



Universidade de São Paulo

Escola Politécnica

Especialização em Engenharia de Dados e Big Data

Disciplina Análises Preditivas

Prova

Análises Preditivas

Érika Carvalho de Aquino

São Paulo
2025

Sumário

1) Criação de variáveis	2
1.1) Crie uma variável chamada: Dia_da_semana e popule-na com (dom, seg, ter, etc)	2
1.2) Crie uma variável chamada: Mês_venda e popule-na com (jan,fev,mar,abr, etc)	3
2) Façam as análises univariadas	4
2.1) Análises descritivas	4
2.2) Análises gráficas: gráfico de barras, histogramas, etc	9
3) Façam as análises bi-variadas	14
3.1) AnaliseS graficas bi variadas: Grafico de dispersão (Vendas(y) x Temperatura (x))	14
3.2) calcule a correlação entre as variaveis Temperatura e total de vendas (R\$)23	
4) Definam a equacao através de uma Regressão Linear	25
5) Prevejam qual será o total a ser faturado com as vendas (em R\$) na data de 30/out/23, sabendo que a previsao do tempo aponta para 18o C	27
6) Escreva a sua conclusão	28
7) Escreva a sua recomendação para o Diretor Fianceiro desta empresa	29

Enunciado: Considere uma indústria que fabrica sorvetes e você está responsável por fazer as estimativas de vendas da área financeira

OBJETIVO: Estimar qual será o volume de vendas (em R\$) de sorvete no dia 30/out/2023

A atividade foi realizada usando os softwares R (R Core Team (2023). *_R: A Language and Environment for Statistical Computing_*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>) e RStudio (RStudio Team (2023). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA. URL: <https://posit.co/>).

A documentação completa pode ser consultada em: <https://github.com/Erika-Aquino/An-lises-Preditivas---Atividade-final.git>

1) Criação de variáveis

1.1) Crie uma variável chamada: Dia da semana e popule-na com (dom, seg, ter, etc)

A variável `Dia_da_semana` foi criada a partir da coluna `Date`, inicialmente convertida para o formato de data sem a informação de hora. Em seguida, utilizou-se a função `weekdays()` para identificar o dia da semana correspondente a cada data, e, por meio da função `case_when()`, os nomes completos foram transformados em abreviações padronizadas, resultando nos valores “dom”, “seg”, “ter”, “qua”, “qui”, “sex” e “sáb” para cada observação do banco de dados (Figura 1).

```
# Converter Date para Date (removendo hora) e criar Dia_da_semana
bd <- bd %>%
  mutate(
    Data = as.Date(Date), # Cria coluna apenas com a data (sem hora)
    Dia_da_semana = weekdays(Date)
  ) %>%
  mutate(
    Dia_da_semana = case_when(
      Dia_da_semana == "domingo" ~ "dom",
      Dia_da_semana == "segunda-feira" ~ "seg",
      Dia_da_semana == "terça-feira" ~ "ter",
      Dia_da_semana == "quarta-feira" ~ "qua",
      Dia_da_semana == "quinta-feira" ~ "qui",
      Dia_da_semana == "sexta-feira" ~ "sex",
      Dia_da_semana == "sábado" ~ "sáb",
      TRUE ~ Dia_da_semana
    )
  )
head(bd)
```

Figura 1 - Script para criação da variável Dia_da_semana a partir da coluna Date.

1.2) Crie uma variável chamada: Mês_venda e popule-na com (jan,fev,mar,abr, etc)

A variável Mês_venda foi criada a partir da coluna Data, utilizando a função month() do pacote lubridate. Essa função permitiu extrair o mês de cada observação, e, com os argumentos label = TRUE e abbr = TRUE, os meses foram apresentados em formato abreviado. Além disso, foi definido o parâmetro locale = "pt_BR.UTF-8", garantindo que os nomes aparecessem em português. Dessa forma, cada linha do banco de dados passou a conter o mês correspondente à data de venda, expresso em abreviações padronizadas como "jan", "fev", "mar", "abr" e assim por diante até "dez" (Figura 2).

```
# Criar a variável Mês_venda com nomes abreviados em português
bd <- bd %>%
  mutate(
    Mês_venda = month(Date, label = TRUE, abbr = TRUE, locale = "pt_BR.UTF-8")
  )
head(bd$Mês_venda)
```

Figura 2 - Script para Criação da variável Mês_venda a partir da coluna Data, utilizando a função month() do pacote lubridate para gerar abreviações dos meses em português (jan, fev, mar, abr, etc.).

2) Façam as análises univariadas

2.1) Análises descritivas

Foram executadas estatísticas descritivas univariadas para todas as variáveis do banco de dados. Primeiramente, utilizou-se o comando `str(bd)` para verificar a estrutura do banco, identificando os tipos de dados presentes. Em seguida, foram calculadas medidas descritivas para as variáveis numéricas por meio das funções `summary()` e `describe()`, permitindo observar medidas de tendência central (média e mediana), dispersão (desvio padrão, mínimo e máximo), além de assimetria e curtose. Para as variáveis categóricas `Dia_da_semana` e `Mês_venda`, foram construídas tabelas de frequência absolutas e relativas (`table()` e `prop.table()`), possibilitando analisar a distribuição percentual dos registros. Também foi feita a verificação de valores ausentes no banco com o comando `colSums(is.na(bd))`. Por fim, empregou-se a função `dfSummary()` do pacote `summarytools` para gerar um relatório detalhado em formato tabular, integrando estatísticas descritivas, frequências e informações sobre dados faltantes (Figura 3).

```
##### 2.1) Análises descritivas

# Ver estrutura do banco
str(bd)

# Estatísticas descritivas para variáveis numéricas
# Selecionar apenas variáveis numéricas
numericas <- bd %>% select(where(is.numeric))

# Descritivas padrão
summary(numericas)

# Descritivas completas (média, desvio padrão, assimetria, curtose etc)
describe(numericas)

# Tabelas de frequência para variáveis categóricas
# Frequência para Dia_da_semana
table(bd$Dia_da_semana)
prop.table(table(bd$Dia_da_semana)) * 100

# Frequência para Mês_venda
table(bd$Mês_venda)
prop.table(table(bd$Mês_venda)) * 100

# Verificar valores ausentes
colSums(is.na(bd))

# Gerar relatório detalhado
view(dfsummary(bd))
citation("rstudio")
```

Figura 3 - Script das análises descritivas univariadas, incluindo estatísticas para variáveis numéricas, frequências de variáveis categóricas, verificação de valores ausentes e geração de relatório detalhado.

Estatísticas descritivas das variáveis numéricas (Tabela 1):

- **Temperature** variou entre 41,82 e 98,88, com média de 73,91. O valor mediano foi 74,31, indicando uma distribuição aproximadamente simétrica em torno da média. Foram observados 7 valores ausentes.

- **Total das vendas (R\$)** apresentou mínimo de R\$ 28.238 e máximo de R\$ 55.023, com média de R\$ 40.155 e mediana de R\$ 39.053. Isso sugere uma distribuição relativamente equilibrada, mas com alguma dispersão. Foram identificados 8 valores ausentes.

- **Temperatura (graus celsius)** oscilou entre 11 e 42 graus, com média de 25,43 e mediana de 24, evidenciando valores centrais próximos e distribuição levemente assimétrica. Houve 7 registros ausentes.

Tabela 1 - Estatísticas descritivas das variáveis numéricas do banco de dados

Variável	Min	1º Quartil	Mediana	Média	3º Quartil	Máx	NA's
Temperature	41,82	62,10	74,31	73,91	86,10	98,88	7
Total das vendas (R\$)	28.238	36.365	39.053	40.155	45.452	55.023	8
Temperatura (graus celsius)	11,00	21,00	24,00	25,43	31,00	42,00	7

Dando continuidade às análises descritivas, também foram avaliadas as variáveis categóricas Dia_da_semana e Mês_venda, por meio de tabelas de frequência absolutas e relativas. Esse procedimento permitiu identificar a distribuição percentual dos registros ao longo dos dias da semana e dos meses do ano, destacando padrões de ocorrência no banco de dados. Além disso, realizou-se a verificação de valores ausentes com a função `colSums(is.na(bd))`, possibilitando quantificar a presença de dados faltantes em cada variável e assegurando maior confiabilidade para as análises subsequentes (Figura 4).

```
# Tabelas de frequência para variáveis categóricas
# Frequência para Dia_da_semana
table(bd$Dia_da_semana)
prop.table(table(bd$Dia_da_semana)) * 100

# Frequência para Mês_venda
table(bd$Mês_venda)
prop.table(table(bd$Mês_venda)) * 100

# Verificar valores ausentes
colSums(is.na(bd))

# Gerar relatório detalhado
view(dfSummary(bd))
citation("rstudio")
```

Figura 4 - Script para cálculo das frequências de variáveis categóricas e verificação de valores ausentes

A análise da variável Dia_da_semana revelou uma distribuição praticamente uniforme das observações entre os sete dias. Cada dia da semana concentrou aproximadamente 1.026 registros, correspondendo a cerca de 14,3% do total. A pequena diferença observada na segunda-feira (1.027 registros, 14,30%) não é estatisticamente relevante, confirmando que os dados encontram-se equilibrados quanto à variável temporal (Tabela 2).

Tabela 2 - Distribuição da variável Dia_da_semana

Dia da semana	Frequência	Percentual (%)
dom	1.026	14,28
seg	1.027	14,30
ter	1.026	14,28
qua	1.026	14,28
qui	1.026	14,28
sex	1.026	14,28
sáb	1.026	14,28
Total	7.183	100,0


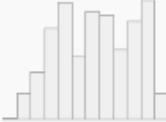
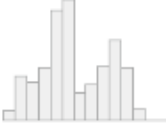
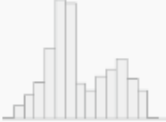



A distribuição da variável Mês_venda mostrou-se relativamente equilibrada entre os doze meses do ano, variando entre 536 registros (fevereiro, 7,46%) e 620 registros (março, maio, junho, julho e agosto, 8,63%). Essa pequena diferença está associada à quantidade de dias de cada mês, especialmente fevereiro, que possui menos dias no calendário. De forma geral, não houve concentração expressiva em nenhum mês específico, indicando uma distribuição temporal estável ao longo do ano (Tabela 3).

Tabela 3 - Distribuição da variável Mês_venda

Mês	Frequência	Percentual (%)
jan	589	8,20
fev	536	7,46
mar	620	8,63
abr	600	8,35
mai	620	8,63
jun	600	8,35
jul	620	8,63
ago	620	8,63
set	600	8,35
out	619	8,62
nov	570	7,94
dez	589	8,20
Total	7.483	100,0

Para complementar as análises descritivas realizadas, o Quadro 1 apresenta a visão consolidada das variáveis do banco de dados, incluindo medidas de posição e dispersão, frequência de valores distintos, representações gráficas e informações sobre dados válidos e ausentes. Destaca-se a baixa proporção de valores ausentes (menos de 0,2%), o que reforça a consistência do banco de dados para as etapas posteriores da análise preditiva.

Quadro 1 - Resumo das estatísticas descritivas das variáveis do banco de dados

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Date [POSIXct, POSIXt]	min : 2004-03-01 med : 2013-12-30 max : 2023-10-30 range : 19y 7m 29d	7183 distinct values		7183 (100.0%)	0 (0.0%)
2	Temperature [numeric]	Mean (sd) : 73.9 (13.7) min ≤ med ≤ max: 41.8 ≤ 74.3 ≤ 98.9 IQR (CV) : 24 (0.2)	7170 distinct values		7176 (99.9%)	7 (0.1%)
3	Total das vendas (R\$) [numeric]	Mean (sd) : 40154.7 (5642.7) min ≤ med ≤ max: 28238.2 ≤ 39052.6 ≤ 55022.6 IQR (CV) : 9086.7 (0.1)	7173 distinct values		7175 (99.9%)	8 (0.1%)
4	Temperatura (graus celsius) [numeric]	Mean (sd) : 25.4 (6.3) min ≤ med ≤ max: 11 ≤ 24 ≤ 42 IQR (CV) : 10 (0.2)	30 distinct values		7176 (99.9%)	7 (0.1%)
5	Data [Date]	min : 2004-03-01 med : 2013-12-30 max : 2023-10-30 range : 19y 7m 29d	7183 distinct values		7183 (100.0%)	0 (0.0%)
6	Dia_da_semana [character]	1. dom 2. qua 3. qui 4. sáb 5. seg 6. sex 7. ter	1026 (14.3%) 1026 (14.3%) 1026 (14.3%) 1026 (14.3%) 1027 (14.3%) 1026 (14.3%) 1026 (14.3%)		7183 (100.0%)	0 (0.0%)
7	Mês_venda [ordered, factor]	1. jan 2. fev 3. mar 4. abr 5. mai 6. jun 7. jul 8. ago 9. set 10. out [2 others]	589 (8.2%) 536 (7.5%) 620 (8.6%) 600 (8.4%) 620 (8.6%) 600 (8.4%) 620 (8.6%) 620 (8.6%) 600 (8.4%) 619 (8.6%) 1159 (16.1%)		7183 (100.0%)	0 (0.0%)

2.2) Análises gráficas: gráfico de barras, histogramas, etc

O trecho do script apresentado na Figura 5 mostra a utilização de ggplot2 para criar um histograma com classes de largura igual a 1 grau Celsius. A função `geom_histogram()` desenha as barras com preenchimento azul claro (`skyblue`) e bordas pretas, enquanto `labs()` define os rótulos dos eixos e o título do gráfico. O tema visual aplicado é o `theme_minimal()`, que confere uma aparência mais limpa ao gráfico.

```
##### 2.2) analise graficas: grafico de barras, histogramas, etc
# Histograma da Temperatura
ggplot(bd, aes(x = `Temperatura (graus celsius)`) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(
    title = "Distribuição das temperaturas registradas",
    x = "Temperatura (°C)",
    y = "Frequência"
  ) +
  theme_minimal()
```

Figura 5 - Script para Histograma da temperatura

A Figura 6 apresenta o histograma da variável Temperatura (°C), revelando a distribuição de frequência das temperaturas registradas diariamente na base de dados analisada. Observa-se uma distribuição bimodal, com dois picos principais: um entre 21 °C e 24 °C, e outro entre 30 °C e 33 °C. O primeiro pico sugere maior concentração de registros em temperaturas amenas, enquanto o segundo revela também uma expressiva ocorrência de dias quentes.

Essa distribuição bimodal pode refletir variações sazonais importantes, como invernos e verões bem definidos, ou diferenças regionais no conjunto de dados.

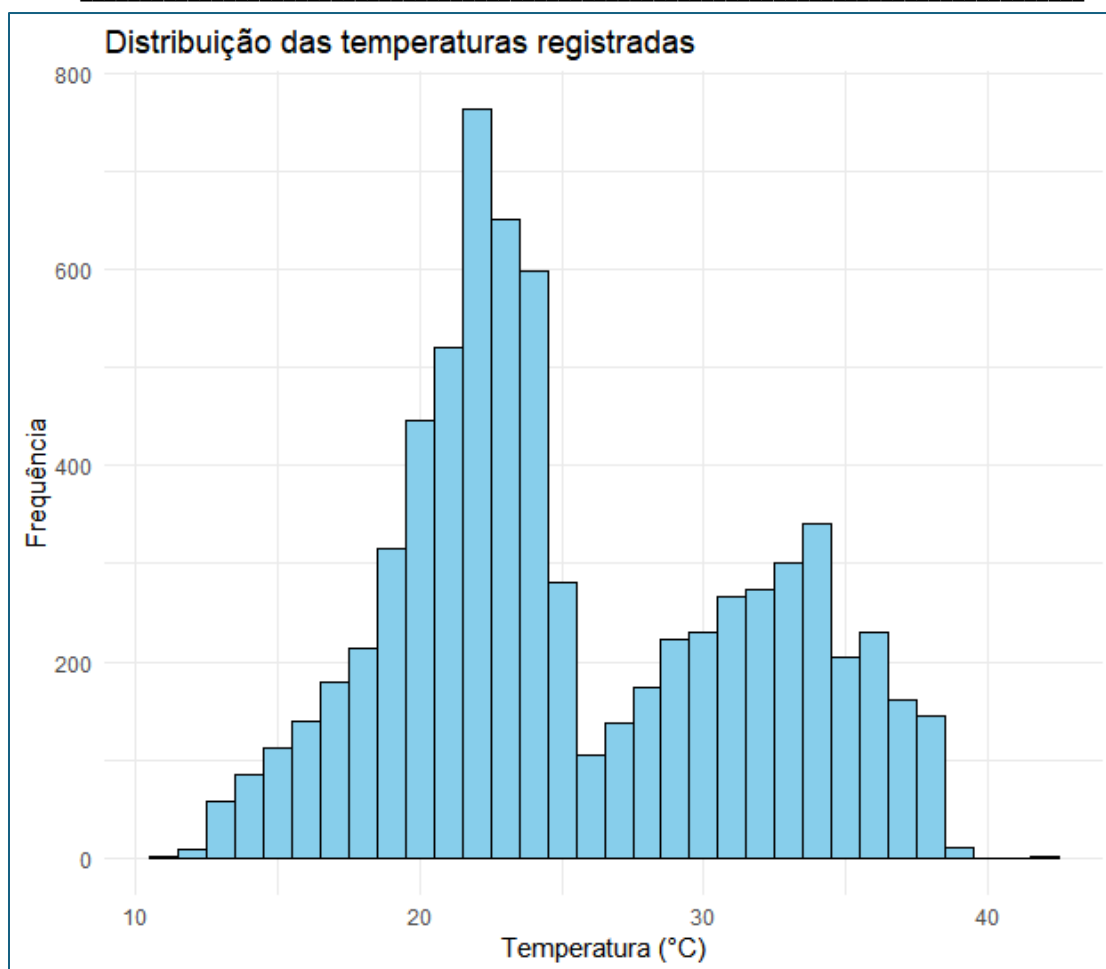


Figura 6 - Distribuição das temperaturas registradas ao longo da série histórica

O script apresentado na Figura 7 realiza a análise univariada da variável "Total das vendas (R\$)" por meio da construção de um histograma com a biblioteca ggplot2. Utiliza-se a base de dados, especificando-se no eixo x a variável de interesse. O histograma é gerado com largura de classe (binwidth) de R\$ 1.000, permitindo visualizar a frequência de dias em que o total diário de vendas se encontra dentro de cada faixa de valor. O gráfico recebe o título "Distribuição do total diário de vendas (R\$)", e os eixos são rotulados com os nomes correspondentes. Por fim, é aplicado o tema `theme_minimal()` para garantir uma apresentação mais limpa e legível.

```
# Histograma do Total de Vendas
ggplot(bd, aes(x = `Total das vendas (R$)`)) +
  geom_histogram(binwidth = 1000, fill = "lightgreen", color = "black") +
  labs(
    title = "Distribuição do total diário de vendas (R$)",
    x = "Total das vendas (R$)",
    y = "Frequência"
  ) +
  theme_minimal()
```

Figura 7 - Script para histograma do total de vendas

A Figura 8 apresenta o histograma da variável "Total das vendas (R\$)", permitindo observar como os valores de vendas se distribuem ao longo dos dias. É possível identificar uma distribuição aparentemente bimodal, com dois picos de frequência principais: um concentrado em torno de R\$ 39.000 e outro em torno de R\$ 48.000. Isso sugere que há grupos distintos de dias com volumes de venda significativamente diferentes, o que pode estar associado a fatores sazonais, variações de temperatura ou comportamento do consumidor em determinados períodos. A análise dessa distribuição auxilia na compreensão da variabilidade das vendas e pode orientar estratégias de previsão e planejamento financeiro.

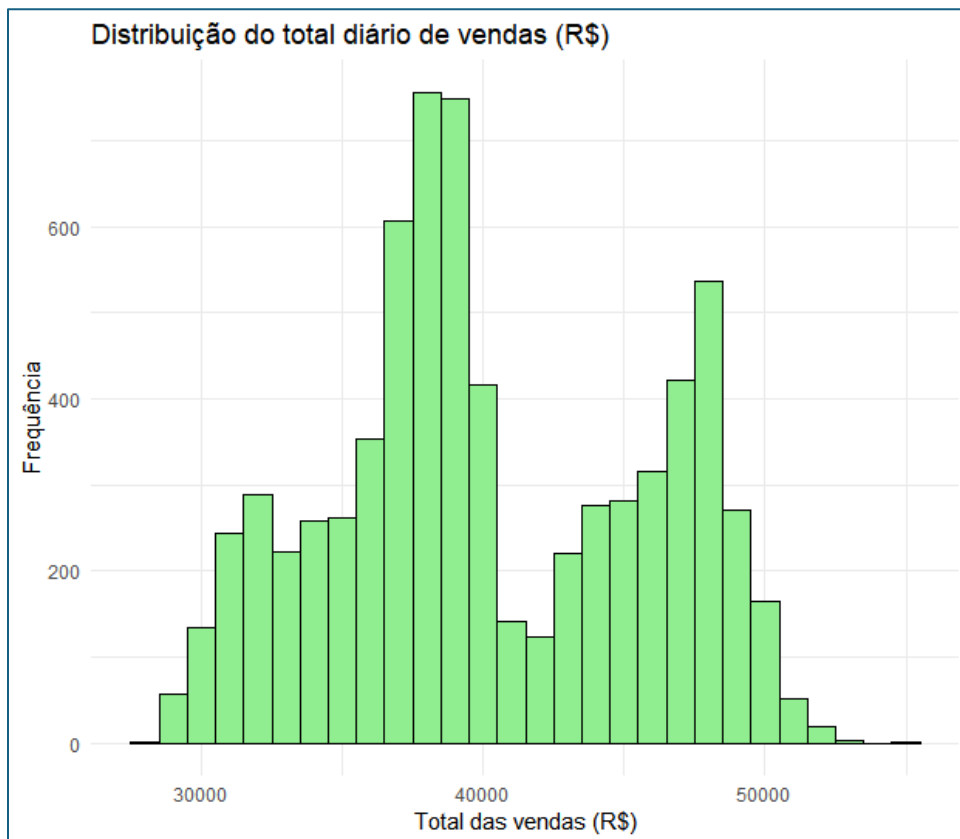


Figura 8 - Distribuição do total de vendas

A Figura 9 apresenta o script que tem como objetivo construir um gráfico de barras com as vendas médias diárias (em R\$) por dia da semana. Primeiramente, os dados são agrupados por Dia_da_semana e, em seguida, é calculada a média da variável Total das vendas (R\$) para cada grupo, ignorando valores ausentes (na.rm = TRUE). O resultado dessa agregação é então visualizado com ggplot2, por meio de barras coloridas em laranja (geom_col(fill = "orange")). O gráfico resultante apresenta no eixo x os dias da semana (por exemplo: segunda, terça, etc.) e no eixo y o valor médio de vendas correspondente, permitindo identificar em quais dias os resultados foram mais expressivos. A estética foi finalizada com theme_minimal() para um visual limpo e claro.

```
# Gráfico de barras - vendas médias por dia da semana
bd %>%
  group_by(Dia_da_semana) %>%
  summarise(venda_media = mean(`Total das vendas (R$)`, na.rm = TRUE)) %>%
  ggplot(aes(x = Dia_da_semana, y = venda_media)) +
  geom_col(fill = "orange") +
  labs(
    title = "vendas médias (R$) por dia da semana",
    x = "Dia da semana",
    y = "Média de vendas (R$)"
  ) +
  theme_minimal()
```

Figura 9 - Script para gráfico de barras de vendas médias diárias (R\$) por dia da semana

A Figura 10 apresenta um gráfico de barras com as médias de vendas diárias por dia da semana. Observa-se que os valores médios são bastante similares entre os dias, indicando que o total vendido diariamente não varia significativamente ao longo da semana. Essa constância pode sugerir um padrão de consumo estável e previsível, independentemente do dia.

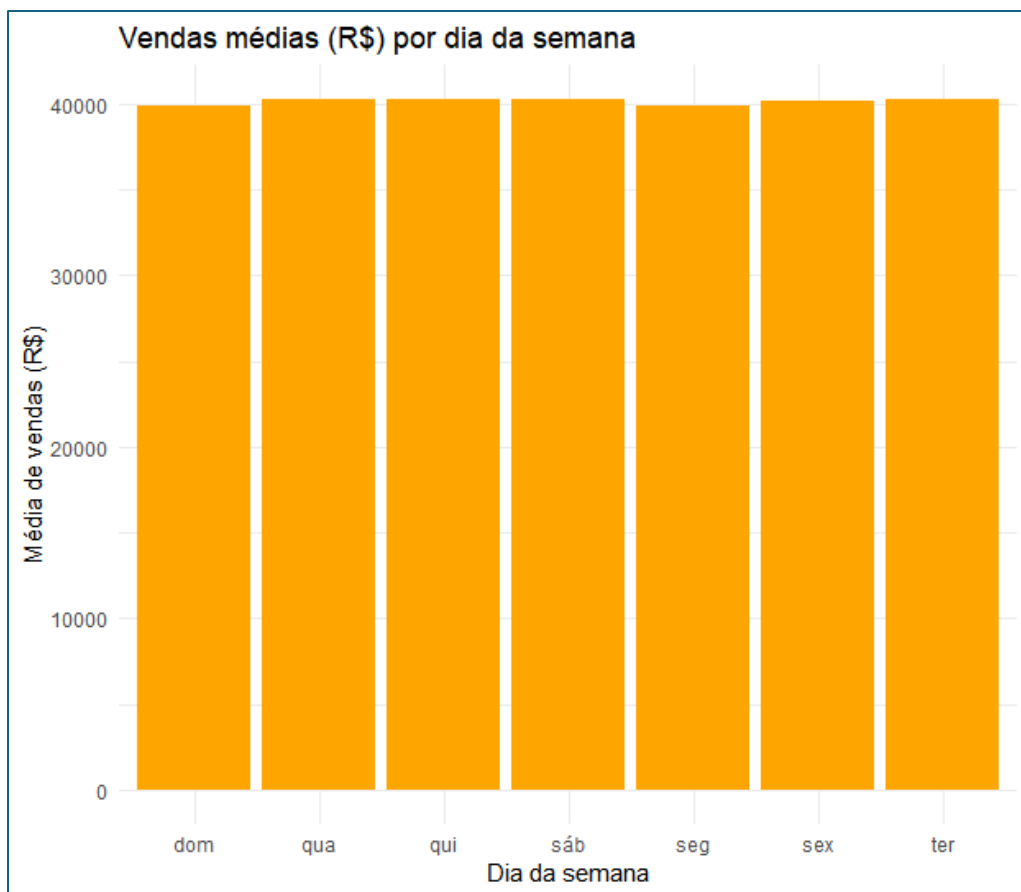


Figura 10 - Médias de vendas diárias (R\$) nos dias da semana

3) Façam as análises bi-variadas

3.1) AnáliseS graficas bi variadas: Grafico de dispersão (Vendas(y) x Temperatura (x))

O código apresentado realiza a análise bivariada entre as variáveis Temperatura (graus celsius) e Total das vendas (R\$). Inicialmente, são selecionadas as colunas relevantes e removidas observações com valores ausentes por meio da função `drop_na()`. Em seguida, calcula-se a correlação de Pearson, que quantifica a força e direção da associação linear entre temperatura e vendas. Por fim, é gerado um gráfico de dispersão com ajuste linear simples sobreposto, utilizando `geom_smooth(method = "lm")`, o que permite visualizar a tendência de crescimento das vendas em função do aumento da temperatura, evidenciando o comportamento esperado para o setor de sorvetes (Figura 11).

```
##### 3) Façam as análises bi-variadas
# =====
##### 3.1) análise graficas bi variadas: Grafico de dispersão (Vendas(y) x Temperatura (x))

## Análise biavariada "Total de vendas" x Temperatura
# Filtrar NAs do que será usado no gráfico
bd_scatter <- bd %>%
  select(`Total das vendas (R$)`, `Temperatura (graus celsius)`, Mês_venda, Dia_da_semana) %>%
  tidyr::drop_na()

# Correlação (para referência no relatório)
cor_pearson <- cor(bd_scatter$`Total das vendas (R$)`,
  bd_scatter$`Temperatura (graus celsius)`,
  use = "complete.obs", method = "pearson")

cor_pearson

# Dispersão com suavização com Ajuste linear simples sobreposto
ggplot(bd_scatter,
  aes(x = `Temperatura (graus celsius)`, y = `Total das vendas (R$)`) +
  geom_point(alpha = 0.35) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(title = "Vendas (R$) vs. Temperatura (°C) - ajuste linear",
    x = "Temperatura (°C)", y = "Total das vendas (R$)") +
  theme_minimal())
```

Figura 11 - Script dispersão entre temperatura (°C) e total das vendas (R\$), com linha de regressão linear ajustada

O gráfico de dispersão evidencia uma relação claramente positiva entre a temperatura (°C) e o total das vendas (R\$). Observa-se que, à medida que a temperatura aumenta, os valores de vendas também tendem a subir de forma consistente, com pouca dispersão em torno da reta ajustada. A linha azul representa a regressão linear simples, que descreve bem a tendência dos dados,

indicando que a temperatura é uma forte preditora do comportamento de vendas no setor de sorvetes (Figura 12).

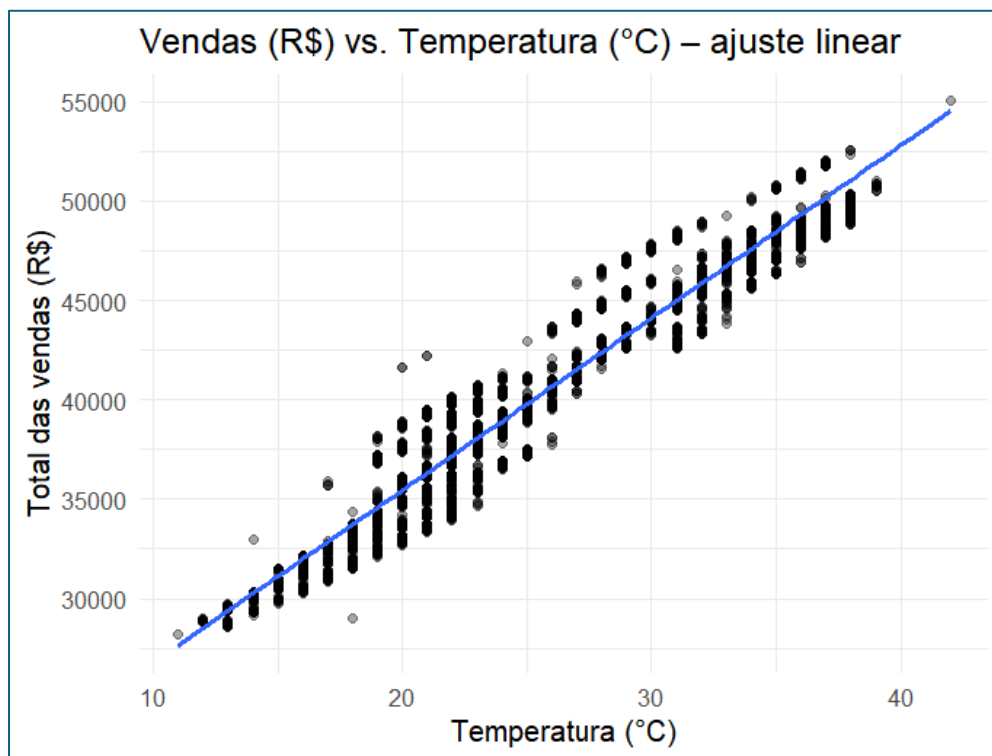


Figura 12 - Relação entre temperatura (°C) e total das vendas (R\$), com linha de tendência linear ajustada

O script apresentado a seguir realiza uma análise bivariada entre o dia da semana e o total de vendas (R\$), por meio de um boxplot. Esse tipo de gráfico permite visualizar a mediana, a dispersão e possíveis outliers das vendas em cada dia da semana. O objetivo é identificar variações sazonais ao longo da semana, como padrões recorrentes de maiores vendas em determinados dias — o que pode refletir o comportamento dos consumidores e apoiar estratégias operacionais e promocionais da indústria (Figura 13).


```
## Análise bivariada "Dia da semana" x "Total de vendas"  
# Boxplot: Vendas por dia da semana  
ggplot(bd_scatter, aes(x = Dia_da_semana, y = `Total das vendas (R$)`)) +  
  geom_boxplot(outlier.alpha = 0.4) +  
  labs(title = "Vendas (R$) por dia da semana",  
        x = "Dia da semana", y = "Total das vendas (R$)") +  
  theme_minimal()
```

Figura 13 - Script distribuição do total de vendas (R\$) por dia da semana

O gráfico representado a seguir apresenta a distribuição do total das vendas (R\$) nos dias da semana. Observa-se que as medianas são bastante próximas entre os dias, com uma leve elevação em quarta e quinta-feira. A dispersão dos valores é relativamente homogênea, indicando que o volume de vendas mantém comportamento consistente independentemente do dia da semana. Apesar de pequenas variações, não há evidência visual de concentração de picos ou quedas acentuadas em dias específicos, sugerindo estabilidade na demanda diária por sorvetes (Figura 14).

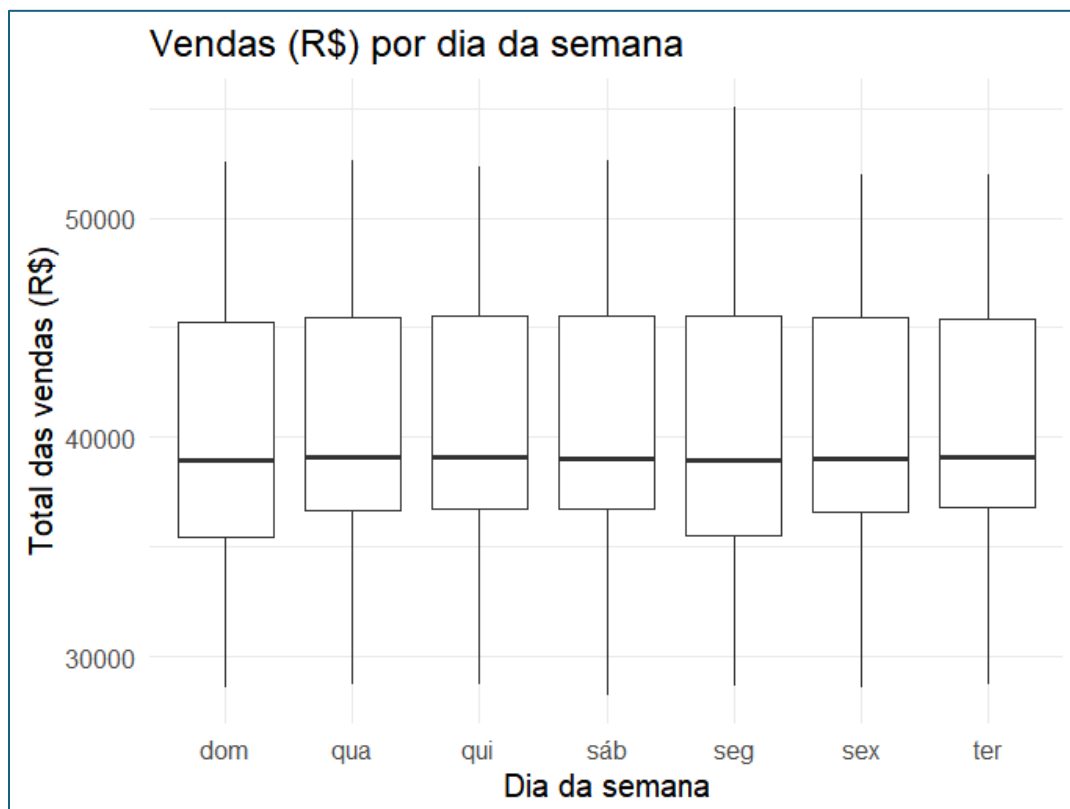


Figura 14 - Distribuição do total de vendas (R\$) por dia da semana

A seguir, o código realiza a análise bivariada entre mês de venda e total das vendas (R\$). Primeiramente, são selecionadas as variáveis relevantes e removidos os registros com dados ausentes. Em seguida, são geradas duas visualizações complementares: a primeira figura apresenta um boxplot que mostra a distribuição das vendas por mês, permitindo observar variações sazonais e dispersão dos valores em cada período; já a segunda figura mostra uma linha com os valores médios mensais de vendas, permitindo identificar tendências ao longo dos meses e comparações diretas entre os períodos. Essas representações visuais são úteis para detectar sazonalidades típicas do setor, como aumento nas vendas em meses mais quentes (Figura 15).

```
## Análise bivariada "Mês venda" x "Total de vendas"
# Filtrar variáveis necessárias e remover NAs
base_mes <- bd %>%
  select(Mês_venda, `Total das vendas (R$)` ) %>%
  drop_na()

# --- Gráfico: Boxplot das vendas por mês
ggplot(base_mes, aes(x = Mês_venda, y = `Total das vendas (R$)`)) +
  geom_boxplot(fill = "skyblue", outlier.alpha = 0.3) +
  labs(title = "Distribuição das vendas (R$) por mês",
       x = "Mês", y = "Total das vendas (R$)") +
  theme_minimal()

# --- Gráfico: Linha da média de vendas ao longo dos meses
ggplot(media_mes, aes(x = Mês_venda, y = venda_media, group = 1)) +
  geom_line(color = "darkblue") +
  geom_point(color = "blue", size = 2) +
  labs(title = "Tendência da média mensal de vendas",
       x = "Mês", y = "Vendas médias (R$)") +
  theme_minimal()
```

Figura 15 - Script das análises bivariadas “Mês venda” x “Total de Vendas”

A figura a seguir apresenta a distribuição do total das vendas (R\$) ao longo dos meses do ano, utilizando boxplots para cada mês. As medianas mensais (representadas pelas linhas horizontais dentro de cada caixa) mostram uma relativa estabilidade entre os meses, com valores ligeiramente mais elevados nos meses de outubro, novembro e dezembro. Já os meses de fevereiro a junho apresentam medianas discretamente menores, sugerindo uma leve redução nas vendas nesse período. A dispersão das vendas é relativamente homogênea entre os meses, com presença ocasional de valores extremos. Em conjunto, o

gráfico sugere variações sazonais suaves, sem indícios de picos de vendas concentrados em apenas um período do ano (Figura 16).

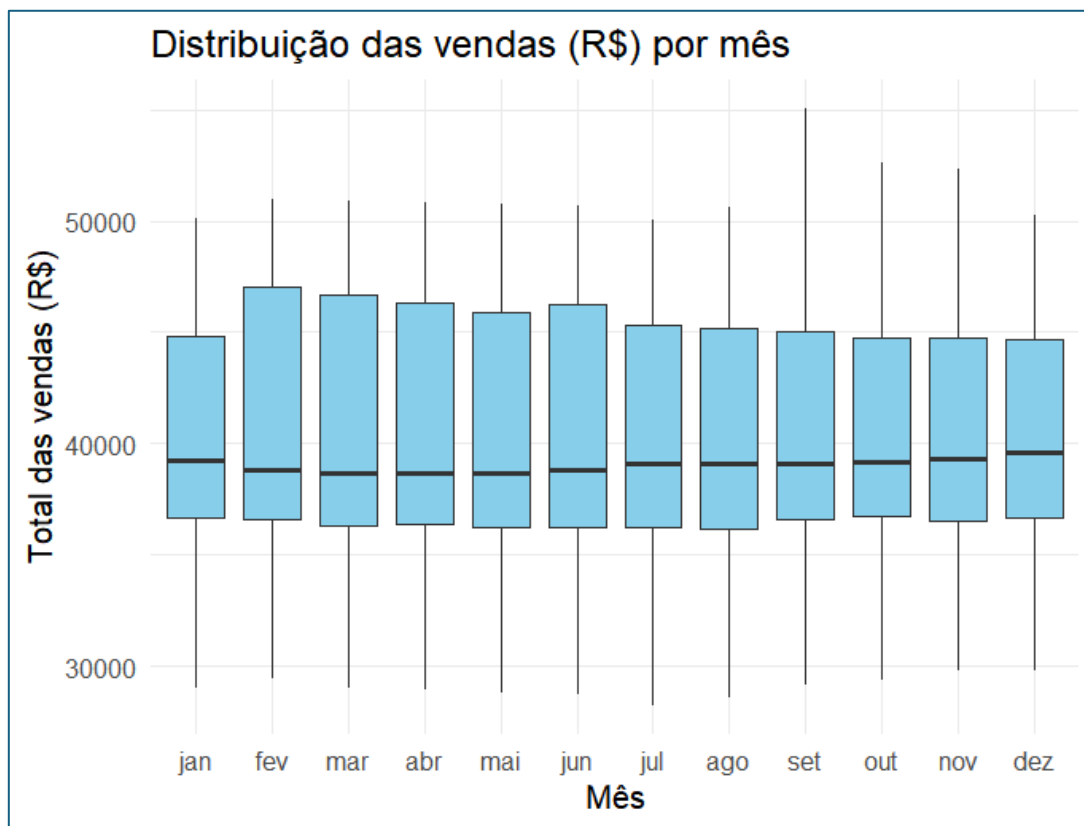


Figura 16 - Distribuição do total de vendas (R\$) por mês do ano

A figura a seguir apresenta a tendência da média mensal de vendas de sorvetes ao longo do ano. Observa-se uma leve queda nas médias de março a agosto, com o ponto mais baixo em agosto, indicando uma possível influência de fatores sazonais nesse período. A partir de setembro, nota-se uma recuperação expressiva nas médias, culminando em valores máximos nos meses de novembro e dezembro. Apesar de a variação entre os meses ser sutil (cerca de R\$ 600 entre o menor e o maior valor), esse comportamento reforça a existência de uma sazonalidade moderada, com desempenho mais forte no último trimestre do ano, possivelmente associado a temperaturas mais elevadas ou aumento da demanda em datas comemorativas (Figura 17).

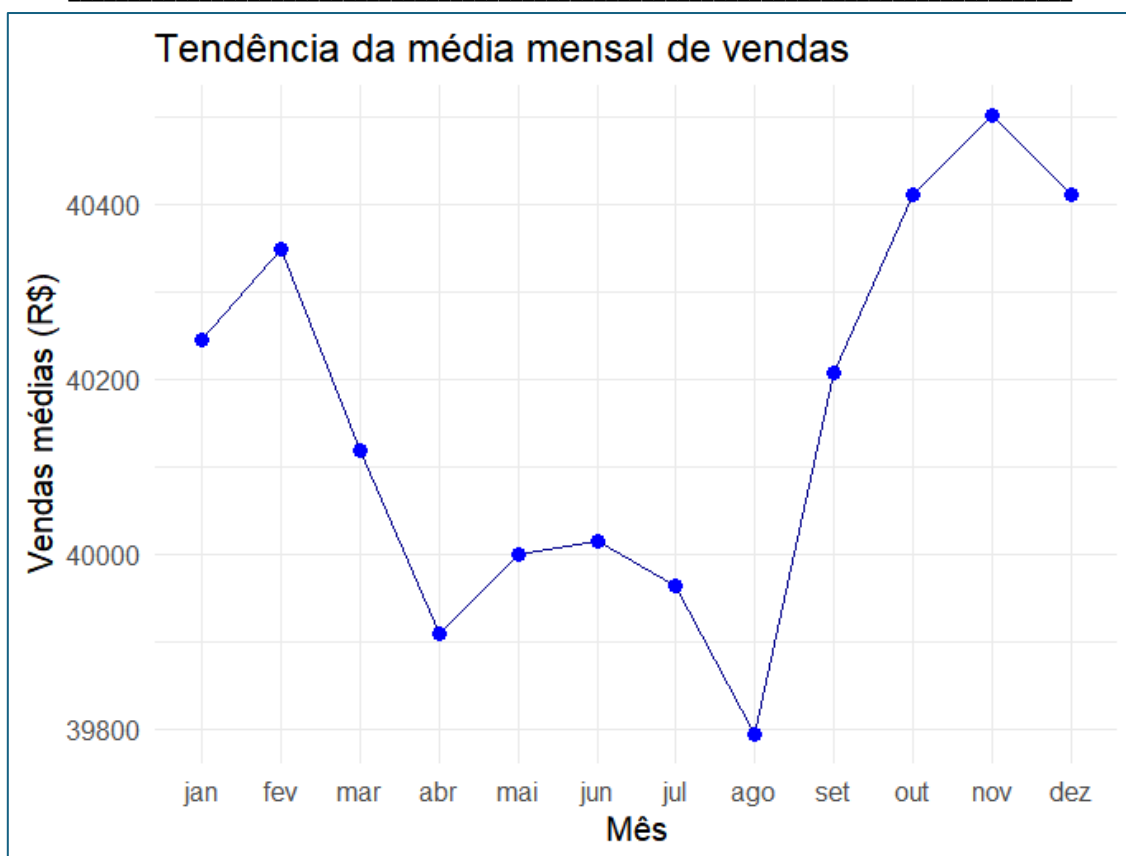


Figura 17 - Tendência da média mensal de vendas (R\$) ao longo do ano

A diferença visual entre o boxplot e o gráfico de linha decorre da natureza estatística de cada representação. Enquanto o boxplot expressa a distribuição completa das vendas por mês (incluindo mediana, dispersão e valores extremos), o gráfico de linha exibe apenas a média mensal, suavizando variações internas. Assim, um mês com vendas irregulares e outliers pode ter boxplots elevados, mas média similar a meses com vendas mais estáveis. Por isso, ambos os gráficos devem ser interpretados de forma complementar.

O trecho de código apresentado abaixo tem como objetivo analisar a evolução temporal das vendas de sorvetes, tanto em nível diário quanto mensal. No primeiro gráfico, a linha representa o comportamento do total de vendas ao longo do tempo, com granularidade diária, permitindo identificar oscilações sazonais, ciclos curtos e eventuais rupturas. Já o segundo gráfico sintetiza a informação por meio da média mensal de vendas ao longo dos anos, agrupando os dados por mês e ano, o que favorece a detecção de tendências de longo prazo, como crescimento, estabilidade ou declínio das vendas. A análise

conjunta desses dois gráficos fornece uma base sólida para avaliar o desempenho histórico da empresa e orientar estimativas futuras (Figura 18).

```
## Evolução das vendas ao longo do tempo
# --- Gráfico: Linha com todas as vendas (dia a dia)
ggplot(bd, aes(x = Data, y = `Total das vendas (R$)`)) +
  geom_line(color = "steelblue", alpha = 0.6) +
  labs(title = "Evolução diária das vendas ao longo do tempo",
       x = "Data", y = "Total das vendas (R$)") +
  theme_minimal()

# --- Gráfico 2: Média mensal de vendas ao longo dos anos
bd_mensal <- bd %>%
  mutate(Ano_mes = floor_date(Data, unit = "month")) %>%
  group_by(Ano_mes) %>%
  summarise(venda_media = mean(`Total das vendas (R$)`, na.rm = TRUE), .groups = "drop")

ggplot(bd_mensal, aes(x = Ano_mes, y = venda_media)) +
  geom_line(color = "darkgreen") +
  geom_point(size = 1.5) +
  labs(title = "Média mensal de vendas ao longo dos anos",
       x = "Ano-Mês", y = "Vendas médias (R$)") +
  theme_minimal()
```

Figura 18 - Script dos gráficos de evolução diária e média mensal do valor das vendas ao longo do tempo (de março de 2004 a outubro de 2023)

A Figura 19 apresenta a evolução diária das vendas ao longo do tempo, é apresentado um gráfico de linha com dados diários, cobrindo o período de 2004 a 2023. Cada ponto da linha representa o total de vendas em um único dia. Essa abordagem mostra com detalhes as flutuações de curto prazo e os períodos de ausência de dados (representados por lacunas na linha), além de permitir identificar sazonalidades, tendências de queda ou crescimento ao longo dos anos. Observa-se uma variação significativa nos níveis de vendas ao longo dos anos. Entre 2004 e 2006, os valores oscilaram entre R\$30.000 e R\$38.000. A partir de 2007, houve um salto expressivo nas vendas diárias, atingindo patamares entre R\$45.000 e R\$50.000, tendência que se manteve até aproximadamente 2011. A partir de então, identifica-se um movimento de queda gradual nas vendas, atingindo valores próximos a R\$30.000 em torno de 2018. Em seguida, observa-se nova elevação entre 2019 e 2020, seguida por nova desaceleração nos últimos anos da série.

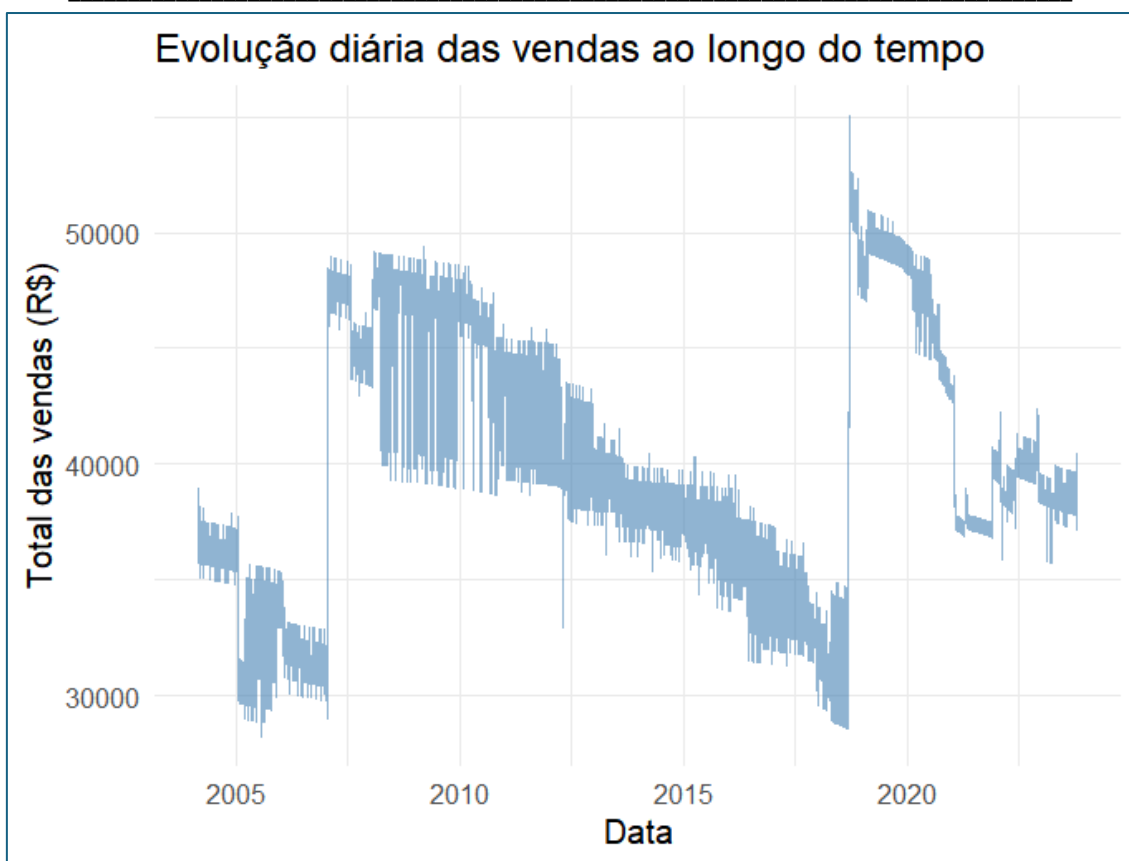
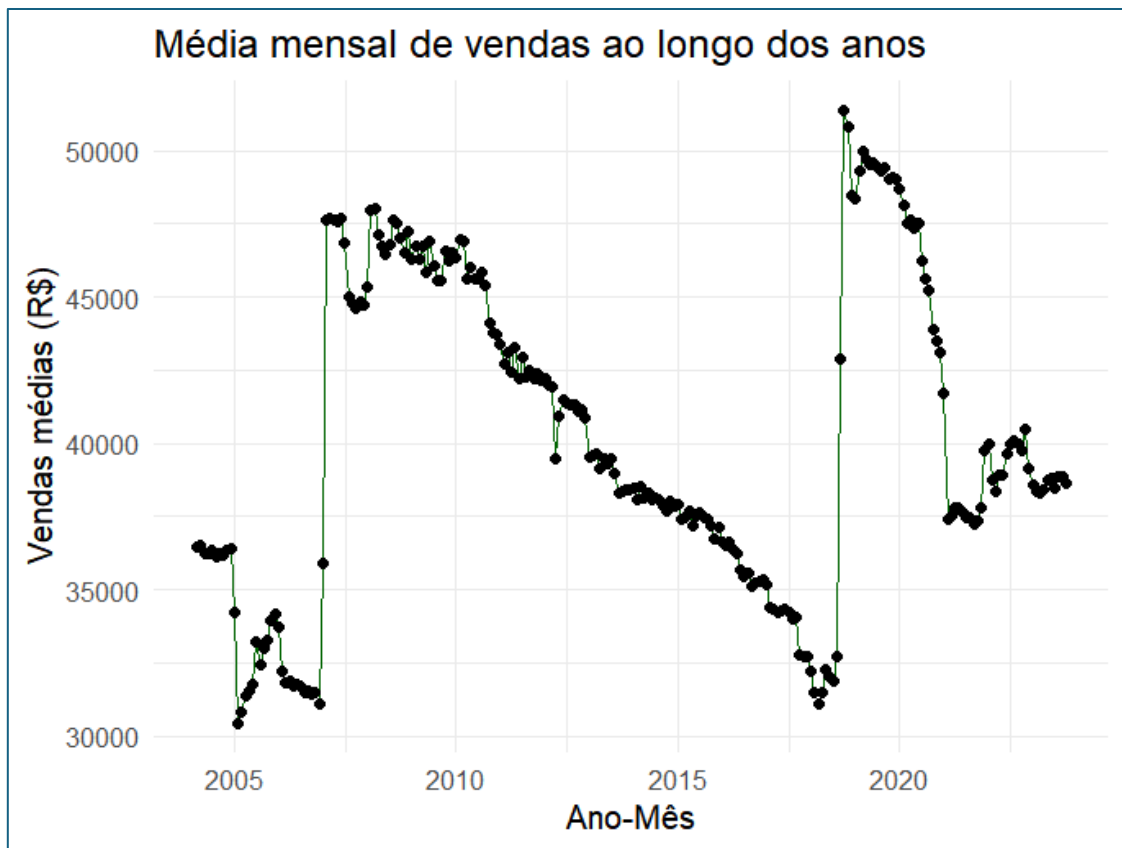


Figura 19 - Evolução das vendas por dia

A Figura 20 apresenta uma linha de tendência baseada na média mensal de vendas. Aqui, os dados foram agregados por mês e ano (ano-mês), de modo que cada ponto representa a média de vendas de todos os dias de determinado mês. Isso suaviza as variações diárias e destaca as tendências mais amplas no tempo. A linha confirma a tendência identificada na Figura 19: após um período inicial de menor faturamento médio mensal, há um crescimento expressivo a partir de 2007, com picos superiores a R\$48.000 entre 2008 e 2011. Após esse período de alta, observa-se uma tendência de declínio contínuo até 2018, quando as médias mensais caem para valores em torno de R\$30.000. A partir de 2019, há nova recuperação, alcançando novamente picos acima de R\$50.000, seguidos por queda a partir de 2021.



3.2) calcule a correlação entre as variáveis Temperatura e total de vendas (R\$)

O trecho de código apresentado a seguir realiza o cálculo da correlação de Pearson entre as variáveis Temperatura (graus celsius) e Total das vendas (R\$), utilizando apenas os casos completos (sem valores ausentes). A função `cor.test()` não apenas calcula o coeficiente de correlação linear, como também fornece um teste de significância estatística associado à hipótese de ausência de correlação ($H_0: \rho = 0$). O resultado gerado pelo comando inclui o valor do coeficiente de correlação `rrr`, o valor-p (`p-value`), o intervalo de confiança e o número de observações válidas. Esse tipo de análise é útil para quantificar a intensidade e a direção da associação linear entre as duas variáveis, contribuindo para fundamentar modelos de previsão ou interpretações estatísticas em relatórios (Figura 21).

```
##### 3.2) calcule a correlação entre as variáveis Temperatura e total de vendas (R$)
# Calcular correlação de Pearson
cor.test(bd$`Temperatura (graus celsius)`|
        bd$`Total das vendas (R$)` ,
        method = "pearson",
        use = "complete.obs")
```

Figura 21 - Script da análise de correlação linear entre temperatura e total de vendas de sorvete (R\$)

A Figura 22 apresenta os resultados do teste de correlação de Pearson aplicado entre a variável temperatura (°C) e o total diário de vendas de sorvetes (R\$). A correlação estimada foi $r = 0,966$, indicando uma associação linear positiva forte e estatisticamente significativa entre as duas variáveis (valor de $p < 2,2e-16$). O intervalo de confiança de 95% para o coeficiente de correlação vai de 0,9645 a 0,9676, reforçando a robustez do achado. Isso sugere que aumentos na temperatura tendem a estar fortemente associados a aumentos nas vendas de sorvete. Essa relação fundamenta o uso da temperatura como preditora no modelo de regressão linear desenvolvido no item anterior.


```
Pearson's product-moment correlation  
data: bd$`Temperatura (graus celsius)` and bd$`Total das vendas (R$)`  
t = 316.96, df = 7173, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.9645280 0.9676135  
sample estimates:  
      cor  
0.9661052
```

Figura 22 - Resultado do teste de correlação de Pearson entre temperatura e total de vendas (R\$)

4) Definam a equação através de uma Regressão Linear

O script apresentado na Figura 23 implementa um modelo de regressão linear simples para investigar a relação entre a temperatura (°C) e o total de vendas de sorvete (R\$). Inicialmente, o modelo é ajustado por meio da função `lm()`, que estima os coeficientes da equação da reta. Em seguida, utiliza-se `summary()` para inspecionar estatísticas do ajuste, como os valores dos coeficientes e o coeficiente de determinação (R^2). A equação da reta é então extraída e formatada com `paste0()` para ser incluída diretamente no gráfico. Por fim, o script gera um gráfico de dispersão com a linha de regressão sobreposta (`geom_smooth`) e a equação exibida no canto inferior direito (`annotate`), facilitando a interpretação visual da relação entre as variáveis. Esse procedimento automatizado permite não apenas visualizar a tendência linear, mas também explicitar, de forma transparente, os parâmetros obtidos no modelo ajustado.

```
# =====  
#### 4) definam a equacao através de uma Regressão Linear  
# =====  
# Ajustar modelo de regressão linear  
modelo_r1 <- lm(`Total das vendas (R$)` ~ `Temperatura (graus celsius)`, data = bd)  
  
# Ver o resumo do modelo  
summary(modelo_r1)  
  
# Ajustar o modelo  
modelo_r1 <- lm(`Total das vendas (R$)` ~ `Temperatura (graus celsius)`, data = bd)  
  
# Extrair coeficientes  
intercepto <- round(coef(modelo_r1)[1], 2)  
inclinação <- round(coef(modelo_r1)[2], 2)  
r2 <- round(summary(modelo_r1)$r.squared, 3)  
  
# Montar equação como texto  
equacao <- paste0("Vendas = ", intercepto, " + ", inclinação, " x Temperatura\nR² = ", r2)  
  
# Gráfico com equação no canto  
ggplot(bd, aes(x = `Temperatura (graus celsius)`, y = `Total das vendas (R$)`)) +  
  geom_point(alpha = 0.3) +  
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  
  annotate("text", x = Inf, y = -Inf, label = equacao,  
    hjust = 1.1, vjust = -0.5, size = 4, color = "black") +  
  labs(title = "Regressão linear: Vendas vs Temperatura",  
    x = "Temperatura (°C)",  
    y = "Total das vendas (R$)") +  
  theme_minimal()
```

Figura 23 - Script do modelo e gráfico de regressão linear entre temperatura (°C) e vendas de sorvete (R\$), incluindo equação da reta

A figura a seguir apresenta o gráfico de dispersão entre a temperatura (°C) e o total de vendas diárias de sorvete (R\$), com a linha de regressão linear ajustada sobreposta. Observa-se uma forte associação positiva entre as duas

variáveis: à medida que a temperatura aumenta, o volume de vendas também tende a crescer. A equação estimada foi $\text{Vendas} = 18129,4 + 866,19 \times \text{Temperatura}$, indicando que, para cada grau Celsius adicional, estima-se um acréscimo médio de R\$ 866,19 no total diário de vendas. O coeficiente de determinação ($R^2 = 0,933$) reforça que o modelo linear ajustado explica aproximadamente 93,3% da variação observada nas vendas, o que evidencia a elevada capacidade preditiva da variável temperatura nesse contexto (Figura 24).

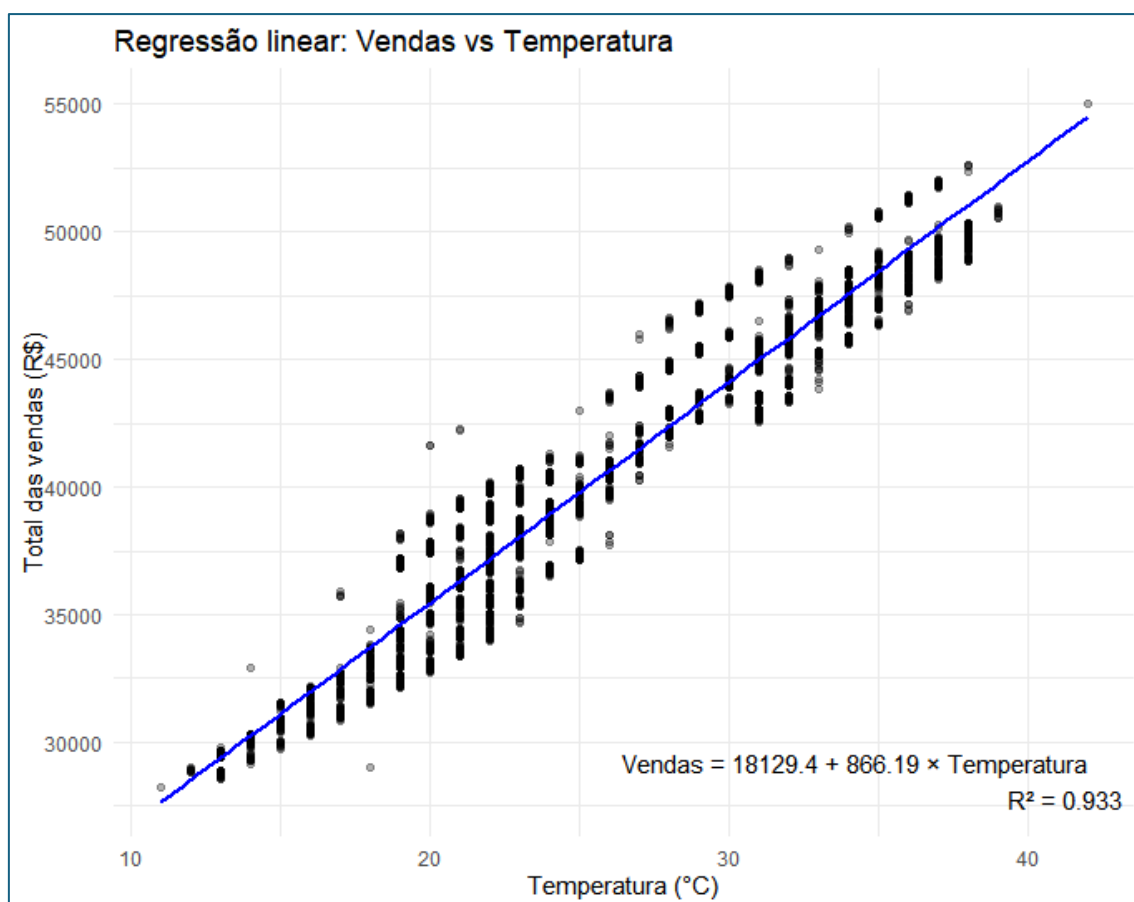


Figura 24 - Regressão linear entre temperatura (°C) e total de vendas (R\$), incluindo equação da reta

5) Prevejam qual será o total a ser faturado com as vendas (em R\$) na data de 30/out/23, sabendo que a previsão do tempo aponta para 18o C

O trecho de código apresentado na Figura 25 realiza a previsão do faturamento diário com base na equação da regressão linear simples entre temperatura e total de vendas. Como a previsão do tempo para o dia 30 de outubro de 2023 é de 18 °C, substituiu-se esse valor na equação estimada anteriormente:

$$\text{Vendas} = 18129,4 + 866,19 \times \text{Temperatura}$$

Substituindo a temperatura por 18, temos:

$$\text{Vendas} = 18129,4 + 866,19 \times 18$$

O resultado armazenado em `previsao` corresponde ao valor esperado de vendas em reais (R\$) para essa data, de acordo com o modelo linear ajustado. Essa estimativa pode orientar o planejamento de estoques e logística em função das condições climáticas previstas.

```
# =====  
#### 5) Prevejam qual será o total a ser faturado com as vendas (em R$) na data de 30/out/23, sabendo que a previsão do tempo aponta para 18o c  
# =====  
# cálculo direto a partir da equação  
previsao <- 18129.4 + 866.19 * 18  
previsao
```

Figura 25 - Script de previsão de vendas com base na temperatura prevista

A Figura 26 apresenta o valor estimado de R\$ 33.720,82 para o total de vendas no dia 30 de outubro de 2023, considerando a previsão de temperatura de 18 °C e com base na equação ajustada por regressão linear simples entre temperatura e faturamento. Esse resultado reforça a forte dependência positiva identificada entre as variáveis.

```
> # Cálculo direto a partir da equação  
> previsao <- 18129.4 + 866.19 * 18  
> previsao  
[1] 33720.82
```

Figura 26 - Resultado da previsão de vendas para 30/out/2023

6) Escreva a sua conclusão

As análises estatísticas e gráficas realizadas evidenciaram uma relação forte e positiva entre a temperatura ambiente e o volume de vendas de sorvetes, com destaque para a correlação de Pearson de 0,97, indicando associação praticamente perfeita. A regressão linear simples confirmou essa relação, permitindo estimar com precisão o total de vendas com base na previsão de temperatura, o que se mostrou útil para projeções de faturamento, como no caso do dia 30 de outubro de 2023.

As análises bivariadas também mostraram que, embora haja certa variabilidade nas vendas ao longo dos dias da semana e dos meses, essas oscilações são menos acentuadas quando comparadas ao impacto da temperatura. A evolução temporal revelou padrões de sazonalidade e possíveis variações estruturais ao longo dos anos, apontando para a importância de considerar também o componente temporal nas projeções futuras.

De forma geral, os resultados confirmam que a temperatura é um dos principais fatores explicativos do comportamento de vendas de sorvetes nesta indústria, representando uma variável-chave para a elaboração de estratégias de planejamento e tomada de decisão.

7) Escreva a sua recomendação para o Diretor Financeiro desta empresa

Com base nas análises realizadas, recomenda-se que a Diretoria Financeira utilize a temperatura como principal variável preditiva para o planejamento de vendas de sorvetes. A forte correlação observada entre temperatura e faturamento ($r = 0,97$) permite prever com elevada confiança os volumes de receita a partir das previsões meteorológicas. A equação da regressão linear identificada ($\text{Vendas} = 18.129,4 + 866,19 \times \text{Temperatura}$) pode ser aplicada de forma direta para estimar o faturamento diário esperado com base na temperatura máxima prevista.

Dessa forma, recomenda-se incorporar esse modelo preditivo nas rotinas de planejamento financeiro, otimizando a alocação de recursos, controle de estoques e definição de metas de vendas. Além disso, períodos com previsão de temperaturas mais baixas, que tendem a apresentar menor faturamento, podem demandar ações específicas, como promoções, diversificação de produtos ou reforço em canais alternativos de venda, a fim de mitigar perdas. Por fim, sugere-se o acompanhamento contínuo das séries temporais para ajustes futuros no modelo, considerando eventuais mudanças estruturais nos padrões de consumo ao longo dos anos.