Multivariate data analysis project

Data Analysis

Degree in data science and engineering

UPC

David González

Angel Morales

David Torrecilla

May 2025

Table of contents

1. Introduction

This document is the report of the study of *1985 Auto Imports Database*, obtained from the UC Irvine Machine Learning Repository. It is a dataset with 26 variables which describe some properties of cars, from physical properties, (such as the height or engine-location) efficiency and performance parameters (like city-mpg, horsepower...) and even price and a variable called symboling, a categorical variable which tells the risk given by an insurance company to that car.

The data seems to be Originally created by Jeffrey C. Schlimmer, using data from 1985 Ward's Automotive Yearbook, 1985 Model Import Car and Truck Specifications, Personal Auto Manuals, Insurance Services Office and Insurance Collision Report, Insurance Institute for Highway Safety.

The analysis has consisted on preprocessing the data, performing descriptive statistics, applying principal component analysis, multidimensional scaling, correspondence analysis, clustering, discriminant analysis and MANOVA, with the goal of reducing the dimensionality of the dataset and discovering relations between the variables and the individuals themselves, that allow a deep comprehension of what is the information that this dataset hides. R language has been used to perform the analyses.

Some general variables of the dataset are the *symboling* (categorical), the *normalized-losses* that the insurance company estimates (in $), the *price* of the car (in $) and the *make*: the manufacturer (categorical). There are some variables related to the shape and weight of the car: *num-of-doors* and *body-style* (both categorical), *length*, *wheel-base*, *width* and *height* (all four in inches) and *curb-weight* (the weight without passengers nor personal belongings, in pounds).

The rest of the variables refer to efficiency, performance and technical aspects: *fuel-type, fuel-system, aspiration, drive-wheels, engine-location, engine-type* and *num-of-cylinders,* which are categorical, and *engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg* and *highway-mpg*, which are continuous.

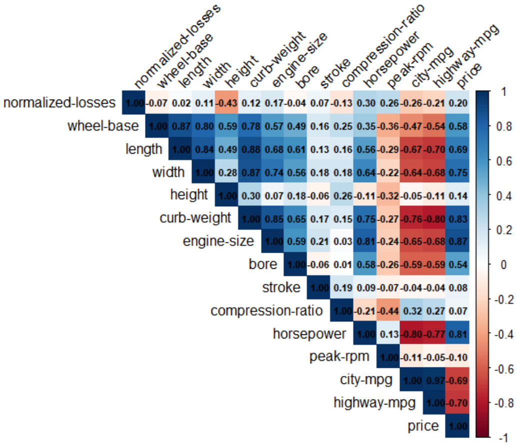
We will take an in-depth look to all variables in the results.

2. Results

Descriptive statistics

The first thing to do is descriptive statistics and preprocessing. After loading the dataset to R, the number of individuals (cars) and variables is checked. There are 205 cars and 26 variables. The scale of the numerical variables and the number of individuals of each group of the categorical ones give an idea of the next steps.

The variable *engine-location* shows to be unnecessary, because there are only 2 classes: front and rear, and only 3 cars are in the rear class, which we can identify: Porsche brand cars with hardtop and convertible body types are the ones which have rear engines. This means that the information of *engine-location* is contained inside those other variables (*make* and *body-type*, and the variable is erased. No other variable shows to be unnecessary at this point, so the correlation is studied, to get an idea of some ways to reduce dimensionality.



	normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
normalized-losses	1.00	-0.07	0.02	0.11	-0.43	0.12	0.17	-0.04	0.07	-0.13	0.30	0.26	-0.26	-0.21	0.20
wheel-base		1.00	0.87	0.80	0.59	0.78	0.57	0.49	0.16	0.25	0.35	-0.36	-0.47	-0.54	0.58
length			1.00	0.84	0.49	0.88	0.68	0.61	0.13	0.16	0.56	-0.29	-0.67	-0.70	0.69
width				1.00	0.28	0.87	0.74	0.56	0.18	0.18	0.64	-0.22	-0.64	-0.68	0.75
height					1.00	0.30	0.07	0.18	-0.06	0.26	-0.11	-0.32	-0.05	-0.11	0.14
curb-weight						1.00	0.85	0.65	0.17	0.15	0.75	-0.27	-0.76	-0.80	0.83
engine-size							1.00	0.59	0.21	0.03	0.81	-0.24	-0.65	-0.68	0.87
bore								1.00	-0.06	0.01	0.58	-0.26	-0.59	-0.59	0.54
stroke									1.00	0.19	0.09	-0.07	-0.04	-0.04	0.08
compression-ratio										1.00	-0.21	-0.44	0.32	0.27	0.07
horsepower											1.00	0.13	-0.80	-0.77	0.81
peak-rpm												1.00	-0.11	-0.05	-0.10
city-mpg													1.00	0.97	-0.69
highway-mpg														1.00	-0.70
price															1.00

Figure 1 - Correlation matrix of all the original numerical variables

The correlation matrix reveals what we could suspect: some variables share information. It is necessary to check this further before attempting to join any variables.

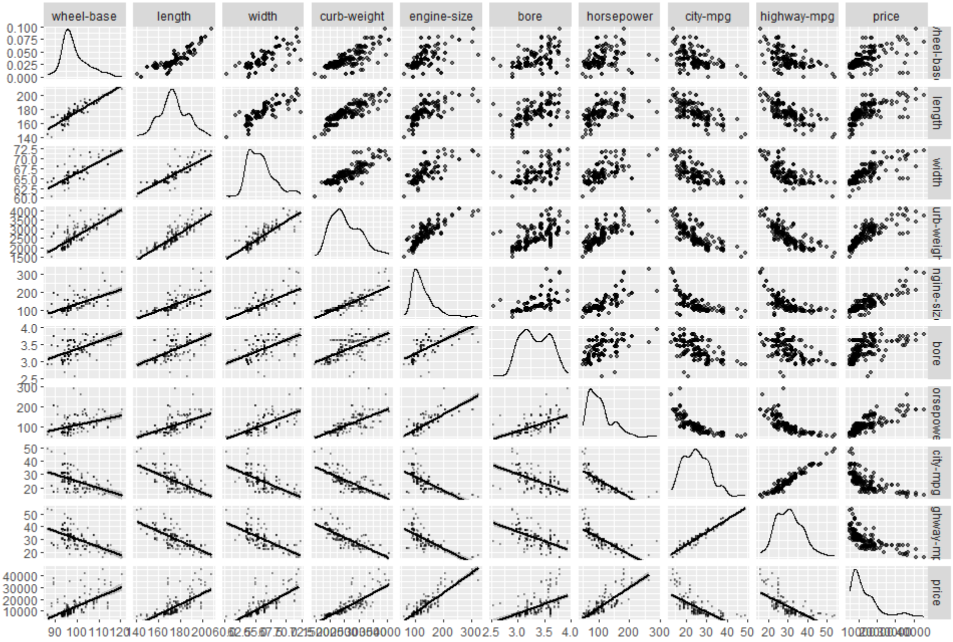


Figure 2 - Scatterplot matrix of some numerical variables

The scatterplot matrix reveals the relations between the variables. Although they are not fully linearly related, it seems necessary to join them in some way, otherwise the information that we share will take too much importance when doing further analysis, and will hide the information given by the rest of the variables of the dataset. But before applying dimensionality reduction techniques, there are some things to be done.

Considering that *wheel-base* is the distance between front and rear wheels, its correlation with the *length*, and the fact that its behaviour with other variables is very similar to *length's* but sparser, we have decided to erase the first variable, as it does not give any information that *length* does not, and it is not as intuitive. We have considered this variable to be unnecessary.

By using different metrics applied on categorical variables, the relation between those and with numerical variables can be studied.

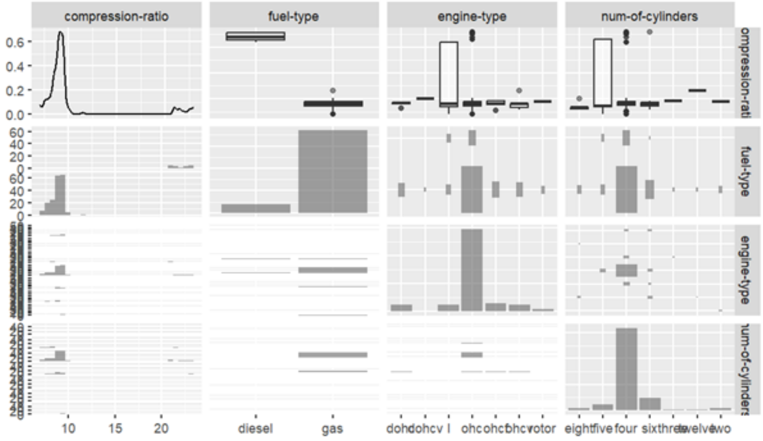


Figure 3 - Scatterplot matrix of some categorical and numerical variables

Fuel-type, which has two classes (diesel and gas) is very related to the *compression ratio*, which clearly has two clusters (one around 0.6 and one around 0.1). These two clusters are clearly related to the *fuel-type* and allow us to predict it. This means that *fuel-type* is redundant, and should be removed.

The next step is to treat missing values.

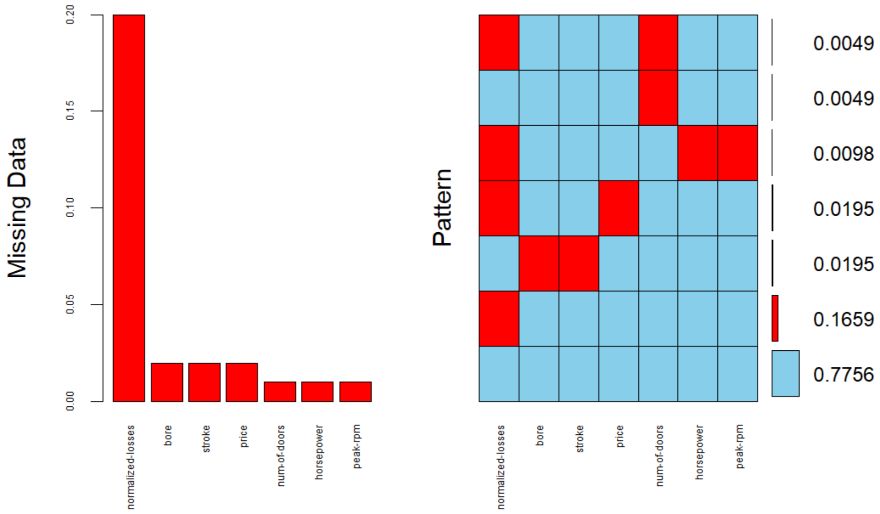


Figure 4 - Missing values per variable (left) and by combination of variables (right)

There are some missing values, mainly for the *normalized-losses*. Because erasing a 20% of variables is too much, and because we have a lot of other variables that allow prediction techniques, we have used k Nearest Neighbors imputation to generate the missing values.

Right after the imputation, *city-mpg* and *highway-mpg* have been joined into *combined-mpg* using USA EPA standard formula: combined = 0.55 city + 0.45 highway. This has been done because of the very high linear correlation of these two variables and their similar meaning.

It is very important to check the normality of the variables and the outliers. Doing those in this same order results in having less observations considered as outliers, if the data is not gaussian, which is the case. For example, we can look at the price.

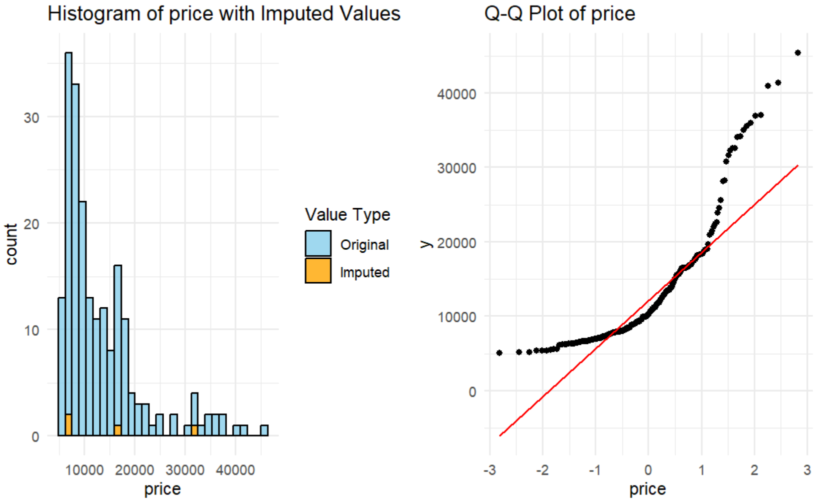


Figure 5 - Histogram of price (left) and Q-Q plot of price (right)

The histogram and Q-Q plot are enough to see that price is not gaussian.

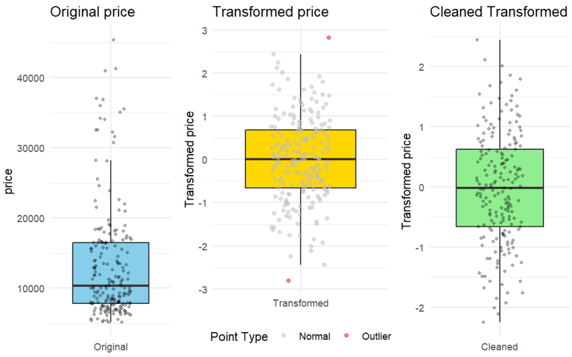


Figure 7 - From left to right: boxplots of: original price, transformed price and cleaned transformed price

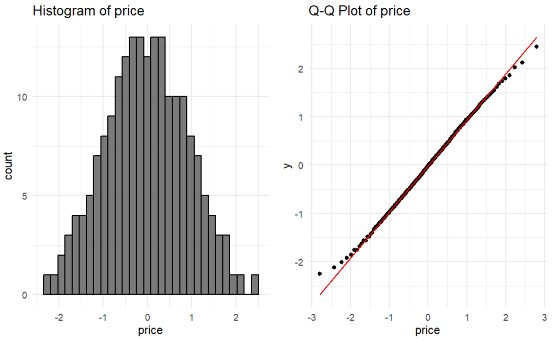


Figure 6 - Histogram of price (left) and Q-Q plot of price (right) after applying the transformation

The package bestNormalize automatically transforms the data with a suitable method (such as Box-Cox, the most known one). The transformed data has less outliers than the original, and those are the ones cleaned. Because there are very few, we have decided to erase the observations which contain any outlier after applying the transformation.

Dimensionality reduction

The group of variables found before is perfect to apply principal component analysis or factor analysis to reduce the dimensionality, because all those variables are numerical. These are *length, width, curb-weight, engine-size, bore, horsepower, price,* and *combined-mpg*.

In this case, because we do not want to search for latent relations, as it seems trivial that larger cars will be wider and heavier, and big engines will be more powerful and pricey but consume more fuel (among other predictable relationships), we will not focus on understanding this relation, but on reducing all those variables to one or two while keeping most of their variation. It is convenient to apply PCA. It has been applied on the transformed variables.

As expected, these eight variables can be joined into one or two. By using the common criteria, which are taking at least an 80% variance, taking as much components as eigenvalues

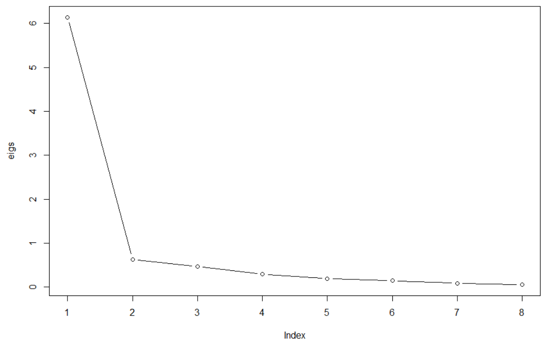


Figure 8 – Eigenvalues of the correlation matrix of the 8 high-correlated variables

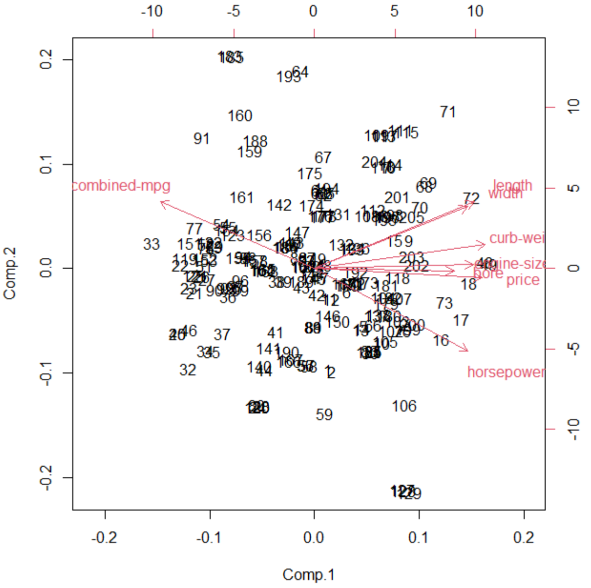


Figure 9 - Biplot of the PCA of the 8 high-correlated variables

greater than 1, or taking all components up to the elbow point, it is fair to say that one dimension is enough (although it only explains a 76% of the variation), given our goal of reduce these high-correlated variables as much as possible, to allow further study with the other fifteen variables of the dataset.

The resulting component will be used in the dataset as SWaPC, which is used as an acronym for Size, Weight, Power and Cost in some industries. This is not related to car manufacturing, though it is a common goal to minimize those as a good practice. In our case, minimizing the SWaPC will mean having better efficiency (less mpg). We will be able to understand which other variables (such as car companies or body types) are related to a lower SWaPC.

Multiple Correspondence Analysis

MCA is used instead of CA because the dataset includes several categorical variables, not just a two-way contingency table. While CA is limited to two categorical variables, MCA allows us to explore relationships across multiple categorical dimensions simultaneously.

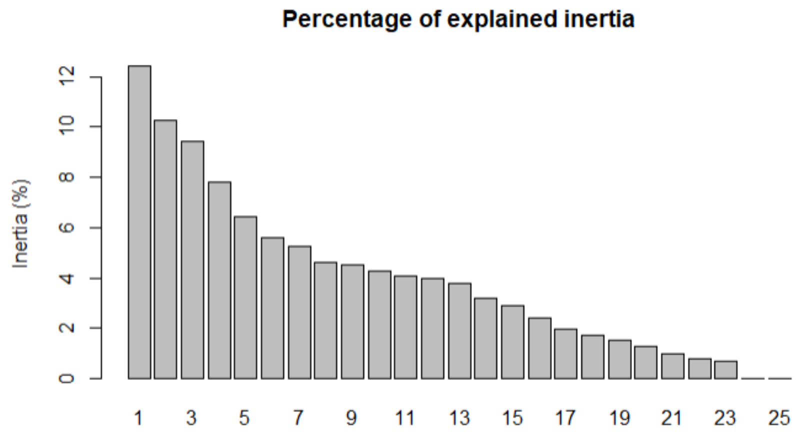


Figure 10 – Percentage of explained inertia

The first two dimensions explain approximately 22.67% of the total inertia (12.42% and 10.25%, respectively). Although this proportion may seem relatively low, it is common in MCA due to the high number of categories across multiple variables. These two dimensions are still useful for visualizing the main associations among the categorical variables in a reduced space.

Several categories show strong contributions and high cos² values in Dimension 1, such as "rotor", "4bbl", and "num-of-cylinders_two", indicating that these are highly distinctive configurations in the dataset. In Dimension 2, the categories "num-of-cylinders_eight" and "ohcv" stand out due to their strong contribution and high quality of representation.

Categories such as "body-style_hatchback" and "drive-wheels_rwd" also contribute notably to both dimensions, supporting their relevance in differentiating car profiles. Conversely, categories with low cos² values, such as "mpfi" or "ohc", are less informative in the first dimensions and contribute more weakly to the overall structure.

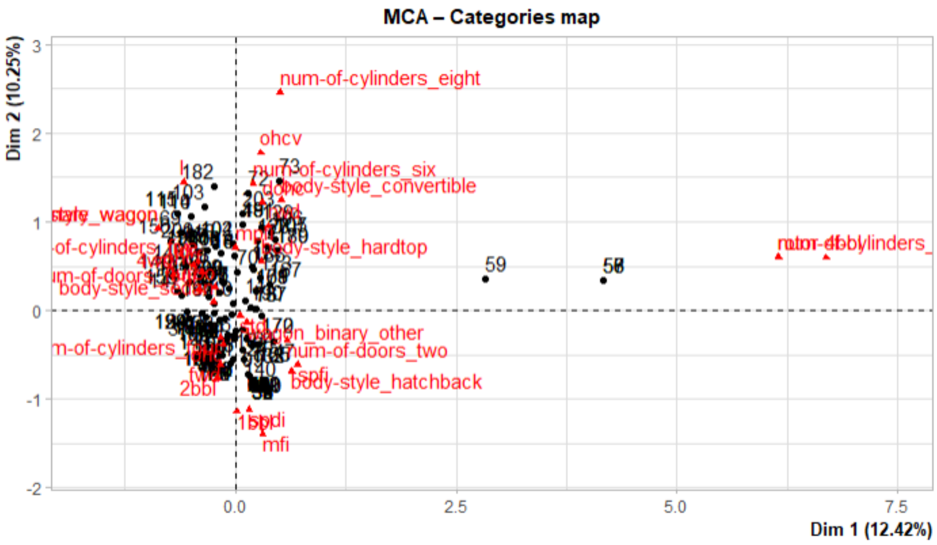


Figure 11 – MCA categories map

The factor map displays the variable categories projected onto the two-dimensional space defined by the MCA. Dimension 1 (12.42% of variance) mainly contrasts rare engine configurations such as "rotor" or "4bbl" and "num-of-cylinders_two" (far right), with more common or moderate setups clustered near the origin. Dimension 2 (10.25%) differentiates categories like "num-of-cylinders_eight" and "ohcv" (top), which are associated with more powerful engines, from simpler configurations like "1bbl" or "mfi" (bottom).

Categories located near the origin, such as "ohc" or "body-style_sedan", represent more typical profiles and contribute less to the variance. In contrast, those far from the center—especially "rotor" and "4bbl"—highlight unique or extreme configurations that help define the factor structure. This spatial arrangement allows us to detect which categorical features distinguish certain car types and which ones reflect standard industry configurations.

The coordinates, contributions, and squared cosine values (cos²) of the individuals were also analyzed. Most cars show low contributions and moderate cos² values, as expected in a dense central cluster. However, a few individuals, such as 58 and 59, exhibit high contributions to Dimension 1 and high cos², confirming their role as outliers with atypical categorical profiles. These observations help define the structure of the first dimension and provide insight into rare vehicle configurations.

The map of individuals shows that the majority of cars are densely clustered near the origin, indicating typical and frequent combinations of categorical attributes. However, a few cars—such as individuals 58, 59, and 73—are clearly separated from the center, suggesting they possess rare configurations (e.g., unusual engine types or drive systems). These outliers contribute significantly to the variance captured by the first two dimensions. While no clear clusters are visible, the spread along Dimension 1 reflects variation in engine and transmission features, whereas Dimension 2 seems to differentiate less common design elements such as body style or cylinder count.

Hierarchical clustering

To apply hierarchical clustering, it is necessary to compute the pairwise dissimilarities between observations. We used the Euclidean distance on the scaled version of the numerical variables to ensure that all features contribute equally to the clustering process. This distance matrix serves as the input for the hierarchical clustering algorithms.

We then applied various hierarchical clustering methods to the distance matrix and visualized the resulting dendrograms to compare the clustering structures produced by each linkage strategy.

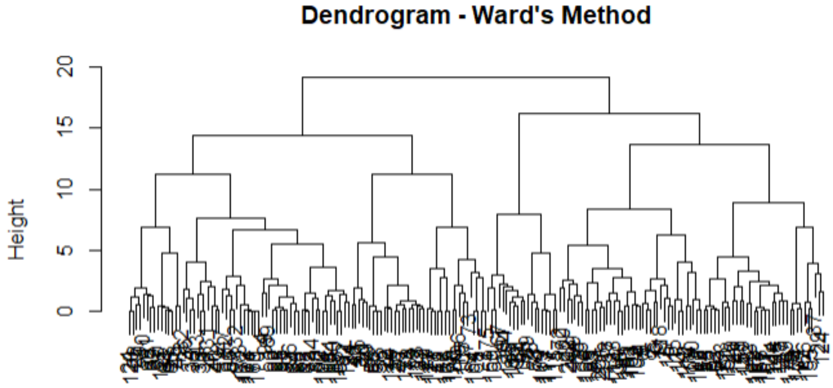


Figure 12 – Dendrogram obtained using Ward's method

Among the five linkage methods tested (Ward's, single, complete, average, and centroid), Ward's method was selected as the most interpretable for this dataset. Its dendrogram displayed compact and well-separated branches, suggesting clearly defined clusters. In contrast, methods like single linkage showed significant chaining effects, while centroid and average linkages produced less distinguishable groupings. Ward's method minimizes the total within-cluster variance, which aligns well with the structure of the scaled dataset and justifies the choice of k = 4 or k = 5.

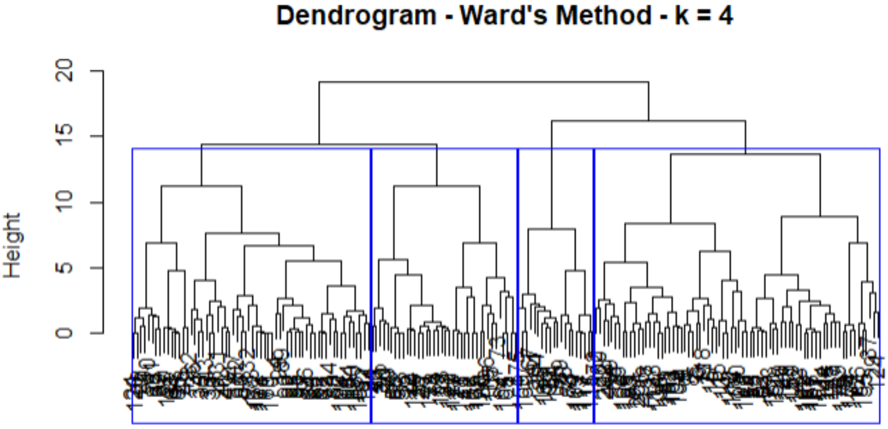


Figure 13 – Dendrogram with 4 clusters defined using Ward's method

Both k = 4 and k = 5 provide meaningful clusterings with slight differences in group composition. We will further evaluate their interpretability and cohesion using visual inspection and summary statistics.

To support the choice of the optimal number of clusters, we compute the Pseudo-F index for different values of k. This index evaluates the ratio of between-cluster variance to within-cluster variance, helping to identify the number of clusters that best separates the data.

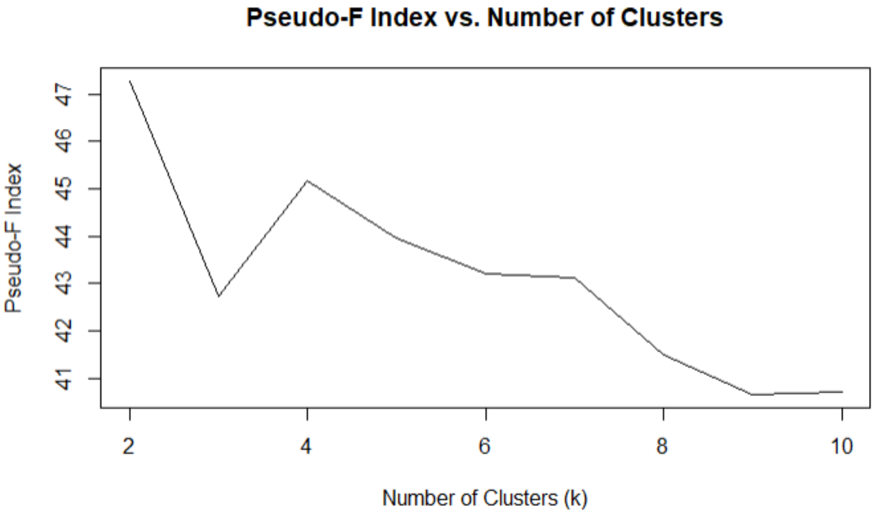


Figure 14 – Pseudo-F index for different numbers of clusters

The Pseudo-F index reaches a maximum at k=4, indicating the best trade-off between between-cluster and within-cluster variance. This peak suggests that the data structure is most clearly separated into four distinct groups. Therefore, we select k=4 as the optimal number of clusters for further analysis.

To further analyze the structure of the data, we apply the k-means clustering algorithm using k=4, as suggested by the dendrogram and confirmed by the peak in the Pseudo-F index. To interpret the k-means clustering results, we examine the average values of each variable within the four clusters. This allows us to identify the key characteristics that differentiate groups of vehicles based on their technical specifications and performance-related attributes.

The table of the numerical variables within each of the four k-means clusters. Cluster 1 is characterized by high compression ratios and stroke values, but low peak RPM, possibly indicating vehicles with more powerful engines operating at lower speeds. Cluster 3, by contrast, displays lower-than-average height and stroke, but higher RPM, which may correspond to lighter or sportier vehicles. These profiles suggest distinct groups of vehicles based on their technical specifications and performance characteristics.

Now, we perform a Principal Component Analysis (PCA) on the selected variables to reduce dimensionality while retaining most of the variance in the data. Then, we apply hierarchical clustering on the PCA scores to obtain an alternative clustering solution and compare it with previous results.

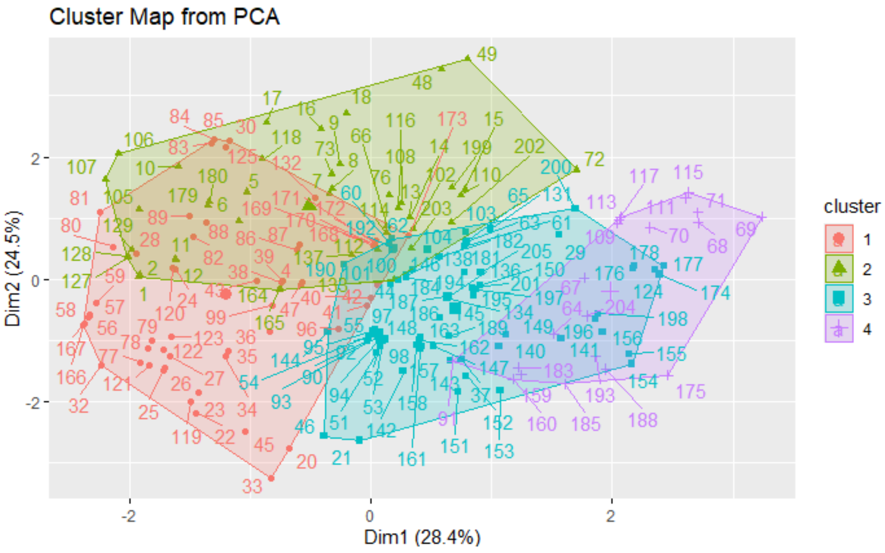


Figure 15 – Cluster visualization on the first two PCA dimensions

The HCPC method applied to the principal component scores revealed four clusters of vehicles. The first two principal components account for approximately 52.9% of the total variance, providing a partial yet informative projection of the data. Some overlap between clusters is observed in the 2D PCA factor map, which is expected due to the loss of information inherent to dimensionality reduction. Despite this visual superposition, the hierarchical clustering captures meaningful groupings of vehicles with distinct profiles.

- Cluster 1 (red) includes vehicles with low engine sizes, horsepower, and price, likely representing compact or economy cars aimed at affordability and fuel efficiency.
- Cluster 2 (green) features cars with moderate specifications across most dimensions, forming a balanced middle segment of the market.
- Cluster 3 (blue) stands out with higher values in engine performance variables such as horsepower, RPM, and compression ratio, suggesting a group of high-performance or sports cars.
- Cluster 4 (purple) appears to group heavier vehicles with higher curb weight and larger dimensions, possibly corresponding to luxury sedans or SUVs.

These clusters reflect meaningful differences in vehicle design and performance strategy, and provide valuable insights for segmenting the automotive market.

Between the two hierarchical clustering approaches—directly on the scaled numerical data using Ward's method, and on the principal component scores using HCPC—both produced a consistent solution with k=4 clusters and showed similar dendrogram structures and cluster separation. However, we choose the HCPC-based approach as the preferred method. This is because HCPC combines dimensionality reduction with clustering, helping to denoise the data and reduce the influence of multicollinearity among variables. Moreover, it provides clearer interpretability through principal components, which summarize the underlying structure of the dataset more efficiently than the raw variables.

Multidimensional Scaling

Multidimensional Scaling (MDS) is a statistical technique used to visualize the similarities or dissimilarities between observations in a dataset. The basic idea is to compute a distance value between all observations and to represent them in a low-dimensional space while preserving the distances as much as possible. Observations with smaller distance values should therefore appear close together in this new space.

Since our dataset contains both categorical and numerical variables, we used the Gower distance metric, which handles mixed variable types effectively. This choice ensures that the dissimilarities in the final projection accurately reflect differences across all variable types.

After applying MDS, we obtained a scatter plot where each point represents an observation. To extract meaningful insights, we started by labeling the points according to various categorical variables. This approach allows us to see how these variables contribute to the overall structure. Some of those plots are shown here:

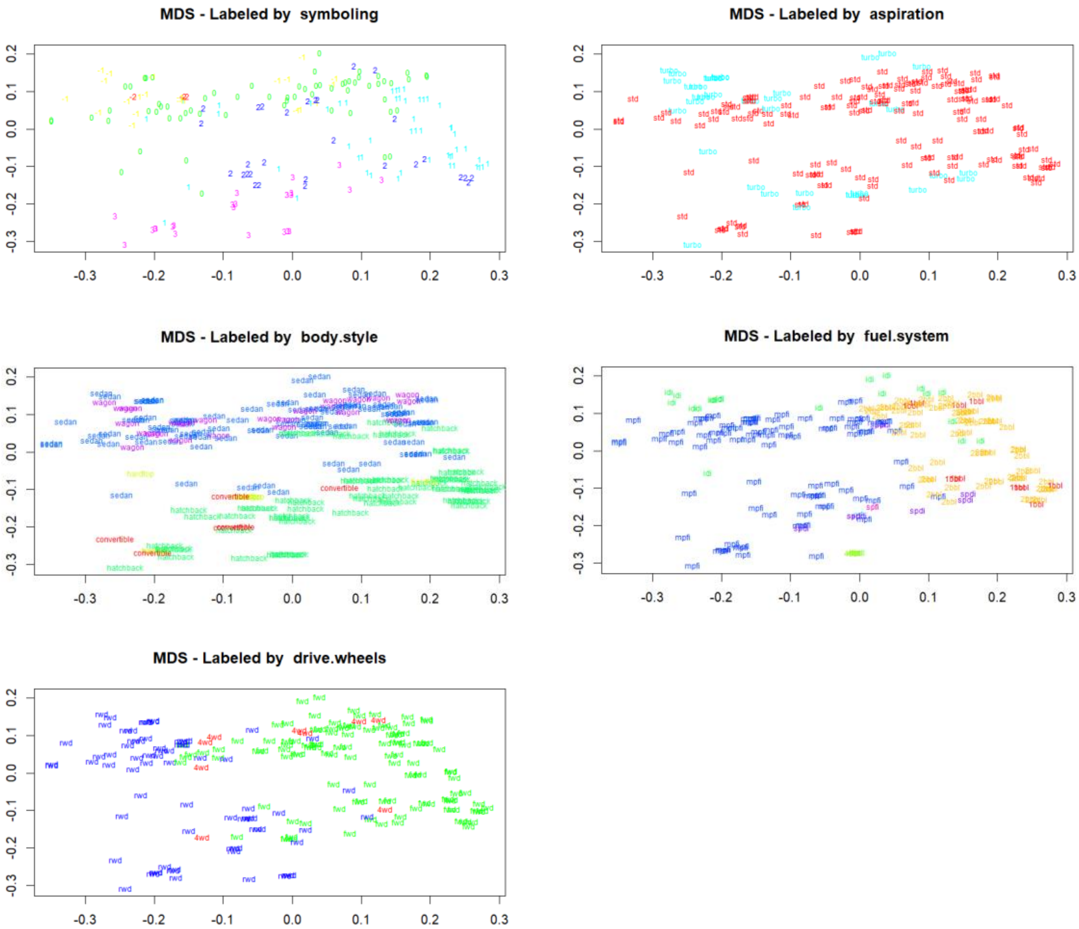
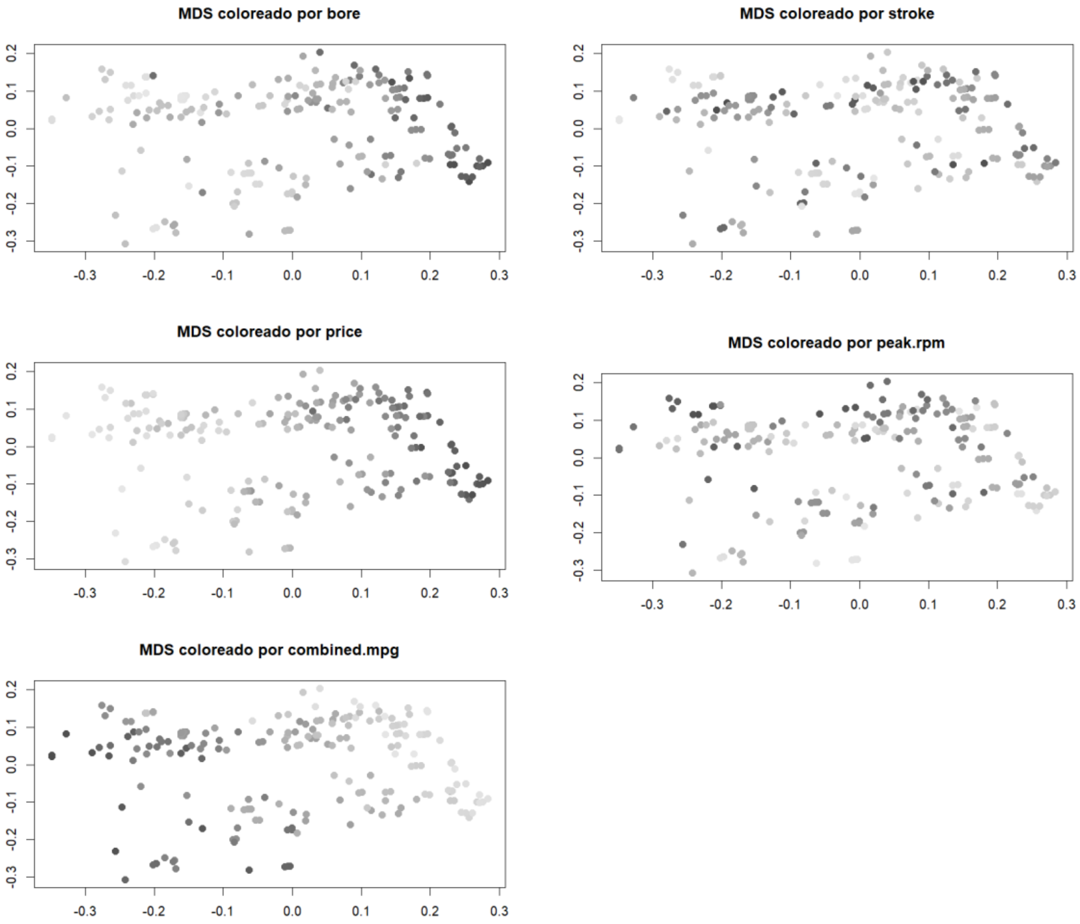


Figure 16 – MDS plots labeled for different categorical variables

Most of the categorical variables produce coherent clusters in the plot. Variables like "num-doors," "body-style," or "fuel-system" present plots where most clusters can be easily identified within specific areas of the plot, although these areas can sometimes overlap and include some outlier value. This suggests that these variables are strongly related to the rest of the data.

On the other hand, variables like "aspiration" do not produce clear groupings, and certain categories such as "4wd" for "drive-wheels" or the "hardtop" body style appear somewhat arbitrary in their distribution. This may indicate that these variables have weaker relationships with the rest of the dataset.

In addition to categorical labeling, we also explored the data using color gradients to represent the values of numerical variables. Darker colors represent lower values, while lighter colors represent higher values. In these plots, we focus less on clusters and more on the presence of color gradients. We can see some of the plots here:



MDS coloreado por bore

MDS coloreado por stroke

MDS coloreado por price

MDS coloreado por peak.rpm

MDS coloreado por combined.mpg

Figure 17 – MDS colored by numerical variables

Variables such as "bore," "horsepower," "price," and "combined.mpg" display very clear gradients, suggesting they have a strong correlation with the rest of the dataset. In contrast, variables like "stroke" and "peak-rpm" show scattered color patterns, indicating weaker relationships.

Interestingly, many variables exhibit higher variance along the horizontal axis, meaning their values change significantly from left to right. Variables such as “drive-wheels,” “fuel-system,” “engine-type,” “num-of-cylinders,” “bore,” “horsepower,” “price,” and “combined.mpg” show this trend most clearly. This suggests that the horizontal axis represents a general “size” or “magnitude” dimension for the cars: larger vehicles, with more robust engines and higher prices, tend to cluster on one side, while smaller, simpler cars cluster on the other.

Conversely, variables like “num-doors” and “body-style” show more variance along the vertical axis. It is less clear what this vertical axis represents, but it could be related to the shape or general design of the car. For instance, hatchbacks, which typically have two doors, appear lower on the plot, while sedans and wagons, usually with four doors, tend to appear higher.

Finally, we can visualize the MDS projection using car make as the labeling variable.

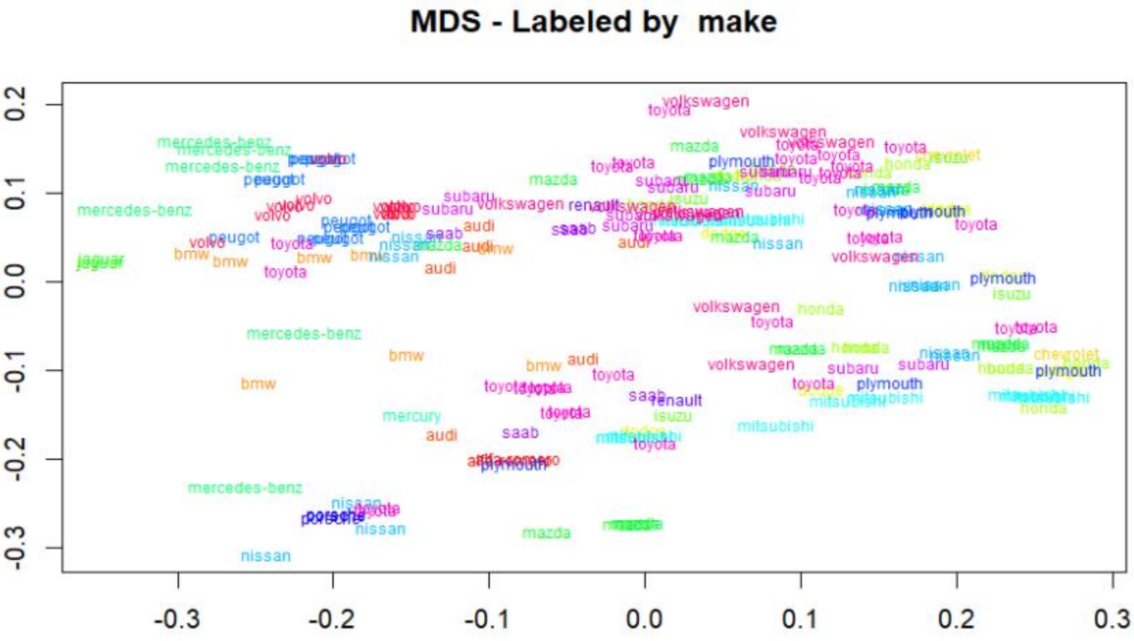


Figure 18 – MDS labeled by make

At first glance, the plot may appear somewhat arbitrary, since each brand typically produces a wide range of car models that do not necessarily resemble one another. As a result, the makes do not form clearly separated clusters. However, a closer inspection reveals several interesting patterns.

- For instance, Mercedes-Benz models tend to cluster toward the upper left of the plot. This suggests that these vehicles are generally larger, more powerful, and more expensive compared to others. They also tend to share features such as having four doors and being either sedans or wagons, which contributes to their similarity in the MDS space.

- Other brands like BMW, Peugeot, and Audi show a similar trend, although their clusters are less distinct and more centrally located. This dispersion likely reflects a broader variety of models within each brand, indicating a balance between premium features and broader market reach.

- Brands such as Subaru, Toyota, Nissan, and Mazda appear to be more widely scattered across the MDS space. This distribution suggests that these manufacturers produce a more diverse range of vehicles, varying significantly in size, configuration, and price.

- In contrast, brands like Honda, Mitsubishi, Volkswagen, or Plymouth are more concentrated toward the upper right. This region of the MDS space seems to be associated with more affordable, compact, and family-oriented vehicles.
- Finally, premium sports brands like Jaguar or Porsche are represented by very few observations that cluster tightly together. This indicates that their models are highly similar to one another, reinforcing the idea of a consistent, niche design focus.

While not all makes form clearly defined groups, this visualization still offers a rich way to explore brand-level patterns. With careful examination, many subtle relationships can be discovered.

MANOVA and discriminant analysis

In order to understand some relationships between our variables, MANOVA has been used to try to find differences between the classes of two categorical variables: num-of-doors (binary) and body-style (multiclass). Specifically, it has been checked weather the means are equal for some of our continuous variables.

This analysis assumes that our continuous variables are gaussian, and also assumes that the covariances are equal for the different classes. So, the first steps have been to check normality and the covariance. By looking at the histograms and Q-Q plots, it's fair to say that SWaPC and normalized-losses look gaussian (after applying a transformation on SWaPC, to fix the tails of the Q-Q plot). Height, stroke compression-ratio and peak-rpm do not have as much of a good behavior, but they are acceptable. The Shapiro-Wilk test considers that, as we suggested, SWaPC and normalized-losses are gaussian, as well as the stroke. It rejects normality for the compression-ratio and the height, so those are the less important variables for our next analyses.

The Box’s M-test for Homogeneity of Covariance Matrices fails applied on all those continuous variables, for num-of-doors as well as for body-style. It has been necessary to reduce the input variables. For num-of-doors, the test does not reject homogeneity when using the SWaPC, the normalized-losses and the stroke. For the body-style, the same variables, excluding the stroke. So in the end we are not going to work with the variables which did not look quite gaussian and failed the test; we are working with variables that can fairly be considered gaussian.

Those are the variables to use for the MANOVA. After fitting the model, the Wilks, Pillai, Hotelling-Lawley and Roy tests are performed, and all of them agree that the combined dependent variables significantly differ across the different classes of num-of-doors (all the p-values are extremely low), and the same result for the MANOVA on body-style. The ANOVA is used to detect which variables exactly have different means depending on the class of the categorical variables.

About num-of-doors, ANOVA reveals that there is strong evidence that this variable affects the mean of the SWaPC and has a significant effect on the normalized-losses. There is no evidence of significant effect of num-of-doors on the stroke variable. This results match with the expected behavior: cars with more doors (4 instead of 2) will tend to be longer and heavier, so have a greater SWaPC.

About body-style, ANOVA reveals that this variable has a strong effect on both the SWaPC and the normalized-losses, too. Again, a wagon or a sedan will be longer and heavier than a hatchback or a convertible, for example, so this results make sense. In this case, we can go deeper. Tukey has been performed to understand the effect of the body style on the SWaPC and normalized-losses (independently) to see which categories differ significantly. It is found that SWaPC has statistically different means between two pairs: sedan-hatchback and wagon-hatchback; and normalized-losses has statistically different means between four pairs: wagon-convertible, wagon-hardtop, wagon-hatchback and wagon-sedan.

MANOVA and ANOVA have allowed to see that hatchback seems the most differentiable class in terms of SWaPC, and wagon in terms of normalized-losses (in fact wagon should be totally distinguishable from the rest). This information is very important to be able to select the proper variables for a linear discriminant analysis.

A LDA model has been fit to predict num-of-doors (2 or 4; binary classification) depending on the SWaPC and normalized-losses. The resulting model has aa 70% of correct classification rate, which is a regular performance. It is necessary to check some other metrics. The CCR across the 2 categories is 78% for the 4 doors, and 56% for the 2 doors, so we can see that our model has problems predicting one of the classes. The prior probabilities sum of squares is similar (a bit lower) than the posterior probabilities sum of squares (0.517 vs 0.514), although checking the posteriors individually, we see that most of the rows are very different from a 50 50. We can conclude that our model is fairly better than random guessing, although not extremely good. A train / test split has been used to train the model and reassure its performance, and the test accuracy has been of a 69%, similar to the test accuracy.

Table 1 - Confusion matrix for the train data; LDA on num-of-doors

Actual > Predicted v	4 doors	2 doors
4 doors	64	23
2 doors	14	34

Table 2 - Confusion matrix for the test data; LDA on num-of-doors

Actual > Predicted v	4 doors	2 doors
4 doors	26	7
2 doors	11	15

Table 3 - Group mean son the transformed variables

Group means	SWaPC	normalized-losses
4 doors	0.16	-0.37
2 doors	-0.23	0.42

The group means, although they are not intuitive because they are on the transformed variable, allow to see how cars with more doors tend to have more SWaPC (as predicted) but less normalized losses than cars with two doors.

A more complex problem has been tried to solve using quadratic discriminant analysis: predict the body-style (hatchback, sedan, sportscar, wagon) using the six continuous variables (SWaPC, normalized-losses, stroke, compression-ratio, height and peak-rpm). Notice that it has been necessary to join the categories convertible and hardtop into sportscar, because those categories had too little observations in our train split to fit a QDA model. Also, it is true that not all convertible and hardtop cars are sportscars, but this allows to make an intuitive category without placing hardtop inside of convertible.

The QDA has a considerable performance for a multiclass problem. The correct classification rate is a 77% for the train split and a 64% for the test split, which indicates some overfitting (our model is taking too much into account each of the observations of the dataset).

The sum of squares of the posteriors is 0.706, much higher than the priors' 0.335. This model is not performing bad, though the main problem is having too few samples for some groups (sportscar has 13 samples).

Table 4 - Confusion matrix for the train data;
Table 5 - Confusion matrix for the test data
QDA on body-style
QDA on body-style

Actual > Predicted v	hatchback	sedan	sportscar	wagon
hatchback	37	13	0	0
sedan	7	45	2	6
sportscar	1	1	9	0
wagon	0	1	0	13

Actual > Predicted v	hatchback	sedan	sportscar	wagon
hatchback	12	9	0	0
sedan	4	20	0	2
sportscar	1	0	2	0
wagon	0	5	0	4

The confusion matrices allow to see how the QDA easily predicts sedans as hatchbacks, and sometimes wagons and sedans. Nevertheless, this is not a bad model, because body styles do not follow strict rules or thresholds on size, weight, height or any other variable of our dataset, instead, they follow some styling standards such as the shape of the front or the back, aspects that we do not know about the cars of this dataset.

3. Conclusions

This dataset contains some diverse but precise information about these cars. It is clear that some variables share a lot of information, and it is convenient to reduce the dimensionality to do a proper analysis. Also, some variables have all their information contained in others and are redundant.

Multiple Correspondence Analysis (MCA) was applied to examine the relationships among several categorical variables in the dataset. The method effectively reduced the dimensionality of the data while preserving relevant associations between categories and individuals. Through the analysis of contributions and quality of representation ($\cos^2$), we identified which variables were most influential in defining the primary dimensions. The resulting factor maps revealed both common and distinctive patterns across the dataset, demonstrating MCA's utility in uncovering latent structures within categorical data.

Hierarchical clustering proved to be an effective method for uncovering structure within the dataset. By computing pairwise distances and applying various linkage methods, we identified Ward's method as the most interpretable due to its ability to produce compact and well-separated clusters. The analysis was further supported by the Pseudo-F index, which suggested an optimal partition with four clusters. Additionally, applying hierarchical clustering on the principal component scores via HCPC yielded consistent results while improving interpretability through dimensionality reduction. Overall, hierarchical clustering successfully revealed meaningful groupings of vehicles based on their technical characteristics.

Performing the MANOVA analysis has allowed to see some very specific relations between some categories of variables num-of-doors and bodystyles with some continuous variables. We have been able to see that our created variable SWaPC (size, weight, power and cost), along with the normalized-loss allow to differentiate cars with two or four doors and some main body styles: wagon, sedan, hatchback and sportscar (convertible and hardtop joined).

Then, the linear discriminant analysis has revealed that the SWaPC and the normalized-losses allow to predict with a 70% of correct classification rate if cars have two or four doors, based on the fact that the SWaPC mean is higher on cars with four doors, and the normalized-losses are lower.

Lastly, the quadratic discriminant analysis has shown up to be a good way to predict the body styles using the continuous variables SWaPC, normalized-losses, stroke, compression-ratio, height and peak-rpm, with a good correct classification rate, but having some overfitting of the too few observations present in our model.

This dataset could also allow deeper analyses on specific relationships, for example, predict the price using some other variables, or as suggested by the creator of the dataset, predict the symbolling. Nevertheless, this study has been focused on finding general relationships between variables, and we can conclude there are, in fact, a lot of hidden relationships.

4. Bibliography

UC Irvine Machine Learning Repository, available online at https://archive.ics.uci.edu/

Page of the *Automobile* dataset on UCI ML Repository available online at https://archive.ics.uci.edu/dataset/10/automobile

Page of the *Automobile Dataset* on Kaggle, available online at https://www.kaggle.com/datasets/toramky/automobile-dataset/data

The Automobile Data on Furman University page, Andy Liaw, available online at http://math.furman.edu/~dcs/courses/math47/R/library/randomForest/html/imports85.html

United States Environmental Protection Agency, webpage about Gasoline Label, available online at https://www.epa.gov/fueleconomy/text-version-gasoline-label

Bae Systems: What is SWaP-C?, available online at https://www.baesystems.com/en-us/definition/what-is-swap-c

1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.

Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038

Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037