

PROJET DE GROUPE

- Conditions :
 - 2 à 4 personnes par groupe
- 1 repo github :
 - Les fichiers python ou notebooks
 - Les données
 - 1 fichier requirements.txt (avec la liste des librairies nécessaires)
 - 1 présentation powerpoint avec des graphique, app streamlit ou dashboard Power bi

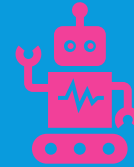
ORGANISATION



1 chef de groupe



Une liste de tâche à accomplir



kanban (trello ou sur github ou autre)



Répartition des tâches (ex :
ingestion / stockage /
transformation / requêtage &
graphiques / dashboards)

PYTHON ETL (EXTRACT TRANSFORM LOAD)

Extraction

- Excel / CSV / Parquet / API (exemple : API kaggle/datagouv)

Transformation

- Colonnes (création, suppression, filtrage, calcul)
- Valeurs (manquantes, remplacement)
- Aggregation / Jointures

Chargement

- Excel / CSV / BDD / Parquet

PROPOSITION STRUCTURE PROJET

- Repo Github
 - 1 fichier notebook (ipynb) ou 1 fichier python (.py)
 - 1 fichier pptx correspondants à présentation
- 1 dossier « data » :
 - 1 dossier « raw » : fichiers source brutes
 - 1 dossier « interim » : fichiers nettoyés
 - 1 dossier « processed » : fichiers avec à l'intérieur uniquement les données utilisées
 - 1 dossier « external » (facultatif) : fichiers externes
- 1 dossier « src »
- 1 fichier requirements.txt (pip freeze ...)

ANALYSE DE DONNÉES

- Statistiques globales
- Description du jeu de données
- Etat des lieux concernant la situation de la population du dataset
 - Existe-t-il des groupes ? (minorité, majorité, plusieurs profils)
 - Existe-t-il un problème à analyser dans ce dataset ?
 - Pouvez-vous proposer des solutions à ces problèmes ?
 - Que peut-on tirer de ce dataset ?



Création d'1 client fictif (nom, charte graphique, domaine, taille entreprise, problématique)



1 Sujet Global avec 1 ou des problématiques auxquelles vous répondez



Présentation Professionnelle (utiliser le jargon du domaine en question s'il en existe un)



Préparez une FAQ (anticipez les questions)



Être prêt pour les questions

STORYTELLING POUR LA PRÉSENTATION

LES LIBRAIRIES/OUTILS À UTILISER (OU PAS)

- Pandas/Polars
- SQL
- Numpy
- Matplotlib
- Seaborn
- Plotly
- Requests
- Streamlit
- Duckdb/Sqlite/Postgresql
- Docker
- CSV/Parquet
- (Power bi)
- IA (au besoin mais soyez maitre de votre code)

LES SOURCES DE DONNÉES

<https://datasetsearch.research.google.com/>

<https://www.mavenanalytics.io/data-playground>

- Github Public Dataset (tous les thèmes) <https://github.com/awesomedata/awesome-public-datasets>
- Kaggle (tous les thèmes) <https://www.kaggle.com/datasets?fileType=csv>
- BuzzFeed News Dataset (données sociales, politiques, démographiques)
<https://github.com/orgs/BuzzFeedNews/repositories>
- Inside Airbnb (données airbnb - location de logement) <https://insideairbnb.com/get-the-data/>
- Données publiques françaises (données liées à la France) <https://www.data.gouv.fr/fr/>
- Données Economique en France (emploi, masse salariale du secteur privé, embauches, taux d'impayés, délais de paiement)
open.urssaf.fr
- Omdena (beaucoup de données liées à la santé) <https://www.omdena.com/projects>
- Data World (tous les thèmes) <https://data.world/datasets/open-data>

!!! Privilégier Kaggle

LES SOURCES DE DONNÉES

- Google data search <https://datasetsearch.research.google.com/>
- Maven Analytics (tous les thèmes) <https://www.mavenanalytics.io/data-playground>
- Github Public Dataset (tous les thèmes) <https://github.com/awesomedata/awesome-public-datasets>
- Kaggle (tous les thèmes) <https://www.kaggle.com/datasets?fileType=csv>
- BuzzFeed News Dataset (données sociales, politiques, démographiques) <https://github.com/orgs/BuzzFeedNews/repositories>
- Inside Airbnb (données airbnb - location de logement) <https://insideairbnb.com/get-the-data/>
- Donnée publiques françaises (données liées à la France) <https://www.data.gouv.fr/fr/>
- Données Economique en France (emploi, masse salariale du secteur privé, embauches, taux d'impayés, délais de paiement) open.urssaf.fr
- Omdena (beaucoup de données liées à la santé) <https://www.omdena.com/projects>
- Data World (tous les thèmes) <https://data.world/datasets/open-data>

EXEMPLES DE PROJETS

- Analyse boutiques retails (3 tables: magasins, caractéristiques et ventes)
<https://www.kaggle.com/datasets/manjeetsingh/retaildataset?select=Features+data+set.csv>
- Carte Shopping (4 tables: clients, commandes, ventes, produits)
<https://www.kaggle.com/datasets/ruchi798/shopping-cart-database?select=customers.csv>
- Données Économiques (les pays sont les clefs primaires) <https://donnees.banquemondiale.org/>
- DB Chinook (11 tables: chanson, genre, employé, artiste, clients, etc...)
<https://www.sqlitetutorial.net/sqlite-sample-database/>
- Baseball (20 tables : salaire, joueurs, hall of fame, equipes, manager, awards, etc..)
<https://www.kaggle.com/datasets/open-source-sports/baseball-databank>
- Football Européen (7 tables: equipes, joueurs, ligue, matchs, etc...)
<https://www.kaggle.com/datasets/hugomathien/soccer>
- NBA (14 tables: equipe, joueur, draft, score, etc...)
<https://www.kaggle.com/datasets/wyattowalsh/basketball>
- E-commerce (9 tables: clients, geolocalisation, produits, paiements, vendeurs, etc...)
<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>