



**UNIVERSITÉ
DE REIMS
CHAMPAGNE-ARDENNE**

Master 1 Informatique

Rapport de Projet d'INFO0808 :

Visualisation des performances académiques des étudiants

Mars 2023

Réalisé par :

Kheireddine SATOURI

Youssef Anis DAHLOUK

Table des matières

1. Introduction	1
1.1. Objectifs du projet.....	1
1.2. Méthodologie utilisée	1
2. Description de la base de données	2
2.1. Origine et nature des données	2
2.1. Informations sur le jeu de données et liste des attributs	2
3. Structure de l'interface de visualisation et les techniques utilisées.....	3
3.1. Types de graphiques employés et justification de leur utilisation	3
3.2. Organisation des tableaux de bord.....	4
4. Analyse des graphes	5
4.1. Les Histogrammes	5
4.2. Les graphiques à barres	6
4.3. Les graphiques à barres empilées.....	8
4.4. La matrice de corrélation	10
4.5. Graphiques de nuage des points	10
4.6. Graphique en diagramme à points	11
4.7. Graphique en boîte.....	11
5. Discussion et conclusion	13
5.1. Discussion des résultats.....	13
5.2. Conclusion	14

1. Introduction

1.1. Objectifs du projet

L'objectif de ce projet est d'effectuer une analyse de données en utilisant des techniques de visualisation vues en cours. Pour ce faire, nous avons choisi une base de données pertinente et représentative portant sur le rendement scolaire des étudiants, sur laquelle nous avons appliqué différentes techniques de visualisation, en fonction de nos objectifs d'analyse et des caractéristiques des données. Notre objectif est de présenter ces données de manière claire et compréhensible, en utilisant des graphiques interactifs et des tableaux de bord qui permettent de synthétiser et d'explorer les résultats obtenus. Enfin, nous visons à proposer une expérience utilisateur optimale, en concevant une interface web de visualisation, qui offre des fonctionnalités adaptées aux besoins des utilisateurs. À travers ce projet, nous souhaitons démontrer l'importance et l'intérêt des techniques de visualisation de données pour l'analyse et la compréhension des données complexes, et leur potentiel pour aider les entreprises et les organisations à prendre des décisions éclairées.

1.2. Méthodologie utilisée

Dans le cadre de ce projet, nous avons utilisé le logiciel R pour effectuer l'analyse des données et créer les graphiques de visualisation. Nous avons commencé par importer les données dans R et effectué des analyses descriptives pour mieux comprendre les caractéristiques de l'ensemble de données. Cela a inclus des résumés statistiques, des analyses de données manquantes et d'autres techniques pour explorer les attributs de l'ensemble de données.

Ensuite, nous avons suivi une structure de développement basée sur « shinydashboard », un outil de construction d'applications web de type tableaux de bord, pour intégrer les séries de graphiques construits par ggplot2 dans des tableaux de bord. Cette approche nous a permis de créer une interface de visualisation interactive pour nos données, offrant une expérience utilisateur intuitive et efficace.

Nous avons également utilisé d'autres packages R tels que dplyr, tidyr et ggplot2 pour manipuler et visualiser les données, en fonction des besoins spécifiques de notre projet. Enfin, nous avons documenté notre code et nos analyses de manière systématique et claire, afin de faciliter la reproductibilité et la compréhension de notre travail par d'autres utilisateurs.

2. Description de la base de données

2.1. Origine et nature des données

<https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data?datasetId=436&language=R&outputs=Visualization>

Il s'agit d'un ensemble de données de rendement scolaire des étudiants importée de la plateforme Kaggle et est issue de l'université de Jordanie, à Amman. Les données sont collectées à partir du système de gestion de l'apprentissage appelé Kalboard 360, un LMS multi-agents conçu pour faciliter l'apprentissage grâce à l'utilisation d'une technologie de pointe. Les données sont collectées à l'aide d'un outil de suivi de l'activité de l'apprenant appelé Experience API (xAPI), qui permet de surveiller les progrès d'apprentissage et les actions de l'apprenant, telles que lire un article ou regarder une vidéo de formation.

L'ensemble de données se compose de 480 dossiers d'étudiants et de 16 caractéristiques, classées en trois grandes catégories : les caractéristiques démographiques, les caractéristiques des antécédents académiques et les caractéristiques comportementales. Les données ont été collectées sur deux semestres d'enseignement, comprenant 245 dossiers d'étudiants au cours du premier semestre et 235 dossiers d'étudiants au cours du second semestre.

Les étudiants sont originaires de différents pays, tels que le Koweït, la Jordanie, la Palestine, l'Irak, le Liban, la Tunisie, l'Arabie Saoudite, l'Égypte, la Syrie, les États-Unis, l'Iran, la Libye, le Maroc et le Venezuela. Enfin, la base de données comprend également des informations sur la participation des parents dans le processus éducatif, telles que le sondage de réponse des parents et la satisfaction des parents à l'école.

2.1. Informations sur le jeu de données et liste des attributs

Le jeu de données est caractérisé comme étant multivarié et contenant 480 instances. Il appartient au domaine de l'apprentissage en ligne et de l'éducation, avec une utilisation prédictive des modèles et de l'extraction de données éducatives. Les attributs sont à la fois entiers et catégoriques et le jeu de données comprend un total de 16 attributs.

Le jeu de données est associé à une tâche de classification et ne contient pas de valeurs manquantes. Les données sont stockées sous forme de fichier CSV avec le nom de fichier "xAPI-Edu-Data.csv".

Les attributs comprennent :

1. **Genre** : le genre de l'élève (nominal: 'Male' or 'Female')
2. **Nationality** : la nationalité de l'élève (nominal: ' Kuwait', ' Lebanon', ' Egypt', ' SaudiArabia', ' USA', ' Jordan', 'Venezuela', ' Iran', ' Tunis', ' Morocco', ' Syria', ' Palestine', ' Iraq', ' Lybia')
3. **Place of birth**: le lieu de naissance de l'élève (nominal: ' Kuwait', ' Lebanon', ' Egypt', ' SaudiArabia', ' USA', ' Jordan', 'Venezuela', ' Iran', ' Tunis', ' Morocco', ' Syria', ' Palestine', ' Iraq', ' Lybia')
4. **Educational Stages**: le niveau d'éducation de l'élève (Ordinal: 'lowerlevel', 'MiddleSchool', 'HighSchool')

5. **Grade Levels**: le niveau scolaire de l'élève (Ordinal: 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12')
6. **Section ID** : la classe à laquelle appartient l'élève (Ordinal: 'A', 'B', 'C')
7. **Topic** : le sujet du cours (nominal: 'English', 'Spanish', 'French', 'Arabic', 'IT', 'Math', 'Chemistry', 'Biology', 'Science', 'History', 'Quran', 'Geology')
8. **Semester** : le semestre de l'année scolaire (Ordinal: 'First', 'Second')
9. **Parent responsible for student** : le parent responsable de l'élève (nominal : « mom », « father »)
10. **Raised hand** : le nombre de fois où l'élève a levé la main en classe (numérique : 0-100).
11. **Visited resources** : le nombre de fois où l'élève a consulté le contenu du cours (numérique : 0-100).
12. **Viewing announcements** : le nombre de fois où l'élève a consulté les nouvelles annonces (numérique : 0-100).
13. **Discussion groups** : le nombre de fois où l'élève a participé à des groupes de discussion (numérique : 0-100).
14. **Parent Answering Survey** : les parents ont-ils répondu aux enquêtes fournies par l'école ou non (nominal : 'Yes', 'No').
15. **Parent School Satisfaction** : Satisfaction des parents vis-à-vis (nominal : 'Yes', 'No')
16. **Student Absence Days** : le nombre de jours d'absence pour chaque étudiant (Ordinal : above-7, under-7)

Les élèves sont classés en trois intervalles numériques en fonction de leur note/note totale :

Niveau bas « **L** » : l'intervalle comprend des valeurs de 0 à 69,

Niveau intermédiaire « **M** » : l'intervalle comprend des valeurs de 70 à 89,

Niveau élevé « **H** » : l'intervalle comprend des valeurs comprises entre 90 et 100.

3. Structure de l'interface de visualisation et les techniques utilisées

3.1. Types de graphiques employés et justification de leur utilisation

Dans cette étude, plusieurs types de graphiques ont été employés pour visualiser les données et en tirer des conclusions significatives. Les types de graphiques utilisés comprennent des histogrammes, des graphiques à barres, des graphiques à barres empilées, une matrice de corrélation des variables qualitatives, des graphiques de nuage de points, des graphiques en diagramme à points et des graphiques en boîte.

- **Les histogrammes** ont été utilisés pour représenter la distribution des variables continues telles que "Raised Hand", "Visited Resources", "Viewing Announcements" et "Discussion Groups". Ils montrent la fréquence de chaque valeur de la variable dans un intervalle spécifique et permettent de visualiser la forme de la distribution. Les histogrammes sont

particulièrement utiles pour détecter les valeurs aberrantes ou les valeurs extrêmes qui pourraient avoir un impact sur l'analyse.

- **Les graphiques à barres** ont été utilisés pour représenter les variables catégorielles telles que "Gender", "Nationality", "Place of Birth", "Educational Stages", "Grade Levels", "Section ID", "Topic", "Semester", "Parent responsible for student", "Parent Answering Survey" et "Parent School Satisfaction". Ils permettent de comparer les fréquences de chaque catégorie et de mettre en évidence les différences entre les groupes. Les graphiques à barres sont particulièrement utiles pour visualiser les différences entre les groupes de taille différente.
- **Les graphiques à barres empilées** ont été utilisés pour comparer les proportions des différentes catégories pour chaque variable. Par exemple, pour comparer la proportion de chaque nationalité pour chaque niveau d'éducation. Cela permet de mettre en évidence les différences dans la composition des groupes pour chaque variable.
- **La matrice de corrélation** des variables qualitatives a été utilisée pour visualiser les relations entre les différentes variables quantitatives. Elle permet de détecter les relations entre les variables et les comparées.
- **Les graphiques de nuage de points** ont été utilisés pour visualiser les relations entre deux variables continues telles que "Raised Hand" et "Visited Resources". Ils permettent de voir s'il y a une corrélation entre les deux variables et de détecter les valeurs aberrantes ou les valeurs extrêmes qui pourraient avoir un impact sur l'analyse.
- **Les graphiques en diagramme à points** ont été utilisés pour visualiser la distribution de chaque variable quantitative selon chaque niveau scolaire. Par exemple, les graphiques en diagramme à points ont été utilisés pour visualiser la distribution de "Raised Hand" pour chaque niveau d'éducation. Cela permet de voir les différences dans la distribution entre les groupes.
- **Les graphiques en boîte** ont été utilisés pour visualiser la distribution de chaque variable quantitative pour groupe et de détecter les valeurs aberrantes. Ils permettent de voir la dispersion des données, la médiane, le quartile inférieur et le quartile supérieur pour chaque groupe. Les graphiques en boîte sont particulièrement utiles pour détecter les différences entre les groupes et les valeurs aberrantes.

En conclusion, l'utilisation de plusieurs types de graphiques dans cette étude nous a permis de visualiser les données sous différents angles et de détecter les relations et les différences entre les variables et les groupes. Les histogrammes, les graphiques

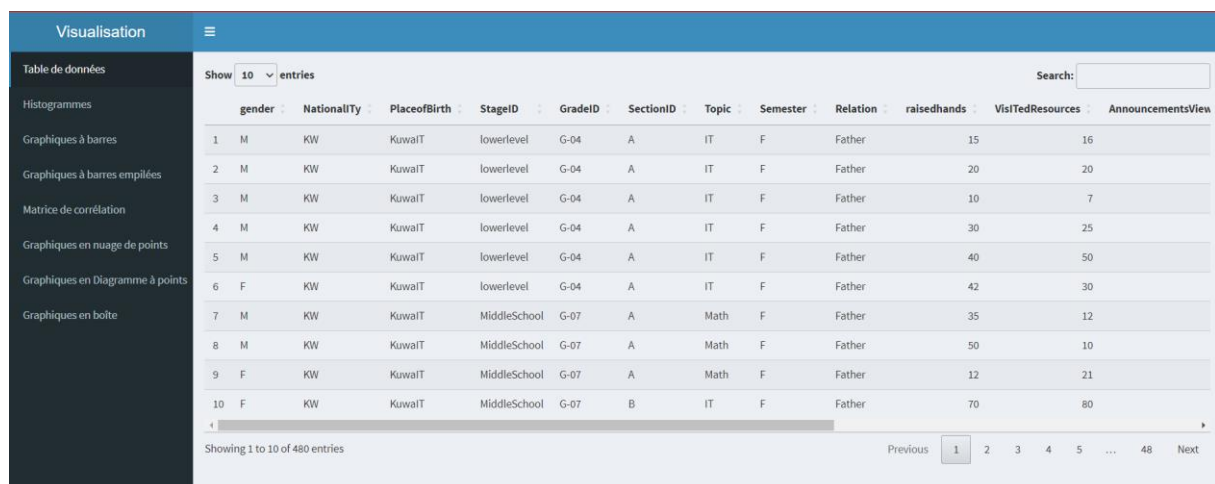
3.2. Organisation des tableaux de bord

La section d'organisation des tableaux de bord présente une approche pratique pour visualiser les données à l'aide de différents types de graphiques. Les différents types de graphiques mentionnés dans la partie 3.1 ont été présentés sous forme de tableaux de bord organisée dans une interface web « https://kheireddine-satouri.shinyapps.io/Projet_Visualisation_R/ », où chaque section dans le menu présente une série de graphes selon un type spécifique.

La première section dans le menu contient la table de données qui montre les informations de base du jeu de données. Cette table de données est utile pour explorer les différentes variables et les valeurs associées à chacune. Elle est également utile pour trier ou filtrer les données selon des critères spécifiques.

Les fonctions fournies par la librairie "shinydashboard" ont été employées pour la conception de l'interface web, notamment la fonction "dashboardPage" qui permet d'élaborer la structure de la page. Les différents paramètres de cette fonction permettent de définir les différentes parties de la page, telles que "dashboardSidebar" pour spécifier les sections de la barre de menu et "dashboardBody" qui fait appel aux différents graphiques selon chaque section. Les graphiques sont représentés sous forme d'objets et construits dans une fonction nommée "server" qui prend deux paramètres : une entrée qui représente le nom de l'objet et une sortie, c'est-à-dire le graphique spécifique à l'objet appelé. Les graphiques sont créés à l'aide de la librairie "ggplot".

Pour lancer l'application, il suffit d'exécuter la fonction "shinyApp" qui prend en paramètre la fonction "dashboardPage" et la fonction "server".



The screenshot shows a Shiny dashboard with a sidebar on the left and a main content area. The sidebar contains a 'Visualisation' menu with options like 'Table de données', 'Histogrammes', 'Graphiques à barres', etc. The main area displays a table of data with columns: gender, Nationality, Place of Birth, Stage ID, Grade ID, Section ID, Topic, Semester, Relation, raised hands, Visited Resources, and Announcements View. The table shows 10 entries, and a search bar is visible at the top right.

	gender	Nationality	Place of Birth	Stage ID	Grade ID	Section ID	Topic	Semester	Relation	raised hands	Visited Resources	Announcements View
1	M	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	15	16	
2	M	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	20	20	
3	M	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	10	7	
4	M	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	30	25	
5	M	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	40	50	
6	F	KW	Kuwait	lowerlevel	G-04	A	IT	F	Father	42	30	
7	M	KW	Kuwait	MiddleSchool	G-07	A	Math	F	Father	35	12	
8	M	KW	Kuwait	MiddleSchool	G-07	A	Math	F	Father	50	10	
9	F	KW	Kuwait	MiddleSchool	G-07	A	Math	F	Father	12	21	
10	F	KW	Kuwait	MiddleSchool	G-07	B	IT	F	Father	70	80	

Figure 1: Structure de l'interface de visualisation

4. Analyse des graphes

4.1. Les Histogrammes

Les histogrammes ont été utilisés pour représenter la distribution des 4 variables continues de la base de données notamment, "Raised Hand", "Visited Resources", "Viewing Announcements" et "Discussion Groups" pour visualiser la fréquence de levée de main, la fréquence de ressources visitées, la fréquence d'annonce vues et la fréquence de discussion initiées par intervalle.

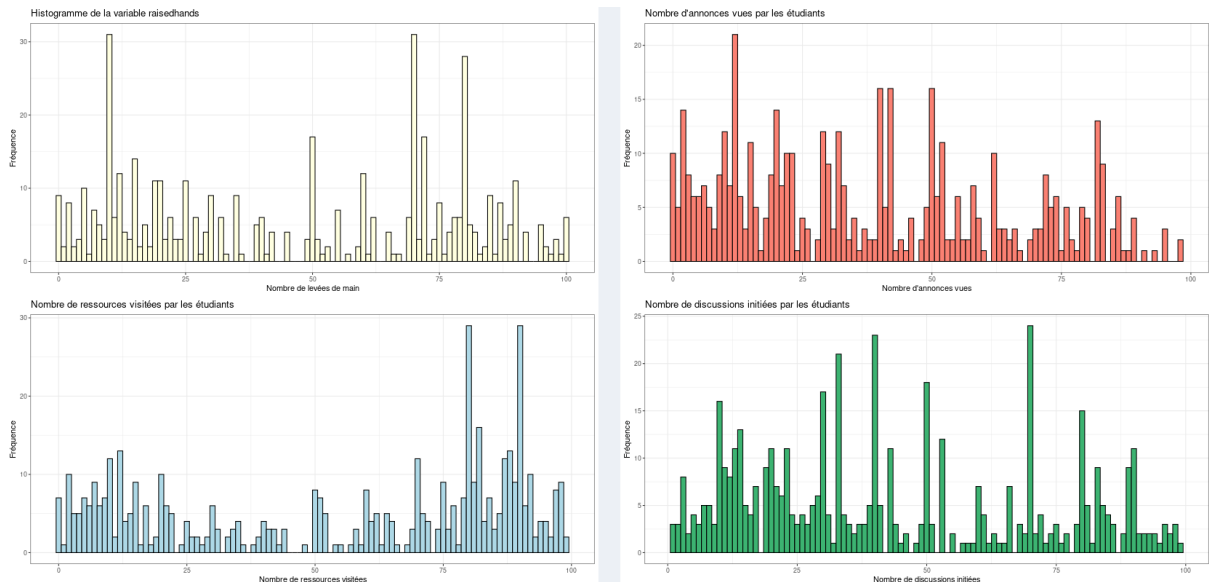


Figure 2: Visualisation de la distribution des variables quantitatives

Après avoir analysé les données, il n'a pas été constaté la présence de formes de distributions particulières selon les lois de distributions. Cependant, il est possible de noter que la distribution du nombre de ressources visitées est étalée vers la droite, tandis que la distribution du nombre de discussions initiées présente une forme relativement symétrique. Ce dernier constat suggère que la moyenne, la médiane et le mode de cette distribution sont relativement proches et situés autour du centre de la distribution, c'est-à-dire au niveau de 50. Il convient de souligner que ces résultats sont importants car ils permettent de mieux comprendre la distribution des données et d'adapter les analyses en conséquence.

4.2. Les graphiques à barres

Vues le nombre important des variables catégorielles, les graphiques à barres nous a permis de comparer la fréquence des étudiants selon des modalités spécifique, par exemple on peut remarquer que le niveau intermédiaire est le plus dominant dans la classification des étudiants :

Étant donné le grand nombre de variables catégorielles, l'utilisation de graphiques à barres nous a permis de comparer la fréquence des étudiants selon des modalités spécifiques. Par exemple, il est observé que le niveau intermédiaire est la classification la plus représentée chez les étudiants :

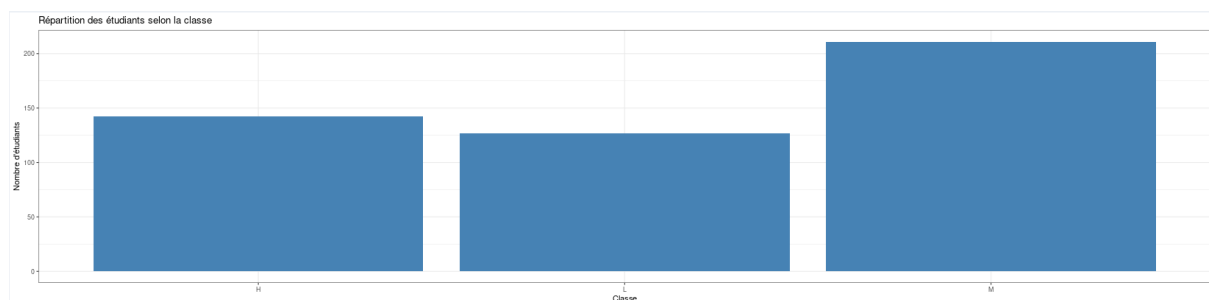


Figure 3: Représentation des étudiants selon la classe

On peut montrer aussi que le nombre de garçon est plus importants que les filles dans l'ensemble de données :

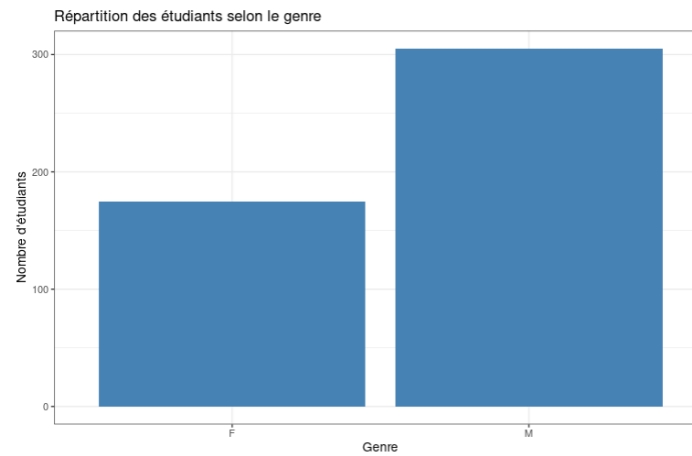


Figure 4: Répartition des étudiants selon le genre

On peut remarquer aussi que l'IT est le sujet le plus étudié :

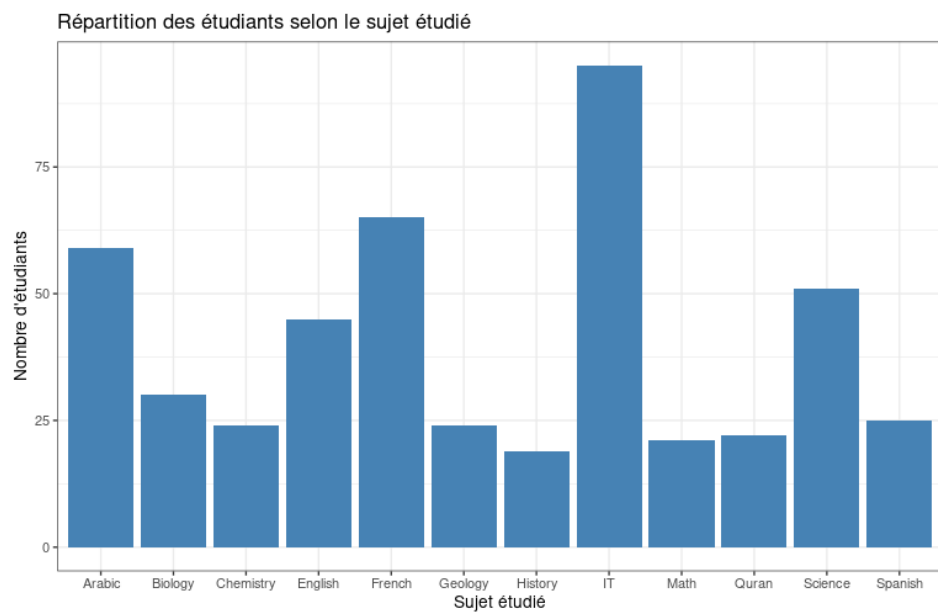


Figure 5: Répartition des étudiants selon le sujet étudié

Nous avons tenté d'inclure un maximum de graphiques à barres afin de mieux comprendre les facteurs qui influencent la performance des étudiants :

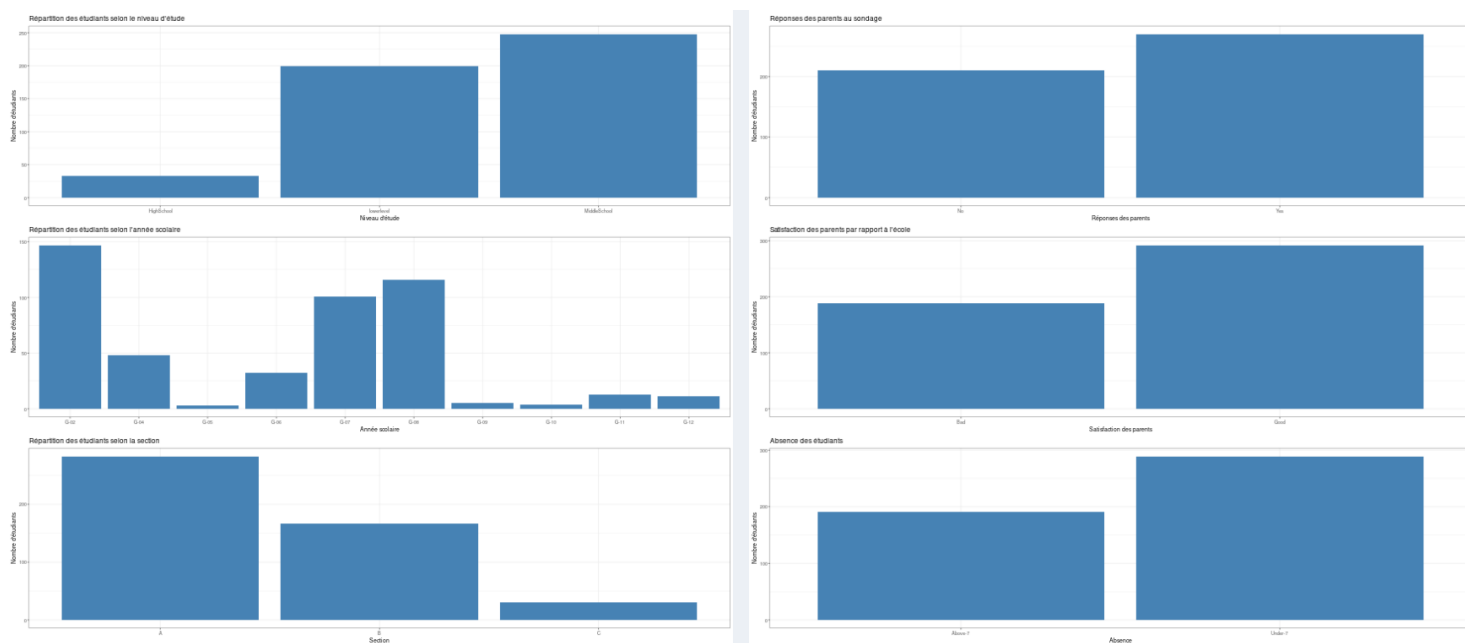


Figure 6: Autres Graphes à barres

4.3. Les graphiques à barres empilées

Pour aller plus loin dans la compréhension de l'influence des caractéristiques catégorielle sur le rendement des étudiants, les graphiques à barres empilées permet de comparer les proportions des différentes catégories pour chaque variable selon un critère supplémentaire .

Par exemple on peut remarquer que les filles sont mieux classées que les garçons :

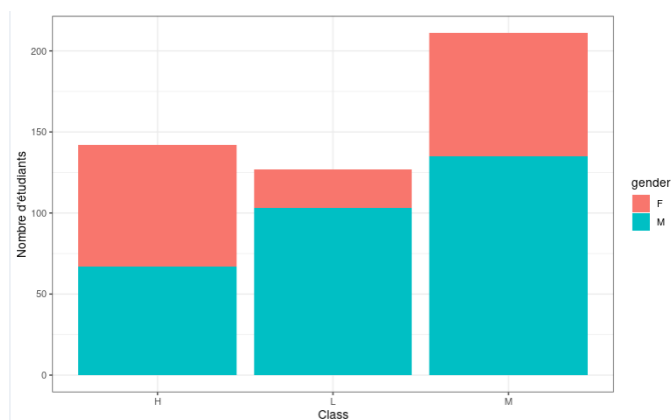


Figure 7: classification des étudiants par genre

Et que les mères sont les responsables des élèves les mieux classées pour la plupart entre eux :

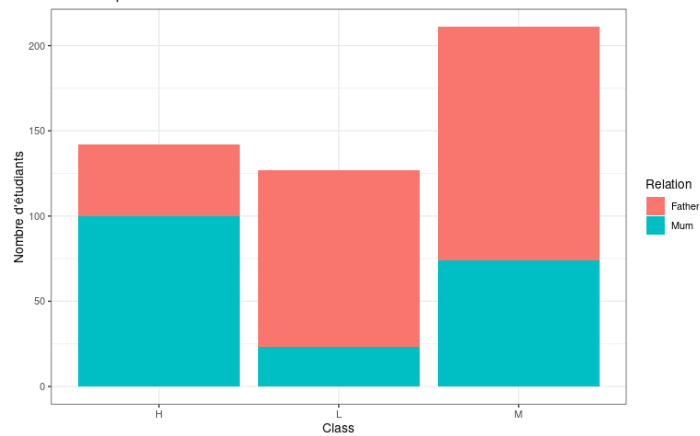


Figure 8: Influence parentale sur la classification finale des étudiants

On peut remarquer que plus l'étudiant a un taux d'absence élevé, moins il obtient un bon classement :

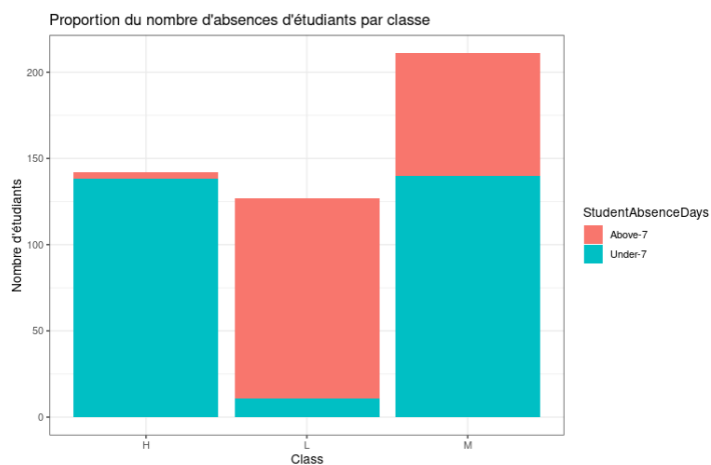


Figure 9: Proportion du nombre d'absence d'étudiant par classe

La figure ci-dessous montre le taux de classification des étudiants par sujet d'étude :

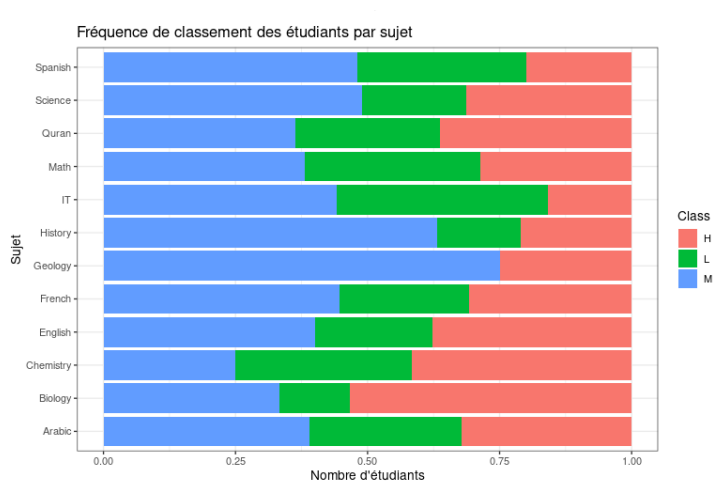


Figure 10: taux de classification des étudiants par sujet d'étude

4.4. La matrice de corrélation

La matrice de corrélation nous permet de vérifier et de comparer les relations entre les variables quantitatives de notre base de données. Étant donné qu'il y a quatre variables quantitatives, les corrélations entre ces variables sont représentées dans la matrice ci-dessous :

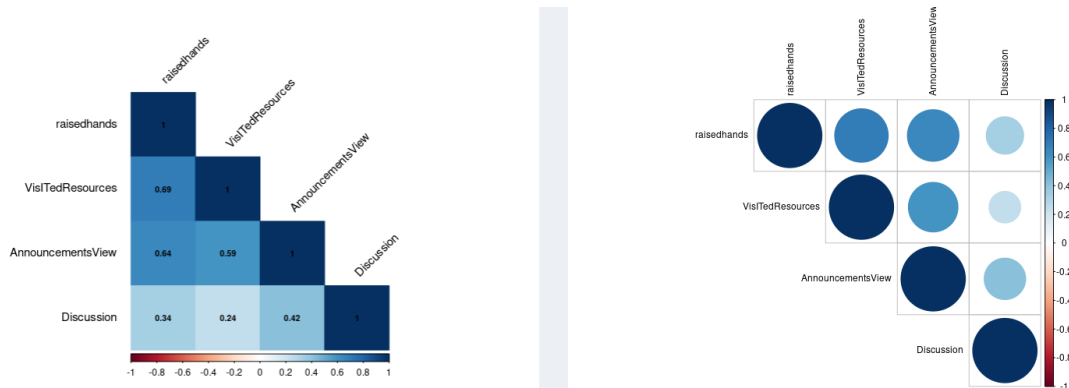


Figure 11: Matrice de corrélation

Il est remarqué que le taux de discussion entre les étudiants reste faible, même si ces derniers ont visité des ressources ou vérifié de nouvelles informations. Ainsi, on peut en déduire que les étudiants ont tendance à travailler individuellement et à ne pas partager leurs connaissances.

En revanche, la majorité des étudiants ayant visité des ressources ou consulté de nouvelles annonces ont levé la main pour participer.

4.5. Graphiques de nuage des points

Le graphique de nuage de points est un outil utile pour détecter la présence d'une corrélation entre deux variables quantitatives et identifier les éventuelles valeurs aberrantes ou extrêmes qui pourraient biaiser l'analyse. En effet, il permet de visualiser la relation entre les deux variables en traçant des points correspondant à chaque observation. Ainsi, si les points ont une forme linéaire ou suivent une tendance générale, on peut conclure à l'existence d'une corrélation. De plus, le graphique de nuage de points permet de mettre en évidence les points qui s'éloignent de cette tendance, pouvant correspondre à des valeurs extrêmes ou des observations atypiques qui pourraient influencer l'analyse.

En analysant les différentes relations entre les 4 variables quantitatives, il n'y a pas de linéarité remarquable entre certaines d'entre elles. Mais on peut remarquer la dispersion des points de la relation entre le nombre de levée de main et le nombre de ressource visitée :

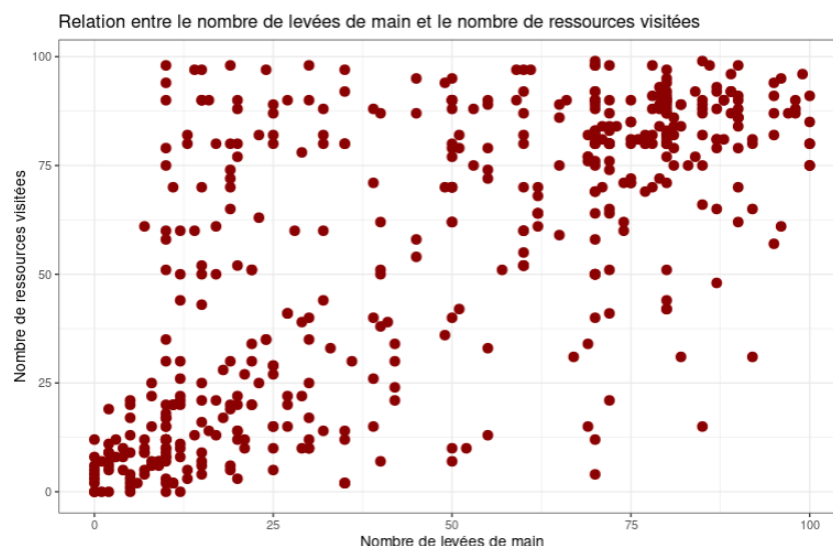


Figure 12: Relation entre le nombre de levée de main et le nombre de ressources visitées

4.6. Graphique en diagramme à points

Les graphiques en diagramme à points ont été utilisés pour visualiser la distribution de chaque variable quantitative selon chaque niveau scolaire. On peut remarquer que les étudiants de niveaux supérieurs ont généralement un nombre plus élevé de mains levées, de discussions, d'annonces vues et de recherches de ressources que les étudiants de niveaux inférieurs :

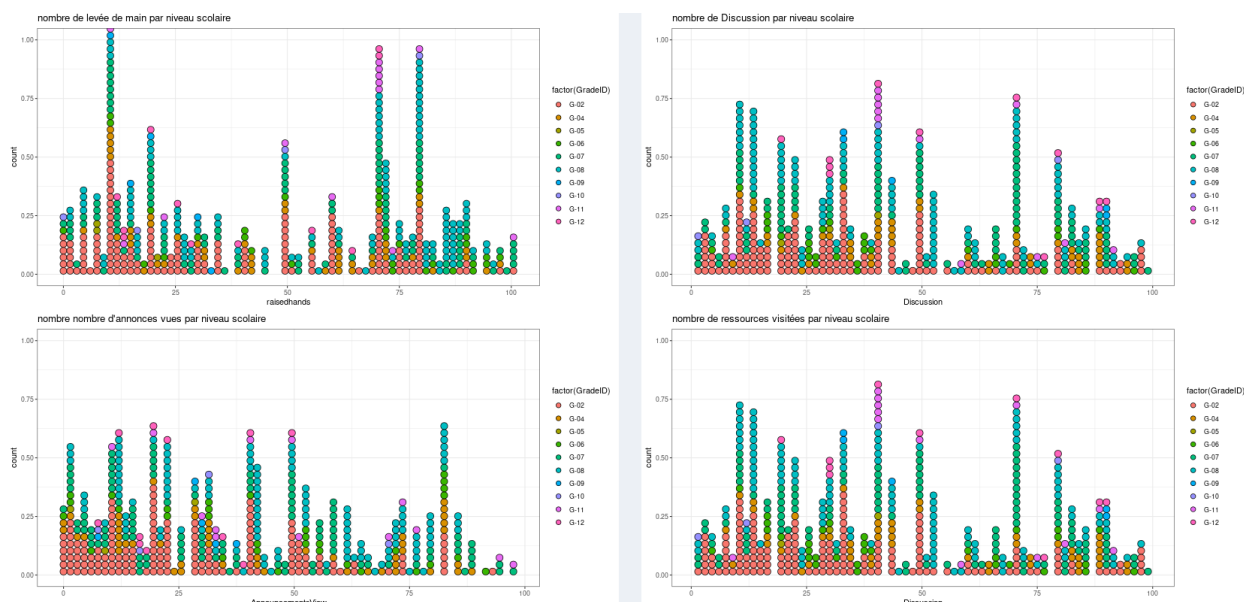


Figure 13: Fréquence des mains levées, de discussions, d'annonces vues et de recherches de ressources selon le niveau d'études

Cette forme de graphique est très facile à visualiser, surtout lorsqu'on souhaite analyser une variable quantitative en fonction d'une variable qualitative ordinaire ayant plusieurs modalités.

4.7. Graphique en boîte

Le dernier type de graphique généré est le graphique en boîte, il a été utilisés pour comparer la distribution de chaque variable quantitative selon le genre de l'étudiant,

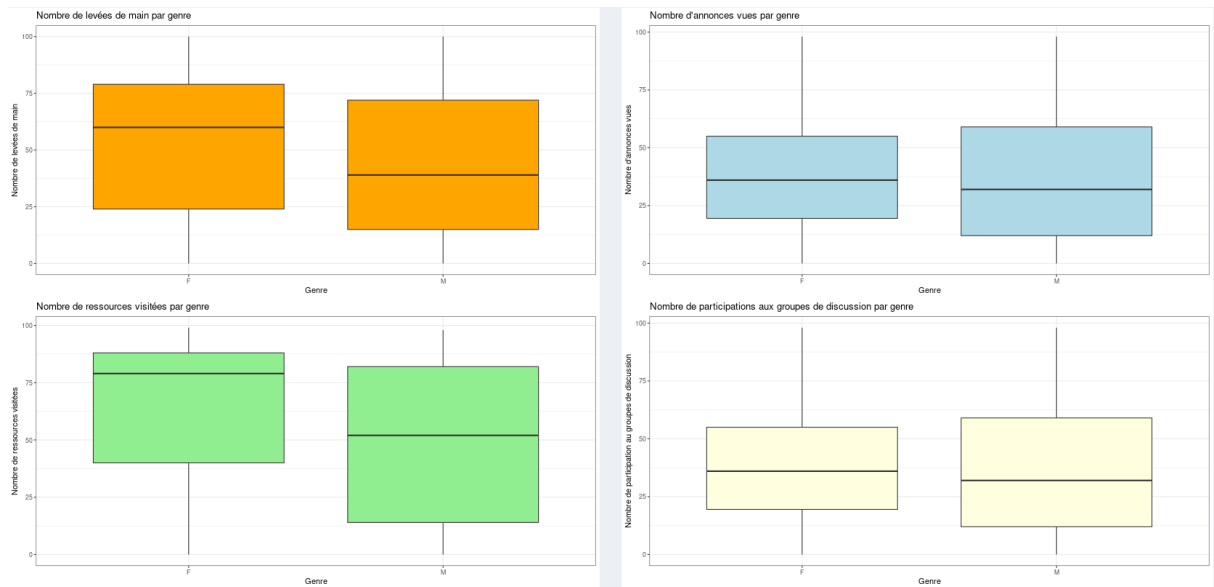


Figure 14: Distribution des variables quantitatives selon le genre

La figure ci-dessus montre que les filles participent davantage aux cours et visitent plus de ressources que les garçons. En ce qui concerne le nombre d'annonces vues et le nombre de discussions initiées, la distribution est approximativement identique entre les deux sexes.

5. Discussion et conclusion

5.1. Discussion des résultats

Les graphes ont permis une analyse visuelle des données collectées dans cette étude. Les différentes techniques de visualisation des données ont permis d'extraire plusieurs informations importantes concernant le sujet de performances académiques des étudiants.

Tout d'abord, en utilisant les histogrammes, nous avons eu une idée générale à propos la distribution des données quantitatives

En utilisant les graphes à barres, nous avons pu comparer la fréquence des étudiants selon les modalités spécifiques de certaines variables qualitatives telles que le niveau, le genre, la présence... Vue le nombre importants de ces variables, nous avons pu constater plusieurs choses tels que le niveau intermédiaire qui a été le plus représenté parmi les étudiants, que le nombre de garçons est plus important que les filles, que l'IT est le sujet le plus étudié, que la plus pats des étudiants sont dans la section A, et plein d'autres...

En utilisant les graphes à barres empilées, nous avons pu comparer les fréquences de différentes variables catégorielles, en mettant en évidence les différences entre les différents groupes. Cela a permis de faire une analyse approfondie de l'influence des caractéristiques catégorielle sur le rendement des étudiants, nous avons constaté par exemple que les filles sont mieux classées que les garçons, que les mères sont les responsables des élèves les mieux classées pour la plupart entre eux, et que plus l'étudiant a un taux d'absence élevé, moins il obtient un bon classement...

En utilisant la matrice de corrélation, nous avons pu visualiser les relations entre les différentes variables quantitatives. Les résultats ont montré qu'il n'y avait pas de corrélation linéaire forte entre les variables étudiées. Nous avons constaté une corrélation très faible avec la variable qui représente la participation en groupe de discussion, nous avons conclure ainsi que les étudiants préféraient travailler individuellement plutôt que de partager leurs connaissances

Le nuage de points nous a confirmé qu'il n'y a pas une linéarité forte entre les différents paires de valeurs quantitatives .

En utilisant les graphiques en diagramme à points, nous avons constaté que plus l'étudiant est dans un niveau supérieur, plus il est actif en participant au groupe de discussion, en levées des mains, en recherches de ressources et en vues d'annonces.

Enfin, les graphes en boîte ont montré que les filles participent davantage aux cours et visitent plus de ressources que les garçons.

5.2. Conclusion

En conclusion, l'utilisation des graphes est essentielle pour analyser les données de manière efficace et faciliter la compréhension des résultats. Cette étude a montré que les différentes techniques de visualisation des données, telles que les histogrammes, les graphes à barres, les graphes à barres empilées, la matrice de corrélation, le nuage de points, les graphiques en diagramme à points et les graphes en boîte, offrent une grande variété de perspectives sur les données, permettant ainsi de découvrir des tendances, des modèles et des relations qui seraient autrement difficiles à repérer. Grâce à l'utilisation de ces outils, nous avons pu extraire plusieurs informations importantes concernant les performances académiques des étudiants. En outre, la visualisation des données peut aider les décideurs à prendre des décisions éclairées basées sur les résultats obtenus. Ainsi, il est important de mettre en œuvre des techniques de visualisation des données dans toute analyse statistique pour faciliter l'interprétation et la communication des résultats.