[SLIDES 1 à 4]

Ces notes présentent le compte-rendu d'activité des deux premiers exercices comptables du site de vente en ligne de la librairie Lapage. Après avoir brièvement présenté quelques hypothèses de départ, on s'attachera à passer en revue les principaux résultats de l'étude, avant de détailler la méthodologie en présentant les actions menées pour rendre le jeu de données utilisable, et d'analyser quelques indicateurs de vente au travers du chiffre d'affaires, de l'offre produits et des profils clients. On caractérisera enfin plus spécifiquement les comportements d'achat des clients au travers de l'analyse de la relation entre le genre et l'âge et plusieurs autres variables qualitatives et quantitatives. On conclura avec quelques recommandations pour améliorer la qualité des données et la pertinence de l'analyse.

[SLIDE 5]

Tout d'abord, quelques considérations méthodologiques générales. Au cours de cette présentation, plusieurs références à des moyennes nationales seront faites ; il s'agit des statistiques tirées du baromètre IPSOS / CNL sur les Français et la lecture en 2021 disponible ici :

https://centrenationaldulivre.fr/donnees-cles/les-francais-et-la-lecture-en-2021

Lorsqu'il sera fait mention de tests statistiques, et sauf mention contraire, il s'agira de tests non-paramétriques, ni la normalité de la distribution des variables étudiées ni leur homoscédasticité n'ayant pu être établies (voir les annexes de la présentation pour le détail des tests de normalité).

[SLIDE 6]

Quelques chiffres tout d'abord. Nos résultats se sont montrés solides malgré une conjoncture défavorable, avec une hausse globale du chiffre d'affaires de 3.26% entre les 2 exercices et une hausse de 1.41% du nombre de livres vendus, quand les moyennes nationales s'établissent à -2.7% et -3.4% respectivement. Les achats annuels moyens de nos clients augmentent de plus de 21 euros quand la moyenne nationale n'est que de +12 euros, et le nombre de nos clients reste relativement stable à -0.88%. Le panier moyen est en hausse chez les clients particuliers comme chez les professionnels.

[SLIDE 7]

Intéressons-nous à présent au nettoyage des données. Une fois importés, le contenu des fichiers .csv a été inspecté pour s'assurer que les données étaient correctement typées et les tables (ou dataframes) et colonnes étaient nommées explicitement par rapport à leur contenu et provenance. La colonne date de la table des transactions (renommée sales) renvoie un message d'erreur quand on essaye de la caster en format datetime.

[SLIDE 8]

Après un examen plus approfondi et l'élimination de 126 lignes de doublons de la table sales (qui étaient en fait des données de test qui n'ont rien à faire dans une base de prod), le casting de cette variable en format datetime fonctionne. Pour éviter de polluer les analyses avec des données fictives, on retire également les clients test de la base client et le produit test de la base produit. L'analyse des 2 autres tables (clients et books) ne fait pas apparaître de doublons.

[SLIDE 9]

Une autre information dupliquée apparaît au niveau des références produit dans la table books, dont le premier caractère est égal au contenu de la colonne catégorie. On a vérifié que c'est toujours le cas sans exception, et comme ça n'est pas de nature à nous poser problème dans l'analyse, on a laissé les références produit inchangées mais en pratique, ce format de stockage des données ne respecte pas la première forme normale car la référence produit n'est pas atomique et elle est en outre dupliquée avec la colonne catégorie ; il faudrait donc envisager d'utiliser une autre référence produit, comme le numéro ISBN par exemple, pour éliminer ces doublons dans les données stockées.

[SLIDE 10]

L'utilisation de la librairie missingno de Python ne fait apparaître aucune valeur manquante dans les 3 tables, i.e. il n'y a pas de NaNs ou autres « trous » dans les données. Le croisement des identifiants uniques des produits et clients présents dans les différentes tables révèle 21 clients n'ayant jamais effectué d'achats en ligne, 21 produits n'ayant jamais été vendus, et un produit vendu non-référencé dans la table books : on a donc imputé à ce produit (0_2245) le prix médian des produits de sa catégorie, qui est indiquée comme on l'a dit par le premier caractère de sa référence produit (ici 0). La recherche d'outliers dans la table sales_trim (dont on a retiré les données de test) fera par ailleurs apparaître des données manquantes en octobre 2021, on y reviendra.

[SLIDE 11]

On s'intéresse à présent à la recherche d'outliers dans nos 3 tables. Le boxplot des années de naissance des clients ne fait pas apparaître de valeurs atypiques ou aberrantes.

[SLIDE 12]

Si on croise les tables sales_trim et clients, la somme des ventes par client fait apparaître 4 grands clients que nous considèrerons comme des professionnels pour la suite de l'analyse. Le boxplot de droite est un « zoom » permettant d'observer le reste de la distribution des ventes individuelles par client de façon plus précise.

[SLIDE 13]

Dans la table books:

- 36 livres ont des prix supérieurs aux moustaches hautes de leurs catégories respectives : 22 livres de catégorie 0, 10 de catégorie 1 et 4 de catégorie 2 ont des prix supérieurs à la moustache haute de leurs catégories respectives ;
- En l'absence d'informations détaillées sur ces produits, il n'est pas possible de déterminer si leurs prix sont aberrants ou juste atypiques, ils ont donc tous été conservés dans la table books pour l'analyse ;
- On note l'existence de 3 gammes de prix très différentes pour ces 3 catégories de produits.

[SLIDE 14]

Dans la table sales, on note un « décrochement » du nombre de ventes en octobre 2021 et 25 jours qui ont un nombre de sessions quotidiennes inférieures à la moustache basse.

[SLIDE 15]

Ces 25 jours se trouvent tous correspondre aux 25 jours d'octobre 2021, presque tous consécutifs, pour lesquels aucune vente de produits de catégorie 1 n'a été enregistrée – c'est ce que montre la table de droite. Il semblerait dont que nous ayons un problème d'enregistrement ou de restitution des données sur notre site web. Compte tenu de la brévité de l'historique de données, aucune imputation n'a été effectuée pour combler les données manquantes.

[SLIDE 16]

En ce qui concerne les clés primaires de nos tables :

- L'identifiant client (client_id_clients) est une clé primaire pour la table clients ;
- La référence produit (product_id_books) est une clé primaire pour la table books ;
- Aucune des colonnes de la table sales ne peut être considérée individuellement comme une clé candidate, sauf date_sales qui est l'horodatage de la transaction à la nanoseconde près (ce qui a peu d'utilité pour notre analyse) :
    - Les colonnes product_id_sales et client_id_sales sont des clés étrangères pour la table sales_trim vers les tables books et clients respectivement et on utilisera une combinaison de ces clés pour l'analyse (avec le numéro de session).

[SLIDE 17]

Enfin on ajoute quelques données utiles, comme le calcul de l'âge des clients, la discrétisation de l'âge en tranches (choisies d'après l'étude IPSOS / CNL mentionnée précédemment) dans la table clients, et des colonnes permettant d'identifier l'exercice comptable, le jour de la semaine, le trimestre et l'année calendaire dans la table sales et qui serviront de support à plusieurs de nos analyses.

[SLIDES 18 et 19]

Intéressons-nous à présent à quelques indicateurs de vente. En premier lieu le chiffre d'affaires total quotidien, représenté sur ce graphique avec ses moyennes glissantes hebdomadaire, mensuelle, trimestrielle et annuelle. Vous trouverez en annexe 4 une analyse similaire mais par catégorie de produits cette fois.

- Ni le chiffre d'affaires quotidien, ni aucune des moyennes glissantes calculées ne semble mettre en évidence de tendances saisonnières ou cycliques dans les ventes ;
- On fait la même observation pour le chiffre d'affaires par catégorie en Annexe 4.

[SLIDE 20]

Quelques données de synthèse à présent sur notre chiffre d'affaires et notre volume de ventes. Le chiffre d'affaires total de nos 2 premiers exercices s'élève à un peu plus de 11,850,000 euros, avec un taux de croissance comme on l'a dit de 3.26% entre nos 2 premiers exercices comptables, soit 3.35% d'augmentation chez les particuliers et 2.22% chez les professionnels. Ce chiffre d'affaires a été réalisé sur 679,332 ventes de produits au total, représentant une évolution en volume de 1.41% entre l'année 1 et l'année 2.

[SLIDE 21]

Si on regarde la répartition de ces ventes entre les différentes catégories de produits, on s'aperçoit que les clients professionnels, comme les particuliers, achètent majoritairement des livres de catégorie 0.

[SLIDE 22]

La moyenne quotidienne s'établit à 942 livres vendus, 60.4% de catégorie 0, 34.3% de catégorie 1 et 5.3% de catégorie 2.

[SLIDE 23]

Si on regarde à présent le chiffre d'affaires quotidien de chaque catégorie de livres, on observe :

- Un chute des ventes due à l'absence de ventes de produits de catégorie 1 en octobre 2021 ;

- Et que les ventes de produits de catégories 0 et 2 semblent négativement corrélées d'après la lecture graphique, ce que l'on tentera de vérifier mathématiquement plus tard.

[SLIDES 24 et 25]

Caractérisons à présent notre offre produits :

- Les contributions respectives des catégories des produits au chiffre d'affaires sont très inégales :
  - Les produits de catégorie 0 représentent plus de 70% de notre catalogue mais à peine plus de 37% du chiffre d'affaires ;
  - Les produits de catégorie 2 représentent moins de 8% de notre catalogue mais plus de 23% du chiffre d'affaires.

[SLIDE 26]

- L'indice de Gini pour les produits est de 0.7425, ce qui confirme la grande inégalité de contribution des produits au chiffre d'affaires :
  - Les 78.50% de nos produits les moins vendus contribuent seulement 20% du chiffre d'affaires ;
  - Le CA médial atteint avec 92.9% des produits, en d'autres termes les 8% des produits les plus vendus contribuent plus de 50% du chiffre d'affaires ;
  - 98.34% des produits contribuent 80% du chiffre d'affaires.

[SLIDE 27]

Caractérisons à présent la relation entre les ventes de produits de catégorie 0 et de catégorie 2.

Comme on l'a dit, la distribution du nombre de ventes quotidiennes n'est pas Normale.

- on utilise donc un test non-paramétrique pour évaluer la corrélation entre ces 2 variables quantitatives: on calcule le rho de Spearman, qui est un coefficient de corrélation sur les rangs d'échelles ordinales (c'est-à-dire non-linéaire, contrairement au coefficient de Pearson qui est son équivalent paramétrique). Le rho de Spearman mesure la relation monotone entre deux variables i.e. le fait que lorsqu'une variable augmente, l'autre diminue toujours (relation monotone décroissante) ou augmente toujours (relation monotone croissante). Ce coefficient est proche de zéro si la relation n'est pas monotone, i.e. si lorsqu'une variable augmente, l'autre augmente parfois et diminue parfois ; un coefficient de Spearman proche de zéro ne signifie pas une absence de relation entre les variables, juste que cette relation n'est pas monotone (cf. relations parfaitement quadratiques comme $y = x^2$) ;

- Le rho de Spearman est (tout comme le coefficient de Pearson) toujours compris entre -1 et 1 et s'interprète de la même façon : plus le résultat est proche de 1 et plus les variables sont corrélées (resp. -1 / inversement corrélées) ; ici, rho vaut -0.36 et on peut donc conclure à l'existence d'une corrélation négative d'intensité faible entre les ventes de produits de classe 0 et 2, caractérisée par une relation monotone décroissante ;
- La p-value est indiquée comme étant zéro, ce qui signifie que ce résultat est significatif. Elle correspond à la p-value du test H_0: il n'existe pas de corrélation monotone entre ces variables contre l'alternative H_1: il existe une corrélation monotone entre ces variables. Au seuil de risque $\alpha = 1\%$, on rejette donc H_0, et on considère le coefficient de Spearman comme significatif de l'existence d'une corrélation monotone décroissante faible entre les ventes de ces 2 catégories de livres.

[SLIDE 28]

Regardons à présent le top 10 des ventes : en volume, il est dominé par les produits de catégorie 1, et en valeur par les produits de catégorie 2.

[SLIDE 29]

Si on regarde à présent les pires ventes (ce que j'ai appelé le « flop 10 »), les pires ventes en volume et en valeur sont dominées par les produits de catégorie zéro, donnée qu'il faut également rapprocher des 21 produits jamais vendus, dont 16 sont aussi de catégorie zéro. En résumé, si les produits de catégorie zéro représentent la majorité de notre catalogue, beaucoup d'entre eux en fait ne se vendent pas, et s'ils dominent les ventes en nombre chez les particuliers et les professionnels, en fait ces ventes sont concentrées sur un nombre de références relativement restreint. Les produits de catégorie zéro contribuent donc peu au chiffre d'affaires par rapport à leur nombre de références au catalogue (+ de 70% des références comme on l'a dit) non seulement parce que leur prix moyen est bien plus faible que ceux des autres catégories, mais aussi et surtout parce que beaucoup d'entre eux se vendent très peu.

Au total, près de 33% des produits de notre catalogue se vendent très peu, i.e. moins d'une unité par mois.

[SLIDE 30]

L'analyse des associations a révélé que 9 des produits du top 10 en volume sont souvent associés à 23 autres produits dans les paniers de nos clients, ce que nous pourrons utiliser pour faire des recommandations en phase de pré-validation du panier. Un de ces produits (le 1_403) ne figure pas dans les résultats car il n'a pas passé l'étape de "prunage : de l'algorithme d'analyse des associations, étant vendu seul dans plus de 99.8% des cas.

[SLIDES 31 et 32]

Intéressons-nous à présent aux profils de nos clients : qui sont-ils ? Ils sont composés de 52.1% de femmes et 47.9% d'hommes, ce qui est un peu supérieur à la moyenne nationale du dernier recensement de la population 2019 publié par l'INSEE en 2022 qui faisait état de 51.6% de femmes, donc nous avons cherché à savoir via un test statistique si cette différence était significative. Ici, être une femme est le résultat d'une épreuve de Bernoulli, et une succession de variables de Bernoulli indépendantes suit une loi Binomiale de paramètres (p, 1-p) et on sait que d'après le théorème centrale limite X barre, qui est un bon estimateur de p, converge en loi vers une loi normale de paramètres mu et racine de p fois 1-p divisé par n, si n est suffisamment grand (ici n vaut 8600 ce qui est le cas), donc le test utilisé ici est paramétrique (en application du théorème de la limite centrale) et fait partie des exceptions que j'ai mentionnées en introduction. On pose H_0 : les proportions de femmes sont égales dans nos clients et dans la population française (i.e. la probabilité que nos clients soient des femmes est de 51.6%) et H_1 : la probabilité que nos clients soient des femmes est > à 51.6%. On pose aussi $\alpha = 5\%$

- On trouve statistique de test $Z = 0.8961$ et p-value = 0.1851

p-value > α, donc au seuil de risque donné, nous n'avons pas suffisamment d'informations pour rejeter H_0 et nous devons accepter que la différence entre les proportions de femmes dans nos clients et dans la population française en général est due au hasard de l'échantillonnage

[SLIDE 33]

En termes de chiffre d'affaires, les femmes représentent 49.4% tous clients confondus et 52% une fois les clients professionnels exclus, et leur panier moyen s'élève à 18.95 €, soit 0.5% de moins que les hommes.

[SLIDE 34]

La pyramide des âges révèle un très grand nombre de clients âgés de 19 ans. La classe d'âge la plus représentée parmi nos clients est celle des 35-49 ans avec plus de 30%, suivie des 50-64 ans avec un peu plus de 23%.

[SLIDE 35]

Si on observe le chiffre d'affaires réalisé avec le top 10 de nos clients, on retrouve les 4 clients professionnels mentionnés plus tôt. En incluant ces clients, on obtient un indice de Gini de 0.4463…

[SLIDE 36]

… et 0.4025 en les excluant. Notre chiffre d'affaires est donc réalisé de façon relativement inégale et concentrée sur certains clients. Il faut plus de 48% de nos clients les moins acheteurs pour attendre 20% de notre chiffre d'affaires, et le chiffre d'affaires médial est atteint avec plus de 76%

des clients, ce qui est une autre façon de dire que plus de la moitié du chiffre d'affaires est réalisé avec les 25% des plus gros clients particuliers.

[SLIDE 37]

Si on compare à présent la représentation des classes d'âge parmi les clients acheteurs et non-acheteurs, on constate que

- Les 15-24 ans sont la classe d'âge majoritaire parmi les clients non-acheteurs ; et que
- La classe des 35-49 ans n'est pas représentée parmi les clients non-acheteurs.

[SLIDE 38]

Compte-tenu de la brévité de l'historique (rappelons que nous n'avons que 2 exercices comptables ici) et de la nature des produits vendus (des livres, qui sont des biens culturels et d'acquisition, par opposition aux biens de consommation), il ne nous a pas semblé pertinent de procéder à une analyse RFM en bonne et due forme, mais nous avons cependant segmenté nos clients en catégories en fonction du nombre de livres achetés par an :

- Petits lecteurs : < 5 livres / an ;
- Lecteurs moyens : entre 5 et 20 ;
- Grands lecteurs : entre 21 et 52 ;
- Très grands lecteurs : entre 53 et 104 ;
- Collectionneurs : entre 105 et 999 ;
- Clients professionnels : 1000 et plus.

Le résultat de cette segmentation révèle que :

- Plus 60% de nos clients sont des grands ou très grands lecteurs, ou des collectionneurs (plus de 52 livres / an) ;
- Près de 40% de nos clients sont composés de petits ou de moyens lecteurs
    - nos actions marketing devraient probablement se concentrer sur ces derniers

[SLIDES 39 et 40]

Nous allons à présent nous intéresser aux comportements d'achat de nos clients, à commencer par vérifier l'hypothèse selon laquelle le genre des clients n'impacte pas les catégories de livres achetées. Ces 2 variables étant qualitatives, un test du khi-2 de contingence est approprié. Il s'agit ici encore d'un test paramétrique, dont la statistique de test, qui est les écarts réduits entre les distributions théorique et empirique de nos 2 variables qualitatives, suit une loi du khi-2 : On admet ici que, n étant suffisamment grand (ici n vaut 679,332), l'espérance de chacune des 2 variables

converge en loi vers une loi Normale et le carré de leurs différences (mesure par les écarts réduits) converge donc vers une loi du khi-2. On pose H_0 : les variables sont t indépendantes, i.e. la distribution des observations est identique pour les 3 catégories de livres aux fréquences théoriques attendues, i.e. le genre du client n'affecte pas les catégories de livres achetées ; et H_1 : les variables ne sont pas indépendantes, i.e. la proportion de livres achetés dans chaque catégorie dépend du genre du client et les fréquences observées différent significativement des fréquences théoriques attendues.

On pose également $\alpha = 5\%$

La valeur du chi2 calcule est : 147.0 – il « sort de la table » par la droite

La p-value est $1.2*10^{-32}$

Nombre de degrés de liberté : 2

La valeur du khi-2 théorique est : 0.1026

On rejette donc H_0. Les variables ne sont pas indépendantes

On utilise le V de Cramer pour quantifier cette non-indépendance : il vaut 0.0147

> Le V de Cramer étant très proche de 0, on en conclut que, bien que les différences entre les catégories de livres achetées par les hommes et les femmes soient statistiquement significatives, la dépendance entre ces deux variables est en fait extrêmement faible et les livres de catégorie 2 contribuent le plus à cette très faible non-indépendance, comme le montre la heatmap.

[SLIDE 41]

Nous allons à présent tenter d'appréhender l'influence de l'âge sur plusieurs variables quantitatives…

[SLIDE 42]

… à commencer par le montant total des achats.

- Les 15-24 ans et les 25-34 ans dépensent le plus par achat (panier moyen) mais ils réalisent moins de transactions que les 35-49 ans et les 50-64 ans, probablement car leurs effectifs totaux dans notre pool de clients sont moindres que ces 2 classes ;
- Il pourrait donc être pertinent d'organiser des campagnes marketing ciblant les 2 classes de moins de 35 ans pour augmenter leur proportion dans le total de nos clients.

[SLIDE 43]

- L’âge des clients n'étant pas distribué selon une loi normale, ici encore on utilise le rho de Spearman pour évaluer la corrélation entre ces 2 variables, à savoir l’âge du client et ses dépenses totales. Ici rho vaut -0.87 et on peut donc conclure à l'existence d'une corrélation négative d’intensité très forte entre l’âge des clients et le montant total des achats, caractérisée par une relation monotone décroissante.
- La p-value est indiquée comme étant zéro, ici encore ce résultat est significatif

[SLIDES 44 et 45]

Voyons à présent l’influence de l’âge sur la fréquence d’achat. Il semblerait que toutes les classes d'âge n'aient pas les mêmes habitudes d'achat en termes de fréquence

[SLIDE 46]

- En particulier, il semble que le nombre moyen de livres lus par an n'est pas identique parmi nos 5 classes d’âge. Vérifions cette hypothèse par un test statistique. Comme l’âge des clients n'étant pas distribué selon une loi normale et que l’on a plus de 2 populations, ni l’utilisation du T-test de comparaison de populations (ou test de Student) ni ANOVA ne sont appropriés ici, et nous utiliserons le test de non-paramétrique de Kruskal-Wallis, qui est une généralisation du test de Mann-Whitney à plus de 2 populations (les populations étant ici nos groupes de lecteurs dans chaque classe d’âge).
- On pose H_0 : le nombre moyen de livres lus par an est le même pour toutes les classes d’âge, que nous testons contre l'hypothèse alternative H_1 : le nombre moyen de livres d'au moins une classe d’âge est diffèrent de celui des 2 autres, à un seuil $\alpha = 5\%$.
- La statistique de test vaut 1834.95
- La p-value vaut 0.000000e+00
- On rejette donc l'hypothèse nulle, au moins une classe d’âge a une moyenne du nombre de livres lus par an différente des 4 autres.
- Le test de Dunn donnant des p-values pour toutes les différences entre les classes d’âge très proches de zéro, nous pouvons en conclure que toutes nos classes d’âge ont des moyennes du nombre de livres lus par an différentes au seuil $\alpha = 5\%$.

[SLIDES 47 et 48]

Il semblerait également que les classes de lecteurs n’aient pas le même âge moyen. Vérifions cette hypothèse par un test statistique. Ici encore on utilise le test de Kruskal-Wallis comparer nos populations pour les mêmes raisons que citées précédemment.

On pose H_0 : l'âge moyen est le même pour tous les reader tiers, que nous testons contre l'hypothèse alternative H_1 : l'âge moyen d'au moins un reader tier est différent de celui des 2 autres, avec un seuil α = 5%.

- La statistique de test vaut 211.11
- La p-value vaut : $1.53*10^{e-44}$

On rejette donc l'hypothèse nulle, au moins un reader tier a un âge moyen d'achat différent des 4 autres.

Le test de Dunn donnant des p-values pour toutes les différences entre les reader tiers très proches de zéro sauf entre les catégories 1 et 3, nous pouvons en conclure que tous nos reader tiers ont des moyennes différentes au seuil α = 5%, sauf les grands et les très grands lecteurs, que l'on pourra considérer comme homogènes du point de vue de l'âge moyen.

[SLIDES 49 et 50]

Vérifions à présent s'il existe une relation entre l'âge et la taille du panier moyen mesurée par le nombre de livres.

- Coefficient de corrélation de Spearman : -0.23
- p-value : 0.00

Le coefficient de Spearman reste relativement proche de zéro. La p-value est significative, donc nous pouvons conclure qu'il existe une relation monotone décroissante faible entre l'âge des clients et le nombre de livres dans leur panier

[SLIDE 51]

Si l'on s'intéresse à présent au montant du panier moyen, il semble décroître avec la classe d'âge. Le Rho de Spearman vaut -0.34 et la p-value est significative, donc on peut ici conclure à une relation monotone décroissante faible entre l'âge d'un client et le montant de son panier moyen.

[SLIDE 52]

Si l'on s'intéresse à présent au nombre total d'achats en fonction de l'âge, et toujours en faisant appel au rho de Spearman, on trouve rho = -0.67 et une p-value significative, donc nous pouvons conclure à l'existence d'une relation monotone décroissante forte entre l'âge et le nombre total d'achats.

[SLIDES 53 et 54]

En ce qui concerne les catégories de livres achetées, il semble que l'âge moyen des lecteurs n'est pas identique parmi nos 3 catégories de produits. Ici encore on utilise le test de Kruskal-Wallis pour comparer nos populations ; bien que l'on ait une variable quantitative et une variable qualitative, ANOVA ne s'applique pas en raison de l'absence de normalité des données.

- On teste :
    - H_0 : l'âge moyen est le même pour toutes les catégories de livres
    - H_1 : l'âge moyen d'au moins une catégorie de livres est différent de celui des 2 autres
    - $\alpha = 5\%$

  On trouve :

  - Statistique de test : 72214.83433330593
  - p-value : 0.000000e+00

On rejette donc l'hypothèse nulle, au moins une catégorie de livres a un âge moyen d'achat différent des 2 autres

[SLIDES 55 et 56]

Pour finir, mentionnons rapidement quelques autres indicateurs.

- En premier lieu, le temps médian passé en ligne, qui est relativement proche pour toutes les classes d'âges de moins de 50 ans et s'établit autour de 1,500 secondes (environ 25 minutes) ;
- Ce résultat est biaisé par le fait que notre système de vente en ligne ne permet de calculer une durée de session que pour les paniers comportant plus d'un article ;
- Il conviendrait de changer les modalités d'horodatage de notre magasin en ligne afin d'enregistrer la durée de toutes les sessions (même celles ne résultant que dans l'achat d'un seul article) pour pouvoir raffiner cette analyse ainsi que l'analyse sur le lien entre la durée des sessions et l'âge.

[SLIDE 57]

- Il ne semble pas exister de lien entre le jour de la semaine et le montant du panier ; les paniers moyens et médian sont extrêmement proches pour tous les jours de la semaine, ce qui peut sembler surprenant compte-tenu du fait que même les boutiques en ligne enregistrent des pics de trafic le weekend ;
- Ici encore, il conviendrait donc de tester en amont la fiabilité des données collectées par notre site de vente en ligne.

[SLIDE 58]

- Enfin les montants moyen et médian du panier semblent croître avec la durée des sessions ;

- Ici encore, l'analyse est faussée par le système d'horodatage qui ne prend en compte que les paniers multi- produits pour le calcul de la durée des sessions.

[SLIDES 59 et 60]

En conclusion, quelques recommandations pour améliorer la qualité des données et donc de nos analyses.

- Il y a des problèmes à corriger dans l'acquisition des données : ventes non enregistrées, produits absents du référentiel, données de test dans une base de prod, clients achetant de multiples exemplaires du même ouvrage (cf. analyse des associations), stockage des données non conforme à la 1NF ;

- Il faudrait enrichir l'information client pour affiner l'analyse : ajouter le niveau d'études, la CSP, les loisirs favoris hors lecture, la zone géographique, la fréquentation de nos magasins physiques pour n'en citer que quelques-uns ;

- Il faudrait mettre en place un suivi individualisé pour les clients professionnels ;

- Il faut enrichir l'historique : 2 exercices comptables ne suffisent pas pour permettre la mise en place d'une analyse RFM (à supposer qu'on juge une telle analyse pertinente pour un bien culturel comme le livre), ni pour évaluer le coût d'acquisition de nouveaux clients, le taux d'attrition et la durée de vie d'un client de façon significative ;

- Il faut revoir notre offre de produits : près de 33% de notre catalogue se vend peu, l'offre de catégorie 2 est à élargir (up-selling), l'offre de catégorie 0 est à repenser (peu de succès pour beaucoup de références), l'analyse des associations est à utiliser pour faire des recommandations (cross-selling) ; et enfin

- L'information produit est à enrichir : il faudrait analyser séparément les e-books et les livres papier, enrichir le référentiel produit avec des catégories détaillées par genre littéraire, l'existence d'une suite ou série, l'auteur, l'éditeur, la date de sortie, les prix littéraires par exemple, pour affiner l'analyse et améliorer les recommandations aux clients