Université Sultan Moulay Slimane
Ecole Supérieure de Technologie
Fkih Ben Salah

Mini-Projet

Module : Programmation Python

Filière Licence BIG DATA, Semestre : 5, AU : 2024-2025

Objectif principal :

L'objectif principal de ce mini projet, est de concevoir une approche dont le but principal est la réalisation d'un système d'aide à la décision basé sur le processus d'extraction de connaissances à partir de données afin d'avoir une vue globale sur l'ensemble de données, et pour prédire l'attrition (perte de clients) dans les contextes de e-commerce.

Ce mini projet a pour objectif principal de se familiariser avec les bibliothèques PANDAS et Numpy pour assurer l'étape de préparation, transformation, manipulation et l'analyse de données (rendre les données de qualité).

Pour atteindre les résultats souhaités, nous allons adopter le modèle de référence CRISP-DM (Cross-Industry Standard Process for Data mining).

Description de la base de données

L'entreprise de commerce en ligne de notre étude utilise les sources de données suivantes :

- **La base de données des clients :** Cette base de données contient les coordonnées, et les données démographiques sur les clients (nom, sexe, âge, adresse, ville, pays, profession, ..)
- **Les données sur les transactions :** Lorsque le client valide une commande, toutes les informations liées à cette action sont enregistrées (identifiant du client, référence de produit, date de transaction, montant, quantité achetée, le mode de paiement, adresse de livraison, ville de livraison ...).

- **Les données sur les sessions :** Toutes les sessions ouvertes par les clients sur le site e-commerce sont enregistrées mêmes qu'elles n'ont pas terminées avec une transaction. Ces informations comprennent l'identification du client, la date de connexion et de déconnexion, la durée passée lors d'une session, le statut de session (session avec

consultation de produits, session avec ajout au panier, session avec check-out et session avec transaction), l'adresse ip, le terminal utilisé, système d'exploitation utilisé, etc...

La société dispose aussi une base de données des produits qui contient des informations détaillées sur ses produits (référence produit, libellé, description, prix, quantité disponible, catégorie). Mais dans notre projet nous allons focaliser seulement sur les coordonnées de clients, les habitudes de navigations et les données sur les achats.

Travail N°1 :

Partie 1

Période d'étude : Du 1er novembre 2013 au 28 février 2015

On se basant sur les données de transactions effectuées par les clients sur le site e-commerce, et qui sont enregistrées dans les fichiers (les deux fichiers : dataset_P1(Nov2013-June2014).csv et dataset_P2(July2014-January2015) :

1. Faire les traitements nécessaires pour charger et fusionner les fichiers.
2. Pour chaque colonne donner le nombre de valeurs manquantes.
3. Afficher un data frame avec toutes les lignes qui ont au moins une valeur manquante.
4. Supprimer ensuite toutes les lignes avec des valeurs manquantes.
5. Vérifier les dates des transactions afin de garder que celles qui sont effectuées par les clients dans la période d'étude.
6. Afficher les statistiques descriptives de base.
7. Vérifier s'il y a des valeurs aberrantes dans la colonne *Price*, si oui comment vous pouvez expliquer l'existence de ces valeurs généralement ? quelle sera votre approche pour traiter ces valeurs ?
8. Quel est le mois qui a enregistré le plus grand nombre de ventes ? et quel est le montant total de ce mois ?
9. Afficher un graphe qui montre l'évolution du montant total par mois
10. Pour chaque client, Créer les variables suivantes :
    a. Fréquence : Nombre de transactions observées au cours de la période analysée.
    b. Récence : le nombre de jours entre le premier jour de la période d'étude et le jour du dernier achat.
    c. Longueur : nombre de jours entre le premier et le dernier achat (Longueur de la relation client/site e-commerce)
    d. Montant : le montant total dépensé par le client au cours de la période analysée. (MAD)
    e. NbrP1 : indique le nombre des transactions observées dans la première période.
    f. NbrP2 : indique le nombre des transactions observées dans la deuxième période.
    g. Inter_achat : Nombre moyen de jours entre les achats.
11. Afficher les statistiques descriptives de base (min, max, moyenne, ecart type) pour les variables Longueur, Récence, Fréquence et Montant.