Practical Work P1

Data Mining GEI, Q2 2024-2025

Car sales in US states

Alex Beauchamp

Otman Ezzayat Maid

Joan Gómez Català

Matija Jakovac

Álvaro Monclús Muñoz

Delivery date : 24th of March 2025

Index

Motivation of the work and general description of the problem 1
Data source presentation 2
Formal description of Data structure and metadata 3
Description of data matrix 3
Description of metadata 3
Metadata Table 4
Complete Data Mining process performed 5
Detailed description of Preprocessing and Data preparation 6
Basic statistical descriptive analysis 7
Univariate variables 7
General description 15
PCA analysis for numerical variables 16
Individuals projections 17
Interpretation of relationships 26
Conclusions 27
Hierarchical Clustering on original data 28
Description of the data and clustering method used 28
Analysis of final number of clusters 28
Resulting dendrogram 29
Profiling of clusters 30
Global discussion and general conclusions 38
Working plan 40
Initial Gantt Diagram 40
Final Gantt Diagram 41
Tasks assignment grid 42
Deviations from the Original Schedule 43
Risk Management: Avoided and Unexpected Risks 44
R Scripts 45

Motivation of the work and general description of the problem

The second-hand market is a vast ecosystem all around the globe where consumers and retailers aim to make the correct decision based on several factors and vehicle characteristics. With the increasing focus on data-driven approaches to solve real world problems, analyzing consumer patterns will ultimately help us understand the market and provide valuable insight into trends, pricing, demand and general market dynamics. This study aims to explore trends and patterns within a dataset of used car sales in the US market to understand which factors influence vehicle prices and sales likelihood.

The dataset consists of entries that describe a listing of a car sale, with the technical characteristics of each car. It also includes ownership history, a greatly important factor to consider when studying consumer behaviour. A very important piece of information in the dataset is whether a car was ultimately sold or not. By analyzing these features we can understand the patterns we have discussed, helping us identify the key determinants of a successful sale, and highlight regional differences in pricing and demand.

The primary objectives of this study are the following:

1. Understanding pricing factors: We aim to examine how attributes such as mileage, year of manufacture, fuel type, among other factors influence the price retailers set when listing a car.

2. Predicting sales likelihood: We aim to identify which characteristics increase the chances of a car being sold.

3. Consumer and seller behaviour: We aim to analyze the trends that consumers and different seller types (dealers vs individual sellers) follow.

4. Regional market variations: We aim to investigate how location affects car prices and sales trends.

Data source presentation

The data was sourced from Kaggle uploaded by the user “Shubam Kumar” (https://www.kaggle.com/datasets/shubham1kumar/usedcar-data), an online platform commonly used for obtaining datasets. Since the dataset contained over 5,000 records, we used R, programming language for statistical computing and data visualization, to randomly select 5,000 records for our analysis. The data is stored in CSV format, with a semicolon (;) as the delimiter when loading it. This dataset represents used car sales by an Indian automobile company across various states in the United States from 1994 to 2020. Each transaction includes details about the distributors, car specifications, and location.

Formal description of Data structure and metadata

Description of data matrix

This dataset consists of 5,000 randomly selected records from a total of 8,128 entries, representing used car sales by an Indian automobile company across various states in the United States from 1994 to 2020. Each transaction includes details about the distributors, car specifications, and location.

Description of metadata

Each row in the dataset represents a unique car listing, capturing several characteristics of the vehicle, the seller and the transaction details. These records contain a combination of numerical and qualitative variables, including the physical properties of the car (e.g. mileage, engine capacity, max power), sales information (e.g. price, seller type, owner) and contextual information (e.g. region, state, city) Additionally, it contains a binary variable that shows whether the car was sold or not.

The scope of the study focuses on identifying patterns in consumer behaviour and factors that influence the success rate of a used car sale. This dataset will be analyzed in great detail using several data mining techniques to further understand how the variables in the dataset influence the factors we have named previously.

Inclusion criteria:

- All vehicle listings with complete data for the key attributes needed for the study.
- Listings that include a valid sale outcome.
- Listings that contain information that seems plausible and correct.
- Vehicles manufactured between 1994 and 2020.

Exclusion criteria:

- Listings with missing or inconsistent values in essential columns.
- Listings with data that seems incompatible with real data, including outliers.
- Duplicate entries, if any are found, to ensure the integrity of the data.

Metadata Table

Variable	Short Name	Modalities	Short Mod Name	Meaning	Type	Units	Missing Code	Measuring Procedure	Range	Role
Sales_ID	Id			ID of the sale	Quali					Explanatory
name	Name			Name of the car	Quali					Explanatory
year	Year			Year of fabrication	Num	years			[1994, 2020]	Explanatory
selling_price	Price			Price it was sold at	Num	euros			[330, 71753]	Explanatory
km_driven	Km			Km driven	Num	km			[1000, 577414]	Explanatory
region	Region			Region where it was sold	Quali					Explanatory
state	State			State or province where it was sold	Quali					Explanatory
city	City			City where it was sold	Quali					Explanatory
fuel	Fuel			Type of fuel it uses	Quali					Explanatory
seller_type	Seller			Type of seller	Quali					Explanatory
transmission	Trans			Type of transmission	Boolean					Explanatory
		Manual	M							
		Automatic	A							
owner	Owner			Number of owner	Quali					Explanatory
mileage	Mileage			Mileage	Num	mpg			[9, 33.44]	Explanatory
engine_size	Engine			Engine displacement	Num	cm3			[624, 3604]	Explanatory
max_power	MaxPower			Maximum power of the engine	Num	hp			[32.80, 282.00]	Explanatory
seats	Seats			Number of seats	Num	seats			[4, 10]	Explanatory
sold	Sold			Sold status	Boolean					Explanatory
		Yes	Y							
		No	N							

Complete Data Mining process performed

The full data mining process consisted of 6 primary steps:

- Data acquisition
- Data selection
- Preprocessing
- PCA
- Clustering
- Profiling and interpretation

The process began with obtaining the data from the popular Kaggle.com website. From here, the data pertaining to car sales in the US was extracted and set into our RStudio working directory. The next step was to select an appropriate amount of data, as the original dataset contained over 8.000 entries, far beyond the scope of this study. As there seemed to be no missing data or strange values, 5.000 rows of data were randomly selected as per the specification of the project guidelines. The next step was preprocessing the data. At first we generated an initial descriptive to identify the variables that were in need of being preprocessed and how they could be modified. Here some variables were selected, such as city, state, price and power and some aspects such as the number of modalities or the units of the variable were modified. At this point the outliers and wrong data that were found were dealt with.

Once the data was properly preprocessed the next step was applying PCA to understand the relationships between our variables, reduce dimensionality and identify the most significant features. This was an essential process to truly understand the data and to be able to extrapolate conclusions on the used car market and its trends. Following the PCA, the clustering phase was conducted to segment the data into meaningful groups. We opted for the K-Means algorithm with K=5. Finally, the profiling and interpretation stage involved examining the PCA and characteristics of each cluster to obtain insight on the data. Key factors were analyzed to understand their relationships and extrapolate trends in the market.

Detailed description of Preprocessing and Data preparation

Firstly, no data imputing algorithms were used in preprocessing as there was no missing data. However, we did find some columns that were either of no use (X, Sales_ID) or had data in an unusable state (Torque). The decision was made to remove these columns to maximize the quality of the data. Three columns were also identified as having too many modalities to be interesting (names, provinces, cities). These columns were kept, but their visualisations were either reduced or deleted in the descriptive as they may serve for statistical purposes.

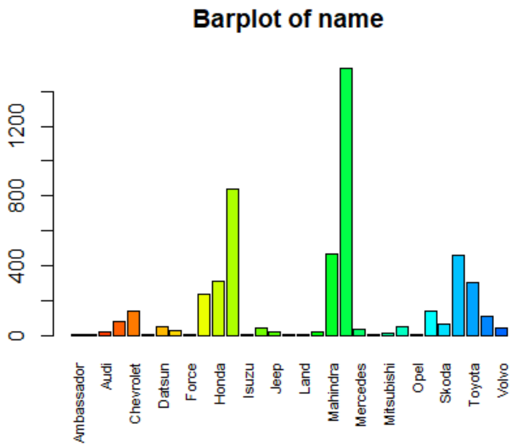
Some decisions were also made for specific columns. First, to simplify presentation of results and adapt them to our environment, we converted the values in the “selling_price” column from rupees to euros. For the “state” column, we have replaced state names with their state codes. In the same interest of shortening qualitative modalities, we have changed “name” so that all names have been cut to a max of 5 letters with a “.” at the end when cut. In the “region” column, "Central" has been shortened to "Centr." For “seller_type”, "Individual" is shortened to "Indiv" and "Trustmark_Dealer" to "Trust." In the “transmission” column, "Manual" is shortened to "Man" and "Automatic" to "Auto." For the “owner” column, "First_Owner" is shortened to "First", "Second_Owner" to "Second", "Third_Owner" to "Third", "Fourth_Above_Owner" to "FourthM" and "Test_Drive_Car" becomes "Test".

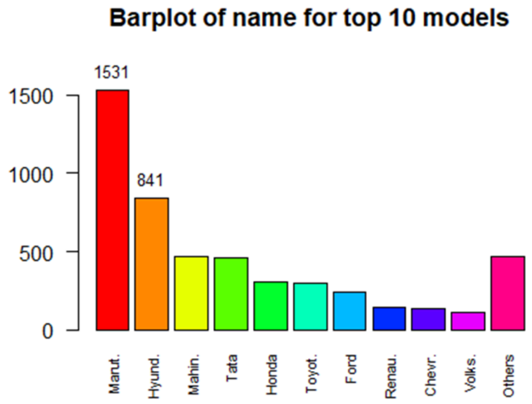
The rest of the columns had numerical values and as such did not need any apparent preprocessing. The only interesting aspect we noticed was some slight outliers in selling_price. A better visualization may necessitate an exponential scale, as while most of them were fairly low value (< 1e+6 rupees), quite a few would reach up to 6e+6 rupees, which slightly break the visualisation. There are a few outliers in the “mileage” column with values of 0, which we replaced with the average of cars of the same brand in the dataset. There are also two outliers in the “km_driven” column. Whereas every other value is around or below 500k (which is high but possible), the top two are 2.3 and 1.5 millions respectively, which is incredibly high but still below the 4.8 million km world record so we choose to keep them as-is.

Basic statistical descriptive analysis

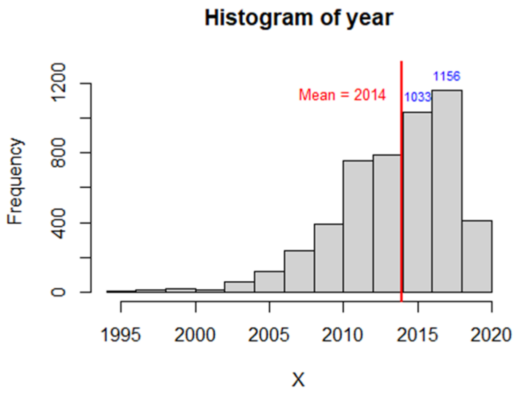
Univariate variables

For the basic statistical analysis, we have mainly only kept the plots after preprocessing. For the plots that underwent interesting changes, the before and after can be found on the left and right, respectively.

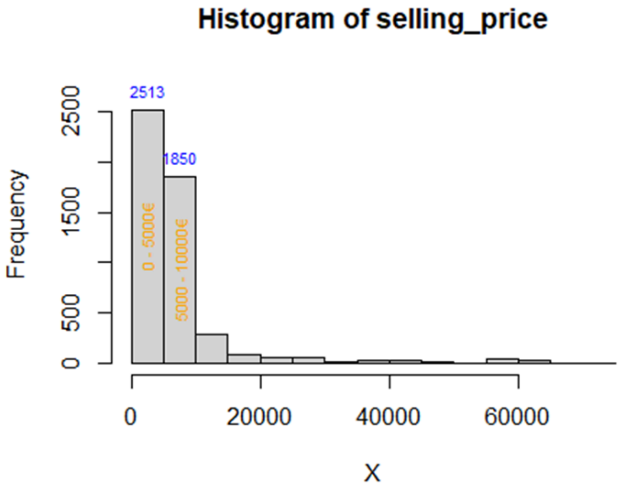

Barplot of name


Barplot of name for top 10 models

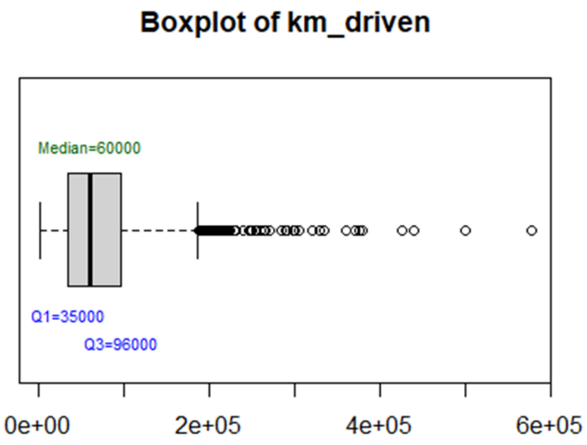
These are barplots of a qualitative variable describing the model name of the car. There are 31 different models. As we can see on the left (before preprocessing), the number of models was too high for the graph to have any meaning, so, as a high number of models had a small number of appearances, we included only the top 10 models and put the rest in the “Others” category. It is interesting that the most common models of cars are sold multiple times more than the lesser sold cars.



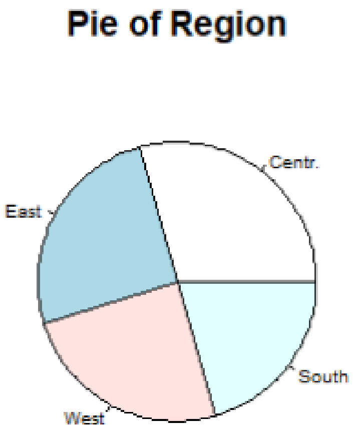
This is an histogram of the quantitative variable indicating the year the car was sold. It seems to follow close to a normal distribution, though with a slight right deviation, indicating that newer cars tend to be sold more.



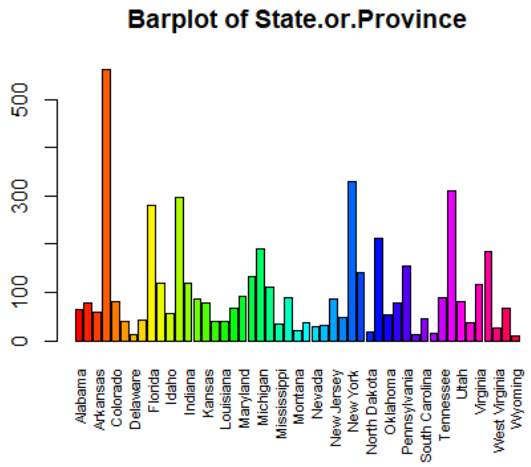
This is a histogram of the quantitative variable describing the price paid for the car in euros. As we can see, the vast majority of values are actually quite low, as to be expected with the second hand market, with some specific listings having a higher price.

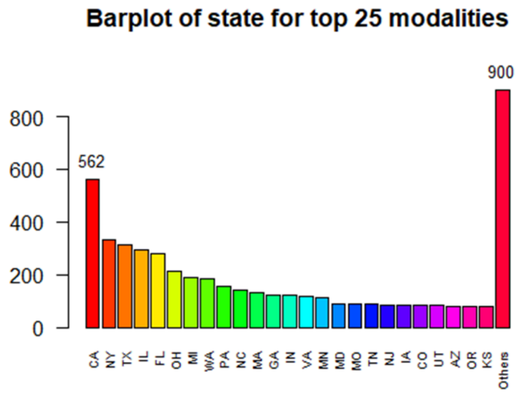


This is a boxplot of the quantitative variable describing the number of kilometers driven at the moment of sale. As we can see, most values are contained beneath 500 000 kms, but there are a few outliers, including some vastly superior to that which is average. It should be noted that the outliers are still plausible in terms of how many km a car can drive, so we did not count it as an error or missing data, simply as an extreme value.

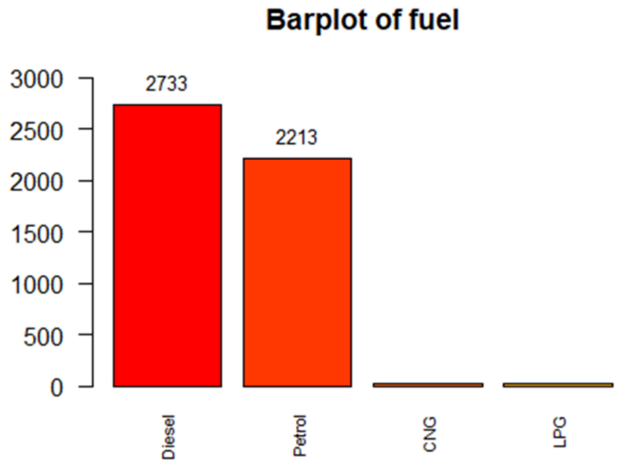


Here we have a qualitative variable in a box plot describing the region of the US in which the sale was conducted. It is actually quite balanced, with each section representing about a quarter of sales.

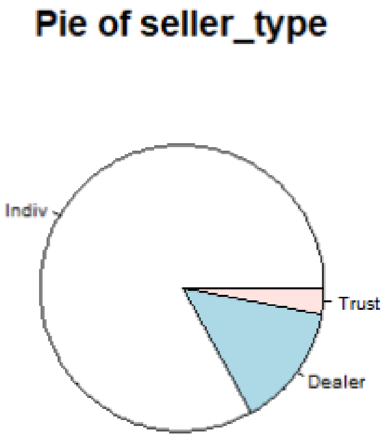




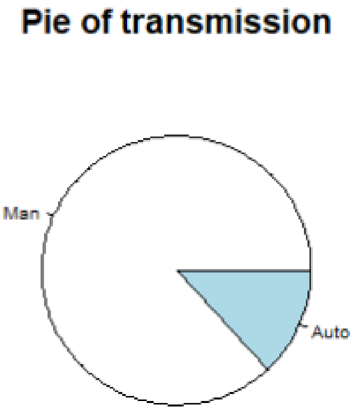
The title of the barplot is StateOrProvince but represented here are only states. Due to the high number of modalities (as you can see on the left), here are the top 25 most represented states. There are still states that are overly represented compared to others. You can also view the impact of preprocessing on the state names, which have been replaced by their state codes.



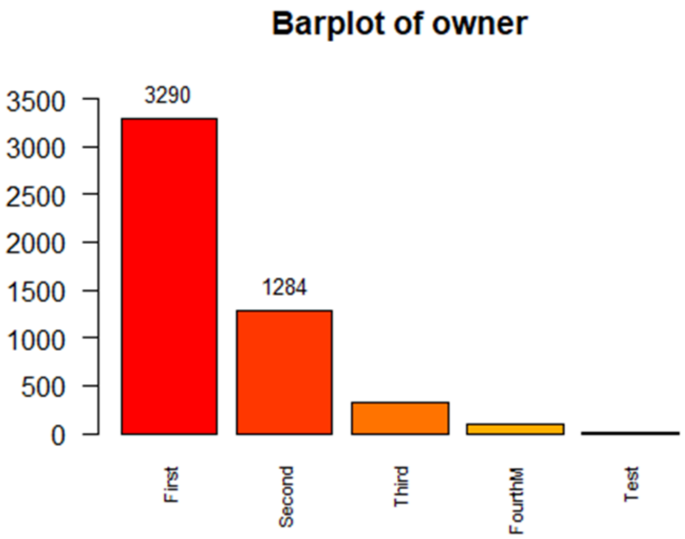
Here is a variable representing the type of fuel in each car. We can notice that Diesel and Petrol are much more common than the other two, with a slight preference for diesel.



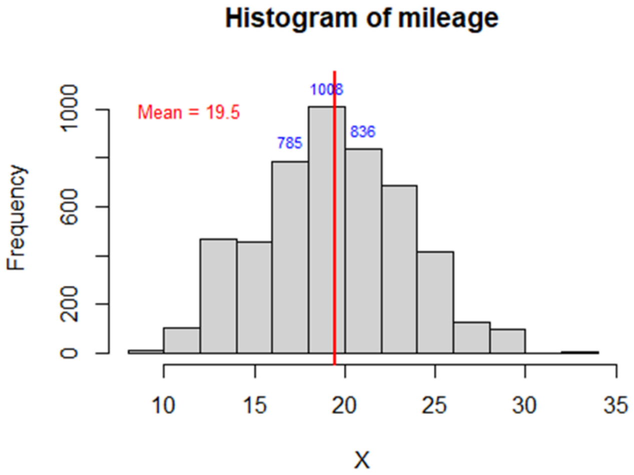
This variable represents the type of seller of the car in a pie chart, with only three modalities: Individual, dealer and Trustmark dealer. The most common, individual, represents more than 80% of sales. On the other hand, Trustmark dealer seems to be the least favourite option by a significant margin.



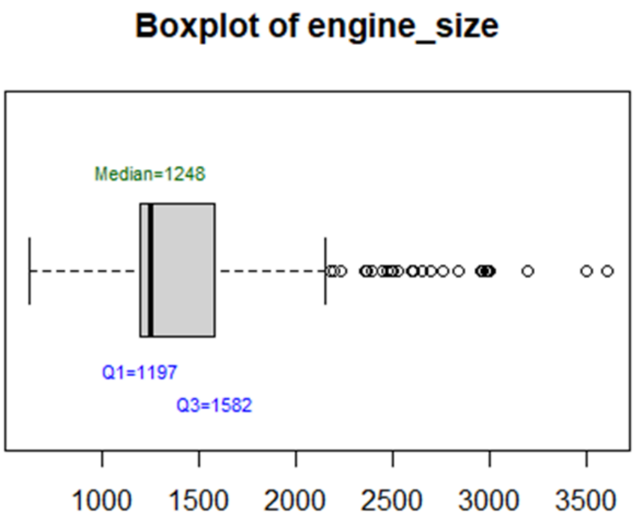
Here is a pie chart describing the type of transmission variable which could be either manual or automatic. Manual is the most sold type by a very high margin (>80%), making Automatic the less preferred option on sales.



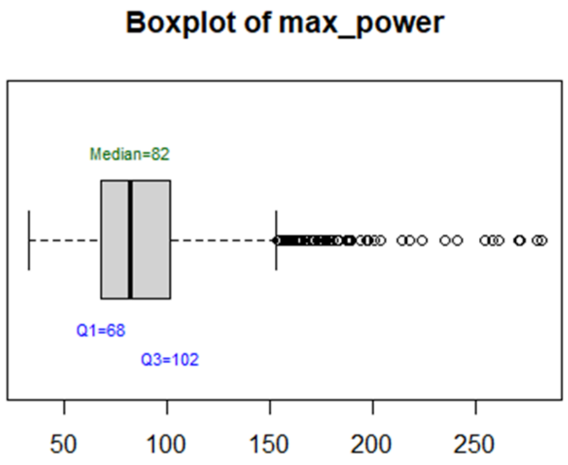
This is a histogram of the qualitative variable describing how many cars the buyer has had beforehand (First means they have never had a car before and FourthM means that they had four or more cars before). Test simply describes a test driver. We can observe that most sales were done to buyers who had never had a car before.



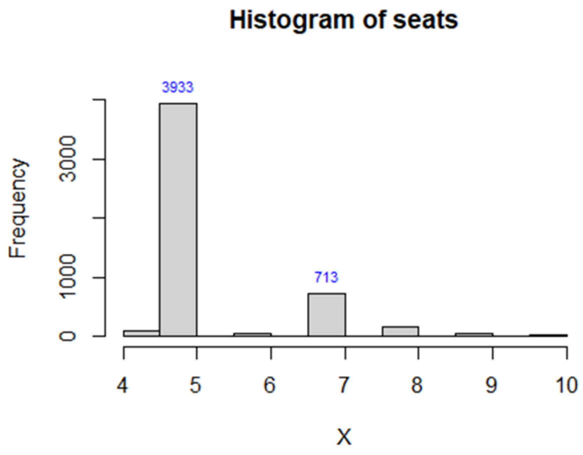
This is an histogram describing the quantitative variable of the number of kilometers a car can travel using one liter of gas. It is relatively close to following a normal distribution.



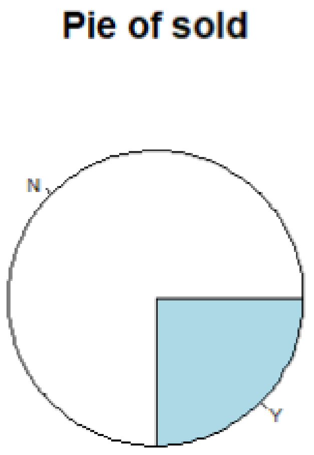
This is a boxplot describing the variable of the size of the engine in the car. We can notice quite a few outliers, as well as a very slim second quarter, indicating a very high concentration of values within that smaller range.



This is a boxplot of the quantitative variable describing the max power the engine can output. We can see that the values seem to follow a relatively normal distribution, as the main quarters are all quite similar in size. However, we can notice quite a great number of outliers.



This is an histogram of the quantitative variable describing the number of seats in the car. We can notice that the standard seems to be 5, but that 7 seems to be the second most popular, whereas there have not been that many sales regarding cars of other numbers of seats.



This is a pie chart of the qualitative variable describing whether or not the car ended up being sold. “No” represents about 75% of the values, meaning that only a quarter of the cars were actually sold.

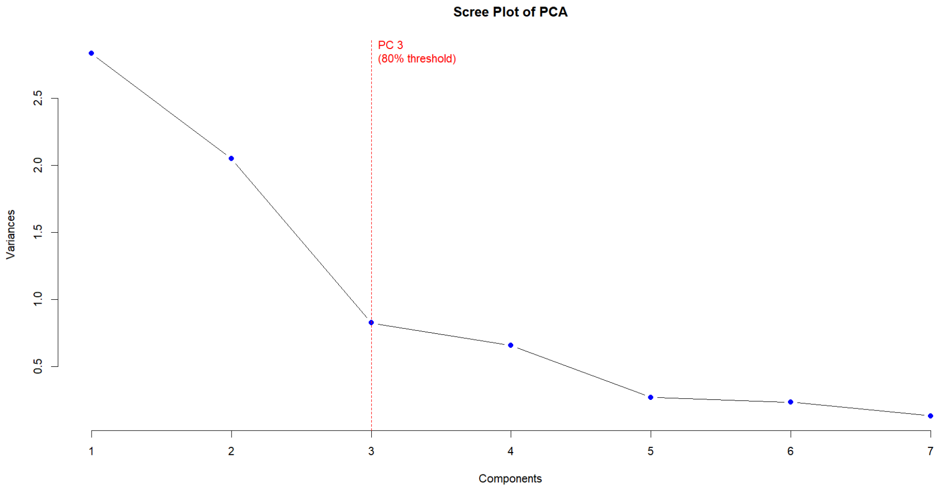
General description

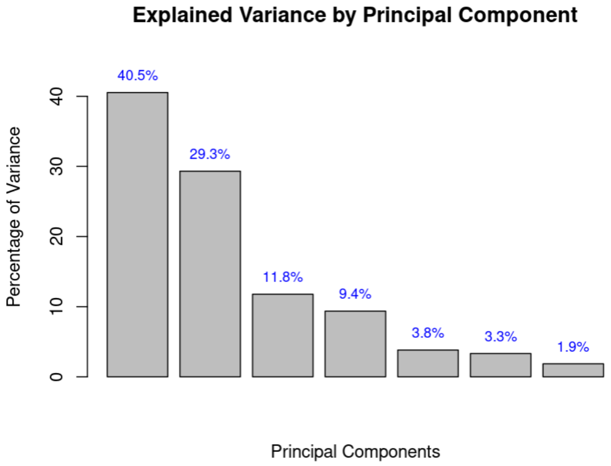
The dataset offers a broad understanding of car sales since it contains most relevant information about the qualities of a car using both qualitative and quantitative variables.

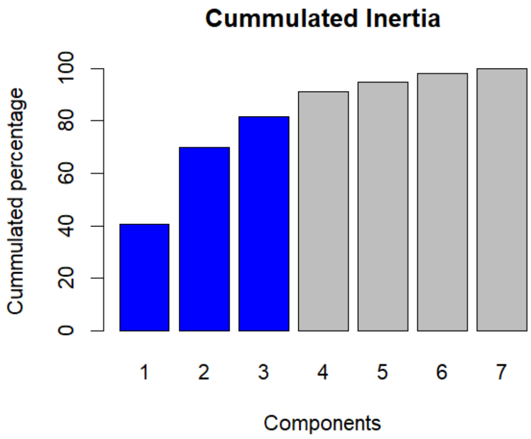
In the case of numerical variables, we used either histograms for variables that had a low range of values, such as “Seats” and “Years”, and boxplots for variables that could acquire a lot of values, like the variables “engine_size” and “max_power”. On the other hand, for categorical variables we used histograms and even some pie plots, since they generally had fewer unique values except in the case of the variables “Models” and “States/provinces” which had a much larger variety.

Most of the data is meaningful and well-structured showing distributions that we had expected. However, several variables have extreme values which are still reasonable. These features guarantee that the dataset will always be helpful in analysis.

PCA analysis for numerical variables







Observing the plots we have selected **3 principal components** as they capture around 80% of the total variance which is a good balance between dimensionality reduction and retaining meaningful information.

The scree plot shows how much variance each component explains, guiding us to choose a cutoff where the curve usually begins to flatten. The explained variance bar chart indicates proportion of variance each individual principal component contributes having PC1 contributing the most with 40.5%, PC2 with 29.5% and PC3 with 11.8% of total variance. The cumulative inertia bar chart shows how the explained variance adds up across all 7 components. Together, these visuals confirm that the first three PCs provide enough information on the data to continue our analysis.

Individuals projections

Across these individual projections on the principal components, most data points form a relatively continuous cloud with a few denser regions and some outliers. Since the first three principal components capture a large portion of the total variance, visualizing the data in these reduced dimensions can help uncover patterns or groupings.

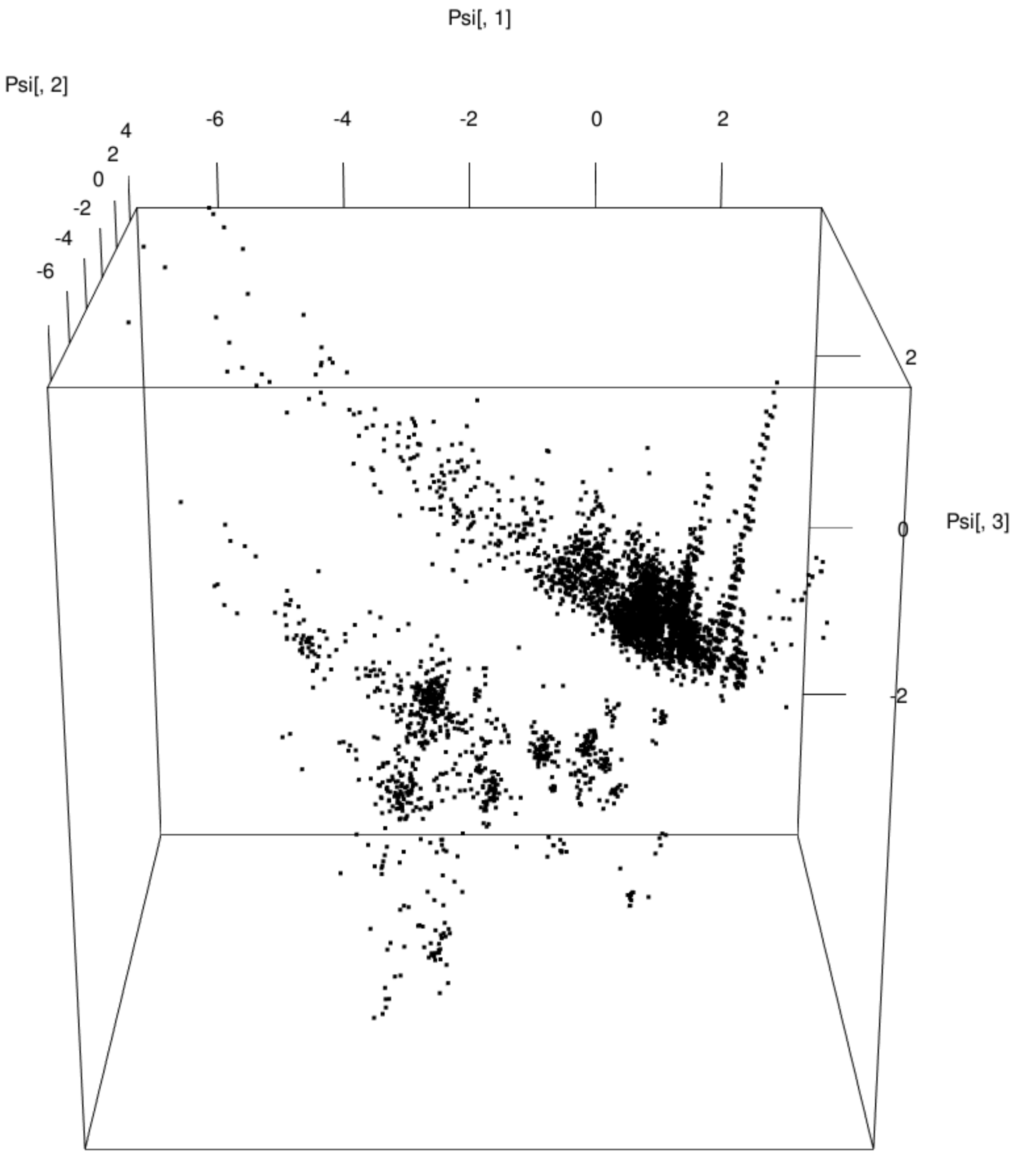
PC1 vs. PC2: Most individuals appear aligned more strongly along the PC1 axis (the x-axis), suggesting a higher correlation with PC1 than with PC2. The majority of points lie on the positive side of PC1.

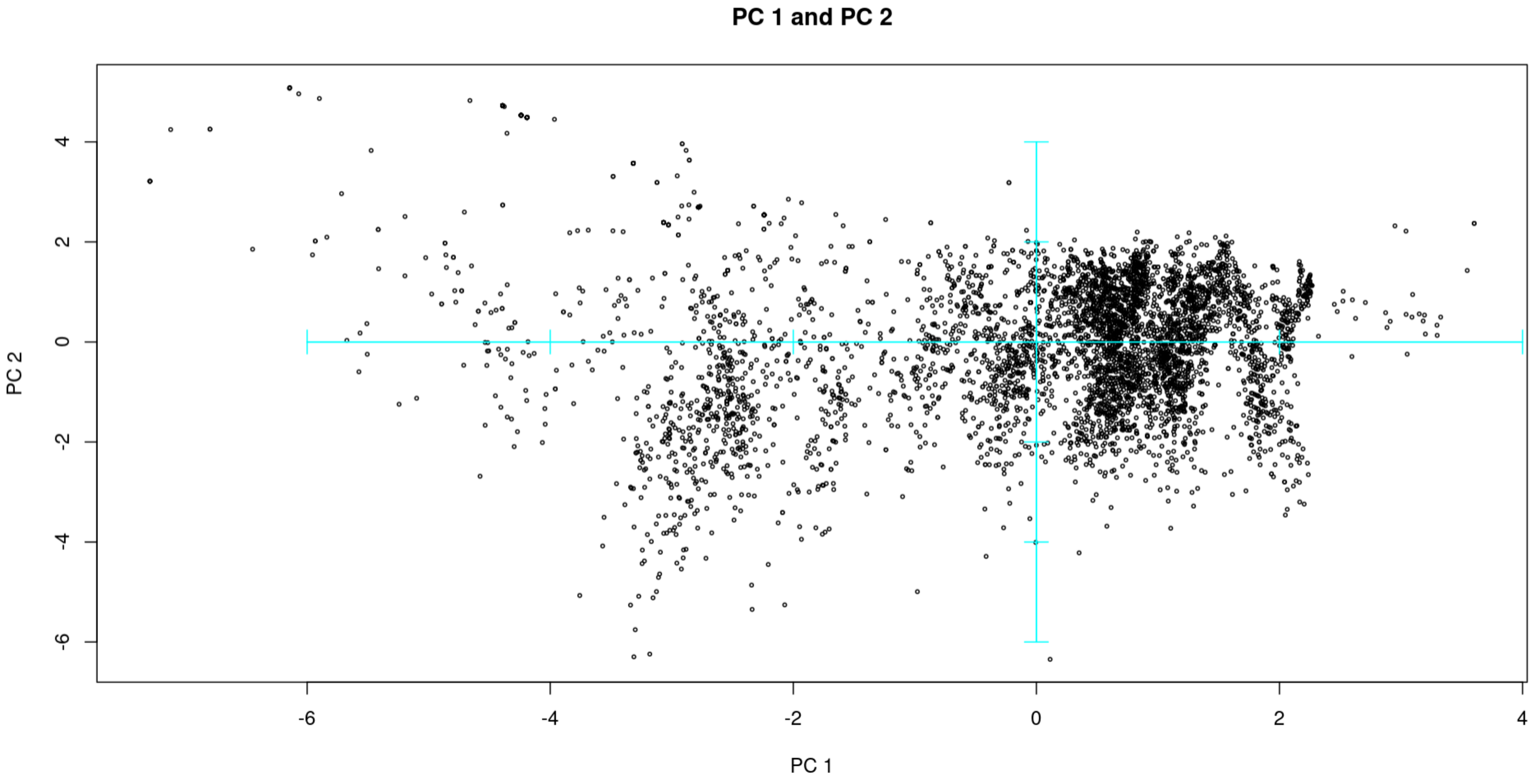
PC1 vs. PC3: Points still cluster toward the positive side of PC1, but they are not as strictly aligned along the x-axis. A rough trend ($y = -0.5x$) could separate the cloud into two apparent subgroups.

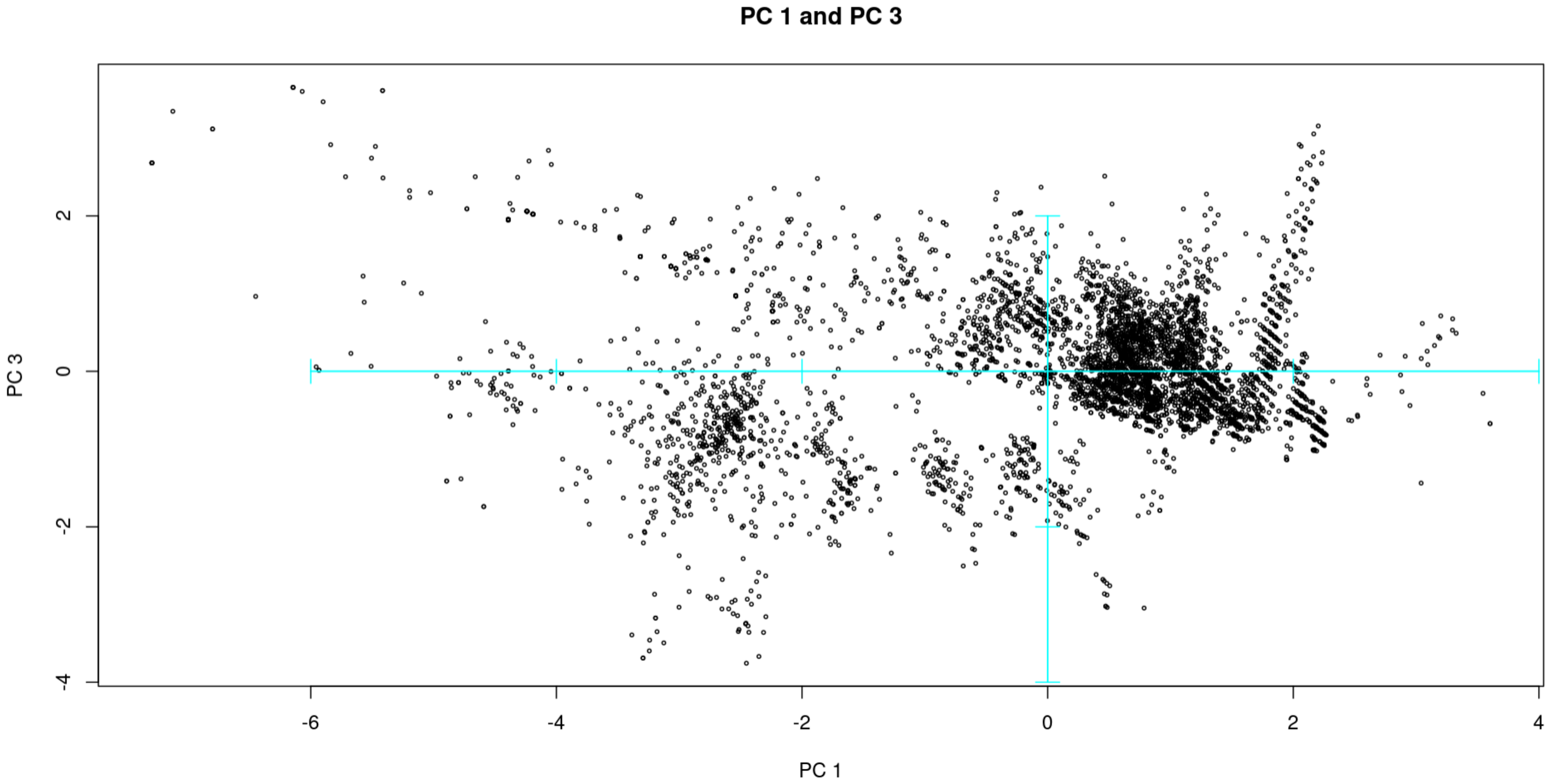
PC2 vs. PC3: The distribution is somewhat similar, with points clustering around a general trend line. However, there are no clear, well-separated clusters; the data remains fairly continuous.

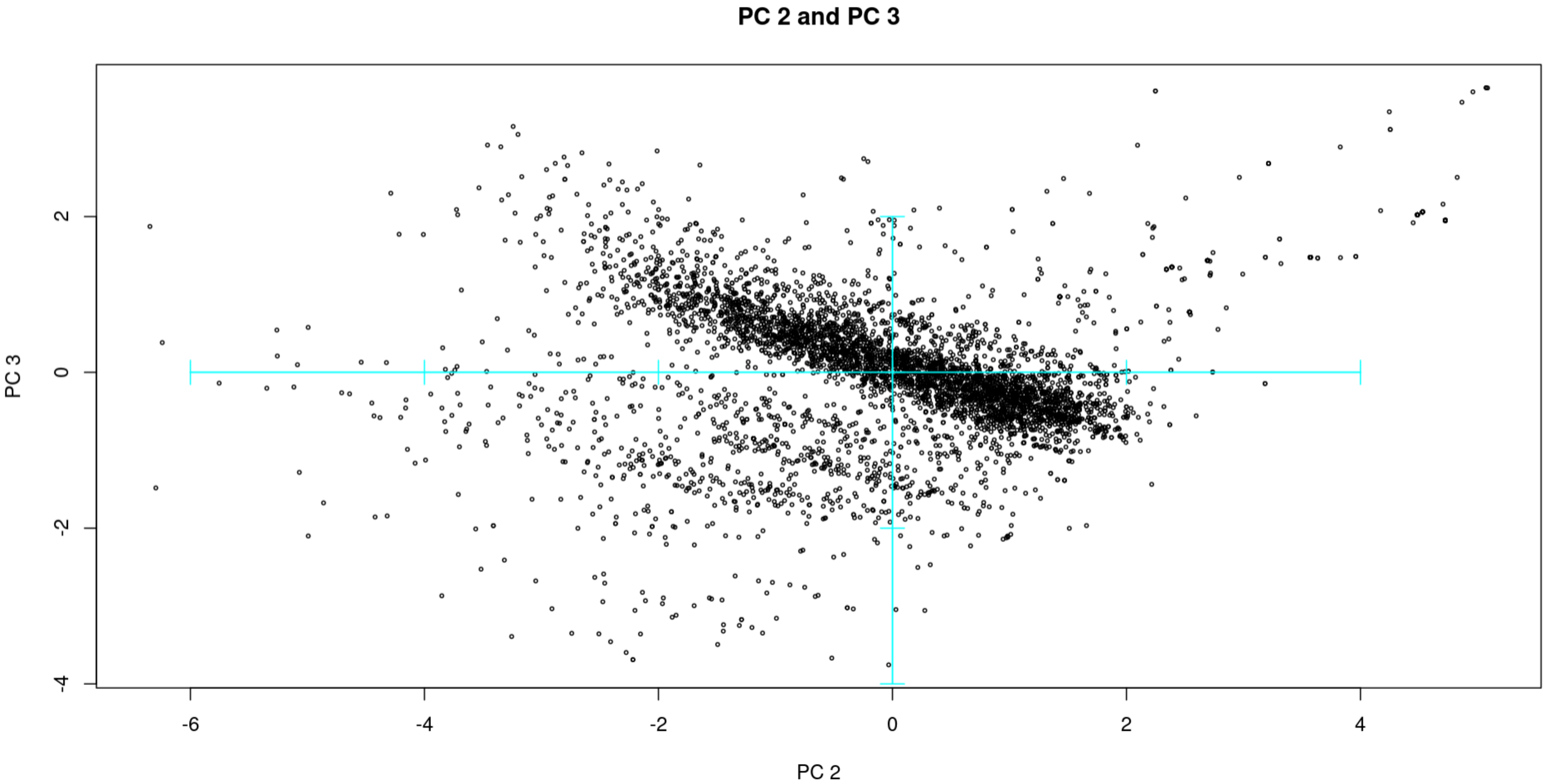
3D Plot (PC1, PC2, PC3): This 3D visualization reinforces the relationships seen in the 2D plots, showing how individuals spread out across all three components simultaneously.

Overall, these factorial maps indicate that while the data varies meaningfully along the principal components—especially PC1—there are no defined clusters. Instead, the individuals occupy a continuous space, reflecting different degrees of correlation across the first three principal components.

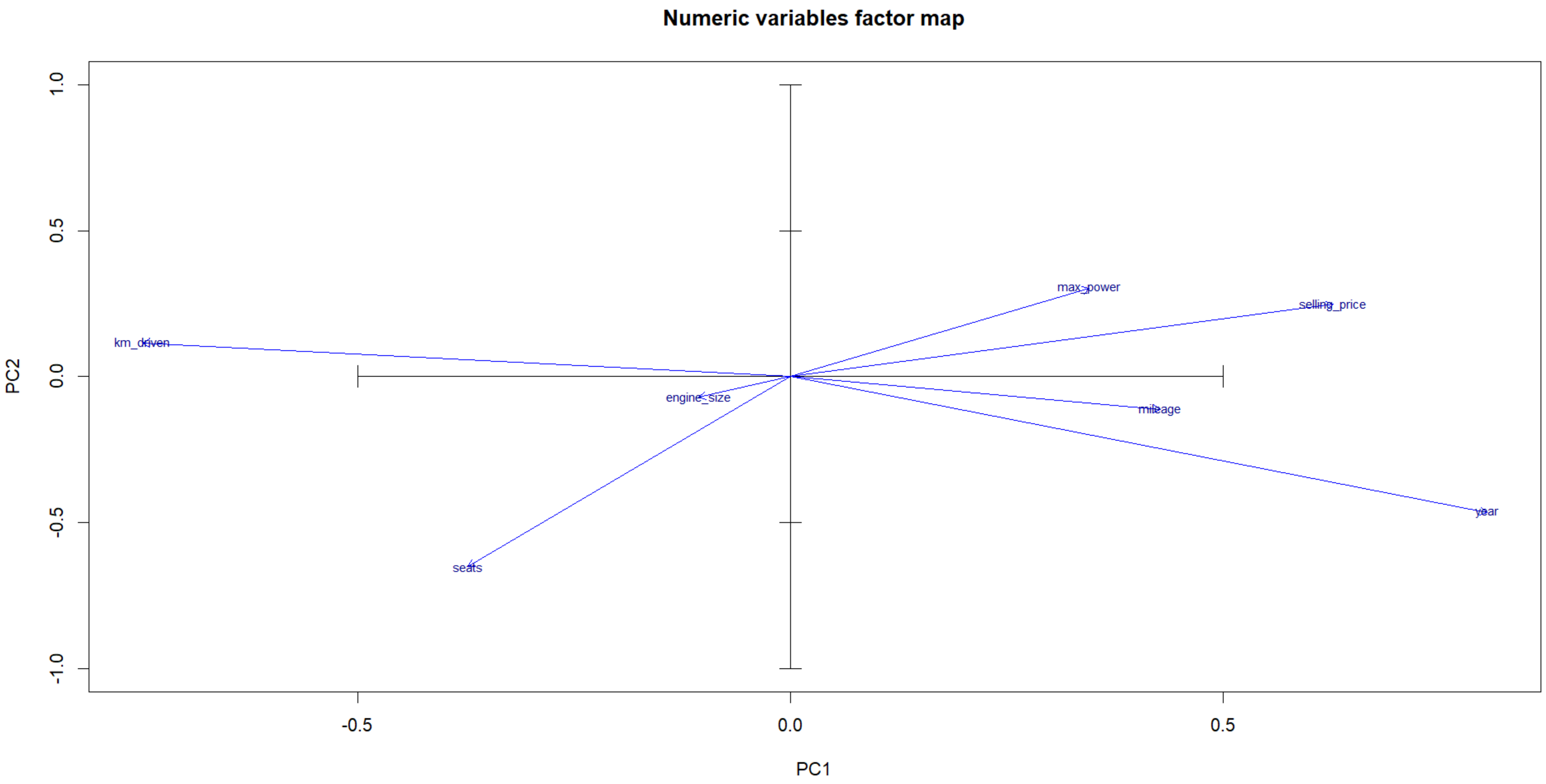

Psi[, 1]
-6
-4
-2
0
2
Psi[, 2]
4
2
0
-2
-4
-6
Psi[, 3]
2
0
-2


PC 1 and PC 2
PC 2
4
2
0
-2
-4
-6
PC 1
-6
-4
-2
0
2
4


PC 1 and PC 3
PC 3
2
0
-2
-4
-6
-4
-2
0
2
4
PC 1


PC 2 and PC 3
PC 3
2
0
-2
-4
-6
-4
-2
0
2
4
PC 2

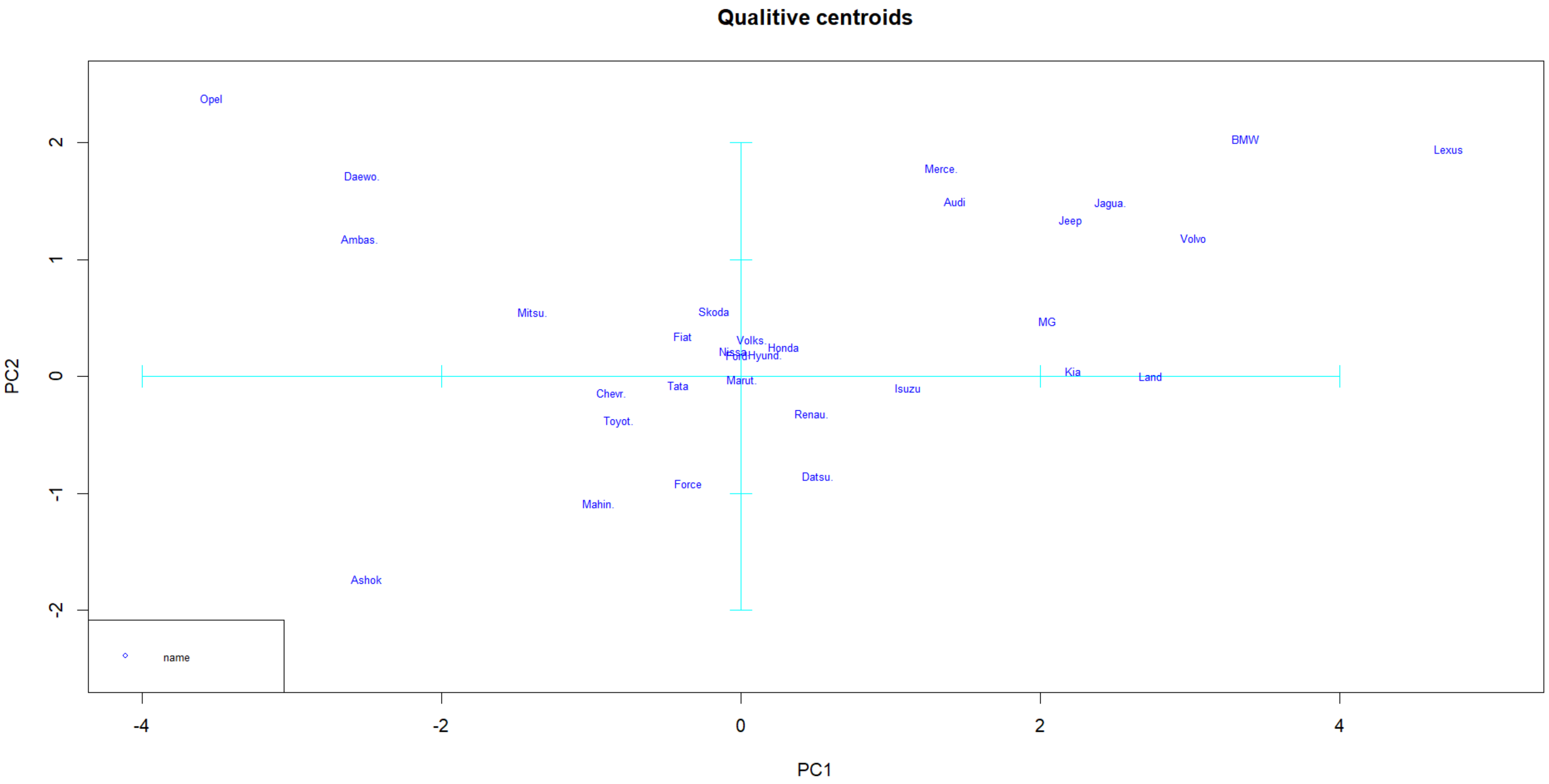
i. Common projection of numerical variables and modalities of qualitative variables



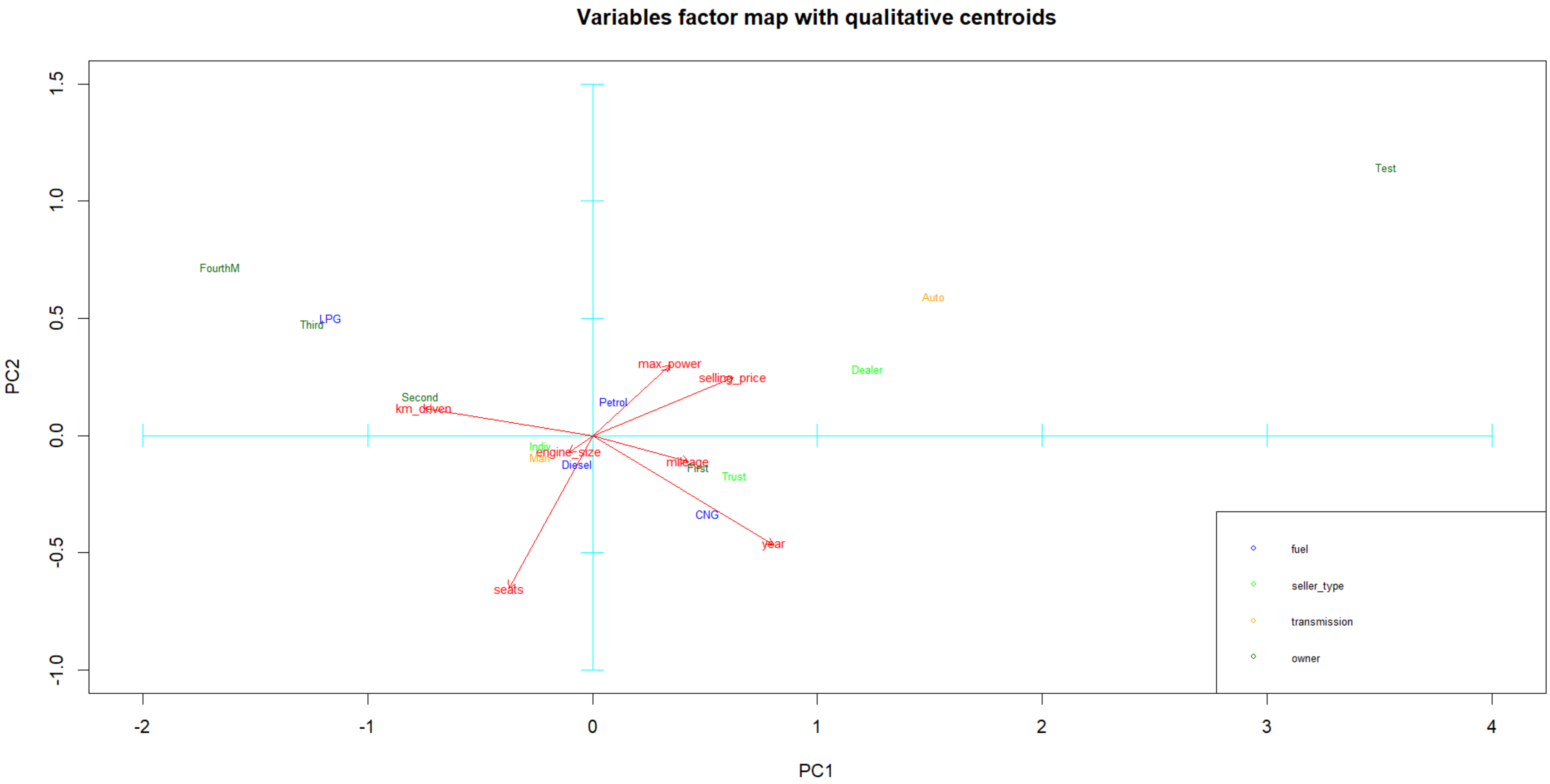
There are quite a few notable correlations between the different numerical variables. First of all, the strongest inverse correlation is between the year and selling price, and the amount of kilometers traveled. This makes complete sense since an older car will have had more years to travel, as well as sell for a generally lower price. Second, the strongest direct correlation is between the selling price and both mileage and max power. Generally speaking, a car with a higher performance will sell for a higher price. This trend is also true for newer cars, especially so for mileage, which may be explained with the release of new technologies.

There is also a fairly strong inverse correlation between the number of seats and the max power. This could indicate a trend where more utility focused vehicles that have lots of seats (SUVs, vans) tend to have less power than more sports-oriented cars. The number of seats also seems to be inversely proportional to the selling price, indicating once again that the upper-right seems to be reserved for less utility-oriented, more luxurious sports cars. Finally, engine size doesn’t have very significant correlations, but it is relatively aligned with kilometers driven and directly opposite to max power.

Generally speaking, PC1 seems to separate higher end, more expensive cars on the right, cheaper, older cars on the left. PC2 also seems to separate cars with more seats on the bottom and more power on top, though this distinction is notably weaker than with PC1.



Here the PC2 separation becomes more apparent. On the bottom end, we can see more utility oriented brands (Ashok vans, Mahindra SUVs), whereas the top of the graph is filled with more luxurious brands (Mercedes, BMW). The trend of cheaper, older cars on the left and newer, more expensive cars on the right also seems to be maintained.


Variables factor map with qualitative centroids
PC1
PC2
-2
-1
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
1.5
Test
FourthM
Third
LPG
Auto
Dealer
max_power
selling_price
Second
km_driven
Petrol
Indiv
engine_size
Man
Diesel
mileage
First
Trust
CNG
year
seats
fuel
seller_type
transmission
owner

From the previous graph we can extract several takeaways. Firstly, there is a clear positive correlation between the number of owners of a vehicle and the kilometers driven, which is aligned with the expectations. Additionally, we can observe that cars with fewer owners tend to have a higher mileage. Moreover, CNG vehicles, being a relatively newer technology compared to diesel, petrol and LPG, are generally more recent in manufacturing.

The PCA graph also reveals that diesel cars typically have larger engines than petrol vehicles, something that aligns with the reality of these types of cars. Furthermore, cars sold by dealers tend to have higher selling prices than those sold by individual owners, something that can be due to higher fees and quality provided by dealers compared to individuals. Lastly, automatic vehicles are generally priced higher than their manual counterparts, something common in the market.

Interpretation of relationships

As mentioned a few times in the above discussions, PC1 seems to separate older, cheaper cars on the left, and newer, more expensive cars on the right. This is especially apparent from the inverse correlation between selling price and kilometers driven. The owner variable also contributes to this observation, as dealer cars (on the right) tend to be newer and more expensive, whereas individual sellers tend to sell older, cheaper cars (on the left).

As for PC2, it seems to separate utility focused cars on the bottom and more sporty cars on the top. The main clue for this observation is the names that can be found, with luxury brands such as Mercedes and BMW in the upper bounds, and vans or SUVs like Ashok and Mahindra on the lower ends of the graph. There are also two numerical variables that contribute to this observation : number of seats on the bottom, and max power towards the top.

Conclusions

The key takeaways of this analysis are as follow :

- The highest inverse correlation is between year/selling price and kilometers driven; Newer cars have less kilometers driven and sell for more.
- The strongest direct correlation is between selling price, max power and mileage; More expensive cars have generally higher performance.
- There is a clear separation between the left and right side, with newer, more expensive cars on the right, and older, cheaper cars on the left.
- There is also a separation between the lower and upper side, with more performance-oriented cars on the top and more utility oriented cars on the bottom.

In general, this already gives us a pretty good idea on the distribution of the data. There definitely seems to be a trend separating different types of cars, but clustering will be necessary before true conclusions can be made.

Hierarchical Clustering on original data

Description of the data and clustering method used

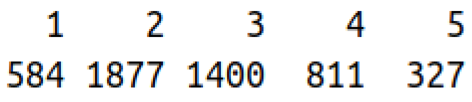
We decided to include every variable in our dataset because we believe they are all relevant. To ensure that each variable contributes equally to the distance calculation, we standardized the data. Since our dataset contains both categorical and numerical variables, we computed the dissimilarity matrix using the Gower metric, which performs well with mixed data.

After obtaining the dissimilarity values, we squared them to generate the distance matrix, as Ward's method operates on squared distances. Finally, we performed hierarchical clustering using Ward's method, which minimizes the total within-cluster variance at each iteration of the agglomerative process. This method is effective for producing well-defined and separated clusters.

Analysis of final number of clusters

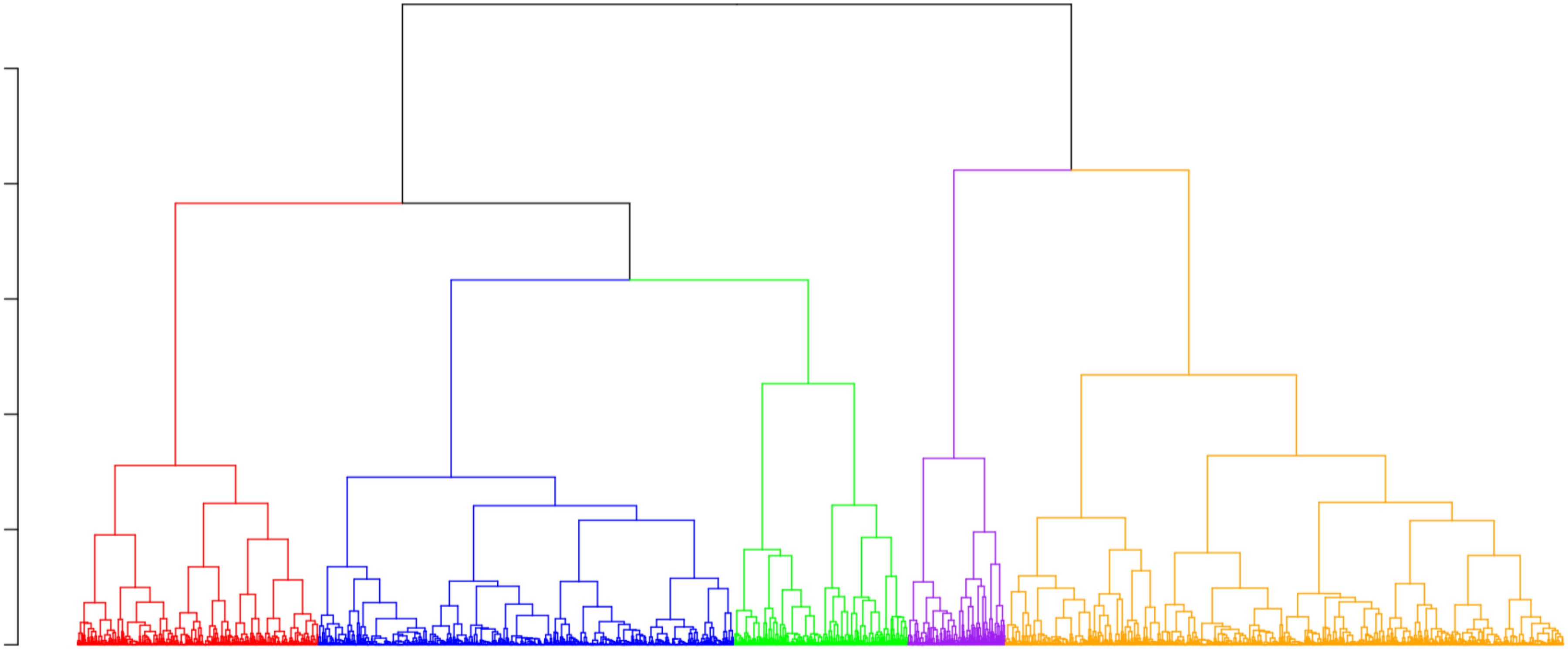
By carefully examining the dendrogram, there appears to be a distinct increase in the vertical distance between the merges at heights around 3 to 4. For this reason, we decided to cut the dendrogram at height 3, resulting in five clusters.

We did question whether this was the best cut point, because the final merge in cluster 3 does not occur much higher than when cluster 2 merges with cluster 3. However, after comparing the alternative of defining only two clusters, we opted for five clusters as a more balanced solution.

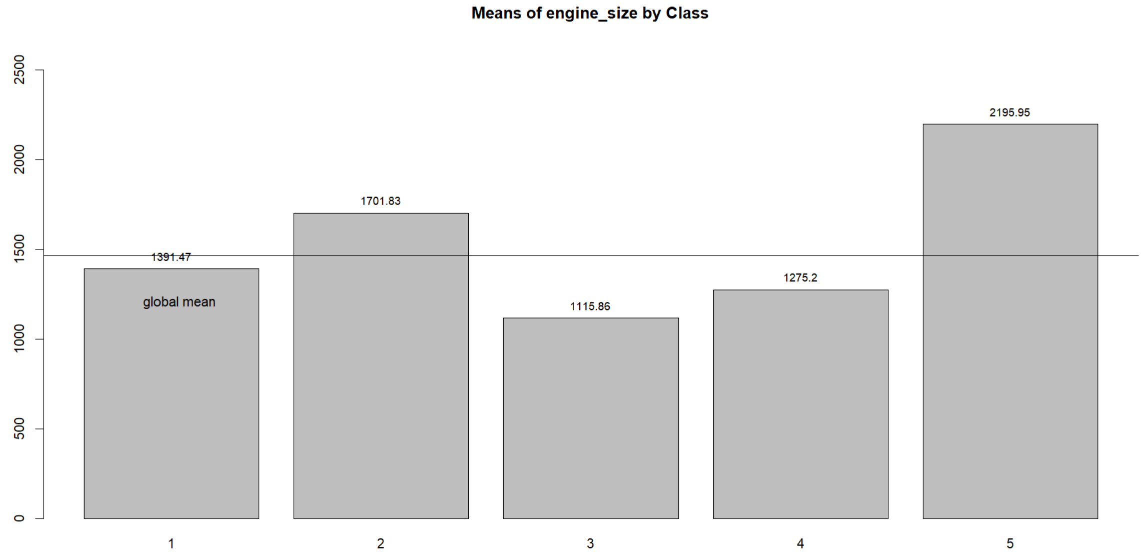


```
   1    2    3    4    5 
 584 1877 1400  811  327 
```

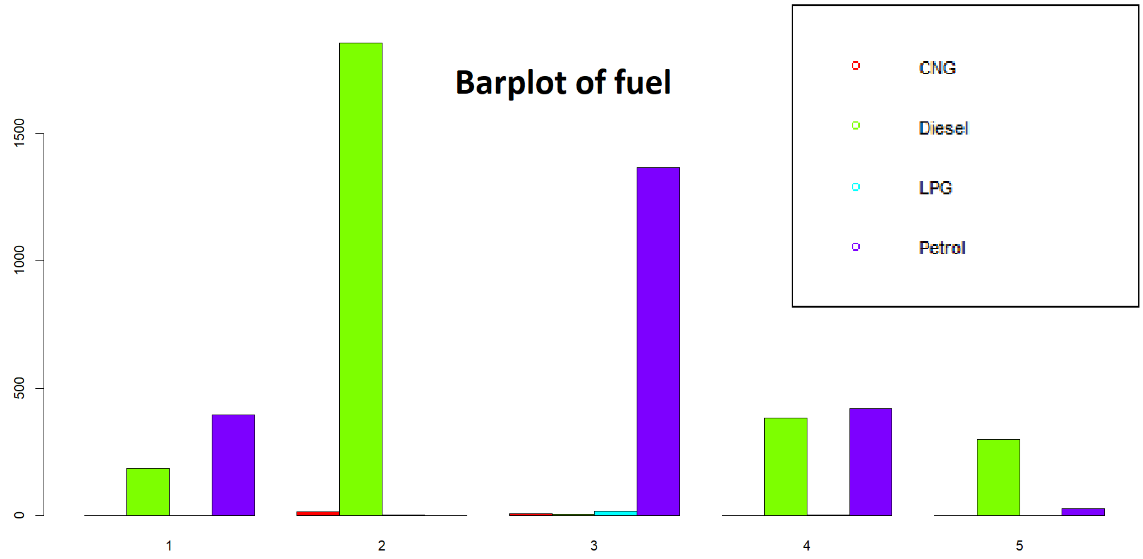
Resulting dendrogram



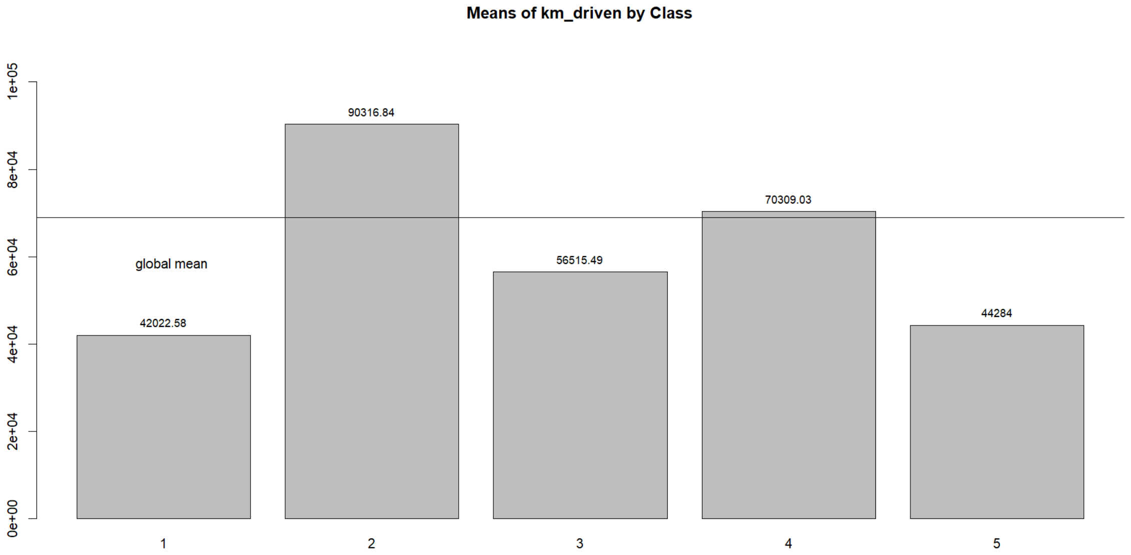
Profiling of clusters



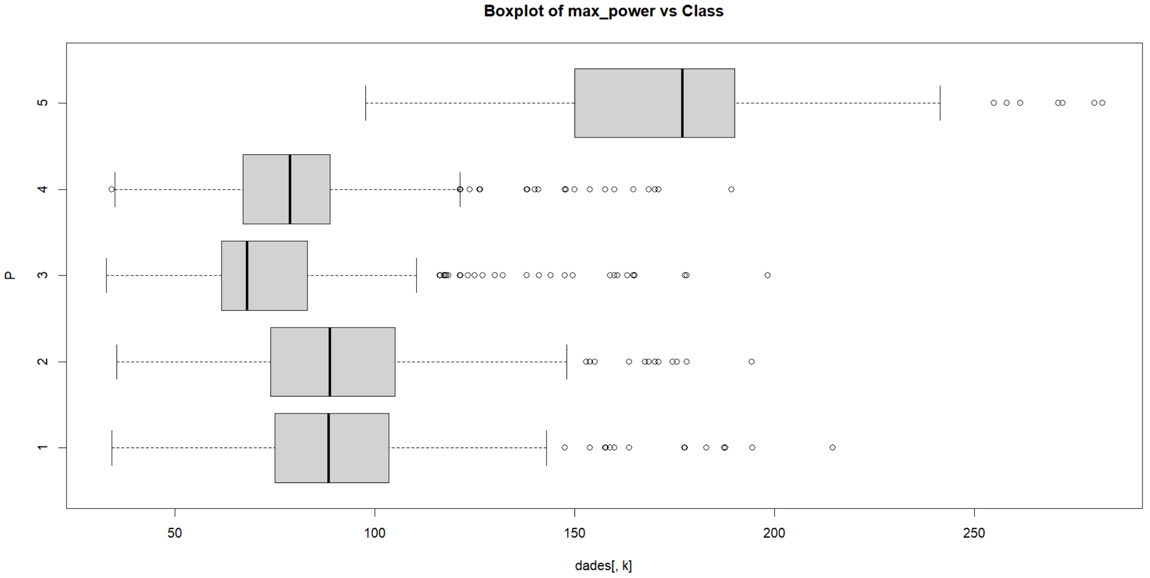
The variable **engine_size** separates clusters in two different categories : clusters 1, 3 and 4 have lower engine sizes whereas clusters 2 and 5 have generally higher engine sizes.



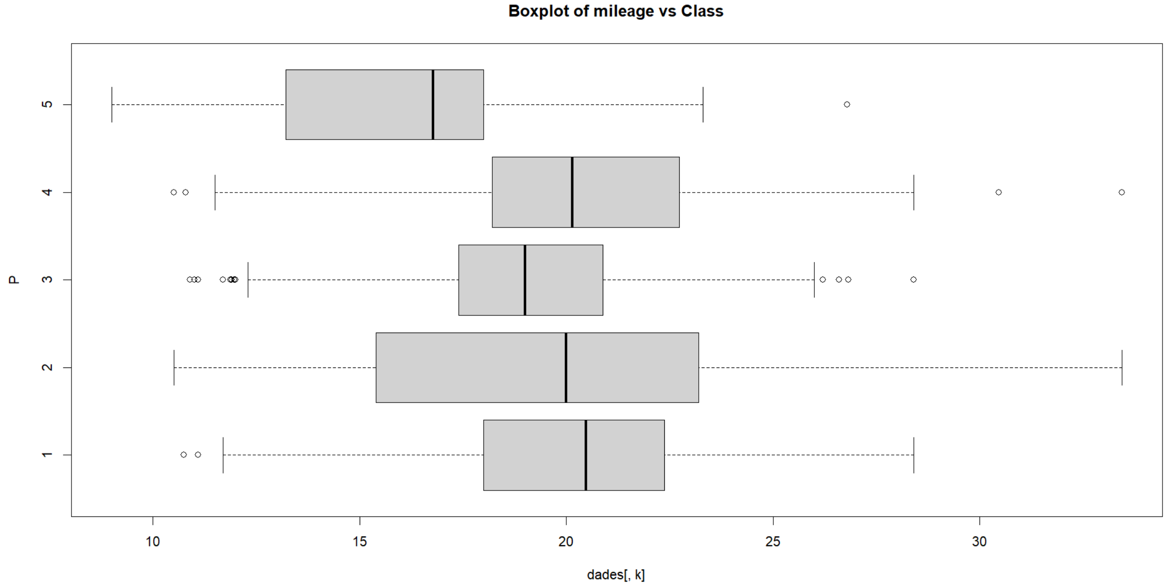
Clusters 2 and 3 have very distinct preferences for the type of **fuel** (Diesel and Petrol respectively) but the rest of the clusters give little information.



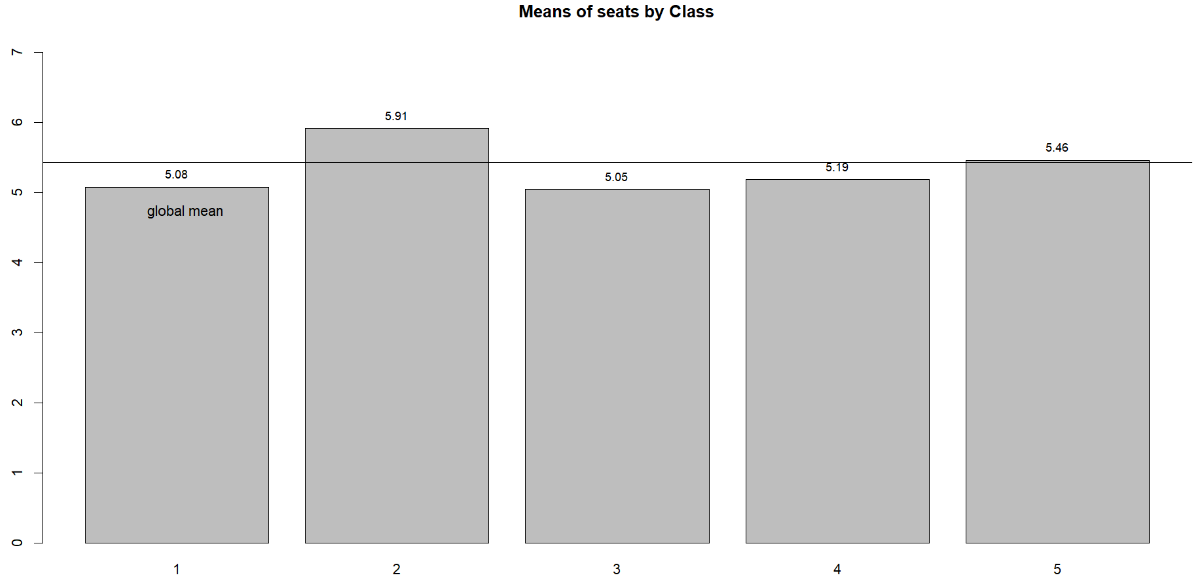
We can notice here that cluster 2 has notably high **kilometers driven** and that clusters 1 and 5 are noticeably lower.



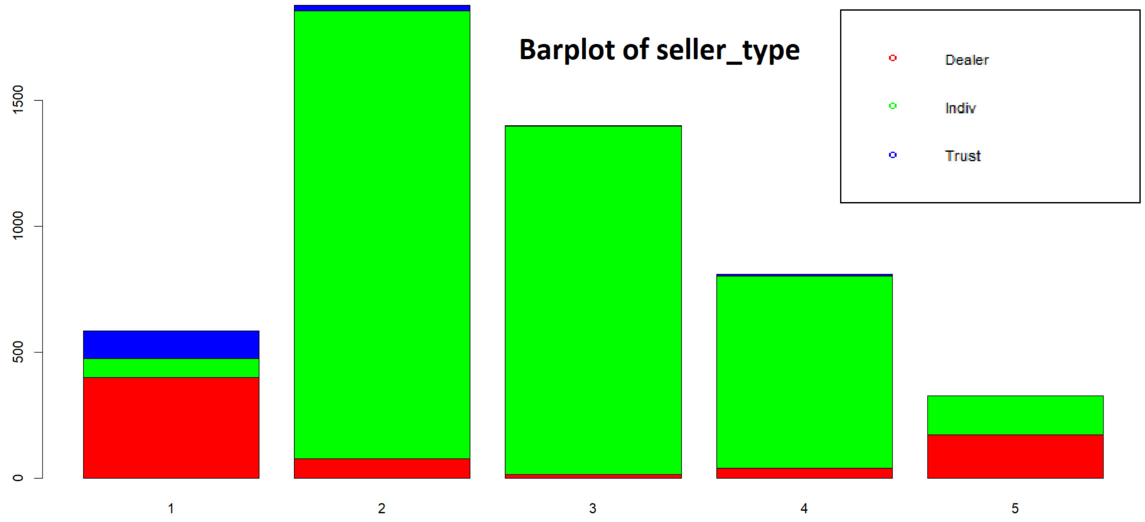
The most interesting aspect of this graph is the clear separation between cluster 5 and the other clusters in terms of **max power**. Clearly, cluster 5 has generally much more powerful cars. Cluster 3 also seems to have lower power than the other ones, though the distinction is much less noticeable.



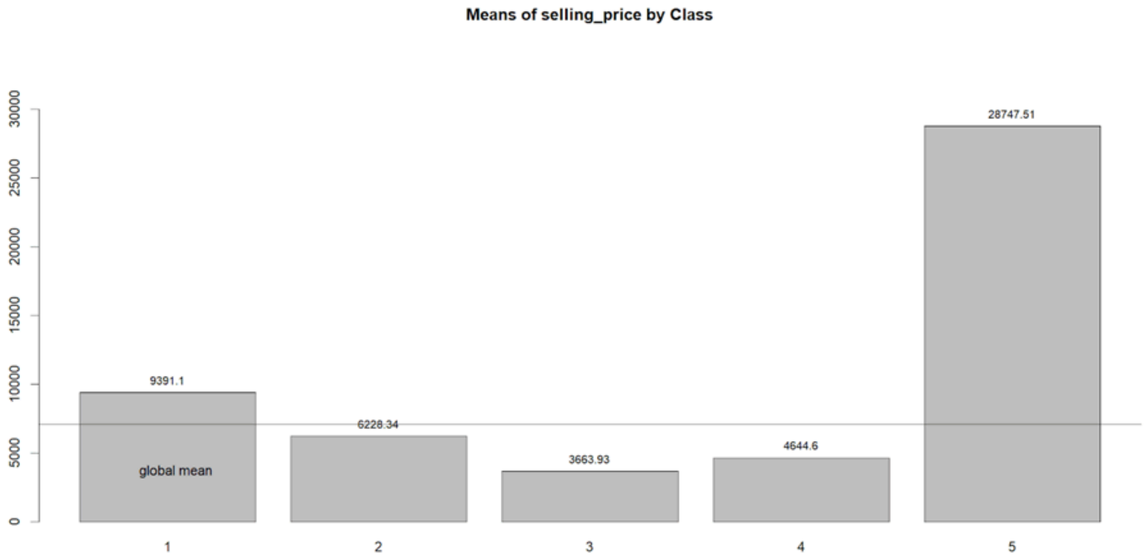
For mileage, again cluster 5 distinguishes itself, though in this case for having notably lower **mileage** than most. The only other interesting part of this graph is cluster 2 which has a noticeably wide range of values compared to the others.



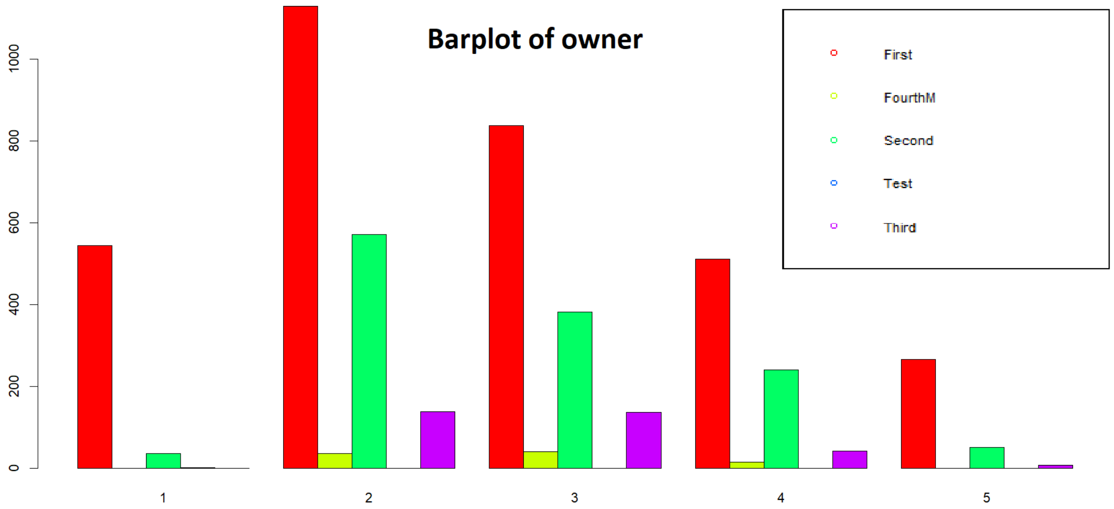
The graph for **seats** seems more uniform, but only because the minimum amount of seats a car can have is generally 2 or 3. Having this in mind allows us to notice that cluster 2 very much has a larger average amount of seats per car.



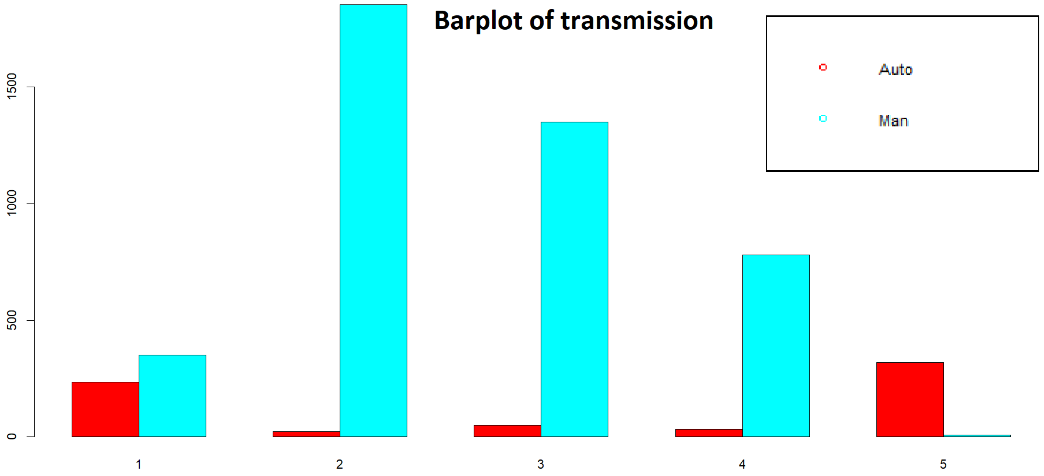
This graph shows that most clusters highly prefer Individual **seller types** (2, 3 and 4) but that cluster 1 specifically prefers Dealers. Cluster 5 shows no preference between the two.



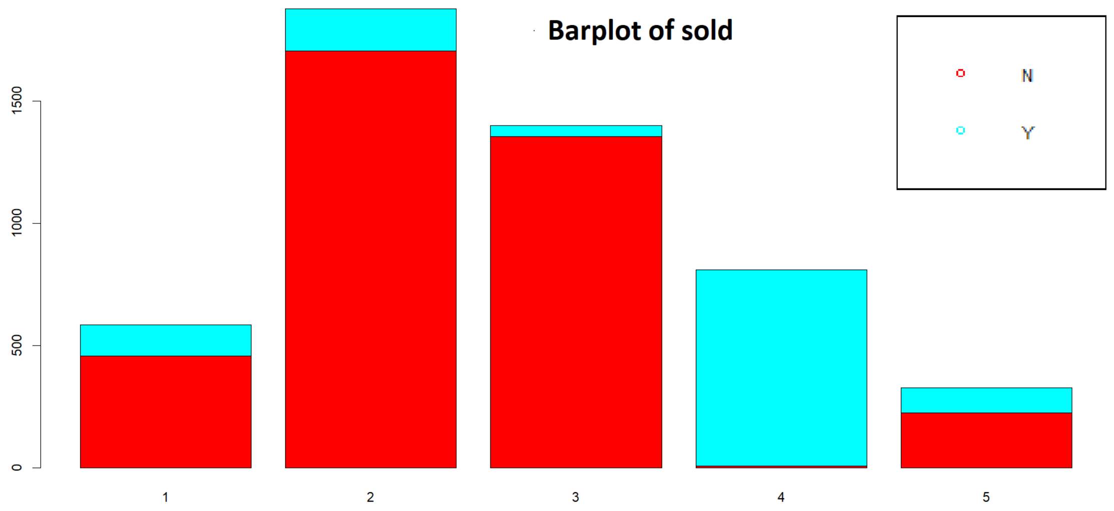
This graph shows that cluster 5 buys much higher priced cars than all others. Taking a closer look at the other clusters also tells us that cluster 1 also buys higher priced cars, though not to the same extent as cluster 5, and that cluster 3 **selling prices** are much lower.



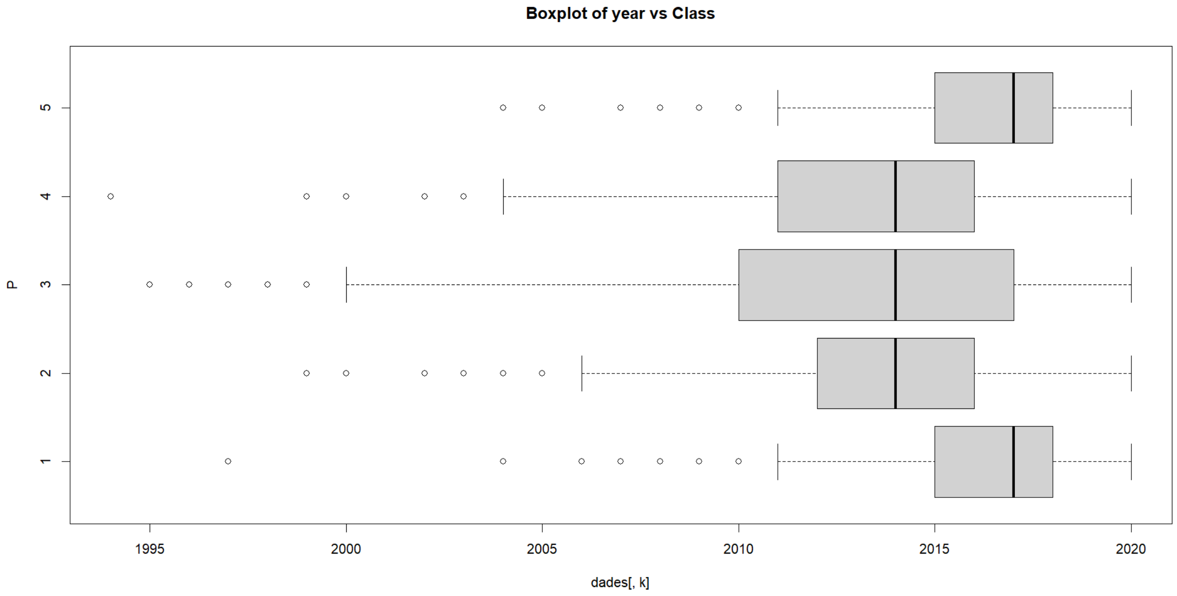
For there to be **owners** who are buying their second, third, fourth car, there needs to have been first owners before. This explains why the “First” modality is represented so much more. The distribution for each cluster is generally the same, with the exception of cluster 1 which has an even clearer preference for “First”.



For **Transmission**, we can notice that clusters 2, 3 and 4 have a clear preference for Manual transmissions, whereas cluster 5 has one for Automatic. Cluster 1 does not seem to have much of a preference either way.



We can notice that most clusters are unable to **sell** their car, with the clear exception of cluster 4 which was practically always able to.



We can notice here that clusters 1 and 5 tend to favor much **newer** cars, whereas cluster 3 has noticeably **older** cars. Clusters 2 and 4 have the same median as cluster 3 but their interval is much shorter than for cluster 3.

Observing the previous plots, the following table can be concluded.

Variable/Cluster	1	2	3	4	5
Engine Size	Average	/	Small	/	Big
Fuel	/	Diesel	Petrol	Half petrol/ half diesel	/
Kilometers driven	Low	Largest	/	Average	Low
Max power	/	/	Lowest	/	Largest
Seats	/	Most	/	/	/
Mileage	/	Largest range	/	/	Lowest
Owner	Only first	/	/	/	/
Selling price	Expensive	/	Cheap	/	Overly expensive
Seller type	Dealer	Individual	Individual	Individual	/
Transmission	/	Manual	Manual	Manual	Automatic
Sold	/	No	No	Yes	/
Year	New	/	Old	/	New

Cluster 1 appears to represent brand-new cars people would find on a dealership lot. They have average engine size which would probably correlate to midsize sedans from mainstream brands (Honda Civic, Hyundai Tucson, etc.). Because they are new, the price is relatively high compared to used cars, but not necessarily “luxury” expensive.

Cluster 2 could represent high-mileage diesel vehicles that seat more passengers than usual. Types of cars that fit that description are probably larger SUVs or passenger vans often used for family or commercial purposes (Ford Transit or Renault Traffic). They are typically sold by private owners rather than dealerships. Because of the very high mileage, they can be harder to sell, so they might stay on the market longer.

Cars that are budget-friendly and older small ones are represented in the **cluster 3**. They have a low horsepower and small engines which could represent entry-level hatchbacks or compact sedans (Ford Fiesta, Volkswagen Polo, etc.) from 5-10 years ago. That is also correlated to the fact they are generally manual transmissions. Because they are older and less powerful, they tend to be the cheapest cars on the used market.

Attributes that are only unique for **cluster 4** cars are the only ones being sold on the market. They are being sold by individuals and are also manual transmission. That could point to the cars that are family-orientated, reliable and used on a daily basis. They have been driven an average number of kilometers so they could offer 2-3 reliable years of service to buyers. Their fuel type is divided half-half with diesel and petrol which points out that buyers do not care what fuel type cars have as long as they are reliable.

Cluster 5 cars are premium, high-performance or luxury vehicles typically labeled as “supercars”. They could represent sports cars (Porsche 911, Chevrolet Corvette) or luxury SUVs (BMW X5/X7, Mercedes G-Class) with large engines (V6, V8 or higher). Because of their high power they have a low mileage which correlates to “supercars”. Also, automatic transmission is common in high-end modern luxury/performance cars with very high price tags.

Global discussion and general conclusions

Our main goal in this study was to evaluate the used car sales industry in the USA using a dataset of 5000 records that were selected randomly. The main objectives were to understand the factors that influence car prices, find characteristics that increase the probability of a sale and analyse regional differences in the market.

Preprocessing statistical analysis, PCA and clustering were among the structured data mining procedures we employed to get relevant information. Throughout the preprocessing phase, we ensured that the dataset was cleaned and standardized for analysis. After applying these techniques, we could achieve data that was clearer and which may give information about the sales at first sight.

The market for used cars showed important trends according to the statistical analysis. One of the most noteworthy conclusions was that the majority of transactions were conducted by individual sellers and manual transmissions were far more prevalent than automatic ones. Furthermore, the pricing distribution was as predicted, with the majority of cars being offered at comparatively low prices and only a small number having much higher prices. However, the data was still too general, and we could not see exactly what variables were related, which was important to take into account as not all buyers and sales are the same. For this reason, PCA Analysis and Clustering were still required.

PCA analysis revealed different patterns in the dataset, primarily highlighting the separation between older, lower-priced cars with higher mileage and newer, more expensive vehicles. The highest inverse correlation was found between the year/selling price and kilometers driven, which basically meant that the oldest and cheapest cars tend to have the most kilometers driven, which makes sense as they have been used the most, making their price lower due to wear and tear. Additionally, high-performance cars with greater max power tend to be positioned in the higher price range, while utility vehicles, such as vans and SUVs, go towards the lower range. With this data, sellers can already try to decide factors like pricing based on the qualities of the car taking into account other factors, for example, giving high prices to old cars with a lot of kilometers would not give good results. While

PCA provided a clear structure within the data, a clustering analysis was needed to identify well-defined groups.

So, after we applied hierarchical clustering using Ward’s method, we could group the data into five clusters based on vehicle attributes. Cluster 1 includes mainly new but average dealership lot cars, while Cluster 2 captures high-mileage diesel SUVs or passenger vans. Cluster 3 consists of very cheap, old but compact cars, and Cluster 4 represents reliable, family vehicles. Lastly, Cluster 5 groups high-performance luxury cars with large engines and premium pricing. These clusters may help sellers to decide on a group to focus their sales on, depending on the type of cars that will be sold.

There are some interesting parallels to draw between the information gathered from performing the PCA and the clustering / profiling. First of all, the PCA showed a strong correlation between selling price and max power. This correlation was found within clusters 1 and 5, where cars were generally more powerful and expensive. The PCA also showed a clear distinction between older, cheaper cars and newer, more expensive cars. The clustering reflected this distinction by separating cluster 3 from clusters 1 and 5. The other axis of the PCA showed a distinction between utility and power. The clustering once again reflected this distinction by separating clusters 2 and 3 from clusters 1 and 5.

However, there was some information found in the PCA that was not reflected in the profiling. The biggest is a very strong relation between kilometers driven and the year / selling price of the car. The clustering had the oldest, cheapest cars in cluster 3, but the cars with the most kilometers driven in cluster 2. The PCA also showed that cars with more seats had generally lower max power, but this correlation was not found within any of the clusters.

In conclusion, this research offered important views into the market for used cars highlighting patterns among vehicle attributes, cost and sales. These results could help sellers and buyers who want to understand the market better or optimize pricing strategies.

Working plan

Initial Gantt Diagram

	10/02							17/02							24/02							03/03							10/03							17/03						
TASK	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
D3																																										
Initial working plan(Gantt, Task division, Risk Plan)								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
Metadata file								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
Basic initial univariate descriptive								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
Enumeration of steps of the preprocessing								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
Justification of decisions for preprocessing								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
Additional descriptive statistics of variables								🟩	🟩	🟩	🟩	🟩	🟩	🟩																												
D4																																										
Motivation of the work and general description																						🟩	🟩	🟩	🟩																	
Cover + index + Data Source presentation																						🟩	🟩	🟩	🟩																	
Formal description of Data structure and metadata																										🟩	🟩	🟩	🟩													
Complete Data Mining process performed																						🟩	🟩	🟩	🟩																	
Detailed description of Preprocessing and data preparation																										🟩	🟩	🟩	🟩													
Basic statistical descriptive analysis																														🟩	🟩	🟩	🟩									
PCA analysis for numerical variables																										🟩	🟩	🟩	🟩													
Hierarchical Clustering on original data																										🟩	🟩	🟩	🟩													
Profiling of clusters																														🟩	🟩	🟩	🟩									
Global discussion and general conclusions																																		🟩	🟩	🟩	🟩					
Working plan																						🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
Making the ppt																																					🟩	🟩	🟩	🟩		

Final Gantt Diagram

	10/02							17/02							24/02							03/03							10/03							17/03						
TASK	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
D3																																										
Initial working plan(Gantt, Task division, Risk Plan)								🟩	🟩	🟩																																
Metadata file								🟩	🟩	🟩	🟩	🟩																														
Basic initial univariate descriptive								🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩																									
Enumeration of steps of the preprocessing								🟩	🟩	🟩																																
Justification of decisions for preprocessing								🟩	🟩	🟩	🟩	🟩																														
Additional descriptive statistics of variables								🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩																								
D4																																										
Motivation of the work and general description																					🟩	🟩	🟩																			
Cover + index + Data Source presentation																					🟩	🟩	🟩																			
Formal description of Data structure and metadata																					🟩	🟩	🟩																			
Complete Data Mining process performed																					🟩	🟩	🟩	🟩																		
Detailed description of Preprocessing and data preparation																											🟩	🟩	🟩	🟩	🟩	🟩										
Basic statistical descriptive analysis																													🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
PCA analysis for numerical variables																																				🟩	🟩	🟩	🟩			
Hierarchical Clustering on original data																																				🟩	🟩	🟩	🟩			
Profiling of clusters																																				🟩	🟩	🟩	🟩			
Global discussion and general conclusions																																			🟩	🟩	🟩	🟩	🟩			
Working plan																					🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩			
Making the ppt																																						🟩	🟩	🟩		
Preparing the presentation																																									🟩	🟩

Tasks assignment grid

Participant	Álvaro	Alex	Joan	Matija	Otman
Definition and projects assignment	x	x	x	**X**	x
Mail for team consolidation	**X**				
D3					
Initial working plan(Gantt, Task division, Risk Plan)					**X**
Metadata file			**X**		x
Basic initial univariate descriptive				**X**	
Enumeration of steps of the preprocessing	x	x	**X**		
Justification of decisions for preprocessing	x	**X**			
Additional descriptive statistics of variables				**X**	
D4					
Motivation of the work and general description of the problem to be analyzed	**X**				
Cover + index + Data Source presentation		**X**			
Formal description of Data structure and metadata	**X**				
Complete Data Mining process performed	**X**				
Detailed description of Preprocessing and data preparation			**X**		x
Basic statistical descriptive analysis		**X**	x		x
PCA analysis for numerical variables	x	x		**X**	
Hierarchical Clustering on original data			**X**	x	
Profiling of clusters		x		**X**	
Global discussion and general conclusions of the whole work	x	x	x	x	**X**
Working plan					**X**
Making the ppt	x	**X**	x	x	x

Deviations from the Original Schedule

At the beginning of the project, we established a Gantt Diagram, setting one week for all tasks in D3 and four days for each task in D4. However, as the project progressed, some tasks from D3, like the descriptive analysis of variables and data preprocessing, took longer than we had expected, more specifically, handling outliers, transforming categorical variables, and ensuring that all the data was correct required more time than anticipated, affecting the time to do the next tasks.

Another significant deviation was encountered during the clustering and profiling phases. Initially planned as a straightforward step, these phases were delayed because we had done an incorrect descriptive analysis, which the teacher later corrected. As a result, we needed extra time to revise it in the project, which in turn delayed tasks such as selecting the optimal number of clusters and interpreting results. Also, since we had been working with an incorrect version, we had to regenerate some steps for clustering and profiling, which required more steps and more time. Despite this, adjustments were made to be able to finish these tasks without compromising the overall timeline.

On the other hand, some tasks were completed faster than expected, especially in D4, as they involved work that had already been done in D3. This allowed us to use the existing information in the report rather than starting over and ensuring that the overall workflow remained steady.

Risk Management: Avoided and Unexpected Risks

Throughout the project, we prevented several risks while also managing some other challenges.

Firstly, preprocessing required careful decisions as any inappropriate transformation would have the risk of changing the data completely, therefore we were verifying in each step that the data was still correlated to the previous version. Similarly, incorrect data types were identified and corrected, ensuring a better numerical analysis. We also prevented the risk of losing information by carefully evaluating preprocessing decisions, preserving important information in categorical variables. Finally, some other errors in plot generation were avoided by validating plots with descriptive statistics before reaching any conclusions.

Some issues required both prevention and management of the risk. For outliers, we used boxplots to detect extreme values, but still had to manage specific cases. Instead of removing them directly, we applied transformations where necessary to be able to interpret them. Apart from this, most variables had clear definitions and units, but some others like mileage, engine, and max_power lacked them, therefore we assigned appropriate units based on context in order to continue our study.

On the other hand, an unexpected problem appeared when two team members left the project, but even if it meant more workload for each person it has been managed by doing an equal distribution of tasks.

To conclude, a consistent validation in each transformation of the data and team coordination allowed us to prevent or manage all identified risks.

R Scripts

preprocessing.r

```
dd<- read.csv("PreprocessedUserCarData.csv",header=T, sep=",", stringsAsFactors = TRUE)
dd <- dd[, -1]  # Removes the first column
dd <- dd[-4857,] # Removes the row with a lot of km_driven
names(dd)[names(dd) == "engine..cm3."] <- "engine_size"
names(dd)[names(dd) == "Region"] <- "region"
names(dd)[names(dd) == "State.or.Province"] <- "state"
write.csv(dd, "PreprocessedUserCarData_modified.csv", row.names = FALSE)
```

desc.Rmd

```
dd<- read.csv("PreprocessedUserCarData_modified.csv",header=T, sep=",", stringsAsFactors = TRUE)
class(dd)
n<-dim(dd)[1]
K<-dim(dd)[2]
print(paste("Entries: ", n))
print(paste("Variables: ", K))
names(dd)
descriptiva<-function(X, nom){
  if (!(is.numeric(X) || class(X)=="Date")){
    frecs<-table(as.factor(X), useNA="ifany")
    frecs<-sort(frecs, decreasing = TRUE)
    proportions<-frecs/n
    #ojo, decidir si calcular porcentages con o sin missing values
    if (!nom %in% c("name", "state", "fuel")) {
      pie(frecs, cex = 0.6, main = paste("Pie of", nom))

    }
    name_size = 0.7
    if (nom=="state"){
      name_size = 0.6
    }
    if (!nom %in% c("name", "state")) {
      upper_limit <- max(frecs) * 1.1
      bp<-barplot(frecs, las=2, cex.names=name_size, ylim=c(0,upper_limit), main=paste("Barplot of", nom), col=listOfColors)
      if (nom %in% c("fuel", "owner")){
        text(bp[1:2], frecs[1:2], labels = frecs[1:2], pos = 3, cex = 0.8, col = "black")
      }

    } else {

      top_n <- 10
      if (nom=="state"){
        top_n <- 25
      }
      top_freqs <- frecs[1:top_n]
      others_sum <- sum(frecs[(top_n+1):length(frecs)])
      final_freqs <- c(top_freqs, Others = others_sum)

      barplot_title = paste("Barplot of", nom, "for top", top_n, "modalities")
      if (nom=="name"){
        barplot_title <- paste("Barplot of", nom, "for top", top_n, "models")
      }

      upper_limit <- max(final_freqs) * 1.1
      bp <- barplot(final_freqs, las = 2, cex.names = name_size,
            col = rainbow(length(final_freqs)), main = barplot_title, ylim=c(0,upper_limit))

      if (nom=="name" ){
        text(bp[1:2], final_freqs[1:2], labels = final_freqs[1:2], pos = 3, cex = 0.8, col = "black")
      }

      if (nom=="state"){
        text(bp[1], final_freqs[1], labels = final_freqs[1], pos = 3, cex = 0.8, col = "black")
        text(bp[26], final_freqs[26], labels = final_freqs[26], pos = 3, cex = 0.8, col = "black")
      }
    }

    print(paste("Number of modalities: ", length(frecs)))
    print("Frequency table")
```

```
 print(frecs)
 print("Relative frequency table (proportions)")
 print(proportions)
 print("Frequency table sorted")
 print(sort(frecs, decreasing=TRUE))
 print("Relative frequency table (proportions) sorted")
 print(sort(proportions, decreasing=TRUE))
}else{
  if(class(X)=="Date"){
    print(summary(X))
    print(sd(X))
    #decide breaks: weeks, months, quarters...
    hist(X,breaks="weeks")
  }else{
    h<-hist(X, main=paste("Histogram of", nom), plot=FALSE)
    hist(X, main = paste("Histogram of", nom), ylim = c(0, max(h$counts)*1.1))
    if (nom %in% c("year", "mileage")){
      mean_val <- mean(X)
      abline(v=mean_val, col="red", lwd=2)
      x_align <- 4
      if (nom=="mileage"){
        x_align <- 8
      }
      text(x = mean_val-x_align, y = max(h$counts) * 0.9,
           labels = paste("Mean =", round(mean_val, 1)),
         pos = 3, col = "red", cex = 0.8)
    }
    if (nom=="selling_price"){
      range_texts <- paste0(h$breaks[1:2], " - ", h$breaks[2:3], "€")
      text(
           x      = h$mids[1:2],            # bin midpoints
           y      = h$counts[1:2] / 2,      # half the bin height
           labels = range_texts,
           col    = "orange",
           srt    = 90,
           cex    = 0.7
         )
    }
    count_indices<-2
    if (nom=="mileage"){
      count_indices<-3
    }
    highest_indices <- order(h$counts, decreasing = TRUE)[1:count_indices]
      text(x = h$mids[highest_indices],           # x-coordinates of the bar midpoints
            y = h$counts[highest_indices],         # y-coordinates = bar heights
            labels = h$counts[highest_indices],    # what to label with
            pos = 3,              # position 3 = above
            cex = 0.7,            # size of the text
            col = "blue")
    boxplot(X, horizontal=TRUE, main=paste("Boxplot of",nom))

    stats <- boxplot.stats(X)$stats

     q1_val     <- stats[2]
     median_val <- stats[3]
     q3_val     <- stats[4]
     mean_val   <- mean(X)

     text(x = q1_val,
          y = 0.7,
          labels = paste0("Q1=", round(q1_val, 0)),
          cex=0.7,
          col = "blue")

     # Q3
     text(x = q3_val,
          y = 0.6,
          labels = paste0("Q3=", round(q3_val, 0)),
          cex=0.7,
          col = "blue")
      text(x = median_val,
      y = 1.3,
      labels = paste0("Median=", round(median_val, 0)),
      cex=0.7,
      col = "darkgreen")
      print("Extended Summary Statistics")
      print(summary(X))
      print(paste("sd: ", sd(X, na.rm=TRUE)))
      print(paste("vc: ", sd(X, na.rm=TRUE)/mean(X, na.rm=TRUE)))
     }
  }
}
dataset<-dd
actives<-c(1:K)
colDate<-1
```

PCA.r

```
dd<- read.csv("PreprocessedUserCarData_modified.csv",header=T, sep=",", stringsAsFactors = TRUE)
objects()
attributes(dd)

# VISUALISATION OF DATA
# PRINCIPAL COMPONENT ANALYSIS OF CONTINUOUS VARIABLES
# CREATION OF THE DATA FRAME OF CONTINUOUS VARIABLES

attach(dd)
names(dd)

#is R understanding well my factor variables?
sapply(dd,class)
numeriques<-which(sapply(dd,is.numeric))
dcon<-dd[,numeriques]
sapply(dcon,class)

# PRINCIPAL COMPONENT ANALYSIS OF dcon
# Instruction of the PCA - magic line
pc1 <- prcomp(dcon, scale=TRUE)

# Screeplot
screeplot(pc1, type = "lines", main = "Scree Plot of PCA")
points(pc1$sdev^2, col = "blue", pch = 19)
title(xlab = "Components")

eigenvalues <- pc1$sdev^2
cumvar <- cumsum(eigenvalues) / sum(eigenvalues)

# SELECTION OF THE SINGIFICNT DIMENSIONS (keep 80% of total inertia)
# How many values describe 80% of total inertia
nd <- min(which(cumvar >= 0.8))
cat("Number of principal components to retain (80% variance):", nd, "\n")

abline(v = nd, col = "red", lty = 2)
text(nd, max(eigenvalues),
     labels = paste("PC", nd, "\n(80% threshold)"),
     pos = 4, col = "red")

class(pc1)
attributes(pc1)
print(pc1)
str(pc1)

# WHICH PERCENTAGE OF THE TOTAL INERTIA IS REPRESENTED IN SUBSPACES?
pc1$sdev
inerProj<- pc1$sdev^2
inerProj
totalIner<- sum(inerProj)
totalIner
pinerEix<- 100*inerProj/totalIner
pinerEix
barplot(pinerEix)
#Cummulated Inertia in subspaces, from first principal component to the 7th dimension subspace
bar_colors <- rep("gray", dim(dcon)[2])
bar_colors[1:nd] <- "blue"
barplot(100*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2], main="Cummulated Inertia",
        xlab="Components", ylab="Cummulated percentage", names.arg=1:dim(dcon)[2],
        col=bar_colors)
percInerAccum<-100*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2]
percInerAccum
```

```
print(pc1)
pc1$rotation

# STORAGE OF THE EIGENVALUES, EIGENVECTORS AND PROJECTIONS IN THE nd DIMENSIONS
View(pc1$x)
dim(pc1$x)
dim(dcon)
dcon[2000,]
pc1$x[2000,]

Psi = pc1$x[,1:nd]
dim(Psi)
Psi[2000,]

# STORAGE OF LABELS FOR INDIVIDUALS AND VARIABLES
iden = row.names(dcon)
etiq = names(dcon)
ze = rep(0,length(etiq)) # WE WILL NEED THIS VECTOR AFTERWARDS FOR THE GRAPHICS

# PLOT OF INDIVIDUALS
#select your axis
#which domination would you like to show?
eje1<-1
eje2<-2

combinations <- combn(1:3, 2)
n <- ncol(combinations)

for (i in 1:n) {
  comp <- combinations[, i]  # This is a vector of two component indices
  eje1 <- comp[1]
  eje2 <- comp[2]

  # Create the plot for the chosen pair
  plot(Psi[, eje1], Psi[, eje2], type = "n",
       main = paste("PC", eje1, "and PC", eje2),
       xlab = paste("PC", eje1), ylab = paste("PC", eje2))

  # Plot points with a smaller radius
  points(Psi[, eje1], Psi[, eje2], pch = 19, cex = 0.4)

  # Add axes with custom styling
  axis(side = 1, pos = 0, labels = FALSE, col = "cyan")
  axis(side = 3, pos = 0, labels = FALSE, col = "cyan")
  axis(side = 2, pos = 0, labels = FALSE, col = "cyan")
  axis(side = 4, pos = 0, labels = FALSE, col = "cyan")
}
library(rgl)
plot3d(Psi[,1],Psi[,2],Psi[,3])

#Projection of variables

Phi = cor(dcon,Psi)
View(Phi)

#select your axis
# shows on first two components arrows of numerical variables
#zoomed in
X<-Phi[,eje1]
Y<-Phi[,eje2]
plot(Psi[,eje1],Psi[,eje2],main="Numeric variables factor map", xlab="PC1", ylab="PC2"
     ,type="n",xlim=c(min(X,0),max(X,0)), ylim=c(-1,1))
axis(side=1, pos= 0, labels = F)
axis(side=3, pos= 0, labels = F)
axis(side=2, pos= 0, labels = F)
axis(side=4, pos= 0, labels = F)
arrows(ze, ze, X, Y, length = 0.07,col="blue")
text(X,Y,labels=etiq,col="darkblue", cex=0.7)

# From here we add qualitative of centroids to the graphs
# PROJECTION OF ILLUSTRATIVE qualitative variables on individuals' map
# 10th column is owner
#all qualitative together
# if you put name in plot then xlim=c(-4,5), ylim=c(-2.5,2.5)
# plot(Psi[,eje1],Psi[,eje2],main="Qualitive centroids", xlab="PC1", ylab="PC2",type="n",xlim=c(-4,5), ylim=c(-2.5,2.5))

#if all else xlim=c(-2,4), ylim=c(-1,1.5)
plot(Psi[,eje1],Psi[,eje2], main="Qualitive centroids", xlab="PC1", ylab="PC2",type="n",xlim=c(-2,4), ylim=c(-1,1.5))

axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

#nominal qualitative variables
# Which qualitative to plot in graph
#dcat<-c(1)
dcat<-c(7,8,9,10)
#divide categoricals in several graphs if joint representation saturates

#build a palette with as much colors as qualitative variables

colors<-c("blue","green","orange","darkgreen")
#alternative
#colors<-rainbow(length(dcat))

c<-1
for(k in dcat){
  seguentColor<-colors[c]
  fdic1 = tapply(Psi[,eje1],dd[,k],mean)
  fdic2 = tapply(Psi[,eje2],dd[,k],mean)

  text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)
  c<-c+1
}
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)

#represent numerical variables in background
#if you put name in plot then xlim=c(-4,5), ylim=c(-2.5,2.5)
#plot(Psi[,eje1],Psi[,eje2],main="Variables factor map with qualitative centroids", xlab="PC1", ylab="PC2",type="n",xlim=c(-4,5), ylim=c(-2.5,2.5))

# if all else xlim=c(-2,4), ylim=c(-1,1.5)
plot(Psi[,eje1],Psi[,eje2],main="Variables factor map with qualitative centroids",
     xlab="PC1", ylab="PC2",type="n",xlim=c(-2,4), ylim=c(-1,1.5))
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

#add projections of numerical variables in background
arrows(ze, ze, X, Y, length = 0.07,col="red")
text(X,Y,labels=etiq,col="red", cex=0.7)

#add centroids
c<-1
for(k in dcat){
  seguentColor<-colors[c]

  fdic1 = tapply(Psi[,eje1],dd[,k],mean)
  fdic2 = tapply(Psi[,eje2],dd[,k],mean)
  #points(fdic1,fdic2,pch=16,col=seguentColor, labels=levels(dd[,k]))
  text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)
  c<-c+1
}
legend("bottomright",names(dd)[dcat],pch=1,col=colors, cex=0.6)
```

clustering_profiling.R

```
dd<- read.csv("PreprocessedUserCarData_modified.csv",header=T, sep=",", stringsAsFactors = TRUE)

names(dd)
dim(dd)
summary(dd)

attach(dd)
names(dd)

#hierarchical clustering
#euclidean distance si totes son numeriques
```

```
dcon <- data.frame (year,selling_price,km_driven,mileage,engine_size,max_power,seats)

# euclidean distances between first 15 rows
d<-dist(dcon[1:15,])

#move to Gower mixed distance to deal
#simultaneously with numerical and qualitative data
library(cluster)
#dissimilarity matrix
#do not include in actives the identifier variables nor the potential response variable

actives<-c(1:15)
dissimMatrix <- daisy(dd[,actives], metric = "gower", stand=TRUE)

distMatrix<-dissimMatrix^2

h1 <- hclust(distMatrix,method="ward.D2")# NOTICE THE COST
#versions noves "ward.D" i abans de plot: par(mar=rep(2,4)) si se quejara de los margenes del plot

plot(h1, labels=FALSE)

#number of classes to cut, change this
k<-5
c1 <- cutree(h1,4)
table(c1)
c2 <- cutree(h1,k)
#class sizes
table(c2)

#comparing with other partitions
table(c1,c2)

# LETS SEE THE PARTITION VISUALLY
# cut is better with 5
c1<-c2
table(c1)
plot(year,km_driven,col=c1,main="Clustering of cars in 5 classes")
legend("topleft",c("class1","class2","class3","class4","class5"),pch=1,col=c(1:k),cex=0.6)

plot(max_power,engine_size)
plot(max_power,engine_size,col=c1,main="Clustering of cars in 5 classes")
legend("topright",c("class1","class2","class3","class4","class5"),pch=1,col=c(1:k), cex=0.6)

plot(selling_price,max_power,col=c1,main="Clustering of cars in 5 classes")
legend("topleft",c("class1","class2","class3","class4","class5"),pch=1,col=c(1:k), cex=0.6)

#pairs(dcon[,1:7], col=c1)

# LETS SEE THE QUALITY OF THE HIERARCHICAL PARTITION

cdg <- aggregate(as.data.frame(dcon),list(c1),mean)
cdg
plot(cdg[,1], cdg[,7], xlab=names(cdg)[1], ylab=names(cdg)[7])
names(dd)
cdg <- aggregate(as.data.frame(dcon),list(c2),mean)
cdg
plot(selling_price, max_power, col= c2)
points(cdg[,3],cdg[,7],pch=16,col="yellow")
text(cdg[,3],cdg[,7], labels=cdg[,1], pos=2, font=2, cex=0.8, col="yellow")

potencials<-c(1,2,3,4,5,6,7)
#pairs(dcon[,potencials],col=c2)

tt<-table(seats,c2)
tt

#Profiling plots, very important from here

ValorTestXnum <- function(Xnum,P){
  #freq dis of fac
  nk <- as.vector(table(P));
  n <- sum(nk);
  #mitjanes x grups
  xk <- tapply(Xnum,P,mean);
  #valors test
  txk <- (xk-mean(Xnum))/(sd(Xnum)*sqrt((n-nk)/(n*nk)));
  #p-values
  pxk <- pt(txk,n-1,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){if (pxk[c]>0.5){pxk[c]<-1-pxk[c]}}
  return (pxk)
}
ValorTestXquali <- function(P,Xquali){
  taula <- table(P,Xquali);
  n <- sum(taula);
  pk <- apply(taula,1,sum)/n;
  pj <- apply(taula,2,sum)/n;
  pf <- taula/(n*pk);
  pjm <- matrix(data=pj,nrow=dim(pf)[1],ncol=dim(pf)[2], byrow=TRUE);
  dpf <- pf - pjm;
  dvt <- sqrt(((1-pk)/(n*pk))%*%t(pj*(1-pj)));
  #i hi ha divisions iguals a 0 dona NA i no funciona
  zkj <- dpf
  zkj[dpf!=0]<-dpf[dpf!=0]/dvt[dpf!=0];
  pzkj <- pnorm(zkj,lower.tail=F);
  for(c in 1:length(levels(as.factor(P)))){for (s in 1:length(levels(Xquali))){if (pzkj[c,s]> 0.5){pzkj[c,s]<-1- pzkj[c,s]}}}
  return (list(rowpf=pf,vtest=zkj,pval=pzkj))
}

#dades contain the dataset
dades<-dd
K<-dim(dades)[2]
par(ask=TRUE)
#P must contain the class variable
#P<-dd[,3]
P<-c2
#P<-dd[,18]
nameP<-"classe"
#P<-df[,33]
nc<-length(levels(factor(P)))
nc
pvalk <- matrix(data=0,nrow=nc,ncol=K, dimnames=list(levels(P),names(dades)))
nameP<-"Class"
n<-dim(dades)[1]

for(k in 1:K){
  if (is.numeric(dades[,k])){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))
    bxp <- boxplot(dades[,k]~P, main=paste("Boxplot of", names(dades)[k], "vs", nameP ), horizontal=TRUE)
    bar_heights <- tapply(dades[[k]], P, mean)
    bp <- barplot(tapply(dades[[k]], P, mean),main=paste("Means of", names(dades)[k], "by", nameP ),
                  ylim = c(0, max(bar_heights) * 1.2))
    text(x = bp,
         y = bar_heights,
         labels = round(bar_heights, 2),  # or use formatC() for custom formatting
         pos = 3, cex = 0.8)
    abline(h=mean(dades[[k]]))
    legend(0,mean(dades[[k]]),"global mean",bty="n")
    print("Estadístics per groups:")
    for(s in levels(as.factor(P))) {print(summary(dades[P==s,k]))}
    o<-oneway.test(dades[,k]~P)
    print(paste("p-valueANOVA:", o$p.value))
    kw<-kruskal.test(dades[,k]~P)
    print(paste("p-value Kruskal-Wallis:", kw$p.value))
    pvalk[,k]<-ValorTestXnum(dades[,k], P)
    print("p-values ValorsTest: ")
    print(pvalk[,k])
  }else{
    if(class(dd[,k])=="Date"){
      print(summary(dd[,k]))
      print(sd(dd[,k]))
      #decide breaks: weeks, months, quarters...
      hist(dd[,k],breaks="weeks")
    }else{
      #qualitatives
      print(paste("Variable", names(dades)[k]))
      table<-table(P,dades[,k])
      #   print("Cross-table")
      #   print(table)
      rowperc<-prop.table(table,1)
      colperc<-prop.table(table,2)
      #  print("Distribucions condicionades a files")
      # print(rowperc)
      dades[,k]<-as.factor(dades[,k])
      marg <- table(as.factor(P))/n
      print(append("Categories=",levels(as.factor(dades[,k]))))
      #from next plots, select one of them according to your practical case
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }
      #with legend
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }
      legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)
       #condicionades a classes
      print(append("Categories=",levels(dades[,k])))
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(rowperc[,c],col=paleta[c]) }
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))
      paleta<-rainbow(length(levels(dades[,k])))
      for(c in 1:length(levels(dades[,k]))){lines(rowperc[,c],col=paleta[c]) }
      legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)

      #amb variable en eix d'abcisses
      marg <-table(dades[,k])/n
      print(append("Categories=",levels(dades[,k])))
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)
      #x<-plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), xaxt="n")
      #text(x=x+.25, y=-1, adj=1, levels(CountryName), xpd=TRUE, srt=25, cex=0.7)
      paleta<-rainbow(length(levels(as.factor(P))))
      for(c in 1:length(levels(as.factor(P)))){lines(rowperc[c,],col=paleta[c]) }

      #with legend
      plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)
      for(c in 1:length(levels(as.factor(P)))){lines(rowperc[c,],col=paleta[c])}
      legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)

      #condicionades a columna
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)
      paleta<-rainbow(length(levels(as.factor(P))))
      for(c in 1:length(levels(as.factor(P)))){lines(colperc[c,],col=paleta[c]) }

      #with legend
      plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)
      for(c in 1:length(levels(as.factor(P)))){lines(colperc[c,],col=paleta[c])}
      legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)

      table<-table(dades[,k],P)
      print("Cross Table:")
      print(table)
      print("Distribucions condicionades a columnes:")
      print(colperc)
      paleta<-rainbow(length(levels(dades[,k])))
      barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta )

      barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta )
      legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)
      barplot(table(dades[,k], as.factor(P)), beside=TRUE,col=paleta )
      barplot(table(dades[,k], as.factor(P)), beside=TRUE,col=paleta)
      legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)
      print("Test Chi quadrat: ")
      print(chisq.test(dades[,k], as.factor(P)))

      print("valorsTest:")
      print( ValorTestXquali(P,dades[,k]))
      #calcular els pvalues de les quali
    }
  }
}#endfor
```