

PROJET DE DATAMING

DATAMINIG SUR EBAY- KLEINANZEIGEN

**Statistiques descriptives ,
clustering et prédiction**

RÉALISÉ PAR :

HANNACHI MAJDI



majdi_hannachi@outlook.fr

Sommaire:

- Contexte.....3**
- Énoncé du projet.....4**
- Source de données.....5**
- Environnement de travail.....7**
- RapidMiner.....8**
- Prétraitement des données.....11**
- Statistiques descriptives.....14**
- Etude statistique.....18**
- Extraction des connaissances à partir des données.....22**
- Comparaison et interprétation.....30**

Contexte:

Selon **Forbes**, le secteur de l'**automobile** a connu une croissance massive de 68% depuis son passage à un creux lors de la crise financière mondiale de 2009, selon un rapport publié par la société de **ventes aux enchères automobile** Manheim au début de cette année.

Le troisième trimestre de 2016 a clôturé avec 9,8 millions de véhicules vendus sur le marché des véhicules d'occasion, soit une augmentation de 3,3% par rapport à l'année précédente. En outre, le véhicule moyen vendu au détail s'est vendu 19,232 \$ au troisième trimestre de 2016, soit une augmentation de 4,3% par rapport à l'année dernière. Cette augmentation des valeurs globales est alimentée par le plus jeune âge (4,0 ans en moyenne) des véhicules vendus chez les concessionnaires franchisés.

Les changements de comportement d'achat de voitures neuves commencent à modifier le paysage des véhicules d'occasion franchisés.

Ainsi, les entreprises de véhicules d'occasion franchisées et d'autres grandes places de marché en ligne comme **E-Bay** tirent parti du taux de croissance de l'industrie des voitures d'usage. Nous

à l'occasion de faire un projet collective par binôme suggéré par notre enseignante à l'**Institut supérieur des études technologiques de Radès** Madame Hind eloui, on a donc essayé de comprendre ce marché et sa dynamique à l'aide du dataset 'Used Car Database' publié sur la plateforme de DataScience de **Kaggle.com** en exécutant des algorithmes de Datamining **descriptifs** et **prédictifs** en se mettant à la place d'un acheteur et un annonceur.

Enoncé du Projet:

1- Objectif:

Appliquer le processus d'Extraction des Connaissances à partir des données (ECD) sur un cas pratique en utilisant différents algorithmes de Data Mining.

2- Travail à faire:

- Choisir une base (assez volumineuse)
- Décrire la base (en faisant des recherches)
- Etude statistique:
 - Appliquer une méthode statistique pour l'analyse de données.
 - Visualiser les résultats.
 - Commenter, interpréter les résultats.
 - Tirer les conclusions
- Extraction des connaissances à partir des données:
 - Identifier le problème.
 - Préparer les données (prétraitements)
 - Utiliser plusieurs techniques de Data Mining pour extraire les connaissances et explorer des modèles.
 - Visualisation des résultats.
 - Comparaison et interprétation des résultats obtenus.
- Comparaison entre la méthode statistique et les techniques de Data Mining.
- Rédaction de rapport.



Source de données:

1- Fichier source:

Le dataset est un fichier csv intitulé '**autos**' contenant plus de 370 000 véhicules qui ont été vendus au enchères sur ebay de taille 65 MB.

Le dataset était publié le **28/11/2016** sur **kaggle** par le data analyste et le mathématicien chez **Statistik Beratung , Orges Leka**.
Profil: www.kaggle.com/orgesleka



2- Kaggle:



Kaggle est une plateforme web organisant des compétitions en science des données. L'entreprise a été créée en 2010 par Anthony Goldbloom. Sur cette plateforme, les entreprises proposent des problèmes en science des données et offrent un prix aux datalogistes obtenant les meilleures performances.

3- eBay:



eBay est une entreprise américaine de courtage en ligne, connue par son site web de ventes aux enchères du même nom. Elle a été créée en 1995 par le Français Pierre Omidyar. Parmi les articles qui se vendent on trouve les voitures d'occasion

4- description du fichier:

Le fichier source Autos.csv est un fichier de type CSV contenant un tableau de données concernant 371 824 lignes et 20 colonnes décrivant les caractéristiques d'une voiture ainsi que l'annonce dont on trouve sur ebay de l'Allemagne.

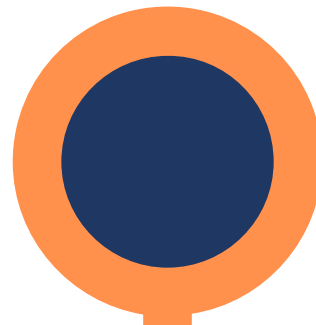
Ces attribues ce traduit selon ci-dessous:

- **dateCrawled** : la date de publication
- **name** : nom de la voiture (titre de l'annonce)
- **seller** : (privé ou concessionnaire)
- **offerType**: offre ou pétition
- **price** : prix de la voiture en euro
- **abtest**: test ou contrôle
- **vehicleType**: type de véhicule
- **yearOfRegistration** : l'année dont la voiture est construite
- **gearbox**: type de transmission
- **powerPS** : puissance de la voiture
- **model**: modèle de la voiture
- **kilometer** : kilométrés traversés par la voiture
- **monthOfRegistration** :le mois dans la voiture est construite
- **fuelType**: type de carburant
- **brand**: marque
- **notRepairedDamage** :si la voiture est endommagé
- **dateCreated** : date de création de l'annonce
- **nrOfPictures** : nombre d'images
- **postalCode**
- **lastSeenOnline** : dernière vu (la dernière vu est considéré comme date de vente de la voiture)

Environnement de travail:



Majdi Hannachi
junior Data-Scientist
Etudiant en M2BI à ISET RADES



LenovoTM

Ram 4 Go
Disque dur 1T
Processeur 6^{ème}
generation i5



rapidminer

RapidMiner:

1- Présentation:



RapidMiner est une **plate-forme logicielle** de science des données développée par la société du même nom, **payante**, qui fournit un environnement intégré pour la préparation de données, l'apprentissage automatique, l'apprentissage en profondeur, l'exploration de texte et l'analyse prédictive.

2- Services:

RapidMiner possède un outil de **préparation de données multifonctionnel** qui aide à faire tout type de préparation de données que ce soit nettoyage de données ou transformation de données, ainsi que des outils de visualisation de avec une grandes diversité de diagrammes.



RapidMiner propose une interface graphique conviviale avec un ensemble de fonctionnalités qui peut traiter et appliquer tout type **d'apprentissage automatiques supervisé et non supervisé** afin de **classifier** les données, **prédire** et même **détecter les valeurs aberrantes**.

RapidMiner présente aussi des outils pour la **validation** des algorithmes interprétés avec diverse type de validation tout dépend du type de l'algorithme, et offre aussi une interface de simulation afin de **déployer** le clustering ou la prediction en tant que utilisateur



Fonctionnalités de RapidMiner:

1-Principe:

L'interface de RapidMiner est basé sur le principe de Drag & Drop et la modélisation du démarche de dataming, ce principe est défini par Visual workflow designer avec lequel on fait le désigne de notre prototype.

2-Fonctionnalités:

Les différents fonctionnalités de RapidMiner se traduit dans ces six modules:



Data Access

Accès à tous type de données que ce soit des données enregistré sur des fichiers ou bien des données hébergées sur une base de données externe sql et NoSql.



Data Exploration

Des outils de découverte de données avec une bibliothèques riches de diagrammes pour l'exploration de données et pour assister la statistique descriptives



Data Blending

Multiples outils et fonctionnalités pour entrer les préparations nécessaires à fin de transformer les données et les organiser pour le déploiement des algorithmes de l'apprentissage automatique.



Data Cleansing

Ensemble de composantes de worlflow pour faire tout type de filtrage et de nettoyage des données à fin de les préparer pour les algorithmes.



Modeling

Faire la construction des modelés et les paramétrer selon les besoins du modèle choisi et faire la livraison des modèles plus rapide.



Validation

Valider le modèle avec les nombreux outils de validation disponibles pour voir la performance du modèle et les taux d'erreur produites.





Pourquoi RapidMiner?

1- Avantages:

- **Visual workflow designer** accélère le prototypage et la validation des modèles prédictifs, avec des connexions prédéfinies, des modèles intégrés et des flux de travail répétables.
- Bibliothèque riche de plus de **1500 algorithmes** et fonctions d'apprentissage automatique pour créer le modèle prédictif le plus puissant possible, quel que soit le cas d'utilisation
- Ouvert et extensible pour une intégration facile avec les applications existantes, les données et les langages de programmation tels que **Python et R**
- Désormais avec Auto Model et Turbo Prep: RapidMiner Auto Model utilise l'apprentissage automatique pour accélérer chaque étape du cycle de vie de la construction d'un modèle. Avec RapidMiner Turbo Prep, tout le monde, des analystes aux scientifiques de données, peut facilement transformer, faire pivoter et mélanger des données provenant de sources multiples en quelques clics.

2- Comparaison avec autres outils:

Overall Comparison

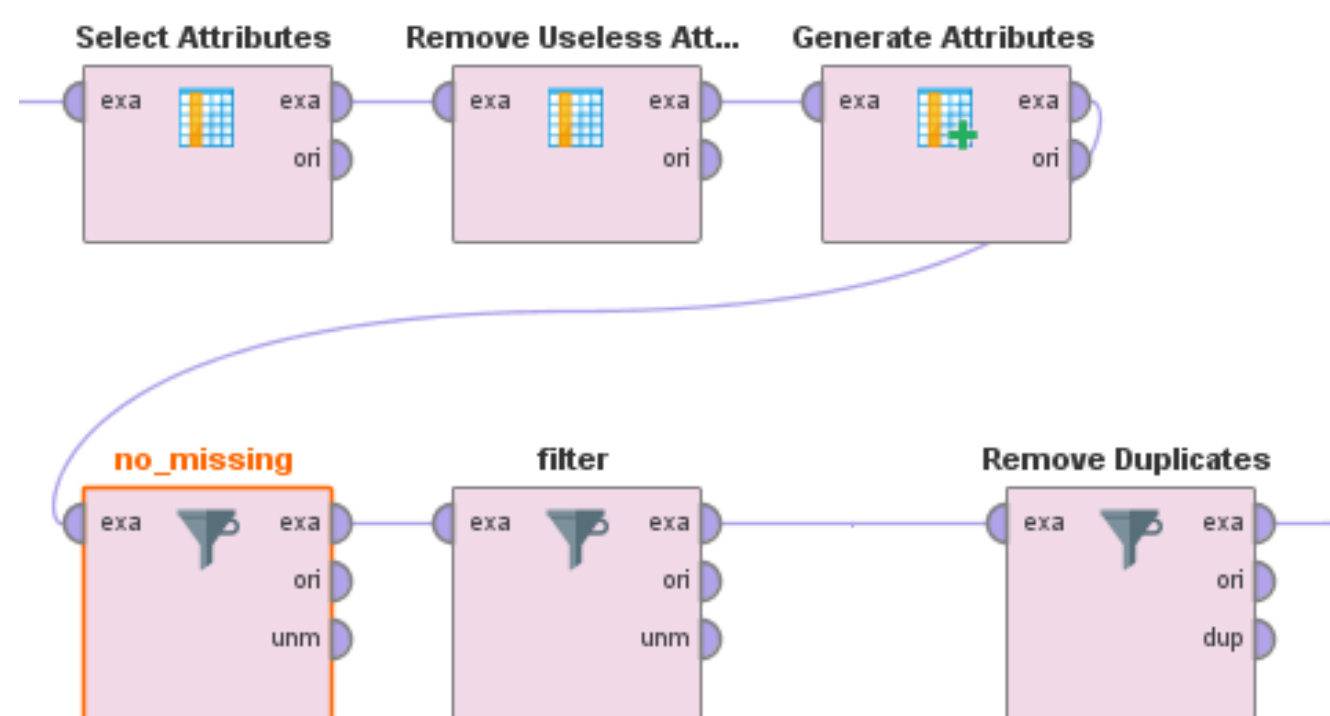
				
Procedure	R-Programming	RapidMiner	Weka	Orange
Partitioning of dataset into training and testing sets.	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)
Descriptor scaling	Pass	Pass	Fail (cannot save parameters for scaling to apply to future datasets)	Fail (no scaling methods)
Descriptor selection	Fail (no wrapper methods)	Pass	Pass (but is not part of KnowledgeFlow)	Fail (no wrapper methods)
Parameter optimization of machine learning/statistical methods	Fail (not automatic)	Pass	Fail (not automatic)	Fail (not automatic)
Model validation using cross-validation and/or independent validation set	Pass (but limited error measurement methods)	Pass	Pass (but cannot save model so have to rebuild model for every future dataset)	Pass (but cannot save model so have to rebuild model for every future dataset)

Prétraitement des données:

1-Principe:

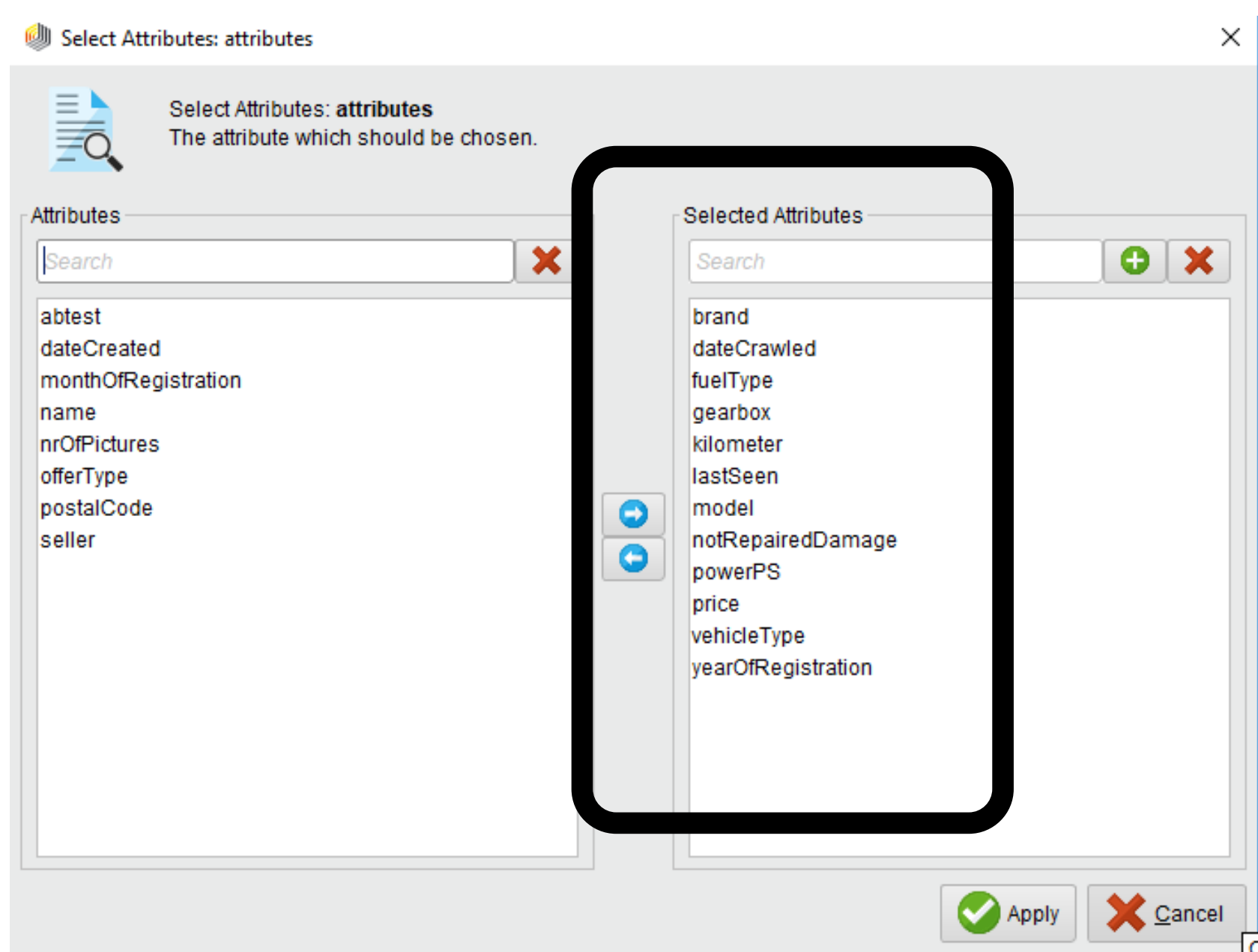
Le prétraitement concerne la mise en forme des données entrées selon leur type (numérique, symbolique, image, texte, son), ainsi que le nettoyage des données, le traitement des données manquantes, la sélection d'attributs ou la sélection d'instances. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillons d'apprentissage peut faire échouer l'opération.

2-Nettoyage des données:




Processus de nettoyage

Pour le dataset 'autos' , on doit faire la sélection des attributs pertinentes qui se coïncide avec l'analyse de la valeur du voiture ainsi que de l'annonce, Avec le composant '**Select Attributes**' on fait la sélection des attributs suivants:





On abandonne tout les attributs qui avec une stabilité plus de 95% avec le filtre **Remove useless Attributes** et on fait la génération des jours de ventes.




Edit Parameter List: **function descriptions**

List of functions to generate.









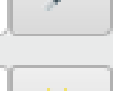

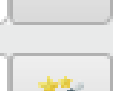

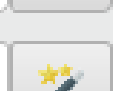

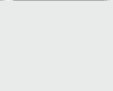

attribute name	function expressions	
<input type="text" value="sold_days"/>	<input type="text" value="ceil(date_diff(dateCrawled,lastSeen)/1000/60/60/24)"/>	
<input type="text" value="age"/>	<input type="text" value="2018-[yearOfRegistration]"/>	

Puis on se fait débarrasser des valeurs manquantes grâce au composent 'no_missing_values'.



Create Filters: **filters**


Defines the list of filters to apply.


<input type="text" value="yearOfRegistration"/>	<input type="text" value="≤"/>	<input type="text" value="2017"/>		
<input type="text" value="yearOfRegistration"/>	<input type="text" value="≥"/>	<input type="text" value="1990"/>		
<input type="text" value="kilometer"/>	<input type="text" value="≤"/>	<input type="text" value="300000"/>		
<input type="text" value="kilometer"/>	<input type="text" value="≥"/>	<input type="text" value="5000"/>		
<input type="text" value="price"/>	<input type="text" value="≤"/>	<input type="text" value="300000"/>		
<input type="text" value="price"/>	<input type="text" value="≥"/>	<input type="text" value="1000"/>		
<input type="text" value="powerPS"/>	<input type="text" value=">"/>	<input type="text" value="30"/>		
<input type="text" value="powerPS"/>	<input type="text" value="<"/>	<input type="text" value="500"/>		


☒ Match all

☐ Match any

☒ Preselect comparators

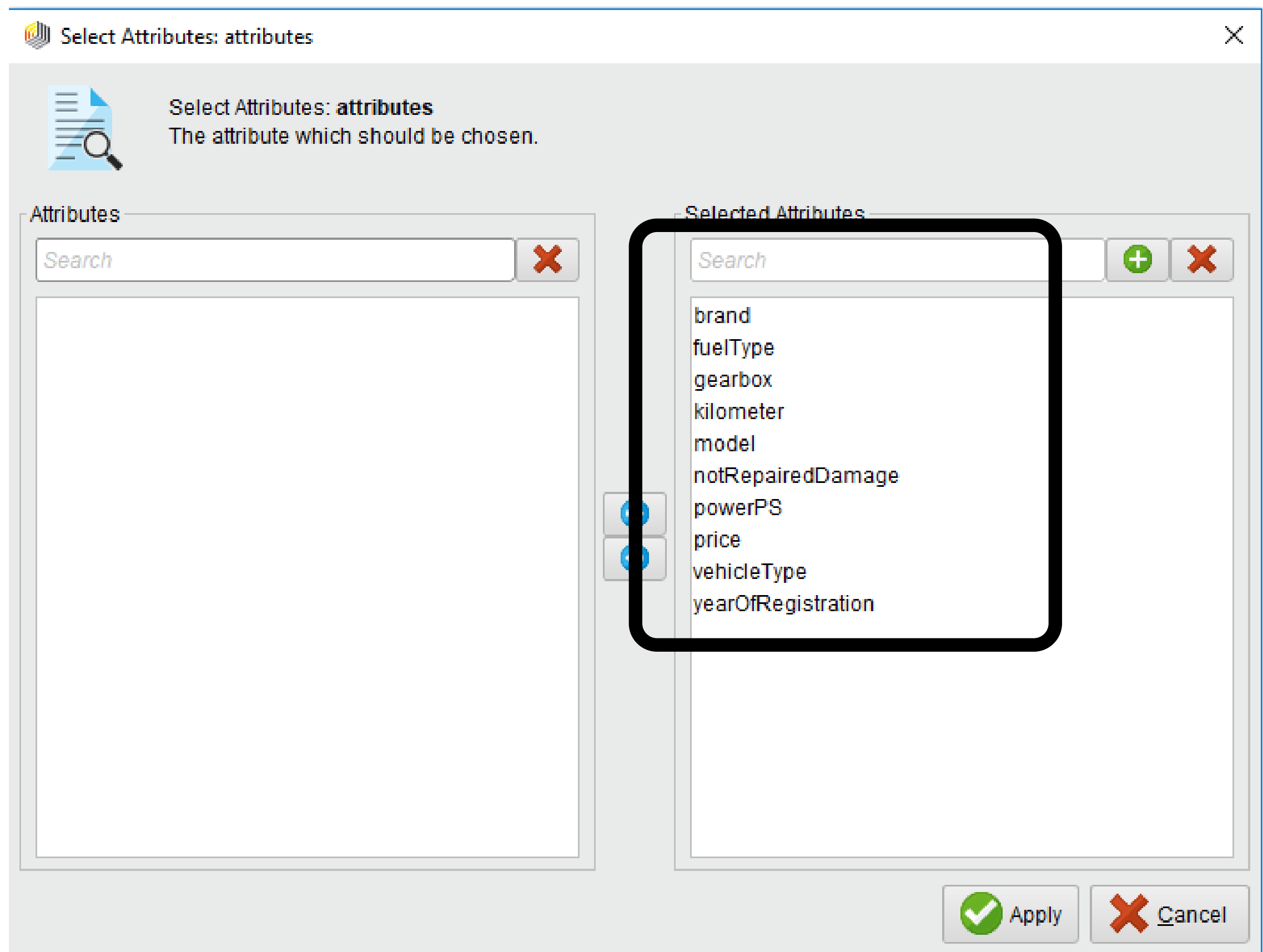
 Add Entry

 OK

 Cancel

Le filtrage ici est basé sur la logique d'un acheteur qui ne voulait pas une voiture très âgée , avec un nombre de kilomètres traversés qui est ne traduit pas une voiture neufs ou une voiture en état endommagé avec un prix raisonnable et une puissance acceptable.Ce processus nous aide à se débarrasser un valeur aberrantes.

Enfin , on termine le processus de nettoyage par la suppression des lignes dupliqués selon les attributs suivantes:



3- Résultat:

Le processus de prétraitement nous a aidé à se débarrasser de plus 180 000 valeurs manquantes, supprimer 8 colonnes qui n'ont aucun intérêt pour l'analyse.

Donc notre dataset s'est transformé en:

- 12 colonnes
- 218 699 ligne
- 8.7 Mb de taille

Statistiques descriptives:

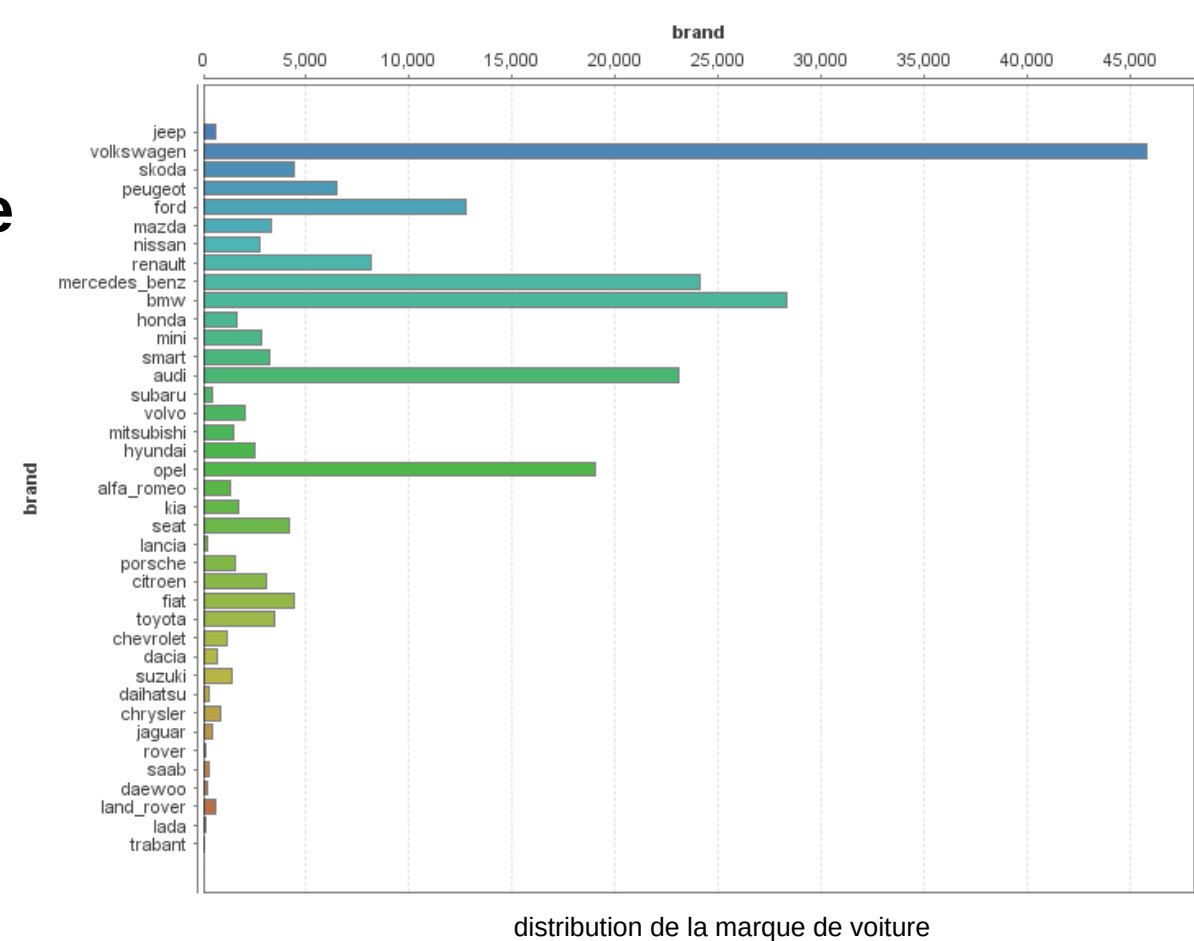
1- Objectif:

L'objectif de cette analyse descriptif des données et de mieux connaître les différentes aspects du dataset et de mieux comprendre les différentes tendances présentées par les variable pertinentes.

2- Analyse des variables:

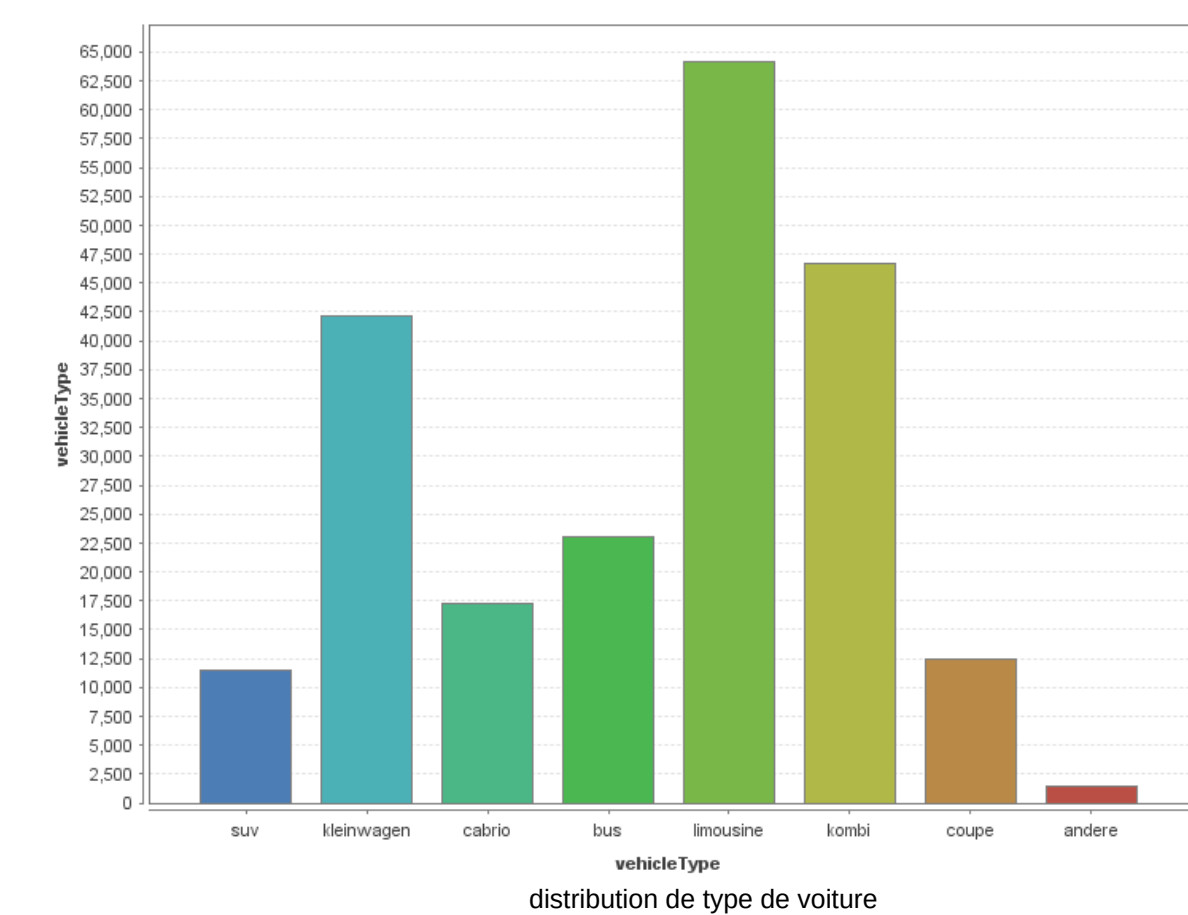
a- Brand:

Ce diagramme nous montre que la marque plus publié sur le site de ebay est **volkswagen** avec plus de **45 000** annonce puis on trouve **BMW** et **mercedes benz**. Dans ce data set on trouve 40 marques de voiture ceci est traduit par le nombre des modalités de l'attribut **Brand**.



distribution de la marque de voiture

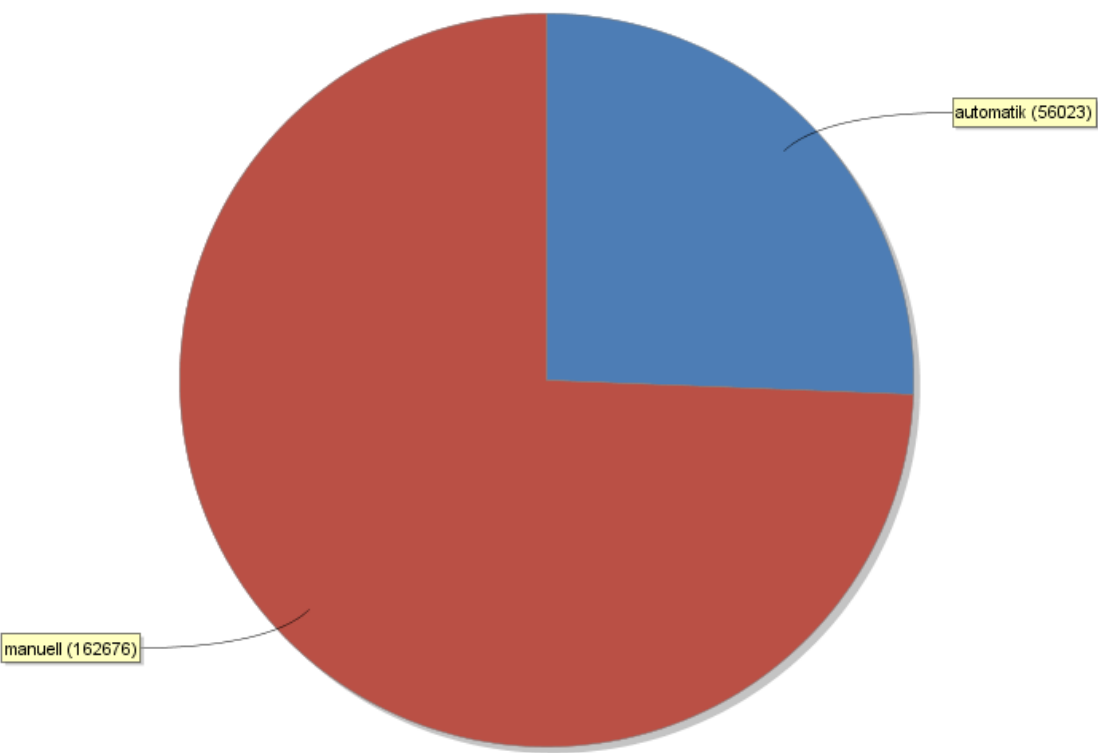
b- VehiculeType:



distribution de type de voiture

On trouve alors que les voiture **berline (limousine), break (kombi) et les petites voitures (keinwagen)** sont les types les plus publiés et les plus vendu sur ebay avec un nombre totale de plus de 150 000 publication.

c- GearBox:

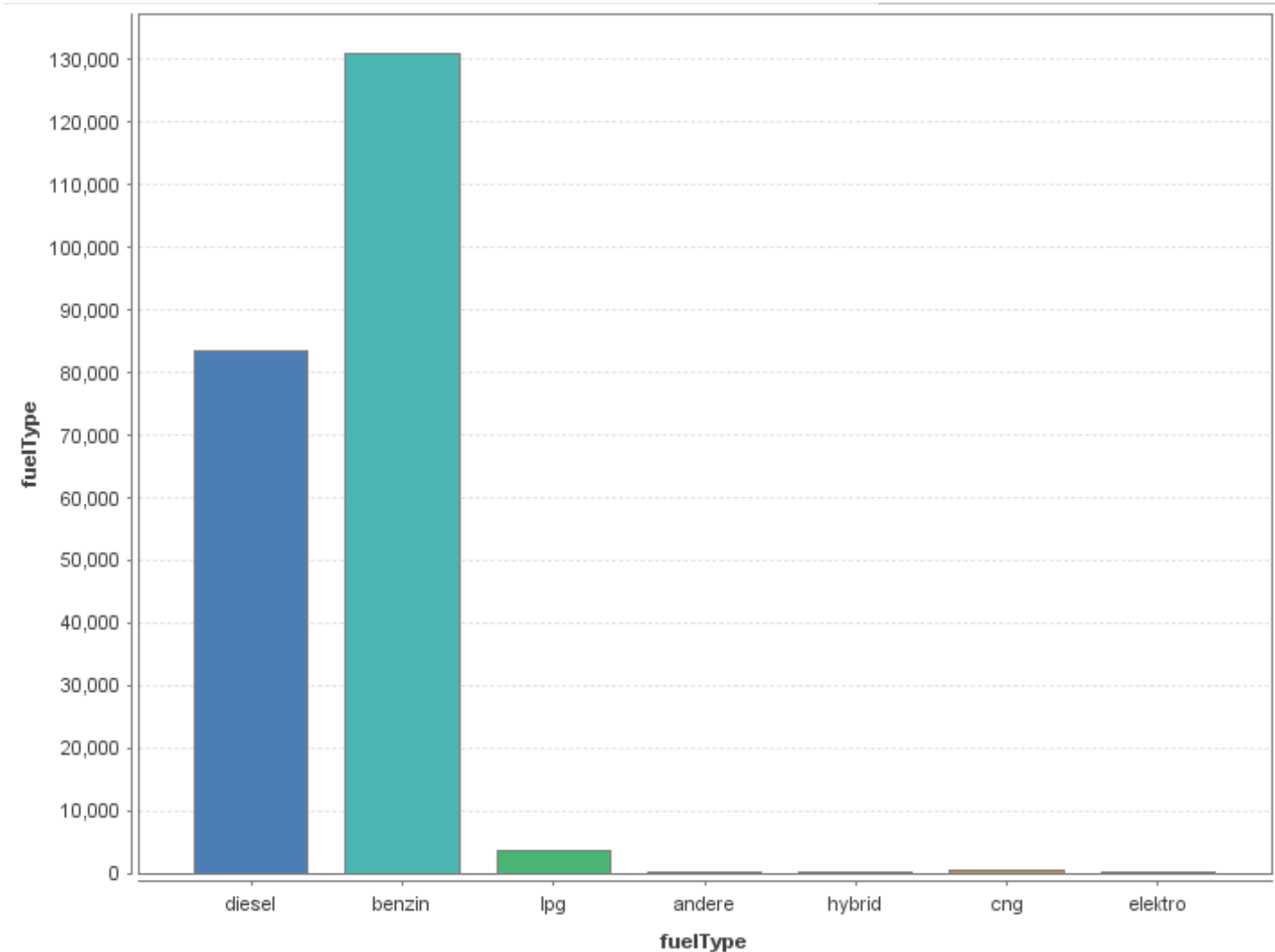


répartition du Gearbox sur la population

Le graphique fait clarifie que la majorité de population (3/4 des annonces) ont la transmission manuelle et le reste de transmission automatique.

c- FuelType:

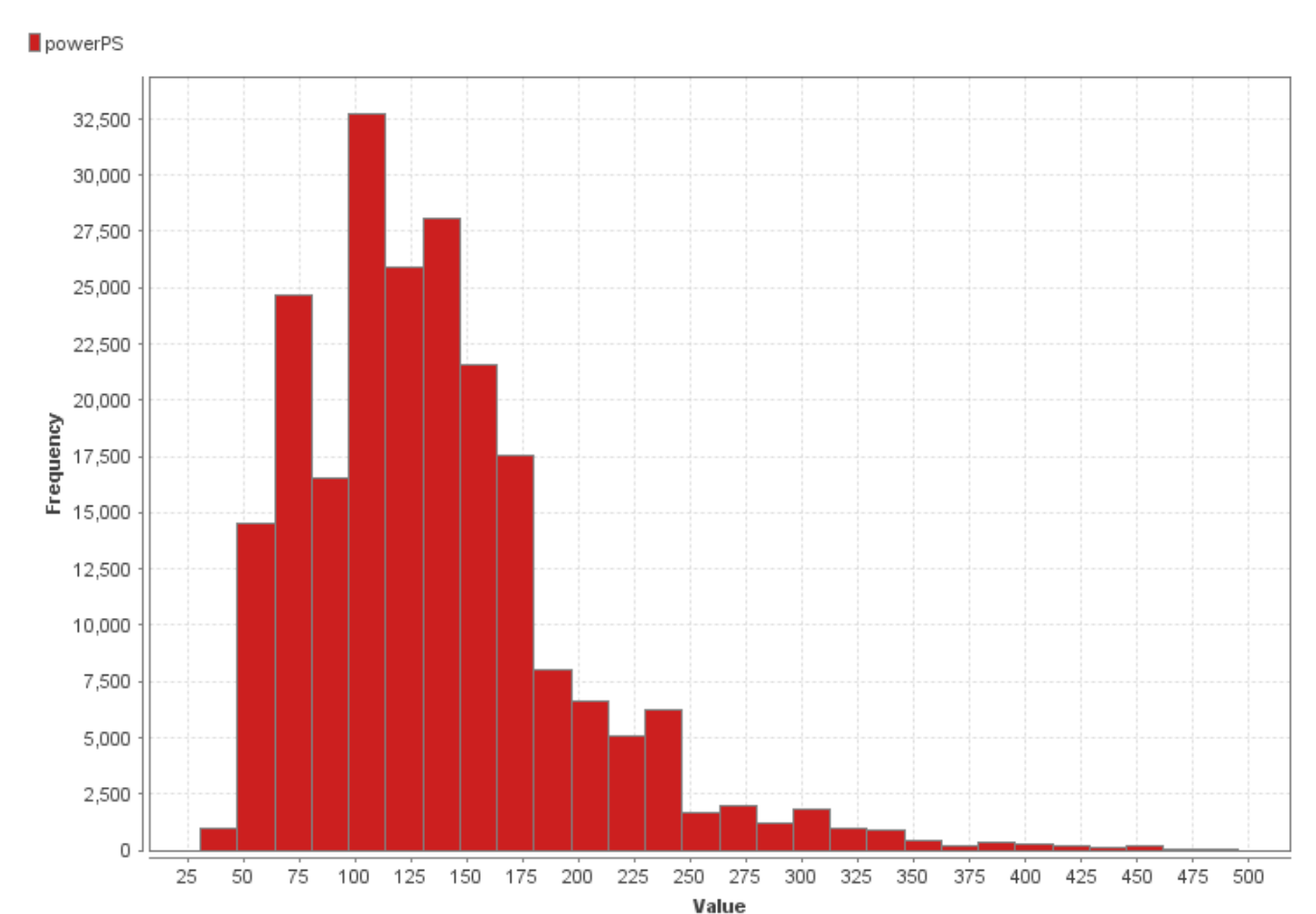
Le bar-chart nous présente que le type de carburant les utilisés sont l'essence (benzin) et le diesel sur une somme de plus de 200 000 lignes.



Distribution de type de carburant sur la population

d- PowerPs:

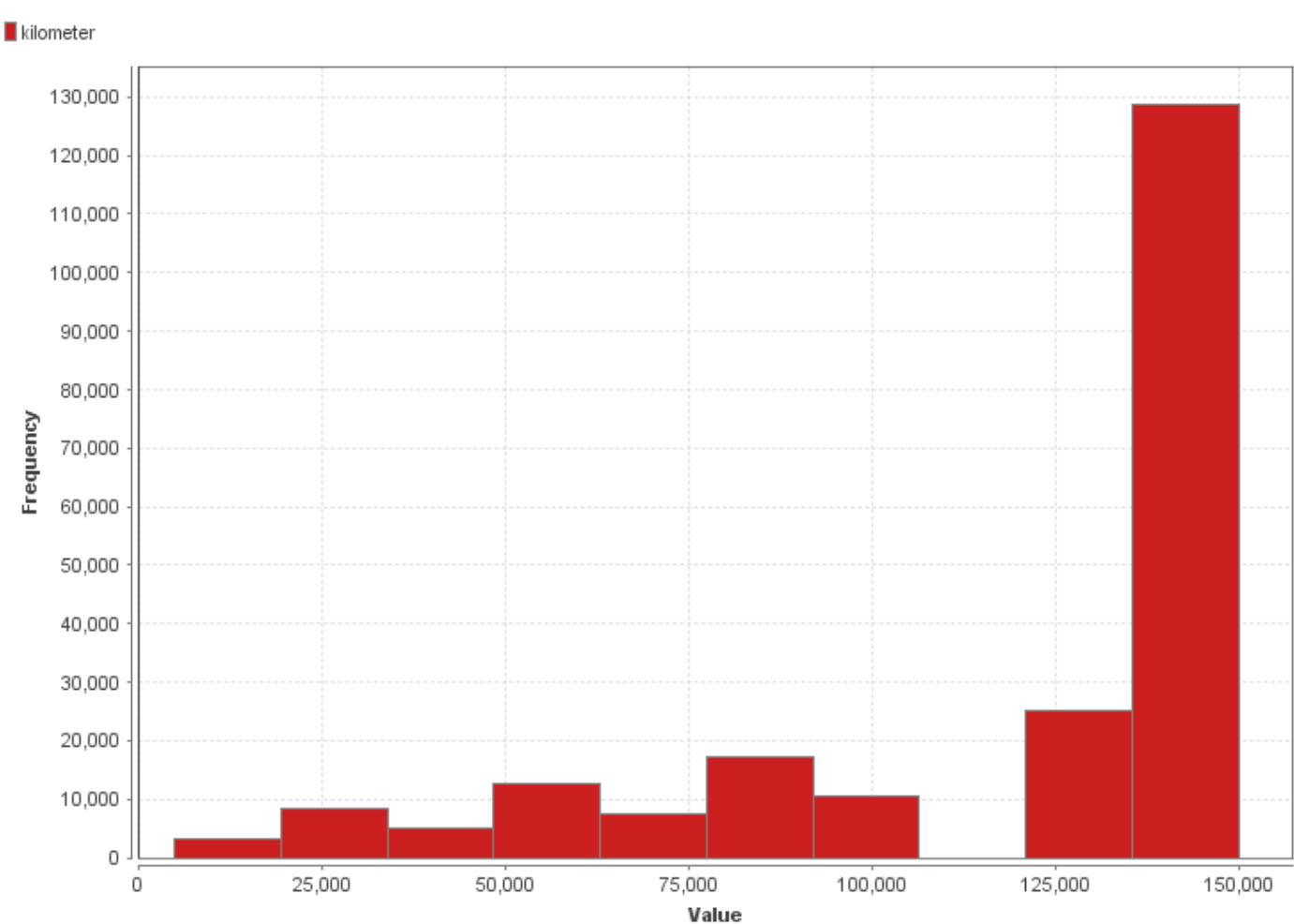
La moyenne et le médian de cette distribution est autour de 105 ainsi que la plupart de la population se focalise entre 30 et 250 chevaux vapeurs.



histogramme de répartition de puissance

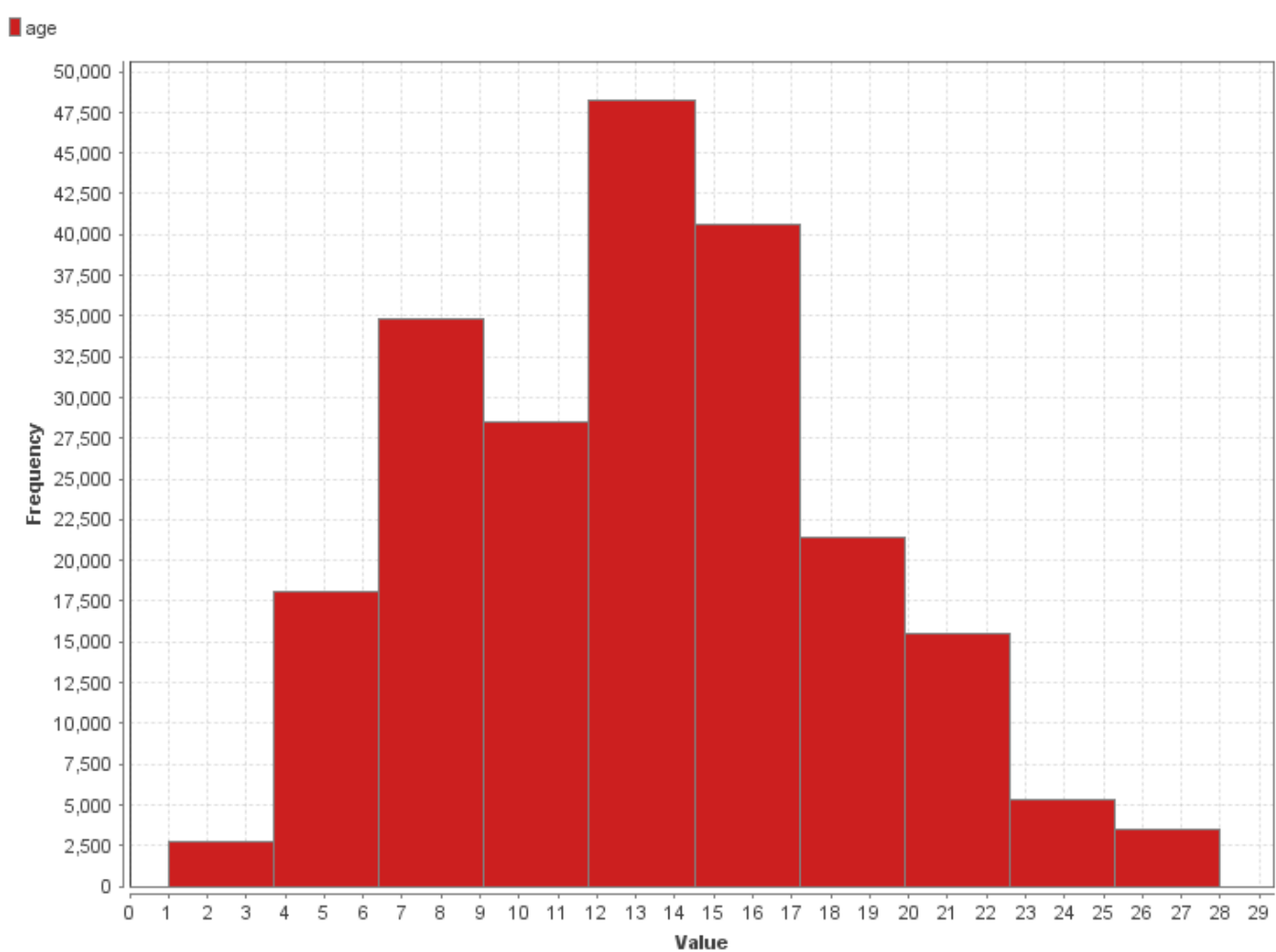
e- Kilometer:

On voit ici que la concentration des données et vocalisé sur plus de 100 000 km et plus particulièrement les voitures qui ont 150 000 de km dans sur le compteur.



Histogramme des kilométrés traversé par les voitures

e- age:

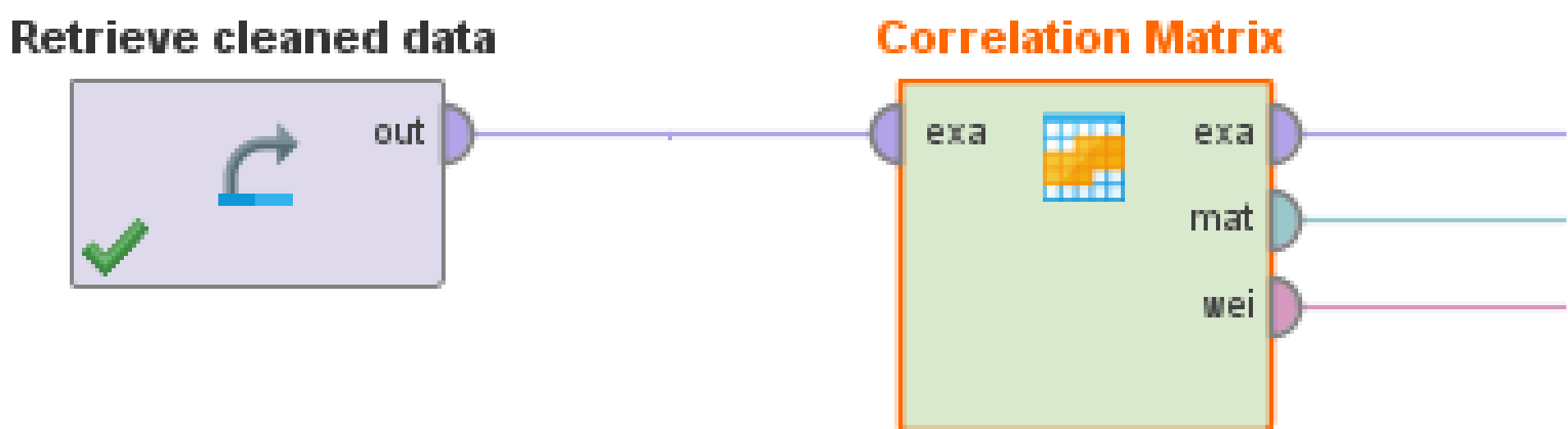


Distribution selon l'age

On trouve ici une distribution normal dont la moyenne et le médian sont autout de 14 ans.

2- Analyse de la corrélation:

Une matrice de corrélation est utilisée pour évaluer la dépendance entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres. Ceci se fait par le composant de RapidMiner Corrélation Matrix comme indiqué ci dessous:



workflow de visualisation de la matrice de corrélation

Le paramétrage se fait par la sélection des attributs à apparaître dans la matrice. Le résultat obtenu est le tableau suivant:

Attribut...	price	powerPS	kilometer	sold_da...	age
price	1	0.551	-0.447	0.115	-0.545
powerPS	0.551	1	0.036	0.058	-0.121
kilometer	-0.447	0.036	1	-0.084	0.537
sold_days	0.115	0.058	-0.084	1	-0.079
age	-0.545	-0.121	0.537	-0.079	1

matrice de corrélation

Interprétation:

- On a une faible corrélation entre la durée de vente et le prix. Pas de relation entre ses deux variables.
- Une corrélation moyenne négative entre le nombre de kilomètres parcourus et le prix. Plus le Kilométrage augmente plus le prix diminue.
- Une corrélation positive moyenne entre la puissance et le prix. Plus la puissance augmente le prix augmente.

Etude statistique:

1- Objectif:

L'objectif absolu d'une étude statistique est d'identifier un phénomène, le décrire et même parfois énumérer ses raisons. Pour ce cas, on veut caractériser le phénomène de vente au enchère des voitures et avoir une explication sur les différentes catégories de prix.

2- L'analyse:

Pour cela, on va faire une **Analyse en Composantes Principales** (ACP) sur notre ensemble de données.

a- Définition:

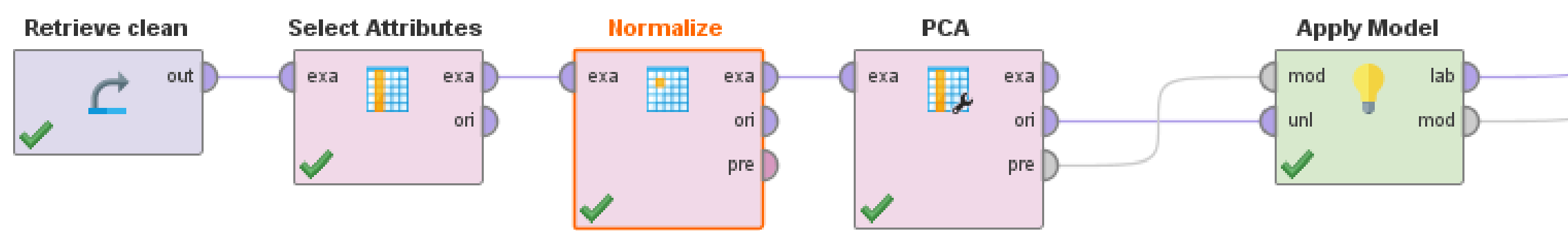
L'Analyse en Composantes Principales (ACP) est l'une des méthodes d'analyse de données multivariées les plus utilisées. Elle permet d'explorer des jeux de données multidimensionnels constitués de variables quantitatives.

b- Principe:

L'Analyse en Composantes Principales peut être considérée comme une méthode de projection qui permet de projeter les observations depuis l'espace à p dimensions des p variables vers un espace à k dimensions ($k < p$) tel qu'un maximum d'information soit conservée (l'information est ici mesurée au travers de la variance totale du nuage de points) sur les premières dimensions. Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du nuage de points, on pourra représenter les observations sur un graphique à 2 ou 3 dimensions, facilitant ainsi grandement l'interprétation.

c- Application et resultats:

Sur RapidMiner pour mener une ACP(PCA en anglais) il faut choisir le composant PCA :



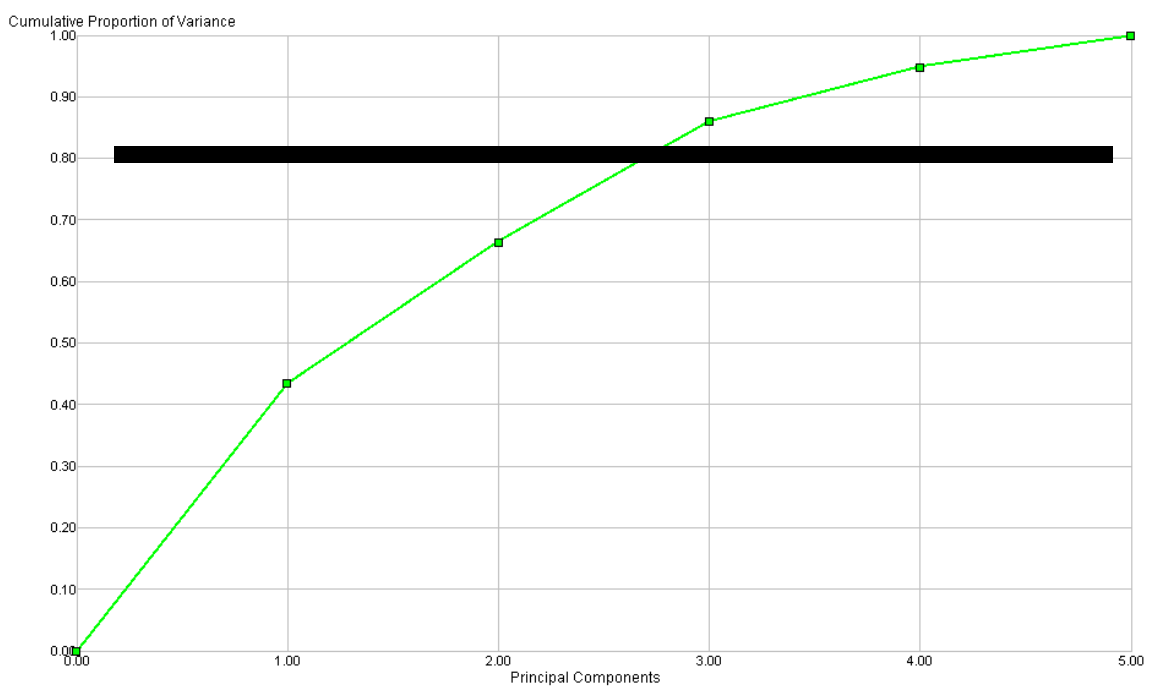
Workflow de ACP

L'exécution de ce workflow nous a montrer les resultat suivantes:

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.475	0.435	0.435
PC 2	1.073	0.230	0.665
PC 3	0.987	0.195	0.861
PC 4	0.666	0.089	0.949
PC 5	0.504	0.051	1.000

tableau de la variance cumulatives

- Le premier axe conserve 44% de l'inertie du nuage. Il est peu probable qu'il soit dû au hasard.
- Le second axe conserve une part de l'inertie totale de 22%.
- La chute est importante dès le troisième axe qui ne conserve plus que 22% de l'inertie totale.



test de coude: variance cumulative

- On peut décider de ne retenir que les trois premiers axes (le premier plan factoriel) car il compréhensible par l'œil (c'est un plan) et ne déforme pas trop le nuage (il explique 82% de l'inertie du nuage)

Maintenant on va analyser les principaux composantes de chaque vecteur à l'aide du table suivant:

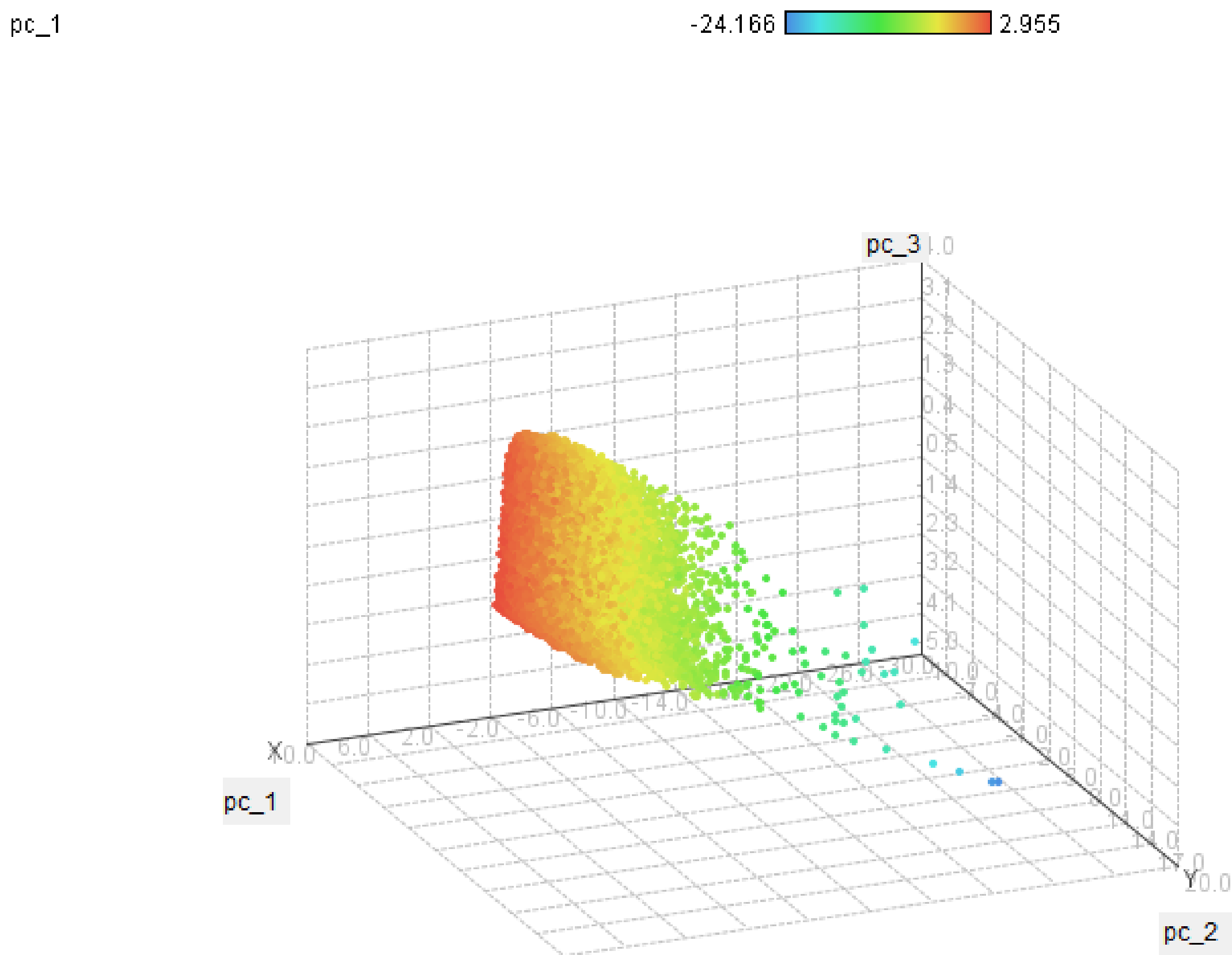
Attribute	PC 1	PC 2	PC 3
price	-0.598	0.242	-0.082
powerPS	-0.329	0.769	-0.083
kilometer	0.474	0.509	0.040
sold_days	-0.145	0.037	0.988
age	0.538	0.300	0.089

Eigenvectors

L'interprétation des nouvelles variables (des axes factoriel) se fera à l'aide des individus et variables contribuant le plus à l'axe avec la règle suivante : si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

- Pour PC1: l'age (0.538) et le prix (price , -0.598) contribuent avec l'axe; plus la valeur est élevée plus le prix diminue et plus l'age augmente.
- Pour PC2: la puissance (powerPS, 0.769) contribue avec l'axe; plus la valeur est élevée plus la puissance est élevée.
- Pour PC3: la durée de vente (sold_days , 0.988) contribue bien avec l'axe; valeur est élevée plus le nombre de jours est élevée.
- Pour la variable kilometer, on ne peut pas juger car les deux valeurs sur les deux premiers axes sont presque la même (0.474 , 0.509).

Pour présenter le resultat, on besoin d'un graphique de trois dimensions comme présenté ci-dessous:



Présentation du nuage des données

- Le nuage des points est présenté au milieu de l'axe d'ou les voitures sont de valeur autour de la moyenne (8010 euro) et l'age varie autour de 12 ans.
- La puissance des voitures est concentré dans la marge de l'origine 0 d'ou aussi il y'a une concentration des données autour du moyenne 105 avec une présence de quelques valeurs aberrantes présentées à gauche de l'axe avec les voitures puissantes.
- La concentration du nuage est présentée autour de -1 donc les voitures en générale se vendent vite dès la première semaine (moyenne = 9 jours).
- Les véhicules puissantes se vente mal puisque il sont chère et même interprétation pour les véhicules âgées puissantes.

Extraction des connaissances à partir des données

1- Présentation:

a- Datamining:

Forage de données, explorations de données ou fouilles de données, ce sont les traductions possibles du data mining en Français. En règle générale, le terme Data Mining désigne l'analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns. Ces informations peuvent ensuite être utilisées par les entreprises pour augmenter un chiffre d'affaires ou pour réduire des coûts. Elles peuvent également servir à mieux comprendre une clientèle afin d'établir de meilleures stratégies marketing...

b- ECD:

L'ECD est un processus complexe qui se déroule suivant une suite d'opérations. Des étapes de prétraitement ont lieu avant le « data mining » proprement dit. Le prétraitement porte sur l'accès aux données en vue de construire des « datamarts », des corpus de données spécifiques. Le prétraitement concerne la mise en forme des données entrées selon leur type (numérique, symbolique, image, texte, son), ainsi que le nettoyage des données, le traitement des données manquantes, la sélection d'attributs ou la sélection d'instances. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillons d'apprentissage peut faire échouer l'opération.

2- Problématique:

Le dataset utilisé ait une quantité de données vaste qui décrit le marché des enchères pour la ventes des voitures d'occasion. Comprendre le domaine et tous ses atouts est un facteur majeur pour tout type de participant direct ou indirect dans le marché.

Pour cela on va mener un processus de ECD dont on va appliquer diverse méthodes de machine learning et de dataming à fin de :

- Connaitre les différents segments du marchés.
- Classifier les voitures selon la valeur à fin d'optimiser l'achat.
- Prédire la durée de vente pour tout annonceur.

Alors on va utiliser des algorithmes de classification (k-means , règles d'association) et des algorithmes de prédiction (régression linéaire..)

3- Mise en oeuvre:

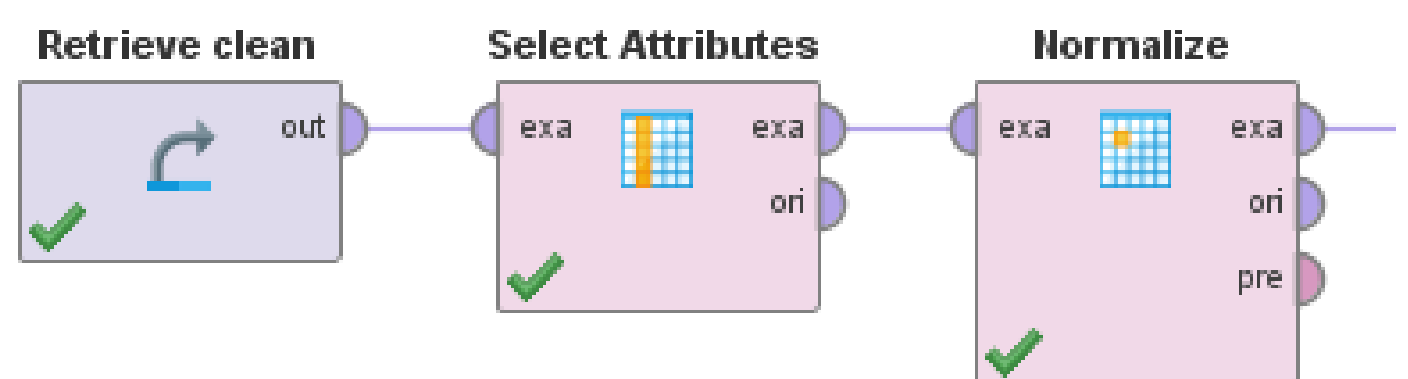
a- Clustering K-means:

La méthode des k-means est un outil de classification classique qui permet de répartir un ensemble de données en classes homogènes. Dans le cadre de la classification non supervisée, on cherche généralement à partitionner l'espace en classes concentrées et isolées les unes des autres. Dans cette optique, l'algorithme des k-means vise à minimiser la variance intra-classe.

Prétraitement:

Pour avoir une segmentation k-means fiable, on doit travailler avec des attributs numériques normalisés qui sont pertinente à l'analyse. Les attributs sont:

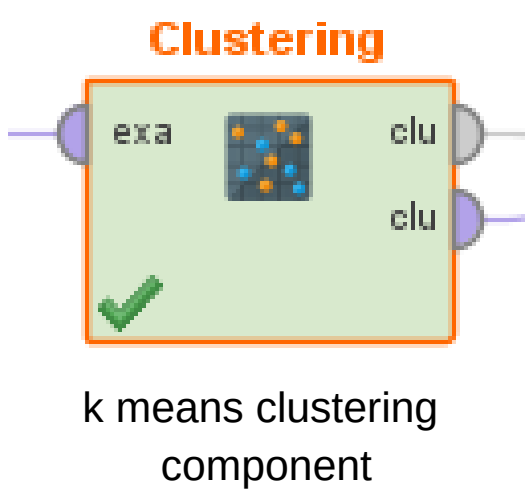
- l'age de la voiture
- le prix
- le kilométrage parcouru
- la puissance



Workflow de prétraitement pour le clustering k-means

Exécution:

Sur Rapidminer, la interprétation du modèle se fait grace au composent 'clustering' dont on paramètre la methode de calcule de distance (euclidienne) .



résultat:

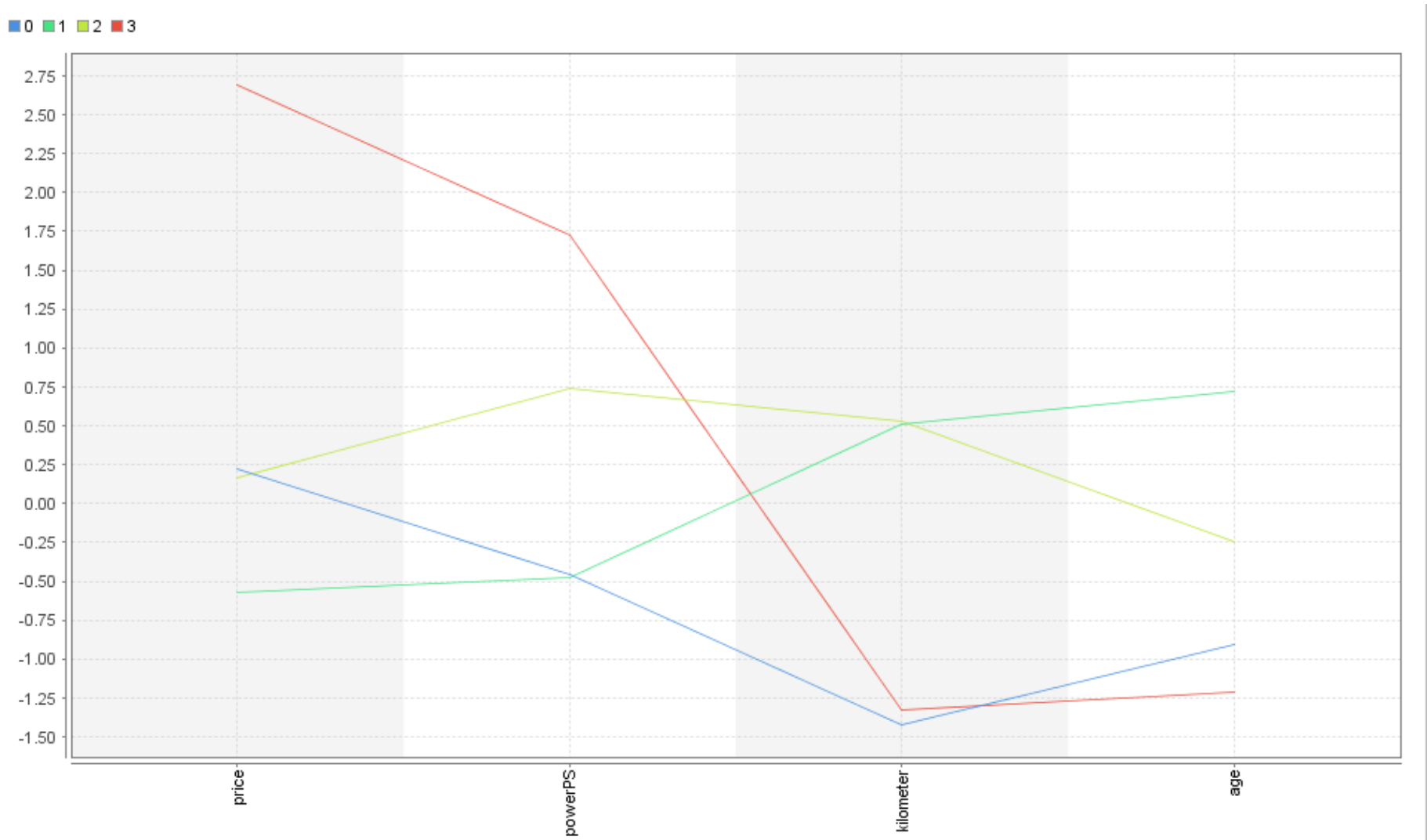
Cluster Model

```
Cluster 0: 45025 items
Cluster 1: 100058 items
Cluster 2: 59885 items
Cluster 3: 13731 items
Total number of items: 218699
```

résultat du clustering

- Cluster 0 : 45 025 instances
- Cluster 1 : 100 058 instances
- Cluster 2 : 59 885 instances
- Cluster 3 : 13 731 instances

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
price	0.225	-0.570	0.165	2.692
powerPS	-0.453	-0.478	0.743	1.725
kilometer	-1.423	0.507	0.526	-1.325
age	-0.905	0.721	-0.247	-1.208



- Cluster 0 : prix moyen , puissance faible , kilométrage faible, age faible : ceux sont des voiture 'daily' qui sont récemment achetées
- Cluster 1 : voitures âgées avec un kilométrage élevé de prix faible
- Cluster 2 : voitures puissantes avec un kilométrage élevé
- Cluster 3 : voiture puissante et nouvelles (sport et luxueuse)

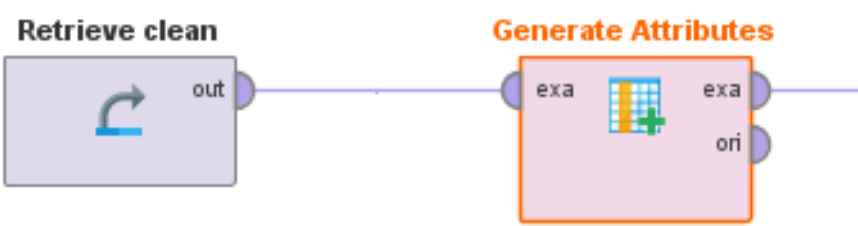
b- Règles d'association:

La recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données.

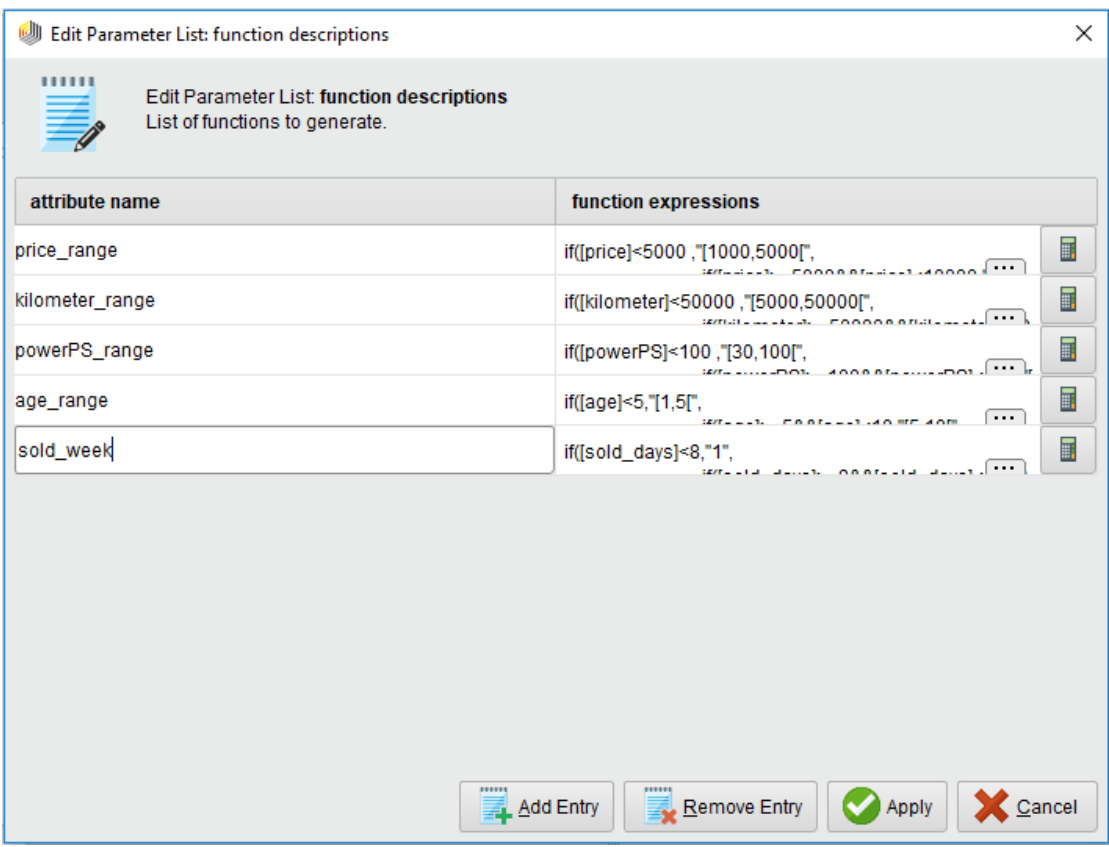
- $X \rightarrow Y$ où X et Y sont des ensembles d'items disjoints.
- Condition: Partie gauche de la règle.
- Conséquence: Partie droite de la règle.
- Support, fréquence: « Partie gauche et droite sont présentes ensemble dans la base ».
- Confiance: « Si partie gauche de la règle est vérifiée, probabilité que la partie droite de la règle soit vérifiée ».

Prétraitement:

Le prétraitement ici est de transformer tous les attributs numériques en attributs catégoriques.



workflow de prétraitement

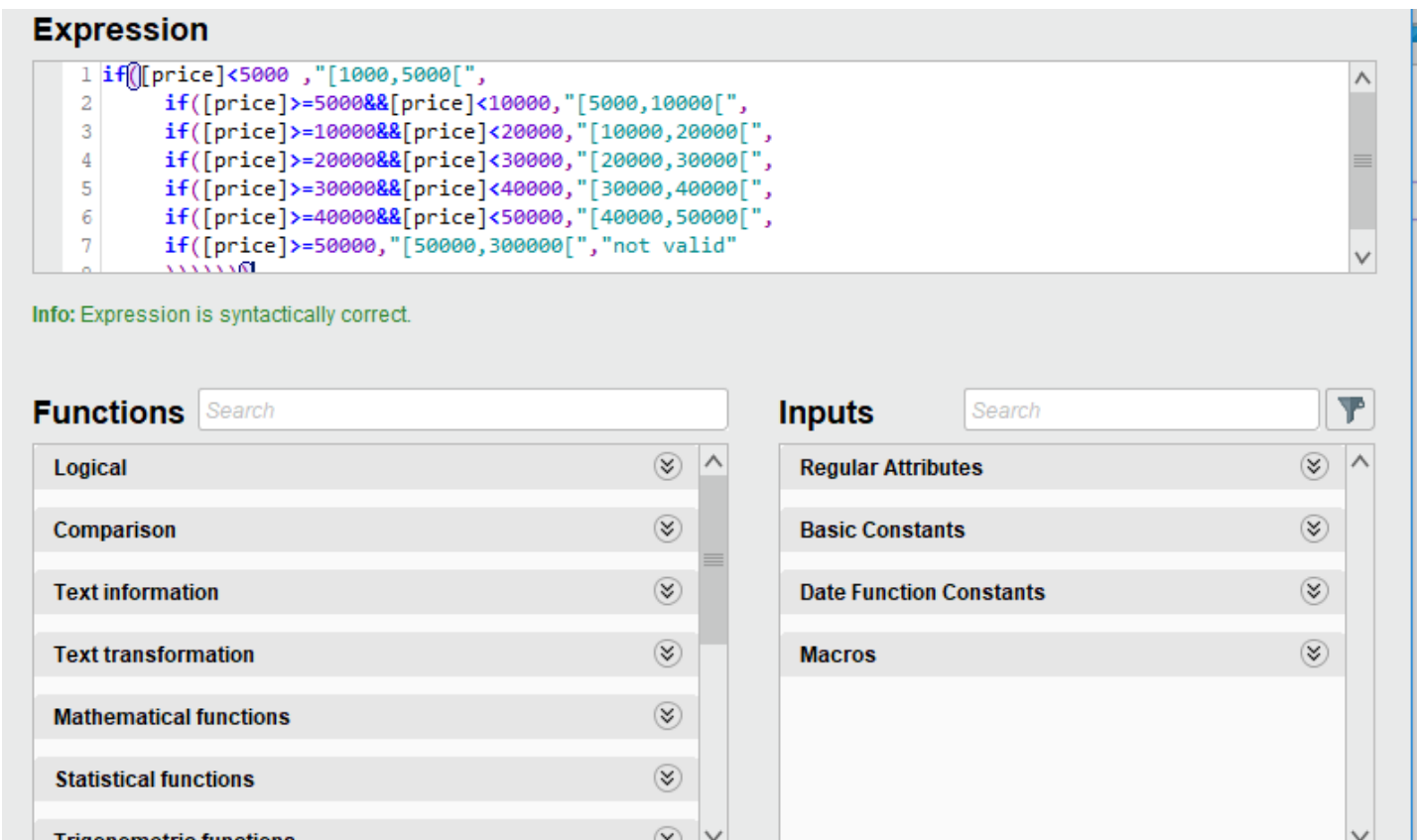


Paramétrage des attributs catégoriques

Les variables numérique qu'on doit transformer sont:

- le prix
- nombre des kilomètres parcourus
- puissance
- l'age
- durée de vente

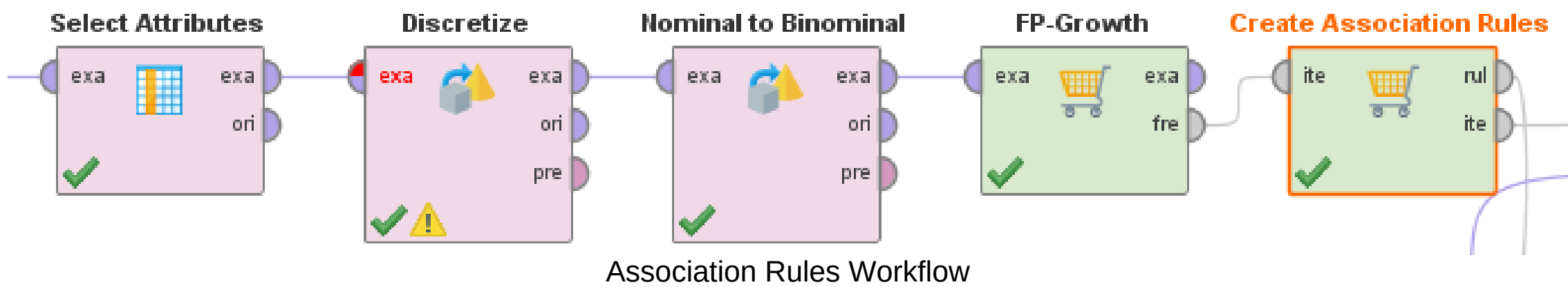
Le paramétrage des intervalles choisie se fait à l'aide de l'expression 'si alors' de RapidMiner



Paramétrage des intervalles de prix

Exécution:

L'exécution du modèle se fait grâce à une succession d'opérations qui vise à discrétiser les données et les transformer en binomiale afin de les passer sur l'interprétation des règles.



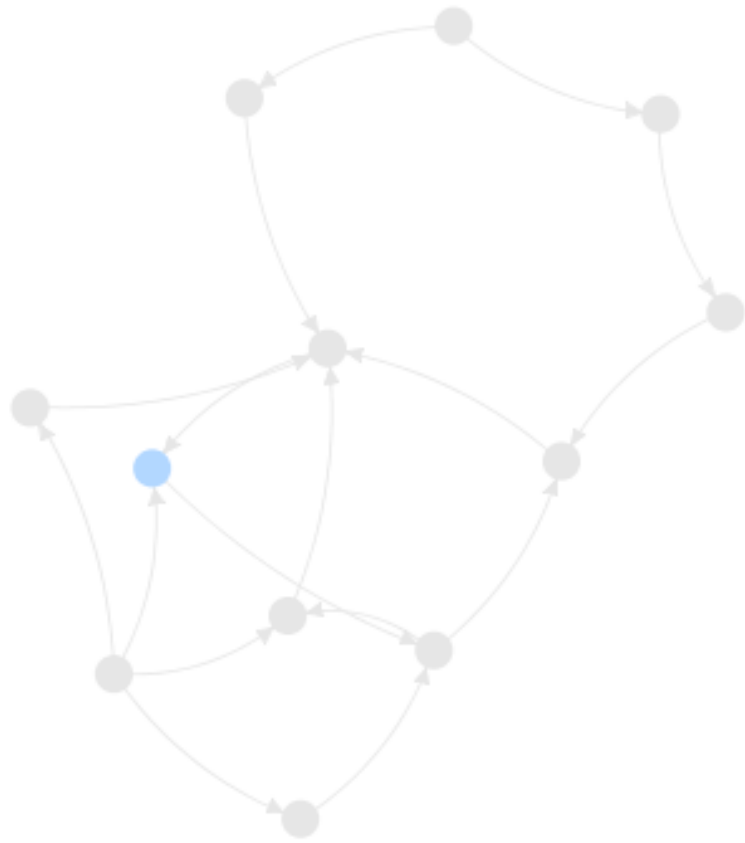
Parameters window for 'Create Association Rules':

- Criterion: confidence
- Min confidence: 0.75
- Gain theta: 2.0
- Laplace k: 1.0

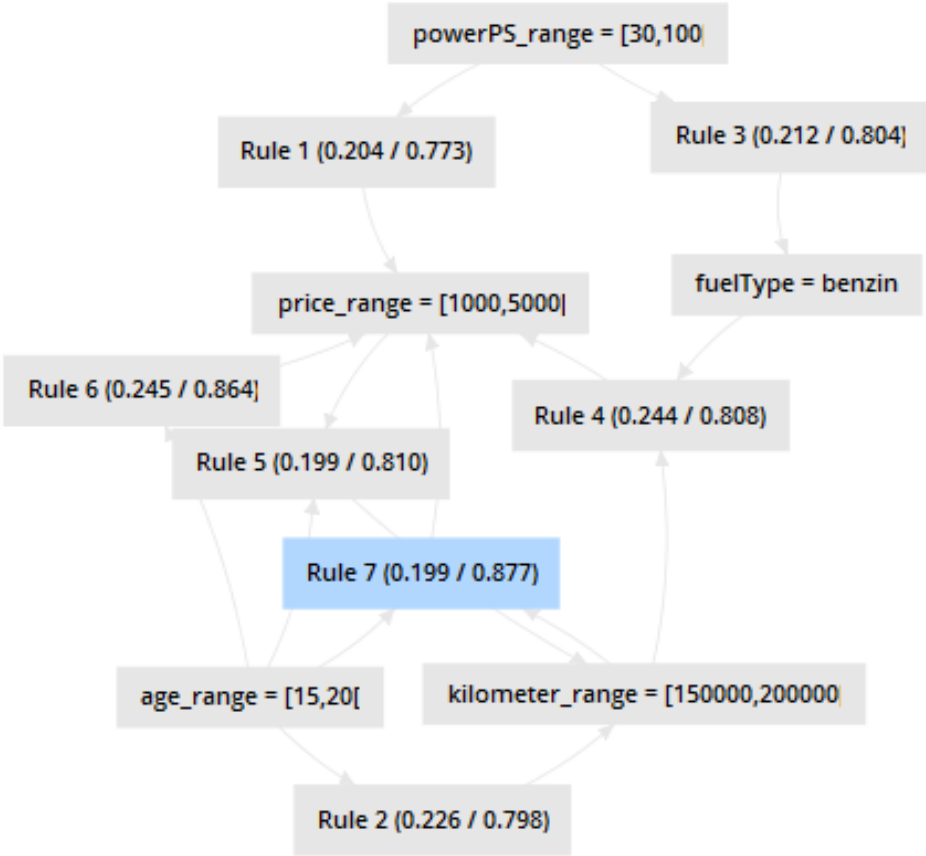
Association rules confidence

On a fait le paramétrage des règles qui vont être interprétées avec un minimum de confiance de 0.75. Travailler avec des niveaux de confiance élevés nous aide à avoir des associations logiques et pertinentes.

Résultat:



Nœuds des règles d'association



Nœuds des règles d'association avec label

Le résultat obtenu se définit par le tableau suivant:

No.	Premises	Conclusion	Confiden... ↓
7	kilometer_range = [150000,200000[, age_range = [15,20[price_range = [1000,5000[0.877
6	age_range = [15,20[price_range = [1000,5000[0.864
5	price_range = [1000,5000[, age_range = [15,20[kilometer_range = [150000,200000[0.810
4	fuelType = benzin, kilometer_range = [150000,200000[price_range = [1000,5000[0.808
3	powerPS_range = [30,100[fuelType = benzin	0.804
2	age_range = [15,20[kilometer_range = [150000,200000[0.798
1	powerPS_range = [30,100[price_range = [1000,5000[0.773

Règles d'association et les niveaux de confiance reliés

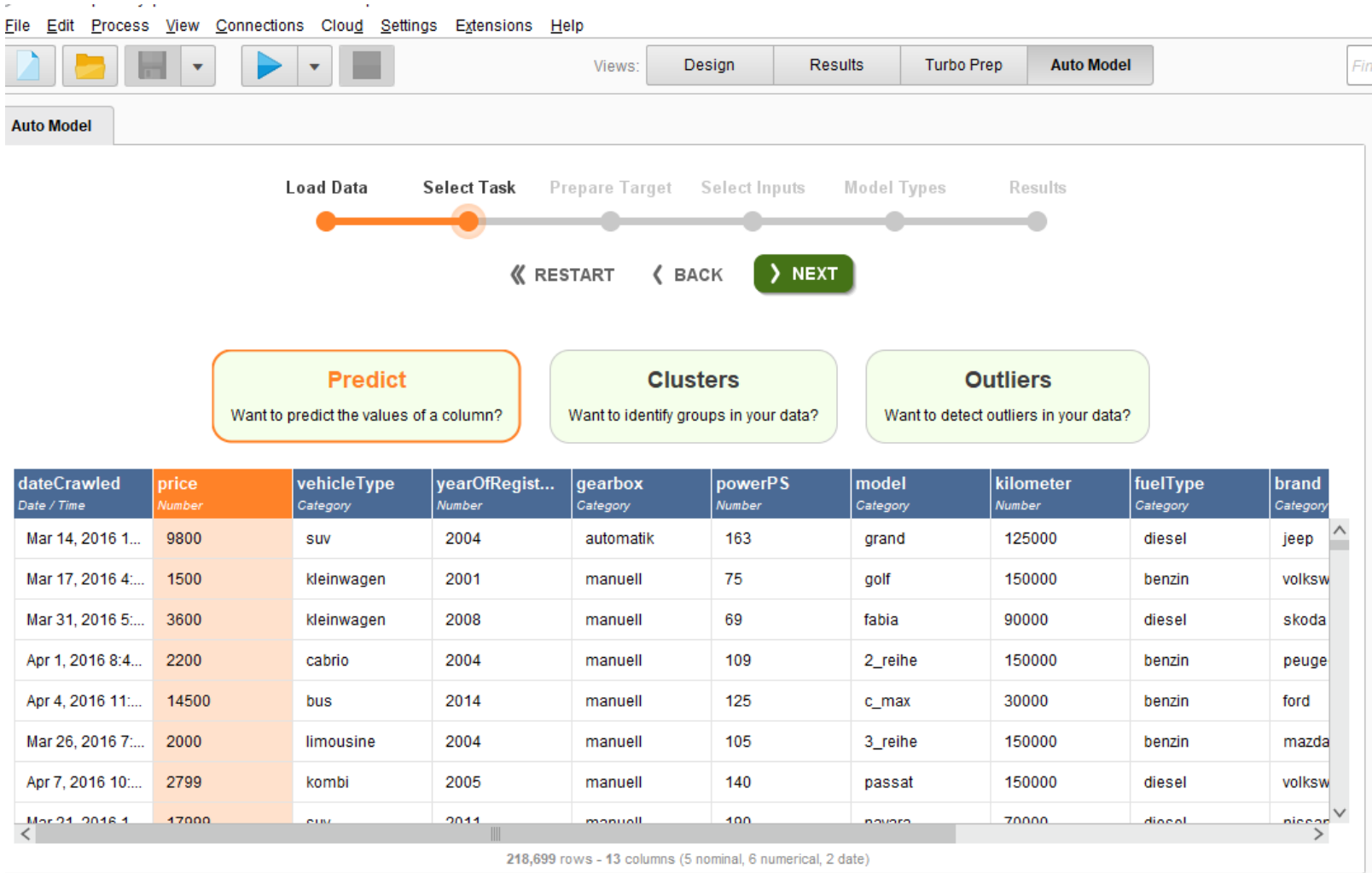
- le résultat obtenu est 7 règles d'association dont on interprète que:
- Si le kilométrage est plus de 150 000 et l'age plus 15 ans le prix est entre [1000,5000] euros.
 - Si l'age est plus de 15 ans alors le prix est entre [1000,5000[
 - Les règles 7 et 6 sont combinés dans la règle 5.
 - Si le carburant est Essence(benzin) et le kilométrage est élevé alors le prix est moins de 5000 euros.
 - Si l'age est élevé le kilométrage est élevé.
 - Si la puissance est moins de 100 chevaux alors elle est essence.

c- Régression linéaire:

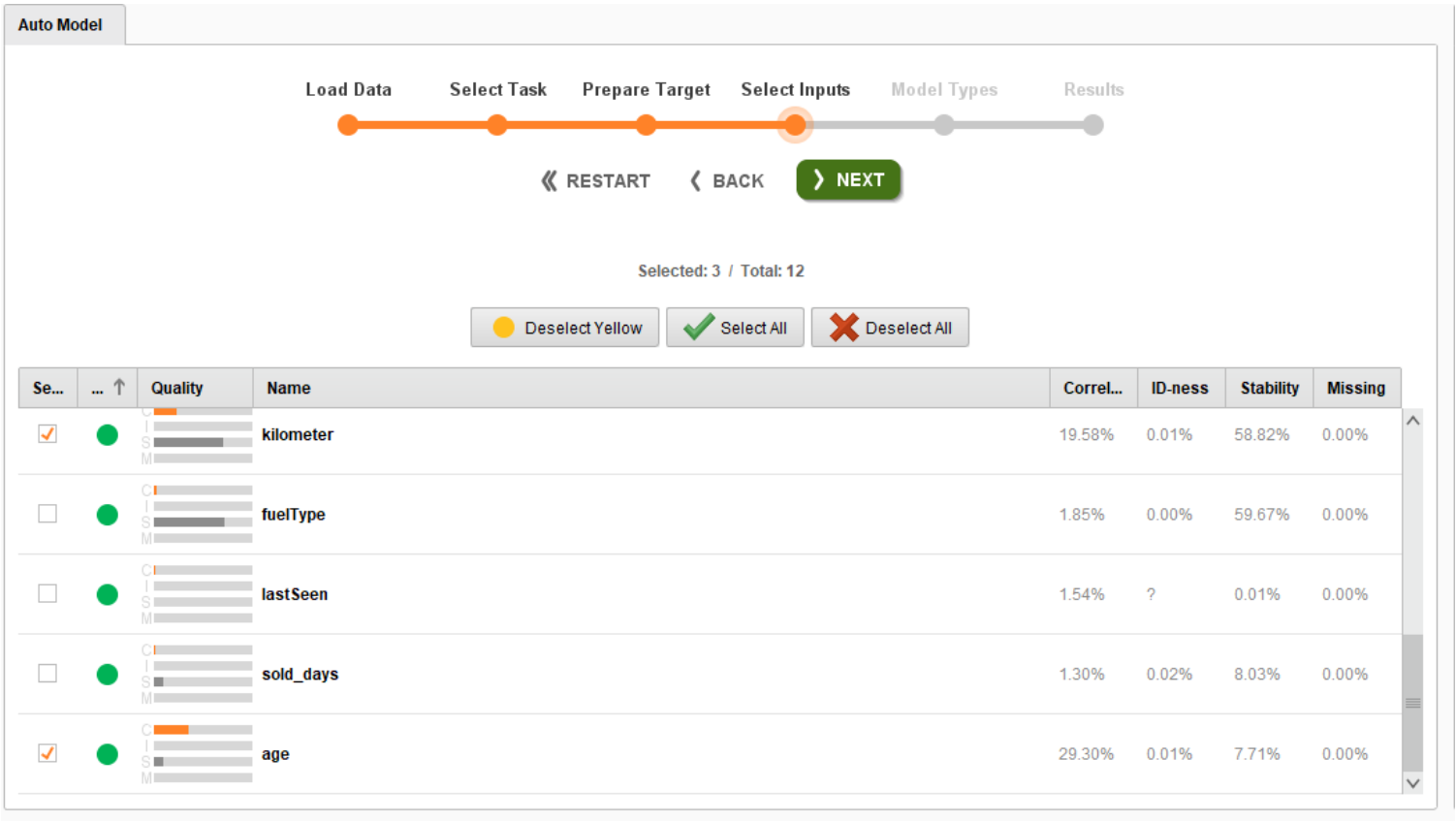
La régression linéaire est une modélisation linéaire qui permet d'établir des estimations dans le futur à partir d'informations provenant du passé. Dans ce modèle de régression linéaire, on a plusieurs variables dont une qui est une variable explicative et les autres qui sont des variables expliquées.

Exécution:

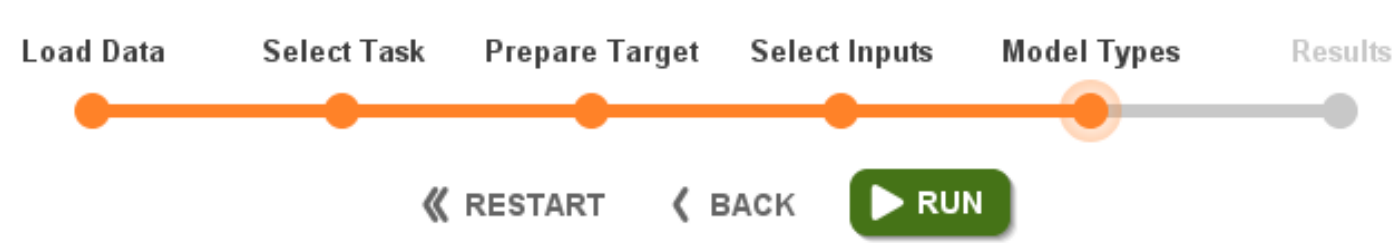
l'exécution de la régression linéaire pour RapidMiner est facile ; il se fit de naviguer vers l'onglet AutoModel et choisir 'prédicit" et choisir le champs à prédire :



Choisir la prédiction



Paramétrer les entrées



Models

mns ☒ Generalized Linear Model (GLM) - Warning: long computation time on this data!
☐ Use Regularization ☒ Calculate p-Values

Choisir le modèle de prédiction

Résultat:

Le résultat se traduit par le tableau des coefficient suivant :

Generalized Linear Model - Model

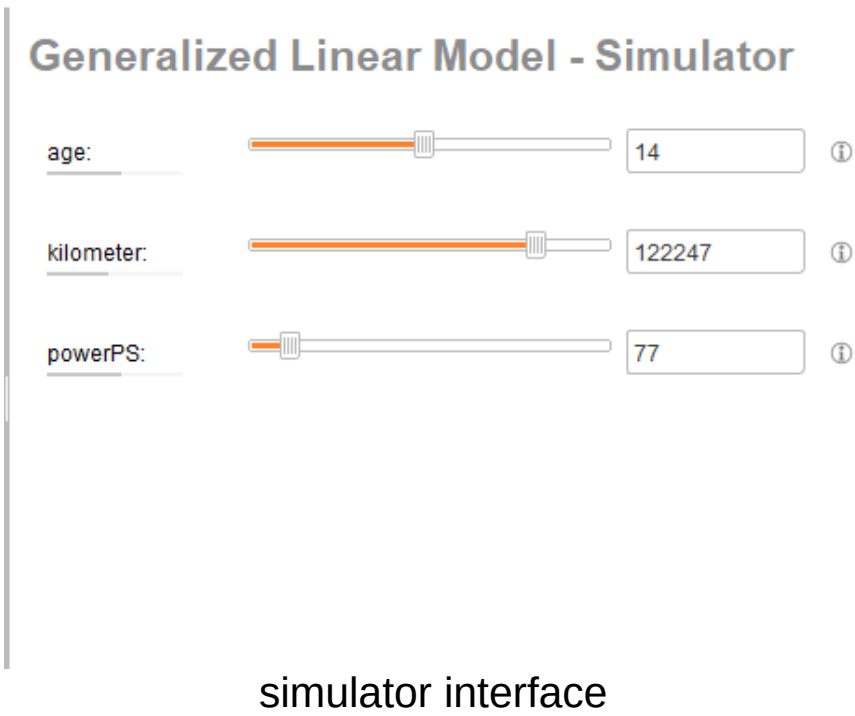
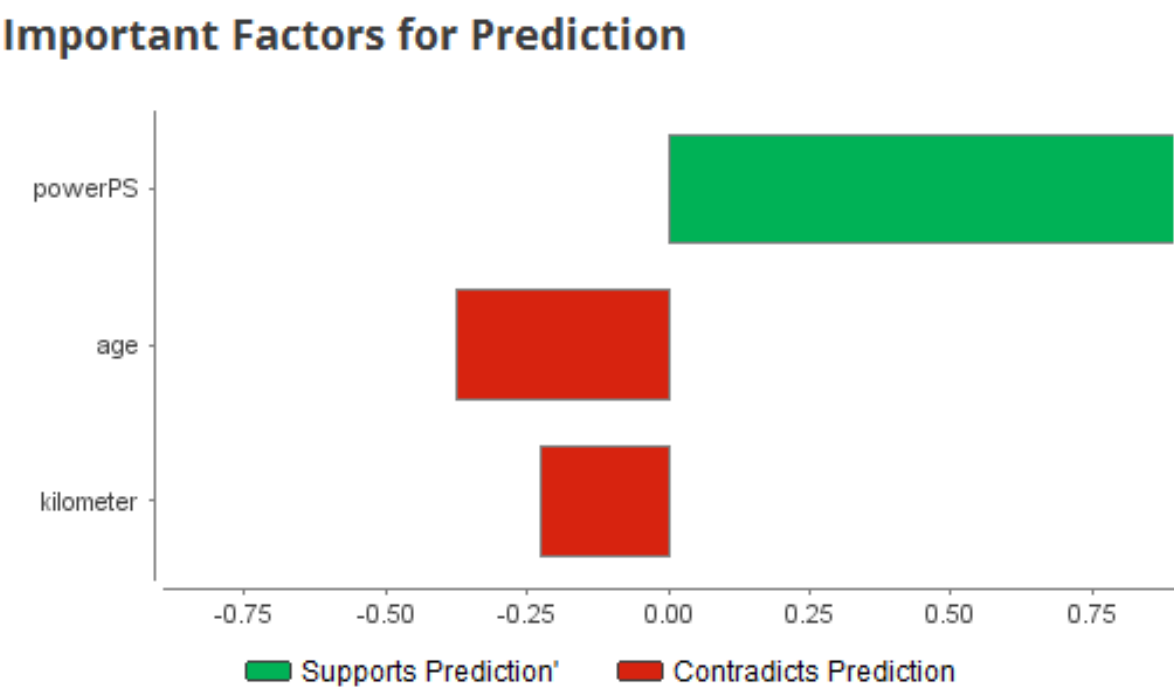
Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
age	-539.065	-2757.028	2.980	-180.920	0
kilometer	-0.060	-2425.680	0.000	-160.272	0
powerPS	72.767	4363.661	0.214	339.464	0
Intercept	12101.823	7536.582	20.334	595.143	0

tableau des coefficient

le tableau montre que tout les attributs choisis sont significatif(p-value = 0).
Donc on peut retenir les valeurs des coefficients pour la fonction linéaire suivante:

Prix= 12 101.823 + 72.767 Puissance - 0.060 kilometrage - 539.065 age

L'effet de la puissance et de l'age sont contradictoire sur la prédiction. Ceci est traduit par le graphique à droite.



l'un des avantages de l'auto Model est d'avoir un simulateur pour prédire des conditions souhaitées.

- L'age et le kilométrage augmente le prix diminue.
- la puissance augmente le prix augmente aussi.

Comparaison et interprétations:

La validation de chaque se fait différente de l'autre puisque on fait différentes type de datamining. Le résultat obtenu joue un rôle important pour la détermination de la validité de chaque model interprété mais pas par rapport aux autres. Les interprétations des différents modèles peut faire la comparaison entre ces différents algorithmes.

Critères	ACP	ECD		
		K-means	Association rules	Linear regression
Prétraitement	<ul style="list-style-type: none">Normalisation des valeursTrie des valeurs	<ul style="list-style-type: none">Normalisation des valeurs	<ul style="list-style-type: none">Transformation des données numériques en données catégoriques	-
Attributs analysés	Seulement des attributs numériques: Age , kilometer , price , powerPs , Sold_days	Seulement des attributs numériques: Age , kilometer , price , powerPs	Tout les attributs	Age , kilometer , price , powerPs
Validation	-	Distance intra-cluster = - 1,658	Niveau de confiance est plus de 0,75	Coefficient de corrélation = 0,526 Acceptable pour la prédiction
Temps d'exécution sur RapidMiner	1 s	3 min 49 s	15 s	35 s

L'analyse des composantes principales nous aider à représenté les données sur trois dimensions , le résultat obtenu de l'ACP n'a pas pratiquement expliqué la variation de prix. Le but de l'ACP de réduire la taille des données en faisant un fusion entre les différentes variables, ce qui a engendré trois dimensions qui n'ont pas bien expliqué la concentration de données.

Les Clustering k-means et les règles d'association nous a aider mieux segmenter et classifier les données, à avoir des groupes homogènes et bien identifier les différentes règles à les segmenter.

La régression linéaire nous a aidé à avoir une idée et clarifier de quoi dépend le prix d'une voiture.

Tout les modèles confirment que plus la voiture est âgée plus sa valeur diminue, plus le kilométrage augmente plus le prix diminue , plus la voiture est puissante plus elle est chère.

La limousine, la kombi et la kleinwagen sont les types de véhicules les plus populaires sur le marché de l'occasion. Les voitures les plus chères sont les SUV, tandis que les moins chères sont les kleinwagens.

En moyenne, le type de véhicule Kleinwagen est le moins cher et le moteur le moins puissant. Mais cela montre aussi les valeurs les plus aberrantes - peut-être en raison de la diversité des modèles de marques.

Les marques les plus populaires sont Volkswagen, BMW, Opel, Mercedes, Audi, Ford, Renault, Peugeot, Fiat et Seat. Ces 10 marques correspondent à près de 80% des voitures. (À l'origine, notre jeu de données contient environ 40 marques).

Selon notre analyse de régression, l'âge (39%), le kilomètre (23%) et la puissance du moteur (% 19) sont les facteurs les plus importants expliquant le prix de l'occasion.

Sur le marché de l'occasion, la plupart des voitures dépassent 100 000 km, voire 150 000 km. Les gens ne changent pas souvent de voiture selon notre ensemble de données.

La majorité des voitures d'occasion ne sont vendues que dans les 35 jours. Le ratio des 10 premiers jours (le jour 0 correspond à la vente le jour même) est assez élevé. Cela nous montre qu'Ebay-Kleinanzeigen réussit très bien à cibler les clients ou que le marché de l'occasion est plus fluide que nous le pensions.

Références:

- <https://rapidminer.com/>
- <https://www.kaggle.com/orgesleka/used-cars-database/home>
- [https://www.researchgate.net/publication/304265468_Comparaison_de_tech
hniques_de_Data_Mining_pour_l'adaptation_statistique_des_previsions_d'
ozone_du_modele_de_chimie-transport_MOCAGE](https://www.researchgate.net/publication/304265468_Comparaison_de_tech_niques_de_Data_Mining_pour_l'adaptation_statistique_des_previsions_d' ozone_du_modele_de_chimie-transport_MOCAGE)
- [https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-des-
correspondances-multiples-acm-ou-afcm](https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-des-correspondances-multiples-acm-ou-afcm)
- [https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-
information-th9/bases-de-donnees-42309210/extraction-de-connaissances-
a-partir-de-donnees-eed-h3744/](https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/bases-de-donnees-42309210/extraction-de-connaissances-a-partir-de-donnees-eed-h3744/)
- <https://www.lebigdata.fr/data-mining-definition-exemples>
- <http://iml.univ-mrs.fr/~reboul/ADD3-MAB.pdf>
- [https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-en-
composantes-principales-acp](https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-en-composantes-principales-acp)
- <https://explorable.com/fr/la-correlation-statistique>