Assignement 1: PCA – CA
Select Topic

Teacher: Miguel Reyes
Student: 403715 BARBE Victor, 403719 Antoine Bregeon

In this assignement, we are going to apply Principal Component Analysis (PCA) and Correspondence Analysis (CA) on two different datasets. We have chosen two distinct datasets because to work properly, PCA and CA requires specific formats of values.

I – Principal Component Analysis

A – Dataset

Breast cancer is a plague that unfortunately hits too many people. We found interesting to use such a dataset to find conclusion about tumors and patients measures.
The dataset is used to determine whether a tumor is malignant or benign. The values have been found following this process:
For each patient, 30 photos of the tumor have been taken. Each value in the dataset comes from the mean measure of these 30 photos.
The variables are the following:

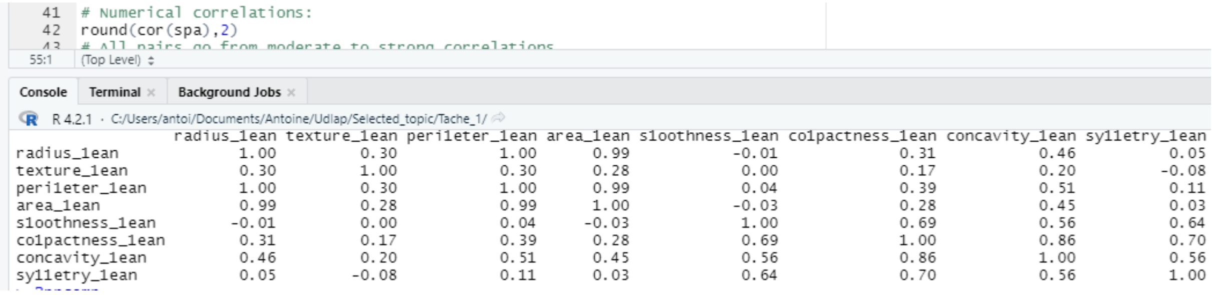
- radius_mean: mean radius from center of the tumor to the perimeter
- texture_mean: standard deviation of gray-scale values. When the tumor's picture is taken, it is in black and white. The different textures are recognizable by their gray-scale values.
- perimeter_mean: mean size of the core tumor
- area_mean: mean area covered by the tumor
- smoothness_mean: mean of local variation in radius length, gives an idea of the tumor's shape
- compactness_mean: mean of perimeter^2 / area -1
- concavity_mean: mean of severity of concave portions of the contour of the tumor
- symmetry_mean: gives a ratio of the tumor's symmetry, giving as smoothness_mean an idea of its shape

B – Why to apply PCA

We have chosen this dataset because he is shaped to perform PCA. He contains only continuous numerical values and gives a binary prediction: the tumor is benign or malignant. Also, there is a huge correlation between variables. Finally, the field of study if interesting and useful for the common good. We currently have nine columns in our dataset. Thanks to PCA, we will reduce this size. With this process, we will be able to predict much more easily the tumor's state. It will allow us to answer this question: which anatomical characteristics allows us to determine whether a tumor is benign or malignant?

C - Explain PCA

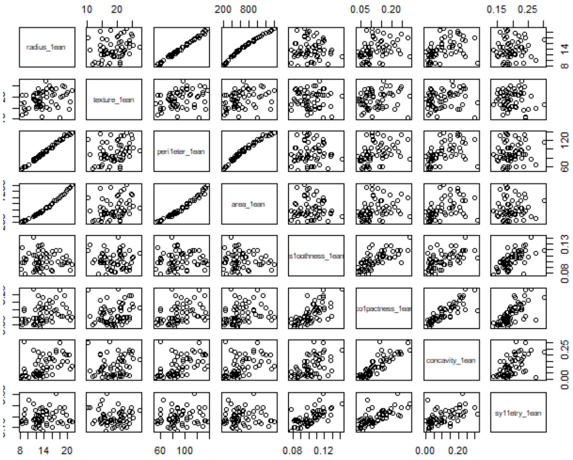
- The first step of PCA is to calculate the covariance matrix of the whole dataset:



```
# Numerical correlations:
round(cor(spa),2)
# All pairs go from moderate to strong correlations
```

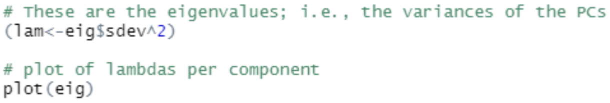
55:1
(Top Level)
Console
Terminal
Background Jobs
R 4.2.1 · C:/Users/antoi/Documents/Antoine/Udlap/Selected_topic/Tache_1/

```
                 radius_1ean texture_1ean peri1eter_1ean area_1ean s1oothness_1ean co1pactness_1ean concavity_1ean sy11etry_1ean
radius_1ean             1.00         0.30           1.00      0.99           -0.01             0.31           0.46          0.05
texture_1ean            0.30         1.00           0.30      0.28            0.00             0.17           0.20         -0.08
peri1eter_1ean          1.00         0.30           1.00      0.99            0.04             0.39           0.51          0.11
area_1ean               0.99         0.28           0.99      1.00           -0.03             0.28           0.45          0.03
s1oothness_1ean        -0.01         0.00           0.04     -0.03            1.00             0.69           0.56          0.64
co1pactness_1ean        0.31         0.17           0.39      0.28            0.69             1.00           0.86          0.70
concavity_1ean          0.46         0.20           0.51      0.45            0.56             0.86           1.00          0.56
sy11etry_1ean           0.05        -0.08           0.11      0.03            0.64             0.70           0.56          1.00
```



The covariance matrix is a square matrix 8*8 in this case. It’s a symmetric matrix and its diagonal is 1. It determines the correlation between each variable of the dataset. For example, radius_mean and concavity_mean have a covariance of 0.46. The value goes from -1 to 1. The higher is the values higher the correlation between the two variable is strong.

- The second step it’s to compute eigenvectors and the corresponding eigenvalues



```
# These are the eigenvalues; i.e., the variances of the PCs
(lam<-eig$sdev^2)

# plot of lambdas per component
plot(eig)
```

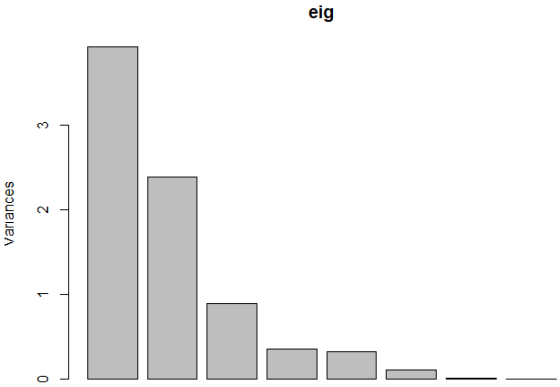
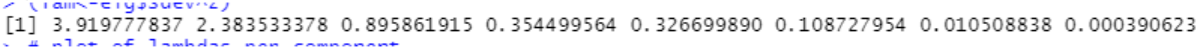


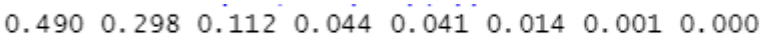
Fig : scree plot



```
> (lam<-eig$sdev^2)
[1] 3.919777837 2.383533378 0.895861915 0.354499564 0.326699890 0.108727954 0.010508838 0.000390623
> # plot of lambdas per component
```

The result is the 8 eigenvalues corresponding to the eight dimensions of the dataset. The value of each eigenvalue gives the global variation. We plot this value to have a better overview of the values.

After we can calculate the importance in percentage



```
0.490 0.298 0.112 0.044 0.041 0.014 0.001 0.000
```

On this result:

49% of the total variation is explained just by z1!

29,8% of the total variation is explained by z2

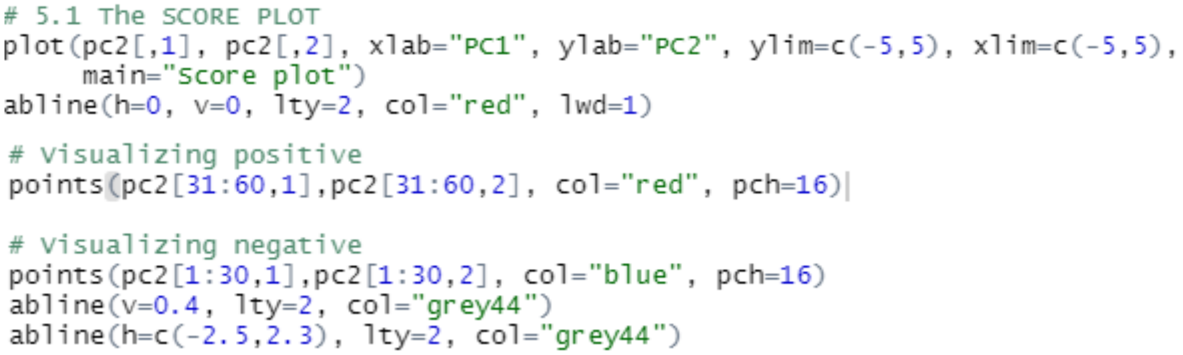
if we take z1 y z2 it only represented 78,8% de la variation. We can represent the whole data in 8 dimensions in basically 1 dimension with around 50% of accuracy. We can represent the whole data in just 2 dimensions with 82% of accuracy. We might take z3 also but is it to small that it can be considered as noise.

According to the scree test, we identify the point where the eigenvalues become too small, we can identify that after the two first components the importance of other components become too small.

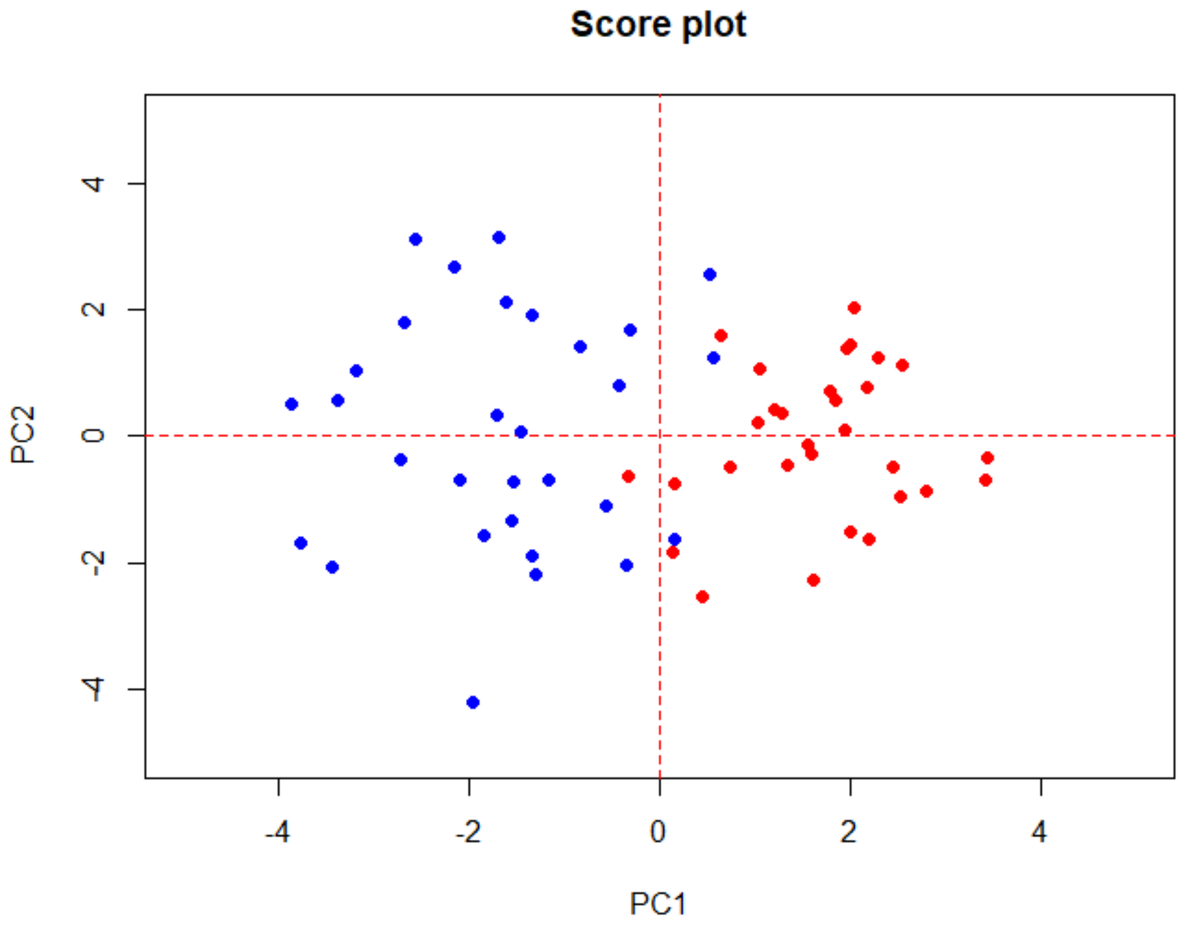
- Plotting two-dimension PC, PC2 to find characteristic

In this case we chose to keep the first to eigen vectors, PC1 and PC2.

In the original configuration we couldn't plot the 8-dimension variable. But now the original 8-dimension data can be plot with 2 dimensions. Doing so involves distortion. In this case the distortion is less than 20%. This projection is 80% accurate!



```
# 5.1 The SCORE PLOT
plot(pc2[,1], pc2[,2], xlab="PC1", ylab="PC2", ylim=c(-5,5), xlim=c(-5,5),
     main="Score plot")
abline(h=0, v=0, lty=2, col="red", lwd=1)

# visualizing positive
points(pc2[31:60,1],pc2[31:60,2], col="red", pch=16)

# visualizing negative
points(pc2[1:30,1],pc2[1:30,2], col="blue", pch=16)
abline(v=0.4, lty=2, col="grey44")
abline(h=c(-2.5,2.3), lty=2, col="grey44")
```



We can see a clear difference between the two groups. The biggest difference is in the PC1, has we can see a huge difference on the X axis.The cancer patients have positive values of PC1 and concentrates around 0 of value on PC2. They have higher means contents.

In the other side the non diagnostics patient have lower means constant their value of PC2 are more extend and extreme.

- Biplot

The plot of the score is quite like the previous one. The vectors represent the original variables and their importance on each PC. The length of arrows is proportional to its variance. The direction of vector gives us information on the correlation:

- two vectors with similar or opposite direction have a strong correlation.
- If the vectors are orthogonal to other, they have a small correlation

To notice the influence of each variable on PC1 and PC2, you have to measure the distance projecting the vector on axis

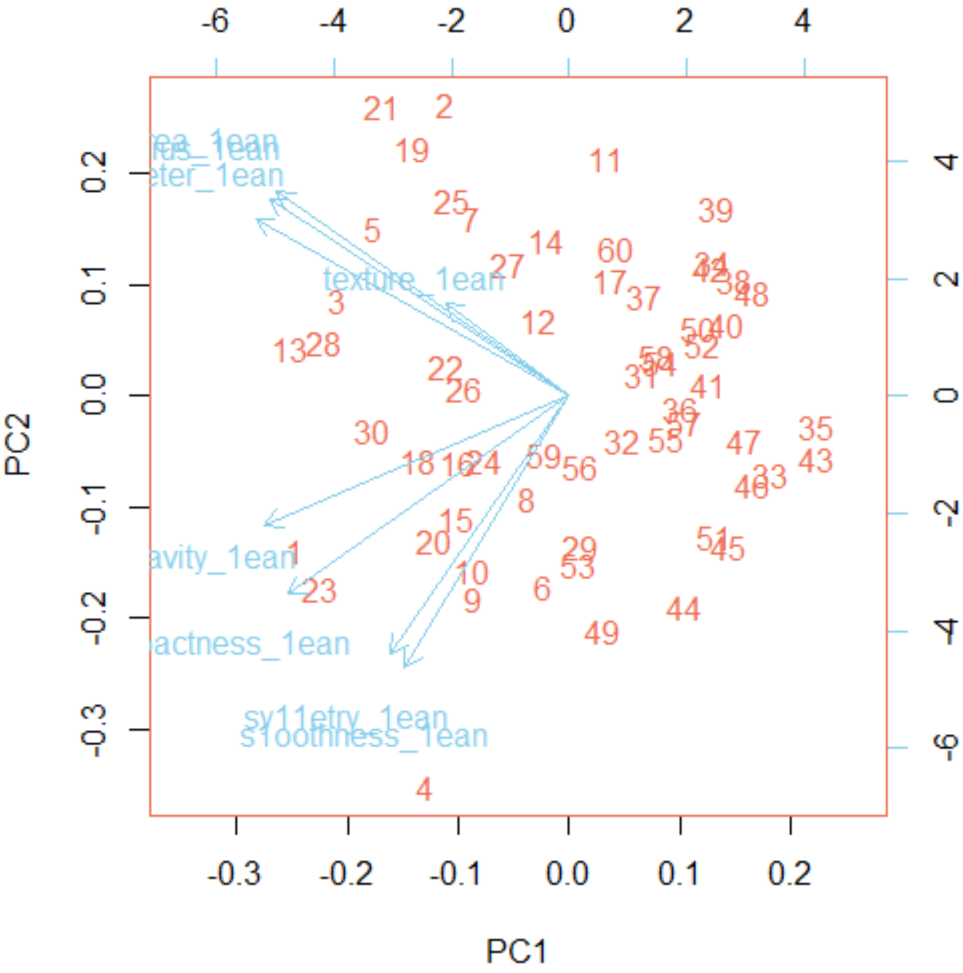


Fig biplot

D - Answer the question

We can see a clear distinction between cancer tumor and benign tumor. The difference can't not be reduced to the presence of some factor or not but on the intensity of them. Meaning if the tumor characteristics like radius perimeter and area are high the chance of cancer tumor is higher. And it works the same for the opposite.

2 – CA

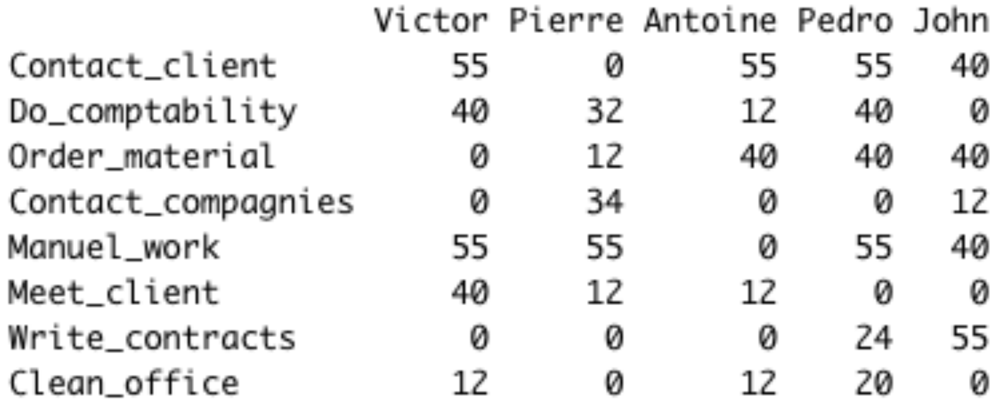
A – What is the data set

For the second part of PA, we will be using a dataset we will create that will be similar to the one we used in class with the chores. The idea is that we will create a contingency table about a family-owned company. There will be different tasks that have to be done in the company (contact clients, …) and we will see the repartition of the tasks among the different members of the family working there, which means how many times each person performed a task.

B – Why are we applying PA, what are the questions we want to answer

Here, it is relevant to apply PA because we are in the case of categorical values and not continuous data. Here we have different categories on the row and the columns, and the contingency table we defined is giving us the frequency distribution of the variables. As a result of the PA, we will get a score factor, that will allow us to see the correlation between row and columns in a low dimension space.

Here is the contingency table we decided to define:

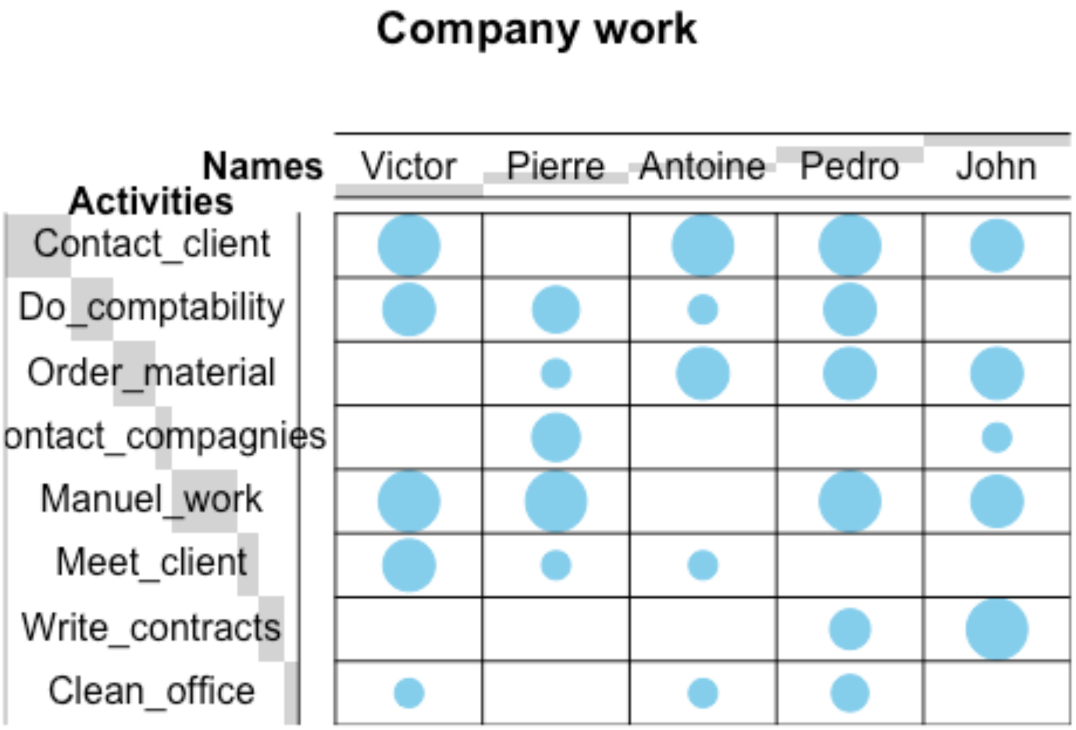


```
                    Victor Pierre Antoine Pedro John
Contact_client          55      0      55    55   40
Do_comptability         40     32      12    40    0
Order_material           0     12      40    40   40
Contact_compagnies       0     34       0     0   12
Manuel_work             55     55       0    55   40
Meet_client             40     12      12     0    0
Write_contracts          0      0       0    24   55
Clean_office            12      0      12    20    0
```

It is giving us the frequency repartition of each task depending on the member of the family who is working in the company. Our goal will be to use PA to reduce the dimensionality and create a 2D plot.

<u>What are the questions we want to ask ourselves threw this project?</u>

In the end, we will want to highlight the relation between different activities and some people working in the company. This project should help us to determine a profile for each attribute (which task is performed by who, what are the correlation between different people and tasks …). Here at the beginning of our code, we can see a first view of the data to highlight the repartition.



C – Explanation of the process of CA

Before explaining the code, and how the theoretical process applies to this example, we will summarize shortly the principle of CA. It is an extension of PCA that applies only to categorical data, like we can see in the graph above. The goal of using CA will be to display a two-dimension biplot that will highlight structure hidden in the multivariate data. The biplot will allow us to visualize geometrically the hidden patterns.

The CA is generally applied to a contingency table, which here will be the dataset we showed above: a contingency table displays the frequency distribution of the variable, here the frequency being who did what task and how many times.

The principle of CA is then to compute the residual of each sell in the table, which is the difference between the value we would observe is the variables were uncorrelated, and the value we actually observe. The bigger the residual is (negative or positive), the bigger the correlation is. We can then also compute the row and column mass, which are the added values of that column or row. Then to get the expected value of a cell, we just multiply the row mass by the column mass.

Now that we have the expected and the actual value of each cell, we can compute the table of residuals. This table is going to give us more information about the data. For our example, if we were to get a negative value in the residual table between Pedro and meet clients, this means that Pedro is less likely than other people to perform this activity.

Now to get more accurate information, we can compute the indexed Residual. This will compensate for the difference in the sample size of attributes (if someone did more activities as an example). To do so, we divide the residual value we calculated and divide it by the expected value.

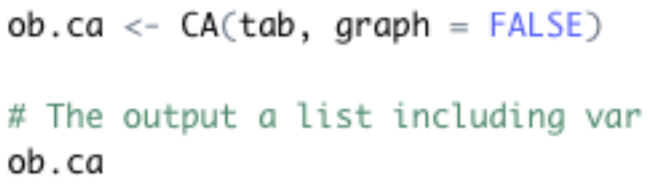
Using this method, we now get a percentage relation between attribute, hence if we get 20% in the case of Antoine and meet_client it means that Antoine is 17% more likely to perform this task then if there was no relationship between persons and the task they perform.

Then to compute the CA, we take the SVD of this index residual matrix. The result of SVD we get will give us a vector, and two matrices that will give us the information needed about the dimensions, columns and rows to get the CA. Using the left and right singular vector, we can now compute the CA and plot the result. Let's now see the application on our dataset.

D – Answering the questions

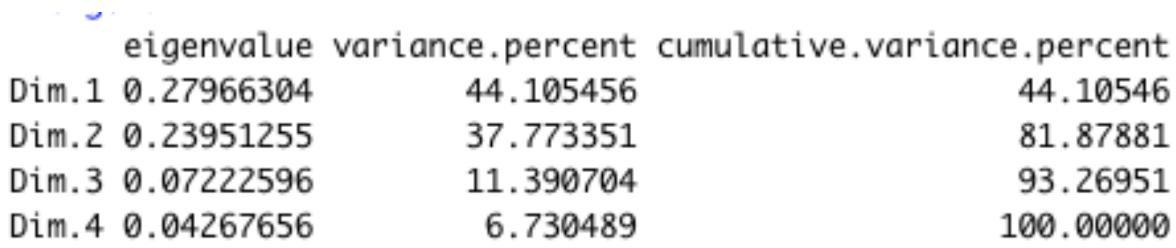
Now we start programming in R. First, we define the data we described above, and we pass the Pearson's Chi-squared test. We get as an output "X-squared = 570.04, df = 28, p-value < 2.2e-16". With such a low p-value, we can make the hypotheses that there will be a relationship between the different categories. This means that there is a point of using CA to reduce the dimension, since the data is corelated.

We then compute the correspondence analysis using a function and display the output. This function will follow the previously described process of CA.



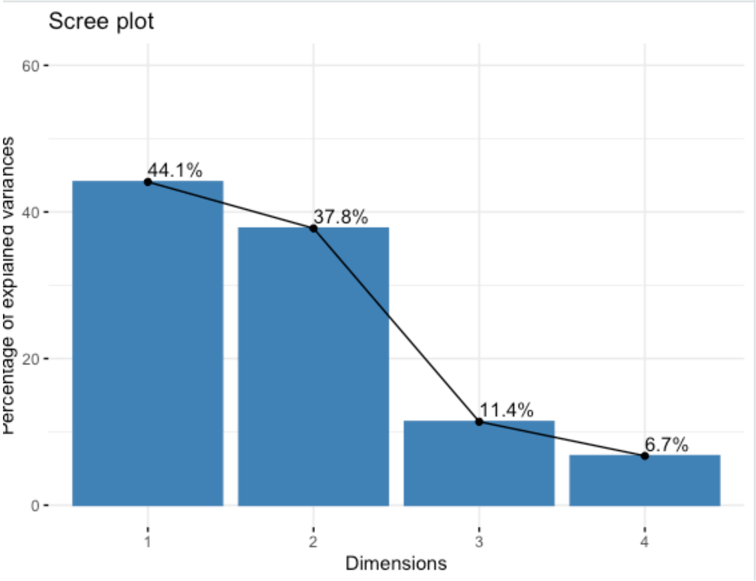
```
ob.ca <- CA(tab, graph = FALSE)

# The output a list including var
ob.ca
```

Now to see the number of dimensionalities we will need to consider when representing the data in low dimension, we will output the eigenvalues of our CA. We get this output:

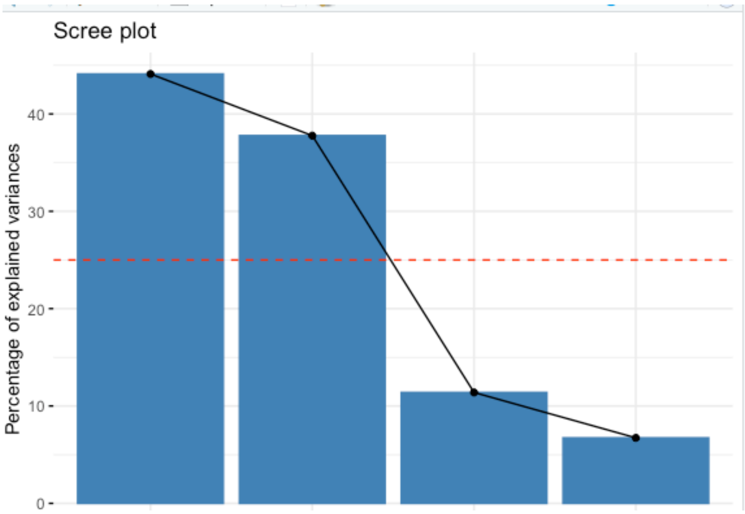


```
      eigenvalue variance.percent cumulative.variance.percent
Dim.1 0.27966304        44.105456                    44.10546
Dim.2 0.23951255        37.773351                    81.87881
Dim.3 0.07222596        11.390704                    93.26951
Dim.4 0.04267656         6.730489                   100.00000
```

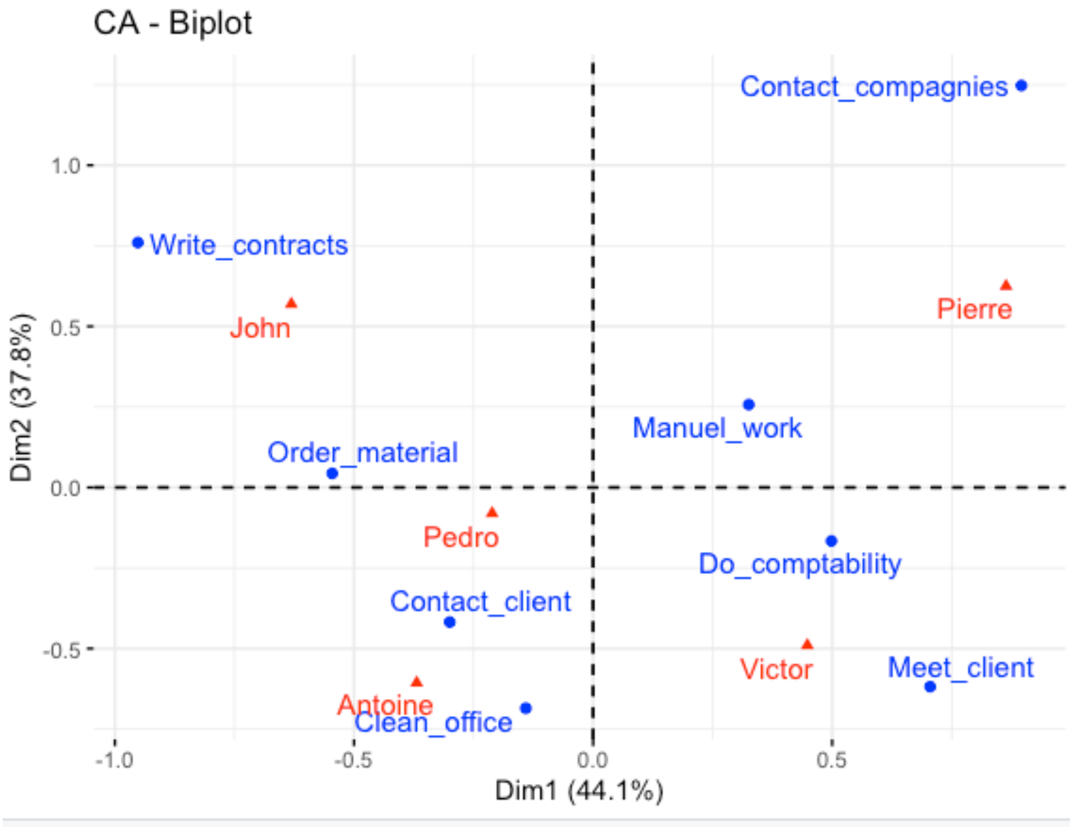
This means that our contingency table is described at 81% using the first 2 dimensions, which is a pretty good description of the dataset. We will be able to plot 2D graphs that describe at 81% the variance in our dataset. The third dimension explains 11% more, and the last one 6. We will see later how this can be visualized using a graph. We then apply a few plots seen in class, which are describing the way the dimensions seen with the eigenvalues are describing our data.



This graph shows the different dimensions of our CA, and how they represent our dataset using the variance just like we explained above. We can now add a red line displaying the average eigenvalue. We have 4 dimensions, so each should cover 25% of the variance. When plotting the graph again, we can see that the 2 first dimensions are way above the average, and the other 2 dimensions are way under, thus confirming our choice of a 2D solution.



Then we can create a symmetric plot of the CA, which goal is to display in two dimensions only the different variables (row in blue, columns in red). This plot will give us a better idea of the relation between the data and will show the different “profiles” we wanted to extract from the data.

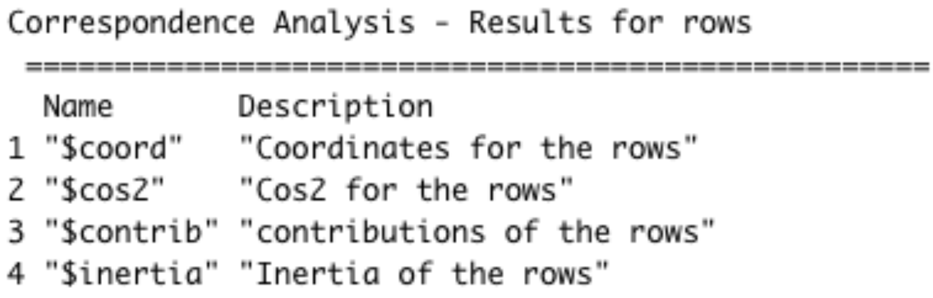


Here, the distance between two attributes of the same dimension (row or column) is showing their similarity. We cannot measure directly the distance between row or columns, we will have to use a asymmetric biplot to get information and interpret the distance between row and columns. To do so, columns would have to be displayed in a "row space" and the other way around. However, it still gives a very good idea of the different areas that exist in the data.

Indeed, we can see on this plot that Victor's profile is more oriented around meeting clients and doing computability. We can also see that Victor and John profiles are very different as they are located in the opposite quartile, which we could have seen looking at the previous balloon graph since they nearly have no activities in common.

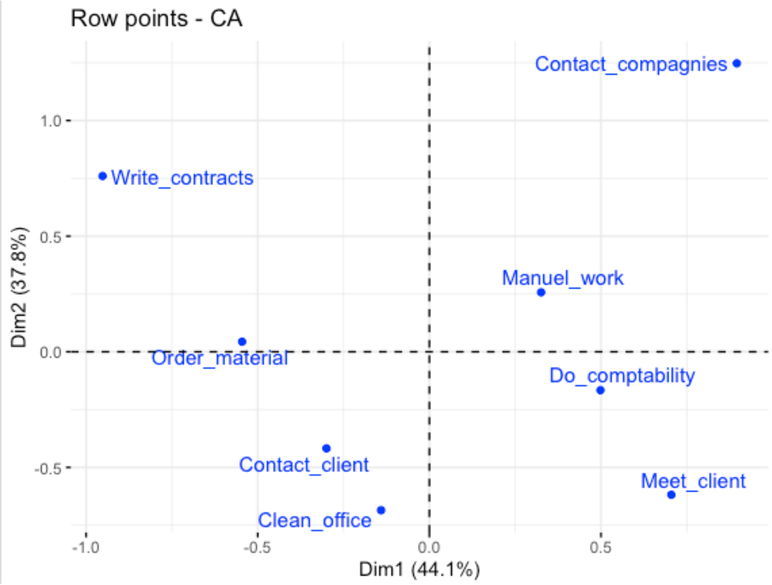
We can also see that Antoine and Pedro's profiles are similar since they are located in the same quartile and often perform the same tasks. In general, profiles in the same quartile will have similarity, and the one diagonally opposed will show a lot of differences.

Then, we can create other plots that will be showing different aspects about the rows of our data (the different activites). We will first get some information about the rows using "row <- get_ca_row(ob.ca)"



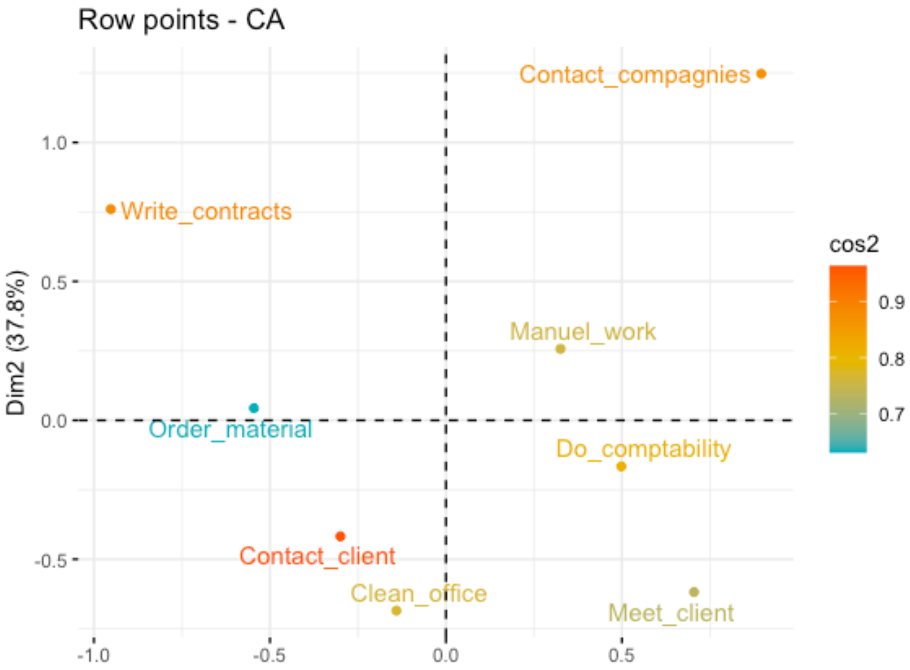
```
Correspondence Analysis - Results for rows
 ===================================================
  Name       Description
1 "$coord"   "Coordinates for the rows"
2 "$cos2"    "Cos2 for the rows"
3 "$contrib" "contributions of the rows"
4 "$inertia" "Inertia of the rows"
```

Then we can display some of that information in tabs. Here we can also display a symmetric plot of the rows only, that will give us extra information. The rows which are grouped together have a similar profile. When rows are correlated negatively (an increase in one means a decrease in the other) they are positioned in opposite quartiles.

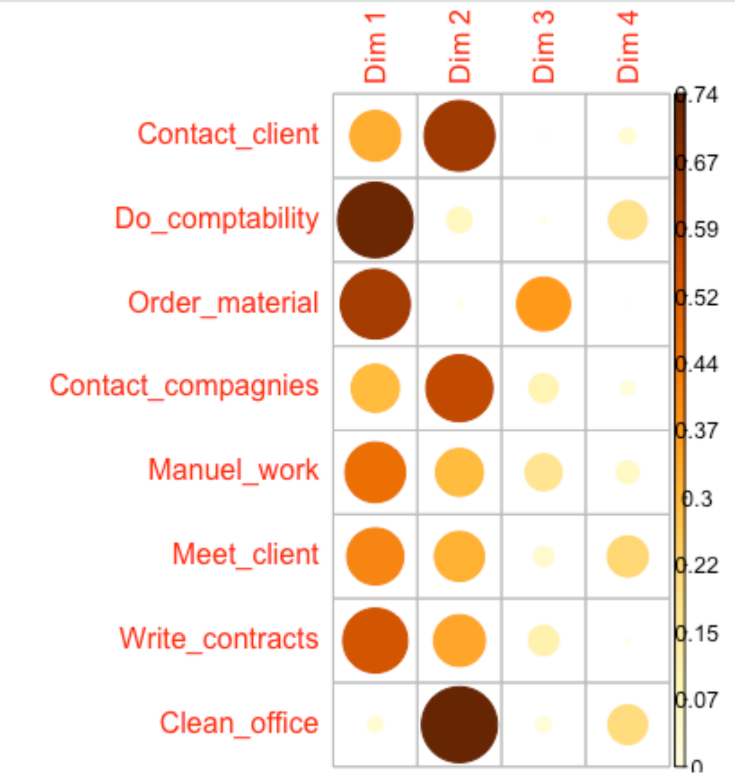


Now we would like to see how well our data is represented in the 2D spaces, and how different attributes are correlated to some of the 4 dimensions.

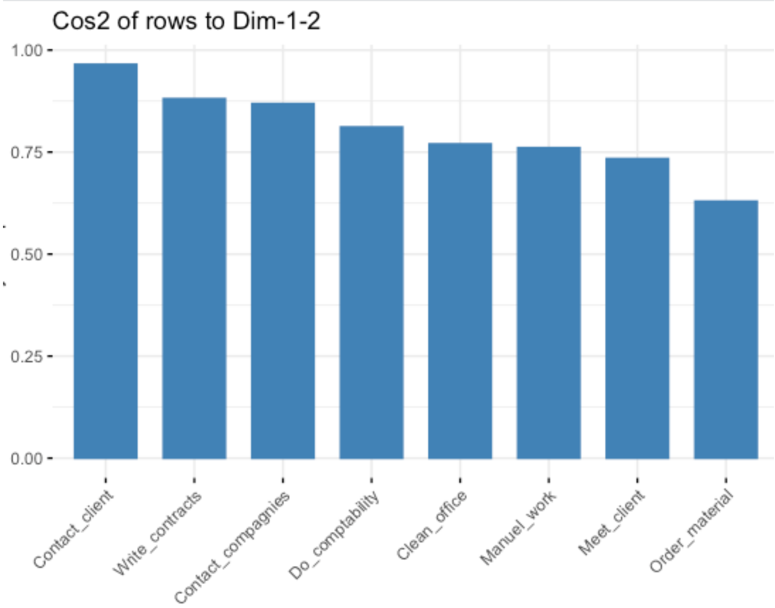
Using cos2 values here, we can see how well our data is represented in the 2D space. Here we can see that most values are well represented, but order material isn't that much.



After that, we can make a corr plot showing the cos2 of points depending on their dimensions. We can see that Order_material has a high value in dimension 3, which cannot appear in the 2 dimensions plot. This is why the cos2 value of Order_material on the 2d plot was lower.

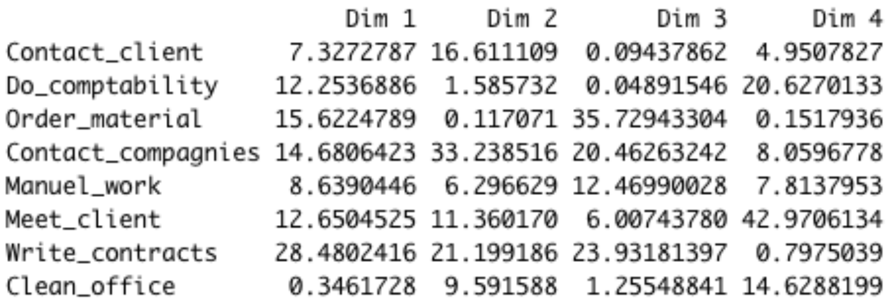


We can represent a bar plot of the rows' cos2 over dimension 1 and 2.



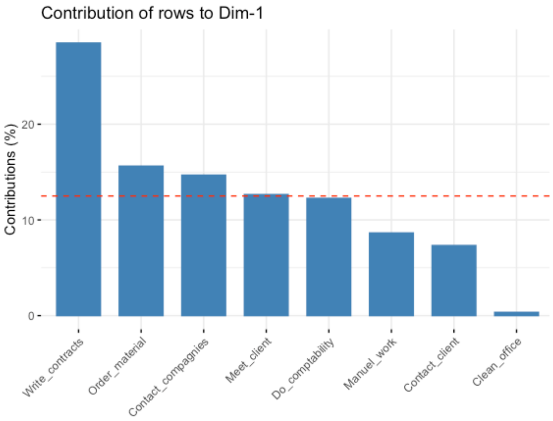
We can see again that the value of order_material is low. This means that the position of the point order_material on the biplot should be interpreted with some caution, as we would need a 3d dimension for that point to be described accurately.

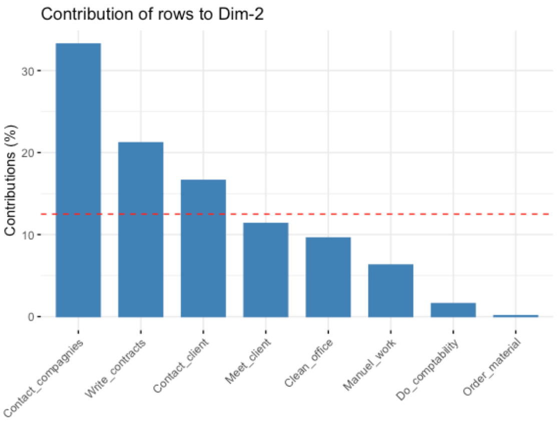
We can also get the contribution of rows (in %) to the definition of the dimensions. We can see again that order material participates a lot to dimension 3. The rows that contribute the most to Dim.1 and Dim.2 are the most important in explaining the variability in the data set.



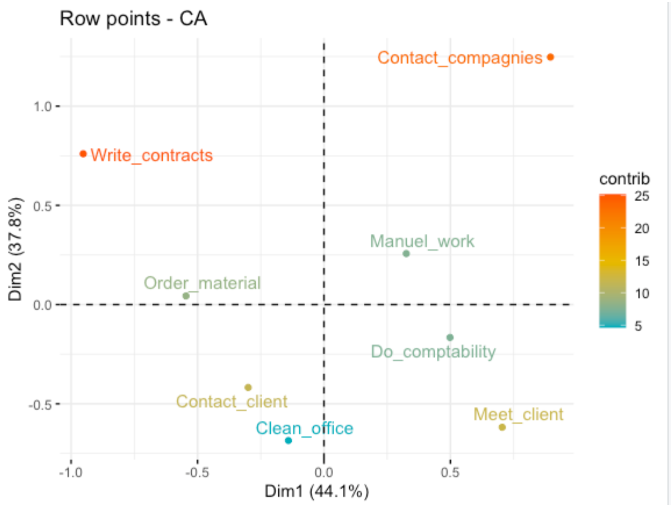
```
                        Dim 1     Dim 2       Dim 3      Dim 4
Contact_client      7.3272787 16.611109  0.09437862  4.9507827
Do_comptability    12.2536886  1.585732  0.04891546 20.6270133
Order_material     15.6224789  0.117071 35.72943304  0.1517936
Contact_compagnies 14.6806423 33.238516 20.46263242  8.0596778
Manuel_work         8.6390446  6.296629 12.46990028  7.8137953
Meet_client        12.6504525 11.360170  6.00743780 42.9706134
Write_contracts    28.4802416 21.199186 23.93181397  0.7975039
Clean_office        0.3461728  9.591588  1.25548841 14.6288199
```

Here we can see more easily the contribution of each row to the two first dimensions. Write_contract and order_material contribute most to dimension 1, contact_companies and write_contract the most to dimension 2.

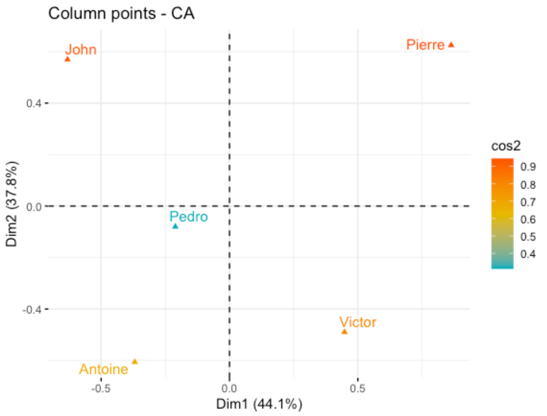


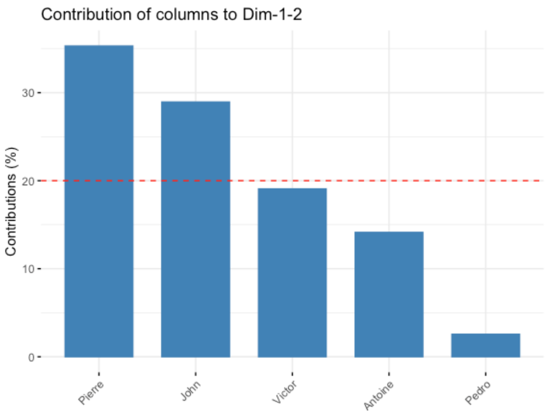


Here is a two dimension plot showing the contribution of each row point on the 2D space. Contact_companies and manuel_work have the most contribution to the positive side of the first dimension. Write_contracts and order_material have the most contribution to the negative side of the first dimension. This means that dimension is defined by the opposition of contact_companies, manuel_work and write_contracts, order_material. Same goes for the other dimension with the other two quartiles.



Now we will take a closer look to the columns. Just like the rows, we can create a simple 2D plot, or a more interesting plot with the cos2 values to see how well represented the different rows are. When the cos2 value is close to one, the column is well represented on the 2d space. When it goes much lower, it might not be well represented on those dimensions. Here we can see that Pedro is not very well represented on two dimension and might need the 3rd or 4th one to be described properly.





To conclude, we can say here that CA allowed us to reduce the number of dimensions of our dataset, to allow us to visualize it in two dimensions only. The different plots we saw allow us to understand why some variables are better represented into two dimensions, because other depend more on the 3rd and 4th dimension. In the end as we saw with the eigenvalues, we still managed to represent 80% of the variance of the dataset using 2 dimensions, and most of the rows and columns are described properly.

We also managed to show in the two dimensions plot the relation between different rows and columns, and to determine different profile of workers in the company.