U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Professor Maria Paula Brito

Cars Sales Study

Statistics and Data Analysis

June, 2024

José Rodrigues - 202006455

Pedro Leite - 201906697

Table of Contents

1. Introduction

The increased dependence on a commute through our own vehicles is a part of life that generally increases as we grow older. The commitment of buying that specific vehicle is one that is affected by a variety of factors, mainly due to our preferences and our economic availability. In this project we analyze a dataset that was found in [1] Kaggle containing information about different vehicles and their sales. Through the statistical analysis of the observations in the dataset, we want to address some questions:

- What are the cheapest and/or most costly brands and models?
- What are the brands and models with the highest and lowest sales?
- Is there any correlation between original prices and their resale values?
- What are the variables (or combination of variables) that have the highest effect on the vehicle prices and/or sales?
- What is more important in the sale of a car, marketability/name brand or the car's specific attributes?

2. Materials

2.1. Dataset

The dataset has various vehicle models, with a lot of information about their specs and value. It has 156 observations, each representing a different model. From the 16 variables, 4 are categorical and 12 are numerical:

- "Manufacturer": Nominal categorical variable, that represents the brand of the car. We renamed it as "Brand".
- "Model": Nominal categorical variable, that represents the model of the car.
- "Vehicle_type": Nominal categorical variable, that represents the type of vehicle. It can be either "Car", a vehicle of a smaller scale, for personal or private use, or "Passenger", a vehicle of a bigger scale, that can be used to transport several people. We renamed it as "Type".
- "Latest_Launch": Ordinal categorical variable, that represents the date of the launch of the vehicle model.

- “Sales_in_thousand": Continuous numerical variable (ratio scale), that represents the amount of sales, in thousands.
- “Price_in_thousands”: Continuous numerical variable (ratio scale), that represents the price, in thousands of dollars.
- “X__year_resale_value”: Continuous numerical variable (ratio scale), that represents the price in thousands of dollars, after X amount of years. We renamed it as “Resale_Value”.
- “Engine_size”: Continuous numerical variable (ratio scale), that represents the size of the engine, in liters.
- “Wheelbase”: Continuous numerical variable (ratio scale), that represents the distance between the wheel, in inches.
- “Width”: Continuous numerical variable (ratio scale), that represents the width, in inches.
- “Length”: Continuous numerical variable (ratio scale), that represents the length, in inches.
- “Curb_weight": Continuous numerical variable (ratio scale), that represents the total weight, in thousands of pounds.
- “Fuel_capacity”: Continuous numerical variable (ratio scale), that represents the maximum volume of fuel that the vehicle can take, in gallons.
- “Fuel_efficiency”: Continuous numerical variable (ratio scale), that represents due distance the vehicle can travel, in miles per gallon.
- “Horsepower”: Continuous numerical variable (ratio scale), that represents the amount of horsepower.
- “Power_perf_factor”: Continuous numerical variable (ratio scale), that represents the factor between power and performance. We renamed it as “Power_Performance_Factor”.

2.2. Packages

For this project, we used the programming language “R”. With the packages: “tidyverse”, “dplyr”, “ggplot2”, “ggrepel”, “forcats”, “sccales”, “gridExtra”, “corrplot”, “e1071”, “MASS”, “caret”, “tidyr”, “FactoMineR”, “cluster”, “dendextend”, “caTools”, “psych”, “car”.

3. Data Cleaning

We started by creating a new categorical variable, "Brand_Model", that is the result of the concatenation of "Brand" and "Model". In order to have all the information in just 1 variable. We've also converted the variable "Type" to binary, since it only has 2 possible options. Where "Passenger" is 1, and "Car" is 0. "Latest_Launch", was converted to a discrete numerical variable that represents the years since launch. So, we subtracted the "Latest_Launch" from the current date, and converted the result to years. The name of the variable is now "Years_Launch".

Brand_Model	Years_Launch	Type
Acura Integra	12	1
Acura TL	13	1
Acura CL	12	1
Acura RL	13	1
Audi A4	13	1
Audi A6	13	1

We've checked for rows with missing values. And noticed that row 34, is the only one with more than 1 null value, with 10. So we removed it.

```
> na_rows_count
  [1]  0  0  2  0  0  0  0  1  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0
 [34] 10  1  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [67]  1  0  0  0  0  0  1  0  1  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1
[100]  1  1  0  0  0  0  0  1  1  0  2  1  0  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  1  1  0  0  0
[133]  1  1  1  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1
```

We've also checked for variables with missing values. And replaced these values, in the variables that have them, with its mean.

```
> na_variables_count
                 Brand                  Model        Sales_in_Thousands           Resale_Value
                     0                      0                        0                     36
     Price_in_Thousands            Engine_Size             Horsepower              Wheelbase
                     1                      0                        0                      0
                 Width                 Length            Curb_Weight          Fuel_Capacity
                     0                      0                        1                      0
       Fuel_Efficiency Power_Performance_Factor            Brand_Model           Years_Launch
                     2                      1                        0                      0
                  Type
                     0
```

Then, we checked for values equal to 0 in the numerical variables, since in the context of the problem, it wouldn't make sense. But no value is equal to 0 ("Type" is a binary variable).

```
> zero_values
      Sales_in_Thousands             Resale_Value       Price_in_Thousands              Engine_Size
                       0                        0                        0                        0
              Horsepower                Wheelbase                    Width                   Length
                       0                        0                        0                        0
             Curb_Weight            Fuel_Capacity      Fuel_Efficiency Power_Performance_Factor
                       0                        0                        0                        0
            Years_Launch                     Type
                       0                       40
```

4. Univariate Analysis

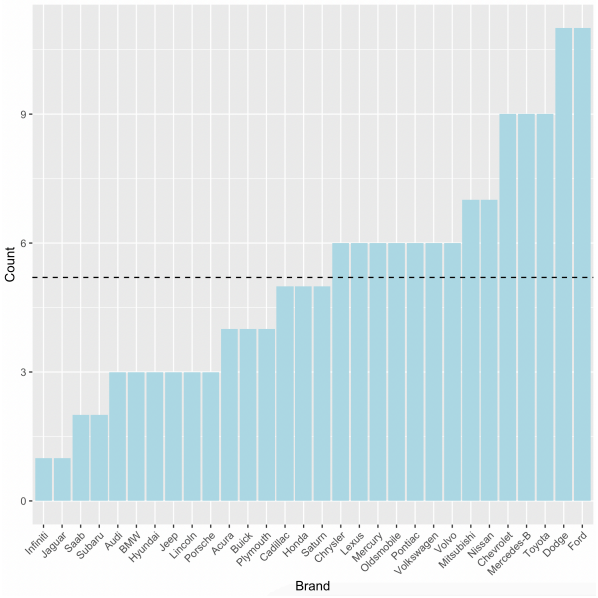
4.1. Categorical Variables

There's no need to visualize "Model" count, since every row represents a different model.

```
> num_model
[1] 155
```

The variable "Brand" has 30 different values. We visualized the count of each one, using a bar-plot in ascending order. With a horizontal line, representing the mean.

```
> num_man
[1] 30
```



"Ford", "Dodge", "Toyota", "Mercedes-Benz" and "Chevrolet", are the brands with the most models. While "Infiniti", "Jaguar", "Saab" and "Subaru" are the brands with the least models. The mean gives a reference point to compare the results. There are 14 brands in the top half, and the remaining 16 are in the bottom half.

To visualize the frequency of each brand, we created a table, ordered by descending order of frequency. We've also added a variable that represents the cumulative frequency, and one with the total values, seen previously.

	Brand	Frequency	Cumulative_Frequency	Total
1	Dodge	7.1	7.1	11
2	Ford	7.1	14.2	11
3	Chevrolet	5.8	20.0	9
4	Mercedes-B	5.8	25.8	9
5	Toyota	5.8	31.6	9
6	Mitsubishi	4.5	36.1	7
7	Nissan	4.5	40.6	7
8	Chrysler	3.8	44.4	6
9	Lexus	3.8	48.2	6
10	Mercury	3.8	52.0	6

Just 10 brands out of the 30, hold more than 50% of the different types of models.

4.2. Numerical Variables

4.2.1. Location

Using the function "summary(df)", we calculated the location measures of the numerical variables. At first glance each variable shows us that it seems to have a symmetrical distribution in terms of point location with the exception of the variables "Sales_in_Thousands", "Horsepower" and "Engine_Size". These variables contain a difference between max values and 3rd quartiles much larger than the difference between min and 1st quartiles. Looking at the differences between trimmed means (5%) and normal means of the variables, most of the trimmed values are lower than the means showing us that as expected the larger outliers are more significant than the lower ones. Most of the trimmed means, depending on how close to the regular means, indicate that the initial mean is not heavily affected by outliers. However, further analysis needs to be done since the relationship between trimmed means and regular means can vary depending on the distribution and characteristics of the data. Another way of calculating the robustness of the central tendency of each variable is comparing the trimmed mean with the median. Looking at the results almost all variables contained similar values with the exception of "Sales in Thousands" where there is a difference of about 10k this could mean that either the percentage trimmed did not encompass all the targeted outliers or the presence of a skewed distribution.

	In Thousands		
	Sales	Resale Value	Price
Min	0.11	5.16	9.24
1st Qu.	14.04	12.53	18.08
Median	29.21	17.62	23.10
Mean	52.99	18.06	27.39
3rd Qu.	68.07	18.08	31.94
Max	540.56	67.55	85.50
Trimmed Mean (5%)	44.16	16.82	25.91

	Engine Size	Wheelbase	Width	Length	Curb_Weight
Min	1.00	92.6	62.60	149.4	1.895
1st Qu.	2.30	103.0	68.40	177.6	2.973
Median	3.00	107.0	70.55	187.9	3.355
Mean	3.06	107.5	71.15	187.2	3.378
3rd Qu.	3.58	112.2	73.42	196.1	3.789
Max	8.00	138.7	79.90	224.5	5.572
Trimmed Mean (5%)	3.00	107.15	71.02	187.41	3.35

	Horsepower	Fuel_Capacity	Fuel_Efficiency	Power_Performance_Factor
Min	55.0	10.30	15.00	23.28
1st Qu.	149.5	15.80	21.00	60.57
Median	177.5	17.20	24.00	72.16
Mean	185.9	17.95	23.84	77.04
3rd Qu.	215.0	19.57	26.00	89.41
Max	450.0	32.00	45.00	188.14
Trimmed Mean (5%)	183.30	17.72	23.75	75.58

4.2.2. Dispersion

Dispersion measures were also calculated in order to further understand our dataframe. Looking at the IQR and range values, the data shows us that most variables contain the middle 50% of the data points close together with the full extent of the data being much larger, this is clearly shown in "Sales_in_Thousands" where the IQR is much smaller (54) than the range (540), and is then further indicated in the rest of the dispersion measures. The variance measures the average squared deviation of the data points, and the variables show a variety of results. Variables with a higher range like "Horsepower2 (3214.19) and "Sales_in_Thousands" (4657.86) are among the variables with highest variance showing us greater dispersion among the data points but also higher levels of uncertainty. Variables with lower ranges like

"Engine_Size" and "Curb_Weight" contain low variances. The standard deviation provides a measure of the spread of the data points around the mean, looking at the values in conjunction with the variance gives us a better understanding of the dispersion of the data points.

The coefficient of variation is calculated as the ratio of standard deviation to the mean in a variable, and is used to compare the variabilities between variables as long as they contain all points with the same sign. Since, from the location measure values we can see that every variable only contains positive points, the CV of all numerical variables was calculated. Comparing these values we can see that most of the variables contain a moderate CV in between 20-40%, meaning that most variables have typical levels of variability. On the other hand there are also variables with extremes such as "Sales_in_Thousands" containing 128.75% levels of variability and "Wheelbase" only containing 7.11% (indicating low variability and a stable variable).

	Sales	Resale Value	Price
IQR	54.03	5.55	13.86
Range	540.45	62.39	76.27
Variance	4657.86	101.54	204.64
Standard-deviation	68.25	10.08	14.31
Coefficient of variation	128.78	55.79	52.23

	Engine Size	Wheelbase	Width	Length	Curb_Weight
IQR	1.28	9.20	5.03	18.55	0.82
Range	7.00	46.10	17.3	75.10	3.67
Variance	1.09	58.39	11.92	180.41	0.39
Standard-deviation	1.04	7.64	3.45	13.43	0.63
Coefficient of variation	34.13	7.11	4.85	7.17	18.60

	Horsepower	Fuel_Capacity	Fuel_Efficiency	Power_Performance_Factor
IQR	65.5	3.76	5.00	28.84
Range	395.0	21.70	30.0	164.87
Variance	3214.93	15.12	18.10	628.08
Standard-deviation	56.70	3.89	4.25	25.06
Coefficient of variation	30.49	21.66	17.84	32.53

4.2.3. Distribution

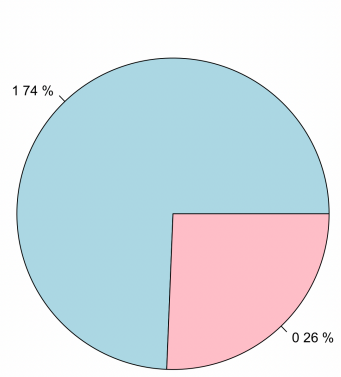
The kurtosis of each numerical variable was also calculated using the R library e1071. It measures the peakedness of the distribution, giving us a better representation of outlier occurrences. In the library e1071 a kurtosis's distribution value of 3 is equivalent to an excess kurtosis of 0. Analyzing each value, the variable with the "tails" closest to one with a normal distribution is "Fuel_Efficiency" with 3.098, while the variable furthest from a normal distribution is "Sales _in_Thousands" with 16.59. Comparing all values, most of the variables have lighter tails, with only "Resale_Value", "Price_in_Thousands" and the previously mentioned "Sales_in_Thousands" having heavier tails. This means that most of the variables contain a more stable distribution with fewer outliers than a normal distribution would have.

	Sales	Resale Value	Price
Kurtosis	16.59	7.80	3.43

	Engine Size	Wheelbase	Width	Length	Curb_Weight
Kurtosis	2.17	2.68	-0.36	0.21	1.16

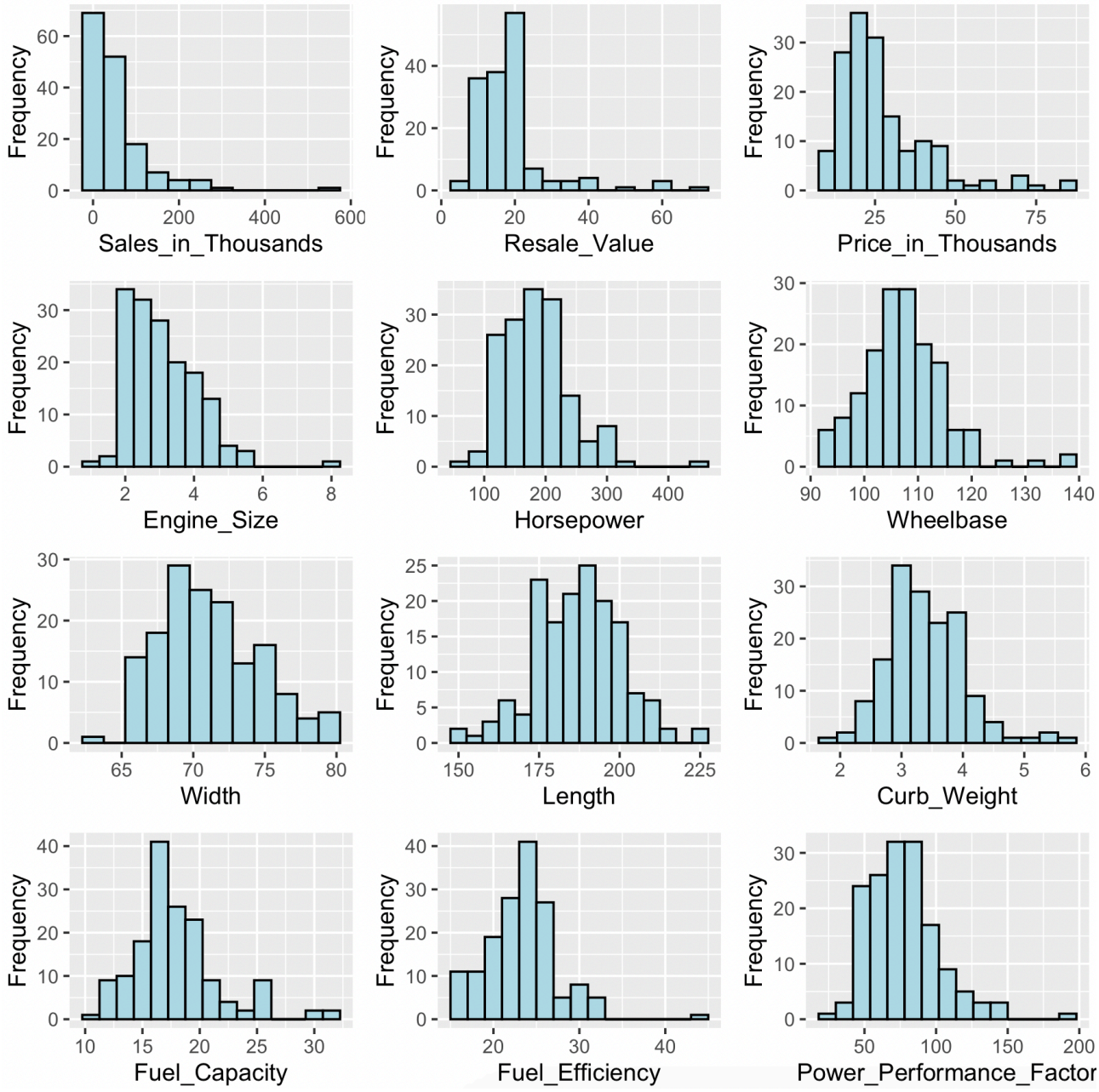
	Horsepower	Fuel_Capacity	Fuel_Efficiency	Power_Performance_Factor
Kurtosis	2.22	1.91	3.098	1.94

We used a pie chart, in order to visualize the frequency of each category in the binary variable "Type". Where 1 represents "Passenger", and 0 represents "Car".

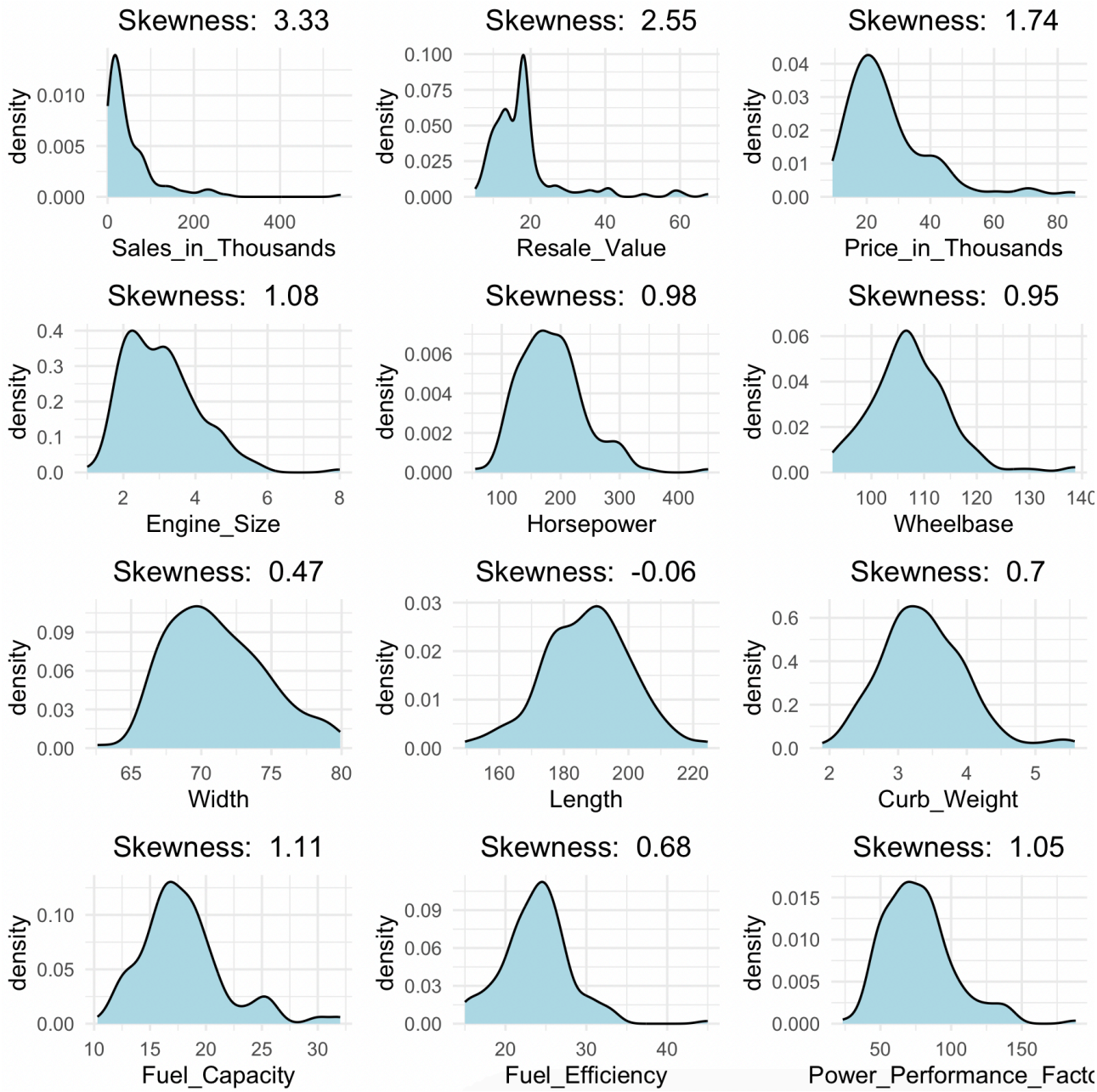


Vehicles of the type "Passenger" represent 3/4 of the observations. While vehicles of the type "Car" represent 1/4.

To represent the distribution of the frequency for each numerical variable (but the "Years_Launch"), we used histograms. Displaying every histogram in a grid.



Analyzing each histogram, it's clear that most present high skewness and some outliers. So we calculated the skewness level, for every numerical variable, and plotted the density distribution. Displaying every plot in a grid.



We know that, when variables have negative skewness, typically their mean is smaller than the median and mode. When the variables have positive skewness, typically their mean is larger than the median an mode. And when the variables have an approximately symmetrical skewness, their mean, median and mode are close to equal.

"Sales_in_Thousands", show high positive skewness, meaning that most models represent a lower amount of sales.

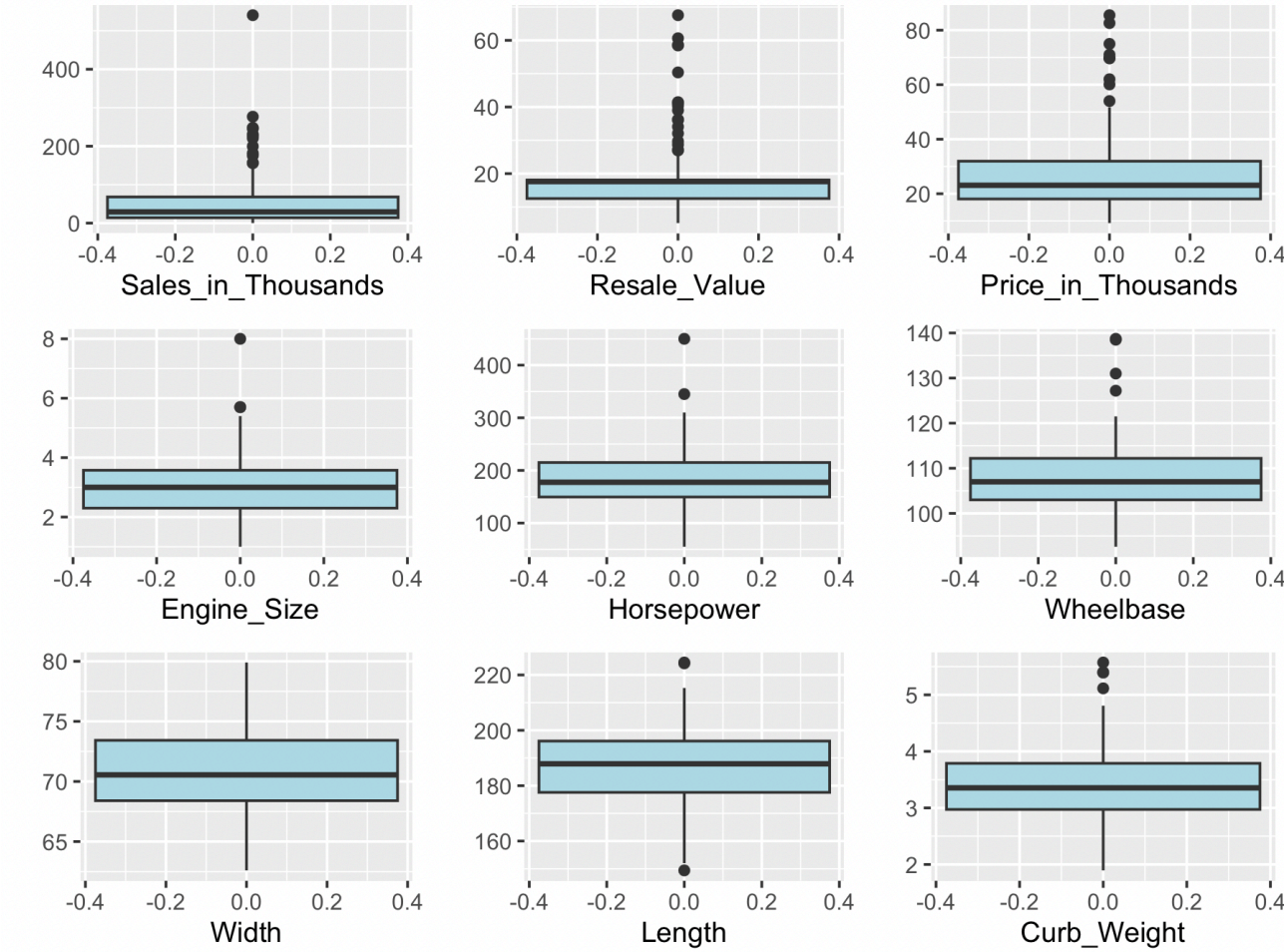
"Resale_Value" and "Price_in_Thousands", also show high positive skewness, meaning that most models represent a lower price.

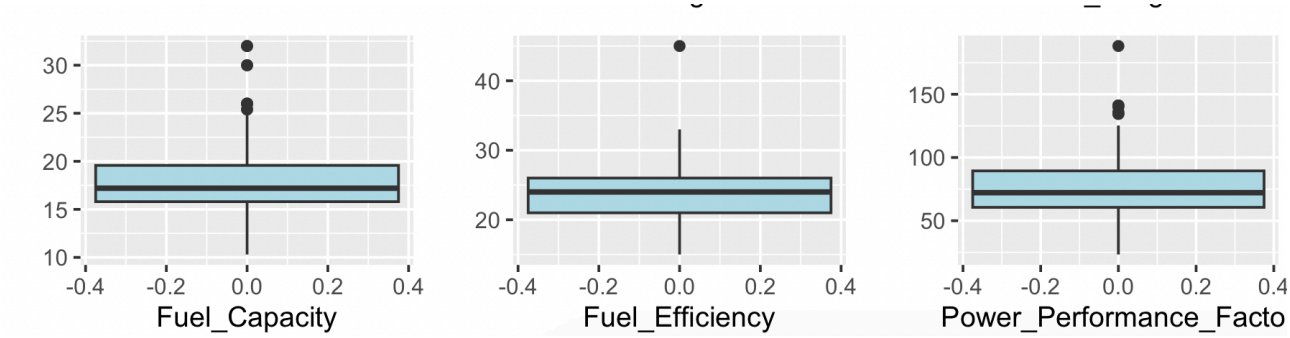
"Engine_Size", "Horsepower", "Wheelbase", "Fuel_Capacity" and "Power_Performance_Factor", also show high positive skewness, meaning that most models have a smaller engine, horsepower, wheelbase, fuel capacity and power performance factor.

"Width", "Length", "Curb_Weight" and "Fuel_Efficiency", show a more balanced symmetrical distribution, meaning that most models represent a more balanced distribution of the values.

4.2.4. Outliers

Afterwards, we plotted the box-plots for each numerical variable, in order to analyze the position, dispersion and outliers. Every outlier presented is order from biggest to smallest.





"Sales_in_Thousands", has several positive outliers. These outliers represent models with a lot more sales than the rest, and they are: "Ford F-Series" (by far the biggest outlier), "Ford Explorer", "Toyota Camry", "Ford Taurus", "Honda Accord", "Dodge Ram Pickup", "Ford Ranger", "Honda Civic", "Dodge Caravan", "Ford Focus", "Jeep Grand Cherokee" and "Ford Windstar". "Ford" is by far the brand with the most outliers in this variable.

"Resale_Value", also has several outliers. These outliers represent models much more expensive when resold than the rest, and they are: "Porsche Carrera Cabrio", "Porsche Carrera Coupe", "Mercedes-B SL-Class", "Dodge Viper" (the most severe outliers), "Mercedes-B S-Class", "Mercedes-B E-Class", "Porsche Boxter", "Lexus LS400", "Audi A8", "Chevrolet Corvette", "BMW 528i", "Toyota Land Cruiser", "Lexus GS300", "Acura RL", "BMW 328i", "Toyota Land Cruiser", "Lexus GS300", "Acura RL", "BMW 328i", "Cadillac Seville" and "Lexus ES300".

"Price_in_Thousands", also has several outliers. These outliers represent models much more expensive than the rest, and they are: "Mercedes-B CL500", "Mercedes-B SL-Class", "Porsche Carrera Cabrio", "Porsche Carrera Coupe", "Dodge Viper", "Mercedes-B S-Class", "Audi A8", "Lexus LX470" and "Lexus LS400". "Mercedes", "Porsche" and "Lexus" have at least 2 outliers each.

"Engine_Size", only has 3 outliers. These outliers represent models with engines much bigger than the rest, and they are: "Dodge Viper" (severe outlier), "Cadillac Escalade" and "Chevrolet Corvette" (moderate outliers, they have the same size).

“Horsepower”, only has 2 outliers. These outliers represent models with much bigger horsepower than the rest, and they are: “Dodge Viper” (severe outlier) and “Chevrolet Corvette” (moderate outlier). These 2 models also are outliers in terms of engine size.

“Wheelbase”, only has 4 outliers. These outliers represent models with much bigger wheelbase than the rest, and they are: “Dodge Ram Pickup”, “Ford F-Series”, “Dodge Dakota” and “Dodge Ram Van” (all moderate outliers). “Dodge” contains almost all outliers for this variables.

“Width” has no outliers.

“Length”, only has 3 outliers. From this outliers, 2 represent models with much bigger length than the rest, and they are: “Ford F-Series” and “Dodge Ram Pickup” (both are moderate outliers). This models also are positive outliers in the “Wheelbase” variable. And the other one represents a model, “Chevrolet Metro”, with a much smaller length than the rest.

“Curb_Weight”, only has 4 outliers. These outliers represent models much heavier than the rest, and they are: “Cadillac Escalade”, “Lexus LX470”, “Lincoln Navigator” and “Toyota Land Cruiser” (all moderate outliers).

“Fuel_Capacity”, has several outliers. These outliers represent models with a much bigger fuel capacity than the rest, and they are: “Dodge Ram Wagon”, “Dodge Ram Van”, “Cadillac Escalade”, “Lincoln Navigator”, “Dodge Ram Pickup”, “Ford Windstar” and “Ford Expedition” (all are moderate outliers).

“Fuel_Efficiency” only has 1 outlier, “Chevrolet Metro” (severe and positive).

“Power_Performance_Factor”, has several outliers. These outliers represent models with a much bigger factor than the rest, and they are: “Dodge Viper”, “Chevrolet Corvette”, “Mercedes-B CL500”, “Mercedes-B SL-Class” (severe outliers), “Porsche Carrera Cabrio”, “Audi A8” and “Porsche Carrera Coupe” (moderate outliers).

There are 38 different models that represent outliers.

```
> all_outliers
 [1] "Ford F-Series"          "Ford Explorer"          "Toyota Camry"           "Ford Taurus"
 [5] "Honda Accord"           "Dodge Ram Pickup"       "Ford Ranger"            "Honda Civic"
 [9] "Dodge Caravan"          "Ford Focus"             "Jeep Grand Cherokee"    "Ford Windstar"
[13] "Porsche Carrera Cabrio" "Porsche Carrera Coupe"  "Mercedes-B SL-Class"    "Dodge Viper"
[17] "Mercedes-B S-Class"     "Mercedes-B E-Class"     "Porsche Boxter"         "Lexus LS400"
[21] "Audi A8"                "Chevrolet Corvette"     "BMW 528i"               "Toyota Land Cruiser"
[25] "Lexus GS300"            "Acura RL"               "BMW 328i"               "Cadillac Seville"
[29] "Lexus ES300"            "Mercedes-B CL500"       "Lexus LX470"            "Cadillac Escalade"
[33] "Dodge Dakota"           "Dodge Ram Van"          "Chevrolet Metro"        "Lincoln Navigator"
[37] "Dodge Ram Wagon"        "Ford Expedition"
```

5. Bivariate Analysis

5.1. Categorical x Numerical Variables

5.1.1. "Type"

For the bivariate analysis, we started by plotting a pie-chart for each numerical variable. This pie-charts, display the mean value of those variables for "Passenger" (blue) and "Car" (pink), from the categorical variable "Type".

"Mean Sales" shows that car vehicles sell a lot better than passenger vehicles.

“Mean Resale Value”, “Mean Price” are vey well balanced, but skew towards car vehicles a little bit. Showing that car vehicle have prices a bit bigger than passenger vehicles.

“Mean Engine Size”, “Mean Fuel Capacity” and “Mean Weight”, skew towards passenger vehicles. Showing that passenger vehicles, have bigger engines, fuel capacity and are more heavy.

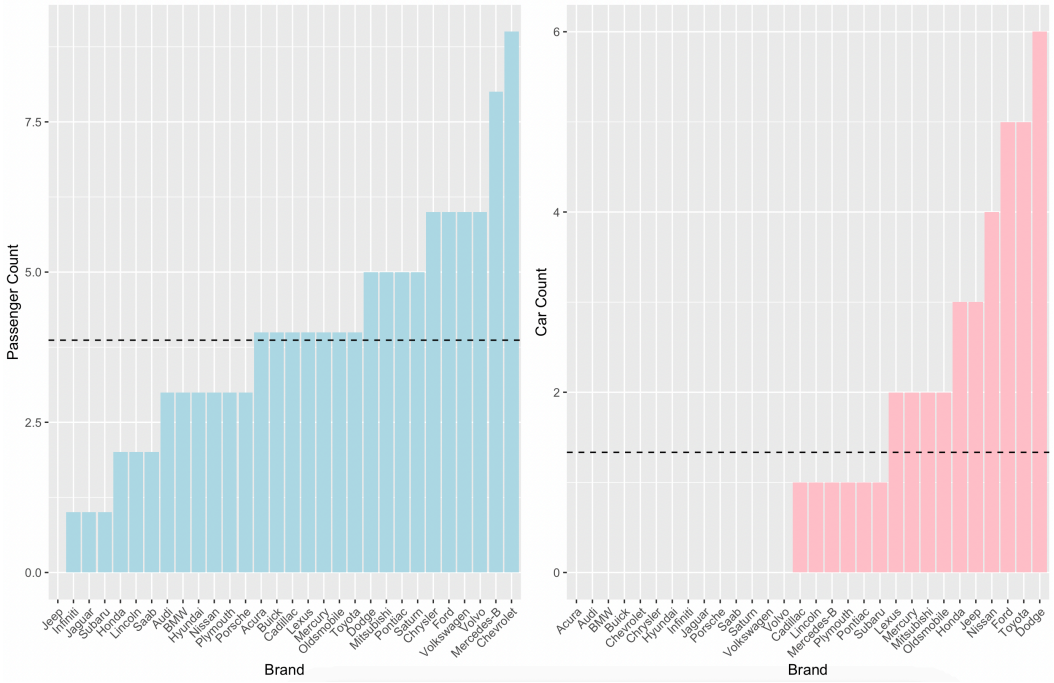
“Mean Horsepower” , “Mean Wheelbase”, “Mean Width”, “Mean Length”, “Mean Factor” are almost perfectly balanced. Showing that car vehicles and passenger vehicles, have very balanced power, power performance factor, wheelbase size, width and length.

“Mean Fuel Efficieny”, skews toward car vehicles. Showing that car vehicles, can travel further without having to stop for fuel.

In order to further explore the relationship between the categorical variable "Type" and every numerical variable of our dataset, we conducted a series of statistical tests known as t-tests, in order to find out if there’s a significant difference in means between two groups. For each t-test conducted, we obtained a p-value, which indicates the probability of observing the obtained difference in means (or more extreme) if the null hypothesis were true. After conducting the t-tests for each numerical variable, we examined the results to identify any statistically significant differences between "Passenger" and "Car" types. Variables with p-values less than 0.05 were considered to have significant differences between the two groups. And this variables were: “Sales_in_Thousands”, “Engine_Size”, “Wheelbase”, “Width”, “Curb_Weight”, “Fuel_Capacity” and “Fuel_Efficiency”.

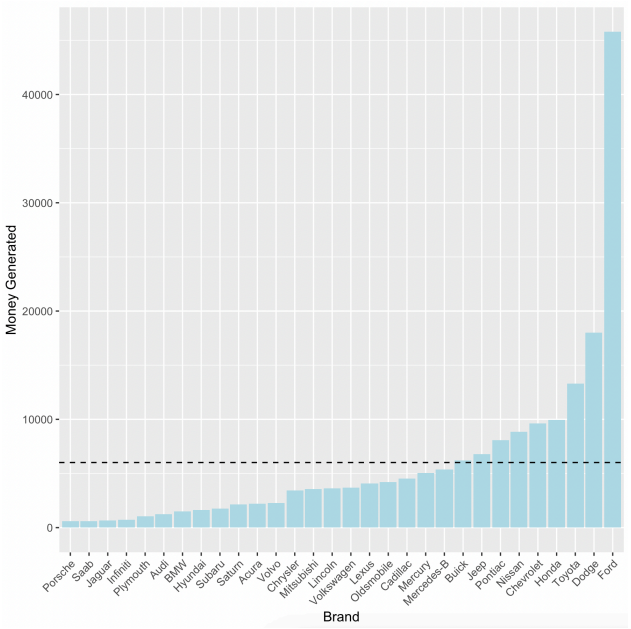
5.1.2. “Brand” Bar-Plots

Afterwards, we visualized the number of each type of vehicle, per brand, using bar-plots, with the results in ascending order. We also plotted an horizontal line with the mean, for each.



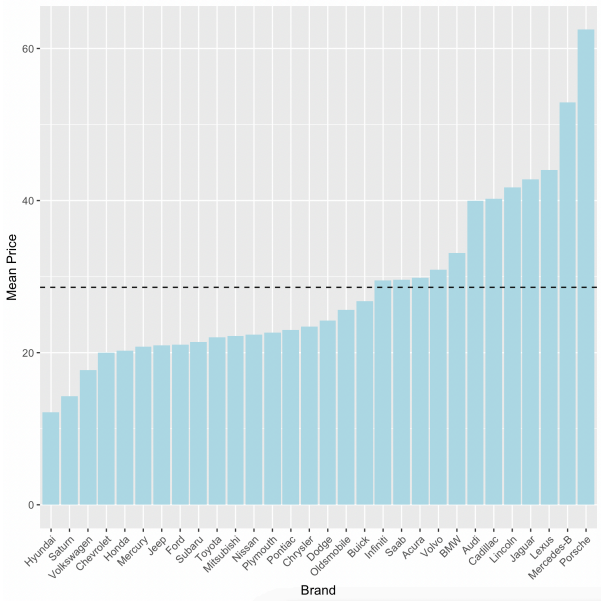
“Chevrolet” and “Mercedes-Benz”, are the brands with more passenger vehicles, while “Jeep”, doesn’t have a single passenger vehicle. “Dodge”, “Toyota” and “Ford”, are the brands with more car vehicles, while “Acura”, “Audi”, “BMW”, “Buick”, “Chevrolet”, “Chrysler”, “Hyundai”, “Infiniti", “Jaguar”, “Porsche", “Saab”, “Saturn”, “Volkswagen” and “Volvo”, don’t have a single car vehicle.

Then we checked which brands generated more money. To do that we summed for each, the price times the sales, for every model. And displayed the results using a bar-plot, with the results in ascending order. We also plotted an horizontal line with the mean.



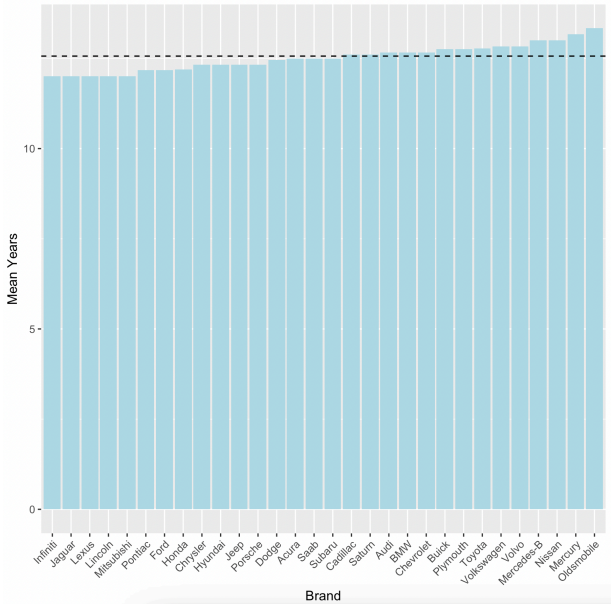
"Ford" is by far the most successful brand. "Dodge" and "Toyota" complete the top 3. They're also the 3 brands with more models. While "Porsche", "Saab" and "Jaguar" are the brands that generated the least amount of money.

We also plotted the mean price, for each brand, in ascending order and plotted an horizontal line with the mean.



The most successful brands, in terms of money generated, are in the bottom half of prices. While the least successful brands, in terms of money generated, are in the top half of prices. "Porsche" is both, the least successful brand and the most expensive.

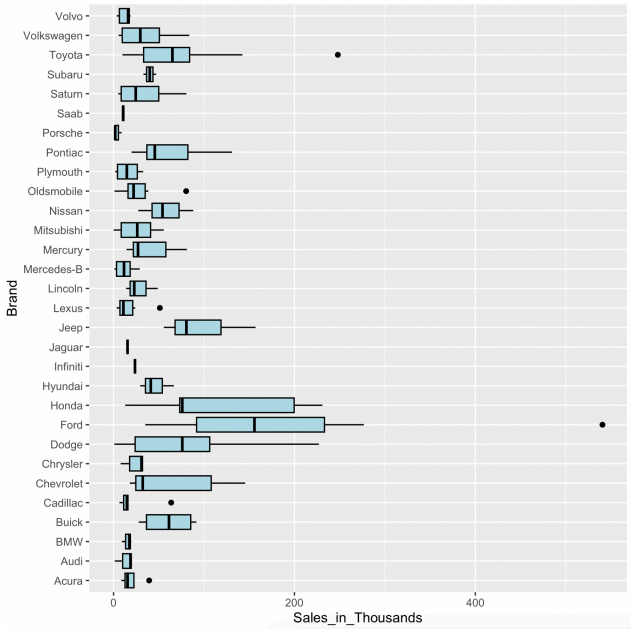
We also plotted the average “Years_Launch” per brand, in ascending order and with a horizontal line with the mean.



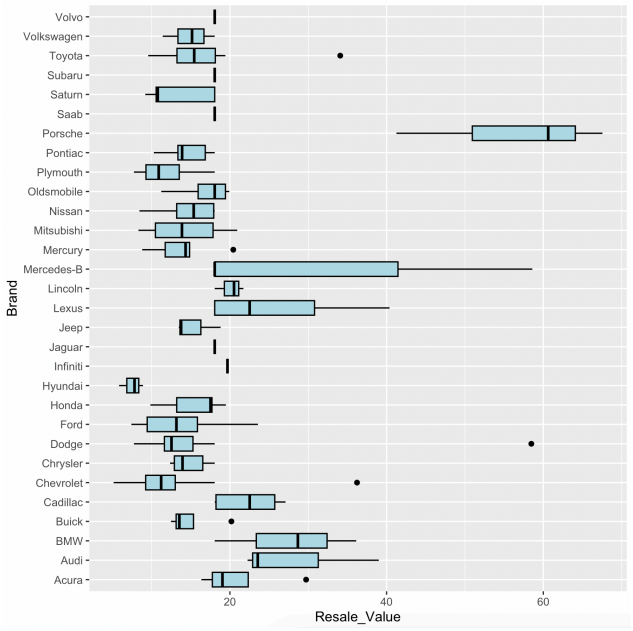
“Oldsmobile”, “Mercury”, “Nissan” and “Mercedes-Benz” are the brands with on average older cars. “Infiniti”, “Jaguar”, “Lexus”, “Lincoln" and “Mitsubishi” are the brands with on average younger cars.

5.1.3. “Brand" Box-Plots

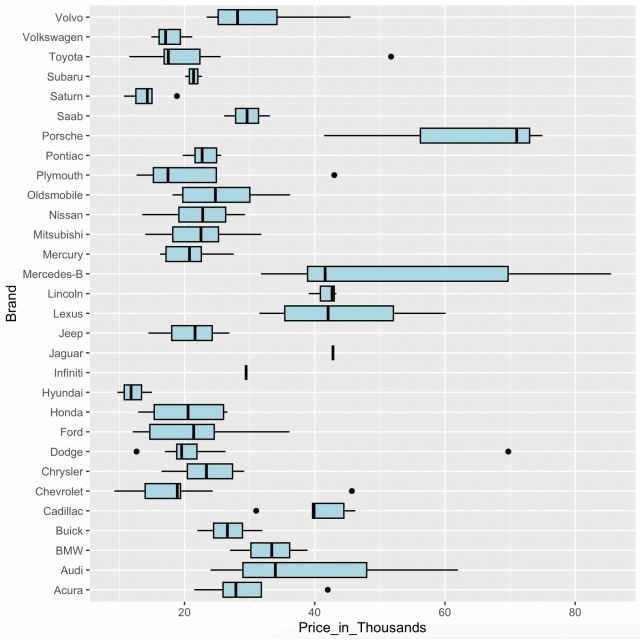
We decided to use box-plots, to visualize the position and dispersion of every numerical variable in relation to each brand.



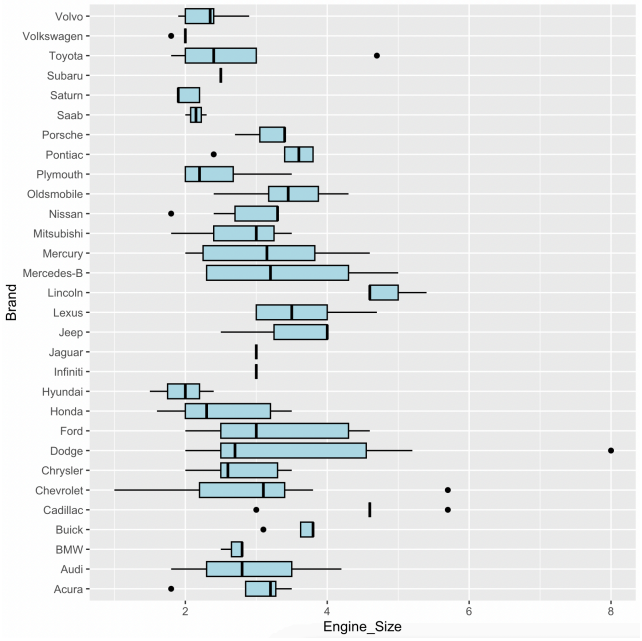
In the “Sales_in_Thousands” box-plot, we can see that “Toyota”, “Oldsmobile”, “Lexus”, “Cadillac”, “Acura” and mainly “Ford”, have 1 outlier each, representing a much more expensive model. Every model has a pretty balanced and high amount of sales, except “Subaru” which has low number sales, this is mainly due to the fact that it has a low amount of models. “Jeep" and “Ford” also have a substantial amount of sales per model.



In the “Resale_Value” box-plot, we can see that “Toyota”, “Mercury”, “Chevrolet”, “Buick”, “Acura” and mainly “Dodge”, have 1 outlier each, representing a much higher resale value. “Porsche” although is the least bought brand, it shows the better resale value, separating itself from the rest of the group. “Mercedes-Benz” comes as second, followed by “Lexus”, “Lincoln”, “Jaguar”, “Infiniti”, “Cadillac”, “BMW”, “Audi” and “Acura”. Not surprisingly, the brands with higher resale value are also the most expensive brands.

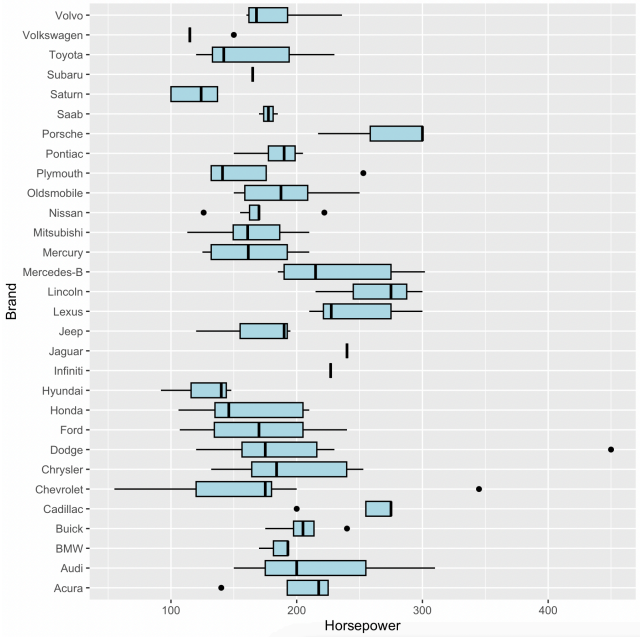


In the “Price_in_Thousands” box-plot, we can see that “Toyota", “Plymouth”, “Chevrolet”, “Acura” and mainly “Dodge”, have 1 outlier each, representing a much more expensive model. “Dodge” and “Cadillac” also have 1 outlier each representing a much less expensive model. Although “Porsche” is on average the most expensive brand, “Mercedes-Benz” has the most expensive model.

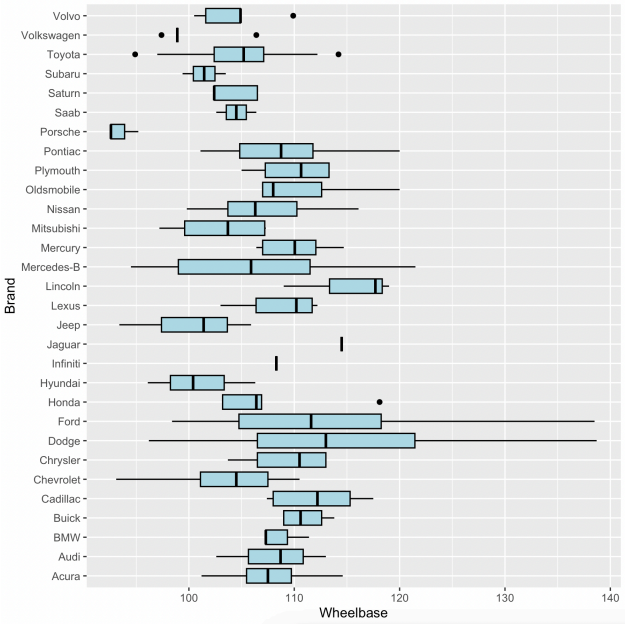


In the “Engine_Size” box-plot, we can see that “Toyota”, “Chevrolet”, “Cadillac” and mainly “Dodge", have 1 outlier each, representing a model with a much bigger engine. “Volkswagen”, “Pontiac”, “Nissan”, “Cadillac", “Buick” and “Acura", also have 1 outlier each but represent a model with a much smaller engine. “Lincoln” and

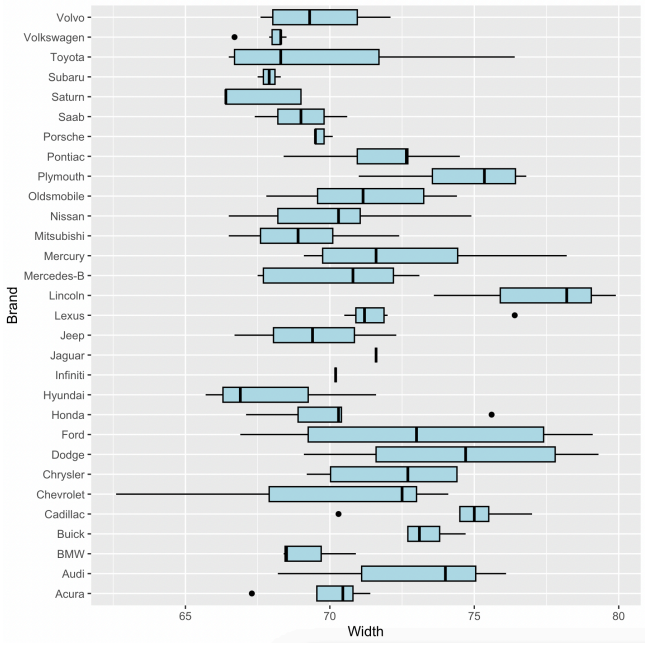
“Cadillac” seem to be the brands with the bigger engines. “Hyundai” and “Chevrolet” seem to be the brands with the smaller engines.



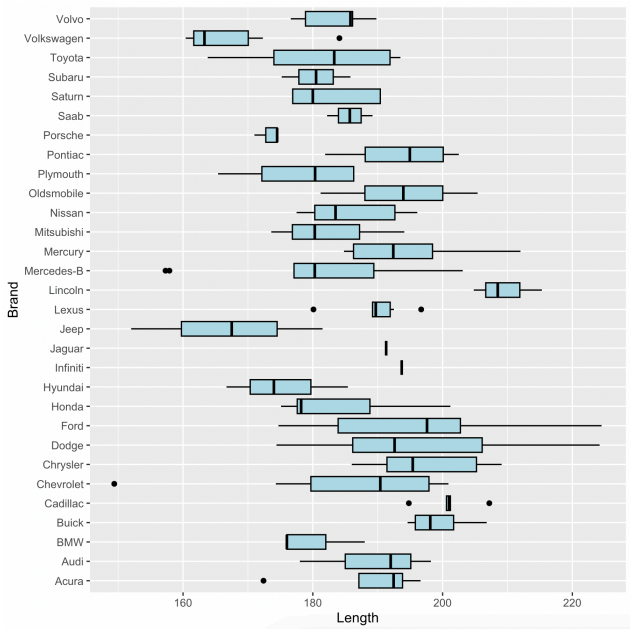
In the “Horsepower” box-plot, we can see that “Volkswagen”, “Plymouth”, “Nissan”, “Chevrolet”, “Buick” and mainly “Dodge", have 1 outlier each, representing a model with much higher horsepower than the rest. “Nissan” and “Cadillac”, have 1 outlier each representing a model with much lower horsepower than the rest. “Saab”, “Mercedes-Benz”, “Lincoln”, “Lexus", “Jaguar”, “Infiniti” and “Cadillac” are the brands with more horsepower. “Volkswagen”, “Saturn”, “Hyundai” and “Chevrolet” are the brands with the least horsepower.



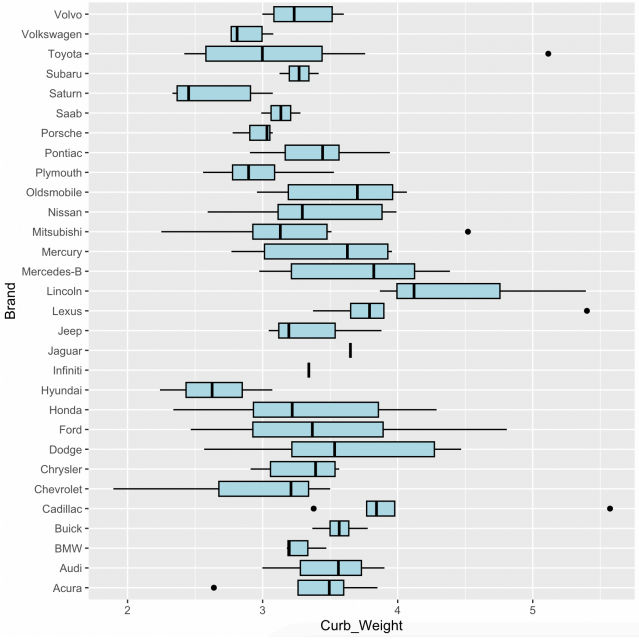
In the “Wheelbase” box-plot, we can see that “Volvo", “Volkswagen”, “Toyota” and mainly “Honda”, have 1 outlier each representing a model with a much bigger wheelbase. “Volkswagen” and “Toyota”, have 1 outlier, representing a model with a much smaller wheelbase. “Porsche” is by far the brand with on average the smaller wheelbase. "Lincoln” seems to be the one with the bigger wheelbases, although “Ford” and “Dodge” have larger spectrums of wheelbases size.



In the “Width” box-plot, we can see that “Lexus” and “Honda”, have 1 outlier, representing a model with a much bigger width. “Volkswagen”, “Cadillac” and “Acura", have 1 outlier each, representing a model with a much smaller width. As expected, since “Lincoln" is the brand with the bigger wheelbases, is also the brand with bigger width. But surprisingly, “Porsche” is not the smallest, but “Saturn” and “Hyundai”. “Chevrolet” has a large spectrum of widths.

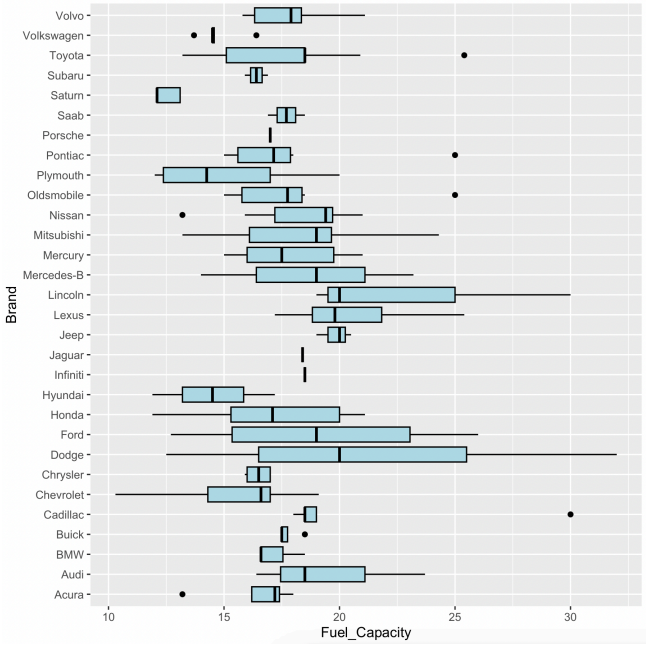


In the “Length” box-plot, we can see that “Volkswagen”, “Lexus" and “Cadillac”, have 1 outlier each, representing a model with a much bigger length. “Mercedes-Benz”, “Lexus”, “Cadillac”, “Acura” and mainly “Chevrolet", have outliers representing models with a much smaller length. “Lincoln”, not only has bigger wheelbases and width, but is also the lengthiest brand. “Jeep" and “Volkswagen” are the smallest. “Ford” and “Dodge" have a large spectrum of lengths.

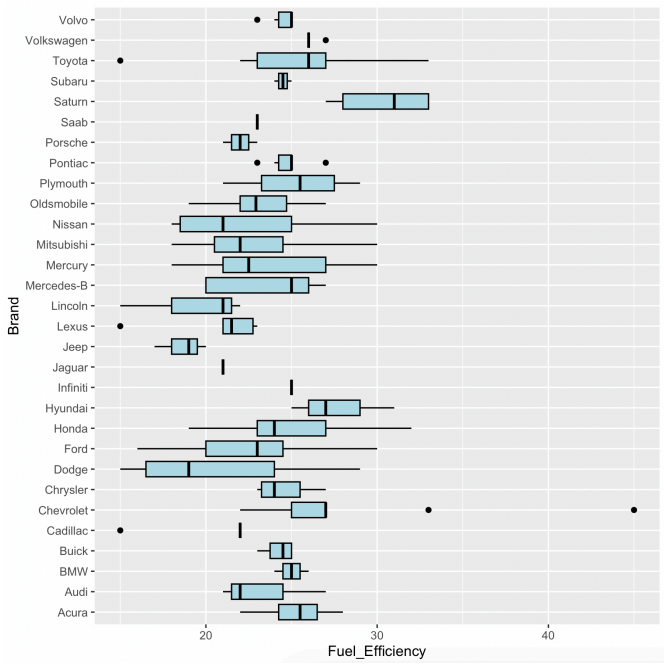


In the “Curb_Weight” box-plot, we can see that “Mitsubishi”, “Lexus”, “Cadillac” and mainly “Toyota” have 1 outlier, representing a much heavier model. “Cadillac” and “Acura" also have 1 outlier, representing a model, with a much lighter model.

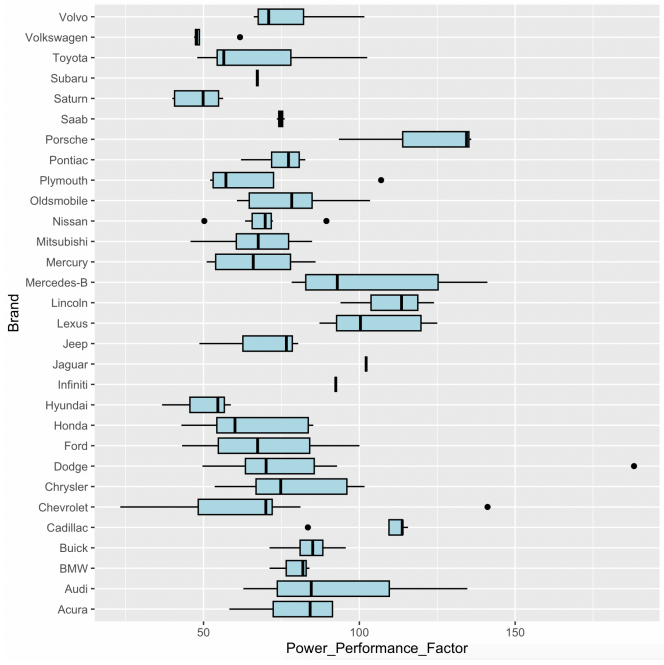
"Lincoln", predictably, since it's the biggest brand in terms of wheelbase, width and length, is also the heaviest. "Saturn" seems to be the lightest, although "Chevrolet", has a range of a few models that are lighter.



In the "Fuel_Capacity" box-plot, "Volkswagen", "Toyota", "Pontiac", "Oldsmobile", "Buick" and mainly "Cadillac" have 1 outlier, representing a model with a much bigger fuel capacity. "Volkswagen", "Nissan" and "Acura" also have 1 outlier, representing a model with a much smaller fuel capacity. "Saturn" is the brand that has on average models with lower fuel capacity. "Ford" and "Dodge" have a large spectrum of values. "Dodge" has a range of models, with the highest fuel capacity. Followed by "Lincoln", which makes sense, since "Lincoln" vehicles are the biggest so they need more fuel.



In the “Fuel_Efficiency” box-plot, “Volkswagen”, “Toyota" and “Pontiac” have 1 outlier, representing a model with much bigger fuel efficiency. “Chevrolet” has 2, one of them being much higher than the others. “Pontiac", “Toyota”, “Lexus" and “Cadillac” have 1 outlier, representing a model with a much smaller fuel efficiency. The outlier from “Toyota”, “Lexus” and “Cadillac” represent the minimum value found. The brand with the biggest fuel efficiency on average is “Saturn”. “Toyota” also has a range of vehicles, with a big fuel efficiency.

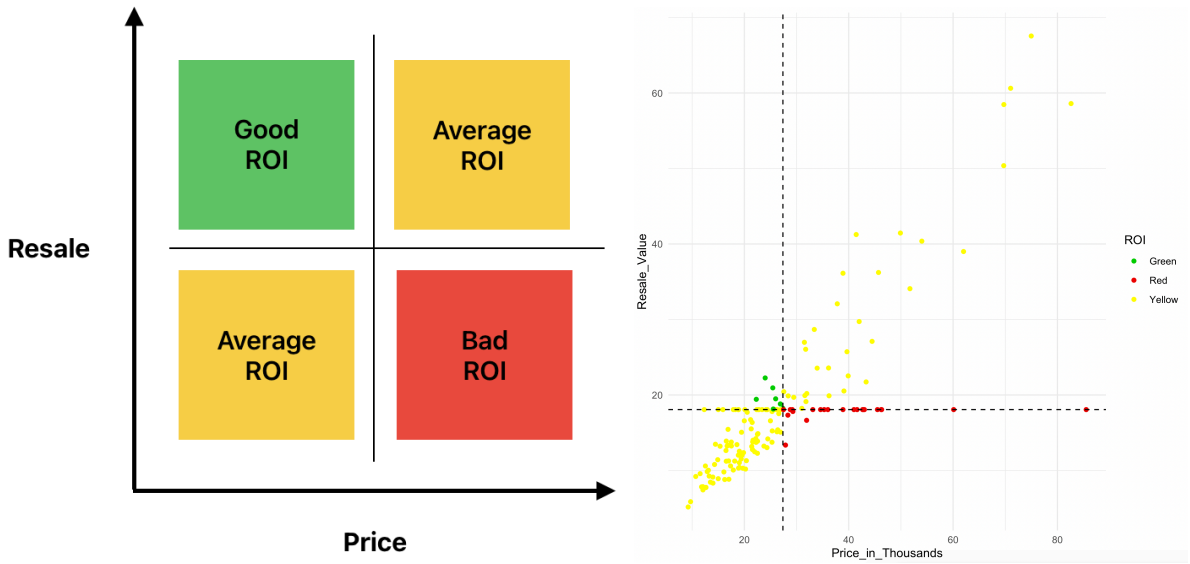


In the “Power_Performance_Factor” box-plot, “Volkswagen”, “Plymouth”, “Nissan”, “Chevrolet” and mainly “Dodge" have 1 outlier each, representing a model with a much higher factor. “Dodge” has the model with, by far, the highest factor. “Nissan” and “Cadillac", have 1 outlier each, representing a model with a much smaller factor. “Porsche” seems to be the brand with the highest factor, although “Audi” also has a range of a few models with a high factor.

5.2. Numerical x Numerical Variables

5.2.1. Return on Investment

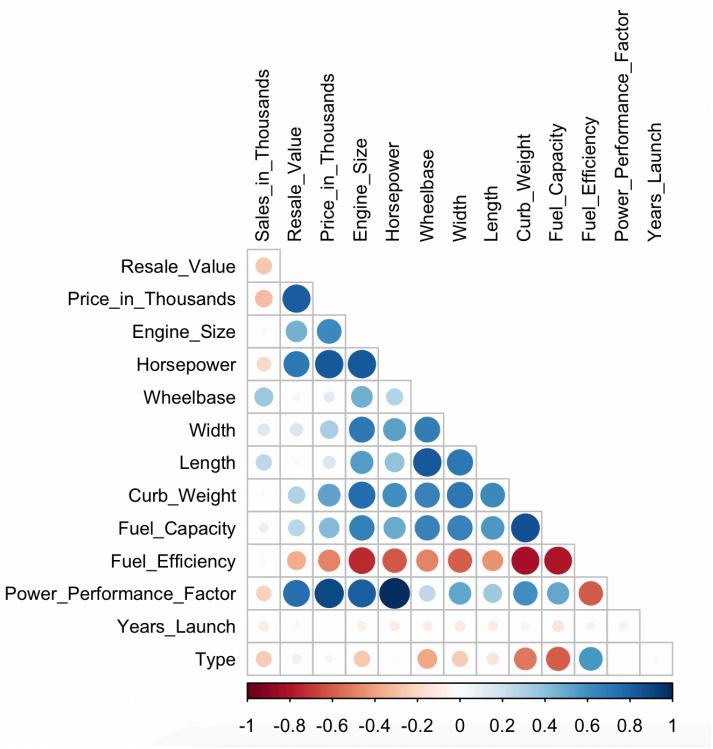
In the bivariate analysis with only numerical variables, we started by plotting the relation between “Resale_Value” and “Price_in_Thousands”. In order to find out, what models are the smartest buys in terms of ROI (Return on Investment). To do that we used a scatterplot, divided by half both in the x-axis and y-axis, forming 4 different groups, that can be either: green (good ROI), yellow (average ROI) and red (bad ROI).



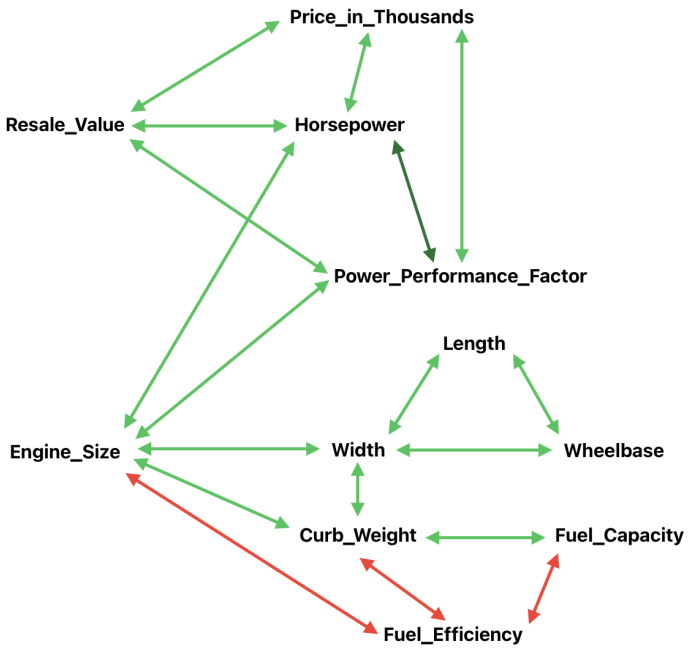
Most of the models represent an average ROI. We have very few that represent a good ROI, about 7 out of the 155, and they are: “Acura CL”, “Audi A4”, “Honda Odyssey”, “Jeep Grand Cherokee”, “Mitsubishi 3000GT”, “Toyota Avalon” and “Toyota 4Runner”.

5.2.2. Correlation

Afterwards we calculated and plotted the correlation matrix between every numerical variable.



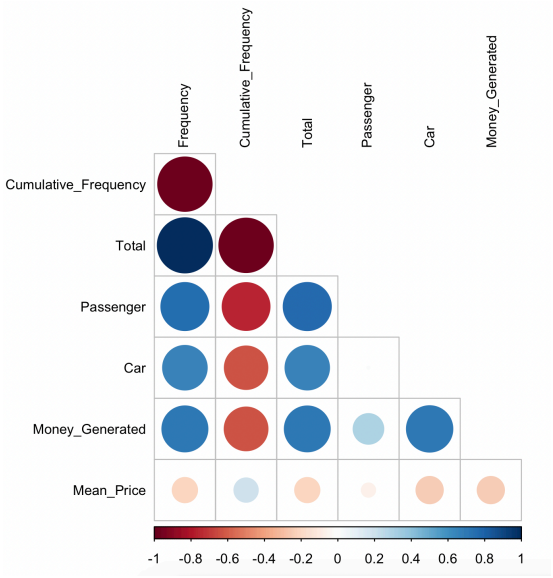
We manually built a graph displaying every correlation, positive or negative, bigger than 0.7, in order to simplify the interpretation. The green arrows represent the positive correlations, the red arrows represent the negative correlations. The darker green arrow represents a correlation very close to 1.



Looking at the graph, we can see that the main attributes that correlate with the “Resale_Price” are the initial “Price_in_Thousands”, the “Horsepower” and the “Power_Performance_Factor”. This factor shows an almost perfect correlation with “Horsepower", making this variable redundant and unnecessary. The main attributes that define the price of the model is the “Engine_Size” and the “Horsepower”, 2 variables that are also, predictably, correlated between them. “Engine_Size” is also correlated with the “Width” and “Weight” of the car, 2 variables that are also correlated between each other. The variables that represent sizes, “Length”, “Width” and “Wheelbase”, are all correlated between them. The “Curb_Weight” is predictably correlated with the “Fuel_Capacity”, since the “Curb_Weight” includes the weight of the tank filled. “Fuel_Efficiency” is negatively correlated with the “Engine_Size, “Curb_Weight" and “Fuel_Capacity”.

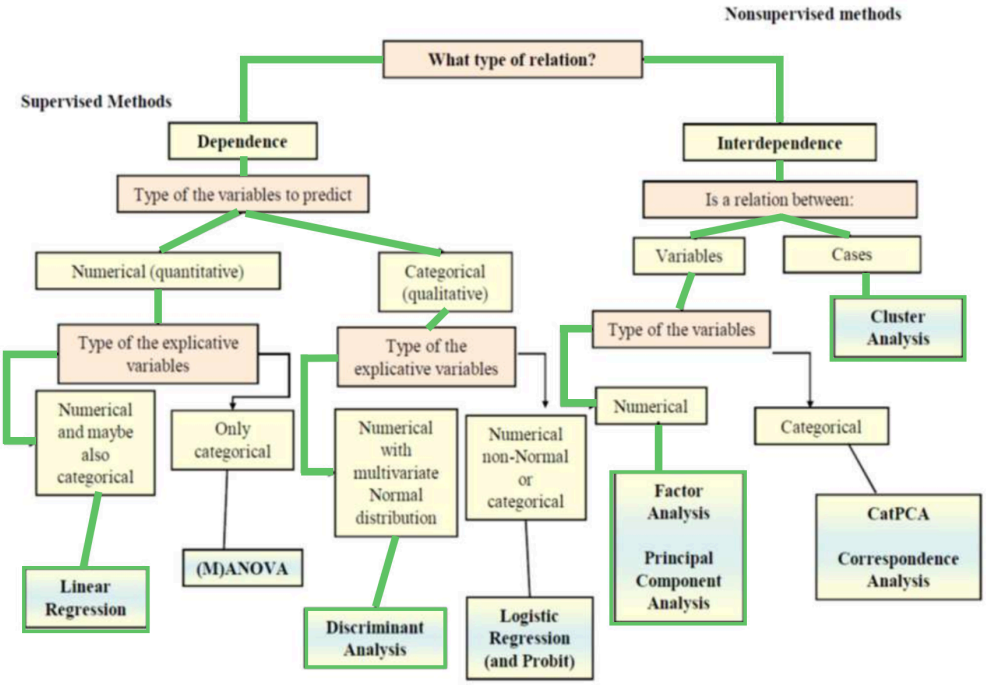
We also calculated and plotted the correlation matrix, of the dataframe, for each brand, that has as variables: the frequency and cumulative frequency (in ascending order), the total number of models, the total number of passenger vehicles, the total number of car vehicles and the money generated.

	Brand	Frequency	Cumulative_Frequency	Total	Passenger	Car	Money_Generated	Mean_Price
1	Acura	2.6	79.4	4	4	0	2229.5532	29.82269
2	Audi	1.9	86.5	3	3	0	1212.4650	39.98000
3	BMW	1.9	88.4	3	3	0	1523.0872	33.09667
4	Buick	2.6	82.0	4	4	0	6219.4866	26.78125
5	Cadillac	3.2	70.4	5	4	1	4541.0673	40.25400
6	Chevrolet	5.8	20.0	9	9	0	9610.6533	20.02278
7	Chrysler	3.8	44.4	6	6	0	3461.3130	23.43083
8	Dodge	7.1	7.1	11	5	6	18000.6084	24.21364
9	Ford	7.1	14.2	11	6	5	45776.1289	21.04727
10	Honda	3.2	73.6	5	2	3	9940.3056	20.27700



The main interpretation that we can take out of this matrix, is that there's a much higher correlation between the "Money_Generated" with "Car", than with "Passenger". Meaning that a brand generates more money when they have more car vehicles, than passenger vehicles.

6. Multivariate Analysis



6.1. Principal Component Analysis

6.1.1. Applying PCA

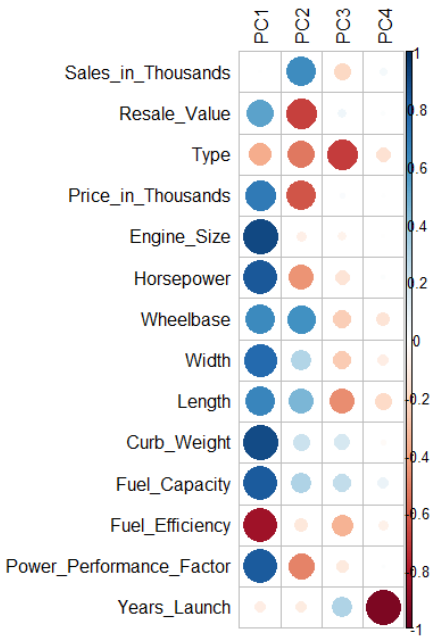
The main objective in conducting Principal Component Analysis (PCA) is to reduce the dimensions of the data by creating an approximation that represents a small subset of variables while removing the least amount of information possible. After the correlation in between variables showed us that PCA is possible we standardized the data so that we can conduct a normed PCA. To understand how many components we should keep we calculated the cumulative variance of each component analyzed.

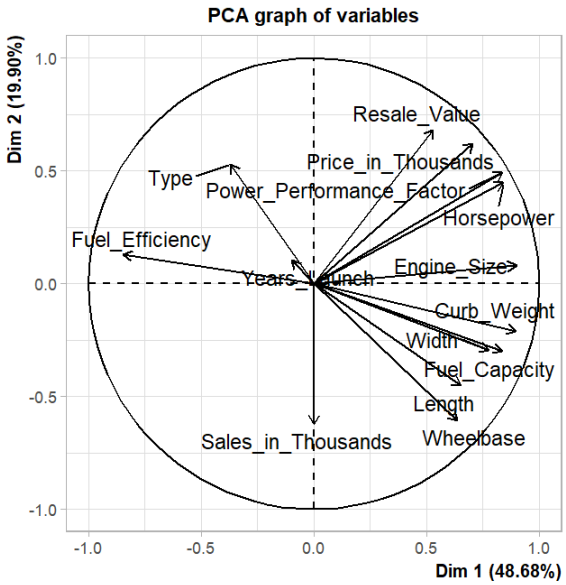
	Inertia_Explained	Cumulative_Inertia
1	0.487	0.487
2	0.199	0.686
3	0.086	0.772
4	0.070	0.842

Pearson's Criteria says that we should keep the PCs that explain at least 80% of the total. Looking at the table above we can see that the dispersion can be done in the first 4 PCs.

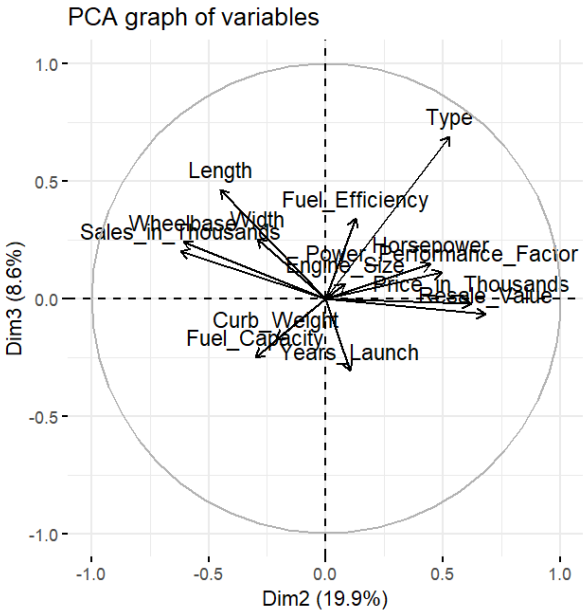
6.1.2. Variables Interpretation

A correlation plot between variables and PC was then made. Looking at the plot, PC1 shows a correlation with almost every variable with the exception of "Type", "Years_Launch" and "Sales_in_Thousands", "Resale_Value" has low correlation. PC2 shows a positive high correlation with "Sales_in_Thousands" and "Wheelbase", and a negative one with "Resale_Value" and "Price_in_Thousands". PC3 shows a negative high negative correlation with "Type" and PC4 with "Years_Launch". At first glance, PC1 represents the general measure of the vehicle characteristics, PC2 seems to capture a contrast between sales and vehicle size versus value and cost making it represent a trade-off between popularity and size. PC3 differentiates vehicle types while PC4 represents differences in the launch year.





Looking at the graphical representation for the first 2 PCs we can see the representation of the correlation above, but a closer look shows us three distinct groups that have positive correlations between them. Group 1 consisting of "Wheelbase", "Length", "Fuel_Capacity", "Width" and "Curb_Weight" being more focused on the car and its distinct features. Group 2 consisting of "Engine_Size", "Horsepower", "Power_Performance_Factor", "Price_in_Thousands" and "Resale_Value" making it more prevalent the fact that one of the car's biggest appeals were it speed and durability. The final cluster group consisted of the "Type" of car, the "Year_Launch" and "Fuel_Efficiency". But there are caveats since both Group 1 and 3 as expected have negative correlations and "Years_Launch" has almost no impact in these PCs.



Another interesting plot we made was between PC2 and PC3. In this case, we can see also three groups that have a positive correlation between them. Group 1 with "Years_Launch", "Fuel_Capacity" and "Curb_Weight" indicating that newer models may have larger fuel capacities and higher curb weights; Group 2 with "Sales_in_Thousands", "Wheelbase", "Width", "Length" suggesting larger vehicles have higher sales; and Group 3 with "Fuel_Efficiency", "Type", "Engine_Size", "Horsepower", "Power_Performance_Factor", "Price_in_Thousands" and "Resale_Value" correlating the engine size and horsepower with how expensive the car is and their resale value. Between these groups the correlation is almost none with the exception of the relationship between groups 2 and 3 which have a negative correlation. In these 2 PCs the "Type" of car is the most important while "Engine_Size" is the least.

A graph of variables was also made with PC3 and PC4 but showed very similar results between them with 2 groups being formed one of them containing most of the variables while the other group contained the variable that is negatively correlated with the PC in question (PC3 with "Type" and PC4 with "Years_Launch").

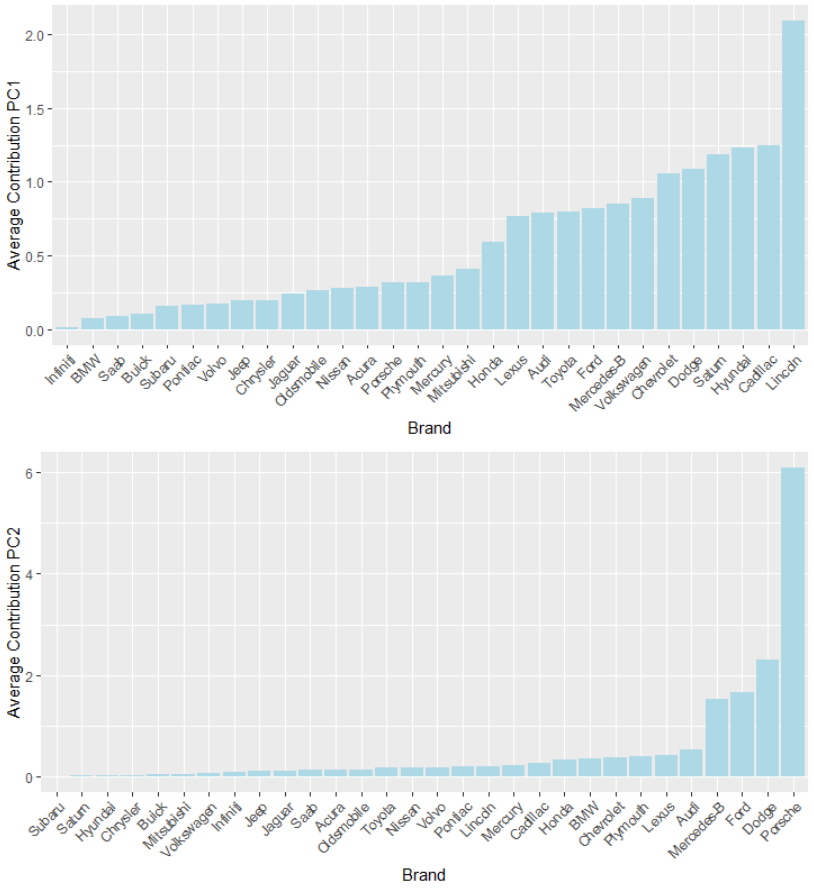
	PC1	PC2	PC3	PC4
Sales_in_Thousands	"0.0000001408209"	"0.1384421080033"	"0.0340778851112"	"0.0019338428052"
Resale_Value	"0.0414524667616"	"0.1670441116393"	"0.0037503834087"	"0.0003612229045"
Type	"0.0199898560959"	"0.1002072769656"	"0.3960034344713"	"0.0232150007604"
Price_in_Thousands	"0.0734280065617"	"0.1381100557701"	"0.0003733398606"	"0.0000485016396"
Engine_Size	"0.1188579788879"	"0.0023565943281"	"0.0032771028090"	"0.0000011766133"
Horsepower	"0.1040304240287"	"0.0719321892795"	"0.0181506846845"	"0.0002120524806"
Wheelbase	"0.0600119184238"	"0.1318825042803"	"0.0479386414748"	"0.0211223134272"
Width	"0.0888572242498"	"0.0310000686748"	"0.0541421998297"	"0.0097208381324"
Length	"0.0625414305908"	"0.0731566199161"	"0.1786866835242"	"0.0400410671776"
Curb_Weight	"0.1183102261887"	"0.0160487520701"	"0.0253473640375"	"0.0007401921930"
Fuel_Capacity	"0.1027827376290"	"0.0323154552242"	"0.0517612189285"	"0.0081861886649"
Fuel_Efficiency	"0.1049695523830"	"0.0058930615225"	"0.0968883479891"	"0.0058272300974"
Power_Performance_Factor	"0.1033476652678"	"0.0877703963400"	"0.0103107529728"	"0.0001179903410"
Years_Launch	"0.0014203721104"	"0.0038408059861"	"0.0792919608983"	"0.8884723827630"

The contributions of each variable was then calculated. For PC2, "Engine_Size" and "Fuel_Efficiency" had almost 0 contribution. "Width", "Curb_Weight" and "Fuel_Capacity", also had a low impact. It is important to note that the variables with low contribution, are all correlated between each other, like seen previously. The contribution of the rest of the variables varied between 7% and 16%. Just like PC1, any of the variables had a big impact in this PC, where the contribution was well distributed. PC3 was mainly defined by "Type" and "Length" with 57.4% of the total contribution. "Resale_Value", "Engine_Size" and "Price_in_Thousands" had almost no impact. PC4 is mainly defined by "Years_Launch", contributing to 88.8% of the PC. PC3 explained 8.6% of the total inertia, and is mainly defined by "Type" and

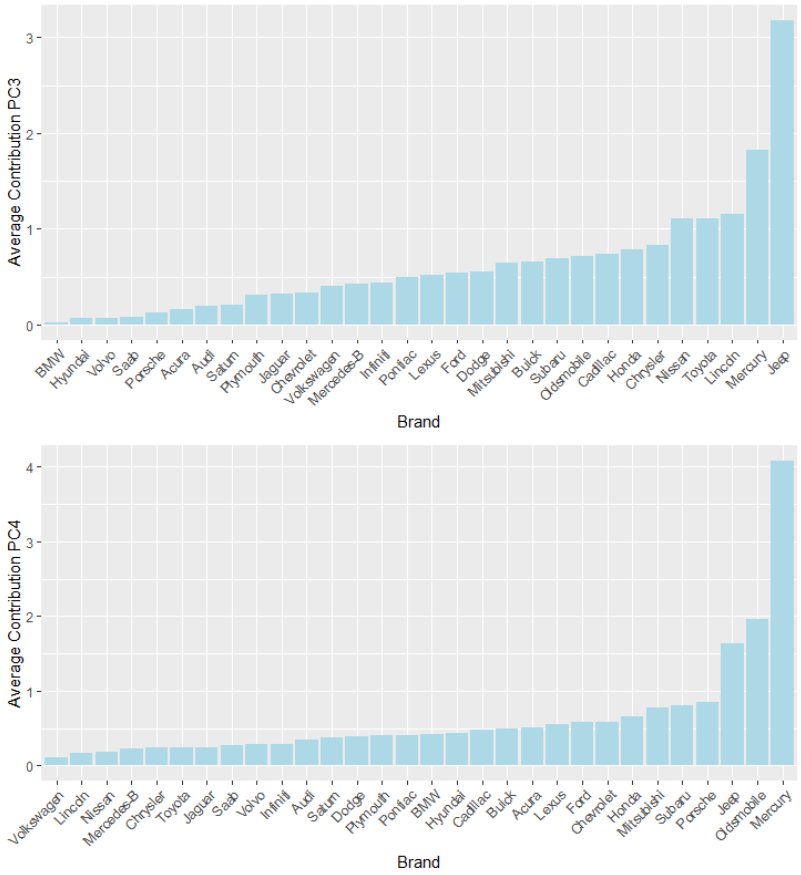
“Length”, like seen previously. We can see by the correlation matrix between the variables and PCs, that both showed a negative correlation with PC3.This meant that vehicles of the type 0, in this case "car", tend to have lower lengths. PC4 explained 7% of the total inertia, and was mainly defined by “Years_Launch”. We also know that this variable and PC4 had a negative correlation, meaning that PC4 mainly captured characteristics from younger cars.

6.1.3. “Brand”

We then investigated how each brand influenced each PC. For PC1, Lincoln was by far the most influential brand. This was a brand defined by its “Resale_Value”, “Engine_Size", “Horsepower”, “Wheelbase”, “Width”, “Length”, “Curb_Weight”. On the other hand, Infiniti was the least influential brand in PC1, a brand defined by its very little amount of cars, high “Resale_Value” and “Horsepower". For PC2, “Porsche” was by far the most influential brand. This brand generated tittle amount of money, was the most expensive, presented the best “Resale_Value”, the smaller wheelbase and the biggest “Power_Performance_Factor”. “Dodge", “Ford" and “Mercedes-Benz”, also showed a high contribution, with a massive drop to every other brand.

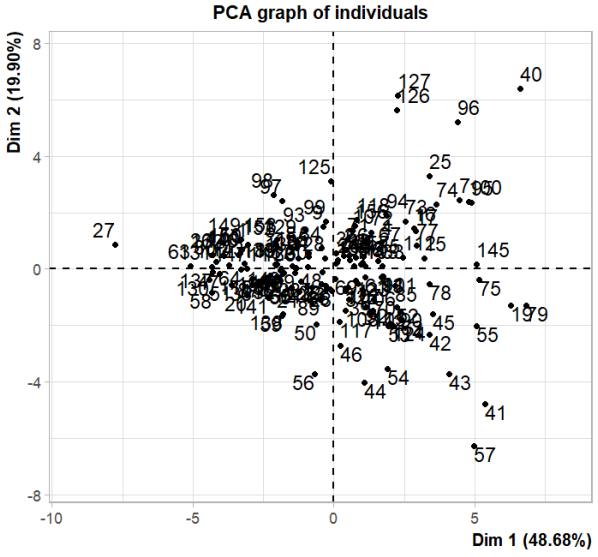


For PC3, that was mainly defined by "Type" and "Length", "Jeep" was by far the most influential brand "Jeep" was the only brand without a passenger vehicle (type 1) and was one of the brands with the most car vehicles (type 0). "Jeep" was also the smallest brand in terms of "Length". Since this variable had such a big impact in the PC, it showed negative correlation, explaining "Jeep" having the biggest impact. PC4 was mainly defined by "Years_Launch", a variable negatively correlated with the PC, "Mercury" was by far the most influential brand. This makes sense since it was one of the oldest brands.

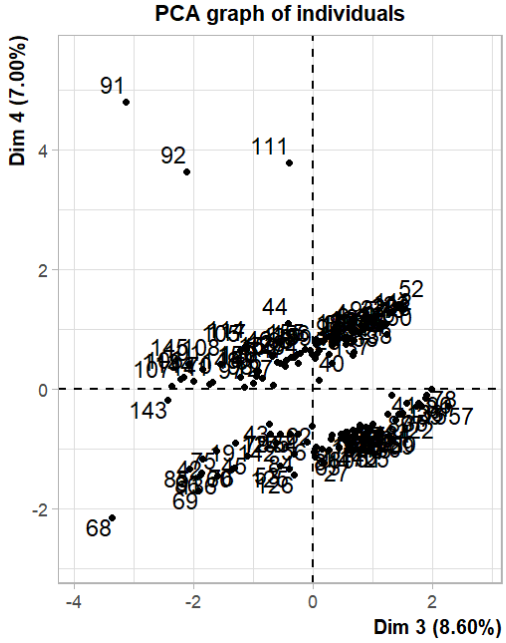


6.1.4. Individuals Interpretation

Investigating the individual cases, we calculated the coordinates, contributions and cos2 of each car to each Principal Component.



In terms of their coordinates for PC1 the most extreme cases were 79 ("Mitsubishi Mirage") and 40 ("Dodge Ram Pickup"), which had high valued PC1 coordinates with 6.82 and 6.60 respectively. 27 ("Chevrolet Metro") was the lowest with -7.75. The high positive scores on cars 79 and 40 made us believe that these cars may be high-performing with attributes like powerful engines and high curb weight while the coordinates of car 27 made us believe it was the opposite. In PC2 the standout scores were cars 57 ("Honda Civic" with -6.27), 127 ("Saab 5-Sep" with 6.16) and 40 (6.39). From the plot of individuals, we discovered that the cars that had high coordinates either negative or positive for PC2 generally had high positive coordinates for PC1. The vehicle in position 40 indicated that the car most likely represented a high-performance, large and popular vehicle.



The plot of individuals concerning PC3 and PC4 had interesting results. The first thing we could see was two groups of data that have an identical slight trend upward along the PC relationship. This combination of data suggests that many vehicles types had similar launch dates and also a correlation between the type of vehicle and the time period they were introduced. Another interesting point we saw was that vehicles 91 ("Mercury Villager" with 4.8), 92 ("Mercerdes-Benz C-Class" with 3.64) and 111 ("Oldsmobile Aurora" with 3.78) had significantly higher PC4 coordinates meaning that they differed the most in launch years.

For the contributions of the individual cars analyzed, we know that there were 157 individuals meaning that on average each individual should have a contribution of

around 1.57, any value above it meant the individual was relatively important to the component. For PC1 the number of individuals with a contribution above the average was 16 cars. Unsurprisingly, since they also had the most extreme coordinates the cars with the highest contributions for PC1 was 27,79 and 40. For PC2, 15 cars were above the average with the most extreme cases being 40 with 9.4% ,57 with 9.02% and 127 with 8.73%. For PC3 the number of contributions higher than 1.57 was 23 with the highest contribution being around 6% with vehicle 68 ("Jeep Cheroke"). Finally, PC4 only had 4 vehicles with a higher value than the average but vehicle 92 and 111 both had a contribution higher than 8%.

The Cos2 was then analyzed, showing us how each component or pair of components represented the individual. To identify the best representation, we studied if any pair of components or single component would represent more than 80% of the individual. Looking at each component, we can see that many of the vehicles analyzed were represented in PC1 with 20 of the vehicles having more than 80%. The most extreme case was vehicle 102 in which PC1 represented 97% of the vehicle. Despite this, only PC3 could also represent a vehicle in our dataset, this one being 107, with 83%. Looking at pairs of components, the results showed that the pair that could represent the most amount of individuals was pair PC1-PC2 with 13 vehicles. This was then followed by PC1-PC3 with 12, PC1-PC4 with 7, PC3-PC4 and PC2-PC3 both had 2 vehicles and PC2-PC4 could represent only one. From the results we could see that PC1 represented most of the vehicles either alone or in a pair.

6.2. Factor Analysis

6.2.1. Applying FA

Factor analysis was then conducted, with the purpose of identify underlying factors that help us explain relationships between hidden variables. Our first step was done by conducting a correlation matrix and performing the Kaiser-Meyer-Olkin Test. This helped us understand if the dataset was adequate for factor analysis, by examining the proportion of variance among variables that was caused by underlying factors.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = df_numeric)
Overall MSA =  0.82
MSA for each item =
      Sales_in_Thousands            Resale_Value                    Type      Price_in_Thousands             Engine_Size
                    0.73                    0.91                    0.67                    0.67                    0.95
              Horsepower              Wheelbase                   Width                  Length              Curb_Weight
                    0.72                    0.81                    0.96                    0.77                    0.88
           Fuel_Capacity       Fuel_Efficiency Power_Performance_Factor           Years_Launch
                    0.93                    0.95                    0.71                    0.48
```

We got an overall KMO score of 0.83, which indicated us that it was possible to do factor analysis. Almost every variable got at least a 0.67, making most of the variables suitable except for the variable “Years_Launch" which had 0.48 so it was left out of the model.

```
    Sales_in_Thousands            Resale_Value                    Type      Price_in_Thousands             Engine_Size
             0.9862319               0.8181926               0.9236878               0.8957330               0.8269493
            Horsepower              Wheelbase                   Width                  Length              Curb_Weight
             0.9348126               0.8545120               0.7865877               0.9003753               0.9018848
         Fuel_Capacity       Fuel_Efficiency Power_Performance_Factor
             0.8649217               0.8541995               0.9674944
```

Overall, the variables in the dataset had high communalities, suggesting that they are well represented by the extracted factors. Variables with lower communalities like “Width”, may have had unique variance that was not explained by the extracted factors, indicating that there may have been other factors or variables influencing them.

```
Loadings:
                        RC2    RC1    RC3    RC4
Sales_in_Thousands     -0.201  0.191          0.950
Resale_Value            0.894 -0.127
Type                    0.127         -0.938 -0.158
Price_in_Thousands      0.924         0.102  -0.158
Engine_Size             0.676  0.514  0.323
Horsepower              0.894  0.344
Wheelbase                      0.850  0.298  0.208
Width                   0.281  0.807  0.239
Length                         0.935         0.110
Curb_Weight             0.395  0.599  0.614 -0.103
Fuel_Capacity           0.311  0.525  0.702
Fuel_Efficiency        -0.451 -0.380 -0.711
Power_Performance_Factor 0.925 0.297  0.104 -0.108

                 RC2   RC1   RC3   RC4
SS loadings    4.366 3.560 2.546 1.044
Proportion Var 0.336 0.274 0.196 0.080
Cumulative Var 0.336 0.610 0.805 0.886
```

The loadings showed us that RC2 had high correlations with ”Resale_Value”, “Price_in_Thousands”, “Horsepower" and “Power_Performance_Factor”

representing attributes related to vehicle cost and performance. RC1 with “Wheelbase”, “Width” and “Length" representing the physical dimensions of the vehicle. The RC3 had a negative correlation only with “Type” while RC4 had a high correlation with ”Sales_in_Thousands” distinguishing the vehicles in terms of their price. The variability of each rotative component shows us that RC2 represents the largest amount of variability in the model. The cumulative variability indicated that the first 4 RC show 86% of the total variability. The residuals were then calculated, and the results showed values close to zero, with all of them having at least 10-2 significant figures. This indicates a generally good fit of the model to the data, as the residuals are minimal.

3 different methods of factor extraction were then made: With Principal axis, minimal residuals and finally with the maximum likelihood.

6.2.2. Principal Axis

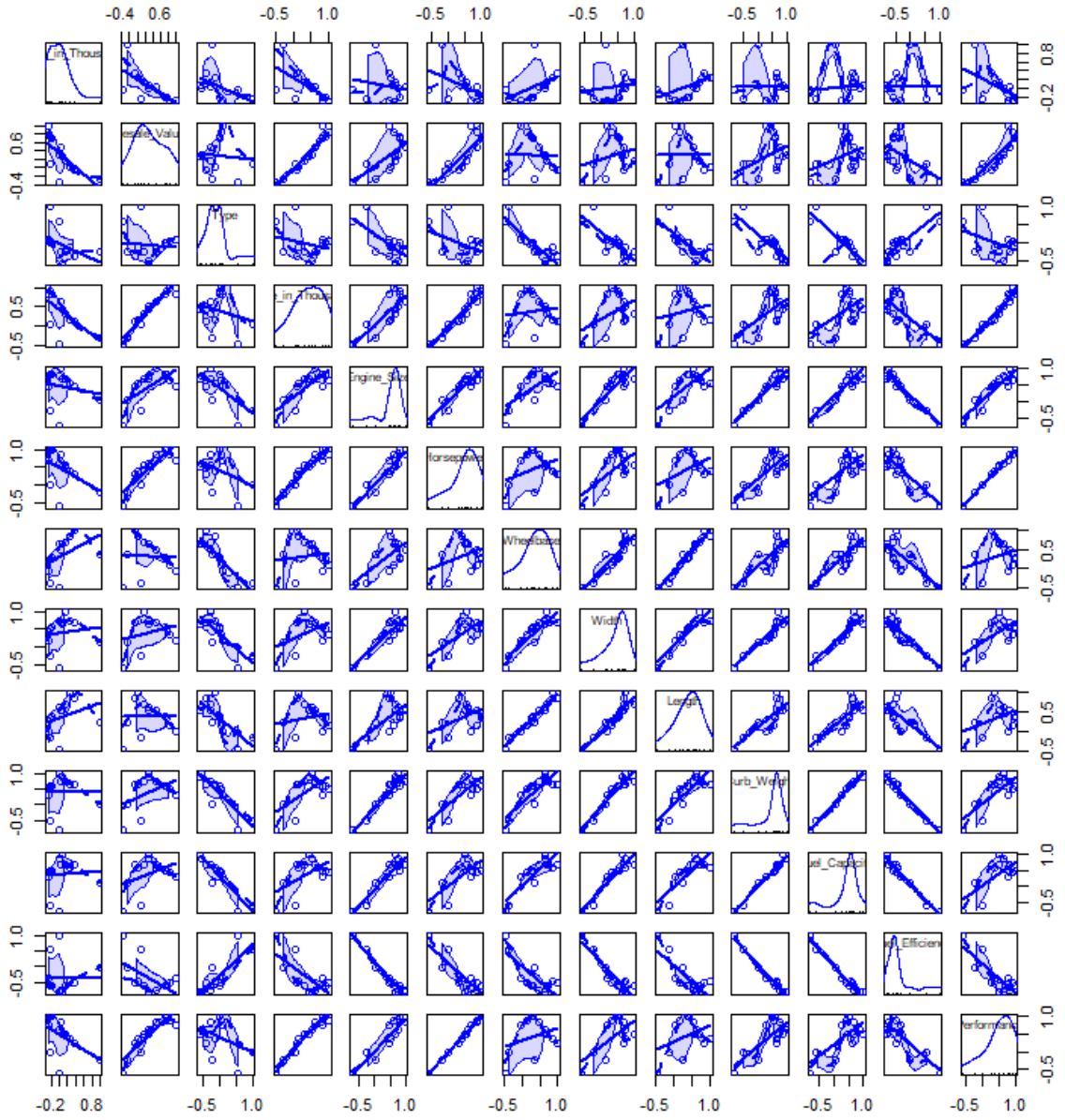
In terms of PAs with a higher correlation than 0.8: “Fuel_Capacity”, “Price_in_Thousands”, “Horsepower” and “Power_Performance_Factor” had a significant higher correlation in PA2, “Wheelbase” and “Length” in PA1, “Type” had a negative correlation in PA3 while “Sales_in_Thousands” was the only high correlation in PA4. In terms of variance explained by each factor, PA2 (39%) was the factor that performed better, followed by PA1 (30%), PA3 (22%), and then PA4 (0.09%). “Curb_Weight” showed the biggest complexity. Mean item complexity indicated that, on average, each variable loaded on approximately 1.6 factors. Root Mean Square of the Residuals represented the average discrepancy between the observed values and the values predicted by the model. In our case, 0.02 indicated a very good fit.

6.2.3. Manual Residuals

For MR1 “Wheelbase” and “Length” had significantly high correlations. “Resale_Value", “Price_in_Thousands”, “Horsepower”, “Power_Performance_Factor” for MR2, MR3 again had a highly negative “Type” and there was no correlation in MR4. In terms of variance MR2 (39%) was the factor that performed better, followed by MR1 (31%), MR3 (25%), and then MR4

(0.05%). “Engine_Size” had the greatest complexity. Mean item complexity showed that, on average, each variable loaded on approximately 1.8 factors, a little above the factor analysis with the principal axis. Root Mean Square of the Residuals was identical to principal axis.

6.2.4. Maximum Likelihood



A scatterplot of the df_numeric without “Years_Launch” was then made to check if the variables were under normal distribution and see if it was available for feature extraction with maximum likelihood. The diagonal of the scatterplot showed us a

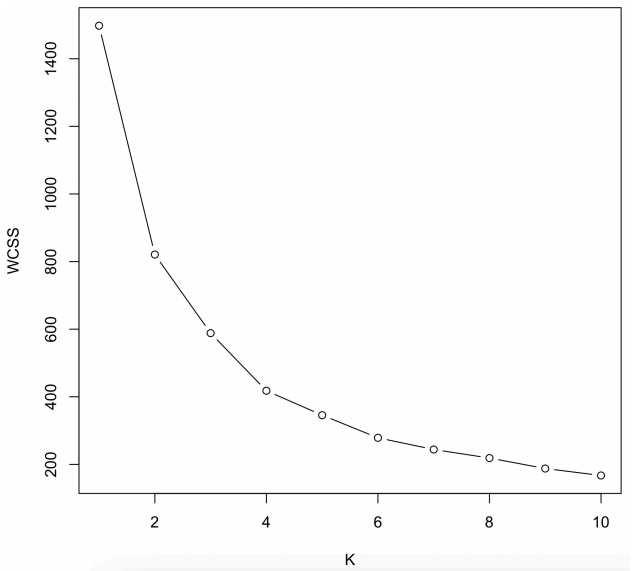
smooth histogram that demonstrated the distribution of each variable. The plot indicated us that most of the variables either had a positive or negative skewness, meaning that they are not well adjusted for maximum likelihood.

6.3. Clustering Analysis

6.3.1. Non-Hierarchical Clustering

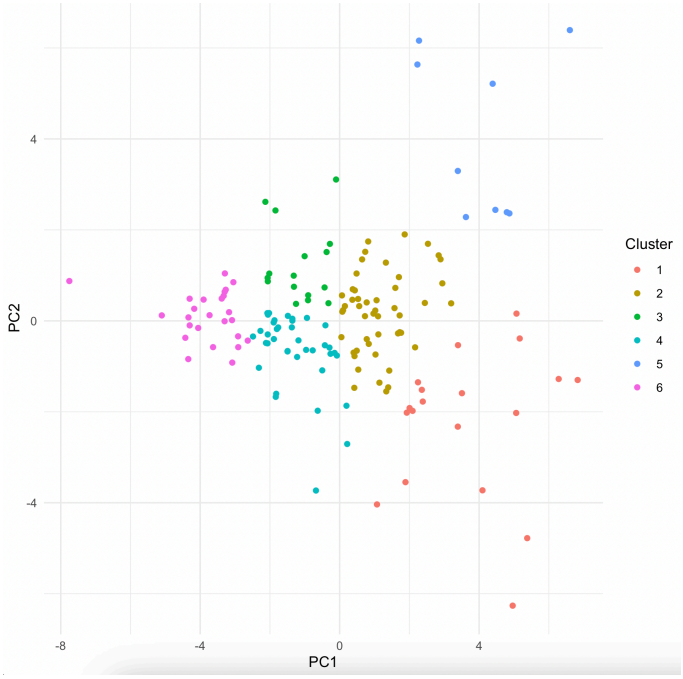
We began the cluster analysis by standardizing the dataset to ensure that each variable contributed equally to the analysis. Principal Component Analysis (PCA) was then applied to reduce the dimensionality of the dataset while retaining most of the variability present in the original variables. The resulting dataframe contained the coordinates of the observations on the first two principal components, which explained a substantial portion of the total variance. These principal components were subsequently used for clustering, as they provide a simplified yet informative representation of the data.

To determine the optimal number of clusters, we utilized the elbow method. This technique involves running k-means clustering with varying numbers of clusters (ranging from 1 to 10) and plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters. The WCSS measures the total variance within each cluster, and the goal is to find the point at which adding another cluster does not significantly reduce the WCSS, indicating diminishing returns. The plot revealed that the WCSS stabilized around six clusters, suggesting that k=6. is an appropriate choice for our data.



Using k=6 clusters as determined by the elbow method, we performed k-means clustering on the first two principal components. K-means clustering partitions the data into k clusters by minimizing the variance within each cluster. The algorithm iteratively assigns each observation to the nearest cluster centroid, then recalculates the centroids based on the new assignments, repeating this process until convergence.

The cluster assignments and centroids were computed and visualized using a scatter plot of the first two principal components. Each point was colored according to its cluster assignment, providing a clear visual representation of the cluster separation. This plot helped us to understand the spatial distribution of the clusters and the separation between them.



To evaluate the clustering quality, we calculated the silhouette coefficient, which measures how similar each point is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, with higher values indicating better-defined clusters. Our k-means clustering resulted in a silhouette coefficient of 0.54, indicating reasonably good clustering quality. This suggests that the clusters are well-separated, and the data points within each cluster are relatively close to each other compared to points in other clusters.

After determining the cluster assignments, we combined the original dataframe with the cluster labels to analyze the descriptive statistics of each cluster. This involved calculating various metrics such as the count of observations, average sales, average

resale value, most common type, average price, and other relevant attributes for each cluster. This detailed analysis provided insights into the characteristics of each cluster.

Cluster 1:

- High PC1 and highest PC2.
- Contains the most observations.
- Exhibits the highest average sales.
- Has the lowest average resale value, average price, average engine size, and average horsepower.
- Displays low average width and length.
- Cluster 1 is characterized by high sales but low resale value and vehicle specifications, indicating it might consist of vehicles that are popular in terms of sales but not necessarily high-end or performance-focused.

Cluster 2:

- Low PC1.
- Contains a low number of observations.
- Exhibits low average sales.
- Shows high average resale value and high average price.
- High average horsepower.
- Cluster 2 consists of fewer vehicles that have high resale values and prices, indicating these might be luxury or high-performance vehicles that are less commonly sold but retain their value well.

Cluster 3:

- Contains the fewest observations.
- Shows low average price and low average horsepower.
- Displays a low average wheelbase.
- Cluster 3 might represent a niche segment of vehicles with lower prices and smaller sizes, possibly compact or economy cars.

Cluster 4:

- Lowest PC1.
- Exhibits the highest average resale value and highest average price.
- Shows the highest average engine size, highest average horsepower, highest average wheelbase, highest average width, and highest average length.
- Cluster 4 is characterized by high-end vehicles with significant size and power, likely representing luxury or high-performance segments.

Cluster 5:

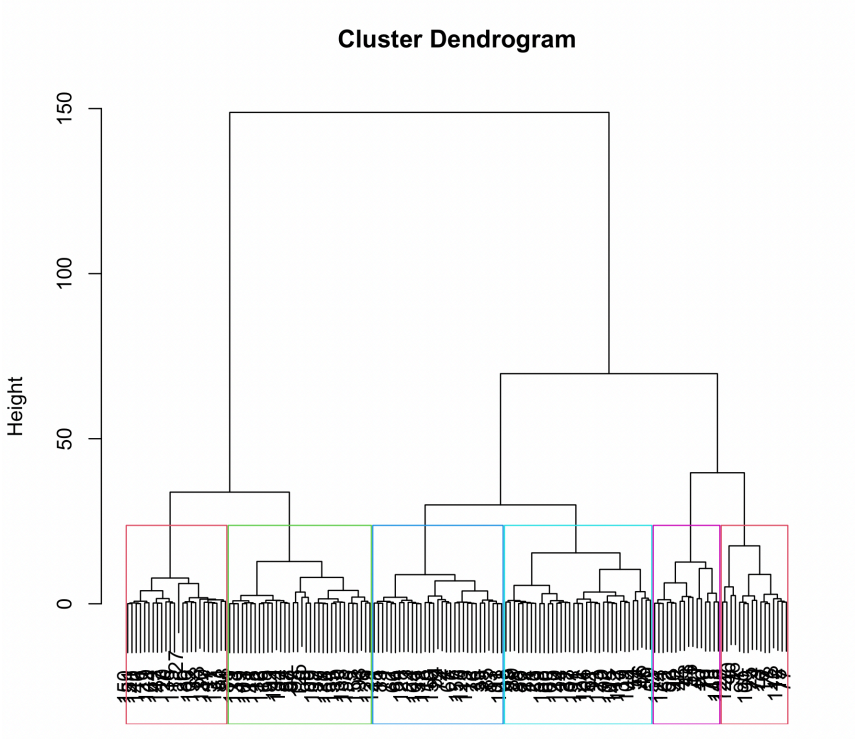
- High PC1 and lowest PC2.
- Exhibits the lowest average sales.
- Shows the lowest average wheelbase, width, and length.
- Displays low average engine size.
- Cluster 5 contains vehicles that are less popular in terms of sales and have smaller dimensions, possibly indicating a segment of less popular or more specialized vehicles.

Cluster 6:

- High PC1.
- Contains a high number of observations.
- Exhibits high average sales.
- Shows low average resale value.
- Displays high average engine size, high average wheelbase, high average width, and high average length.
- Cluster 6 is characterized by high sales and large vehicle dimensions but lower resale value, possibly indicating popular vehicles that might not retain their value as well as luxury or performance vehicles.

6.3.2. Hierarchical Clustering

For comparison, we also applied hierarchical clustering using Ward's method on the PCA-transformed dataset. Ward's method minimizes the total within-cluster variance, producing more compact clusters. The resulting dendrogram was cut to form six clusters, consistent with the number identified by the elbow method for k-means clustering.



However, the silhouette coefficient for hierarchical clustering was 0.30, indicating worse performance compared to k-means clustering. This lower silhouette coefficient suggests that the clusters formed by hierarchical clustering were not as well-separated and had more overlapping compared to those formed by k-means clustering.

The k-means clustering approach with k=6 provided a better fit for the data compared to hierarchical clustering, as evidenced by the higher silhouette coefficient. The cluster profiles offered valuable insights into the characteristics of different vehicle segments, highlighting variations in sales, resale value, price, and other key metrics. This analysis demonstrates the utility of PCA for dimensionality reduction and k-means clustering for segmenting complex datasets. The results suggest that k-means clustering on PCA-transformed data is effective for identifying distinct vehicle segments with specific characteristics, which can be valuable for targeted marketing and strategic decision-making.

6.4. Linear Discriminant Analysis

In order to perform a LDA we first need to make sure the assumptions are met. One of these assumptions concerns the distribution of the variables, being multinormality due to its necessity in statistical test. To check the distribution, we used explore in the IBM SPSS statistics viewer, and calculated the Kolmogorov-Smirnov and Shapiro-Wilks for each predictor variables. To perform LDA we believed the grouping variable should be either "Type" or "Brand" meaning all other numerical variables were tested. Looking at the table below, the p-values for Shapiro-Wilks are all significant with the exception of "Length" meaning none of the variables are normally distributed and the LDA assumptions are not met.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Sales_in_Thousands	,219	156	<,001	,667	156	<,001
Resale_Value	,263	156	<,001	,725	156	<,001
Price_in_Thousands	,162	156	<,001	,838	156	<,001
Engine_Size	,121	156	<,001	,930	156	<,001
Horsepower	,077	156	,026	,949	156	<,001
Wheelbase	,082	156	,013	,945	156	<,001
Width	,080	156	,017	,968	156	,001
Length	,051	156	,200*	,993	156	,704
Curb_Weight	,067	156	,089	,968	156	,001
Fuel_Capacity	,120	156	<,001	,926	156	<,001
Fuel_Efficiency	,107	156	<,001	,949	156	<,001
Power_Performance_Factor	,087	156	,006	,941	156	<,001
Years_Launch	,294	156	<,001	,654	156	<,001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

6.5. Linear Regression

6.5.1. Initial Model

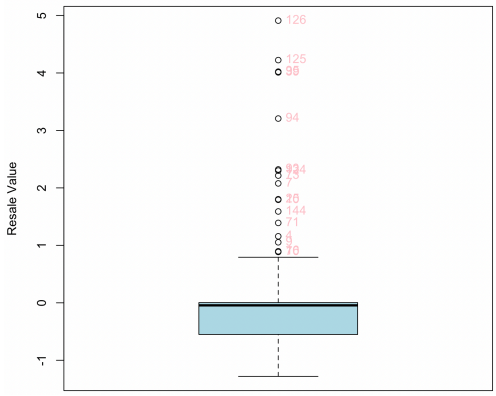
To predict the "Resale_Value" of vehicles, we employed a Linear Regression model using our dataset of numerical variables. First, we split our dataset into training and testing sets with an 80/20 ratio to ensure a robust evaluation of our model's performance.

We trained the Linear Regression model on the training dataset and made predictions on the test dataset. The Root Mean Squared Error (RMSE) for our initial model was 4.6, indicating the average prediction error. Analyzing the residuals, we found that the model underestimated the target variable by up to approximately 31 units and overestimated by up to approximately 13 units. The median residual was close to zero, suggesting relatively accurate predictions on average. However, the considerable variability in residuals highlighted potential outliers and areas where the model's performance could be improved.

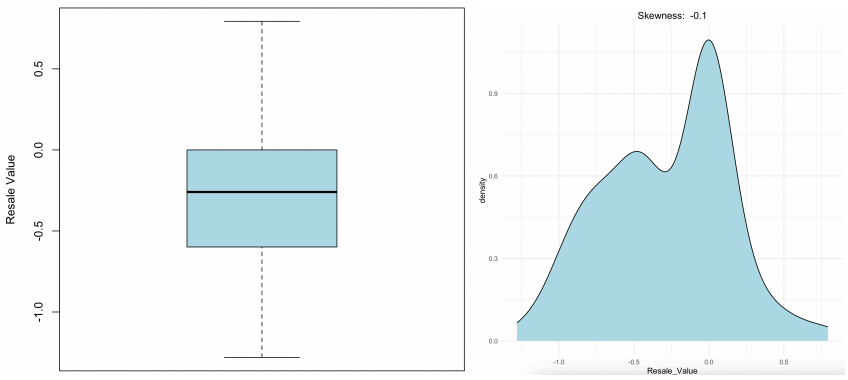
The significance of variables showed that “Price_in_Thousands”, “Curb_Weight", and “Fuel_Capacity” had significant p-values, indicating their substantial impact on “Resale_Value”. The Multiple R-Squared value of 0.70 indicated that 70% of the variance in “Resale_Value” was explained by the predictor variables. The Residual Standard Error (RSE) was 5.7, representing the average deviation of the model's predictions from the actual values.

6.5.2. Standardized Model Without the Outliers

To enhance model performance, we standardized the data and removed outliers, as indicated by the box-plot of the target variable.



Outliers were identified and removed based on the standardized “Resale_Value”. After removal, the skewness of the target variable was balanced, with a skewness level close to 0.



The standardized dataset was split into training and testing sets with an 80/20 ratio. We trained a new Linear Regression model on the standardized and outlier-free training dataset and made predictions on the test dataset. The RMSE for the standardized model was lower, indicating more accurate predictions with less variability.

In the standardized model, the variables “Engine_Size", “Width", and “Curb_Weight“ had significant p-values. However, the Multiple R-Squared value dropped to 0.54, suggesting a reduction in the variance of “Resale_Value” explained by the predictor variables. This reduction is due to the removal of outliers and standardization, which can reduce the model's ability to capture extreme variations. The RSE was also lower, indicating a tighter fit of the model to the data.

7. Conclusion

In conclusion, the analysis explained in this report provides us insight into the vehicle sales data identifying factors that had an influence on these results. This was done through an initial phase of univariate analysis where we were able to understand each variable in the data, understanding both its distribution and other specific measures, setting the stage for further analysis.

The bivariate analysis consisted focused mainly on the impact of categorical variables such as “Type” and “Brand” on other numerical variables. In terms of “Brands”, “Ford” was the most successful while “Porsche" had the highest mean price per vehicle. The analysis focusing on vehicle type demonstrated that “Cars" sold significantly better than “Passenger” vehicles. The ROI plot made for the Numerical

x Numerical analysis demonstrated that only 7 of 155 vehicles had an above average ROI.

The multivariate analysis consisted of PCA, Factor Analysis, Cluster analysis, LDA and Linear Regression exploring various relationships between variables and individuals. For PCA, the data was reduced to 4 Principal Components, the first two contributing to almost 70% of the variance and representing the vehicle characteristics and the trade-off between vehicle size and sales. Factor Analysis showed us the underlying relationship between vehicle cost and performance, with various extraction methods being tested. The Clustering analysis identified 6 clusters each containing insights into the characteristics of the vehicles and their corresponding sales/resale values. LDA was not able to be performed due the variable distribution, while a linear regression model indicated variables that had an impact on resale value. However, the standardized model showed a reduced explanatory power, emphasizing the importance of outliers in capturing market extremes.

Overall, the findings underscore the complexity of the market, where multiple factors influence sales and resale values. Future research could extend these insights by incorporating additional variables and data, while also exploring other analytical methods.

8. References

[1] https://www.kaggle.com/datasets/gagandeep16/car-sales/data