Exercice de Régression : Prédiction du Prix des Voitures

Objectif

L'objectif de cet exercice est d'utiliser une **régression linéaire** pour prédire le **prix des voitures** en fonction de leurs caractéristiques techniques. Nous travaillerons avec le dataset **Automobile** disponible sur Kaggle.

Description du Dataset

Le dataset contient plusieurs caractéristiques des voitures, telles que :

- **Caractéristiques générales** : marque (`make`), type de carburant (`fuel-type`), type de carrosserie (`body-style`), type de roues motrices (`drive-wheels`), etc.
- **Caractéristiques dimensionnelles** : empattement (`wheel-base`), longueur (`length`), largeur (`width`), hauteur (`height`), poids (`curb-weight`).
- **Caractéristiques du moteur** : type de moteur (`engine-type`), nombre de cylindres (`num-of-cylinders`), taille du moteur (`engine-size`), système d'alimentation (`fuel-system`), puissance (`horsepower`), consommation en ville (`city-mpg`) et sur autoroute (`highway-mpg`).
- **Prix du véhicule (`price`)**, qui sera notre **variable cible (y)**.

Tâches à réaliser sur Google Colab

1. **Chargement des données** : Importer le dataset depuis Kaggle et le charger dans un DataFrame Pandas.
2. **Prétraitement des données** :
   - Gérer les valeurs manquantes (`?`).
   - Convertir les colonnes nécessaires en format numérique.
   - Supprimer les valeurs aberrantes éventuelles.
3. **Exploration des données** :
   - Afficher les statistiques du dataset (`describe()`).
   - Visualiser les corrélations entre les variables (matrice de corrélation, diagrammes de dispersion).

4. **Sélection des variables explicatives** : Identifier les variables ayant une forte corrélation avec `price` pour la régression.
5. **Création du modèle de régression linéaire** :
    - Séparer les données en ensembles d'entraînement et de test.
    - Entraîner un modèle de régression linéaire avec Scikit-Learn.
    - Évaluer les performances avec des métriques comme le **R^2, RMSE, MAE**.
6. **Visualisation des prédictions** : Comparer les prix réels et prédits avec un **graphique de dispersion**.

Livrable attendu

Un **notebook Google Colab** contenant :

- Le code proprement commenté.
- Les analyses exploratoires et graphiques.
- Le modèle de régression linéaire et son évaluation.