Relatório Old Town Road

Consultores Responsáveis:
Davi Folha Desseaux

Requerente:
João Sábio

Brasília, 2 de novembro de 2025.

ESTAT
consultoria estatística

Sumário

1 Introdução

O seguinte projeto tem como objetivo realizar análises estatísticas que visam compreender melhor o mercado e comércio no Faroeste. Serão feitos estudos sobre correlações, distribuições, padrões e outras métricas relevantes para se obter soluções embasadas estatisticamente. Além disso, gráficos, tabelas e quadros que ajudem na visualização de tudo que foi citado. O nível de significância utilizado será de 5%.

A base de dados representa uma amostra não probabilística por conveniência, composta pelos registros disponíveis no sistema da empresa. Embora não tenha sido realizada uma seleção aleatória, o conjunto de dados é suficientemente amplo e variado, o que permite análises representativas do perfil dos clientes. O banco de dados foi cedido pelo próprio cliente e abrange variáveis de todo tipo (quantitativas e qualitativas), englobando informações sobre lojas, cidades, clientes, vendas e muito mais.

O software utilizado será o R na versão 4.3.0.

2 Referencial Teórico

Este relatório é composto por técnicas estatísticas que serão descristas a seguir de acordo com o que foi utilizado em tal estudo.

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.

2.4 Variância

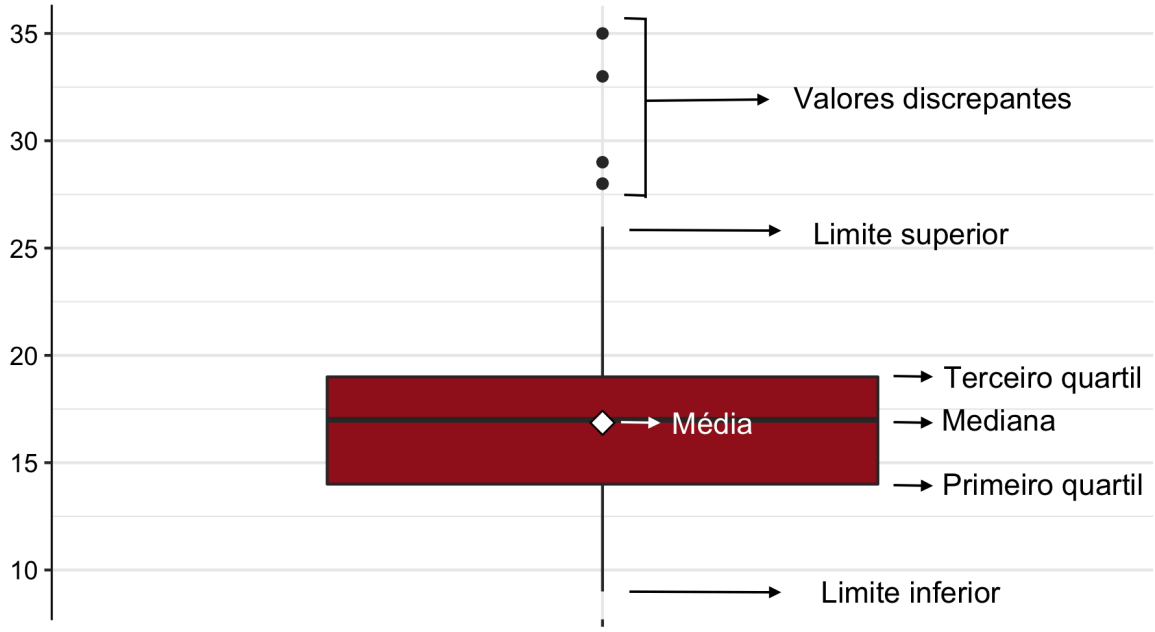
A variância é uma medida que avalia o quanto os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

2.5 Desvio Padrão

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

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

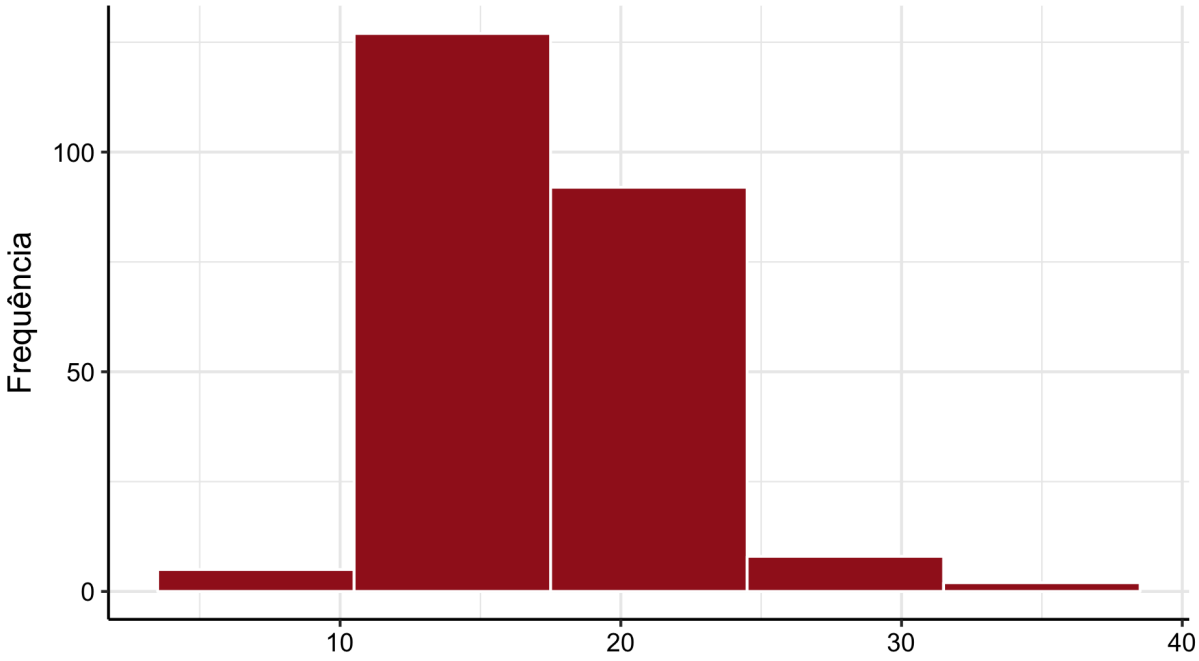


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada 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.

2.6 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.

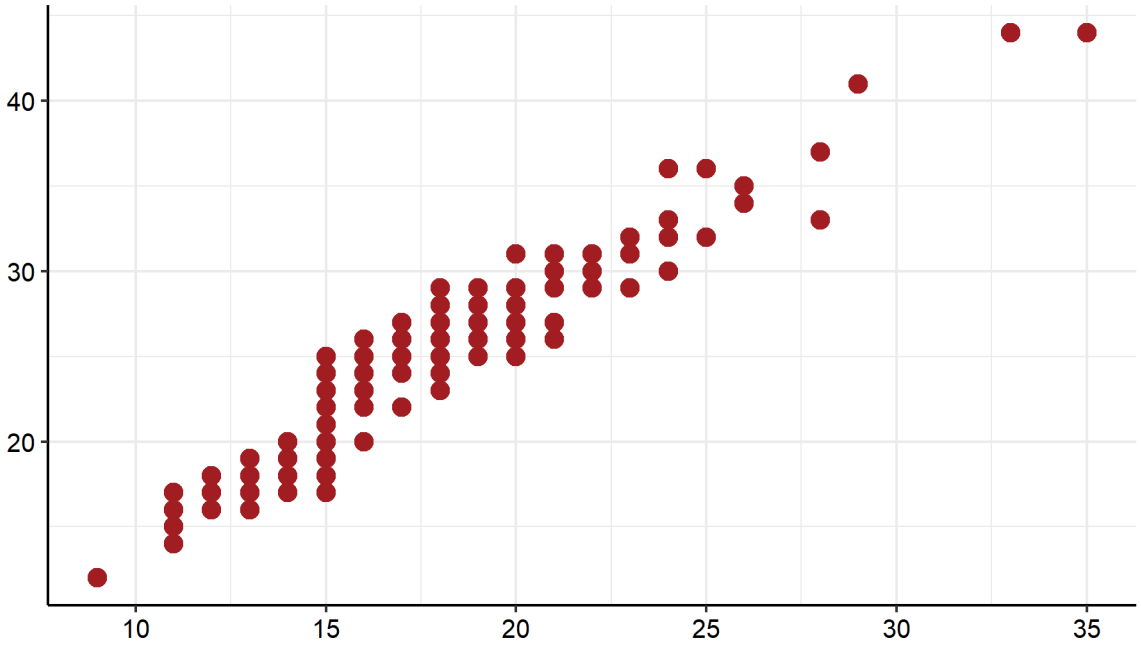
Figura 2: Exemplo de histograma



2.7 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



2.8 Tipos de Variáveis

2.8.1 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)

2.8.2 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)

2.9 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.

2.9.1 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.

2.9.2 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%.

2.9.3 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.

2.9.4 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.

2.9.5 Teste de Correlação de Postos de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica, por meio de postos de variáveis quantitativas ou qualitativas ordinais, o grau de correlação linear entre duas variáveis. Esse 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 ter uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, afirma-se que as duas variáveis são diretamente proporcionais.

O coeficiente é calculado da seguinte maneira:

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

no qual

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $R(x_i) =$ posto atribuído a x_i, quando comparado a outros valores de x
- $R(y_i) =$ posto atribuído a y_i, quando comparado a outros valores de y
- $n =$ número total de observações na amostra
- $r_{Spearman} =$ coeficiente de correlação de postos de Spearman amostral

Observação: ao ordenar de forma crescente as observações $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$ afirma-se que $R(X_1) = 1$ se x_1 é o menor valor da amostra, $R(X_3) = 2$ se x_3 é o segundo menor valor da amostra, $R(X_4) = 3$ se x_4 é o terceiro menor valor, e assim sucessivamente. Quando há empates nas observações, o posto atribuído a elas é a média dos postos que teriam se não houvesse empate. Por exemplo, se X assume os valores 9, 5, 6 e 9, tem-se duas observações com mesmo valor e, assim, seus postos serão obtidos por meio da média entre 3 e 4, que seriam seus postos se não houvesse empate.

Por ser um método não paramétrico, não há suposições para o teste.

Para a realização do teste, são feitas as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não há correlação de postos entre as variáveis } X \text{ e } Y \\ \qquad (\rho_{Spearman} = 0) \\ H_1 : \text{Há correlação de postos entre as variáveis } X \text{ e } Y \\ \qquad (\rho_{Spearman} \neq 0) \end{cases}$$

em que $\rho_{Spearman}$ é o coeficiente de correlação de postos populacional (parâmetro a ser testado com base em $r_{Spearman}$).

Além disso, a hipótese alternativa também pode ser escrita para testar se a correlação dos postos é positiva ($\rho_{Spearman} > 0$) ou negativa ($\rho_{Spearman} < 0$).

A estatística T do teste ($T = r_{Spearman}$), se H_0 é verdadeira, tem distribuição:

- **a) Pequenas amostras e sem/poucos empates:** exata com os valores apresentados em uma tabela.
- **b) Grandes amostras ou muitos empates:** aproximada pela Normal Padrão (Normal com média 0 e variância 1), tal que:

$$w_p = \frac{z_p}{\sqrt{(n-1)}}$$

no qual

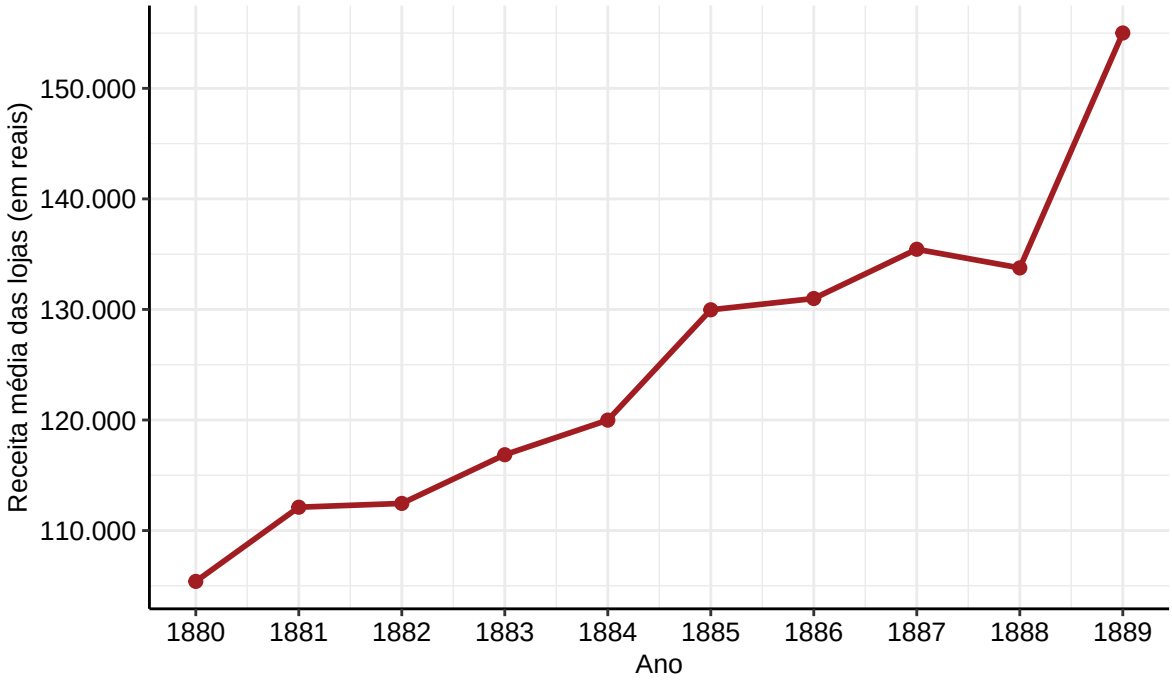
- $w_p =$ quantil de ordem p da distribuição que a estatística T segue
- $z_p =$ quantil de ordem p da distribuição Normal padrão
- $n =$ número total de observações na amostra

3 Análises

3.1 Receita média das lojas registrada nos anos de 1880 até 1889

Essa análise tem como objetivo visualizar a evolução da receita média total das lojas, em reais, na região do faroeste nos últimos 10 anos (1880-1889). Uma abordagem simples e objetiva foi adotada, onde foi calculada a média, por ano, de faturamento das lojas. As variáveis utilizadas foram data, id (produto, loja, venda), quantidade de produtos vendidos e preço. Além disso, foi utilizada uma taxa de conversão de 5,31 reais para 1 dólar.

Figura 4: Receita média das lojas por ano (em reais)



A **Figura 4** evidencia que houve um constante aumento na receita média das lojas, indo de R$105.399,00 no primeiro ano até R$155.009,10 no último ano (com uma pequena queda entre 1887 e 1888, de R$135.444,80 para R$133.757,60). Portanto, ao longo dos 10 anos, a receita média aumentou R$49.610,10. Assim, um aumento médio de, aproximadamente, R$5.000,00.

3.2 Variação Peso por Altura

O intuito dessa análise é entender a relação entre o peso (em quilogramas) e altura (em centímetros) dos clientes. A partir dela, concluiremos se a medida que o peso aumenta: a altura também aumenta, o contrário ou se não tem diferença. O nível de significância utilizado será de 5%.

Foi escolhido o teste de correlação de Pearson, já que o coeficiente de correlação linear de Pearson indica a força e a direção do relacionamento linear entre duas variáveis quantitativas. As hipóteses do teste são:

$$\begin{cases} H_0 : \text{Não há correlação linear entre altura e peso} \\ H_1 : \text{Há correlação linear entre altura e peso} \end{cases}$$

Ao realizar o teste, foram encontradas as seguintes conclusões:

Figura 5: Gráfico de dispersão do peso pela altura

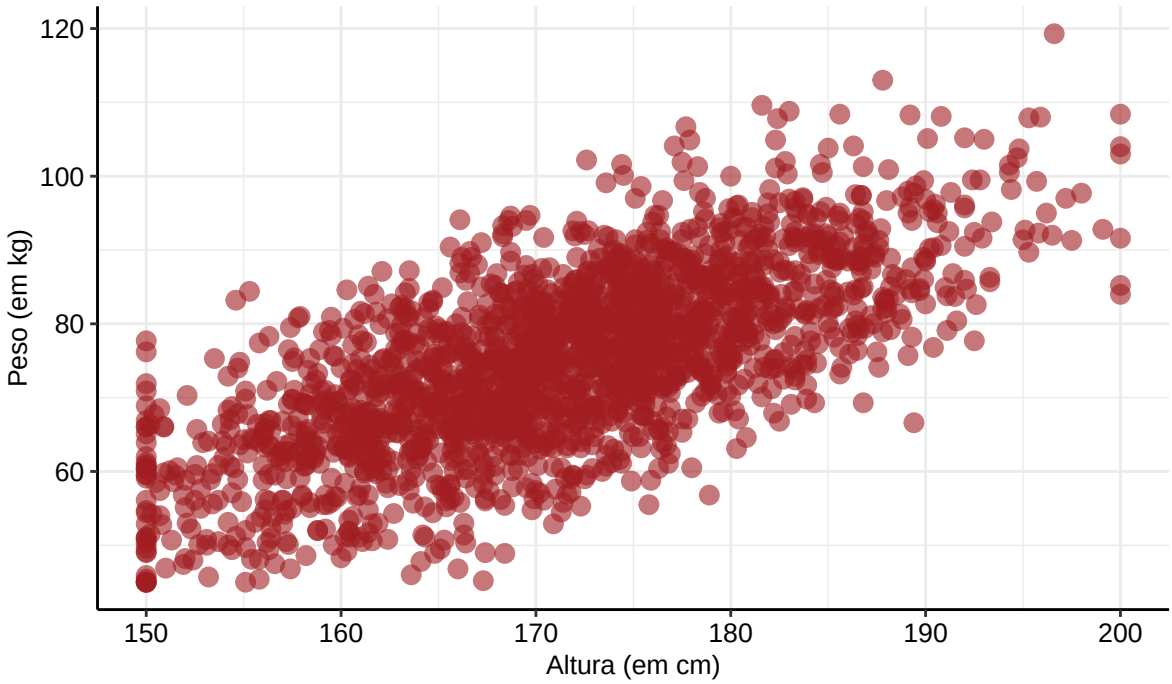


Tabela 1

Tabela 1 – Teste de correlação de Pearson entre Peso e Altura (Adaptado)

Variáveis	Coeficiente (p)	P-valor	Decisão do teste
Peso e Altura	0,70	< 0,0001	Rejeita H0

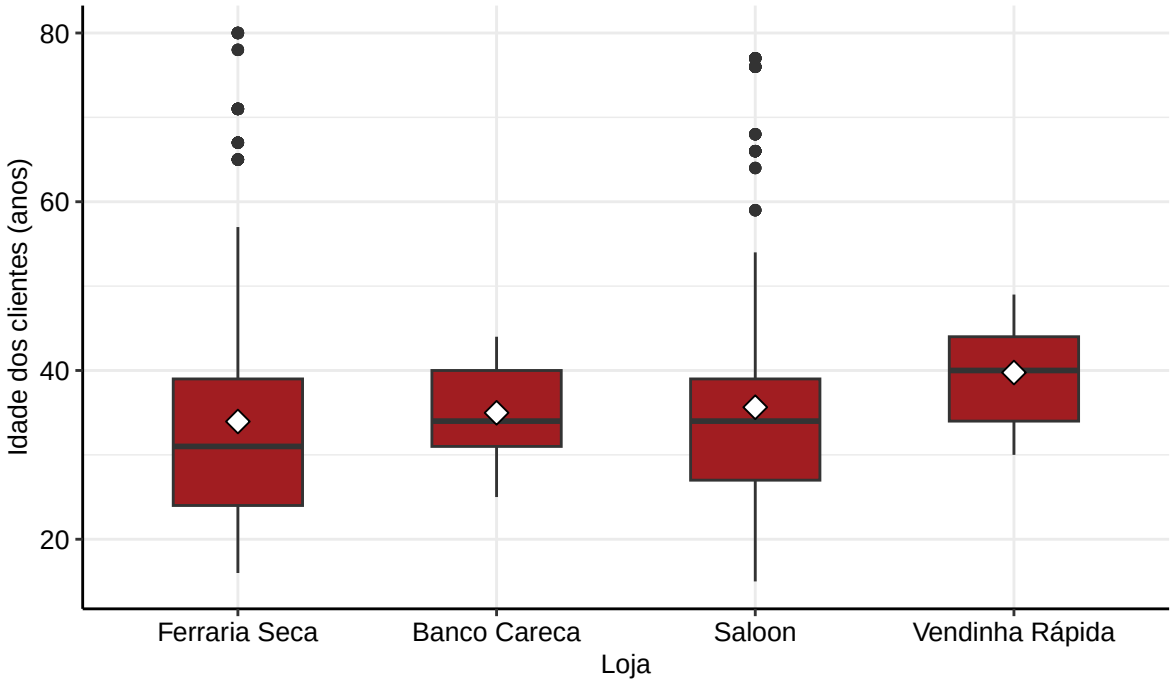
A **Figura 5** apresentada acima mostra a relação entre altura (em centímetros) e peso (em quilogramas) dos clientes. Cada ponto representa um indivíduo, permitindo visualizar a distribuição conjunta das duas variáveis. Observa-se uma relação positiva clara: à medida que a altura aumenta, o peso também tende a ser maior. Já a **Tabela 1** mostra os resultados: a correlação de Pearson indicou que há evidência estatisticamente significativa de que existe uma correlação linear entre peso e altura.($\rho \approx 0.70$).

Em resumo, a **Figura 5** e a **Tabela 1** sugerem fortemente que clientes mais altos tendem a ter maior peso.

3.3 Idade dos clientes de Âmbar Seco a depender da loja

Com o intuito de entender melhor o perfil das idades dos clientes nas diferentes lojas da cidade de Âmbar Seco, serão descritas as características dentro do banco de dados e mostrados quais são os perfis das idades dos clientes para cada loja da cidade.

Figura 6: Boxplot da idade dos clientes por loja



Quadro 1: Vendinha Rápida

Estatística	Valor
Média	39,76
Desvio Padrão	6,05
Variância	36,64
Mínimo	30,00
1º Quartil	34,00
Mediana	40,00
3º Quartil	44,00
Máximo	49,00

Quadro 2: Saloon

Estatística	Valor
Média	35,64
Desvio Padrão	12,00
Variância	144,12
Mínimo	15,00
1º Quartil	27,00
Mediana	34,00
3º Quartil	39,00
Máximo	77,00

Quadro 3: Ferraria Seca

Estatística	Valor
Média	33,95
Desvio Padrão	13,51
Variância	182,54
Mínimo	16,00
1º Quartil	24,00
Mediana	31,00
3º Quartil	39,00
Máximo	80,00

Quadro 4: Banco Careca

Estatística	Valor
Média	34,99
Desvio Padrão	5,47
Variância	29,91
Mínimo	25,00
1º Quartil	31,00
Mediana	34,00
3º Quartil	40,00
Máximo	44,00

Os boxplots apresentados no **Figura 6** ilustram a distribuição das idades dos clientes nas quatro lojas de Âmbar Seco: Ferraria Seca, Banco Careca, Saloon e Vendinha Rápida. Observa-se que as idades dos clientes variam entre as lojas, tanto em termos de centralidade quanto de dispersão.

De forma geral, observa-se que as idades dos clientes variam consideravelmente entre as lojas, com medianas situadas entre 31 e 40 anos. A seguir, destacam-se os principais pontos de cada loja:

3.3.1 Ferraria Seca

Apresenta a maior amplitude de idades entre as lojas, com valores que variam dos 15 aos 80 anos. O grande número de outliers indica a presença de clientes mais velhos que o público típico da loja. A mediana é de 31 anos, sugerindo um público relativamente jovem, mas com presença notável de clientes bem mais velhos.

3.3.2 Banco Careca

Possui uma distribuição mais concentrada, com faixa etária entre 25 e 45 anos. Tanto o desvio-padrão quanto a variância são menores que nas demais lojas, refletindo uma clientela mais homogênea em termos de idade. A média e a mediana próximas indicam simetria na distribuição.

3.3.3 Saloon

Mostra uma dispersão intermediária, com idades variando de cerca de 15 a 77 anos. A mediana e a média sugerem que a maior parte dos clientes tem aproximadamente 34 anos e o desvio-padrão relativamente alto reflete a presença de clientes bem mais jovens ou mais velhos do que o perfil típico.

3.3.4 Vendinha Rápida

A loja apresenta a maior mediana, indicando que seu público é relativamente mais maduro. A dispersão das idades é pequena, sugerindo um perfil de clientes mais homogêneo, com idades concentradas principalmente entre 34 e 44 anos (1º e 3º quartis). A faixa total vai de 30 a 49 anos, sem valores extremos muito distantes.

3.3.5 Resumo final

Em termos de tendência central, a média das idades acompanha de perto as medianas, o que indica distribuições relativamente simétricas, exceto na Ferraria Seca, onde há uma leve assimetria à direita devido à presença dOS clientes mais velhos.

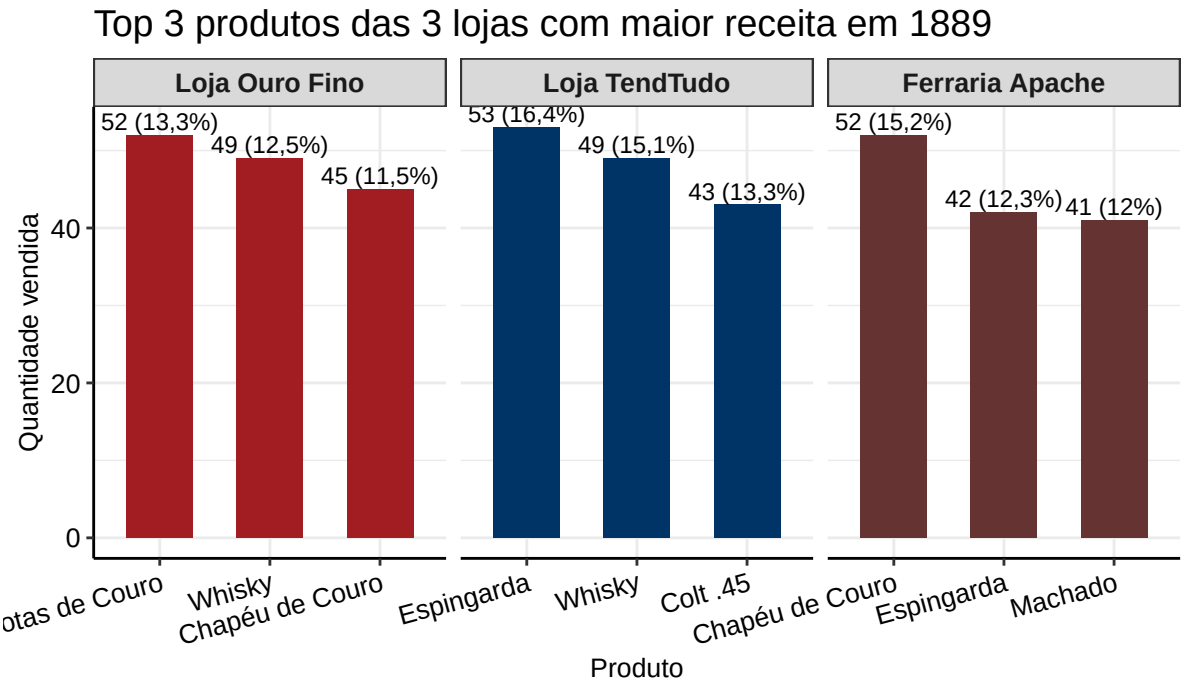
No conjunto, os resultados sugerem que o público das lojas de Âmbar Seco é majoritariamente adulto jovem a meia-idade e que há diferenças no perfil etário conforme a loja, com a Vendinha Rápida atendendo clientes mais velhos e a Ferraria Seca apresentando maior heterogeneidade;

Essas variações podem refletir diferentes tipos de produtos, localizações ou estratégias comerciais de cada estabelecimento.

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

Essa análise tem o objetivo de encontrar e visualizar quais são os 3 produtos mais vendidos nas 3 lojas que tiveram a maior receita no ano de 1889. Dessa forma, entender quais foram os produtos, a quantidade vendida e as lojas que mais venderam neste ano.

Figura 7: Gráfico de barras bivariado com frequências



A **Figura 7** mostra os 3 produtos mais vendidos das 3 lojas com maior receita, em ordem (da maior loja para menor, com os produtos mais vendidos para os menos).Os únicos produtos que se repetem são o whisky (Loja Ouro Fino e Loja TendTudo), o chapéu de couro (Loja Ouro Fino e Ferraria Apache) e a espingarda (Loja TendTudo e Ferraria Apache). Porém, todas as lojas tem pelo menos um produto em comum com as outras, o que sugere uma popularidade desses produtos no Faroeste, independentemente da loja.

A porcentagem em cima de cada coluna é a frequência relativa daquele produto em relação à receita total de cada loja. É interessante observar que em nenhuma dessas lojas os 3 produtos mais vendidos somam mais de 45% da receita. Ou seja, há uma variedade de produtos e vendas, não dependendo de algum específico. Esse fato é positivo para as lojas, pois não dependem de um produto único e conseguem captar o capital de forma variada e dispersa.

(AINDA QUERO COLOCAR UMA TABELA COM AS RECEITAS DESSAS TOP3 LOJAS A TITULO DE CURIOSIDADE)

4 Conclusões

O objetivo desse projeto era entender melhor o cenário do comércio no Faroeste para estudá-lo e entender dores, possibilidades de expansão, tendências, etc. Compreender características e extrair informaçõe relevantes para o negócio, baseando-se em dados de todo tipo e relacionados à diferentes pontos (receita média, características físicas, idade, produtos).

Em relação à evolução do mercado, as análises de receita média das lojas indicam um crescimento constante ao longo da última década (1880-1889). Houve um aumento médio de aproximadamente R$ 5.000,00 na receita média anual, sugerindo um mercado em expansão e saudável e com possibilidades de investimentos.

No que tange ao perfil do consumidor, a análise da correlação entre peso e altura demonstrou haver uma relação positiva estatisticamente significativa. Isso implica que clientes mais altos tendem a ter um peso maior, uma informação útil para estratégias relacionadas a produtos que dependem de dimensões físicas, podendo direcionar melhor os produtos para cada região (ou até cada loja).

RESUMO DA ANÁLISE 3 VAI FICAR PENDENTE PORQUE AINDA ESTÁ ERRADO

Por fim, ao analisar os produtos de maior sucesso nas três lojas com maior receita em 1889, verificou-se que produtos como chapéu de couro, whisky e espingarda são populares em todo o Faroeste, pois se repetem entre as líderes de vendas. No entanto, é um ponto positivo que nenhuma das lojas dependa excessivamente de um único item, já que o trio de produtos mais vendidos não somou mais de 45% da receita em nenhum dos casos. Isso demonstra uma variedade saudável de vendas e uma entrada de capital de forma dispersa.