2022-2023

Module : Analyse de données

Prof : Mr. A. Ouaarab

Rapport d'analyse de données

Réalisé par:

AHDDAD Imad (IRSI2)

TRAORE Issiaka (IRSI2)

TRAORE Koudadim Olivier (IFA2)

Table des matières

Introduction

Dans le but de connaitre le comportement des étudiants de la FASTG, une collecte de données a été menée auprès de la population cible. Cette collecte d'information est rendue possible grâce à un formulaire google form de vingt-neuf (29) questions d'intérêt fermées et à choix multiples. Le nombre d'étudiants ayant participé au sondage est de 128.

L'étude vise deux objectifs : étude comportementale et étude de la personnalité des étudiants de la FSTG. Ainsi les dix-neuf premières questions ont été conçues pour répondre au premier objectif et les dix dernières au dernier objectif.

1. Etude comportementale

Pour mener à bien l'étude du comportement des étudiants, les données collectées sont traitées et analysées à l'aide de différentes méthodes pour répondre à des questions bien précises. Dans cette section, nous évoquerons les méthodes ainsi les questions auxquelles elles permettent de répondre.

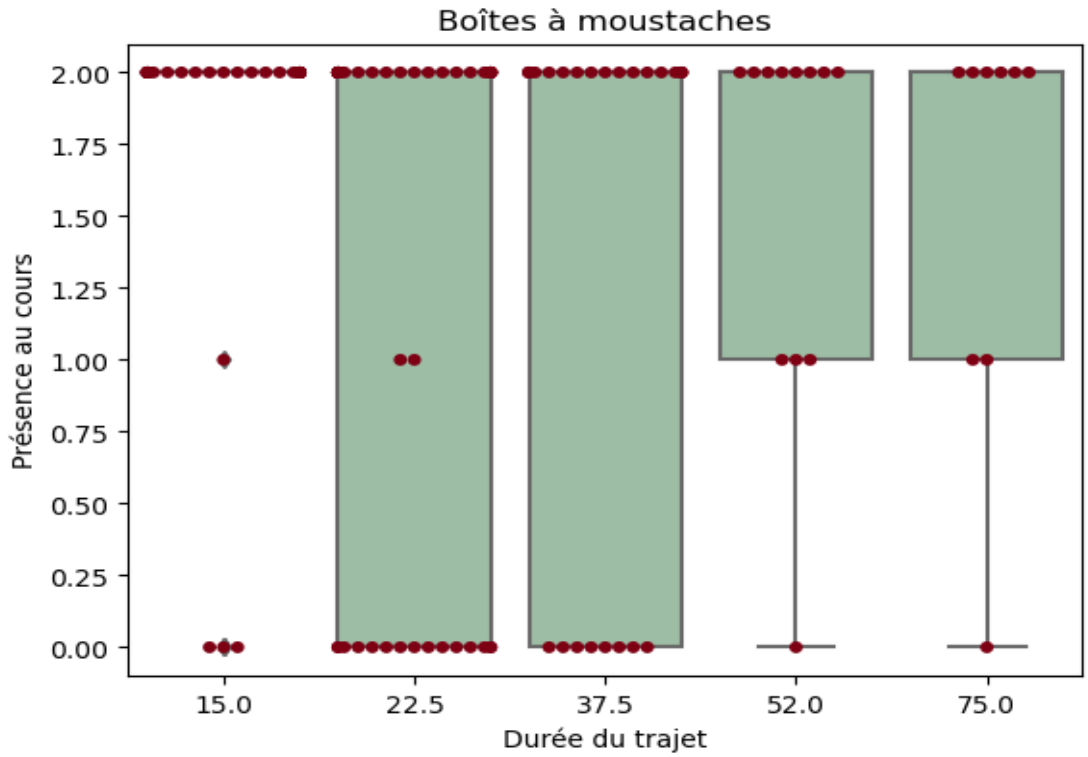
1.1. ANOVA à un facteur

Par à l'ANOVA, nous répondrons à deux questions.

***Question 1** : La durée de trajet Domicile-FSTG à parcourir par un étudiant, a-t-elle un impact sur son assiduité aux cours ?*

Considérons la boîte à moustaches de la figure 1 ci-dessous.

D'après elle, il y a une variation de la « **Présence au cours** » (Y) selon les modalités de la « **Durée de trajet** » (Q7). Car d'une modalité à une, il y a une dispersion différente des valeurs de Y. Donc l'ANOVA est bien applicable aux données.



Box-plot (Fig.1)

Après application d'ANOVA, on obtient le tableau suivant :

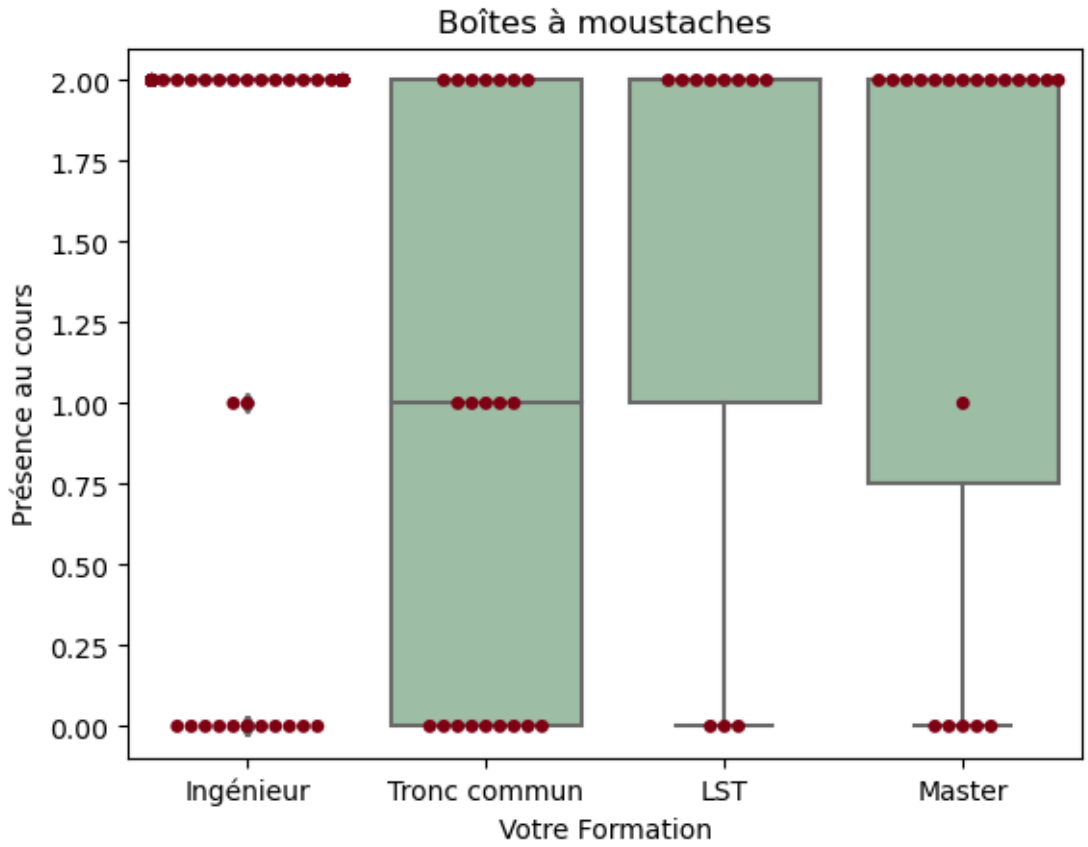
	df	sum_sq	mean_sq	F	PR(>F)
C(Q7)	4.0	4.483919	1.120980	1.650946	0.165772
Residual	123.0	83.516081	0.678993	NaN	NaN

Selon ce tableau, pour un seuil de 5%, on ne rejette donc pas l'hypothèse H0 selon laquelle $\alpha 1 = \alpha 2 = 0$. On peut conclure que la durée du trajet n'a aucune incidence sur la présence de l'étudiant, tout au moins pas d'effet significatif. Car 0.16>0.05.

***Question 2** : La nature de la formation d'un étudiant peut-elle expliquer son assiduité aux cours ?*

Considérons la boîte à moustaches de la figure 2 ci-dessous.

D'après elle, Il y a une variation de la « **Présence au cours** » (Y) selon les modalités de la « **Formation** » (Q3). Car d'une modalité à une, il y a une dispersion différente des valeurs de Y. Donc l'ANOVA est bien applicable aux données.



Box-plot (Fig.2)

Après application d'ANOVA, on obtient le tableau suivant. Il montre l'état des relations entre les variables Q3 et Q7 prises à part sur la présence(Q8) au cours et prises ensemble.

	df	sum_sq	mean_sq	F	PR(>F)
Q3	3.0	10.092151	3.364050	5.462128	0.001489
Q7	1.0	0.000436	0.000436	0.000708	0.978811
Q3:Q7	3.0	4.001049	1.333683	2.165470	0.095651
Residual	120.0	73.906363	0.615886	NaN	NaN

Selon ce tableau, pour un seuil de 5%, on peut conclure que la durée du trajet (Q7) n'a aucune incidence sur la présence de l'étudiant. Car 0.16>0.05.

Par contre, puisque 0.001489 < 0.05, la formation de l'étudiant influence sa présence.

Notons cependant que la combinaison des variables **Formation & Durée** de trajet n'a pas d'effet sur l'assiduité de l'étudiant.

1.2. Analyse Factorielle des Correspondances (AFC)

A l'aide de l'AFC, nous chercherons à vérifier s'il existe un lien entre le niveau d'étude et le sexe d'un étudiant donné.

Pour ce faire, on considère le graphe suivant représentant les différentes modalités en jeu.

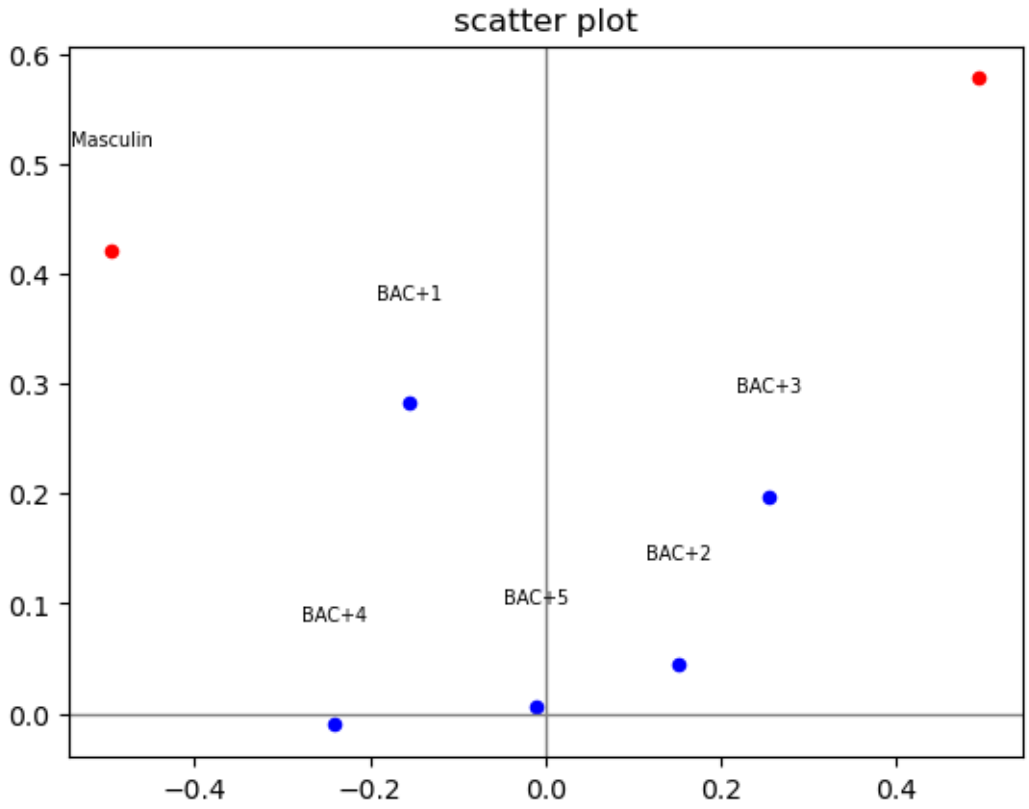


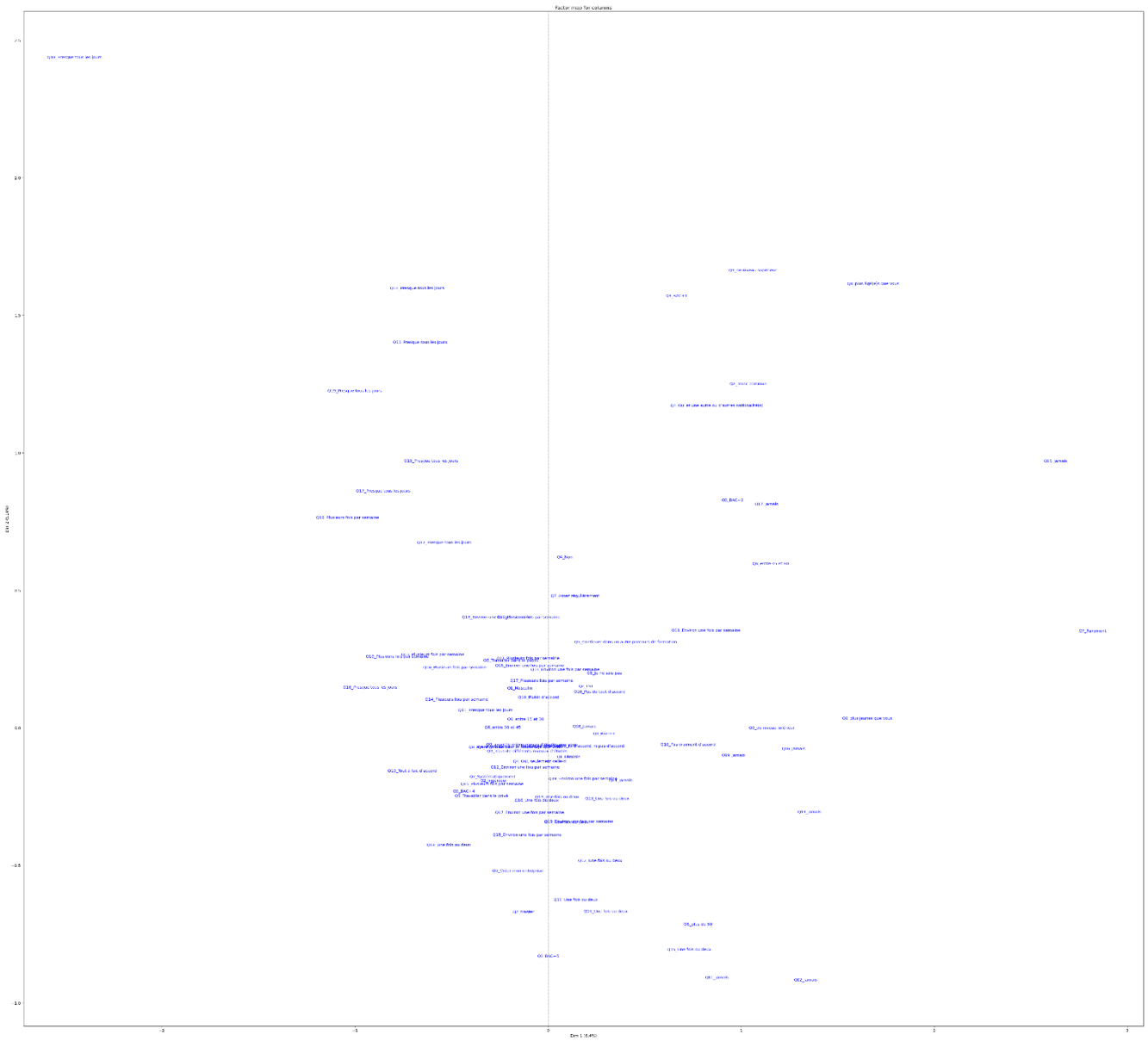
Fig.3

On peut conclure que dans le cadre notre étude, il y a plus d'étudiants de niveau BAC+1, BAC+4 et BAC+5 que d'étudiantes. Par contre, on enregistre plus de filles que de garçons dans les niveaux BAC+2 et BAC+3.

1.3. Analyse en Composantes Multiples (ACM)

ETUDE DES RELATIONS ENTRE LES MODALITES

Par cette méthode, l'on obtient la représentation ci-dessous des différentes modalités.



Graphe représentant les différentes modalités des variables étudiées (fig. 4)

Cette figure permet de répondre à des questions portant sur les relations entre modalités.

Plus deux modalités sont proches l'une de l'autre, plus elles sont similaires ; Plus deux modalités sont éloignés l'une de l'autre, moins elles sont prises simultanément par les individus étudiés. A titre illustratif, on a :

- Quel sexe a le plus participé au sondage ?
  Les filles ! Car la modalité *Féminin* est plus proche du centre que *Masculin*.
  Cela est confirmé par le diagramme sur la figure 5.
- Quels sont les niveaux d'étude et la formation où les filles sont dominantes en nombre ?
  Filles : BAC+3, BAC+4, cycle d'ingénieur.
  Garçons : LST, BAC+1, BAC+2, …

- Quelles sont les préférences d'orientation après les études, selon le sexe ?
  En général, les Garçons préfèrent travailler dans le public alors que les filles préfèrent faire carrière dans le privé. Par ailleurs, il existe plus de garçons que de filles ne sachant pas dans quelle carrière professionnelle s'engager.
- Quel est le niveau d'étude dominant chez les élève-ingénieurs ?
  Elève-ingénieurs : BAC+4.
- Existe-t-il plus ou moins d'étudiants en Master de niveau BAC+5 que d'élèves-ingénieurs ?
  Oui !
- Existe-t-il un effet entre le sexe et l'intégration des étudiants ?
  Les filles se sentent beaucoup plus intégrées que les garçons.
- Quel est le niveau d'étude des étudiants qui ont le plus participé au sondage ?
  BAC+4 : Car il est beaucoup plus proche du centre que les autres. La confirmation est donnée par la figure 6.

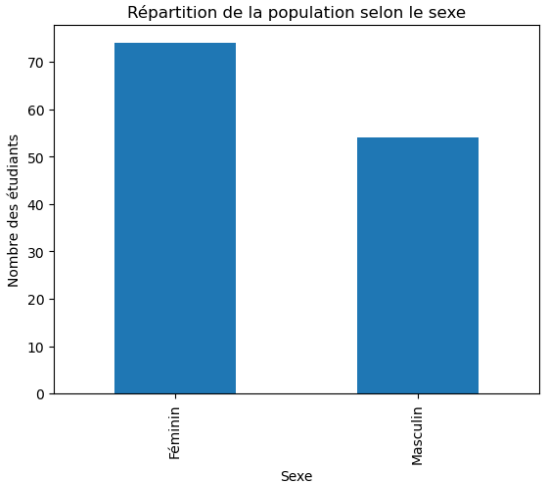


Fig.5

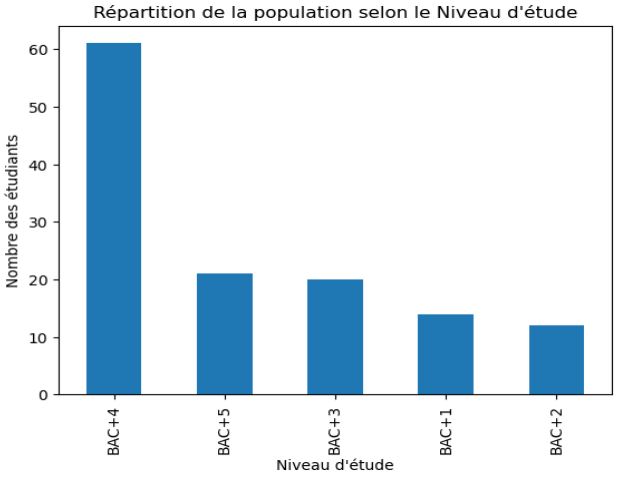


Fig.6

LES VALEURS PROPRES

D'où proviennent les deux axes factoriels constituant le plan factoriel dans lequel sont représentées les modalités ?

- Le premier axe factoriel provient de la plus grande valeur propre de tableau des données étudiées.
- Le deuxième axe factoriel est issu de la deuxième plus grande valeur propre.

Il existe 66 valeurs propres pour les tableaux de données étudiées. Elles sont représentées par le diagramme suivant :

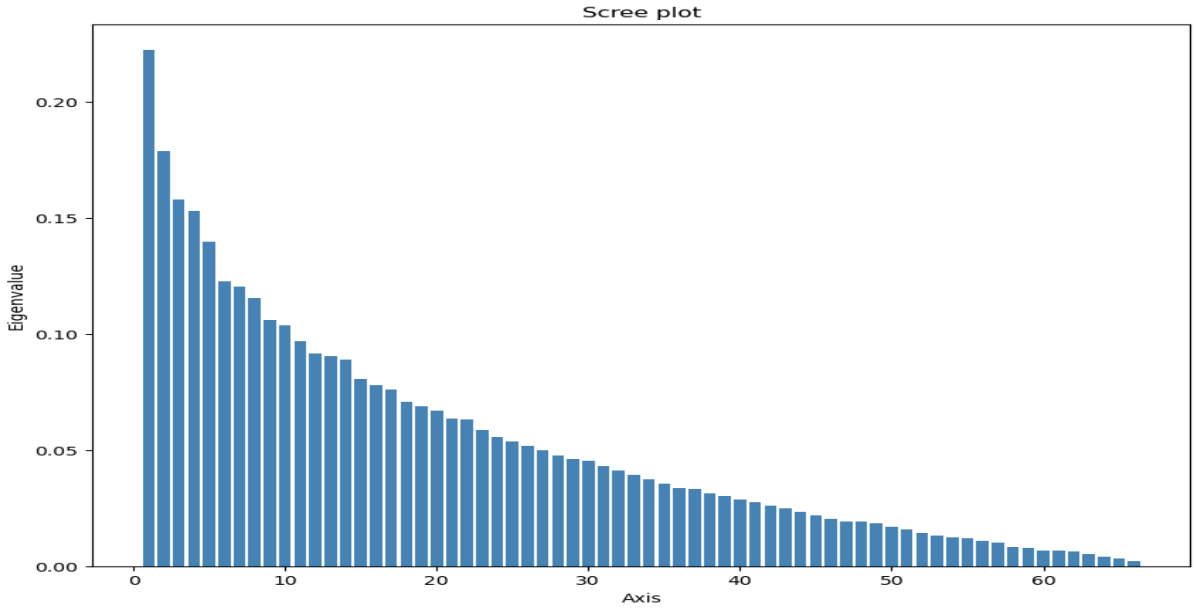
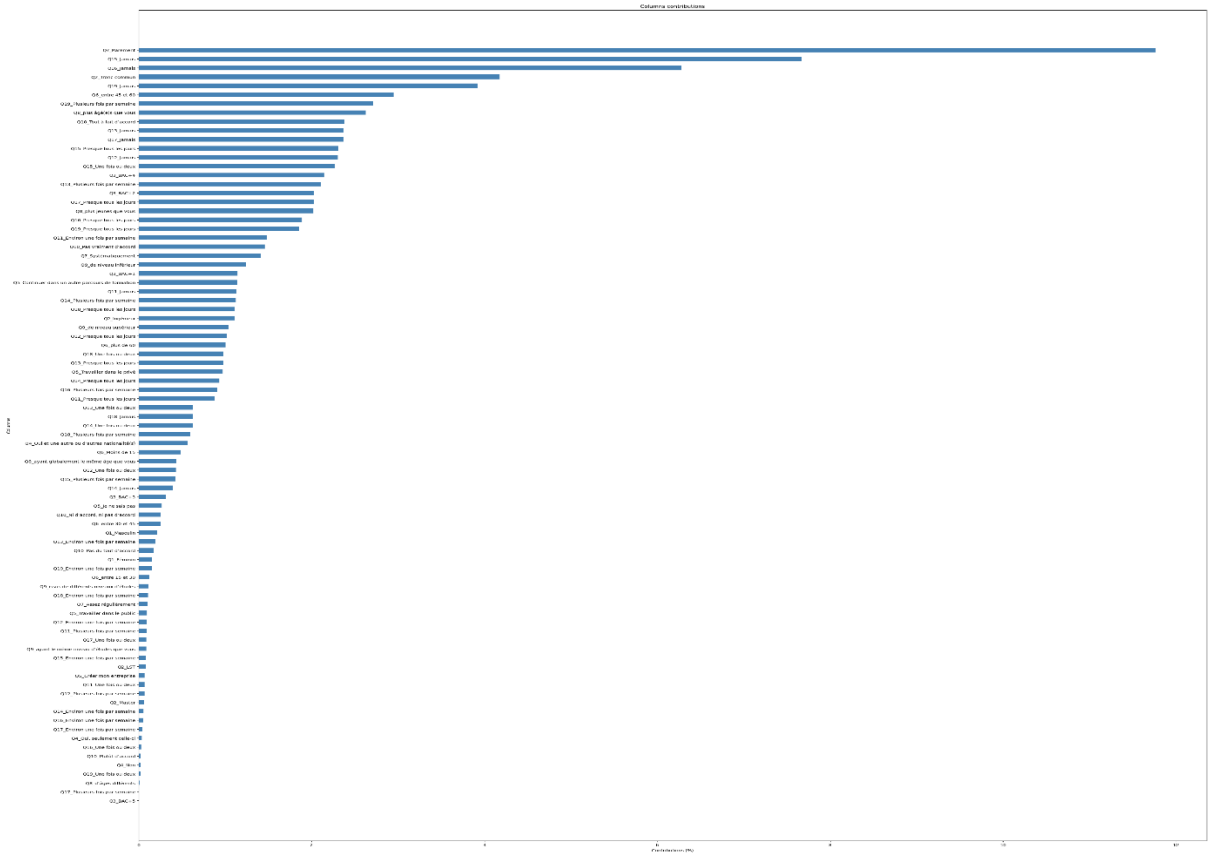


Diagramme des valeurs propres (Fig.7)

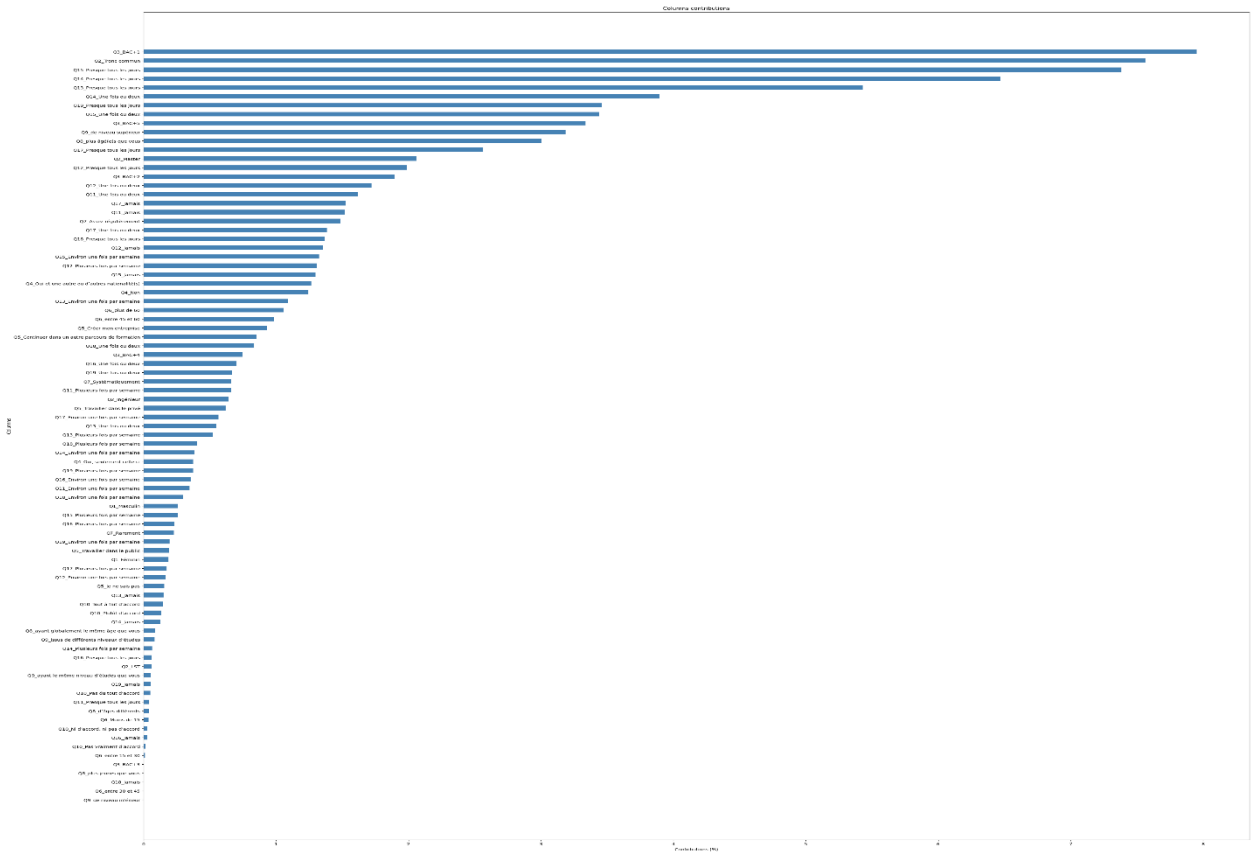
CONTRIBUTIONS DES MODALITES A LA FORMATION DES AXES FACTORIELS

Les différentes contributions des modalités à la formation des axes 1 & 2 sont résumées par les diagrammes suivants.



Contributions à la formation de l'axe 1 (fig.8)

On peut constater que des modalités des variables Q8, Q16 et Q17 sont respectivement les trois premières qui contribuent le plus à la formation du l'axe 1.



Contributions à la formation de l'axe 2 (fig.9)

On peut constater que des modalités des variables Q4, Q3 et Q16 sont respectivement les trois premières qui contribuent le plus à la formation du l'axe 2.

2. Etude de la personnalité des étudiants

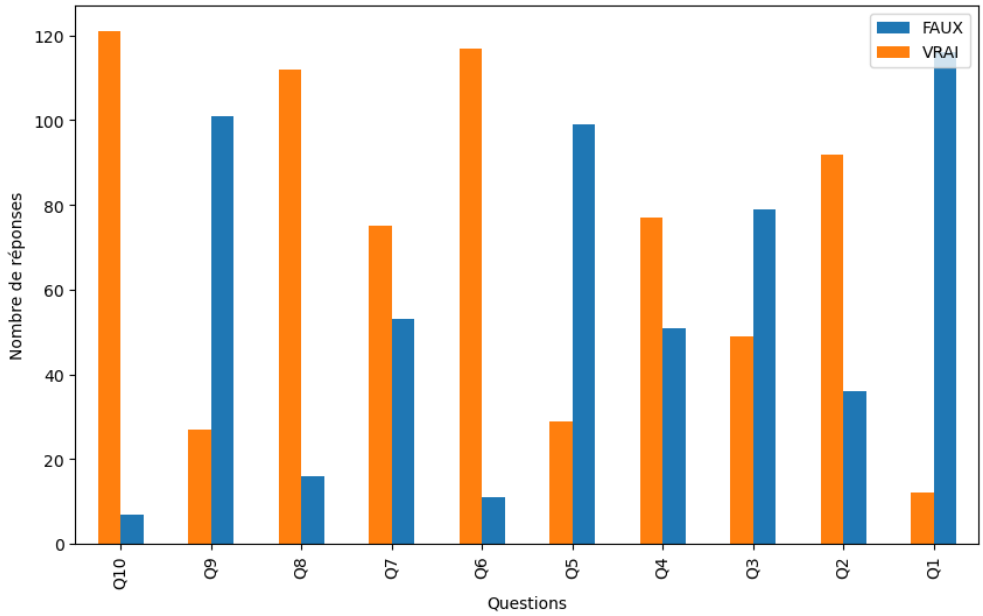
2.1. La Classification

Commençons par une description du jeu de données qui concerne la personnalité des étudiants.
Cette partie de jeu de données est constituée de :

- 10 colonnes qui représentent des questions à choix multiples visant à déterminer la personnalité des étudiants.
- 128 lignes qui représentent les réponses de 128 étudiants de la FSTM.

On a changé les noms des colonnes pour que les affichages soient lisibles.

- Q1 : Si je dois travailler dur à quelque chose, cela veut dire que je ne suis pas intelligent(e).
- Q2 : J'aime essayer des choses difficiles.
- Q3 : Lorsque je fais une erreur, j'ai honte ou je suis gêné.
- Q4 : J'aime qu'on me dise que je suis intelligent(e)
- Q5 J'abandonne généralement quand quelque chose devient difficile ou frustrant.
- Q6 : Ça ne me dérange pas de faire des erreurs car elles m'aident à apprendre
- Q7 : Il y a des choses pour lesquelles je ne serai jamais doué(e).
- Q8 : N'importe qui peut apprendre certaines choses s'il y travaille dur.
- Q9 : Les gens naissent soit stupides, moyens ou intelligents et ils ne peuvent changer cela par la suite
- Q10 : Je suis fier(e) quand je fais de mon mieux, même si ce n'est pas parfait



Répartition des réponses pour chaque questions (fig.10)

Afin d'appliquer la classification des étudiants selon leur esprit, on a fait une préparation des données, on a :

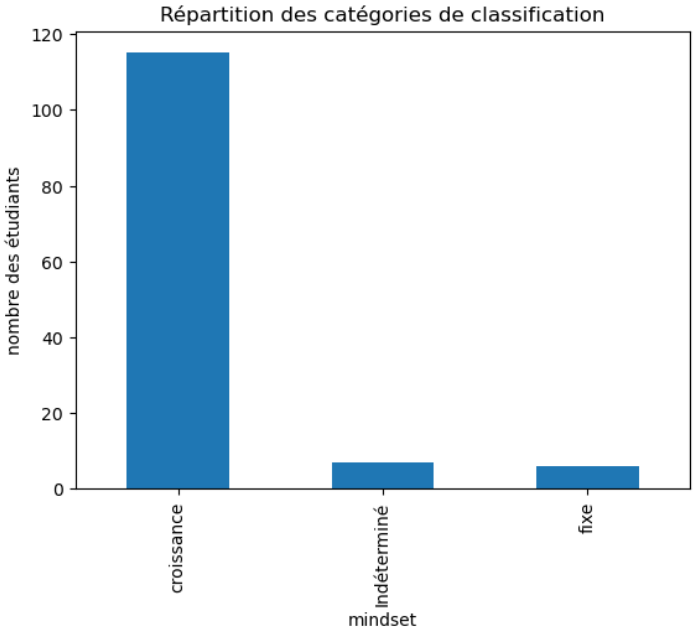
- mis toutes les réponses en majuscule
- transformé VRAI en 1 et FAUX en 0
- ajouté les trois colonnes fixe_score, croissance_score et mindset

Après, on a classifié les questions selon la logique en esprit fixe et esprit de croissance, la classification est comme suit :

	VRAI	FAUX
Q1	fixe	croissance
Q2	croissance	fixe
Q3	fixe	croissance
Q4	fixe	croissance
Q5	fixe	croissance

	VRAI	FAUX
Q6	croissance	fixe
Q7	fixe	croissance
Q8	croissance	fixe
Q9	fixe	croissance
Q10	croissance	fixe

Maintenant en se basant sur cette classification, on peut classer les étudiants selon leurs réponses en esprit fixe ou esprit de croissance.



Répartition des étudiants selon leur état d'esprit (fig.11)

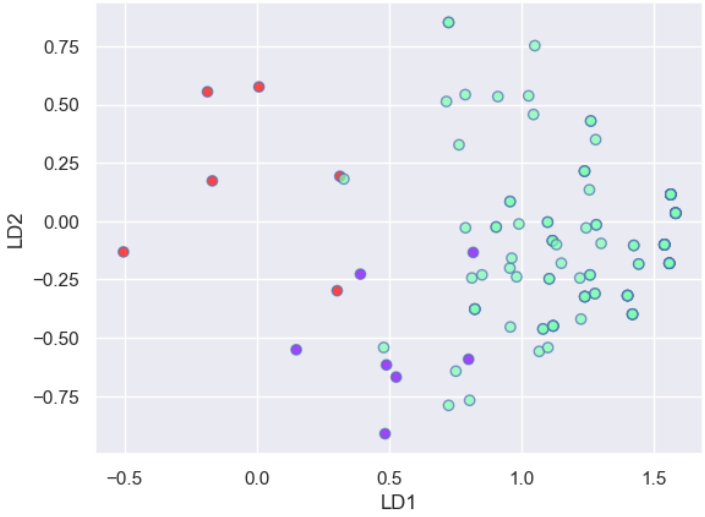
Les étudiants se répartissent selon l'état d'esprit comme suit :

- fixe : 6
- de croissance : 115
- indéterminé : 7

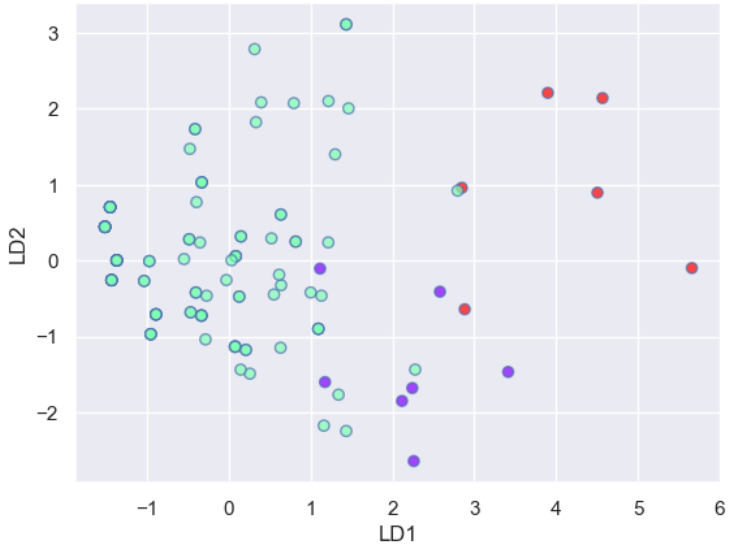
Total : 128

2.2. L'AFD

Maintenant, on a notre jeu de données classé suivant la colonne mindset, on peut donc appliquer une AFD.



Le nuage de points la colonne de mindset avant l'application de l'AFD (fig.12)



Le nuage de points la colonne de mindset après l'application de l'AFD (fig.13)