PROJET VALIDÉ

Analysez les ventes de votre entreprise

70 heures

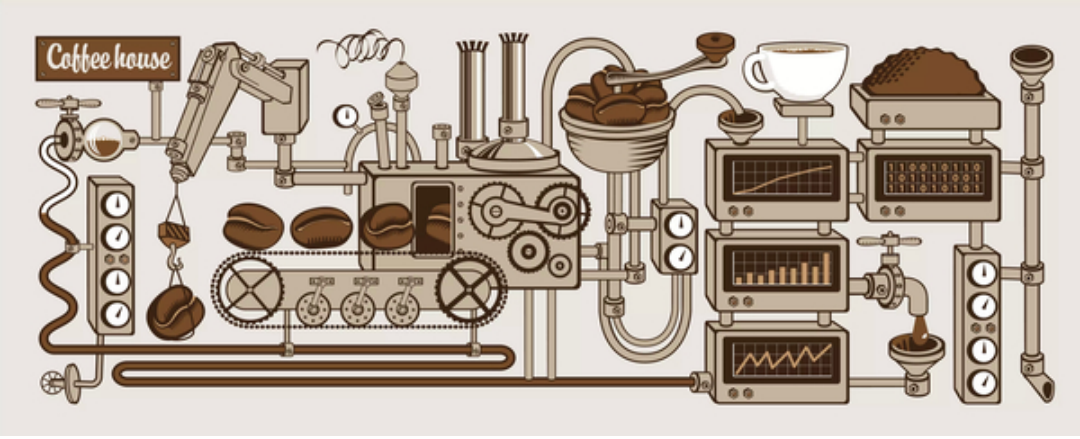
Mis à jour le mardi 9 novembre 2021

Prérequis

Les prérequis pour ce projet sont de connaître au choix les langages **R** ou **Python**, en sachant manipuler principalement les *Dataframes* (natifs en R, ou disponibles par la librairie Pandas en Python). Il vous faudra également connaître les bases de la statistique descriptive (moyenne, médiane, variance, représentations graphiques, tests de corrélation, analyse bivariée, etc.).

Scénario

Vous êtes data analyst d'une grande chaîne de librairie, fraîchement embauché depuis une semaine ! Vous avez fait connaissance avec vos collègues, votre nouveau bureau, mais surtout, la machine à café high-tech :



Rien que ça !

Mais revenons à votre mission : 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 vous "faire la main".

Comme vous l'avez appris dans vos recherches avant de postuler, votre entreprise, "Rester livres" 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 des algorithmes de recommandation, lui a valu un franc succès !

Les données

Vous avez accès à ces données, extraites directement de la base de l'entreprise vers les fichiers CSV. Voici les fichiers à votre disposition :

- les ventes (appelées "Transactions") ;
- la liste des clients ;
- la liste des produits.

Téléchargez le jeu de données à **cette adresse**.

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.

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.

Pour les corrélations, pas besoin d'effectuer en entier les tests (chi-2, ANOVA, etc.). Seul le calcul des statistiques de test est demandé (r^2, η^2, ξ_n).

Quelques précisions

- Vous avez le choix entre 2 langages : R ou Python. Dans les deux cas, vos donnés devront être manipulées via les structures Dataframe ou Matrice (présentes nativement sous R, ou dans la librairie Pandas sous Python).
- Pour plus de simplicité, nous considérons ici que les prix des articles ne varient pas en fonction du temps.

Livrables

Un fichier .zip contenant ces fichiers :

- le **script** destiné à nettoyer le jeu de données ;
- le **script** contenant les différentes analyses effectuées ;
- les **graphiques** dans un format image (PNG ou JPG) ;
- un court **fichier README**, contenant les explications pour lancer vos scripts.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé "*P4_nom_prenom*", tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "*P4_01_scriptdonnées*", "*P4_02_scriptanalyse*", et ainsi de suite.

Soutenance

Vous serez amené à créer des graphiques. Il y a certains pièges dans lesquels il ne faut pas tomber : veillez bien à respecter les règles données dans **ce chapitre**. Pensez également à votre examinateur : si votre présentation n'est pas donnée en plein écran, et si l'écran de votre examinateur est plus petit que le vôtre, il risque de ne pas pouvoir lire correctement les graphiques, les légendes ou les noms des axes.

La soutenance, d'une durée de 30 minutes, se déroule en 3 étapes :

1. Détail du nettoyage des données : quelles valeurs aberrantes et manquantes avez-vous trouvées, comment les avez-vous traitées, avez-vous effectué d'autres nettoyages ? (10 minutes)
2. Présentation de l'analyse demandée, à l'aide d'un logiciel de présentation adapté. (10 minutes)
3. Présentez votre analyse des corrélations, en les interprétant. (10 minutes)
4. Séance de questions-réponses éventuelle.

Compétences évaluées


Décrire un jeu de données par la statistique descriptive


Maîtriser les concepts statistiques fondamentaux


Nettoyer un jeu de données

OPENCLASSROOMS

OPPORTUNITÉS

AIDE

POUR LES ENTREPRISES

EN PLUS

Français

Télécharger dans l'App Store