

Sumário

| | Página |
|--|---------------|
| 1 Introdução | 2 |
| 2 Referencial Teórico | 3 |
| 2.1 Frequência Relativa | 3 |
| 2.2 Média | 3 |
| 2.3 Mediana | 3 |
| 2.4 Quartis | 4 |
| 2.5 Variância | 4 |
| 2.5.1 Variância Populacional | 4 |
| 2.5.2 Variância Amostral | 5 |
| 2.6 Desvio Padrão | 5 |
| 2.6.1 Desvio Padrão Populacional | 5 |
| 2.6.2 Desvio Padrão Amostral | 6 |
| 2.7 Coeficiente de Variação | 6 |
| 2.8 Boxplot | 6 |
| 2.9 Histograma | 7 |
| 2.10 Gráfico de Dispersão | 8 |
| 2.11 Tipos de Variáveis | 9 |
| 2.11.1 Qualitativas | 9 |
| 2.11.2 Quantitativas | 9 |
| 2.12 Coeficiente de Correlação de Pearson | 10 |
| 2.13 Teste de Hipóteses | 10 |
| 2.14 Tipos de teste: bilateral e unilateral | 11 |
| 2.15 Nível de significância (α) | 11 |
| 2.16 Estatística do Teste | 12 |
| 2.17 P-valor | 12 |
| 2.18 Intervalo de Confiança | 12 |
| 2.19 Teste de Correlação de Pearson | 13 |
| 3 Análises | 14 |
| 3.1 A receita média das lojas registrada nos anos de 1880 até 1889 | 14 |
| 3.2 Variação Peso por Altura | 15 |
| 3.3 Idade dos clientes de Âmbar Seco a depender da loja | 18 |
| 3.4 Os top 3 produtos mais vendidos nas top 3 lojas com maior receita em 1889 | 19 |
| 4 Conclusão | 23 |
| 5 Anexo | 24 |

1 Introdução

O objetivo deste relatório é apresentar uma análise dos dados, visando compreender o comportamento das vendas e dos clientes da região. O escopo das análises foi a **receita média** das lojas, a variação do **peso** por **altura**, a idade dos clientes de **Âmbar Seco** e os três **produtos mais vendidos** nas lojas com maior receita. As técnicas empregadas incluíram análises exploratórias, medidas descritivas, gráficos de dispersão e boxplots, testes de normalidade, correlação de Pearson e intervalos de confiança.

Os dados foram fornecidos pela ESTAT – Consultoria Estatística. O conjunto de dados é composto por 19.885 observações de compras realizadas por 1.990 clientes, envolvendo 10 tipos de produtos comercializados em 18 lojas distribuídas entre 5 cidades da região de interesse. Para mais informações sobre as variáveis, consulte o **Anexo 1**. A análise dos dados foi realizada na versão 4.5.0 do R, mais informações sobre os pacotes e ambiente utilizados estão no **Anexo 2**.

2 Referencial Teórico

2.1 Frequência Relativa

A frequência relativa é utilizada para a comparação entre classes de uma variável categórica com c categorias, ou para comparar uma mesma categoria em diferentes estudos.

A frequência relativa da categoria j é dada por:

$$f_j = \frac{n_j}{n}$$

Com:

- $j = 1, \dots, c$
- n_j = número de observações da categoria j
- n = número total de observações

Geralmente, a frequência relativa é utilizada em porcentagem, dada por:

$$100 \times f_j$$

2.2 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Com:

- $i = 1, 2, \dots, n$
- n = número total de observações

2.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é o valor que deixa metade das observações abaixo dela e metade dos dados acima.

Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, \text{ para } n \text{ ímpar} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{ para } n \text{ par} \end{cases}$$

2.4 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) delimita os 25% menores valores, o segundo representa a mediana, e o terceiro delimita os 25% maiores valores. Inicialmente deve-se calcular a posição do quartil:

- Posição do primeiro quartil P_1 :

$$P_1 = \frac{n+1}{4}$$

- Posição da mediana (segundo quartil) P_2 :

$$P_2 = \frac{n+1}{2}$$

- Posição do terceiro quartil P_3 :

$$P_3 = \frac{3 \times (n+1)}{4}$$

Com n sendo o tamanho da amostra. Dessa forma, $X_{(P_i)}$ é o valor do i -ésimo quartil, onde $X_{(j)}$ representa a j -ésima observação dos dados ordenados.

Se o cálculo da posição resultar em uma fração, deve-se fazer a média entre o valor que está na posição do inteiro anterior e do seguinte ao da posição.

2.5 Variância

A variância é uma medida que avalia o quanto os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

2.5.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Com:

- $X_i = i\text{-ésima observação da população}$
- $\mu = \text{média populacional}$
- $N = \text{tamanho da população}$

2.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Com:

- $X_i = i\text{-ésima observação da amostra}$
- $\bar{X} = \text{média amostral}$
- $n = \text{tamanho da amostra}$

2.6 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Ele avalia o quanto os dados estão dispersos em relação à média.

2.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Com:

- $X_i = i\text{-ésima observação da população}$
- $\mu = \text{média populacional}$
- $N = \text{tamanho da população}$

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Com:

- X_i = i-ésima observação da amostra
- \bar{X} = média amostral
- n = tamanho da amostra

2.7 Coeficiente de Variação

O coeficiente de variação fornece a dispersão dos dados em relação à média. Quanto menor for o seu valor, mais homogêneos serão os dados. O coeficiente de variação é considerado baixo (apontando um conjunto de dados homogêneo) quando for menor ou igual a 25%. Ele é dado pela fórmula:

$$C_V = \frac{S}{\bar{X}} \times 100$$

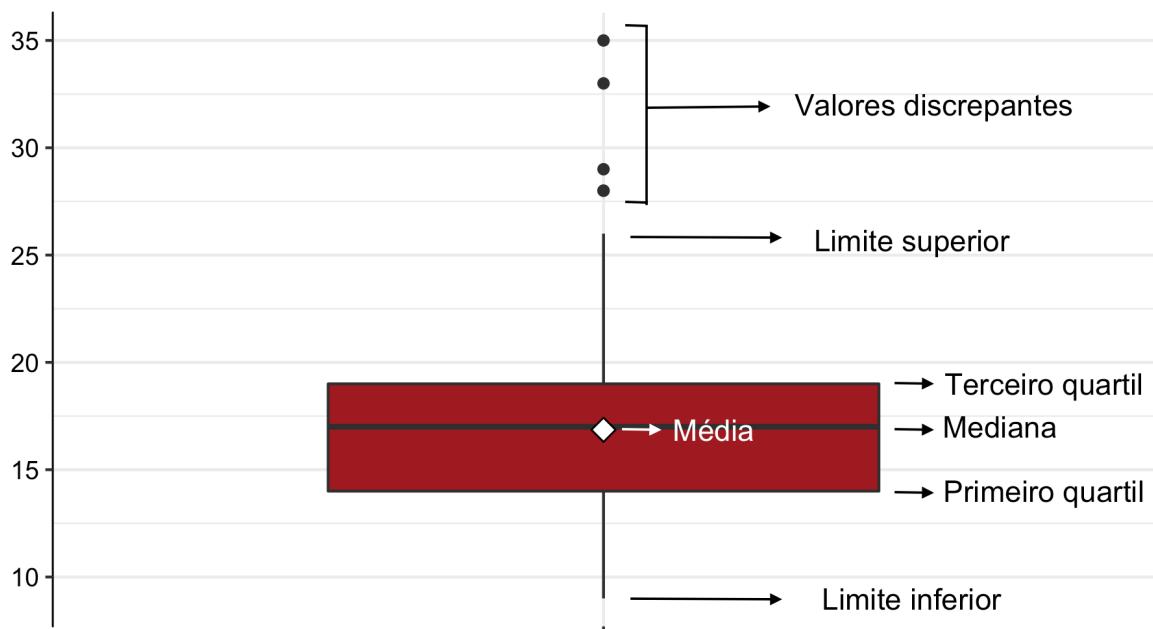
Com:

- S = desvio padrão amostral
- \bar{X} = média amostral

2.8 Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos. A figura abaixo ilustra um exemplo de boxplot.

Figura 1: Exemplo de boxplot

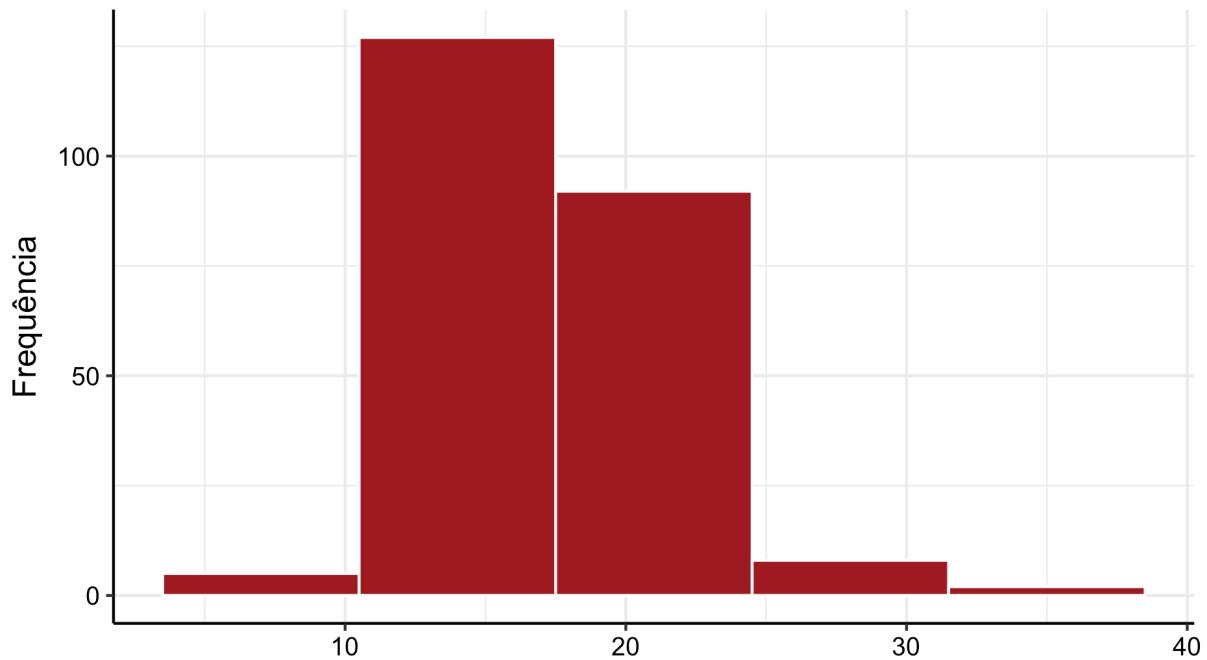


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

2.9 Histograma

O histograma é uma representação gráfica utilizada para a visualização da distribuição dos dados e pode ser construído por valores absolutos, frequência relativa ou densidade. A figura abaixo ilustra um exemplo de histograma.

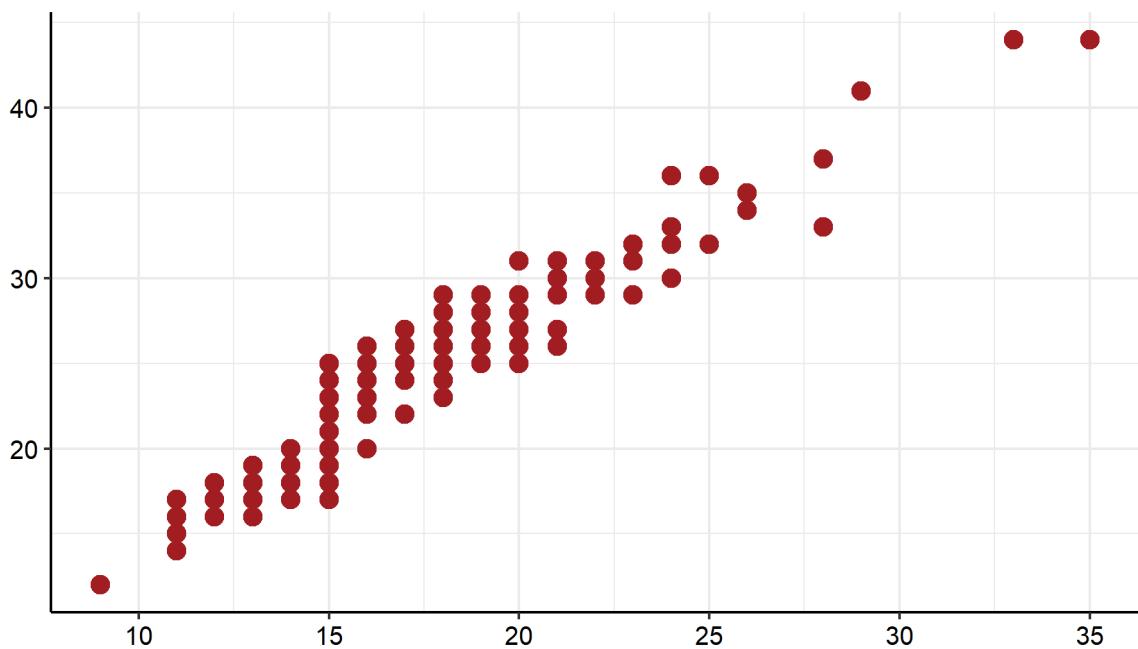
Figura 2: Exemplo de histograma



2.10 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

Figura 3: Exemplo de Gráfico de Dispersão



2.11 Tipos de Variáveis

2.11.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- **Nominais:** quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)
- **Ordinais:** quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

2.11.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- **Discretas:** quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- **Contínuas:** quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

2.12 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}}$$

Onde:

- x_i = i-ésimo valor da variável X
- y_i = i-ésimo valor da variável Y
- \bar{x} = média dos valores da variável X
- \bar{y} = média dos valores da variável Y

Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

2.13 Teste de Hipóteses

O teste de hipóteses tem como objetivo fornecer uma metodologia para verificar se os dados das amostras possuem indicativos que comprovem, ou não, uma hipótese previamente formulada. Ele é composto por duas hipóteses:

$$\begin{cases} H_0 : \text{hipótese a ser testada (chamada de hipótese nula)} \\ H_1 : \text{hipótese alternativa que será aceita caso a hipótese nula seja rejeitada} \end{cases}$$

Essa decisão é tomada por meio da construção de uma região crítica, ou seja, região de rejeição do teste.

2.14 Tipos de teste: bilateral e unilateral

Para a formulação de um teste, deve-se definir as hipóteses de interesse. Em geral, a hipótese nula é composta por uma igualdade (por exemplo, $H_0 : \theta = \theta_0$). Já a hipótese alternativa depende do grau de conhecimento que se tem do problema em estudo. Assim, tem-se três formas de elaborar H_1 que classificam os testes em duas categorias:

- **Teste Bilateral:**

Esse é o teste mais geral, em que a hipótese alternativa consiste em verificar se existe diferença entre os parâmetros de interesse, independentemente de um ser maior ou menor que o outro. Dessa forma, tem-se:

$$H_1 : \theta \neq \theta_0$$

- **Teste Unilateral:**

Dependendo das informações que o pesquisador possui a respeito do problema e os questionamentos que possui, a hipótese alternativa pode ser feita de forma a verificar se existe diferença entre os parâmetros em um dos sentidos. Ou seja:

$$H_1 : \theta < \theta_0$$

ou

$$H_1 : \theta > \theta_0$$

Tipos de Erros Ao realizar um teste de hipóteses, existem dois erros associados: **Erro do Tipo I** e **Erro do Tipo II**.

- **Erro do Tipo I:**

Esse erro é caracterizado por rejeitar a hipótese nula (H_0) quando essa é verdadeira. A probabilidade associada a esse erro é denotada por α , também conhecido como nível de significância do teste.

- **Erro do Tipo II:**

Ao não rejeitar H_0 quando, na verdade, é falsa, está sendo cometido o **Erro do Tipo II**. A probabilidade de se cometer este erro é denotada por β .

2.15 Nível de significância (α)

O nível de significância do teste é o nome dado à probabilidade de se rejeitar a hipótese nula quando essa é verdadeira; essa rejeição é chamada de **erro do tipo I**.

O valor de α é fixado antes da extração da amostra e, usualmente, assume 5%, 1% ou 0,1%.

Por exemplo, um nível de significância de $\alpha = 0,05$ (5%) significa que, se for tomada uma grande quantidade de amostras, em 5% delas a hipótese nula será rejeitada quando não havia evidências para essa rejeição, isto é, a probabilidade de se tomar a decisão correta é de 95%.

2.16 Estatística do Teste

A estatística do teste é o estimador que será utilizado para testar se a hipótese nula (H_0) é verdadeira ou não. Ela é escolhida por meio das teorias estatísticas.

2.17 P-valor

O **P-valor**, ou nível descritivo, é uma medida utilizada para sintetizar o resultado de um teste de hipóteses. Ele também pode ser chamado de *probabilidade de significância* do teste e indica a probabilidade de se obter um resultado da estatística de teste mais extremo do que o observado na presente amostra, considerando que a hipótese nula é verdadeira. Dessa forma, rejeita-se H_0 quando $P\text{-valor} < \alpha$, porque a chance de uma nova amostra possuir valores tão extremos quanto o encontrado é baixa, ou seja, há evidências para a rejeição da hipótese nula.

2.18 Intervalo de Confiança

Quando calcula-se um estimador pontual para o parâmetro, não é possível definir qual a possível magnitude do erro que se está cometendo. Com o objetivo de associar um erro à estimativa, são construídos os intervalos de confiança que se baseiam na distribuição amostral do estimador pontual.

Dessa forma, considere T um estimador pontual para θ e que a distribuição amostral de T é conhecida. O intervalo de confiança para o parâmetro θ será dado por t_1 e t_2 , tal que:

$$P(t_1 < \theta < t_2) = \gamma$$

A probabilidade γ é estabelecida no início do estudo e representa o nível de confiança do intervalo. A interpretação desse resultado é que, se forem tiradas várias amostras de mesmo tamanho e forem calculados intervalos de confiança para cada uma, $100 \times \gamma\%$ dos intervalos irão conter o parâmetro θ . Assim, ao calcular um intervalo, pode-se dizer que há $100 \times \gamma\%$ de confiança de que o intervalo contém o parâmetro de interesse.

2.19 Teste de Correlação de Pearson

O coeficiente de correlação linear de Pearson indica a força e a direção do relacionamento linear entre duas variáveis quantitativas. É um índice adimensional com valores situados entre -1 e 1, no qual o valor -1 representa total correlação linear negativa entre as variáveis (quando o valor de uma variável cresce, o valor da outra diminui) e o valor 1 representa total correlação linear positiva entre elas (ambas crescem simultaneamente). Esse coeficiente é obtido por meio da fórmula:

$$r_{Pearson} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}}$$

em que

- x_i = i-ésimo valor da variável X
- y_i = i-ésimo valor da variável Y
- \bar{x} = média dos valores da variável X
- \bar{y} = média dos valores da variável Y
- $r_{Pearson}$ = coeficiente de correlação linear de Pearson amostral

Para o teste de correlação de Pearson, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não há correlação linear entre as variáveis } X \text{ e } Y \\ \quad (\rho_{Pearson} = 0) \\ H_1 : \text{Há correlação linear entre as variáveis } X \text{ e } Y \\ \quad (\rho_{Pearson} \neq 0) \end{cases}$$

em que $\rho_{Pearson}$ é o parâmetro a ser testado: coeficiente de correlação linear populacional.

Se X e Y tem distribuição normal, tem-se que a estatística do teste é dada por:

$$t_{Pearson} = \frac{r_{Pearson} \sqrt{n-2}}{\sqrt{1 - r_{Pearson}^2}} \sim t_{n-2}$$

Assim, sob H_0 , $t_{Pearson}$ segue uma distribuição t -Student com $(n - 2)$ graus de liberdade.

3 Análises

3.1 A receita média das lojas registrada nos anos de 1880 até 1889

O objetivo desta análise é compreender o comportamento da região ao longo da série temporal por meio da **receita média**. Essa variável foi calculada pela soma da receita anual da região dividida pelo número de lojas (18). Receita média é uma variável **quantitativa contínua**, por esse motivo, para facilitar a interpretação dos resultados, foram utilizados gráficos de dispersão e tabelas. A Figura 4 apresenta a evolução da receita média ao longo dos anos, com o tempo no eixo x e a receita média no eixo y, enquanto a Tabela 1 exibe os valores correspondentes (em reais) no período de 1880 a 1889.

Figura 4: Receita média da região

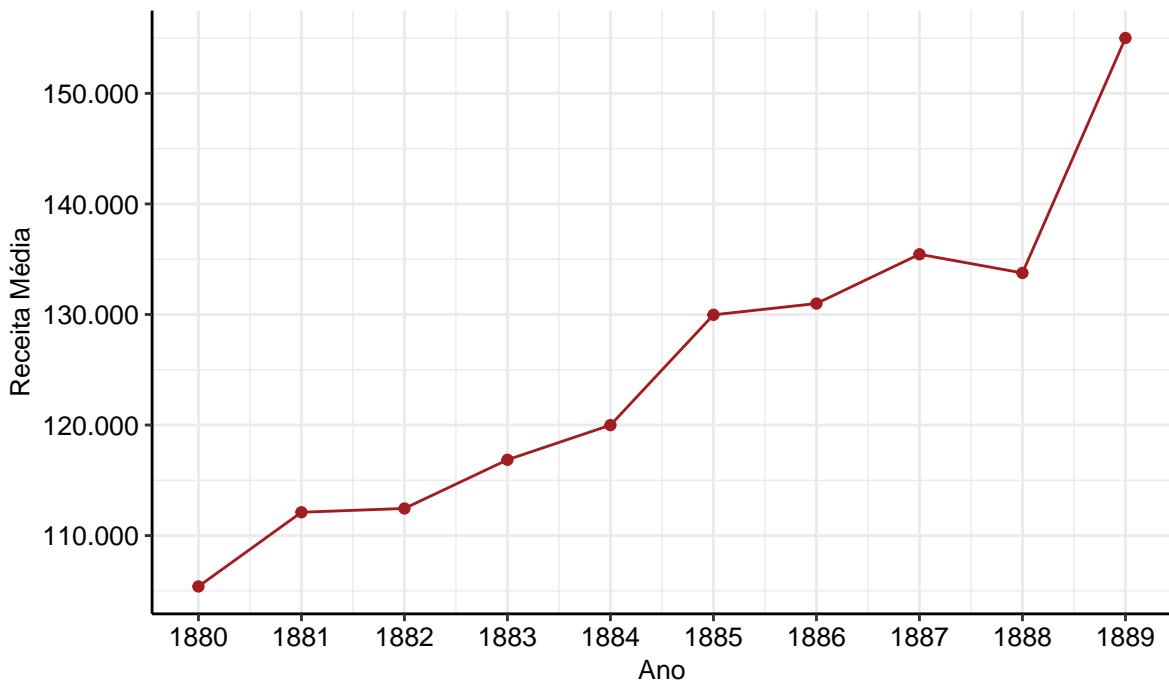


Tabela 1: Receita méida da região nos anos de 1880 a 1889

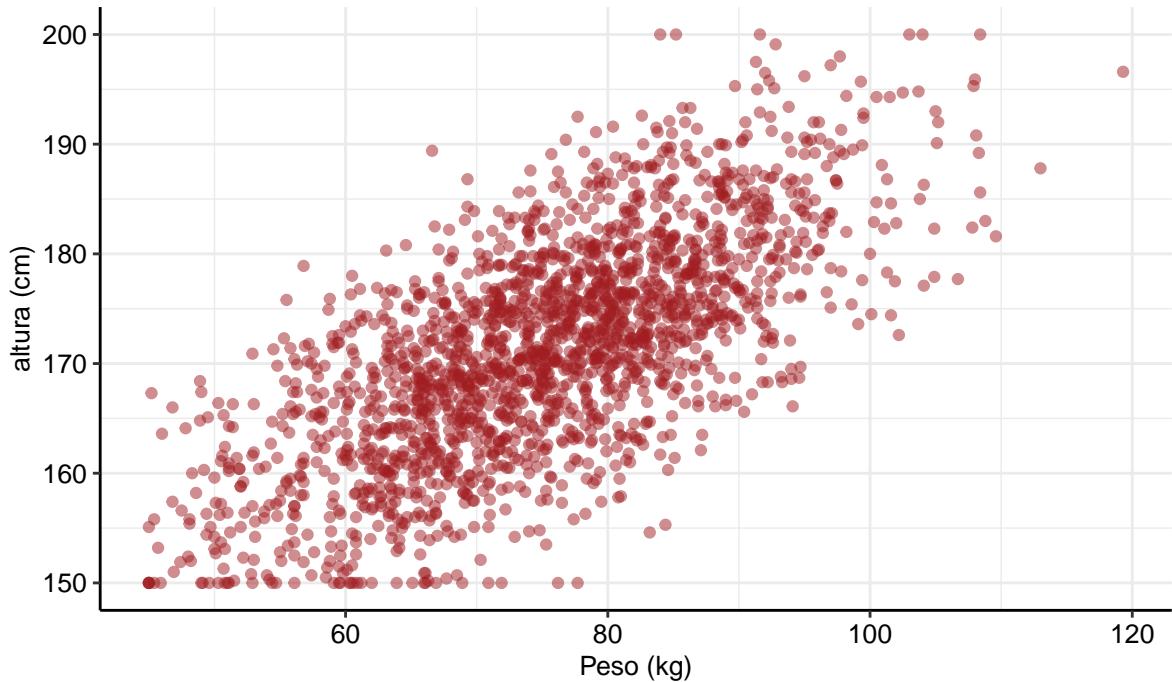
| Ano | Receita média (R\$) |
|------|---------------------|
| 1880 | 105399,0 |
| 1881 | 112110,0 |
| 1882 | 112452,4 |
| 1883 | 116856,9 |
| 1884 | 119989,8 |
| 1885 | 129969,0 |
| 1886 | 130989,2 |
| 1887 | 135444,8 |
| 1888 | 133757,6 |
| 1889 | 155009,1 |

A receita média da região apresentou um crescimento expressivo ao longo do período analisado, com um aumento total de R\$49.610,10 desde o início da série histórica. É importante destacar que, em 1888, ocorreu a primeira redução na receita média da região durante todo o período analisado, entretanto, no ano seguinte (1889), observou-se um aumento explosivo na receita média. Esses resultados indicam que a região apresenta **grande potencial** de investimento.

3.2 Variação Peso por Altura

O objetivo desta análise é investigar se existe relação linear entre o **peso** e a **altura** dos clientes, em que o peso foi convertido para quilogramas (multiplicando o valor em libras por 0,453592) e a altura convertido para centímetros (multiplicando o valor em decímetros por 10). Ambas variáveis são **quantitativas contínuas**.

Figura 5: Dispersão entre peso e altura dos clientes



Quadro 1: Principais métricas do peso dos clientes

| Estatística | Valor |
|---------------|--------|
| Média | 75,19 |
| Desvio Padrão | 11,92 |
| Variância | 142,00 |
| Mínimo | 45,00 |
| 1º Quartil | 66,90 |
| Mediana | 75,30 |
| 3º Quartil | 83,20 |
| Máximo | 119,30 |

Quadro 2: Principais métricas da altura dos clientes

| Estatística | Valor |
|---------------|--------|
| Média | 171,48 |
| Desvio Padrão | 9,87 |
| Variância | 97,38 |
| Mínimo | 150,00 |
| 1º Quartil | 164,80 |
| Mediana | 171,75 |
| 3º Quartil | 178,00 |
| Máximo | 200,00 |

Ao observar o gráfico de dispersão entre **peso** e **altura**, nota-se de imediato uma relação positiva entre as variáveis, assim como uma alta variabilidade. A medida que o peso aumenta a altura tende a aumentar e os pontos estão bem espalhados, indicando que, embora exista uma correlação positiva, o peso varia bastante mesmo entre clientes com estaturas semelhantes.

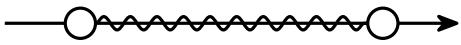
A normalidade das variáveis **peso** e **altura** foi verificada a fim de assegurar a adequação dos dados às suposições do teste. Em seguida, realizou-se o teste de correlação de Pearson, considerando as seguintes hipóteses.

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Tabela 2: Resultados da correlação de Pearson entre peso e altura

| | |
|---------------|-----------|
| $\hat{\rho}$ | 0,6971007 |
| p-valor | < 0,0001 |
| Estatística t | 43,35117 |

Sob H_0 , foi obtido um p-valor $< 2,2 \times 10^{-16}$. Ao nível de significância de 5% há evidencia estatística suficiente para rejeitar a hipótese nula, indicando que há correlação linear entre as variáveis. O **estimador** do coeficiente de correlação, o **p-valor** e a **estatística do teste** estão apresentados na Tabela 2. O intervalo de confiança, ao nível de 95%, é mostrado a seguir.

| Parâmetro | Intervalo de Confiança (95%) |
|-----------|---|
| ρ |  0.6737991 0.7190173 |

O estimador da correlação de Pearson ($0,5 < \hat{\rho} < 0,7$) indica que na população de clientes há uma relação linear positiva moderada, quase alta, entre o **peso** e a **altura**. Isso é, à medida que a altura dos clientes aumenta, o peso também tende a aumentar. O p-valor muito baixo fornece evidência estatística que a correlação linear é não nula, sugerindo que que a relação observada dificilmente ocorreu ao acaso. O intervalo de confiança indica que com 95% de confiança no procedimento

de estimação, o valor verdadeiro da correlação populacional estaria dentro do intervalo $[0,6737991; 0,7190173]$. Assim, podemos concluir que em geral, **clientes mais altos pesam mais**.

3.3 Idade dos clientes de Âmbar Seco a depender da loja

Para esta análise, o objetivo é examinar a distribuição da **idade** dos **clientes** por **lojas** na cidade de **Âmbar Seco**. Idade é variável **quantitativa discreta** e Clientes é variável **qualitativa nominal**.

Figura 6: Boxplot da idade dos clientes em Âmbar Seco

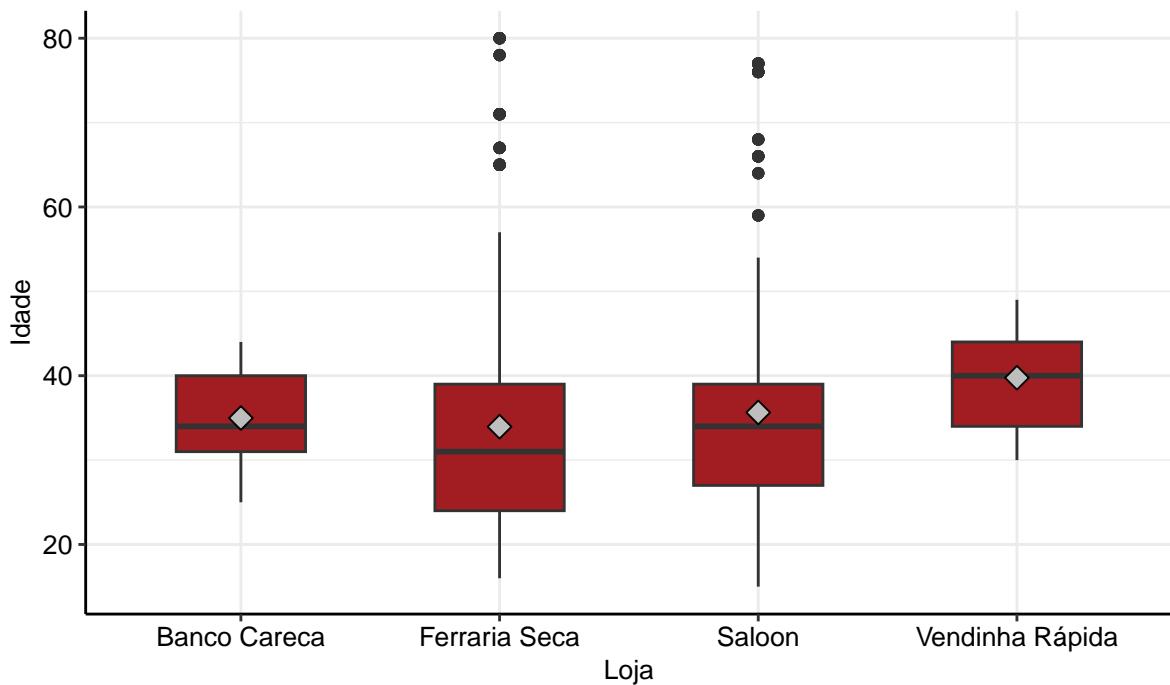


Tabela 3: Resumo da idade dos clientes em Âmbar Seco

| Loja | Média | Mediana | DP | DIQ |
|-----------------|-------|---------|-------|-----|
| Vendinha Rápida | 39,76 | 40 | 6,05 | 10 |
| Saloon | 35,64 | 34 | 12,00 | 12 |
| Banco Careca | 34,99 | 34 | 5,47 | 9 |
| Ferraria Seca | 33,95 | 31 | 13,51 | 15 |

Da Figura 6, e da distância interquantílica na Tabela 3, podemos observar que as lojas **Vendinha Rápida** e **Banco Careca** apresentam uma distribuição de idades de clientes mais concentrada. Isso pode ser visto pelo menor tamanho das caixas no boxplot, e pelos menores valores de DIQ. Enquanto as lojas **Saloon** e **Ferraria Seca**

possuem maior variação na idade dos clientes. Possuem caixas maiores e maiores valores de DIQ. A loja **Vendinha Rápida** se destaca por apresentar a maior média (39,76) e mediana (40) de idade, indicando que seus clientes tendem a ser mais velhos em comparação às demais lojas na cidade de Âmbar Seco. Por outro lado, a **Ferraria Seca** apresentou menor média (33,95) e mediana (31), sugerindo um público mais jovens, no entanto, o maior desvio padrão (13,51) e maior disntância interquantílica (15), indicam que o perfil étario dos clientes é mais vasto e heterogêno.

3.4 Os top 3 produtos mais vendidos nas top 3 lojas com maior receita em 1889

Por fim, o objetivo desta análise é identificar os três **produtos mais vendidos** nas três **lojas com maior receita** no ano de 1889. As lojas que apresentaram as maiores receitas nesse ano foram: **Loja Ouro Fino** (R\$ 197.312,50), **Loja TendTudo** (R\$ 196.340,30) e **Ferraria Apache** (R\$ 181.689,10).

Tabela 4: Top 3 lojas com maior receita em 1889

| Loja | Receita (R\$) |
|-----------------|---------------|
| Loja Ouro Fino | 197312,5 |
| Loja TendTudo | 196340,3 |
| Ferraria Apache | 181689,1 |

Tabela 5: Produtos mais vendidos nas top 3 lojas com maior receita

| Produto | Quantidade Vendida | Receita Gerada (R\$) |
|-----------------|--------------------|----------------------|
| Whisky | 139 | 7498,99 |
| Espingarda | 133 | 130850,29 |
| Chapéu de Couro | 130 | 19135,12 |
| Botas de Couro | 119 | 27202,86 |
| Colt .45 | 110 | 88508,67 |
| Sela | 101 | 54623,17 |
| Municao | 90 | 1686,99 |
| Pá | 81 | 6077,45 |
| Cavalo | 78 | 232595,20 |
| Machado | 76 | 7163,19 |

Tabela 6: Top 3 produtos nas top 3 lojas com maior receita em 1889

| Loja | Produto | Quantidade Vendida | Receita Gerada (R\$) |
|-----------------|-----------------|--------------------|----------------------|
| Loja Ouro Fino | Botas de Couro | 52 | 11886,97 |
| | Whisky | 49 | 2643,53 |
| | Chapéu de Couro | 45 | 6623,69 |
| Loja TendTudo | Espingarda | 53 | 52143,35 |
| | Whisky | 49 | 2643,53 |
| | Colt .45 | 43 | 34598,84 |
| Ferraria Apache | Chapéu de Couro | 52 | 7654,05 |
| | Espingarda | 42 | 41321,15 |
| | Machado | 41 | 3864,35 |

Figura 7: Top 3 produtos mais vendidos na Loja Ouro Fino

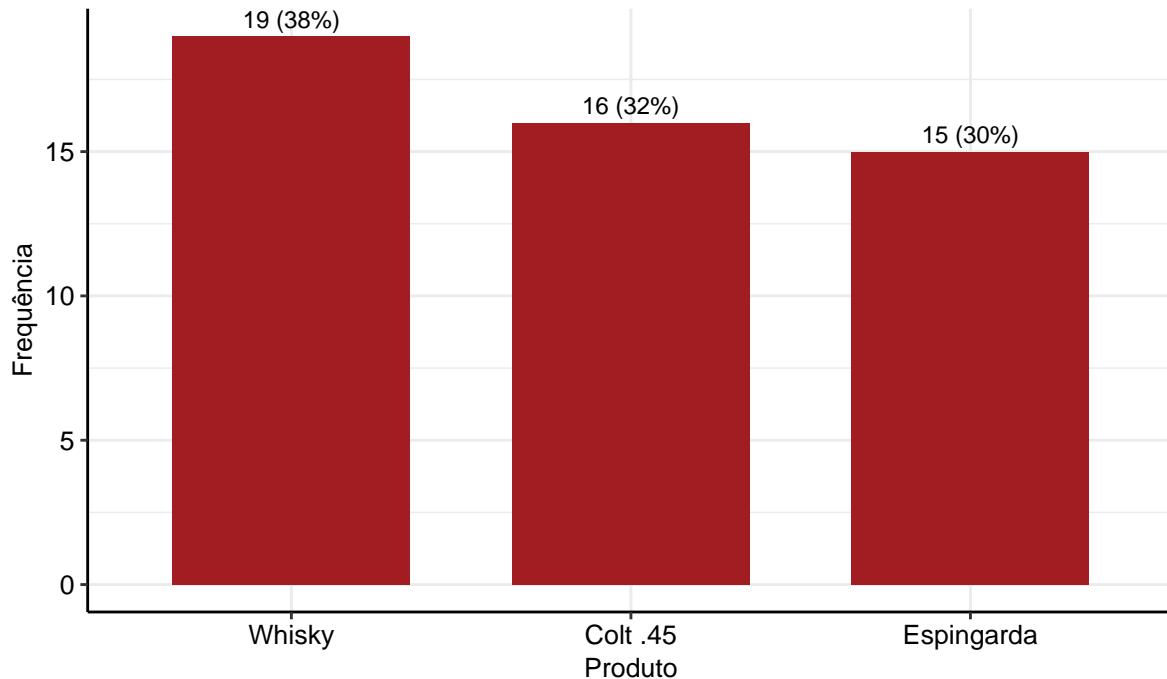


Figura 8: Top 3 produtos mais vendidos na Loja TendTudo

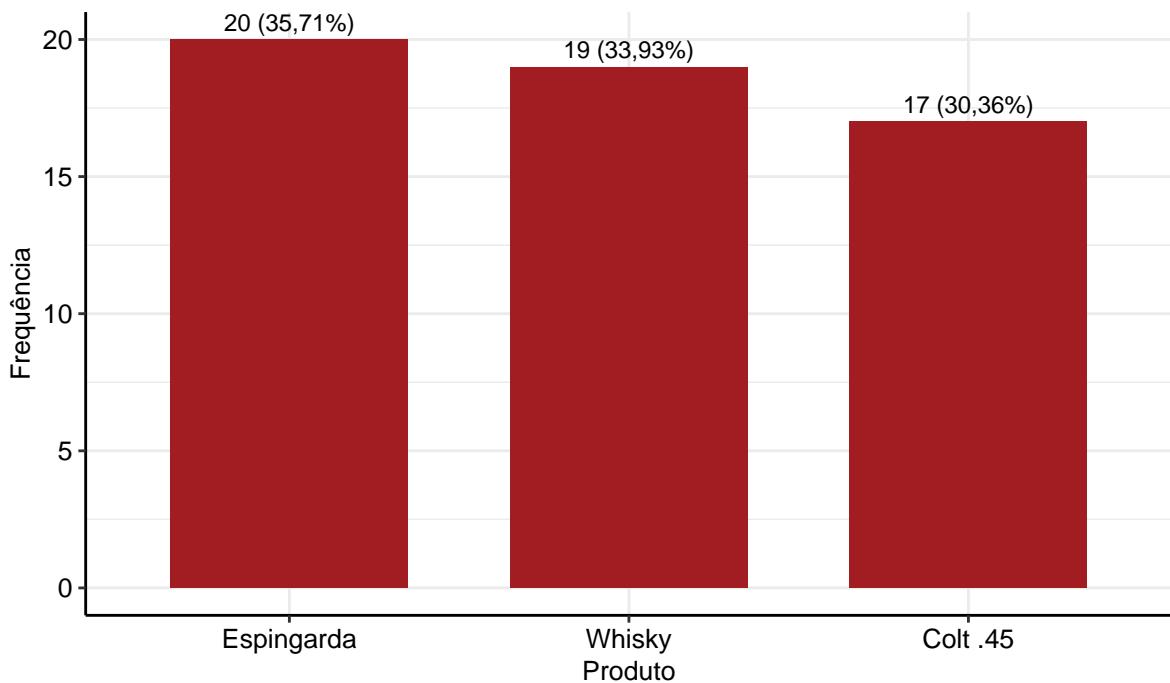
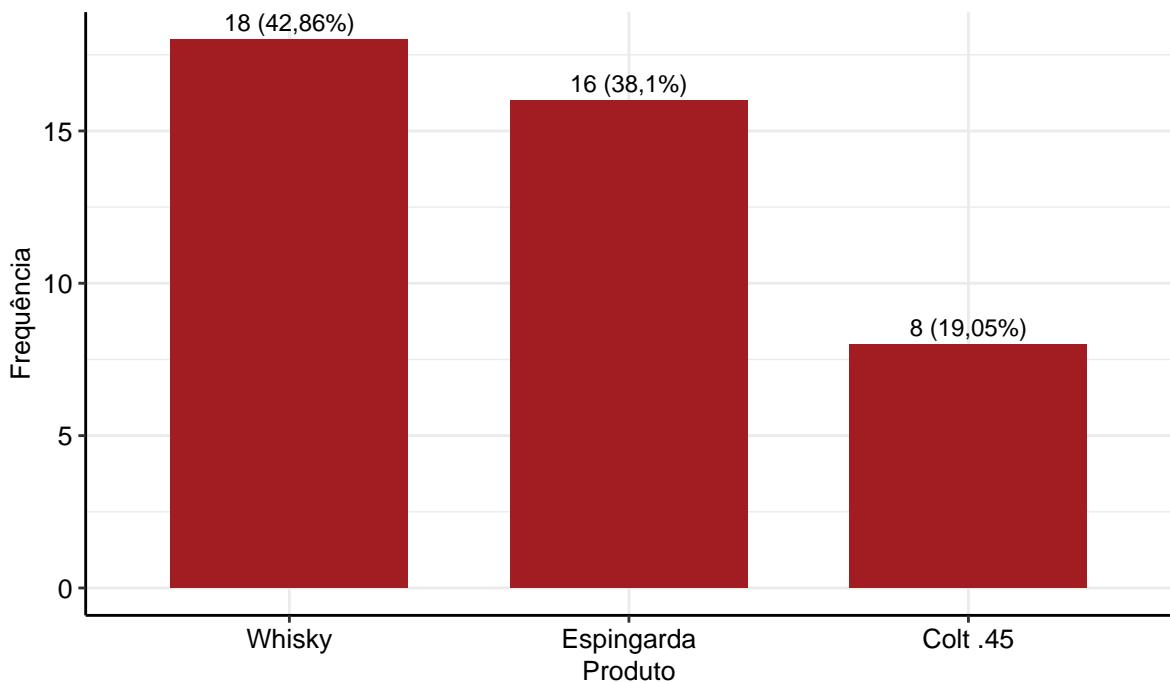


Figura 9: Top 3 produtos mais vendidos na Ferraria Apache



Pode-se observar da tabela 4 que o **Whisky** foi o produto mais vendido das três lojas em análise. O desempenho desse produto nas diferentes lojas sugere que é um item popular entre os consumidores da região, assim como o **Chapeu de Coro** e a **Espingarda** que aparecem mais de uma vez na Tabela 5. Da Tabela 5, observa-se que o produto mais vendido na loja com maior receita em 1889 (**Loja Ouro Fino**) foram as **Botas de Couro**, com 52 unidades vendidas, o que corresponde a 6,02% da

receita anual. Entre as três lojas analisadas, o produto com maior volume de vendas foi a **Espingarda**, da **Loja TendTudo**, com 53 unidades vendidas.

4 Conclusão

Ao longo deste relatório, foram exploradas diversas análises estatísticas para compreender o comportamento das vendas e dos clientes da região durante o período de 1880 a 1889. A análise da receita média revelou um crescimento significativo ao longo dos anos, indicando um potencial econômico promissor.

A investigação da relação entre **peso** e **altura** dos clientes revelou alta variabilidade, indicando que indivíduos com pesos semelhantes podem apresentar alturas bastante distintas. Além disso, observou-se uma correlação positiva moderada, sugerindo que, de modo geral, clientes mais altos tendem a pesar mais.

O perfil etário dos clientes em **Âmbar Seco** variou significativamente entre as lojas, com as lojas **Ferraria Seca** e **Saloon** apresentando uma faixa etária mais ampla, enquanto **Banco Careca** e **Vendinha Rápida** uma faixa etária mais concentrada.

Finalmente, a análise dos produtos mais vendidos nas lojas com maior receita em 1889 destacou itens populares entre os consumidores, como o **Whisky**, a **Espingarda** e o **Chapeu de Couro**.

5 Anexo

1. Dicionário das variáveis do banco de dados

- ClientID: Chave do cliente
- Name: O nome do cliente
- Age: A idade do cliente em anos
- Sex: Sexo do Cliente
- Height_dm: A altura do cliente em decímetros
- Weight_lbs: O peso do cliente, em libras
- Anual_Income_usd: Renda do cliente anualmente em dólares
- CityID: Chave da cidade
- NameCity: Nome da cidade
- EmployeeID: Chave do funcionário
- EmployeeName: Nome do funcionário
- StoreID: Chave da loja
- StoreName: Nome da loja
- CityID: Chave da cidade
- ItemID: Chave do produto
- NameProduct: Nome do produto
- UnityPrice: Preço unitário do produto em dólares
- SaleID: Chave da venda
- ItemID: Chave do produto
- SaleID: Chave da venda
- Date: A data da ocorrência da venda (YYYY-MM-DD)
- StoreID: Chave da loja
- ClientID: Chave do cliente
- Quantity: Quantidade comprada deste produto

2.

```
sessionInfo()
```

```
R version 4.5.0 (2025-04-11 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)
```

```
Matrix products: default
```

```
LAPACK version 3.12.1
```

```

locale:
[1] LC_COLLATE=Portuguese_Brazil.utf8  LC_CTYPE=Portuguese_Brazil.utf8
[3] LC_MONETARY=Portuguese_Brazil.utf8 LC_NUMERIC=C
[5] LC_TIME=Portuguese_Brazil.utf8

time zone: America/Sao_Paulo
tzcode source: internal

attached base packages:
[1] stats      graphics   grDevices utils      datasets   methods    base

other attached packages:
[1] kableExtra_1.4.0   knitr_1.50          wordcloud_2.6       abjutils_0.3.2
[5] ggrepel_0.9.6     sf_1.0-21           geojsonr_1.9.1      xtable_1.8-4
[9] skimr_2.2.1       xlsx_0.6.5          nortest_1.0-4       scales_1.4.0
[13] RColorBrewer_1.1-3 cowplot_1.2.0      ggcorrplot_0.1.4.1  readxl_1.4.5
[17] pROC_1.19.0.1    data.table_1.17.8   lubridate_1.9.4    forcats_1.0.1
[21] stringr_1.5.2    dplyr_1.1.4          purrr_1.1.0         readr_2.1.5
[25] tidyverse_1.3.1   tibble_3.3.0        ggplot2_4.0.0       tidyverse_2.0.0
[29] pacman_0.5.1

loaded via a namespace (and not attached):
[1] gtable_0.3.6      xfun_0.53          rJava_1.0-11        tzdb_0.5.0
[5] vctrs_0.6.5       tools_4.5.0         generics_0.1.4      curl_7.0.0
[9] proxy_0.4-27     pkgconfig_2.0.3    KernSmooth_2.23-26 S7_0.2.0
[13] lifecycle_1.0.4   compiler_4.5.0    farver_2.1.2        textshaping_1.0.3
[17] tinytex_0.57     repr_1.1.7          codetools_0.2-20   htmltools_0.5.8.1
[21] class_7.3-23     yaml_2.3.10        pillar_1.11.1       classInt_0.4-11
[25] tidyselect_1.2.1  digest_0.6.37    stringi_1.8.7       fastmap_1.2.0
[29] grid_4.5.0       cli_3.6.5          magrittr_2.0.4      base64enc_0.1-3
[33] xlsxjars_0.9.0   e1071_1.7-16    withr_3.0.2         timechange_0.3.0
[37] rmarkdown_2.30    cellranger_1.1.0  hms_1.1.3          evaluate_1.0.5
[41] viridisLite_0.4.2 rlang_1.1.6        Rcpp_1.1.0          glue_1.8.0
[45] DBI_1.2.3        xml2_1.4.0        svglite_2.2.1       rstudioapi_0.17.1
[49] jsonlite_2.0.0   R6_2.6.1          systemfonts_1.3.0  units_0.8-7

```