RAPPORT
MÉTHODE D'ANALYSES DE DONNÉES

Noa Sans [1]

[1] *Université de Lille, Département d'Ingénierie et Management de la Santé,*
1ère année de Master Data Science en Santé,
noa.sans.etu@univ-lille.fr

Résumé.
Ce projet analyse l'ensemble de données MTCARS, contenant des informations techniques et de performance sur divers modèles de voitures. L'objectif est d'explorer les relations entre les variables, de réduire la complexité des données tout en préservant l'essentiel et d'identifier des groupes homogènes de véhicules. Plusieurs techniques statistiques ont été utilisées : l'analyse de corrélation, l'analyse en composantes principales (ACP) et la classification ascendante hiérarchique (CAH). L'analyse de corrélation a révélé des relations significatives, telles que l'association entre le poids et la consommation de carburant, et entre le nombre de cylindres et la cylindrée. L'ACP a réduit la dimensionnalité en identifiant deux dimensions principales : l'une liée aux performances et à l'efficacité énergétique, l'autre aux caractéristiques mécaniques. La CAH a permis d'identifier des groupes distincts de véhicules similaires. L'utilisation combinée de ces techniques a fourni des informations précieuses sur les relations entre les variables et la segmentation des modèles de voitures.

Mots-clés.
Analyse des données, analyse de corrélation, analyse en composantes principales (ACP), classification ascendante hiérarchique (CAH), réduction de dimensionnalité, clustering

Abstract.
This project analyses the MTCARS dataset, containing technical and performance information on various car models. The goal is to explore the relationships between variables, reduce the complexity of the data while preserving the essentials, and identify homogeneous groups of vehicles. Several statistical techniques were used: correlation analysis, principal component analysis (PCA), and hierarchical ascending clustering (HAC). Correlation analysis revealed significant relationships, such as the association between weight and fuel consumption, and between the number of cylinders and displacement. PCA reduced the dimensionality by identifying two main dimensions: one related to performance and fuel efficiency, the other to mechanical characteristics. HAC allowed to identify distinct groups of similar vehicles. The combined use of these techniques provided valuable insights into the relationships between variables and the segmentation of car models.

Keywords.
Data analysis, correlation analysis, principal component analysis (PCA), hierarchical ascending clustering (HAC), dimensionality reduction, clustering

Sommaire

1. Introduction

Dans ce projet, nous allons étudier le jeu de données MTCARS, qui regroupe des informations techniques et de performance de divers modèles de voitures. L'objectif de ce travail est d'analyser les relations entre les différentes variables présentes dans les données et de regrouper les voitures en clusters homogènes. Cela permettra d'identifier les facteurs clés influençant les performances automobiles et de classer les modèles selon leurs caractéristiques communes.

Pour atteindre ces objectifs, plusieurs étapes d'analyse seront réalisées. Nous commencerons par une analyse de corrélation pour explorer les liens entre les variables et les représenter à l'aide de visualisations comme les corrélogrammes. Ensuite, une analyse en composantes principales (ACP) sera menée afin de réduire la dimensionnalité des données et mettre en évidence les axes principaux qui expliquent la majorité de la variance. Enfin, une classification ascendante hiérarchique (CAH) permettra d'identifier et de visualiser les regroupements d'observations similaires sous forme de dendrogramme.
En combinant ces différentes approches, ce projet vise à fournir une analyse complète et structurée des données MTCARS. Les résultats obtenus permettront de mieux comprendre les relations entre les variables et de segmenter les modèles automobiles en groupes significatifs.

2. Analyse des corrélations et structuration des données

- **Analyse de corrélation**

Un test de corrélation permet de mesurer la force et la direction de la relation entre deux variables, tout en vérifiant sa significativité statistique. Il aide à identifier des relations linéaires, à éviter la redondance dans les modèles et à orienter les décisions dans l'analyse des données.

Nous avons construit la matrice qui regroupe les coefficients de corrélation pour chaque paire de variables. Une matrice de corrélation est un outil statistique essentiel pour analyser les relations linéaires entre plusieurs variables quantitatives. Elle est composée de coefficients de corrélation, qui varient entre -1 et +1, indiquant respectivement une corrélation négative ou positive parfaite, tandis qu'une valeur proche de 0 suggère l'absence de relation linéaire.

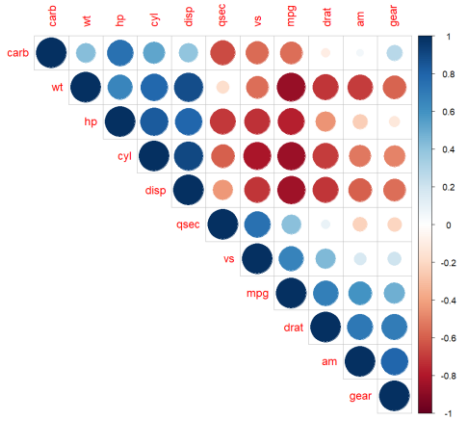
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.000	-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.600	0.480	-0.551
cyl	-0.852	1.000	0.902	0.832	-0.700	0.782	-0.591	-0.811	-0.523	-0.493	0.527
disp	-0.848	0.902	1.000	0.791	-0.710	0.888	-0.434	-0.710	-0.591	-0.556	0.395
hp	-0.776	0.832	0.791	1.000	-0.449	0.659	-0.708	-0.723	-0.243	-0.126	0.750
drat	0.681	-0.700	-0.710	-0.449	1.000	-0.712	0.091	0.440	0.713	0.700	-0.091
wt	-0.868	0.782	0.888	0.659	-0.712	1.000	-0.175	-0.555	-0.692	-0.583	0.428
qsec	0.419	-0.591	-0.434	-0.708	0.091	-0.175	1.000	0.745	-0.230	-0.213	-0.656
vs	0.664	-0.811	-0.710	-0.723	0.440	-0.555	0.745	1.000	0.168	0.206	-0.570
am	0.600	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168	1.000	0.794	0.058
gear	0.480	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	1.000	0.274
carb	-0.551	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274	1.000

La matrice de corrélation permet donc d'identifier les relations linéaires entre les différentes variables. On observe une corrélation négative forte entre la consommation de carburant (mpg) et le poids du véhicule (wt) (-0.868), indiquant que les voitures plus lourdes consomment davantage de carburant. Une corrélation positive est également présente entre le nombre de cylindres (cyl) et la cylindrée (disp)

(0.902), des moteurs avec plus de cylindres ont une cylindrée plus élevée. D'autres relations, comme la corrélation forte entre la puissance du moteur (hp) et le nombre de carburateurs (carb) (0.750), montrent des associations directes liées aux caractéristiques mécaniques des véhicules. À l'inverse, certaines variables, comme le temps au quart de mile (qsec) et le rapport du pont arrière (drat) (0.091), n'ont quasiment aucune relation linéaire.

Afin de mieux visualiser les groupes de variables qui partagent des patterns similaires et de voir clairement quelles variables sont fortement corrélées entre elles, un clustering hiérarchique peut être réalisé. C'est une technique qui permet de regrouper les variables qui sont les plus fortement corrélées entre elles. Ainsi, les variables les plus similaires en termes de corrélation seront regroupées et placées à proximité les unes des autres sur le graphique.

Les couleurs et tailles des cercles sont accompagnées d'une échelle de couleur qui montre les valeurs de corrélation, allant de -1 (corrélation négative parfaite) à +1 (corrélation positive parfaite).



L'hypothèse nulle (H_0) dans le cadre de ce test de corrélation est qu'il n'y a pas de relation linéaire entre les variables. L'hypothèse nulle suppose qu'aucune association linéaire significative n'existe entre les deux variables.

Lors d'une analyse de corrélation, le calcul de la p-value est essentiel pour évaluer la significativité statistique du coefficient de corrélation.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	0.0000000	0.0000000	0.0000000	0.0000002	0.0000178	0.0000000	0.0170820	0.0000342	0.0002850	0.0054009	0.0010844
cyl	0.0000000	0.0000000	0.0000000	0.0000000	0.0000082	0.0000001	0.0003661	0.0000000	0.0021512	0.0041733	0.0019423
disp	0.0000000	0.0000000	0.0000000	0.0000001	0.0000053	0.0000000	0.0131440	0.0000052	0.0003662	0.0009636	0.0252679
hp	0.0000002	0.0000000	0.0000001	0.0000000	0.0099888	0.0000415	0.0000058	0.0000029	0.1798309	0.4930119	0.0000008
drat	0.0000178	0.0000082	0.0000053	0.0099888	0.0000000	0.0000048	0.6195826	0.0116755	0.0000047	0.0000084	0.6211834
wt	0.0000000	0.0000001	0.0000000	0.0000415	0.0000048	0.0000000	0.3388683	0.0009798	0.0000113	0.0004587	0.0146386
qsec	0.0170820	0.0003661	0.0131440	0.0000058	0.6195826	0.3388683	0.0000000	0.0000010	0.2056621	0.2425344	0.0000454
vs	0.0000342	0.0000000	0.0000052	0.0000029	0.0116755	0.0009798	0.0000010	0.0000000	0.3570439	0.2579439	0.0006670
am	0.0002850	0.0021512	0.0003662	0.1798309	0.0000047	0.0000113	0.2056621	0.3570439	0.0000000	0.0000001	0.7544526
gear	0.0054009	0.0041733	0.0009636	0.4930119	0.0000084	0.0004587	0.2425344	0.2579439	0.0000001	0.0000000	0.1290291
carb	0.0010844	0.0019423	0.0252679	0.0000008	0.6211834	0.0146386	0.0000454	0.0006670	0.7544526	0.1290291	0.0000000

Dans la majorité de nos cas, la p-value obtenue dans le test de corrélation est inférieure au seuil de significativité $\alpha = 5\%$ que nous avons pris, alors l'hypothèse nulle est rejetée, ce qui signifie qu'il existe une corrélation linéaire significative entre les variables.

Après avoir déterminé les relations de corrélation entre les variables, il est pertinent de poursuivre l'analyse en réalisant l'Analyse en Composantes Principales (ACP).

- ## Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) est une méthode statistique multivariée utilisée pour réduire la dimensionnalité d'un jeu de données tout en préservant autant que possible l'information originale. Cette technique permet de transformer un grand nombre de variables corrélées en un plus petit nombre de composantes principales qui sont linéairement indépendantes.

Pour évaluer la qualité des composantes principales, optimiser la réduction dimensionnelle et mieux comprendre la structure des données, il est nécessaire de calculer les valeurs propres.

Histogramme des valeurs propres

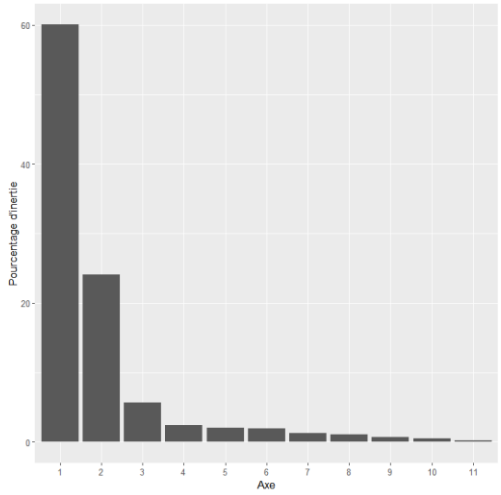


Tableau des valeurs propres

Axe	%	Cum. %
1	60.1	60.1
2	24.1	84.2
3	5.7	89.9
4	2.5	92.3
5	2.0	94.4
6	1.9	96.3
7	1.2	97.5
8	1.1	98.6
9	0.7	99.3
10	0.5	99.8
11	0.2	100.0

Une valeur propre élevée indique qu'une composante principale explique une grande part de la variance des données, ce qui signifie qu'elle est importante pour l'analyse.

Une méthode est couramment utilisée pour déterminer le nombre de composantes principales à retenir, la règle des 80 %. Elle consiste à sélectionner suffisamment de composantes pour expliquer au moins 80 % de la variance totale des données. En cumulant ces valeurs, on identifie le nombre minimal de dimensions nécessaires pour représenter l'essentiel de l'information initiale.
Dans notre cas, les deux dimensions les plus importantes représentent 84,2% de la variance totale, ce sont ces deux dimensions qui seront par la suite utilisées pour la poursuite de l'analyse.
A noter que ces deux dimensions répondent également à la règle de Kaiser qui stipule que seules les composantes principales ayant une valeur propre supérieure ou égale à 1 doivent être retenues.

Maintenant que les valeurs propres ont été calculées, nous pouvons nous concentrer sur l'analyse des deux dimensions retenues, qui expliquent une grande partie de la variance des données. Nous conservons les variables dont la contribution est supérieure à la moyenne. La moyenne des contributions est calculée en divisant 100 par le nombre total de variables (ici 11). Cela permet de déterminer un seuil au-delà duquel les variables sont considérées comme ayant un impact significatif sur les composantes principales.

Variable	Coord	Contrib	Cos2	Cor
cyl	0.961	13.98	0.924	0.961
disp	0.946	13.56	0.896	0.946
mpg	-0.932	13.14	0.869	-0.932
wt	0.890	11.98	0.792	0.89
hp	0.848	10.89	0.720	0.848
vs	-0.788	9.39	0.621	-0.788
drat	-0.756	8.65	0.572	-0.756
am	-0.604	5.52	0.365	-0.604
carb	0.550	4.58	0.303	0.55
gear	-0.532	4.28	0.283	-0.532
qsec	-0.515	4.02	0.266	-0.515

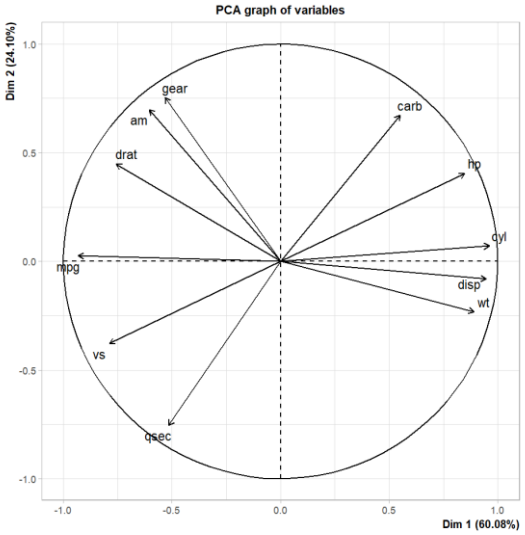
La première dimension de l'analyse en composantes principales est principalement définie par les variables ayant une contribution supérieure à la moyenne, soit 9.09 %. Les variables retenues sont cyl (13.98 %), disp (13.56 %), mpg (13.14 %), wt (11.98 %), hp (10.89 %) et vs (9,39%). Ces variables

jouent un rôle central dans la définition de cet axe. Les variables cyl, disp, wt, et hp sont fortement corrélées positivement avec la dimension, ce qui indique qu'elles représentent des caractéristiques similaires, probablement liées à la taille et à la puissance des véhicules. En revanche, mpg et vs sont fortement corrélées négativement, ce qui reflète une opposition avec les autres variables, traduisant une tendance inverse liée à l'efficacité énergétique des véhicules.
Ainsi, cette dimension peut être interprétée comme un contraste entre des véhicules puissants et imposants, d'une part, et d'autre part des véhicules économes en carburant.

Variable	Coord	Contrib	Cos2	Cor
qsec	-0.754	21.47	0.569	-0.754
gear	0.753	21.38	0.567	0.753
am	0.699	18.44	0.489	0.699
carb	0.673	17.10	0.453	0.673
drat	0.447	7.55	0.200	0.447
hp	0.405	6.19	0.164	0.405
vs	-0.377	5.37	0.142	-0.377
wt	-0.233	2.05	0.054	-0.233
disp	-0.080	0.24	0.006	-0.08
cyl	0.071	0.19	0.005	0.071
mpg	0.026	0.03	0.001	0.026

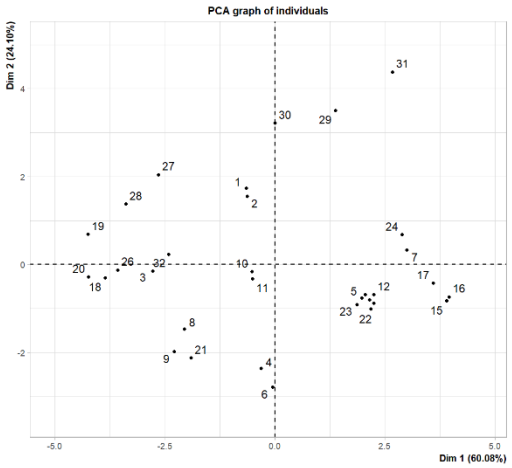
Pour la seconde dimension, les variables retenues sont qsec (21.47 %), gear (21.38 %), am (18.44 %) et carb (17.10 %). Ces variables sont les plus influentes dans la définition de cet axe. Les variables gear, am, et carb sont positivement corrélées avec cette dimension, indiquant une association entre ces caractéristiques, tandis que qsec présente une corrélation négative, traduisant une tendance opposée. Cette dimension semble refléter une distinction entre des variables liées à la transmission et aux caractéristiques mécaniques d'un véhicule et des aspects liés au temps d'accélération (qsec). L'axe représente un contraste entre la rapidité d'un véhicule et ses spécificités mécaniques telles que le type de transmission et le nombre de carburateurs.

Après avoir exploré les dimensions principales, il est essentiel de visualiser les observations dans l'espace des composantes principales pour analyser les regroupements et les relations entre les individus.



Ce graphique, appelé cercle des corrélations, illustre comment les variables d'origine se projettent dans les nouvelles dimensions créées par l'ACP. L'interprétation de ce graphique repose principalement sur l'angle entre les variables. Plus l'angle entre deux flèches représentant deux variables est petit, plus ces variables sont fortement corrélées. De plus, si les flèches sont proches du cercle, cela signifie que ces variables sont bien représentées dans l'espace des composantes principales et ont une bonne qualité de représentation.
En examinant ce graphique, on peut identifier les groupes de variables fortement corrélées, qui apparaissent proches les unes des autres sur le cercle.

Ce graphique, quant à lui, permet d'examiner comment les différentes observations se distribuent selon les composantes principales et permet d'analyser leur similarité. Les individus proches les uns des autres sur ce graphique partagent des profils similaires, c'est-à-dire que leurs caractéristiques mesurées dans les variables d'origine sont comparables dans l'espace réduit des composantes principales. Ce type de représentation permet également d'identifier des regroupements naturels ou des clusters dans les données, où des individus similaires sont regroupés dans des zones spécifiques de l'espace des composantes principales.

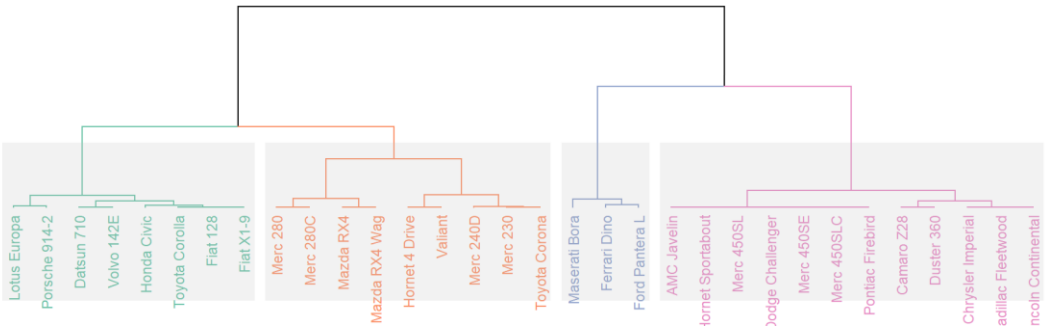


Ces deux types de graphiques offrent des perspectives différentes mais complémentaires sur les résultats de l'ACP. Le premier aide à comprendre les relations entre les variables et leur contribution à l'ACP, tandis que le second permet d'explorer la distribution et les similitudes des individus dans l'espace des composantes principales.

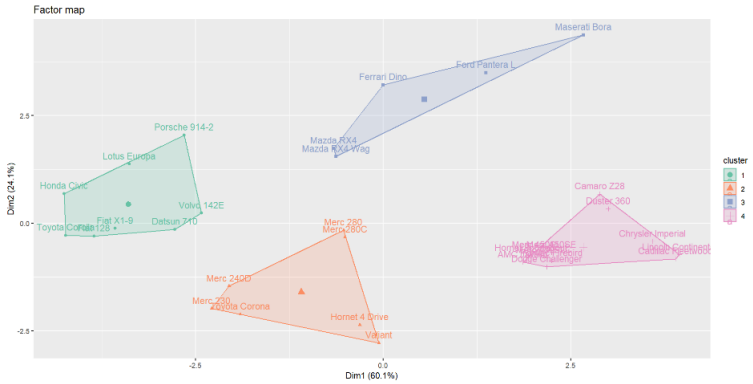
- **Classification Ascendante Hiérarchique (CAH)**

La Classification Ascendante Hiérarchique (CAH) repose sur une approche itérative et hiérarchique. Elle débute en considérant chaque observation comme un cluster individuel, puis fusionne progressivement les clusters les plus similaires jusqu'à ce qu'il n'en reste qu'un seul contenant toutes les observations.

La méthode de Ward est une technique d'agrégation utilisée dans la classification ascendante hiérarchique (CAH) qui vise à minimiser la variance intra-cluster lors du regroupement des individus. Elle fonctionne en fusionnant les groupes de manière à ce que l'augmentation de la variance à l'intérieur des nouveaux groupes soit la plus faible possible. Les branches de l'arbre montrent comment les individus (voitures) sont regroupés en fonction de leur similarité. Plus les branches se rejoignent bas sur l'axe vertical, plus les groupes sont similaires entre eux.



Le dendrogramme, illustre les regroupements entre différents modèles de voitures en fonction de leurs similitudes. Chaque branche représente un groupe d'individus similaires, avec des fusions successives indiquant des proximités croissantes. Par exemple, des voitures comme Lotus Europa et Porsche 914/2 appartiennent au même groupe, en raison de caractéristiques partagées. De la même manière, des modèles comme Maserati Bora et Ferrari Dino forment un groupe qui pourrait refléter des caractéristiques propres aux voitures de sport.



Ce graphique montre les clusters projetés sur l'espace factoriel. Les individus (voitures) sont répartis en groupes colorés en fonction de leur appartenance à un cluster. Les centres des clusters sont également visibles, donnant ainsi une idée de leur position moyenne dans l'espace des composantes principales. Les proximités entre points au sein d'un cluster indiquent des similarités, tandis que la distance entre clusters reflète leurs différences.

3. Résultats et interprétations

L'analyse des données a révélé plusieurs relations significatives entre les variables.

Tout d'abord, la matrice de corrélation a montré une forte corrélation négative entre le poids des véhicules (wt) et leur consommation de carburant (mpg), indiquant qu'un poids plus élevé est généralement associé à une consommation accrue. De plus, une corrélation positive a été observée entre le nombre de cylindres (cyl) et la cylindrée du moteur (disp), suggérant une relation directe entre ces deux variables qui mesurent la taille du moteur.

L'Analyse en Composantes Principales (ACP) a permis de réduire la dimensionnalité des données tout en conservant l'essentiel de l'information. Les deux premières dimensions extraites expliquent la majorité de la variance. La première est associée à des variables telles que la consommation de carburant (mpg), le poids (wt) et la puissance (hp), traduisant la performance et l'efficacité énergétique des véhicules. La deuxième dimension capture davantage les caractéristiques mécaniques, telles que le type de transmission (am) et la configuration du moteur (vs).

En combinant les résultats de l'ACP et de la Classification Ascendante Hiérarchique (CAH), nous avons identifié quatre clusters distincts de véhicules, chacun ayant des caractéristiques techniques et de performance spécifiques.
Un premier cluster (violet), caractérisé par des valeurs positives sur les deux premières dimensions, regroupe des modèles comme la Maserati Bora, la Ferrari Dino et la Mazda RX4. Ces véhicules puissants, dotés de moteurs sophistiqués et lourds, sont peu économes en carburant et représentent des voitures sportives alliant performance et consommation.
Le cluster vert, avec une dimension 1 négative et une dimension 2 positive, comprend des modèles comme la Lotus Europa et la Volvo 142E. Ces véhicules sont économes en carburant, bien équipés mécaniquement, ils allient performance modérée et efficacité énergétique.
Le cluster orange, marqué par des valeurs négatives sur les deux dimensions, regroupe des véhicules comme la Mercedes 280C, la Toyota Corona et la Valiant. Ces modèles, économes en carburant mais

manquant de performance, privilégient l'efficacité énergétique.
Enfin, le cluster rose, avec une dimension 1 positive et une dimension 2 négative, inclut des modèles comme la Camaro Z28, la Chrysler Imperial et la Dodge Challenger. Ces véhicules allient puissance et accélération rapide, caractéristiques des voitures de sport, mais avec un compromis sur l'efficacité énergétique.

4. Conclusion

Ce projet a permis de mener une analyse approfondie des données automobiles à l'aide de techniques telles que l'Analyse en Composantes Principales (ACP) et la Classification Ascendante Hiérarchique (CAH). L'ACP a été efficace pour réduire la dimensionnalité des données tout en préservant l'essentiel de l'information. Elle a mis en évidence des relations significatives entre des variables clés telles que la consommation de carburant, le poids et la puissance des véhicules. La CAH, de son côté, a permis d'identifier des groupes de véhicules partageant des caractéristiques communes en termes de puissance, de consommation et de propriétés mécaniques. Ces résultats confirment l'efficacité de ces méthodes pour classer et segmenter les données, offrant ainsi une base solide pour des analyses futures et une meilleure compréhension des modèles de véhicules.