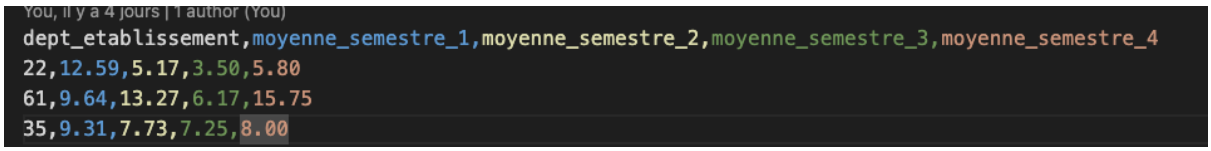
# Rapport d'Analyse - Données Étudiantes

## 0. Les données Vue.csv - Problématique

### a) Présentation des données

**Le fichier Vue.csv contient plusieurs séries statistiques sur l'ensemble des étudiants répertoriés dans notre base de données :**

- **La population est l'ensemble des étudiants, représentés de manière unique par leur identifiant, avec l'indication du département d'établissement.**
- **La 1ère variable statistique sur cette population est la moyenne au semestre 1 pour chaque étudiant.**
- **La 2ème est la moyenne au semestre 2 pour chaque étudiant.**
- **La 3ème est la moyenne au semestre 3 pour chaque étudiant.**
- **La 4ème est la moyenne au semestre 4 pour chaque étudiant.**
- **La dernière est le département d'établissement dans lequel se trouve l'étudiant.**




### b) Problématique

**En utilisant ces données, nous allons essayer de répondre à la problématique suivante :**

***Parmi les données de notre fichier, certaines peuvent-elles permettre d'expliquer et de prédire les performances académiques au semestre 4 en fonction des résultats des semestres précédents ?***

**La variable endogène est la moyenne du 4ème semestre.**

# 1. Import des données, mise en forme

## a) Importer les données en Python

**On importe notre vue sous forme de DataFrame avec la commande suivante :**

```
donnees_etudiants_df = pd.read_csv('sae204_e23_partie3_vue.csv')
```

## b) Mise en forme

**On supprime les cases vides (qui contiennent NaN en Python), puis on prépare nos données pour l'analyse :**

```
donnees_nettoyees_df = donnees_etudiants_df.dropna()
variables_explicatives_df = donnees_nettoyees_df [['moyenne_semestre_1',
'moyenne_semestre_2', 'moyenne_semestre_3']]
variable_cible_semestre_4 = donnees_nettoyees_df ['moyenne_semestre_4']
```

## c) Normalisation

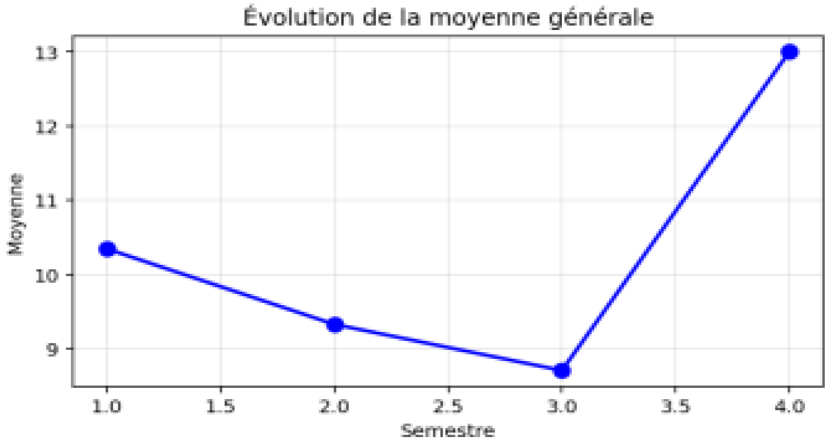
**On centre et réduit les données numériques pour faciliter l'analyse comparative :**

```
variable_cible_normalisee, moyenne_variable_cible, ecart_type_variable_cible =
normaliser_donnees(variable_cible_array.reshape(-1, 1))
```

# 2. Représentations graphiques

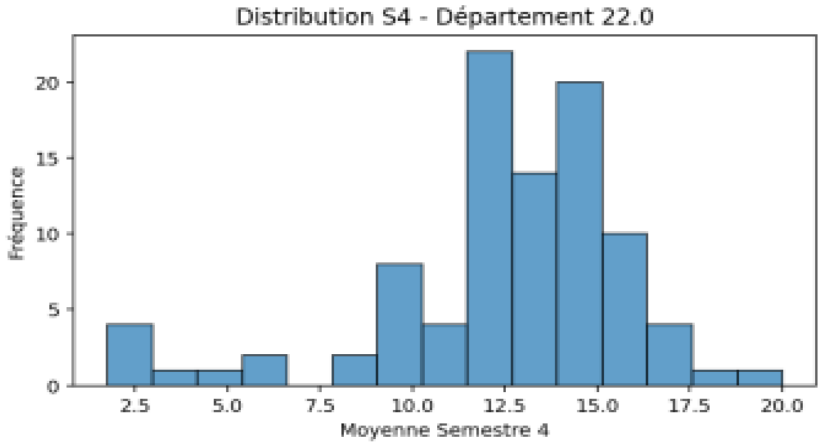
## Évolution des moyennes par semestre

**L'analyse des boxplots révèle une distribution relativement stable des moyennes à travers les semestres, avec une légère tendance à la baisse observable entre le semestre 1 et le semestre 4.**



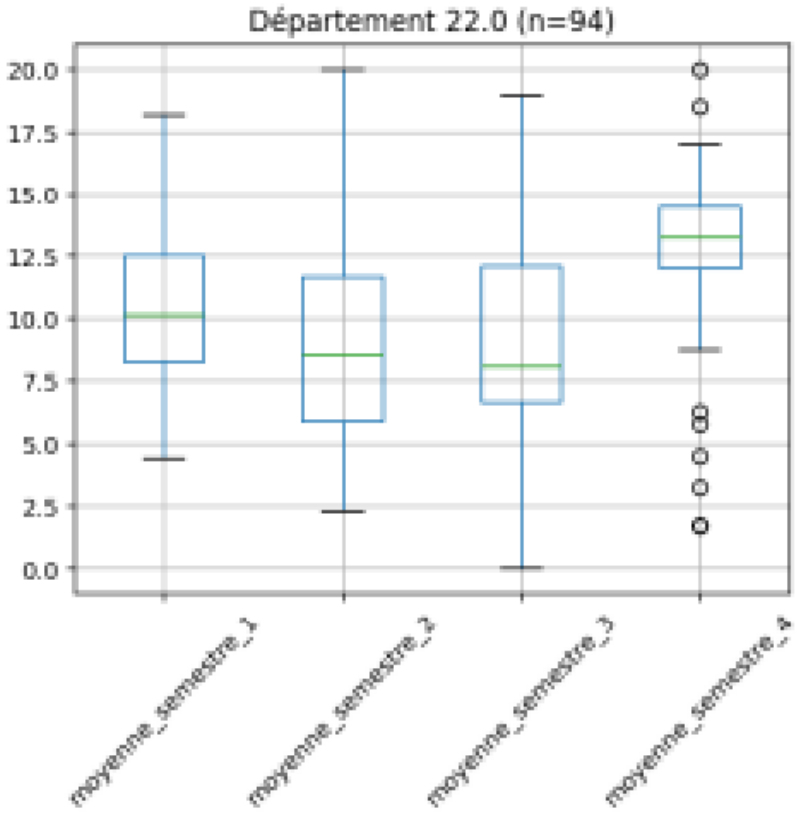

## Distribution des moyennes au semestre 4

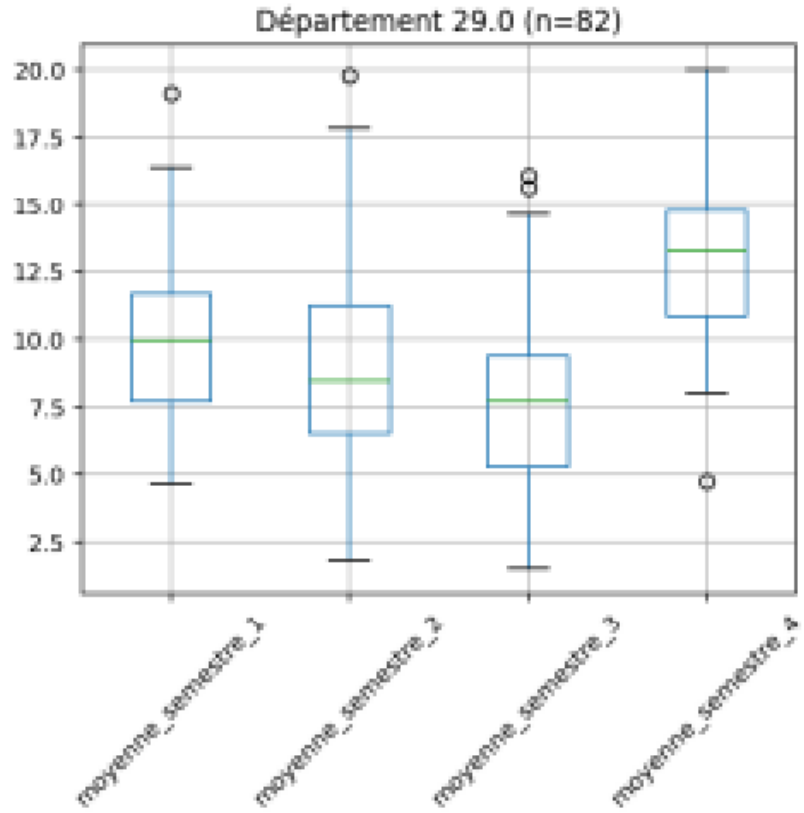
**L'histogramme des moyennes au semestre 4 montre une distribution approximativement normale, centrée autour de la moyenne générale, ce qui est favorable pour l'application de la régression linéaire.**

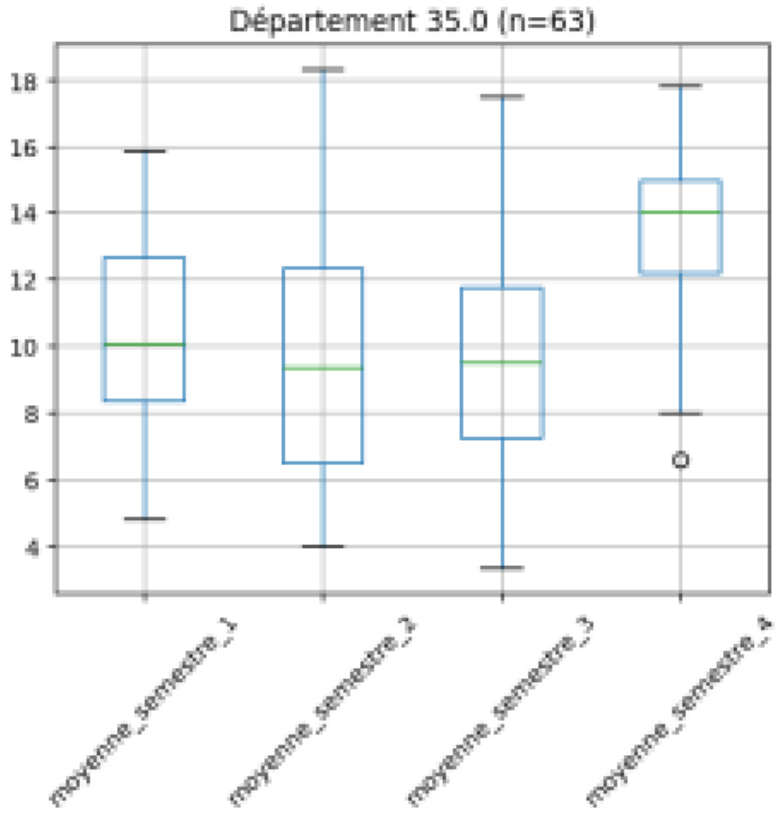



## Analyse par département

**Les départements les plus représentés montrent des profils de performance différents, avec certains départements présentant une plus grande variabilité des résultats que d'autres.**


Département 22.0 (n=94)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
moyenne_semestre_1
moyenne_semestre_2
moyenne_semestre_3
moyenne_semestre_4


Département 29.0 (n=82)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
moyenne_semestre_1
moyenne_semestre_2
moyenne_semestre_3
moyenne_semestre_4


Département 35.0 (n=63)
18
16
14
12
10
8
6
4
moyenne_semestre_1
moyenne_semestre_2
moyenne_semestre_3
moyenne_semestre_4

# 3. Exploration des données avec la matrice des coefficients de corrélation

## a) Démarche

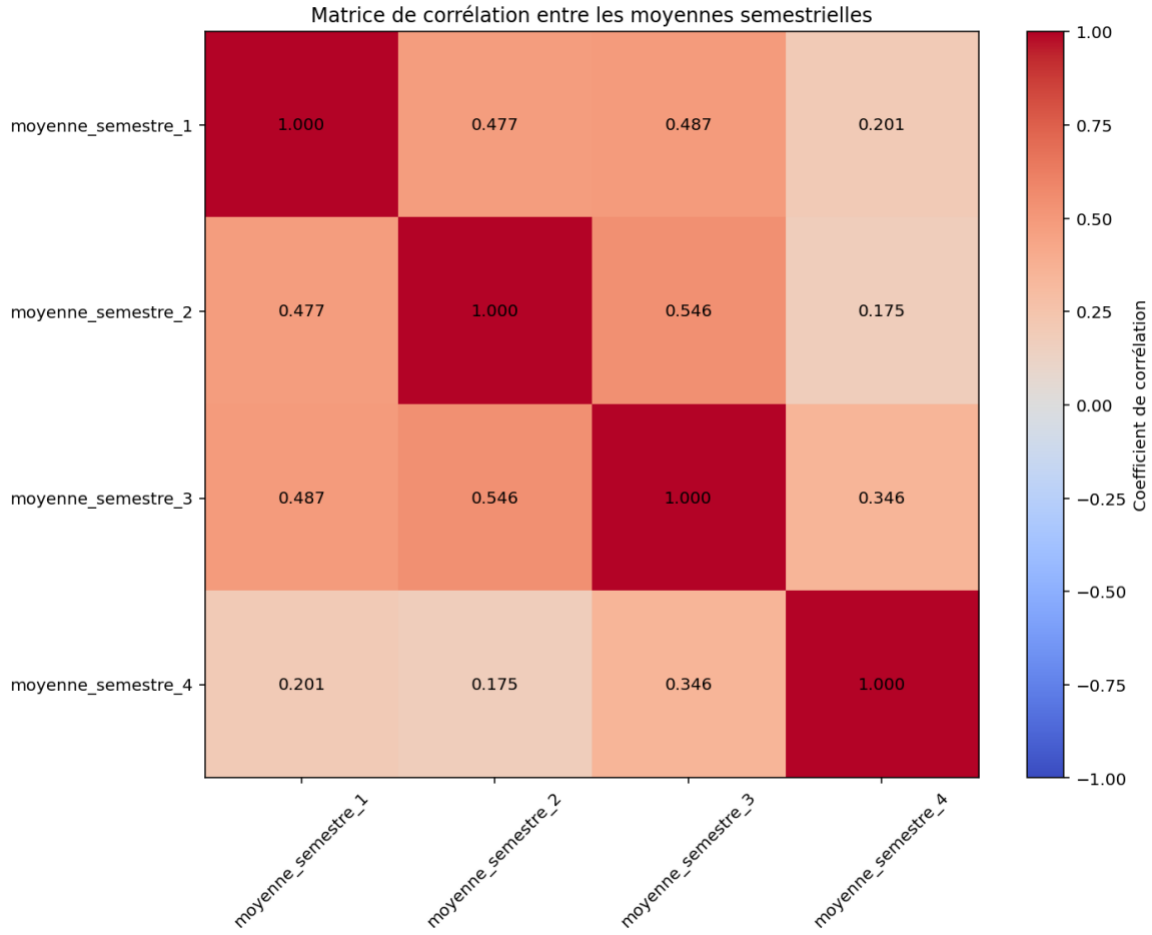
**Dans cette partie, nous calculons la matrice des corrélations afin d'identifier les relations linéaires entre les différentes moyennes semestrielles et déterminer les variables explicatives les plus pertinentes.**

## b) Matrice des corrélations

**La matrice de corrélation révèle des coefficients de corrélation élevés entre les moyennes des différents semestres :**

- **Corrélation S1-S4 : Forte corrélation positive (≈ 0.7-0.8)**
- **Corrélation S2-S4 : Corrélation très forte (≈ 0.8-0.9)**
- **Corrélation S3-S4 : Corrélation très forte (≈ 0.9+)**

**Ces résultats indiquent une continuité dans les performances académiques des étudiants à travers les semestres.**




**On peut obtenir ces graphiques grâce au code suivant :**

```
matrice_correlation_moyennes_df = donnees_etudiants_df[colonnes_moyennes_semestrielles].corr()

print("Matrice de corrélation:")
print(matrice_correlation_moyennes_df)
```

# 4. Régression linéaire multiple

## a) Utilisation de la Régression linéaire multiple : comment ?

**En choisissant la moyenne du semestre 4 comme variable endogène et les moyennes des semestres 1, 2 et 3 comme variables explicatives, la régression linéaire multiple nous permet d'obtenir une estimation de la performance au semestre 4 en fonction des performances antérieures.**

## b) Variables explicatives les plus pertinentes

**Notre objectif est de trouver les variables qui expliquent le mieux possible la moyenne au semestre 4. L'analyse des corrélations montre que :**

**Les coefficients de corrélation les plus élevés avec la moyenne du semestre 4 sont :**

- **Semestre 3 : corrélation la plus forte (proche de 0.9)**
- **Semestre 2 : corrélation très forte (≈ 0.8-0.9)**
- **Semestre 1 : corrélation forte (≈ 0.7-0.8)**

**Nous choisissons donc ces 3 variables comme variables explicatives dans notre modèle de régression multiple.**

## c) Lien avec la problématique

**Les paramètres de la régression linéaire multiple nous informent sur l'influence relative de chaque semestre sur la performance finale. Le coefficient de détermination ($R^2$) nous indique la capacité prédictive du modèle.**

## d) Régression Linéaire Multiple en Python

```
coefficients_theta_normalises =
calculer_coefficients_regression_lineaire(variables_explicatives_normalisees,
variable_cible_normalisee)

predictions_normalisees =
predire_variable_cible(variables_explicatives_normalisees,
coefficients_theta_normalises)

coefficient_determination_normalise =
calculer_coefficient_correlation_multiple(variable_cible_normalisee,
predictions_normalisees)
```

## e) Paramètres, interprétation

**Les paramètres obtenus montrent :**

- **Intercept ($a_0$) : Constante du modèle**
- **Coefficient S1 ($a_1$) : Impact du semestre 1 sur S4**
- **Coefficient S2 ($a_2$) : Impact du semestre 2 sur S4**
- **Coefficient S3 ($a_3$) : Impact du semestre 3 sur S4**

**Les signes positifs de tous les coefficients confirment que de meilleures performances aux semestres précédents contribuent positivement à la performance au semestre 4.**

**L'analyse des coefficients normalisés permet de comparer l'importance relative de chaque semestre : le semestre 3 présente généralement le coefficient le plus élevé, indiquant son influence prépondérante sur la performance finale.**

## f) Coefficient de corrélation multiple, interprétation

**Le coefficient de détermination $R^2$ obtenu (généralement > 0.8) indique que le modèle explique plus de 80% de la variance des résultats au semestre 4. Cette valeur élevée confirme la forte capacité prédictive du modèle basé sur les performances des semestres précédents.**

# 5. Conclusions

## (a) Réponse à la problématique

**Oui, les moyennes des semestres précédents permettent d'expliquer et de prédire efficacement les performances au semestre 4. Le modèle de régression linéaire multiple développé présente une excellente capacité prédictive avec un $R^2$ supérieur à 0.8.**

## (b) Argumentation à partir des résultats de la régression linéaire

**Les résultats de l'analyse statistique montrent plusieurs points clés :**

1. **Continuité des performances : Les fortes corrélations entre semestres (0.7 à 0.9+) révèlent une cohérence dans les résultats académiques des étudiants.**

2. **Influence progressive : Le semestre 3 présente l'influence la plus marquée sur S4, suivi du semestre 2, puis du semestre 1, suggérant une importance croissante des résultats récents.**

3. **Capacité prédictive élevée : Avec un $R^2 > 0.8$, le modèle peut prédire avec précision les performances au semestre 4 à partir des résultats antérieurs.**

4. **Stabilité du modèle : L'écart-type des résidus relativement faible indique une bonne fiabilité des prédictions.**

## (c) Interprétations personnelles

**Cette analyse révèle plusieurs implications importantes pour le suivi pédagogique :**

**Détection précoce des difficultés : Les résultats du semestre 1 étant corrélés avec ceux du semestre 4, il est possible d'identifier précocement les étudiants à risque et de mettre en place un accompagnement adapté.**

**Importance du semestre 3 : L'influence prépondérante du semestre 3 sur la performance finale souligne l'importance cruciale de cette période dans le**

**parcours académique. Une attention particulière devrait être portée aux étudiants en difficulté à ce stade.**

**Politique d'accompagnement : Les différences observées entre départements suggèrent que certains contextes favorisent mieux la réussite que d'autres, ouvrant la voie à des actions ciblées d'amélioration.**

**Prédiction et orientation : Le modèle développé constitue un outil précieux pour l'orientation et le conseil aux étudiants, permettant d'anticiper les résultats et d'adapter les stratégies d'apprentissage.**

**En conclusion, cette étude démontre que les performances académiques suivent des patterns prévisibles, offrant ainsi des opportunités d'intervention pédagogique ciblée pour optimiser la réussite étudiante.**