Analyse Statistique de la Performance Scolaire avec le Langage R

FOOVI Komivi
Bineta Ndour

Université Numérique Cheikh Hamidou Kane
Département de Statistiques et Sciences des Données

Année académique 2024–2025

Encadrant : PhD Mamadou Korka DIALLO

Résumé

L'objectif de ce projet est d'identifier les facteurs qui influencent la performance académique des étudiants, à partir du jeu de données *Student Performance* comprenant 25 000 observations et 16 variables.
Les données ont d'abord été explorées afin d'analyser leur structure, détecter les valeurs manquantes et examiner la distribution des variables. Des analyses statistiques unidimensionnelles et bidimensionnelles ont ensuite été réalisées, avant de construire un modèle de régression linéaire multiple pour expliquer le score global des étudiants.
Les analyses montrent que les heures d'étude et le taux de présence ont un effet positif et significatif sur la performance académique, tandis que certaines variables sociodémographiques présentent un impact plus limité. Ces résultats soulignent l'importance de l'engagement scolaire dans la réussite des étudiants.

Introduction : Contexte et problématique

La performance scolaire constitue un enjeu majeur dans les systèmes éducatifs, car elle conditionne la réussite académique et professionnelle des apprenants.
L'analyse statistique des données éducatives permet de mieux comprendre les facteurs influençant les résultats scolaires et d'orienter les décisions pédagogiques et éducatives.
Dans ce projet, nous analysons les déterminants de la performance académique des élèves à partir de données individuelles portant sur leurs caractéristiques personnelles, familiales et scolaires.

Problématique

Quels sont les facteurs qui influencent significativement la performance globale des élèves et dans quelle mesure peut-on prédire cette performance à partir des variables disponibles ?

Objectifs du projet

Les objectifs principaux de ce projet sont :

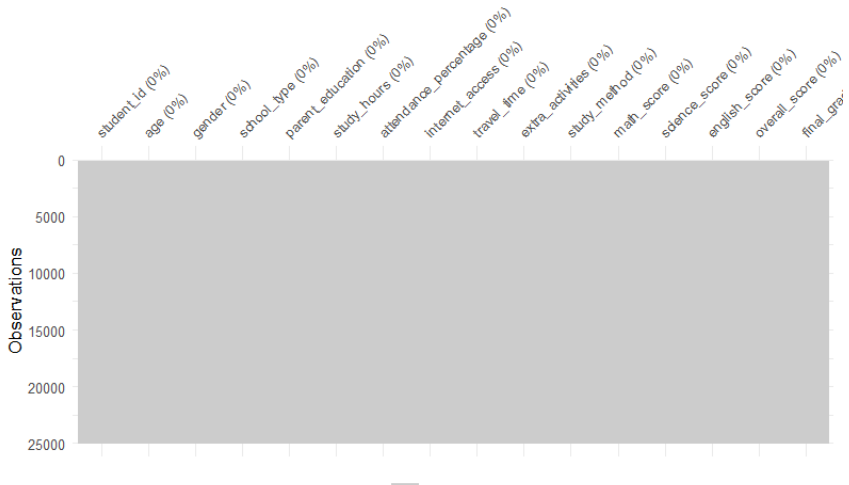
- Explorer et comprendre la structure du jeu de données.
- Décrire les variables à l'aide de statistiques unidimensionnelles.
- Étudier les relations entre les caractéristiques des élèves et leurs performances scolaires.
- Construire un modèle de régression linéaire multiple pour expliquer le score global des élèves.
- Interpréter les résultats statistiques de manière claire et rigoureuse.

Présentation du jeu de données

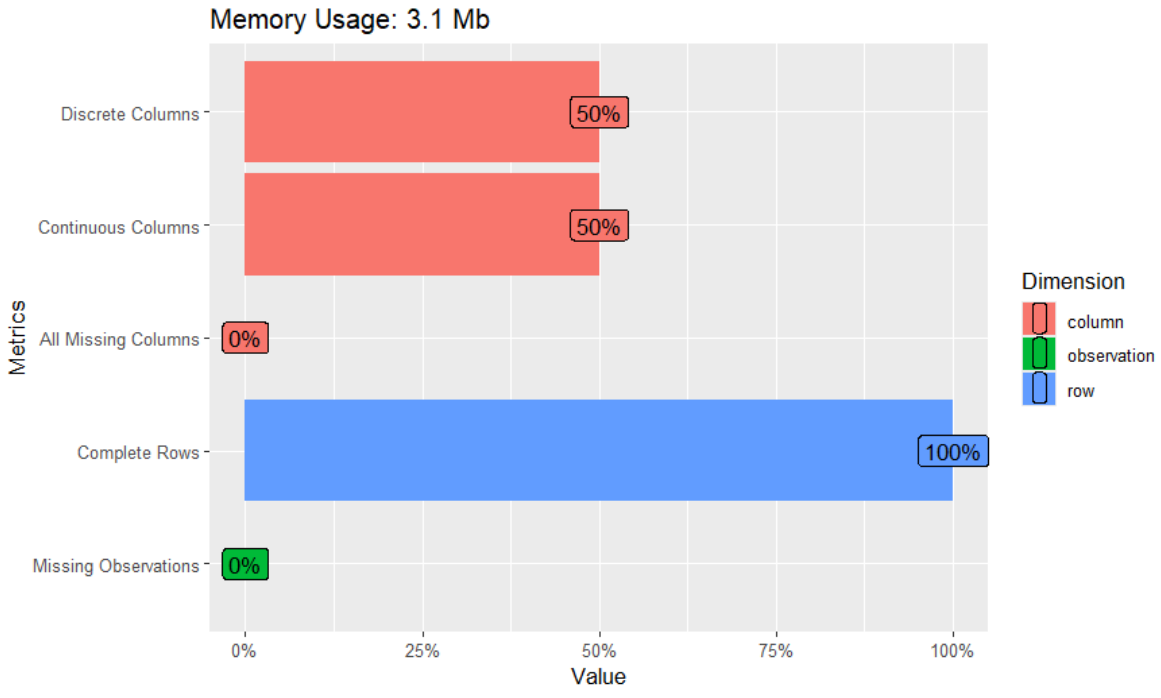
- **Nom du dataset :** Student Performance
- **Source :** Données éducatives à usage pédagogique, téléchargées sur la plateforme Kaggle
- **Nombre d'observations :** Environ 25 000 élèves
- **Nombre de variables :** 16
- **Variables quantitatives :**
  - âge, heures d'étude, taux de présence,score en mathématiques,score en sciences,score en anglais,score global
- **Variables qualitatives :**
  - genre,type d'école,niveau d'éducation des parents, accès à Internet, activités extra scolaires, méthode d'étude, note finale

Analyse exploratoire – Valeurs manquantes

- L'analyse exploratoire des données ne révèle aucune valeur manquante dans les variables étudiées.
- Toutes les observations sont complètes, ce qui garantit une bonne qualité des données.
- Aucun traitement spécifique (suppression ou imputation) n'est nécessaire avant les analyses statistiques et la modélisation.
- D'après les visualisations, il n'y a pas de valeurs manquantes.



Vue d’ensemble et qualité des données

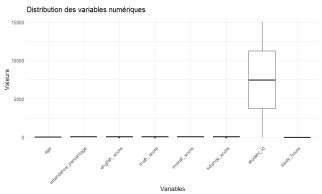


Interprétation :

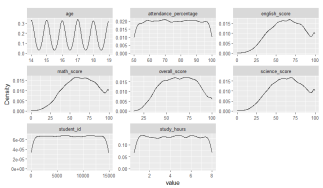
L’ensemble des observations est complet (100% de lignes complètes) et aucune valeur manquante n’est détectée. Cela indique une excellente qualité des données et élimine le besoin de techniques d’imputation.

Analyse descriptive des variables numériques

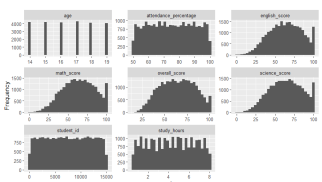
Distribution globale



Densités



Histogrammes



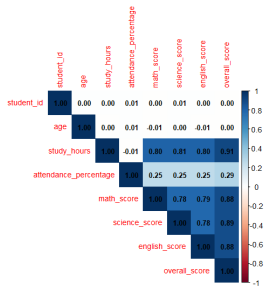
Interprétation globale

1) Boxplots Les distributions montrent une forte hétérogénéité d'échelle entre les variables. La variable *student_id* présente naturellement une dispersion élevée, tandis que les scores académiques sont concentrés dans des intervalles cohérents.

2) Densités Les scores scolaires présentent des formes proches d'une loi normale, légèrement asymétriques à droite. Cela suggère que la majorité des élèves obtiennent des scores moyens à bons, avec peu de valeurs extrêmes.

3) Histogrammes Les scores académiques se situent majoritairement entre 40 et 80. Le temps d'étude est réparti de manière variée, traduisant une diversité de comportements d'apprentissage.

Matrice de corrélation linéaire et conclusion



Interprétation :
Les performances en mathématiques, sciences et anglais sont fortement liées. Le score global résume fidèlement le niveau de l'élève. Les heures d'étude constituent le facteur explicatif le plus fort observé. La présence influence positivement mais modestement la réussite. L'âge n'a pas d'impact significatif.

Conclusion générale :

- Données complètes et de bonne qualité, Distributions cohérentes et interprétables,Peu de valeurs aberrantes critiques,Analyse exploratoire validée pour la modélisation

Analyse des mesures de tendance centrale

- **Âge des élèves :** moyenne = 16,48 ans, médiane = 16 ans. La légère différence suggère une distribution légèrement asymétrique à droite.
- **Temps d'étude quotidien :** moyenne = 4,25 h, médiane = 4,3 h. La proximité des valeurs indique une distribution symétrique.
- **Temps de transport :** la moyenne n'a pas pu être calculée en raison de valeurs non numériques ou manquantes. Un nettoyage des données est nécessaire.
- **Scores en mathématiques :** moyenne = 63,79, médiane = 64,1. Les scores sont distribués de manière équilibrée.
- **Scores en sciences :** moyenne = 63,75, médiane = 64,1. Distribution également équilibrée.

Analyse des mesures de dispersion

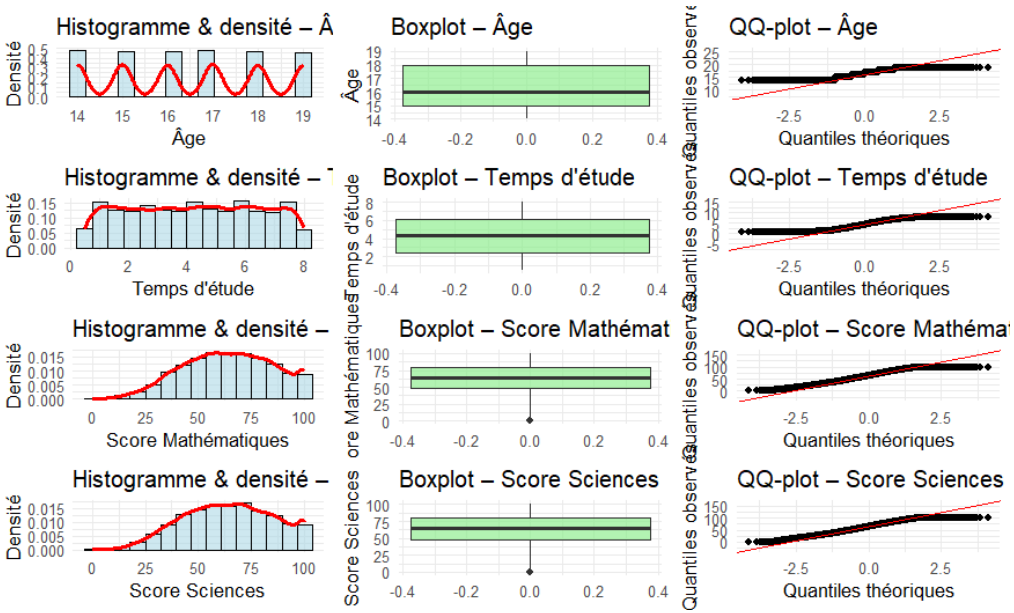
- **Âge des élèves :** écart-type = 1,70, variance = 2,90, étendue = 5, CV = 0,10. Les âges sont faiblement dispersés, avec 50% des élèves âgés de 15 à 18 ans, indiquant une population homogène.
- **Temps d'étude quotidien :** écart-type = 2,17, variance = 4,70, CV = 0,51. La dispersion est relativement élevée, montrant une forte variabilité des habitudes d'étude entre les élèves.
- **Scores en mathématiques :** écart-type = 20,88, variance = 435,78, CV = 0,33. Les performances sont très dispersées, avec 50% des élèves ayant des scores compris entre 48,3 et 80.
- **Scores en sciences :** écart-type = 20,97, variance = 439,76, CV = 0,33. Les résultats présentent une variabilité élevée, traduisant une hétérogénéité importante des niveaux en sciences.

Forme de la distribution

- **Asymétrie (skewness) :** âge = 0.01 ; temps d'étude = -0.007 ; score en mathématiques = -0.15 ; score en sciences = -0.14. Ces valeurs très proches de zéro indiquent des distributions globalement symétriques.
- **Aplatissement (kurtosis) :** âge = 1.74 ; temps d'étude = 1.80 ; score en mathématiques = 2.30 ; score en sciences = 2.29. Les coefficients étant inférieurs à 3, les distributions sont platicurtiques, avec une dispersion relativement étalée.
- **Test de normalité de Shapiro–Wilk :** ce test n'a pas pu être appliqué car la taille de l'échantillon dépasse la limite autorisée ($n > 5000$).
- L'analyse conjointe des valeurs de skewness, de kurtosis et des visualisations graphiques (histogrammes, boxplots et QQ-plots) permet de considérer les variables comme approximativement normales.

Visualisations des variables quantitatives

- **Interprétation générale :**
  - Les distributions des variables sont globalement symétriques (skewness proche de 0).
  - Les boxplots montrent une dispersion cohérente avec les mesures statistiques et quelques valeurs extrêmes isolées.
  - Les QQ-plots indiquent que les distributions peuvent être considérées comme approximativement normales, ce qui justifie l'utilisation de méthodes statistiques paramétriques.



Visualisations des variables qualitatives – Interprétation

- **Interprétation générale :**
  - Les distributions des variables qualitatives sont globalement équilibrées, avec des fréquences relatives proches entre les différentes modalités.
  - La variable *gender* présente une répartition homogène entre les catégories *female*, *male* et *other*, sans dominance marquée.
  - Le type d'école (*school_type*) montre une légère prédominance des établissements privés, tout en conservant une distribution quasi symétrique.
  - Le niveau d'éducation des parents (*parent_education*) affiche une répartition très homogène entre les modalités, indiquant une diversité équilibrée des profils éducatifs.
  - Cette structure des données limite les biais liés à la sur-représentation de certaines catégories et permet des comparaisons fiables dans les analyses ultérieures.

Visualisations des variables qualitatives – Graphiques

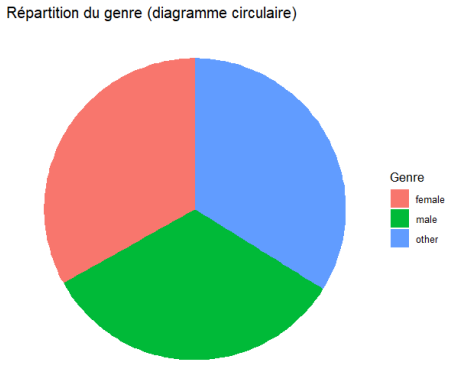


Figure – *

(a) Genre

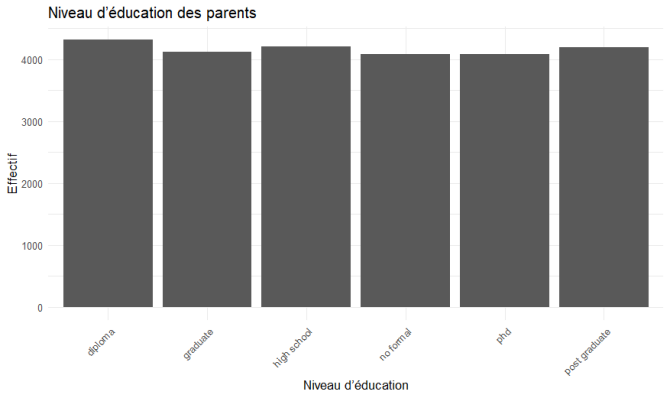


Figure – *

(b) Type d'école

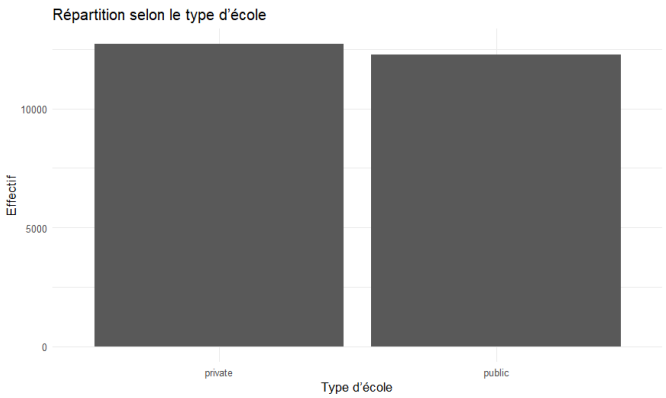


Figure – *

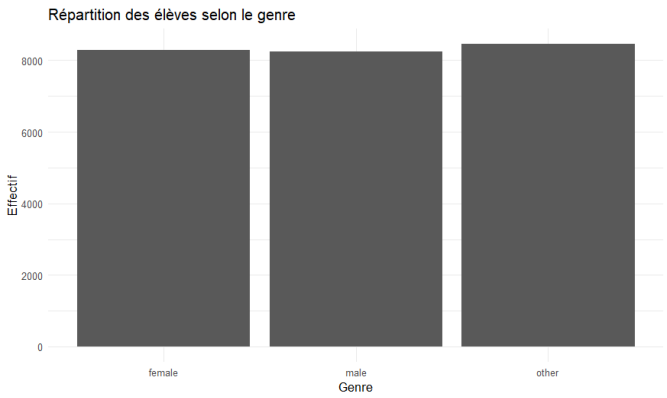


Figure – *

Analyse quantitative vs quantitative
Âge et score en mathématiques

Objectif : Évaluer la relation entre l'âge des élèves et leur performance en mathématiques.

Méthode :

- Coefficient de corrélation de Pearson
- Test de significativité (`cor.test`)

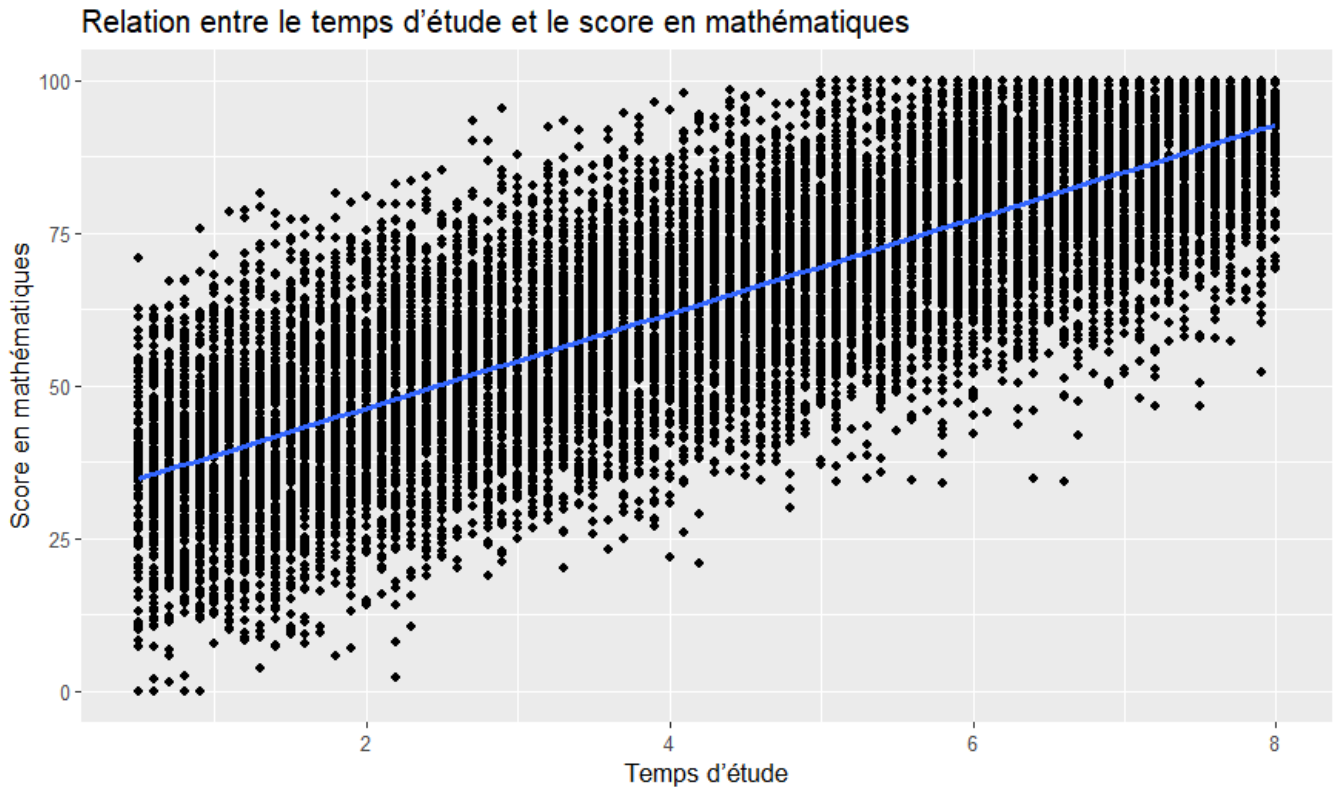
Résultats :

- Coefficient de corrélation : $r = -0{,}0052$
- Statistique de test : $t = -0{,}83$
- p-value : $p = 0{,}408 > 0{,}05$
- Intervalle de confiance à 95% :

$$IC_{95\%} = [-0{,}0176 \ ; \ 0{,}0072]$$

Visualisation : Âge et score en mathématiques

Nuage de points avec droite de régression



La visualisation confirme l'absence de relation linéaire entre l'âge et le score en mathématiques.

3.3.2 Analyse quantitative vs qualitative

Variables étudiées :

- Variable quantitative : Score en mathématiques
- Variable qualitative : Genre

Statistiques descriptives par groupe :

- Moyennes des scores très proches :
  - Femmes : 64,0
  - Hommes : 63,9
  - Autre : 63,4
- Médianes et quartiles similaires entre les groupes

Constat descriptif : Aucune différence marquée des scores n'est observée entre les modalités du genre.

Comparaison des scores selon le genre

Tests de comparaison :

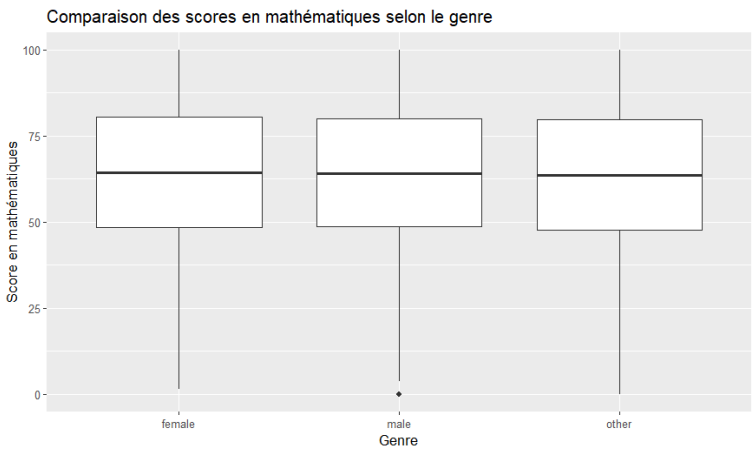
- ANOVA : $F = 1{,}83$, $p = 0{,}161$
- Kruskal–Wallis : $\chi^2 = 3{,}01$, $p = 0{,}223$

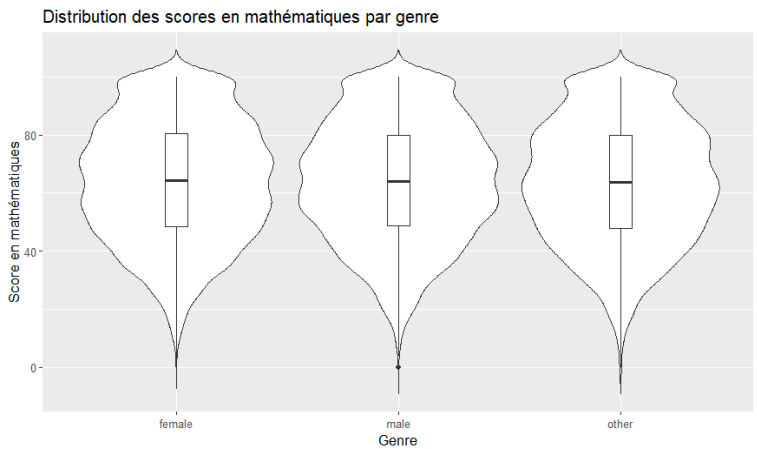
Interprétation : Les tests paramétrique et non paramétrique conduisent à la même conclusion : il n'existe pas de différence statistiquement significative des scores en mathématiques selon le genre.

Conclusion : Le genre n'a pas d'effet significatif sur la performance en mathématiques dans l'échantillon étudié.

Visualisations : Score en mathématiques selon le genre

Comparaison graphique des distributions





Lecture graphique :

- Distributions similaires entre les groupes
- Médianes proches et dispersions comparables
- Absence de différence marquée visuellement

Analyse des relations : Genre et Type d'école

1. **Genre et accès à Internet**
   - La majorité des élèves, quel que soit le genre, ont accès à Internet :
     - Female : 85% ont accès
     - Male : 85% ont accès
     - Other : 85% ont accès
   - Test du Chi-2 ($\chi^2 = 1.474$, $p = 0.4785$) : pas de relation significative.
   - Conclusion : L'accès à Internet est indépendant du genre.
2. **Type d'école et participation aux activités extrascolaires**
   - Les effectifs sont proches entre écoles privées et publiques pour la participation aux activités.
   - Test du Chi-2 ($\chi^2 = 3.2273$, $p = 0.072$) : pas de relation significative.
   - Résidus de Pearson faibles (< 1), confirmant l'absence d'écarts notables.

Méthode d'étude et note finale

- La répartition des notes varie selon la méthode d'étude :
  - Coaching et group study : légèrement plus de bonnes notes (a, b)
  - Notes et online videos : plus d'élèves avec des notes moyennes (c, d)
- Proportions par ligne montrent les différences de performances selon la méthode d'étude.
- Test du Chi-2 ($\chi^2 = 65.751$, $p < 0.001$) : relation significative.
- Conclusion : La méthode d'étude influence les résultats scolaires, certaines méthodes étant associées à de meilleures performances.

Visualisation : Heatmap ou barplots peuvent illustrer ces écarts.

Modèle de régression linéaire multiple (1/2)

Objectif : expliquer la performance académique globale des étudiants (`overall_score`) à partir de facteurs scolaires et socio-démographiques.

Variable dépendante :

- Score académique global (`overall_score`)

Variables explicatives :

- Heures d'étude hebdomadaires (`study_hours`)
- Taux de présence (`attendance_percentage`)
- Niveau d'éducation des parents (`parent_education`)
- Type d'école (`school_type`)

Modèle de régression linéaire multiple (2/2)

Forme du modèle :

$$\text{overall\_score} = \beta_0 + \beta_1\,\text{study\_hours} + \beta_2\,\text{attendance\_percentage} + \beta_3\,\text{parent\_education} + \beta_4\,\text{school\_type} + \varepsilon$$

Qualité globale du modèle :

- Coefficient de détermination : $R^2 = 0.9089$
- Coefficient ajusté : $R^2_{ajust} = 0.9089$
- Test F global : $p < 2.2 \times 10^{-16}$

Interprétation : le modèle explique environ **91 %** de la variabilité du score académique global.

Résultats, validation et sélection du modèle

Résultats principaux :

- Heures d'étude : effet fortement positif et significatif
- Taux de présence : effet positif et significatif
- Éducation des parents et type d'école : effets modérés

Erreurs de prédiction :

- RMSE = 5.71
- MAE = 4.93

Validation des hypothèses :

- Linéarité : résidus sans structure apparente
- Normalité : QQ-plot satisfaisant (test non applicable, $n > 5000$)
- Homoscédasticité : hétéroscédasticité détectée (Breusch-Pagan, $p < 0.05$)
- Multicolinéarité : aucune (VIF ≈ 1)