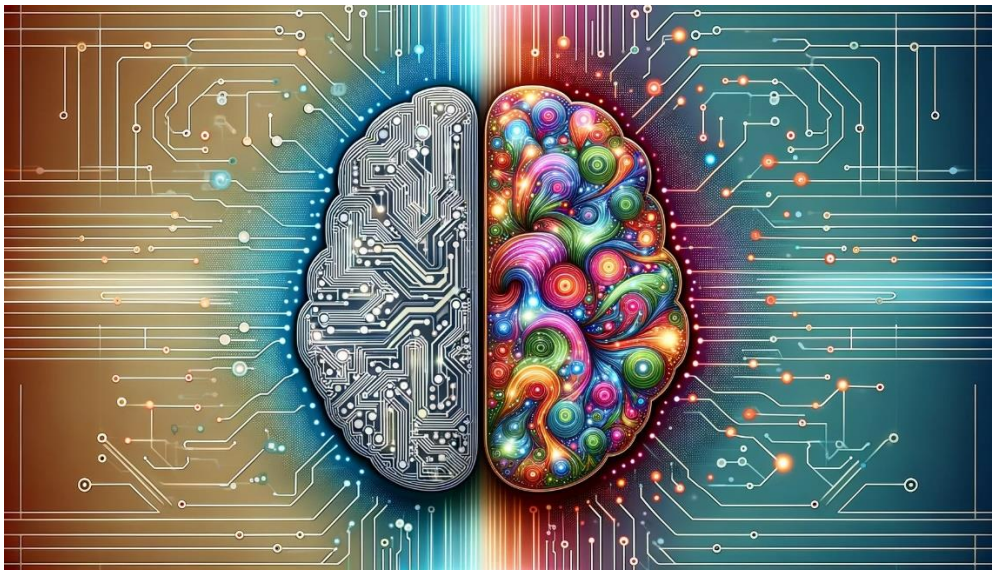
UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE
ÉCOLE D'ÉCONOMIE
DE LA SORBONNE

Master 1 Économétrie-Statistiques

Rapport du projet Machine Learning et Analyse de données



Sous la direction de Monsieur Joseph RYNKIEWICZ

Réalisée par :

Maeva N'GUESSAN

Roland DUTAUZIET

Bethuel ASSE

2023-2024

ABSTRACT

Dans le cadre de notre projet d'analyse de données et Machine Learning nous avons abordé quatre exercices distincts, en utilisant des techniques variées pour explorer et interpréter des ensembles de données complexes. L'exercice 1 a impliqué une régression logistique pour modéliser une variable dépendante binaire à partir de deux prédicteurs, révélant une précision satisfaisante pour l'une des classes. L'ACP de l'exercice 2 a permis de décomposer la variance de caractéristiques de véhicules divers et a abouti à des interprétations significatives des principaux facteurs de variation. Les méthodes de clustering k-means et hiérarchique de l'exercice 3 ont offert des perspectives sur la segmentation des véhicules, tandis que l'analyse des correspondances multiples de l'exercice 4 a donné un aperçu des associations entre les races de chiens et leurs fonctions. Ce rapport synthétise les résultats de chaque approche, illustrés par des graphiques, et conclut sur leur pertinence et leur potentiel d'application.

SOMMAIRE

I. INTRODUCTION

L'analyse de données et l'apprentissage automatique représentent des outils puissants et indispensables dans la compréhension de phénomènes complexes et la prise de décisions éclairées dans de nombreux domaines. À travers ce projet, nous explorons l'application de méthodes statistiques et algorithmiques à une variété de jeux de données pour en extraire des informations pertinentes et des modèles prédictifs. Le projet se compose de quatre exercices distincts, chacun ciblant une compétence spécifique et exploitant une méthode d'analyse différente.

Le premier exercice met en œuvre la régression logistique et teste différents modèles afin d'identifier le meilleur dans le cadre d'une régression binaire. Cette approche nous permet de discerner des relations non linéaires et d'évaluer la force et la direction des associations entre les variables explicatives continues.

Dans le deuxième exercice, nous appliquons l'analyse en composantes principales (ACP) pour réduire la dimensionnalité d'un ensemble de données sur les véhicules, en identifiant les principaux facteurs qui capturent la majeure partie de la variance.

La classification non supervisée est explorée dans le troisième exercice à travers les méthodes k-means et la classification hiérarchique, offrant un aperçu de la structuration naturelle des données. Enfin, le quatrième exercice se concentre sur l'analyse des correspondances multiples, une technique adaptée à l'examen des relations entre des variables catégorielles, ici appliquée aux caractéristiques des races de chiens.

Ce rapport documente les méthodes utilisées, les analyses effectuées, et les interprétations dérivées de chaque exercice, illustrant la manière dont ces techniques peuvent être employées pour extraire des connaissances à partir de données variées. Il vise à fournir une compréhension globale des applications de l'analyse de données et de l'apprentissage automatique, ainsi qu'à montrer comment ces méthodes peuvent être utilisées pour aborder des questions pratiques et théorique.

II. EXERCICES ET ANALYSES

1) Régression binaire

Cet exercice implique l'analyse et la modélisation d'un ensemble de données simulées pour prédire une variable catégorielle binaire. Nous avons une variable Y à expliquer à partir d'observations de deux variables explicatives X1 et X2. Mais quel modèle serait le plus adéquat ?

Nous commencerons par faire une brève exploration des données, ensuite l'application et l'évaluation de différents modèles et enfin nous prédirons des nouvelles observations à partir de nouvelles données

1.1 Exploration et preprocessing des données

- **Chargement des librairies**
    - Numpy et Pandas pour les matrices et dataframes
    - Matplotib et Seaborn pour la visualisation
    - Différentes fonctions de sklearn pour l'entrainement, l'évaluation et l'optimisation des différents modèles

- **Quelques statistiques descriptives**

	X1	X2	Y
0	-1.681427	-1.534811	1
1	-0.690532	0.710814	1
2	4.676125	-1.624768	2
3	0.211525	3.657683	2
4	0.387863	0.522408	2

```
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X1      2000 non-null   float64
 1   X2      2000 non-null   float64
 2   Y       2000 non-null   int64
dtypes: float64(2), int64(1)
memory usage: 47.0 KB
X1    0
X2    0
Y     0
dtype: int64
```

	X1	X2	Y
count	2000.000000	2000.000000	2000.000000
mean	0.087890	-0.043909	1.414000
std	3.001684	2.955444	0.492672
min	-9.143583	-9.387265	1.000000
25%	-1.931862	-1.944058	1.000000
50%	0.086439	-0.074521	1.000000
75%	2.125775	1.945073	2.000000
max	10.171112	10.263284	2.000000

Il y a 2000 observations de 2 variables explicatives X1 et X2 et une variable expliquée Y qui est binaire (valeurs= 1 ou 2). Pas de valeurs manquantes.

- **Division des données**

Vu la nature du problème (régression binaire), des modèles tels que la régression logistique, les arbres de décision ou le Random Forest pourraient être appropriés. On va commencer par des modèles simples, puis monter en complexité. Tout d'abord, on divise nos données en 80% pour l'entrainement et 20% pour le test.

1.2 Application et évaluation de différents modèles (Question 1)

*Pour chaque modèle, nous procédons à l'initialisation, l'optimisation et entrainement du modèle, son évaluation et son interprétation. On initialise le modèle et on l'optimise à l'aide de recherche d'hyperparamètres (**GridSearchCV**). Puis, on entraine le modèle optimisé sur l'ensemble d'entrainement (**.fit**) et nous l'évaluons à l'aide des prédictions du modèle sur l'ensemble de test (**.predict**) et les vraies prédictions de l'ensemble de test. On fait tout cela sur chaque modèle en s'assurant chaque fois de la reproductibilité des résultats (**random_state=42**).*

- **Régression linéaire** : Cette méthode statistique modélise la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes en ajustant une équation linéaire aux données observées.

Nos scores

MSE : 0.22721287305046942
R2 : 0.06410102749266033

Interprétation

Ici, ce modèle n'est pas adapté à notre problème de classification binaire car notre variable expliquée est discrète. Cela se confirme avec nos très maigres scores à la suite de l'évaluation du modèle. Vu la binarité de notre explicative, nous allons procéder à une régression logistique.

- **Régression logistique :** Elle est utilisée pour modéliser la probabilité d'une variable dépendante catégorielle, généralement binaire, en fonction d'une ou plusieurs variables indépendantes, en utilisant une fonction logistique :

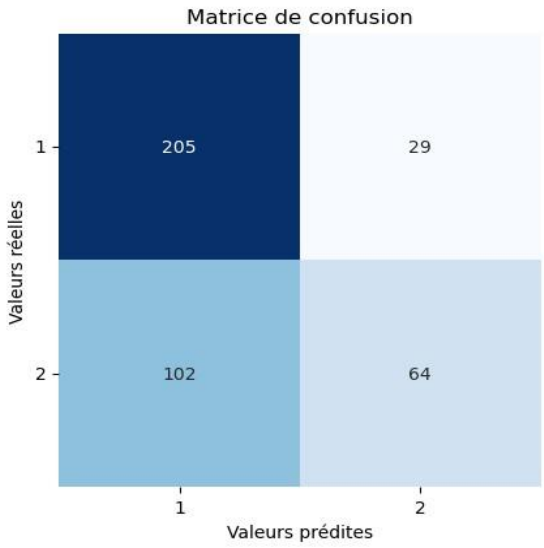
$$P(Y = 1 \mid X_1, X_2) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}}$$

On utilise la régularisation Ridge pour prévenir l'overfitting mais ici ce n'est pas vraiment nécessaire

Nos scores

Accuracy : 0.6725
Rapport de classification :

	precision	recall	f1-score	support
1	0.67	0.88	0.76	234
2	0.69	0.39	0.49	166
accuracy			0.67	400
macro avg	0.68	0.63	0.63	400
weighted avg	0.68	0.67	0.65	40



Interprétation :

On a une **accuracy** de 67,25% : La régression logistique est plutôt performante.

On a une meilleure sensibilité (capacité du modèle à retrouver les Y=1)(**recall**=0.88) mais une faible spécificité (0.39).

Le modèle se trompe beaucoup sur la classe 2. Cela peut être dû au fait que les deux classes sont un peu déséquilibrées.

- **Arbre de décision :** Il effectue des prédictions en divisant les données en sous-ensembles de plus en plus homogènes, selon des critères de décision définis à chaque nœud de l'arbre, formant une structure arborescente.

Nos hyperparamètres :

On a donc le critère d'entropy ($\sum_1^2 p_i ln\,(1 - p_i)$) et la profondeur maximale de l'arbre qui est de 7.

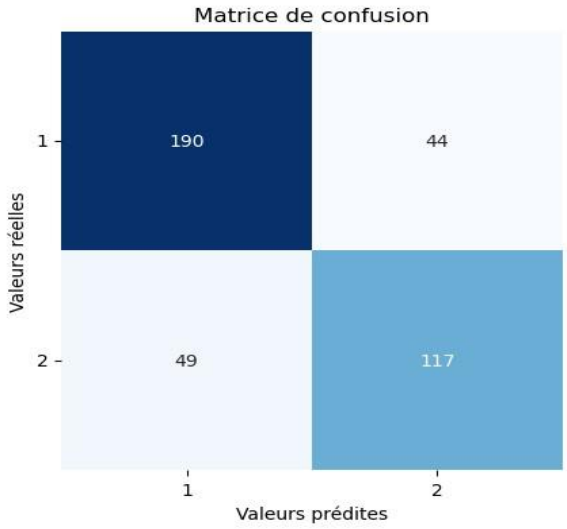
Meilleurs paramètres : {'criterion': 'entropy', 'max_depth': 7}

Nos scores

Accuracy: 0.7675

Rapport de classification:

	precision	recall	f1-score	support
1	0.79	0.81	0.80	234
2	0.73	0.70	0.72	166
accuracy			0.77	400
macro avg	0.76	0.76	0.76	400
weighted avg	0.77	0.77	0.77	400



Interprétation :

L'arbre de décision est facilement interprétable.

On a une accuracy de 76,75% : Donc, notre arbre de décision est plus performant que la régression logistique.

On a d'assez bons recall (0.81 pour 1 et 0.70 pour 2)

Cependant, il y a de fortes variations de l'arbre si on change les données d'entrainement

- **Bagging :** Il est une technique d'ensemble qui entraîne plusieurs modèles sur des sous-ensembles aléatoires des données d'entraînement et en combinant leurs prédictions.

Nos hyperparamètres :

On a 50 arbres de 50% au maximum des données pour chaque bootstrap.

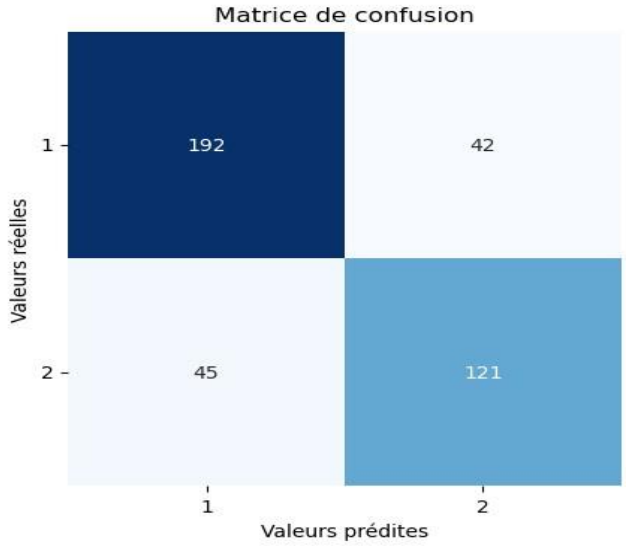
Meilleurs paramètres : {'max_samples': 0.5, 'n_estimators': 50}

Nos scores

Accuracy: 0.7825

Rapport de classification:

	precision	recall	f1-score	support
1	0.81	0.82	0.82	234
2	0.74	0.73	0.74	166
accuracy			0.78	400
macro avg	0.78	0.77	0.78	400
weighted avg	0.78	0.78	0.78	400



Interprétation :

On a une accuracy de 78,25% : Le modèle de Bagging des arbres est un peu plus performant que l'arbre de décision tout seul, d'après nos métriques.

On a de bons recall (0.82 pour 1 et 0.73 pour 2)

- **Random-forest :** C’est un algorithme d'apprentissage supervisé qui construit un ensemble de plusieurs arbres de décision entraînés sur des sous-ensembles de données différents, améliorant la robustesse et la précision par rapport à un seul arbre de décision.

Nos hyperparamètres :

On a 100 arbres, de profondeur maximale 10 et 20% au maximum des données pour chaque bootstrap.

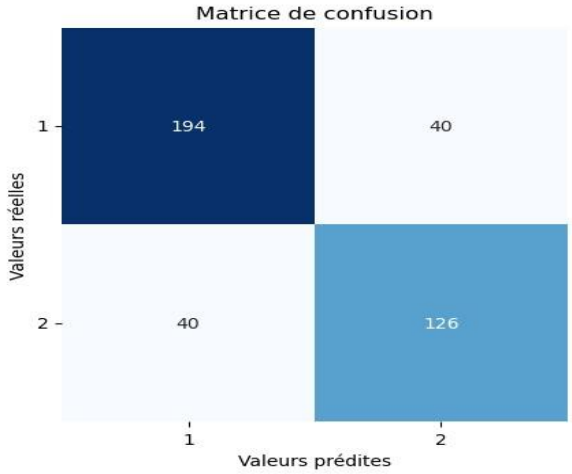
Meilleurs paramètres : {'max_depth': 10, 'max_samples': 0.2, 'n_estimators': 100}

Nos scores

Accuracy: 0.8

Rapport de classification:

	precision	recall	f1-score	support
1	0.83	0.83	0.83	234
2	0.76	0.76	0.76	166
accuracy			0.80	400
macro avg	0.79	0.79	0.79	400
weighted avg	0.80	0.80	0.80	400



Interprétation :

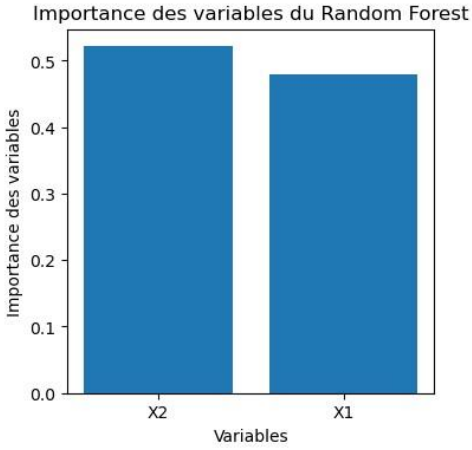
Notre modèle a une précision globale de 80%, il surpasse ainsi les modèles précédents en termes de performance.

Le recall, qui mesure la capacité du modèle à identifier correctement les instances d'une classe est également très bon (83% pour 1 et de 76% pour 2).

On a aussi de bons scores en général. (F1-score : 83% et 76%)

Cependant, d’autres réglages des hyperparamètres pourraient changer ces résultats.

L'importance des variables :



Les deux variables ont quasiment le même niveau d'importance.
On aurait eu des résultats plus intéressants si on avait un bon nombre de variables.

- **Gradient Boosting :** Le Gradient Boosting est une méthode d'apprentissage en ensemble qui construit séquentiellement des modèles, généralement des arbres de décision, en se concentrant successivement sur les erreurs des modèles précédents pour améliorer la précision des prédictions.

Nos hyperparamètres :

On a 100 arbres, de profondeur maximale 3 et 0,1 en learning_rate.

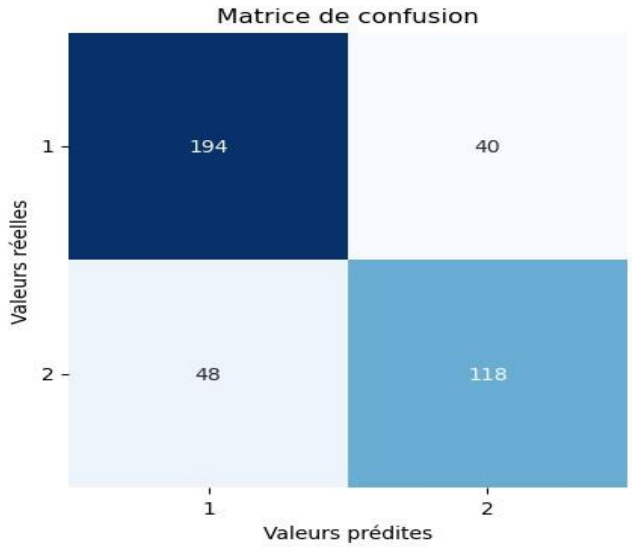
Meilleurs paramètres {'learning_rate': 0.1, 'max_depth': 3, 'n_e stimators': 100}

Nos scores

Accuracy: 0.78

Rapport de classification:

	precision	recall	f1-score	support
1	0.80	0.83	0.82	234
2	0.75	0.71	0.73	166
accuracy			0.78	400
macro avg	0.77	0.77	0.77	400
weighted avg	0.78	0.78	0.78	400



Interprétation

Le modèle a correctement classé environ 78% de l'échantillon. C'est un bon score mais moins performant que le Random Forest et légèrement moins que le Bagging.
Le recall est de 83% pour la classe 1 et de 71% pour la classe 2 : ce qui est correct mais moins bon que ceux du Random Forest
En comparaison, le modèle Random Forest semblait avoir de meilleures performances. Cependant, il est important de considérer le choix des hyperparamètres qui peut jouer sur la performance.

Le modèle du **Random Forest** semble être le plus approprié, compte tenu de tout ce qui a été dit plus haut.

1.3 Nouvelles prédictions (Question 2)

- Une fois le modèle choisi, on établit des prédictions de nouvelles variables. On stocke ce vecteur prédiction dans un fichier texte.

- **Les nouvelles données**

	X1	X2
0	1.533001	5.794728
1	5.423978	-1.849424
2	-5.107845	-1.687702
3	0.862346	-2.969889
4	-0.807342	8.193683

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X1      1000 non-null   float64
 1   X2      1000 non-null   float64
dtypes: float64(2)
memory usage: 15.8 KB
X1    0
X2    0
dtype: int64
```

	X1	X2
count	1000.000000	1000.000000
mean	0.059809	-0.143525
std	3.133790	2.947674
min	-10.402470	-10.206071
25%	-2.098294	-2.132507
50%	0.020087	-0.263866
75%	2.182233	1.861905
max	11.543303	9.202502

Les prédictions

	2
0	2
1	1
2	2
3	2
4	1
...	...
994	2
995	1
996	1
997	2
998	1

999 rows × 1 columns

2) Analyse en composantes principales (ACP)

2.1 Définitions

L'ACP consiste en une réduction de dimensionnalité. C'est une méthode statistique descriptive permettant de résumer le maximum de l'information contenue dans un tableau de données constitué de n individus et p variables quantitatives. Après standardisation des variables pour les rendre comparables, l'ACP a été appliquée pour extraire les axes principaux qui expliquent la majeure partie de la variance.

L'ACP s'effectue en plusieurs étapes :

1. **Standardisation des données** : Il est courant de commencer par standardiser les données pour chaque variable afin qu'elles aient une moyenne de 0 et un écart type de 1.
2. **Calcul des valeurs propres et des vecteurs propres** : Ces valeurs propres et vecteurs propres sont obtenus à partir de la matrice de covariance. Les vecteurs propres indiquent les directions des axes principaux, tandis que les valeurs propres indiquent leur magnitude (c'est-à-dire la quantité de variance capturée).
3. **Choix des composantes principales** : Les composantes principales sont sélectionnées en ordonnant les vecteurs propres selon leurs valeurs propres correspondantes, en décroissant. Généralement, on sélectionne les premières composantes qui expliquent une proportion suffisante de la variance (par exemple, 95%).
4. **Transformation des données** : Les données originales sont transformées en un nouveau jeu de données basé sur les composantes principales sélectionnées.
5. **Interprétation des graphiques**

DATASET: "Voitures"

	Modele	CYL	PUIS	LON	LAR	POIDS	VITESSE	ACCEL	CO2
0	ALPHAMITO	875	105	406	172	1130	184	11.4	98
1	AUDIA1	999	95	397	174	1065	186	10.9	103
2	CITROENC4	1199	130	442	182	1280	196	10.1	115
3	JAGUARF	2995	340	447	192	1587	260	5.7	234
4	PEUGEOTRCZ	1997	160	428	184	1370	220	8.2	130
5	LANDROVER	2993	256	483	191	2570	180	9.3	203
6	RENAULTCLIO	898	90	406	173	1092	182	12.2	105
7	BMWS3	1995	116	462	181	1570	198	11.1	109
8	DACIA	898	90	406	173	962	175	11.1	116
9	HYUNDAI	1995	136	447	185	1751	184	10.9	139
10	LANCIA	2776	177	522	200	2315	193	11.5	207
11	RENAULTCAPTUR	898	90	412	178	1180	171	13.0	113
12	FORDMUSTANG	4951	421	272	192	1720	250	4.8	299
13	FIAT500	1242	69	355	163	905	160	12.9	115
14	HONDA	2199	150	472	184	1632	212	9.4	138
15	FERRARI	6262	660	491	195	1880	335	4.1	380
16	SUBARU	1998	147	445	178	1440	198	9.3	141
17	MAZDA	1560	115	458	175	1490	180	13.7	138
18	VOLKSWAGEN	1598	105	425	179	1220	192	10.7	99
19	JAGUARPACE	1999	180	473	194	1775	208	8.7	139

Avec :

- **CYL** (Nombre de cylindres)
- **PUIS** (Puissance)
- **LON** (Longueur), **LAR** (Largeur), **POIDS** (Poids)
- **VITESSE** (Vitesse maximale) et **ACCEL** (Accélération)
- **CO2** (Émissions de dioxyde de carbone)

2.2 Pourcentages d'inertie expliquée par les trois premiers facteurs ? Par le premier plan factoriel ? (Question 2)

En ACP, l'inertie d'un facteur (ou composante principale) correspond à la part de la variance totale des données qu'il explique. Plus une composante principale a une inertie élevée, plus elle est significative pour expliquer la variabilité des données.

Pour chaque composante principale, on divise sa valeur propre par la somme totale des valeurs propres. On Multiplie le résultat par 100 pour obtenir le pourcentage d'inertie expliqué par cette composante.

$$\text{Pourcentage d'Inertie des 3 premiers facteurs} = \left(\frac{\lambda 1+\lambda 2+\lambda 3}{\sum_{i=1}^{n}\lambda i}\right) \times 100$$

Où :

- $\lambda 1 + \lambda 2 + \lambda 3$ sont les valeurs propres (variances expliquées) des trois premières composantes principales.
- $\sum_{i=1}^{n}\lambda i$ est la somme des valeurs propres de toutes les composantes (la variance totale des données).
- *n* est le nombre total de composantes principales.

Ainsi, le pourcentage d'inertie expliquée par les trois premiers facteurs est de 94.75%

Pour calculer le pourcentage d'inertie expliqué par le premier plan factoriel dans une Analyse en Composantes Principales (ACP), on se concentre uniquement sur les deux premières composantes principales. Le premier plan factoriel est en effet représenté par ces deux composantes, qui capturent ensemble la plus grande partie de la variance (ou inertie) des données parmi toutes les combinaisons possibles de deux axes.

$$\text{Pourcentage d'Inertie des 3 premiers facteurs} = \left(\frac{\lambda 1+\lambda 2}{\sum_{i=1}^{n}\lambda i}\right) \times 100$$

Ainsi, le pourcentage d'inertie expliquée par le premier plan factoriel est de 88.74%

Qualité de représentation des variables et contribution à la formation des axes

Lors de l'analyse en composantes principales (ACP), il est important de mesurer la qualité de représentation des variables sur les axes principaux et leur contribution à la formation de ces axes.

Qualité de représentation des variables (cosinus carré)

La qualité de représentation des variables est souvent mesurée à l'aide du carré du cosinus de l'angle entre les vecteurs représentant les variables et les axes principaux. Le cosinus carré est également appelé **cos2**. Pour une variable donnée, la somme des cos2 sur tous les axes principaux est égale à 1.

La formule pour calculer le cos2 pour une variable i sur un axe principal k est la suivante :

$$\cos^2(i,k) = \frac{\text{coord}(i,k)^2}{\|\text{coord}(i)\|^2}$$

où coord(i,k) est la coordonnée de la variable i sur l'axe principal k.

Un seuil couramment utilisé pour interpréter la qualité de représentation des variables est 0,5. Si le cos2 d'une variable est supérieur à 0,5 sur un axe principal, cela signifie que cet axe représente bien la variable.

Contribution des variables à la formation des axes (CTR)

La contribution des variables à la formation des axes indique l'importance relative de chaque variable pour expliquer la variabilité des données sur un axe principal donné. Cette contribution est souvent exprimée en pourcentage et est calculée à l'aide de la formule suivante :

$$\text{CTR}(i,k) = \frac{\text{coord}(i,k)^2}{n \cdot \lambda_k}$$

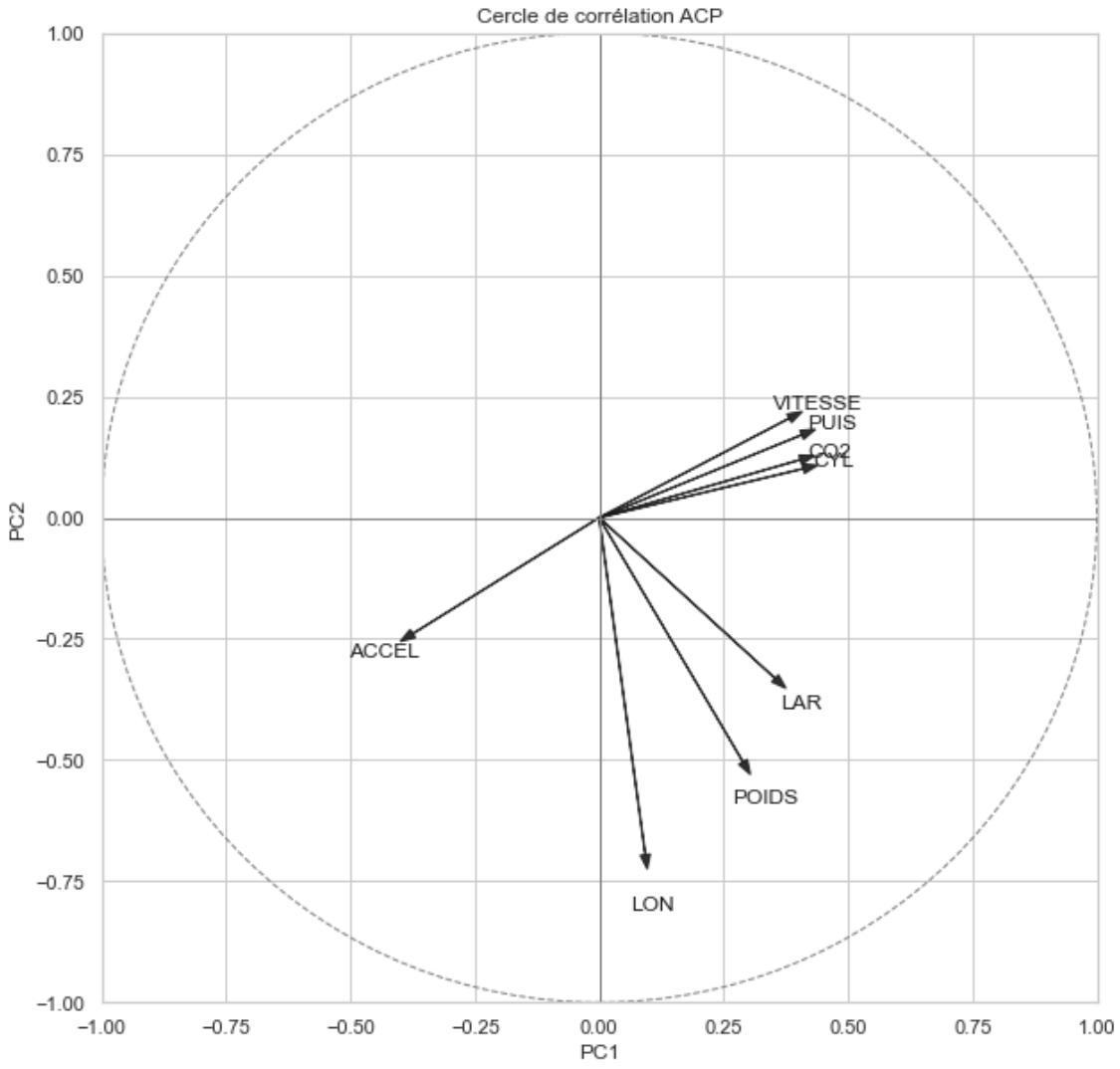
où :

- coord(i,k) est la coordonnée de la variable i sur l'axe principal k
- n est le nombre total de variables
- λk est la valeur propre associée à l'axe principal k

Il est important de noter que la somme des contributions des variables sur un axe principal donné est égale à 100 %.

En interprétant les contributions des variables à la formation des axes, **il est courant de considérer les variables ayant une contribution supérieure à la contribution moyenne (100 % / nombre de variables)** comme étant importantes pour la formation de l'axe principal considéré.

2.3 Interpréter les 2 axes principaux à partir des corrélations des variables avec ces axes. (Question 3)

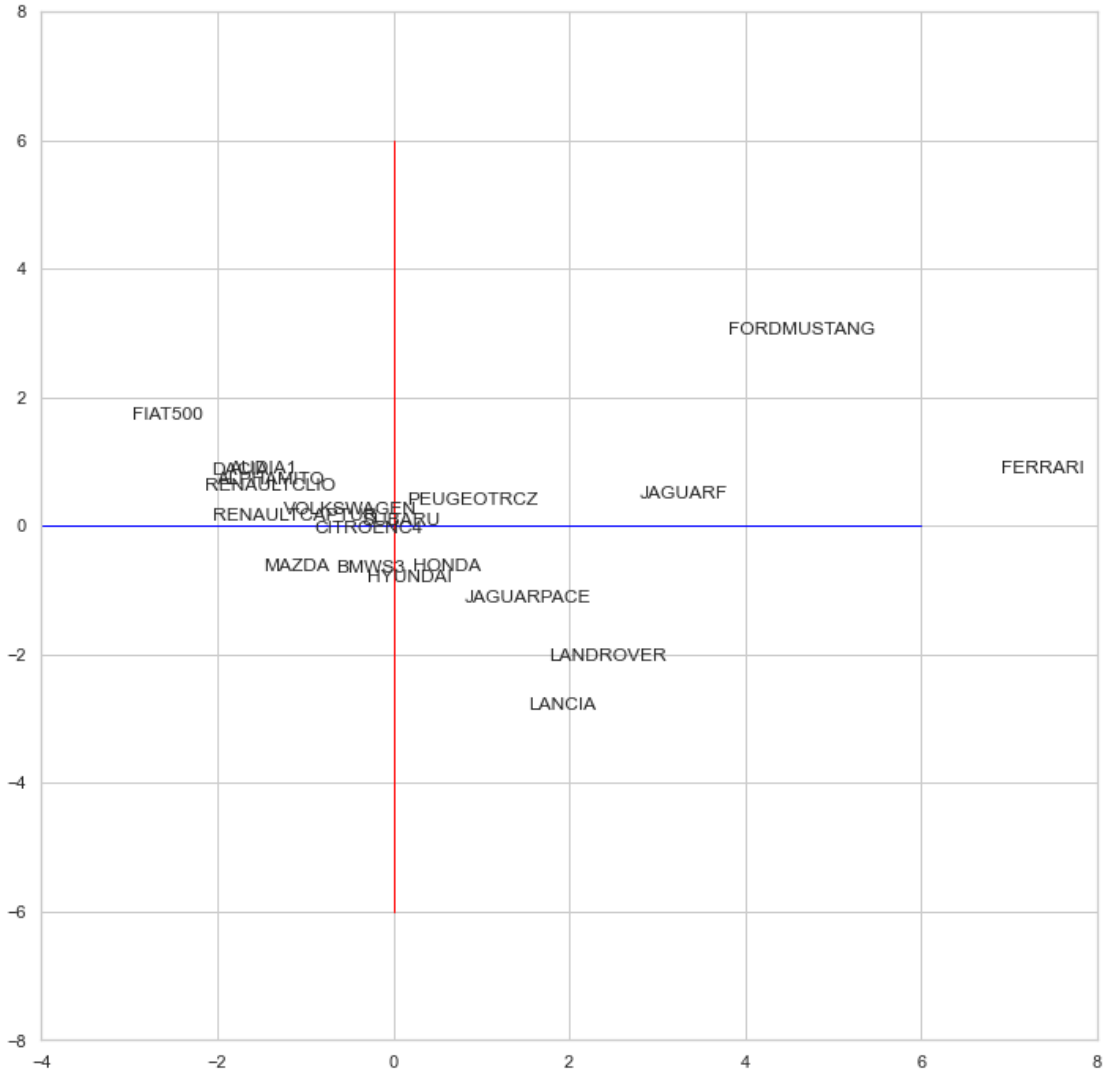


Interprétation :

Premier Axe Principal : Le PC1 met en lumière **un effet de taille** notable, caractérisé par une corrélation positive avec des variables telles que CYL, PUIS, VITESSE et CO2. Les voitures qui se distinguent par des caractéristiques de performance élevées—moteurs volumineux, puissance considérable, vitesse supérieure—et qui sont également associées à des émissions de CO2 plus importantes sont proéminentes sur cet axe. Cela suggère que le PC1 capture un profil de "haute performance environnementale", où les véhicules à forte cylindrée et puissants définissent un segment distinct au sein du jeu de données.

Deuxième Axe Principal : Le PC2 révèle **un effet de forme**, comme en témoignent les corrélations négatives avec des variables telles que LON, LAR et POIDS. Cet axe semble distinguer les véhicules plus grands et plus lourds, indiquant que des dimensions plus étendues et un poids accru caractérisent les véhicules situés négativement sur cet axe. Il est intéressant de noter que l'accélération, bien que modérément corrélée, participe également à cet effet de forme, séparant potentiellement les véhicules performants en termes d'accélération de ceux qui privilégient l'espace et le confort.

2.4 Représentez les individus sur le premier plan factoriel (Question 3)



Interprétation :

- Les voitures situées à droite ont tendance à avoir des valeurs élevées pour les variables corrélées positivement avec le premier axe (performance et impact environnemental).
- Les voitures situées en haut ou en bas ont des caractéristiques distinctes en termes de taille et de poids, comme indiqué par leurs corrélations avec le deuxième axe.

Représentation sur le Premier Plan Factoriel :

L'analyse indique que les individus (ici, les modèles de voitures) sont généralement bien représentés sur le premier plan factoriel. Une proportion substantielle de l'inertie (88.74%) est capturée par les deux premiers axes, ce qui suggère une bonne qualité de représentation.

Caractéristiques des Individus en Haut du Graphique :

Les modèles de voitures situés en haut du graphique sont probablement caractérisés par des spécificités ou des attributs qui ont une forte corrélation positive avec le deuxième axe. Cela pourrait inclure des éléments tels que le luxe, la sportivité, ou d'autres caractéristiques non spécifiées qui sont mesurées par cet axe.

Caractéristiques des Individus à Droite du Graphique :

Les voitures positionnées à droite du graphique tendent à avoir des scores élevés sur les variables corrélées positivement avec le premier axe, ce qui pourrait correspondre à des caractéristiques telles que la puissance, la vitesse, et potentiellement une plus grande émission de CO2.

Caractéristiques des Individus en Bas à Gauche du Graphique :

Les voitures en bas à gauche sont associées à des valeurs faibles sur les axes principaux, suggérant qu'elles pourraient être plus petites, moins puissantes, moins rapides et éventuellement plus écologiques.

Comparaison entre PEUGEOTRCZ et JAGUARF :

Bien que la PEUGEOTRCZ et la JAGUARF se trouvent toutes deux sur le côté positif du premier axe, la JAGUARF est significativement plus éloignée, indiquant des différences en termes de performance ou d'impact environnemental. Elles partagent des caractéristiques mais à des degrés différents.

Comparaison entre LANDROVER et LANCIA :

LANDROVER et LANCIA se situent tous deux dans le quadrant négatif du deuxième axe et positif du premier axe, ce qui pourrait signifier qu'ils partagent des attributs de performance élevée mais sont également plus grands et lourds, avec une possible diminution de l'agilité.

3) Classification

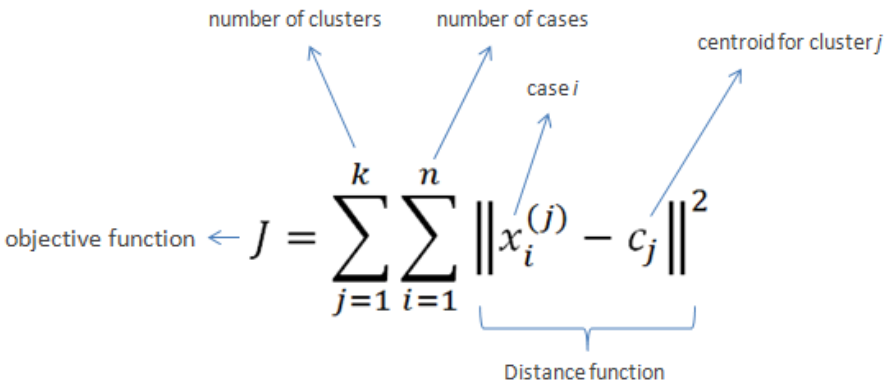
3.1 Définitions

La méthode des k-means est une technique d'analyse de clustering (regroupement) utilisée en apprentissage automatique et en statistique. Elle a pour objectif de diviser un ensemble de données en k groupes (ou clusters) distincts. Voici une explication détaillée de cette méthode, ainsi que son application dans le cas spécifique de données sur les voitures.

La méthode des K-means se procède en 4 étapes principales :

1. **Initialisation** : Choisissez k points comme centres de clusters initiaux. Ces points peuvent être choisis aléatoirement ou selon une certaine stratégie.
2. **Attribution à un cluster** : Pour chaque point de données, trouvez le centre de cluster le plus proche (en mesurant la distance, généralement la distance euclidienne) et attribuez le point à ce cluster.
3. **Mise à jour des centres de cluster** : Recalculez les centres de chaque cluster comme étant le barycentre (moyenne) de tous les points attribués à ce cluster.
4. **Répétition** : Répétez les étapes 2 et 3 jusqu'à ce que les centres des clusters ne changent plus ou que le changement soit inférieur à un seuil prédéfini.

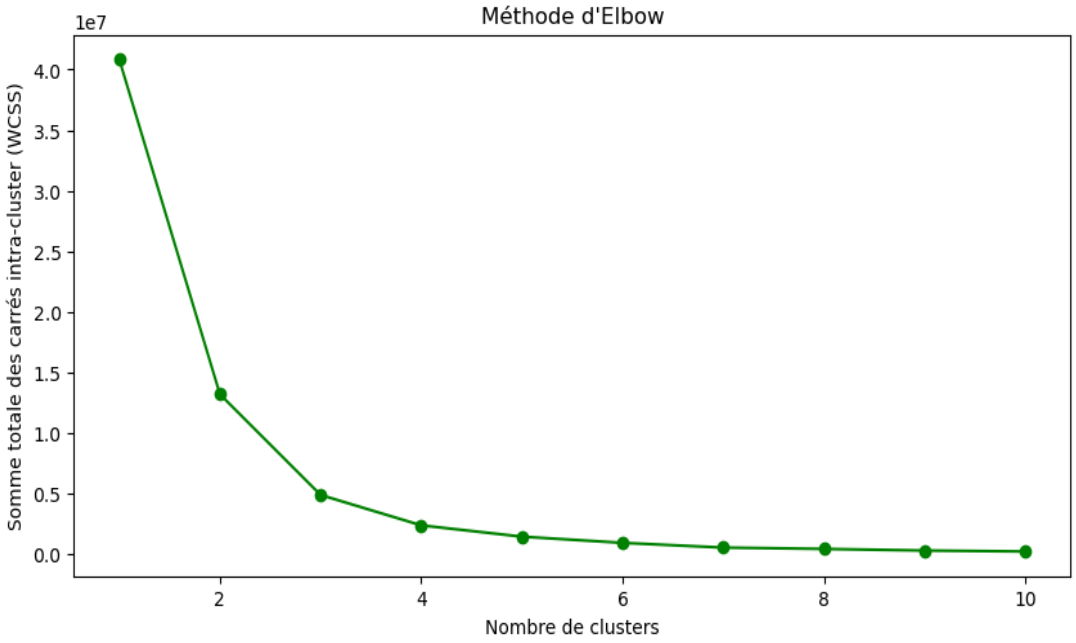
La formule pour calculer la distance euclidienne entre deux points *x* et *y* dans un espace à n dimensions est la suivante :



Dataset : "Voitures" (Cf page 7)

3.2 Question 1 : Clustering par K-Means

Pour la première question de la partie 3 sur la classification, nous avons effectué un clustering k-means sur un ensemble de données de 20 voitures. Nous avons d'abord nettoyé les données en supprimant une colonne non pertinente, puis utilisé la méthode du coude pour déterminer le nombre optimal de clusters.



Nous avons ainsi choisi de regrouper les données en trois clusters. Nous avons attribué à chaque voiture un label de cluster et calculé les moyennes de différentes caractéristiques automobiles pour chaque groupe, ce qui a permis d'identifier et de décrire les profils distincts de ces classes.

cluster	CYL	PUIS	LON	LAR	POIDS	VITESSE	ACCEL	CO2
0	2327.444444	184.666667	464.333333	187.666667	1778.888889	205.888889	9.344444	160.000000
1	5606.500000	540.500000	381.500000	193.500000	1800.000000	292.500000	4.450000	339.500000
2	1129.666667	98.777778	411.888889	174.333333	1147.111111	180.666667	11.777778	111.333333

La

La répartition des voitures dans les clusters se présente de cette façon :

Unnamed: 0	ALPHAMITO	AUDIA1	CITROENC4	JAGUARF	PEUGEOTRCZ	LANDROVER	RENAULTCLIO	BMWS3	DACIA
cluster	2	2	2	0	0	0	2	0	2

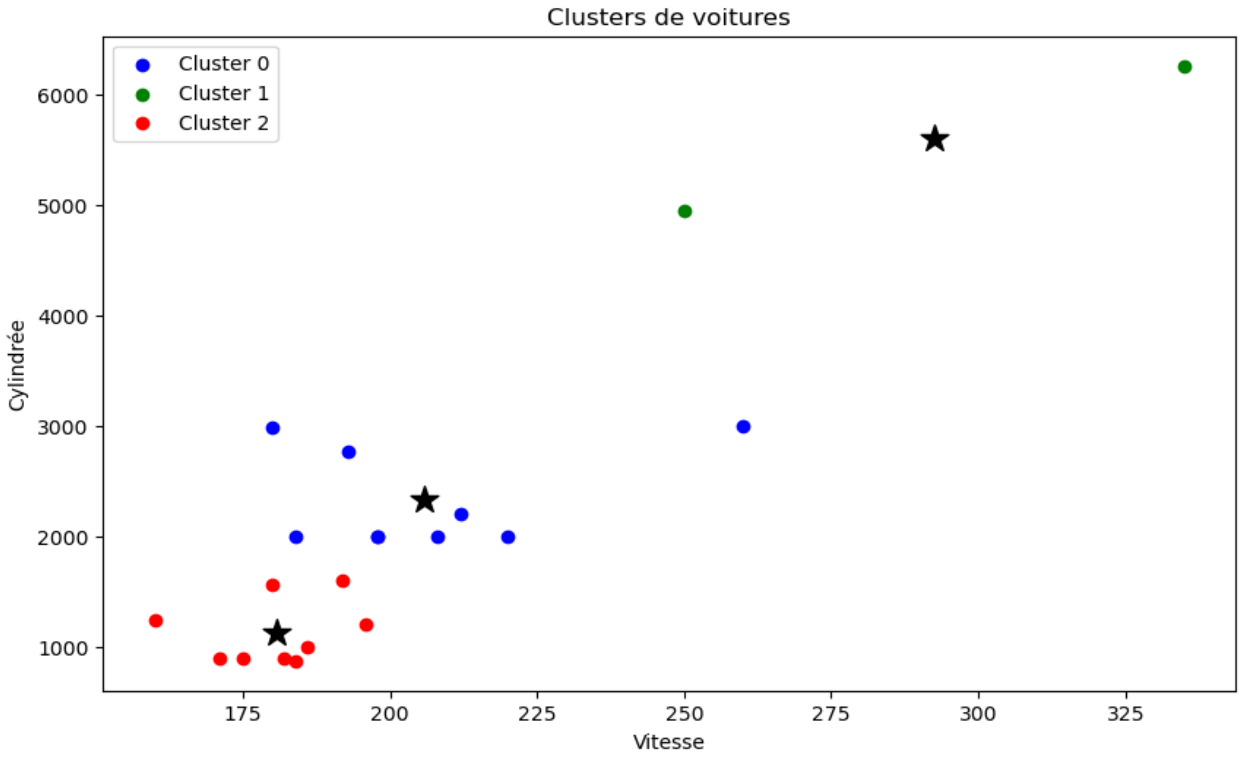
DACIA	HYUNDAI	LANCIA	RENAULTCAPTUR	FORDMUSTANG	FIAT500	HONDA	FERRARI	SUBARU	MAZDA	VOLKSWAGEN
2	0	0	2	1	2	0	1	0	2	2

JAGUARPACE
0

Ces caractéristiques nous ont permis d'identifier trois différentes classes de véhicules :

- Voitures économiques/citadines
- Berlines/SUVs
- Voitures de sport de luxes

Visualisation des clusters prédits par le K-Means clustering



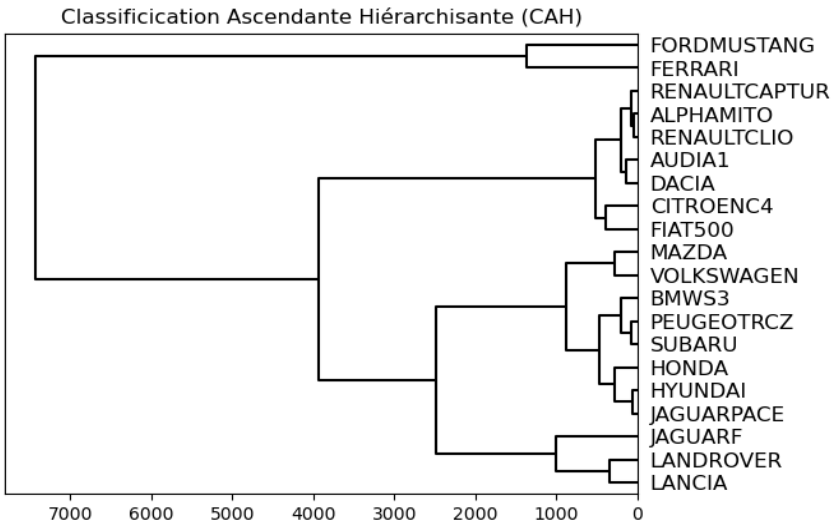
Interprétation des clusters :

On distingue 3 classes de voitures via les clusters :

- Cluster 2 : Ce cluster comprend des véhicules avec une faible cylindrée et puissance d'où les faibles émissions de CO2 et une vitesse maximale assez faible d'où une accélération plus lente. Ce type de véhicules se rapproche généralement des voitures "citadines" ou "économiques" qui sont généralement conçues pour les déplacements en ville car économiques en termes de consommation de carburant. Leur assez faible vitesse maximale est également appropriée pour des voyages sur faible distance.
- Cluster 1 : Il est caractérisé par des voitures disposant d'une cylindrée et d'une puissance et d'une vitesse maximale très élevées. Ces voitures ont une accélération très rapide et des émissions de CO2 élevées. Les voitures de ce cluster sont clairement des voitures de sport de luxe, convenables pour des trajets de très longues distances ou sur des autoroutes.
- Cluster 0 : Il regroupe des voitures avec une cylindrée et une puissance modérée, une bonne vitesse maximale et une accélération équilibrée. Ces véhicules ont des émissions de CO2 modérées. Ce type de véhicules s'apparente aux berlines ou au SUVs qui sont conçues pour être polyvalentes, adaptées à la fois à la conduite en ville et aux voyages sur de longues distances. Les caractéristiques de ce type de voitures sont intermédiaires entre celles du cluster 2 et celles du cluster 1, d'où le fait qu'elles soient relativement équilibrées.

3.3 Question 2 : Classification Ascendante Hiérarchique avec la Méthode de Ward

Pour la deuxième question, nous avons procédé à une classification hiérarchique en utilisant la méthode de Ward. Nous avons d'abord créé un dendrogramme pour visualiser la manière dont les voitures ont été regroupées en clusters à différents niveaux de similarité.



Nous avons analysé le dendrogramme pour déterminer en combien de classes il serait idéal de diviser les données, en supposant initialement trois clusters comme pour la question précédente. Finalement, nous avons plutôt opté pour un ajustement de quatre clusters.

Interprétation du dendrogramme :

Via ce dendrogramme, on retrouve d'abord en haut :

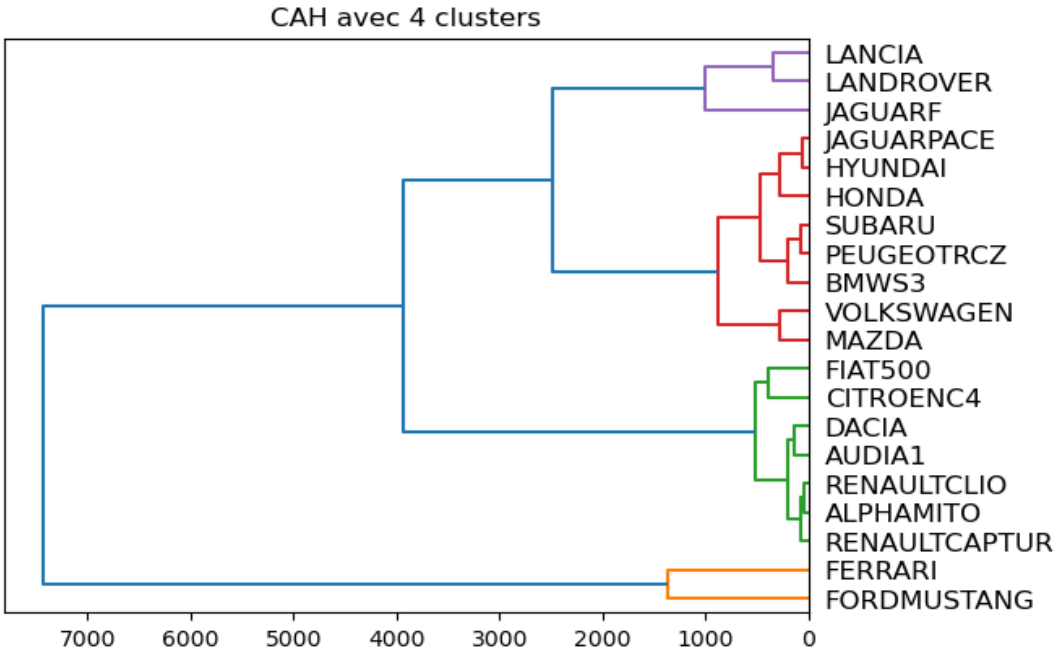
- Les voitures FORDMUSTANG ET FERRARI qui selon les interprétations précédentes semblent appartenir à la classe des voitures de sport de luxe. Elles se démarquent par une cylindrée, une puissance et une vitesse très élevées. Elles émettent une grande quantité de CO2. Elles conviennent pour des voyages de long trajet.
- On distingue un autre groupe de véhicules allant de RENAULTCAPTUR à FIAT500, une variété de voitures dites citadines ou économiques, caractérisées par leurs faibles émissions de CO2, leurs cylindrées, puissances et vitesses peu fortes. Elles sont adaptées pour des déplacements à courte distance.
- Le troisième groupe comptant les voitures MAZDA à JAGUARPACE est quant à lui composé de berlines/SUV moyennes, appropriées pour des voyages à court ou long trajet. Elles ont une vitesse abordable. Elles sont puissantes et possèdent de fortes cylindrées.
- Le quatrième et dernier groupe est composé de trois voitures : JAGUARF, LANDROVER et LANCIA. Ces véhicules sont des berlines/SUV de haute gamme. Elles possèdent également les caractéristiques des berlines et SUV moyennes mais en meilleures. En effet, ces trois voitures sont beaucoup plus rapides que les moyennes et disposent aussi d'une puissance et d'une vitesse maximale supérieure pouvant avoisiner quelques voitures de luxe. Par exemple, la JAGUARF a une vitesse maximale supérieure à celle de la FORDMUSTANG qui est elle une voiture de sport de luxe. Ce type de véhicules est tout aussi adapté aux trajets longs et courts.

Globalement, on remarque que le cluster des berlines/SUV dans l'exercice 1 s'est en quelque sorte subdivisé en deux pour marquer la distinction entre les moyennes et les hauts de gamme. Il existe également deux voitures qui sont passées de voitures économiques à berlines/SUV "moyennes".

En résumé, on aurait donc envie de couper ce dendrogramme en 4 classes différentes :

- Les voitures économiques
- Les berlines/SUV moyennes
- Les berlines/SUV haut de gamme

- Les voitures de sport de luxe



4) Analyse des correspondances multiples

4.1 Définitions

L'analyse des correspondances (AC) est une méthode statistique multivariée qui permet d'analyser et de visualiser des tableaux de contingence. C'est une technique d'exploration de données qui sert à détecter des structures dans des données catégorielles (qualitatives), souvent sous forme de tableaux croisant deux ou plusieurs variables qualitatives (par exemple, les réponses de personnes à différents items d'un questionnaire).

Dataframe : Chiens

```
> df_chiens
          TAI POI VEL INT AFF AGR FON
beaucero    3   2   3   2   2   2   3
basset      1   1   1   1   1   2   2
bergeral    3   2   3   3   2   2   3
boxer       2   2   2   2   2   2   1
buldog      1   1   1   2   2   1   1
bullmast    3   3   1   3   1   2   3
caniche     1   1   2   3   2   1   1
chihuahu    1   1   1   1   2   1   1
cocker      2   1   1   2   2   2   1
colley      3   2   3   2   2   1   1
dalmatie    2   2   2   2   2   1   1
doberman    3   2   3   3   1   2   3
dogueall    3   3   3   1   1   2   3
epagneub    2   2   2   3   2   1   2
epagneuf    3   2   2   2   1   1   2
foxhound    3   2   3   1   1   2   2
foxterri    1   1   2   2   2   2   1
gbdeasco    3   2   2   1   1   2   2
labrador    2   2   2   2   2   1   2
levrier     3   2   3   1   1   1   2
mastiff     3   3   1   1   1   2   3
pekinois    1   1   1   1   2   1   1
pointer     3   2   3   3   1   1   2
saintber    3   3   1   2   1   2   3
setter      3   2   3   2   1   1   2
teckel      1   1   1   2   2   1   1
terreneu    3   3   1   2   1   1   3
```

On réalise l'ACM grâce à la fonction **MCA() du package FactoMineR.**

```
> resultats <- MCA(df_chiens, quali.sup=7, graph=FALSE)
> print(resultats)
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 27 individuals, described by 7 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. of the categories"
4  "$var$cos2"          "cos2 for the categories"
5  "$var$contrib"       "contributions of the categories"
6  "$var$v.test"        "v-test for the categories"
7  "$var$eta2"          "coord. of variables"
8  "$ind"               "results for the individuals"
9  "$ind$coord"         "coord. for the individuals"
10 "$ind$cos2"          "cos2 for the individuals"
11 "$ind$contrib"       "contributions of the individuals"
12 "$quali.sup"         "results for the supplementary categorical variables"
13 "$quali.sup$coord"   "coord. for the supplementary categories"
14 "$quali.sup$cos2"    "cos2 for the supplementary categories"
15 "$quali.sup$v.test"  "v-test for the supplementary categories"
16 "$call"              "intermediate results"
17 "$call$marge.col"    "weights of columns"
18 "$call$marge.li"     "weights of rows"
```

Nous allons nous intéresser principalement aux objets 1 (valeurs propres), 3 (coordonnées des modalités), 5 (contributions des modalités), 8 (coordonnées des individus) et 12 (coordonnées des modalités qualitatives supplémentaires).

Combien de facteurs faut-il retenir ?

Dans notre exemple, on va donc obtenir au maximum 10 facteurs (16 modalités - 6 variables principales), classés du plus important au moins important (ou du plus explicatif au moins explicatif) et à chacun de ces 10 facteurs va être associée une « valeur propre » qui traduit la quantité d'information expliquée par le facteur.

On obtient :

```
> print(resultats$eig)
       eigenvalue percentage of variance cumulative percentage of variance
dim 1  0.481606165              28.896370                           28.89637
dim 2  0.384737288              23.084237                           51.98061
dim 3  0.210954049              12.657243                           64.63785
dim 4  0.157554025               9.453242                           74.09109
dim 5  0.150132670               9.007960                           83.09905
dim 6  0.123295308               7.397718                           90.49677
dim 7  0.081462460               4.887748                           95.38452
dim 8  0.045669757               2.740185                           98.12470
dim 9  0.023541911               1.412515                           99.53722
dim 10 0.007713034               0.462782                          100.00000
```

L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.

Les 2 premiers axes de l'analyse expriment **51.98%** de l'inertie totale du jeu de données ; cela signifie que 51.98% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important, et le premier plan représente donc convenablement la variabilité contenue dans une grande part du jeu de données actif.

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 2 premiers axes. Ces composantes révèlent un taux d'inertie supérieur à celle du quantile 0.95-quantile de distributions aléatoires (51.98% contre 41.17%). Cette observation suggère que seuls ces axes sont porteurs d'une véritable information. En conséquence, la description de l'analyse sera restreinte à ces seuls axes.

Du fait de ces observations, il serait tout de même probablement préférable de considérer également dans l'analyse les dimensions supérieures ou égales à la troisième.

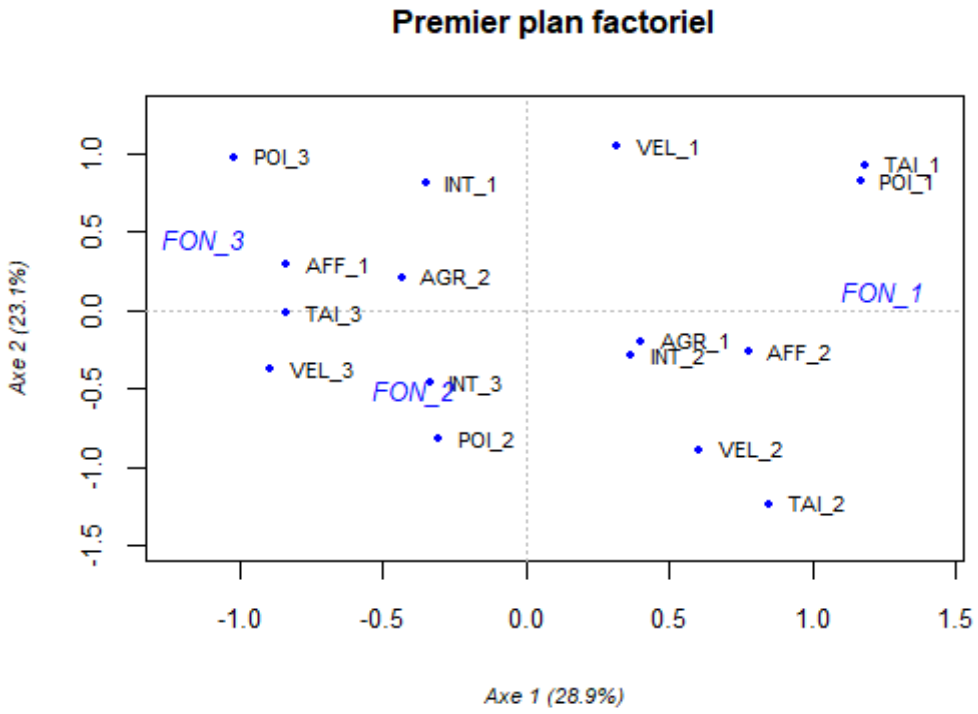
Nous avons les coordonnées des différentes modalités selon les axes 1 et 2 et les cos2 de chaque modalité selon les axes 1 et 2 :

```
> print(var_coord)
           Dim 1       Dim 2
TAI_1  1.1849557  0.92389650
TAI_2  0.8510880 -1.23171972
TAI_3 -0.8366753 -0.02057846
POI_1  1.1689180  0.82434462
POI_2 -0.3054053 -0.81887572
POI_3 -1.0151341  0.97390062
VEL_1  0.3199406  1.04490006
VEL_2  0.6036867 -0.88781355
VEL_3 -0.8920999 -0.37183247
INT_1 -0.3490450  0.80855486
INT_2  0.3694426 -0.28550314
INT_3 -0.3350656 -0.45948302
AFF_1 -0.8351500  0.28746968
AFF_2  0.7754964 -0.26693613
AGR_1  0.4007145 -0.19425299
AGR_2 -0.4315386  0.20919553
```

```
> print(var_cos2)
           Dim 1        Dim 2
TAI_1 0.49144201 0.2987546600
TAI_2 0.16462520 0.3448030588
TAI_3 0.87503205 0.0005293413
POI_1 0.57531341 0.2861238116
POI_2 0.10044717 0.7221387844
POI_3 0.23420393 0.2155641859
VEL_1 0.06021292 0.6422447857
VEL_2 0.15344741 0.3318791146
VEL_3 0.39792110 0.0691296921
INT_1 0.05129787 0.2752677726
INT_2 0.12673870 0.0756897524
INT_3 0.03207684 0.0603213262
AFF_1 0.64765585 0.0767360421
AFF_2 0.64765585 0.0767360421
AGR_1 0.17292377 0.0406368567
AGR_2 0.17292377 0.0406368567
```

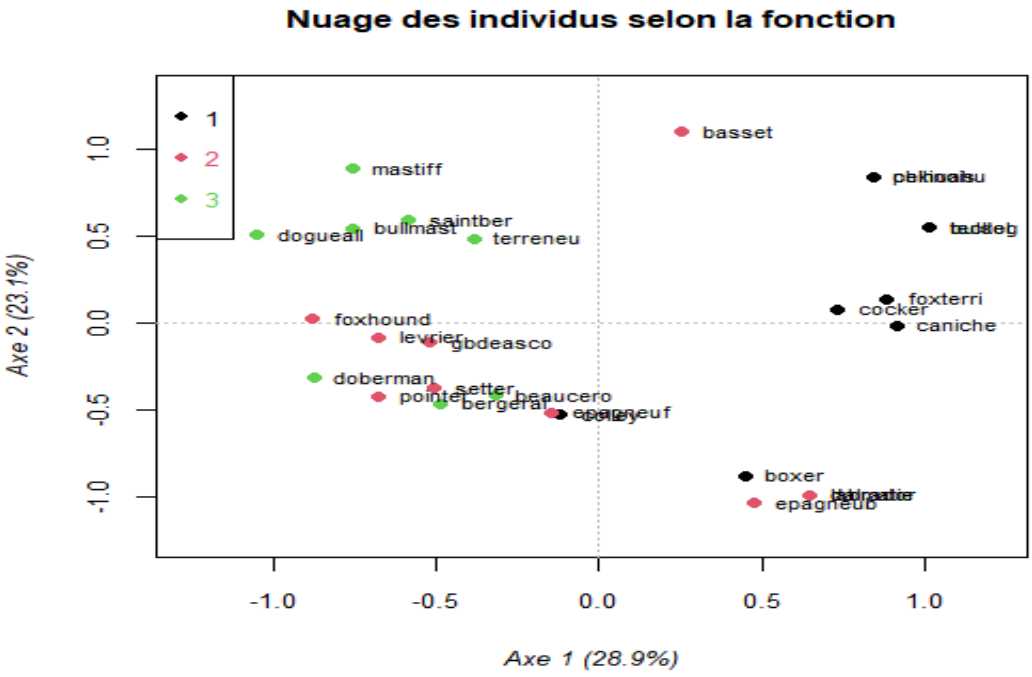
On décide à présent de représenter la corrélation entre modalités et axes. Nous choisissons de représenter les points et les étiquettes de toutes les modalités, qu'elles soient actives (suffisamment représentées) ou pas.

- **GRAPHIQUE DES MODALITÉS :**



- **NUAGE DE POINTS DES INDIVIDUS :**

Il est souvent très instructif d'observer le nuage des individus :



4.2 Question 1 : En prenant la variable FON comme variable supplémentaire, faire une analyse des correspondances multiples de ces données.

- **Axe 1 (Horizontal)** : Il représente la première dimension factorielle et explique 28.9% de l'inertie dans les données. C'est l'axe le plus informatif de l'ACM.

Les modalités "TAI_1" (petite taille) et "POI_1" (léger poids), situées à l'extrême droite du graphique sont positivement corrélées avec l'axe 1.
Une petite taille et un poids léger sont généralement corrélés entre eux et qu'elles distinguent un groupe spécifique d'individus sur cet axe.

La modalité AFF_2 (grande affectuosité) semble être positivement corrélée à l'axe 1 tandis que AFF_1 (faible affectuosité) semble y être négativement corrélée.
À l'opposé par rapport à l'axe, les modalités "TAI_3" (grande taille) et "VEL_3" (rapide vélocité) sont négativement corrélées à l'axe 1 et sont situées à l'extrême gauche. Cela indique une éventuelle association entre grande taille et rapidité.

Les modalités "FON_3" (fonction de garde) et "FON_2" (fonction de chasse) se trouvent également du côté gauche, ce qui pourrait signifier que ces fonctions peuvent être liées à des chiens de plus grande taille et plus rapides. Les chiens de chasse ("FON_3") ont généralement une faible affectuosité ("AFF_1").

- **Axe 2 (Vertical)** : Il représente quant à lui la deuxième dimension factorielle, et explique 22.8% de l'inertie.

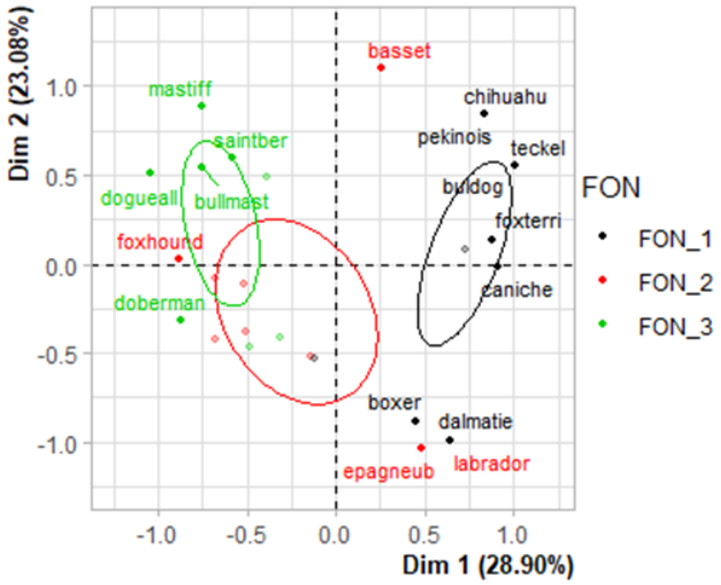
Les modalités INT_1(faible intelligence) et VEL_1 (lent), situées en haut, sont corrélées positivement à l'axe 2. Une faible intelligence et une faible vitesse sont corrélés entre elles.
Tandis qu'en bas, nous avons POI_2(poids moyen) et VEL_2(vitesse moyenne) et TAI_2(taille moyenne) qui sont corrélés négativement à l'axe 2. Cela indique une éventuelle association entre taille et rapidité.
Ainsi, la **dimension 2** oppose des individus tels que *pekinois* et *basset* (en haut du graphe, caractérisés par de faibles tailles, poids et vitesse) à des individus comme *dalmatie*, *labrador*, *epagneub* (en bas du graphe, caractérisés par des taille, vitesse et poids moyens).

NB : Pour toutes les autres modalités dont la somme des cos2 sur les deux dimensions est inférieure à 0.5, nous pouvons conclure qu'elles ne sont pas suffisamment représentées par ces deux dimensions. Ainsi, leur interprétation nécessite des dimensions supplémentaires pour une meilleure analyse.

4.3 Question 2 : En déduire une description des différentes races de chiens

On peut distinguer plusieurs catégories de races :

- **Les chiens de compagnie** : de petite taille et d'un poids léger, ces chiens sont assez affectueux (comme le basset ou le chihuahua)
- **Les chiens de garde** : chiens grands et rapides comme le Berger Allemand ou le Doberman, souvent utilisés pour la garde ou le travail policier. Ils sont peu affectueux.
- **Les chiens de chasse** : chiens dont la taille varie entre moyen et grand et de poids moyen, ce sont des chiens assez rapides et intelligents.



CONCLUSION

Ce rapport a exploré les diverses facettes de l'analyse de données et du machine learning, en soulignant leur rôle indispensable dans l'extraction de connaissances à partir de vastes ensembles de données complexes. Chaque technique employée a révélé des aspects uniques et a contribué à une compréhension globale des modèles sous-jacents dans les données examinées.

Tout d'abord, la régression binaire a démontré l'efficacité des méthodes ensemblistes face à la variabilité des données. L'approche Random Forest s'est distinguée, offrant une amélioration notable en termes de précision et de résistance au surajustement, comparée aux méthodes traditionnelles de régression.

Ensuite, l'Analyse en Composantes Principales (ACP), a joué un rôle dans la réduction de la dimensionnalité. Cette méthode a permis de simplifier la complexité des données tout en préservant l'intégrité des informations. Les visualisations découlant de l'ACP ont grandement aidé à l'interprétation des orientations et des relations entre les variables, révélant ainsi des patterns significatifs.

Par ailleurs, la classification avec k-means a démontré l'efficacité des algorithmes non supervisés dans la segmentation des données en clusters naturels. Cette approche a mis en lumière des structures et des regroupements.

Enfin, l'analyse des correspondances est une méthode efficace pour explorer les relations entre les catégories de variables qualitatives. Son application a permis de découvrir des liaisons et des corrélations subtiles, enrichissant notre compréhension des interdépendances au sein des données.

Nous avons également rencontré des limitations, telles que la sensibilité de l'ACP aux échelles de mesure et la dépendance du Random Forest à la sélection d'hyperparamètres optimaux. Ces défis suggèrent des pistes pour des recherches ultérieures, notamment l'optimisation des hyperparamètres via des techniques de recherche en grille et l'expérimentation avec d'autres algorithmes de clustering pour comparer les résultats avec ceux obtenus via k-means.

Pour conclure, ce projet nous a aidé à explorer les méthodes de machine learning supervisées et non supervisées et nous a permis d’approfondir nos connaissances en l'analyse de données.