Projet SB 12

Rapport de projet statistique : Prédiction des prix de vente des voitures d'occasion

Réalisé par :
Abdeljalil FARID
Ahmed Taha LAMRANI
Mohamed BOUHMID

Encadré par :
Mr. Saad BENJELLOUN

TABLE DES MATIÈRES

REMERCIEMENT

Nous tenons à exprimer notre profonde gratitude envers notre encadrant, Monsieur Saad BENJELLOUN, pour son soutien constant, son expertise et ses conseils précieux tout au long de ce projet. Sa disponibilité, son engagement et sa patience ont été essentiels pour nous guider à travers les défis méthodologiques et techniques rencontrés lors de la réalisation de ce travail. Son expertise a grandement contribué à notre apprentissage et à la réussite de ce projet. Nous lui sommes reconnaissants pour son dévouement et son encouragement, et nous le remercions sincèrement pour sa contribution à notre développement académique et professionnel.

1. Introduction

- Contexte du projet

Dans le contexte de l'industrie automobile marocaine, caractérisée par sa croissance rapide et ses fluctuations constantes, la prédiction des prix des voitures d'occasion revêt une importance stratégique. Ce projet académique en statistique s'inscrit dans cette dynamique en proposant une analyse approfondie et des modèles prédictifs pour anticiper les variations de prix sur ce marché.

L'industrie automobile au Maroc connaît une expansion significative, alimentée par une économie en croissance et une demande croissante en matière de transport. Dans ce paysage, le marché des véhicules d'occasion joue un rôle crucial en offrant une alternative économique aux consommateurs et en constituant un segment important du secteur automobile.

La volatilité des prix dans ce marché représente un défi majeur pour les acteurs économiques. Les vendeurs doivent fixer des prix compétitifs pour attirer les acheteurs tout en maximisant leurs profits, tandis que les acheteurs cherchent à évaluer la juste valeur des véhicules proposés. Dans ce contexte, disposer de modèles prédictifs précis devient essentiel pour informer les décisions commerciales et garantir une transaction équitable.

Ce projet s'attache donc à développer des modèles statistiques robustes pour prédire les prix des voitures d'occasion, en utilisant des techniques avancées d'analyse de données et de modélisation statistique. En combinant des données réelles avec des méthodes analytiques rigoureuses, nous visons à fournir aux acteurs de l'industrie des outils fiables pour évaluer et anticiper les tendances du marché.

Au-delà de son importance pratique, ce projet revêt également une dimension académique, offrant à notre groupe une opportunité d'appliquer nos connaissances théoriques à des problèmes réels. En explorant les complexités du marché des voitures d'occasion au Maroc et en proposant des solutions innovantes, nous contribuons à enrichir notre compréhension de l'industrie automobile et à développer des compétences essentielles pour notre future carrière professionnelle.

2. Les données

- Scrapping

Les données utilisées dans ce projet ont été extraites du site web moteur.ma, une plateforme spécialisée dans l'achat et la vente de voitures neuves et d'occasion au Maroc. Moteur.ma propose divers services, notamment un guide d'achat complet pour les particuliers et les professionnels de l'automobile, ainsi qu'un calculateur de cote reconnu comme une référence dans le secteur.

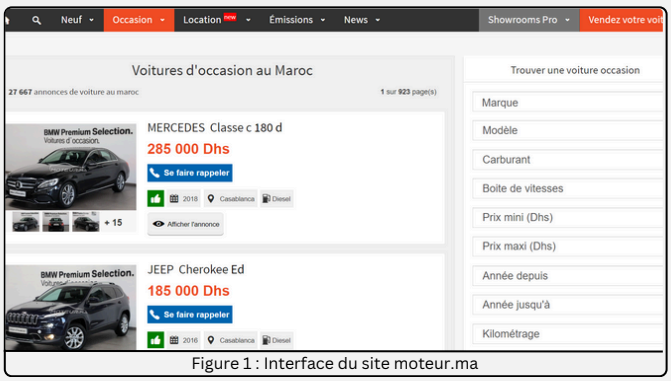

Figure 1 : Interface du site moteur.ma

Pour collecter les données nécessaires à notre analyse, nous avons utilisé l'outil de web scraping de webscraper.io. Ce choix s'est justifié par sa facilité d'utilisation et sa capacité à extraire efficacement des données à partir de sites web complexes. L'outil nous a permis de configurer un scraper pour collecter les informations pertinentes sur les voitures d'occasion disponibles sur moteur.ma.

Le processus de collecte des données s'est déroulé en plusieurs étapes :

1. Configuration du scraper : Nous avons utilisé l'interface conviviale de webscraper.io pour configurer notre scraper en sélectionnant les éléments à extraire sur le site moteur.ma, tels que les caractéristiques des véhicules et leurs prix.

2. Extraction des données : Une fois configuré, le scraper a automatiquement parcouru le site web, extrait les informations désirées et les a structurées dans un format exploitable pour notre analyse.

Grâce à cet outil, nous avons pu collecter efficacement un ensemble de données représentatif du marché des voitures d'occasion au Maroc. Ces données constituent la base de notre analyse statistique visant à prédire les prix des voitures d'occasion et à fournir des insights précieux aux acteurs de l'industrie automobile.

- # Base de données brute

Notre base de données brutes est composé de 17028 lignes et 22 colonnes, la figure de côté montre un aperçu sur les 5 premières lignes de note base de données :

	web-scraper-order	web-scraper-start-url	voitures	voitures-href	typ
0	1713692489-1	https://www.moteur.ma/fr/voiture/achat-voiture...	Afficher l'annonce	https://www.moteur.ma/fr/voiture/achat-voiture...	VOLV Xc9
1	1713692496-2	https://www.moteur.ma/fr/voiture/achat-voiture...	Afficher l'annonce	https://www.moteur.ma/fr/voiture/achat-voiture...	RENAUL Fluenc
2	1713692502-3	https://www.moteur.ma/fr/voiture/achat-voiture...	Afficher l'annonce	https://www.moteur.ma/fr/voiture/achat-voiture...	RENAUL Symbo
3	1713692508-4	https://www.moteur.ma/fr/voiture/achat-voiture...	Afficher l'annonce	https://www.moteur.ma/fr/voiture/achat-voiture...	DACI Sandero
4	1713692515-5	https://www.moteur.ma/fr/voiture/achat-voiture...	Afficher l'annonce	https://www.moteur.ma/fr/voiture/achat-voiture...	LAND ROVER Rang rove spor

Figure 2 : Aperçu de la base de données brute

Les colonnes de notre base de données comprennent les éléments suivants :

1. web-scraper-order : Ordre de la ligne dans la sortie du scraper.
2. web-scraper-start-url : URL de départ à partir de laquelle les données ont été extraites.
3. voitures : Marque et modèle de la voiture.
4. voitures-href : Lien URL vers la page détaillée de la voiture sur le site moteur.ma.
5. type : Type de véhicule.
6. price : Prix du véhicule.
7. Kilométrage : Kilométrage parcouru par le véhicule.
8. Année : Année de fabrication du véhicule.
9. Boite de vitesses : Type de boîte de vitesses du véhicule.
10. Carburant : Type de carburant utilisé par le véhicule.
11. Date : Date de publication de l'annonce.
12. Puissance fiscale : Puissance fiscale du véhicule.
13. Nombre de portes : Nombre de portes du véhicule.
14. Couleur : Couleur du véhicule.
15. Carrosserie : Type de carrosserie du véhicule.
16. Véhicule en garantie : Indique si le véhicule est encore sous garantie.
17. Voiture personnalisée (tuning) : Indique si le véhicule a subi des modifications de tuning.
18. Première main : Indique si le véhicule est de première main.
19. add1 : Informations supplémentaires.
20. add2 : Informations supplémentaires.
21. add3 : Informations supplémentaires.
22. add4 : Informations supplémentaires.

Ces colonnes fournissent une variété d'informations sur chaque véhicule répertorié dans notre base de données. Cependant, ces colonnes ne sont pas toutes utiles, pour cette raison nous nous contenterons dans le reste de l'étude aux colonnes suivantes: 'Kilométrage', 'Année', 'Boite de vitesses', 'Carburant', 'Date', 'Puissance fiscale', 'Nombre de portes', 'Couleur', 'Carrosserie', 'Voiture personnalisée (tuning)', 'Première main', 'type', 'price'.

• Base de données après nettoyage :

La base de données brute présentait plusieurs problèmes nécessitant un nettoyage approfondi, notamment en ce qui concerne le format des données. En effet, toutes les données étaient initialement textuelles, ce qui rendait difficile toute analyse quantitative ou statistique.

L'un des défis majeurs rencontrés lors du nettoyage des données était le problème de structure, notamment en ce qui concerne les types de données. En effet, lors du scraping des données à partir du site moteur.ma, nous avons constaté que les utilisateurs qui publient des annonces ne renseignent pas toujours les mêmes informations de manière cohérente. Par conséquent, une même colonne pouvait contenir à la fois des données sur le kilométrage, la puissance fiscale et d'autres informations.

Pour remédier à ce problème, nous avons mis en place une solution consistant à ajouter un préfixe spécifique devant chaque valeur avant de les scraper. La figure ci-dessous illustre notre méthodologie :

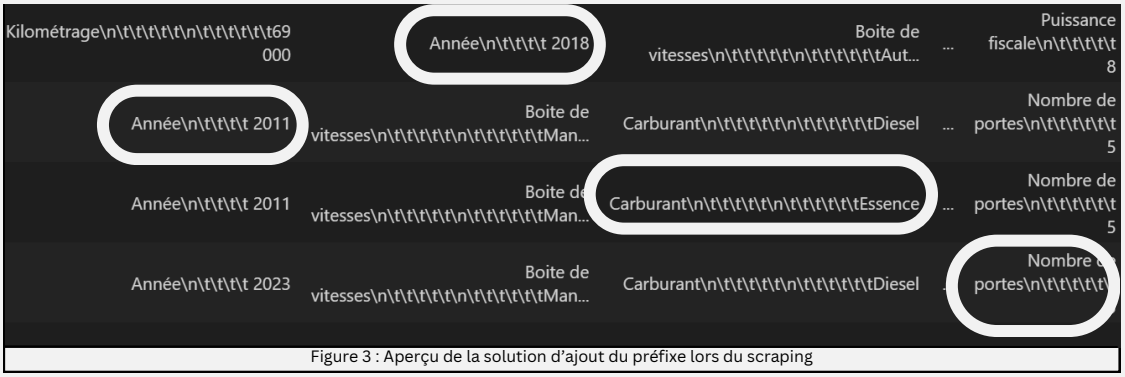

Figure 3 : Aperçu de la solution d'ajout du préfixe lors du scraping

La base de données brute présentait plusieurs problèmes nécessitant un nettoyage approfondi, notamment en ce qui concerne le format des données. En effet, toutes les données étaient initialement textuelles, ce qui rendait difficile toute analyse quantitative ou statistique.
Ainsi, pour récupérer la valeur appropriée d'une ligne donnée, il suffit de parcourir les valeurs de cette ligne et de rechercher le préfixe correspondant à la colonne d'intérêt. Cette méthode nous a permis d'harmoniser et de structurer efficacement notre base de données, facilitant ainsi son analyse ultérieure.
Une fois les données sont structurées, la prochaine étape consiste à convertir les données numériques appropriées dans un format adéquat. Parmi les données à convertir, nous trouvons le prix, la puissance fiscale, l'année, le kilométrage et le nombre de portes. Cette conversion permettra une analyse plus précise et des calculs statistiques pertinents.
La figure ci-dessous illustre la différence avant et après la conversion des types de données :

```
Kilométrage                        object
Année                              object
Boite de vitesses                  object
Carburant                          object
Date                       datetime64[ns]
Puissance fiscale                  object
Nombre de portes                   object
Couleur                            object
Carrosserie                        object
Voiture personnalisée (tuning)     object
Première main                      object
type                               object
price                              object
dtype: object
```
Figure 4 : Avant la conversion

```
Kilométrage                        float64
Année                              float64
Boite de vitesses                   object
Carburant                           object
Date                       datetime64[ns]
Puissance fiscale                  float64
Nombre de portes                   float64
Couleur                             object
Carrosserie                         object
Voiture personnalisée (tuning)      object
Première main                       object
type                                object
price                              float64
dtype: object
```
Figure 5 : Après la conversion

La dernière étape du nettoyage des données consiste à traiter les valeurs manquantes, également appelées "missing values". Ces valeurs manquantes peuvent être problématiques car elles peuvent fausser les analyses statistiques et les modèles prédictifs. La suppression des valeurs manquantes est donc une étape importante pour garantir la qualité et la fiabilité de notre base de données.

Les figures ci-dessous illustre respectivement l'histogramme et la matrice des données manquantes :

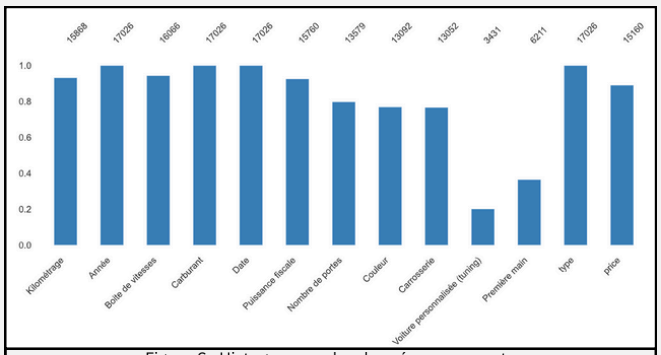

Figure 6 : Histogramme des données manquantes

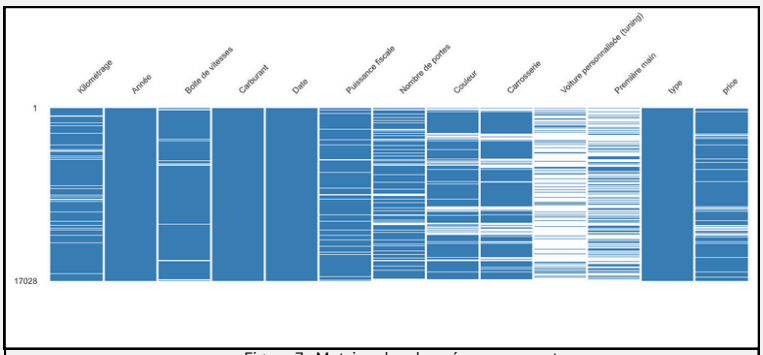

Figure 7 : Matrice des données manquantes

Après avoir effectué les étapes de nettoyage des données, notamment le traitement des valeurs manquantes, le nombre de lignes de notre base de données a été réduit de manière significative. Initialement composée de 17 028 lignes, notre base de données est maintenant constituée de 9 240 lignes, soit une réduction d'environ 45%.

Cette réduction du nombre de lignes s'explique principalement par le nettoyage des données manquantes et la suppression des enregistrements incomplets. Bien que cette réduction puisse sembler importante, elle est essentielle pour garantir la qualité et la fiabilité de notre ensemble de données.

La base de données résultante, composée de 9 240 lignes après nettoyage, est désormais prête à être utilisée pour des analyses statistiques approfondies et des modélisations prédictives précises. En priorisant la qualité des données, nous nous assurons de produire des résultats fiables et significatifs pour notre projet d'analyse des prix des voitures d'occasion au Maroc.

3. Analyse des données

- Histogrammes (profiling avancée)

Après avoir complété le processus de nettoyage de notre base de données, éliminant toute donnée manquante, nous nous engageons maintenant dans une Analyse Exploratoire des Données (EDA). Cette phase, essentielle, vise à obtenir une compréhension approfondie de la structure et des caractéristiques de notre ensemble de données. Notre objectif principal dans cette section est de visualiser les distributions de chaque variable à l'aide d'histogrammes.

Pour ce faire, nous examinons de manière individuelle chaque variable, traçons les histogrammes correspondants et analysons les distributions obtenues. Cette approche nous permet de détecter d'éventuelles tendances, schémas ou anomalies dans nos données, fournissant ainsi des insights précieux pour notre analyse.

Avant de débuter cette analyse exploratoire, nous présentons ci-dessous un tableau descriptif de l'état global de nos données, fournissant une vue d'ensemble des différentes variables et de leurs caractéristiques. Cette présentation initiale nous permettra d'aborder la suite de l'analyse avec une compréhension claire de notre ensemble de données.

	Kilométrage	Date	Puissance fiscale	Nombre de portes	price	car_age	Year_of_annonce	Month_of_annonce
count	9.240000e+03	9240	9240.000000	9240.000000	9.240000e+03	9240.000000	9240.000000	9240.000000
mean	9.412440e+05	2023-01-08 00:41:36.623376640	7.673701	4.798052	1.588810e+05	12.019481	2022.530736	6.398268
min	1.000000e+00	2021-04-21 00:00:00	2.000000	2.000000	1.000000e+00	1.000000	2021.000000	1.000000
25%	8.900000e+04	2022-07-24 00:00:00	6.000000	5.000000	8.100000e+04	8.000000	2022.000000	4.000000
50%	1.450000e+05	2022-12-15 12:00:00	8.000000	5.000000	1.250000e+05	11.000000	2022.000000	6.000000
75%	2.000000e+05	2023-08-02 00:00:00	8.000000	5.000000	1.900000e+05	16.000000	2023.000000	9.000000
max	2.147484e+09	2024-04-21 00:00:00	60.000000	10.000000	2.700000e+06	85.000000	2024.000000	12.000000
std	3.888787e+07	NaN	2.567905	0.597027	1.342143e+05	6.159939	0.796360	3.102296

Figure 8 : Description des données numériques de la base de données

Ce tableau descriptif nous permet d'identifier plusieurs problèmes potentiels dans notre ensemble de données :

1. Problèmes dans la colonne des prix : Nous observons que la valeur minimale dans la colonne des prix est de 1 MAD, ce qui est incohérent et probablement le résultat d'annonces non sérieuses sur le site. Avant de tracer l'histogramme des prix, il est impératif de filtrer cette colonne pour éliminer de telles valeurs aberrantes.
2. Problèmes dans la colonne du kilométrage : De manière similaire, la valeur minimale de la colonne du kilométrage est de 1 km, ce qui est peu probable pour une voiture d'occasion. Encore une fois, il est nécessaire de filtrer cette colonne pour éliminer les annonces non sérieuses avant d'effectuer toute analyse.
3. Nouvelles colonnes ajoutées : Nous avons introduit de nouvelles colonnes, telles que "car_age", qui représente l'âge de la voiture en années. Cependant, nous remarquons que la valeur maximale de cette colonne est de 85 ans, ce qui soulève des questions sur la validité de cette donnée. Il est donc important de filtrer cette colonne également avant d'utiliser ces données pour notre analyse.

Pour remédier aux problèmes identifiés dans notre ensemble de données, nous prévoyons de supprimer les valeurs aberrantes (outliers). Cette approche nous permettra d'éliminer les annonces non sérieuses et les valeurs incohérentes dans chaque colonne, assurant ainsi que nos analyses reposent sur des données fiables et cohérentes. En éliminant les outliers, nous pourrons obtenir des résultats plus précis et significatifs lors de notre Analyse Exploratoire des Données (EDA), nous permettant ainsi de tirer des insights pertinents et exploitables pour notre projet.

	Kilométrage	Année	Date	Puissance fiscale	Nombre de portes	price	car_age
count	7574.000000	7574.000000	7574	7574.000000	7574.000000	7574.000000	7574.000000
mean	148257.185899	2013.422894	2022-12-30 21:11:44.409822976	7.516768	4.807367	144137.347769	11.577106
min	24000.000000	1940.000000	2021-04-21 00:00:00	2.000000	2.000000	46000.000000	2.000000
25%	96000.000000	2010.000000	2022-07-18 00:00:00	6.000000	5.000000	87000.000000	8.000000
50%	146000.000000	2014.000000	2022-12-05 00:00:00	7.000000	5.000000	126000.000000	11.000000
75%	196000.000000	2017.000000	2023-07-26 00:00:00	8.000000	5.000000	180000.000000	15.000000
max	303000.000000	2023.000000	2024-04-21 00:00:00	60.000000	8.000000	375000.000000	85.000000
std	67797.541783	5.046809	NaN	2.315740	0.584806	73536.722148	5.046809

Figure 9 : Description des données numériques de la base de données après la suppression des données aberrantes

Le tableau ci-dessus montre qu'après la suppression des données aberrantes toutes la valeurs sont devenues cohérentes.

Dans cette section, nous effectuons une analyse avancée de nos données en examinant de manière approfondie les caractéristiques de chaque variable. Nous présentons ci-dessous une analyse détaillée des graphiques pour les colonnes importantes telles que le prix, le kilométrage, l'âge du véhicule, la puissance fiscale, ainsi que la marque.

Prix (Price)

Le premier graphique représente une figure illustrant différentes statistiques des prix des voitures d'occasion au Maroc. Nous observons que la distribution des prix est étalée sur une large plage, allant de 83,000 MAD à 2,700,000 MAD. Les quantiles nous permettent de visualiser la dispersion des données autour de la médiane qui est égale à 127,000 MAD, tandis que d'autres informations telles que la moyenne qui est égale à 160,439.25 MAD et l'écart type qui est égale à 132,556.51 MAD nous donnent une idée de la tendance centrale et de la dispersion des prix.

D'après l'histogramme la distribution des prix ressemble à une loi de Gama

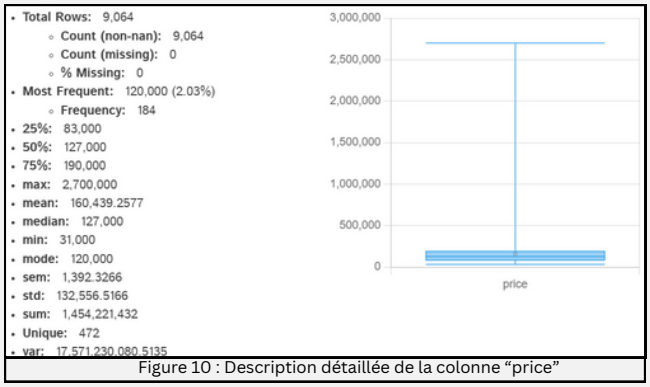

Figure 10 : Description détaillée de la colonne "price"

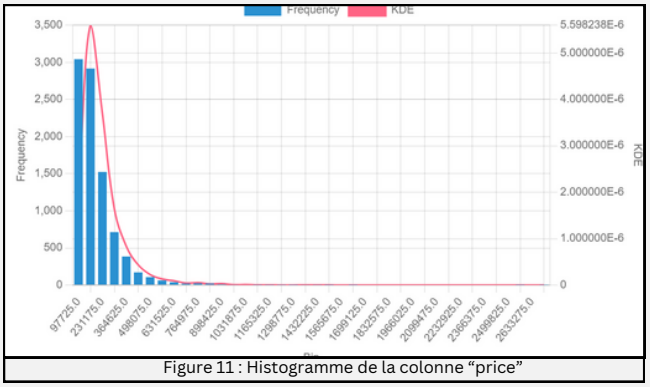

Figure 11 : Histogramme de la colonne "price"

Kilométrage (Kilometrage)

Le deuxième graphique présente une analyse similaire pour la variable du kilométrage des voitures. Nous observons une distribution variée des kilométrages, avec une concentration significative autour de 200,000 Km . Cette distribution nous donne un aperçu de l'utilisation et de l'usure des véhicules d'occasion sur le marché marocain. Nous nous attendons qu'il y aura une forte corrélation entre cette variable avec l'âge du véhicule.

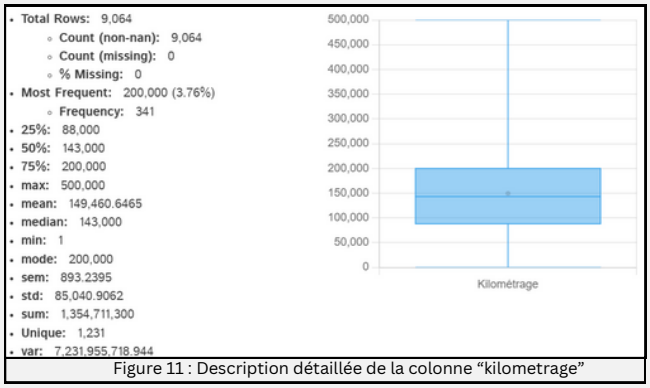

Figure 11 : Description détaillée de la colonne "kilometrage"

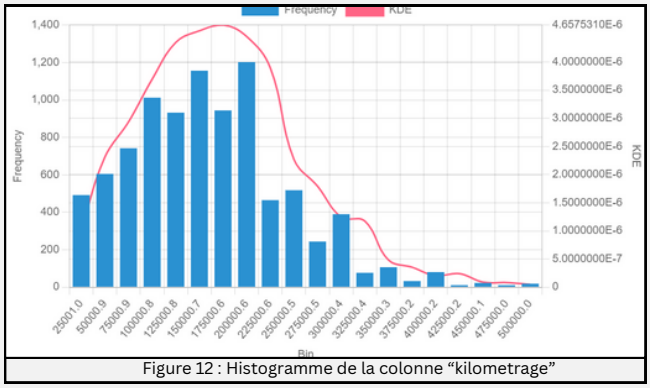

Figure 12 : Histogramme de la colonne "kilometrage"

Âge du véhicule (Car Age)

Le troisième graphique représente l'âge des véhicules en années. Nous observons une distribution relativement uniforme, avec une concentration autour la valeur 13 ans. Cela nous donne une indication de la composition de l'ensemble des voitures d'occasion disponibles sur le marché, en termes d'ancienneté des modèles.

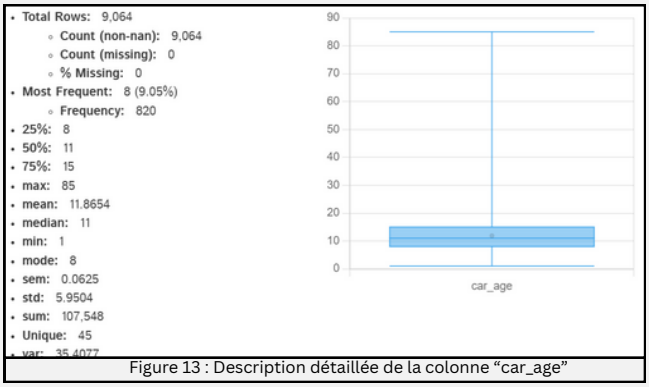

Figure 13 : Description détaillée de la colonne “car_age”

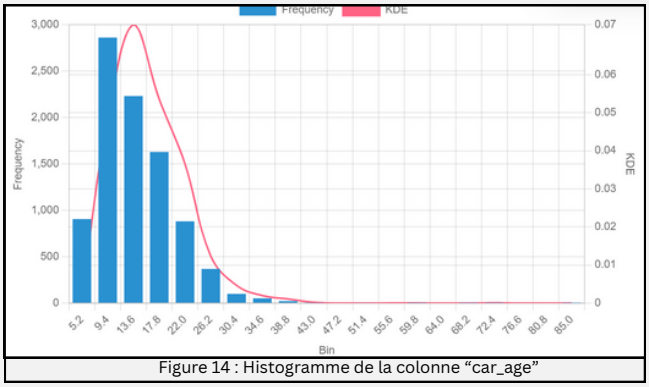

Figure 14 : Histogramme de la colonne “car_age”

Puissance fiscale (Puissance fiscale)

Le quatrième graphique présente une analyse similaire pour la puissance fiscale des voitures. Nous observons une distribution variée des puissances fiscales, avec une concentration autour de 10 chevaux fiscaux. Ce qui correspond à une vignette de 3000 MAD (essence) / 6000 MAD (diesel). Nous nous attendons à une forte corrélation de cette variable avec le prix de la voiture parce qu’elle représente une charge annuelle pour les acheteurs potentiels.

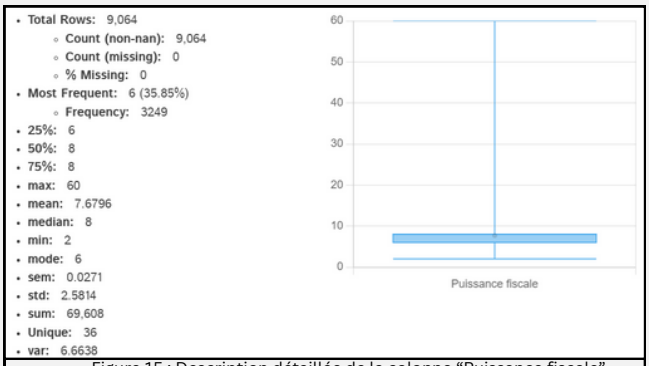

Figure 15 : Description détaillée de la colonne “Puissance fiscale”

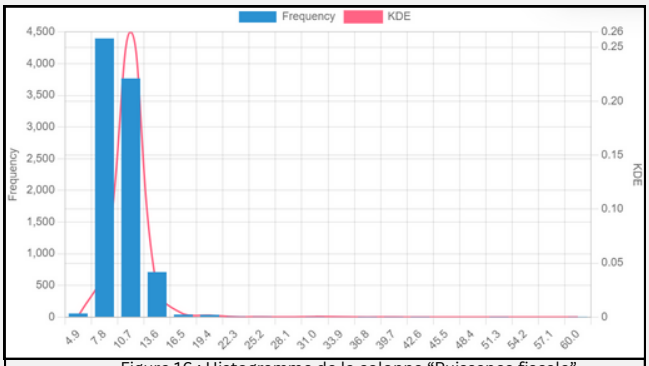

Figure 16 : Histogramme de la colonne “Puissance fiscale”

Marque (Marque)

Pour la variable catégorique de la marque, nous utilisons un nuage de mots (word cloud) pour visualiser les marques les plus fréquentes dans notre ensemble de données. De plus, un tableau illustrant la fréquence des marques les plus fréquentes est également fourni. Cela nous donne un aperçu clair des marques les plus populaires sur le marché des voitures d'occasion au Maroc.

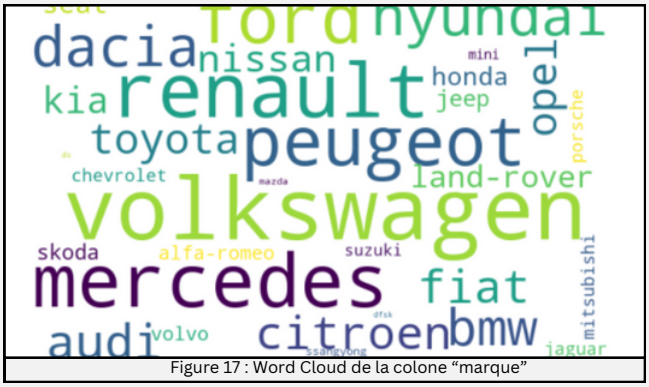

Figure 17 : Word Cloud de la colone “marque”

Value	Count	Frequency (%)
volkswagen	1128	12.4%
renault	809	8.9%
mercedes	754	8.3%
ford	655	7.2%
peugeot	633	7.0%
hyundai	542	6.0%
dacia	533	5.9%

Figure 18 : Tableau des fréquences de la colonne “marque”

En conclusion, cette analyse avancée nous permet de mieux comprendre les caractéristiques de nos données et de détecter d'éventuelles tendances ou particularités. Ces insights nous seront précieux pour la suite de notre analyse et nous aideront à prendre des décisions éclairées dans notre projet d'analyse des prix des voitures d'occasion. L'étude de la corrélation va nous permettre de mieux comprendre la relation entre les différentes variables.

- ## Corrélations:

Dans cette section, nous explorons les relations entre les différentes variables de notre ensemble de données afin de comprendre les interactions et les associations potentielles entre celles-ci. Nous utilisons principalement la matrice de corrélation et des graphiques de dispersion pour examiner ces relations.

Matrice de Corrélation :

La matrice de corrélation nous permet d'identifier les corrélations linéaires entre chaque paire de variables de notre ensemble de données. Les valeurs de la matrice de corrélation varient de -1 à 1, où :

- Une valeur proche de 1 indique une corrélation positive forte,
- Une valeur proche de -1 indique une corrélation négative forte,
- Une valeur proche de 0 indique une faible corrélation ou l'absence de corrélation.

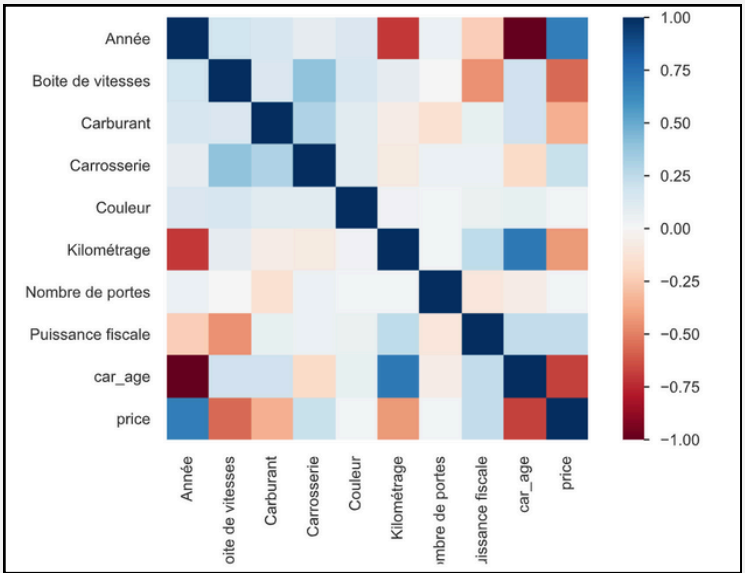

Figure 19 : Matrice de corrélation

En analysant la matrice de corrélation nous faisons les remarques suivantes:

- Il y a une corrélation de -0.424 entre le prix du véhicule et le kilométrage parcourut . Cette forte corrélation est expliqué par le faite que les l'étape du véhicule se dégrade avec la distance.
- Pour la même raison, la corrélation entre le prix du véhicule et son âge est égale à -0.672.
- Il y a une corrélation de 0.231 entre la puissance fiscale et le prix du véhicule . Contrairement à ce qui était prévu, la corrélation entre ces deux variables n'est pas aussi forte. Cela peut être expliquer par le faite qu'il y a d'autres paramètres plus importants que la puissance fiscale de la voiture, qui impactent le prix de cette dernière.
- Il y a une corrélation de 0.698 entre l'âge du véhicule et le kilométrage parcourut . Cette forte corrélation naturelle, et reflète le faite que la véhicule parcours plus de distance cumulée au fil des années.

Graphiques de Dispersion

En complément de la matrice de corrélation, nous utilisons des graphiques de dispersion pour visualiser les relations entre des paires spécifiques de variables.

Ces graphiques nous permettent de détecter visuellement les tendances, les schémas ou les anomalies dans les données.
Pour chaque paire de variables d'intérêt, nous traçons un nuage de points où chaque point représente une observation dans notre ensemble de données. En examinant la répartition des points dans le graphique, nous pouvons identifier les tendances linéaires ou non linéaires entre les variables et évaluer la force de la relation entre elles.

Prix en fonction de la Puissance Fiscale

le premier graphique de dispersion représente le prix des voitures d'occasion en fonction de la puissance fiscale. Chaque point dans le nuage de points représente une observation de notre ensemble de données, où l'axe des x représente la puissance fiscale et l'axe des y représente le prix de la voiture. En examinant ce graphique, nous ne pouvons pas identifier s'il existe une relation linéaire ou non linéaire entre le prix et la puissance fiscale. Cela vient du faite que la puissance fiscale est une valeur entière discontinue, un représentation en box plot, sera plus adéquate a cette situation.
Même après avoir tracer le diagramme en boxe plot, on ne peut pas affirmer qu'il existe une relation linéaire entre le puissance fiscale et le prix de la voiture.

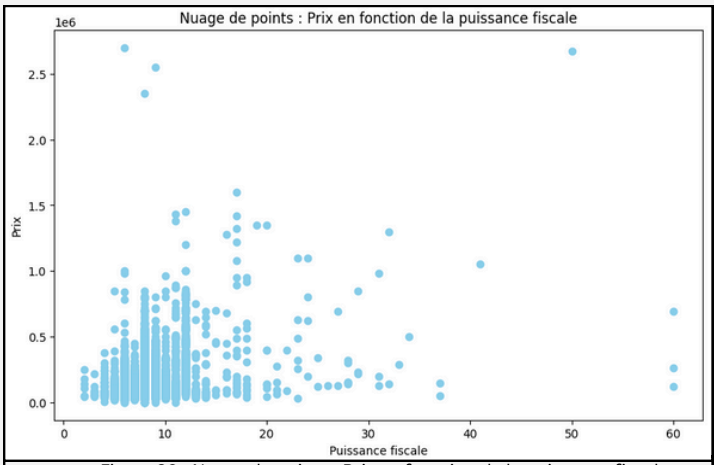

Figure 20 : Nuage de points : Prix en fonction de la puissance fiscale

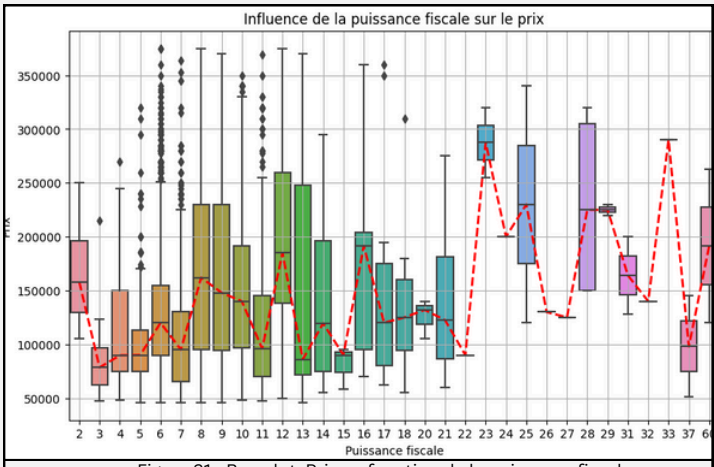

Figure 21 : Box plot: Prix en fonction de la puissance fiscale

Prix en fonction du Kilométrage

Le deuxième graphique de dispersion montre le prix des voitures d'occasion en fonction du kilométrage. Encore une fois, nous remarquons qu'il n'y a pas une relation linéaire entre le prix et le kilométrage .
Les nuages des points obtenus ne sont pas claire, puisque la valeur de la corrélation linéaire entre le prix et les autres variables est généralement faible.

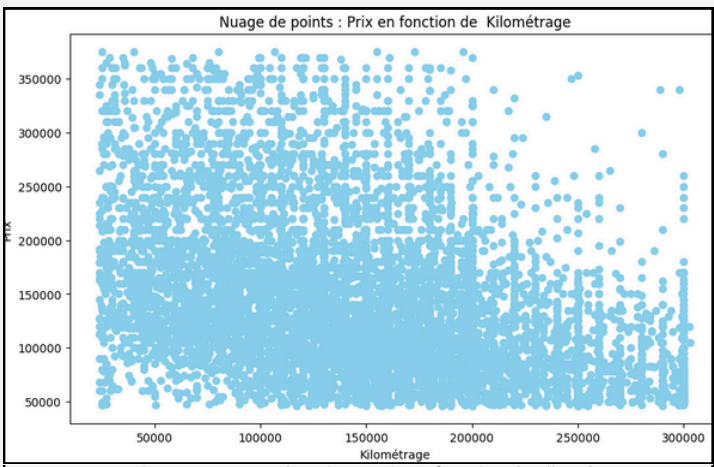

Figure 22 : Nuage de points : Prix en fonction de Kilométrage

4. Modèles

Dans cette section, nous abordons les différents modèles utilisés dans notre projet d'analyse des prix des voitures d'occasion au Maroc. Nous commençons par présenter la régression linéaire multiple, suivie par d'autres modèles tels que les modèles linéaires généralisés (GLM), les arbres de décision, ou encore les méthodes d'apprentissage automatique comme les réseaux de neurones. Chaque modèle est présenté avec ses principes fondamentaux, ses hypothèses et sa mise en œuvre pratique sur nos données.

- ## Régression Multilinéaire:

La régression linéaire multiple est l'un des modèles les plus couramment utilisés pour prédire une variable dépendante à partir de plusieurs variables indépendantes. Son principe repose sur l'ajustement d'une équation linéaire qui décrit la relation entre la variable cible et les variables explicatives. Cette équation prend la forme :
$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn+\varepsilon$
où :

- Y est la variable cible (le prix des voitures dans notre cas),
- X1,X2,...,Xn sont les variables explicatives (telles que l'année du modèle, le kilométrage, etc.),
- β0,β1,...,βn sont les coefficients de régression qui représentent l'effet de chaque variable explicative sur la variable cible,
- ε est le terme d'erreur qui capture les différences entre les valeurs observées et les valeurs prédites par le modèle.

Principe du modèle:

Le principe de la régression linéaire multiple consiste à estimer les coefficients ,β1 ,...,βn qui minimisent la somme des carrés des résidus (la différence entre les valeurs observées et les valeurs prédites par le modèle). Cela se fait généralement à l'aide de la méthode des moindres carrés ordinaires (MCO).

Hypothèses du modèle:

Pour que les estimations des coefficients de régression soient efficaces et non biaisées, la régression linéaire multiple repose sur plusieurs hypothèses fondamentales :

- Linéarité : La relation entre les variables explicatives et la variable cible est linéaire.
- Homoscédasticité : La variance des résidus est constante pour toutes les valeurs des variables explicatives.
- Indépendance des résidus : Les résidus ne sont pas corrélés entre eux.
- Normalité : Les résidus suivent une distribution normale.
- Absence de colinéarité : Les variables explicatives ne sont pas fortement corrélées entre elles.

En vérifiant et en respectant ces hypothèses, nous pouvons obtenir des estimations précises et fiables des coefficients de régression, ce qui nous permettra d'interpréter correctement les effets des variables explicatives sur la variable cible.

Application du modèle aux données:

Dans cette section, nous appliquons notre modèle de prédiction des prix des voitures d'occasion aux données collectées. Les variables indépendantes sont divisées en deux catégories :

- Caractéristiques Numériques (Numerical Features) : Comprend les variables telles que l'âge du véhicule (car_age), le kilométrage (Kilométrage) et la puissance fiscale (Puissance fiscale).
- Caractéristiques Catégoriques (Categorical Features) : Comprend les variables telles que le type de boîte de vitesses (Boite de vitesses), le type de carburant (Carburant), l'année de l'annonce (Year_of_annonce) et le modèle de voiture (car_model). Nous appliquons la méthode get_dummies() aux caractéristiques catégoriques pour les convertir en variables binaires.

La variable dépendante dans notre modèle est le prix (price). Cependant, pour améliorer la stabilité et la distribution des données, nous appliquons une transformation log-normale sur les prix : y=log(log(prix))y=log(log(prix)). Cette transformation permet de réduire la dispersion des données et d'améliorer la performance du modèle.

Après avoir appliqué notre modèle, nous obtenons les résultats suivants :

- R-squared (Entraînement) : 0.9048552541772494
- R-squared (Test) : 0.8775230514322861

Ces valeurs de R-carré indiquent que notre modèle explique environ 90,49 % de la variance des prix dans l'ensemble d'entraînement et 87,75 % de la variance dans l'ensemble de test. Cela suggère que notre modèle a une bonne capacité à prédire les prix des voitures d'occasion au Maroc.

Dans le même contexte, nous appliquons également une régression linéaire à l'aide de la bibliothèque statsmodels. Les résultats obtenus sont similaires à ceux du modèle principal, mais dans certains cas, ils fournissent des résultats plus avancés.

Statistiques de performance du modèle:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.905
Model:	OLS	Adj. R-squared:	0.897
Method:	Least Squares	F-statistic:	114.0
Date:	Mon, 13 May 2024	Prob (F-statistic):	0.00
Time:	01:27:54	Log-Likelihood:	17188.
No. Observations:	6379	AIC:	-3.339e+04
Df Residuals:	5887	BIC:	-3.007e+04
Df Model:	491		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.5090	0.587	7.676	0.000	3.358	5.661
car_age	-0.0054	6.2e-05	-87.587	0.000	-0.006	-0.005
Kilométrage	-4.607e-08	3.03e-09	-15.185	0.000	-5.2e-08	-4.01e-08
Puissance fiscale	-0.0006	0.000	-3.857	0.000	-0.001	-0.000
Year_of_annonce	-0.0010	0.000	-3.311	0.001	-0.002	-0.000
Boite de vitesses_Manuelle	-0.0118	0.001	-16.872	0.000	-0.013	-0.010
Carburant_Electrique	0.0357	0.018	2.032	0.042	0.001	0.070
Carburant_Essence	-0.0164	0.001	-17.797	0.000	-0.018	-0.015
Carburant_Hybride	0.0214	0.004	5.232	0.000	0.013	0.029
car_model_ABARTH 595 turismo 165ch	3.27e-11	1.09e-11	2.990	0.003	1.13e-11	5.41e-11
car_model_ABARTH 695	-0.0079	0.020	-0.402	0.688	-0.046	0.031

Figure 23 : Statistiques de performance du modèle

- R-squared (R^2) : Le coefficient de détermination R^2 est de 0.905, ce qui signifie que 90.5% de la variance de la variable dépendante (prix) est expliquée par le modèle de régression.
- P-values (P>|t|) : Les P-values pour les coefficients individuels indiquent la probabilité de l'observation d'une valeur t (t-statistic) aussi extrême que celle observée, sous l'hypothèse nulle selon laquelle le coefficient est nul. Les valeurs proches de zéro indiquent que les variables sont significatives. Par exemple, pour la variable car_age, la P-value est pratiquement nulle (0.000), ce qui suggère une forte corrélation entre l'âge de la voiture et son prix.

- Coefficients : Les coefficients indiquent la magnitude de l'effet de chaque variable indépendante sur la variable dépendante, toutes choses étant égales par ailleurs. Par exemple, un coefficient négatif pour car_age (-0.0054) suggère qu'à mesure que l'âge de la voiture augmente, son prix diminue.
- R-squared (Test) : Un deuxième R^2 est fourni pour le test, avec une valeur de 0.877, qui peut être utilisé pour évaluer la performance prédictive du modèle sur des données non utilisées lors de l'apprentissage.

Les résultats fournis sont issus d'une régression linéaire multiple appliquée à notre modèle de prédiction des prix des voitures d'occasion. Voici l'interprétation des résultats, en mettant l'accent sur les valeurs p (P>|t|) :

Interprétation des statistiques de performances :

1. const (Constante) :
 - Coefficient (coef) : 4.5090
 - Valeur p (P>|t|) : 0.000
 - Interprétation : La constante représente le terme d'interception de la régression. La valeur p très faible ($p < 0.05$) indique que la constante est statistiquement significative.

2. car_age (Âge du véhicule) :
 - Coefficient : -0.0054
 - Valeur p : 0.000
 - Interprétation : L'âge du véhicule a une influence significative sur le prix, avec une diminution de 0.0054 dans le prix pour chaque unité d'augmentation de l'âge du véhicule.

3. Kilométrage :
 - Coefficient : -4.607e-08
 - Valeur p : 0.000
 - Interprétation : Le kilométrage a une influence significative sur le prix, avec une diminution de 4.607e-08 dans le prix pour chaque unité d'augmentation du kilométrage.

4. Puissance fiscale :
 - Coefficient : -0.0006
 - Valeur p : 0.000
 - Interprétation : La puissance fiscale a une influence significative sur le prix, avec une diminution de 0.0006 dans le prix pour chaque unité d'augmentation de la puissance fiscale.

5. Year_of_annonce (Année de l'annonce) :
 - Coefficient : -0.0010
 - Valeur p : 0.001
 - Interprétation : L'année de l'annonce a une influence significative sur le prix, avec une diminution de 0.001 dans le prix pour chaque unité d'augmentation de l'année de l'annonce.

6. Autres variables :
 - Les variables Boite de vitesses_Manuelle, Carburant_Essence, Carburant_Hybride, et car_model_ABARTH 595 turismo 165ch ont toutes des valeurs p significativement faibles ($p < 0.05$), indiquant qu'elles contribuent de manière significative à la prédiction du prix.

Comparaison des données prédites avec les données réelles

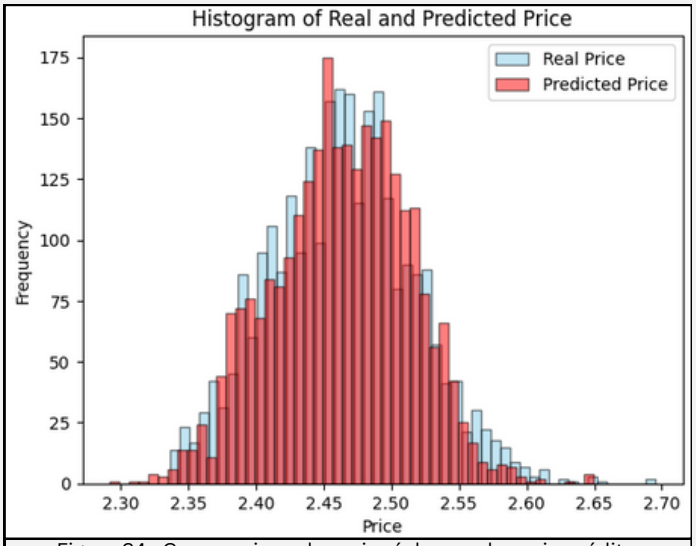

Figure 24 : Comparaison des prix réels avec les prix prédits

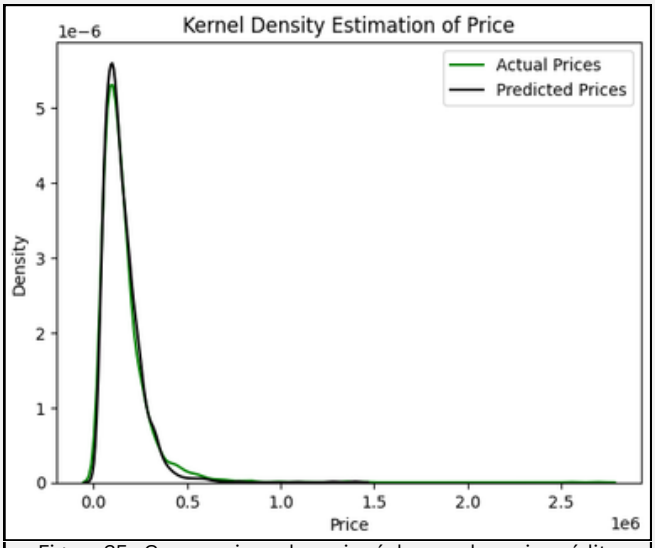

Figure 25 : Comparaison des prix réels avec les prix prédits

En résumé, notre modèle de régression semble bien ajusté, comme en témoigne son R^2 élevé et la significativité des variables.

b.GLM (Modèle linéaire généralisé)

Principe du modèle

Le Modèle Linéaire Généralisé (GLM) étend le concept de régression linéaire multiple en permettant la modélisation de variables dépendantes qui ne suivent pas nécessairement une distribution normale. Contrairement au modèle de régression linéaire classique, le GLM permet de modéliser des variables de réponse qui peuvent être continues ou discrètes et qui peuvent suivre différentes distributions de probabilité, telles que la distribution de Poisson, la distribution binomiale, ou la distribution gamma. Le GLM utilise une fonction de lien pour relier la moyenne de la variable de réponse à un ensemble de prédicteurs. Les hypothèses de linéarité, d'indépendance et d'homoscédasticité sont toujours importantes dans le cadre du GLM, mais la distribution de la variable dépendante peut être spécifiée de manière plus flexible.

Hypothèses du modèle

Les hypothèses clés du modèle linéaire généralisé incluent :

1. Linéarité : La relation entre les prédicteurs et la variable de réponse est linéaire dans l'espace des paramètres.
2. Indépendance : Les observations sont indépendantes les unes des autres.
3. Homoscédasticité : La variance de la variable de réponse est constante pour toutes les valeurs des prédicteurs.
4. Spécification correcte de la distribution et de la fonction de lien.

Application du modèle aux données

Nous avons appliqué le même modèle avec les mêmes variables indépendantes que précédemment .
Après avoir appliqué notre modèle, nous obtenons les résultats suivants :

- Pseudo R-squared (Entraînement): 0.9998
- R-squared (Test) : 0.8828423873579977

Statistiques de performance du modèle:

Generalized Linear Model Regression Results

Dep. Variable:	price	No. Observations:	7290
Model:	GLM	Df Residuals:	6777
Model Family:	Gamma	Df Model:	512
Link Function:	log	Scale:	4.7653e-05
Method:	IRLS	Log-Likelihood:	19656.
Date:	Mon, 13 May 2024	Deviance:	0.31896
Time:	02:15:05	Pearson chi2:	0.323
No. Iterations:	29	Pseudo R-squ. (CS):	0.9998
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.8784	0.221	8.501	0.000	1.445	2.312
car_age	-0.0022	2.38e-05	-94.030	0.000	-0.002	-0.002
Kilométrage	-1.836e-08	1.17e-09	-15.700	0.000	-2.06e-08	-1.61e-08
Puissance fiscale	-0.0002	5.31e-05	-2.915	0.004	-0.000	-5.08e-05
Year_of_annonce	-0.0005	0.000	-4.239	0.000	-0.001	-0.000
Boite de vitesses_Manuelle	-0.0049	0.000	-18.347	0.000	-0.005	-0.004
Carburant_Electrique	0.0133	0.007	1.867	0.062	-0.001	0.027
Carburant_Essence	-0.0068	0.000	-19.582	0.000	-0.007	-0.006
Carburant_Hybride	0.0069	0.002	4.381	0.000	0.004	0.010
car_model_ABARTH 595 turismo 165ch	-4.146e-11	8.2e-12	-5.056	0.000	-5.75e-11	-2.54e-11
car_model_ABARTH 695	-0.0032	0.008	-0.397	0.691	-0.019	0.012
...						

Figure 26 : Statistiques de performance du modèle

Comparaison des données prédites avec les données réelles

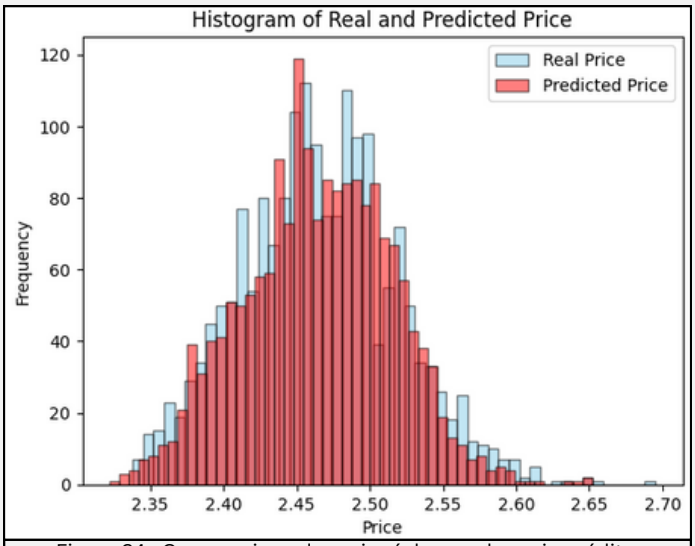

Figure 24 : Comparaison des prix réels avec les prix prédits

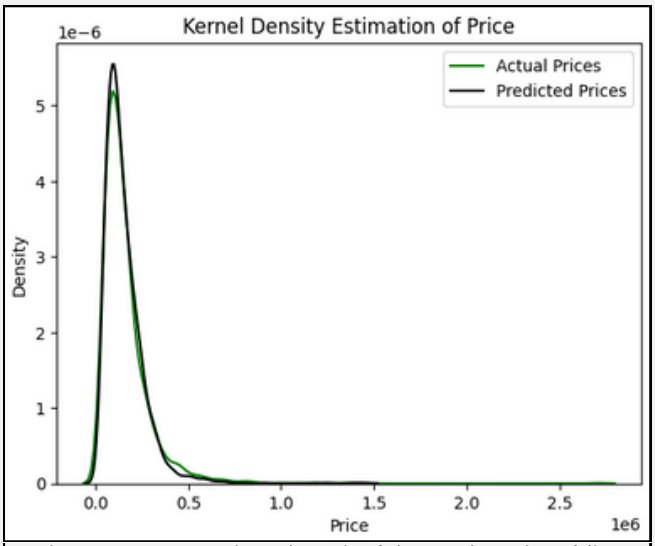

Figure 25 : Comparaison des prix réels avec les prix prédits

Remarques et Conclusion :

Nous remarquons que l'application du modèle de régression linéaire généralisée n'a pas présenté de différence significative en termes de performance par rapport à la régression linéaire multiple. Cette observation s'explique par le fait que la prédiction n'est pas directement basée sur le prix ou sur le logarithme du prix, mais sur le logarithme du logarithme du prix. Ce choix de transformation nous a permis d'atteindre une performance maximale même avec un modèle moins complexe tel que la régression linéaire multiple.
Par ailleurs, toutes les p-values obtenues étaient quasiment faibles, ce qui indique une forte significativité des variables utilisées dans le modèle. Cette constatation suggère que le site moteur.ma fournit des données cohérentes et fiables, ce qui renforce la crédibilité de notre modèle de prédiction des prix des voitures d'occasion au Maroc.
En conclusion, notre approche méthodologique, combinant à la fois la régression linéaire multiple et la régression linéaire généralisée, a permis de développer un modèle robuste et performant pour prédire les prix des voitures d'occasion. Ces résultats démontrent l'efficacité de notre méthodologie dans l'analyse et la modélisation de données complexes, et soulignent l'importance d'une sélection rigoureuse des variables et d'une compréhension approfondie des données pour obtenir des prévisions précises et fiables.