Relatório do Projeto de Disciplina

Análise Exploratória de Dados

Eraldo N. Ferreira Pinto Júnior

07 abril, 2024

Sumário

Introdução

Para uma melhor organização deste projeto, foram criadas pastas com propósitos específicos para armazenamento dos arquivos.

A estrutura é dividida da seguinte forma:

- / - Pasta raiz do projeto:
  - Dataset - Pasta com os arquivos de extensão CSV.
    - Brent
    - Cambio
    - Combustivel
    - PPI
  - Image - Pasta com os arquivo de extensão PNG.

Preparando o Ambiente

Pacotes

Durante a análise foi verificado a necessidade de utilização de alguns pacotes. A lista de pacotes utilizados encontra-se abaixo:

```
# Lista com todos os pacotes
package <- c("tidyverse", "ggplot2", "summarytools", "data.table", "knitr", "dlookr", "ggpubr", "naniar
```

Com a lista de pacotes mapeados, a próxima etapa tem como foco verificar se todos os pacotes necessários para a análise se encontram instalados. Caso não estejam instalados, o processo de instalação será automático.

Para verificação dos pacotes instalados, a variável *is_installed* receberá o resultado *TRUE* para os pacotes instalados da lista *package* ou *FALSE* para os pacotes não instalados da lista *package*.

```
# Veritifica se o Pacote está instalado e o instala se for necessário.
is_installed <- package %in% rownames(installed.packages())

if(any(is_installed == FALSE)){
  install.packages(package[!is_installed])
}
```

Com os pacotes instalados, agora é necessário o carregamento deles, para que assim possamos iniciar as tratativas necessárias com os dados da análise.

```
# Carregando os Pacotes
invisible(lapply(package, library, character.only = TRUE))
```

Pacotes carregados. Agora vamos remover as variáveis desnecessárias para as próximas etapas.

```
# Removendo variáveis desnecessárias
rm(list=ls())
```

Funções Criadas

Durante a elaboração da análise foi identificado a necessidade de realizar o tratamento das datas, uniformizando-as a partir das diferentes fontes de dados. Foram criadas funções para este tratamento.

A função *transform_date_one* transforma no formato AAAA-mm-dd todos os valores com o formato 12.10.2023 ou 12/10/2023.

A função *transform_date_two* transforma no formato AAAA-mm-dd todos os valores com o formato 12102023.

A função *transform_date_three* transforma no formato AAAA-mm-dd todos os valores com o formato Apr 2023.

Foi criada uma função específica que retornará a data formatada com padrão único, chamada *format_data*. Esta função será utilizada por todas as outras funções *transform_date_*.

```
transform_date_one <- function(data){
  partes_da_data <- strsplit(data, "[/.]")
  ano <- as.numeric(sapply(partes_da_data, `[`,3))
  mes <- as.numeric(sapply(partes_da_data, `[`,2))

  data_formatada <- format_data(ano, mes)

  return(data_formatada)
}

transform_date_two <- function(data){
  ano <- as.numeric(substr(data, nchar(data) - 3, nchar(data)))
  mes <- as.numeric(substr(data, nchar(data) - 5, nchar(data) - 4))

  data_formatada <- format_data(ano, mes)

  return(data_formatada)
}

transform_date_three <- function(data){
  ano <- as.numeric(substr(data, nchar(data) - 3, nchar(data)))
  mes_abreviado <- substr(data, nchar(data) - 7, nchar(data) - 5)
  mes <- as.integer(match(mes_abreviado, month.abb))

  data_formatada <- format_data(ano, mes)

  return(data_formatada)
}

format_data <- function(ano_data, mes_data){
  data_formatada <- as.Date(sprintf("%04d-%02d-01", ano_data, mes_data))
  return(data_formatada)
}
```

Dando continuidade a necessidade de funções específicas, alguns dados foram disponibilizados pelas suas fontes em arquivos distintos. Para uma carga de dados mais eficiente foram criadas funções que possibilitarão uma agilidade neste processo.

```r
extractor_csv2 = function(dados){
  readr::read_csv2(dados, locale = locale(encoding = 'UTF-8'), show_col_types = FALSE)
}

extractor_csv = function(dados){
  read.csv(dados, header = FALSE, sep = ";", dec = ",")
}
```

Abaixo é descrita as funções geradoras de binwidths.

```r
fd <- function(x) {
  n <-length(x)
  return((2*IQR(x))/n^(1/3))
}


sr <- function(x) {
  n <-length(x)
  return((3.49*sd(x))/n^(1/3))
}
```

Dos Dados

Contexto

O valor de venda dos derivados de petróleo aos consumidores brasileiros é sempre um assunto polêmico. Há muitas variáveis que influenciam na flutuação do valor de venda. Para o consumidor final o que importa é o quanto estas flutuações impactam no orçamento mensal da família.

Os meios de comunicação frequentemente noticiam o aumento ou a redução dos derivados do petróleo diante da flutuação de algumas variáveis, como por exemplo o brent e o câmbio.

A flutuação destas variáveis e de outras são oriundas de acontecimentos mundiais. Os grandes canais de comunicação noticiam periodicamente estes eventos.

- O "Petróleo sobe mais de 3% em meio a tensões no Oriente Médio" (CNN-Brasil, 2024).
- A "Guerra e petróleo: veja reações mais drásticas da commodity a grandes conflitos" (CNN-Brasil, 2023).
- A "Gerra no Oriente Médio pode aumentar preço do diesel, diz Petrobras" (AgênciaBrasil-EBC, 2023).

Além das variáveis, uma sigla que foi introduzida na vida dos brasileiros diante a mudança da política de preço praticada pela petrolífera brasileira (Petrobras). Esta sigla é conhecida como o Preço de Paridade de Importação - PPI.

Um breve histórico da adoção do PPI pela Petrobras e seus desdobramentos políticos pode ser lido na matéria "Gasolina cara, lucro recorde: como foi o PPI, antiga política da Petrobras" (Economia UOL, 2023)

O fim da adoção do PPI pela Petrobras em 16 de maio de 2023 repercutiu nacionalmente.

- A "Petrobras anuncia fim da paridade internacional de preços do petróleo" (CNN Brasil, 2023). "Para Ineep, fim do PPI na Petrobras trouxe maior estabilidade de preço dos combustíveis" (InfoMoney, 2024).

- A “Gasolina da Petrobras está 17% mais barata que preço internacional” (Metrópoles, 2024).

Em resumo, o objetivo desta análise é compreender quais, como e o quanto as variáveis influenciam na vida dos consumidores e na política de preço dos derivados de petróleo, em específico a Gasolina no Estado do Rio de Janeiro no período de Janeiro de 2020 até Fevereiro de 2024.

Fonte dos dados

Um fator crucial para qualquer análise é a busca de fontes de dados abertos confiáveis. Portanto, buscou-se através de sites oficiais de governo e instituições renomadas os dados necessários para a respectiva análise.

O primeiro dado a ser obtido foi a “Série Histórica de Preços de Combustíveis e de GLP” (Dados Abertos-ANP, 2024). Esta fonte de dados possui os dados das pesquisas realizadas até a penúltima semana de março de 2024. Os dados utilizandos para esta análise foram os dados oriundos das pesquisas realizadas até fevereiro de 2024, pois, no momento da coleta dos dados, as pesquisas realizadas no mês de março de 2024 ainda não foram finalizadas.

Ainda no site da ANP, foi utilizado os “Preços de paridade de importação” (PPI-ANP, 2024).

O Brent foi obtido através da *U.S. Energy information Administration - EIA*. Os dados obtidos fazem parte da visão histórica dos dados em *PETROLEUM & OTHER LIQUIDS* (EIA, 2024).

A série histórica da taxa cambial foi obtida através do site do Banco Central do Brasil, em sua área “Cotações e boletins” (BCB, 2024).

Realizando o carregamento dos dados

Neste momento será realizado o carregamento dos dados obtidos através das fontes de dados supracitadas. Os dados foram armazenados nas subpastas da pasta Dataset.

Para realização desta etapa, duas estratégias foram adotada.

A primeira estratégia, de forma recursiva, se utilizou o *list.files* para localizar todos os arquivos a partir de um *pattern* (padrão) no nome dos arquivos. Uma variável com a lista contendo o nome do arquivo e o caminho foi criada para armazená-las. Posteriormente foi utilizada a função map_dfr para aplicar cada elemento (arquivos) na função criada para extração dos dados. Esta estratégia envolve a carga de dados histórica dos combustíveis e taxa de câmbio que são constituídas de vários arquivos.

```
arquivos <- list.files(pattern = "^ca-", recursive = TRUE)
combustivel_agg <- map_dfr(arquivos, extractor_csv2)
message("Dados carregados dos arquivos CSV.")

rm("arquivos")
```

Iniciando a carga dos dados da taxa de câmbio.

```
arquivos <- list.files(pattern = "^CotacoesMoedasPeriodo", recursive = TRUE)
cambio_agg <- map_dfr(arquivos, extractor_csv)

rm("arquivos")
```

A segunda estratégia, foi mais simples, pois se refere a extração de um único arquivo com todos os dados históricos do Brent.

```
brent <- read.table("Dataset/Brent/Europe_Brent_Spot_Price_FOB.csv", sep=",",
                    header = TRUE)
colnames(brent) <- c("Data", "Brent USD/Barril")
```

A segunda estratégia também foi adotada para a extração de um único arquivo com todos os dados históricos do PPI.

Atenção: Os dados do PPI foram disponibilizados em vários sheets em uma única planilha do Excel, com extensão XLSX. Foi necessário tratar os dados diretamente no Excel, possibilitando assim um carregamento mais célere.

```
ppi <- read.table("Dataset/PPI/ppi.csv", sep=";", dec = ",",
                  header = TRUE)
```

Visão Inicial dos Dados

Após a importação das bases de dados, vamos apresentar as primeiras observações para conhecimento das variáveis.

```
head(combustivel_agg)
```

```
## # A tibble: 6 x 16
##   `Regiao - Sigla` `Estado - Sigla` Municipio Revenda          `CNPJ da Revenda`
##   <chr>            <chr>            <chr>     <chr>            <chr>
## 1 SE               SP               GUARULHOS AUTO POSTO SAKA~ 49.051.667/0001-~
## 2 SE               SP               GUARULHOS AUTO POSTO SAKA~ 49.051.667/0001-~
## 3 SE               SP               GUARULHOS AUTO POSTO SAKA~ 49.051.667/0001-~
## 4 SE               SP               GUARULHOS AUTO POSTO SAKA~ 49.051.667/0001-~
## 5 NE               BA               SALVADOR  PETROBRAS DISTR~ 34.274.233/0015-~
## 6 NE               BA               SALVADOR  PETROBRAS DISTR~ 34.274.233/0015-~
## # i 11 more variables: `Nome da Rua` <chr>, `Numero Rua` <chr>,
## #   Complemento <chr>, Bairro <chr>, Cep <chr>, Produto <chr>,
## #   `Data da Coleta` <chr>, `Valor de Venda` <dbl>, `Valor de Compra` <dbl>,
## #   `Unidade de Medida` <chr>, Bandeira <chr>
```

```
head(cambio_agg)
```

```
##        V1  V2 V3  V4     V5     V6 V7 V8
## 1 2012020 220  A USD 4.0207 4.0213  1  1
## 2 3012020 220  A USD 4.0516 4.0522  1  1
## 3 6012020 220  A USD 4.0548 4.0554  1  1
## 4 7012020 220  A USD 4.0835 4.0841  1  1
## 5 8012020 220  A USD 4.0666 4.0672  1  1
## 6 9012020 220  A USD 4.0738 4.0744  1  1
```

```r
head(brent)
```

```
##       Data Brent USD/Barril
## 1 Feb 2024            83.48
## 2 Jan 2024            80.12
## 3 Dec 2023            77.63
## 4 Nov 2023            82.94
## 5 Oct 2023            90.60
## 6 Sep 2023            93.72
```

```r
head(ppi)
```

```
##         Data Santos Duque.de.Caxias Cubatão   Mauá Paulínia São.José.dos.Campos
## 1 01/01/2020 1.8941          1.9595  1.9027 1.9201   1.9275              1.9253
## 2 01/01/2020 1.8608          1.9262  1.8694 1.8869   1.8943              1.8920
## 3 01/01/2020 1.8544          1.9198  1.8630 1.8804   1.8878              1.8856
## 4 01/01/2020 1.8224          1.8878  1.8310 1.8485   1.8559              1.8536
## 5 01/01/2020 1.7296          1.7951  1.7383 1.7557   1.7631              1.7608
## 6 01/02/2020 1.6884          1.7539  1.6971 1.7145   1.7219              1.7196
##    Produto Unidade.de.Medida
## 1 Gasolina          R$/litro
## 2 Gasolina          R$/litro
## 3 Gasolina          R$/litro
## 4 Gasolina          R$/litro
## 5 Gasolina          R$/litro
## 6 Gasolina          R$/litro
```

Analisando os tipos de cada variável nas bases

Após a carga dos dados, foi necessário identificar o tipo de cada variável nas bases. Utilizar-se-á a função diagnose do pacote dlookr que reportará o tipo em todas as bases.

Base de Dados Combustíveis - Visão da pesquisa nacional

```r
combustivel_agg %>% dlookr::diagnose()
```

```
## # A tibble: 16 x 6
##    variables        types missing_count missing_percent unique_count unique_rate
##    <chr>            <chr>         <int>           <dbl>        <int>       <dbl>
##  1 Regiao - Sigla   char~             0               0            5  0.00000145
##  2 Estado - Sigla   char~             0               0           27  0.00000783
##  3 Municipio        char~             0               0          469  0.000136
##  4 Revenda          char~             0               0        18505  0.00536
##  5 CNPJ da Revenda  char~             0               0        19906  0.00577
##  6 Nome da Rua      char~             0               0        12003  0.00348
##  7 Numero Rua       char~          1531          0.0444         5141  0.00149
```

```
##  8 Complemento      char~       2677681         77.6            3570 0.00103
##  9 Bairro           char~          8553          0.248          7681 0.00223
## 10 Cep              char~             0          0             13888 0.00403
## 11 Produto          char~             0          0                 6 0.00000174
## 12 Data da Coleta   char~             0          0              1028 0.000298
## 13 Valor de Venda   nume~             0          0              4976 0.00144
## 14 Valor de Compra  nume~       3293286         95.5            22767 0.00660
## 15 Unidade de Medi~ char~             0          0                 3 0.000000870
## 16 Bandeira         char~             0          0                79 0.0000229
```

```
combustivel_agg %>% dplyr::select(Produto) %>% base::unique()
```

```
## # A tibble: 6 x 1
##   Produto
##   <chr>
## 1 GASOLINA
## 2 ETANOL
## 3 DIESEL S10
## 4 GNV
## 5 DIESEL
## 6 GASOLINA ADITIVADA
```

```
#str(combustivel_agg)
```

As variáveis deste contjunto de dados pode ser classificadas conforme a tabela abaixo.

\begin{center} Table 1: Classificação variáveis base combustíveis. \end{center

Variável	Classificação
Regiao - Sigla	Qualitativa nominal
Estado - Sigla	Qualitativa nominal
Municipio	Qualitativa nominal
Revenda	Qualitativa nominal
CNPJ da Revenda	Qualitativa nominal
Nome da Rua	Qualitativa nominal
Numero Rua	Qualitativa nominal
Complemento	Qualitativa nominal
Bairro	Qualitativa nominal
Cep	Qualitativa nominal
Produto	Qualitativa nominal
Data da Coleta	Qualitativa nominal
Valor de Venda	Quantitativa contínua
Valor de Compra	Quantitativa contínua
Unidade de Medida	Qualitativa nominal
Bandeira	Qualitativa nominal

É possível identificar a existência na base de 16 variáveis. Um total de 14 variáveis são qualitativas, sendo estas nominais. Sobre as variáveis quantitativas, temos Valor de Venda e Valor de Compra, ambas variáveis contínuas.

Observa-se que as variáveis Numero da rua, Complemento, Bairro e Valor de Venda possuem missing.

No Brasil, a pesquisa foi realizada:

- Em 5 regiões.
- Em 27 estados.
- Em 469 municípios.
- Em 19.906 revendas por CNPJ.
- Considerando 6 produtos comercializados.
  - GASOLINA
  - ETANOL
  - DIESEL S10
  - GNV
  - DIESEL
  - GASOLINA ADITIVADA
- E o missing do Valor de Compra foi de 3.293.286, o qual representa 95,5% da base de dados.
- Em 79 bandeiras diferentes.
- As variáveis de interesse nesta base de dados são:
  - Regiao - Sigla
  - Estado - Sigla
  - Municipio
  - CNPJ da Revenda
  - Produto
  - Data da Coleta
  - Valor de Venda
  - Valor de Compra
  - Unidade de Medida
  - Bandeira

Para a respectiva análise um conjunto de variáveis foram selecionadas.

```
combustiveis <- combustivel_agg[,c(1,2,3,5,11,12,16,13,14)]
```

Foram inseridos novos nomes para as variáveis.

```
colnames(combustiveis) <- c("Regiao",
                            "UF",
                            "Municipio",
                            "CNPJ_Revenda",
                            "Produto",
                            "Data",
                            "Bandeira",
                            "Valor_de_Venda",
                            "Valor_de_Compra"
                            )
```

Base de Dados Câmbio

Dando continuidade, será realizada a análise da próxima base de dados, taxa de câmbio.

```r
str(cambio_agg)
```

```
## 'data.frame':    1043 obs. of  8 variables:
##  $ V1: int  2012020 3012020 6012020 7012020 8012020 9012020 10012020 13012020 14012020 15012020 ...
##  $ V2: int  220 220 220 220 220 220 220 220 220 220 ...
##  $ V3: chr  "A" "A" "A" "A" ...
##  $ V4: chr  "USD" "USD" "USD" "USD" ...
##  $ V5: num  4.02 4.05 4.05 4.08 4.07 ...
##  $ V6: num  4.02 4.05 4.06 4.08 4.07 ...
##  $ V7: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ V8: num  1 1 1 1 1 1 1 1 1 1 ...
```

```r
cambio_agg %>% dlookr::diagnose()
```

```
## # A tibble: 8 x 6
##   variables types     missing_count missing_percent unique_count unique_rate
##   <chr>     <chr>             <int>           <dbl>        <int>       <dbl>
## 1 V1        integer               0               0         1043    1
## 2 V2        integer               0               0            1    0.000959
## 3 V3        character             0               0            1    0.000959
## 4 V4        character             0               0            1    0.000959
## 5 V5        numeric               0               0          991    0.950
## 6 V6        numeric               0               0          992    0.951
## 7 V7        numeric               0               0            1    0.000959
## 8 V8        numeric               0               0            1    0.000959
```

Ao inspecionar a base de dados, é possível identificar na base a existência de 8 variáveis. Neste primeiro momento não foi possível identificar claramente o propósito das variáveis. Para isso, foi utilizado a função head para leitura dos primeiros dados da base.

É possível observar que a variável V1 é do tipo integer, contudo, ela expressa a data de cotação do câmbio, portanto, é uma variável qualitativa nominal. As variáveis V5 e V6 expressão, respectivamente, cotação de compra e venda na moeda real. Portanto, são quantitativas e ambas são contínuas. A variável V2, V3 e V4, através da informação contida na descrição da base, são: V2 - Código da Moeda, V3 - Tipo da Moeda e V4 - Símbolo da Moeda. Ambas são qualitativas nominais. As variáveis V7 e V8 não possuem descrição na base.

Para o objetivo da análise, foram selecionadas as variáveis V1 e V5

Estes dados foram inseridos em uma nova variável de ambiente, de nome cambio, permitindo que o carregamento original não seja descartado, possibilitando análises futuras.

```r
cambio <- cambio_agg[, c(1, 5)]
colnames(cambio) <- c("Data", "Taxa_Cambio")
```

Base de Dados Brent

A análise de variável desta base de dados foi mais simples. A base é composta por duas variáveis. A data é uma variável qualitativa nominal e a variável Preco USD/Barril é quantitativa contínua.

```r
str(brent)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ Data            : chr  "Feb 2024" "Jan 2024" "Dec 2023" "Nov 2023" ...
##  $ Brent USD/Barril: num  83.5 80.1 77.6 82.9 90.6 ...
```

```
brent %>% dlookr::diagnose()
```

```
## # A tibble: 2 x 6
##   variables        types  missing_count missing_percent unique_count unique_rate
##   <chr>            <chr>          <int>           <dbl>        <int>       <dbl>
## 1 Data             chara~             0               0           50           1
## 2 Brent USD/Barril numer~             0               0           50           1
```

Base de Dados PPI

Considerando que esta base de dados passou por um tratamento prévio e externo, as variáveis existentes são os valores do PPI por portos e pontos de entrega. A data é uma variável qualitativa nominal e as variáveis que representam os portos e os pontos de entrega são quantitativas contínuas.

```
str(ppi)
```

```
## 'data.frame':    654 obs. of  9 variables:
##  $ Data               : chr  "01/01/2020" "01/01/2020" "01/01/2020" "01/01/2020" ...
##  $ Santos             : num  1.89 1.86 1.85 1.82 1.73 ...
##  $ Duque.de.Caxias    : num  1.96 1.93 1.92 1.89 1.8 ...
##  $ Cubatão            : num  1.9 1.87 1.86 1.83 1.74 ...
##  $ Mauá               : num  1.92 1.89 1.88 1.85 1.76 ...
##  $ Paulínia           : num  1.93 1.89 1.89 1.86 1.76 ...
##  $ São.José.dos.Campos: num  1.93 1.89 1.89 1.85 1.76 ...
##  $ Produto            : chr  "Gasolina" "Gasolina" "Gasolina" "Gasolina" ...
##  $ Unidade.de.Medida  : chr  "R$/litro" "R$/litro" "R$/litro" "R$/litro" ...
```

```
ppi %>% dlookr::diagnose()
```

```
## # A tibble: 9 x 6
##   variables         types missing_count missing_percent unique_count unique_rate
##   <chr>             <chr>         <int>           <dbl>        <int>       <dbl>
## 1 Data              char~             0               0           50     0.0765
## 2 Santos            nume~             0               0          639     0.977
## 3 Duque.de.Caxias   nume~           218            33.3          436     0.667
## 4 Cubatão           nume~           218            33.3          433     0.662
## 5 Mauá              nume~           218            33.3          434     0.664
## 6 Paulínia          nume~           218            33.3          432     0.661
## 7 São.José.dos.Cam~ nume~           218            33.3          435     0.665
## 8 Produto           char~             0               0            3     0.00459
## 9 Unidade.de.Medida char~             0               0            2     0.00306
```

Para a respectiva análise utilizar-se-á a variável Duque.de.Caxias, a qual, será renomeada para PPI_DC.

```
ppi <- ppi %>%
   dplyr::filter(Produto == "Gasolina") %>%
   dplyr::select(Data, Duque.de.Caxias) %>%
   dplyr::rename("PPI_DC" = "Duque.de.Caxias")
```

Esta variável contém valores referente ao Ponto de entrega chamado Duque de Caxias. Foi considerado esta variavél pois trata do ponto de entrega mais próximo para o estado do Rio de Janeiro.

Tratamento das Datas

As quatro bases de dados dispostas nesta análise (combustiveis, cambio, brent e ppi) possuem datas com formatos e características diferentes.

A uniformização das datas possibilitará mesclar estes dados em uma única base de dados.

Para esta uniformização, as funções de transformação das datas serão chamadas, passando como parâmetro o campo data das bases de dados.

```
combustiveis$Data <- transform_date_one(combustiveis$Data)
cambio$Data <- transform_date_two(cambio$Data)
brent$Data <- transform_date_three(brent$Data)
ppi$Data <- transform_date_one(ppi$Data)
```

Análise de frequências de variáveis qualitativas

Na base de dados combustiveis, as variáveis Produto, Região, UF e Município são variáveis qualitativas nominais na base. São variáveis interessantes para extração das frequências.

Para esta primeira análise de frequência, analisaremos a variável produto utilizando a função freq() do pacote summarytools.

```
combustiveis %>%
   dplyr::select(Produto) %>%
   summarytools::freq(., style = 'rmarkdown', order = "freq", plain.ascii = FALSE)
```

```
## ### Frequencies
## #### combustiveis$Produto
## **Type:** Character
##
## |                   |    Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-----------------------:|--------:|--------:|-------------:|--------:|-------------:|
## |           **GASOLINA** |  921556 |   26.72 |        26.72 |   26.72 |        26.72 |
## |             **ETANOL** |  808287 |   23.43 |        50.15 |   23.43 |        50.15 |
## |          **DIESEL S10** |  695662 |   20.17 |        70.32 |   20.17 |        70.32 |
## | **GASOLINA ADITIVADA** |  554522 |   16.08 |        86.40 |   16.08 |        86.40 |
## |             **DIESEL** |  405268 |   11.75 |        98.14 |   11.75 |        98.14 |
## |                **GNV** |   64003 |    1.86 |       100.00 |    1.86 |       100.00 |
## |             **\<NA\>** |       0 |         |              |    0.00 |       100.00 |
## |              **Total** | 3449298 |  100.00 |       100.00 |  100.00 |       100.00 |
```

Nas colunas Freq, temos a frequência absoluta, mostrando um grau de heterogeneidade. Através da coluna % Valid, que apresenta a frequência relativa por produto, é possível ratificar esta observação. Observa-se um destaque para os Produtos GASOLINA, ETANOL e DIESEL S-10, portanto, sendo os produtos que mais foram coletados durante as pesquisas de preços realizadas.

Para esta segunda análise de frequência, analisaremos a variável Regiao.

```
combustiveis %>%
   dplyr::select(Regiao) %>%
   summarytools::freq(., style = 'rmarkdown', order = "freq", plain.ascii = FALSE)
```

```
## ### Frequencies
## #### combustiveis$Regiao
## **Type:** Character
##
## |       |    Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-----------:|--------:|--------:|-------------:|--------:|-------------:|
## |     **SE** | 1674880 |   48.56 |        48.56 |   48.56 |        48.56 |
## |     **NE** |  675512 |   19.58 |        68.14 |   19.58 |        68.14 |
## |      **S** |  586858 |   17.01 |        85.16 |   17.01 |        85.16 |
## |     **CO** |  295137 |    8.56 |        93.71 |    8.56 |        93.71 |
## |      **N** |  216911 |    6.29 |       100.00 |    6.29 |       100.00 |
## | **\<NA\>** |       0 |         |               |    0.00 |       100.00 |
## |  **Total** | 3449298 |  100.00 |       100.00 |  100.00 |       100.00 |
```

A coluna Freqs demonstra uma heterogeneidade. É possível observar através da coluna (% of Valid) a frequência relativa por região. A observação possibilita demonstrar que a região Sudeste contribui com quase metade das fontes de postos de combustíveis pesquisados.

É possível observar que a ordem do resultado da frequência relativa por região é igual a análise de do quantitativo de missing da variável Valor_de_Compra por região.

Para a análise de frequência dos Estados que mais contribuíram para a pesquisa, analisaremos a variável UF utilizando a função freq() do pacote summarytools.

```
combustiveis %>%
   dplyr::select(UF) %>%
   summarytools::freq(., style = 'rmarkdown', order = "freq", plain.ascii = FALSE)
```

```
## ### Frequencies
## #### combustiveis$UF
## **Type:** Character
##
## |       |    Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-----------:|--------:|--------:|-------------:|--------:|-------------:|
## |     **SP** | 1007206 |   29.20 |        29.20 |   29.20 |        29.20 |
## |     **MG** |  334763 |    9.71 |        38.91 |    9.71 |        38.91 |
## |     **RJ** |  263852 |    7.65 |        46.56 |    7.65 |        46.56 |
## |     **RS** |  223506 |    6.48 |        53.03 |    6.48 |        53.03 |
## |     **PR** |  220206 |    6.38 |        59.42 |    6.38 |        59.42 |
## |     **BA** |  185710 |    5.38 |        64.80 |    5.38 |        64.80 |
## |     **SC** |  143146 |    4.15 |        68.95 |    4.15 |        68.95 |
## |     **CE** |  126553 |    3.67 |        72.62 |    3.67 |        72.62 |
## |     **GO** |  122929 |    3.56 |        76.19 |    3.56 |        76.19 |
```

```
## |     **PE** |  107256 |    3.11 |        79.30 |    3.11 |        79.30 |
## |     **MT** |   84661 |    2.45 |        81.75 |    2.45 |        81.75 |
## |     **ES** |   69059 |    2.00 |        83.75 |    2.00 |        83.75 |
## |     **PA** |   63064 |    1.83 |        85.58 |    1.83 |        85.58 |
## |     **MA** |   61979 |    1.80 |        87.38 |    1.80 |        87.38 |
## |     **MS** |   50729 |    1.47 |        88.85 |    1.47 |        88.85 |
## |     **RN** |   43185 |    1.25 |        90.10 |    1.25 |        90.10 |
## |     **PB** |   42627 |    1.24 |        91.34 |    1.24 |        91.34 |
## |     **AM** |   42110 |    1.22 |        92.56 |    1.22 |        92.56 |
## |     **PI** |   41903 |    1.21 |        93.77 |    1.21 |        93.77 |
## |     **RO** |   40072 |    1.16 |        94.93 |    1.16 |        94.93 |
## |     **AL** |   38645 |    1.12 |        96.05 |    1.12 |        96.05 |
## |     **DF** |   36818 |    1.07 |        97.12 |    1.07 |        97.12 |
## |     **TO** |   28775 |    0.83 |        97.95 |    0.83 |        97.95 |
## |     **SE** |   27654 |    0.80 |        98.76 |    0.80 |        98.76 |
## |     **AC** |   23196 |    0.67 |        99.43 |    0.67 |        99.43 |
## |     **RR** |   10504 |    0.30 |        99.73 |    0.30 |        99.73 |
## |     **AP** |    9190 |    0.27 |       100.00 |    0.27 |       100.00 |
## | **\<NA\>** |       0 |         |              |    0.00 |       100.00 |
## |  **Total** | 3449298 |  100.00 |       100.00 |  100.00 |       100.00 |
```

O resultado demonstra que o estado de SP (frequência relativa 29.2%), MG (frequência relativa 9.71%) e RJ (frequência relativa 7.65%) se destacam com a quantidade de postos de combustíveis pesquisados.

É possível observar que a ordem do resultado da frequência relativa por estado é muito semelhante a análise de do quantitativo de missing da variável Valor_de_Compra. Os únicos estados que sofreram mudanças de posição, entre eles, são AM e PI.

A próxima análise de frequência visa analisar os municípios do Estado do Rio de Janeiro que mais contribuíram para a pesquisa, analisaremos a variável Municipio utilizando a função freq() do pacote summarytools.

```
combustiveis %>%
   dplyr::filter(UF == "RJ") %>%
   dplyr::select(Municipio) %>%
   summarytools::freq(., style = 'rmarkdown', order = "freq", plain.ascii = FALSE)
```

```
## ### Frequencies
## #### combustiveis$Municipio
## **Type:** Character
##
## |                            |   Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |--------------------------------:|-------:|--------:|-------------:|--------:|-------------:|
## |              **RIO DE JANEIRO** |  57233 |  21.691 |       21.691 |  21.691 |       21.691 |
## |            **DUQUE DE CAXIAS** |  17272 |   6.546 |       28.237 |   6.546 |       28.237 |
## |                **SAO GONCALO** |  13901 |   5.268 |       33.506 |   5.268 |       33.506 |
## |                    **NITEROI** |  13648 |   5.173 |       38.679 |   5.173 |       38.679 |
## |                **NOVA IGUACU** |  12856 |   4.872 |       43.551 |   4.872 |       43.551 |
## |                 **PETROPOLIS** |  11401 |   4.321 |       47.872 |   4.321 |       47.872 |
## |       **CAMPOS DOS GOYTACAZES** |   9038 |   3.425 |       51.297 |   3.425 |       51.297 |
## |               **NOVA FRIBURGO** |   8251 |   3.127 |       54.424 |   3.127 |       54.424 |
## |          **SAO JOAO DE MERITI** |   8028 |   3.043 |       57.467 |   3.043 |       57.467 |
## |                **BARRA MANSA** |   8021 |   3.040 |       60.507 |   3.040 |       60.507 |
## |               **BELFORD ROXO** |   7332 |   2.779 |       63.286 |   2.779 |       63.286 |
```

VOLTA REDONDA	6970	2.642	65.927	2.642	65.927
ARARUAMA	6913	2.620	68.548	2.620	68.548
RESENDE	6601	2.502	71.049	2.502	71.049
CABO FRIO	5953	2.256	73.305	2.256	73.305
MARICA	5591	2.119	75.424	2.119	75.424
TERESOPOLIS	5530	2.096	77.520	2.096	77.520
ITABORAI	5405	2.048	79.569	2.048	79.569
SAQUAREMA	5290	2.005	81.574	2.005	81.574
VALENCA	4706	1.784	83.357	1.784	83.357
RIO BONITO	4503	1.707	85.064	1.707	85.064
ANGRA DOS REIS	4235	1.605	86.669	1.605	86.669
MACAE	4185	1.586	88.255	1.586	88.255
ITAGUAI	4173	1.582	89.837	1.582	89.837
ITAPERUNA	4168	1.580	91.416	1.580	91.416
TRES RIOS	4043	1.532	92.949	1.532	92.949
NILOPOLIS	3808	1.443	94.392	1.443	94.392
BARRA DO PIRAI	3761	1.425	95.817	1.425	95.817
SAO FRANCISCO DE ITABAPOANA	3597	1.363	97.181	1.363	97.181
SANTO ANTONIO DE PADUA	3442	1.305	98.485	1.305	98.485
MAGE	2103	0.797	99.282	0.797	99.282
SAPUCAIA	1775	0.673	99.955	0.673	99.955
MESQUITA	119	0.045	100.000	0.045	100.000
\<NA\>	0			0.000	100.000
Total	263852	100.000	100.000	100.000	100.000

É possível observar que a ordem do resultado da frequência relativa por município é muito semelhante a análise de do quantitativo de missing da variável Valor_de_Compra dos 6 primeiros, ocorrendo uma mudança de posição entre São Gonçalo e Niterói. A partir dos próximos, o que se destaca é a cidade de Campos dos Goytacazes. Ele surge como o próximo município que mais contribuiu com postos de combustíveis, contudo, na análise de missing ele surge em décimo lugar. Tal análise é ratificada pela porcentagem de missing deste município. Ele foi o município que possui a menor porcentagem de missing dos municípios pesquisados.

A próxima análise possibilita visualizar a contribuição, por produto, de cada município do estado do Rio de Janeiro para a pesquisa.

```r
df <- combustiveis %>% dplyr::filter(UF == "RJ")

summarytools::ctable(x = df$Municipio,
                     y = df$Produto,
                     prop = "t",
                     style = "rmarkdown")
```

```
## ### Cross-Tabulation, Total Proportions
## #### Municipio * Produto
## **Data Frame:** df
##
## |                            |         |                |                |                |                |                    |
## |----------------------------:|--------:|---------------:|---------------:|---------------:|---------------:|-------------------:|--
## |                            | Produto |         DIESEL |      DIESEL S10 |          ETANOL |        GASOLINA | GASOLINA ADITIVADA |
## |                  Municipio |         |                |                |                |                |                    |
## |              ANGRA DOS REIS |         |   373 (0.141%) |   853 ( 0.323%) |   922 ( 0.349%) |  1170 ( 0.443%) |      861 ( 0.326%) |
## |                   ARARUAMA |         |   652 (0.247%) |  1439 ( 0.545%) |  1666 ( 0.631%) |  1671 ( 0.633%) |     1075 ( 0.407%) |
## |              BARRA DO PIRAI |         |   491 (0.186%) |   566 ( 0.215%) |   904 ( 0.343%) |   905 ( 0.343%) |      580 ( 0.220%) |
## |                 BARRA MANSA |         |  1143 (0.433%) |  1181 ( 0.448%) |  1873 ( 0.710%) |  1896 ( 0.719%) |     1185 ( 0.449%) |
## |                BELFORD ROXO |         |   506 (0.192%) |  1379 ( 0.523%) |  1771 ( 0.671%) |  1811 ( 0.686%) |     1011 ( 0.383%) |
## |                   CABO FRIO |         |   437 (0.166%) |   995 ( 0.377%) |  1320 ( 0.500%) |  1324 ( 0.502%) |     1071 ( 0.406%) |
## |       CAMPOS DOS GOYTACAZES |         |  1009 (0.382%) |  1663 ( 0.630%) |  2171 ( 0.823%) |  2302 ( 0.872%) |     1194 ( 0.453%) |
## |             DUQUE DE CAXIAS |         |  1691 (0.641%) |  2987 ( 1.132%) |  3919 ( 1.485%) |  3972 ( 1.505%) |     2466 ( 0.935%) |
## |                   ITABORAI |         |   770 (0.292%) |  1020 ( 0.387%) |  1142 ( 0.433%) |  1138 ( 0.431%) |      600 ( 0.227%) |
## |                    ITAGUAI |         |   457 (0.173%) |   780 ( 0.296%) |   967 ( 0.366%) |   966 ( 0.366%) |      590 ( 0.224%) |
## |                  ITAPERUNA |         |   703 (0.266%) |   787 ( 0.298%) |   865 ( 0.328%) |   978 ( 0.371%) |      734 ( 0.278%) |
## |                      MACAE |         |   451 (0.171%) |   876 ( 0.332%) |   967 ( 0.366%) |  1015 ( 0.385%) |      635 ( 0.241%) |
## |                       MAGE |         |   301 (0.114%) |   401 ( 0.152%) |   503 ( 0.191%) |   509 ( 0.193%) |      273 ( 0.103%) |
## |                     MARICA |         |   547 (0.207%) |  1035 ( 0.392%) |  1349 ( 0.511%) |  1482 ( 0.562%) |      958 ( 0.363%) |
## |                    MESQUITA |         |     7 (0.003%) |    17 ( 0.006%) |    24 ( 0.009%) |    24 ( 0.009%) |       24 ( 0.009%) |
## |                    NILOPOLIS |         |     8 (0.003%) |   371 ( 0.141%) |  1267 ( 0.480%) |  1268 ( 0.481%) |      736 ( 0.279%) |
## |                     NITEROI |         |   484 (0.183%) |  2600 ( 0.985%) |  3370 ( 1.277%) |  3497 ( 1.325%) |     2463 ( 0.933%) |
## |                NOVA FRIBURGO |         |  1040 (0.394%) |  1157 ( 0.439%) |  2278 ( 0.863%) |  2286 ( 0.866%) |     1436 ( 0.544%) |
## |                   NOVA IGUACU |         |  1262 (0.478%) |  2310 ( 0.875%) |  2875 ( 1.090%) |  2882 ( 1.092%) |     2156 ( 0.817%) |
## |                   PETROPOLIS |         |   966 (0.366%) |  1749 ( 0.663%) |  2825 ( 1.071%) |  2893 ( 1.096%) |     2219 ( 0.841%) |
## |                      RESENDE |         |   498 (0.189%) |  1166 ( 0.442%) |  1581 ( 0.599%) |  1582 ( 0.600%) |     1201 ( 0.455%) |
## |                   RIO BONITO |         |   549 (0.208%) |   778 ( 0.295%) |  1111 ( 0.421%) |  1160 ( 0.440%) |      857 ( 0.325%) |
## |                RIO DE JANEIRO |         |  2994 (1.135%) |  9121 ( 3.457%) | 13877 ( 5.259%) | 14081 ( 5.337%) |    10762 ( 4.079%) |
```

```
## |      SANTO ANTONIO DE PADUA |         |   594 (0.225%) |   562 ( 0.213%) |   813 ( 0.308%) |   811 ( 0.307%) |      590 ( 0.224%) |
## | SAO FRANCISCO DE ITABAPOANA |         |   867 (0.329%) |   520 ( 0.197%) |   917 ( 0.348%) |   929 ( 0.352%) |      270 ( 0.102%) |
## |                 SAO GONCALO |         |  1669 (0.633%) |  2236 ( 0.847%) |  3271 ( 1.240%) |  3281 ( 1.244%) |     2137 ( 0.810%) |
## |          SAO JOAO DE MERITI |         |   834 (0.316%) |  1351 ( 0.512%) |  1765 ( 0.669%) |  1829 ( 0.693%) |     1359 ( 0.515%) |
## |                   SAPUCAIA |         |   440 (0.167%) |   317 ( 0.120%) |   458 ( 0.174%) |   477 ( 0.181%) |       83 ( 0.031%) |
## |                  SAQUAREMA |         |   667 (0.253%) |   934 ( 0.354%) |  1242 ( 0.471%) |  1247 ( 0.473%) |      917 ( 0.348%) |
## |                TERESOPOLIS |         |   222 (0.084%) |  1155 ( 0.438%) |  1535 ( 0.582%) |  1543 ( 0.585%) |      924 ( 0.350%) |
## |                  TRES RIOS |         |   534 (0.202%) |   734 ( 0.278%) |   881 ( 0.334%) |   975 ( 0.370%) |      644 ( 0.244%) |
## |                    VALENCA |         |   548 (0.208%) |   924 ( 0.350%) |  1143 ( 0.433%) |  1184 ( 0.449%) |      801 ( 0.304%) |
## |               VOLTA REDONDA |         |   773 (0.293%) |  1211 ( 0.459%) |  1563 ( 0.592%) |  1664 ( 0.631%) |     1257 ( 0.476%) |
## |                      Total |         | 24487 (9.281%) | 45175 (17.121%) | 63135 (23.928%) | 64752 (24.541%) |    44069 (16.702%) | 2
```

```
rm("df")
```

A próxima análise possibilita visualizar a contribuição, por produto, para cada município na pesquisa.

```
df <- combustiveis %>% dplyr::filter(UF == "RJ")

summarytools::ctable(x = df$Municipio,
                     y = df$Produto,
                     prop = "r")
```

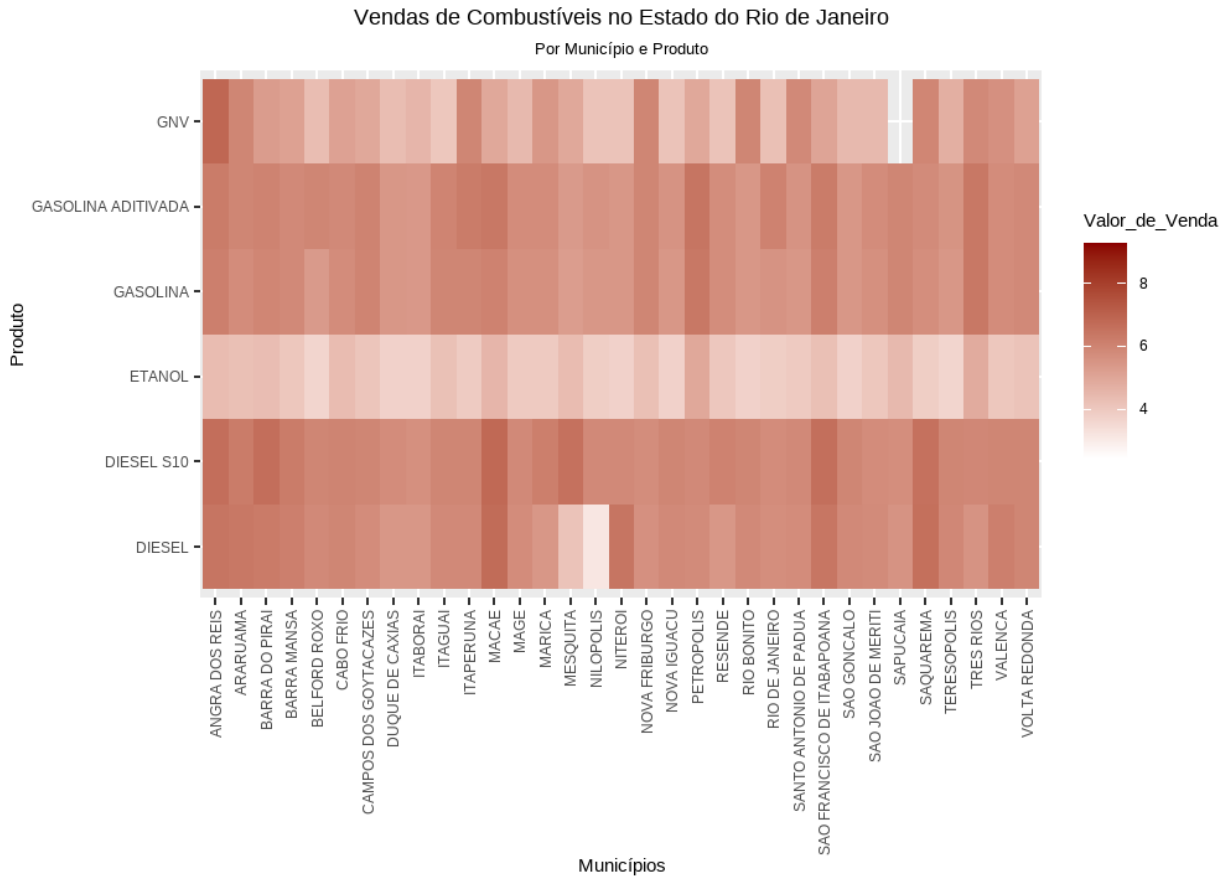
```
## Cross-Tabulation, Row Proportions
## Municipio * Produto
## Data Frame: df
##
## ---------------------------- --------- --------------- --------------- --------------- --------------- -------------------- ----------
##                              Produto           DIESEL      DIESEL S10          ETANOL        GASOLINA   GASOLINA ADITIVADA
##                    Municipio
##                ANGRA DOS REIS               373 ( 8.8%)     853 (20.1%)     922 (21.8%)    1170 (27.6%)          861 (20.3%)      56 (
##                     ARARUAMA               652 ( 9.4%)    1439 (20.8%)    1666 (24.1%)    1671 (24.2%)         1075 (15.6%)     410 (
##                BARRA DO PIRAI               491 (13.1%)     566 (15.0%)     904 (24.0%)     905 (24.1%)          580 (15.4%)     315 (
##                   BARRA MANSA              1143 (14.3%)    1181 (14.7%)    1873 (23.4%)    1896 (23.6%)         1185 (14.8%)     743 (
##                  BELFORD ROXO               506 ( 6.9%)    1379 (18.8%)    1771 (24.2%)    1811 (24.7%)         1011 (13.8%)     854 (1
##                     CABO FRIO               437 ( 7.3%)     995 (16.7%)    1320 (22.2%)    1324 (22.2%)         1071 (18.0%)     806 (1
##          CAMPOS DOS GOYTACAZES              1009 (11.2%)    1663 (18.4%)    2171 (24.0%)    2302 (25.5%)         1194 (13.2%)     699 (
##                DUQUE DE CAXIAS              1691 ( 9.8%)    2987 (17.3%)    3919 (22.7%)    3972 (23.0%)         2466 (14.3%)    2237 (1
##                      ITABORAI               770 (14.2%)    1020 (18.9%)    1142 (21.1%)    1138 (21.1%)          600 (11.1%)     735 (1
```

```
##                          ITAGUAI              457 (11.0%)     780 (18.7%)     967 (23.2%)     966 (23.1%)          590 (14.1%)     413 (
##                        ITAPERUNA              703 (16.9%)     787 (18.9%)     865 (20.8%)     978 (23.5%)          734 (17.6%)     101 (
##                            MACAE              451 (10.8%)     876 (20.9%)     967 (23.1%)    1015 (24.3%)          635 (15.2%)     241 (
##                             MAGE              301 (14.3%)     401 (19.1%)     503 (23.9%)     509 (24.2%)          273 (13.0%)     116 (
##                           MARICA              547 ( 9.8%)    1035 (18.5%)    1349 (24.1%)    1482 (26.5%)          958 (17.1%)     220 (
##                         MESQUITA                7 ( 5.9%)      17 (14.3%)      24 (20.2%)      24 (20.2%)           24 (20.2%)      23 (1
##                        NILOPOLIS                8 ( 0.2%)     371 ( 9.7%)    1267 (33.3%)    1268 (33.3%)          736 (19.3%)     158 (
##                         NITEROI              484 ( 3.5%)    2600 (19.1%)    3370 (24.7%)    3497 (25.6%)         2463 (18.0%)    1234 (
##                    NOVA FRIBURGO             1040 (12.6%)    1157 (14.0%)    2278 (27.6%)    2286 (27.7%)         1436 (17.4%)      54 (
##                      NOVA IGUACU             1262 ( 9.8%)    2310 (18.0%)    2875 (22.4%)    2882 (22.4%)         2156 (16.8%)    1371 (1
##                       PETROPOLIS              966 ( 8.5%)    1749 (15.3%)    2825 (24.8%)    2893 (25.4%)         2219 (19.5%)     749 (
##                          RESENDE              498 ( 7.5%)    1166 (17.7%)    1581 (24.0%)    1582 (24.0%)         1201 (18.2%)     573 (
##                       RIO BONITO              549 (12.2%)     778 (17.3%)    1111 (24.7%)    1160 (25.8%)          857 (19.0%)      48 (
##                    RIO DE JANEIRO             2994 ( 5.2%)    9121 (15.9%)   13877 (24.2%)   14081 (24.6%)        10762 (18.8%)    6398 (1
##            SANTO ANTONIO DE PADUA              594 (17.3%)     562 (16.3%)     813 (23.6%)     811 (23.6%)          590 (17.1%)      72 (
##       SAO FRANCISCO DE ITABAPOANA              867 (24.1%)     520 (14.5%)     917 (25.5%)     929 (25.8%)          270 ( 7.5%)      94 (
##                         SAO GONCALO             1669 (12.0%)    2236 (16.1%)    3271 (23.5%)    3281 (23.6%)         2137 (15.4%)    1307 (
##                SAO JOAO DE MERITI              834 (10.4%)    1351 (16.8%)    1765 (22.0%)    1829 (22.8%)         1359 (16.9%)     890 (1
##                         SAPUCAIA              440 (24.8%)     317 (17.9%)     458 (25.8%)     477 (26.9%)           83 ( 4.7%)       0 (
##                        SAQUAREMA              667 (12.6%)     934 (17.7%)    1242 (23.5%)    1247 (23.6%)          917 (17.3%)     283 (
##                      TERESOPOLIS              222 ( 4.0%)    1155 (20.9%)    1535 (27.8%)    1543 (27.9%)          924 (16.7%)     151 (
##                        TRES RIOS              534 (13.2%)     734 (18.2%)     881 (21.8%)     975 (24.1%)          644 (15.9%)     275 (
##                          VALENCA              548 (11.6%)     924 (19.6%)    1143 (24.3%)    1184 (25.2%)          801 (17.0%)     106 (
##                    VOLTA REDONDA              773 (11.1%)    1211 (17.4%)    1563 (22.4%)    1664 (23.9%)         1257 (18.0%)     502 (
##                            Total            24487 ( 9.3%)   45175 (17.1%)   63135 (23.9%)   64752 (24.5%)        44069 (16.7%)   22234 (
## ---------------------------- --------- --------------- --------------- --------------- --------------- -------------------- ----------
```

```
rm("df")
```

A análise a seguir visa apresentar um mapa de calor a partir do valor de venda dos produtos por município do RJ.

```
combustiveis %>%
   dplyr::filter(UF == "RJ") %>%
   ggplot(aes(x = Municipio,
              y = Produto,
              fill = Valor_de_Venda)) +
   geom_tile() +
   labs(x = "Municípios",
        y = "Produto",
        title = "Vendas de Combustíveis no Estado do Rio de Janeiro",
        subtitle = "Por Município e Produto") +
   theme(axis.text.x = element_text(angle = 90,
                                    hjust = 1,
                                    vjust = 0.5),
         plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5,
                                      size = 10)) +
   scale_fill_gradient(low = "white", high = "darkred")
```



Gerando dados calculados

Para a realização da próxima etapa da análise, a qual envolve o *merge* entre as bases de dados até aqui apresentada, será necessário realizar filtros e cálculos.

A variável taxa de câmbio da base de dados cambio é composta por valores cotados diariamente.

Para uma uniformização mensal dos dados, foi realizado cálculos estatísticos do câmbio praticado mensalmente a partir do valor diário.

```
cambio_calc <- cambio %>%
   dplyr::group_by(Data) %>%
   dplyr::summarise(Cambio = mean(Taxa_Cambio),.groups = 'drop')
#   dplyr::summarise(Cambio = mean(Taxa_Cambio),
#                    Max_Cambio = max(Taxa_Cambio),
#                    Min_Cambio = min(Taxa_Cambio),
#                    SD_Cambio = sd(Taxa_Cambio),.groups = 'drop')
```

Foi gerada uma nova base de dados, tendo como descrição o termo calc de calculada. Utilizou-se a função group_by() do pacote dplyr para agrupar os dados pela data e posteriormente a função summarise() do pacote dplyr para os cálculos.

A variável PPI_DC da base de dados ppi é fruto da análise semanal. As datas semanais foram tratadas para uniformização mensal dos dados. Foram realizados cálculos estatísticos a partir dos valores semanais para o respectivo cálculo mensal.

```
ppi_calc <- ppi %>%
   dplyr::group_by(Data) %>%
   dplyr::summarise(PPI_DC = mean(PPI_DC), .groups = 'drop')
#   dplyr::summarise(PPI_DC = mean(PPI_DC),
#                    Max_PPI_DC = max(PPI_DC),
#                    Min_PPI_DC = min(PPI_DC),
#                    SD_PPI_DC = sd(PPI_DC),.groups = 'drop')
```

Foi gerada uma nova base de dados, tendo como descrição o termo calc de calculada. Utilizou-se a função group_by() do pacote dplyr para agrupar os dados pela data e posteriormente a função summarise() do pacote dplyr para os cálculos.

Foi constatado durante as análises que no mês 09 de 2020 não ocorreu nenhuma pesquisa nos postos de combustíveis do Brasil. Por este motivo, foi necessário remover o respectivo mês das bases cambio_calc, ppi_calc e brent.

```
cambio_calc <- cambio_calc %>% dplyr::filter(!Data == "2020-09-01")
ppi_calc <- ppi_calc %>% dplyr::filter(!Data == "2020-09-01")
brent <- brent %>% dplyr::filter(!Data == "2020-09-01")
```

Merge das bases de dados

Após a realização do filtro e os devidos cálculos, realizar-se-á nesta etapa o *merge* das bases de dados.

```
combustiveis <- base::merge(combustiveis,
                            brent,
                            by = "Data",
                            all = TRUE)
combustiveis <- base::merge(combustiveis,
                            cambio_calc,
                            by = "Data",
                            all = TRUE)
combustiveis <- base::merge(combustiveis,
                            ppi_calc,
                            by = "Data",
                            all = TRUE)

combustiveis <- combustiveis %>%
   dplyr::mutate(Brent_Real_Barril = `Brent USD/Barril` * Cambio)

combustiveis <- combustiveis %>%
   dplyr::select(-`Brent USD/Barril`)
```

Calculando as estatísticas descritiva dos dados

Uma outra forma de visualização das estatísticas descritivas dos dados é apresentada abaixo considerando o ano da análise como filtro. As seguintes funções e pacotes foram utilizados:

- Função mutate() do pacote dplyr para setar uma variável ano;
- Função filter() do pacote dplyr para filtrar pela variável ano;
- Função group_by() do pacote dplyr para agrupar por ano;
- Função descr() do pacote summarytools para cálculo da estatística descritiva.

Abaixo é apresentado as possibilidades de análise da estatística descritiva do cambio.

```
ano_filtro = "2020"
mes_filtro = "01"

# Estatística descritiva anual do cambio
cambio %>%
   dplyr::mutate(ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro) %>%
   dplyr::group_by(ano) %>%
   summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
## #### Taxa_Cambio by ano
## **Data Frame:** cambio
## **N:** 251
```

```
##
## |            | ano = 2020 |
## |----------------:|-----------:|
## |        **Mean** |       5.16 |
## |     **Std.Dev** |       0.47 |
## |         **Min** |       4.02 |
## |          **Q1** |       5.07 |
## |      **Median** |       5.28 |
## |          **Q3** |       5.46 |
## |         **Max** |       5.94 |
## |         **MAD** |       0.29 |
## |         **IQR** |       0.39 |
## |          **CV** |       0.09 |
## |    **Skewness** |      -1.01 |
## | **SE.Skewness** |       0.15 |
## |    **Kurtosis** |       0.01 |
## |     **N.Valid** |     251.00 |
## |   **Pct.Valid** |     100.00 |
```

```
# Estatística descritiva mensal do cambio
cambio %>%
   dplyr::mutate(mes = format(Data, "%m"), ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro, mes == mes_filtro) %>%
   group_by(ano, mes) %>%
   summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
## #### cambio$Taxa_Cambio
## **Group:** ano = 2020, mes = 01
## **N:** 22
##
## |            | Taxa_Cambio |
## |----------------:|------------:|
## |        **Mean** |        4.15 |
## |     **Std.Dev** |        0.07 |
## |         **Min** |        4.02 |
## |          **Q1** |        4.07 |
## |      **Median** |        4.17 |
## |          **Q3** |        4.20 |
## |         **Max** |        4.27 |
## |         **MAD** |        0.06 |
## |         **IQR** |        0.12 |
## |          **CV** |        0.02 |
## |    **Skewness** |       -0.24 |
## | **SE.Skewness** |        0.49 |
## |    **Kurtosis** |       -1.17 |
## |     **N.Valid** |       22.00 |
## |   **Pct.Valid** |      100.00 |
```

Abaixo é apresentado as possibilidades de análise da estatística descritiva dos combustíveis.

```
ano_filtro = "2020"
mes_filtro = "01"
```

```r
# Estatística descritiva anual da gasolina no Brasil
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA") %>%
   dplyr::select(Valor_de_Venda, Valor_de_Compra, Data) %>%
   dplyr::mutate(ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro) %>%
   group_by(ano) %>%
   summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
## #### combustiveis
## **Group:** ano = 2020
## **N:** 214043
##
## |            | Valor_de_Compra | Valor_de_Venda |
## |----------------:|----------------:|---------------:|
## |        **Mean** |            3.81 |           4.28 |
## |     **Std.Dev** |            0.39 |           0.40 |
## |         **Min** |            2.63 |           2.87 |
## |          **Q1** |            3.52 |           3.99 |
## |      **Median** |            3.86 |           4.30 |
## |          **Q3** |            4.07 |           4.58 |
## |         **Max** |            5.00 |           5.90 |
## |         **MAD** |            0.39 |           0.44 |
## |         **IQR** |            0.55 |           0.59 |
## |          **CV** |            0.10 |           0.09 |
## |    **Skewness** |           -0.13 |           0.01 |
## | **SE.Skewness** |            0.01 |           0.01 |
## |    **Kurtosis** |           -0.51 |          -0.40 |
## |     **N.Valid** |        49906.00 |      214043.00 |
## |   **Pct.Valid** |           23.32 |         100.00 |
```

```r
# Estatística descritiva anual da gasolina no RJ
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Valor_de_Venda, Valor_de_Compra, Data) %>%
   dplyr::mutate(ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro) %>%
   group_by(ano) %>%
   summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
## #### combustiveis
## **Group:** ano = 2020
## **N:** 15595
##
## |            | Valor_de_Compra | Valor_de_Venda |
## |----------------:|----------------:|---------------:|
## |        **Mean** |            4.22 |           4.76 |
## |     **Std.Dev** |            0.35 |           0.31 |
## |         **Min** |            2.63 |           3.86 |
## |          **Q1** |            3.94 |           4.55 |
```

```
## |      **Median** |            4.27 |           4.80 |
## |          **Q3** |            4.54 |           5.00 |
## |         **Max** |            5.00 |           5.89 |
## |         **MAD** |            0.43 |           0.30 |
## |         **IQR** |            0.60 |           0.45 |
## |          **CV** |            0.08 |           0.06 |
## |    **Skewness** |           -0.39 |          -0.25 |
## | **SE.Skewness** |            0.03 |           0.02 |
## |    **Kurtosis** |           -1.03 |          -0.33 |
## |     **N.Valid** |         6050.00 |       15595.00 |
## |   **Pct.Valid** |           38.79 |         100.00 |
```

```r
# Estatística descritiva mensal da gasolina no Brasil
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA") %>%
   dplyr::select(Valor_de_Venda, Valor_de_Compra, Data) %>%
   dplyr::mutate(mes = format(Data, "%m"), ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro, mes == mes_filtro) %>%
   group_by(ano, mes) %>% summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
## #### combustiveis
## **Group:** ano = 2020, mes = 01
## **N:** 25996
##
## |            | Valor_de_Compra | Valor_de_Venda |
## |----------------:|----------------:|---------------:|
## |        **Mean** |            4.13 |           4.62 |
## |     **Std.Dev** |            0.24 |           0.28 |
## |         **Min** |            3.54 |           3.80 |
## |          **Q1** |            3.96 |           4.40 |
## |      **Median** |            4.09 |           4.60 |
## |          **Q3** |            4.27 |           4.80 |
## |         **Max** |            5.00 |           5.90 |
## |         **MAD** |            0.22 |           0.30 |
## |         **IQR** |            0.31 |           0.40 |
## |          **CV** |            0.06 |           0.06 |
## |    **Skewness** |            0.57 |           0.18 |
## | **SE.Skewness** |            0.03 |           0.02 |
## |    **Kurtosis** |           -0.36 |          -0.27 |
## |     **N.Valid** |         9337.00 |       25996.00 |
## |   **Pct.Valid** |           35.92 |         100.00 |
```

```r
# Estatística descritiva mensal da gasolina no RJ
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Valor_de_Venda, Valor_de_Compra, Data) %>%
   dplyr::mutate(mes = format(Data, "%m"), ano = format(Data, "%Y")) %>%
   dplyr::filter(ano == ano_filtro, mes == mes_filtro) %>%
   group_by(ano, mes) %>%
   summarytools::descr(., style = 'rmarkdown')
```

```
## ### Descriptive Statistics
```

```
## #### combustiveis
## **Group:** ano = 2020, mes = 01
## **N:** 1887
##
## |            | Valor_de_Compra | Valor_de_Venda |
## |----------------:|----------------:|---------------:|
## |        **Mean** |            4.57 |           5.05 |
## |     **Std.Dev** |            0.10 |           0.17 |
## |         **Min** |            4.30 |           4.50 |
## |          **Q1** |            4.52 |           4.90 |
## |      **Median** |            4.58 |           5.00 |
## |          **Q3** |            4.63 |           5.18 |
## |         **Max** |            5.00 |           5.89 |
## |         **MAD** |            0.08 |           0.15 |
## |         **IQR** |            0.11 |           0.28 |
## |          **CV** |            0.02 |           0.03 |
## |    **Skewness** |           -0.42 |           0.48 |
## | **SE.Skewness** |            0.08 |           0.06 |
## |    **Kurtosis** |            0.35 |           1.04 |
## |     **N.Valid** |         1007.00 |        1887.00 |
## |   **Pct.Valid** |           53.37 |         100.00 |
```

```
rm(ano_filtro)
rm(mes_filtro)
```

Análise descritiva e de histogramas das variáveis contínuas

A partir desta etapa, a análise será realizada sobre a base de dados criada com intuito de compreender a relação do preço da gasolina no municípios dos Rio de Janeiro pesquisados. Portanto, a base a ser analisada será a gasolina_rj.

Análise descritiva das variáveis contínuas

Para a variável Valor_de_Venda, podemos analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

Para realizar essa análise, utilizar-se-á a função descr do pacote summarytools, e posteriormente realizar a leitura desses dados.

```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Valor_de_Venda) %>%
   summarytools::descr()
```

```
## Descriptive Statistics
## combustiveis$Valor_de_Venda
## N: 64752
##
##                     Valor_de_Venda
```

```
## ----------------- ----------------
##              Mean             5.82
##           Std.Dev             1.00
##               Min             3.86
##                Q1             5.00
##            Median             5.59
##                Q3             6.40
##               Max             8.99
##               MAD             0.91
##               IQR             1.40
##                CV             0.17
##          Skewness             0.64
##       SE.Skewness             0.01
##          Kurtosis            -0.46
##           N.Valid         64752.00
##         Pct.Valid           100.00
```

É possível ver pelo critério de skewness, que o valor está entre 0.5 e 1 para assimetria, nos permitindo interpretar que esta distribução possui assimetria moderada, com cauda à direita

Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor levemente superior, o que aponta que os valores mais distantes do centro da distribuição puxam o valor da média pra cima.

Já a mediana por ser uma estatística de ordem, não é sensível a dados que apresentam alto valor na distribuição, o que é reforçado por seu valor levemente mais baixo que a média.

Se houve-se outliers nesta distribuição a média se descolaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

Para a variável Cambio, podemos analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Cambio) %>%
   summarytools::descr()
```

```
## Descriptive Statistics
## combustiveis$Cambio
## N: 64752
##
##                      Cambio
## ----------------- ----------
##              Mean       5.13
##           Std.Dev       0.32
##               Min       4.15
##                Q1       4.94
##            Median       5.16
##                Q3       5.29
##               Max       5.65
##               MAD       0.30
##               IQR       0.35
##                CV       0.06
##          Skewness      -0.79
```

```
##       SE.Skewness       0.01
##          Kurtosis       1.28
##           N.Valid   64752.00
##         Pct.Valid     100.00
```

É possível ver pelo critério de skewness, que o valor está entre -0.5 e -1 para assimetria, nos permitindo interpretar que esta distribução possui assimetria moderada, com cauda à esquerda.

Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor levemente inferior.

Se houve-se outliers representativos nesta distribuição a média se descolaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

Para a variável Brent R$/Barril, podemos analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Brent_Real_Barril) %>%
   summarytools::descr()
```

```
## Descriptive Statistics
## combustiveis$Brent_Real_Barril
## N: 64752
##
##                     Brent_Real_Barril
## ----------------- -------------------
##              Mean              384.59
##           Std.Dev              123.22
##               Min               97.87
##                Q1              293.33
##            Median              393.69
##                Q3              462.64
##               Max              619.51
##               MAD              102.22
##               IQR              169.31
##                CV                0.32
##          Skewness               -0.38
##       SE.Skewness                0.01
##          Kurtosis               -0.28
##           N.Valid            64752.00
##         Pct.Valid              100.00
```

É possível ver pelo critério de skewness, que o valor está entre -0.5 e 0.5 para assimetria, nos permite interpretar que esta distribução possui assimetria fraca, com cauda leve a esquerda.

Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor levemente inferior.

Se houve-se outliers representativos nesta distribuição a média se descolaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

Para a variável PPI_DC, podemos analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

```r
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(PPI_DC) %>%
   summarytools::descr()
```

```
## Descriptive Statistics
## combustiveis$PPI_DC
## N: 64752
##
##                      PPI_DC
## ----------------- ----------
##              Mean       2.92
##           Std.Dev       0.87
##               Min       0.94
##                Q1       2.40
##            Median       3.03
##                Q3       3.50
##               Max       4.78
##               MAD       0.70
##               IQR       1.09
##                CV       0.30
##          Skewness      -0.26
##       SE.Skewness       0.01
##          Kurtosis      -0.07
##           N.Valid   64752.00
##         Pct.Valid     100.00
```

É possível ver pelo critério de skewness, que o valor está entre -0.5 e 0.5 para assimetria, nos permite interpretar que esta distribução possui assimetria fraca, com cauda leve a esquerda.

Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor levemente inferior.

Se houve-se outliers representativos nesta distribuição a média se descolaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

Histogramas das variáveis contínuas

Um conceito para esta etapa da análise é fundamental compreendermos.

O que é uma distribuição normal?

Podemos conceituar como sendo uma distribuição estatística no formato de um sino e simétrica em relação a média.

O que simboliza o formato de um sino?

A maioria dos dados estão concentrados no centro, diminuindo a quantidade destes dados em ambas as direções.

O que é a simetria em relação a média?

O termo simetria em relação a média é nada mais do que os valores da mediana e moda concidirem com o valor da média.

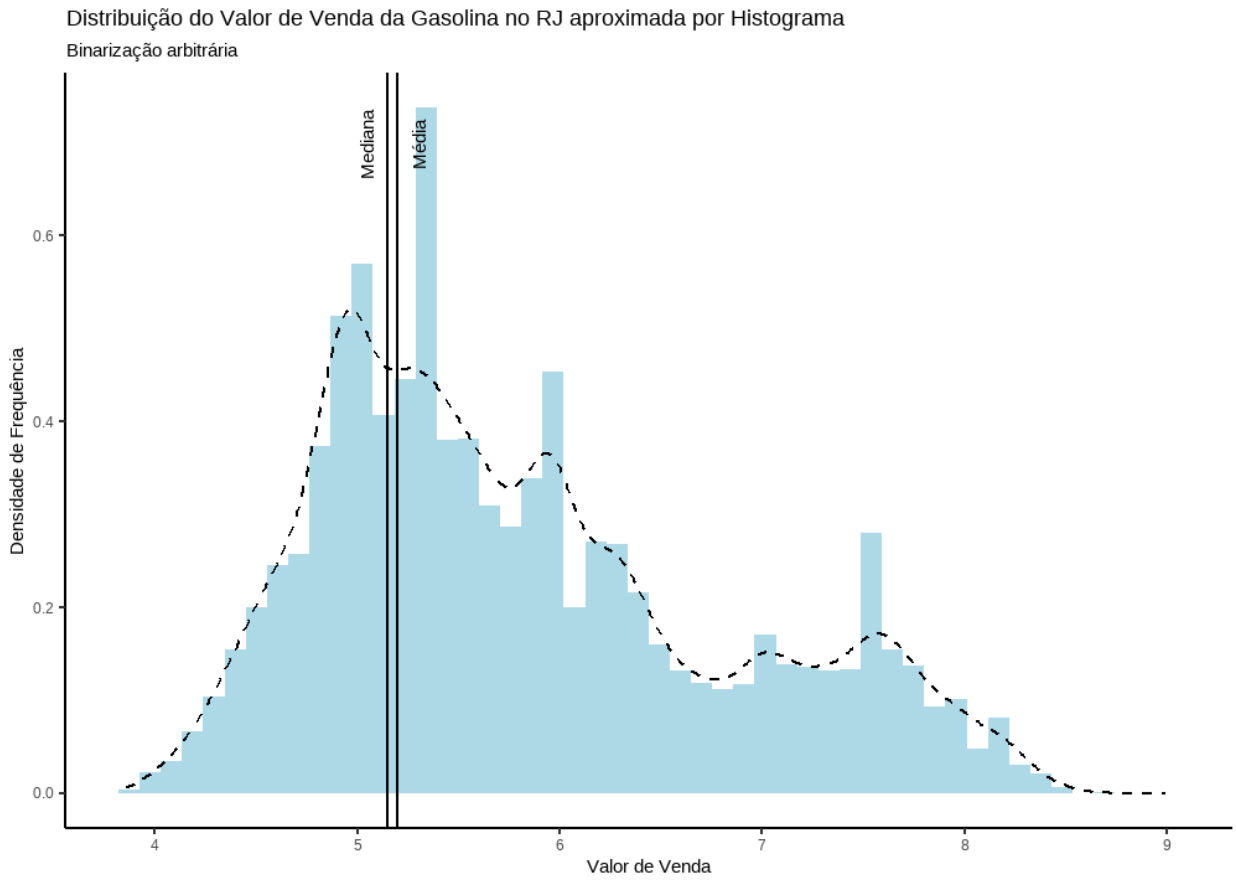
Portanto, pode-se dizer que para que haja simetria e o formato de sino é necessário que a média, mediana e moda possuam o mesmo valor e o quantitativo de valores do lado esquerdo e direito da média são iguais.

A partir do histograma apresentado abaixo é possível compreender o valor de venda da gasolina praticado durante o período de 2020 e 02/2024 nos municípios do Estado do RJ.

A escolha do número de bins para o histograma foi uma escolha arbitrária, assumindo a ocultação da realidade dos dados.

A primeira variável a ser gerado do histograma é o Valor de Venda da Gasolina nos municípios do Rio de Janeiro.

```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Valor_de_Venda) %>%
   ggplot(aes(x=Valor_de_Venda)) +
   geom_histogram(aes(y = after_stat(density)) , bins=50, fill = 'lightblue') +
   xlab('Valor de Venda') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Valor de Venda da Gasolina no RJ aproximada por Histograma",
        subtitle = "Binarização arbitrária") +
   geom_vline(xintercept=c(median(combustiveis$Valor_de_Venda),
                           mean(combustiveis$Valor_de_Venda))) +
   annotate("text", x=median(combustiveis$Valor_de_Venda) +
               -0.15, y=0.7, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Valor_de_Venda) +
               0.15, y=0.7, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```



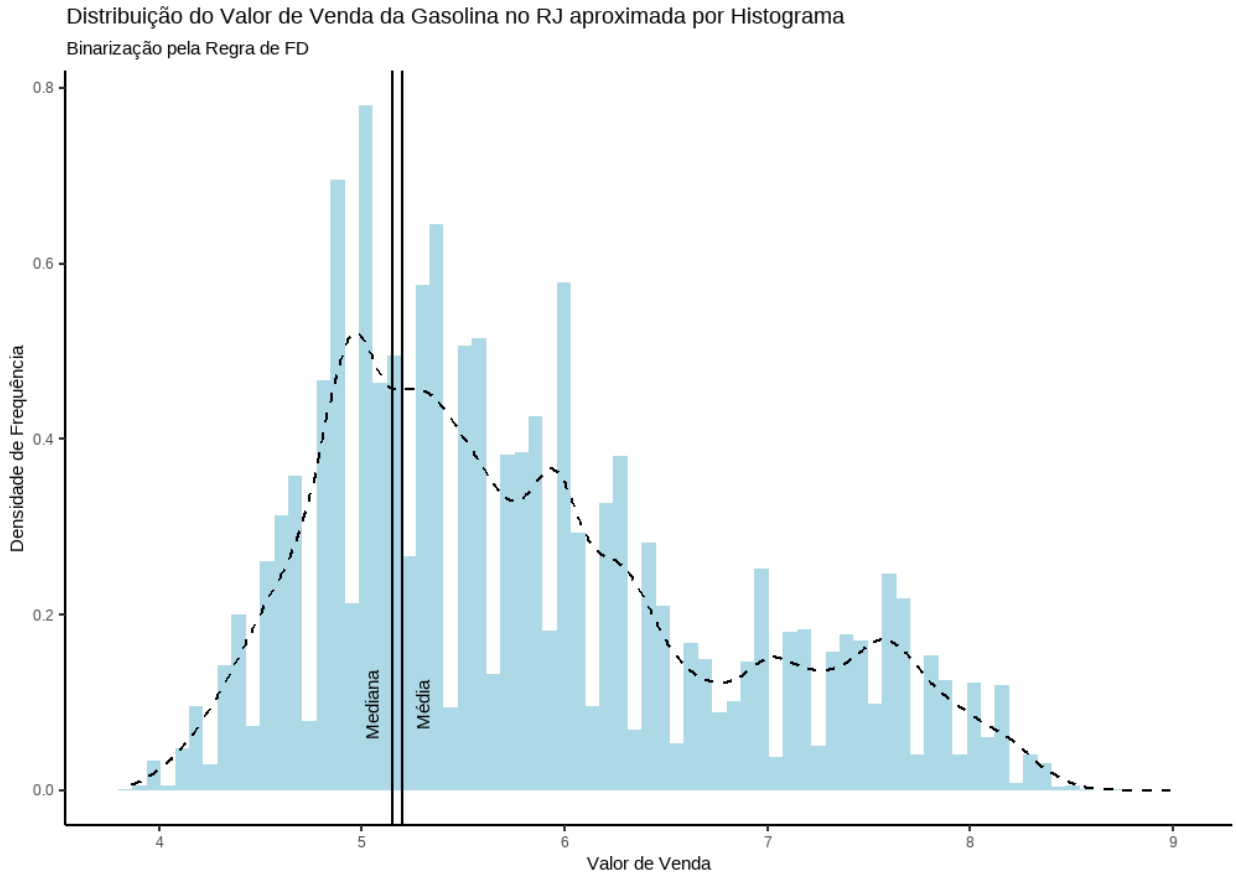
É possível observar nesta visualização que leva a conclusões semelhantes a análise das estatísticas descritas.

1. Assimetria moderada, com cauda à direita.
2. Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor superior, o que aponta que os valores mais distantes do centro da distribuição puxam o valor da média pra cima.

Em relação a binarização, é possível considerar outras regras de binarização levando em consideração regras disponíveis na literatura, como a regra de Freedman-Diaconis, bem como a regra de Sturge, como segue:

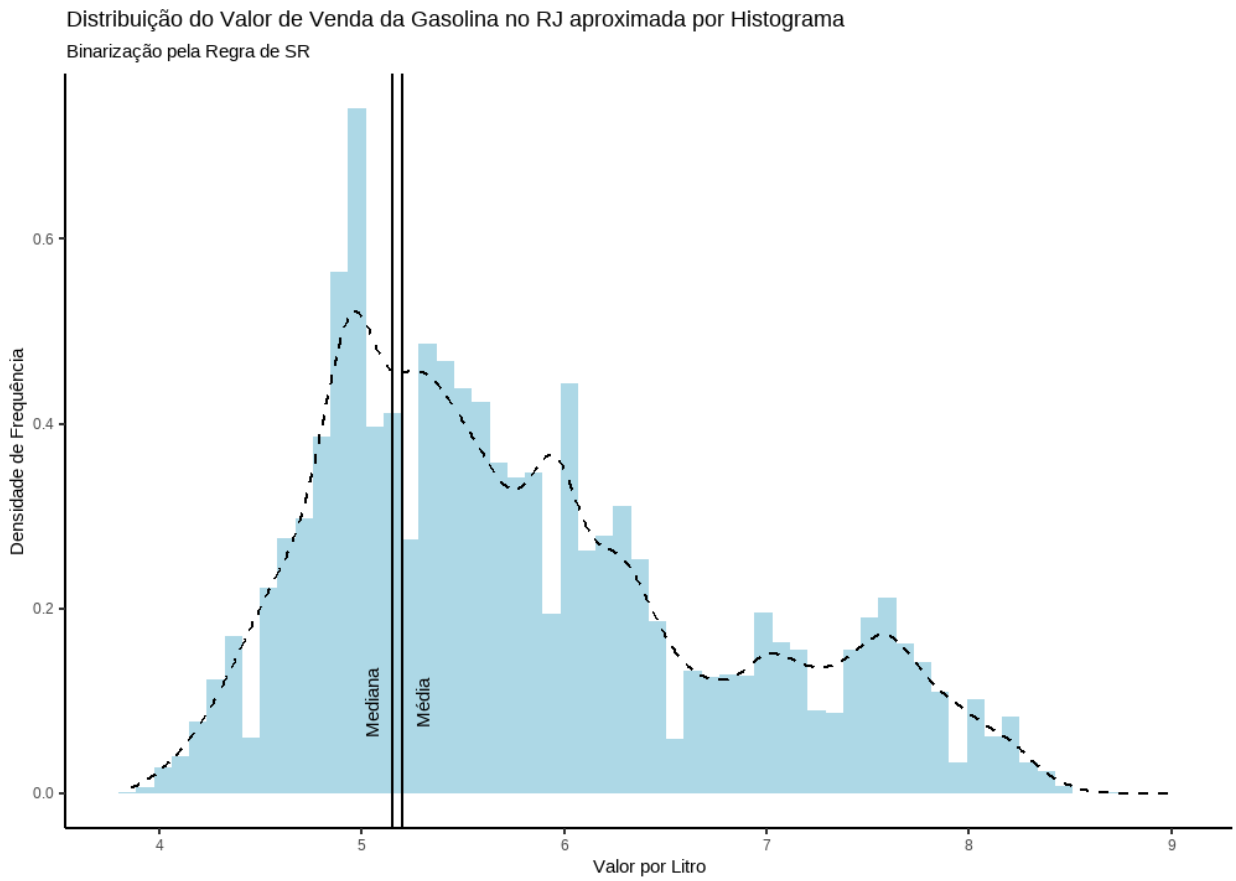
```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Valor_de_Venda) %>%
   ggplot(aes(x=Valor_de_Venda)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=fd,
                  fill = 'lightblue') +
   xlab('Valor de Venda') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Valor de Venda da Gasolina no RJ aproximada por Histograma",
        subtitle = "Binarização pela Regra de FD") +
   geom_vline(xintercept=c(median(combustiveis$Valor_de_Venda),
                           mean(combustiveis$Valor_de_Venda))) +
   annotate("text", x=median(combustiveis$Valor_de_Venda) +
```

```
            -0.15, y=0.1, label="Mediana", angle=90) +
  annotate("text", x=mean(combustiveis$Valor_de_Venda) +
            0.15, y=0.1, label="Média", angle=90) +
  geom_density(linetype = 2) +
  theme_classic()
```



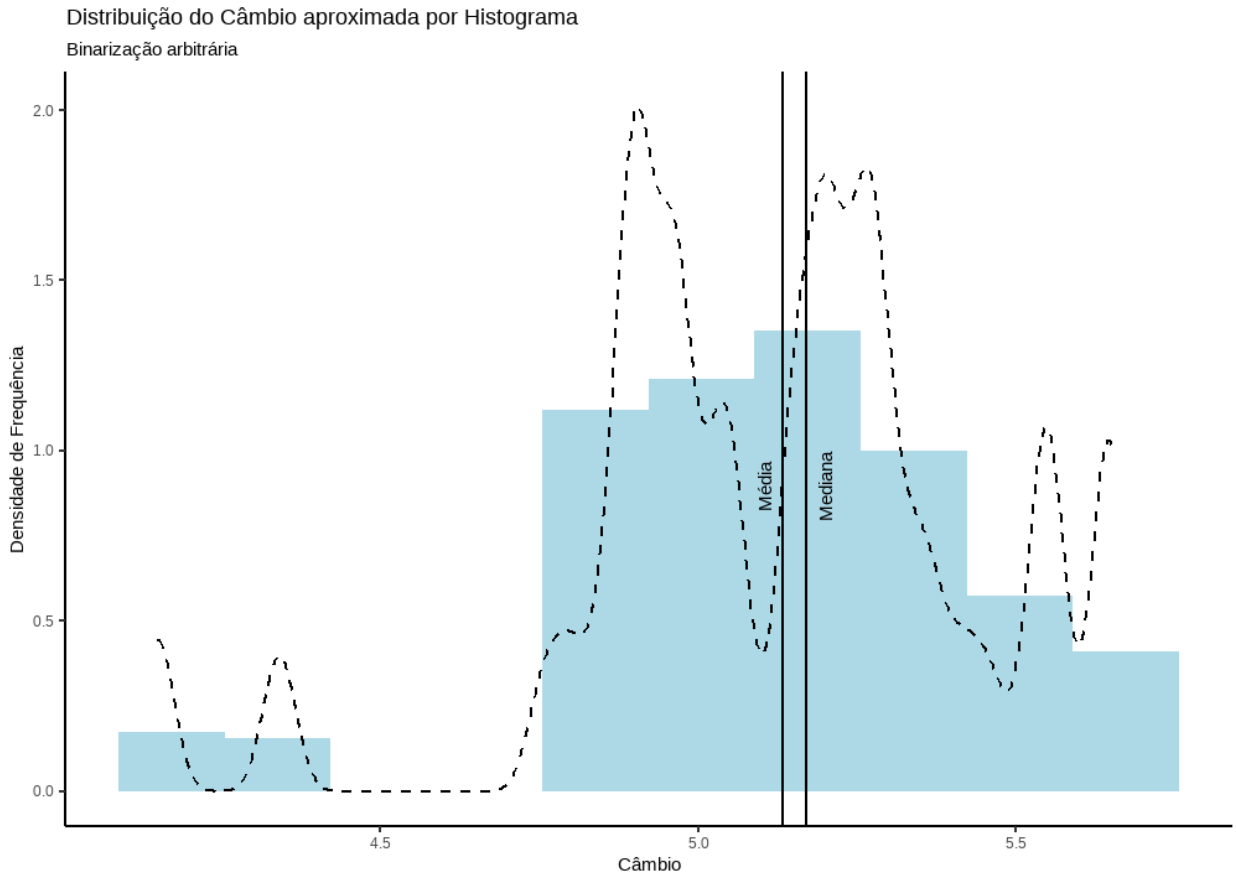

```
combustiveis %>%
  dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
  dplyr::select(Valor_de_Venda) %>%
  ggplot(aes(x=Valor_de_Venda)) +
  geom_histogram(aes(y = after_stat(density)) ,
                 binwidth=sr,
                 fill = 'lightblue') +
  xlab('Valor por Litro') +
  ylab('Densidade de Frequência') +
  labs(title = "Distribuição do Valor de Venda da Gasolina no RJ aproximada por Histograma",
       subtitle = "Binarização pela Regra de SR") +
  geom_vline(xintercept=c(median(combustiveis$Valor_de_Venda),
                          mean(combustiveis$Valor_de_Venda))) +
  annotate("text", x=median(combustiveis$Valor_de_Venda) +
            -0.15, y=0.1, label="Mediana", angle=90) +
  annotate("text", x=mean(combustiveis$Valor_de_Venda) +
            0.15, y=0.1, label="Média", angle=90) +
```

```
   geom_density(linetype = 2) +
   theme_classic()
```

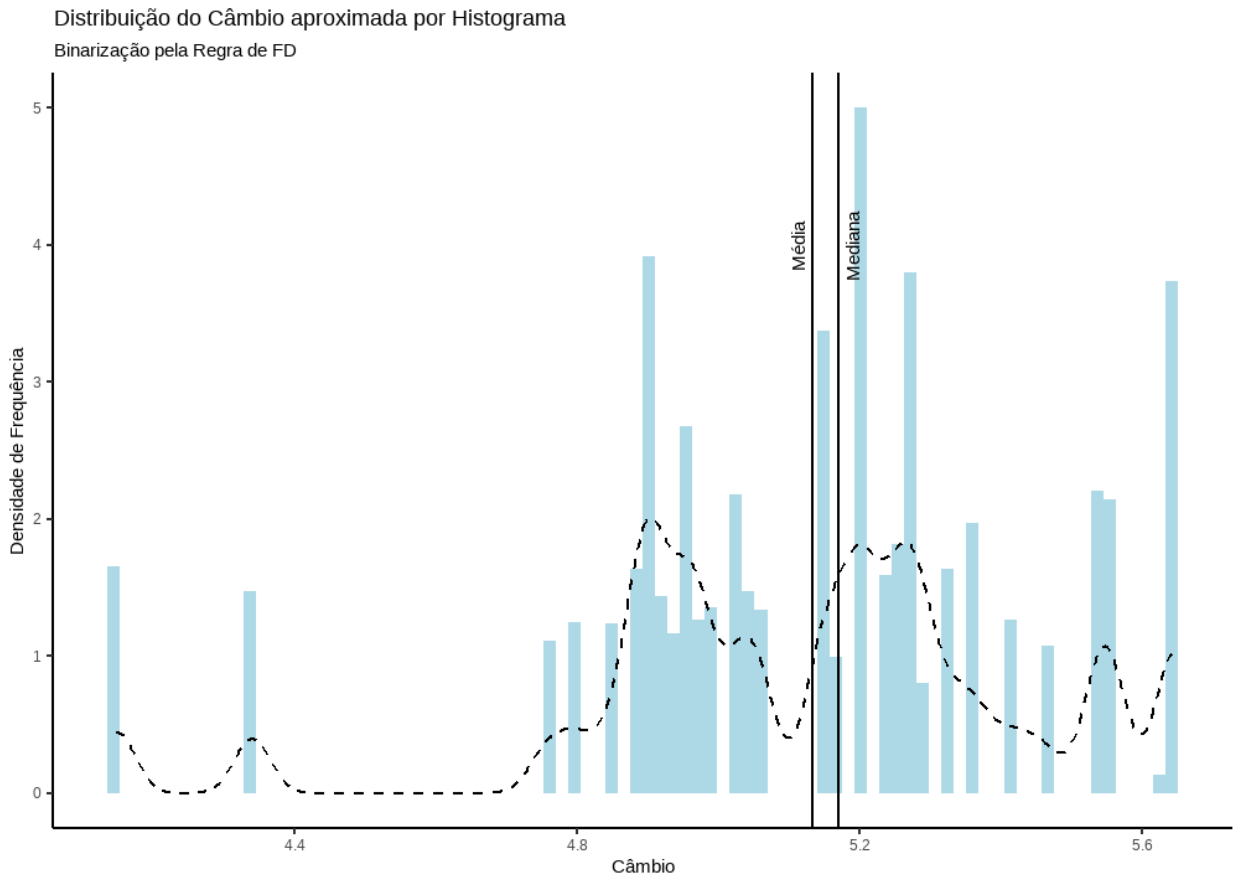


A segunda variável a ser gerado do histograma é o câmbio.

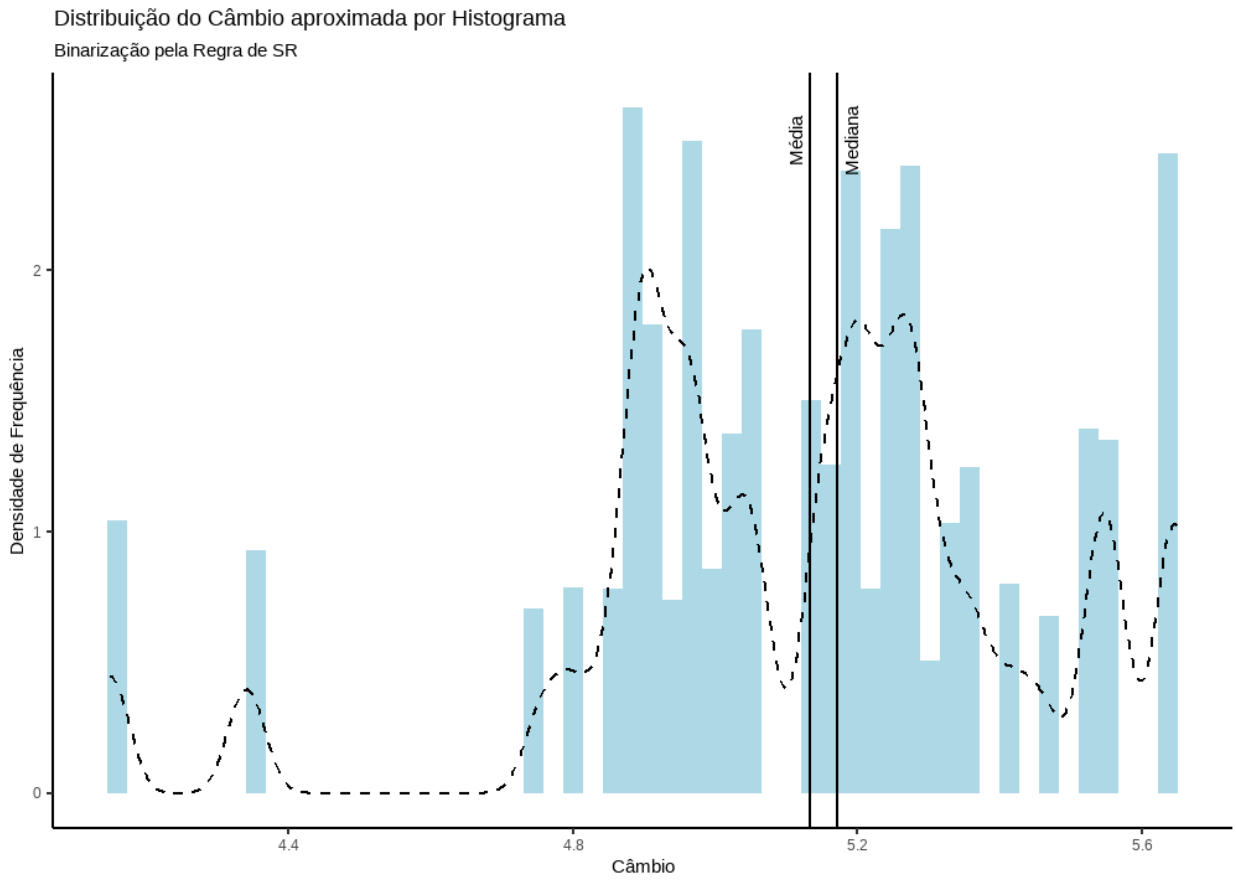
```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Cambio) %>%
   ggplot(aes(x=Cambio)) +
   geom_histogram(aes(y = after_stat(density)) , bins=10, fill = 'lightblue') +
   xlab('Câmbio') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Câmbio aproximada por Histograma",
        subtitle = "Binarização arbitrária") +
   geom_vline(xintercept=c(median(combustiveis$Cambio),
                           mean(combustiveis$Cambio))) +
   annotate("text", x=median(combustiveis$Cambio) +
               0.03, y=0.9, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Cambio) +
               -0.03, y=0.9, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```




```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Cambio) %>%
   ggplot(aes(x=Cambio)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=fd,
                  fill = 'lightblue') +
   xlab('Câmbio') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Câmbio aproximada por Histograma",
        subtitle = "Binarização pela Regra de FD") +
   geom_vline(xintercept=c(median(combustiveis$Cambio),
                           mean(combustiveis$Cambio))) +
   annotate("text", x=median(combustiveis$Cambio) +
               0.02, y=4, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Cambio) +
               -0.02, y=4, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```

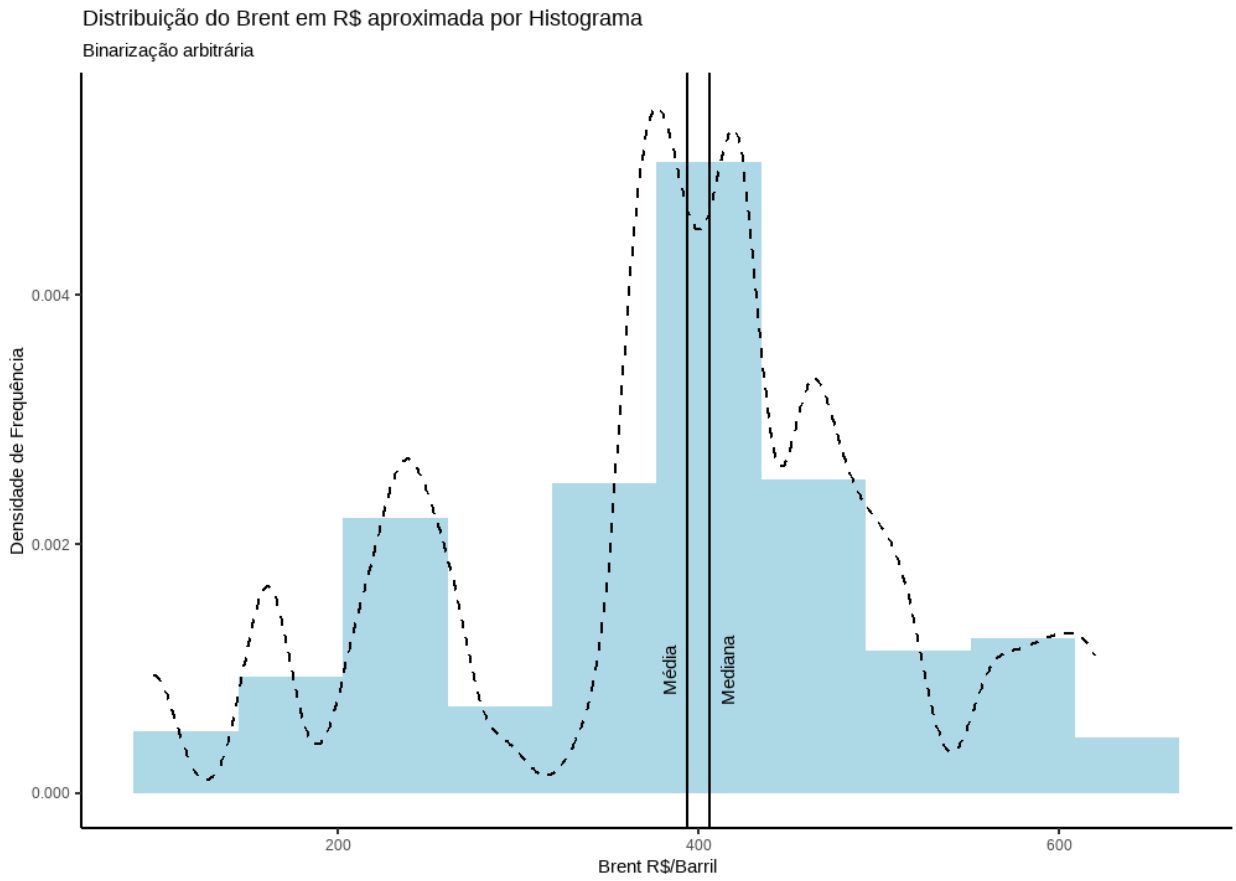



```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Cambio) %>%
   ggplot(aes(x=Cambio)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=sr,
                  fill = 'lightblue') +
   xlab('Câmbio') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Câmbio aproximada por Histograma",
        subtitle = "Binarização pela Regra de SR") +
   geom_vline(xintercept=c(median(combustiveis$Cambio),
                           mean(combustiveis$Cambio))) +
   annotate("text", x=median(combustiveis$Cambio) +
               0.02, y=2.5, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Cambio) +
               -0.02, y=2.5, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```

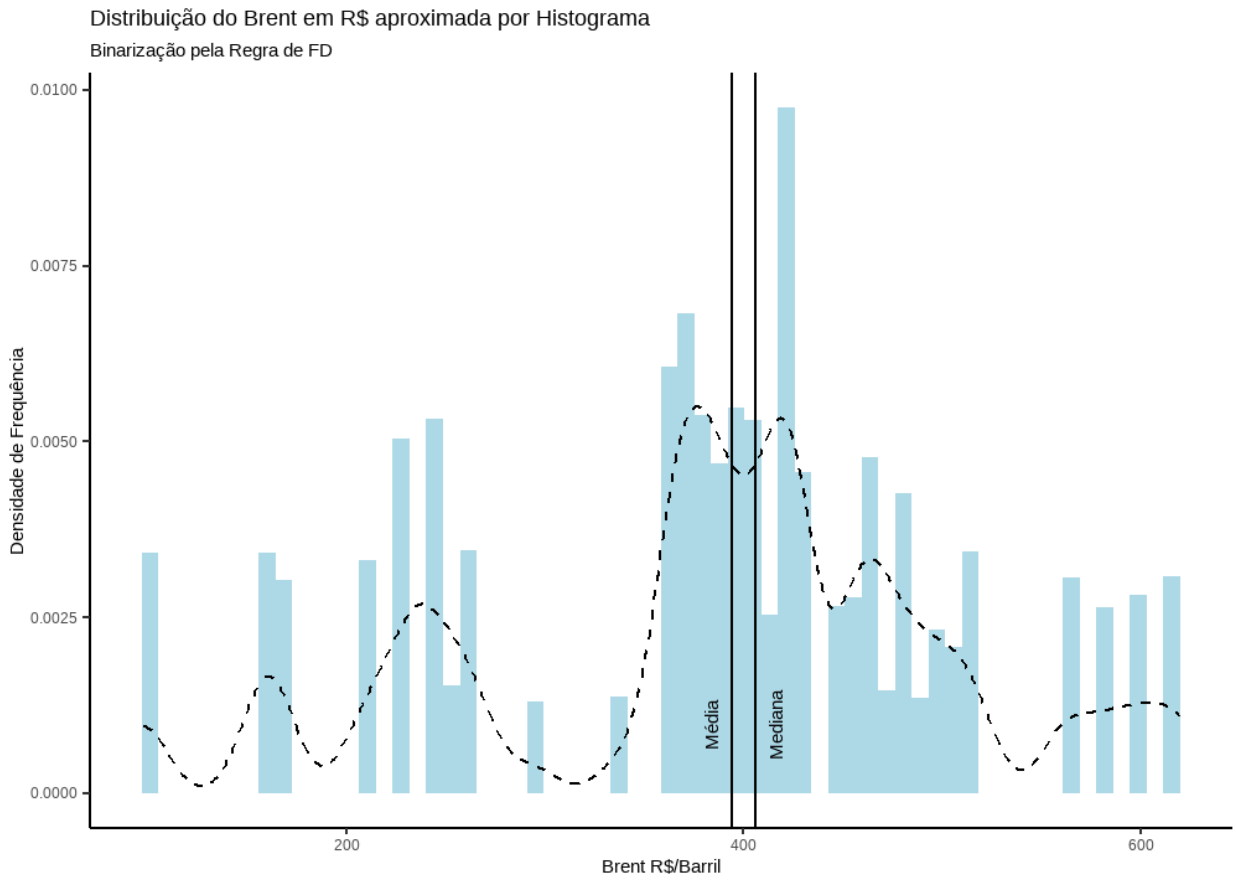


A terceira variável a ser gerado do histograma é o brent.

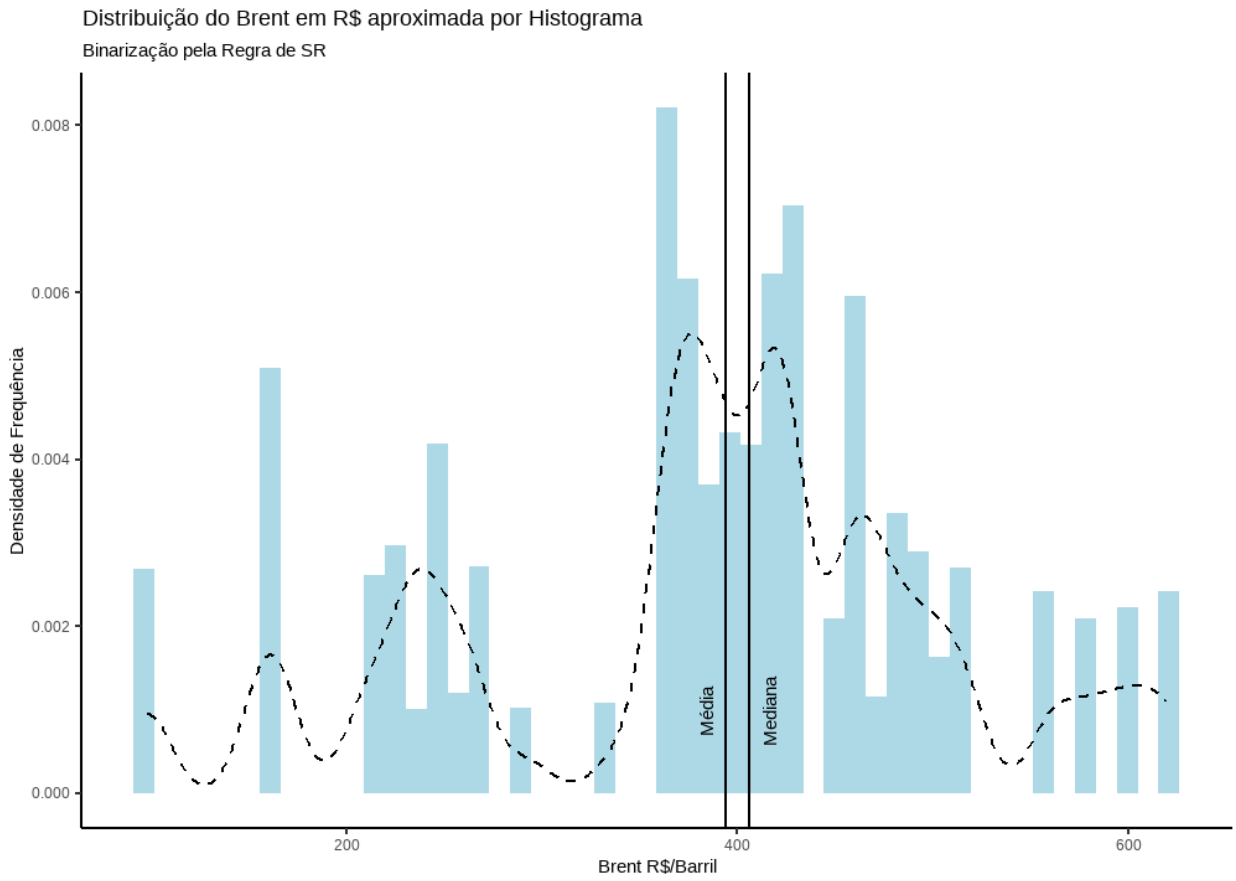
```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Brent_Real_Barril) %>%
   ggplot(aes(x=Brent_Real_Barril)) +
   geom_histogram(aes(y = after_stat(density)) , bins=10, fill = 'lightblue') +
   xlab('Brent R$/Barril') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Brent em R$ aproximada por Histograma",
        subtitle = "Binarização arbitrária") +
   geom_vline(xintercept=c(median(combustiveis$Brent_Real_Barril),
                           mean(combustiveis$Brent_Real_Barril))) +
   annotate("text", x=median(combustiveis$Brent_Real_Barril) +
               10, y=0.001, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Brent_Real_Barril) +
               -10, y=0.001, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```




```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Brent_Real_Barril) %>%
   ggplot(aes(x=Brent_Real_Barril)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=fd,
                  fill = 'lightblue') +
   xlab('Brent R$/Barril') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Brent em R$ aproximada por Histograma",
        subtitle = "Binarização pela Regra de FD") +
   geom_vline(xintercept=c(median(combustiveis$Brent_Real_Barril),
                           mean(combustiveis$Brent_Real_Barril))) +
   annotate("text", x=median(combustiveis$Brent_Real_Barril) +
               10, y=0.001, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Brent_Real_Barril) +
               -10, y=0.001, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```

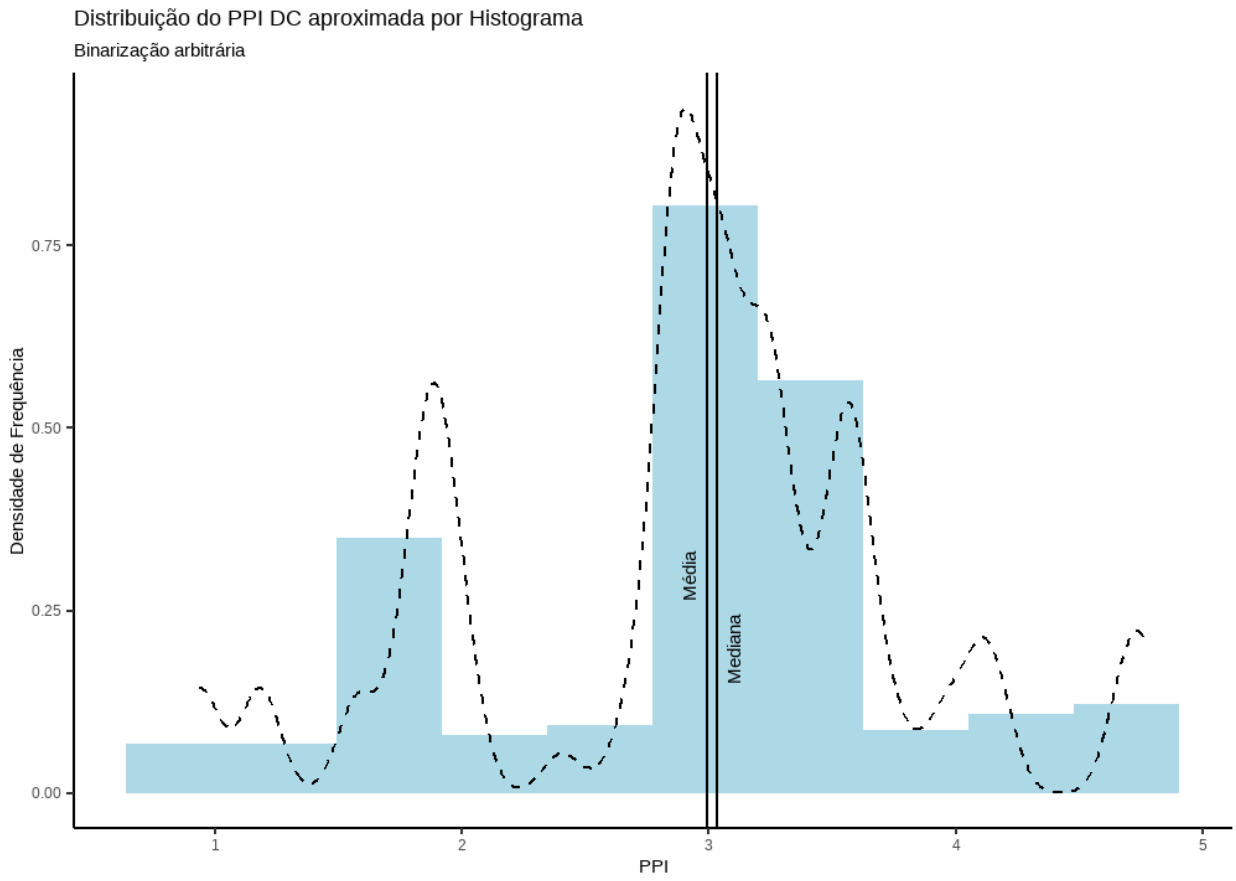



```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(Brent_Real_Barril) %>%
   ggplot(aes(x=Brent_Real_Barril)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=sr,
                  fill = 'lightblue') +
   xlab('Brent R$/Barril') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do Brent em R$ aproximada por Histograma",
        subtitle = "Binarização pela Regra de SR") +
   geom_vline(xintercept=c(median(combustiveis$Brent_Real_Barril),
                           mean(combustiveis$Brent_Real_Barril))) +
   annotate("text", x=median(combustiveis$Brent_Real_Barril) +
               10, y=0.001, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$Brent_Real_Barril) +
               -10, y=0.001, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```

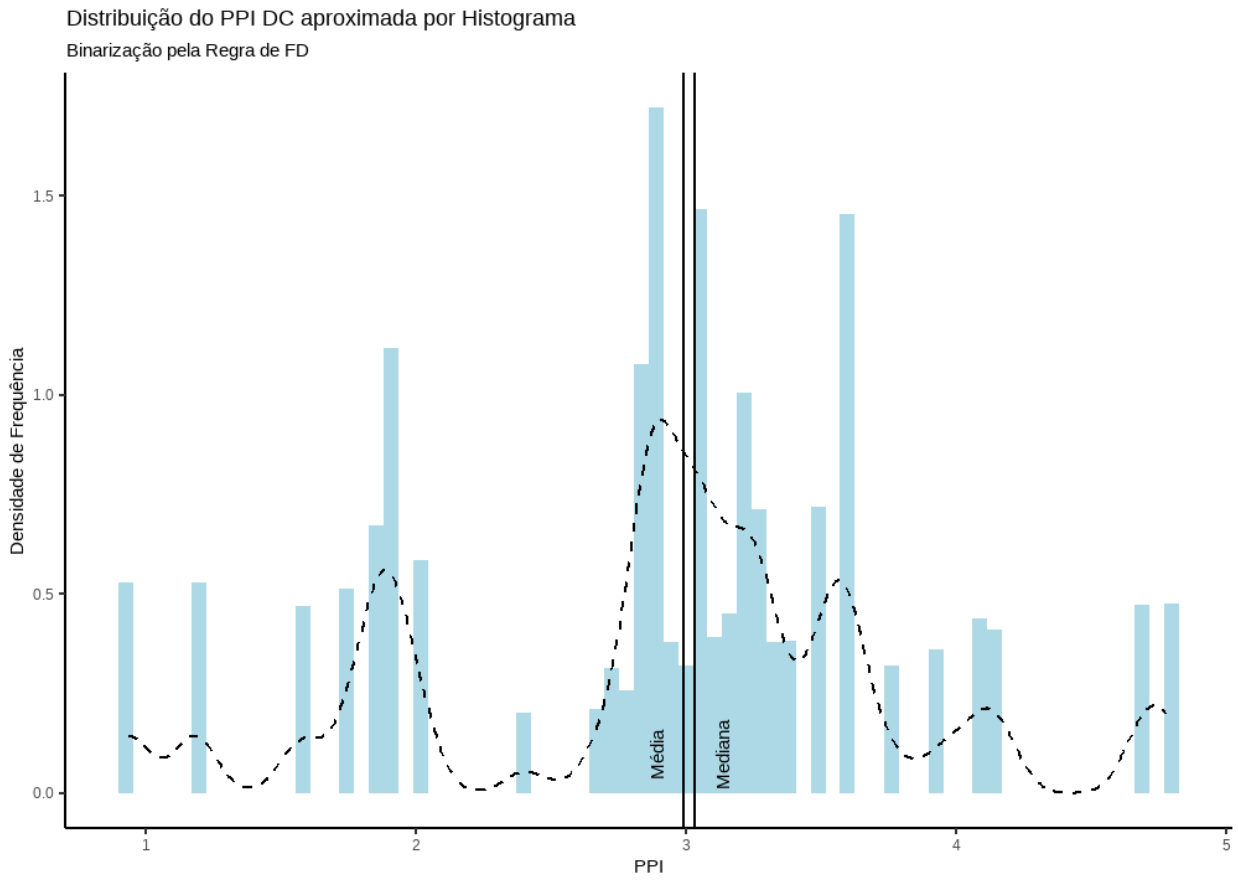


A quarta variável a ser gerado do histograma é o PPI.

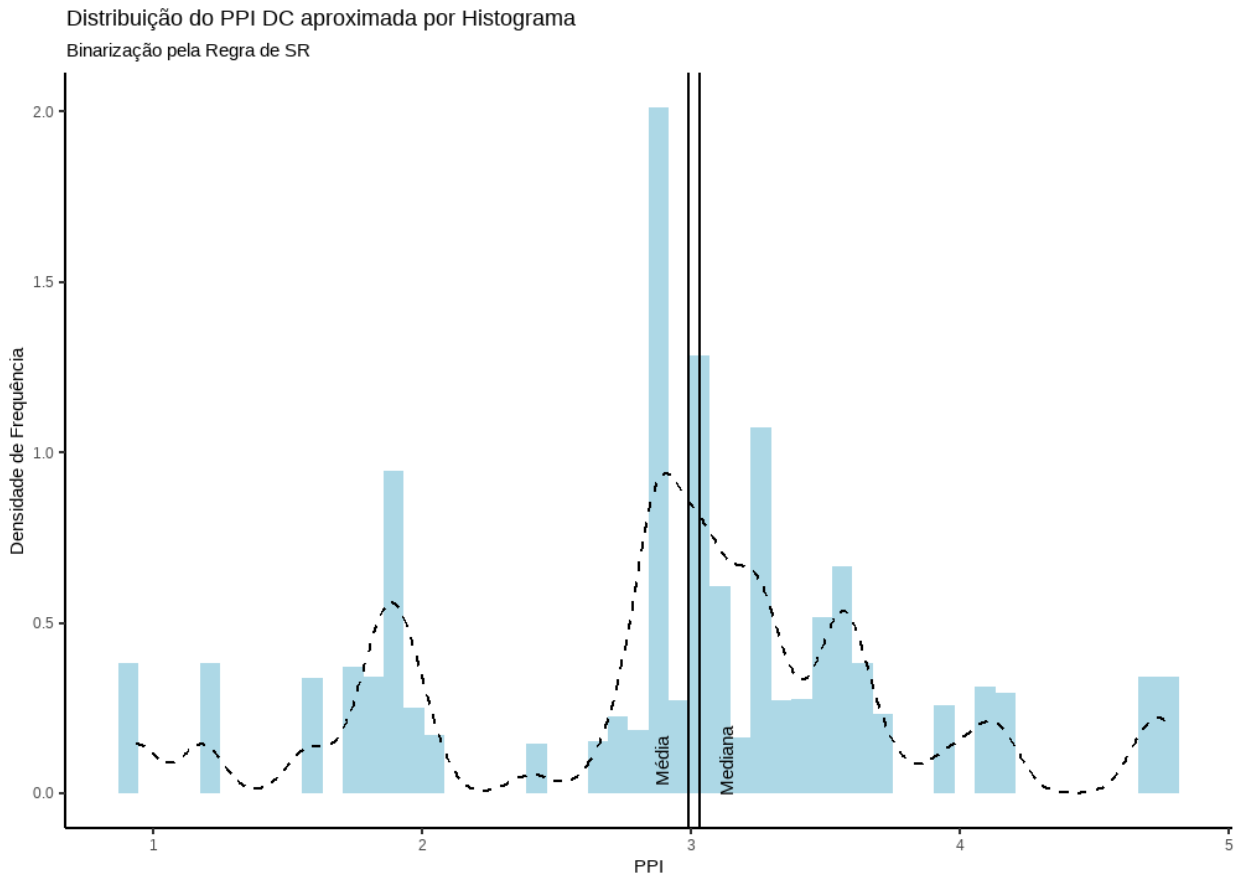
```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(PPI_DC) %>%
   ggplot(aes(x=PPI_DC)) +
   geom_histogram(aes(y = after_stat(density)) , bins=10, fill = 'lightblue') +
   xlab('PPI') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do PPI DC aproximada por Histograma",
        subtitle = "Binarização arbitrária") +
   geom_vline(xintercept=c(median(combustiveis$PPI_DC),
                           mean(combustiveis$PPI_DC))) +
   annotate("text", x=median(combustiveis$PPI_DC) +
               0.07, y=0.2, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$PPI_DC) +
               -0.07, y=0.3, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```




```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(PPI_DC) %>%
   ggplot(aes(x=PPI_DC)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=fd,
                  fill = 'lightblue') +
   xlab('PPI') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do PPI DC aproximada por Histograma",
        subtitle = "Binarização pela Regra de FD") +
   geom_vline(xintercept=c(median(combustiveis$PPI_DC),
                           mean(combustiveis$PPI_DC))) +
   annotate("text", x=median(combustiveis$PPI_DC) +
               0.1, y=0.1, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$PPI_DC) +
               -0.1, y=0.1, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```




```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(PPI_DC) %>%
   ggplot(aes(x=PPI_DC)) +
   geom_histogram(aes(y = after_stat(density)),
                  binwidth=sr,
                  fill = 'lightblue') +
   xlab('PPI') +
   ylab('Densidade de Frequência') +
   labs(title = "Distribuição do PPI DC aproximada por Histograma",
        subtitle = "Binarização pela Regra de SR") +
   geom_vline(xintercept=c(median(combustiveis$PPI_DC),
                           mean(combustiveis$PPI_DC))) +
   annotate("text", x=median(combustiveis$PPI_DC) +
               0.1, y=0.1, label="Mediana", angle=90) +
   annotate("text", x=mean(combustiveis$PPI_DC) +
               -0.1, y=0.1, label="Média", angle=90) +
   geom_density(linetype = 2) +
   theme_classic()
```

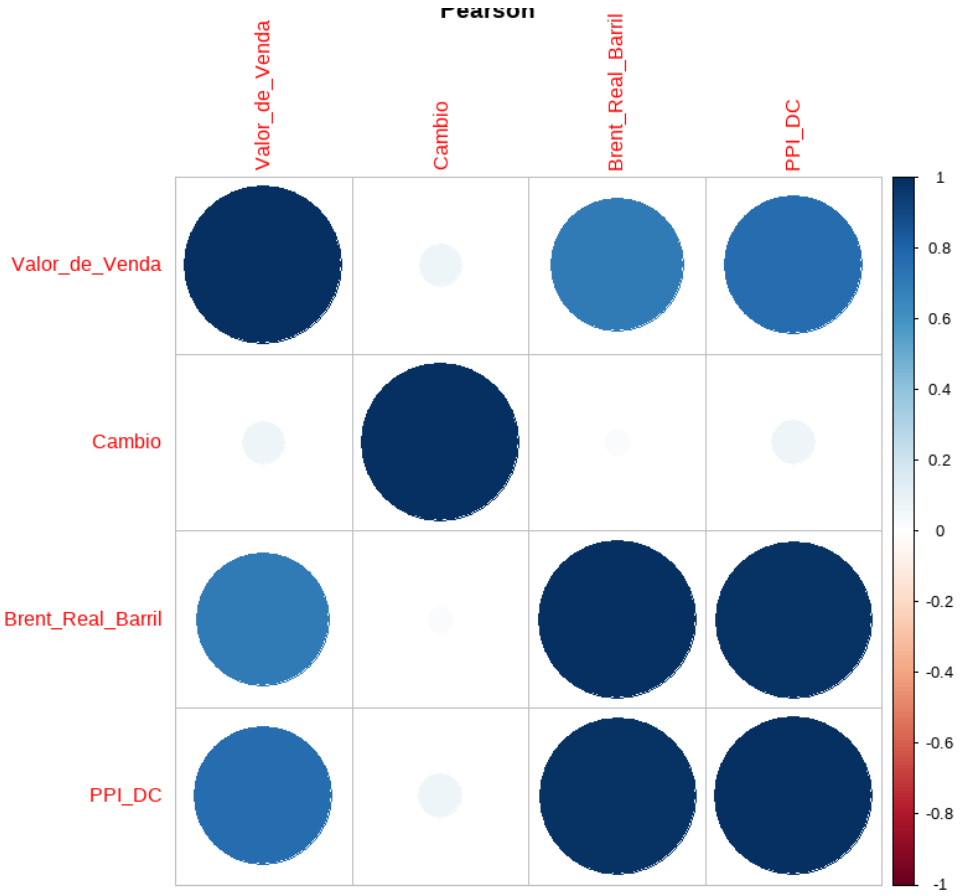


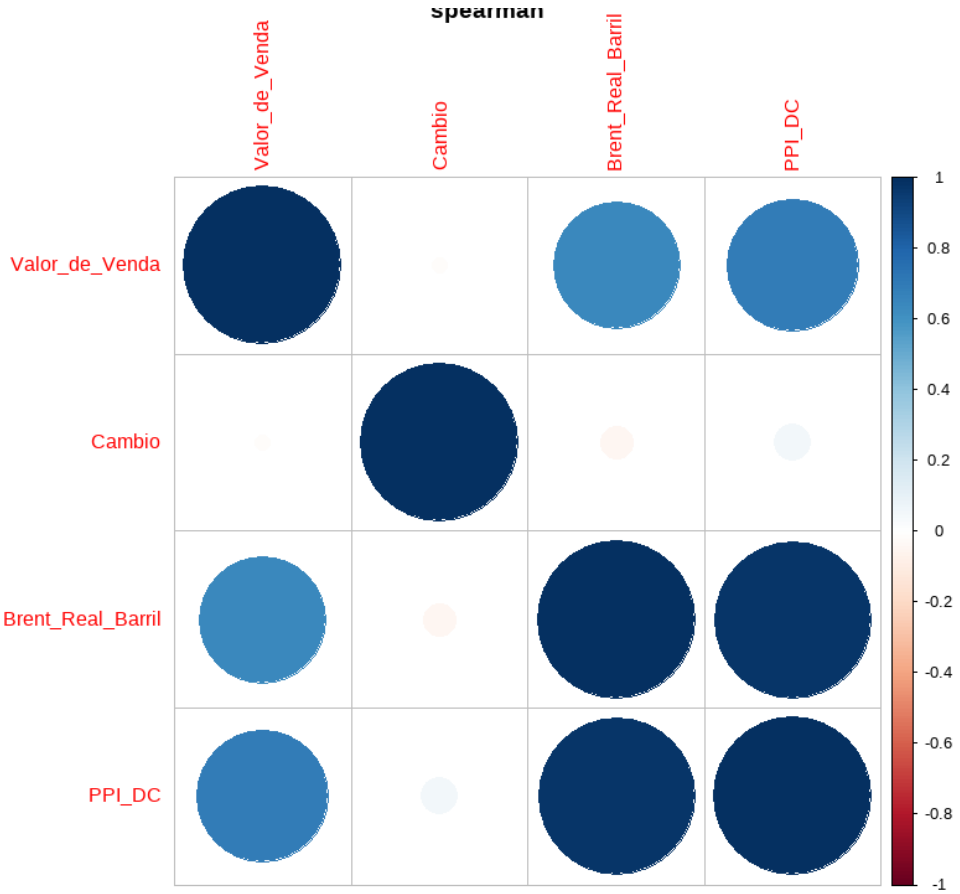
Calculando a dispersão e as correlações

Vamos calcular a correlação entre as variáveis, para verificar se há evidência de relação entre elas, e o quão intensa é essa relação.

```
cor(combustiveis %>%
      dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
      dplyr::select(Valor_de_Venda, Cambio, Brent_Real_Barril, PPI_DC))
```

```
##                   Valor_de_Venda     Cambio Brent_Real_Barril     PPI_DC
## Valor_de_Venda        1.00000000 0.07559102        0.70787134 0.76009481
## Cambio                0.07559102 1.00000000        0.02596631 0.07997226
## Brent_Real_Barril     0.70787134 0.02596631        1.00000000 0.98060222
## PPI_DC                0.76009481 0.07997226        0.98060222 1.00000000
```

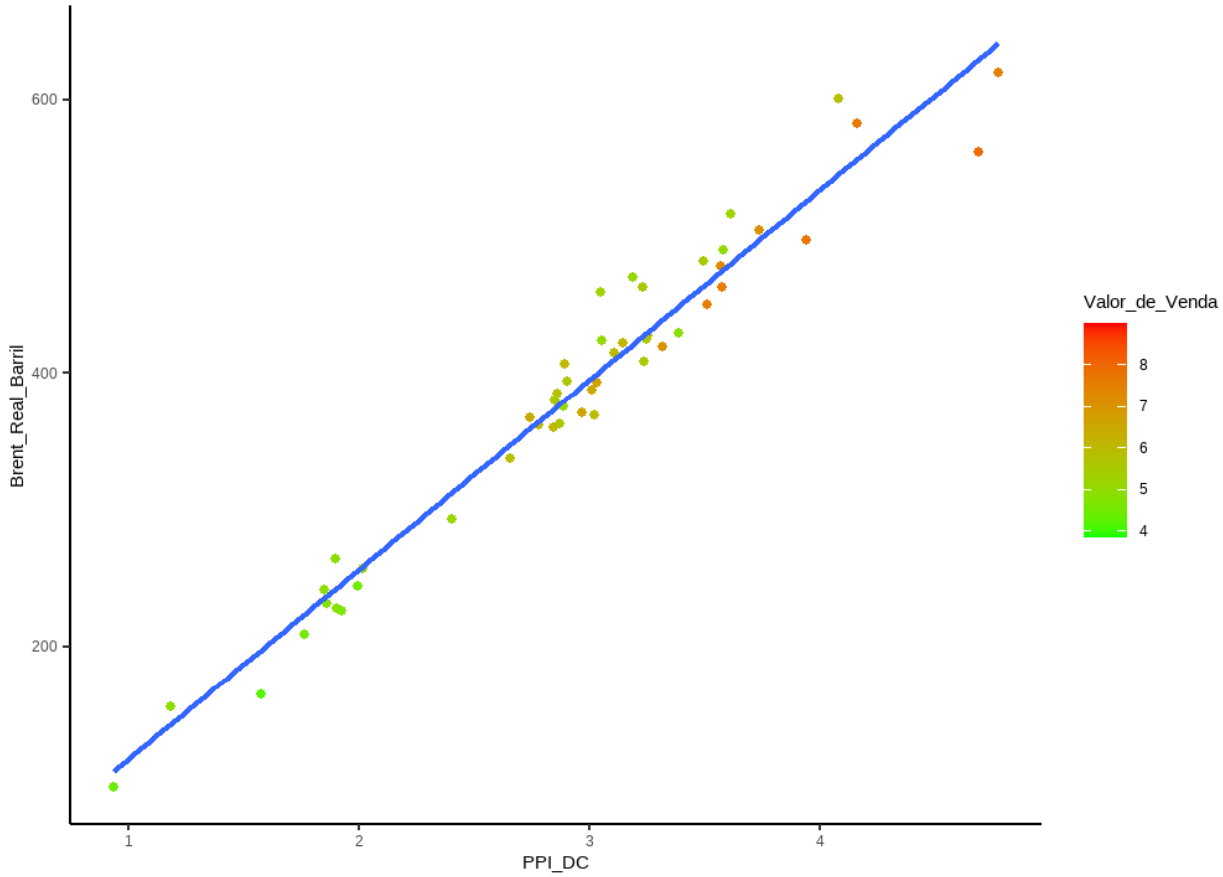

Pearson
Valor_de_Venda
Cambio
Brent_Real_Barril
PPI_DC
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



É possível observar que o valor de venda da gasolina possui uma correlação expressiva com o valor do Brent e PPI DC. Tal análise permite concluir que o valor da venda da gasolina nos postos de combustíveis variam com alguma frenquência conforme a variação do custo do Brent e do PPI.

O scatter plot pode nos auxiliar na visualização da proporcionalidade de crescimento entre as variáveis mais correlacionadas.

```
combustiveis %>%
   dplyr::filter(Produto == "GASOLINA", UF == "RJ") %>%
   dplyr::select(PPI_DC, Brent_Real_Barril, Valor_de_Venda) %>%
   ggplot(aes(x = PPI_DC, y = Brent_Real_Barril, color = Valor_de_Venda)) +
   geom_point() +
   scale_colour_gradient(low = "green", high = "red") +
   geom_smooth(method = "lm", se = FALSE) +
   theme_classic()
```



É possível observar uma alta correção entre Valor_de_Venda em relação PPI_DC e Brent R$/Barril. Diante do exposto é possível afirmar que o PPI_DC e o Brent influenciam no valor do preço da Gasolina.

Esta parte falta

Analisando a normalidade dos dados

Considerando os resultados apresentados anteriormente pelos histogramas, o próximo passo visa realizar alguns testes para checar normalidade das variáveis.

Manipulando base de dados com dados faltantes e outliers

O que é completude de dados?

Completude de dados se refere a ausência de dados em um conjunto de dados. Quando os dados estão completos em um conjunto de dados e sua consistência pode ser validada, dizemos que há qualidade dos dados a serem utilizados em uma análise, possibilitando assim *insights* confiáveis.

Qual o impacto que os dados faltantes podem ter em uma análise?

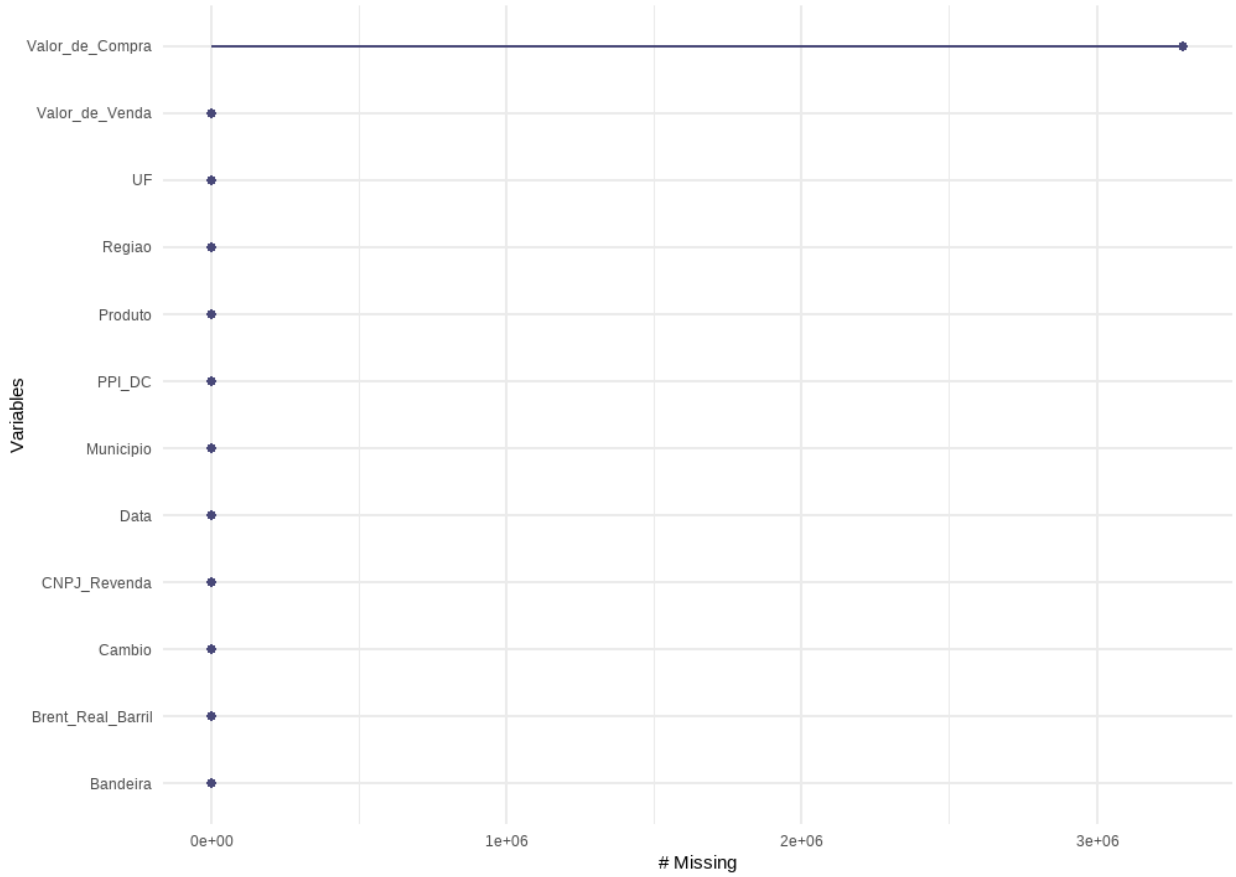
A ausência de dados ou lacunas no conjunto de dados em uma análise, além de proporcionar uma baixa qualidade dos dados, poderá impossibilitar a obtenção de *insights* confiáveis e precisos. A inconfiabilidade nos dados podem gerar interpretações equivocadas e propor decisões errôneas.

Índice de completude

```
combustiveis %>% naniar::miss_var_summary()
```

```
## # A tibble: 12 x 3
##    variable           n_miss pct_miss
##    <chr>               <int>    <num>
##  1 Valor_de_Compra   3293286     95.5
##  2 Data                    0      0
##  3 Regiao                  0      0
##  4 UF                      0      0
##  5 Municipio               0      0
##  6 CNPJ_Revenda            0      0
##  7 Produto                 0      0
##  8 Bandeira                0      0
##  9 Valor_de_Venda          0      0
## 10 Cambio                  0      0
## 11 PPI_DC                  0      0
## 12 Brent_Real_Barril       0      0
```

```
combustiveis %>% gg_miss_var()
```



A única variável de interesse que possui missing é a Valor de Compra, intitulada Valor_de_Compra.

Base de Dados Combustíveis - Visão de missing nacional

Para uma melhor compreensão dos missing existentes nesta variável, algumas investigações foram feitas.

A primeira análise visa visualizar o missing da variável por Regiao.

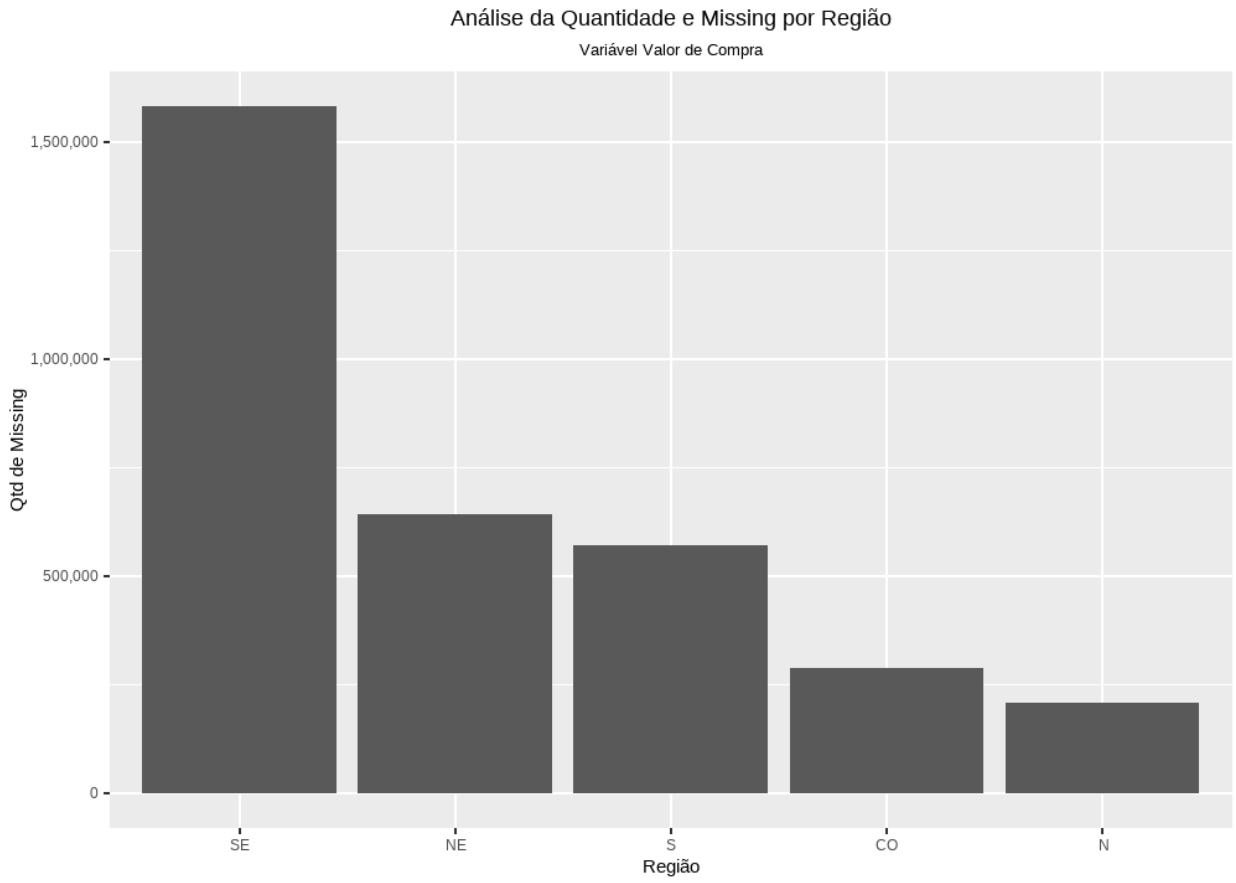
```
df <- combustiveis %>%
  dplyr::group_by(Regiao) %>%
  dplyr::mutate(n_linhas = 1) %>%
  dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra))) %>%
  dplyr::arrange(desc(n_missing))

#knitr::kable(df, format="latex")
knitr::kable(df,
             format="latex",
             caption = "Quantidade de missing da variável Valor de Compra por Região.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = ""
)
```

Table 2: Quantidade de missing da variável Valor de Compra por Região.

Regiao	n_missing
SE	1582901
NE	642656
S	570709
CO	288051
N	208969

É possível observar que a região sudeste possui maior quantidade de missing. O gráfico abaixo demonstra esta análise visualmente.

```
ggplot(df, aes(x = reorder(Regiao,
                           n_missing, decreasing = TRUE),
               y = n_missing)) +
   geom_bar(stat = "identity") +
   labs(x = "Região",
        y = "Qtd de Missing",
        subtitle = "Variável Valor de Compra") +
   scale_y_continuous(labels = scales::comma_format()) +
   ggtitle("Análise da Quantidade e Missing por Região") +
   theme(plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5,
                                      size = 10)) +
   scale_fill_brewer(palette = "Set1")
```




```
rm("df")
```

Foi feita uma segunda análise considerando a porcentagem de missing sobre a observação desta variável para cada região. É possível observar uma uniformização nas porcentagens entre as regiões.

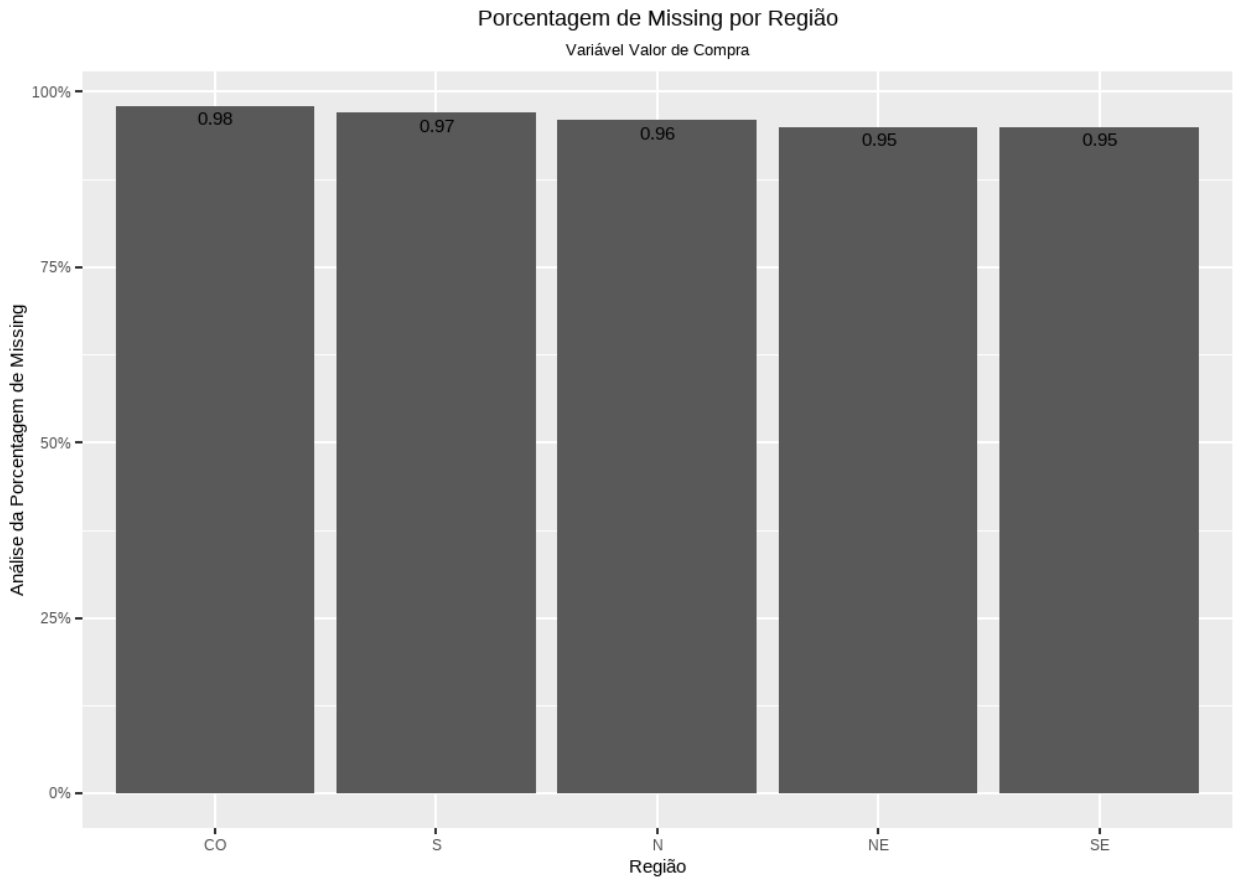
```
df <- combustiveis %>%
  dplyr::group_by(Regiao) %>%
  dplyr::mutate(n_linhas = 1) %>%
  dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra)),
            n_linhas = sum(n_linhas),
            porcentagem_missing = round(n_missing / n_linhas, 2)) %>%
  dplyr::arrange(desc(porcentagem_missing))

#kable(df, format="latex")
knitr::kable(df,
             format="latex",
             caption = "Porcentagem de missing da variável Valor de Compra por Região.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = "",)
```

Table 3: Porcentagem de missing da variável Valor de Compra por Região.

Regiao	n_missing	n_linhas	porcentagem_missing
CO	288051	295137	0.98
S	570709	586858	0.97
N	208969	216911	0.96
NE	642656	675512	0.95
SE	1582901	1674880	0.95

O gráfico abaixo demonstra visualmente esta análise.

```
ggplot(df, aes(x = reorder(Regiao, porcentagem_missing, decreasing = TRUE),
               y = porcentagem_missing)) +
   geom_bar(stat = "identity") +
   labs(x = "Região",
        y = "Análise da Porcentagem de Missing",
        subtitle = "Variável Valor de Compra") +
   scale_y_continuous(labels = scales::percent_format()) +
   ggtitle("Porcentagem de Missing por Região") +
   theme(plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5,
                                      size = 10)) +
   geom_text(aes(label = porcentagem_missing),
             vjust = 1.5)
```




```
rm("df")
```

A mesma análise realizada por estado, foi feita por Estado.

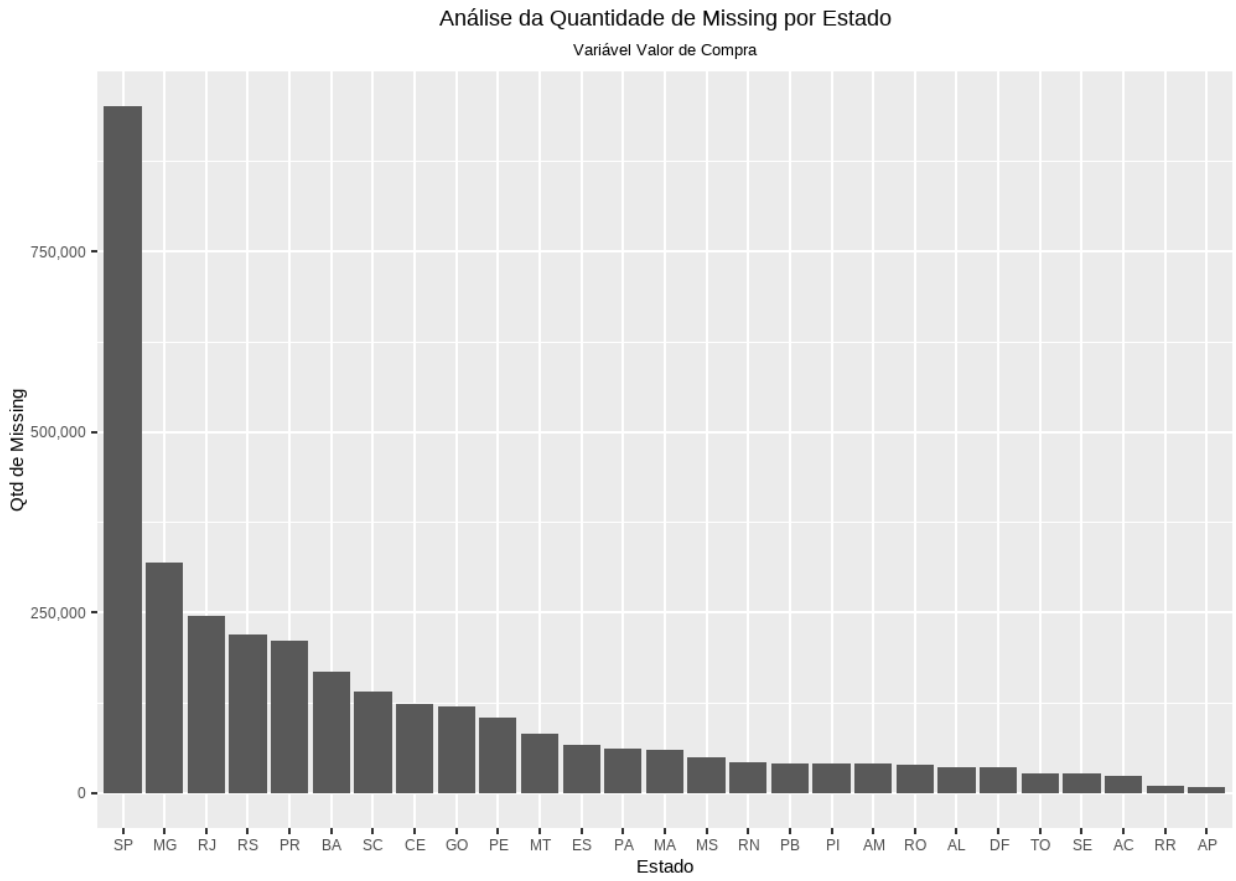
```
df <- combustiveis %>%
   dplyr::group_by(UF) %>%
   dplyr::mutate(n_linhas = 1) %>%
   dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra))) %>%
   dplyr::arrange(desc(n_missing))

#kable(df, format="latex")

knitr::kable(df,
             format="latex",
             caption = "Quantidade de missing da variável Valor de Compra por Estado.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = "",)
```

Table 4: Quantidade de missing da variável Valor de Compra por Estado.

UF	n_missing
SP	951532
MG	319240
RJ	244726
RS	218907
PR	210907
BA	167253
SC	140895
CE	122589
GO	120176
PE	104438
MT	82896
ES	67403
PA	60808
MA	60711
MS	48877
RN	42117
PB	41490
PI	40982
AM	40418
RO	38941
AL	36200
DF	36102
TO	27602
SE	26876
AC	23022
RR	9260
AP	8918

É possível observar que o missing da respectiva variável é expressivo para os postos de combustíveis de SP, MG e RJ. O gráfico abaixo demonstra esta representatividade.

```
ggplot(df, aes(x = reorder(UF, n_missing, decreasing = TRUE), y = n_missing)) +
   geom_bar(stat = "identity") +
   labs(x = "Estado", y = "Qtd de Missing", subtitle = "Variável Valor de Compra") +
   scale_y_continuous(labels = scales::comma_format()) +
   ggtitle("Análise da Quantidade de Missing por Estado") +
   theme(plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5, size = 10)) +
   scale_fill_brewer(palette = "Set1")
```




```
rm("df")
```

Foi feita uma segunda análise considerando a porcentagem de missing sobre a observação desta variável para cada estado. É possível observar valores próximos.

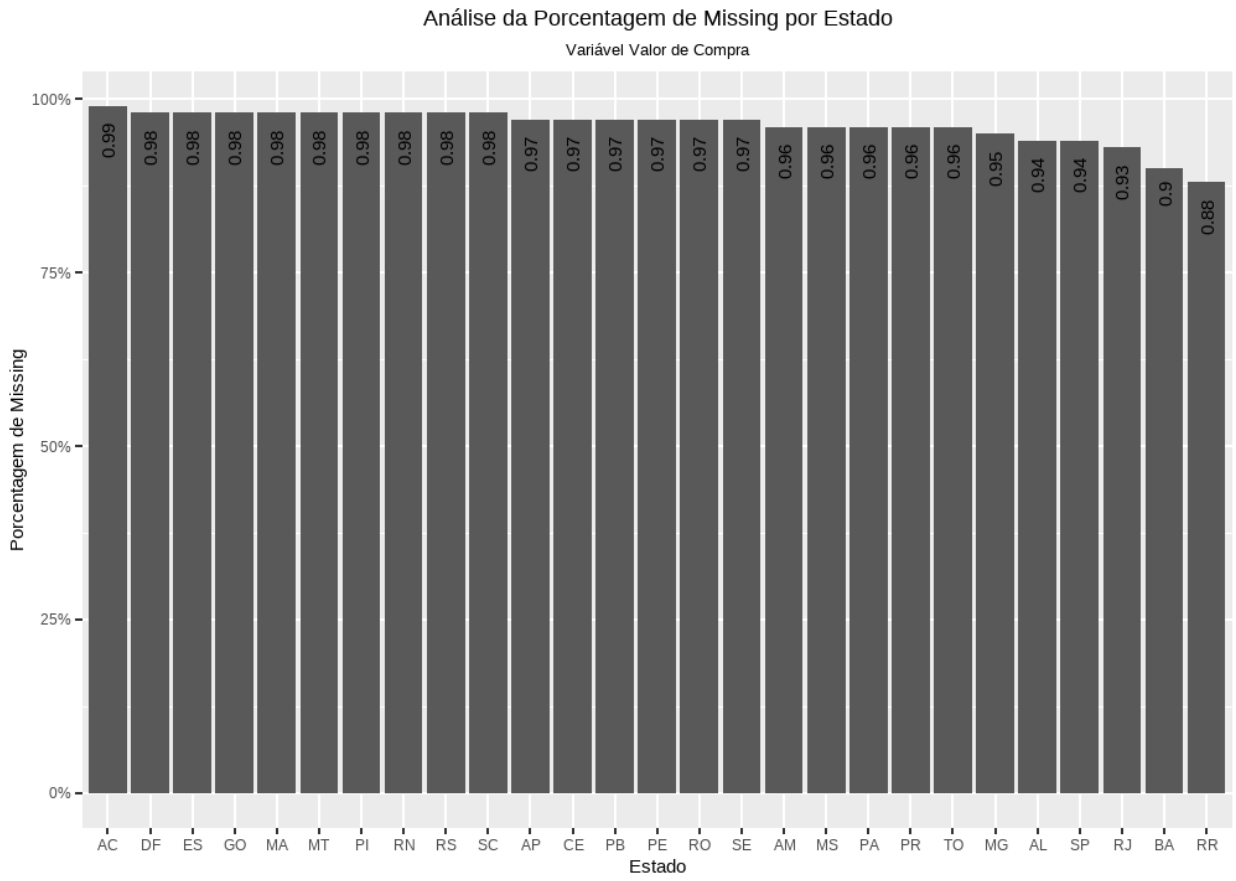
```
df <- combustiveis %>%
  dplyr::group_by(UF) %>%
  dplyr::mutate(n_linhas = 1) %>%
  dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra)),
            n_linhas = sum(n_linhas),
            porcentagem_missing = round(n_missing / n_linhas, 2)) %>%
  dplyr::arrange(desc(porcentagem_missing))

#kable(df, format="latex")
knitr::kable(df,
             format="latex",
             caption = "Porcentagem de missing da variável Valor de Compra por Estado.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = "",)
```

Table 5: Porcentagem de missing da variável Valor de Compra por Estado.

UF	n_missing	n_linhas	porcentagem_missing
AC	23022	23196	0.99
DF	36102	36818	0.98
ES	67403	69059	0.98
GO	120176	122929	0.98
MA	60711	61979	0.98
MT	82896	84661	0.98
PI	40982	41903	0.98
RN	42117	43185	0.98
RS	218907	223506	0.98
SC	140895	143146	0.98
AP	8918	9190	0.97
CE	122589	126553	0.97
PB	41490	42627	0.97
PE	104438	107256	0.97
RO	38941	40072	0.97
SE	26876	27654	0.97
AM	40418	42110	0.96
MS	48877	50729	0.96
PA	60808	63064	0.96
PR	210907	220206	0.96
TO	27602	28775	0.96
MG	319240	334763	0.95
AL	36200	38645	0.94
SP	951532	1007206	0.94
RJ	244726	263852	0.93
BA	167253	185710	0.90
RR	9260	10504	0.88

O gráfico abaixo demonstra visualmente estes valores.

```
ggplot(df, aes(x = reorder(UF,
                           porcentagem_missing, decreasing = TRUE),
               y = porcentagem_missing)) +
  geom_bar(stat = "identity") +
  labs(x = "Estado",
       y = "Porcentagem de Missing",
       subtitle = "Variável Valor de Compra") +
  scale_y_continuous(labels = scales::percent_format()) +
  ggtitle("Análise da Porcentagem de Missing por Estado") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5, size = 10)) +
  geom_text(aes(label = porcentagem_missing),
            angle = 90,
            vjust = 0.5,
            hjust = 1.5)
```




```
rm("df")
```

Base de Dados Combustíveis - Visão de missing estado RJ

A próxima análise visa compreender a quantidade de missing da variável Valor de Compra nos municípios do Estado do RJ.

```
df <- combustiveis %>% dplyr::filter(UF == "RJ") %>%
  dplyr::group_by(Municipio) %>%
  dplyr::mutate(n_linhas = 1) %>%
  dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra))) %>%
  dplyr::arrange(desc(n_missing))


#kable(df, format="latex")
knitr::kable(df,
             format="latex",
             caption = "Quantidade de missing da variável Valor no Estado do RJ.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = "",)
```

Table 6: Quantidade de missing da variável Valor no Estado do RJ.

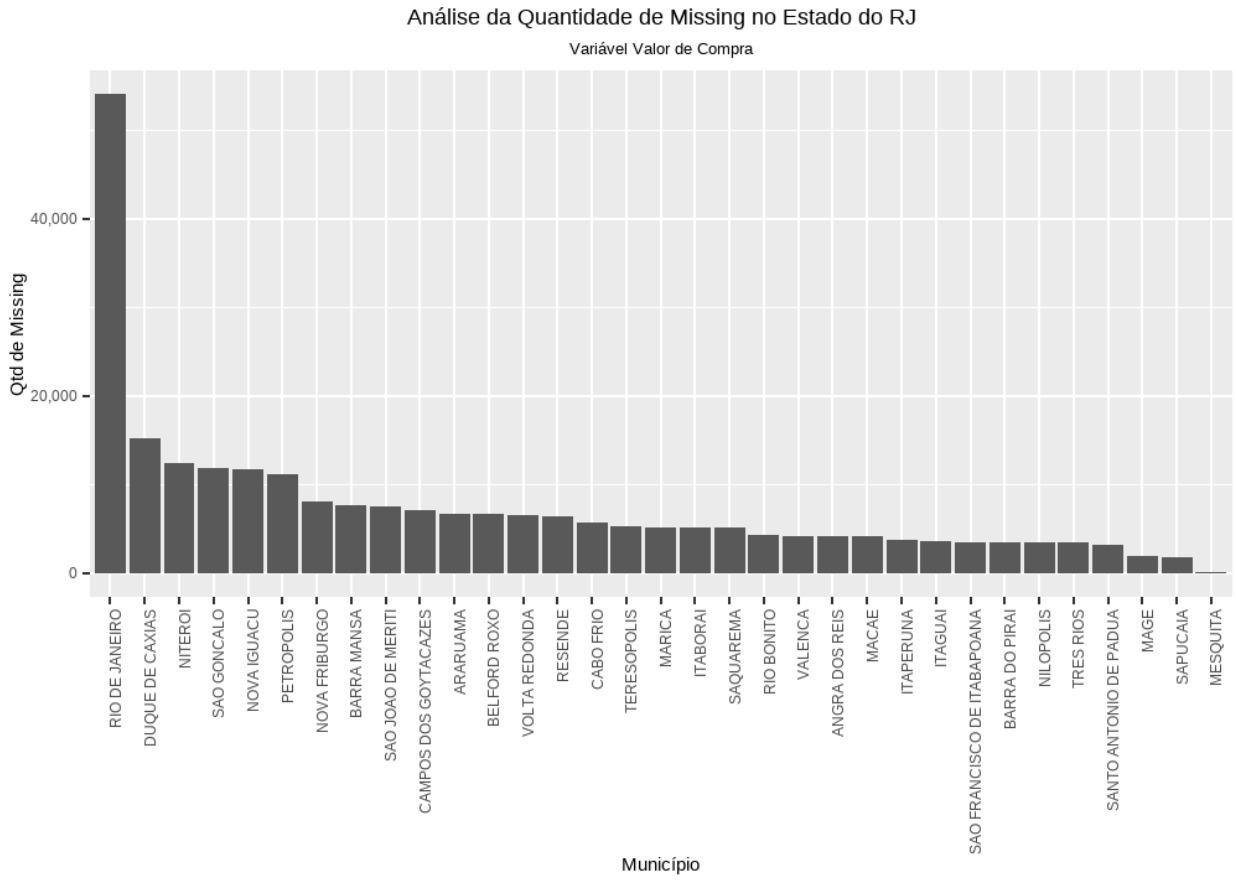
Municipio	n_missing
RIO DE JANEIRO	54054
DUQUE DE CAXIAS	15180
NITEROI	12462
SAO GONCALO	11933
NOVA IGUACU	11712
PETROPOLIS	11151
NOVA FRIBURGO	8066
BARRA MANSA	7605
SAO JOAO DE MERITI	7564
CAMPOS DOS GOYTACAZES	7131
ARARUAMA	6732
BELFORD ROXO	6646
VOLTA REDONDA	6507
RESENDE	6448
CABO FRIO	5715
TERESOPOLIS	5278
MARICA	5199
ITABORAI	5135
SAQUAREMA	5113
RIO BONITO	4293
VALENCA	4226
ANGRA DOS REIS	4139
MACAE	4114
ITAPERUNA	3770
ITAGUAI	3631
SAO FRANCISCO DE ITABAPOANA	3501
BARRA DO PIRAI	3487
NILOPOLIS	3436
TRES RIOS	3407
SANTO ANTONIO DE PADUA	3247
MAGE	1972
SAPUCAIA	1753
MESQUITA	119

É possível observar os cinco municípios com maior quantidade de missing encontram-se na região metropolitana do Rio de Janeiro.

O gráfico a seguir desmonstra esta analise visualmente.

```
ggplot(df, aes(x = reorder(Municipio, n_missing, decreasing = TRUE),
               y = n_missing)) +
   geom_bar(stat = "identity") +
   labs(x = "Município",
        y = "Qtd de Missing",
        subtitle = "Variável Valor de Compra") +
   scale_y_continuous(labels = scales::comma_format()) +
   ggtitle("Análise da Quantidade de Missing no Estado do RJ") +
   theme(plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5,
```

```
                                    size = 10),
        axis.text.x = element_text(angle = 90,
                                   hjust = 1)
  ) +
  scale_fill_brewer(palette = "Set1")
```




```
rm("df")
```

Foi feita uma segunda análise considerando a porcentagem de missing sobre a observação desta variável para cada Município do Rio de Janeiro.

```
df <- combustiveis %>% dplyr::filter(UF == "RJ") %>%
  dplyr::group_by(Municipio) %>%
  dplyr::mutate(n_linhas = 1) %>%
  dplyr::summarise(n_missing = sum(is.na(Valor_de_Compra)),
            n_linhas = sum(n_linhas),
            porcentagem_missing = round(n_missing / n_linhas, 2)) %>%
  dplyr::arrange(desc(porcentagem_missing))

df
```

```
## # A tibble: 33 x 4
```

```
##    Municipio                  n_missing n_linhas porcentagem_missing
##    <chr>                          <int>    <dbl>               <dbl>
##  1 MESQUITA                         119      119                1
##  2 SAPUCAIA                        1753     1775                0.99
##  3 ANGRA DOS REIS                  4139     4235                0.98
##  4 MACAE                           4114     4185                0.98
##  5 NOVA FRIBURGO                   8066     8251                0.98
##  6 PETROPOLIS                     11151    11401                0.98
##  7 RESENDE                         6448     6601                0.98
##  8 ARARUAMA                        6732     6913                0.97
##  9 SAO FRANCISCO DE ITABAPOANA     3501     3597                0.97
## 10 SAQUAREMA                       5113     5290                0.97
## # i 23 more rows
```

```
#kable(df, format="latex")
knitr::kable(df,
             format="latex",
             caption = "Porcentagem de missing da variável
             Valor de Compra por Município RJ.",
             align = "c",
             booktabs = TRUE,
             longtable = TRUE,
             linesep = "",)
```

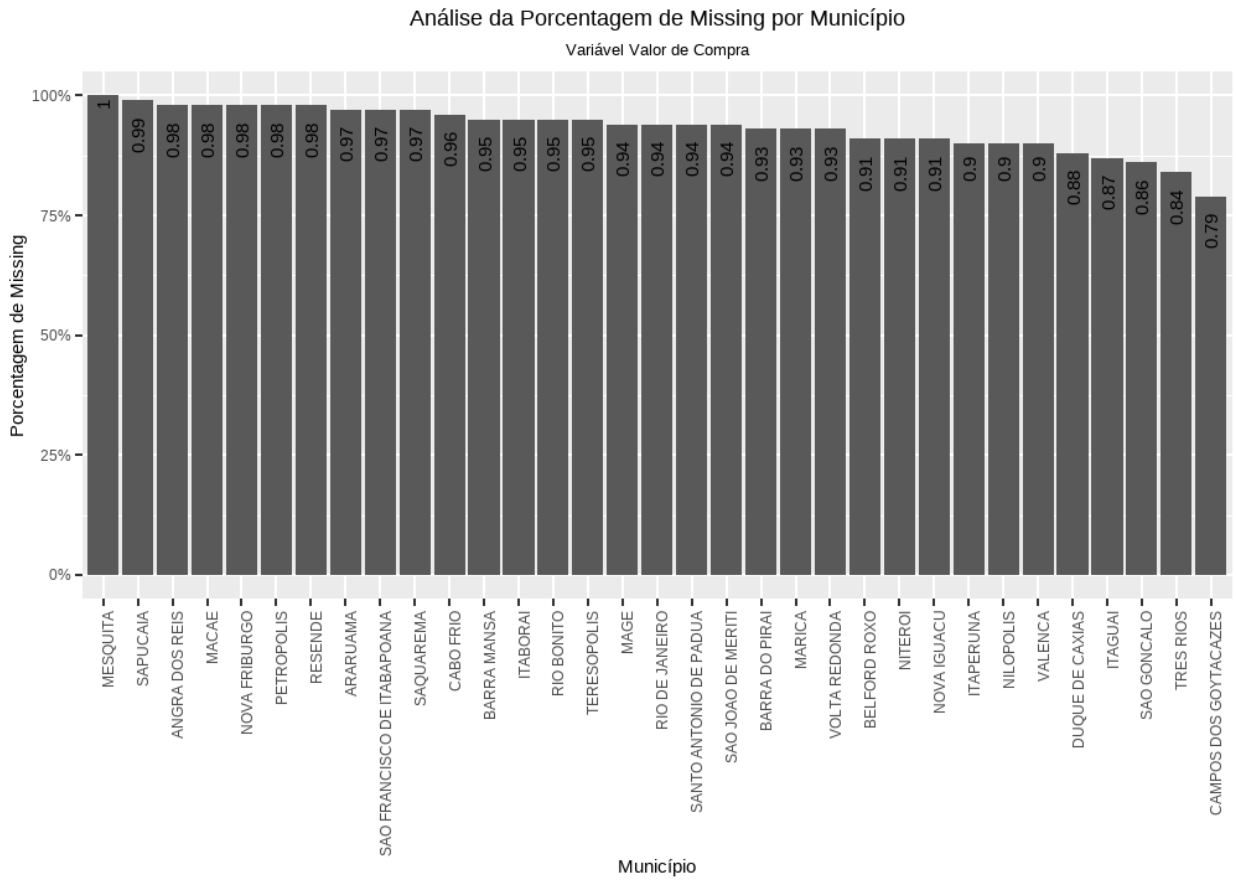
Table 7: Porcentagem de missing da variável Valor de Compra por Município RJ.

Municipio	n_missing	n_linhas	porcentagem_missing
MESQUITA	119	119	1.00
SAPUCAIA	1753	1775	0.99
ANGRA DOS REIS	4139	4235	0.98
MACAE	4114	4185	0.98
NOVA FRIBURGO	8066	8251	0.98
PETROPOLIS	11151	11401	0.98
RESENDE	6448	6601	0.98
ARARUAMA	6732	6913	0.97
SAO FRANCISCO DE ITABAPOANA	3501	3597	0.97
SAQUAREMA	5113	5290	0.97
CABO FRIO	5715	5953	0.96
BARRA MANSA	7605	8021	0.95
ITABORAI	5135	5405	0.95
RIO BONITO	4293	4503	0.95
TERESOPOLIS	5278	5530	0.95
MAGE	1972	2103	0.94
RIO DE JANEIRO	54054	57233	0.94
SANTO ANTONIO DE PADUA	3247	3442	0.94
SAO JOAO DE MERITI	7564	8028	0.94
BARRA DO PIRAI	3487	3761	0.93
MARICA	5199	5591	0.93
VOLTA REDONDA	6507	6970	0.93
BELFORD ROXO	6646	7332	0.91
NITEROI	12462	13648	0.91

NOVA IGUACU	11712	12856	0.91
ITAPERUNA	3770	4168	0.90
NILOPOLIS	3436	3808	0.90
VALENCA	4226	4706	0.90
DUQUE DE CAXIAS	15180	17272	0.88
ITAGUAI	3631	4173	0.87
SAO GONCALO	11933	13901	0.86
TRES RIOS	3407	4043	0.84
CAMPOS DOS GOYTACAZES	7131	9038	0.79

o gráfico abaixo demonstra visualmente esta análise.

```
ggplot(df, aes(x = reorder(Municipio,
                           porcentagem_missing,
                           decreasing = TRUE),
               y = porcentagem_missing)) +
   geom_bar(stat = "identity") +
   labs(x = "Município",
        y = "Porcentagem de Missing",
        subtitle = "Variável Valor de Compra") +
   scale_y_continuous(labels = scales::percent_format()) +
   ggtitle("Análise da Porcentagem de Missing por Município") +
   theme(plot.title = element_text(hjust = 0.5),
         plot.subtitle = element_text(hjust = 0.5, size = 10),
         axis.text.x = element_text(angle = 90, hjust = 1)) +
   geom_text(aes(label = porcentagem_missing), angle = 90, vjust = 0.5, hjust = 1.5)
```




```
rm("df")
```

É possível observar nesta análise que o município de Mesquita que estava em último na análise de quantidade de missing agora é o primeiro. Para compreender melhor, é necessário entender a frequência relativa deste município na pesquisa. O que se pode deduzir neste primeiro momento é que 100% dos postos pesquisados não informaram o valor da variável alvo desta análise. Tal afirmação é observada pela quantidade de valores em n_linhas igual a n_missing, onde n_linhas representa a quatidade de postos pesquisados.

É possível observar que somente Mesquita, dentre os cinco primeiros municípios, faz parte dos municípios da região metropolitana do Rio de Janeiro.

Realizando teste de Little para checar se os dados faltantes são completamente aleatórios

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1  2.27e-21     0       0                2
```

Realizando a imputação de dados

```
## Class: mids
```

```
## Number of multiple imputations:  1
## Imputation methods:
##              Data            Regiao                UF         Municipio
##                ""                ""                ""                ""
##      CNPJ_Revenda           Produto          Bandeira    Valor_de_Venda
##                ""                ""                ""                ""
##   Valor_de_Compra            Cambio            PPI_DC Brent_Real_Barril
##    "norm.predict"                ""                ""                ""
## PredictorMatrix:
##              Data Regiao UF Municipio CNPJ_Revenda Produto Bandeira
## Data            0      0  0         0            0       0        0
## Regiao          1      0  0         0            0       0        0
## UF              1      0  0         0            0       0        0
## Municipio       1      0  0         0            0       0        0
## CNPJ_Revenda    1      0  0         0            0       0        0
## Produto         1      0  0         0            0       0        0
##              Valor_de_Venda Valor_de_Compra Cambio PPI_DC Brent_Real_Barril
## Data                      1               1      1      1                 1
## Regiao                    1               1      1      1                 1
## UF                        1               1      1      1                 1
## Municipio                 1               1      1      1                 1
## CNPJ_Revenda              1               1      1      1                 1
## Produto                   1               1      1      1                 1
## Number of logged events:  6
##   it im dep     meth          out
## 1  0  0     constant       Regiao
## 2  0  0     constant           UF
## 3  0  0     constant    Municipio
## 4  0  0     constant CNPJ_Revenda
## 5  0  0     constant      Produto
## 6  0  0     constant     Bandeira
```