Modelo Projeto - Quarto

Consultores Responsáveis:
Laura Mello

Requerente:
João Sábio, Old Town Road.Ltda

Brasília, 9 de novembro de 2025.

ESTAT
consultoria estatística

Sumário

1 Introdução

O seguinte projeto tem como objetivo fornecer uma visão detalhada sobre o desempenho das lojas da região entre os anos de 1880 e 1889, por meio de análises estatísticas descritivas. Foi avaliado como as vendas das lojas evoluíram ao longo do tempo, o comportamento de vendas de produtos e características do perfil de clientes, com ênfase em identificar quais itens impulsionam o faturamento em cada loja.

O banco de dados utilizado foi fornecido pela Old Town Road Ltda. e contém registros completos das vendas realizadas na região nos últimos dez anos, incluindo informações detalhadas sobre produtos, quantidades vendidas, receita por item, além de características dos clientes, como idade, peso e altura. Com essa abrangência, a base permite analisar com precisão tanto padrões de consumo quanto o impacto financeiro de cada produto em diferentes lojas.

As análises foram realizadas utilizando o software do R.Studio, que permitiu a manipulação e visualização dos dados, realização de cálculos estatísticos e geração de tabelas e gráficos de forma clra e confiável.

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.

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.

2.4.1 Variância Populacional

Para uma população, a variância é dada por:

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

Com:

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

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.

2.5.1 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:

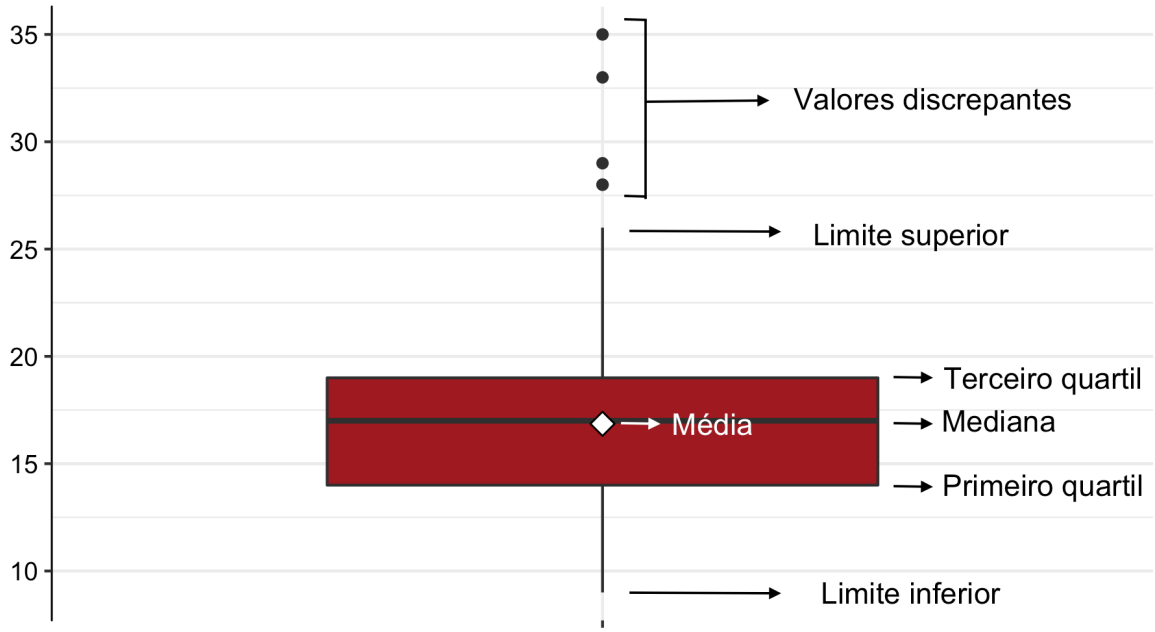
- $X_i =$ i-ésima observação da população

- $\mu =$ média populacional
- $N =$ tamanho da população

2.6 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

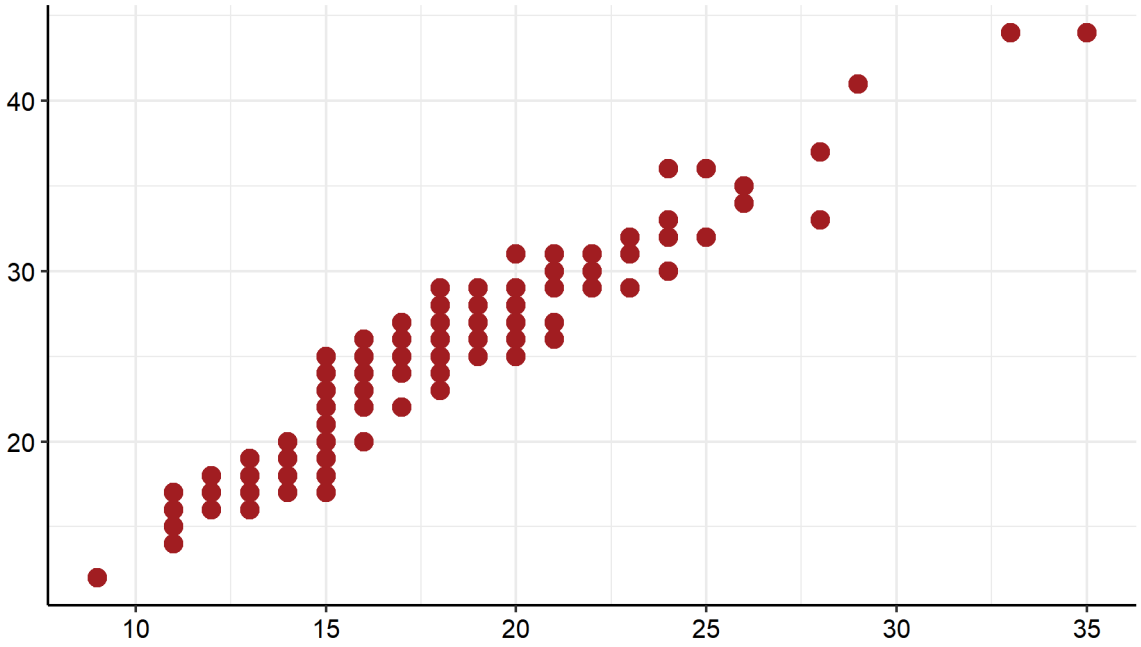


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada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

2.7 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 2: 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 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:

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

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.

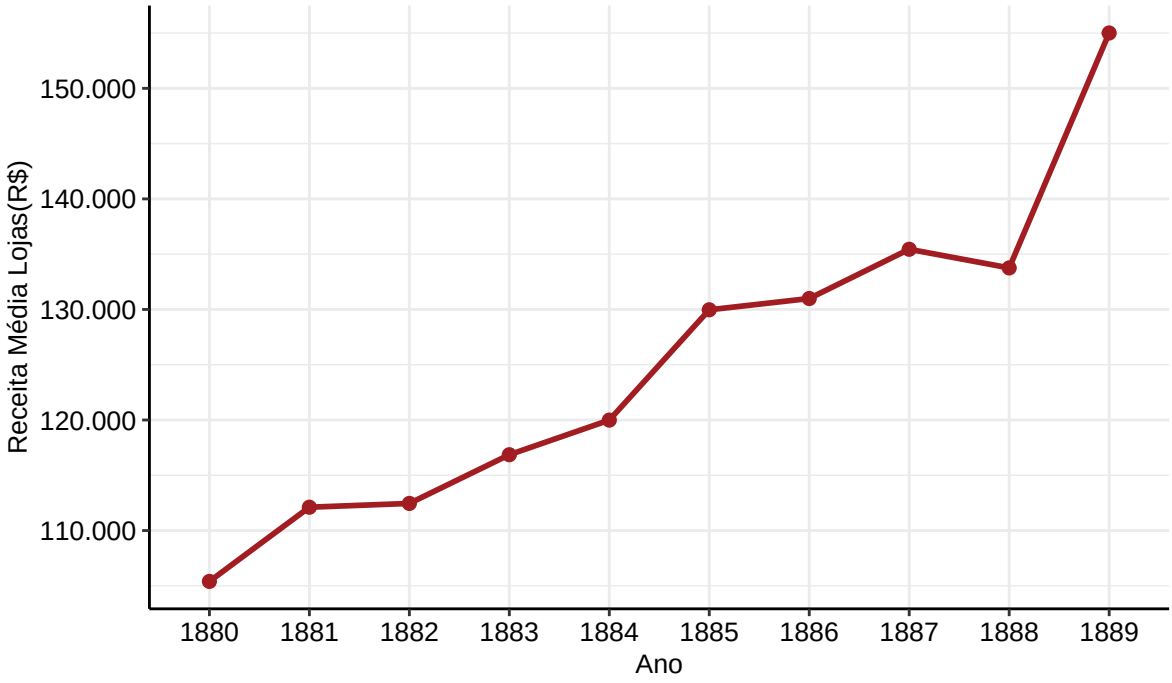
3 Análises

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

Esta análise tem como objetivo compreender a variação da receita média das lojas da região no período entre 1880 e 1889. Para isso, foram consideradas duas variáveis principais: “ano”, que é quantitativa discreta e “ReceitaMedia”, que é quantitativa contínua, expressa em reais.

Os valores monetários foram convertidos de dollar para real (1 dollar = 5,31 reais) a fim de padronizar a análise financeira e facilitar a interpretação dos resultados.

Figura 3: Gráfico de linhas da receita média da lojas ao decorrer dos últimos dez anos

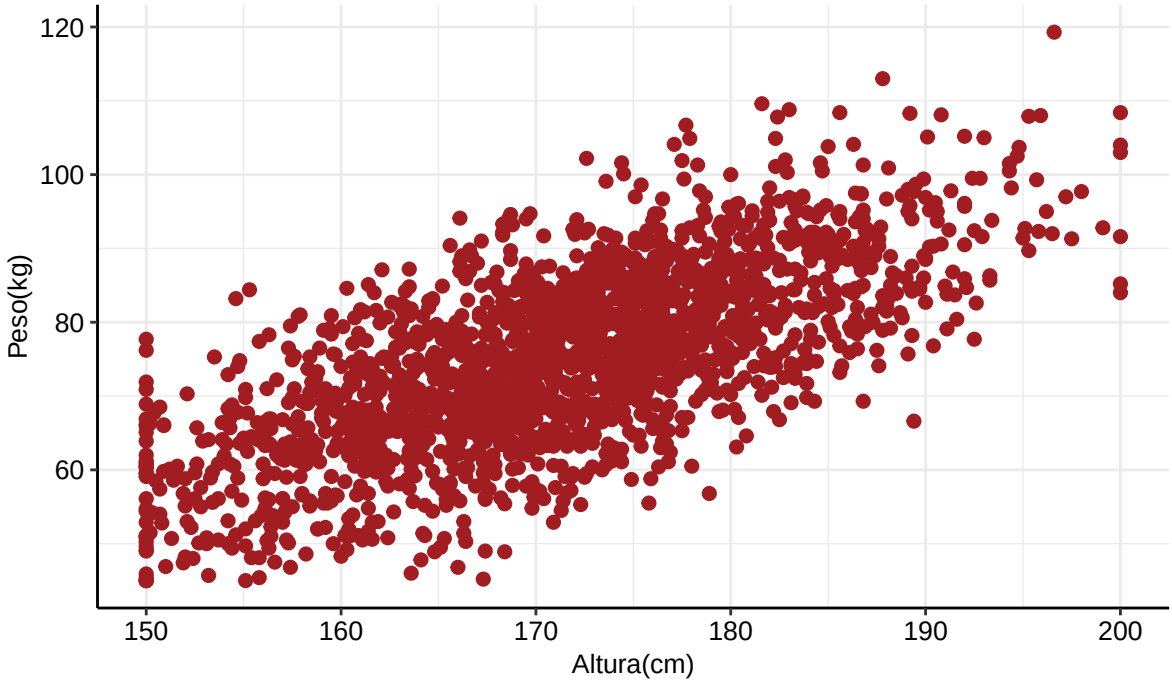


Como mostra a **Figura 3**, a receita média das lojas vem crescendo desde 1880, com exceção de 1888, ano em que houve uma pequena queda de R$ 1.687,20 em relação ao período anterior. Por outro lado, o crescimento mais significativo ocorreu em 1889, com um aumento de R$ 21.251,50 na receita média das lojas da região.

3.2 Variação Peso por Altura

Nesta análise busca-se compreender a relação entre a altura dos indivíduos, registrada originalmente em decímetros (dm) e convertida para centímetros (cm), e o peso, registrado em libras (lbs) e convertido para quilogramas (kg). Ambas as variáveis são quantitativas contínuas. O objetivo é verificar se há associação entre as medidas, avaliando se indivíduos mais altos tendem a apresentar maior peso.

Figura 4: Gráfico de dispersão do peso (kg) em função da altura (cm) dos indivíduo



Quadro 1: Medidas resumo do 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

Medidas resumo do peso (kg)

Quadro 2: Medidas resumo da 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

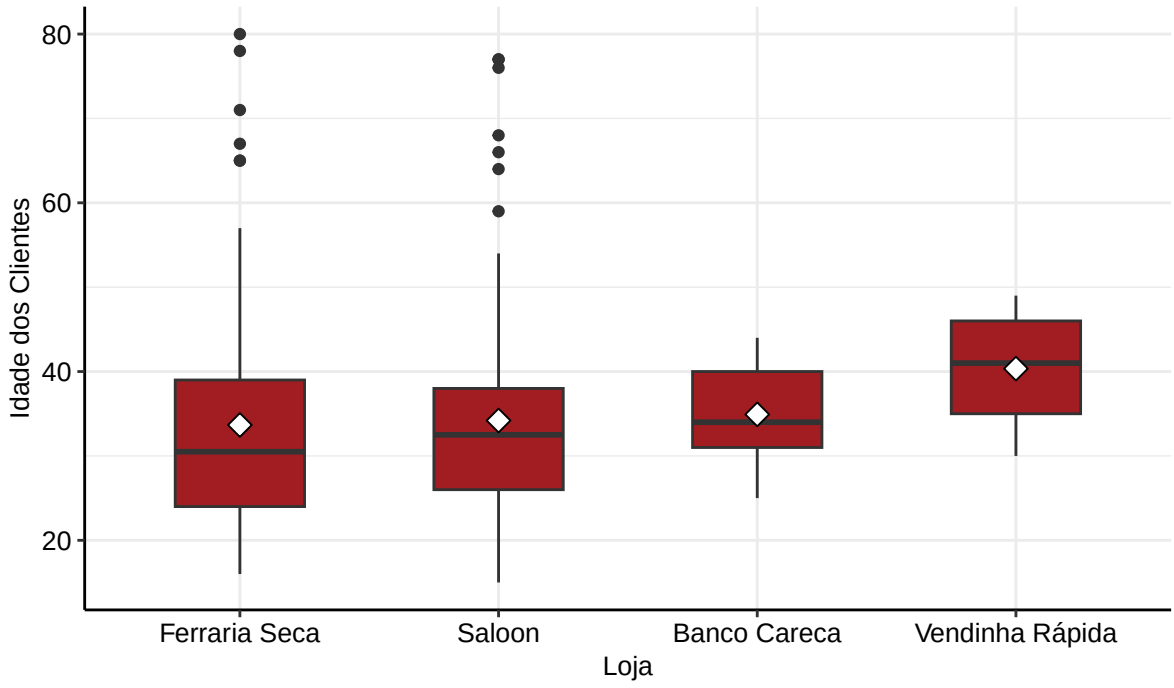
Medidas resumo do altura (cm)

A Figura **Figura 4** evidencia que indivíduos mais altos tendem a apresentar valores de peso maiores, embora exista variação considerável dentro de cada faixa de altura. As observações analisadas variam entre 150 cm e 200 cm para altura e entre 45 kg e 115 kg para peso. O **Quadro 1** e o **Quadro 2** permitem identificar que a maior concentração de indivíduos ocorre entre 170 cm e 180 cm, com pesos situados predominantemente entre 60 kg e 85 kg, caracterizando a região mais densa da distribuição. O coeficiente de correlação de Pearson calculado entre peso e altura é de aproximadamente $r \approx 0{,}75$, indicando uma correlação linear positiva forte. Isso significa que, em geral, quanto maior a altura do indivíduo, maior tende a ser o peso. Apesar dessa tendência, a dispersão observada indica que indivíduos com a mesma altura podem apresentar pesos bastante distintos, sugerindo a influência de outros fatores sobre a variável peso.

3.3 Idade dos clientes de Âmbar Seco a depender da loja

Esta análise tem como objetivo investigar o perfil etário dos clientes das lojas localizadas na cidade de Âmbar Seco. Para isso, foram consideradas as variáveis Idade dos Clientes uma variável quantitativa contínua e Loja, variável qualitativa nominal que identifica cada estabelecimento.

Figura 5: Boxplot das idades dos clientes por loja em Âmbar Seco



Quadro 3: Medidas resumo da idade dos clientes por loja

Estatística	Banco Careca	Ferraria Seca	Saloon	Vendinha Rápida
Média	34,92	33,67	34,20	40,35
Desvio Padrão	5,57	13,31	12,70	6,03
Variância	31,06	177,18	161,23	36,39
Mínimo	25,00	16,00	15,00	30,00
1o Quartil	31,00	24,00	26,00	35,00
Mediana	34,00	30,50	32,50	41,00
3o Quartil	40,00	39,00	38,00	46,00
Máximo	44,00	80,00	77,00	49,00

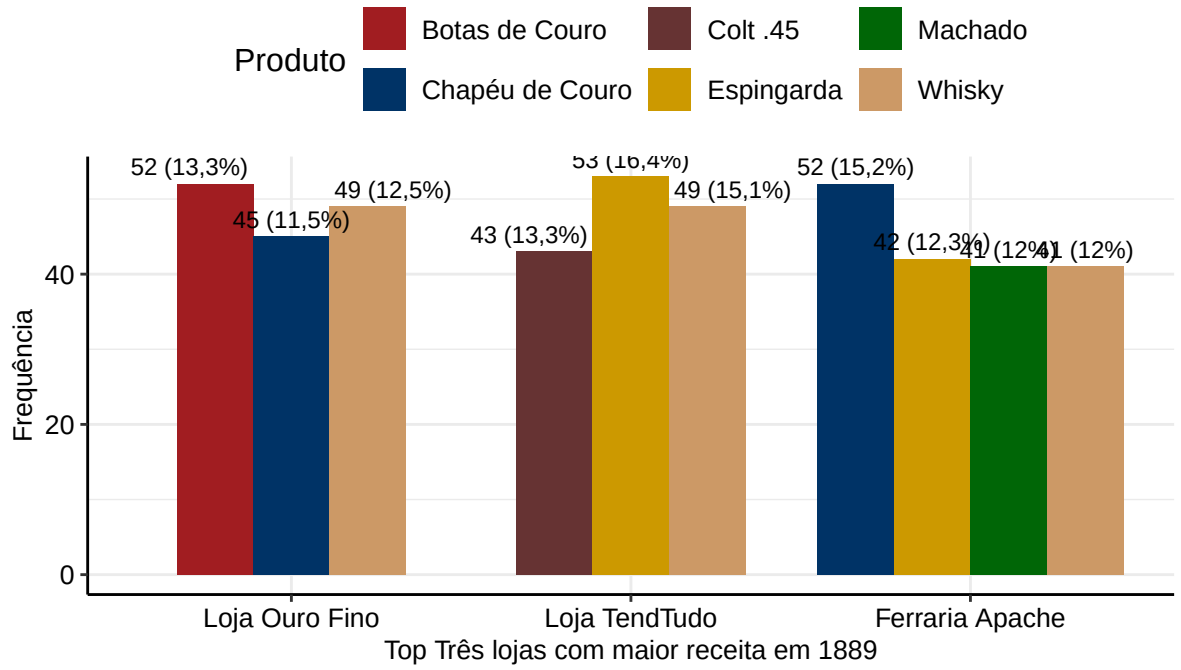
Medidas resumo das Idades dos Clientes por Loja

Conforme evidenciado pelo **Quadro 3** e na **Figura 5**, a distribuição das idades varia significativamente entre as lojas analisadas. A “Vendinha Rápida” apresenta a maior média (40,35 anos), seguida pelo “Banco Careca” (34,92), “Saloon” (34,2) e “Ferraria Seca” (33,67), indicando uma ordem crescente clara das médias. As idades na “Vendinha Rápida” e no “Banco Careca” são relativamente homogêneas, com desvios padrão de 6,03 e 5,57, concentrando-se entre 35 e 46 anos e 31 e 40 anos, respectivamente, enquanto “Saloon” e “Ferraria Seca” apresentam maior variabilidade (12,7 e 13,31), com quartis mais dispersos (1º quartil 26 e 24, 3º quartil 38 e 39) e extremos que vão de 15 a 77 anos e de 16 a 80 anos. Esses dados evidenciam que algumas lojas atendem faixas etárias mais homogêneas e outras apresentam clientes com idades bastante diversificadas.

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

Esta análise tem como objetivo identificar os três produtos mais vendidos nas lojas que apresentaram as maiores receitas em 1889. Para isso, foram consideradas as variáveis Loja e Produto (qualitativas nominais) e Quantidade Vendida (quantitativa discreta). A partir da frequência de vendas de cada produto, buscou-se compreender o comportamento de consumo nas lojas de maior desempenho, permitindo observar quais itens se destacam e como o mix de produtos varia entre os estabelecimentos. Essa análise contribui para identificar preferências do público e orientar decisões estratégicas relacionadas ao foco comercial.

Figura 6: Gráfico de colunas da frenquência dos produtos mais vendidos pelas lojas coma maior receita no ano de 1889



Conforme evidenciado pela Figura **Figura 6**, que apresenta os três produtos mais vendidos nas lojas com maior receita em 1889, observa-se que cada estabelecimento possui um padrão distinto de preferência de produtos. Na Loja Ouro Fino, o produto com maior frequência foi “Botas de Couro” (52 unidades), seguido de “Whisky” (49) e “Chapéu de Couro” (45), indicando predominância de itens relacionados a vestimenta e acessórios. Já na Loja TendTudo, o destaque é a “Espingarda” (53 unidades), que supera os demais produtos, enquanto “Whisky” (49) e “Colt .45” (43) aparecem em seguida, caracterizando um perfil de vendas mais orientado a armamentos. Por fim, na Ferraria Apache, a distribuição é mais equilibrada: “Chapéu de Couro” lidera com 52 unidades, seguido por “Espingarda” (42) e por “Machado” e “Whisky”, ambos com 41 unidades, sugerindo ausência de um único produto dominante. Essa variação evidencia que, embora as lojas possuam alta receita, cada estabelecimento tem um perfil de consumo específico.

4 Conclusões

O projeto desenvolvido para a Old Town Road Ltda., tem como objetivo compreender o funcionamento do mercado na região em que a empresa pretende investir. As quatro análises realizadas possibilitaram uma visão ampla sobre o comportamento dos clientes, o desempenho das lojas e os produtos mais relevantes para o faturamento.

A primeira análise abordou a evolução da receita média das lojas entre 1880 e 1889, revelando uma tendência geral de crescimento ao longo do período, com destaque para 1889, ano de maior expansão. Essa conclusão indica um cenário de mercado favorável e crescimento consistente, o que reforça o potencial econômico da região e apoia a decisão de João em investir.

Na segunda análise, que explorou a relação entre altura e peso dos clientes, foi observada uma tendência de associação entre as duas variáveis, indivíduos mais altos tendem a ter maior peso, embora exista uma grande variação. Esse resultado, ainda que descritivo, reforça a importância de conhecer o perfil físico médio dos clientes, o que pode auxiliar no planejamento de produtos específicos e em decisões sobre o público-alvo em potenciais investimentos.

A terceira análise, sobre a distribuição das idades dos clientes por loja, mostrou que há diferenças marcantes no perfil etário do público atendido. Algumas lojas concentram clientes em faixas etárias mais homogêneas, o que sugere um público-alvo bem definido, enquanto outras apresentam maior diversidade de idades, indicando um atendimento mais amplo. Esses resultados ajudam a entender quais lojas devem fazer campanhas mais segmentadas e quais podem se beneficiar de estratégias mais generalistas.

Por fim, a quarta análise avaliou os produtos mais vendidos nas três lojas com maior receita em 1889. Observou-se que cada estabelecimento apresenta um mix de vendas distinto, com algumas lojas concentradas em categorias específicas e outras com distribuição mais equilibrada entre os produtos. Esses resultados permitem identificar quais itens realmente atraem os clientes em cada loja e reforçam que estratégias de venda devem considerar o perfil de consumo local. Assim, a escolha de produtos a serem priorizados no investimento não deve ser uniforme, mas adaptada ao comportamento observado em cada contexto comercial.