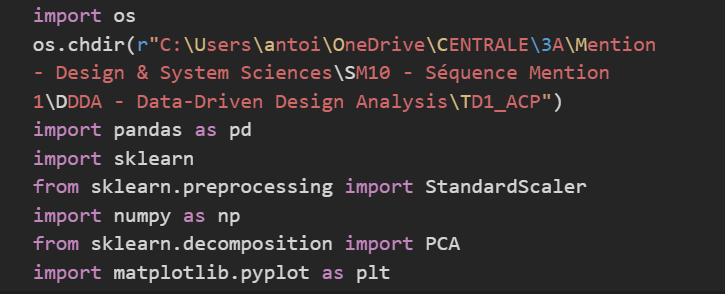
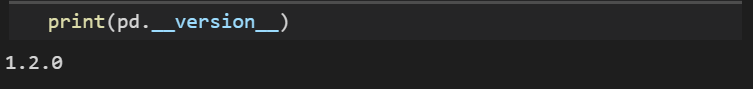
TD1 : Data-Driven Design Analysis

ANALYSE EN COMPOSANTES PRINCIPALES

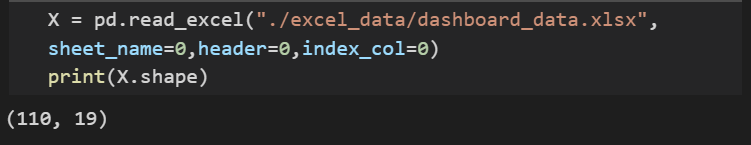
Pour commencer, nous importons toutes les fonctions et librairies qui seront nécessaires à l’exécution du code par la suite.



Nous vérifions la version de Pandas

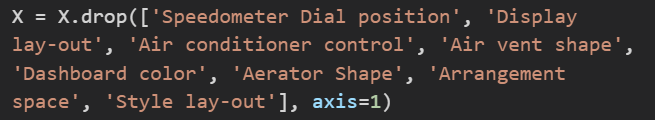


Il faut tout d’abord importer la donnée, avec la fonction **read\_excel** de pandas



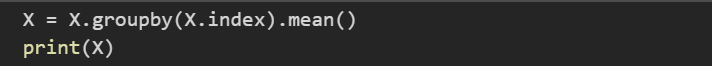
Nous obtenons un Dataframe de 19 colonnes et 110 lignes. Affichons les 5 premières lignes du Dataframe avec la méthode **.head()**

Nous notons que nous devons supprimer dans un premier temps les variables illustratives pour construire notre modèle. Nous utilisons la méthode .drop() qui permet de renseigner le nom des colonnes à supprimer du Dataframe.



Nous avons une quantité de donnée importante, avec plusieurs index identiques un certain nombre de fois. Nous travaillons sur des individus voiture. Il y a 10 types de voiture différentes, avec une occurrence de 11 fois chacun. Il s’agit de réduire le nombre de lignes de notre Dataframe (pour la visualisation future des individus dans le plan factoriel, qui risque de ne plus être lisible avec 110 occurrences distinctes).

Nous allons donc moyenner les variables actives sur chacun des individus distincts.

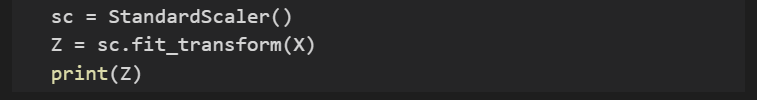


Nous entrons le nombre de variables et d’observations dans les variables p et n (que nous réutiliserons par la suite)

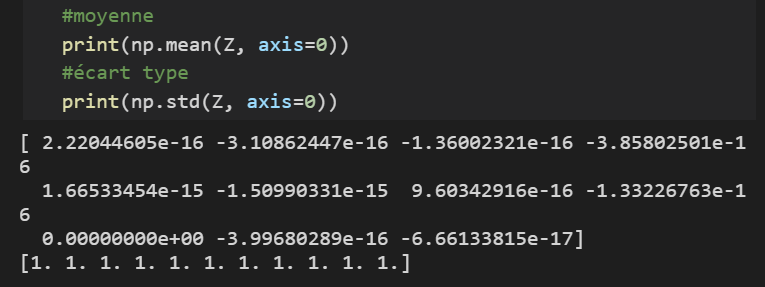


Pour réaliser notre ACP normée, nous devons avoir des variables centrées réduites (moyenne 0, variance 1).

Nous utilisons la classe StandardScaler, et la méthode **fit\_transform**

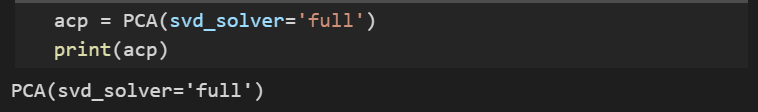


Nous pouvons vérifier que notre nouvelle donnée Z est bien centrée réduite



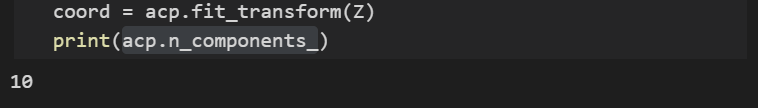
Notons que dans l’output de la moyenne, nous n’obtenons pas exactement 0 mais des valeurs de l’ordre de , qui sont dues à des erreurs négligeables de calcul.

Nous créons ensuite l’objet à partir de la classe PCA



La fonction **fit\_transform()** renvoie en sortie les coordonnées factorielles que nous collectons dans la variable coord. Nous affichons le nombre de composantes générées, il n’est pas égal à 11 mais à 10.

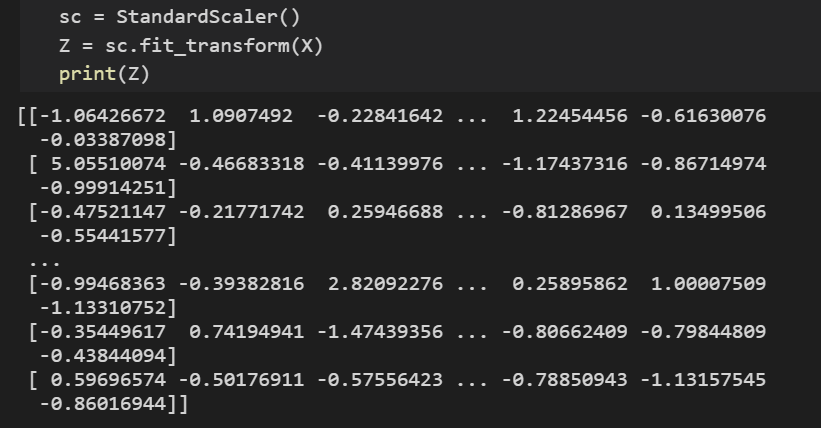
En effet, nous avons 10 individus et 11 colonnes, or le nombre de composantes principales ne peut pas excéder 10.



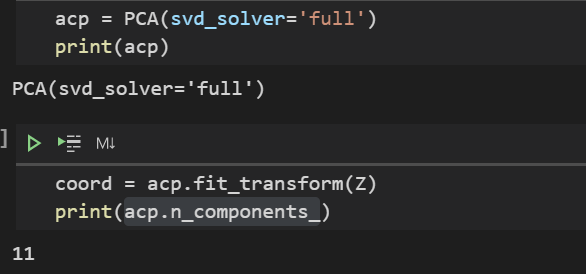
Nous devons donc réaliser l’ACP sur l’ensemble des individus, et effectuer notre moyenne après pour la visualisation

Reprenons depuis le début.

Nous redéfinissons notre donnée Z centrée réduite

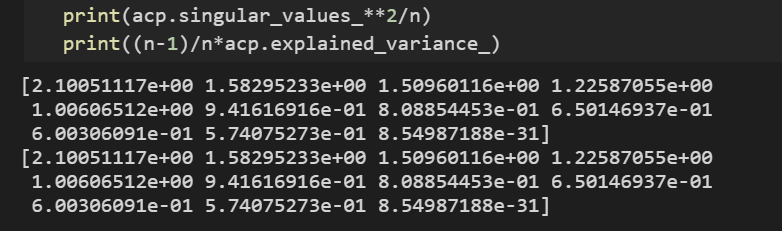


Nous réinitialisons notre objet acp

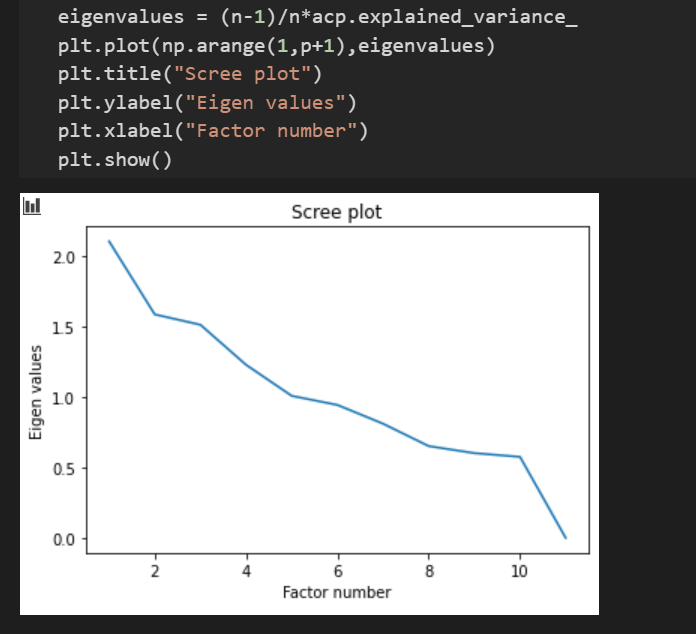


Cette fois-ci, le nombre de composantes est bien égal à 11.

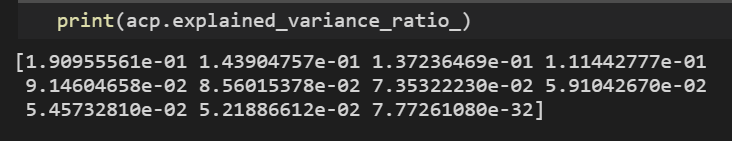
L’attribut **.explained\_variance\_** pondéré par un coefficient nous donne les variances associées aux axes factoriels, que nous pouvons retrouver en passant par les valeurs singulières via l’attribut **.singular\_values\_**



Voici les valeurs propres affichées en fonction des axes :



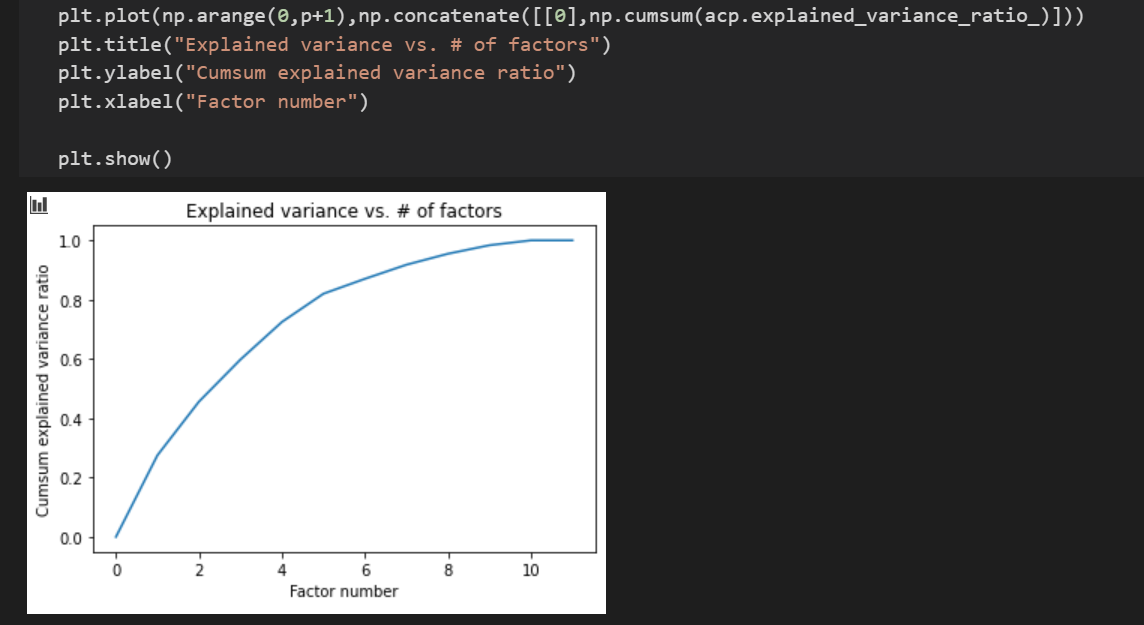
Afin d’obtenir les proportions de variance associées aux axes, nous utilisons l’attribut **.explained\_variance\_ratio\_.**



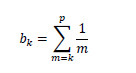
|  |  |
| --- | --- |
| Axe | Pourcentage de variance |
| 1 | 19.1% |
| 2 | 14.4% |
| 3 | 13.7% |
| 4 | 11.1% |
| 5 | 9.15% |
| 6 | 8.56% |

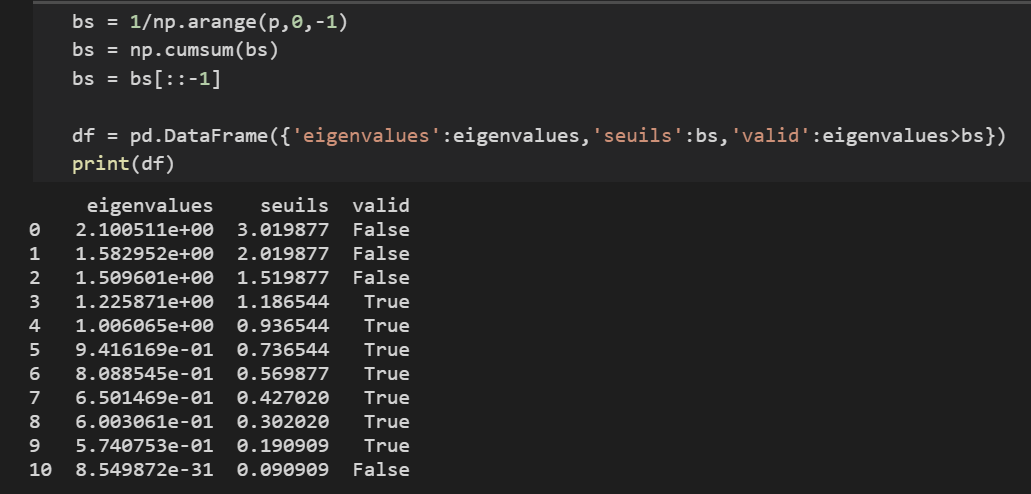
Nous notons que l’information est répartie assez équitablement sur les premières variables.

Afin de visualiser la quantité de variance expliquée en fonction du nombre de facteurs, nous pouvons réaliser ceci en cumulant la variance sur les axes factoriels (voir ci-après).



Afin de déterminer le nombre de variables à retenir, nous pouvons utiliser le test des bâtons brisés qui consiste à vérifier si les valeurs propres sont supérieures à un seuil défini par la formule suivante :

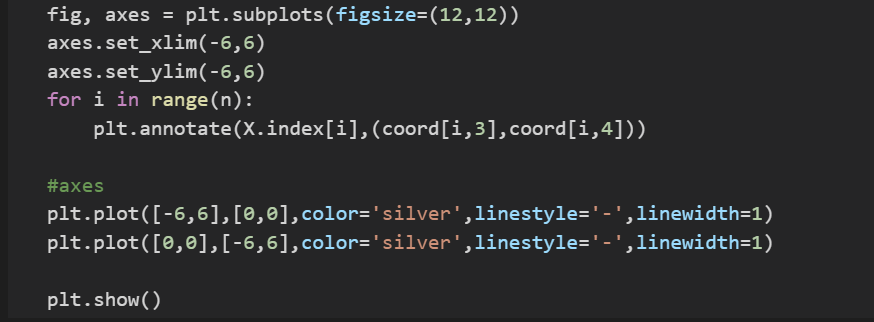


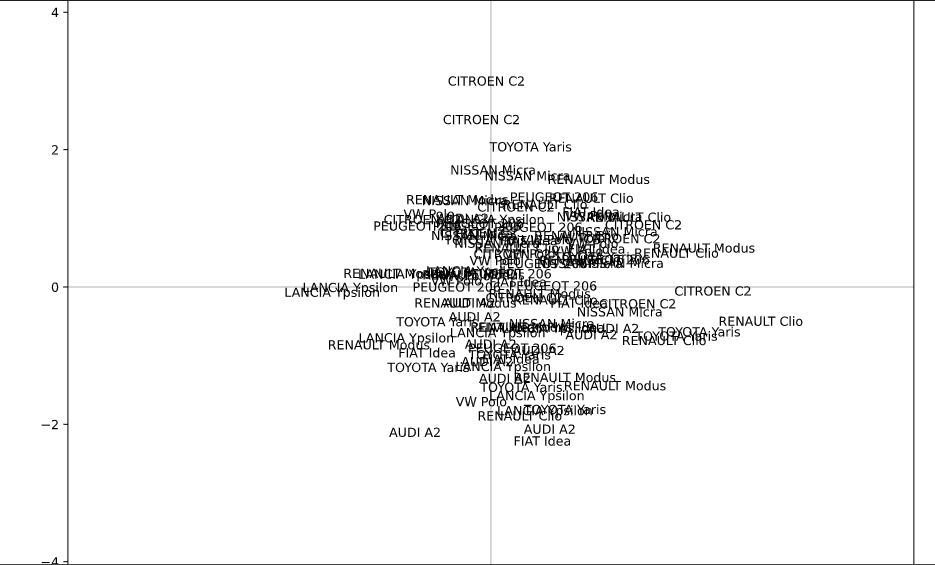


Nous voyons que 6 facteurs sur 10 sont valides, à savoir du 4e jusqu’au 9e

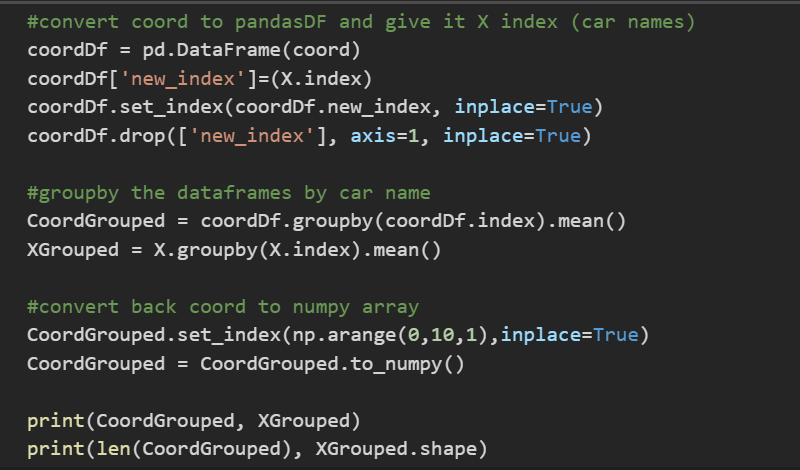
Il s’agit de représenter les individus dans un plan :

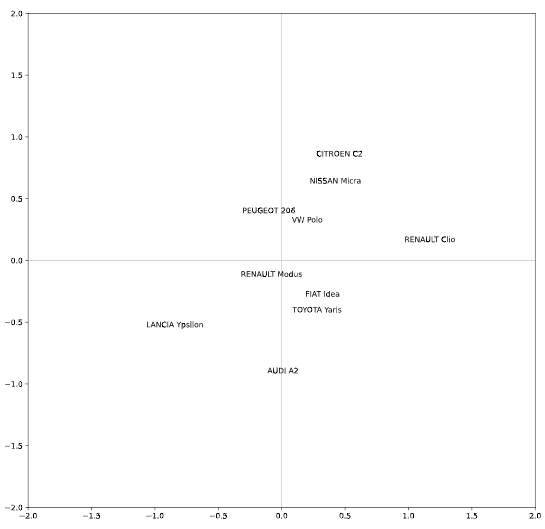
Nous pouvons par exemple choisir le plan (3 ; 4), valide selon le critère des bâtons brisés.





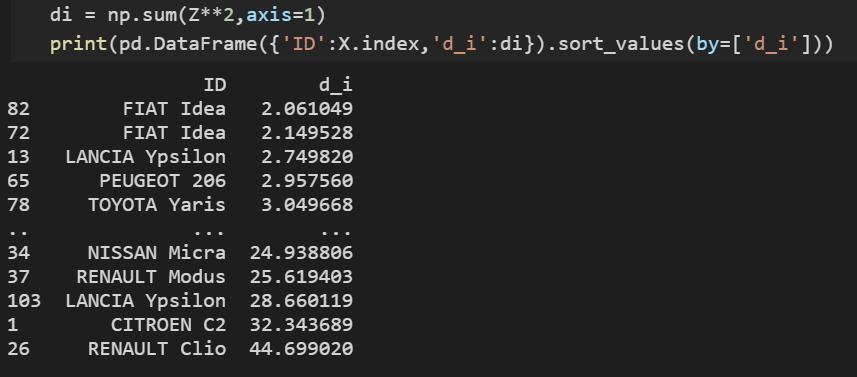
Nous pouvons grouper le tableau coord et X par voiture, pour avoir une seule occurrence de chaque modèle.



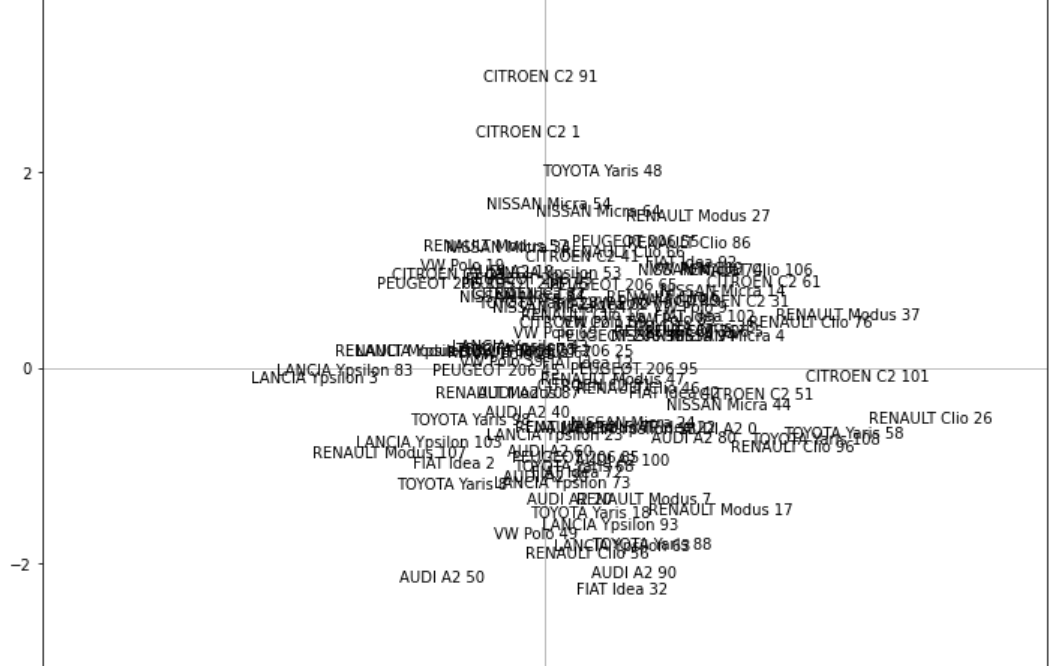


Nous obtenons la figure suivante dans le plan factoriel (3,4)

Analysons la qualité de la représentation en calculant les carrés des distances à l’origine des individus.

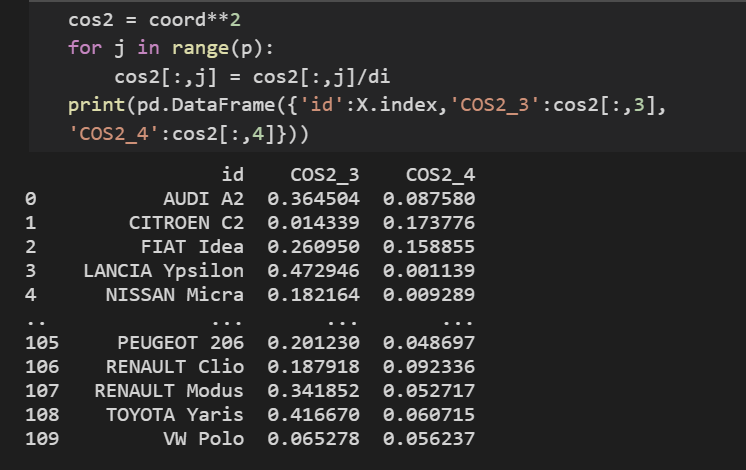


Les véhicules qui se démarquent sont identifiables dans la représentation graphique sur l’axe factoriel (3,4). Nous avons ajouté le numéro d’index pour distinguer les modèles de voiture identiques.

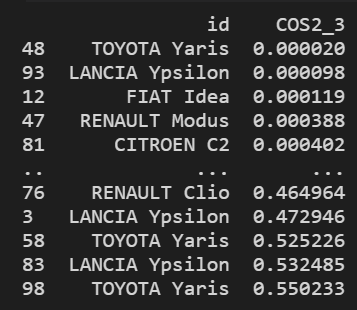


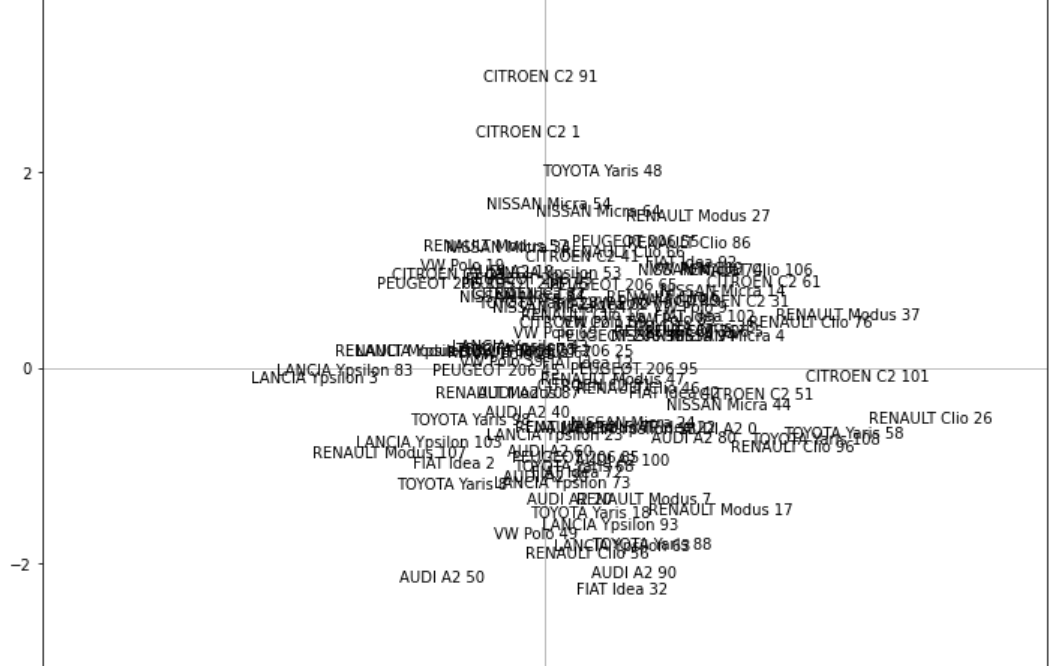
Nous identifions notamment la Renault Clio 26, au bout de l’axe factoriel 3.

Regardons à présent la qualité de représentation des individus en calculant les cosinus carrés des individus sur les différents axes.

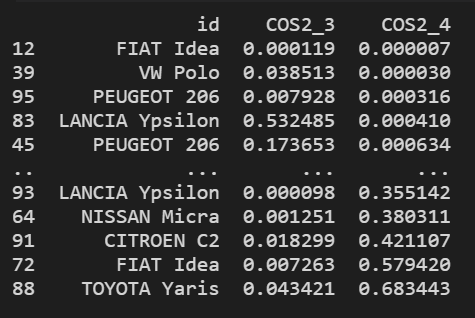
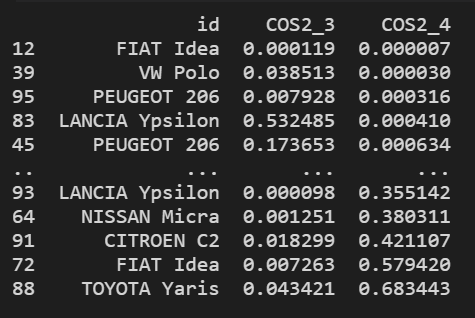


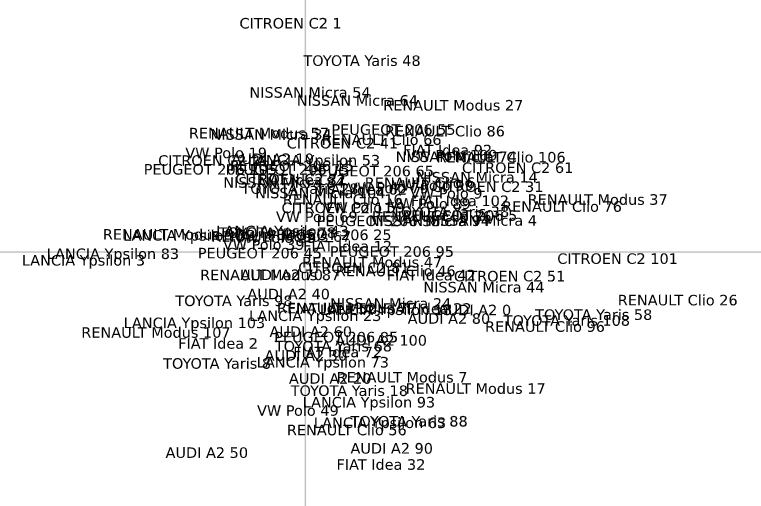
Si nous trions sur l’axe 3 :



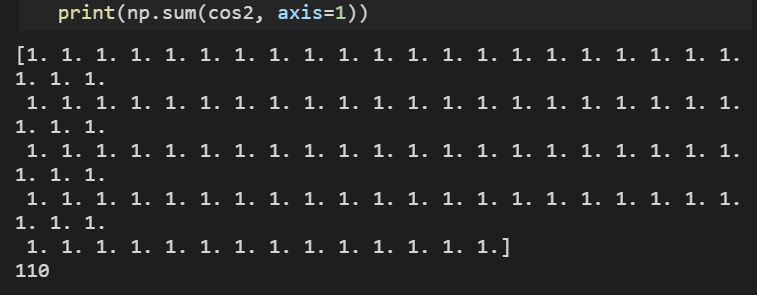


Si nous trions sur l’axe 4 :

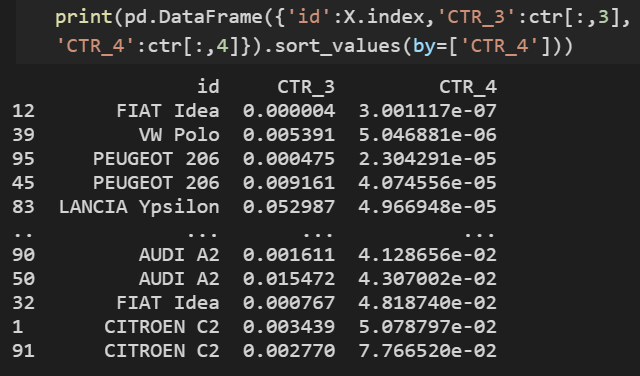
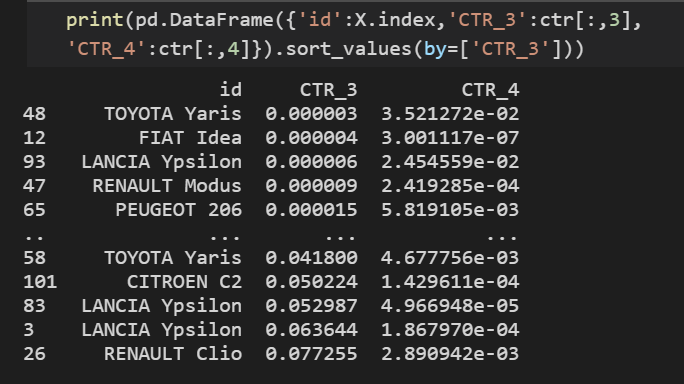




On vérifie que la somme des COS² est bien égale à 1 pour chaque individu

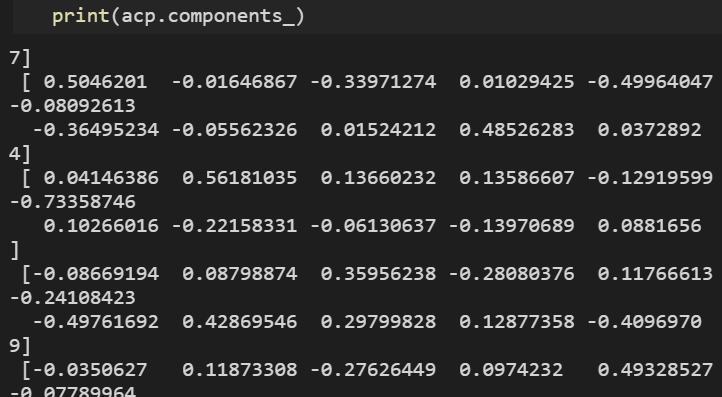


On regarde maintenant quels individus participent le plus à chaque facteur de l’ACP

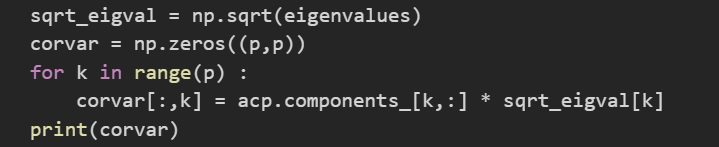


On retrouve bien les individus aux extrémités sur la représentation du plan factoriel (3,4)

Pour analyser les variables nous récupérons les vecteurs propres (axes factoriels) grâce à l’attribut **.components\_**

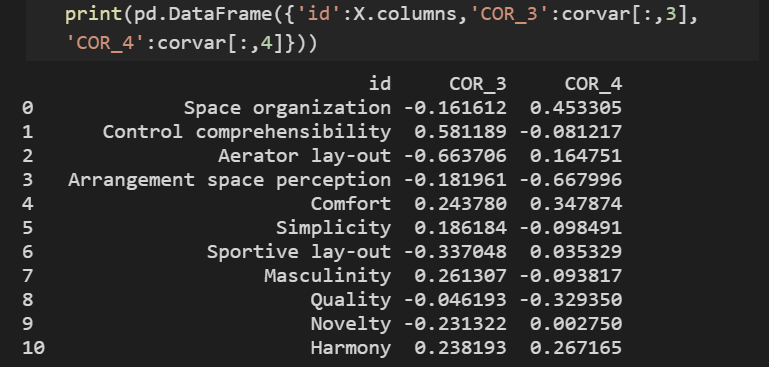


Nous créons une matrice de corrélation entre variables et facteurs (variables en ligne, facteurs en colonne) :

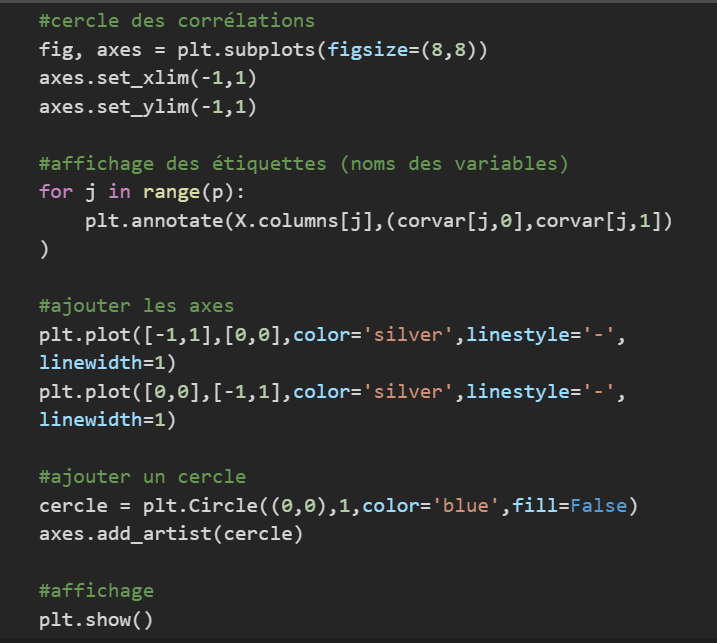


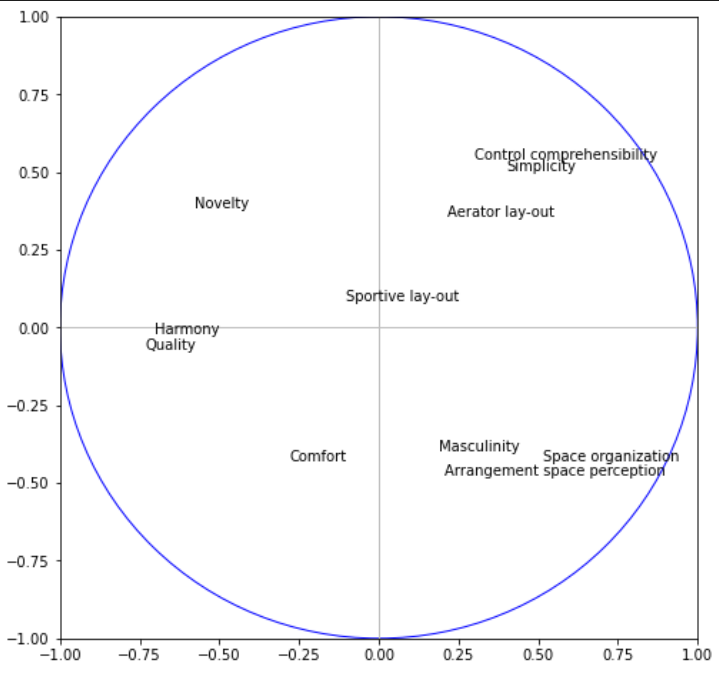
On obtient une matrice 11x11.

SI on regarde uniquement les facteurs 3 et 4 (comme précédemment) :



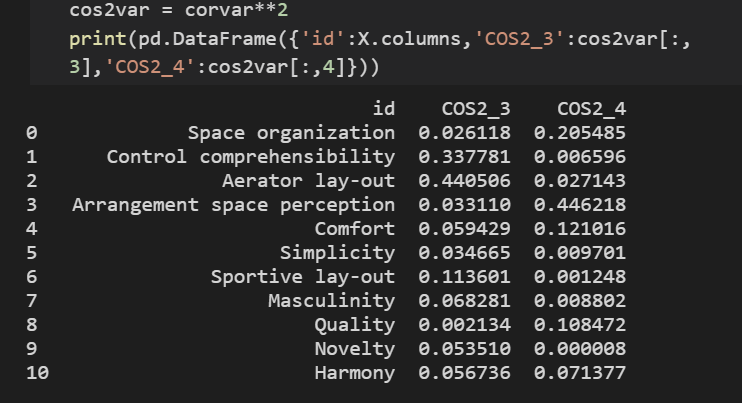
On peut afficher toutes nos variables dans un cercle de corrélation :

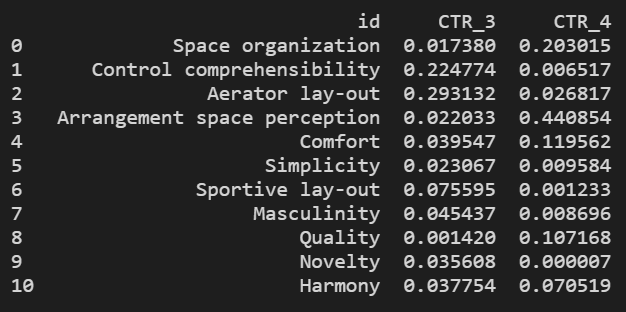




On remarque des corrélations fortes entre les variables telles Space organization / space perception par exemple.

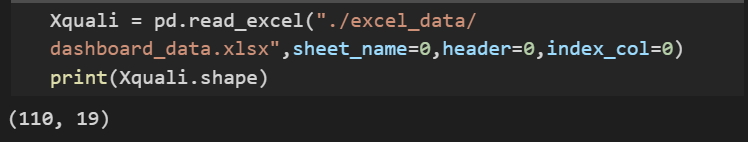
De même que précédemment, on peut regarder la qualité de représentation des variables sur les axes via les cosinus carrés puis les contributions des variables aux axes.



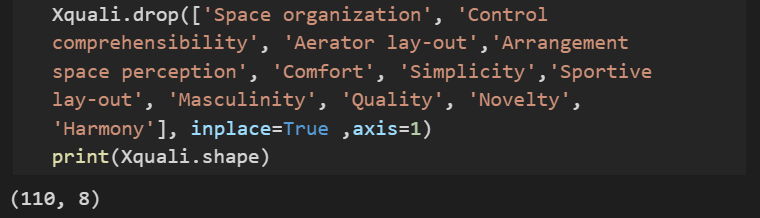


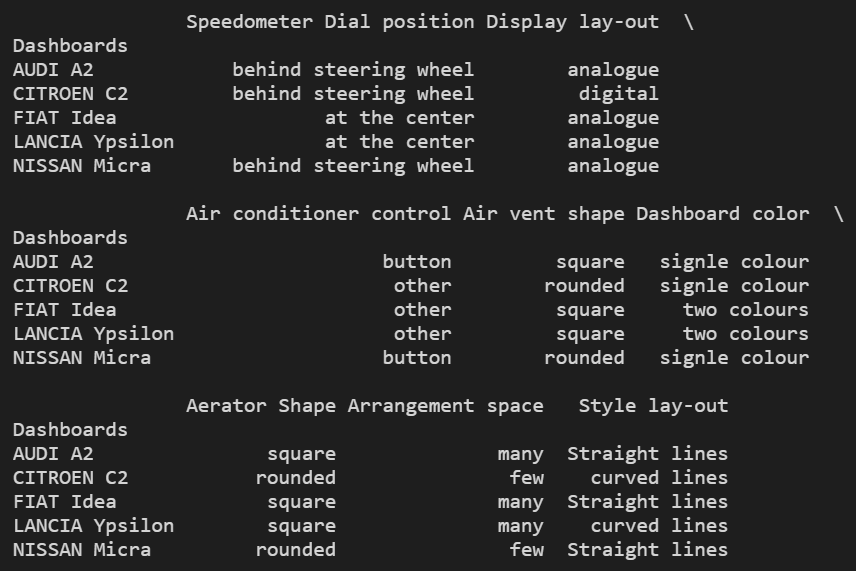
On remarque que les contributions principales sur l’axe 3 sont Control comprehensibility et Aerator lay-out. Sur l’axe 4, c’est arrangement space perception.

Créons un nouveau Dataframe pour les variables qualitatives.



Nous supprimons les variables quantitatives



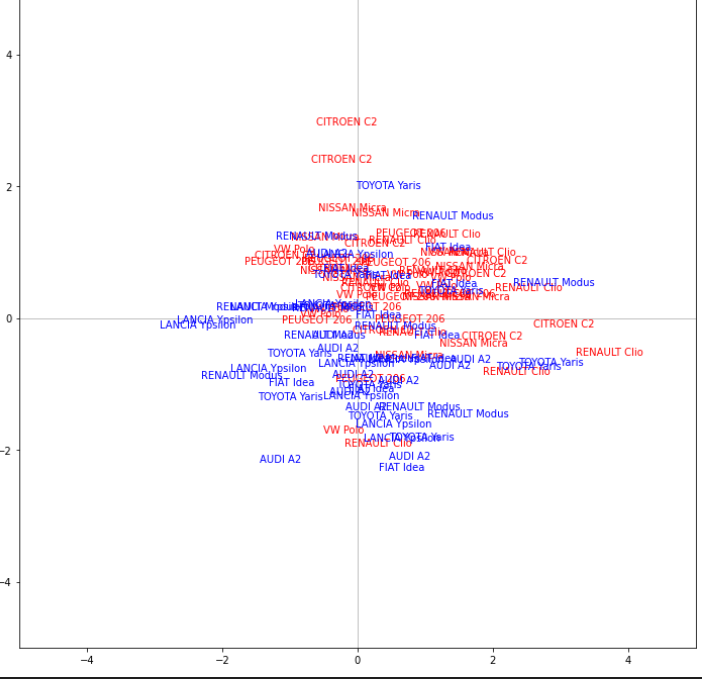


Nous pouvons sélectionner une variable illustrative et la représenter dans le plan factoriel (3,4)



Cette variable prend deux valeurs distinctes à savoir « many » ou « few ». Nous pouvons afficher nos individus et voir à quelle feature appartient chaque individu





Nous pouvons également afficher les barycentres de chacune des classes (few et many), ce qui peut donner lieu à des interprétations multiples.

