Projet 2 – Étude de données libre guidée (question-driven)

Cours Big Data – L3

Objectif du projet

Le **Projet 2** consiste à conduire une étude de données **de bout en bout** à partir d'une question claire, en mobilisant les outils vus en cours (Python, Jupyter, bibliothèques de data science). L'objectif n'est pas de produire un projet "technique" complexe, mais une analyse **rigoureuse, lisible, et argumentée**.

Organisation

— Travail en **monôme**, **binôme** ou **trinôme**.

— Le nombre de pages du rapport et la durée de la présentation orale dépendent de la taille du groupe, comme indiqué ci-dessous :

	Nombre de pages	**Présentation**
Monôme	5	5 minutes
Binôme	7–8	7 minutes
Trinôme	9–10	10 minutes

— Le projet est réalisé lors des deux dernières séances de TP et complété par un travail personnel en dehors des séances.

Principe directeur : une question, des données, une réponse

Le projet doit être construit autour d'une question du type :

— *Quels facteurs influencent la propagation d'une maladie dans un pays ?*

— *Quels facteurs influencent le prix des actions d'une entreprise sur une période donnée ?*

— *Quels hashtags génèrent le plus d'engagement sur Twitter pour un événement donné ?*

— *Comment les températures évoluent-elles selon les villes et les saisons ?*

Important : la question peut être adaptée selon les données choisies, mais elle doit rester **spécifique** (périmètre, période, variables, population/zone).

Choix du jeu de données

Le choix du dataset est **libre** (open data, API, données publiques, données collectées), à condition de respecter les contraintes suivantes :

— données **licites** et **réutilisables** (licence claire) ;

— dataset d'une taille **raisonnable** (pas besoin d'être énorme) ;

— présence d'au moins une dimension structurante (temps, espace, catégories, groupes, etc.) ;

— capacité à produire **au moins 3 visualisations pertinentes**.

Note : l'utilisation de données sensibles (santé nominative, informations personnelles, etc.) est à éviter.

Étapes attendues (pipeline)

Le projet suit un pipeline simple et explicite :

1. **Formulation de la question** (1 paragraphe clair, hypothèses éventuelles).
2. **Présentation des données** (source, variables principales, limites).
3. **Prétraitement / nettoyage** (valeurs manquantes, types, doublons, incohérences).
4. **Analyse exploratoire** (statistiques descriptives, comparaisons, tendances).
5. **Visualisations** (figures propres, lisibles, commentées).
6. **Interprétation et discussion** (réponse à la question + limites + pistes).

Validation du sujet (obligatoire)

Afin d'éviter les sujets trop vagues ou irréalisables, une **validation courte** est demandée :

— titre du projet ;

— question (1–2 phrases) ;

— source(s) de données ;

— variables d'intérêt pressenties + 2–3 analyses envisagées.

Cette validation se fait au début de la période projet (modalités précisées en séance).

Livrables

— **Notebook Jupyter** propre, exécutable, commenté (et structuré en sections).

— **Rapport** (6 à 10 pages max) :

  — question et contexte,

  — description des données,

  — méthodologie (nettoyage + analyse),

  — résultats et figures,

  — discussion + limites,

  — références (sources de données, éventuelles sources externes).

— **Présentation orale** (support libre : slides, notebook, etc.).

Évaluation

Le Projet 2 compte pour **1/2 de la note finale** de l'UE.

Critère	**Poids**
Maîtrise Python et qualité du code	20%
Méthodes Big Data et pipeline d'analyse	20%
Analyse et interprétation des résultats	20%
Présentation orale	20%
Rendu écrit : clarté, structure, professionnalisme	20%

Présentation orale : attentes

Lors de l'oral, il est attendu :

— la question et le choix des données en 30 secondes ;
— 2–3 résultats majeurs (avec figures) ;
— une discussion honnête des limites ;
— une conclusion claire : *que peut-on affirmer, et que ne peut-on pas affirmer ?*

Pistes de sujets (idées prêtes à démarrer)

Climat / environnement

— Évolution des températures (villes, régions, saisons) ; anomalies et tendances.
— Qualité de l'air (PM2.5, NO_2) : tendances, comparaisons entre zones.
— Événements extrêmes (canicules, pluies) : fréquence et saisonnalité.

Économie / finance

— Volatilité d'actions : variations, périodes, comparaison entreprises/secteurs.
— Inflation / pouvoir d'achat : tendances temporelles, comparaison pays/régions.

Mobilité / territoires

— Trafic / transport : pics horaires, différences jour/semaine, saisons.
— Accidents : facteurs (météo, heure, zone), cartographie simple.

Réseaux sociaux / web

— Hashtags et engagement : volume, temporalité, comparaison événements.
— Analyse simple de texte : fréquence de mots, hashtags, longueur (sans NLP avancé).

Culture / divertissement

— Films/musiques : genres, popularité, tendances temporelles, corrélations simples.
— Jeux vidéo : notes, plateformes, années, genres (évolution et comparaisons).

Utilisation de l'intelligence artificielle (IA)

— L'utilisation d'outils d'intelligence artificielle est autorisée dans le respect de bonnes pratiques. Ils peuvent être utilisés comme aide à l'apprentissage ou au débogage, mais les étudiants restent responsables du contenu rendu. Ces outils sont aujourd'hui très efficaces, et nous en avons pleinement conscience ; ils ne sont donc pas l'objet de l'évaluation. Une compréhension insuffisante sera en revanche identifiable, aussi bien à l'écrit que lors de la présentation orale.

Conseils pratiques (pour réussir)

— Mieux vaut une question **simple et bien traitée** qu'une question ambitieuse et inachevée.
— Garder des figures sobres : peu de graphiques, mais **tous commentés**.
— Toujours distinguer : **observations** (ce que montrent les données) et **interprétations**.
— Les limites font partie du travail : elles améliorent la qualité du projet.