ECOLE NATIONALE D'INGENIEUR DE CARTHAGE

DÉPARTEMENT GENIE ELECTRIQUE

Rapport de Projet

Analyse de données d'une base de données de voitures

DAGHSNI Saif Eddine

ANTAR Mohamed Aziz

Sous la direction de

Mrs. Imen Kammoun

Table des matières

Table des figures

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers Dieu Tout-Puissant, qui m'a donné la force, la patience et la volonté d'achever ce travail.

Je remercie chaleureusement mon encadrant universitaire, Mme.Imen Kammoun, pour son accompagnement, ses conseils avisés et sa disponibilité tout au long de ce projet. Son expertise et ses orientations ont été précieuses pour la réussite de ce travail.

Je tiens également à exprimer ma reconnaissance à l'ensemble du corps enseignant , pour la qualité de l'enseignement et la rigueur académique qui m'ont permis d'acquérir les connaissances nécessaires à la réalisation de ce projet.

Un grand merci à mes collègues et amis pour leur soutien moral, leur aide et leurs encouragements durant cette période.

Enfin, je dédie ce travail à ma famille, notamment à mes parents, pour leur amour inconditionnel, leur soutien constant et leurs prières, qui m'ont toujours accompagné dans mon parcours universitaire.

Merci à tous.

Introduction générale

Dans un contexte où l'analyse de données devient essentielle pour mieux comprendre des phénomènes complexes, ce rapport présente une étude approfondie basée sur plusieurs méthodes statistiques exploratoires.

L'objectif est d'analyser, de réduire et de structurer l'information afin de mieux interpréter les données disponibles.
Pour cela, nous avons d'abord eu recours à l'Analyse en Composantes Principales (ACP) afin de réduire la dimensionnalité des variables quantitatives tout en conservant l'essentiel de l'information.

Ensuite, l'Analyse des Correspondances Multiples (ACM) a été utilisée pour étudier les relations entre variables qualitatives.

En complément, deux méthodes de classification ont été mises en œuvre pour segmenter les individus : la Classification Ascendante Hiérarchique (CAH), qui permet de visualiser l'organisation des groupes sous forme d'arbre, et l'algorithme des k-moyennes (K-means), basé sur l'optimisation de la variabilité intra-groupes.

Ce rapport présente les résultats obtenus pour chacune de ces étapes, compare les méthodes de classification et discute la pertinence des regroupements identifiés.

Chapitre 1

Présentation du Dataset

Le jeu de données utilisé dans ce projet est constitué de **60 observations** représentant différents modèles de voitures. Chaque observation contient des informations techniques et économiques permettant d'analyser les performances et caractéristiques des véhicules.

1.1 Structure du dataset

Le dataset comprend **8 variables** décrivant chaque voiture :

- **Car Name** : Nom du modèle de la voiture.
- **Engine Power (HP)** : Puissance du moteur exprimée en chevaux.
- **Price (USD)** : Prix de la voiture en dollars américains.
- **Weight (kg)** : Poids total du véhicule en kilogrammes.
- **Top Speed (km/h)** : Vitesse maximale atteignable par la voiture.
- **0–100 km/h Acceleration (s)** : Temps nécessaire pour atteindre 100 km/h depuis l'arrêt.
- **Fuel Type** : Type de carburant (essence, diesel, etc.).
- **Transmission Type** : Type de transmission (manuelle ou automatique).

1.2 Aperçu du dataset

Un extrait des premières lignes du dataset est présenté ci-dessous :

	Engine Power (HP)	Price (USD)	Weight (kg)	Top Speed (km/h)	Accel. (s)
count	60	60	60	60	60
mean	223.46	33683.33	1526.91	204.93	7.71
std	113.79	18316.18	360.15	33.69	2.28
min	67	9000	920	150	3.1
25%	149.75	21000	1277.5	180	6.07
50%	187	29000	1585	197.5	8.0
75%	287.75	46250	1752.5	236.25	8.52
max	670	99000	2200	305	13.5

Table 1.1 – Matrice descriptive du dataset

1.3 Statistiques descriptives

L'analyse statistique du dataset permet de mieux comprendre la distribution des variables numériques (puissance, prix, poids, vitesse, accélération).

```
       Engine Power (HP)   Price (USD)  Weight (kg)  Top Speed (km/h)  \
count          60.000000     60.000000    60.000000         60.000000
mean          223.466667  33683.333333  1526.916667        204.933333
std           113.798136  18316.188696   360.154265         33.694976
min            67.000000   9000.000000   920.000000        150.000000
25%           149.750000  21000.000000  1277.500000        180.000000
50%           187.000000  29000.000000  1585.000000        197.500000
75%           287.750000  46250.000000  1752.500000        236.250000
max           670.000000  99000.000000  2200.000000        305.000000

       0-100 km/h Acceleration (s)
count                    60.000000
mean                      7.716667
std                       2.280363
min                       3.100000
25%                       6.075000
50%                       8.000000
75%                       8.525000
max                      13.500000
Dimensions : (60, 7)
```

Figure 1.1

Conclusion

Ce dataset est riche et varié, car il combine des informations **techniques** (puissance, poids, vitesse, accélération) et **économiques** (prix) ainsi que des caractéristiques **qualitatives** (type de carburant, type de transmission). Il est donc particulièrement adapté pour des analyses comparatives, statistiques ou des applications de machine learning liées à l'automobile.

Chapitre 2

ACP

Introduction

L'Analyse en Composantes Principales (ACP) est une méthode statistique puissante utilisée pour explorer, simplifier et visualiser des ensembles de données complexes. Dans ce chapitre, nous allons définir l'ACP, analyser les principales corrélations entre les variables étudiées, identifier les axes principaux extraits de notre jeu de données automobile, et interpréter les résultats obtenus. Cette approche nous permettra de mieux comprendre la structure des données et de mettre en évidence les relations essentielles entre les caractéristiques des véhicules.

2.1 Définition

L'ACP est une méthode statistique utilisée pour réduire la dimensionnalité d'un jeu de données tout en conservant un maximum d'information. Elle transforme les variables initiales corrélées en un nouveau jeu de variables indépendantes appelées **composantes principales**, permettant ainsi de visualiser les données plus facilement.

	valprop	inertie	inertiecum
0	3.966352	0.780049	78.004919
1	0.737475	0.145037	92.508597
2	0.180731	0.035544	96.062972
3	0.161464	0.031755	99.238434
4	0.038724	0.007616	100.000000

FIGURE 2.1 – Les résultats du calcul ACP

On remarque que les deux premières composantes expliquent par exemple **92,5%** de la variance totale, ce qui signifie qu'on peut représenter les données initiales dans un espace à deux dimensions sans trop de perte d'information.

2.2 Corrélations observées

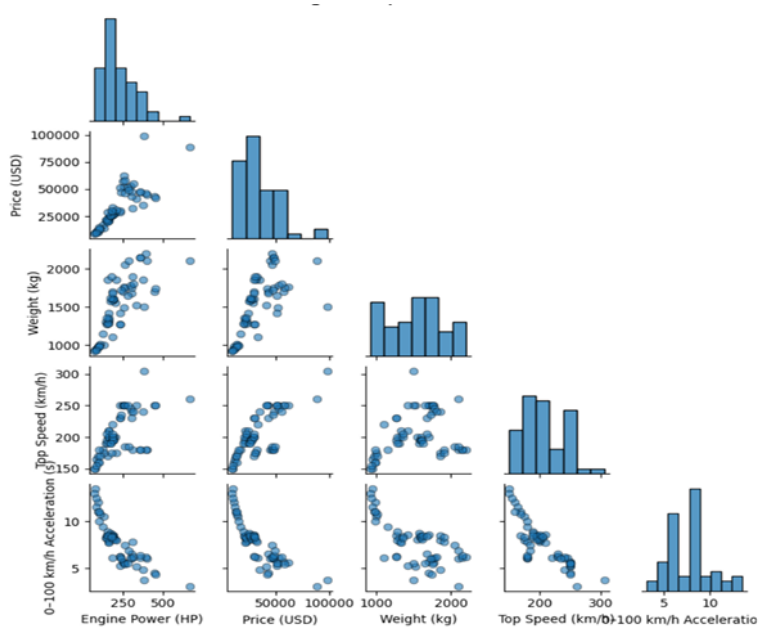

FIGURE 2.2 – Matrice des nuages de points entre les variables

- **Engine Power ↔ Price** : Corrélation positive
  Plus la puissance est élevée, plus le prix est élevé.
- **Engine Power ↔ Weight** : Corrélation positive légère
  Les moteurs plus puissants sont souvent dans des voitures plus lourdes.
- **Engine Power ↔ Top Speed** : Corrélation positive forte
  Les moteurs plus puissants permettent une vitesse maximale plus élevée.
- **Engine Power ↔ Acceleration (0–100 km/h)** : Corrélation négative forte
  Plus la puissance est grande, plus l'accélération est rapide (temps plus bas).
- **Price ↔ Weight** : Corrélation faible
  Le prix dépend peu du poids directement.
- **Price ↔ Top Speed** : Corrélation positive
  Les voitures plus chères vont en général plus vite.
- **Weight ↔ Top Speed** : Corrélation négative légère
  Plus une voiture est lourde, plus sa vitesse maximale tend à diminuer.
- **Price ↔ Acceleration (0–100 km/h)** : Corrélation positive
  Plus une voiture est lourde, plus elle met de temps à atteindre 100 km/h.

2.3 Axes principaux

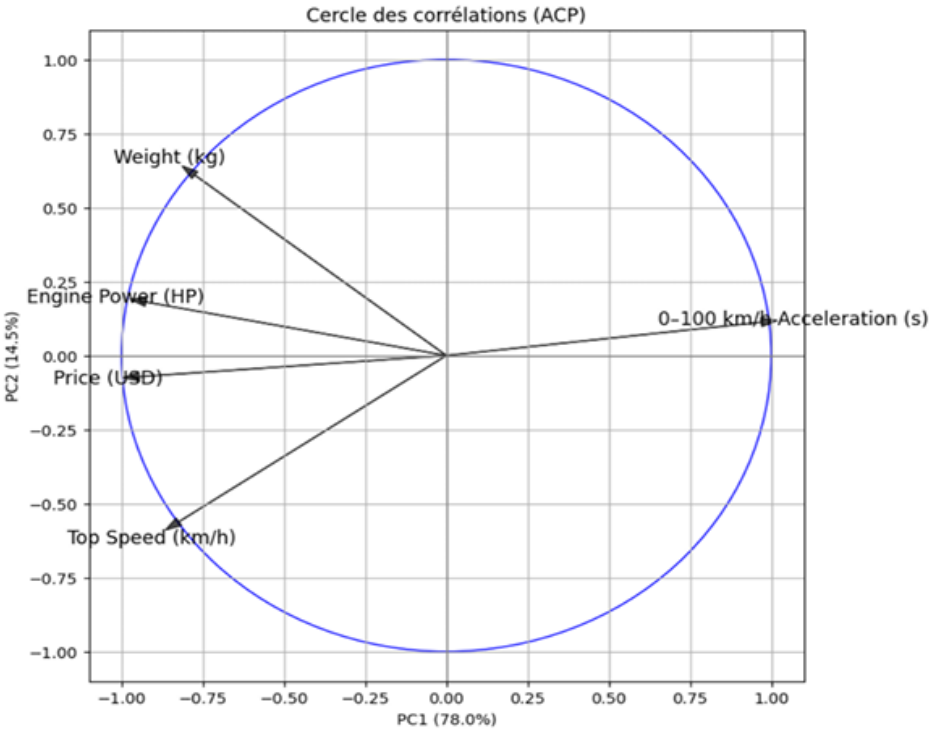

FIGURE 2.3 – Cercle des corrélation ACP

- **PC1 (78,0%) → 1er axe principal**

  - Explique 78% de l'information contenue dans les données.
  - C'est l'axe qui sépare fortement :
    - **Négativement** : voitures chères et puissantes (prix élevé, puissance élevée).
    - **Positivement** : voitures moins chères, moins puissantes , ou avec un accélération élevée (en secondes, donc plus lente).

- **PC2 (14,5%) → 2 axe principal**

  - Il explique 14,5% de l'information.
  - Cet axe est plus lié à des variations de poids et de vitesse .

2.4 Analyse des variables

- **Price (USD)** et **Engine Power (HP)**

- Très corrélés négativement à PC1.
- Plus le prix est maximal et la puissance moteur augmentent, plus la valeur de PC1 est petite.

- **0–100 km/h Acceleration (s)**

  Plus **l'accélération en secondes est élevée, plus c'est lent**. Donc une valeur élevée indique une **mauvaise performance**.

  - Ce qui est cohérent : plus la puissance est **grande**, plus le 0–100 km/h est rapide (moins de secondes).
- **Weight (kg)** et **Top Speed(km/h)**

  - Principalement projeté entre PC1 et PC2.
  - Cela signifie que le **poids** et la **vitesse** ne varient pas forcément avec la puissance ou l'accélération mais représente **un facteur indépendant**.

2.5 Positionnement des voitures

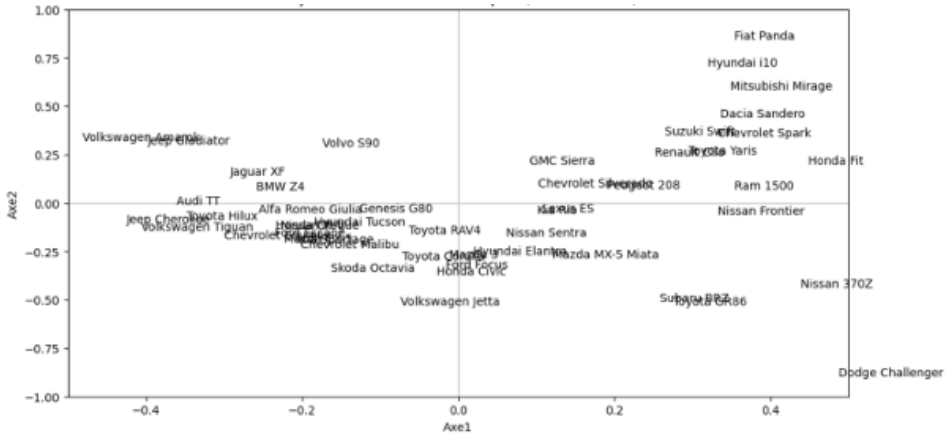

FIGURE 2.4 – Projection des individus sur le plan (Axe 1 et Axe 2)

- **Haut droite** (Fiat Panda, Hyundai i10, Mitsubishi Mirage) : Forte contribution sur les 2 axes donc se sont des voitures lentes et de poids léger .
- **Bas droite** (Dodge Challenger, Nissan 370Z) : voitures rapides avec un poids moyen.
- **À gauche** (Volkswagen Amarok, Jeep Gladiator, Audi TT) : Contribution négative sur l'axe 1 donc se sont des voitures puissantes et de prix élevés .

- **Centre** (Honda CR-V, Toyota Corolla, Toyota RAV4) : voitures proches de la moyenne sur les axes étudiés donc ils ne bien contribuent sur aucun des axes .

2.6 Informations supplémentaires

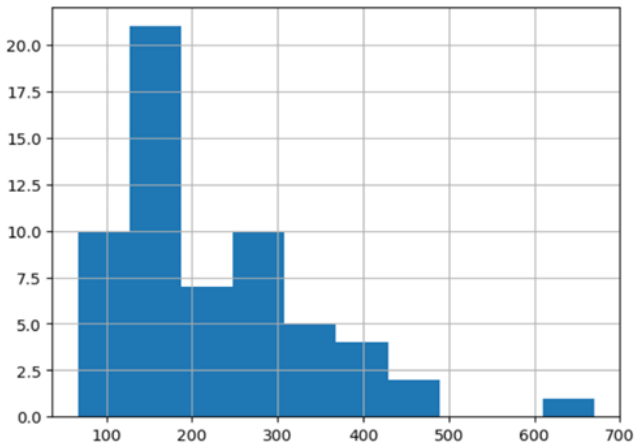

FIGURE 2.5 – Puissance du moteur

- Plus de 20 voitures possèdent une puissance entre 100 HP et 200 HP.
- Environ 10 voitures ont une puissance proche de 300 HP.
- Environ 10 voitures aussi possèdent une puissance proche de 100 HP.

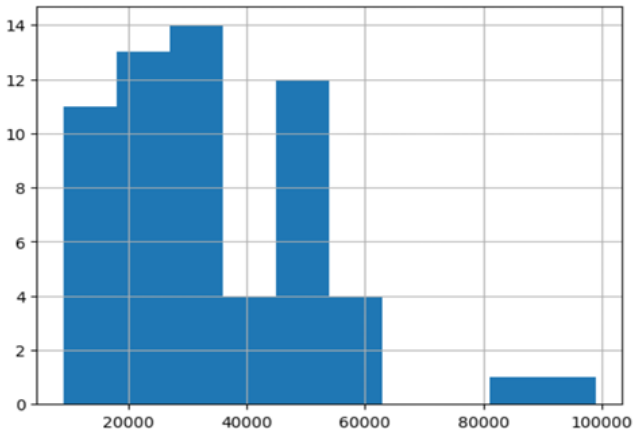

FIGURE 2.6 – Prix(USD)

- La majorité des voitures coûtent moins de 60 000 USD, sauf Porsche 911 (99 000 USD) et Tesla Model S (89 000 USD).

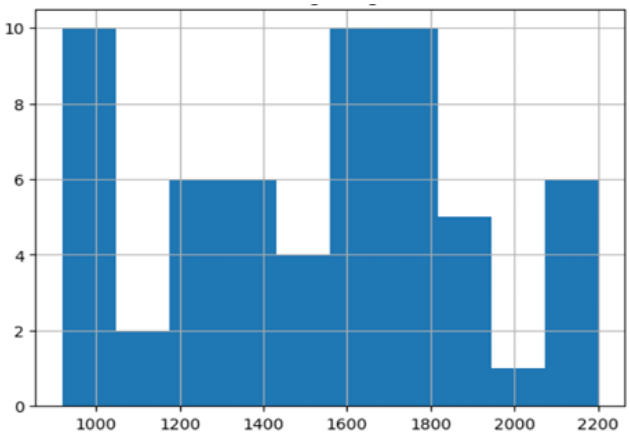

FIGURE 2.7 – Poids(kg)

- Le poids varie entre 500 kg et 2200 kg, avec deux groupes majoritaires : entre 1000 kg et 1200 kg ou entre 1600 kg et 1800 kg.

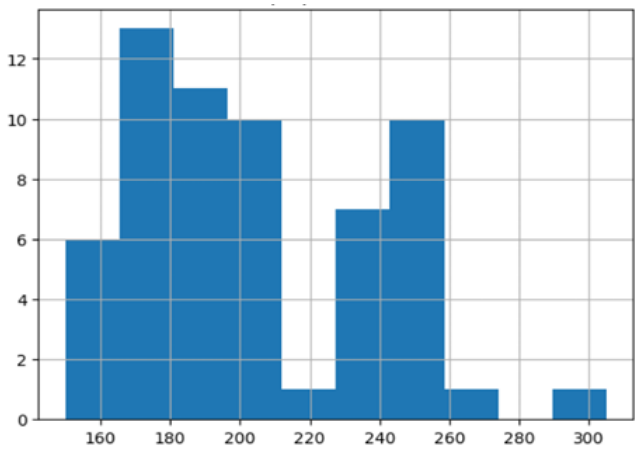

FIGURE 2.8 – vitesse maximale(km/h)

- Les vitesses maximales se répartissent principalement entre 170–210 km/h et 230–260 km/h.

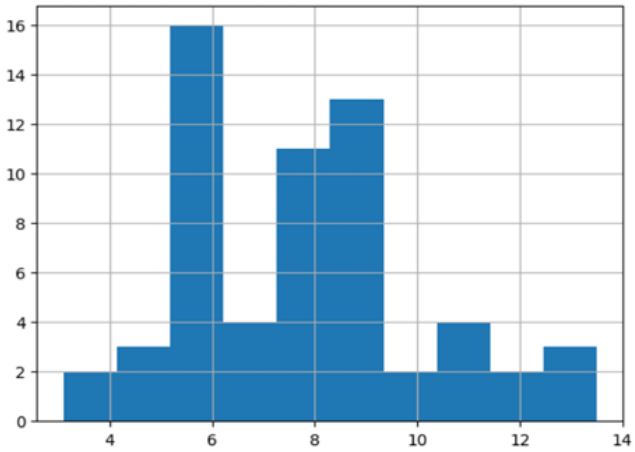

FIGURE 2.9 – 0-100 km/h acceleration(s)

- La majorité des voitures atteignent 100 km/h en moins de 10 secondes :

- 16 voitures environ en 6 secondes.
- Une autre portion en environ 8 secondes.

Conclusion

À travers l'application de l'ACP, nous avons pu réduire la dimensionnalité de notre jeu de données tout en conservant l'essentiel de l'information. Les deux premières composantes principales expliquent une grande partie de la variance, rendant possible une représentation efficace en deux dimensions. L'analyse des corrélations et du positionnement des véhicules a mis en évidence des regroupements logiques basés sur la puissance, le prix, la vitesse maximale et le poids. Cette étude démontre l'intérêt de l'ACP pour simplifier l'analyse de données complexes et en extraire des tendances significatives.

Chapitre 3

ACM

Introduction

L'Analyse des Correspondances Multiples (ACM) est une méthode d'analyse factorielle destinée à explorer les relations entre plusieurs variables qualitatives. Elle offre une représentation graphique qui facilite la compréhension des liaisons entre différentes modalités et individus. Dans ce chapitre, nous appliquons l'ACM sur les variables qualitatives de notre base de données automobile afin d'interpréter les principaux axes de variation et de mieux visualiser les structures sous-jacentes.

3.1 Définition

L'ACM (Analyse des Correspondances Multiples) est une extension de l'Analyse Factorielle des Correspondances (AFC) adaptée aux variables qualitatives (catégorielles). Elle permet de représenter graphiquement des individus et des modalités sur un même plan, facilitant l'interprétation des relations entre plusieurs variables catégorielles.

```
Dimensions du tableau (lignes, colonnes) :
(60, 5)
Nombre d'observations : 60
Nombre de variables qualitatives après transformation : 5
```

FIGURE 3.1 – Dimensions du tableau

Pour faire l'analyse ACM on a transformer nos 2 variables qualitatives (Fuel type et Transmission type) en 5 variables nommés :

- Fuel Type Diesel
- Fuel Type Electric

- Fuel Type Petrol
- Transmission Type Automatic
- Transmission Type Manual

	valprop	inertie	inertiecum
0	5.006759e-01	4.934278e-01	49.342780
1	4.649902e-01	4.582588e-01	95.168656
2	4.902313e-02	4.831344e-02	100.000000
3	1.135880e-16	1.119436e-16	100.000000
4	2.518989e-17	2.482523e-17	100.000000

FIGURE 3.2 – Les résultats du calcul ACM

On remarque que les deux premières composantes expliquent par exemple **95,16%** de la variance totale, ce qui signifie qu'on peut représenter les données initiales dans un espace à 2 dimensions sans trop de perte d'information.

3.2 Analyse des axes principaux

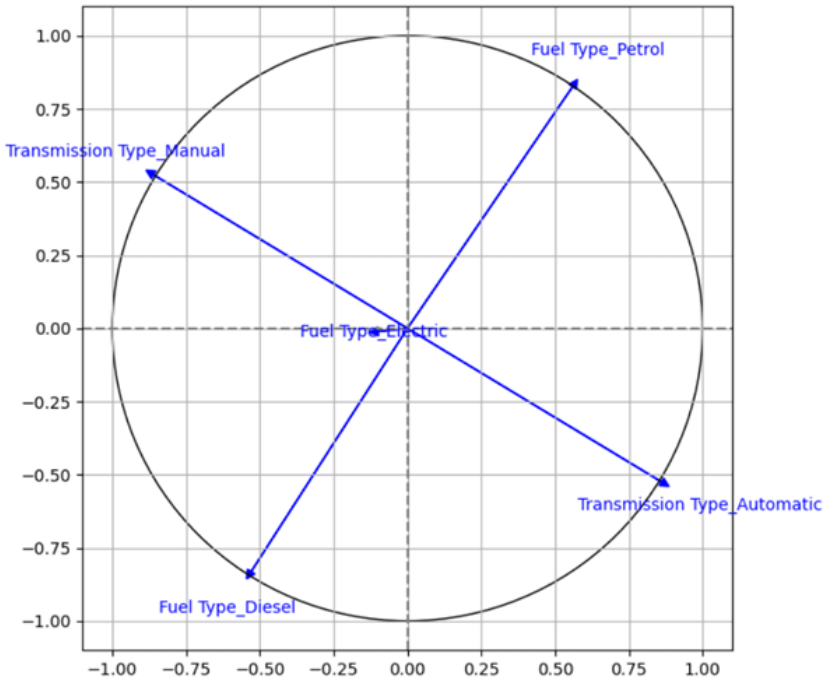

FIGURE 3.3 – Le cercle des corrélations ACM

Selon le cercle de corrélation obtenu :

- **Axe 1** :
  - Représente positivement les voitures à essence (pétrole) et à transmission automatique.

- Représente négativement les voitures diesel et manuelles.

- **Axe 2** :
  - Représente positivement les voitures à essence et à transmission manuelle.
  - Représente négativement les voitures diesel et automatiques.

- Les voitures qui se trouvent prés du centre sont des voitures electrique.

3.3 Positionnement des voitures

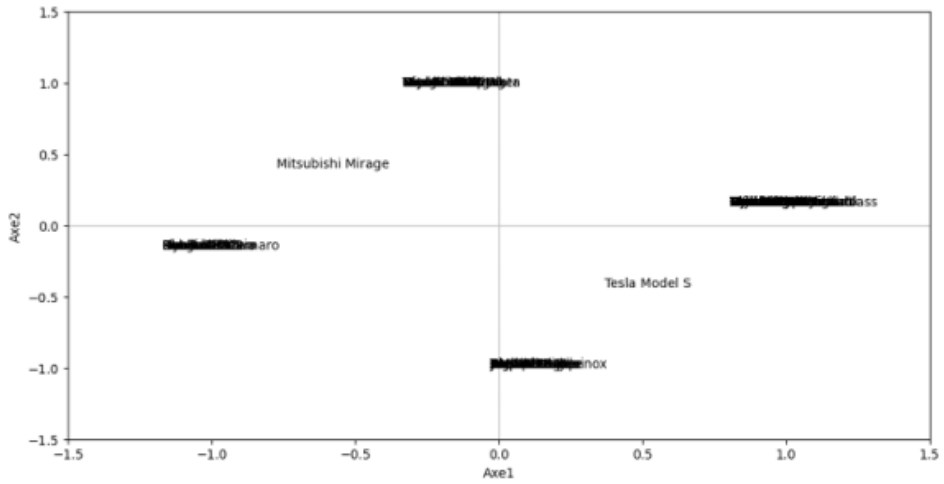

FIGURE 3.4 – Projection des individus ACM

Selon le cercle de corrélation précédente on peut dire que :

- **Partie positive de l'axe 1** :
  - Voitures automatiques à essence, telles que Honda Civic, Volkswagen Tiguan, Toyota RAV4, etc.
- **Partie négative de l'axe 1** :
  - Voitures manuelles et diesel, telles que Kia Rio, Skoda Octavia, Chevrolet Camaro, etc.
- **Partie positive de l'axe 2** :
  - Voitures manuelles à essence, comme Ford Focus, Ford Mustang, Toyota GR86, etc.
- **Partie négative de l'axe 2** :
  - Voitures automatiques et diesel, telles que Toyota Corolla, Kia Sportage, Jeep Cherokee, etc.

À noter que les véhicules Mitsubishi Mirage et Tesla Model S, bien qu'automatiques, n'ont pas de forte contribution sur les deux premiers axes.

Conclusion

L'application de l'ACM nous a permis de mettre en évidence les structures et les associations entre les types de carburant et de transmission des véhicules. Les deux premiers axes expliquent une grande partie de l'information et révèlent des regroupements cohérents selon les combinaisons de motorisation et de boîte de vitesses. Cette analyse confirme la pertinence de l'ACM pour explorer des données qualitatives complexes et facilite la visualisation des tendances principales au sein de notre échantillon automobile.

Chapitre 4

CAH et K-means

Introduction

La segmentation des données est essentielle pour identifier des groupes d'individus homogènes et révéler des structures sous-jacentes. Dans ce chapitre, nous présentons deux méthodes de classification non supervisée : la Classification Ascendante Hiérarchique (CAH) et l'algorithme K-means. Leur complémentarité et la forte cohérence de leurs résultats renforcent la pertinence de notre approche analytique.

4.1 Classification Ascendante Hiérarchique (CAH)

4.1.1 Définition

La CAH (Classification Ascendante Hiérarchique) est une méthode de classification qui regroupe progressivement les individus en clusters selon leur ressemblance, construisant ainsi un arbre (dendrogramme). À chaque étape, les deux groupes les plus similaires sont fusionnés, aboutissant à une hiérarchie complète des regroupements.

4.1.2 Choix du nombre de classes

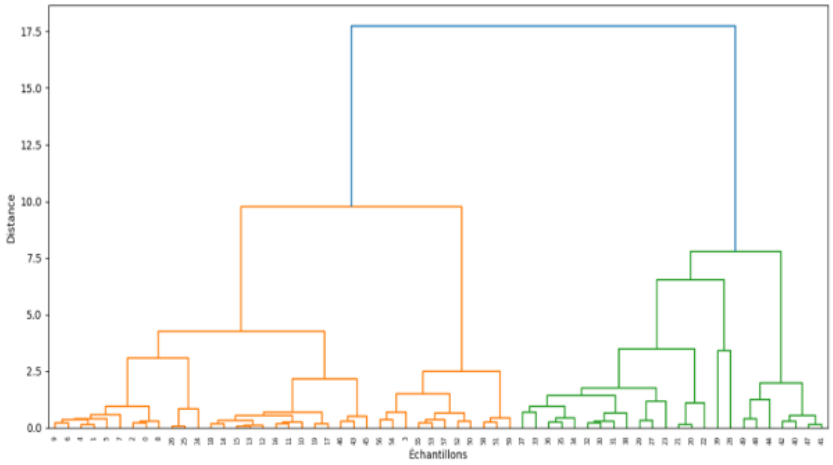

FIGURE 4.1 – Dendrogramme de la CAH (quantitatives uniquement)

Le choix de **5 classes** a été déterminé par l'observation du dendrogramme. Un saut significatif de la distance inter-classes est visible à ce niveau, indiquant une structure naturelle de cinq groupes distincts. Ce seuil garantit des classes à la fois homogènes en interne et bien séparées entre elles.

4.1.3 Interprétation des classes obtenues

Après application de la CAH avec 5 classes, nous obtenons les regroupements suivants :

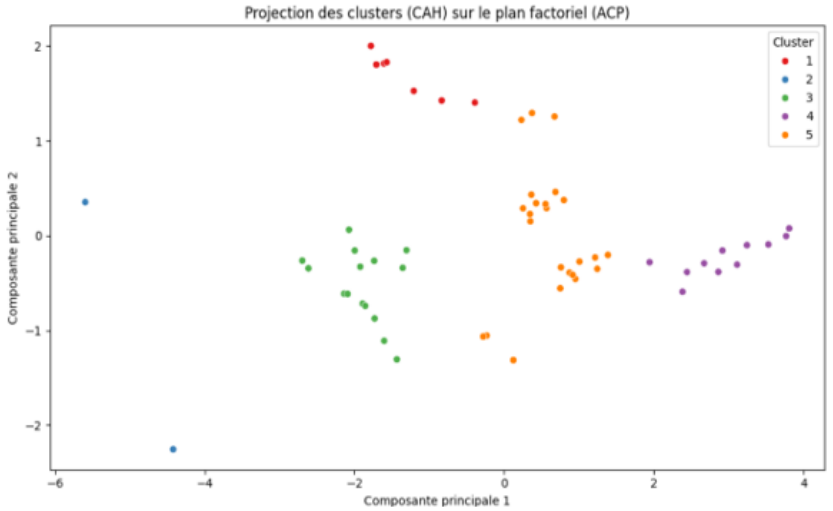

FIGURE 4.2 – Projection des clusters CAH sur le plan factoriel ACP

- **Classe 1** : Voitures lourdes (poids élevé).
- **Classe 2** : Voitures de prix élevé.
- **Classe 3** : Voitures rapides.

- **Classe 4** : Voitures lentes (accélération 0–100 km/h élevée).
- **Classe 5** : Voitures de forte puissance moteur (Engine Power élevé).

4.2 K-means

4.2.1 Définition

Le K-means (Algorithme des k-moyennes) est un algorithme de clustering qui partitionne les données en k groupes en minimisant la variance intra-classe. Chaque individu est assigné au centre de cluster le plus proche, et les centres sont ajustés itérativement jusqu'à stabilisation des regroupements.

4.2.2 Résultats obtenus

L'application de la méthode K-means, avec $k = 5$, aboutit à des regroupements très similaires à ceux obtenus par la CAH.

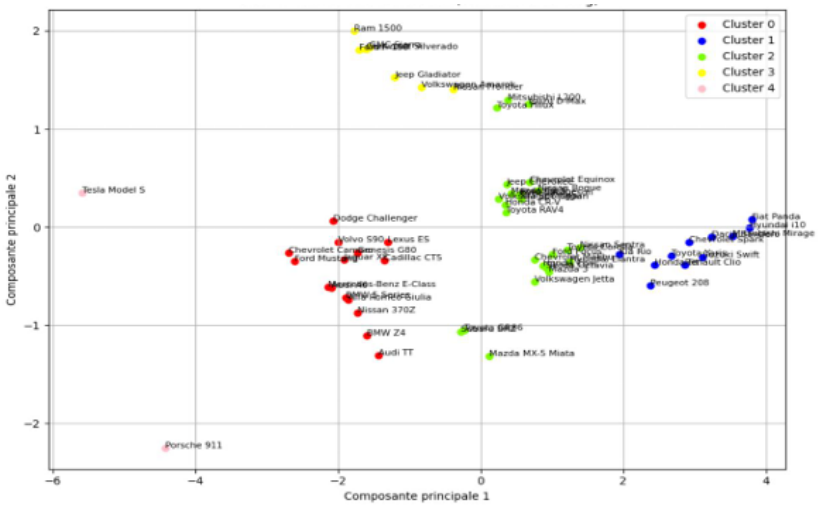

FIGURE 4.3 – Visualtisation ACP des individus (K-means clustering)

La projection sur le plan factoriel issu de l'ACP montre une correspondance étroite entre les deux segmentations, aussi bien en termes de répartition spatiale que de composition des groupes. Cela valide la robustesse et la cohérence de notre segmentation.

4.2.3 Mesure de la similarité entre CAH et K-means

Pour évaluer quantitativement la similarité entre les deux méthodes, nous avons utilisé l'indice ARI (Adjusted Rand Index) avec la commande suivante :

```
ari = adjusted_rand_score(df_quant['Cluster_CAH'], df_quant['Cluster_KMeans'])
```

L'interprétation de l'ARI est la suivante :

- 1 : les deux clusterings sont identiques,
- 0 : les regroupements sont aléatoires,
- < 0 : pire qu'un tirage aléatoire (cas très rare).

Dans notre cas, $\mathbf{ARI = 1}$, confirmant que les deux méthodes produisent exactement les mêmes classes.

4.2.4 Choix optimal du nombre de clusters

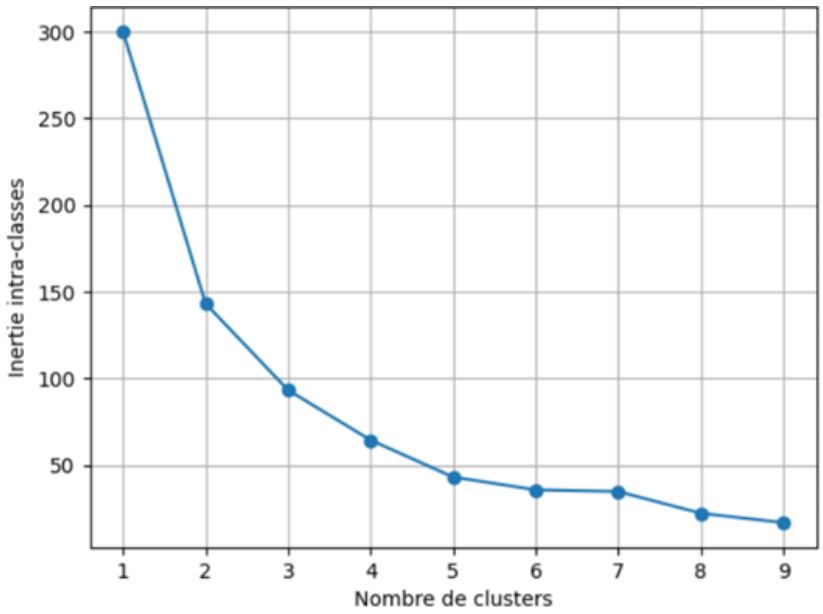

FIGURE 4.4 – Evolution de l'inertie intra-classes selon le nombre de clusters

L'évolution de l'inertie intra-classe montre une forte diminution entre 1 et 5 clusters, indiquant une amélioration significative de la qualité du regroupement. Au-delà de 5 clusters, la baisse devient plus faible, révélant un "effet de coude". Ainsi, le choix de 5 clusters apparaît comme un compromis optimal entre homogénéité des groupes et complexité du modèle.

Conclusion

L'application combinée de la CAH et du K-means a permis d'identifier cinq groupes distincts au sein de notre jeu de données automobile. La forte concordance entre les deux méthodes, validée par l'ARI, confirme la robustesse de la segmentation obtenue. Ces résultats offrent une base solide pour une analyse plus fine des caractéristiques spécifiques à chaque cluster.

Conclusion générale

À travers ce travail, nous avons exploré différentes méthodes d'analyse statistique afin d'extraire des informations pertinentes d'un ensemble de données complexes. L'Analyse en Composantes Principales (ACP) nous a permis de réduire la dimensionnalité tout en visualisant efficacement la structure des variables quantitatives. L'Analyse des Correspondances Multiples (ACM) a, quant à elle, facilité l'interprétation des relations entre variables qualitatives. Pour la classification, la mise en œuvre de la Classification Ascendante Hiérarchique (CAH) et de l'algorithme des k-moyennes (K-means) a permis de segmenter les individus de manière cohérente. La comparaison entre les deux méthodes, validée par des indicateurs tels que l'Adjusted Rand Index (ARI), a montré une forte similarité entre les regroupements obtenus. Globalement, ces analyses complémentaires ont offert une compréhension riche et structurée des données, tout en mettant en évidence les grandes tendances sous-jacentes. Elles constituent une base solide pour toute prise de décision ou étude plus approfondie