Rapport d'Analyse Exploratoire des Données (EDA)

Prédiction de la Performance des Étudiants Marocains

Jeu de données	Morocco_Student_Data_Cleaned.csv
Nombre d'observations	10 000 étudiants
Nombre de variables	269 colonnes
Notebook source	EDA_Advanced.ipynb

8 février 2026

Ce rapport est généré automatiquement à partir de l'analyse EDA réalisée dans le notebook `EDA_Advanced.ipynb`.

Table des matières

1 Introduction et Vue d'ensemble

Ce rapport présente les résultats détaillés de l'analyse exploratoire des données (EDA) effectuée sur le jeu de données **Morocco_Student_Data_Cleaned.csv**. L'objectif principal est de comprendre la structure des données, détecter les anomalies, identifier les variables les plus pertinentes et préparer le terrain pour la phase de modélisation prédictive.

1.1 Caractéristiques du jeu de données

Table 1 – Vue d'ensemble du jeu de données

Caractéristique	Valeur
Nombre de lignes (observations)	10 000
Nombre de colonnes (variables)	269
Variables numériques (`float64`)	50
Variables numériques (`int64`)	48
Variables catégorielles (`object`)	171
Total variables numériques	**98**
Total variables catégorielles	**171**

Le jeu de données est **riche et multidimensionnel**, couvrant des informations académiques, sociodémographiques, comportementales, familiales et institutionnelles pour chaque étudiant.

2 Qualité des données

2.1 Valeurs manquantes

L'analyse révèle que **262 colonnes sur 269 sont complètes** (aucune valeur manquante). Seules **7 colonnes** présentent des données incomplètes :

Table 2 – Colonnes avec valeurs manquantes

Variable	Nb manquantes	Pourcentage (%)
`prise_notes`	10 000	100,00
`date_collecte`	10 000	100,00
`pays_cible`	8 344	83,44
`etablissement_precedent`	5 925	59,25
`niveau_allemand`	2 953	29,53
`type_travail`	2 239	22,39
`type_maladie`	2 204	22,04

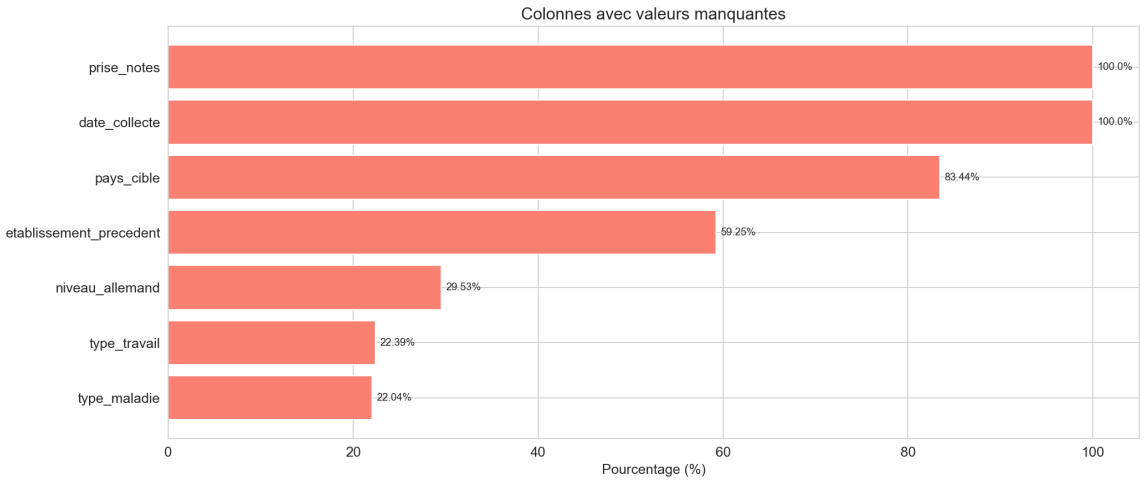

FIGURE 1 – Visualisation des colonnes avec valeurs manquantes

Interprétation :

— prise_notes et date_collecte sont **entièrement vides** (100%) et doivent être **supprimées**.

— `pays_cible` (83,44%) : concerne probablement les étudiants souhaitant poursuivre leurs études à l'étranger. Le taux élevé de données manquantes est logique car seule une minorité est concernée.

— `etablissement_precedent` (59,25%) : information historique non renseignée pour les étudiants sans transfert.

— `niveau_allemand` (29,53%) : langue optionnelle, les valeurs manquantes correspondent aux étudiants n'apprenant pas l'allemand.

— `type_travail` (22,39%) et `type_maladie` (22,04%) : non applicables pour les étudiants ne travaillant pas ou n'ayant pas de maladie chronique.

2.2 Doublons

La vérification des doublons a été effectuée selon trois critères :

TABLE 3 – Résultats de la vérification des doublons

Critère de vérification	Nombre de doublons
Lignes entièrement dupliquées	**0**
Doublons sur `id_etudiant`	**0**
Doublons sur `code_massar`	**0**

✓ **Aucun doublon détecté.** Le jeu de données contient bien 10 000 observations uniques.

3 Variables cibles

Le jeu de données contient trois variables cibles, chacune représentant un aspect différent de la performance de l'étudiant :

3.1 Description des variables cibles

TABLE 4 – Variables cibles du jeu de données

Variable	Type	Description
`moyenne_annuelle`	Numérique continue	Variable cible principale pour la régression. Représente la note moyenne annuelle de l'étudiant (échelle 0–20).
`performance_cible`	Numérique continue	Score dérivé de la moyenne annuelle (entre 0,08 et 0,99). À exclure des features car dérivée de la cible.
`probabilite_reussite`	Catégorielle	Classification en 4 niveaux de probabilité de réussite. Utilisable pour la classification.

3.2 Analyse de `performance_cible` (numérique)

— **Type :** `float64` – variable continue

— **Nombre de valeurs uniques :** 90 (entre 0,08 et 0,99)

— **Moyenne :** 0,41 **Médiane :** 0,41 **Écart-type :** 0,19

— **Distribution :** Légèrement asymétrique à droite (positive skewness)

— **Corrélation avec `moyenne_annuelle` :** $r = 0{,}936$ – corrélation très forte

3.3 Analyse de `probabilite_reussite` (catégorielle)

TABLE 5 – Distribution de la variable `probabilite_reussite`

Classe	Effectif	Pourcentage (%)
Élevé	6 508	65,1
Moyen	2 707	27,1
Faible	718	7,2
Très Faible	67	0,7

Constat important : Le jeu de données présente un **déséquilibre de classes significatif**. La classe *Élevé* représente 65,1% des observations tandis que la classe *Très Faible* ne compte que 67 individus (0,7%). Des techniques de rééquilibrage (SMOTE, sous-échantillonnage) seront nécessaires pour la classification.

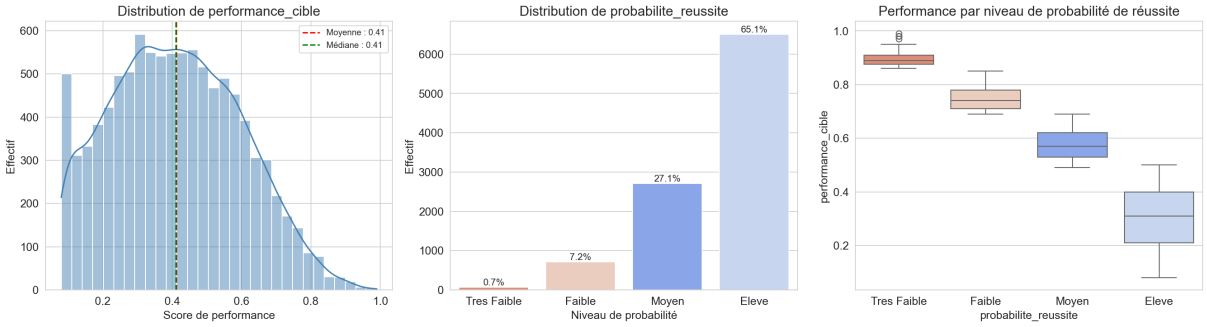

FIGURE 2 – Analyse de la variable cible : distribution de `performance_cible` (gauche), équilibre des classes de `probabilite_reussite` (centre), et relation entre les deux (droite)

4 Distribution des variables

4.1 Variables catégorielles principales

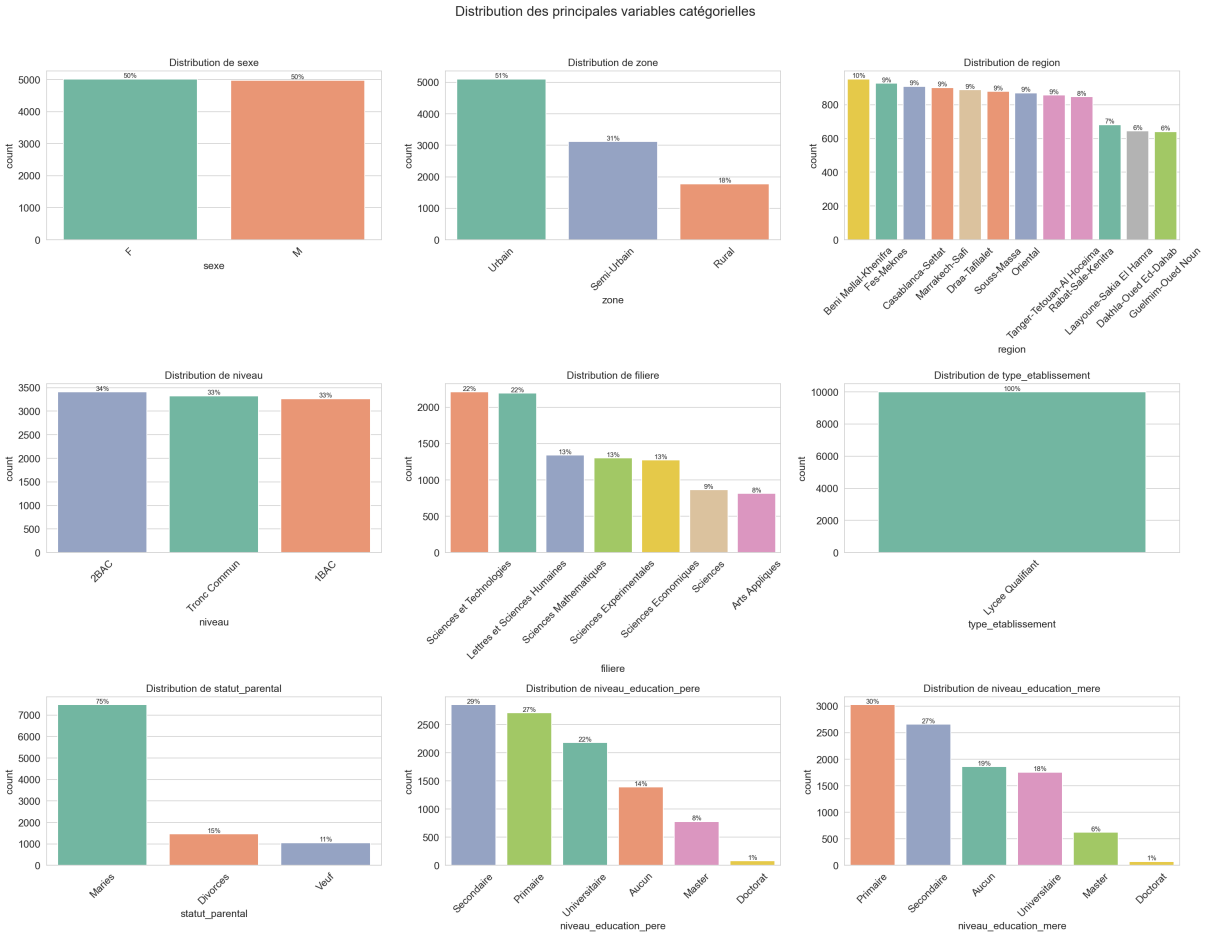

FIGURE 3 – Distribution des 9 principales variables catégorielles

Les observations clés sur les variables catégorielles :

— **Sexe :** Distribution quasi équilibrée – Féminin : 5 017 (50,2%), Masculin : 4 983 (49,8%).

— **Zone :** Prédominance urbaine – Urbain : 5 072 (50,7%), Semi-Urbain : 3 133 (31,3%), Rural : 1 795 (18,0%).

— **Niveau scolaire :** Répartition assez homogène – 2BAC : 3 389 (33,9%), Tronc Commun : 3 283 (32,8%), 1BAC : 3 328 (33,3%).

— **Région :** Couverture de 12 régions du Maroc, avec Casablanca-Settat comme la plus représentée.

4.2 Variable cible principale : `moyenne_annuelle`

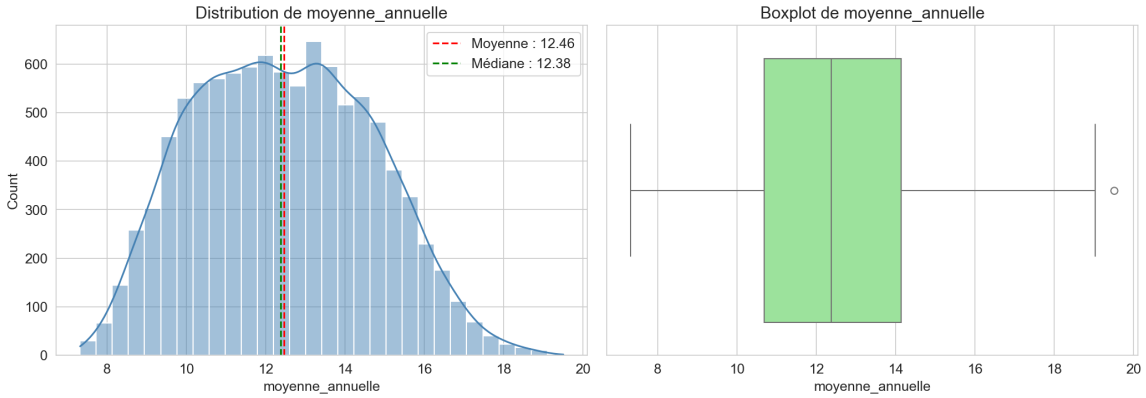

FIGURE 4 – Distribution et boxplot de la variable `moyenne_annuelle`

TABLE 6 – Statistiques descriptives de `moyenne_annuelle`

Statistique	Valeur
Moyenne	12,46
Médiane	12,42
Écart-type	2,24
Minimum	7,32
1er quartile (Q1)	10,73
3e quartile (Q3)	14,13
Maximum	19,51

La distribution est **approximativement symétrique** (skewness $\approx 0{,}13$), légèrement étalée vers la droite. La moyenne et la médiane sont très proches, confirmant une distribution quasi-normale. Cela est favorable pour les modèles de régression linéaire.

5 Analyse de l'asymétrie (Skewness)

L'asymétrie mesure le degré de déformation de la courbe de distribution par rapport à la loi normale. Une variable est considérée fortement asymétrique lorsque $|\text{skewness}| > 1$.

5.1 Variables fortement asymétriques

L'analyse identifie **8 variables** avec une asymétrie prononcée :

TABLE 7 – Variables avec |skewness| > 1

Variable	Asymétrie (Skewness)	Aplatissement (Kurtosis)
`montant_tayssir`	1,577	0,546
`annees_redoublement`	1,484	0,202
`revenu_mensuel_mere`	1,446	3,536
`heures_travail_semaine`	1,373	0,636
`absences_non_justifiees`	1,327	1,771
`revenu_mensuel_pere`	1,306	2,611
`cout_mensuel_soutien`	1,116	0,280
`revenu_familial`	1,113	2,205

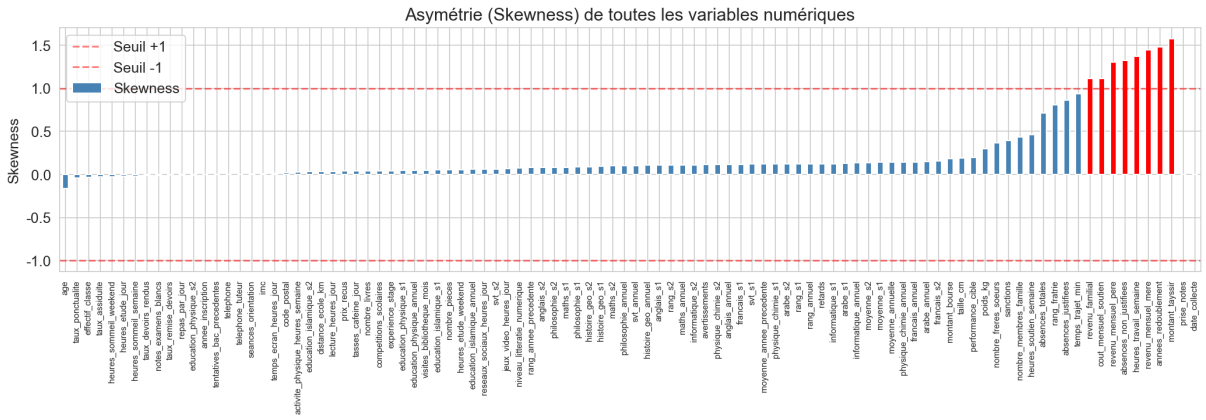

FIGURE 5 – Asymétrie de toutes les variables numériques. Les barres rouges indiquent les variables dépassant le seuil |skewness| > 1.

Interprétation :

— Toutes les 8 variables ont une **asymétrie positive** (queue étalée vers la droite), ce qui signifie qu'il y a une concentration de valeurs faibles avec quelques valeurs élevées exceptionnelles.

— Les variables liées aux **revenus** (`revenu_familial`, `revenu_mensuel_pere`, `revenu_mensuel_mere`) sont naturellement asymétriques – c'est un phénomène classique en sciences sociales.

— `revenu_mensuel_mere` présente le **kurtosis le plus élevé** (3,536), indiquant des queues de distribution lourdes (valeurs extrêmes).

— **Recommandation :** Envisager une transformation logarithmique ($\log(x+1)$) ou de Box-Cox pour ces variables avant la modélisation.

6 Détection des valeurs aberrantes (Outliers)

La détection des outliers a été réalisée avec la méthode de l'intervalle interquartile (IQR). Un point est considéré comme aberrant si $x < Q_1 - 1{,}5 \times IQR$ ou $x > Q_3 + 1{,}5 \times IQR$.

TABLE 8 – Résumé des valeurs aberrantes par variable

Variable	Nb outliers	%	Borne inf.	Borne sup.
revenu_familial	214	2,14	−6 019,75	32 880,25
moyenne_s1	4	0,04	5,55	19,28
moyenne_annuelle	1	0,01	5,51	19,32
moyenne_s2	1	0,01	5,49	19,33
age	0	0,00	14,00	22,00
absences_totales	0	0,00	−18,50	33,50
heures_etude_jour	0	0,00	−1,80	7,80
distance_ecole_km	0	0,00	−8,05	25,95
imc	0	0,00	11,75	34,55

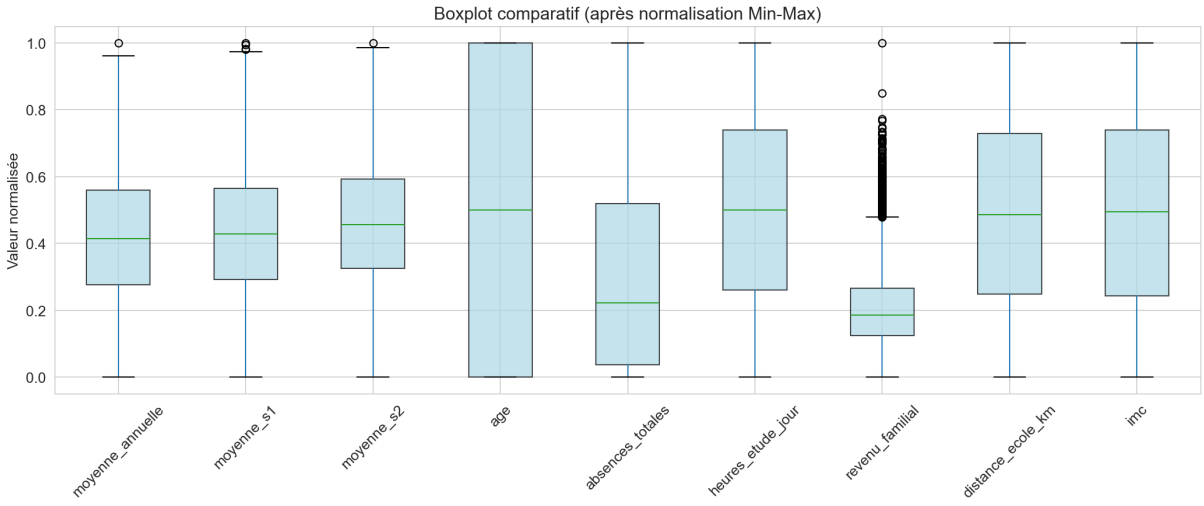

FIGURE 6 – Boxplot comparatif des variables clés après normalisation Min-Max

Constat principal :

— `revenu_familial` est la **seule variable significativement affectée** par les outliers avec 214 observations (2,14%) au-delà de la borne supérieure de 32 880,25 DH.

— Les variables académiques (`moyenne_s1`, `moyenne_annuelle`, `moyenne_s2`) ont un nombre **négligeable** d'outliers ($< 0{,}05\%$).

— Les variables `age`, `absences_totales`, `heures_etude_jour`, `distance_ecole_km` et `imc` ne présentent **aucun outlier** selon la méthode IQR.

— **Recommandation :** Traiter les outliers de `revenu_familial` par *winsorisation* ou transformation logarithmique.

7 Analyse des corrélations

7.1 Corrélations avec la variable cible `moyenne_annuelle`

Table 9 – Top 15 des corrélations positives avec `moyenne_annuelle`

Variable	Coefficient de corrélation (r)
`moyenne_s2`	0,991
`moyenne_s1`	0,991
`moyenne_annee_precedente`	0,958
`performance_cible`	0,936
`francais_annuel`	0,918
`education_physique_annuel`	0,918
`physique_chimie_annuel`	0,918
`maths_annuel`	0,917
`svt_annuel`	0,917
`anglais_annuel`	0,915
`education_islamique_annuel`	0,914
`informatique_annuel`	0,913
`arabe_annuel`	0,913
`histoire_geo_annuel`	0,911
`philosophie_annuel`	0,910

Table 10 – Corrélations négatives notables avec `moyenne_annuelle`

Variable	Coefficient de corrélation (r)
`nombre_membres_famille`	−0,020
`poids_kg`	−0,025
`age`	−0,039
`montant_tayssir`	−0,146
`rang_annee_precedente`	−0,771
`rang_s2`	−0,823
`rang_s1`	−0,847
`rang_annuel`	−0,847

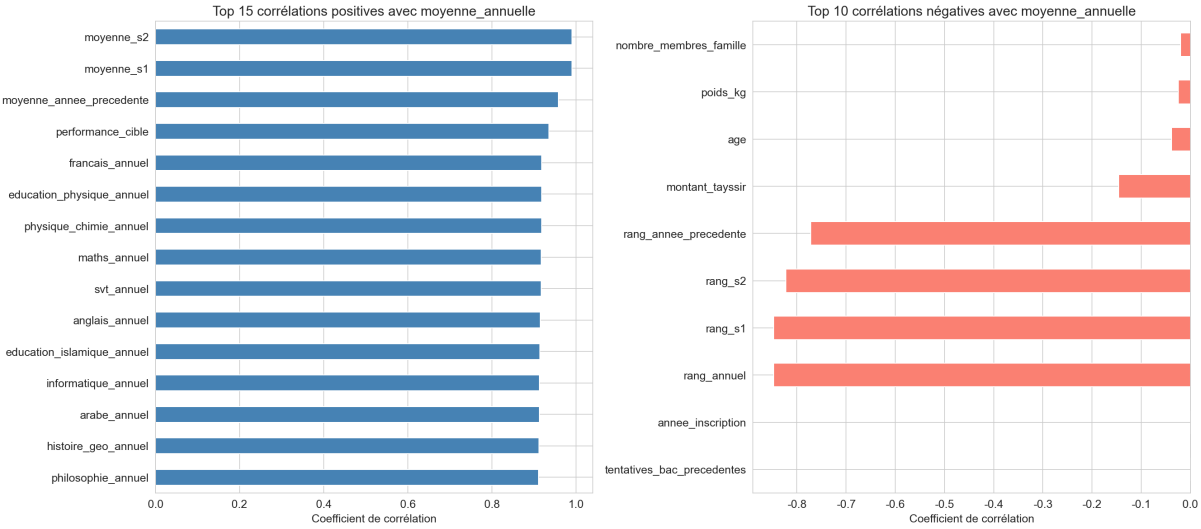

FIGURE 7 – Top 15 des corrélations positives (gauche) et top 10 des corrélations négatives (droite) avec `moyenne_annuelle`

7.2 Heatmap de corrélation

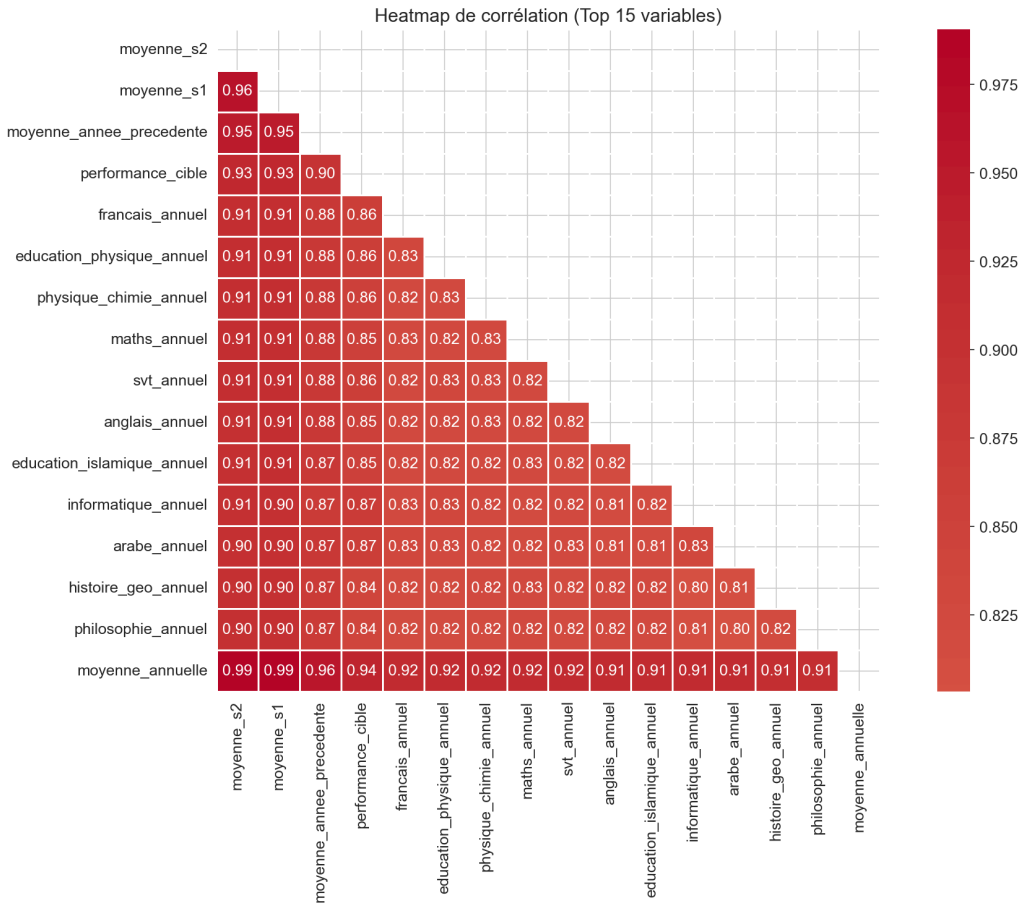

FIGURE 8 – Matrice de corrélation (heatmap) des 15 variables les plus corrélées avec `moyenne_annuelle`

7.3 Interprétation des corrélations

7.3.1 Variables fortement corrélées positivement ($r > 0{,}9$)

1. **moyennes semestrielles** ($r = 0{,}991$) : `moyenne_s1` et `moyenne_s2` sont très fortement corrélées car la `moyenne_annuelle` est leur moyenne arithmétique. **Risque de fuite de données (data leakage).**
2. **moyenne année précédente** ($r = 0{,}958$) : forte continuité dans la performance académique d'une année à l'autre. **Variable dérivée – à évaluer soigneusement.**
3. **performance_cible** ($r = 0{,}936$) : variable directement **dérivée** de `moyenne_annuelle`. **Doit être exclue des features.**
4. **Notes annuelles par matière** ($r \approx 0{,}91$–$0{,}92$) : les notes de français, éducation physique, physique-chimie, maths, SVT, anglais, etc., sont les **composantes directes** de la moyenne. **Variable dérivée – à exclure pour éviter la fuite.**

7.3.2 Variables fortement corrélées négativement ($r < -0{,}7$)

Les variables de **classement** (`rang_annuel`, `rang_s1`, `rang_s2`) montrent une forte corrélation négative avec la moyenne : plus la moyenne est élevée, meilleur est le rang (plus petit en valeur). Ces variables sont **redondantes** car elles sont une transformation directe de la moyenne.

7.3.3 Variables faiblement corrélées

Les variables `nombre_membres_famille` ($r = -0{,}020$), `poids_kg` ($r = -0{,}025$) et `age` ($r = -0{,}039$) n'ont **pratiquement aucune corrélation linéaire** avec la performance académique when considérées isolément.

△ **Alerte – Risque de fuite de données (Data Leakage) :**

Les variables suivantes sont **dérivées** de `moyenne_annuelle` et ne doivent **pas** être utilisées comme features dans le modèle prédictif :

- `moyenne_s1`, `moyenne_s2` (composantes de la moyenne)
- `performance_cible` (transformation directe)
- `rang_annuel`, `rang_s1`, `rang_s2` (classement basé sur la moyenne)
- Toutes les notes annuelles par matière (`*_annuel`)

8 Analyse bivariée

8.1 Variables numériques vs. `moyenne_annuelle`

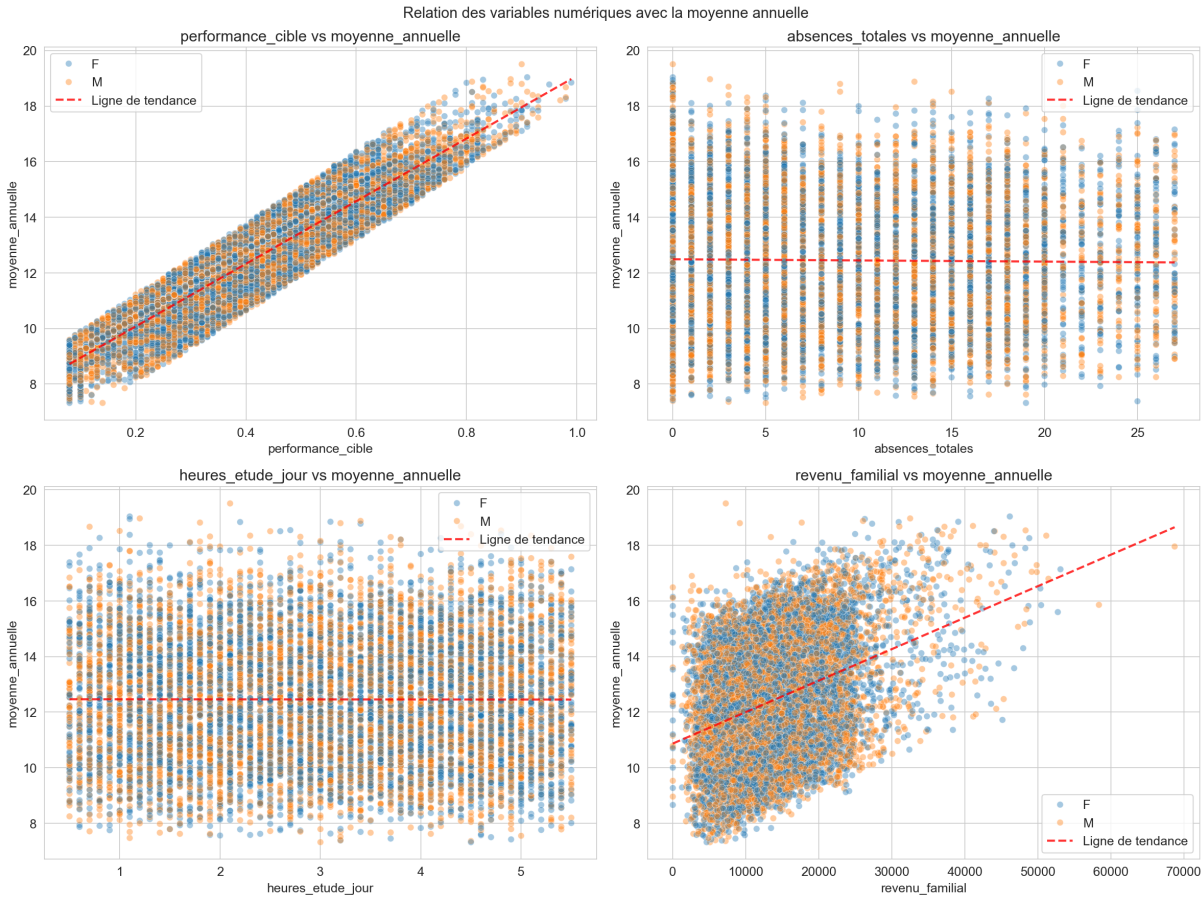

FIGURE 9 – Nuages de points des variables numériques en fonction de `moyenne_annuelle`, colorés par sexe avec ligne de tendance

Observations :

— `performance_cible` montre une relation linéaire **très forte** avec `moyenne_annuelle` (confirmation de la corrélation $r = 0{,}936$), car il s'agit d'une variable dérivée.

— `absences_totales` ne montre **pas de relation linéaire claire** avec la moyenne. La dispersion est uniforme, ce qui suggère que les absences seules ne suffisent pas à expliquer la performance.

— `heures_etude_jour` présente une dispersion **homogène**, sans tendance marquée.

— `revenu_familial` montre une relation **quasi inexistante** avec la moyenne annuelle.

— Aucune différence visible entre les sexes dans les nuages de points.

8.2 Variables catégorielles vs. `moyenne_annuelle`

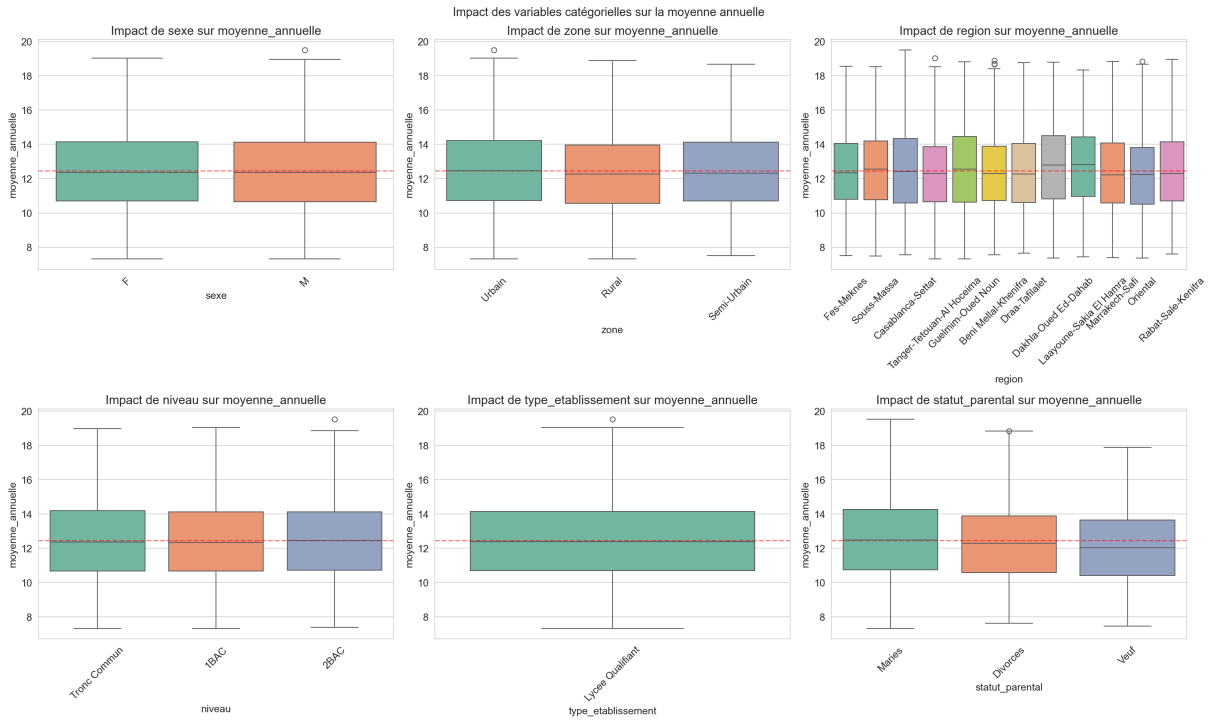

FIGURE 10 – Boxplots montrant l'impact des variables catégorielles sur `moyenne_annuelle`. La ligne rouge horizontale représente la moyenne globale.

Observations :

— **Sexe :** Les distributions sont **quasi identiques** entre les garçons et les filles, confirmant l'absence de biais de genre dans les performances académiques.

— **Zone :** Les médianes sont **très similaires** entre les zones urbaine, semi-urbaine et rurale.

— **Région :** Quelques variations entre les 12 régions, mais les différences restent **modestes**.

— **Niveau scolaire :** Les 3 niveaux (Tronc Commun, 1BAC, 2BAC) montrent des distributions **comparables**.

— **Type d'établissement :** Différence **minimale** entre les types.

— **Statut parental :** Distribution **homogène**, le statut marital des parents n'influence pas significativement la moyenne.

9 Analyse croisée (Cross-Tabulation)

Les tableaux croisés examinent la relation entre les variables catégorielles et la variable `probabilite_reussite`.

9.1 Sexe × Probabilité de réussite

Table 11 – Tableau croisé : Sexe × Probabilité de réussite (%)

Sexe	Élevé	Faible	Moyen	Très Faible
Féminin (F)	65,2	6,9	27,3	0,6
Masculin (M)	64,9	7,5	26,9	0,7
Total	**65,1**	**7,2**	**27,1**	**0,7**

Constat : Les proportions sont **quasi identiques** entre les sexes. Le sexe de l'étudiant ne semble pas être un facteur discriminant pour la probabilité de réussite.

9.2 Zone × Probabilité de réussite

Table 12 – Tableau croisé : Zone × Probabilité de réussite (%)

Zone	Élevé	Faible	Moyen	Très Faible
Rural	68,3	6,2	24,9	0,6
Semi-Urbain	65,2	7,3	26,9	0,6
Urbain	63,9	7,5	27,9	0,7

Constat : Étonnamment, les étudiants en zone **rurale** présentent un taux légèrement plus élevé de probabilité "Élevée" (68,3% vs 63,9% en urbain). Cela pourrait être un artefact des données synthétiques ou refléter un effet de sélection.

9.3 Niveau scolaire × Probabilité de réussite

Table 13 – Tableau croisé : Niveau × Probabilité de réussite (%)

Niveau	Élevé	Faible	Moyen	Très Faible
1BAC	66,6	7,2	25,7	0,6
2BAC	64,0	7,5	27,8	0,7
Tronc Commun	64,7	6,9	27,7	0,7

Constat : Les distributions sont **très homogènes** entre les trois niveaux scolaires. Le niveau d'études n'est pas un facteur fortement discriminant.

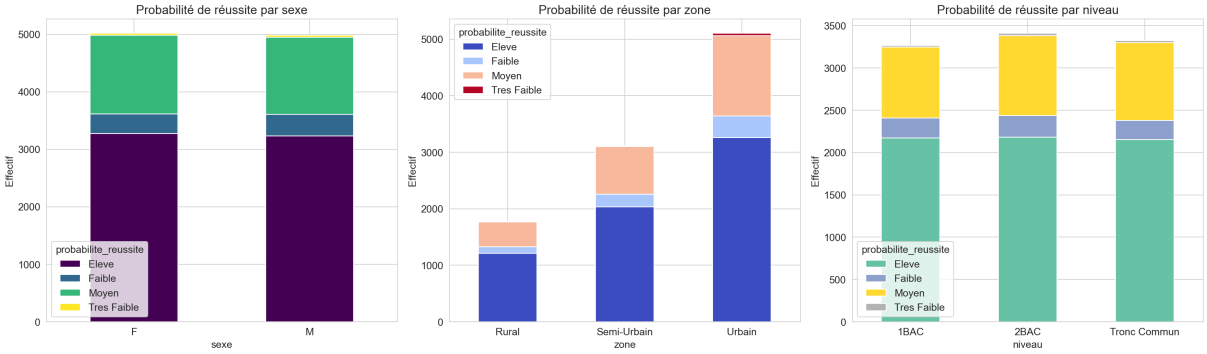

FIGURE 11 – Distribution de la probabilité de réussite par sexe, zone et niveau scolaire

10 Synthèse et recommandations

10.1 Résumé des constats principaux

TABLE 14 – Synthèse des résultats de l'EDA

Aspect	Constat
Taille	10 000 observations × 269 variables (98 numériques, 171 catégorielles)
Qualité	Très bonne. Seulement 7/269 colonnes avec valeurs manquantes. 0 doublon.
Variable cible	`moyenne_annuelle` (régression) : distribution quasi-normale, $\mu = 12{,}46$, $\sigma = 2{,}24$
Déséquilibre	`probabilite_reussite` : classe *Élevé* = 65,1%, classe *Très Faible* = 0,7%
Corrélations	Variables dérivées très corrélées ($r > 0{,}9$). Variables socio-économiques faiblement corrélées.
Asymétrie	8 variables avec |skewness| > 1 (revenus, absences, Tayssir)
Outliers	Significatifs uniquement pour `revenu_familial` (2,14%)
Discrimination	Sexe, zone et niveau ont un impact **minimal** sur la performance

10.2 Recommandations pour la modélisation

10.2.1 1. Nettoyage préalable

— **Supprimer** les colonnes `prise_notes` et `date_collecte` (100% manquantes).

— **Supprimer** les colonnes identifiantes : `id_etudiant`, `prenom`, `nom`, `nom_complet`, `code_massar`, `telephone`.

— **Supprimer** les variables à valeur unique : `annee_inscription`, `intervention_necessaire`, `id_collecteur`, `statut_verification`.

10.2.2 2. Gestion de la fuite de données (Data Leakage)

— **Exclure impérativement** des features : `moyenne_s1`, `moyenne_s2`, `performance_cible`, `rang_annuel`, `rang_s1`, `rang_s2`, `rang_annee_precedente`, et toutes les notes annuelles par matière (`*_annuel`).
— **Évaluer soigneusement** : `moyenne_annee_precedente` (informative mais potentiellement dérivée).

10.2.3 3. Ingénierie des features (Feature Engineering)

— **Transformation logarithmique** : appliquer $\log(x+1)$ aux variables à forte asymétrie (`revenu_familial`, `revenu_mensuel_pere`, `revenu_mensuel_mere`, `montant_tayssir`).
— **Encodage** : One-Hot Encoding pour les variables catégorielles à faible cardinalité, Label Encoding ou Target Encoding pour celles à haute cardinalité.
— **Binning de l'âge** : regrouper en catégories (17, 18, 19) si nécessaire.
— **Nouvelles variables** : ratio revenus/membres de famille, taux d'absences justifiées, interaction heures d'étude × participation aux cours de soutien.

10.2.4 4. Rééquilibrage des classes (pour la classification)

— Appliquer **SMOTE** (Synthetic Minority Over-sampling Technique) pour augmenter les classes minoritaires.
— Ou utiliser un **sous-échantillonnage stratifié** de la classe majoritaire.
— Utiliser des **poids de classe** dans les algorithmes supportant cette fonctionnalité.

10.2.5 5. Traitement des outliers

— **Winsorisation** de `revenu_familial` au 95^{e} percentile.
— Conserver les outliers modérés des variables académiques (peu nombreux).

10.2.6 6. Modèles recommandés

— **Régression :** Random Forest Regressor, Gradient Boosting (XGBoost/LightGBM), Ridge/Lasso pour la prédiction de `moyenne_annuelle`.
— **Classification :** Random Forest Classifier, XGBoost Classifier, SVM pour la prédiction de `probabilite_reussite`.
— **Baseline :** Régression linéaire et Logistic Regression pour établir des performances de référence.

11 Annexe : Erreurs détectées et corrections apportées

Lors de la revue de l'EDA originale, **6 erreurs** ont été identifiées et corrigées dans le notebook `EDA_Advanced.ipynb`. Les corrections suivantes ont été appliquées pour garantir la fiabilité des résultats :

Table 15 – Erreurs détectées et corrections apportées au notebook

#	Erreur originale	Correction appliquée
1	`performance_cible` décrite comme catégorielle	Corrigé : c'est une variable **numérique continue** (float64, 90 valeurs uniques)
2	`probabilite_reussite` décrite comme numérique	Corrigé : c'est une variable **catégorielle** (4 classes)
3	Corrélation calculée par rapport à `probabilite_reussite` (catégorielle ⇒ résultats incorrects)	Corrigé : corrélation calculée par rapport à `moyenne_annuelle`
4	Analyse bivariée utilisant `performance_cible` comme cible	Corrigé : utilisation de `moyenne_annuelle` comme cible
5	Tableau croisé avec `performance_cible` (90 valeurs float = tableaux illisibles)	Corrigé : utilisation de `probabilite_reussite` (4 catégories)
6	Pair plot avec hue = `performance_cible` (dégradé incohérent)	Corrigé : hue = `probabilite_reussite`

Fin du rapport – Analyse Exploratoire des Données (EDA)
Données : Morocco_Student_Data_Cleaned.csv | Source : EDA_Advanced.ipynb