Estatística e Análise de Dados

Estudo de Venda de Carros

Projeto Realizado por José Rodrigues e Pedro Leite

Professora Maria Paula Brito

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Tabela de Conteúdos

Variáveis Categóricas

- Manufacturer
- Model
- Vehicle_Type

Nominais

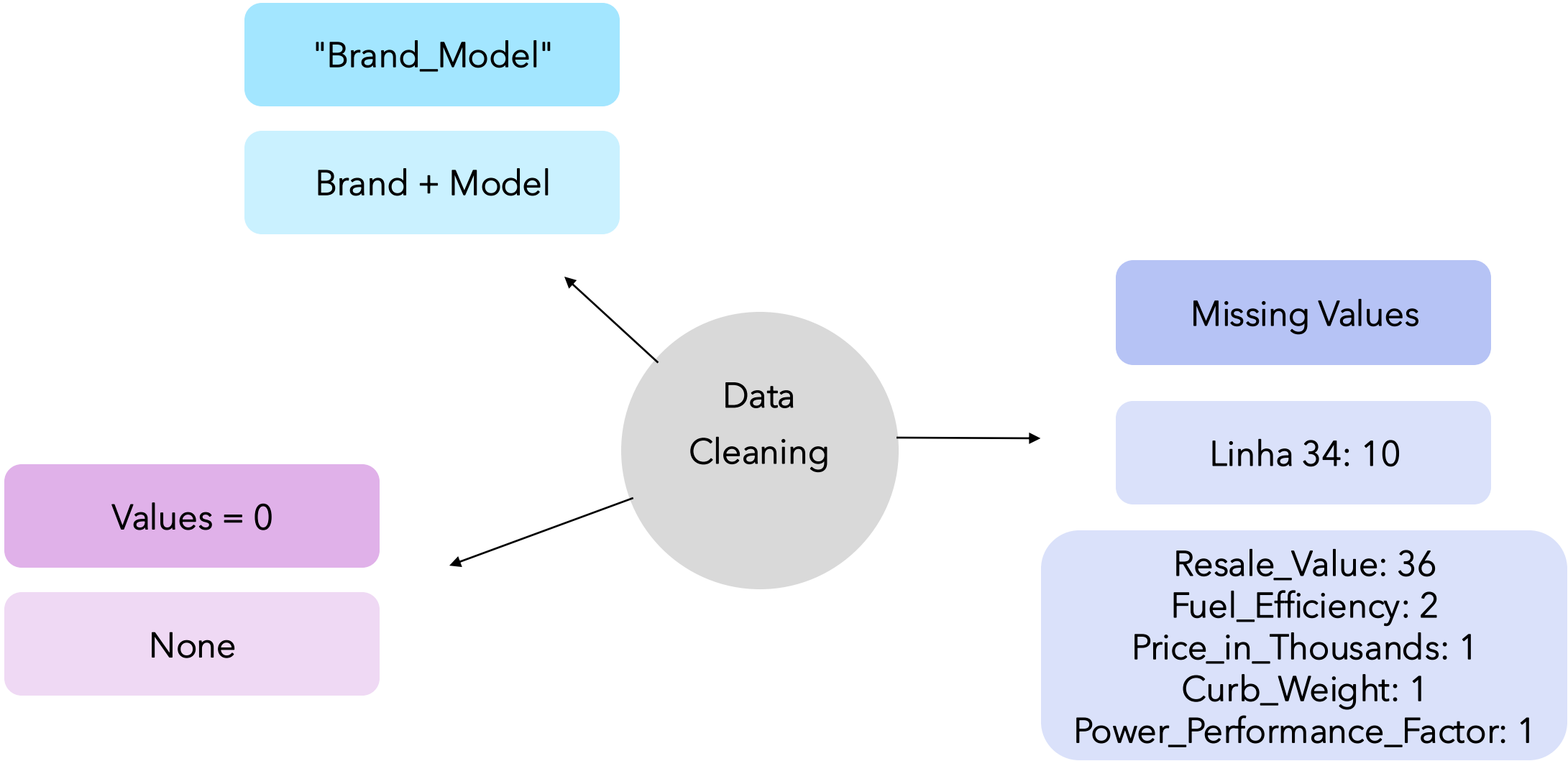
- Latest_Launch

Ordinal

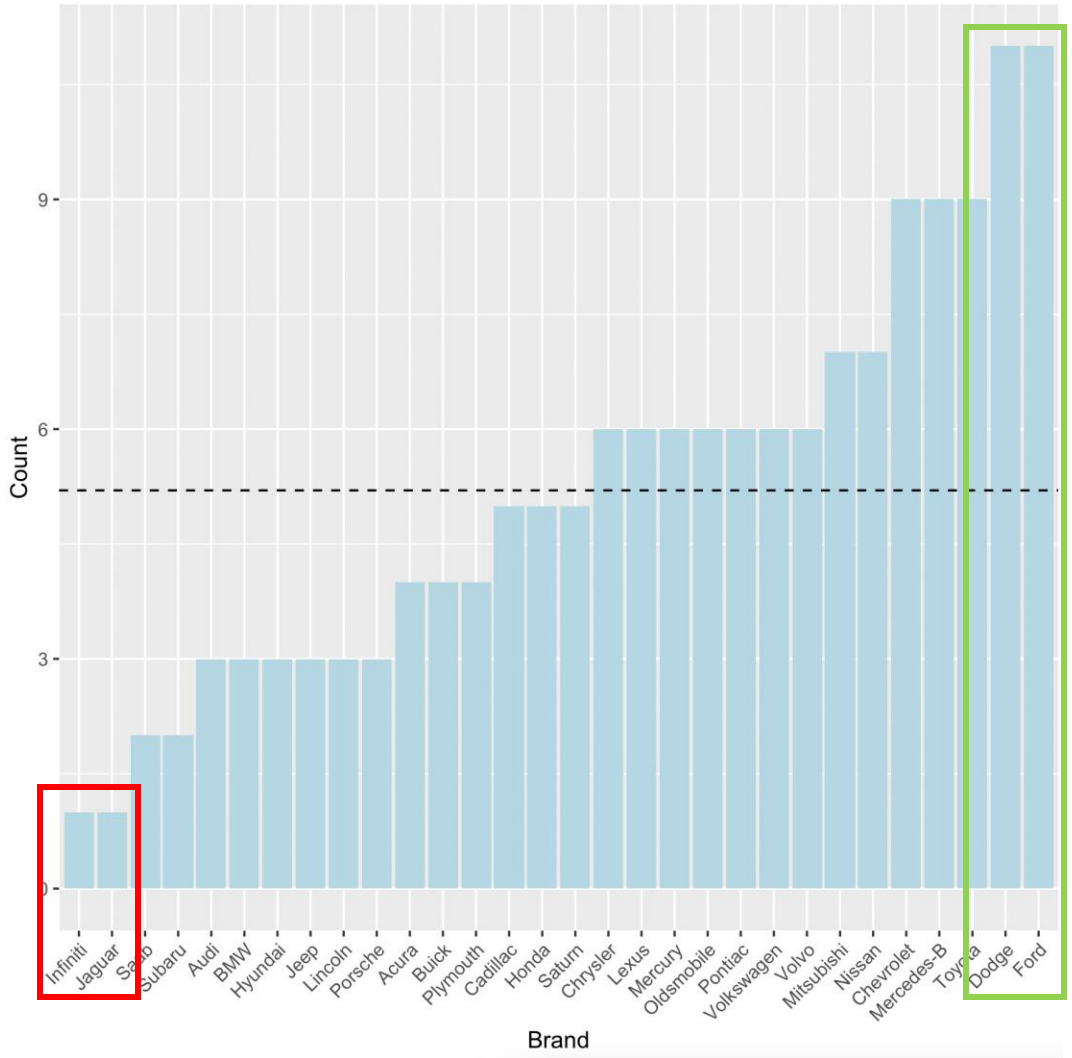
Variáveis Numéricas

- Sales_in_Thousands
- Price_in_Thousands
- Resale_Value
- Engine_Size
- Wheelbase
- Width
- Length
- Curb_Weight
- Fuel_Capacity
- Fuel_Efficiency
- Horsepower
- Power_Performance_Factor

Contínuas


"Brand_Model"
Brand + Model
Data Cleaning
Missing Values
Linha 34: 10
Resale_Value: 36
Fuel_Efficiency: 2
Price_in_Thousands: 1
Curb_Weight: 1
Power_Performance_Factor: 1
Values = 0
None

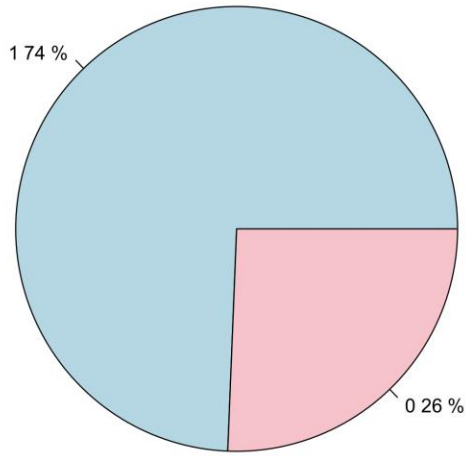
ANÁLISE UNIVARIADA: VARIÁVEIS CATEGÓRICAS

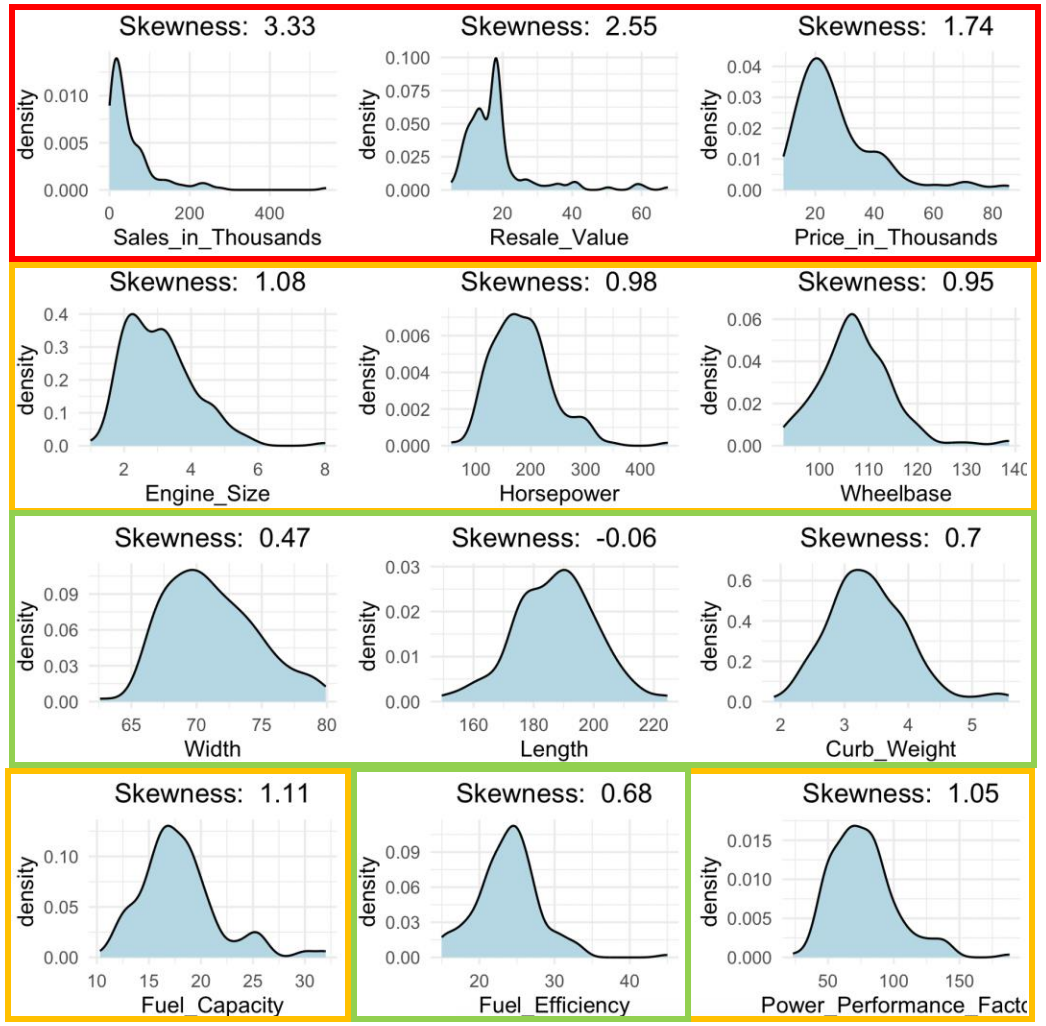


	Brand	Frequency	Cumulative_Frequency	Total
1	Dodge	7.1	7.1	11
2	Ford	7.1	14.2	11
3	Chevrolet	5.8	20.0	9
4	Mercedes-B	5.8	25.8	9
5	Toyota	5.8	31.6	9
6	Mitsubishi	4.5	36.1	7
7	Nissan	4.5	40.6	7
8	Chrysler	3.8	44.4	6
9	Lexus	3.8	48.2	6
10	Mercury	3.8	52.0	6

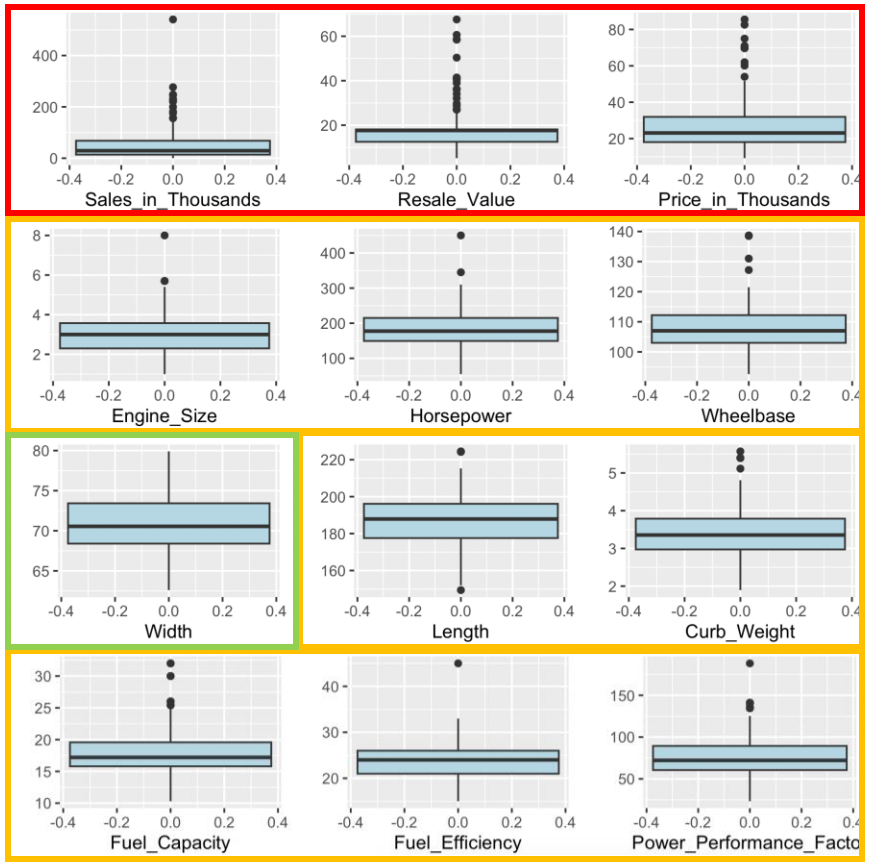
ANÁLISE UNIVARIADA: VARIÁVEIS NUMÉRICAS

"TYPE":
BLUE - "PASSENGER"
PINK - "CAR"





ANÁLISE UNIVARIADA: VARIÁVEIS NUMÉRICAS

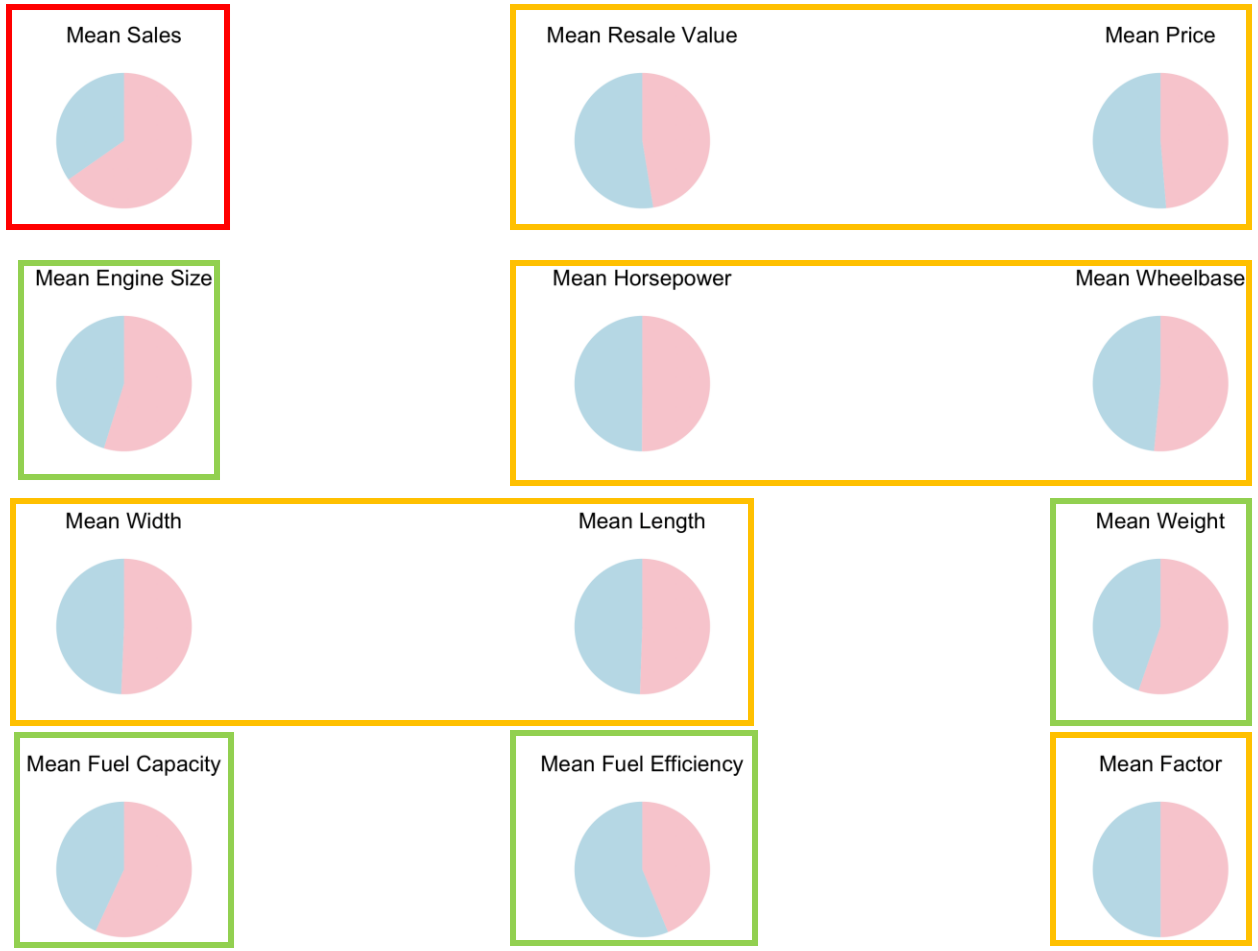



```
> all_outliers
 [1] "Ford F-Series"          "Ford Explorer"          "Toyota Camry"           "Ford Taurus"
 [5] "Honda Accord"           "Dodge Ram Pickup"       "Ford Ranger"            "Honda Civic"
 [9] "Dodge Caravan"          "Ford Focus"             "Jeep Grand Cherokee"    "Ford Windstar"
[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"
```

38 MODELOS

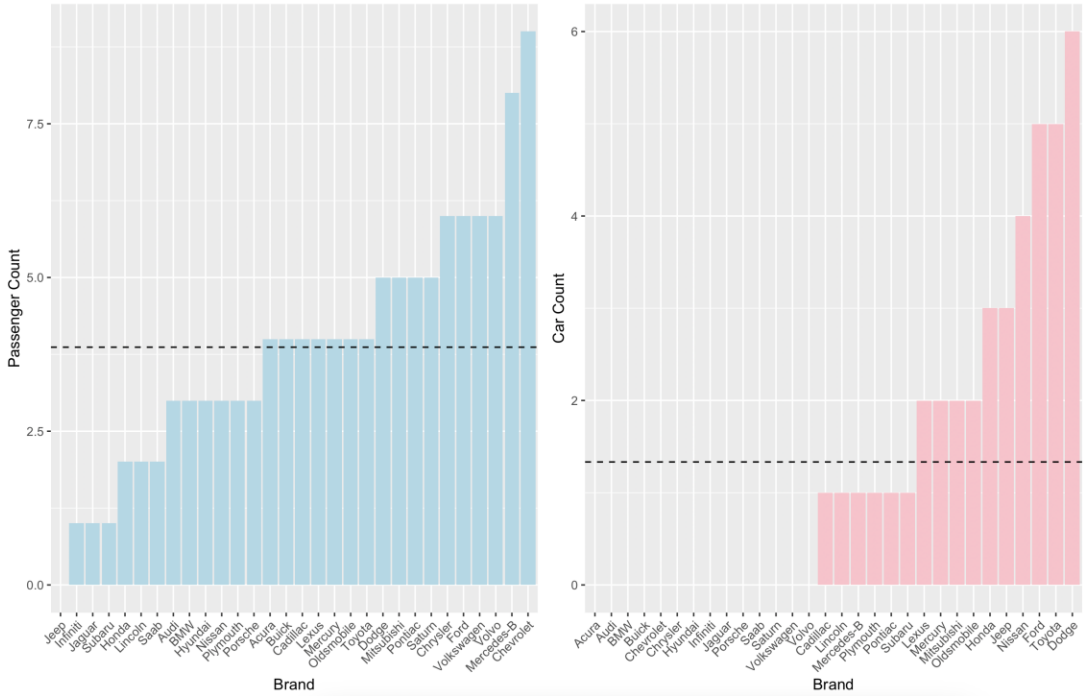
ANÁLISE BIVARIADA: VARIÁVEIS CATEGÓRICAS X NUMÉRICAS

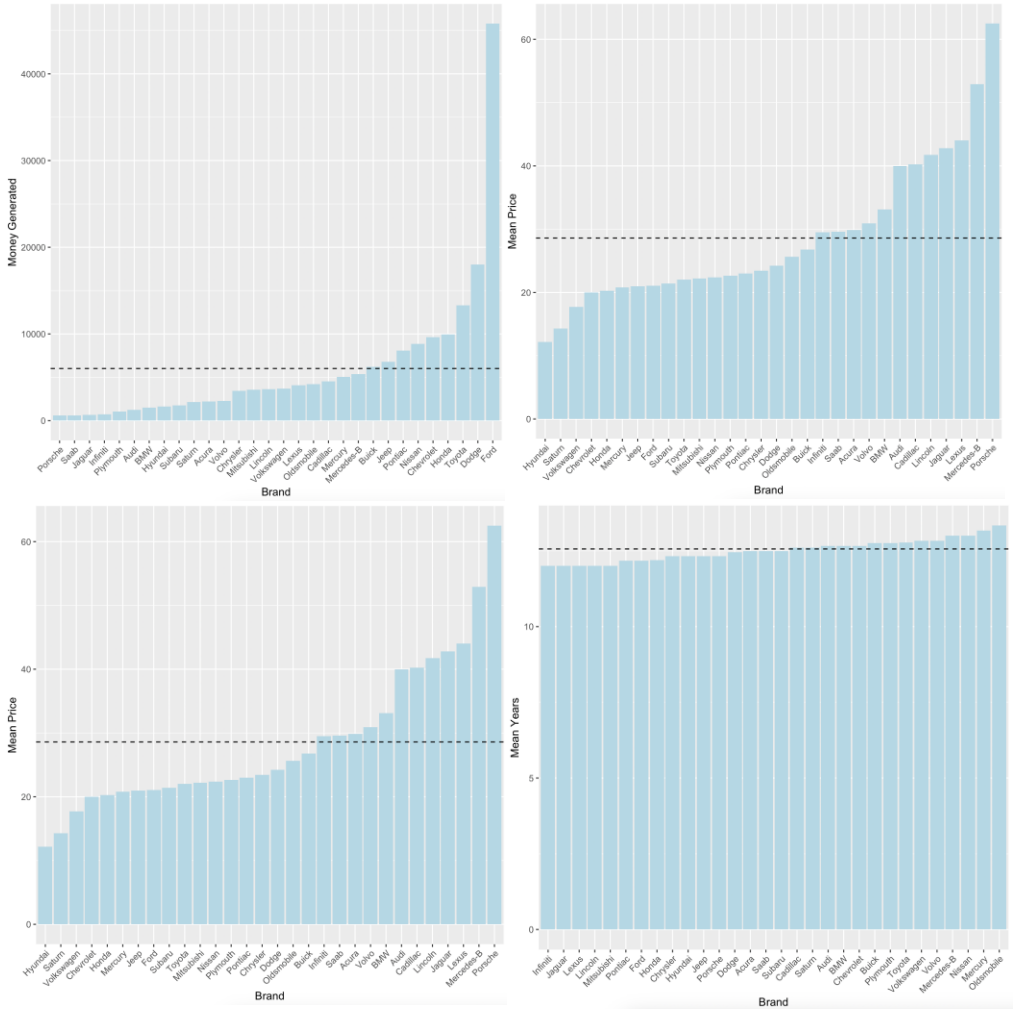
"TYPE"



ANÁLISE BIVARIADA: VARIÁVEIS CATEGÓRICAS X NUMÉRICAS

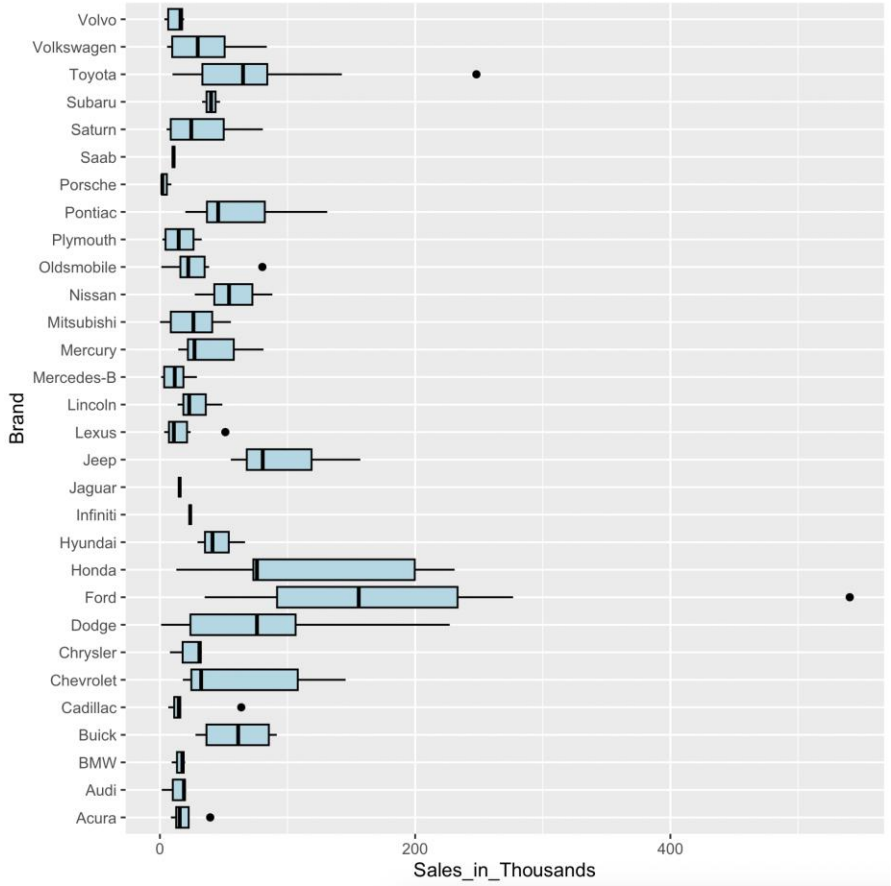
"BRAND"





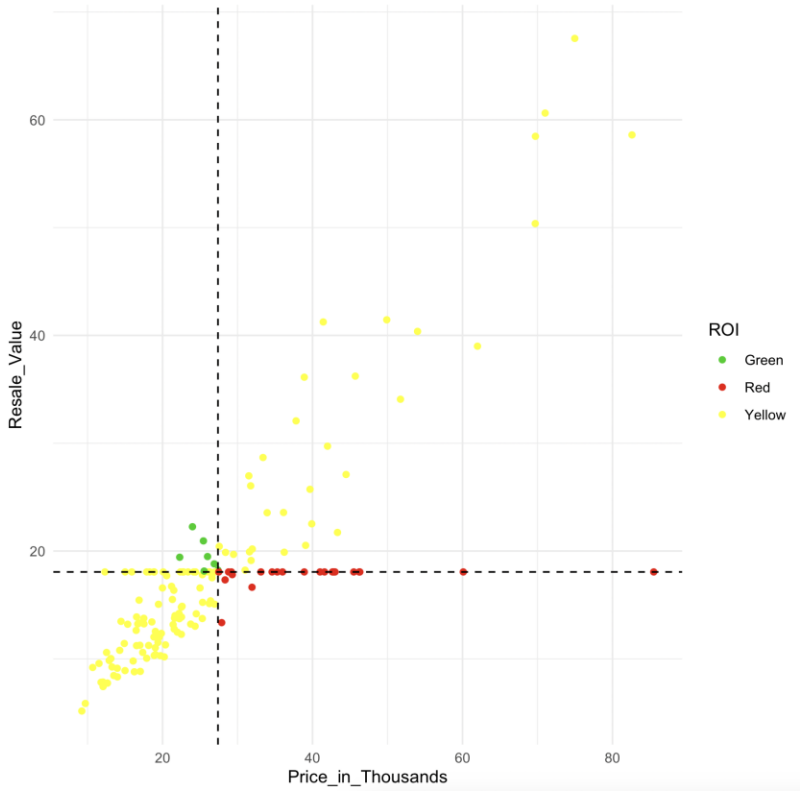
ANÁLISE BIVARIADA: VARIÁVEIS CATEGÓRICAS X NUMÉRICAS

"BRAND"

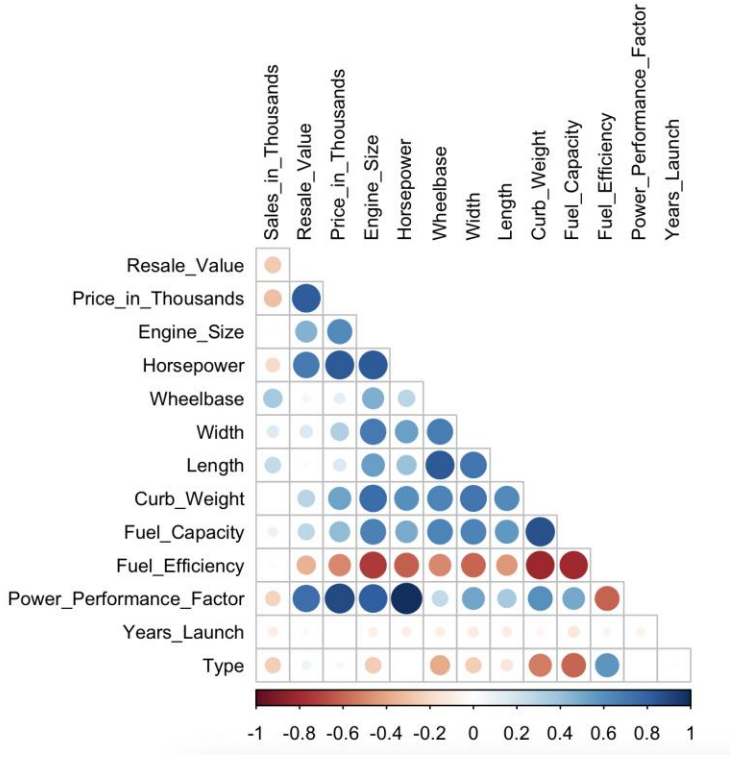


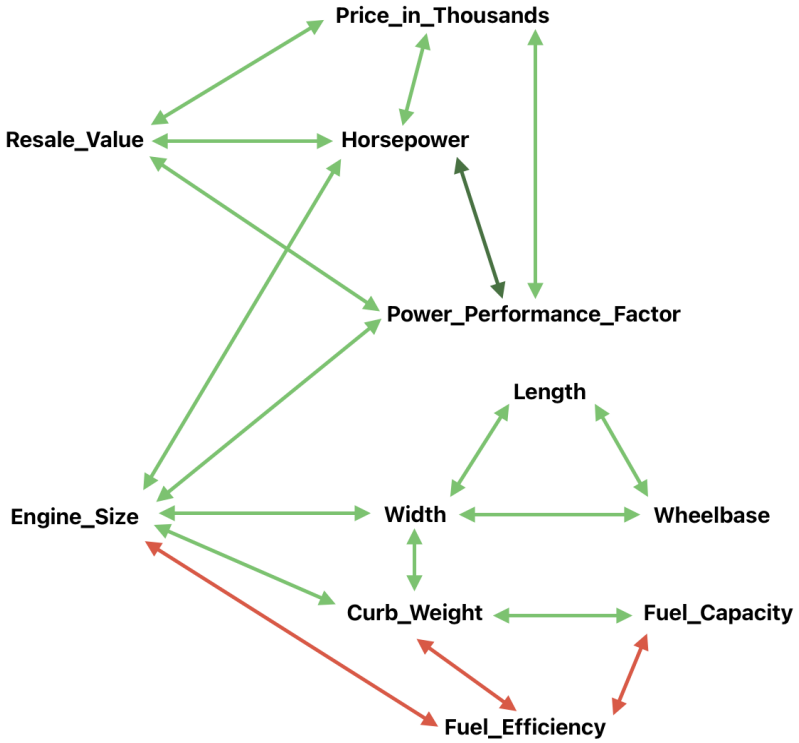
FORAM FEITOS E ANALISADOS BOX-PLOTS DA VARIÁVEL "BRAND" COM TODAS AS VARIÁVEIS NUMÉRICAS

ANÁLISE BIVARIADA: VARIÁVEIS NUMÉRICAS X NUMÉRICAS



ANÁLISE BIVARIADA: VARIÁVEIS NUMÉRICAS X NUMÉRICAS





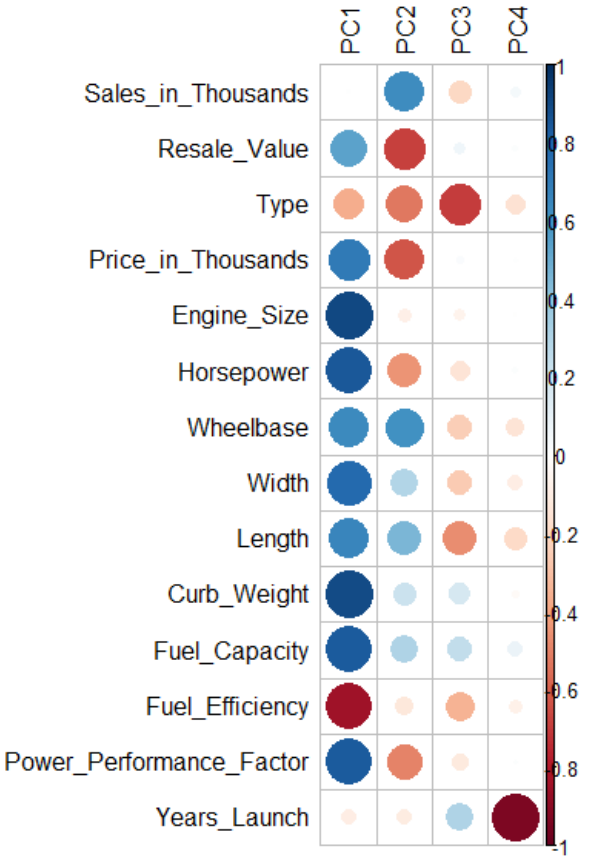
ANÁLISE MULTIVARIADA: NORMED PCA

Inércia

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

PC1: Características do veículo
PC2: Contraste entre vendas e tamanho do veículo
PC3: Diferenciar "Passenger" de "Car"
PC4: Representa "Launch_Year"

Correlação

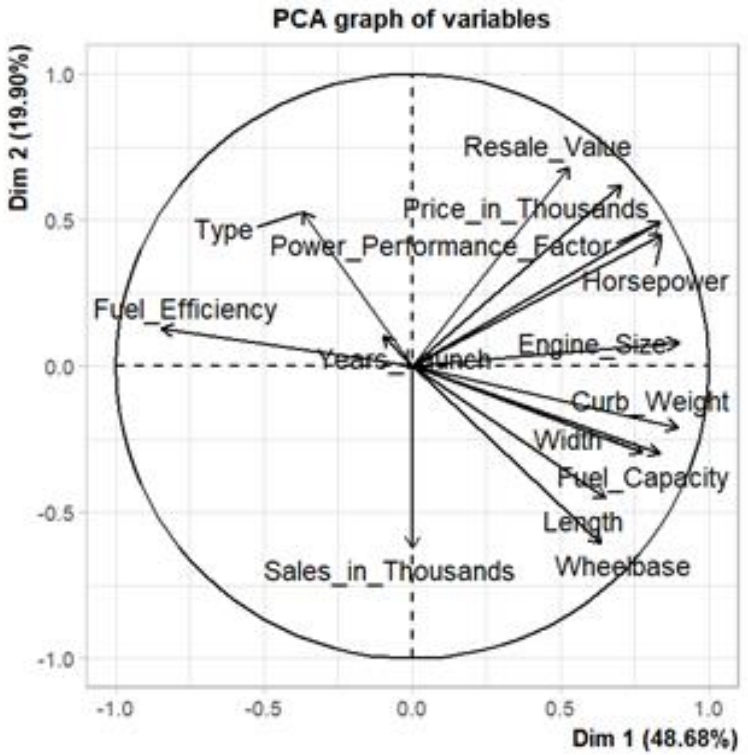


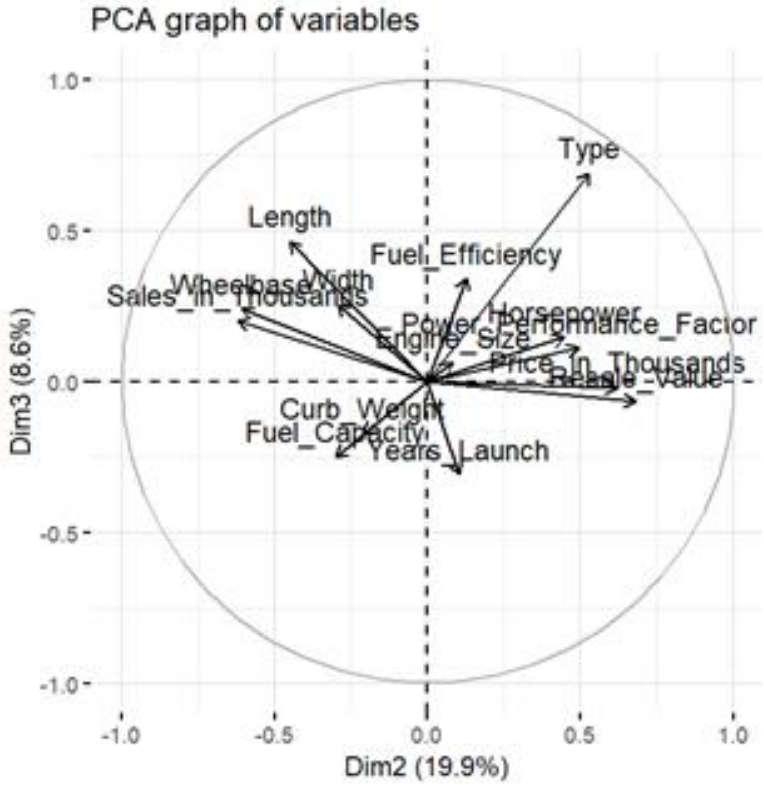
ANÁLISE MULTIVARIADA: NORMED PCA (CONTRIBUIÇÕES)

	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"

62% 66% 56% 89%

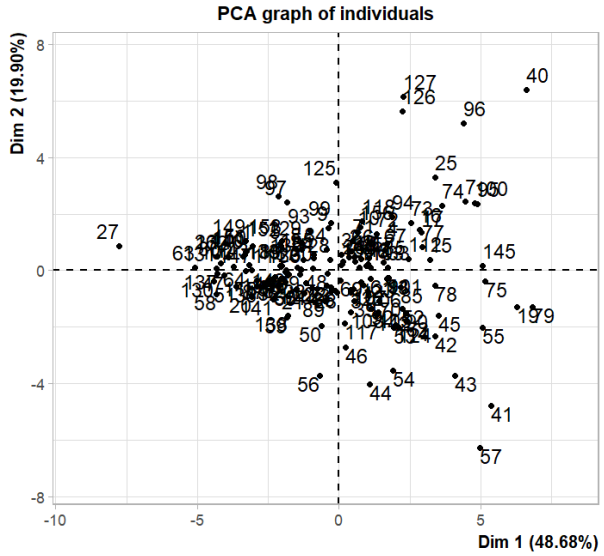
ANÁLISE MULTIVARIADA: NORMED PCA (VARIÁVEIS)

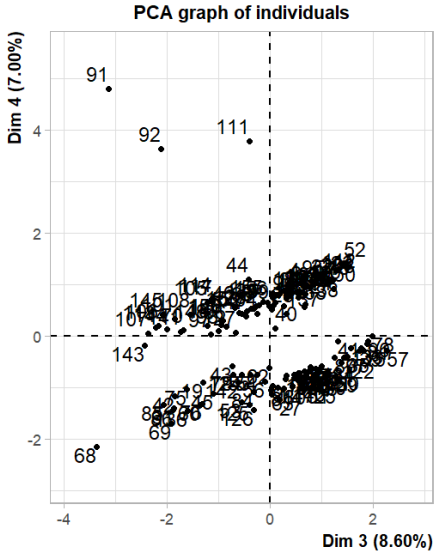




ANÁLISE MULTIVARIADA: NORMED PCA (INDIVIDUAL)

Coordenadas:





Contribuições:
Veículos com uma contribuição acima de 1.57 são significativos.
PC1 (16), PC2(15), PC3(23), PC4(4)

Cos2 bem representados acima de 80%:
PC1(20), PC3 (1: 102)
Pares:
PC1/PC2 (13), PC1/PC3(12), PC1/PC4(7), PC3/PC4(2) PC2/PC4(2)

ANÁLISE MULTIVARIADA: FACTOR ANALYSIS

Teste KMO: MAS = 0.82
"Years_Launch" removido por um valor MAS de 0.48

```
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
```

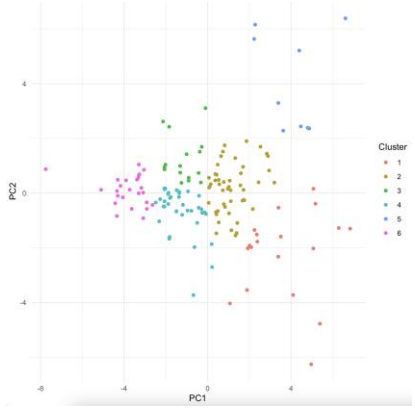
RC2: Atributos relacionados com custo e performance
RC1: Dimensões físicas do veículo
RC3: Relacionado com o tipo de veículo
RC4: Distinguir veículos atraves do Preço

Valores residuais perto do 0

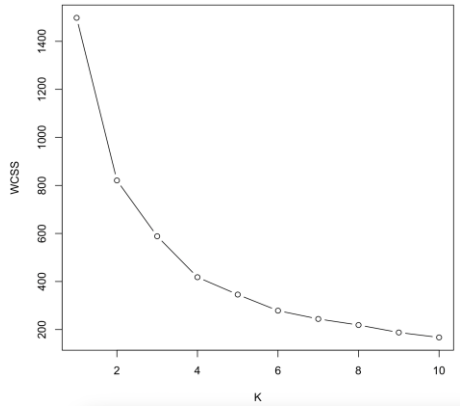
Outros Metodos de extração :

- Principal Axis
- Manual Residuals
- Maximum likelihood

ANÁLISE MULTIVARIADA: CLUSTERING ANALYSIS



Standardizar → PCA → Elbow Method → K-Means → Sillhouette Coefficient → Analisar Clusters



0.54

ANÁLISE MULTIVARIADA: LDA

Two main assumptions : Multinormality; Equal covariance matrices

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

P-values for Shapiro-Wilks are all significant with the exception of “Length”

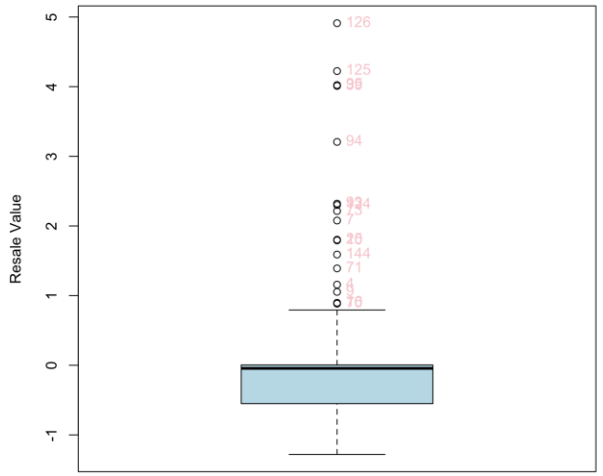
ANÁLISE MULTIVARIADA: LINEAR REGRESSION

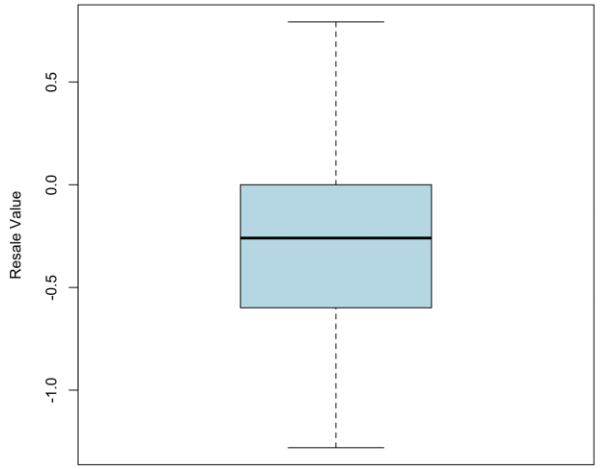
Standardizar → Remover Outliers → Split 80/20 → Testar e Avaliar

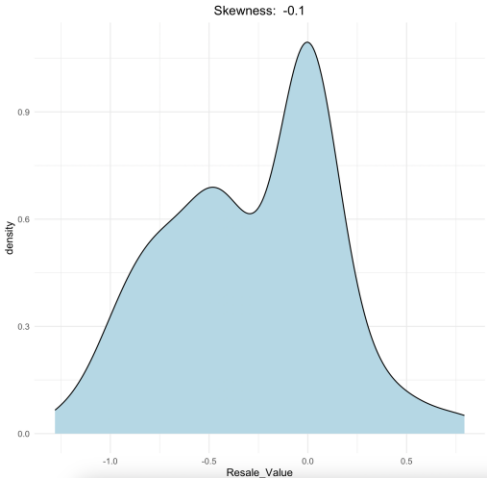
P-VALUES SIGNIFICATIVOS:
"ENGINE_SIZE"
"WIDTH"
"CURB_WEIGTH"

RMSE: 4.6

TARGET VARIABLE:
"RESALE VALUE"







CONCLUSÕES

ANÁLISE UNIVARIADA

- Perceber o comportamento e a distribiuição de cada variável

ANÁLISE BIVARIADA

- Ford foi a marca mais bem- sucessida
- Carros venderam mais do que veículos passageiros
- 7 de 155 veículos tiveram um ROI maior do que a média

ANÁLISE MULTIVARIADA

- PCA
  - Reduzido a 4 componentes com os dois primerios explicarem quase 70% da variadade
- FA
  - Identificação de uma relação entre performance e custo de um veículo
- Cluster
  - Identificação de 6 clusters
- LDA
- Linear regression
  - Variaveis com impacto no "Resale _Value"