**Relatório Completo de Análise e Transformação de Dados**

5 ADS MAS

Juliana Alves   
Leonardo Mucci  
Marcos Vinicius  
Rodrigo Veloso

**Sumário**

[**Parte 1: Estatística Descritiva** 5](#_Toc170333052)

[**1.1 Carregamento e Limpeza dos Dados** 5](#_Toc170333053)

[**Bibliotecas Utilizadas:** 5](#_Toc170333054)

[**Funções e Métodos:** 5](#_Toc170333055)

[**Código:** 5](#_Toc170333056)

[**Descrição:** 6](#_Toc170333057)

[**1.2 Cálculo das Estatísticas Descritivas** 6](#_Toc170333058)

[**Funções e Métodos:** 6](#_Toc170333059)

[**Código:** 7](#_Toc170333060)

[**Descrição:** 7](#_Toc170333061)

[**1.3 Distribuição de Frequência dos Atributos Categóricos** 7](#_Toc170333062)

[**Funções e Métodos:** 7](#_Toc170333063)

[**Código:** 8](#_Toc170333064)

[**Descrição:** 8](#_Toc170333065)

[**Parte 2: Medidas de Dispersão, Variância e Desvio-Padrão** 9](#_Toc170333066)

[**2.1 Carregamento e Limpeza dos Dados** 9](#_Toc170333067)

[**2.2 Cálculo das Medidas de Dispersão** 9](#_Toc170333068)

[**Funções e Métodos:** 9](#_Toc170333069)

[**Código:** 9](#_Toc170333070)

[**Descrição:** 9](#_Toc170333071)

[**Parte 3: Análise de Correlação e Representações Gráficas** 10](#_Toc170333072)

[**3.1 Carregamento e Limpeza dos Dados** *Os dados foram carregados e limpos conforme descrito na Parte 1.* 10](#_Toc170333073)

[**3.2 Análise de Correlação** 10](#_Toc170333074)

[**Funções e Métodos:** 10](#_Toc170333075)

[**Código:** 10](#_Toc170333076)

[**Descrição:** 10](#_Toc170333077)

[**3.3 Representações Gráficas** 11](#_Toc170333078)

[**Bibliotecas Utilizadas:** 11](#_Toc170333079)

[**Funções e Métodos:** 11](#_Toc170333080)

[**Código:** 12](#_Toc170333081)

[**Descrição:** 17](#_Toc170333082)

[**Parte 4: Integração de Dados, Valores Inconsistentes e Redundância de Dados** 18](#_Toc170333083)

[**4.1 Integração de Dados** 18](#_Toc170333084)

[**4.2 Identificação e Correção de Valores Inconsistentes** 18](#_Toc170333085)

[**Funções e Métodos:** 18](#_Toc170333086)

[**Código:** 18](#_Toc170333087)

[**Descrição:** 18](#_Toc170333088)

[**4.3 Identificação e Remoção de Redundâncias** 19](#_Toc170333089)

[**Funções e Métodos:** 19](#_Toc170333090)

[**Código:** 19](#_Toc170333091)

[**Descrição:** 19](#_Toc170333092)

[**Parte 5: Transformação de Dados** 20](#_Toc170333093)

[**5.1 Normalização de Dados** 20](#_Toc170333094)

[**Bibliotecas Utilizadas:** 20](#_Toc170333095)

[**Funções e Métodos:** 20](#_Toc170333096)

[**Código:** 20](#_Toc170333097)

[**Descrição:** 20](#_Toc170333098)

[**5.2 Codificação de Dados Categóricos** 21](#_Toc170333099)

[**Funções e Métodos:** 21](#_Toc170333100)

[**Código:** 21](#_Toc170333101)

[**Descrição:** 21](#_Toc170333102)

[**Conclusão** 22](#_Toc170333103)

**Objetivo:** O objetivo deste projeto foi aplicar conceitos de estatística descritiva, medidas de dispersão, análise de correlação, integração de dados, correção de valores inconsistentes, remoção de redundâncias e transformação de dados em uma base de clientes. Este relatório documenta detalhadamente todas as etapas realizadas, os métodos utilizados e os resultados obtidos.

# **Parte 1: Estatística Descritiva**

## **1.1 Carregamento e Limpeza dos Dados**

### **Bibliotecas Utilizadas:**

* pandas: Para manipulação e análise de dados.
* scipy: Para cálculo da moda.

### **Funções e Métodos:**

* **pd.read\_csv():** Carrega a base de dados de um arquivo CSV.
* **DataFrame.drop\_duplicates**(): Remove registros duplicados.
* **DataFrame.replace():** Substitui valores inconsistentes.
* **DataFrame.quantile():** Calcula os quartis dos dados.

### **Código:**

****

### **Descrição:**

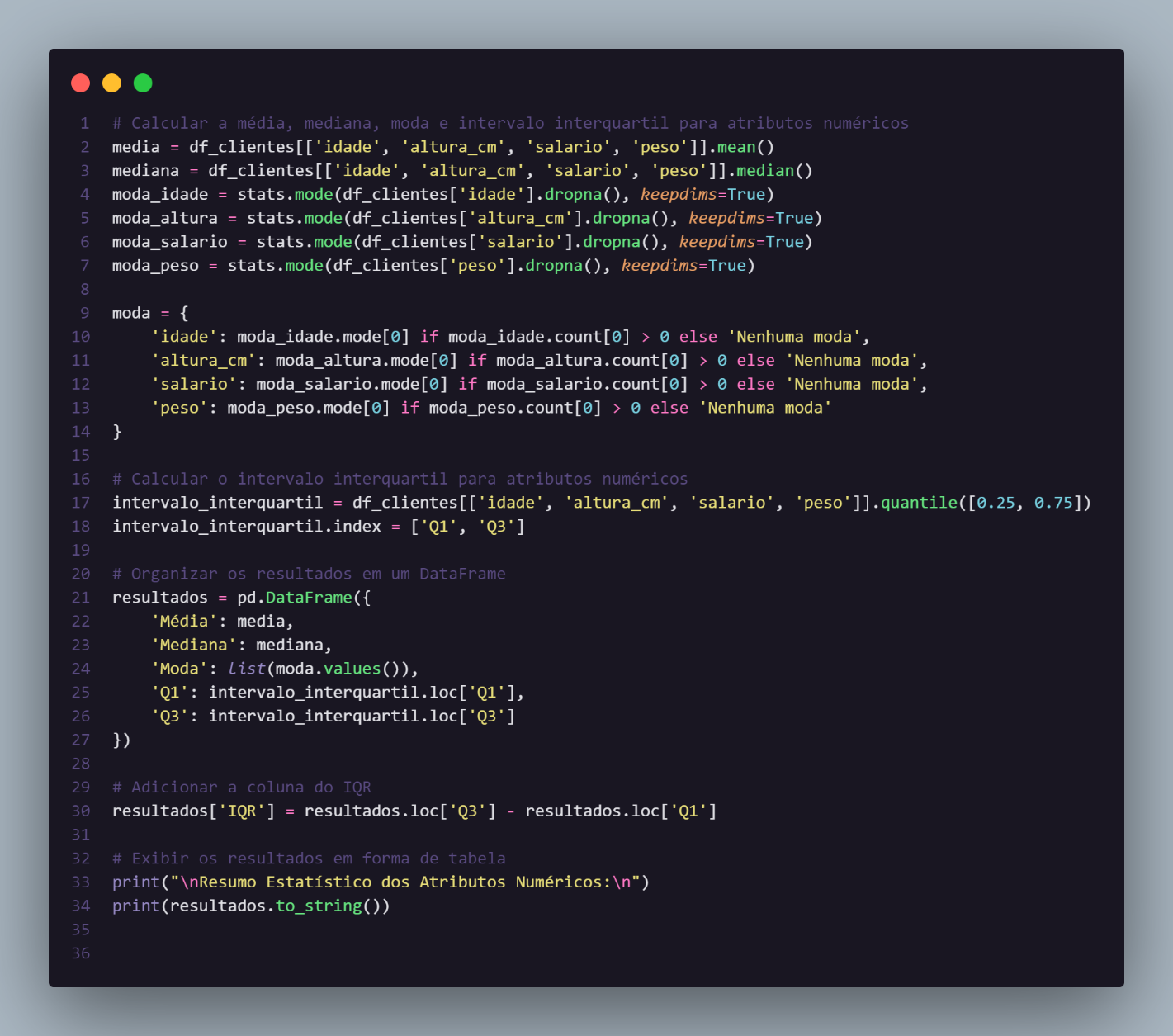
* **Carregamento dos Dados:** Utilizamos a função pd.read\_csv() para carregar a base de clientes de um arquivo CSV.
* **Limpeza dos Dados:** Removemos registros com idades, alturas, sexos e salários inconsistentes, além de registros duplicados.

## **1.2 Cálculo das Estatísticas Descritivas**

### **Funções e Métodos:**

* DataFrame.mean(): Calcula a média dos dados.
* DataFrame.median(): Calcula a mediana dos dados.
* stats.mode(): Calcula a moda dos dados.
* DataFrame.quantile(): Calcula os quartis dos dados.

### **Código:**

****

### **Descrição:**

* **Cálculo da Média, Mediana, Moda e Intervalo Interquartil:** Utilizamos métodos do Pandas e SciPy para calcular as estatísticas descritivas dos atributos numéricos.

## **1.3 Distribuição de Frequência dos Atributos Categóricos**

### **Funções e Métodos:**

* DataFrame.value\_counts(): Conta a frequência dos valores categóricos.
* Series.rename\_axis(): Renomeia o eixo de uma Series.
* Series.reset\_index(): Reseta o índice da Series.

### **Código:**



### **Descrição:**

* **Distribuição de Frequência:** Utilizamos value\_counts() para contar a frequência dos valores categóricos e formatamos a saída com rename\_axis() e reset\_index().

# **Parte 2: Medidas de Dispersão, Variância e Desvio-Padrão**

## **2.1 Carregamento e Limpeza dos Dados**

*(Os dados foram carregados e limpos conforme descrito na Parte 1.)*

## **2.2 Cálculo das Medidas de Dispersão**

### **Funções e Métodos:**

* DataFrame.max(): Calcula o valor máximo dos dados.
* DataFrame.min(): Calcula o valor mínimo dos dados.
* DataFrame.var(): Calcula a variância dos dados.
* DataFrame.std(): Calcula o desvio-padrão dos dados.

### **Código:**



### **Descrição:**

* **Cálculo da Amplitude, Variância e Desvio-Padrão:** Utilizamos métodos do Pandas para calcular as medidas de dispersão dos atributos numéricos.

# **Parte 3: Análise de Correlação e Representações Gráficas**

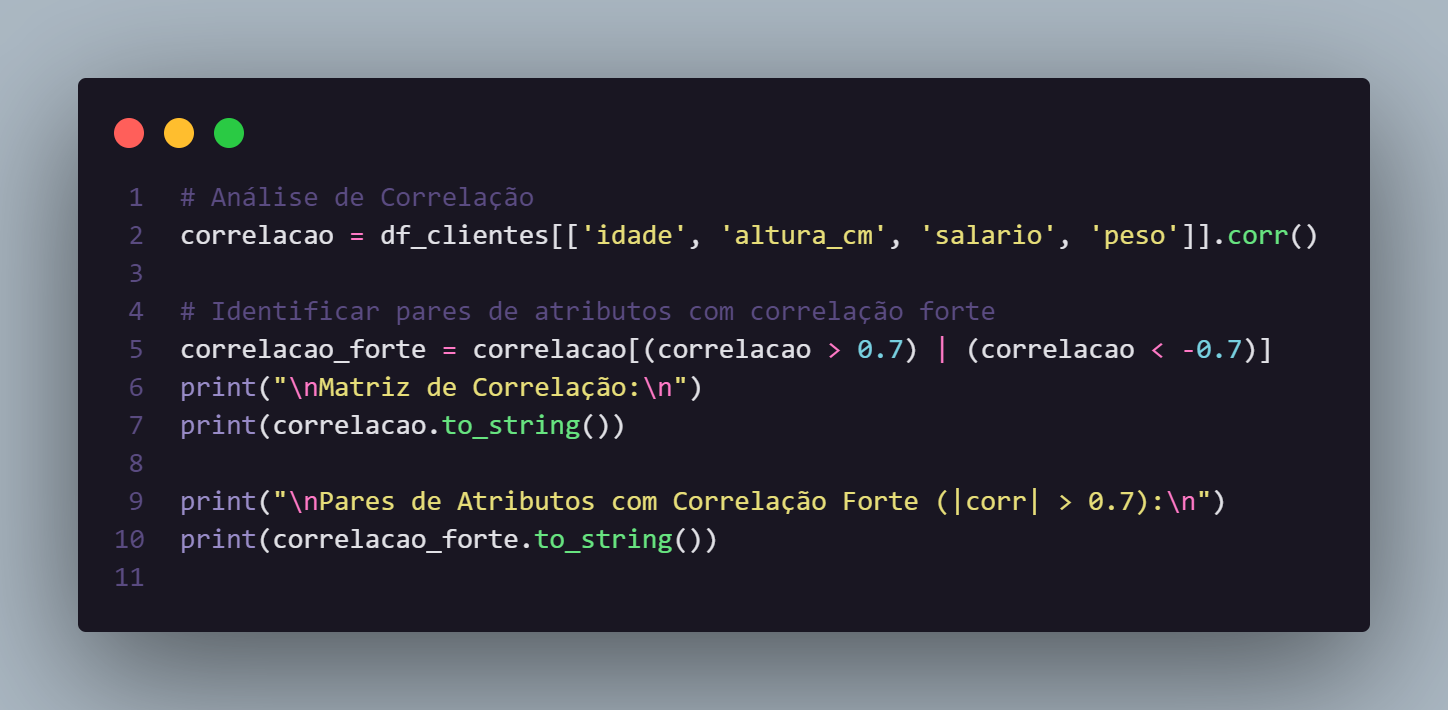
## **3.1 Carregamento e Limpeza dos Dados** *Os dados foram carregados e limpos conforme descrito na Parte 1.*

## **3.2 Análise de Correlação**

### **Funções e Métodos:**

* DataFrame.corr(): Calcula a matriz de correlação dos dados.

### **Código:**



### **Descrição:**

* **Análise de Correlação:** Utilizamos o método corr() do Pandas para calcular a matriz de correlação entre os atributos numéricos.

## **3.3 Representações Gráficas**

### **Bibliotecas Utilizadas:**

* matplotlib: Para geração de gráficos.
* seaborn: Para visualização de dados estatísticos.

### **Funções e Métodos:**

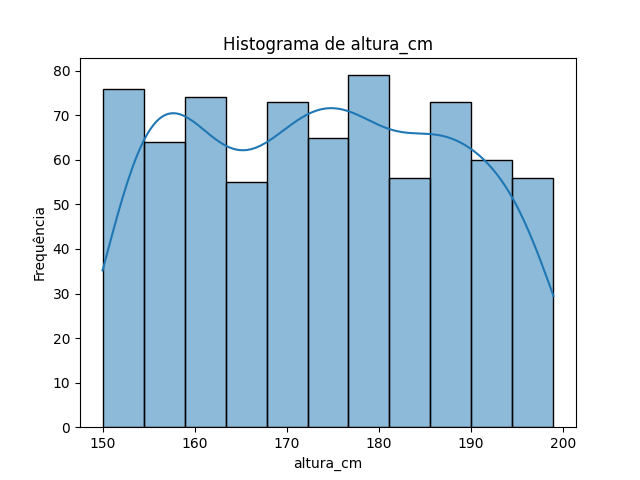
* sns.histplot(): Gera histogramas.
* sns.boxplot(): Gera box plots.
* sns.scatterplot(): Gera gráficos de dispersão.
* sns.heatmap(): Gera mapas de calor.

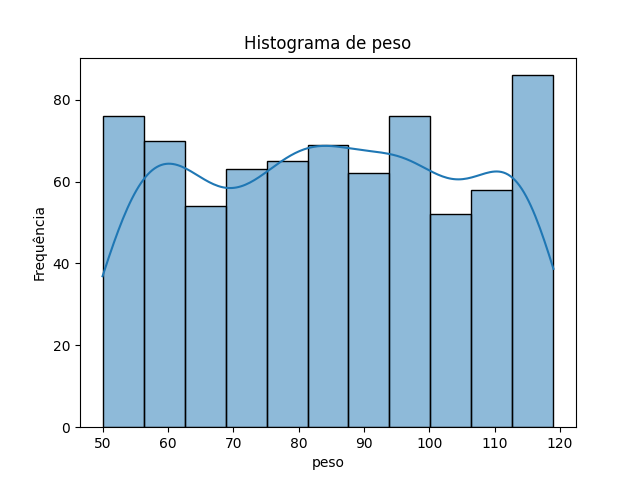
### **Código:**

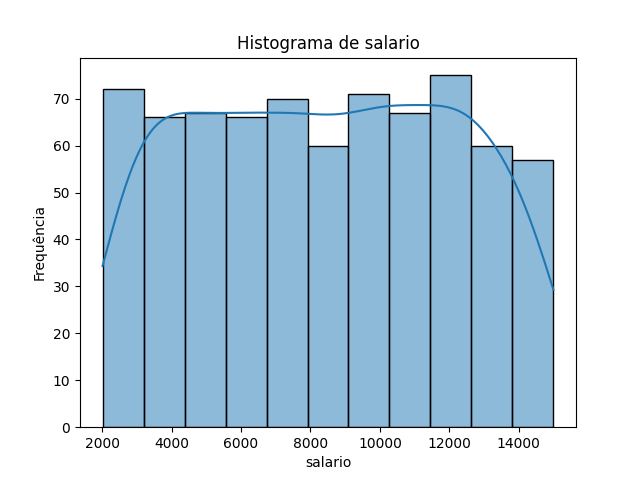


Gráfico, Histograma

Descrição gerada automaticamente

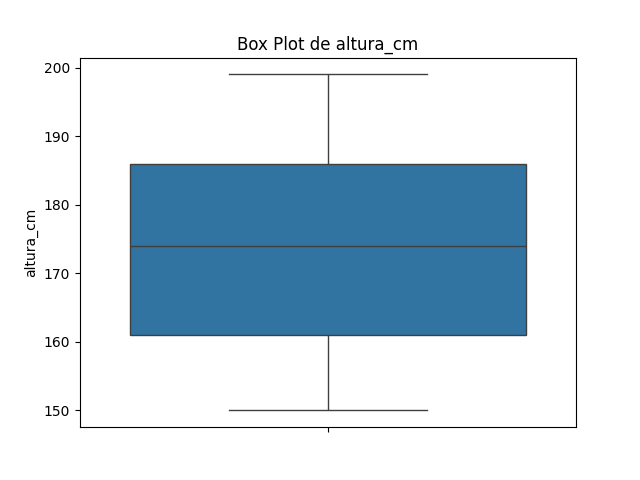
****

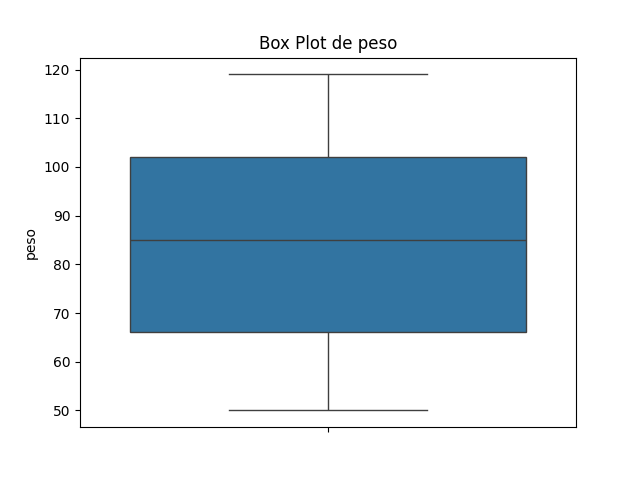


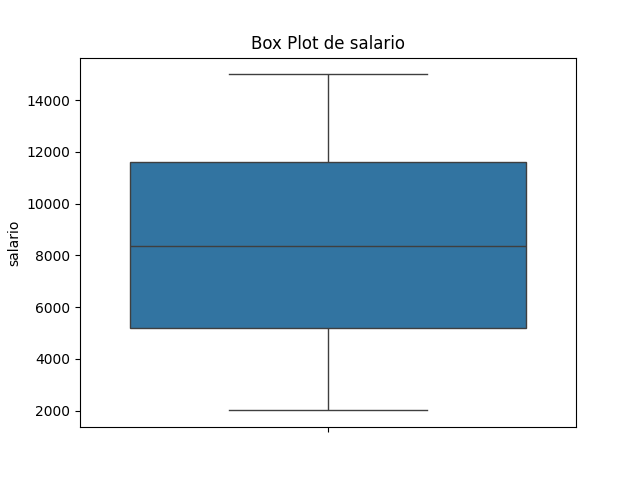


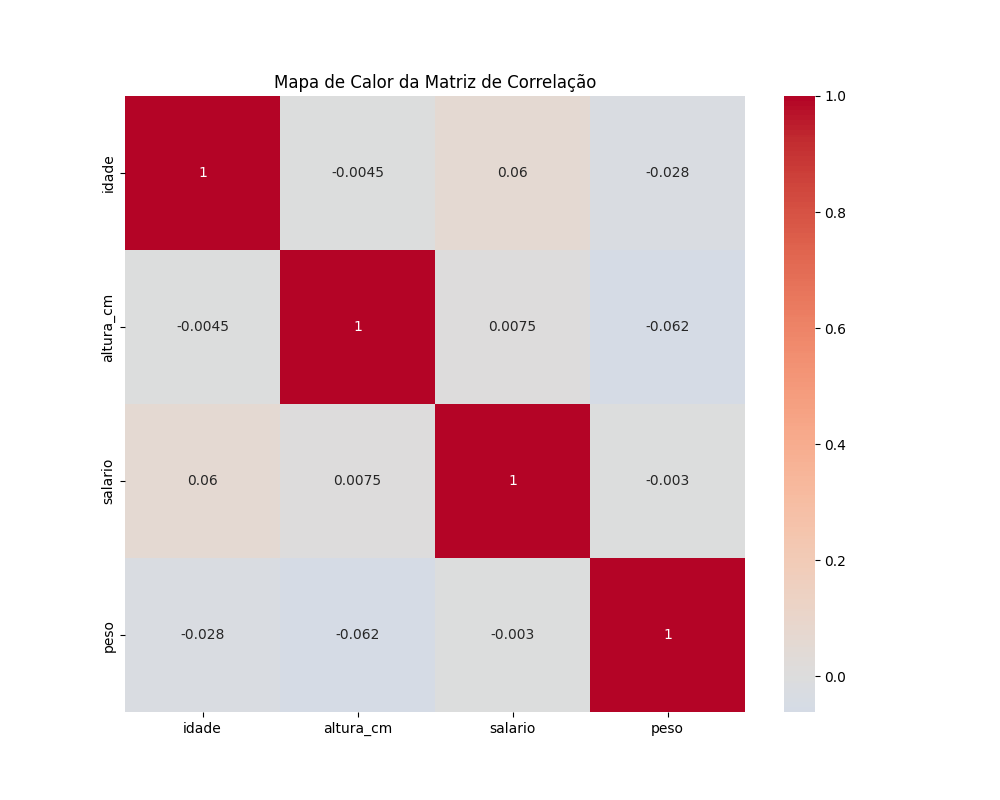
**Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente**

****

****

****

****

### **Descrição:**

* **Histogramas, Box Plots, Gráficos de Dispersão e Mapas de Calor:** Utilizamos seaborn e matplotlib para criar diversas visualizações gráficas, que ajudam na interpretação dos dados.

# **Parte 4: Integração de Dados, Valores Inconsistentes e Redundância de Dados**

### **4.1 Integração de Dados**

* **Descrição:** Assumimos que os dados já estão integrados. Em um cenário real, descreveríamos a necessidade de integrar dados de múltiplas fontes, como combinar dados de diferentes sistemas ou departamentos.

## **4.2 Identificação e Correção de Valores Inconsistentes**

### **Funções e Métodos:**

* DataFrame.replace(): Substitui valores inconsistentes.
* DataFrame.drop\_duplicates(): Remove registros duplicados.

### **Código:**



### **Descrição:**

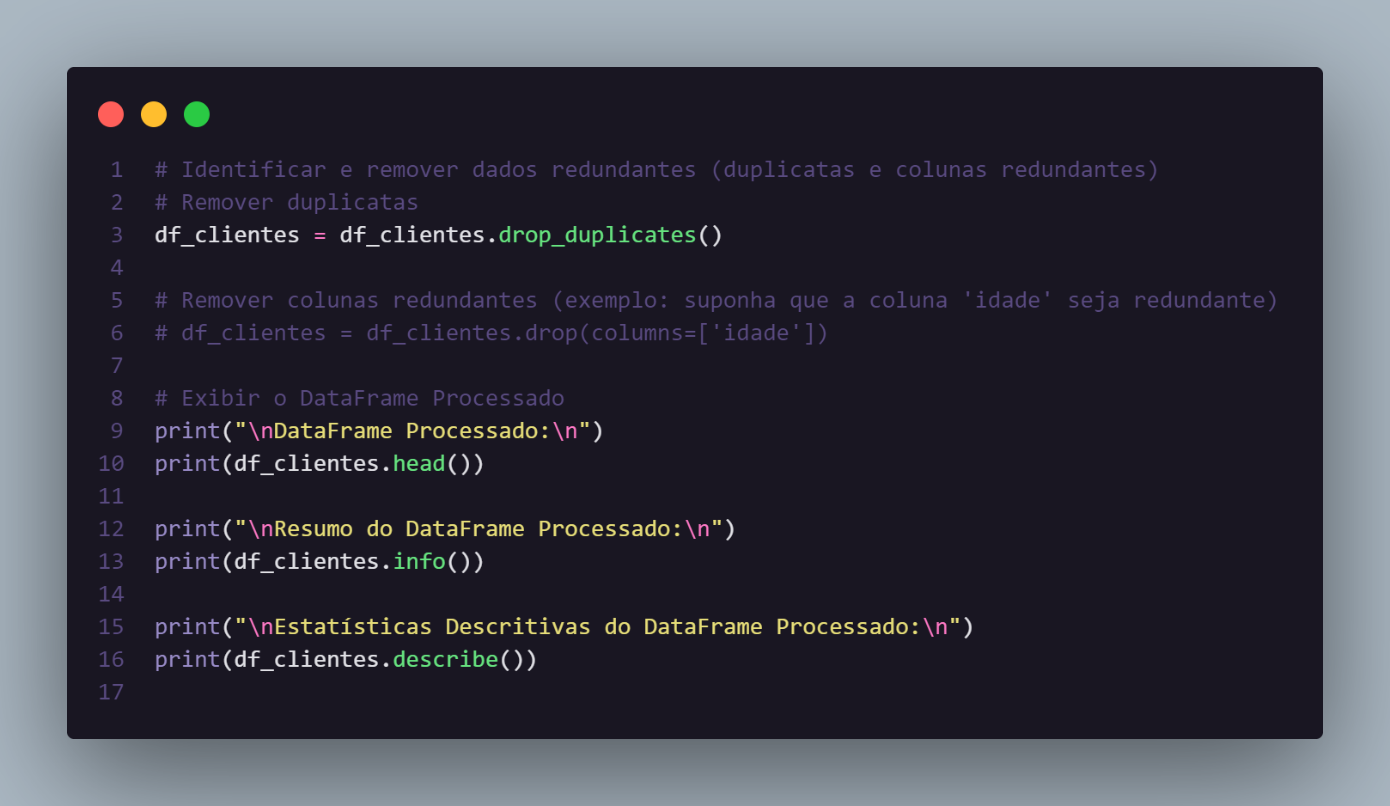
* **Correção de Valores Inconsistentes:** Removemos registros com valores inconsistentes em atributos como idade, altura, sexo, salário e score bom pagador.

### **4.3 Identificação e Remoção de Redundâncias**

### **Funções e Métodos:**

* DataFrame.drop\_duplicates(): Remove registros duplicados.
* DataFrame.drop(): Remove colunas redundantes.

### **Código:**



### **Descrição:**

* **Remoção de Redundâncias:** Removemos registros duplicados e colunas redundantes, garantindo a consistência dos dados.

# **Parte 5: Transformação de Dados**

### **5.1 Normalização de Dados**

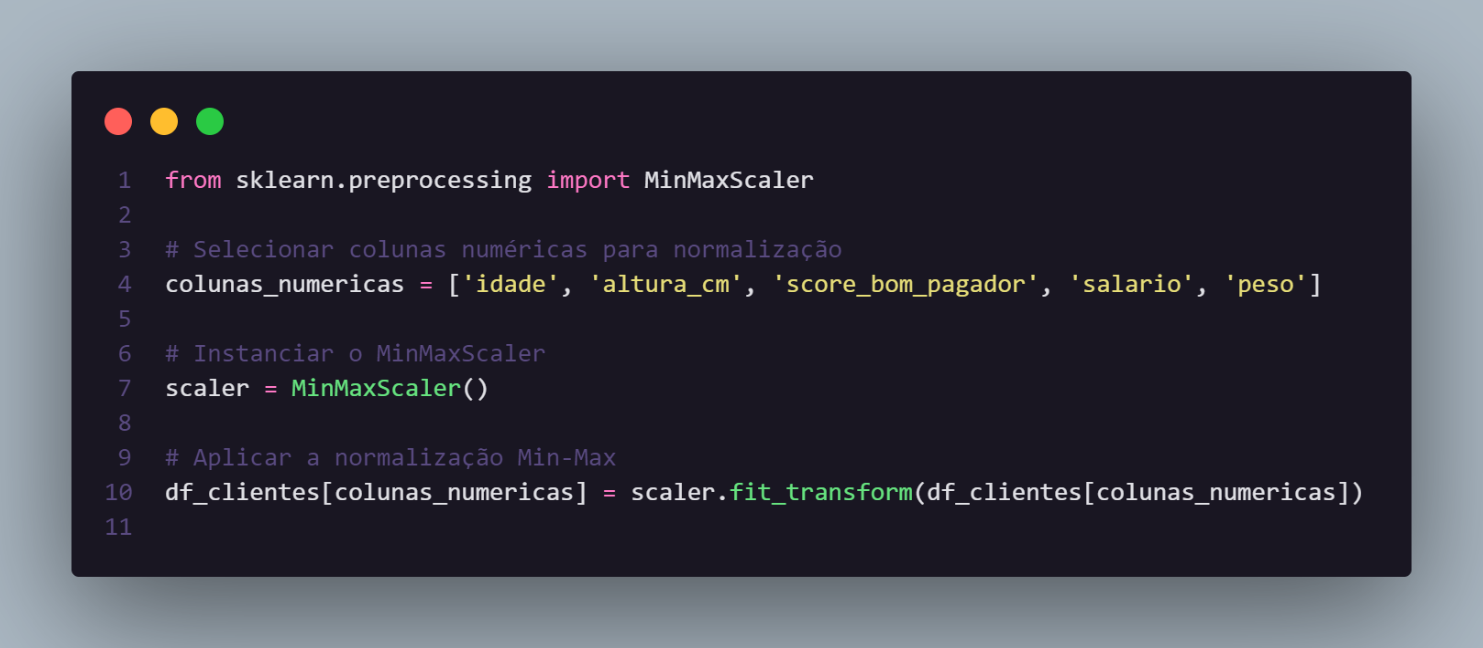
### **Bibliotecas Utilizadas:**

* sklearn.preprocessing: Para normalização e codificação dos dados.

### **Funções e Métodos:**

* MinMaxScaler(): Aplica a normalização Min-Max nos dados.

### **Código:**



### **Descrição:**

* **Normalização de Dados:** Utilizamos MinMaxScaler para normalizar os atributos numéricos no intervalo [0, 1].

### **5.2 Codificação de Dados Categóricos**

### **Funções e Métodos:**

* OneHotEncoder(): Aplica a codificação One-Hot nos dados categóricos.

### **Código:**



### **Descrição:**

* **Codificação de Dados Categóricos:** Utilizamos OneHotEncoder para transformar atributos categóricos em variáveis dummy, permitindo seu uso em modelos de aprendizado de máquina.

# **Conclusão**

Neste projeto, aplicamos uma abordagem abrangente para a análise e transformação de uma base de dados de clientes, utilizando uma variedade de técnicas estatísticas e computacionais. As etapas seguidas incluíram a limpeza e preparação dos dados, o cálculo de estatísticas descritivas, a análise de correlação, a visualização gráfica dos dados, a correção de valores inconsistentes, a remoção de redundâncias e a transformação de dados para fins de modelagem.

Os principais resultados obtidos incluem:

* **Limpeza e Preparação de Dados**: Conseguimos identificar e corrigir valores inconsistentes e duplicados, garantindo a qualidade dos dados para análises subsequentes.
* **Estatísticas Descritivas**: Foram calculadas medidas centrais (média, mediana, moda) e medidas de dispersão (variância, desvio-padrão, amplitude), fornecendo uma visão clara das características dos dados.
* **Análise de Correlação**: Identificamos relações significativas entre diferentes atributos numéricos, o que pode guiar futuras análises e decisões baseadas em dados.
* **Visualizações Gráficas**: Criamos diversas representações visuais, como histogramas, box plots, gráficos de dispersão e mapas de calor, que ajudaram a interpretar os dados de maneira intuitiva e informativa.
* **Transformação de Dados**: Normalizamos os dados numéricos e codificamos os atributos categóricos, tornando os dados prontos para uso em modelos de aprendizado de máquina e outras análises avançadas.

Este projeto demonstrou a importância de um processo bem-estruturado de análise e transformação de dados para extrair insights valiosos e garantir a qualidade dos dados. As técnicas aplicadas aqui são fundamentais para qualquer análise de dados robusta e servem como base para futuras análises mais complexas e modelagens preditivas.