CERGY PARIS UNIVERSITÉ

MOOC

Analyse de l'engagement des apprenants dans les MOOC en fonction du genre et de l'indice de développement humain

MAURAN Marion

Diplôme universitaire

Analyste de données

2024 - 2025

Résumé

Cette étude a exploré les facteurs influençant l'engagement des apprenants dans un MOOC, en se concentrant sur la consommation de vidéos, le genre, l'IDH et la réussite académique. Une analyse quantitative a été réalisée à l'aide de tests de corrélation et d'analyses de survie pour examiner ces relations. Les résultats ont montré que les femmes consommaient légèrement plus de vidéos que les hommes, et que l'engagement augmentait avec le niveau d'IDH. Toutefois, ni le genre ni l'IDH n'ont eu d'effet significatif sur la réussite académique, ce qui suggère que des facteurs individuels, comme la motivation, pourraient être déterminants. L'étude a également révélé une forte association entre la réalisation de quiz et la consommation de vidéos, mettant en avant l'importance d'une approche intégrée dans les MOOC. Ces résultats indiquent que l'engagement des apprenants est multifactoriel et que des éléments personnels jouent un rôle clé dans leur interaction avec le contenu éducatif.

Summary

This study explored factors influencing learner engagement in a MOOC, focusing on video consumption, gender, HDI and academic achievement. Quantitative analysis was conducted using correlation tests and survival analyses to examine these relationships. The results revealed that women consumed slightly more videos than men, and that engagement increased with HDI level. However, neither gender nor HDI showed a significant effect on academic success, suggesting that other individual factors, such as motivation, may be determinant. The study also found a strong association between quiz completion and video consumption, underlining the importance of an integrated approach in MOOCs. These results indicate that learner engagement is multifactorial, and that personal elements play a key role in their interaction with educational content.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

1.1 Engagement des apprenants dans les MOOC

Les MOOC (Massive Open Online Courses) ont transformé l'accès à l'éducation en offrant une meilleure flexibilité d'apprentissage et en rendant l'enseignement supérieur plus accessible.

Cependant, cette démocratisation soulève des problématiques sociétales importantes, notamment en ce qui concerne les facteurs influençant l'engagement des apprenants.

Certaines études suggèrent des disparités en fonction du genre, du statut socio-économique ou de l'accès à la technologie, mais leur impact exact sur la participation et le taux d'abandon reste encore à évaluer ([1] Kizilcec, Piech, & Schneider, 2013).

Il est donc essentiel d'examiner dans quelle mesure ces facteurs influencent les différents types d'apprenants et de déterminer comment les concepteurs de MOOC peuvent adapter leurs cours pour favoriser une plus grande inclusion.

1.2 Facteurs influençant l'implication des apprenants

Il est essentiel de mieux comprendre comment les facteurs socio-économiques et de genre influencent l'engagement et la réussite des apprenants dans les MOOC. Bien que des recherches aient permis de catégoriser les apprenants selon leur niveau d'interaction avec le cours, il reste encore des lacunes concernant l'impact de variables telles que le statut socio-économique, le genre ou l'accès à la technologie sur ces typologies.

Sans une telle compréhension, il devient difficile d'adapter les contenus et les méthodes pédagogiques pour réduire les inégalités d'accès et d'engagement, limitant ainsi le potentiel des MOOC à démocratiser l'éducation. Ce manque de connaissance pourrait également accentuer les disparités sociales existantes, empêchant une large partie de la population de tirer pleinement parti des opportunités d'apprentissage offertes par cette modalité.

1.3 Lien entre IDH, genre et engagement dans les MOOC

La question centrale de cette étude porte sur l'exploration de l'Indice de Développement Humain (IDH) et de genre dans le contexte des MOOC.

Plus précisément, il s'agit de comprendre comment ces facteurs influencent l'engagement, la progression et la complétion des apprenants dans ces cours en ligne.

Nous chercherons à déterminer quelles variables socio-démographiques peuvent expliquer les différentes typologies d'apprenants, telles que les apprenants ayant complété le cours, les "désengagés", les "auditeurs" et les "spectateurs".

La question à laquelle nous allons tenter de répondre est par conséquent la suivante : "*Comment l'Indice de Développement Humain (IDH) et le genre influencent-ils l'engagement, la progression et la complétion des apprenants dans les MOOC ?*"

1.4 Objectifs et cadre de l'étude

L'étude utilise une approche quantitative, s'appuyant sur des données collectées à partir des rapports d'activité des étudiants, des carnets de notes et des réponses aux enquêtes.

Les analyses portent sur la consommation de vidéos, les soumissions de quiz et les données démographiques (IDH et genre).

Ces données sont traitées à l'aide d'outils statistiques pour explorer les relations entre les variables socio-économiques, le genre et les types d'engagement des apprenants dans les MOOC.

1.5 Hypothèses et approche de recherche

Deux hypothèses principales ont été formulées.

La première postule que les apprenants issus de pays avec un IDH élevé seront plus engagés et compléteront davantage les MOOC, en raison de l'accès plus facile à la technologie et aux ressources éducatives.

La deuxième hypothèse suggère que des différences notables existent entre les hommes et les femmes en termes d'engagement dans les MOOC, probablement influencées par des facteurs socio-culturels et des stéréotypes de genre.

Ces hypothèses sont en adéquation avec des résultats précédents qui montrent une relation entre le statut socio-économique, le genre et les comportements d'engagement en ligne ([2] Ho et al., 2014).

Des méthodes statistiques et analytiques seront appliquées pour tester ces hypothèses, en analysant les données d'engagement et les profils des apprenants à travers les différentes itérations des MOOC.

2 Méthodes et données

2.1 Source des données

Les données utilisées proviennent du MOOC *Effectuation*, un cours de cinq semaines en entrepreneuriat proposé par l'EMLYON Business School.

Nous avons analysé les données de trois itérations successives de ce MOOC, avec un total de 8 996 apprenants inscrits lors de la première itération, 6 200 lors de la deuxième, et 4 236 lors de la troisième. Ce cours a été hébergé sur la plateforme Canvas.

La fiabilité des données repose sur un suivi détaillé des activités des apprenants sur la plateforme Canvas, ainsi que sur la collecte systématique d'informations via des enquêtes administrées à des échantillons représentatifs des inscrits, assurant ainsi la validité et la représentativité des résultats.

L'analyse repose sur des variables catégorielles comme le genre (masculin, féminin) et l'IDH des pays d'origine des apprenants (très élevé, intermédiaire, faible). Elle se base également sur des variables quantitatives, telles que le nombre de vidéos regardées et de quiz soumis.

TABLE 1 – Distribution des apprenants selon l'IDH et le genre

IDH	Homme (nb)	Homme (%)	Femme (nb)	Femme (%)
B (Bas)	883	14.7%	147	5.0%
I (Intermédiaire)	432	7.2%	233	8.0%
TH (Très Haut)	4716	78.2%	2546	87.0%
Total	6031	100%	3926	100%

Une Table de contingence (Table 1) a été réalisée pour analyser la répartition des apprenants selon leur IDH et leur genre, montrant des différences notables dans les comportements d'engagement. La Figure 1 montre la répartition des apprenants en fonction de leur genre et de l'IDH de leur pays d'origine.

Les résultats du test du chi² ($\chi^2(2, N=10000) = 178$, $p < 0.001$) montrent une statistique chi² de 178 et une p-value < 0.001, bien inférieure au seuil de significativité habituel de 0.05. Cela nous permet de rejeter l'hypothèse nulle selon laquelle les deux variables (genre et IDH) sont indépendantes, ce qui indique qu'il existe une relation statistiquement significative entre le genre et l'IDH.

Le V de Cramer, calculé à 0.14, suggère une association modérée entre le genre et l'IDH. Cette métrique quantifie l'intensité de la relation : une valeur proche de 0 indique une faible association, tandis qu'une valeur proche de 1 indique une forte association. Un V de Cramer de 0.14 indique donc qu'il existe des différences suffisamment significatives dans la répartition des sexes selon l'IDH.

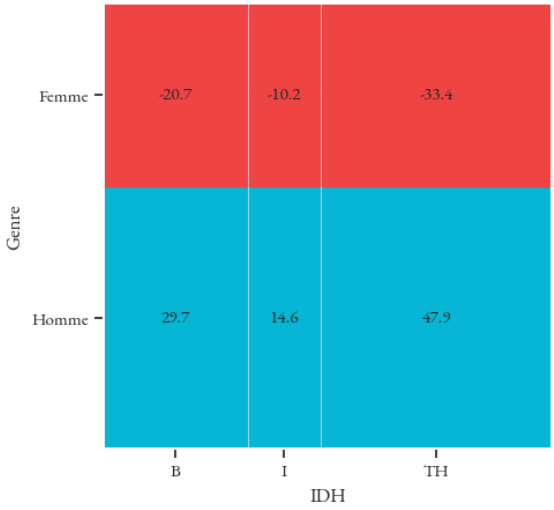

FIGURE 1 – Distribution des apprenants selon l'IDH et le genre.

Concernant les résidus, ceux-ci montrent que les hommes sont surreprésentés dans les groupes B, I et TH, tandis que les femmes sont sous-représentées par rapport aux valeurs attendues. Les écarts sont particulièrement marqués dans le groupe TH, où les hommes sont surreprésentés avec un excédent de 47,9 individus, tandis que les femmes sont sous-représentées avec un déficit de 33,4 individus. Ces écarts sont systématiques, avec des excédents d'hommes dans chaque groupe et des déficits de femmes, ce qui suggère une dépendance entre le genre et l'IDH.

Ces résultats pourraient être le reflet de plusieurs facteurs socio-économiques ou culturels influençant la participation à ce type de MOOC. Par exemple, il est possible que des inégalités de genre existent dans l'accès à la formation, l'intérêt pour ce type de programme, ou la disponibilité des ressources nécessaires pour suivre un MOOC sur l'*Effectuation*.

Les hommes pourraient être davantage encouragés ou avoir plus de facilités à s'engager dans ce type de formation, notamment si des stéréotypes de genre influencent leur participation à des programmes axés sur l'entrepreneuriat ou l'innovation.

À l'inverse, les femmes pourraient rencontrer des obstacles supplémentaires, comme des responsabilités familiales ou des attentes sociales qui limitent leur disponibilité ou leur accès à ce genre de cursus.

2.2 Gestion des outliers, anomalies, données manquantes

Une analyse des données manquantes a été réalisée pour chaque variable utilisée.

La Table 2 présente le nombre de valeurs manquantes par variable. Les lignes comportant des valeurs manquantes ont été supprimées, notamment pour la variable IDH.

Après cette suppression, il restait 8 924 lignes dans le jeu de données, contre 17 297 initialement.

Aucune valeur aberrante (outlier) n'a été identifiée, suite à une analyse des valeurs uniques des colonnes, des valeurs minimales et maximales, ainsi qu'à une analyse visuelle à l'aide d'une boîte à moustaches.

Aucun doublon n'a été détecté après une vérification exhaustive des lignes, confirmant ainsi l'intégrité des données sans duplications.

Ce traitement garantit une qualité élevée des données utilisées.

TABLE 2 – Taux de valeurs manquantes par variables

Variable	Nombre de valeurs manquantes
ID de l'étudiant	2 %
IDH	48 %
Genre	47 %
Vidéos vues	0 %
Quiz réalisés	0 %
Examens réalisés	12 %
Devoirs soumis	12 %
Examen fait ou devoir remis	0 %
Certification obtenue	62 %
Type d'apprenant	13 %
Iteration du MOOC	0 %

Concernant le nombre de valeurs manquantes dans la colonne « Certification obtenue », celui-ci est trop élevé pour envisager une suppression immédiate (62%). Par conséquent, lors de l'utilisation de cette variable, le tableau de données sera filtré sur les valeurs disponibles et s'appliquera à une portion plus restreinte de l'échantillon.

2.3 Techniques utilisées pour l'analyse

L'analyse des données repose sur plusieurs approches statistiques et graphiques pour explorer et modéliser les relations entre les variables.

Une première étape a consisté à appliquer un test du χ^2 pour évaluer l'indépendance entre les variables IDH et genre, accompagné d'un diagramme en mosaïque pour visualiser les résidus du test.

Ensuite, des statistiques descriptives (moyenne, médiane, écart-type) ont été calculées pour résumer les données, et des représentations graphiques comme les boîtes à moustaches et histogrammes ont été utilisées pour observer la distribution des variables.

Les modèles statistiques utilisés incluent le test de Mann-Whitney U pour comparer les comportements de consommation de vidéos entre hommes et femmes, et le test de Kruskal-Wallis pour examiner les différences en fonction de l'IDH. Le test de Spearman a mesuré la corrélation entre le nombre de vidéos vues et de quiz réalisés.

Une régression logistique a été appliquée pour étudier l'impact du genre et de l'IDH sur la probabilité d'obtenir un certificat. Enfin, un modèle de Cox a été utilisé pour analyser les comportements d'engagement vidéo en fonction du genre et de l'IDH. Ces modèles permettent de comprendre les relations entre les variables et les comportements des apprenants.

2.4 Outils utilisés

L'analyse des données a été réalisée principalement avec le langage Python.

Pour la manipulation des données et la gestion des valeurs manquantes, pandas a été utilisé, tandis que numpy a permis d'effectuer des calculs numériques.

Les visualisations graphiques ont été réalisées grâce à matplotlib et seaborn (histogrammes, boîtes à moustaches).

En ce qui concerne les analyses statistiques, scipy et statsmodels ont été utilisés pour effectuer des tests de corrélation (Pearson, Spearman), des tests d'indépendance (χ^2) et des régressions logistiques.

3 Résultats

Les résultats montrent des différences significatives dans le comportement des apprenants en fonction du genre et de l'IDH. Les hommes et les femmes présentent des habitudes de consommation de vidéos distinctes, tandis que l'IDH influence l'engagement vidéo, avec une tendance à un engagement plus homogène dans les groupes à IDH élevé. Des différences d'engagement sont également observées selon les typologies d'apprenants, avec une diminution progressive de l'engagement au fil des itérations. Les analyses supplémentaires révèlent des corrélations entre les vidéos vues et les quiz réalisés, sans effet significatif du genre ou de l'IDH sur la probabilité de réussite.

3.1 Lien entre nombre de vidéos vues et genre

Étant donné que le nombre de vidéos vues ne suit pas une distribution normale, le test paramétrique de Student n'est pas approprié. Comme l'illustre la Figure 2, les données montrent une répartition asymétrique. On observe ainsi :

— un premier pic à 2924 apprenants pour 0 à 4 vidéos visionnées,
— un fléchissement du nombre d'apprenants pour un nombre de vidéos vues compris entre 5 et 9, 10 et 14, 15 et 19, 20 et 24 (passant de 1271 à 583),
— un nouveau pic pour 25 à 29 vidéos visionnées (1363 apprenants).

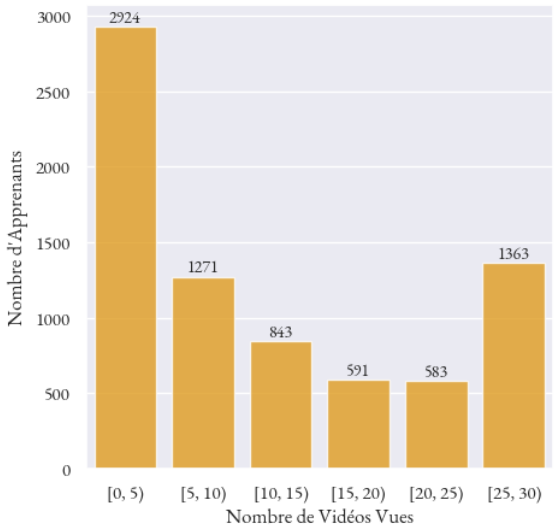

FIGURE 2 – Distribution du nombre de vidéos vues par apprenants

Par conséquent, nous avons utilisé le test non-paramétrique de Mann-Whitney U (ou test de Wilcoxon pour deux échantillons indépendants), qui ne nécessite pas de

satisfaire à l'hypothèse de normalité. Celui-ci permet de comparer les distributions des deux groupes indépendants (hommes et femmes) de manière plus fiable.

Les résultats du test de Mann-Whitney U, avec une statistique U de 8 373 795 et une p-value $< 0{,}001$, révèlent une différence statistiquement significative dans les comportements de consommation de vidéos entre les hommes et les femmes. De plus, la valeur élevée de la statistique U suggère que cette différence est marquée.

La Figure 3 illustre la différence de consommation de vidéos entre les sexes, montrant que les femmes ont tendance à visionner plus de vidéos que les hommes. En moyenne 14,5 vidéos sont vues pour les femmes contre 13,6 vidéos pour les hommes.

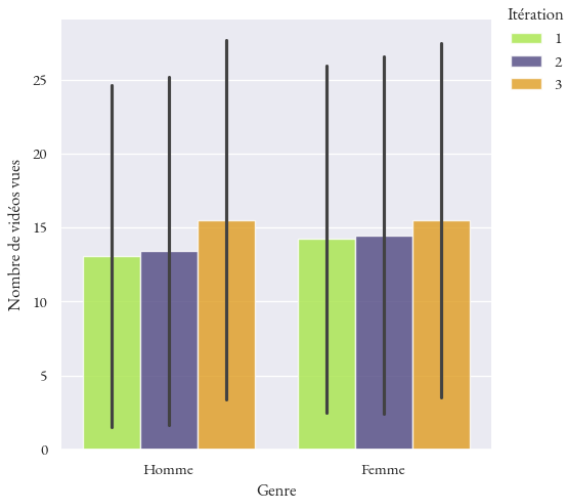

FIGURE 3 – Distribution du nombre de vidéos vues par genre et par itérations de MOOC

Par ailleurs, il est important de noter que les comportements de consommation de vidéos ont évolué au fil des itérations du MOOC. Les moyennes des vidéos vues ont augmenté de 14,2 (version 1) à 15,5 (version 3) chez les femmes, et de 13,1 (version 1) à 15,5 (version 3) chez les hommes, indiquant une tendance à l'augmentation.

Enfin, les écarts-types ont quant à eux varié entre 11,7 et 12,1 pour les femmes, et 11,5 et 12,2 pour les hommes, avec des coefficients de variation compris entre 74,3% à 88,2%. Cela révèlent qu'il existe une variabilité substantielle dans les comportements de consommation des vidéos parmi les apprenants.

3.2 Lien entre nombre de vidéos vues et IDH

Étant donné que les données ne suivent pas une distribution normale, nous avons utilisé un test non-paramétrique, le test de Kruskal-Wallis qui est adapté lorsque plusieurs groupes sont à comparer.

Le test de Kruskal-Wallis a révélé une différence statistiquement significative dans le nombre de vidéos vues en fonction de l'IDH ($H(2) = 556$, $p < 0{,}001$).

La Figure 4 illustre cette tendance en montrant que la moyenne des vidéos vues augmente avec le niveau d'IDH. En effet, les participants issus des groupes à faible IDH ont visionné en moyenne 6,5 vidéos, tandis que ce chiffre passe à 10,8 pour le groupe intermédiaire, et atteint 15,2 pour ceux ayant un IDH très élevé.

Les coefficients de variation diminuent à mesure que le niveau d'IDH augmente, passant de 145 pour le groupe B à 72 pour le groupe TH, ce qui suggère une plus grande homogénéité dans les comportements au sein des groupes ayant un IDH plus élevé. Les écarts-types augmentent au fil des itérations dans tous les groupes IDH :

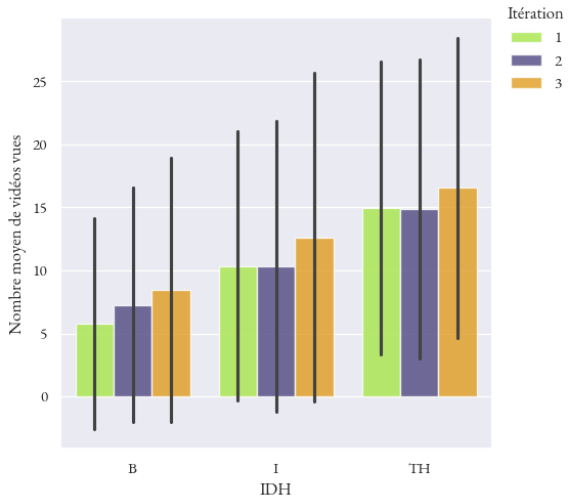

FIGURE 4 – Distribution du nombre de vidéos vues selon l'IDH et l'itération du MOOC

de 8,4 à 10,5 pour le groupe B, de 10,7 à 13,0 pour le groupe I, et de 11,6 à 11,9 pour le groupe TH. Cela indique une augmentation de la variabilité dans la consommation de vidéos à mesure que les itérations du MOOC progressent.

De manière générale, les comportements de consommation des vidéos ont évolué au fil des itérations du MOOC. On observe une augmentation des moyennes de vidéos vues, passant de 5,8 (version 1) à 8,5 (version 3) dans le groupe à faible IDH, et de 15,0 (version 1) à 16,6 (version 3) dans le groupe à très haut IDH. Cette tendance suggère un accroissement global de la consommation de vidéos au fil des versions du MOOC.

3.3 Typologies d'apprenants

Les comportements des apprenants varient en fonction du genre, de l'IDH et de l'itération, ce qui permet d'établir une typologie d'engagement avec quatre catégories : auditeurs, spectateurs, compléteurs et désengagés.

TABLE 3 – Distribution des typologies d'apprenants

Typologie	Itération 1	Itération 2	Itération 3	Total	Taux
Auditeur	74	45	33	152	1.7%
Spectateur	936	469	235	1640	18.4%
Compléteur	2378	951	854	4183	46.9%
Désengagé	1836	681	432	2949	33.0%
Total	5224	2146	1554	8924	100.0%

Selon la Table 3, les auditeurs, représentant 1,7 % des apprenants, voient leur nombre diminuer au fil des itérations, passant de 74 à 33. Leur proportion reste stable au fil des itérations.

Les spectateurs, qui composent 18,4 % du total, présentent également une baisse de participation, de 936 à 235. Leur proportion diminue entre l'itération 2 et 3 (passant de 22% à 15%).

Les compléteurs, représentant la majorité avec 46,9 %, connaissent aussi une diminution, de 2378 à 854 au fil des itérations. Leur proportion augmente entre l'itération 1 et 3 (+10 points, 55% de l'échantillon).

Enfin, les désengagés, représentant 33 % des apprenants, subissent un déclin marqué, passant de 1836 à 432 entre la première et la troisième itération. Leur proportion diminue également passant de 35% à 28% de l'échantillon.

De manière générale, le nombre d'apprenants a fortement diminué entre l'itération 1 (5224) et l'itération 3 (1554), avec une baisse de 70 pts.

3.4 Lien entre quiz réalisés et nombre de vidéos vues

Après avoir analysé la typologie des apprenants, il est essentiel d'examiner la relation entre le nombre de quiz réalisés et le nombre de vidéos visionnées.

Étant donné que les tests de corrélation de Pearson et la régression linéaire supposent que les données sont normalement distribuées, il est préférable d'utiliser un test de corrélation non paramétrique, comme le test de Spearman, qui ne repose pas sur l'hypothèse de normalité. La corrélation de Spearman mesure la relation entre les variables, indépendamment de leur distribution.

En effectuant le test de corrélation de Spearman, nous obtenons un coefficient de corrélation de 0,71, ce qui indique une relation positive modérée à forte entre le nombre de quiz réalisés et le nombre de vidéos vues. Ce coefficient suggère que, à mesure que le nombre de quiz réalisés augmente, le nombre de vidéos vues tend également à augmenter, bien que cette relation ne soit pas nécessairement linéaire.

La p-value associée à ce test est $< 0{,}001$, ce qui est largement inférieur au seuil de signification habituel de 0,05. Cela signifie que la relation observée entre ces deux variables est hautement significative et qu'il est très improbable que cette relation

soit due au hasard.

Ainsi, même en l'absence de normalité des données, le test de la corrélation de Spearman indique une forte association positive entre le nombre de quiz réalisés et le nombre de vidéos vues.

3.5 Effet du genre et de l'IDH sur la probabilité de réussite

L'objectif de la régression logistique est ici d'examiner l'impact de certaines variables explicatives (genre et IDH) sur une variable binaire représentant l'obtention du certificat et/ou la réalisation de l'examen final.

La régression logistique a été utilisée pour estimer les odds ratios (OR) des différentes modalités de ces variables, sans prendre en compte les interactions entre genre et IDH.

Table 4 – Odds ratios

Variable	OR	CI 95% min	CI 95% max	p-value
Réf. (Femme & IDH B)	2.04	1.47	2.83	< 0.001
Genre (Homme)	1.15	0.96	1.38	0.12
IDH (I)	0.95	0.61	1.48	0.82
IDH (TH)	0.97	0.71	1.32	0.82

Les résultats de la Table 4 montrent que, par rapport au groupe de référence (femmes et IDH Bas), le genre et les différentes catégories d'IDH n'ont pas d'effet significatif sur l'obtention du certificat et/ou de l'examen final.

Pour les hommes, l'odd ratio est de 1,15 (IC 95% : [0,96, 1,38], p = 0,12), indiquant une légère augmentation des chances d'obtenir le certificat et/ou l'examen final, mais cette différence n'est pas statistiquement significative, car la p-value est supérieure à 0,05. Il n'y a donc pas de preuve suffisante pour conclure à un effet du genre sur l'obtention du certificat.

Concernant l'IDH, les résultats montrent que l'odd ratio pour un IDH Intermédiaire est de 0,95 (IC 95% : [0,61, 1,48], p = 0,82), ce qui suggère une légère réduction des chances d'obtenir le certificat et/ou l'examen final, mais cette relation n'est pas significative en raison d'une p-value élevée.

De même, pour un IDH Très haut, l'odd ratio est de 0,97 (IC 95% : [0,71, 1,32], p = 0,82), suggère une légère diminution des chances d'obtenir le certificat et/ou l'examen final, sans effet significatif, puisque la p-value est également supérieure à 0,05.

3.6 Proportion d'apprenants selon le nombre de vidéos vues, le genre et l'IDH

La modélisation par Poisson, bien qu'initialement pertinente, est remise en cause par la distribution des données. Un modèle de Cox, basé sur l'analyse de survie, apparaît alors comme une méthode plus adaptée.

La Figure 5 montre que la proportion de femmes ayant visionné des vidéos est légèrement plus élevée que celle des hommes, avec une moyenne de 14,5 vidéos contre 13,6 pour les hommes.

Cette proportion décroît progressivement entre 2 et 28 vidéos vues, avec trois baisses notables :

— entre 0 et 1 vidéo,
— entre 4 et 5 vidéos,
— et entre 28 et 30 vidéos vues.

Ces variations suggèrent des comportements différents d'engagement selon le nombre de vidéos visionnées, quel que soit le genre.

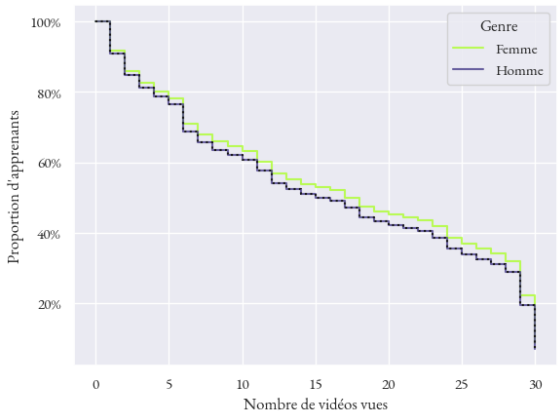

FIGURE 5 – Évolution de la proportion d'apprenants en fonction du nombre de vidéos vues et du genre

La Figure 6 montre que le nombre moyen de vidéos vues varie selon le groupe d'IDH. Les apprenants du groupe à faible IDH (B) ont en moyenne regardé 6,5 vidéos, tandis que ceux du groupe à IDH intermédiaire (I) ont visionné en moyenne 10,8 vidéos, et ceux du groupe à très haut IDH (TH) ont regardé en moyenne 15,2 vidéos. Cette tendance suggère que l'engagement vidéo augmente avec le niveau d'IDH. Cette proportion décroît progressivement entre 2 et 28 vidéos vues, avec trois baisses notables :

— entre 0 et 1 vidéo,
— entre 4 et 5 vidéos,
— et entre 28 et 30 vidéos vues.

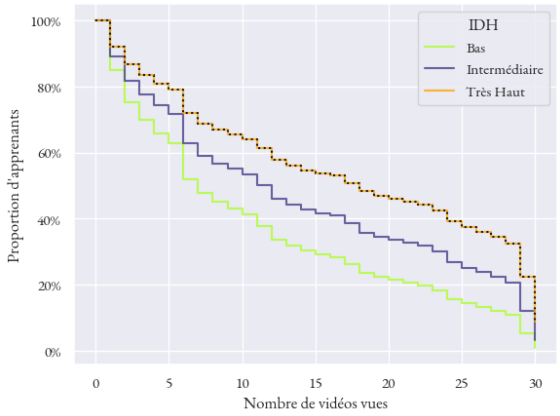

FIGURE 6 – Évolution de la proportion d'apprenants en fonction du nombre de vidéos vues et de l'IDH

Ces variations suggèrent des comportements différents d'engagement selon le nombre de vidéos visionnées, quel que soit l'IDH.

Les analyses des hazard ratios (HR) ont révélé des différences significatives dans la consommation de vidéos entre les auditeurs et les apprenants désengagés.

Pour les auditeurs, le HR était de 1,78, indiquant qu'ils consommaient plus de vidéos que les apprenants désengagés (HR = 1,66 pour les apprenants désengagés).

En revanche, les apprenants désengagés avaient un HR de 2,33, suggérant une consommation plus faible de vidéos par rapport aux auditeurs.

Les résultats ont montré une différence statistiquement significative, avec une p-value $< 0{,}001$, ce qui indique que ces différences de consommation sont significatives sur le plan statistique.

4 Discussion

L'analyse des comportements des apprenants révèle des variations importantes en fonction du genre et du niveau d'IDH, avec un engagement global croissant au fil des itérations du MOOC. Cependant, ni le genre ni l'IDH ne semblent influencer de manière significative la réussite académique, ce qui suggère que d'autres facteurs, tels que la motivation ou les stratégies d'apprentissage, jouent un rôle clé.

4.1 Interprétations

Les résultats de cette étude montrent des variations significatives dans les comportements des apprenants en fonction de facteurs tels que le genre et le niveau d'IDH. Les femmes consomment légèrement plus de vidéos que les hommes, bien que cette tendance soit marquée par une grande variabilité au sein de chaque groupe d'IDH. L'engagement général, mesuré par le nombre de vidéos vues, augmente au fur et à mesure des itérations du MOOC, ce qui suggère une évolution positive de l'engagement des apprenants, mais avec des comportements de consommation qui restent influencés par des facteurs individuels, tels que la motivation ou les préférences d'apprentissage.

Concernant l'IDH, les résultats indiquent que les apprenants avec un IDH plus élevé ont tendance à regarder davantage de vidéos, et ce comportement est plus homogène que chez les apprenants à faible IDH. Cela peut refléter des différences d'accès aux ressources ou des compétences d'apprentissage plus développées, facilitant l'engagement dans le contenu. Cependant, malgré ces variations dans la consommation des vidéos, ni le genre ni l'IDH n'ont montré d'effet significatif sur la réussite académique, mesurée par l'obtention du certificat ou la réussite de l'examen final. Cela suggère que d'autres facteurs, tels que la persévérance ou les stratégies d'apprentissage, peuvent jouer un rôle plus déterminant dans la réussite des apprenants.

Enfin, la forte association entre la réalisation de quiz et la consommation de vidéos confirme que les apprenants engagés dans une activité ont tendance à s'investir également dans l'autre. Cette interaction met en lumière l'importance d'une approche intégrée du contenu dans les MOOC, où les vidéos et les quiz se complètent pour favoriser l'apprentissage. En somme, cette étude met en évidence l'importance des facteurs individuels dans l'engagement et la réussite des apprenants, tout en soulignant que le genre et l'IDH n'expliquent pas à eux seuls la performance académique.

4.2 Conclusion

Cette étude offre des insights intéressants sur les facteurs influençant l'engagement des apprenants dans un MOOC, en explorant les relations entre la consommation de vidéos, le genre, l'IDH et la réussite académique. Toutefois, plusieurs limites doivent être prises en compte. Premièrement, la collecte de données n'a pas permis de contrôler tous les facteurs individuels susceptibles d'influencer l'engagement, comme la motivation intrinsèque des apprenants ou leur environnement social. De plus, la distribution inégale des données sur certaines variables, notamment le faible nombre d'apprenants dans certains groupes, peut avoir affecté la robustesse des résultats.

En termes de perspectives, des analyses futures pourraient explorer plus en profondeur l'interaction entre différentes variables individuelles (par exemple, la motivation ou les habitudes d'apprentissage) et leur impact sur l'engagement. De plus, une étude longitudinale permettrait d'évaluer l'évolution de l'engagement des apprenants sur plusieurs cycles de MOOC et d'analyser comment l'expérience d'apprentissage peut influencer les comportements à long terme. L'intégration d'analyses de données comportementales, telles que les patterns de navigation ou d'interaction avec le contenu, pourrait également enrichir la compréhension des mécanismes d'engagement dans un contexte d'apprentissage en ligne.

Références

[1] **Kizilcec, R. F., Piech, C., & Schneider, E. (2013).** Deconstructing disengagement : Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 170-179. https://doi.org/10.1145/2460296.2460330

[2] **Ho, A. D., Chuang, I., Reich, J., & Koller, D. (2014).** HarvardX and MITx : The first year of open online courses. *Research & Practice in Assessment,* 9(1), 23–29. https://dx.doi.org/10.2139/ssrn.2586847

A Annexes

A.1 Modèle linéaire

L'ANOVA (Analyse de la Variance) est une méthode statistique utilisée pour tester si des différences significatives existent entre les moyennes de plusieurs groupes.

Dans notre étude, l'objectif est d'évaluer l'effet de deux variables indépendantes catégorielles, à savoir l'IDH (avec ses différentes catégories) et le genre (homme vs femme), sur une variable dépendante continue, le nombre de vidéos vues.

Les résultats de l'ANOVA sans interaction (voir Table 5) montrent que l'effet de l'IDH est très significatif ($F(2, 8920) = 284$, $p < 0.001$), ce qui indique que les groupes d'IDH expliquent une grande partie de la variance.

Comparé au groupe de référence, le groupe I augmente de 4.2 unités ($p < 0.001$), et le groupe TH de 8.7 unités ($p < 0.001$), tandis que l'intercept est de 6.6 ($p < 0.001$).

En revanche, l'effet du genre est non significatif ($F(1, 8920) = 0.39$, $p = 0.54$), avec une estimation de -0.16 pour les hommes ($p = 0.54$), suggérant qu'il n'y a pas de différence entre les sexes.

Table 5 – Résultats de l'ANOVA sans interaction entre l'IDH et le genre

	Ddl	**Som. carrés**	**Moy. carrés**	**F**	**P-value**	η^2
IDH	2	74 008	37 004	284	< 0.001	0.06
Genre	1	50	50	0.39	0.535	0.00
Résidus	8920	1 162 745	130	-	-	-

Une ANOVA à deux facteurs a également été réalisée avec le facteur d'interaction "IDH et genre". Les résultats sont présentés dans la Table 6.

Dans l'ANOVA avec interaction, l'effet de l'IDH reste significatif ($F(2, 89) = 4674$, $p < 0.001$, $^2 = 0.06$), mais l'effet du genre reste non significatif ($F(1, 89) = 0.41$, $p = 0.52$, $^2 = 0$).

De plus, l'interaction entre IDH et genre n'est pas significative ($F(2, 89) = 1.42$, $p = 0.243$, $^2 = 0$), ce qui indique qu'il n'y a pas d'effet d'interaction.

Table 6 – Résultats de l'ANOVA avec interaction entre l'IDH et le genre

	Ddl	**Som. carrés**	**Moy. carrés**	**F**	**P-value**	η^2
IDH	2	74 008	37 004	284	< 0.001	0.06
Genre	1	50	50	0.39	0.54	0.00
Interaction	2	369	185	1.42	0.24	0.00
Résidus	8918	1 162 375	130	-	-	-

La Table présentée différemment nous permet d'obtenir de nouvelles informations (voir Table 7).

TABLE 7 – Résultats de l'ANOVA avec statistiques inférentielles

	Estimate	**std error**	**t-value**	**PR (> t)**
Intercept	6.6	0.42	15.7	< 0.001
IDH (I)	4.2	0.57	7.4	< 0.001
IDH (TH)	8.7	0.38	22.7	< 0.001
Genre (Homme)	-0.16	0.26	-0.62	0.54

Les résultats montrent que l'intercept est significatif (= 6,62, SE = 0,4, t = 15,7, p < 0,001). Le facteur IDH a un effet positif significatif pour les niveaux Intermédiaire (= 4,24, SE = 0,57, t = 7,43, p < 0,001) et pour un IDH = TH (= 8,71, SE = 0,39, t = 22,65, p < 0,001), indiquant des valeurs plus élevées de la variable dépendante pour ces groupes par rapport au groupe de référence. En revanche, le coefficient pour Homme est négatif (= -0,16, SE = 0,26, t = -0,62) et non significatif.

Comme les résultats sont identiques avec et sans interactions entre variables indépendantes, nous poursuivons l'étude en considérant qu'il n'y a pas d'interaction entre les variables.

A.2 Diagnostic du modèle linéaire

L'analyse de régression linéaire est une méthode couramment utilisée pour évaluer la relation entre une variable dépendante continue (nombre de vidéos vues) et une ou plusieurs variables indépendantes (IDH et genre).

Cependant, avant d'interpréter les résultats d'un modèle linéaire, il est essentiel de vérifier que les hypothèses du modèle sont respectées.

Ces hypothèses incluent l'homoscédasticité, la normalité des résidus, l'absence de points influents et la colinéarité entre les variables indépendantes.

Cette annexe présente les étapes de diagnostic du modèle linéaire utilisé pour analyser l'impact de l'IDH (Indice de Développement Humain) et du genre sur le nombre de vidéos vues.

L'homoscédasticité est l'hypothèse selon laquelle les résidus du modèle ont une variance constante pour toutes les valeurs ajustées.

Pour vérifier cette hypothèse, nous avons tracé un graphique des résidus contre les valeurs ajustées (voir Figure 7).

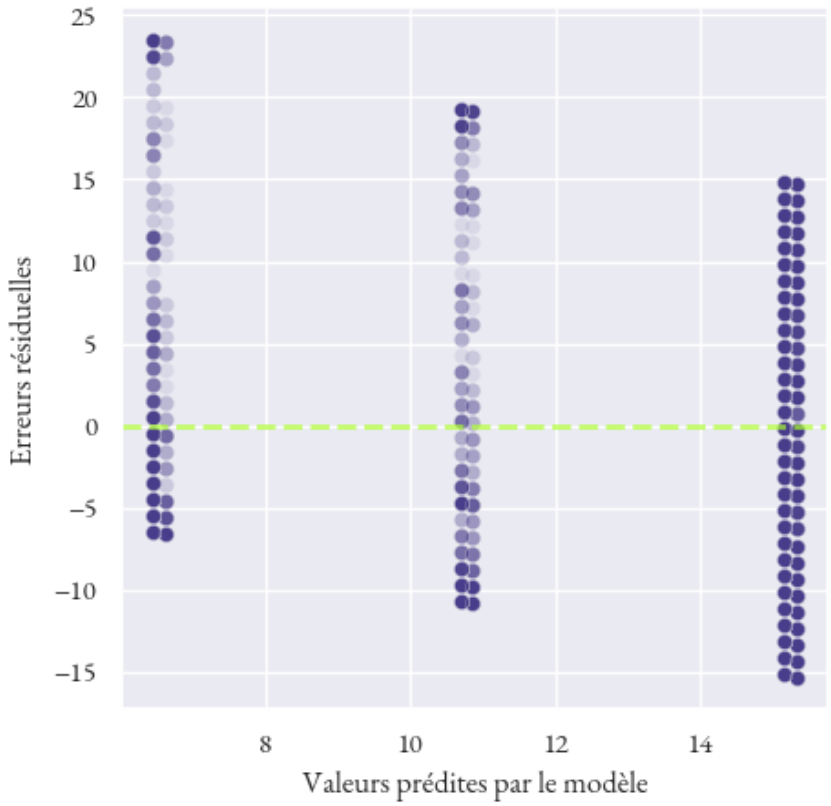

FIGURE 7 – Test de l'homoscédasticité des résidus

La Figure 7 suggère une non-homoscédasticité probable, bien que la hauteur des résidus semble constante.

En effet, les résidus sont plus élevés à gauche et plus bas à droite, ce qui indique un biais systématique dans le modèle.

Cette différence dans la magnitude des résidus à différentes valeurs ajustées suggère que la variance des erreurs n'est pas constante, ce qui pourrait être un signe d'hétéroscédasticité.

Ce phénomène peut résulter d'une mauvaise spécification du modèle, notamment si la relation entre les variables est non linéaire.

Il est donc recommandé de tester des modèles non linéaires, comme l'ajout de termes polynomiaux, ou d'utiliser des transformations sur les variables indépendantes pour mieux ajuster le modèle.

Le test de Breusch-Pagan est particulièrement adapté pour tester l'hétéroscédasticité, car il permet d'examiner si la variance des erreurs dépend des valeurs des variables indépendantes, ce qui est essentiel pour identifier les problèmes d'hétéroscédasticité dans le modèle.

Le test de Breusch-Pagan a été utilisé pour évaluer l'homoscédasticité des résidus. La statistique du test est de 328,4 avec une p-value « 0,001.

Ce résultat indique une hétéroscédasticité significative, c'est-à-dire que la variance des résidus n'est pas constante à travers les valeurs ajustées.

Par conséquent, l'hypothèse d'homoscédasticité est rejetée, et il est recommandé d'explorer des modèles ou des transformations qui prennent en compte cette hétéroscédasticité.

En combinant au test de Breusch-Pagan un diagramme Quantile-Quantile, on peut avoir une vue d'ensemble plus complète de la validité du modèle, en s'assurant que les résidus respectent les deux hypothèses clés de variance constante et de normalité.

Le diagramme Quantile-Quantile permet en effet de vérifier si les résidus suivent une distribution normale.

Si les points suivent une ligne droite, cela suggère que les résidus sont normalement distribués. Sur la Figure 8, le diagramme Quantile-Quantile des résidus présente une courbe en S, ce qui suggère une asymétrie et une possible non-normalité des erreurs.

Cette déviation de la ligne droite indique également une non-linéarité dans la relation entre les variables, suggérant que le modèle pourrait être mal spécifié.

Il serait pertinent de réévaluer le modèle, en envisageant l'ajout de termes non linéaires ou en appliquant des transformations aux variables afin de mieux capturer la relation entre celles-ci et d'améliorer la validité du modèle.

Réaliser une Table des VIF (Variance Inflation Factors) après avoir testé l'homoscédasticité et réalisé un diagramme Quantile-Quantile des résidus permet d'obtenir une compréhension complète de la validité du modèle de régression et de s'assurer que les résultats sont fiables.

L'évaluation des VIF après l'examen des résidus permet d'identifier un autre type de problème possible dans le modèle (la multicolinéarité), même si l'homoscédasticité

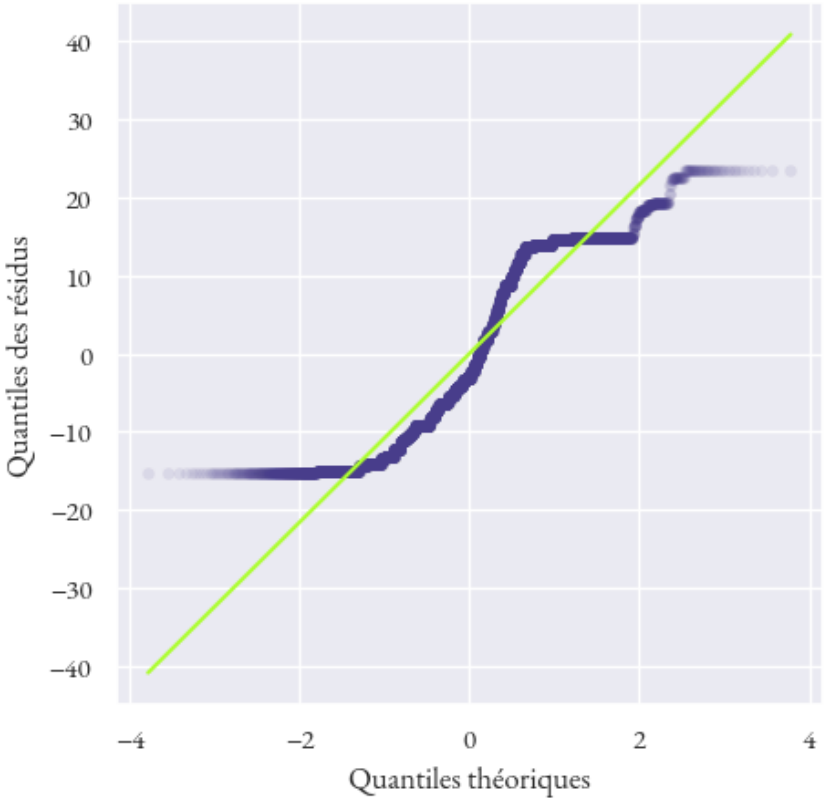

FIGURE 8 – Test de la normalité des résidus

et la normalité des résidus sont respectées.

C'est une étape cruciale pour garantir que les relations entre les variables indépendantes et la variable dépendante sont correctement estimées et interprétées (voir Table 8)

TABLE 8 – Variance Inflation Factors (VIF) pour les variables IDH et Genre

Variable	**VIF**
Intercept	12.09
IDH (I)	1.54
IDH (TH)	1.55
Genre (Homme)	1.02

Les VIF pour toutes les variables sont faibles (inférieurs à 5, et souvent inférieurs à 2), ce qui suggère qu'il n'y a pas de multicolinéarité problématique dans le modèle.

Les variables sont indépendantes les unes des autres et il n'y a pas de risque de surinflation des coefficients de régression à cause de la colinéarité.

L'analyse de régression linéaire permet d'évaluer la relation entre une variable dépendante (nombre de vidéos vues) et plusieurs variables indépendantes (IDH et genre).

La vérification des hypothèses du modèle, y compris l'homoscédasticité, la normalité des résidus et la multicolinéarité, a révélé plusieurs problèmes qui remettent en question la validité de l'utilisation de l'ANOVA dans ce contexte.

L'hétéroscédasticité identifiée par le test de Breusch-Pagan et la déviation des résidus par rapport à la normalité observée dans le diagramme Quantile-Quantile indiquent que les résidus ne suivent pas une distribution normale et que la variance n'est pas constante.

Ces violations des hypothèses de l'ANOVA suggèrent que cette méthode n'était pas appropriée pour analyser l'impact de l'IDH et du genre sur le nombre de vidéos vues.

Par conséquent, une approche alternative, comme l'utilisation de modèles non linéaires ou de régressions robustes, serait plus appropriée pour tirer des conclusions fiables.

A.3 Régression de type Poisson

La distribution du nombre de vidéos vues dans un MOOC ne suit pas une loi de Poisson en raison de son asymétrie et des valeurs extrêmes à droite. Beaucoup d'apprenants regardent peu de vidéos, ce qui génère une forte concentration à gauche de l'histogramme. En outre, les comportements des apprenants sont très variés, certains étant plus engagés que d'autres. La loi de Poisson ne convient pas à cette dispersion élevée des données (voir Figure 2). Les résultats de ce modèle sont affichés dans la Table 9 .

TABLE 9 – Résultats de la régression de type Poisson

Variable	**estimate**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	1.88	0.01	141.2	< 0.001	1.85	1.90
IDH (I)	0.50	0.01	29.5	< 0.001	0.47	0.54
IDH (TH)	0.85	0.01	67.1	< 0.001	0.83	0.88
Genre (Homme)	-0.01	0.01	-1.9	0.06	-0.02	0.00

Les résultats de la régression de type Poisson indiquent que les groupes IDH = I et IDH = TH ont un impact significatif et positif sur le nombre de vidéos vues.

Ces effets sont statistiquement significatifs ($p < 0.001$) et précis, ce qui suggère que l'IDH joue un rôle important dans la quantité de vidéos vues.

En revanche, le genre n'a pas d'impact significatif sur cette variable, ce qui suggère que les différences de genre ne sont pas un facteur déterminant dans le nombre de vidéos visionnées.

La modélisation par Poisson est meilleure que celle par l'ANOVA, mais le petit pic autour de 25-30 vidéos vues invalide l'utilisation d'une loi de Poisson. Il y a de toute façon trop d'individus n'ayant regardé aucune vidéo, ce qui nous amène en théorie sur des modèles dits "zéro-inflation".