

## Sumário

	<b>Página</b>
1    Introdução . . . . .	3
2    Referencial Teórico . . . . .	4
2.1    Frequência Relativa . . . . .	4
2.2    Média . . . . .	4
2.3    Mediana . . . . .	4
2.4    Quartis . . . . .	5
2.5    Variância . . . . .	5
2.5.1    Variância Populacional . . . . .	5
2.5.2    Variância Amostral . . . . .	6
2.6    Desvio Padrão . . . . .	6
2.6.1    Desvio Padrão Populacional . . . . .	6
2.6.2    Desvio Padrão Amostral . . . . .	7
2.7    Coeficiente de Variação . . . . .	7
2.8    Boxplot . . . . .	7
2.9    Histograma . . . . .	8
2.10    Gráfico de Dispersão . . . . .	9
2.11    Tipos de Variáveis . . . . .	10
2.11.1    Qualitativas . . . . .	10
2.11.2    Quantitativas . . . . .	10
2.12    Coeficiente de Correlação de Pearson . . . . .	11
2.13    Teste de Hipóteses . . . . .	11
2.14    Tipos de teste: bilateral e unilateral . . . . .	12
2.15    Nível de significância ( $\alpha$ ) . . . . .	12
2.16    Estatística do Teste . . . . .	13
2.17    P-valor . . . . .	13
2.18    Intervalo de Confiança . . . . .	13
3    Teste de Normalidade . . . . .	14
3.1    Teste de Normalidade de Lilliefors . . . . .	14
3.2    Teste de Correlação de Pearson . . . . .	14
4    Método . . . . .	16
5    Análises . . . . .	17
5.1    Análise 1 . . . . .	17
5.2    Análise dos produtos (Extra) . . . . .	21
5.3    Análise das cidades (Extra) . . . . .	23
5.4    Análise 2 . . . . .	24
5.5    Análise do IMC (Extra) . . . . .	29

5.6	Análise 3 . . . . .	32
5.7	Análise 4 . . . . .	33
6	Conclusão . . . . .	35
7	Anexo . . . . .	36

## 1 Introdução

João Sábio, proprietário da Old Town Road.Ltda, uma holding que controla diversas empresas ligadas ao comércio no faroeste, nos contratou para realizar análises estatísticas que visam compreender melhor o mercado em certa região em que ele tem o interesse de investir. Nesse relatório serão abordadas as análises de interesse do cliente: a receita média das lojas registrada nos anos de 1880 até 1889, variação Peso por Altura, idade dos clientes de Âmbar Seco a depender da loja e o top 3 produtos mais vendidos nas top 3 lojas com maior receita em 1889.

## 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:

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

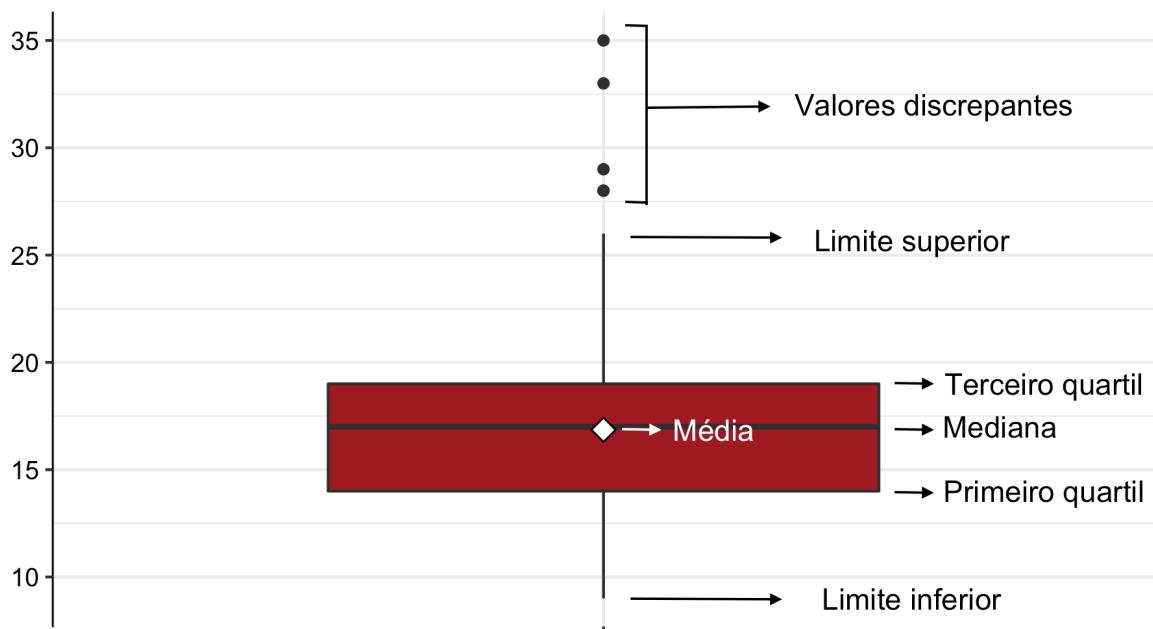
Com:

- $S$  = 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

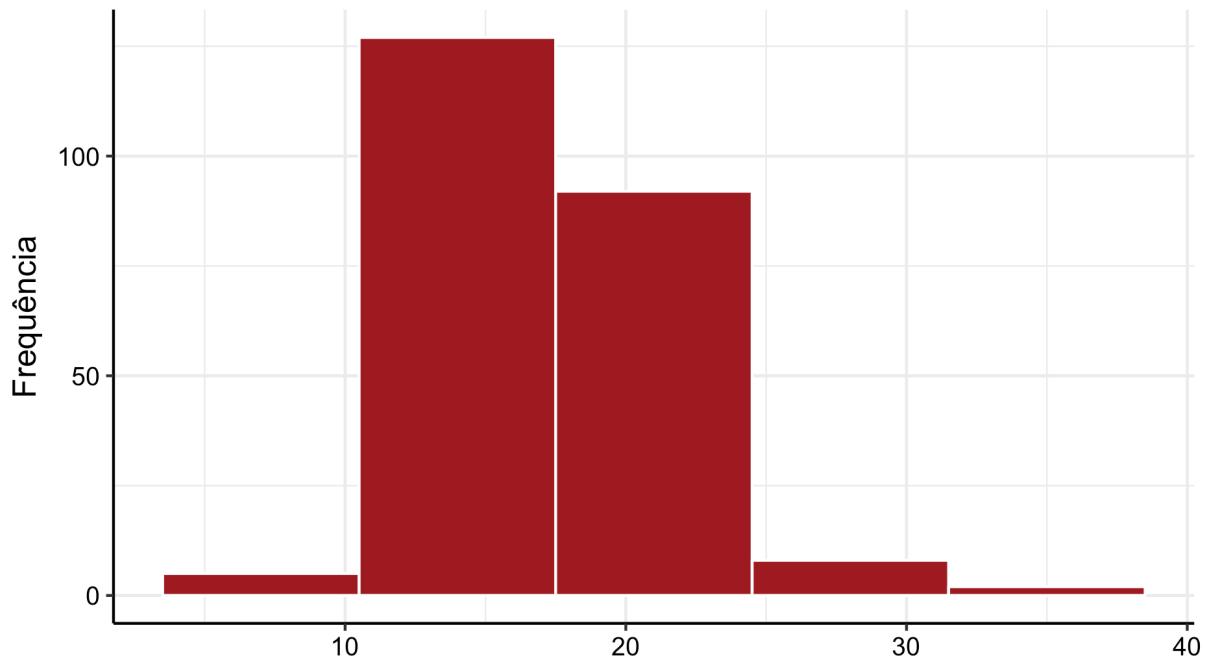


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada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

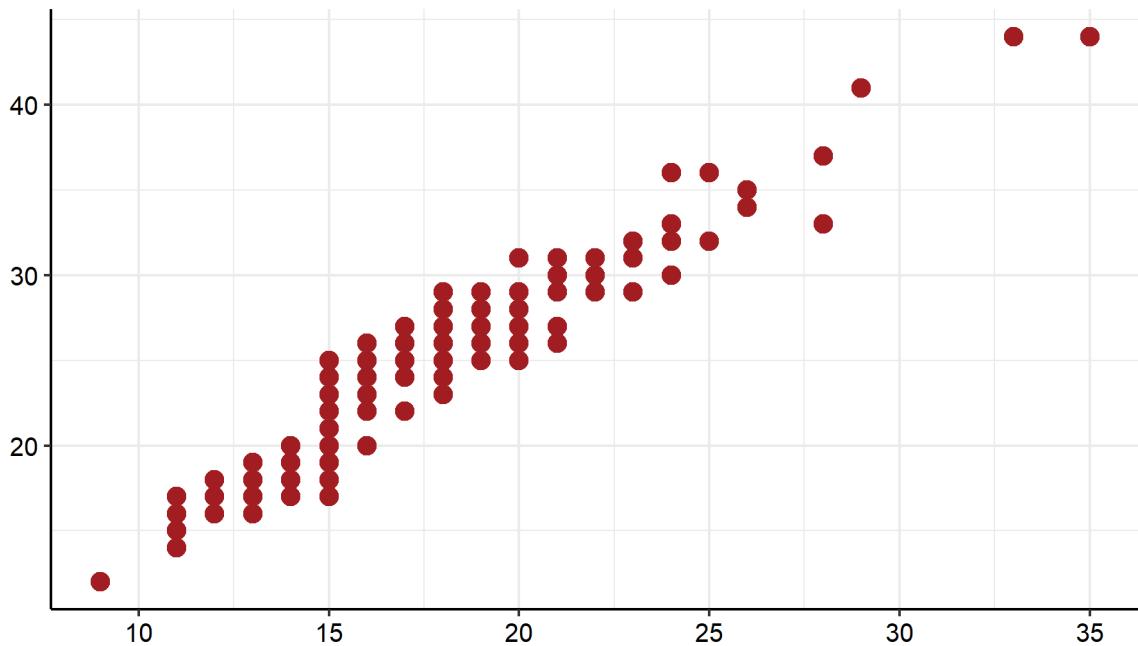
Figura 2: Exemplo de histograma



## 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  $\alpha$ , 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  $\beta$ .

## 2.15 Nível de significância ( $\alpha$ )

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  $\alpha$  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  $\theta$  e que a distribuição amostral de  $T$  é conhecida. O intervalo de confiança para o parâmetro  $\theta$  será dado por  $t_1$  e  $t_2$ , tal que:

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

A probabilidade  $\gamma$  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  $\theta$ . 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 Lilliefors

Assim como o teste de Kolmogorov-Smirnov, o **teste de Lilliefors** é utilizado para verificar se um conjunto de dados  $X_1, X_2, \dots, X_n$  de tamanho  $n$  segue determinada distribuição. A estatística de teste para este teste é dada por:

$$T_1 = \sup_x |F^*(x) - S(X)|$$

A diferença entre a estatística  $T$  de Kolmogorov e  $T_1$  de Lilliefors é que a função de distribuição acumulada  $S_n(x)$  é obtida através dos dados padronizados da amostra, ou seja,  $Z_1, Z_2, \dots, Z_n$ , com  $Z_i = \frac{X_i - \bar{X}}{S}$ . O teste acima é recomendado para amostras grandes com presença de valores discrepantes (*outliers*).

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue o modelo proposto} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula é verdadeira, espera-se que as diferenças entre  $F_0(x)$  e  $S_n(x)$  sejam pequenas e estejam dentro dos limites dos erros aleatórios.

#### 3.2 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.

## 4 Método

Neste estudo, serão utilizados dados fornecidos pela ESTAT - Consultoria Estatística. Conferir Anexo 1 com o dicionário das variáveis em estudo.

## 5 Análises

O conjunto de dados consiste em 19.885 observações de compras feitas por 1990 **clientes** de 10 tipos de **produtos** em 18 **lojas** distintas nas 5 **cidades** da região de interesse.

### 5.1 Análise 1

O objetivo desta análise é compreender o comportamento das **lojas** ao longo da série temporal por meio da **receita anual** e da **receita média**. A receita anual corresponde à soma da quantidade vendida multiplicada pelo preço unitário de cada item comercializado em determinado ano, enquanto a receita média é a soma da receita anual dividido pela quantidade de lojas na região (18), ambas são variáveis **quantitativas contínuas**, por esse motivo foi utilizado gráficos de dispersão. Por outro lado, a loja é uma variável **qualitativa nominal**. Foram utilizados gráficos de dispersão para facilitar a visualização e a análise dos dados. Na Figura 1, o gráfico é apresentado em facetas e mostra a receita anual de cada loja, com o ano no eixo x e a receita total anual no eixo y. Já na Figura 2, o gráfico apresenta a receita média. A Tabela 1 apresenta a média, a mediana e o desvio-padrão da receita de cada loja em todo o período, enquanto a Tabela 2 mostra a receita média em reais ao longo dos anos de 1880 a 1889.

Figura 4: Receita anual das lojas

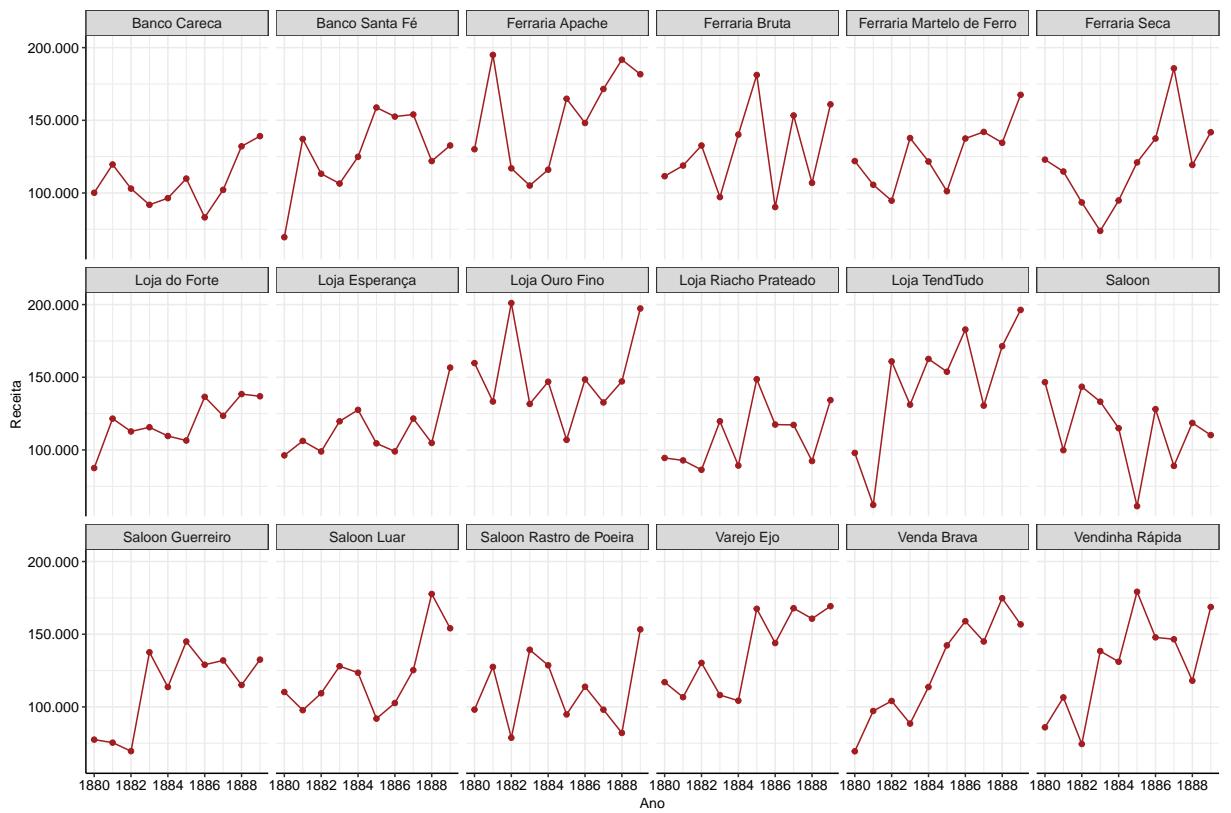


Tabela 1: Resumo da receita das lojas em todo o período

Loja	Média	Mediana	DP
Ferraria Apache	15211,02	6438,59	19009,86
Loja Ouro Fino	15052,68	6402,90	17799,87
Loja TendTudo	14496,37	5506,92	18169,25
Varejo Ejo	13756,91	5727,71	15719,38
Vendinha Rápida	12967,56	5204,73	16064,40
Ferraria Bruta	12931,75	5173,67	15770,30
Banco Santa Fé	12713,51	6438,59	13874,71
Ferraria Martelo de Ferro	12645,38	5519,74	14845,16
Venda Brava	12507,55	4857,38	15424,63
Saloon Luar	12205,33	5090,43	14329,95
Ferraria Seca	12052,96	4457,61	14572,85
Loja do Forte	11888,41	5351,92	12931,86
Saloon	11453,42	3922,36	13910,80
Loja Esperança	11352,05	4562,99	13116,79
Saloon Guerreiro	11272,70	3886,12	13556,47
Saloon Rastro de Poeira	11145,99	4343,31	12951,11
Loja Riacho Prateado	10926,29	4457,61	12936,53
Banco Careca	10776,12	4564,21	12315,83

As 18 lojas da região apresentam comportamentos diversos, sendo que a maioria registrou crescimento na receita anual ao final do período em relação ao primeiro ano da série. A loja **Saloon** destaca-se das demais, em que sua receita anual apresentou uma queda significativa durante o mesmo intervalo. As lojas que tiveram maior receita média nos 10 anos em estudo foram **Ferraria Apache**, **Loja Ouro Fino** e **Loja TendTudo**, já as que tiveram a menor receita média foram **Banco Careca**, **Loja Riacho Prateado** e **Saloon Rastro de Poeira**.

Figura 5: Receita média da região

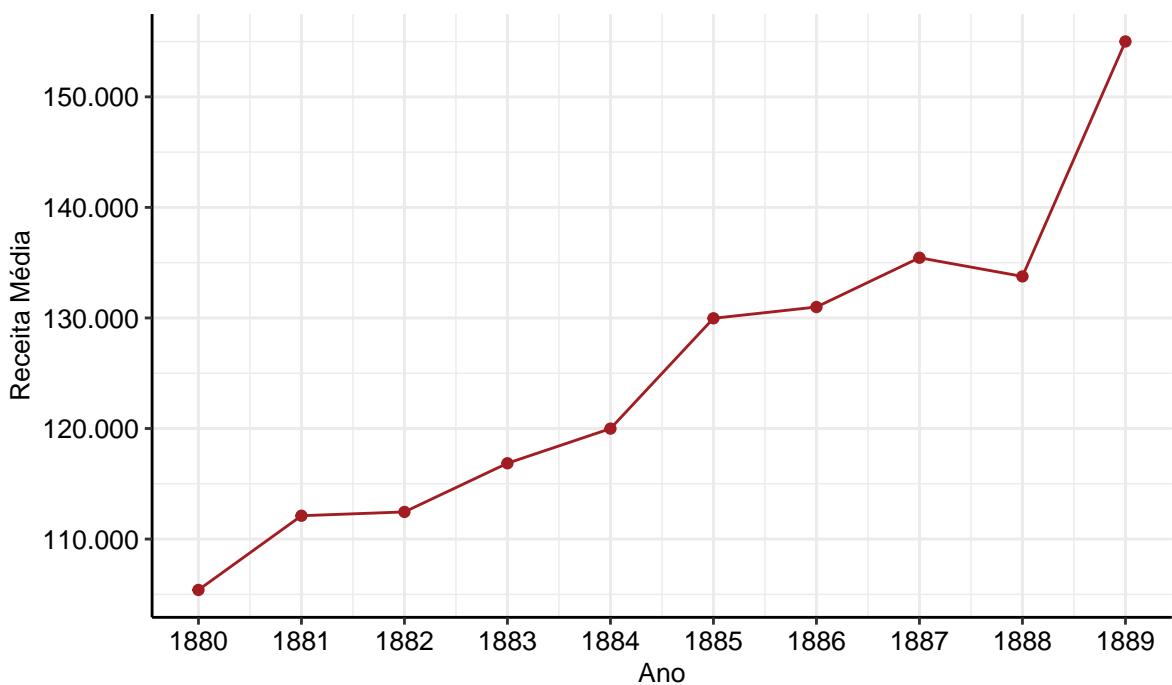


Tabela 2: Receita méida da região nos anos de 1880 a 1889

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. Esses resultados indicam que a região apresenta **grande potencial** de investimento.

## 5.2 Análise dos produtos (Extra)

O objetivo desta análise é examinar mais detalhadamente a **receita anual** e a **receita média por produto** na região, bem como a **quantidade vendida**, produtos é variável qualitativa nominal.

Figura 6: Receita anual por produto

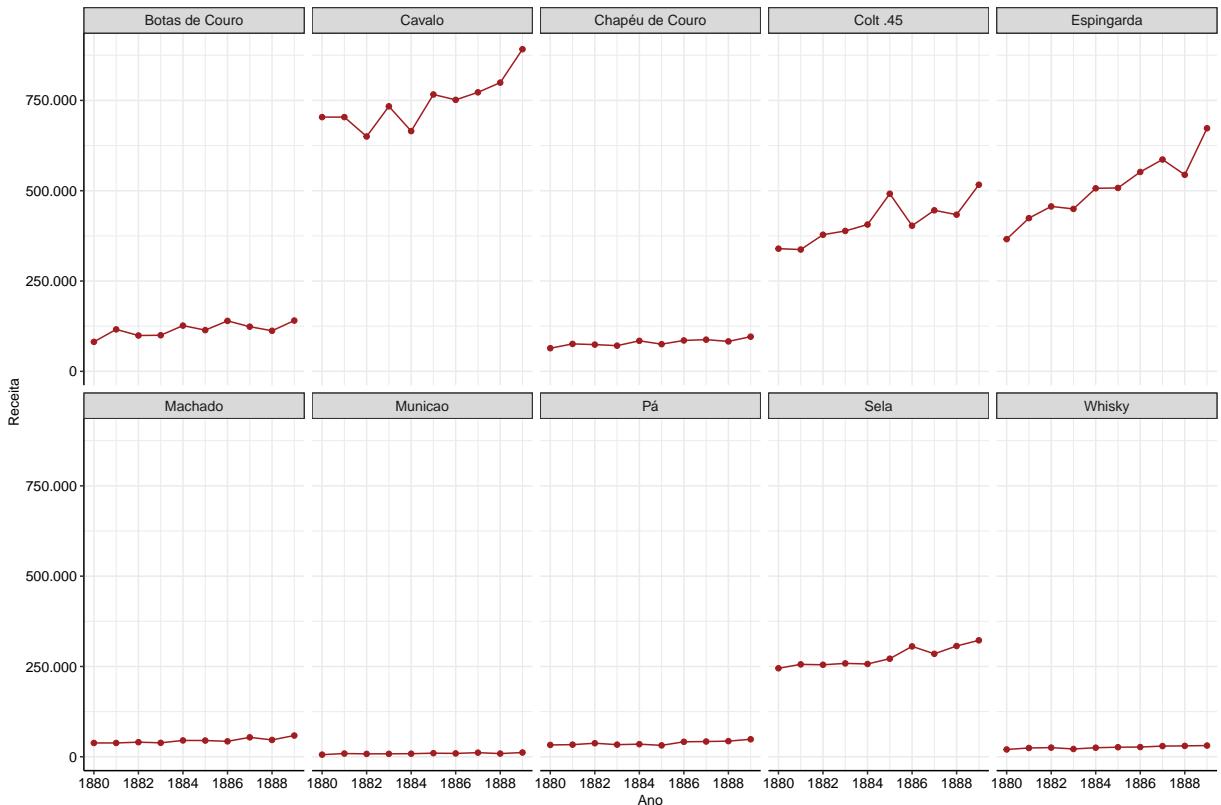


Tabela 3: Quantidade de produtos vendidos nos anos de 1880 a 1889

Produto	Quantidade
Chapéu de Couro	5414
Espingarda	5149
Colt .45	5146
Sela	5107
Pá	5058
Botas de Couro	5047
Municao	4862
Whisky	4838
Machado	4752
Cavallo	2494

Tabela 4: Resumo da receita dos produtos nos anos de 1880 a 1889

Produto	Média	Mediana	DP
Cavalo	41317,13	38765,87	14932,00
Espingarda	28143,20	28531,27	10984,38
Colt .45	23003,31	22529,48	9185,08
Sela	15344,36	14872,65	5486,53
Botas de Couro	6409,56	5943,48	2463,87
Chapéu de Couro	4427,24	4268,60	1677,21
Machado	2488,27	2450,56	996,16
Pá	2108,35	2025,82	810,09
Whisky	1450,05	1456,64	581,60
Municao	506,30	477,98	207,73

Dos 10 produtos disponíveis, o **Cavalo** apresentou a maior média de receita, mesmo sendo o menos vendido. Esse resultado se deve ao elevado preço unitário do produto, de R\$ 2.981,99. O **Chapéu de Couro** foi o item mais vendido na região, seguido pela **Espingarda** e pela **Colt .45**. É importante observar que a soma das quantidades vendidas das duas armas foi de 10.295 unidades, enquanto o total de **Munições** vendidas foi de apenas 4.862. Esse contraste pode indicar que os consumidores adquirem munições em outras regiões ou não sentem necessidade de comprá-las com frequência após a aquisição da arma. Além disso, nota-se que foram vendidas mais do que o dobro de **Selas** em relação ao número de **Cavalos**, o que pode sugerir que as **Selas** vendidas na região são de baixa qualidade ou mal cuidadas, dentre outras possibilidades.

### 5.3 Análise das cidades (Extra)

Nessa análise iremos aprofundar o entendimento de cada **cidade** observando a **receita anual** e a **receita média** assim como a **taxa de crescimento econômico**, que foi calculada pela diferença das receitas nos anos de 1889 e 1880, dividido pela receita em 1880, cidade é variável **qualitativa nominal** e taxa de crescimento econômico é variável **quantitativa contínua**.

Figura 7: Receita anual por cidade

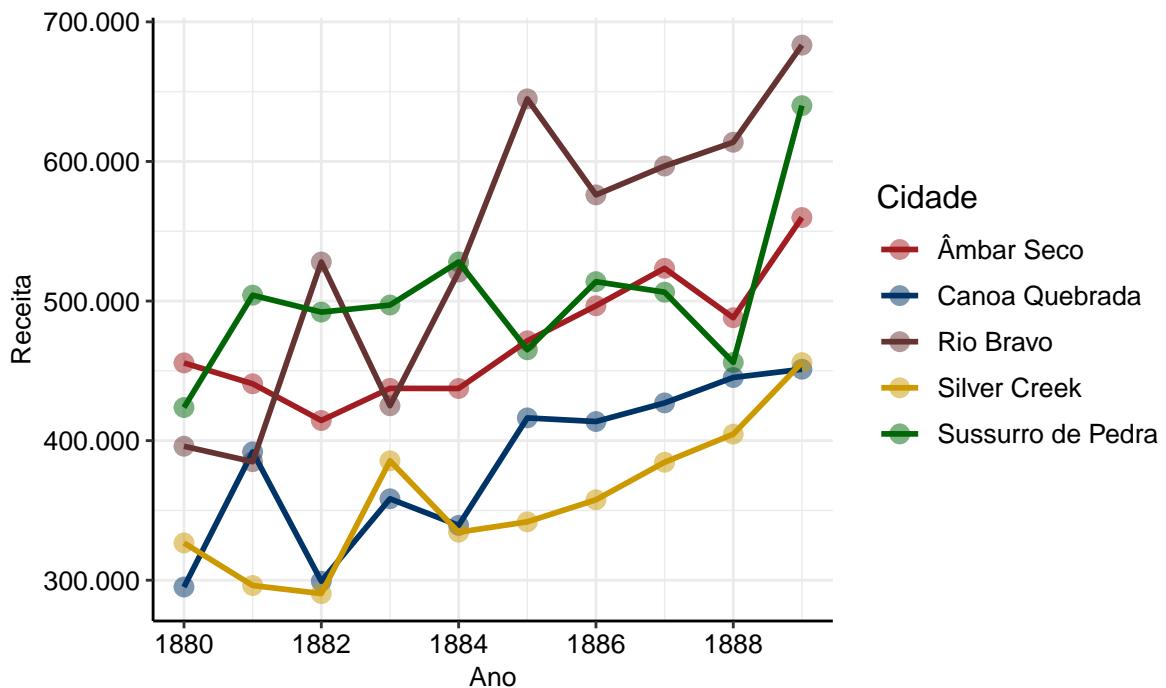


Tabela 5: Resumo da receita das cidades nos anos de 1880 a 1889

Cidade	Média	Mediana	DP
Rio Bravo	13423,15	5257,70	16264,93
Canoa Quebrada	12790,71	5204,73	15455,60
Sussurro de Pedra	12566,06	5486,29	14598,46
Silver Creek	11925,67	4869,64	14032,26
Âmbar Seco	11812,52	4379,56	14248,61

Tabela 6: Taxa de crescimento econômico (%) das cidades entre 1880 e 1889

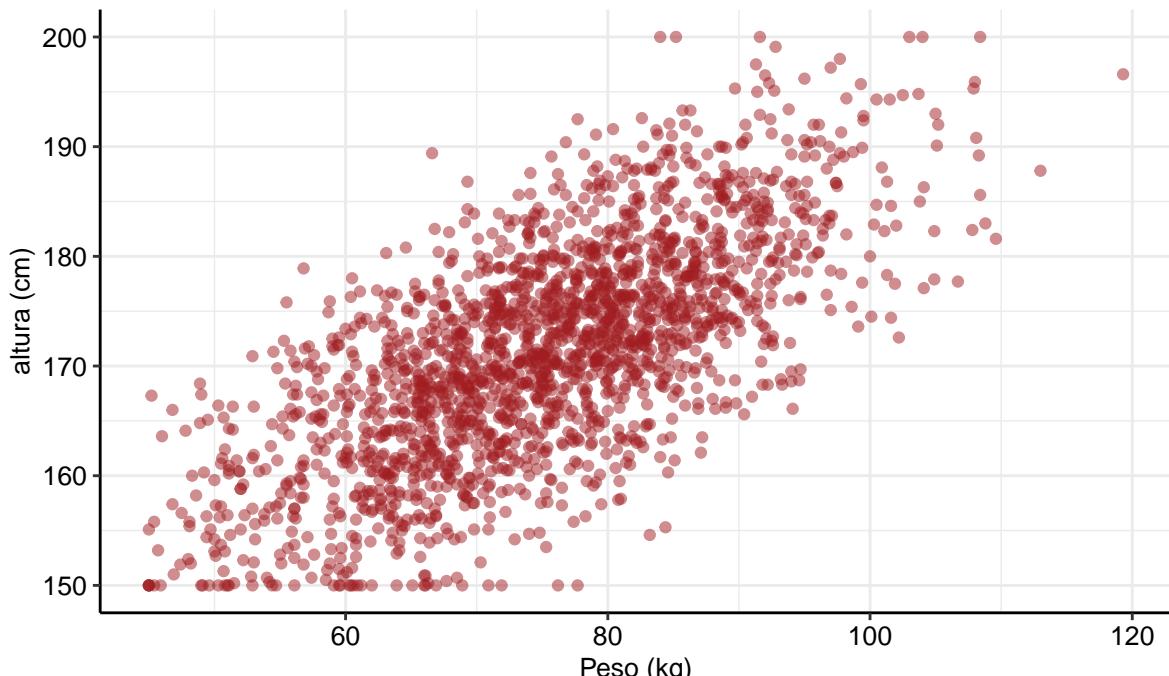
Cidade	Taxa de crescimento (%)
Rio Bravo	72,56
Canoa Quebrada	52,86
Sussurro de Pedra	51,04
Silver Creek	39,56
Âmbar Seco	22,86

A cidade de **Rio Bravo** apresentou o maior crescimento em receita anual ao longo do período analisado, destacando-se também por possuir o maior número de vendas (11290) e a maior receita média (R\$2.527,90). Observa-se que **Canoa Quebrada** realizou apenas 8.031 vendas nos 10 anos em análise, mas registrou a segunda maior receita média. Todas As cidades demonstram bom crescimento econômico e se mostram **promissoras** para investimentos futuros, porém as cidades com maior taxa de crescimento econômico foi **Rio Bravo, Canoa Quebrada e Sussuro de Pedra**.

## 5.4 Análise 2

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 8: 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.

Para realizar a análise, primeiro foram realizados dois testes de aderência de Lilliefors, a fim de verificar normalidade das variáveis **peso** e **altura**. As hipóteses nulas  $H_0$  consideram que o peso dos clientes provém de uma população normalmente distribuída com média igual a 75,18849 e variância igual a 141,9977, e que a altura dos clientes provém de uma população normalmente distribuída com média igual a 171,4808 e variância igual a 97,37976. As hipóteses alternativas  $H_1$  afirmam que os dados não seguem essas distribuições propostas. Os parâmetros de média e variância utilizados nessas distribuições correspondem às médias e variâncias amostrais da respectiva variável (ver Quadros 1 e 2).

$$\begin{cases} H_0 : \text{Peso} \sim \text{Normal}(\mu = 75, 18849; \sigma^2 = 141, 9977) \\ H_1 : \text{Peso segue outra distribuição} \end{cases}$$

$$\begin{cases} H_0 : \text{Altura} \sim \text{Normal}(\mu = 171, 4808; \sigma^2 = 97, 37976) \\ H_1 : \text{Altura segue outra distribuição} \end{cases}$$

Sob  $H_0$ , os p-valores encontrados foram 0,24463150 e 0,3278868, respectivamente, ao nível de significância de 5% não há evidência suficiente para rejeitar as hipóteses nulas. Sugerindo que ambos os dados seguem uma distribuição aproximadamente normal.

Tabela 7: Resultados dos testes

Variável	p-valor	Estatística D	Decisão
Peso	0,2446315	0,0160580	Não rejeitar H0
Altura	0,3278868	0,0151383	Não rejeitar H0

Além disso, a suposição de normalidade também foram avaliadas gráficamente por meio da comparação da densidade amostral com a densidade teórica, e a análise dos quantis observados com os quantis teóricos da distribuição normal.

Figura 9: Distribuição empírica e normal teórica do peso

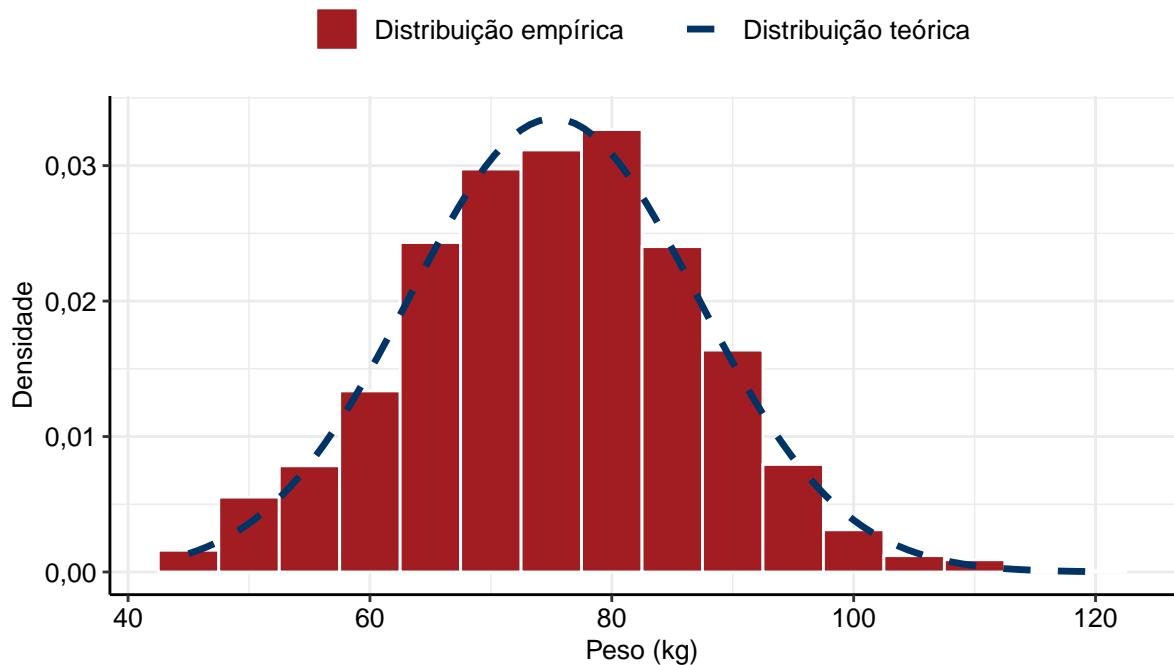


Figura 10: Distribuição empírica e normal teórica da altura

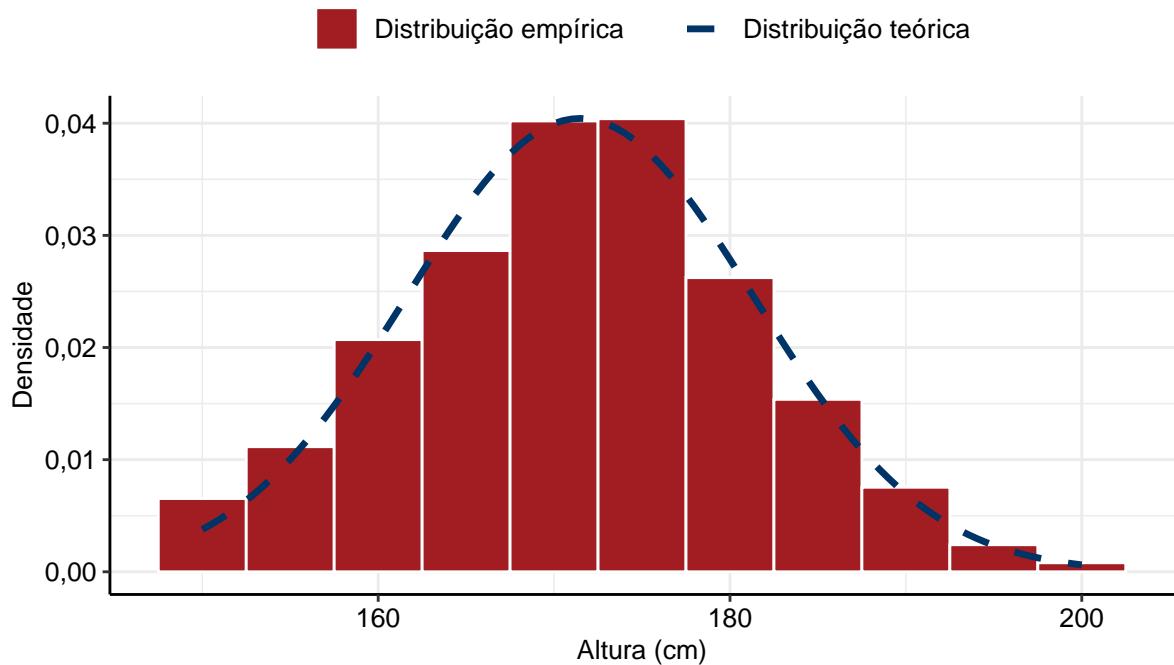


Figura 11: Q-Q plot do peso

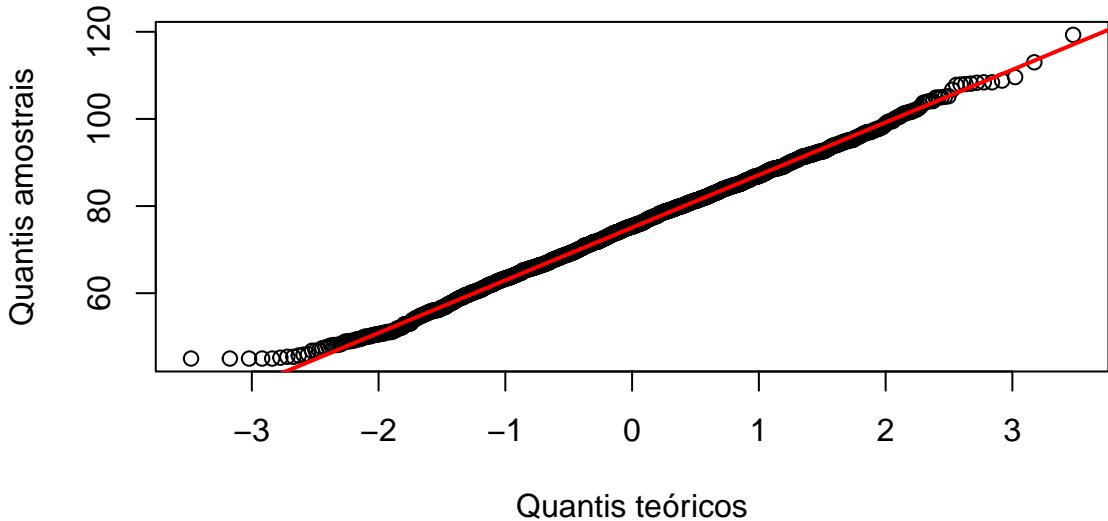
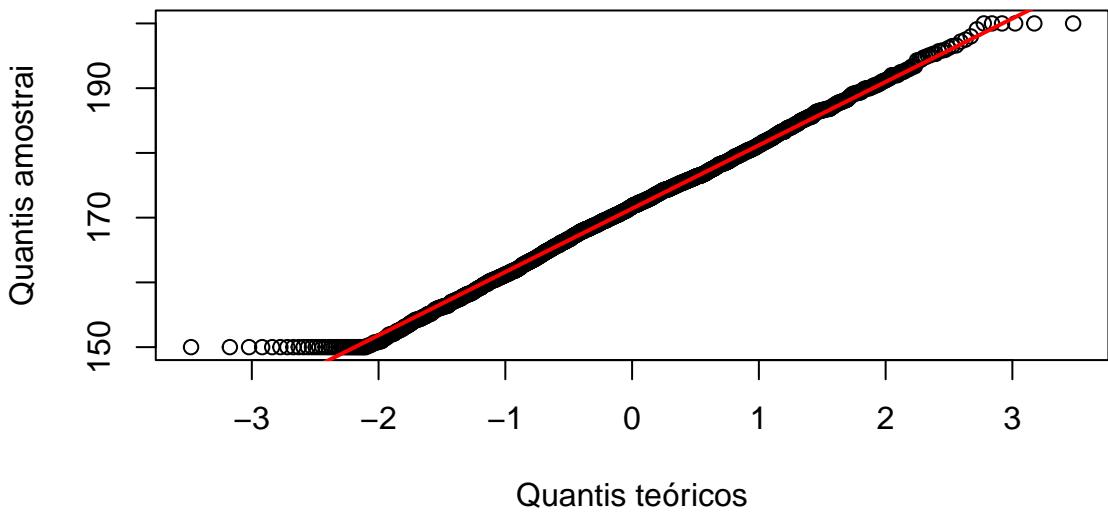


Figura 12: Q-Q plot da altura



Em seguida, após a verificação da normalidade das variáveis, foi realizado o teste de correlação de Pearson sob as seguintes hipóteses.

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

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, a estatística do teste e o intervalo de confiança a 95% estão na seguinte tabela.

Tabela 8: Resultados da correlação de Pearson entre peso e altura

$\hat{\rho}$	0,6971007
p-valor	0,0000000
Estatística t	43,3511704

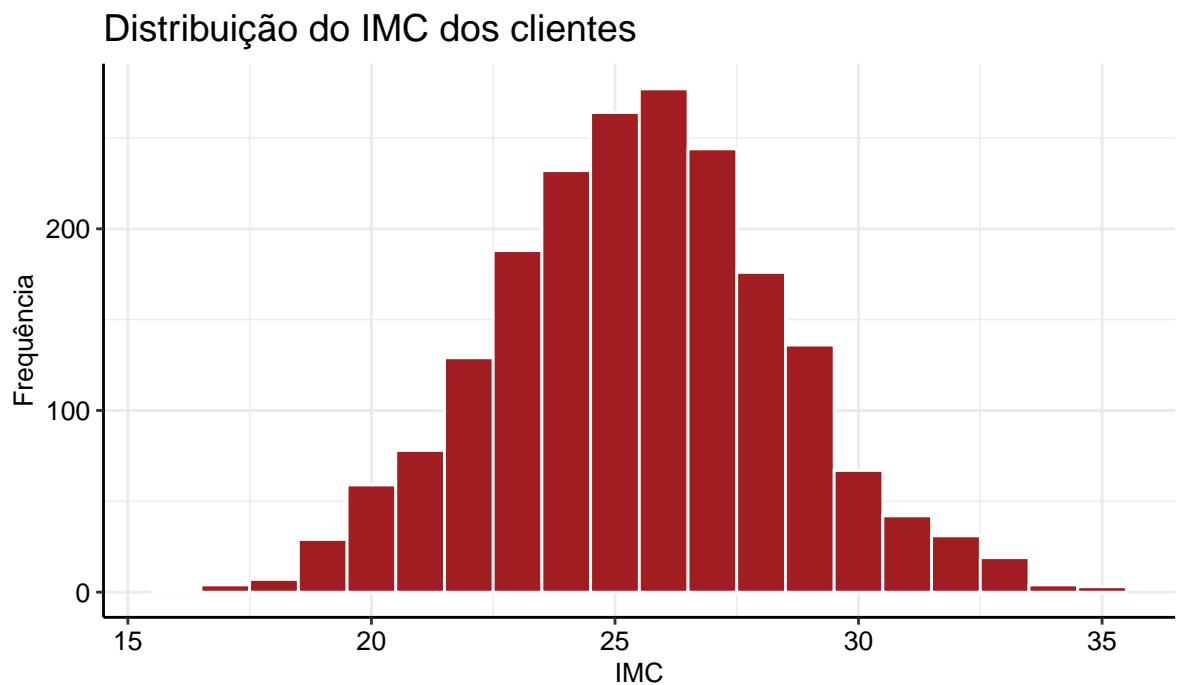
Parâmetro	Intervalo de Confiança (95%)
$\rho$	 0.6737991      0.7190173

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 a relação observada dificilmente ocorreu ao acaso. O intervalo de confiança indica que com 95% de confiança no procedimento de estimativa, 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**.

## 5.5 Análise do IMC (Extra)

O objetivo desta análise é examinar em maior detalhe o **Índice de Massa Corporal (IMC)** e a **Faixa de Peso** dos clientes, obtido a partir da relação entre o peso em quilogramas e o quadrado da altura em metros (conferir Anexo 3 e 4), IMC é variável **quantitativa contínua** e Faixa de Peso é **qualitativa ordinal**.

Figura 13: Distribuição do IMC dos clientes



Quadro 3: Principais métricas do IMC dos clientes

Estatística	Valor
Média	25,50
Desvio Padrão	2,96
Variância	8,75
Mínimo	16,15
1º Quartil	23,51
Mediana	25,51
3º Quartil	27,39
Máximo	34,99

Figura 14: Distribuição das faixas de peso dos clientes

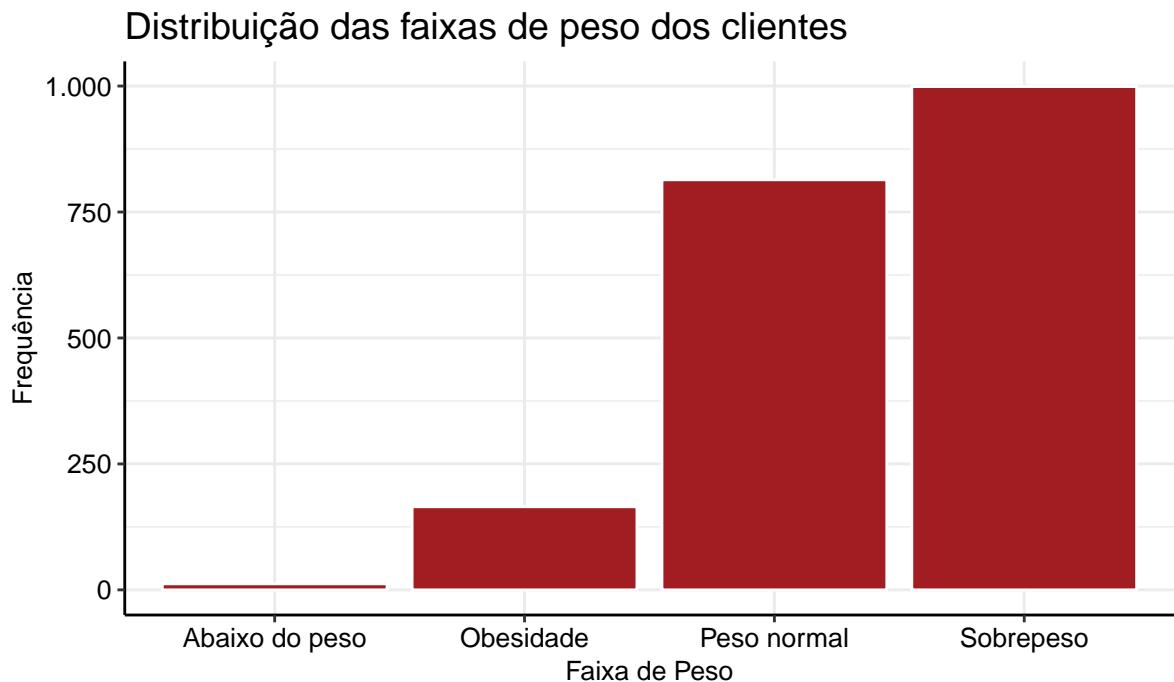


Tabela 9: Frequência e percentual dos clientes por faixa de peso

Faixa de Peso	frequencia	percentual (%)
Sobre peso	999	50,20
Peso normal	814	40,90
Obesidade	165	8,29
Abaixo do peso	12	0,60

A mediana do **IMC** foi de 25,51, indicando que pelo menos metade da população da região encontra-se na faixa de **sobre peso**. Observou-se também um número

reduzido de casos de **obesidade** (165) e de **abaixo do peso** (12), correspondendo a aproximadamente 8,29% e 0,60% da amostra de clientes, respectivamente.

## 5.6 Análise 3

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

Figura 15: Boxplot da idade dos clientes em Âmbar Seco

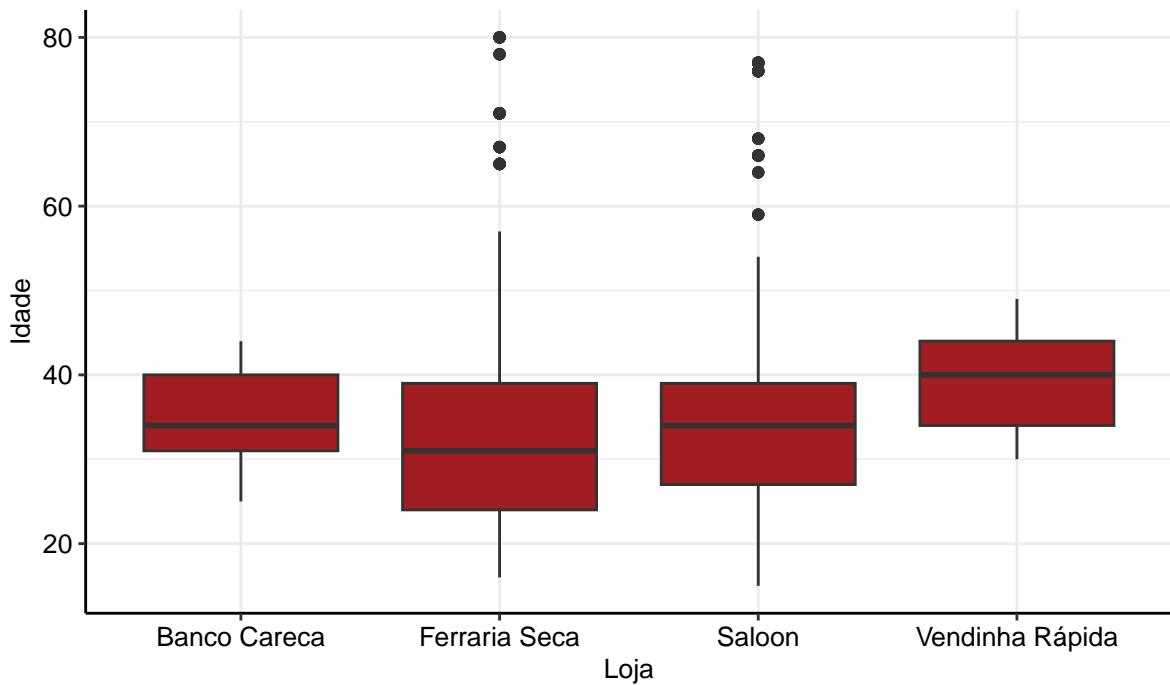


Tabela 10: Resumo da idade dos clientes em Âmbar Seco

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 10, e da disntância interquantílica na Tabela 11, 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 e mediana 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 e mediana, sugerindo um público mais jovens, no entanto, o maior desvio padrão e maior disntância interquantílica, indicam que o perfil étario dos clientes é mais vasto e heterogêno.

## 5.7 Análise 4

Por fim, o objetivo desta análise é identificar os três **produtos mais vendidos** nas três **lojas com maior receita** no ano de 1889.

Tabela 11: Lojas com maior receita em 1889

Loja	Receita
Ferraria Apache	1521102
Loja Ouro Fino	1505268
Loja TendTudo	1449637

Tabela 12: Produtos mais vendidos nas lojas com maior receita em 1889

Produto	Quantidade Vendida
Chapéu de Couro	32
Botas de Couro	30
Espingarda	30
Sela	30
Whisky	30
Colt .45	27
Municao	27
Pá	27
Machado	26
Cavalo	9

Tabela 13: Produtos mais vendidos separado por lojas

Loja	Produto	quantidadeVendida
Ferraria Apache	Botas de Couro	10
	Chapéu de Couro	10
	Espingarda	10
	Sela	10
	Whisky	10
Loja Ouro Fino	Chapéu de Couro	11
	Botas de Couro	10
	Espingarda	10
	Sela	10
	Whisky	10
Loja TendTudo	Chapéu de Couro	11
	Botas de Couro	10
	Espingarda	10
	Sela	10
	Whisky	10

As lojas com maior receita em 1889 foram **Ferraria Apache**, **Loja Ouro Fino** e **Loja TendTudo**. O produtos mais vendido nessas lojas foi o **Chapéu de Couro**, que contabilizou 32 vendas. Empatados para segundo lugar foram as **Botas de Couro**, a **Espingarda**, a **Sela** e o **Whisky**, com 30 vendas. O **Chapéu de Couro** foi o produto mais popular (ou empatou com o produto mais popular) nas três lojas com maior receita no ano de 1889. Esses resultados sugerem que os clientes têm preferência pelo chapéu, porém é importante notar que a diferença da quantidade de itens vendidos é baixa, variando apenas por uma ou duas unidades entre o item mais vendido e o segundo mais vendido (ver Tabelas 13 e 14).

## 6 Conclusão

A região mostrou-se muito promissora para investimentos, com crescimento expressivo na receita média ao longo do período analisado. A análise dos produtos revelou que itens como o **Cavalo** e o **Chapéu de Couro** tiveram destaque em termos de receita e quantidade vendida, respectivamente. As cidades da região também apresentaram crescimento econômico significativo, com **Rio Bravo** liderando nesse aspecto. Também foi verificado que o peso e a altura dos clientes está positivamente correlacionada. As lojas de **Âmbar Seco** possuem lojas com maior concentração de idade, e lojas mais heterogêneas quanto a faixa etária. E por fim, podemos observar que no ultimo ano dá série historica em análise, o **Chapeu de Coro** foi o item mais vendido das lojas que tiverem maior receita no ano, mesmo que apenas por duas unidades.

## 7 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. Cálculo da taxa de crescimento econômico

$$\text{Taxa de crescimento econômico} = \frac{\text{Receita}_{final} - \text{Receita}_{inicial}}{\text{Receita}_{inicial}} \times 100$$

### 3. Cálculo do Índice de Massa Corporal (IMC)

$$IMC = \frac{\text{peso}(kg)}{\text{altura}(m)^2}$$

### 4. Classificação do IMC segundo a Organização Mundial da Saúde (OMS)

- Abaixo do peso: IMC < 18,5
- Peso normal:  $18,5 \leq \text{IMC} < 24,9$
- Sobre peso:  $25 \leq \text{IMC} < 29,9$
- Obesidade:  $\text{IMC} \geq 30$