Case Study Alexis Laks

Alexis Laks

Presentation of the case study

EasyKost

A common approach to determine the cost of products is the **should cost** method. It consists in estimating what a product should cost based on materials, labor, overhead, and profit margin. Although this strategy is very accurate, it has the drawback of being tedious and it requires expert knowledge of industrial technologies and processes. To get a quick estimation, it is possible to build a statistical model to predict the cost of products given their characteristics. With such a model, it would no longer be necessary to be an expert or to wait several days to assess the impact of a design modification, a change in supplier or a change in production site. Before builing a model, it is important to explore the data which is the aim of this case study.

Die Casting

This study was carried out for a company that sells parts for the car industry. They build many parts themselves, but because they don't have foundries, they don't make die-cast parts and they need to buy them. To bid on tenders, they usually ask their supplier how much the die-cast part will cost them. However, suppliers may take time to respond and the company may lose the tender. Therefore, they want to try to use the data to estimate the price of die-casting accurately and quickly without consulting the supplier, and thus be able to respond to the call for tenders.

Some explanation for some variables. "EXW cost" : unit price, (ex-works price: no transport) "Yearly Volume": Annual order volume: number of items ordered.

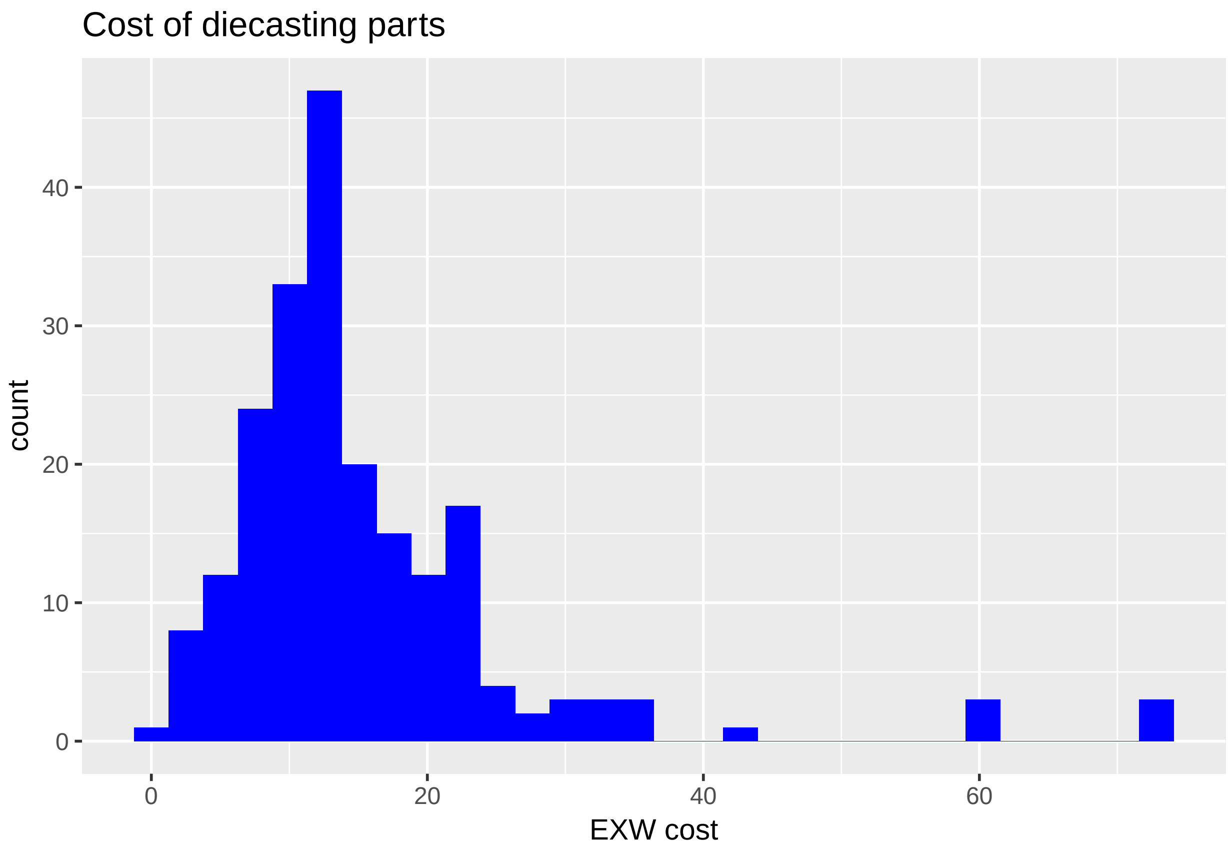
This allows for an identical line in the data except for the volume to have a different price, since in general, the purchase volume is an important cost-driver.

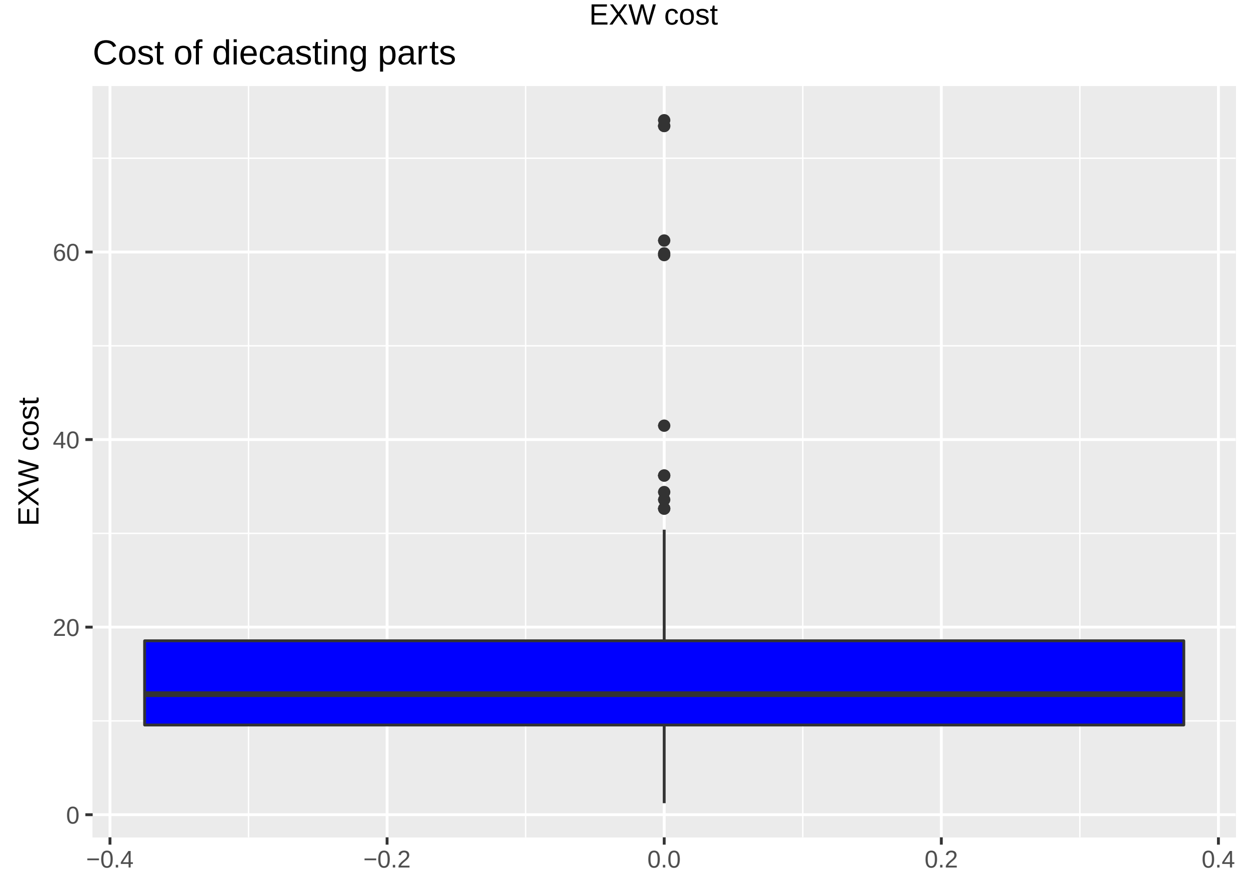
1) Import and summarize the data.

the diecasting dataset is a dataframe containing 19 variables and 211 observations, the varaibles are information on these 211 diecasting parts from various suppliers. Information on these parts range from where the part came from to how it was cooled, so we have a vast amount of info on each part. We have both quantitative and qualitative variable, an important feature to keep in mind when we go further in our analysis.

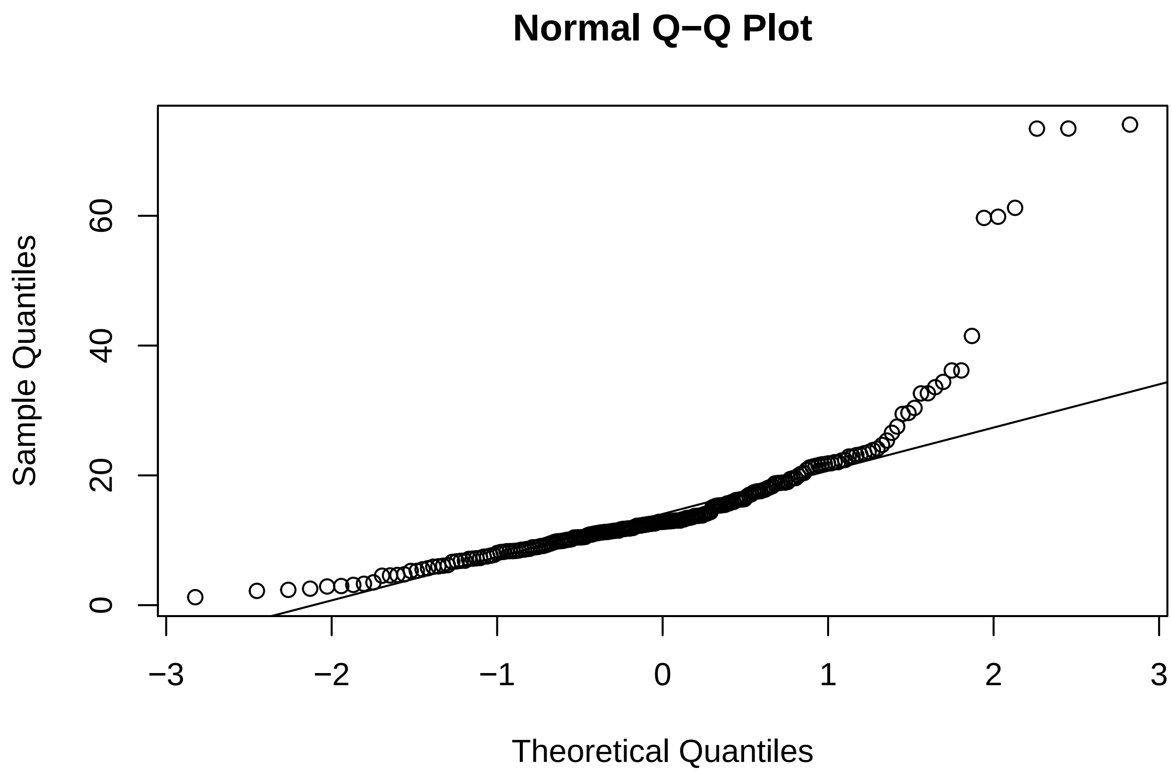
2) We start with univariate and bivariate descriptive statistics. Using appropriate plot(s) or summaries answer the following questions

2.1 How is the distribution of the cost? Comment your plot with respect to the quartiles of the cost.





The histogram above seems to ressemble a skewed normal distribution, with data centeredaround ~17 let's check the quantiles if they match.



Although the tails are heavy (to be expected when we have skewness) the distribution does seem to be approximately normal.

2.2 Which are the most frequent suppliers?

The most frequent suppliers are those with the biggest yearly volume. We thus have:

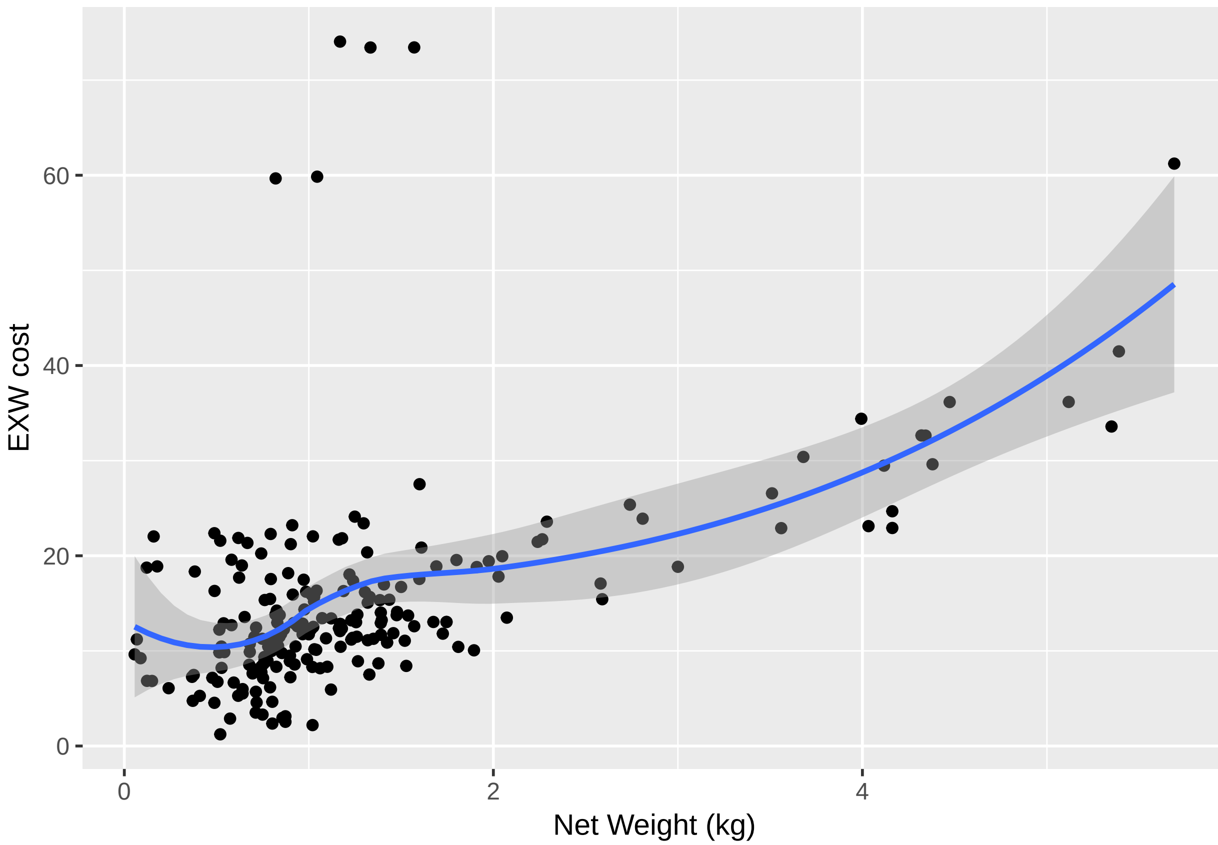
```
## [1] 6e+05
```

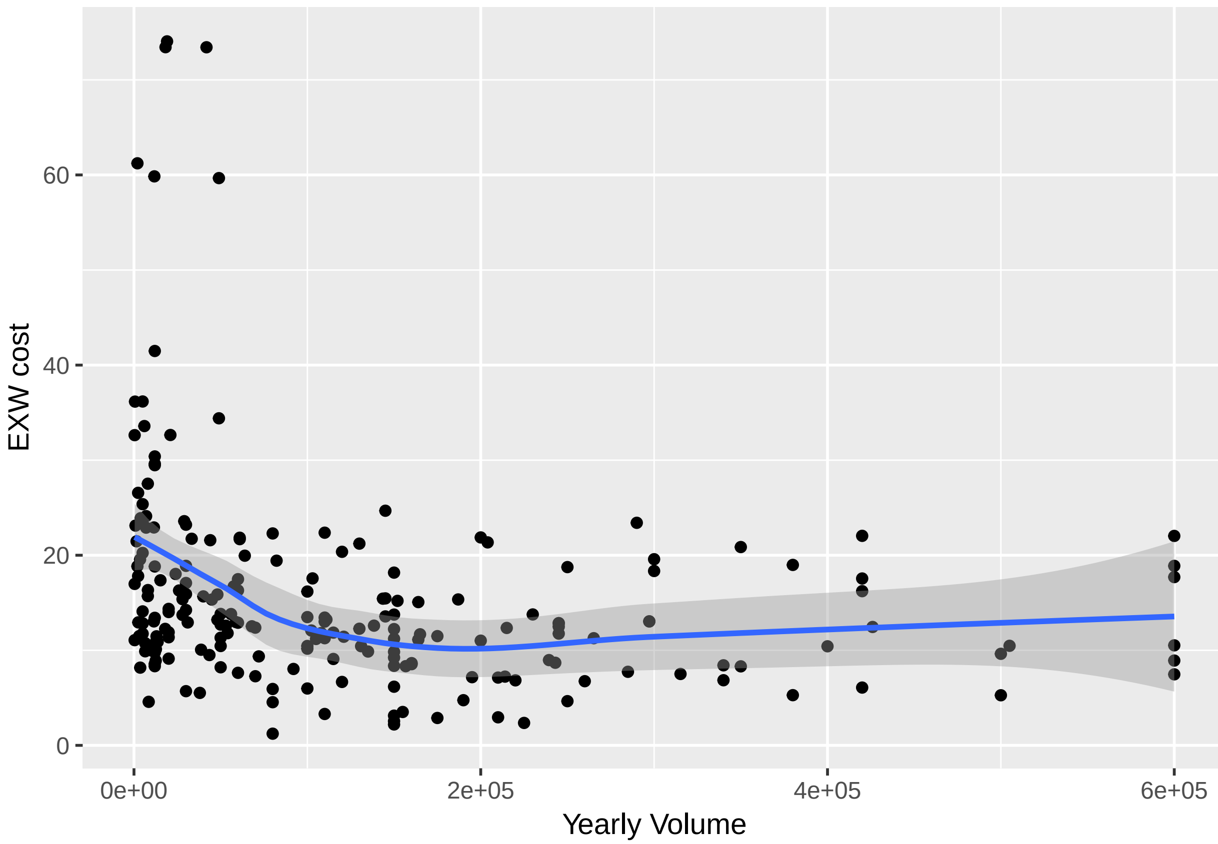
```
## # A tibble: 20 x 2
##    Supplier                supply
##    <chr>                    <dbl>
##  1 Admiral Supplier       2871529
##  2 Les espaces Supplier   2288915
##  3 Excalibur Supplier     2039880
##  4 Optima Supplier        1860202
##  5 Convergence Supplier   1813009
##  6 OneUp Supplier         1638515
##  7 Imaginaire Supplier    1616548
##  8 Hollywood Supplier     1465071
##  9 Conception Supplier    1461810
## 10 Galileo Supplier       1354036
## 11 Conduit Supplier       1280822
## 12 Full house Supplier     946550
## 13 Sedona Supplier         938719
## 14 Carcajou Supplier       937565
## 15 Downtown Supplier       890291
## 16 Chanceux Supplier       875211
## 17 World Supplier          807745
## 18 Nord Supplier           668480
```

```
## 19 Alcyon Supplier         561552
## 20 MillionDollar Supplier  529652
```

So the top 3 suppliers are Admiral, les espaces and Excalibur.

2.3 _Does the cost depend on the Net weight? on Yearly Volume? Does this make sense to you? Can you explain (from a business point of view) the form of the relationship for high volume values.






```
## [1] -0.2388312
```

```
## [1] 0.5045601
```

Seems there is a positive relationship between Net Weight and cost which just seems logical (the bigger the piece the higher the price). As for Yearly Volume and cost it makes sense as well given the property of economies of scale. The more production there is the more costs decrease, bigger lot sizes given overall economic profit bring costs down.

2.4 Let $n = 25$. Generate variables X and Y by drawing observations from independent gaussian distributions with mean $\mu = (0)_{1\times 2}$ and covariance matrix $\mathrm{Id}_{2\times 2}$. Compute the value of the correlation coefficient. Repeat the process 100 times and take the quantile at 95% of this empirical distribution (under the null hypothesis of no linear relationship) of the correlation coefficient. Comment the results. What should be learned from this experience?

```
# library(MASS)
n <- 25
XY <- MASS::mvrnorm(n, c(0,0), matrix(c(1,0,0,1),2,2))
X <- rnorm(25,0,1)
Y <- rnorm(25,0,1)
cor(XY[,1],XY[,2])
```

```
## [1] -0.1165405
```

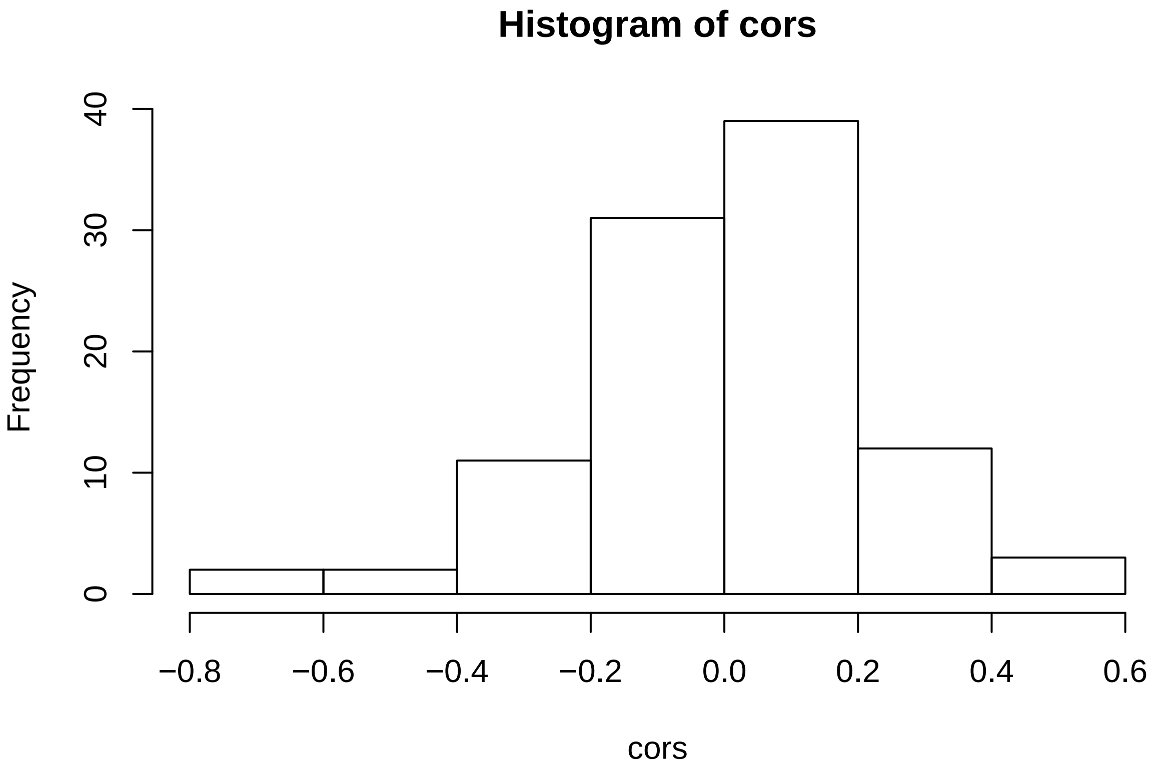
```
samples <- lapply(1:100, function(i) MASS::mvrnorm(n, c(0,0), matrix(c(1,0,0,1),2,2)))
cors <- sapply(samples, function(xy) cor(xy[,1],xy[,2]))
var(cors)
```

```
## [1] 0.04507112
```

```r
mean(cors)
```

```
## [1] 0.00153397
```

```r
hist(cors)
```




```r
quantile(cors, 0.95)
```

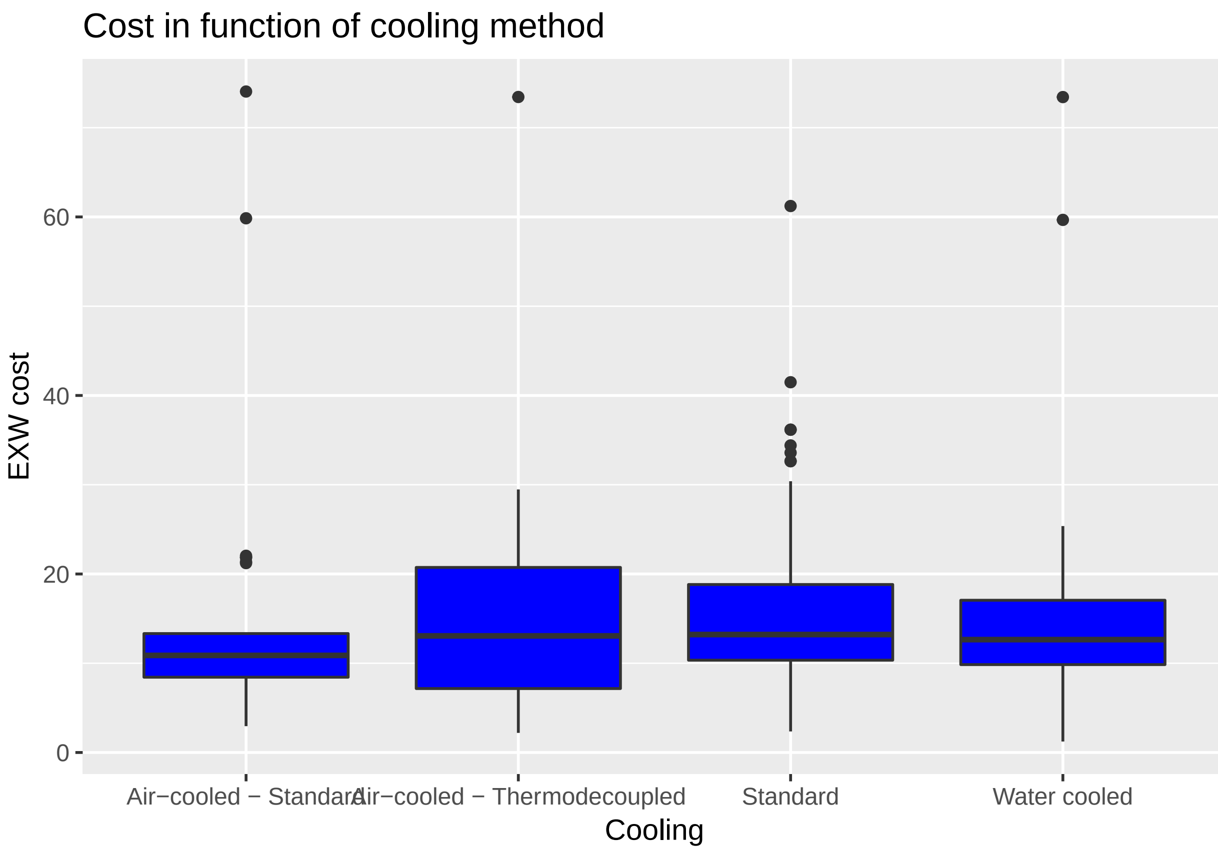
```
##       95%
## 0.3381779
```

Given their independence the correlation coefficient of the two random variables should be very close to 0 (cov(x,y) = 0 from the properties of gaussian vectors whose components are independent. We see there are slight deviations from that exact property of independance despite having imposed it when generating random variables, this means that when we do see links between varaibles we shouldn't be too hasty in interpreting them as really correlated.

2.5 Does the cost depend on the Cooling ?

```
##
## Call:
## lm(formula = data$`EXW cost` ~ data$Cooling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.739  -5.981  -2.681   3.169  58.871
##
## Coefficients:
##                                         Estimate Std. Error t value
## (Intercept)                              15.1787     2.0423   7.432
## data$CoolingAir-cooled - Thermodecoupled   1.7602     3.3697   0.522
## data$CoolingStandard                       0.1975     2.2991   0.086
```

```
## data$CoolingWater cooled                    0.4722     2.6423   0.179
##                                           Pr(>|t|)
## (Intercept)                               2.77e-12 ***
## data$CoolingAir-cooled - Thermodecoupled     0.602
## data$CoolingStandard                         0.932
## data$CoolingWater cooled                     0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.37 on 207 degrees of freedom
## Multiple R-squared:  0.001602,   Adjusted R-squared:  -0.01287
## F-statistic: 0.1107 on 3 and 207 DF,  p-value: 0.9538
```



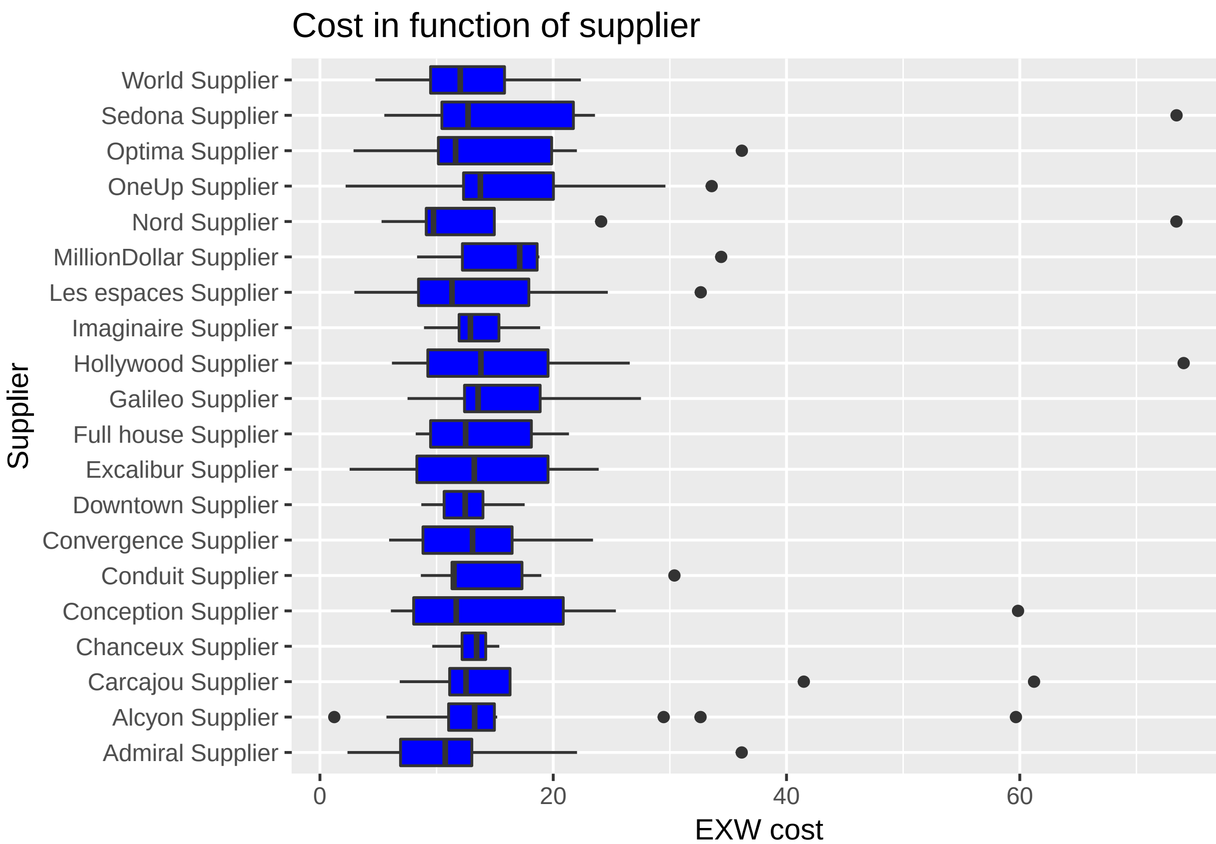
We see no real difference in costs in function of different cooling methods since they don't vary much in distribution across categories.

2.6 Which is the less expensive Supplier?

We can't just rely on least EXWcost, since this will vary in funcion of the quantity. So we'll approximate the expensiveness of suppliers by the average cost per unit of volume.

```
## # A tibble: 20 x 4
##    Supplier               total_volume total_cost  av_price
##    <chr>                         <dbl>      <dbl>     <dbl>
##  1 Admiral Supplier            2871529      168.  0.0000586
##  2 Imaginaire Supplier         1616548       95.2 0.0000589
##  3 Chanceux Supplier            875211       51.9 0.0000593
##  4 Optima Supplier             1860202      148.  0.0000797
##  5 Downtown Supplier            890291       75.7 0.0000850
```

```
##   6 Full house Supplier          946550       83.0 0.0000877
##   7 Convergence Supplier        1813009      165.  0.0000909
##   8 Les espaces Supplier        2288915      255.  0.000112
##   9 Conception Supplier         1461810      164.  0.000112
##  10 Excalibur Supplier          2039880      233.  0.000114
##  11 Conduit Supplier            1280822      161.  0.000126
##  12 World Supplier               807745      105.  0.000130
##  13 Galileo Supplier            1354036      176.  0.000130
##  14 OneUp Supplier              1638515      245.  0.000150
##  15 Hollywood Supplier          1465071      225.  0.000154
##  16 Carcajou Supplier            937565      184.  0.000196
##  17 MillionDollar Supplier       529652      107.  0.000201
##  18 Nord Supplier                668480      152.  0.000228
##  19 Sedona Supplier              938719      245.  0.000261
##  20 Alcyon Supplier              561552      240.  0.000428
```

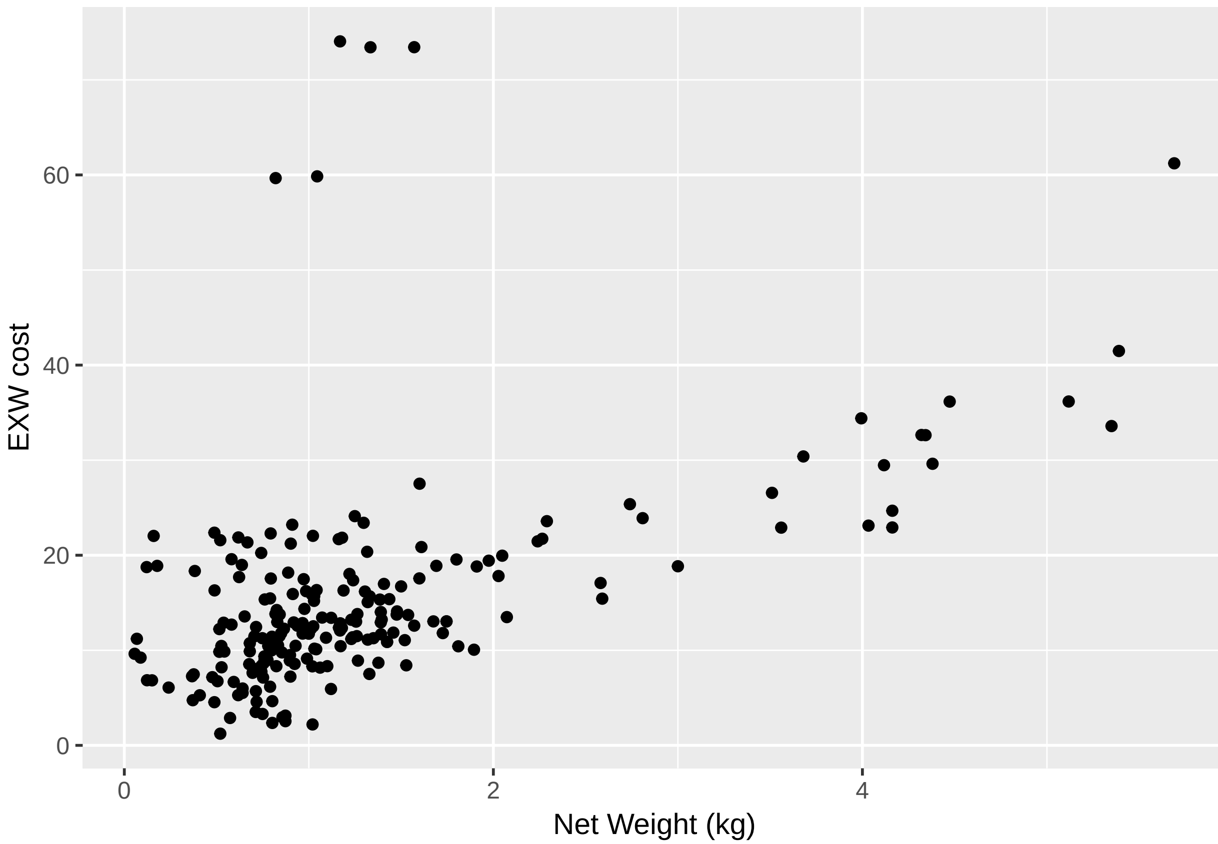


Seems the less expensive supplier is Admiral.

3) One important point in exploratory data analysis consists in identifying potential outliers.

3.1 Could you give points which are suspect regarding the Cost variable. Give the characteristics (other features) of the observations. We could keep them but keep in mind their presence and check if results are not too affected by these points.

I'll show suspicious points by looking at the cost in function of the most obvious and first varaible we should check



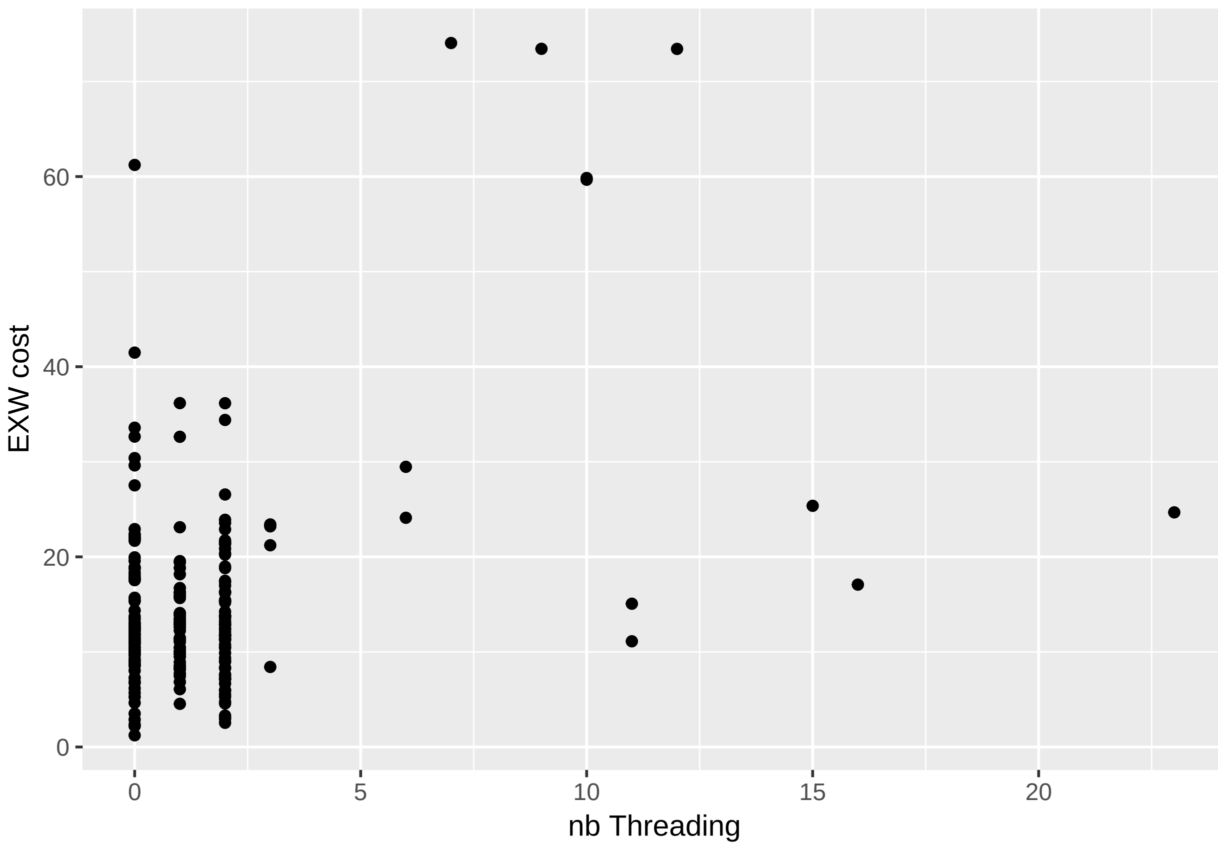
There are 5 points which seem suspicious, they don't fit the general increasing line we can easily imagine with the above data let's look at them.

```
##    ID       idPhoto       Date            Supplier Supplier Country
## 1 199 dieCasting-11 27/07/2016     Alcyon Supplier         Vietnam
## 2 152  dieCasting-8 29/02/2016 Conception Supplier           China
## 3 198  dieCasting-3 13/01/2016  Hollywood Supplier           China
## 4 109  dieCasting-6 15/04/2015       Nord Supplier           China
## 5 108  dieCasting-4 10/11/2016     Sedona Supplier           China
##   Yearly Volume Raw material Net Weight (kg)    Finishing
## 1         49000      Al 5371           0.820        Other
## 2         11760      Al 5371           1.045 Shotblasting
## 3         19088      Al 5371           1.169 Shotblasting
## 4         41830      Al 5371           1.334 Shotblasting
## 5         18200      Al 5371           1.571        Other
##   Surface envelop (LG x lg) (mm2) nb Machining Surfaces nb Threading
## 1                           45407                     5          10
## 2                           53556                     2          10
## 3                           39716                     2           7
## 4                           48753                     3          12
## 5                           20291                     2           9
##   Over molding Assembly nb Cavities                      Cooling   Process
## 1           No       No          10                 Water cooled       GDC
## 2          Yes       No          12        Air-cooled - Standard      HPDC
## 3          Yes      Yes          12        Air-cooled - Standard      HPDC
## 4           No       No          12                 Water cooled      HPDC
## 5          Yes      Yes          10 Air-cooled - Thermodecoupled Sand Cast
```

```
##   nb Cores EXW cost
## 1        4    59.67
## 2        1    59.85
## 3        1    74.05
## 4        4    73.43
## 5        2    73.44
```

There isn't any redundant feature in regards to all the other variables considered in our data so this either could be an error in registering the data (either weight isn't appropriate or cost etc.) or there is another characteristic not mentionned in our data.

3.2 Inspect the variable nb Threading, in views of its values of what could you suggest?




```
##  [1] "0"  "1"  "2"  "3"  "6"  "7"  "9"  "10" "11" "12" "15" "16" "23"
```

We find the same setting as Net Weight, so nb threading isn't behind this increase in cost. If it was we could expect higher costs for higher number of threading. Or it may be that there is an optimal number of threading which makes the product exceptional or that it is very rare.

4) Perform a PCA on the dataset DieCast.

```
data <- data.frame(diecasting) %>% mutate(ID = as.character(ID))

class <- as.data.frame(sapply(data,class))
class
```

```
##                                  sapply(data, class)
## ID                                         character
## idPhoto                                    character
## Date                                       character
```

```
## Supplier                                    character
## Supplier.Country                            character
## Yearly.Volume                                 numeric
## Raw.material                                character
## Net.Weight..kg.                               numeric
## Finishing                                   character
## Surface.envelop..LG.x.lg...mm2.               numeric
## nb.Machining.Surfaces                         numeric
## nb.Threading                                  numeric
## Over.molding                                character
## Assembly                                    character
## nb.Cavities                                   numeric
## Cooling                                     character
## Process                                     character
## nb.Cores                                      numeric
## EXW.cost                                      numeric
```

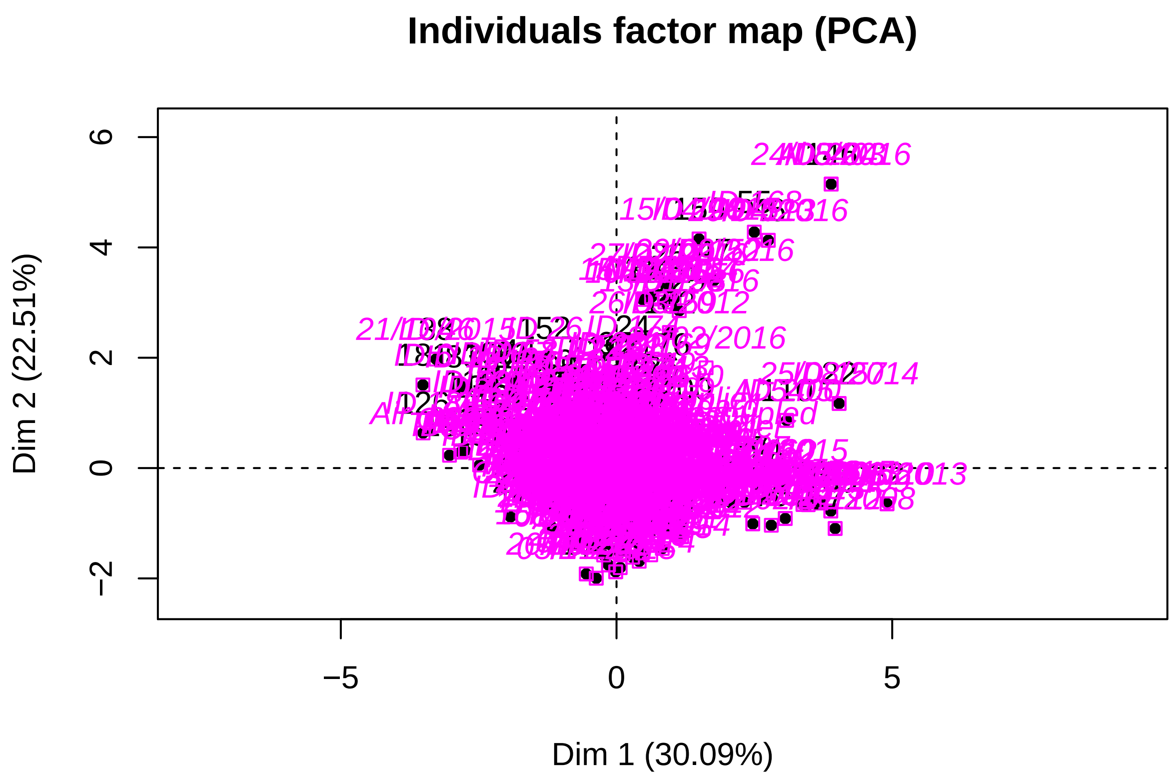
```
# We need to take into account that we have string vectors etc.

# ID was defined as numeric so we needed to transform it so the PCA wouldn't "take it into account"
# don_pca <- don %>% select(-ID)

strings <- c(which(class$`sapply(data, class)`!="numeric"))
# Defined a vector containing indexes of all the string vectors in our data
don_num <- data %>%
  select_if(is.numeric)

estim_ncp(don_num, method = "GCV")
```

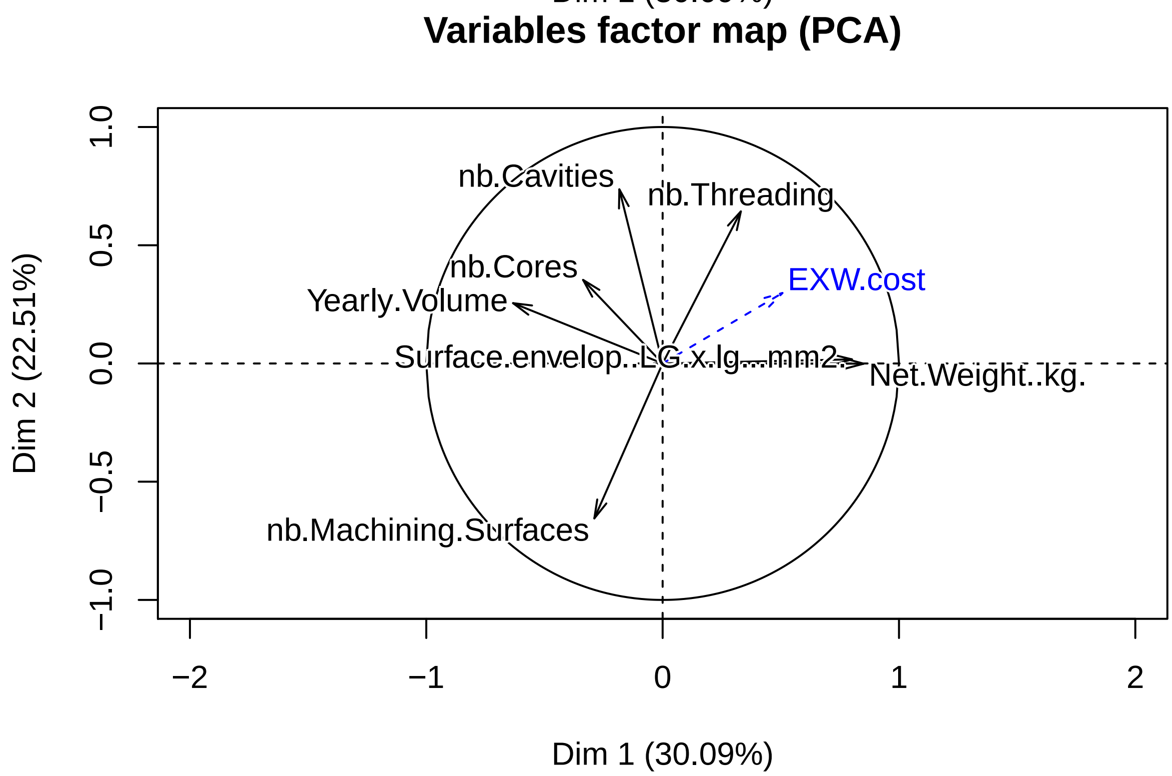
```
## $ncp
## [1] 2
##
## $criterion
## [1] 1.0000000 0.9093720 0.8735131 0.9385071 1.0744063 1.3125296 1.6808361
## [8] 2.6055416
```

```
estim_ncp(don_num, method = "Smooth")
```

```
## $ncp
## [1] 2
##
## $criterion
## [1]   1.0000000   0.8801871   0.8762108   1.0471334   1.6148836   3.1838768
## [7]  51.8851064 136.7400994
```

```
# We check the best number of dimensions to be kept when running our PCA, here it recommends considerin

res.pca <- PCA(data, quali.sup=strings, quanti.sup = 19, ncp = 2, scale=T)
```






```
# We scale here to take into account difference in scales in our data (YearlyVolume goes up to nx10000 a
```

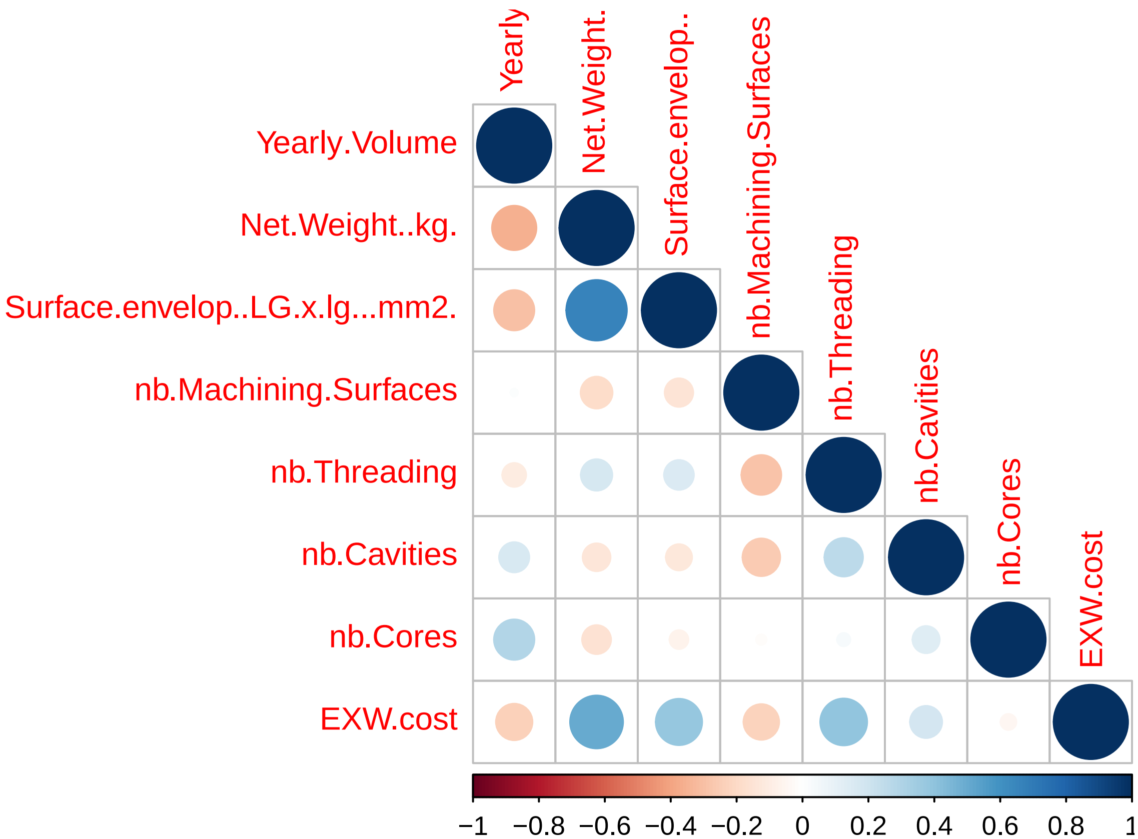
We see some points that are detached from the others in the individuals plot which could correspond to those wierd points we pointed out earlier!

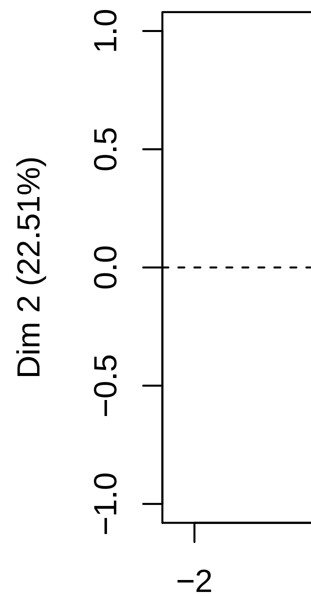
4.1 Explain briefly what are the aims of PCA and how categorical variables are handled?_

The aim of PCA is finding the best representation of a cloud of data in multiple dimensions in 2 dimensions as to be readable/interpretable by the human eye. Concerning categorical variables, PCA will project the categories at the point which minimizes the distance between that point and all observations which fall within the given catogery/ies, it's a sort of center of gravity of observations from a same category.

4.2 Compute the correlation matrix between the variables and comment it with respect to the correlation circle.

```
##                                  Yearly.Volume Net.Weight..kg.
## Yearly.Volume                             1.00           -0.80
## Net.Weight..kg.                          -0.80            1.00
## Surface.envelop..LG.x.lg...mm2.          -0.76            0.92
## nb.Machining.Surfaces                     0.13           -0.40
## nb.Threading                             -0.38            0.30
## nb.Cavities                               0.25           -0.37
## nb.Cores                                  0.54           -0.56
## EXW.cost                                 -0.71            0.76
##                                  Surface.envelop..LG.x.lg...mm2.
## Yearly.Volume                                              -0.76
## Net.Weight..kg.                                             0.92
## Surface.envelop..LG.x.lg...mm2.                             1.00
## nb.Machining.Surfaces                                      -0.37
## nb.Threading                                                0.25
## nb.Cavities                                                -0.40
## nb.Cores                                                   -0.49
## EXW.cost                                                    0.65
##                                  nb.Machining.Surfaces nb.Threading
## Yearly.Volume                                     0.13        -0.38
## Net.Weight..kg.                                  -0.40         0.30
## Surface.envelop..LG.x.lg...mm2.                  -0.37         0.25
## nb.Machining.Surfaces                             1.00        -0.67
## nb.Threading                                     -0.67         1.00
## nb.Cavities                                      -0.49         0.35
## nb.Cores                                         -0.01        -0.22
## EXW.cost                                         -0.60         0.60
##                                  nb.Cavities nb.Cores EXW.cost
## Yearly.Volume                           0.25     0.54    -0.71
## Net.Weight..kg.                        -0.37    -0.56     0.76
## Surface.envelop..LG.x.lg...mm2.        -0.40    -0.49     0.65
## nb.Machining.Surfaces                  -0.49    -0.01    -0.60
## nb.Threading                            0.35    -0.22     0.60
## nb.Cavities                             1.00     0.16     0.06
## nb.Cores                                0.16     1.00    -0.47
## EXW.cost                                0.06    -0.47     1.00
```





To compare the two plots I'll focus first on strong correlations identified for both and see if they coincide through a few examples: - Surface envelop and Net Weight are positively highly correlated for the correlation plot, and are almost aligned in the correlation circle from the PCA - Net weight and Surface envelop are strongly correlated with EXW Cost from the correlation plot, this is reflected in the correlation circle fiven the angle between their projections is tight. - Yearly volume and Nb machining surface have correlation of 0 and they are orthogonal in the correlation plot.

The PCA seems to have conserved the actual correlations between variables in our data, although the projection of EXW cost isn't as good as we would like, the length of the vector being a bit small.

4.3 On what kind of relationship PCA focuses? Is it a problem?

PCA focuses on linear relationships of our data, and although it might seem restrictive as their exists many other ways to consider the relationships between data (log, quadratic, etc.) considering linear relationships is very reasonnable for an initial approximation.

4.4 Give the the R object with the two principal components which are the synthetic variables the most correlated to all the variables.

We sam before that the best number of dimensions to represent the variability of our data was 2, so in any case I only have those two components to show you... They would be also those in a PCA where we wouldn't have limited the PCA to 2 dimensions.

```
##                                      Dim.1         Dim.2
## Yearly.Volume                   -0.6329578  0.2551483858
## Net.Weight..kg.                  0.8509078 -0.0009213585
## Surface.envelop..LG.x.lg...mm2.  0.8011105  0.0173780368
## nb.Machining.Surfaces           -0.2889087 -0.6553092910
## nb.Threading                     0.3310270  0.6432060238
## nb.Cavities                     -0.1831341  0.7365883304
```

```
## nb.Cores                          -0.3366753  0.3534213169
##                                  Correlation PC1 Correlation PC2 Cos2 PC1
## Yearly.Volume                              -0.63            0.26     0.40
## Net.Weight..kg.                             0.85            0.00     0.72
## Surface.envelop..LG.x.lg...mm2.             0.80            0.02     0.64
## nb.Machining.Surfaces                      -0.29           -0.66     0.08
## nb.Threading                                0.33            0.64     0.11
## nb.Cavities                                -0.18            0.74     0.03
## nb.Cores                                   -0.34            0.35     0.11
##                                  Cos2 PC2 Contribution PC1 Contribution PC2
## Yearly.Volume                        0.07            19.02             4.13
## Net.Weight..kg.                      0.00            34.37             0.00
## Surface.envelop..LG.x.lg...mm2.      0.00            30.47             0.02
## nb.Machining.Surfaces                0.43             3.96            27.25
## nb.Threading                         0.41             5.20            26.25
## nb.Cavities                          0.54             1.59            34.43
## nb.Cores                             0.12             5.38             7.93
```

5) Clustering

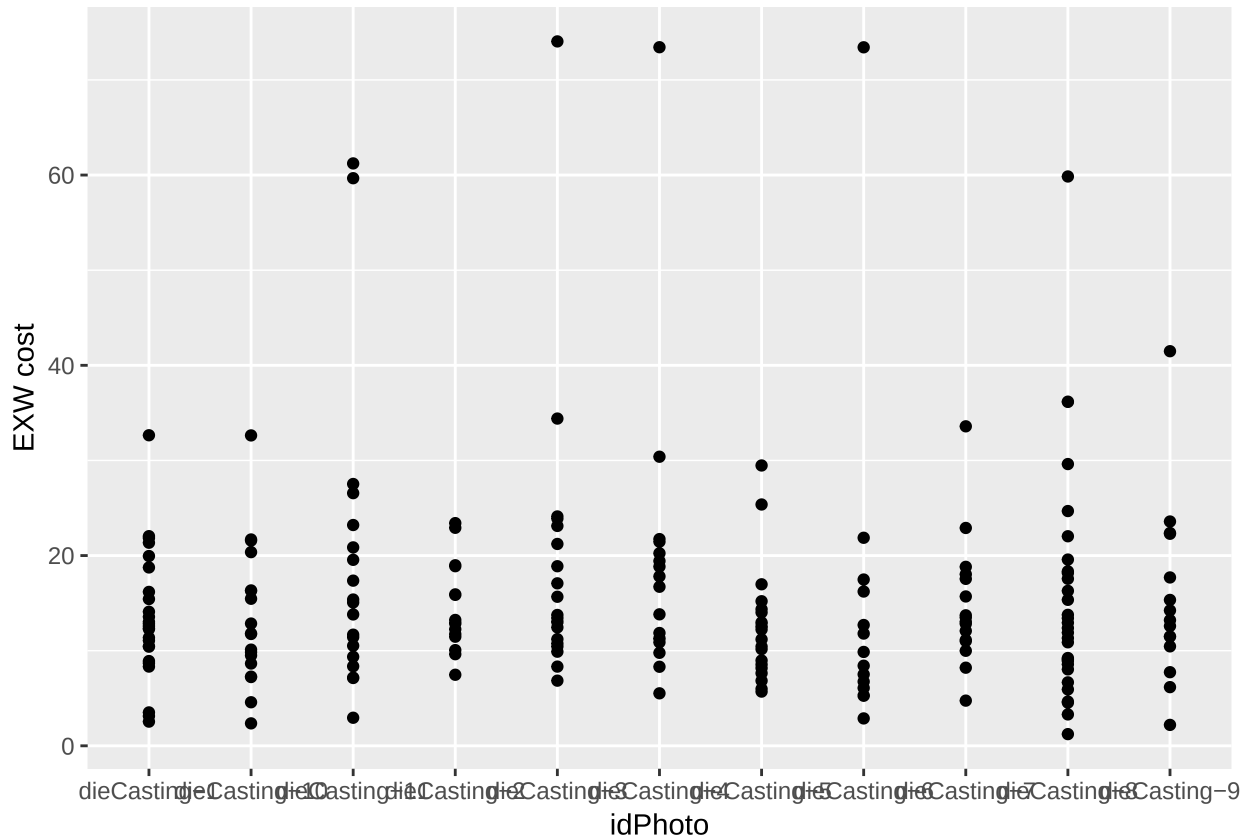
5.1) Principal components methods such as PCA is often used as a pre-processing step before applying a clustering algorithm, explain the rationale of this approach and how many components you should keep.

PCA can be peroformed on a dataset before going through clustering methods when there is a large amount of varaibles. It denoises our data to allow a more stable clustering by keeping only a the first principal components such that we keep 95% of the inertia (we don't want to lose too much information). In addition, combining both methods gives us plots which allow better interpretation so it's only benefitial if we're cautious about restricting the number of dimensions we keep.

5.2) To simultaneously take into account quantitative and categorical variables in the clustering you should use the clustering on the results of the FAMD ones. FAMD stands for Factorial Analysis of Mixed Data and is a PCA dedicated to mixed data. Explain what will be the impacts of such an analysis on the results?_

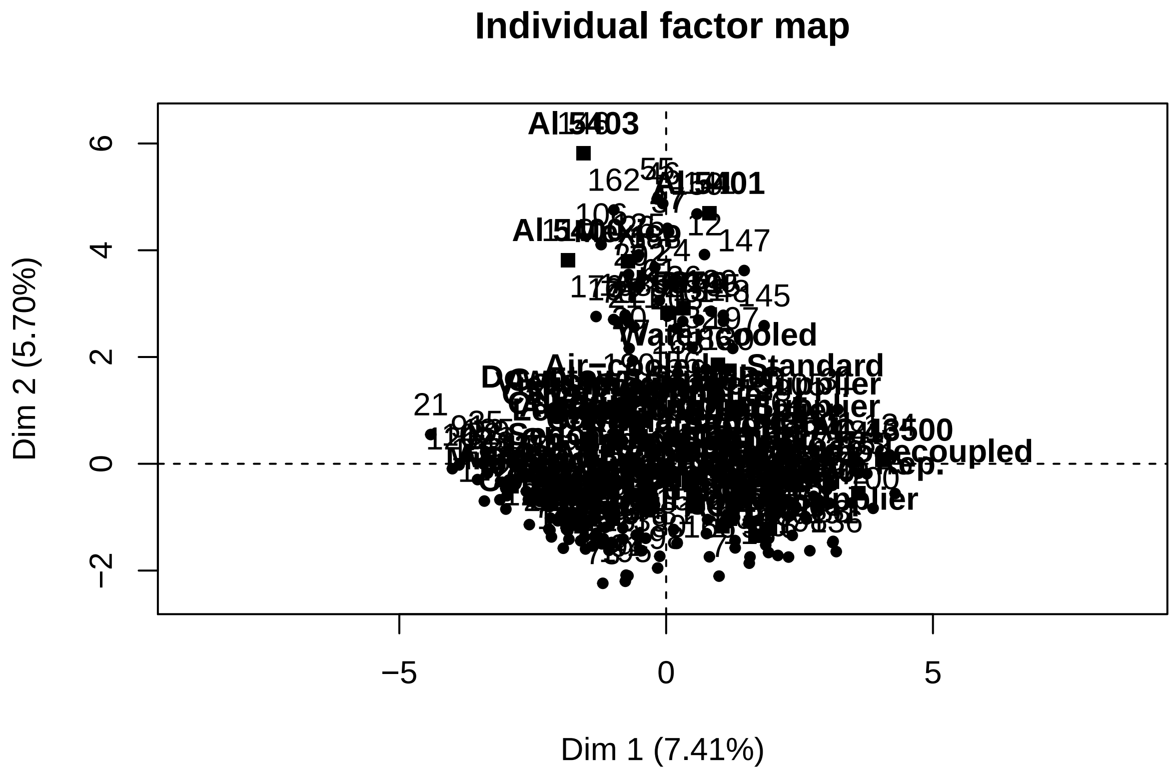
Obviously the principal components will change since FAMD will take into account qualitative variables instead of calculating the barycenter of data that fall within the classes of the qualitative variables. Here FAMD will balance the influence of each varaible when computing the distance between an individual when projected and the center of gravity of the cloud.

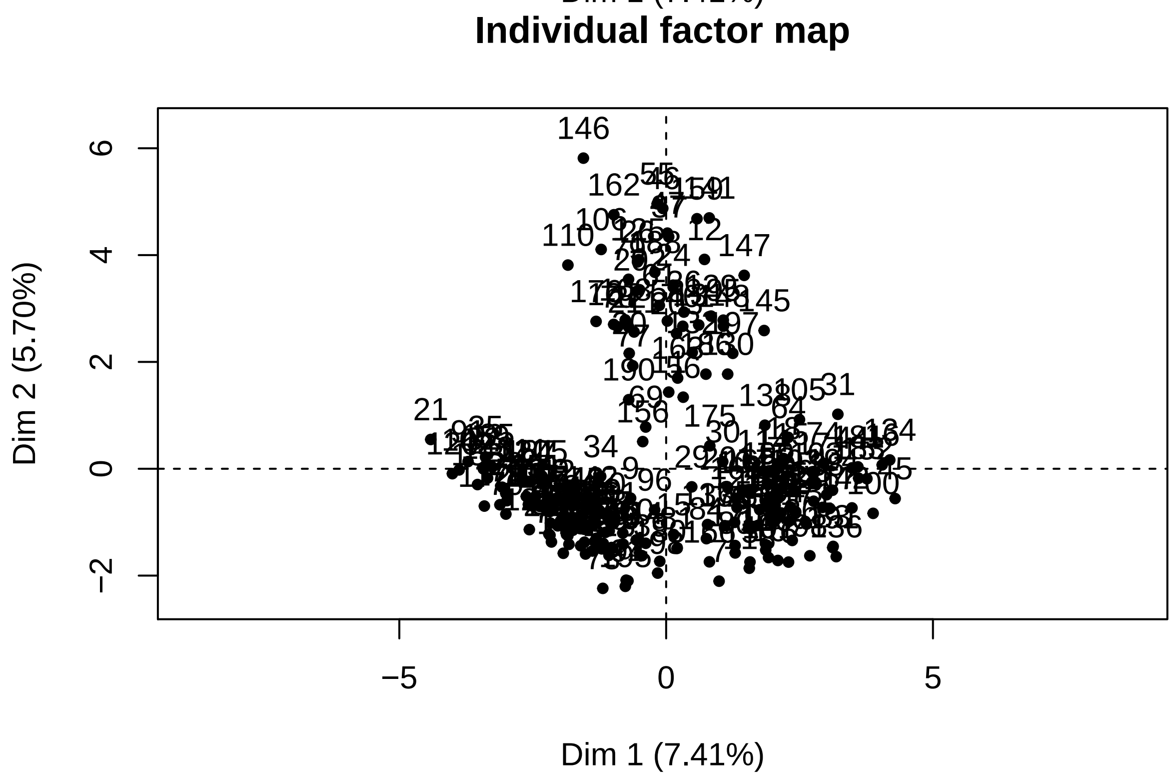
5.3) Perform the FAMD, and keep the principal components you want for the clustering.



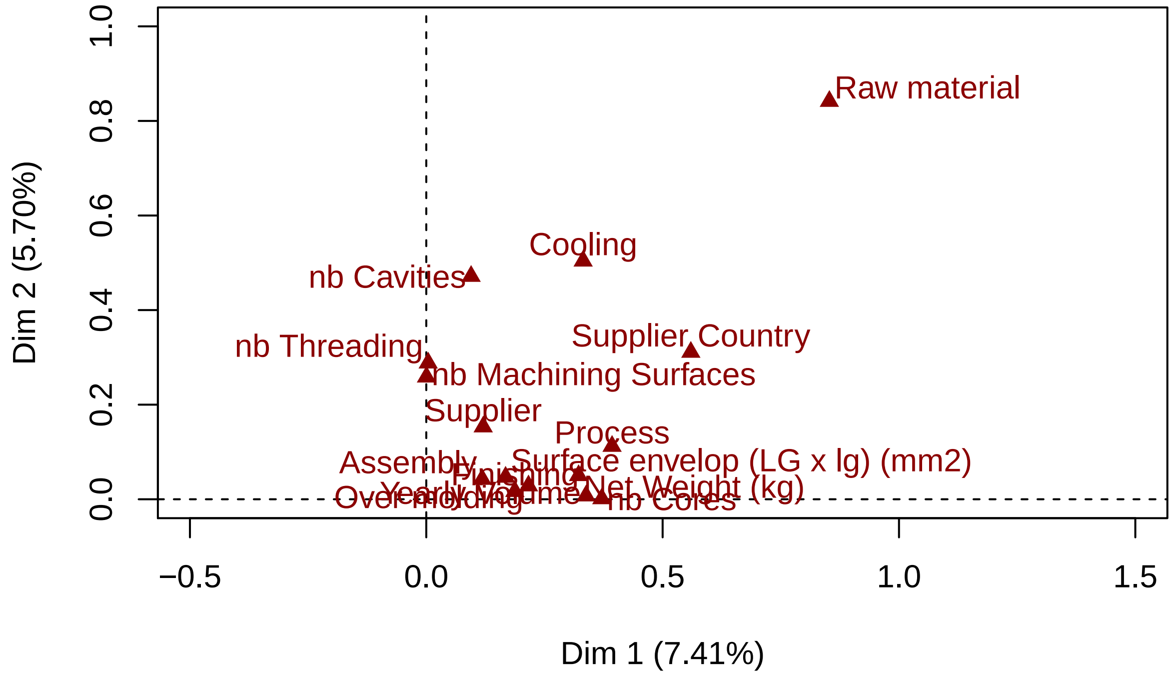
No relation between idphoto and cost, I decide to discard it for my kmmeans, I get rid of date as well as I don't know if the cost was determined after the transatcion or if it is the cost at the date where it was made. Il also get rid of ID since it's just an identifier.

```
##                                    sapply.diecasting..class.
## ID                                                   numeric
## idPhoto                                            character
## Date                                               character
## Supplier                                           character
## Supplier Country                                   character
## Yearly Volume                                        numeric
## Raw material                                       character
## Net Weight (kg)                                      numeric
## Finishing                                          character
## Surface envelop (LG x lg) (mm2)                      numeric
## nb Machining Surfaces                                numeric
## nb Threading                                         numeric
## Over molding                                       character
## Assembly                                           character
## nb Cavities                                          numeric
## Cooling                                            character
## Process                                            character
## nb Cores                                             numeric
## EXW cost                                             numeric
```

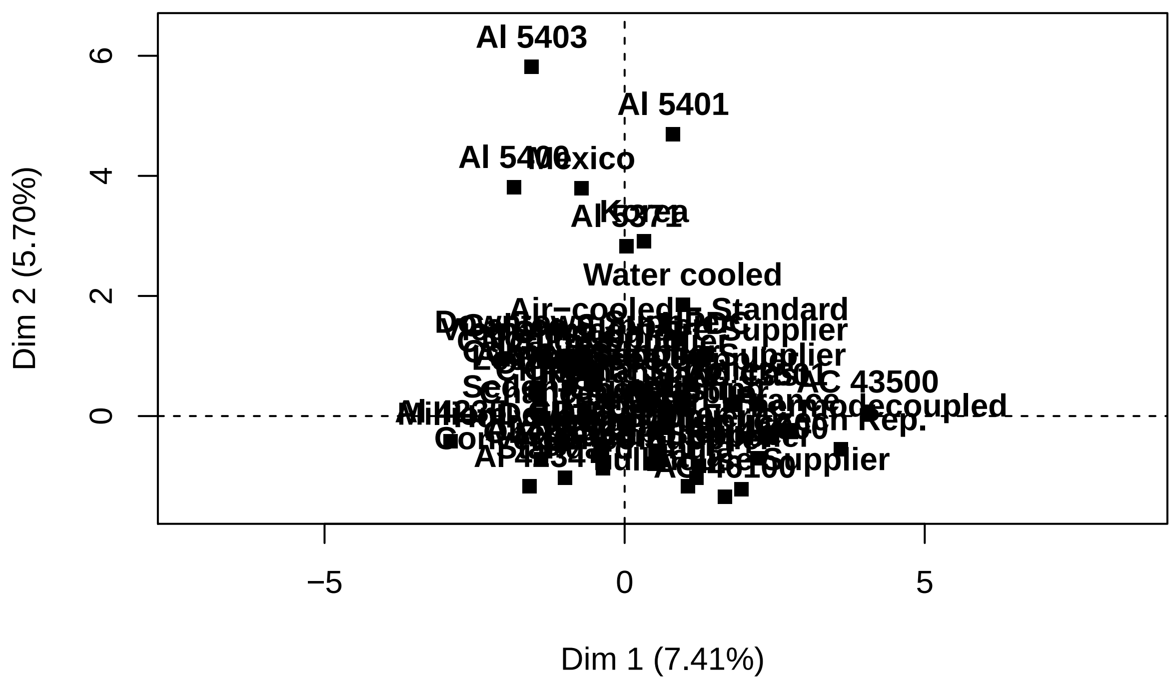

Individual factor map
Dim 2 (5.70%)
Dim 1 (7.41%)
−2
0
2
4
6
−5
0
5

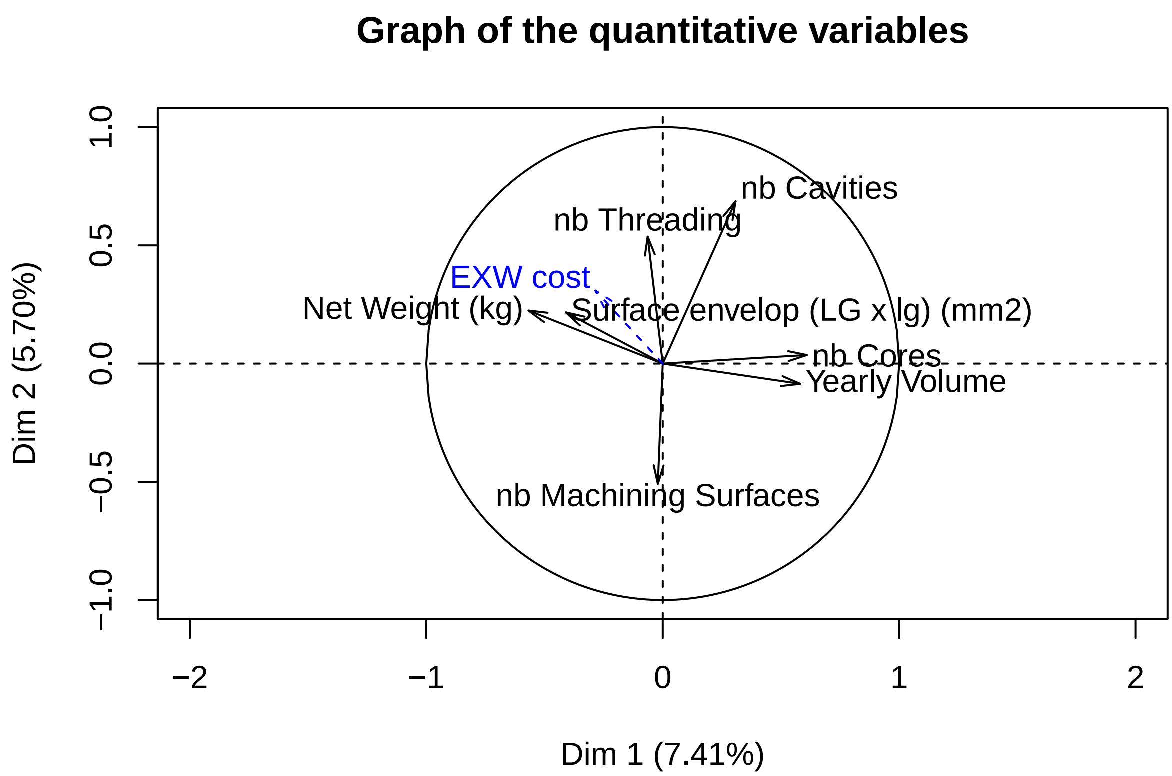

Individual factor map
Dim 2 (5.70%)
Dim 1 (7.41%)
−2
0
2
4
6
−5
0
5

Graph of the variables



Individual factor map

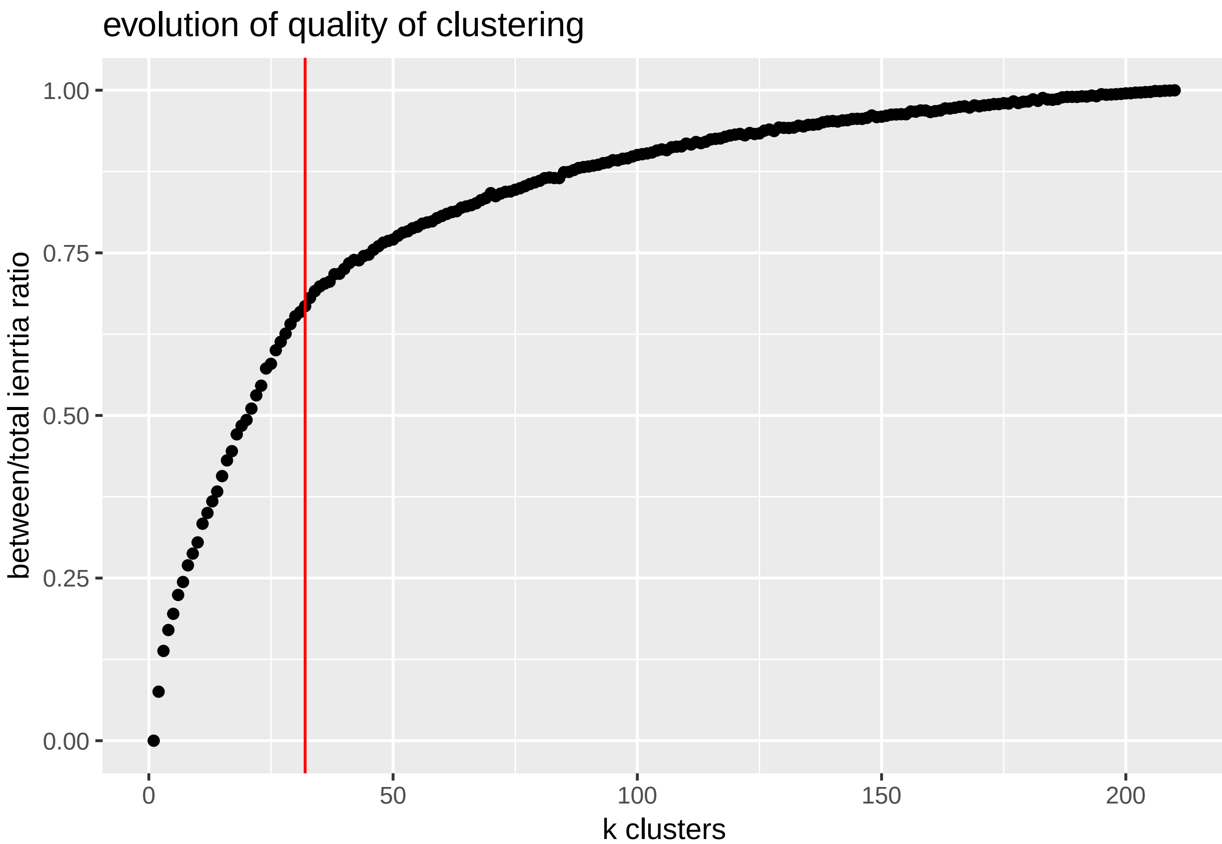




Here a compromise is to be done between the amount of inertia we want to keep which increases with the number of dimensions, and the actual number of dimensions which we don't want to be too high. I think keeping at least 80% of the inertia is a good compromise, which corresponds to keeping 31 ncp's.

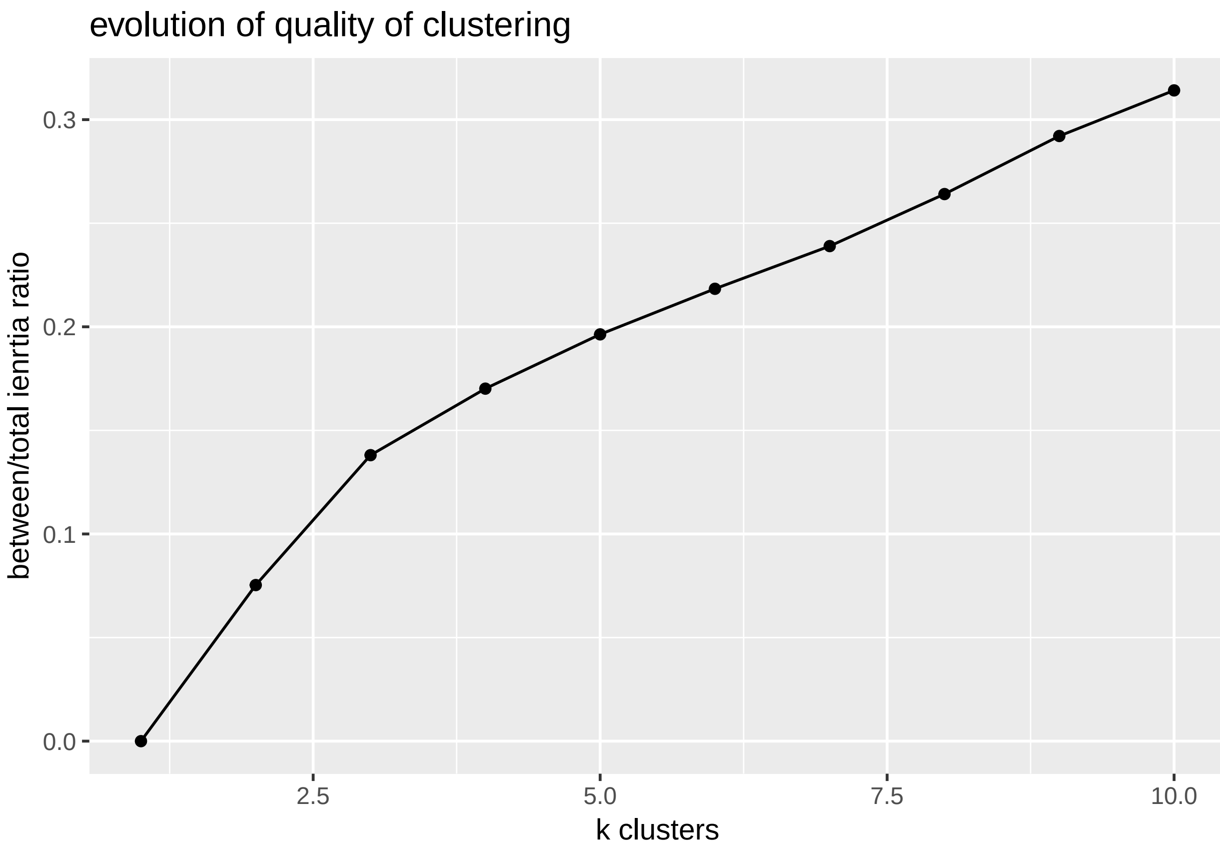
5.4) Perfom a kmeans algorithm on the selected principal components of FAMD. To select how many cluster you are keeping, you can represent the evolution of the ratio between/total intertia. Justify your choices.

```
pc <- data.frame(res.famd$ind$coord)

res.kmeanss <- lapply(1:210, function(i) kmeans(res.famd$ind$coord,centers = i,nstart = 10))
qual_kmeans <- sapply(1:210, function(i) (res.kmeanss[[i]]$betweens)/(res.kmeanss[[i]]$totss))
ggplot(data = NULL,aes(x = 1:210,y = qual_kmeans)) + geom_point() + labs(x = "k clusters", y = "between
```

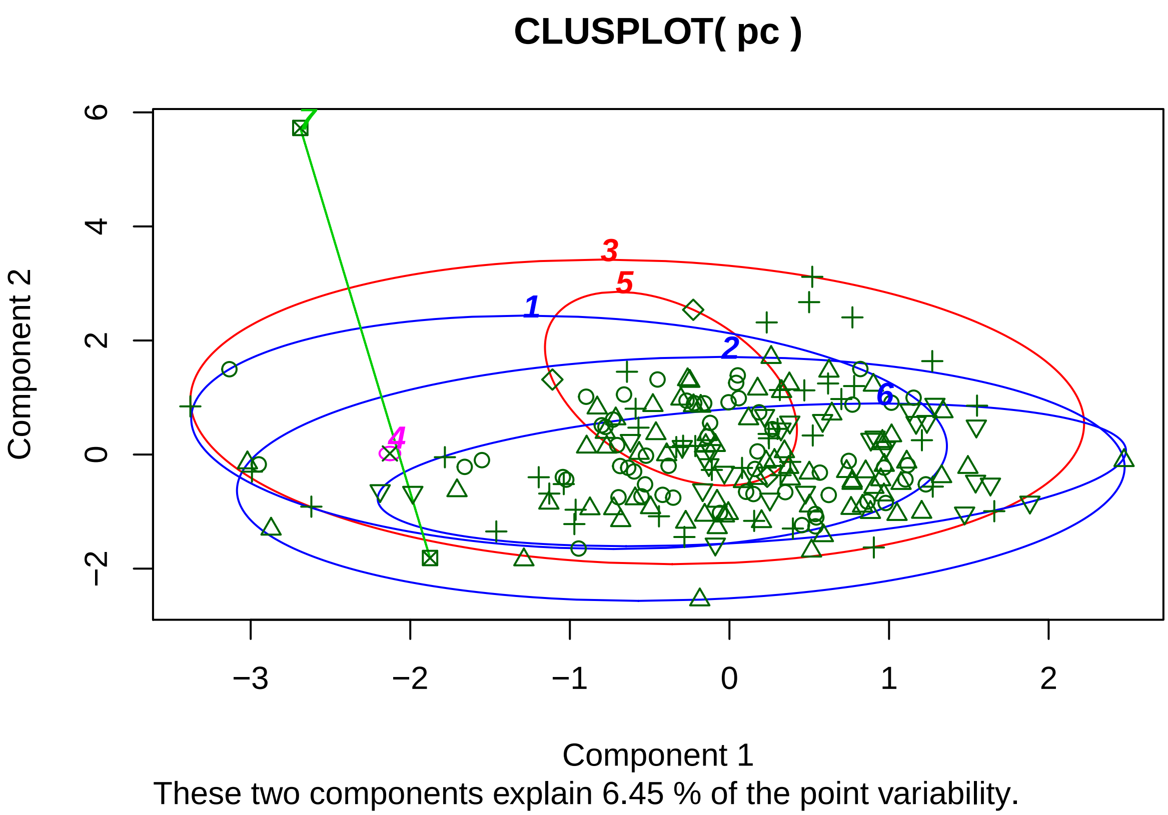


Here we need to make a choice, we want to choose a certain number of clusters such that we keep a an acceptable percentage of within cluster inertia, but not take too much clusters as it goes against the purpose of clusters since we would end up with as much clusters as observations. We see here that the marginal increase in percentage of between inertia/ total inertia decreases a lot when reaching approx. 30 clusters, but choosing the corresponding number of clusters would make us take way too many clusters. I'll re-iterate my analysis but on a closer interval:

```
res.kmeanss <- lapply(1:10, function(i) kmeans(res.famd$ind$coord,centers = i,nstart = 10))
qual_kmeans <- sapply(1:10, function(i) (res.kmeanss[[i]]$betweens)/(res.kmeanss[[i]]$totss))
ggplot(data = NULL,aes(x = 1:10,y = qual_kmeans)) + geom_point() + labs(x = "k clusters", y = "between/
```



I'm hesitating between between clustering ranging from 3 to 7. Let's see what they look like:



```
## # A tibble: 7 x 2
##   classe   obs
##   <fct>  <int>
## 1 1         58
## 2 2         76
## 3 3         43
## 4 4          1
## 5 5          3
## 6 6         28
## 7 7          2
```

The clusters might be good, but we get this one cluster with only one observation which either means this a real outlier within our data or that we've chosen just a bit too much clusters and kmeans found a way to minimize between class inertia by attributing this one observation to a cluster. Let's see the point in question:

```
##                  Supplier Supplier Country Yearly Volume Raw material
## 1        Admiral Supplier            China        284908      Al 4234
## 2        Admiral Supplier          Romania        200000      Al 4234
## 3        Admiral Supplier           Italia        505000      Al 4234
## 4         Alcyon Supplier           Italia         30000      Al 4234
## 5         Alcyon Supplier           Italia        152000      Al 4234
## 6       Carcajou Supplier            China         67980      Al 4234
## 7       Carcajou Supplier            Korea         54000      Al 4234
## 8       Chanceux Supplier           Italia         50000      Al 4234
## 9     Conception Supplier            China          8050      Al 4234
## 10    Conception Supplier            China        165000      Al 4234
## 11    Conception Supplier          Romania         92000      Al 4234
## 12       Conduit Supplier            China        103000      Al 4234
## 13       Conduit Supplier            China         12000      Al 4234
## 14       Conduit Supplier           Italia        121000      Al 4234
## 15   Convergence Supplier         Slovakia        100000      Al 4234
## 16   Convergence Supplier         Slovakia        245000      Al 4234
## 17   Convergence Supplier            India          3608      Al 4234
## 18   Convergence Supplier            China         44885      Al 4234
## 19      Downtown Supplier            China         20000      Al 4234
## 20      Downtown Supplier           Italia          5000      Al 4234
## 21     Excalibur Supplier         Slovakia        245000      Al 4234
## 22     Excalibur Supplier            China          3500      Al 4234
## 23     Excalibur Supplier         Slovakia        245000      Al 4234
## 24     Excalibur Supplier           Italia         30000      Al 4234
## 25    Full house Supplier          Romania         50000      Al 4234
## 26       Galileo Supplier            China         12000      Al 4234
## 27       Galileo Supplier            China         31000      Al 4234
## 28     Hollywood Supplier           Italia        187000      Al 4234
## 29     Hollywood Supplier            China         48113      Al 4234
## 30     Hollywood Supplier            China         18000      Al 4234
## 31     Hollywood Supplier          Vietnam          6470      Al 4234
## 32   Les espaces Supplier           Italia          5000      Al 4234
## 33   Les espaces Supplier            China          2000      Al 4234
## 34   Les espaces Supplier            China        110000      Al 4234
## 35   Les espaces Supplier            India           400      Al 4234
## 36   Les espaces Supplier            China         12000      Al 4234
## 37 MillionDollar Supplier           Italia         24000      Al 4234
## 38 MillionDollar Supplier            China         12000      Al 4234
## 39          Nord Supplier            China         12650      Al 4234
```

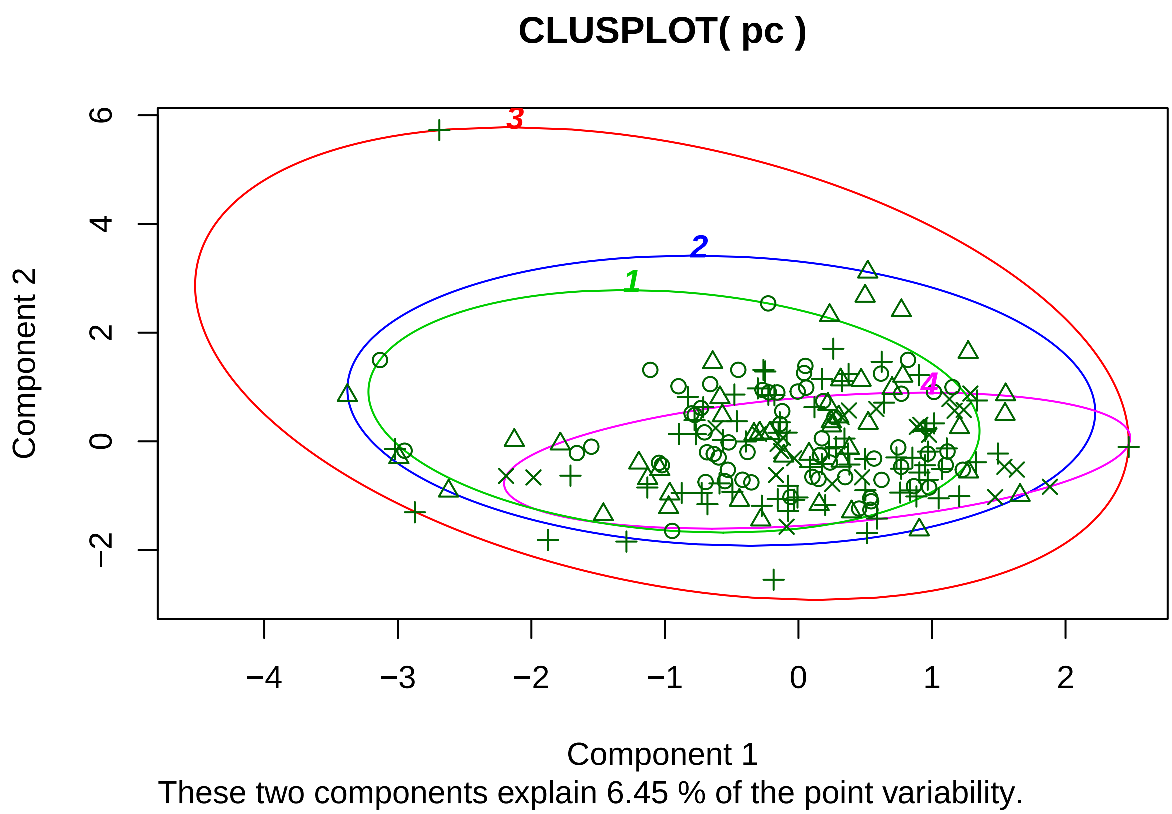
```
## 40          Nord Supplier          China        115000      Al 4234
## 41          Nord Supplier       Slovakia         20000      Al 4234
## 42          Nord Supplier          China         20000      Al 4234
## 43         OneUp Supplier          China        150000      Al 4234
## 44         OneUp Supplier       Slovakia        245000      Al 4234
## 45         OneUp Supplier        Vietnam          6470      Al 4234
## 46         OneUp Supplier          China          7000      Al 4234
## 47         OneUp Supplier         Italia        215000      Al 4234
## 48        Optima Supplier          China          1500      Al 4234
## 49        Optima Supplier         Italia          5000      Al 4234
## 50         World Supplier          China         53000      Al 4234
## 51         World Supplier          China         13000      Al 4234
## 52         World Supplier         Italia         20000      Al 4234
## 53        Sedona Supplier          China        100000      Al 4234
## 54        Sedona Supplier          China         10000      Al 4234
## 55        Sedona Supplier        Romania        195000      Al 4234
## 56        Sedona Supplier          India         38000      Al 4234
## 57        Sedona Supplier          China         82250      Al 4234
## 58        Sedona Supplier          China          8000      Al 4234
##    Net Weight (kg)    Finishing Surface envelop (LG x lg) (mm2)
## 1            0.743 Shotblasting                           24685
## 2            0.790 Shotblasting                           16346
## 3            0.928 Shotblasting                           51536
## 4            0.826        Other                           54007
## 5            1.028        Other                           34652
## 6            1.024     Tumbling                           42146
## 7            0.850     Tumbling                           21921
## 8            0.819     Tumbling                           10612
## 9            1.042     Tumbling                           48743
## 10           1.391 Shotblasting                           57346
## 11           0.705 Shotblasting                           39025
## 12           1.599        Other                           57950
## 13           0.854        Other                           18109
## 14           0.800     Tumbling                           21397
## 15           1.304        Other                           43209
## 16           0.966 Shotblasting                           17821
## 17           1.061        Other                           10190
## 18           1.385 Shotblasting                           23025
## 19           0.976        Other                           43539
## 20           1.170 Shotblasting                           41012
## 21           0.966        Other                           28435
## 22           1.800        Other                           37791
## 23           0.966        Other                           26803
## 24           0.913        Other                           11140
## 25           0.527        Other                           10196
## 26           1.121     Tumbling                           56173
## 27           0.918 Shotblasting                           33687
## 28           0.761        Other                           43917
## 29           1.016        Other                           39588
## 30           1.011     Tumbling                           15710
## 31           0.680        Other                           51497
## 32           0.742 Shotblasting                           15395
## 33           3.000     Tumbling                           22657
## 34           1.256 Shotblasting                           12727
```

```
## 35           1.520     Tumbling                           29924
## 36           0.923        Other                           15381
## 37           1.220 Shotblasting                           50682
## 38           0.824        Other                           25413
## 39           1.040        Other                           38610
## 40           0.760 Shotblasting                           13676
## 41           0.997 Shotblasting                           37860
## 42           0.990     Tumbling                           46933
## 43           1.476        Other                           24331
## 44           0.966     Tumbling                           20715
## 45           0.680     Tumbling                           35972
## 46           3.560 Shotblasting                           32282
## 47           1.179 Shotblasting                           34340
## 48           2.240     Tumbling                           58835
## 49           1.170 Shotblasting                           51926
## 50           0.938 Shotblasting                           20143
## 51           0.705     Tumbling                           27493
## 52           1.390 Shotblasting                           20116
## 53           0.526     Tumbling                           23299
## 54           0.798 Shotblasting                           47852
## 55           0.477     Tumbling                           15188
## 56           0.641        Other                           10709
## 57           1.975     Tumbling                           48543
## 58           1.029 Shotblasting                           19144
##    nb Machining Surfaces nb Threading Over molding Assembly nb Cavities
## 1                     10            1           No      Yes           1
## 2                      8            0          Yes       No           0
## 3                     24            2          Yes      Yes           4
## 4                     18            2           No       No           2
## 5                     23            2          Yes       No           2
## 6                     29            0           No      Yes           1
## 7                     16            0           No      Yes           2
## 8                     21            1           No      Yes           0
## 9                     17            2           No      Yes           1
## 10                    20            2           No      Yes           1
## 11                    10            0          Yes       No           1
## 12                    32            0          Yes       No           1
## 13                    14            1           No       No           1
## 14                    10            2           No       No           2
## 15                    16            1           No       No           2
## 16                    16            2          Yes      Yes           0
## 17                    13            1           No      Yes           2
## 18                    30            0           No       No           1
## 19                    21            0          Yes      Yes           1
## 20                    28            1          Yes      Yes           2
## 21                    16            2          Yes       No           0
## 22                     7            1           No      Yes           1
## 23                    16            2          Yes      Yes           0
## 24                    27            1          Yes       No           2
## 25                    14            1          Yes      Yes           0
## 26                    13            1           No      Yes           0
## 27                    29            1          Yes      Yes           1
## 28                    25            2          Yes      Yes           2
## 29                    21            1          Yes       No           1
```

```
## 30                   16            1          No        No            1
## 31                   16            2         Yes       Yes            2
## 32                   25            2          No       Yes            2
## 33                   13            1          No        No            1
## 34                   24            1         Yes       Yes            2
## 35                   15            1          No       Yes            0
## 36                   11            1         Yes       Yes            1
## 37                   18            0          No       Yes            2
## 38                    8            2          No        No            1
## 39                   19            0          No        No            1
## 40                   10            2          No        No            0
## 41                   17            0          No        No            0
## 42                    8            0         Yes        No            0
## 43                   29            2          No       Yes            1
## 44                   16            2         Yes        No            0
## 45                   15            2         Yes       Yes            2
## 46                   14            2          No        No            1
## 47                   14            2          No        No            2
## 48                    6            2         Yes        No            1
## 49                   22            0          No       Yes            2
## 50                   27            0          No        No            1
## 51                   10            1          No        No            1
## 52                   16            1          No        No            2
## 53                   19            1         Yes       Yes            2
## 54                   11            1          No        No            1
## 55                   13            2         Yes        No            1
## 56                    8            2          No        No            0
## 57                   19            1          No        No            2
## 58                   21            0         Yes       Yes            1
##     Cooling Process nb Cores EXW cost classe
## 1  Standard     GDC        1    7.738      1
## 2  Standard     GDC        1   11.012      1
## 3  Standard     GDC        2   10.471      1
## 4  Standard     GDC        4   14.234      1
## 5  Standard     GDC        3   15.198      1
## 6  Standard     GDC        2   12.524      1
## 7  Standard     GDC        1   11.801      1
## 8  Standard     GDC        2   13.823      1
## 9  Standard     GDC        3   16.334      1
## 10 Standard     GDC        1   11.683      1
## 11 Standard     GDC        1    8.036      1
## 12 Standard     GDC        5   17.562      1
## 13 Standard     GDC        1    9.776      1
## 14 Standard     GDC        1   11.426      1
## 15 Standard     GDC        2   16.181      1
## 16 Standard     GDC        1   11.756      1
## 17 Standard     GDC        1    8.175      1
## 18 Standard     GDC        2   15.335      1
## 19 Standard     GDC        1   14.356      1
## 20 Standard     GDC        1   12.830      1
## 21 Standard     GDC        1   12.852      1
## 22 Standard     GDC        1   19.553      1
## 23 Standard     GDC        1   12.852      1
## 24 Standard     GDC        5   15.918      1
```

```
## 25 Standard     GDC         1    8.212       1
## 26 Standard     GDC         2   13.416       1
## 27 Standard     GDC         2   12.930       1
## 28 Standard     GDC         3   15.347       1
## 29 Standard     GDC         3   15.869       1
## 30 Standard     GDC         3   12.253       1
## 31 Standard     GDC         1    9.894       1
## 32 Standard     GDC         3   20.239       1
## 33 Standard     GDC         2   18.839       1
## 34 Standard     GDC         2   13.018       1
## 35 Standard     GDC         2   11.067       1
## 36 Standard     GDC         1    8.568       1
## 37 Standard     GDC         1   18.031       1
## 38 Standard     GDC         1    8.327       1
## 39 Standard     GDC         2   10.111       1
## 40 Standard     GDC         1    9.088       1
## 41 Standard     GDC         1   11.896       1
## 42 Standard     GDC         1    9.120       1
## 43 Standard     GDC         2   13.754       1
## 44 Standard     GDC         1   12.491       1
## 45 Standard     GDC         1   10.746       1
## 46 Standard     GDC         2   22.903       1
## 47 Standard     GDC         3   12.349       1
## 48 Standard     GDC         1   21.462       1
## 49 Standard     GDC         1   12.830       1
## 50 Standard     GDC         1   12.586       1
## 51 Standard     GDC         1   11.470       1
## 52 Standard     GDC         2   14.025       1
## 53 Standard     GDC         1   10.456       1
## 54 Standard     GDC         1    9.991       1
## 55 Standard     GDC         1    7.172       1
## 56 Standard     GDC         1    5.518       1
## 57 Standard     GDC         2   19.427       1
## 58 Standard     GDC         1   15.701       1
```

From the exploratory data analysis we led previously, this point is far from being an outlier. In order to get the optimal number of clusters where this point is indeed intergrated to one of the main clusters instead of being on itself, we need to set the kmeans on 4 clusters:




```
## # A tibble: 4 x 2
##   classe   obs
##   <fct>  <int>
## 1 1         63
## 2 2         45
## 3 3         77
## 4 4         26
```

5.5) To Describe the clusters, you can use catdes function, by concatenating your dataset to the variable specifying in which cluster each observation is and indicating that you want to describe this variable (that must be as a factor).

5.6) Comment the results and describe precisely one cluster._

```
##                                       Cla/Mod   Mod/Cla    Global
## Raw.material=Al 4234                96.666667 92.063492 28.436019
## Cooling=Standard                    53.448276 98.412698 54.976303
## Process=GDC                         47.244094 95.238095 60.189573
## Over.molding=Yes                    44.262295 42.857143 28.909953
## Supplier.Country=Romania            64.285714 14.285714  6.635071
## Assembly=Yes                        40.789474 49.206349 36.018957
## Finishing=Other                     42.000000 33.333333 23.696682
## Supplier.Country=India              12.000000  4.761905 11.848341
## Assembly=No                         23.703704 50.793651 63.981043
## Over.molding=No                     24.000000 57.142857 71.090047
## Cooling=Air-cooled - Thermodecoupled  0.000000  0.000000  8.530806
## Supplier.Country=France              0.000000  0.000000 10.426540
## Raw.material=Al 4235                 0.000000  0.000000 11.374408
## Cooling=Air-cooled - Standard        3.225806  1.587302 14.691943
## Process=HPDC                         4.444444  3.174603 21.327014
```

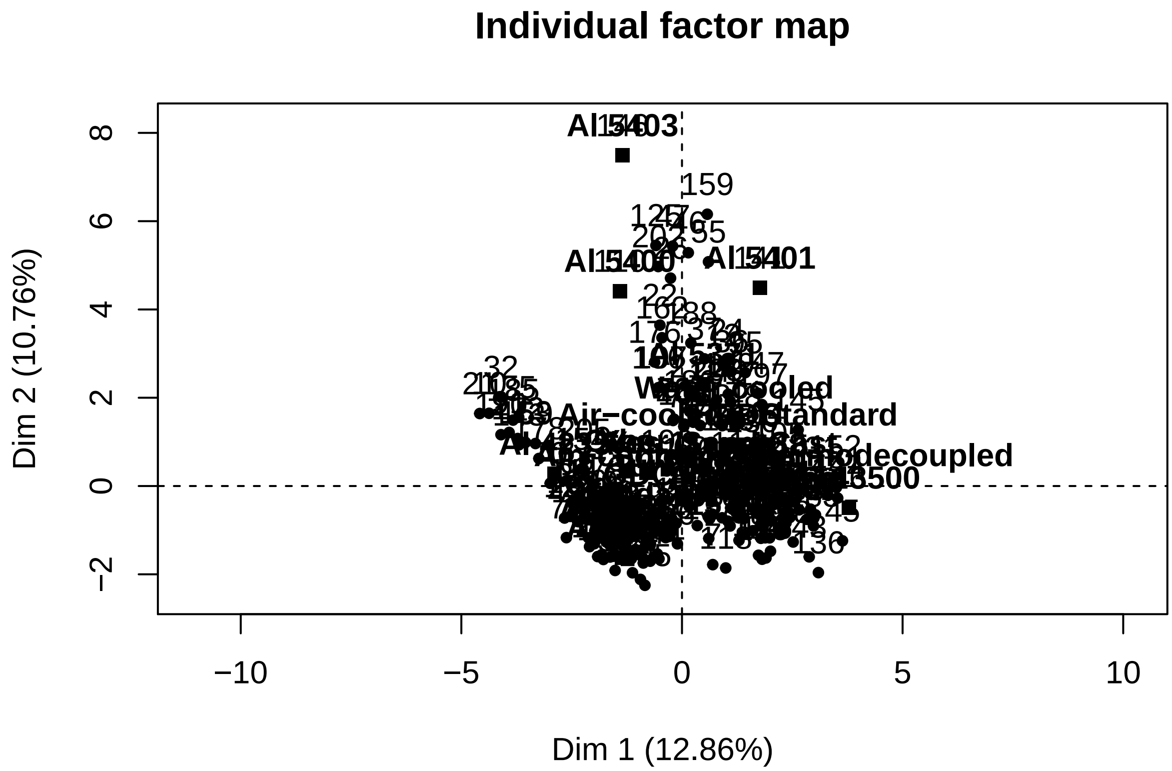
```
## Process=Sand Cast                         2.564103  1.587302 18.483412
## Raw.material=Al 5371                      2.272727  1.587302 20.853081
## Cooling=Water cooled                      0.000000  0.000000 21.800948
## Raw.material=AC 46000                     0.000000  0.000000 32.227488
##                                               p.value    v.test
## Raw.material=Al 4234                     2.463849e-43 13.802436
## Cooling=Standard                         1.025325e-19  9.086230
## Process=GDC                              2.324041e-13  7.328695
## Over.molding=Yes                         4.613978e-03  2.832817
## Supplier.Country=Romania                 7.306546e-03  2.682540
## Assembly=Yes                             1.070729e-02  2.552109
## Finishing=Other                          3.714986e-02  2.084113
## Supplier.Country=India                   3.318842e-02 -2.129796
## Assembly=No                              1.070729e-02 -2.552109
## Over.molding=No                          4.613978e-03 -2.832817
## Cooling=Air-cooled - Thermodecoupled 1.212940e-03 -3.235820
## Supplier.Country=France                  2.458899e-04 -3.666503
## Raw.material=Al 4235                     1.089937e-04 -3.869645
## Cooling=Air-cooled - Standard            1.074737e-04 -3.873068
## Process=HPDC                             4.274296e-06 -4.597579
## Process=Sand Cast                        4.118965e-06 -4.605287
## Raw.material=Al 5371                     4.746904e-07 -5.036270
## Cooling=Water cooled                     6.458596e-09 -5.804428
## Raw.material=AC 46000                    6.165617e-14 -7.504516
```

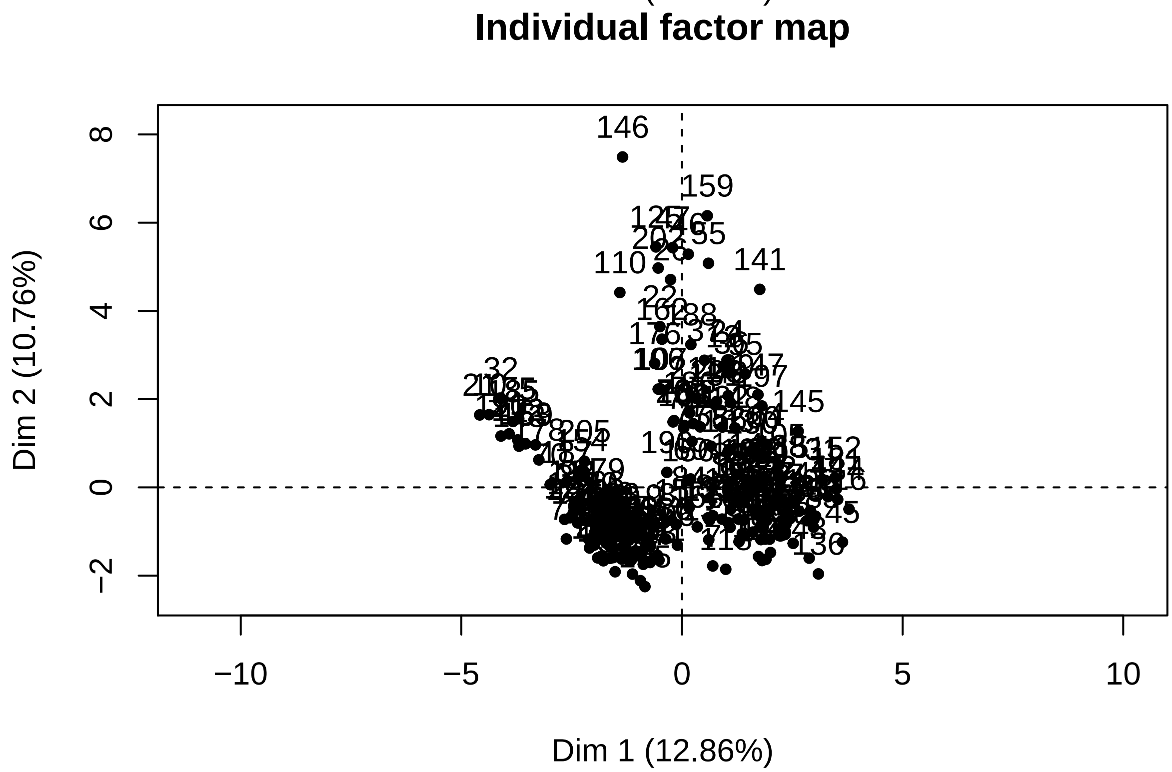
Since this cluster is big, I'll give the main characterisics. We can see for example that we will find 100% of the diecasting parts coming from mexico fall within our first cluster. We can also see that 95% of the parts produced from the AL 5371 material are in this same cluster. We are also certain that parts from Italy, India, made from either Al 4234, Al 4235, AC 46000 and gone through standard cooling are absolutely not within the first cluster. We can repeat this analysis based on the Cla/Mod column which gives us the percentage of observation with a specific characteristic which belong to a certain cluster.

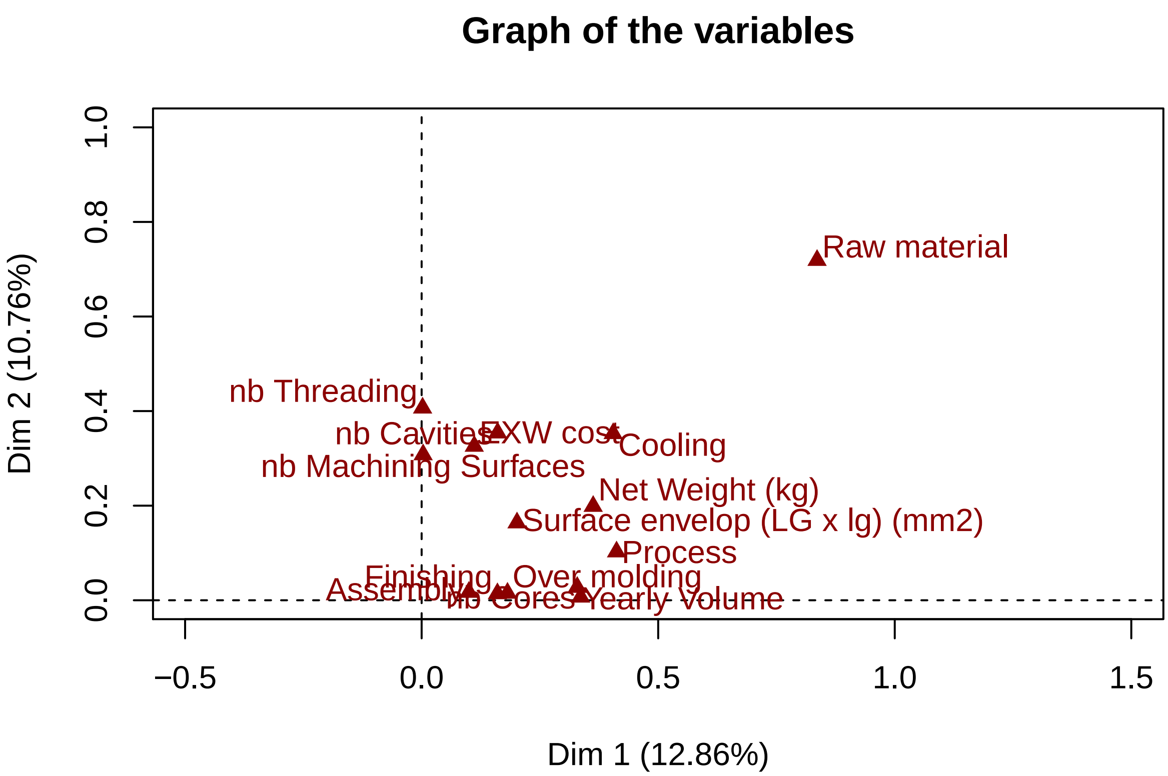
5.7) If someone asks you why you have selected k components to perform the clustering and not k+1 or k-1, what is your answer? (could you suggest a strategy to assess the stability of the approach? are there many differences between the clustering obtained on k components or on the initial data). You can have a look at the Rand Index.

We chose beforehand to compromise between a minimum amount of principle components in our FAMD in order to keep our dimension reductions and denoising to what it was meant for, but sufficiently enough of them to keep enough inertia to describe our data correctly. This led us to choosing 31 dimensions on which we would project our data, where 80% of the inertia was kept. Theoretically, we could decide to use k+1 or k-1 components but our choice of threshold was made on keeping 80% of the inertia. This depends how much you're ready to lose in inertia in order to denoize. To assess the stability of this approach we can compare the clustering on the raw data vs the denoized one for several level of inertias kept(i.e several thresholds of ncps). To compare this we will use the rand index which computes a ratio of similarities/similarities+dissimilarities to assess the differences we mentionned before.

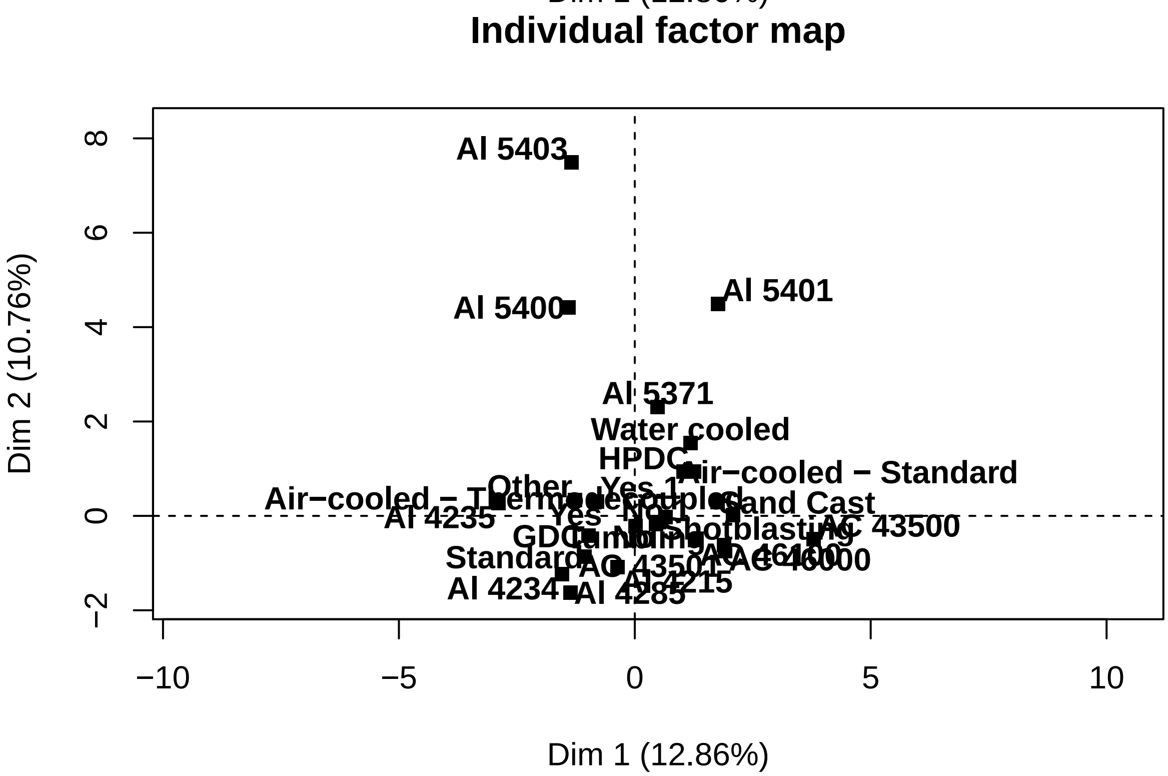
```r
# Different levels of ncp for FAMD

pc0 <- FAMD(don_famd, graph = TRUE, sup.var = c(1,2,18), ncp = 31)$ind$coord
```


Individual factor map
Dim 2 (10.76%)
Dim 1 (12.86%)
−10
−5
0
5
10
−2
0
2
4
6
8
Al1403
Al1400
Al1401
Water-cooled
Air-cooled
Standard
Decoupled
159
146
141
110
32
136

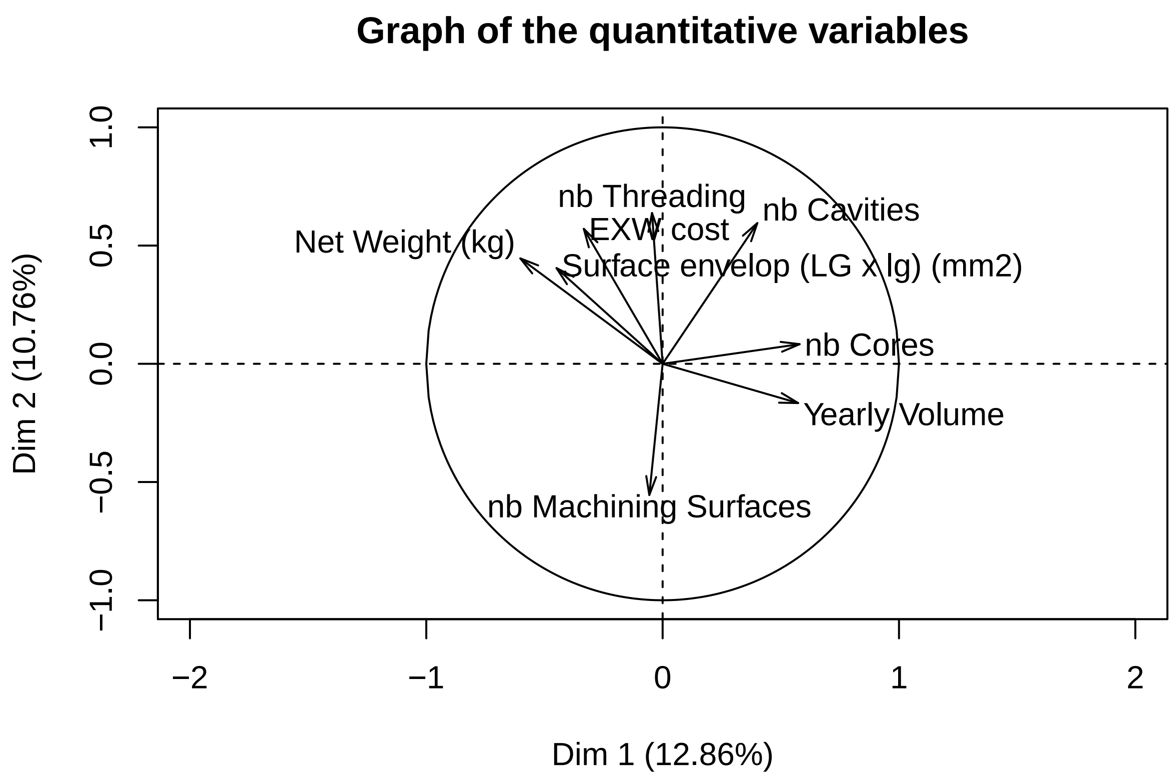

Individual factor map
Dim 2 (10.76%)
Dim 1 (12.86%)
−10
−5
0
5
10
−2
0
2
4
6
8
146
159
141
110
145
32
205
136



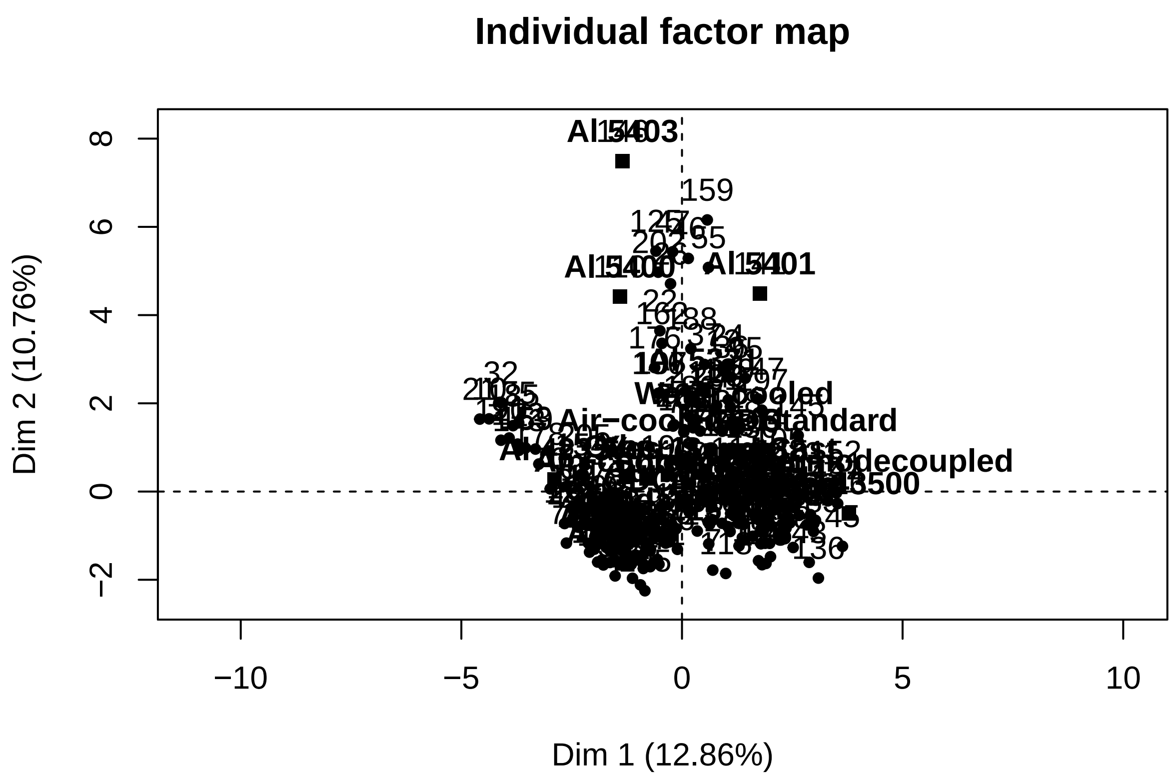
Graph of the variables
Raw material
nb Threading
nb Cavities
EXW cost
Cooling
nb Machining Surfaces
Net Weight (kg)
Surface envelop (LG x lg) (mm2)
Process
Finishing
Over molding
Assembly
nb Cores
Yearly Volume
Dim 2 (10.76%)
Dim 1 (12.86%)

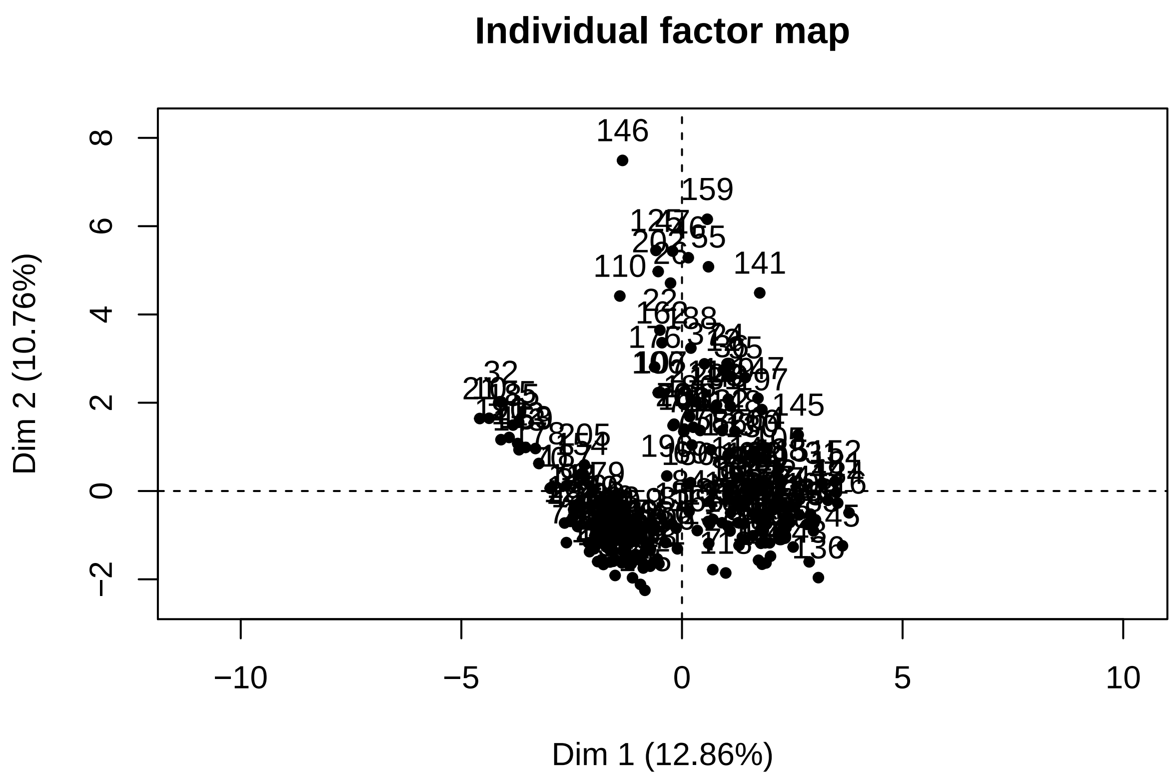


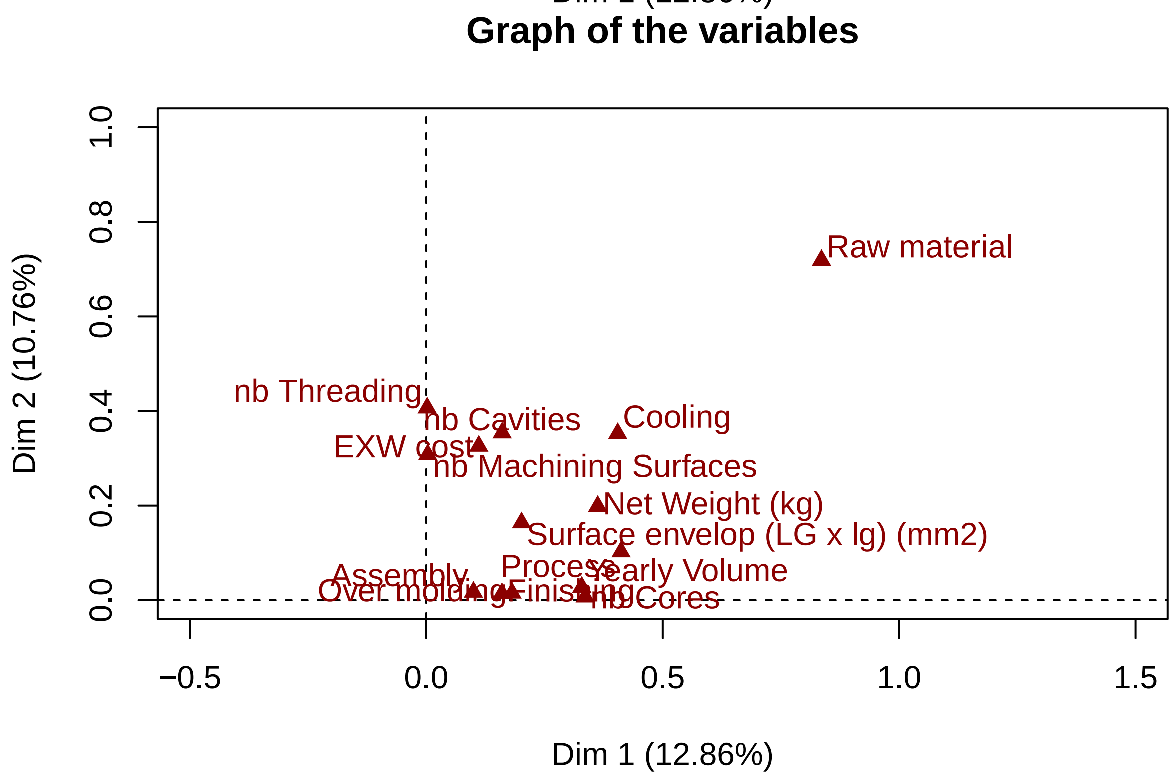
Individual factor map
Al 5403
Al 5400
Al 5401
Al 5371
Water cooled
HPDC
Air−cooled − Standard
Sand Cast
AC 43500
Al 4235
GDC
Standard
Al 4234
Al 4285
AC 46000
Dim 2 (10.76%)
Dim 1 (12.86%)

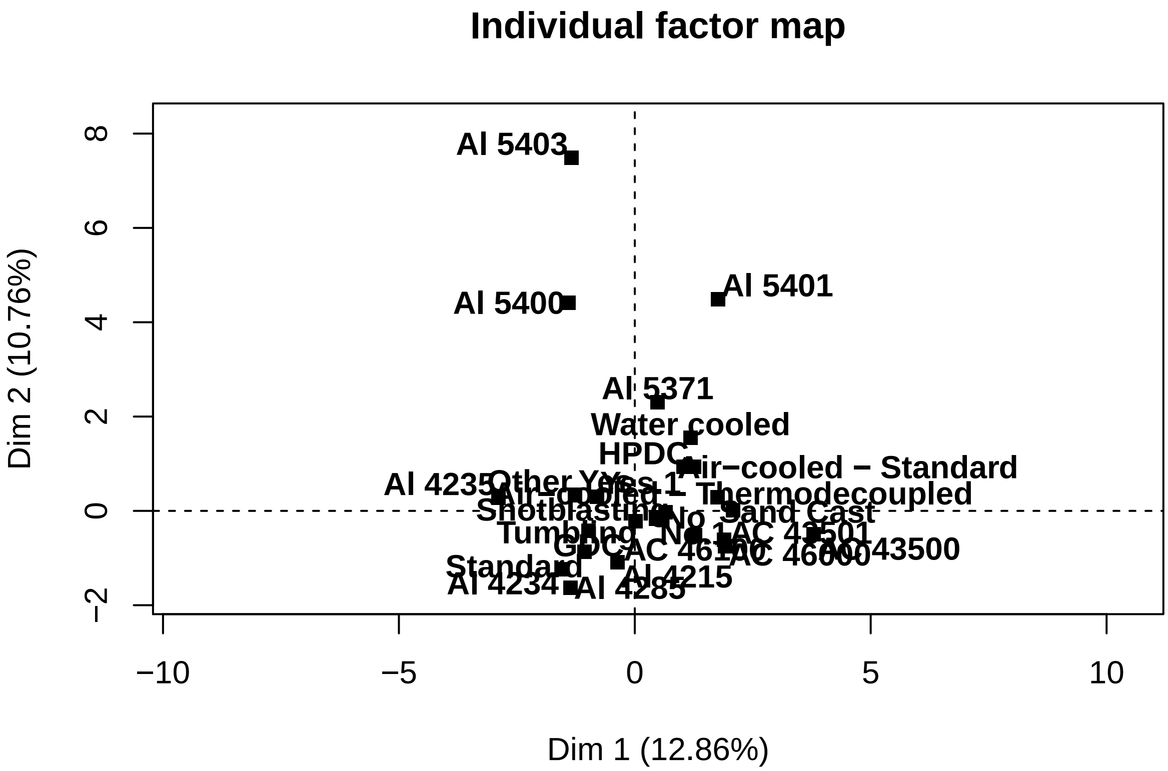


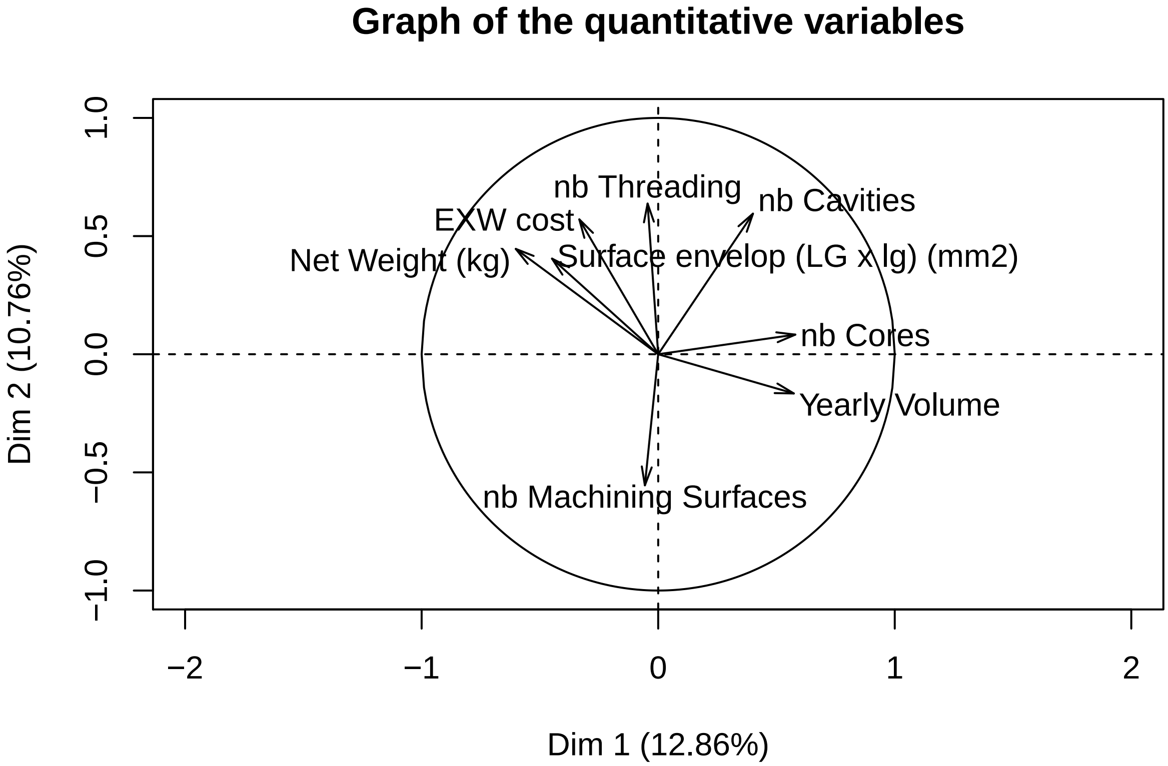

```
pc_low <- FAMD(don_famd, graph = TRUE, sup.var = c(1,2,18), ncp = 15)$ind$coord # Our original choice
```



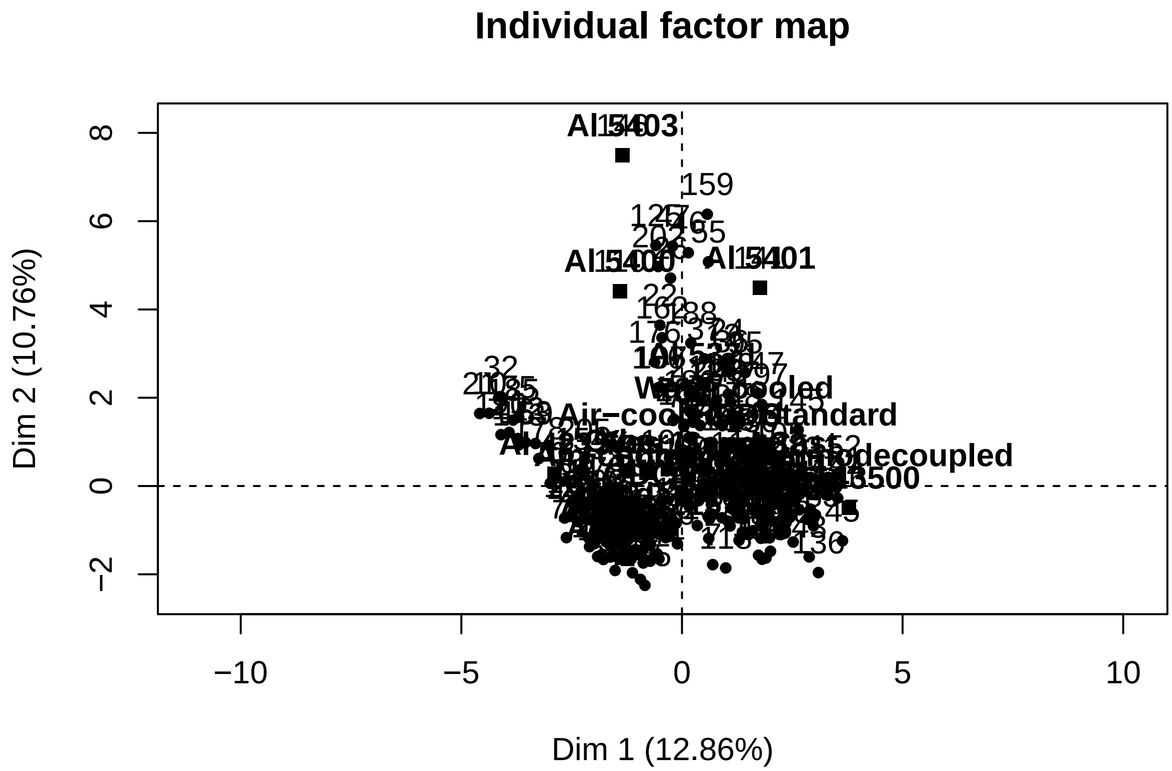

Individual factor map
Dim 2 (10.76%)
Dim 1 (12.86%)

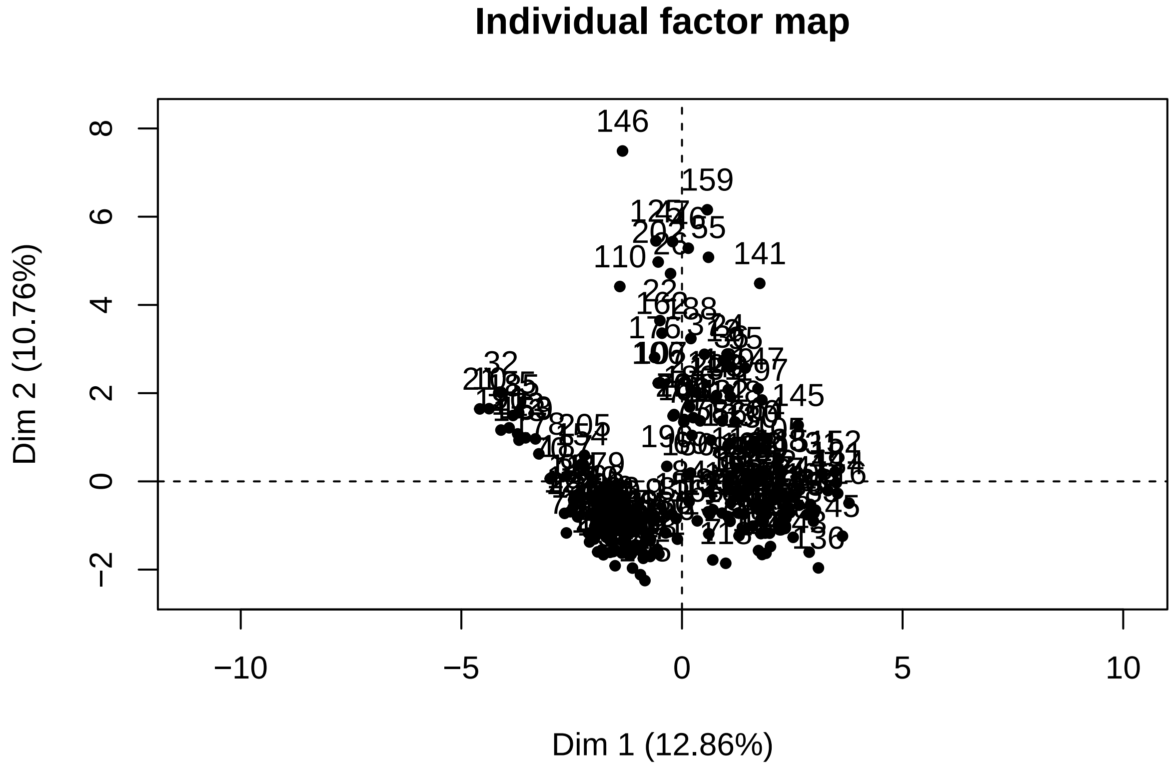

Graph of the variables
Dim 2 (10.76%)
Dim 1 (12.86%)
Raw material
nb Threading
nb Cavities
Cooling
EXW cost
nb Machining Surfaces
Net Weight (kg)
Surface envelop (LG x lg) (mm2)
Process
Yearly Volume
Assembly
Over molding
Finishing
nb Cores

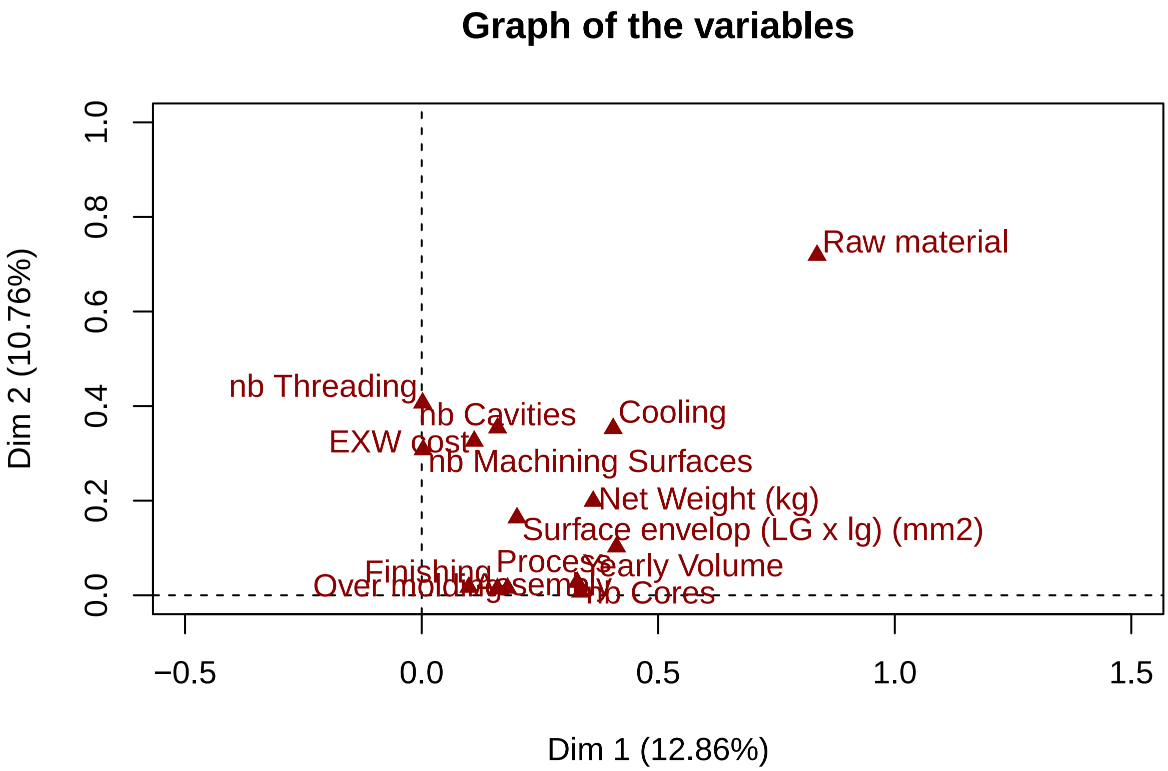


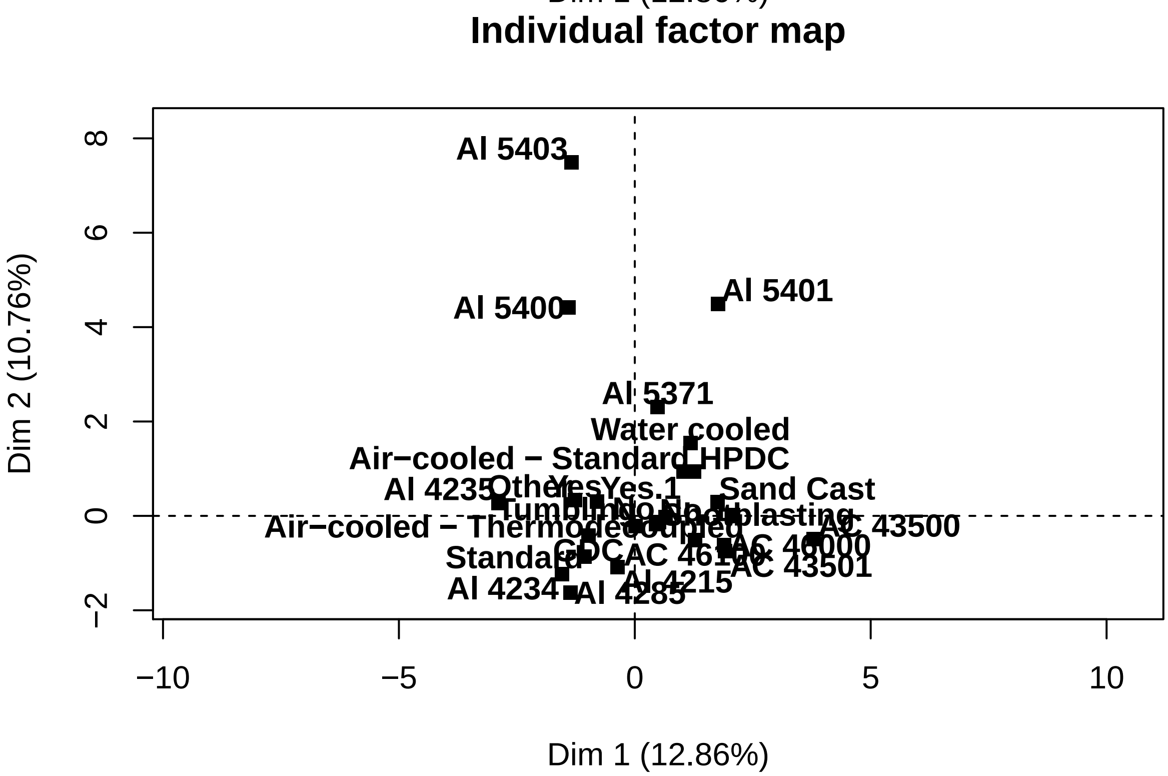


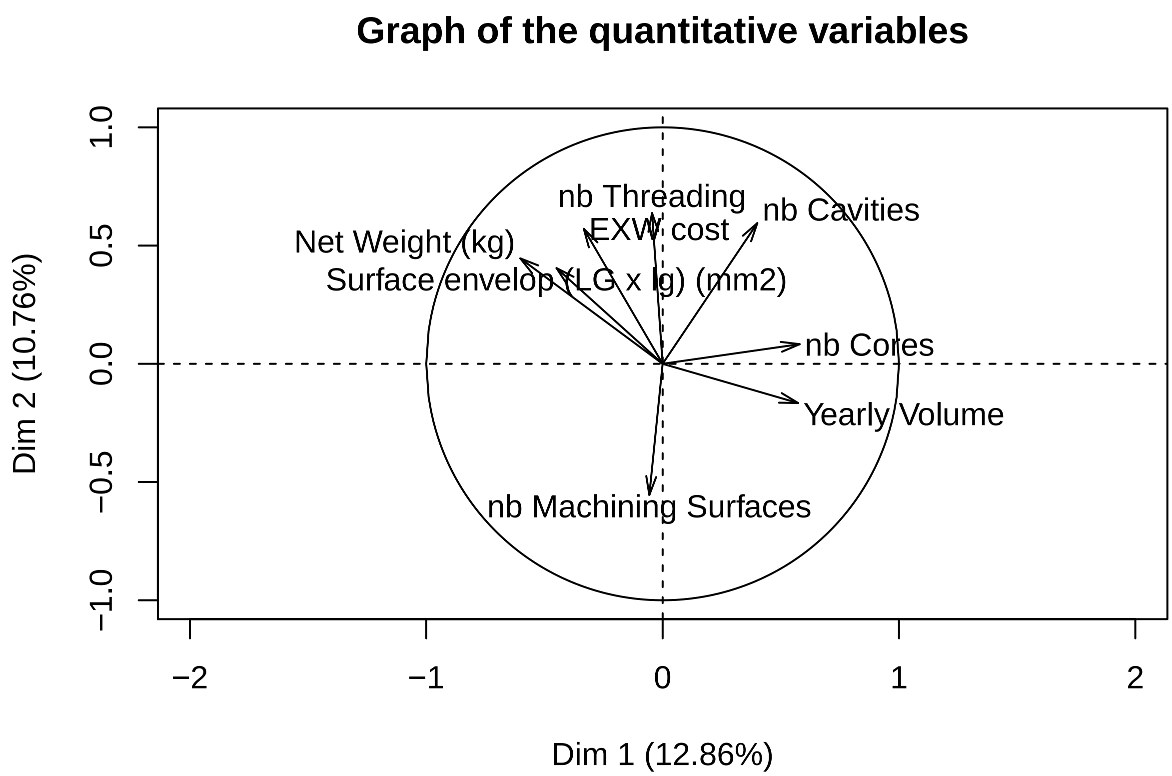

```
pc_high <- FAMD(don_famd, graph = TRUE, sup.var = c(1,2,18), ncp = 45)$ind$coord
```


Individual factor map
Dim 2 (10.76%)
Dim 1 (12.86%)
Water-cooled
Air-cooled
Standard
159
136


Individual factor map
Dim 2 (10.76%)
Dim 1 (12.86%)
146
159
110
141
145
136


Graph of the variables
Raw material
nb Threading
nb Cavities
Cooling
EXW cost
nb Machining Surfaces
Net Weight (kg)
Surface envelop (LG x lg) (mm2)
Process
Yearly Volume
Finishing
nb Cores
Dim 2 (10.76%)
Dim 1 (12.86%)
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
0.0
0.5
1.0
1.5


Individual factor map
Al 5403
Al 5400
Al 5401
Al 5371
Water cooled
Air−cooled − Standard
HPDC
Al 4235
Sand Cast
AC 43500
Air−cooled − Thermocoupled
AC 46000
AC 43501
Standard
Al 4234
Al 4215
Al 4285
Dim 2 (10.76%)
Dim 1 (12.86%)
−2
0
2
4
6
8
−10
−5
0
5
10




```
# Comparing the according clusterings:

rand.index(kmeans(pc0, centers = 4, nstart=100)$cluster, kmeans(pc_low, centers = 4, nstart=100)$cluste
```

```
## [1] 0.9905665
```

```
rand.index(kmeans(pc0, centers = 4, nstart=100)$cluster, kmeans(pc_high, centers = 4, nstart=100)$clust
```

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

We see that choosing much less principal components or much more does not change much in terms of clustering from the rand index. At least with our threshold we're not advancing blindfolded and are sure that we decided of ncp=31 in order to keep at least 80% of the inertia.

6) The methodology that you have used to describe clusters can also be used to describe a categorical variable, for instance the supplier country. Use the function catdes and explain how this information can be useful for the company.

This can be very interesting for the company indeed since we have a representation of the percentage of diecasting parts with a specific characteristic fall within parts coming from each country. If they are looking for oa diecasting part with a specific material used for its fabrication, they could find it using the info we have from the catdes function. To illustrate this let's look at the the parts coming from china :

```
## NULL
```

For example, if the company wants specifically parts made out of the material Al 4235, they will find 100 % of our data regarding these parts come from china, thus reducing the search and focusing on other characteristics.

7) Perform a model to predict the cost. Explain how the previous analysis can help you interpret the results.

First we need to transform some of the quantitative variables (scale them) in order to get better results from our model since they will act alongside categorical variables that will have levels that surely don't surpass the dozen:

Now that our varaibles are scaled, my idea is to make a regression per cluster. My though being that since we only have certain levels of a categorical variable in each cluster (for example china in cluster 1, not at all in 2,3 etc..) this could avoid running a regression on the whole data and having to pick varaibles (obviously they won't all fit in the regression) from this huge package. By runing models per cluster, we are focusing on specific data and the link between their characteristics and cost without biasing our view, since we aren't discarding varaibles, only discarding certain levels that aren't relevant to our data. I'll use the stepwise method to select among var-levels in each cluster, even though it's a greedy method it can give a first step of a model:

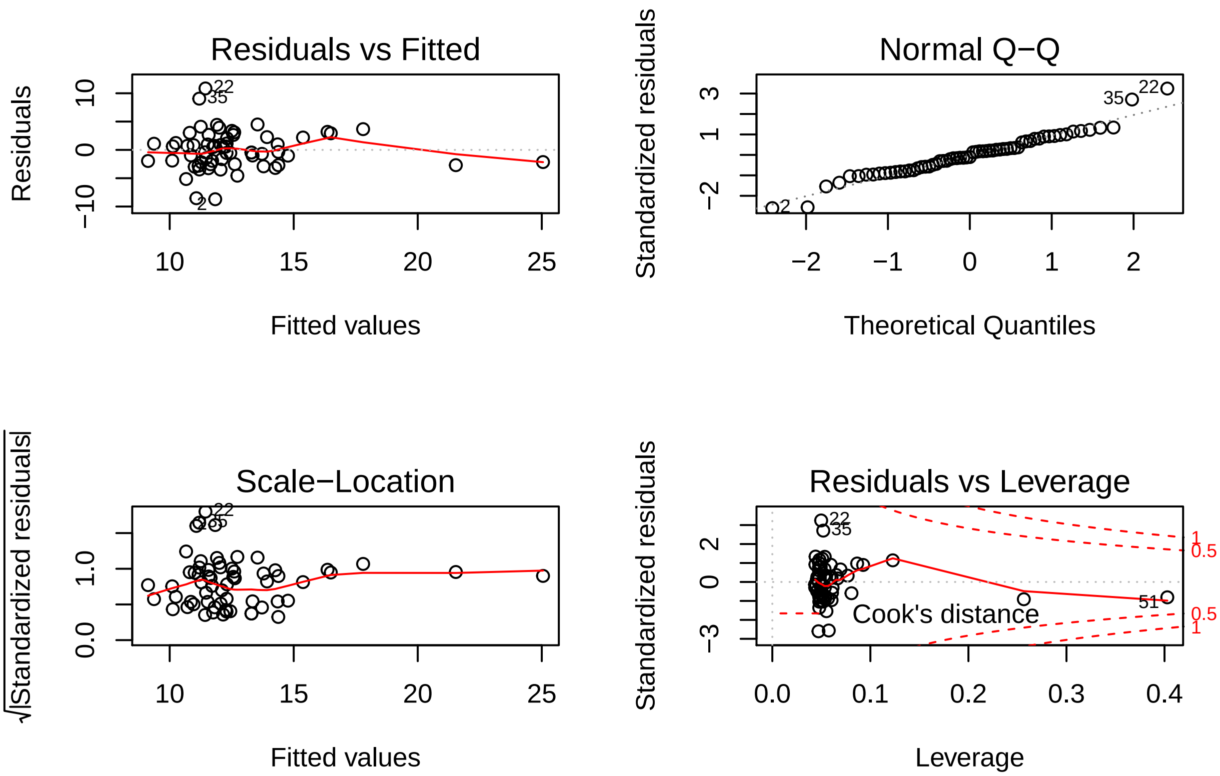
Model 1

```
## Start:  AIC=182.25
## `EXW cost` ~ 1
##
##                                     Df Sum of Sq     RSS    AIC
## + `Net Weight (kg)`                  1    399.36  701.95 155.88
## + `nb Cores`                         1    351.56  749.76 160.03
## + `Raw material`                     4    312.01  789.30 169.26
## + Process                            2    204.01  897.30 173.34
## + `nb Machining Surfaces`            1    105.13  996.19 177.93
## + `Supplier Country`                 6    249.50  851.82 178.07
## + `Surface envelop (LG x lg) (mm2)`  1     93.81 1007.50 178.64
## + `nb Cavities`                      1     80.18 1021.14 179.49
## + `Yearly Volume`                    1     60.09 1041.22 180.72
## <none>                                         1101.32 182.25
## + `nb Threading`                     1     14.29 1087.02 183.43
## + Cooling                            1      3.23 1098.09 184.07
## + Assembly                           1      0.10 1101.22 184.25
## + Finishing                          2      5.15 1096.16 185.96
## + Supplier                          18    232.63  868.69 203.30
##
## Step:  AIC=155.88
## `EXW cost` ~ `Net Weight (kg)`
##
##                                     Df Sum of Sq     RSS    AIC
## + `nb Cores`                         1    267.86  434.09 127.60
## + `Raw material`                     4    297.65  404.31 129.12
## + Process                            2    173.29  528.66 142.01
## + `nb Machining Surfaces`            1     98.22  603.74 148.38
## + `nb Cavities`                      1     72.56  629.39 151.00
## + `Supplier Country`                 6    163.50  538.45 151.17
## + `Surface envelop (LG x lg) (mm2)`  1     33.07  668.88 154.84
## <none>                                          701.95 155.88
## + `nb Threading`                     1     14.27  687.69 156.58
## + `Yearly Volume`                    1     13.28  688.68 156.67
## + Assembly                           1     12.57  689.38 156.74
## + Cooling                            1     11.87  690.08 156.80
## + Finishing                          2      7.33  694.62 159.21
## + Supplier                          18    133.11  568.85 178.63
## - `Net Weight (kg)`                  1    399.36 1101.32 182.25
##
## Step:  AIC=127.6
```

```
## `EXW cost` ~ `Net Weight (kg)` + `nb Cores`
##
##                                    Df Sum of Sq    RSS    AIC
## + `Raw material`                    4   117.106 316.99 115.79
## + Process                           2    78.095 356.00 119.10
## + `Supplier Country`                6    87.744 346.35 125.37
## + Assembly                          1    24.676 409.42 125.91
## <none>                                          434.09 127.60
## + `nb Cavities`                     1    12.032 422.06 127.83
## + `Surface envelop (LG x lg) (mm2)` 1    10.651 423.44 128.03
## + `Yearly Volume`                   1    10.304 423.79 128.08
## + `nb Threading`                    1     9.386 424.71 128.22
## + `nb Machining Surfaces`           1     5.735 428.36 128.76
## + Cooling                           1     3.959 430.13 129.02
## + Finishing                         2     8.327 425.77 130.38
## + Supplier                         18    80.838 353.25 150.62
## - `nb Cores`                        1   267.861 701.95 155.88
## - `Net Weight (kg)`                 1   315.664 749.76 160.03
##
## Step:  AIC=115.79
## `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material`
##
##                                    Df Sum of Sq    RSS    AIC
## + `Supplier Country`                6     83.21 233.78 108.61
## + `nb Machining Surfaces`           1     19.21 297.78 113.85
## + Assembly                          1     13.38 303.60 115.07
## + `Surface envelop (LG x lg) (mm2)` 1     11.77 305.22 115.41
## <none>                                          316.99 115.79
## + `nb Cavities`                     1      7.20 309.79 116.34
## + `Yearly Volume`                   1      4.80 312.18 116.83
## + `nb Threading`                    1      0.54 316.44 117.68
## + Finishing                         2      3.17 313.81 119.16
## - `Raw material`                    4    117.11 434.09 127.60
## - `nb Cores`                        1     87.32 404.31 129.12
## + Supplier                         18     60.08 256.91 138.55
## - `Net Weight (kg)`                 1    322.23 639.22 157.98
##
## Step:  AIC=108.61
## `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material` +
##     `Supplier Country`
##
##                                    Df Sum of Sq    RSS    AIC
## + `Yearly Volume`                   1    23.622 210.16 103.90
## + Assembly                          1    12.853 220.93 107.05
## <none>                                          233.78 108.61
## + `nb Machining Surfaces`           1     5.052 228.73 109.23
## + `nb Threading`                    1     2.762 231.02 109.86
## + `Surface envelop (LG x lg) (mm2)` 1     2.098 231.68 110.04
## + `nb Cavities`                     1     0.003 233.78 110.61
## + Finishing                         2     1.270 232.51 112.27
## - `Supplier Country`                6    83.205 316.99 115.79
## - `nb Cores`                        1    40.514 274.30 116.68
## - `Raw material`                    4   112.567 346.35 125.37
## + Supplier                         18    47.213 186.57 130.40
```

```
## - `Net Weight (kg)`                  1   285.756 519.54 156.92
##
## Step:  AIC=103.9
## `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material` +
##     `Supplier Country` + `Yearly Volume`
##
##                                     Df Sum of Sq    RSS    AIC
## + Assembly                           1    18.304 191.86 100.16
## + `nb Machining Surfaces`            1     7.142 203.02 103.72
## <none>                                           210.16 103.90
## + `Surface envelop (LG x lg) (mm2)`  1     2.823 207.34 105.05
## + `nb Cavities`                      1     2.385 207.77 105.18
## + `nb Threading`                     1     0.067 210.09 105.88
## + Finishing                          2     0.734 209.43 107.68
## - `Yearly Volume`                    1    23.622 233.78 108.61
## - `nb Cores`                         1    41.132 251.29 113.16
## - `Supplier Country`                 6   102.022 312.18 116.83
## - `Raw material`                     4   105.206 315.37 121.47
## + Supplier                          18    42.234 167.93 125.76
## - `Net Weight (kg)`                  1   258.946 469.11 152.49
##
## Step:  AIC=100.16
## `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material` +
##     `Supplier Country` + `Yearly Volume` + Assembly
##
##                                     Df Sum of Sq    RSS    AIC
## <none>                                           191.86 100.16
## + `Surface envelop (LG x lg) (mm2)`  1     2.535 189.32 101.32
## + `nb Cavities`                      1     2.168 189.69 101.44
## + `nb Machining Surfaces`            1     1.230 190.63 101.75
## + `nb Threading`                     1     0.320 191.54 102.05
## + Finishing                          2     0.798 191.06 103.89
## - Assembly                           1    18.304 210.16 103.90
## - `Yearly Volume`                    1    29.072 220.93 107.05
## - `nb Cores`                         1    51.601 243.46 113.16
## - `Raw material`                     4    88.024 279.88 115.95
## - `Supplier Country`                 6   106.661 298.52 116.01
## + Supplier                          18    39.359 152.50 121.69
## - `Net Weight (kg)`                  1   271.932 463.79 153.77

##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material` +
##     `Supplier Country` + `Yearly Volume` + Assembly, data = don_cluster_1)
##
## Coefficients:
##               (Intercept)           `Net Weight (kg)`
##                 1.301e+01                   4.502e+00
##                `nb Cores`       `Raw material`Al 4215
##                 1.109e+00                  -1.308e+01
##     `Raw material`Al 4234       `Raw material`Al 4285
##                -7.123e+00                  -8.677e+00
##     `Raw material`Al 5371     `Supplier Country`India
##                -9.172e+00                  -4.637e+00
```

```
##   `Supplier Country`Italia     `Supplier Country`Korea
##                  1.571e+00                   2.220e-01
##  `Supplier Country`Romania  `Supplier Country`Slovakia
##                 -3.711e-01                   2.340e+00
##  `Supplier Country`Vietnam             `Yearly Volume`
##                 -8.944e-01                  -8.432e-06
##                AssemblyYes
##                  1.212e+00

##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)` + Finishing, data = don_cluster_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.710 -2.169 -0.353  2.117 10.848
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.51022    1.12457   6.678 9.46e-09 ***
## `Net Weight (kg)`       4.91801    0.84211   5.840 2.38e-07 ***
## FinishingShotblasting   0.03175    1.04062   0.031    0.976
## FinishingTumbling      -0.72899    1.09512  -0.666    0.508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.431 on 59 degrees of freedom
## Multiple R-squared:  0.3693, Adjusted R-squared:  0.3372
## F-statistic: 11.51 on 3 and 59 DF,  p-value: 4.823e-06
```



The residuals aren't the best but they are ok, independance and constant variance seems to be approximately verified, we see some points with very high leverage but they remain within cook's distance so not much to worry about, and the residuals are approximately normally distributed according to the qqplot. Our model makes sense according to the p-value of the F-statistic, although the R2 is low this doesn't mean much in our case.

Model 2:

```
## Start:  AIC=261.65
## `EXW cost` ~ 1
##
##                                    Df Sum of Sq   RSS    AIC
## + `nb Threading`                    1    3768.0 10657 250.03
## + `nb Machining Surfaces`           1    2005.6 12419 256.92
## + `Yearly Volume`                   1    1953.3 12472 257.10
## + Cooling                           2    2138.4 12286 258.43
## + `nb Cavities`                     1     884.5 13540 260.81
## + `Supplier Country`                6    3507.0 10918 261.12
## + `Surface envelop (LG x lg) (mm2)`  1     721.3 13704 261.34
## <none>                                          14425 261.65
## + `Over molding`                    1     107.5 14317 263.31
## + `Net Weight (kg)`                 1      83.2 14342 263.39
## + Assembly                          1       3.8 14421 263.64
## + `nb Cores`                        1       0.5 14424 263.65
## + Process                           2     613.3 13812 263.70
## + Finishing                         2     320.3 14104 264.64
## + Supplier                         17    7153.9  7271 264.82
## + `Raw material`                    3      42.1 14383 267.52
##
## Step:  AIC=250.03
## `EXW cost` ~ `nb Threading`
##
##                                    Df Sum of Sq     RSS    AIC
## + `Yearly Volume`                   1    1093.4  9563.4 247.16
## + Cooling                           2    1423.0  9233.8 247.58
## + `nb Machining Surfaces`           1     765.3  9891.5 248.68
## + `nb Cavities`                     1     648.9 10007.9 249.20
## + `Over molding`                    1     598.9 10057.9 249.43
## + `Raw material`                    3    1358.2  9298.6 249.89
## <none>                                          10656.8 250.03
## + Process                           2     782.9  9873.9 250.59
## + `Net Weight (kg)`                 1     263.0 10393.8 250.90
## + `nb Cores`                        1     202.7 10454.1 251.16
## + `Surface envelop (LG x lg) (mm2)`  1     156.7 10500.1 251.36
## + Finishing                         2     561.5 10095.3 251.59
## + Assembly                          1       7.4 10649.4 252.00
## + `Supplier Country`                6    1783.8  8873.0 253.78
## + Supplier                         17    5066.1  5590.7 255.00
## - `nb Threading`                    1    3768.0 14424.8 261.65
##
## Step:  AIC=247.16
## `EXW cost` ~ `nb Threading` + `Yearly Volume`
```

```
##
##                                    Df Sum of Sq     RSS    AIC
## + Cooling                           2    1207.2  8356.2 245.08
## + `Over molding`                    1     637.3  8926.1 246.05
## + `nb Machining Surfaces`           1     540.8  9022.6 246.54
## + `Net Weight (kg)`                 1     438.7  9124.6 247.04
## <none>                                          9563.4 247.16
## + `nb Cavities`                     1     359.3  9204.1 247.43
## + `nb Cores`                        1     309.0  9254.4 247.68
## + Finishing                         2     651.3  8912.0 247.98
## + `Raw material`                    3     929.1  8634.2 248.56
## + `Surface envelop (LG x lg) (mm2)` 1      23.3  9540.1 249.05
## + Assembly                          1       2.4  9561.0 249.15
## - `Yearly Volume`                   1    1093.4 10656.8 250.03
## + Process                           2     219.8  9343.6 250.11
## + `Supplier Country`                6    1715.3  7848.0 250.26
## + Supplier                         17    4134.2  5429.1 255.68
## - `nb Threading`                    1    2908.1 12471.5 257.10
##
## Step:  AIC=245.08
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling
##
##                                    Df Sum of Sq     RSS    AIC
## + `Net Weight (kg)`                 1     888.5  7467.7 242.03
## + `Over molding`                    1     555.0  7801.2 243.99
## + `nb Cavities`                     1     426.5  7929.7 244.73
## + `nb Machining Surfaces`           1     415.5  7940.7 244.79
## <none>                                          8356.2 245.08
## + `Supplier Country`                6    1814.6  6541.6 246.07
## + `Raw material`                    3     773.3  7582.9 246.71
## + `Surface envelop (LG x lg) (mm2)` 1      11.2  8345.0 247.02
## + `nb Cores`                        1       9.3  8346.9 247.03
## + Assembly                          1       0.1  8356.1 247.08
## + Finishing                         2     359.6  7996.6 247.10
## - Cooling                           2    1207.2  9563.4 247.16
## - `Yearly Volume`                   1     877.6  9233.8 247.58
## + Process                           2     115.5  8240.7 248.46
## + Supplier                         17    3616.2  4740.0 253.57
## - `nb Threading`                    1    2642.3 10998.5 255.45
##
## Step:  AIC=242.03
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)`
##
##                                    Df Sum of Sq     RSS    AIC
## + `Surface envelop (LG x lg) (mm2)` 1     695.1  6772.6 239.63
## + `nb Machining Surfaces`           1     646.8  6820.9 239.95
## + `nb Cores`                        1     496.8  6970.9 240.93
## + `Raw material`                    3    1022.0  6445.7 241.40
## <none>                                          7467.7 242.03
## + `Over molding`                    1     280.3  7187.4 242.30
## + `Supplier Country`                6    1711.9  5755.8 242.31
## + `nb Cavities`                     1     169.8  7297.9 242.99
## + Finishing                         2     398.5  7069.1 243.56
## + Assembly                          1      17.2  7450.4 243.92
```

```
## + Process                          2     322.7  7145.0 244.04
## - `Net Weight (kg)`                1     888.5  8356.2 245.08
## - `Yearly Volume`                  1    1005.5  8473.2 245.71
## - Cooling                          2    1657.0  9124.6 247.04
## + Supplier                        17    3309.2  4158.5 249.68
## - `nb Threading`                   1    3382.6 10850.2 256.84
##
## Step:  AIC=239.63
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)`
##
##                                   Df Sum of Sq     RSS    AIC
## + `nb Cores`                       1     438.6  6334.0 238.62
## + `Raw material`                   3     964.1  5808.4 238.72
## + `nb Machining Surfaces`          1     385.8  6386.8 238.99
## <none>                                          6772.6 239.63
## + `Over molding`                   1     264.3  6508.3 239.84
## + `nb Cavities`                    1     214.5  6558.1 240.18
## + Assembly                         1      33.5  6739.1 241.41
## + Process                          2     286.9  6485.7 241.68
## - `Yearly Volume`                  1     661.5  7434.1 241.82
## - `Surface envelop (LG x lg) (mm2)`  1     695.1  7467.7 242.03
## + `Supplier Country`               6    1296.0  5476.6 242.07
## + Finishing                        2     217.3  6555.3 242.16
## + Supplier                        17    3186.1  3586.5 245.02
## - Cooling                          2    1881.1  8653.7 246.66
## - `Net Weight (kg)`                1    1572.4  8345.0 247.02
## - `nb Threading`                   1    3504.0 10276.6 256.39
##
## Step:  AIC=238.62
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores`
##
##                                   Df Sum of Sq     RSS    AIC
## + `nb Machining Surfaces`          1     823.6  5510.4 234.35
## + Supplier                        17    3506.5  2827.5 236.32
## + `Raw material`                   3    1028.6  5305.4 236.64
## <none>                                          6334.0 238.62
## + `nb Cavities`                    1     169.9  6164.1 239.39
## + `Over molding`                   1     137.6  6196.4 239.63
## - `nb Cores`                       1     438.6  6772.6 239.63
## + Assembly                         1      20.6  6313.4 240.47
## - `Surface envelop (LG x lg) (mm2)`  1     636.9  6970.9 240.93
## + Process                          2     217.0  6116.9 241.05
## - `Yearly Volume`                  1     812.2  7146.1 242.04
## + Finishing                        2      55.7  6278.3 242.22
## + `Supplier Country`               6     966.7  5367.3 243.16
## - Cooling                          2    2182.4  8516.3 247.94
## - `Net Weight (kg)`                1    1986.6  8320.5 248.89
## - `nb Threading`                   1    3895.6 10229.5 258.19
##
## Step:  AIC=234.35
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces`
```

```
##
##                                   Df Sum of Sq    RSS    AIC
## + Supplier                        17    3317.8 2192.6 226.88
## + `Raw material`                   3     908.4 4602.0 232.24
## <none>                                          5510.4 234.35
## - `Surface envelop (LG x lg) (mm2)`  1     269.6 5780.0 234.50
## + `nb Cavities`                    1     171.2 5339.2 234.93
## + Process                          2     349.9 5160.5 235.40
## + `Over molding`                   1      31.8 5478.6 236.09
## + Assembly                         1       7.0 5503.4 236.29
## + Finishing                        2      37.8 5472.6 238.04
## - `Yearly Volume`                  1     805.4 6315.8 238.49
## - `nb Machining Surfaces`          1     823.6 6334.0 238.62
## - `nb Cores`                       1     876.4 6386.8 238.99
## + `Supplier Country`               6     782.3 4728.1 239.46
## - Cooling                          2    2537.3 8047.7 247.39
## - `Net Weight (kg)`                1    2422.4 7932.8 248.74
## - `nb Threading`                   1    3382.4 8892.8 253.88
##
## Step:  AIC=226.88
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces` +
##     Supplier
##
##                                   Df Sum of Sq    RSS    AIC
## + `Raw material`                   3     848.2 1344.4 210.87
## + `Supplier Country`               6     701.8 1490.8 221.52
## + `Over molding`                   1     213.1 1979.5 224.28
## + Process                          2     255.2 1937.4 225.31
## <none>                                          2192.6 226.88
## + Assembly                         1      79.2 2113.4 227.22
## + `nb Cavities`                    1       5.7 2186.9 228.76
## + Finishing                        2       9.9 2182.7 230.68
## - `Yearly Volume`                  1     385.8 2578.5 232.17
## - `Surface envelop (LG x lg) (mm2)`  1     422.6 2615.2 232.81
## - Supplier                        17    3317.8 5510.4 234.35
## - `nb Machining Surfaces`          1     634.9 2827.5 236.32
## - `nb Cores`                       1    1219.0 3411.6 244.77
## - `nb Threading`                   1    1840.1 4032.7 252.30
## - Cooling                          2    2434.6 4627.3 256.49
## - `Net Weight (kg)`                1    2317.0 4509.6 257.33
##
## Step:  AIC=210.87
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces` +
##     Supplier + `Raw material`
##
##                                   Df Sum of Sq    RSS    AIC
## + Process                          2     411.3  933.1 198.43
## + `Supplier Country`               6     549.2  795.2 199.24
## + Assembly                         1     100.4 1244.0 209.38
## <none>                                          1344.4 210.87
## + `Over molding`                   1      21.5 1322.9 212.14
## + `nb Cavities`                    1       0.0 1344.4 212.87
```

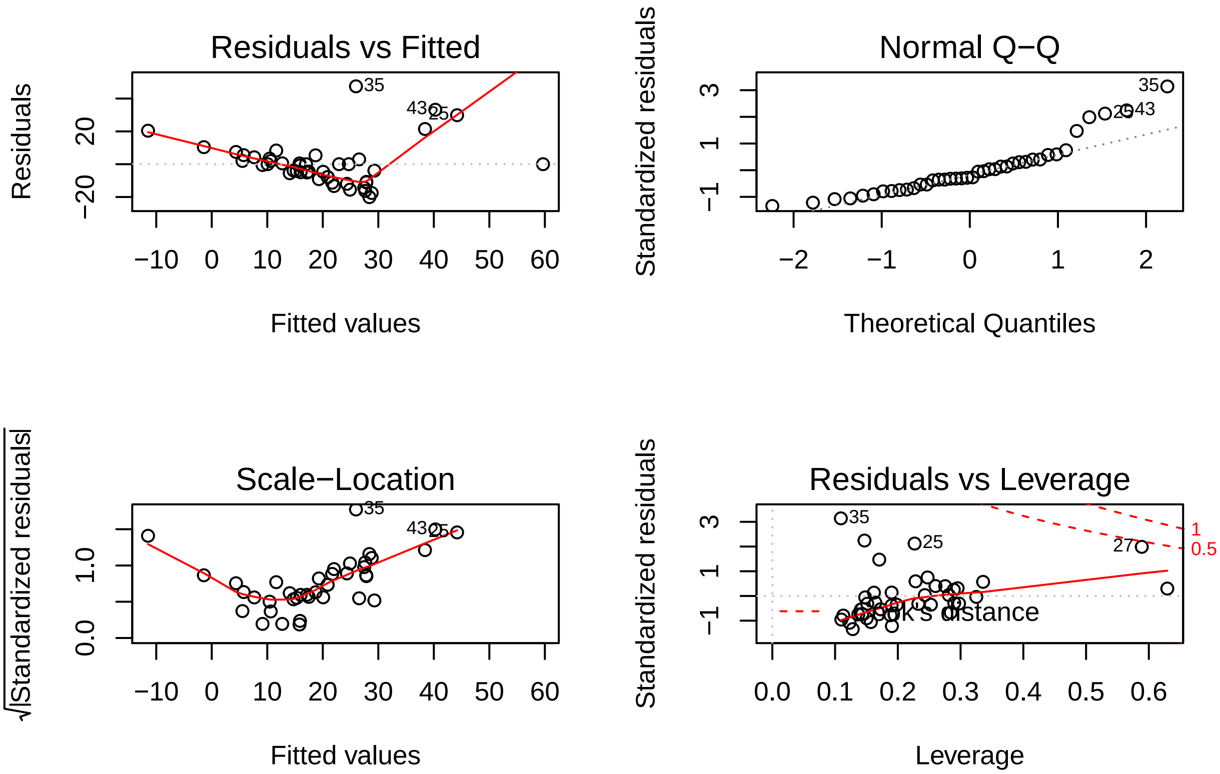
```
## + Finishing                         2      26.8 1317.7 213.96
## - `Yearly Volume`                   1     281.7 1626.1 217.43
## - `Surface envelop (LG x lg) (mm2)`  1     561.0 1905.5 224.56
## - `nb Machining Surfaces`           1     643.8 1988.3 226.48
## - `Raw material`                    3     848.2 2192.6 226.88
## - Supplier                         17    3257.5 4602.0 232.24
## - `nb Cores`                        1    1373.2 2717.6 240.54
## - `nb Threading`                    1    1839.5 3183.9 247.66
## - Cooling                           2    2723.4 4067.8 256.69
## - `Net Weight (kg)`                 1    2691.7 4036.1 258.34
##
## Step:  AIC=198.43
## `EXW cost` ~ `nb Threading` + `Yearly Volume` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces` +
##     Supplier + `Raw material` + Process
##
##                                    Df Sum of Sq    RSS    AIC
## - `Yearly Volume`                   1       5.5  938.6 196.70
## <none>                                           933.1 198.43
## + Assembly                          1      20.8  912.3 199.42
## + `Over molding`                    1      17.6  915.5 199.57
## + `nb Cavities`                     1       4.4  928.7 200.22
## + Finishing                         2      32.3  900.8 200.85
## + `Supplier Country`                6     166.4  766.7 201.59
## - Process                           2     411.3 1344.4 210.87
## - `Surface envelop (LG x lg) (mm2)`  1     437.1 1370.2 213.72
## - `Raw material`                    3    1004.3 1937.4 225.31
## - `nb Machining Surfaces`           1     933.4 1866.5 227.63
## - Supplier                         17    3283.5 4216.6 232.31
## - `nb Cores`                        1    1247.9 2181.0 234.64
## - Cooling                           2    2062.4 2995.5 246.92
## - `nb Threading`                    1    2018.1 2951.2 248.25
## - `Net Weight (kg)`                 1    2260.6 3193.7 251.80
##
## Step:  AIC=196.7
## `EXW cost` ~ `nb Threading` + Cooling + `Net Weight (kg)` + `Surface envelop (LG x lg) (mm2)` +
##     `nb Cores` + `nb Machining Surfaces` + Supplier + `Raw material` +
##     Process
##
##                                    Df Sum of Sq    RSS    AIC
## <none>                                           938.6 196.70
## + `Over molding`                    1      17.2  921.3 197.86
## + Assembly                          1      13.3  925.2 198.05
## + `Yearly Volume`                   1       5.5  933.1 198.43
## + `nb Cavities`                     1       2.6  936.0 198.57
## + Finishing                         2      34.1  904.4 199.03
## + `Supplier Country`                6     124.9  813.7 202.27
## - `Surface envelop (LG x lg) (mm2)`  1     432.3 1370.8 211.74
## - Process                           2     687.6 1626.1 217.43
## - `Raw material`                    3    1033.4 1971.9 224.10
## - `nb Machining Surfaces`           1     977.1 1915.6 226.80
## - Supplier                         17    3482.9 4421.5 232.44
## - `nb Cores`                        1    1265.9 2204.5 233.12
## - Cooling                           2    2124.1 3062.6 245.92
```

```
## - `nb Threading`                   1    2015.4 2954.0 246.29
## - `Net Weight (kg)`                1    2255.2 3193.8 249.80
```

```
##
## Call:
## lm(formula = `EXW cost` ~ `nb Threading` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces` +
##     Supplier + `Raw material` + Process, data = don_cluster_2)
##
## Coefficients:
##                        (Intercept)                       `nb Threading`
##                          9.299e+00                            2.460e+00
## CoolingAir-cooled - Thermodecoupled                 CoolingWater cooled
##                          2.294e+01                           -4.140e+01
##                  `Net Weight (kg)`    `Surface envelop (LG x lg) (mm2)`
##                         -2.726e+01                            4.235e-04
##                         `nb Cores`              `nb Machining Surfaces`
##                          1.369e+01                           -1.178e+00
##            SupplierAlcyon Supplier            SupplierCarcajou Supplier
##                          2.000e+01                            2.342e+01
##          SupplierChanceux Supplier          SupplierConception Supplier
##                          2.590e+01                            9.117e+00
##           SupplierConduit Supplier         SupplierConvergence Supplier
##                         -3.324e+00                            1.959e+01
##          SupplierDowntown Supplier           SupplierExcalibur Supplier
##                          1.520e+01                            2.332e+01
##           SupplierGalileo Supplier           SupplierHollywood Supplier
##                          1.108e+01                            4.574e+01
##        SupplierImaginaire Supplier         SupplierLes espaces Supplier
##                          1.752e+00                            1.785e+01
##              SupplierNord Supplier              SupplierOneUp Supplier
##                          2.455e+01                           -1.599e+00
##            SupplierOptima Supplier             SupplierSedona Supplier
##                          1.836e+01                            1.259e+01
##             SupplierWorld Supplier              `Raw material`Al 5400
##                         -9.722e+00                            4.177e+01
##              `Raw material`Al 5401              `Raw material`Al 5403
##                         -1.793e+01                           -3.706e+00
##                        ProcessHPDC                      ProcessSand Cast
##                          5.507e+00                            1.556e+01
```

```
##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)` + `nb Cores` + `Raw material` +
##     `Supplier Country` + `Yearly Volume` + Assembly, data = don_cluster_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.061  -7.837  -0.498   3.390  47.453
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.429e+01  9.902e+00   4.473 9.65e-05 ***
## `Net Weight (kg)`         -1.712e+00  4.612e+00  -0.371  0.71306
## `nb Cores`                -3.293e+00  1.936e+00  -1.701  0.09901 .
```

```
## `Raw material`Al 5400      -2.049e-01  1.963e+01  -0.010  0.99174
## `Raw material`Al 5401      -1.171e+01  1.695e+01  -0.691  0.49495
## `Raw material`Al 5403       1.719e+01  2.090e+01   0.822  0.41710
## `Supplier Country`France   -2.505e+01  1.732e+01  -1.446  0.15815
## `Supplier Country`Korea    -1.471e+01  8.663e+00  -1.698  0.09956 .
## `Supplier Country`Mexico   -1.960e+01  9.524e+00  -2.058  0.04807 *
## `Supplier Country`Romania  -2.291e+01  9.379e+00  -2.442  0.02049 *
## `Supplier Country`Slovakia -1.483e+01  7.150e+00  -2.075  0.04639 *
## `Supplier Country`Vietnam   3.330e+01  1.721e+01   1.935  0.06215 .
## `Yearly Volume`            -6.844e-05  2.313e-05  -2.960  0.00586 **
## AssemblyYes                 6.499e+00  5.593e+00   1.162  0.25408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.01 on 31 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.2182
## F-statistic: 1.945 on 13 and 31 DF,  p-value: 0.06374
```



Same as model one, we get some good posteriori confirmation of our residuals, a lot of the variables within our model are strongly individually significant, the p-value of the F statistic is very low so our model makes sense.

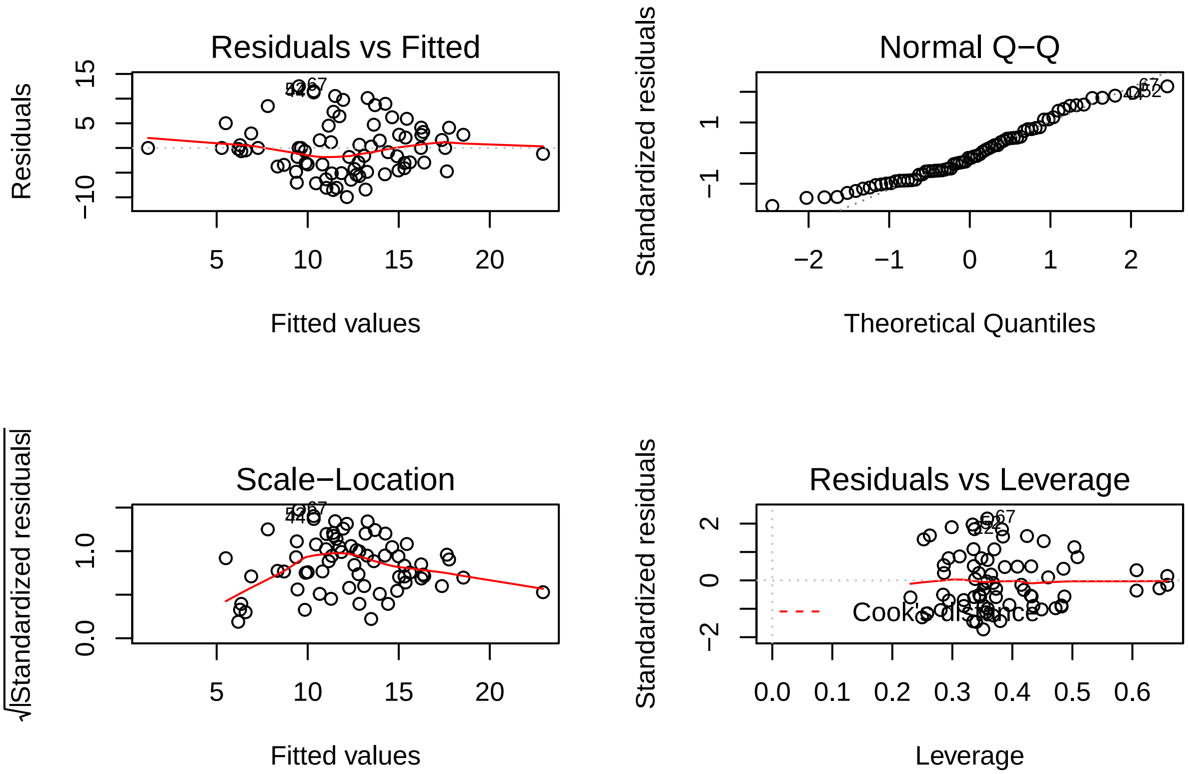
Model 3:

```
## Start:  AIC=289.04
## `EXW cost` ~ 1
##
##                                       Df Sum of Sq     RSS     AIC
```

```
## + `Net Weight (kg)`                    1    154.16 3047.9 287.24
## <none>                                             3202.0 289.04
## + `Surface envelop (LG x lg) (mm2)`  1     57.74 3144.3 289.63
## + Process                            2    137.89 3064.1 289.65
## + `Over molding`                     1     46.22 3155.8 289.92
## + `nb Threading`                     1     40.19 3161.8 290.06
## + `Yearly Volume`                    1     19.36 3182.7 290.57
## + `nb Cores`                         1     19.19 3182.8 290.57
## + Finishing                          2     96.69 3105.3 290.68
## + `nb Cavities`                      1     11.20 3190.8 290.77
## + `nb Machining Surfaces`            1      6.19 3195.8 290.89
## + Assembly                           1      1.49 3200.5 291.00
## + `Raw material`                     4    171.33 3030.7 292.80
## + `Supplier Country`                 5    246.07 2956.0 292.88
## + Cooling                            3     32.86 3169.2 294.24
## + Supplier                          19    638.03 2564.0 309.93
##
## Step:  AIC=287.24
## `EXW cost` ~ `Net Weight (kg)`
##
##                                     Df Sum of Sq    RSS    AIC
## + Finishing                          2    158.01 2889.9 287.14
## + `Over molding`                     1     78.53 2969.3 287.23
## <none>                                             3047.9 287.24
## + `Yearly Volume`                    1     55.72 2992.1 287.81
## + `Surface envelop (LG x lg) (mm2)`  1     33.48 3014.4 288.39
## + `nb Cavities`                      1     29.94 3017.9 288.48
## + Process                            2    104.18 2943.7 288.56
## + `nb Cores`                         1     16.15 3031.7 288.83
## - `Net Weight (kg)`                  1    154.16 3202.0 289.04
## + `nb Threading`                     1      5.39 3042.5 289.10
## + `nb Machining Surfaces`            1      4.47 3043.4 289.12
## + Assembly                           1      2.33 3045.5 289.18
## + `Supplier Country`                 5    282.48 2765.4 289.75
## + `Raw material`                     4    135.34 2912.5 291.74
## + Cooling                            3     20.58 3027.3 292.71
## + Supplier                          19    607.68 2440.2 308.11
##
## Step:  AIC=287.14
## `EXW cost` ~ `Net Weight (kg)` + Finishing
##
##                                     Df Sum of Sq    RSS    AIC
## + `Over molding`                     1     74.83 2815.0 287.12
## <none>                                             2889.9 287.14
## - Finishing                          2    158.01 3047.9 287.24
## + `Surface envelop (LG x lg) (mm2)`  1     22.05 2867.8 288.55
## + `Yearly Volume`                    1     13.38 2876.5 288.78
## + `nb Cores`                         1     12.51 2877.3 288.80
## + Process                            2     82.79 2807.1 288.90
## + `nb Cavities`                      1      7.35 2882.5 288.94
## + `nb Threading`                     1      6.52 2883.3 288.96
## + `nb Machining Surfaces`            1      5.07 2884.8 289.00
## + Assembly                           1      4.08 2885.8 289.03
## + `Supplier Country`                 5    270.03 2619.8 289.58
```

```
## - `Net Weight (kg)`                  1    215.48 3105.3 290.68
## + `Raw material`                     4    135.96 2753.9 291.43
## + Cooling                            3     34.40 2855.5 292.21
## + Supplier                          19    686.30 2203.6 304.26
##
## Step:  AIC=287.12
## `EXW cost` ~ `Net Weight (kg)` + Finishing + `Over molding`
##
##                                    Df Sum of Sq    RSS    AIC
## <none>                                          2815.0 287.12
## - `Over molding`                     1     74.83 2889.9 287.14
## - Finishing                          2    154.31 2969.3 287.23
## + `Surface envelop (LG x lg) (mm2)`  1     22.91 2792.1 288.49
## + `nb Threading`                     1     11.47 2803.6 288.80
## + `nb Cores`                         1      9.93 2805.1 288.85
## + `nb Machining Surfaces`            1      8.28 2806.8 288.89
## + `nb Cavities`                      1      2.64 2812.4 289.05
## + Assembly                           1      1.83 2813.2 289.07
## + `Yearly Volume`                    1      0.98 2814.0 289.09
## + Process                            2     70.78 2744.2 289.16
## + `Supplier Country`                 5    263.16 2551.9 289.56
## + `Raw material`                     4    127.34 2687.7 291.55
## - `Net Weight (kg)`                  1    253.15 3068.2 291.75
## + Cooling                            3     33.78 2781.2 292.19
## + Supplier                          19    630.79 2184.2 305.58

##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)` + Finishing + `Over molding`,
##     data = don_cluster_3)
##
## Coefficients:
##          (Intercept)      `Net Weight (kg)`  FinishingShotblasting
##               15.224                  4.405                 -7.309
##    FinishingTumbling      `Over molding`Yes
##               -6.067                  8.939

##
## Call:
## lm(formula = `EXW cost` ~ `nb Threading` + Cooling + `Net Weight (kg)` +
##     `Surface envelop (LG x lg) (mm2)` + `nb Cores` + `nb Machining Surfaces` +
##     Supplier + `Raw material` + Process, data = don_cluster_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9512 -4.1008 -0.2044  2.9443 12.5115
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.052e+01  1.031e+01   1.020    0.313
## `nb Threading`                      6.266e-02  1.098e+00   0.057    0.955
## CoolingAir-cooled - Thermodecoupled  3.932e-01  3.231e+00   0.122    0.904
## CoolingStandard                     8.909e-02  2.883e+00   0.031    0.975
## CoolingWater cooled                -1.110e+00  3.142e+00  -0.353    0.726
## `Net Weight (kg)`                   3.097e+00  5.312e+00   0.583    0.563
```

```
## `Surface envelop (LG x lg) (mm2)`   -6.283e-06  1.391e-04  -0.045    0.964
## `nb Cores`                           3.745e-02  5.204e-01   0.072    0.943
## `nb Machining Surfaces`             -4.006e-02  1.224e-01  -0.327    0.745
## SupplierAlcyon Supplier             -1.336e+00  6.363e+00  -0.210    0.835
## SupplierCarcajou Supplier            1.354e+00  5.334e+00   0.254    0.801
## SupplierChanceux Supplier            4.098e+00  8.485e+00   0.483    0.632
## SupplierConception Supplier          8.321e-01  5.037e+00   0.165    0.870
## SupplierConduit Supplier             7.288e+00  6.237e+00   1.168    0.249
## SupplierConvergence Supplier         3.952e-01  5.078e+00   0.078    0.938
## SupplierDowntown Supplier            1.166e+01  8.223e+00   1.418    0.163
## SupplierExcalibur Supplier           2.651e+00  3.947e+00   0.672    0.505
## SupplierFull house Supplier          5.744e+00  4.710e+00   1.220    0.229
## SupplierGalileo Supplier             5.700e+00  5.724e+00   0.996    0.325
## SupplierHollywood Supplier           1.265e+00  4.504e+00   0.281    0.780
## SupplierImaginaire Supplier          7.681e+00  5.490e+00   1.399    0.169
## SupplierLes espaces Supplier        -8.635e-01  4.376e+00  -0.197    0.844
## SupplierMillionDollar Supplier       3.359e+00  8.397e+00   0.400    0.691
## SupplierNord Supplier               -5.057e+00  8.377e+00  -0.604    0.549
## SupplierOneUp Supplier               4.053e+00  4.212e+00   0.962    0.341
## SupplierOptima Supplier              1.199e+00  4.723e+00   0.254    0.801
## SupplierSedona Supplier              7.116e+00  5.851e+00   1.216    0.231
## SupplierWorld Supplier               7.398e+00  5.118e+00   1.446    0.156
## `Raw material`AC 43501              -4.161e+00  1.154e+01  -0.361    0.720
## `Raw material`AC 46000              -1.893e+00  9.391e+00  -0.202    0.841
## `Raw material`AC 46100              -2.687e+00  1.009e+01  -0.266    0.791
## `Raw material`Al 5371               -6.495e+00  1.260e+01  -0.515    0.609
## ProcessHPDC                         -3.257e+00  2.453e+00  -1.328    0.191
## ProcessSand Cast                     6.688e-01  2.368e+00   0.282    0.779
##
## Residual standard error: 7.165 on 43 degrees of freedom
## Multiple R-squared:  0.3106, Adjusted R-squared:  -0.2185
## F-statistic: 0.587 on 33 and 43 DF,  p-value: 0.9424
```



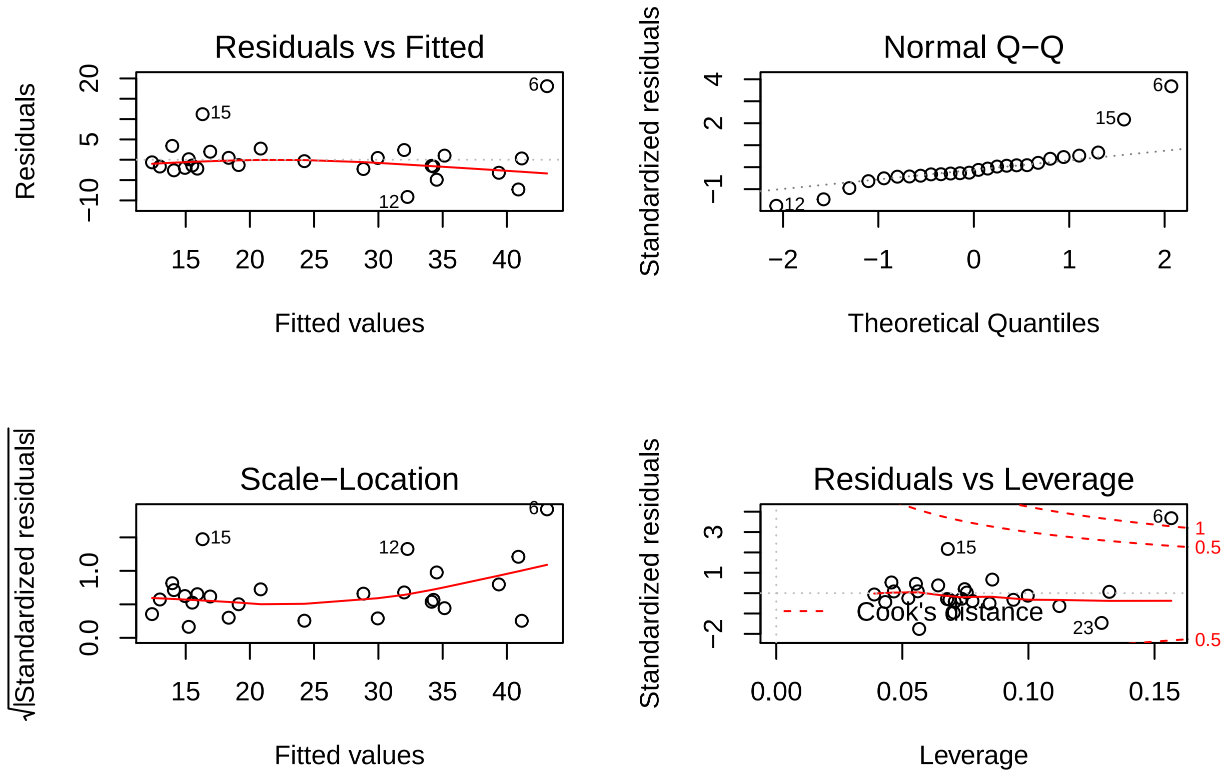
Here we could have the same comments as before, although the residuals are not good at all. The regression model may not fit this particular cluster, this could be due to some very influential and with high leverage points as we see in the residuals plots.

Model 4:

Here a specific problem with models per cluster comes up. All the variables that only have one level are unusable, we can't estimate the effect on the cost of a change in that variable if it doesn't change at all in our data. So we need to remove them, and this constitutes the weak point of this method in general. Although I think it works well since a lot of our categorical variables divide into multiple levels and we're left with tons of variables on which our regression is supposed to fit.

```
## Start:  AIC=128.65
## `EXW cost` ~ 1
##
##                                      Df Sum of Sq    RSS     AIC
## + `Net Weight (kg)`                   1   2704.02  687.9  89.164
## + `Surface envelop (LG x lg) (mm2)`   1   1749.53 1642.4 111.791
## + `Raw material`                      1    620.15 2771.8 125.398
## + Supplier                           13   2256.61 1135.3 126.191
## + `nb Machining Surfaces`             1    464.65 2927.3 126.817
## + `nb Threading`                      1    260.55 3131.4 128.569
## <none>                                            3391.9 128.647
## + `nb Cavities`                       1    199.08 3192.8 129.075
## + `Yearly Volume`                     1    136.34 3255.6 129.581
## + `Over molding`                      1     12.40 3379.5 130.552
## + Assembly                            1      6.51 3385.4 130.597
## + `nb Cores`                          1      3.13 3388.8 130.623
```

```
## + Process                           1      2.91 3389.0 130.625
## + Finishing                         2     62.41 3329.5 132.165
##
## Step:  AIC=89.16
## `EXW cost` ~ `Net Weight (kg)`
##
##                                    Df Sum of Sq    RSS     AIC
## <none>                                           687.9  89.164
## + `Raw material`                    1     49.50  638.4  89.222
## + `nb Cores`                        1     35.52  652.4  89.785
## + `Surface envelop (LG x lg) (mm2)`  1     16.01  671.9  90.552
## + Process                           1      7.88  680.0  90.864
## + `Over molding`                    1      6.88  681.0  90.903
## + Assembly                          1      3.62  684.3  91.027
## + `nb Cavities`                     1      2.89  685.0  91.054
## + `Yearly Volume`                   1      2.83  685.1  91.057
## + `nb Threading`                    1      2.52  685.4  91.068
## + `nb Machining Surfaces`           1      0.97  686.9  91.127
## + Finishing                         2     22.52  665.4  92.298
## + Supplier                         13    367.75  320.1  95.277
## - `Net Weight (kg)`                 1   2704.02 3391.9 128.647

##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)`, data = don_cluster_4)
##
## Coefficients:
##       (Intercept)  `Net Weight (kg)`
##             5.825              6.556

##
## Call:
## lm(formula = `EXW cost` ~ `Net Weight (kg)`, data = don_cluster_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1497 -2.1431 -0.9738  0.8754 18.0965
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.8255     2.2588   2.579   0.0165 *
## `Net Weight (kg)`   6.5557     0.6749   9.713 8.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.354 on 24 degrees of freedom
## Multiple R-squared:  0.7972, Adjusted R-squared:  0.7887
## F-statistic: 94.34 on 1 and 24 DF,  p-value: 8.655e-10
```



The only appropriate model seems to be the one corresponding to cluster 2, which is the second most filled (in bn of observations) cluster when looking both at the residuals and the overall and individual significance of the variables. If I were to propose a model in any case, I think the above one is one way to approach the problem. Althoug we can't say for sure since we don't have much data (only 211 observation for more than 15 variables) which makes this analysis and the construction of a prediction model very complicated.

8) If someone asked you why you did one global model and not one model per supplier, what would be your answer?

That would have been omitting a valuable predictor for each. Suppliers are competing on the international scene, or even local, discarding other suppliers on the market as a variable would be ignoring the forces that drive quantities and prices on markets worldwide. The objective of estimating the "Should cost" of a product hasn't changed, only we wouldn't be taking into account the influence that the market has on the overall price if we were constructing separate models for each supplier. This is actually a regression problem, omitted-variable-bias, where the residuals are correlated with the outcome varaible. Generally to counter this problem an instrumental variables method can be implemented (or equivalent method 2stage least squares) when we have no idea of what this omitted variable is.

9) These data contained missing values. One representative in the compagny suggests either to put 0 in the missing cells or to impute with the median of the variables. Comment. For the categorical variables with missing values, it is decided to create a new category ???missing???. Comment.

Replacing NA's by zeros or the median of a given varaible can very rarely be a good idea. Althoug using the median is already better than replacing by zeros, we are still far from any optimal way of hanling missing data. The problem in imputing data by this same value is that it will drastically affect the covariance and correlation that exists within our data. By replacing them by this arbitrary value, we are also completely ignoring the potential reason why they were missing to begin with, which is a valuable information in itself. This falls within the underlying properties of missing values, they can either be MCAR, MAR or MNAR. Once we've determined the reason why they are missing, we can then assess what strategy we may want to implement to handle the NA's. In general, it is assumed that the data are at least MAR, in which case there are several more optimal ways to handle this problem. Iterative PCA is one example (classic/regularized/soft dependening on level of noise in the data) if we wish to do point estimates, we could also consider multiple imputation methods like joint/conditional modelling and bootstrap PCA. As for creating a new variable Missing which takes value zero for observations containing na's and ones otherwise, in addition to taking the risk of being left with very few data, we are also taking the risk of ignoring a potential subsample representative of a whole portion of the population studied. Not only would it be then impossible to extend our analysis/prediction on that ignored subpopulation, but it will also bias our analysis for the data we do consider.