Royaume du Maroc
Université Mohammed Premier
Ecole Nationale des Sciences Appliquée
Oujda

المملكة المغربية
جامعة محمد الأول
المدرسة الوطنية للعلوم التطبيقية
وجدة

RAPPORT DU PROJET

MODÉLISATION STOCHASTIQUE

Sujet :

Application de la Régression Linéaire Multiple sur la Performance des étudiants

Encadré par : Mme Rachida ELMEHDI

Réalisé par :

- Abdelaali LAMRANI
- Chaymae MANSOURI
- Wissal ZAZOU

Filière Data Science and Cloud Computing – 2ème année

ENSAO

Email: abdelaali.lamrani20@ump.ac.ma | chaymae.mansouri@ump.ac.ma | wissal.zazou@ump.ac.ma

REMERCIMENT

À notre chère professeur Rachida ELMEHDI,

Nous tenons à exprimer notre sincère gratitude envers vous. Vos conseils éclairés et votre générosité pédagogique ont été une source constante d'inspiration tout au long de ce semestre. Nous sommes profondément reconnaissants de l'opportunité que vous nous avez accordée pour mener à bien ce projet. Votre gentillesse et votre expertise sont pour nous un modèle permanent. Veuillez accepter, chère Professeur, l'assurance de notre plus profond respect et de notre estime dans ce travail.

Sommaire

I. Introduction à la prédiction avec la Régression Linéaire Multiple dans la Modélisation Stochastique

La prédiction revêt une importance cruciale en modélisation stochastique, offrant une méthode pour comprendre les relations complexes entre variables. La régression linéaire multiple, en modélisant la relation entre une variable dépendante et plusieurs variables indépendantes, émerge comme un outil puissant. Fondée sur l'idée que la variable dépendante est une combinaison linéaire de multiples variables indépendantes, cette approche systématique permet de saisir la complexité des phénomènes influencés par plusieurs facteurs. Les avantages majeurs incluent la capacité à capturer cette complexité et à interpréter les relations spécifiques entre les variables explicatives, offrant ainsi une compréhension approfondie des mécanismes sous-jacents.

II. Présentation du projet : Modélisation de la Performance des Étudiants

1. Contextualisation du Projet

La performance académique des étudiants, cruciale tant pour les institutions éducatives que pour le développement individuel des apprenants, suscite un intérêt prépondérant. Comprendre les facteurs influençant cette performance est impératif pour mettre en œuvre des interventions ciblées en vue d'améliorer les résultats scolaires. Dans cette optique, notre projet se focalise sur la modélisation de la performance des étudiants. En utilisant la régression linéaire multiple comme outil prédictif au sein de la modélisation stochastique, nous explorons les diverses composantes de la performance académique pour identifier les variables clés influençant le Performance Index des étudiants.

2. Variables du Projet

Notre projet intègre plusieurs variables clés pour la modélisation de la performance académique ; *Les heures d'étude*, indicateur crucial, prévoient le succès scolaire, tandis que *les scores précédents* offrent une perspective sur le potentiel futur. *Les activités parascolaires* équilibrent éducation et social, influençant la performance. *Les heures de sommeil* impactent la concentration, *les exercices types* préparent aux évaluations, et le *Performance Index* synthétise la réussite académique. Cette approche vise à fournir des insights pour optimiser les politiques éducatives grâce à des modèles stochastiques.

III. Méthodologie de Modélisation Stochastique et Application de la Régression Linéaire Multiple

1. Choix Méthodologiques et Utilisation de Python

Notre approche méthodologique repose sur l'utilisation du langage de programmation *Python*, réputé pour sa flexibilité et sa richesse en bibliothèques spécialisées. *Python* offre un environnement propice à la manipulation de données et à l'implémentation efficace d'algorithmes statistiques. Les bibliothèques clés mobilisées dans notre projet comprennent :

- **pandas** : Cette bibliothèque facilite la manipulation et l'analyse des données, permettant un traitement efficace des jeux de données complexes.
- **numpy** : Fondamentale pour les calculs numériques, numpy est incontournable pour la manipulation des matrices et tableaux, élément central dans la modélisation statistique.
- **matplotlib.pyplot** et **seaborn** : Ces bibliothèques sont utilisées pour la visualisation des données, offrant des outils graphiques essentiels à la compréhension des relations entre variables.
- **scikit-learn** : Avec des modules tels que StandardScaler, LabelEncoder, et train_test_split, scikit-learn propose des outils performants pour la préparation des données et la création de jeux d'entraînement et de tests.
- **statsmodels** : Cette bibliothèque joue un rôle central dans l'application de la régression linéaire multiple, permettant une analyse approfondie des relations entre les variables.
- **scipy.stats** et **statsmodels.graphics.gofplots** : Utilisées pour des tests statistiques avancés et la vérification des hypothèses du modèle.

2. Application Pratique en Python

L'application de la méthodologie repose sur la mise en œuvre pratique de bibliothèques *Python* clés. Tout d'abord, *Pandas* est utilisé pour le chargement et la manipulation des données, offrant une structure de données tabulaire adaptée à notre ensemble d'étudiants. Ensuite, *NumPy* est exploité pour effectuer des opérations numériques fondamentales, assurant la préparation des données et les calculs nécessaires à notre analyse.

La visualisation des données est ensuite réalisée à l'aide des bibliothèques *Matplotlib* et *Seaborn*, permettant une compréhension visuelle des relations entre

différentes variables telles que les heures d'étude, les scores précédents, les activités parascolaires, etc. Ces bibliothèques graphiques contribuent à l'analyse exploratoire de manière significative.

Le processus de modélisation fait appel à la bibliothèque *Statsmodels*, où la régression linéaire multiple est implémentée pour comprendre les relations entre les variables et prédire le Performance Index des étudiants. En parallèle, *l'équation normale* est appliquée comme une approche alternative pour évaluer le modèle, soulignant la flexibilité de notre méthodologie.

IV. Cadre théorique

1. Définition

La régression linéaire multiple est une technique statistique puissante qui vise à comprendre et prédire la relation linéaire entre une variable dépendante, que l'on cherche à expliquer ou prédire, et plusieurs variables indépendantes, qui représentent les facteurs influençant la variable dépendante. Le modèle général de régression linéaire multiple peut être formulé comme suit :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad \text{pour i = 1, … ,n.}$$

- X_{ij} : sont *variables* ***explicatives***, sont *observées* et *non aléatoires*.
- y : est *la variable* ***à expliquer*** (à valeurs dans R), est *observée* et *aléatoire*.
- **β0, β1, . . ., βp :** sont les *paramètres* ***à estimer***.
- ε_i le ***terme d'erreur*** est une variable *aléatoire*, *non observée*.

La forme matricielle du modèle : $Y = X\beta + \varepsilon$

où

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{et} \quad \epsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

- Y désigne le vecteur à expliquer de taille n,
- X la matrice explicative de taille $n \times (p+1)$,
- ϵ le vecteur d'erreurs de taille n.

2. Hypothèses

Avant de plonger dans le détail de chaque hypothèse, il est essentiel de comprendre que les fondements de la régression linéaire multiple reposent sur des présuppositions statistiques critiques. Ces hypothèses garantissent les conditions requises pour une estimation fiable des paramètres du modèle et permettent la réalisation de tests d'hypothèses valables. Ces hypothèses fondamentales englobent des principes tels que *l'indépendance* et *la normalité des erreurs*, la *moyenne nulle des perturbations*, *leur dispersion constante indépendamment des variables explicatives*, ainsi que leur *indépendance mutuelle et vis-à-vis des variables exogènes*. Qui sont indispensables pour une interprétation correcte et la formulation de conclusions fiables lorsqu'on analyse des données comprenant plusieurs prédicteurs.

Passons maintenant en revue les hypothèses individuelles :

- **Hypothèse de normalité:** Les erreurs (ε) sont supposées être *indépendantes* et *distribuées normalement* avec une *moyenne de zéro* et une *variance* σ^2 *constante*. Cela implique que *la distribution des erreurs est normale dans toute la population*.

- **Hypothèse de moyenne nulle:** La moyenne des erreurs est égale à zéro$(E(\varepsilon)=0)$, indiquant que les erreurs sont *centrées autour de zéro* et qu'il *n'y a pas de biais systématique* dans les prédictions.

- **Hypothèse d'homoscédasticité:** La variance des erreurs est constante $(V(\varepsilon)=\sigma^2)$, ce qui signifie que la dispersion ou la variabilité des erreurs est la même quelle que soit la valeur de la variable indépendante.

- **Hypothèse d'indépendance des erreurs avec les variables exogènes:** Il n'existe *aucune corrélation* entre *les erreurs* et *les variables exogènes* $(Cov(x, \varepsilon)=0)$, assurant que les fluctuations dans les variables indépendantes ne sont pas systématiquement liées aux erreurs.

- **Hypothèse d'absence de corrélation entre les erreurs :** Les erreurs pour différentes observations sont *non corrélées* $(Cov(\varepsilon_i, \varepsilon_j)=0$ *pour tout* $i\neq j)$, ce qui signifie que *la présence d'une erreur pour une observation donnée n'influence pas celle des autres*.

3. Estimation des paramètres β0, β1,..,βp

Pour estimer $\beta = (\beta_0, \beta_1,..., \beta_p)$, on peut utiliser la *méthode des moindres carrés* qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de ε_i, contrairement à la méthode du maximum de *vraisemblance* qui est fondée sur la normalité de ε_i .

On cherche $\hat{\beta} \in R^{p+1}$ qui minimise la somme des erreurs quadratiques

$\varepsilon_i = (yi-\beta_0-\beta_1 x_{i1}-.....-\beta_p x_{ip})^2$

On doit donc résoudre le problème d'optimisation suivant :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} [y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]^2.$$

Le problème d'optimisation est :

$$\min_{\beta \in \mathbb{R}^{p+1}} F(\beta),$$

avec

$$\begin{aligned} F(\beta) &= \sum_{i=1}^{n} [y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]^2 \\ &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta \end{aligned}$$

Le minimum est atteint pour : $\frac{\delta F(\beta)}{\delta \beta} = 0$

Solution du problème d'optimisation on en déduit après quelques manipulations :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Sous réserve que $X^T X$ soit inversible.

4. Matrice variance-covariance

La matrice de variance-covariance est un outil statistique clé dans l'analyse des données multidimensionnelles. Elle est carrée, de dimension $n \times n$, où n représente le nombre de variables. Chaque élément de cette matrice fournit une mesure de la variance (sur la diagonale) ou de la covariance (hors diagonale) entre les paires de variables. Elle est définie par :

où

$$\Sigma = \begin{pmatrix} s_1^2 & \cdots & s_{1j} & \cdots & s_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{j1} & \cdots & s_j^2 & \cdots & s_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p1} & \cdots & s_{pj} & \cdots & s_p^2 \end{pmatrix}$$

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad ; \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

5. Matrice de corrélation

La matrice de corrélation est une matrice carrée $n \times n$ qui mesure les relations linéaires entre paires de variables. Chaque élément de la matrice indique le coefficient de corrélation entre deux variables, variant de -1 à 1. C'est une la matrice de covariance standardisée, définie par :

$$\mathrm{R} = \begin{pmatrix} 1 & \cdots & r_{1j} & \cdots & r_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{j1} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p1} & \cdots & r_{pj} & \cdots & 1 \end{pmatrix}$$

Sachant que le coefficient de corrélation est défini par :

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

6. Coefficient de détermination

Le coefficient de détermination, noté R^2, est une mesure statistique de la *qualité de l'ajustement* d'un modèle de régression linéaire. Il est défini comme le rapport entre la somme des carrés des écarts expliqués (SCE) et la somme des carrés totaux (SCT), soit la formule de R^2:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Sachant que :

SCE : *la somme des carrés des erreurs*

SCR: *la somme des carrés de la régression*

SCT= SCR+SCE *: la somme des carrés totale*

Où :

$$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$$

$$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SCT = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$$

Ce coefficient varie entre 0 et 1, où une valeur proche de 1 indique que le modèle explique une grande partie de la variabilité de la variable dépendante.

Dans certains cas, il est plus judicieux d'utiliser le coefficient de détermination ajusté ($R^2_{ajusté}$) plutôt que le coefficient de détermination classique (R^2) en raison des différences fondamentales dans leurs propriétés. Pour comprendre quand opter pour l'un plutôt que pour l'autre, il est crucial d'examiner et de comparer leurs caractéristiques spécifiques.

Les critères R ² et R ² ajusté :

- Le coefficient $R^2 = 1 - \frac{SCR}{SCT}$
  - Mesure l'ajustement du modèle aux données.
  - Augmente lorsque le nombre de variables incluses dans le modèle augmente.
  - Permet de comparer des modèles ayant le même nombre de variables.

- Le coefficient $R^2_{ajuste} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-p-1)}$
  - Estime le $R^2_{population} = 1 - \frac{V(\varepsilon)}{V(Y)} = 1 - \frac{\sigma^2}{\sigma_Y^2}$
  - N'augmente pas forcément lorsque le nombre de variables introduites dans le modèle augmente.
  - Permet de comparer des modèles ayant un nombre de variables différent.

7. Problème de multi-colinéarité

La présence de multi-colinéarité pose un problème majeur dans l'analyse de régression. Elle se produit lorsque *les* ***variables explicatives*** *sont fortement* ***interdépendantes***, ce qui peut biaiser l'estimation des paramètres du modèle. Pour détecter cette multi-colinéarité, une première étape consiste à examiner la matrice de corrélation entre les variables explicatives. Des corrélations élevées entre certaines paires de prédicteurs peuvent révéler des interactions potentielles, et dans certains cas, il peut être nécessaire de retirer du modèle des prédicteurs fortement corrélés. Cela est important car des variables fortement corrélées fournissent souvent des *informations redondantes*, et leur suppression ne diminue généralement pas considérablement le coefficient de détermination (R^2). Parfois, il peut également être nécessaire de tester *l'interaction* entre trois variables explicatives, ce qui implique de régresser l'effet combiné de deux d'entre elles sur la troisième. Cependant, cette approche peut devenir complexe et lourde, en particulier lorsque le nombre de

variables explicatives est important. Il est donc essentiel de gérer la multi-colinéarité de manière appropriée pour garantir des résultats fiables dans l'analyse de régression.

8. Intervalles de confiance

Lorsqu'on souhaite estimer les valeurs des paramètres inconnus, une simple estimation ponctuelle ne suffit généralement pas pour prendre des décisions robustes. C'est pourquoi on préfère construire des intervalles de confiance. Pour ce faire, il est essentiel de définir la distribution des estimations des coefficients. En pratique, on suppose généralement un niveau de confiance de 95% (ce qui correspond à α= 0,05). Cela signifie que l'on souhaite construire un intervalle de confiance à 95% de certitude pour les valeurs des paramètres inconnus on a :

$$\beta_j \in \left[\hat{\beta}_j \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}\right]$$

d'où

$$Pr\left(\hat{\beta}_j - t_{n-(p+1),1-\frac{\alpha}{2}}\, \hat{\sigma}_\varepsilon \sqrt{[(X^t.\, X)^{-1}]_{jj}} \le \beta_j \le \hat{\beta}_j + t_{n-(p+1),1-\frac{\alpha}{2}}\, \hat{\sigma}_\varepsilon \sqrt{[(X^t.\, X)^{-1}]_{jj}}\right) = 1-\alpha$$

9. Tests d'hypothèse sur les paramètres

Il paraît raisonnable de se poser les questions suivantes :

- *Est-ce-que la variable x_j a-t-elle réellement une influence sur Y ?*
- *Tester la significativité globale du modèle, c'est à dire est-ce-que que tous les coefficients sont supposés nuls, excepté la constante ?*

Nous pouvons expliciter les questions précédentes en termes de test d'hypothèse correspond à :

- **H_0 : $\beta_i = 0$,** contre **H_1 : $\beta_i \neq 0$.** Cela revient au **test de Student** à n−p−1 degré de liberté. La *statistique du test Student t est* définie par :

$$t_{cal} = \frac{\widehat{\beta_j} - \beta_{j0}}{\hat{s}_{\hat{\beta}_j}} = \frac{(\widehat{\beta_j} - \beta_j)/s_{\hat{\beta}_j}}{\sqrt{\chi^2_{(n-p-1)}/(n-p-1)}}$$

Etant un test bilatéral, la règle de décision du test est :

Si $|t_{cal}| > t_{n-p-1,1-\frac{\alpha}{2}}$ **on rejette H_0** au niveau de risque α.

- **H_0 : tous les $\beta_i = 0$, $j = 1, \ldots, p$** contre **H_1 : $\exists j \; tel \; que \; \beta_j \neq 0$.** La *statistique du test Fisher F* est définie par :

$$F_{cal} = \frac{\frac{SCReg}{p}}{\frac{SCE}{n-p-1}}$$

Etant un test bilatéral, la règle de décision du test est :

Si $F_{cal} > F_{p,n-p-1,1-\alpha}$ **on rejette H_0** au niveau de risque α.

10. ANOVA

A. Définition

L'Analyse de la Variance, plus communément connue sous l'acronyme ANOVA, est une technique statistique fondamentale qui joue un rôle crucial dans l'analyse des données expérimentales. Elle est utilisée pour déterminer si les différences entre les moyennes de différents groupes sont statistiquement significatives. Cette méthode est particulièrement utile dans des situations où l'on souhaite comparer trois groupes ou plus, car des tests traditionnels comme le test t de Student ne sont pas adaptés à ces cas de figure.

B. Table ANOVA

La table d'ANOVA, un outil central en analyse statistique, a été développée bien avant l'ère des ordinateurs. Elle résume tous les calculs nécessaires pour effectuer un test d'hypothèse spécifique en statistiques : celui qui compare les moyennes de plusieurs groupes.

Ce test, souvent formulé comme *H0 : μ1=μ2=⋯=μI contre H1 : il existe au moins un couple de moyennes (μi, μi') qui sont différentes*, est crucial lorsque la variance des données est inconnue, une situation fréquente dans la recherche pratique. La table ANOVA est structurée de manière spécifique pour refléter les sources de variation dans les données :

Source de variation	Somme des carrés	ddl	carré moyen	F
régression (expliquée)	$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$	p	$\frac{1}{p}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$	$\frac{SCE/p}{SCR/(n-p-1)}$
Résiduelle	$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-(p+1)	$\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	
Totale	$SCT = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$	n-1	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$	

Remarques :

- On retrouve la statistique dite de Fisher Fn qui permet de tester l'ajustement du modèle.

- On retrouve la propriété fondamentale SCT = SCE + SCR qui permet de mesurer l'ajustement du modèle par le coefficient de détermination.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

- Le coefficient R^2 donne la proportion de variabilité de y qui est expliquée par le modèle. Plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données.

11. Analyse des résidus

L'analyse des résidus joue un rôle crucial dans l'évaluation de la pertinence d'un modèle de régression. Les *résidus*, également connus sous le terme *d'erreurs observées*, sont définis comme les écarts entre les valeurs réelles observées dans les données (y_i) et les valeurs prédites par le modèle de régression ($\hat{y}$). Formellement, ils sont exprimés par la formule $\varepsilon_i = y_i - \hat{y}$. Ces résidus représentent la portion des données qui n'est pas expliquée par le modèle de régression. L'objectif de l'analyse des résidus est de *tester la validité du modèle de régression utilisé*. Cette analyse est essentielle, car elle permet de révéler des problèmes dans le modèle qui pourraient ne pas être apparents lors de l'examen initial des variables. Les résidus peuvent mettre en évidence divers problèmes tels que les valeurs aberrantes, les non-linéarités et l'omission de groupes d'observations significatifs. En somme, l'analyse des résidus est un outil diagnostique indispensable pour évaluer si les hypothèses de base du modèle de régression sont respectées.

Il existe principalement quatre hypothèses clés dans l'analyse des résidus, qui doivent être vérifiées pour assurer la validité du modèle :

- *Normalité des erreurs:* Cette hypothèse peut être vérifiée en examinant la distribution des résidus, par exemple, à l'aide d'un histogramme ou d'un graphe quantile-quantile normal (Normal Q-Q plot).
- *Homoscédasticité:* Cette hypothèse implique que la variance des erreurs est constante. L'hétéroscédasticité, où la variance des erreurs varie, peut être un signe de problème dans le modèle.
- *Absence de Multi-colinéarité :* Dans un modèle de régression linéaire, il est supposé que les variables indépendantes ne sont pas fortement corrélées entre elles. La présence d'une forte corrélation, ou multi- colinéarité, peut fausser les résultats.
- *Absence d'Autocorrélation :* Les termes d'erreur ne doivent pas être corrélés les uns avec les autres. L'autocorrélation est souvent un problème dans les données de séries temporelles.

Si un modèle de régression ne satisfait pas à ces hypothèses, sa fiabilité et sa validité peuvent être remises en question, et il pourrait ne pas être approprié de l'utiliser pour des prédictions ou des analyses supplémentaires.

Autocorrélation_: signifie que les termes d'erreur ne doivent pas être corrélés les uns avec les autres.

V. Cadre pratique

1. Importation des données

Les bibliothèques :

Préparons notre environnement avec les librairies dont nous aurons besoin

```
import pandas as pd
import numpy as np
import  matplotlib.pyplot  as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
from scipy.stats import f
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.anova import anova_lm
```

Importation des données

On utilise pandas pour charger l'ensemble de données dans ce bloc-notes. En utilisant pandas, nous pouvons lire notre fichier de données (Student_Performancee.csv)

```
df=pd.read_csv("Student_Performance.csv",header=0)
```

Visualiser le type des données des colonnes et quelques informations

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 6 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Hours Studied                   10000 non-null  int64  
 1   Previous Scores                 10000 non-null  int64  
 2   Extracurricular Activities      10000 non-null  object 
 3   Sleep Hours                     10000 non-null  int64  
 4   Sample Question Papers Practiced  10000 non-null  int64  
 5   Performance Index               10000 non-null  float64
dtypes: float64(1), int64(4), object(1)
memory usage: 468.9+ KB
```

Dimension du dataset

	Nombre de lignes	Nombre de colonnes
0	10000	6

Notre jeu de données contient 1000 Lignes et 6 colonnes.

Renommer les colonnes en français

	Nom de la colonne
0	Heures_détude
1	Scores_précédents
2	Activités_parascolaires
3	Heures_de_sommeil
4	Exercices_pratiqués_sur_des_questions_types
5	Performance Index

2. Nettoyage du dataset

```
1 df.isna().sum()
```

```
Heures_détude                                0
Scores_précédents                            0
Activités_parascolaires                      0
Heures_de_sommeil                            0
Exercices_pratiqués_sur_des_questions_types  0
Performance Index                            0
const                                        0
dtype: int64
```

Le dataset ne contient aucune valeur manquante

```
1 df.duplicated().sum()
```

```
127
```

Le dataset contient 127 lignes dupliquées.

```
1 df.drop_duplicates(inplace=True)
2 df.duplicated().sum()
```

Supprimer les duplications puis compter le nombre total des duplications dans l'ensemble du DataFrame

Il semble clair que nous n'avons aucune valeur manquante/nulle dans notre ensemble de données !

3. Analyse exploratoire des données

```
1 df.head()
```

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
0	7	99	Yes	9	1	91.0
1	4	82	No	4	2	65.0
2	8	51	Yes	7	2	45.0
3	5	52	Yes	5	2	36.0
4	7	75	No	8	5	66.0

En scrutant les cinq premières lignes de données, il est essentiel d'acquérir une compréhension approfondie de chaque colonne. Cette approche nous permettra de poursuivre notre exploration de l'ensemble de données de manière efficace. Il est crucial de saisir clairement le rôle de chaque caractéristique dans l'ensemble de données et de comprendre l'information qu'elle cherche à transmettre.

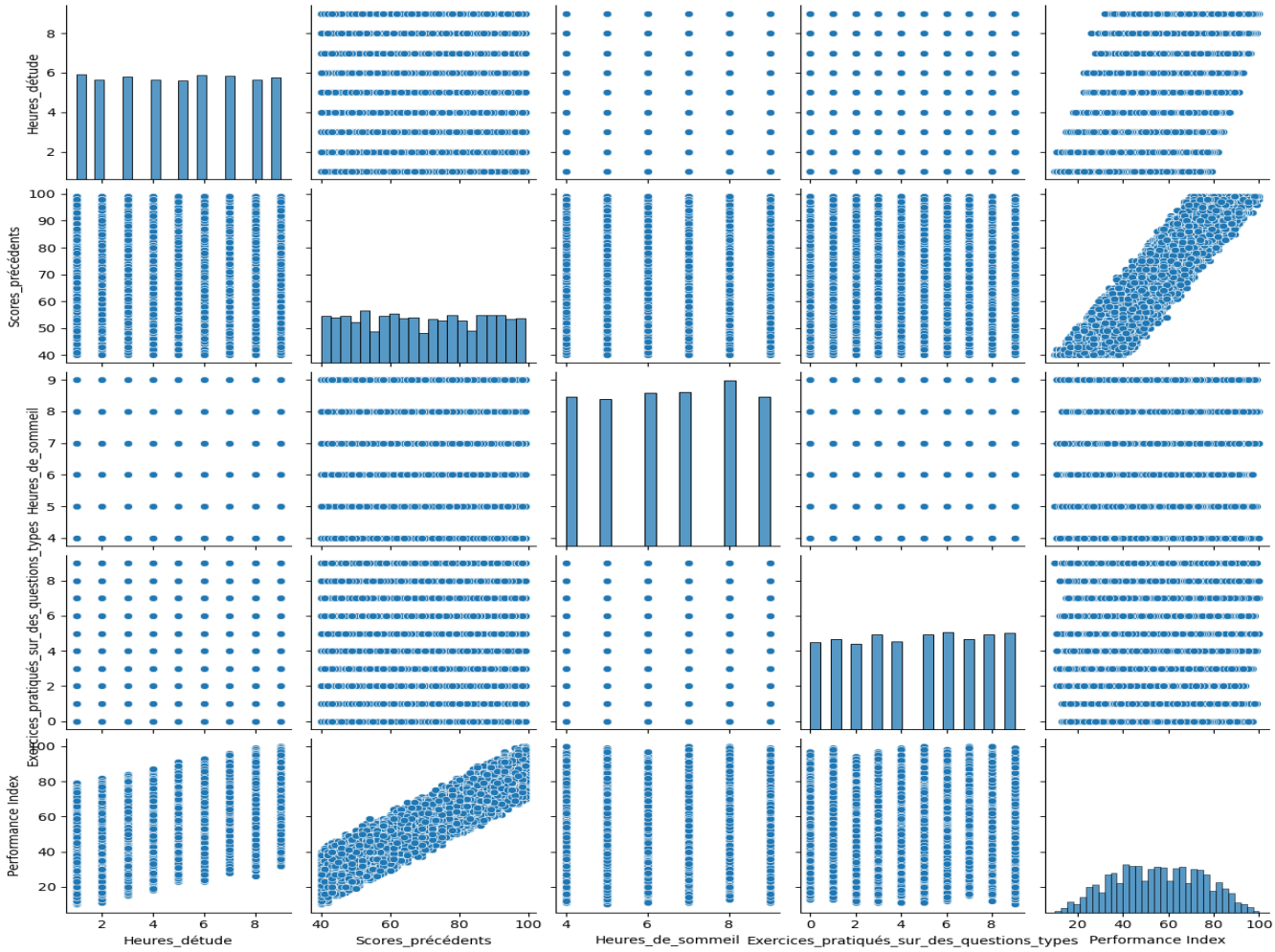
```
1 df.describe()
```

	Heures_détude	Scores_précédents	Activités_parascolaires	Heures_de_sommeil	Exercices_pratiqués_sur_des_questions_types	Performance Index	const
count	9873.000000	9873.000000	9873.000000	9873.000000	9873.000000	9873.000000	9873.0
mean	4.992100	69.441102	0.494986	6.531652	4.583004	55.216651	1.0
std	2.589081	17.325601	0.500000	1.697683	2.867202	19.208570	0.0
min	1.000000	40.000000	0.000000	4.000000	0.000000	10.000000	1.0
25%	3.000000	54.000000	0.000000	5.000000	2.000000	40.000000	1.0
50%	5.000000	69.000000	0.000000	7.000000	5.000000	55.000000	1.0
75%	7.000000	85.000000	1.000000	8.000000	7.000000	70.000000	1.0
max	9.000000	99.000000	1.000000	9.000000	9.000000	100.000000	1.0

Voir des informations sur les valeurs numériques :

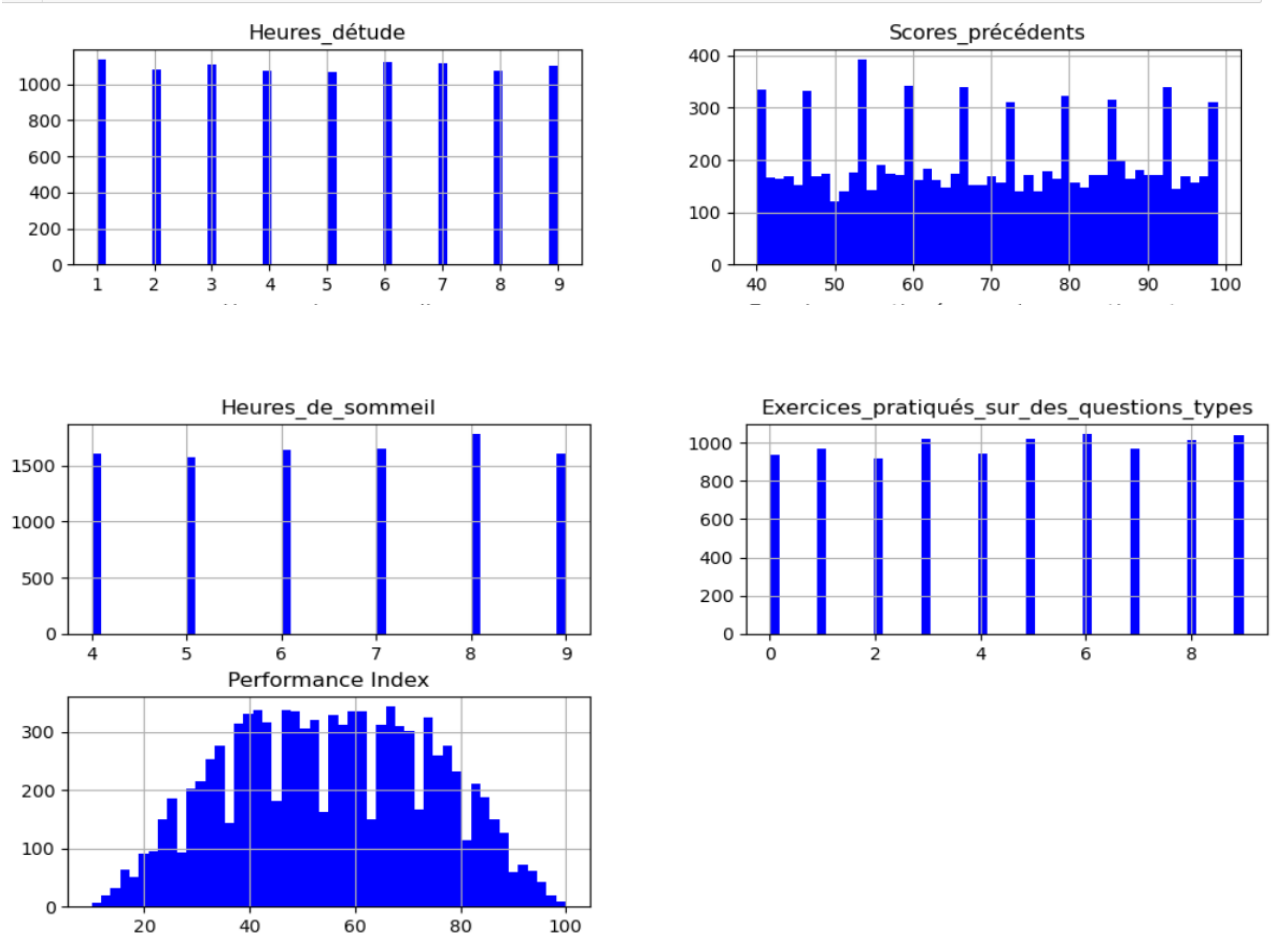
Les statistiques descriptives du DataFrame fournissent une vue d'ensemble rapide mais informative de notre ensemble de données. Elles révèlent la diversité des habitudes d'étude, des performances antérieures, et de la participation à des activités parascolaires parmi les étudiants. Ces chiffres permettent d'appréhender rapidement la moyenne, l'écart-type, les valeurs minimales et maximales de chaque variable, donnant ainsi un aperçu global des tendances centrales et de la variabilité. Cette compréhension préliminaire est essentielle pour orienter davantage l'exploration de données et identifier les domaines clés qui pourraient nécessiter une attention particulière dans une analyse plus approfondie.

4. Visualisation des données



La matrice que nous observons est composée de scatter plots ou graphiques de dispersion qui montre la relation entre plusieurs variables. Les histogrammes sur la diagonale indiquent la distribution de chaque variable individuelle, et les plots hors diagonale montrent les relations entre elles. La corrélation est suggérée par la formation de tendances dans ces scatter plots: Par exemple, une relation linéaire positive est évidente entre "Heures_d'étude" et "Performance index", où les points s'alignent le long d'une ligne ascendante, indiquant que plus les élèves étudient, mieux ils performent. D'autres plots montrent un nuage de points dispersé sans tendance claire, comme entre "Heures_de_sommeil" et "Performance index", suggérant l'absence de corrélation.

Histogrammes pour chaque variable



Ces graphiques à barres illustrent la distribution des différentes variables du notre dataset.

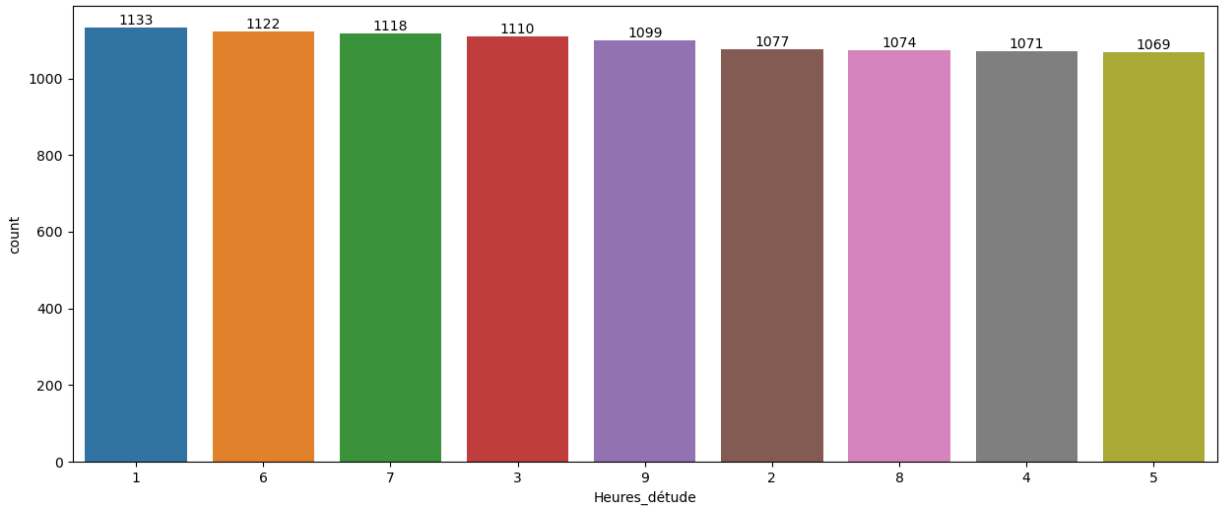
"Heures d'étude": Distribution uniforme, suggérant une répartition équitable d'étudiants pour chaque catégorie d'heures d'étude.

"Scores précédents": Répartition uniforme, indiquant une diversité considérable dans les résultats académiques passés sans prédominance de clusters spécifiques.

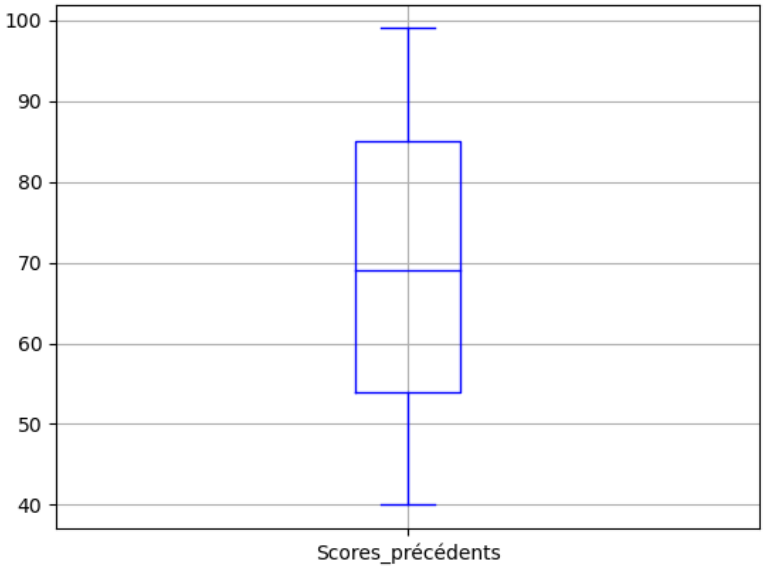
"Heures de sommeil": Légère inclinaison vers des valeurs plus élevées, suggérant que la majorité des étudiants dorment un nombre d'heures plus élevé.

"Exercices pratiqués sur des questions types": Distribution uniforme, suggérant une pratique fréquente et équivalente indépendamment du nombre d'exercices.

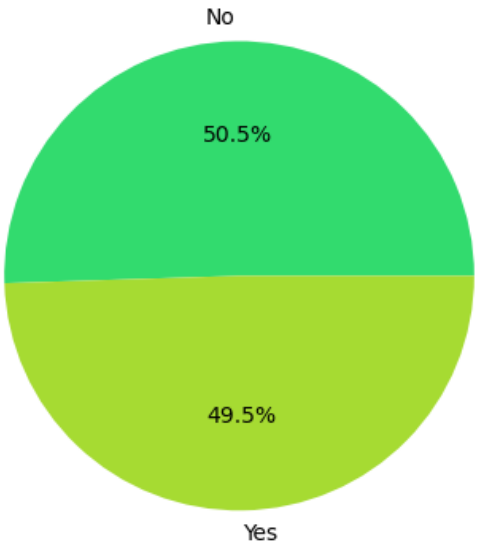
"Index de Performance": Distribution normale, montrant la concentration autour d'un score moyen avec moins d'extrêmes, illustrant la répartition des résultats académiques.



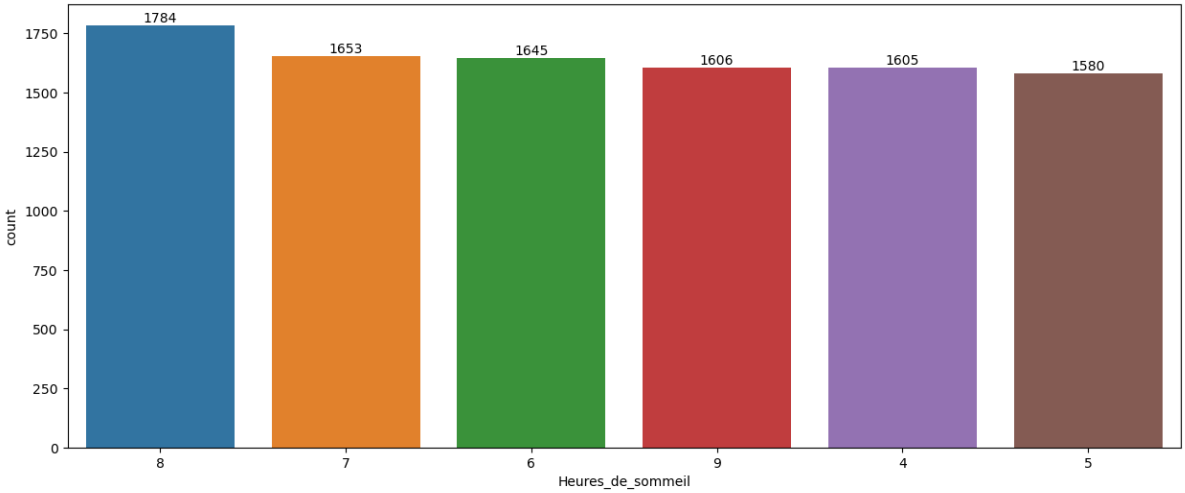
Le graphique à barres illustre le nombre d'étudiants répartis par le nombre d'heures d'étude par jour. Les données montrent une distribution relativement uniforme des étudiants s'étalant de 1 à 9 heures d'étude, avec une légère préférence pour 1 heure (1133 étudiants) et une légère baisse pour 5 heures (1069 étudiants). Les différences entre le nombre d'étudiants pour chaque durée d'étude sont minimes, suggérant qu'il n'y a pas de durée d'étude nettement plus privilégiée que les autres parmi cette population étudiante.



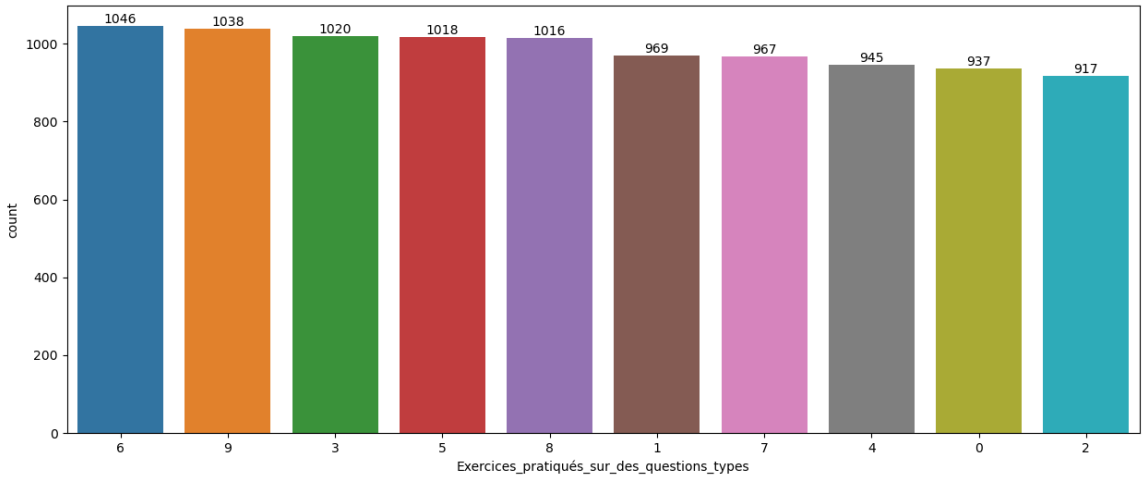
Le diagramme en boîte pour les "Scores précédents" révèle une médiane de score environ 70, ce qui signifie que 50% des étudiants ont des scores au-dessous de cette valeur et les autres 50% au-dessus. Le premier quartile se situe aux alentours de 53, indiquant que 25% des étudiants ont des scores en dessous de ce chiffre, et le troisième quartile est approximativement à 85, montrant que 75% des étudiants ont des scores en dessous de ce niveau. L'écart entre Q1 et Q3 révèle que la majorité des étudiants ont des scores qui se situent dans cette gamme, avec peu de dispersion en dehors de ces limites, ce qui indique une concentration des scores sans grande variabilité extrême.



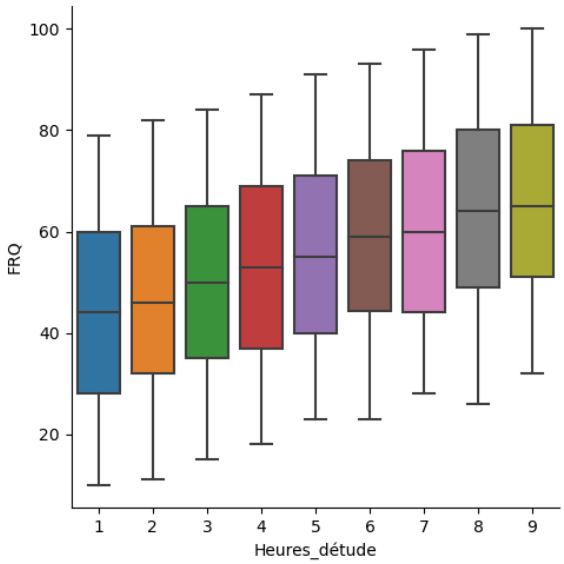
L'autopct montre que les étudiants sont presque également répartis en ce qui concerne leur participation à des activités parascolaires. Avec 49,5% des étudiants qui s'engagent dans des activités parascolaires (indiqué par "Yes") et 50,5% qui ne participent pas (indiqué par "No"), on observe une division presque parfaite au sein de la population étudiée. Cette répartition quasi équilibrée suggère que l'engagement dans les activités parascolaires est aussi commun que son absence parmi les étudiants concernés.



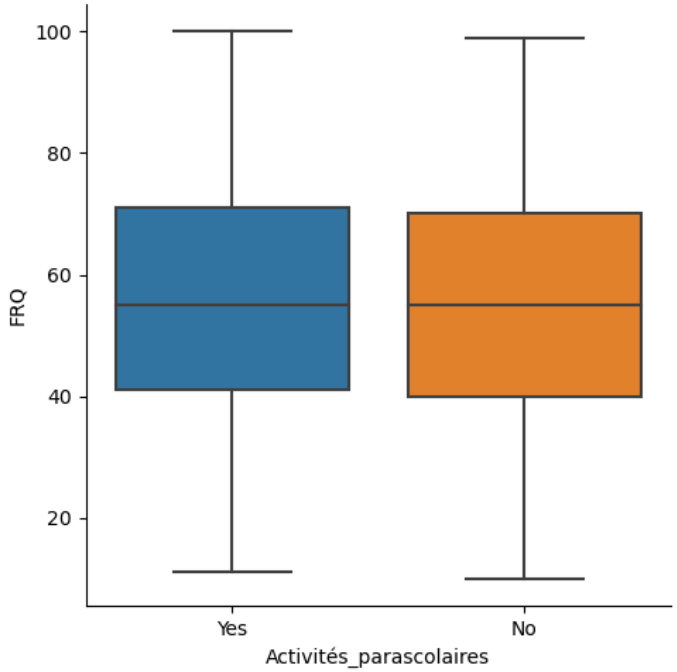
Le graphique à barres affiche le nombre d'étudiants selon le nombre d'heures de sommeil par nuit, avec une tendance montrant que la majorité des étudiants dorment 8 heures (1784 étudiants), suivi de près par ceux qui dorment 7 heures (1653 étudiants). Il y a une légère diminution du nombre d'étudiants qui dorment 6 heures (1645 étudiants) et ceux qui dorment 9 heures (1606 étudiants). Le graphique révèle également que moins d'étudiants dorment 5 heures (1580 étudiants) et le moins d'étudiants dorment 4 heures (1605 étudiants). Cette distribution suggère que, dans cette population étudiante, la durée de sommeil la plus commune est de 8 heures, et il y a une préférence générale pour une plage de sommeil de 6 à 9 heures par nuit.



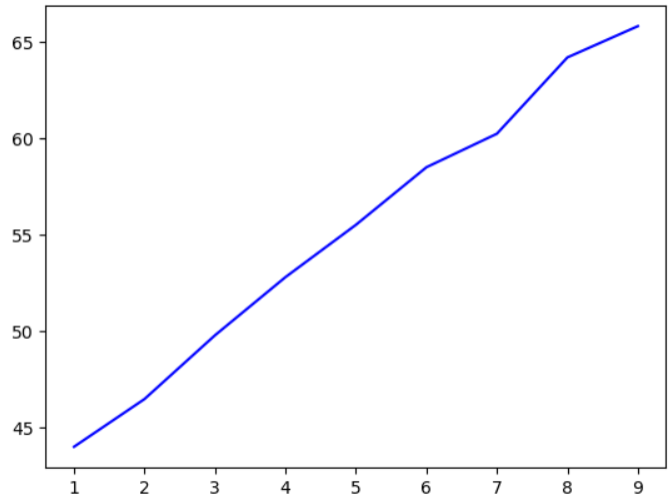
La répartition des pratiques d'exercices montre une diversité significative dans les habitudes des étudiants. La pratique de 6 exercices est la plus fréquente, suivie de près par ceux qui en font 9, 3 , 5 exercices, 8 exercices, et 1 exercice. Les variations dans les nombres d'étudiants pratiquant 7, 4, 0, et 2 exercices indiquent une dispersion dans les préférences d'exercices. En conclusion, cette distribution révèle une diversité d'approches dans les pratiques d'exercices parmi les étudiants, suggérant une adaptation aux préférences individuelles en matière d'études.



Le diagramme en boîte à moustache dépeint la relation entre le nombre d'heures d'étude et la fréquence des performances académiques élevées chez les étudiants. Les médianes semblent indiquer une tendance générale où les performances augmentent avec le nombre d'heures d'étude, comme en témoigne l'augmentation graduelle des valeurs médianes à mesure que le nombre d'heures d'étude s'accroît. Malgré la variabilité significative dans les données, soulignée par la longueur des moustaches, il n'y a pas de valeurs aberrantes marquées qui indiqueraient des anomalies. Cette distribution suggère que, en général, les étudiants qui consacrent plus de temps aux études ont tendance à avoir de meilleures performances académiques.



Le diagramme en boîte compare la fréquence des performances académiques (FRQ) des étudiants qui participent à des activités parascolaires (Yes) et ceux qui n'y participent pas (No). Les deux boîtes ont des médianes proches l'une de l'autre, indiquant que la participation à des activités parascolaires n'affecte pas significativement la performance moyenne des étudiants.



Ce graphique linéaire montre une tendance positive entre le nombre d'heures d'étude et les performances académiques. L'axe des abscisses représente les heures d'étude, et l'axe des ordonnées indique la performance. On peut observer une augmentation régulière et constante de la performance à mesure que le nombre d'heures d'étude augmente, allant de la valeur la plus basse à gauche (pour 1 heure d'étude) jusqu'à la valeur la plus haute à droite (pour 9 heures d'étude). Cette corrélation positive suggère que les étudiants qui étudient plus longtemps tendent à avoir de meilleures performances académiques.

Corrélation

La matrice de corrélation est un instrument clé en statistique qui évalue la force et la direction des *relations entre variables*, variant de *-1* pour *une corrélation négative parfaite* à *+1* pour *une corrélation positive parfaite*, avec *0* indiquant *l'absence de corrélation*. Dans une telle matrice, une variable est toujours parfaitement corrélée avec elle-même (corrélation de 1 sur la diagonale), et les variations de couleur donnent une indication visuelle de la force de la corrélation entre les variables distinctes. Dans notre cas, l'analyse révèle que les "Heures d'étude" et le "Performance Index" sont modérément corrélés positivement, ce qui suggère que plus les étudiants étudient, mieux ils peuvent performer. Les "Scores précédents" ont une forte corrélation positive avec le "Performance Index", ce qui renforce l'idée que les performances passées pourraient prédire les futures réussites académiques. Par contre, les "Heures de sommeil" ne montrent pas de corrélation notable avec les performances académiques, indiquant que le sommeil n'est pas directement lié aux résultats scolaires dans cette matrice. De façon inattendue, la pratique à travers des "Exercices sur des questions types" n'est que faiblement corrélée avec le "Performance Index", ce qui peut remettre en cause la supposition commune que plus d'exercice conduit à de meilleures performances. Globalement, cette matrice suggère que si certaines pratiques d'études sont significativement liées au succès académique, d'autres, à notre surprise, semblent avoir peu d'impact direct sur les performances.

5. Préparation des données

Encoder les valeurs de la colonne "Activités parascolaires" de (Yes, No) à (1, 0) afin de rendre l'ensemble de données entièrement numérique.

	Heures_détude	Scores_précédents	Activités_parascolaires	Heures_de_sommeil	Exercices_pratiqués_sur_des_questions_types	Performance Index
0	7	99	1	9	1	91.0
1	4	82	0	4	2	65.0
2	8	51	1	7	2	45.0
3	5	52	1	5	2	36.0
4	7	75	0	8	5	66.0

Séparation des données en variables indépendantes et dépendante.

```
X = df.drop(columns = "Performance Index")#Train
y = df["Performance Index"]#Target
```

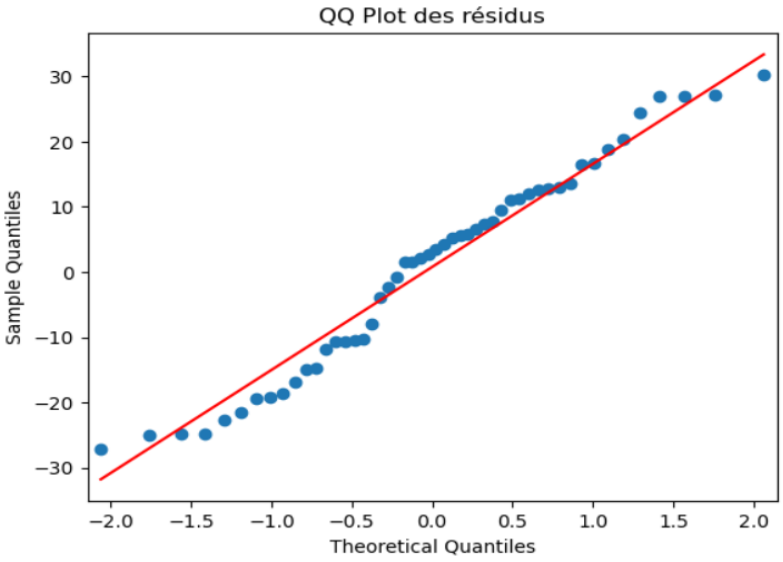
La division des données en variables indépendantes et dépendantes consiste à séparer les caractéristiques utilisées pour prédire ou expliquer (indépendantes) de la caractéristique que l'on cherche à prédire ou expliquer (dépendante)

6.Régression linéaire simple sur chaque variable

Regression sur 'Heures_détude'

```
1 lm('Heures_détude')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Performance Index   R-squared:                       0.141
Model:                            OLS   Adj. R-squared:                  0.141
Method:                 Least Squares   F-statistic:                     1619.
Date:                Thu, 04 Jan 2024   Prob (F-statistic):               0.00
Time:                        20:41:55   Log-Likelihood:                -42437.
No. Observations:                9873   AIC:                         8.488e+04
Df Residuals:                    9871   BIC:                         8.489e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          41.3156      0.389    106.148      0.000      40.553      42.079
Heures_détude   2.7846      0.069     40.232      0.000       2.649       2.920
==============================================================================
```



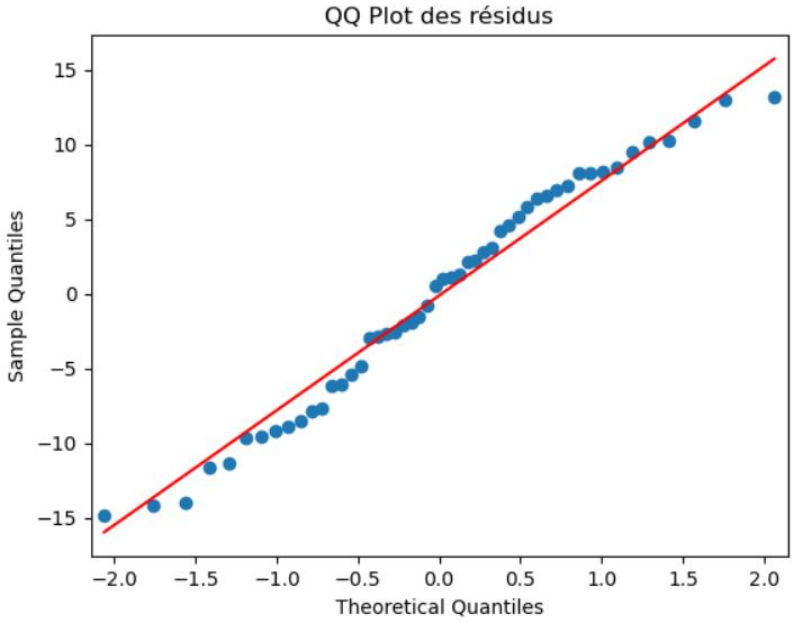
	Heures_détude
R2	14.087413
R2_adj	14.078710
Fcalc	1618.585346
Ftbl_model	3.842401

*On a **Fcal >> Ftbl** donc on rejette au niveau de risque 5% l'hypothèse H0 d'où : **Le modèle est globalement significatif***

Regression sur 'Scores_précédents'

```
1 lm('Scores_précédents')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Performance Index   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                 5.086e+04
Date:                Thu, 04 Jan 2024   Prob (F-statistic):               0.00
Time:                        20:41:55   Log-Likelihood:                -34218.
No. Observations:                9873   AIC:                         6.844e+04
Df Residuals:                    9871   BIC:                         6.845e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            -15.2378      0.322    -47.326      0.000     -15.869     -14.607
Scores_précédents  1.0146      0.004    225.529      0.000       1.006       1.023
==============================================================================
```



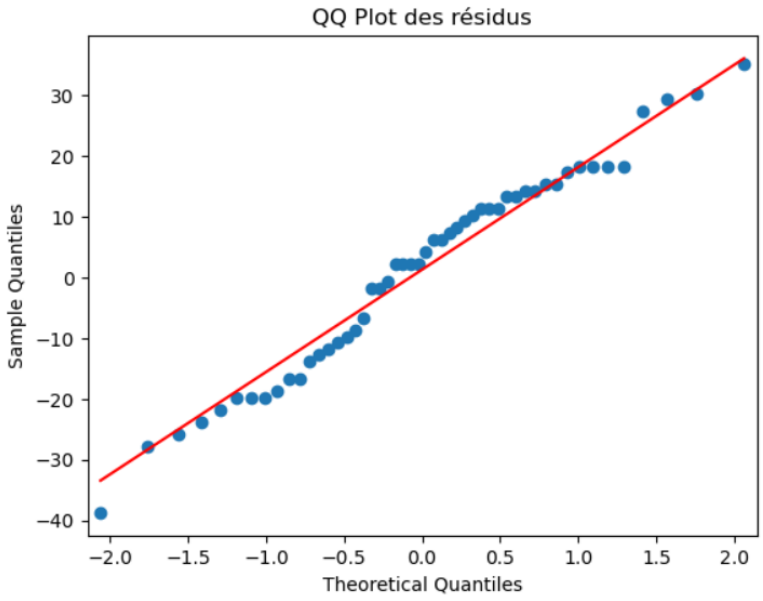
	Scores_précédents
R2	83.747221
R2_adj	83.745575
Fcalc	50863.230088
Ftbl_model	3.842401

*On a **Fcal >> Ftbl** donc on rejette au niveau de risque 5% l'hypothèse H0 d'où : **Le modèle est globalement significatif***

Regression sur 'Activités_parascolaires'

```
1 lm('Activités_parascolaires')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Performance Index   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     6.716
Date:                Thu, 04 Jan 2024   Prob (F-statistic):            0.00957
Time:                        20:41:56   Log-Likelihood:                -43184.
No. Observations:                9873   AIC:                          8.637e+04
Df Residuals:                    9871   BIC:                          8.639e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           54.7208      0.272    201.215      0.000      54.188      55.254
Activités_parascolaires          1.0017      0.387      2.591      0.010       0.244       1.759
==============================================================================
```



:

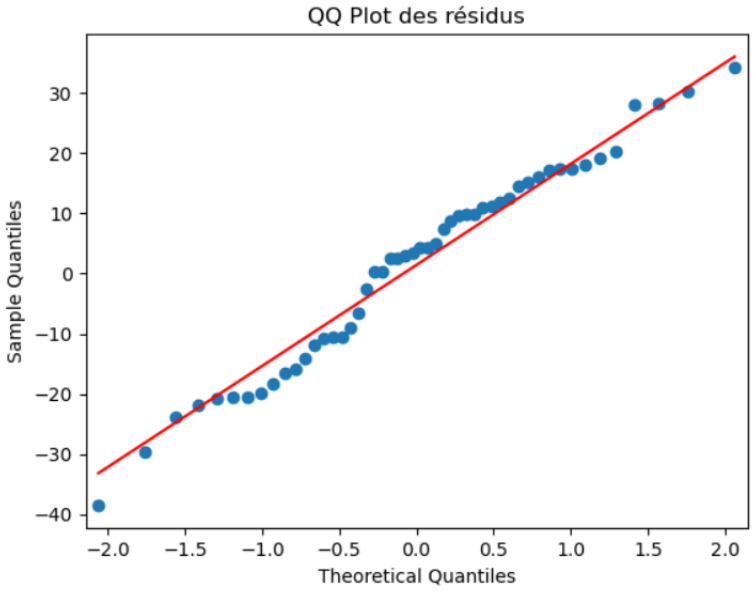
	Activités_parascolaires
R2	0.067988
R2_adj	0.057865
Fcalc	6.715704
Ftbl_model	3.842401

*On a **Fcal > Ftbl** donc on rejette au niveau de risque 5% l'hypothèse H0 d'où : **Le modèle est globalement significatif***

Regression sur 'Heures_de_sommeil'

```
1 lm('Heures_de_sommeil')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Performance Index   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     25.09
Date:                Thu, 04 Jan 2024   Prob (F-statistic):           5.57e-07
Time:                        20:41:56   Log-Likelihood:                -43174.
No. Observations:                9873   AIC:                         8.635e+04
Df Residuals:                    9871   BIC:                         8.637e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         51.4955      0.768     67.088      0.000      49.991      53.000
Heures_de_sommeil     0.5697      0.114      5.009      0.000       0.347       0.793
==============================================================================
```



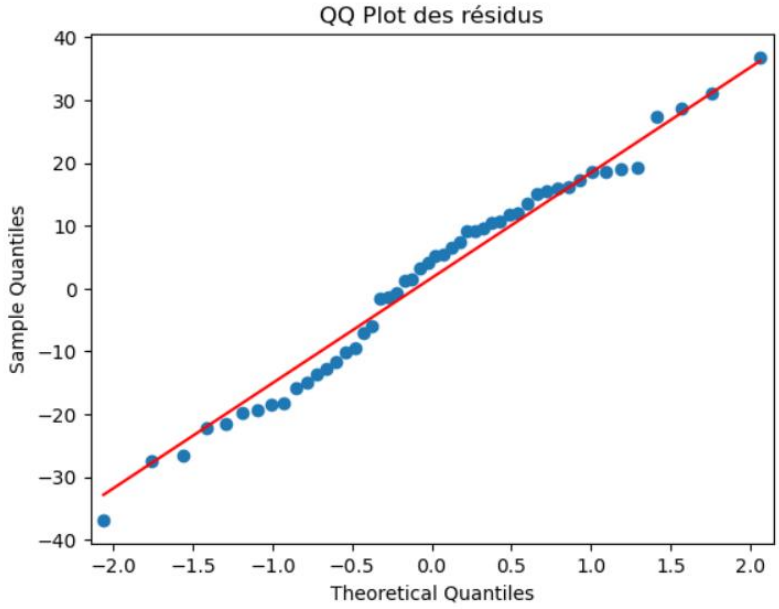
	Heures_de_sommeil
R2	0.253537
R2_adj	0.243432
Fcalc	25.090267
Ftbl_model	3.842401

*On a **Fcal > Ftbl** donc on rejette au niveau de risque 5% l'hypothèse H0 d'où : **Le modèle est globalement significatif***

Regression sur 'Exercices_pratiqués_sur_des_questions_types'

```
1 lm('Exercices_pratiqués_sur_des_questions_types')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Performance Index   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     18.66
Date:                Thu, 04 Jan 2024   Prob (F-statistic):           1.58e-05
Time:                        20:41:57   Log-Likelihood:                -43178.
No. Observations:                9873   AIC:                          8.636e+04
Df Residuals:                    9871   BIC:                          8.637e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================================================
                                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                         53.8830      0.364    147.957      0.000      53.169      54.597
Exercices_pratiqués_sur_des_questions_types     0.2910      0.067      4.320      0.000       0.159       0.423
==============================================================================
```



	Exercices_pratiqués_sur_des_questions_types
R2	0.188666
R2_adj	0.178555
Fcalc	18.658435
Ftbl_model	3.842401

*On a **Fcal > Ftbl** donc on rejette au niveau de risque 5% l'hypothèse H0 d'où : **Le modèle est globalement significatif***

7. Régression linéaire multiple sur toutes les variables

A. Modèle avec statsmodels

a) Construction et entraînement du modèle

```
def lm1(X,y):
    return sm.OLS(y, sm.add_constant(X)).fit()
```

```
X = df.drop(columns = ["const","Performance Index"])#Train
y = df["Performance Index"]#Target
model=lm1(X,y)
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	-34.0689	0.128	-265.875	0.000	-34.320	-33.818
Heures_détude	2.8527	0.008	358.940	0.000	2.837	2.868
Scores_précédents	1.0183	0.001	857.427	0.000	1.016	1.021
Activités_parascolaires	0.6167	0.041	14.981	0.000	0.536	0.697
Heures_de_sommeil	0.4803	0.012	39.623	0.000	0.457	0.504
Exercices_pratiqués_sur_des_questions_types	0.1939	0.007	27.017	0.000	0.180	0.208

Le tableau issu de la régression linéaire multiple offre des informations essentielles. Les coefficients mesurent l'influence moyenne des prédicteurs, tandis que la "statistique t" teste leur différence significative de zéro. Des valeurs de "P>|t|" inférieures à 0,05 indiquent une forte probabilité de significativité. Les faibles erreurs standard renforcent la fiabilité des résultats. L'intervalle de confiance ([0.025 0.975]) fournit une plage probable pour les coefficients, avec des intervalles excluant zéro renforçant la significativité. Malgré la significativité globale, le coefficient constant mérite une attention particulière en raison de son impact négatif sur la variable dépendante.

b) Prédiction pour les 5 premières valeurs

	Exemple1
Valeur prédite	91.847305
Valeur réelle	91.000000
	Exemple2
Valeur prédite	63.153299
Valeur réelle	65.000000
	Exemple3
Valeur prédite	45.053968
Valeur réelle	45.000000
	Exemple4
Valeur prédite	36.553459
Valeur réelle	36.000000
	Exemple5
Valeur prédite	67.086264
Valeur réelle	66.000000

On remarque que la valeur réelle et la valeur prédite ne présentent pas de grande différence, on peut conclure que Notre modèle de prédiction est précis et fiable ce qui signifie que le modèle que nous avons utilisé est capable de capturer efficacement la relation entre les données d'entrée et les données de sortie pour ce scénario spécifique. Donc cela peut être considéré comme un signe de succès pour notre modèle de prédiction.

c) Evaluation du modèle

	Model
R2	98.868138
R2_adj	98.867565

Les coefficients de détermination (R^2) et ajusté (R^2 ajusté) évaluent la capacité du modèle à expliquer la variance de la variable cible (Performance Index). Les résultats, avec un R^2 de 98,87% et un R^2 ajusté de 98,87%, démontrent une excellente capacité du modèle à rendre compte de la variabilité de la cible.

d) Tests d'hypothèse sur chaque paramètre du modèle : Test de Student

$$H_0 : \beta_j = \beta_{j0} \text{ versus } H_1 : \beta_j \neq \beta_{j0} \quad , \quad j \in \{0, 1, 2, \dots, p\}$$

	|tvalue|	ttabl
const	265.875182	1.645008
Heures_détude	358.939822	1.645008
Scores_précédents	857.427191	1.645008
Activités_parascolaires	14.981236	1.645008
Heures_de_sommeil	39.622553	1.645008
Exercices_pratiqués_sur_des_questions_types	27.017306	1.645008

Avec ***|tcal| > ttbl*** pour tous les coefficients, *l'hypothèse nulle H0 est rejetée* au niveau de risque de 5%. Par conséquent, ***tous les coefficients sont statistiquement significatifs.***

e) Tests d'hypothèse global du modèle : Test de Fisher

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = \cdots = \beta_p = 0 \text{ versus } H_1 : \exists\, j \; tel \; que \; \beta_j \neq 0$$

	Model
Ftbl	2.010515
fcal	172376.499436

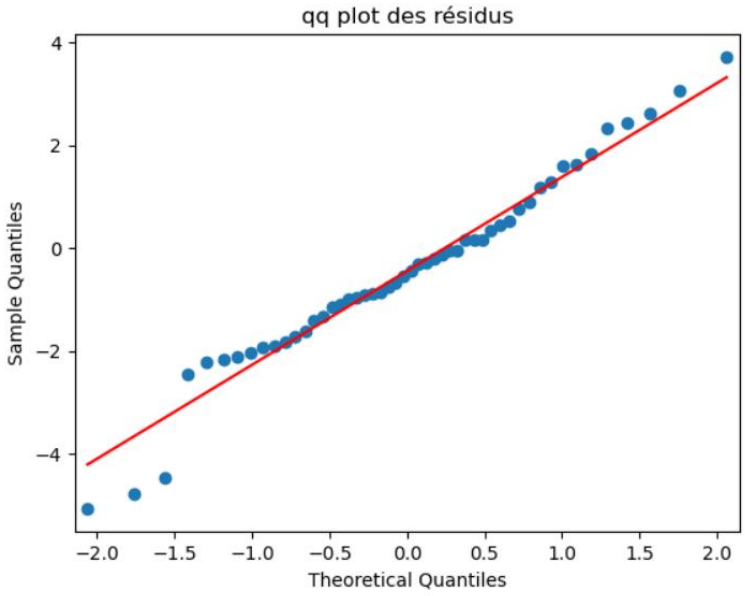
Avec un ***Fcal > Ftbl***, l'hypothèse nulle H0 est rejetée au niveau de risque de 5%. Cela confirme que ***le modèle est globalement significatif***, indiquant que l'ensemble des variables indépendantes contribue de manière significative à l'explication de la variance de la variable dépendante.

f) Intervalles de confiance des paramètres à 95 % de certitude

	min	Max
Intercept	-34.320093	-33.817737
Heures_détude Coef.	2.837149	2.868308
Scores_précédents Coef.	1.015991	1.020647
Activités_parascolaires Coef.	0.536003	0.697384
Heures_de_sommeil Coef.	0.456559	0.504084
Exercices_pratiqués_sur_des_questions_types Coef.	0.179841	0.207979

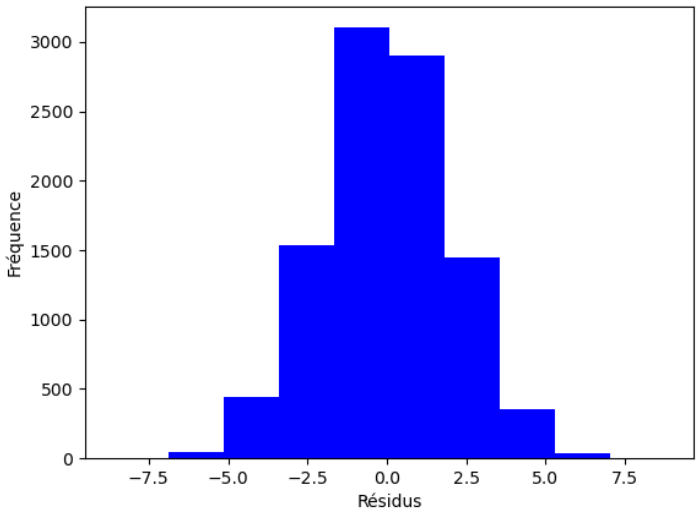
Les intervalles de confiance à 95 % pour les coefficients du modèle fournissent des indications sur la fiabilité des estimations. Si l'intervalle de confiance d'un coefficient n'inclut pas zéro, cela indique que le coefficient est significatif, contribuant de manière significative au modèle. D'après nos résultats, aucun des intervalles de confiance n'inclut zéro, ce qui confirme que tous les coefficients sont significatifs et ont une contribution importante au modèle.

g) Analyse des résidus



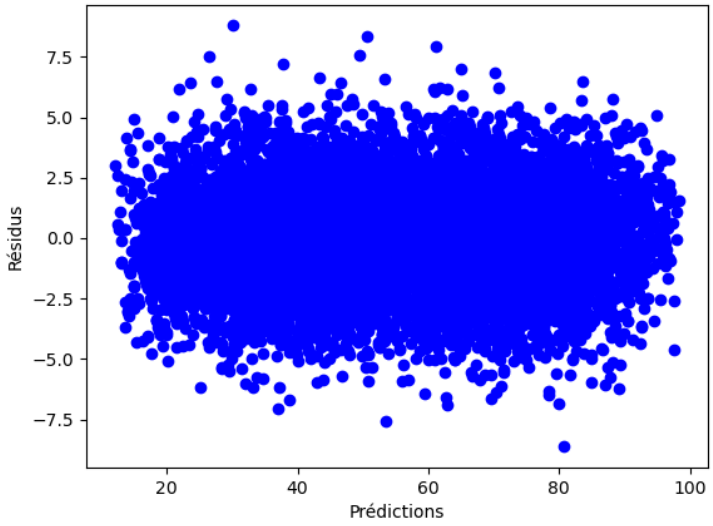
Le graphique montre un QQ plot, c'est un outil qui aide à voir si les erreurs d'un modèle statistique (les résidus) sont réparties normalement. Les points bleus représentent les erreurs du modèle et la ligne rouge montre comment ces erreurs se comporteraient si elles étaient parfaitement normales. Ici, les points suivent assez bien la ligne rouge, ce qui suggère que les erreurs de notre modèle sont proches de ce qu'on attend d'une distribution normale, un bon signe pour la validité du modèle.

Plot de la distribution des résidus



L'histogramme des résidus dépeint une distribution harmonieusement équilibrée autour de l'axe central, illustrant le profil classique d'une distribution normale. Cette symétrie ancrée autour du zéro, révèle un modèle statistique raffiné dont les prédictions sont non seulement précises en moyenne, mais également bien distribuées, signalant des erreurs minimes qui s'équilibrent avec une délicate précision.

Plot de la régression des résidus



Le Plot de régression des résidus affiche un nuage de points de résidus d'un modèle statistique, avec les prédictions sur l'axe horizontal et les résidus sur l'axe vertical. La concentration des points autour de la ligne zéro indique que les prédictions sont en moyenne exactes. L'absence de motifs distincts suggère qu'il n'y a pas de biais systématique dans les prédictions. La distribution uniforme des résidus autour de zéro témoigne d'une erreur constante, ce qui implique une bonne adéquation du modèle aux données observées.

h) Table ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
Heures_détude	1.0	5.131289e+05	5.131289e+05	122806.952473	0.000000e+00
Scores_précédents	1.0	3.077585e+06	3.077585e+06	736557.202994	0.000000e+00
Activités_parascolaires	1.0	8.666067e+02	8.666067e+02	207.404665	1.497719e-46
Heures_de_sommeil	1.0	6.605961e+03	6.605961e+03	1581.002225	8.013745e-321
Exercices_pratiqués_sur_des_questions_types	1.0	3.049914e+03	3.049914e+03	729.934822	3.721290e-155
Residual	9867.0	4.122766e+04	4.178337e+00	NaN	NaN

Le modèle ANOVA évalue l'importance de chaque variable dans le modèle. Une p-value inférieure à 0,05 indique que la variable est significative et apporte une contribution significative au modèle, tandis qu'une p-value supérieure à 0,05 suggère le contraire.

Les résultats de la table ANOVA démontrent que toutes les variables, notamment "Heures d'étude", "Scores précédents", "Activités parascolaires", "Heures de sommeil" et "Exercices pratiqués sur des questions types", présentent des p-values bien inférieures à 0,05. Ces valeurs significativement faibles confirment l'importance significative de toutes les variables dans le modèle. Ainsi, chacune de ces variables contribue de manière substantielle à l'explication de la variance dans la variable cible (Performance Index).

B. Modèle from scratch : avec L'Equation Normale

a) Définition du modèle

Dans le contexte de la régression linéaire multiple, les équations normales représentent une approche mathématique visant à déterminer explicitement les coefficients optimaux du modèle. Ces équations émergent en annulant le gradient de la fonction de coût par rapport aux coefficients. La fonction de coût, dans le contexte de la modélisation stochastique, mesure l'écart entre les valeurs prédites par le modèle ($\hat{y}$) et les valeurs réelles de la variable dépendante (y). Elle est formulée comme suit :

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Ainsi, les équations normales offrent une solution directe et non itérative pour ajuster les paramètres du modèle, offrant une compréhension claire du processus de minimisation de l'erreur dans la régression linéaire multiple.

Dans le cadre de l'équation normale pour la régression linéaire multiple, la recherche des paramètres ϑ qui minimisent cette fonction coût est formulée par l'expression :

$$\theta = (X^T X)^{-1} X^T y$$

où :

- θ représente le vecteur des paramètres du modèle,

- X est la matrice des caractéristiques,

- y est le vecteur des valeurs réelles de la variable dépendante,

- $(X^T X)^{-1}$ est l'inverse de la matrice produit $X^T X$.

Lorsque la matrice résultant du calcul de $X^T X$ est non inversible, l'équation normale est ajustée en utilisant une régularisation, définie comme suit : $\theta = (X^T X + \lambda I) X^T y$ où $\lambda > 0$ est un terme de régularisation, et I est la matrice identité.

Cette approche, souvent appelée régularisation de Tikhonov ou régularisation de Ridge, permet de stabiliser l'inversion de la matrice lorsque celle-ci n'est pas pleine rangée, en introduisant un terme de pénalité dans l'équation normale.

b) Construction du modèle

- On a créé un modèle from scratch nommé `LinearRegressionMultipleNormale`.

Ce modèle propose une implémentation personnalisée de la régression linéaire multiple basée sur l'équation normale.

Il comporte trois méthodes essentielles :

- *train* : pour l'entraînement et l'initialisation des coefficients.
- *predict* : pour les prédictions.
- *evaluate* : pour obtenir les valeurs R^2, R^2_{adj} et F_{calc} permettant d'évaluer la performance du modèle.

c) Evaluation du modèle

	Model_normale
R2	98.781204
R2_adj	98.780339

	Model_normale
Ftbl	2.010515
Fcalc	114220.011780

Notre modèle démontre une excellente capacité prédictive, expliquant 98.78% de la variance de la variable cible. De plus, la statistique F élevée confirme la significativité globale du modèle, renforcée par le rejet de l'hypothèse nulle à un niveau de risque de 5%.

C. Comparaison des modèles

	model	model_normale
R2	98.868138	98.781204
R2_adj	98.867565	98.780339

La comparaison entre le modèle de ***Statsmodel*** et le modèle *from scratch* basé sur ***l'équation normale*** montre des résultats très similaires. Les coefficients de détermination R^2 et R^2_{adj} sont presque identiques pour les deux modèles, indiquant que les performances prédictives sont équivalentes. Cela suggère une bonne concordance entre les implémentations, validant ainsi la précision du modèle from scratch par rapport à l'approche de Statsmodel.

D. Problème de colinéarité

	Heures_détude	Scores_précédents	Activités_parascolaires	Heures_de_sommeil	Exercices_pratiqués_sur_des_questions_types
Heures_détude	1.000000	-0.010676	0.004899	0.002131	0.015740
Scores_précédents	-0.010676	1.000000	0.009534	0.007975	0.008719
Activités_parascolaires	0.004899	0.009534	1.000000	-0.024008	0.013839
Heures_de_sommeil	0.002131	0.007975	-0.024008	1.000000	0.004907
ués_sur_des_questions_types	0.015740	0.008719	0.013839	0.004907	1.000000

La matrice de corrélation entre les variables suggère des relations peu significatives. Aucune corrélation marquée n'est observée entre les différentes paires de variables, indiquant une faible interdépendance. En résumé, les variables semblent être relativement indépendantes les unes des autres, soulignant une faible colinéarité.

VI. Conclusion

En résumé, la modélisation stochastique par régression linéaire multiple a été une méthode efficace pour prédire la performance des étudiants. En utilisant diverses variables telles que les heures d'étude, les scores précédents et d'autres, les modèles ont montré une forte capacité de prédiction, comme indiqué par les coefficients de détermination élevés (R^2 et R^2_{adj}).

L'analyse des résidus et les tests d'hypothèse ont confirmé la validité des modèles. Malgré quelques faibles corrélations entre certaines variables, la faible colinéarité suggère que chaque variable contribue distinctement à la prédiction.

La comparaison entre le modèle Statsmodels et l'implémentation from scratch des équations normales a montré une concordance étroite. En conclusion, cette approche offre une méthode fiable pour anticiper la performance académique des étudiants, avec des implications importantes pour les décideurs éducatifs.