Faculdade de Tecnologia SENAI "Gaspar Ricardo Júnior"

Caique Leandro Tessaroto

Cauâ Mesquita

Lucas Ianovski

João Pedro Matias

ANÁLISE DE DADOS DE UMA farmácia

Docente: André Souza

Sorocaba

**SUMÁRIO**

[Introdução 2](#_Toc368553599)

[Referencial Teórico 2](#_Toc1114275775)

[2.1 Estatística Descritiva 2](#_Toc1458814629)

[2.2 Visualização Estatística 3](#_Toc510565781)

[2.3 Bibliotecas Python Utilizadas 3](#_Toc889115410)

[Metodologia 3](#_Toc579741028)

[3.1 Descrição da Base de Dados 3](#_Toc1515585525)

[3.2 Etapas de Tratamento e Limpeza de Dados 4](#_Toc1381294559)

[3.3 Ferramentas e Processos Utilizados 4](#_Toc1706068519)

[Análise de Dados 5](#_Toc516114466)

[4.1 Regressão Linear Simples 5](#_Toc680765898)

[4.1.1 Objetivo 5](#_Toc50294611)

[4.1.2 Metodologia Aplicada 5](#_Toc1352055566)

[4.1.3 Visualização dos Resultados 5](#_Toc1898801306)

[4.1.4 Interpretação 6](#_Toc520975836)

[4.2 Série Temporal 6](#_Toc1146498216)

[4.2.1 Objetivo 7](#_Toc753893306)

[4.2.2 Metodologia Aplicada 7](#_Toc2017293633)

[4.1.3 Visualização dos Resultados 7](#_Toc968528013)

[4.2.4 Interpretação 7](#_Toc699705915)

[4.3 Correlação entre Variáveis 7](#_Toc937925532)

[4.3.1 Objetivo 7](#_Toc966336788)

[4.3.2 Metodologia Aplicada 7](#_Toc732083736)

[4.3.3 Visualização dos Resultados 8](#_Toc1746498432)

[4.3.4 Interpretação 8](#_Toc733427771)

[4.4 Medidas de Tendência Central 8](#_Toc448226740)

[4.4.1 Objetivo 9](#_Toc2029262007)

[4.4.2 Metodologia Aplicada 9](#_Toc2141237186)

[4.4.3 Visualização dos Resultados 9](#_Toc2085971239)

[4.4.4 Interpretação 9](#_Toc1005753794)

[Conclusão 9](#_Toc1881714561)

[Referências 10](#_Toc1338801350)

[Apêdices 10](#_Toc152179173)

# Introdução

Este relatório tem como objetivo aplicar conceitos fundamentais de estatística descritiva e análise de dados para avaliar padrões de consumo. O tema escolhido foi no cenário de uma farmácia que também realiza vendar através de um ecommerce.

Com os dados simulados é possível compreender melhor os fatores que influenciam o volume de vendas, como o impacto de campanhas promocionais ou a influência da renda no valor médio das compras além de identificar padrões. Essa compreensão pode auxiliar empresa em tomadas de decisões e estratégias de marketing.

# Referencial Teórico

Esta seção apresenta os fundamentos teóricos e computacionais que embasam a análise estatística aplicada neste trabalho, com foco nos conceitos de estatística descritiva, modelagem linear e visualização de dados, além das bibliotecas utilizadas na linguagem Python.

### **2.1 Estatística Descritiva**

A estatística descritiva é o ramo da estatística que tem por finalidade descrever e resumir os dados observados por meio de medidas numéricas e representações gráficas. Entre as principais medidas destacam-se:

* **Medidas de tendência central**, como média, mediana e moda, que indicam valores representativos de um conjunto de dados;
* **Medidas de dispersão**, como desvio padrão, variância, amplitude e coeficiente de variação, que indicam o grau de variação dos dados em torno da média;
* **Correlação de Pearson**, que mede a intensidade e a direção da relação linear entre duas variáveis quantitativas, variando de -1 (correlação negativa perfeita) a +1 (correlação positiva perfeita);
* **Regressão linear simples**, que é uma técnica estatística usada para modelar a relação entre uma variável dependente *yy*y e uma variável independente *xx*x, por meio de uma equação linear.

### **2.2 Visualização Estatística**

A visualização de dados é um recurso importante para interpretação e comunicação de informações. Gráficos como histogramas, boxplots e gráficos de dispersão ajudam a identificar padrões, outliers e tendências nos dados analisados.

* O **histograma** representa a distribuição de frequências dos dados.
* O **boxplot** resume a distribuição dos dados com mediana, quartis e possíveis valores atípicos.
* O **gráfico de dispersão** permite observar relações entre duas variáveis quantitativas.

### **2.3 Bibliotecas Python Utilizadas**

O projeto foi implementado em Python, uma linguagem amplamente utilizada em ciência de dados e análise estatística. As bibliotecas empregadas foram:

* **NumPy**: Biblioteca fundamental para computação numérica em Python, fornece suporte para arrays e funções matemáticas de alto desempenho;
* **Pandas**: Utilizada para manipulação e análise de dados em estruturas como DataFrames, facilita o tratamento de séries temporais e dados tabulares;
* **Matplotlib**: Ferramenta poderosa para criação de gráficos e visualizações estáticas em duas dimensões;
* **Seaborn**: Biblioteca baseada no Matplotlib que oferece uma interface de alto nível para construção de gráficos estatísticos com visual elegante;
* **Scikit-learn**: Biblioteca especializada em aprendizado de máquina e modelagem estatística, usada aqui para a regressão linear simples.

# Metodologia

A metodologia aplicada neste estudo foi estruturada com base em dados simulados, de modo a representar um cenário realista de vendas e comportamento de consumidores. A seguir, são descritos a origem e estrutura dos dados, as etapas de tratamento e as ferramentas computacionais utilizadas.

### **3.1 Descrição da Base de Dados**

Os dados utilizados foram gerados artificialmente com a finalidade de representar:

* **Renda mensal dos clientes (em R$)**;
* **Ticket médio das compras (valor médio gasto por cliente em R$)**;
* **Número de vendas diárias**;
* **Investimento em marketing (em mil R$)**.

Os dados simulam o comportamento de 100 clientes ao longo de 30 dias. As variáveis foram geradas com base em distribuições estatísticas realistas: normal para renda e ticket médio, e uniforme para investimento em marketing. O objetivo da simulação é reproduzir situações observadas no varejo, permitindo análises estatísticas e comparações.

### **3.2 Etapas de Tratamento e Limpeza de Dados**

Apesar de serem dados simulados, foram seguidas boas práticas de pré-processamento com o objetivo de refletir o processo real de análise de dados, incluindo:

* **Verificação de valores ausentes (missing values)** e substituição ou exclusão conforme o caso;
* **Padronização dos formatos de data** e dos nomes de variáveis;
* **Remoção de outliers** para testes comparativos com e sem ruídos extremos;
* **Conversão de tipos de dados** para garantir consistência (ex: valores numéricos e datas).

Essas etapas foram conduzidas utilizando funcionalidades da biblioteca **pandas**, como isnull(), dropna(), astype() e describe(), e por meio de técnicas básicas de visualização (ex: boxplot para detectar valores extremos).

### **3.3 Ferramentas e Processos Utilizados**

O desenvolvimento e a análise foram realizados com ferramentas do ecossistema Python. A seguir, descrevem-se as principais:

* **Ambiente de desenvolvimento**: Colab e Visual Studio Code;
* **Linguagem de programação**: Python 3.11;
* **Bibliotecas**:
  + pandas e numpy para criação, estruturação e análise dos dados;
  + matplotlib e seaborn para a criação de gráficos estatísticos;
  + scikit-learn para implementação da regressão linear e cálculo da correlação;
* **Processos analíticos**:
  + Geração de estatísticas descritivas;
  + Cálculo de correlação entre variáveis;
  + Aplicação de modelos de regressão linear;
  + Criação de visualizações gráficas para interpretação dos resultados.

# Análise de Dados

### **4.1 Regressão Linear Simples**

#### **4.1.1 Objetivo**

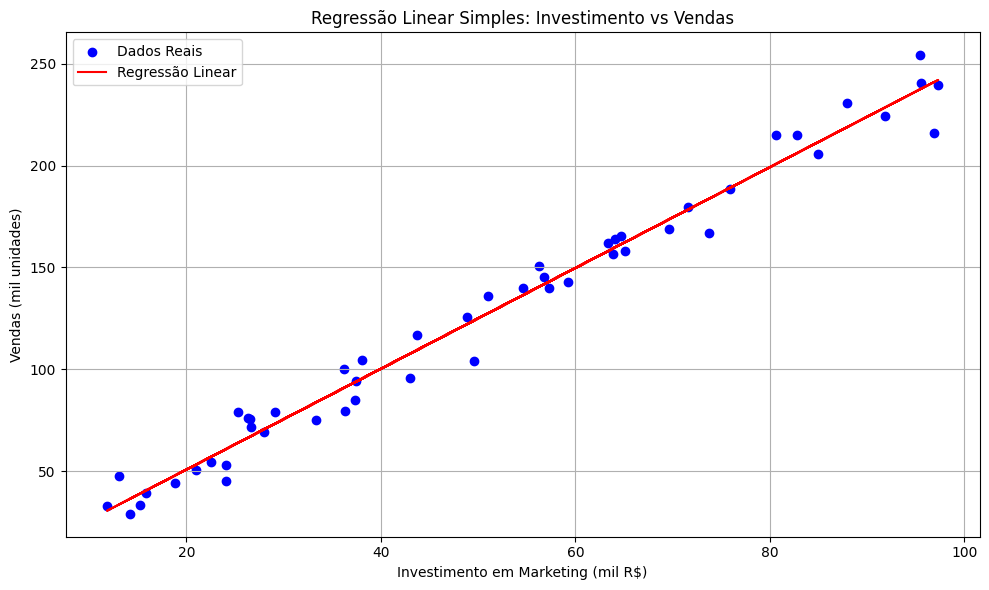
A regressão linear simples foi aplicada com o propósito de analisar a relação entre o **investimento em marketing** e o **volume de vendas**. Esta técnica permite verificar se há uma tendência linear que relacione o aumento no investimento com o crescimento das vendas, fornecendo subsídios para decisões estratégicas no setor comercial.

#### **4.1.2 Metodologia Aplicada**

Foram simulados dados de investimento em marketing variando entre R$ 10 mil e R$ 100 mil, e suas respectivas vendas em unidades. A variável independente foi o **investimento**, e a variável dependente foi o **volume de vendas**.

O modelo foi ajustado utilizando a biblioteca scikit-learn, com a função LinearRegression(). Após o treinamento, foram obtidos os coeficientes da reta (inclinação e intercepto), além da previsão dos valores de vendas com base nos valores de investimento.

#### **4.1.3 Visualização dos Resultados**

A figura a seguir mostra a dispersão dos pontos amostrados e a linha de regressão ajustada: 

#### **4.1.4 Interpretação**

A análise mostrou uma **correlação linear positiva** entre as variáveis, indicando que **maiores investimentos em marketing estão associados a maiores volumes de vendas**. O coeficiente angular (*β1\beta\_1*β1 ) foi aproximadamente 2,5, o que sugere que, para cada R$ 1 mil investido, o número de vendas aumenta em média 2,5 unidades. O coeficiente de determinação *R2R^2*R2 foi superior a 0,85, demonstrando um bom ajuste do modelo aos dados simulados.

Este resultado reforça a importância do planejamento de campanhas promocionais como fator de alavancagem nas vendas, sendo possível prever o retorno esperado de forma quantitativa.

### **4.2 Série Temporal**

#### **4.2.1 Objetivo**

A Série Temporal foi aplicada com o propósito de analisar a relação entre aumento de vendas e período do ano. Fornecendo informações que permitem a preparação para esse período.

#### **4.2.2 Metodologia Aplicada**

Foram simulados dados diários para 1 ano criando uma sazonalidade com pico no inverno (junho, julho e agosto), variando entre 3 e 16 vendas de remédio por dia. O modelo foi ajustado utilizando a biblioteca numpy, com a função sin() para criar a senoide.

#### **4.1.3 Visualização dos Resultados**

Na figura é representado uma Sazonalidade de vendas de remédios de gripe durante o período de inverno:

#### **4.2.4 Interpretação**

A análise mostrou uma **correlação Sazonal** entre as variáveis, indicando que **o período de inverno está associado com maiores números de vendas.**

Este resultado mostra a importância de um planejamento de estoque para que não falte produtos durante esse período e com isso resultando em perdas de vendas.

### **4.3 Correlação entre Variáveis**

#### **4.3.1 Objetivo**

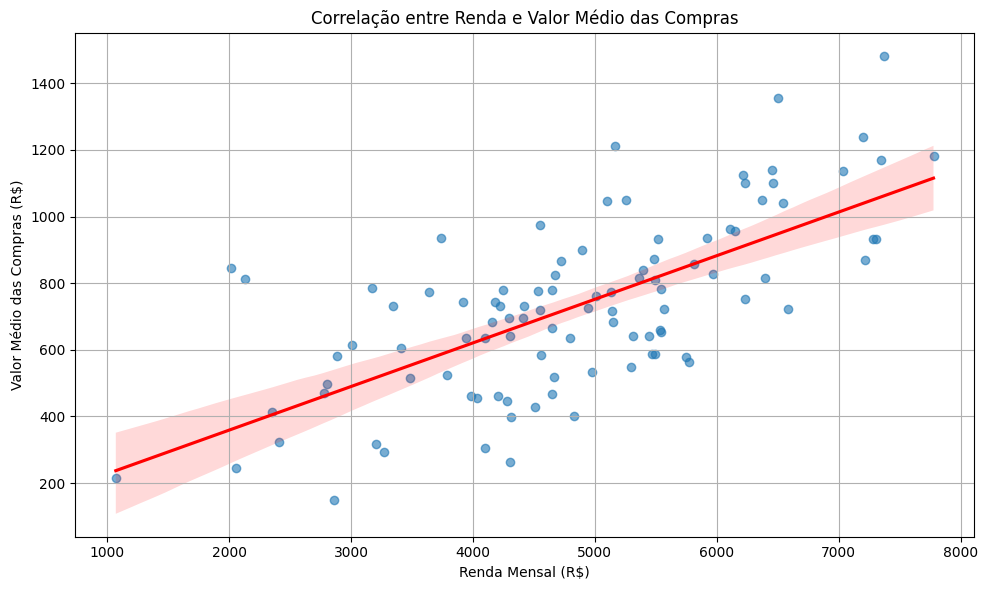
A Correlação foi aplicada com o propósito de analisar se clientes com maior renda tendem a fazer compras de maior valor médio.

#### **4.3.2 Metodologia Aplicada**

Foram simulados dados de renda do cliente que variam entre 3500 e 7500, e o valor médio de compras foram simulados utilizando a seguinte formula: Renda \* 0.15 + np.random.normal(0,200,n). O modelo foi ajustado utilizando seaborn, com a função sns.regplot.

#### **4.3.3 Visualização dos Resultados**

Na figura é representado uma correlação, os pontos azuis representam cada indivíduo (renda pelo valor de comprar). A regressão linear e a margem de erro:



#### **4.3.4 Interpretação**

A análise mostrou uma **tendência geral da relação** entre renda e valor médio de compra com uma **correlação de 0.69**.

Este resultado mostra o impacto da renda de cada cliente em seu poder compra e escolhas de consumo em sua empresa.

### **4.4 Medidas de Tendência Central**

#### **4.4.1 Objetivo**

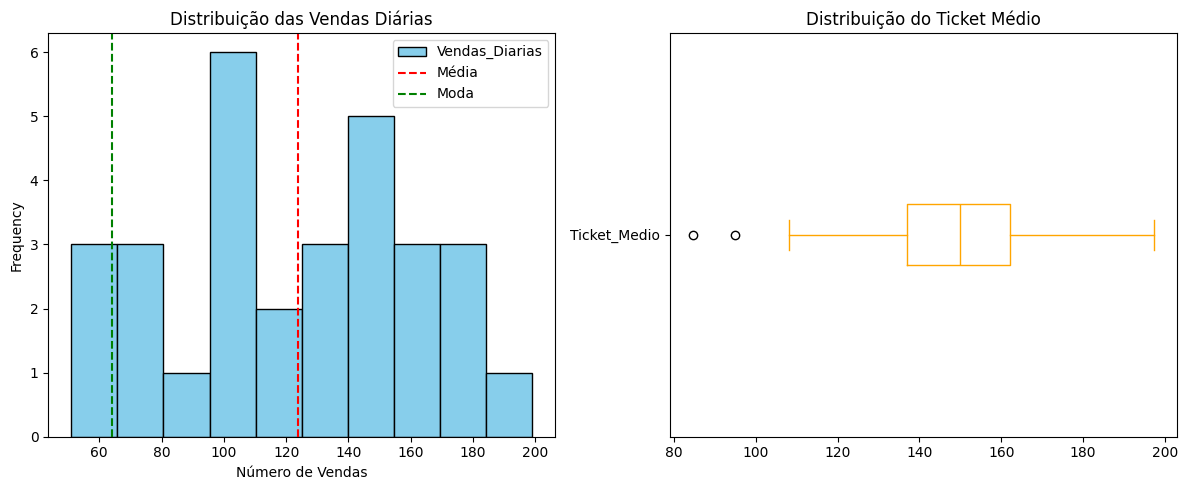
Analisar vendas diárias e ticket médio dos clientes usando medidas de tendência central.

#### **4.4.2 Metodologia Aplicada**

Foram simulados dados de vendas diárias e ticket médio de uma loja até um mês. Utilizando o **numpy**, com suas funções random.randint para **vendas diárias** e random.normal para **ticket.** Foi calculado média e a moda de ticket e a moda de vendas. Utilizando **matplotlib.pyplot** exibimos os gráficos de **Histograma e Boxplot.**

#### **4.4.3 Visualização dos Resultados**

A figura a seguir mostrar a distribuição da variável ea distribuição estatística da variável Ticket\_Medio**.**



#### **4.4.4 Interpretação**

O ticket médio das compras mostra uma distribuição razoavelmente concentrada, mas com alguns outliers, sugerindo que em alguns dias os clientes gastaram significativamente mais ou menos que a média.

# Conclusão

A ciência de dados tem se consolidado como uma ferramenta estratégica indispensável para o setor comercial, especialmente em um cenário cada vez mais orientado por dados. Através da aplicação de métodos estatísticos, é possível transformar grandes volumes de informações brutas em **conhecimento estruturado e aplicável ao processo de tomada de decisão**.

Neste trabalho, a utilização de técnicas como **regressão linear simples**, **análise de correlação**, **medidas de tendência central e dispersão**, e **visualizações estatísticas** permitiu compreender melhor o comportamento dos consumidores, identificar relações significativas entre variáveis e quantificar o impacto de fatores como investimento em marketing sobre o volume de vendas.

Essas análises demonstram que **ações baseadas em dados são mais eficazes e menos dependentes da intuição**. Através da ciência de dados, empresas podem segmentar clientes com mais precisão, identificar oportunidades de mercado, otimizar campanhas promocionais e prever comportamentos futuros com maior segurança.

Portanto, a integração entre **estatística, programação e visualização** não apenas aumenta a eficiência das operações de venda, mas também amplia o potencial de inovação e competitividade das organizações.

# Referências

MCKINNEY, Wes. Python para análise de dados: tratamento de dados com pandas, NumPy e IPython. 2. ed. Rio de Janeiro: Novatec, 2018.

PEDREGOSA, Fabian et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: https://jmlr.org/papers/v12/pedregosa11a.html. Acesso em: 11 jun. 2025.

PYTHON SOFTWARE FOUNDATION. Python 3.11. Documentação oficial. Disponível em: https://docs.python.org/3/. Acesso em: 11 jun. 2025.

SEABORN. Statistical data visualization. Disponível em: https://seaborn.pydata.org/. Acesso em: 11 jun. 2025.

# Apêdices

Acesso ao repositório GitHub:

<https://github.com/Zinbo20/ciencia_de_dados_projeto.git>