Consultores Responsáveis:

Estatiano 1

Estatiano 2

Estatiano 3

Requerente:

ESTAT

Sumário

1 Introdução

O presente relatório tem como objetivo realizar uma análise estatística exploratória a partir do conjunto de dados disponibilizado pela empresa Old Town Road, que contém informações sobre vendas, clientes, lojas e produtos de uma pequena cidade chamada Âmbar Seco. A proposta é compreender o comportamento e as relações entre diferentes variáveis, por meio de gráficos e medidas resumo, possibilitando uma visão mais ampla do desempenho das lojas e do perfil de seus clientes ao longo do tempo.

Inicialmente, foi analisada a variação das receitas médias anuais das lojas no período de 1880 a 1889, com o intuito de observar possíveis tendências de crescimento ou diminuição do faturamento médio ao longo dos anos. Para isso, foi utilizado um gráfico de linhas univariado, que permite acompanhar o comportamento contínuo da variável “receita média” em função do tempo. Além disso, os valores originalmente em dólares foram convertidos para reais, utilizando a cotação de R$5,31 por dólar.

Na sequência, foi investigada a relação entre o peso e a altura dos clientes, variáveis quantitativas contínuas que podem indicar padrões corporais distintos entre indivíduos. Foram analisados 1.990 clientes, e para garantir maior interpretabilidade, o peso foi convertido de libras para quilogramas e a altura de decímetros para centímetros. A relação entre essas variáveis foi representada por meio de um gráfico de dispersão bivariado, adequado para observar associações e possíveis correlações entre variáveis quantitativas. A partir disso, foi calculado o coeficiente de correlação de Pearson, o que permitiu avaliar a intensidade e o sentido dessa relação.

Em seguida, buscou-se compreender o perfil etário dos clientes em cada loja da cidade de Âmbar Seco, por meio da variável “idade”. Para isso, foi elaborado um boxplot bivariado, relacionando a variável quantitativa discreta “idade” com a variável qualitativa nominal “loja”. Além do gráfico, foi construído um quadro de medidas resumo (média, mediana, quartis, variância e extremos), possibilitando uma análise mais detalhada da distribuição das idades por loja e permitindo identificar diferenças entre os públicos atendidos.

Posteriormente, foi feita uma análise das três lojas com maior receita total no ano de 1889, com o intuito de identificar quais estabelecimentos apresentaram o melhor desempenho comercial no período. Essa análise foi realizada através de um gráfico de colunas, que facilita a comparação direta dos valores de receita entre as lojas. Em seguida, foi avaliada a quantidade dos produtos mais vendidos nas principais lojas, também utilizando gráficos de colunas, que permitem observar de forma clara quais itens contribuíram mais para o faturamento de cada loja.

2 Referencial Teórico

2.1 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum\limits_{i=1}^{n} X_i}{n}$$

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.

2.5.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

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

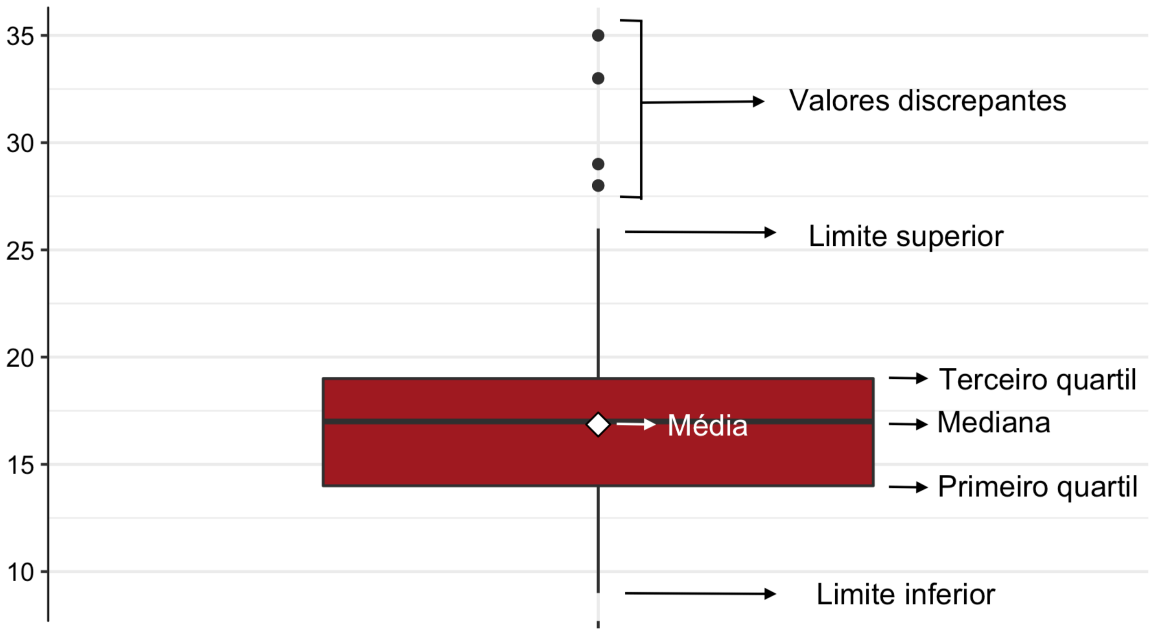
Com:

- $X_i =$ 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 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.

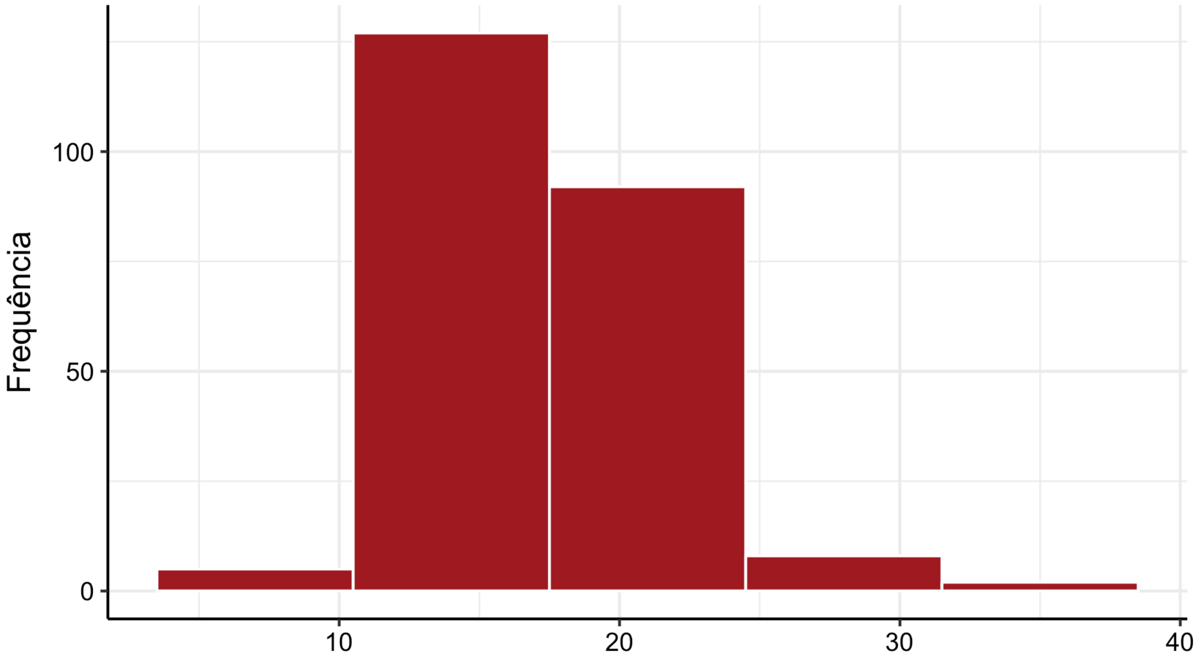
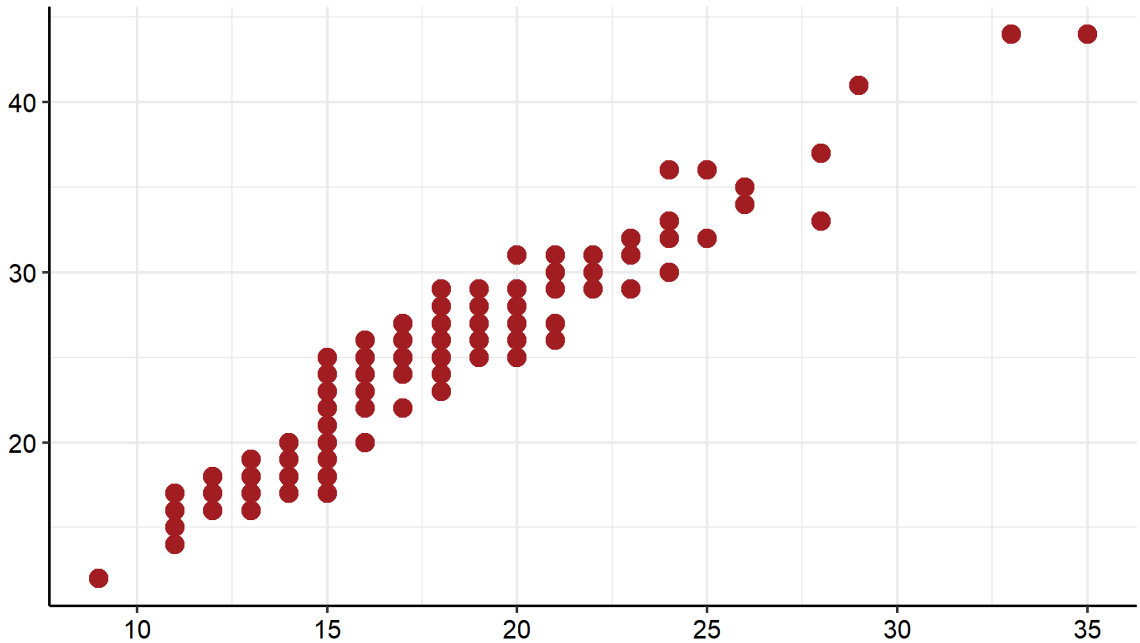


Gráfico de Dispersão

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.

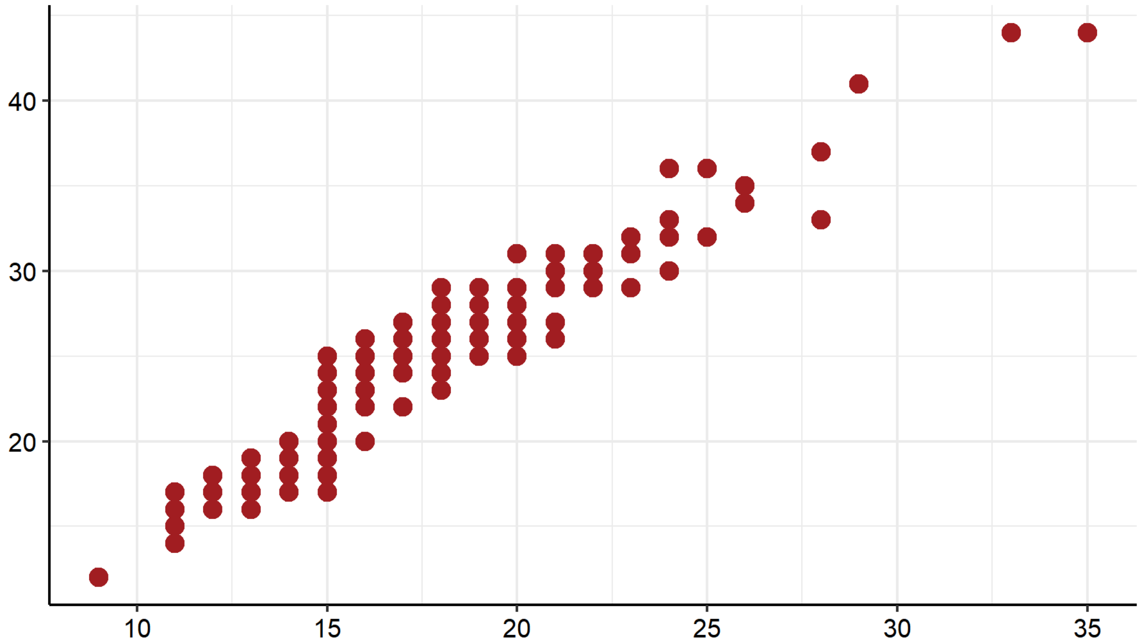
Figura 2: Exemplo de Gráfico de Dispersão



2.8 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



2.9 Tipos de Variáveis

2.9.1 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)

2.9.2 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)

2.10 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

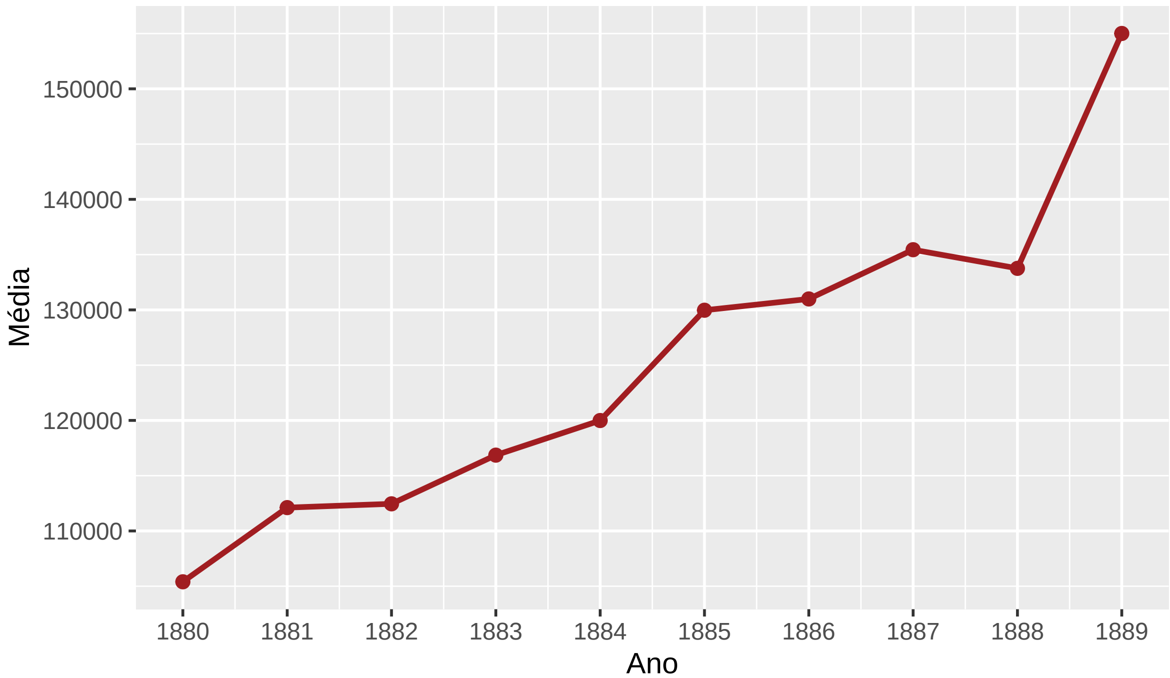
Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

3 Análises

3.1 Análise da média das receitas das lojas por ano

O objetivo da análise é observar a variação das médias das receitas ao longo dos anos de 1880 e 1889. Para isso, foi utilizado um gráfico de linhas univariado, que permite uma visualização melhor do comportamento contínuo das duas variáveis quantitativas: a média das receitas e os anos observados. Além disso, os dados foram agrupados pela data e foi feita a conversão de dólares para reais pela cotação de 5,31.

Figura 4: Gráfico de linhas univariado da média da receita das lojas por ano



A partir desse gráfico foi possível observar um crescimento dos valores ao longo dos anos, sendo o maior índice de aumento da receita entre os anos de 1888 e 1889, o que pode indicar alguma mudança marcante durante esse período que pode ter ocasionado certo crescimento da média. Além disso, é possível concluir que predomina um crescimento constante sem mudanças drásticas recorrentes.

3.2 Análise da relação entre o peso e a altura dos clientes

O objetivo da análise é observar a relação entre as variáveis quantitativas contínuas peso e altura dos clientes. Para isso, foram analisados 1990 clientes. Além disso, para melhor compreensão dos valores a unidade de medida do peso foi alterada de libras para quilogramas e a unidade de medida da altura foi alterada de decímetros para