Mise en place d’un processus de Datamining

Mamouni Ayoub/Moulin Vincent

BUT 3A VCOD

Groupe5/Cas 2

1. Table des matières

2. Vérification des données

Variable
make
fuel-type
aspiration
num-of-doors
body-style
drive-wheels
length
width
height
num-of-cylinders
engine-size
horsepower
city-mpg
highway-mpg
price

Le set de données utiliser est un tableau Excel comportant les informations de voiture utiliser au États-Unis qui vont de la marque jusqu'au prix.

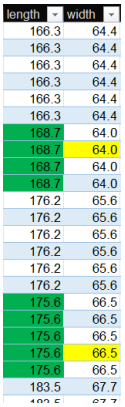
Dans un premier temps nous avons due vérifier si le tableau comporter des erreurs ou bien des valeurs manquantes qui peuvent changer ou biaiser les résultats finals.

1) Valeur à virgule

Après vérification une grande partie des valeurs de type quantitatives comporter soit des valeurs à virgules pour la colonne horsepower et price qui n'avais pas lieux d'être pour le cas horsepower qui n'ont pas de nombre à virgule et price généralement les prix d'achat ne comporte pas de nombre à virgule pour de gros achat comme des voiture donc les valeurs devait être modifier en enlevant les chiffres se situant après la virgule afin d'avoir une meilleur lisibilité sur le jeu de donnée en général même si pour l'analyse cela n'aurait pas eu de grand impact.

2) Valeur manquante

Pour la colonne largeur 3 cellules ne comportent pas de valeur afin de savoir quelle valeur doit être ajouter pour compléter les cellules nous avons décidé d'utiliser les valeurs d'autres voiture pour voir quelle donnée est intégrer généralement pour ce genre de voiture, pour les deux valeurs manquante la longueur était utilisée en comparaison avec les voitures ayant la même longueur que les cellules vides



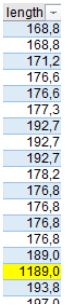
length	width
166.3	64.4
166.3	64.4
166.3	64.4
166.3	64.4
166.3	64.4
166.3	64.4
168.7	64.0
168.7	64.0
168.7	64.0
168.7	64.0
176.2	65.6
176.2	65.6
176.2	65.6
176.2	65.6
176.2	65.6
176.2	65.6
175.6	66.5
175.6	66.5
175.6	66.5
175.6	66.5
175.6	66.5
183.5	67.7

Et pour la troisième valeur les longueurs ne pouvaient pas être utiliser car plusieurs voitures ayant la même longueur mais pas la même largeur nous avons dû choisir d'autre données à croiser, les données était l'aspiration et la carrosserie. (En jaune la valeur manquante, vert foncer la valeur carrosserie et en vert clair la valeur aspiration).

aspiration	num-of-doors	body-style	drive-wheels	length	width
std	four	sedan	fwd	171.7	65.5
std	four	sedan	fwd	171.7	65.5
std	four	sedan	fwd	171.7	65.5
turbo	four	sedan	fwd	171.7	65.5
std	four	sedan	fwd	171.7	65.5
std	two	convertible	fwd	159.3	64.2
std	two	hatchback	fwd	165.7	64.0
std	four	sedan	fwd	180.2	66.9
turbo	four	sedan	fwd	180.2	66.9
std	four	wagon	fwd	183.1	66.9
std	four	sedan	rwd	188.8	67.2
std	four	wagon	rwd	188.8	67.2
std	four	sedan	rwd	188.8	67.2
std	four	wagon	rwd	188.8	67.2
turbo	four	sedan	rwd	188.8	68.9
turbo	four	wagon	rwd	188.8	67.2
std	four	sedan	rwd	188.8	68.9
turbo	four	sedan	rwd	188.8	68.8
std	four	sedan	rwd	188.8	68.9
turbo	four	sedan	rwd	188.8	68.9
turbo	four	sedan	rwd	188.8	68.9

Dans les valeurs manquantes la valeur de nombre de roues motrices ne comporte pas de valeurs nous avons croisé cette valeur avec le nombre de portes et la carrosserie, on peut voir ci-dessous que la valeur « fwd » a été choisies car les voiture avec le même type de carrosserie et du nombre de portes utilise ce type roues motrices.

num-of-doors	body-style	drive-wheels
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd
four	hatchback	fwd



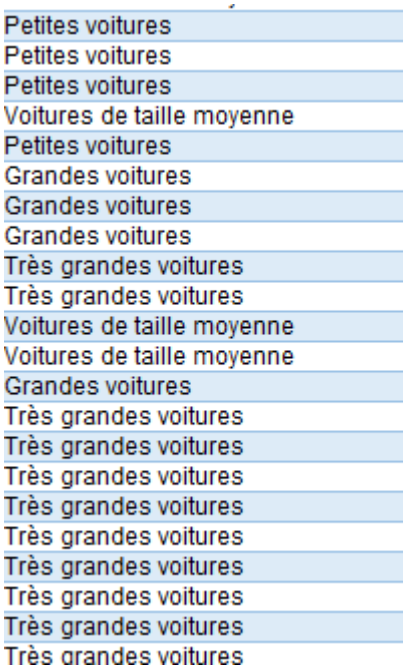
3) Valeur aberrante

Parmi les valeurs aberrantes il y avait la colonne longueur, nous avons constaté que la valeur étant de 1189 cm n'été pas une valeur normale pour la longueur d'une voiture après réflexion nous avons pensé que cela était une erreur de frappe et avons enlevé un 1 de trop pour concorder avec les autres valeurs.

3. Discrétisation des données

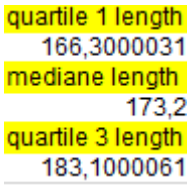
Suite à la correction des différentes données qui peuvent potentiellement erroné la partie statistique, il a fallu discrétiser les données qu'on appelle quantitatives pour séparer les voitures en fonction de leurs données, pour cela on a dû définir ce qu'on appelle le quartile 1, la médiane et le quartile 3 qui vont nous servir à définir des seuils dès lors qu'une donnée se situe en dessous ou au-dessus pour indiquer une description qu'on a au préalablement ajouter.

Par exemple pour la longueur d'une voiture on indiquera parmi quatre descriptions qui sont Petites voitures, Voitures de taille moyenne, Grande voitures et Très grande voitures.



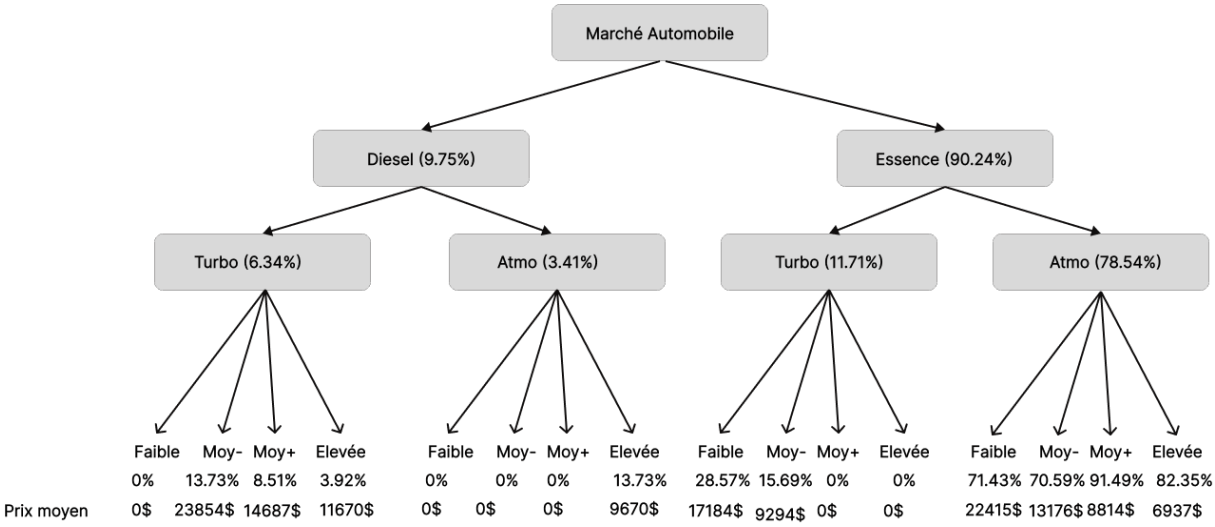
Pour c'est différents description les seuils utiliser sont ceux-ci :

Ce système ci-dessus est utilisé pour les modalités horsepower, Price et consommation moyenne, les descriptions sont appliquées en fonction de la modalité à discrétiser.



Toutes cette partie discrétisation nous sera utile pour calculer les pourcentages en fonction des différentes modalités.

4. Arbre pondérer

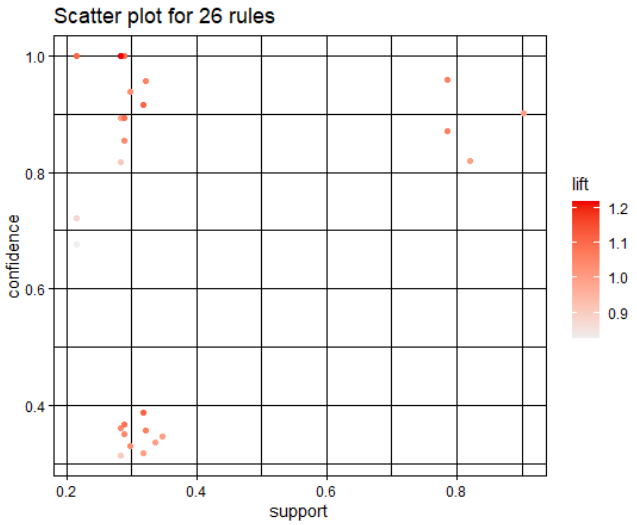


Nous constatons que le marché automobile est largement dominé par les véhicules à essence, représentant 90,24% des ventes, avec une nette préférence pour les moteurs atmosphériques 78,54%, particulièrement dans les segments de prix faible et moyen-, ce qui témoigne d'une forte demande pour des voitures accessibles et économiques. Les moteurs Essence Turbo 11,71% offrent également des opportunités intéressantes dans des gammes légèrement plus performantes. En revanche, le Diesel, avec seulement 9,75% de part de marché, est en recul, bien qu'il conserve une niche dans les catégories moyennes et haut de gamme. Nous avons donc tout intérêt à concentrer nos efforts sur le développement de voitures essence abordables, tout en explorant les niches premium pour le Diesel et en innovant dans les moteurs Turbo pour mieux répondre à ces besoins.

Grâce à l'analyse de cet arbre décisionnel, nous avons pu identifier les niches potentielles du marché automobile. Il met en évidence la domination des moteurs Essence Atmosphériques dans les segments économiques, tout en soulignant des opportunités pour les moteurs Essence Turbo dans les gammes performantes. De plus, le Diesel, bien que minoritaire, conserve une place intéressante dans les catégories premium. Cet arbre nous a permis de comprendre clairement la segmentation du marché et d'orienter nos stratégies vers les segments les plus prometteurs.

5. Script R

L'analyse des données réalisée avec R a permis d'extraire des règles d'association pertinentes pour segmenter efficacement le marché automobile. En utilisant des paramètres tels que le support et la confiance, nous avons identifié des liens significatifs entre des caractéristiques clés des véhicules (consommation, taille, prix) et les segments de marché (économique, gamme moyenne, luxe).



Nous allons analyser plus précisément trois de ces règles pour identifier les marchés à fort potentiel.

1) Interprétation des règles

a) Règle 1

La première règle que nous avons pu déduire indique que les véhicules avec une consommation moyenne plus élevée (entre 15 et 24,5 litres pour 100 km) ont une très forte probabilité d'utiliser de l'essence comme carburant. Cela suggère que les voitures plus gourmandes en carburant sont généralement alimentées par de l'essence.

b) Règle 2

La deuxième règle montre que les véhicules avec une consommation modérée (entre 24,5 et 30,5 litres pour 100 km) qui fonctionnent à l'essence sont très souvent des modèles sans turbo.

c) Règle 3

La dernière règle identifie que les voitures les plus économes (avec une consommation élevée, entre 30,5 et 51,5 litres pour 100 km) et fonctionnant à l'essence sont systématiquement des modèles avec une aspiration standard. Cela reflète une relation forte et cohérente dans ces cas précis.

d) Conclusion

En conclusion les analyses révèlent trois marchés à fort potentiel. D'abord, le segment des véhicules à consommation moyenne élevée, majoritairement à essence, offre des opportunités pour des produits premium comme des carburants performants ou des assurances haut de gamme, ciblant une clientèle exigeante.

Ensuite, les véhicules à consommation moyenne avec aspiration standard se positionnent comme des choix polyvalents, idéaux pour des offres économiques sur l'entretien et les accessoires, visant des consommateurs recherchant le meilleur rapport qualité/prix.

Enfin, le segment des véhicules économe, avec des produits abordables.

Chaque segment présente des axes spécifiques pour maximiser les performances marketing. Ce qui correspond aux trois types de consommateur Ces trois segments reflètent les besoins distincts des consommateurs et présentent des opportunités marketing spécifiques. Ils s'alignent ainsi avec les profils des trois principaux types de consommateurs identifiés dans le marché automobile.