Projet 4 : Analyser les ventes de votre entreprise

SOMMAIRE

Nettoyage des données

Produits et clients tests

Premièrement, j'ai trouvé des valeurs qui ne respectées pas les contraintes de forme (dont le nom était aberrant par rapport aux autres). Les produit test (T_0) et les clients tests (ct_0 et ct_1).

J'ai considéré que ses valeurs avaient été créé au départ pour faire des tests et qu'aujourd'hui elle ne faisait que polluer les données, je les ai donc supprimées.

J'ai réalisé une restriction sur les trois tables.

J'ai enlevé T_0 des tables products et transactions et j'ai enlevé ct_0 et ct_1 de la table customers

Importation de données manquante dans la table transactions

Deuxièmement, j'ai join ma table transaction à ma table produit pour avoir les prix et les catégorie de produit affiché dans ma table transations

j'ai remarqué que le produit 0_2245 n'avait ni prix ni catégorie. En fait le livre n'était pas dans la table products.

Au vu de son nom je l'ai mis dans la categ 0 et je lui ai donné le prix median des produits de cette catégorie.

Dans la pratique :

J'ai d'abord regardé si ma table contenait des valeurs non renseignées. Il y en avait.

J'ai ensuite extrait ces valeurs avec une restriction dans une table que j'ai nommé na_table.

J'ai remplacé les na par les valeurs adéquates que j'ai cité précédemmment.

Puis j'ai enlever les lignes contenant les Na de ma table transaction et je l'ai collé à la table que je viens de créer.

Enfin, j'ai ajouté le produit à ma table products.

Quatre très gros clients

Troisièmement, j'ai détecté 4 clients aberrants, en moyenne, ils achètent sur notre site plus de 3 fois par jour et ont dépensés plus de 50 000 euros chez nous cette année.

En comparaison le 5em plus gros acheteur dépense seulement 2564.25 euro chez nous et il achète un livre une fois tout les 5 jours.

Je les ai filtré quand j'ai voulu étudier les comportements des clients, mais je les ai gardé pour le reste (gain de l'entreprise, étude des produits…)

Dans la pratique je les ai repéré au moment ou j'ai ordonné ma table par montant total.

Et pour pouvoir les filtrer ou non selon mes besoin j'ai créé un vecteur contenant leurs 4 id_client

Clients inactifs pendant 1 an

Quand j'ai joint les transactions et les produits, j'ai remarqué que certains clients n'avaient rien achetés cette année.

J'ai supprimé les clients qui n'ont rien dépensé cette année de la liste des clients.

Je les ai supprimé en faisant une jointure à gauche au lieu d'une jointure pleine.

Exucative Summary

Voila maintenant un an que nous stockons les données dans notre

boutique en ligne.

Quelle bilan peut-on en tirer ?

On va étudier les différents insights analysé et et je vous ferai part de certaine recommandation

L'entreprise trébuche en octobre

Alors que l'argent gagné par notre bibliothèque en ligne semblait stable, 15830 euros par jour en moyenne, avec un minimum de 13792, il y a eu une chute impressionnante des gains de l'entreprise en octobre. 9187 euros de moyenne entre le 02/10 et le 27/10.

Mais la courbe reprend son court normal peu de temps après, le 27 octobre.

Les gains par jour recommencent à grimper à partir de février et atteignent leur maximum mi-février 20313 euros.

Zoom sur les catégories de livre

Avant d'étudier d'où vient la chûte des gains de la boutique, nous devons analyser les différentes catégories de livres.

La boutique en ligne en possède trois.

Globalement les livres de catégorie 0 sont très peu chers, les livres de catégorie 1 sont un peu plus chers et les livres de catégorie 2 sont chers.

Il faut que je me renseigne auprès des développeurs pour savoir à quoi c'est catégorie correspondent afin de mieux comprendre les données.

La catégorie 1, première responsable

Cette chute a eu lieu parce qu'aucun livre de la catégorie 1 n'a été vendu entre le 02 et le 27/10/2021.

Oui j'ai bien dit aucun, alors qu'on en vendait 283 par jour en moyenne, et avant ça le minimum qu'on en ai vendu en un jour c'était 209.

On observe aussi que le 01/10 il y a déjà une chute du nombre de livre vendu de la catégorie 0 assez importante.

Ces deux chutes interviennent juste après la ventes records en terme de nombre de livre vendu, le 30/09/2021.

L'explication la plus plausible est un bug informatique dû au trop grand nombre de livre commandé fin septembre. mais peut être il y a eu un problème ailleurs, il faut se renseigner auprès du service de comptabilité voir si leurs chiffres sont différent, et ensuite auprès du service informatique voir ce qu'il s'est passé.

L'entreprise est envahie par les clients de 18 ans

Le nombre de client de 18 ans est incroyablement élevé.

437 au total alors que la médiane du nombre d'individu par année est de 136

L'âge requis pour s'inscrire sur notre site doit être 18 ans. Du coup ce chiffre doit venir du fait que des clients ayant moins de 18 ans mentent sur leur âge pour s'inscrire sur notre site.

Alors soit on trouve un moyen pour empêcher les mineurs de s'inscrire

-Message d'avertissement

-Vérification carte d'identité, carte bleu, photo etc…

Ou plutôt autoriser les mineurs à s'inscrire, ils auraient par exemple simplement une case à cocher pour dire qu'ils ont l'autorisation de leurs parents.

Et cela nous permettrai d'avoir des données plus fiables

1 produit sur 2 ne rapporte rien

50 % des livres de la boutique en ligne ne rapportent que 4 % du chiffre d'affaire. Le coefficient de Gini est très élevé 0.74.

Il faut essayer de vendre ces livres soit en les suggérant directement au client.

Soit en créant un onglet par exemple :« livre méconnu qui pourrait vous plaire » qui établirai pour chaque client une liste « adapté » composé de ces livres .

L’entreprise doit fidéliser plus encore

Cette fois l’indice de Gini est de 0.44 il est donc largement plus faible que pour la courbe précédente en effet les gains par clients sont plus équilibrés.

La plupart de nos client sont donc déjà fidèle à notre boutique.

Néanmoins, 10% des clients ne rapportent que 1.3% du chiffre d’affaire de la boutique. Il faut encore améliorer tout ce qui touche à la fidélité du client, carte de fidélité, abonnements, etc..(avantage premium, créer un contact avec le client via des messages durant la session d’achat..)

Anomalie détectée

Les livres des catégories 0 et 2 semble n’attirer que des clients d’une certaine tranche d’âge. 30 à 50 ans pour la catégorie 0.

18 à 30 ans pour la catégorie 2.

Ce graphique révèle un comportement suspect.

Pourquoi un client de 31 ans n’achèterai pas de livre de catégorie 2 alors que les clients de 30 n’achètent quasiment que ça.

Ou pourquoi un client de 30 ans n’achèterai pas un livre de catégorie 0 alors que les clients de 31 ans achètent majoritairement les livres de cette catégorie.

Peut être est-ce à cause de l’algorithme de suggestion.

Il tiendrait compte de façon trop importante de la catégorie d’âge du client.

Dans ce cas il faudrait l’optimiser, pour qu’il agisse de façon plus continue, c'est-à-dire avec des catégorie d’âge plus petite ou même pour chaque âge, afin que nos clients achètent d’avantage.

Point à part :

Un nombre plus grand de catégorie devrait permettre à l’algorithme de faire des suggestions plus personnalisées à nos clients

Conclusion

Il faut revoir certains points de la boutique en ligne :

- Comprendre la cause de l'effondrement des ventes en octobre
- Autoriser les mineurs à s'inscrire sur notre site
- Mettre en valeur les livres moins vendus
- Fidéliser encore plus le client
- Perfectionner l'algorithme de suggestion

Malgré tous ces problèmes, il n'y a pas d'inquiétude à avoir le chiffre d'affaire augmente, il a atteint son maximum le 15/02/2022.

Analyse des corrélations

Dans toute l'étude qui va suivre j'ai filtré les 4 gros client ainsi que les achats entre le 02 et le 27/10/2021

Sexe Catégorie de produit

J'ai effectué un test du chi-2 car j'ai deux variables qualitatives et j'ai ensuite calculé le coefficient v de cramer.

Les variables sexe et catégorie de produit acheté ne sont pas indépendantes car la p-value est extrêmement faible.

Mais V = 0.016 donc la corrélation est extrêmement faible

Age Catégorie de produit

Je fais une ANOVA pour voir l'effet des catégorie sur l'âge du client

La p-value est très faible donc il y a une corrélation.

la catégorie de produit achetée à un effet moyen sur l'âge des clients, $eta^2 = 0.12$, mais il est plus faible que ce à quoi je m'attendais.

Au vue du graphique je vais m'intéresser à définir des catégorie d'âge

Analyse factorielle des correspondances

J'effectue donc une Analyse factorielle des correspondances pour déterminer si il y a bien des groupes et comment ils se composent.

J'observe très facilement trois groupes les 18-30 ans, les 30-50 ans et les 50-93 ans.

J'ai ensuite fait un test du chi-2 entre categ_age et categ et j'obtient une corrélation forte entre les 2

V = 0.48 (mais je suis dans la partie variable quali et quanti donc revenons à nos moutons)

Part de dépense par catégorie Catégorie d'âge

Pour faire ce graphique j'ai d'abord calculé pour chaque client le pourcentage d'argent qu'il dépense dans chaque catégorie. Puis j'ai regroupé les client par âge, j'ai utilisé la moyenne pour les regrouper (mes données sont assez bien centré les moyennes et les médianes sont très proche)

Ensuite j'ai fait une ANOVA pour calculer l'effet des catégorie d'âge sur la part de dépense dans chacune des catégorie

La catégorie d'âge explique presque totalement la variance des proportions d'achat de chacune des catégorie.

Categ 0 : Eta^2 = 0.74 / Categ 1 : Eta^2 = 0.61 / Categ 2 : Eta^2 = 0.82

Variables quantitatives

Études des corrélations

J'ai d'abord réalisé des tests de corrélation de pearson sur les clients, entre leur âge et trois variable,

Panier moyen(nombre d'article acheté par session),

Fréquence(nombre de session conclu par un achat en moyen sur un mois),

montant total(nombre d'argent dépensé dans notre boutique sur toute la période)

les p-value sont très faible (<0.01) il y a donc dépendances pour ces 3 relations,

mais les corrélation linéaire sont faible.

Une corrélation anti linéaire faible pour le montant total et le panier et une corrélation linéaire faible avec la fréquence.

Pour plus de lisibilité j'ai regroupé les clients par âge comme précédemment cela m'a permis de tracer les courbes des trois variables en fonction de l'âge. (C'est moins précis que les nuages de point mais c'est bien plus lisible)

Paniers moyen en fonction de l'âge

Stable entre 18 et 30 puis ca monte c'est a nouveau stable entre 30 et 50 puis ça descend et ca se stabilise à nouveau.

On peut voir trois plateaux.

Montant total en fonction de l'âge idem

Fréquence en fonction de l'âge idem

Il semble que les catégories d'âge ont un effet énorme sur chacune des courbes

Test par catégorie d'âge

Du coup j'ai décider de refaire les test de corrélation de Pearson à l'intérieur des groupes d'âge.

Malheureusement mes p-value sont trop élevé je n'ai rien de significatif.

On n'a donc pas de raison de douter de l'hypothèse nul. Les trois variables sont indépendantes de l'âges à l'intérieur des groupes.

De plus aucune tendance linéaire même non significative ne semble se dessiner.

Il n'y a visiblement pas de corrélation entre les trois variables et l'âge dans les groupes.

Montant total Catégorie d'âge $Eta^2 = 0.09$

Panier moyen Catégorie d'âge $Eta^2 = 0.38$

Fréquence Catégorie d'âge $Eta^2 = 0.16$

Du coup je regarde la corrélation avec les groupes.

Je réalise maintenant trois test ANOVA pour calculer l'effet des catégorie d'âge sur ces trois variables.

La catégorie d'âge a un effet très important sur le panier moyen, important sur la fréquence et moyen sur le montant total

Pour aller plus loin..

Je vais revenir sur des variables quantitatives

Je commence par affiché ma matrice des corrélation en ajoutant les proportion d'achat pour chaque catégorie.

Elles sont fortement corrélé à l'âge. Et a priori elles sont influencé par l'algorithme

J'ai donc un algorithme qui influence les client selon leur âge

Corrélations partielles

Du coup J'ai décidé de calculer la corrélation avec l'âge de mes trois variable sans tenir compte des proportion d'achat de chaque catégorie pour effacer l'influence de l'algorithme

En supprimant l'effet des proportions d'achat j'ai

une corrélation anti linéaire partiel entre montant total et âge très faible

Une corrélation anti linéaire partielle moyenne entre panier moyen et âge

Et j'ai a priori indépendance entre fréquence et âge, la p valu est de 0,5

Bilan

Ce projet m'a permis d'approfondir mes connaissances en statistique et m'a fait découvrir

les nombreuses possibilités qu'offre le logiciel R en terme de représentation graphique et de

calcul de test statistique .