Data Management, Data Viz & Text Mining

SDA 2023-2024

Sujet du projet : (En binôme)

Réalisation d'une application Streamlit qui permet d'afficher tout ce qu'il y a à savoir sur votre jeu de données. L'application devra contenir les informations suivantes :

- Description du jeu de données : l'origine des données, nombre d'observations, nombre de variables, types de variables, signification de chaque variable, nombre de valeurs manquantes par variable
- Statistiques descriptives
- Visualisations : Au moins cinq graphiques avec quelques filtres interactifs.

Etapes à suivre :

1. Choisir un jeu de données sur Kaggle ou OpenData qui contient à la fois des données catégorielles, des données numériques et des données temporelles.
    a. *https://www.kaggle.com/datasets*
    b. *Sites OpenData :*
        i. *https://opendata.paris.fr/*
        ii. *https://www.data.gouv.fr/fr/datasets/*
2. Réaliser toutes les étapes d'exploration des données & de nettoyage des données pour gérer les valeurs manquantes
3. Création d'au moins deux nouvelles variables en fonction des autres variables

4. Visualisation des tendances des différentes colonnes choisies en utilisant plusieurs types de graphiques
5. Création d'un wordcloud basé soit sur le texte de la description de votre jeu de données une fois traité (les étapes du pré-processing vues en cours) ou sur un texte que vous allez générer en utilisant une ou plusieurs colonnes.

Fichiers à rendre :

Une archive tar qui contient :

1. Un fichier requirements.txt avec toutes les librairies & leurs versions utilisées dans votre environnement virtuel.
2. Le lien vers le jeu de données utilisé
3. Votre code :
    a. Le Notebook Jupyter qui contiendra toutes vos étapes de Data Management
    b. Vous devrez ensuite charger votre dataframe traité dans un fichier csv que vous allez utiliser sur Streamlit pour la visualisation
    c. Le/s fichiers .py de votre application Streamlit qui permettra d'afficher toutes les informations citées en haut.