CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
FACULDADE DE TECNOLOGIA DE BOTUCATU CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JOÃO MARCOS MARTINS JÚNIOR

ANÁLISE DA BASE MTCARS UTILIZANDO O SOFTWARE RSTUDIO

Botucatu – SP
Novembro – 2018

CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
FACULDADE DE TECNOLOGIA DE BOTUCATU CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JOÃO MARCOS MARTINS JÚNIOR

ANÁLISE DA BASE MTCARS UTILIZANDO O SOFTWARE RSTUDIO

Orientador: Prof. Dr. Carlos Roberto Pereira Padovani

Artigo apresentado à FATEC - Faculdade de Tecnologia de Botucatu, como exigência para cumprimento do Trabalho de Conclusão de Curso no Curso Superior de Análise e Desenvolvimento de Sistemas.

Botucatu – SP
Novembro – 2018

ANÁLISES ESTATÍSTICAS UTILIZANDO O SOFTWARE RSTUDIO

João Marcos Martins Júnior[1] Carlos Roberto Pereira Padovani[2]

RESUMO

Este trabalho teve como objetivo realizar diferentes análises estatísticas, utilizando como foco as ferramentas do *software RStudio*, além de utilizar outros *softwares* padrão como o Microsoft Word, sendo a base de dados utilizada fornecida pelo próprio *software.* A fim de demonstrar seu uso, foram analisadas informações de estatística descritiva, pertinentes aos dados contidos na base nativa *mtcars*. Os resultados foram obtidos e disponibilizados através de diferentes gráficos, adequados para os diferentes tipos de variáveis.
PALAVRAS-CHAVE: análise. dados. mtcars. rstudio.

ABSTRACT

This work aimed to perform different statistical analyzes, using as a focus the tools of *RStudio software*, in addition to using other standard *software* such as Microsoft Word, and the database used provided by the *software* itself. In order to demonstrate its use, descriptive statistical information, relevant to the data contained in the native *mtcars* database, was analyzed. The results were obtained and made available through different graphs, suitable for the different types of variables.
KEY WORDS: analysis. data. mtcars. rstudio.

[1] Graduando em Análise e Desenvolvimento de Sistemas– Fatec Botucatu, joao.m.martins@live.com
[2] Professor da Fatec Botucatu. Email. cpadovani@fatecbt.edu.br

1. INTRODUÇÃO

Com o advento da tecnologia e a chegada dos computadores pessoais, que garantiram um espaço de destaque, facilitou o desenvolvimento do conhecimento estatístico para profissionais de diferentes áreas. Nos dias atuais, existem *softwares* e dispositivos que possibilitam a utilização e o estudo de grandes volumes de dados, assim também como difundiram o uso dos métodos e fórmulas estatísticas (IGNÁCIO, 2012). Segundo Spiegel e Stephens (2000, p. 21):

> A estatística está relacionada aos métodos científicos para coleta, organização, resumo, apresentação e análise de dados, bem como à obtenção de conclusões válidas e à tomada de decisões razoáveis baseadas em tais análises.

A Análise Descritiva é a primeira etapa do estudo estatístico. Consiste numa organização dos dados, ressaltando os fatores de importância no conjunto, ou mesmo, fazer o comparativo de um determinado fator entre outros conjuntos. Esse tipo de análise usa como ferramenta a apresentação dos dados através de gráficos e tabelas. Como estes são componentes primários de uma análise, costumam ser utilizados para facilitar a compreensão de dados brutos, devido à perda que se tem ao condensar os dados (REIS, 2001).

R é um ambiente de *software* livre para computação estatística e gráficos. (R, 2018). É multiplataforma, sendo disponibilizado para diferentes sistemas operacionais, como *Windows*, *Mac OS X* e *Linux*. Possui gráficos excelentes, porém apresenta como ponto negativo sua linguagem de programação, também chamada de linguagem R, pois requer o conhecimento de uma série de comandos para que se tenha uma boa utilização do programa. (VERZANI, 2008). A principal vantagem da utilização de *softwares* como este, é a facilidade que ele agrega no cálculo estatístico, tanto pelo fator de fórmulas, quanto pelo fator de tamanho do conjunto de dados a ser estudado.

RStudio é uma IDE (*Integrated Developer Environment*), ou seja, é um ambiente de desenvolvedor integrado, que utiliza e combina a linguagem estatística R, com outros pacotes do programa R, como o *knitr* e *rmarkdown*, que permitem combinar análises estatísticas e a apresentação dos resultados em um documento na forma de um relatório. Também trabalham com outras linguagens como *Bash*, *Python* e *Ruby*; Além de *mark-ups*, que são instruções sobre como formatar um documento de apresentação (GANDRUD, 2015). Assim como o *software R*, também é *open source* (RACINE, 2012) e seu funcionamento e sua interface

gráfica, são de modo geral bastante similares aos do *software R* (HENNING, 2013). É o principal ambiente de desenvolvimento integrado da linguagem de programação *R*. Está disponível em edições de código aberto e comerciais na área de trabalho (*Windows, Mac e Linux*) e de um navegador da *Web* em um servidor *Linux* executando o *RStudio Server* ou o *RStudio Server Pro* (RSTUDIO, 2018).

A base de dados que foi utilizada nos testes e experimentos deste artigo é a *mtcars* que é fornecido pelo próprio *RStudio,* e contém dados da revista *Motor Trend* nos Estados Unidos de 1974, e compreendem dados com 32 observações divididas em 11 variáveis (modelos 1973-1974). As variáveis incluem aspectos como milhas por galão, número de cilindros dos veículos, tipo de transmissão (automática ou manual) e número de carburadores. Tais dados serão utilizados com a finalidade de testes e exemplificações.

Este artigo objetivou descrever o funcionamento básico do *software RStudio*, através do uso da base supracitada, a fim de realizar análises e testes estatísticos simples, assim como a criação de gráficos básicos.

2. MATERIAL E MÉTODOS

2.1 Material

Para a realização deste artigo foi utilizado um computador *notebook*, com a seguinte configuração:

- Processador Intel Core i5, 6ª geração com 2,3 GHz
- Memória de 8 GB RAM
- HDD de 1 TB

Os *softwares* nas seguintes versões:

- Microsoft Word 2016
- Sistema Operacional: Windows 7
- Software RStudio versão 1.0.136

A base de dados, que foi utilizada para a realização deste artigo:

- Base *mtcars* é fornecida pelo próprio *software*, e contém dados da revista Motor Trend nos Estados Unidos de 1974, e compreendem dados com 32 observações divididas em 11 variáveis (modelos 1973-1974). As variáveis incluem aspectos como milhas por galão, número de cilindros dos veículos, tipo de transmissão (automática ou manual) e número de carburadores. Tais dados foram utilizados com a finalidade de testes e exemplificações.

2.2 Metodologia de Pesquisa

A metodologia utilizada consistiu em pesquisa bibliográfica em livros, artigos acadêmicos e revistas científicas, e simulações estatísticas realizadas em *softwares* próprios, tendo como base para os testes as informações contidas na base *mtcars*. Foram utilizadas as seguintes análises estatísticas mais utilizadas, como Medidas de Tendência Central (Média Aritmética Simples e Mediana) e Medidas de Posição (Quartis).

Média Aritmética Simples: situa-se entre o valor máximo e o valor mínimo da distribuição. Não pode, portanto, ser inferior ou superior ao valor mínimo e ao máximo da distribuição. A média aritmética é um valor que pretende ser o resumo de todos os valores da distribuição. Dessa forma, pode vir a ser um valor não presente na distribuição, e possibilita inferir o resultado numa comparação de dois ou mais grupos, mostrando as nuances de cada um.

Cálculo da Média Aritmética Simples: A média aritmética simples é obtida somando todos os valores e dividindo essa soma pelo número de observações.

Fórmula da Média Aritmética Simples:

$$\overline{x} = \frac{x1 + x2 + \cdots + xn}{n}$$

Mediana: a mediana é o valor médio de uma distribuição ordenada, o qual apresenta o mesmo número de valores abaixo e acima desse valor.

Cálculo da Mediana: para calcular o valor da mediana, é necessário ordenar os dados. O valor da mediana é aquele abaixo e acima do qual encontra-se a mesma quantidade de valores da série.

Conjunto de dados é representado por N: N: 4, 2, 2, 1, 3, 5, 6, 3, 1, 1

Ordenar N: 1, 1, 1 ,2 ,2 ,3, 3, 4, 5, 6

Fórmula da Mediana:

$$\mathrm{Md} = \frac{\mathrm{N}+1}{2} = \frac{11}{2} = 5{,}5° \quad \mathrm{Md} = 2{,}5$$

Quartis: Os quartis dividem-se em 3 partes, que dividem um conjunto em 4 partes iguais.

Q1 - Primeiro Quartil - separa os primeiros 25% dos valores mais baixos da amostra ordenada;

Q2 - Segundo Quartil - separa os dados pela metade, ou seja, 50% dos valores mais baixos de uma amostra ordenada, dos 50% dos valores mais altos;

Q3 - Terceiro Quartil - separa os 75% dos valores mais baixos de uma amostra ordenada, dos 25% dos valores mais altos.

Cálculo do Quartil: para calcular o valor do quartil, é necessário ordenar os dados

Amostra: 6, 47, 49, 15, 42, 41, 7, 39, 43, 40, 36

Amostra ordenada: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Fórmulas do Quartil:

$$\mathrm{Q1} = \frac{1.\mathrm{N}}{4}$$

$$\mathrm{Q2} = \frac{2.\mathrm{N}}{4}$$

$$\mathrm{Q3} = \frac{3.\mathrm{N}}{4}$$

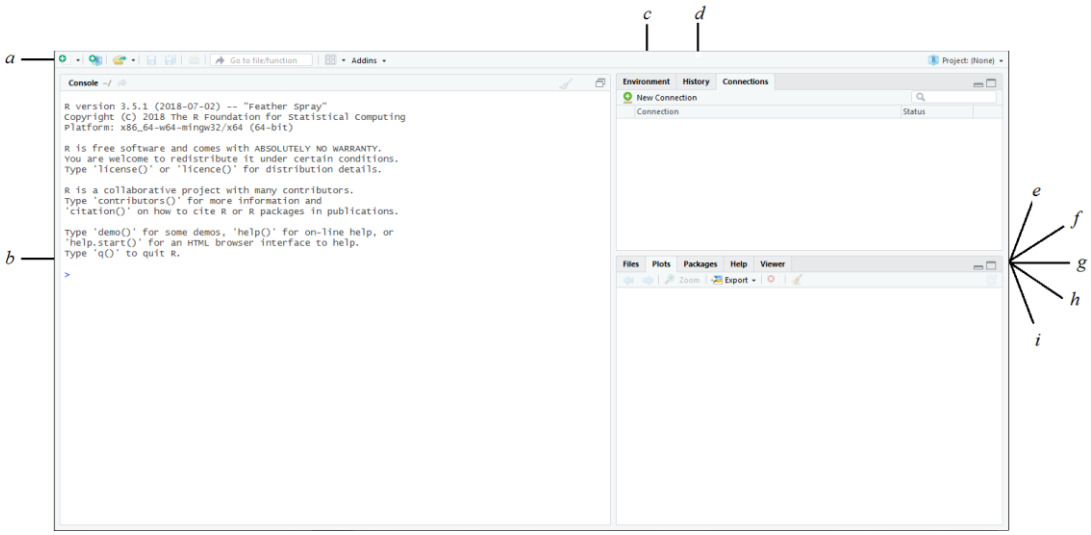
Q1 = 11/4 = 2,5, porém, deve-se arredondar para mais, no caso o quartil 1 será o valor na posição 3 da amostra. Logo, o Q1 será 15.

Q2 = 22/4 = 5,5, porém, deve-se arredondar para mais, no caso o quartil 2 será o valor na posição 6 da amostra. Logo, o Q2 será 40.

Q4 = 33/4 = 8,25, porém, deve-se arredondar para mais, no caso o quartil 1 será o valor na posição 9 da amostra. Logo, o Q3 será 43. (FEIJOO, 2010).

Para baixar e instalar o *RStudio,* acesse o link https://www.rstudio.com/products/rstudio/download/ e escolha a opção *RStudio Desktop.* Feito o *download* e instalação, a tela inicial do *RStudio* é dividida em algumas janelas menores, conforme ilustra a Figura 1.

Figura 1 - Tela e subtelas iniciais do RStudio



Fonte: Próprio Autor (2018)

a) O *Script* localizado no canto superior a esquerda é apresentada a janela *Source*, na qual são disponibilizados os scripts (códigos de programação previamente redigidos e salvos em arquivo com extensão. R). Para acessar a janela de script, basta clicar no ícone de retângulo com um sinal de + em verde, ou através do atalho Ctrl+Shift+N.

b) O *Console* do *software RStudio* fica no lado esquerdo, e é onde são digitados os comandos.

c) A ferramenta *Environments* é a aba que está localizada no lado direito, no bloco superior, sendo a aba primária do canto superior. Nela, ficam armazenados todos

os objetos criados e as bases de dados importadas, funções criadas e equações salvas.

d) O *History* é a ferramenta localizada no mesmo lugar que a aba *Environment*, é a aba secundária. Armazena o histórico de todos os comandos digitados.

No lado inferior direito existem 5 abas:

e) *Files*: são listados os arquivos constantes do diretório de trabalho (ou *working directory*). *Working directory* é uma pasta no computador que armazena os dados do *R/RStudio*. Ela guarda as bases de dados e os arquivos gerados e que serão utilizados pelo R.

f) *Plots*: aba que exibe os gráficos gerados no *R/RStudio.*

g) *Packages*: nessa aba são listados todos os pacotes instalados. Nela também é possível procurar novos pacotes, e atualizar os já instalados.

h) *Help*: aba que fornece um guia de ajuda sobre tópicos do *RStudio.*

i) *Viewer*: é utilizada para a visualização de arquivos interativos produzidos (SANTANA, 2017).

O próprio *RStudio* oferece ajuda nos comandos, contendo um glossário das equações, explicando o cálculo que cada comando faz. Para acessar o comando *help*, basta digitar no console o comando ? ou *help*( ) seguido do nome da função. Exemplo: ?*mean* ou *help(mean)*, que neste caso seria o comando de ajuda para média. Um comando muito útil, por exemplo é o *summary*, que realiza a parte de medidas de posição, como média, mediana, moda, além de exibir os valores mínimos e máximos contidos nas variáveis. A sintaxe no *RStudio* é *summary(mtcars).*

2.2.1 *RStudio*

O *RStudio* necessita para seu funcionamento que o *R* esteja previamente instalado. Sua instalação está disponível em http://cran.r-project.org/ , sendo que atualmente encontra-se na versão 3.5.1, que foi lançada em 2 de Julho de 2018 (HENNING, 2013).

3. RESULTADOS E DISCUSSÃO

A partir dos testes estatísticos realizados com o a base nativa do *RStudio mtcars*, foram possíveis obter resultados quanto as variáveis contidas e analisadas na base: Milhas por galão, Potência bruta, Transmissão (automática ou manual) e Número de Carburadores. Os resultados foram agrupados na seguinte tabela, de autoria própria, que mostrou, de acordo com as variáveis escolhidas, os valores mínimos e máximos de cada uma delas, além de valores de média, mediana e quartis:

Tabela 1 - Resumo das Estatísticas da base mtcars

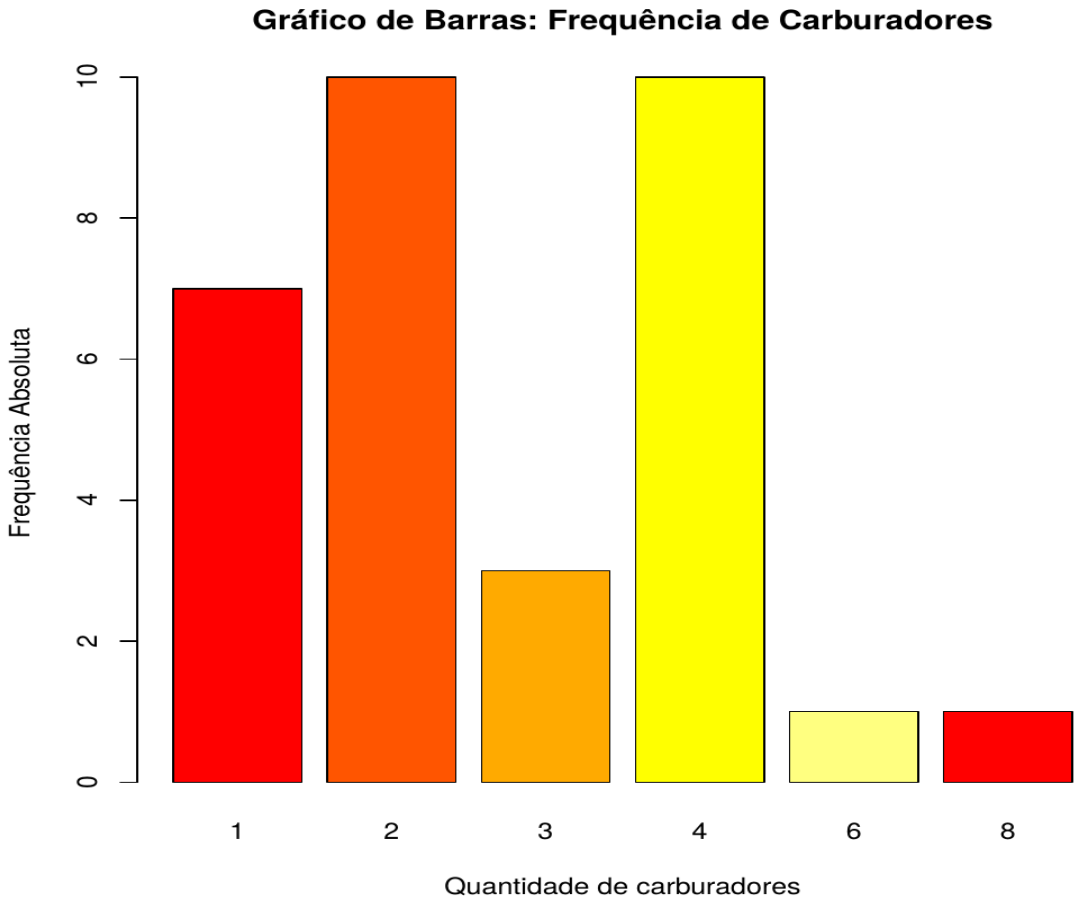
	Milhas p/ galão	Potência	Transmissão (automática ou manual)	Nº de carburadores
Valor mínimo	10.40	52.0	0	1.000
1º Quartil	15.43	96.5	0	2.000
Mediana	19.20	123.0	0	2.000
Média	20.09	146.7	0.40625	2.812
3º Quartil	22.80	180.0	1	4.000
Valor máximo	33.90	335.0	1	8.000

3.1 Gráficos

3.1.1 Gráficos para Variáveis Quantitativas

3.1.1.1 Barras: Constituído por retângulos ou barras, onde a dimensão apresenta a proporção a frequência a ser exibida (frequência absoluta (*ni*), ou como a frequência relativa (*fi*). São dispostas de forma paralela, umas as outras, podendo ser na vertical ou na horizontal, conforme ilustra a Figura 2.

Figura 2 - Gráfico de Barras



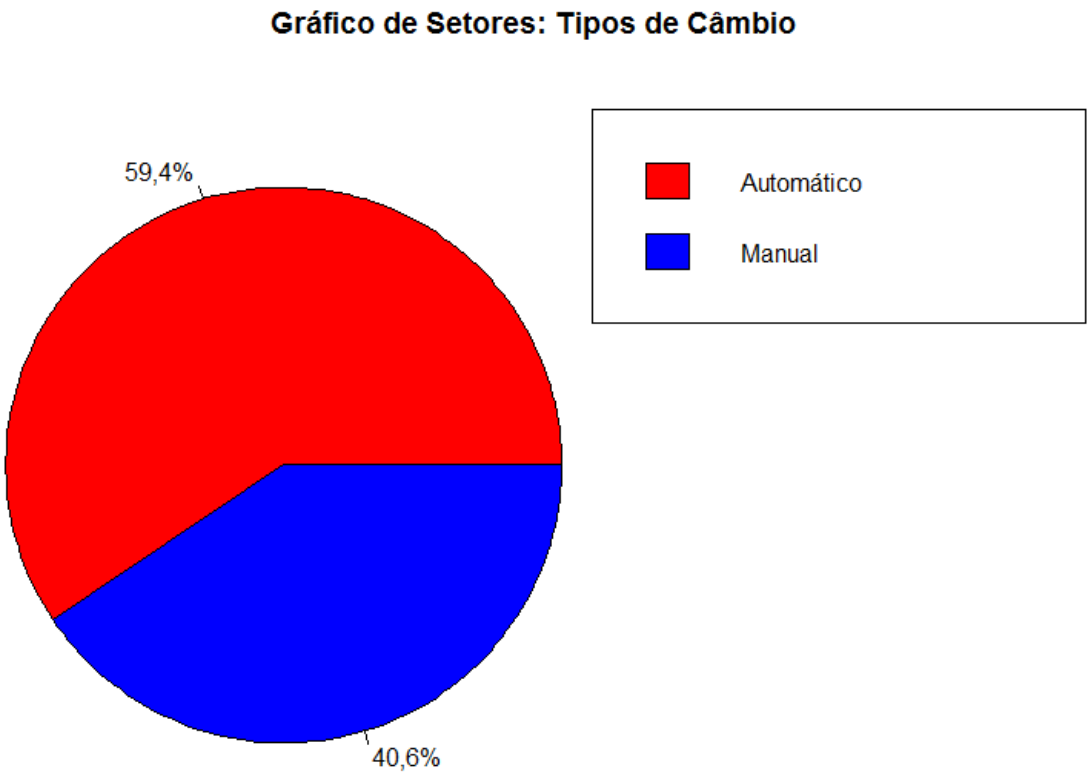
Fonte: Próprio Autor (2018).

Interpretação - Gráfico de Barras:

Dos 32 automóveis analisados, podemos observar a repetição da frequência de 10 veículos com 2 carburadores e com 4 carburadores; e apenas 1 veículo com 6 carburadores, e com 8 carburadores.

3.1.1.2 Setores/Pizza: muito utilizado para representações percentuais, de partes de um todo. É formado de um círculo, que representa o total, e dividido em setores (fatias), de maneira proporcional, conforme ilustra a Figura 3.

Figura 3 - Gráfico de Setores



Fonte: Próprio Autor (2018).

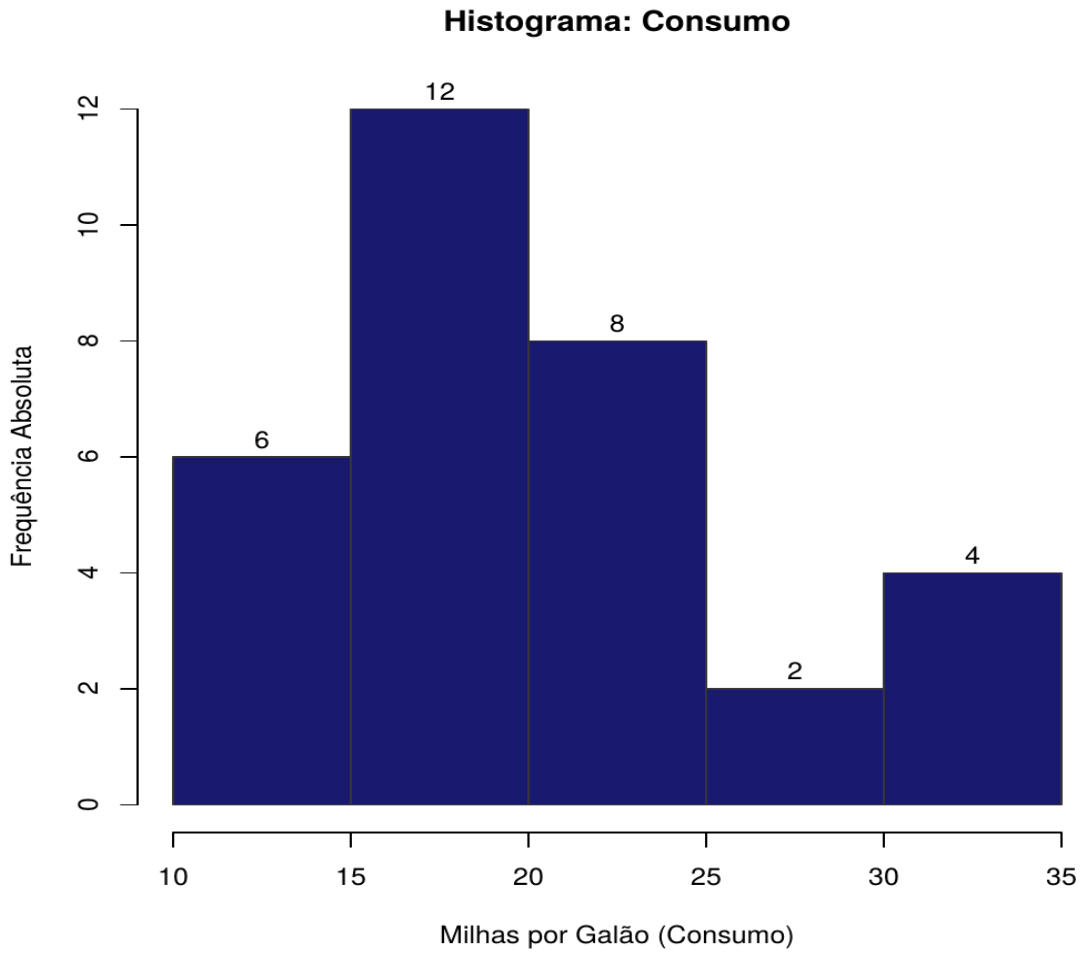
Interpretação - Gráfico de Setores:

Dos 32 automóveis analisados, 59,4% possuem câmbio do tipo Automático, e 40,6% possuem câmbio do tipo Manual.

3.1.2 Gráficos para Variáveis Qualitativas

3.1.2.1 Histograma: Similar ao gráfico de barras, porém não têm espaçamento entre as barras. Cada uma de suas barras apresenta bases de igual proporção aos intervalos de classes, e a área de cada retângulo proporcional à respectiva frequência. Permite que seja utilizada tanto a frequência absoluta (*ni*), como a frequência relativa (*fi*), conforme ilustra a Figura 4.

Figura 4 - Histograma



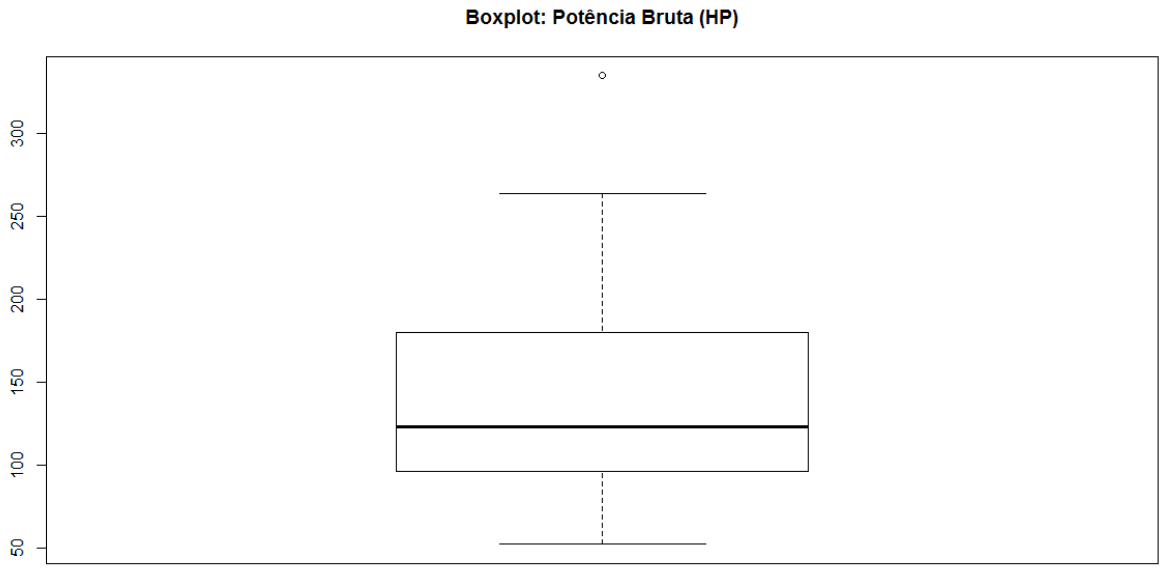
Fonte: Próprio Autor (2018).

Interpretação - Histograma:

De um total de 32 automóveis analisados, 12 apresentaram um consumo relativo entre 15 a 20 milhas por galão de combustível.

3.1.2.2 ***Boxplot:*** apresenta ideias de posição, dispersão, dentro outras medidas, além de dados discrepantes. Além disso, contém pontos específicos, que são os limites, superiores e inferiores. Os valores que ficam entre esses limites são os chamados valores adjacentes, e os valores acima ou abaixo dos limites são conhecidos como pontos exteriores (comumente em algumas literaturas são denominados pontos discrepantes), sendo graficamente representados por um asterisco (*). A posição central do *Boxplot* é a mediana; as posições relativas de Q1, Q2 e Q3 (quartis) fornecem uma noção sobre a assimetria da distribuição (MORETTIN; BUSSAB, 2017), conforme ilustra a Figura 5.

Figura 5 - Boxplot



Fonte: Próprio Autor (2018).

Interpretação - Boxplot:

Dos 32 automóveis analisados, 25% deles apresentaram uma potência inferior a 96,5 cavalos (HP); 50% deles apresentaram uma potência inferior a 123 cavalos (HP), e 75% apresentaram uma potência inferior a 180 cavalos (HP), e possui um ponto discrepante (ou limite superior) de 305,25.

4. CONCLUSÃO

Durante todo o tempo de experimento, o *software* se mostrou responsivo, de fácil utilização, com um consumo básico de recursos do computador, acarretando numa utilização plena de seus recursos.

Além disso, o *software* possui uma interface intuitiva e seus comandos, de modo geral, são de fácil compreensão, possibilitando assim um rápido aprendizado sobre como utilizar as ferramentas e posteriormente, alcançar um grande conhecimento sobre o mesmo.

REFERÊNCIAS BIBLIOGRÁFICAS

FEIJOO, A. M. L. C. **A pesquisa e a estatística na psicologia e na educação**. 2010. Disponível em:<http://books.scielo.org/id/yvnwq>. Acesso em: 18 jan. 2019.

GANDRUD, C. **Reproducible Research with R and RStudio Second Edition**. Disponível em: <https://www.taylorfrancis.com/books/9781498715386>. Acesso em: 03 jul. 2018.

HENNING, E. et al. **RStudio como suporte no ensino de planejamento de experimentos.** In: Congresso Brasileiro de Educação em Engenharia-COBENGE, Gramado. 2013. Disponível em: <http://www.fadep.br/engenharia-eletrica/congresso/pdf/117891_1.pdf>. Acesso em: 03 jul. 2018.

IGNÁCIO, S. A. **Importância da estatística para o processo de conhecimento e tomada de decisão**. Revista Paranaense de Desenvolvimento-RPD, n. 118, p. 175-192, 2012. Disponível em:<http://www.ipardes.gov.br/ojs/index.php/revistaparanaense/article/view/89/645>. Acesso em: 18 jan. 2019.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. Editora Saraiva, 2017. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=vDhnDwAAQBAJ&oi=fnd&pg=PA4&dq=ESTAT%C3%8DSTICA+B%C3%81SICA+Por+PEDRO+ALBERTO+MORETTIN,WILTON+OLIVEIRA+BUSSAB+boxplot&ots=DXf2gfcX_e&sig=ltSjwwscbYoYQSB4BSDETGDikLo#v=onepage&q=ESTAT%C3%8DSTICA%20B%C3%81SICA%20Por%20PEDRO%20ALBERTO%20MORETTIN%2CWILTON%20OLIVEIRA%20BUSSAB%20boxplot&f=false. Acesso em: 30 out. 2018.

R: The R Project for Statistical Computing: Getting Started. Version 3.5.1. 2018. Disponível em: <https://www.r-project.org/>. Acesso em: 05 jul. 2018.

RACINE, J. S. RStudio: A Platform-Independent IDE for R and Sweave. **Journal of Applied Econometrics**, v. 27, n. 1, p. 167-172, 2012. Disponível em:<https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1278>. Acesso em: 03 jul. 2018.

REIS, E. A.; REIS, I. A. Análise Descritiva de Dados. **Síntese numérica Estatística**, 2001. Disponível em:<http://www.est.ufmg.br/portal/arquivos/rts/rte0202.pdf>. Acesso em: 06 jul. 2018.

RSTUDIO: Desktop Version 1.1.453. 2018. Disponível em: <https://www.rstudio.com/products/rstudio/download2/>. Acesso em: 03 jul. 2018.

SANTANA, V. **Tutorial R/RStudio.** 2017. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/2996937/mod_resource/content/1/Tutorial.pdf>. Acesso em: 01 jan. 2017.

SPIEGEL, M. R.; STEPHENS, L. J. **Estatística: Coleção Schaum**. Bookman, 2000. Disponível em:<https://books.google.com.br/books?hl=pt-BR&lr=&id=h70pDwAAQBAJ&oi=fnd&pg=PR4&dq=conceito+de+estat%C3%ADstica&ot

s=WGRlDPrjEC&sig=k6fdk9VFAeXMeCzaaRjWR4zC-CI#v=onepage&q=conceito%20de%20estat%C3%ADstica&f=false>. Acesso em: 05 jul. 2018.

VERZANI, J. Using R in Introductory Statistics Courses with the pmg Graphical User Interface. **Journal of Statistics Education**, v. 16, n. 1, 2008. Disponível em: <https://amstat.tandfonline.com/doi/pdf/10.1080/10691898.2008.11889558?needAccess=true>. Acesso em: 03 jul. 2018.