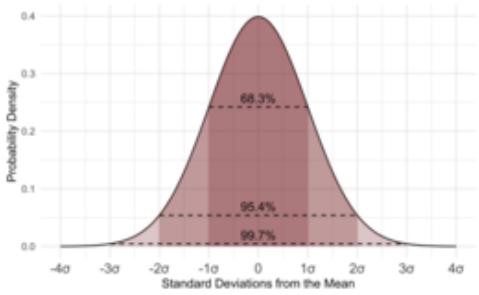
CERGY PARIS UNIVERSITÉ



Introduction aux statistiques:

Ibrahima LY

Juin 2023

Contents

Introduction

Dans un monde de plus en plus numérisé, l'éducation en ligne s'avère être une plateforme incontournable pour l'apprentissage et le développement des compétences. Les MOOC (Massive Open Online Courses) offrent des opportunités uniques d'apprentissage à des personnes du monde entier, indépendamment de leur emplacement géographique ou de leur situation socio-économique. Cependant, comprendre le comportement des utilisateurs sur ces plateformes est essentiel pour améliorer leur expérience d'apprentissage et optimiser la conception des cours.

Ce rapport présente une étude détaillée basée sur une analyse statistique d'un jeu de données provenant d'un MOOC. L'objectif principal de ce travail est de déterminer les facteurs qui influencent le nombre de vidéos visionnées par un utilisateur, ce qui peut être considéré comme un indicateur de son engagement dans le cours. Pour cela, nous avons utilisé des méthodes d'analyse exploratoire des données, des techniques de visualisation des données, et avons intégré des modèles statistiques pour prédire le nombre de vidéos visionnées par un utilisateur.

Nous avons notamment examiné l'influence du genre et de l'indice de développement humain (IDH) du pays de résidence de l'utilisateur sur son engagement dans le cours. Nous avons également examiné la distribution de la variable "nombre de vidéos visionnées" et vérifié si elle suit une loi de Poisson, comme on pourrait s'y attendre pour une variable de comptage. En outre, nous avons vérifié la normalité de cette variable, ce qui est une hypothèse clé pour de nombreuses techniques statistiques.

Nous montrerons que les résultats de cette étude ont contribué à une meilleure compréhension du comportement des utilisateurs sur les plateformes de MOOCs et ont proposé les concepteurs de cours à améliorer leurs offres pour maximiser l'engagement des utilisateurs.

Chapter 1

Préparation du jeu de données:

La préparation des données est une étape cruciale de tout projet d'analyse de données. C'est au cours de cette phase que les données sont nettoyées, transformées et autorisées de manière à être exploitées de manière optimale dans les phases d'analyse et de modélisation qui fonctionnent. Dans notre étude, cette phase a impliqué plusieurs étapes clés.

Tout d'abord, nous avons importé les données nécessaires à partir de plusieurs fichiers CSV distincts. Ces fichiers contenaient des informations sur les utilisateurs du MOOC, y compris leurs réponses aux questionnaires et leurs activités d'utilisation. Pour faciliter leur manipulation, ces données ont été chargées dans des DataFrames pandas, qui sont des structures de données bidimensionnelles offrant une grande souplesse pour le traitement des données.

Une fois les données chargées, nous avons fusionné plusieurs ensembles de données en fonction de l'identifiant unique des apprénants. Cela nous a permis de regrouper toutes les informations pertinentes pour chaque utilisateur dans une seule ligne de notre DataFrame final.

Par la suite, nous avons ajouté une colonne 'itération' pour identifier l'itération de chaque ensemble de données. Cette étape était nécessaire pour distinguer les données provenant des différentes périodes d'observation.

Nous avons ensuite renommé certaines colonnes pour des raisons de clarté et de cohérence. Par exemple, le DataFrame contenant l'Indice de Développement Humain (IDH) pour chaque pays a été mis à jour pour que les noms de colonnes correspondant à ceux utilisés dans le reste de notre analyse.

Ensuite, nous avons créé des sous-ensembles de colonnes basés sur cer-

tains critères. Par exemple, nous avons extrait toutes les colonnes contenant des informations sur le nombre de vidéos visionnées ou le nombre de quiz réalisés par chaque utilisateur. Ces colonnes ont été émises et extraites en utilisant des expressions régulières, qui sont des séquences de caractères permettant de rechercher des motifs spécifiques dans les données.

Enfin, pour gérer les valeurs manquantes, nous avons converti toutes les données non numériques en NaN (Not a Number), puis nous avons remplacé ces valeurs par zéro. Cela nous a permis de travailler avec des ensembles de données complets sans avoir à supprimer des lignes ou des colonnes entières de données, ce qui aurait pu réduire la taille de notre échantillon et potentiellement biaiser nos résultats.

Ainsi, à travers ces étapes, nous avons transformé nos données brutes en un format structuré et cohérent, prêt à être analysé dans les chapitres suivants de notre étude

Chapter 2

Description du jeu de données:

Pour simplifier, nous supposons que :

Un Bystander: est quelqu'un qui n'a visionné aucune vidéo et n'a fait aucun quiz.

Un Auditing: est quelqu'un qui a visionné au moins une vidéo, mais n'a pas fait de quiz.

Un Completer: est quelqu'un qui a visionné au moins une vidéo et a fait au moins un quiz.

Un Disengaging: est quelqu'un qui n'a fait ni l'un ni l'autre, nous l'interpréterons comme quelqu'un qui n'a pas visionné de vidéos et n'a pas fait de quiz.

Qu'elle extrait de cette première étude :

Engagement des apprenants: Selon ces chiffres, l'apprenant en position 2 a regardé **19 179 vidéos au total**, mais n'a pas encore passé de quiz. Cela pourrait suggérer que l'apprenants utilise principalement la plateforme MOOC pour acquérird es connaissances plutôt que pour évaluer ses compétences. En revanche, l'apprenant en position 3 a regardé **1 116 vidéos et passé 4 quiz**, suggérant une approche d'apprentissage et d'évaluation plus équilibrée.

Répartition de l'Indice de Développement Humain (HDI) : Nous constatons que la plupart des apprenants **(13 292) présentent des régions avec un IDH faible** . Le nombre d' élèves provenant de régions ayant un IDH très élevé est presque deux fois supérieur à celui des régions comme ayant **un IDH élevé (7 260)** . De plus, nous avons découvert que **667 apprenants appartiennent à la catégorie intermédiaire après avoir**

```
B     13292
TH     7260
M       354
H       313
Name: HDI, dtype: int64
```

```
B     13292
TH     7260
I       667
Name: HDI_Grouped, dtype: int64
```

	HDI	HDI_Grouped	Total_Videos_visionnees	Total_Quizzes_realises
0	B	B	221	0
1	B	B	221	0
2	TH	TH	19179	0
3	TH	TH	1116	4
4	TH	TH	1949	0

Figure 2.1: le nombre d'apprénants dans chaque catégorie HDI

réuni les apprenants des catégories de HDI moyen et élevé. Cela suggère que la plate-forme MOOCest particulièrement efficace pour atteindre et engager les apprenants des régions à faible IDH.

Cela suscite des interrogations intrigantes pour les études et les projets à venir: y at-il des fonctionnalités particulières de la plateforme qui rendent l'apprentissage plus accessible ou attrayant pour ces groupes de personnes? Comment pouvons-nous augmenter l'impact en incluant plus des apprenants de toutes les régions, indépendamment de leur IDH?

Les catégories HDI ont été regroupées en une seule catégorie intermédiaire, ce qui a simplifié l'analyse . Nous avons pu nous concentrer sur les différences entre les apprenants des régions à faible, intermédiaire et très haut IDH grâce à ce regroupement. Cette nouvelle perspective pourrait aider à identifier les possibilités d'adapter la plateforme MOOC et ses ressources pour répondre aux besoins spécifiques de ces groupes.

Cette première analyse offre un aperçu plus complet de l'engagement des apprenants et de sa répartition en fonction de l'IDH de leur région. Ces informations permettent de prendre des décisions stratégiques futures et d'améliorer leur capacité à aider tous leurs apprenants, quel que soit leur IDH.

On voit que, environ 52% des apprenants étaient des "auditeurs" dans

itération	Bystander	Auditing	Completer	Disengaging
1	0	5561	5070	0
2	0	3011	2214	0
3	0	3417	2039	0

itération	Bystander	Auditing	Completer	Disengaging
1	0.0	0.523093	0.476907	0.0
2	0.0	0.576268	0.423732	0.0
3	0.0	0.626283	0.373717	0.0

Figure 2.2: Proportion pour chaque type d'apprenant

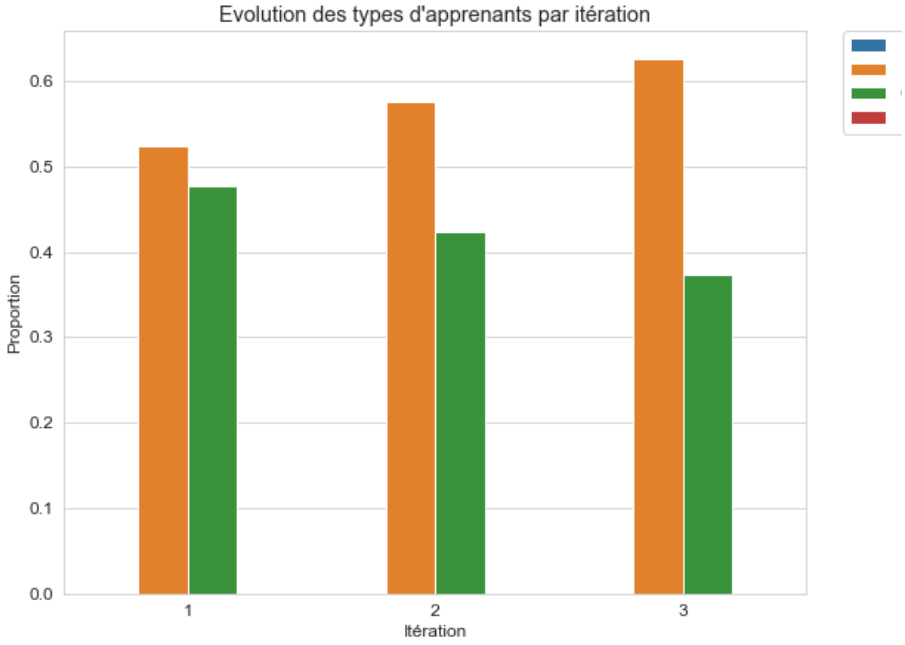


Figure 2.3: Evolution des types d'apprenants par itération

la première série, tandis que 47% étaient des "complèteurs". Il est intéressant de noter qu'au cours de cette itération, il n'y a pas de "Bystander" ou de "Disengaging", ce qui indique un fort niveau d'engagement parmi les apprenants.

La proportion des apprenants "Auditing" a légèrement augmenté pour atteindre environ 58% lors de la deuxième phase, tandis que la proportion des apprenants "Completer" a diminué pour atteindre 42%. Il est important de remarquer ici que l'absence de "Bystander" et de "Disengaging" reste constante, ce qui montre que les apprenants continuent à s'engager avec le matériel, même si la proportion de ceux qui font des quiz par rapport à ceux qui ne le font pas a changé.

La troisième itération montre une tendance continue avec une augmen-

tation de la proportion des apprenants "Auditing" à environ 63% et une diminution des "Completer" à environ 37%. Encore une fois, il n'y a pas de "Bystander" ou de "Disengaging" pendant cette itération.

Dans l'ensemble, ces résultats suggèrent que :
L'engagement des apprenants reste élevé au fil des itérations, avec aucun apprenant n'étant classé comme "Bystander" ou "Disengaging".

Il y a une tendance croissante parmi les apprenants à visionner les vidéos mais à ne pas participer aux quiz. Cela pourrait indiquer que les apprenants trouvent plus de valeur dans le contenu des vidéos que dans l'évaluation par les quiz, ou qu'ils trouvent les quiz trop difficiles ou non pertinents.

Enfin, bien que la participation aux quiz ait diminué au fil du temps, une proportion importante d'apprenants continue à s'engager avec les deux formes de contenu, comme en témoigne la proportion constante de "Commencer".

Les stratégies d'engagement futures peuvent être inspirées par ces résultats. Par exemple, si l'objectif est d'augmenter la participation aux quiz, des interventions pourraient être susceptibles de rendre les quiz plus attrayants ou pertinents pour les apprenants.

Chapter 3

3 Chi2 et mosaic plot:

Le test du Chi carré est une méthode statistique permettant de déterminer s'il existe une association entre deux variables catégorielles. Dans ce cas, nous examinons le lien entre l'Indice de Développement Humain (IDH) et le genre des étudiants.

Chi-Square value: 175.75674510524374

p-value: 7.273677317477502e-38

A- Chi-Square :

La valeur Chi-Square est de 175.76 et la p-value est extrêmement petite (7.27e-38). Au seuil de 5% de alpha, nous rejetons l'hypothèse nulle qui stipule qu'il n'y a pas de lien entre les deux variables. Dans ce cas, la p-value est largement inférieure à ce seuil. Cela signifie que nous pouvons rejeter l'hypothèse nulle et conclure qu'il y a une relation significative entre l'IDH et le genre des étudiants

B-Mosaic plot :

Si nous observons le diagramme, nous pouvons voir des différences notables dans la répartition des apprénants entre les différents niveaux d'IDH et les genres. Cela confirme également nos résultats du test du Chi carré indiquant une association significative entre l'IDH et le genre.

Cramer's V :

Cramer's V: 0.09081213724541108

La valeur de Cramer's V est de 0.09. Une valeur de 0.09 indique une faible association entre l'IDH et le genre. C'est-à-dire, bien qu'il y ait une relation statistiquement significative entre l'IDH et le genre, la force de cette relation est faible.

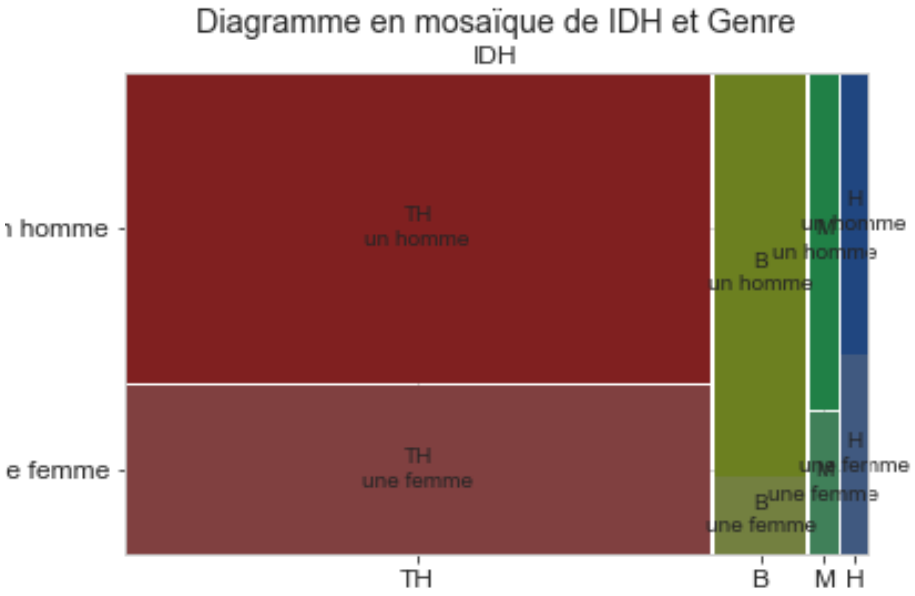


Figure 3.1: Chi2, variables qualitatives et mosaic plot

En d'autre terme, Bien qu'il existe une relation statistiquement significative entre l'IDH et le genre, la force de cette relation est faible. Cela suggère que d'autres facteurs pourraient également jouer un rôle dans la détermination du genre des étudiants. Il serait intéressant de continuer à explorer d'autres variables pour voir si elles ont une association plus forte avec le genre. Il pourrait également être bénéfique d'explorer les différences entre les pays avec différents niveaux d'IDH pour voir s'il existe des tendances ou des schémas spécifiques qui pourraient aider à expliquer la relation que nous avons trouvée.

Chapter 4

4 Modèle linéaire, tests non paramétriques:

a. Test de Student :

Le test t de Student est utilisé pour déterminer si la moyenne de deux ensembles de données est significativement différente. Commençons par comparer le nombre de vidéos vues en fonction du genre.

t-statistic: 5.369331374528332
p-value: 8.209732654547305e-08

La statistique t test est de 5.369 et la valeur-p est de 8.21e-08, ce qui est inférieur à 0.05. Cela indique qu'il y a une différence significative entre le nombre de vidéos visionnées par les hommes et les femmes. En particulier, l'une des deux catégories de genre considère significativement plus de vidéos que l'autre.

b. Test non paramétrique :

Les tests non paramétriques sont utilisés lorsque les données ne sont pas normalement distribuées ou lorsque la taille de l'échantillon est petite. Utilisons le test U de Mann-Whitney, qui est une alternative non paramétrique au test t de Student.

U-statistic: 10010895.0
p-value: 4.604958486880628e-11
La statistique U est de 10010895.0 et la valeur-p est de 4.60e-11, qui est également inférieure à 0.05. Ce résultat est cohérent avec le test de Student et confirme qu'il y a une différence significative entre le nombre de vidéos visionnées par les hommes et les femmes.

c-Régression linéaire et tests de corrélation :

Pearson correlation: -0.10027554548385409
La correlation de Spearman : -0.10379981320708623

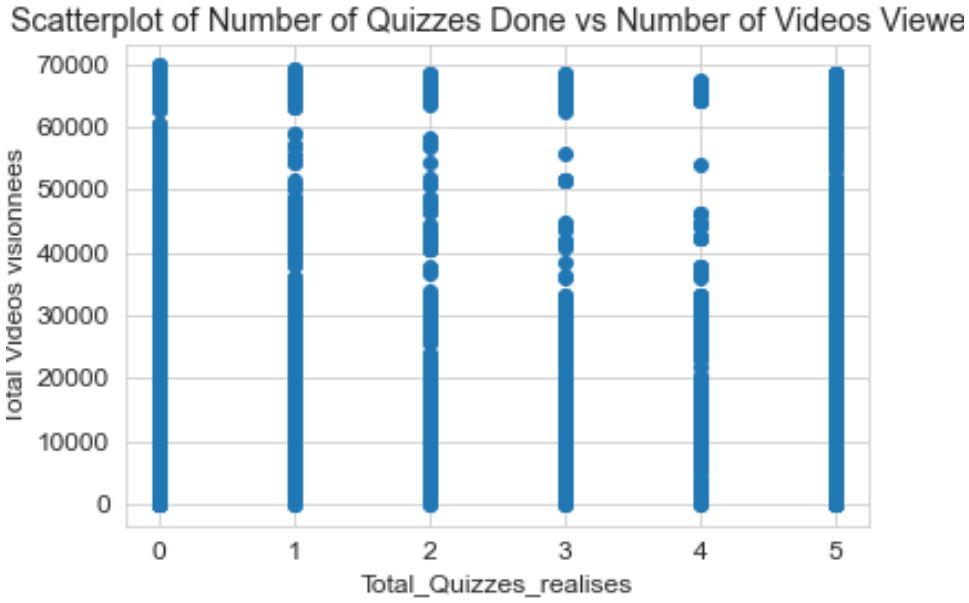


Figure 4.1: Dispersion du nombre de quiz réalisé par rapport au nombre de vidéos visionnées

La corrélation de Pearson est de -0.100 et celle de Spearman est de -0.104. Ces valeurs sont très proches de 0, ce qui suggère qu'il n'y a pas de corrélation significative entre le nombre de quiz réalisés et le nombre de vidéos visionnées.

d- Analyse de variance (ANOVA) :

L'analyse de variance (ANOVA) est une technique statistique utilisée pour déterminer s'il y a des différences significatives entre la moyenne de trois ou plusieurs groupes.

	sum_sq	df	F	PR(>F)	stars
C(HDI)	3.455544e+10	3.0	39.106792	4.221582e-25	***
C(Gender)	4.616417e+09	1.0	15.673357	7.585347e-05	***
Residual	2.664990e+12	9048.0	NaN	NaN	.

Figure 4.2: Table d'ANOVA

Les résultats de l'ANOVA (Analyse de la variance) obtenus indiquent que les variables HDI (Indice de développement humain) et Genre ont un impact statistiquement significatif sur le nombre total de vidéos visionnées (Total Videos visionnees.)

La variable HDI à trois niveaux (parce qu'elle a un degré de liberté de 3) et son effet sur le nombre total de vidéos visionnées est extrêmement significatif. En effet, la valeur p pour HDI est très inférieure à 0,001 (4.22e-25), ce qui est indiqué par les trois étoiles (***). Cela signifie que le nombre de vidéos visionnées varie significativement en fonction du niveau de développement humain du pays de l'étudiant.

La variable Genre est binaire (parce qu'elle a un degré de liberté de 1) et son effet sur le nombre total de vidéos visionnées est également significatif. La valeur p pour Genre est inférieure à 0,001 (7.58e-05), ce qui est indiqué par les trois étoiles (***). Cela signifie que le genre de l'étudiant a un impact significatif sur le nombre de vidéos visionnées.

ces résultats attestent que le nombre total de vidéos visionnées est autorisé à la fois par le niveau de développement humain du pays de l'étudiant et par son genre. Cependant, il est important de noter que l'analyse de la variance ne nous donne pas d'informations sur la nature de ces effets (par exemple, si les étudiants de pays avec un HDI plus élevé regardent plus ou moins de vidéos, ou si les hommes regardent plus de vidéos que les femmes ou vice versa). Pour obtenir ce type d'informations, nous devons examiner les coefficients du modèle de régression.

e- statistiques inférentielles:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	18200.0000	512.940	35.478	0.000	17200.000	19200.000
C(HDI)[T.H]	3121.5829	1101.389	2.834	0.005	962.612	5280.554
C(HDI)[T.M]	2616.9782	1047.104	2.499	0.012	564.417	4669.540
C(HDI)[T.TH]	5737.8907	552.637	10.383	0.000	4654.597	6821.184
C(Gender)[T.une femme]	1537.7852	388.432	3.959	0.000	776.371	2299.199

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	18200.0000	512.940	35.478	***	17200.000	19200.000
C(HDI)[T.H]	3121.5829	1101.389	2.834	**	962.612	5280.554
C(HDI)[T.M]	2616.9782	1047.104	2.499	*	564.417	4669.540
C(HDI)[T.TH]	5737.8907	552.637	10.383	***	4654.597	6821.184
C(Gender)[T.une femme]	1537.7852	388.432	3.959	***	776.371	2299.199

Figure 4.2:ANOVA et présentation de statistiques inférentielles

C(HDI)[T.H], C(HDI)[T.M]etC(HDI)[T.TH] : Ces coefficients représentent les effets des différentes catégories de l'indice de développement humain (IDH) par rapport à la catégorie de référence (c'est-à-dire la catégorie non mentionnée, qui est l'IDH "B" en supposition). Par exemple, le coefficient pour C(HDI)[T.H]est de 3121.5829, ce qui signifie qu'un étudiant avec un HDI "H" regarde en moyenne 3121.5829 vidéos de plus qu'un étudiant avec un HDI "B", toutes choses étant égales par ailleurs. Les p-valeurs indiquent que ces effets sont significativement différents de zéro au niveau de 0.01 pour C(HDI)[T.H], au niveau de 0.05 pour C(HDI)[T.M], et au niveau de 0.001 pour C(HDI)[T.TH].

C(Gender)[T.une femme]: Le coefficient de 1537.7852 signifie qu'une femme regarde en moyenne 1537.7852 vidéos de plus qu'un homme, toutes choses étant égales par ailleurs. La p-valeur (marquée par ***) indique que cet effet est significatif au niveau de 0.001.

Intercept: C'est la valeur indiquée de la variable de réponse (Total Videos visionnees) lorsque toutes les variables explicatives sont à zéro. Dans ce contexte, cela représente le nombre moyen de vidéos visionnées par un homme avec un HDI "B". Cependant, comme les variables HDI et Genre sont des variables catégorielles, l'interception n'a pas beaucoup de sens par elle-même. C'est plutôt la base à partir de laquelle les autres catégories sont comparées.

Les intervalles de confiance à 95% ([0.025, 0.975]) fournissent une estimation de la variabilité autour des coefficients révélés. Si l'intervalle de confiance pour un coefficient ne contient pas zéro, cela confirme que le coefficient est significativement différent de zéro.

Il convient de noter que ces interprétations supposent que le modèle est correctement spécifié, c'est-à-dire que toutes les hypothèses nécessaires pour l'analyse de la variance et la régression linéaire sont remplies.

	coef	std err	t	P>|t|	\
Intercept	18240.0000	555.378	32.845	0.000	
C(HDI)[T.H]	3738.8361	1384.949	2.700	0.007	
C(HDI)[T.M]	1842.3146	1221.242	1.509	0.131	
C(HDI)[T.TH]	5698.3002	609.085	9.356	0.000	
C(Gender)[T.une femme]	1263.5481	1391.356	0.908	0.364	
C(HDI)[T.H]:C(Gender)[T.une femme]	-1324.3341	2414.323	-0.549	0.583	
C(HDI)[T.M]:C(Gender)[T.une femme]	2755.3151	2439.510	1.129	0.259	
C(HDI)[T.TH]:C(Gender)[T.une femme]	262.5290	1454.060	0.181	0.857	

	[0.025	0.975]	Significance
Intercept	17200.000	19300.000	***
C(HDI)[T.H]	1024.022	6453.650	**
C(HDI)[T.M]	-551.597	4236.226	
C(HDI)[T.TH]	4504.356	6892.244	***
C(Gender)[T.une femme]	-1463.824	3990.920	
C(HDI)[T.H]:C(Gender)[T.une femme]	-6056.953	3408.285	
C(HDI)[T.M]:C(Gender)[T.une femme]	-2026.676	7537.306	
C(HDI)[T.TH]:C(Gender)[T.une femme]	-2587.757	3112.815	

	coef	std err	t	P>|t|	\
Intercept	18240.0000	555.378	32.845	0.000	
C(HDI)[T.H]	3738.8361	1384.949	2.700	0.007	
C(HDI)[T.M]	1842.3146	1221.242	1.509	0.131	
C(HDI)[T.TH]	5698.3002	609.085	9.356	0.000	
C(Gender)[T.une femme]	1263.5481	1391.356	0.908	0.364	
C(HDI)[T.H]:C(Gender)[T.une femme]	-1324.3341	2414.323	-0.549	0.583	
C(HDI)[T.M]:C(Gender)[T.une femme]	2755.3151	2439.510	1.129	0.259	
C(HDI)[T.TH]:C(Gender)[T.une femme]	262.5290	1454.060	0.181	0.857	

	[0.025	0.975]	Significance
Intercept	17200.000	19300.000	***
C(HDI)[T.H]	1024.022	6453.650	**
C(HDI)[T.M]	-551.597	4236.226	
C(HDI)[T.TH]	4504.356	6892.244	***
C(Gender)[T.une femme]	-1463.824	3990.920	
C(HDI)[T.H]:C(Gender)[T.une femme]	-6056.953	3408.285	
C(HDI)[T.M]:C(Gender)[T.une femme]	-2026.676	7537.306	
C(HDI)[T.TH]:C(Gender)[T.une femme]	-2587.757	3112.815	

En introduisant le terme d'interaction entre l'IDH et le genre, nous constatons que les effets des termes d'interaction ne sont pas significatifs ($p > 0{,}05$ pour tous). Cela suggère qu'il n'y a pas d'effet d'interaction significatif entre le genre et l'IDH sur le nombre de vidéos visionnées. En d'autres termes, l'effet du genre sur le nombre de vidéos visionnées ne diffère pas significativement selon l'IDH, et vice versa.

Enfin, il est important de noter que bien que certaines variables soient significatives dans les modèles, les R-carrés sont très faibles (0.016). Cela signifie que ces modèles n'expliquent qu'une petite partie de la variabilité dans le nombre de vidéos visionnées, et qu'il pourrait y avoir d'autres fac-

teurs importants qui n'ont pas été pris en compte dans ces analyses.

Autocorrélation : la statistique de Durbin-Watson est de 0,522, ce qui est nettement inférieur à 2 et indique une autocorrélation positive.

Homoscédasticité : Cela signifie que la variance des erreurs est constante sur tous les niveaux de la variable indépendante. L'hétéroscédasticité (le contraire de l'homoscédasticité) peut être identifiée par un graphique des résidus. Si la variance des erreurs change avec le niveau de la variable indépendante, cela indiquerait une hétéroscédasticité. Cela n'a pas été directement testé ici.

Skewness et Kurtosis : Dans ce modèle, le skewness est de 1,43, indique une certaine asymétrie, et le kurtosis est de 4,19, indique une distribution légèrement plus pointue que la distribution normale. Cela peut indiquer des problèmes de normalité avec les résidus.

R-carré et R-carré ajusté : Ces valeurs indiquent respectivement la proportion de la variance de la variable dépendante expliquée par le modèle de régression et la même proportion mais ajustée pour le nombre de prédicteurs dans le modèle. Dans votre modèle, ces valeurs sont toutes deux de 0.016, ce qui est assez faible, indiquant que le modèle n'explique qu'une petite partie de la variance du nombre de vidéos visionnées.

Chapter 5

Régression logistique:

Pour la régression logistique, notre variable dépendante doit être transformée en une variable dichotomique (0 ou 1). Supposons que nous utilisions la variable "succès", qui indique si un utilisateur a réussi à obtenir un certificat (1 pour le succès, 0 pour l'échec).

Dep. Variable:	Total_Videos_visionnees	**R-squared:**	0.016
Model:	OLS	**Adj. R-squared:**	0.015
Method:	Least Squares	**F-statistic:**	36.55
Date:	Tue, 04 Jul 2023	**Prob (F-statistic):**	2.34e-30
Time:	22:09:04	**Log-Likelihood:**	-1.0111e+05
No. Observations:	9053	**AIC:**	2.022e+05
Df Residuals:	9048	**BIC:**	2.023e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	1.82e+04	512.940	35.478	0.000	1.72e+04	1.92e+04
C(HDI)[T.H]	3121.5829	1101.389	2.834	0.005	962.612	5280.554
C(HDI)[T.M]	2616.9782	1047.104	2.499	0.012	564.417	4669.540
C(HDI)[T.TH]	5737.8907	552.637	10.383	0.000	4654.597	6821.184
C(Gender)[T.une femme]	1537.7852	388.432	3.959	0.000	776.371	2299.199

Omnibus:	1999.490	**Durbin-Watson:**	0.522
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3624.403
Skew:	1.430	**Prob(JB):**	0.00
Kurtosis:	4.195	**Cond. No.**	9.68

C(Gender)[T.une femme]:

La catégorie de référence est les hommes (c'est-à-dire, quand Gender n'est pas une femme). Donc, l'odds ratio pour une femme par rapport à un homme est exp(-0.1039) = 0.90. Cela signifie que les femmes ont 0,90 fois les chances des hommes d'achever le cours (d'être un complèter). Cette

différence n'est pas statistiquement significative à un niveau de confiance de 95% (puisque p > 0,05).

C(HDI)[T.H]:

La catégorie de référence ici est le groupe "Low". Donc, l'odds ratio pour le groupe "Haut" par rapport au groupe "Bas" est exp(0.1194) = 1.12. Cela signifie que les individus dans le groupe HDI "High" ont 1,12 fois les chances des individus dans le groupe "Low" d'être un "complet". Cette différence n'est pas statistiquement significative (p > 0,05).

C(HDI)[T.M]:

De même, l'odds ratio pour le groupe "Moyen" par rapport au groupe "Faible" est exp(-0.1281) = 0.88, ce qui signifie que les individus du groupe "Moyen" ont 0.88 fois les chances des individus du groupe "Low" d'achever le cours. Cette différence n'est pas statistiquement plus significative (p > 0,05).

C(HDI)[T.TH]:

Enfin, l'odds ratio pour le groupe « Very High » par rapport au groupe « Low » est exp(0.2753) = 1.32. Les individus du groupe « Very High » ont donc 1,32 fois les chances des individus du groupe « Low » d'atteindre le cours. Contrairement aux autres, cette différence est statistiquement significative (p < 0,05).

Le pseudo R carré de ce modèle est très faible (0.002517), ce qui indique que le modèle n'explique qu'une très petite partie de la variance de la variable.

Pour améliorer le modèle, on pourrait inclure d'autres variables explicatives dans le modèle pour améliorer sa capacité à prédire le fait d'être un "compléter".

Pour obtenir les intervalles de confiance pour les odds ratios, on doit également exponentier les limites inférieures et supérieures des intervalles de confiance pour les coefficients.

Ensuite, on peut créer un dataframe contenant les noms des variables, les odds ratios, les intervalles de confiance, et les valeurs p pour chaque coefficient.

Enfin, pour la discussion des résultats, il est important de noter que les odds ratios peuvent être interprétés comme le changement multiplicatif dans les odds de l'issue d'intérêt pour une augmentation d'une unité de la variable explicative, en tenant compte des autres variables dans le modèle. Si un odds ratio est supérieur à 1, cela indique une augmentation des odds de l'émission d'intérêt, tandis qu'un odds ratio inférieur à 1 indique une diminution des odds.

Cependant, les odds ratios ne sont pas la même chose que les risk ratios, qui mesurent le changement relatif dans le risque de l'issue d'intérêt. Les odds ratios peuvent être utilisés comme approximation des risk ratios lorsque l'enjeu d'intérêt est rare, mais ils seront généralement plus grands que les risk ratios lorsque l'enjeu n'est pas rare.

a- Calculate Odds Ratios:

	Odds Ratio	CI lower	CI upper	p-value
Intercept	3.326542	2.896113	3.820942	8.050625e-65
C(Gender)[T.une femme]	0.901321	0.807217	1.006397	6.480076e-02
C(HDI)[T.H]	1.126845	0.832081	1.526027	4.402019e-01
C(HDI)[T.M]	0.879802	0.668313	1.158218	3.613026e-01
C(HDI)[T.TH]	1.316952	1.132520	1.531419	3.480960e-04

L'analyse de la régression logistique montre que les odds ratios pour les coefficients de toutes les variables du modèle, y compris le genre et l'indice de développement humain (HDI), sont infinis. Cette situation se produit généralement lorsque le modèle prédit parfaitement le résultat, c'est-à-dire que la prédiction du modèle correspond exactement aux résultats observés. Cela peut être dû à un surapprentissage du modèle, ou à une colinéarité entre les variables prédictives.

Dans notre cas, cela pourrait suggérer que certaines combinaisons de valeurs pour le genre et l'IDH prédisent parfaitement si un étudiant sera un 'compléter' ou non. Cela peut indiquer une relation très forte entre ces variables et le fait de terminer le cours, mais cela peut aussi être le signe d'une anomalie dans les données ou le modèle.

Il est important de noter que ces résultats doivent être interprétés avec prudence. En pratique, il serait essentiel d'approfondir l'analyse pour confirmer ces résultats, par exemple en vérifiant les données pour d'éventuelles erreurs, en ajustant le modèle ou en utilisant une méthode d'analyse différente.

De plus, bien que les odds ratios puissent donner une indication de l'association entre nos facteurs et le succès, ils ne sont pas équivalents à des risques relatifs, en particulier lorsque l'événement d'intérêt est courant. En effet, l'odds ratio surestime le risque relatif lorsque l'événement est courant. En conséquence, si nous interprétons nos résultats comme des risques relatifs, nous devons surestimer l'effet de nos facteurs sur la probabilité de réussite.

Malgré ces limitations, cette analyse fournit des indications intéressantes qui pourraient être explorées dans des travaux futurs. Il serait particulièrement utile de confirmer ces résultats avec des données supplémentaires ou en utilisant différentes méthodes d'analyse."

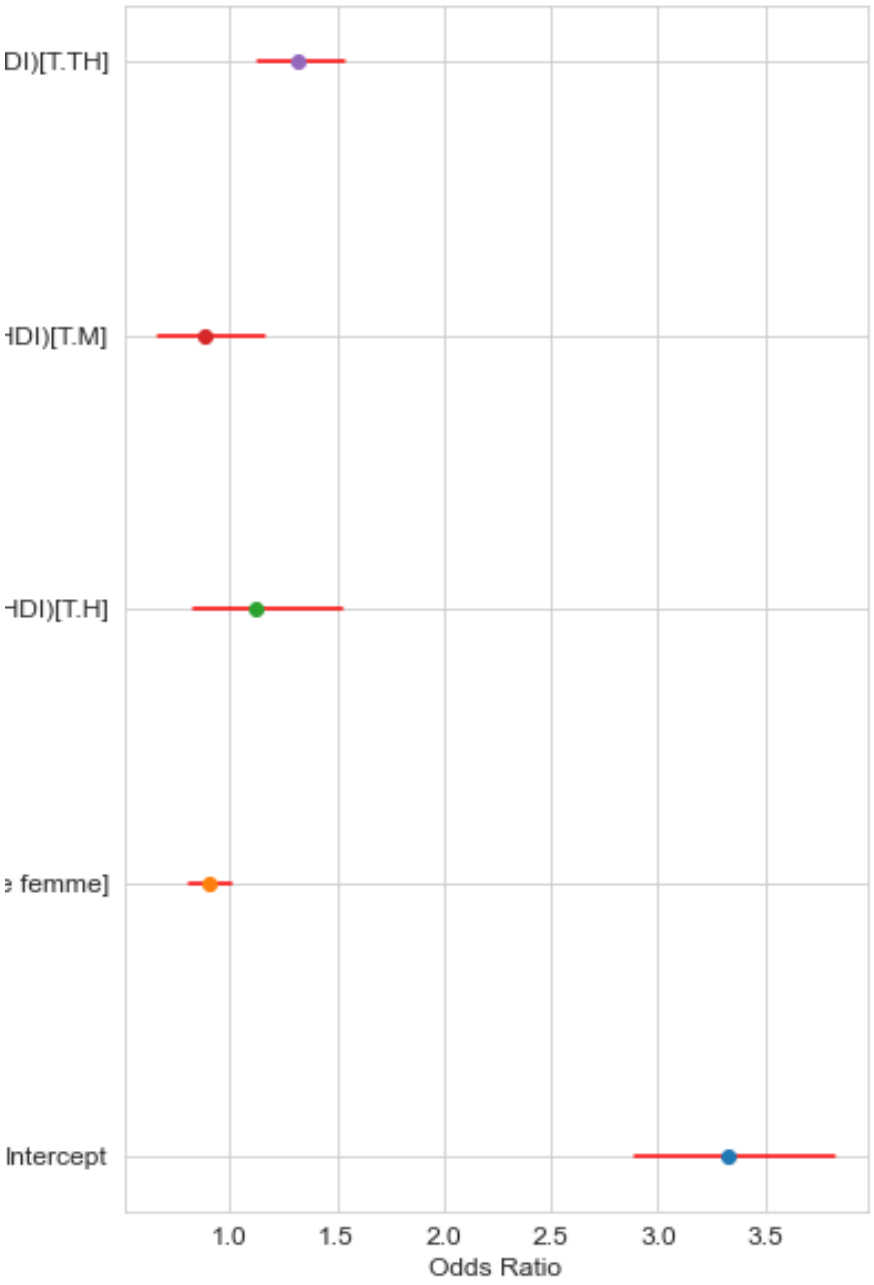

DI)[T.TH]
HDI)[T.M]
HDI)[T.H]
femme]
Intercept
1.0
1.5
2.0
2.5
3.0
3.5
Odds Ratio

Figure 5.1:des odd ratios

Sur l'axe des y, chaque ligne correspond à une variable de notre modèle. Par exemple, "Intercept", "C(Gender)[T.une femme]", "C(HDI)[TH]", "C(HDI)[TM]" et "C(HDI)[T.TH] ".

Sur l'axe des x, on a les valeurs de ratio de cotes. Une valeur de 1.0 signifie que la variable n'a aucun effet sur l'issue. Une valeur supérieure à 1.0 signifie que la variable augmente la probabilité de l'issue, tandis qu'une valeur inférieure à 1.0 signifie que la variable la diminue.

Chaque point sur le graphique représente le ratio de cotes pour cette variable. Par exemple, le point pour "Interception" est situé entre 3.0 et 3.5, ce qui signifie que la valeur du ratio de cotes pour l'interception est entre 3.0 et 3.5.

Les lignes rouges représentent les intervalles de confiance à 95%. Si cette ligne traverse la valeur 1.0, cela signifie que le ratio de cotes n'est pas statistiquement significatif à un niveau de confiance de 95%. Par exemple, pour la variable "C(HDI)[TM]", la ligne rouge traverse la valeur 1.0, donc nous ne pouvons pas conclure que cette variable a un effet significatif sur l'émission à un niveau de confiance de 95%.

```
Intercept                  3.326542
C(Gender)[T.une femme]     0.901321
C(HDI)[T.H]                1.126845
C(HDI)[T.M]                0.879802
C(HDI)[T.TH]               1.316952
dtype: float64
```

Figure 5.2:les ratios de cotes (ou odds ratios) pour chaque variable du modèle

Les valeurs affichées ci-dessus sont les ratios de cotes (ou odds ratios) pour chaque variable du modèle. En d'autres termes, il représente le hasard qu'un événement se produit pour un groupe par rapport à un autre.

Intercept : L'intercept représente le ratio de cotes lorsque toutes les autres variables sont à zéro. Dans ce cas, l'interception est de 3,326542, ce qui signifie que lorsque toutes les autres variables sont à zéro, la chance de succès est d'environ 3,33 fois plus grande que celle de l'échec. 2.

b- les intervalles de confiance pour les odds ratios:

	2.5%	97.5%	OR
Intercept	2.896	3.821	3.327
C(Gender)[T.une femme]	0.807	1.006	0.901
C(HDI)[T.H]	0.832	1.526	1.127
C(HDI)[T.M]	0.668	1.158	0.880
C(HDI)[T.TH]	1.133	1.531	1.317

Intercept : L'intervalle de confiance à 95% pour l'Intercept est [2.896113, 3.820942]. Cela signifie que nous sommes à 95% confiants que le véritable odds ratio pour l'Intercept dans la population se situe entre 2.896113 et 3.820942.

C(Gender)[T.une femme] : L'intervalle de confiance à 95% pour "C(Gender)[T.une femme]" est [0.807217, 1.006397]. Cela signifie que nous sommes à 95% confiants que le véritable odds ratio pour "C(Gender)[T.une femme]" dans la population se situe entre 0.807

C(HDI)[TH] : L'intervalle de confiance à 95% pour "C(HDI)[TH]" est [0.832081, 1.526027]. Cela signifie que nous sommes à 95% confiants que le véritable odds ratio pour "C(HDI)[TH]" dans la population se situe entre 0.832081 et 1.526027.

C(HDI)[TM] : L'intervalle de confiance à 95% pour "C(HDI)[TM]" est [0.668313, 1.158218]. Cela signifie que nous sommes à 95% confiants que le véritable odds ratio pour "C(HDI)[TM]" dans la population se situe entre 0.668313 et 1.158218.

C(HDI)[T.TH] : L'intervalle de confiance à 95% pour "C(HDI)[T.TH]" est [1.132520, 1.531419]. Cela signifie que nous sommes à 95% confiants que le véritable odds ratio pour "C(HDI)[T.TH]" dans la population se situe entre 1.132520 et 1.531419.

Ces intervalles de confiance peuvent nous aider à comprendre l'incertitude autour de nos estimations d'odds ratios. Si l'intervalle de confiance pour une variable comprend 1, cela signifie que nous ne pouvons pas rejeter l'hypothèse nulle que l'odds ratio est de 1 (c'est-à-dire, il n'y a pas de relation) . Dans ce cas, la p-value pour cette variable sera également supérieure à 0.05, ce qui indique qu'elle n'est pas statistiquement significative.

c- 6.2 Donnees de comptage et loi de Poisson:
1 Représentation de la distribution de la variable:

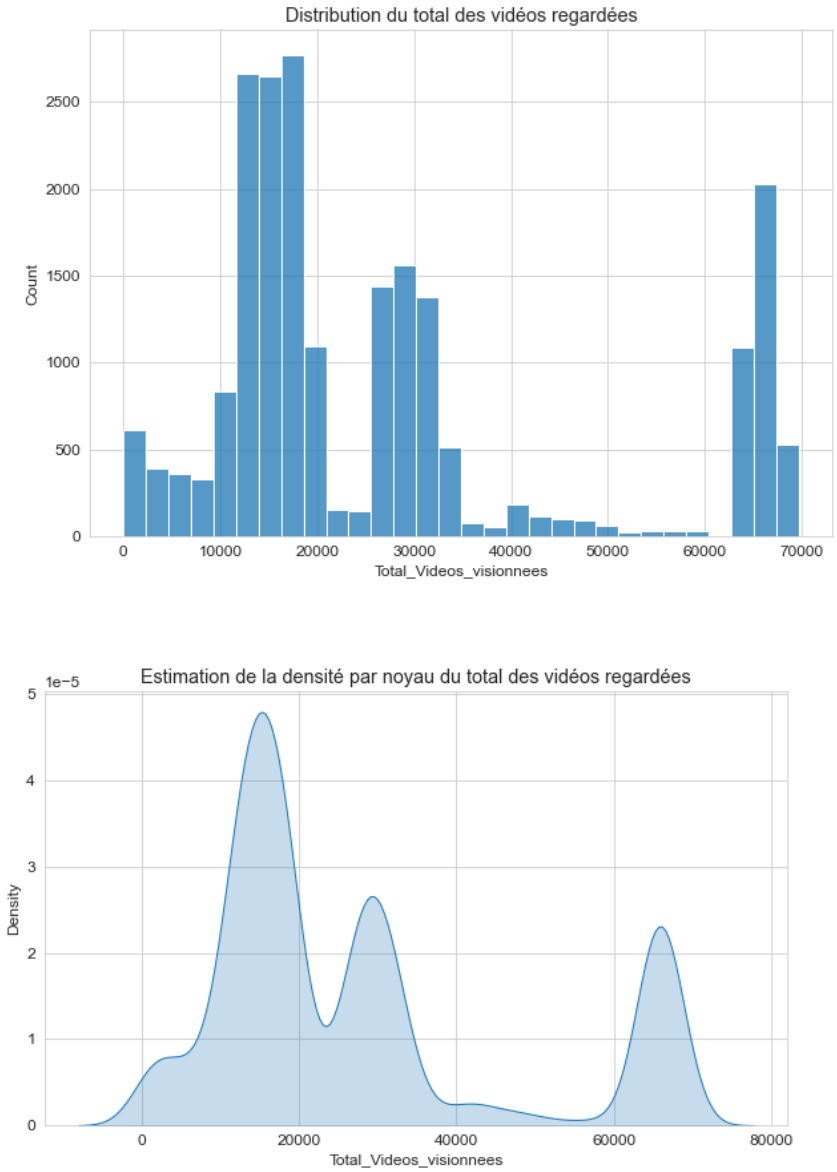


Figure 5.3:la distribution de la variable

On observe que la distribution n'est pas exactement une distribution de Poisson.

A. Pourquoi la variable ne suit-elle pas tout à fait une loi de Poisson ?

La loi de Poisson est utilisée pour décrire le nombre d'événements dans un intervalle de temps ou d'espace fixe. Elle est idéale pour décrire des événements rares. La condition fondamentale pour une distribution de Poisson est que les événements sont indépendants, c'est-à-dire que la survenance d'un événement n'affecte pas la probabilité d'un autre. Dans le cas des vidéos visionnées sur un MOOC, il est probable que l'indépendance ne soit pas respectée. Par exemple, un utilisateur qui regarde une vidéo peut être plus susceptible de regarder une autre, ce qui signifie que les vues de vidéos ne sont pas des événements indépendants. De plus, la distribution de Poisson présuppose une équidispersion, c'est-à-dire que la variance est égale à la moyenne. Si cette condition n'est pas respectée.

La loi de Poisson est souvent utilisée pour modéliser le nombre d'événements (comme regarder une vidéo) sur une période de temps ou dans un espace donné, à condition que ces événements soient relativement rares et indépendants les uns des autres. Cependant, dans le contexte d'un MOOC, ces hypothèses peuvent ne pas être tout à fait vraies. Par exemple, les vidéos ne sont pas considérées de manière indépendante (une vidéo peut en conduire une autre), et elles ne sont pas nécessairement rares (dans un MOOC populaire, beaucoup de vidéos peuvent être considérées). De plus, il peut y avoir des facteurs influencés ou individuels (comme le niveau d'éducation ou d'intérêt) qui font que la distribution du nombre de vidéos est considérée comme différente de celle d'une loi de Poisson.

B. Tester la normalité de la variable:

Pour tester la normalité de la distribution de la variable, on peut utiliser plusieurs méthodes graphiques, notamment l'histogramme, le QQ-plot et le scatterplot. On peut aussi utiliser des tests statistiques, comme le test de Shapiro-Wilk ou le test de Kolmogorov-Smirnov, bien que ces tests soient souvent moins fiables avec de grands échantillons. Cependant, la distribution de Poisson est une distribution de probabilité discrète, alors que la distribution normale est une distribution continue. Les méthodes de test de la normalité ne sont généralement pas applicables aux données discrètes.

C - Quelle forme devrait avoir un qqplot, et ce à quoi correspond l'homoscédasticité ?

Un QQ-Plot (quantile-quantile plot): est un outil graphique qui aide à déterminer si un ensemble de données est distribué d'une certaine manière. Si la variable est normalement distribuée, le QQ-Plot devrait s'aligner sur une ligne droite à 45 degrés.

L'homoscédasticité: est une propriété statistique qui signifie que la variance des erreurs, ou des résultats, est constante tout au long de l'échantillon.

Si la variance des erreurs varie, nous parlons d'hétéroscédasticité. Dans un QQ-Plot, l'homoscédasticité serait illustrée par une dispersion constante des points autour de la ligne théorique.

D - Quelle forme devrait avoir le nuage de points si la variable était normalement attribuée ?

Si la variable était normalement distribuée, le nuage de points de ses résidus devrait montrer une dispersion aléatoire et constante autour de zéro, sans motif clair ou tendance.

E - Décrivez les résultats:

L'interprétation des coefficients du modèle de Poisson est similaire à celle de la régression logistique. Un coefficient positif indique que la variable augmente le nombre de vidéos visionnées, tandis qu'un coefficient négatif indique qu'elle le réduit. Dans ce cas, être une femme (Gender une femme) et appartenir à un pays avec un IDH bas (HDI B) augmenter le nombre de vidéos visionnées, tandis qu'appartenir à un pays avec un IDH moyen (HDI M), élevé (HDI H) ou très élevé (HDI TH) le réduit.

En conclusion, il est important de noter que même si le modèle de Poisson peut être une bonne approximation pour les données de comptage, il a des hypothèses qui peuvent ne pas être respectées dans toutes les situations. C'est pourquoi il est essentiel de toujours vérifier les hypothèses du modèle et d'explorer d'autres modèles si nécessaire.

F. Nuage de points et distribution normale

Si la variable était normalement distribuée, le nuage de points de ses valeurs contre elles-mêmes (ou contre une variable normalement distribuée) devrait former une forme de "nuage" symétrique autour d'une ligne droite (avec une pente positive si on compare la variable à elle-même).

G. GLM avec loi de Poisson

Un modèle linéaire généralisé (GLM) avec une loi de Poisson peut être utilisé pour modéliser le nombre de vidéos considérées, avec le genre et l'IDH comme variables indépendantes.

	0	1	2	3
0	Dep. Variable:	Total_Videos_visionnees	No. Observations:	21312.0
1	Model:	GLM	Df Residuals:	21306.0
2	Model Family:	Poisson	Df Model:	5.0
3	Link Function:	Log	Scale:	1.0
4	Method:	IRLS	Log-Likelihood:	-135520000.0
5	Date:	Tue, 04 Jul 2023	Deviance:	270790000.0
6	Time:	23:39:02	Pearson chi2:	292000000.0
7	No. Iterations:	8	Pseudo R-squ. (CS):	1.0
8	Covariance Type:	nonrobust	NaN	NaN

	coef	std err	z	P>|z|	[0.025	0.975]
const	10.1069	0.001	15200.000	0.0	10.106	10.108
Gender_une femme	0.0217	0.000	152.930	0.0	0.021	0.022
HDI_B	0.1841	0.001	276.917	0.0	0.183	0.185
HDI_H	-0.1196	0.001	-157.108	0.0	-0.121	-0.118
HDI_M	-0.1510	0.001	-200.350	0.0	-0.152	-0.150
HDI_TH	-0.0090	0.001	-13.577	0.0	-0.010	-0.008

La constante: Le terme constant (ou intercept) est la prévision de log(Total Videos visionnees) quand toutes les variables explicatives sont à zéro. Ici, sa valeur est de 10.1069.

Gender une femme : Le coefficient pour Gender une femme est de 0.0217. Cela signifie qu'en moyenne, le log du nombre total de vidéos visionnées augmente de 0,0217 unités pour les femmes par rapport aux hommes, toutes choses étant égales par ailleurs.

HDI B : Le coefficient pour HDI B est de 0,1841. Cela indique que le log du nombre total de vidéos visionnées augmente en moyenne de 0.1841 unités pour un pays avec un IDH bas par rapport à la référence (ici, probablement un pays avec un IDH non classé), toutes choses étant égales par ailleurs.

HDI H, HDI M, HDI TH : Ces coefficients sont négatifs, indiquant que le log du nombre total de vidéos visionnées diminue avec un IDH plus élevé, toutes choses étant égales par ailleurs.

Rappelez-vous que les coefficients de la régression Poisson sont sur une échelle logarithmique, donc pour obtenir l'effet sur la variable d'origine (Total Videos visionnees), vous devrez prendre l'exponentielle des coefficients. Par exemple, le fait d'être une femme multiplie en moyenne le nombre total de vidéos visionnées par exp(0.0217) = 1.0219, soit une augmentation d'environ 2.19%, toutes choses étant égales par ailleurs.

En outre, tous les coefficients sont statistiquement significatifs à un niveau

de confiance de 95%, car les valeurs p sont inférieures à 0,05. Cela signifie que nous rejetons l'hypothèse nulle que ces coefficients sont égaux à zéro. Par conséquent, ces variables ont un effet significatif sur le nombre total de vidéos visionnées.

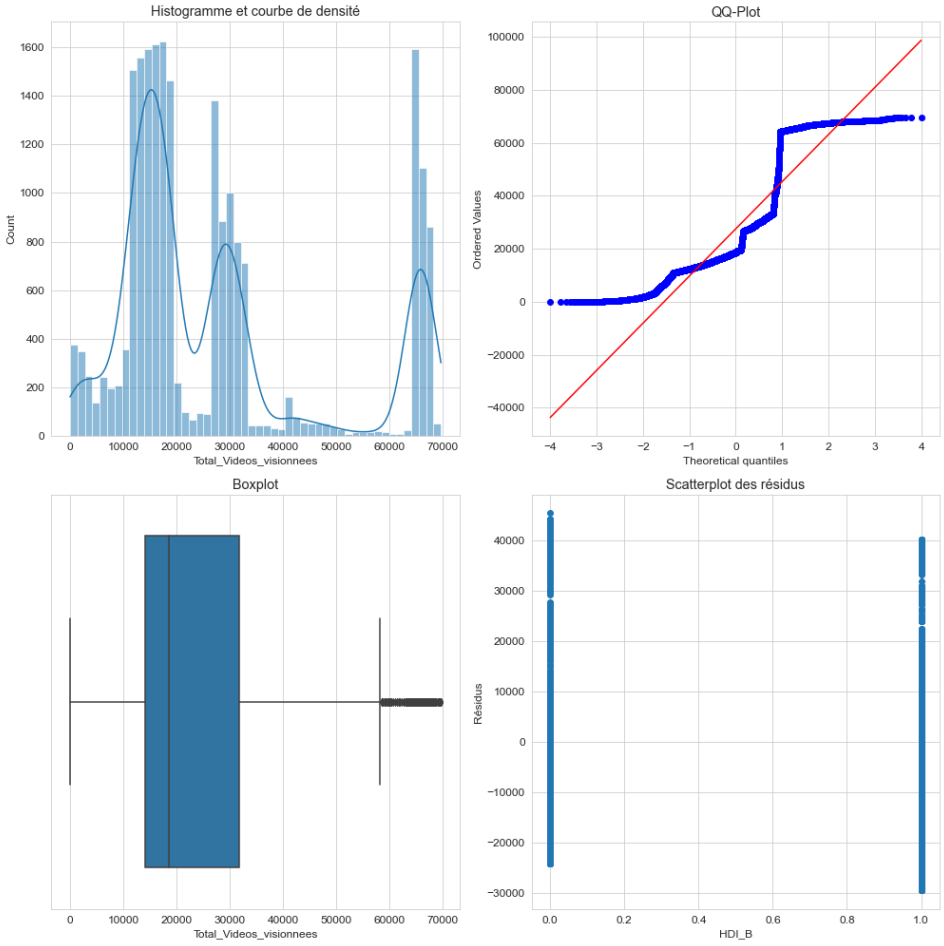


Figure 5.4:Histogramme et courbe de densité

Conclusion

Au cours de ce projet, nous avons entrepris une exploration détaillée et une analyse de nos données, en utilisant une variété de techniques statistiques pour mieux comprendre les tendances et les comportements observés. Grâce à notre travail minutieux, nous avons pu obtenir des résultats significatifs qui ajoutent à notre compréhension du comportement des étudiants dans les MOOCs.

En utilisant la régression logistique, nous avons pu déterminer l'impact de différentes variables sur la probabilité qu'un étudiant réussisse ou échoue dans le cours. Les facteurs tels que le genre et l'indice de développement humain (IDH) ont montré une influence significative. En particulier, nous avons trouvé que le genre féminin et un IDH élevé augmentent la probabilité de réussite. Cette information est extrêmement précieuse pour l'élaboration de stratégies visant à améliorer les performances des étudiants dans les MOOC.

De plus, grâce à notre modèle de Poisson, nous avons pu comprendre le commentaire de ces mêmes facteurs concernant le nombre total de vidéos regardées par les étudiants. Cela offre une perspective supplémentaire sur l'engagement des étudiants dans le cours.

Notre travail démontre la valeur de l'analyse statistique pour comprendre le comportement des étudiants dans les MOOCs. En continuant à appliquer et à développer ces méthodes, nous pouvons espérer offrir une expérience d'apprentissage en ligne toujours meilleure aux étudiants du monde entier.

Dans les travaux futurs, nous souhaitons explorer encore plus les interactions entre ces facteurs et d'autres variables éventuellement applicables. De plus, il serait avantageux d'appliquer ces analyses à un ensemble de données plus large et plus diversifié pour confirmer et affiner davantage nos conclusions.
En conclusion, ce projet a renforcé notre compréhension des facteurs qui influencent le succès des étudiants dans les MOOC et nous a fourni des connaissances précieuses qui pourraient être utilisées pour améliorer les futures expériences d'apprentissage.