Sujet 2 : Analyse statistique sur data et e-commerce avancée

Introduction

Dans un contexte de forte concurrence dans le secteur du e-commerce, la fidélisation client représente un enjeu stratégique majeur. Comprendre les comportements qui favorisent ou freinent cette fidélité permet aux entreprises de mieux cibler leurs actions marketing, optimiser l'expérience utilisateur, et augmenter la valeur vie client. L'objectif de cette étude est d'analyser, à partir de données comportementales, les principaux facteurs liés à la fidélisation, et de construire des modèles statistiques permettant de prédire cette fidélité.

1. Analyse exploratoire des données

Objectif
Cette première étape vise à comprendre la structure générale du dataset, identifier d'éventuelles anomalies ou tendances, et préparer le terrain pour les analyses statistiques et prédictives à venir. Il s'agit également d'observer comment les différentes variables sont réparties et de détecter des motifs intéressants liés à la fidélisation client.

Méthodologie et outils utilisés
L'analyse a été réalisée avec Python, en utilisant :

- la bibliothèque pandas pour charger et manipuler les données ;
- seaborn et matplotlib pour créer des visualisations statistiques ;
- les fonctions describe(), isnull(), value_counts() et des graphes pour explorer les données.
- ChatGPT pour rédiger le code et modifier les erreurs

Structure du dataset
Le fichier contient 800 clients d'un site e-commerce. Chaque ligne représente un client avec plusieurs caractéristiques comme son âge, son revenu annuel, le temps moyen passé par session, le montant moyen de ses paniers, son niveau de satisfaction, son nombre de visites via mobile ou web, ainsi que la variable cible : la fidélisation (valeur 1 pour client fidèle, 0 sinon).

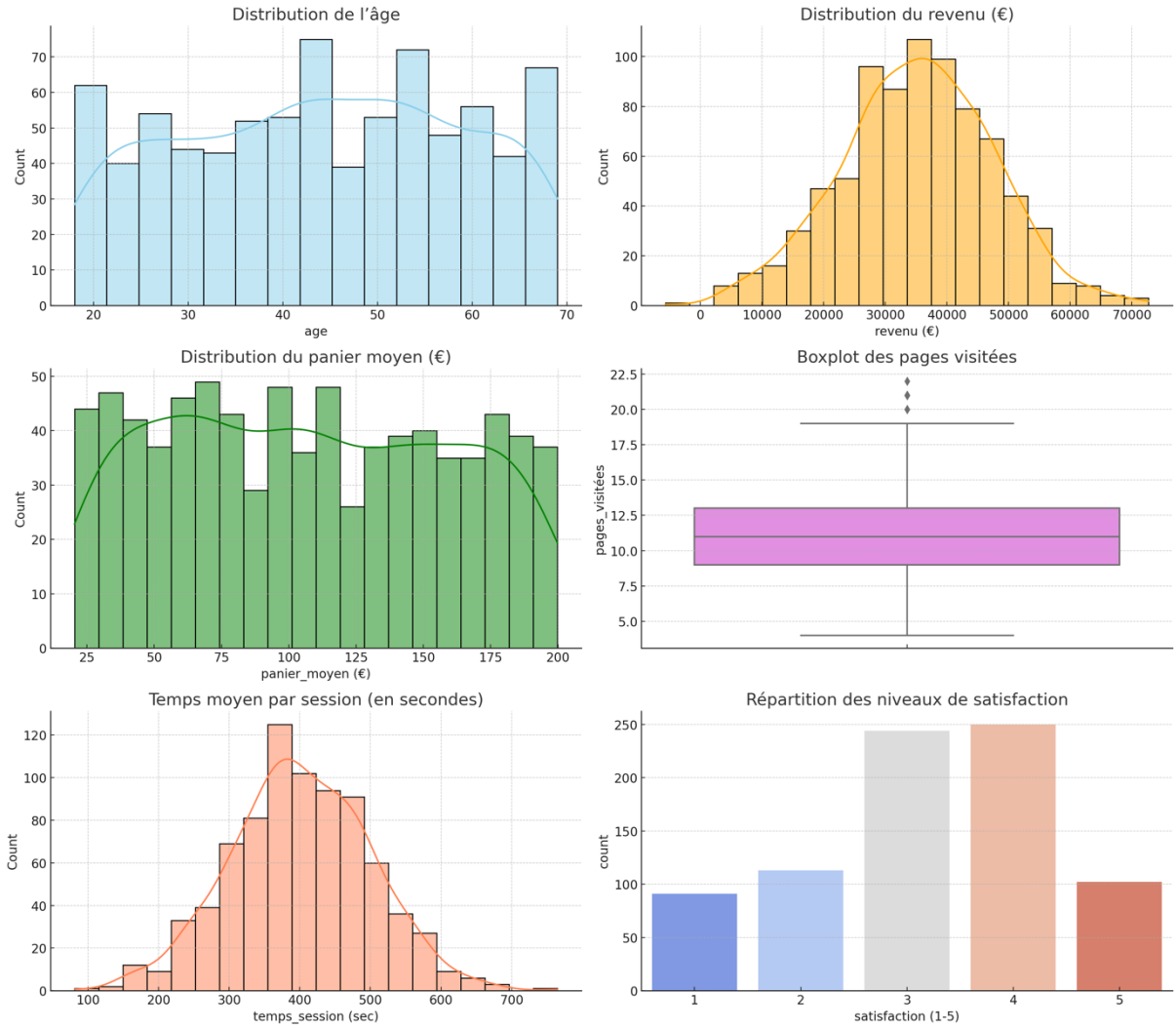
Il n'y a aucune valeur manquante dans le fichier, ce qui permet de travailler directement sans traitement préliminaire.

Analyse des distributions
L'âge des clients est assez bien réparti, avec une concentration entre 30 et 50 ans. La distribution du revenu est asymétrique, avec des cas de revenus très élevés qui peuvent être considérés comme des valeurs extrêmes. Le montant moyen du panier présente une distribution similaire, avec une majorité de clients entre 50 et 150 euros, et quelques clients au-delà de 300 euros.

Le nombre de pages visitées varie beaucoup selon les clients, ce qui peut refléter différents comportements d'achat. Le temps passé sur le site (en secondes) est également très dispersé, allant de sessions brèves à très longues.

La variable de satisfaction, qui va de 1 à 5, est centrée sur les notes intermédiaires : la majorité des clients ont donné une note de 3 ou 4, tandis que très peu ont noté 1 (forte insatisfaction).

Enfin, la répartition de la variable cible "fidélisation" montre un léger déséquilibre, avec un peu plus de clients fidèles que non fidèles, mais les deux groupes sont suffisamment représentés pour entraîner un modèle de classification sans rééchantillonnage pour l'instant.



Conclusion de cette étape
Les données sont globalement bien structurées, complètes, et prêtes pour une analyse statistique plus poussée. On remarque déjà quelques variables potentiellement explicatives de la fidélisation, comme la satisfaction, le nombre de pages visitées, ou le montant du panier moyen. La prochaine étape consistera à tester ces relations plus formellement.

2. Tester les relations entre variables

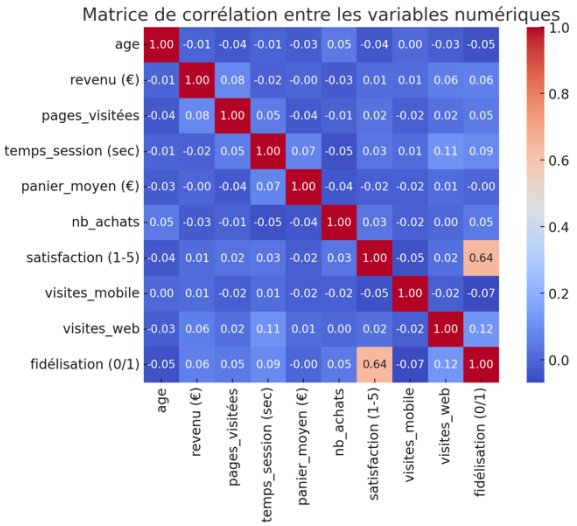
Objectif
L'objectif est d'identifier les variables ayant une influence statistiquement significative sur la fidélisation client, en examinant leurs relations avec la variable cible. Cette étape permet de sélectionner les variables pertinentes pour la modélisation prédictive à venir.

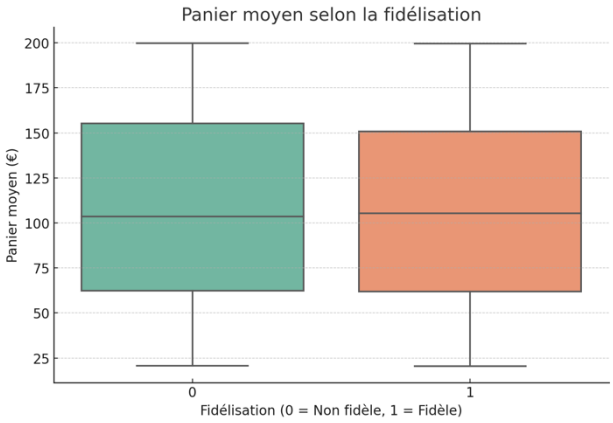
Méthodologie
L'analyse a été conduite avec Python en utilisant :

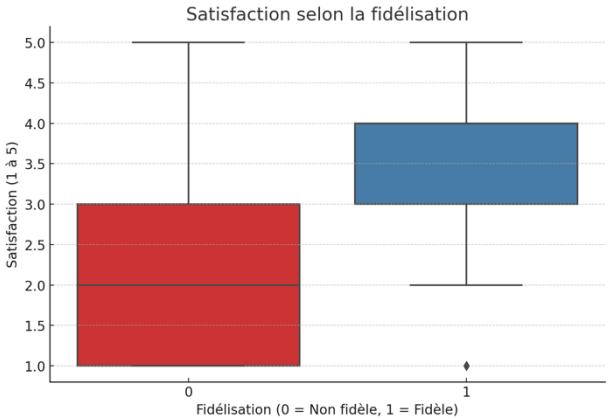
- La fonction corr() pour mesurer les corrélations de Pearson entre variables numériques
- Le test du Khi^2 (chi2_contingency) pour analyser les relations entre variables qualitatives
- Le test t de Student (ttest_ind) pour comparer les moyennes de variables quantitatives entre deux groupes (clients fidèles vs non fidèles)

- Les bibliothèques seaborn et matplotlib pour les visualisations
- ChatGPT pour rédiger le code et modifier les erreurs

Corrélations avec la fidélisation
Une matrice de corrélation a été calculée pour l'ensemble des variables numériques. Le graphique ci-dessous en donne un aperçu visuel.







La variable "satisfaction" se distingue par une corrélation élevée (≈ 0.64) avec la fidélisation. Cela indique qu'un niveau de satisfaction plus élevé est fortement associé à la fidélité du client.

Les variables "visites web" (≈ 0.12) et "temps de session" (≈ 0.09) montrent une corrélation plus modérée mais positive. Elles semblent donc également jouer un rôle dans la fidélisation, bien qu'à un degré moindre.

Les autres variables comme l'âge, le revenu, le nombre de pages visitées ou le panier moyen n'affichent pas de corrélation significative avec la fidélité.

Test du Khi^2 : Sexe et fidélisation
Nous avons ensuite vérifié s'il existait un lien entre le sexe du client et la fidélisation à l'aide d'un test du Khi^2.

Voici la table de contingence utilisée :

	Non fidèle	Fidèle
Femme (F)	116	305
Homme (M)	95	284

Le test donne une p-value de 0.47. Ce résultat indique qu'il n'existe pas de relation significative entre le sexe du client et sa fidélité. Hommes et femmes sont représentés de manière équivalente dans les deux groupes.

Test t de Student : Panier moyen
Un test t de Student a été mené pour comparer la moyenne du panier d'achat entre les clients fidèles et non fidèles. Le graphique ci-dessous montre la distribution des paniers selon la fidélisation.

La p-value obtenue est de 0.95, ce qui est largement supérieur au seuil de significativité de 5 %. Il n'y a donc aucune différence significative entre les deux groupes. Cela signifie que le montant moyen dépensé par achat n'est pas un bon indicateur de fidélisation dans ce dataset.

Comparaison des niveaux de satisfaction
Le graphique suivant compare les niveaux de satisfaction des clients fidèles et non fidèles.

On observe très clairement que les clients fidèles ont des niveaux de satisfaction plus élevés. Cela confirme la forte corrélation positive détectée plus tôt. Cette variable est donc un facteur déterminant de la fidélisation.

Conclusion de cette étape
L'analyse met en évidence trois enseignements clés :

- La satisfaction est le facteur le plus fortement lié à la fidélisation.
- Le temps passé sur le site et le nombre de visites web jouent également un rôle, mais dans une moindre mesure.
- Aucune relation significative n'a été détectée avec le sexe ou le panier moyen, qui semblent donc être des variables peu explicatives.

Ces résultats guideront le choix des variables à intégrer dans les modèles statistiques de la prochaine section.

3. Modèles statistiques adaptés

Objectif
L'objectif de cette partie est de modéliser la fidélisation client en mobilisant des méthodes statistiques complémentaires. Nous combinons ici deux approches :

- une régression logistique, permettant de prédire la probabilité de fidélisation à partir de variables quantitatives ;
- une analyse factorielle des correspondances (AFC), visant à explorer les liens entre les modalités de variables qualitatives et la fidélisation.

Cette approche conjointe permet à la fois d'expliquer la fidélisation en termes de probabilité prédictive et de profil comportemental.

L'analyse a été réalisée avec le langage Python, en mobilisant plusieurs bibliothèques adaptées aux traitements statistiques et à la modélisation :

- pandas : pour la gestion et la manipulation des données
- scikit-learn (sklearn) : pour la régression logistique, le découpage en jeu d'entraînement/test, la normalisation des variables, et l'évaluation des performances via la matrice de confusion et les scores de classification
- matplotlib et seaborn : pour la visualisation des résultats (graphique de corrélation, boxplots, etc.)
- prince (ou équivalent) : pour la réalisation de l'Analyse Factorielle des Correspondances (AFC)
- ChatGPT pour rédiger le code et modifier les erreurs

Régression logistique

Méthodologie
La régression logistique est adaptée à la prédiction d'une variable binaire. Elle permet d'estimer la probabilité qu'un client soit fidèle (1) ou non (0), à partir d'un ensemble de variables explicatives.

Nous avons sélectionné trois variables quantitatives pour alimenter le modèle :

- la satisfaction (note de 1 à 5)
- le temps moyen passé sur le site (en secondes)
- le nombre de visites via le site web

Ces variables ont été choisies pour leur corrélation positive avec la fidélisation, identifiée dans les étapes précédentes.

Les données ont été normalisées, puis divisées en deux ensembles : 70 % pour l'entraînement du modèle, 30 % pour l'évaluation.

Résultats
Le modèle atteint une précision globale d'environ 86 %. Il détecte très bien les clients fidèles, bien que légèrement moins performant pour identifier les non fidèles.

La variable la plus influente est la satisfaction, suivie par le temps de session et les visites web. Le modèle confirme ainsi que l'expérience utilisateur joue un rôle clé dans la fidélisation.

Matrice de confusion

	Non fidèle (prédit)	Fidèle (prédit)
Non fidèle (réel)	46	24
Fidèle (réel)	11	159

Scores De Classification (Régression Logistique)

		precision	recall	f1-score	support
1	0	0.8070175438596491	0.6571428571428571	0.7244094488188977	70.0
2	1	0.8688524590163934	0.9352941176470588	0.9008498583569404	170.0
3	accuracy	0.8541666666666666	0.8541666666666666	0.8541666666666666	0.8541666666666666
4	macro avg	0.8379350014380212	0.796218487394958	0.812629653587919	240.0
5	weighted avg	0.8508172754290096	0.8541666666666666	0.849388072241678	240.0

Analyse Factorielle des Correspondances (AFC)

Méthodologie
L'AFC est une méthode descriptive utilisée pour analyser les correspondances entre des variables qualitatives. Elle permet de représenter visuellement les relations entre les modalités et de faire émerger des profils types de clients.

L’analyse a été réalisée à partir des modalités suivantes :

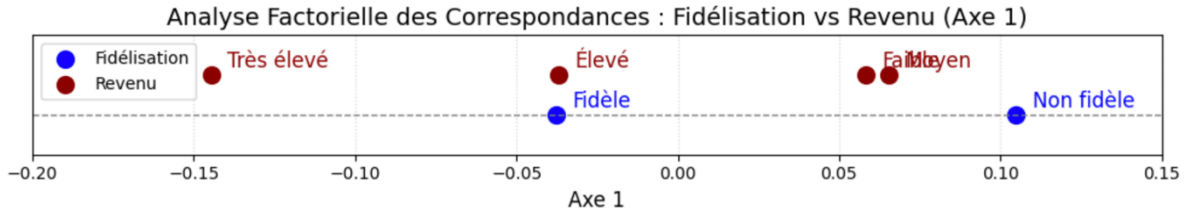
- fidélisation (0 ou 1)
- satisfaction (discrétisée)
- éventuellement d’autres variables catégorisées (temps de session, type de visite, etc.)

Le graphique factoriel obtenu permet d’interpréter la proximité entre modalités et de détecter les regroupements structurels.

Résultats
L’AFC révèle une association forte entre la fidélisation et les niveaux élevés de satisfaction. Les modalités correspondant à une satisfaction faible ou à une faible activité sur le site sont éloignées de la fidélisation.

Cela permet d’identifier visuellement des profils clients fidèles (satisfaits, actifs sur le site web) et des profils à risque de non-fidélisation.

Cette méthode complète la régression logistique en apportant une lecture qualitative et visuelle des comportements de fidélité.



Conclusion
La combinaison de ces deux approches permet d’obtenir une vue complète de la fidélisation :

La régression logistique fournit un modèle prédictif robuste, utile pour des décisions opérationnelles automatisées (score de fidélité).
L’AFC permet une lecture exploratoire riche, utile pour segmenter les clients, adapter les messages ou cibler les actions marketing.
Ces deux méthodes convergent vers le même constat : la satisfaction client est le facteur central de la fidélisation, renforcé par un usage fréquent et prolongé du site.

Conclusion

Cette étude a permis de démontrer que la fidélisation client peut être prédite avec une bonne précision à partir de variables comportementales simples. La variable de satisfaction s’avère être le facteur déterminant, renforcée par le temps passé sur le site et les visites web. Les analyses statistiques, combinées à des méthodes de modélisation comme la régression logistique et l’AFC, ont confirmé que la fidélité s’appuie à la fois sur des critères mesurables (quantitatifs) et des profils comportementaux (qualitatifs). Ces résultats peuvent être utilisés pour construire un score de fidélité ou pour orienter les campagnes de personnalisation et de rétention client.