Compte-rendu du TP n°3 – Analyse factorielle des composantes

**Noms du groupe : AMINI Nada, ESSAYEGH Nour**

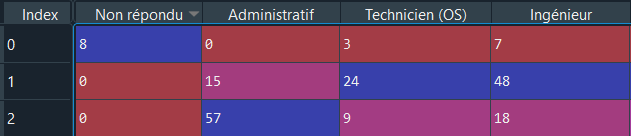
**----------------------------------------------------------------------------------------------------------------**

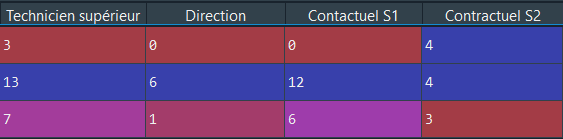
**Il s’agit de :**

1. **Mettre au point un script qui permet successivement de :**

* **Faire les transformations de données à partir des données fournies et des matrices nécessaires à la réduction.**

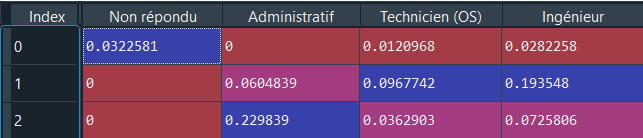
Pour cela, on commence d’abord par importer le fichier de données .csv, on ne retient que les colonnes « Sexe » et « Fonction » qui feront l’objet de cette étude, et ensuite on procède à la construction de la matrice de contingence.

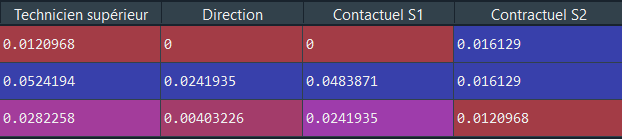




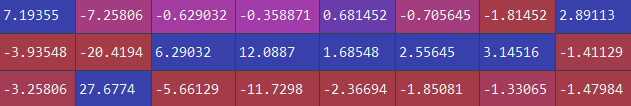
Pour ce qui est de la colonne des indices, 0 correspond à une personne n’ayant pas précisé son sexe, 1 correspond à un homme et 2 à une femme.

On peut maintenant construire la matrice des fréquences :





On calcule la matrice de distribution des correspondances si les deux variables qualitatives sont indépendantes pour avoir une première idée sur le niveau de corrélation de ces deux dernières. On calcule ensuite la matrice R = T – T0 (vue en cours) des écarts entre les données de la matrice de contingence et la matrice de contingence sous hypothèse d'indépendance. On obtient la matrice suivante :



On remarque que cette matrice est différente de la matrice nulle, par conséquent on dit qu'on a dépendance entre les différentes variables (les fonctions), mais il faut vérifier ce propos avec la statistique du .

**Test du :**

>freq\_ind = T0 / total

>khi = 0

>for i in range(3):

> for j in range(n):

> khi += ((mat\_contingence.loc[i,"fonction {}".format(j)] - freq\_ind.loc[i,"fonction >{}".format(j)])\*\*2)/freq\_ind.loc[i,"fonction {}".format(j)]

>

>khi \*= total

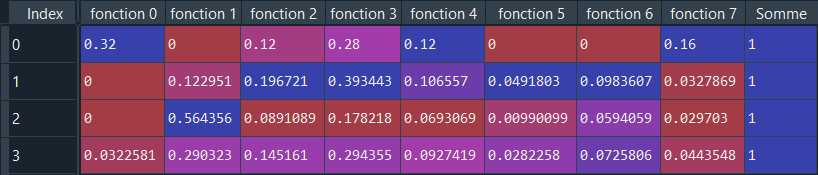
>print("La statistique du Khi 2 vaut: " + str(khi))

La statistique du vaut : 24120966.007713992

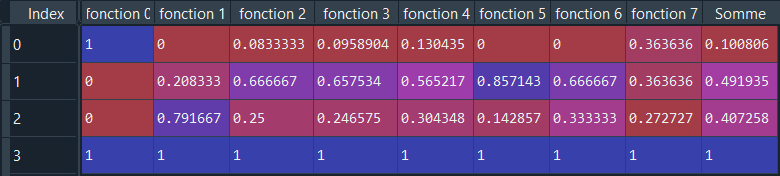
L'intensité de liaison est très élevée, on en conclut qu'on a dépendance.

En plus, la statistique du est d'autant plus grande ce qui veut dire que nos fonctions sont très dépendantes.

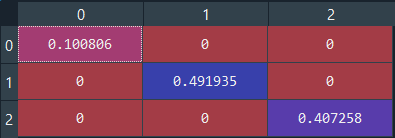
Ensuite, on construit les profils lignes, les profils colonnes ainsi que les matrices et des marges en lignes et en colonnes :

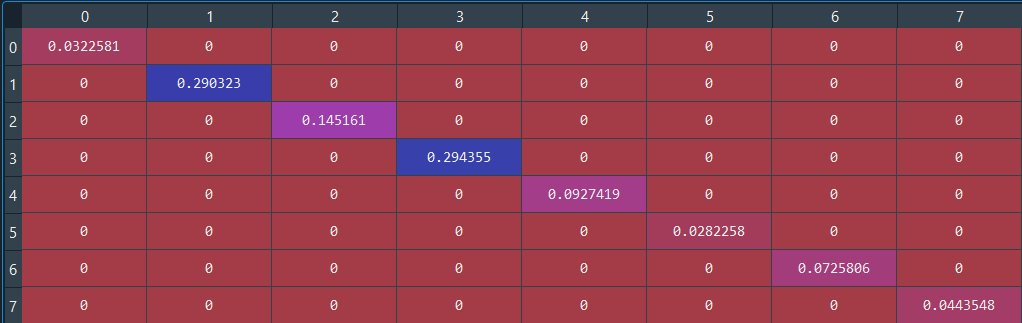


-Profile lignes-

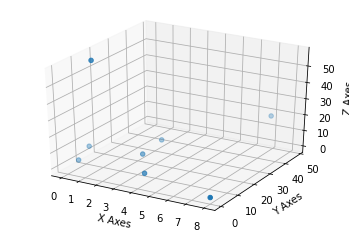


-Profile colonnes-

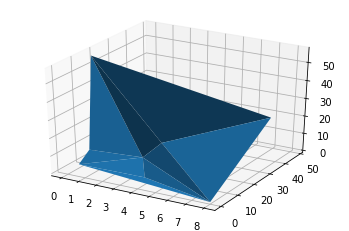
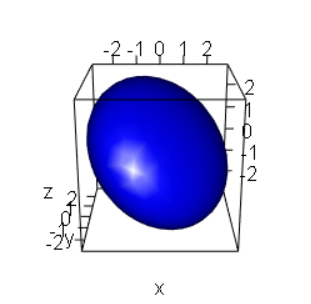




On visualise maintenant le nuage des points dans l’espace « Sexe » avant de procéder à la réduction de dimension dans :



Quand on approche la surface recouvrant le nuage de points avec une ellipse, on voit que c’est une ellipse bien définie et plate.

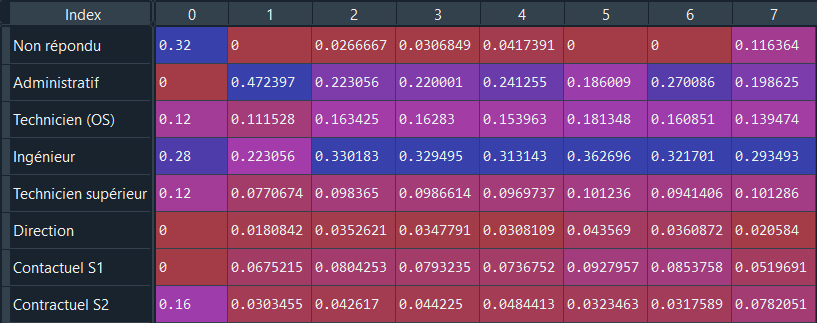
La partie d’approximation de la triangulation par une ellipse a été faite sous R, on n’est pas parvenu à bien la coder sous Python.

**Conclusion :** D’après la matrice R des écarts entre les données de la matrice de contingence et la matrice de contingence sous hypothèse d'indépendance, le test du et l’approximation du nuage de points par une ellipse, on conclut que la réduction de dimension se fera sans perte d’informations.

* **Faire la réduction et la projection du nuage des points lignes dans l’espace afin d’identifier les hyperplans pour lesquels l’inertie projetée est maximale : les hyperplans sont associés aux *p* directions de l’espace, chaque axe factoriel est de vecteur propre 𝑢 (respectivement matrice U) et une valeur propre (respectivement matrice Λ).**

Il s’agit maintenant d’effectuer une AFC dans l’espace . Pour ce faire, commençons d’abord par construire la matrice S qu’on va diagonaliser. Elle a une forme différente de la matrice à diagonaliser vue dans l’ACP qui n’est autre que la matrice de variance-covariance du tableau de contingence. La différence revient ici à un usage d’une métrique différente qui est non pas la métrique euclidienne mais la métrique du .

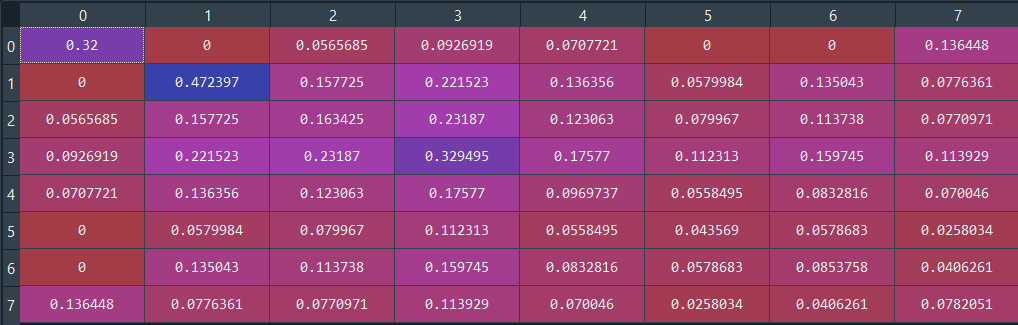
Si l’on chercher à réduire les distances entre les points des profils lignes de la matrice de contingence, la matrice à diagonaliser est :***S =***



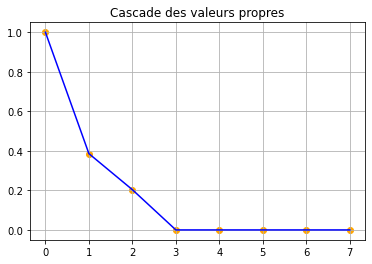
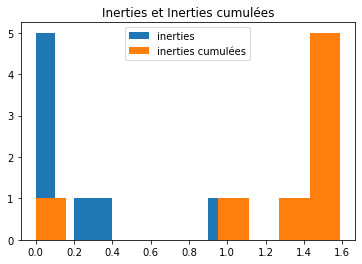
On voit bien que cette matrice est réelle mais pas symétrique donc non diagonalisable. On va utiliser, par conséquent, le complément de cours pour la transformer en matrice diagonalisable :

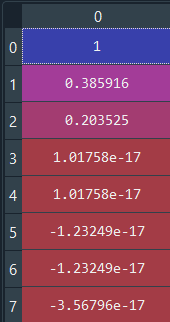
>A\_chapeau = np.transpose(mat\_frequence).dot(np.linalg.inv(Dn).dot(mat\_frequence))

>A = np.sqrt(np.linalg.inv(Dp)).dot(A\_chapeau.dot(np.sqrt(np.linalg.inv(Dp))))



Cette fois-ci, la matrice est bien symétrique. En plus, elle a les mêmes valeurs propres de S, et une simple relation linéaire nous permet de remonter aux vecteurs propres de S.

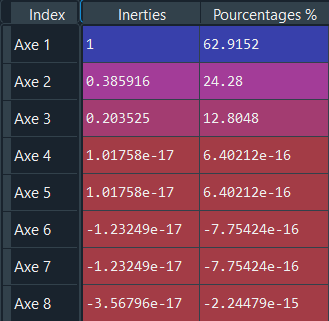


D’après le graphe de la cascade des valeurs propres ainsi que le tableau des valeurs propres, on choisit les deux premières valeurs propres qui sont plus prépondérantes que le reste. Ensuite, on calcule les nouvelles coordonnées des individus dans le nouveau plan factoriel et on représente le tout.

Une image contenant table

Description générée automatiquement

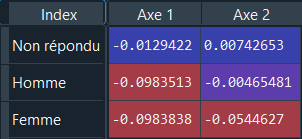
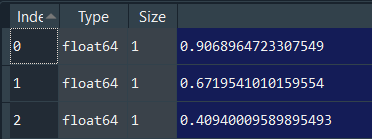
Pour consolider davantage nos choix, on construit une matrice qui montre l’inertie et le pourcentage de variance pour chaque axe :



La colonne de droite représente l’inertie et celle de gauche le pourcentage de variance pour chaque axe. Et on voit bien que les deux premières inerties expliquent le mieux nos données.

1. **Pour ce qui est de la qualité de l’AFC et de la qualité de projections des points, la similitude entre ACP et AFC vous permet de réutiliser les développements fait dans le TP précédent.**

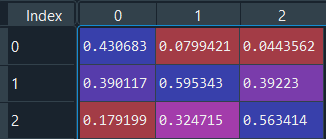
Ensuite, on reprend le tableau des nouvelles coordonnées ainsi que la qualité de projection dont la formule a été donnée dans le premier TP de l’ACP.

On en conclut que la qualité de projection est moyennement bonne car le nombre d’individus est plutôt faible (3).

1. **Vous pouvez envisager aussi en parallèle le cas de l’AFC sur l’espace de .**

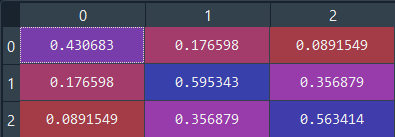
On reprend la même méthode et les mêmes calculs que dans l’AFC sur , la seule différence étant dans la matrice à diagonaliser : ***T =***



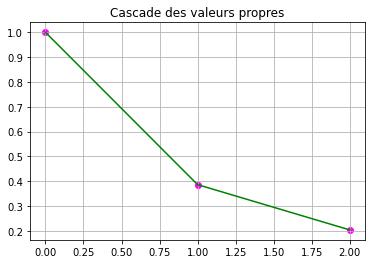
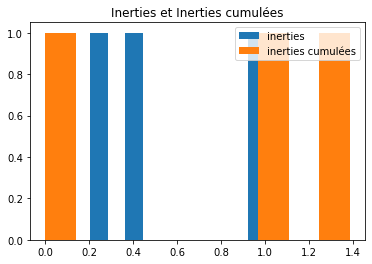
Encore une fois, on voit que cette matrice est réelle mais pas symétrique. On reprend la même approche que tout à l’heure, quitte à invertir les rôles de F et F’ d’une part, et et d’autre part.

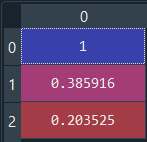
>M\_chapeau = mat\_frequence.dot(np.linalg.inv(Dp).dot(np.transpose(mat\_frequence)))

>M =np.sqrt(np.linalg.inv(Dn)).dot(M\_chapeau.dot(np.sqrt(np.linalg.inv(Dn))))



On peut maintenant diagonaliser et choisir les axes factoriels.

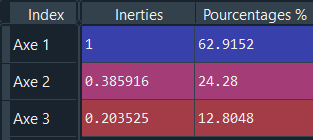


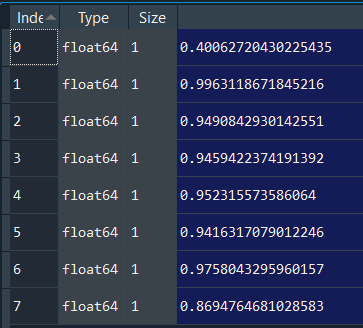
On choisit les deux premières valeurs propres et on trace ensuite le nuage des fonctions dans le nouveau plan factoriel.

Une image contenant table

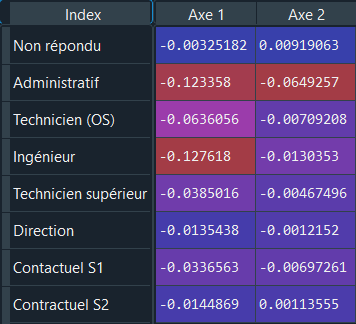
Description générée automatiquement

Matrice des inerties et pourcentage de variance pour chaque axe :





La qualité de projection est très élevée pour toutes les fonctions sauf la première qui correspond à « Non répondu ». On peut conclure que l’AFC dans l’espace se fait sans perte d’information.



Matrice des nouvelles coordonnées dans l’espace factoriel réduit de à

1. **Plot simultané des deux nuages :**

Cette représentation simultanée des genres et des fonctions nous permettra, entre autres, de savoir quel genre a tendance à avoir quelle fonction dans l’entreprise.

Une image contenant table

Description générée automatiquement

D’après la représentation graphique simultanée, on peut conclure que les hommes ont plus tendance à occuper les fonctions d’ingénieur ou technicien/technicien supérieur dans l’entreprise et que les femmes ont plus tendance à occuper les postes administratifs de l’entreprise.

1. **Comparer avec la CA (usage : res.ca <- CA(don[,1:3])) de FactoMineR**

Cette partie du TP sera codé sous R. On a essayé de trouver l’équivalent de la fonction CA dans une des librairies de Python, mais les seules qu’on a trouvé sont FactorAnalysis de la librairie sklearn.decomposition et FactorAnalyzer de la librairie factor\_analyzer. Mais ces deux fonctions ne donnaient malheureusement pas les résultats voulus.

On commence d’abord par l’installation est l’import des paquets nécessaires.

>install.packages(c("FactoMineR","factoextra"))

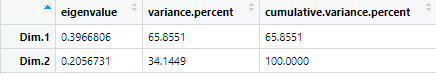
>library("FactoMineR")

>library("factoextra")

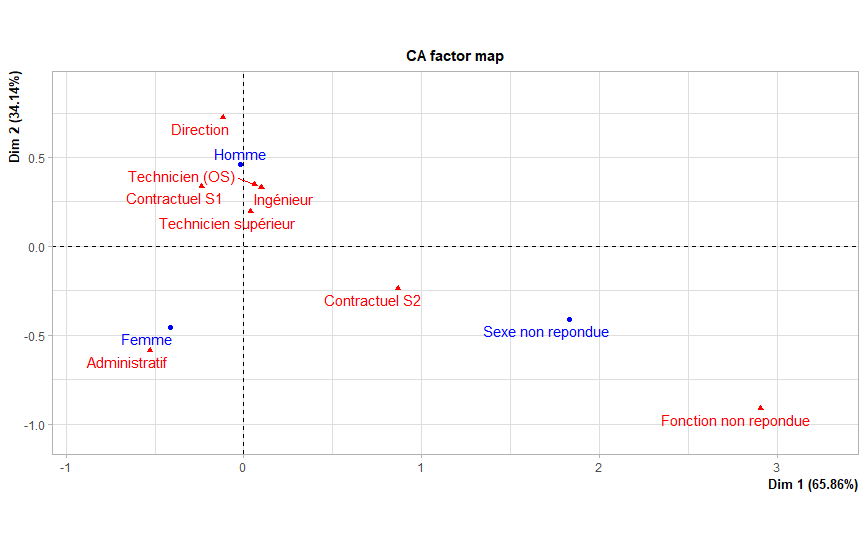
On construit la matrice de contingence et on la passe en argument à la fonction CA().

>res.ca <- CA(mat\_contingence)

>eig.val <- get\_eigenvalue(res.ca)



On obtient les mêmes résultats que ceux par AFC codé à la main au niveau des valeurs propres et des pourcentages de variance expliquée, toutefois, la fonction CA a choisi les valeurs propres : 0.38 et 0.20 au lieu de 1 et 0.38.



De même, on obtient la même allure de graph et on parvient aux mêmes conclusions. Les hommes dans cette entreprise ont plus tendance à être ingénieurs ou techniciens, et les femmes dans des postes administratifs. Cependant, on n’a pas tout à fait les mêmes coordonnées pour les points et cela reviendrait peut-être au fait que CA choisit les valeurs propres 0.38 et 0.20 au lieu de 1 et 0.38 pour la construction des axes factoriels.

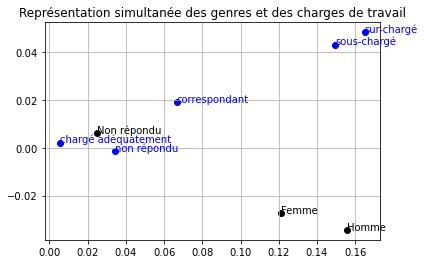
1. **La mise en œuvre se fera sur les données d’enquête réalisée sur lors d’un sondage dans une entreprise sur la perception du travail réalisé : fichier TP\_AFCmajeur1718travail.xls**

**Cette enquête avait pour cadre la pénibilité du travail dans une entreprise : vous disposez de deux questions à choisir pour la mise en œuvre de l’AFC :**

**1) En fonction du type de Fonction exercée et du sexe de l’individu, qui concerne l’adéquation entre temps de travail et travail à réaliser**

**2) En fonction du type de Fonction exercée et du sexe de l’individu, qui concerne la qualité de vie au travail exprimée.**

Pour conclure ce rapport, on a choisi de travailler sur l’adéquation entre temps de travail et travail à réaliser en fonction du sexe et de la fonction dans l’entreprise. Il suffit de reprendre les codes précédents et les réadapter aux variables choisies :



**Conclusion 1 :** Ni les hommes ni les femmes ne sont surchargées dans leurs fonctions.

Une image contenant table

Description générée automatiquement

**Conclusion 2** : les fonctions administratives et direction, les ingénieurs et les techniciens OS sont des fonctions largement réalisables dans le temps accordé par l’entreprise. Les fonctions contractuelles S1 et de technicien supérieur quant à elles correspondent parfaitement au temps fourni par l’entreprise.

***Conclusion Finale :***

Que l’analyse soit faite suivant la fonction occupée dans l’entreprise ou suivant le genre de l’employé, on conclut que l’entreprise fournit à ses employés un temps suffisant pour la réalisation de leurs tâches.