UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Presidente Prudente

LEANDRO YUJI KOGA

ANÁLISE ESTATÍSTICA DA SATISFAÇÃO DE PROPRIETÁRIOS DE VEÍCULOS POR MEIO DO USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA

PRESIDENTE PRUDENTE
2021/2022

LEANDRO YUJI KOGA

ANÁLISE ESTATÍSTICA DA SATISFAÇÃO DE PROPRIETÁRIOS DE VEÍCULOS POR MEIO DO USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística, da Faculdade de Ciências e Tecnologia FCT/ de Presidente Prudente - UNESP, para obtenção do grau do Curso.

Orientador: Prof. Dr. Mário Hissamitsu Tarumoto

PRESIDENTE PRUDENTE
2021/2022

K78a

Koga, Leandro Yuji
Análise estatística da satisfação de proprietários de veículos por meio do uso de técnicas de análise multivariada / Leandro Yuji Koga. -- Presidente Prudente, 2022
32 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Mário Hissamitsu Tarumoto

1. Análise Multivariada. 2. Análise de Cluster. 3. Classificação de carros. 4. Web scrapping. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

LEANDRO YUJI KOGA

ANÁLISE ESTATÍSTICA DA SATISFAÇÃO DE PROPRIETÁRIOS DE VEÍCULOS POR MEIO DO USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores
Departamento de Estatística

_____

Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

Presidente Prudente, 26 de março de 2022.

RESUMO

Os carros mudaram completamente a forma como os seres humanos se locomovem e isso impacta diretamente na compra de um veículo, pois o consumidor quer que o carro reflita seus gostos e personalidade. Existe uma enorme variedade de fabricantes e cada um conta com diversos modelos. Buscar em *sites* especializados as informações a respeito não seria uma amostra aleatória, porém, existem alguns *sites* que fazem levantamento da opinião dos proprietários. Esses *sites* possibilitam a troca de opiniões, o que é positivo, e dificulta o acesso a informações, devido ao volume imensurável de dados disponíveis. Neste trabalho, a obtenção dos dados foi realizada através da técnica de *web scraping*, após houve a depuração dos dados onde foram selecionados algumas marcas e modelos para possibilitar a comparação entre eles. Após a aplicação de algumas técnicas multivariadas, foi possível a construção de classificação dos modelos de carros, a fim de estabelecer critérios que ajudem um possível comprador a escolher um fabricante e um modelo de carro, que é o intuito deste trabalho. Foram realizados estudos de formas de aquisição de dados da *internet* de maneira automatizada, utilizando-se de ferramentas computacionais, e a aplicação de técnicas estatísticas para a análise dos dados. Concluiu-se que a extração de dados via *web scrapping* se mostrou útil e rápida para a coleta de dados, facilitando o processo da aquisição das avaliações. As análises podem ser aprimoradas com um aumento na base de dados e com a inclusão de mais fabricantes para refletir a realidade.

Palavras-chave: Análise multivariada; Análise Conglomerado, Classificação de Carros; Web scraping.

ABSTRACT

Cars have completely changed the way humans get around and this directly impacts the purchase of a vehicle, as consumers want the car to reflect their tastes and personality. There is a huge variety of manufacturers and each one has different models. Searching on specialized sites for information about it would not be a random sample, however, there are some sites that survey the opinion of the owners. These sites make it possible to exchange opinions, which is positive, and make it difficult to access information, due to the immeasurable volume of data available. In this work, the data collection was performed using the web scraping technique, after the data was debugged where some brands and models were selected to enable the comparison between them. By applying some multivariate techniques, it was possible to construct a classification of car models, in order to establish criteria that help a possible buyer to choose a manufacturer and a car model, which is the purpose of this work. Studies were carried out on ways of acquiring data from the internet in an automated way, using computational tools, and the application of statistical techniques for data analysis. It was concluded that data extraction via web scrapping proved to be useful and fast for data collection, facilitating the process of acquiring evaluations. The analysis can be improved with an increase in the database and with the inclusion of more manufacturers to reflect the reality.

Keywords: Multivariate analysis; Cluster Analysis, Vehicles classification; Web scraping

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

Sumário

1. INTRODUÇÃO

Os carros mudaram completamente a forma como os seres humanos se locomovem, e como aponta Schor (1999), ligar um carro se tornou mais intuitivo do que montar em um cavalo. Isso impacta diretamente na compra de um veículo, pois o consumidor quer que o carro reflita seus gostos e personalidade, porém, a decisão de como escolher um carro para se comprar é bem complexa, pois como argumenta Porto e Torres (2012), o próprio indivíduo pode não entender exatamente a sua preferência.

Existe uma enorme de variedade de fabricantes e cada um conta com diversos modelos. Uma forma de escolha poderia ser a de ver a propaganda das marcas e modelos, no entanto, estas informações poderiam dificultar a tomada de decisão, tendo em vista que a tendência de cada marca é ressaltar somente as qualidades, e os defeitos ficarem como informações omissas. Outra forma, seria fazer uma pesquisa com o intuito de levantar informações dos proprietários dos carros, no entanto, isso seria muito caro e levaria muito tempo.

Uma possibilidade seria buscar em *sites* especializados as informações a respeito. Naturalmente esta informação não seria uma amostra aleatória, no entanto, teríamos ideia comparativa entre os donos de carros em relação à opinião a cada quesito avaliado, e, cada dono pode ter uma opinião que nem sempre é a mesma entre os donos de carros da mesma marca, modelo e ano.

Existem alguns *sites* que fazem levantamento da opinião dos proprietários, incluindo vários quesitos. Entre os *sites*, a página *https://bestcars.com.br/bc/*, reúne a opinião de donos e ex-donos de modelos de carros junto com a avaliação por meio de variáveis quantitativas em que cada participante atribui uma nota de zero a dez para cada uma delas. As variáveis constantes nesta página são: estilo, acabamento, posição de dirigir, instrumentos, itens de conveniência, espaço interno, capacidade de bagagem, motor, desempenho, consumo, câmbio, freios, suspensão, estabilidade, segurança passiva e custo-benefício. Além das informações do ano, modelo, tempo que é dono deste carro e a quilometragem.

Existem algumas variáveis qualitativas como: prós, contras, defeitos apresentados e opinião geral. Para o momento não serão estudadas estas últimas variáveis, pois, envolve outras técnicas de análise.

Sites como esse (*https://bestcars.com.br/bc/*), apesar de possibilitarem troca

de opiniões, o que é positivo, dificulta o acesso a informações, devido ao grande volume de dados disponíveis. O usuário se perde na quantidade de variáveis e se encontra sem saber como equilibrar tantas covariáveis ao mesmo tempo. Idealmente, teria um carro superior em todos os sentidos, mas não é o que ocorre, há fatores inversamente proporcionais, como será mostrado posteriormente, e, portanto, há de se flexibilizar em algumas categorias.

Outro passo importante em qualquer trabalho estatístico é a obtenção dos dados. Neste caso, e extração será automatizada, pois a obtenção dos dados manualmente é trabalhosa e demorada, além do risco de erro de digitação. Uma alternativa viável é a técnica chamada de *web scraping*, que consiste na extração automatizada de dados de *sites* da *internet*, pois a captação é mais rápida e confiável (PONTOLIO, 2014).

Após a obtenção e depuração dos dados, considerando que o interesse neste trabalho é a possível escolha de um carro, foram selecionados algumas marcas e modelos para possibilitar a comparação entre eles. Entre as várias possíveis técnicas estatísticas para a análise destes dados, foram aplicadas algumas técnicas multivariadas. Segundo Hair, et al. (2009, p.), podemos definir análise multivariada como “todas as técnicas estatísticas que simultaneamente analisam múltiplas medidas sobre indivíduos ou objetos sob investigação”.

Por exemplo, para o processo de classificação, para ver qual quesito foi o mais importante para a escolha de um determinado carro, para cada modelo, poderá ser aplicado o modelo de regressão multivariado. Com base em algumas destas técnicas de análise multivariada, foi possível a construção de classificação dos modelos de carros, a fim de estabelecer critérios que ajudem um possível comprador a escolher um fabricante e um modelo de carro.

Portanto, este trabalho de conclusão de curso, teve como intuito realizar estudos de formas de aquisição de dados da *internet* de forma automatizada, utilizando-se de ferramentas computacionais, além do estudo e aplicação de técnicas estatísticas para a análise dos dados.

A estrutura deste trabalho está disposta da seguinte forma: Introdução, que descreve a toda a parte inicial, a temática abordada; o Desenvolvimento que consta de três capítulos, onde o primeiro discorre sobre a aquisição dos dados; o segundo capítulo descreve a análise fatorial dos dados coletados, quais as técnicas que foram utilizadas para o desenvolvimento, e, o terceiro capítulo enfatiza a aplicação, a

análise exploratória de alguns testes estatísticos realizados. Após do Desenvolvimento do trabalho foi realizada as Considerações Finais.

2. DESENVOLVIMENTO

Para o desenvolvimento deste trabalho foi realizado uma breve descrição das atividades desenvolvidas, dividindo-se entre a parte teórica e prática. Na parte teórica foi realizada descrição breve da teoria estudada; na parte prática foi realizada a obtenção de dados, de forma automatizada, utilizando o procedimento conhecido como *Webscrapping.*

2.1 Aquisição de dados

A primeira etapa para a realização deste trabalho foi o estudo das formas de aquisição automática de dados de *internet.* Esses dados estão disponíveis de forma pública, que, após serem coletados, extraídos da *internet*, foram trabalhados.

Entender, mesmo que superficialmente, como as páginas de *internet* são construídas se faz necessário para dar continuidade neste trabalho. Cada página tem um endereço virtual chamado de *URL*, que vem do inglês *Uniform Resource Locator*, traduzido seria um Localizador Uniforme de Recursos.

Muitas páginas utilizam um código chamado *Hyper Text Markup Language* (*HTML*) que serve para estruturar os parágrafos e documentos. O HTML não chega a ser uma linguagem de programação, visto que serve apenas como uma maneira de escrever um conteúdo para a *internet*. Porém, isso não significa que não seja importante, ao contrário, é extremamente utilizada, devido a sua funcionalidade e facilidade de acesso. A maioria das páginas utilizam o *HTML* e quando necessário acrescentam uma linguagem de programação, como o *PHP* ou *JavaScript*, mas não é o caso deste trabalho.

Figura 1 – Exemplo de um código-fonte em HTML

Fonte: código-fonte do *site Bestcars*

A página do *site Bestcars* é composta por *tags* (Figura 1), que são delimitadas pelos sinais < e >. A *tag div* serve para definir uma parte da página, com o objetivo de facilitar a organização. Os textos que serão vistos pelos usuários estão entre os seguintes comandos, - para indicar o início e, - para indicar o fim.

O entendimento de *HTML* já é suficiente para extrair os dados. Essa extração é chamada de *webscrapping*. Todas as páginas são escritas em um texto chamado código-fonte, e, a maioria está disponível para os usuários.

Na Figura 1, no código-fonte, percebe-se que as notas estão dentro da classe "*content-inner*", essa informação é extremamente útil para localizar os dados de interesse. Visto o código-fonte, o *software R* foi utilizado para a captação das avaliações dos carros. Nesse *software* há vários pacotes feitos para auxiliar a *webscrapping*, como o "*rvest*", que dado a *URL* do *site,* coleta as informações.

Após a extração dos dados, outro pacote foi utilizado, o "*tokenizers*", que transforma o código-fonte da página em uma lista em que cada entrada é uma palavra do código-fonte. A partir dessa lista, basta isolar as notas dos usuários em uma matriz. Enfim, os dados estão coletados e prontos para serem analisados estatisticamente.

Na Figura 2 é apresentado, como exemplo, a captação dos dados do modelo do carro HB20, da marca *Hyundai*.

Figura 2 - Exemplo do modelo de carro HB20, marca Hyundai

```
> url <- 'https://bestcars.com.br/bc/participe/opinioes-proprietarios/teste-do-leitor-hyundai/hyundai-hb20/'
> webpage <- read_html(url)
> rank<- html_nodes(webpage,'.content-inner')
> data <- html_text(rank)
> words <- tokenize_words(data)
> words
[[1]]
   [1] "nome"          "eduardo"      "lima"          "cidade"       "cotia"
   [6] "estado"        "sp"           "modelo"        "hyundai"      "hb20"
  [11] "versão"        "x"            "confort"       "style"        "motor"
  [16] "1,6"           "16v"          "ano"           "modelo"       "2014"
  [21] "quilometragem" "atual"        "70.000"        "km"           "combustível"
  [26] "gasolina"      "tempo"        "há"            "que"          "possui"
```

Fonte: elaborado pelo autor (2022)

2.2 Breve descrição da teoria: análise multivariada

Após a aquisição e depuração dos dados, percebeu-se que há dados faltantes. Há avaliações que não estão completas, o usuário não preencheu todas as variáveis por motivos desconhecidos, porém há uma quantidade pequena de dados ausentes, desta forma, os indivíduos que não preencheram todos os campos serão retirados da análise neste momento, pois a imputação de dados não faz parte do escopo desse trabalho.

Para a análise e interpretação das 16 variáveis quantitativas serão utilizadas técnicas de análise multivariada.

Podemos citar a análise fatorial e a de componentes principais para simplificar os dados e facilitar a interpretação, pois não é recomendado por Possoli (2012) usar uma grande quantidade de variáveis.

2.2.1 Componentes principais

A análise de componentes principais tem como objetivo reduzir a quantidade de covariáveis, criando componentes que sejam combinações lineares das covariáveis iniciais, dessa forma a dimensionalidade dos dados é diminuída e ainda pode-se julgar a importância das variáveis originais (NETO; MOITA, 1998). Essa técnica não requer normalidade e depende apenas da matriz de covariâncias ou correlações. Considerando $X_1, X_2 \ldots, X_p$ um conjunto de p variáveis aleatórias e X' o vetor $\{X_1, X_2, \ldots, X_P\}$ cuja matriz de covariâncias é representada por $\sum$, em que os p autovalores λ seguem a ordem $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. As combinações lineares são dadas por:

$$Y_1 = a'_1\boldsymbol{X}$$
$$Y_2 = a'_2\boldsymbol{X}$$
$$.$$
$$.$$
$$.$$
$$Y_{\boldsymbol{p}} = a'_p\boldsymbol{X}$$

Para encontrar os vetores, algumas suposições foram feitas. Os componentes principais devem ser não correlacionados, a variância de $a_i\boldsymbol{X}$ deve estar maximizada para *i = 1,2,..., p*, o tamanho do vetor *a* deve ser unitário e $cov(a_i\boldsymbol{X}, a_j\boldsymbol{X}) = 0$, para $i \neq j$.

$$\max_{a\neq 0}\frac{a'\sum a}{a'a} \text{ é equivalente a maximizar } a'\sum a, \text{ pois } a'a = 1$$

Considerando os *p* pares (λ_i, e_i) como o *i-ésimo* autovalor e autovetor da matriz de covariâncias, $\sum$. Conclui-se que:

$$\max_{a\neq 0}\frac{a_i'\sum a_i}{a_i'a_i} = \lambda_i$$

Portanto, pode-se escrever os componentes principais como:

$$Y_1 = e'_1\boldsymbol{X}$$
$$Y_2 = e'_2\boldsymbol{X}$$
$$.$$
$$.$$
$$.$$
$$Y_{\boldsymbol{p}} = e'_p\boldsymbol{X}$$

Denotando por σ_{11}, σ_{22}, ..., σ_{pp} as variâncias das *p* variáveis, que são os elementos da diagonal principal da matriz $\sum$, podemos escrever a seguinte igualdade:

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} var(X_i) = traço(\sum) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} var(Y_i).$$

Consequentemente, a proporção da variância populacional explicada pelo *k-ésimo* componente principal é:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, k = 1, 2, \ldots, p.$$

Para auxiliar a interpretação dos componentes principais, calcula-se a correlação entre o *k-ésimo* componente e *a i-ésima* variável pela seguinte fórmula:

$$\rho_{Yk,Xi} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}.$$

O $\rho_{Yk,Xi}$ mostra o quanto a variável X_i contribui para o componente Y_k.

Não há um método exato para determinar a quantidade de componentes principais a ser utilizada, mas, uma ferramenta útil é a utilização de um tipo de gráfico chamado de "*scree plot*" onde, geralmente, fica claro quantos componentes utilizar, visto que se pretende escolher poucos componentes, mas que expliquem grande parte da variabilidade dos dados.

No eixo horizontal temos a quantidade de componentes *e* no eixo vertical a proporção da variância explicado. Procura-se no *screeplot* um ponto de mudança brusca, chamado de ponto de cotovelo.

2.2.2 Análise fatorial

A análise fatorial foi desenvolvida por volta de 1900 por Francis Galton e Charles Spearman (POSSOLI, 2012) para estudar a inteligência. O objetivo era entender que fatores influenciavam a inteligência humana. Essa técnica consiste em agrupar covariáveis correlacionadas em fatores que explicam a variância dos dados. A análise fatorial pode ser vista como uma extensão dos componentes principais. O modelo estatístico da análise fatorial ortogonal é o seguinte:

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1p}F_p + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2p}F_p + \varepsilon_2$$
$$\vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pp}F_p + \varepsilon_p$$

A variável *p* representa a quantidade de variáveis; *m* representa a quantidade de fatores, e, *X* é um vetor aleatório de dimensão *(px1); μ* é o vetor de médias de *X* de dimensão *(px1);* $l_{11}, l_{12}, \ldots, l_{pp}$ são coeficientes chamados de carga ou peso dos fatores; $F_1, F_2, \ldots, F_p$ são vetores aleatórios de dimensão *(mx1)* chamados de fatores comuns e, por último, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ são os erros de dimensão *(px1)*, também chamados de fatores específicos.

Em notação matricial, o modelo pode ser escrito como:

$$X - \mu = LF + \varepsilon.$$

Algumas premissas são feitas para o desenvolvimento da análise fatorial.

$$E[F] = 0_{mx1}$$
$$cov(F) = E[FF'] = I_{mxm}$$
$$E[\varepsilon] = 0_{px1}$$
$$cov(\varepsilon) = E[\varepsilon\varepsilon'] = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix},$$

Os elementos $\psi_1, \psi_2, \ldots, \psi_p$ são chamados de variância específica. Além disso ε e F são independentes, ou seja, $\mathrm{cov}(\varepsilon, \mathrm{F}) = 0_{\mathrm{pxm}}$.

A proporção de variabilidade de cada variável que é explicada pelos fatores é chamada de comunalidade, representada por *h²*.

$$h_i{}^2 = l^2{}_{i1} + l^2{}_{i1} + l^2{}_{i1}$$

Uma maneira de estimar a matriz *L* é pelo método de componentes principais. Decompõe-se espectralmente a matriz $\sum$.

$$\sum = \left[\sqrt{\lambda_1}e_1 \middle| \sqrt{\lambda_2}e_2 \middle| \ \ldots \ \sqrt{\lambda_p}e_p\right] \begin{bmatrix} \sqrt{\lambda_1}e_1 \\ \sqrt{\lambda_2}e_2 \\ . \\ . \\ . \\ \sqrt{\lambda_p}e_p \end{bmatrix}$$

Permitindo uma aproximação, podemos escrever:

$$\sum = \left[\sqrt{\lambda_1}e_1 \middle| \sqrt{\lambda_2}e_2 \middle| \ \ldots \ \sqrt{\lambda_m}e_m\right] \begin{bmatrix} \sqrt{\lambda_1}e_1 \\ \sqrt{\lambda_2}e_2 \\ . \\ . \\ . \\ \sqrt{\lambda_m}e_m \end{bmatrix},$$

em que m é a quantidade de fatores e $m \leq p$.

A matriz de variâncias específicas pode ser obtida por:

$$\boldsymbol{\psi} = \sum - \boldsymbol{LL'}$$

A proporção da variância explicada pelos fatores pode ser calculada da seguinte forma:

$$proporção = \begin{cases} \dfrac{\lambda_j}{s_{11} + s_{22} + \cdots + s_{pp}}, \text{se for utilizado a matriz de covariâncias} \\ \dfrac{\lambda_j}{p}, \text{ se for utilizado a matriz de correlações} \end{cases}$$

Para estimar os fatores comuns, se tiverem sido calculados pelo método de componentes principais, é aconselhado por Johnson, Wichern (2007) a utilizar o método dos mínimos quadrados ordinários.

$$\hat{f_j} = (LL')^{-1}L(x_j - \bar{x})$$

Após estimar os fatores, a interpretação nem sempre é óbvia. Para ajudar, pode-se aplicar uma rotação, a fim de simplificar a estrutura dos dados e facilitar a interpretação.

2.2.3 Análise de conglomerado

Outra técnica relevante é a análise de conglomerado ou *clusters*. Técnica interesse que agrupa veículos semelhantes em grupos de modo que os indivíduos no mesmo *cluster* sejam parecidos entre si e diferentes dos indivíduos de outros *clusters* (HAIR, et al., 2009). Assim, pode-se apresentar opções para o comprador. Primeiro, temos que definir uma distância, que é uma medida sempre maior ou igual a 0 e só se mede 0 *se e* somente os dois elementos são iguais. E, a partir da distância entre os elementos, constrói-se uma matriz que contém todas as distâncias, essa matriz é chamada de matriz de parecença.

Sejam *x* e *y* dois vetores de *p* dimensões, define-se a distância euclidiana como:

$$d(x, y) = \sqrt{(x - y)'(x - y)}$$

Há dois tipos de agrupamento, os modelos hierárquicos e os não hierárquicos. A diferença entre eles é que os modelos hierárquicos são inflexíveis, isto é, se um item for incluído em um *cluster*, ele não pode ser retirado. Outra diferença importante, é que os métodos não hierárquicos dependem de uma partição

inicial que servirão de centros para a construção de centroides.

Os hierárquicos começam com *N* agrupamentos, um para cada item. A partir disso, é calculada uma matriz contendo as distâncias de todos os elementos, dois a dois. Depois o par que tiver a menor distância é agrupado. Em seguida, a matriz de distância é recalculada, considerando o par do passo passado. Repetindo esse processo até todos os itens estarem em um único grupo.

Johnson e Wichern (2007) propõe que para minimizar a perda de informação pode-se utilizar o método de *Ward*, que minimiza a Soma dos Erros ao Quadrado (ESS). Segundo Hair, et al. (2009, p.), o critério de agrupamento é “a soma de quadrados entre os dois agrupamentos somados sobre todas as variáveis”. No estágio final, em que todos os itens estão no mesmo grupo, o valor do ESS é:

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x})'(x_j - \bar{x})$$

Os estágios de agrupamento podem ser apresentados em um dendrograma, que mostra as junções em cada passo.

Johnson e Wichern (2007) apresentam outra técnica chamada de *K-means*, um procedimento não hierárquico, que inicia com uma quantidade arbitrária de *K* partições. Em seguida, deve-se construir *K* centroides e cada elemento pertencerá ao centroide em que estiver mais perto. Cada vez que um centroide receber ou perder um elemento, ele deverá ser recalculado. Esse processo deverá ser repetido até não haver mais atualizações. É recomendável rodar o algoritmo com diversas quantidades de partições iniciais.

3. APLICAÇÃO

Após a coleta de dados, foram obtidas 190 críticas. Porém, várias eram pertinentes ao mesmo modelo. Então, as avaliações de carros do mesmo ano e mesmo motor foram agrupados, a fim de que cada linha na matriz de dados representasse um modelo único. Para essa junção de avaliações, foi utilizada a média das avaliações em cada variável. O resultado pode ser visto no Quadro 1, onde disponibilizadas avaliações de dois veículos, como ilustração:

Quadro 1 - - Exemplo de resultados de dois modelos de carros

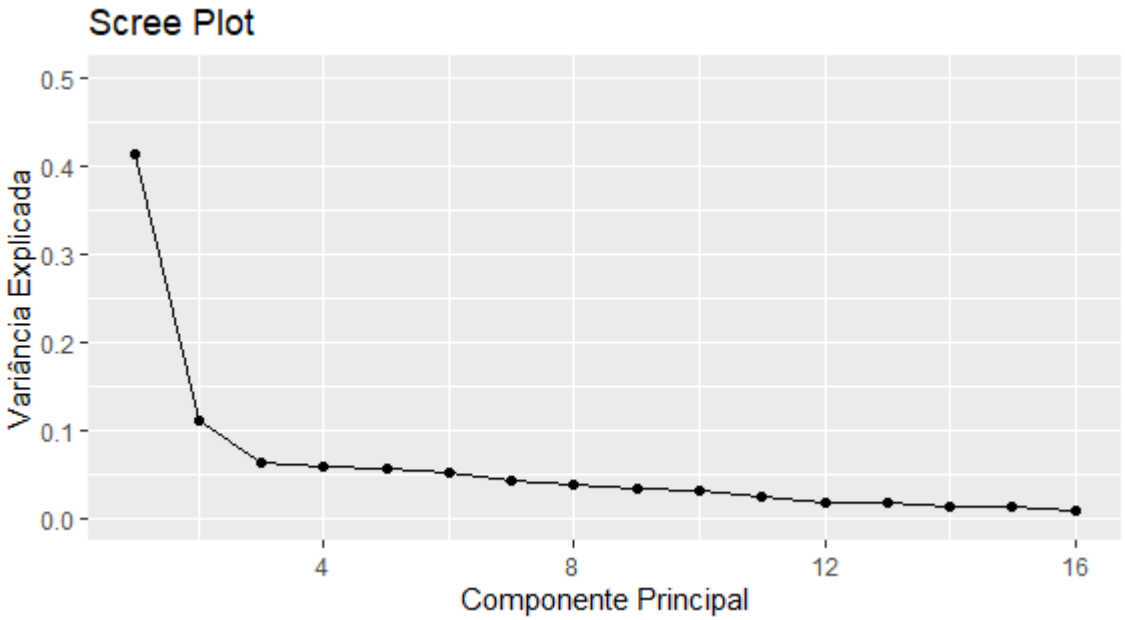
FABRICANTE	Ford	Fiat
MODELO	Ka	Strad
PATÊ	1.0	1.4
ANO	2000	2010
ESTILO	2,5	5,0
ACABAMENTO	2,0	3,5
DIRIGIR	4,5	4,0
INSTRUMENTOS	2,5	4,0
CONVENIÊNCIA	2,0	2,5
INTERNO	2,5	4,0
BAGAGEM	1,5	4,5
MOTOR	3,5	5,0
DESEMPENHO	3,5	4,0
CONSUMO	5,0	4,5
CÂMBIO	3,0	4,0
FREIOS	3,0	4,5
SUSPENSÃO	2,5	5,0
ESTABILIDADE	4,0	3,5
PASSIVA	3,5	3,5
BENENEFÍCIO	5,0	4,0

Fonte: elaborado pelo autor, 2022

As variâncias e covariâncias foram calculadas e seus autovalores e

autovetores, para aplicar a técnica de análise de componentes principais. O gráfico de *scree plot* (Figura 3) foi construído e percebeu-se que o ponto de cotovelo indica que os dois primeiros componentes principais são indicados para o modelo final.

Figura 3 - Gráfico Scree Plot



Fonte: elaborado pelo autor, 2022.

Os coeficientes dos componentes principais são obtidos como os autovetores da matriz das variâncias e cada componente pode ser escrito como:

$$
\begin{aligned}
Y_1 = \; & -0.16Estilo - 0.28Acabamento - 0.29Dirigir - 0.29Instrumentos \\
& - 0.29Conveniência - 0.32Interno - 0.16Bagagem - 0.28Motor \\
& - 0.23Desempenho - 0.18Consumo - 0.19Câmbio - 0.23Freios \\
& - 0.27Suspensão - 0.28Estabilidade - 0.27Passiva \\
& - 0.20Benefício \\
Y_2 = \; & -0.05Estilo - 0.16Acabamento - 0.11Dirigir - 0.21Instrumentos \\
& - 0.21Conveniência - 0.23Interno - 0.54Bagagem + 0.31Motor \\
& + 0.31Desempenho + 0.41Consumo + 0.24Câmbio - 0.01Freios \\
& + 0.00Suspensão + 0.05Estabilidade + 0.06Passiva \\
& + 0.32Benefício
\end{aligned}
$$

Johnson e Wichern (2007) apresentam um critério de escolha que consiste

em escolher os componentes que têm autovalor maior ou igual a *um*. Tanto o critério do ponto do cotovelo quando o do valor dos autovalores maiores do que *um* sugerem escolher dois componentes principais. Os autovalores e variância explicada apresentados na Tabela 1.

Tabela 1 - Principais Resultados dos Componentes Principais

Componentes	Autovalor	Proporção explicada	Proporção acumulada
Componente 1	4,61	0,4	0,41
Componente 2	1,25	0,11	0,53
Componente 3	0,71	0,06	0,59
Componente 4	0,66	0,06	0,65
Componente 5	0,64	0,05	0,70
Componente 6	0,58	0,05	0,75
Componente 7	0,48	0,04	0,80
Componente 8	0,43	0,03	0,83
Componente 9	0,37	0,03	0,87
Componente 10	0,35	0,03	0,91
Componente 11	0,27	0,02	0,94
Componente 12	0,21	0,01	0,96
Componente 13	0,19	0,01	0,97
Componente 14	0,16	0,01	0,98
Componente 15	0,14	0,01	0,99
Componente 16	0,09	0,01	1,00

Fonte: elaborado pelo autor, 2022.

Os componentes Y_1 e Y_2 são combinações lineares das variáveis originais e explicam cerca de 53% da variância dos dados. Apesar do valor parecer baixo, o acréscimo de componentes adicionais não traz uma diferença significativa. O terceiro componente acrescentaria apenas 6%.

Para a análise fatorial foi construído um quadro (Tabela 2) para comparar os dois métodos de estimação apresentados anteriormente.

Tabela 2 - Quadro comparativo dos fatores

	COMPONENTES PRINCIPAIS			MÁXIMA VEROSSIMILHANÇA		
	Fatores estimados		Variância específica	Fatores estimados		Variância específica
VARIÁVEIS	F_1	F_2		F_1	F_2	
Estilo	0.34	-0.05	0.12	0.34	0.05	0.12
Acabamento	0.58	-0.18	0.36	0.55	0.28	0.38
Dirigir	0.57	-0.12	0.34	0.53	0.27	0.36
Instrumentos	0.60	-0.21	0.41	0.57	0.28	0.40
Conveniência	0.62	-0.24	0.44	0.58	0.30	0.45
Interno	0.65	-0.22	0.47	0.60	0.30	0.45
Bagagem	0.31	-0.38	0.24	0.26	0.34	0.18
Motor	0.58	0.32	0.44	0.66	-0.33	0.55
Desempenho	0.48	0.35	0.36	0.56	-0.36	0.44
Consumo	0.37	0.35	0.26	0.40	-0.22	0.21
Câmbio	0.39	0.21	0.20	0.41	-0.12	0.18
Freios	0.49	-0.01	0.24	0.47	0.10	0.30
Suspensão	0.55	-0.01	0.30	0.54	0.10	0.30
Estabilidade	0.58	0.02	0.34	0.56	0.13	0.33
Passiva	0.55	0.04	0.30	0.53	0.12	0.30
Benefício	0.42	0.33	0.28	0.43	-0.14	0.20

Fonte: elaborado pelo autor, 2022

Observando a Tabela 2, não foi possível a realização de uma interpretação simples, por isso, optou-se por uma rotação dos fatores. Estes resultados são apresentados na Tabela 3.

Tabela 3 - Quadro dos fatores rotacionados

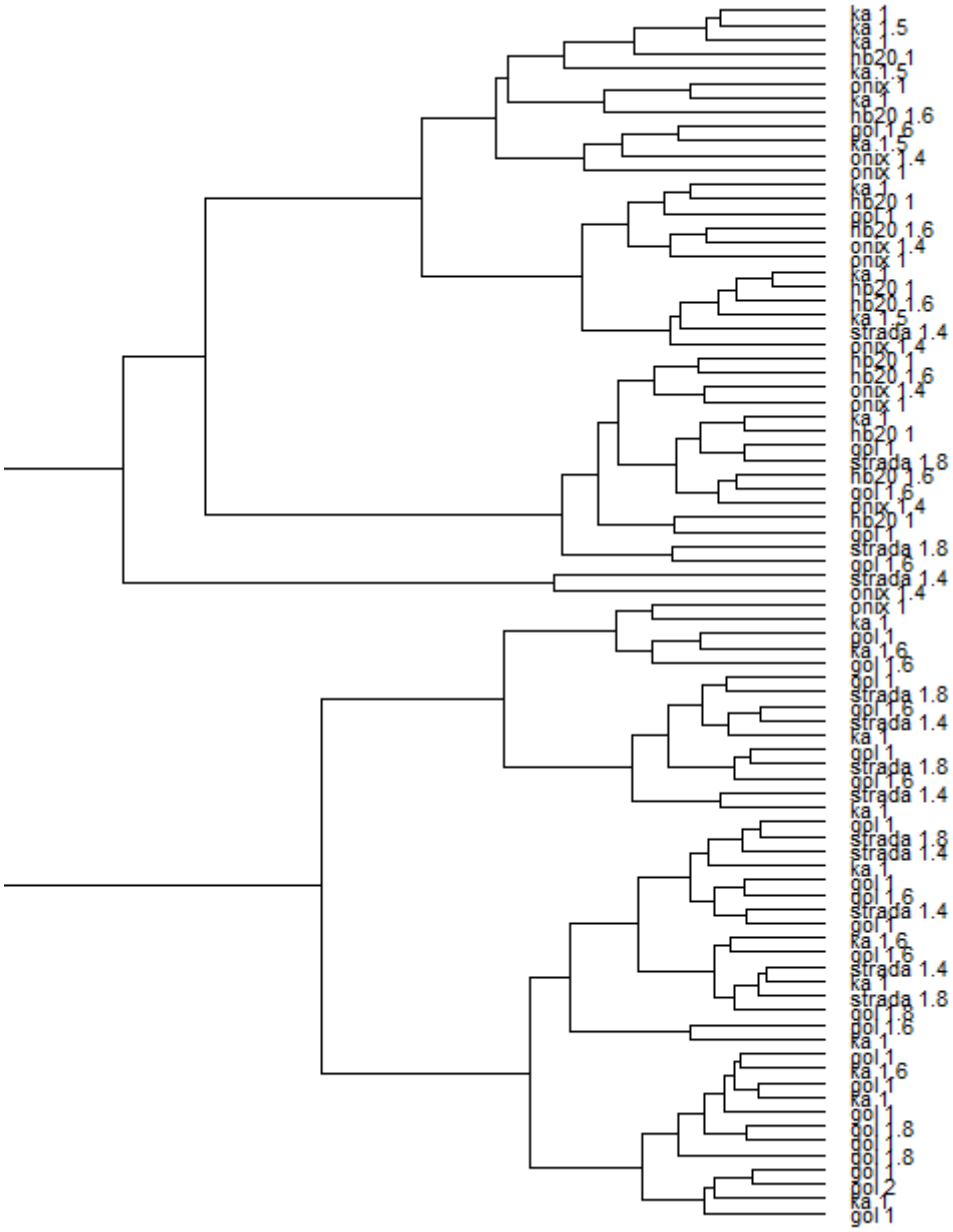
	COMPONENTES PRINCIPAIS (ROTAÇÃO)			MÁXIMA VEROSSIMILHANÇA (ROTAÇÃO)		
	Fatores estimados		Variância específica	Fatores estimados		Variância específica
VARIÁVEIS	F_1	F_2		F_1	F_2	
Estilo	0.29	0.08	0.12	0.25	0.13	0.12
Acabamento	0.58	0.04	0.36	0.62	-0.01	0.38
Dirigir	0.52	0.11	0.34	0.60	0.00	0.36
Instrumentos	0.63	0.01	0.41	0.63	0.01	0.40
Conveniência	0.67	-0.01	0.44	0.66	-0.01	0.40
Interno	0.67	0.03	0.47	0.67	0.00	0.45
Bagagem	0.57	-0.29	0.24	0.51	-0.23	0.18
Motor	0.13	0.59	0.44	0.05	0.71	0.55
Desempenho	0.03	0.58	0.36	-0.03	0.68	0.44
Consumo	-0.06	0.54	0.26	0.01	0.45	0.21
Câmbio	0.10	0.39	0.20	0.12	0.34	0.18
Freios	0.37	0.19	0.24	0.39	0.14	0.23
Suspensão	0.4	0.22	0.30	0.42	0.19	0.30
Estabilidade	0.40	0.26	0.34	0.47	0.16	0.33
Passiva	0.37	0.26	0.30	0.44	0.16	0.33
Benefício	0.00	0.53	0.28	0.11	0.38	0.20

Fonte: elaborado pelo autor, 2022

Com os fatores rotacionados, pode-se perceber que o primeiro fator é responsável principalmente pelas variáveis: estilo, acabamento, dirigir, instrumentos, conveniência, interno, bagagem, freios, suspensão, estabilidade e passiva. Essas variáveis estão mais relacionadas às questões estéticas e de conforto. Enquanto o segundo fator é significativo nas variáveis: motor, desempenho, consumo, câmbio e benefício. Essas variáveis são mais técnicas. Pode-se então reduzir as 16 variáveis originais em dois fatores.

Para a análise de *cluster*, aplicou-se um modelo hierárquico e um não hierárquico. O modelo hierárquico escolhido foi o modelo de *Ward*. O resultado pode ser visto no dendrograma apresentado na Figura 4.

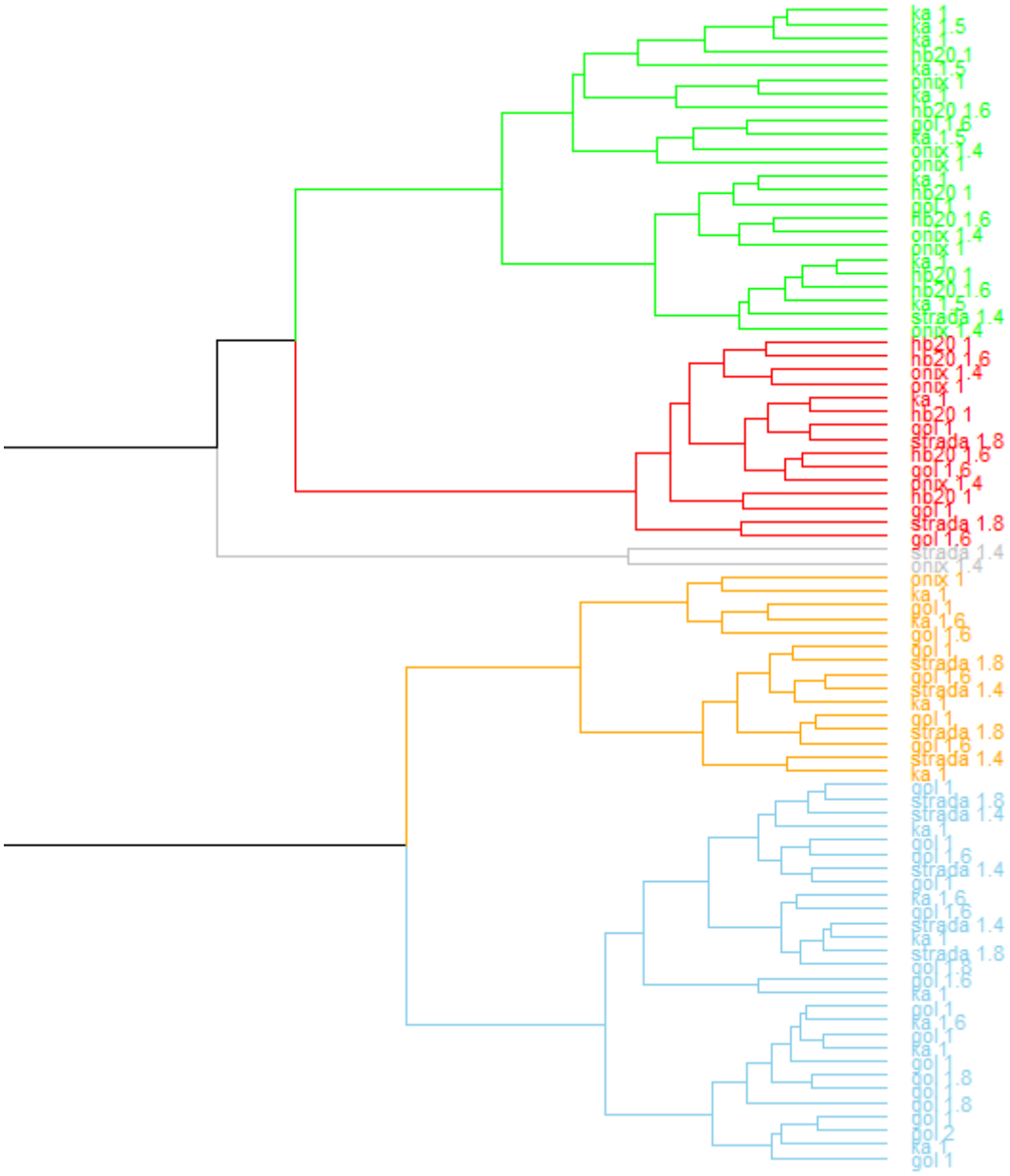
Figura 4 – Dendrograma



Fonte: elaborado pelo autor, 2022

Escolhendo uma quantidade arbitrária de *cluster*, por exemplo dividindo se amostra em cinco, obteve-se o dendograma apresentado na Figura 5.

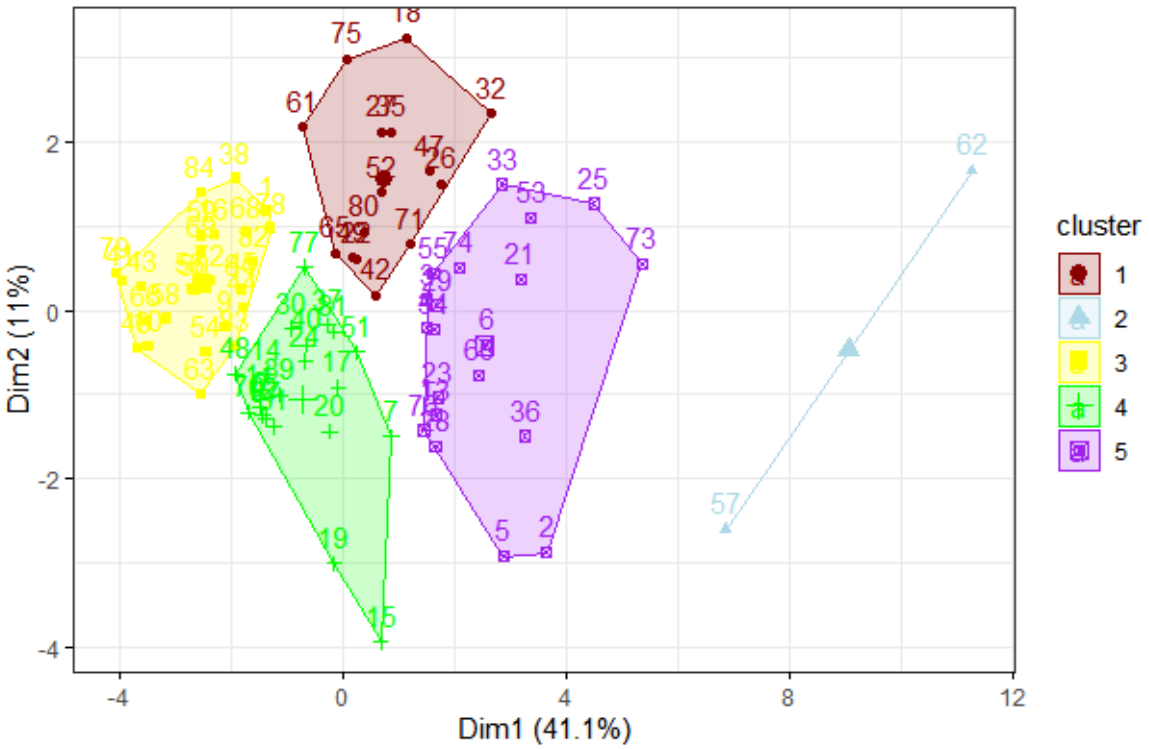
Figura 5 - Dendrograma com separação dos clusters



Fonte: elaborado pelo autor, 2022

Percebe-se que um *cluster* tem apenas dois veículos, 57 e 62, que são o Onix de 2015 e Strada de 2016, respectivamente. Algo similar acontece aplicando um modelo não hierárquico, como o *k-means,* apresentado na Figura 6.

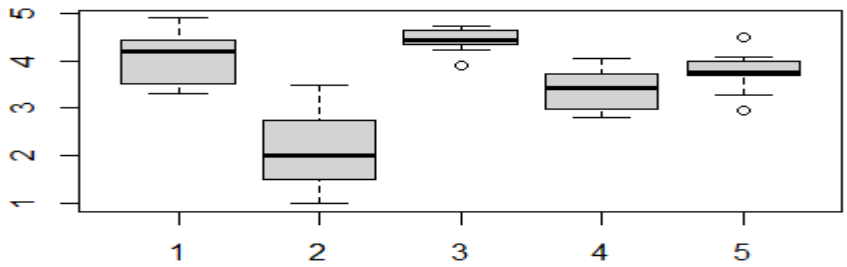
Figura 6 - Gráfico dos clusters k-means



Fonte: elaborado pelo autor, 2022

Ao analisar as médias em cada variável por *cluster*, percebe-se que o terceiro *cluster* apresenta os melhores resultados, pois tem boas notas em todas as variáveis, enquanto o *cluster* dois é o pior. O primeiro *cluster* apresenta uma alta variabilidade nas médias, o que o torna pouco confiável. Estes resultados são apresentados na Figura 7.

Figura 7 - Box-plot das médias



Fonte: elaborado pelo autor, 2022

Na Tabela 4 são apresentadas as estatísticas descritivas por cluster, podendo-se observar os resultados numéricos apresentados no Gráfico 2. Percebe-

se que o terceiro cluster apresenta os melhores resultados.

Tabela 4 - Estatisticas descritivas dos Clusters

	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Cluster1	3,30	3,53	4,19	4,07	4,43	4,88
Cluster2	1,00	1,50	2,00	2,09	2,62	3,50
Cluster3	3,88	4,36	4,43	4,45	4,63	4,73
Cluster4	2,81	2,98	3,44	3,40	3,71	4,05
Cluster5	2,96	3,69	3,75	3,79	3,99	4,59

Fonte: elaborado pelo autor,2022

Para a análise por fabricante e cluster, a distribuição de frequências são apresentados na Tabela 5.

Tabela 5 - Distribuição de frequências dos fabricantes por cluster k-means

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Chevrolet	1	1	6	2	1
Fiat	2	1	2	1	7
Ford	4	0	9	5	3
Hyundai	2	0	3	2	4
Volkswagen	12	0	6	10	0

Fonte: elaborado pelo autor,2022

Pecebe-se que a Chevrolet apresenta mais de 50% dos seus modelos no *cluster* três (3). Enquanto a Fiat tem 55% dos seus modelos no *cluster* cionco (5). A Ford apresenta cerca de 43% dos seus carros no *cluster* três (3). A Hyundai está dividida entre os *clusters* um, três, quarto e cinco. A Volkswagen entre os *cluster*s um, três e quatro.

Por outro lado, observa-se que o *cluster* um (1) é composto principlamente pela Volkswagen, o *cluster* três (3) pela Chevrolet, Ford e Hyundai; o *cluster* quatro (4) pela Volkswagen e o *cluster* cinco (5) pela Fiat, Ford e Hyundai.

4. CONSIDERAÇÕES FINAIS

Os objetivos dos trabalhos foram atingidos com sucesso. A extração de dados via *web scrapping* se mostrou útil e rápida para a coleta de dados, facilitando o processo da aquisição das avaliações.

As técnicas de componentes principais e análise fatorial conseguiram reduzir as 16 variáveis para apenas dois componentes e dois fatores. A análise fatorial foi além e ajudou a entender que um fator está relacionado a uma questão de conforto enquanto o outro a um quesito mais técnico, de custo-benefício.

Com a análise de agrupamentos foi possível categorizar os veículos em cinco *clusters*, alguns melhores do que os outros, permitindo assim, recomendar carros, principalmente, do terceiro agrupamento.

As análises podem ser aprimoradas com um aumento na base de dados e com a inclusão de mais fabricantes para refletir a realidade.

REFERÊNCIAS

HAIR; et al. *Análise multivariada de dados.* 6ed, Porto Alegre: Bookman, 2009.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis.* 6ed. Upper Sadde River: Pearson Prentice Hall, 2007.

NETO, José M.M.; MOITA, Graziella C. *Uma introdução à análise exploratória de dados multivariados.* Teresina, 1998. Disponível em: <https://www.scielo.br/pdf/qn/v21n4/3193.pdf> Acesso em: maio de 2021.

SCHOR, Tatiana. *O automóvel e o desgaste social.* São Paulo, 1999. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-88391999000300014> Acesso em: maio de 2021.

PORTO, Rafael B.; TORRES, Claudio V. *Comparações entre preferência e posse de carro:* predições dos valores humanos, atributos do produto e variáveis sociodemográficas. São Paulo, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-21072012000100011.> Acesso em: maio de 2021.

POSSOLI, Silvio. *Técnicas de análise multivariada para avaliação das condições de saúde dos municípios do Rio Grande do Sul*, *Brasil*. Rio Grande do Sul, 2012. Disponível em: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89101984000400004. > Acesso em: maio de 2021.

PONTOLIO, Luan S. *Plataforma de extração e recuperação de dados na Web no contexto de Big Data.* Marília, 2014. Disponível em: <https://aberto.univem.edu.br/bitstream/handle/11077/1006/Luan%20Silveira%20Pontolio.pdf?sequence=1&isAllowed=y.> Acesso em: maio de 2021.

SCREE PLOT. In: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation, 2010]. Disponível em: < https://en.wikipedia.org/w/index.php?title=Scree_plot&oldid=949546588.> Acesso: 22 setembro de 2021.

SITE. *Auto livraria:* best cars. Disponível em: <*https://bestcars.com.br/bc/*> Acesso: janeiro 2022.

BIBLIOGRAFIA CONSULTADA

ANDERSON, T. W. *An introduction to multivariate statistical analysis*. 2ed. New York: John Wiley & Sons, 1984.

APOSTILA DE NORMATIZAÇÃO DOCUMENTÁRIA: com base nas normas da ABNT. Org. DEGASPARI, S.D.; VANALLI, T.R.; MOREIRA, M.R.G. Presidente Prudente, UNESP, 2021.

BARROSO, L. P.; ARTES, R. *Análise Multivariada*. Lavras: UFLA, 2003.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. *Multivariate analysis.* London: Academic Press, 1992.

MIGUEZ, F. *Introduction to R for Multivariate.* Data Analysis, 2007. Disponível em: <http://miguezlab.agron.iastate.edu/OldWebsite/Teaching/MultivariateRGGobi.pdf.> Acesso em: maio de 2021.

MINGOTI, S.A. *Análise de dados através de métodos de estatística multivariada:* uma abordagem aplicada. Belo Horizonte: UFMG, 2005.

RAM, Nilam. *Cluster Analysis*: an example. Disponível em: <https://quantdev.ssri.psu.edu/tutorials/cluster-analysis-example.> Acesso em: setembro de 2021.

ZELTERMAN, Daniel. *Applied multivariate statistics with R.* Springer, 2015. Disponível em: <https://web.uniroma1.it/memotef/sites/default/files/file%20lezioni/102b_textbook.pdF > Acesso em: maio de 2021.