Rapport de projet

Data Mining

Membres du groupe :

Cheddadi Radja
Abichou Nour Elhouda

2022/2023

Table de matière :

Introduction

1. **Analyse exploratoire des données**

   1.1 Analyse univariée et bivariée

   1.2 Matrice de corrélation des variables quantitatives

   1.3 Analyse statistique du chi deux

   1.4 Preprocessing et sélection des variables

2. **Analyse multidimensionnelle**

   2.1 Analyse par composantes principales

   2.2 Analyse par composantes multiples

3. **Modélisation**

   3.1 Modèle non supervisé

   3.2 Modèles supervisés

Conclusion

Annexes

Introduction

L'apprentissage automatique est de plus en plus utilisé dans de nombreux domaines pour analyser de grandes quantités de données. Dans ce projet, nous allons utiliser des algorithmes d'apprentissage automatique pour analyser un ensemble de données de voitures d'occasion disponibles à la vente sur Craigslist. L'objectif est de regrouper les véhicules en fonction de leur prix, puis de prédire le prix d'un véhicule en fonction de ses caractéristiques telles que le modèle, l'année, le fabricant, le type de transmission, etc.

Ce projet est basé sur un ensemble de données de voitures d'occasion disponibles à la vente sur Craigslist, qui comprend chaque entrée de voiture d'occasion aux États-Unis. Nous allons donc utiliser ces données pour développer nos modèles de prédiction.

Pour atteindre cet objectif, nous allons suivre une méthodologie en plusieurs étapes. Tout d'abord, nous allons effectuer une analyse exploratoire des données pour mieux comprendre la structure de notre ensemble de données. Nous allons examiner la distribution des variables et les relations entre elles en utilisant une analyse univariée et bivariée et une analyse statistique du chi-carré

Ensuite, nous allons utiliser une analyse multidimensionnelle pour explorer la structure sous-jacente de l'ensemble de données. Nous allons utiliser une analyse par composantes principales (ACP) et une analyse par composantes multiples (ACM) pour identifier les variables les plus importantes qui contribuent à la variabilité de l'ensemble de données.

Enfin, nous allons construire des modèles non supervisés et supervisés pour regrouper les véhicules en fonction de leur prix et prédire le prix d'un véhicule en fonction de ses caractéristiques. Nous allons utiliser une régression multiple simple, une régression logistique et un algorithme de gradient boosting pour construire des modèles de prédiction.

Dans cette étude, nous allons répondre à la problématique suivante : Comment utiliser des algorithmes d'apprentissage automatique pour regrouper les véhicules en fonction de leur prix et prédire le prix d'un véhicule en fonction de ses caractéristiques telles que le modèle, l'année, le fabricant, le type de transmission, etc. ?

1. Analyse exploratoire des données

Afin de comprendre les relations entre les variables de notre ensemble de données nous allons effectuer une analyse exploratoire de données. Nous allons commencer par vérifier les types de données de chaque colonne du dataset. Cela nous aidera à déterminer si des conversions de type de données sont nécessaires.

En effet, pour les modèles d'apprentissage supervisé, nous devons généralement convertir toutes les variables catégorielles en données numériques à l'aide de techniques telles que l'encodage à un coup ou l'encodage des étiquettes. En effet, la plupart des algorithmes d'apprentissage automatique fonctionnent avec des données numériques, et les variables catégorielles sous leur forme brute ne peuvent pas être utilisées en entrée.

Pour les modèles d'apprentissage non supervisé, la nécessité de convertir les types de données dépend de l'algorithme spécifique utilisé et de la nature des données. Par exemple, si nous travaillons avec des algorithmes de clustering, nous devrons peut-être mettre à l'échelle ou normaliser des données numériques pour nous assurer que les différentes caractéristiques sont sur une échelle comparable.

On constate que la plupart des variables sont de type character, ce qui signifie qu'elles contiennent des valeurs textuelles. Les variables numeriques représentent des valeurs numériques tandis que integer représente des entiers.

Corrélation de l'année avec le reste variables quantitatives

L'année présente une faible corrélation négative avec l'identifiant (-0,059), ce qui est normal puisque l'identifiant n'est qu'un identifiant et n'a pas de lien réel avec l'année.

Le compteur kilométrique présente une faible corrélation négative avec l'année (-0,157), ce qui suggère que les voitures plus anciennes ont tendance à avoir un kilométrage plus élevé.

La latitude et la longitude ont une très faible corrélation avec l'année (-0,015 et -0,001, respectivement), ce qui suggère qu'il n'y a pas de véritable modèle géographique pour l'âge des voitures sur Craigslist.

Corrélation du compteur kilométrique avec le reste variables quantitatives

Le compteur kilométrique présente une faible corrélation positive avec l'ID (0,011), ce qui suggère qu'il pourrait y avoir un lien entre l'ordre dans lequel les voitures ont été répertoriées et leur kilométrage.

La latitude et la longitude ont une très faible corrélation avec le compteur kilométrique (0,0098 et 0,010, respectivement), ce qui suggère qu'il n'y a pas de véritable modèle géographique pour le kilométrage des voitures sur Craigslist.

Corrélation de la Latitude et longitude avec le reste variables quantitatives :

La latitude et la longitude ont une corrélation négative modérée l'une avec l'autre (-0,128), ce qui est attendu puisqu'elles sont toutes deux des mesures de l'emplacement géographique.

La latitude et la longitude ont toutes deux de faibles corrélations négatives avec l'ID (-0,069 et -0,122, respectivement), ce qui suggère qu'il peut y avoir un certain regroupement géographique des données sur les ventes de voitures.

Dans l'ensemble, ces corrélations sont relativement faibles, ce qui signifie qu'il n'existe pas de relation linéaire forte entre les variables. Cela suggère que d'autres facteurs, tels que l'état de la voiture, la localisation du vendeur et la période de l'année, peuvent être plus importants pour déterminer le prix et d'autres caractéristiques des voitures sur Craigslist.

1.1 Analyse univariée et bivariée

Nous allons ensuite effectuer une analyse univariée et bivariée. Le but de l'analyse univariée étant d'explorer et de décrire les caractéristiques d'une seule variable de notre dataset à la fois, tandis que le but de l'analyse bivariée est de déterminer s'il existe une relation entre deux variables différentes, dans notre cas nous allons nous concentrer sur l'existence d'une association entre le prix et le reste des variables.

La première étape d'une analyse univariée sera de visualiser les données afin de mieux comprendre leur répartition. Une fois que nous aurons une compréhension visuelle des données, nous pourrons procéder à une analyse plus approfondie.

Il est à noter que nous avons filtré les valeurs aberrantes des prix, en supprimant tous les prix supérieurs à 250 000 $, ainsi que les observations vides et nulles. Il y a 426880 observations dont 95 avec des prix supérieurs à 250 000 $, nous considérerons donc les 95 observations comme des valeurs aberrantes. En supprimant les valeurs aberrantes, nous avons limité l'impact de données extrêmes sur les résultats de l'analyse. Les valeurs aberrantes peuvent être causées par des erreurs de saisie, des données manquantes ou des observations atypiques, et leur inclusion peut fausser les résultats de l'analyse en donnant une vision faussée de la distribution des données.

En supprimant les observations vides et nulles, nous avons également amélioré la qualité de nos données en éliminant les données manquantes. Les données manquantes peuvent être causées par des erreurs de saisie, des erreurs de mesure ou des erreurs d'enregistrement, et leur inclusion peut rendre les résultats de l'analyse moins fiable

Analyse univariée de la variable Prix

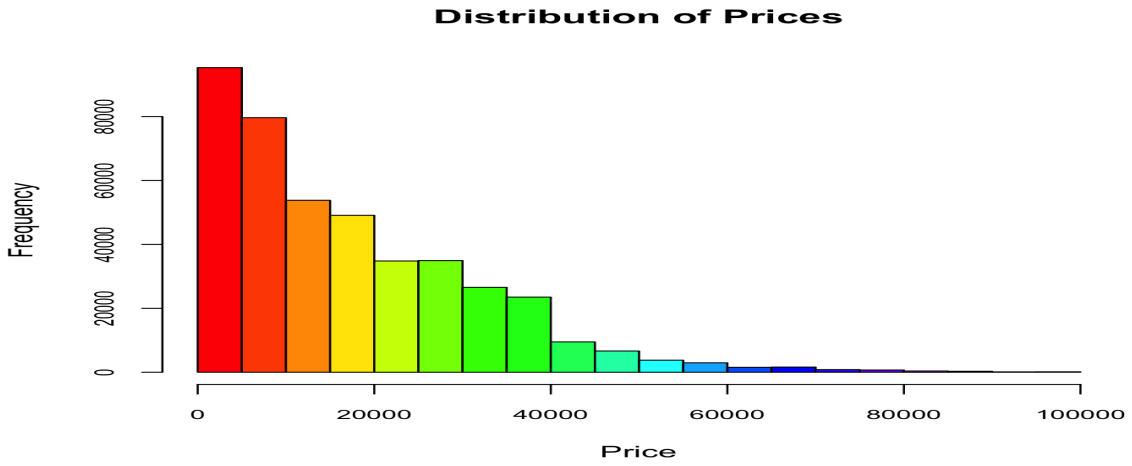
Les graphiques 01 et 02 nous montrent que le prix minimum pour une voiture est de 0, ce qui est probablement un résultat non valide ou un manque d'informations sur le prix. Par ailleurs, le prix médian pour une voiture est de 13 950, ce qui signifie que la moitié des voitures sont vendues à un prix inférieur à cette valeur et l'autre moitié à un prix supérieur.

La moyenne des prix pour une voiture est de 75 200, ce qui est beaucoup plus élevé que le prix médian. Cela peut indiquer la présence d'un petit nombre de voitures très chères sur le marché qui tirent la moyenne vers le haut.

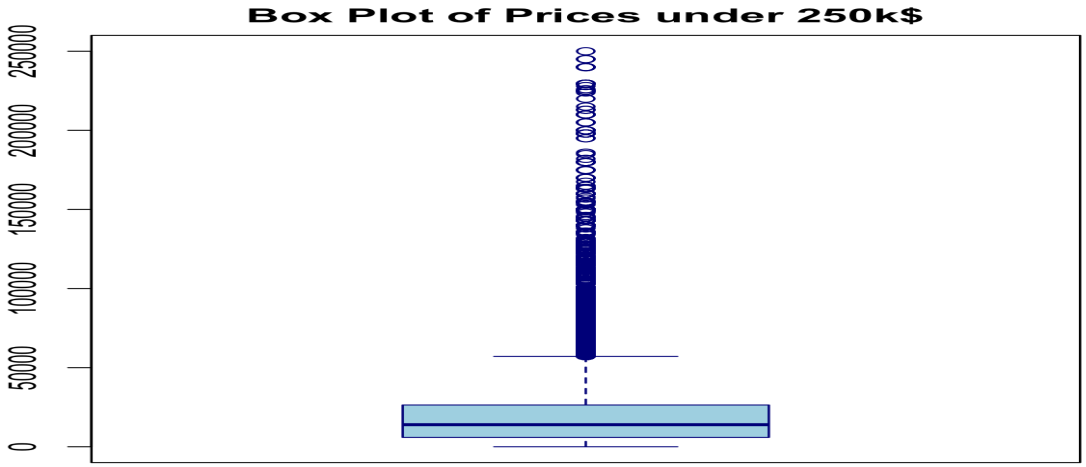
Le 1er quartile des prix pour une voiture est de 5 900, ce qui indique que 25 % des voitures sont vendues à un prix inférieur à cette valeur. Le 3ème quartile des prix pour une voiture est de 26 490, ce qui indique que 75 % des voitures sont vendues à un prix inférieur à cette valeur. Le prix maximum

pour une voiture est de 3,737 milliards, ce qui est probablement une erreur ou une anomalie, car cela représente un prix extrêmement élevé pour une voiture sur Craigslist.

Graphique 02 : distribution des prix



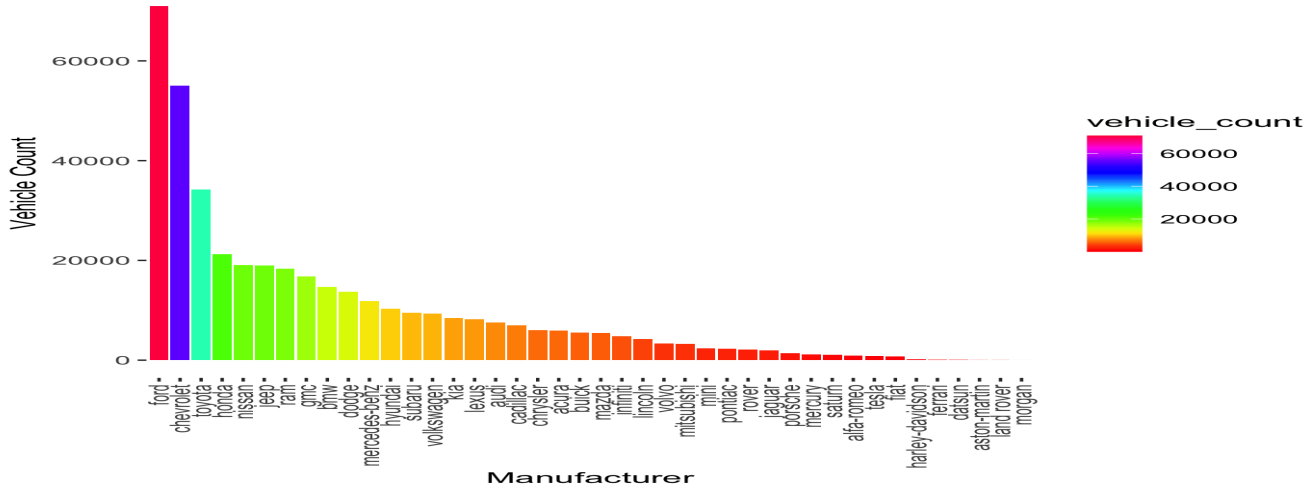
Graphique 03 : Boxplot des prix



Analyse univariée de la variable Manufacturer

Le graphique 04 nous montre qu'Il y a un nombre important de voitures américaines en vente sur Craigslist, avec des fabricants tels que Ford, Chevrolet, Dodge et Jeep représentant une part importante du marché. Les constructeurs allemands BMW, Audi, Mercedes-Benz et Volkswagen sont également bien représentés, tandis que les constructeurs japonais comme Toyota, Honda et Nissan sont également populaires. Les fabricants italiens Alfa Romeo et Ferrari sont également présents, bien que dans une mesure beaucoup moins importante. Enfin, il y a un certain nombre de fabricants avec un nombre très limité de voitures en vente, comme Morgan et Aston Martin.

Graphique 04 : Distribution des voitures par marque



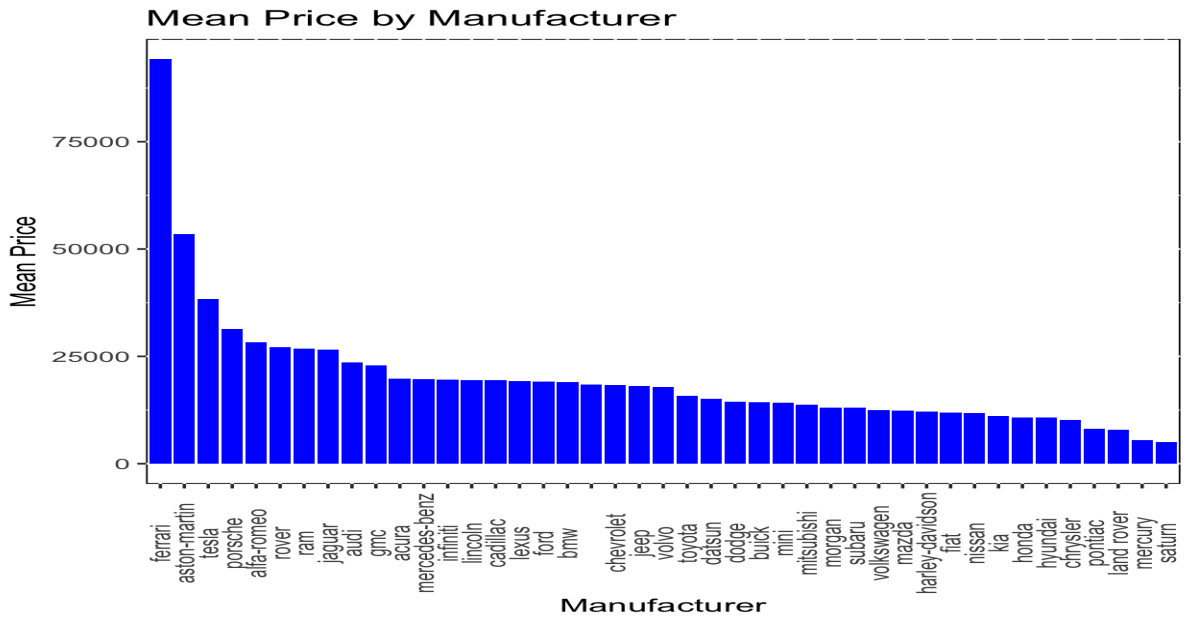
Analyse bivariée de la variable Manufacturer par rapport au prix moyen

Le graphique 05 montre que les prix moyens des véhicules disponibles sur Craigslist varient considérablement d'un constructeur à l'autre. Le constructeur le plus cher en moyenne est Aston Martin, avec un prix moyen de plus de 53 000 $. Ferrari a également un prix moyen très élevé d'environ 94 000 $. En revanche, Saturn a le prix moyen le plus bas, avec environ 5 000 dollars, tandis que Land Rover est le deuxième constructeur le moins cher, avec un peu plus de 7 900 dollars.

Nous pouvons également constater que certaines des marques automobiles les plus populaires, telles que Ford et Toyota, se situent quelque part au milieu de la fourchette de prix, avec des prix moyens d'environ 19 000 et 15 000 dollars respectivement. Il est intéressant de noter que certaines marques de voitures de luxe, telles que Mercedes-Benz et Lexus, se situent également dans cette même fourchette de prix, ce qui indique qu'il existe un large éventail de prix, même au sein d'une même catégorie de voitures.

Dans l'ensemble, ce graphique peut fournir des informations précieuses aux acheteurs et vendeurs potentiels de voitures sur Craigslist en donnant un aperçu des prix moyens des différentes marques de voitures.

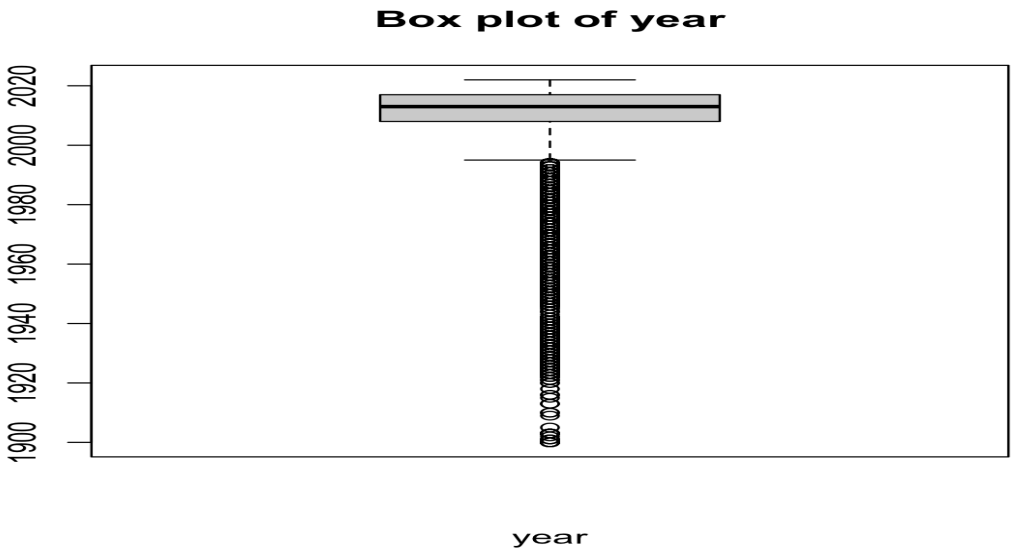
Graphique 05 : Distribution des marques par prix moyen



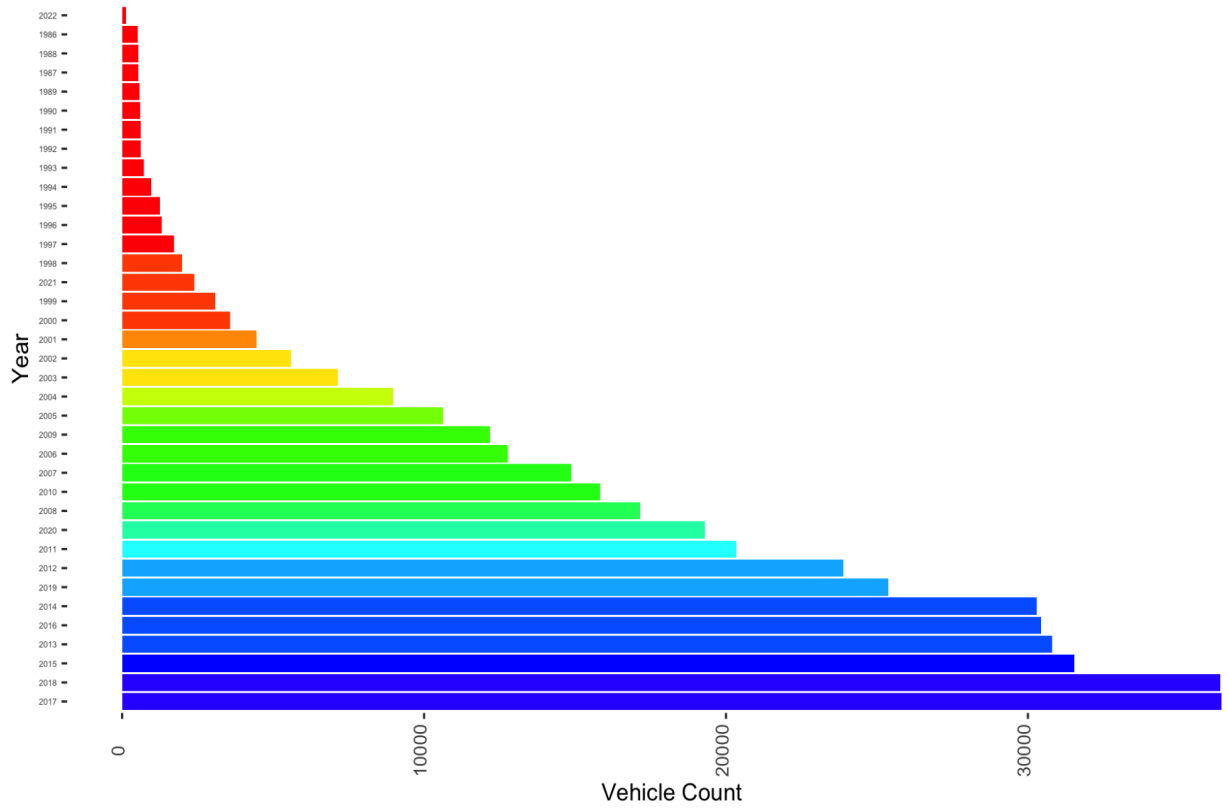
Analyse univariée de la variable Year

Nous pouvons constater que la répartition des véhicules vendus sur Craigslist par année de fabrication est biaisée vers la droite, la majorité des véhicules datant des années les plus récentes et se condense à partir du début des années 2000 avec une valeur maximale du nombre de voitures vendues et qui ont été fabriquées en 2017. Dans l'ensemble, ces données peuvent donner un aperçu des tendances et de la popularité des différents types de véhicules au fil du temps, ainsi que du marché des voitures d'occasion sur Craigslist.

Graphique 06 : boxplot du nombre de véhicules par année



Graphique 07 : Distribution du nombre de véhicules par année



Analyse bivariée de la variable Year par rapport au prix moyen

Le graphique 08 montre que les prix moyens des voitures varient considérablement d'une décennie de fabrication à l'autre. Les voitures fabriquées entre 1910 et 1920 ont le prix moyen le plus élevé, soit 21 890 $. Cela s'explique probablement par la rareté et l'importance historique des voitures de cette époque. En revanche, les voitures fabriquées entre 1900 et 1910 affichent le prix moyen le plus bas (4 188 $). Cela s'explique probablement par le fait que les voitures de cette époque sont extrêmement rares et difficiles à trouver en bon état.

Les voitures fabriquées entre 1930 et 1940 ont le deuxième prix moyen le plus élevé, soit 26 314 $. Cela s'explique probablement par le fait que les voitures de cette époque sont considérées comme des voitures classiques et sont très recherchées par les collectionneurs.

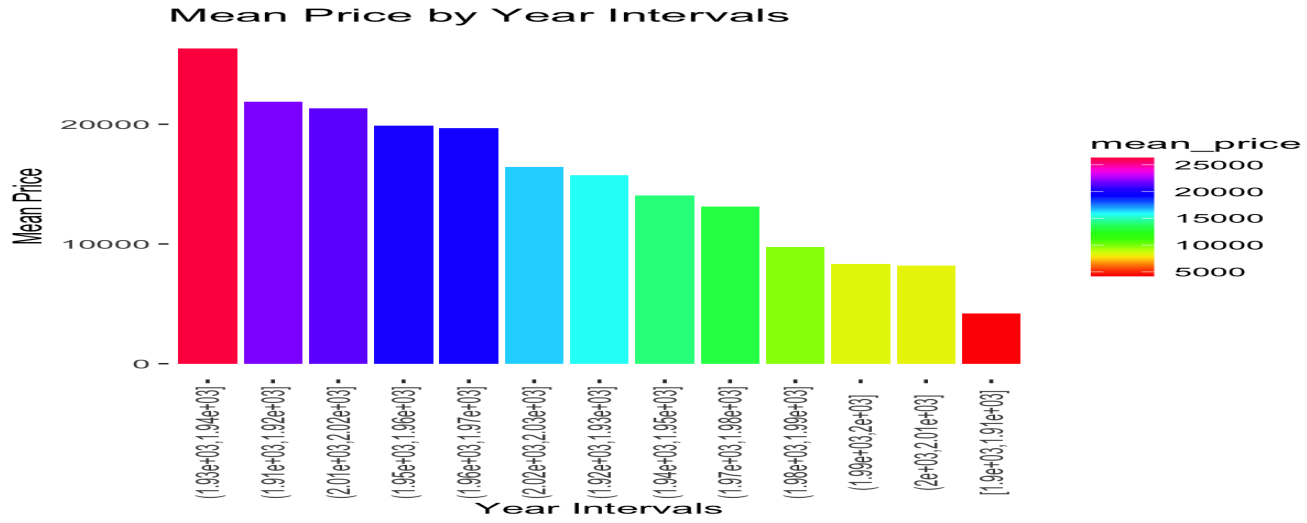
Il est intéressant de noter que les voitures fabriquées entre 1970 et 1980 ont un prix moyen relativement bas de 13 110 $. Cela s'explique probablement par le fait que les voitures de cette époque ne sont pas encore considérées comme des voitures classiques et n'ont donc pas autant de valeur que les voitures des décennies précédentes.

Les voitures fabriquées entre 2000 et 2010 ont un prix moyen de 8 194 $, ce qui est relativement bas par rapport aux voitures des autres décennies. Cela s'explique probablement par le fait que les voitures de cette époque sont encore relativement récentes et n'ont pas encore atteint le statut de voitures classiques.

Enfin, il convient de noter que les voitures fabriquées entre 2010 et 2022 ont un prix moyen relativement élevé de 21 347 $. Cela peut s'expliquer par le fait que ces voitures sont encore relativement nouvelles et donc très demandées, ou par d'autres facteurs tels que les progrès technologiques et l'amélioration des dispositifs de sécurité.

Dans l'ensemble, ces données fournissent des informations précieuses sur la valeur des voitures au cours des différentes décennies de fabrication et peuvent être utilisées par les passionnés d'automobile et les collectionneurs pour prendre des décisions plus éclairées en matière d'achat et de vente de voitures.

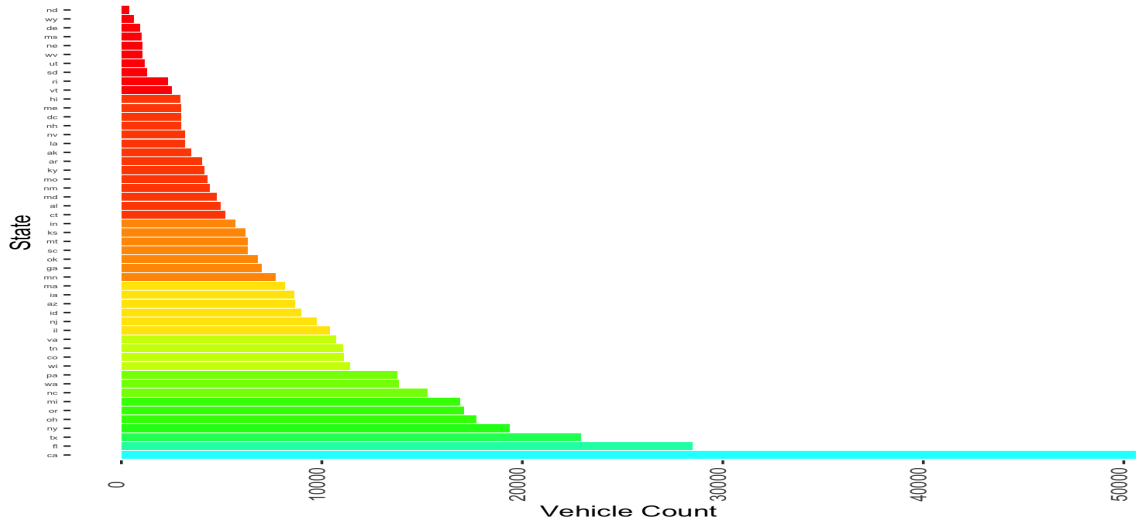
Graphique 08 : Distribution du prix moyen de véhicules par intervalle d'année de 10 ans



Analyse univariée de la variable State

La répartition du nombre de véhicules vendus sur le site Craigslist par état montre une grande variabilité entre les différents états américains. Les États les plus peuplés comme la Californie, New York et la Floride ont vendu le plus grand nombre de véhicules, avec respectivement 50 614, 19 386 et 28 511 unités vendues. D'autres États avec des populations plus modestes ont également connu un volume de ventes élevé, comme le Colorado, l'Oregon et la Pennsylvanie, avec respectivement 11 088, 17 104 et 13 753 unités listées pour vente. En revanche, certains états ont vu un nombre de voitures proposées à la vente sur Craigslist relativement faibles, avec moins de 2 000 unités vendues dans les cas les plus extrêmes. En somme, cette répartition montre que le marché des voitures d'occasion est dynamique et varie considérablement selon l'endroit où l'on se trouve.

Graphique 09 : Distribution du nombre de véhicules par Etat

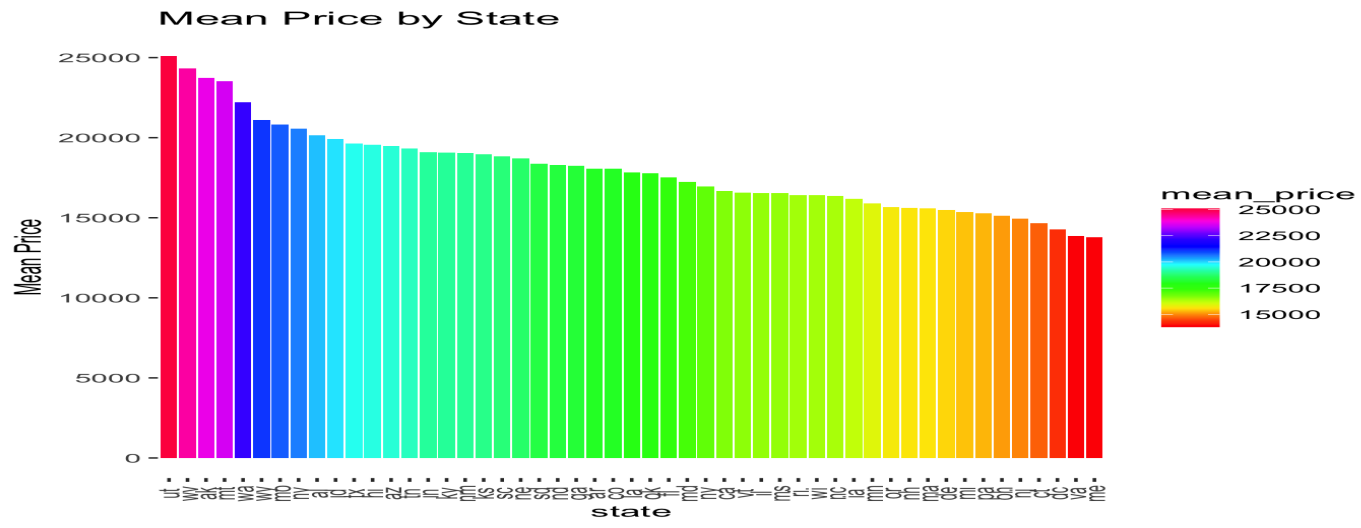


Analyse bivariée de la variable state par rapport au prix moyen

Les données relatives aux prix moyens des véhicules vendus sur Craigslist par État et observées dans le graphique 10 permettent de faire les observations suivantes. L'État de l'Utah a le prix moyen le plus élevé pour les véhicules vendus, tandis que le Maine a le prix moyen le plus bas. Les États où les prix moyens sont les plus élevés sont généralement situés dans l'ouest des États-Unis, tandis que les États où les prix moyens sont les plus bas sont généralement situés dans l'est des États-Unis.

L'État du Montana a le prix moyen le plus élevé parmi les États de l'Ouest, tandis que l'État du Vermont a le prix moyen le plus élevé parmi les États de l'Est. Les États de Californie et de New York, qui sont tous deux situés sur les côtes, ont des prix moyens relativement plus bas que les autres États de leurs régions respectives. Les prix moyens varient considérablement d'un État à l'autre, le prix moyen le plus élevé étant presque le double du prix moyen le plus bas. Dans l'ensemble, ces observations suggèrent que les prix moyens des véhicules vendus sur Craigslist varient considérablement d'un État à l'autre.

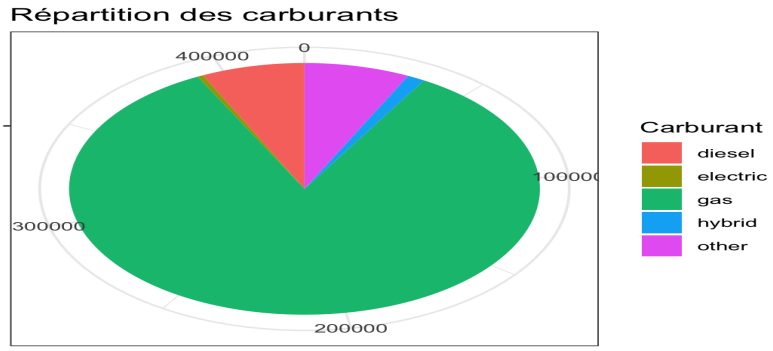
Graphique 10 : Distribution du prix moyen de véhicules par Etat



Analyse univariée de la variable Fuel

Sur la base des données ressorties (**Graphique 11**), on peut observer que les véhicules à essence sont les plus vendus sur craigslist, avec un total de 356 209. Les véhicules à moteur diesel arrivent en deuxième position, avec un total de 30 062. Les véhicules hybrides et les autres types de carburant sont relativement moins nombreux, avec respectivement 5 170 et 30 728 véhicules vendus. Les véhicules électriques sont les moins nombreux, avec seulement 1 698 véhicules vendus.

Graphique 11 : Distribution du prix moyen de véhicules par Etat



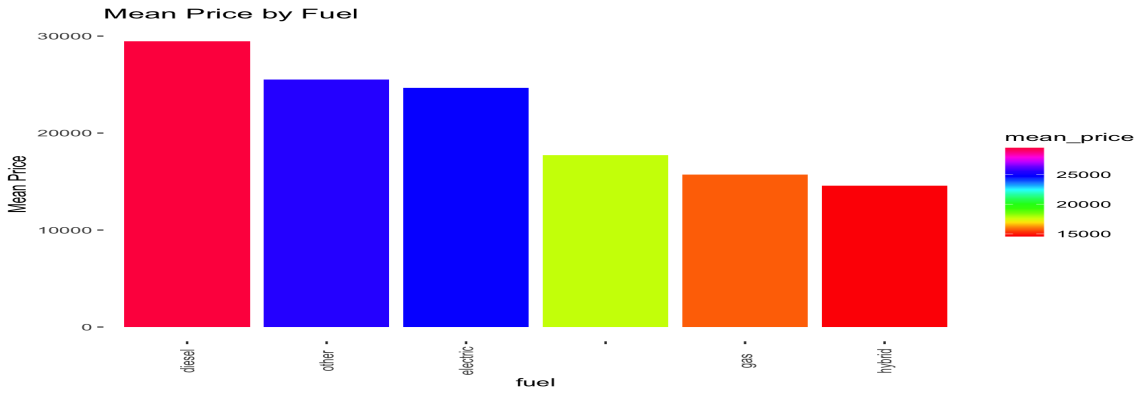
Analyse bivariée de la variable fuel par rapport au prix moyen

En analysant les données fournies, on peut constater que le prix moyen des voitures postées sur Craigslist varie considérablement en fonction du type de carburant utilisé. Les voitures diesel ont le prix moyen le plus élevé à 29477,86 dollars, tandis que les voitures hybrides ont le prix moyen le plus bas à 14582,43 dollars.

Les voitures électriques ont un prix moyen de 24648,36 dollars, ce qui est plus élevé que celui des voitures hybrides mais plus bas que celui des voitures diesel. Les voitures à essence ont un prix moyen de 15714,64 dollars, ce qui est considérablement plus bas que le prix moyen des voitures diesel, électriques et autres types de carburant.

En conclusion, l'analyse des données indique que le type de carburant utilisé a une influence significative sur le prix moyen des voitures postées sur Craigslist. Les voitures diesel ont tendance à être les plus chères, suivies des voitures de type "autre", électriques, à essence et hybrides. Ces informations peuvent être utiles pour les acheteurs et les vendeurs de voitures qui cherchent à déterminer le prix approprié pour leurs transactions.

Graphique 12 : Distribution du prix moyen de véhicules par Etat

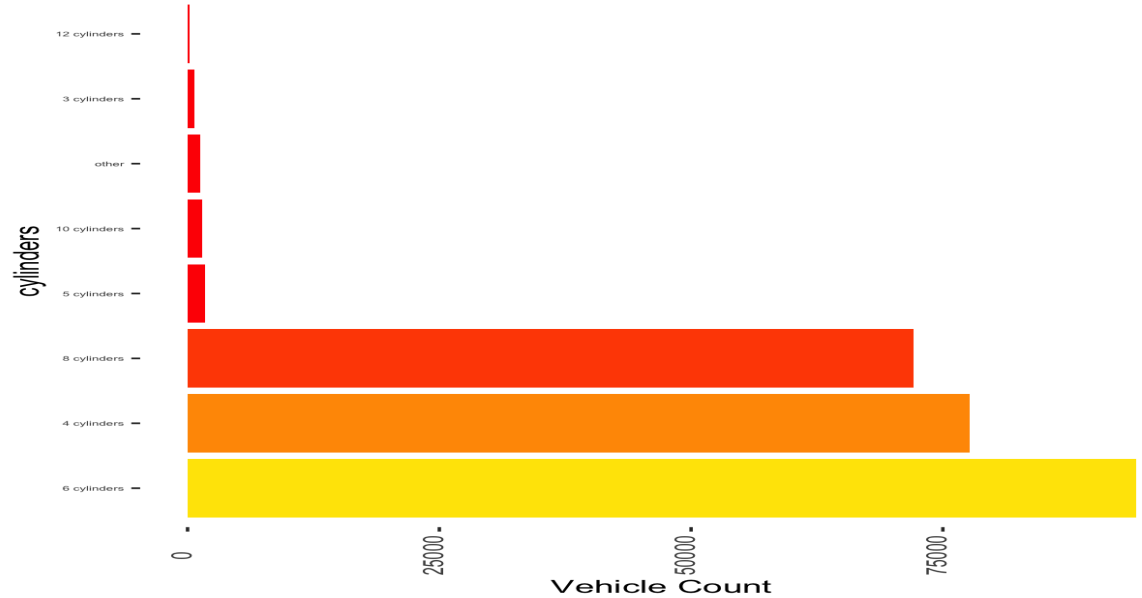


Analyse univariée de la variable Cylinders

D'après le graphique 13 type de cylindre le plus courant pour les véhicules vendues sur craigslist est le 6 cylindres, avec un nombre de 94 169. Vient ensuite le 8 cylindres, avec 72 062 cylindres. Les types de cylindres les moins courants sont le 12 cylindres, avec seulement 209 cylindres, et le 10 cylindres, avec 1 455 cylindres. Le 4 cylindres est également un choix populaire pour les véhicules vendus sur craigslist, avec un nombre de 77 642.

Le nombre de véhicules vendus sur craigslist avec 3 cylindres, 5 cylindres et d'autres types de cylindres est faible mais notable, avec respectivement 655, 1 712 et 1 298. Dans l'ensemble, il semble que la plupart des personnes qui vendent leur véhicule sur craigslist préfèrent les moteurs à 6 et 8 cylindres, mais il existe également un marché important pour les moteurs à 4 cylindres. Les types de cylindres les moins courants ne représentent qu'une petite partie de l'ensemble des véhicules vendus sur craigslist.

Graphique 13 : Distribution nombre de véhicules par cylindres



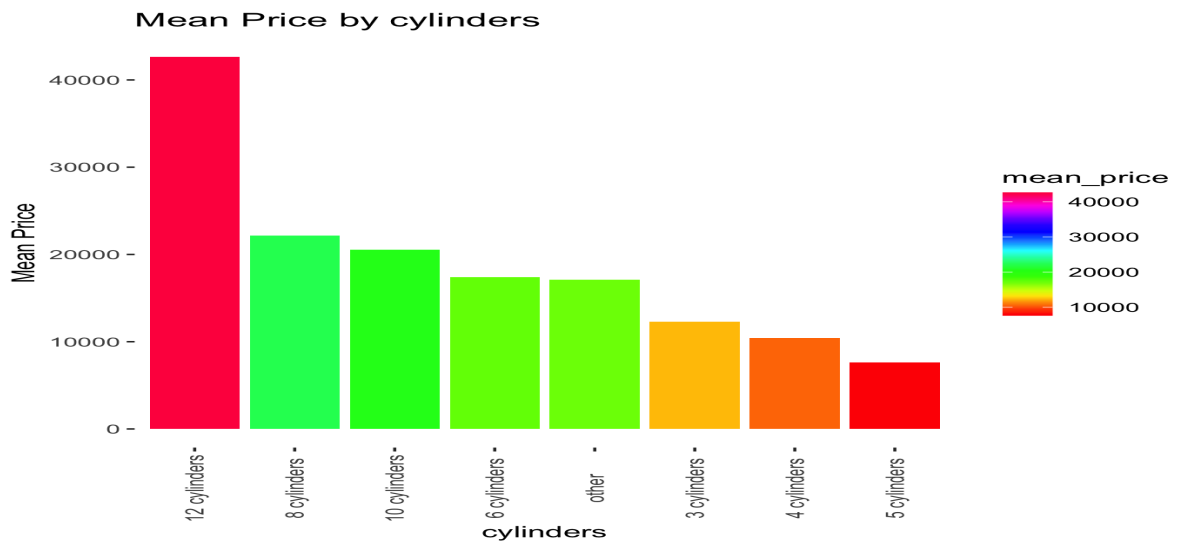
Analyse bivariée de la variable Cylinders par rapport au prix moyen

Les véhicules à 12 cylindres ont le prix moyen le plus élevé, soit 42 659,22 $ (graphique 14), ce qui n'est pas surprenant puisque les voitures à plus grand nombre de cylindres ont généralement des moteurs plus puissants et sont considérées comme des véhicules de luxe.

Le deuxième prix moyen le plus élevé est celui des véhicules à 10 cylindres, qui s'élève à 20 557,85 dollars. Ces véhicules sont également considérés comme des véhicules de luxe, mais ils sont moins courants que les véhicules à 12 cylindres.

Les véhicules à 8 cylindres ont le troisième prix moyen le plus élevé, à 22 158,70 $. Ces véhicules sont généralement plus grands et plus puissants que ceux à 6 cylindres, mais ne sont pas aussi rares ou chers que ceux à 10 ou 12 cylindres.Les véhicules à 6 cylindres ont un prix moyen de 17 417,44 $, ce qui est plus élevé que les véhicules à 4 et 5 cylindres, mais moins élevé que les véhicules à 8 cylindres ou plus. Les véhicules à 4 cylindres ont le prix moyen le plus bas, soit 10 429,71 $, ce qui n'est pas surprenant car ces véhicules sont généralement plus petits et moins puissants que ceux qui ont plus de cylindres. Cependant, ils sont aussi généralement plus économes en carburant et plus abordables.Les véhicules à 5 cylindres ont un prix moyen légèrement plus élevé que ceux à 4 cylindres, mais restent relativement abordables à 7 613,51 $. Les véhicules à 3 cylindres ont le prix moyen le plus bas après ceux à 4 cylindres, à 12 297,00 $, ce qui n'est pas surprenant non plus car ces véhicules sont généralement de petites citadines bon marché conçues pour être économes en carburant.

Graphique 14 : Distribution du prix moyen de véhicules par cylindres



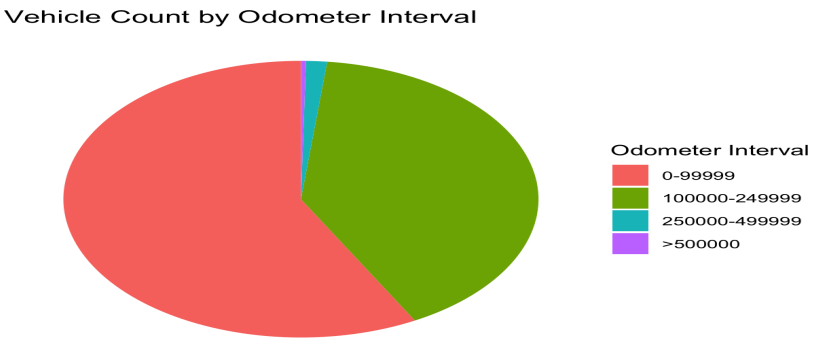
Analyse univariée de la variable Odometer

Il apparaît que la majorité des véhicules vendus sur Craigslist ont un kilométrage inférieur à 100 000 miles. Plus précisément, 244 878 véhicules ont été vendus avec un kilométrage compris entre 0 et 99 999 miles. Un nombre important de véhicules ont également été vendus avec un kilométrage compris entre 100 000 et 249 999 miles, soit 170 150 véhicules.

Le nombre de véhicules vendus diminue de manière significative à mesure que le kilométrage augmente. Seuls 6 029 véhicules ont été vendus avec un kilométrage compris entre 250 000 et 499 999 miles, et encore moins, 1 423 véhicules, ont été vendus avec un kilométrage supérieur à 500 000

miles. Dans l'ensemble, ces données suggèrent que la majorité des véhicules vendus sur Craigslist ont un kilométrage relativement faible et que le nombre de véhicules vendus avec un kilométrage plus élevé est plus faible.

Graphique 15: Distribution du nombre de véhicules par intervalle d'Odomertre

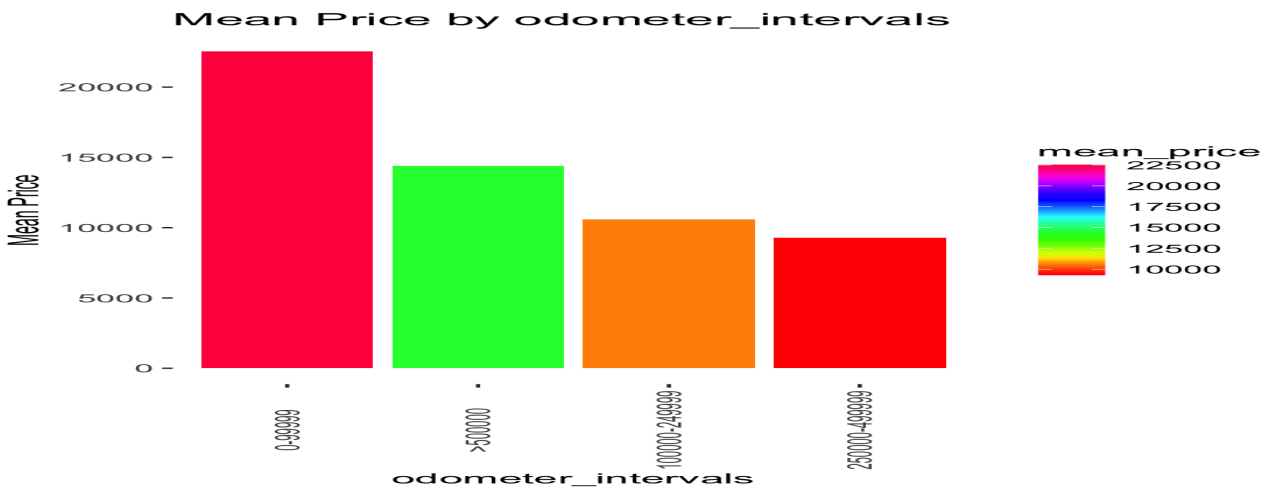


Analyse bivariée de la variable Odometer par rapport au prix moyen

Il apparaît que le prix moyen des véhicules vendus sur Craigslist diminue généralement à mesure que le kilométrage augmente. Plus précisément, les données montrent que les véhicules dont le kilométrage est compris entre 0 et 99 999 miles ont le prix moyen le plus élevé, soit 22 530,307 $. En revanche, les véhicules dont le kilométrage est compris entre 250 000 et 499 999 miles ont le prix moyen le plus bas, soit 9 294,382 dollars.

Les véhicules dont le kilométrage est compris entre 100 000 et 249 999 miles et ceux dont le kilométrage est supérieur à 500 000 miles ont un prix moyen de 10 602,188 $ et 14 391,061 $, respectivement. Dans l'ensemble, ces données suggèrent que le prix moyen des véhicules vendus sur Craigslist a tendance à être plus élevé pour ceux dont le kilométrage est faible, et qu'il diminue au fur et à mesure que le kilométrage augmente. Cela s'explique probablement par le fait que les véhicules dont le kilométrage est élevé sont perçus comme plus usés et donc moins précieux. Toutefois, il est important de noter qu'il s'agit d'une tendance générale et que certains véhicules peuvent ne pas suivre ce schéma en raison d'autres facteurs qui influencent leur valeur.

Graphique 16 : Distribution du prix moyen de véhicules par cylindres



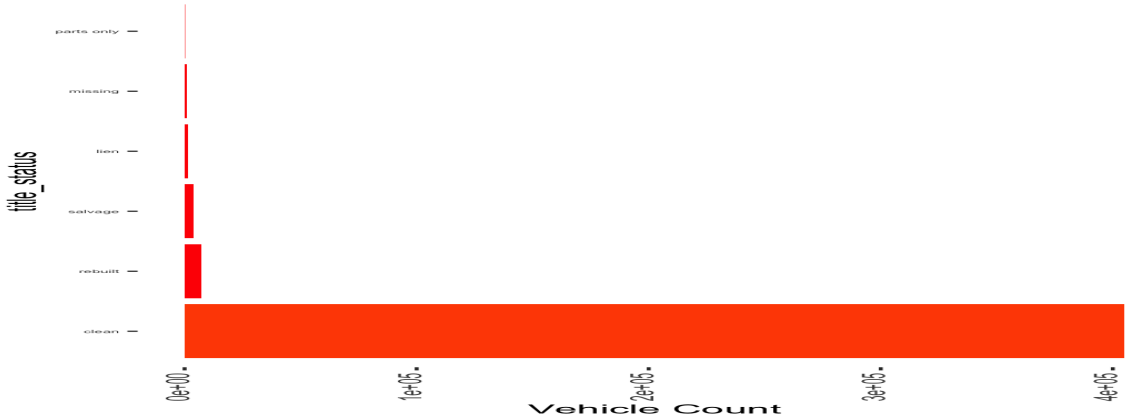
Analyse univariée de la variable Title status

En analysant les données fournies, on peut constater que le statut du titre a une influence significative sur le nombre de voitures postées sur Craigslist. Les voitures avec un titre original (clean) représentent la grande majorité des voitures postées, avec un total de 405117 voitures. Les voitures avec un titre reconstruit arrivent en deuxième position avec un total de 7219 voitures postées, suivies des voitures avec un titre de récupération avec 3868 voitures postées.

Les voitures avec un titre de lien et les voitures pour pièces seulement représentent des nombres beaucoup plus faibles, avec respectivement 1422 et 198 voitures postées. Les voitures avec tes titres de propriété manquants sont également peu courantes, avec seulement 814 voitures postées.

En conclusion, l'analyse des données montre que le statut du titre est un facteur important pour les acheteurs et les vendeurs sur Craigslist. Les voitures avec un titre propre sont les plus courantes et sont susceptibles d'avoir une valeur plus élevée que les voitures avec des titres de reconstruction ou de récupération. Les voitures avec un titre de lien ou destinées aux pièces sont moins courantes et peuvent être plus difficiles à vendre en raison de leur statut particulier.

Graphique 17 : Distribution du nombre de véhicules par Title status



Analyse bivariée de la variable Odometer par rapport au prix moyen

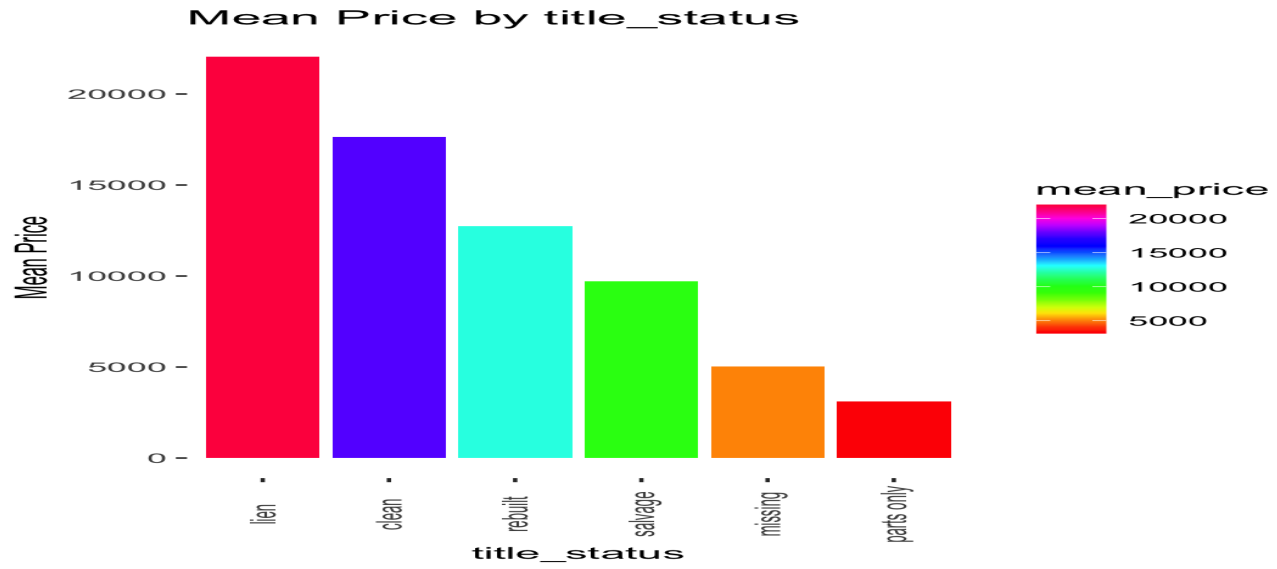
Nous remarquons dans le graphique 18 que les voitures avec le titre "lien" ont le prix moyen le plus élevé de 22049.285 dollars, tandis que les voitures avec le titre "parts only" ont le prix moyen le plus bas de 3101.660 dollars. Cela peut s'expliquer par le fait que les voitures avec le titre "lien" sont probablement en meilleur état et ont moins de problèmes mécaniques ou de dommages que les voitures avec le titre "parts only" qui ne sont vendues que pour leurs pièces.

De plus, nous remarquons que les voitures avec le titre "rebuilt" ont un prix moyen de 12739.880 dollars, ce qui est significativement inférieur à celui des voitures avec le titre "clean" ou "lien". Cela peut s'expliquer par le fait que les voitures avec le titre "rebuilt" ont été endommagées auparavant et ont donc perdu une partie de leur valeur.

Enfin, nous voyons que les voitures avec le titre "missing" ont un prix moyen de seulement 5035.012 dollars. Cela pourrait être dû au fait que l'absence de titre de propriété peut rendre la vente plus difficile car le nouveau propriétaire devra obtenir un titre de remplacement auprès des autorités locales.

En conclusion, l'analyse de ces données montre que le titre d'une voiture peut avoir un impact significatif sur son prix moyen sur Craigslist.

Graphique 18: Distribution du prix moyen de véhicules par Title status



Analyse univariée de la variable Paint color

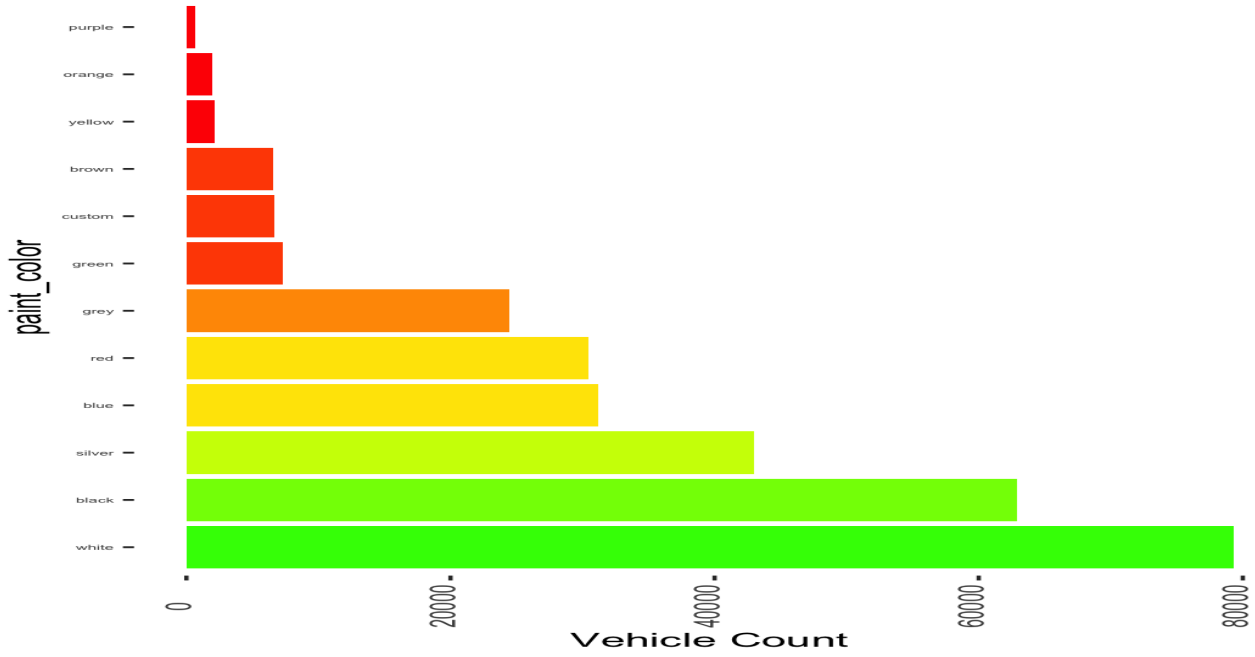
En tant que data analyste, la première étape serait de visualiser les données afin de mieux comprendre la distribution des couleurs des voitures postées sur Craigslist. Nous pouvons utiliser un histogramme ou un graphique à barres pour cela.

Une fois que nous avons une compréhension visuelle des données, nous pouvons procéder à une analyse plus approfondie. Nous remarquons que le blanc est la couleur de voiture la plus fréquemment postée sur Craigslist, avec un total de 79285 voitures. Le noir et l'argent suivent de près, avec respectivement 62861 et 42970 voitures. Ces trois couleurs représentent donc plus de la moitié de toutes les voitures postées sur Craigslist.

En revanche, le violet et l'orange sont les couleurs de voiture les moins fréquentes, avec respectivement 687 et 1984 voitures postées. Cela peut s'expliquer par le fait que ces couleurs sont moins populaires auprès des acheteurs de voitures.

De plus, nous pouvons remarquer que les couleurs les plus populaires sont souvent des couleurs neutres, telles que le blanc, le noir et l'argent. Cela peut s'expliquer par le fait que ces couleurs sont considérées comme plus "sûres" ou "traditionnelles" et sont donc plus faciles à vendre.

Graphique 19: Distribution du nombre de véhicules par couleur



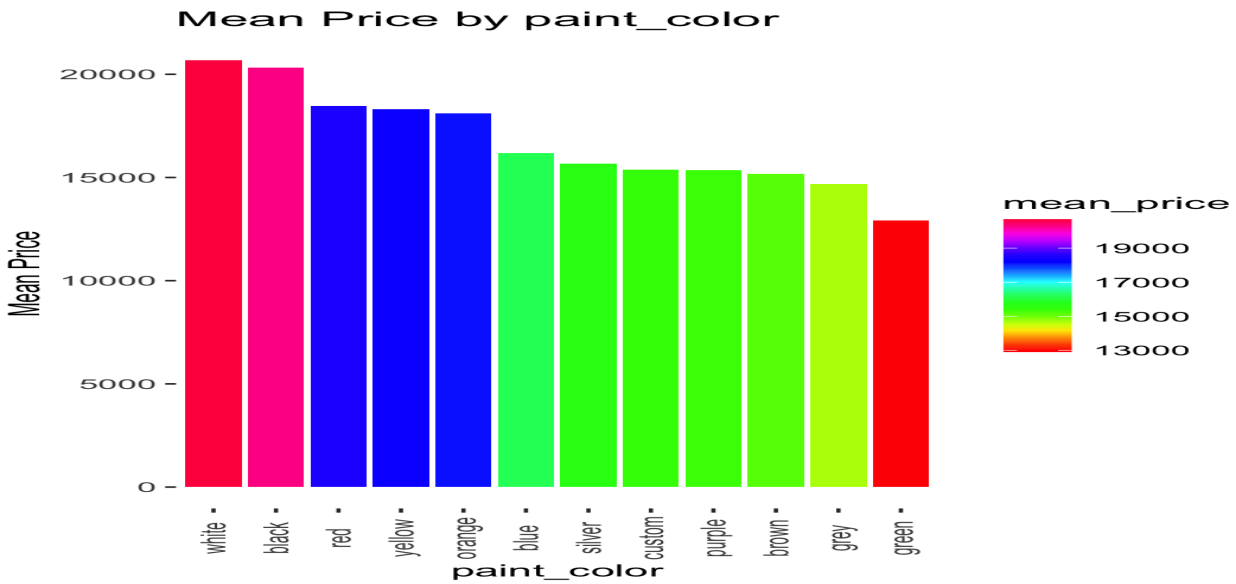
Analyse bivariée de la variable Odometer par rapport au couleur

Nous remarquons que le prix moyen des voitures varie considérablement en fonction de la couleur, allant de 12912.40 $ pour les voitures vertes à 20690.71 $ pour les voitures blanches. Les voitures noires, blanches et rouges sont les trois couleurs les plus chères en moyenne, avec des prix moyens respectifs de 20319.67 $, 20690.71 $ et 18472.16 $.

En revanche, les voitures vertes et grises sont les moins chères en moyenne, avec des prix moyens respectifs de 12912.40 $ et 14692.43 $. Cela peut s'expliquer par le fait que ces couleurs sont moins populaires auprès des acheteurs de voitures et donc moins en demande.

De plus, nous pouvons remarquer que les couleurs les plus chères sont souvent des couleurs vives et voyantes, telles que le rouge, le jaune et l'orange. Cela peut s'expliquer par le fait que ces couleurs sont souvent associées à des voitures de sport ou de luxe, ce qui peut augmenter leur valeur perçue.

Graphique 20: Distribution du prix moyen de véhicules par couleur



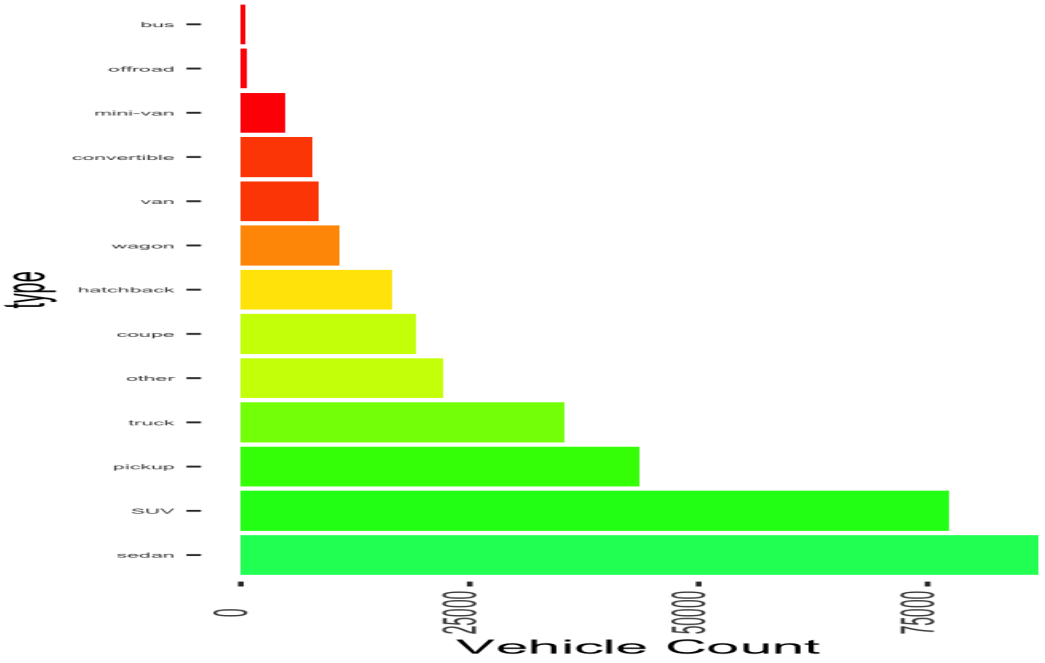
Analyse univariée de la variable type

Nous remarquons que les types de voitures les plus populaires sont les SUV, les sedans et les pickups, avec des nombres respectifs de 77284, 87056 et 43510 véhicules postés sur Craigslist. Cela peut s'expliquer par le fait que ces types de voitures sont populaires auprès des acheteurs et donc en demande.

En revanche, les types de voitures les moins populaires sont les bus, les offroads et les vans, avec des nombres respectifs de seulement 517, 609 et 8548 véhicules postés sur Craigslist. Cela peut s'expliquer par le fait que ces types de voitures sont plus spécialisés ou moins couramment utilisés, et donc moins en demande.

De plus, nous pouvons remarquer que les types de voitures les plus couramment utilisés pour les déplacements quotidiens, comme les SUV et les sedans, sont également les plus populaires sur Craigslist. En revanche, les types de voitures plus spécialisés, comme les convertibles et les coupes, ont des nombres moins élevés de véhicules postés.

Graphique 21: Distribution du nombre de véhicules par type



Nou

Analyse bivariée de la variable Odometer par rapport au type de voiture

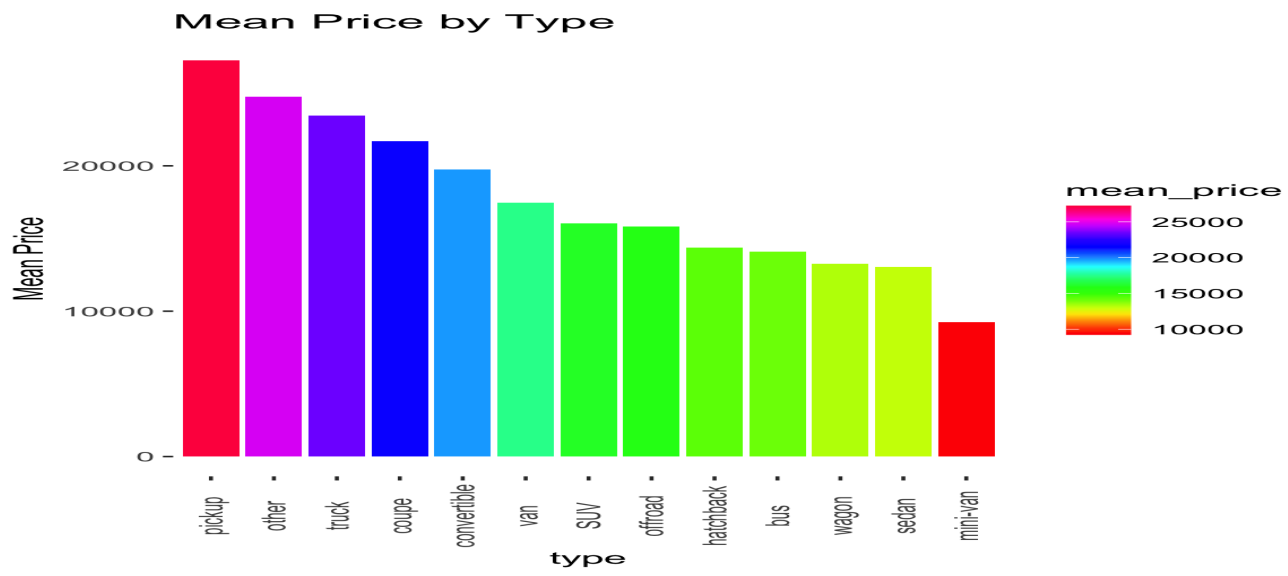
Il y a une grande variation dans les prix moyens selon le type de voiture. Les pickups ont le prix moyen le plus élevé de 27244,912 $, suivi des camions avec un prix moyen de 23461,677 $. Les autres types de voitures ont des prix moyens inférieurs à ces deux types de voitures.

Les types de voitures avec des prix moyens les plus bas sont les mini-fourgonnettes (9234,088 $) et les berlines (13051,534 $). Les hatchbacks ont également un prix moyen relativement bas de 14384,513 $. Les types de voitures les plus chers sont les coupés avec un prix moyen de 21680,948 $, les fourgonnettes avec un prix moyen de 17455,713 $ et les convertibles avec un prix moyen de 19758,508 $.

Les SUV et les offroads ont des prix moyens similaires à environ 16051,475 $ et 15813,094 $ respectivement. Les bus ont également un prix moyen similaire à 14105,617 $.Le type de voiture "other" a le prix moyen le plus élevé de tous les types de voitures à 24743,588 $. Cependant, sans informations supplémentaires sur ce type de voiture, il est difficile de tirer des conclusions solides sur les raisons de ce prix élevé.

En conclusion, il y a une grande variation dans les prix moyens des voitures postées sur Craigslist selon le type de voiture. Les pickups et les camions ont les prix moyens les plus élevés, tandis que les mini-fourgonnettes et les berlines ont les prix moyens les plus bas. Les types de voitures les plus chers sont les coupés, les fourgonnettes et les convertibles.

Graphique 22: Distribution du prix moyen des véhicules par type



Analyse univariée de la variable Condition

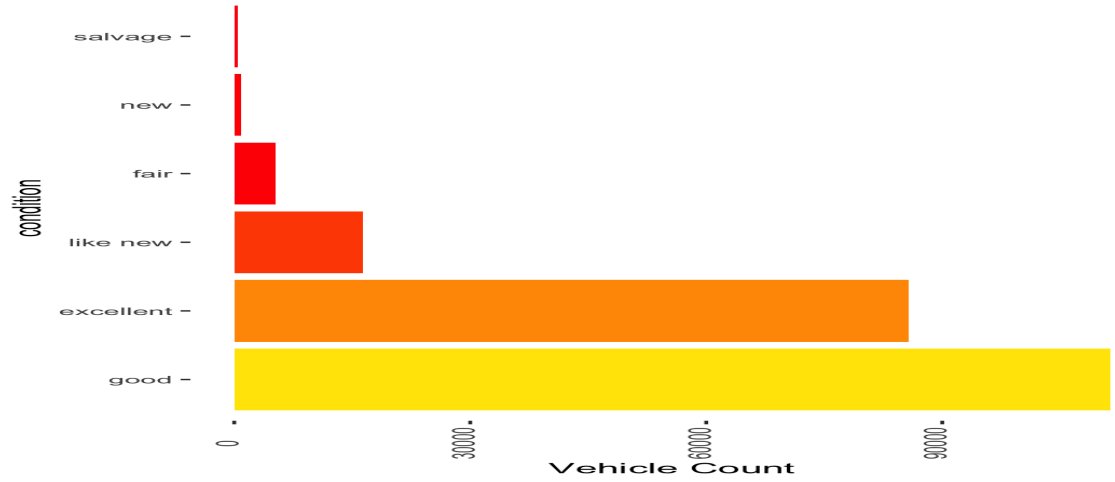
La majorité des voitures postées sur Craigslist sont dans une condition "bonne" avec un total de 111383 voitures (graphique 23). Cela représente environ 55,25% de toutes les voitures postées. La deuxième condition la plus courante est "excellente" avec un total de 85780 voitures postées, représentant environ 42,59% de toutes les voitures postées.

La condition "comme neuve" a un nombre de voitures postées relativement faible de 16309, représentant environ 8,09% de toutes les voitures postées. La condition "acceptable" a un nombre encore plus faible de voitures postées avec seulement 5259, représentant environ 2,61% de toutes les voitures postées.

Il y a un nombre très faible de voitures postées dans des conditions "neuves" et "salvage" avec 830 et 440 voitures respectivement. Cela représente seulement 0,41% et 0,22% de toutes les voitures postées.

En résumé, la majorité des voitures postées sur Craigslist sont en bonnes ou excellentes conditions, tandis que les conditions moins courantes (comme neuves, acceptables, neuves et salvage) représentent ensemble une petite proportion de toutes les voitures postées.

Graphique 23: Distribution du nombre de véhicules par condition

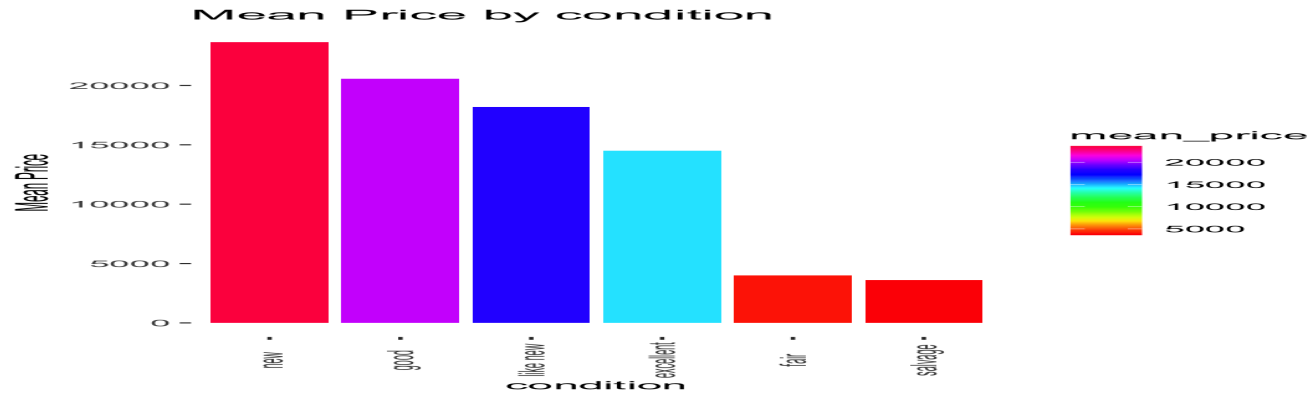


Analyse bivariée de la variable Odometer par rapport à la condition des voitures

Les voitures postées dans la condition "nouvelle" ont le prix moyen le plus élevé de 23657,267 $, ce qui est attendu étant donné leur état pratiquement neuf.Les voitures postées dans la condition "bonne" ont le deuxième prix moyen le plus élevé de 20570,625 $, suivi de près par les voitures postées dans la condition "comme neuve" avec un prix moyen de 18196,041 $. Ces conditions reflètent des voitures bien entretenues et présentant peu d'usure, ce qui explique leur prix relativement élevé.

Les voitures postées dans la condition "excellente" ont un prix moyen de 14500,763 $, ce qui est également relativement élevé. Les voitures dans cette condition sont considérées comme étant en excellent état, mais peuvent présenter une certaine usure ou des défauts mineurs. Les voitures postées dans la condition "acceptable" ont un prix moyen relativement bas de 4011,226 $. Cela peut être dû à leur état de maintenance inférieur ou à leur âge et leur kilométrage élevés. Enfin, les voitures postées dans la condition "salvage" ont le prix moyen le plus bas de toutes les conditions, avec seulement 3605,534 $. Ces voitures ont été considérablement endommagées et ont été reconstruites, ce qui peut expliquer leur prix bas. Il existe donc une corrélation entre le prix moyen des voitures postées sur Craigslist et leur condition.

Graphique 23: Distribution du prix moyen de véhicules en fonction de la condition de la voiture



Analyse univariée de la variable Condition periode of year

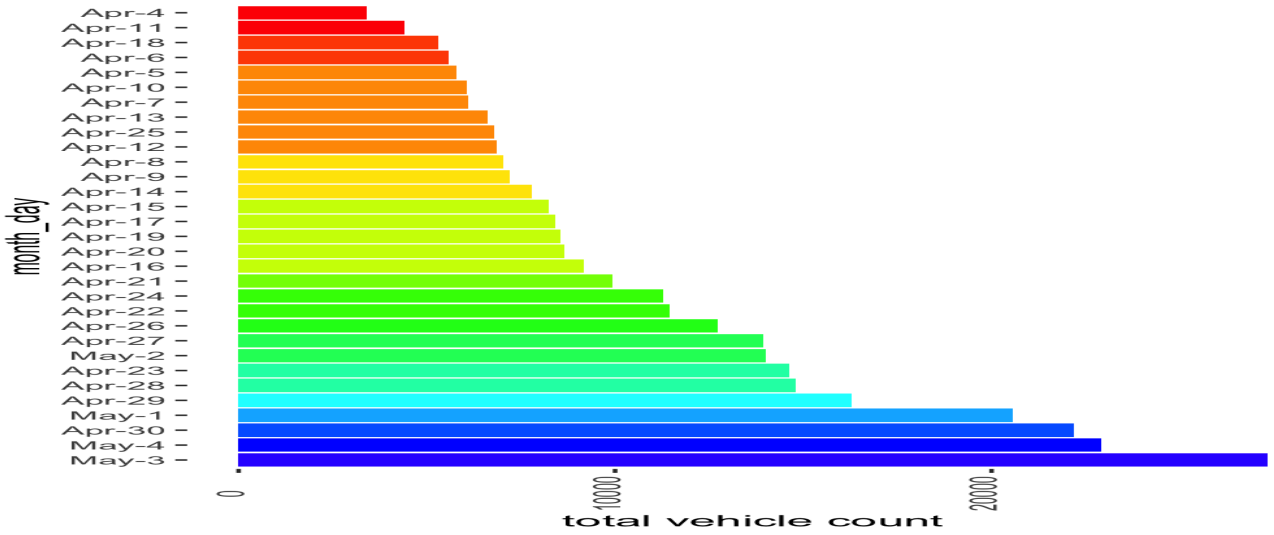
En analysant le graphique 24, on peut constater que le nombre de voitures postées sur Craigslist pour la vente a augmenté au fil du temps et a atteint un sommet le 30 avril avec un total de 22173 voitures postées. La moyenne quotidienne pour le mois d'avril était d'environ 9477 voitures postées, avec une médiane de 8399.

En comparant les données des week-ends (4-5 avril, 11-12 avril, 18-19 avril, 25-26 avril) avec celles des jours de la semaine, on peut voir qu'il y a une tendance à une diminution du nombre de voitures postées les week-ends. Il est intéressant de noter que le 30 avril, le nombre de voitures postées a augmenté considérablement par rapport aux autres jours, suggérant peut-être un pic d'activité.

En outre, la moyenne quotidienne pour le mois de mai était nettement plus élevée que celle d'avril, à environ 21106 voitures postées, avec une médiane de 20580. Cela peut suggérer une tendance à la hausse de l'activité de vente de voitures sur Craigslist.

En conclusion, l'analyse des données indique une tendance à la hausse de l'activité de vente de voitures sur Craigslist au fil du temps, avec des pics d'activité et une baisse pendant les week-ends. Ces informations peuvent être utiles pour les vendeurs de voitures qui cherchent à maximiser leur exposition et leur taux de réussite sur la plateforme.

Graphique 23: Distribution du prix moyen de véhicules en fonction de la condition de la voiture

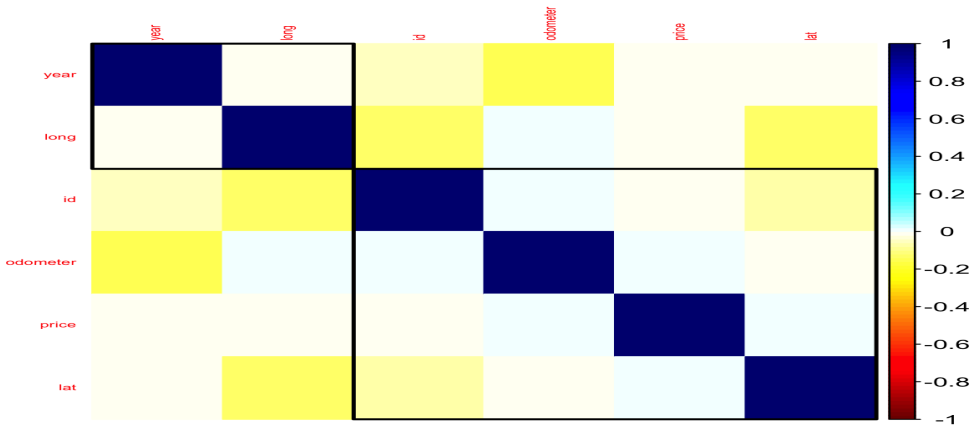


1.2 Matrice de corrélation des variables quantitatives

Nous allons à présent explorer les corrélations existantes entre les variables numériques et le prix. L'exploration des corrélations entre les variables est une étape cruciale dans l'analyse des données. Dans notre cas, nous allons nous intéresser aux corrélations entre les variables numériques et le prix

des voitures. En effet, comprendre comment les différentes variables numériques telles que l'année de fabrication, le kilométrage et autres, sont liées au prix des voitures peut fournir des informations précieuses pour la modélisation des données.

Graphique 01 : Matrice de corrélation



Corrélation du prix avec le reste variables quantitatives

Le prix présente une faible corrélation négative avec l'ID (-0,003), ce qui est normal puisque l'ID n'est qu'un identifiant et n'a pas de lien réel avec le prix.

L'année a une faible corrélation négative avec le prix (-0,059), ce qui suggère que les voitures plus récentes ont tendance à se vendre plus cher que les voitures plus anciennes.

Le compteur kilométrique présente une faible corrélation positive avec le prix (0,010), ce qui suggère que les voitures ayant un kilométrage élevé tendent à se vendre à un prix légèrement inférieur. Toutefois, cette corrélation est assez faible et pourrait ne pas être très significative.

La latitude et la longitude ont une très faible corrélation avec le prix (0,0004 et -0,0004, respectivement), ce qui suggère qu'il n'y a pas de véritable modèle géographique pour les prix des voitures sur Craigslist.

1.3 Statistique du chi deux

Nous avons 16 variables catégorielles et le calcul de la statistique du chi carré pour toutes les combinaisons possibles de deux variables donnerait 82 combinaisons. Il s'agit d'un grand nombre de tests à effectuer et il y a un risque d'observer des faux positifs (erreurs de type I) en raison de tests multiples.

Pour résoudre ce problème, nous avons décidé de nous concentrer sur les paires de variables que nous soupçonnons d'être fortement associées. Ce faisant, nous pouvons réduire le nombre de tests effectués, ce qui peut augmenter la puissance de notre analyse pour détecter des associations réelles tout en réduisant le risque d'observer des faux positifs.

Notre approche consiste à donner la priorité aux variables qui sont susceptibles d'être liées sur la base de connaissances du sujet de notre base de données. Par exemple, le "constructeur" et le "modèle" sont deux variables susceptibles d'être associées parce que certains constructeurs automobiles ont tendance à produire des modèles spécifiques qui peuvent avoir des caractéristiques uniques. Un constructeur de voitures de luxe comme BMW peut produire des modèles plus chers et plus performants qu'un constructeur de voitures bon marché comme Kia. Il est donc raisonnable de supposer que les variables relatives au constructeur et au modèle peuvent être associées.

Le "carburant" et les "cylindres" sont tous deux liés à la puissance et à l'efficacité d'un véhicule, de sorte qu'il peut être logique d'examiner l'association entre ces variables. De même, "drive" et "transmission" sont tous deux liés à la manière dont la puissance est transmise aux roues, de sorte que ces variables pourraient également être liées. Enfin, les var.

Enfin, Les variables "region" et "state" sont deux variables catégorielles qui indiquent toutes deux la localisation géographique d'un véhicule. Comme ces deux variables sont toutes deux liées à la localisation, il est possible qu'elles soient fortement associées. Il peut donc être utile de calculer la statistique du khi-deux pour toutes ces associations.

Tableau 01 : Statistique du chi carré pour les variables "constructeur" et le "modèle"

	X^2	**df**	**P(> X^2)**
Likelihood Ratio	652133	446840	0
Pearson	4332364	446840	0
Contingency Coeff.	0.987	NA	NA
Cramer's V	0.987	NA	NA

Les résultats du test pour les variables constructeur et le modèle montrent qu'il existe une forte association entre le fabricant et les variables du modèle, comme l'indiquent les valeurs élevées des mesures d'association. La valeur p est très faible, ce qui indique que l'association est statistiquement significative. Le coefficient de contingence et le coefficient V de Cramer sont tous deux proches de 1, ce qui indique une forte association entre les deux variables.

Par conséquent, la suppression de l'une de ces variables peut ne pas affecter de manière significative les résultats de nos modèles, étant donné que l'information capturée par une variable est en grande partie capturée par l'autre variable.

Tableau 02 : Statistique du chi carré pour les variables "carburant" et les "cylindres"

	X^2	**df**	**P(> X^2)**
Likelihood Ratio	7517.9	28	0

Pearson	32117.4	28	0
Contingency Coeff	0.474	NA	NA
Cramer's V	0.269	NA	NA

Le tableau présente les résultats d'un test du chi-carré pour l'association entre deux variables catégorielles "carburant" et les "cylindres". Le test du rapport de vraisemblance et le test du chi-carré de Pearson ont tous deux été effectués et, dans les deux cas, la valeur p est très faible (inférieure à 0,05), ce qui indique qu'il existe une association significative entre les deux variables.

Le coefficient de contingence est de 0,474, ce qui signifie qu'il existe une association modérée entre les deux variables. Ce coefficient varie de 0 à 1, 0 indiquant une absence d'association et 1 une association parfaite. Le coefficient du V de Cramer est de 0,269, ce qui indique également une association modérée entre les deux variables. Ce coefficient est compris entre 0 et 1, 0 indiquant une absence d'association et 1 une association parfaite.

Dans l'ensemble, ces résultats suggèrent qu'il existe une association significative et modérée entre les deux variables catégorielles.

Tableau 03 : Statistique du chi carré pour les variables "drive" et les "transmission"

Test	**X^2**	**df**	**P(> X^2)**
Likelihood Ratio	3731.7	4	0
Pearson	3559.9	4	0
Contingency Coeff.	0.176	NA	NA
Cramer's V	0.127	NA	NA

Le tableau présente les résultats d'un test du chi-carré pour l'association entre deux variables catégorielles "drive" et les "transmission". La statistique du chi carré pour le test du rapport de vraisemblance et le test de Pearson sont tous deux élevés (respectivement 3731,7 et 3559,9), ce qui indique une forte association entre les deux variables.

Le coefficient de contingence (0,176) suggère que la force de l'association est modérée, ce qui signifie que les deux variables ne sont pas complètement indépendantes, mais qu'elles ne sont pas non plus parfaitement liées.

La valeur du V de Cramer (0,127) indique également une association modérée entre les deux variables. Dans l'ensemble, ces résultats suggèrent qu'il existe une association modérée, mais statistiquement significative, entre les variables de conduite et de transmission.

Tableau 04 : Statistique du chi carré pour les variables "state" et les "region"

	X^2	**df**	**P(> X^2)**
Likelihood Ratio	2991182	20150	0
Pearson	20720148	20150	0
Contingency Coeff.	0.99	NA	NA
Cramer's V	0.985	NA	NA

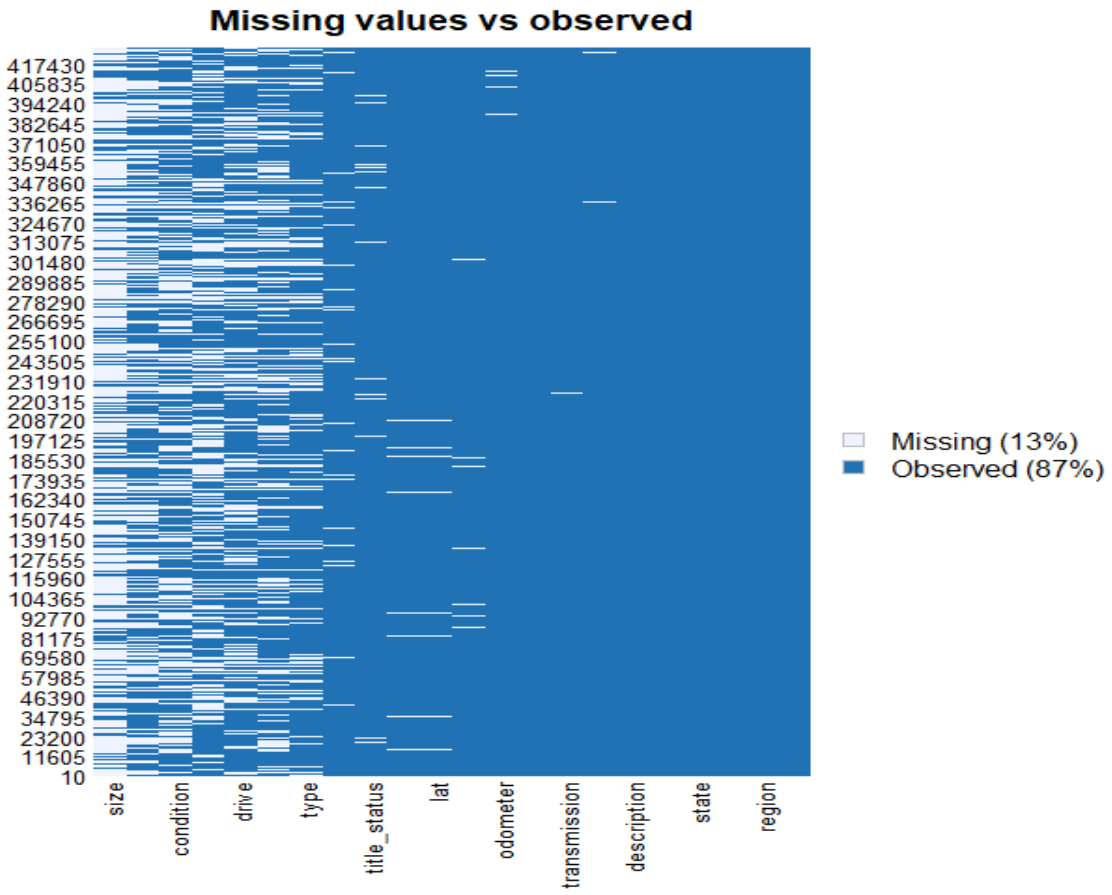
Le tableau présente les résultats d'un test du chi-carré pour l'association entre deux variables catégorielles "state" et les "region". La valeur P est extrêmement faible (proche de zéro), ce qui indique qu'il existe une association statistiquement significative entre les deux variables.

Le coefficient de contingence est de 0,99, ce qui indique une association très forte entre les deux variables. De même, la valeur du V de Cramer est également élevée (0,985), ce qui indique une forte association entre les deux variables.

1.4 Preprocessing et sélection des variables

Avant de mettre en œuvre un modèle d'apprentissage automatique supervisé ou non supervisé, il est indispensable de traiter les données pour garantir sa qualité et sa plus value décisionnelle.

Graphique 24 : Représentation des variables manquantes



Considérons le graphique 24 et tableau 5, nous pourrions interpréter la sortie comme suit : la prévalence globale des données manquantes est notable (13 % manquantes : soit un nombre total de 1 228 386) et il y a des valeurs manquantes principalement dans septs variables : **Size**: 71.7674 % ,**Cylinders**: 41.62 %, **Condition**: 40.79 % ,**VIN**: 37.73 %, **Drive** : 30.59%, **Paint_color** : 30.51 %, **Type** : 21.5%. Les autres pourcentages sont faibles et varient entre 0.016% et 4.14%.

Dans le cas de notre dataset, toutes les observations possèdent au moins une valeur manquante dans l'une des variables.

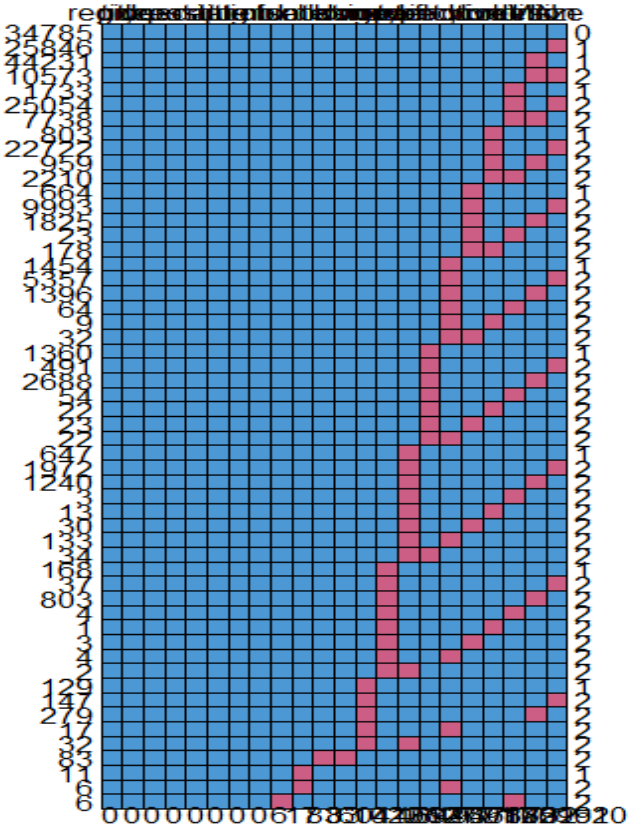
Tableau 05 : Nombre et pourcentage des valeurs manquantes

Variable	Nombre de VM	En pourcentage
size	306361	71.8
cylinders	177678	41.6
condition	174104	40.8
VIN	161042	37.7
drive	130567	30.6
paint_color	130203	30.5

type	92858	21.8
manufacturer	17646	4.13
title_status	8242	1.93
lat	6549	1.53
long	6549	1.53
model	5276	1.24
odometer	4400	1.03
fuel	3013	0.706
transmission	2556	0.599
year	1205	0.282
description	69	0.0162
posting_date	68	0.0159
id	0	0
region	0	0
price	0	0
state	0	0

Nous essayons maintenant de déterminer la typologie de nos valeurs manquantes(VM). le graphique ci-dessous (Graphique 25) permet de présenter le profit de nos valeurs manquantes tout en créant une matrice dans laquelle chaque ligne correspond à un modèle de données manquantes (1=observé, 0=manquant). Plus précisément on peut conclure que le modèle de nos VM est plus un modèle non connecté dont il y a pas vraiment une relation entre les différents VM. Cette conclusion est basée sur la forme de graphique ou il y n'a pas des VM manquantes qui sont connectés entre eux.

Graphique 25 :Profil des valeurs manquantes



Nous appuyons notre conclusion via la visualisation des combinaisons fréquentes de données manquantes. Le but est de déterminer approximativement quelles variables ont tendance à être manquantes simultanément(co-occurrence) (Graphique 26). Il faut en distinguer les causes, surtout si elles ne sont pas le simple fruit du hasard.

Il se passe beaucoup de choses, voici les principales pièces :

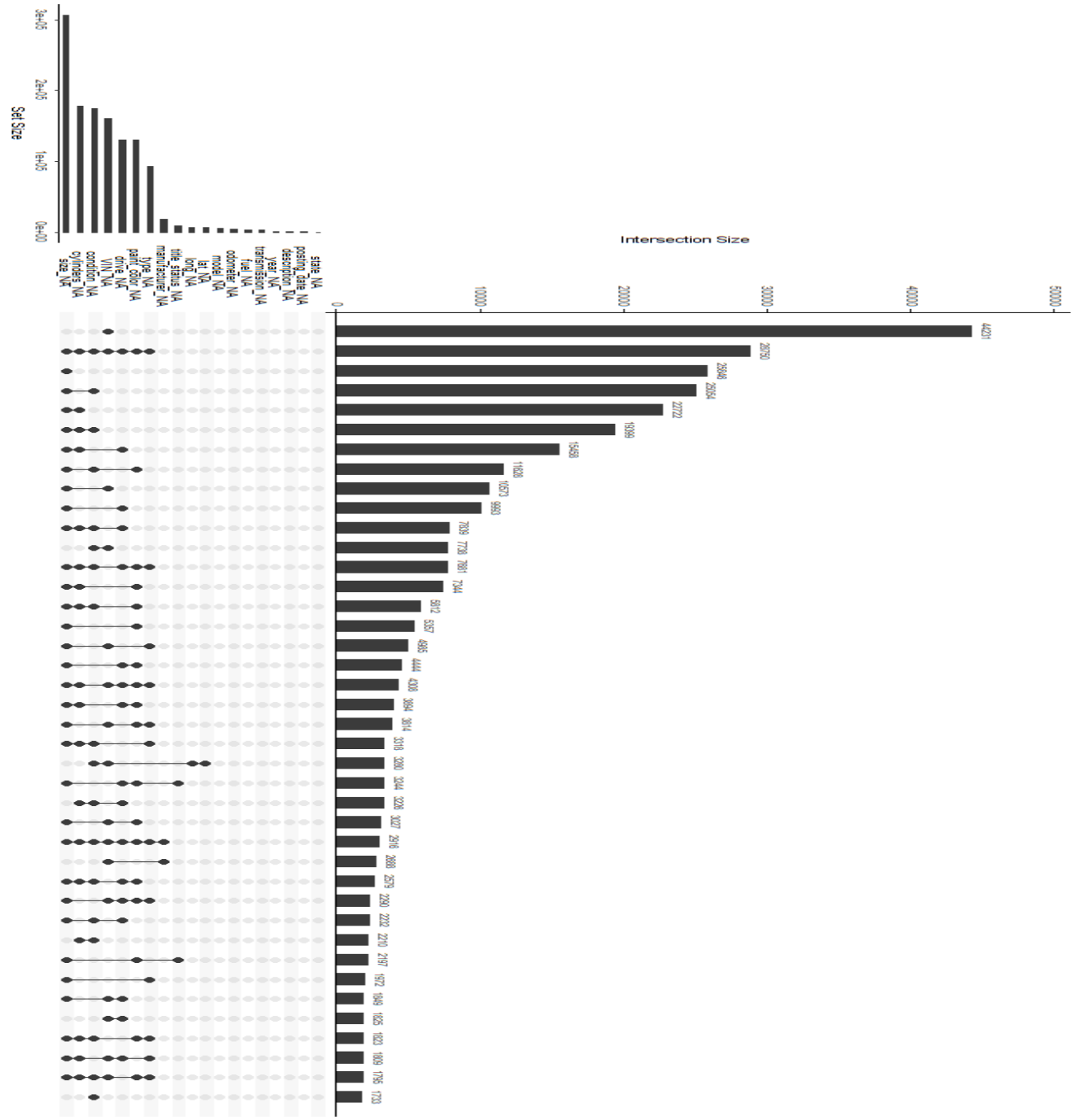
- Les barres verticales indiquent la fréquence des combinaisons manquantes uniques, indiquées par les points sous chaque barre qui correspondent aux noms de variables à leur gauche
- Les barres horizontales en bas à gauche indiquent le nombre total de valeurs manquantes pour chaque variable

On peut voir que sur plus que 400k observations il n'existe que 44231 observations dont il y a une relation avec l'ensemble des VM. Donc on ne trouve pas des relations fréquentes entre les VM. En se basant sur la typologie qui a été développée par Little & Rubin (1987) : on peut affirmer avec prediance que la catégorie de nos VM sont MAR c'est-à-dire que les données ne manquent pas de façon complètement aléatoire. Autrement dit, la probabilité d'absence est liée à une ou plusieurs autres variables observées.

Graphique 26 :les combinaisons fréquentes de données manquantes

Suppression de variables

Nous avons supprimé certaines colonnes qui n'étaient pas pertinentes et/ou offrent une information redondante pour l'analyse des prix des voitures.



Les colonnes supprimées sont "region", "url", "id", "region_url", "model", "image_url", "county", "ID", "VIN", "description", et "size". Il est justifié de supprimer des variables avant l'analyse en composantes principales (ACP) et l'analyse en composantes multiples (ACM), lorsque les variables ne sont pas pertinentes pour l'objectif de notre analyse. En effet, la présence de variables redondantes ou non pertinentes peut entraîner une surcharge des données et affecter la qualité de l'analyse.

La suppression des identifiants des observations est justifiée par les exigences spécifiques de notre analyse et des modèles utilisés. Notre analyse avec un algorithme non supervisé ne nécessite pas

d'identifiants pour identifier les tendances. De plus, pour notre deuxième étape qui consiste à utiliser un algorithme supervisé pour prédire les prix futurs voitures, la suppression des ’ID’ ne devrait pas poser de problème majeur car les modèles supervisés peuvent apprendre les tendances à partir des données sans avoir besoin des identifiants des observations.

La variable 'url' contient simplement l'URL de la page Web où l'annonce a été publiée. Cette variable n'est pas liée à la prédiction des prix des vehicules et n'affecte pas les résultats de notre modèle. Par conséquent, elle peut être supprimée sans perte d'informations utiles.

La variable 'region_url' est également une URL, qui contient des informations sur la région où l'annonce a été publiée et la variable ‘région’ contient de la région où a lieu la vente. Bien que ces variables puissent être utilisées pour regrouper les données en fonction de la région, elles ne sont pas nécessaires pour notre analyse, car les variables ‘lat’, ‘long’ et ‘state’ fournissent déjà des informations sur l’emplacement géographique de la vente du véhicule. Par ailleurs, les résultats des tests de chi-squared (tableau 04) indiquent une forte association entre la région et l’Etat, le degré d’information fourni par région existe donc deja dans Etat.

La variable 'image_url' contient l'URL de l'image associée à l'annonce. Bien que cela puisse être utile pour afficher l'image correspondante, cela n'affecte pas la prédiction des prix des vehicules. Par conséquent, cette variable peut également être supprimée sans perte d'informations utiles.

La variable "size" possède 70% de valeurs manquantes, il est peu probable que l'on puisse imputer ces valeurs manquantes à partir des 30% de données restantes de manière fiable et sans risquer d’introduire un biais important dans les résultats de nos modèles. Dans ce cas, il est préférable de supprimer la variable plutôt que d'utiliser des techniques d'imputation des données qui peuvent introduire des biais dans les analyses.

La variable VIN (Vehicle Identification Number) est une valeur unique attribuée à chaque voiture et elle ne fournit pas d'information utile pour prédire le prix d'une voiture. En d'autres termes, la variable VIN n'a pas de relation directe avec la variable cible (le prix de la voiture) ou avec d'autres variables descriptives de la voiture (modèle, type, constructeur, etc.). Par conséquent, il est justifié de supprimer la variable VIN car elle ne contribue pas à l'analyse et peut même ajouter du bruit aux données.

La suppression de la variable Id de notre dataset est justifiée car elle est spécifique à chaque vente et n'affecte pas directement le prix de vente de la voiture. Par ailleurs la matrice de corrélation (graphique 01) a permis de voir que l’id n'était effectivement pas corrélé avec le prix. Par conséquent, il est raisonnable de supprimer cette variable étant donné que notre l'objectif principal de l'analyse est de prédire ou d'explorer les facteurs qui influencent le prix des voitures sur Craigslist.model

En conclusion, la suppression de ces variables ne réduit pas la qualité de notre modèle de prédiction prix des véhicules, car ces variables ne sont pas pertinentes pour notre analyse.

Nous avons ensuite retiré des données aberrantes dans la colonne prix en ne gardant que les voitures dont le prix est inférieur à 250 000 dollars et supérieur à zéro. En effet, en observant les boxplots de la variable "prix" de notre dataset, nous avions constaté une concentration des points en dessous de 250 000 dollars.

Nous avons donc décidé de ne garder que les voitures dont le prix est inférieur à 250 000 dollars et supérieur à zéro. Cette décision a été prise pour éviter d'inclure des données aberrantes qui pourraient

fausser nos résultats et nuire à la qualité de nos modèles. En supprimant ces valeurs extrêmes, nous avons pu nous concentrer sur un échantillon plus homogène et représentatif de la population totale de voitures vendues sur Craigslist.

A ce stade de notre retraitement le jeu de données contenait 393 890 observations, ce qui est beaucoup trop pour effectuer des analyses efficaces. Par conséquent, nous avons créé un échantillon avec zéro valeur manquante par observation. Cela nous a permis d'obtenir un échantillon plus petit et plus facile à manipuler contenant 112 264 observations.

En supprimant les observations incomplètes, nous nous sommes assurés de disposer d'un échantillon plus fiable et représentatif pour effectuer nos analyses. Cela nous permettra d'obtenir des résultats plus précis et plus pertinents pour nos modèles de prédiction de prix des voitures. En outre, cela permettra également de simplifier le traitement des données et de rendre les analyses plus rapides et plus efficaces.

Nous avons choisi de ne pas imputer les valeurs manquantes dans ce cas car les variables avec des valeurs manquantes étaient principalement des variables catégorielles, pour lesquelles l'imputation aurait pu introduire des biais et des erreurs dans les résultats des analyses. En conclusion, nous avons opté pour une approche plus conservatrice en n'utilisant que les observations complètes, ce qui a limité les risques d'erreurs dans les analyses ultérieures.

Cependant, même avec cet échantillon, nous avons encore beaucoup d'observations. Ainsi, nous avons choisi de travailler sur un sous-échantillon aléatoire de 20% de l'échantillon initial. Cela nous a permis de travailler sur un ensemble de données plus gérable, contenant 22 453 observations.

Notre choix de prendre un échantillon aléatoire de 20% de l'échantillon initial est justifié par plusieurs raisons. Tout d'abord, cela nous permet de réduire la taille de l'échantillon et donc d'alléger la charge de calcul nécessaire pour effectuer des analyses. En effet, avec l'ensemble initial de 112 264 observations, il est difficile de manipuler efficacement les données et de les visualiser de manière adéquate. De plus, le choix d'un échantillon aléatoire nous permet de préserver la représentativité de l'échantillon initial.

En effet, si nous avions choisi de prendre les 22 453 premières observations de l'échantillon initial, cela aurait pu biaiser les résultats si ces observations avaient été choisies selon un critère particulier, tel que la date de publication de l'annonce. En revanche, en choisissant un échantillon aléatoire, nous nous assurons que chaque observation a la même probabilité d'être incluse dans l'échantillon, ce qui réduit le risque de biais de sélection. Enfin, un échantillon aléatoire permet également de réduire les risques de sur-représentation ou de sous-représentation de certaines sous-populations dans l'échantillon, ce qui peut être important dans le cadre d'analyses statistiques.

3. Analyse multidimensionnelle :

Dans le cadre de notre analyse multidimensionnelle nous allons adopter la méthode d’analyse en composantes principales pour les variables quantitatives et l'analyse des correspondances multiples pour les variables qualitatives.

3.1 L’analyse en composantes principales (ACP) :

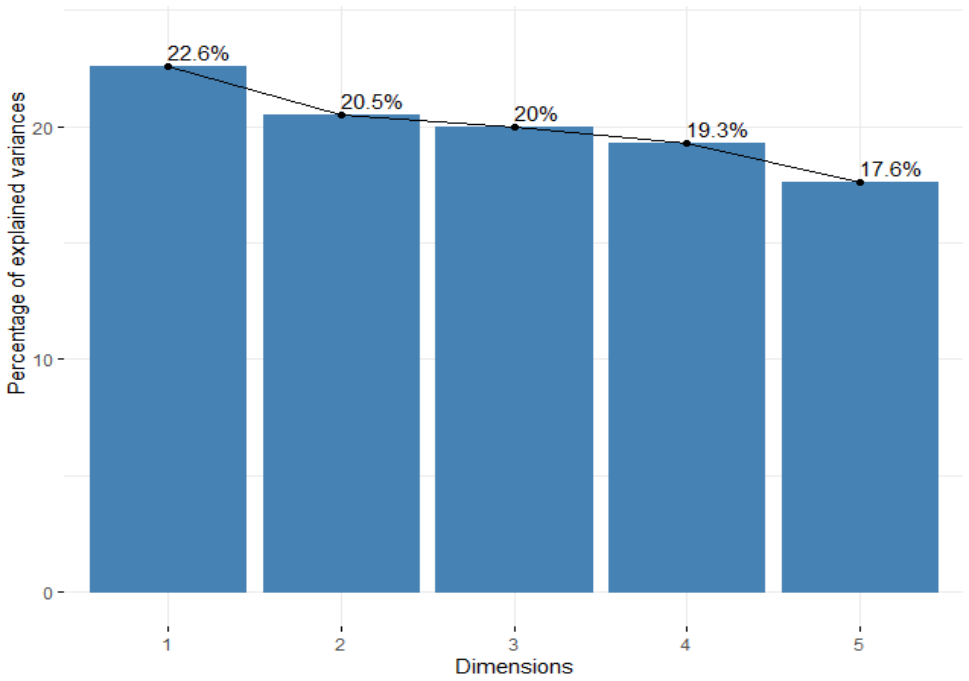
Dans un premier temps, nous allons analyser un modèle d'analyse en composantes principales. Contrairement à l'analyse univariée, l'analyse multidimensionnelle permet d'étudier les liens entre plusieurs variables et/ou plusieurs individus. Commençons par le choix de nombre de composantes principales à utiliser et pour cela on va se baser sur les méthodes existantes dans la littérature qui sont globalement trois solutions. les trois tois règles fréquemment utilisés sont les suivantes :

- 1ere règle : on regarde la valeur propre si la valeur est supérieure à 1 on le garde (Kaiser 1961).
- 2ème règle : « **méthode du coude** » qui consiste à repérer l'endroit à partir duquel le pourcentage d'inertie diminue beaucoup plus lentement lorsque l'on parcourt le diagramme des éboulis de gauche à droite.
- 3ème règle : on regarde la var cumulé en gardant les dimensions qui englobent une variance cumulée jusqu'à 80%.

Malheureusement, il n'existe pas de méthode objective bien acceptée pour décider du nombre d'axes principaux qui suffisent. Cela dépendra du domaine d'application spécifique et du jeu de données spécifiques.

Dans le graphique ci-dessus, nous pourrions vouloir toutes les composantes. En effet, les informations (variances) contenues dans les données sont reparties avec des pourcentages très proches sur les cinq composantes principales. On remarque que la première dimension ne représente pas la variance expliquée la plus élevée alors donc on peut proposer la non existence d'une forte corrélation entre les variables explicatives quantitatives dans notre modèle. Il faut bien noter que moins les variables sont corrélées plus la variance va être dispersée sur les différents axes.

Graphique 27 : Pourcentage de variances expliquées de chaque axes

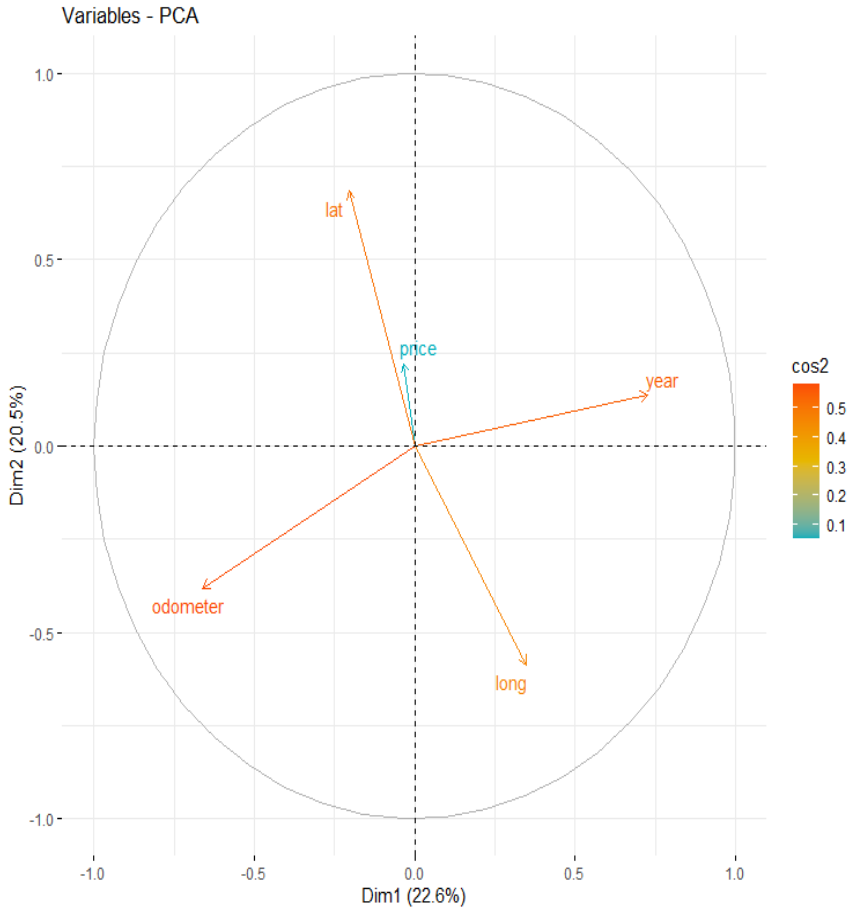


Passons à présent à l'analyse simultané du cercle de corrélations et le tableau représentant les contributions de chaque variable dans la construction des axes 1 et 2. L'examen du cercle de corrélation permet de détecter les éventuels groupes de variables qui se ressemblent ou au contraire qui s'opposent donnant ainsi un sens aux axes principaux. Dans un premier lieu, les variables sont représentées selon leurs corrélations avec les axes. La valeur de la corrélation dépend de la longueur de flesh.

On remarque que les variables qui corrèlent le plus à la dimension 1 sont **year** avec une corrélation positive (une contribution de 46.841675), **odometer** (38.597545) avec une corrélation négative avec l'axe et **long**(10.773947) avec une corrélation positive. La variable **prix** est faiblement corrélé négativement avec l'axe 1 avec une très faible contribution à l'ordre de 0.099003 (un taille réduit du flèche). La variable **prix** est plus dominante dans la construction du troisième axe (9.463299e+01).

La deuxième dimension est principalement corrélée positivement avec les variables **lat**(45.548036) et négativement avec **long**(33.651384). Il faut noter que, plus la coordonnée d'une variable sera importante, plus la variable contribue au concept représenté par ces axes.

Graphique 28 : Cercle de corrélations



Selon la représentation des variables sur la cercle de corrélations, on observe aussi :

- une corrélation négative entre les variables year et odometer => une colinéarité négative élevée (une angle plate égale presque 180 degré).
- une colinéarité positive entre les variables price et lat. Elles peuvent présenter un groupe de variables qui varient ensemble.
- L'angle formé par les vecteurs colonnes renseignent la corrélation sur les variables

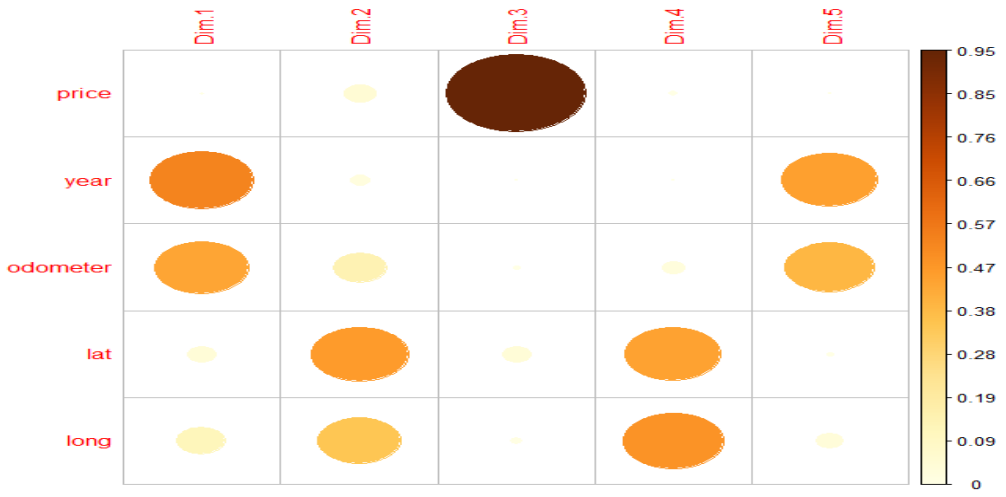
En se basant sur l'interprétation de la cercle de corrélation nous conclurons qu' on aura pas un problème de multicolinéarité entre les variables quantitatives indépendantes dans nos modèles.

Tableau 6 :Contributions des variables dans la construction des axes

	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
price	0.099003	4.790155	9.463299e+01	4.703664e-01	0.00748868
year	46.841675	1.851489	4.922587e-04	2.613853e-04	51.30608237
odometer	38.597545	14.158936	3.472356e-01	2.684680e+00	44.21160390
lat	3.687830	45.548036	4.224247e+00	4.625778e+01	0.28210422

long	10.773947	33.651384	7.950385e-01	5.058691e+01	4.19272083

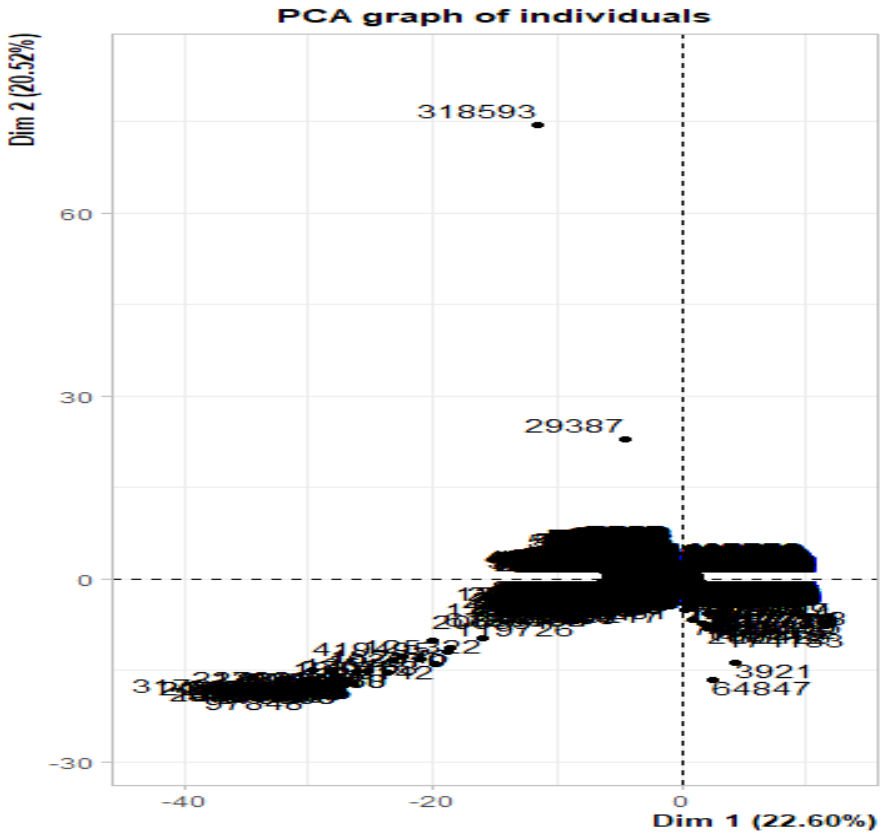
Graphique 28 : Contributions des variables dans la construction des axes 1 et 2



On s'intéresse maintenant aux individus et leurs représentations sur les axes. Le graphique ci-dessous affiche les dimensions et les coordonnées des observations : on regarde s' il y a des observations qui se rapprochent et on peut voir aussi la contribution de chaque observation sur les axes 1 et 2. Il faut noter que les observations qui sont proches de 0 n'ont pas beaucoup contribué à l'axe. Aussi, on va regarder la qualité de représentation de chaque observation dans la construction des axes 1 et 2. Les observations qui ont des coordonnées élevées sur l'axe 1 sont caractérisées par un valeur élevée des variables qui sont corrélés significativement à cet axe notamment les variables year, odometer et long . Pour l'axe 2 variables qui sont corrélés significativement à cet axe sont les variables lat et long. Donc l'axe 2 c'est plutôt un axe qui divise les observations en termes de localisations géographique. L'axe 1 présente des caractéristiques liées à l'état de la vésicule en termes d'ancienneté et la distance parcourue. En d'autre terme, les observations prochent ces deux axes sont des véhicules qui partagent les mêmes caractéristiques vis-à-vis des variables quantitatives étudiées. Par exemple, le groupe des véhicules présentent en bas à gauche sont des observations qui ont la distance parcourue la plus élevée.

Dans notre cas on constate que nos individus présentent des coordonnées condensés autour de la moyenne au même temps on constate aussi des contributions excessive qui constituent un facteur d'instabilité comme dans le cas des observations 29387, 318593. Le retirer de ces derniers peut modifier profondément le résultat de l'analyse. On a alors intérêt à effectuer l'analyse en éliminant puis à le rajouter et de comparer ensuite les résultats.

Graphique 29 : Contributions des variables dans la construction des axes 1 et 2



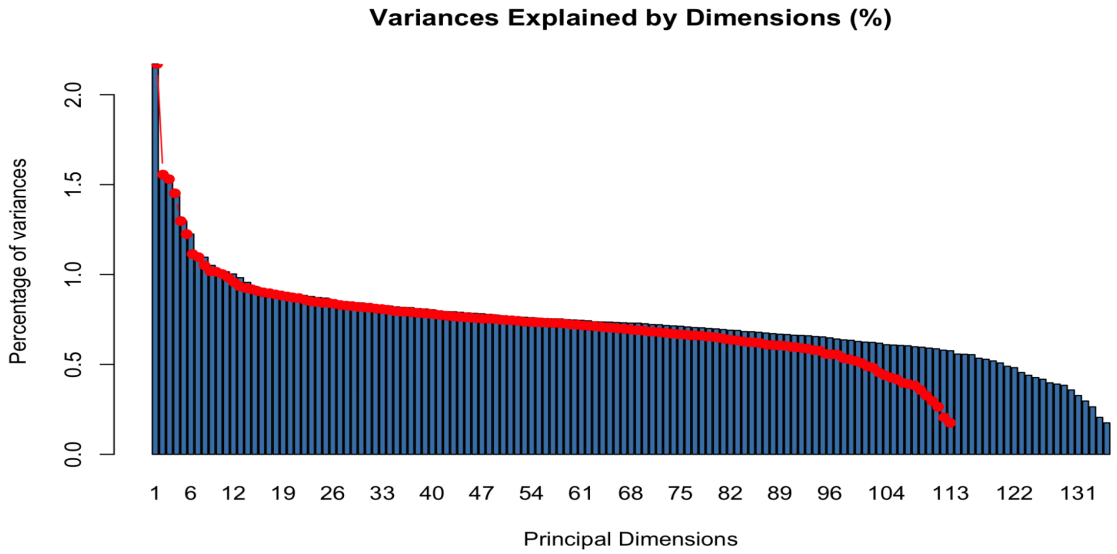
3.2 L'analyse des correspondances multiples (ACM) :

Nous allons à présent utiliser L'ACM pour analyser plus en profondeur les données catégorielles et identifier les associations entre les variables catégorielles ainsi que de savoir quelles sont les modalités corrélées entre elles. Cela nous permettra par ailleurs, de visualiser des données à haute dimension, ce qui facilite leur interprétation et permet d'en tirer des conclusions significatives.

Les résultats dans les graphiques 30 et 31 montrent que l'analyse en composantes multiples (ACM) permet d'expliquer une partie de la variance des données en utilisant plusieurs dimensions. Dans ce cas précis, la première dimension explique 2,17 % de la variance totale, la deuxième dimension explique 1,56 % de la variance totale, et ainsi de suite. Les dimensions suivantes expliquent une proportion de plus en plus faible de la variance totale.

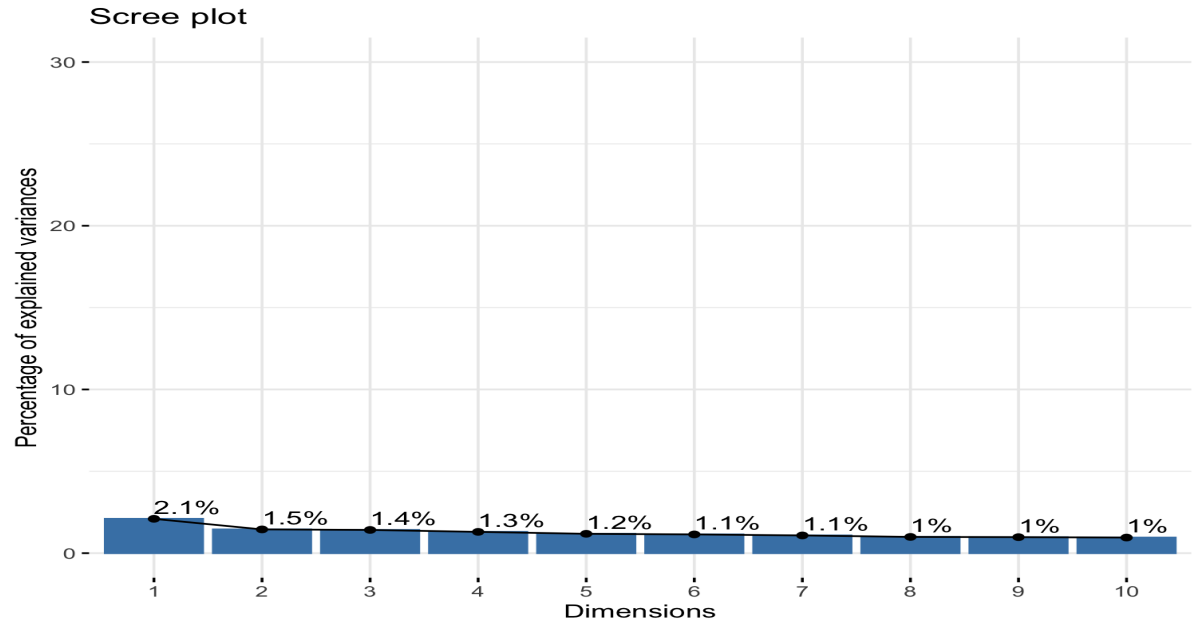
Il est important de noter que chaque fois qu'on rajoute un facteur à l'analyse, on gagne en inertie totale expliquée, mais on perd en parcimonie. Cela signifie que si l'on inclut trop de dimensions, on risque de surinterpréter les données et de perdre la capacité de généralisation. Cependant, il est intéressant de noter que toutes les dimensions cumulées expliquent 100% de la variance des données.

graphique 30: Valeurs propres ou inertie de chaque axe



Le pourcentage d'inertie expliqué par les 10 premières dimensions est de 13%(graphique 31), ce qui est relativement faible. Cela indique que l'ACM n'est peut-être pas le meilleur outil pour comprendre la structure des données, ou que les variables catégorielles utilisées sont très hétérogènes et ne peuvent pas être facilement représentées par un petit nombre de dimensions.

graphique 31: Valeurs propres ou inertie des dix premiers axes

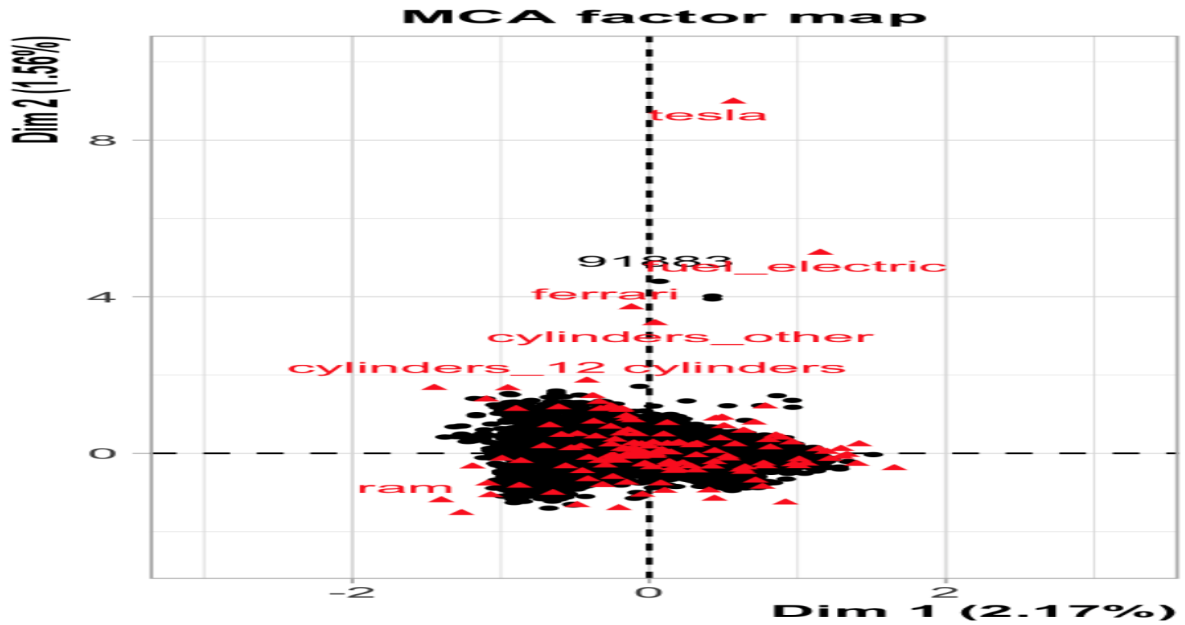


La condensation des modalités vers le centre de la carte factorielle en MCA est tout à fait normale et indique une forte corrélation entre ces modalités. Cela signifie que ces modalités partagent des caractéristiques communes. Les modalités qui se trouvent loin du centre indiquent une spécificité ou une singularité par rapport aux autres modalités.

Les modalités "Tesla" et "Fuel electric" en haut à droite, ainsi que la modalité "Ferrari" en haut au milieu, indiquent une similarité ou une proximité entre ces modalités. Cela peut être dû à des

caractéristiques communes entre ces modalités, telles que le luxe, la performance, l'innovation technologique, etc.

graphique 31: Nuages des observations et des modalités



Le graphique 32 représente les carrés de rapport de corrélation et nous permet de savoir quelles sont les variables qui sont les plus liées à chacunes des dimensions 1 et 2. La variable "title status" est très proche de l'origine des axes, ce qui peut indiquer qu'elle ne contribue pas beaucoup à la variabilité totale des données.

La variable "paint color" a une composante modérée sur l'axe de dim1, mais presque pas sur l'axe de dim2. Cela peut signifier que la couleur de la peinture est un facteur important pour expliquer la variabilité totale des données, mais que cette variable n'est pas corrélée avec d'autres variables importantes sur l'axe de dim2. La variable "cylinder" est placée dans l'extrême droite mais en bas. Cela peut signifier que cette variable a une forte influence sur l'axe de dim1 mais pas sur l'axe de dim2.

Les variables "state", "condition", "fuel" et "transmission" sont un peu éloignées du centre et plus vers le haut. Cela peut signifier que ces variables ont une influence sur les deux axes de manière similaire.

Les variables "type", "manufacturer" et "drive" possèdent les liaisons les plus fortes et les plus proches de 1. Ceci n'est pas très élevé mais compte tenu du nombre très important d'individus, les résultats peuvent être considérés comme significatifs. Cela peut signifier que ces variables ont une forte influence sur la variabilité totale des données et qu'elles sont également corrélées avec d'autres variables importantes.

Par ailleurs, les chiffres dans la sortie du code qui représentent les contributions de chaque catégorie à chaque dimension nous permettent de faire une analyse plus poussée des contributions des variables et de leurs modalités aux dimensions.

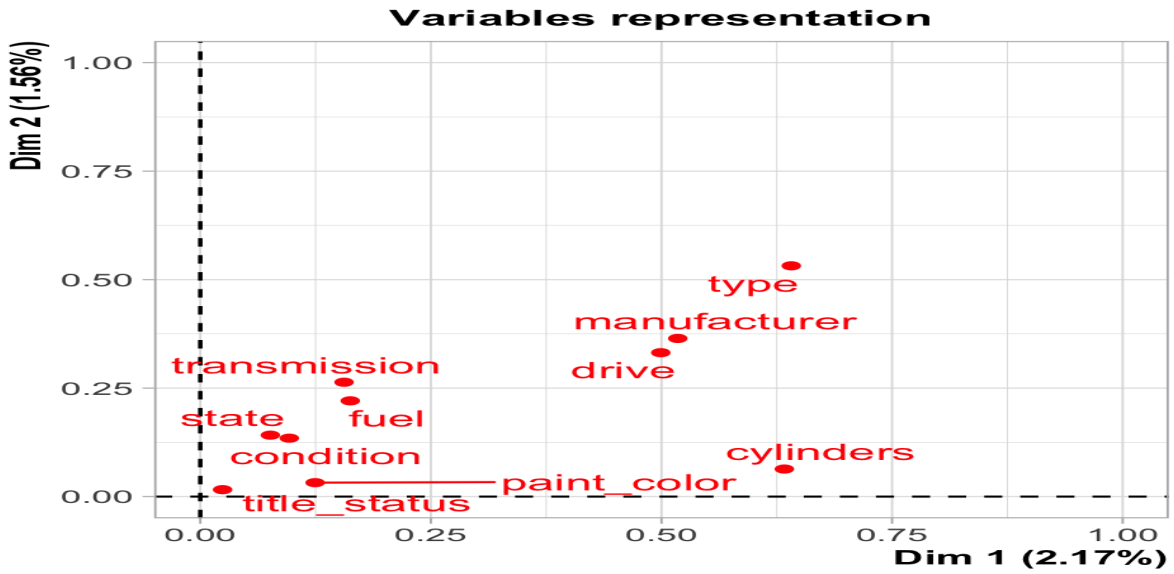
Les dimensions sont des axes latents qui ont été créés à partir de l'analyse des correspondances multiples (ACM) des données d'origine. Les contributions mesurent l'importance de chaque catégorie dans la création de chaque dimension.

Plus précisément, pour chaque catégorie, les valeurs indiquées sous les dimensions 1, 2 et 3 représentent la contribution de cette catégorie à chacune des trois dimensions. Les chiffres sont normalisés pour que la somme des carrés des contributions de chaque catégorie soit égale à 1.

En analysant les chiffres, on peut voir que certaines catégories ont des contributions importantes à une ou plusieurs dimensions, tandis que d'autres ont des contributions faibles ou nulles. Par exemple, les catégories "ford", "gmc", "subaru" et "dodge" ont des contributions importantes à la première dimension, tandis que les catégories "tesla", "nissan", "hyundai" et "honda" ont des contributions importantes à la deuxième dimension. Les catégories "jeep", "cadillac" et "acura" ont des contributions importantes à la troisième dimension.

Ces informations peuvent être utilisées pour comprendre la structure sous-jacente des données d'origine. Par exemple, les catégories qui ont des contributions importantes à une même dimension peuvent être regroupées ensemble dans une analyse ultérieure. Les dimensions peuvent également être utilisées pour créer des visualisations qui aident à interpréter les données.

Graphique 32: Représentation des variables



3. Modélisation

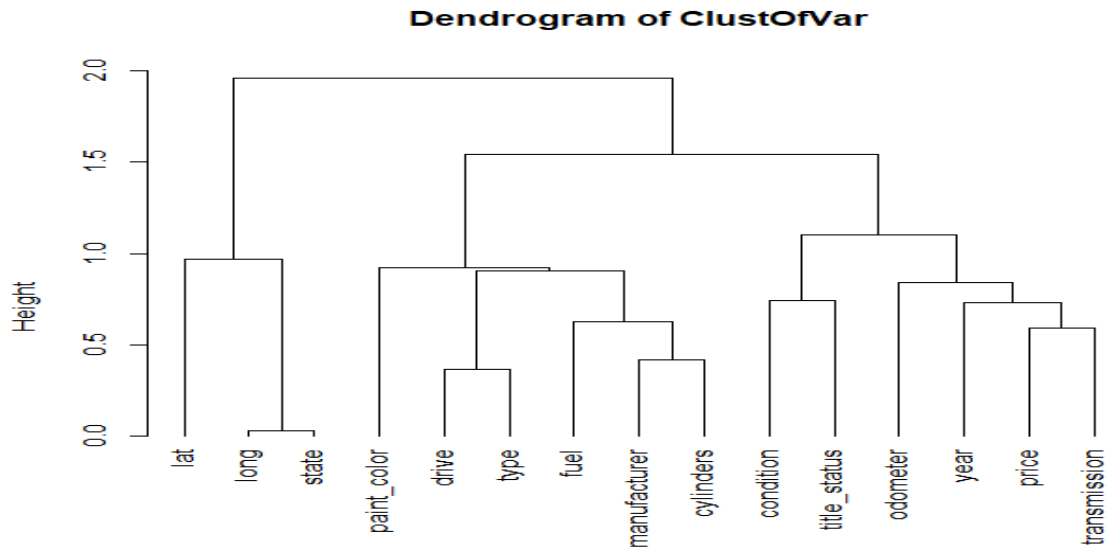
3.1 Modèle non supervisé : La classification hiérarchique

Nous allons dans un premier temps faire une classification hiérarchique des variables. Nous nous permettons de rappeler à ce stade à quel point l'objectif est de rassembler dans une même classe des variables qui sont fortement liées les unes aux autres et fournissant le même type d'information.

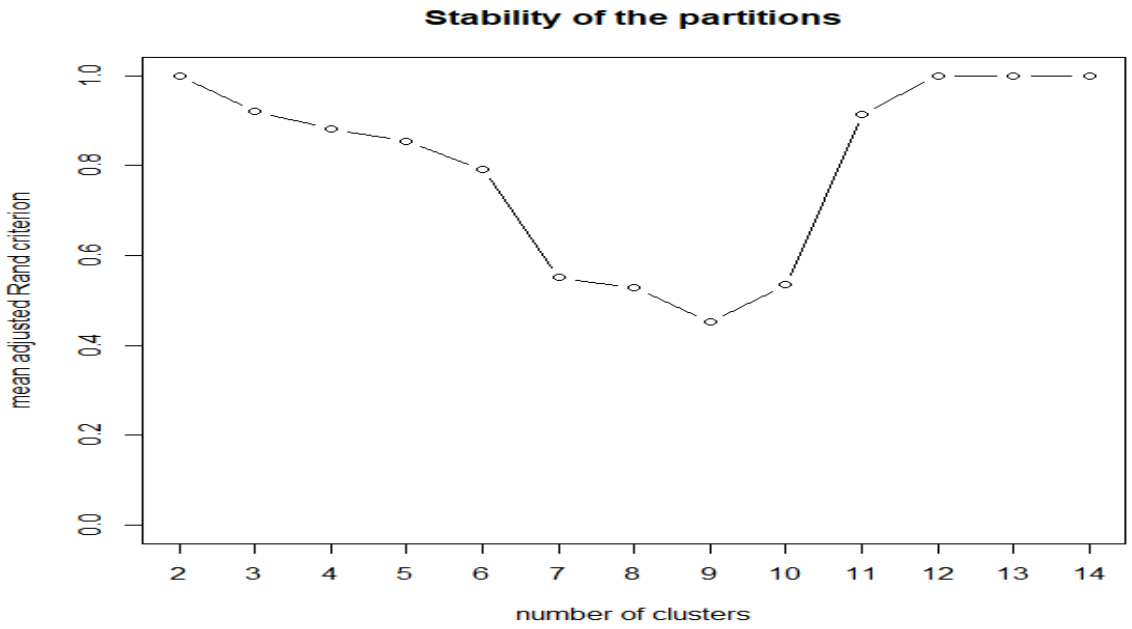
Notre Base de données est constituée principalement par des variables quantitatives et majoritairement par des variables qualitatives. Donc il faut choisir un modèle non supervisé qui peut créer des classes tout en combinant ces deux types de variables. Dans ce cadre nous avons utilisé le package R "ClustOfVar" qui a été développé pour répondre au problème de la classification des variables mixtes. Le critère d'homogénéité d'une classe est la somme des carrés des corrélations entre les variables quantitatives et les ratios de corrélation entre les variables qualitatives.

En effet, la classification hiérarchique des données sur les véhicules d'occasions considérées est déterminée via la fonction **hclustvar** du package R **ClustOfVar**.

Figure : Dendrogramme de la classification hiérarchique

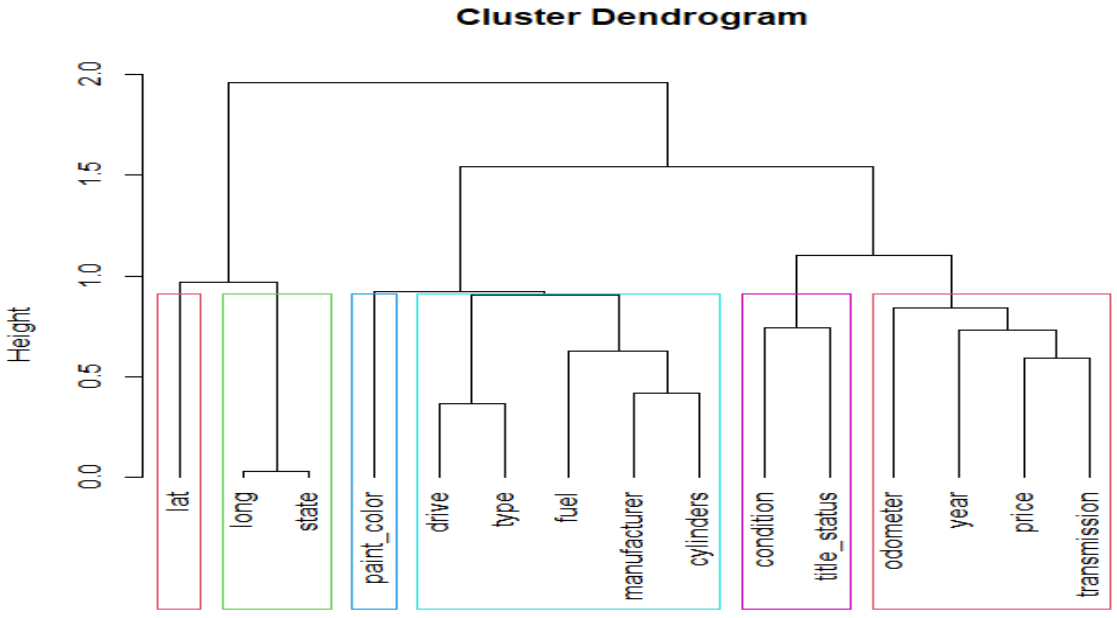


Afin de déterminer un nombre K raisonnable de classes à considérer pour la suite de l'étude, nous pouvons utiliser le critère de stabilité. La figure x permet d'avoir une représentation graphique de ce critère.



En se basant notre analyse sur graphique ci-dessus, nous décidons de retenir k = 6. c'est le nombre de classe le plus raisonnable en termes d'interprétation et stabilité. Un nombre plus élevé de classes n'est pas plus stable sauf si on fixe a nombre supérieurs ou égale 12 or ce dernier n'est pas efficace en termes de synthèse d'informations.

Figure : Dendrogramme de la classification hiérarchique après fixation de cluster



Les tableaux ci-dessous permettent de présenter les contenus et surtout la relation existant entre une variable et son cluster autrement dit avec une variable synthétique synthétisant la classe. Cette liaison est mesuré par deux indices : **squared loading et la corrélation.**

On cherche a trouver un squared loading proche de 1 que se soit pour les variables quantitatives et qualitatives afin de valider une bonne représentation dans cette classe. Par contre, l'indice de corrélation est exclusif pour les variables quantitatives et il permet de terminer la corrélation ainsi que le sens de relation entre ce dernier et la variable de synthèse (soit une relation positive ou bien négative).

Il faut noter qu' à l'intérieur de chaque cluster, les variables sont classées selon leur qualité de représentation dans le groupe.

Commençons par le premier cluster qui contient notre variable qu'on cherche à prédire. Selon ce groupe de variable on constate que la variable prix est la variable la plus représentée dans cette classe et elle très corrélé avec la variable de synthèse (0.77). Cette classe présente un regroupement des variables caractérisant un véhicule ainsi que l'âge avec le prix. Par exemple, on peut dire que plus la distance parcourue (odometer) par une voiture augmente, le prix de cette dernière va diminuer.

Le 3eme cluster est un regroupement des véhicules qui partagent les mêmes localisations géographiques. Même chose pour la classe.

4eme cluster regroupe plutôt les aspects du véhicule donc il y a un ensemble de véhicules qui se ressemblent en termes de caractéristiques .

Le 5eme cluster présente un regroupement de véhicules qui partagent les mêmes états c-a-d une voiture en bon état(condition) va surement avoir un type de titre qui signifie que la voiture n'a pas subi de dommages majeur et donc il aura aucun obstacle juridique à la vente de cette dernière.

Tableau 7 : Cluster 1

	squared loading	**correlation**
price	0.59	0.77
transmission	0.51	NA
year	0.44	0.66
odometer	0.29	-0.54

Tableau 8 : Cluster 2

	squared loading	correlation
paint_color	1	1

Tableau 9: Cluster 3

	squared loading	correlation
long	0.59	0.77
state	0.51	NA

Tableau 10 : Cluster 4

	squared loading	correlation
type	0.68	NA
cylinders	0.65	NA
drive	0.58	NA
manufacturer	0.58	NA
fuel	0.20	NA

Tableau 11 : Cluster 5

	squared loading	correlation
condition	0.59	0.77
title_status	0.51	NA

Tableau 12: cluster 6

	squared loading	correlation
lat	1	1

Finalement, Gain in cohesion est le pourcentage d'homogénéité que représente chaque partition et il est égale à 55.36 %. Alors on peut dire que les véhicules de chaque groupe sont semblables à l'ordre de 55.36 %.

3.2 Modèles supervisés

Pour prédire les prix des voitures vendues sur Craigslist, nous avons sélectionné plusieurs variables avec diverses caractéristiques telles que l'année, l'état, le kilométrage et la localisation du véhicule. Nous avons ensuite élaboré trois modèles supervisés différents pour prédire les prix à l'aide de ces caractéristiques. Dans le modèle de régression multiple, nous avons utilisé toutes les caractéristiques disponibles après retraitement de la base de données lors de la phase exploratoire pour prédire le prix. Pour le modèle de régression logistique, nous avons utilisé un seuil de prix supérieur ou inférieur à 20 000$ pour classer la voiture comme chère ou pas chère. Enfin, dans le modèle de gradient boosting, nous avons utilisé un ensemble d'arbres de décision pour prédire le prix.

Nous avons évalué les performances des modèles à l'aide de trois mesures différentes : R-carré, RMSE et précision.

Tableau 12: Résultats de nos modèles

	Régression multiple	**Régression logistique**	**Gradient boosting**
accuracy	NA	0.895	NA
R squared	0.4751695	0.938	0.7522249
RMSE	9457.646	0.277	6613.138

Le tableau ci-dessus montre que le modèle de régression logistique est le plus performant, avec une valeur R-carré de 0,938, une RMSE de 0,277 et une précision de 0,895. Le modèle de régression logistique a été capable de classifier si une voiture était chère ou non sur la base du seuil que nous avions fixé, et cette classification a donné de très bons résultats, avec une précision de 0,895. Le modèle de régression multiple présentait la valeur R-carré la plus faible (0,475), ce qui indique qu'il expliquait le moins de variance dans la variable dépendante qui est le prix. Le modèle de gradient boosting avait un R-carré de 0,752, ce qui était mieux que le modèle de régression multiple mais moins bien que le modèle de régression logistique.

En conclusion, le modèle de régression logistique a été le plus performant pour prédire les prix des voitures vendues sur Craigslist, car il présentait la plus grande précision, le plus faible RMSE et la valeur R au carré la plus élevée.

Conclusion

En conclusion, ce projet a permis d'explorer l'utilisation d'algorithmes d'apprentissage automatique pour l'analyse de grandes quantités de données de voitures d'occasion disponibles à la vente sur Craigslist. À travers une méthodologie en plusieurs étapes, nous avons pu comprendre la structure de notre ensemble de données en utilisant une analyse exploratoire univariée et bivariée, ainsi qu'une analyse multidimensionnelle par composantes principales et multiples.

Nous avons également développé des modèles de prédiction pour regrouper les véhicules en fonction de leur prix et prédire le prix d'un véhicule en fonction de ses caractéristiques. Les modèles non supervisés et supervisés ont donné des résultats prometteurs, mettant en évidence l'importance des caractéristiques telles que le modèle, l'année, le type de transmission et odomètre dans la prédiction des prix des voitures d'occasion.

En somme, ce projet a démontré la pertinence et l'efficacité de l'utilisation de l'apprentissage automatique dans l'analyse de grandes quantités de données dans le domaine de la vente de voitures d'occasion. Ces résultats pourraient avoir des implications importantes pour les vendeurs et les acheteurs de voitures, en aidant à établir des prix plus précis et équitables pour les transactions de voitures d'occasion.

Annexes

Annexe 1 : statistiques des variables numériques

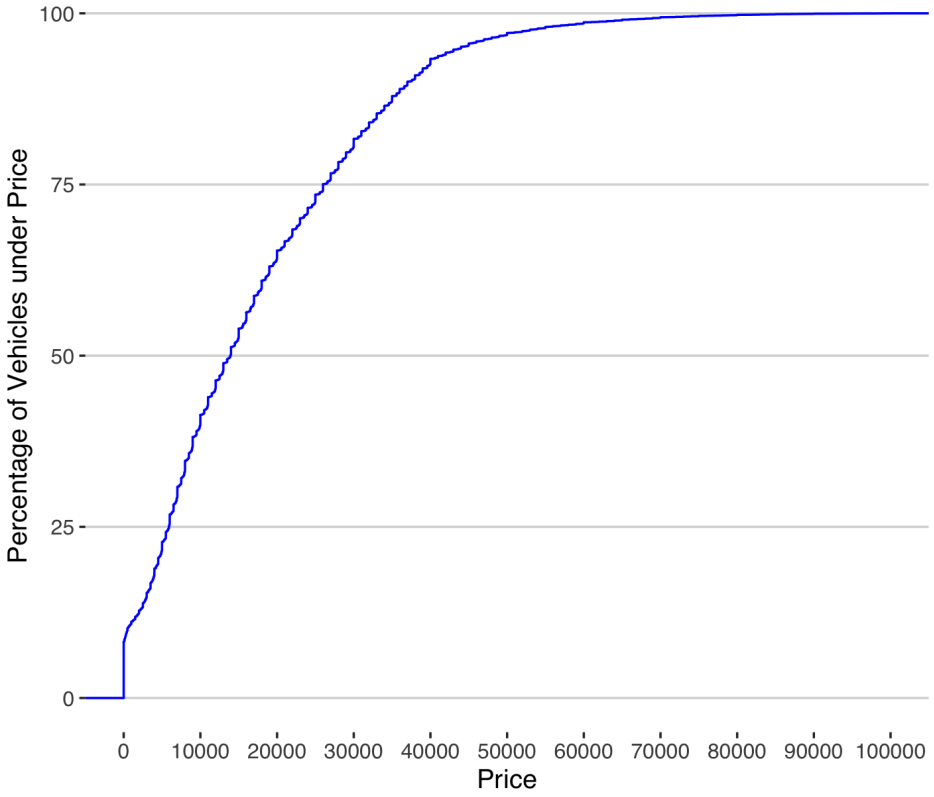
Column Name	Data Type
id	numeric
region	character
price	numeric
year	integer

manufacturer	character
model	character
condition	character
cylinders	character
fuel	character
odometer	integer
title_status (carte grise)	character
transmission	character
VIN	character
drive	character
size	character
type	character
paint_color	character
description	character
county	logical
state	character
lat	numeric
long	numeric
posting_date	character

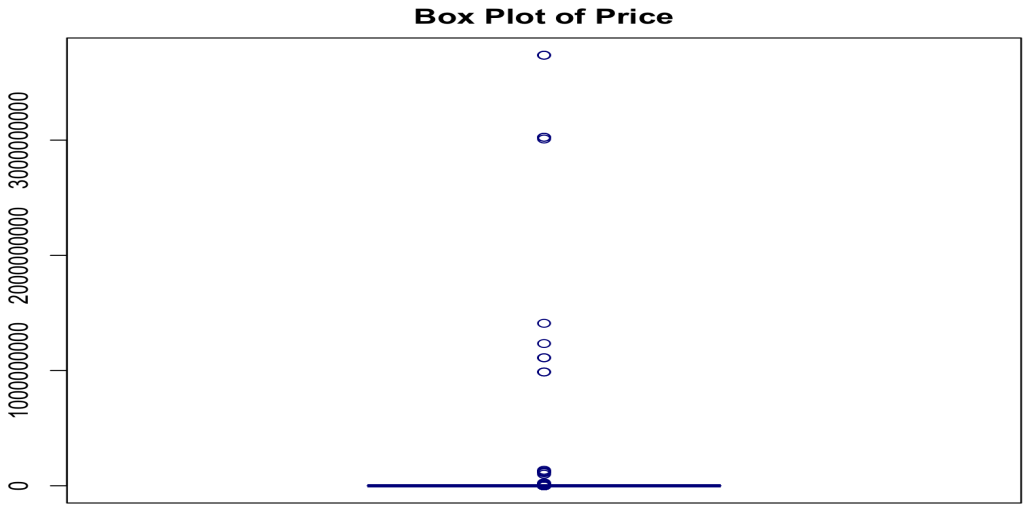
Annexe 02: tableau de corrélation

	id	price	year	odometer	lat	long
id	1.000000000	-0.0027794659	-0.059039506	0.010721063	-0.069388249	-0.1218643951
price	-0.002779466	1.0000000000	-0.004925082	0.010032267	0.000357334	-0.0004084214
year	-0.059039506	-0.0049250819	1.000000000	-0.157215107	-0.014676947	-0.0014102574
odometer	0.010721063	0.0100322667	-0.157215107	1.000000000	-0.001459185	0.0098072957
lat	-0.069388249	0.0003573340	-0.014676947	-0.001459185	1.000000000	-0.1280881640
long	-0.121864395	-0.0004084214	-0.001410257	0.009807296	-0.128088164	1.0000000000

Annexe 03: pourcentage de véhicules de moins de 100000 $



Annexe 04: Boxplot des prix avec les valeurs aberrantes



Annexe 05: Nombre de voitures par marque

manufacturer	vehicle_count
acura	5978
alfa-romeo	897

aston-martin	24
audi	7573
bmw	14699
buick	5501
cadillac	6953
chevrolet	55064
chrysler	6031
datsun	63
dodge	13707
ferrari	95
fiat	792
ford	70985
gmc	16785
harley-davidson	153
honda	21269
hyundai	10338
infiniti	4802
jaguar	1946
jeep	19014
kia	8457
land rover	21
lexus	8200
lincoln	4220
mazda	5427
mercedes-benz	11817
mercury	1184
mini	2376
mitsubishi	3292
morgan	3
nissan	19067
pontiac	2288
porsche	1384
ram	18342
rover	2113

saturn	1090
subaru	9495
tesla	868
toyota	34202
volkswagen	9345
volvo	3374

Annexe 06: Prix moyen de voitures par marque

manufacturer	**mean_price**
acura	19842.870
alfa-romeo	28237.349
aston-martin	53494.542
audi	23574.850
bmw	19022.882
buick	14344.954
cadillac	19439.521
chevrolet	18344.448
chrysler	10197.386
datsun	15149.667
dodge	14413.883
ferrari	94214.300
fiat	11890.518
ford	19154.818
gmc	22962.120
harley-davidson	12129.791
honda	10751.364
hyundai	10719.659
infiniti	19561.849
jaguar	26549.176
jeep	18038.534
kia	11083.325
land rover	7911.095
lexus	19206.153
lincoln	19514.347
mazda	12345.608
mercedes-benz	19647.345

mercury	5482.536
mini	14163.475
mitsubishi	13743.071
morgan	13100.000
nissan	11747.186
pontiac	8176.903
porsche	31352.907
ram	26802.966
rover	27182.899
saturn	5051.507
subaru	13055.513
tesla	38354.456
toyota	15774.541
volkswagen	12537.099
volvo	17854.130

Annexe 07: Nombre de voitures par année de fabrication

year	**vehicle_count**
1910	2
1913	2
1915	1
1916	2
1918	1
1920	2
1921	2
1922	3
1923	36
1924	9
1925	8
1926	16
1927	37
1928	38
1929	57
1930	68
1931	58
1932	57

1933	25
1934	44
1935	23
1936	43
1937	70
1938	38
1939	54
1940	81
1941	67
1942	15
1943	1
1944	3
1945	2
1946	58
1947	71
1948	100
1949	89
1950	106
1951	98
1952	110
1953	105
1954	100
1955	226
1956	160
1957	174
1958	73
1959	93
1960	120
1961	82
1962	138
1963	234
1964	270
1965	365
1966	424
1967	357
1968	425
1969	409

1970	345
1971	312
1972	409
1973	334
1974	280
1975	204
1976	242
1977	273
1978	345
1979	387
1980	272
1981	214
1982	217
1983	257
1984	387
1985	469
1986	523
1987	532
1988	528
1989	571
1990	599
1991	608
1992	626
1993	712
1994	968
1995	1246
1996	1302
1997	1724
1998	1988
1999	3094
2000	3572
2001	4443
2002	5587
2003	7151
2004	8971
2005	10622
2006	12763

2007	14873
2008	17150
2009	12185
2010	15829
2011	20341
2012	23898
2013	30794
2014	30283
2015	31538
2016	30434
2017	36420
2018	36369
2019	25375
2020	19298
2021	2396
2022	133

Annexe 08: Prix moyen de voitures par année de fabrication

year_intervals	**mean_price**
[1900, 1910)	4188.972
(1910, 1920)	21890.000
(1920, 1930)	15744.355
(1930, 1940)	26314.631
(1940, 1950)	14042.727
(1950, 1960)	19894.215
(1960, 1970)	19677.288
(1970, 1980)	13109.577
(1980, 1990)	9748.076
(1990, 2000)	8348.253
(2000, 2010)	8194.959
(2010, 2020)	21347.586
(2020, 2022)	16436.353

Annexe 09: Nombre de voitures par Etat

state	vehicle_count
ak	3474
al	4955
ar	4038
az	8679
ca	50614
co	11088
ct	5188
dc	2970
de	949
fl	28511
ga	7003
hi	2964
ia	8632
id	8961
il	10387
in	5704
ks	6209
ky	4149
la	3196
ma	8174
md	4778
me	2966
mi	16900
mn	7716
mo	4293
ms	1016
mt	6294
nc	15277
nd	410
ne	1036
nh	2981
nj	9742
nm	4425
nv	3194

ny	19386
oh	17696
ok	6792
or	17104
pa	13753
ri	2320
sc	6327
sd	1302
tn	11066
tx	22945
ut	1150
va	10732
vt	2513
wa	13861
wi	11398
wv	1052
wy	610

Annexe 10 : Prix moyen de voitures par Etat

state	**mean_price**
ak	23745.82
al	20144.24
ar	18061.28
az	19482.62
ca	16667.41
co	18060.01
ct	14659.31
dc	14284.37
de	15492.34
fl	17521.80
ga	18230.80
hi	19560.16
ia	16183.89
id	19923.51

il	16536.23
in	19088.37
ks	18954.53
ky	19061.85
la	17841.27
ma	15582.39
md	17227.75
me	13782.82
mi	15360.03
mn	15895.26
mo	20807.76
ms	16532.39
mt	23515.94
nc	16373.67
nd	18305.96
ne	18701.60
nh	15606.13
nj	14952.82
nm	19041.87
nv	20574.81
ny	16963.89
oh	15112.72
ok	17767.57
or	15670.48
pa	15280.65
ri	16424.11
sc	18823.20
sd	18369.45
tn	19335.31
tx	19636.52
ut	25099.98
va	13849.88
vt	16579.10
wa	22221.34
wi	16418.50

wv	24317.42
wy	21104.07

Annexe 11: Nombre de voitures par type de fuel utilisé

fuel	**vehicle_count**
diesel	30062
electric	1698
gas	356209
hybrid	5170
other	30728

Annexe 12: Prix moyen de voitures par type de fuel utilisé

fuel	**mean_price**
diesel	29477.86
electric	24648.36
gas	15714.64
hybrid	14582.43
other	25531.63

Annexe 13: Nombre de voitures par cylindres

cylinders	**vehicle_count**
10 cylinders	1455
12 cylinders	209
3 cylinders	655
4 cylinders	77642
5 cylinders	1712
6 cylinders	94169
8 cylinders	72062
other	1298

Annexe 14 : Prix moyen de voitures par cylindres

cylinders	**mean_price**

10 cylinders	20557.85
12 cylinders	42659.22
3 cylinders	12297.00
4 cylinders	10429.71
5 cylinders	7613.51
6 cylinders	17417.44
8 cylinders	22158.70
other	17110.08

Annexe 15: Nombre de voitures par Odometer

odometer_interval	**vehicle_count**
0-99999	244878
100000-249999	170150
250000-499999	6029
>500000	1423

Annexe 16: Prix moyen de voitures par intervalle d'Odomètre

odometer_intervals	**mean_price**
0-99999	22530.307
100000-249999	10602.188
250000-499999	9294.382
>500000	14391.061

Annexe 17: Nombre de voitures par title status

title_status	**vehicle_count**
clean[1]	405117

[1]Ce type de titre signifie que la voiture n'a pas subi de dommages majeurs et n'a pas été impliquée dans un accident important. Il n'y a aucun obstacle juridique à la vente de cette voiture.

lien[2]	1422
missing[3]	814
parts only[4]	198
rebuilt[5]	7219
salvage[6]	3868

Annexe 18: Prix moyen de voitures par Title status

title_status	**mean_price**
clean	17653.089
lien	22049.285
missing	5035.012
parts only	3101.660
rebuilt	12739.880
salvage	9713.573

Annexe 19 : Nombre de voitures par couleur

paint_color	**vehicle_count**
black	62861
blue	31223
brown	6593
custom	6700

---

[2] Un titre de lien est utilisé lorsqu'un propriétaire a un prêt sur la voiture. Le prêteur (banque, société de crédit, etc.) est inscrit sur le titre en tant que détenteur du privilège de prêt jusqu'à ce que le prêt soit remboursé.

[3]Cela signifie que le propriétaire n'a pas de titre pour la voiture.

[4]Cela signifie que la voiture ne peut être vendue que pour des pièces détachées et non pour être conduite sur la route. Ce type de titre est souvent utilisé pour les voitures qui ont été considérées comme "perte totale" par les compagnies d'assurance.

[5] Un titre de reconstruction est attribué à une voiture qui a été endommagée et réparée. Cela signifie que la voiture a subi des dommages majeurs, mais a été réparée et jugée apte à être conduite sur la route après une inspection de sécurité.

[6] Ce type de titre est attribué à une voiture qui a subi des dommages importants (comme une collision, une inondation, etc.) et qui a été jugée économiquement irréparable par l'assureur. La voiture peut être réparée, mais elle peut être considérée comme présentant un risque accru de problèmes mécaniques et de sécurité.

green	7343
grey	24416
orange	1984
purple	687
red	30473
silver	42970
white	79285
yellow	2142

Annexe 20: Prix moyen de voitures par couleur

paint_color	**mean_price**
black	20319.67
blue	16180.25
brown	15175.67
custom	15383.07
green	12912.40
grey	14692.43
orange	18116.32
purple	15347.15
red	18472.16
silver	15669.61
white	20690.71
yellow	18302.65

Annexe 21: Nombre de voitures par type

type	vehicle_count
SUV	77284
bus	517
convertible	7731
coupe	19204

hatchback	16598
mini-van	4825
offroad	609
other	22110
pickup	43510
sedan	87056
truck	35279
van	8548
wagon	10751

Annexe 22: Prix moyen de voitures par type

type	mean_price
SUV	16051.475
bus	14105.617
convertible	19758.508
coupe	21680.948
hatchback	14384.513
mini-van	9234.088
offroad	15813.094
other	24743.588
pickup	27244.912
sedan	13051.534

truck	23461.677
van	17455.713
wagon	13273.910

Annexe 23: Nombre de voitures par condition

condition	**vehicle_count**
excellent	85780
fair	5259
good	111383
like new	16309
new	830
salvage	440

Annexe 24: Prix moyen de voitures par condition

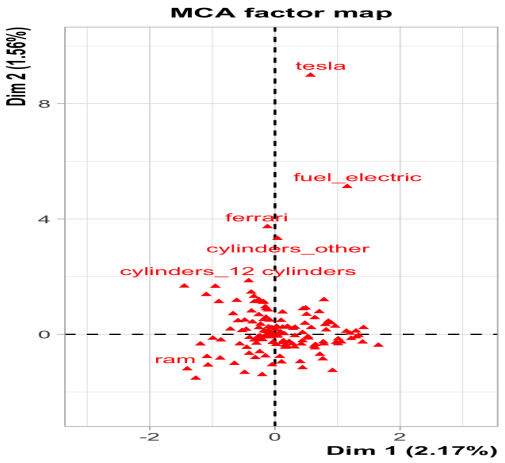
condition	**mean_price**
excellent	14500.763
fair	4011.226
good	20570.625
like new	18196.041
new	23657.267
salvage	3605.534

Annexe 25 : Nombre de voitures vendues par jour

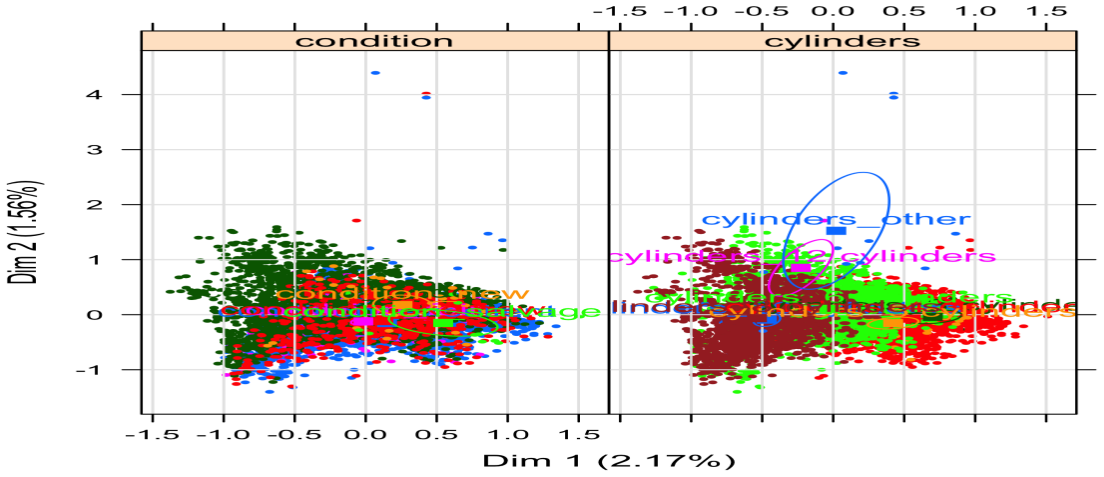
month_day	**total_vehicle_count**
Apr-10	6075
Apr-11	4422
Apr-12	6866
Apr-13	6598
Apr-14	7794
Apr-15	8231

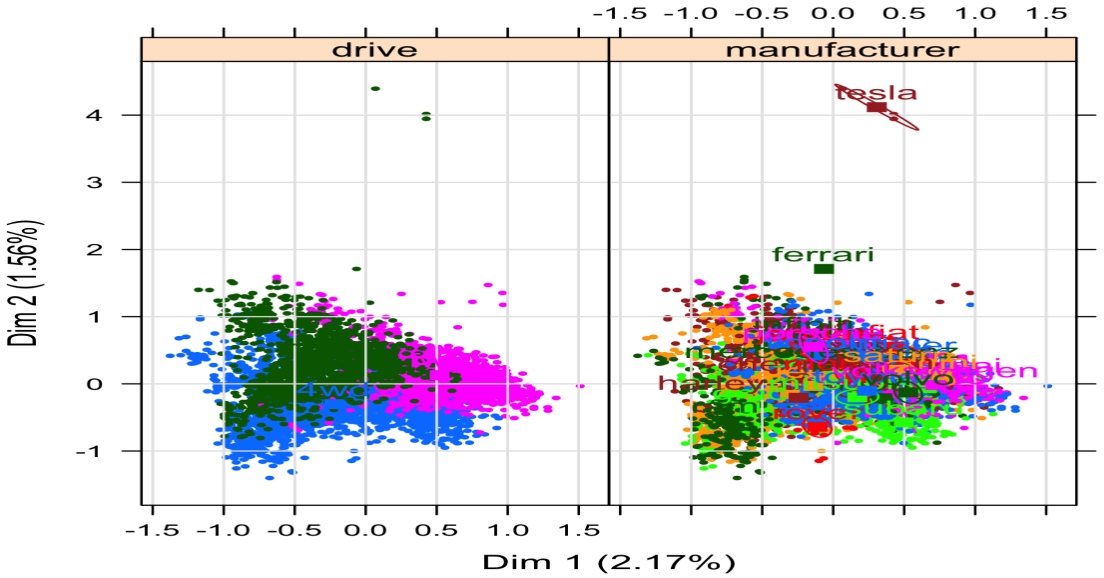
Apr-16	9167
Apr-17	8399
Apr-18	5319
Apr-19	8559
Apr-20	8659
Apr-21	9921
Apr-22	11459
Apr-23	14637
Apr-24	11279
Apr-25	6783
Apr-26	12727
Apr-27	13938
Apr-28	14803
Apr-29	16269
Apr-30	22173
Apr-4	3403
Apr-5	5802
Apr-6	5578
Apr-7	6111
Apr-8	7030
Apr-9	7194
May-1	20580
May-2	14010
May-3	27337
May-4	22899

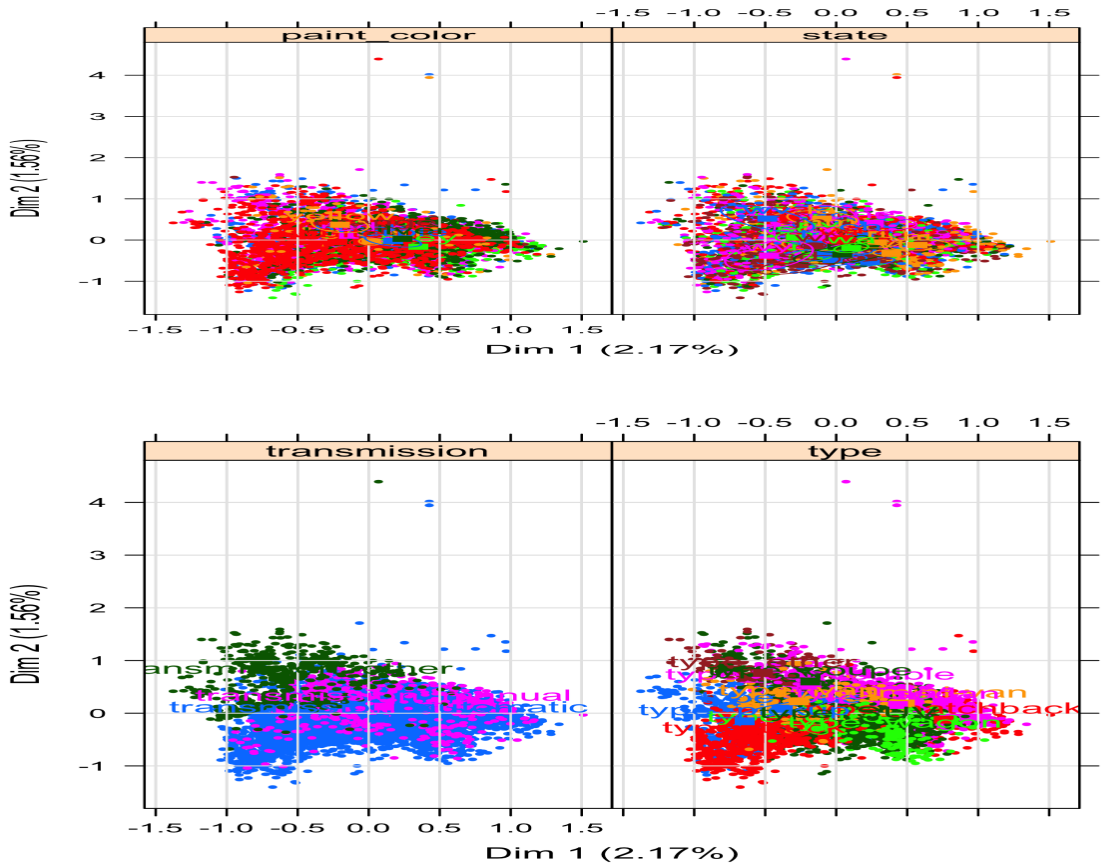
Annexe 26: factor map des modalités



Annexe 28: Coloration des individus par groupes et avec les ellipses de confiance autour de la moyenne des groupes







Annexe: Summary de la régression multiple

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-7.151e+05	2.119e+04	-33.754	< 2e-16	***
year	3.768e+02	1.022e+01	36.859	< 2e-16	***
conditionfair	-7.558e+03	4.724e+02	-16.000	< 2e-16	***
conditiongood	-1.591e+03	1.930e+02	-8.246	< 2e-16	***
conditionlikenew	3.898e+03	2.765e+02	14.100	< 2e-16	***
conditionnew	6.758e+03	1.277e+03	5.290	1.24e-07	***
conditionsalvage	-7.073e+03	1.693e+03	-4.177	2.97e-05	***
cylinders12 cylinders	7.308e+03	3.950e+03	1.850	0.064354	.

cylinders3 cylinders	-3.275e+03	2.131e+03	-1.537	0.124329	
cylinders4 cylinders	-2.182e+03	9.302e+02	-2.346	0.019006	*
cylinders5 cylinders	-2.849e+03	1.316e+03	-2.165	0.030397	*
cylinders6 cylinders	2.013e+02	9.150e+02	0.220	0.825901	
cylinders8 cylinders	3.380e+03	9.118e+02	3.707	0.000211	***
cylindersother	-4.007e+02	1.927e+03	-0.208	0.835296	
fuelelectric	1.526e+02	2.198e+03	0.069	0.944651	
fuelgas	-9.736e+03	3.927e+02	-24.792	< 2e-16	***
fuelhybrid	-8.492e+03	8.667e+02	-9.798	< 2e-16	***
fuelother	-4.750e+03	8.155e+02	-5.825	5.85e-09	***
odometer	-5.682e-03	3.927e-04	-14.467	< 2e-16	***
title_statusliен	2.006e+03	8.862e+02	2.264	0.023576	*
title_statusmissing	1.022e+02	1.909e+03	0.054	0.957324	
title_statusparts only	-3.755e+03	3.693e+03	-1.017	0.309265	
title_statusrebuilt	-4.853e+02	4.644e+02	-1.045	0.295960	
title_statussalvage	-2.772e+03	7.479e+02	-3.706	0.000212	***
transmissionmanual	2.673e+03	3.387e+02	7.892	3.21e-15	***
transmissionother	1.115e+04	3.366e+02	33.115	< 2e-16	***
drivefwd	-4.065e+03	2.569e+02	-15.821	< 2e-16	***
driverwd	-8.870e+02	2.603e+02	-3.408	0.000656	***

typeconvertible	-8.223e+02	2.624e+03	-0.313	0.753973	
typecoupe	5.933e+02	2.602e+03	0.228	0.819651	
typehatchback	-6.475e+03	2.622e+03	-2.469	0.013544	*
typemini-van	-5.614e+03	2.640e+03	-2.126	0.033507	*
typeoffroad	-2.013e+03	2.906e+03	-0.693	0.488504	
typeother	8.090e+02	2.638e+03	0.307	0.759126	
typepickup	-1.775e+03	2.598e+03	-0.683	0.494619	
typesedan	-6.190e+03	2.588e+03	-2.391	0.016795	*
typeSUV	-5.419e+03	2.592e+03	-2.091	0.036551	*
typetruck	-1.937e+03	2.589e+03	-0.748	0.454372	
typevan	-3.103e+03	2.621e+03	-1.184	0.236377	
typewagon	-7.423e+03	2.634e+03	-2.819	0.004832	**
paint_colorblue	-1.162e+03	3.109e+02	-3.738	0.000186	***
paint_colorbrown	-2.051e+03	5.258e+02	-3.902	9.60e-05	***
paint_colorcustom	-6.026e+02	5.894e+02	-1.023	0.306561	
paint_colorgreen	-1.606e+03	4.997e+02	-3.214	0.001312	**
paint_colorgrey	-1.527e+03	3.149e+02	-4.850	1.25e-06	***
paint_colororange	1.860e+03	1.074e+03	1.731	0.083424	.
paint_colorpurple	6.109e+02	1.541e+03	0.396	0.691756	
paint_colorred	-2.010e+02	3.103e+02	-0.648	0.517132	
paint_colorsilver	-2.397e+03	2.824e+02	-8.490	< 2e-16	***

paint_colorwhite	-1.802e+02	2.556e+02	-0.705	0.480797	
paint_coloryellow	3.759e+02	8.940e+02	0.420	0.674160	
stateal	-7.682e+03	2.018e+03	-3.807	0.000142	***
statear	-9.468e+03	2.017e+03	-4.695	2.70e-06	***
stateaz	-7.947e+03	1.753e+03	-4.532	5.90e-06	***
stateca	-7.445e+03	1.540e+03	-4.836	1.34e-06	***
stateco	-9.219e+03	1.622e+03	-5.682	1.36e-08	***
statect	-1.226e+04	2.018e+03	-6.078	1.25e-09	***
statedc	-1.267e+04	2.106e+03	-6.017	1.83e-09	***
statede	-1.169e+04	2.382e+03	-4.910	9.22e-07	***
statefl	-1.157e+04	2.094e+03	-5.523	3.39e-08	***
stategа	-1.010e+04	2.027e+03	-4.984	6.30e-07	***
statehi	9.508e+02	2.526e+03	0.376	0.706584	
stateia	-1.094e+04	1.686e+03	-6.487	9.07e-11	***
stateid	-1.038e+04	1.447e+03	-7.168	8.00e-13	***
stateil	-1.106e+04	1.771e+03	-6.247	4.31e-10	***
statein	-1.061e+04	1.823e+03	-5.820	6.01e-09	***
stateks	-9.554e+03	1.773e+03	-5.388	7.24e-08	***
stateky	-1.005e+04	1.913e+03	-5.253	1.52e-07	***
statela	-1.038e+04	2.264e+03	-4.583	4.62e-06	***
statema	-1.281e+04	1.970e+03	-6.506	8.01e-11	***
statemd	-1.039e+04	2.033e+03	-5.110	3.27e-07	***
stateme	-1.257e+04	2.135e+03	-5.889	3.99e-09	***
statemi	-1.129e+04	1.750e+03	-6.452	1.14e-10	***
statemn	-1.149e+04	1.628e+03	-7.057	1.79e-12	***
statemo	-8.835e+03	1.886e+03	-4.686	2.82e-06	***
statems	-1.296e+04	2.365e+03	-5.478	4.38e-08	***
statemt	-8.240e+03	1.567e+03	-5.260	1.46e-07	***
statenc	-9.220e+03	1.947e+03	-4.737	2.20e-06	***

statend	-1.260e+04	2.062e+03	-6.109	1.03e-09	***
statene	-8.331e+03	2.080e+03	-4.006	6.20e-05	***
statenh	-1.221e+04	2.113e+03	-5.778	7.73e-09	***
statenj	-1.311e+04	1.959e+03	-6.692	2.29e-11	***
statenm	-8.619e+03	1.832e+03	-4.703	2.58e-06	***
statenv	-7.682e+03	1.857e+03	-4.137	3.53e-05	***
stateny	-1.207e+04	1.876e+03	-6.432	1.30e-10	***
stateoh	-1.305e+04	1.806e+03	-7.228	5.18e-13	***
stateok	-9.792e+03	1.842e+03	-5.316	1.08e-07	***
stateor	-9.319e+03	1.367e+03	-6.816	9.75e-12	***
statepa	-1.204e+04	1.898e+03	-6.344	2.31e-10	***
stateri	-1.435e+04	2.133e+03	-6.728	1.79e-11	***
statesc	-9.857e+03	2.040e+03	-4.832	1.37e-06	***
statesd	-9.938e+03	2.031e+03	-4.893	1.00e-06	***
statetn	-9.135e+03	1.905e+03	-4.794	1.65e-06	***
statetx	-9.300e+03	1.872e+03	-4.967	6.89e-07	***
stateut	-4.398e+03	2.461e+03	-1.787	0.073996	.
stateva	-1.164e+04	1.953e+03	-5.957	2.63e-09	***
statevt	-1.136e+04	1.984e+03	-5.726	1.05e-08	***
statewa	-9.662e+03	1.370e+03	-7.055	1.81e-12	***
statewi	-1.095e+04	1.692e+03	-6.470	1.01e-10	***
statewv	-1.097e+04	2.412e+03	-4.550	5.41e-06	***
statewy	-7.411e+03	2.397e+03	-3.092	0.001989	**
lat	1.292e+02	4.885e+01	2.645	0.008177	**
long	6.499e+01	2.114e+01	3.074	0.002116	**