

**Projet de Fin de Module :**

**Développement d’un modèle de prédiction des prix de voitures pour le marché marocain.** 

**Réalisé par : Encadré par :**

**Mouad Radouani & Abdelkarim Narjiss Mme Laila ELJIANI**

**Date : Avril 2025**

# Introduction Générale

L’intelligence artificielle et le machine learning connaissent aujourd’hui une croissance fulgurante et transforment progressivement de nombreux secteurs, de la santé à la finance, en passant par le commerce, la logistique ou encore l’automobile. Dans ce contexte, les données deviennent un actif stratégique permettant d’automatiser, d’optimiser et de prédire des phénomènes complexes avec une précision inégalée. Le présent projet de fin de module (PFM) s’inscrit dans cette dynamique, en appliquant les méthodes d’apprentissage automatique à un domaine concret et proche du quotidien : **la prédiction des prix des voitures d’occasion**.

Le marché de l’automobile d’occasion, en particulier au Maroc, est un marché en pleine expansion. Il est caractérisé par une forte hétérogénéité des offres, une grande diversité de marques et de modèles, ainsi qu’une forte subjectivité dans la fixation des prix. Cette absence de standardisation rend les décisions d’achat ou de vente complexes, parfois injustes, et souvent inefficaces. Dès lors, une question essentielle se pose : **peut-on estimer de manière fiable et automatisée le prix d’un véhicule d’occasion à partir de ses caractéristiques techniques ?**

C’est à cette problématique que ce projet ambitionne de répondre. En combinant la puissance des algorithmes de machine learning à l’exploitation de données issues de plateformes d’annonces automobiles, nous visons à concevoir un **modèle prédictif performant, robuste et interprétable**, capable d’offrir une estimation réaliste et instantanée du prix d’un véhicule. Ce projet, au-delà de sa dimension technique, représente un cas d’usage emblématique des apports de la science des données dans la résolution de problèmes concrets, à forte valeur ajoutée pour les citoyens comme pour les professionnels.

Ce rapport présente ainsi, étape par étape, le déroulement du projet : de la compréhension du besoin métier à la collecte des données, en passant par la modélisation, l’évaluation et les perspectives d’intégration. Il s’inscrit dans une démarche rigoureuse, fondée sur la méthode scientifique, tout en restant ancré dans une finalité pratique et accessible.

Table des matières

[Introduction Générale 1](#_Toc195213773)

[Partie I : Présentation Générale du Projet 4](#_Toc195213774)

[1. Contexte : Une Révolution Silencieuse dans le Marché de l’Automobile 4](#_Toc195213775)

[2. Problématique : Comment Évaluer Justement un Véhicule d’Occasion ? 5](#_Toc195213776)

[3. Objectif Général : Mettre l’IA au Service d’une Estimation Juste et Transparente 6](#_Toc195213777)

[4. Objectifs Spécifiques : Décomposer pour Mieux Construire 6](#_Toc195213778)

[5. Méthodologie Globale : Une Approche Itérative et Centrée sur les Données 7](#_Toc195213779)

[Partie II : Collecte de données 8](#_Toc195213780)

[1. Description de la source de données 8](#_Toc195213781)

[2. Méthodologie de web scraping 8](#_Toc195213782)

[3. Justification des choix technologiques 9](#_Toc195213783)

[Partie III : Préparation des données 10](#_Toc195213784)

[1. Nettoyage des données : de l'extraction brute à un corpus exploitable 10](#_Toc195213785)

[2. Encodage des variables catégorielles 12](#_Toc195213786)

[3. Normalisation des variables numériques 13](#_Toc195213787)

[Partie IV : Analyse Exploratoire des Données (EDA) 14](#_Toc195213788)

[1. Distribution des variables clés 14](#_Toc195213789)

[2. Visualisations multivariées 15](#_Toc195213790)

[3. Corrélation entre variables numériques 16](#_Toc195213791)

[4. Étude du prix selon des variables catégorielles 17](#_Toc195213792)

# Partie I : Présentation Générale du Projet

## 1. Contexte : Une Révolution Silencieuse dans le Marché de l’Automobile

Le marché de l’automobile d’occasion connaît, depuis plusieurs années, une croissance soutenue au Maroc comme à l’échelle internationale. Cette dynamique est alimentée par de nombreux facteurs : la pression inflationniste sur les véhicules neufs, l’évolution des modes de consommation, et l’accessibilité accrue des plateformes de vente en ligne. Pourtant, malgré cette expansion, l’une des problématiques majeures persiste : **l’incertitude liée à l’évaluation du prix d’un véhicule d’occasion**.

En effet, les acheteurs comme les vendeurs se trouvent confrontés à un manque de transparence et de standardisation dans la fixation des prix. Les évaluations sont souvent empiriques, basées sur l’intuition, l’expérience personnelle ou la comparaison aléatoire d'annonces. Cette subjectivité engendre non seulement des transactions déséquilibrées, mais alimente également une méfiance croissante dans le processus d’achat/vente.

Dans ce contexte, **l’émergence de solutions technologiques basées sur l’intelligence artificielle (IA)** représente une opportunité stratégique pour repenser les mécanismes d’évaluation. L’exploitation des données massives disponibles sur les plateformes de petites annonces, couplée à la puissance prédictive du Machine Learning, ouvre la voie à une estimation des prix **plus rigoureuse, objective et fiable**.

Ce projet de fin de module (PFM) s’inscrit dans cette logique d’innovation en proposant **un modèle prédictif des prix des voitures d’occasion basé sur l’apprentissage automatique**, visant à apporter une réponse concrète, accessible et performante à une problématique réelle du marché.

## 2. Problématique : Comment Évaluer Justement un Véhicule d’Occasion ?

La question de l’estimation du prix d’un véhicule d’occasion soulève plusieurs problématiques interdépendantes :

* **Variabilité des critères d’influence** : Le prix dépend d’une multitude de facteurs – marque, modèle, année, kilométrage, état, type de carburant, boîte de vitesses, etc. – dont l’impact peut varier considérablement selon les tendances du marché.
* **Manque de standardisation** : Chaque vendeur fixe son prix selon ses propres critères. Les bases de données officielles sont limitées ou obsolètes, et les plateformes ne fournissent pas d’outil analytique avancé.
* **Risques pour les acteurs** :
  + L’acheteur peut surpayer un véhicule sous-évalué.
  + Le vendeur peut sous-évaluer son bien par manque de connaissance.
  + Les professionnels (concessionnaires, garages, experts) perdent un temps considérable dans l’évaluation manuelle.

Face à ces défis, une question centrale émerge :  
 **Peut-on concevoir un système intelligent capable d’estimer le prix d’un véhicule d’occasion de manière précise, rapide et automatisée, à partir de ses caractéristiques techniques ?**

## 3. Objectif Général : Mettre l’IA au Service d’une Estimation Juste et Transparente

L’objectif principal du projet est de développer un outil prédictif fiable et interprétable permettant d’estimer le prix de vente d’un véhicule d’occasion au Maroc à partir de ses attributs techniques et commerciaux.

Ce modèle vise à:

* Réduire l’incertitude liée à l’évaluation manuelle.
* Offrir une aide à la décision aux acheteurs, vendeurs et professionnels du secteur.
* Démontrer la pertinence de l’intelligence artificielle dans un contexte concret et applicable.

## 4. Objectifs Spécifiques : Décomposer pour Mieux Construire

Pour atteindre cet objectif général, le projet est structuré autour d’objectifs spécifiques et opérationnels :

* **Compréhension métier**:
  + Étudier les dynamiques du marché marocain de l’automobile.
  + Identifier les variables influentes et les comportements des acteurs.
* **Collecte et prétraitement des données** :
  + Extraire des milliers d’annonces réelles depuis des plateformes comme Avito.ma.
  + Nettoyer, filtrer, structurer les données pour assurer leur qualité analytique.
* **Conception et entraînement du modèle** :
  + Choisir les algorithmes de régression les plus pertinents (Linear Regression, Random Forest, XGBoost, etc.).
  + Optimiser les performances du modèle à travers un processus itératif (cross-validation, hyperparameter tuning…).
* **Évaluation des performances** :
  + Utiliser des métriques pertinentes (RMSE, MAE, R²) pour juger de la précision du modèle.
* **Déploiement et accessibilité** :
  + Intégrer le modèle dans une interface simplifiée, orientée utilisateur.
  + Permettre une prédiction en temps réel sur la base d’un formulaire.

## 5. Méthodologie Globale : Une Approche Itérative et Centrée sur les Données

La méthodologie adoptée dans ce projet repose sur une **approche empirique, progressive et centrée sur l’expérience utilisateur**. Elle se décline en cinq étapes clés :

### a. Phase exploratoire (compréhension métier & utilisateurs)

* Analyse des pratiques d’évaluation existantes.
* Enquête qualitative auprès d’utilisateurs réels (vendeurs, acheteurs).

### b. Collecte des données (Web Scraping)

* Utilisation de Python avec requests et BeautifulSoup.
* Extraction massive d’annonces automobiles structurées.

### c. Prétraitement et préparation des données

* Nettoyage, traitement des valeurs manquantes, standardisation des formats.
* Encodage des variables catégorielles et normalisation des données numériques.

### d. Modélisation prédictive

* Comparaison de plusieurs algorithmes supervisés.
* Validation croisée et sélection du modèle optimal.

### e. Évaluation & déploiement

* Visualisation des résultats, tests de robustesse, intégration dans une interface simple.

# Partie II : Collecte de données

## 1. Description de la source de données

Pour constituer notre base de données, nous avons ciblé la plateforme **Avito.ma**, un site de petites annonces en ligne largement utilisé au Maroc, notamment pour la vente de véhicules d’occasion. Ce choix s’explique par plusieurs facteurs :

* **Richesse des annonces** : Avito.ma propose une vaste sélection de voitures avec des informations détaillées (marque, modèle, année, kilométrage, prix, carburant, etc.).
* **Fréquence de mise à jour** : De nouvelles annonces sont publiées quotidiennement, assurant ainsi une base de données représentative du marché en temps réel.
* **Accessibilité publique** : Les informations affichées sur les annonces sont librement accessibles, ce qui facilite la collecte de données via le web scraping.

Nous avons volontairement limité notre étude à Avito.ma afin de garantir la cohérence des données collectées et de se concentrer sur la qualité du traitement plutôt que sur la fusion de sources hétérogènes.

## 2. Méthodologie de web scraping

La collecte de données a été réalisée à l’aide d’un script Python développé spécifiquement pour ce projet. Le processus de scraping a été structuré autour de deux grandes étapes :

* **Extraction des liens d'annonces** : Le script commence par parcourir un ensemble de pages de résultats sur Avito.ma afin de collecter les URL individuelles des annonces automobiles.
* **Extraction des informations détaillées** : Pour chaque lien collecté, une requête est envoyée pour extraire les caractéristiques du véhicule. Parmi les données récupérées figurent : la marque, le modèle, l’année de mise en circulation, le type de carburant, le kilométrage, la puissance fiscale, le nombre de portes, la boîte à vitesses, le prix affiché, etc.

Le scraping a été encadré par plusieurs bonnes pratiques :

* **Rotation des agents utilisateurs (user-agents)** : Afin d’éviter d’être bloqué par les mécanismes anti-bot du site cible, nous avons intégré une rotation aléatoire de plusieurs user-agents (simulateurs de navigateurs).
* **Temporisation entre les requêtes** : Des délais aléatoires ont été insérés entre chaque requête pour simuler un comportement humain et réduire les risques de détection.
* **Gestion des erreurs et des pages vides** : Le script inclut des mécanismes de vérification de statut des réponses, de gestion des exceptions réseau, ainsi que de filtrage des pages incomplètes ou inaccessibles.

Les données ont été stockées dans un fichier CSV structuré, avec un enregistrement effectué à chaque itération afin de limiter les pertes potentielles en cas d’interruption du processus.

## 3. Justification des choix technologiques

Les outils et bibliothèques suivants ont été utilisés pour leur efficacité et leur pertinence dans le contexte du web scraping :

* **Python** : Langage de programmation flexible et bien adapté aux tâches d’automatisation et de traitement de données.
* **Requests** : Pour l’envoi des requêtes HTTP, permettant une récupération fluide du contenu HTML des pages.
* **BeautifulSoup** : Pour l’analyse et l’extraction des données depuis le code HTML de manière rapide et structurée.
* **CSV** : Format de stockage simple et facilement exploitable pour les traitements ultérieurs (nettoyage, analyse, modélisation).
* **Random et Time** : Pour la gestion du comportement simulé d’un utilisateur humain (delais entre les requêtes, rotation d’headers).

Ce choix d’outils a permis de garantir un processus de collecte robuste, scalable et reproductible, tout en minimisant le risque d’interruption ou de blocage. Il a également facilité l’intégration des données dans les étapes suivantes du projet (prétraitement, modélisation, etc.)

# Partie III : Préparation des données

La préparation des données constitue une étape fondamentale du processus d’apprentissage automatique. Elle vise à transformer les données brutes issues du scraping en un ensemble cohérent, propre et directement exploitable par les algorithmes de modélisation. Cette phase repose sur trois piliers essentiels : le **nettoyage**, l’**encodage** des variables catégorielles et la **normalisation** des variables numériques.

## 1. Nettoyage des données : de l'extraction brute à un corpus exploitable

Les données issues du web scraping sont souvent hétérogènes et incomplètes. Ainsi, un processus rigoureux de nettoyage a été mis en place afin d’éliminer les biais, réduire les sources d’erreur et améliorer la qualité statistique du jeu de données.

### a. Suppression des colonnes non pertinentes et des doublons

Dans un souci de pertinence analytique, certaines colonnes ont été supprimées. Il s'agit notamment des champs **"Première main"** et **"Origine"**, dont les données étaient soit largement manquantes, soit peu corrélées aux variables cibles, réduisant ainsi leur utilité dans le cadre de la prédiction des prix.

Concernant les doublons, une vérification systématique a été effectuée à l’aide d’un processus de détection automatisé. Aucune duplication n’a été détectée dans l’ensemble de données, garantissant l’unicité des observations.

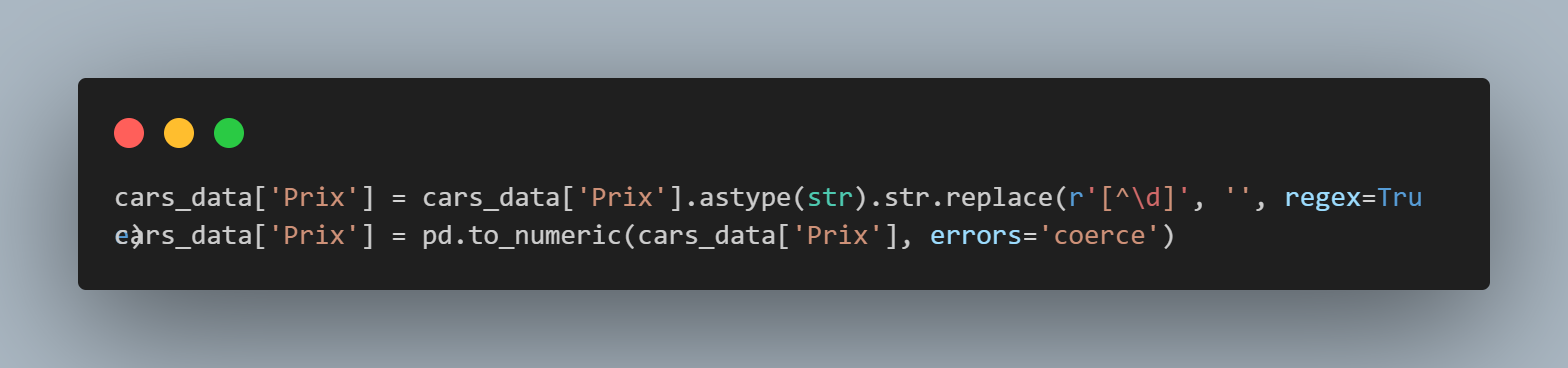
### b. Gestion des valeurs manquantes

Les valeurs manquantes ont été traitées de manière stratégique, en fonction de leur criticité pour l’analyse :

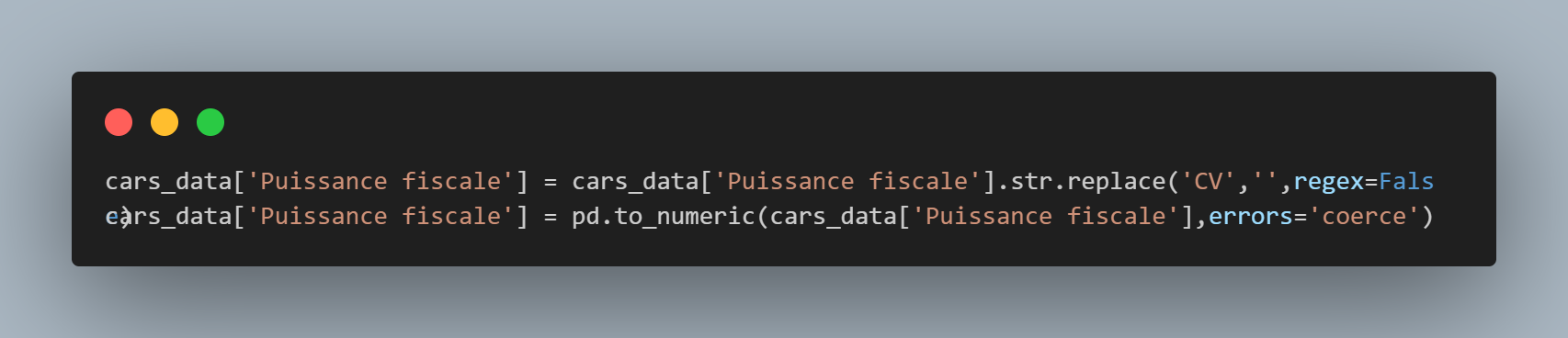
* Les observations ne comportant pas de valeurs pour des variables clés comme le **prix**, l’**état** ou le **nombre de portes** ont été éliminées afin de préserver la fiabilité des analyses futures.
* Des alternatives comme l’imputation par la moyenne ou la médiane ont été envisagées, mais écartées pour éviter l’introduction d’un biais statistique non négligeable.

### c. Harmonisation des formats

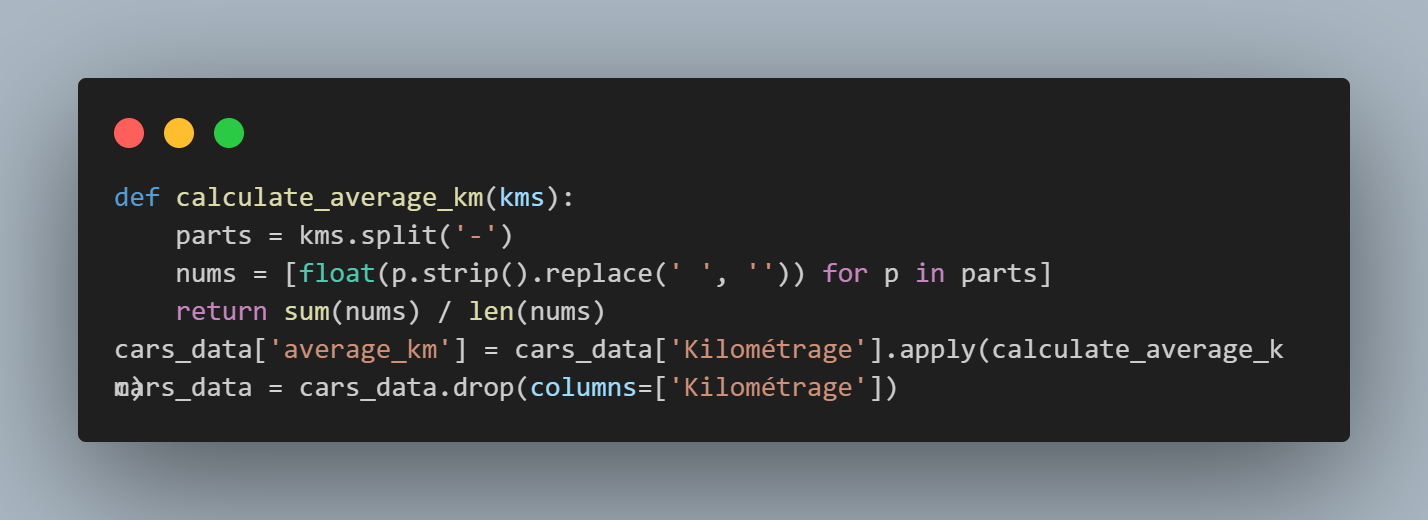
Plusieurs champs contenaient des chaînes de caractères nécessitant une transformation pour permettre leur exploitation numérique :



* Le champ **prix** a été nettoyé par suppression des symboles monétaires et des espaces, puis converti en entier.



* La **puissance fiscale** a été convertie en valeur numérique après suppression de l’unité "CV".



* Le **kilométrage**, souvent exprimé sous forme d’intervalles (par exemple "10 000 - 20 000 km"), a été transformé en une valeur moyenne représentative afin de conserver l’information tout en la rendant exploitable.

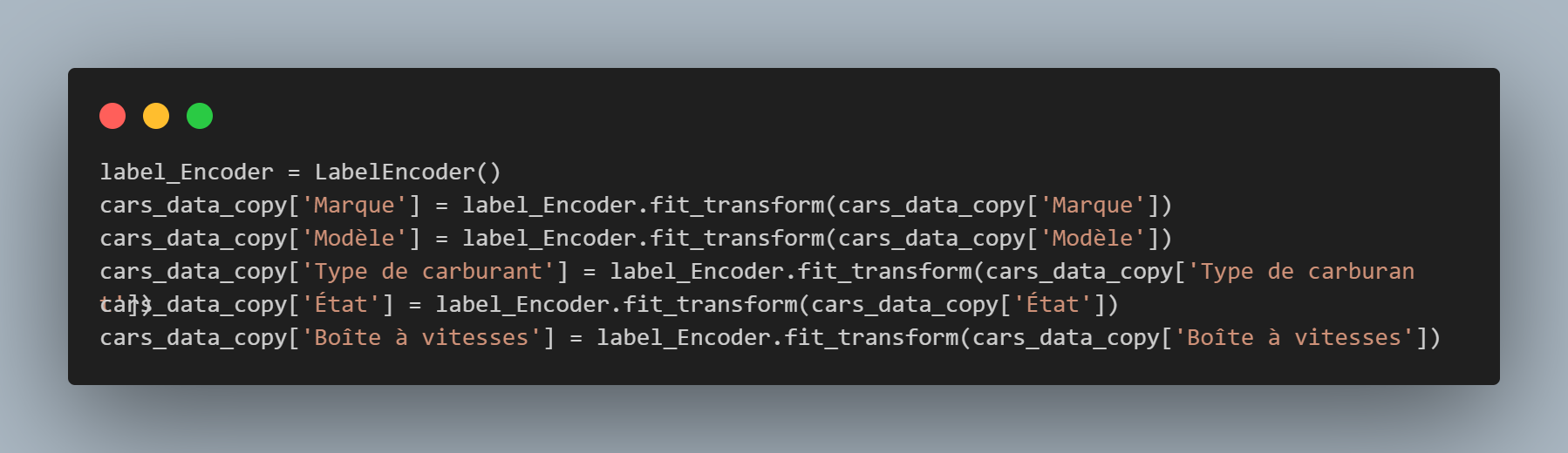
## 2. Encodage des variables catégorielles

Les algorithmes de machine learning nécessitent que toutes les données soient exprimées sous forme numérique. Il a donc été indispensable d’encoder les variables qualitatives présentes dans le jeu de données.

### a. Motivation de l'encodage

Des variables telles que la **marque**, le **modèle**, le **type de carburant**, l’**état du véhicule** ou encore la **boîte à vitesses** sont de nature catégorielle. Sans encodage adéquat, elles seraient inexploitables par les algorithmes d’apprentissage supervisé.

### b. Méthode utilisée : Encodage par étiquetage (Label Encoding)

****

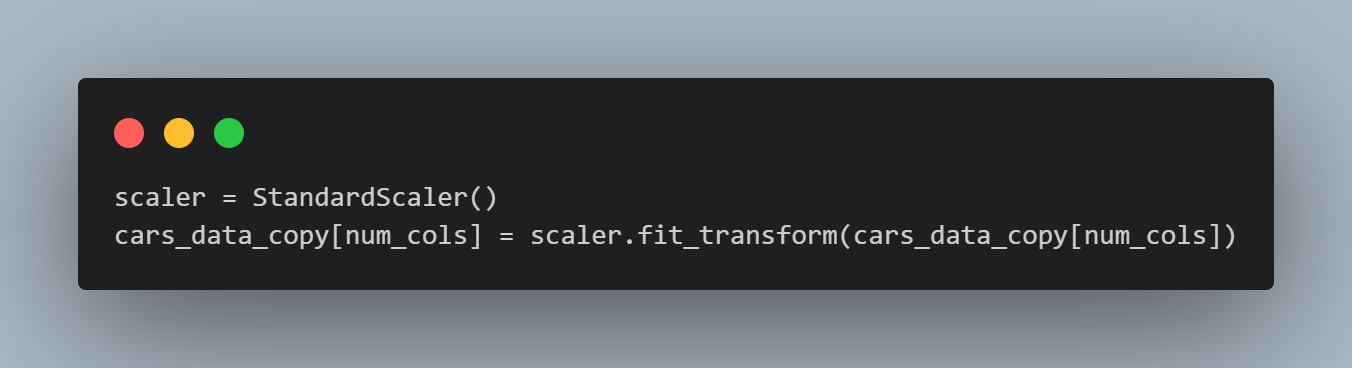
L’approche retenue repose sur l’utilisation de la méthode d’**encodage par étiquetage**, qui attribue un entier unique à chaque modalité présente dans la variable. Cette technique a été appliquée à l’ensemble des variables catégorielles susmentionnées.

Toutefois, ce type d'encodage introduit une hiérarchisation implicite entre les catégories (ex. : "Peugeot" = 5 > "BMW" = 1), qui ne reflète pas une réalité sémantique. Bien que le **One-Hot Encoding** ait été envisagé pour éviter ce biais, il aurait considérablement accru la dimensionnalité du jeu de données, compromettant la performance de certains modèles. Le **Label Encoding** a donc été retenu comme compromis entre simplicité et efficacité.

## 3. Normalisation des variables numériques

Les variables numériques présentent souvent des échelles de valeurs très différentes (ex. : le prix d’une voiture peut varier de 30 000 à plus de 500 000 DH, tandis que la puissance fiscale varie généralement entre 5 et 12). Une mise à l’échelle est donc essentielle pour éviter que certaines variables ne dominent les autres lors de la modélisation.

### a. Standardisation (Z-score)

****

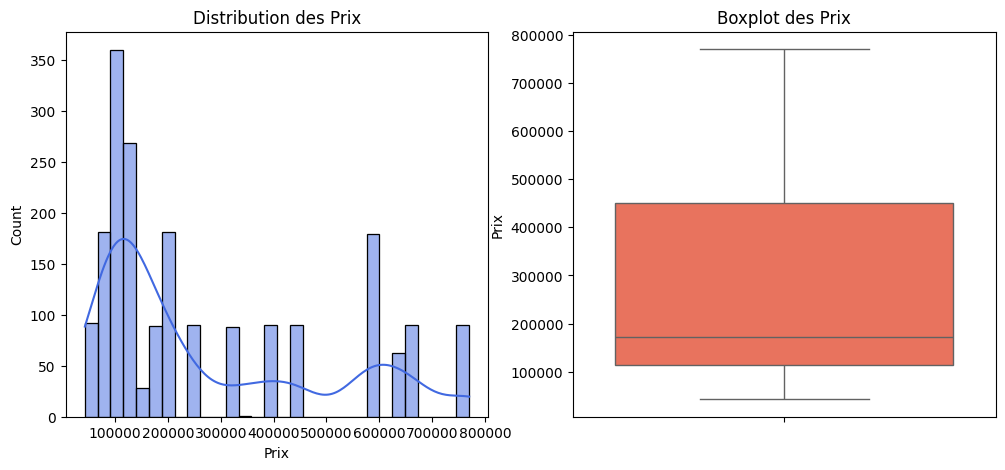
La méthode de **standardisation** a été choisie pour homogénéiser les variables numériques. Cette technique consiste à recentrer les données autour de la moyenne et à les diviser par leur écart-type, de manière à obtenir une distribution normalisée avec une moyenne nulle et une variance unitaire.

# Partie IV : Analyse Exploratoire des Données (EDA)

L’analyse exploratoire des données (Exploratory Data Analysis - EDA) constitue une étape fondamentale dans tout projet de machine learning. Elle permet d’identifier les tendances générales, d’évaluer la qualité des données, de détecter d’éventuelles anomalies, et d’étudier les relations entre les différentes variables du jeu de données. Dans le cadre de notre projet, cette phase a servi à mieux comprendre les dynamiques du marché automobile marocain tel qu’il est représenté à travers les données collectées depuis Avito.ma.

## 1. Distribution des variables clés

### a. Distribution des prix

****

L’analyse de la variable cible « Prix » a révélé une distribution fortement asymétrique, orientée à droite. La majorité des véhicules se situent dans une fourchette de prix comprise entre **100 000 et 300 000 dirhams**, avec un pic notable entre **100 000 et 150 000 dirhams**, reflétant une dominance des véhicules d’entrée et de moyenne gamme. Cependant, une queue étendue vers les valeurs plus élevées indique la présence de modèles premium ou de collection, allant jusqu’à **800 000 dirhams**.

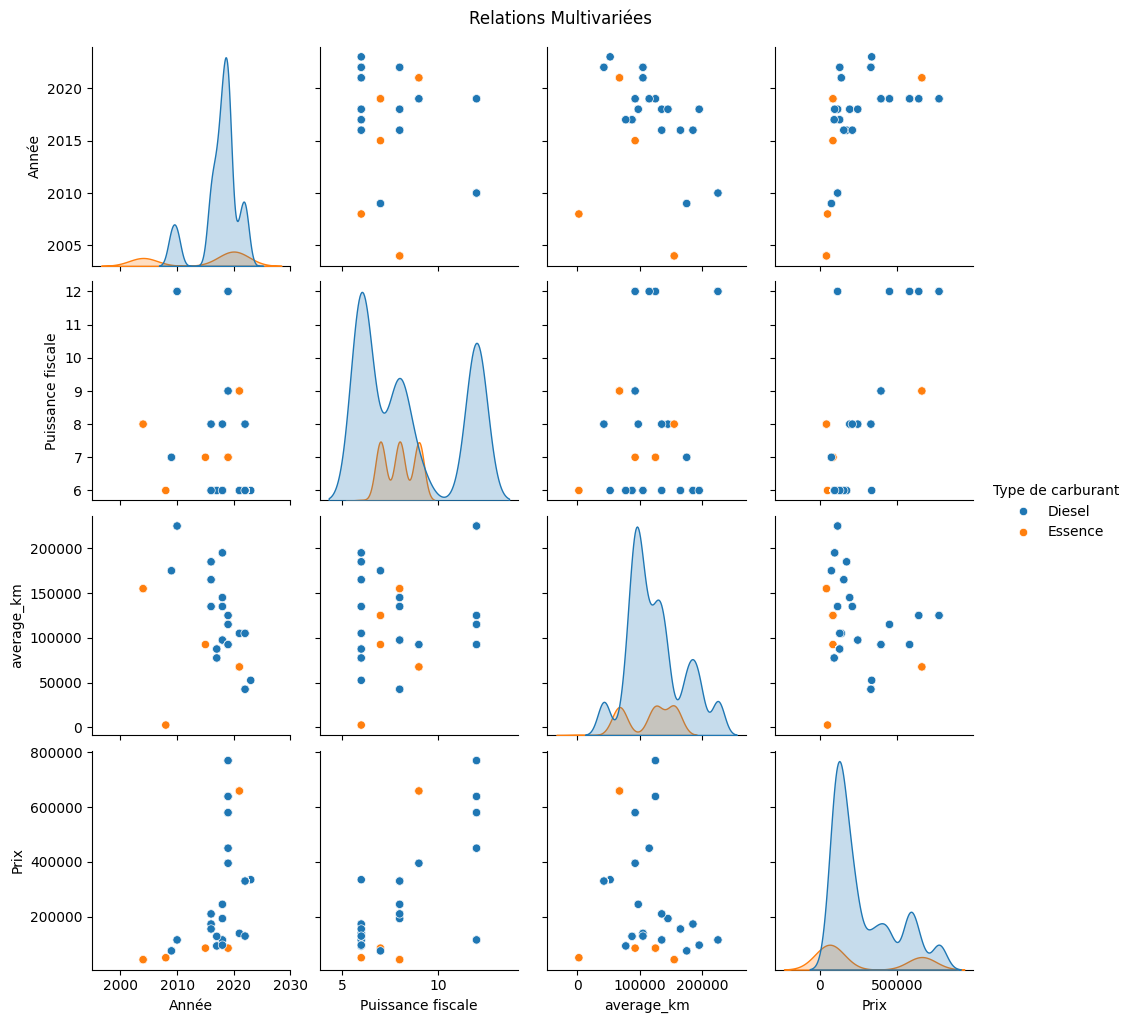
Un **boxplot** a permis de visualiser la dispersion des prix :

* **Médiane** aux alentours de **200 000 DH**
* **Intervalle interquartile** situé entre **100 000 et 450 000 DH**
* **Présence significative d’outliers**, notamment au-delà de **600 000 DH**

Une telle distribution suggère une asymétrie positive caractéristique des données économiques. Une **transformation logarithmique** pourrait s’avérer pertinente lors de l’entraînement des modèles afin de réduire la variance et stabiliser la distribution.

## 2. Visualisations multivariées

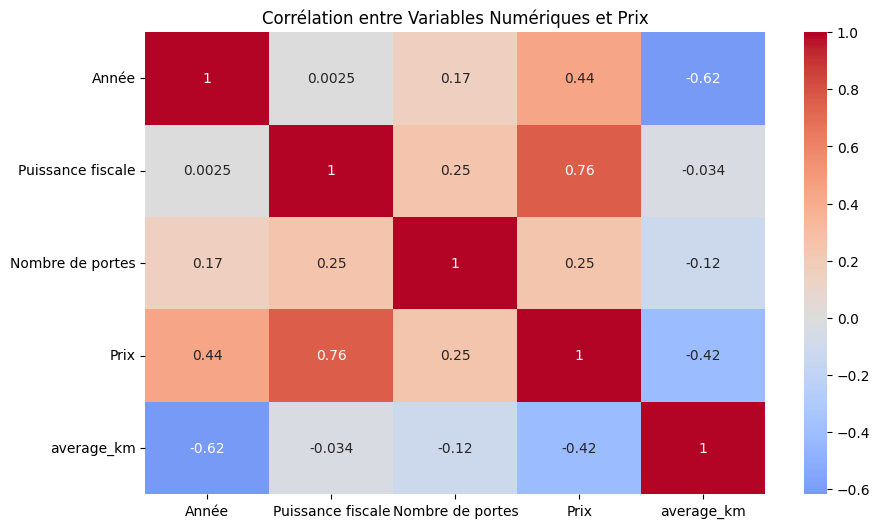
### a. Analyse croisée : Type de carburant vs. autres variables

****

Une visualisation matricielle a été employée pour représenter les interactions entre **année**, **puissance fiscale**, **kilométrage** et **prix**, selon le type de **carburant** :

* Les véhicules **Diesel** dominent dans les modèles récents et montrent un **kilométrage plus élevé**, suggérant une utilisation plus intensive ou une meilleure longévité.
* La **puissance fiscale** des véhicules Diesel est légèrement plus élevée que celle des véhicules à essence.
* Les Diesel présentent également une **plage de prix plus large**, avec une médiane plus élevée, ce qui confirme leur position dominante sur le marché.

## 3. Corrélation entre variables numériques

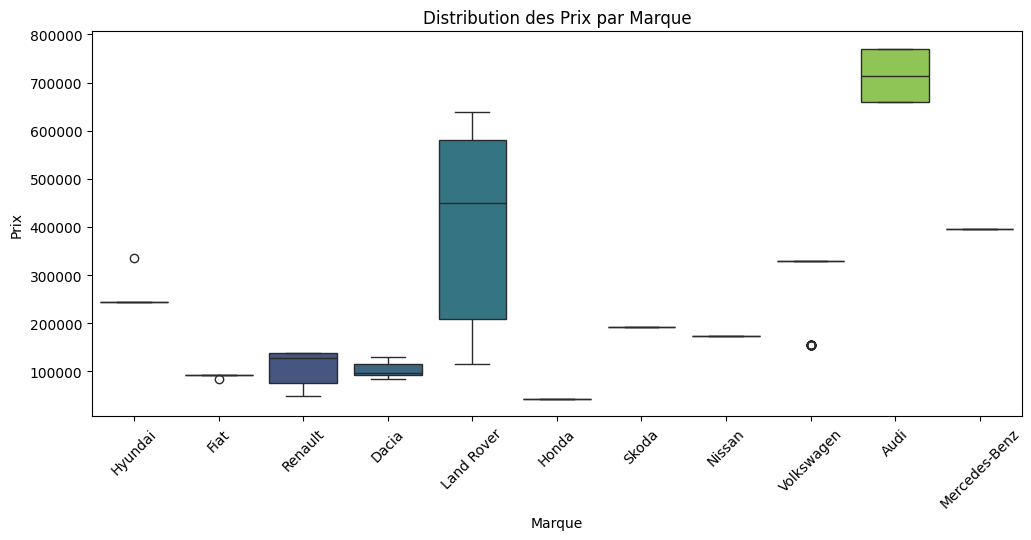
****

Une **matrice de corrélation (heatmap)** a été générée afin d’examiner les relations linéaires entre les variables numériques :

* **Prix ↔ Puissance fiscale** : Corrélation forte positive (**+0.76**)
* **Prix ↔ Année** : Corrélation modérée positive (**+0.44**)
* **Prix ↔ Kilométrage** : Corrélation négative significative (**−0.42**)
* **Année ↔ Kilométrage** : Corrélation négative marquée (**−0.62**), ce qui est cohérent : les voitures récentes sont en général moins roulées.

Ces résultats montrent que les voitures puissantes, récentes et peu utilisées sont naturellement mieux valorisées sur le marché de l’occasion.

## 4. Étude du prix selon des variables catégorielles

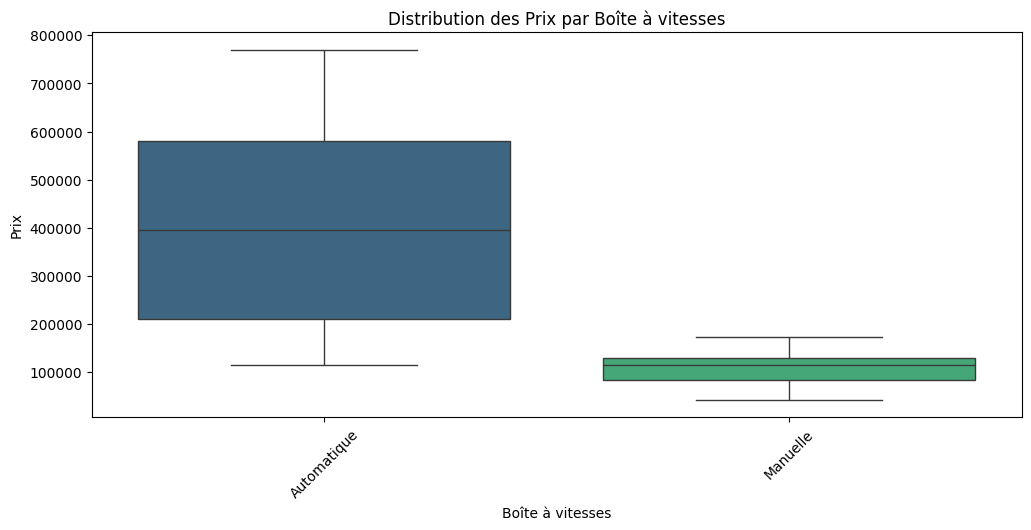
a. Prix par marque****

L’analyse des prix par marque via un **boxplot** révèle des différences notables :

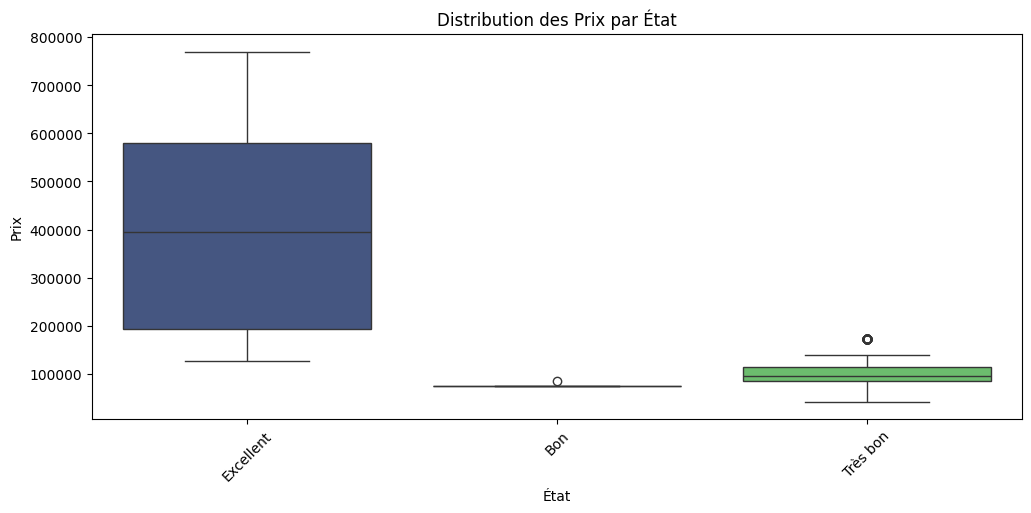
* **Audi, Land Rover, Mercedes-Benz** : Prix élevés, grande dispersion → Segment premium
* **Fiat, Honda, Dacia** : Prix bas et homogènes → Positionnement économique
* **Renault, Volkswagen, Hyundai** : Positionnement intermédiaire

Cette hétérogénéité reflète le positionnement stratégique de chaque marque sur le marché, allant du luxe à l’entrée de gamme.

### b. Prix par type de boîte à vitesses

****

Les véhicules **automatiques** présentent une médiane de prix **nettement plus élevée** que les véhicules **manuels**, accompagnée d’une plus grande dispersion. Cela reflète leur association fréquente avec des modèles modernes et haut de gamme.

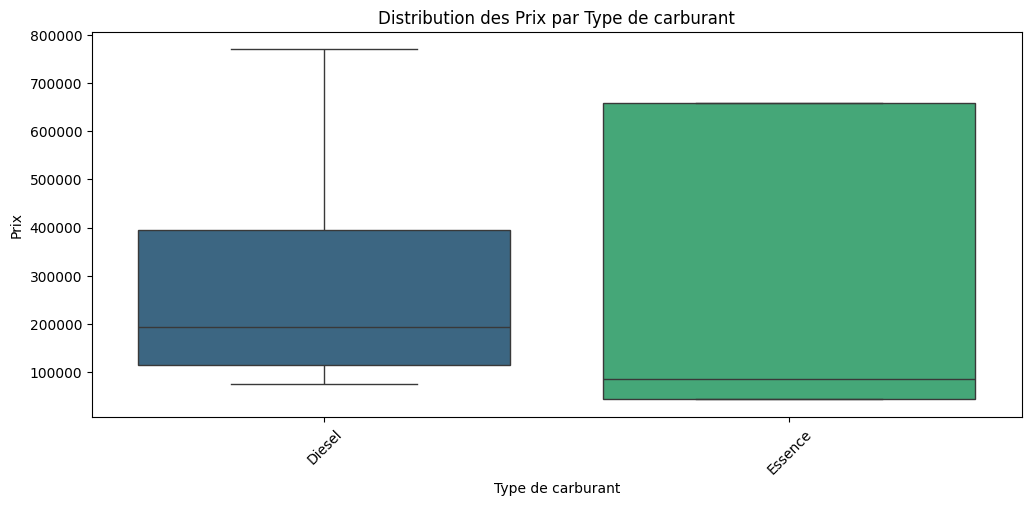
c. Prix par état du véhicule****

L’état du véhicule constitue un critère clé de valorisation :

* **État "Excellent"** : Prix les plus élevés et dispersion maximale
* **État "Très bon"** : Prix intermédiaires avec une distribution modérée
* **État "Bon"** : Prix faibles, faible variation

Plus le véhicule est déclaré en bon état, plus sa valeur marchande est élevée, ce qui témoigne de l’importance perçue de la qualité d’entretien.

### d. Prix par type de carburant

****

Les véhicules **Diesel** sont généralement plus chers que leurs équivalents **Essence**, avec une distribution plus large. Ce constat pourrait être lié à une perception accrue de la fiabilité et de la durabilité du Diesel sur le marché marocain.

**Conclusion de l’EDA**

L’analyse exploratoire a permis de dégager des tendances fondamentales dans la valorisation des voitures d’occasion :

* **La puissance fiscale**, **l’année de mise en circulation** et **le kilométrage** ressortent comme des facteurs prédictifs majeurs du prix.
* Les variables catégorielles telles que **la marque**, **le type de boîte**, **le type de carburant** et **l’état général** modulent fortement la valeur finale du véhicule.
* La présence de valeurs extrêmes dans les prix, typiques du segment premium, justifie l’utilisation de techniques de transformation pour modéliser ces données avec plus de robustesse.