# [Réalisez une analyse exploratoire de données](https://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees)

Table des matières

[Réalisez une analyse exploratoire de données 1](#_Toc57704744)

[Êtes-vous prêt à suivre ce cours ? 3](#_Toc57704745)

[Prérequis 3](#_Toc57704746)

[Quelle est la forme des données que nous allons utiliser ? 3](#_Toc57704747)

[Et si j'ai besoin d'aide ? 4](#_Toc57704748)

[Découvrez l'intérêt de l’analyse multidimensionnelle 5](#_Toc57704749)

[Statistiques exploratoires multidimensionnelles 5](#_Toc57704750)

[Analyse exploratoire multidimensionnelle 5](#_Toc57704751)

[Quel est l'intérêt d'une étude multidimensionnelle ? 6](#_Toc57704752)

[Découvrez les méthodes factorielles et la classification non supervisée 9](#_Toc57704753)

[Supervisé ou non supervisé ? Telle est la question. 10](#_Toc57704754)

[Téléchargez les jeux de données analysés dans ce cours 12](#_Toc57704755)

[Échantillon n°1 : Les cours OpenClassrooms que vous avez suivis 12](#_Toc57704756)

[Échantillon n°2 : Le texte de cours OpenClassrooms 13](#_Toc57704757)

[Échantillon n°3 : surprise ! 15](#_Toc57704758)

[Représentez vos données dans un espace 17](#_Toc57704759)

[La notion d'espace euclidien 17](#_Toc57704760)

[La notion de distance 19](#_Toc57704761)

[La notion de nuage de points 20](#_Toc57704762)

[La notion d’inertie 20](#_Toc57704763)

[Avez-vous compris l'intérêt de l'analyse exploratoire multidimensionnelle ? 22](#_Toc57704764)

[Compétences évaluées 22](#_Toc57704765)

[ Question 1 22](#_Toc57704766)

[ Question 2 22](#_Toc57704767)

[ Question 3 23](#_Toc57704768)

[ Question 4 25](#_Toc57704769)

[ Question 5 25](#_Toc57704770)

[ Question 6 25](#_Toc57704771)

[ Question 7 26](#_Toc57704772)

[Comprenez l'enjeu de l'Analyse en Composantes Principales 27](#_Toc57704773)

[Qu'allons-nous voir dans cette 2e partie ? 27](#_Toc57704774)

[L'enjeu de l'ACP 27](#_Toc57704775)

[Le principe de l'ACP 30](#_Toc57704776)

[Centrer et réduire 33](#_Toc57704777)

[Découvrez les espaces que nous utiliserons 37](#_Toc57704778)

[Répondre au premier objectif par le nuage des individus **NI** : 37](#_Toc57704779)

[Répondre au second objectif par le nuage des variables **NK** : 37](#_Toc57704780)

[Un résultat remarquable 41](#_Toc57704781)

[En résumé : 41](#_Toc57704782)

[Aller plus loin : Calcul des composantes principales du nuage **NI**, valeurs propres et vecteurs propres 42](#_Toc57704783)

[Interprétez le cercle des corrélations 44](#_Toc57704784)

[Le cercle des corrélations 44](#_Toc57704785)

[Analyse de votre jeu de données 46](#_Toc57704786)

[Attention ! 50](#_Toc57704787)

[Variables et angles : qualité de la représentation 50](#_Toc57704788)

[Aller plus loin : Que dire des angles entre les variables initiales ? 51](#_Toc57704789)

[Aller plus loin : Pourquoi un cercle ? 52](#_Toc57704790)

[Représentez les individus sur les plans factoriels 54](#_Toc57704791)

[Qualité de la représentation et contributions 56](#_Toc57704792)

[Aller plus loin : Contribution d’un individu à l’inertie d’un axe 57](#_Toc57704793)

[Aller plus loin : Qualité de représentation d’un individu par un axe 58](#_Toc57704794)

[Choisissez le nombre de composantes 60](#_Toc57704795)

[Réfléchissons 60](#_Toc57704796)

[Combien de composantes analyser ? 61](#_Toc57704797)

[TP : Réalisez une ACP 66](#_Toc57704798)

[Les cours OpenClassrooms que vous avez suivis 66](#_Toc57704799)

[Le jeu de données mystère 67](#_Toc57704800)

[L'échantillon en bag of words 70](#_Toc57704801)

[Annexe : les fonctions outils 76](#_Toc57704802)

[Soyez attentif aux spécificités de l'ACP 81](#_Toc57704803)

[Points d'attention 81](#_Toc57704804)

[Limites de l’ACP 81](#_Toc57704805)

[Pratiquez l'ACP 83](#_Toc57704806)

[Compétences évaluées 83](#_Toc57704807)

[ Question 1 83](#_Toc57704808)

[ Question 2 85](#_Toc57704809)

[ Question 3 86](#_Toc57704810)

[ Question 4 87](#_Toc57704811)

[ Question 5 87](#_Toc57704812)

[ Question 6 87](#_Toc57704813)

[ Question 7 88](#_Toc57704814)

[Recherchez une bonne partition 91](#_Toc57704815)

[Évaluer la qualité d'une partition 92](#_Toc57704816)

[Découvrez l’algorithme k-means 95](#_Toc57704817)

[Fonctionnement de l'algorithme 95](#_Toc57704818)

[Effectuez une classification hiérarchique 98](#_Toc57704819)

[Classification hiérarchique ascendante et descendante 99](#_Toc57704820)

[Méthode des liens et méthode de Ward 100](#_Toc57704821)

[Avantages et inconvénients 101](#_Toc57704822)

[Interprétez votre partition 102](#_Toc57704823)

[TP : Partitionnez vos données 104](#_Toc57704824)

[L'échantillon en bag of words 104](#_Toc57704825)

[Le chat 108](#_Toc57704826)

[Annexe : les fonctions outils 111](#_Toc57704827)

[Entraînez-vous à réaliser un détecteur de slides grâce à de l'analyse d'image 113](#_Toc57704828)

[À vous de jouer 113](#_Toc57704829)

[Vérifiez votre travail 116](#_Toc57704830)

Vous avez un important volume de données ? Il est important de savoir les synthétiser !

Dans ce cours, vous apprendrez à effectuer une **analyse exploratoire multidimensionnelle.**Nous utiliserons des méthodes populaires pour**analyser rapidement votre échantillon** en réduisant la dimension du nombre d'individus ou de variables.

Nous aborderons des méthodes emblématiques comme l'**Analyse en Composantes Principales**ou encore le fameux ***clustering :***

L'**Analyse en Composantes Principales** (ACP ou **PCA** en anglais) permet de dégager rapidement les principales tendances de votre échantillon, en diminuant le nombre de variables nécessaires à la représentation de vos données tout en perdant le moins d'informations possible.

Nous aborderons également les deux méthodes de clustering les plus populaires : l'algorithme du **k-means** et la **classification hiérarchique**. Celles-ci permettent de regrouper vos individus selon leurs similarités.

À la fin de ce cours, vous aurez ajouté à votre boîte à outils les méthodes classiques de tout bon Data Analyst !

**Objectifs pédagogiques :**

* Appréhender les espaces vectoriels euclidiens
* Réaliser une réduction de dimension
* Interpréter une Analyse en Composantes Principales (ACP)
* Réaliser une ACP
* Utiliser un algorithme de clustering (Kmeans)
* Réaliser une classification hiérarchique

**Prérequis** :

* Maîtriser les statistiques descriptives uni et bidimensionnelles, notamment :
  + connaître le [**vocabulaire de base et savoir représenter un échantillon**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/decouvrez-les-statistiques-vocabulaire-et-tour-dhorizon-1) ;
  + les différents [**types de variables**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/decouvrez-les-4-types-de-variables) ;
  + la notion de [**distribution**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/representez-la-distribution-empirique-dune-variable) (et comment la représenter) ;
  + la notion de [**corrélation**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/recherchez-les-correlations) et plus précisément de [**corrélation linéaire**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/comment-presenter-les-donnees-en-analyse-bivariee).
* Être familier avec la notion de vecteur (écriture, représentation graphique), de droite, d'axe.
* Savoir utiliser le langage Python dans le cadre de la Data Science, ou le langage R.

Si vous avez besoin de vous rafraîchir la mémoire, voici quelques cours :

* [**Statistiques descriptives uni et bidimensionnelles**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees)
* [**Langage Python (initiation)**](https://openclassrooms.com/courses/demarrez-votre-projet-avec-python)
* [**Langage Python appliqué à la data science**](https://openclassrooms.com/courses/decouvrez-les-librairies-python-pour-la-data-science)

## Êtes-vous prêt à suivre ce cours ?

### 

### Prérequis

Ce cours fait partie du parcours [Data Analyst](https://openclassrooms.com/paths/data-analyst). C’est la suite du cours sur les statistiques descriptives intitulé [Nettoyez et décrivez votre jeu de données](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees). Il est vivement conseillé de le suivre ou de vérifier si vous connaissez bien les concepts qui y sont présentés avant de vous lancer.

### 

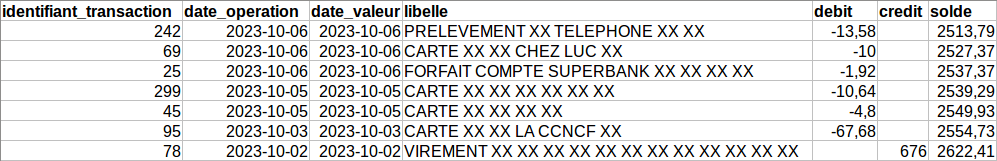
### Quelle est la forme des données que nous allons utiliser ?

Pour commencer, remettons-nous rapidement en tête le vocabulaire que nous avons vu dans [ce chapitre du cours de statistiques descriptives](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/decouvrez-les-statistiques-vocabulaire-et-tour-dhorizon-1) :

Nous souhaitons étudier une **population** composée d’**individus**. Ces individus ne sont pas forcément des personnes, mais peuvent être des objets, des animaux, des relevés effectués par des capteurs (relevés de température, par exemple), ou beaucoup d’autres choses !

Souvent, une population est difficile à étudier dans sa globalité (surtout quand elle contient un grand nombre d’individus et qu’il est impossible de tous les observer). Pour cela, on extrait un **échantillon** de la population, c’est-à-dire que l’on sélectionne certains individus pour les étudier précisément. Étudier un individu, c’est observer ses caractéristiques : chaque caractéristique est décrite par une **variable**.

Ainsi, il est possible de stocker nos observations dans un tableau dans lequel chaque ligne représente un individu, et chaque colonne représente une variable. Dans le cours précédent, par exemple, nous avons étudié des relevés bancaires : chaque individu (en ligne) est une opération bancaire, et chaque variable (en colonne) est une caractéristique de l’opération (comme sa date, son libellé, son montant, etc.) :



Une précision importante : nous n'utiliserons dans ce cours que des variables [**quantitatives**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/decouvrez-les-4-types-de-variables). L'analyse en composantes principales n'est, par nature, applicable qu'à de telles variables. Nous utiliserons donc très peu les variables qualitatives, sauf parfois pour interpréter les résultats finaux. L'équivalent de l'ACP pour des variables qualitatives s'appelle l'Analyse en Composantes Multiples, que nous n'aborderons pas ici.

### Et si j'ai besoin d'aide ?

Vous pouvez poser vos questions sur le [forum du cours](https://openclassrooms.com/forum/sujet/cours-analyse-exploratoire-de-donnees). Vous y trouverez d'autres étudiants capables d'y répondre.

## Découvrez l'intérêt de l’analyse multidimensionnelle

### 

### Statistiques exploratoires multidimensionnelles

Comme nous l’avons vu dans un cours [précédent](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/decouvrez-les-statistiques-vocabulaire-et-tour-dhorizon-1), le domaine des statistiques est vaste ! Alors où nous situons-nous donc dans ce cours d'analyse exploratoire ?

Le cours [Nettoyez et analysez votre jeu de données](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees) traitait déjà de **statistiques descriptives**. Comme leur nom l’indique, les statistiques descriptives ont pour vocation à décrire des données, c’est tout.

En statistiques descriptives, on étudie un échantillon, c’est-à-dire une proportion plus ou moins grande (souvent même assez petite) d’une population totale.

### Analyse exploratoire multidimensionnelle

Dans ce cours, nous ferons de l’**analyse exploratoire multidimensionnelle**.

Le terme d’analyse exploratoire vous évoque peut-être les explorateurs tels Indiana Jones ou Christophe Colomb partant explorer des territoires inconnus. Mais ici, « exploratoire» sera juste synonyme de « descriptif ».

Que signifie le terme **multidimensionnel** ?

Quand on fait une analyse unidimensionnelle, on étudie les variables une par une, séparément. C'est ce que vous avez vu dans les 2 premières parties de [ce cours](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees). Dans la 3e partie, nous avons étudié les relations entre 2 variables, grâce notamment au concept de corrélation. Ici, il s'agit d'analyse multidimensionnelle, mais on se limitait à 2 variables à la fois.

Ici, on ira donc plus loin : on étudiera les relations entre plus de 2 variables à la fois !

Et c’est ici qu’il faudra réveiller votre âme d’explorateur ! En effet, vous partirez à la découverte d'espaces inexplorés, des espaces que le cerveau de l’homme ne peut même pas appréhender, car ce sont des espaces qui possèdent souvent plus de 3 dimensions…



Eh oui ! Sur un papier ou sur un écran, c’est encore facile de faire un graphique à 2 dimensions : avec 2 axes, un horizontal et un vertical. En 3 dimensions, on ajoute la profondeur. Là, on peut encore s’arranger avec des écrans 3D, ou faire une maquette avec des balles de ping-pong pour représenter les points. Mais dès que l’on passe à 4 dimensions, le cerveau humain n’est plus capable de se représenter cet espace, car nous vivons tous dans un espace à 3 dimensions spatiales.

### 

### Quel est l'intérêt d'une étude multidimensionnelle ?

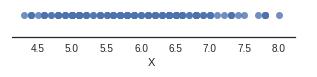
La réponse est simple : on voit beaucoup plus de choses quand on étudie les relations entre les variables que quand on étudie les variables séparément (une à une).

Prenons quelques exemples :

#### Premier exemple

Supposons un échantillon décrit par 2 variables quantitatives. Peu importe ce qu'elles représentent, nous les appellerons donc simplement x et y.

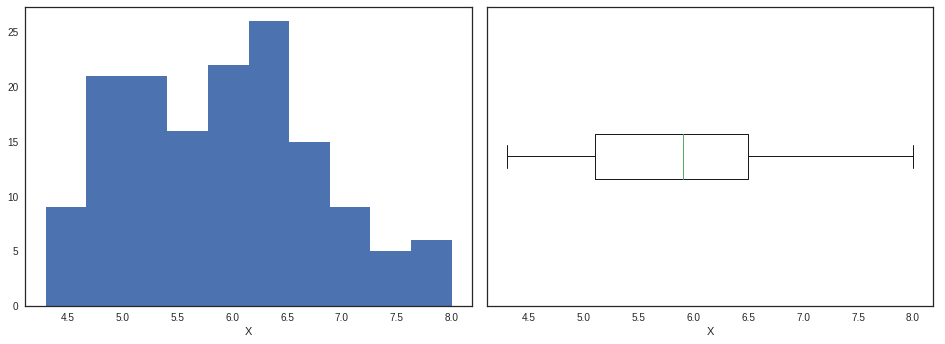
Si l'on commence par une analyse monodimensionnelle (ou univariée), il nous est possible de représenter chaque individu par un point le long d'un axe : c'est un graphique à 1 dimension. Voici celui de la variable x :



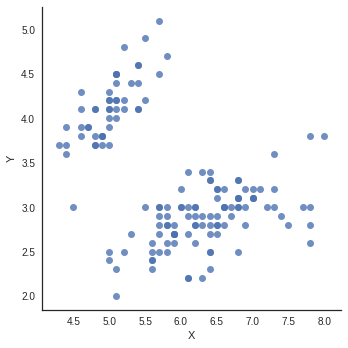
Et celui de la variable y :



Ce n'est pas très lisible, n'est-ce pas ? Ne vous inquiétez pas, l'histogramme (à gauche) et la boîte à moustaches (à droite) ont été créés justement pour remédier à ce problème :



Maintenant, passons à une analyse bivariée, avec un graphique de dispersion sur lequel nous représentons les variables x et y :



Tout à coup, on peut voir une chose que l'on ne voyait pas avant : les individus sont séparés en 2 groupes bien distincts ! Ceci, nous n'étions pas capables de le voir en analyse univariée !

#### Deuxième exemple : un questionnaire de satisfaction

Nous demandons aux étudiants qui suivent ce cours s'ils sont satisfaits, à travers 5 critères sur lesquels ils doivent se positionner sur une échelle de 1 à 5, 1 correspondant à "très insatisfait" et 5 à "très satisfait". Voici ces 5 critères :

1. Clarté du cours écrit
2. Fluidité de la lecture du cours écrit
3. Les exemples du cours écrit sont-ils faciles à comprendre ?
4. Clarté des vidéos
5. Êtes-vous satisfait du rasage du professeur dans la vidéo ?

8 étudiants ont répondu :

|  | **critère 1** | **critère 2** | **critère 3** | **critère 4** | **critère 5** |
| --- | --- | --- | --- | --- | --- |
| individu 1 | 3 | 3 | 3 | 5 | 5 |
| ind. 2 | 2 | 3 | 2 | 5 | 4 |
| ind. 3 | 2 | 3 | 3 | 4 | 5 |
| ind. 4 | 1 | 1 | 1 | 1 | 1 |
| ind. 5 | 5 | 5 | 4 | 3 | 3 |
| ind. 6 | 4 | 5 | 5 | 2 | 3 |
| ind. 7 | 5 | 5 | 5 | 3 | 3 |
| ind. 8 | 1 | 1 | 1 | 1 | 1 |
| **moyenne** | **2.875** | **3.25** | **3** | **3** | **3.125** |

Si vous regardez les moyennes pour chaque critère, vous faites une analyse univariée, car vous étudiez chaque variable (chaque colonne) une à une. Cela nous amène à penser que les étudiants ont globalement un niveau de satisfaction qui tourne autour de 3, et ce quel que soit le critère.

Mais ce qui nous intéresse ici, c'est le **profil** des étudiants : leur profil, c'est l'ensemble de leurs 5 réponses. Ici, un étudiant, c'est un "paquet" de 5 valeurs (on dit plutôt un vecteur de dimension 5). Comme c'est le profil des étudiants qui nous intéresse, on ne peut pas se contenter d'étudier chaque variable séparément, il faut étudier les paquets de 5 valeurs qui caractérisent les étudiants.

Ainsi, on voit par exemple que 2 étudiants ont répondu "1" à tous les critères (les individus 4 et 8). C'est un phénomène classique : à chaque fois que l'on propose un questionnaire de satisfaction, certaines personnes souhaitant exprimer leur mécontentement répondent "très insatisfait" (ou l'inverse) à toutes les questions sans trop regarder l'intitulé de celles-ci.

Ce phénomène, nous ne sommes pas capables de le détecter en analyse univariée (nous ne l'avons pas vu en étudiant la moyenne par colonne, par exemple). À vous de décider ensuite si vous souhaitez garder ces individus "revendicatifs" pour la suite des traitements statistiques.

Nous ne pouvons pas repérer une autre information en analyse univariée. Hormis les "revendicatifs", il existe 2 groupes bien distincts de personnes : celles qui sont très satisfaites des vidéos, mais moyennement du texte (individus 1, 2 et 3), et celles qui sont au contraire très satisfaites du texte, mais moyennement des vidéos (individus 5, 6 et 7).

En résumé, l'analyse multivariée est utile quand on souhaite étudier des profils, c'est-à-dire un ensemble de caractéristiques d'un individu.

## Découvrez les méthodes factorielles et la classification non supervisée

Rentrons progressivement dans le sujet, en découvrant l'enjeu des traitements statistiques que nous allons aborder.

Nous allons évoquer 2 familles de méthodes :

* les méthodes **factorielles ;**
* les méthodes de **classification non supervisée**, aussi appelées de partitionnement de données (plus connues sous le terme anglophone de **clustering**).

Chacune de ces 2 familles possède une méthode emblématique :

* l'**analyse en composantes principales** (ACP) ou Principal component analysis (PCA) en anglais, qui est la plus connue des méthodes factorielles ;
* l'algorithme **k-means** (en français "K-moyennes"), qui est le plus connu des algorithmes de clustering.

#### En quoi sont-elles intéressantes ?

Dans les 2 cas, ces 2 familles de méthodes ont un intérêt commun : celui de simplifier les données pour faciliter ensuite leur analyse. Comment ? En trouvant des stratagèmes pour réduire les dimensions d'un tableau de données.

Si votre échantillon est représenté par un tableau à 100 000 lignes et 1 000 colonnes, c'est un peu difficile à analyser ! De plus, certains traitements statistiques effectués par ordinateur seront très longs à exécuter (parfois plusieurs jours, plusieurs mois, ou plus !).

Alors arrivent l'ACP et le clustering : la première permet de réduire le nombre de variables en trouvant de nouvelles variables qui en **synthétisent** plusieurs. Trouver une variable synthétique permet de remplacer plusieurs colonnes du tableau par une seule. Malheureusement, cette transformation nous fera perdre un peu d'information.

Le clustering quant à lui se chargera de **regrouper** des individus similaires, c'est-à-dire qu'il va **partitionner** l'ensemble des individus. Regrouper des individus est ici synonyme de regrouper des lignes. Parfois, il est possible de regrouper 100 000 lignes en 3 groupes assez homogènes pour n'étudier finalement que le profil général de chacun de ces 3 groupes, c'est-à-dire 3 lignes !

Mais ce n'est pas tout ! Au-delà de la réduction des dimensions du tableau de données, ces méthodes ont d'autres intérêts.

L'ACP, tout d'abord, permet d'étudier :

* la variabilité entre les individus, c'est-à-dire quelles sont les différences et les ressemblances entre les individus ;
* les liaisons entre les variables : y a-t-il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

Retenez bien ces 2 points. Ils constituent une "grille" d'analyse. Si vous sortez de celle-ci, vous risquez de vous égarer et de produire des analyses douteuses.

#### Le clustering a de multiples applications

Il est par exemple très utilisé en marketing pour segmenter une base de données de clients. Le fait de former des "groupes" de clients et d'étudier leurs caractéristiques (en termes d'âge, de centres d'intérêt, etc.) permet aux marketeurs de cibler leurs campagnes de marketing.

Mais la classification non supervisée a bien d'autres applications, par exemple en analyse d'image : lorsque 2 pixels d'une photo sont très similaires en termes de couleur, il est possible de les regrouper en une seule couleur. Ainsi, on réduit de manière optimale le nombre de couleurs d'une image, et on réduit donc son poids ([voir cet exemple réalisé avec sklearn](http://scikit-learn.org/stable/auto_examples/cluster/plot_color_quantization.html#sphx-glr-auto-examples-cluster-plot-color-quantization-py)).

On peut citer également la classification des espèces animales (introduite pour la première fois au XVIIIe siècle par Linné, naturaliste suédois), qui est l'une des classifications les plus célèbres.

En fait, la classification non supervisée existe dans de nombreux domaines, mais sous des noms différents :

* En sciences naturelles : taxonomie ou systématique (classification des formes vivantes).
* En médecine : nosologie (classification des maladies).
* Etc.

### Supervisé ou non supervisé ? Telle est la question.

En statistiques, on distingue les traitements **supervisés** des traitements **non supervisés**.

L'approche non supervisée consiste à explorer des données sans guide, alors que l'approche supervisée apprend pour prévoir (une variable quantitative, dans le cas d'une régression ; ou une variable qualitative, dans le cas d'une classification).

"Prédisez une variable en fonction de la valeur des autres." (Vous souvenez-vous de la [régression linéaire](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/analysez-2-variables-quantitatives-par-regression-lineaire), par exemple ?)

Alors, dans ce cours, on fait du supervisé ou pas ?

Non ! Toutes les méthodes que nous verrons dans ce cours sont non supervisées.

On avait déjà un indice, car j'avais dit que nous faisions ici des statistiques **exploratoires** (descriptives).

La **classification non supervisée** consiste en l'organisation d'individus en groupes homogènes. En gros, on définit des classes que l'on ne connaît pas à l'avance.

La **classification supervisée** consiste à "ranger" les individus dans des classes connues. Ici, il y a une question préalable. Vous rencontrerez peut-être le terme de classement, qui est synonyme de classification supervisée.

Comme nous l'avons vu, la classification non supervisée est appelée en anglais clustering. Jusque là, vous me suivez ; mais attention, quand un anglophone vous parlera de classification, il vous parlera en fait de... classification supervisée ! Vous êtes un peu perdu ? Ne vous inquiétez pas. ;)

|  | **En anglais** | **En français** |
| --- | --- | --- |
| **supervisé** | classification | classification supervisée |
| **non supervisé** | clustering | classification non supervisée |

## Téléchargez les jeux de données analysés dans ce cours

Nous illustrerons les chapitres qui suivent à l’aide de 3 jeux de données.

### 

### Échantillon n°1 : Les cours OpenClassrooms que vous avez suivis

Nous allons utiliser des données que vous connaissez bien : les cours OpenClassrooms que vous avez suivis. La liste de ceux-ci est disponible sur votre page d'accueil : <https://openclassrooms.com/dashboard>

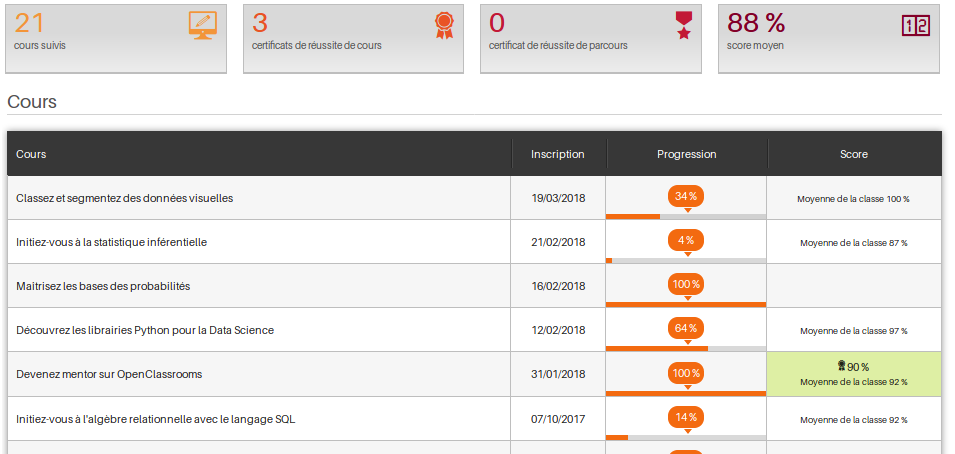
À l'aide du code Javascript donné ci-dessous, vous allez télécharger les données au format CSV.

Si cela ne fonctionne pas, ou si vous n'avez pas suivi assez de cours (l'idéal serait d'en avoir au moins 15), vous pouvez toujours télécharger l'échantillon qui provient de mon profil sur le[**git du cours**](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire) (fichier my\_courses.csv).

Dans cet échantillon, chaque individu est un cours que vous avez suivi. Voici le détail des variables :

* titreCours : le titre du cours.
* idCours : l'identifiant du cours.
* inscription : nombre de jours écoulés depuis votre inscription au cours.
* progression : votre progression sur le cours (en pourcentage).
* moyenneDeClasse : moyenne de la classe aux évaluations (en pourcentage).
* duree : durée estimée du cours (en heures).
* difficulte : difficulté estimée du cours (1 : facile... 3 : difficile).
* nbChapitres : nombre de chapitres.
* nbEvaluations : nombre d'évaluations dans le cours (comprend les quiz et les activités).
* ratioQuizEvaluation : proportion de quiz par rapport au nombre total d'évaluations (nombre d'évaluations : nombre de quiz + nombre d'activités).

À partir de votre navigateur web (Firefox, Chrome, Safari, Edge, ou autre), rendez-vous sur cette page : <https://openclassrooms.com/dashboard>. Vous devriez y voir la liste des cours que vous avez suivis. Cela devrait ressembler à ceci :



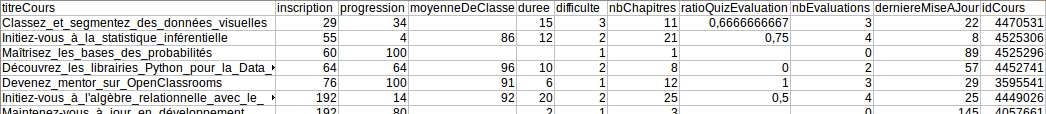
La moyenne de classe n'est plus accessible sur le dashboard depuis 2018. Si vous utilisez ce code javascript, la colonne moyenneDeClasse sera donc replie de 0.

Ensuite, tout en restant sur cette page d'accueil, ouvrez la console du navigateur. En général, il suffit d'appuyer sur la touche F12 pour qu'elle apparaisse en bas de l'écran.

Si cela ne fonctionne pas avec votre navigateur, tapez dans votre moteur de recherche favori votre nom de navigateur suivi de cette phrase : "how to open the console".

La console du navigateur vous permet d’exécuter du code Javascript. Copiez-collez-y le code présent dans le fichier extract\_my\_courses.js présent sur le [git du cours](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire).

Après quelques secondes d’exécution, une page blanche devrait s'afficher, vous proposant de télécharger un fichier CSV. Téléchargez-le, et retenez l'endroit où vous l'enregistrez sur votre ordinateur. Il devrait ressembler à cela :



### 

### Échantillon n°2 : Le texte de cours OpenClassrooms

Le second provient de la plateforme OpenClassrooms. J’ai tout simplement récupéré le texte de différents cours.

Vous pouvez les télécharger sur le [**git du cours**](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire) (fichiers bag\_of\_words.csv et courses\_info.csv).

Plus précisément, j'ai récupéré les cours de 12 parcours de formation. Un parcours (ex. : le parcours [Data Analyst](https://openclassrooms.com/paths/data-analyst)) est composé de plusieurs cours.

Voici la liste de ces parcours, qui est détaillée plus en détail dans le fichier courses\_info.csv :

* Thématique Data :
  + Parcours Data Analyst
  + Parcours Data Architect
  + Parcours Data Scientist
* Thématique Développement :
  + Développeur·se d'application - Python
  + Développeur·se web junior
  + Développeur·se d'application - PHP / Symfony
* Thématique Marketing :
  + Community manager
  + Responsable marketing opérationnel et communication
  + Expert·e en stratégie marketing et communication
* Thématique Ressources Humaines :
  + Gestionnaire de paie
  + Manager ressources humaines
  + Chargé·e de gestion des ressources humaines

Dans tous ces textes, nous nous intéresserons aux mots qui les composent.

Mais comment représenter des textes au format dont nous avons l’habitude, c’est-à-dire un tableau avec des lignes et des colonnes ?

En utilisant l’approche « [sac de mots](https://openclassrooms.com/courses/exploitez-des-donnees-textuelles/representez-votre-corpus-en-bag-of-words) », ou « bag of words » en anglais. Elle consiste à représenter les textes en un tableau dans lequel chaque ligne correspond à un texte, et chaque colonne correspond à un mot. Dans chaque case, on indique l’effectif ou la fréquence du mot en question dans le texte. Par exemple, prenons ces 3 textes :

J’aime les statistiques. Surtout les statistiques inférentielles.

J’aime le chocolat noir.

Bonjour !

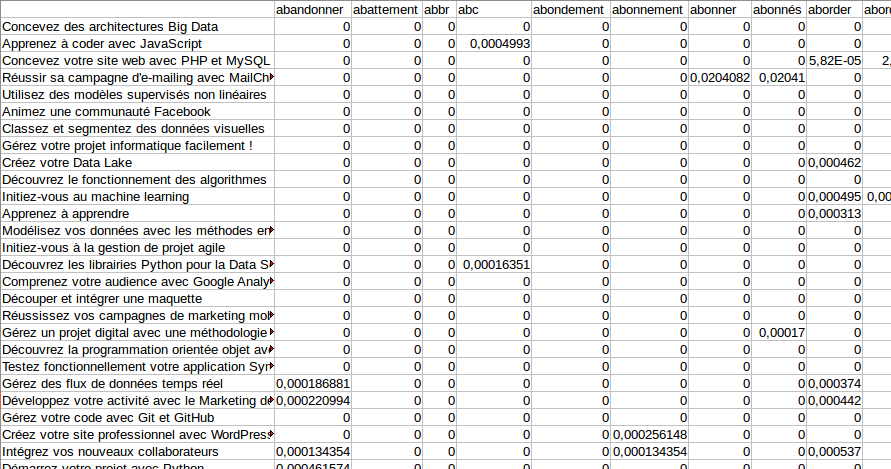
La représentation en bag of words sera la suivante, en représentant les effectifs des mots :

|  | j | aime | les | statistiques | surtout | inférentielles | le | chocolat | noir | bonjour |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| texte 1 | 1 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| texte 2 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| texte 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

L'ordre des colonnes n'a ici aucune importance !

Pour avoir la représentation avec les fréquences des mots, il faut diviser chaque nombre par le nombre total de mots que contient le texte. Par exemple, la fréquence de « statistiques » dans le texte 1 est de 2/8 = 0,25.

Voici un aperçu du jeu de données bag\_of\_words.csv, avec les fréquences des mots :



Ce jeu de données contient 9 343 variables (soit 9 343 colonnes) et 105 individus (correspondant aux 105 cours).

Les mots qui sont très fréquents dans la langue française, et qui n'ont pas de réel sens, ont été retirés. Par exemple : "la", "les", "je", "elles", "est", "en", etc.

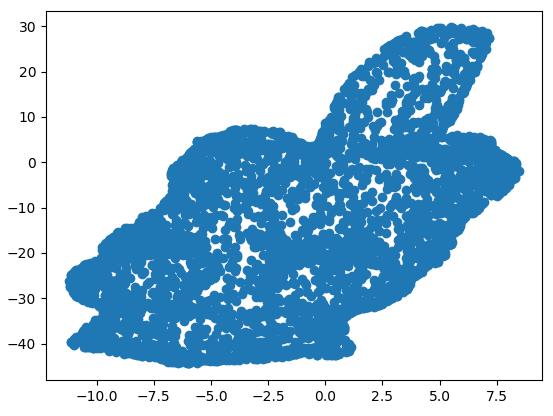
### Échantillon n°3 : surprise !

Pour le 3e échantillon… j’ai envie de vous laisser la surprise pour plus tard ! Je ne vous dis pas à quoi il correspond. Je vous dis juste qu’il est composé de 5 000 individus et de 3 variables qualitatives, que nous appellerons x, y et z.

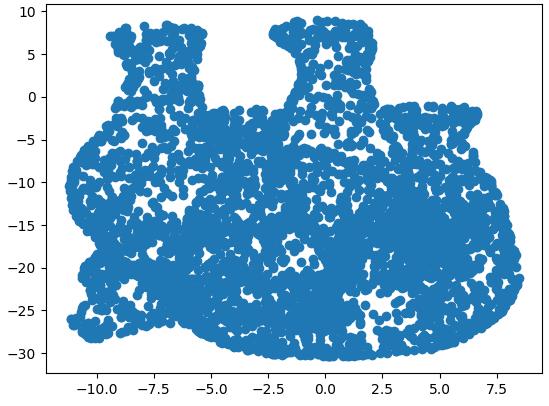
Vous pouvez le télécharger [**sur le git du cours**](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire) (fichier mystery.csv).

Voici juste deux indices, les graphiques de dispersion de :

* x et y :



* x et z :



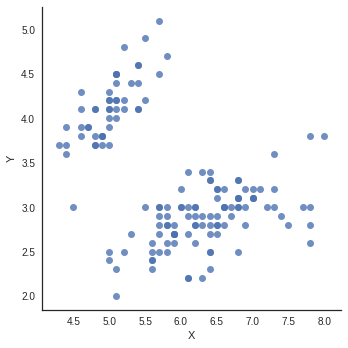
 Une petite idée ? Vous découvrirez la réponse lors de la partie sur l'ACP. ;)

## Représentez vos données dans un espace

### 

### La notion d'espace euclidien

Reprenons le graphique de dispersion d'un précédent chapitre :



Graphique de dispersion

Observons-le un peu. Sur celui-ci, on représente les individus par des points ayant chacun 2 coordonnées : une abscisse et une ordonnée. On dit donc ici que les données sont représentées dans un espace à 2 dimensions, car pour placer les points, on a sélectionné 2 des variables qui décrivent les individus. En quelque sorte, on a associé la notion de **variable** à celle de **dimension**.

**Rappel** :

Quand on étudie un espace à 2 dimensions, on fait de l'[**algèbre linéaire**](https://fr.wikipedia.org/wiki/Alg%C3%A8bre_lin%C3%A9aire). L’algèbre linéaire, c’est une branche des mathématiques dont l’objet principal que l’on étudie est appelé **espace vectoriel**. Vous n'avez pas forcément à savoir ce que c'est, mais retenez juste que c'est un espace dans lequel tous les objets sont des vecteurs.

Sur mon graphique de dispersion, il n’y a pas de vecteurs, il n’y a que des points !

Oui, c’est vrai. Mais on peut en fait considérer que chaque point est équivalent à un vecteur. Pour un point A, par exemple, on dira qu’il est équivalent au vecteur  où **O** est l’origine du repère (c'est-à-dire le point de coordonnées (0,0) dans un espace à 2 dimensions). Cette petite astuce nous permet d’utiliser toute la puissance de l’algèbre linéaire en statistiques : vous vous en rendrez rapidement compte dès les prochains chapitres ! Ainsi, dans la suite, on ne fera pas de distinction entre la notion de **point** et de **vecteur**, et on pourra noter **A** indifféremment, comme ceci : **A** ,  ou même .

 Comment représenter un vecteur ?

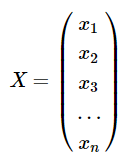
Le plus souvent, on représente un vecteur en colonne. Pour le cas du point **A**, qui a deux coordonnées (abscisse x et ordonnée y), on le note comme ceci :



Bon, mais si vecteur et point sont ici équivalents, et que l’on représente un individu par un point, cela veut dire que l’on peut aussi représenter un individu par un vecteur, n’est-ce pas ?

Tout à fait ! Si un individu est décrit par 4 variables, alors on peut le représenter par un vecteur à 4 dimensions.

Vous voyez poindre à l’horizon des espaces à plus de 2 dimensions ; il arrive régulièrement d’avoir des échantillons décrits par beaucoup de variables (parfois 100, 1 000 ou plus !). On préfère donc noter le vecteur X comme ceci :



**x1, x2 … xn** sont appelées les **composantes** du vecteur X.

Voilà, nous avons posé un peu le cadre : nous travaillons dans un espace vectoriel avec un nombre fini de dimensions (2, 4, 100, 1 000 ou beaucoup plus), où chaque individu est représenté par un vecteur, ce vecteur ayant autant de dimensions que l’espace vectoriel en question.

Si maintenant on rajoute la contrainte que chaque composante d’un vecteur doit être un nombre réel, et que l’on associe à cet espace vectoriel un [produit scalaire](https://fr.wikipedia.org/wiki/Produit_scalaire), alors on dit que l’on travaille dans un**espace euclidien**.

 « On associe à cet espace un produit scalaire », cela veut dire quoi ?

Un produit scalaire est une opération algébrique entre 2 vecteurs. Dans notre cas, cette opération associe à 2 vecteurs un nombre réel.

Dire que l’on « associe » un produit scalaire à un espace vectoriel signifie que l’on va souvent utiliser celui-ci dans les calculs que nous effectuerons : c'est ce produit scalaire qui permettra de calculer des distances, des longueurs, des projections et des angles.

Monsieur Wikipédia a une petite précision à vous apporter :

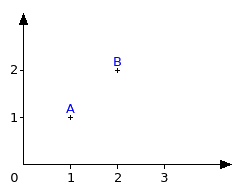
Un espace euclidien est un objet algébrique permettant de généraliser de façon naturelle la géométrie traditionnelle développée par Euclide dans ses Éléments. ([**Source**](https://fr.wikipedia.org/wiki/Espace_euclidien))

... c’est-à-dire toute la géométrie que vous avez étudiée à l’école !

### 

### La notion de distance

Si je vous demande la distance entre 2 points A et B sur un graphique à 2 dimensions comme celui ci-dessous, qu’allez-vous faire ?



Graphique à 2 dimensions

Certains seront allés chercher une règle graduée pour mesurer, d’autres auront été plus aventuriers et auront calculé la distance à partir des coordonnées des 2 points. Mais, dans les 2 cas, vous obtiendrez tous le même résultat (ici, √2 , soit environ 1.41).

Ce que vous avez mesuré instinctivement s’appelle la **distance euclidienne**, que vous connaissez tous.

Tu sous-entends qu’il y a plusieurs types de distances ?

Tout à fait. Mais vous le savez déjà, sans vous en rendre compte. Quand dans une ville, vous demandez à quelle distance se trouve un bâtiment donné, on vous répondra soit avec une distance « à vol d’oiseau », soit avec une distance en suivant les rues (car vous ne pouvez pas voler, je pense). En mathématiques, c’est un peu le même principe : il y a plusieurs types de distances. Pour reprendre l’exemple de la ville, sachez qu’il existe par exemple la distance de Manhattan.

La distance de Manhattan (appelée aussi taxi-distance) est la distance entre deux points parcourue par un taxi lorsqu'il se déplace dans une ville où les rues sont agencées selon un réseau ou quadrillage. Un taxi-chemin est le trajet fait par un taxi lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau. ([*Source*](https://fr.wikipedia.org/wiki/Distance_de_Manhattan)))

Sur le graphique précédent, on calcule la distance de Manhattan en se déplaçant d’abord parallèlement à l’axe des abscisses (on trouve donc 1), puis en se déplaçant parallèlement à l’axe des ordonnées (on trouve encore 1), ce qui nous donne une distance de 2.

Il existe aussi une distance qui calcule des proximités entre des mots (plus généralement des chaînes de caractères) : c’est la distance de [Levenshtein](https://fr.wikipedia.org/wiki/Distance_de_Levenshtein), selon laquelle la distance entre « Bonjour » et « Bonsoir » est de 2.

### La notion de nuage de points

Lorsque l’on représente les individus d’un échantillon par des points dans un espace euclidien, l’ensemble de ces points est appelé **nuage de points**. Poétique, non ? Comme durant les après-midis d’été, allongés sur l’herbe, où nous regardons les nuages dans le ciel. Vous avez d’ailleurs sûrement déjà joué à ce jeu : essayer de comparer la forme des nuages à des animaux ou à d’autres objets connus.

En statistiques, on fait la même chose, on décrit des nuages de points : quelle forme ont-ils ? sont-ils étalés, resserrés, denses, gros, petits ? quelle est leur position ?

Un nuage étalé dans l’espace traduira par exemple des individus très différents les uns des autres. Peut-être y a-t-il des amas dans un nuage, c’est-à-dire des zones plus denses que d’autres. Dans ce cas, cela signifie qu’il y a des groupes d’individus similaires entre eux, et plutôt différents des autres groupes.

### 

### La notion d’inertie

Comme nous étudions la dispersion d’un nuage (étalé ou resserré), nous avons besoin d’une notion qui définit ce concept : c’est la notion d’**inertie**.

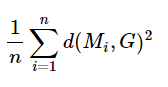
La notion d’inertie est similaire à celle que rencontrent nos amis physiciens lorsqu’ils étudient le mouvement des objets (cette discipline s’appelle la mécanique) : un objet avec une forte inertie est un objet difficile à mettre en mouvement, ou à faire entrer en rotation.

Si vous avez deux objets de même masse, mais pas de même taille, l’objet qui sera plus grand (donc plus étalé dans l’espace) sera plus difficile à faire tourner autour de son centre de gravité.

Nous avons déjà vu ce parallèle en statistique descriptive monodimensionnelle quand nous avons parlé des moments empiriques et des [moments d’inertie](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/formez-vous-sur-les-mesures-de-forme#r-4739831). L’un des moments que nous avons étudiés était la variance.

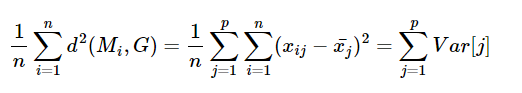
Pour une variable donnée, la variance empirique est calculée à partir de la somme des carrés des distances entre les observations et leur moyenne. Cela, c’est pour une variable, c’est-à-dire pour un nuage de points à 1 dimension. La généralisation de ce concept à un nuage de points à pp dimensions nous donne l’inertie du nuage de points **NI** . Ainsi, l’inertie de **NI** est la moyenne des carrés des distances entre les points **Mi** et leur centre de gravité **G**. On note la distance entre le point **i** et **G** comme ceci : **d(Mi,G)**.

L'inertie totale de **NI** par rapport à **G** est donc égale à :



Cette définition suppose que l’on ne souhaite pas attribuer des poids différents aux points : on considère donc qu’ils ont tous la même importance dans l’analyse.

Plus le nuage sera dispersé (étalé), plus son inertie sera grande. Nous avons déjà vu cette notion de dispersion avec la variance empirique, que nous avons qualifiée d’indicateur de dispersion. D’ailleurs, l’inertie du nuage de points **Ni** est aussi la somme des variances sur toutes les **p** dimensions :



# Avez-vous compris l'intérêt de l'analyse exploratoire multidimensionnelle ?

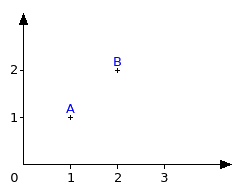
Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Appréhender la notion d'espace vectoriel euclidien

### Question 1

**Quelle est la distance euclidienne entre A et B ?**

****

* + 

1

* + 

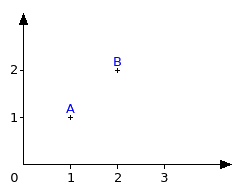
2

* + 

 √2

### Question 2

**Quelle est la distance de Manhattan entre A et B ?**

****

* + 

1

* + 

2

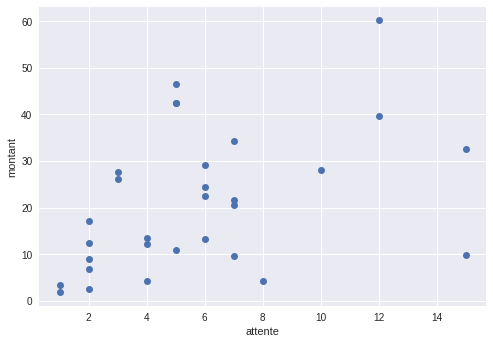
* + 

 √2

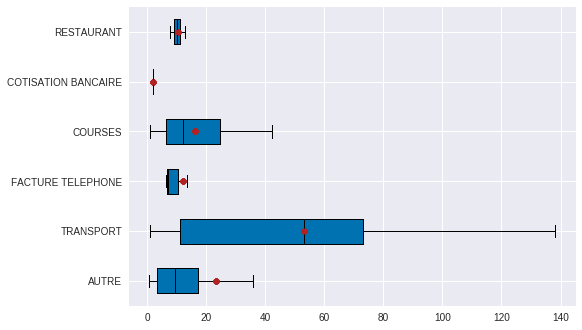
### Question 3

**Laquelle de ces 3 illustrations représente un espace euclidien à 2 dimensions ?**

* + 



* + 

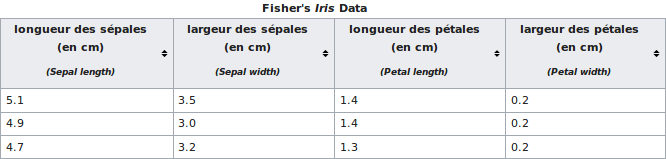


* + 



*Le graphique 1 est un graphique de dispersion avec 2 variables ("attente" et "montant"), donc 2 dimensions. De plus, ces 2 variables sont quantitatives réelles, d'où l'espace euclidien. Le graphique 2 présente également 2 variables mais l'une d'elles est qualitative. L'image 3 représente un satellite dans l'espace; l'espace dans lequel nous vivons est à 3 dimensions spatiales.*

### Question 4

****

**Pour représenter les 3 individus donnés (en ligne) de l'échantillon donné sur cette image, la solution la plus logique (dans le cadre de ce cours) sera une représentation par :**

* + 

Un espace euclidien à 4 dimensions

* + 

Un espace euclidien à 12 dimensions

* + 

Un espace euclidien à 3 dimensions

### Question 5

**Que permet l'ACP ?**

*Attention, plusieurs réponses sont possibles.*

* + 

Pas grand chose malheureusement

* + 

Étudier les liaisons (linéaires) entre les variables d'un échantillon

* + 

Réduire les dimensions d'un échantillon

* + 

Étudier les ressemblances et la variabilité entre individus

### Question 6

**Le clustering, c'est :**

*Attention, plusieurs réponses sont possibles.*

* + 

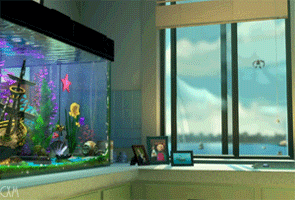
Le fait de regrouper des individus d'un échantillon en différents groupes

* + 

Le fait de partitionner les individus d'un échantillon

* + 

Le fait de projeter des individus sur un plan à 2 dimensions :



### Question 7

**Pour détecter les individus de mauvaise foi ayant répondu "très mécontent" à l'ensemble des questions d'un questionnaire de satisfaction de 100 questions, il faut une analyse...**

* + 

monodimensionnelle

* + 

multidimensionnelle

## Comprenez l'enjeu de l'Analyse en Composantes Principales

### 

### Qu'allons-nous voir dans cette 2e partie ?

Cette partie est assez conséquente, mais passionnante ! Vous y découvrirez la plus emblématique des méthodes factorielles : l'**Analyse en Composantes Principales**, très souvent appelée**PCA** (pour Principal Component Analysis). Ici, nous l’appellerons ACP.

L’ACP est cruciale et stratégique pour un·e data analyst. Il est donc très important de bien la comprendre. Elle nécessite un minimum de pratique pour analyser les données correctement. Si la compréhension est approximative, elle mène alors très facilement à des analyses erronées, approximatives.

L'ACP a deux objectifs principaux. Retenez-les bien, nous y ferons référence tout au long de la partie. Elle permet d'étudier :

* La variabilité entre les individus, c'est-à-dire quelles sont les différences et les ressemblances entre individus.
* Les liaisons entre les variables : y a-t-il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

Cette partie est composée de :

* 1 chapitre, simple à appréhender : Comprenez l'enjeu de l'ACP.
* 1 chapitre un peu plus mathématique, où l'on plonge dans la « machinerie » interne de l’ACP : il est important pour la compréhension des 2 chapitres suivants.
* 4 chapitres qui traitent du vif du sujet :
  + 1 chapitre sur le cercle des corrélations.
  + 1 chapitre sur la projection des individus sur le premier plan factoriel.
  + 1 chapitre sur le choix du nombre de composantes.
  + 1 chapitre de pratique : il est important pour la pratique, car il donne quelques exemples concrets d'analyses.

Un souci de compréhension sur l'ACP ? Posez votre question sur le [**forum du cours**](https://openclassrooms.com/forum/sujet/cours-analyse-exploratoire-de-donnees) ;)

### 

### L'enjeu de l'ACP

Nous avons parlé précédemment du [nuage de points](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/representez-vos-donnees-dans-un-espace#r-5145716) par lequel on représente les individus.

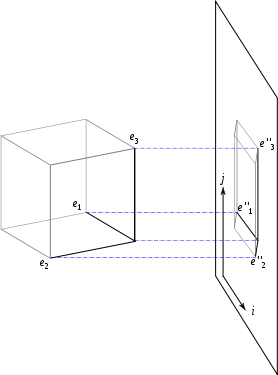
C’est ce nuage que nous souhaitons étudier, et donc visualiser. Mais comme nous l’avons déjà évoqué, nous avons un petit souci pour visualiser ce nuage de points. En effet, il a souvent plus de 2 dimensions, alors que nos écrans sont en 2 dimensions. Un écran ou une feuille de papier sont des [plans](https://fr.wikipedia.org/wiki/Plan_(math%C3%A9matiques)).

Pour visualiser des points dans un espace à **n** dimensions (avec **n>2**) sur un plan, la solution est d’effectuer une **projection orthogonale**.

Le mot **projection** vous évoque peut-être une projection de cinéma, ou bien le fait de lancer un objet sur un mur, comme le font certains artistes quand ils projettent de la peinture sur une toile.  
Dans les 2 cas, il s’agit de la même chose : pour le cinéma, on projette une image (l’image d’un acteur). Dans la réalité, cet acteur est à 3 dimensions (non, un être humain n’est jamais plat !), mais l’image projetée sur l’écran est en 2 dimensions, car l’écran de cinéma est un plan. De même, la toile du peintre est aussi plane, alors que les gouttes de peinture qui y sont projetées sont à peu près sphériques, donc en 3D.

L’espace dans lequel nous vivons est à 3 dimensions, que l’on a coutume d’appeler **largeur, hauteur** et **profondeur.**

La [projection](https://fr.wikipedia.org/wiki/Projection_affine) mathématique, c’est la même chose : c’est représenter des points à **p** dimensions dans un espace plus « petit », c’est-à-dire à **q** dimensions avec**q<p**.



Projection orthogonale d'un cube (à 3 dimensions) sur un plan (à 2 dimensions)

(Source : [Wikipedia, projection orthogonale](https://fr.wikipedia.org/wiki/Projection_orthogonale))

Malheureusement, quand on projette des points, on perd de l’information.

Eh oui ! Reprenons l’exemple du cinéma. Imaginons que l’acteur soit face à la caméra et qu’il tienne dans ses mains une règle graduée en centimètres, face à la caméra. La règle est à la même distance de la caméra que l’acteur.

Sur l’écran de cinéma, vous pouvez facilement savoir quelle distance sépare les 2 yeux de l’acteur grâce à la règle graduée. Vous pouvez également mesurer l’espacement entre ses 2 oreilles ou même la taille de l’acteur. Vous avez donc une bonne appréhension d’au moins 2 dimensions : la largeur et la hauteur.

Cependant, pour la 3e dimension (la profondeur), c’est plus compliqué. En effet, vous ne pouvez pas mesurer avec précision la distance entre le ventre de l’acteur et son dos ni même la distance entre le bout de son nez et le reste de son visage. Pour mesurer ces 2 longueurs, il faut placer la règle dans le sens de la profondeur, et il vous est impossible d’appréhender avec précision cette profondeur sur un écran de cinéma. C’est pour cela que l’on dit qu’il y a une perte d’information : vous perdez l’information de la profondeur.

Perdre de l’information, c’est frustrant !

Oui ! Il faut essayer de perdre le moins d’information possible.

Cela tombe bien, car pour un même objet (l’acteur, par exemple), il y a plusieurs projections possibles : en fonction de là où se place la caméra. La caméra peut se placer au-dessus de l’acteur, ou de profil, de face, ou même en contre-plongée, un peu en biais, etc.

Mais avez-vous déjà vu un film dans lequel les acteurs sont filmés constamment du dessus, ou même du dessous ?



Vue de dessous



Vue du dessus

« Non », me direz-vous, car on ne verrait rien ! Rien, ou en tout cas, pas grand-chose.

Effectivement, on verrait moins bien que quand les acteurs sont filmés de face. Dire « on ne voit pas grand-chose » équivaut à dire « on perd beaucoup d’information ». Vu de dessus, on perd par exemple l’information de l’expression du visage : l’acteur est-il heureux ? triste ?



Vues de face, les expressions du visage sont bien plus visibles.

Il y aurait donc des projections qui seraient meilleures que d’autres ?

Tout à fait, il y a des projections pour lesquelles on voit moins bien que d’autres, avec lesquelles on perd plus d’information que d’autres.

Tout l’enjeu est donc de trouver une projection pour nos données qui perde le moins d’information possible.

Pourquoi voit-on mieux un acteur de face que de dessus ?

La principale raison est que la forme de l’être humain est allongée : notre hauteur est plus grande que notre largeur. Ainsi, l’image d’un acteur sera plus « allongée » s’il est filmé de face plutôt que d'en haut. Son image sera plus « étalée » à l’écran.

Une image « étalée » ? Cela me rappelle quelque chose !

Oui, nous avons déjà vu cette notion précédemment : c’est la notion d’[inertie](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/representez-vos-donnees-dans-un-espace#r-5148848).  
Des points très étalés ont une grande inertie.

C’est ici la clé de l’ACP : rechercher la projection pour laquelle l’inertie des points est maximale.

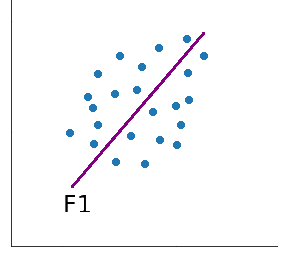
### 

### Le principe de l'ACP

Pour commencer simplement, nous chercherons une projection sur un axe (à 1 dimension) plutôt que sur un plan.

La question est de savoir comment positionner cet axe dans l’espace pour que la projection orthogonale du nuage de points sur celui-ci soit la plus étalée possible.

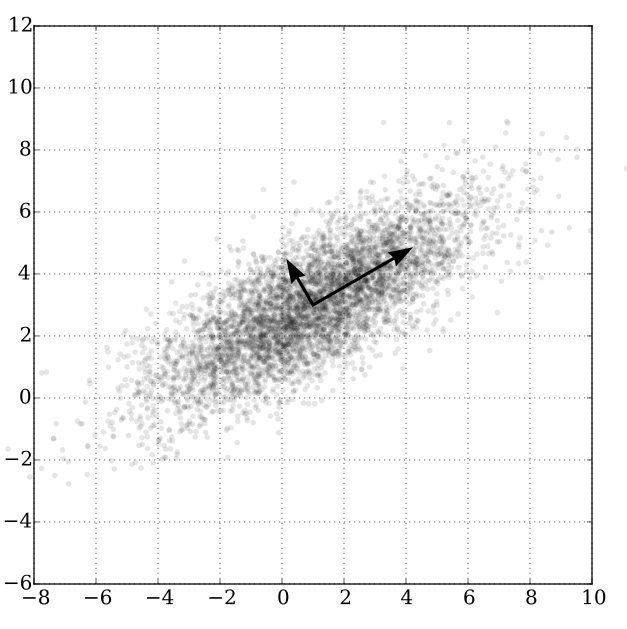
Nous appellerons cet axe **premier axe principal d’inertie**, que l'on note **F1**.



Par cette projection sur le premier axe **F1**, nous perdons de l’information. Pour en perdre moins, on peut ensuite chercher un second axe d’inertie principal **F2**.  
Mais pour ce second axe, nous fixons une règle : il doit être « perpendiculaire » au premier.

« Perpendiculaire » est un terme qui s’emploie sur un plan. Dans un espace à **n** dimensions, on dit plutôt **« orthogonal »**.

Si l’espace de nos données est à 2 dimensions, alors nous n’avons pas le choix : il n’y a qu’une seule direction possible pour ce second axe (c’est la direction donnée par votre équerre quand vous la posez sur le 1er axe d’inertie). Mais si nous sommes à plus de 2 dimensions, il y a plusieurs solutions.



Les deux axes principaux d'inertie perpendiculaires, sur des données en 2 dimensions

(<https://en.wikipedia.org/wiki/Principal_component_analysis#/media/File:GaussianScatterPCA.svg>)

Une fois la direction de ce second axe trouvé, on cherchera le 3e ( F3F3 ), avec la contrainte qu’il soit orthogonal à tous les précédents : au second, mais aussi au premier. Ensuite, nous chercherons le 4e, orthogonal à tous les précédents, etc.

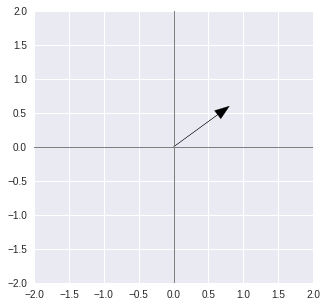
Une chose très importante : ces axes principaux d’inertie peuvent être vus comme de nouvelles variables.

Pas des variables qui sortent de nulle part, non ! Mais des variables qui sont calculées à partir des variables déjà existantes (c’est-à-dire les variables de notre échantillon, que nous appelons « variables initiales »).

Il faut savoir que les axes principaux d'inertie sont des combinaisons linéaires des variables initiales. Par exemple, si nous travaillons avec un échantillon décrit par 2 variables **x** et **y** , alors on peut obtenir un axe principal d’inertie que l’on peut considérer comme une nouvelle variable **F1**, qui sera par exemple de cette forme **F1=0.8×x+0.6×y** :

|  | x | y | F1 |
| --- | --- | --- | --- |
| individu 1 | 1 | 2 | 2 |
| individu 2 | 0 | 3 | 1.8 |
| individu 3 | -1 | 1 | 0.2 |
| ... | ... | ... | ... |

Pour représenter cet axe principal d’inertie, il suffit de tracer le vecteur de coordonnées (0.8, 0.6) :



### 

### Centrer et réduire

N’aurions-nous pas oublié quelque chose ?

Observez ces 2 échantillons, représentant tous deux les mêmes pommes :

* Échantillon 1 :

| **poids (g)** | **diamètre (mm)** |
| --- | --- |
| 100 | 70 |
| 95 | 65 |

* Échantillon 2 :

| **poids (g)** | **diamètre (cm)** |
| --- | --- |
| 100 | 7 |
| 95 | 6.5 |

Ces deux échantillons sont identiques, sauf que l’un exprime le diamètre des pommes en millimètres et l’autre en centimètres.

Que ce soit sur le premier échantillon ou sur le second, l’inertie pour la variable poids est la même (la variance empirique est de 6,25). Cependant, l’inertie pour la variable diamètre est bien plus grande dans le premier cas que dans le second (la variance empirique est de 6,25 dans le premier cas et 0,0625 dans le second).

Dans le premier cas, quand on va chercher le premier axe principal d’inertie, les variables poids et diamètre influenceront de manière égale le calcul de l’axe (elles ont toutes deux une variance de 6,25). Mais dans le second cas, la variable poids « pèsera beaucoup plus lourd » que la variable diamètre dans le calcul, car 6,25 est bien plus grand que 0,00625 ! C’est problématique, car finalement, le premier et le second cas représentent exactement les mêmes pommes !

Ce problème est très classique, car il est très fréquent que les variables d’un échantillon ne soient pas exprimées dans la même unité (simplement parce qu’elles ne représentent pas la même chose) ! Mais on peut ruser pour comparer des variables qui représentent des quantités différentes : faire en sorte que leurs moyennes soient toutes égales et que leurs variances le soient aussi.

Classiquement, on s’arrange pour transformer nos variables de telle manière que leur moyenne soit égale à 0 et que leur variance soit égale à 1.

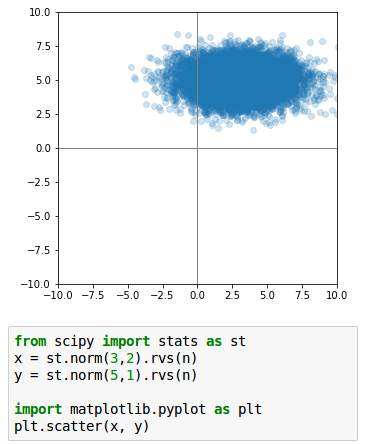
La première opération s’appelle le **centrage**. Pour effectuer un centrage sur des observations, il faut soustraire à toutes ces observations leur moyenne.

Exemple : La moyenne de ces observations **{4,5,6,5,4,6}** est 5. Si l'on centre ces données, on obtient **{−1,0,1,0,−1,1}**. Si l'on calcule la moyenne de ces données centrées, on obtient bien 0. ;)

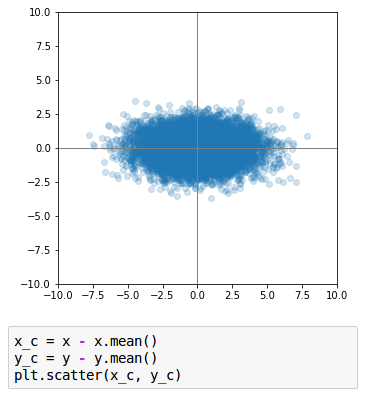
Le fait de centrer les données ne change en rien la forme du nuage de points : ce n’est donc pas dérangeant pour ce que nous souhaitons étudier ici. Centrer les données ne fait que déplacer (par une translation) le nuage de points de telle manière à ce que son centre de gravité coïncide avec l’origine du repère.

La seconde opération s’appelle la **réduction**. Après avoir centré les données, si on les divise par leur écart-type (l’écart-type est la racine carrée de la variance), alors on obtient des valeurs dont la variance vaut 1.

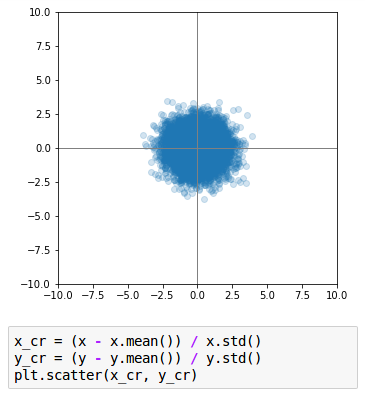
Voici une petite illustration : ces points sont disposés selon une forme elliptique. Une ellipse est un "cercle aplati". Ici, il est aplati selon l’axe horizontal, ce qui signifie que la variance de la variable représentée en abscisses est plus grande que celle de la variable représentée en ordonnées. En effet, elle est de 2 en abscisses et de 1 en ordonnées :



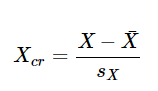
Après centrage des données, le centre de l’ellipse a pour coordonnées (0,0) :



Quand on réduit, l’ellipse devient alors un cercle, car la variable en abscisses a maintenant la même importance que celle en ordonnées : il n’y a plus d’aplatissement !



Voici donc la formule du centrage-réduction, où l'on centre et réduit **X** pour obtenir **Xcr** :



(  est la moyenne de **X** , et  son écart-type)

Appliquons ce calcul au diamètre de nos pommes. Si le numérateur de cette fraction est exprimé en centimètres, alors le dénominateur l’est également. Si l'on veut maintenant convertir en millimètres, on est obligé de multiplier le numérateur par 10, et le dénominateur par 10. Finalement, ces deux 10 s’annulent ! Le résultat est donc le même, quelle que soit l’unité choisie.

## 

## Découvrez les espaces que nous utiliserons

Maintenant que nous avons compris le principe de l’ACP, passons au concret !

Tout d’abord, rappelons-nous les 2 objectifs de l’ACP :

1. Étudier la **variabilité des individus** (leurs ressemblances et différences) ;
2. Étudier les **liaisons entre les variables** (et au besoin, regrouper les variables liées en nouvelles variables synthétiques pour réduire le nombre de colonnes de nos données).

Nous avons 2 objectifs. Pour chacun d’eux, nous allons utiliser un espace différent, ainsi qu’un type de graphique différent.

### 

### Répondre au premier objectif par le nuage des individus **NI** :

Nous avons déjà parlé du nuage **NI** des individus précédemment : vous le connaissez donc déjà.

Pour rappel, si notre échantillon a **p** variables quantitatives, alors on se place dans un espace à **p** dimensions, et nous plaçons les individus dans cet espace. Chaque individu est représenté par un point.

Cet espace est .

On a donc un espace à **p** dimensions qui contient **n** individus.

Pour placer les points, il s’agit de considérer que chaque colonne de notre échantillon correspond à une dimension, et de placer les individus en fonction de leurs coordonnées dans .

L’ensemble des points placés dans l'espace  forme le **nuage des individus**, que l'on notera **NI**.

Pour répondre à l’objectif numéro 1 (étudier la variabilité des individus), on recherche les axes d’inertie maximum, comme expliqué au chapitre précédent.

### 

### Répondre au second objectif par le nuage des variables **NK** :

Pour le second objectif, nous nous plaçons dans un espace totalement différent !

Pour l’objectif 1, chaque colonne de l’échantillon correspondait à une dimension de l’espace, et chaque ligne (chaque individu) correspondait à un point dans cet espace. Mais ici, c’est l’inverse :waw: !!!

Chaque ligne correspondra à une dimension, et chaque colonne à un point ! On est donc dans un espace totalement différent ! Celui-ci aura donc autant de dimensions que d’individus, **n** ; et autant de points que de variables, **p**.

#### L'espace

Nous avons donc :

* **NK**, le nuage des variables ;
* et  , l'espace à **n** dimensions dans lequel il est placé.

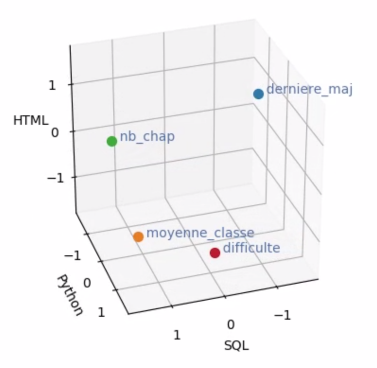
Voici un petit exemple avec un échantillon de 3 individus et 4 variables. Les individus sont des cours OpenClassrooms :

|  | Dernière mise à jour | Moyenne de classe | Nombre de chapitres | Difficulté |
| --- | --- | --- | --- | --- |
| Cours SQL | 20 jours | 85% | 28 | 2 |
| Cours Python | 40 jours | 83% | 10 | 3 |
| Cours HTML | 60 jours | 80% | 19 | 1 |

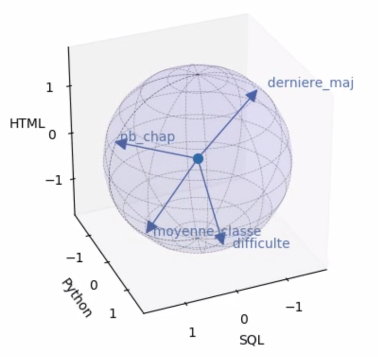
Comme nous l’avons évoqué au chapitre précédent, nous centrons puis réduisons les données, ce qui nous donne ce nouveau tableau (qui n’a plus d’unités à la suite de la réduction) :

|  | Dernière mise à jour | Moyenne de classe | Nombre de chapitres | Difficulté |
| --- | --- | --- | --- | --- |
| Cours SQL | -1.225 | 1.136 | 1.225 | 0 |
| Cours Python | 0 | 0.162 | -1.225 | 1.225 |
| Cours HTML | 1.225 | -1.300 | 0 | -1.225 |

Avec ce tableau, nous aurons 4 points en 3 dimensions :



Pour représenter le nuage des variables, on préfère tracer des flèches entre l’origine et le point en question. Les flèches sont souvent utilisées pour représenter des vecteurs. Mais nous avons dit [**précédemment**](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/representez-vos-donnees-dans-un-espace#r-5148837) que, dans les espaces vectoriels, les notions de point et de vecteur étaient équivalentes. ;)



Représenter le nuage de variables dans cet espace apporte certains avantages très intéressants.

Tout d’abord, si l'on prend deux points (correspondant donc à 2 variables **u** et **v**), et que l’on calcule l’angle  (où O est l'origine du repère), alors cet angle est lié au coefficient de corrélation des variables **u** et **v** !  
C’est pour cette raison que l’on préfère représenter les points par des flèches : il est plus simple visuellement de regarder un angle entre deux flèches.

Plus précisément, c’est le cosinus de cet angle qui est égal au coefficient de corrélation entre **u** et **v**.

C’est une aubaine ! On souhaitait justement étudier les liaisons entre les variables : ici, elles sont tout simplement représentées par les angles entre les flèches !

Cette propriété est valable si les données sont centrées, ce qui est toujours le cas en ACP.

Ensuite, si les données sont centrées et réduites, les longueurs de toutes les flèches sont égales ! On peut même connaître cette longueur : c’est **√n**, la racine carrée du nombre d’individus dans notre échantillon.

En 2 dimensions, l'ensemble des points qui sont à égale distance de l'origine forment un cercle. En 3 dimensions, ils forment une sphère. Dans un espace à plus de 3 dimensions, on dit qu'ils décrivent une [**hypersphère**](https://fr.wikipedia.org/wiki/N-sph%C3%A8re). L'hypersphère, c'est simplement la généralisation de la sphère à un espace de plus de 3 dimensions ;) .

Comme toutes nos flèches sont de longueurs égales, et qu'elles partent toutes de l'origine, alors leurs extrémités sont toutes situées sur une même hypersphère de rayon **√n**. C'est pour cela que, dans l'illustration ci-dessus, j'ai représenté les 3 flèches dans une sphère.

#### Les axes principaux d'inertie du nuage des variables

Rappelons-nous l’objectif numéro 2 : étudier les liaisons entre les variables dans le but d’en regrouper en créant de nouvelles variables synthétiques.

Imaginons un échantillon dans lequel chaque individu est décrit par 15 variables. Ces 15 variables sont regroupables en 2 groupes : un groupe de 10 variables et un autre de 5 variables. Les 10 premières sont très corrélées entre elles, et les 5 restantes sont également corrélées entre elles, mais ne sont pas corrélées aux 10 premières.

Quand on représente le nuage de ces 15 variables, nous aurons 15 flèches.

L’angle entre 2 flèches est lié à la corrélation entre les 2 variables qu’elles représentent. Si cet angle est petit, la corrélation est forte. On aura donc 10 flèches séparées d’un angle très petit. Autrement dit, ces 10 flèches pointent à peu près dans la même direction. Quant aux 5 autres variables, leurs 5 flèches pointeront aussi à peu près dans la même direction, mais cette direction sera différente du groupe des 10 flèches précédentes.

Imaginons que nous construisons une maquette en 3D. Cette maquette représentera le nuage des variables d'un échantillon composé de 3 individus décrits par 15 variables.

Sur celle-ci, vous représentez ces 15 variables par 15 flèches de bois. Si vous étudiez l’équilibre de cette maquette, vous vous rendez compte que les 10 bouts de bois qui pointent dans la même direction déséquilibrent la maquette dans une certaine direction. Autrement dit, ils créent une inertie dans une certaine direction. La direction d’inertie maximale est donnée par ces 10 bâtons ; il y aura également une autre direction d’inertie secondaire induite par les 5 autres bâtons.

On en revient ici aussi au concept de direction d’[inertie maximale](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/comprenez-lenjeu-de-lanalyse-en-composantes-principales#r-5280367), comme quand nous parlions du nuage des individus !  
Ainsi, pour le second objectif, nous chercherons aussi les axes principaux d’inertie, non pas pour le nuage des individus, mais pour le nuage des variables.

### Un résultat remarquable

Attention, vous allez être bluffé !  
À ce stade, nous avons 2 espaces totalement différents :  et . Le premier est à **p** dimensions et contient le nuage d’individus **NI**, et l’autre est à **n** dimensions et contient le nuage des variables **NK**.

Dans chacun des 2 espaces, on a cherché les axes principaux d’inertie.  
Dans , nous avons vu que l’on pouvait considérer les axes principaux comme de nouvelles variables calculables à partir des variables initiales.  
Et bien, figurez-vous que, si l'on place ces nouvelles variables dans , alors celles-ci coïncident exactement avec les axes principaux d’inertie du nuage des variables **NK** !

Autrement dit, étudier les axes d’inertie des individus est équivalent à étudier les axes principaux d’inertie des variables !

Ce résultat est très important en ACP. Il nous dit que, quand on analyse un échantillon, étudier les individus ou étudier les variables, c’est en fait étudier deux facettes d’une même chose !  
Il y a donc une dualité.

### 

### En résumé :

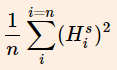
|  | **espace** | **nuage** | **représentation du nuage** | **particularité** |
| --- | --- | --- | --- | --- |
| **Objectif 1** | RpRp | NINI : nuage des individus | par des points |  |
| **Objectif 2** | RnRn | NKNK : nuage des variables | par des vecteurs (flèches) partant de l'origine | Toutes les flèches ont la même longueur si les données sont centrées-réduites. |

### Aller plus loin : Calcul des composantes principales du nuage **NI**, valeurs propres et vecteurs propres

Si vous souhaitez vous plonger dans les calculs, c'est par ici. ;)

Notons **us**, un vecteur unitaire de l'axe de rang **s**. Pour le moment, nous ne connaissons pas encore les composantes de **us** : c'est ce que nous allons chercher ici.

La seule chose que l'on sait, c'est qu'il doit être placé de manière à maximiser l'inertie de la projection (sur **us**) du nuage des individus, c'est-à-dire cette quantité :



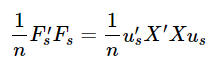
avec la contrainte que u**\_s** soit orthogonal aux **s−1** directions déjà trouvées (c'est-à-dire **us−1,us−2,...,u1**).

 est la projection du point **i** sur **us**)

On peut ranger les projections  dans le vecteur **Fs** , et comme **||Hsi||** s'obtient par produit scalaire entre **us** et la i-ième ligne de nos données , on peut alors écrire sous forme matricielle :



La quantité à maximiser (donnée dans l'encadré ci-dessus) peut donc être écrite sous cette forme matricielle :



Remarque : Si les variables sont centrées-réduites, alors  est la matrice des corrélations. Si les variables sont uniquement centrées, alors il s'agit de la matrice des covariances.

Maximiser la quantité   signifie trouver quel **us** rend cette quantité maximale. En replongeant dans vos souvenirs d'algèbre linéaire, vous verrez que l'on résout ce type de problème par une diagonalisation de matrice.

 Rappel :

Une matrice carrée M est dite diagonalisable s'il existe une matrice inversible P et une matrice diagonale D satisfaisant la relation :



Dans ce cas, chaque vecteur colonne Y de la matrice P est un **vecteur propre** pour la matrice M, c'est-à-dire qu'il existe un scalaire **λ** (appelé **valeur propre**) sur la diagonale de D tel que MY=λY.

On peut montrer mathématiquement (nous ne le ferons pas ici, mais vous pouvez consulter la démonstration [ici](https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales#Projection)) que le vecteur **u\_s** vérifie :



À partir de ce résultat, on voit que **us** est le vecteur propre unitaire associé à la valeur propre **λs** de la matrice **X′X**, les valeurs propres étant rangées dans l'ordre décroissant sur la diagonale.

De plus, on voit que **λs** est égal à l'inertie portée par l'axe **s** (c'est l'inertie de la projection du nuage des individus sur **us**).

Pour obtenir les vecteurs propres **us**, on a donc diagonalisé la matrice des corrélations !

(Source : Analyse factorielle multiple avec R, Jérôme Pagès, éditions EDP Sciences)

## Interprétez le cercle des corrélations

Bon, c’est vrai que le chapitre précédent peut sembler complexe. Mais si vous n’avez pas tout saisi, ce ne devrait pas être un problème pour la suite. ;)

Résumons où nous en sommes.  
Nous avons 2 espaces :

1. l'espace  à p dimensions où l'on a placé le nuage **NI** des individus ;
2. l'espace  à n dimensions où l'on a placé le nuage **NK** des variables.

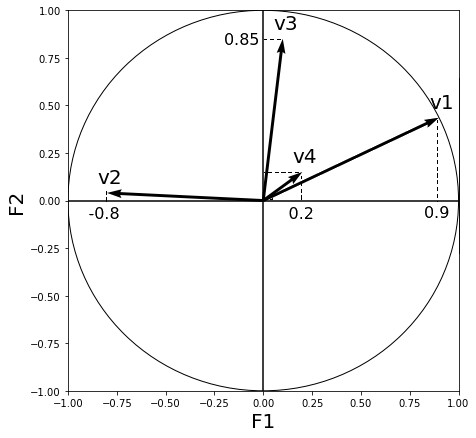
Pour ces 2 objectifs, nous allons créer 2 graphiques. Chacun d’eux sera à 2 dimensions :

1. Pour l'objectif 1, ce sera la projection du nuage des individus **NI** sur les 2 premiers axes d’inertie, c’est-à-dire sur le premier plan factoriel. Le second s’appelle le cercle des corrélations.
2. Pour l'objectif 2, ce sera la projection du nuage des variables **NK** sur le premier plan factoriel.

### Le cercle des corrélations

Dans ce chapitre, nous étudions le second graphique. Il s'appelle **le cercle des corrélations**.

En voici un exemple :



Qu’y voit-on ?  
On y voit un cercle, de rayon 1. De plus, l’axe des abscisses représente le premier axe d'inertie. L’axe des ordonnées représente **F2**.  
À l’intérieur du cercle, il y a des flèches qui partent du centre. Elles sont plus ou moins grandes, et peuvent aller jusqu’à toucher le cercle, sans jamais le dépasser.

Les flèches ne dépassent jamais le cercle à condition que les données soient centrées (mais c’est toujours le cas en ACP) ET réduites (là, ce n’est pas toujours le cas). Une ACP sur des données centrées-réduites est dite **ACP normée**.

Comment l’interpréter ?

Rappelons-nous l’objectif numéro 2 : étudier les liaisons entre les variables. On va donc chercher à savoir s’il y a des groupes de variables qui sont fortement corrélées entre elles (deux à deux).

Si de tels groupes existent, alors toutes les variables d’un groupe donné seront « synthétisables » (résumables) par une variable synthétique.

Ces variables synthétiques sont en fait les composantes principales **F1**, **F2**, **F3**, etc.

Étudions donc les corrélations entre les variables initiales et les composantes principales !

Mais comment détecter des variables fortement corrélées aux variables synthétiques F1, F2 ?

Il suffit de savoir une chose : la projection de la flèche (représentant la variable**v**) sur **F1** correspond au coefficient de corrélation entre **v** et **F1**. Rappelons qu'un coefficient de corrélation est compris entre -1 et 1. Cela tombe bien : le cercle est justement de rayon 1 !

Regardons **v1** : sa projection sur **F1** vaut 0.9.

Cela signifie que ****=0.9.

De même, on a une variable **v2** qui est corrélée négativement à **F1** (on dit aussi anticorrélée), c’est-à-dire que, quand **v2** croît, alors **F1** décroît. Cela se traduit par un coefficient de corrélation proche de -1 (ici -0.8).

Par contre, **v3** est très peu corrélée à **F1**, mais l’est fortement à **F2**. **V4**, quant à elle, est très peu corrélée à la fois à **F1** et à **F2**.

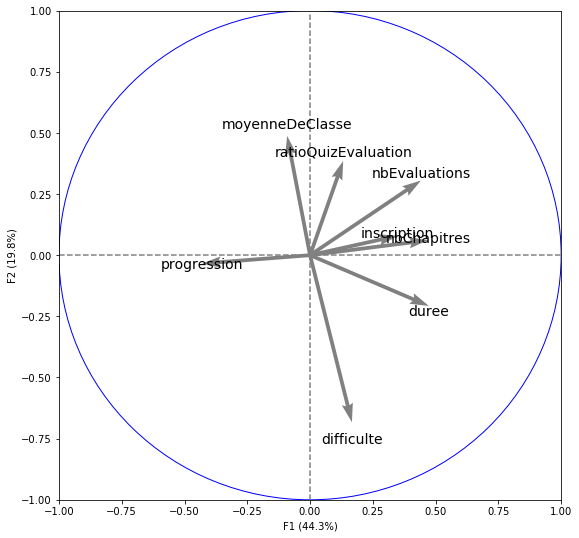
Comme les composantes principales synthétisent les variables initiales, on a l'habitude de les interpréter, c'est-à-dire à chercher à quelle notion elles correspondent. Nous allons voir ceci un peu plus bas. ;)

Au fait, pourquoi dessiner un cercle ?

On verra un peu plus bas que la longueur des flèches est importante. De plus, cette longueur ne dépasse jamais 1 (avec des données centrées-réduites). Le cercle indique donc la longueur maximale des flèches. Si vous avez bien compris le chapitre précédent, alors lisez la section Aller plus loin, et tout s'éclaircira !

### Analyse de votre jeu de données

Pour le jeu de données des cours OpenClassrooms, voici le cercle des corrélations que j'obtiens avec mes données :



Les variables les plus corrélées à F1 sont :

* la durée ;
* le nombre de chapitres (nbChapitres) ;
* le nombre d’évaluations du cours (nbEvaluations).

Elles sont corrélées positivement à F1, mais on a aussi la variable progression qui est corrélée négativement à F1.

Ce qui est intéressant ici, c’est d’interpréter l’axe F1. Ici, il se trouve que toutes ces variables ont un “mode” commun, c’est-à-dire une notion qui les unit.

Ici, c’est la notion de la longueur du cours .

Pour les variables durée et nombre de chapitres, c’est évident.

Pour la variable nombre d’évaluations, on comprend facilement que plus un cours est long, plus il a d’évaluations.

La variable progression a une corrélation négative avec F1.

Intuitivement, on comprend que, PLUS un cours est long, moins on le finit rapidement, ou moins on a de chances de le finir, d’où la corrélation négative.

Si l'on fait le même travail pour la seconde composante, on voit que les variables les plus corrélées à F2 sont :

* la difficulté (avec une corrélation négative) ;
* la moyenne de classe ;
* la proportion de quiz par rapport au nombre total d’évaluations.

On peut donc interpréter F2 comme la facilité du cours.

C’est évident pour la variable difficulté, qui lui est anticorrélée.

C’est également compréhensible pour la variable moyenne de classe : plus un cours est facile, meilleures seront les notes des étudiants.

Pour le ratio quiz/évaluation, il faut savoir qu’il y a deux types d’évaluations : les quiz et les activités. Vous avez sûrement remarqué que les quiz sont en général plus simples que les activités. Donc plus la proportion de quiz est importante, plus simple est le cours !

Ça y est, nous approchons de notre objectif : celui de regrouper les variables en variables synthétiques !

En effet, on peut résumer notre jeu de données comme ceci :

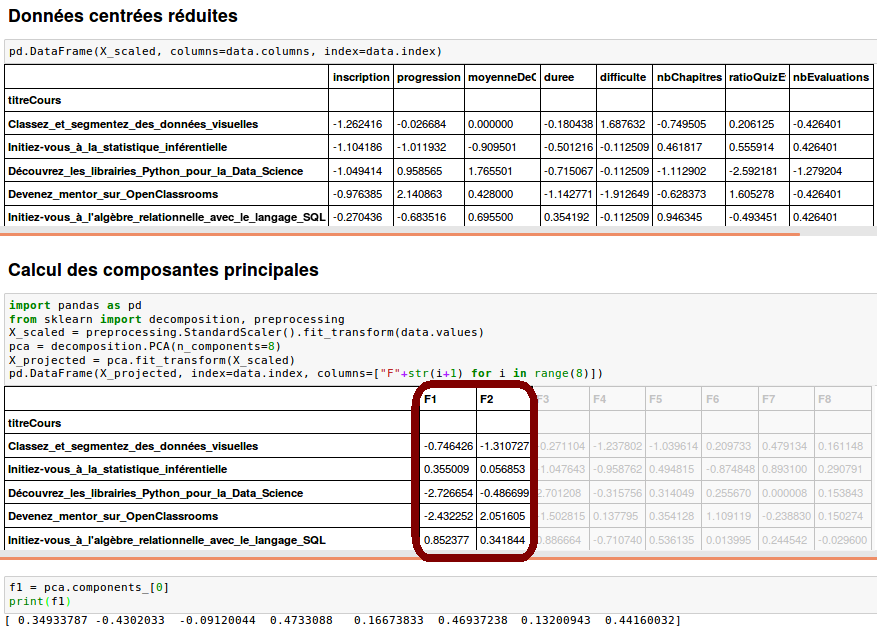
Ce qui différencie le plus les cours de notre échantillon, c’est leur longueur. Ensuite, on trouve une deuxième tendance : il y a les cours qui sont faciles, et ceux qui sont difficiles.

Bien entendu, on ne peut pas rendre compte de la complexité de la réalité de nos **n** individus en 2 phrases ! Quand on résume, on perd de l’information, forcément. Cependant, ces 2 phrases sont les 2 phrases « optimales », c’est-à-dire que ce sont celles qui résument le mieux l’échantillon. Si on avait choisi d’autres phrases pour caractériser notre échantillon, alors on aurait perdu plus d’information ; notre résumé aurait été moins bon.

Ici, on a résumé avec des phrases, mais on peut aussi résumer avec des chiffres !

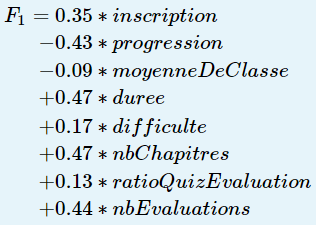
Vous savez que les composantes principales peuvent être vues comme de nouvelles variables, c’est-à-dire comme de nouvelles colonnes de notre tableau de données. Ainsi, on pourrait supprimer toutes les variables initiales et les remplacer par les colonnes F1 et F2.

En résumant **p** variables en 2 variables F1 et F2, on perd forcément de l’information ! Mais quitte à en perdre, autant le faire de manière intelligente : c'est ce que l'on a fait ici.



On crée les nouvelles variables F1 et F2.

**F1** est une combinaison linéaire des autres variables. Ici,



Au passage, on voit, en regardant les coefficients, les variables qui contribuent le plus à **F1**. ;)

Bien sûr, on peut aller plus loin et rajouter des phrases à notre résumé, ou des colonnes à notre tableau, en analysant les composantes F3, F4, etc.

### Attention !

Faut-il interpréter les flèches qui sont petites ?

Une flèche qui est petite sur le premier plan factoriel, cela signifie qu’elle est faiblement corrélée à la première composante principale F1, et faiblement corrélée aussi à F2. Mais elle peut très bien l’être à F3, F4, F5, etc.

Ainsi, il est préférable de n’interpréter que les flèches les plus longues, car les flèches les plus petites correspondent à des variables dites « mal représentées » sur le premier plan factoriel.

Mal représentées ?

Oui, c’est encore une histoire de projection. Vous savez que, quand on projette, on perd de l’information. Une variable bien représentée est une variable pour laquelle on perd très peu d’information sur le plan factoriel. Les variables bien représentées auront une flèche de longueur proche de 1, et leur extrémité sera donc proche du cercle des corrélations. Pour mieux appréhender cette notion, rendez-vous à la section Aller plus loin.

Si certaines variables sont mal représentées, on peut tenter d’afficher le second plan factoriel, avec F3 et F4. On peut aussi continuer avec F5, F6, etc.

Nous verrons prochainement comment bien sélectionner le nombre de plans factoriels à analyser.

Peut-on interpréter les angles entre les flèches ?

Mieux vaut être prudent, et ne pas prendre cette habitude. En effet, le cercle des corrélations n’est pas fait pour interpréter la corrélation entre 2 variables initiales, mais plutôt entre une variable initiale et l’un des axes d’inertie. Pour plus de précisions, descendez cette page jusqu’à la section Aller plus loin.

### 

### Variables et angles : qualité de la représentation

Il faut bien garder en tête que les flèches que vous voyez sur le cercle des corrélations sont en fait des projections.

En effet, on a projeté le nuage des variables **NK** d'un espace  à **n** dimensions vers le plan factoriel, qui est un plan en 2D.

Une flèche, c'est un objet très fin et très long ! Un peu comme un bâton ou une épée.

Prenons cette analogie de l'épée : imaginez-vous au cinéma, en train de regarder un film avec des chevaliers qui se battent à l'épée.

Disons que les épées ont une longueur de 1 mètre.

Sur l’écran de cinéma, la longueur de l’épée apparaîtra maximale si l’acteur la place sur un plan parallèle à l’écran. Mais s’il la déplace et qu’il pointe l’épée directement vers la caméra, alors l’épée vous apparaîtra toute petite, puisqu’elle pointe vers vous.



L’épée pointe presque vers vous, sa longueur est donc mal représentée.

On peut ainsi dire que, quand la longueur de l’épée est maximale, elle est bien représentée, car vous pouvez appréhender toute sa longueur, alors que, lorsqu’elle pointe vers vous, elle est mal représentée.

Sur le cercle des corrélations, c’est pareil. Une flèche qui est de longueur 1 (et qui touche le cercle) est parfaitement bien représentée : elle est parallèle au plan factoriel. Alors qu’une flèche qui sera parfaitement orthogonale à ce plan aura une longueur de 0. Ainsi, plus la flèche est petite, plus on a perdu de l’information dans la projection.

### Aller plus loin : Que dire des angles entre les variables initiales ?

Au chapitre précédent, nous avons vu que, dans l’espace , on représente les variables par des flèches, et que toutes les flèches ont la même longueur (à condition que les données soient centrées-réduites).

Nous avons vu aussi que, dans , le cosinus de l’angle entre 2 flèches correspond au coefficient de corrélation entre les 2 variables correspondantes. Si 2 flèches sont très proches, l’angle qui les sépare est proche de 0°, et cos(0)=1 , donc leur coefficient de corrélation est proche de 1 : ces 2 variables sont très corrélées. De même, si 2 flèches sont orthogonales (perpendiculaires), alors l’angle qui les sépare est de 90°. Le cosinus de cet angle valant 0, deux flèches orthogonales correspondent à des variables non corrélées (indépendantes).

Mais ça, c'est dans . Attention, le cercle des corrélations est une projection dans . Comme c'est une projection, les angles apparaissent déformés !

Attention, ce n’est pas toujours vrai !

On ne peut déduire cela QUE si leurs 2 extrémités sont proches du cercle. Si ce n’est pas le cas, c’est parce qu’elles sont mal représentées sur le plan factoriel. Si elles sont mal représentées, alors il ne faut pas les interpréter ni interpréter l’angle qui les sépare.

De même, on peut croire que 2 flèches perpendiculaires correspondent à des variables indépendantes (non corrélées). Là encore, on ne peut déduire cela QUE si ces 2 variables sont bien représentées sur le plan factoriel !

De manière générale, le cercle des corrélations n'est pas fait pour interpréter les corrélations entre les variables initiales. Il est fait pour interpréter les axes d'inertie F1, F2, etc.

Si cependant vous pensez voir sur le cercle des variables fortement corrélées (ou au contraire non corrélées), alors calculez absolument leurs coefficients de corrélation pour vérifier. Sachez qu'il existe une représentation bien pratique qui donne tous les coefficients de corrélation des variables deux à deux. Elle s’appelle la [matrice des corrélations](http://www.sthda.com/french/wiki/matrice-de-correlation-guide-simple-pour-analyser-formater-et-visualiser).

### Aller plus loin : Pourquoi un cercle ?

Nous avons vu au chapitre précédent que le cercle, la sphère et l'[hypersphère](https://fr.wikipedia.org/wiki/N-sph%C3%A8re) étaient de la même famille. En effet, le cercle est composé de tous les points situés à égale distance du centre du cercle, dans un espace à 2 dimensions. La sphère est son équivalent dans un espace en 3D, et l'hypersphère en est l'équivalent pour tout espace de plus de 3 dimensions.

Par exemple, quand vous prenez un compas et que vous réglez l’espacement entre la pointe et la mine sur une longueur de 5 cm, vous tracez un cercle de 5 cm de rayon. Le compas vous aura permis de dessiner tous les points qui sont à 5 cm de là où vous avez planté la pointe du compas.

Ensuite, il faut savoir que la projection d'une sphère ou d'une hypersphère sur un espace en 2D donne un cercle. En effet, quand une caméra filme un ballon de foot, le ballon vous apparaît comme un cercle sur l'écran de cinéma. Mais en réalité, le ballon de foot est une sphère, pas un cercle !

Revenons à notre nuage des variables **NK**.

Les extrémités des flèches du nuage des variables **NK** sont donc placées sur une hypersphère, car elles sont toutes de même longueur.

Mais comme le graphique du cercle des corrélations est en 2D, on projette cette hypersphère sur le premier plan factoriel. La projection de cette hypersphère sur un plan donne… un cercle ! Ainsi, le cercle des corrélations est simplement la projection de l'hypersphère sur le premier plan factoriel !

Le cercle des corrélations que nous avons vu dans ce chapitre est de rayon 1. Mais l’hypersphère du chapitre précédent a pour rayon **√n**. Qu’à cela ne tienne, divisons la longueur de toutes les flèches par **√n** pour nous ramener une hypersphère de rayon 1 : c’est bien plus facile à interpréter ;). (En algèbre linéaire, on dit plutôt que l’on utilise une métrique définie par une matrice diagonale dont les termes sont **√n**.)

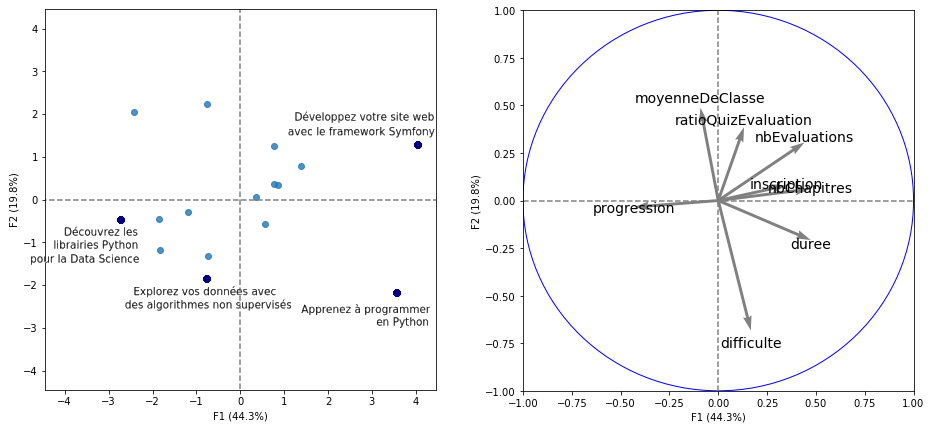
Ainsi, le cercle des corrélations est la projection de l’hypersphère sur le plan factoriel ! De même, les variables du nuage des variables **NK** sont aussi projetables sur ce plan : leur projection donne les flèches qui sont à l’intérieur du cercle des corrélations.

## Représentez les individus sur les plans factoriels

Nous avons répondu à l'objectif n°2, alors passons maintenant à l'objectif n°1 : celui d'étudier la variabilité entre individus.

Nous nous intéressons maintenant à l’espace , dans lequel se situe le nuage des individus **NI**. Comme vous vous en doutez, nous allons projeter ce nuage sur le premier plan factoriel, c’est-à-dire sur un plan composé des 2 premières composantes principales F1 et F2.

Voici ce que cela donne avec le jeu de données des cours OpenClassrooms que j'ai suivis :



Bon, OK, mais on fait quoi avec cela ?

En fait, il faut interpréter ce graphique en parallèle du cercle des corrélations. En effet, le cercle des corrélations nous indique quelles variables sont très corrélées (ou anticorrélées) à F1 et F2.

La question que nous nous posons ici est la suivante : qu’est-ce qui différencie les individus qui ont une abscisse grande de ceux qui en ont une petite ?

Par exemple, qu’est-ce qui différencie cet individu :

* [le cours n°3619856 parlant du framework Symfony](https://openclassrooms.com/courses/developpez-votre-site-web-avec-le-framework-symfony)

de cet individu ?

* [le cours n°4452741 parlant des librairies Python pour la data science](https://openclassrooms.com/courses/decouvrez-les-librairies-python-pour-la-data-science)

La réponse nous est donnée par le cercle des corrélations. En effet, nous avons dit que nous pouvions voir les axes principaux d’inertie comme des « nouvelles variables » qui synthétisent des variables déjà existantes. Ainsi, F1 peut être vue comme une nouvelle variable que l’on ajoute sous forme de nouvelle colonne à notre échantillon.  
Quand on se déplace le long de l’axe des abscisses de gauche à droite, c’est-à-dire dans le sens des abscisses croissantes, alors on se déplace vers les points pour lesquels la valeur de F1 est grande.

Comme F1 est très corrélée aux variables durée, nombre de chapitres et nombre d’évaluations, alors il y a de grandes chances pour que ces individus aient aussi de grandes valeurs pour ces variables.

Ainsi, se déplacer le long des abscisses dans le sens croissant, c’est un peu se déplacer vers les cours qui sont longs.

Cette interprétation est vraie de manière générale, mais peut cependant être fausse pour certains individus pris séparément, surtout pour les individus mal représentés.

De même, la variable progression est anticorrélée à F1, car la flèche de cette variable pointe vers les abscisses décroissantes sur le cercle des corrélations. Ainsi, se déplacer dans le sens des abscisses croissantes signifie se déplacer vers les individus ayant de faibles valeurs pour la variable progression.

Si l'on revient à nos 2 objectifs, on se rend compte que l’on a répondu au premier : celui d’étudier la variabilité des individus, ce qui les différencie, et ce qui fait qu’ils sont semblables.

Ici, on a vu que ces 2 cours différaient par leur longueur : le cours sur [Symfony](https://openclassrooms.com/courses/developpez-votre-site-web-avec-le-framework-symfony) est très long, alors que celui sur les [librairies Python pour la data science](https://openclassrooms.com/courses/decouvrez-les-librairies-python-pour-la-data-science) est court.

Allez vérifier, c'est un très bon réflexe à avoir, surtout quand on débute en ACP. Revenez toujours à vos données pour vérifier vos analyses !

Ensuite, on peut passer à F2. Là, c’est le même principe. On se pose cette question :

« Qu’est-ce qui différencie deux individus qui ont à peu près la même abscisse, mais des ordonnées très différentes ?»

ou bien

« Mis à part la longueur du cours, qu’est-ce qui différencie 2 individus qui ont des ordonnées très différentes ?»

On y répond en regardant les variables les plus corrélées à F2 sur le cercle des corrélations. Ici, difficulté et moyenne de classe.

Deux des cours ont une ordonnée basse :

* [Le cours n°4379436](https://openclassrooms.com/courses/explorez-vos-donnees-avec-des-algorithmes-non-supervises)
* [Le cours n°235344](https://openclassrooms.com/courses/apprenez-a-programmer-en-python)

Ils correspondent donc à des cours plus difficiles que les autres. Le second s'intitule [Apprenez à programmer en Python](https://openclassrooms.com/courses/apprenez-a-programmer-en-python).

Oh ! Mais Python est un langage très simple, enfin !

Oui, certes, mais l'auteur du cours l'a noté comme « Difficile ». De plus, la moyenne de classe est certes bonne (84 %), mais elle reste la plus basse de l'échantillon ! On remarque par ailleurs que ce cours a une abscisse élevée : c'est donc un cours à la fois long et difficile. L'auriez-vous remarqué à la simple lecture de votre échantillon initial, dont le tableau contient 10 colonnes ? Pas sûr ! ;) Alors merci l'ACP !

Si l'on affiche ensuite le deuxième plan factoriel (composé de F3 et F4), alors on se posera cette question : « Qu’est-ce qui différencie 2 individus qui ont des composantes à peu près égales pour F1 et F2, mais très différentes pour F3 ? » Ou bien : « Mis à part la longueur et la difficulté, qu’est-ce qui différencie deux individus qui ont des composantes très différentes pour F3 ? »

… et ainsi de suite avec F4, F5… F100000 ! Enfin… sachez vous arrêter quand même ! Nous verrons ceci au chapitre suivant.

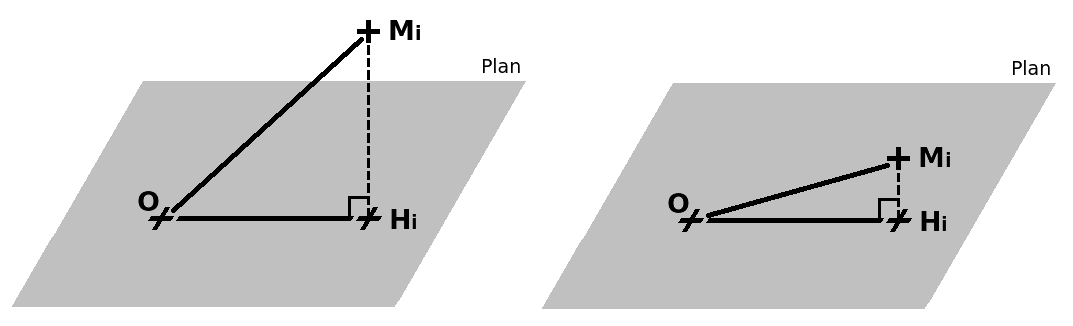
### Qualité de la représentation et contributions

De même qu’au chapitre précédent nous avions des variables (des flèches) bien ou mal représentées, nous avons ici aussi des individus bien et mal représentés.

C’est en fait une histoire de proximité du point avec le plan factoriel.

Un point peut être proche ou loin du plan factoriel sur lequel on le projette. Chaque point **Mi** (c’est-à-dire chaque individu) peut être projeté sur le plan factoriel ; cette projection nous donne le point **Hi**. Quand on ne regarde que le plan factoriel, on ne voit que la position **Hi** ; on ne connaît pas les vraies positions des **Mi**. Les **Hi** sont les images des **Mi**.

* Si **Mi** est proche du plan factoriel, alors la distance entre **Mi** et **Hi** est faible. On peut alors dire : « Où est **Mi** ? Je ne sais pas, mais en tout cas, il n’est vraiment pas loin de **Hi**. Comme on connaît **Hi**, on sait à peu près où est **Mi**. **Mi** est donc bien représenté. »
* Mais si **Mi** est loin du plan, la distance **MiHi** est grande. On se dit alors :  
  « Où est **Mi** ? Je ne sais pas. Certes, je connais sa projection **Hi**, mais comme **Mi** est loin de **Hi**, cela ne m’avance pas beaucoup :(. **Mi** est mal représenté par g. »



À gauche, cas où M est loin du plan. À droite, cas où M est proche du plan.

Graphiquement, il arrive que l’on représente les individus (projetés sur le plan factoriel) par un disque plutôt que par un point, la taille du disque étant grande si l’individu est bien représenté, petite dans le cas contraire.  
Pour calculer la qualité de la représentation d’un individu, rendez-vous au bas de ce chapitre, dans la section Aller plus loin.

Sur le cercle des corrélations, on s’intéressait aux angles entre les flèches, car ces angles étaient liés à la corrélation entre les variables. Ici, on s’intéresse aux distances entre les individus. En effet, deux individus qui sont proches dans sont similaires, et s’ils sont éloignés, ils sont différents

.

Mais n’oublions pas qu’en représentant les individus sur un plan factoriel, nous visualisons une projection ! Analyser les distances entre 2 points sur un plan factoriel n’est pertinent que si ces individus sont bien représentés tous les 2.

Prenons un exemple. Par une belle nuit d’été, vous regardez les étoiles. Vous avez l’impression qu’elles sont toutes à la même distance, n’est-ce pas ? Vous avez l’impression qu’elles sont toutes situées sur une sphère, que l’on appelle voûte céleste. Mais en réalité, les étoiles ne sont pas toutes à la même distance de vous : vous n’êtes pas le centre de l’univers, quand même !

Si nous les percevons toutes à la même distance, c’est qu’elles sont tellement loin que notre cerveau n’arrive pas à évaluer la distance qui nous sépare d’elles. Pour les objets qui sont proches, le cerveau analyse la différence entre les 2 images perçues par chacun de nos 2 yeux pour estimer les distances. Mais les étoiles sont tellement loin que la [triangulation](https://fr.wikipedia.org/wiki/Triangulation) ne fonctionne plus. Votre cerveau se dit : « Je n’y arrive pas ! Mettons tout à la même distance, et n’en parlons plus ! »

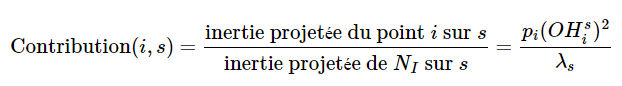
Ainsi, si vous regardez une toute petite surface de la voûte céleste, vous pouvez considérer que celle-ci est (presque) plane. Votre cerveau effectue donc une projection de l’espace (qui a 3 dimensions) vers un plan à 2 dimensions. Sur cette petite surface de la voûte céleste, vous remarquez deux étoiles très proches l’une de l’autre : vous pensez donc qu’elles sont voisines. Mais en réalité, l’une d’elles peut très bien être 100 fois plus lointaine que la seconde ! Voilà ce qui se passe quand on interprète une distance sur des points projetés : on peut faire des erreurs !

En astronomie, c’est cette notion qui différencie les « vraies étoiles doubles » des « [**fausses étoiles doubles**](https://fr.wikipedia.org/wiki/Binaire_visuelle) ».

### Aller plus loin : Contribution d’un individu à l’inertie d’un axe

Lors du calcul des composantes principales, chaque individu est pris en compte. Cependant, certains influent plus que d’autres sur le calcul de certaines composantes.

Il arrive parfois qu’un axe ne soit dû principalement qu’à un tout petit groupe d’individus, voire qu’à un seul. Cela signifie que quelques individus « attirent » fortement un axe dans leur direction, quasiment sans laisser les autres individus contribuer à la formation de cet axe. Sa formulation n’est pas très compliquée :



**λs** est l'inertie projetée du nuage **NI** sur l'axe de rang **s** (par exemple, **F1** si **s=1**). **pi** est le poids associé au point **i** (ici, on considère que tous les points ont le même poids), et  est la distance entre l'origine et l'image du point **i** par projection sur l'axe de rang **s**.

**λs** est fixe quel que soit l’individu. Si, en plus, tous les individus ont le même poids **pi**, alors les différences de contributions entre les individus ne sont contenues que dans le . Comme et  sont visibles sur le premier plan factoriel (ce sont respectivement l’abscisse et l’ordonnée du point i), alors on voit directement les individus qui contribuent le plus à l’axe. Ceux qui sont le plus loin de l’origine à droite ou à gauche du graphique sont ceux qui contribuent le plus à F1. De même, ceux qui sont les plus éloignés de l’origine en haut ou en bas du graphique sont ceux qui contribuent le plus à F2. Comme on voit donc tout sur le graphique, il n’est en fait intéressant de calculer ces contributions que quand les poids **pi** des individus sont différents. En effet, là, le graphique ne peut plus nous fournir cette information.

La somme des contributions de tous les individus à un axe donné est égale à 100 %.

La notion de contribution est aussi valable pour les variables. Les variables qui contribuent fortement à une composante principale sont celles qui sont très corrélées à celle-ci. On le voit donc directement sur le cercle des corrélations.

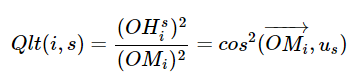
### Aller plus loin : Qualité de représentation d’un individu par un axe

Calculer la qualité de représentation d’un individu sur un plan factoriel (le premier, par exemple), c’est calculer la qualité de représentation du point par l’axe F1, puis par l’axe F2.  
Cette qualité s’exprime par le **pourcentage d’inertie**du point qui est expliqué par l’axe.

Reprenons les notations de ce chapitre, où **Mi** est un point, et où est la projection de **Mi** sur l’axe de rang **s**. Par exemple, **Mi** projeté sur l’axe **F1** donne un point que l’on note.

Il ne s’agit absolument pas de « puissance 1 » ! C’est juste une notation : comme il n’y avait plus de place en bas du H, on a mis le 1 au-dessus, c’est tout. ;)

Comme nous l’avons déjà vu, l’inertie du point **Mi**, c’est .  
Ensuite, l’inertie de **Mi** expliquée par l’axe **F1**, c’est l’inertie de la projection de **Mi** sur **F1**. C’est donc l’inertie du point , soit . Le pourcentage d’inertie de l’individu ii expliqué par l’axe usus est donc :



Cette quantité s’additionne sur plusieurs axes ! Ainsi, le pourcentage d’inertie de l’individu ii expliqué par le premier plan factoriel est de . Et si l'on additionne sur tous les axes principaux d’inertie, on obtient 100 %.

## Choisissez le nombre de composantes

Depuis le début de ce cours, on parle de composantes principales.  
Nous avons vu comment les étudier, mais nous n’avons pas vu combien en étudier : doit-on étudier seulement le premier plan factoriel, ou doit-on aller plus loin ?

### Réfléchissons

Tout d’abord, posons-nous la question suivante :

Combien d’axes d’inertie peut-on trouver ?

Prenons un plan sur lequel vous tracez un axe au hasard. Si vous devez en tracer un deuxième, orthogonal au premier, vous n’avez pas le choix dans la direction : c'est la direction que vous indique votre équerre. Si maintenant on vous demande de tracer un 3e axe orthogonal aux 2 premiers, vous verrez que cette tâche est impossible ! La raison est que, dans un espace à **p** dimensions, vous ne pouvez trouver que **p** axes orthogonaux au maximum.

Mais de combien d’axes a-t-on besoin pour décrire parfaitement **n** individus ?

Plaçons 2 points dans un espace à **p=3** dimensions. Pour capter 100 % de l’inertie de ce nuage de points, un seul axe suffit : ce sera la droite qui passe par ces 2 points. Donc ici, 1 axe suffit pour 2 points.

De même, prenons 3 points dans un espace en 3D. Pour capter 100 % de l’information, il suffit de faire passer un plan par ces 3 points, et un plan c’est en 2D. Ici, 2 axes suffisent donc pour 3 points. Pour décrire parfaitement **n** individus, on a donc besoin au maximum de **n−1** axes.

Mais comme nous avons dit tout à l’heure, on ne peut pas trouver plus de **p** axes.

Si l'on assemble ces deux contraintes, on arrive à cette conclusion :

Le nombre maximal de composantes d’une ACP sera le minimum entre **p** et **n-1.**

Ainsi, si vous avez un échantillon de 2 000 individus décrits par 1 000 variables, vous trouverez 1 000 composantes.

Mais vous vous doutez qu’il ne faudra pas analyser chacune de ces 1 000 composantes. Non, ce serait trop de travail !

### Combien de composantes analyser ?

Alors, combien de composantes analyser ?

Vous savez qu’en ACP, on projette les données sur les axes principaux d’inertie, et que ceux-ci sont ordonnés selon l’inertie du nuage projeté : de la plus grande à la plus petite.

Quand on additionne les inerties associées à tous les axes, on obtient l’inertie totale du nuage des individus.

On peut donc afficher un diagramme qui décrit le pourcentage d’inertie totale associé à chaque axe.

On appelle ce diagramme l’**éboulis des valeurs propres**. En voici un exemple :

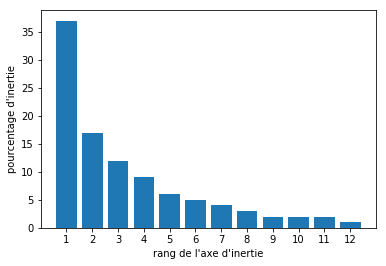


Diagramme des éboulis de valeurs propres

Pourquoi "valeurs propres" ?

Parce que les inerties portées par chaque axe **us** sont égales aux valeurs propres **λs** de la matrice de covariance des données. Nous en avons parlé dans la section Aller plus loin de[**ce chapitre**](https://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees/5278610-calculez-les-composantes-principales).

On peut également afficher la somme cumulée des inerties, c’est une courbe qui part de l’origine et qui arrive à 100 % après avoir parcouru tous les axes :

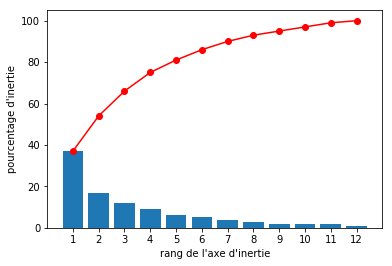


Diagramme des éboulis de valeurs propres avec somme cumulée

Quand on affiche un plan factoriel, on a l’habitude d’écrire l’inertie associée à chacun des 2 axes.

#### Comment interpréter ces pourcentages d’inertie ?

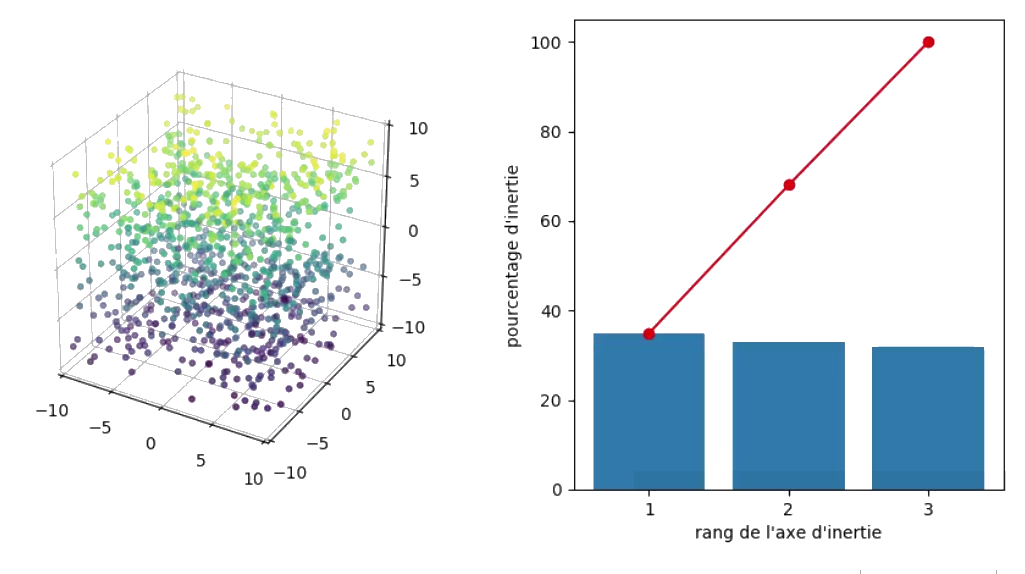
Les pourcentages d’inertie nous donnent une information sur la « structure » de nos données. Prenons les 2 cas extrêmes : des données sans aucune structure et des données très structurées.

##### Cas extrême 1

Un échantillon sans aucune structure est un échantillon pour lequel les variables n’ont aucune corrélation entre elles : elles sont toutes indépendantes deux à deux.

Ainsi, pour un individu donné, on ne peut absolument pas se faire une idée de la valeur d’une variable, même en connaissant la valeur de toutes les autres. C’est assez rare, car cela signifie que les caractéristiques (les variables) qui décrivent vos individus ne sont absolument pas liées entre elles ! Dans ce cas, l’inertie totale est équitablement répartie entre les axes, qui sont tous à (**100/p**) pour cent de l’inertie totale.

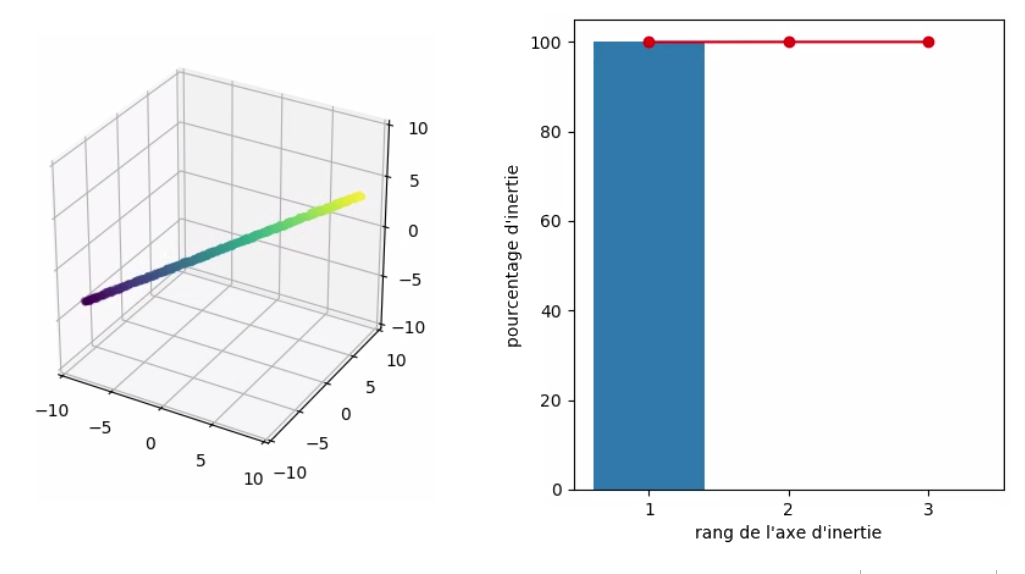
Voici un exemple pour des données en 3 dimensions, où chaque axe est associé à (environ) 100/3 = 33 % de l'inertie totale :



Cas où les 3 variables sont indépendantes

##### Cas extrême 2

L’extrême inverse serait un échantillon pour lequel toutes les variables sont corrélées deux à deux avec un coefficient de corrélation de 1. Dans ce cas, pour un individu donné, il est possible de connaître avec exactitude la valeur d’une variable rien qu’en connaissant la valeur de n’importe quelle autre variable. C’est le cas où tous les points sont parfaitement alignés. Comme tous les points sont alignés, il n’y a besoin que d’un seul axe pour capter 100 % de l’inertie totale : il suffit de le placer dans l'alignement des points !



Cas où toutes les variables sont corrélées deux à deux avec un coefficient de corrélation de 1

Ce second cas est idéal quand on veut réduire les dimensions, car on peut supprimer toutes nos variables initiales pour les remplacer par… une unique variable sans perdre aucune information. Le rêve ! :soleil:

Bref, plus les données auront une structure, mieux on pourra les résumer en ne gardant que les premiers axes principaux d’inertie.

Ainsi, projeter le nuage des individus **NI** sur le premier plan factoriel, c’est bien, mais encore faut-il savoir si cette projection représente bien les données ou pas. Si 75 % de l’inertie totale sont associés à F1, et 25 % à F2, alors c’est parfait : le premier plan factoriel représente 100 % de l’inertie totale. La représentation est donc parfaite, pas besoin d’afficher F3, F4, etc., car ils ne contiendront aucune information (du coup, ils n’existent même pas).

Si, au contraire, on a des données en 10 dimensions, et que 20 % de l’inertie totale sont associés à F1, et 15 % à F2, le premier plan factoriel ne représente alors que 35 % de l’inertie totale. Si l'on considère que 35 % sont trop peu, alors on peut analyser les composantes suivantes (F3, etc.).

#### Finalement, combien de composantes analyser ?

Il n’y a pas de réponse toute faite : tout dépend de vos objectifs et du contexte. Tout dépend jusqu’où vous souhaitez pousser votre analyse.

Cependant, voici quelques pistes.

Tout d’abord, on a tendance à ne pas considérer comme importants les axes dont l’inertie associée est inférieure à **(100/p)%**, car ils représentent moins de variabilité qu’une seule variable initiale. La valeur de **(100/p)%** est celle obtenue quand toutes les variables sont indépendantes deux à deux. Ce critère est appelé **critère de Kaiser**.

Ensuite, il existe la « [méthode du coude](https://en.wikipedia.org/wiki/Elbow_method_(clustering)) » qui consiste à repérer l’endroit à partir duquel le pourcentage d’inertie diminue beaucoup plus lentement lorsque l’on parcourt le diagramme des éboulis de gauche à droite.

En pratique, on a plutôt tendance à s’arrêter dès que l’on n’arrive plus à interpréter les axes principaux d’inertie, c’est-à-dire quand on n’arrive plus à dire « le groupe de variables est très corrélé à **Fs**, et il correspond à la notion de [...] ». Il est même fréquent de n’analyser que le premier plan factoriel.

## TP : Réalisez une ACP

C'est le moment de plonger dans le code !

Nous allons ici analyser les jeux de données que nous avons détaillés dans un [précédent chapitre](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/telechargez-les-jeux-de-donnees-que-nous-etudierons).

Retrouvez tout le code sur le [**git du cours**](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire). ;)

Un souci dans l'exécution du code ? Posez votre question sur le [**forum du cours**](https://openclassrooms.com/forum/sujet/cours-analyse-exploratoire-de-donnees) !

### 

### Les cours OpenClassrooms que vous avez suivis

Dans les précédents chapitres, nous avons illustré le cercle des corrélations et la projection des individus à l'aide du jeu de données des cours OpenClassrooms que j'ai suivis (que vous pouvez aussi télécharger au [chapitre dédié](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/telechargez-les-jeux-de-donnees-que-nous-etudierons)).

Voici donc le code Python qui vous permettra de faire de même avec votre propre échantillon. Ce code est également accessible dans le fichier **pca\_my\_courses.py** sur le git :

import pandas as pd

import numpy as np

from sklearn import decomposition

from sklearn import preprocessing

from functions import \*

# choix du nombre de composantes à calculer

n\_comp = 6

# import de l'échantillon

data = pd.read\_csv("my\_courses.csv",decimal=".",index\_col=0)

# selection des colonnes à prendre en compte dans l'ACP

data\_pca = data[["inscription","progression","moyenneDeClasse","duree","difficulte","nbChapitres","ratioQuizEvaluation","nbEvaluations"]]

# préparation des données pour l'ACP

data\_pca = data\_pca.fillna(data\_pca.mean()) # Il est fréquent de remplacer les valeurs inconnues par la moyenne de la variable

X = data\_pca.values

names = data["idCours"] # ou data.index pour avoir les intitulés

features = data.columns

# Centrage et Réduction

std\_scale = preprocessing.StandardScaler().fit(X)

X\_scaled = std\_scale.transform(X)

# Calcul des composantes principales

pca = decomposition.PCA(n\_components=n\_comp)

pca.fit(X\_scaled)

# Eboulis des valeurs propres

display\_scree\_plot(pca)

# Cercle des corrélations

pcs = pca.components\_

display\_circles(pcs, n\_comp, pca, [(0,1),(2,3),(4,5)], labels = np.array(features))

# Projection des individus

X\_projected = pca.transform(X\_scaled)

display\_factorial\_planes(X\_projected, n\_comp, pca, [(0,1),(2,3),(4,5)], labels = np.array(names))

plt.show()

Pour exécuter ce code, vous aurez besoin des fonctions display\_scree\_plot, display\_circles et display\_factorial\_planes, disponibles dans le fichier functions.py sur le git, ou tout en bas de ce chapitre.

### 

### Le jeu de données mystère

C'est maintenant le moment de découvrir ce que représente l'échantillon mystère. Et pour cela, nous allons projeter les individus sur le premier plan factoriel, comme d'habitude !

import pandas as pd

from sklearn import decomposition, preprocessing

from functions import \*

# choix du nombre de composantes à calculer

n\_comp = 3

# import de l'échantillon

data = pd.read\_csv("mystery.csv")

X = data.values

# Réduire n'est ici pas nécessaire car les variables sont exprimées dans la même unité.

# On se contente juste de centrer les données, ce qui est obligatoire pour une ACP.

X = preprocessing.StandardScaler(with\_std=False).fit\_transform(X)

# Calcul des composantes principales

pca = decomposition.PCA(n\_components= n\_comp)

pca.fit(X)

# Eboulis des valeurs propres

display\_scree\_plot(pca)

# projection des individus

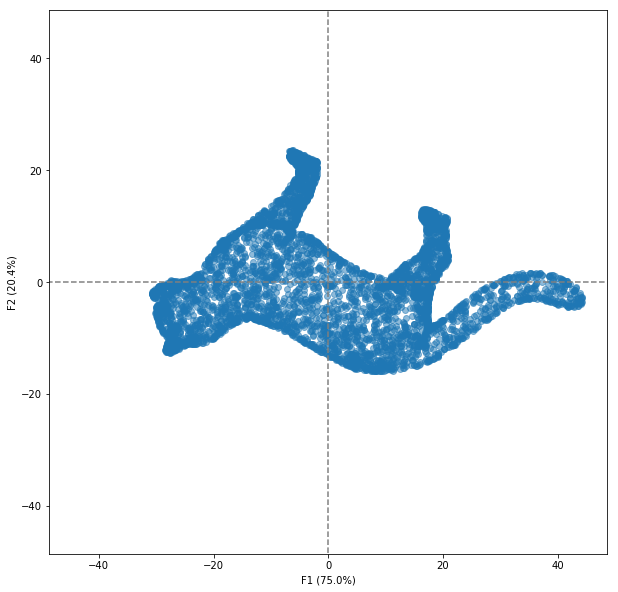
X\_projected = pca.transform(X)

display\_factorial\_planes(X\_projected, n\_comp, pca, [(0,1),(1,2),(0,2)], alpha = 0.2)

plt.show()

Ce code est disponible dans le fichier**pca\_mystery.csv**. Pour l'exécuter, vous aurez besoin des fonctions disponibles dans le fichier functions.py sur le git, ou tout en bas de ce chapitre.

Et voici la projection des individus sur le premier plan factoriel :



Il s'agit d'un chat !

Dans cet échantillon, chaque individu est un point du chat. Bien entendu, vous aurez rarement un échantillon statistique où les points seront disposés selon la forme d'un animal ou d'un objet. En effet, cet échantillon se rapproche plutôt de ce qu'utilisent les personnes qui pratiquent la modélisation 3D (architectes, vidéastes, ingénieurs en mécanique, physiciens, etc.). Mais, depuis le début, nous avons souvent fait le parallèle entre nos nuages de points de statisticiens et l'inertie des objets réels qu'étudient les physiciens. Alors, continuons ici cette analogie.

Ce qui est affiché ici, c'est une projection orthogonale en 2D d'un objet en 3D.

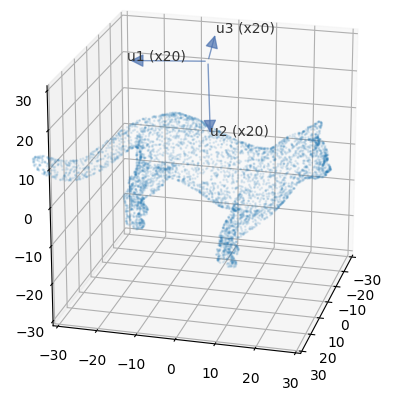
Les 2 graphiques de dispersion que je vous avais donnés dans [ce chapitre](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/telechargez-les-jeux-de-donnees-que-nous-etudierons) étaient aussi des projections 2D de points en 3D. Cependant, sur ceux-ci, nous n'arrivions pas à distinguer la forme du chat. Pourquoi ?

Parce que l'ACP a trouvé la **meilleure projection :** celle qui montre la plus grande inertie. Autrement dit, si vous voulez prendre une photo de votre chat, le meilleur angle pour avoir le plus de détails possibles sera donc de profil, comme vous le voyez ci-dessus.

Sur le premier plan factoriel, on voit que l'axe d'inertie principal d'un chat part du bout de sa queue pour aller jusqu'à sa tête : c'est l'axe des abscisses du graphique.

**Bonus :** dans le fichier **pca\_mystery.csv**, vous trouverez aussi le code pour afficher le chat en 3D, en mode interactif (vous pouvez le tourner grâce à la souris, à condition de ne pas utiliser le notebook, mais d’exécuter le programme en ligne de commande) :

python pca\_mystery.py



### 

### L'échantillon en bag of words

Pour réaliser l'ACP de l'échantillon en bag of words, c'est exactement le même principe. Vous retrouverez ce code dans le fichier **pca\_bag\_of\_words.py** :

import pandas as pd

import numpy as np

from sklearn import decomposition

from sklearn import preprocessing

from functions import \*

# choix du nombre de composantes à calculer

n\_comp = 50

# import de l'échantillon et des informations relatives aux cours

data = pd.read\_csv('bag\_of\_words.csv', index\_col = 0)

courses\_info = pd.read\_csv('courses\_info.csv',index\_col = 0)

# Theme du ou des parcours auxquels appartient le cours (data, developpement, marketing, etc.)

theme = [courses\_info.loc[course\_id, "theme"] for course\_id in data.index]

# préparation des données pour l'ACP

X = data.values

features = data.columns

# Centrage et Réduction

std\_scale = preprocessing.StandardScaler().fit(X)

X\_scaled = std\_scale.transform(X)

# Calcul des composantes principales

pca = decomposition.PCA(n\_components=n\_comp)

pca.fit(X\_scaled)

# Eboulis des valeurs propres

display\_scree\_plot(pca)

# Cercle des corrélations

pcs = pca.components\_

display\_circles(pcs, n\_comp, pca, [(0,1),(2,3),(4,5)])

display\_circles(pcs, n\_comp, pca, [(0,1)], lims=[.0155, .019, 0.053, .057], labels = np.array(features))

display\_circles(pcs, n\_comp, pca, [(2,3)], lims=[-.035, -.026, -.03,-.016], labels = np.array(features))

# Projection des individus

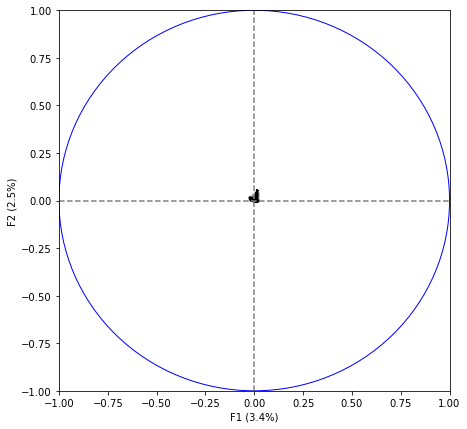
X\_projected = pca.transform(X\_scaled)

display\_factorial\_planes(X\_projected, n\_comp, pca, [(0,1),(2,3),(4,5)], illustrative\_var = theme, alpha = 0.5)

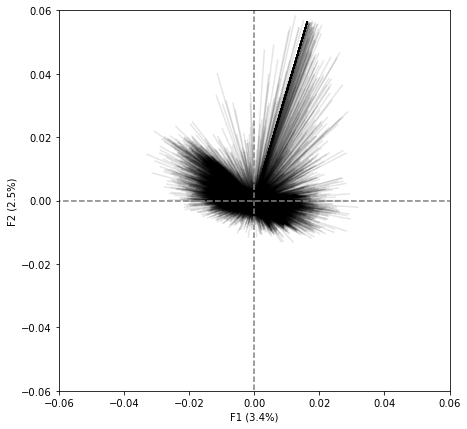
plt.show()

Pour exécuter ce code, vous aurez besoin des fonctions disponibles dans le fichier functions.py sur le git, ou tout en bas de ce chapitre.

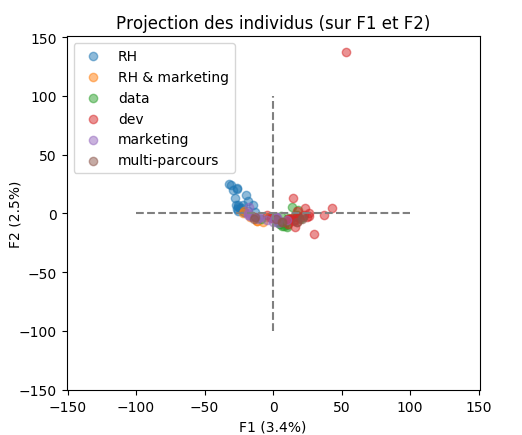
Bon... sauf que là, il y a plus de 9 000 variables. Si l'on affiche le cercle des corrélations avec 9 000 flèches ainsi que leur nom, cela va être le bazar. N'affichons donc pas leur nom ni les extrémités des flèches :



Les flèches sont toutes petites ! Zoomons sur la petite "tache noire" au centre du cercle :

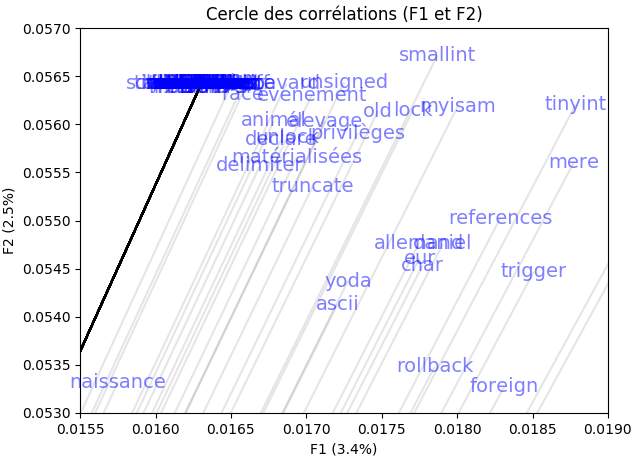


Affichons également la projection des individus, en les colorant selon le thème du cours :



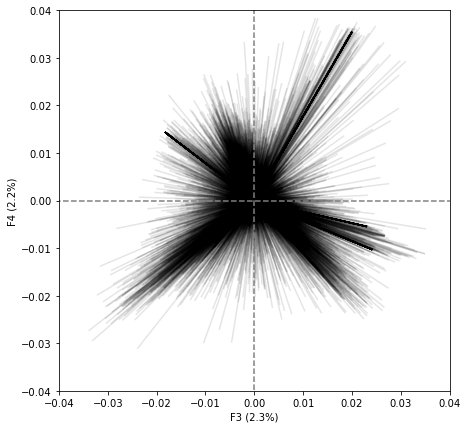
Tout d'abord, on voit que les cours en rouge (cours de développement informatique) ont tendance à se regrouper vers les abscisses importantes, alors que les cours en bleu (cours de ressources humaines) ont tendance à se regrouper vers les abscisses faibles.

De plus, il y a clairement un individu qui se distingue des autres : en haut à droite (en rouge). Pour trouver ce qui le différencie des autres cours, zoomons sur le cercle des corrélations vers les variables qui ont une abscisse et une ordonnée élevées. Pour rappel, chaque variable de cet échantillon correspond à un mot :

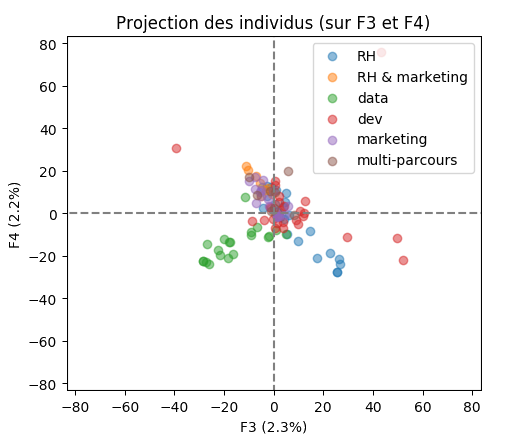


On voit ici que le cours auquel nous nous intéressons a probablement de grandes valeurs pour des variables telles :"trigger", "rollback", "myisam", "smallint", "references". Ceux qui ont déjà utilisé les bases de données auront peut-être reconnu quelques mots-clés propres aux bases de données MySQL ! En vérifiant, on s'aperçoit que l'individu rouge en haut à droite est effectivement le cours [Administrez vos bases de données avec MySql](https://openclassrooms.com/courses/1959476-administrez-vos-bases-de-donnees-avec-mysql) .

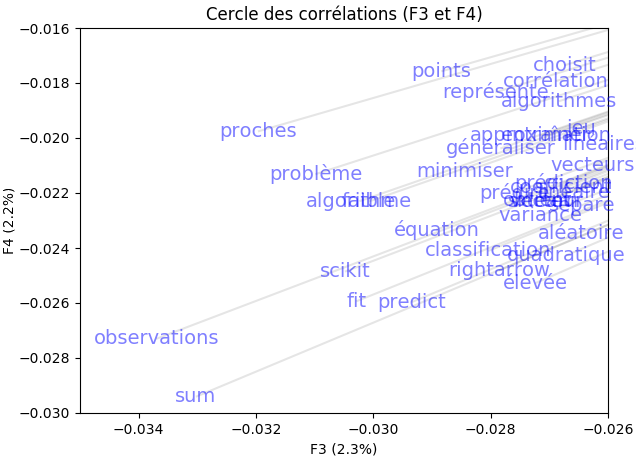
Analysons maintenant le second plan factoriel :



Ici encore, les flèches sont toutes petites, nous avons donc zoomé. Le cercle de rayon 1 n'est donc plus visible.



Intéressons-nous ici aux individus verts, qui ont des abscisses et des ordonnées faibles. En zoomant dans le cercle des corrélations vers les flèches pointant vers des abscisses et des ordonnées faibles, on obtient ceci :



Les termes "observations", "aléatoire", "corrélation", "variance", etc., devraient vous évoquer le domaine des statistiques, n'est-ce pas ? On trouve également des mots-clés propres à la librairie [scikit-learn](http://scikit-learn.org/) (dédiée au machine learning, mais que nous utilisons dans ce cours également) tels "scikit", "fit", "predict". Vous aurez donc deviné qu'il s'agit des cours de data !

### Annexe : les fonctions outils

Voici la définition des fonctions utilisées dans ce chapitre, disponible dans le fichier **functions.py** :

import matplotlib.pyplot as plt

from matplotlib.collections import LineCollection

import numpy as np

import pandas as pd

def display\_circles(pcs, n\_comp, pca, axis\_ranks, labels=None, label\_rotation=0, lims=None):

for d1, d2 in axis\_ranks: # On affiche les 3 premiers plans factoriels, donc les 6 premières composantes

if d2 < n\_comp:

# initialisation de la figure

fig, ax = plt.subplots(figsize=(7,6))

# détermination des limites du graphique

if lims is not None :

xmin, xmax, ymin, ymax = lims

elif pcs.shape[1] < 30 :

xmin, xmax, ymin, ymax = -1, 1, -1, 1

else :

xmin, xmax, ymin, ymax = min(pcs[d1,:]), max(pcs[d1,:]), min(pcs[d2,:]), max(pcs[d2,:])

# affichage des flèches

# s'il y a plus de 30 flèches, on n'affiche pas le triangle à leur extrémité

if pcs.shape[1] < 30 :

plt.quiver(np.zeros(pcs.shape[1]), np.zeros(pcs.shape[1]),

pcs[d1,:], pcs[d2,:],

angles='xy', scale\_units='xy', scale=1, color="grey")

# (voir la doc : https://matplotlib.org/api/\_as\_gen/matplotlib.pyplot.quiver.html)

else:

lines = [[[0,0],[x,y]] for x,y in pcs[[d1,d2]].T]

ax.add\_collection(LineCollection(lines, axes=ax, alpha=.1, color='black'))

# affichage des noms des variables

if labels is not None:

for i,(x, y) in enumerate(pcs[[d1,d2]].T):

if x >= xmin and x <= xmax and y >= ymin and y <= ymax :

plt.text(x, y, labels[i], fontsize='14', ha='center', va='center', rotation=label\_rotation, color="blue", alpha=0.5)

# affichage du cercle

circle = plt.Circle((0,0), 1, facecolor='none', edgecolor='b')

plt.gca().add\_artist(circle)

# définition des limites du graphique

plt.xlim(xmin, xmax)

plt.ylim(ymin, ymax)

# affichage des lignes horizontales et verticales

plt.plot([-1, 1], [0, 0], color='grey', ls='--')

plt.plot([0, 0], [-1, 1], color='grey', ls='--')

# nom des axes, avec le pourcentage d'inertie expliqué

plt.xlabel('F{} ({}%)'.format(d1+1, round(100\*pca.explained\_variance\_ratio\_[d1],1)))

plt.ylabel('F{} ({}%)'.format(d2+1, round(100\*pca.explained\_variance\_ratio\_[d2],1)))

plt.title("Cercle des corrélations (F{} et F{})".format(d1+1, d2+1))

plt.show(block=False)

def display\_factorial\_planes(X\_projected, n\_comp, pca, axis\_ranks, labels=None, alpha=1, illustrative\_var=None):

for d1,d2 in axis\_ranks:

if d2 < n\_comp:

# initialisation de la figure

fig = plt.figure(figsize=(7,6))

# affichage des points

if illustrative\_var is None:

plt.scatter(X\_projected[:, d1], X\_projected[:, d2], alpha=alpha)

else:

illustrative\_var = np.array(illustrative\_var)

for value in np.unique(illustrative\_var):

selected = np.where(illustrative\_var == value)

plt.scatter(X\_projected[selected, d1], X\_projected[selected, d2], alpha=alpha, label=value)

plt.legend()

# affichage des labels des points

if labels is not None:

for i,(x,y) in enumerate(X\_projected[:,[d1,d2]]):

plt.text(x, y, labels[i],

fontsize='14', ha='center',va='center')

# détermination des limites du graphique

boundary = np.max(np.abs(X\_projected[:, [d1,d2]])) \* 1.1

plt.xlim([-boundary,boundary])

plt.ylim([-boundary,boundary])

# affichage des lignes horizontales et verticales

plt.plot([-100, 100], [0, 0], color='grey', ls='--')

plt.plot([0, 0], [-100, 100], color='grey', ls='--')

# nom des axes, avec le pourcentage d'inertie expliqué

plt.xlabel('F{} ({}%)'.format(d1+1, round(100\*pca.explained\_variance\_ratio\_[d1],1)))

plt.ylabel('F{} ({}%)'.format(d2+1, round(100\*pca.explained\_variance\_ratio\_[d2],1)))

plt.title("Projection des individus (sur F{} et F{})".format(d1+1, d2+1))

plt.show(block=False)

def display\_scree\_plot(pca):

scree = pca.explained\_variance\_ratio\_\*100

plt.bar(np.arange(len(scree))+1, scree)

plt.plot(np.arange(len(scree))+1, scree.cumsum(),c="red",marker='o')

plt.xlabel("rang de l'axe d'inertie")

plt.ylabel("pourcentage d'inertie")

plt.title("Eboulis des valeurs propres")

plt.show(block=False)

## Soyez attentif aux spécificités de l'ACP

Discutons un peu de l’ACP. C’est une méthode qui nécessite un peu d’entraînement et pour laquelle il faut être prudent, surtout au début. Je vous donne donc ici quelques points d’attention. Nous présenterons ensuite les inconvénients de cette méthode, et nous verrons comment y remédier.

### 

### Points d'attention

Voici quelques points sur lesquels il faut être prudent, surtout au début.

Il est fréquent d’être un peu perdu dans toutes ces flèches, ces points, les plans factoriels, les axes d’inertie, etc. Cela conduit parfois à des interprétations un peu incertaines ou erronées. Heureusement, vous avez toujours la possibilité de vérifier vos analyses en revenant aux données initiales.

Par exemple, si les variables d’un groupe vous semblent corrélées, alors calculez les coefficients de corrélation entre elles (ou la matrice de corrélation, c’est plus rapide), et vous en aurez le cœur net !

De même, si certains individus vous semblent similaires, car ils ont des abscisses ou des ordonnées à peu près égales sur un plan factoriel (par exemple F1 en abscisses et F2 en ordonnées), alors vérifiez-le sur vos données initiales. Par exemple, si ces individus ont des abscisses similaires, alors prenez les variables fortement corrélées à F1, et vérifiez si vos individus ont des valeurs semblables pour ces variables. Pour ne pas produire d’analyse erronée, essayez de ne pas sortir de ces 2 objectifs que nous avons rappelés tout du long :

1. Étudier la variabilité des individus (leurs ressemblances et leurs différences).
2. Étudier les liaisons entre variables, et trouver de nouvelles variables qui synthétisent les groupes de variables très liées.

### 

### Limites de l’ACP

Depuis le début, nous parlons de corrélations. Cependant, en ACP, nous sommes limités aux corrélations linéaires.

La corrélation linéaire, c’est celle mesurée par **r\_{X,Y}** , coefficient de Pearson (pour vous rafraîchir la mémoire, c’est [**par ici**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees/comment-presenter-les-donnees-en-analyse-bivariee)).

Comme l’ACP utilise ce coefficient **r**, elle ne peut donc mesurer que les liaisons linéaires entre les variables. Pour passer outre ce problème, on peut utiliser l’[ACP avec noyau](https://openclassrooms.com/courses/explorez-vos-donnees-avec-des-algorithmes-non-supervises/plongez-dans-une-variete-mds-et-tsne), ou kernel PCA en anglais.

Le coefficient **r** est très sensible aux outliers : c’est donc le cas aussi pour l’ACP. Avantage ou inconvénient ? Tout dépend de la situation, mais si vous affichez les plans factoriels, vous verrez très facilement les outliers : sur l’un des axes principaux (F1, F2, etc.), ils se distingueront beaucoup des autres.

En fait, nous en avons déjà parlé dans [ce chapitre](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/representez-les-individus-sur-les-plans-fatoriels), quand nous parlions des contributions des individus à un axe. Nous avons dit que, parfois, un axe d’inertie n’était dû qu’à un petit groupe d’individus (ou même à un seul) : ce sont les outliers. Un individu situé très loin de tous les autres a tendance à « attirer » dans sa direction l’un des axes d’inertie (bien souvent le premier).  
Si cet outlier (ou ce groupe d’outliers) ne présente pas d’intérêt dans votre analyse, alors il suffit de ne pas analyser l’axe auquel il contribue fortement, et de n’analyser que les autres.

Attention à ne faire ceci que si l’outlier (ou le groupe d’outliers) contribue à quasiment 100 % à cet axe.

La non-robustesse de l’ACP aux outliers n’est donc pas vraiment problématique. Au contraire, elle permet de détecter facilement les outliers, et qu’ils soient intéressants ou non, et de rapidement « passer à autre chose » sans avoir à relancer le calcul de l’ACP.

Autre inconvénient de l’ACP : elle ne se limite qu’aux variables quantitatives ! Heureusement, il existe d’autres méthodes factorielles permettant de remédier à cela, comme l'[Analyse des Correspondances Multiples](https://fr.wikipedia.org/wiki/Analyse_des_correspondances_multiples) pour des variables qualitatives, ou l'[Analyse Factorielle des Données Mixtes](https://fr.wikipedia.org/wiki/Analyse_factorielle_de_donn%C3%A9es_mixtes).

# Pratiquez l'ACP

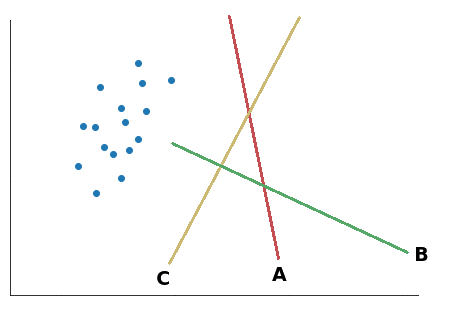
Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Interpréter une ACP

### Question 1

**Lequel de ces 3 axes contiendra le plus d'inertie lorsque l'on projette sur lui le nuage de points ?**

****

* + 

L'axe A

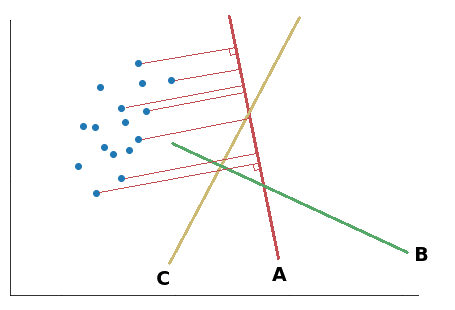
* + 

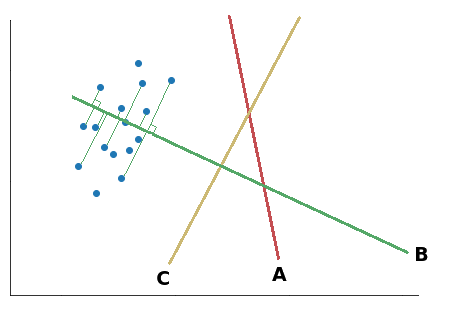
L'axe B

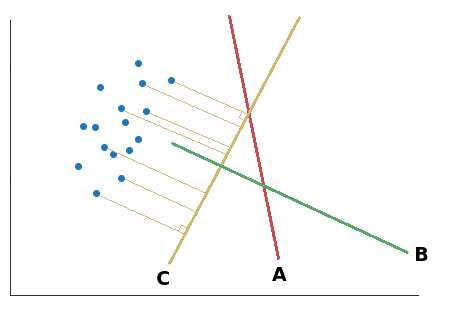
* + 

L'axe C

*Voici les 3 projections différentes. La projection qui est globalement la plus "étalée" (celle qui a le plus d'inertie) est celle de l'axe C. C'est en quelques sortes la "direction générale" des points.*

**

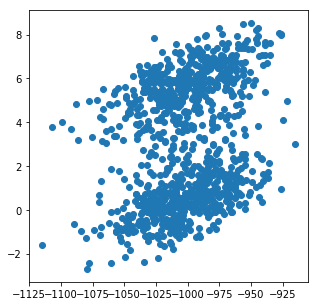
**

**

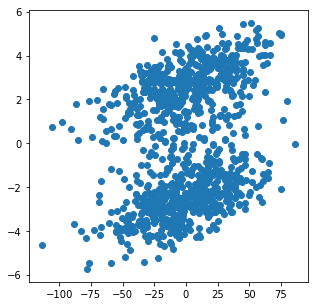
### Question 2

**Lequel de ces 4 diagrammes représente des données centrées-réduites ?**

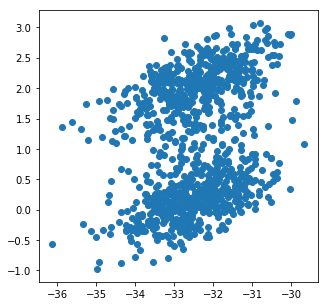
* + 



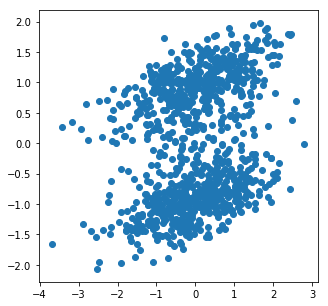
* + 



* + 



* + 



*Le graphique 1 donne les données originelles (ni centrées ni réduites). Le graphique 2 présente des données centrées, car la moyenne des points en abcisses et en ordonnées est de 0. Le graphique 3 représente des données réduites car leur écart-type est de 1 (Certes, leur étendue est supérieure à 1, mais leur écart-type est bien de 1 ;) ). Le graphique 4 combine les caractéristiques du 2 et du 3, les données sont donc centrées réduites*

### Question 3

**Dans le cadre d'une ACP, si votre échantillon contient**nn**individus décrits par**pp**variables quantitatives, comment représente-t-on le nuage des individus ?**

* + 

Par un nuage de nn points dans l'espace RnRn

* + 

Par un nuage de nn points dans l'espace RpRp

* + 

Par un nuage de pp points dans l'espace RnRn

* + 

Par un nuage de pp points dans l'espace RpRp

### Question 4

**Dans le cadre d'une ACP, si votre échantillon contient n individus décrits par p variables quantitatives, comment représente-t-on le nuage des variables ?**

* + 

Par un nuage de pp points dans l'espace RnRn

* + 

Par un nuage de nn points dans l'espace RnRn

* + 

Par un nuage de  nn points dans l'espace RpRp

* + 

Par un nuage de pp points dans l'espace RpRp

### Question 5

**Dans le cadre d'une ACP, si votre échantillon contient n individus décrits par p variables quantitatives, le diagramme appelé cercle des corrélations est :**

* + 

La projection du nuage des variables d'un espace à nn dimensions RnRn vers un espace à 2 dimensions R2R2 , où les 2 dimensions sont des axes principaux d'inertie.

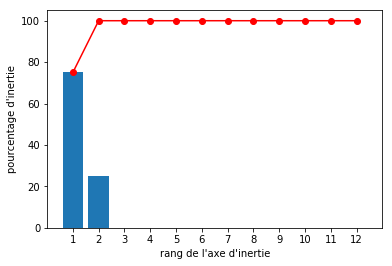
* + 

La projection du nuage des variables d'un espace à pp dimensions RpRp vers un espace à 2 dimensions R2R2 , où les 2 dimensions sont des axes principaux d'inertie.

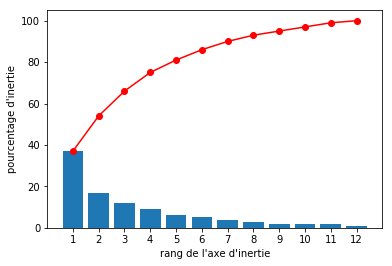
### Question 6

**En analysant ces 2 éboulis de valeurs propres, désignez le cas dans lequel le premier plan factoriel représentera le plus de variance.**

* + 



* + 



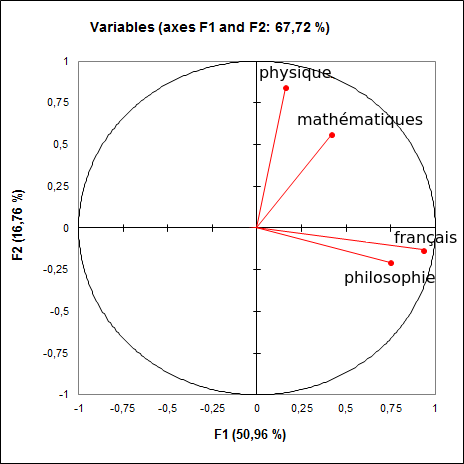
*Dans le cas 1, les 2 premières composantes contiennent 100% de la variance. Dans le second cas, elles n'en contiennent qu'un peu moins de 60%.*

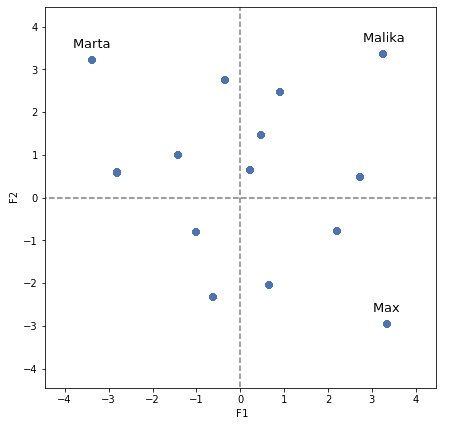
### Question 7

**Un échantillon contient 14 élèves d'une classe de lycée, chacun d'eux étant caractérisé par sa note obtenue aux examens dans des matières telles : philosophie, physique, mathématiques, français, etc. Ils sont notés sur une échelle de 0 à 20, 20 étant la meilleure note.**

**Après avoir calculé l'ACP, voici le cercle des corrélations ainsi que la projection des individus sur le premier plan factoriel.**

**Seules 4 matières sont représentées sur le cercle. Par soucis des simplicité, les matières non représentées ne sont ni des matières littéraires ni des matières scientifiques.**

****

****

**Trouvez la réponse correcte :**

* + 

L'axe F1 mesure la performance dans les matières scientifiques, et F2 la performance dans les matières littéraires. De plus, il semble que :

* + - Marta soit parmi les meilleurs élèves dans les matières littéraires mais parmi les moins bons en sciences
    - Malika soit parmi les meilleurs élèves dans les 4 matières à la fois
    - Max soit parmi les meilleurs en sciences, mais a eu des notes inférieures à la moyenne dans les matières littéraires
  + 

L'axe F1 mesure la performance dans les matières littéraires, et F2 la performance dans les matières scientifiques. De plus, il semble que :

* + - Marta soit parmi les meilleurs élèves dans les matières scientifiques mais parmi les moins bons dans les matières littéraires
    - Malika soit parmi les meilleurs élèves dans les 4 matières à la fois
    - Max soit parmi les meilleurs dans les matières littéraires, mais avec des notes inférieures à la moyenne en sciences

*Les projection de Francais et Philosophie sur F1 sont grandes mais petites sur F2, c'est l'inverse pour Mathématiques et Physique.*

## Recherchez une bonne partition

Dans la partie précédente, nous avons travaillé à synthétiser les variables, c'est-à-dire à réduire le nombre de colonnes de notre tableau de données.

Dans cette 3e partie, nous allons regrouper les lignes, c'est-à-dire que nous allons créer des groupes d'individus : nous allons partitionner les données !

Voici un petit exemple. Nous souhaitons regrouper les individus en 3 groupes.

À l’œil, on voit ces 3 groupes : l'un est très distinct, et les 2 autres se superposent un peu :



Au niveau du vocabulaire, les termes **partitionnement** et **clustering** sont synonymes. Aussi, les **groupes** recherchés sont souvent appelés des **classes**. Quand tous les individus sont affectés à l'un des groupes, on a une **partition**.

Déterminer des groupes peut être l'objectif premier, mais on peut aussi vouloir réduire la taille de notre jeu de données en regroupant certaines lignes.

Dans l'échantillon ci-dessus, on a 150 individus répartis en 3 groupes différents. Chaque individu est caractérisé par **p** variables. Au sein d'un même groupe, les individus sont à peu près similaires, car ils sont proches dans l'espace.

Comme les individus d'un même groupe sont similaires, on peut parfois se contenter d'étudier les caractéristiques du groupe plutôt que d'étudier chacun des individus qui le composent.

Autrement dit, cela revient à étudier les caractéristiques de "l'individu moyen" de chaque groupe.

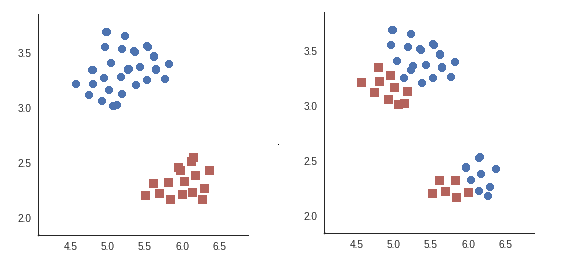
Ainsi, on peut se ramener à un tableau à **p** colonnes (les **p** variables) et à 3 lignes. Chaque ligne correspond à l'individu moyen des 3 groupes.

En faisant cela, on perd forcément de l'information, car même si 2 individus d'un même groupe sont similaires, ils ne sont pas totalement identiques : il y a de la variabilité au sein des groupes.

Le challenge sera donc de trouver des méthodes qui font perdre le moins d'information possible ! Il nous faut donc un critère pour évaluer la qualité d'une partition.

### Évaluer la qualité d'une partition

Reprenons nos données, et déterminons 2 partitions différentes, chacune en 2 classes (une classe bleue et une classe rouge) :



2 partitions (en 2 classes chacune)

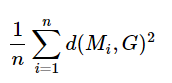
Qu'en pensez-vous ? La partition de gauche paraît meilleure que celle de droite, n'est-ce pas ? Pourquoi ? Parce que, sur l'image de droite, les points des groupes bleus et rouges sont chacun très étalés. De plus, les groupes bleus et rouges sont très proches l'un de l'autre.

Au contraire, sur l'image de gauche, le groupe rouge et le groupe bleu sont chacun peu étalés, et ils sont loin l'un de l'autre.

Je pense que, maintenant, vous avez le réflexe : dès je parle de points "étalés", vous pensez **inertie !**

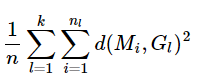
Bingo ! En fait, on ne va pas mesurer l'inertie du nuage de points tout entier, mais celle des différents nuages correspondant à chaque groupe.

Rappelons-nous la formule de l'inertie d'un nuage de points que nous avons vue [précédemment](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/representez-vos-donnees-dans-un-espace) :



où **n** est le nombre d'individus, et  la distance entre le point ii et le centre de gravité **G** du nuage.

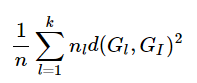
On peut calculer cette quantité pour chaque groupe, puis les additionner (en leur donnant un poids proportionnel au nombre d'individus qu'ils contiennent). On obtient ainsi l**'inertie intraclasse :**

****

où **k** est le nombre de clusters, **nl** le nombre d'individus dans le cluster **l**, et **Gl** le centre de gravité du cluster **l**.

On l'appelle aussi **variance intraclasse** ou **variation intraclasse.**

Ensuite, pour voir si les groupes sont éloignés les uns des autres, on s'imagine un nouveau nuage de points qui est composé uniquement des centres de gravité des différents groupes. Si ces groupes sont éloignés, alors leurs centres de gravité respectifs le seront aussi. L’inertie de ce nouveau nuage imaginaire sera donc grande. On l'appelle inertie interclasse, et elle est donnée par :



où **nl** est l'effectif de la classe **l**, **Gl** le centre de gravité du cluster **l**, et **G** le centre de gravité du nuage tout entier.

On l'appelle aussi **variance interclasse** ou **variation interclasse**.

Rappelons-nous les critères d'une bonne partition ; on veut que les groupes (aussi appelés **clusters**) soient :

1. Resserrés sur eux-mêmes : deux points qui sont proches devraient appartenir au même groupe.
2. Loin les uns des autres, c'est-à-dire qu'ils soient fortement différenciés.

Le premier critère correspond à une inertie inertie intraclasse faible, et le second critère correspond à une inertie interclasse forte.

En fait, on se rend compte que chercher à minimiser l'inertie intraclasse est mathématiquement équivalent à maximiser l'inertie interclasse. Donc pas besoin de se casser la tête avec les 2 critères, car ils sont équivalents !

Cela est dû au fait que : inertie totale = inertie intraclasse + inertie interclasse

Or, quelle que soit la partition, l'inertie totale est constante. Minimiser l'inertie intraclasse est donc équivalent à maximiser l'inertie interclasse ! Cette formule est appelée Théorème de Huygens ou équation d'analyse de la variance.

Voilà ! Maintenant que nous savons qu'il faut minimiser l'inertie intraclasse, voyons 2 méthodes de partitionnement qui peuvent poursuivre cet objectif !

## Découvrez l’algorithme k-means

Commençons avec l'algorithme le plus célèbre en clustering : l'algorithme **k-means**, ou algorithme des centres mobiles en français.

### Fonctionnement de l'algorithme

Vous allez voir, il est très intuitif et facile à comprendre.

Il faut tout d'abord déterminer combien de groupes on souhaite trouver : on appelle ce nombre **K**.

L'objectif : trouver des groupes en faisant en sorte de minimiser l'inertie intraclasse.

L'algorithme du k-means travaille avec les centres de gravité des groupes. Seulement voilà, au départ, on ne connaît pas les groupes ! On ne peut donc pas connaître leur centre de gravité (que l'on appelle souvent des **centroïdes**). Mais ce n'est pas grave : on va supposer qu'on les connaît quand même, et les placer aléatoirement dans l'espace. Enfin... pas trop aléatoirement quand même : on va plutôt "piocher" au hasard des points qui font partie du nuage, pour y placer les centroïdes.

Une fois les centroïdes placés, on prend chaque point du nuage et on lui associe le cluster du centroïde dont il est le plus proche. On obtient donc un nuage de points dont chaque point appartient à l'un des K groupes.

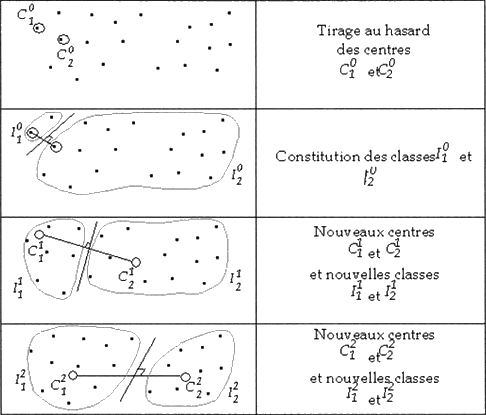
Une fois cette opération faite, on calcule le centre de gravité de chaque groupe. Pour un groupe donné, le centre de gravité n'est généralement pas exactement au même endroit que là où l'on avait placé le centroïde initialement. On déplace donc ce centroïde sur le nouveau centre de gravité calculé.

Mais comme le centroïde a bougé, il faut recalculer, pour chacun des points du nuage, quel est le centroïde le plus proche, pour l'associer au bon cluster. Puis on recalcule le nouveau centre de gravité, et ainsi de suite... jusqu'à la convergence de l'algorithme.

Convergence ?

Ici, on dit que l'algorithme converge quand plus rien ne bouge d'une itération à l'autre, c'est-à-dire quand le centroïde reste immobile même après avoir recalculé pour chaque point son cluster.

Rien de tel qu'un petit schéma pour illustrer tout cela !



Source : Data mining et statistique décisionnelle : l'intelligence des données, Stéphane Tufféry

En version vidéo, voilà ce que ça donne :

(source : [**https://www.youtube.com/watch?v=5I3Ei69I40s)**](https://www.youtube.com/watch?v=5I3Ei69I40s))

Et pour comprendre encore mieux, rien de mieux que de manipuler l'algorithme, itération après itération ;) : c'est [**par ici**](https://kkevsterrr.github.io/K-Means/) !

#### Avantages et inconvénients du k-means

L'algorithme du k-means converge en général très rapidement : il n'est pas rare qu'il atteigne la convergence au bout de 10 itérations, même avec beaucoup de points.

Cependant, il faut faire attention à un point. Avec votre échantillon, il y aura toujours (pour un nombre de clusters donné) une partition qui donnera une inertie intraclasse minimale, mais le k-means ne la trouvera pas toujours. En effet, cela dépend de l'initialisation des centroïdes : comme on les place aléatoirement, le résultat donné peut être différent quand on relance l'algorithme. Il est donc conseillé de relancer plusieurs fois l'algorithme pour sélectionner la partition dont l’inertie intraclasse sera la plus petite.

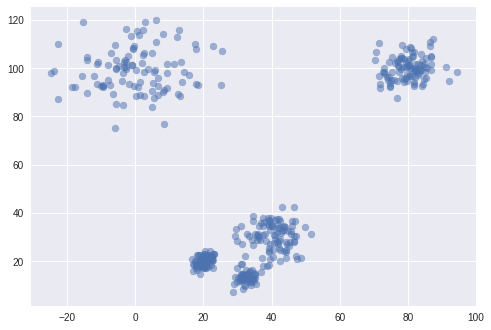
Malheureusement, le k-means n'est pas capable de déterminer le nombre de classes optimal : on est obligé de le lui spécifier au départ. Si on lui demande de trouver 3 clusters alors que vos données sont très clairement regroupées en 5 clusters, alors le k-means vous donnera 3 clusters, même si ce n'est visiblement pas la solution la meilleure.

Pour choisir le nombre optimal de clusters, on peut aussi lancer le k-means plusieurs fois, avec différentes valeurs de **K**. Pour chacune d'entre elles, on note l’inertie intraclasse obtenue. Bien entendu, l’inertie intraclasse diminue forcément quand K augmente, mais en observant la valeur de K au-delà de laquelle la diminution est plus faible, on peut déterminer une bonne valeur de K. C'est la [méthode du coude](https://en.wikipedia.org/wiki/Elbow_method_(clustering)).

## Effectuez une classification hiérarchique

Nous allons voir un autre type de partitionnement : la classification hiérarchique.

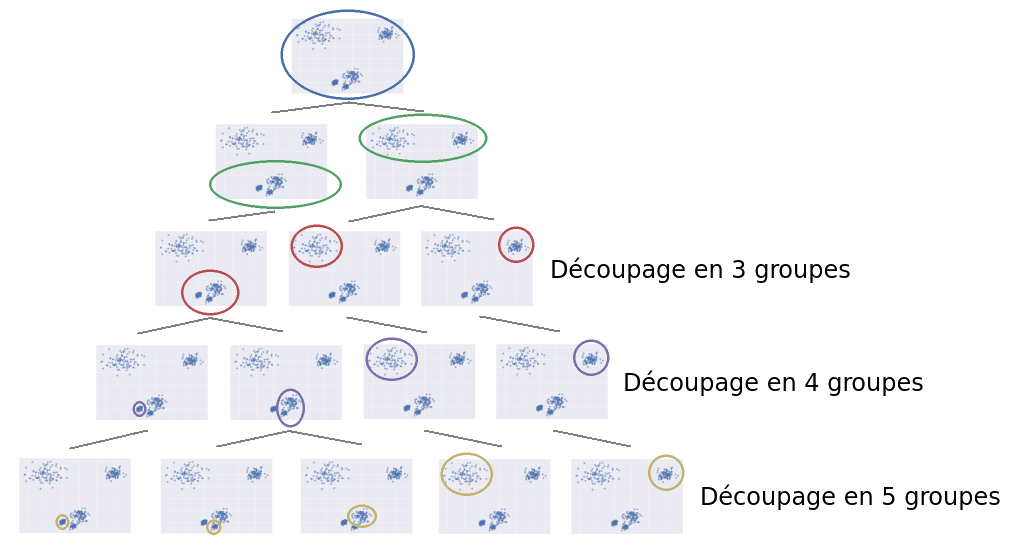
Regardez ces données : combien de clusters y voyez-vous ?



Certains diront 5, d'autres 4, ou encore 3. En effet, il y a 5 groupes bien distincts, mais deux d'entre eux sont très rapprochés : on peut donc considérer qu'ils ne forment qu'un seul cluster et répondre 4. Ces deux clusters très proches sont également relativement proches d'un troisième : si on les regroupe, on considère donc qu'il y a 3 clusters.

En fait, il n'y a pas vraiment de bonne réponse : tout dépend du contexte et des objectifs poursuivis.

On comprend ici l'aspect hiérarchique : ceux qui auront répondu 3 ont eu une vue très générale, alors que ceux qui ont répondu 5 ont préféré avoir une analyse plus fine, plus en profondeur. C'est donc le niveau de profondeur qui détermine une hiérarchie : avec une analyse générale, on voit 3 clusters, mais si l'on creuse un peu plus, on peut diviser l'un de ces 3 clusters en 2 sous-clusters. En allant un peu plus loin, on peut encore diviser l'un de ces 2 sous-clusters en 2 autres. Cette notion de "sous-clusters" est résumable par un arbre :



La notion de profondeur est visible ici à travers l'axe vertical. En effet, on peut "couper" l'arbre plus ou moins haut. Si on le coupe peu profondément, on trouve 3 clusters. Si l'on descend un peu plus bas, on en trouve 5.

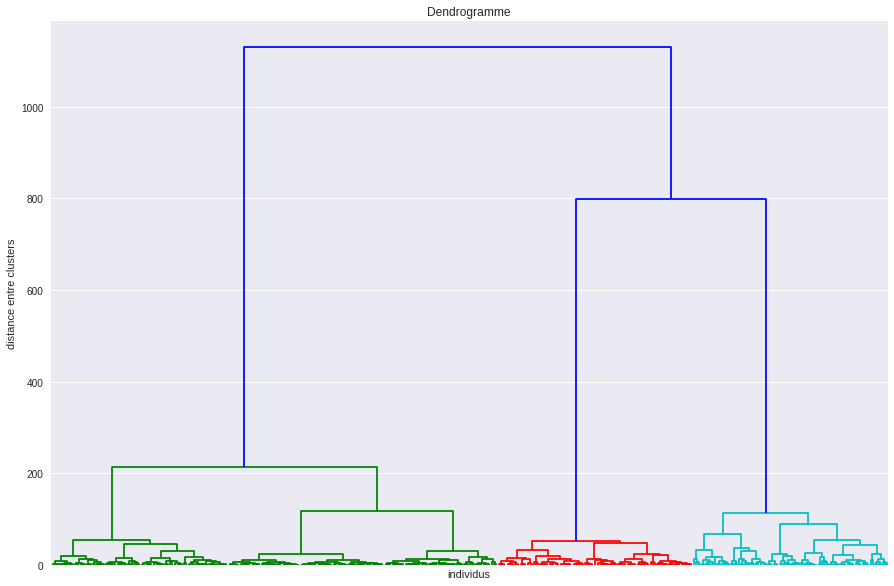
### Classification hiérarchique ascendante et descendante

Il y a deux approches pour créer une partition hiérarchique :

* L'approche **ascendante**, aussi appelée clustering agglomératif : on considère tout d'abord que chaque point est un cluster. Il y a donc autant de clusters que de points. Ensuite, on cherche les deux clusters les plus proches, et on les agglomère en un seul cluster. On répète cette étape jusqu'à ce que tous les points soient regroupés en un seul grand cluster.
* L'approche **descendante**, aussi appelée clustering divisif : c'est l'inverse. On part d'un grand cluster contenant tous les points, puis on le divise successivement jusqu'à obtenir autant de clusters que de points.

On obtient donc une arborescence qui a un cluster tout en haut, et qui se divise petit à petit jusqu'à avoir autant de clusters que de points. On appelle cette arborescence un **dendrogramme**.

Voici le dendrogramme des données ci-dessus :



Si vous avez beaucoup de points, le bas du dendrogramme risque d'être illisible, c'est pourquoi on ne représente parfois que le haut de l'arborescence.

### Méthode des liens et méthode de Ward

Que ce soit avec l'approche ascendante ou descendante, on a besoin de mesurer la distance entre 2 clusters.

La distance entre 2 points, c'est assez intuitif. Mais comme un cluster est composé de plusieurs points, il y a différentes manières de considérer une distance entre 2 clusters. On les appelle les **méthodes de lien** (linkage methods), car ce sont elles qui permettent de lier les clusters lorsque l'on construit petit à petit l'arborescence.

On a :

* Le**lien simple** (simple linkage) : on considère que la distance entre 2 clusters est la distance entre leurs 2 points les plus proches. Cela est équivalent à dire "deux clusters sont proches si au moins deux de leurs points sont proches".
* Le **lien complet** (complete linkage) : on considère que la distance entre 2 clusters est la distance entre leurs 2 points les plus éloignés. Cela équivaut donc à dire "deux clusters sont proches si tous leurs points sont proches".
* Le **lien moyen :** on considère que la distance entre 2 clusters est la moyenne de toutes les distances entre les points d'un cluster et les points de l'autre cluster. Pour la calculer, on énumère toutes les paires de points possibles d'un cluster à l'autre, puis on calcule la distance de chaque paire, puis on calcule la moyenne.
* Le**lien centroïdal :** on considère que la distance entre 2 clusters est la distance entre les centroïdes de ceux-ci.

Les méthodes de lien permettent de garantir que les clusters sont bien séparés, car on prend en compte la distance entre les clusters. Cependant, elles ne garantissent pas que les clusters sont resserrés sur eux-mêmes (c'est la notion d'inertie intraclasse). Pour résoudre cela, il existe la [méthode de Ward](https://fr.wikipedia.org/wiki/M%C3%A9thode_de_Ward) qui, à chaque itération, c'est-à-dire à chaque fois que 2 clusters sont regroupés en 1, cherche à minimiser l'augmentation d'inertie intraclasse due au regroupement des 2 clusters. Par défaut, on préfère en général utiliser la méthode de Ward.

### Avantages et inconvénients

Contrairement au k-means, la classification hiérarchique ne nécessite pas de déterminer un nombre de classes au préalable. En effet, en jouant sur la profondeur de l'arbre, on peut explorer différentes possibilités et choisir le nombre de classes qui nous convient le mieux.

Ce choix peut se faire en regardant les inerties intraclasse et interclasse, ou en utilisant d'autres critères. Nous pouvons choisir un nombre de classes en observant le dendrogramme.

Cependant, il y a des chances pour que l'algorithme de classification hiérarchique mette du temps avant de vous fournir son résultat. En effet, à chaque itération (c'est-à-dire à chaque fois que l'on divise un cluster en 2 pour l'approche descendante, ou que l'on regroupe 2 clusters pour l'approche ascendante), il faut recalculer les distances de toutes les paires de points possibles entre les 2 clusters en question !

Cela nécessite beaucoup de temps et beaucoup d'espace mémoire. On dit que cet algorithme a une forte [*complexité algorithmique*](https://openclassrooms.com/courses/decouvrez-le-fonctionnement-des-algorithmes/comprenez-la-complexite-algorithmique) en temps et en espace, c'est-à-dire que vous manquerez peut-être de temps ou d'espace mémoire, même avec un jeu de données de taille moyenne. Le clustering hiérarchique est donc plus adapté aux petits échantillons.

Oui, mais tu nous disais que le clustering pouvait justement servir à réduire les dimensions du jeu de données (en regroupant des lignes) si celui-ci est trop gros. Si l'on ne peut pas utiliser la classification hiérarchique sur un gros échantillon, à quoi cela sert-il ?

Effectivement. Mais il y a une petite astuce qui permet de combiner les avantages d'un algorithme rapide, mais qui offre peu de flexibilité comme le k-means, avec le confort d'une classification hiérarchique. Si par exemple vous avez un échantillon contenant 10 000 individus, vous pouvez commencer par les regrouper en 50 clusters, en extraire les 50 centroïdes, puis appliquer un clustering hiérarchique sur ces 50 centroïdes. Avec le dendrogramme, vous pourrez ainsi choisir le nombre de clusters (compris entre 1 et 50) qui vous convient !

## Interprétez votre partition

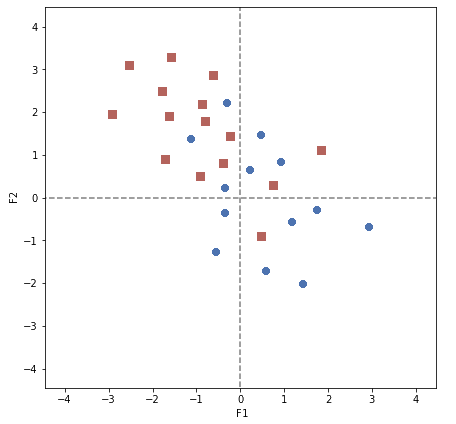
Nous avons vu comment partitionner nos données. Mais pour présenter les données, il serait bien d’avoir une représentation graphique adaptée, n’est-ce pas ?

« Oui, bien sûr », me répondrez-vous. Ici, nous sommes face au même problème qu’au début de la partie 2 de ce cours : nous voulons représenter des données à **p** dimensions de manière intelligible.

Après le clustering, nos données sont toujours à **p** dimensions. La seule chose qui a changé, c’est que nous avons attribué à chaque point une classe (un cluster).

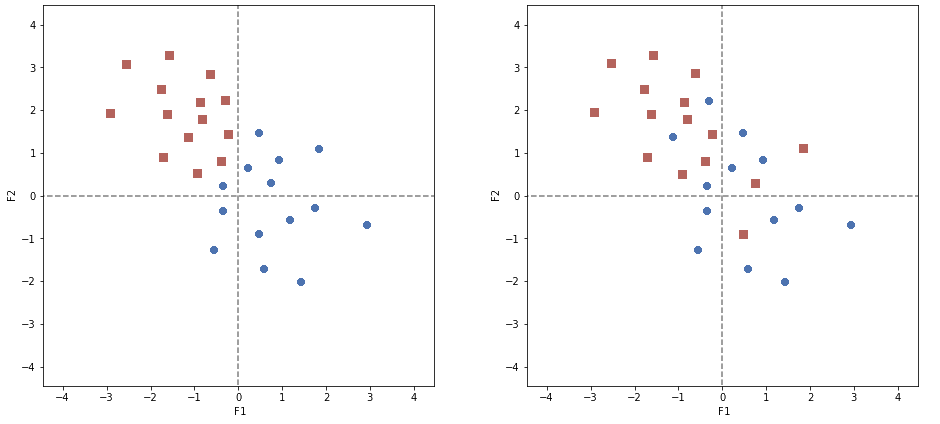
Comme nous avons le même problème qu’à la partie 2, nous pouvons de la même manière utiliser une ACP !

Ainsi, nous affichons les points dans le premier plan factoriel. Ensuite, pour représenter les classes que nous avons déterminées, il suffit de colorer les points en fonction de leur groupe (ou alternativement d’afficher les points d’une même classe avec un même symbole : carré, rond, etc., ce qui est plus pratique pour les personnes souffrant de daltonisme) :



Projection d'une partition sur le premier plan factoriel

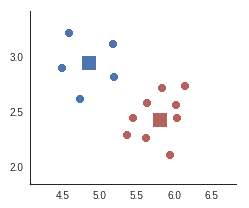
Effectuer un clustering sur les points, puis les projeter sur le premier plan factoriel n’est pas équivalent à effectuer un clustering sur la projection des points sur ce plan ! En effet, dans le premier cas, vous faites votre clustering sur des points à p dimensions (p>2), alors que, dans le second, vous faites un clustering sur des points à 2 dimensions. Dans le second cas, vous avez perdu de l’information lors de la projection, le résultat du clustering n’est donc pas forcément identique. Il se peut donc que l’affichage des couleurs soit donc contre-intuitif :



À gauche, le résultat du k-means APRÈS avoir projeté les données sur le premier plan factoriel. À droite, application du k-means, PUIS projection sur le premier plan factoriel : le résultat du clustering est visuellement moins intuitif.

Et nos centroïdes ?

Les centroïdes de chaque classe sont eux aussi des points à pp dimensions. Nous pouvons donc aussi les représenter sur le premier plan factoriel :



Les centres de classes sont représentés sous forme de carrés.

Attention, cependant, à ne pas prendre en compte les centroïdes lors du calcul des axes principaux d’inertie. En effet, seuls les individus doivent entrer en compte dans ce calcul, pas les centroïdes. Il faut donc juste se contenter de calculer la projection de chaque centroïde sur le premier plan factoriel pour l’afficher.

## TP : Partitionnez vos données

Appliquons la classification hiérarchique et l'algorithme k-means sur 2 de nos échantillons : bag\_of\_words.csv, ainsi que notre chat mystery.csv.

Retrouvez tout le code sur[**le git du cours**](https://github.com/OCCourses/realisez-une-analyse-de-donnees-exploratoire). ;)

Un souci dans l'exécution du code ? Posez votre question sur le [**forum du cours**](https://openclassrooms.com/forum/sujet/cours-analyse-exploratoire-de-donnees).

### L'échantillon en bag of words

Nous allons ici réaliser une classification hiérarchique sur l'échantillon en bag of words, que vous pouvez télécharger à [ce chapitre](https://openclassrooms.com/courses/realisez-une-analyse-de-donnees-exploratoire/telechargez-les-jeux-de-donnees-que-nous-etudierons).

La particularité de cet échantillon est qu'il contient énormément de variables : plus de 9 000 !

Partitionner, c'est regrouper les individus similaires. Dans cet échantillon, les individus sont des cours OpenClassrooms. Chacun est décrit par plus de 9 000 variables, qui correspondent chacune à un mot. Ainsi, deux cours seront similaires si leurs textes contiennent beaucoup de mots en commun.

Voici le code de la classification hiérarchique :

import pandas as pd

from functions import plot\_dendrogram

from scipy.cluster.hierarchy import linkage, fcluster

from sklearn import preprocessing

# import de l'échantillon et des informations relatives aux cours

data = pd.read\_csv('bag\_of\_words.csv', index\_col = 0)

courses\_info = pd.read\_csv('courses\_info.csv',index\_col = 0)

# Theme du ou des parcours auxquels appartient le cours (data, developpement, marketing, etc.)

theme = [courses\_info.loc[course\_id, "theme"] for course\_id in data.index]

# préparation des données pour le clustering

X = data.values

names = data.index

# Centrage et Réduction

std\_scale = preprocessing.StandardScaler().fit(X)

X\_scaled = std\_scale.transform(X)

# Clustering hiérarchique

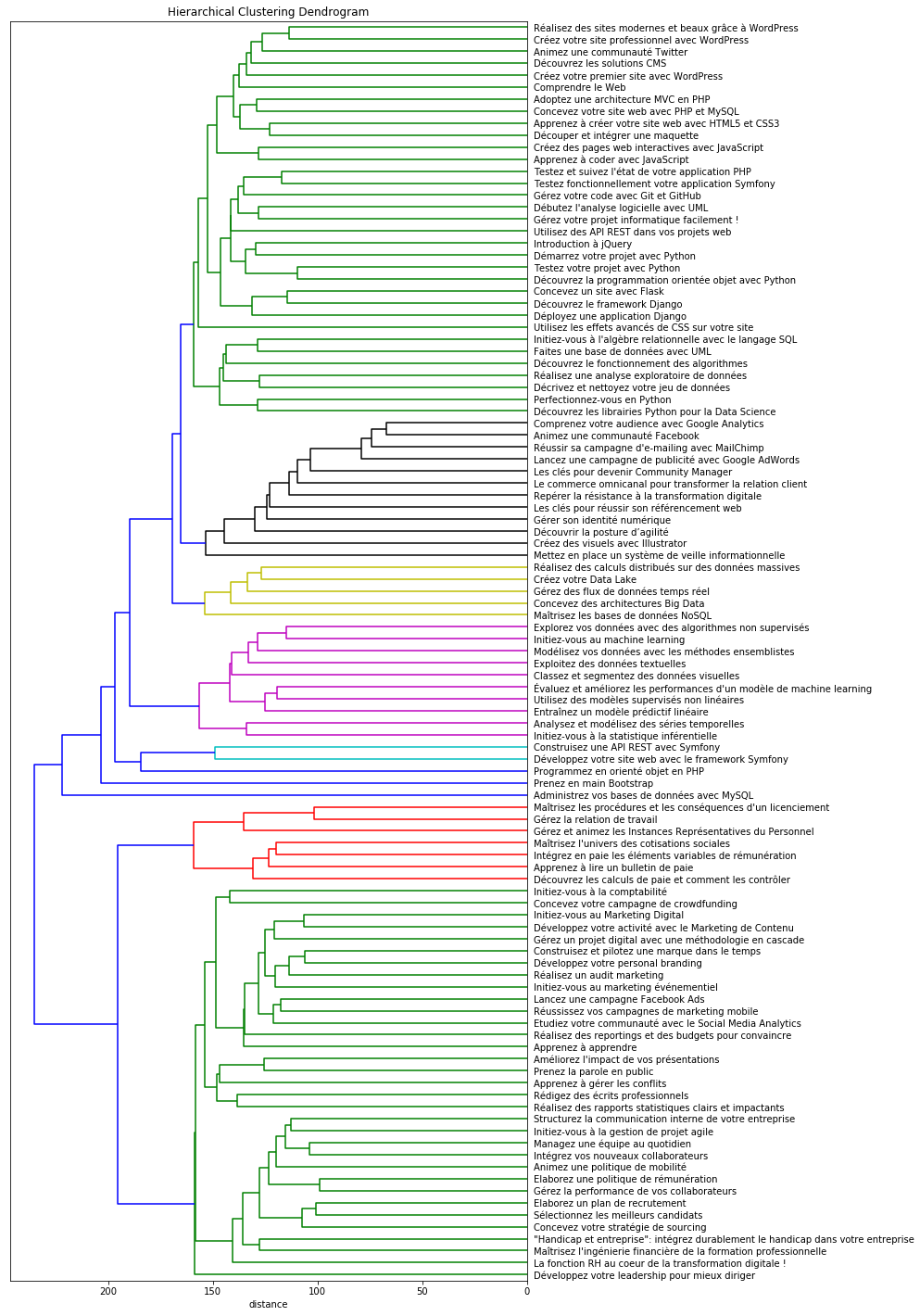
Z = linkage(X\_scaled, 'ward')

# Affichage du dendrogramme

plot\_dendrogram(Z, names)

Pour exécuter ce code, vous aurez besoin de la fonction plot\_dendrogram, disponible dans le fichier functions.py sur le git, ou tout en bas de ce chapitre.

Voici le dendrogramme obtenu (cliquez sur l'image pour zoomer) :



On peut ensuite couper l'arbre de manière à obtenir 12 clusters. À partir de ces 12 clusters, il peut être intéressant, par curiosité, de voir si les cours regroupés en un même cluster ont tendance à appartenir à une même thématique ou non. La thématique des cours est connue, et elle est stockée dans la variable theme. Pour ceci, nous afficherons le tableau de contingence entre la variable cluster et theme :

# Coupage du dendrogramme en 12 clusters

clusters = fcluster(Z, 12, criterion='maxclust')

#clusters = fcluster(Z, 159, criterion='distance') # ligne équivalente à la précédente

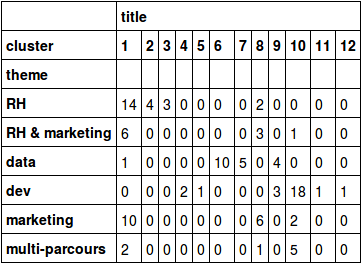
# Comparaison des clusters trouvés avec les classification des cours en différents thèmes

courses = pd.DataFrame({"theme": theme, "cluster": clusters, "title": names})

print(courses)

print(courses.pivot\_table(index="theme", columns="cluster", aggfunc=len, fill\_value=0))

Voici le tableau de contingence, affiché grâce à la ligne de code n° 8 :



Tout d'abord, on y voit que le cluster 1 a "capté" la grande majorité des cours de RH et de marketing : en effet, il en contient 30 (14 + 6 + 10) alors qu'il y a au total 51 cours de marketing et de RH. Et, surprise, un cours de data s'est glissé dans ce cluster ! À y regarder de plus près , on voit qu'il s'agit du cours [Réalisez des rapports statistiques clairs et impactants](https://openclassrooms.com/courses/4525336-realisez-des-rapports-statistiques-clairs-et-impactants). Sur le dendrogramme, on voit qu'il est proche de cours tels que [Rédigez des écrits professionnels](https://openclassrooms.com/courses/4929676-redigez-des-ecrits-professionnels) ou [Améliorez l'impact de vos présentations](https://openclassrooms.com/courses/3013891-ameliorez-limpact-de-vos-presentations), qui sont des cours destinés aux formation RH et marketing. Finalement, c'est assez compréhensible : le cours Réalisez des rapports statistiques clairs et impactants, même s'il est destiné aux data analysts, est plus similaire aux cours de communication écrite ou orale qu'aux cours de statistiques purement mathématiques. J'en profite au passage pour vous conseiller vivement de le suivre, car si vous ne savez pas rendre compte de vos analyses, vous aurez presque travaillé pour rien. ;)

Ensuite, on voit que le reste des cours de data sont répartis entre les clusters 6, 7 et 9. En regardant de plus près, le cluster 7 est composé exclusivement de cours du parcours [Data Architect](https://openclassrooms.com/paths/64-data-architect). Ils parlent donc bien de data, mais plus sous un angle logiciel que mathématique, et sont donc bien distincts des cours du cluster 6, qui appartiennent aux parcours [Data Analyst](https://openclassrooms.com/fr/paths/data-scientist) et [Data Scientist](https://openclassrooms.com/fr/paths/164-data-scientist). Quant au cluster 9, il contient également des cours de développement informatique. En fait, ceux-ci parlent de thématiques qui sont à mi-chemin entre la data et le développement : bases de données, librairies Python adaptées à la data science, algorithmes, etc.

### Le chat

Appliquons maintenant l'algorithme k-means sur notre chat, en demandant 6 clusters :

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn import decomposition

# Nombre de clusters souhaités

n\_clust = 6

# import de l'échantillon

data = pd.read\_csv('mystery.csv')

# préparation des données pour le clustering

X = data.values

# Réduire n'est ici pas nécessaire car les variables sont exprimées dans la même unité

# X\_scaled = preprocessing.StandardScaler().fit\_transform(X)

# Clustering par K-means

km = KMeans(n\_clusters=n\_clust)

km.fit(X)

# Récupération des clusters attribués à chaque individu

clusters = km.labels\_

# Affichage du clustering par projection des individus sur le premier plan factoriel

pca = decomposition.PCA(n\_components=3).fit(X)

X\_projected = pca.transform(X)

plt.scatter(X\_projected[:, 0], X\_projected[:, 1], c=clusters.astype(np.float), cmap = 'jet', alpha=.2)

plt.title("Projection des {} individus sur le 1e plan factoriel".format(X\_projected.shape[0]))

plt.show(block=False)

Voici le résultat :



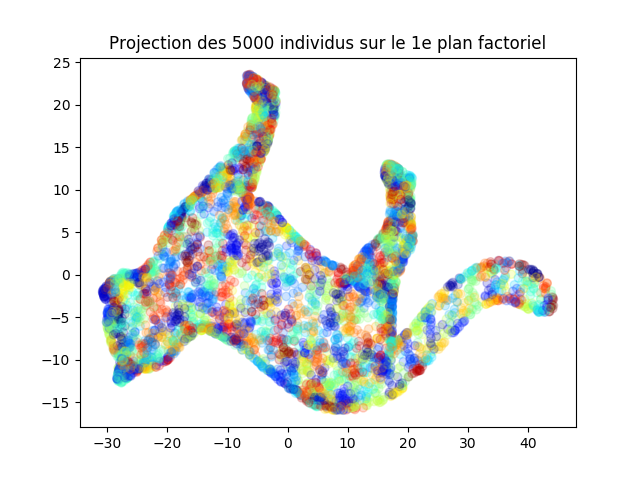
Ici, le k-means a identifié les parties du corps du chat : on a un cluster correspondant à la queue, un autre aux 2 pattes avant, un aux 2 pattes arrière, etc.

Remarquons qu'ici, nous avons affiché les données sur le premier plan factoriel, puis coloré les points en fonction de leur cluster.

OK, mais que fait-on avec cela ?

Sur un jeu de données qui représente un "objet" géométrique, le k-means peut par exemple servir à faire de la quantification. Ici, nous avons un chat composé de 5 000 points. Il peut y avoir des applications pour lesquelles les 5 000 points ne sont pas requis, et pour lesquelles on veut réduire intelligemment le nombre de points de l'objet sans en perdre la forme.

On peut donc faire un k-means en 500 clusters, ce qui donne ceci :



Pour obtenir cette image, reprenez le code ci-dessus en modifiant n\_comp à 500.

Ensuite, on ne garde que le centre de chaque cluster, et on reconstitue un nouvel échantillon à partir de ces 500 centres. On obtient un échantillon de 500 individus, on a donc réduit les dimensions des données. Avec ces 500 points, la forme reste (plus ou moins) inchangée :

# Affichage des positions des centres de classes

plt.figure()

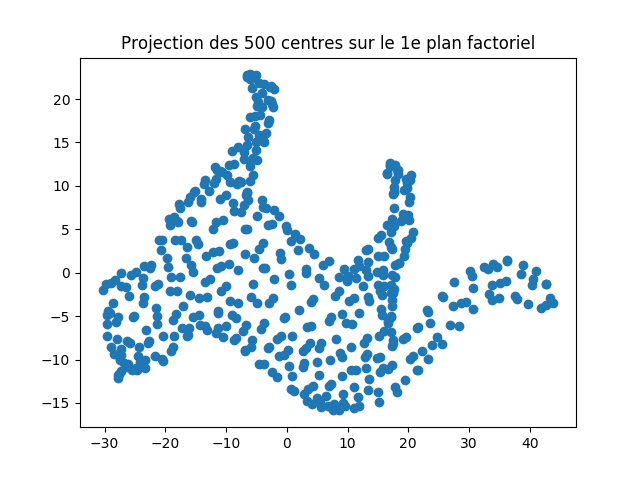
centroids = km.cluster\_centers\_

centroids\_projected = pca.transform(centroids)

plt.scatter(centroids\_projected[:,0],centroids\_projected[:,1])

plt.title("Projection des {} centres sur le 1e plan factoriel".format(len(centroids)))

plt.show()



Le même principe peut être utilisé en traitement d'image pour la quantification des couleurs. Allez voir ce lien, il est intéressant et le code est très similaire à celui ci-dessus : [**http://scikit-learn.org/stable/auto\_examples/cluster/plot\_color\_quantization.html**](http://scikit-learn.org/stable/auto_examples/cluster/plot_color_quantization.html)

### 

### Annexe : les fonctions outils

Voici la définition des fonctions utilisées dans ce chapitre, disponible dans le fichier functions.py :

import matplotlib.pyplot as plt

from scipy.cluster.hierarchy import dendrogram

def plot\_dendrogram(Z, names):

plt.figure(figsize=(10,25))

plt.title('Hierarchical Clustering Dendrogram')

plt.xlabel('distance')

dendrogram(

Z,

labels = names,

orientation = "left",

)

plt.show()

## Entraînez-vous à réaliser un détecteur de slides grâce à de l'analyse d'image

### 

### À vous de jouer

Pour vous entraîner, réalisez cet exercice étape par étape. Une fois terminé, vous pouvez comparer votre travail avec les pistes que je vous propose.

Exécutez le code donné. 4 lignes de code ont été remplacées par  ###[...]###  . Il vous faudra les remplacer par le bon code. Vous ne serez évalué que sur ces 4 lignes.

Téléchargez :

* [L'échantillon](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525281/P2P/video.csv.zip) de 1 934 lignes et 27 648 colonnes (à décompresser) ;
* le code Python (format [notebook](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525281/P2P/detecteur_slides.ipynb) ou [html](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525281/P2P/detecteur_slides.html)), ou bien
* le [code](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525281/P2P/slide_detector.r) en langage R

Nous allons utiliser l'ACP et le clustering pour traiter une vidéo ! Vous la trouverez [ici](https://www.youtube.com/watch?v=uV5hmpzmWsU) (Il n'est pas nécessaire de la télécharger).

Nous allons programmer un "détecteur de slides". La vidéo que nous utilisons est celle d'un cours magistral donné en amphithéâtre, dans laquelle le professeur appuie son propos grâce à un diaporama, composé de plusieurs slides.   
Notre objectif sera d'extraire les différentes slides de la présentation, grâce à une analyse d'image.

Une vidéo est un enchaînement rapide d'images fixes. Quand une slide est affichée à l'écran durant 20 secondes, alors toutes les images contenues dans ces 20 secondes sont semblables (même si le professeur bouge un peu). On a donc des images similaires (dues aux slides immobiles), même si un peu différentes (dues au mouvement du professeur).

On va donc détecter les images similaires pour en extraire les différentes slides de la présentation ! Pour détecter les images similaires, nous utiliserons du **clustering**.

Mais nous serons également obligés de**réduire les dimensions** de nos données : nous ferons ceci grâce à l'**ACP**.

**Les données**

La vidéo dure 1 h 36, et contient 174 053 images. Une image est composée de pixels. Chaque image mesure 640 pixels en largeur et 360 en hauteur, soit 230 400 pixels ! En plus, la couleur d'un pixel est encodée selon 3 niveaux : rouge, vert et bleu. On a donc 174 053 \* 230 400 \* 3 = 120 305 433 600 unités d'information dans cette vidéo ! C'est un peu lourd pour nos traitements statistiques, on devra donc réduire la taille des données.

Pour en réduire la taille, nous ne sélectionnons qu'une image sur 90 (une toutes les 3 secondes). Ensuite, on réduit la taille des images à 72 pixels de hauteur et 128 de largeur. Nous réduisons également la palette des couleurs possibles : originellement, les niveaux des 3 couleurs (RGB) sont encodés sur une échelle de 0 à 256 : nous réduirons cette échelle à une échelle de 0 à 64.

On obtient donc un tableau où chaque **ligne** (chaque individu) correspond à 1 image. Il y a **1 934** images. Chaque image est caractérisée par 72\*128\*3 = **27 648** unités d'information, donc autant de **colonnes**. On ajoute en plus une colonne qui indique la position de l'image dans le temps (par exemple, "cette image s'affiche à la150e seconde").

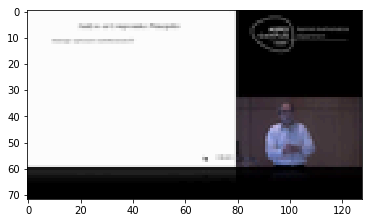
Vous n'avez pas à reprogrammer ceci, le fichier csv que vous avez téléchargé est déjà le résultat des opérations citées ci-dessus. Mais si vous êtes curieux, le code correspondant est : [**ici**](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525281/P2P/create_csv.py).

**Chargement des données**

Le code des étapes suivantes est donné dans le notebook.

Nous importons le CSV et stockons la matrice dans un array numpy appelé X. Cette opération peut prendre quelques secondes. Ensuite, nous affichons l'image de la 150e seconde.

<https://youtu.be/uV5hmpzmWsU?t=150s>

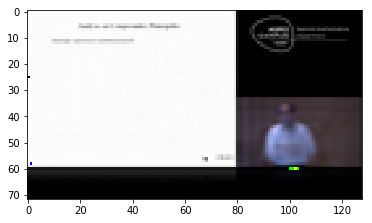


**ACP et réduction de dimensions**

Comme il y a beaucoup de colonnes, nous réduisons les données grâce à une ACP après centrage-réduction des données.

En fait, nous effectuons ici une compression de vidéo ! Il existe des formats informatiques d'images ou de vidéos compressées (jpeg, mp4, etc.). Certains de ceux-ci se basent sur des méthodes de compressions similaires à l'ACP !

Après compression de la vidéo, on a forcément perdu en qualité. Affichons l'image de la 150e seconde pour voir cette perte de qualité :



En affichant les individus projetés sur les premiers plans factoriels, on voit déjà quelques groupes d'images similaires.  
Pour les visionner, rendez-vous à cette URL, [https://www.youtube.com/watch?v=uV5hmpzmWsU?t=[...]s](https://www.youtube.com/watch?v=uV5hmpzmWsU?t=%5B...%5Ds), en remplaçant le [...] par l'identifiant du point à visualiser. Cet identifiant correspond à la position de l'image dans la vidéo ; il est donné en secondes. Vous pouvez vérifier si deux points proches correspondent à deux images qui se ressemblent.

**Clustering**

Nous allons regrouper les images similaires. L'idéal serait d'avoir un groupe par slide.

Nous allons donc effectuer une **classification hiérarchique** et afficher le dendrogramme pour choisir le nombre de classes.

Cependant, il y a trop d'individus (trop d'images) pour effectuer une C.H. rapidement. Nous allons donc d'abord effectuer un clustering avec **k-means** afin de trouver 300 clusters, en espérant que le professeur n'ait pas plus de 300 slides. Il est fort probable qu'après l'application du k-means, plusieurs clusters correspondent à une même slide : c'est là que la classification hiérarchique sera utile afin de regrouper les clusters correspondant à une même slide.

Nous allons ensuite effectuer la classification hiérarchique sur les 300 clusters. Le professeur a environ 40 slides différentes. Coupons donc l'arbre en 40 clusters.

**Résultat**

On affiche aussi l'image moyenne de chaque cluster. Elle représente une "image moyenne" composée de plusieurs images. Sur celles-ci, les objets fixes seront nets, et les objets mobiles seront flous. Les slides seront donc nettement visibles !

### 

### Vérifiez votre travail

Voici un [**exemple**](https://static.oc-static.com/activities/2657/evaluation_resources/realisez-un-detecteur-de-slides-grace-a-de-lanalyse-dimage_exemple-2019-06-06T172018.zip) pour vous permettre de vérifier votre travail.