Modelo Projeto - Quarto

Consultores Responsáveis:
Gabriel Cunha Gonçalves Garcia

Requerente:
João Sábio

Brasília, November 9, 2025.

ESTAT
consultoria estatística

Sumário

1 Introdução

Este relatório apresenta as análises realizadas para o cliente João Sábio, proprietário da empresa Old Town Road Ltda., com o objetivo de identificar padrões, propor melhorias e gerar insights baseados em dados. As investigações foram organizadas em quatro etapas principais, cada uma abordando um aspecto específico do conjunto de dados disponibilizado, a fim de oferecer uma compreensão ampla sobre o desempenho das lojas e o perfil de seus clientes ao longo do período analisado.

As análises foram realizadas a partir do banco de dados relatório_old_town_road, composto pelas abas Relatório de Vendas, Informação de Vendas, Produto, Cliente, Funcionário, Cidade e Loja. Para o processamento, tratamento e exploração dos dados, foi utilizado o software RStudio, que possibilitou a aplicação de métodos estatísticos e de visualização para obtenção dos resultados apresentados.

2 Referencial Teórico

2.1 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:

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

2.2 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:

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

2.3 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. ## Desvio Padrão

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

2.3.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

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

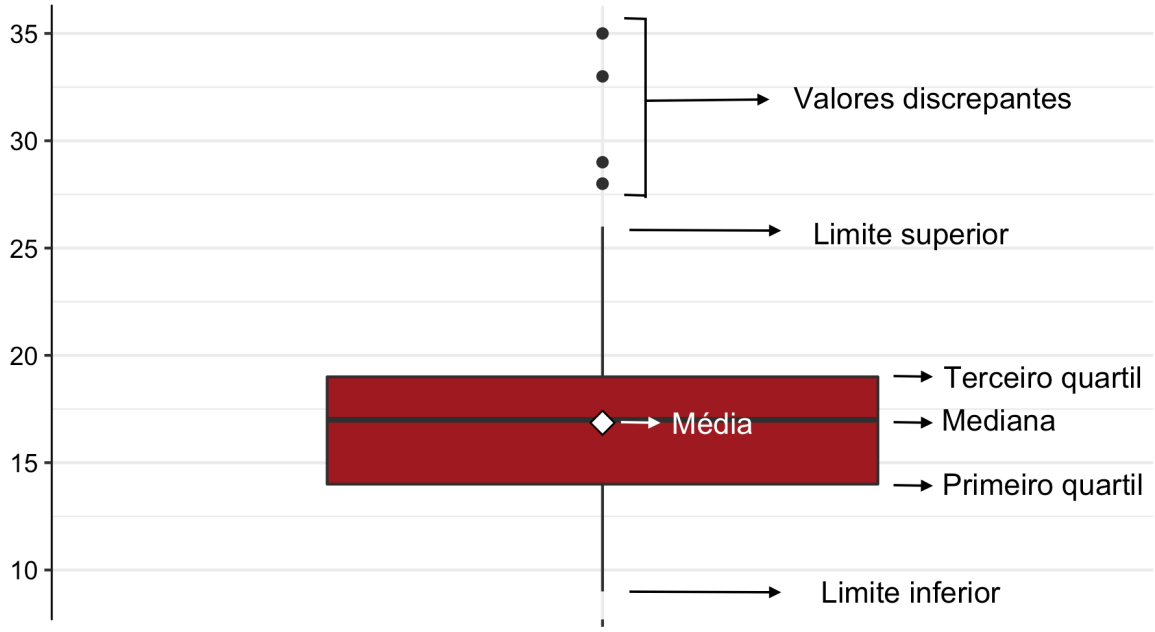
Com:

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

2.4 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.

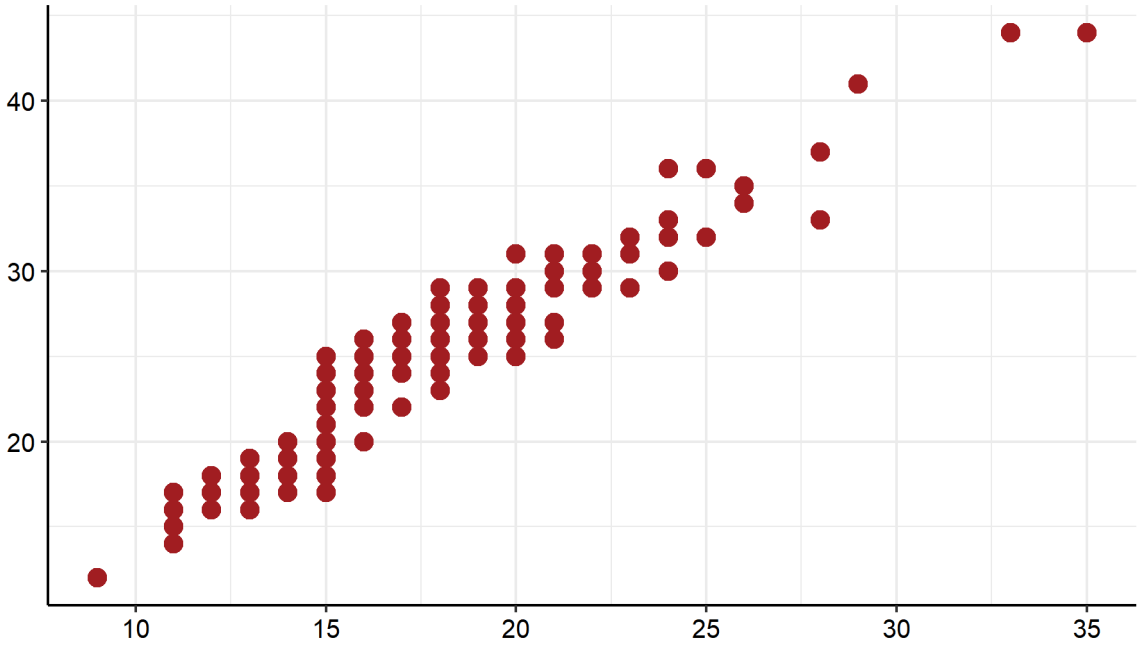
Figura 1: Exemplo de boxplot



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. ## 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 2: Exemplo de Gráfico de Dispersão



2.5 Tipos de Variáveis

2.5.1 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)

2.5.2 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) ## Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica, através de postos de variáveis quantitativas ou qualitativas ordinais, o grau de relação linear entre duas variáveis. 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 ρ é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando ρ é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente é calculado da seguinte maneira:

$$\rho_{Spearman} = \frac{\sum_{i=1}^{n}\left[\left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(R(x_i)^2\right) - n\left(\frac{n+1}{2}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(R(y_i)^2\right) - n\left(\frac{n+1}{2}\right)^2}}$$

Onde:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $R(x_i) =$ posto relativo à observação i de X
- $R(y_i) =$ posto relativo à observação i de Y
- $n =$ número total de observações na amostra # Definição para Testes

2.6 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.

2.7 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:

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

2.8 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%.

2.9 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.

2.10 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-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.

2.11 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.

3 Teste de Normalidade

Os testes de normalidade são utilizados para verificar se uma variável aleatória segue um distribuição Normal de probabilidade ou não. Eles são muito importantes, pois impactam em qual teste deve ser utilizado em uma análise futura. Se o resultado do teste confirmar que a variável segue uma distribuição normal, procedimentos paramétricos podem e devem ser utilizados. Caso contrário, os métodos não paramétricos são mais recomendados.

3.1 Teste de Normalidade de Shapiro-Wilk

O **Teste de Shapiro-Wilk** é utilizado para verificar a aderência de uma variável quantitativa ao modelo da Distribuição Normal, sendo mais recomendado para amostras pequenas. A suposição de normalidade é importante para a determinação do teste a ser utilizado. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

A amostra deve ser ordenada de forma crescente para que seja possível obter as estatísticas de ordem. A estatística do teste é dada por:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)}\right)\right]$$

Com:

- K aproximadamente $\frac{n}{2}$
- $X_{(i)} =$ estatística de ordem *i*
- $D = \sum_{i=1}^{n}(X_i - \bar{X})^2$, em que $\bar{X}$ é a média amostral
- $a_i = $ constantes que apresentam valores tabelados

4 Análises

4.1 Receita média total das lojas (1880-1889)

Nesta análise, busca-se compreender a evolução da receita média total das lojas entre os anos de 1880 e 1889. Os valores estão expressos em reais , considerando a cotação de 1 dólar = R$5,31.

Foram utilizadas as seguintes variáveis:

Faturamento médio das lojas: variável quantitativa contínua, obtida pela soma do faturamento total anual de cada loja (quantidade vendida × preço unitário em reais), dividida pelo número de lojas ativas em cada ano. Essa medida permite avaliar o desempenho médio das lojas ao longo do período analisado.

Ano: variável qualitativa ordinal, que indica o período de referência das observações.

Abaixo é apresentada um gráfico e uma tabela com medidas estatísticas das receitas no período analisado:

Figura 3: Gráfico de faturamento médio das lojas por ano

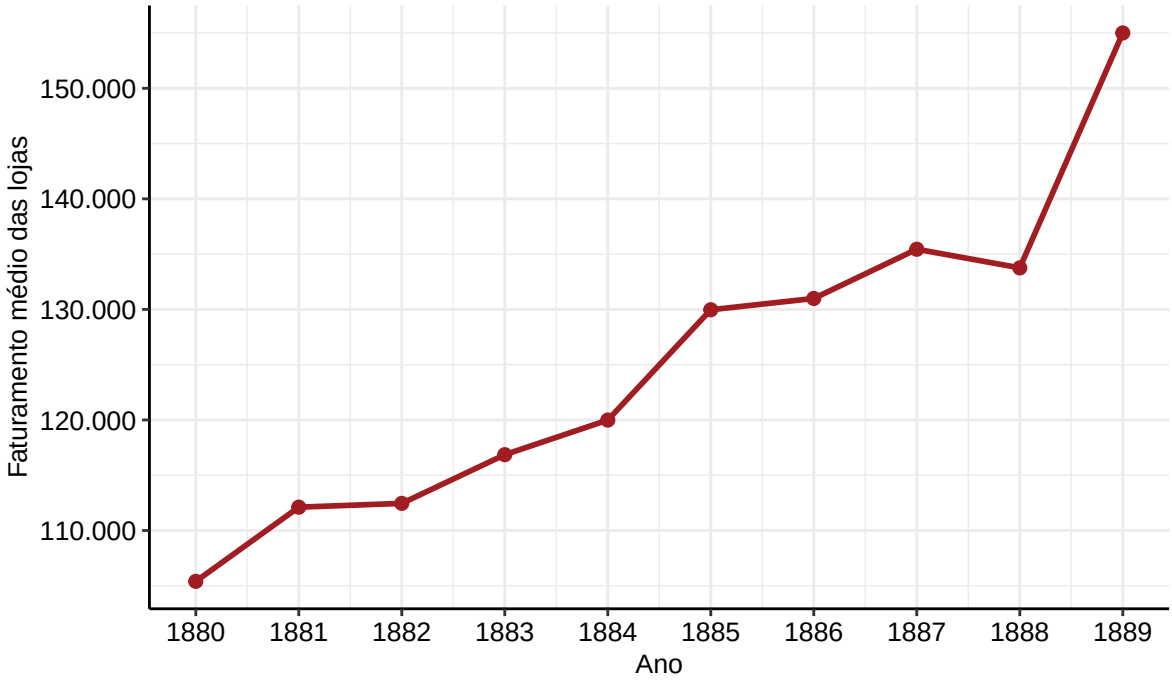


Tabela 1: Faturamento médio das lojas por ano

Ano	Faturamento Médio (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

Como mostrado na **Figura 3** e na **Tabela 1**, as receitas médias variaram entre R$105.399,00 e R$155.009,10, com uma tendência de crescimento ao longo da década. Observa-se um crescimento contínuo da receita média das lojas ao longo dos anos, com pequenas oscilações em 1887–1888, mas uma tendência geral positiva. O maior crescimento ocorreu entre 1888 e 1889, quando o faturamento médio aumentou aproximadamente 15,9%, evidenciando um expressivo avanço no desempenho das lojas. A análise descritiva demonstra, portanto, um aumento consistente na receita média total entre 1880 e 1889.

4.2 Variação Peso por Altura

Nesta análise, busca-se compreender a relação entre o peso (em quilogramas) e a altura (em centímetros) dos clientes. O objetivo é verificar se há uma associação estatisticamente significativa entre essas variáveis — isto é, se indivíduos com maior peso tendem a apresentar maior altura ou se não existe uma relação consistente entre elas.

Antes da aplicação dos testes de correlação, foi avaliada a normalidade dos dados utilizando o teste de Shapiro–Wilk. Os resultados indicaram que tanto a variável Peso quanto a variável Altura não seguem uma distribuição normal. Diante disso, optou-se pela utilização de um teste não paramétrico, o teste de correlação de Spearman, adequado para esse tipo de dado e considerando um nível de confiança de 95%.

Foram utilizadas as seguintes variáveis:

Peso (kg): variável quantitativa contínua, obtida a partir da conversão dos valores originais em libras para quilogramas, utilizando a equivalência de 1 libra = 0,453592 kg.

Altura (cm): variável quantitativa contínua, obtida pela conversão de decímetros para centímetros, considerando 1 dm = 10 cm.

A seguir, apresenta-se um gráfico de dispersão que ilustra o comportamento

conjunto das variáveis Altura (cm) e Peso (kg), bem como uma tabela resumo dos teste de hipótese aplicado:

Figura 4: Gráfico de dispersão entre Peso e Altura dos clientes

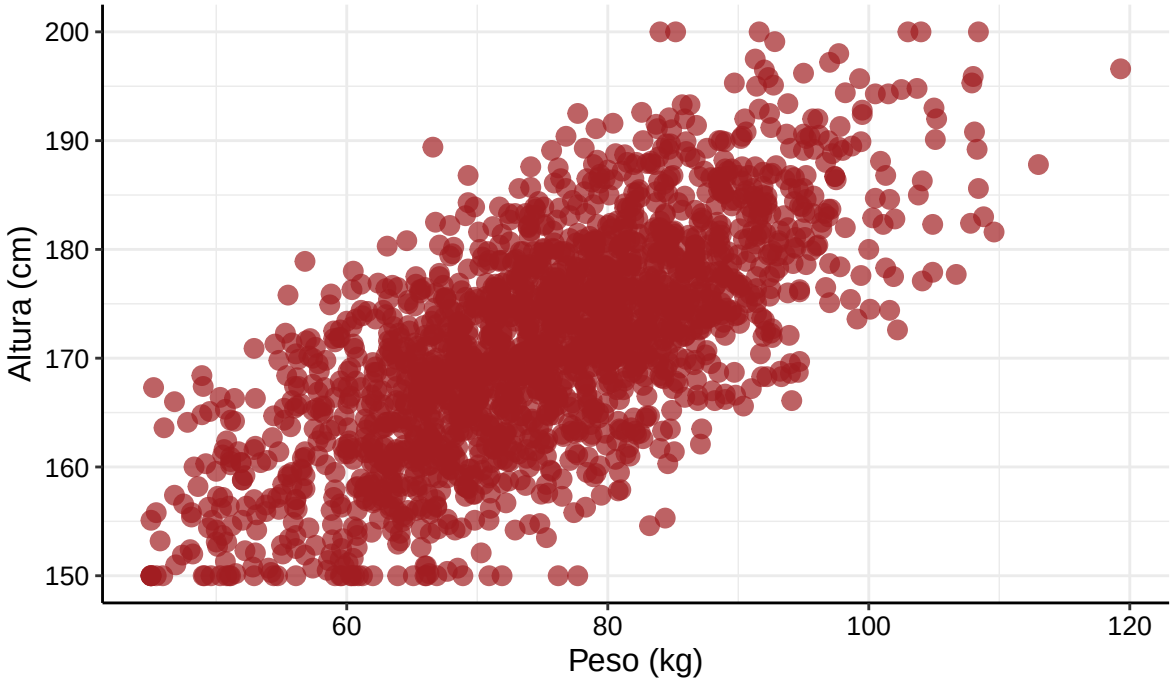


Tabela 2: Teste de hipótese de Spearman

Teste	Variável	Estatística	p-valor	Decisão	Interpretação
Spearman	Peso × Altura	0.68	0.00	Rejeita H0	Correlação significativa

Quadro 1: Medidas resumo da variável Peso (kg)

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: Medidas resumo da variável Altura (cm)

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

Observa-se, pela **Figura 4**, uma relação positiva clara entre as variáveis, evidenciando que, de forma geral, à medida que o peso aumenta, a altura também tende a aumentar. Essa tendência linear crescente é confirmada pelos resultados do teste de correlação de Spearman apresentados no **Tabela 2** , o qual revelou uma correlação positiva significativa ($\rho = 0{,}6865$; p-valor $< 0{,}000001$).

A análise foi realizada com base em um total de 1.990 clientes, após a aplicação dos critérios de filtragem, garantindo a consistência e representatividade dos dados utilizados.

Esses resultados indicam a existência de uma associação estatisticamente relevante entre peso e altura, sugerindo que indivíduos com maior peso tendem, em média, a apresentar maior estatura. A análise conjunta — visual e estatística — reforça a consistência dessa relação nas amostras observadas.

4.3 Distribuição das Idades dos Clientes por Loja — Âmbar Seco

Nesta análise, busca-se compreender a distribuição das idades dos clientes nas quatro lojas da cidade de Âmbar Seco: Ferraria Seca, Saloon, Banco Careca e Vendinha Rápida. O objetivo é identificar padrões etários entre os consumidores, avaliando se há diferenças significativas no perfil de idade dos clientes que frequentam cada estabelecimento.

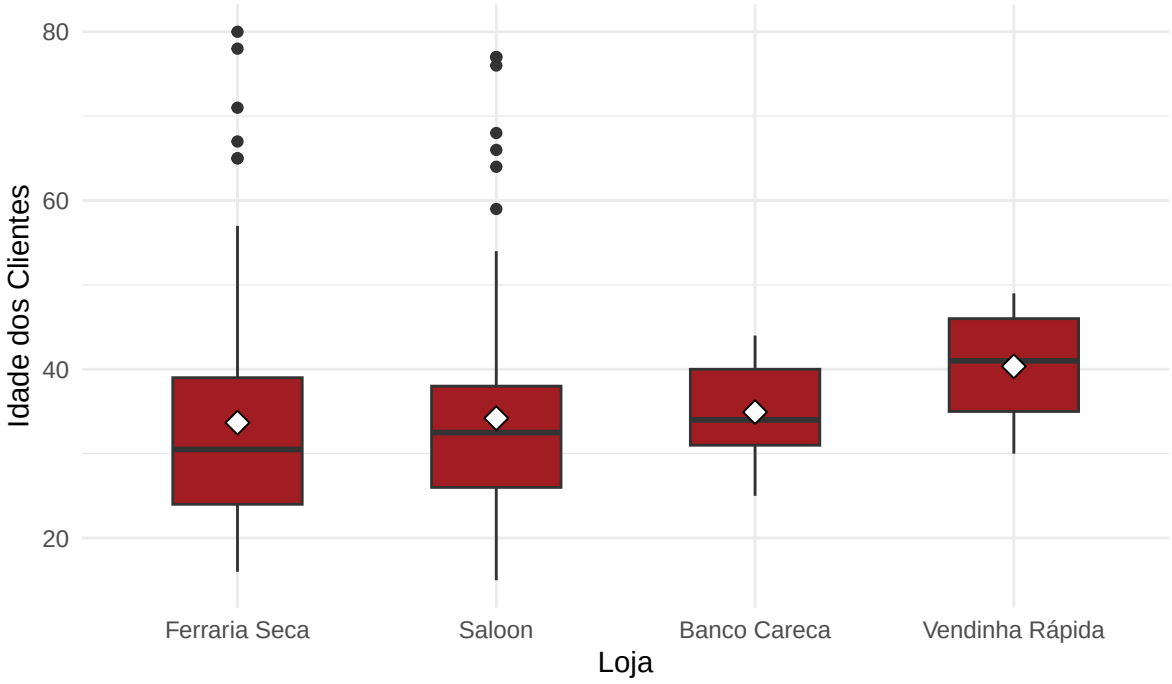
Foram utilizadas as seguintes variáveis:

Idade: variável quantitativa discreta , representando a idade dos clientes (em anos).

Loja: variável qualitativa nominal, correspondente ao nome da loja na qual a venda foi registrada.

Abaixo, é apresentado o gráfico boxplot que demonstra a dispersão das idades em cada loja:

Figura 5: Distribuição das Idades dos Clientes por Loja - Âmbar Seco



Quadro 3: Medidas resumo da variável Idade — Loja: Banco Careca

Estatística	Valor
Média	34,92
Desvio Padrão	5,57
Variância	31,06
Mínimo	25,00
1º Quartil	31,00
Mediana	34,00
3º Quartil	40,00
Máximo	44,00

Quadro 4: Medidas resumo da variável Idade — Loja: Ferraria Seca

Estatística	Valor
Média	33,67
Desvio Padrão	13,31
Variância	177,18
Mínimo	16,00
1º Quartil	24,00
Mediana	30,50
3º Quartil	39,00
Máximo	80,00

Quadro 5: Medidas resumo da variável Idade — Loja: Saloon

Estatística	Valor
Média	34,20
Desvio Padrão	12,70
Variância	161,23
Mínimo	15,00
1º Quartil	26,00
Mediana	32,50
3º Quartil	38,00
Máximo	77,00

Quadro 6: Medidas resumo da variável Idade — Loja: Vendinha Rápida

Estatística	Valor
Média	40,35
Desvio Padrão	6,03
Variância	36,39
Mínimo	30,00
1º Quartil	35,00
Mediana	41,00
3º Quartil	46,00
Máximo	49,00

A Figura **Figura 5**, complementada pelos **Quadros de medida resumo** da variável Idade por loja, apresenta a distribuição das idades dos clientes em cada loja da cidade de Âmbar Seco. Após a aplicação do filtro, foram analisados 415 clientes.

De forma geral, nota-se que as idades médias variam entre aproximadamente 33,7 e 40,3 anos, indicando que o público das lojas é predominantemente adulto. A Vendinha Rápida se destaca com a maior média e mediana (40,3 e 41 anos, respectivamente), além de baixa dispersão (desvio-padrão ≈ 6,0). Isso sugere um público mais homogêneo e ligeiramente mais velho.

O Banco Careca apresenta média (34,9) e mediana (34) muito próximas, com pequena variabilidade (desvio-padrão ≈ 5,6), indicando um perfil etário equilibrado e consistente — majoritariamente composto por adultos jovens.

Em contraste, as lojas Ferraria Seca e Saloon exibem as maiores dispersões (desvios-padrão ≈ 13,3 e 12,7, respectivamente) e amplitudes interquartis mais largas (Q1–Q3 variando de 24 a 39 e 26 a 38). Essa característica, aliada à presença de outliers no gráfico, evidencia que atendem tanto a clientes jovens quanto a faixas etárias mais elevadas, refletindo maior diversidade de público.

Por fim, observa-se que, em todas as lojas, média e mediana permanecem próximas, indicando distribuições aproximadamente simétricas. Ainda assim, os valores máximos registrados (até 80 anos em algumas lojas) reforçam a presença de clientes idosos ocasionais.

Em síntese, a análise conjunta da Figura **Figura 5** e das **Medidas resumo da variável Idade** revela diferenças sutis, mas relevantes, no perfil etário dos consumidores de Âmbar Seco. Enquanto Vendinha Rápida atrai um público mais maduro e homogêneo, Ferraria Seca e Saloon destacam-se pela maior heterogeneidade etária, e o Banco Careca mantém um perfil centrado em jovens adultos.

4.4 Top 3 produtos mais vendidos nas top 3 lojas com maior receita em 1889:

Nesta análise, busca-se identificar os produtos mais vendidos nas três lojas com maior faturamento no ano de 1889: Ferraria Apache, Loja Ouro Fino e Loja TendTudo. O objetivo é compreender o comportamento das vendas por produto dentro de cada estabelecimento e avaliar a relação entre o volume de vendas e o desempenho financeiro das lojas.

Foram consideradas as seguintes variáveis:

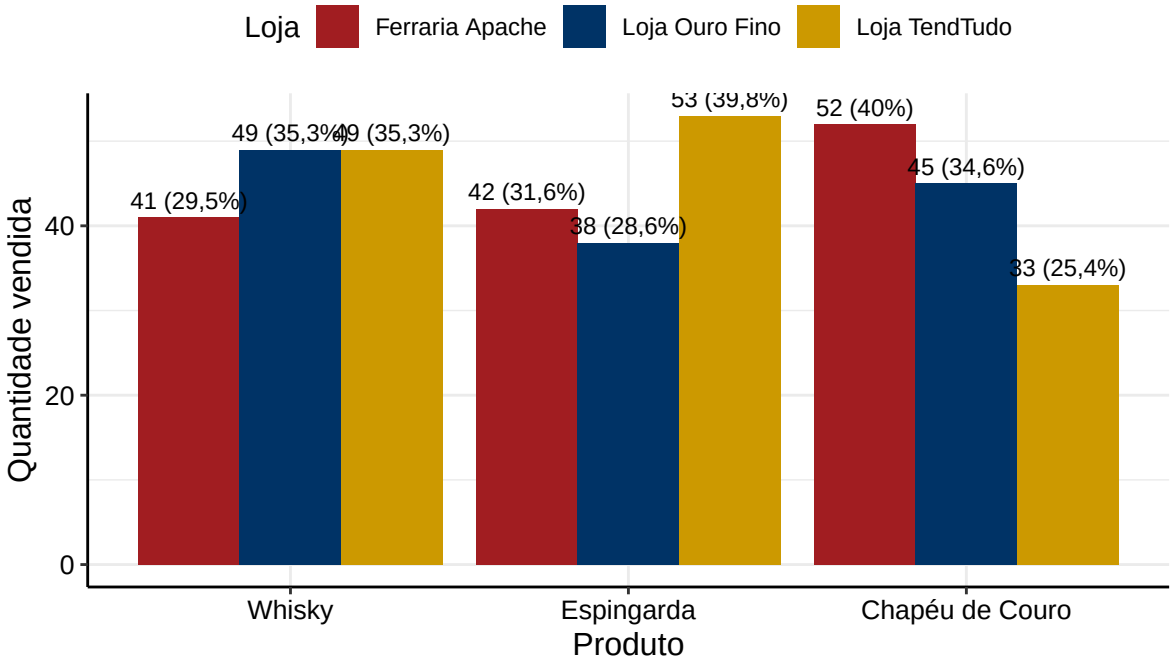
Produto: variável qualitativa nominal, representando o item comercializado.

Quantidade Vendida: variável quantitativa discreta, correspondendo ao número de unidades vendidas de cada produto.

Loja: variável qualitativa nominal, identificando o estabelecimento responsável pela venda.

Abaixo, apresenta-se os 3 produtos mais vendidos em cada loja no ano de 1889:

Figura 6: Top 3 Produtos Mais Vendidos em Cada Loja (Ano 1889)



A **Figura 6** evidencia que, embora o volume de vendas varie entre os produtos e lojas, há padrões claros de preferência dos consumidores e eficiência de comercialização.

A Loja Ouro Fino e a Loja TendTudo apresentaram desempenho bastante semelhante em volume total de vendas, enquanto a Ferraria Apache manteve resultados próximos, indicando uma concorrência equilibrada entre as três unidades.

Na Ferraria Apache, o destaque foi o Chapéu de Couro, com 52 unidades vendidas, seguido pela Espingarda (42) e pelo Whisky (41).

Na Loja Ouro Fino, os itens de maior saída foram o Whisky (49), o Chapéu de Couro (45) e a Espingarda (38).

Já a Loja TendTudo liderou em volume individual com a Espingarda (53), seguida pelo Whisky (49) e pelo Chapéu de Couro (33).

De forma geral, observa-se que Whisky e Espingarda aparecem entre os mais vendidos em mais de uma loja, o que indica preferência generalizada por esses produtos no mercado de 1889.

Em síntese, a **Figura 6** demonstra que as estratégias de vendas diferenciadas das três lojas resultaram em volumes similares, mas com perfis de consumo específicos.

4.5 Conclusão

As análises desenvolvidas para o cliente João Sábio, proprietário da Old Town Road Ltda., permitiram compreender de forma ampla o desempenho comercial das lojas e o perfil dos consumidores na região de Âmbar Seco.

De modo geral, os resultados mostraram que:

A receita média das lojas aumentou de R$105.399,00 em 1880 para R$155.009,10 em 1889, representando um crescimento acumulado de aproximadamente 47% ao longo da década. O maior avanço ocorreu entre 1888 e 1889, quando o faturamento médio subiu 15,9%, evidenciando um período de expansão mais intensa do comércio local.

A relação entre peso e altura dos clientes apresentou correlação positiva significativa ($\rho = 0{,}6865$; p-valor $< 0{,}000001$), indicando que, em média, indivíduos mais altos tendem também a possuir maior peso. Esse resultado reforça uma tendência linear clara entre as variáveis analisadas.

A distribuição etária dos clientes variou conforme a loja: a Vendinha Rápida concentrou o público mais maduro (média ≈ 40,3 anos), enquanto o Banco Careca apresentou perfil mais jovem e homogêneo (média ≈ 34,9 anos). Já Ferraria Seca e Saloon exibiram maior dispersão etária, com clientes entre 18 e 80 anos, refletindo maior diversidade de público.

Por fim, a análise dos produtos mais vendidos revelou que Whisky e Espingarda se destacaram como os itens de maior popularidade, aparecendo entre os três mais vendidos em todas as lojas de maior faturamento. Esses produtos contribuíram de forma decisiva para a composição da receita total de 1889.