

Task One



1 2 3 4 5

Exemple de réponse

Bon travail! Jetez un œil à l'exemple de réponse ci-dessous pour voir comment un professionnel aurait tenté cette tâche. Pensez à ce que vous avez bien fait et à la façon dont vous pouvez vous améliorer.

Cher [Point de contact client],

Merci de nous avoir fourni les trois ensembles de données de Sprocket Central Pty Ltd. Le tableau ci-dessous met en évidence les statistiques récapitulatives des trois ensembles de données reçus. Veuillez nous faire savoir si les chiffres ne correspondent pas à votre compréhension.

Nom de la table	Nombre d'enregistrements	Identifiants clients distincts	Date de réception des données
Données démographiques du client	-insérer la valeur-	-insérer la valeur-	-insérer la valeur-
Adresse du client	-insérer la valeur-	-insérer la valeur-	-insérer la valeur-
Données de transaction	-insérer la valeur-	-insérer la valeur-	-insérer la valeur-

Les problèmes notables de qualité des données rencontrés et les méthodes utilisées pour atténuer les incohérences des données identifiées sont les suivants. En outre, des recommandations ont été formulées pour éviter la réapparition de problèmes de qualité des données et améliorer l'exactitude des données sous-jacentes utilisées pour prendre des décisions commerciales.

- Customer_ids supplémentaires dans la « Table des transactions » et la « Table des adresses du client », mais pas dans la « Fiche client (données démographiques du client) ».**
Atténuation : Veuillez vous assurer que toutes les tables datent de la même période. Seuls les clients de la liste principale des clients seront utilisés comme ensemble de formation pour notre modèle.
Cela indique que les données reçues peuvent ne pas être synchronisées les unes avec les autres, ce qui peut fausser les résultats de l'analyse s'il manque des enregistrements de données. Veuillez vous référer au fichier Excel « data_outliers.xlsx » pour la liste des valeurs aberrantes entre les tableaux.

Des

Comple

- **Diverses colonnes, telles que la marque d'un achat ou l'intitulé du poste, ont des valeurs vides dans certains enregistrements.**

Atténuation : si seul un petit nombre de lignes sont vides, filtrez entièrement l'enregistrement de l'ensemble d'entraînement à des fins de prédiction. Sinon, s'il s'agit d'un domaine principal, imputez en fonction de la distribution dans l'ensemble de données de formation.

Pour les ensembles de données clés, tels que les transactions, moins de 1 % des transactions (totalisant moins de 0,1 % des revenus) comportent des champs manquants. Ces enregistrements ont été supprimés de l'ensemble de données de formation.

- **Valeurs incohérentes pour le même attribut (par exemple, Victoria étant représentée par « V », « Vic » et « Victoria »).**

Atténuation : utilisez une expression régulière pour remplacer les valeurs étendues par des abréviations afin de garantir la cohérence entre les adresses.

Recommandation : imposez une liste déroulante pour l'utilisateur qui saisit les données plutôt qu'un champ de texte libre. Afin de construire des variables significatives pour le modèle, les données ont été nettoyées pour éviter plusieurs représentations de la même valeur. De plus, les enregistrements de genre où « U » ont été remplacés en fonction de la distribution de l'ensemble de données de formation.

- **Type de données incohérent pour le même attribut (par exemple, valeurs numériques pour certains champs et chaînes pour d'autres).**

Atténuation : Convertissez les enregistrements sélectionnés en caractères en numériques. Supprimez les caractères non numériques de la chaîne. Recommandation : assurez-vous que les tables de faits dans la base de données donnée ont des contraintes sur les types de données.

Le fait d'avoir différents types de données pour un champ donné rend difficile l'interprétation des résultats à un stade ultérieur. Par conséquent, des transformations de données appropriées sont effectuées pour garantir des types de données cohérents pour un champ donné. Nom de la table Nombre d'enregistrements ID client distincts Date de réception Données démographiques du client -insérer la valeur- -insérer la valeur- -insérer la valeurAdresse du client -insérer la valeur- -insérer la valeur- -insérer la valeurDonnées de transaction -insérer la valeur- -insérer la valeur- -insérer la valeurRemarque : Les données et informations contenues dans ce document reflètent une situation et un client hypothétiques. Ce document doit être utilisé uniquement à des fins de stage virtuel KPMG.

À l'avenir, l'équipe poursuivra le processus de nettoyage, de normalisation et de transformation des données à des fins d'analyse du modèle. Des questions seront soulevées tout au long du processus et les hypothèses seront documentées. Une fois cette opération terminée, ce serait formidable de passer un peu de temps avec votre SME de données pour vous assurer que toutes les hypothèses sont alignées sur la compréhension de Sprocket Central.

Cordialement,
[Nom du consultant junior]

Remarque : Les données et informations contenues dans ce document reflètent une situation et un client hypothétiques. Ce document doit être utilisé uniquement à des fins de stage virtuel KPMG.

