Projeto Fantasma

Consultores Responsáveis:
Estatiano 1
Estatiano 2
Estatiano 3

Requerente:
ESTAT

Brasília, 31 de outubro de 2025.

ESTAT
consultoria estatística

Sumário

tlmgr install multibib

1 Introdução

Este relatório estatístico foi elaborado para a Old Town Road.Ltda, uma holding sob a liderança de João Sábio, com o objetivo de apoiar sua estratégia de expansão e investimento em uma nova região do comércio no faroeste. O objetivo geral deste trabalho é analisar estatisticamente diversos aspectos do mercado-alvo, visando identificar padrões de consumo e características demográficas relevantes. Especificamente, o relatório apresentará quatro análises estatísticas distintas focadas na receita média das lojas no período de 10 anos, a relação entre a altura e o peso dos clientes, a idade dos clientes das lojas da cidade de Âmbar Seco e os produtos mais vendidos nas lojas que mais faturam. A relevância deste estudo reside na sua capacidade de mitigar riscos de investimento, otimizar a oferta de produtos e serviços e, consequentemente, maximizar o retorno da Old Town Road.Ltda na nova região. Os dados utilizados para as análises foram obtidos a partir do arquivo fornecido pelo cliente (relatorio_old_town_road.xlsx).

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 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.

2.3 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. ### 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

2.3.1 Variância Amostral

Para uma amostra, a variância é dada por:

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

Com:

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

2.4 Desvio Padrão

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

2.4.1 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

2.4.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

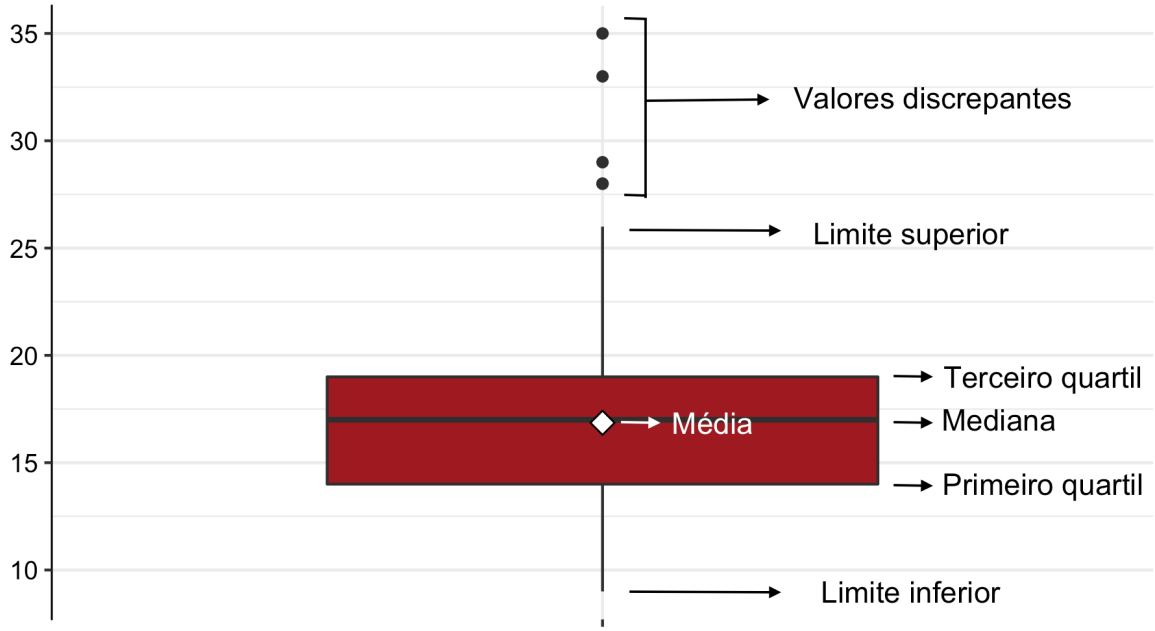
$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

Com:

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

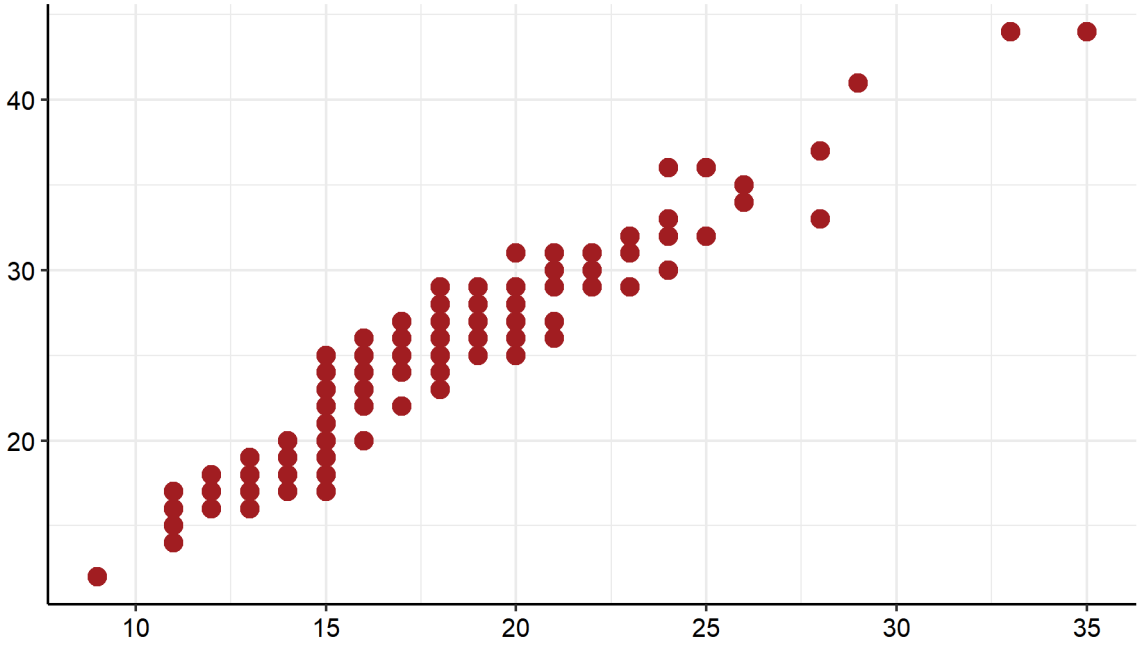


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada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

2.6 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

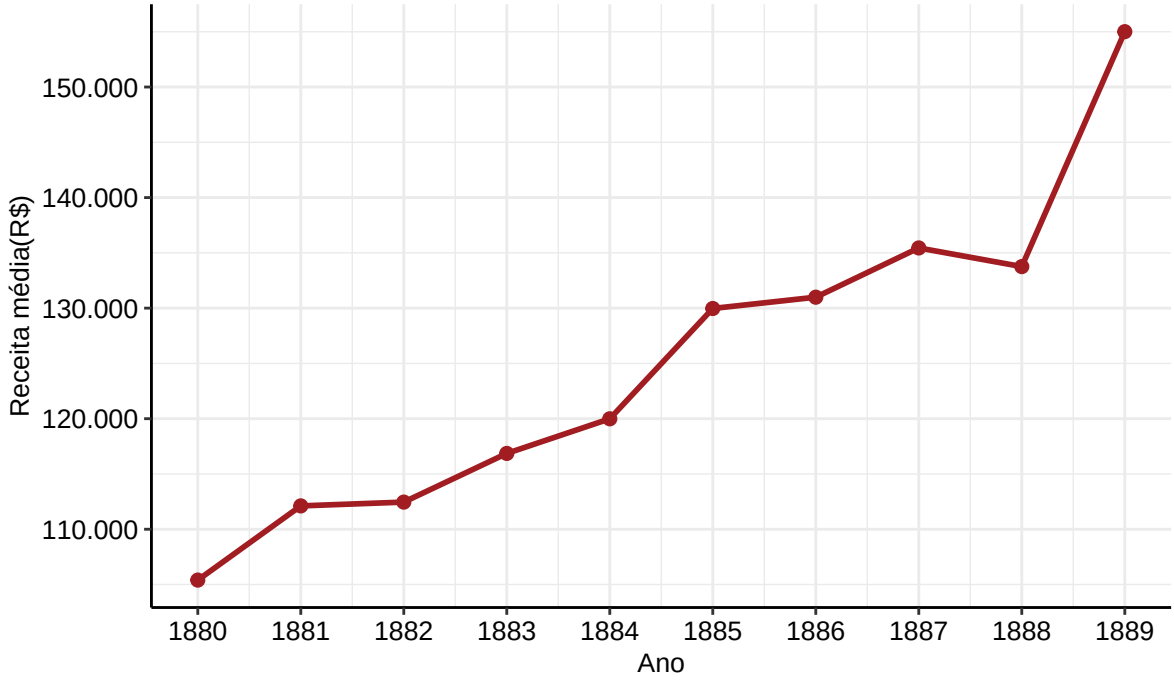
Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

3 Análises

3.1 Análise da receita média das lojas entre os anos de 1880 até 1889

O objetivo dessa análise é entender o comportamento da receita média dos anos ao longo dos anos, se há ou não um crescimento ou decrescimento durante o período de tempo. Para essa análise, transformamos os dados dos valores das ventas que estavam em dólares, para reais, multiplicando-os por 5,31. Além disso, os dados foram obtidos por meio da junção das tabelas “infos_vendas”, “infos_produtos” e “relatorio_vendas”. Como ano é uma varíavel quantitativa discreta, uma vez que está definido em um certo intervalo de tempo, e a receita média é uma variável quantitativa discreta, já que depende de uma medição, nesse caso o real, é feito um gráfico de linhas univariado em função dessas duas variáveis. Também há uma tabela com o ano e a receita total média das lojas, a fim de esclareccer os valores.

Figura 3: Receita total média das lojas ao longo dos anos(1880-1889)

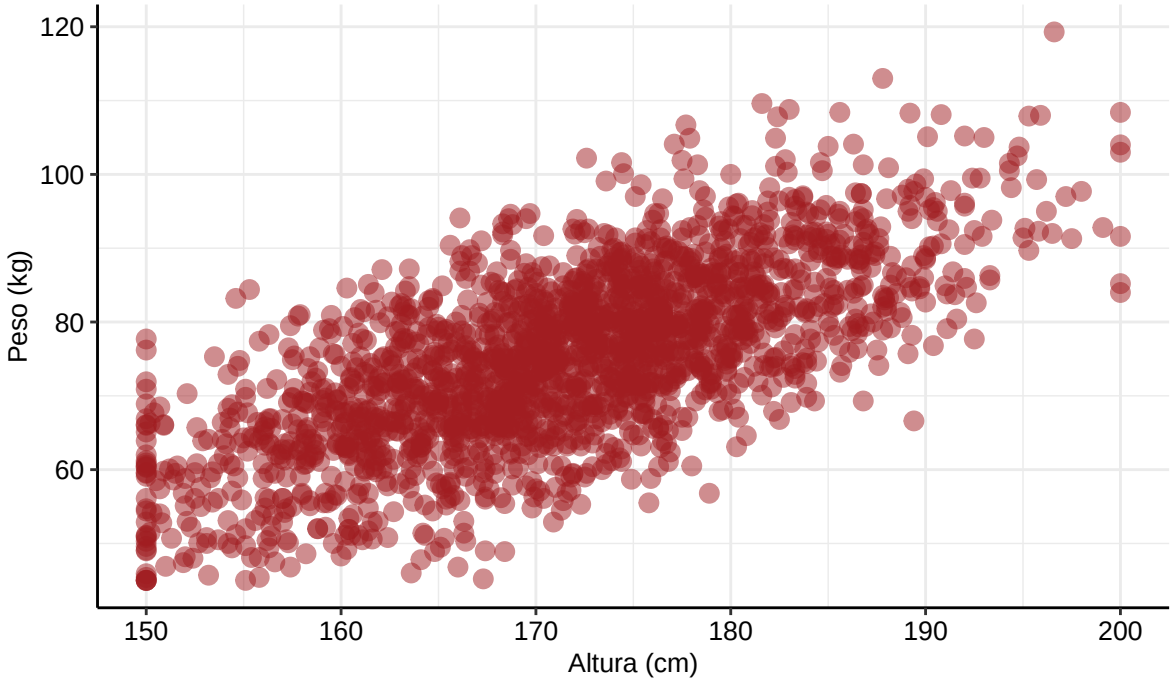


A partir do gráfico, podemos deduzir que a receita média das lojas vêm crescendo ao longo dos anos em um ritmo quase que constante, o valor da receita média só tem uma leve queda no ano de 1888, de 1.687,00 reais, e logo após essas quedas a receita volta a crescer. Em especial, houve um grande crescimento entre os anos de 1888 e 1889, onde a receita média creceu de 21.251,00 reais. Pode-se observar que houve um aumento de 147,06% da receita total média em dez anos, e a espectativa para os próximos anos é que essa continue a crescer, se seguir os mesmos padrões.

3.2 Análise da variação de peso pela altura dos clientes

Para analisar a relação entre as duas variáveis quantitativas contínuas , peso e altura, é feito um gráfico de dispersão. O objetivo é compreender como uma variável influencia a outra e se existe uma correlação linear, ou seja, se elas crescem ou decrescem juntas.Não foi realizado nenhum agrupamento de dados para essa análise.

Figura 4: Gráfico de Dispersão do Peso Pela Altura



A partir do gráfico, observa-se uma tendência de correlação positiva, pois conforme o valor da altura cresce o valor do peso também cresce, assim as variáveis se movem na mesma direção. Para verificação da correlação e sua força, foi calculado o coeficiente de correlação de Pearson, que verifica o grau de relação linear entre as duas variáveis quantitativas, obtendo-se o valor de 0,6971. Isso significa que, realmente, há uma correlação positiva e direta entre as variáveis, além disso, indica que a força de relação linear é de moderada a forte.

3.3 Análise dos perfis das idades dos clientes para cada loja da cidade de Âmbar Seco.

Com o objetivo de entender o perfil da idade dos clientes de Âmbar Seco por loja, foi realizada uma análise descritivas. Os dados foram obtidos por meio da junção das tabelas “infos_vendas”, “infos_produtos”, “infos_clientes”, “infos_lojas”, “infos_cidades”. Adicionalmente, o banco de dados foi filtrado para incluir apenas clientes de Âmbar Seco e ajustado para considerar somente clientes únicos, garantindo que a análise

do perfil etário não fosse enviesada pelo volume de transações. Além disso, como se tratam de uma variável quantitativa discreta (Idade) e uma variável qualitativa nominal (Nome da Loja), foi feito um diagrama de caixas bivariado (Figura 3), em função do número da Loja, e uma tabela de medidas resumo (Tabela 2), contendo: média, mediana, quartis, desvio-padrão.

Figura 5: Boxplot da idade (anos) pelo nome da loja de Âmbar Seco

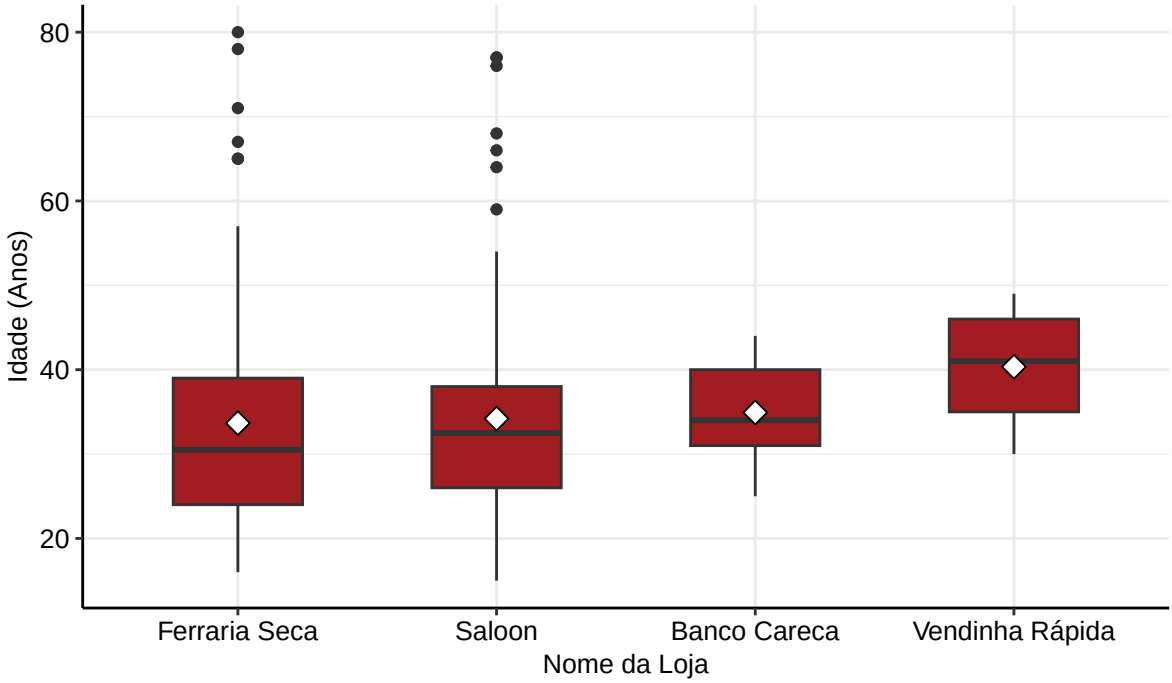


Tabela 1

Quadro 1: Medidas resumo da idade dos clientes por loja

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

O diagrama foi construido, de modo que, está ordenado pela ordem crescente da média, que é representada pelo losango branco. A Loja Ferraria Seca possui a menor

média de idade das lojas, 33,67, e o segundo menor mínimo, 16 anos, além disso, metade das pessoas que vão a essa Loja têm entre 24 e 39 anos, e possui o maior limite superior, além de vários outliers. Já a Loja Saloon, possui a segunda menor média entre as lojas, 34,2, e o menor mínimo, 15 anos, metade de seus consumidores tem entre 26 e 38 anos e apresenta valores átipicos maiores do que o seu limite superior. A Loja Banco Careca não possui valores discrepantes, e 50% dos seus clientes tem entre 31 e 40 anos, seu limite infeirior é 25 anos, o que já é maior que a média da Loja Ferraria Seca, e seu máximo é 44 anos, diferente das Lojas Saloon e Ferraria Seca que possuem máximos maiores e outliers, isso indica que a idade é mais próxima entre os clientes que a frequentam, o que é confirmado pelo seu desvio-padrão baixo, de 5 anos, o menor desvio padrão, além de ter a menor variância, 31,06. Por fim, a Loja Vendinha Rápida, possui a maior média, 40,35 anos de idade, e assim como a Loja Saloon, não possui outliers e sua assimetria é muito menor que nas Lojas Ferraria Seca e Saloon, e metade de seus clientes tem entre 35 e 46 anos. Portanto, as Lojas Banco Careca e Vendinha Rápida , com menores desvios padrão, 5,47 e 6,03, respectivamente, demonstram um público alvo etário mais homogêneo e definido, concentrado entre aproximadamente 30 e 45 anos. Em contraste, as Lojas Ferraria Seca e Saloon, com altos desvios-padrão, 12,7 e 13,31, indicam uma grande heterogeneidade etária devido à presença de valores atípicos e maior dispersão.

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

O objetivo dessa análise é compreender quais são os produtos mais vendidos nas três lojas com maior receita no ano de 1889. Para isso, juntamos as tabelas “relatorio_vendas”, “infos_produtos”, “infos_clientes”, “infos_lojas” e “infos_cidades”, a fim de organizar uma só tabela que contenha todos os dados necessários para a analíse como o nome da loja, o nome do produto, a quantidade de produtos vendidos, o valor dos produtos e o ano. A partir do valor dos produtos, da quantidade, o ano e do nome da loja conseguimos obter quanto cada loja faturou no ano de 1889, como o preo estava em dólar, ele foi multiplicado por 5,31, considerando a cotação do dólar à 5,31 reais. A seguinte tabela mostra na ordem decrescente o nome da loja e sua receita.

A partir da tabela, podemos notar que as três lojas com maior faturamento do maior para o menos foram: Loja Ouro Fino, Loja TendTudo e Ferraria Apache, com a receita de 197.313,00 reias, 196.340,00 reais e 181.689,00 reais, respectivamente. Então, são elas que foram analizadas. A seguinte tabela mostra em ordem decrescente os produtos e a quantidade de itens que foram comprados na Loja Ouro Fino em 1889.

Diante disso, observa-se que os produtos mais vendidos na Loja Ouro Fino foram: Botas de Couro, Whisky e Chapéu de Couro.

A próxima tabela mostra em ordem decrescente os produtos e a quantidade de itens que foram vendidos na Loja TendTudo em 1889.

Sendo assim, os três produtos mais vendidos na Loja TendTudo foram a Espingarda, o Whisky e o Colt. 45.

A seguinte tabela mostra em ordem decrescente os produtos e a quantidade de itens que foram vendidos na Ferraria Apache em 1889.

Como resultado, é observado que os produtos mais vendidos na Ferraria Apache foram: Chapéu de Couro, Espingarda e Machado.

Prtanto, pode-se observar que os únicos dois itens que apareceram em duas lojas com as maiores receitas foram o Chapéu de Couro e o Whisky, fora esses nenhum outro item se repetiu, o que indica que não há muita relação entre o item os mais vendidos e o faturamento da loja.

4 Conclusões