DEADLINE : Vendredi 23 Juillet 2021 16h sur la plateforme SimplonOnLine
Les rendus après l'heure et la date ne seront pas évalués et les compétences non validées.

Vous êtes Développeur IA indépendant chez un marchand de vins de renom qui dispose d'un site internet pour ses ventes. Votre rôle est d'accompagner votre client dans sa transformation digitale. En effet pour le moment plusieurs painpoint sont présents. Par exemple, l'ERP utilisé n'est pas connecté au site internet : l'analyse des ventes sur internet est impossible, les outils sont rudimentaires, ect.. Votre mission se décline donc en plusieurs points :

- Faire un rapprochement entre l'export de la table de CMS qui contient les infos sur les produits vendus sur internet (nombre de ventes depuis la création du site internet, nom, ect..) et l'export de l'ERP (références produits, stock, ect..)
- Après cette centralisation, le client souhaite avoir le chiffre d'affaire par produit et le total de chiffre d'affaire réalisé
- Le client se pose également des questions sur certains prix produits pour voir des éventuelles erreurs de saisies. Sa demande consiste à détecter d'éventuelles valeurs aberrantes et d'en faire une représentation graphique
- Vous décidez de construire des profils de vins autour de chaque cluster. Tester donc l'algorithme de clustering kmeans et décrivez les cluster obtenus

NB : Un membre de l'équipe du client a créé un tableau Excel qui permet d'établir le lien entre la référence du produit dans l'ERP (product_id) et la référence du même produit dans la base de la boutique en ligne (SKU). La liste des product_id est exhaustive, mais pour les références côté Web, c'est moins sûr...Il a peiné à rapprocher certaines références.

L'analyse et l'identification des outliers se fera en utilisant les 4 méthodes suivantes:

- L'approche statistique : Méthode du z score (supérieur à 2 standard deviation) et méthode des interquartiles
- La méthode graphique : nuage de points (les outliers seront d'une autre couleur que le reste des points) et le boxplot

Bonus 1 : Choisir le nombre optimal de cluster grâce aux indicateurs Score de silhouette, le critère de coude, R2, ect… et décriver les cluster

Bonus 2 : Tester l'ACP pour décrire les cluster de manière adéquate

MODALITES ET CRITERES DE PERFORMANCE

-Le projet est individuel.

-Le code sera exclusivement du Python (Jupyter notebook) : notebook enrichi de commentaires, d'analyses , de conclusions et respectant PEP8.

-Le livrable final est un dépôt individuel Github. Il devra être **mis à jour tous jours durant le projet (Git add, git commit, git push)**