Test Data Scientist

Adopte - BI

2023

Consignes

Le fichier à rendre devra être sous format PDF. Il devra intégrer le code utilisé pour répondre aux questions ainsi que les graphiques demandés. Nous vous conseillons d'utiliser un notebook et de le convertir en format PDF.

Contexte

Julie du service marketing a mené une campagne dans la ville d'Angers pendant les 10 premiers jours de Mars 2023. Cette campagne avait pour objectif **d'augmenter le panier moyen des moins de 25 ans** lors de leurs **3 premiers achats**. Elle aimerait savoir si sa campagne a eu l'effet escompté au 31 Mars 2023.

En tant que stagiaire Data Scientist dans l'équipe BI, vous êtes en charge de répondre à l'interrogation de Julie. Pour cela vous disposez deux tables suivantes :

users		
Nom du champ	*Type*	*Description*
account_id	INTEGER	ID de l'utilisateur
birth_date	DATE	Date de naissance de l'utilisateur
zip_code	STRING	Code postal de l'utilisateur
register_ts	INTEGER	Timestamp d'inscription de l'utilisateur

purchases		
Nom du champ	*Type*	*Description*
account_id	INTEGER	ID de l'utilisateur
amount	INTEGER	Montant de l'achat (en centimes)
purchase_ts	INTEGER	Timestamp d'achat de l'utilisateur

Question 1 : Requête SQL

Ecrivez une requête SQL permettant d'extraire dans une nouvelle table la somme des 3 premiers achats des personnes s'étant inscrit à Angers entre le 01/03/2023 et le 10/03/2023.

Cette nouvelle table devra avoir les colonnes suivantes :

- ***ca*** : la somme des 3 premiers achats de l'individu.
- ***birth_date*** : la date de naissance de l'individu.
- ***account_id*** : l'ID de l'individu.
- ***zip_code*** : le code postal de l'individu lors de son inscription.
- ***register_ts*** : le timestamp d'inscription de l'individu.

Seules les personnes ayant fait au moins 3 achats devront être listé dans cette table.

Question 2 : Exploration et visualisation des données

Le résultat de votre requête SQL a été exporté dans le fichier "df.csv"

1. Lisez le fichier "df.csv" et créez la colonne *age*, l'âge des individus au 31 Mars 2023.
2. Explorez et analysez brièvement le jeu de donnée.
3. Afficher sur un graphique les dépenses des individus par rapport à leur âge. Avez-vous une première intuition de réponse à la question de Julie ?

Question 3 : Classification non supervisée

Pour confirmer votre intuition, vous décidez de mener une classification non supervisée sur le jeu de données.

1. Créez un nouveau jeu de données ne contenant que les colonnes *ca* et *age*.
2. Appliquez un algorithme du k-means sur ce nouveau jeu de données. Quel est le nombre optimal de cluster ? Proposez une méthodologie mathématique et appliquez la pour confirmer vos dires.
3. Afficher sur un graphique les dépenses des individus par rapport à leur âge en modifiant les couleurs des individus en fonction de leur cluster. Qu'allez-vous répondre à Julie ?