Projet 1

Cours Big Data – L3

OVNI - Observations Volumineuses Non Interprétées

Objectif du projet

Ce projet a pour objectif de mettre en œuvre un **pipeline complet d'analyse de données** à partir d'un jeu de données fourni, afin de répondre à des questions d'analyses définies.

Organisation

- Travail en monôme ou binôme.
- Le projet est réalisé pendant deux séances de TP et aussi chez soi.
- Les échanges et questions pendant les séances sont encouragés **MAIS** à un niveau sonore acceptable.
- Le rapport ne doit pas dépasser 4 pages **maximum**, et est à rendre pour le **28 février** 23h59 et 59 secondes **au plus tard**.
- L'utilisation d'outils d'intelligence artificielle est autorisée, voire encouragée, uniquement dans une démarche d'apprentissage (aide à la compréhension, au débogage ou à l'exploration d'idées). L'objectif n'étant en aucun cas d'évaluer l'IA elle-même, toute utilisation se limitant à du copier-coller, y compris de conclusions, sera sanctionnée par une réduction significative de la note finale. Avec un minimum d'expérience, ce type de pratique est relativement facile à détecter.

Jeu de données

Le jeu de données utilisé pour ce projet est mis à disposition sur Community. Il s'agit d'un jeu de données réel, susceptible de contenir des valeurs manquantes ou incohérentes.

Il se présente sous la forme d'une table regroupant plus de 80 000 observations d'apparitions d'OVNI entre 1949 et 2014, comprenant notamment la ville, l'État, le pays, les coordonnées géographiques, la forme observée, la durée, la date et l'heure de l'événement, ainsi que des commentaires descriptifs.

Questions d'analyse

Ce projet vise à explorer le jeu de données des observations d'OVNI à travers plusieurs axes complémentaires : analyse descriptive, temporelle, géographique et textuelle.

Les questions suivantes structurent l'analyse :

1. **Quelles formes d'OVNI sont observées le plus fréquemment ?** Peut-on identifier un nombre limité de formes dominantes dans les signalements ?
2. **Comment le nombre d'observations d'OVNI a-t-il évolué depuis l'an 2000 ?** Observe-t-on une tendance particulière au fil des années (augmentation, diminution, stabilité) ?
3. **Aux États-Unis, quels États rapportent le plus d'observations d'OVNI ?** Comment peut-on interpréter ces résultats (population, biais de déclaration, contexte culturel, etc.) ?
4. **Quelle est la distribution de la durée des observations d'OVNI ?** Que révèle l'analyse statistique de cette variable après nettoyage des données ?
5. **Quelles informations intéressantes peut-on extraire des commentaires textuels ?** Peut-on faire émerger des thèmes, mots-clés ou tendances qualitatives à partir des descriptions ?

Des analyses complémentaires pertinentes pourront être valorisées dans l'évaluation.

Travail attendu

Le travail attendu suit les différentes étapes d'un pipeline classique d'analyse de données.

1. Compréhension et présentation des données

— présentation générale du jeu de données ;
— description des principales variables utilisées ;
— identification des problèmes de qualité des données (valeurs manquantes, formats incohérents, valeurs aberrantes).

2. Nettoyage des données

Cette étape est essentielle pour la suite de l'analyse.

— gestion des valeurs manquantes ;
— traitement des incohérences et des valeurs aberrantes ;
— transformation des variables si nécessaire (dates, durées, catégories, texte) ;
— justification claire et argumentée des choix effectués.

3. Analyse exploratoire

a) Analyse des formes d'OVNI

— identification des formes les plus fréquemment observées ;
— visualisation adaptée (diagramme en barres, par exemple) ;
— interprétation des résultats.

b) Analyse temporelle (depuis 2000)

— étude de l'évolution annuelle du nombre d'observations ;
— mise en évidence de tendances ou de variations notables ;
— discussion des résultats obtenus.

c) Analyse géographique (États-Unis)

— comparaison du nombre d'observations entre les États ;
— visualisation des résultats ;
— interprétation critique des différences observées.

4. Analyse de la durée des observations

— nettoyage spécifique de la variable de durée ;
— analyse de la distribution des durées ;
— construction d'un histogramme ou d'une visualisation équivalente ;
— interprétation des résultats (asymétrie, valeurs extrêmes, etc.).

5. Analyse des commentaires textuels

— exploration de la colonne `comments` ;
— extraction d'informations pertinentes (mots fréquents, thèmes, longueur des textes, etc.) ;
— présentation d'au moins une analyse originale issue des données textuelles.

6. Visualisation et interprétation

— production de graphiques clairs et lisibles ;
— titres, axes et légendes correctement renseignés ;
— interprétation systématique des figures produites.

7. Conclusion

— synthèse des principaux résultats obtenus ;
— discussion des limites de l'analyse et du jeu de données ;
— pistes d'amélioration ou d'analyses complémentaires possibles.

Livrables

— un notebook Jupyter propre et structuré ;
— un court rapport (4 pages maximum) présentant l'analyse et les résultats.

Évaluation

Ce projet est noté et compte pour **1/2 de la note finale** de l'UE.

Critère	**Poids**
Qualité du code et du pipeline (Python)	20%
Méthodes d'analyse	20%
Analyse et interprétation des résultats	20%
Qualité des visualisations	20%
Rapport écrit (clarté et structure)	20%

Remarques

— La clarté du raisonnement est privilégiée par rapport à la complexité technique.
— Les résultats doivent être interprétés et discutés.
— Ce projet sert de base méthodologique pour le Projet 2.