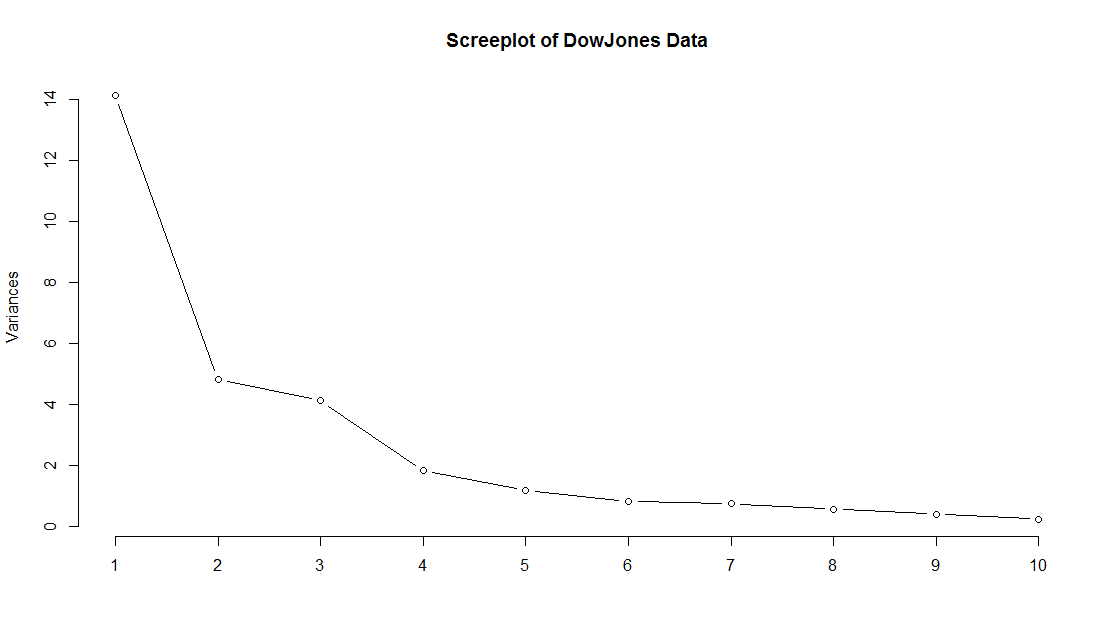
* Variances of the indices are widely different. Hence scaling of variables necessary for further analysis
* Variables are not necessarily normally distributed.
* From the scatterplot it is evident that several variables are strongly correlated.
* Several pairwise correlations are over 70%.

Hence this data set is a prime candidate for factor analysis.

1. Principal component analysis is done on the scaled data. Partial summary of the results are given below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Standard deviation | 3.757 | 2.190 | 2.029 | 1.354 | 1.087 | 0.909 | 0.859 |
| Proportion of variance | 0.470 | 0.160 | 0.137 | 0.061 | 0.039 | 0.028 | 0.025 |
| Cumulative proportion | 0.470 | 0.630 | 0.768 | 0.829 | 0.868 | 0.896 | 0.920 |

Screeplot of the variances are shown in the following graph.



Important Observations:

* In the scree plot the elbow effect is not pronounced
* After PC5 the eigenvalue falls below 1
* The first 5 principal components explain 87% of the total variances. However the first 6 components explain 90%, even though the eigenvalue is less than 1.

Eigenvalue is the variance of the corresponding PC. Square root of the eigenvalue is the sd(PC) which is obtained from R output. For factor extraction either 6 or 7 factors need to be considered.

1. Investigating the correlation between the variables and the first 7 components (see Appendix) it is observed that PC1 has high correlation (≥ 0.7) with 17 of the 30 indices, PC2 has high correlation with only one index and PC3 has high correlation with 2 indices. The other PCs do not have high correlation with any indices at all. This does not render well to proper interpretation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| MMM | 0.95 | 0.13 | -0.16 | 0.09 | -0.04 | 0.07 | -0.12 |
| T | 0.40 | -0.22 | 0.76 | -0.32 | -0.09 | -0.17 | 0.04 |
| AXP | 0.87 | 0.22 | 0.15 | 0.17 | 0.08 | 0.00 | -0.25 |
| BA | 0.65 | -0.34 | -0.47 | 0.39 | -0.11 | 0.01 | -0.10 |
| CAT | -0.08 | 0.87 | -0.01 | -0.11 | 0.32 | -0.03 | 0.06 |
| CVX | 0.76 | 0.46 | 0.05 | 0.31 | -0.16 | -0.18 | 0.11 |
| CSCO | 0.46 | 0.37 | -0.55 | -0.12 | 0.16 | 0.11 | 0.45 |
| KO | 0.75 | -0.04 | 0.27 | -0.40 | -0.15 | -0.27 | -0.06 |
| DD | 0.82 | 0.35 | -0.23 | 0.15 | 0.00 | -0.10 | -0.19 |
| XOM | 0.74 | 0.45 | 0.29 | 0.20 | -0.03 | -0.14 | 0.02 |
| GE | 0.64 | -0.64 | 0.10 | 0.11 | 0.09 | -0.24 | -0.01 |
| GS | 0.80 | -0.11 | 0.15 | 0.20 | 0.47 | 0.09 | 0.06 |
| HD | 0.63 | -0.50 | 0.36 | -0.32 | 0.18 | 0.10 | 0.03 |
| INTC | 0.41 | 0.63 | -0.34 | -0.49 | 0.08 | -0.12 | -0.11 |
| IBM | 0.06 | 0.31 | 0.71 | 0.26 | 0.23 | -0.24 | 0.31 |
| JPM | 0.78 | -0.24 | 0.03 | 0.14 | 0.50 | 0.00 | 0.12 |
| JNJ | 0.84 | 0.28 | -0.34 | -0.04 | -0.21 | 0.03 | 0.06 |
| MCD | 0.04 | -0.38 | -0.71 | 0.12 | -0.12 | -0.46 | 0.22 |
| MRK | 0.66 | 0.47 | 0.08 | -0.24 | -0.25 | 0.25 | 0.22 |
| MSFT | 0.72 | 0.19 | -0.34 | -0.42 | 0.18 | -0.21 | -0.13 |
| NKE | 0.49 | -0.50 | -0.64 | -0.08 | 0.13 | -0.09 | -0.14 |
| PFE | 0.64 | -0.59 | 0.06 | 0.24 | -0.20 | 0.17 | 0.16 |
| PG | 0.84 | 0.39 | -0.15 | 0.07 | -0.10 | 0.06 | -0.03 |
| TRV | 0.88 | -0.13 | 0.06 | -0.22 | -0.11 | 0.02 | 0.00 |
| UTX | 0.80 | 0.23 | -0.07 | 0.49 | 0.03 | 0.08 | -0.08 |
| UNH | 0.51 | -0.25 | -0.61 | -0.21 | 0.02 | 0.28 | 0.14 |
| VZ | 0.80 | -0.21 | 0.30 | -0.01 | -0.35 | -0.07 | 0.15 |
| V | 0.72 | -0.49 | 0.28 | -0.14 | 0.20 | 0.13 | -0.12 |
| WMT | 0.74 | 0.30 | 0.50 | 0.05 | -0.15 | 0.17 | -0.14 |
| DIS | 0.80 | -0.41 | 0.12 | -0.22 | -0.03 | 0.03 | 0.13 |

1. Extraction of factors is the next stage of the analysis. As mentioned before either a 6-factor model or a 7-factor model is to be considered.

The table below compares the communality arrived at in 6-factor and 7-factor model.

|  |  |  |
| --- | --- | --- |
|  | Communality | |
|  | 6-Factor | 7-Factor |
| MMM | 0.96 | 0.98 |
| T | 0.93 | 0.93 |
| AXP | 0.87 | 0.93 |
| BA | 0.92 | 0.93 |
| CAT | 0.89 | 0.89 |
| CVX | 0.95 | 0.96 |
| CSCO | 0.7 | 0.9 |
| KO | 0.9 | 0.91 |
| DD | 0.88 | 0.92 |
| XOM | 0.89 | 0.89 |
| GE | 0.92 | 0.92 |
| GS | 0.94 | 0.94 |
| HD | 0.93 | 0.93 |
| INTC | 0.93 | 0.94 |
| IBM | 0.78 | 0.87 |
| JPM | 0.93 | 0.95 |
| JNJ | 0.95 | 0.95 |
| MCD | 0.89 | 0.94 |
| MRK | 0.84 | 0.89 |
| MSFT | 0.92 | 0.93 |
| NKE | 0.93 | 0.95 |
| PFE | 0.89 | 0.91 |
| PG | 0.9 | 0.9 |
| TRV | 0.85 | 0.85 |
| UTX | 0.95 | 0.96 |
| UNH | 0.82 | 0.84 |
| VZ | 0.89 | 0.91 |
| V | 0.91 | 0.93 |
| WMT | 0.94 | 0.96 |
| DIS | 0.87 | 0.89 |

Except for the index CSCO 6-factor model is performing well. For CSCO the 6 factors are able to explain 70% of the variability. With 7-factor model the minimum communality achieved is 84% for UNH.

Final conclusion: To use 7 factor model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | RC1 | RC4 | RC3 | RC2 | RC5 | RC6 | RC7 |
| MMM | 0.83 | 0.34 | 0.28 | 0.12 | 0.25 | 0.03 | 0.12 |
| T | 0.04 | 0.85 | -0.38 | 0.02 | -0.04 | -0.18 | -0.12 |
| AXP | 0.82 | 0.34 | 0.01 | 0.13 | 0.31 | -0.13 | -0.1 |
| BA | 0.55 | 0.15 | 0.52 | -0.28 | 0.29 | 0.41 | 0.01 |
| CAT | 0.21 | -0.48 | -0.39 | 0.6 | 0.03 | -0.27 | 0.2 |
| CVX | 0.92 | 0.13 | -0.23 | 0.07 | 0.04 | 0.12 | 0.16 |
| CSCO | 0.38 | -0.13 | 0.18 | 0.35 | 0.2 | 0.19 | 0.71 |
| KO | 0.42 | 0.79 | -0.04 | 0.32 | -0.05 | 0.05 | -0.02 |
| DD | 0.87 | 0.11 | 0.17 | 0.28 | 0.19 | 0.1 | 0.02 |
| XOM | 0.82 | 0.25 | -0.34 | 0.14 | 0.12 | -0.07 | 0.05 |
| GE | 0.2 | 0.68 | 0.13 | -0.22 | 0.41 | 0.38 | -0.16 |
| GS | 0.48 | 0.4 | -0.03 | -0.01 | 0.73 | -0.03 | 0.11 |
| HD | 0.04 | 0.87 | 0.1 | -0.02 | 0.37 | -0.12 | 0.04 |
| INTC | 0.42 | -0.03 | 0.15 | 0.83 | -0.09 | -0.05 | 0.22 |
| IBM | 0.15 | 0.09 | -0.89 | -0.06 | 0.17 | -0.12 | -0.01 |
| JPM | 0.38 | 0.44 | 0.04 | 0.01 | 0.76 | 0.13 | 0.13 |
| JNJ | 0.79 | 0.22 | 0.29 | 0.23 | 0 | 0.13 | 0.35 |
| MCD | -0.06 | -0.07 | 0.36 | -0.04 | -0.01 | 0.89 | 0.11 |
| MRK | 0.62 | 0.3 | -0.02 | 0.23 | -0.17 | -0.28 | 0.5 |
| MSFT | 0.45 | 0.32 | 0.3 | 0.67 | 0.2 | 0.17 | 0.14 |
| NKE | 0.13 | 0.25 | 0.72 | 0.09 | 0.35 | 0.48 | 0.01 |
| PFE | 0.31 | 0.59 | 0.26 | -0.55 | 0.24 | 0.17 | 0.13 |
| PG | 0.87 | 0.17 | 0.13 | 0.21 | 0.1 | -0.03 | 0.23 |
| TRV | 0.52 | 0.69 | 0.21 | 0.12 | 0.13 | 0.02 | 0.16 |
| UTX | 0.9 | 0.06 | 0.05 | -0.11 | 0.35 | 0.03 | 0.05 |
| UNH | 0.18 | 0.21 | 0.68 | 0.08 | 0.23 | 0.18 | 0.46 |
| VZ | 0.54 | 0.75 | -0.04 | -0.19 | -0.04 | 0.09 | 0.13 |
| V | 0.2 | 0.78 | 0.2 | -0.08 | 0.46 | -0.11 | -0.07 |
| WMT | 0.73 | 0.45 | -0.21 | 0.01 | 0.04 | -0.43 | -0.01 |
| DIS | 0.29 | 0.81 | 0.2 | -0.03 | 0.25 | 0.09 | 0.19 |

1. Final interpretation: There is strong evidence that the 30 DowJones indices can be explained by a set of 7 common factors. Of the total variability in the data, about 92% is explained by the factors. Communality is high for all the indices, indicating that individual factors are modelled well by the factors.

It is also possible to model each index as a function of a subset of the factors. For example, if we agree to use a threshold cut-off at 70%

* MMM, AXP, CVX are functions of Factor 1 only
* T, KO, XD are functions of Factor 2 only

However a few indices, such as TRV, are functions of multiple factors, since they show medium dependency on Factor 1 and Factor 2.

The usefulness of factor analysis for portfolio managers is that, instead of modeling each of the 30 indices separately to predict the behavior of Dow Jones, it is enough to deal with the seven common factors.

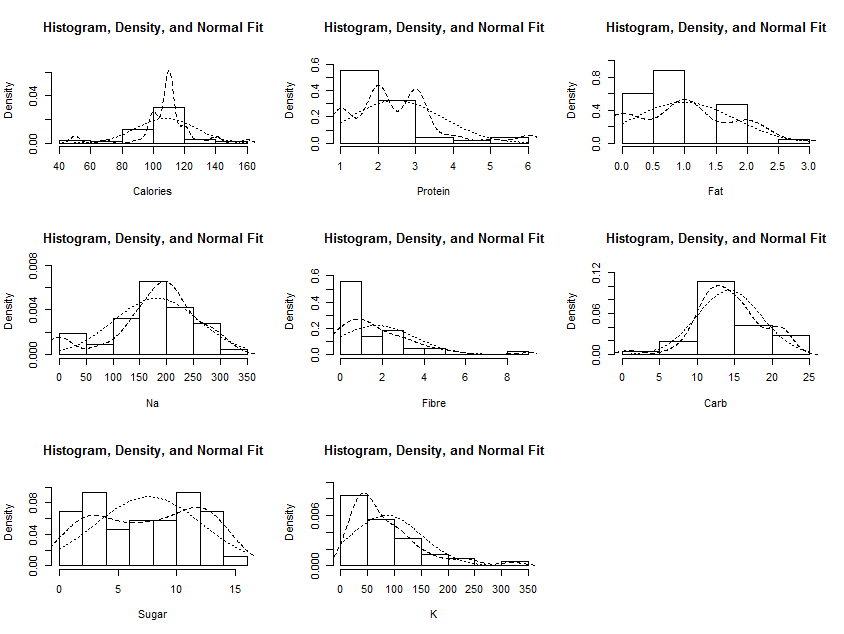
**Practice Assignment II**

Breakfast cereal market is big business in the US. There are a number of giants in the boxed cereal business and every one of them claim that theirs is the best nutritious. To refute (or otherwise) their claims a consumer welfare group decided to check whether there is really any difference among the cereals. They used the data collected on 43 cereals by three manufacturers and wanted to check if there is any reason to believe that cereals produced by different manufacturers are, in fact, different. Develop a linear discriminant analysis to check whether the cereals possess good classification property. The data for model development is provided in the file CerealData.csv. After the model is developed, its predictive ability can be tested on a different data set NewCerealData.csv.

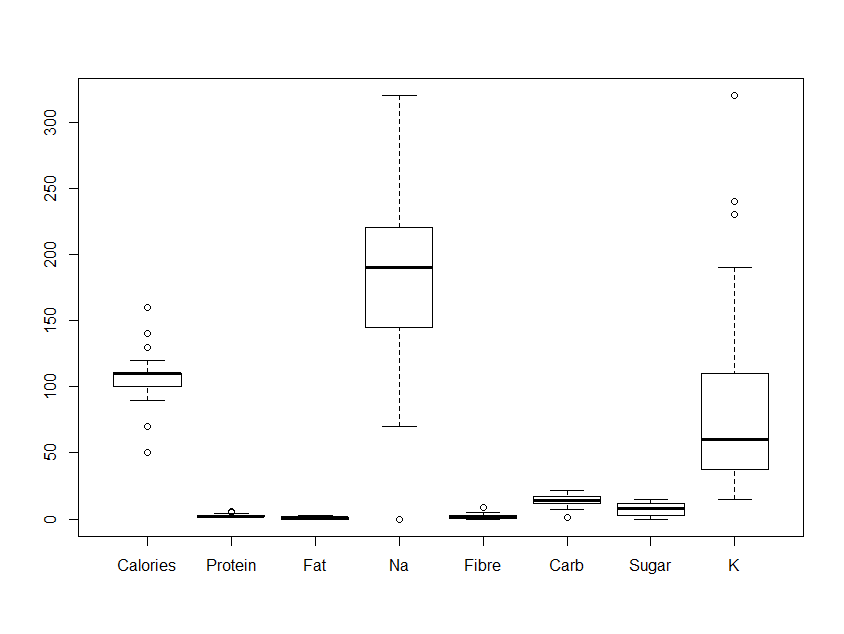
**Solution:**

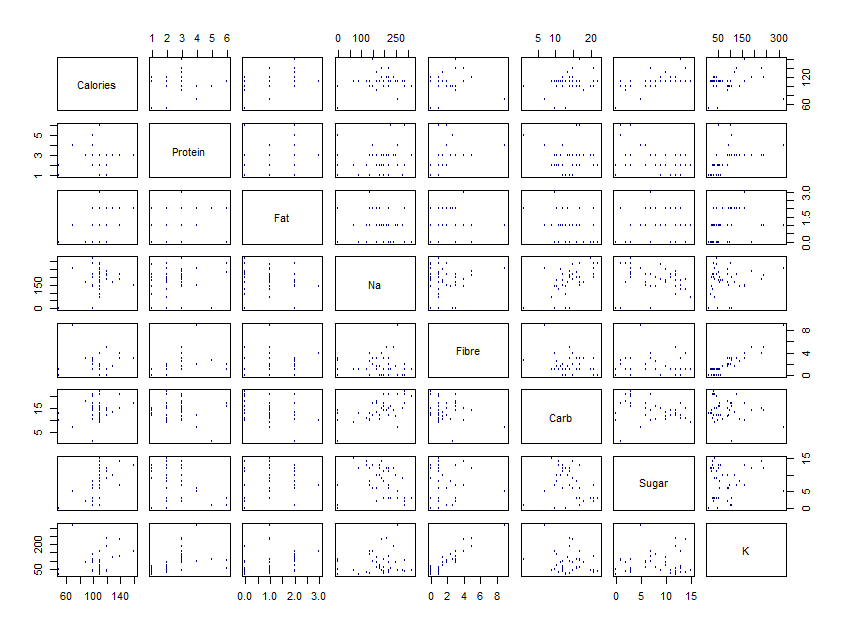
1. All analyses must begin with exploration of the data. Typically univariate exploration includes computation of mean, median and standard deviations of the variables and bivariate analyses include computation of pairwise correlations. Graphical analyses include construction of histograms and scatterplots. In this instance, there are 8 variables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | vars | n | mean | sd | median | min | max |
| Calories | 1 | 43 | 107.91 | 18.97 | 110 | 50 | 160 |
| Protein | 2 | 43 | 2.47 | 1.22 | 2 | 1 | 6 |
| Fat | 3 | 43 | 0.98 | 0.8 | 1 | 0 | 3 |
| Na | 4 | 43 | 180.47 | 79.21 | 190 | 0 | 320 |
| Fibre | 5 | 43 | 1.71 | 1.8 | 1 | 0 | 9 |
| Carb | 6 | 43 | 14.26 | 4.26 | 14 | 1 | 22 |
| Sugar | 7 | 43 | 7.6 | 4.54 | 8 | 0 | 15 |
| K | 8 | 43 | 84.42 | 66.11 | 60 | 15 | 320 |



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Calories | Protein | Fat | Na | Fibre | Carb | Sugar | K |
| Calories | 1.00 | 0.03 | 0.39 | 0.34 | -0.02 | 0.26 | 0.58 | 0.14 |
| Protein | 0.03 | 1.00 | 0.21 | 0.09 | 0.51 | -0.08 | -0.40 | 0.50 |
| Fat | 0.39 | 0.21 | 1.00 | 0.01 | 0.16 | -0.32 | 0.19 | 0.31 |
| Na | 0.34 | 0.09 | 0.01 | 1.00 | 0.04 | 0.57 | -0.05 | 0.11 |
| Fibre | -0.02 | 0.51 | 0.16 | 0.04 | 1.00 | -0.24 | -0.03 | 0.93 |
| Carb | 0.26 | -0.08 | -0.32 | 0.57 | -0.24 | 1.00 | -0.32 | -0.22 |
| Sugar | 0.58 | -0.40 | 0.19 | -0.05 | -0.03 | -0.32 | 1.00 | 0.08 |
| K | 0.14 | 0.50 | 0.31 | 0.11 | 0.93 | -0.22 | 0.08 | 1.00 |





Important Observations:

* Variances of the nutritive factors are widely different. Hence scaling of variables necessary for further analysis
* Variables are not necessarily normally distributed.
* Pairwise correlation does not seem to be a problem in this data

1. A linear discriminant model is fit. Results shown below.

> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")

> ldfit.c <- lda(Manufacturer ~ ., data=CD1)

> ldfit.c

Call:

lda(Manufacturer ~ ., data = CD1)

Prior probabilities of groups:

G K Q

0.3953488 0.4651163 0.1395349

Group means:

Calories Protein Fat Na Fibre Carb Sugar K

G 110.5882 2.352941 1.235294 203.52941 1.294118 14.58824 8.117647 85.00000

K 111.0000 2.600000 0.650000 185.50000 2.250000 15.25000 7.950000 91.75000

Q 90.0000 2.333333 1.333333 98.33333 1.116667 10.00000 5.000000 58.33333

Coefficients of linear discriminants:

LD1 LD2

Calories 0.022344104 0.045417606

Protein 0.369109646 -0.332405063

Fat -0.837675738 -0.386499597

Na -0.000763493 -0.006017311

Fibre 1.420281838 1.039957871

Carb 0.202200109 -0.203863959

Sugar 0.195246015 -0.235306430

K -0.030687468 -0.026966644

Proportion of trace:

LD1 LD2

0.7745 0.2255

> prdcd <- predict(ldfit.c, newdata=CD1)

> table(prdcd$class, CD1$Manufacturer)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Manufacturer | | | |
| Predicted Class |  | G | K | Q |
| G | 15 | 3 | 3 |
| K | 2 | 16 | 0 |
| Q | 0 | 1 | 3 |

Important Observations:

* Observed prior is used. Pr(G) = 39.5%, Pr(K) = 46.5%, Pr(Q) =14%
* Means of the different nutritive quantities show marked difference across manufacturers. Eg. G and K do not differ substantially in calories, carb or sugar; but show difference in fat and Na (sodium). Q is very different in calories and K from both G and K. Other differences also exist.
* Comparison of predicted class (Manufacturer) and actual class is given in the table above. Out of 43 observations 34 has been correctly classified. Hence misclassification rate is 9/43 = 21%.

1. Performance of the model is checked using a new set of 53 cereals from the same 3 manufacturers.

> View(NewCerealData)

> prdcdnew <- predict(ldfit.c, newdata=NewCerealData)

> table(prdcdnew$class, NewCerealData$Manufacturer)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Manufacturer | | | |
| Predicted Class |  | G | K | Q |
| G | 19 | 3 | 4 |
| K | 3 | 19 | 0 |
| Q | 0 | 1 | 4 |

Important Observations:

* Misclassification probability is 20%. The model is robust in the sense that the predictability of the model holds in a new set.
* Posterior probabilities (shown in the appendix) show unambiguous classification. That is to say, posterior probability of belonging to one class is very high compared to the other two. In only a few instances posterior probabilities of belonging to two different classes are close.

1. Final interpretation: The linear discriminant analysis has limited performance in the data. That is possibly due to the fact that the differences are more among cereal variants, not among the manufacturers.

**Appendix I:**

In this section the R commands and raw outputs are shown for PCA and FA.

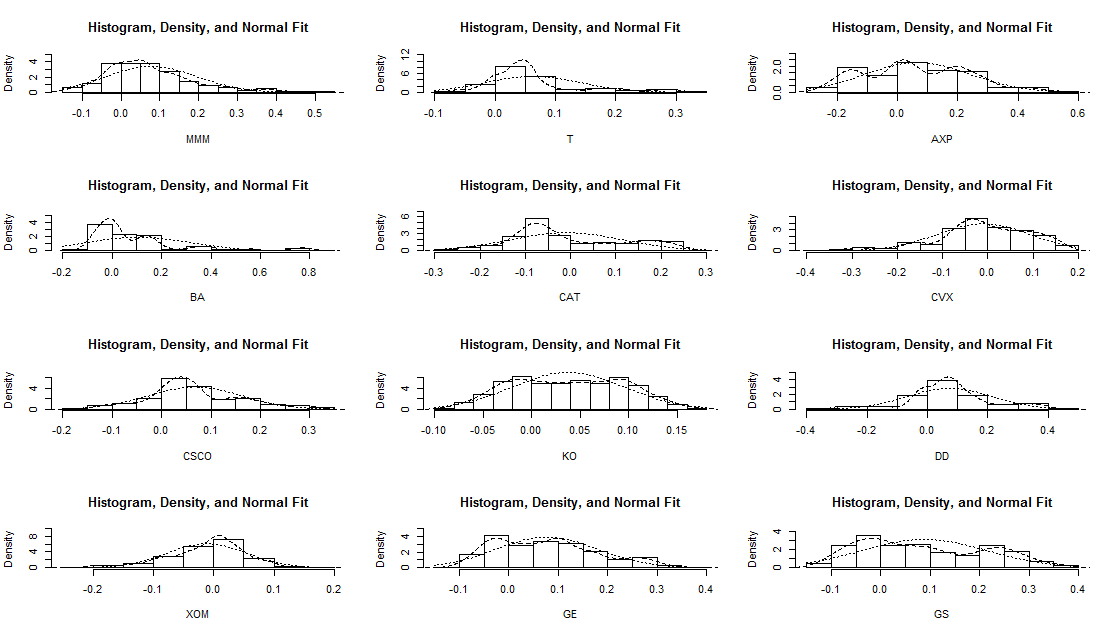
> library("psych", lib.loc="~/R/win-library/3.3")

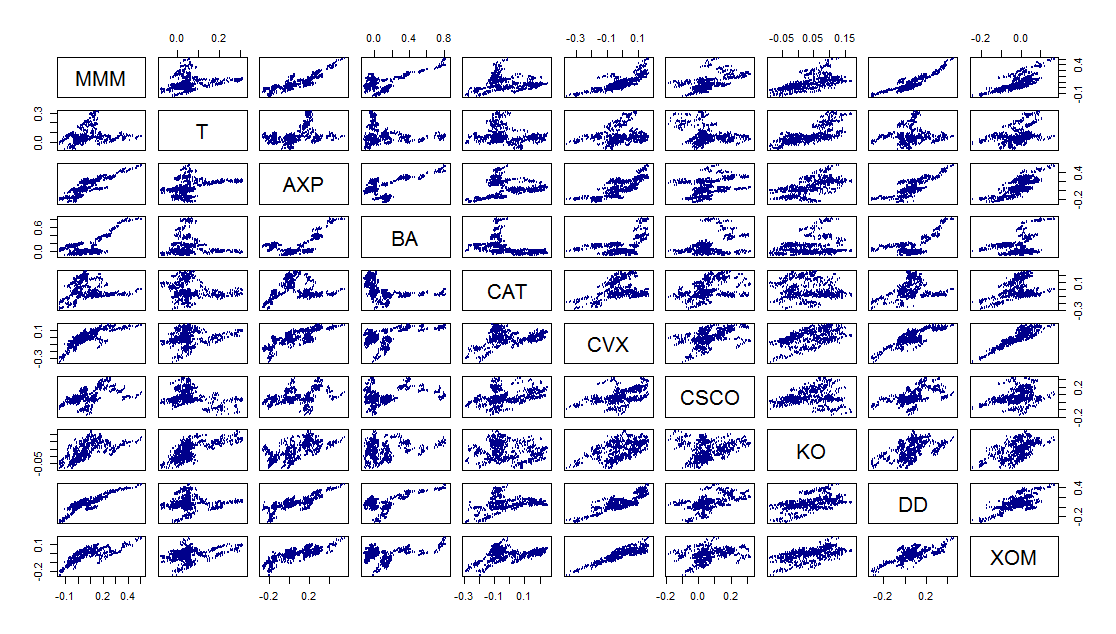
> describe(DowJones[,2:31])

> multi.hist(DowJones[,2:13])

> plot(DowJones[,2:11], pch=21, col="darkblue", cex=0.4)

Consult DowJones.xls for output of describe() and cor().





> DJpc1 <- prcomp(DowJones[,2:31], scale=T)

> summary(DJpc1)

> plot(DJpc1, type="l", mtext(""))

> title("Screeplot of DowJones Data", cex=0.6)

> DJpc1

> DJ1 <- scale.default(DowJones[,2:31])

> cor(DJ1, DJpc1$x)

Consult DowJones.xls for output of prcomp() and cor().

> DJfact6 <- principal(DJ1, nfactors=6, rotate="varimax")

> DJfact6

> DJfact7 <- principal(DJ1, nfactors=7, rotate="varimax")

> DJfact7

**Appendix II:**

In this section the R command and raw output are shown for LDA

> prdcdnew$posterior

|  |  |  |  |
| --- | --- | --- | --- |
|  | G | K | Q |
| 1 | 0.02 | 0.00 | 0.98 |
| 2 | 0.13 | 0.87 | 0.00 |
| 3 | 0.00 | 1.00 | 0.00 |
| 4 | 0.72 | 0.10 | 0.18 |
| 5 | 0.01 | 0.99 | 0.00 |
| 6 | 0.44 | 0.56 | 0.00 |
| 7 | 0.88 | 0.02 | 0.10 |
| 8 | 0.92 | 0.08 | 0.00 |
| 9 | 0.54 | 0.00 | 0.46 |
| 10 | 0.81 | 0.07 | 0.12 |
| 11 | 0.96 | 0.02 | 0.03 |
| 12 | 0.11 | 0.89 | 0.00 |
| 13 | 0.01 | 0.99 | 0.00 |
| 14 | 0.97 | 0.01 | 0.02 |
| 15 | 0.46 | 0.06 | 0.48 |
| 16 | 0.02 | 0.98 | 0.00 |
| 17 | 0.89 | 0.05 | 0.06 |
| 18 | 0.09 | 0.91 | 0.00 |
| 19 | 0.02 | 0.98 | 0.00 |
| 20 | 0.00 | 1.00 | 0.00 |
| 21 | 0.00 | 1.00 | 0.00 |
| 22 | 0.96 | 0.02 | 0.02 |
| 23 | 0.66 | 0.21 | 0.13 |
| 24 | 0.84 | 0.15 | 0.01 |
| 25 | 0.58 | 0.40 | 0.03 |
| 26 | 0.01 | 0.99 | 0.00 |
| 27 | 0.93 | 0.06 | 0.01 |
| 28 | 0.84 | 0.07 | 0.09 |
| 29 | 0.96 | 0.02 | 0.02 |
| 30 | 0.08 | 0.92 | 0.00 |
| 31 | 0.67 | 0.30 | 0.03 |
| 32 | 0.81 | 0.16 | 0.03 |
| 33 | 0.07 | 0.93 | 0.00 |
| 34 | 0.01 | 0.99 | 0.00 |
| 35 | 0.95 | 0.03 | 0.02 |
| 36 | 0.21 | 0.79 | 0.00 |
| 37 | 0.03 | 0.00 | 0.97 |
| 38 | 0.02 | 0.00 | 0.98 |
| 39 | 0.77 | 0.21 | 0.01 |
| 40 | 0.00 | 0.00 | 1.00 |
| 41 | 0.40 | 0.60 | 0.00 |
| 42 | 0.84 | 0.01 | 0.15 |
| 43 | 0.42 | 0.53 | 0.06 |
| 44 | 0.60 | 0.40 | 0.00 |
| 45 | 0.14 | 0.86 | 0.01 |
| 46 | 0.08 | 0.92 | 0.00 |
| 47 | 0.85 | 0.13 | 0.02 |
| 48 | 0.52 | 0.48 | 0.00 |
| 49 | 0.18 | 0.81 | 0.01 |
| 50 | 0.97 | 0.01 | 0.01 |
| 51 | 0.73 | 0.23 | 0.04 |
| 52 | 0.16 | 0.84 | 0.00 |
| 53 | 0.62 | 0.37 | 0.01 |