Rapport

Dans ce TP, j'ai travaillé sur l'intégration des données AdventureWorks pour créer un mini pipeline ETL en Python. J'ai rencontré plusieurs problèmes de qualité : des valeurs manquantes dans les colonnes importantes (dates, prix, quantités), des types incorrects (dates en texte, prix en string), des incohérences comme des quantités à zéro ou des remises négatives, et des doublons dans les ventes.

Pour nettoyer, j'ai converti les types pour éviter les erreurs, supprimé les lignes avec des valeurs manquantes critiques, filtré les incohérences (quantité > 0 et remise >= 0) et enlevé les doublons. Le pipeline est simple : 4 fichiers CSV → Extraction avec pandas → Nettoyage (types, valeurs, doublons) → Jointures → Table finale (clean_adventureworks_sales_2020.csv).

Les indicateurs clés que j'ai calculés sont le chiffre d'affaires par catégorie de produit et par année, le top 10 des clients par chiffre d'affaires, et le chiffre d'affaires par territoire. Ce travail m'a permis de comprendre comment préparer des données avant l'analyse.