Atividade programação semana 1 - Módulo 11 - Sku_price

Livia Coutinho

2024-08-10

Introdução

Este relatório apresenta uma análise de um conjunto de dados contendo informações sobre preços de produtos, incluindo IDs de produtos e datas de início e término dos preços (sku_price).

ANÁLISE

```
# Carregar pacotes
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(gridExtra)
```

```
##
## Anexando pacote: 'gridExtra'
##
## O seguinte objeto é mascarado por 'package:dplyr':
##
##     combine
```

```
library(readr)
library(skimr)
library(psych)
```

```
##
## Anexando pacote: 'psych'
##
## Os seguintes objetos são mascarados por 'package:ggplot2':
##
##     %+%, alpha
```

```
library(ggplot2)

# Carregar o conjunto de dados
arquivo <- "C:/Users/Inteli/Desktop/GitHub (módulo 11)/progS1/sku_price - Sheet1.csv"
dados <- read.csv(arquivo)

# Mostrar as primeiras linhas dos dados
head(dados)
```

```
##         SKU_ID   START_DT     END_DT PRICE_AMT
## 1 134784109083 2019-01-01 2019-10-13     46.99
## 2 134784109083 2019-10-14 2020-05-30     42.99
## 3 134784109083 2020-05-31 2021-04-14     53.99
## 4 134784109083 2021-04-15 2021-06-15     59.99
## 5 965005157985 2019-01-01 2019-03-19     42.99
## 6 965005157985 2019-03-20 2019-11-28     48.99
```

```
print(dados)
```

```
##           SKU_ID   START_DT     END_DT PRICE_AMT
## 1   134784109083 2019-01-01 2019-10-13     46.99
## 2   134784109083 2019-10-14 2020-05-30     42.99
## 3   134784109083 2020-05-31 2021-04-14     53.99
## 4   134784109083 2021-04-15 2021-06-15     59.99
## 5   965005157985 2019-01-01 2019-03-19     42.99
## 6   965005157985 2019-03-20 2019-11-28     48.99
## 7   965005157985 2019-11-29 2020-11-04     59.99
## 8   965005157985 2020-11-05 2021-09-19     50.99
## 9    24111701977 2019-01-01 2019-12-03     76.99
## 10   24111701977 2019-12-04 2020-09-16     72.99
## 11   24111701977 2020-09-17 2021-08-19     78.99
## 12   24111701977 2021-08-20 2022-03-24     84.99
## 13   24111701977 2022-03-25 2022-08-30     80.99
## 14   24111701977 2022-08-31 2023-01-29     86.99
## 15   24111701977 2023-01-30 2023-09-11     97.99
## 16  689001624189 2019-01-01 2019-12-20     16.99
## 17  689001624189 2019-12-21 2020-02-22     27.99
## 18  689001624189 2020-02-23 2020-05-23     18.99
## 19  689001624189 2020-05-24 2020-12-03      9.99
## 20  194090141801 2019-01-01 2019-03-14     34.99
## 21  194090141801 2019-03-15 2019-08-30     30.99
## 22  194090141801 2019-08-31 2019-11-08     41.99
## 23  194090141801 2019-11-09 2020-08-13     37.99
## 24  593180751003 2019-01-01 2019-07-20     53.99
## 25  593180751003 2019-07-21 2019-10-16     64.99
## 26  593180751003 2019-10-17 2020-10-01     75.99
## 27  593180751003 2020-10-02 2021-09-25     81.99
## 28  593180751003 2021-09-26 2022-05-07     87.99
## 29  725969988687 2019-01-01 2019-11-21     50.99
## 30  725969988687 2019-11-22 2020-10-28     41.99
## 31  725969988687 2020-10-29 2021-10-15     47.99
## 32  725969988687 2021-10-16 2022-01-26     38.99
## 33  725969988687 2022-01-27 2022-07-09     44.99
## 34  725969988687 2022-07-10 2023-06-17     50.99
## 35  298825663061 2019-01-01 2019-04-17     34.99
## 36  298825663061 2019-04-18 2019-09-29     40.99
## 37  298825663061 2019-09-30 2020-05-08     31.99
## 38  298825663061 2020-05-09 2021-02-23     27.99
## 39  401200387128 2019-01-01 2019-10-27     62.99
## 40  401200387128 2019-10-28 2020-08-20     73.99
## 41  401200387128 2020-08-21 2021-05-10     79.99
## 42  401200387128 2021-05-11 2021-12-23     75.99
## 43  401200387128 2021-12-24 2022-06-10     66.99
## 44  401200387128 2022-06-11 2022-12-05     57.99
## 45  401200387128 2022-12-06 2023-04-13     63.99
## 46  401200387128 2023-04-14 2023-09-06     59.99
## 47  304604508758 2019-01-01 2019-09-24     57.99
## 48  304604508758 2019-09-25 2020-04-07     48.99
## 49  304604508758 2020-04-08 2020-10-03     54.99
## 50  304604508758 2020-10-04 2021-08-30     65.99
## 51  509299973007 2019-01-01 2019-10-23     66.99
## 52  509299973007 2019-10-24 2019-12-25     57.99
## 53  509299973007 2019-12-26 2020-04-15     63.99
## 54  509299973007 2020-04-16 2021-03-02     69.99
## 55  509299973007 2021-03-03 2021-11-10     80.99
## 56  509299973007 2021-11-11 2022-02-22     91.99
## 57  509299973007 2022-02-23 2022-08-31     82.99
## 58  380925956416 2019-01-01 2019-09-21     74.99
## 59  380925956416 2019-09-22 2020-02-24     65.99
## 60  380925956416 2020-02-25 2020-05-18     76.99
## 61  380925956416 2020-05-19 2021-05-02     72.99
## 62  380925956416 2021-05-03 2021-11-01     68.99
## 63  380925956416 2021-11-02 2022-03-31     64.99
## 64  380925956416 2022-04-01 2022-10-16     75.99
## 65   88306780994 2019-01-01 2019-06-15     94.99
## 66   88306780994 2019-06-16 2020-01-03    105.99
## 67   88306780994 2020-01-04 2020-03-13    101.99
## 68   88306780994 2020-03-14 2020-07-15     97.99
## 69   88306780994 2020-07-16 2021-06-14    103.99
## 70   88306780994 2021-06-15 2021-10-01     99.99
## 71   88306780994 2021-10-02 2021-12-08    110.99
## 72   88306780994 2021-12-09 2022-02-27    116.99
## 73  526136230581 2019-01-01 2019-11-19     26.99
## 74  526136230581 2019-11-20 2020-05-18     37.99
## 75  526136230581 2020-05-19 2021-04-01     48.99
## 76  526136230581 2021-04-02 2021-10-29     59.99
## 77  526136230581 2021-10-30 2022-02-18     70.99
## 78  526136230581 2022-02-19 2022-10-21     66.99
## 79  526136230581 2022-10-22 2023-07-23     62.99
## 80  923366404315 2019-01-01 2019-10-16     70.99
## 81  923366404315 2019-10-17 2020-09-06     66.99
## 82  923366404315 2020-09-07 2021-07-30     72.99
## 83  923366404315 2021-07-31 2022-01-06     63.99
## 84  923366404315 2022-01-07 2022-03-13     69.99
## 85  923366404315 2022-03-14 2022-06-22     80.99
## 86  923366404315 2022-06-23 2023-02-05     76.99
## 87  923366404315 2023-02-06 2023-12-26     67.99
## 88  112757389373 2019-01-01 2019-03-30     55.99
## 89  112757389373 2019-03-31 2019-07-19     51.99
## 90  112757389373 2019-07-20 2019-10-28     47.99
## 91  112757389373 2019-10-29 2020-08-29     58.99
## 92  112757389373 2020-08-30 2021-05-29     54.99
## 93  112757389373 2021-05-30 2021-10-10     45.99
## 94  145338904399 2019-01-01 2019-09-23     60.99
## 95  145338904399 2019-09-24 2019-12-03     66.99
## 96  145338904399 2019-12-04 2020-06-27     57.99
## 97  145338904399 2020-06-28 2020-12-17     63.99
## 98  145338904399 2020-12-18 2021-10-10     74.99
## 99  145338904399 2021-10-11 2022-01-03     65.99
## 100 145338904399 2022-01-04 2022-10-16     61.99
## 101 145338904399 2022-10-17 2023-04-30     72.99
## 102 397072998268 2019-01-01 2019-04-02     58.99
## 103 397072998268 2019-04-03 2019-09-08     54.99
## 104 397072998268 2019-09-09 2020-06-30     45.99
## 105 397072998268 2020-07-01 2020-11-11     36.99
## 106 571123564033 2019-01-01 2019-03-17     84.99
## 107 571123564033 2019-03-18 2019-05-18     80.99
## 108 571123564033 2019-05-19 2019-09-20     71.99
## 109 571123564033 2019-09-21 2020-07-23     62.99
## 110 571123564033 2020-07-24 2021-01-19     58.99
## 111 571123564033 2021-01-20 2021-06-20     49.99
## 112 571123564033 2021-06-21 2022-01-27     45.99
## 113 290853023558 2019-01-01 2019-03-31     85.99
## 114 290853023558 2019-04-01 2019-06-05     96.99
## 115 290853023558 2019-06-06 2020-04-03    107.99
## 116 290853023558 2020-04-04 2021-02-06    118.99
## 117 290853023558 2021-02-07 2021-09-11    129.99
```

```
nrow(dados)
```

```
## [1] 117
```

```
# Verificação da estrutura dos dados
str(dados)
```

```
## 'data.frame':    117 obs. of  4 variables:
##  $ SKU_ID   : num  1.35e+11 1.35e+11 1.35e+11 1.35e+11 9.65e+11 ...
##  $ START_DT : chr  "2019-01-01" "2019-10-14" "2020-05-31" "2021-04-15" ...
##  $ END_DT   : chr  "2019-10-13" "2020-05-30" "2021-04-14" "2021-06-15" ...
##  $ PRICE_AMT: num  47 43 54 60 43 ...
```

```
glimpse(dados)
```

```
## Rows: 117
## Columns: 4
## $ SKU_ID    <dbl> 134784109083, 134784109083, 134784109083, 134784109083, 9650…
## $ START_DT  <chr> "2019-01-01", "2019-10-14", "2020-05-31", "2021-04-15", "201…
## $ END_DT    <chr> "2019-10-13", "2020-05-30", "2021-04-14", "2021-06-15", "201…
## $ PRICE_AMT <dbl> 46.99, 42.99, 53.99, 59.99, 42.99, 48.99, 59.99, 50.99, 76.9…
```

```
#RESUMO ESTATÍSTICO

# Resumo estatístico das variáveis
summary(dados)
```

```
##      SKU_ID            START_DT            END_DT            PRICE_AMT     
##  Min.   :2.411e+10   Length:117         Length:117         Min.   :  9.99  
##  1st Qu.:1.453e+11   Class :character   Class :character   1st Qu.: 49.99  
##  Median :3.971e+11   Mode  :character   Mode  :character   Median : 63.99  
##  Mean   :4.097e+11                                         Mean   : 64.81  
##  3rd Qu.:5.711e+11                                         3rd Qu.: 76.99  
##  Max.   :9.650e+11                                         Max.   :129.99  
```

```
# Resumo estatístico das variáveis numéricas
skim(dados)
```

Data summary

Name	dados
Number of rows	117
Number of columns	4
_______________________	
Column type frequency:	
character	2
numeric	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
START_DT	0	1	10	10	0	95	0
END_DT	0	1	10	10	0	112	0

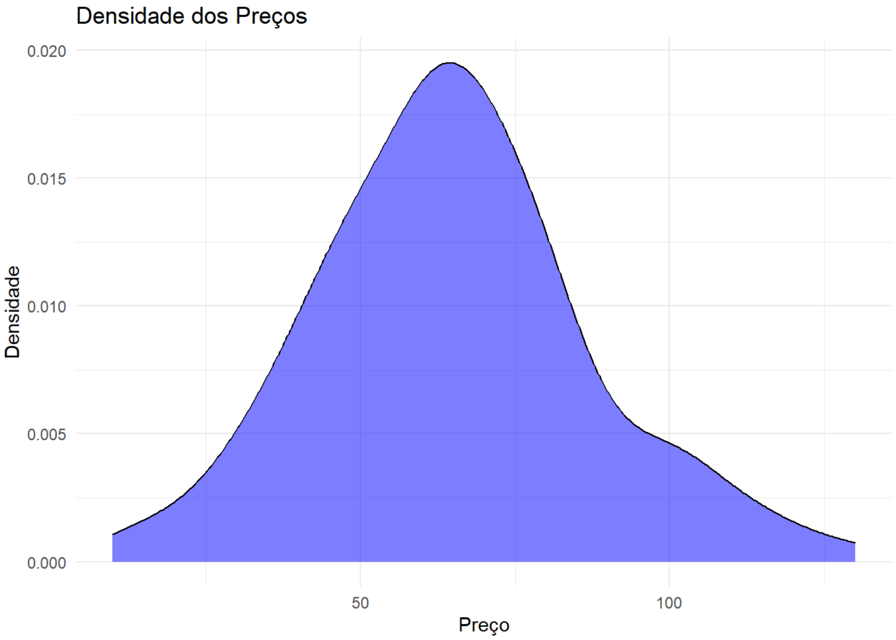
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hi
SKU_ID	0	1	4.096833e+11	2.688745e+11	2.41117e+10	1.453389e+11	3.97073e+11	5.711236e+11	9.650052e+11	■
PRICE_AMT	0	1	6.481000e+01	2.211000e+01	9.99000e+00	4.999000e+01	6.39900e+01	7.699000e+01	1.299900e+02	⌟

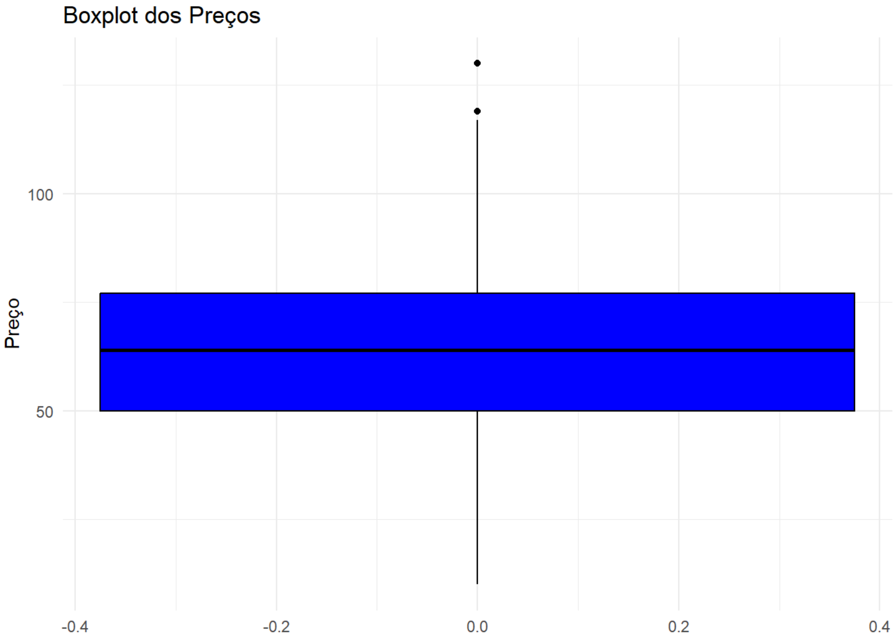
```
# Resumo estatístico com a função describe (psych)
describe(dados)
```

```
##           vars   n         mean           sd      median      trimmed
## SKU_ID       1 117 4.096833e+11 2.688745e+11 3.97073e+11 3.903932e+11
## START_DT*    2 117 3.993000e+01 3.035000e+01 3.80000e+01 3.864000e+01
## END_DT*      3 117 5.638000e+01 3.240000e+01 5.60000e+01 5.636000e+01
## PRICE_AMT    4 117 6.481000e+01 2.211000e+01 6.39900e+01 6.412000e+01
##                   mad         min          max        range skew kurtosis
## SKU_ID    3.009424e+11 2.41117e+10 9.650052e+11 940893456008 0.44    -0.72
## START_DT* 4.151000e+01 1.00000e+00 9.500000e+01           94 0.20    -1.30
## END_DT*   4.151000e+01 1.00000e+00 1.120000e+02          111 0.02    -1.24
## PRICE_AMT 1.927000e+01 9.99000e+00 1.299900e+02          120 0.29     0.24
##                     se
## SKU_ID    2.485745e+10
## START_DT* 2.810000e+00
## END_DT*   3.000000e+00
## PRICE_AMT 2.040000e+00
```

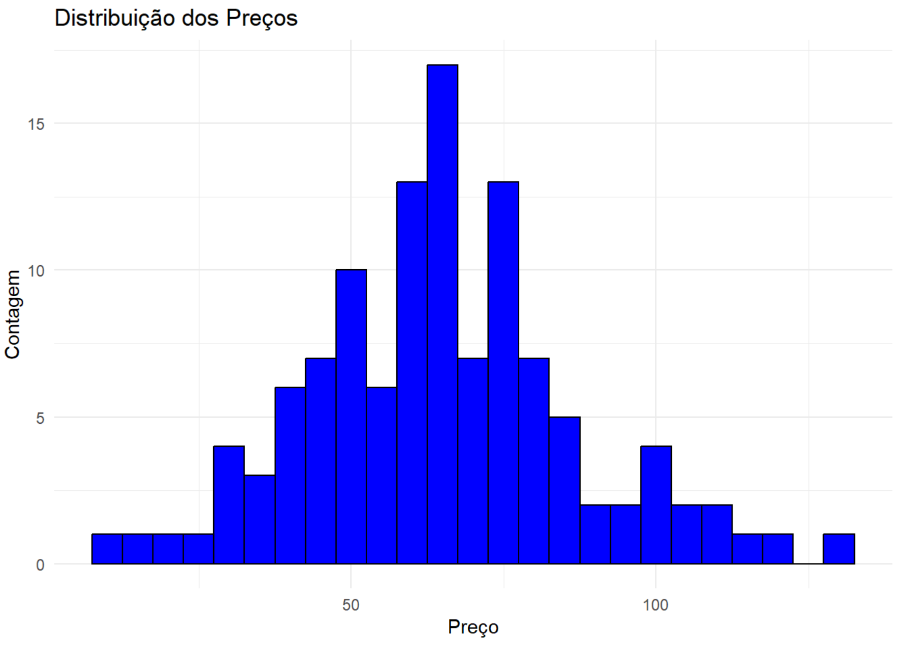
```
#VISUALIZAÇÃO DAS DISTRIBUIÇÕES

# Formatar as datas (O histograma não estava dando certo sem formatar a data)
dados$START_DT <- as.Date(dados$START_DT, format = "%Y-%m-%d")
dados$END_DT <- as.Date(dados$END_DT, format = "%Y-%m-%d")

# Gráfico de densidade - variável PRICE_AMT
ggplot(dados, aes(x = PRICE_AMT)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Densidade dos Preços", x = "Preço", y = "Densidade") +
  theme_minimal()
```



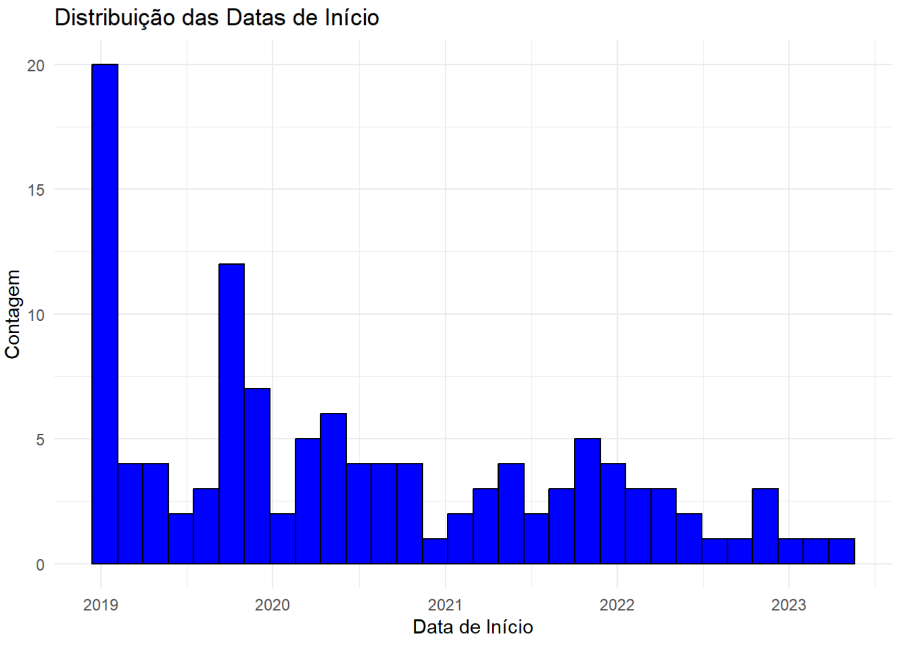
```
# Boxplot - variável PRICE_AMT
ggplot(dados, aes(y = PRICE_AMT)) +
  geom_boxplot(fill = "blue", color = "black") +
  labs(title = "Boxplot dos Preços", y = "Preço") +
  theme_minimal()
```



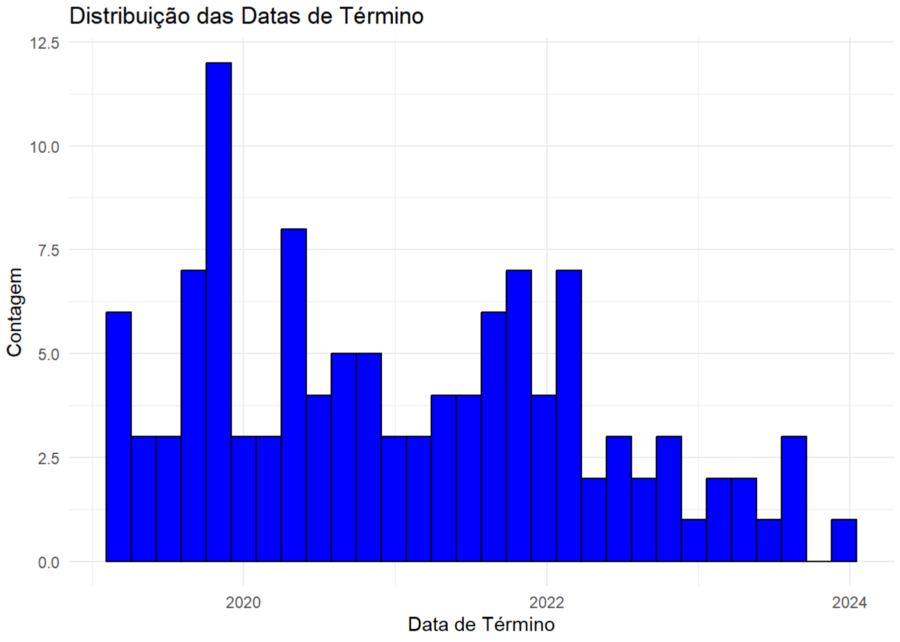
```
# Histograma - variável PRICE_AMT
ggplot(dados, aes(x = PRICE_AMT)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribuição dos Preços", x = "Preço", y = "Contagem") +
  theme_minimal()
```



```
# Histograma - variável START_DT
ggplot(dados, aes(x = START_DT)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Distribuição das Datas de Início", x = "Data de Início", y = "Contagem") +
  theme_minimal()
```



```
# Histograma - variável END_DT
ggplot(dados, aes(x = END_DT)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Distribuição das Datas de Término", x = "Data de Término", y = "Contagem") +
  theme_minimal()
```

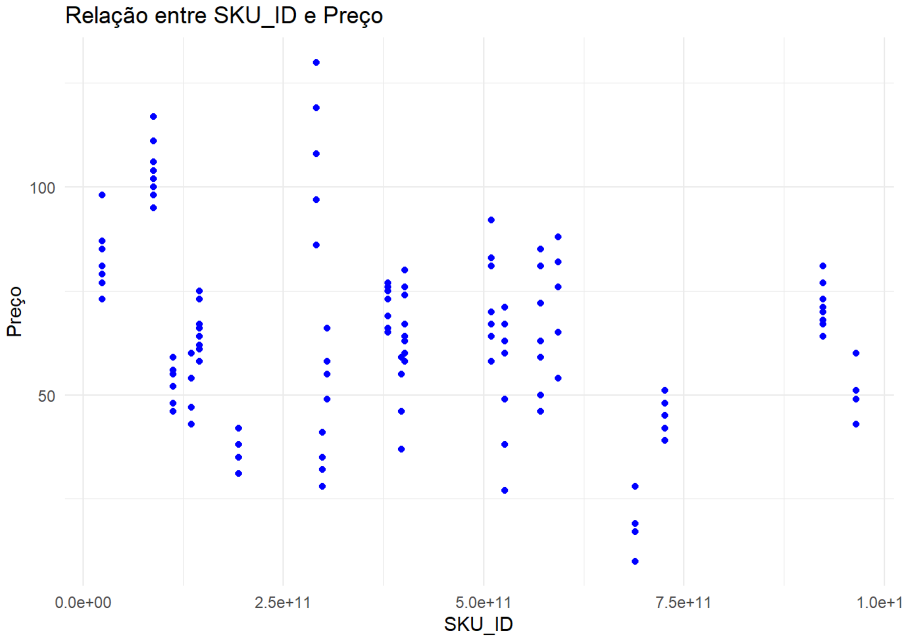


Distribuição das Datas de Término

```
#IDENTIFICAÇÃO DE OUTLIERS

# Calcular estatísticas descritivas para PRICE_AMT
summary_stats <- summary(dados$PRICE_AMT)
IQR_value <- IQR(dados$PRICE_AMT)

# Identificar outliers
lower_bound <- summary_stats["1st Qu."] - 1.5 * IQR_value
upper_bound <- summary_stats["3rd Qu."] + 1.5 * IQR_value

outliers <- dados$PRICE_AMT[dados$PRICE_AMT < lower_bound | dados$PRICE_AMT > upper_bound]
outliers
```

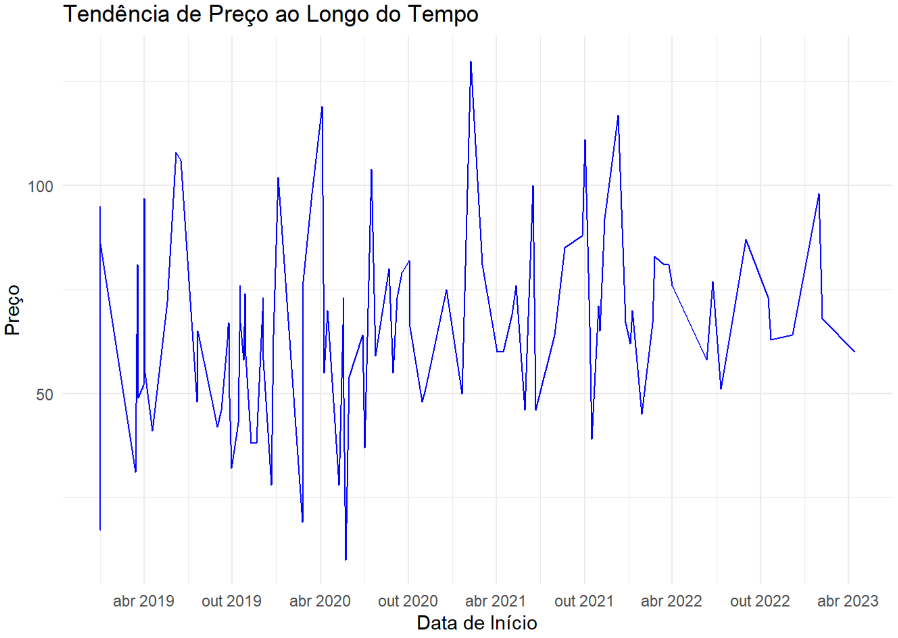
```
## [1] 118.99 129.99
```

```
#ANÁLISE BIVARIADA

# Gráfico de dispersão entre PRICE_AMT e SKU_ID
ggplot(dados, aes(x = SKU_ID, y = PRICE_AMT)) +
  geom_point(color = "blue") +
  labs(title = "Relação entre SKU_ID e Preço", x = "SKU_ID", y = "Preço") +
  theme_minimal()
```

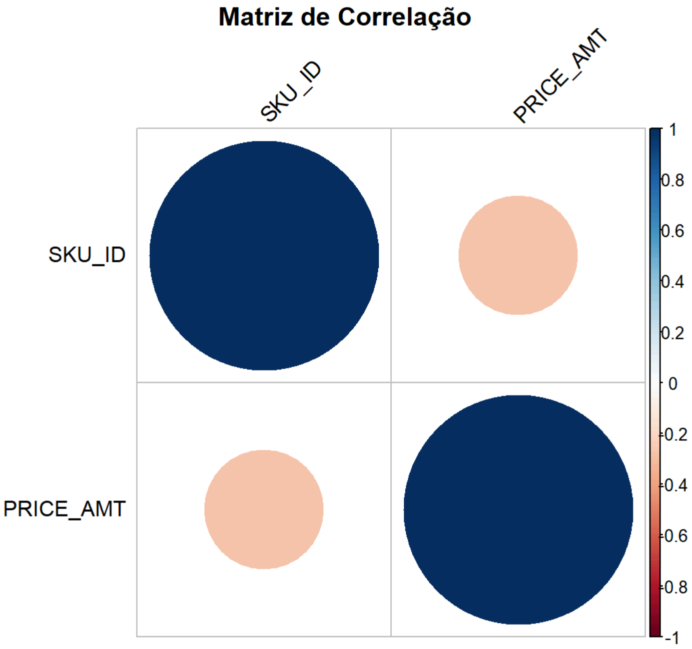


Relação entre SKU_ID e Preço

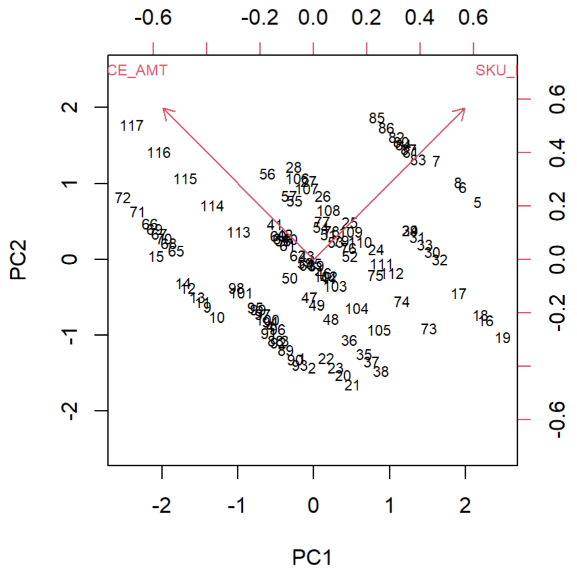
```
# Gráfico de linha para visualizar a tendência de preço ao longo do tempo
ggplot(dados, aes(x = START_DT, y = PRICE_AMT)) +
  geom_line(color = "blue") +
  labs(title = "Tendência de Preço ao Longo do Tempo", x = "Data de Início", y = "Preço") +
  theme_minimal() +
  scale_x_date(date_breaks = "6 months", date_labels = "%b %Y")
```



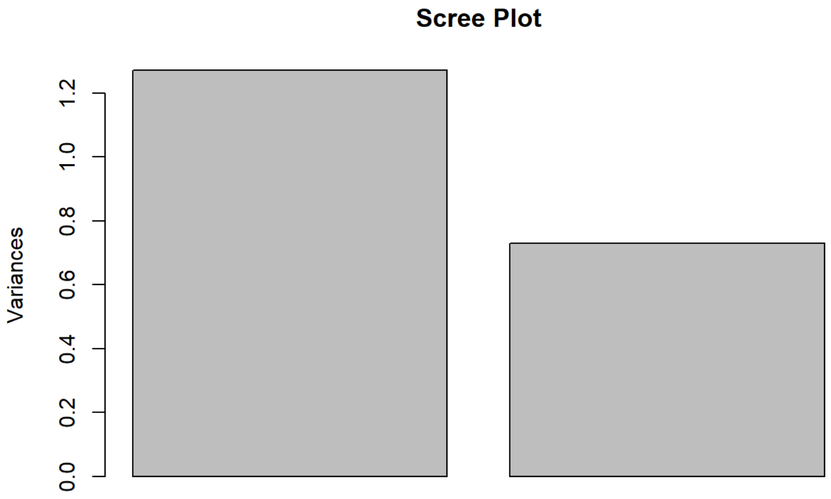
```
# Análise de correlação
matriz_correlacao <- cor(dados %>% select_if(is.numeric), use = "complete.obs")
corrplot(matriz_correlacao, method = "circle", type = "full",
         tl.col = "black", tl.srt = 45,
         title = "Matriz de Correlação",
         mar = c(0,0,1,0))
```



```
#ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

# Seleção e padronização das variáveis numéricas
dados_numericos <- dados %>% select_if(is.numeric)
dados_pca <- scale(dados_numericos)

# Realizar PCA
pca_result <- prcomp(dados_pca, center = TRUE, scale. = TRUE)

# Biplot para visualizar os resultados
biplot(pca_result, scale = 0, cex = 0.7)
```



```
# Scree Plot para plotar a variância explicada por cada componente
screeplot(pca_result, main = "Scree Plot")
```



```
# Resumo dos componentes principais
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2
## Standard deviation     1.1274 0.8538
## Proportion of Variance 0.6355 0.3645
## Cumulative Proportion  0.6355 1.0000
```

Sumário e discussão

Dados

O conjunto contém preços de produtos com IDs, datas de início e término dos preços.

Distribuição dos preços

- A maioria dos preços está em uma faixa específica.
- Identificamos alguns valores extremos como outliers (118.99 e 129.99).

Distribuição das datas

- As datas de início e término estão bem distribuídas ao longo do período.

Relação entre variáveis

- Não há uma correlação clara entre o ID do produto e o preço.
- Os preços mostram variações ao longo do tempo.

PCA

- Os primeiros componentes principais explicam a maior parte da variância nos dados.

Limitações e melhorias

Limitações

- **Dados**: Presença de outliers e possíveis erros.
- **Variáveis Categóricas**: Não foram incluídas na análise.

Melhorias

- **Limpeza de Dados**: Tratar outliers e erros.
- **Incluir Mais Variáveis**: Adicionar variáveis categóricas na análise.
- **Explorar Tendências**: Analisar padrões de preços com mais detalhes.