



RAPPORT DE

Projet Superstore

Analytics

Préparation des données

LAILA BAHIJ

26/02/2026

SOMMAIRE

1 DÉFINITION DU PROJET

2 OBJECTIF DU PROJET

3 SOLUTION ET APPROCHE

4 ÉTAPES DÉTAILLÉES

5 LIVRABLES

6 CONCLUSION

DÉFINITION DU PROJET

DÉFINITION

Le projet vise à transformer les données de ventes brutes d'une entreprise de distribution multi-catégories (mobilier, fournitures de bureau, technologie) en un dataset propre, structuré et exploitable.

Le problème principal : le fichier CSV brut présente de multiples anomalies :

- Valeurs manquantes
- Incohérences de formats (dates, types numériques)
- Doublons
- Colonnes mal structurées
- Données non normalisées
- Absence d'indicateurs métier calculés

OBJECTIF DU PROJET

OBJECTIF DU PROJET

produire un dataset fiable permettant aux équipes métiers de comprendre la performance commerciale et financière, d'analyser les ventes par catégorie, région et segment, et de poser les bases pour un dashboard et une base PostgreSQL

SOLUTION ET APPROCHE

SOLUTION ET APPROCHE

Le projet suit le cycle de vie de la donnée :
Ingestion → Nettoyage → Transformation & Feature
Engineering → Préparation finale & Export

- **Ingestion et exploration :** lecture et analyse du fichier CSV pour identifier les anomalies et comprendre la structure des données.
- **Nettoyage :** correction des valeurs manquantes, suppression des doublons, harmonisation des types et formats.
- **Transformation et feature engineering :** création de métriques métiers (marge, taux de profit, ratio quantité/vente), extraction des informations temporelles (année, mois, trimestre), catégorisation et ajout de colonnes dérivées.
- **Préparation finale et export :** vérification de la cohérence, ajout des colonnes nécessaires pour SQL et visualisation, export du dataset nettoyé (superstore_clean.csv).

ÉTAPES DÉTAILLÉES

SOLUTION ET APPROCHE

3.1 Importation et exploration

- Lire le fichier CSV (store.csv) avec Pandas
- Inspecter les colonnes et types de données
- Identifier les valeurs manquantes et les anomalies
- Analyser les distributions et repérer les outliers

3.2 Nettoyage

- **Valeurs manquantes** : remplacement ou suppression selon le contexte métier
- **Doublons** : suppression ou fusion des enregistrements identiques
- **Types de colonnes** : conversion des dates, des colonnes numériques et catégorielles
- **Harmonisation** : uniformisation des formats de dates et des textes

SOLUTION ET APPROCHE

3.3 Transformation et feature engineering

- **Extraction des informations temporelles :** année, mois, trimestre
- **Création de nouvelles métriques :**
 - Marge = Profit / Ventes
 - Taux de profit
 - Ratio quantité / ventes
- **Catégorisation :** segmentation clients, catégories de produits
- Ajout de colonnes dérivées pour analyses **futures** : délais de livraison, indicateurs de performance
- **Agrégation pour analyse :** ventes par produit, par région, par catégorie

3.4 Préparation finale et export

- Vérification de la cohérence globale du dataset
- Confirmation que toutes les colonnes nécessaires pour SQL et visualisation sont présentes
- **Export du dataset final nettoyé :** superstore_clean.csv

LIVRABLES

LIVRABLES

- Dataset propre et transformé : superstore_clean.csv
- Document explicatif (PDF) des étapes de nettoyage et transformation
- Tableau de suivi Jira des tâches
- Notebook ou script Python documenté

CONCLUSION

CONCLUSION

Le projet a transformé un fichier CSV brut et incohérent en un dataset propre, structuré et exploitable pour les analyses métiers.

Les problèmes de qualité ont été corrigés et des indicateurs commerciaux et financiers ont été calculés.

Des colonnes dérivées et des transformations ont permis une segmentation par clients, catégories et périodes.

Le dataset final `superstore_clean.csv` est prêt à être utilisé dans une base SQL ou des outils de visualisation pour soutenir la prise de décision.