GRAU D'ENGINYERIA INFORMÀTICA (UPC). CURS 20-21 Q1 –QUIZ 1
Anàlisi de Dades i Explotació de la Informació (ADEI) .
(Data: 10/11/2020 14:00-15:30 h A5-202 Room)

Professor:	Lídia Montero Mercadé
Rules for quiz:	Internet access is required, emailing and chatting is strictly forbidden. Mobile phones should be switched off.
Duration:	1h 30 min
Marks:	Before 17/11/20 Subject ATENEA WEB site.
Open Office:	Email requests.

Problem 1: All questions account for 1 point

The data set for this exercise contains 396 observations for the mean daily values of the variables included in the table referred to a wastewater treatment plant. The plant has measures on the quality of the wastewater at the entrance of the plant, they are the variables xxx.e, from here they go to a first decantation process (Primary Treatment) where it is intended that they settle the solids in suspensions. Then it goes into the Biological Reactor. This is the most critical part. Here is a biological mud that literally "lives by eating organic matter." It is activated by temperature and aeration. This process is carried out by many species of microorganisms. These microorganisms work at different temperatures. If they are not balanced with the composition of the water, they eat each other. The control variables are those that graduate aeration, temperature, recirculation and purge of the bioreactor. This is the most difficult part: if things go well, dirty water and biological mud enter the bioreactor and end up with clean water and colonies of microorganisms that have consumed the organic matter. Then the water goes through a second decanting process where the microorganisms settle because if the water is not aerated they fall to the ground. And then clean can be poured into the river. The purified water, at the end of the process, before being poured into the river must have neither DBO, nor DQO, nor SS nor SSV, absolute zeros are impossible and therefore the current legislation has permissible limits that are not dangerous for the life in rives.

Mean Daily Observations	Input	Decantation	Biological	Output
Flow (Q)	**q.e**		***qb.b***	
Iron pretreatment (FE)	**fe.e**			
Hydrogen potential (pH)	**ph.e**	**ph.d**		**ph.s**
Solid in Suspension (SS)	**ss.e**	**ss.d**		**ss.s**
Suspended Volatile Solids (SSV)	**ssv.e**	**ssv.d**		**ssv.s**
Fraction of degradable organic matter (DQO)	**dqo.e**	**dqo.d**		**dqo.s**
BIOdegradable organic matter fraction (DBO)	**dbo.e**	**dbo.d**		**dbo.s**
Volumetric Analysis (V30.B)			**v30.b**	
Recirculation Flow (QR.G)			**qr.g**	
Purge Flow (QP.G)			**qp.g**	
Air inflow (QA.G)			**qa.g**	
Mixed Liquor Suspended Solids (MLSS.B)			**mlss.b**	
Volatile solids in suspension liquor mixture (MLVSS.B)			**mlvss.b**	
Cell Age (MCRT.B)			**mcrt.b**	

The data technically correspond to daily measurements and there is a temporal correlation that cannot be dealt with in this subject. **You only have to work in this exercise with the data in randomized order.** The response variables are considered the fraction of biodegradable organic matter DBO.S, degradable organic matter DQO.S or solids in suspension, either volatile (SSV.S) or not (SS.S) in the OUTPUT of the plant. **The response variable DQO.S is initially considered.**
The list of the included variables in the dataset contains some additional columns that will not be considered in the exercise.

date	id from 1 to number of observations
dateformated	dd-mm-yy
datenorm	dd/mm/yyyy
q.e	Input Flow
qb.b	Flow after biological reactor
qr.g	Recirculation Flow
qp.g	Purge Flow
qa.g	Air inflow
fe.e	Iron pretreatment
ph.e	Hydrogen potential
ss.e	Input Solid in Suspension
ssv.e	Input Suspended Volatile Solids
dqo.e	Input Fraction of degradable organic matter
dbo.e	Input BIOdegradable organic matter fraction
nkt.e	*Input Hydrogen potential*
nh4.e	*Input Ammonium concentration*
p.e	*Input Phosphor concentration*
ph.d	Decantation Hydrogen potential at the settler
ss.d	Decantation Solid in Suspension at the settler
ssv.d	Decantation Suspended Volatile Solids at the settler
dqo.d	Decantation Fraction of degradable organic matter at the settler
dbo.d	Decantation BIOdegradable organic matter fraction at the settler
nkt.d	*Decantation Hydrogen potential at the settler*
nh4.d	*Decantation Ammonium concentration at the settler*
p.d	*Decantation Phosphor concentration at the settler*
ph.s	Output Hydrogen potential
ss.s	Output Solid in Suspension
ssv.s	Output Suspended Volatile Solids
dqo.s	Output Fraction of degradable organic matter
dbo.s	Output BIOdegradable organic matter fraction
nk.s	*Unknown*
nh4.s	*Output Ammonium concentration*
p.s	*Output Phosphor concentration*
v30.b	Biological Volumetric Analysis
mlss.b	Biological Mixed Liquor Suspended Solids
mlvss.b	Biological Volatile solids in suspension liquor mixture
im.b	*Unknown*
cm1.b	*Unknown*
cm2.b	*Unknown*
mcrt.b	Biological Cell Age
trh.c	*Unknown (non important)*
dbo.dqoe	Input Quocient DBO.E into DQO.E
dbo.dqod	Quocient DBO.D into DQO.D at the settler
dbo.dqos	Output Quocient DBO.S into DQO.S
weekday	Day of the week
season	Year season

1. Produce a randomized dataset to destroy serial correlation.

You have to define a list containing of 396 row names without replacement. This can be done using:

```{r}
set.seed(12345)
llrandom<-sample(1:nrow(df),nrow(df))
df<-df[ llrandom, ]
```

2. Missing data have been treated, but some NA coded as 0 values still remain in fe.e and qp.g and have to be removed by applying imputation tools explained in class.

There are some low value outliers in both variables, in fact 17 observations for fe.e and 6 for qp.g also in fe.e list. These observations have to be set as NA and imputePCA() process in missMDA library. Imputation returns reasonable figures, so we accept them.

```{r}
##### There are outliers fe.e qp.g
calcQ(df$fe.e)
calcQ(df$qp.g)
```

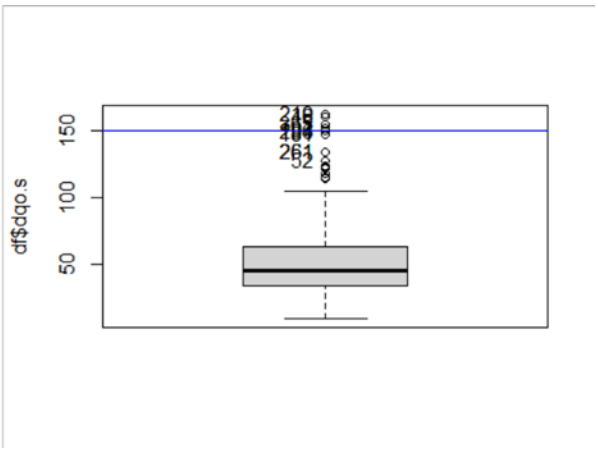
Name:
DNI/Passport:

```
ll<-which(df$fe.e<0.01);length(ll)
llq<-which(df$qp.g<0.01);length(llq)
ll<-which(df$fe.e<4.1);length(ll)
llq<-which(df$qp.g<42.4);length(llq)
ll<-unique(ll,llq);length(ll)
# Solution remove the ll observations or impute them
df$fe.e[ll]<-NA
df$qp.g[llq]<-NA

library(missMDA)
names(df)
res.imp<-imputePCA(df[,4:44],ncp=10)
# Validation is needed
summary(res.imp$completeObs[,6])
summary(res.imp$completeObs[,4])
df$fe.e<-res.imp$completeObs[,6]
df$qp.g<-res.imp$completeObs[,4]
```

3. Univariant outliers for output variable **DQO.S** are also present and have to be treated. Do it.

Mild outliers lie outside -9.5 to 107 units and severe outlier lie out -53 to 150 units for dqo:s. Severe upper outliers are present over 150. There are 4 observations over 150 and since dqo:s is the target variable then observations with univariant upper severe outliers ***are removed from the sample.*** *Other answers are admitted provided outlier thresholds are justified.*

```
library(car)
Boxplot(df$dqo.s)
> sumaux<-summary(df$dqo.s);sumaux
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   34.00   45.50   51.26   63.00  163.00
> mouti<-sumaux[2]-1.5*(sumaux[5]-sumaux[2]);mouti
1st Qu.
   -9.5
> souti<-sumaux[2]-3*(sumaux[5]-sumaux[2]);souti
1st Qu.
    -53
> mouts<-sumaux[5]+1.5*(sumaux[5]-sumaux[2]);mouts
3rd Qu.
  106.5
> souts<-sumaux[5]+3*(sumaux[5]-sumaux[2]);souts
3rd Qu.
    150
> abline(h=calcQ(df$dqo.s)$souts,col="blue")
> ll<-which(df$dqo.s>calcQ(df$dqo.s)$souts);length(ll)
[1] 4
> df<-df[-ll,]
```



4. Are there multivariant outliers? Find them. Try to explain their singularity. Multivariant outliers are not going to be treated in this exercise: keep them as they are.

Define a vector indicating multivariant outlier indicator. This can be done using chemometrics library and Moutlier method. Robust Mahalanobis distance over 250 can be considered a multivariant outlier. Technically robust Mah. Distance over 32.22 are extreme

Name:
DNI/Passport:

outliers (38 observations are over this threshold). Any argumented threshold has been considered correct. This vector can be added as an additional variable in data set and profiling it using catdes() method in FactoMineR library provides a fast charaterization of multivariant outliers. Multivariant outliers are strongly related to quantitative vàlues in mcrt.b, qr.g, qb.b, mlss.b and q.e according to component quanti.var of catdes() output. Taking a look at quanti component for Yes-mout cluster: mcrt.b and mlss.b are significantly over their sample mean, while qr.g, qb.b and q.e are significantly under their mean.

```
> names(df)
 [1] "date"         "dateformated" "datenorm"     "q.e"          "qb.b"         "qr.g"
 [7] "qp.g"         "qa.g"         "fe.e"         "ph.e"         "ss.e"         "ssv.e"
[13] "dqo.e"        "dbo.e"        "nkt.e"        "nh4.e"        "p.e"          "ph.d"
[19] "ss.d"         "ssv.d"        "dqo.d"        "dbo.d"        "nkt.d"        "nh4.d"
[25] "p.d"          "ph.s"         "ss.s"         "ssv.s"        "dqo.s"        "dbo.s"
[31] "nk.s"         "nh4.s"        "p.s"          "v30.b"        "mlss.b"       "mlvss.b"
[37] "im.b"         "cm1.b"        "cm2.b"        "mcrt.b"       "trh.c"        "dbo.dqoe"
[43] "dbo.dqod"     "dbo.dqos"     "weekday"      "season"       "mout"
> vars_con<-names(df)[c(4:44)]
>
> library(chemometrics)
> #summary(df[,vars_con])
> mout<-Moutlier(df[,vars_con],quantile = 0.995, plot = TRUE)
>
> library(car)
> ll<-which(mout$rd>calcQ(mout$rd)$souts);length(ll);calcQ(mout$rd)$souts
[1] 38
 3rd Qu.
32.21727
> ll<-which(mout$rd>250);length(ll)
[1] 3
> par(mfrow=c(1,1))
> Boxplot(mout$rd, col="cyan")
 [1] 294  39 255 131 356 374 103  73 155 105
> df[ll,vars_con]
        q.e    qb.b    qr.g     qp.g   qa.g     fe.e ph.e ss.e ssv.e dqo.e    dbo.e nkt.e nh4.e  p.e ph.d
155 22883.1 22046.0 18343.5 213.0287  96451 21.65231  7.3   93    37   157 81.33333 21.26 10.32 2.56  7.3
153 22752.0 21964.4 18091.9 233.1185 119971  6.50000  7.5   91    19    64 89.66667 24.62 12.54 3.02  7.7
157 23662.9 22891.0 18150.3 289.7194 143151 25.30025  7.6   77    51   180 73.00000 17.90  8.10 2.10  7.7
    ss.d ssv.d dqo.d    dbo.d    nkt.d nh4.d  p.d ph.s ss.s ssv.s dqo.s    dbo.s nk.s nh4.s p.s v30.b
155   67    42    92 51.33333 18.80769  10.3 1.82  7.3  6.4   2.8     9 5.666667 3.16  1.54 1.0   320
153   46    13    36 48.66667 22.51538  11.7 1.74  7.2  6.4   3.6    18 6.333333 4.32  2.58 0.9   383
157   40    30   152 54.00000 15.10000   8.9 1.90  7.6  5.2   4.0    66 5.000000 2.00  0.50 1.1   310
    mlss.b mlvss.b  im.b cm1.b      cm2.b mcrt.b trh.c dbo.dqoe  dbo.dqod dbo.dqos
155   3134    2005 102.1  0.07 0.06333333 307.97  8.69     0.60 0.3933333     0.13
153   2908    1959 131.7  0.02 0.07666667 286.82  8.72     0.79 0.4266667     0.18
157   2936    1878 105.6  0.08 0.05000000 341.99  8.37     0.41 0.3600000     0.08
> #summary(df[,vars_con])
> library(FactoMineR)
> df$mout<-0
> df$mout[ll]<-1
> df$mout<-factor(df$mout, labels=c("No-MOut","Yes-Mout"))
> res.cat<-catdes(df[,c("mout",vars_con)],1)
> res.cat$quanti
$`No-MOut`
              v.test Mean in category Overall mean sd in category   Overall sd      p.value
qr.g       9.873592     4.112148e+04 4.094602e+04   3.480060e+03 4.001248e+03 5.418855e-23
qb.b       7.040746     3.903551e+04 3.890744e+04   3.841983e+03 4.095868e+03 1.912132e-12
q.e        6.346820     4.196465e+04 4.182027e+04   4.869598e+03 5.122072e+03 2.198112e-10
nh4.d      5.862823     2.606816e+01 2.594749e+01   4.442100e+00 4.634597e+00 4.550643e-09
fe.e       4.847931     4.751478e+01 4.728751e+01   1.024828e+01 1.055596e+01 1.247557e-06
cm1.b      4.769390     6.237532e-01 6.194133e-01   1.995943e-01 2.048916e-01 1.847843e-06
dqo.d      4.530012     2.499974e+02 2.487985e+02   5.808430e+01 5.959473e+01 5.898036e-06
qp.g       4.293505     6.282591e+02 6.253282e+02   1.505894e+02 1.537060e+02 1.758747e-05
ssv.d      4.182928     6.516710e+01 6.488520e+01   1.485109e+01 1.517412e+01 2.877786e-05
qa.g       3.917360     2.330442e+05 2.321780e+05   4.896201e+04 4.978964e+04 8.952398e-05
nkt.e      3.881408     4.211546e+01 4.195585e+01   9.110714e+00 9.259087e+00 1.038534e-04
nkt.d      3.849958     3.655075e+01 3.641496e+01   7.815117e+00 7.941639e+00 1.181380e-04
ssv.e      3.333792     1.582416e+02 1.573036e+02   6.268077e+01 6.335794e+01 8.567069e-04
ss.d       3.310616     8.897301e+01 8.868240e+01   1.953489e+01 1.976532e+01 9.309071e-04
dqo.e      3.237344     4.432082e+02 4.408393e+02   1.631090e+02 1.647661e+02 1.206479e-03
dbo.d      3.130410     1.204126e+02 1.198839e+02   3.769079e+01 3.802631e+01 1.745626e-03
dbo.dqos   2.659953     4.096530e-01 4.075128e-01   1.801773e-01 1.811687e-01 7.815149e-03
ss.e       2.508448     2.103856e+02 2.094413e+02   8.439767e+01 8.476119e+01 1.212627e-02
nk.s       2.446254     1.931953e+01 1.919586e+01   1.133894e+01 1.138322e+01 1.443495e-02
dbo.e      2.377406     2.142326e+02 2.132156e+02   9.599642e+01 9.632900e+01 1.743487e-02
nh4.s      2.366534     1.223261e+01 1.215078e+01   7.759331e+00 7.785901e+00 1.795552e-02
p.e        2.254027     9.967899e+00 9.911206e+00   5.647996e+00 5.663353e+00 2.419444e-02
```

Name:
DNI/Passport:

```
dbo.s      2.222928     1.867326e+01 1.857372e+01   1.005721e+01 1.008268e+01 2.622066e-02
p.d        2.045770     6.113397e+00 6.080539e+00   3.610878e+00 3.616446e+00 4.077898e-02
mlvss.b   -3.971322     1.341117e+03 1.345756e+03   2.586361e+02 2.630451e+02 7.147481e-05
trh.c     -5.548017     4.160900e+00 4.194821e+00   1.326422e+00 1.376709e+00 2.889278e-08
mlss.b    -5.914806     1.759126e+03 1.768566e+03   3.441293e+02 3.593773e+02 3.322661e-09
mcrt.b   -16.476621     1.206233e+01 1.435976e+01   1.731057e+01 3.139619e+01 5.402091e-61
$`Yes-Mout`
            v.test Mean in category Overall mean sd in category   Overall sd      p.value
mcrt.b   16.476621     3.122600e+02 1.435976e+01   2.272642e+01 3.139619e+01 5.402091e-61
mlss.b    5.914806     2.992667e+03 1.768566e+03   1.005894e+02 3.593773e+02 3.322661e-09
trh.c     5.548017     8.593333e+00 4.194821e+00   1.583947e-01 1.376709e+00 2.889278e-08
mlvss.b   3.971322     1.947333e+03 1.345756e+03   5.249974e+01 2.630451e+02 7.147481e-05
p.d      -2.045770     1.820000e+00 6.080539e+00   6.531973e-02 3.616446e+00 4.077898e-02
dbo.s    -2.222928     5.666667e+00 1.857372e+01   5.443311e-01 1.008268e+01 2.622066e-02
p.e      -2.254027     2.560000e+00 9.911206e+00   3.755884e-01 5.663353e+00 2.419444e-02
nh4.s    -2.366534     1.540000e+00 1.215078e+01   8.491564e-01 7.785901e+00 1.795552e-02
dbo.e    -2.377406     8.133333e+01 2.132156e+02   6.804138e+00 9.632900e+01 1.743487e-02
nk.s     -2.446254     3.160000e+00 1.919586e+01   9.471360e-01 1.138322e+01 1.443495e-02
ss.e     -2.508448     8.700000e+01 2.094413e+02   7.118052e+00 8.476119e+01 1.212627e-02
dbo.dqos -2.659953     1.300000e-01 4.075128e-01   4.082483e-02 1.811687e-01 7.815149e-03
dbo.d    -3.130410     5.133333e+01 1.198839e+02   2.177324e+00 3.802631e+01 1.745626e-03
dqo.e    -3.237344     1.336667e+02 4.408393e+02   5.014867e+01 1.647661e+02 1.206479e-03
ss.d     -3.310616     5.100000e+01 8.868240e+01   1.157584e+01 1.976532e+01 9.309071e-04
ssv.e    -3.333792     3.566667e+01 1.573036e+02   1.309792e+01 6.335794e+01 8.567069e-04
nkt.d    -3.849958     1.880769e+01 3.641496e+01   3.027318e+00 7.941639e+00 1.181380e-04
nkt.e    -3.881408     2.126000e+01 4.195585e+01   2.743429e+00 9.259087e+00 1.038534e-04
qa.g     -3.917360     1.198577e+05 2.321780e+05   1.906536e+04 4.978964e+04 8.952398e-05
ssv.d    -4.182928     2.833333e+01 6.488520e+01   1.189771e+01 1.517412e+01 2.877786e-05
qp.g     -4.293505     2.452889e+02 6.253282e+02   3.247006e+01 1.537060e+02 1.758747e-05
dqo.d    -4.530012     9.333333e+01 2.487985e+02   4.736619e+01 5.959473e+01 5.898036e-06
cm1.b    -4.769390     5.666667e-02 6.194133e-01   2.624669e-02 2.048916e-01 1.847843e-06
fe.e     -4.847931     1.781752e+01 4.728751e+01   8.140089e+00 1.055596e+01 1.247557e-06
nh4.d    -5.862823     1.030000e+01 2.594749e+01   1.143095e+00 4.634597e+00 4.550643e-09
q.e      -6.346820     2.309933e+04 4.182027e+04   4.020799e+02 5.122072e+03 2.198112e-10
qb.b     -7.040746     2.230047e+04 3.890744e+04   4.188968e+02 4.095868e+03 1.912132e-12
qr.g     -9.873592     1.819523e+04 4.094602e+04   1.075171e+02 4.001248e+03 5.418855e-23

> res.cat$quanti.var
               Eta2       P-value
mcrt.b   0.69431980 2.051110e-102
qr.g     0.24932947  4.140906e-26
qb.b     0.12678288  3.682252e-13
q.e      0.10302333  7.635219e-11
mlss.b   0.08947554  1.518859e-09
nh4.d    0.08790969  2.141256e-09
trh.c    0.07872249  1.592580e-08
fe.e     0.06010853  8.931888e-07
cm1.b    0.05817669  1.353346e-06
dqo.d    0.05248340  4.596665e-06
qp.g     0.04714624  1.443653e-05
ssv.d    0.04474907  2.413053e-05
mlvss.b  0.04033606  6.212894e-05
qa.g     0.03924734  7.846221e-05
nkt.e    0.03853025  9.150412e-05
nkt.d    0.03790838  1.045597e-04
ssv.e    0.02842499  8.039997e-04
ss.d     0.02803115  8.754084e-04
dqo.e    0.02680408  1.141508e-03
dbo.d    0.02506257  1.665122e-03
dbo.dqos 0.01809553  7.653640e-03
ss.e     0.01609287  1.194392e-02
nk.s     0.01530475  1.424679e-02
dbo.e    0.01445540  1.724288e-02
nh4.s    0.01432349  1.776321e-02
p.e      0.01299396  2.400419e-02
dbo.s    0.01263787  2.603290e-02
p.d      0.01070377  4.062470e-02
```

5. Indicate by using exploratory data analysis tools which are apparently the most associated variables with the response variable (use only the indicated variables). Use also **FactoMineR profiling tools**.

Normality assumption seems to fail for dqo:s target, then Spearman correlation has to be calculated. Inversely related to the target are mcrt:b and mlss:b variables and directly correlated are dqo:d ss:s ssv:s dbo:s

Name:
DNI/Passport:

```
> names(df)
 [1] "date"         "dateformated" "datenorm"     "q.e"          "qb.b"         "qr.g"
 [7] "qp.g"         "qa.g"         "fe.e"         "ph.e"         "ss.e"         "ssv.e"
[13] "dqo.e"        "dbo.e"        "nkt.e"        "nh4.e"        "p.e"          "ph.d"
[19] "ss.d"         "ssv.d"        "dqo.d"        "dbo.d"        "nkt.d"        "nh4.d"
[25] "p.d"          "ph.s"         "ss.s"         "ssv.s"        "dqo.s"        "dbo.s"
[31] "nk.s"         "nh4.s"        "p.s"          "v30.b"        "mlss.b"       "mlvss.b"
[37] "im.b"         "cm1.b"        "cm2.b"        "mcrt.b"       "trh.c"        "dbo.dqoe"
[43] "dbo.dqod"     "dbo.dqos"     "weekday"      "season"       "mout"
> hist(df$dqo.s)
> shapiro.test(df$dqo.s)

     Shapiro-Wilk normality test

data:  df$dqo.s
W = 0.92222, p-value = 2.228e-13

> vars_input<-names(df)[c(4:28,30:37,40)]
> #summary(df[,vars_input])
> tt<-round(cor(df[,c("dqo.s",vars_input)],method="spearman"),digits=3)
> sort(tt[1,])
 mcrt.b  mlss.b mlvss.b     q.e     p.s    fe.e    qa.g    qb.b    ph.e     p.e     p.d
nh4.e   nh4.d
 -0.212  -0.138  -0.129  -0.027   0.003   0.004   0.010   0.024   0.030   0.041   0.043
0.054   0.061
   ss.d   nkt.e    ph.d   ssv.d   v30.b   nh4.s    im.b    qr.g   dbo.d   nkt.d    qp.g
ph.s    nk.s
  0.063   0.094   0.095   0.095   0.102   0.122   0.131   0.138   0.145   0.146   0.167
0.179   0.184
  dbo.e    ss.e   dqo.e   ssv.e   dqo.d    ss.s   ssv.s   dbo.s   dqo.s
  0.199   0.200   0.205   0.230   0.291   0.293   0.324   0.359   1.000
```

Factor season is globally associated to dqo.s target, showing spring category 4.77 points over the grand mean and winter -7.23 units under the grand mean. Globally correlated variables with dqo.s target are showing a direct effect ssv.s, ss.v, dbo.s (over 0.4) and inversely mlss.b and mlvss.b (showing less intensity).

```
> library(FactoMineR)
> res.con<-condes(df[,c("dqo.s",vars_input,"weekday","season")],1)
> res.con$quanti
        correlation      p.value
ssv.s     0.4458880 1.514703e-20
ss.s      0.4335064 2.164847e-19
dbo.s     0.4329508 2.433070e-19
dqo.d     0.2769874 2.460702e-08
nk.s      0.2195709 1.148978e-05
im.b      0.2063535 3.839403e-05
ph.s      0.1967719 8.787050e-05
ssv.e     0.1772296 4.223480e-04
nh4.s     0.1747234 5.107259e-04
nkt.d     0.1724304 6.063426e-04
dqo.e     0.1694302 7.565586e-04
qr.g      0.1665652 9.314721e-04
qp.g      0.1541520 2.208787e-03
ss.e      0.1490030 3.104338e-03
dbo.e     0.1199878 1.747104e-02
dbo.d     0.1198802 1.757293e-02
nkt.e     0.1184365 1.899154e-02
ssv.d     0.1155987 2.207306e-02
ph.d      0.1151010 2.265601e-02
v30.b     0.1131964 2.501151e-02
mcrt.b   -0.1431321 4.519074e-03
mlvss.b  -0.1647856 1.058180e-03
mlss.b   -0.1903638 1.496171e-04
> res.con$quali
                R2     p.value
season 0.03123194 0.006342759
> res.con$category
               Estimate     p.value
season=Spring  4.766705 0.040804611
season=Winter -7.235385 0.001210176
```

Name:
DNI/Passport:

6. Define polytomous factors **f.dbo.s, f.dqo.s, f.sst.s** (from SSV.S plus SS.S) for the covariates according to the legal limit (DBO 25 mg/l O2, DQO 125 mg/l O2 and total suspended solids 35 mg/l). **Profile f.dqo.s** factor.

This can be easily done using cut() method: 17.6% of days show dbo.s over legal limits, 1.5% for dqo.s and 18.11% for sst limits.

```
>df$sst.s<-df$ss.s+df$ssv.s
> df$f.dbo.s<-factor(cut(df$dbo.s,breaks=c(0,25,max(df$dbo.s)),include.lowest = T))
> df$f.dqo.s<-factor(cut(df$dqo.s,breaks=c(0,125,max(df$dqo.s)),include.lowest = T))
> df$f.sst.s<-factor(cut(df$sst.s,breaks=c(0,35,max(df$sst.s)),include.lowest = T))
> prop.table(table(df$f.dbo.s,useNA="ifany"))

   [0,25]   (25,84]
0.8239796 0.1760204
> prop.table(table(df$f.dqo.s,useNA="ifany"))

   [0,125]  (125,150]
0.98469388 0.01530612
> prop.table(table(df$f.sst.s,useNA="ifany"))

   [0,35]  (35,233]
0.8188776 0.1811224
```

Globally related to f.dqo.s factor are f.sst.s and f.dbo.s factors. Out of limits ((125,150]) cluster for f.dqo.s factor shows overrepresentation of f.sst.s=(35,233] and f.dbo.s=(25,84] out of limits. Variables ssv.s, ss.s and dbo.s are globally associated to f.dqo.s factor. ssv.s, ssv, dqo.s, dbo.s and im.b means in the (125,150] level of f.dqo.s are above sample means and mlss.b variables lies under its sample mean.

```
> res.cat<-catdes(df[,c("f.dqo.s","f.sst.s","f.dbo.s","dqo.s",vars_input,"weekday","season
")],1)
> res.cat$test.chi2
             p.value df
f.sst.s 2.910019e-05  1
f.dbo.s 1.471759e-03  1
> res.cat$category
$`[0,125]`
                 Cla/Mod  Mod/Cla   Global     p.value    v.test
f.sst.s=[0,35]   99.68847 82.90155 81.88776 0.000920905  3.313638
f.dbo.s=[0,25]   99.38080 83.16062 82.39796 0.010817127  2.548552
f.dbo.s=(25,84]  94.20290 16.83938 17.60204 0.010817127 -2.548552
f.sst.s=(35,233] 92.95775 17.09845 18.11224 0.000920905 -3.313638

$`(125,150]`
                  Cla/Mod  Mod/Cla   Global     p.value    v.test
f.sst.s=(35,233] 7.0422535 83.33333 18.11224 0.000920905  3.313638
f.dbo.s=(25,84]  5.7971014 66.66667 17.60204 0.010817127  2.548552
f.dbo.s=[0,25]   0.6191950 33.33333 82.39796 0.010817127 -2.548552
f.sst.s=[0,35]   0.3115265 16.66667 81.88776 0.000920905 -3.313638

> res.cat$quanti.var
              Eta2      P-value
ssv.s   0.23692737 1.036739e-24
ss.s    0.23481015 1.787418e-24
dqo.s   0.20608643 2.516028e-21
dbo.s   0.09885014 1.925244e-10
im.b    0.02515848 1.630810e-03
mlss.b  0.01128669 3.549566e-02
> res.cat$quanti
$`[0,125]`
          v.test Mean in category Overall mean sd in category Overall sd      p.value
mlss.b  2.100737       1773.32642   1768.56633     358.027347  359.37734 3.566408e-02
im.b   -3.136394        153.77617    155.55816      88.441487   90.11206 1.710394e-03
dbo.s  -6.216945         18.17850     18.57372       8.996493   10.08268 5.069271e-10
dqo.s  -8.976625         48.76399     50.17066      22.288601   24.85345 2.792000e-19
ss.s   -9.581794         15.09326     15.92449      10.986478   13.75873 9.537401e-22
ssv.s  -9.624895         11.25440     11.90561       8.578293   10.73075 6.277106e-22

$`(125,150]`
          v.test Mean in category Overall mean sd in category Overall sd      p.value
ssv.s   9.624895          53.8000     11.90561      31.726120   10.73075 6.277106e-22
```

Name:
DNI/Passport:

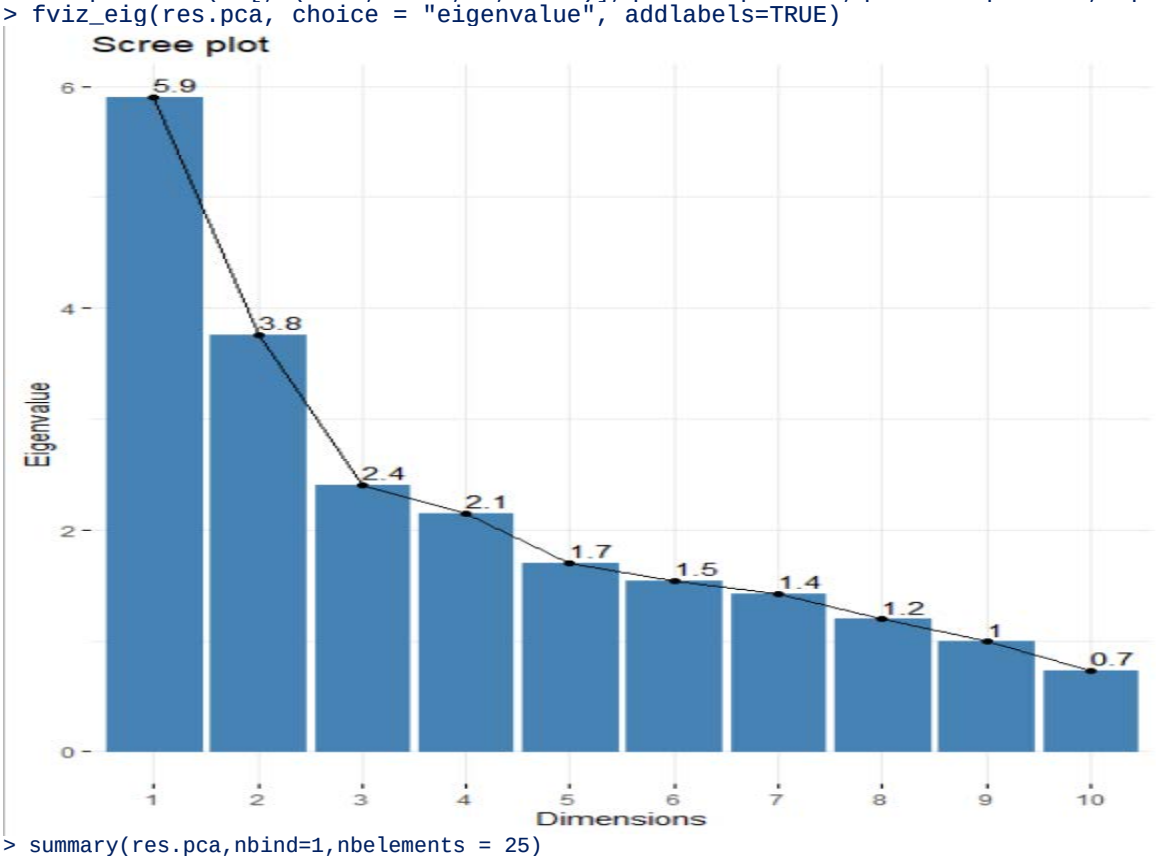
```
ss.s     9.581794        69.4000     15.92449      41.212781   13.75873 9.537401e-22
dqo.s    8.976625       140.6667     50.17066       8.919392   24.85345 2.792000e-19
dbo.s    6.216945        44.0000     18.57372      27.898626   10.08268 5.069271e-10
im.b     3.136394       270.2000    155.55816     118.164687   90.11206 1.710394e-03
mlss.b -2.100737      1462.3333   1768.56633     310.196533  359.37734 3.566408e-02
```

7. A Normalized Principal Component Analysis is addressed using as supplementary variables available factor and xxxx.s output variables. How many axes do you have to retain according to Kaiser criteria? And according to Elbow's rule? What's the inertia explained by retained Kaiser-based principal components? Try to explain the meaning of the axes in the first factorial plane. Which are the 3 variables with the greatest correlation with the first factorial plane?

Kaiser rule for Normalized Principal Components states to take as many dimensions as eigenvalues are greater than 1 in this cases then 8 dimensions that account for 77% of the total inertia, it is enough. Elbow's rule a screeplot is needed: it is difficult to say, but from dimension 5 there is no remarkable descent. First factorial axes seems to be correlated to entrance variables (ss.e, ssv.e, dbo.e, ss.e, q.e and the same variables referred to decantation). Second axis seems to be related to ph, amonium and phosphate variables: these variables are not altered by water treatment at the plant. Cos2 Axis 1 + Cos2 Axis 2 shows the correlation to the First Factorial plane::

- *p.e 0.128+0.525, p.d 0.128+0.492, qp.g 0.105+0.474 and Ssv.d 0.508*

```
> library(FactoMineR)
> names(df)[c(4:30,32:36,40,45:46)]
 [1] "q.e"     "qb.b"    "qr.g"    "qp.g"    "qa.g"    "fe.e"    "ph.e"    "ss.e"    "ssv.e"
[10] "dqo.e"   "dbo.e"   "nkt.e"   "nh4.e"   "p.e"     "ph.d"    "ss.d"    "ssv.d"   "dqo.d"
[19] "dbo.d"   "nkt.d"   "nh4.d"   "p.d"     "ph.s"    "ss.s"    "ssv.s"   "dqo.s"   "dbo.s"
[28] "nh4.s"   "p.s"     "v30.b"   "mlss.b"  "mlvss.b" "mcrt.b"  "weekday" "season"
> res.pca<-PCA(df[,c(4:30,32:36,40,45:46)],quali.sup=34:35,quanti.sup=23:29,ncp=8)
> fviz_eig(res.pca, choice = "eigenvalue", addlabels=TRUE)
```




```
> summary(res.pca,nbind=1,nbelements = 25)

Call:
PCA(X = df[, c(4:30, 32:36, 40, 45:46)], ncp = 8, quanti.sup = 23:29,
     quali.sup = 34:35)
```

Name:

DNI/Passport:

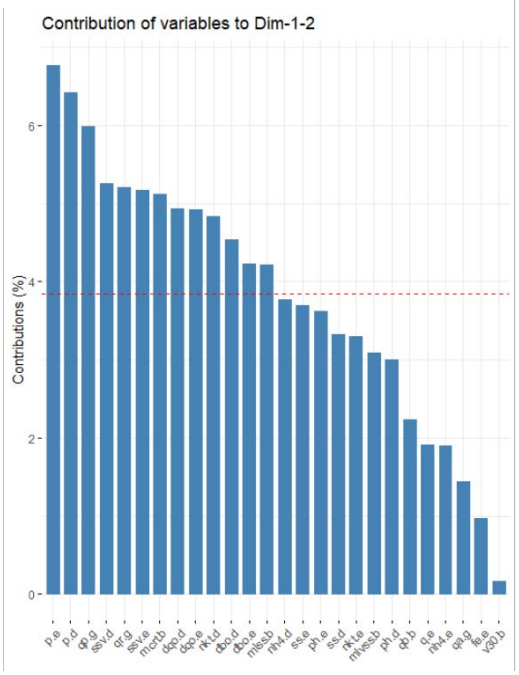
```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance               5.903   3.759   2.404   2.145   1.699   1.538   1.422   1.193   0.990
% of var.             22.704  14.457   9.245   8.250   6.534   5.917   5.471   4.588   3.808
Cumulative % of var.  22.704  37.161  46.406  54.656  61.190  67.107  72.578  77.166  80.975
                      Dim.10  Dim.11  Dim.12  Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
Variance               0.731   0.641   0.594   0.495   0.440   0.401   0.326   0.299   0.218
% of var.              2.810   2.466   2.283   1.903   1.691   1.543   1.255   1.150   0.840
Cumulative % of var.  83.785  86.251  88.535  90.438  92.129  93.671  94.926  96.077  96.916
                      Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24  Dim.25  Dim.26
Variance               0.197   0.174   0.133   0.106   0.077   0.048   0.042   0.025
% of var.              0.758   0.668   0.513   0.407   0.297   0.184   0.161   0.095
Cumulative % of var.  97.675  98.343  98.856  99.263  99.560  99.744  99.905 100.000

Individuals (the 1 first)
             Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
334       |  3.846 |  0.519  0.012  0.018 | -1.515  0.156  0.155 |  0.372  0.015  0.009 |

Variables (the 25 first)
             Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
q.e       |  0.426  3.074  0.181 |  0.058  0.090  0.003 |  0.172  1.226  0.029 |
qb.b      |  0.428  3.099  0.183 | -0.181  0.869  0.033 |  0.158  1.044  0.025 |
qr.g      |  0.702  8.357  0.493 | -0.101  0.273  0.010 | -0.400  6.650  0.160 |
qp.g      |  0.324  1.773  0.105 | -0.688 12.597  0.474 |  0.038  0.060  0.001 |
qa.g      |  0.327  1.810  0.107 |  0.180  0.857  0.032 |  0.240  2.388  0.057 |
fe.e      |  0.306  1.586  0.094 | -0.019  0.010  0.000 | -0.317  4.170  0.100 |
ph.e      | -0.068  0.078  0.005 | -0.588  9.187  0.345 |  0.225  2.111  0.051 |
ss.e      |  0.572  5.546  0.327 |  0.174  0.802  0.030 |  0.277  3.199  0.077 |
ssv.e     |  0.689  8.052  0.475 |  0.156  0.646  0.024 |  0.256  2.719  0.065 |
dqo.e     |  0.679  7.813  0.461 |  0.119  0.380  0.014 |  0.323  4.341  0.104 |
dbo.e     |  0.593  5.962  0.352 |  0.239  1.513  0.057 |  0.259  2.785  0.067 |
nkt.e     |  0.310  1.625  0.096 | -0.472  5.931  0.223 | -0.379  5.966  0.143 |
nh4.e     |  0.257  1.117  0.066 |  0.342  3.116  0.117 | -0.497 10.259  0.247 |
p.e       |  0.358  2.174  0.128 |  0.724 13.964  0.525 | -0.434  7.835  0.188 |
ph.d      | -0.072  0.087  0.005 | -0.534  7.576  0.285 |  0.273  3.110  0.075 |
ss.d      |  0.565  5.400  0.319 |  0.045  0.055  0.002 |  0.336  4.710  0.113 |
ssv.d     |  0.712  8.600  0.508 | -0.017  0.007  0.000 |  0.304  3.850  0.093 |
dqo.d     |  0.683  7.911  0.467 | -0.097  0.253  0.010 |  0.256  2.727  0.066 |
dbo.d     |  0.640  6.948  0.410 |  0.168  0.755  0.028 |  0.122  0.624  0.015 |
nkt.d     |  0.281  1.335  0.079 | -0.623 10.331  0.388 | -0.126  0.656  0.016 |
nh4.d     |  0.587  5.844  0.345 | -0.137  0.499  0.019 | -0.370  5.706  0.137 |
p.d       |  0.358  2.171  0.128 |  0.702 13.092  0.492 | -0.410  6.982  0.168 |
v30.b     |  0.124  0.261  0.015 | -0.023  0.014  0.001 | -0.132  0.729  0.018 |
mlss.b    | -0.317  1.701  0.100 |  0.553  8.148  0.306 |  0.412  7.071  0.170 |
mlvss.b   | -0.204  0.704  0.042 |  0.507  6.839  0.257 |  0.396  6.518  0.157 |

Supplementary continuous variables
             Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
ph.s      |  0.089  0.008 | -0.610  0.372 |  0.040  0.002 |
ss.s      |  0.197  0.039 | -0.092  0.008 | -0.089  0.008 |
ssv.s     |  0.218  0.047 | -0.063  0.004 | -0.094  0.009 |
dqo.s     |  0.222  0.049 | -0.146  0.021 |  0.004  0.000 |
dbo.s     |  0.264  0.070 | -0.151  0.023 |  0.026  0.001 |
nh4.s     |  0.281  0.079 | -0.333  0.111 | -0.313  0.098 |
p.s       |  0.280  0.078 |  0.690  0.476 | -0.387  0.149 |

Supplementary categories
             Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2 v.test
Sunday    |  1.968 | -1.315  0.446 -4.414 | -0.291  0.022 -1.224 | -0.947  0.231 -4.982 |
Thursday  |  0.456 |  0.182  0.158  0.597 |  0.146  0.103  0.602 |  0.102  0.050  0.526 |
Monday    |  0.687 |  0.554  0.650  1.860 | -0.079  0.013 -0.330 |  0.231  0.113  1.215 |
Tuesday   |  0.463 |  0.279  0.364  0.928 | -0.018  0.002 -0.075 |  0.170  0.135  0.885 |
Wednesday |  0.592 |  0.233  0.155  0.759 | -0.063  0.011 -0.258 |  0.255  0.186  1.301 |
Saturday  |  0.841 | -0.537  0.408 -1.804 |  0.224  0.071  0.944 | -0.075  0.008 -0.396 |
Friday    |  0.984 |  0.639  0.422  2.122 |  0.083  0.007  0.347 |  0.289  0.086  1.506 |
Autumn    |  2.458 |  1.178  0.230  5.609 |  1.964  0.638 11.719 | -0.416  0.029 -3.106 |
Spring    |  1.912 |  0.477  0.062  2.149 | -1.176  0.378 -6.642 |  0.949  0.246  6.704 |
Summer    |  1.703 | -0.209  0.015 -1.053 | -0.863  0.257 -5.461 | -0.704  0.171 -5.575 |
Winter    |  1.908 | -1.526  0.640 -6.830 |  0.064  0.001  0.361 |  0.341  0.032  2.394 |
> res.dim<-dimdesc(res.pca,axes=1:2)
> res.dim$Dim.1$quanti
        correlation      p.value
ssv.d     0.7124847 5.602414e-62
qr.g      0.7023582 1.526514e-59
ssv.e     0.6894219 1.408253e-56
dqo.d     0.6833497 3.071176e-55
dqo.e     0.6791216 2.513642e-54
dbo.d     0.6404458 1.229367e-46
dbo.e     0.5932682 1.236734e-38
nh4.d     0.5873327 1.017266e-37
```

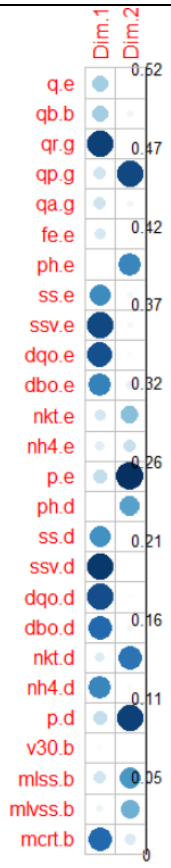
Name:

DNI/Passport:

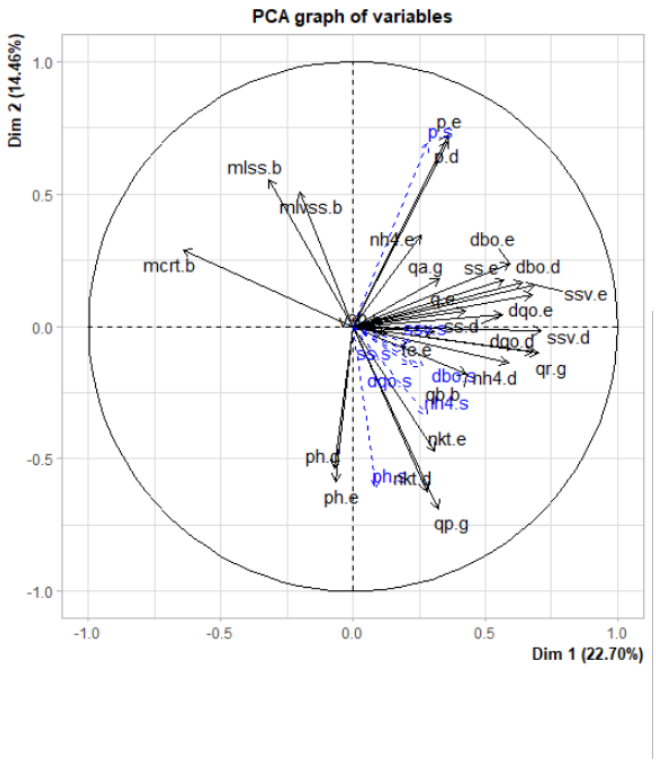
```
ss.e      0.5721938 1.812168e-35
ss.d      0.5645659 2.232342e-34
qb.b      0.4277019 7.258481e-19
q.e       0.4259730 1.036080e-18
p.e       0.3582470 2.584187e-13
p.d       0.3579759 2.700828e-13
qa.g      0.3268374 3.279664e-11
qp.g      0.3235409 5.285378e-11
nkt.e     0.3097485 3.658654e-10
fe.e      0.3059784 6.103125e-10
nkt.d     0.2807423 1.560413e-08
nh4.s     0.2806712 1.574026e-08
p.s       0.2798101 1.748350e-08
dbo.s     0.2641147 1.114250e-07
nh4.e     0.2567452 2.553898e-07
dqo.s     0.2221370 9.011223e-06
ssv.s     0.2176553 1.374927e-05
ss.s      0.1966960 8.843512e-05
v30.b     0.1241087 1.393657e-02
mlvss.b  -0.2039058 4.761320e-05
mlss.b   -0.3169101 1.356469e-10
mcrt.b   -0.6415450 7.698168e-47
> res.dim$Dim.1$quali
                R2      p.value
season  0.16263629 7.141169e-15
weekday 0.07106556 7.473268e-05
> res.dim$Dim.1$category
                  Estimate      p.value
season=Autumn   1.1977677 1.091000e-08
season=Spring   0.4968756 3.147688e-02
weekday=Friday  0.6337188 3.365840e-02
weekday=Sunday -1.3196677 8.137259e-06
season=Winter  -1.5062108 1.995308e-12
>
> res.dim$Dim.2$quanti
        correlation      p.value
p.e       0.7244940 5.244326e-65
p.d       0.7015151 2.408786e-59
p.s       0.6897287 1.202710e-56
mlss.b    0.5534197 7.804297e-33
mlvss.b   0.5070361 5.382807e-27
nh4.e     0.3422442 3.262649e-12
mcrt.b    0.2872585 6.964297e-09
dbo.e     0.2385156 1.784642e-06
qa.g      0.1795002 3.547738e-04
ss.e      0.1736323 5.543295e-04
dbo.d     0.1684703 8.114551e-04
ssv.e     0.1557876 1.978142e-03
dqo.e     0.1194720 1.796434e-02
qr.g     -0.1013290 4.496652e-02
nh4.d    -0.1369596 6.611366e-03
dqo.s    -0.1461306 3.736633e-03
dbo.s    -0.1514422 2.645495e-03
qb.b     -0.1807289 3.225500e-04
nh4.s    -0.3334086 1.245030e-11
nkt.e    -0.4721469 3.705056e-23
ph.d     -0.5336430 3.108022e-30
ph.e     -0.5876380 9.137282e-38
ph.s     -0.6096094 2.965484e-41
nkt.d    -0.6231518 1.521602e-43
qp.g     -0.6881221 2.741497e-56
> res.dim$Dim.2$quali
              R2      p.value
season 0.4033119 3.137887e-43
> res.dim$Dim.2$category
                Estimate      p.value
season=Autumn  1.9662031 1.526992e-38
season=Summer -0.8600331 2.721077e-08
season=Spring -1.1732867 8.590487e-12
>
> plot.PCA(res.pca,choix="var",select="contrib 10")
>
> # Modern facilities
> #
> library(factoextra)
> fviz_pca_contrib(res.pca,choice="var")
```

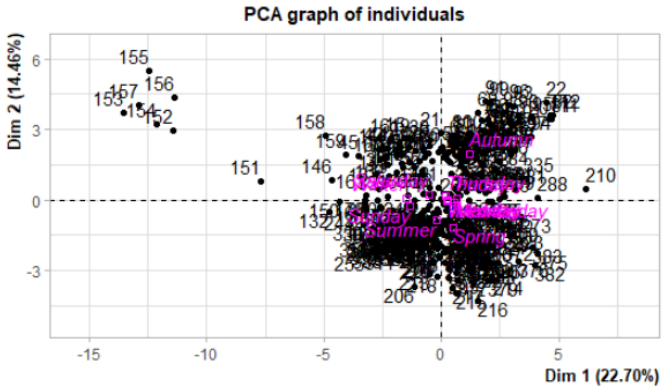
Name:
DNI/Passport:






```
>
> library("corrplot")
> corrplot(res.pca$var$cos2[,1:2], is.corr=F)
```



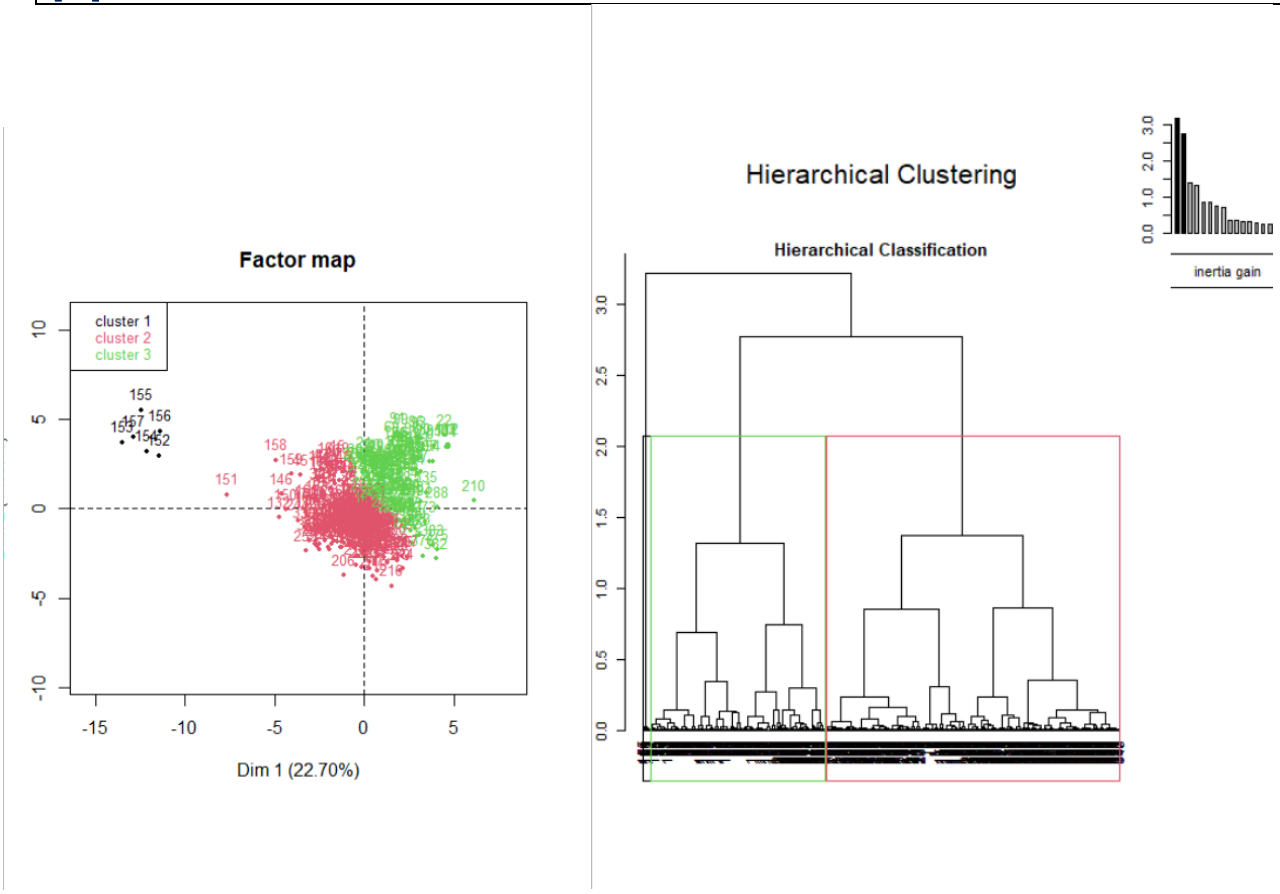


Name:
DNI/Passport:

8. A Hierarchical Clustering is addressed. How many clusters are needed to represent 60% of the total inertia.

According to 60% inertia representation, 10 clusters have to be taken. Default according to R are 3. Take a look at cluster 1: it is a group of outliers already seen in the individual projection of PCA.

```
> res.hcpc<-HCPC(res.pca, nb.clust=-1)
> 100*(res.hcpc$call$t$within[1]-res.hcpc$call$t$within[1:12])/res.hcpc$call$t$within[1]

[1]  0.00000 16.02141 29.81789 36.66827 43.23720 47.52454 51.75035 55.44035 58.87834 60.61508
[11] 62.32264 63.82283
```



9. A nondefault criteria for selecting the number of clusters to 3 has to be set. Explain the characteristics of cluster number 3.

You have to specify res.hcpc(res.pca, nb.clust = 3). It is the green cluster shown at the end of the previous question. It is characterized by remarkable large values for p.e, ssv.e (phosphatus and suspension solids) over the mean and small values for ph group (ph.s, ph.d, ph.e, so more acid water than the average) under the corresponding variable means.

```
> res.hcpc$desc.var$quanti
...

$`3`
          v.test Mean in category Overall mean sd in category   Overall sd      p.value
p.e     11.685195     1.432745e+01 9.911206e+00   5.698221e+00 5.663353e+00 1.517316e-31
ssv.e   11.588475     2.063007e+02 1.573036e+02   6.824545e+01 6.335794e+01 4.714524e-31
p.d     11.421674     8.837020e+00 6.080539e+00   3.564827e+00 3.616446e+00 3.258937e-30
dbo.e   10.525008     2.808741e+02 2.132156e+02   1.157397e+02 9.632900e+01 6.625571e-26
ss.e    10.517681     2.689336e+02 2.094413e+02   9.137954e+01 8.476119e+01 7.161486e-26
p.s     10.454190     3.986405e+00 2.746848e+00   1.757217e+00 1.776778e+00 1.401921e-25
dqo.e   10.118330     5.520944e+02 4.408393e+02   1.812088e+02 1.647661e+02 4.581646e-24
dbo.d    9.678390     1.444441e+02 1.198839e+02   3.200263e+01 3.802631e+01 3.725379e-22
ssv.d    8.835456     7.383217e+01 6.488520e+01   1.448075e+01 1.517412e+01 9.969071e-19
```

Name:
DNI/Passport:

```
ss.d      7.777184     9.894056e+01 8.868240e+01   1.960212e+01 1.976532e+01 7.415657e-15
dqo.d     7.090350     2.769965e+02 2.487985e+02   5.215357e+01 5.959473e+01 1.337731e-12
nh4.e     6.179342     6.343810e+01 4.129155e+01   8.063043e+01 5.370574e+01 6.436929e-10
qa.g      6.173617     2.526906e+05 2.321780e+05   4.974165e+04 4.978964e+04 6.674498e-10
nh4.d     5.806236     2.774325e+01 2.594749e+01   3.725238e+00 4.634597e+00 6.389294e-09
qr.g      4.793757     4.222604e+04 4.094602e+04   2.575331e+03 4.001248e+03 1.636862e-06
mlvss.b   4.506336     1.424860e+03 1.345756e+03   2.762398e+02 2.630451e+02 6.595652e-06
q.e       4.361334     4.331104e+04 4.182027e+04   3.295655e+03 5.122072e+03 1.292721e-05
mlss.b    3.757119     1.858671e+03 1.768566e+03   3.002693e+02 3.593773e+02 1.718808e-04
nkt.e    -2.032657     4.069989e+01 4.195585e+01   9.071656e+00 9.259087e+00 4.208719e-02
qp.g     -3.814115     5.862056e+02 6.253282e+02   1.406147e+02 1.537060e+02 1.366720e-04
nkt.d    -3.816880     3.439212e+01 3.641496e+01   5.871284e+00 7.941639e+00 1.351500e-04
ph.s     -4.912113     7.468531e+00 7.532781e+00   1.702198e-01 1.959998e-01 9.010005e-07
ph.d     -4.926920     7.516084e+00 7.562245e+00   1.426033e-01 1.403965e-01 8.353589e-07
ph.e     -5.780302     7.563986e+00 7.618367e+00   1.335902e-01 1.409795e-01 7.456659e-09

> res.hcpc$desc.var$category
...

$`3`
                Cla/Mod   Mod/Cla   Global      p.value    v.test
season=Autumn  80.00000 55.944056 25.51020 2.234281e-25 10.409912
weekday=Sunday 21.05263  8.391608 14.54082 7.672544e-03 -2.666150
season=Winter  18.68132 11.888112 23.21429 3.517233e-05 -4.137100
season=Summer  13.76147 10.489510 27.80612 1.432413e-09 -6.051794
```

10. Use a partition method to group available data into the selected number of clusters found in Question 8. Determine the quality of the partition and plot the resulting partition in the first factorial plane.

K-Means partition to 10 clusters based on 8 Ortogonal Component in PCa captures 78% of total inertia, instead of 60% represented by Hierarchical Clustering.

```
> dis<-dist(res.pca$ind$coord[,1:8])
> res.km<-kmeans(dis,10)
> res.km$betweenss/res.km$totss
[1] 0.7833207
> table(res.km$cluster)

 1  2  3  4  5  6  7  8  9 10
28 24 30  7 68 49 34 74 36 42
>
> ff<-factor(res.km$cluster)
> plot(res.pca$ind$coord[,1:2],col=ff, pch=19, main= "K-Means - 10 cluster - First Factori
al Plane")
> legend("bottomleft", title="K-Means partition", legend=levels(ff),
+        col=1:10, pch=19, cex=0.8)
```