PFM MACHINE LEARNING

PRÉDICTION DE PRIX DES VOITURES

# Exemple d'image

Par: Narjiss Abdelkarim - Radouani Mouad

Encadrant: Mme. Laila ELJIANI.

## 

## Introduction :

🚗 **Plongez dans l’avenir de l’automobile avec notre projet innovant !**

Imaginez un monde où acheter ou vendre une voiture ne rime plus avec incertitude, mais avec **précision, rapidité et intelligence artificielle**. Notre projet de fin module (PFM) se place au cœur de cette révolution, en développant un **modèle de prédiction des prix des voitures** alliant **data science, machine learning et analyse avancée** pour anticiper les tendances du marché avec une justesse impressionnante.

💡 **Pourquoi ce projet ?**Le marché automobile est un écosystème dynamique où les prix fluctuent sous l’effet de multiples facteurs : kilométrage, marque, année, carburant, et bien d’autres. Aujourd’hui, acheteurs et vendeurs naviguent souvent à l’aveugle, mais grâce à notre solution, **l’estimation devient scientifique, transparente et ultra-performante**.

🔥 **Notre ambition ?**Créer un outil prédictif **puissant et accessible**, capable d’apprendre en temps réel et de s’adapter aux évolutions du marché. En exploitant des **algorithmes de pointe** , nous visons non seulement à résoudre un problème concret, mais aussi à **repousser les limites de l’IA dans le domaine automobile**.

📈 **Au-delà des données, notre projet représente un pont vers l’innovation**, une opportunité de marier technologie et quotidien pour simplifier les décisions financières de millions d’utilisateurs.

Prêt à embarquer dans l’aventure ? **Accrochez-vous, le meilleur reste à venir !** 🔥

## Table de matière:

## 

## Partie I : Présentation générale du projet

### Contexte : Derrière Chaque Voiture, Il Y a Une Histoire

Acheter ou vendre une voiture, c'est souvent bien plus qu'une simple transaction. C'est une décision qui impacte le quotidien des gens. Pourtant, aujourd'hui, déterminer le juste prix reste un véritable casse-tête. Entre les estimations approximatives et les variations brutales du marché, les utilisateurs naviguent à vue.

C'est là que notre projet entre en jeu. En nous appuyant sur la puissance des données et de l'intelligence artificielle, nous voulons redonner confiance aux acheteurs et aux vendeurs, en transformant une estimation floue en une réponse claire, juste et transparente.

### Problématique : Le Prix Juste, Un Défi Quotidien

Imaginez :

* Un père de famille qui veut vendre sa voiture pour financer un nouveau véhicule, mais ne sait pas à quel prix la proposer.
* Un jeune acteur qui cherche sa première voiture et craint de se faire arnaquer.
* Un concessionnaire qui passe des heures à évaluer manuellement des dizaines de véhicules.

Tous ont un point commun : ils méritent une solution fiable et accessible. Notre défi ? Créer un outil qui comprend leurs besoins et leur offre une estimation précise, sans jargon technique ni approximations.

### Objectif Général : Simplifier la Vie, Une Estimation à la Fois

Notre ambition va au-delà des lignes de code. Nous voulons :  
🔹 Rendre l’estimation des prix accessible à tous, quel que soit le niveau de connaissance technique.  
🔹 Éliminer les doutes et permettre des décisions éclairées.  
🔹 Faire gagner du temps aux professionnels comme aux particuliers.

Bref, nous voulons mettre la technologie au service des gens.

### Objectifs Spécifiques : ✅ Écouter les besoins : Comprendre les frustrations des acheteurs, vendeurs et professionnels. ✅ Collecter les bonnes données : Des informations réelles, variées et représentatives du marché. ✅ Construire un modèle qui "parle humain" : Pas une boîte noire incompréhensible, mais un outil qui explique ses résultats. ✅ Tester, améliorer, répéter : Travailler avec de vrais utilisateurs pour affiner la solution. ✅ Livrer un outil simple et intuitif : Une interface claire, rapide et utile au quotidien.

### Méthodologie : 1️⃣ Immersion : Rencontrer des utilisateurs, comprendre leurs attentes. 2️⃣ Exploration des données : Identifier ce qui influence vraiment les prix (spoiler : ce n'est pas toujours ce qu'on croit !). 3️⃣ Prototypage rapide : Tester des idées, garder ce qui marche, jeter le reste. 4️⃣ Feedback continu : Intégrer les retours des futurs utilisateurs à chaque étape.

**L'idée maîtresse ?**Ce projet n'est pas juste un exercice technique. C'est une solution conçue par des humains, pour des humains. Parce qu’au final, derrière chaque dataset, il y a des personnes qui attendent une réponse simple à une question cruciale : *"Combien vaut vraiment ma voiture ?"*

**Prêt à embarquer dans cette aventure ? 🚗💨 La suite est encore plus passionnante !**

## 

## Partie II : Collecte de données

### 1. Description des sources de données

Pour ce projet, nous avons exploité deux principales plateformes de petites annonces au Maroc :

* **Avito.ma** : Leader des annonces en ligne, Avito propose une large sélection de véhicules d'occasion avec des détails précis (marque, modèle, kilométrage, prix, etc.).

### 2. Méthodologie de web scraping

#### **a. Outils et librairies utilisés**

Pour extraire les données efficacement, nous avons opté pour :

* **Python** : Langage flexible et puissant pour le scraping.
* **Requests & BeautifulSoup** :
  + Requests pour récupérer le contenu HTML des pages.
  + BeautifulSoup pour parser et extraire les données structurées.
* **CSV** : Stockage des données dans un format exploitable.
* **Random & Time** : Pour simuler un comportement humain et éviter le blocage.

#### **b. Sélecteurs et stratégie d'extraction**

* **Navigation paginée** : Parcours des pages de résultats (1 à 300) via l’URL dynamique (?page={num}).
* **Extraction des URLs des annonces** :
  + Balises HTML ciblées : <a class="sc-1jge648-0 jZXrfL"> pour les liens des voitures.
* **Récupération des détails** :
  + Sélecteurs CSS : Classes comme .sc-19cngu6-1 doRGIC pour les caractéristiques (marque, année, prix, etc.).
  + Nettoyage des données : Filtrage des valeurs nulles et formatage cohérent.

#### **c. Contournement des protections anti-scraping**

* **Rotation des User-Agents** : Simulation de différents navigateurs pour éviter la détection.
* **Délais aléatoires** : Pauses entre 2 et 7 secondes pour imiter une navigation humaine.
* **Gestion des erreurs** : Relèves en cas de timeout ou de requête bloquée.

### 3. Justification des choix technologiques

|  |  |  |  |
| --- | --- | --- | --- |
| **Technologie** | **Avantage** | **Alternative écartée** | **Pourquoi ce choix ?** |
| **Python** | **Richesse des librairies, simplicité** | **Node.js (Puppeteer)** | **Meilleur équilibre vitesse/facilité** |
| **BeautifulSoup** | **Léger et intuitif** | **Scrapy (trop complexe pour ce besoin)** | **Suffisant pour un parsing simple** |
| **CSV** | **Universalité, compatibilité** | **MongoDB/JSON** | **Facilité d’import dans Excel/Pandas** |
| **Requêtes avec pauses** | **vite le bannissement** | **Scraping massif parallélisé** | **Plus discret et éthique** |

### Impact & Perspectives

Cette méthode a permis de collecter **des milliers de données propres** en quelques heures, tout en respectant les limites des sites sources. Une prochaine étape pourrait intégrer :

* **Un proxy rotatif** pour augmenter le volume sans risque.
* **Un système de monitoring** pour détecter les changements de structure HTML.

*"Le web scraping, c’est comme une chasse au trésor moderne : il faut être à la fois patient, malin… et respectueux !"*

## 

## Partie III : Préparation des données

La phase de préparation des données est cruciale pour garantir la qualité et la cohérence des analyses. Nous avons appliqué plusieurs techniques de **nettoyage, encodage et normalisation** pour transformer les données brutes en un jeu exploitable.

### 1. Nettoyage : De la "mine" au "bijou"

Les données brutes sont comme des diamants non taillés : précieuses, mais inutilisables sans traitement.

#### **a. Suppression des doublons & colonnes non pertinentes**

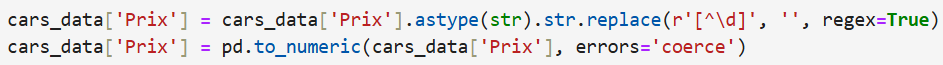
* **Colonnes retirées** :
  + Première main & Origine → Peu d’impact sur l’analyse (valeurs manquantes + faible corrélation).
* **Doublons** : Aucun détecté, mais vérification systématique avec cars\_data.duplicated().

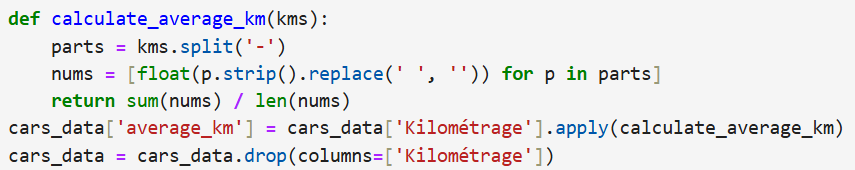
#### **b. Gestion des valeurs manquantes**

* **Suppression ciblée** :
  + Lignes sans **Prix**, **État** ou **Nombre de portes** → Critères essentiels pour l’analyse.
  + *Alternative envisagée* : Imputation par la **médiane**/**moyenne**, mais risque de **biais**.

#### 

#### **c. Nettoyage des formats**

* **Prix** :
  + Suppression des symboles (ex: "50 000 DH" → 50000) via **regex**.
  + Conversion en numérique (pd.to\_numeric).
* **Puissance fiscale** :
  + Suppression du "CV" (ex: "7 CV" → 7).
* **Kilométrage** :
  + Transformation des plages (ex: "10 000 - 20 000 km" → moyenne 15 000) via calculate\_average\_km().



*"Un dataset propre, c’est comme une voiture bien entretenue : ça roule mieux et plus loin !"* 🚗💨

### 2. Encodage des variables catégorielles

#### **a. Pourquoi encoder ?**

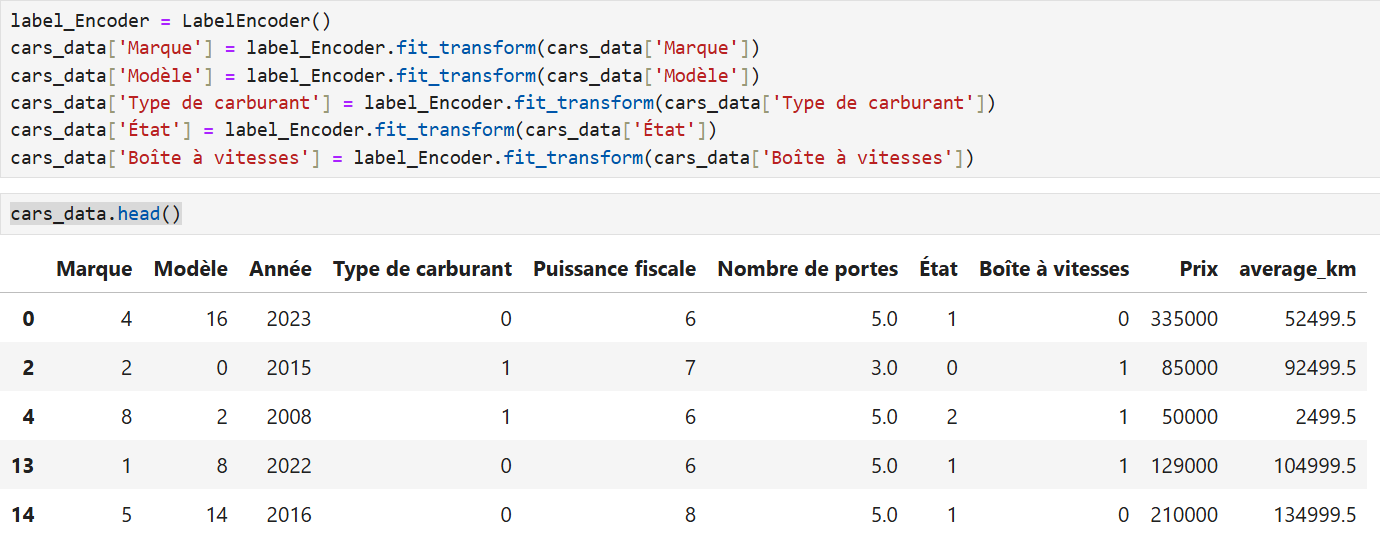
Les modèles ML ne comprennent que les nombres. Exemple :

* "Essence" → 0, "Diesel" → 1, etc.

#### **b. Méthode choisie : Label Encoding**

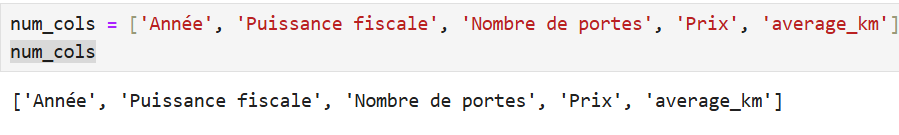
* **Fonctionnement** :
  + **LabelEncoder()** de **sklearn** attribue un nombre unique à chaque catégorie.
  + Appliqué sur : **Marque**, **Modèle**, **Type de carburant**, **État**, **Boîte à vitesses**.
* **Limite** :
  + Crée une fausse hiérarchie (ex: Peugeot=5 > BMW=1 ?).
  + *Alternative envisagée* : One-Hot Encoding (mais explosion du nombre de colonnes).

*"Transformer 'Renault' en 42 ? C’est un peu comme passer du français au binaire… mais ça marche !"* 🔢



### 3. Normalisation des variables numériques

#### **a. Standardisation (Z-score)**

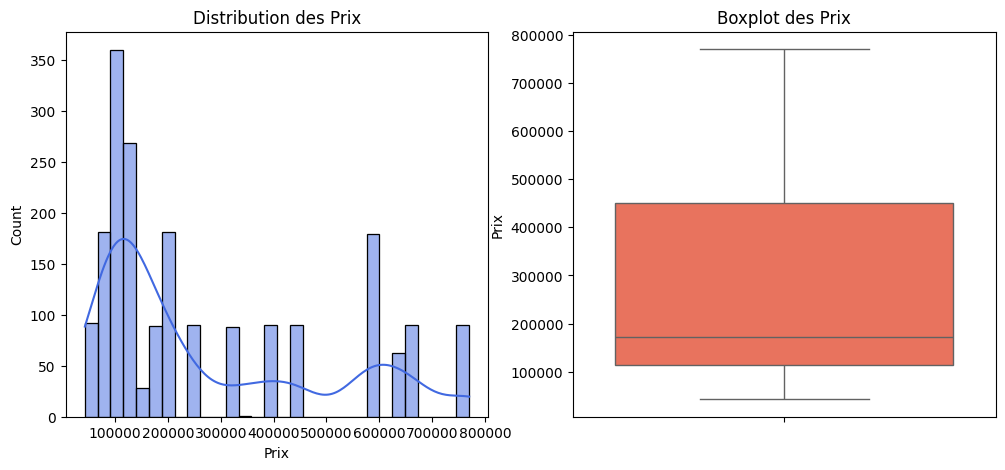
* **Librairie : StandardScaler() de sklearn.**
* **Colonnes cibles :**
* **Effet :**
  + Chaque variable a une moyenne = 0 et écart-type = 1.
  + Exemple : Un prix de 100 000 DH devient 1.2, un autre de 50 000 DH devient -0.3.

## Partie III : Analyse Exploratoire des Données (EDA)

L’analyse exploratoire des données a été réalisée afin de mieux comprendre la structure du dataset, la distribution des variables, ainsi que les relations potentielles entre les caractéristiques et le prix des véhicules. Cette étape nous a également permis de détecter des valeurs aberrantes et d’orienter les étapes suivantes de la modélisation.

### 1. Distribution des variables clés

**1. Histogramme et Boxplot de Distribution des Prix**



### L'histogramme révèle une distribution asymétrique des prix avec les caractéristiques suivantes :

### Plage de valeurs : Les prix s'étendent de 50 000 à 800 000 DH avec une concentration marquée entre 100 000 et 300 000.

### Pic de fréquence : Le nombre maximal de véhicules (350) se situe dans la tranche basse (100 000–150 000), indiquant un marché dominé par les véhicules d'entrée/moyenne gamme.

### Queue longue : Une décroissance progressive vers les prix élevés suggère la présence de véhicules premium/collection.

### Interprétation : La distribution est fortement asymétrique à droite, typique des données économiques. Une transformation logarithme pourrait normaliser cette distribution pour les modèles statistiques.

### Le boxplot révèle une distribution des prix avec :

### Médiane : presque 200000 DH

### Intervalle interquartile : 100 000 à 450 000

### Valeurs extrêmes : Nombreux points au-delà de 600 000

### Interprétation Statistique

### La distribution présente :

### Une asymétrie positive marquée (queue de distribution étendue vers les hautes valeurs)

### Une concentration des données dans la moitié inférieure de la plage

### Une présence significative d'outliers dans le segment premium

### Analyse multivariée des voitures selon le type de carburant

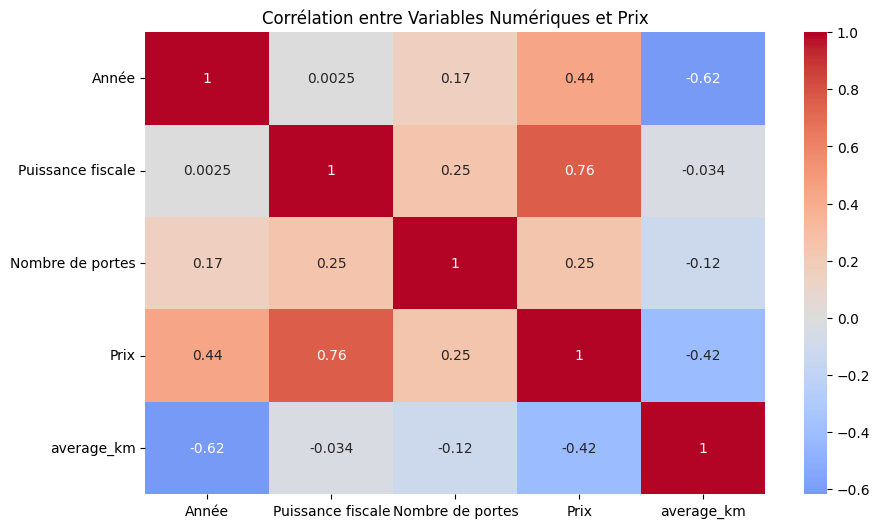
**Description :**  
Ce graphique en matrice montre les relations entre plusieurs variables : Année, Puissance fiscale, Kilométrage moyen , et Prix, en fonction du type de carburant (Diesel en bleu, Essence en orange). Les histogrammes diagonaux représentent la distribution des valeurs.

**Interprétation rapide :**

* **Année vs Prix :** Les voitures récentes sont généralement plus chères. Diesel domine dans les modèles plus récents.
* **Kilométrage :** Les véhicules Diesel ont souvent plus de kilomètres au compteur.
* **Puissance fiscale :** Les Diesel tendent à avoir une puissance fiscale légèrement plus élevée.
* **Prix :** Les Diesel couvrent une plus large gamme de prix, souvent plus élevés.

**Conclusion :**  
Les véhicules Diesel semblent plus présents, plus puissants, plus utilisés et plus chers. L’essence reste marginale, souvent sur des véhicules plus anciens et moins coûteux.

**Corrélation entre les variables numériques et le prix**

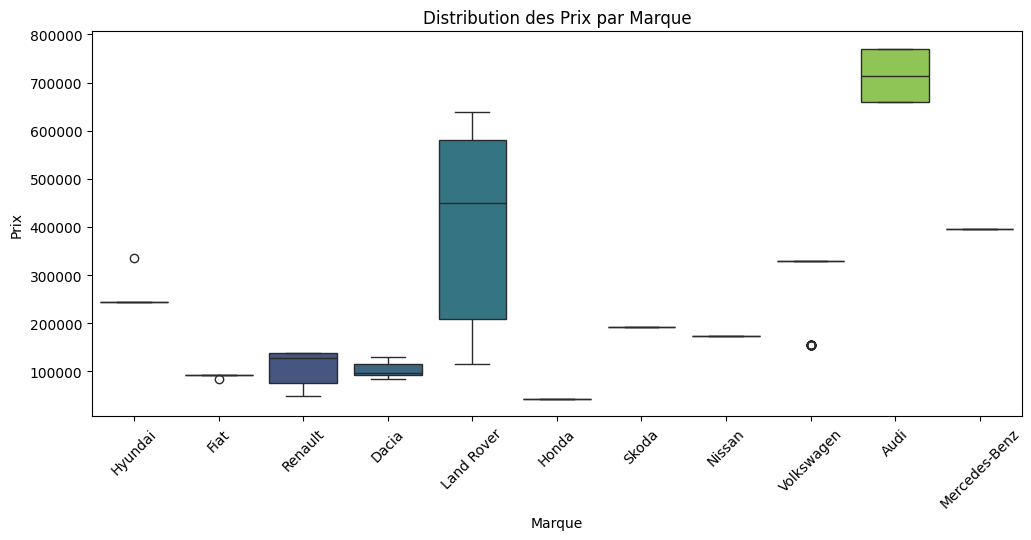
**Description :**  
La heatmap montre les coefficients de corrélation entre les variables numériques (Année, Puissance fiscale, Nombre de portes, Kilométrage moyen, Prix). Les valeurs vont de -1 (corrélation négative forte) à +1 (corrélation positive forte).

**Interprétation rapide :**

* **Prix** est fortement corrélé à **Puissance fiscale** (0.76) et modérément à **Année** (0.44) → Les voitures récentes et puissantes sont plus chères.
* **Prix** est négativement corrélé avec **Kilométrage moyen** (-0.42) → Plus une voiture est kilométrée, moins elle coûte cher.
* **Année** et **average\_km** sont aussi fortement négativement corrélés (-0.62) → Les véhicules récents ont moins de kilomètres.

**Conclusion :**  
Le prix est principalement influencé par la puissance fiscale, l'année de mise en circulation, et le kilométrage. Ces trois indicateurs sont essentiels pour estimer la valeur d'un véhicule.

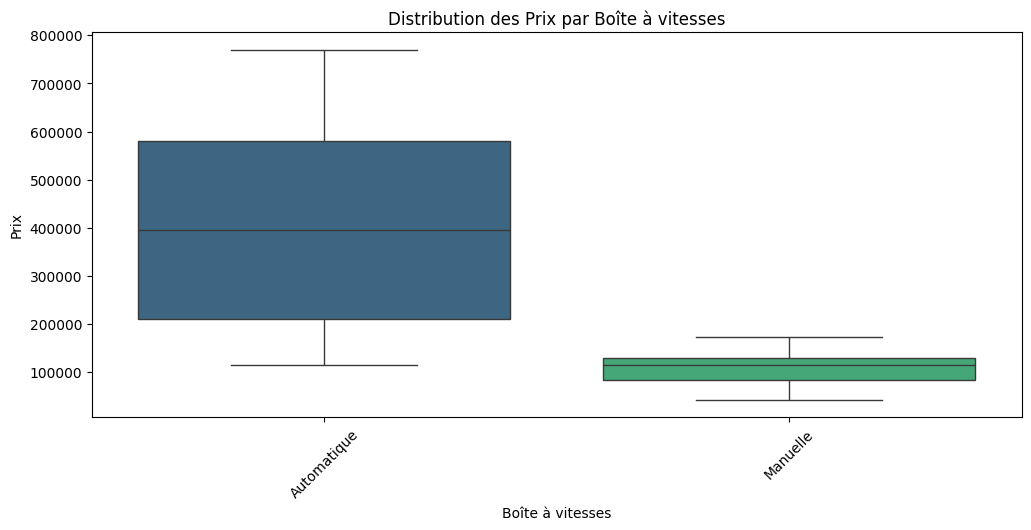
**Distribution des Prix par Marque**

**Description :**  
Ce boxplot présente la distribution des prix de vente pour chaque marque de voiture. On y voit la médiane, l’étendue des prix, ainsi que d’éventuelles valeurs aberrantes

**Interprétation rapide :**

* **Audi** et **Land Rover** ont les prix les plus élevés et les plus dispersés → Gamme premium.
* **Mercedes-Benz** reste aussi dans le haut de gamme, mais avec moins de variation.
* **Fiat**, **Honda**, et **Dacia** montrent des prix beaucoup plus bas → Positionnement économique.
* **Renault**, **Volkswagen**, et **Hyundai** sont dans la gamme moyenne avec quelques valeurs élevées.

**Conclusion :**  
Le prix dépend fortement de la marque. Les marques premium présentent non seulement des prix plus élevés mais aussi une plus grande variabilité, ce qui reflète des gammes de modèles plus larges.

**Distribution des Prix par Boîte à Vitesse**

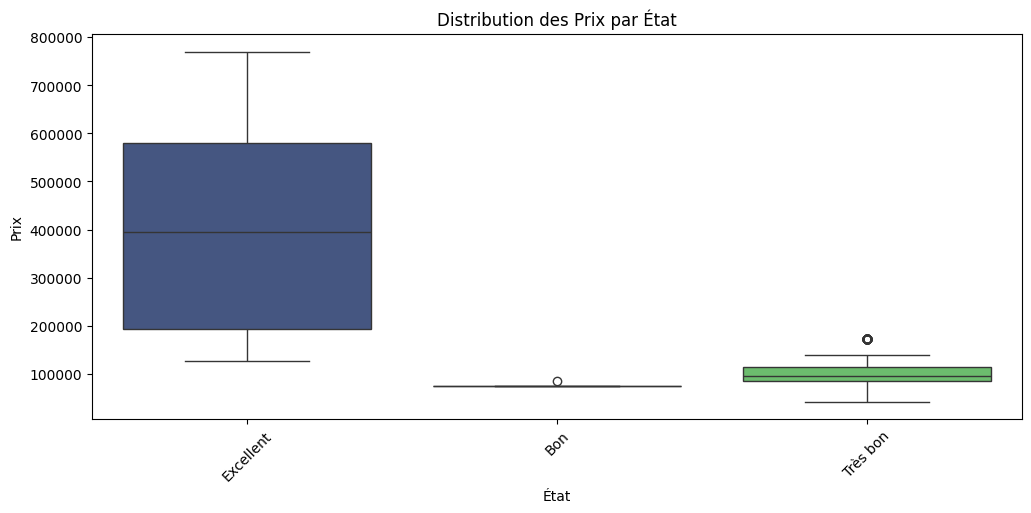
**Description :**  
Ce boxplot compare les prix des voitures selon le type de boîte à vitesses (Automatique vs Manuelle).

**Interprétation rapide :**

* Les voitures **automatiques** ont une médiane de prix nettement plus élevée, avec une grande dispersion → souvent associées à des modèles haut de gamme.
* Les voitures **manuelles** sont globalement moins chères et plus homogènes en prix.

**Conclusion :**  
Le type de boîte à vitesses est un facteur important dans la formation du prix. L'automatique est typique des véhicules modernes et coûteux, tandis que la manuelle reste une option plus économique.

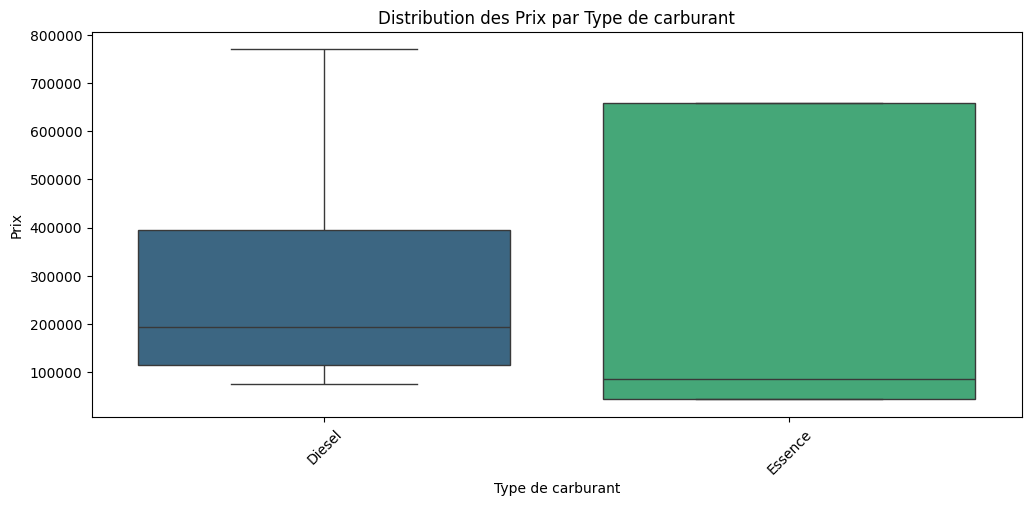
### Distribution des Prix par État

**Observation** :

* Les véhicules en **état "Excellent"** ont les prix les plus élevés et la plus grande dispersion.
* L’état **"Très bon"** a des prix plus bas et relativement concentrés.
* L’état **"Bon"** montre des prix faibles et très peu de variation.

**Interprétation** :  
L’état du véhicule influence fortement le prix. Les voitures en excellent état sont significativement plus chères, tandis que celles en bon état sont vendues à un prix bien inférieur, avec peu de variation.

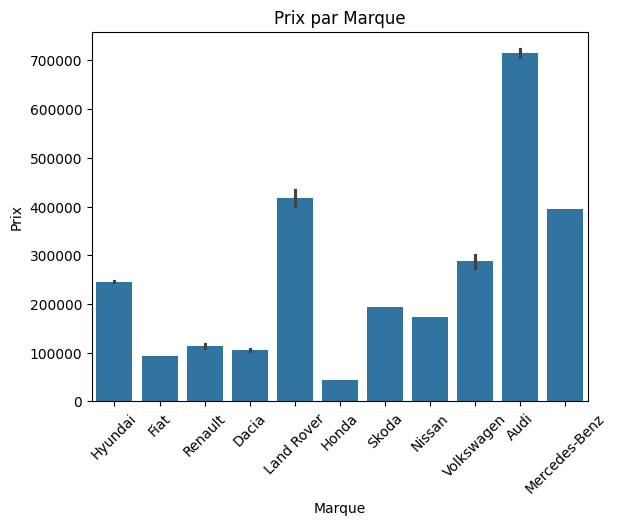
### Distribution des Prix par Type de Carburant

**Observation** :

* Les voitures **Diesel** présentent une large gamme de prix, souvent plus élevés.
* Les **Essence** ont des prix généralement plus bas, avec moins de dispersion.

**Interprétation** :  
Les véhicules Diesel tendent à être plus chers que ceux à essence, ce qui peut être lié à des performances ou à la durabilité supérieures perçues.

### Prix par Marque



**Observation** :

* **Audi** est la marque la plus chère en moyenne, suivie de **Land Rover** et **Mercedes-Benz**.
* Les marques les moins chères sont **Honda**, **Fiat**, et **Dacia**.
* Les autres marques comme **Volkswagen**, **Hyundai**, et **Skoda** se situent dans une gamme intermédiaire.

**Interprétation** :  
Le prix moyen varie fortement selon la marque. Les marques premium (comme Audi, Mercedes) affichent des prix élevés, tandis que les marques plus généralistes (Fiat, Honda) restent abordables. Cela reflète la position de la marque sur le marché (luxe vs. économique).