Description structurée de la base de données Petrol/Gas Prices Worldwide

El Baz Younes
22007219 - CAC2

26 Novembre 2025

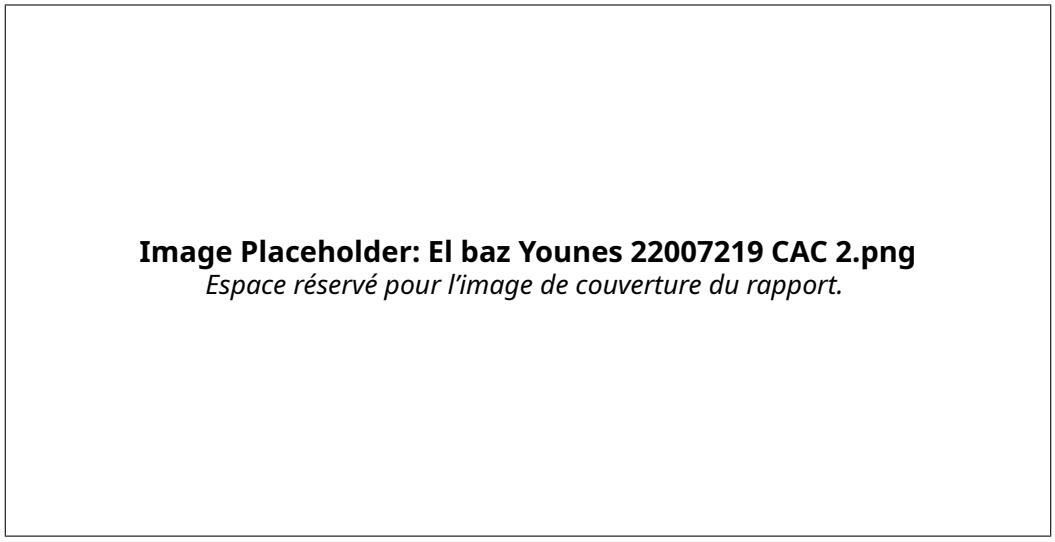

Figure 1: Placeholder pour l'image de couverture du rapport.

Compte Rendu

Analyse du Dataset "Petrol Consumption"

Table des matières

1 Introduction et Contexte

Le projet consiste à analyser un dataset portant sur la **consommation de pétrole (Petrol Consumption)** par région. L'objectif principal est de :

- Comprendre les facteurs influençant la consommation.
- Examiner la relation entre variables économiques et consommation énergétique.
- Construire un modèle prédictif simple afin d'anticiper le niveau de consommation en fonction de différents paramètres.

Ce rapport résume toutes les étapes réalisées dans Google Colab : Chargement, Nettoyage, Visualisation, Analyse statistique, Modélisation et Interprétation des résultats.

2 Chargement & Description des Données

Le dataset a été chargé depuis un fichier CSV. Les premières analyses permettent d'observer :

- **Nombre d'observations :** 48
- **Nombre de variables :** 5
- **Colonnes :**
  - *Petrol_tax* : taxe sur le carburant
  - *Average_income* : revenu moyen
  - *Paved_Highways* : routes pavées (en miles)
  - *Population_Driver* : proportion de conducteurs
  - *Petrol_Consumption* : consommation annuelle (cible)

Les premières lignes du dataset confirment une structure propre et exploitable. Les statistiques descriptives montrent une grande variabilité des revenus, une forte dispersion du kilométrage de routes pavées et des valeurs de consommation non uniformes.

3 Nettoyage et Prétraitement

Après analyse :

- **Aucune valeur manquante** (✔)
- **Aucun doublon détecté**
- Colonnes toutes numériques (✔) $\rightarrow$ idéal pour la régression

Le dataset était donc propre dès le départ, nécessitant seulement quelques vérifications.

4 Analyse Exploratoire (EDA)

Target: comprendre la distribution des variables

Les histogrammes et boxplots ont révélé :

- *Petrol_tax* $\rightarrow$ distribution homogène, peu d'outliers
- *Average_income* $\rightarrow$ très dispersé
- *Paved_Highways* $\rightarrow$ distribution très étalée
- *Petrol_Consumption* $\rightarrow$ distribution non normale, présence de valeurs extrêmes

Ces observations orientent la modélisation vers des algorithmes robustes aux dispersions.

5 Analyse Statistique & Corrélations

Une matrice de corrélation a été produite.

Corrélations observées :

Variable	Corrélation avec la consommation
Petrol_tax	Négative (augmentation taxe $\rightarrow$ baisse consommation)
Average_income	Faible
Paved_Highways	Très faible
Population_Driver	Faible

Table 1: Synthèse des corrélations

Conclusion : La taxe sur le carburant est le facteur le plus déterminant.

6 Modélisation et Prédiction

Un modèle de **régression linéaire simple** a été appliqué.

Étapes suivies :

1. Séparation features (X) et cible (y)
2. Entraînement du modèle avec `LinearRegression()`
3. Prédiction sur l'ensemble
4. Calcul des métriques : R^2, MSE, RMSE
5. Visualisation "Valeurs réelles vs valeurs prédites"

7 Résultats des Modèles

Performance du modèle linéaire :

- **R^2 faible** $\rightarrow$ le modèle explique peu de variance
- **MSE élevé** $\rightarrow$ erreurs importantes
- **RMSE important** $\rightarrow$ grande distance entre valeurs réelles et prédictions

Conclusion intermédiaire :

La relation semble **non linéaire**, ce qui rend la régression linéaire peu adaptée.

8 Analyse, Interprétation & Recommandations

Ce que montre l'analyse :

- La taxe sur le carburant influence nettement la consommation
- Les autres variables ont un poids relativement faible
- Le modèle linéaire ne capture pas bien la complexité des données

Recommandations :

1. Tester des modèles non linéaires (*Random Forest*, *Decision Tree*, *Gradient Boosting*)
2. Tenter des transformations : logarithmes, standardisation
3. Créer de nouvelles variables (ex : *income/tax*)
4. Retirer ou traiter les outliers extrêmes
5. Collecter davantage de données pour affiner la granularité

9 Conclusion

Cette étude a permis d'obtenir :

- ✔ Une compréhension claire des facteurs influençant la consommation
- ✔ Une validation de l'impact de la taxe sur le carburant
- ✔ Un aperçu des limites de la régression linéaire
- ✔ Une base solide pour des modélisations plus avancées

Malgré les performances modestes du modèle, ce travail constitue une étape essentielle pour une analyse prédictive plus précise à l'avenir.