DU DATA ANALYST
Université de Cergy
UE de statistiques

Rapport final

Table des matières

I. Les données

Le jeu de données utilisées dans ce projet porte sur l'analyse de l'apprentissage par l'intermédiaire d'un MOOC, MOOC Effectuation. La focale porte sur l'engagement des apprenants, et notamment sur le visionnage de vidéos et la réalisation de quizz.

On définit les 4 catégories d'apprenant suivant :

Completer	S'ils ont passé l'examen ou obtenus le certificat
Auditing learner	Aucun quiz réalisé, aucun devoir soumis mais plus de 6 vidéos visionnées
Disengaging learners	Un quiz a été réalisé ou un devoir soumis mais le certificat n'a pas été obtenu ou l'examen pas réalisé
Bystander	Aucun quiz réalisé, aucun devoir soumis et 5 ou moins de 6 vidéos visionnées

Tableau 1: Catégorie des apprenants. MOOC Effectuation

II. Préparation des données

Création du jeu de données

Le jeu de données est obtenu par concaténation horizontale dans un premier temps des fichiers de données source dont nous disposons qui correspondent aux 3 itérations du MOOC.

Shape de QUEST1: (8986, 35)
Shape de QUEST2: (4078, 40)
Shape de QUEST3: (4233, 26)
Shape de EFFEC1: (7965, 73)
Shape de EFFEC2: (3798, 74)
Shape de EFFEC3: (3883, 76)

Tableau 2: Caractéristiques des fichiers sources

En premier lieu les colonnes manquantes dans les différentes sources sont complétées (44 colonnes pour QUEST et 76 colonnes pour EFFEC), puis on fait une jointure extérieure des jeux QUEST et EFFEC avec comme clé le numéro d'étudiant. La jointure extérieure permet de prendre en compte les étudiants qui n'apparaitraient pas dans les 2 fichiers en même temps.

Puis on concatène verticalement toutes les lignes en supprimant les données non utilisées par la suite.

Remarque : les doublons de la colonne Student_ID sont conservés car il est possible qu'un étudiant ait suivi plusieurs itérations du même MOOC avec des résultats distincts.

On construit ensuite les colonnes Nb_Videos_Visionnees, sommes des valeurs des colonnes SxLy et Nb_Quiz_Realises somme des valeurs des colonnes 'Quizz.x.bin'. La colonne HDI est ajoutée pour regrouper dans la modalité « I » les modalités « M » et « H ».

Modalités	Nombre
B	1032
I	667
TH	7270

Tableau 3: Comptage des modalités de HDI

III. Description du jeu de données

Création des statuts

Les différents types de statuts sont calculés avec les règles suivantes :

Bystander : Nb_Videos_Visionnees < 6 et Nb_Quiz_Realises = 0 et Assignment.bin = 0

Auditing_Learner :
Nb_Videos_Visionnees > 6 et Nb_Quiz_Realises = 0 et Assignment.bin = 0

Disengaging_Learner :
(Assignment.bin ou Nb_Quiz_Realises >= 1) et (Exam.bin = 0 ou Certif.bin = 0)

Completer : Exam.bin >= 1 ou Certif.bin >= 1

Proportion des apprenants par itérations

La table donnant en lignes les proportions des quatre types d'apprenants que nous avons définis (bystander, auditing, completer, disengaging), et en distinguant en colonne les 3 itérations, avec le nombre total d'étudiant et la proportion de chaque statut pour chacune des itérations.

iteration_quest	**1**	**2**	**3**
auditing_learner	1.05	1.56	0.98
bystander	57.97	46.95	50.75
completer	0.37	24.82	24.05
disengaging_learner	40.61	26.66	24.22
Nb total	5415.00	3529.00	3460.00
Prct total	100.00	100.00	100.00

Tableau 4: Proportions des statuts des apprenants par itération

IV. Chi2 et Mosaic plot

On croise les variables HDI et Gender en faisant une table pivot sur ces colonnes.

Gender	Un homme	Une femme
HDI		
B	883	147
I	432	233
TH	4711	2545

Tableau 5: Données du pivot HDI et Gender

Calcul du Chi2

Faisons l'hypothèse nulle que dans l'étude de l'engagement des étudiants à suivre le MOOC Effectuation les variables HDI et Genre sont indépendantes.

Les prérequis pour le calcul du Chi2 sont respectés car les deux variables sont qualitatives et les effectifs sont suffisants (>5) dans plus de 80% des cas.

Le Chi2 donne un résultat de 179 et une p-value < 0.001. Cela montre une différence assez notable entre les données observées et celles que nous aurions si les variables HDI et Gender étaient indépendantes, c'est-à-dire si dans le cadre de l'utilisation des MOOC l'indice HDI du pays d'une personne ne dépendait pas de son genre. La p value très inférieure à 0.01 indique une probabilité très faible d'avoir un tel écart si l'hypothèse nulle est vraie. Cela nous permet de réfuter l'hypothèse nulle.

On cherche à savoir si, dans le cadre du suivi de l'engagement des apprenants au MOOC Effectuation, il existe une relation de dépendance entre les variables HDI et Gender. En d'autres termes si le niveau d'HDI des pays des apprenants dépend du fait que l'apprenant est un homme ou une femme. Le Chi2 de 179 et la p_value < 0.001 montre qu'il existe une dépendance effective mais ce lien de dépendance est faible (V de Cramer = 0.1, bien inférieur à 1).

Représentation graphique

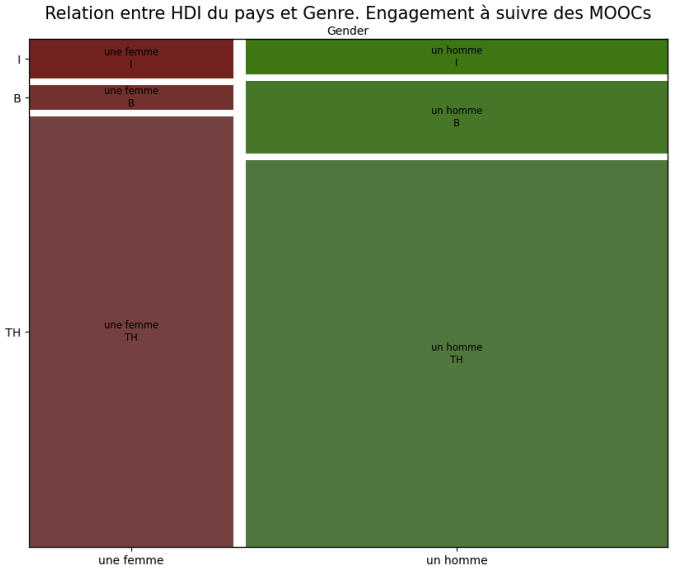

Figure 1:Mosaique du Chi2 de HDI et Gender. Engagement MOOC Effectuatione

La proportion d'hommes qui participent au MOOC Effectuation est nettement plus importante que la proportion de femmes essentiellement dans les pays d'indice de développement humain bas (« B »). On peut l'expliquer par la plus grande facilité d'accès à Internet pour les hommes que pour les femmes dans les pays à HDI bas, ou plus simplement parce que dans ces pays les femmes sont plus occupées à des tâches ménagères à la maison ou au travail dans les campagnes, leur donnant moins de liberté pour se connecter au MOOC.

Ici les couleurs utilisées permettent uniquement de distinguer les hommes des femmes mais nous aurions pu n'utiliser qu'une seule couleur. Nous ne représentons pas les résidus il n'y a donc pas d'information à visualiser sur une « sur » ou « sous » représentation de ces populations par HDI de leur pays ce qui nous obligerait à utiliser plusieurs couleurs.

V. Modèle linéaire. Tests non paramétriques

5.1 Comparaison du # de vidéos selon le genre. Test U de Mann-Whitney. Test non paramétrique

Les distributions du # de vidéos selon le genre ne sont pas normales (la médiane n'est pas située entre les 1er et 3e quartiles dans la figure 2) nous effectuons un test non paramétrique pour effectuer cette comparaison. On suppose toujours que les deux échantillons sont aléatoires.

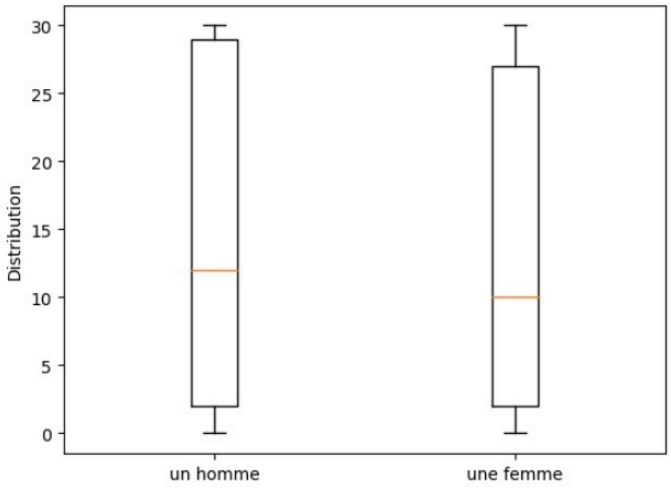

Figure 2: Distribution non normale du # de vidéos par Genre

Le test U de Mann-Whitney (test non paramétrique, équivalent du test t à 2 échantillons) nous donne stats= 8711923.0 et une p-value inférieure à 0.001. On peut rejeter l'hypothèse nulle (Médiane des vidéos visionnées par les hommes – Médiane des vidéos visionnées par les femmes) = 0, et en conclure que les médianes sont différentes. Cela signifie que l'engagement des hommes et des femmes vis-à-vis du MOOC Effectuation n'est pas similaire. Il faut encore trouver la population pour laquelle l'engagement est le plus important et quelle en est la raison.

5.2 Lien entre le # de quiz réalisés et le # de vidéos visionnées

Corrélation de Spearman
La corrélation de Spearman, test non paramétrique pour des valeurs non normalement distribuées, montre la même corrélation fortement positive, (stats = 0.80 et p-value=0).

Illustration par un scatterplot pour illustrer la corrélation

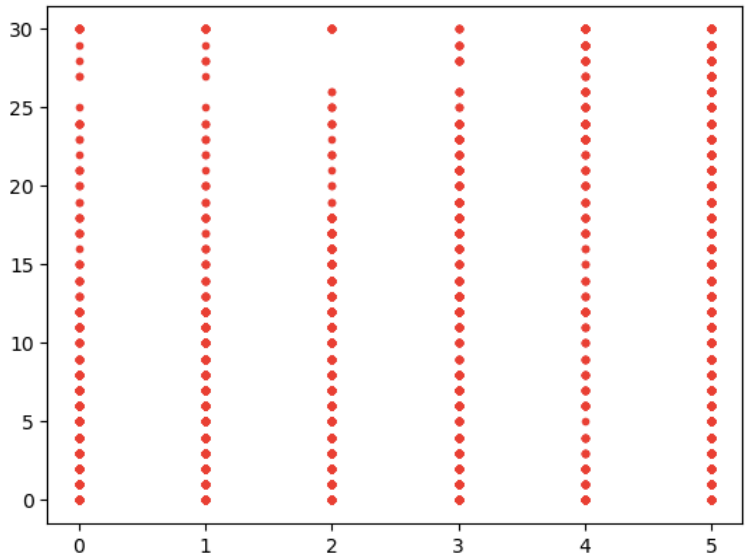

Figure 3: Répartition des # de quiz et des # de vidéos

Test d'ANOVA

L'analyse de l'effet du genre et de l'HDI sur le nombre de vidéos visionnées est faite avec un test d'ANOVA. C'est un test paramétrique il faut donc que les distributions soient de forme normale, que les variances soient égales entre les groupes et que les groupes soient indépendant.

Nous allons effectuer une ANOVA à deux facteurs (Genre et HDI) ayant plusieurs modalités, 2 pour Genre (un homme, une femme) et 3 pour HDI (B, I, TH), dans le premier cas sans prendre en compte l'interaction entre Genre et HDI. Les étapes non représentées ici sont : i) calculer un modèle de régression linéaire (ordinary least square) ii) utiliser la fonction anova_lm de la librairie statsmodels avec ces modèles.

Les hypothèses nulles H0 sont que les Moyennes du Nb de vidéos pour les hommes sont similaires aux Moyennes du Nb de vidéos pour les femmes, et que les Moyennes du Nb de vidéos pour les pays à indice Très Haut soient similaires aux Moyennes du Nb de vidéos pour les pays d'indice Intermédiaire et similaires aux Moyennes du Nb de vidéos pour les pays d'indice Bas.

Les variances du # de vidéos visionnées par les hommes et par les femmes sont similaires (pour les hommes : 137.31, pour les femmes : 140.66). En revanche les variances pour les indices HDI ne sont pas similaires (B :79.5, I : 128, TH : 138). Enfin, les données de Genre sont indépendantes car les données Homme et Femme s'excluent mutuellement, de même que les données d'HDI qui n'accepte qu'un seul statut pour un pays.

Résultat :

	Df	Sum_Sq	Mean_Sq	F	PR (>F)
C(HDI)	2.0	75972.38	37986.2	291.61	<< 0.0001
C(Gender)	1.0	57.4	57.43	0.44	0.51
Residual	8947.0	1165380	130.25	NaN	NaN

Le F représente le rapport entre la variance entre tous les échantillons et la variance à l'intérieur des échantillons. Par exemple, pour l'indice HDI, F = 291 signifie que la variance au sein de la totalité des échantillons est largement supérieure à la variance au sein de chaque échantillon, le « bruit » des

résidus de chaque échantillon est négligeable, ce qui signifie un écart statistique important entre les différentes modalités de la variable HDI. La p-value, PR, nous indique la probabilité d'obtenir le même résultat pour F si l'hypothèse nulle est vérifiée. Ici PR est très inférieur à 0.0001 on peut donc rejeter l'hypothèse nulle et conclure à une différence statistique entre les modalités de la variable HDI.
Pour le Genre en revanche le F est faible (0.44) et la p-value est largement au-dessus de 0.01, on ne peut donc conclure sur l'effet des modalités de Genre sur le nombre de vidéos visionnées.

Le DDL 2 pour HDI est obtenu par la formule (nombre de modalité de HDI – 1), soit 3-1=2. Un étudiant donné habite forcément un pays qui a un indice de développement humain soit Bas soit Intermédiaire soit Très Haut, donc si on connait l'état de 2 indices HDI alors la valeur du 3^{e} est forcément connue (ex : HDI Bas : Faux et HDI Très Haut : Vrai donc HDI Intermédiaire : Faux.
Le DDL 1 pour Gender est obtenu par la formule (nombre de modalité de Gender – 1), soit 2-1=1. Ici de la même manière si on sait qu'un étudiant est un Homme on sait automatiquement qu'il n'est pas une femme, et inversement, d'où le degré de liberté de 1.