**DEADLINE : Mercredi 14 Novembre 2022**



Nous sommes dans une grande chaine de librairie concurrente d’Amazon. Vous n’êtes plus le nouveau arrivé dans l’entreprise et vous avez déjà fait connaissance avec vos collègues et l’ambiance est au rendez-vous.

Mais revenons à nos moutons: il est temps de mettre les mains dans le cambouis ! Le service Informatique vous a donné l’accès à la base de données des ventes. À vous de vous familiariser avec les données, et de les analyser. Votre manager souhaite que vous réalisiez une présentation pour un dossier important !!

Comme vous l'avez appris dans vos recherches et lors d’une précédente mission, l’entreprise, s'est d'abord développée dans une grande ville de France, avec plusieurs magasins, jusqu'à décider d'ouvrir une boutique en ligne. Son approche de la vente de livres en ligne, basée sur votre algorithme de classement google à base de chaines de Markov lui a valu un franc succès 😊 et toute l’équipe vous a félicité pour le travail accompli !

**Les Data**

Voici les fichiers à votre disposition extraites directement de la base de l’entreprise

* les ventes ;
* la liste des clients ;
* la liste des produits

**Vos missions**

#### Mission n° 1

Avant de pouvoir entrer dans le vif du sujet, il vous faudra faire un peu de nettoyage ! Par exemple, vous devrez faire des choix quant au traitement des valeurs manquantes et des valeurs aberrantes. L’analyse et l’identification des outliers sur les prix se fera en utilisant les 4 méthodes suivantes:

\*L’approche statistique : Méthode du z score (supérieur à 2 standard deviation) et méthode des interquartiles

\*La méthode graphique : nuage de points ( les outliers seront d’une autre couleur que le reste des points) et le boxplot

#### Mission n° 2

Ensuite, vous réaliserez l’analyse des données. Une grande liberté vous est laissée sur ce plan, mais à vous de trouver les informations qui ont du sens pour mieux comprendre les ventes.

Vous devrez y utiliser au moins :

* des indicateurs de tendance centrale et de dispersion ;
* une analyse de concentration, via une courbe de Lorenz et un indice de Gini ;
* des représentations graphiques, dont au moins un histogramme, une représentation avec des "boîtes à moustaches", et une représentation de série temporelle (c’est-à-dire un graphique dont l’axe des abscisses représente des dates) ;
* des analyses bivariées.

#### Mission n° 3

Voici quelques questions supplémentaires, que votre manager vous a posées :

1. Y a-t-il une corrélation entre le sexe des clients et les catégories de produits achetés ?
2. Y a-t-il une corrélation entre l'âge des clients et :  
   * Le montant total des achats ;
   * La fréquence d’achat (ie. nombre d'achats par mois par exemple) ;
   * La taille du panier moyen (en nombre d’articles) ;
   * Les catégories de produits achetés.

**MODALITES ET CRITERES DE PERFORMANCE**

-Le code sera exclusivement du Python ( **Jupyter notebook**) : notebook enrichi de commentaires, d’analyses , de conclusions et respectant PEP8.

- Le travail est **en groupe et le plagiat est sanctionné par une note de 0.**

-Le livrable final est un dossier zip. Il devra contenir

* le **fichier jupyternotebook**  sur les chaines de Markov
* le **fichier jupyternotebook** destiné à nettoyer le jeu de données et les différents analyses effectuées ( issues de la base de donnés de ventes du libraire )
* les **graphiques** dans un format image (PNG ou JPG) dans un dossier spécifié