Brief projet
Exploration des données

Compétences

Qualifier les données grâce à des outils d'analyse et de visualisation de données en vue de vérifier leur adéquation avec le projet

Tags

Python, exploration des données

Durée

3 jours

Description rapide et ressources

Lance-toi dans ta première exploration sur des données qui serviront plus tard à alimenter un algorithme d'intelligence artificielle.

Ton travail sera évalué, lors de l'entretien technique que tu auras avec l'équipe Youcode le jour de la sélection.

Pour faciliter le développement de cette tâche d'exploration des données, nous proposons de faire des recherches sur Google, YouTube ces ressources suivantes :

Installation de l'environnement de développement

- **Installation de Python, Jupyter Notebook et les bibliothèques Python avec Miniconda**

Jupyter Notebook

- **Tuto – vidéo sur l'utilisation de Jupyter Notebook**
- **Tuto - utilisation de Jupyter Notebook (en anglais)**

Pandas

- **Tutoriel vidéo de Pandas**

- Pour la première partie du brief, tu pourras trouver la vidéo utile.
- Pour la deuxième partie du brief:
    - Value_counts et graphiques matplotlib
        - Groupby : équivalent des pivot table de Excel dans Pandas

Fiche client

Une entreprise de vente au détail, Sell4All, a ouvert ses portes il y a 6 mois. Leur travail consiste à faire de la vente de vêtements d'occasion sur Internet et souhaitent intégrer leur première fonctionnalité d'IA dans le site Web. Cette fonctionnalité permettra au site Web actuel de suggérer automatiquement certains produits. Les suggestions seront basées sur la similitude des données d'un utilisateur avec d'autres en tenant compte de ses données démographiques, du montant dépensé sur le Web et des produits achetés.

Pour commencer ce projet, les premières étapes nécessaires sont les suivantes :

1. Utiliser les données stockées sur les données démographiques et les dépenses des utilisateurs.
2. Explorer les données disponibles
3. Nettoyer les données
4. Enregistrer les données

Tes missions

Tu viens d'arriver chez Sell4All en tant que Data Développeur Junior et au cours de ta première semaine, tu es chargé d'assumer cet ensemble de tâches nécessaires pour alimenter ensuite l'algorithme d'IA avec les données explorées. Avant de mettre la main à la pâte, tu dois installer et configurer ton environnement de développement :

1. Installer Python via Miniconda
2. Installer Jupyter Notebook
3. Installez les bibliothèques Python :
    a. Pandas
    b. Matplotlib

Ensuite, tu dois créer et exécuter un programme Python sur le Jupyter Notebook qui est capable de :

- Lire les données du fichier CSV 'dataset-sell4all.csv' qui contient les données démographiques et les dépenses des utilisateurs,
- Afficher des informations sur les 5 premières lignes du fichier CSV,
- Afficher un résumé technique des données disponibles dans le fichier CSV avec des informations telles que :
    - nombre de lignes
    - les colonnes du fichier CSV
    - les types de données des champs du fichier CSV

- Expliquer les détails affichés du résumé technique dans une cellule de démarque du bloc-

notes Jupyter,

- Calculer la médiane et la moyenne des colonnes :
    - « Age »
    - « Customer spendings »

- Question bonus: Calculer la médiane d'âge pour chaque pays

- Créer une visualisation des données du graphique à barres qui montre les dépenses des clients par pays,
- Nettoyer les lignes avec moins de 10 € de dépenses par client : supprimez toutes les lignes d'utilisateurs ayant dépensé moins de 10 € sur le site,
- Nettoyer les doublons : supprimer toutes les lignes qui apparaissent plus d'une fois dans les données,
- Écrire les données nettoyées dans un nouveau fichier CSV avec uniquement les colonnes suivantes :
    - « Country »
    - « Âge »
    - « Gender »
    - « Customer spendings »

Livrables

Un fichier Jupyter Notebook contenant le travail effectué. Le candidat apporte son projet sur un support (usb) pour pouvoir présenter son travail et son code source lors de l'entretien technique de la journée de sélection.

Critères d'évaluation

- La syntaxe Python est utilisée de manière cohérente avec les objectifs de briefing ● La bibliothèque libre du pandas est utilisée efficacement pour résoudre les tâches proposées
- L'explication du résumé technique répond aux questions
    - Combien y a-t-il d'entrées dans l'ensemble de données ?
    - Qu'est-ce que « non nul » ?

    - Quels types de données sont présents dans l'ensemble de données ? Quels sont ces types de données ?

- Le candidat suit les prescriptions et les instructions reçues
- Le candidat sélectionne ce qui peut être utile pour tirer des leçons des propositions qui lui sont faites
- Le candidat cherche et applique ce qui pourrait combler les lacunes qu'il a identifiées
- Le candidat décrit les tâches qu'il a effectuées pour obtenir un résultat
- Le candidat organise ses tâches en fonction de différentes variables (planification, contraintes, objectifs, délais...)

Technologies

Python, Miniconda, Jupyter Notebook, pandas, matplotlib