Soutenance Projet 4

Bonjour, je vais vous présenter mes travaux pour le projet 4 du parcours data analyst. Il s’agit ici de réaliser une analyse des ventes d’un site de e-commerce intitulé « rester livres » et qui propose de la vente en ligne de livres.

Pour réaliser cette étude nous avions trois jeux de données fournis :

* Customers, regroupant les clients, avec trois colonnes : l’identifiant d’un client, son genre, et son année de naissance.
* Products, regroupant les produits, avec trois colonnes également : l’identifiant du produit, son prix et sa catégorie, allant de 0 à 2
* Transactions, regroupant les ventes, avec quatre colonnes : le produit vendu, à quelle date, lors de quelle session et à quel client.

Avant de réaliser cette analyse, il fallait nettoyer ces jeux de données et je vais vous présenter le script.

Le script s’articule en deux blocs de codes : un pour l’identification et un pour le nettoyage des données.

L’identification s’effectue en exécutant une fonction pour chaque dataframe qui commence par afficher le type de données de chaque colonne et se termine par la vérification de la présence de doublons et/ou de valeurs manquantes.

Pour le tableau customers, j’ai également vérifié si tous les clients étaient d’abord vivants (âgés de moins de 122 ans) et ensuite majeurs. Puis j’ai vérifié si les clients étaient soit des hommes soient des femmes.

Pour le tableau products, j’ai vérifié si les produits étaient tous classés dans une catégorie 0, 1 ou 2, puis s’il y avait des prix d’une valeur négative.

Et enfin, pour le tableau Transactions, j’ai vérifié si tous les produits et les clients étaient bien enregistrés respectivement dans la table products et la table customers, puis s’il y avait des valeurs tests.

À l’issue de cette identification, j’ai constaté que le tableau customers était exploitable en l’état. Pour la table products, il y a des prix d’une valeur négative à supprimer. Et pour la table transactions, tout d’abord, il faut convertir les données de la colonne date au format datetime. Ensuite, il y a un produit qui n’est pas enregistré dans la table products, il y a des valeurs tests à supprimer, et il y a 126 doublons.

Par rapport à ces doublons, il est à noter que la table transactions ne comporte pas de colonne ‘quantité’. Ce qui signifie qu’en cas d’achat de plusieurs exemplaires du même produit par un même client à la même date et par conséquent lors de la même session, il y aura autant de lignes enregistrées dans la table transactions que d’exemplaires. Ces lignes seront donc perçues comme des doublons. J’ai donc identifié ces doublons comme un achat en plusieurs exemplaires et non pas comme une erreur de saisie de données. Et je ne les ai donc pas supprimés.

Le script de nettoyage était d’abord destiné à corriger ces valeurs aberrantes, rajouter le produit manquant dans la table products, puis convertir les données de la colonne date en datetime. Pour le produit manquant, j’ai identifié sa catégorie grâce au chiffre en préfixe de son identifiant, ici 0, puis je lui ai attribué comme prix, la moyenne des prix des produits de même catégorie.

Au cours de l’analyse, j’ai d’abord constaté que rajouter trois colonnes « année », « mois », « jour », issue de la colonne « date » me faciliterait la tâche. Ensuite, j’ai identifié quatre valeurs aberrantes parmi les clients dont le nombre de sessions d’achat et la dépense annuelle n’étaient pas crédibles.

- c\_1609 avec 162007,34 euros de dépenses en 12855 sessions

- c\_6714 avec 73220,8 euros de dépenses en 4473 sessions

- c\_3454 avec 54466,38 euros de dépenses en 3275 sessions

- c\_4958 avec 144257,21 euros de dépenses en 2562 sessions

J’ai donc rectifié le script de nettoyage en conséquence au cours de l’analyse.

J’en ai terminé avec la présentation du script de nettoyage, je vais passer à la présentation de l’analyse.

Corrélations :