22119378
22314000

RAPPORT DE PROJET

LE PRIX DES VOITURES D'OCCASION EN INDE

SOMMAIRE

INTRODUCTION

L'industrie automobile en Inde, en constante évolution, est le reflet d'une économie dynamique et diversifiée. Dans ce contexte, l'analyse des prix des voitures d'occasion revêt une importance cruciale, non seulement pour les acheteurs et les vendeurs, mais aussi pour les chercheurs et les décideurs politiques. En effet, comprendre les facteurs qui influent sur ces prix peut fournir des insights précieux sur les tendances du marché, les comportements des consommateurs et les implications économiques et sociales.

Dans cette optique, le présent projet s'attache à analyser un jeu de données exhaustif portant sur les prix des voitures d'occasion en Inde, en mettant à profit les logiciels SAS et R. Ce jeu de données comprend une variété de variables socio-économiques et démographiques, ainsi que des caractéristiques spécifiques des véhicules, telles que le kilométrage, l'âge, la marque et le modèle.

L'objectif principal de cette étude est de déterminer les principaux facteurs qui influent sur les prix des voitures d'occasion en Inde et de développer des modèles de prédiction robustes. Pour ce faire, nous procéderons à une analyse approfondie des variables du jeu de données, à une exploration des relations entre ces variables et les prix des voitures, ainsi qu'à la construction de modèles de régression pour évaluer et prédire les prix.

01

ANALYSE DESCRIPTIVE DU JEU DE DONNÉES

02

1.JEUX DE DONNÉES

- Make - La marque de la voiture (ex: Honda, Hyundai).
- Model - Le modèle de la voiture.
- Fuel Type - Type de carburant utilisé par la voiture (ex: Petrol, Diesel).
- Transmission - Type de transmission (ex: Manual, Automatic).
- Location - L'emplacement ou la ville où la voiture est vendue.
- Owner - Information sur le propriétaire (ex: First, Second).
- Seller Type - Type de vendeur (ex: Individual, Dealer).
- Color - Couleur de la voiture.
- Drivetrain - Le type de système de transmission aux roues (ex: FWD pour Front-Wheel Drive)
- Price - Le prix de la voiture.
- Year - L'année de fabrication de la voiture.
- Kilometer - Le nombre de kilomètres parcourus par la voiture.
- Engine - La cylindrée du moteur (en cm^3 ou cc).
- Max Power - La puissance maximale du moteur (en bhp).
- Max Torque - Le couple maximal (en Nm).
- Length - La longueur de la voiture (en mm).
- Width - La largeur de la voiture (en mm).
- Height - La hauteur de la voiture (en mm).
- Seating Capacity - La capacité de sièges.
- Fuel Tank Capacity - La capacité du réservoir de carburant (en litres).

Type de VariableVariables :
Qualitatives :
Make, Model, Fuel Type, Transmission, Location, Owner, Seller Type, Color, Drivetrain
Quantitatives :
Price, Year, Kilometer, Engine, Max Power, Max Torque, Length, Width, Height, Seating Capacity, Fuel Tank Capacity

2. DISTRIBUTION DES VARIABLES QUALITATIVES

Nous avons effectué la procédure freq pour calculer les fréquences et les pourcentages de chaque modalité de la variable "Fuel Type" dans notre jeu de données. Cette analyse nous aide à comprendre la répartition des différentes catégories de carburant utilisées par les véhicules.

La procédure FREQ

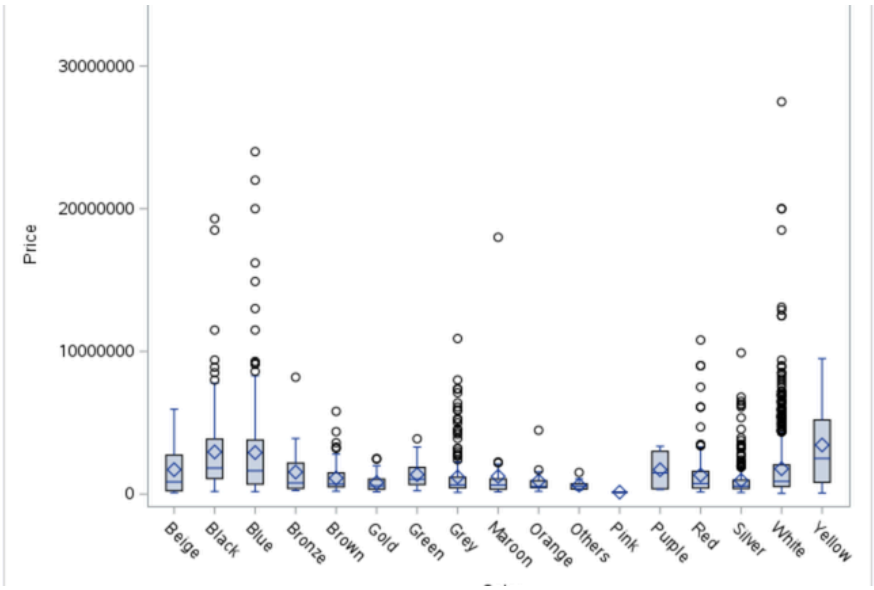
Fuel Type	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
CNG	50	2.43	50	2.43
CNG +	1	0.05	51	2.48
Diesel	1049	50.95	1100	53.42
Electr	7	0.34	1107	53.76
Hybrid	3	0.15	1110	53.91
LPG	5	0.24	1115	54.15
Petrol	944	45.85	2059	100.00

Dans notre jeu de données, la majorité des voitures utilise soit le diesel (50.95%) soit l'essence (petrol, 45.85%). Les autres types de carburant tels que le CNG, CNG+, électrique (Electr), hybride, et LPG sont beaucoup moins fréquents, ne représentant ensemble que 3.21% du total. Cela pourrait indiquer une préférence marquée ou une plus grande disponibilité de voitures diesel et à essence sur le marché concerné par l'étude.

En considérant les pourcentages cumulés, il est évident que les voitures diesel et à essence représentent presque la totalité du marché automobile dans notre ensemble de données, atteignant un pourcentage cumulé de 100%.

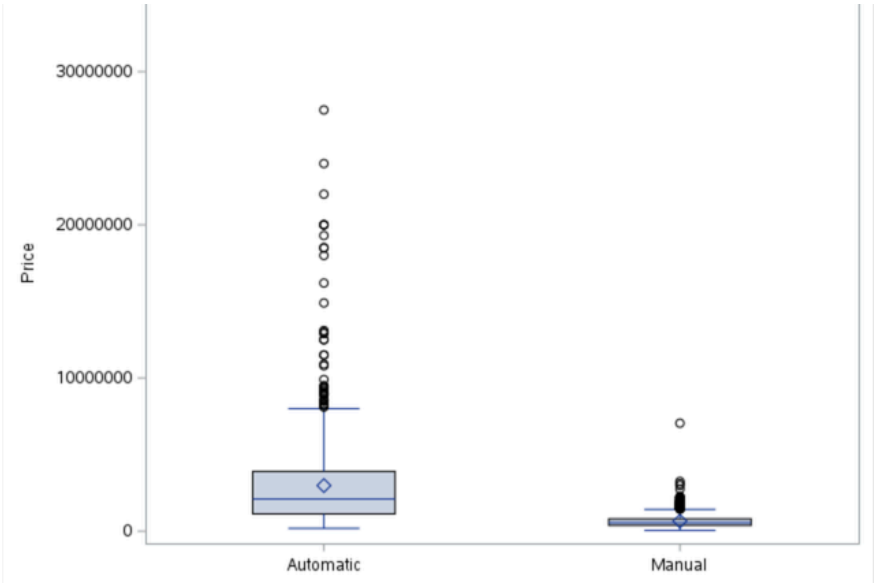
En résumé, cette distribution indique que les véhicules diesel et à essence dominent le marché, tandis que les options de carburant alternatif n'ont pas encore atteint une adoption généralisée en Inde.

Passons maintenant à une analyse approfondie de notre jeu de donnée ; Notre objectif étant de déterminer les variables qui impactent le prix des voitures nous allons nous concentrer sur la variable price. Pour ce faire, nous allons principalement construire nos modèles selon cette variable.
Dans un premier temps, nous avons trié le jeu de données en fonction des différentes observations disponibles, puis, nous allons éliminer les observations aberrantes à l'aide de la procédure « proc sgpplot ».



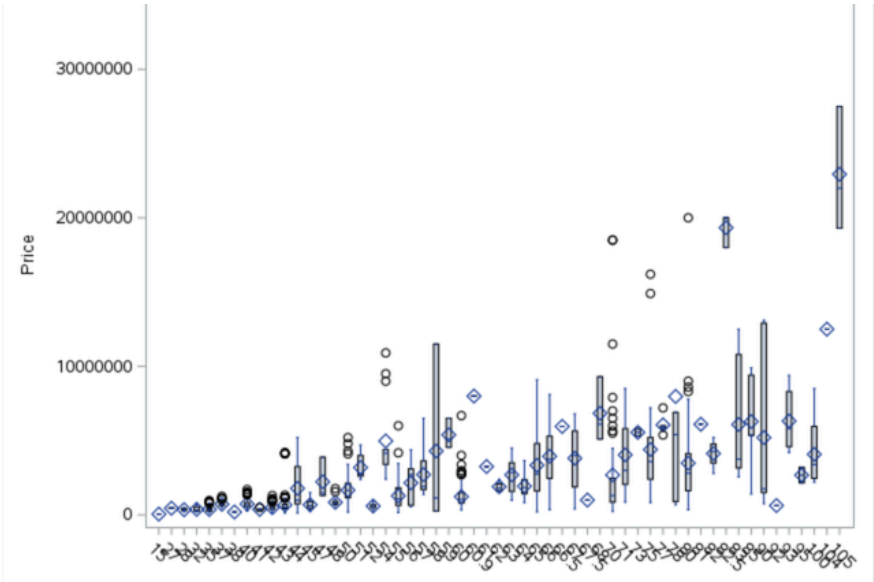
COLOR

Les prix des voitures d'occasion varient considérablement selon la couleur (les plus chères c'est les bleus et les blanches)
Les couleurs peuvent avoir une influence sur le prix des voitures d'occasion en fonction de leur popularité et de leur disponibilité.



TRANSMISSION

Les voitures avec une transmission automatique semblent avoir une distribution de prix plus large que celles avec une transmission manuelle.



FUEL TANK CAPACITY

Il y a une grande variété de distributions de prix en fonction de la capacité du réservoir de carburant.Les voitures avec de plus grands réservoirs peuvent être des modèles plus grands ou plus performants, qui sont généralement plus chers.

Il est actuellement constaté que les variables étudiées influencent le prix des véhicules d'occasion. Cependant, la complexité de leurs interactions ne permet pas de déterminer avec certitude celles qui peuvent être exclues de notre modèle. Face à cette situation, une investigation plus détaillée s'impose. Ainsi, nous envisageons d'approfondir notre analyse pour discerner plus précisément le rôle de chaque variable et affiner notre compréhension de leurs effets sur la valorisation des voitures d'occasion. Cette démarche méthodique est essentielle pour aboutir à des conclusions robustes et fiables.

3. DISTRIBUTION DES VARIABLES QUANTITATIVES

Pour commencer notre analyse des variables quantitatives, nous allons utiliser la procédure « means » afin de regrouper les moyennes, écarts-types, minimums et maximums de chaque variable dans un tableau. Cela nous permettra d'obtenir une vue l'ensemble pour chacune de celles-ci et mieux comprendre leur distribution, que nous examinerons en détail par la suite.

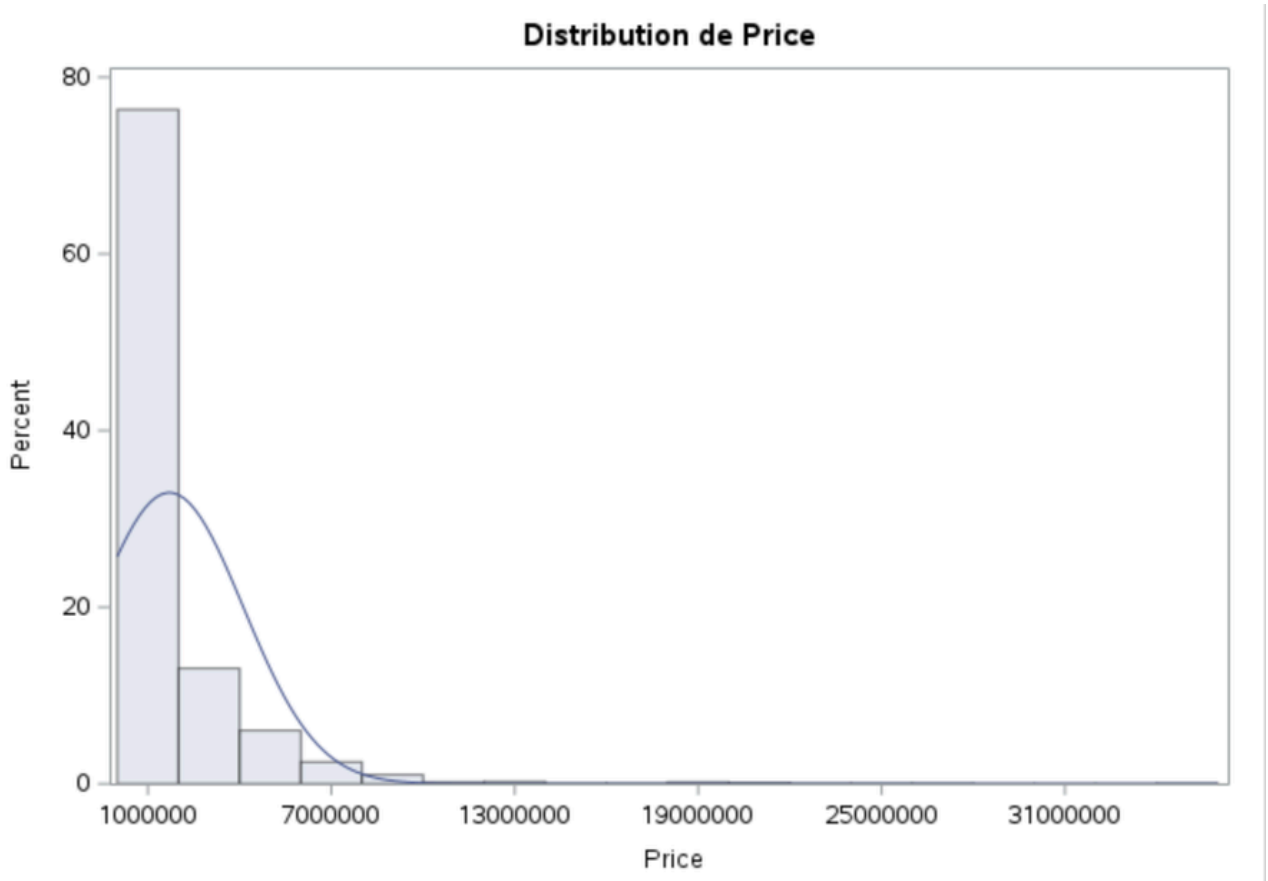
La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum
Price	2059	1702991.70	2419880.64	49000.00	35000000.00
Year	2059	2016.43	3.3635636	1988.00	2022.00
Kilometer	2059	54224.71	57361.72	0	2000000.00
Length	1995	4280.86	442.4585068	3099.00	5569.00
Width	1995	1767.99	135.2658252	1475.00	2220.00
Height	1995	1591.74	136.0739560	1165.00	1995.00
Seating Capacity	1995	5.3062657	0.8221701	2.0000000	8.0000000
Fuel Tank Capacity	1946	52.0022097	15.1101978	15.0000000	105.0000000

L'analyse des données sur les voitures d'occasion révèle une grande diversité du marché, avec une large gamme de prix moyens et extrêmes, reflétant une offre variée allant des véhicules économiques aux modèles de luxe. La majorité des voitures sont récentes, avec un kilométrage moyen indiquant une utilisation modérée. Les dimensions des véhicules montrent une variété de tailles, suggérant une mixité des types de voitures disponibles, tandis que la capacité de sièges et de réservoir indique une gamme allant des citadines aux plus grands véhicules familiaux ou utilitaires. Ensemble, ces mesures suggèrent que les prix des voitures d'occasion sont influencés par un éventail de facteurs, de l'année et du kilométrage aux caractéristiques physiques et fonctionnalités.

En résumé, ces statistiques descriptives révèlent une grande diversité dans les caractéristiques des voitures d'occasion, ce qui nécessite une analyse plus fine pour comprendre comment ces attributs influencent le prix. Les variables avec les plages les plus larges, en particulier le prix, méritent une attention particulière pour identifier les facteurs spécifiques qui contribuent aux valeurs extrêmes.

Maintenant que nous avons effectué une première analyse exploratoire de notre jeu de données, nous allons nous concentrer sur l'étude de la distribution de notre variable dépendante soit « price ».



La procédure UNIVARIATE appliquée aux prix des voitures d'occasion a généré un histogramme avec une courbe de distribution normale superposée. L'histogramme montre que la majorité des prix des voitures d'occasion se concentre dans la tranche la plus basse, avec une chute significative à mesure que le prix augmente. Cela indique que la plupart des voitures d'occasion sont vendues à des prix relativement faibles, avec quelques exceptions à des prix plus élevés. La courbe de distribution normale, cependant, ne s'ajuste pas bien aux données, ce qui suggère que les prix ne suivent pas une distribution normale. En effet, les données présentent une distribution fortement asymétrique avec une concentration de prix dans le bas du spectre et des valeurs extrêmes ou aberrantes vers les prix plus élevés. Cela pourrait indiquer la présence sur le marché d'un petit nombre de voitures d'occasion très coûteuses ou de luxe.

ANALYSE DES VARIABLES DU JEU DE DONNÉE

03

1. TEST DE NORMALITÉ DES VARIABLE QUANTITATIVES

Maintenant que nous avons effectué une première analyse exploratoire de notre jeu de données, nous allons nous concentrer sur l'étude de la distribution de notre variable dépendante soit « price ».

La procédure UNIVARIATE
Fitted Normal Distribution for Price

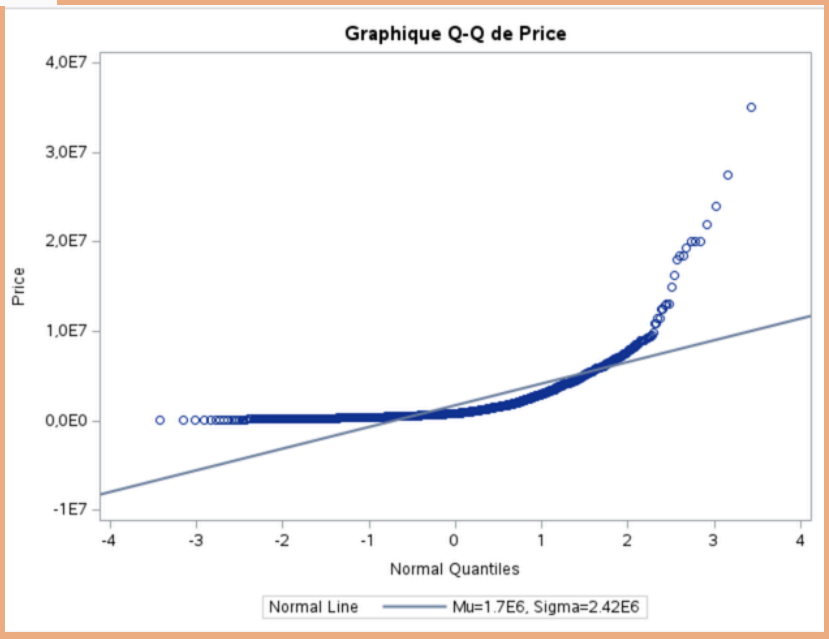
Parameters for Normal Distribution		
Paramètre	Symbole	Estimation
Mean	Mu	1702992
Std Dev	Sigma	2419881

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.254923	Pr > D	<0.010
Cramer-von Mises	W-Sq	45.368794	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	235.052666	Pr > A-Sq	<0.005

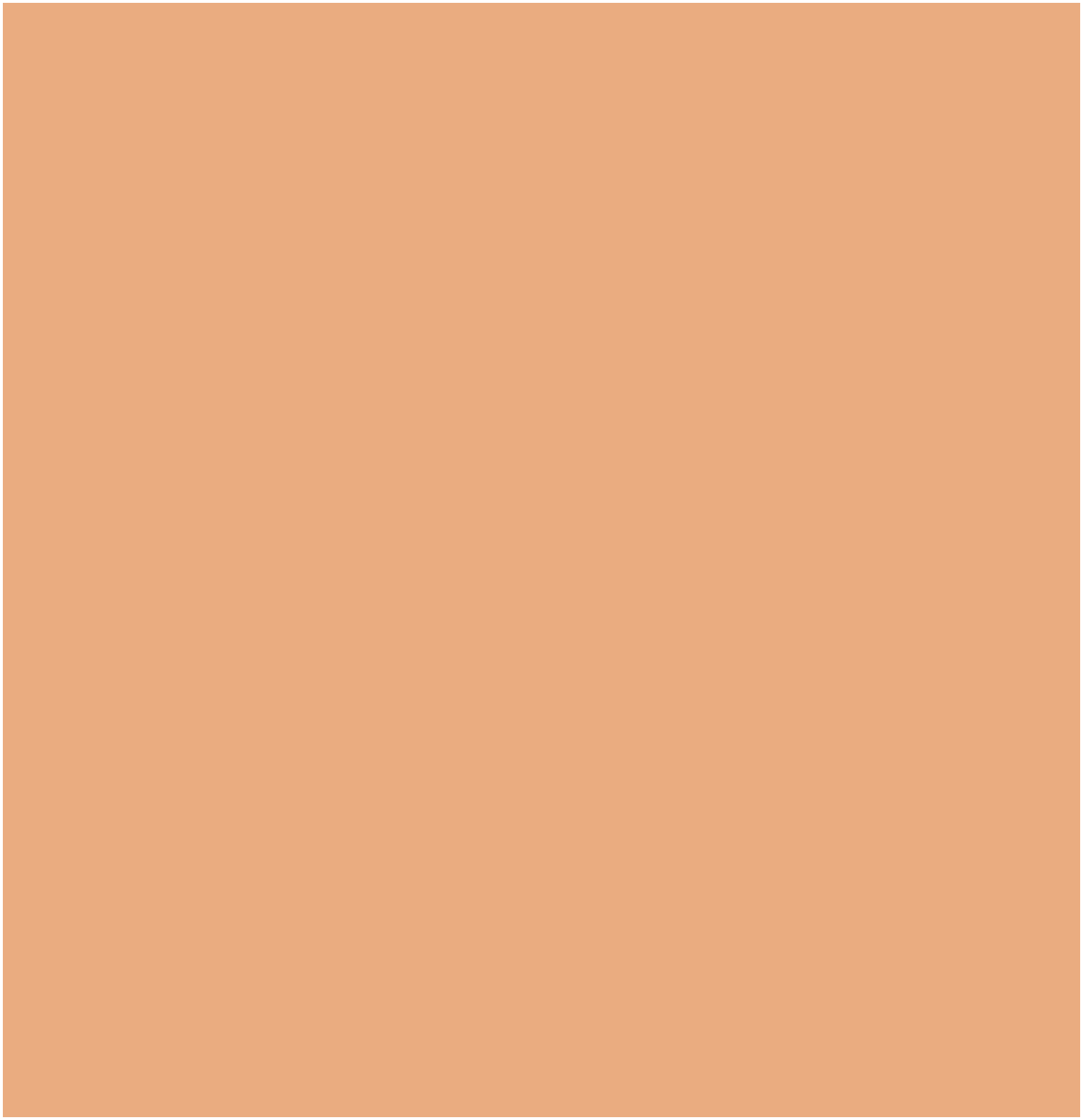
Quantiles for Normal Distribution		
	Quantile	
Pourcentage	Observé	Estimé
1.0	160000	-3926492.5
5.0	250000	-2277357.7
10.0	315000	-1398210.1
25.0	484999	70807.0
50.0	825000	1702991.7
75.0	1925000	3335176.4
90.0	4151000	4804193.5
95.0	5900000	5683341.1
99.0	10900000	7332475.9

nous avons les résultats des tests d'adéquation à une distribution normale (Goodness-of-Fit Tests for Normal Distribution) pour Price. Les tests de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling donnent tous des p-values inférieures au seuil habituel de 0,05, indiquant que la distribution des prix ne suit pas une distribution normale.

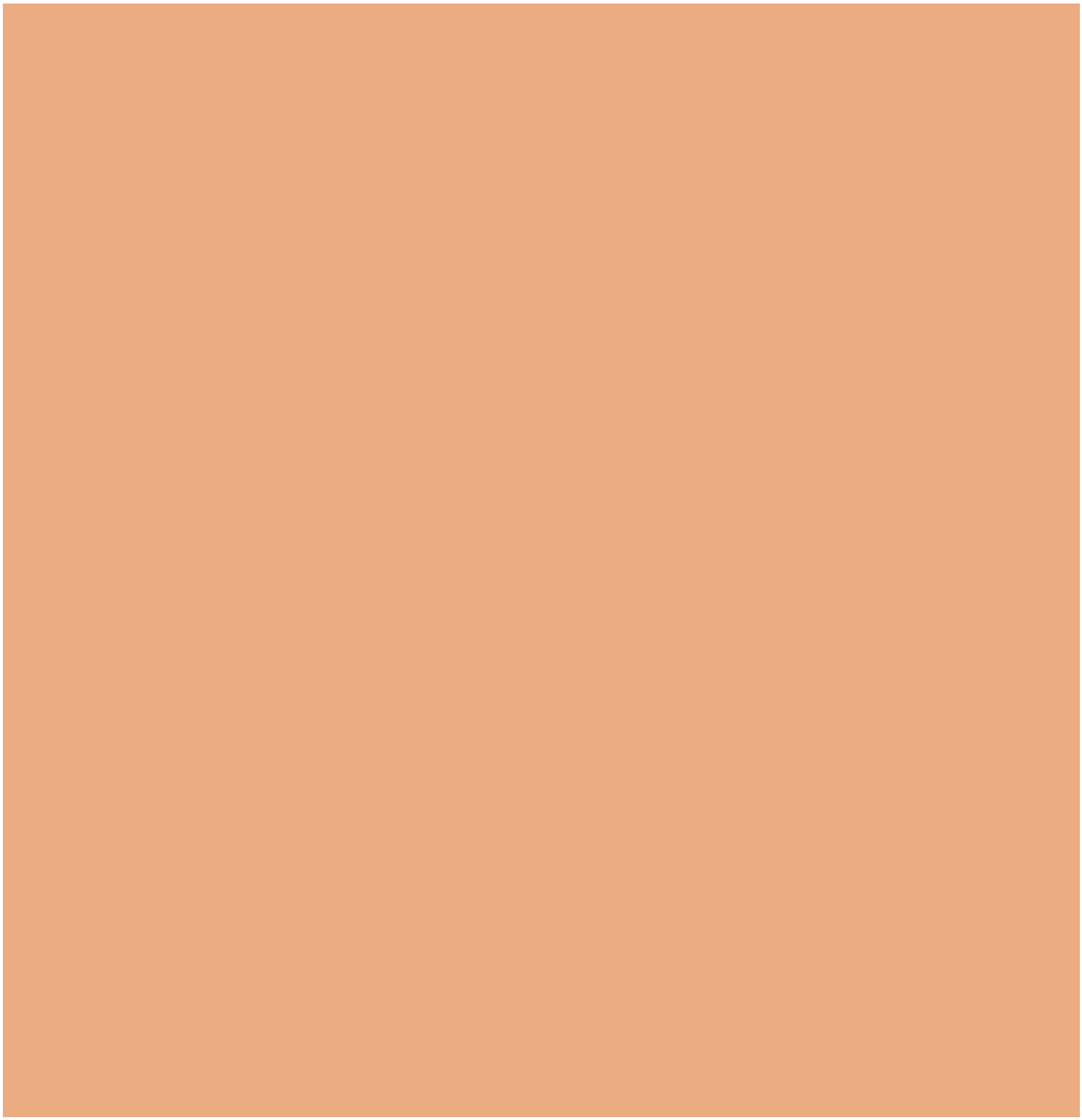
le graphique quantile-quantile (Q-Q Plot) pour Price. Les points ne se situent pas sur la ligne de référence normale, ce qui suggère également que les prix ne sont pas normalement distribués. En particulier, les points forment une courbe plutôt qu'une ligne droite, indiquant une distribution des prix avec une queue lourde - il y a un nombre relativement important de prix très élevés par rapport à ce que prévoirait une distribution normale.



Ensemble, ces résultats suggèrent que Price a une distribution avec une asymétrie positive (skewness) et des valeurs extrêmes sur la droite (kurtosis élevé), typique des données financières où quelques observations peuvent être beaucoup plus hautes que la moyenne, comme c'est souvent le cas avec des biens de luxe ou des articles de collection. En examinant le tableau, il est apparent que les résultats des trois tests de normalité donnent des p-values inférieures à 0,05. Ces résultat conduisent à un rejet del'hypothèse nulle selon laquelle la distribution est normale.



2. TEST D'INDÉPENDANCE ENTRE LES VARIABLES QUALITATIVES



04

LIENS ENTRE « PRICE » ET LES AUTRES VARIABLES

1. CORRÉLATION ENTRE LES VARIABLES QUANTITATIVES

La procédure CORR

8 Avec les variables :	Price Year Kilometer Length Width Height Seating Capacity Fuel Tank Capacity
1 Variables :	Price

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
Price	2059	1702992	2419881	3506459903	49000	35000000
Year	2059	2016	3.36356	4151820	1988	2022
Kilometer	2059	54225	57362	111648687	0	2000000
Length	1995	4281	442.45851	8540317	3099	5569
Width	1995	1768	135.26583	3527144	1475	2220
Height	1995	1592	136.07396	3175512	1165	1995
Seating Capacity	1995	5.30627	0.82217	10586	2.00000	8.00000
Fuel Tank Capacity	1946	52.00221	15.11020	101196	15.00000	105.00000

Coefficients de corrélation de Pearson
Proba > |r| sous H0: Rho=0
Nombre d'observations

	Price
Price	1.00000 2059
Year	0.31140 <.0001 2059
Kilometer	-0.15083 <.0001 2059
Length	0.55674 <.0001 1995
Width	0.56400 <.0001 1995
Height	0.07508 0.0008 1995
Seating Capacity	-0.03852 0.0854 1995
Fuel Tank Capacity	0.58463 <.0001 1946

Les résultats de la corrélation indiquent que les années de fabrication récentes, une longueur et largeur plus importantes, ainsi qu'une plus grande capacité du réservoir de carburant sont associées à des prix plus élevés pour les voitures d'occasion, avec des corrélations significatives allant de modérées à fortes. Inversement, un kilométrage plus élevé est lié à des prix inférieurs. La hauteur et la capacité de sièges montrent des corrélations faibles avec le prix et pourraient ne pas être des indicateurs importants dans la détermination du prix des véhicules d'occasion. Ces variables sélectionnées pourraient être d'excellents points de départ pour une modélisation prédictive plus approfondie des prix des voitures d'occasion.

2. ANALYSE DE VARIANCES APPLIQUÉES AUX DONNÉES CATÉGORIELLES

La procédure GLM (General Linear Model) montre une analyse de la relation entre le type de carburant (Fuel Type) et le kilométrage parcouru (Kilometer) des voitures d'occasion. Le test global de F indique une valeur significative (p < .0001), suggérant qu'il existe une différence statistiquement significative dans le kilométrage moyen en fonction du type de carburant.
Cependant, en regardant les estimations individuelles pour chaque type de carburant, seul le Fuel Type Diesel a une valeur de p significative (< .0001) avec une estimation positive, indiquant que les véhicules diesel ont tendance à avoir un kilométrage plus élevé comparé à la catégorie de référence (le type de carburant "Petrol"). Les autres types de carburant ne montrent pas de différences significatives par rapport au "Petrol" en ce qui concerne le kilométrage, comme en témoignent les valeurs de p supérieures à 0,05.

La procédure GLM

Informations sur les niveaux de classe		
Classe	Niveaux	Valeurs
Fuel Type	7	CNG CNG + Diesel Electr Hybrid LPG Petrol

Nombre d'observations lues	2059
Nombre d'observations utilisées	2059

La procédure GLM

Variable dépendante : Kilometer

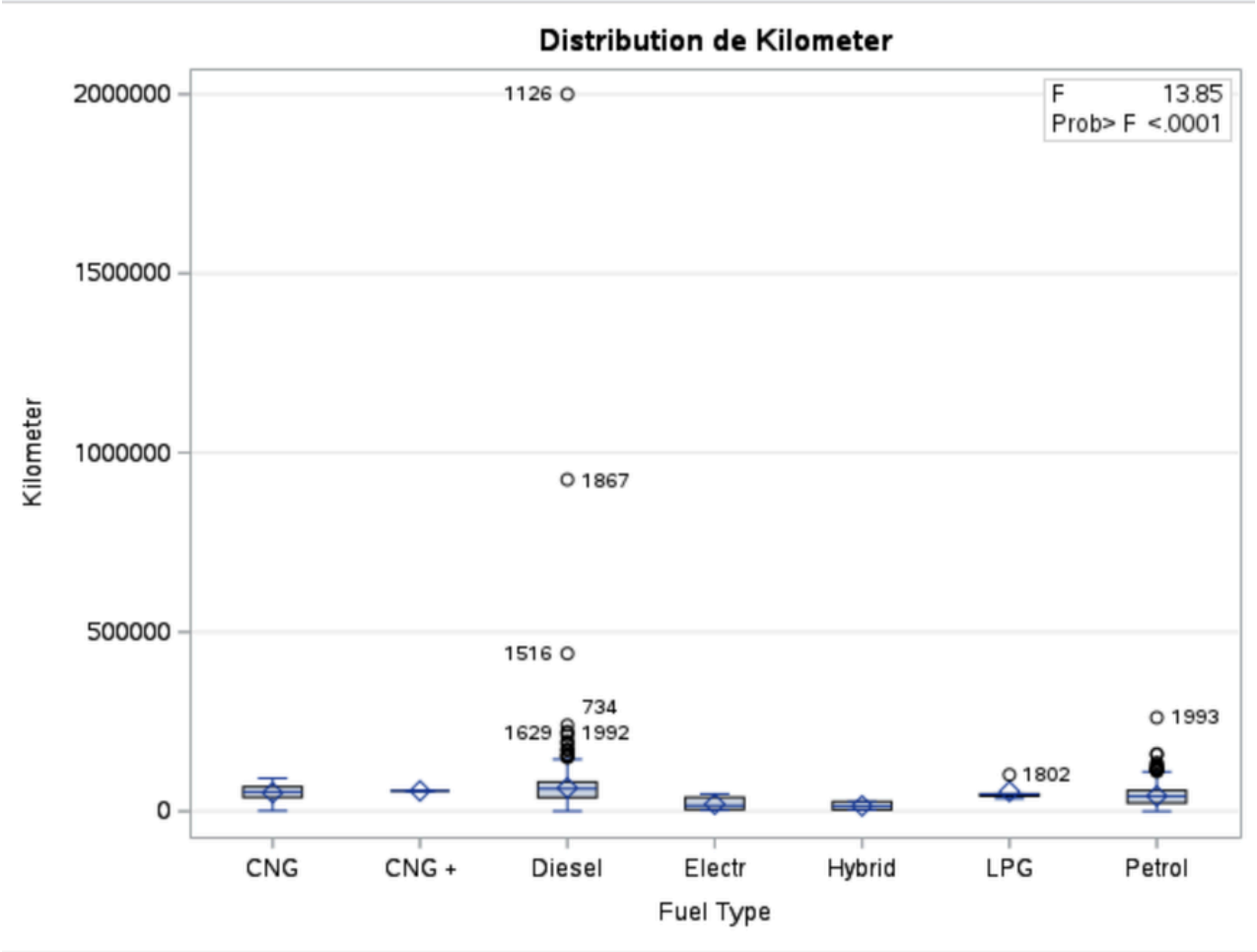
Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	6	263583726393	43930621066	13.85	<.0001
Erreur	2052	6.5079917E12	3171535920.1		
Total sommes corrigées	2058	6.7715754E12			

R-carré	Coef de var	Racine MSE	Kilometer Moyenne
0.038925	103.8574	56316.39	54224.71

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
Fuel Type	6	263583726393	43930621066	13.85	<.0001

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
Fuel Type	6	263583726393	43930621066	13.85	<.0001

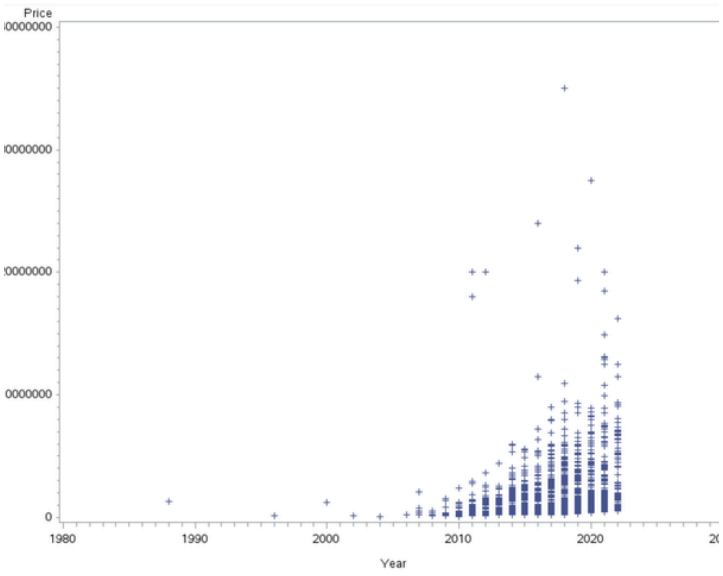
Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	42651.58263	B	1832.94241	23.27	<.0001
Fuel Type CNG	9067.19737	B	8172.53916	1.11	0.2674
Fuel Type CNG +	14191.41737	B	56346.21192	0.25	0.8012
Fuel Type Diesel	22442.49745	B	2526.47337	8.88	<.0001
Fuel Type Electr	-22808.72548	B	21364.36842	-1.07	0.2858
Fuel Type Hybrid	-27651.58263	B	32565.90729	-0.85	0.3959
Fuel Type LPG	12392.61737	B	25252.06649	0.49	0.6237
Fuel Type Petrol	0.00000	B	.	.	.

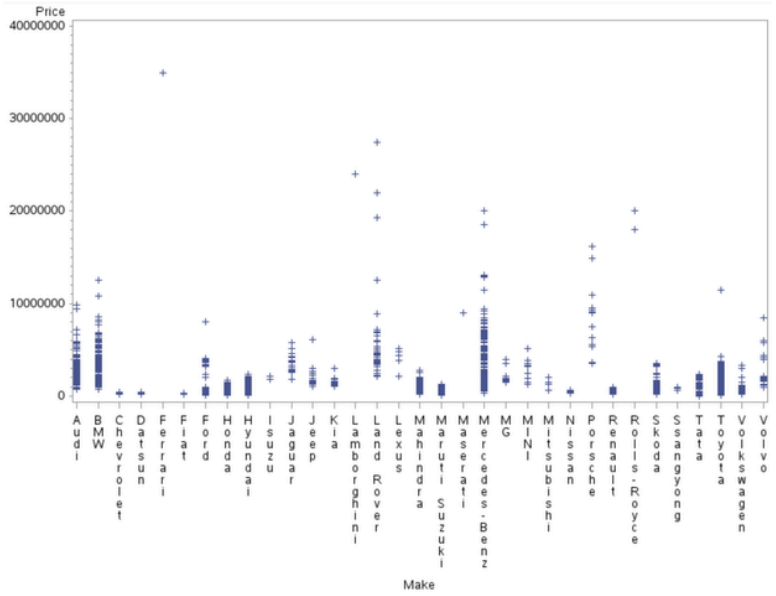


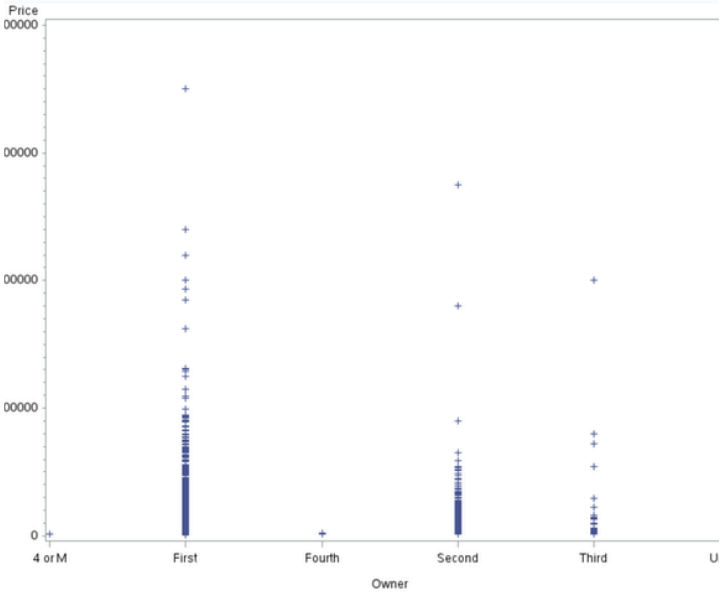
Le boxplot complémentaire illustre visuellement la distribution du kilométrage par type de carburant. Les véhicules diesel semblent avoir une plus grande variabilité dans le kilométrage et des valeurs supérieures par rapport aux autres types de carburant, ce qui corrobore l'analyse statistique indiquant une différence significative pour cette catégorie. En résumé, les véhicules diesel se distinguent par un kilométrage supérieur, ce qui peut refléter leur popularité et leur durabilité dans des utilisations nécessitant de longues distances.

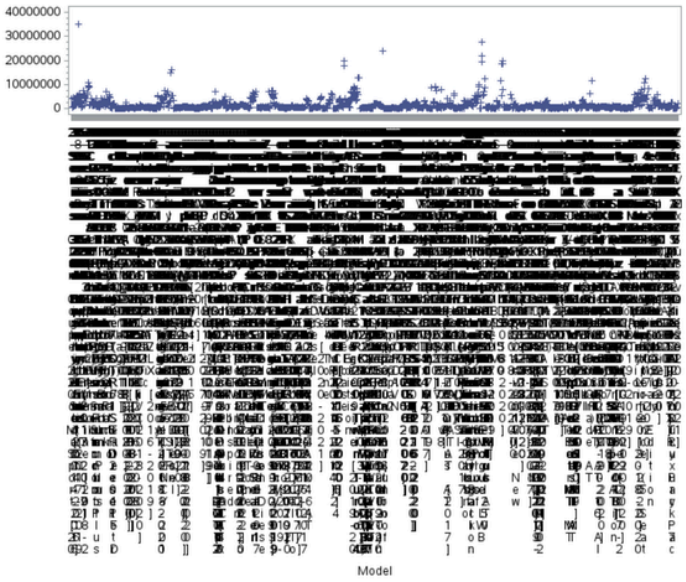
3. ANALYSE BIVARIÉE

En utilisant la fonction « gplot », il est possible d'afficher un graphiqu en nuage de points basé sur deux variables. La variable principale utilisée est le logarithme de price . Cela nous permettra de visualiser d'éventuelles corrélations entre les variables.









MODÈLES DE RÉGRESSION

05

1. RÉGRESSION LINÉAIRE SIMPLE

Le tableau présente les résultats initiaux d'une procédure de régression stepwise, où l'année de fabrication (Year) est la première variable sélectionnée par le modèle. La méthode stepwise est un outil analytique statistique utilisé pour sélectionner des variables significatives dans la construction d'un modèle prédictif. Elle élimine et inclut des variables de manière itérative basée sur des critères statistiques spécifiques.
L'étape 1 indique que la variable Year a été introduite dans le modèle. Il n'y a pas de variable éliminée à cette étape, ce qui suggère que Year est potentiellement un prédicteur important. Le R carré partiel de Year est de 0.0970, ce qui signifie que cette seule variable explique approximativement 9.7% de la variance totale du prix des voitures d'occasion. Le R carré pour le modèle entier est également de 0.0970 à cette étape, ce qui confirme que Year est la seule variable dans le modèle à ce stade.

Synthèse de Sélection Stepwise

Etape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	Year		1	0.0970	0.0970	2.0000	220.89	<.0001

Cependant, l'objectif de parvenir à un R carré de 0,95 n'est pas atteint avec cette seule variable. Cela indique que d'autres variables significatives sont nécessaires pour améliorer la capacité du modèle à expliquer la variance des prix des voitures d'occasion. Le processus de sélection stepwise devrait donc continuer à identifier et inclure d'autres variables significatives pour améliorer la performance du modèle. En éliminant progressivement les variables non significatives et en conservant celles qui sont pertinentes, on peut s'attendre à ce que d'autres variables deviennent importantes dans les étapes suivantes, contribuant ainsi à augmenter le R carré du modèle vers l'objectif visé.

2. RÉGRESSION MULTIPLE

La régression linéaire multiple est une extension de la régression linéaire simple, où le modèle inclut plusieurs variables explicatives pour prédire la variable de sortie. Le principe de la régression linéaire multiple consiste à trouver une relation linéaire entre la variable de sortie et les variables explicatives en ajustant les coefficients du modèle de manière à minimiser l'erreur de prédiction.

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-412063817	24936429	-16.52	<.0001
Height	1	-1179.40008	446.71756	-2.64	0.0084
Year	1	204753	12517	16.36	<.0001
Kilometer	1	-3.18654	0.65399	-4.87	<.0001
Length	1	-142.36507	194.14013	-0.73	0.4635
Width	1	1040.17793	528.53363	1.97	0.0492
Engine_num	1	1852.13536	109.25452	16.95	<.0001
Seating Capacity	1	-664497	68233	-9.74	<.0001
Fuel Tank Capacity	1	40083	5194.17841	7.72	<.0001

Le tableau issu de la procédure REG en SAS présente les résultats d'une régression linéaire multiple pour la variable dépendante Price avec plusieurs variables explicatives. Le R carré du modèle est de 0.5632, ce qui indique que le modèle explique environ 56.32% de la variance de la variable Price. C'est un indice assez élevé, mais pas aussi élevé que le 0.7121 mentionné précédemment dans votre exemple d'analyse.
Le R carré ajusté, qui prend en compte le nombre de prédicteurs dans le modèle, est très proche du R carré non ajusté, suggérant que les variables incluses ont une justification statistique dans le modèle. Les valeurs p associées à chaque coefficient nous indiquent l'importance statistique de chaque variable. Les variables avec une p-value inférieure à 0.05 sont généralement considérées comme ayant une influence significative sur la variable dépendante.
Dans ce cas, Height, Year, Kilometer, Engine_num, Seating Capacity, et Fuel Tank Capacity ont tous des p-valeurs significatives ($p < 0.05$), ce qui suggère que ces prédicteurs ont une influence statistiquement significative sur le prix des voitures d'occasion. En revanche, les variables Length et Width ont des p-valeurs de 0.4635 et 0.0492, respectivement. Bien que la largeur (Width) soit à la limite de la signification statistique, la longueur (Length) n'est pas statistiquement significative dans ce modèle, et on pourrait envisager de l'éliminer dans des analyses ultérieures pour améliorer la pertinence du modèle.

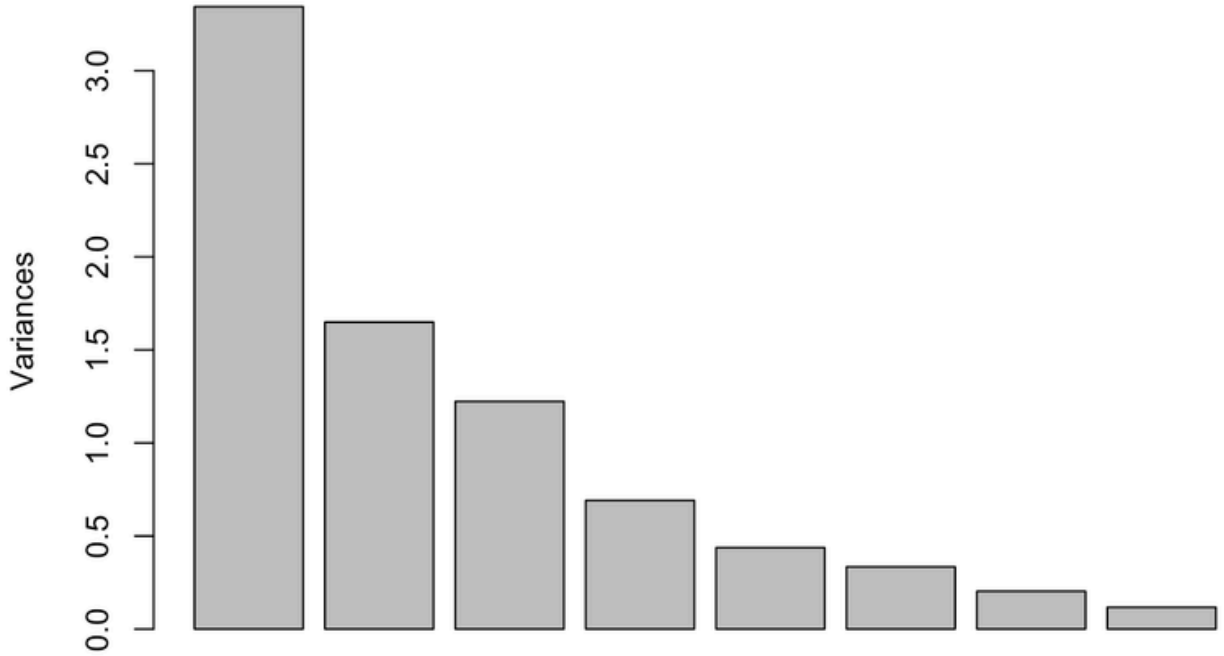
09

ANALYSE DES COMPOSANTES PRINCIPALES ET CLASSIFICATION

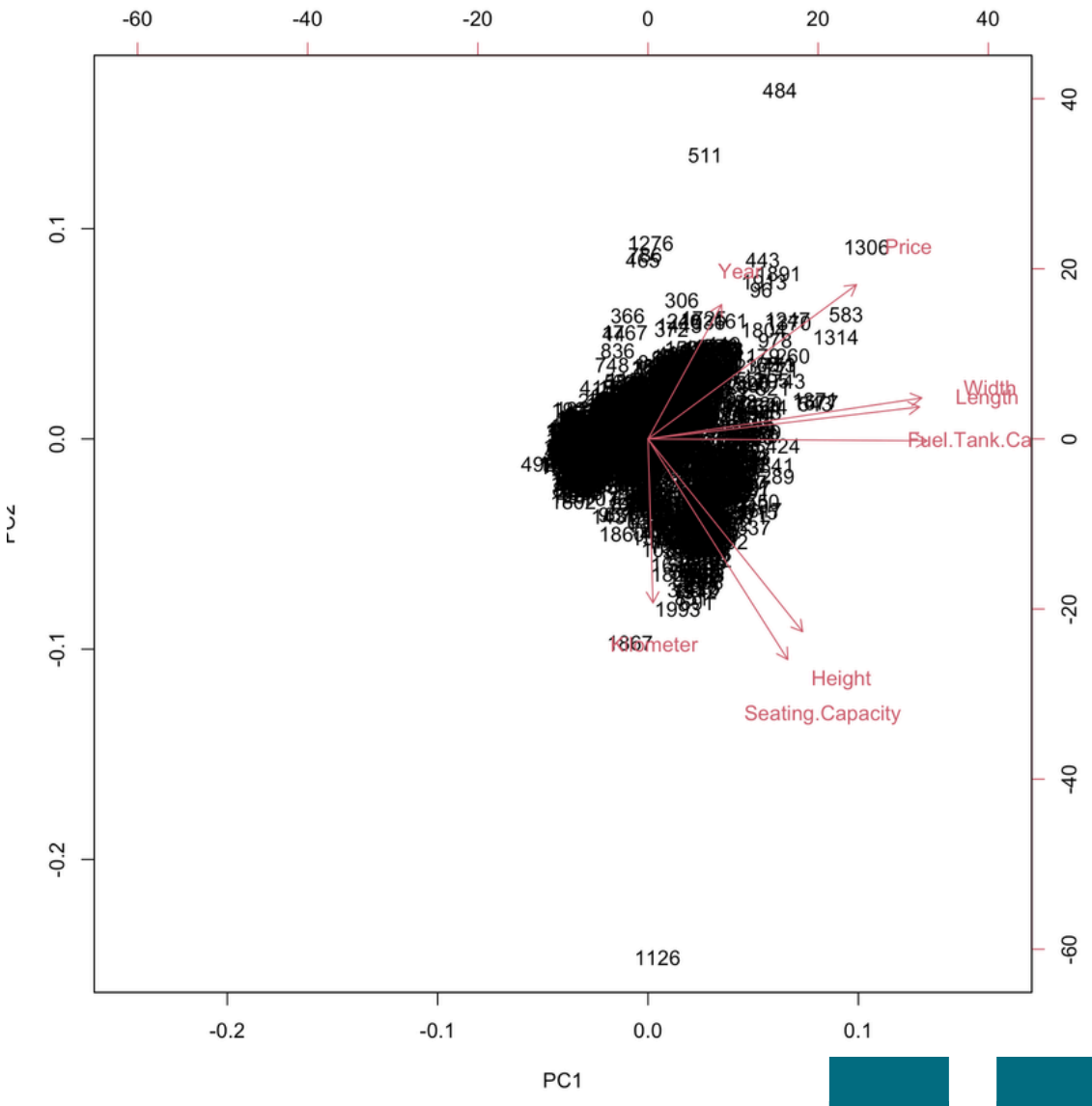
1. REDUCTION DE LA DIMENSIONNALITÉ AVEC L'ACP

Notre objectif dans cette partie est de connaître les variables qui représentent les composantes principales de notre jeu de données. Pour ce faire, nous avons utilisé la commande « prcomp » sur R. Cette commande est uniquement applicable sur les variables quantitatives, de ce fait, nous avons exclu les variables qualitatives

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6     PC7     PC8
Standard deviation     1.829 1.2843 1.1060 0.83135 0.66144 0.5783 0.45061 0.34212
Proportion of Variance 0.418 0.2062 0.1529 0.08639 0.05469 0.0418 0.02538 0.01463
Cumulative Proportion  0.418 0.6242 0.7771 0.86350 0.91819 0.9600 0.98537 1.00000
```

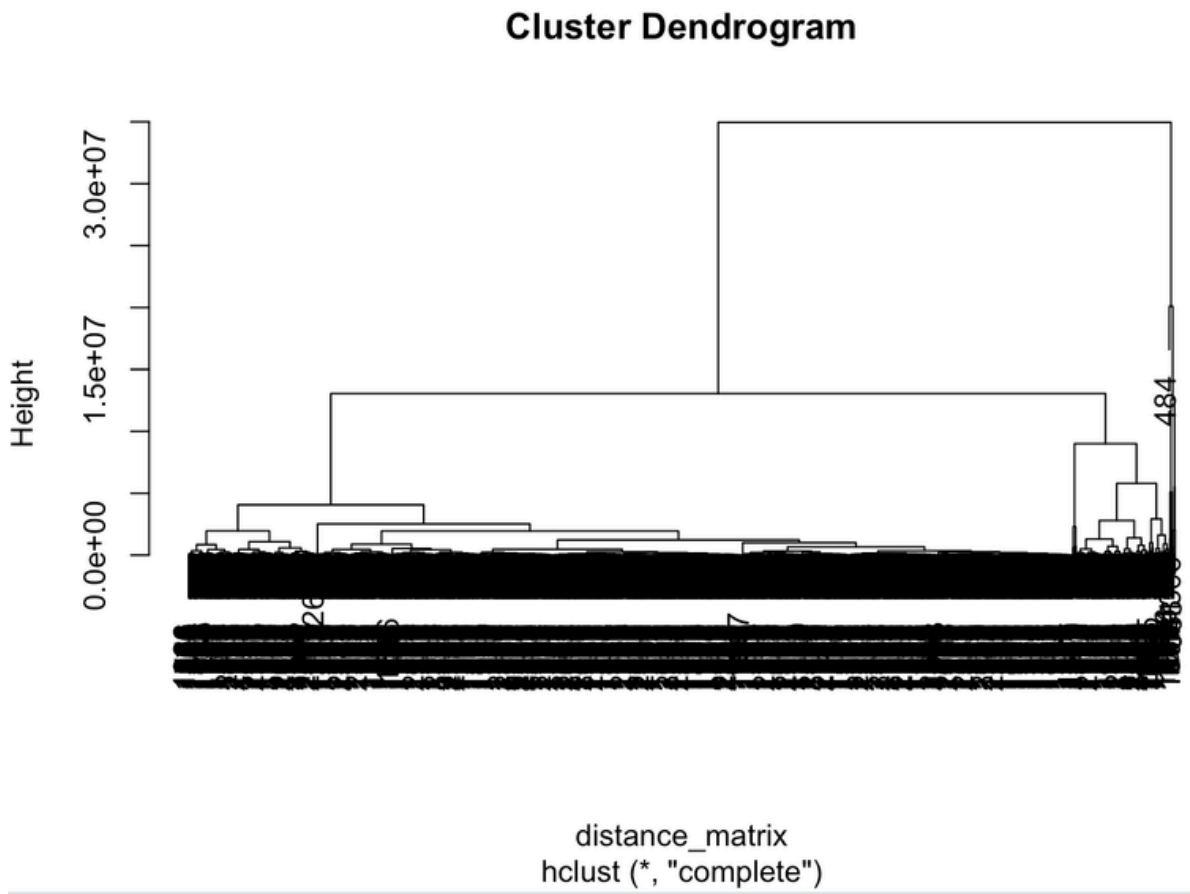


On applique ensuite la commande « biplot » pour créer un graphique contenant à la fois les observations et les analyses factorielles de données. Cela nous permettra de trouver les relations entre les variables et les observations.



2. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

La Classification Ascendante Hiérarchique est une technique statistique visant à mettre une population dans différentes classes ou sous-groupes. Elle nous permettra de regrouper nos observations en fonction des variables.



CONCLUSION

07

OUSMANE TALL

MERCI

KHADIDJA LINA MOULAI