Data Management, Data Visualisation & Text Mining

SDA 2025

Projet de fin de module

Groupe : 2 à 3 personnes

Composition des groupes à envoyer par email à :

malki.fatimaezzahrae@gmail.com

Avant le Vendredi 6 juin

Suivi du projet : - Vendredi 6 juin, 18h30 à 21h

- Vendredi 13 juin, 18h30 à 21h

Soutenance : - Lundi 16 juin, 18h30 à 21h

- Mardi 17 juin, 18h30 à 21h

Les livrables doivent être envoyés par email à l'adresse suivante :
malki.fatimaezzahrae@gmail.com

Avant le dimanche 15 Juin à 22h

Sujet du projet :

Développez une **application interactive avec Streamlit** permettant d'explorer en détail un jeu de données. L'application devra présenter :

- Une présentation du jeu de données :
  - Source et origine des données
  - Nombre d'observations et de variables
  - Types et significations des variables
  - Nombre de valeurs manquantes par variable, etc.
  - Etc.
- Des statistiques descriptives
- Des visualisations interactives :
  - Minimum 5 graphiques
  - Intégration de filtres dynamiques (ex. : menus déroulants, sliders, checkbox, etc.)

Étapes à suivre :

1. Choisir un jeu de données :

Utilisez un dataset contenant **au minimum 200 000 lignes**, avec une diversité de **données numériques, catégorielles et temporelles.**

- Kaggle Datasets https://www.kaggle.com/datasets
- OpenData https://opendata.paris.fr/
- Data.gouv.fr https://www.data.gouv.fr/fr/datasets/

2. Phase de préparation des données :

- Analyse exploratoire des données
- Nettoyage des données (traitement des valeurs manquantes, des doublons, des incohérences)
- Justifiez toutes vos décisions de transformation ou de suppression.

3. Création de variables :

Générez au moins deux nouvelles variables pertinentes, dérivées des données existantes.

4. Visualisation des tendances :

Utilisez différents types de graphiques pour illustrer les tendances et relations

5. Partie text mining :

Trouvez un article de presse en lien avec votre thème de données (ex : transport, santé, climat, économie…).

- Effectuez un prétraitement du texte (nettoyage, tokenisation, suppression des stopwords…)
- Générez un nuage de mots (WordCloud) à partir du texte nettoyé

Fichiers à rendre (dans une archive .zip) :

1. requirements.txt : liste de toutes les librairies **utilisées**, avec leurs versions
2. Un fichier texte avec les noms et prénoms des membres du groupe et un lien vers le jeu de données utilisé
3. Un notebook Jupyter (.ipynb) pour toute l'étape de data management (exploration, nettoyage, création de variables)
4. Un fichier CSV du jeu de données final, nettoyé
5. Un fichier Python .py pour votre application Streamlit complète