

## **RAPPORT DE PROJET - S5.C.01**

Proposer une solution optimisée à partir de données  
internes et externes

SUJET : QUEL EST L'IMPACT DU TEMPS CONSACRÉ AUX JEUX  
VIDÉO, AINSI QUE DES DÉPENDANCES ASSOCIÉES, SUR LES  
PERFORMANCES SCOLAIRES ?

Arsan Abdi, Zen Ahmed-Kamal, Imman Amaladasse, Deraina Andriambala, Victoire Cassirame,  
Zakaria Gueddou, Mehdi Isaf, Rayan Kheroua, Estéban Marie-Louise, Vithushan Vijayatharan

# SOMMAIRE

<b>I. Introduction</b>	<b>4</b>
1 - Contexte et les enjeux du projet	4
2 - Objectifs principaux du projet et démarche adoptée.	4
<b>II. Conception et construction de la bdd relationnelle kaggle</b>	<b>6</b>
1 - Description des données	6
1.1 - Origine des données	6
1.2 - Structure et contenu	6
1.3 - Types de données	6
2 - Modélisation de la bdd	7
2.1 - Les tables	7
2.2 - Les dictionnaires	9
2.3 - Les relations	9
2.4 - Le MCD	9
3 - Implémentation de la bdd	10
3.1 - Création du schéma	10
3.2 - Fusion des données	10
3.3 - Extraction, Transformation, Insertion des données	11
3.4 - Choix de technologies	12
<b>III. Analyse de données Kaggle</b>	<b>14</b>
1 - Introduction	14
1.1 - Présentation	14
1.2 - Contexte	14
1.3 - Objectifs	14
2 - Exploration des données	15
2.1 - Introduction	15
2.2 - Statistiques descriptives clés	15
2.2.1 - Données de mathématiques	15
2.2.2 - Données de portugais	17
2.3 - Visualisation des données	18
2.3.1 - Présentation des graphiques réalisés	18
2.3.2 - Limites de l'analyse	25
2.3.3 - Conclusion	25
3 - Méthodes statistiques, corrélations et résultats	25
3.1 - Introduction	25
3.2 - Calcul des coefficients de corrélations sur le jeu de données “Student Alcohol Consumption”	26
3.2.1 - Première approche exploratoire: Coefficient de corrélation de Pearson	26
3.2.2 - Deuxième approche: Coefficient de corrélation de Spearman	26
3.3 - Test d'indépendance du Chi2	27

<b>3.4 - Test d'entropie</b>	<b>28</b>
<b>3.5 - Résultats des corrélations et interprétations</b>	<b>31</b>
<b>3.5.1 - Matrices de corrélation pour l'ensemble</b>	<b>31</b>
<b>3.5.2 - Matrices de corrélation par groupe</b>	<b>33</b>
<b>4 - Conclusion finale</b>	<b>40</b>
<b>IV. Collecte des données liées aux jeux vidéo</b>	<b>41</b>
<b>1 - Outil utilisé</b>	<b>41</b>
<b>1.1 - Choix de l'outil</b>	<b>41</b>
<b>1.2 - Choix du type de questions</b>	<b>41</b>
<b>1.3 - Items du questionnaire</b>	<b>41</b>
<b>1.2.1 - Items sociodémographiques</b>	<b>42</b>
<b>1.2.2 - Items liés à la performance scolaire</b>	<b>42</b>
<b>1.2.3 - Items liés à l'environnement scolaire</b>	<b>42</b>
<b>1.2.4 - Items liés à la consommation de jeux vidéo</b>	<b>42</b>
<b>1.2.5 - Items liés aux activités quotidiennes</b>	<b>43</b>
<b>1.3 - Rédaction et présentation du questionnaire</b>	<b>43</b>
<b>2 - Echantillon</b>	<b>44</b>
<b>3 - Tests préliminaires et ajustements du questionnaire</b>	<b>44</b>
<b>4 - Limites du questionnaire et de l'échantillon interrogé</b>	<b>45</b>
<b>5 - Axes d'amélioration</b>	<b>46</b>
<b>V. Conception et construction de la bdd relationnelle Jeux vidéo</b>	<b>48</b>
<b>1 - Nettoyage de données</b>	<b>48</b>
<b>1.1 - Nettoyage lié aux zones de réponses de libres</b>	<b>48</b>
<b>1.2 - Nettoyage de données incohérentes</b>	<b>49</b>
<b>2 - Modélisation de la bdd</b>	<b>49</b>
<b>2.1 - Les tables</b>	<b>49</b>
<b>2.2 - Les dictionnaires</b>	<b>52</b>
<b>2.3 - Les relations</b>	<b>52</b>
<b>2.4 - Le MCD</b>	<b>53</b>
<b>3 - Implémentation de la bdd</b>	<b>53</b>
<b>3.1 - Création du Schéma</b>	<b>53</b>
<b>3.2 - Extraction, Transformation, Insertion des Données</b>	<b>54</b>
<b>3.3 - Choix de technologies</b>	<b>54</b>
<b>VI. Analyse de données Jeu vidéo</b>	<b>55</b>
<b>1 - Description des algorithmes et des techniques d'analyse utilisées</b>	<b>55</b>
<b>1.1 - Introduction</b>	<b>55</b>
<b>1.2 - Corrélation de Spearman et premiers résultats</b>	<b>55</b>
<b>1.3 - Résultats</b>	<b>56</b>
<b>2. Analyse comparative</b>	<b>70</b>
<b>VII. Développement de l'interface web</b>	<b>71</b>
<b>1 - Introduction</b>	<b>71</b>
<b>2 - Description de l'interface web</b>	<b>71</b>
<b>2.1 - Onglet "A propos"</b>	<b>71</b>
<b>2.2 - Onglet "Accueil"</b>	<b>72</b>

<b>2.3 - Onglet “Etudes”</b>	<b>73</b>
<b>2.3.1 - Onglet “Etudes” partie “Student Alcohol Consumption”</b>	<b>73</b>
<b>2.3.2 - Onglet “Etudes” partie “Jeux vidéo”</b>	<b>78</b>
<b>3 - Technologies utilisées</b>	<b>85</b>
<b>4 - Charte graphique</b>	<b>85</b>
<b>5 - Hébergement AWS Front</b>	<b>85</b>
<b>6 - Lancer le projet en local</b>	<b>87</b>
<b>VIII. Conclusion</b>	<b>88</b>

# I. Introduction

## 1 - Contexte et les enjeux du projet

Le projet s'inscrit dans un contexte où les jeux vidéo occupent une place centrale dans les loisirs des étudiants, soulevant des interrogations sur leur impact potentiel sur les performances scolaires.

Effectivement, le lien entre jeux vidéo et performance scolaire est encore débattu et la littérature est partagée : si le temps consacré aux jeux vidéo peut réduire celui disponible pour les études, affecter le sommeil ou la concentration, de nombreuses études commencent également à explorer un impact positif; dans certains cas, les jeux vidéo permettent de développer des compétences utiles comme la résolution de problèmes ou la gestion du stress ou bien d'améliorer l'acquisition d'une langue étrangère.

En parallèle, d'autres paramètres, tels que l'organisation du temps liés à différentes activités quotidiennes (réseaux sociaux, tâches ménagères, loisirs divers), la santé physique et mentale ou l'environnement familial peuvent également influencer les résultats académiques. Il devient donc primordial d'explorer non seulement le lien direct entre la consommation de jeux vidéo et les performances scolaires, mais aussi de comprendre comment ces autres facteurs interagissent. Les enjeux sont à la fois scientifiques, pour mieux comprendre ces dynamiques complexes, et pratiques, afin de sensibiliser les étudiants et les éducateurs à l'importance d'un équilibre entre loisirs et études et à la bonne gestion de leur temps et leurs priorités.

## 2 - Objectifs principaux du projet et démarche adoptée.

Le projet vise à explorer le lien entre la consommation de jeux vidéo et les performances académiques. Il a également pour objectif d'identifier d'autres facteurs susceptibles d'influencer ces performances. De manière globale, cette étude vise à constituer une base solide pour de potentielles réutilisations de nos recherches, plus élargies et approfondies à l'avenir.

Par ailleurs, il s'inscrit dans une démarche pédagogique visant à comprendre les différentes étapes d'un processus de recherche : de la collecte et de l'analyse des données jusqu'à leur valorisation à travers le développement d'une interface graphique. Pour cela, notre démarche s'est articulée autour de plusieurs étapes clés :

### 1. Analyse préliminaire de l'étude Student Alcohol Consumption

Nous avons commencé par étudier l'impact de la consommation d'alcool sur les performances académiques à travers l'analyse de l'étude portugaise *Student Alcohol Consumption*. Cette étape nous a permis d'établir une base de comparaison afin de vérifier par la suite si les jeux vidéo influencent les performances de manière similaire. De plus, elle nous a aidés à identifier des paramètres pertinents ayant un impact sur les performances académiques, que nous avons inclus dans notre propre questionnaire (cf. étape 2).

## 2. Conception et réalisation d'un questionnaire

Pour répondre à notre besoin de collecter des données sur les jeux vidéo et performances académiques, nous avons conçu un questionnaire de 45 questions. Pour cela, nous nous sommes basés sur l'analyse de l'étude précédemment mentionnée et sur une recherche dans la littérature scientifique. Notre questionnaire a été proposé à des étudiants de divers établissements et domaines d'études.

Avant le lancement de notre étude, des tests préliminaires ont révélé que le questionnaire, initialement composé de 70 questions, était jugé trop long, intrusif et répétitif. Pour y remédier, nous avons supprimé les questions les moins pertinentes et reformulé certaines pour limiter leur caractère intrusif.

## 3. Analyse des données collectées

Les données obtenues via notre questionnaire ont été analysées avec précaution, en tenant compte des biais liés à notre échantillon (limité en termes d'effectifs). Cette approche rigoureuse visait à limiter les interprétations erronées et à garantir la fiabilité des résultats.

## 4. Structuration des données et développement d'une interface graphique

Parallèlement, nous avons structuré les données issues de l'étude *Student Alcohol Consumption* et celles issues de notre propre questionnaire dans deux bases de données distinctes. Une interface graphique a été développée pour permettre une interaction facilitée avec ces données et une exploration plus intuitive des résultats.

## II. Conception et construction de la bdd relationnelle kaggle

Cette partie vise à expliquer le processus d'extraction et de structuration des données recueillies dans l'étude "Student Alcohol Consumption".

L'objectif d'organiser les données est principalement de faciliter leur compréhension. Il s'agit également de pouvoir utiliser ses données pour des analyses mais aussi de pouvoir les visualiser dans une interface web.

### 1 - Description des données

#### 1.1 - Origine des données

Les données utilisées pour cette SAE proviennent du dataset intitulé "**Student Alcohol Consumption**", disponible sur Kaggle (<https://www.kaggle.com/datasets/uciml/student-alcohol-consumption>). Ce dataset résulte d'une enquête menée auprès d'élèves du secondaire au Portugal. Il contient des informations liées à la consommation d'alcool et aux performances académiques. Initialement conçu pour analyser les effets de la consommation d'alcool des lycéens sur leurs performances scolaires, il offre également un large éventail de variables socio-démographiques, familiales et sociales.

#### 1.2 - Structure et contenu

L'ensemble de données se compose de deux fichiers distincts :

- **student-mat.csv** : données relatives aux étudiants du cours de mathématiques (395 entrées).
- **student-por.csv** : données relatives aux étudiants du cours de portugais (649 entrées).

Au total, ces fichiers contiennent 1044 observations, représentant **674** élèves uniques (certains lycéens apparaissent dans les deux cours).

- **25 étudiants** ont suivi uniquement les cours de mathématiques (395 - **370**).
- **279 étudiants** ont suivi uniquement les cours de portugais (649 - **370**).
- **370 étudiants** ont suivi les 2 cours.

L'obtention de ces données est expliquée dans la partie "3.2 - Fusion des données".

#### 1.3 - Types de données

Les attributs présents dans le dataset sont variés, allant de variables binaires (e.g., **sex** : 'M' ou 'F') à des échelles numériques (e.g., **famrel** : 1 à 5) et des chaînes de caractères (e.g., **reason** : "réputation"). Ces données sont organisées dans les fichiers CSV susmentionnés.

Les données fournies par l'étude peuvent être regroupées en plusieurs catégories selon des thématiques distinctes :

- **Informations sociodémographiques** : sexe, âge, le lieu de résidence (rural ou urbain) et l'école fréquentée.
- **Facteurs familiaux** : taille de la famille, statut marital des parents, niveau d'éducation des parents et qualité des relations familiales etc.
- **Vie sociale** : temps libre, relations sociales, activités extrascolaires, situation amoureuse etc.
- **Performances académiques** : notes obtenues sur trois périodes (G1, G2, G3) en portugais et/ou en mathématiques, et les échecs scolaires.
- **Facteurs liés à la performance scolaire** : soutien scolaire, accès à Internet, motivation à poursuivre des études etc.
- **Habitudes liées à la scolarité** : temps de révision hebdomadaire, temps de trajet, absences.
- **Informations de santé** : consommation d'alcool pendant la semaine (Dalc) et le week-end (Walc), santé générale (health).

En tout 30 variables ont été relevées + 3 notes en portugais et/ou 3 notes en mathématiques.

## 2 - Modélisation de la bdd

Après observation des données de *Student Alcohol Consumption*, il est apparu que chaque enregistrement représente une observation complète et indépendante, et toutes les colonnes sont directement liées à cette observation. Les données ne présentent pas de relation de type *un-à-plusieurs* ou *plusieurs-à-plusieurs* entre les entités. Ainsi, une seule table semblait suffisante pour organiser efficacement les données tout en garantissant leur cohérence, et leur facilité d'accès pour les analyses futures. Seulement, une structure en une seule table n'aurait pas permis une bonne lisibilité ou une bonne compréhension des données. C'est pourquoi nous avons décidé, pour mieux agencer les données, de les organiser selon les catégories susdites.

Cette organisation thématique structure les données pour faciliter l'analyse des interactions entre les facteurs et leur impact sur la consommation d'alcool et les performances scolaires.

### 2.1 - Les tables

Voici les principales entités et leurs relations :

**Table student\_kaggle (étudiant) :**

- **sex** : Sexe de l'étudiant (F pour féminin, M pour masculin).
  - **age** : Âge de l'étudiant.
  - **address** : Type d'adresse de l'étudiant (U pour urbain, R pour rural).
  - **school** : École de l'étudiant (GP pour Gabriel Pereira, MS pour Mousinho da Silveira).
- La table contient également les clés étrangères permettant de relier un étudiant aux autres tables (student\_family, study\_habits, health, support, social\_life, performance).

#### Table performance (performances scolaires) :

- **G1\_por** : Note en portugais du premier trimestre (sur 20).
- **G2\_por** : Note en portugais du deuxième trimestre (sur 20).
- **G3\_por** : Note en portugais du troisième trimestre (sur 20).
- **G1\_math** : Note en mathématiques du premier trimestre (sur 20).
- **G2\_math** : Note en mathématiques du deuxième trimestre (sur 20).
- **G3\_math** : Note en mathématiques du troisième trimestre (sur 20).
- **failures\_por** : Nombre de fois où l'élève a échoué en cours de portugais.
- **failures\_math** : Nombre de fois où l'élève a échoué en cours de mathématiques.

#### Table student\_family (famille) :

- **famsize** : Taille de la famille (LE3 pour famille de 3 membres ou moins, GT3 pour plus de 3 membres).
- **Pstatus** : Statut parental (T pour parents vivant ensemble, A pour parents séparés).
- **mother\_job\_id** : Identifiant de la profession de la mère.
- **father\_job\_id** : Identifiant de la profession du père.
- **guardian\_id** : Identifiant du tuteur principal de l'étudiant.
- **mother\_edu\_id** : Identifiant du niveau d'éducation de la mère.
- **father\_edu\_id** : Identifiant du niveau d'éducation du père.

#### Table study\_habits (habitudes d'étude) :

- **studytime\_id** : Identifiant du temps hebdomadaire consacré à l'étude (<2 heures, 2-5 heures, 5-10 heures, ou >10 heures).
- **absences\_por** : Nombre total d'absences pour le cours de portugais.
- **absences\_math** : Nombre total d'absences pour le cours de maths.
- **travel\_id** : Identifiant du temps de trajet domicile-école (par intervalle : <15 min, 15-30 min, etc.)

#### Table support (environnement de travail / facteurs externes) :

- **schoolsup** : Soutien éducatif supplémentaire à l'école (true/false).
- **famsup** : Soutien éducatif familial (true/false).
- **internet** : Accès à Internet à domicile (true/false).
- **reason\_id** : Identifiant de la raison du choix du lycée (home, reputation, etc.).
- **paid\_math** : Indique si l'étudiant suit des cours particuliers payants pour les mathématiques (true/false).
- **paid\_por** : Indique si l'étudiant suit des cours particuliers payants pour le portugais (true/false).
- **higher** : Indique si l'étudiant souhaite poursuivre des études supérieures (true/false).
- **nursery** : Indique si l'étudiant a fréquenté la maternelle (true/false).

#### Table social\_life (vie sociale) :

- **famrel** : Qualité des relations familiales (1 à 5, 5 étant excellent).
- **freetime** : Temps libre après l'école (1 à 5, 5 étant beaucoup).
- **goout** : Fréquence des sorties avec des amis (1 à 5, 5 étant très souvent).
- **romantic** : Relation amoureuse (true/false).
- **activities** : Participation à des activités extrascolaires (true/false).

#### Table health (santé) :

- **health** : État de santé général (1 à 5, 5 étant excellent).

- **Dalc** : Consommation d'alcool en semaine (1 à 5, 5 étant élevée).
- **Walc** : Consommation d'alcool le week-end (1 à 5, 5 étant élevée).

## 2.2 - Les dictionnaires

Pour éviter la redondance de données, ces tables sont accompagnées de dictionnaires contenant simplement un id et une valeur textuelle (différentes valeurs possibles) :

- **parent\_education** (diplôme des parents)
- **parent\_job** (métier des parents)
- **guardian** (tuteur)
- **travel\_time** (temps de trajet pour se rendre à l'établissement)
- **study\_time** (temps de travail par jour)
- **reason** (raison d'étudier dans l'établissement actuel)

## 2.3 - Les relations

Les entités sont reliées par des clés primaires/étrangères pour garantir l'intégrité référentielle ce qui donne lieu à différentes relations :

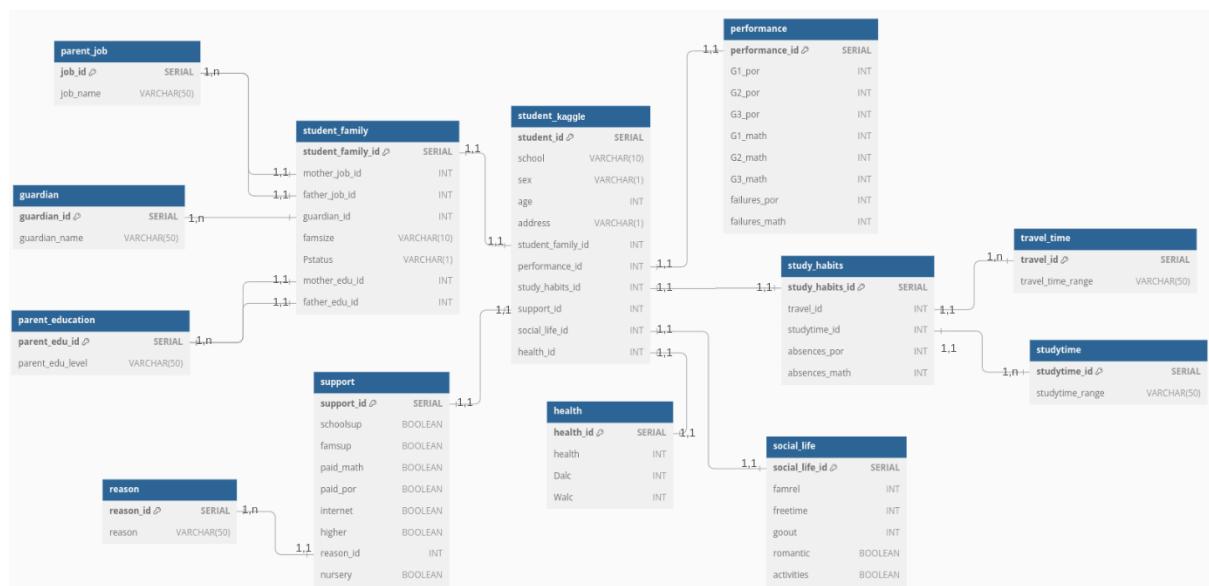
Les relations suivantes sont de type 1,1 - 1,1 :

- Toutes les relations comprenant la table "student". Comme indiqué précédemment, dans les données fournies par l'étude, chaque observation représente un étudiant et est indépendante.

Les relations suivantes sont de type 1,1 - 1,n :

- Toutes les relations comprenant un dictionnaire : une clé étrangère référence forcément une unique valeur de dictionnaire et une valeur de dictionnaire peut être utilisée 1 ou plusieurs fois.

## 2.4 - Le MCD



## 3 - Implémentation de la bdd

### 3.1 - Création du schéma

Les données Kaggle ont été structurées dans phpMyAdmin via un fichier SQL qui a servi à créer les tables. Ces tables sont disposées de manière à maintenir des relations claires, et des contraintes telles que les clés primaires et les clés étrangères assurent l'intégrité des données. (cf. 2 - Modélisation de la base de données).

### 3.2 - Fusion des données

Afin de gérer l'extraction des données, un premier script python nommé '**Kaggle\_Unified\_Csv.py**' fut créé pour pouvoir récupérer les données issus des deux fichiers csv.

Il était plus pratique pour nous de travailler sur un seul fichier mais il était surtout nécessaire de fusionner en un seul fichier les données puisque les 2 fichiers contenaient des observations communes (les étudiants ayant participé aux 2 cours).

Pour cela nous avons utilisé des indications fournies par les chercheurs de l'étude : ils indiquaient 13 colonnes (sex, age...) permettant d'identifier l'unicité d'un élève. En effet, si ces colonnes présentaient les mêmes valeurs dans les 2 csv, cela signifiait qu'il s'agissait d'un même élève ayant participé aux 2 cours.

La première version de ce script présentait des incohérences dans les données. Certaines valeurs censées être uniquement des **valeurs entières** positives avaient des **valeurs flottantes**, (ce qui était impossible car référençant une clé primaire).

De plus, après avoir utilisé les colonnes citées dans l'étude pour le merge des 2 fichiers, nous nous sommes aperçu d'absurdités après l'extraction des données. En effet, les chercheurs de *Student Alcohol Consumption* ont indiqué 382 étudiants ayant participé à la fois au cours de mathématiques et de portugais soit 13 étudiants ayant effectué seulement le cours de mathématiques (395 - 382). Or nous nous ne trouvions pas le même nombre. Après quelques recherches, nous avons découvert que le chiffre fourni par les chercheurs était incorrect et nous avons modifié la liste de colonnes servant à effectuer le merge afin de trouver les étudiants uniques. Nous avons finalement utilisé 27 des 33 variables afin d'identifier les élèves ayant suivi à la fois le cours de mathématiques et celui de portugais (soit 370 lycéens et non pas 382). Les autres étudiants n'ayant suivi qu'un seul cours ont été tout simplement reportés dans le nouveau csv nommé '**student-unified\_finale.csv**'.

Les 6 colonnes non utilisées pour le merge sont les 3 notes ainsi que les colonnes :

- "paid"
- "absences"
- "failures"

En effet, la colonne "paid" (cours particuliers) est décrite comme "paid - extra paid classes within the course subject (Math or Portuguese)". Cela signifie qu'elle dépend du cours de mathématiques ou de portugais.

Cependant, les colonnes “failures” et “absences” ne présentaient pas cette même mention dans la documentation fournie sur Kaggle, ce qui laissait penser que ces données ne dépendent pas du cours et donc que 1 étudiant du cours de portugais et qu’un autre du cours de mathématiques présentant les mêmes données sur 28 attributs mais ayant une réponse différentes au niveau de ces 2 colonnes auraient donc été 2 étudiants différents. Nonobstant, cela aurait été une remarquable conjoncture.

Ainsi, nous avons pris l’initiative de contacter directement les auteurs de l’étude afin d’obtenir les informations nécessaires au bon traitement des données.

Il s’avère finalement que “failures” et “absences” sont aussi propres au cours de mathématiques ou portugais.

D'où l'obtention finale des colonnes “failures\_math” et “failures\_por” (idem pour “paid” et “absences”).

### 3.3 - Extraction, Transformation, Insertion des données

Après l’étape de regroupement des données, il nous a fallu gérer l’extraction et l’insertion de ces dernières en base de données.

Pour cela, nous avons conçu un script ETL nommé ‘**Kaggle\_ETL.py**’ pour intégrer les données d'un fichier CSV dans une base de données locale. Ce script suit les trois étapes classiques du processus ETL : Extraction, Transformation et Load(Chargement). Tout d'abord, les données sont extraites depuis le fichier ‘**student-unified\_finale.csv**’ à l'aide de la bibliothèque Pandas. Elles sont chargées dans un DataFrame, une structure de données bidimensionnelle (lignes {indexé}, colonnes {noms}) similaire à une table relationnelle, les colonnes peuvent accueillir des types de données différents(float, integer, var...). Un DataFrame à 2 lignes et 2 colonnes s'organise comme suit:

	0	1
0	10	11
1	20	21

Les données sont organisées en lignes et colonnes, ce qui facilite leur manipulation et leur exploration. Ce choix permet de détecter et de corriger efficacement les données manquantes ou incorrectes avant leur transformation et leur chargement dans la base de données.

La phase de transformation adapte les données brutes aux contraintes du schéma relationnel de la base de données. Par exemple, des colonnes numériques comme "traveltime" et "studytime" sont converties en libellés textuels (ex. : "1" devient respectivement pour chacune de ces deux tables "<15 min." et "<2 hours"). Les relations entre les tables sont prises en charge grâce à un mécanisme d'insertion ou de récupération : si une valeur de "reason" ou "parent\_education" n'existe pas encore dans une table liée, elle est automatiquement ajoutée et son identifiant est utilisé dans les enregistrements correspondants.

Enfin, les données transformées sont chargées dans les différentes tables du schéma relationnel. Les relations entre les tables sont respectées en insérant les données dans un ordre logique, comme insérer les informations sur les familles et les performances avant de lier ces informations à chaque étudiant puisque c'est ce dernier qui possède les références aux autres tables.

L'implémentation de la base de données a donné lieu à des modifications fréquentes. Pour faciliter notre travail, la première étape du script consiste à vider les tables à l'aide d'un truncate pour garantir une réinsertion propre à chaque modification effectuée. Ce processus assure une intégration cohérente et conforme à la structure relationnelle définie.

Après avoir terminé de saisir les données sur phpMyAdmin, nous avons généré un fichier SQL à partir de phpMyAdmin, qui contient les données importées du Kaggle, afin de préparer notre présentation finale.

### 3.4 - Choix de technologies

**Pour le script ‘Kaggle\_ETL.py’ :** Nous avons choisi Python pour plusieurs raisons convaincantes. Tout d'abord, notre équipe possède une solide expérience avec ce langage, particulièrement dans les domaines de l'extraction et de la transformation de données, ce qui en fait un choix naturel. Python se distingue également par ses bibliothèques robustes qui simplifient grandement la manipulation des données. Parmi elles, Pandas est indispensable : elle nous aide à lire et traiter les données issues de fichiers CSV avec une grande aisance, facilitant des tâches telles que le filtrage et l'organisation des données conformément à nos exigences de base de données. De plus, l'utilisation de mysql.connector enrichit notre script en permettant une interaction fluide et sécurisée avec notre base de données MySQL. Ces outils ensemble rendent notre processus non seulement efficace mais également agréable et facile à gérer pour toute l'équipe.

**Pour le stockage des données :** Nous avons travaillé avec PhpMyAdmin, car c'est une des solutions mise à disposition par l'IUT pour gérer les bases de données. Cet outil était déjà familier à notre équipe ainsi qu'à l'équipe Web, grâce à nos précédents projets communs. Il offrait une simplicité d'utilisation et une continuité dans nos pratiques, ce qui a facilité le travail de tout le monde.

## III. Analyse de données Kaggle

*Analyse du jeu de données sur la consommation d'alcool et les performances scolaires*

### 1 - Introduction

#### 1.1 - Présentation

Cette partie a pour but de montrer l'évolution de notre analyse en mettant en lumière les différentes étapes de traitement des données de l'étude, depuis l'exploration initiale jusqu'à l'identification des corrélations entre la consommation d'alcool et les performances scolaires, tout en posant les bases pour une étude comparative intégrant le temps de jeu vidéo.

#### 1.2 - Contexte

Ce projet s'inscrit dans une démarche exploratoire visant à examiner l'influence de facteurs tels que la consommation d'alcool et le temps consacré aux jeux vidéo sur les résultats académiques.

Pour ce faire, l'analyse s'appuie en premier lieu sur le jeu de données intitulé "**Student Alcohol Consumption**", qui se concentre sur les habitudes des étudiants portugais du secondaire. Ce jeu de données, bien que centré sur la consommation d'alcool, intègre également des variables sociodémographiques et scolaires, permettant ainsi une approche multifactorielle.

#### 1.3 - Objectifs

L'objectif principal de ce projet est d'analyser l'impact de comportements spécifiques sur les performances scolaires des étudiants. Cette analyse vise à mieux comprendre les relations entre ces facteurs et les résultats académiques, tout en identifiant les variables pouvant influencer ces interactions.

Pour atteindre cet objectif global, des sous-objectifs ont été définis :

- Identifier les corrélations entre la consommation d'alcool et les résultats scolaires pour déterminer des tendances et des relations significatives (et pouvoir apporter des réponses à notre étude).
- Utiliser des analyses statistiques pour approfondir la compréhension des données et orienter la conception du questionnaire axé sur d'autres dépendances, notamment le jeu vidéo.

## 2 - Exploration des données

### 2.1 - Introduction

Dans cette section, nous présentons les analyses descriptives initiales réalisées sur le jeu de données intitulé “**Student Alcohol Consumption**”. Cette exploration a pour objectif principal de comprendre la structure des données, d’identifier les tendances globales et de mettre en évidence les variables clés susceptibles d’influencer les performances scolaires.

### 2.2 - Statistiques descriptives clés

#### 2.2.1 - Données de mathématiques

Le fichier de mathématiques contient un total de 395 lignes et 33 colonnes, chacune représentant une variable liée aux étudiants, telles que leur sexe, leur âge, leur statut familial ou encore leurs performances académiques. À titre d’exemple, la grande majorité des étudiants (88 %) est inscrite dans l’école Gabriel Pereira, tandis que 12 % fréquentent Mousinho da Silveira.

#### Analyse globale des variables clés

En étudiant la répartition géographique, on observe que 78 % des étudiants vivent en zone urbaine, tandis que 22 % résident dans des zones rurales. Les étudiants urbains semblent bénéficier d’un meilleur accès aux infrastructures éducatives, ce qui pourrait expliquer leurs performances légèrement meilleures en moyenne.

Les familles nombreuses représentent 71 % de l’échantillon tandis que 29 % des étudiants proviennent de familles plus petites. Bien que les familles nombreuses puissent offrir un soutien émotionnel accru, elles peuvent également limiter les ressources individuelles allouées à chaque enfant, influençant potentiellement les résultats scolaires.

Le niveau d’éducation des parents montre un léger déséquilibre : les mères atteignent un niveau moyen de 2,75 (équivalent à un enseignement secondaire), tandis que les pères affichent un niveau moyen de 2,55. Bien que cet écart soit relativement faible, il reste significatif dans le contexte éducatif. Un niveau d’éducation légèrement supérieur des mères pourrait refléter une implication plus directe dans l’encadrement scolaire et les activités éducatives des enfants.

#### Absences scolaires

Les absences varient significativement parmi les lycéens. La moyenne est de 5,2 jours par an, avec certains étudiants enregistrant jusqu’à 75 jours d’absence. Cependant, 75 % des étudiants accumulent moins de 8 absences. Une corrélation négative significative (-0,34) entre les absences et la note finale (G3) met en évidence l’impact négatif de l’absentéisme sur les performances académiques. Les étudiants ayant des absences fréquentes se retrouvent souvent dans les tranches inférieures de G3.

## **Performances académiques**

Les performances académiques montrent une tendance décroissante entre les évaluations G1, G2 et G3 :

- La moyenne passe de 10,91 pour G1 à 10,42 pour G3, révélant une perte de 0,49 point en moyenne. Cette diminution pourrait indiquer une pression accrue au fil de l'année ou une difficulté à maintenir les performances initiales.
- La dispersion des résultats s'accentue également : l'écart-type progresse de 3,31 pour G1 à 4,58 pour G3. Les notes finales montrent une variabilité accrue, avec des étudiants atteignant des scores aussi bas que 0 et aussi élevés que 20.
- Les femmes surpassent les hommes, avec une moyenne de G3 de 11,1 contre 9,6 pour les hommes. Cette différence, combinée à une dispersion plus faible pour les femmes, suggère une plus grande résilience et une meilleure constance dans leurs performances.

## **Analyse démographique**

L'âge des étudiants est majoritairement compris entre 15 et 18 ans avec une moyenne de 16,7 ans. Les tranches d'âge 16-18 ans regroupent 80 % de l'échantillon.

La proportion de femmes (53 %) est légèrement supérieure à celle des hommes (47 %).

## **Soutien scolaire (Extra scolaire & familial)**

Le soutien scolaire, qu'il soit formel (cours privés) ou informel (aide familiale), joue un rôle clé dans les résultats scolaires. Les étudiants bénéficiant d'un soutien familial ou de cours particuliers présentent des performances légèrement supérieures en G3 par rapport à ceux qui n'en bénéficient pas.

## **Consommation d'alcool**

La consommation d'alcool est mesurée à la fois en semaine (Dalc) et le week-end (Walc) :

- La plupart des étudiants consomment peu d'alcool en semaine, avec une moyenne de 1,5 sur une échelle de 1 à 5. Cela pourrait refléter une priorité donnée aux études pendant la semaine scolaire.
- La consommation d'alcool augmente significativement le week-end, avec une moyenne de 2,8. Cela peut indiquer des comportements sociaux plus marqués pendant les jours non scolaires.

## **Tendances et anomalies**

- Les notes scolaires suivent une distribution légèrement asymétrique, avec une concentration autour de 11 (médiane).
- Quelques étudiants obtiennent des notes de 0 en G2 et G3, probablement liés à des absences prolongées ou à un désengagement scolaire total.
- Les étudiants en zone urbaine (78 %) affichent des résultats légèrement supérieurs à ceux en zone rurale (22 %), ce qui pourrait être lié à un meilleur accès aux ressources éducatives.

## **Conclusion**

L'analyse descriptive met en lumière plusieurs observations clés :

- Une baisse des performances entre les évaluations intermédiaires (G1, G2) et la note finale (G3).
- Une forte corrélation entre le soutien scolaire, l'assiduité et les résultats académiques.
- Un écart notable entre les performances des femmes et des hommes, les femmes obtenant des résultats plus stables et élevés.
- Une consommation d'alcool modérée en semaine mais plus élevée le week-end, pouvant refléter un équilibre entre études et loisirs.
- Une influence significative des caractéristiques familiales, notamment le niveau d'éducation des parents et les dynamiques de soutien familial.

Ces observations constituent un socle robuste pour approfondir l'analyse. Elles permettent non seulement d'explorer plus en détail les liens entre les différentes variables et les performances scolaires, mais aussi de poser de nouvelles questions qui n'auraient peut-être pas été envisagées autrement. Elles offrent également une vision globale éclairant les priorités d'analyse et les aspects nécessitant une attention particulière.

### **2.2.2 - Données de portugais**

Pour rappel, le fichier concernant les étudiants en portugais contient 649 observations et 33 colonnes. Ces colonnes couvrent les mêmes aspects que dans le précédent fichier.

#### **Performances académiques**

Les notes des étudiants en portugais suivent une tendance globalement stable, avec des moyennes légèrement croissantes entre les évaluations :

- La note moyenne à la première évaluation (G1) est de 11,40 avec un écart-type de 2,75.
- La note moyenne à la deuxième évaluation (G2) est de 11,57, suggérant une légère amélioration.
- La note finale (G3) atteint une moyenne de 11,91, montrant une progression positive.

Ces résultats indiquent que les étudiants ont tendance à améliorer leurs performances au fil de l'année, contrairement aux tendances observées dans les données de mathématiques.

#### **Niveau d'éducation des parents**

Le niveau d'éducation des parents présente une moyenne de 2,51 pour les mères et 2,51 pour les pères, ce qui correspond majoritairement à un niveau d'éducation secondaire inférieur. Cette parité suggère une implication potentiellement équivalente des deux parents dans le suivi scolaire des enfants.

## **Soutien scolaire et absences**

Les étudiants ont en moyenne 4 jours d'absence, avec un maximum observé de 32 jours. 75 % des étudiants ont moins de 6 absences, mais une minorité affiche un absentéisme important, ce qui pourrait avoir un impact significatif sur les résultats scolaires.

## **Consommation d'alcool**

La consommation d'alcool des étudiants est modérée en semaine, avec une moyenne de 1,5 sur une échelle de 1 à 5. Cependant, la consommation augmente de manière significative le week-end, atteignant une moyenne de 2,8. Cela met en évidence une différence de comportement entre les jours scolaires et les jours de repos.

## **Observations et tendances**

- Les étudiants en portugais présentent une tendance positive dans leurs notes, avec une amélioration progressive au fil des évaluations.
- L'absence de disparité significative dans le niveau d'éducation des parents pourrait suggérer une influence similaire dans l'encadrement scolaire.
- Une analyse plus approfondie des absences et de la consommation d'alcool pourrait permettre d'identifier des corrélations avec les performances académiques.

## **Conclusion**

L'analyse des données de portugais permet de constater une progression positive des performances académiques au cours de l'année, en contraste avec les tendances observées en mathématiques. Des éléments tels que le soutien familial, les absences et la consommation d'alcool méritent une attention supplémentaire pour comprendre leur impact sur la réussite scolaire.

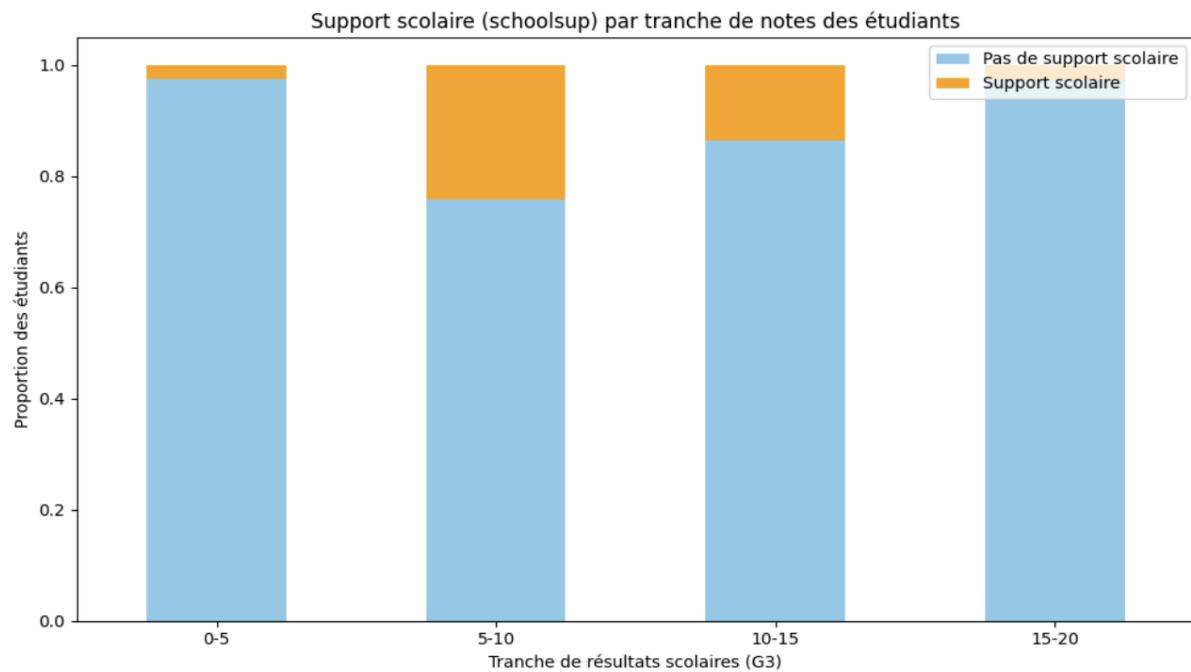
### **2.3 - Visualisation des données**

#### *Des données sur les notes de Mathématiques et de Portugais*

##### **2.3.1 - Présentation des graphiques réalisés**

Les visualisations effectuées permettent de mieux comprendre la répartition des résultats scolaires des étudiants en fonction de différents facteurs démographiques et sociaux. Les graphiques ci-dessous illustrent notamment :

### Soutien scolaire en fonction des tranches de notes (G3)

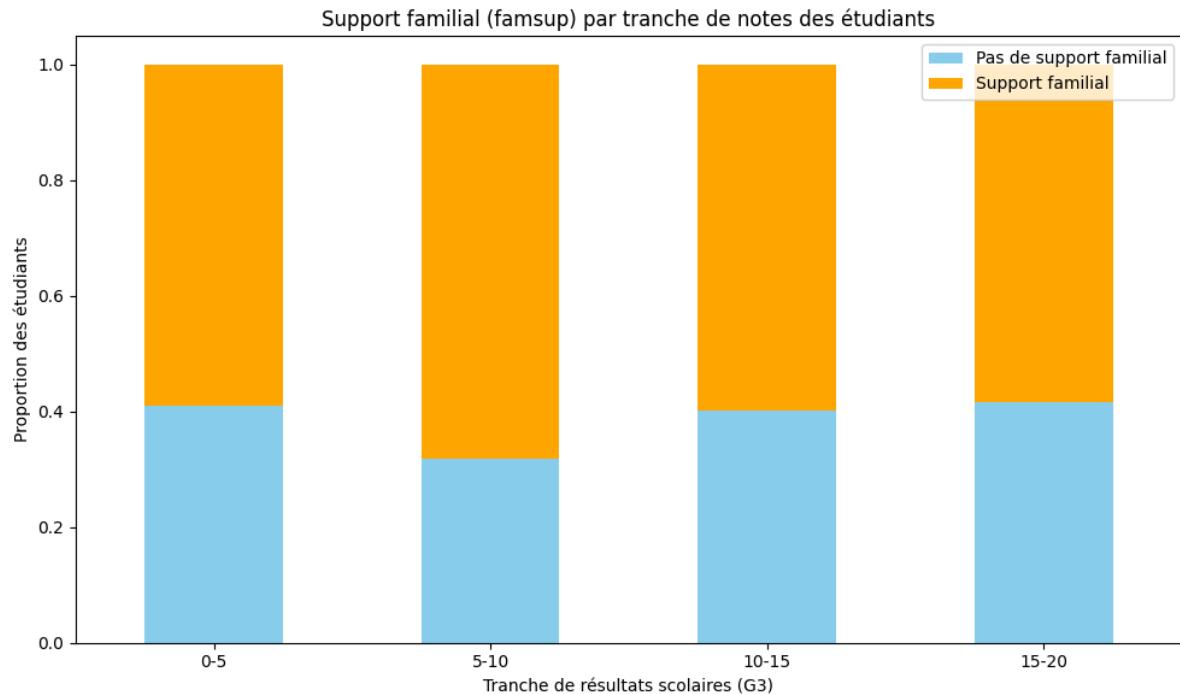


Ce graphique illustre la répartition des élèves selon les tranches de notes et la présence ou l'absence de soutien éducatif extra-scolaire (schoolsup). Une observation claire se dégage : la majorité des étudiants ne bénéficient pas de soutien éducatif extra-scolaire, indépendamment de leur tranche de notes.

- Parmi les élèves ayant des notes très faibles (0 à 5), environ 3 % bénéficient de soutien éducatif. Cette proportion augmente légèrement pour les élèves ayant des notes comprises entre 5 et 10, atteignant environ 25 %.
- Dans la tranche des élèves ayant des notes moyennes (10 à 15), la proportion diminue à environ 15 %.
- Enfin, pour les élèves ayant des notes élevées (15 à 20), le soutien extra-scolaire est presque inexistant.

Ces résultats suggèrent que le soutien éducatif extra-scolaire est principalement ciblé sur les élèves en difficulté (notes faibles à intermédiaires). Une hypothèse plausible est que les parents ou l'école jugent qu'un soutien supplémentaire est essentiel pour les élèves ayant des notes faibles ou moyennes afin de les aider à s'améliorer. À l'inverse, pour les élèves ayant des notes très élevées ou très faibles, il se peut que l'utilité d'un tel soutien soit perçue comme limitée : soit ces élèves n'en ont pas besoin, soit leurs résultats sont jugés trop faibles pour que ce type d'accompagnement soit bénéfique.

### Soutien familial en fonction des tranches de notes (G3)



Ce graphique montre la répartition des élèves selon les tranches de notes et la présence ou l'absence de soutien éducatif familial (famsup). Contrairement à schoolsup, le soutien familial est largement dominant dans toutes les tranches de notes, ce qui reflète un investissement constant des familles dans l'éducation des élèves, quelles que soient leurs performances.

- Pour les élèves ayant des notes très faibles (0 à 5), environ 60 % bénéficient de soutien familial.
- Cette proportion reste constante pour les tranches suivantes (5 à 10, 10 à 15 et 15 à 20), avoisinant les 60 % à 70 % pour chaque catégorie.

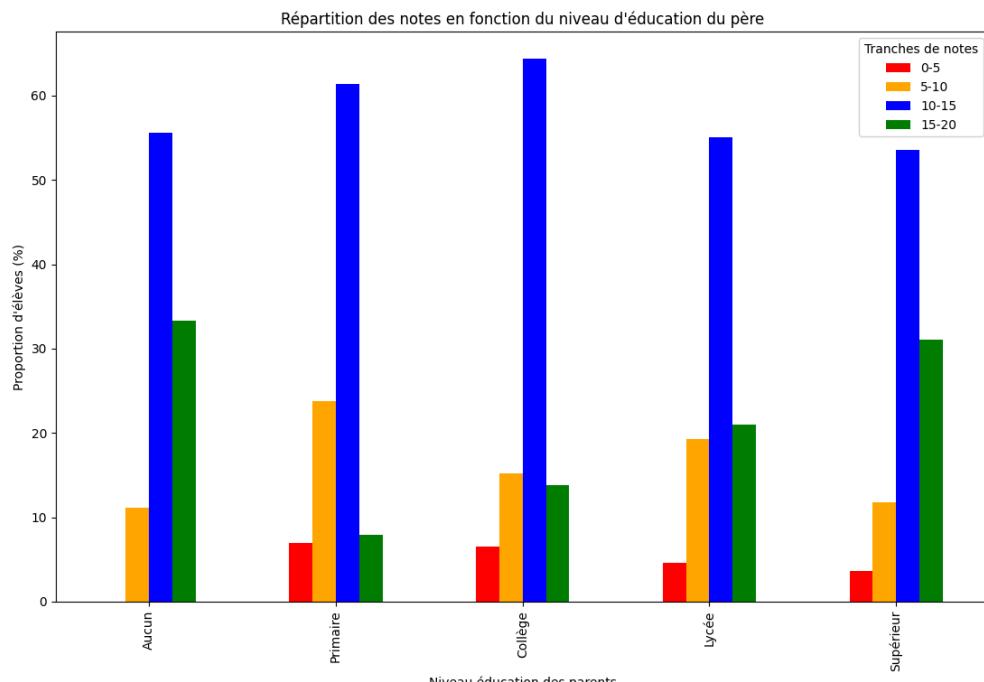
Ces résultats mettent en lumière une implication importante des familles dans l'accompagnement scolaire de leurs enfants, quel que soit leur niveau académique. Une interprétation possible est que les familles perçoivent leur rôle comme essentiel pour soutenir les résultats scolaires de leurs enfants, qu'ils soient faibles ou élevés. Contrairement à schoolsup, le soutien familial ne semble pas ciblé sur un groupe spécifique mais plutôt généralisé. Cela souligne l'importance du cadre familial dans le soutien éducatif global.

Les deux graphiques (**schoolsup & famsup**) révèlent des stratégies de soutien distinctes :

- Schoolsup est principalement orienté vers les élèves ayant des notes faibles à intermédiaires, probablement dans le but d'améliorer leurs performances.
- Famsup, en revanche, est distribué de manière homogène entre les différentes tranches de notes, montrant une implication constante des familles.

Ces analyses mettent en évidence des approches complémentaires entre le soutien éducatif familial et extra-scolaire, chacune visant des objectifs différents pour accompagner les élèves dans leurs performances académiques.

## Influence du niveau d'éducation du père sur les résultats scolaires



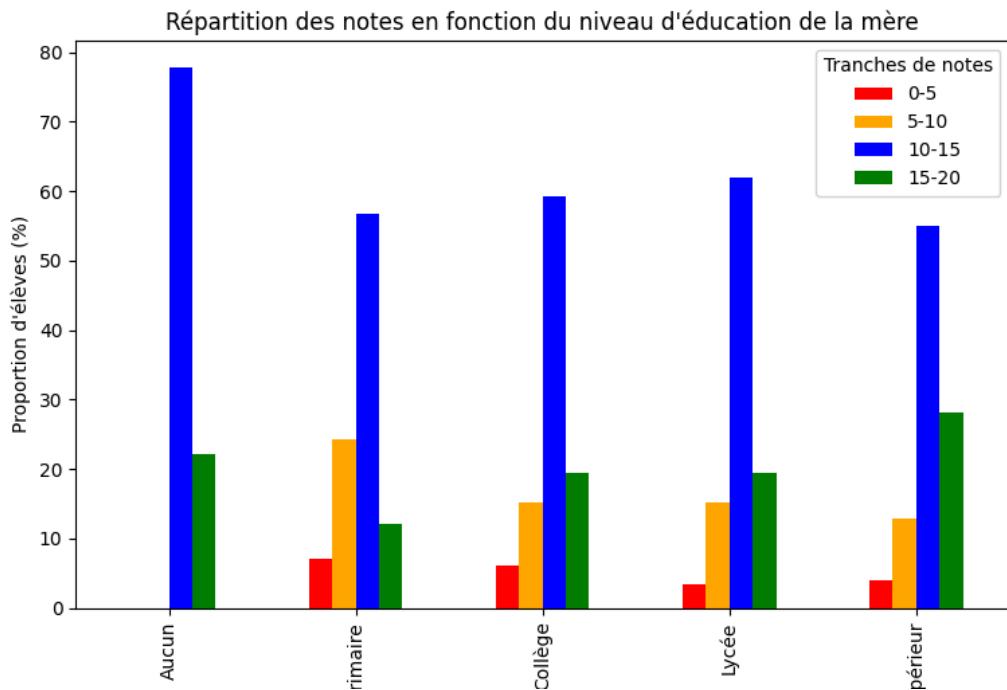
Le graphique ci-dessus illustre la répartition des étudiants selon leurs résultats scolaires (0-5, 5-10, 10-15, et 15-20), en fonction du niveau d'éducation de leur père. Une tendance claire se dégage : plus le niveau d'éducation du paternel est élevé, plus les résultats scolaires sont élevés. On observe une transition des notes faibles (0-5 et 5-10) vers des notes plus élevées (10-15 et 15-20) à mesure que le niveau d'éducation augmente.

Les étudiants dont le père a un faible niveau d'éducation ( primaire ou aucun) se concentrent majoritairement dans les tranches de notes faibles à intermédiaires (0-5 et 10-15). Une proportion notable d'entre eux obtient des notes très faibles (0-5), tandis que les notes élevées (15-20) sont rares dans ce groupe. En revanche, pour les étudiants dont le père possède un niveau d'éducation secondaire ou supérieur, la majorité des résultats se situe dans les tranches élevées (10-15 et 15-20), tandis que les notes faibles deviennent marginales.

La catégorie "Aucun" pour le niveau d'éducation du père est ignorée dans l'analyse en raison du faible nombre d'élèves concernés, ce qui limite la fiabilité des conclusions pouvant en être tirées.

Une corrélation claire apparaît : les étudiants avec un père ayant un niveau d'éducation supérieur ont de meilleures performances académiques. À l'inverse, les élèves issus de familles avec un faible niveau d'éducation paternel pourraient bénéficier d'un soutien supplémentaire pour réduire ces inégalités.

## Influence du niveau d'éducation de la mère sur les résultats scolaires

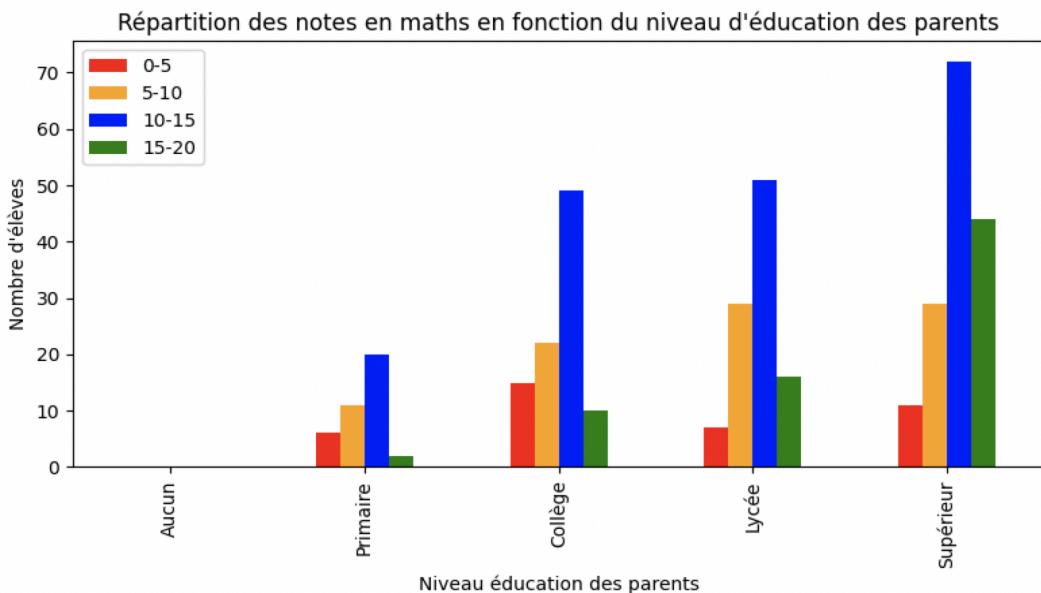


Le graphique ci-dessus illustre la répartition des étudiants selon leurs résultats scolaires (0-5, 5-10, 10-15, et 15-20), en fonction du niveau d'éducation de leur mère. Une tendance générale se dégage : plus le niveau d'éducation maternel est élevé, plus les résultats scolaires des étudiants sont élevés. On observe une transition des notes faibles (0-5 et 5-10) vers des notes plus élevées (10-15 et 15-20) à mesure que le niveau d'éducation augmente.

Les étudiants dont la mère a un faible niveau d'éducation ( primaire ou aucun) se concentrent majoritairement dans les tranches de notes faibles à intermédiaires (0-5 et 10-15). Une proportion notable d'entre eux obtient des notes très faibles (0-5), tandis que les notes élevées (15-20) restent rares. En revanche, pour les étudiants dont la mère possède un niveau d'éducation secondaire ou supérieur, la majorité des résultats se situe dans les tranches élevées (10-15 et 15-20), tandis que les notes faibles deviennent marginales.

Une corrélation claire apparaît : les étudiants avec une mère ayant un niveau d'éducation supérieur ont de meilleures performances académiques. À l'inverse, les élèves issus de familles avec un faible niveau d'éducation maternel pourraient bénéficier d'un soutien supplémentaire pour réduire ces inégalités. Une attention particulière pourrait être portée aux élèves les plus en difficulté afin de favoriser une plus grande équité dans les performances scolaires.

### Influence du niveau d'éducation des parents sur les notes de mathématiques



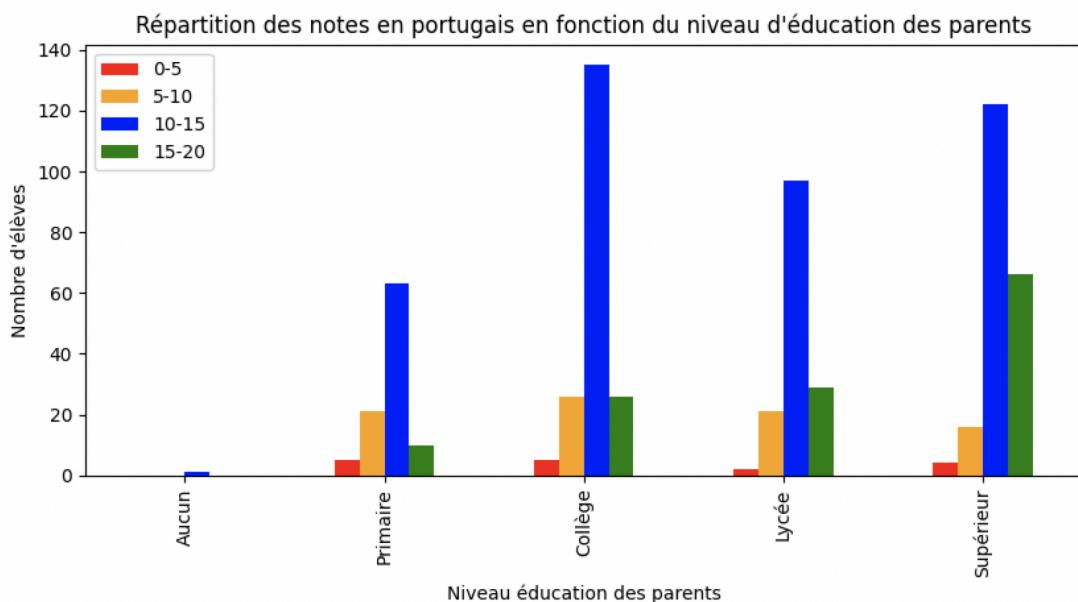
Le graphique ci-dessus met en lumière la relation entre le niveau d'éducation des parents et les performances scolaires des étudiants en mathématiques. On aperçoit qu'à mesure que le niveau d'éducation des parents augmente, la répartition des notes évolue vers des résultats plus élevés. Les étudiants dont les parents ont un niveau d'éducation primaire affichent majoritairement des résultats compris entre 5 et 10, traduisant une certaine difficulté à atteindre des performances supérieures. De plus, une proportion non négligeable d'entre eux obtient des notes très faibles (0-5), ce qui suggère un manque potentiel de ressources éducatives ou de soutien académique à la maison. Cette catégorie est également caractérisée par une quasi-absence d'élèves atteignant des résultats excellents (15-20), révélant une potentielle limite dans l'encadrement éducatif offert par ces parents.

À mesure que le niveau d'éducation des parents progresse vers un niveau intermédiaire (collège et lycée), on observe une répartition plus équilibrée des résultats, avec une montée en puissance de la tranche 10-15 et une diminution des notes faibles. Cependant, malgré cette amélioration, la tranche de notes 5-10 reste encore bien représentée, indiquant que les difficultés ne disparaissent pas totalement, mais que l'environnement familial devient progressivement plus favorable aux apprentissages. Les étudiants issus de parents ayant un niveau d'éducation supérieur affichent, quant à eux, une nette amélioration des résultats. La majorité de ces élèves obtiennent des notes comprises entre 10 et 20, tandis que les performances très faibles deviennent quasi inexistantes. Ces résultats laissent penser que des parents plus instruits sont mieux à même de soutenir efficacement leurs enfants, de leur fournir un encadrement pédagogique approprié et de leur offrir un accès plus large à des ressources éducatives.

Cette analyse met en évidence une corrélation positive entre le niveau d'éducation des parents et la réussite scolaire de leurs enfants. Les parents plus éduqués semblent en effet mieux préparer leurs enfants aux exigences académiques, que ce soit en leur transmettant des méthodes d'apprentissage, en les encourageant dans leur parcours ou en mettant à leur disposition des ressources adéquates. Toutefois, il est important de noter que, même dans la catégorie des parents les plus éduqués, certains étudiants obtiennent encore des notes moyennes, soulignant ainsi l'impact d'autres facteurs, tels que la motivation intrinsèque de l'étudiant, son environnement social ou encore la qualité de l'enseignement reçu.

Face à ces constats, une question se pose : l'analyse séparée du niveau d'éducation du père et de la mère est-elle suffisante pour comprendre pleinement l'impact du cadre familial sur la réussite scolaire ? En effet, combiner le niveau d'éducation des deux parents (*on prend en compte le niveau d'éducation du parent ayant atteint le niveau le plus avancé*) permettrait peut-être d'obtenir une vision plus complète et de mieux identifier les interactions familiales influençant la performance des étudiants. Une telle approche pourrait révéler des corrélations plus significatives et offrir des pistes d'amélioration plus ciblées pour accompagner les élèves dans leur parcours académique.

### Influence du niveau d'éducation des parents sur les notes de portugais



Le graphique ci-dessus illustre la répartition des notes en portugais des étudiants selon le niveau d'éducation de leurs parents. Une tendance générale se dégage : plus le niveau d'éducation parental est élevé, plus les performances scolaires s'améliorent, avec une augmentation notable du nombre d'élèves obtenant des notes élevées (15-20) et une diminution des résultats faibles (0-5).

Les élèves dont les parents ont un niveau d'éducation primaire se concentrent principalement dans la tranche des notes 10-15, avec une présence encore significative dans la tranche 5-10. Cette répartition suggère un encadrement éducatif limité, bien que les notes très faibles (0-5) restent minoritaires. En revanche, à partir du niveau "Collège", la répartition des notes évolue favorablement avec une progression vers la tranche 15-20, illustrant un effet positif du niveau d'instruction parental sur la réussite scolaire.

Lorsque les parents atteignent un niveau d'éducation "Lycée" ou "Supérieur", la majorité des étudiants obtiennent des notes comprises entre 10 et 20, tandis que les résultats les plus faibles deviennent rares. Ces données mettent en évidence l'importance du cadre familial dans la réussite scolaire, suggérant qu'un niveau d'éducation parental plus élevé est associé à un soutien éducatif plus efficace et à un accès élargi aux ressources académiques.

Cette analyse conduit à une réflexion sur la nécessité d'examiner l'impact combiné du niveau d'éducation des deux parents. Une analyse conjointe facilite la mise en évidence de la corrélation entre l'environnement éducatif familial dans son ensemble et la réussite scolaire des étudiants. Cette approche permet d'évaluer l'effet combiné de l'éducation des parents sans devoir analyser les contributions individuelles, qui peuvent parfois se compenser.

### 2.3.2 - Limites de l'analyse

Tout d'abord, la **représentativité de l'échantillon** pose question, car les données concernent uniquement des étudiants portugais, ce qui peut limiter la généralisation des résultats à d'autres contextes. De plus, **les biais de sélection** liés à l'absence de certaines variables clés, telles que le statut socio-économique ou les conditions de travail des parents, peuvent influencer les conclusions.

L'agrégation du niveau d'éducation des parents sans distinction entre le père et la mère constitue également une limite, masquant potentiellement des différences d'impact selon le rôle de chacun dans le suivi scolaire. Enfin, **les données manquantes ou imprécises** compromettent la fiabilité des analyses pour certaines sous-populations.

Pour améliorer les futures études, il serait pertinent d'inclure des variables supplémentaires, comme l'engagement parental dans les études, et d'envisager une approche longitudinale afin de mieux comprendre l'évolution des performances scolaires dans le temps.

### 2.3.3 - Conclusion

L'étude confirme que le niveau d'éducation des parents est un facteur déterminant dans la réussite scolaire des élèves. Les étudiants dont les parents ont un niveau d'éducation supérieur obtiennent de meilleurs résultats, tandis que ceux issus de milieux moins instruits rencontrent plus de difficultés académiques.

## 3 - Méthodes statistiques, corrélations et résultats

### 3.1 - Introduction

L'analyse des corrélations vise à identifier les relations potentielles entre différentes variables afin de mieux comprendre les facteurs influençant les résultats académiques. Après avoir exploré les données de manière descriptive, il est essentiel d'adopter des méthodes statistiques rigoureuses pour valider ces relations et vérifier leur significativité.

Dans cette partie, nous présenterons les différentes approches statistiques utilisées pour évaluer les liens entre les variables, notamment les coefficients de corrélation de Pearson et Spearman, le test d'indépendance du Chi2 ou encore l'entropie. Ces analyses permettront de déterminer si des associations significatives existent entre les habitudes des étudiants, comme leur consommation d'alcool ou leur engagement dans les jeux vidéo, et leurs performances scolaires.

Les résultats obtenus seront discutés en tenant compte des limites des données disponibles.

### 3.2 - Calcul des coefficients de corrélations sur le jeu de données “Student Alcohol Consumption”

#### 3.2.1 - Première approche exploratoire: Coefficient de corrélation de Pearson

Lors de notre analyse initiale, inspirée par les études disponibles sur Kaggle ainsi que par les ressources mises à notre disposition, nous avons calculé une matrice de coefficient de corrélation linéaire (dite de Pearson) comme outil d'exploration de données.

Cependant, au fil de nos recherches, notamment grâce à la lecture de la documentation des différentes librairies Python et par l'exploration de tests statistiques pouvant nous être utiles, nous avons réalisé que ce choix n'était pas pleinement adapté. En effet, l'utilisation du coefficient linéaire ne respectait pas certaines conditions nécessaires à son application.

En effet, le coefficient de Pearson<sup>1</sup> est une mesure permettant d'évaluer la force d'une relation linéaire entre deux variables quantitatives continues (Autrement dit des variables mesurables continues). Or, dans notre cas, nous avons, des variables *principalement ordinaires*; elles sont certes numérisées mais elles traduisent un ordre sans que les écarts entre les valeurs aient une signification claire. Par exemple, Dalc représente une échelle de consommation d'alcool quotidienne allant de “très faible” à “très élevée”, ici la différence entre “très élevée” et “elevée” est difficilement interprétable.

Le fait que le type de nos variables ne remplisse pas la condition d'utilisation du coefficient de Pearson nous a motivé à utiliser un autre outil à savoir le coefficient de Spearman. Dans la section suivante, nous présenterons cet autre coefficient. Nous expliquerons également que bien que l'utilisation du coefficient de Pearson ne soit pas rigoureusement appropriée, cela n'invalider pas totalement notre approche exploratoire.

#### 3.2.2 - Deuxième approche: Coefficient de corrélation de Spearman

Le coefficient de corrélation de Spearman est, en quelque sorte, un équivalent du coefficient de Pearson. A la différence près qu'il s'applique aussi bien que sur des variables ordinaires que des variables mesurables. Le calcul de cette corrélation utilise non pas la valeur de la variable mais son rang.<sup>2</sup> Par exemple, pour une variable sur temps passé sur les réseaux sociaux par jour ayant comme échelle “Inférieur à 1 h”, “Plus de 3 h” et “Entre 1h et 3h”, les rangs associées sont respectivement 1, 3 et 2 (Ici, selon un classement croissant). Ce coefficient mesure la force de concordance de classement entre deux variables.

La formule du coefficient de Spearman est donnée ci-dessous.  $n$  représente le nombre d'observations et  $d_i$  la différence de rang.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

---

<sup>1</sup> [http://www.biostat.ulg.ac.be/pages/Site\\_r/corr\\_pearson.html](http://www.biostat.ulg.ac.be/pages/Site_r/corr_pearson.html)

<sup>2</sup> Saporta, G. (2011). Probabilités, analyse des données et statistique. Editions TECHNIP.

Sur notre jeu de données nous avons 11 variables représentant différentes échelles qui sont donc des variables ordinaires. De plus, à y considérer, les variables représentant les notes ne sont pas des variables quantitatives à strictement parler; en effet une note de 10 en maths ne signifie pas que l'étudiant est deux fois plus fort qu'un autre ayant obtenu 5. (Au contraire de, par exemple, une variable mesurant la largeur en cm). L'utilisation du coefficient de Spearman est donc pertinent ici car il est pleinement adapté à la nature ordinaire de nos variables.

Nous avons comparé les coefficients de Pearson et de Spearman afin de déterminer si des différences significatives auraient pu influencer nos interprétations. Que ce soit pour les données sur les notes en mathématiques ou en portugais, à l'exception de deux couples de variables, les valeurs des coefficients diffèrent de manière négligeable (au maximum  $\pm 0,05$ ) et n'entraîne pas le passage à d'un niveau de corrélation à un autre. (Par exemple, on ne passe pas de corrélation faible à modérée). Les deux exceptions concernent les données sur les notes en mathématiques:

- Le couple de variable temps de trajet et consommation quotidienne d'alcool a comme coefficient de Pearson de 0,14 et 0,06 pour le coefficient de Spearman. Le couple temps de trajet et consommation d'alcool le week-end donne des résultats similaires.
- Le couple de variables notes et consommation quotidienne d'alcool a comme coefficient de Pearson de -0,05 et -0,12 pour le coefficient de Spearman. Le couple notes et consommation d'alcool le week-end donne des résultats similaires.

### 3.3 - Test d'indépendance du Chi2

Un test d'indépendance du Chi2 permet de vérifier l'absence de lien statistique entre deux variables. Nous souhaitons vérifier si les variables Dalc (Consommation d'alcool en semaine) et G3 (note finale en mathématiques) sont indépendantes à l'aide d'un test d'indépendance du Chi2.

**Hypothèse nulle:** Les variables G3 et Dalc sont indépendantes entre elles.

**Hypothèse alternative:** Les variables Dalc et G3 ne sont pas indépendantes.

Le coefficient de corrélation entre ces variables calculé est de -0,12. Le test d'indépendance du Chi2 nous permettra d'évaluer statistiquement cette corrélation.

#### Conditions d'application du test

Pour que ce test soit valide, l'effectif théorique de chaque case du tableau de contingence doit être au moins de 5. Dans notre cas, G3 est une variable pouvant avoir 20 valeurs possibles, Dalc est une variable contenant 5 catégories possibles et les données contiennent 396 observations. Cela amène à un effectif théorique faible d'environ 3.

Les données ont donc été regroupées pour satisfaire ce prérequis, nous obtenons ce tableau.

	0-6	7-9	10-12	13-15	16-20
--	-----	-----	-------	-------	-------

Très Faible	42	41	91	70	32
Faible	16	15	27	12	5
Moyenne	1	8	10	4	3
Élevée à très élevée	2	5	6	7	0

## Résultat du test

Après calcul, la valeur-p, c'est-à-dire la probabilité de rejeter à tort l'hypothèse nulle, est de 18%.

## Conclusion

Avec une valeur-p de 18%, nous ne pouvons rien conclure sur la dépendance ou l'indépendance entre les variables Dalc et G3.

### 3.4 - Test d'entropie

L'entropie est une mesure de la dispersion ou de l'incertitude dans une distribution de données. Elle quantifie le degré d'imprévisibilité ou de désordre au sein d'un ensemble de valeurs.

#### Interprétation de l'entropie :

- **Faible entropie** : Une distribution concentrée sur quelques valeurs (les notes sont similaires ou prévisibles).
- **Haute entropie** : Une distribution répartie sur de nombreuses valeurs (les notes sont variées et imprévisibles).

#### Utilisation de l'entropie

L'analyse de l'entropie a été réalisée en regroupant les étudiants selon leur temps de travail à l'aide d'une fonction qui classe la variable **studytim**e en trois catégories distinctes :

**Faible** :  $\text{studytim} \leq 2$ , **Modéré** :  $\text{studytim} = 3$  et **Intensif** :  $\text{studytim} \geq 4$

Cette classification permet d'analyser la dispersion des notes finales (G3) en fonction du temps d'étude des étudiants.

Pour chaque groupe (Faible, Modéré, Intensif), la distribution des notes finales a été calculée. Par exemple, dans le groupe **Faible**, le nombre d'étudiants ayant obtenu chaque note (de 0 à 20) a été comptabilisé, puis transformé en proportions.

Ainsi, si 10 étudiants ont obtenu la note de 10 sur un total de 50 étudiants dans le groupe, la proportion associée à cette note est calculée de la manière suivante : **10 / 50 = 0.2**.

Ces calculs permettent d'obtenir une distribution normalisée des notes finales au sein de chaque groupe. C'est une étape qu'il faut réaliser au préalable du calcul de l'entropie. Voici comment cela fonctionne :

**Formule de l'entropie de Shannon :**

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

- $x_i$  est la proportion d'une note  $i$  dans le groupe.
- $H$  est l'entropie.

On applique cette formule en utilisant la fonction `scipy.stats.entropy`. Par exemple :

- Si un groupe a des proportions [0.5, 0.5], l'entropie est élevée.
- Si un groupe a des proportions [1.0, 0.0], l'entropie est faible.

Les proportions permettent de mesurer la diversité des valeurs dans un groupe.

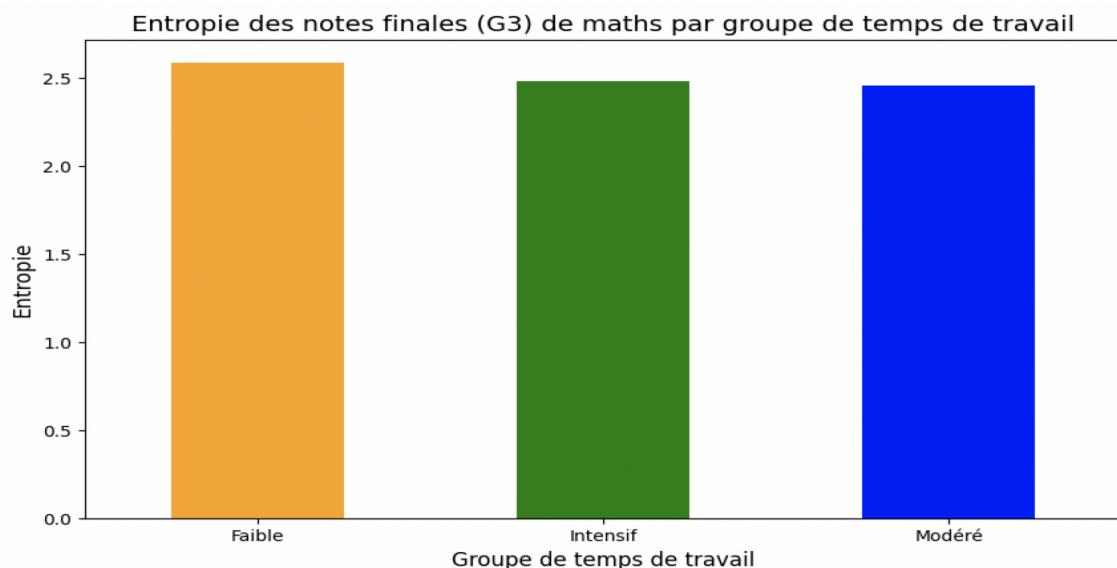
- **Quand toutes les proportions sont égales (par exemple, [0.5, 0.5]),** cela signifie qu'il y a **autant d'occurrences de deux valeurs**. Cela reflète une plus grande incertitude ou diversité.
- **Quand une proportion est dominante (par exemple, [1.0, 0.0]),** cela signifie qu'une seule valeur est présente. Cela reflète une faible incertitude ou diversité.

### **Exemple**

Supposons que les notes dans le groupe Faible soient réparties ainsi :

- Notes = [10,12,15,10,10].
- Les proportions : [0.6 (pour la note 10), 0.2 (pour la note 12), 0.2 (pour la note 15)].
- Entropie:  $H = - [ 0.6 \cdot \log_2(0.6) + 0.2 \cdot \log_2(0.2) + 0.2 \cdot \log_2(0.2) ]$

## Le graphique de notre cas pratique



### Résultat du test

Groupe "Faible" (Temps de travail  $\leq 2$ )

- **Entropie la plus élevée ( $\sim 2.58$ ) :**
  - Ce groupe présente la plus **forte variabilité** des notes.
  - Les étudiants dans ce groupe semblent avoir des performances très hétérogènes, allant des échecs à de bonnes notes.
  - Cela pourrait indiquer que **travailler peu n'aboutit pas nécessairement à des échecs** mais peut entraîner une grande incertitude dans les performances.

Groupe "Modéré" (Temps de travail = 3)

- **Entropie la plus faible bien que élevée ( $\sim 2.46$ ) :**
  - Ce groupe a les notes les **plus concentrées**, indiquant une meilleure homogénéité.
  - On ne peut pas dire que les étudiants dans ce groupe soient plus constants dans leurs performances que les deux autres groupes car l'entropie reste élevée.

Groupe "Intensif" (Temps de travail  $\geq 4$ )

- **Entropie élevée ( $\sim 2.48$ ) :**
  - La variabilité des notes reste présente, mais légèrement moins que dans le groupe "Faible" même si c'est légèrement plus que dans "Modérée".
  - Les étudiants qui travaillent intensément tendent à obtenir des performances un peu plus homogènes que ceux qui "ne travaillent pas".

## Conclusion

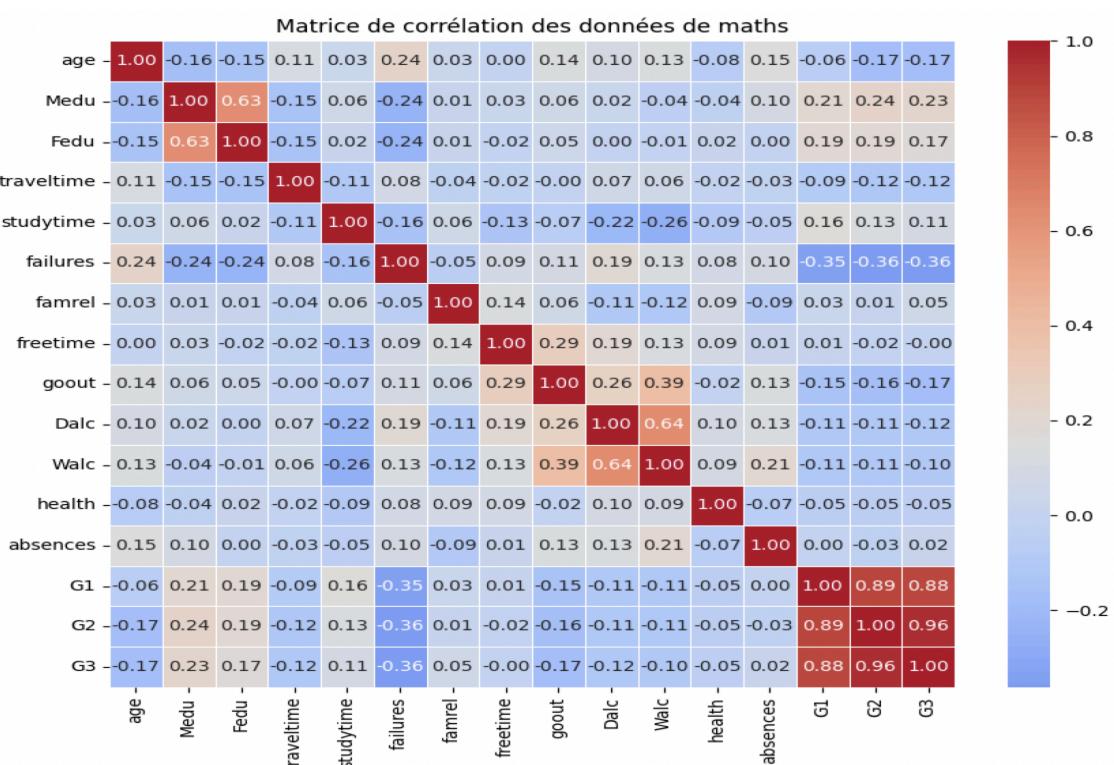
L'analyse de l'entropie appliquée à notre jeu de données a révélé certaines limitations qui remettent en question sa pertinence dans ce contexte. En effet, l'entropie repose sur une distribution équilibrée et un nombre suffisant d'observations pour fournir des résultats significatifs. Or, notre jeu de données présente un déséquilibre important dans la répartition des observations entre les groupes définis.

Par exemple, le groupe "**Faible**" compte environ 300 notes, tandis que le groupe "**Modéré**" n'en possède que 60 et le groupe "**Intensif**", seulement 30. Cette disparité entraîne une estimation biaisée de l'entropie, les effectifs insuffisants dans certaines catégories ne permettant pas de capturer correctement la diversité des notes. En conséquence, les calculs d'entropie risquent de ne pas refléter fidèlement la variabilité des résultats et de conduire à des conclusions erronées.

Ainsi, en l'état actuel des données, l'utilisation de l'entropie ne semble pas adaptée. Une analyse plus pertinente nécessiterait un échantillon plus équilibré et de taille plus importante afin de garantir des estimations fiables et interprétables.

### 3.5 - Résultats des corrélations et interprétations

#### 3.5.1 - Matrices de corrélation pour l'ensemble



## Analyse de la matrice de corrélation des données de mathématiques

**Corrélation entre les notes G1, G2 et G3:** Les corrélations élevées (G1-G2 : 0.89, G1-G3 : 0.88, G2-G3 : 0.96) montrent une forte stabilité des performances scolaires. Cela suggère que les évaluations intermédiaires sont de bons prédicteurs des résultats finaux.

**Éducation parentale et performances scolaires:** Une corrélation modérée est observée entre le niveau d'éducation des parents et les notes finales (Medu-G3 : 0.21, Fedu-G3 : 0.23), ce qui suggère une influence positive, bien que non déterminante, de l'environnement familial.

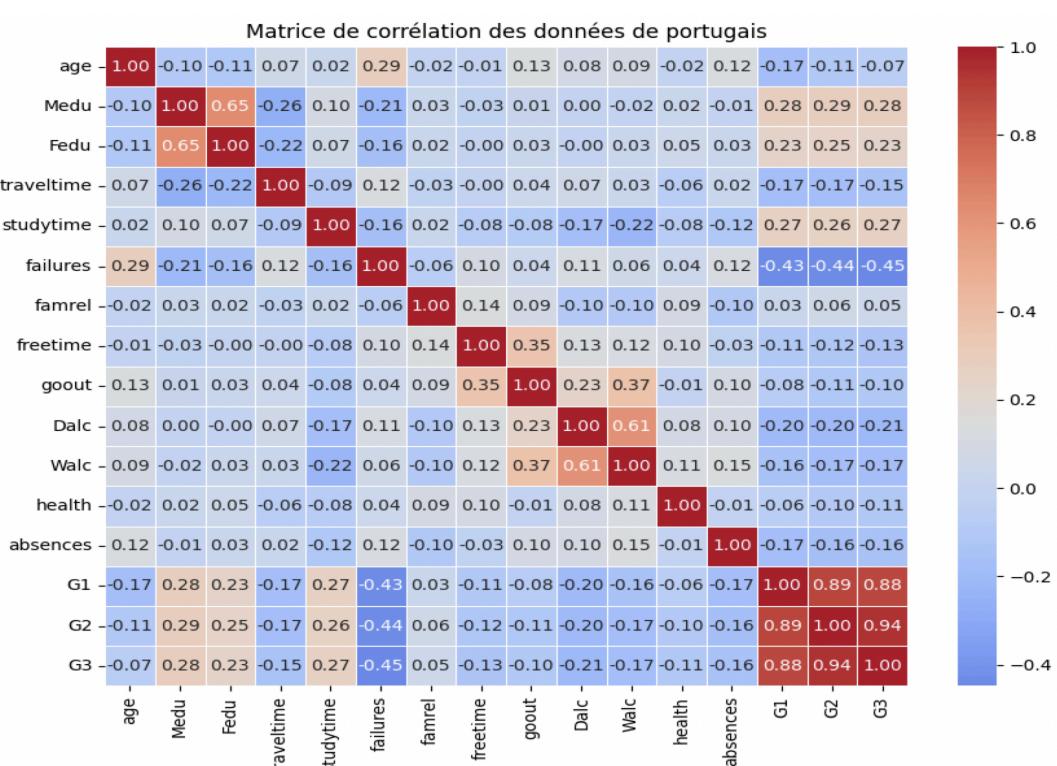
**Impact des échecs scolaires:** Les corrélations négatives avec les notes finales (failures-G1 : -0.35, failures-G2 : -0.36, failures-G3 : -0.36) confirment que l'accumulation d'échecs passés affecte significativement les résultats futurs, soulignant l'importance de la prévention et du soutien scolaire.

**Consommation d'alcool et performances:** Une forte corrélation (0.64) est constatée entre la consommation d'alcool en semaine et le week-end, montrant une constance dans les habitudes. Cependant, son impact sur les résultats reste légèrement négatif, indiquant une possible influence sur les performances.

**Temps d'étude et efficacité:** Les corrélations légèrement négatives (studytime-G3 : -0.26) suggèrent que le volume d'étude seul n'est pas un facteur clé de réussite, mettant en avant l'importance de la qualité de l'apprentissage.

## Conclusion

Les résultats mettent en évidence que la réussite scolaire repose principalement sur la constance des performances, l'historique des échecs et, dans une moindre mesure, le niveau d'éducation des parents. Une attention particulière doit être accordée aux élèves en difficulté pour éviter l'accumulation d'échecs et améliorer l'efficacité du temps d'étude.



## **Analyse de la matrice de corrélation des données de portugais**

**Performance académique stable :** Les corrélations élevées entre G1, G2 et G3 (**0.89, 0.88, 0.94**) montrent une forte cohérence des performances scolaires, indiquant que les notes intermédiaires sont de bons prédicteurs des résultats finaux.

**Impact de l'éducation parentale :** L'éducation des parents (Medu et Fedu, corrélation de **0.65**) influence modérément les résultats scolaires (corrélation avec G3 de **0.28** et **0.23**), suggérant que l'environnement familial peut jouer un rôle positif mais non déterminant.

**Échecs scolaires et résultats finaux :** Une corrélation négative marquée entre le nombre d'échecs scolaires et les notes finales (G3 : **-0.45**) montre que les difficultés antérieures impactent significativement la réussite.

## **Points d'attention**

**Consommation d'alcool :** La corrélation entre la consommation d'alcool en semaine et le week-end (**0.61**) est forte, mais son impact négatif sur les résultats scolaires reste modéré (-0.20 à -0.21), suggérant que les élèves concernés pourraient maintenir un équilibre entre études et loisirs.

**Temps d'étude inefficace :** La corrélation négative entre le temps d'étude et les notes finales (G3 : **-0.15**) suggère que les efforts en termes de quantité d'étude ne se traduisent pas forcément par une amélioration des résultats, mettant en avant l'importance des méthodes d'apprentissage.

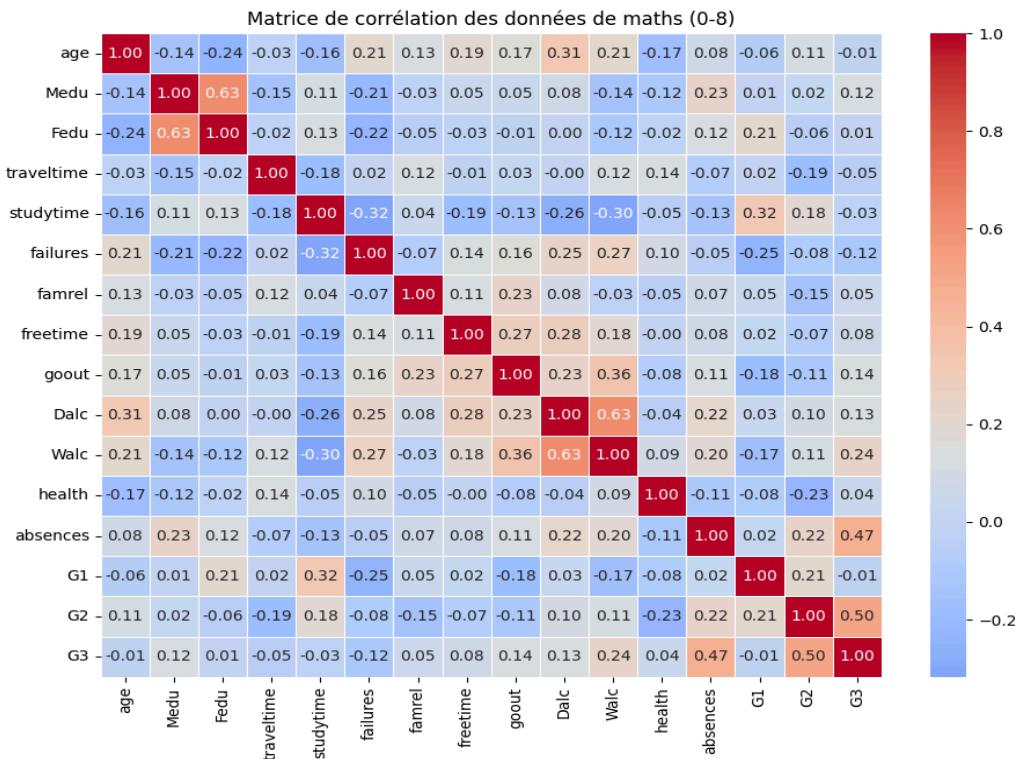
## **Conclusion**

Les résultats scolaires sont fortement influencés par les performances antérieures et l'éducation parentale, tandis que les échecs passés restent un frein majeur. Une attention particulière aux méthodes d'étude et un accompagnement personnalisé pourraient être bénéfiques pour ces élèves.

### **3.5.2 - Matrices de corrélation par groupe**

L'analyse des matrices de corrélation générales offre une vue d'ensemble des relations entre les différentes variables, mais elle peut masquer des disparités importantes entre les étudiants selon leur niveau de performance. En effet, les facteurs influençant les résultats scolaires ne s'appliquent pas de manière uniforme à tous les étudiants. Par exemple, l'impact de l'absentéisme ou du temps d'étude peut différer entre ceux en difficulté et ceux ayant de bons résultats.

Segmenter les élèves en plusieurs groupes permet d'affiner l'analyse en identifiant des tendances et corrélations spécifiques à chaque groupe. Ainsi, nous avons séparé les étudiants en trois groupes distincts selon leurs notes (G3).



### Analyse de la matrice de corrélation (G3 <= 8)

**Corrélations entre les notes G1, G2 et G3:** La corrélation entre G2 et G3 est modérée (0.50), indiquant que les résultats intermédiaires influencent en partie les notes finales. En revanche, G1 ne présente presque aucune corrélation avec G3 (-0.01), suggérant que les performances initiales ne sont pas un bon indicateur des résultats finaux.

### Absentéisme et performances scolaires:

Une corrélation de 0.47 entre les absences et les notes finales montre que l'absentéisme est un facteur clé dans les faibles résultats scolaires. Une meilleure assiduité pourrait améliorer les performances.

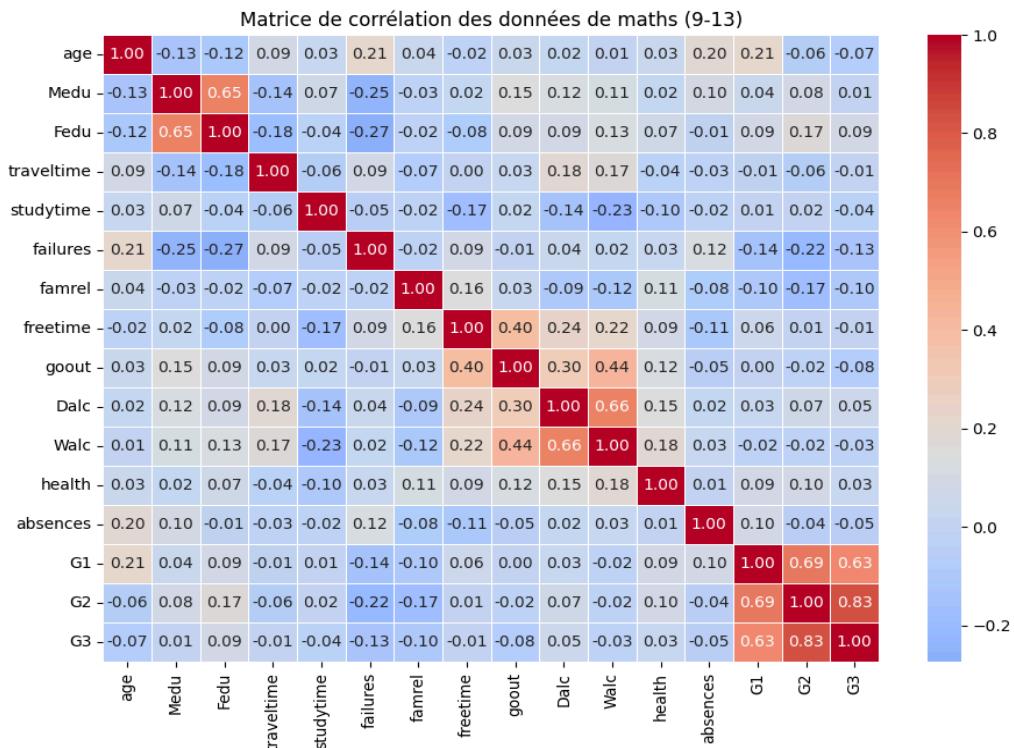
**Consommation d'alcool:** Une forte corrélation (0.63) entre la consommation d'alcool en semaine et le week-end suggère des habitudes de consommation régulières, pouvant potentiellement affecter la concentration et le rendement scolaire.

**Temps d'étude et performances:** Des corrélations négatives (-0.26 à -0.30) entre le temps d'étude et les notes indiquent que les élèves en difficulté étudient davantage sans pour autant améliorer leurs résultats, suggérant un besoin de méthodes d'apprentissage plus efficaces.

**Niveau d'éducation des parents:** Avec des corrélations faibles (entre 0.01 et 0.12), l'impact de l'éducation parentale sur les résultats des élèves en difficulté semble limité, mettant en évidence l'importance de facteurs tels que la motivation personnelle et le soutien externe.

## Conclusion

L'absentéisme et les mauvaises stratégies d'étude apparaissent comme les principaux facteurs contribuant aux faibles performances. Des actions ciblées, telles que l'amélioration de l'encadrement pédagogique et la sensibilisation à l'importance de l'assiduité, sont essentielles pour aider ces élèves.



## Analyse de la matrice de corrélation des données de portugais (13 >= G3 >= 9)

**Corrélation entre les notes G1, G2 et G3:** Les corrélations sont relativement fortes : G1-G2 (0.69), G1-G3 (0.63), G2-G3 (0.83). Cela indique une continuité des performances scolaires, suggérant que les résultats intermédiaires sont de bons indicateurs des résultats finaux pour cette tranche d'élèves.

**Consommation d'alcool (Dalc et Walc):** Une corrélation élevée (0.66) entre la consommation d'alcool en semaine et le week-end montre des habitudes de consommation cohérentes. Cependant, l'impact direct sur les résultats scolaires est faible, ce qui suggère que ces élèves parviennent à gérer leur consommation sans compromettre leur rendement académique.

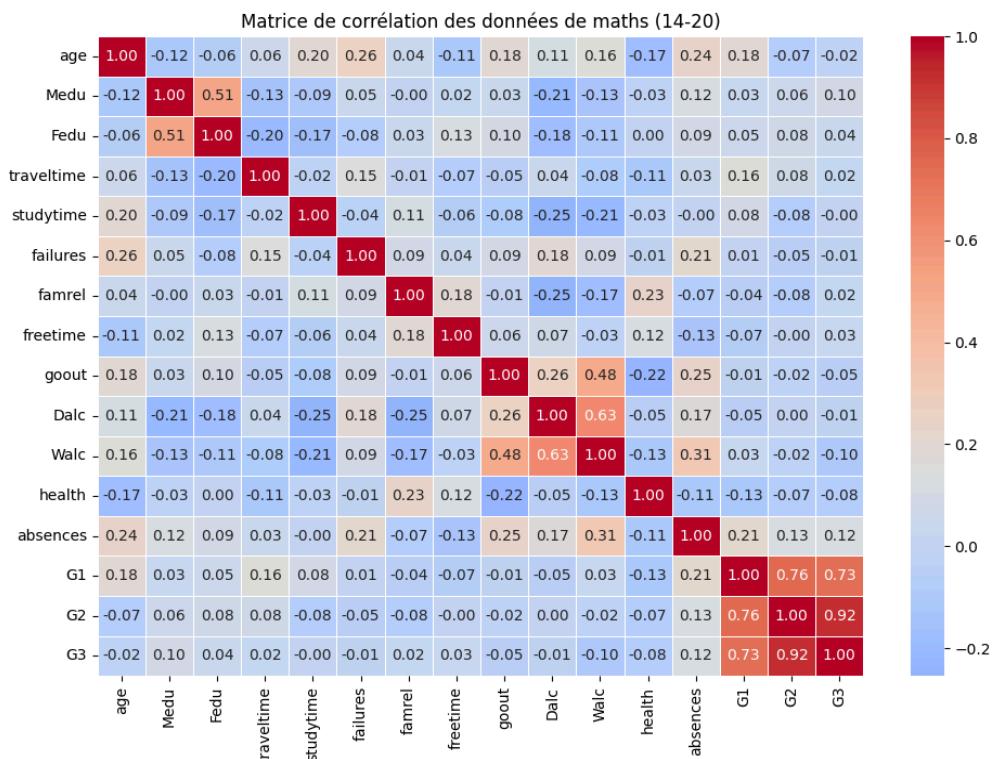
**Échecs scolaires et performances académiques:** Les corrélations négatives entre le nombre d'échecs et les notes sont modérées (jusqu'à -0.22 pour G2), indiquant que ces élèves parviennent à surmonter leurs échecs mieux que les élèves en difficulté.

**Temps d'étude et résultats:** Des corrélations négatives (jusqu'à -0.23) suggèrent que l'augmentation du temps d'étude ne se traduit pas nécessairement par de meilleures performances, mettant en avant l'importance de la qualité des méthodes d'apprentissage.

**Influence parentale limitée:** Le niveau d'éducation des parents présente des corrélations faibles avec les notes finales (Medu : 0.01, Fedu : 0.09), indiquant que dans cette tranche, l'influence familiale est moins marquée sur les résultats.

## Conclusion

Les résultats des élèves dans cette tranche de notes sont relativement stables au fil du temps, suggérant qu'ils maintiennent leurs performances sans subir d'impact majeur de facteurs externes comme l'absentéisme ou l'environnement familial. L'optimisation des méthodes d'apprentissage et de gestion du temps pourrait être une piste plus efficace pour améliorer leur progression académique.



## Analyse de la matrice de corrélation des données de portugais (G3 >= 14)

**Corrélation entre les notes G1, G2 et G3:** Les corrélations élevées entre les notes intermédiaires et finales (G1-G2 : 0.76, G1-G3 : 0.73, G2-G3 : 0.92) indiquent une forte stabilité des performances académiques. Cela confirme que les résultats intermédiaires sont d'excellents prédicteurs des résultats finaux pour ces élèves performants.

**Influence parentale:** Le niveau d'éducation des parents est modérément corrélé entre eux (0.51), mais présente une faible influence sur les résultats des élèves (corrélation de 0.10 et 0.04 avec G3). Ces élèves réussissent donc principalement grâce à leurs propres efforts et stratégies.

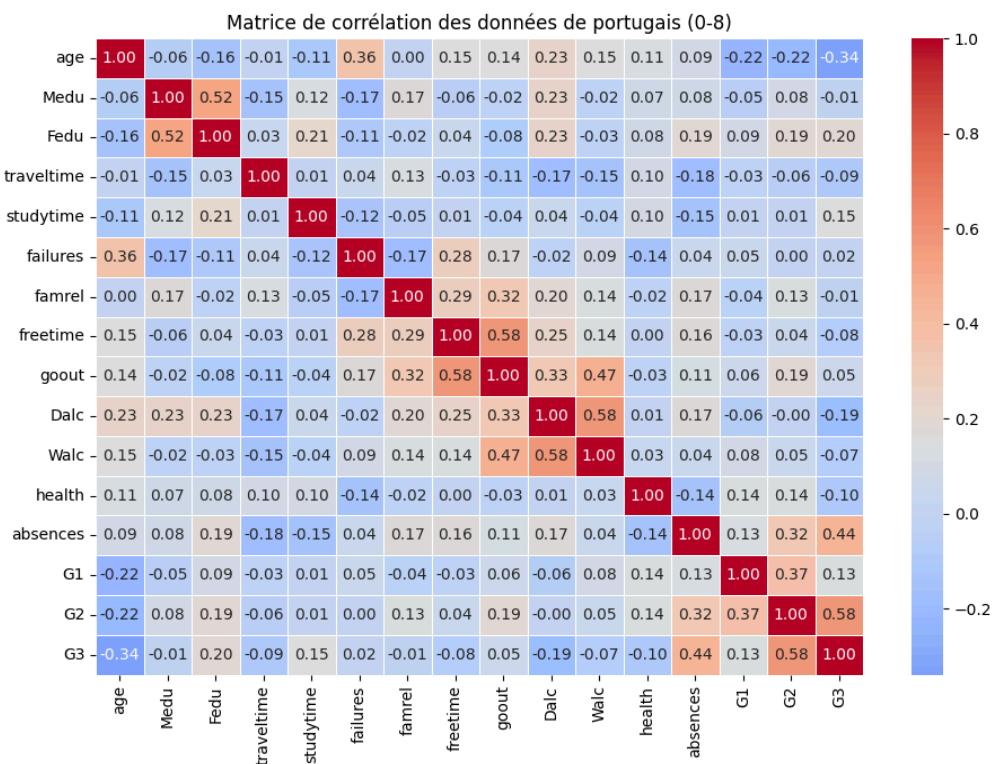
**Absences et résultats scolaires:** Contrairement aux élèves en difficulté, les absences ont une légère corrélation positive avec les notes (0.12), suggérant que ces élèves peuvent gérer leur emploi du temps efficacement et compenser leurs absences.

**Temps d'étude et performances:** La corrélation négative avec G3 (-0.08) indique que ces élèves n'ont pas besoin de consacrer un temps excessif à l'étude pour obtenir de bons résultats, ce qui souligne l'importance de méthodes d'apprentissage efficaces plutôt que la quantité de travail.

**Consommation d'alcool:** Une forte corrélation entre la consommation d'alcool en semaine et le week-end (0.63) est observée, mais son impact sur les résultats est négligeable, montrant que ces élèves parviennent à concilier vie sociale et études sans affecter leurs performances.

## Conclusion

Les élèves ayant des notes élevées présentent une progression stable et prévisible tout au long de l'année, avec une autonomie et une capacité d'organisation importantes. Leur réussite est principalement due à des facteurs internes tels que la discipline personnelle et l'efficacité des stratégies d'apprentissage, plutôt qu'à des influences externes comme l'environnement familial ou le temps d'étude.



## Analyse de la matrice de corrélation des données de portugais (G3 <= 8)

**Corrélation entre les notes G1, G2 et G3:** Les corrélations entre G1 et G2 (0.37) ainsi que G2 et G3 (0.58) indiquent une relation modérée, suggérant une certaine stabilité des performances au fil des évaluations. En revanche, la faible corrélation entre G1 et G3 (0.13)

souligne que les premières évaluations ne permettent pas de prédire avec précision les résultats finaux.

**Absences et performances académiques:** Une corrélation modérée (**0.44**) avec G3 met en évidence l'impact négatif de l'absentéisme sur les résultats scolaires, suggérant que la présence régulière en classe est un levier d'amélioration important.

**Âge et résultats scolaires:** Une corrélation négative de **-0.34** montre que les élèves plus âgés obtiennent des résultats inférieurs, ce qui pourrait être lié à des retards scolaires ou un décrochage progressif.

### Points à retenir

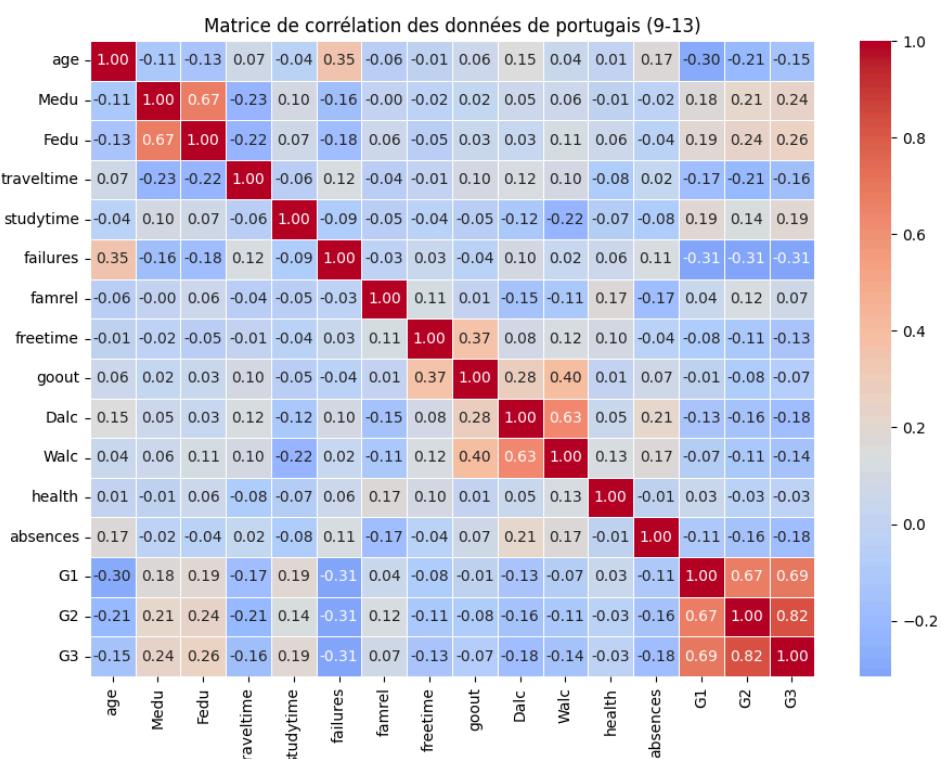
**L'influence parentale est limitée**, avec des corrélations faibles entre le niveau d'éducation des parents et les notes finales, indiquant que les performances des élèves en difficulté sont davantage influencées par des facteurs individuels et scolaires.

**Le temps d'étude montre une faible relation avec les résultats**, suggérant que la qualité de l'apprentissage est plus déterminante que la quantité d'heures étudiées.

### Conclusion

Les résultats soulignent l'importance de lutter contre l'absentéisme et d'apporter un soutien pédagogique adapté aux élèves plus âgés. L'influence des parents étant limitée, des interventions spécifiques en milieu scolaire apparaissent comme la meilleure solution pour améliorer les performances des élèves en difficulté.

### Analyse de la matrice de corrélation des données de portugais (13 >= G3 >= 9)



**Notes intermédiaires et résultats finaux:** Les corrélations entre G1 et G2 (**0.67**) et entre G2 et G3 (**0.82**) indiquent une relation relativement forte, confirmant que les notes intermédiaires sont de bons indicateurs des performances finales. Toutefois, la corrélation entre G1 et G3 (**0.69**) suggère des fluctuations potentielles dans les résultats.

**Impact des échecs scolaires:** La corrélation négative entre les échecs passés et les notes finales (G3 : **-0.31**) montre que les antécédents scolaires influencent négativement la progression des élèves, nécessitant un accompagnement spécifique pour éviter la stagnation.

**Âge et résultats scolaires:** Une corrélation négative avec G3 (**-0.15**) suggère que les élèves plus âgés ont tendance à obtenir des résultats légèrement inférieurs, potentiellement en raison d'un retard académique ou de difficultés d'adaptation.

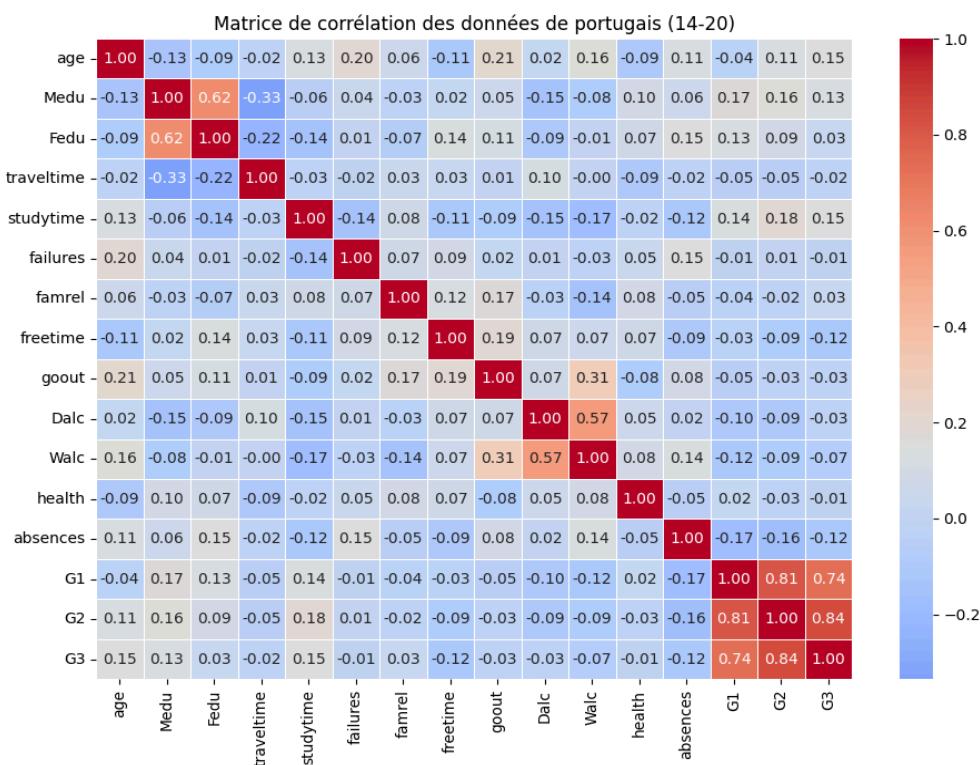
#### Points d'attention :

**Absences et performances:** La corrélation modérée entre les absences et G3 (**-0.18**) indique que l'assiduité a un impact limité sur cette tranche d'élèves, contrairement aux étudiants en difficulté.

**Consommation d'alcool:** Une forte corrélation entre la consommation en semaine et le week-end (**0.63**) montre des habitudes cohérentes, mais l'impact négatif sur les notes reste faible (**-0.18**), suggérant un contrôle relatif de cette consommation.

#### Interprétations

Les performances des élèves de ce groupe sont influencées principalement par les résultats intermédiaires et les antécédents scolaires. Des actions ciblées pour éviter l'accumulation d'échecs et un suivi personnalisé pourraient contribuer à l'amélioration de leurs résultats.



## **Analyse de la matrice de corrélation des données de portugais (G3 >= 14)**

**Cohérence des performances scolaires:** Les corrélations entre G1, G2 et G3 sont fortes (**0.81, 0.74, 0.84**), indiquant une stabilité académique des élèves tout au long de l'année. Les premières évaluations sont de bons prédicteurs des résultats finaux.

**Niveau d'éducation des parents:** Bien que Medu et Fedu soient corrélés entre eux (**0.62**), leur influence sur G3 reste faible (0.13 et 0.03), suggérant que ces élèves s'appuient davantage sur leur autonomie et leur motivation personnelle.

**Absences et performances scolaires:** Une corrélation légèrement négative avec G3 (**-0.12**) indique que les absences ont un impact limité sur ces élèves, qui parviennent à maintenir de bons résultats malgré des absences occasionnelles.

## **Interprétations**

Les résultats montrent que ces étudiants sont caractérisés par une grande régularité dans leurs performances et une bonne capacité à concilier études et vie personnelle. L'absentéisme et le temps d'étude n'ont pas d'effet significatif, indiquant une gestion efficace du travail et une autonomie renforcée.

## **4 - Conclusion finale**

L'analyse des données sur la consommation d'alcool des étudiants et ses corrélations avec les performances académiques a révélé plusieurs tendances intéressantes. Tout d'abord, une relation négative entre la consommation d'alcool et les résultats scolaires a été mise en évidence, soulignant l'impact potentiel de certaines habitudes de vie sur la réussite académique. Par ailleurs, des facteurs tels que le soutien familial et les caractéristiques sociodémographiques jouent un rôle clé dans les performances des étudiants.

Les tests statistiques, notamment le test du Chi2 et l'analyse d'entropie, ont permis de mieux comprendre les liens entre les différentes variables étudiées, bien que certaines limitations, comme la taille de l'échantillon et la distribution des données, aient nécessité des ajustements pour assurer la validité des résultats.

Ces conclusions mettent en lumière l'importance de prendre en compte de multiples facteurs dans l'analyse des performances scolaires, ce qui nous sera utile dans la poursuite de l'étude à travers la réalisation du questionnaire.

## IV. Collecte des données liées aux jeux vidéo

### 1 - Outil utilisé

#### 1.1 - Choix de l'outil

Pour collecter des données sur le temps consacré aux jeux vidéo et les performances scolaires, l'idéal aurait été de réaliser des entretiens directs avec notre échantillon. Cependant, pour des raisons pratiques, cette approche n'était pas envisageable. Nous avons donc opté pour la création d'un questionnaire Google Form composé de 45 questions.

Ce questionnaire couvre plusieurs thématiques : des données sociodémographiques, des informations sur la répartition des activités quotidiennes, des données spécifiques à la consommation de jeux vidéo, ainsi que des éléments relatifs à l'environnement de travail et aux performances scolaires.

#### 1.2 - Choix du type de questions

La majorité des questions du questionnaire ont été rendues obligatoires afin de garantir un volume suffisant de réponses pour nos analyses.

Le questionnaire intègre différents types de questions :

- **Questions à choix unique et multiple** : Ces formats ont été privilégiés pour harmoniser les réponses et réduire la nécessité de nettoyer les données.
- **Champs de réponse courts** : Utilisés de manière limitée, ils permettent aux répondants de fournir des informations précises lorsqu'il est difficile de prévoir toutes les réponses possibles, comme pour indiquer leur âge.
- **Listes déroulantes** : Ces options ergonomiques sont particulièrement utiles lorsqu'il existe de nombreuses réponses possibles, mais qu'il est nécessaire de limiter les choix, par exemple pour sélectionner un domaine d'étude.
- **Échelles linéaires** : Ces questions permettent une représentation claire et graduée de certaines fréquences ou intensités, comme le temps consacré à une activité ou la fréquence des retards.
- **Questions binaires** : De type "Oui/Non", elles constituent une forme simple et efficace de questions à choix unique pour recueillir des réponses directes.

Ce choix diversifié de formats a été réfléchi pour maximiser la clarté, l'ergonomie et la qualité des données collectées, tout en facilitant leur analyse

#### 1.3 - Items du questionnaire

Pour savoir quelles questions seraient pertinentes dans le cadre de notre étude, nous nous sommes basés sur la littérature scientifique puisque l'impact des jeux vidéo sur les performances scolaires n'est pas un thème nouveau.

Mais aussi sur les résultats d'analyse de l'étude *Student Alcohol Consumption*. Cette analyse aura permis d'écartier certains items n'ayant pas de lien (corrélation faible) avec les performances scolaires. Par exemple, la taille de la famille ("famsize") ou le fait d'être dans une relation de couple ("romantic") ne se sont pas révélés comme influence des performances scolaires. En revanche, certains items se sont imposés comme la **volonté de poursuivre ses études** ("higher"). Nous avons donc repris ces items dans notre questionnaire afin de voir si nous obtenons des résultats similaires.

### 1.2.1 - Items sociodémographiques

Nous avons choisi de recueillir des informations socio-démographiques afin de diviser l'échantillon en plusieurs groupes dont certains pourraient se révélés pertinents à l'analyse, notamment le **genre** ou le **niveau d'étude des parents**. En effet, la littérature indique une différence de performances scolaires selon les genre depuis le plus jeune âge mais aussi une différence dans la consommation de jeux vidéos bien que cette différence tende à s'atténuer plus récemment. Par ailleurs, les parents les plus éduqués tendent à être plus invertis dans l'éducation de leurs enfants (Stevenson & Baker, 1997) et une corrélation positive s'observe entre le niveau d'éducation des parents et les performances académiques des enfants (Jimerson, Egeland & Teo, 1999).

### 1.2.2 - Items liés à la performance scolaire

Pour évaluer les performances scolaires, nous avons recueilli des mesures quantitatives telles que les notes qui traduisent la performance directe. Nous avons demandé la **moyenne générale**, ainsi que la **moyenne en anglais**. Nous avons aussi souhaité recueillir d'autres indicateurs moins directs que les notes : les **retards**, les **absences**, le **nombre de matières non validées**, la **fréquence de devoirs rendus en retard ou incomplets** ainsi qu'une évaluation subjective de leurs performances scolaires.

### 1.2.3 - Items liés à l'environnement scolaire

Pour nuancer l'impact des jeux vidéo sur la performance académique, nous avons posé des questions liées à ce qu'on pourrait appeler "environnement de travail" ou "facteurs externes" : **l'accès au ressources** pour travailler, le bénéfice ou non de **soutien scolaire**, la **pression des parents** quant aux performances scolaires ou encore le **temps de trajet** pour se rendre à l'établissement.

Nous avons également regardé l'attitude des individus envers les études : **temps de travail**, **volonté de poursuite d'études**.

### 1.2.4 - Items liés à la consommation de jeux vidéo

Ensuite, nous nous sommes penchés sur la meilleure manière d'**évaluer la consommation de jeux vidéo**.

Incontestablement, nous avons demandé le **temps passé sur les jeux vidéo** en semaine et le week-end (également pour reproduire le type de question de *Student Alcohol Consumption* "Weekly Alcool" et "Daily Alcool" et obtenir des comparaisons précises avec

les résultats obtenus par cette étude), le **type de jeu vidéo**, l'**impact du temps de jeu sur le temps de sommeil**.

Nous avons également consulté la littérature pour formuler de nouvelles questions, par exemple sur les **moments de la journée dédiés aux jeux** vidéo. En effet, selon Aaron Drummond et James D. Sauer (2020), jouer avant l'école serait associé à des performances scolaires plus faibles. Nous avons également demandé **la langue dans laquelle les individus jouent**, cela permettrait de voir si le fait de jouer en anglais améliore les performances en anglais (également d'où le recueil des performances scolaires en anglais) comme suggèrent M. Dhany Winaldo et Lulud Oktaviani (2022).

### 1.2.5 - Items liés aux activités quotidiennes

Afin de contextualiser les réponses et de nuancer l'impact direct des jeux vidéo sur les performances scolaires, nous avons inclus des questions portant sur les **loisirs en général** et le **temps consacré à d'autres activités**, comme les tâches ménagères. Par exemple, certains étudiants vivant seuls doivent nécessairement accorder du temps à des activités telles que le ménage ou la cuisine, ce qui peut réduire le temps disponible pour les études et ainsi avoir un impact non négligeable sur leurs performances académiques.

De plus, deux questions spécifiques ont été consacrées à l'**utilisation des réseaux sociaux**, un passe-temps particulièrement répandu chez les jeunes dont il ne faudrait pas négliger le possible impact sur les performances scolaires lors de notre analyse .

### 1.3 - Rédaction et présentation du questionnaire

De manière globale les items ont été rédigés de manière à ce que les personnes interrogées puissent y répondre le plus facilement possible et que chacun puisse se reconnaître dans une des réponses. Nous avons essayé d'utiliser des échelles précises (en indiquant des quantités, des heures) ou tout du moins plus précises que celles utilisées dans *Student Alcohol Consumption* qui étaient des échelles linéaires (ex : Très faible ... Très haut) qui semblent extrêmement subjectives.

Combien de temps par jour jouez-vous aux jeux vidéo **en week-end** ? \*

- Je ne joue pas
- 1-2 heures
- 2-3 heures
- 4-6 heures
- + 7 heures

figure 1 : exemple d'échelle utilisé dans notre questionnaire

27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

## figure 2 : exemple d'échelle utilisée par Student Alcohol Consumption

Par ailleurs, nous souhaitions éviter que les personnes interrogées identifient précisément l'objectif de notre étude, ce qui aurait pu les amener à ajuster leurs réponses pour donner une image plus favorable ou conforme, risquant ainsi de biaiser les résultats. C'est pourquoi le questionnaire a été présenté comme portant sur "les habitudes de vie étudiante". Certains items, pertinents pour notre sujet principal, ont été ajoutés pour diversifier les thèmes abordés, notamment des questions sur l'utilisation des réseaux sociaux. Cela permettait de détourner l'attention des répondants du lien spécifique entre jeux vidéo et performance scolaire.

## 2 - Echantillon

Pour constituer notre échantillon, nous avions initialement envisagé d'interroger des lycéens, afin de reproduire la méthodologie de l'étude *Student Alcohol Consumption*, qui avait recueilli des données auprès de lycéens. Cependant, en tant qu'étudiants nous-mêmes, il nous était plus simple d'accéder à un public étudiant. De plus, il nous est apparu complexe d'interroger simultanément lycéens et étudiants avec le même questionnaire, car les systèmes d'évaluation diffèrent entre le lycée et l'enseignement supérieur, ce qui aurait compromis la cohérence des données. Par souci de praticité, nous avons donc choisi de nous concentrer uniquement sur les étudiants.

L'échantillon était initialement composé de 287 sujets mais après nettoyage des données, il a été restreint à 269 sujets.

## 3 - Tests préliminaires et ajustements du questionnaire

Avant le lancement de notre étude, nous avons réalisé des tests auprès de quelques camarades pour recueillir des avis extérieurs. Plusieurs critiques ont émergé : le questionnaire était jugé trop long, intrusif (notamment sur les aspects scolaires et familiaux), et répétitif, comme avec les questions "*Combien de temps par jour consaciez-vous à tous vos loisirs en semaine ?*" et "*Combien de temps par jour consaciez-vous à tous vos loisirs le week-end ?*".

Initialement composé d'environ 70 questions, le questionnaire comprenait une section sur l'alcool, inspirée de l'étude *Student Alcohol Consumption*, pour comparer ses effets sur les performances scolaires. Cependant, notre objectif principal étant d'étudier l'impact des jeux vidéo, nous avons décidé de supprimer cette section, de réduire celle sur les réseaux sociaux et d'éliminer des questions jugées redondantes ou peu pertinentes.

Certaines questions ont également été reformulées pour limiter leur caractère intrusif. Par exemple, "*Avez-vous un handicap pouvant influencer vos activités scolaires ou votre apprentissage ?*" est devenu "*Avez-vous rencontré des problèmes de santé ou des limitations physiques qui ont influencé vos activités scolaires ou votre apprentissage ?*". Une option "*Préfère ne pas répondre*" a également été ajoutée à certaines questions pour offrir plus de confort aux répondants.

Malgré ces ajustements, une certaine répétitivité demeure, notamment dans les questions visant à comparer la répartition des activités (comme les tâches ménagères ou les jeux vidéo) entre la semaine et le week-end. Ces éléments ont été conservés pour répondre aux objectifs spécifiques de l'étude.

## 4 - Limites du questionnaire et de l'échantillon interrogé

Plusieurs limites ont été identifiées concernant notre questionnaire et l'échantillon étudié :

### **Sur-représentation de certains domaines d'études et diversité des domaines :**

Une proportion importante des personnes interrogées provient du domaine de l'informatique. Cela peut introduire un biais. De manière générale, le fait d'interroger des personnes de différentes filières peut aussi rendre difficile l'analyse des résultats car les exigences académiques varient selon les filières. Par exemple, pour les étudiants en langues ou en communication, la maîtrise de l'anglais est souvent cruciale, ce qui peut les inciter à réviser davantage et à obtenir de meilleures notes dans cette matière, comparé à des filières où les langues étrangères sont moins valorisées.

### **Diversité des écoles et modalités d'évaluation :**

Les étudiants interrogés proviennent de différentes écoles et établissements, chacun ayant ses propres critères et modalités d'évaluation. Par exemple, un 10/20 dans une école peut correspondre à un 14/20 dans une autre, ce qui complique les comparaisons directes des performances scolaires.

### **Inadéquation pour certains profils :**

Le questionnaire n'est pas adapté à tous les étudiants. Certains, notamment ceux ayant des modalités de travail atypiques (comme l'absence d'horaires fixes, de notions de retard ou d'absence), peuvent avoir des difficultés à répondre à certaines questions de manière pertinente.

### **Conditions de passation non contrôlées :**

S'agissant d'un questionnaire en ligne, nous ne pouvions pas contrôler les conditions dans lesquelles les répondants ont rempli le formulaire. Une personne distraite, occupée ou interrompue pendant qu'elle répond peut fournir des réponses moins fiables, ce qui pourrait biaiser les résultats.

### **Interprétation subjective des questions :**

Bien que nous ayons formulé les questions avec soin pour qu'elles soient aussi claires que possible, certaines restent sujettes à interprétation. Par exemple, pour la question "*Quel est le niveau d'études de vos parents/figures parentales ?*", notre objectif était de connaître le niveau d'études des personnes ayant élevé le répondant. Cependant, certains répondants peuvent avoir été élevés par plusieurs personnes ou par des figures qu'ils ne considèrent ni comme parents ni comme figures parentales, ce qui peut influencer leur réponse.

## 5 - Axes d'amélioration

Plusieurs axes d'amélioration ont été identifiés :

### **Gestion des réponses “Autre” :**

Notre questionnaire permettait aux répondants de cocher une case "Autre" et de fournir une réponse libre. Cela a conduit à des réponses intéressantes, comme "caféine" ou "lecture" pour les addictions, ou encore "cuisine" et "YouTube" pour les loisirs.

Dans certains cas, nous avons pu reclasser ces réponses dans des catégories existantes (ex : "YouTube" a été intégré à la catégorie "Réseaux sociaux"). Cependant, pour d'autres réponses qui ne s'intégraient pas aux catégories préexistantes, nous avons envisagé de créer de nouvelles catégories. Cependant, certaines de ces nouvelles catégories auraient regroupé seulement un ou deux individus, ce qui aurait fragmenté les données en de trop nombreux petits groupes, rendant les analyses peu pertinentes. Pour cette raison, nous avons finalement regroupé toutes ces réponses dans une catégorie unique intitulée "Autre".

**Une solution pourrait être d'anticiper davantage de catégories** : En identifiant à l'avance un plus grand nombre de catégories pertinentes, nous aurions pu proposer des choix plus complets dès le départ. Cela aurait probablement permis à davantage de répondants de sélectionner ces options, constituant ainsi des groupes plus significatifs pour l'analyse. Autrement, nous aurions simplement pu **limiter les réponses libres** : proposer l'option "Autre" sans possibilité de réponse libre, de manière à regrouper directement toutes les réponses atypiques sous une seule catégorie.

Ces ajustements permettraient de mieux structurer les données et d'améliorer la qualité des analyses futures.

### **Structure du questionnaire :**

Notre questionnaire permettait aux répondants de fournir des réponses contradictoires en laissant toutes les questions accessibles à tous, indépendamment de leurs réponses précédentes. Par exemple, un individu pouvait répondre "Je ne joue pas aux jeux vidéo" à la question "Dans quelle langue jouez-vous aux jeux vidéo ?" puis indiquer "2-3 heures" à la question "Combien d'heures par jour jouez-vous aux jeux vidéo ?".

Cette conception du questionnaire a entraîné des incohérences dans les réponses, rendant certaines données inutilisables.

Une solution aurait été de structurer le questionnaire en sections conditionnelles. Par exemple :

1. Poser une question générale, comme "Jouez-vous aux jeux vidéo ?", avec des réponses "Oui" ou "Non".
2. Orienter les répondants vers des questions spécifiques sur les jeux vidéo uniquement s'ils répondent "Oui". Ceux ayant répondu "Non" auraient directement accédé aux questions sur d'autres thèmes, comme la scolarité.

Nous avions envisagé cette approche, mais nous avons choisi de ne pas l'adopter, par crainte que certaines personnes ne se considèrent pas comme des joueurs alors qu'elles le sont. Par exemple, des individus jouant à des petits jeux mobiles pourraient ne pas se reconnaître dans l'image stéréotypée du “gamer” et répondre automatiquement “Non”. Étant donné que les jeux vidéo constituaient le thème principal de notre étude, nous avons préféré laisser toutes les questions accessibles à tous, afin de ne pas exclure involontairement ces profils.

**Bilan de la réflexion rétrospective :**

Avec le recul, une meilleure structuration du questionnaire, incluant des sections conditionnelles, aurait probablement amélioré la cohérence des réponses et aurait permis de préserver la qualité des données recueillies.

En conclusion, nous avons conscience que le questionnaire n'aura pas su répondre à toutes les exigences de chacun et que toute personne y ayant répondu aura perçu les questions posées via sa propre expérience. Nous aurons tenté de faire au mieux pour que le questionnaire permette de récolter des données pertinentes et exploitables.

## V. Conception et construction de la bdd relationnelle Jeux vidéo

Cette section vise à expliquer le processus d'extraction et de structuration des données recueillies via notre questionnaire. En effet, les variables étant nombreuses avec parfois des multiplicités de types 1...n ou 0...n, il est alors essentiel de pouvoir garantir une compréhension des données pour expliquer leurs liens complexes. L'objectif de structuration des données est alors principalement de faciliter leur compréhension.

De plus, notre questionnaire ayant recueilli beaucoup de données redondantes (chaîne de caractère) à travers les différentes questions, une base de données permet d'éviter la répétition de données grâce à l'inclusion des dictionnaires référencés par des clés étrangères.

Enfin, le nettoyage de données effectué par notre équipe est aussi une étape cruciale puisqu'elle garantit la cohérence et la qualité des données en vue des analyses à effectuer. Enfin, il s'agit aussi de présenter des données correctes via notre interface graphique.

### 1 - Nettoyage de données

Les données obtenues sont organisées dans un document csv. Il y avait au départ 287 réponses et 269 après nettoyage.

Un nettoyage a été effectué dans 2 cas :

#### 1.1 - Nettoyage lié aux zones de réponses libres

Puisque notre questionnaire permettait aux répondants de saisir eux-mêmes une réponse textuelle pour certaines questions, il a fallu nettoyer et classer leurs réponses. Un nettoyage manuel a été effectué via des filtres pour sélectionner les réponses à trier pour les questions suivantes :

- Quels sont vos 3 principaux loisirs ?
- Quels sont les 2 types de jeux vidéo auxquels vous jouez le plus ?
- Dans quelle langue jouez-vous aux jeux vidéo ?
- Sur quel support jouez-vous le plus ?
- Vous considérez-vous addict à :
- Quelle est votre principale raison d'étudier dans votre établissement ?
- Si oui, quel niveau d'études souhaitez-vous atteindre ?
- Vous vivez :

Nous avons pensé à créer de nouvelles catégories pour les réponses manuelles des répondants. Par exemple, nous avons observé les réponses "Pâtisserie" et "Cuisiner" pour la question des loisirs, ces 2 réponses auraient pu être regroupées sous la nouvelle catégorie "Cuisine". De même, des réponses uniques auraient pu constituer une catégorie à elles seules mais cela n'aurait pas été pertinent au niveau des analyses puisqu'un groupe constitué de seulement quelques personnes ne montre pas grand chose. Ces réponses ont été regroupées sous la catégorie "Autre".

Seules certaines catégories pertinentes (proposées assez de fois) ont été reconnues par exemple :

- La catégorie "Téléphone" dans les addictions. Nous avons estimé que cette catégorie avait été citée un nombre de fois suffisant pour ne pas être classée dans "Autres" puisque 6 individus l'ont mentionnée comme addiction.
- La catégorie "Aucune" pour les raisons d'étudier dans son établissement scolaire. En effet, 9 individus ont mentionné dans les zones de texte libre n'avoir simplement pas eu le choix de leur établissement.

## 1.2 - Nettoyage de données incohérentes

Notre questionnaire laissait par exemple la possibilité aux répondants de répondre "Je ne joue pas aux jeux vidéo" à une question puis d'indiquer "2-3 heures" de temps de jeu par jour en semaine. Cela a causé une incohérence dans les réponses et engendre une non exploitabilité des données.

Lorsqu'il y avait une seule réponse inconsistante, la case a été vidée. Cela permet de ne pas prendre en compte la réponse incohérente dans les analyses mais de préserver les autres réponses du sujet.

Par exemple, si un individu a indiqué jouer aux jeux vidéo dans la majorité des questions mais qu'à une question, il a répondu "Je ne joue pas aux jeux vidéo", il peut être légitime de penser que le sujet a mal cliqué lorsqu'il a souhaité à la question ou bien qu'il a été distrait etc. Alors, la case incohérente avec le reste des réponses est vidée et sera "null" en base de données. Ainsi, les réponses exploitables du sujet pourront être utilisées lors de l'analyse et ce sujet ne sera pas pris en compte pour les analyses liées à la question vidée (ici, la langue dans laquelle on joue aux jeux vidéo)

À quel moment de la journée jouez-vous en :	Dans quelle langue jouez-vous aux jeux vidé :	Sur quel support jouez-vous le plus :	Combien de temps par jour jouez-vous aux j :
Le soir, après les cours	Je ne joue pas aux jeux vidéo	Ordinateur	1-2 heures
Le soir, après les cours		Ordinateur	1-2 heures

Ici, sur la 1ère ligne de données, on peut voir une donnée incohérente récoltée. La 2ème ligne présente la même observation mais la donnée incohérente a été supprimée.

En revanche, si trop d'incohérences dans les réponses ont été repérées, les réponses du sujet ont été entièrement supprimées car non exploitables.

## 2 - Modélisation de la bdd

### 2.1 - Les tables

De la même manière que pour la base de données de *Student Alcohol Consumption*, les données ont été organisées par thématiques :

- **Informations personnelles** : données socio-démographiques et de santé

- **Informations familiales** : situation familiale
- **Performances académiques** : indicateurs de la performance scolaire (notes, retard...)
- **Environnement de travail / facteurs externes** : informations pouvant influencer la performance scolaire (temps de travail, temps de trajet, soutien scolaire...)
- **Consommation de jeux vidéo** : indicateurs de la consommation (type de jeux, temps de jeu, impact de la consommation...)
- **Loisirs** : loisirs des individus et leurs répartition temporelle
- **Addiction** : potentielles addiction des sujets

**Table student (répondants) :**

- **gender\_id** : Genre de l'étudiant (référencé dans gender\_dict).
  - **age** : Âge de l'étudiant.
  - **student\_schooling\_level\_id** : Niveau d'études (référencé dans student\_schooling\_level\_dict).
  - **study\_field\_id** : Domaine d'étude (référencé dans study\_field\_dict).
  - **household\_type\_id** : Type d'habitation (référencé dans household\_type\_dict).
  - **health\_issue** : Indique si l'étudiant a des problèmes de santé impactant ses études (true/false). La table contient également les clés étrangères permettant de relier un étudiant aux autres tables (family, hobby, vg, academic\_perf, academic\_info).
  - **hobby\_id** : Identifiant des loisirs de l'étudiant (référencé dans hobby).
  - **vg\_id** : Identifiant des informations sur les jeux vidéo de l'étudiant (référencé dans vg).
- La table contient également les clés étrangères permettant de relier un étudiant aux autres tables (family, hobby, vg, academic\_perf, academic\_info).

**Table family (informations familiales) :**

- **parent1\_schooling\_level\_id** : Niveau d'études du premier parent (référencé dans parent\_schooling\_dict).
- **parent2\_schooling\_level\_id** : Niveau d'études du deuxième parent (référencé dans parent\_schooling\_dict).
- **single\_parent** : Indique si la famille est monoparentale (true/false).
- **expectations** : Indique si les parents ont des attentes académiques envers l'étudiant (true/false).

**Table academic\_perf (performances académiques) :**

- **avg\_grade** : Moyenne générale (référencé dans grade\_dict).
- **english\_avg\_grade** : Moyenne en anglais (référencé dans grade\_dict).
- **late** : Fréquence des retards.
- **late\_incomplete\_homework** : Fréquence de devoirs rendus en retard ou incomplets.
- **absences** : Nombre total d'absences.
- **doubling** : Indique si l'étudiant a redoublé une classe (true/false).
- **perf\_evaluation** : Évaluation subjective des performances académiques.
- **classes\_under\_10** : Nombre de matières avec une moyenne inférieure à 10.

#### Table academic\_info (informations scolaires) :

- **school\_type\_id** : Type d'établissement scolaire (référencé dans school\_type\_dict).
- **reason\_id** : Raison d'étudier dans l'établissement (référencé dans reason\_dict).
- **fam\_support** : Soutien scolaire familial (true/false).
- **academic\_support** : Soutien scolaire de l'établissement (true/false).
- **private\_lessons** : Indique si l'étudiant suit des cours particuliers (true/false).
- **resources** : Accès à des ressources éducatives (true/false).
- **studytime\_id** : Temps hebdomadaire de révision (référencé dans studytime\_dict).
- **travel\_time\_id** : Temps de trajet domicile-école (référencé dans travel\_time\_dict).
- **further\_study** : Indique si l'étudiant souhaite poursuivre des études supérieures (true/false).
- **further\_study\_level\_id** : Niveau d'études envisagé (référencé dans further\_study\_level\_dict).

#### Table hobby (loisirs) :

- **weekdays\_hobby\_time\_id** : Temps consacré aux loisirs en semaine (référencé dans hobby\_time\_dict).
- **weekend\_hobby\_time\_id** : Temps consacré aux loisirs le week-end (référencé dans hobby\_time\_dict).
- **weekdays\_chores\_time\_id** : Temps consacré aux tâches ménagères en semaine (référencé dans chores\_sm\_time\_dict).
- **weekend\_chores\_time\_id** : Temps consacré aux tâches ménagères le week-end (référencé dans chores\_sm\_time\_dict).
- **hobby1\_id** : Loisir préféré (référencé dans hobby\_type\_dict).
- **hobby2\_id** : Deuxième loisir préféré (référencé dans hobby\_type\_dict).
- **hobby3\_id** : Troisième loisir préféré (référencé dans hobby\_type\_dict).
- **SM\_time\_id** : Temps consacré aux réseaux sociaux (référencé dans chores\_sm\_time\_dict).
- **sleep\_SM** : Impact des réseaux sociaux sur le sommeil (true/false).

#### Table vg (jeux vidéo) :

- **vg\_type1\_id** : Type de jeu préféré (référencé dans vg\_type\_dict).
- **vg\_type2\_id** : Second type de jeu préféré (référencé dans vg\_type\_dict).
- **violent** : Indique si l'étudiant joue à des jeux violents (true/false).
- **vg\_platform\_id** : Support principal de jeu (référencé dans vg\_platform\_dict).
- **weekdays\_VG\_time\_id** : Temps de jeu en semaine (référencé dans weekdays\_vg\_time\_dict).
- **weekend\_VG\_time\_id** : Temps de jeu le week-end (référencé dans weekend\_vg\_time\_dict).
- **focus\_VG** : Impact des jeux vidéo sur la concentration (true/false).
- **sleep\_VG** : Impact des jeux vidéo sur le sommeil (true/false).
- **gaming\_start\_period\_id** : Période de début des jeux vidéo (référencé dans gaming\_start\_period\_dict).

#### Table addiction (addictions) :

- **student\_id** : Identifiant de l'étudiant (référencé dans student).
- **addiction\_id** : Identifiant de l'addiction (référencé dans addiction\_dict). Un étudiant peut avoir plusieurs addictions, et une addiction peut concerner plusieurs étudiants.

## 2.2 - Les dictionnaires

Ces tables sont accompagnées de 22 dictionnaires contenant, par exemple, les valeurs liées au temps de révision, temps de trajet, type d'habitation, raison d'étudier, types de loisirs, types de jeux vidéo, langues des jeux, plateformes de jeu, etc.

## 2.3 - Les relations

Les entités sont reliées par des clés primaires/étrangères pour garantir l'intégrité référentielle.

Les relations suivantes sont de cardinalité **1,1 - 1,1** :

- Relation entre les tables student et academic\_info
- Relation entre les tables student et hobby
- Relation entre les tables student et vg
- Relation entre les tables student et academic\_perf
- Relation entre les tables student et family

Les relations suivantes sont de cardinalité **0,1 - 1,n** :

- Relation entre les tables vg et vg\_type\_dict : le type de jeux vidéo préféré numéro 2 peut être null et un type de jeux vidéo peut être renseigné plusieurs fois comme type de jeux préféré n°2
- Relation entre les tables hobby et hobby\_type\_dict : les hobby 2 et 3 peuvent être null et un type d'hobby peut être renseigné plusieurs fois comme hobby préféré 2 et 3.
- Relation entre academic\_info et further\_study\_level\_dict : Car il est possible qu'il n'y ait pas n'ait pas un niveau du supérieur.
- 

Les relations suivantes sont de cardinalité **1,1 - 1,n** :

- Relation entre les tables vg et vg\_type\_dict : il y a forcément un type de jeux vidéo préféré et un type de jeu peut être utilisé plusieurs fois en tant que type de jeu préféré.
- Relation entre les tables academic\_info et school\_type\_dict
- Relation entre les tables academic\_info et reason\_dict
- Relation entre les tables academic\_info et studytime\_dict
- Relation entre les tables academic\_perf et grade\_dict
- Relation entre les tables vg et vg\_platform\_dict
- Relation entre les tables vg et weekend\_vg\_time\_dict
- Relation entre les tables vg et weekdays\_vg\_time\_dict
- Relation entre les tables vg et gaming\_start\_period\_dict
- Relation entre les tables student et household\_type\_dict
- Relation entre les tables student et student\_schooling\_level\_dict
- Relation entre les tables student et study\_field\_dict
- Relation entre les tables student et gender\_dict
- Relation entre les tables family et parent\_schooling\_dict

- Relation entre les tables hobby et chores\_sm\_time\_dict

En effet, une valeur de dictionnaire peut être référencée 1 à n fois et une clé étrangère référence forcément une valeur de dictionnaire.

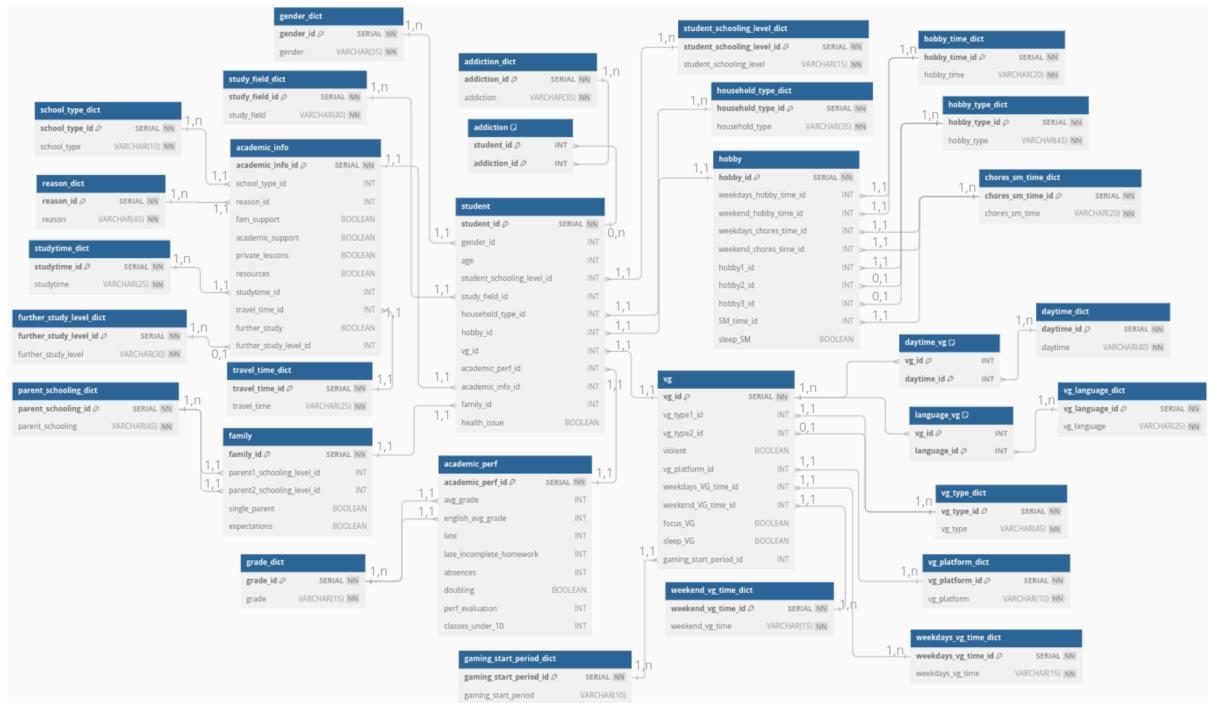
Les relations suivantes sont de cardinalité 0,n - 1,n :

- Relation entre les tables student et addiction : un étudiant peut avoir 0 à plusieurs addictions et une addiction peut être référencée 1 à n fois.

Les relations suivantes sont de cardinalité 1,n - 1,n :

- Relation entre les tables vg (video game/jeux vidéo) et vg\_language\_dict : les jeux sont forcément joué dans 1 langue mais ils peuvent également être joués dans plusieurs langues et une valeur du dictionnaire peut être référencée 1 à n fois.
  - Relation entre les tables vg et daytime\_dict : un étudiant (joueur) joue au moins à 1 moment de la journée ou bien joue à plusieurs moments tandis qu'un moment de la journée peut être référencé 1 à n fois.

## 2.4 - Le MCD



### 3 - Implémentation de la bdd

L'implémentation de la base de données a demandé plusieurs étapes :

### 3.1 - Cration du Schema

Les données concernant le Questionnaire ont été structurées dans phpMyAdmin via un fichier SQL qui a servi à créer les tables. Ces tables sont disposées de manière à maintenir des relations claires, et des contraintes telles que les clés primaires et les clés étrangères assurent l'intégrité des données. (cf. 2 - Modélisation de la base de données).

### 3.2 - Extraction, Transformation, Insertion des Données

Un script Python (ETL) a été conçu pour insérer les données nettoyées dans la base. Les étapes incluent :

La lecture du fichier CSV nettoyé.

La transformation des valeurs pour respecter les contraintes du modèle relationnel :

- Rendre insensible à la casse certaines colonnes qui récupèrent des réponses écrites par des utilisateurs
- Transformation de types de données : certaines questions récupèrent un texte court mais dans le MCD, ces mêmes colonnes sont représentées par un boolean. Ainsi des méthodes ont dû être mises en place pour permettre de convertir un string en bool. Par exemple, la question "Au cours des six derniers mois, avez-vous rencontré des problèmes de santé ou des limitations physiques qui ont influencé vos activités scolaires ou votre apprentissage ?" offrait 4 choix de réponse : 'oui, de manière modérée', 'oui, de manière significative', 'non' et 'Préfère ne pas répondre' mais cette donnée est finalement traitée comme un booléen par notre modèle.
- Renommage des données : dans notre modèle relationnel les variables ont des noms abstraits, cependant dans le notre fichier .csv, certaines colonnes comportent des questions ouvertes. La transformation se fait grâce à une méthode spécifique. Ce renommage a pour but d'éviter la répétition de chaînes de caractères trop longues dans le code et de simplifier sa lisibilité. Par exemple, la question 'quel est votre genre ?' issue du questionnaire devient 'quel\_genre'.
- Insertion des données dans les tables, en commençant par les dictionnaires, puis les tables principales. À chaque fois que ce programme est lancé une méthode effectue la vidange des tables afin de pouvoir déboguer plus facilement et ne pas rajouter des données aberrantes. Après cela, à chaque fois qu'un élément de dictionnaire est lu, s'il n'existe pas déjà, il est ajouté dans un dictionnaire grâce à la méthode `get_or_create` qui vérifie l'existence ou non d'une valeur de dictionnaire puis cette valeur est référencée dans la table nécessaire avec l'id adéquat.

Après avoir terminé de saisir les données sur phpMyAdmin, nous avons généré un fichier SQL à partir de phpMyAdmin, qui contient les données importées du Questionnaire, afin de préparer notre présentation finale.

### 3.3 - Choix de technologies

**Pour le script ‘Questionnaire\_ETL.py’ :** nous avons choisi Python pour sa polyvalence et la richesse de ses bibliothèques. Pandas nous a permis de lire et transformer facilement les données du fichier CSV, en simplifiant des tâches comme le renommage des colonnes ou la conversion des données pour respecter le modèle relationnel (par exemple, transformer des chaînes en booléens). Avec mysql.connector, nous avons pu exécuter des requêtes SQL pour insérer ou vérifier les données dans les bases de données MySQL de manière fluide et fiable. Cette combinaison d'outils a rendu notre processus à la fois efficace, lisible et robuste.

**Pour le stockage des données :** Nous avons travaillé avec PhpMyAdmin, car c'est une des solutions mise à disposition par l'IUT pour gérer les bases de données. Cet outil était déjà familier à notre équipe ainsi qu'à l'équipe Web, grâce à nos précédents projets communs. Il offrait une simplicité d'utilisation et une continuité dans nos pratiques, ce qui a facilité le travail de tout le monde.

# VI. Analyse de données Jeu vidéo

## 1 - Description des algorithmes et des techniques d'analyse utilisées

### 1.1 - Introduction

Pour analyser les données collectées par notre questionnaire, nous nous sommes appuyés sur les méthodes que nous avions utilisées lors de notre analyse des données Kaggle, comme présenté dans la précédente partie. Nous avons écarté les différentes méthodes qui se sont avérées infructueuses lors de l'analyse des données de Student Alcohol Consumption. Nous avons retenu trois outils principaux: la corrélation de Spearman, les tests de khi-deux et enfin différentes visualisations. Cette analyse a deux objectifs : d'une part vérifier certaines hypothèses ou limites identifiées lors de l'élaboration du questionnaire, et d'autre part, examiner le lien entre pratique du jeu vidéo et performances scolaires pour pouvoir répondre à la problématique.

### 1.2 - Corrélation de Spearman et premiers résultats

L'analyse a été réalisée sur un total de 21 variables, dont la majorité sont des variables ordinaires. Les variables ordinaires ont été numérisées et les valeurs représentant une absence de réponse ont été ignorées lors du calcul. Voici les résultats que nous avons trouvé intéressants.

#### **Temps consacré aux jeux-vidéos**

Les deux variables pour mesurer le temps passé sur les jeux vidéo sont weekdays\_vg\_time et weekend\_vg\_time. Ces variables représentent respectivement le temps moyen journalier que les répondants consacrent aux jeux vidéo en semaine et le week-end.

En semaine, les catégories sont : 'Je ne joue pas', '1-2 heures', '2-3 heures' et '+4 heures'. Le week-end, les catégories sont : 'Je ne joue pas', '1-2 heures', '2-3 heures', '4-6 heures' et '+7 heures'.

Nous avons une corrélation négative avec le temps d'études (-0,27) et une corrélation positive avec le fait de rendre en retard des devoirs (0,16).

Nous avons noté une corrélation positive avec le temps consacré aux loisirs, allant de 0,25 à 0,30.

En excluant les non joueurs, la corrélation avec la variable représentant la moyenne générale est nulle. (La raison d'exclusion vient du caractère ordinaire de la variable. Les valeurs de la variable ordinaire doivent être comparables entre elles; pour un non-joueur les valeurs comme '1-2 heures' n'ont pas de sens dans ce contexte)

## **Moyenne générale**

La variable avg\_grade représente la moyenne générale des répondants répartis en 5 catégories: '0 à 8', '8 à 10', '10 à 12', '12 à 16' et '16 à 20'.

Nous avons une corrélation positive avec le niveau d'étude visé (0,17). Cela peut suggérer que le fait d'avoir des objectifs académiques pourrait avoir un effet bénéfique.

Nous avons noté une corrélation négative avec le fait de rendre en retard des devoirs (-0.31), le nombre d'absences (-0.14) et le nombre d'absences (-0.14) ce qui ne semble pas être surprenant.

## **Autres observations**

Nous avons noté une corrélation positive entre le temps consacré aux loisirs le week-end et la moyenne en anglais (0.15). Cela pourrait indiquer que d'autres types de loisirs, non pris en compte dans notre questionnaire, pourraient avoir un impact positif sur l'anglais ou de manière plus générale sur les langues étrangères.

Nous avons observé une corrélation positive entre le temps consacré aux études et le temps passé à accomplir des tâches ménagères (0.21) ainsi qu'une corrélation négative entre le temps consacré aux loisirs et le temps passé à accomplir des tâches ménagères de -0,13. Cela pourrait suggérer que les étudiants privilégièrent souvent le temps d'études au détriment du temps de loisir.

## 1.3 - Résultats

Dans cette partie nous présenterons plus en détails les résultats obtenus en se focalisant sur ce que nous souhaitons vérifier lors de l'élaboration du questionnaire mais également examiner le lien entre pratique du jeu vidéo et performances scolaires.

### **1) Lien entre la note en anglais et la pratique des jeux vidéo en anglais**

Une étude de M. Dhany Winaldo et Lulud Oktaviani (2022) suggère que le fait de jouer en anglais a un impact positif sur les performances scolaires. C'est une des raisons pour lesquelles nous avons inclus cette question à notre questionnaire. Pour vérifier si cela se reproduisait sur notre échantillon, nous avons réalisé d'indépendance du khi-deux.

La variable représentant la moyenne en anglais se divise en cinq catégories : "0 à 8", "8 à 10", "10 à 12", "12 à 16" et "16 à 20". Quant à la variable langue : "Anglais", "Français", et "Anglais et Français". (Les réponses mentionnant d'autres langues ont été ignorées en raison de leur faible nombre.)

Aucune des 16 personnes ayant indiqué la catégorie "Anglais", c'est-à-dire qui joue exclusivement en anglais, n'a une moyenne inférieure à 10. Cependant, afin de respecter l'effectif minimal théorique de 5 requis pour réaliser ce test, nous avons regroupé les

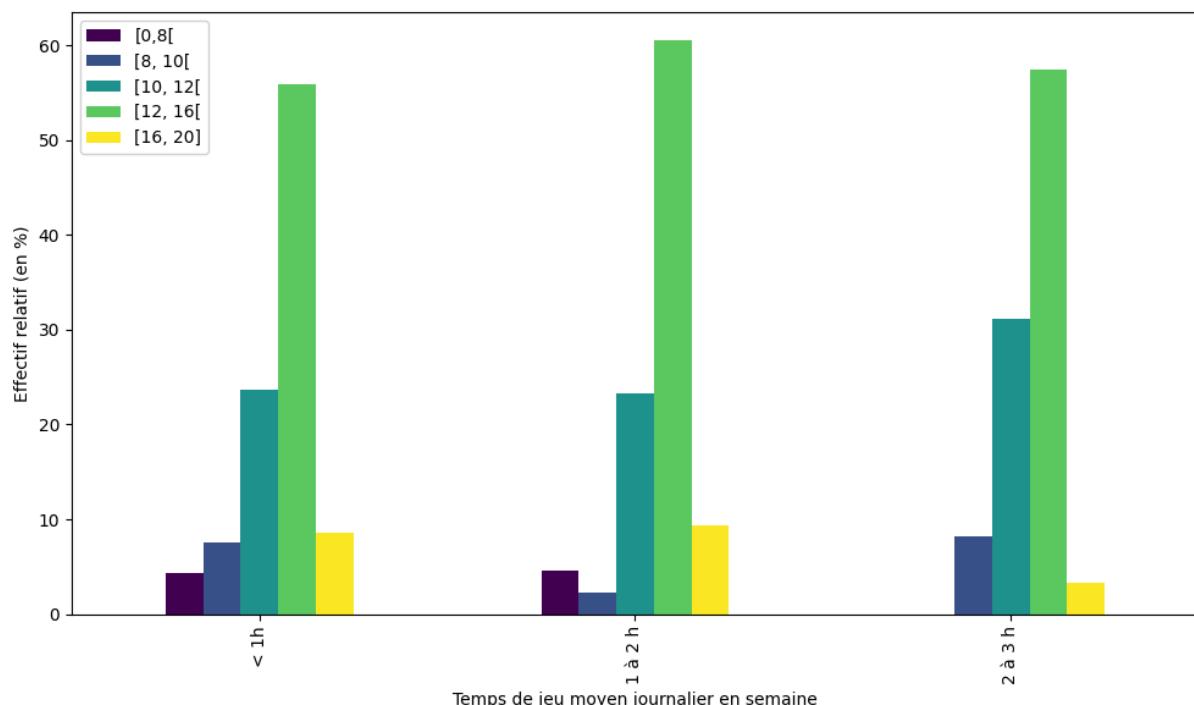
catégories "0 à 8" et "8 à 10" pour la moyenne, ainsi que les catégories "Anglais et Français" et "Anglais" pour la variable langue, ce qui nous donne le tableau d'effectifs ci-dessous.

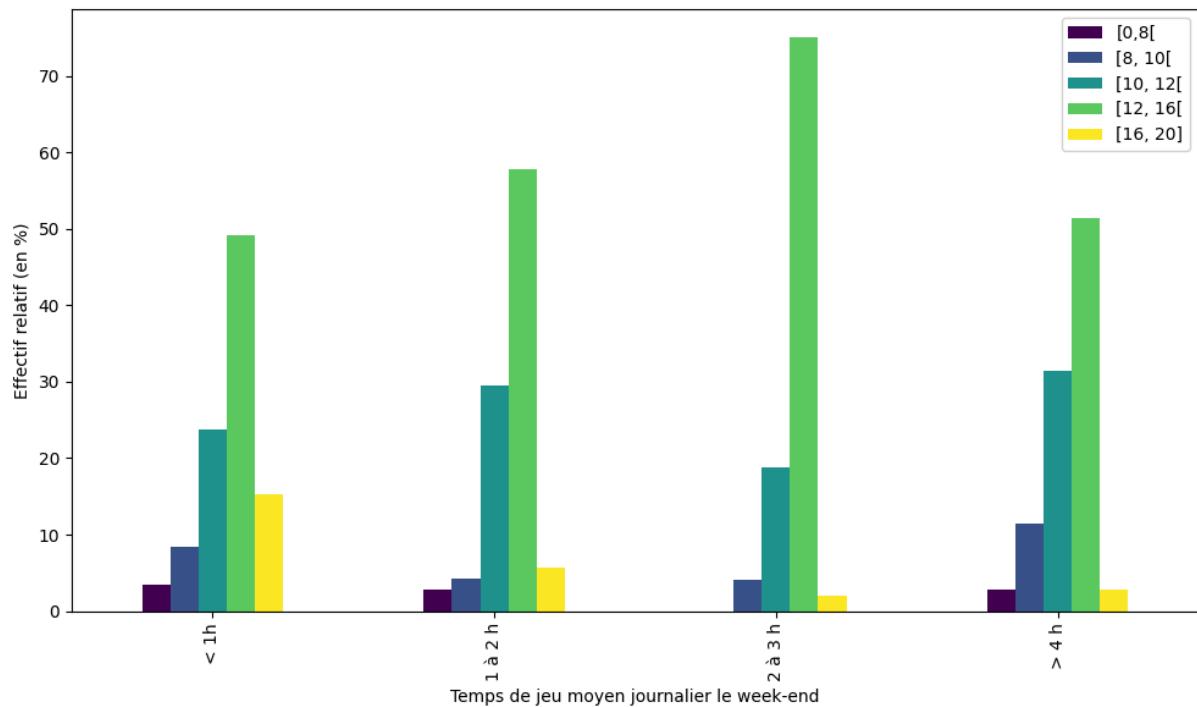
	[0-10[	[10-12[	[12-16[	[16-20[
Français	19	26	24	6
Au moins Anglais	12	14	65	12

La valeur-p est proche de zéro, *l'hypothèse nulle d'indépendance* est donc *rejetée* en faveur de l'hypothèse alternative. Les deux variables sont donc **dépendantes**. Cependant, ce résultat ne permet pas de déterminer si c'est le fait d'être bon en anglais qui incite à jouer en anglais ou, au contraire, si c'est le fait de jouer en anglais qui améliore les performances en anglais.

## 2) Lien entre le temps de jeu vidéo et la moyenne générale

Le coefficient de corrélation entre la moyenne générale et le temps moyen journalier passé sur les jeux vidéo, que ce soit le week-end ou en semaine, est nul. Nous avons réalisé un test statistique de khi-deux d'indépendance pour vérifier s'il existait une relation entre ces variables. Cependant, les valeurs-p obtenues (0,52 pour la semaine et 0,12 pour le week-end) sont supérieures au seuil, nous ne pouvons pas rejeter l'hypothèse nulle d'indépendance et donc en tirer une conclusion. Les histogrammes ne dégagent pas de tendance claire car des répartitions similaires s'observent peu importe le temps de jeu.

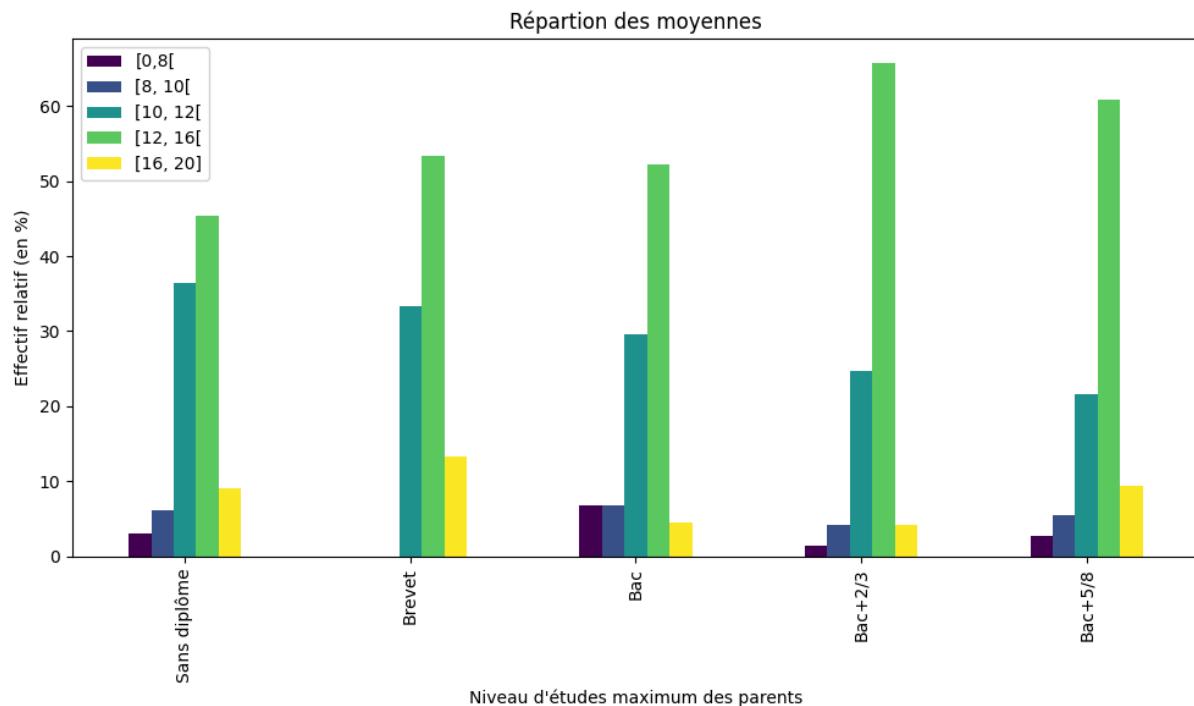




### 3) Influence du niveau d'études des figures parentales sur les notes

Pour examiner s'il existe un lien entre le niveau d'études des parents et la moyenne générale des étudiants de notre échantillon. Nous avons utilisé le niveau d'études le plus élevé entre les deux parents, en tenant compte du cas où un seul parent était renseigné. Le coefficient de corrélation de Spearman obtenu est de 0,10, cela indique une relation faible entre ces deux variables.

En analysant l'histogramme regroupant les moyennes générales en fonction du niveau d'études des figures parentales, on observe une légère diminution de la proportion d'étudiants ayant une moyenne de 10-12 au profit de ceux ayant une moyenne de 12-16, cet effet est cependant modéré.

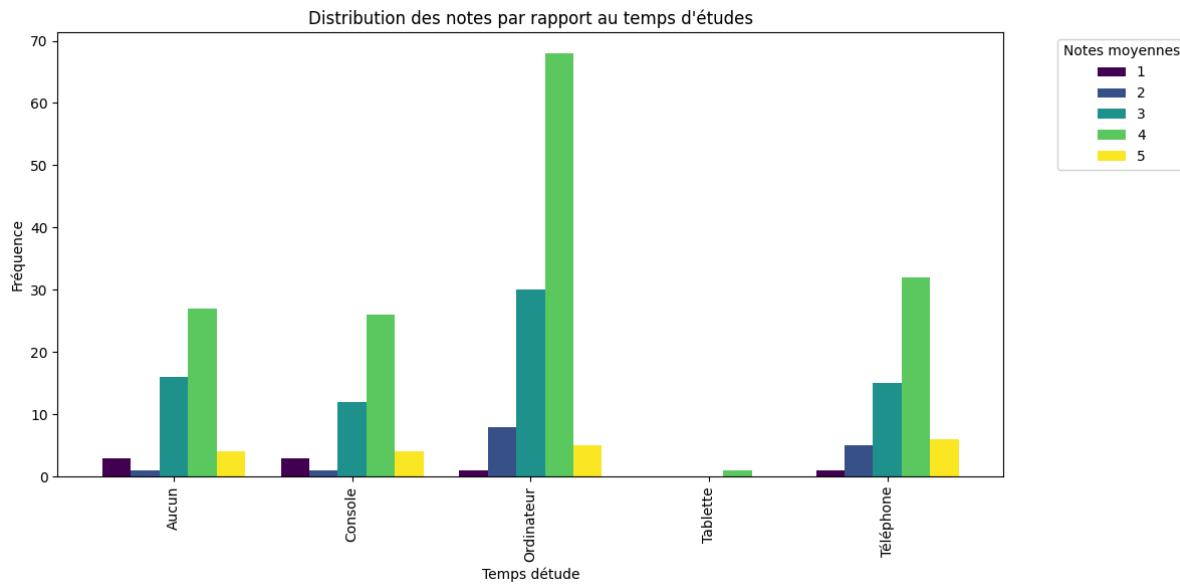


#### 4) Répartition des notes selon le type de joueur

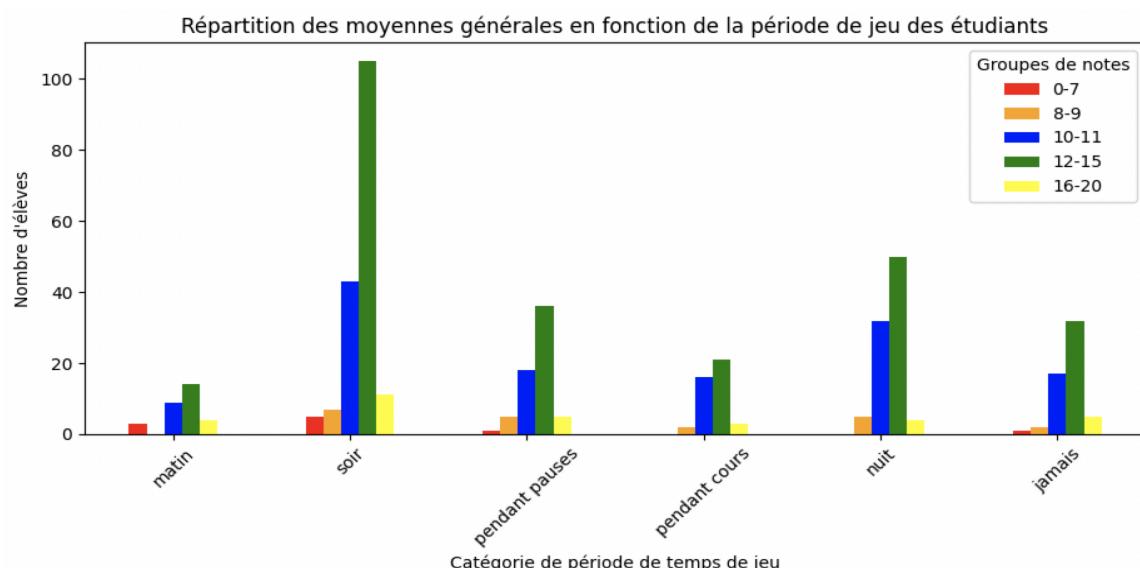
L'analyse du lien entre le support de jeu et les notes n'a pas permis de conclure à une relation significative. Les tentatives d'application de tests statistiques, comme le test de Chi2, ont échoué en raison de la non-conformité des conditions nécessaires, même après des regroupements des catégories de support et des notes. Cette limitation est probablement due à la nature des données disponibles. Une approche exploratoire via des visualisations (histogrammes) a révélé une répartition homogène des notes quel que soit le type de joueur, sans tendance claire.

Plateformes\Notes	[0-10[	[10-12[	[12-20[
Console	4	12	30
Ordinateur	9	30	73
Mobile	6	15	39

(Ici "mobile" est un regroupement de "Téléphone" & "Tablette").



## 5) Impact du jeu vidéo avant l'école sur les performances scolaires



La majorité des étudiants qui **jouent le soir et la nuit** obtiennent des moyennes comprises entre **12 et 15**, suggérant que ces périodes de jeu ne nuisent pas forcément aux performances scolaires.

On peut observer que les catégories "**jouer le matin**" et "**jouer pendant les cours**" ont effectivement le **nombre le plus faible d'élèves** dans la tranche de notes **12-15**, qui est la tranche la plus représentée globalement dans d'autres catégories de jeu (comme le soir et la nuit). Le manque d'effectif dans ces catégories est un frein à l'interprétation des tests.

Les étudiants qui **ne jouent jamais** ont une répartition plus homogène, mais avec une majorité de notes situées dans la plage **12-15**, ce qui peut indiquer que l'absence de jeux vidéo leur permet de se concentrer davantage sur leurs études. Toutefois, une partie d'entre

eux obtient des notes dans les tranches 8-9 et 10-11, ce qui suggère que ne pas jouer aux jeux vidéo ne garantit pas nécessairement des performances scolaires élevées.

Les étudiants qui jouent le matin sont très peu nombreux et se répartissent de manière équilibrée dans toutes les catégories de notes. Cela pourrait indiquer que jouer le matin n'est pas une habitude courante, ou que ce créneau horaire est utilisé de manière ponctuelle. Nos données récoltées ne nous permettent pas vraiment d'en conclure quelque chose de concret.

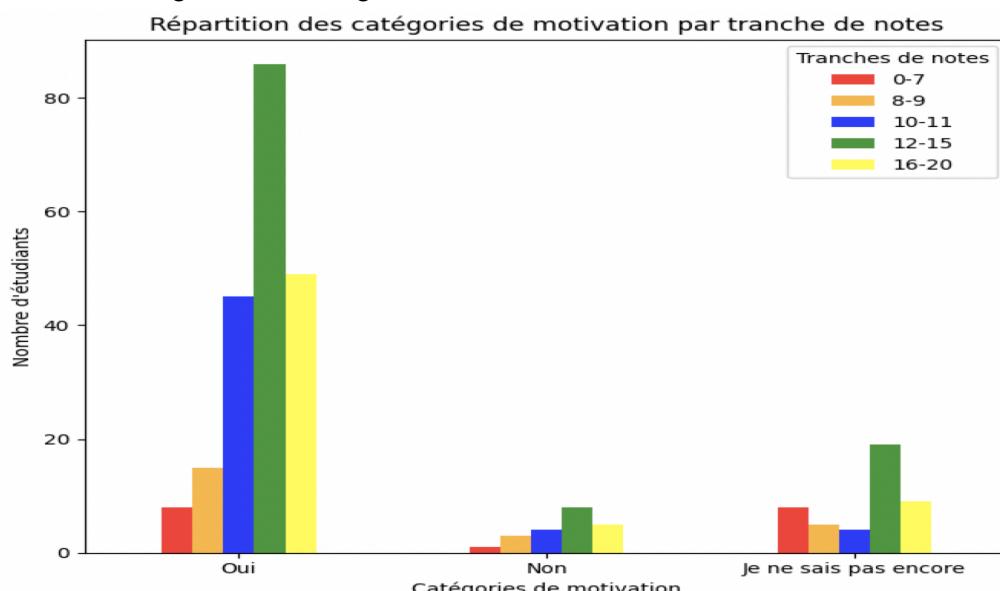
La majorité des étudiants, quelle que soit la période de jeu, se retrouvent dans la catégorie de notes **10-11 et 12-15**, suggérant que le jeu vidéo n'est pas un facteur déterminant unique pour la réussite scolaire, mais qu'il y a d'autres variables en jeu (organisation du temps, matières étudiées, etc.).

Les très bonnes notes (16-20) sont très rares mais réparties de manière similaire entre les différentes périodes de jeu, montrant que certaines personnes parviennent à maintenir un excellent équilibre entre études et loisirs.

L'absence totale de jeux vidéo n'est pas nécessairement synonyme de meilleures performances. Un équilibre entre études et loisirs est probablement la clé de la réussite.

## Conclusion

Certaines catégories de périodes de jeu contiennent un effectif total trop faible, ce qui peut fausser les résultats et limiter la fiabilité des conclusions. En raison de ces effectifs insuffisants, il est difficile de réaliser des tests statistiques supplémentaires, car ils ne seraient pas significatifs. Pour améliorer l'analyse, il serait nécessaire de regrouper certaines catégories ou d'augmenter la taille de l'échantillon.



## 6) Lien entre les indicateurs de motivation et les résultats scolaires

On observe que la majorité des étudiants sont "motivés". En effet, la catégorie "Oui" domine largement avec un grand nombre d'étudiants dans les tranches de notes **16-20** et surtout

**12-15.** Cela suggère que la motivation est un facteur clé pour obtenir de bonnes performances académiques.

On constate que ceux qui ne sont pas motivés ("Non") sont très peu représentés. Il n'y a que 21 étudiants "non motivés", il est donc difficile de concrétiser les résultats.

Parmi le groupe des indécis, une certaine diversité est présente avec des étudiants dans toutes les tranches de notes. On remarque tout de même une présence importante dans la tranche **12-15**, ce qui suggère qu'une certaine incertitude sur la motivation ne signifie pas nécessairement de faibles performances académiques.

Le nombre d'étudiants motivés est nettement supérieur aux autres catégories, montrant que la majorité des étudiants se sentent motivés par leurs études. Leur forte représentation dans les bonnes tranches de notes confirme une corrélation positive entre motivation et succès académique.

On constate que dans la tranche de notes **16-20**, la plupart des étudiants appartiennent au groupe "Oui" et "Je ne sais pas encore", tandis que très peu appartiennent au groupe "Non". Cela confirme que la motivation joue un rôle crucial pour atteindre l'excellence académique.

## Conclusion

Cette étude montre que la motivation est un facteur déterminant dans la réussite académique, avec une forte concentration des étudiants motivés dans les tranches de notes élevées. Les étudiants non motivés sont peu nombreux, rendant difficile une conclusion définitive sur leur impact. Les étudiants indécis affichent des performances variées, suggérant que l'incertitude sur la motivation ne conduit pas nécessairement à de mauvais résultats.

## Test de Chi2

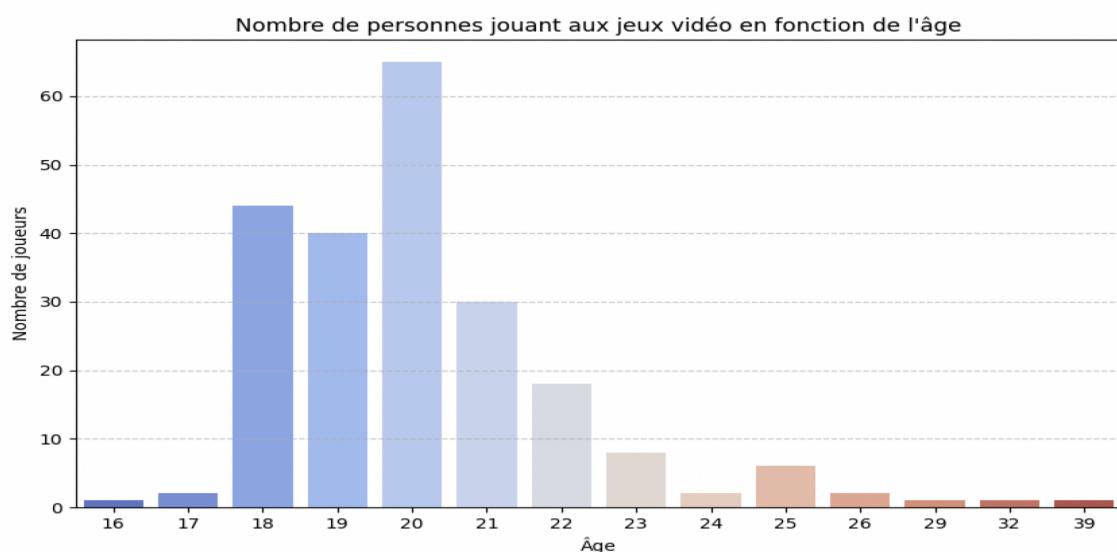
Afin de confirmer statistiquement l'existence d'une relation entre la motivation et les performances académiques, il est pertinent d'effectuer un **test d'indépendance du khi-2**. Ce test permettra de déterminer si la répartition des notes est significativement influencée par le niveau de motivation des étudiants. Nous allons donc procéder à cette analyse pour valider nos observations.

Pour que ce test soit valide, l'effectif théorique de chaque case du tableau de contingence doit être au moins de 5. Dans notre cas, la "**motivation**" est une variable pouvant avoir 3 valeurs possibles, "**notes**" est une variable contenant 5 catégories possibles et les données contiennent 269 observations. Cela amène à un effectif théorique trop faible. Les données ont donc été regroupées pour satisfaire ce prérequis, nous obtenons ce tableau. Cela permet aussi de rééquilibrer les effectifs théoriques.

	[0-8[	[8-10[	[10-12[	[12-16[	[16-20[
Motivés	8	15	45	86	49
Non motivés/ Je ne sais pas	9	8	8	27	14

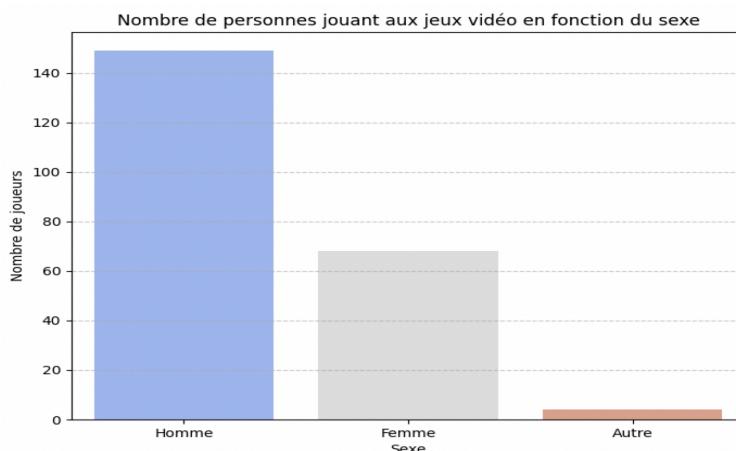
La valeur-p est proche de zéro, l'hypothèse nulle d'indépendance est donc rejetée en faveur de l'hypothèse alternative. Les deux variables sont donc **dépendantes**. Il existe une relation significative entre la motivation et la performance académique.

## 7) Profil des joueurs : Qui sont les étudiants qui jouent aux jeux vidéo ?



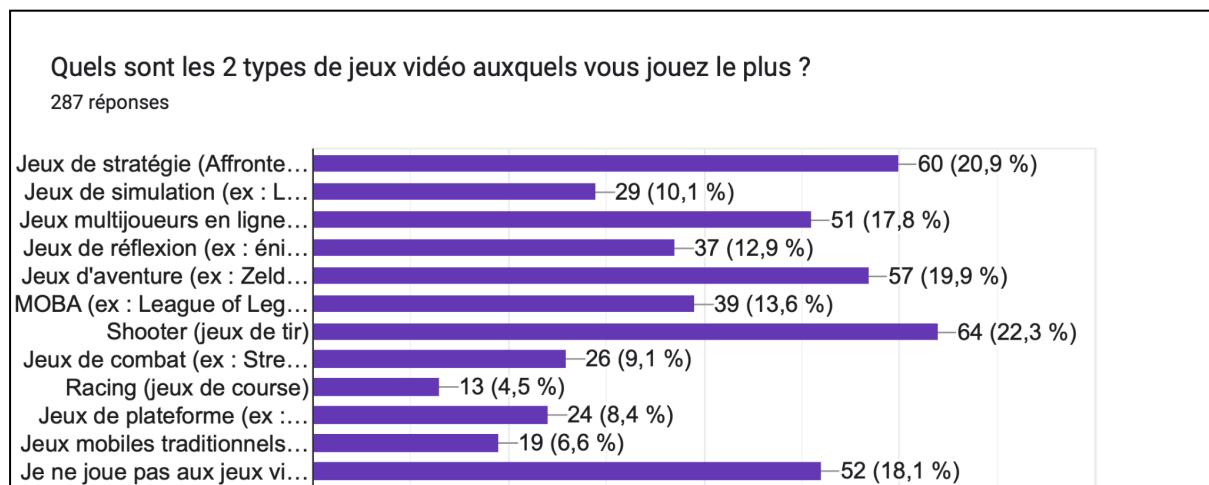
De plus, l'analyse des données révèle que l'**âge moyen des joueurs** est de **20,19 ans**, tandis que celui des non-joueurs est légèrement plus élevé, à **21,38 ans**. Cela suggère que les étudiants plus jeunes sont davantage enclins à jouer aux jeux vidéo.

Le graphique ci-dessous illustre parfaitement que la majorité des joueurs se situent dans la tranche d'âge de 18 à 20 ans, avec une concentration particulièrement élevée à 20 ans. À mesure que l'âge augmente, le nombre de joueurs diminue progressivement, indiquant une baisse de l'intérêt ou de la disponibilité pour les jeux vidéo chez les étudiants plus âgés.



L'analyse des données montre que parmi les étudiants qui jouent aux jeux vidéo, **149 sont des hommes**, **68 sont des femmes**, et **4 étudiants** appartiennent à la catégorie "Autre" ou n'ont pas précisé leur genre. Ces chiffres révèlent une présence majoritaire d'hommes parmi les joueurs.

Il n'est pas pertinent d'analyser la répartition des joueurs par domaine d'études, car la majorité des données collectées proviennent d'étudiants en informatique, issus de notre promotion ou d'autres promotions du même domaine. Cette forte représentation du domaine informatique fausserait les résultats et les rendrait non représentatifs de l'ensemble de la population étudiante. Cela a été calculé au préalable pour vérifier notre hypothèse, avec une simple manipulation des données, et notre hypothèse s'est avérée bonne.



L'enquête révèle que les **jeux de tir (22,3%)**, de **stratégie (20,9%)**, et d'**aventure (19,9%)** sont les plus populaires parmi les étudiants, montrant une préférence pour les expériences compétitives et immersives. Les **jeux de simulation (17,8%)** et les **MOBA (13,6%)** restent également prisés, tandis que les jeux de **combat (9,1%)**, de **plateforme (8,4%)**, et les **jeux mobiles (6,6%)** attirent moins d'étudiants.

À noter que **18,1% des répondants** déclarent ne pas jouer, soulignant qu'une part importante de la population étudiante ne pratique pas cette activité.

## **8) Comparaison des performances entre les étudiants en informatique et ceux d'autres filières**

Les analyses ont révélé des distributions similaires entre les étudiants en informatique et ceux des autres filières, tant pour les notes moyennes générales que pour les notes d'anglais. Les histogrammes et les courbes gaussiennes montrent un chevauchement important, avec des moyennes centrées autour de 12 et 16 (/20) pour les deux groupes. Ces résultats ne mettent pas en évidence de différence significative entre les performances académiques des deux populations, que ce soit pour les notes générales ou pour les performances en anglais.

Cependant, il est important de noter que, bien que des histogrammes aient été utilisés pour visualiser ces distributions, il n'a pas été possible d'appliquer un test de Chi2 pour évaluer

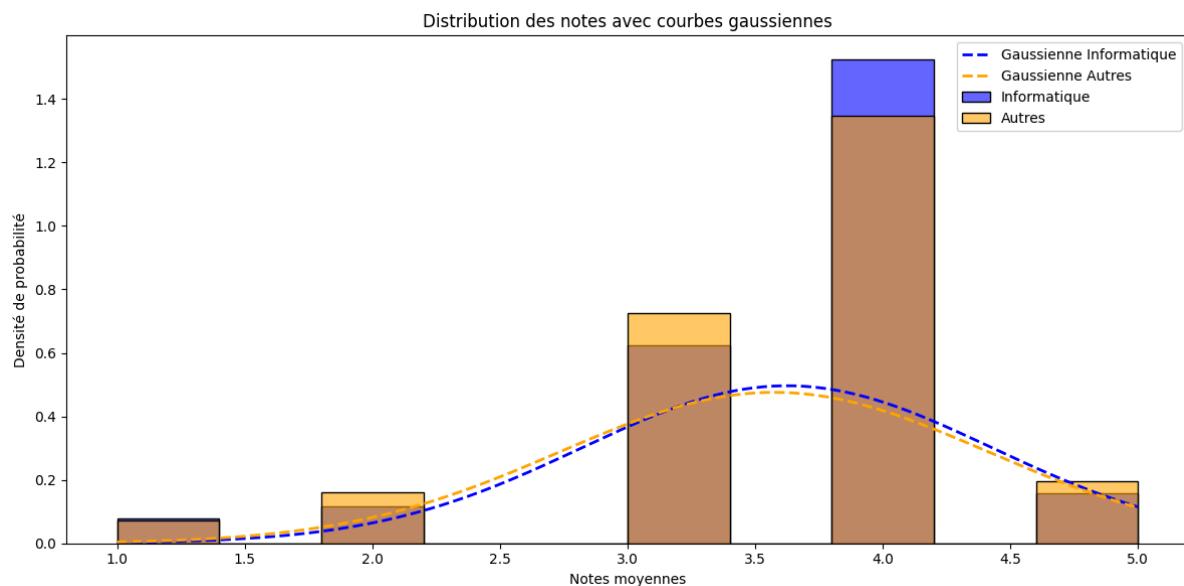
l'indépendance des variables en raison de la non-conformité des conditions nécessaires (effectifs insuffisants dans certaines catégories). Ainsi, les conclusions reposent uniquement sur les observations visuelles des distributions.

Les résultats montrent également que :

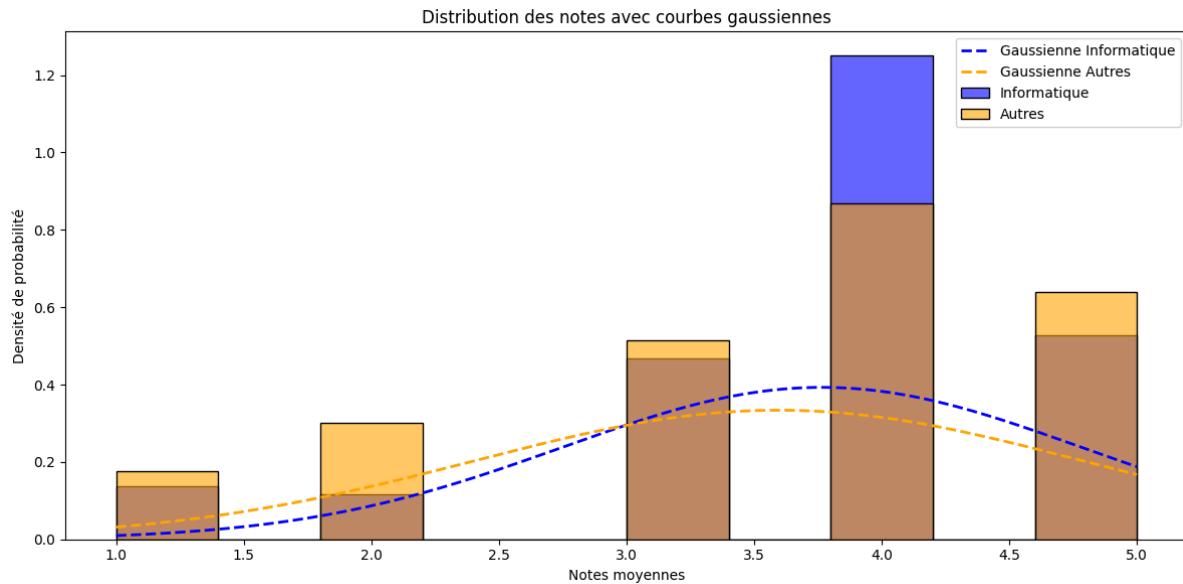
- Les étudiants en informatique ont une répartition légèrement plus homogène autour de l'intervalle [12-16[ (représenté par la valeur 4), tant pour les notes générales que pour les notes d'anglais.
- Les étudiants des autres filières affichent une variabilité légèrement plus importante, notamment avec une présence accrue dans les catégories extrêmes (notes de 1 et 5).

Ces analyses permettent de valider partiellement la fiabilité des données collectées sur les performances académiques. Cependant, cette absence de différenciation pourrait être liée à des biais dans la conception ou la diffusion du questionnaire.

Notes moyennes générales :



Notes moyennes d'anglais :



Méthode de réalisation du graphique

Pour réaliser le graphique, les données ont été divisées en deux groupes :

1. Les étudiants en informatique.
2. Les étudiants d'autres filières.

La répartition est supposée gaussienne, ses paramètres (moyenne et écart-type) ont été interpolés à partir des données en utilisant la bibliothèque `scipy`. Ces courbes ont ensuite été superposées à des histogrammes représentant les distributions des notes pour chaque groupe. Les barres des histogrammes et les courbes gaussiennes sont colorées distinctement pour faciliter la comparaison.

## 9) Relation entre le temps d'étude et les résultats scolaires

L'analyse de la relation entre le temps d'étude déclaré et les performances scolaires, mesurées par les notes moyennes et les notes d'anglais, n'a pas permis de mettre en évidence de corrélation significative. Les résultats obtenus montrent :

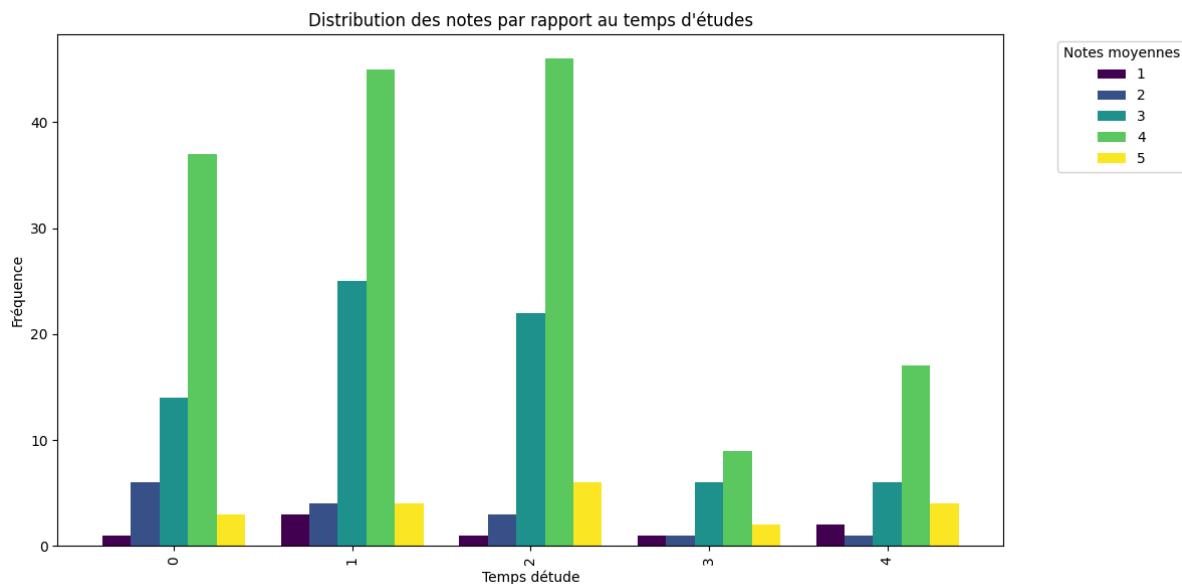
Entre le temps d'étude et les notes moyennes générales, le coefficient de Spearman est très faible ( $r_s = 0.05$ ) avec une p-valeur élevée, indiquant l'absence de relation significative.

Pour les notes d'anglais, la corrélation est légèrement négative ( $r_s = -0.06$ ) mais reste également non significative.

Après regroupement des catégories de temps d'étude et des notes en classes, le test d'indépendance du Chi2 n'a pas permis de rejeter l'hypothèse d'indépendance (p-valeur = 0.52). On ne peut donc rien conclure sur la dépendance des notes vis-à-vis du temps d'études.

Ces résultats indiquent que, dans ce questionnaire, le temps d'étude, tel qu'il est mesuré, n'apparaît pas comme un facteur déterminant pour expliquer les performances académiques. Cependant, ces conclusions doivent être nuancées par certaines limites, par

exemple le temps d'étude est auto-déclaré, ce qui peut introduire des biais.



## 10) Impact du temps consacré aux activités obligatoires sur les notes

L'analyse de l'impact du temps consacré aux activités obligatoires, telles que les tâches ménagères en semaine et le weekend, sur les résultats scolaires n'a révélé aucun effet significatif. Les corrélations calculées entre le temps dédiés aux tâches ménagères et les notes moyennes générales ( $r_s = 0.05$  en semaine,  $r_s = 0.02$  le weekend) ou les notes d'anglais ( $r_s = 0.09$  en semaine,  $r_s = 0.00$  le weekend) sont faibles, indiquant une absence de lien clair. Les tests d'indépendance du Chi2 et d'homogénéité ont confirmé cette conclusion, avec des p-valeurs élevées ne permettant pas de rejeter l'hypothèse d'indépendance. Ces résultats suggèrent que le temps consacré aux activités obligatoires, tel que mesuré dans ce questionnaire, n'a pas d'impact significatif sur les performances académiques. Toutefois, ces observations pourraient être influencées par des limites du questionnaire, comme l'absence de précision sur la nature ou l'intensité des tâches réalisées.

## 11) Comparaison des performances entre les étudiants en retard et ceux ponctuels

Une corrélation négative de  $r_s = -0.14$  a été observée entre les retards et les notes moyennes, indiquant qu'une augmentation des retards est associée à une légère baisse des performances académiques. Bien que cette tendance soit faible, elle est cohérente avec l'idée que les retards peuvent refléter un manque d'organisation ou des obstacles qui affectent les performances scolaires. Aucune différence marquée n'a été observée pour les autres métriques étudiées, et les visualisations montrent une homogénéité dans la répartition des notes entre ces groupes.

## 12) Lien entre jeux vidéo violents et performances académiques

Différentes études suggèrent que le fait de jouer à des jeux violents pourrait avoir un impact négatif sur les performances scolaires. Nous avons donc inclus une question dans notre questionnaire pour vérifier ce lien. Parmi les répondants joueurs, 35 % ont indiqué jouer à

des jeux violents, tandis que 65 % ont répondu par la négative. Le test d'indépendance a donné une valeur-p de 0,28, ce qui ne permet pas de tirer de conclusion statistique.

En analysant les résultats, nous n'observons pas de différence marquée entre les deux groupes de notre échantillon concernant l'impact sur les notes. Pour chaque catégorie de notes, les proportions sont similaires, à l'exception d'une différence : parmi ceux qui ne jouent pas à des jeux violents, 10 % obtiennent une excellente note, contre seulement 5 % parmi ceux qui jouent à des jeux violents.

### **13) Effet des réseaux sociaux et des jeux vidéo sur la concentration et le sommeil**

Nous avons inclus trois questions pour évaluer la perception des étudiants de l'impact de leur temps de jeu sur le sommeil et la concentration, ainsi que celui de l'impact de leur temps passé sur les réseaux sociaux sur leur sommeil. Les questions étaient fermées. ('Oui', 'Non', 'Je ne sais pas')

Le manque de sommeil ou une difficulté de concentration a un impact sur les performances, ce qui pourrait potentiellement se refléter sur les notes.

Bien que ces questions reposent sur le ressenti subjectif des répondants, notre but était de déterminer s'il était possible d'observer une variation sur la répartition de la moyenne générale entre le groupe ayant estimé que cela n'avait pas d'impact et celui ayant estimé que oui.

Nous avons réalisé des tests d'indépendance de khi-deux pour examiner les relations entre les perceptions des répondants et leurs notes. Nous avons regroupé les catégories de notes 0-8 et 8-10, et avons exclu les réponses 'Je ne sais pas'. Les couples de variables étudiées sont:

- 1) Perception de l'impact du temps de jeu vidéo sur la concentration et les notes.
- 2) Perception de l'impact du temps de jeu vidéo sur le sommeil et les notes.
- 3) Perception sur l'impact du temps de jeu vidéo sur la concentration et les notes.

Pour chacun des trois couples, les valeurs-p sont largement au-dessus du seuil significatif (De 0,28 à 0,66). Pour chacun des trois couples, nous notons que la répartition des notes entre le groupe 'Oui' et 'Non' est quasi-identique. Nous observons quelques nuances, non significatif en raison du faible échantillon, dans les notes extrêmes pour les deux premiers couples de variables:

- Pour (1), dans les notes supérieures à 16, 9 personnes pour le 'Non' et 4 pour 'Oui'.
- Pour (2), dans les notes supérieures à 16, 10 personnes pour le 'Non' et 5 pour 'Oui' et dans les notes inférieures à 10, 16 personnes pour le 'Non' et 13 pour 'Oui'

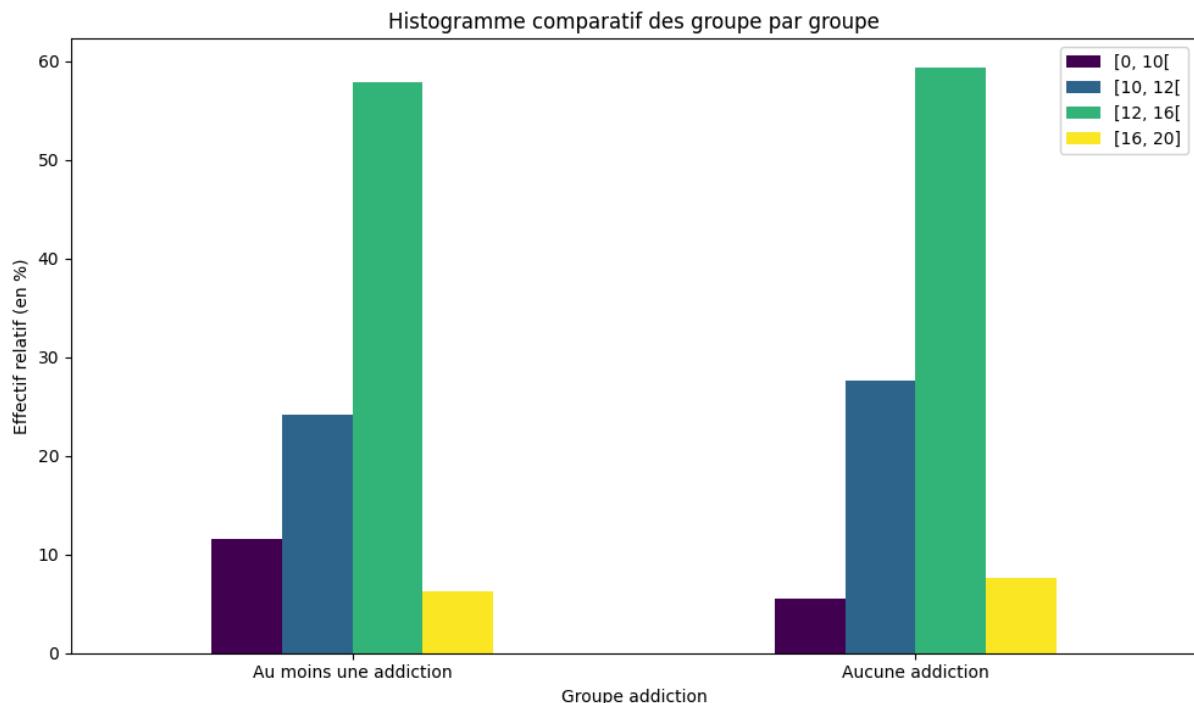
En définitive, nous ne notons aucune différence notable. Il est à noter qu'il est difficile de mesurer avec une certaine justesse l'impact d'un comportement ou une habitude sur soi; il est tout à fait possible de surestimer ou sous-estimer l'impact d'un comportement.

#### 14) Lien entre addiction et moyenne générale

Nous avons inclus une question dans le questionnaire afin d'évaluer le lien entre addiction et moyenne générale. La question était formulée ainsi : "Vous considérez-vous addict à:". Les options proposées étaient : 'Alcool', 'Tabac', 'Sport', 'Jeux vidéos', 'Aucune addiction', avec aussi une option libre.

La majorité des personnes ont déclaré n'avoir aucune addiction. Vu le faible échantillon pour les autres catégories, nous avons regroupé les différentes addictions en une seule catégorie, après avoir exclu les réponses de type 'Préfère ne pas répondre'. Pour pouvoir évaluer le lien avec la moyenne générale nous avons procédé à un test du khi-deux. (Les catégories '0-8' et '8-10' ont été regroupées pour satisfaire l'exigence d'effectif minimal théorique de 5)

La valeur-p est de 0,38; aucune conclusion statistique n'est possible. En consultant l'histogramme comparatif, nous ne constatons aucune différence notable entre les deux groupes.



Comme pour les questions sur la perception des étudiants de l'impact de leur temps de jeu sur le sommeil et la concentration, la question de l'addiction peut être délicate (Les personnes peuvent hésiter/ne pas être conscientes à admettre une telle dépendance, et la société elle-même peut minimiser l'ampleur de ce phénomène). La manière dont les questions sont posées est donc importante. Initialement, le questionnaire comportait une section sur l'alcool, cela a été supprimé afin d'éviter de réduire le nombre de réponses, dans le but d'obtenir un maximum de réponses.

## 2. Analyse comparative

Les deux études ont des similarités dans la mesure où elles cherchent à analyser l'impact de certains facteurs sur les performances scolaires. Cependant, nous souhaitons rappeler qu'il y a des différences: les populations étudiées (Pour les données Kaggle des lycéens portugais et pour notre questionnaire des étudiants) ne sont pas les mêmes (pays différents, lycéens d'un côté et étudiants de l'autre), et les questions communes posées ne sont pas à strictement pareils. A ce titre, il nous semble plus correct de considérer une comparaison de nos interprétations des corrélations observées plutôt qu'une comparaison directe des valeurs de corrélation.

Chez les lycéens, le niveau d'éducation semble avoir une influence positive sur les notes, avec une corrélation d'environ 0,15. Concernant les étudiants, une corrélation de 0,10 est observée. Bien que cet effet semble moins fort en comparaison, nous pourrions soupçonner qu'il existe une certaine transmission de la méthodologie ou des habitudes des parents vers leurs enfants.

Aussi, il semble y avoir des similarités sur ce que nous pourrions qualifier "d'être consciencieux". Chez les lycéens, ceux obtenant de meilleures notes semblent moins consommer d'alcool (Corrélation négative de l'ordre de -0,2). Et côté étudiants, nous notons qu'il y a l'air d'avoir une tendance à privilégier le temps d'étude au détriment du temps de loisir, avec une corrélation négative de -0,2.

Chez les lycéens, nous observons une corrélation négative de l'ordre de -0,3 entre l'âge et les notes; comme mentionné précédemment, cela pourrait traduire une forme de perte d'engagement. (Nous supposons que cela a un lien avec le redoublement) Côté des étudiants, la corrélation entre l'âge et les notes est nulle. Cependant, il semble y avoir aussi une forme de perte d'engagement car nous observons une corrélation positive de l'ordre de 0,3 est observée entre les notes et le fait d'être en retard ou de rendre des exercices en retard. Il semblerait qu'il y a une forme de perte d'engagement au sein des deux populations même si les variables qui les révèlent ne sont pas les mêmes bien que les causes ne soient pas connues.

# VII. Développement de l'interface web

## 1 - Introduction

Dans le cadre d'un projet de si grande ampleur, les données récoltées venant de l'étude "Student Alcohol Consumption" et/ou de notre recherche sur l'impact sur les jeux vidéo, se devaient de pouvoir être visualisées.

Dans ce cadre, nous avons mis en place une interface web permettant à l'équipe d'analyse et à des utilisateurs lambda du site de pouvoir visualiser et interagir avec les données.

L'interface web permet également aux utilisateurs d'obtenir des informations sur notre groupe, notre projet et sur les données récoltées.

## 2 - Description de l'interface web

Notre site comporte 3 onglets :

- Accueil
- Etudes
- A propos

Le contenu de ces onglets va être explicité dans les sections suivantes.

### 2.1 - Onglet "A propos"



Figure n°1 : Logo

Cet onglet permet aux utilisateurs de découvrir l'histoire liée à notre logo afin d'en comprendre le choix. Ce logo a été généré via une intelligence artificielle.

Nous pouvons y voir un bateau de la Royal Navy (pendant la Seconde Guerre Mondiale) coloré en rose car en 1940 lors d'un convoi maritime, le général Mountbatten a remarqué qu'un de ses bateaux, peint en rose, était moins visible que les autres lors du coucher de soleil. Il a alors fait repeindre toute sa flotte en rose, pensant que cela lui donnerait un avantage stratégique. Cette couleur a été renommée le rose "Mountbatten".

De plus, cet onglet présente aussi les rôles des trois équipes (extraction, analyse et interface web) en décrivant brièvement l'objectif de chaque équipe et les tâches réalisées.

## 2.2 - Onglet "Accueil"

La page d'accueil a pour but d'interpeller l'utilisateur via quelques chiffres clés, présentés dans des number cards, trouvés sur le web. Ensuite, l'utilisateur pourra porter son regard sur un titre introductif accompagné d'un texte permettant d'introduire le sujet de la consommation de jeux vidéo en France.



Figure n°2 : Accueil cards

En dessous de cela nous trouvons un court texte explicatif / résumé du projet.



Figure n°3 : Explication du projet

En bas de la page vous trouverez la section "Nos résultats" qui permet de télécharger l'entièreté des analyses effectuées par le groupe d'analyse.

## Nos résultats

Découvrez nos résultats détaillés en téléchargeant le document PDF ci-dessous.

[Télécharger le PDF](#)

Figure n°4 : Bouton de téléchargement des analyses

### 2.3 - Onglet "Etudes"

Comme expliqué dans l'accueil de notre site, nous disposons de 2 sets de données. Lorsque vous cliquez sur l'onglet de navigation 'Études', vous aurez l'opportunité de visualiser, à travers des graphiques interactifs, les données de deux études : "Student Alcohol Consumption" par P. Cortez et A. Silva. et notre étude concernant l'influence des jeux vidéo sur les performances scolaires.

Cette page présente un ensemble de graphiques (histogrammes et jauge) permettant d'observer les notes et la répartition des élèves selon différents critères. Les histogrammes nous permettent de visualiser les notes selon des catégories (intervalle de notes) tandis que les jauge vont nous permettre d'observer une note moyenne selon le critère choisi.

[Student Alcohol Consumption](#)

[Jeux Vidéos](#)

Figure n°5 : Boutons de choix de l'étude à visualiser

#### 2.3.1 - Onglet "Etudes" partie "Student Alcohol Consumption"

Lorsque le bouton "Student Alcohol Consumption" est sélectionné un court texte explicatif nous introduit l'étude qui a été faite. Pour une question d'ergonomie, un bouton 'en savoir plus' a été mis en place afin d'afficher plus en détail l'explication de l'étude.

### Étude : Student Alcohol Consumption

L'étude **Student Alcohol Consumption** est une recherche menée au Portugal, visant initialement à explorer le lien entre consommation d'alcool et performances scolaires. Pour cela, 674 lycéens provenant de 2 lycées ont été interrogés. [En savoir plus](#)

Figure n°6 : Texte explicatif de l'étude student alcohol consumption

En dessous du texte, plusieurs graphiques sont visibles.

L'histogramme permettra de voir la répartition des étudiants par tranches de notes selon différents critères :

- Aucun filtre
- Niveau académique des parents
- Fréquence des sorties
- Consommation d'alcool

Tout d'abord, un histogramme (diagramme en bâton) :

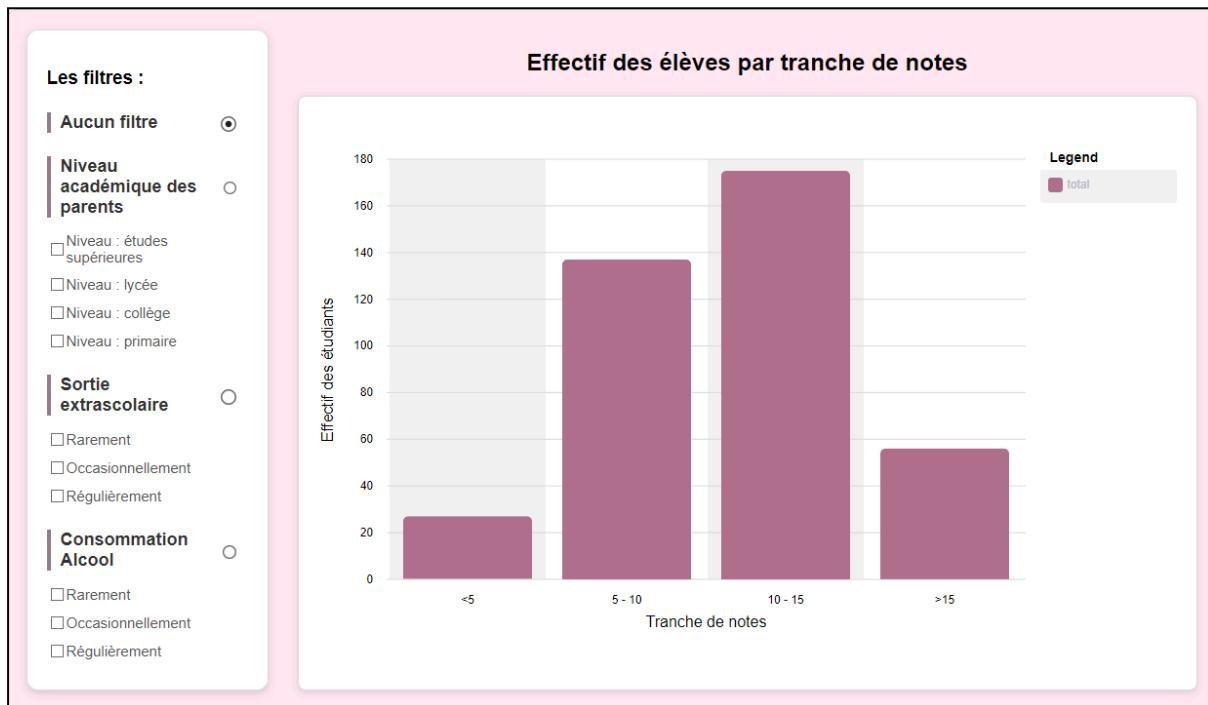


Figure n°7 : Filtre: 'Aucun filtre'

Si vous placez votre curseur sur une des barres, elle vous affichera l'effectif précis.

Ce diagramme ci-dessus présente l'effectif des élèves par tranche de notes. L'utilisateur va pouvoir affiner les données via un filtre qu'il peut sélectionner dans le menu de gauche. Ici aucun filtre n'est sélectionné, ainsi le bar chart présente simplement les effectifs complets par tranche de note.

Lors de la sélection d'un filtre, par exemple le filtre "niveau académique des parents", tous les niveaux académiques sont automatiquement sélectionnés :

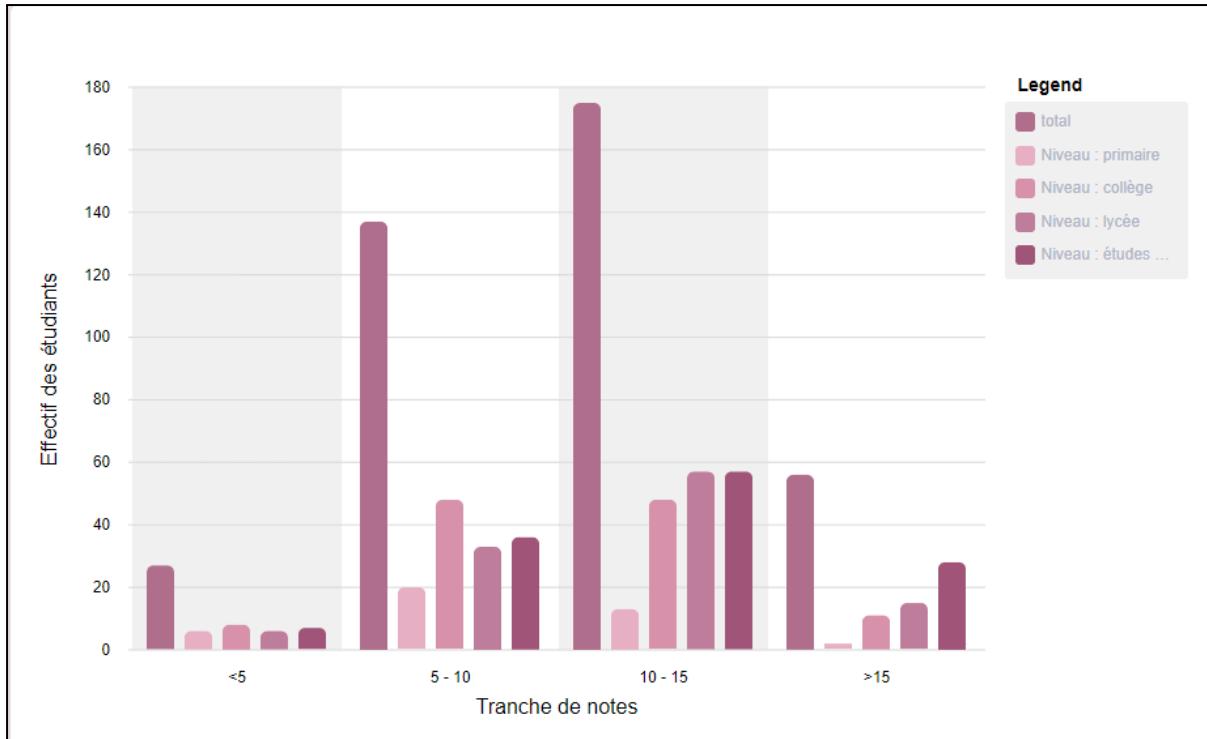


Figure n°8 : Filtre ‘niveau d’étude des parents’

Pour chaque tranche, la barre de gauche affichera toujours le nombre total d’élèves dans cet intervalle de notes. Ensuite, puisqu'il y a ici 4 niveaux d'éducation des parents possibles, pour chaque tranche, 4 nouvelles barres apparaîtront, correspondant respectivement à un niveau d'éducation des parents.

Il est également possible de sélectionner le filtre selon la fréquence de sorties extrascolaires:

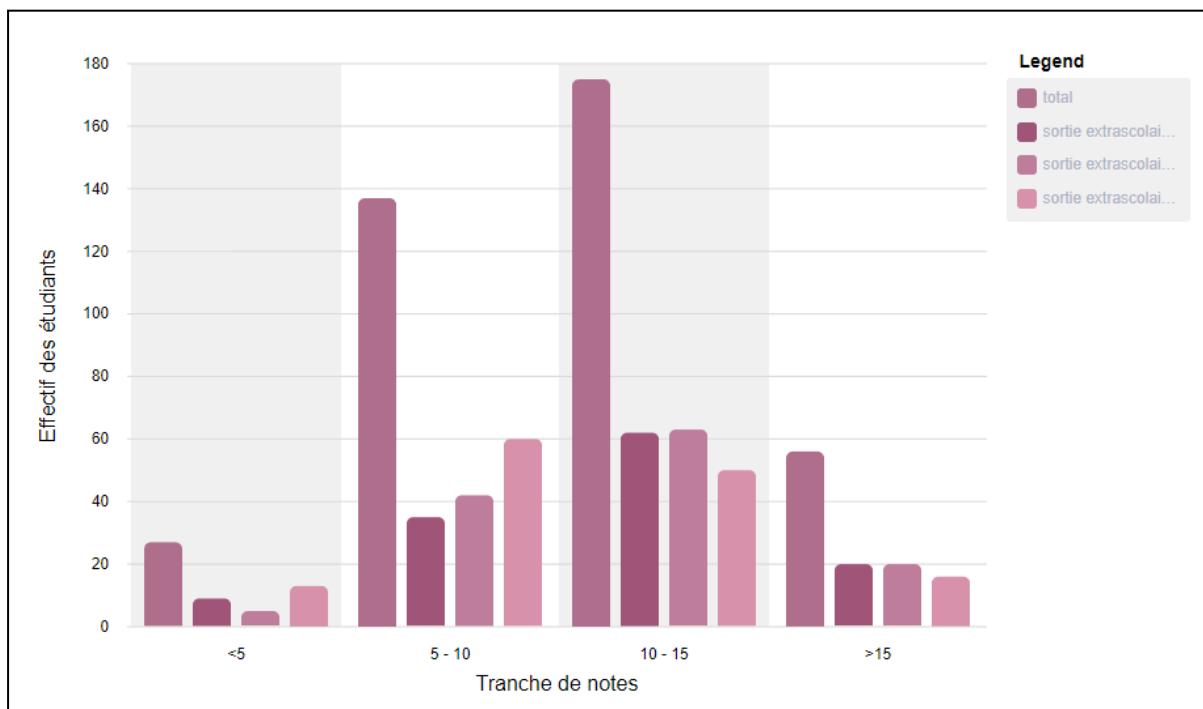


Figure n°9 : Filtre ‘sorties extrascolaire’

Ici, il y a 3 niveaux de sorties extrascolaires (Rarement, Occasionnellement, Régulièrement).

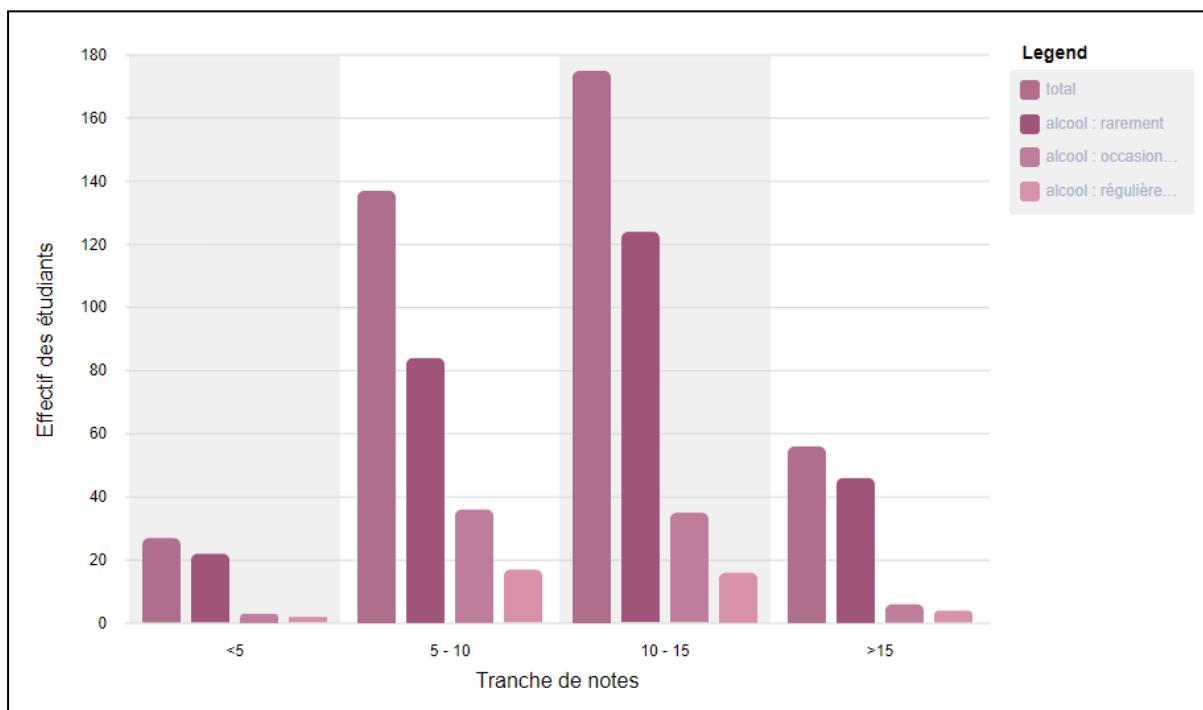


Figure n°10 : Filtre ‘consommation d’alcool’

Tout comme le filtre 'sorties extrascolaires' 3 niveaux de consommation d'alcool ont été mis en place (Rarement, Occasionnellement, Régulièrement).

Ensuite, lorsque vous scrollez vers le bas de la page vous arriverez sur un second graphique. Il s'agit d'une jauge permettant d'observer la note moyenne (moyenne des 3 à 6 notes fournies par individu). Cette note peut être visualisée selon 2 critères :

- Temps de trajet
- Temps de révision.

Vous pourrez aussi observer un slider qui permet de déterminer la moyenne des notes des étudiants en filtrage dynamique. Le déplacement de ce slider permet d'affiner les données analysées en filtrant uniquement les étudiants ayant un temps d'activité supérieur à la condition affichée (inverse d'un effet cumulé). Par exemple, un étudiant ayant un temps de trajet supérieur à 2h sera aussi compté parmi les étudiants ayant un temps de trajet supérieur à 1h et ce jusqu'au 1er cran tandis qu'il ne sera pas compté dans les cas supérieurs.

Ainsi cette jauge permet d'observer la tendance de l'évolution des notes.

4 crans ont été implémentés concernant le temps de trajet, avec de la gauche vers la droite :

- < 15 minutes
- < 30 minutes
- > 30 minutes
- > 1 heure

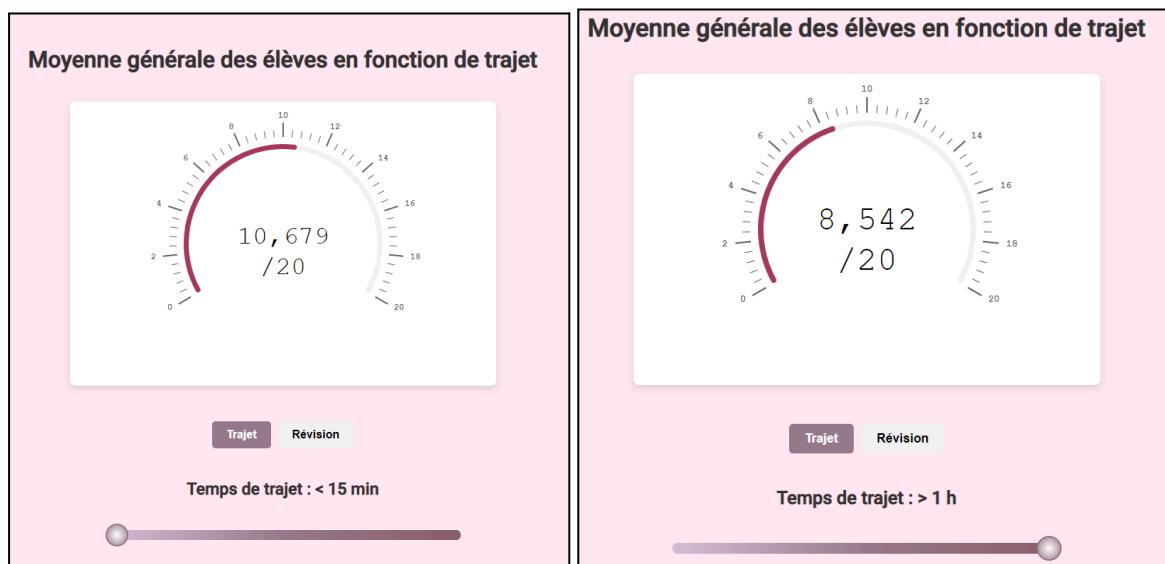


Figure n°11 : Filtre 'Jauge temps de trajet'

Le second bouton “Révision” nous permet de voir une note moyenne en fonction du temps de révision par semaine. De même, 4 crans ont été implémentés avec de la gauche vers la droite :

- < 2 heures
- > 2 heures
- < 5 heures
- > 10 h

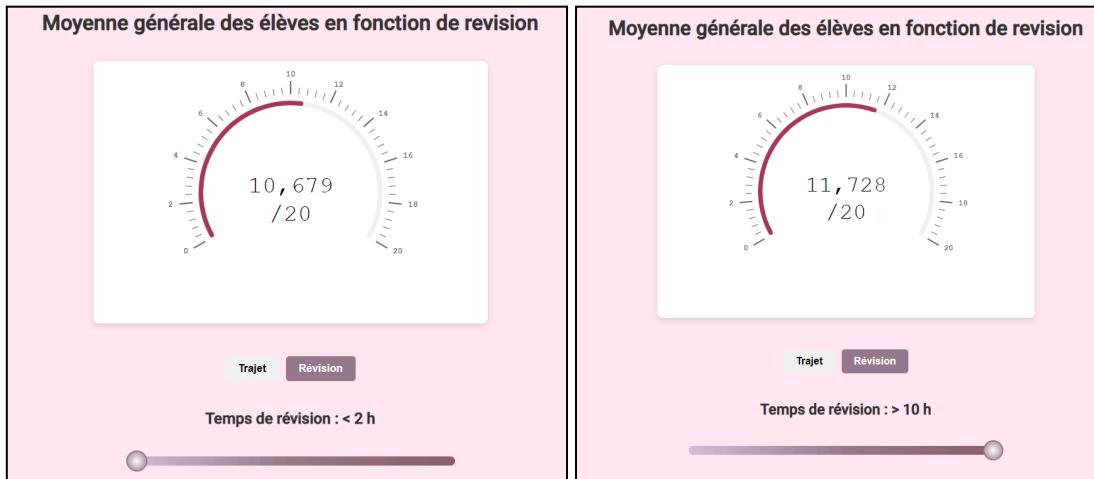


Figure n°12 : Filtre ‘Jauge temps de révision’

### 2.3.2 - Onglet “Etudes” partie “Jeux vidéo”

Le second bouton “Jeux vidéos” (Cf. figure n°5) permet de visualiser les graphiques obtenus via l’analyse des données du questionnaire.

Comme pour la section “Student Alcohol Consumption”, un texte d’introduction a été mis en place avec, toujours dans un but d’ergonomie un bouton “en savoir plus” afin d’avoir le détail de notre recherche sur les jeux vidéo.

#### Étude : Consommation de jeux vidéo et performances académiques

Notre groupe avait pour but de recueillir des données afin d’analyser le lien entre consommation de jeux vidéo et performances académiques. Pour cela, un questionnaire a été conçu, permettant de récolter 269 réponses d’étudiants français en études supérieures.

[En savoir plus](#)

Figure n°13 : Texte explicatif ‘Jeux vidéos et performances académiques’

Tout comme la page ‘Student Alcohol Consumption’ cette page présente un ensemble de graphiques (bar charts et jauge) permettant d’observer les notes des élèves selon différents critères. Les bar chart nous permettent de visualiser les notes selon des catégories (intervalles de notes) tandis que les jauge vont nous permettre d’observer une note moyenne selon le critère choisi.

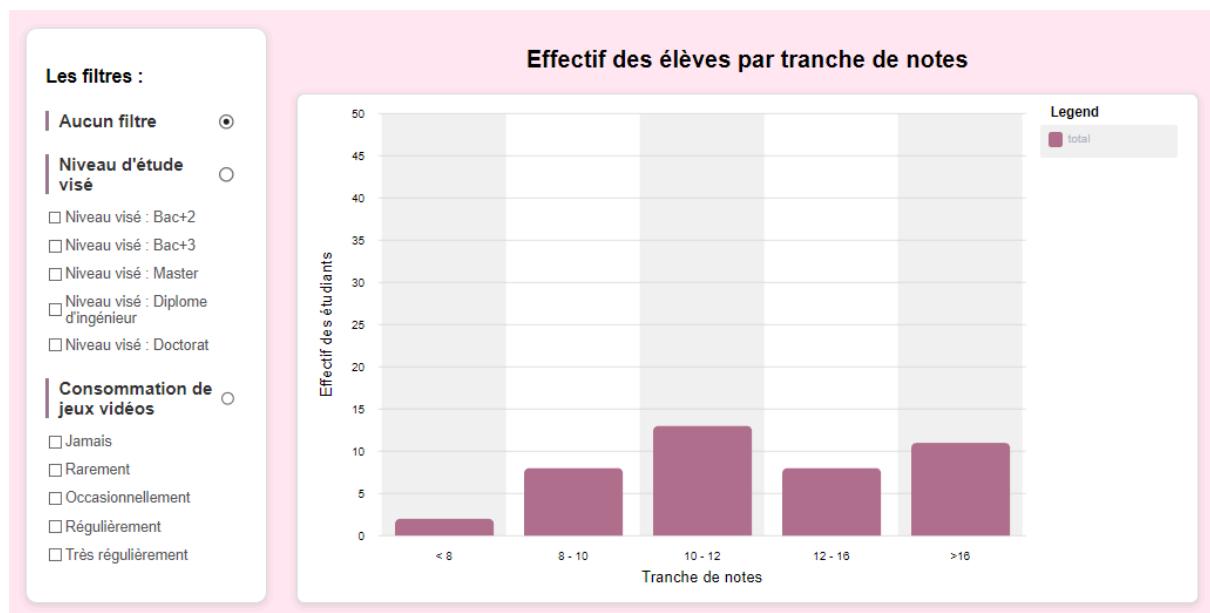


Figure n°14 : Filtre ‘Jauge temps de révision’

Ce diagramme présente l'effectif des élèves par tranche de notes. L'utilisateur va pouvoir affiner les données via un filtre qu'il peut sélectionner dans le menu de gauche. Ici aucun filtre n'est sélectionné, ainsi le bar chart présente simplement les effectifs complets par tranche de note.

Le bar char permettra de voir la répartition des étudiants par tranches de notes selon différents critères :

- Niveau d'étude visé
- Consommation de jeux vidéos

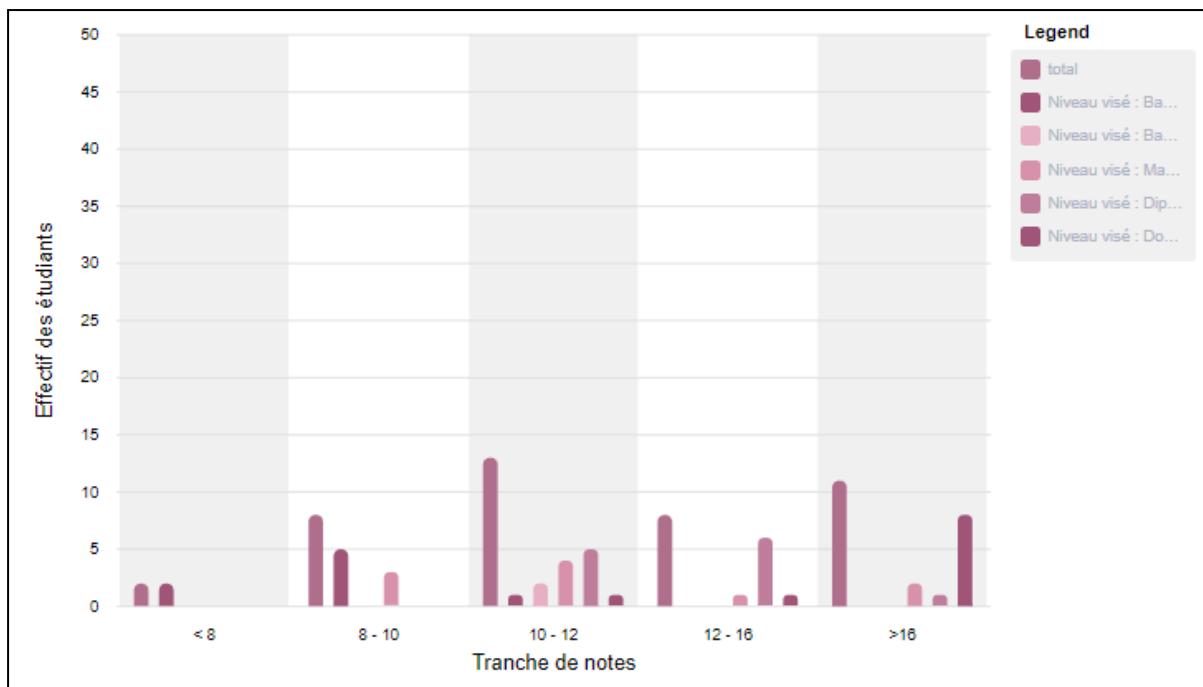


Figure n°15 : Filtre: ‘Niveau d’étude visé’

Ce diagramme présente l'effectif des élèves par tranche de notes. L'utilisateur va pouvoir affiner les données via un filtre qu'il peut sélectionner dans le menu de gauche.

Ici le filtre sélectionné est le filtre ‘Niveau d’étude visé’.

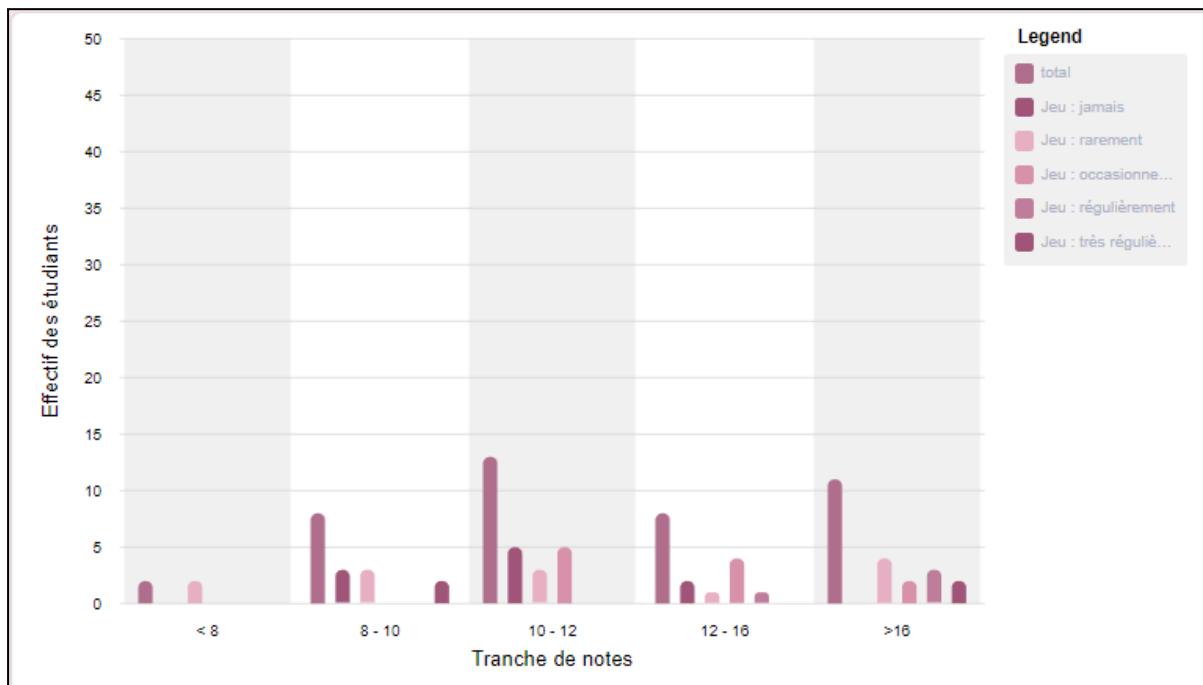


Figure n°16 : Filtre: ‘Consommation de jeux vidéos’

Ce diagramme présente l'effectif des élèves par tranche de notes. L'utilisateur va pouvoir affiner les données via un filtre qu'il peut sélectionner dans le menu de gauche.

Ici le filtre sélectionné est le filtre ‘Consommation de jeux vidéo’. Le graphique affichera donc des données par tranche de notes en fonction de sa consommation de jeux vidéos.



Figure n°17 : boutons des moyennes

Lorsque vous scrollez vers le bas de la page vous arriverez sur un second graphique. Il s’agit d’une jauge permettant d’observer la note moyenne (moyenne des 3 à 6 notes fournies par individu) des étudiants. Vous pourrez visualiser une moyenne minimale et une moyenne maximale sur 2 points :

- Moyenne générale
- Anglais

En dessous du graphique vous pouvez observer 3 boutons :

- Temps de jeu
- Temps de loisirs
- Autres

Ce sont des ‘filtres’ permettant de visualiser les moyennes en fonction du bouton sélectionné. Vous pourrez observer qu’il y'a 2 jauge de résultats, la jauge la plus petite (jaune) représente la moyenne minimale tandis que la jauge plus grande en rose Mountbatten représente la moyenne haute

Pour le temps de jeu plusieurs crans ont été implémentés concernant le temps de jeu, avec de la gauche vers la droite :

- Ne joue pas
- > 1 heure
- > 2 heures
- > [3-9] heures
- > 10 heures

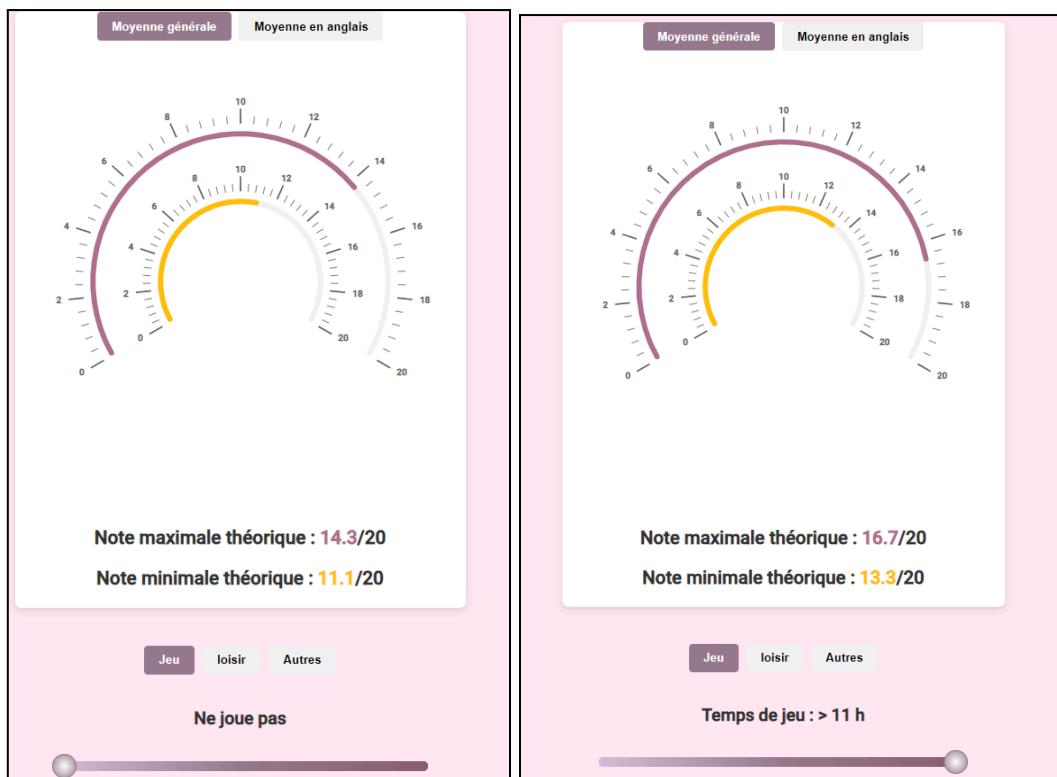


Figure n°18 : Filtre ‘Jauge temps de jeu moyenne générale’

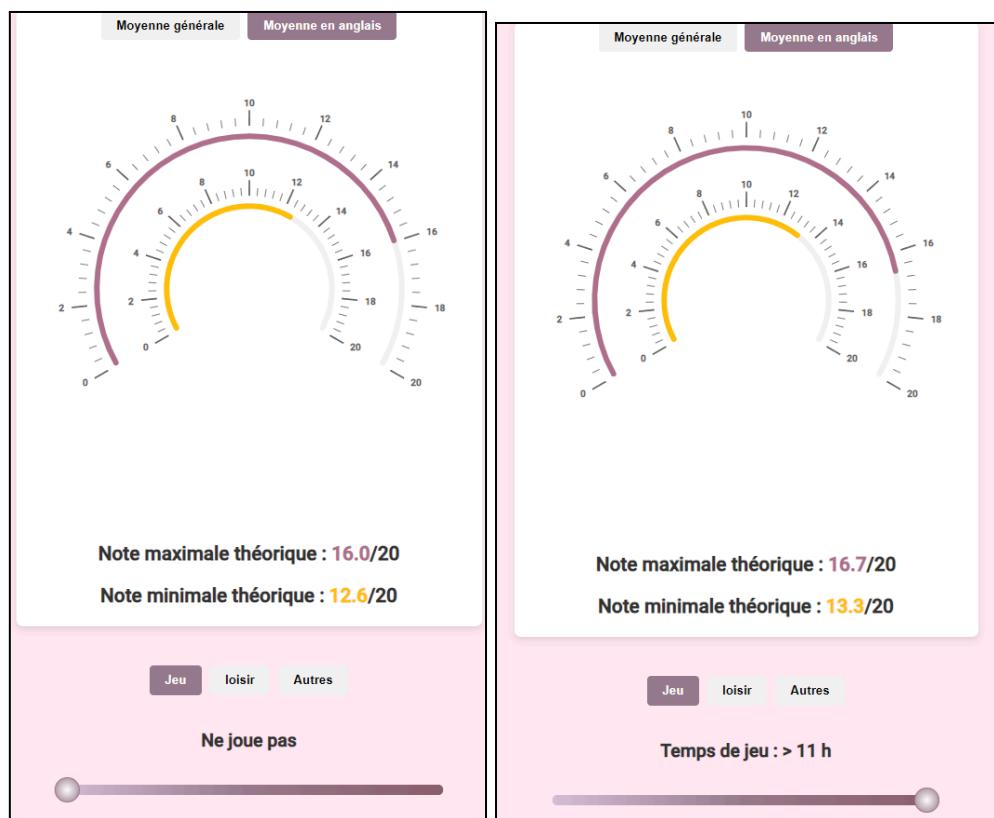


Figure n°19 : Filtre ‘Jauge temps de jeu moyenne anglais’

Pour le temps de loisirs plusieurs crans ont été implémentés concernant le temps de jeu, avec de la gauche vers la droite :

- < 1 heure
- > 1 heure
- > 2 heures
- > [3-9] heures
- > 10 heures
- 

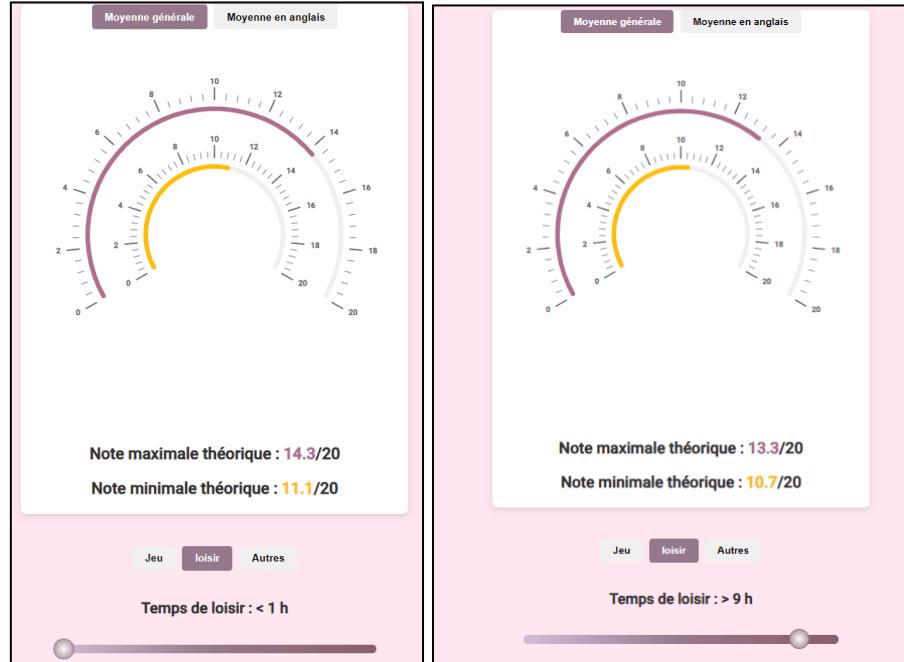


Figure n°20 : Filtre ‘Jauge temps de loisirs moyenne générale’

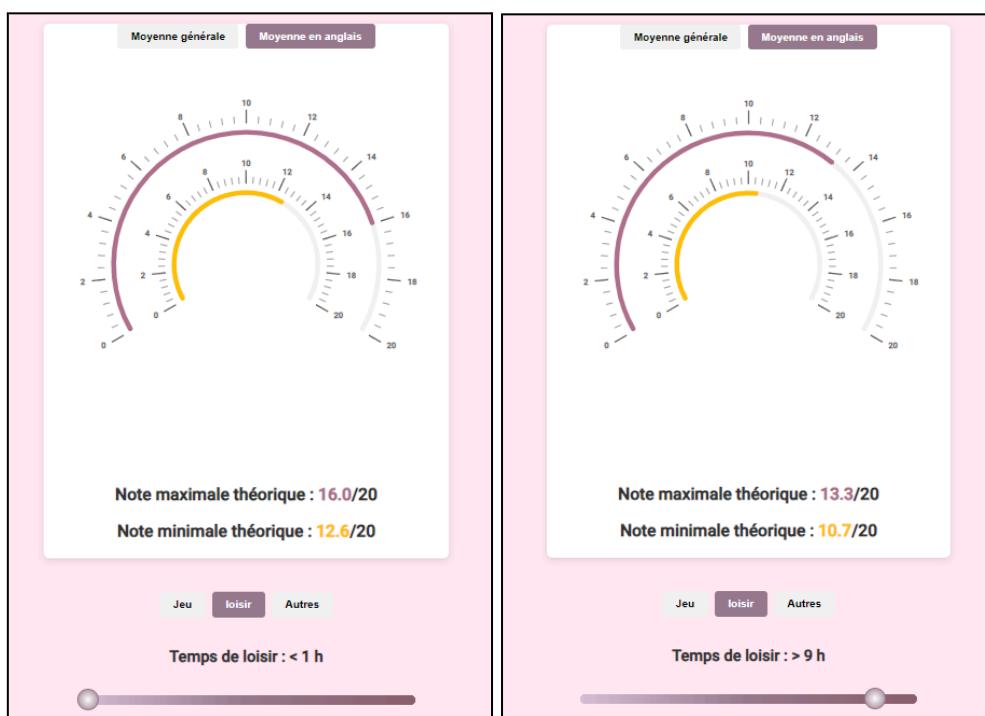


Figure n°21 : Filtre ‘Jauge temps de loisirs moyenne anglais’

Pour le temps de loisirs plusieurs crans ont été implémentés concernant le temps de jeu, avec de la gauche vers la droite :

- < 1 heure
- > 1 heure
- > 2 heures
- > [3-7] heures
- > 8

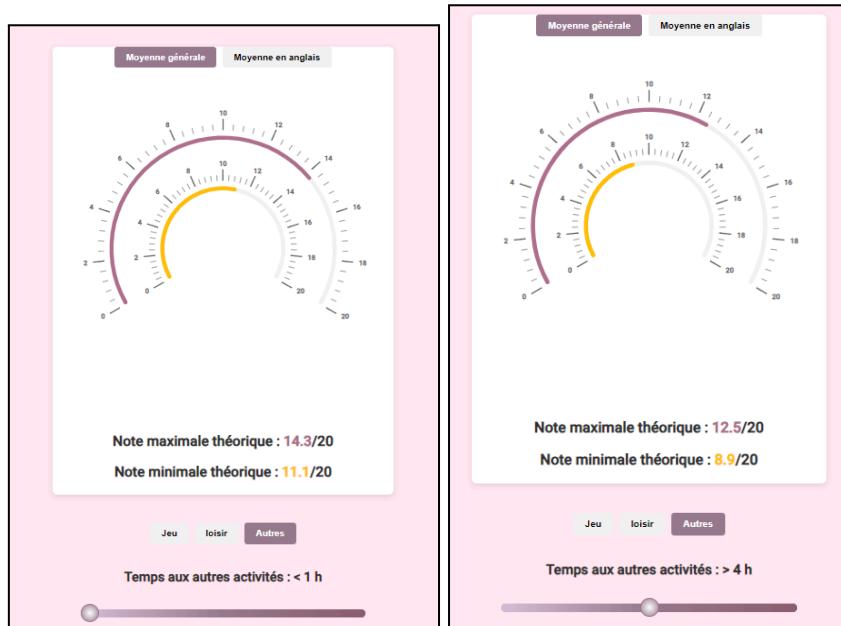


Figure n°22 : Filtre ‘Jauge temps autres moyenne générale’

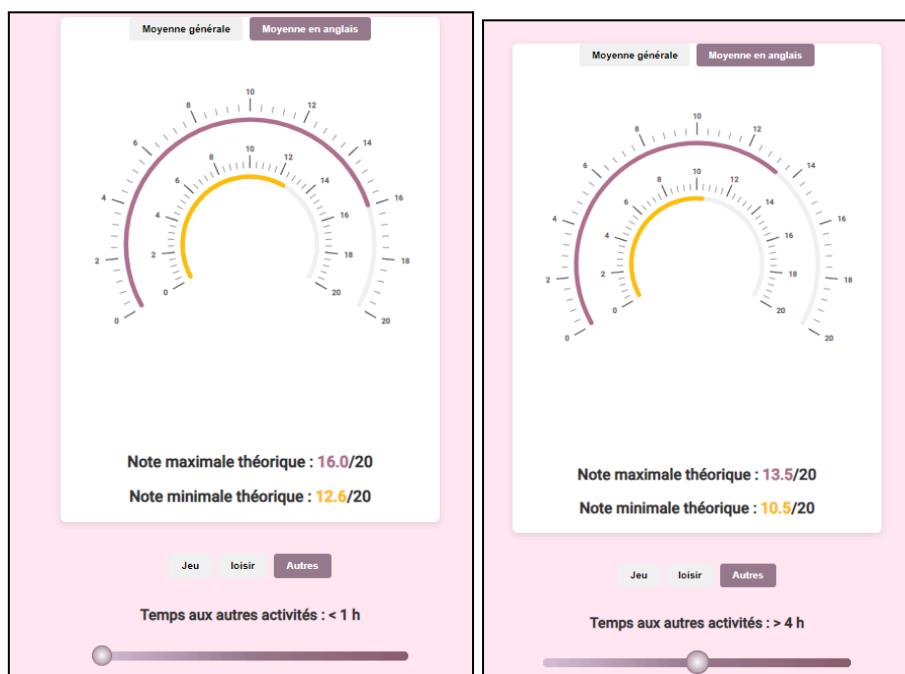


Figure n°23 : Filtre ‘Jauge temps autres moyenne anglais’

## 3 - Technologies utilisées

Pour le front nous avons utilisé la technologie *Angular* qui a été développée et est maintenue par Google, est un framework front-end robuste et complet conçu pour créer des applications web dynamiques et interactives. De plus, cette technologie est maîtrisée par notre équipe.

Pour le back nous avons utilisé la technologie *Flask* qui est un framework Python permettant de créer simplement des API rest.

Via Flask nous avons réussi à nous connecter aux deux bases de données (une pour l'étude Student Alcohol Consumption, l'autre sur l'étude des jeux vidéos et des performances scolaires )

## 4 - Charte graphique

Bien que notre interface web soit avant tout là pour présenter des données, nous avions envie que le design de notre interface soit épuré, lisible et esthétique pour cela nous avions comme couleur principale de notre site le rose Mountbatten.

Nous avons donc mis en place une charte graphique permettant de faire ressortir cette couleur

### Palette de couleurs

- Rose Mountbatten (#997a8d) :
  - Utilisée comme couleur principale du site web
- Rose pâle (#f7d7e0) :
  - Utilisée pour le fond principal des pages.
- Rose foncé(#d9b8c6) :
  - Utilisée pour le hover des cards de la page d'accueil.
- Gris foncé (#333333) :
  - Utilisé pour les titres et le texte principal.
- Gris clair (#555555) :
  - Utilisé pour les sous-titres, légendes ou détails secondaires.
- Blanc (#ffffff)

### Typographie

- Georgia Serif
- Arial Sans Serif

## 5 - Hébergement AWS Front

Dans le cadre du projet, l'hébergement de notre site web était requis.

Nous avons donc essayé de l'héberger sur plusieurs hébergeurs Firebase, Vercel mais sans grande conclusion.

En nous renseignant plus en détail sur les hébergeurs, nous sommes tombés sur AWS, le service d'hébergement d' Amazon.

Afin d'héberger son site web sur aws, il faut tout d'abord se créer un compte.

Une fois le compte créé, il faut :

- Se logger dans la console puis cliquer sur S3.
- Cliquer sur create bucket → entrer un nom
- DÉcocher les checkbox “block public access” (lorsque vous scrollez)
- Cliquer sur “Create bucket”
- Cliquer sur le bucket que l'on vient de créer
- Cliquer sur Permissions y entrer un bucket policies

```
{  
  "Version": "2012-10-17",  
  "Statement": [  
    {  
      "Sid": "PublicReadGetObject",  
      "Effect": "Allow",  
      "Principal": "*",  
      "Action": "s3:GetObject",  
      "Resource": "arn:aws:s3:::sae-s5/*"  
    }  
  ]  
}
```

Figure n°24: ‘Bucket policies’

Au niveau de Angular il faut créer un dossier dist:

- Dans le terminal taper la commande **ng build** (cela va créer les dossiers dont nous avons besoin pour héberger notre site).

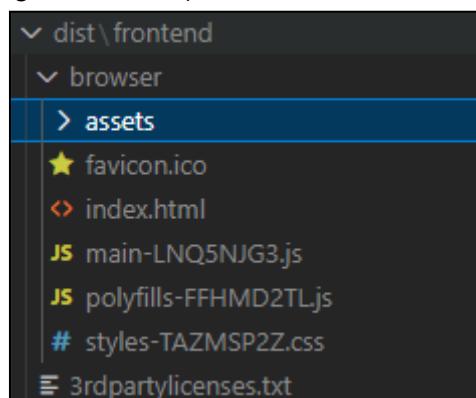


Figure n°25 : Dossier distant’

De retour sur AWS

- Cliquer sur *Upload*
- Insérer tous les documents qui ont été mis dans le dossier dist

Une fois les fichiers mis, le lien du site web est disponible dans Properties → scroller tout en bas.

*Lien du site web :*

<http://sae-s5.s3-website.eu-north-1.amazonaws.com>

NB : Le site qui est hébergé est un site vitrine. Vous ne pourrez pas visualiser les données car elles n'ont pas été hébergées en public (la base de données et le backend).

## 6 - Lancer le projet en local

Récuperer le projet github : <https://github.com/dandriambala/SAE-S5-Parcours-C.git>

Pour le frontend :

- Se déplacer dans le dossier frontend → cd src/frontend
- Installer le projet (installation des dépendances angular) → npm install
- Lancer un serveur local → ng serve
- Le serveur s'ouvrira au serveur local suivant → <http://localhost:4200/>

Pour le backend :

- Se déplacer dans le dossier backend → cd src/backend
- Installation des dépendances / librairies python qui sont dans requirements.txt → pip install -r requirements.txt
- Une fois les dépendances installées → flask run ou python3 app.py

Ces étapes peuvent être trouvées sur le Readme du proje.

## VIII. Conclusion

Les analyses de l'étude *Student Alcohol Consumption* et de notre propre recherche ont permis de mettre en lumière des liens intéressants entre différents paramètres dont certains pouvant avoir un impact potentiel sur les résultats scolaires. Toutefois, il est essentiel de souligner que ces résultats doivent être interprétés avec prudence. Les corrélations observées ne traduisent pas nécessairement des relations de causalité, et il est probable que d'autres facteurs non pris en compte dans notre étude interviennent également.

En explorant les facteurs influençant les performances scolaires, notre recherche a permis d'identifier plusieurs éléments significatifs. Notamment entre le niveau d'études des parents. Ces résultats corroborent les conclusions de la littérature existante : des parents ayant un niveau d'éducation élevé semblent mieux à même d'accompagner leurs enfants dans leur parcours scolaire, en leur offrant un soutien plus efficace. En outre, le niveau d'études visé par les étudiants semble aussi avoir une influence. Cela pourrait s'expliquer par le fait que cette variable reflète une motivation intrinsèque qui peut jouer un rôle crucial.

En ce qui concerne les jeux vidéo, nous avons observé une corrélation négative entre le temps de jeu et le temps d'étude : plus le temps passé à jouer est élevé, moins les étudiants semblent consacrer de temps à leurs études. Cependant, le temps de révision ne s'est pas révélé être un facteur influençant les notes. De plus, cet impact du temps de jeu sur le temps de révision doit être relativisé, car des corrélations similaires ont été trouvées entre les performances académiques et le temps alloué à d'autres activités extrascolaires, comme les autres loisirs.

Par ailleurs, le temps moyen consacré aux jeux vidéo, en semaine comme le week-end, n'a pas montré de lien clair avec la moyenne générale des étudiants, il semble donc que le temps passé à jouer n'ait pas d'impact sur les notes.

Enfin, ni le type de jeu vidéo ni la pratique de jeux vidéo violents n'ont montré d'impact significatif sur les performances scolaires dans le cadre de notre étude.

Concernant la consommation d'alcool, notre analyse a révélé une augmentation notable de cette dernière durant le week-end. Cependant, aucun lien clair n'a pu être établi entre la consommation d'alcool et les performances académiques.

Concernant les comportements d'addiction, nos analyses n'ont pas permis de tirer une conclusion sur leur impact sur les performances scolaires. Dans notre échantillon, les notes ne sont pas influencées par le fait d'être addictive ou pas. De plus, il est important de rappeler que l'évaluation de l'addiction était une évaluation subjective via la question "Vous considérez-vous addictive à" et il est probable qu'une partie de l'échantillon a eu tendance à minimiser certaines addictions. En effet, une grande partie des personnes interrogées avaient répondu "Aucune addiction".

En dehors des performances académiques, notre recherche a mis en évidence des tendances dans les habitudes de vie des étudiants interrogés. Par exemple, nous avons relevé une corrélation positive entre le temps consacré aux jeux vidéo et celui consacré aux loisirs en général. De plus, des liens ont été identifiés entre les absences, les retards, le fait

de rendre des devoirs en retard, et la présence de moyennes en dessous de 10. Ces éléments pourraient indiquer un désengagement global de certains étudiants, dont les causes mériteraient d'être explorées plus en profondeur car d'autres facteurs pourraient également intervenir.

Finalement, cette recherche a souligné les défis liés à l'analyse statistique sur un échantillon relativement restreint et hétérogène. Une meilleure maîtrise de l'échantillonnage, tant en termes d'effectif que d'homogénéité, ou une problématique plus ciblée, aurait probablement permis des conclusions plus robustes et nuancées. Dans notre cas, les corrélations identifiées doivent être interprétées avec prudence, car d'autres variables non mesurées dans notre étude pourraient jouer un rôle déterminant et mériteraient d'être explorées dans des recherches futures. En outre, il est à noter que les objets étudiés dans ce projet sont de nature sociale. Contrairement aux sciences naturelles et physiques, où les phénomènes sont souvent plus quantifiables et mesurables, les phénomènes sociaux sont complexes et plurifactoriels. Il nous apparaît que les outils statistiques et d'analyse de données ne semblent pas suffisants pour pleinement apprécier et interpréter nos résultats; un apport des sciences sociales aurait sans doute été grandement bénéfique.

En conclusion, nous n'avons pas établi de lien précis entre la consommation de jeux vidéo ou différentes dépendances et les performances scolaires. Cependant, il apparaît tout de même essentiel de sensibiliser les étudiants à une gestion équilibrée de leur temps entre études et activités personnelles pour favoriser leur réussite scolaire. Par ailleurs, il serait également pertinent de sensibiliser le personnel éducatif à l'importance du soutien scolaire pour certains groupes d'élèves.