PROJET DATA SCIENCE

Yapi Aho Luc-Aymar*,
Kouassi Kouamé Christian Serge**

*luc-aymar@outlook.fr,
**kouassichristian@gmail.com

Résumé. La Data Science est une discipline qui vise à extraire des connaissances ou des informations dans un ensemble de données. Elle s'appuie sur des outils de Mathématiques, de statistique et d'informatique. L'une des techniques la plus utilisée par les data scientist est sans nul doute l'apprentissage statistique. Système d'apprentissage automatique (en anglais : *Machine Learning*), cette technique traite du problème de la compréhension et de la recherche d'une fonction prédictive basée sur des données. Cet article présente les méthodes utilisées pour faire face à un problème d'apprentissage statistique supervisé, celui de prédire les prix des véhicules d'occasion en Inde dont les données sont issues de *kaggle*.

1 Introduction

1.1 Sujet choisi

Nous avons choisi le sujet « Used Cars Price Prediction » qui consiste à prédire les prix des voitures d'occasion en Inde. Ce sujet nous place face à un problème d'apprentissage statistique supervisé c'est-à-dire nous devons mettre en place un modèle de prédiction à partir d'exemples étiquettés.

Afin de mettre ce modèle de prédiction en place, Un jeu de données dit de "train" est présent afin de permettre aux algorithmes d'apprendre et un jeu de données dit de "test" qui permet aux algorithmes d'appliquer leur apprentissage. La différence entre ces deux jeux de données réside dans le fait que le second (test) ne contient pas les vraies valeurs de la variable à prédire, seul *Kaggle* en à la connaissance.

1.2 Étape de résolution

Pour résoudre ce problème d'apprentissage statistique, il nous a été conseillé d'utiliser la méthode CRISP-DM qui reste aujourd'hui la seule méthode éfficacement pour tous les projets Data Science. Cette méthode comprend 4 parties qui sont :

1. **Compréhension du problème** ;
2. **Exploration et préparation des données** ;
3. **Modélisation et Évaluation** ;
4. **Déploiement**.

1.3 Environnement de travail

Nous étions libres de choisir le langage que nous désirions entre le langage Python ou R. Malgré la multitude de bibliothèques présentes dans Python, nous avons finalement opter pour le langage R. Il semble judicieux pour nous de lister toutes les librairies qui ont été utilisées pour réaliser ce projet sur R. Ce sont : **skimr, stringr, visdat, gridExtra, ggplot2, GGally, Lubridate, Glmnet, RandomForets, Xgboost**.

Dans la suite nous détaillerons le travail accompli dans chacune des étapes de résolution excepté l'étape de déploiement. Les codes R détaillés et commentés seront disponibles dans le dossier Zip.

1.3.1 Compréhension du problème

La première étape avant d'explorer les données est de comprendre le problème que nous essayons de résoudre et les données disponibles. Dans le cadre de ce projet, nous travaillerons avec les données des voitures d'occasions Indiennes. L'objectif est d'utiliser ces données pour construire un modèle qui peut prédire le prix d'une voiture d'occasion en fonction des caractéristiques de cette voiture. Les données d'entraînement incluent le prix des voitures d'occasions, ce qui en fait une tâche d'apprentissage automatique de régression supervisée.

Nous voulons développer un modèle à la fois précis, c'est-à-dire qui peut faire une prédiction proche de la valeur réelle, et interprétable, c'est-à-dire que nous pouvons comprendre les prédictions du modèle.

1.3.2 Exploration et Préparation des données

Dans cette partie on essaiera de se mettre à l'aise avec le jeu de donnée, évaluer sa qualité et comprendre au mieux les différentes variables afin de choisir une stratégie de modélisation pour atteindre notre objectif. Pour simplifier et bien structurer cette étape nous allons utiliser la méthode "QCRD" (Question-Comment-Résultat-Décision).

Cette étape sera divisée en 2 analyses :

- Analyse de la forme :
  Dans l'analyse de la forme on se posera les questions suivantes :
  — Comment le jeu de données est structuré (nombre de lignes, nombre de colonnes) ?
  — Les variables sont-elles dans le bon format ?
  — Combien y-a-t-il de variables qualitatives et quantitatives dans le jeu de données ?
  — Quelle est la variable cible ?
  — Y-a-t-il des valeurs manquantes ?
- Analyse du fond
  Dans l'analyse du fond on se posera les questions suivantes :
  — Que représente chacune des variables du jeu de données ?
  — Quel type de distribution décrit la variable cible ?
  — Quels types de distributions décrivent les autres variables du jeu de données ?
  — Y-a-t-il des valeurs abérrantes dans les différentes distributions ?
  — Quelles sont les variables corrélées à la variable cible ?

— Quelles sont les variables qui sont corrélées entre elles ?

1. **Analyse de la forme**

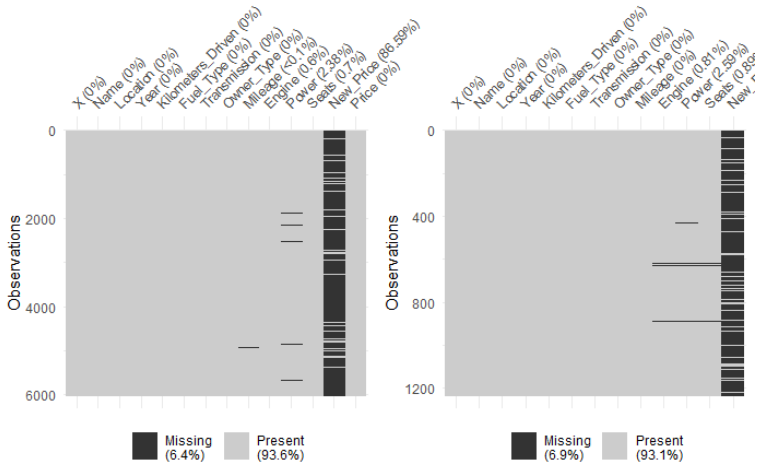
   - **Comment le jeu de données est structuré (nombre lignes, nombre de colonnes) ?**
     Nos disposons d'un jeu de données divisé en deux parties , un train set qui compte **6019 lignes et 14 variables** et un test set qui compte **1234 lignes et 13 colonnes**. La différence du nombre de colonnes est dûe au fait que le train set contient la variable à prédire.
     Les variables du jeu de données sont : **X, Name, Location, Year, Kilometers-Driven, Fuel-Type, Transmission, Owner-Type, Mileage, Engine, Power, Seats, New-Price et Price**.
   - **Les variables sont-elles dans le bon format ?**
     Apres importation voici le format dans lequel le logiciel R a mis chaque variable

Variables	Format après importation
ID	Integer
Name	Factor
Location	Factor
Year	Factor
Kilometers-driven	Integer
Fuel-Type	Factor
Transmission	Factor
Owner-Type	Factor
Mileage	Factor
Engine	Factor
Power	Factor
Seats	Numeric
New-Price	Factor
Price	Numeric

Le contenu des variables et la description des variables disponibles nous permet d'affirmer que toutes les variables ne sont pas dans le format surtout les variables quantitatives. Cela est dû au fait que les unités de mesures des variables accompagnent chaque observation. Pour rectifier cela nous avons utiliser la fonction ***str-replace-all*** de la librairie ***Stringr*** et les fonctions basiques de R telles ***as.numeric*** et ***as.factor***. Le résultat obtenu est le suivant

Variables	Format après importation
ID	Integer
Name	Factor
Location	Factor
Year	Factor
Kilometers-driven	Numeric
Fuel-Type	Factor
Transmission	Factor
Owner-Type	Factor
Mileage	Numeric
Engine	Numeric
Power	Numeric
Seats	Integer
New-Price	Numeric
Price	Numeric

- **Combien y-a-t-il de variables qualitatives et quantitatives dans le jeu de données ?**
  En utilisant la commande ***summary*** ou la fonction ***Skim*** du package ***Skimr*** nous pouvons remarquer que nous disposons exactement de 6 variables qualitatives qui sont : **Name, Location, Year, Fuel-Type, Transmition, Owner-type** et de 8 variables quantitatives qui sont : **X, Kilometers-Driven, Mileage, Engine, Power, Seats, New-Price, Price**.
- **Quelle est la variable cible ?**
  La variable que nous cherchons à Prédire est intitulée ***Price*** dans le jeu de données. C'est une variable quantitative continue.
- **Y-a-t-il des valeurs manquantes ?**
  Nous préférons commencer par une approche graphique pour voir de façon globale l'état du train set et du test set au niveau des valeurs manquantes avant de calculer des pourcentages par colonnes et tirer des conclusions. Pour cela on utilise la fonction ***vis-miss*** de la librairie ***visdat***.
  On obtient les deux graphiques suivants pour le train set et le test set.



Le train set contient peu de valeurs manquantes (6,4%). Cependant la variable ***New Price*** occupe une grande partie dans les 6,4% de valeurs manquantes. Quant au test set il contient 6.9% de valeurs manquantes. Maintenant calculons le pourcentage de valeurs manquantes pour chaque variable. Pour cela on n'utilise pas de package, on a écrit une procédure que vous trouverez dans le fichier R. On obtient le résultat suivant :

Variables	% valeurs manquantes-train set	% valeurs manquantes-test set
ID	0	0
Name	0	0
Location	0	0
Year	0	0
Kilometers-driven	0	0
Fuel-Type	0	0
Transmission	0	0
Owner-Type	0	0
Mileage	0.03	0
Engine	0.59	0.89
Power	2.37	2.59
Seats	0.69	0.89
New-Price	86.59	85.49
Price	0	-

La variables ***New Price*** a plus de 85% de valeurs manquantes dans le train set comme dans le test set ce qui est énorme, les variables ***Mileage, Engine, Seats*** et ***power*** ont très de peu de valeurs manquantes, une méthode d'imputation par la médiane sera utilisée pour résoudre ce problème. Concernant la variable ***New Price*** nous l'examinerons à la fin de l'exploration des données.

2. **Analyse du fond (Train set)**
   - **Que représente chacune des variables ?**

Variables	descriptions
ID	l'identifiant de la voiture
Name	La marque et le modèle de la voiture
Location	le lieu dans lequel la voiture a été vendue ou est disponible pour achat
Year	l'année ou l'édition du modèle
Kilometers-driven	le nombre de kilomètres parcourus par le précédent propriétaire
Fuel-Type	le type de carburant utilisé par la voiture (Petrol, Diesel, Electric, CNG, LPG)
Transmission	le type de transmission utilisé par la voiture (Automatic / Manual)
Owner-Type	Nombre de propriétaires précédents
Mileage	Le kilométrage standard offert par le fabricant de la voiture en kmpl or km/kg
Engine	le volume du déplacement dans le moteur en CC
Power	le maximum de puissance de la voiture en cc
Seats	le nombre de place dans la voiture
New-Price	le prix d'une nouvelle voiture du même modèle
Price	le prix de la voiture d'occasion (target)

TAB. 1 – *Signification des variables*

- **Quel type de distribution décrit la variable cible ?**
  La variable ***Price*** étant une variable quantitative continue, l'une des représentations pouvant nous aider à analyser sa distribution est l'histogramme. Nous avons aussi récapitulé ses indicateurs statistiques.

Min	1^{er} quartile	Mediane	3^{ieme} quartile	Max	Moyenne	Ecart-type
0.440	3.5005	5.640	9.950	160.000	9.479	11.18792

TAB. 2 – *Indicateurs statistiques-Price.*

On remarque que la distribution de la variable cible est asymétrique à gauche (médiane<moyenne). Vu la forme de la distribution et ses caractéristiques, une transformation logarithmique permet de la rendre plus normale (fig 2).

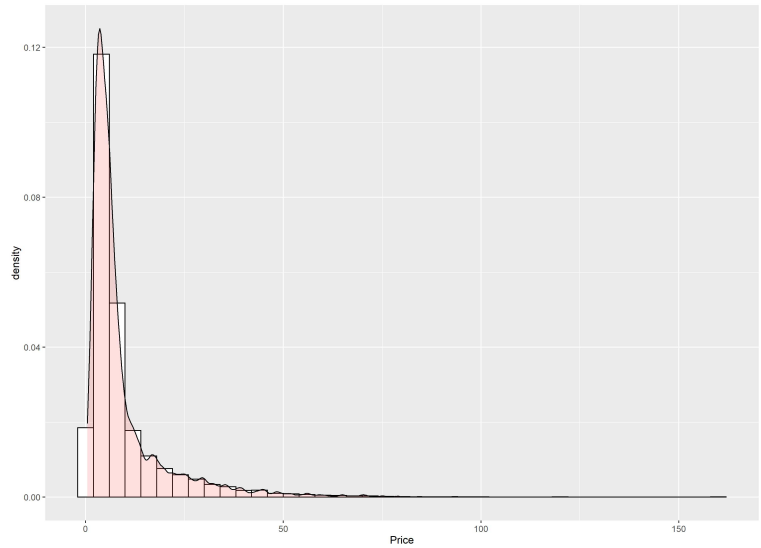


FIG. 1 – *Histogramme de la variable Price.*

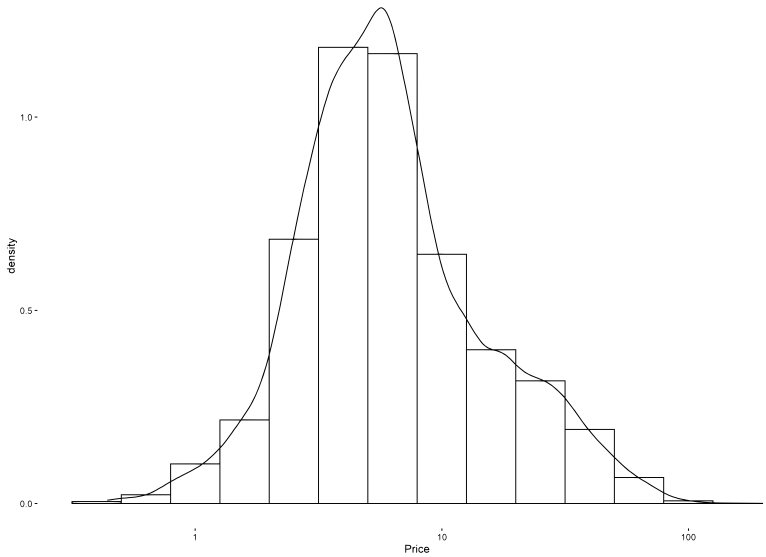


FIG. 2 – *Variable Price après transformation logarithmique.*

- **Quels types de distributions décrivent les autres variables du jeu de données et ont-elles un impact sur la variable cible (Price) ?**

  > Variables qualitatives

La meilleure façon de visualiser une variable qualitative est sans doute de la représenter par un diagramme en Barre. Représentons par un diagramme en Barre toutes les variables qualitatives du train set.

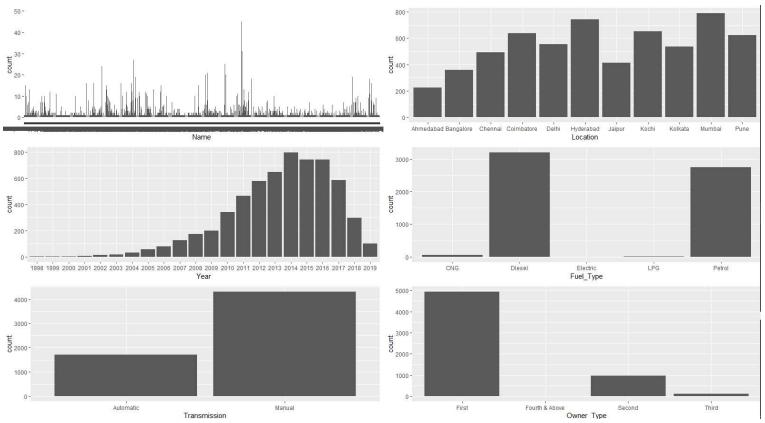


FIG. 3 – *Diagrammes en barre des variables qualitatives.*

Variables qualitatives	Nombre de modalités	Top 3 des modalités
Name	6019	-
Location	11	Mumbai-Hyderabad-Kochi
Year	22	2014-2015-2016
Fuel-Type	5	Diesel-Petrol-CNG
Transmission	2	Manual-Automatic
Owner-Type	4	First-Second-Third

Commentaire : la variable ***Name*** qui représente la marque et le modèle de la voiture compte exactement 6019 modalités autant que le nombre de lignes de notre train set. Il parait judicieux pour nous d'extraire uniquement que les marques et ainsi créer une autre variable. Pour cela on utilise la fonction ***str-split*** du package ***stringr***.

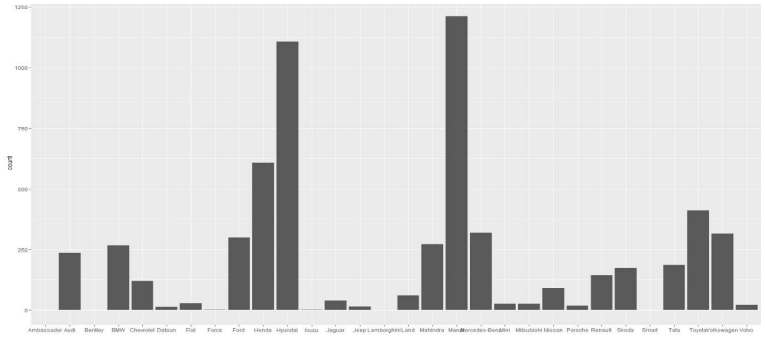


FIG. 4 – *Diagrammes en barre de la variable marque.*

Corrélation entre les variables qualitatives et la variable cible

Pour statuer sur la relation entre les variables qualitatives et le prix nous allons utiliser deux approches : une approche graphique dans laquelle nous allons observer les boîtes à moustaches pour chaque modalité et une autre dans laquelle nous allons utiliser le test de l'anova.

— **Approche graphique**

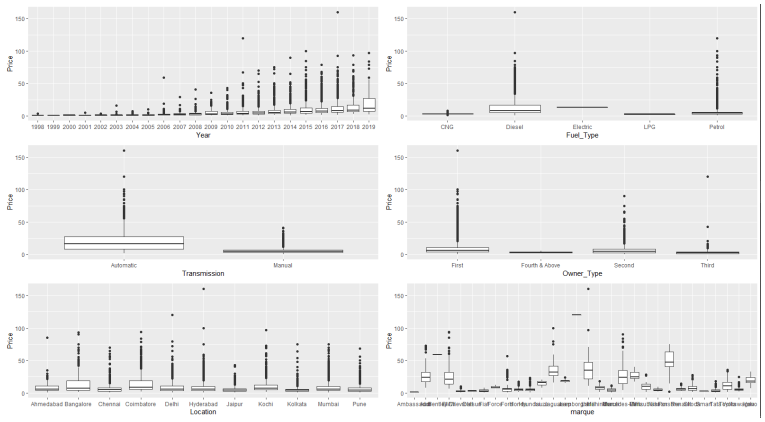


FIG. 5 – *Boîtes à moustache de chaque variable en fonction du prix.*

Commentaire

— En regardant la différence des boxplots liées à chaque modalité on se rend compte que toutes les variables influencent le prix mais certaines plus que d'autres ;

— Les différences paraissent plus prononcées sur le graphique des variables :

- Year : les nouvelles voitures ont un prix plus haut que celui des anciennes voitures ;
- Marque : le graphique nous permet d'observer deux types de voitures : les voitures de marque de luxe comme Lamborghini, Bentley qui ont un prix relativement élevé et les voitures de marques modestes telle que Maruti, Hyundai qui ont des prix relativement faibles.

— **Test de l'Anova**

Variables	P-value
Location	4.6577356381172e-77
Year	4.34149161352524e-120
Fuel-Type	8.14602687971666e-14
Transmission	2.2e-16
Owner-type	1.93493111970881e-12
Marque	2.2e-16

On compare les différentes p-value au seuil de 0.05.

— On remarque que le test de l'anova entre chaque variable qualitative et la variable Price nous produit des p-values qui sont toutes inférieures à 0.05 donc toutes les variables contiennent au moins une modalité dont la moyenne des prix est significativement différente des autres ;

— La p-value de la variable **Year** est largement inférieure à 0.05 ce qui confirme la remarque faite sur l'approche graphique.

> Variables quantitatives

La meilleure façon de visualiser une variable quantitative est sans doute de représenter son histogramme ou une estimation de sa densité (variables continues) et un diagramme en barre (discret). La fonction que nous avons pour le faire est ***ggpairs*** du package ***GGally***. Cette fonction nous permet également de visualiser la matrice de corrélation. On utilisera les résumés statistiques pour l'analyse.

Variables	Min	1^{er} quartile	Mediane	3^{ieme} quartile	Max	Moyenne	Ecart-type
kilometers-Driven	171	34000	53000	73000	6500000	58738	91268.84
Mileage	0.00	15.17	18.15	21.10	33.54	18.13	4.581528
Engine	72	1198	1493	1969	5998	1621	599.6355
Power	34.2	78.0	97.7	138.0	560.0	112.9	53.2837
New-Price	3.91	7.95	11.56	24.45	99.92	20.72	20.21917
Age	1.00	4.000	6.000	9.000	22.00	6.642	3.269742

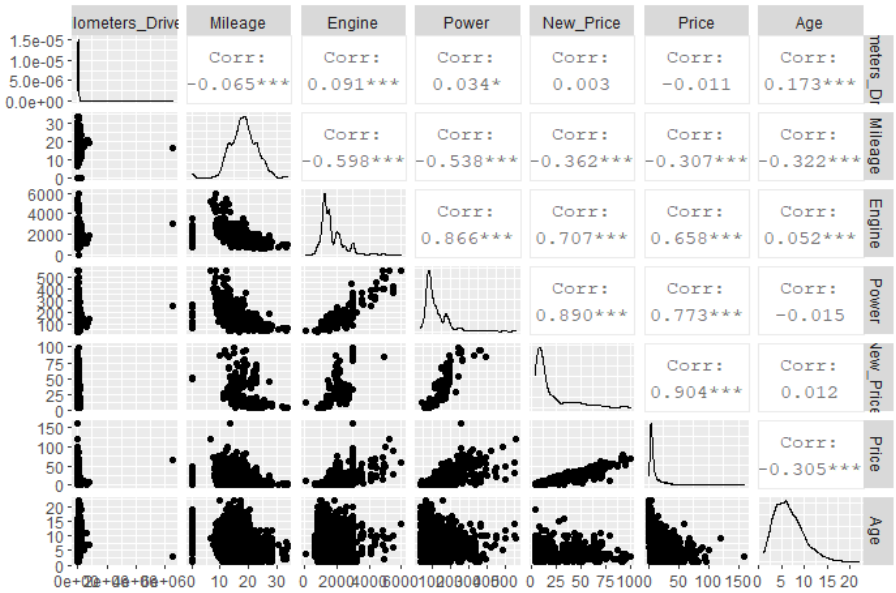


FIG. 6 – *Matrice de corrélation et distribution des variables quantitatives*

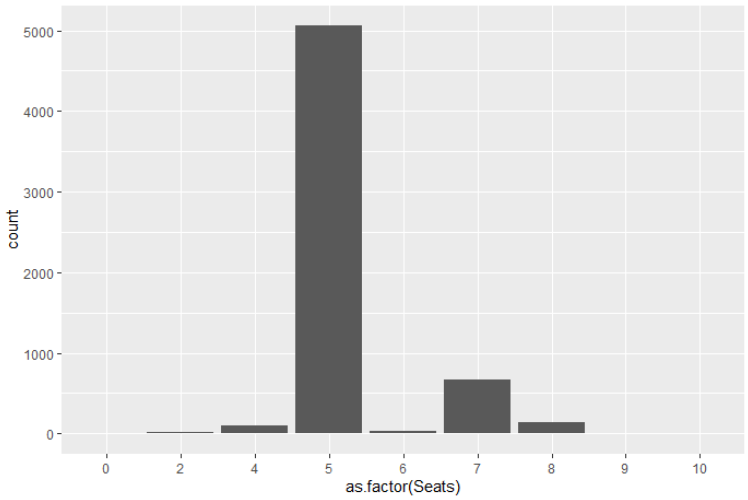


FIG. 7 – *Diagramme en bâtons de la variable Seats*

Commentaire et identification des valeurs aberrantes :

— On remarque que nous avons des voitures d'occasion qui n'ont pas de places assises **(Seats)**. C'est sans doute une erreur. Cette observation sera retirée

Variable qualitative	Nombre de d'occurences	Top 3 des occurences
Seats	10	5-7-8

du train set ;
- On remarque que la variable **kilometers-Driven** est très asymétrique à gauche et prend de très grandes valeurs. Le minimum de cette variable est acceptable mais le maximum laisse planer un peu de doute **(max=650000)** ;
- La variable **Mileage** a un minimum de 0. C'est invraisemblable de voir une voiture d'occasion avec un kilométrage 0. Ces observations seront retirées du train set. Elle a une distribution proche de celle d'une loi normale ;
- Les variables **Engine** et **Power** ont des distributions très similaires. On peut penser que ces variables sont corrélées (asymétrique à gauche).

Corrélation à la variable cible et corrélation entre variables

- La variable la plus corrélée à la variable cible est **New-Price** (0.891), ensuite nous avons dans l'ordre les variables **Power** (0.81) et **Engine** (0,60).Les variables **Mileage** et **Kilometers-Driven** sont faiblement corrélées à **Price** ;
- Les variables **Power** et **Engine** sont fortement corrélées entre elles (0,8). Utiliser les deux variables peut créer un problème de multicolinéarité dans le modèle de régression ;

NB : Malgré le fait que la variable **New-Price** est très corrélée au prix, elle ne sera pas utilisée dans la modélisation car elle a trop de valeurs manquantes 86.59%

1.3.3 Modélisation et évaluation

Pour mettre en œuvre des modèles de prédictions et ensuite les comparer, Il a été nécessaire de :

- Recoder les variables qualitatives en utilisant la méthode **one-hot-encoding** qu'on a mise en œuvre sur R en utilisant la fonction **model.matrix** de R.
  Mais pour essayer de réduire le problème de multicolinéarité auquel on peut être confronté lors de la mise en place d'un modèle linaire, on a créé une autre modalité appelée **'other'** dans les variables **Fuel-Type** et **Marque** qui regroupe les modalités de faibles effectifs. Par exemple les modalités **'LPG'** et **'Electric'** de **Fuel-Type** se regroupent dans la modalité **'other'** ;
- Diviser le train-data en deux **(train set et valid set)** afin de faire l'apprentissage pour choisir le meilleur des modèles ;
- Choisir une métrique pour évaluer les modèles, nous utiliserons le **MSE** (Mean Square Error).

Les modèles mis en œuvre sont décrits ci-dessous

- **Modèle 1 :** Régression linéaire multiple
  La partie d'exploration des données nous a fourni énormément d'informations sur nos variables et surtout leurs liens avec la variable Price. Nous construisons un modèle de régression linéaire avec la fonction **lm** de R avec les variables suivantes :
  - Age **(Year)** et **Marque** : ces deux variables ont un impact significatif sur la variable cible (Price),

— **Power** : c'est la variable quantitative la plus corrélée au prix.

— **Modèle 2 :** Régression Lasso avec la librairie **glmnet**
Nous avons voulu mettre en place ce modèle de regression linéaire multiple avec toutes les variables mais en gérant en même temps le problème de multicolinéarité auquel on peut être confronter apres le recodage des variables qualitatives possédant des modalités de faibles effectifs. Pour cela nous utilisons un modèle de regression Lasso avec lambda choisi par validation croisée ;

— **Modèle 3 :** Forêt aléatoire avec la librairie **randomForest**
L'algorithme de RandomForest donne souvent de bons résultats dans les problèmes de prédictions. Nous l'avons implémenté en utilisant toutes les variables avec la librairie **randomForest** en considérant 100 arbres de décisions ;

— **Modèle 4 :** Gradiant boosting avec la librairie **Xgboost**
L'algorithme de Gradient Boosting donne également de bons résultats en matière de prédiction. Nous l'avons implémenté en utilisant toutes les variables avec la librairie **Xgboost**. Pour les hyperparamètres utilisés voir le code R.

Nb : Le logarithme de base 10 de la variable **Price** sera utilisé pour les modélisations pour optimiser les résultats.
Les résultats obtenus sont les suivants :

Variables	MSE (Train set)	MSE (valid set)	R^2
Modèle 1	0.015	0.020	0.891
Modèle 2	0.010	0.014	0.927
Modèle 3	0.014	0.0058	0.959
Modèle 4	0.0026	0.00966	0.981

Commentaire :

— Les modèle 3 (**random Forest**)et 4 (**Xgboost**) nous donnent les meilleurs résultats avec les score respectifs de 0.95 et 0.98. les resultats du MSE sur les données de validation sont relativement bon. ;

— Nous remarquons aussi qu'une régression linéaire en utilisant les variables Age, Marque et Power, donne d'assez bonnes prédictions et qui sont plus interprétables que les autres modèles ;
nous nous sommes fixés comme objectif de mettre en place un modèle simple et aussi interpretable.Le modèle 1 semble être le modèle adéquat. une étude additionnelle montre que ce modèle respecte assez bien les conditions de validités d'un modèle linéaire (voir script R).
Effectuons pour terminer des prédictions en utilisant les données de test,voici les résultats obtenus.

Id-voiture	prédictions
1	3.6563747
2	2.9434249
3	18.8294518
4	6.3854833
5	4.4709594
6	11.2301354

Summary

Data Science is a discipline that aims to extract knowledge or information from a set of data. It is based on mathematical, statistical and computer tools. One of the techniques most used by data scientists is without a doubt statistical learning. Machine learning system, this technique addresses the problem of understanding and finding a predictive function based on data. This article deploys the methods used to face a problem of supervised statistical learning, that of predicting prices of used vehicles in India whose data come from kaggle.