Diapo 1 :

Je suis Naziha, data analyst au sein du service Marketing de l’entreprise Lapage qui initialement était une librairie physique avec différents points de ventes.

Depuis 2 ans Lapage a ouvert ses ventes en ligne et nous avons souhaité faire le point sur l’activité.

Diapo 2

* D’une part, nous avons souhaité analyser et mettre en lumière le chiffre d’affaires et
* D’autre part, nous avons souhaité mieux comprendre le comportement de nos clients en ligne

Diapo 3

Pour ce faire j’ai utilisé Visual studio code pour accueillir le notebook et les librairies de Python : Pandas et Numpy pour le nettoyage et l’analyse des données, SciPy pour les tests statistiques, et matplotlib, seaborn et plotly pour la visualisation des résultats.

Diapo 4

Ce projet d’analyse se découpent en 4 parties :

* La première partie englobe l’Exploration et le nettoyage des données
* Les parties 2 et 3 ciblent ***nos objectifs business*** qui sont
  + d’une part l’analyse de notre chiffre d’affaires telle que l’évolution par année, le taux de croissance et le taux de variation
  + et d’autre part, l’analyse comportementale des consommateurs sur le site en ligne à travers l’étude statistiques de 5 corrélations que nous verrons en détails à la suite de cette présentation
* et nous finirons par une synthèse de nos résultats

**Diapo 5**

Le dataset se compose de 3 fichiers .csv : clients, produits, et transactions :

J’ai d’abord effectué un premier repérage visuel des valeurs manquantes avec le package missingno qui ne montre aucune valeur manquante dans les 3 BDD.

* Les bases de données clients et produits ne contiennent ni valeurs manquantes, ni doublons, les variables ont un format adapté.
* En revanche, la base de données des Transactions dont la dimension est de : 687534 lignes et 4 variables comporte des ambiguïtés telles que le format de la date en objet que nous avons convertie en datetime avec le module datetime de la bibliothèque Python et de l’horaire «24H» convertie en 00h, la variable id\_product comporte 3265 valeurs uniques, et 587516 doublons. Après vérification, il s’agissait de faux doublons : 1 produit pouvant être acheté une ou plusieurs fois, la suppression ne s’imposait pas.
* La fusion des fichiers s’est réalisée via une jointure interne par les clés étrangères client\_id et id\_prod dans la table d’association « Transactions\_df »
* Les Observations relevées à la suite de cette fusion :
  + Les prix ne contiennent aucune valeurs négatives
  + 2022 est une année complète
  + 2021 et 2023 sont incomplètes, nous n’avons pas les données de janvier et février 2021 et nous n’avons que les mois de janvier et février 2023
  + Pour faciliter nos analyses : des colonnes supplémentaires ont été crées par l’extraction des années, mois, jours et âges des variables « date » et « birth »
  + Nous obtenons 4 dataframes :
    - Merge\_2023 conserve les données de 2021 à 2023
    - Merging\_final ne conserve que les années complètes et semie-complète 2021 et 2022
    - Remaining\_clients : exclus les clients B2B et ne conserve que la clientèle B2C
    - Top\_customers\_data : ne conserve que les 4 clients pro.

**Diapo 6**

Ce graphique présente l’évolution du chiffre d’affaires global sur nos ventes en lignes de mars 2021 à février 2023

**3 phases de croissances se dessinent grâce à la courbe des moyennes mobiles en rouge :**

* La période de mars à juillet 2021 : représente une phase de latence, une courbe douce qui se dessine discrètement juste un peu au-dessus de la barre des 480K.
* La période de août 2021 à février 2022 est une courbe exponentielle avec un taux d’augmentation de +11%
* Et une phase de décroissance de mars 2022 à février 2023 avec une baisse de 11,40%

En comparant le taux de croissance des périodes de août 2021 à février 2022 avec celle de 2022-2023, nous observons un taux de croissance de 11,05% pour la première période et PRESQUE 10% pour la seconde.

**Diapo 7**

Le taux de variation du CA entre 2021 et 2022 est globalement positif avec une augmentation de +23% du CA. Néanmoins, il nous manque les données de janvier et février 2021 et on observe des résultats négatifs en septembre, novembre et décembre : une période qui en 2021 était florissante.

*Comment se répartie notre chiffre d’affaires parmi les clients ?*

**Diapo 8**

La courbe de Lorenz nous montre une répartition inégale certaine du CA au sein de notre clientèle et cela est illustré par la courbe bleue qui est éloignée de la ligne d’égalité parfaite (bissectrice en pointillée).

44% de l’espace entre l’égalité parfaite et la courbe de Lorenz est observé par le coefficient de Gini qui suggère une inégalité modérée.

50% du CA est détenu par 78% des clients et 7% du CA est détenue par 1% des clients

*Qui sont les détenteurs de ces 7% de CA ?*

**Diapo 9**

Parmi les 8600 clients, 4 clients se détachent du reste de la clientèle. Il s’agit de professionnels.

**Diapo 10 : repartition globalement égalitaire par catégorie vs chez les B2C qui consomment moins la cat 2**

**Diapo 11 : Répartition du CA égale entre les femmes et les hommes chez les B2C vs B2B**

**Diapo 12 : répartition du CA très inégale, largement concentré chez la population âgé de 30 à 55 ans**

**Diapo 11**

Nous avons fait un classement des 10 meilleures et moins bonnes ventes.

TEST STATISTIQUES

Diapo 1 : j’ai d’abord testé la normalité de nos données via la méthode Kolmogorov-smirnov sur les variables prix et âges pour déterminer si nos données suivent une distribution gaussienne et pour aider à choisir les bonnes méthodes statistiques par la suite.

Diapo 2 :

La première corrélation que nous avons souhaité tester est celle de la catégorie de livre acheté avec le genre du client.

* Graphiquement nous ne pouvons pas confirmer ou infirmer l’hypothèse nulle qu’il n’existe pas de relation entre le sexe et la catégorie du produit, d’où le test du chi2
* Pour ce faire, j’ai utilisé le test statistiques du Chi2 sur 2 variables catégorielles qui évalue si les différences entre les fréquences observées dans un tableau de contingence et les fréquences qui seraient attendues sous l’hypothèse nulle sont suffisamment importantes pour rejeter cette hypothèse.

Diapo 3,4,5 : Pour l’analyse de corrélation entre

* l’âge et le montant total des achats
* l’âge et la fréquence d’achat
* l’âge et la taille du panier

j’ai utilisé le test de corrélation de Pearson pour évaluer la force et la direction de la relation linéaire entre 2 variables continues.

Le test de Pearson confirme que :

1. qu’il existe une relation linéaire inverse entre l’âge et le montant total des achats : plus un client est jeune, plus le total d’achats est élevé.
2. Qu’il existe une corrélation négative modérée entre l’âge et la fréquence d’achats. Cela suggère qu'à mesure que l'âge augmente, la fréquence d'achat a tendance à diminuer, et vice versa, bien que cela ne soit pas nécessairement une relation causale.
3. Qu’il existe une corrélation négative modérée entre l'âge des clients et la taille du panier moyen. Lorsque l'âge augmente, la taille du panier moyen tend à diminuer, mais cela ne signifie pas nécessairement une relation linéaire parfaite.

Diapo 6 : Et enfin pour répondre à la question : existe t -il une relation entre l’âge et la catégorie des livres achetés, j’ai utilisé le test ANOVA pour évaluer s’il existe des différences significatives entre les moyennes d’âges des différentes catégories. En d’autres termes ANOVA mesure la variation entre les moyennes des différents groupes par raport à la variation à l’interieur des groupes.

Dans notre cas, nous obtenons une p-valeur de 0 qui indique une forte preuve contre l’hypothèse nulle selon laquelle toutes les moyennes des groupes sont égales.

* Il y a dont des différences statistiques significatives entre les moyennes d’âges des différentes catégories testées. Les moyennes d’âges varient de manière significatives entre les différentes catégories