Projet P3 : Consommation de carburant des véhicules

Hana FEKI

20 avril 2025

Résumé

Ce projet étudie la consommation de carburant des véhicules à partir du jeu de données mtcars en utilisant la régression multiple et l'analyse en composantes principales (ACP). Après une analyse descriptive des variables, plusieurs modèles de régression ont été comparés, avec et sans sélection de variables. L'ajout d'une nouvelle observation a permis d'évaluer l'impact sur les coefficients. L'ACP a été utilisée pour réduire la dimension et mieux visualiser les relations entre variables et individus. Enfin, une régression sur les composantes principales a été réalisée et comparée au modèle classique.

Table des matières

Table des figures

1 Introduction

Ce projet vise à appliquer des méthodes de **statistiques inférentielles** pour analyser un jeu de données réel (`mtcars`). L'objectif est d'étudier les **relations entre la consommation de carburant** (`Miles_per_Gallon`) et diverses **variables explicatives techniques** des véhicules (puissance, poids, cylindrée, etc.).

Nous utiliserons principalement les méthodes suivantes :

— **Régression linéaire multiple**, pour modéliser la relation entre la variable cible `Miles_per_Gallon` et les caractéristiques techniques du véhicule ;

— **Analyse en composantes principales (ACP)**, pour réduire la dimensionnalité du jeu de données et identifier les structures sous-jacentes entre les variables.

Avant d'entamer l'analyse approfondie, il est essentiel de **comprendre la structure des données** et de **vérifier leur qualité**. Cette étape préliminaire permet d'identifier d'éventuelles **incohérences**, **valeurs extrêmes** ou **distributions atypiques** pouvant influencer les résultats des analyses statistiques.

2 Exploration statistique préliminaire

2.1 Acquisition des données

2.1.1 Exploration initiale du jeu de données

Lors de l'importation initiale du fichier `mtcars.csv`, nous avons d'abord vérifié la dimension du jeu de données, qui comprend 32 lignes et 12 colonnes. Afin de mieux comprendre le contenu, j'ai ensuite analysé les premières lignes à l'aide de la commande `head(df)` et la structure du dataframe avec `str(df)`.

Les résultats obtenus ont révélé que la colonne `model` était de type chaîne de caractères, tandis que les autres colonnes étaient numériques. Voici le résultat de ces premières commandes :

```
> print(head(df))
                  model  mpg cyl disp  hp drat    wt  qsec vs am gear carb
1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

FIGURE 1 – Les 5 premières lignes des données

```
> str(df)
'data.frame':   32 obs. of  12 variables:
 $ model: chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
 $ mpg  : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl  : int  6 6 4 6 8 6 8 4 4 6 ...
 $ disp : num  160 160 108 258 360 ...
 $ hp   : int  110 110 93 110 175 105 245 62 95 123 ...
 $ drat : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt   : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec : num  16.5 17 18.6 19.4 17 ...
 $ vs   : int  0 0 1 1 0 1 0 1 1 1 ...
 $ am   : int  1 1 1 0 0 0 0 0 0 0 ...
 $ gear : int  4 4 4 3 3 3 3 4 4 4 ...
 $ carb : int  4 4 1 1 2 1 4 2 2 4 ...
```

FIGURE 2 – Structure des données

2.1.2 Identification de la variable cible

À ce stade, nous avons également réfléchi à l'identification de la variable cible pour l'analyse statistique. Le but de cette analyse étant de prédire la consommation de carburant (colonne `Miles_per_Gallon`), cette variable a donc été retenue comme cible de prédiction.

2.1.3 Renommage des colonnes

Étant donné que la signification des colonnes du jeu de données n'était pas suffisamment explicite, nous avons dû effectuer des recherches pour en comprendre le sens. Ensuite, nous avons renommé les colonnes avec des intitulés plus explicites afin de faciliter l'interprétation des données. Le tableau ci-dessous récapitule les informations que nous avons trouvées.

TABLE 1 – Signification des variables

Nom initial	Nouveau nom	Signification
model	Model	Nom du modèle du véhicule.
mpg	Miles_per_Gallon	Consommation de carburant du véhicule, mesurée en miles par gallon (mpg). C'est la variable cible de l'analyse.
cyl	Cylinders	Nombre de cylindres du moteur.
disp	Displacement_cuin	Cylindrée du moteur, mesurée en pouces cubes (cuin).
hp	Horsepower	Puissance du moteur du véhicule, mesurée en chevaux-vapeur (hp).
drat	Rear_Axle_Ratio	Ratio entre la vitesse de rotation des roues arrière et celle du moteur.
wt	Weight_lb_per_1000	Poids du véhicule, exprimé en livres par 1000.
qsec	Quarter_Mile_Time	Temps qu'il faut pour parcourir un quart de mile (400 mètres).
vs	Engine_Shape	Forme du moteur du véhicule.
am	Transmission	Type de transmission du véhicule (manuelle ou automatique).
gear	Forward_Gears	Nombre de vitesses avant dans la transmission du véhicule.
carb	Carburetors	Nombre de carburateurs présents dans le véhicule.

2.1.4 Création d'une nouvelle variable numérique

Pour faciliter la manipulation des données, nous avons créé une nouvelle variable, **num**, qui exclut la variable `model`, étant donné qu'elle est de type chaîne de caractères. Cette exclusion permet de se concentrer uniquement sur les variables numériques pour l'analyse statistique, tandis que la variable `model` sera étudiée plus tard lors de l'analyse univariée.

2.1.5 Résumé statistique des variables

Afin de compléter notre compréhension initiale du jeu de données, nous avons utilisé la commande `summary(df)` pour obtenir un aperçu statistique de chaque variable. Cette commande fournit des informations sur les mesures de tendance centrale, telles que le minimum, la médiane, la moyenne, le maximum, ainsi que les quartiles.

```
> summary(num)
 Miles_per_Gallon   Cylinders     Displacement_cuin   Horsepower    Rear_Axle_Ratio Weight_lb_per_1000
 Min.   :10.40    Min.   :4.000   Min.   : 71.1     Min.   : 52.0   Min.   :2.760   Min.   :1.513
 1st Qu.:15.43    1st Qu.:4.000   1st Qu.:120.8     1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581
 Median :19.20    Median :6.000   Median :196.3     Median :123.0   Median :3.695   Median :3.325
 Mean   :20.09    Mean   :6.188   Mean   :230.7     Mean   :146.7   Mean   :3.597   Mean   :3.217
 3rd Qu.:22.80    3rd Qu.:8.000   3rd Qu.:326.0     3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610
 Max.   :33.90    Max.   :8.000   Max.   :472.0     Max.   :335.0   Max.   :4.930   Max.   :5.424
 Quarter_Mile_Time  Engine_Shape     Transmission    Forward_Gears    Carburetors
 Min.   :14.50     Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:16.89     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :17.71     Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
 Mean   :17.85     Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:18.90     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :22.90     Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

FIGURE 3 – Résumé statistique des variables

L'analyse du résumé statistique des variables révèle plusieurs éléments intéressants :

— **Miles_per_Gallon** : La consommation moyenne de carburant des véhicules est de 20.09 mpg, avec un minimum de 10.4 et un maximum de 33.9. Cette diversité dans les valeurs permet d'obtenir une analyse plus équilibrée et non biaisée, en prenant en compte la variété des modèles présents dans l'échantillon, allant des modèles économiques aux modèles plus gourmands en carburant.

— **Cylinders** : Le nombre moyen de cylindres est de 6.188, avec des valeurs allant de 4 à 8. Cela montre que la majorité des véhicules ont entre 4 et 8 cylindres, reflétant une tendance vers des moteurs de taille moyenne.

— **Displacement_cuin** : La cylindrée moyenne des moteurs est de 230.7 pouces cubes, avec des valeurs extrêmes allant de 71.1 à 472.

— **Horsepower** : La puissance des véhicules varie de 52 à 335 chevaux, avec une médiane à 123 chevaux. Cela montre une distribution asymétrique à droite, où la majorité des véhicules ont une puissance plus faible, mais quelques-uns atteignent des valeurs très élevées. Nous observerons cette distribution sur l'histogramme présenté dans la figure 6.

— **Rear_Axle_Ratio** : Le rapport moyen du pont arrière est de 3.597, avec des valeurs variant entre 2.76 et 4.93.

— **Weight_lb_per_1000** : Le poids des véhicules varie de 1.513 à 5.424 milliers de livres, avec une moyenne de 3.217. Cette large gamme de poids montre une hétérogénéité importante entre les véhicules compacts et les modèles plus lourds.

— **Quarter_Mile_Time** : Les temps au quart de mile varient de 14.5 à 22.9 secondes, avec une médiane à 17.71 secondes. Ces temps sont relativement symétriques, suggérant une distribution proche de la normale. Nous observerons cette distribution sur l'histogramme présenté dans la figure 9.

— **Engine_Shape** : La forme du moteur est une variable qualitative. Les valeurs peuvent correspondre à différentes configurations, mais un résumé statistique direct n'est pas applicable ici.

— **Transmission** : Le type de transmission, qui peut être manuelle ou automatique, montre une répartition de 15 véhicules avec une transmission manuelle et 17 véhicules avec une transmission automatique. La moyenne de cette variable est donc un indicateur du type de transmission **dominante**.

— **Forward_Gears** : Le nombre moyen de vitesses avant dans la transmission est de 3.688, avec des valeurs variant entre 3 et 5. Cela indique une certaine uniformité dans les véhicules testés, bien que quelques modèles aient plus de vitesses pour de

meilleures performances.

— **Carburetors** : Le nombre de carburateurs varie de 1 à 4, avec une médiane à 2. Cela indique que la majorité des véhicules ont un système de carburateurs plus simple, mais certains modèles en possèdent plusieurs pour améliorer les performances.

Cette première analyse descriptive nous permet de mieux cerner les ordres de grandeur des variables, d'identifier la présence d'éventuelles valeurs extrêmes, et de guider les étapes suivantes de l'analyse statistique.

2.2 Analyse univariée

L'analyse univariée examine chaque variable individuellement afin de comprendre sa distribution.
À partir de l'analyse de la structure du jeu de données (figure 2) et du, il apparaît que les variables numériques peuvent être regroupées en deux sous-catégories : les variables continues et celles à valeurs discrètes.
Nous commencerons alors par l'analyse des variables numériques continues, avant de traiter les variables numériques discrètes.

2.2.1 Variables numériques continues

a.Histogrammes Nous commençons notre analyse par une étude des histogrammes, qui sont des outils essentiels pour examiner la distribution des données.

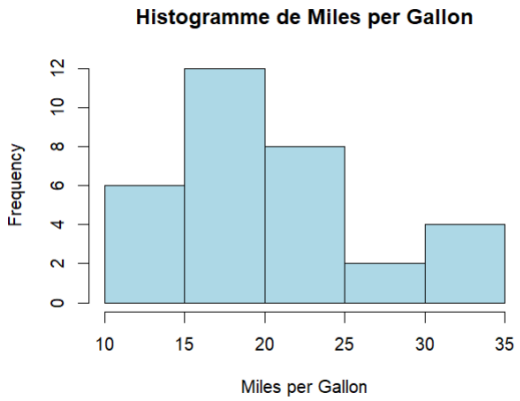

FIGURE 4 – Miles per Gallon

La distribution est légèrement asymétrique à droite. La majorité des véhicules consomment entre 15 et 25 miles par gallon. Quelques valeurs plus élevées représentent des modèles plus économes.

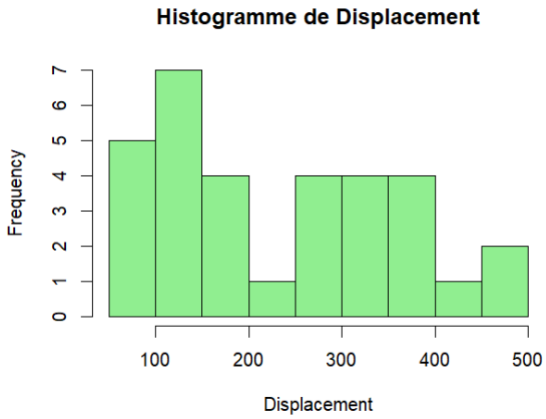

FIGURE 5 – Displacement

La distribution est fortement asymétrique à droite. La plupart des véhicules ont une cylindrée inférieure à 250 pouces cubes, mais certaines valeurs extrêmes vont jusqu'à 500, ce qui peut refléter la présence de moteurs puissants ou de véhicules sportifs.

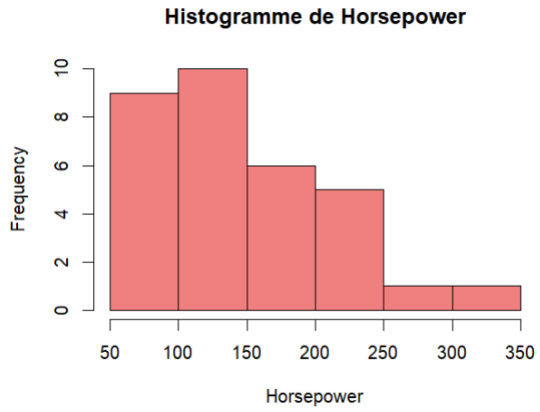

FIGURE 6 – Horsepower

La distribution est également asymétrique à droite. La majorité des voitures ont une puissance entre 90 et 180 chevaux. Quelques voitures très puissantes (au-delà de 250 chevaux) sont présentes dans le jeu de données.

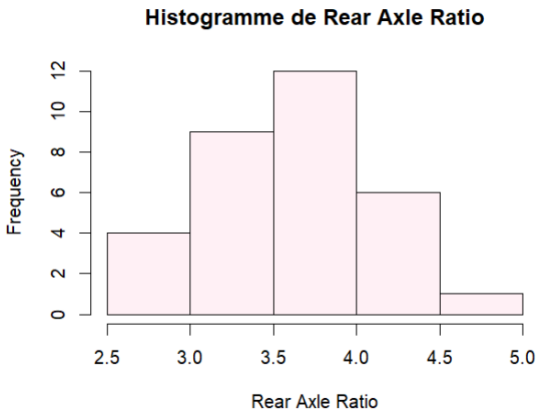

FIGURE 7 – Rear Axle Ratio

Cette variable est plus symétriquement distribuée, avec un pic autour de 3.6. Cela montre une certaine homogénéité des rapports de pont arrière dans le jeu de données.

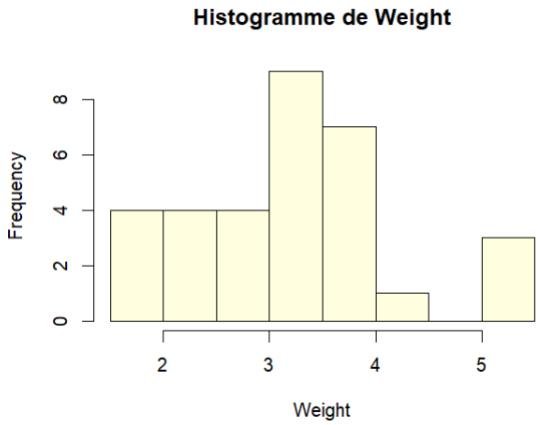

FIGURE 8 – Weight (lb per 1000)

Distribution asymétrique à droite. La plupart des véhicules ont un poids compris entre 2.5 et 3.5 milliers de livres. Quelques modèles plus lourds constituent des valeurs extrêmes.

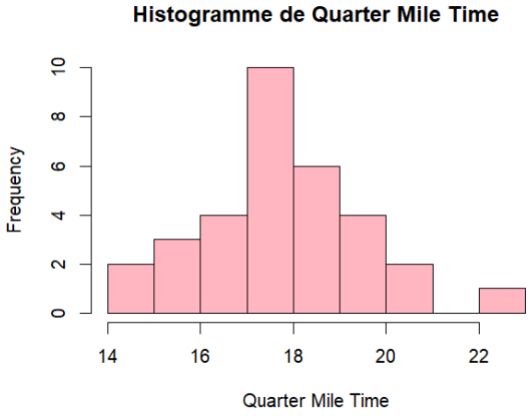

FIGURE 9 – Quarter Mile Time

Cette variable présente une distribution relativement symétrique, avec une majorité des temps situés autour de 18 secondes, sans valeurs aberrantes notables.

La majorité des variables continues présentent une asymétrie positive, ce qui suggère la présence de valeurs extrêmes élevées. Ces valeurs extrêmes pourraient avoir un impact important sur les analyses statistiques futures, en particulier les modèles de régression. Afin de réduire l'influence de ces valeurs atypiques et de stabiliser la variance, une transformation logarithmique a été envisagée pour certaines de ces variables. Cette transformation permet de rendre les distributions plus symétriques et d'atténuer l'effet des valeurs extrêmes.

Nous avons, alors, tracé les histogrammes des variables transformées et les avons comparés côte à côte avec ceux des variables d'origine, afin de déterminer si la transformation améliore la symétrie des distributions et réduit l'impact des valeurs extrêmes.

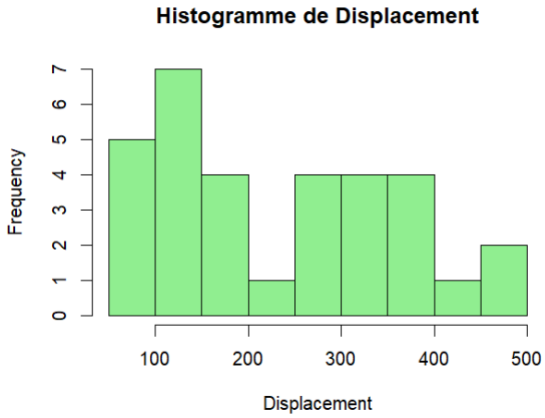

FIGURE 10 – Displacement

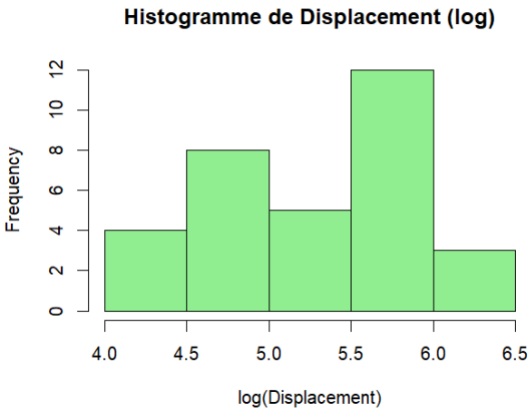

FIGURE 11 – log(Displacement)

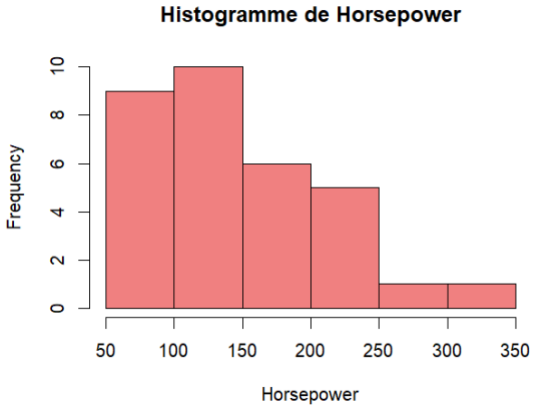

FIGURE 12 – Horsepower

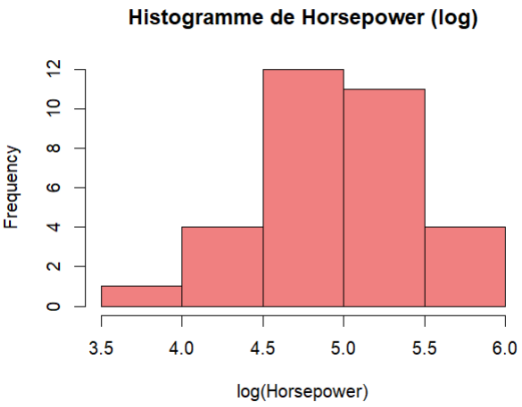

FIGURE 13 – log(Horsepower)

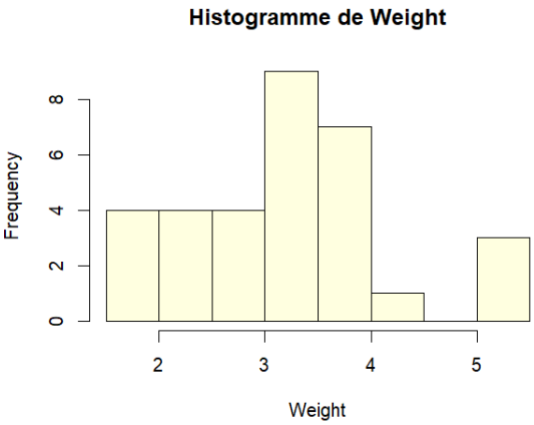

FIGURE 14 – Weight

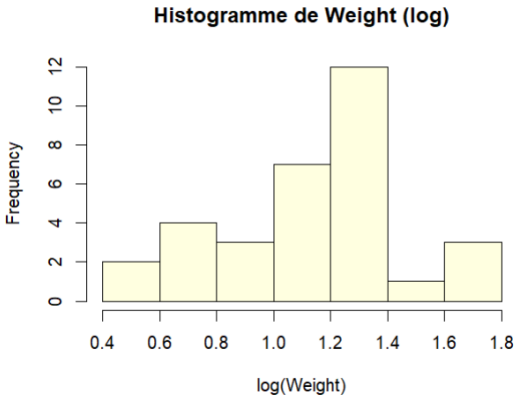

FIGURE 15 – log(Weight)

Afin d'améliorer la symétrie des distributions des variables explicatives (Displacement, Horsepower, Weight) et potentiellement stabiliser la variance pour le modèle de régression,

nous allons utiliser leurs versions transformées par le logarithme naturel. Ces nouvelles colonnes seront nommées en ajoutant le suffixe '_log' à leur nom d'origine. L'analyse comparative des résultats obtenus avec les variables originales et leurs versions logarithmiques permettra d'évaluer l'impact de cette transformation sur la performance et l'interprétation du modèle final.

b.Boxplots Nous utilisons maintenant les boxplots pour valider visuellement le résumé statistique précédemment établi dans la figure 3. Ces boxplots nous permettent de confirmer la distribution des données, d'identifier les valeurs aberrantes et d'évaluer la symétrie. Contrairement aux histogrammes, ils montrent la médiane, les quartiles et les valeurs extrêmes, offrant ainsi une vue plus concise des données. Les résultats obtenus confirment que les valeurs observées sont conformes à nos attentes.

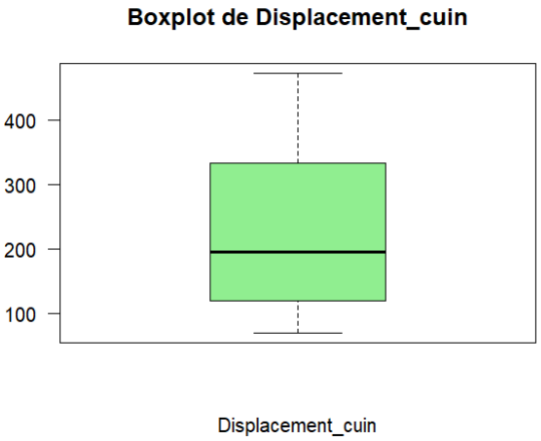

FIGURE 16 – Displacement

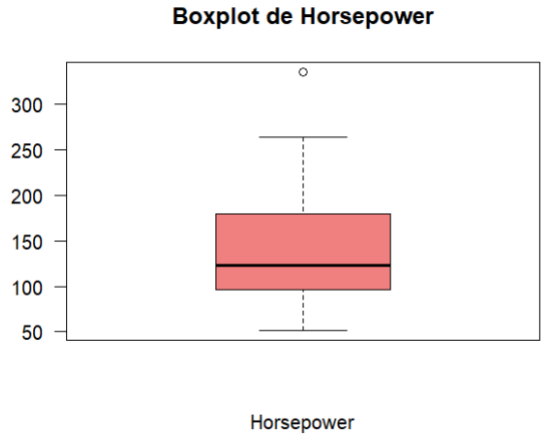

FIGURE 17 – Horsepower

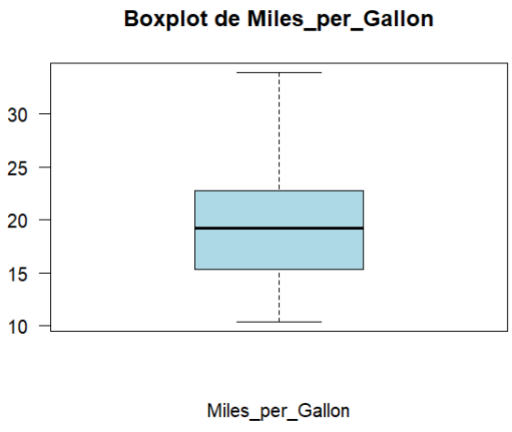

FIGURE 18 – Miles per Gallon

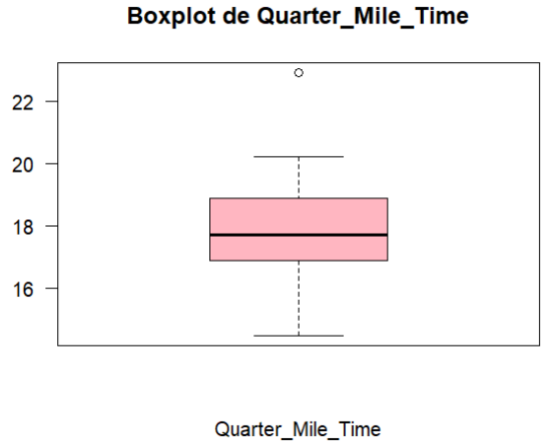

FIGURE 19 – Quarter Mile Time

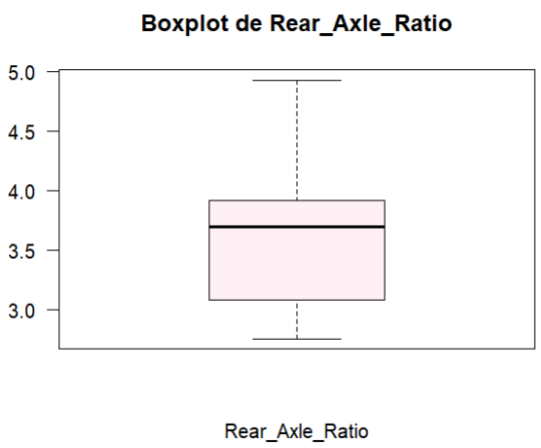

FIGURE 20 – Rear Axle Ratio

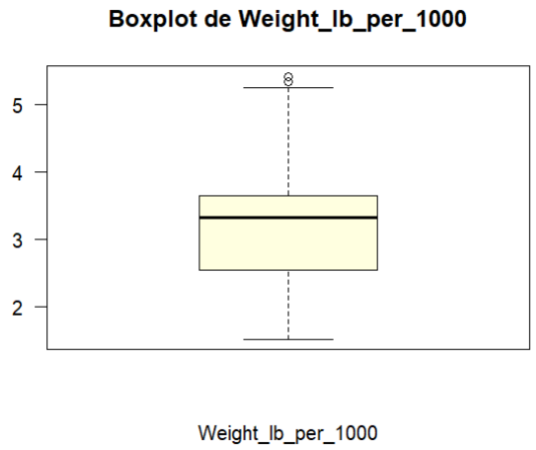

FIGURE 21 – Weight (lb per 1000)

2.2.2 Variables catégoriques

Maintenant que l'analyse des variables numériques continues est terminée, nous passons à l'étude des variables discrètes et la variable catégorique `model`. Dans cette partie, nous allons explorer leur distribution à l'aide de diagrammes en barres (barplots), qui permettent de visualiser efficacement les fréquences de chaque modalité.

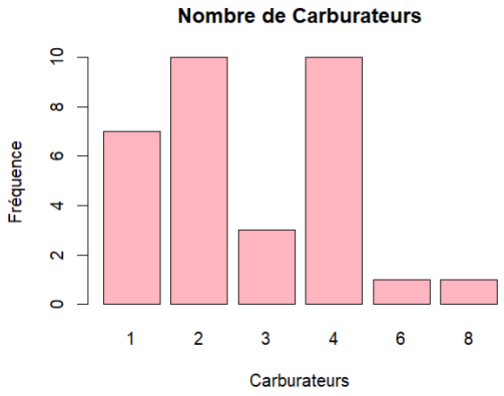

FIGURE 22 – Carburateurs

La majorité des véhicules dans le jeu de données utilise entre 1 et 4 carburateurs. Les valeurs plus élevées sont moins fréquentes.

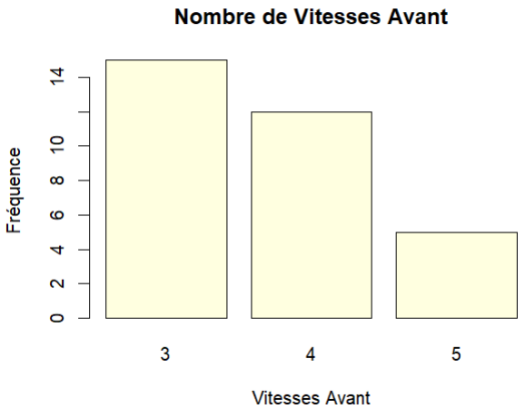

FIGURE 23 – Gear

Le nombre de vitesses avant des véhicules est majoritairement de 3 et 4. Les modèles à 5 vitesses sont moins courants, indiquant une tendance générale vers des véhicules avec moins de vitesses.

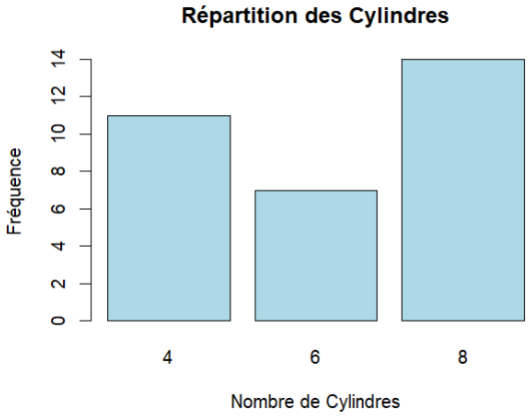

FIGURE 24 – Cylindres

La variable Cylinders montre une prédominance de véhicules ayant 4 ou 8 cylindres. Les véhicules à 6 cylindres sont relativement peu nombreux.

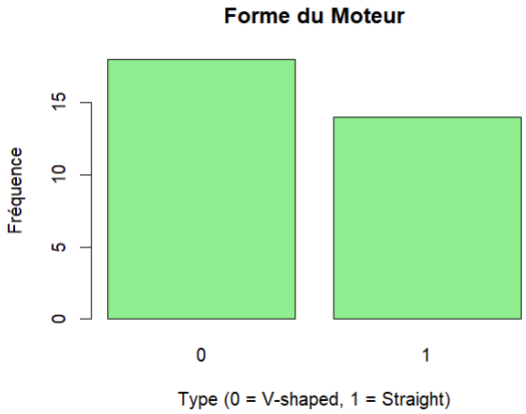

FIGURE 25 – Engine shape

La distribution des moteurs est dominée par les moteurs en forme de "V" (valeur 0), tandis que les moteurs en ligne (valeur 1) sont moins fréquents.

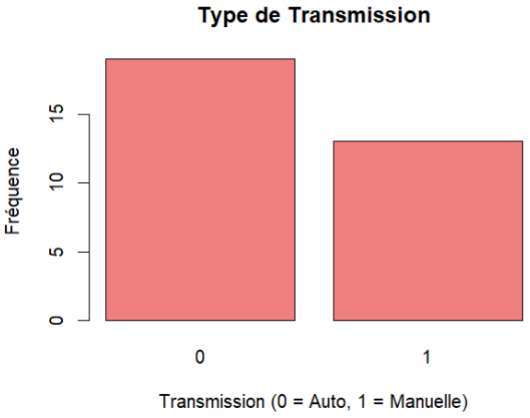

FIGURE 26 – Transmission

Les véhicules avec transmission automatique (valeur 0) sont plus représentés dans le jeu de données.

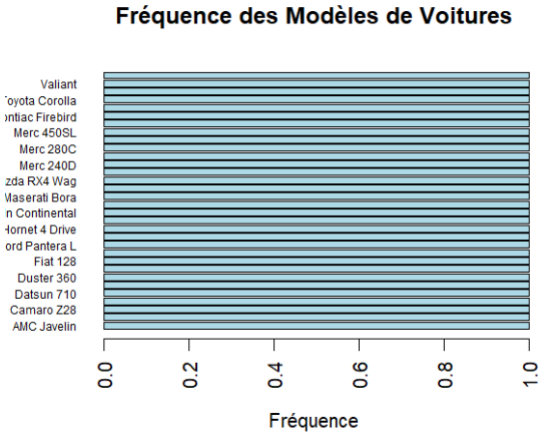

FIGURE 27 – Model

Chaque barre représente un modèle unique de voiture dans notre jeu de données.

2.3 Analyse bivariée

2.3.1 Scatter plots

Nous explorons maintenant les relations entre les variables quantitatives à l'aide des diagrammes de dispersion (scatter plots). Le graphique correspondant est présenté dans la Figure 28, illustrant les relations entre les différentes variables numériques.

FIGURE 28 – Scatter plots

L'analyse de ce scatter plots révèle plusieurs relations intéressantes entre les variables automobiles.

— **mpg (Miles_per_Gallon) :** Une forte relation négative est observée entre notre variable cible et les variables explicatives : cyl (Cylinders), disp (Displacement_cuin), hp (Horsepower), wt (Weight_lb_per_1000).
— **Relations positives notables :**
  — disp (Displacement_cuin) et cyl (Cylinders)
  — disp (Displacement_cuin) et hp (Horsepower)
  — disp (Displacement_cuin) et wt (Weight_lb_per_1000)
  — hp (Horsepower) et wt (Weight_lb_per_1000)

2.3.2 Matrice de coorélation

Pour confirmer et quantifier ces observations, nous allons maintenant analyser la matrice de corrélation, qui permettra d'évaluer précisément la force et la direction des relations entre les différentes variables.

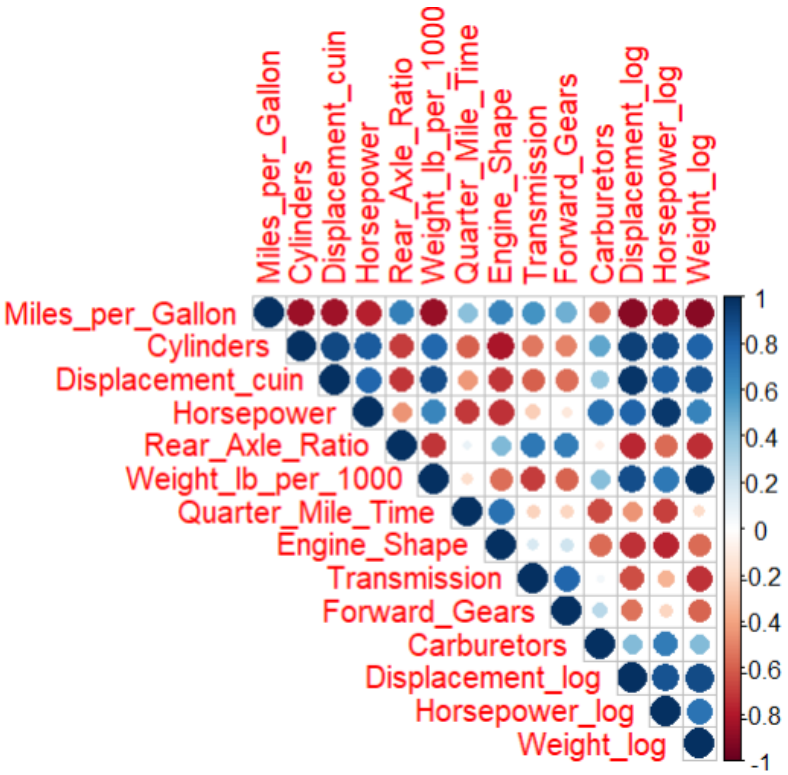

FIGURE 29 – Heatmap

	Miles_per_Gallon	Cylinders	Displacement_cuin	Horsepower	Rear_Axle_Ratio	Weight_lb_per_1000
Miles_per_Gallon	1.0000000	-0.8521620	-0.8475514	-0.7761684	0.68117191	-0.8676594
Cylinders	-0.8521620	1.0000000	0.9020329	0.8324475	-0.69993811	0.7824958
Displacement_cuin	-0.8475514	0.9020329	1.0000000	0.7909486	-0.71021393	0.8879799
Horsepower	-0.7761684	0.8324475	0.7909486	1.0000000	-0.44875912	0.6587479
Rear_Axle_Ratio	0.6811719	-0.6999381	-0.7102139	-0.4487591	1.00000000	-0.7124406
Weight_lb_per_1000	-0.8676594	0.7824958	0.8879799	0.6587479	-0.71244065	1.0000000
Quarter_Mile_Time	0.4186840	-0.5912421	-0.4336979	-0.7082234	0.09120476	-0.1747159
Engine_Shape	0.6640389	-0.8108118	-0.7104159	-0.7230967	0.44027846	-0.5549157
Transmission	0.5998324	-0.5226070	-0.5912270	-0.2432043	0.71271113	-0.6924953
Forward_Gears	0.4802848	-0.4926866	-0.5555692	-0.1257043	0.69961013	-0.5832870
Carburetors	-0.5509251	0.5269883	0.3949769	0.7498125	-0.09078980	0.4276059
Displacement_log	-0.9071119	0.9318804	0.9725003	0.8020950	-0.75645083	0.8845389
Horsepower_log	-0.8487707	0.8806646	0.8278959	0.9687353	-0.56387082	0.7158277
Weight_log	-0.9000811	0.8016184	0.8637147	0.6636911	-0.73124442	0.9788215

	Quarter_Mile_Time	Engine_Shape	Transmission	Forward_Gears	Carburetors	Displacement_log	Horsepower_log
Miles_per_Gallon	0.41868403	0.6640389	0.59983243	0.4802848	-0.55092507	-0.9071119	-0.8487707
Cylinders	-0.59124207	-0.8108118	-0.52260705	-0.4926866	0.52698829	0.9318804	0.8806646
Displacement_cuin	-0.43369788	-0.7104159	-0.59122704	-0.5555692	0.39497686	0.9725003	0.8278959
Horsepower	-0.70822339	-0.7230967	-0.24320426	-0.1257043	0.74981247	0.8020950	0.9687353
Rear_Axle_Ratio	0.09120476	0.4402785	0.71271113	0.6996101	-0.09078980	-0.7564508	-0.5638708
Weight_lb_per_1000	-0.17471588	-0.5549157	-0.69249526	-0.5832870	0.42760594	0.8845389	0.7158277
Quarter_Mile_Time	1.00000000	0.7445354	-0.22986086	-0.2126822	-0.65624923	-0.4479810	-0.6849506
Engine_Shape	0.74453544	1.0000000	0.16834512	0.2060233	-0.56960714	-0.7281298	-0.7617319
Transmission	-0.22986086	0.1683451	1.00000000	0.7940588	0.05753435	-0.6438649	-0.3457917
Forward_Gears	-0.21268223	0.2060233	0.79405876	1.0000000	0.27407284	-0.5465753	-0.2190341
Carburetors	-0.65624923	-0.5696071	0.05753435	0.2740728	1.00000000	0.4383136	0.6996473
Displacement_log	-0.44798103	-0.7281298	-0.64386495	-0.5465753	0.43831361	1.0000000	0.8617723
Horsepower_log	-0.68495060	-0.7617319	-0.34579172	-0.2190341	0.69964733	0.8617723	1.0000000
Weight_log	-0.18053957	-0.5647599	-0.72167808	-0.5840993	0.43787983	0.8992145	0.7314521

FIGURE 30 – Matrice de coorélation

L'analyse de la matrice de corrélation numérique vient étayer les tendances visuelles identifiées dans la matrice de nuages de points :

- Forte corrélation négative de Miles_per_Gallon (mpg) avec Cylinders (-0.85), Displacement_cuin (-0.85), Horsepower (-0.78), et Weight_lb_per_1000 (-0.87). Les corrélations avec les versions logarithmiques sont encore plus fortes : Displacement_log (-0.91), Horsepower_log (-0.85), et Weight_log (-0.90).
- Forte corrélation positive entre (Cylinders, Displacement_cuin) (0.90), (Cylinders, Horsepower) (0.83), (Cylinders, Weight_lb_per_1000) (0.78), (Displacement_cuin, Horsepower) (0.79), (Displacement_cuin, Weight_lb_per_1000) (0.89), et (Horsepower, Weight_lb_per_1000) (0.66).
- Corrélation positive de la variable Tranmission avec Rear_Axle_Ratio (0.71), et négative avec Cylinders (-0.52), Displacement_cuin (-0.59), et Weight_lb_per_1000 (-0.69).

3 Régression multilinéaire

3.1 Analyse du modèle global

Voici le résultat que j'ai trouvé suite à la première approche de régression multiple sur l'ensemble des données avec toutes les variables explicatives fournies :

3.1.1 Régression modèle global

```
Call:
lm(formula = Miles_per_Gallon ~ ., data = num_no_log)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       12.30337   18.71788   0.657   0.5181
Cylinders         -0.11144    1.04502  -0.107   0.9161
Displacement_cuin  0.01334    0.01786   0.747   0.4635
Horsepower        -0.02148    0.02177  -0.987   0.3350
Rear_Axle_Ratio    0.78711    1.63537   0.481   0.6353
Weight_lb_per_1000 -3.71530   1.89441  -1.961   0.0633 .
Quarter_Mile_Time  0.82104    0.73084   1.123   0.2739
Engine_Shape       0.31776    2.10451   0.151   0.8814
Transmission       2.52023    2.05665   1.225   0.2340
Forward_Gears      0.65541    1.49326   0.439   0.6652
Carburetors       -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

FIGURE 31 – Résultat de régression modèle initial

Le modèle initial explique une part substantielle de la variance de la consommation de carburant (R-squared ajusté de 0.8066), et le test F global indique que le modèle est statistiquement significatif ($p < 0.001$). Cependant, l'examen des coefficients individuels révèle que la plupart des variables explicatives ne sont pas statistiquement significatives au seuil de 0.05. Seule la variable Weight_lb_per_1000 approche la significativité ($p = 0.0633$), suggérant qu'un poids plus élevé tend à être associé à une consommation de carburant plus faible dans ce modèle. La non-significativité des autres coefficients pourrait être due à la **multicolinéarité** entre les variables explicatives.

3.1.2 Analyse de la variance (ANOVA)

Afin d'évaluer la contribution individuelle de chaque variable explicative au modèle de régression linéaire, ANOVA a été réalisée, dont les résultats sont présentés dans le tableau suivant :

```
Analysis of Variance Table

Response: Miles_per_Gallon
                  Df Sum Sq Mean Sq  F value    Pr(>F)
Cylinders          1 817.71  817.71 116.4245 5.034e-10 ***
Displacement_cuin  1  37.59   37.59   5.3526  0.030911 *
Horsepower         1   9.37    9.37   1.3342  0.261031
Rear_Axle_Ratio    1  16.47   16.47   2.3446  0.140644
Weight_lb_per_1000 1  77.48   77.48  11.0309  0.003244 **
Quarter_Mile_Time  1   3.95    3.95   0.5623  0.461656
Engine_Shape       1   0.13    0.13   0.0185  0.893173
Transmission       1  14.47   14.47   2.0608  0.165858
Forward_Gears      1   0.97    0.97   0.1384  0.713653
Carburetors        1   0.41    0.41   0.0579  0.812179
Residuals         21 147.49    7.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 32 – Anova modèle 1

Selon cette table ANOVA, les variables qui ont un effet statistiquement significatif sur la consommation de carburant semblent être : Cylinders, Displacement_cuin, et Weight_lb_per_1000. La variable Horsepower n'est pas significative au seuil de 0.05. Les autres variables ne semblent pas apporter une contribution significative supplémentaire à l'explication de la variance de la consommation de carburant une fois que les effets des premières variables sont pris en compte.

3.1.3 Normalité des résidus

Afin d'évaluer visuellement l'hypothèse de normalité des résidus de notre modèle, nous avons effectué un graphique Q-Q (quantile-quantile).

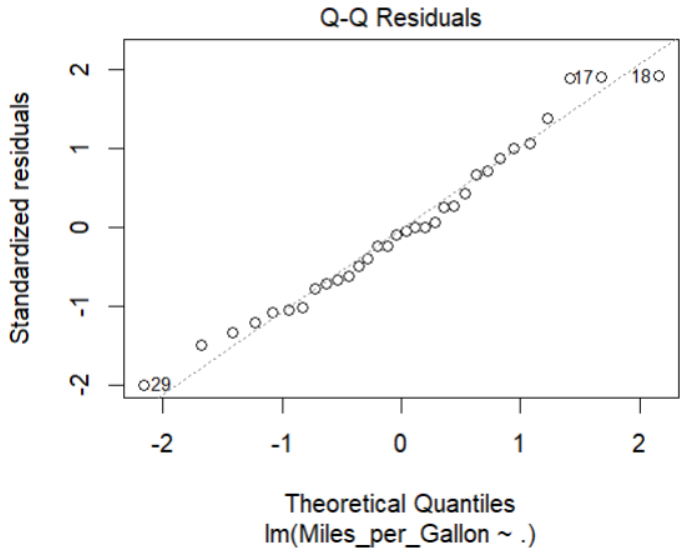

FIGURE 33 – QQ-plot modèle global

On constate que la majorité des points suivent raisonnablement la droite de normalité théorique. Cependant, nous notons une légère déviation aux extrémités, en particulier pour les résidus correspondant aux observations 29, 17 et 18. Bien que cela indique un léger écart par rapport à une distribution parfaitement normale, la déviation ne semble pas suffisamment sévère pour invalider les conclusions de notre modèle.

3.2 Modèle 2 : Variables fortement corrélées

3.2.1 Régression modèle 2

En se basant sur la forte corrélation observée entre Miles_per_Gallon et les variables Cylinders, Displacement_cuin, Horsepower, et Weight_lb_per_1000, il serait logique de considérer un modèle de régression incluant principalement ces variables. La variable Transmission semblait également montrer une relation distincte avec mpg.

```
Call:
lm(formula = Miles_per_Gallon ~ Weight_lb_per_1000 + Cylinders +
    Displacement_cuin + Horsepower + Transmission, data = num_no_log)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5952 -1.5864 -0.7157  1.2821  5.5725

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       38.20280    3.66910  10.412 9.08e-11 ***
Weight_lb_per_1000 -3.30262    1.13364  -2.913  0.00726 **
Cylinders         -1.10638    0.67636  -1.636  0.11393
Displacement_cuin  0.01226    0.01171   1.047  0.30472
Horsepower        -0.02796    0.01392  -2.008  0.05510 .
Transmission       1.55649    1.44054   1.080  0.28984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.505 on 26 degrees of freedom
Multiple R-squared:  0.8551,    Adjusted R-squared:  0.8273
F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

FIGURE 34 – Résultat de régression du modèle 2

Ce modèle réduit montre une amélioration de l'ajustement en termes de R-squared ajusté par rapport au modèle complet. La variable Weight_lb_per_1000 est statistiquement significative, indiquant un impact négatif significatif du poids sur la consommation. La variable Horsepower approche la significativité. Les autres variables ne sont pas statistiquement significatives dans ce modèle. Toutefois, le modèle global reste significatif.

3.2.2 ANOVA modèle 2

```
Analysis of Variance Table

Response: Miles_per_Gallon
                   Df Sum Sq Mean Sq  F value    Pr(>F)
Weight_lb_per_1000  1 847.73  847.73 135.1206 8.458e-12 ***
Cylinders           1  87.15   87.15  13.8910 0.0009487 ***
Displacement_cuin   1   2.68    2.68   0.4271 0.5191516
Horsepower          1  18.05   18.05   2.8767 0.1018130
Transmission        1   7.32    7.32   1.1675 0.2898430
Residuals          26 163.12    6.27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 35 – ANOVA modèle 2

Les résultats de cette analyse indiquent que Weight_lb_per_Gallon et Cylinders ont un impact statistiquement très significatif sur le Miles_per_Gallon ($p < 0.001$). En revanche, les autres variables présentent pas d'effet significatif sur la variable cible.

3.2.3 Normalité des résidus

Après la sélection de variables, nous avons réévalué la normalité des résidus de notre modèle réduit à l'aide d'un nouveau graphique Q-Q.

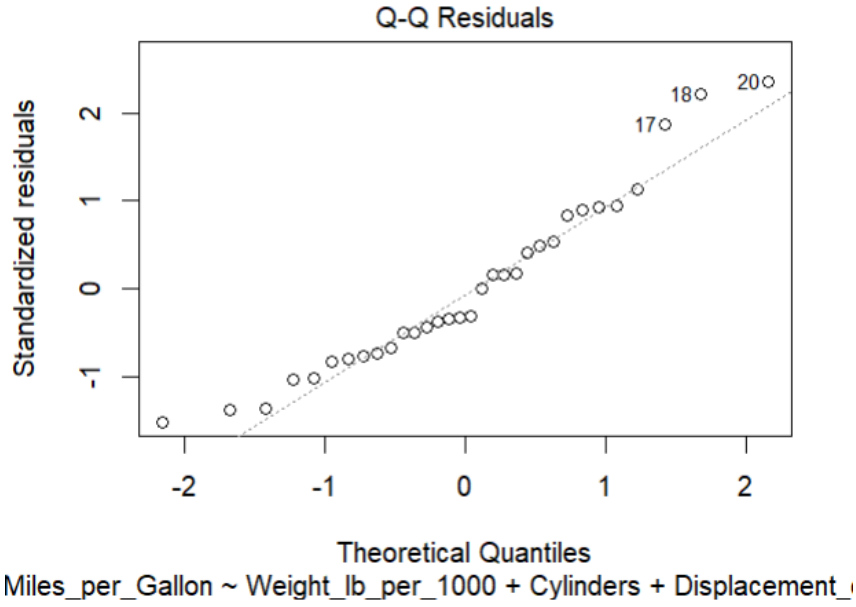

FIGURE 36 – QQ-plot modèle 2

Similaire au graphique précédent, la plupart des points suivent une tendance linéaire, suggérant une distribution des erreurs approximativement normale. Cependant, nous remarquons que les points aux extrémités, notamment les observations 17, 18 et 20, s'éloignent légèrement de la droite. Toutefois, l'amélioration par rapport au modèle initial n'est pas drastique en termes de normalité des résidus.

3.3 Modèle 3 : Variables logarithmiques

3.3.1 Régression modèle 3

Nous allons maintenant explorer un modèle où les variables explicatives fortement corrélées avec notre variable cible (Weight_log + Cylinders + Displacement_log + Horsepower_log + Transmission) sont utilisées dans leur forme transformée logarithmiquement.

```
Call:
lm(formula = Miles_per_Gallon ~ Weight_log + Cylinders + Displacement_log +
    Horsepower_log + Transmission, data = num)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5608 -1.5606 -0.5872  1.1168  4.5117

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      60.19430    9.38034   6.417 8.47e-07 ***
Weight_log       -9.39068    3.14888  -2.982  0.00614 **
Cylinders         0.04034    0.67441   0.060  0.95275
Displacement_log -1.21899    2.77960  -0.439  0.66461
Horsepower_log   -4.82952    2.05110  -2.355  0.02637 *
Transmission      0.49250    1.31825   0.374  0.71173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.176 on 26 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.8697
F-statistic: 42.37 on 5 and 26 DF,  p-value: 1.095e-11
```

FIGURE 37 – Résultat régression modèle 3

Nous constatons une amélioration notable par rapport au modèle précédent, comme en témoignent un R-squared ajusté plus élevé (0.8697 vs 0.8273) et une erreur standard des résidus réduite (2.176 vs 2.505). De plus, les variables transformées logarithmiquement du poids et de la puissance deviennent des prédicteurs statistiquement significatifs de la consommation de carburant.

3.3.2 ANOVA modèle 3

En comparant les résultats de l'ANOVA avec celui du modèle 2, nous constatons une amélioration dans la significativité des prédicteurs après l'application des transformations logarithmiques. L'ajout des transformations logarithmiques a amélioré la significativité, rendant la puissance transformée (Horsepower_log) également significative ($p < 0.05$), en plus du poids transformé (Weight_log, $p < 0.001$) et du nombre de cylindres (Cylinders, $p < 0.01$). Cependant, la cylindrée (Displacement_cuin ou Displacement_log) et le type de transmission (Transmission) n'ont pas montré de significativité dans les deux modèles.

```
Analysis of Variance Table

Response: Miles_per_Gallon
                 Df Sum Sq Mean Sq  F value    Pr(>F)
Weight_log        1 912.26  912.26 192.7067 1.556e-13 ***
Cylinders         1  53.77   53.77  11.3586  0.002355 **
Displacement_log  1   7.16    7.16   1.5117  0.229897
Horsepower_log    1  29.11   29.11   6.1501  0.019941 *
Transmission      1   0.66    0.66   0.1396  0.711733
Residuals        26 123.08    4.73
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 38 – ANOVA modèle 3

3.3.3 Normalité des résidus

Pour évaluer la normalité des erreurs de notre modèle avec les variables transformées, nous avons examiné le graphique Q-Q des résidus standardisés.

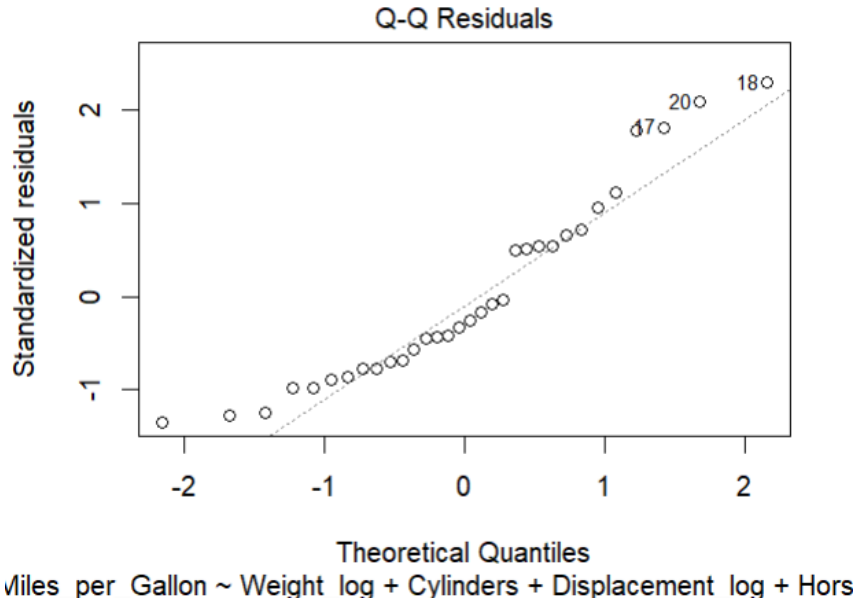

FIGURE 39 – QQ-plot modèle 3

Globalement, l'amélioration de la normalité des résidus après la transformation logarithmique semble minime, voire presque identique à celle observée précédemment (Figure 36).

3.4 Modèle 4 : Variables significatives seulement

3.4.1 Régression modèle 4

Suite à l'analyse de la significativité des variables dans le Modèle 3, nous allons maintenant explorer un Modèle 4 plus parcimonieux. Ce modèle ne retiendra que les variables

qui se sont avérées statistiquement significatives dans l'ANOVA du Modèle 3 : le logarithme du poids (Weight_log), le nombre de cylindres (Cylinders), et le logarithme de la puissance (Horsepower_log). L'objectif est d'obtenir un modèle plus interprétable en se concentrant sur les prédicteurs ayant un impact démontré sur la consommation de carburant. Les résultats de la régression pour ce Modèle 4 sont les suivants :

```
Call:
lm(formula = Miles_per_Gallon ~ Weight_log + Cylinders + Horsepower_log,
    data = num)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8962 -1.4480 -0.5048  1.1993  4.9336

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    56.7223     5.9516   9.531 2.75e-10 ***
Weight_log    -11.0200     2.0099  -5.483 7.42e-06 ***
Cylinders      -0.1996     0.5143  -0.388   0.7009
Horsepower_log  -4.7185     1.6946  -2.784   0.0095 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.116 on 28 degrees of freedom
Multiple R-squared:  0.8887,    Adjusted R-squared:  0.8768
F-statistic: 74.53 on 3 and 28 DF,  p-value: 1.833e-13
```

FIGURE 40 – Résultat régression modèle 4

En passant du Modèle 3 (R-squared ajusté de 0.870) au Modèle 4, nous constatons une légère amélioration du R-squared ajusté à 0.877. Dans le Modèle 4, le logarithme du poids (Weight_log, $p < 0.001$) et le logarithme de la puissance (Horsepower_log, $p < 0.01$) sont statistiquement très significatifs, tandis que le nombre de cylindres (Cylinders) ne l'est pas ($p > 0.05$). L'erreur standard des résidus diminue légèrement de 2.18 à 2.12. Globalement, le Modèle 4 offre un modèle plus meilleur avec un pouvoir explicatif légèrement amélioré et se concentre sur les facteurs ayant un impact statistiquement robuste sur la consommation de carburant.

3.4.2 ANOVA modèle 4

```
Analysis of Variance Table

Response: Miles_per_Gallon
               Df Sum Sq Mean Sq  F value    Pr(>F)
Weight_log      1 912.26  912.26 203.8350 2.229e-14 ***
Cylinders       1  53.77   53.77  12.0145  0.001721 **
Horsepower_log  1  34.70   34.70   7.7533  0.009504 **
Residuals      28 125.31    4.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 41 – ANOVA modèle 4

En comparaison avec l'ANOVA du Modèle 3, toutes les variables retenues dans le Modèle 4 restent significatives, et la variance des erreurs (Mean Sq des Residuals) a

légèrement diminué (de 4.73 à 4.48), suggérant un meilleur ajustement.

3.4.3 Normalité des résidus

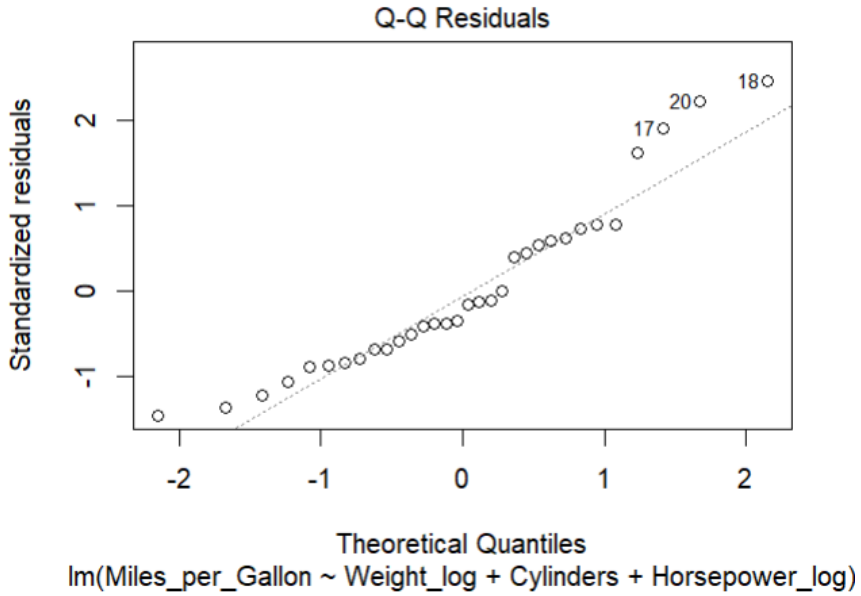

FIGURE 42 – QQ-plot modèle 4

La normalité des résidus ne semble pas avoir été significativement améliorée en passant du Modèle 3 au Modèle 4.

3.5 AIC : Variables explicatives optimales

Après avoir exploré différentes combinaisons de variables et leurs transformations, constatant des améliorations progressives mais parfois modestes de nos modèles, nous allons maintenant recourir au critère d'information d'Akaike (AIC) pour identifier un ensemble de variables explicatives optimal qui équilibre l'ajustement du modèle et sa complexité. L'AIC nous permettra de sélectionner le modèle qui minimise la perte d'information relative.

En appliquant la sélection de variables basée sur l'AIC, nous remarquons une inclusion simultanée de la variable Displacement_cuin et de sa transformation logarithmique (Displacement_log) dans le modèle retenu. Cette présence conjointe est inhabituelle et soulève des questions de multicolinéarité et de redondance d'information.

```
Call:
lm(formula = Miles_per_Gallon ~ Displacement_cuin + Weight_lb_per_1000 +
    Displacement_log + Horsepower_log, data = num)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7605 -1.4953 -0.3428  1.5271  3.0832

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       94.75106   11.21048   8.452 4.59e-09 ***
Displacement_cuin  0.03991    0.01295   3.082  0.00469 **
Weight_lb_per_1000 -2.90496   0.83219  -3.491  0.00167 **
Displacement_log  -10.43242   3.02455  -3.449  0.00186 **
Horsepower_log     -3.97176   1.52885  -2.598  0.01501 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.011 on 27 degrees of freedom
Multiple R-squared:  0.9031,    Adjusted R-squared:  0.8887
F-statistic: 62.87 on 4 and 27 DF,  p-value: 2.746e-13
```

FIGURE 43 – Résultat AIC

3.6 Effet d'ajout de lignes sur la régression

Nous allons, d'ici la fin du projet, nous concentrer sur le Modèle 4. Afin d'illustrer l'impact potentiel de l'ajout de nouvelles données sur la robustesse de ce modèle, nous allons simuler l'ajout d'une observation représentant différents scénarios typiques, aberrants ou étendant la plage de nos données existantes.

3.6.1 Observation typique

```
Call:
lm(formula = formula(model4), data = num_plus1_typique_m4)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9084 -1.4096 -0.3143  1.1806  4.9302

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    56.6509     5.8098   9.751 1.17e-10 ***
Weight_log    -11.0218     1.9753  -5.580 5.07e-06 ***
Cylinders      -0.2052     0.5026  -0.408  0.68599
Horsepower_log  -4.6949     1.6503  -2.845  0.00807 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.079 on 29 degrees of freedom
Multiple R-squared:  0.8887,    Adjusted R-squared:  0.8772
F-statistic: 77.16 on 3 and 29 DF,  p-value: 6.252e-14
```

FIGURE 44 – Effet observation typique

L'ajout d'une observation typique a légèrement modifié les coefficients du Modèle 4, notamment en rendant la variable Cylinders non significative. L'ajustement global du modèle s'est légèrement amélioré, comme indiqué par une augmentation du R-squared ajusté (de 0.867 à 0.877) et une diminution de l'erreur standard des résidus (de 2.255 à 2.079).

3.6.2 Valeur aberrante

```
Call:
lm(formula = formula(model4), data = num_plus1_aberrante_m4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0105 -1.4140 -0.3517  1.1757  4.9760

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     55.5210     4.0800  13.608 4.03e-14 ***
Weight_log     -10.8343     1.8681  -5.800 2.76e-06 ***
Cylinders       -0.3155     0.3029  -1.042 0.306137
Horsepower_log  -4.3661     1.1234  -3.887 0.000544 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.082 on 29 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.8868
F-statistic: 84.55 on 3 and 29 DF,  p-value: 1.924e-14
```

FIGURE 45 – Effet valeur aberrante

On constate que l'ajout d'une valeur aberrante a entraîné des changements notables dans les coefficients et une amélioration de l'ajustement global du modèle (R-squared ajusté plus élevé), bien que cela puisse être un résultat trompeur en raison de l'influence disproportionnée de l'aberration.

3.6.3 Observation étendant la plage

```
Call:
lm(formula = formula(model4), data = num_plus1_etendue_m4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2450 -1.4951 -0.6377  1.1614  5.2301

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     51.5300     5.5201   9.335 3.07e-10 ***
Weight_log      -9.9469     2.0139  -4.939 3.01e-05 ***
Cylinders       -0.6104     0.4872  -1.253   0.2203
Horsepower_log  -3.3597     1.6040  -2.095   0.0451 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.209 on 29 degrees of freedom
Multiple R-squared:  0.8811,    Adjusted R-squared:  0.8688
F-statistic: 71.61 on 3 and 29 DF,  p-value: 1.623e-13
```

FIGURE 46 – Effet observation étendant la plage

On remarque un impact modéré sur le modèle. L'ajustement global s'est légèrement amélioré, et la significativité de Horsepower_log a augmenté. Cela suggère que l'extension de la plage des données peut aider à mieux définir la relation entre les variables.

4 Analyse en composantes principales (ACP)

4.1 Résultats ACP

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.400 1.628 0.77280 0.51914 0.47143 0.45839 0.36458 0.28405 0.23163
Proportion of Variance 0.576 0.265 0.05972 0.02695 0.02223 0.02101 0.01329 0.00807 0.00537
Cumulative Proportion  0.576 0.841 0.90071 0.92766 0.94988 0.97089 0.98419 0.99226 0.99762
                          PC10
Standard deviation     0.15426
Proportion of Variance 0.00238
Cumulative Proportion  1.00000
```

FIGURE 47 – Résultats ACP

On voit clairement que les deux premières composantes principales (PC1 et PC2) capturent à elles seules 84.1% de la variance totale des données, suggérant qu'une grande partie de l'information est concentrée dans ces deux dimensions. Les composantes suivantes expliquent individuellement une portion de plus en plus faible de la variance, indiquant une importance décroissante pour la représentation globale des données. Ceci est montré visuellement dans le scree plot de la figure 50.

```
                             PC1         PC2         PC3           PC4         PC5         PC6
Cylinders            0.4029711  0.03901479 -0.13874360 -8.040022e-05 -0.06148048  0.18206407
Displacement_cuin    0.3959243 -0.05393117 -0.01633491 -2.646304e-01 -0.33851109 -0.35738419
Horsepower           0.3543255  0.24496137  0.18225874  6.000387e-02 -0.52828704  0.03269674
Rear_Axle_Ratio     -0.3155948  0.27847781  0.13057734 -8.528509e-01 -0.10299748  0.23386885
Weight_lb_per_1000   0.3668004 -0.14675805  0.38579961 -2.527210e-01  0.14410292 -0.43201764
Quarter_Mile_Time   -0.2198982 -0.46066271  0.40307004 -7.174202e-02  0.21341845 -0.29265169
Engine_Shape        -0.3333571 -0.22751987  0.41252247  2.119502e-01 -0.62369179  0.11710663
Transmission        -0.2474991  0.43201042 -0.23493804  3.190779e-02 -0.04930286 -0.60874338
Forward_Gears       -0.2214375  0.46516217  0.27929375  2.623809e-01 -0.02039816 -0.24560902
Carburetors          0.2267080  0.41169300  0.56172255  1.233534e-01  0.36576403  0.25782743
                            PC7         PC8          PC9          PC10
Cylinders            0.04257067  0.07041306 -0.863268748 -0.1670687388
Displacement_cuin   -0.19767431 -0.14361684 -0.020039738  0.6838300858
Horsepower           0.08503414  0.58708325  0.291428365 -0.2462606844
Rear_Axle_Ratio     -0.03226657  0.04010725 -0.086765162 -0.0544414772
Weight_lb_per_1000   0.03368560 -0.36605124  0.075971836 -0.5318885631
Quarter_Mile_Time    0.03797611  0.59621869 -0.244573292  0.1545795278
Engine_Shape         0.23387904 -0.36246041 -0.182200371  0.0055443849
Transmission         0.54631997  0.02588771 -0.154149509  0.0003995261
Forward_Gears       -0.69429321 -0.01069942 -0.198369367 -0.0741152014
Carburetors          0.33623769 -0.08067483  0.003086198  0.3585136181
```

FIGURE 48 – Résultats ACP

La PC1 est fortement corrélée positivement avec Cylinders et Displacement_cuin, et négativement avec Rear_Axle_Ratio et Quarter_Mile_Time. La PC2 est principalement influencée positivement par Rear_Axle_Ratio et Transmission (potentiellement manuelle), et négativement par Quarter_Mile_Time.

4.2 Visualisation des résultats

4.2.1 Scree plot

Afin de déterminer le nombre optimal de dimensions à retenir suite à notre ACP, nous avons réalisé un scree plot.

Nous remarquons que les deux premières composantes principales expliquent cumulativement 84.1% de la variance totale (CP1 : 57.6%, CP2 : 26.5%). La diminution brutale de la variance expliquée au-delà de la deuxième composante suggère un coude clair. En conséquence, nous allons ne considérer que les deux premières composantes principales pour nos analyses ultérieures, car elles capturent l'essentiel de la variabilité des données tout en réduisant significativement la dimensionnalité.

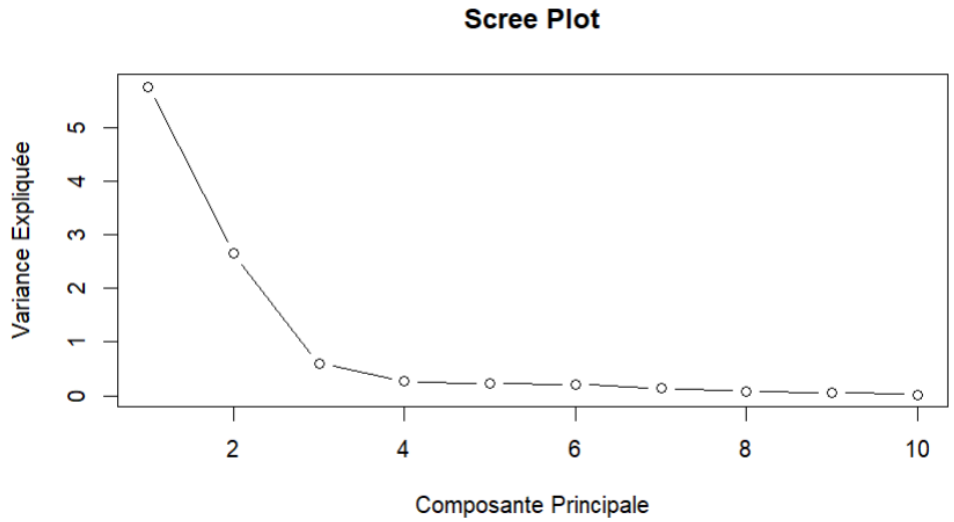

FIGURE 49 – Scree plot

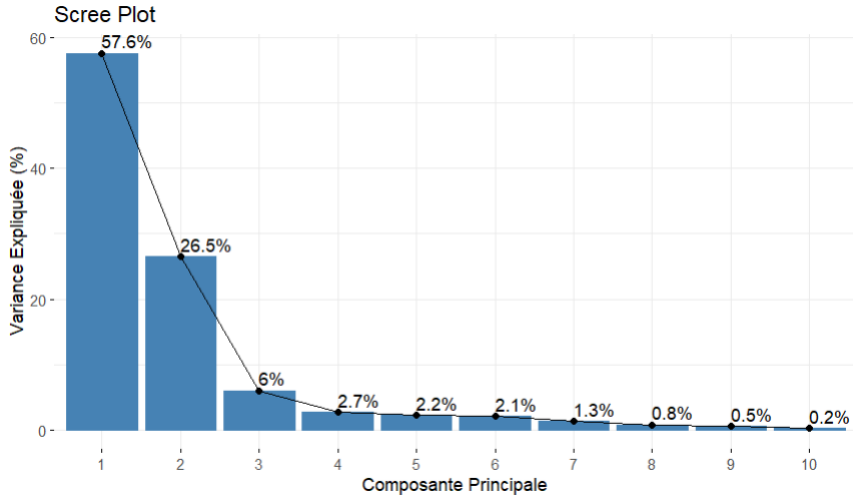

FIGURE 50 – Scree plot avec les pourcentages

4.2.2 Regroupement des variables

Sur le biplot ainsi que le cercle de corrélation, on note une orientation similaire pour les flèches représentant 'Forward_Gears' et 'Transmission', indiquant une possible corrélation positive entre ces caractéristiques. De même, 'Cylinders', 'Displacement_cuin', et 'Weight_lb_per_1000' pointent dans des directions proches, suggérant qu'elles varient ensemble. Inversement, 'Carburetors' s'oppose à 'Quarter_Mile_Time', et 'Horsepower' tend à s'opposer à 'Engine_Shape', ce qui révèle des relations inverses dans la manière dont ces variables contribuent aux deux premières composantes principales.

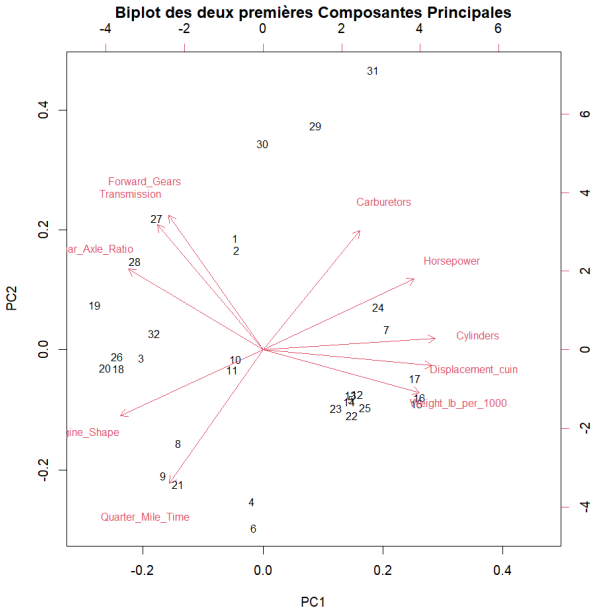

FIGURE 51 – Biplot

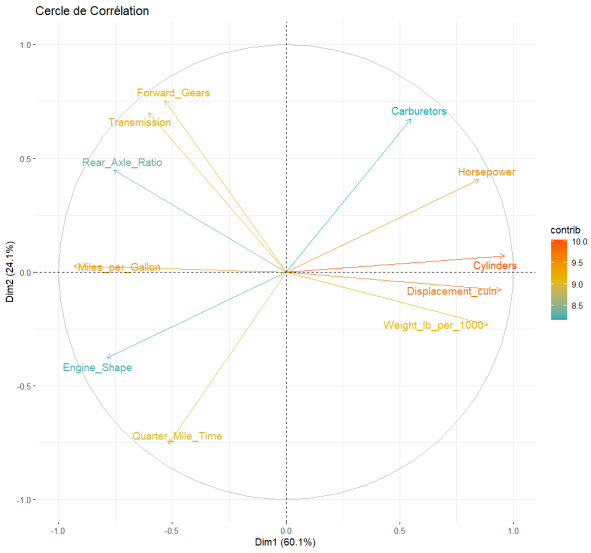

FIGURE 52 – Cercle de corrélations

4.3 Regroupement des individus

En superposant l'origine des voitures sur cette projection, nous distinguons des regroupements par pays : les voitures américaines, japonaises et européennes pourraient former des clusters distincts, reflétant des profils de caractéristiques différents mis en évidence par l'ACP.

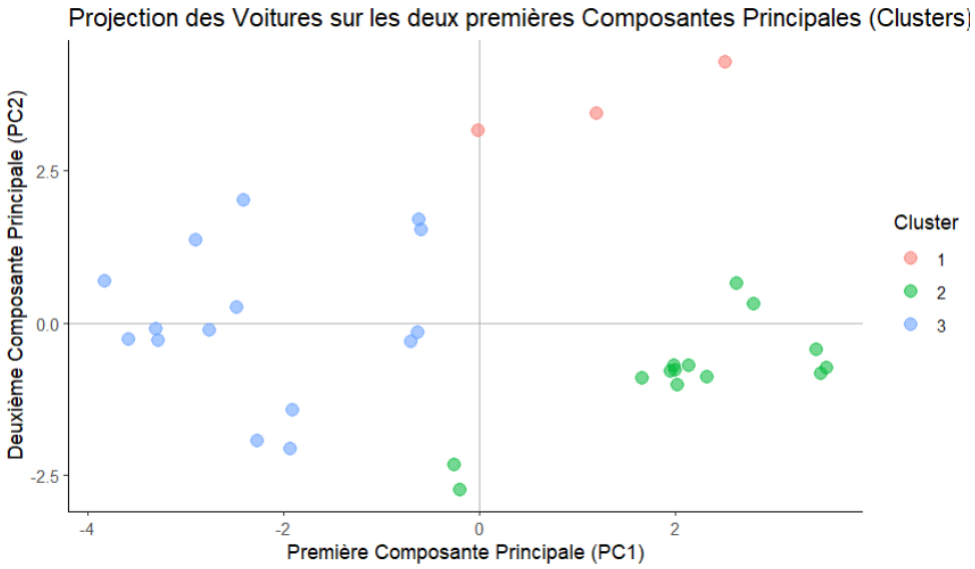

FIGURE 53 – Clusters

4.4 Régression sur les composantes principales

4.4.1 Toutes les composantes principales

Le modèle de régression sur toutes les composantes principales montre un R-squared et un R-squared ajusté de 1, indiquant un ajustement parfait aux données d'entraînement, ce qui suggère un **surajustement** et une fiabilité douteuse pour la généralisation.

```
Call:
lm(formula = y ~ ., data = scores_cp)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.09062    0.46849  42.884  < 2e-16 ***
Comp.1      -2.28131    0.19833 -11.503 1.58e-10 ***
Comp.2       0.11632    0.29242   0.398   0.6948
Comp.3      -1.29925    0.61593  -2.109   0.0471 *
Comp.4       0.09002    0.91688   0.098   0.9227
Comp.5      -0.31279    1.00966  -0.310   0.7598
Comp.6      -0.38410    1.03840  -0.370   0.7152
Comp.7      -0.26029    1.30558  -0.199   0.8439
Comp.8       1.10156    1.67575   0.657   0.5181
Comp.9      -1.28202    2.05496  -0.624   0.5394
Comp.10      3.51367    3.08562   1.139   0.2676
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

FIGURE 54 – Résultat régression

Tous les coefficients des composantes principales sont statistiquement significatifs (p < 2e-16), mais leur interprétation directe en termes des variables d'origine nécessite de considérer les chargements de l'ACP. Ce modèle, bien que décrivant parfaitement les données actuelles, n'est probablement pas le plus pertinent pour prédire de nouvelles observations.

4.4.2 Modèle à deux composantes

```
Call:
lm(formula = y ~ Comp.1 + Comp.2, data = scores_cp)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3611 -1.7263 -0.3322  1.3208  5.6763

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.0906     0.4591  43.760  < 2e-16 ***
Comp.1       -2.2813     0.1944 -11.738 1.55e-12 ***
Comp.2        0.1163     0.2866   0.406    0.688
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.597 on 29 degrees of freedom
Multiple R-squared:  0.8263,    Adjusted R-squared:  0.8143
F-statistic: 68.97 on 2 and 29 DF,  p-value: 9.493e-12
```

FIGURE 55 – Régression 2 CP

Comparativement au modèle utilisant toutes les composantes principales (ajustement parfait mais surajusté), ce modèle à deux composantes offre une explication robuste de

la variance (84.2%) avec moins de prédicteurs et sans signe de surajustement. La Comp.1 (composante 1) s'avère un prédicteur significatif du Miles_per_Gallon, contrairement à Comp.2 qui n'apporte pas d'amélioration significative au modèle. Ce modèle est préférable pour une interprétation et une potentielle généralisation.

4.4.3 Représentation graphique

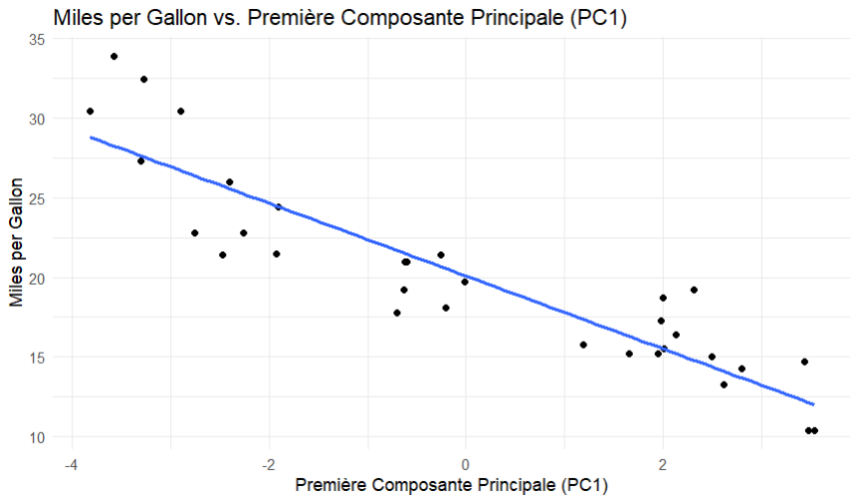

FIGURE 56 – Régression avec PC1

Pour PC1, nous observons que la ligne de régression linéaire semble raisonnablement bien ajustée aux données, suggérant que PC1 capture une part notable de la variance du Miles per Gallon à travers une relation linéaire.

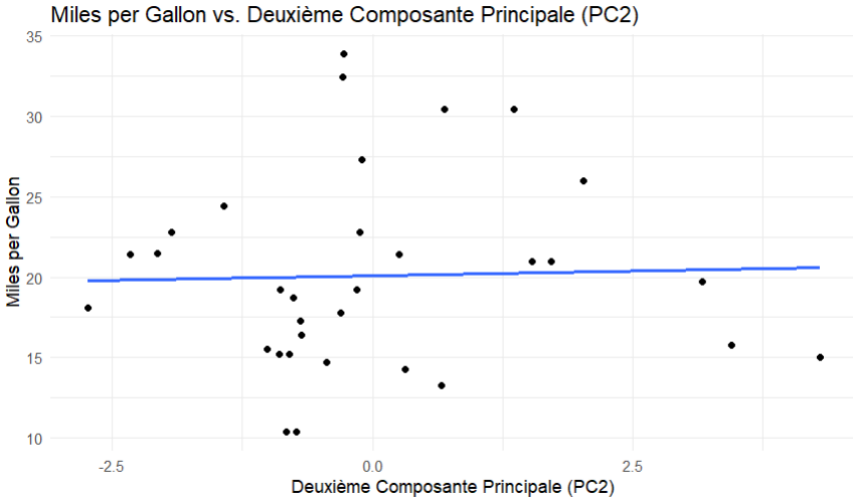

FIGURE 57 – Régression avec PC2

Contrairement à PC1, le nuage de points pour PC2 montre une dispersion plus aléatoire autour de la ligne de régression presque horizontale, indiquant un faible pouvoir explicatif linéaire de cette composante sur le Miles per Gallon. L'absence de tendance

claire suggère que la variation capturée par PC2 n'est pas fortement associée à la consommation de carburant selon un modèle linéaire simple.

5 Conclusion : Avantages et limites des approches

La régression multiple sur mtcars offre une modélisation directe de la consommation, mais souffre de la sélection des variables et de la colinéarité. L'ACP réduit la dimensionnalité et élimine la colinéarité, simplifiant l'analyse et révélant des structures. La RCP combine les deux, utilisant les composantes principales pour la régression, évitant la colinéarité mais perdant en interprétation directe. Le succès de la RCP dépend de la liaison entre les composantes et la consommation. Le choix entre ces méthodes dépend des objectifs : interprétation des variables originales (régression multiple) ou simplification et contournement de la colinéarité (ACP/RCP), en acceptant une interprétation moins directe avec l'ACP/RCP.

A Code R utilisé

```r
setwd("C:/Users/hanof/OneDrive/Desktop/STA202_TPs") # Repertoire de travail
rm(list=objects()); graphics.off()

# Lire le fichier CSV
df <- read.csv2("mtcars.csv", sep=",", dec=".", header=TRUE)

head(df) # Afficher les premieres lignes
dim(df) # Dimensions des donnees
str(df) # Structure des donnees
names(df) # Noms des colonnes

# Analyse univariee
num <- df[,-1] # Selectionner les colonnes numeriques

# Renommer les colonnes
colnames(num) <- c(
  "Miles_per_Gallon",
  "Cylinders",
  "Displacement_cuin",
  "Horsepower",
  "Rear_Axle_Ratio",
  "Weight_lb_per_1000",
  "Quarter_Mile_Time",
  "Engine_Shape",
  "Transmission",
  "Forward_Gears",
  "Carburetors"
)

summary(num) # Statistiques descriptives

#On remarque la presence de variables categoriques discretes et continues donc on va les separer
num_cont <- num[, c("Miles_per_Gallon", "Displacement_cuin", "Horsepower", "Rear_Axle_Ratio",
                    "Weight_lb_per_1000", "Quarter_Mile_Time")]

num_dis <- num[, c("Miles_per_Gallon","Cylinders","Engine_Shape","Transmission","Forward_Gears","Carburetors")]

#Histogrammes:
# Miles per Gallon (variable cible)
hist(df$mpg, main="Histogramme de Miles per Gallon", xlab="Miles per Gallon", col="lightblue", border="black")

# Displacement
hist(df$disp, main="Histogramme de Displacement", xlab="Displacement", col="lightgreen", border="black")

# Horsepower
hist(df$hp, main="Histogramme de Horsepower", xlab="Horsepower", col="lightcoral", border="black")
```

```r

# Weight
hist(df$wt, main="Histogramme de Weight", xlab="Weight", col="
    lightyellow", border="black")

# Quarter Mile Time
hist(df$qsec, main="Histogramme de Quarter Mile Time", xlab="Quarter
    Mile Time", col="lightpink", border="black")

# Rear Axle Ratio
hist(df$drat, main="Histogramme de Rear Axle Ratio", xlab="Rear Axle
    Ratio", col="lavenderblush", border="black")

# Histogrammes logarithmiques:
# Displacement
hist(log(df$disp), main="Histogramme de Displacement (log)", xlab="log(
    Displacement)", col="lightgreen", border="black")

# Horsepower
hist(log(df$hp), main="Histogramme de Horsepower (log)", xlab="log(
    Horsepower)", col="lightcoral", border="black")

# Weight
hist(log(df$wt), main="Histogramme de Weight (log)", xlab="log(Weight)",
     col="lightyellow", border="black")

# Quarter Mile Time
hist(log(df$qsec), main="Histogramme de Quarter Mile Time (log)", xlab="
    log(Quarter Mile Time)", col="lightpink", border="black")

# Appliquer la transformation logarithmique aux colonnes specifiees
num$Displacement_log <- log(num$Displacement_cuin)
num$Horsepower_log <- log(num$Horsepower)
num$Weight_log <- log(num$Weight_lb_per_1000)

# Boite a moustaches
# Miles per Gallon
boxplot(num_cont$Miles_per_Gallon,
        main = "Boxplot de Miles_per_Gallon",
        xlab="Miles_per_Gallon",
        col = "lightblue",
        las = 2)

# Displacement
boxplot(num_cont$Displacement_cuin,
        main = "Boxplot de Displacement_cuin",
        xlab="Displacement_cuin",
        col = "lightgreen",
        las = 2)

# Horsepower
boxplot(num_cont$Horsepower,
        main = "Boxplot de Horsepower",
        xlab="Horsepower",
        col = "lightcoral",
        las = 2)
```

```r

# Rear Axle Ratio
boxplot(num_cont$Rear_Axle_Ratio,
        main = "Boxplot de Rear_Axle_Ratio",
        xlab="Rear_Axle_Ratio",
        col = "lavenderblush",
        las = 2)

# Weight
boxplot(num_cont$Weight_lb_per_1000,
        main = "Boxplot de Weight_lb_per_1000",
        xlab="Weight_lb_per_1000",
        col = "lightyellow",
        las = 2)

# Quarter Mile Time
boxplot(num_cont$Quarter_Mile_Time,
        main = "Boxplot de Quarter_Mile_Time",
        xlab="Quarter_Mile_Time",
        col = "lightpink",
        las = 2)


# Diagrammes en barres
# Cylindres
barplot(table(df$cyl),
        main = "Repartition des Cylindres",
        xlab = "Nombre de Cylindres",
        ylab = "Frequence",
        col = "lightblue",
        border = "black")

# Engine Shape (V/S)
barplot(table(df$vs),
        main = "Forme du Moteur",
        xlab = "Type (0 = V-shaped, 1 = Straight)",
        ylab = "Frequence",
        col = "lightgreen",
        border = "black")

# Transmission
barplot(table(df$am),
        main = "Type de Transmission",
        xlab = "Transmission (0 = Auto, 1 = Manuelle)",
        ylab = "Frequence",
        col = "lightcoral",
        border = "black")

# Forward Gears
barplot(table(df$gear),
        main = "Nombre de Vitesses Avant",
        xlab = "Vitesses Avant",
        ylab = "Frequence",
        col = "lightyellow",
        border = "black")

```

```r
# Carburetors
barplot(table(df$carb),
        main = "Nombre de Carburateurs",
        xlab = "Carburateurs",
        ylab = "Frequence",
        col = "lightpink",
        border = "black")


# Modeles de voitures
barplot(table(df$model),
        main = "Frequence des Modeles de Voitures",
        xlab = "Frequence",
        col = "lavenderblush",
        border = "black",
        horiz = TRUE,
        las = 1)



pairs(num) # Matrice de nuages de points pour toutes les variables numeriques

# Affichage du triangle superieur de la matrice de correlation
library(corrplot)
cor(num) # Matrice de correlation
corrplot(cor(num), type = "upper") # Afficher la matrice de correlation (triangle superieur)

# Supprimer les colonnes contenant "log" dans leur nom
num_no_log <- num[, !grepl("log", colnames(num))]

#Regression mutilineaire

# Modele 1 : lineaire avec toutes les variables (sans log)
model1 <- lm(Miles_per_Gallon ~ ., data = num_no_log)
summary(model1)
anova(model1)
plot(model1,2) # QQ-plot


# Modele 2: lineaire avec des variables specifiques (sans log)
model2 <- lm(Miles_per_Gallon ~ Weight_lb_per_1000 + Cylinders + Displacement_cuin + Horsepower + Transmission, data = num_no_log)
summary(model2)
anova(model2)
plot(model2,2)

# Modele 3: lineaire avec des variables transformees (log)
model3 <-lm(Miles_per_Gallon ~ Weight_log + Cylinders + Displacement_log + Horsepower_log + Transmission, data = num)
summary(model3)
anova(model3)
plot(model3,2)

# Modele 4 : lineaire avec variables significatives (log)
```

```r
model4 <- lm(Miles_per_Gallon ~ Weight_log + Cylinders + Horsepower_log, data = num)
summary(model4)
anova(model4)
plot(model4,2)

# Modele 5: complet avec toutes les variables numeriques
full_model_aic <- lm(Miles_per_Gallon ~ ., data = num)
library(MASS)
model5<- stepAIC(full_model_aic, direction = "both") # Selection de variables par AIC
summary(model5)
# Formule du meilleur modele
formula(model5)

# Modele 6 : lineaire avec des variables specifiques (mix de log et non log)
model6 <- lm(Miles_per_Gallon ~ Weight_lb_per_1000 + Displacement_cuin + Horsepower_log, data = num)
summary(model6)


# Effet ajout de lignes

# Exemple 1 : creation d'une nouvelle observation typique
nouvelle_ligne_typique_m4_complet <- data.frame(
  Miles_per_Gallon = 20,
  Cylinders = 6,
  Displacement_cuin = 200,
  Horsepower = 150,
  Rear_Axle_Ratio = 3.5,        # Valeur typique
  Weight_lb_per_1000 = 3.0,
  Quarter_Mile_Time = 17.0,       # Valeur typique
  Engine_Shape = 0,             # Supposons 0 pour V-shape (valeur frequente)
  Transmission = 0,             # Supposons 0 pour automatique (valeur frequente)
  Forward_Gears = 3,            # Valeur typique pour automatique
  Carburetors = 2,              # Valeur frequente
  Displacement_log = log(200),
  Horsepower_log = log(150),
  Weight_log = log(3.0)
)

num_plus1_typique_m4 <- rbind(num, nouvelle_ligne_typique_m4_complet)

modele_plus1_typique_m4 <- lm(formula(model4), data = num_plus1_typique_m4)
summary(modele_plus1_typique_m4)

# Exemple 2: Creation d'une nouvelle observation aberrante
nouvelle_ligne_aberrante_m4_complet <- data.frame(
  Miles_per_Gallon = 10,
  Cylinders = 4,
  Displacement_cuin = 50,
  Horsepower = 500,
```

```
  Rear_Axle_Ratio = 4.5,        # Valeur plausible
  Weight_lb_per_1000 = 5.0,
  Quarter_Mile_Time = 12.0,       # Valeur plausible pour haute
    performance
  Engine_Shape = 1,             # Supposons 1 pour straight (moins
    frequent)
  Transmission = 1,             # Supposons 1 pour manuel (moins frequent
    )
  Forward_Gears = 5,            # Valeur plausible pour manuel
  Carburetors = 4,              # Valeur plausible pour haute performance
  Displacement_log = log(50),
  Horsepower_log = log(500),
  Weight_log = log(5.0)
)

num_plus1_aberrante_m4 <- rbind(num, nouvelle_ligne_aberrante_m4_complet
    )

modele_plus1_aberrante_m4 <- lm(formula(model4), data =
    num_plus1_aberrante_m4)
summary(modele_plus1_aberrante_m4)

# Exemple 3: creation d'une nouvelle observation dans l'espace etendu
    des variables
nouvelle_ligne_etendue_m4_complet <- data.frame(
  Miles_per_Gallon = 12,
  Cylinders = 8,
  Displacement_cuin = 450,
  Horsepower = 400,
  Rear_Axle_Ratio = 3.0,        # Valeur plausible
  Weight_lb_per_1000 = 6.0,
  Quarter_Mile_Time = 13.5,       # Valeur plausible pour grosse voiture
    puissante
  Engine_Shape = 0,             # V-shape
  Transmission = 0,             # Automatique
  Forward_Gears = 3,            # Automatique
  Carburetors = 4,              # Puissant
  Displacement_log = log(450),
  Horsepower_log = log(400),
  Weight_log = log(6.0)
)

num_plus1_etendue_m4 <- rbind(num, nouvelle_ligne_etendue_m4_complet)

modele_plus1_etendue_m4 <- lm(formula(model4), data =
    num_plus1_etendue_m4)
summary(modele_plus1_etendue_m4)

#ACP

# --- 1. Preparation des donnees pour l'ACP ---

# Nom des variables numeriques (sans la variable cible potentielle)
noms_variables_numeriques <- c("Cylinders", "Displacement_cuin", "
    Horsepower",
```

```
                                "Rear_Axle_Ratio", "Weight_lb_per_1000",
    "Quarter_Mile_Time",
                                "Engine_Shape", "Transmission", "
    Forward_Gears", "Carburetors") # Liste des noms des variables pour l'
    ACP

variables_acp <- num_no_log[, noms_variables_numeriques]

# Standardisation des donnees (important pour l'ACP avec des echelles
    differentes)
variables_standardisees <- scale(variables_acp)

# --- 2. Realisation de l'Analyse en Composantes Principales (ACP) ---

# Effectuer l'ACP
acp_result <- prcomp(variables_standardisees, center = FALSE, scale =
    FALSE)

# --- 3. Exploration des resultats de l'ACP ---

# 3.1. Resume de l'ACP (variance expliquee, importance des composantes,
    etc.)
summary(acp_result)

# 3.2. Affichage des chargements (loadings) des composantes principales
print(acp_result$rotation)

# 3.3. Scree Plot (graphique des valeurs propres)
plot(acp_result$sdev^2, type = "b",
      main = "Scree Plot",
      xlab = "Composante Principale",
      ylab = "Variance Expliquee")

# 3.4. Biplot (visualisation des individus et des variables sur les deux
     premieres CP)
biplot(acp_result, choices = 1:2, cex = 0.8,
       main = "Biplot des deux premieres Composantes Principales")

# --- 4. Analyse des scores des composantes principales et Clustering
    ---

# 4.1. Extraction des scores des deux premieres composantes principales
scores_acp <- as.data.frame(acp_result$x[, 1:2])
colnames(scores_acp) <- c("Comp.1", "Comp.2")

# 4.2. Clustering
# Il est important de choisir un nombre de clusters approprie en amont
set.seed(123)
kmeans_result <- kmeans(variables_standardisees, centers = 3)
scores_acp$cluster <- as.factor(kmeans_result$cluster)

# 4.4. Visualisation des clusters dans l'espace PCA avec ggplot2
ggplot(scores_acp, aes(x = Comp.1, y = Comp.2)) +
  theme_classic() +
  geom_hline(yintercept = 0, color = "gray70") +
  geom_vline(xintercept = 0, color = "gray70") +
```

```
  geom_point(aes(color = cluster), alpha = 0.55, size = 3) +
  xlab("Premiere Composante Principale (PC1)") +
  ylab("Deuxieme Composante Principale (PC2)") +
  # xlim(-5, 6) + # Optionnel : ajuster les limites de l'axe x
  ggtitle("Projection des Voitures sur les deux premieres Composantes Principales (Clusters)") +
  scale_color_discrete(name = "Cluster")

# --- 5. Utilisation de la librairie factoextra pour des visualisations avancees ---

install.packages("factoextra")
library(factoextra)

# 5.1. Scree Plot avec les valeurs sur les barres
fviz_eig(acp_result,
         main = "Scree Plot",
         xlab = "Composante Principale",
         ylab = "Variance Expliquee (%)",
         addlabels = TRUE)

# 5.2. Cercle de correlation
fviz_pca_var(acp_result,
             col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,      # Eviter le chevauchement des etiquettes
             title = "Cercle de Correlation des Variables")



# --- 6. Regression Lineaire sur les Composantes Principales ---

# Preparation des donnees :
X <- variables_standardisees # Donnees standardisees
y <- num$Miles_per_Gallon # Variable cible

# Creer un dataframe des scores de toutes les composantes principales
scores_cp <- as.data.frame(acp_result$x)
colnames(scores_cp) <- paste0("Comp.", 1:ncol(scores_cp))

# 6.1. Modele de RCP avec toutes les composantes principales
modele_rcp_complet <- lm(y ~ ., data = scores_cp)
summary(modele_rcp_complet)

# 6.2. Modele de RCP avec les deux premieres composantes principales
modele_rcp_2cp <- lm(y ~ Comp.1 + Comp.2, data = scores_cp)
summary(modele_rcp_2cp)

# 6.3. Visualisation de la relation entre la variable cible et les premieres CP
ggplot(scores_cp, aes(x = Comp.1, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Miles per Gallon vs. Premiere Composante Principale (PC1)",
       x = "Premiere Composante Principale (PC1)",
```

```
       y = "Miles per Gallon") +
  theme_minimal()

ggplot(scores_cp, aes(x = Comp.2, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Miles per Gallon vs. Deuxieme Composante Principale (PC2)",
       x = "Deuxieme Composante Principale (PC2)",
       y = "Miles per Gallon") +
  theme_minimal()
```