

Ficha de desafio do Projeto Final

Disciplina: Aquisição, Pré-processamento e Exploração de Dados

Curso: Pós-graduação em Ciência de Dados

Cenário: Indicadores Socioeconômicos e Votação Brasil 2022

Integrantes: A. Cristiane R. Lima, Claudio Sampaio, Felipe Botero e José Henrique

Junho de 2025

Definição do Problema e Perguntas Analíticas

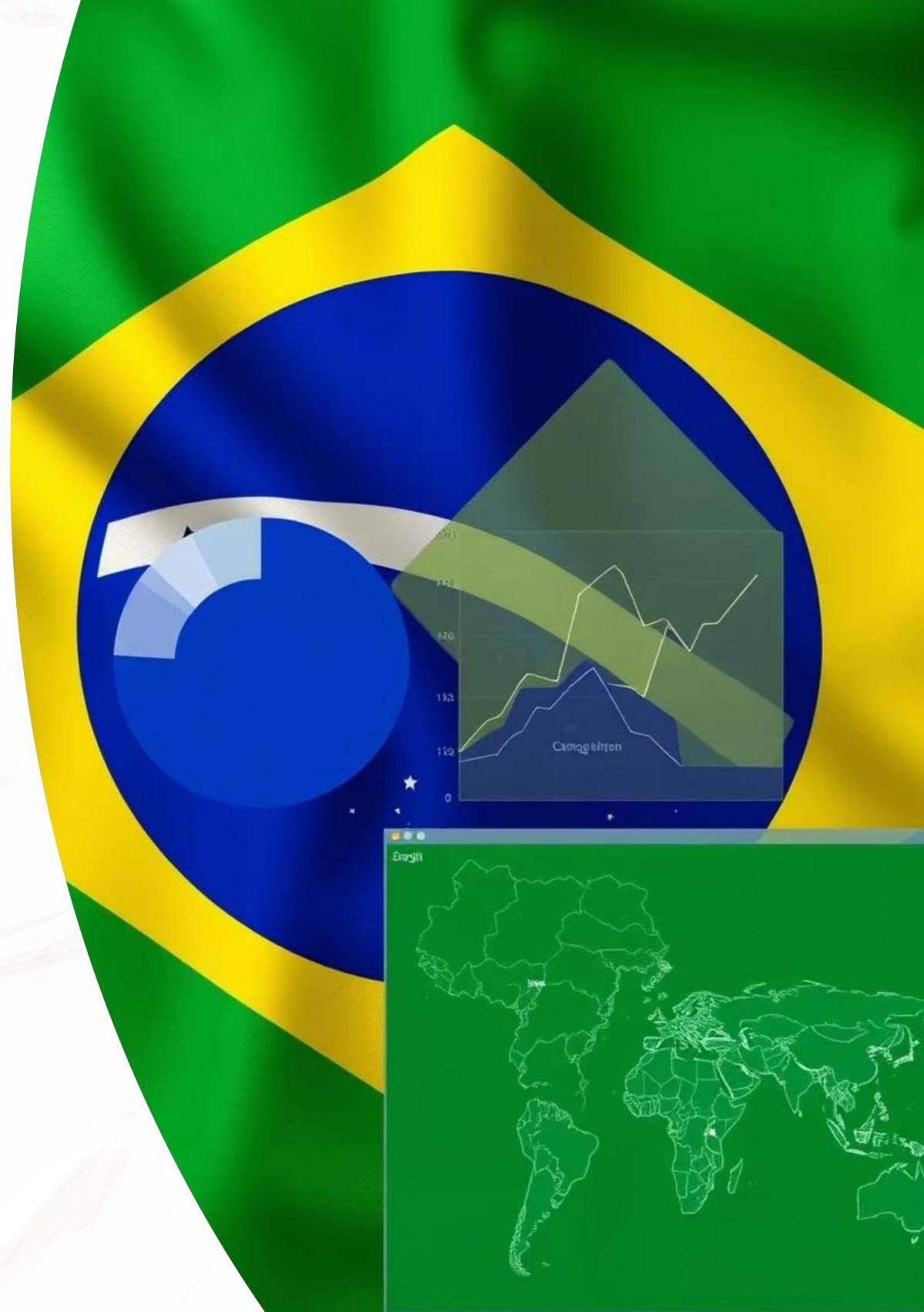
Problema Central

Existe relação entre indicadores socioeconômicos dos estados brasileiros e padrões de votação ideológica nas eleições de 2022?

Perguntas Analíticas

- Estados com maior IDHM votaram mais em partidos de direita, esquerda ou centro?
- Existe correlação entre analfabetismo e votos em um bloco ideológico específico?
- PIB per capita influencia a ideologia predominante dos votos por estado?

A pesquisa busca desvendar a complexa interação entre o desenvolvimento socioeconômico e as escolhas eleitorais, utilizando dados de 2022 para traçar um panorama das tendências políticas no Brasil.



Visão Geral das Fontes dos Dados

Resultados de Votação

Votos nominais válidos do 1º turno de 2022, agrupados por estado e ideologia partidária.

Dados Socioeconômicos

Indicadores por estado de 2022, incluindo PIB, IDHM, esperança de vida e mais.

Justificativas

Todas as fontes de dados utilizadas neste projeto são de abrangência nacional por Unidade de Federação, públicas e gratuitas, promovendo a transparência e a replicabilidade da análise.

Consolidar informações cruciais para entender as relações entre o perfil socioeconômico das UFs e o comportamento eleitoral resulta em um arquivo CSV pronto para análises estatísticas e modelos de Machine Learning.





Coleta e Justificativa dos Dados

Fontes dos Dados Brutos



IBGE

Dados de Contas Regionais 2022, PNAD Contínua (Gini, Analfabetismo) e Projeções Demográficas (Esperança de vida, Mortalidade infantil).



TSE

Resultados Eleitorais 2022, especificamente os votos nominais válidos, garantindo a precisão dos dados de votação.



Limitações

A análise se restringe a dados de 2022, sem granularidade por município ou perfil sociodemográfico individual.

Pré-processamento de Dados

Preparação dos Dados: Indicadores Socioeconômicos

A primeira etapa envolve o carregamento e a padronização dos indicadores socioeconômicos. Isso inclui a leitura do arquivo de indicadores por UF e o renomeamento de colunas para maior clareza e consistência.

Além disso, são realizadas verificações para completar dados ausentes, como o Índice de Gini e a porcentagem de analfabetismo, utilizando arquivos CSV dedicados para essas informações, garantindo um dataset completo e robusto.



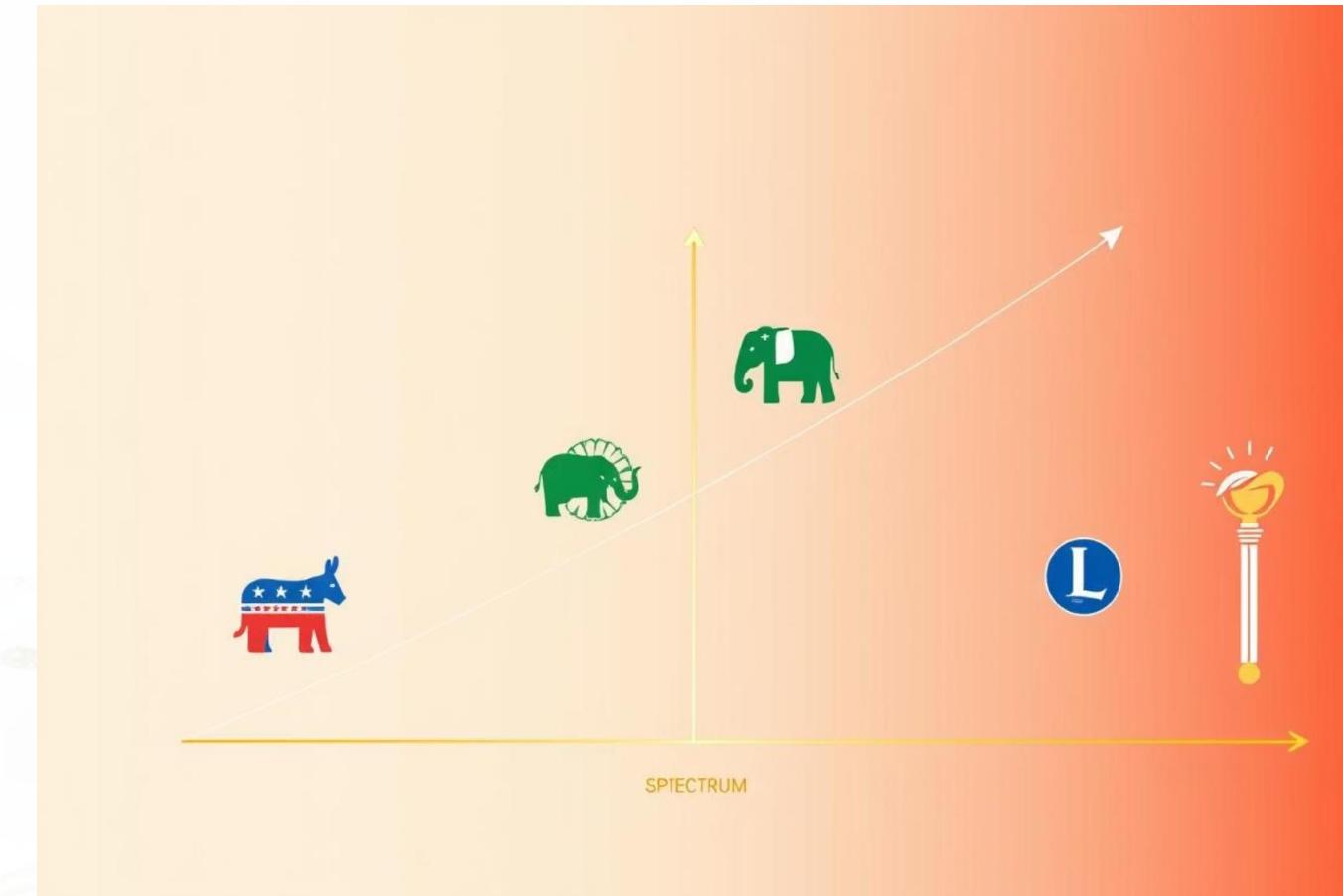
Este processo assegura que todos os indicadores estejam no formato correto e que não haja lacunas significativas nos dados antes da combinação com os resultados de votação.

Pré-processamento de Dados

Carregamento e Atribuição de Ideologia Partidária

Os votos nominais válidos são carregados de um arquivo CSV, filtrando-se apenas os dados do 1º turno e votos válidos. Uma etapa crucial é o mapeamento de cada partido político à sua respectiva ideologia (esquerda, centro-esquerda, centro, centro-direita, direita ou outros).

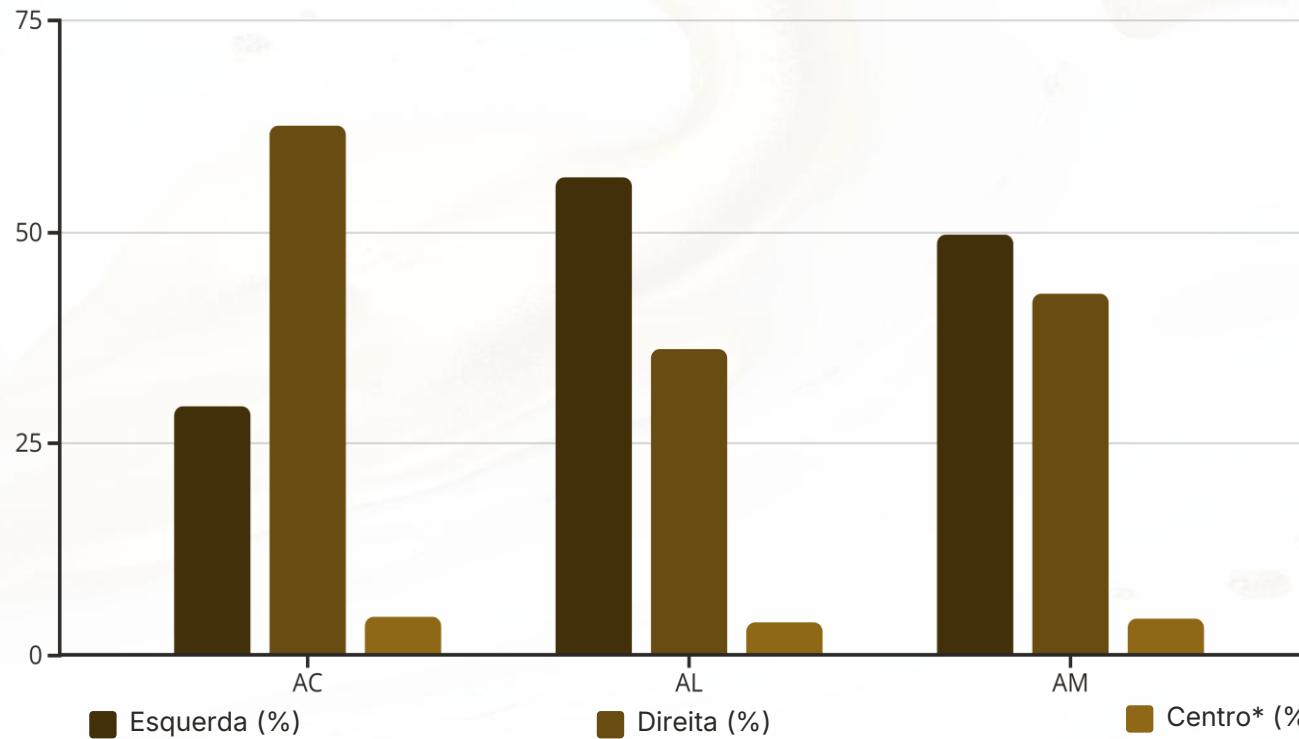
Este mapeamento permite agrupar os votos por blocos ideológicos, facilitando a análise do comportamento eleitoral em relação aos indicadores socioeconômicos.



A precisão neste mapeamento é fundamental para as análises subsequentes, garantindo que os resultados reflitam as tendências ideológicas de votação em cada UF.

Pré-processamento de Dados

Aggregação de Votos por UF e Ideologia



Após o mapeamento, os votos são agregados por UF e ideologia, somando-se os votos nominais válidos para cada bloco ideológico

Em seguida, é calculada a participação percentual de cada ideologia no total de votos de cada UF.

Essa agregação permite uma visão clara da distribuição ideológica dos votos em cada estado, essencial para a combinação final dos dados.

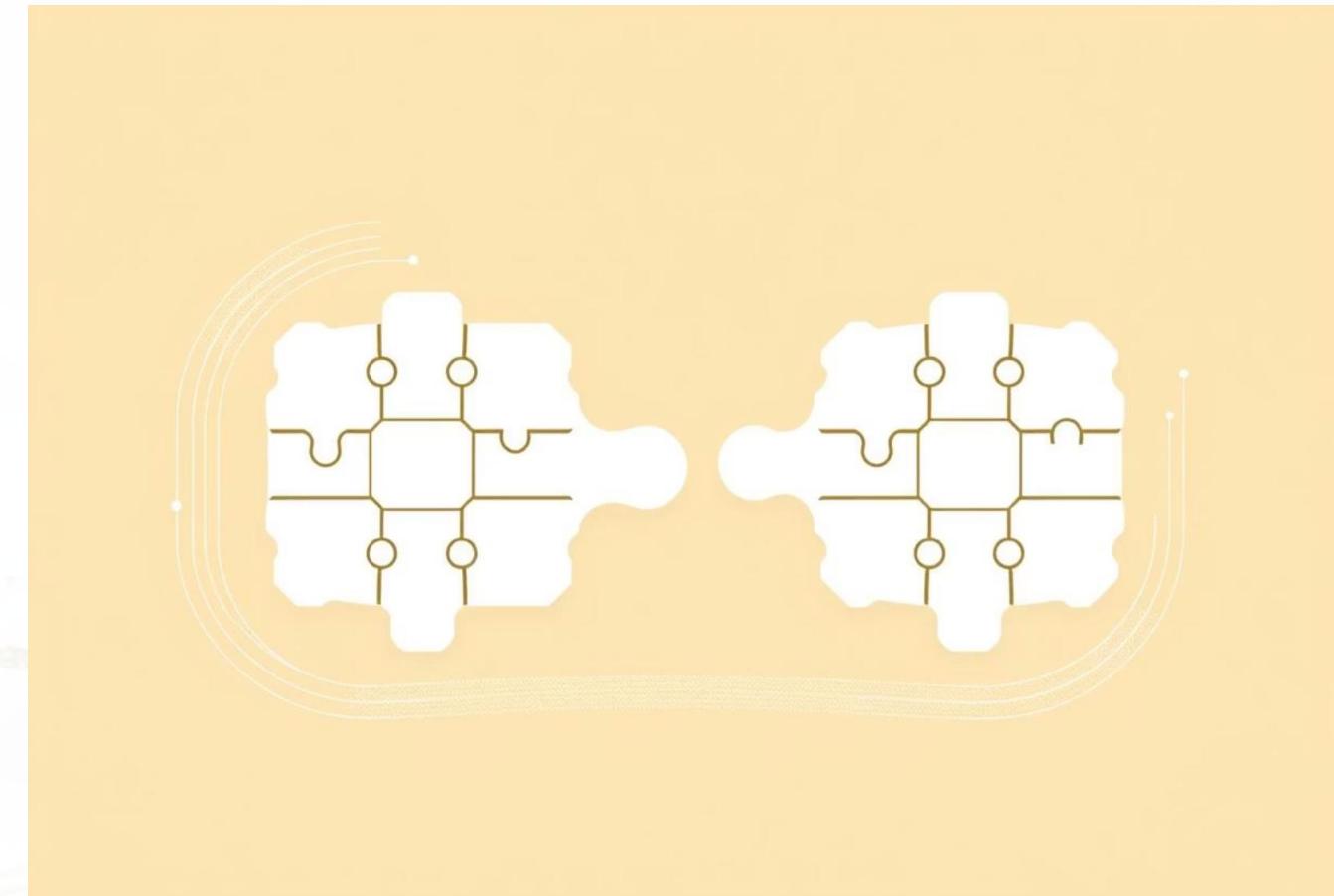
* Para simplificar a imagem, Centro-esquerda, Centro-direita e Centro estão informados como Centro

Pré-processamento de Dados

Combinação Final dos Dados

A etapa final do processo é a combinação dos indicadores socioeconômicos com os dados de votação agregados. Isso é feito através da junção dos dois datasets, utilizando a Unidade da Federação (UF) como chave comum.

O resultado é um dataset unificado que contém tanto as informações socioeconômicas quanto a distribuição percentual dos votos por ideologia para cada UF do Brasil.



Este dataset combinado é a base para futuras análises, permitindo explorar as relações entre as características socioeconômicas e os padrões de votação em nível estadual.

Pré-processamento de Dados

Salvando o Dataset Combinado

Exportação para CSV

O dataset unificado é salvo em um arquivo CSV, nomeado "uf_votos_ideologia_socioeco_2022.csv", garantindo a persistência dos dados.

Pronto para Análise

O arquivo CSV está agora disponível para diversas aplicações, como análises estatísticas, visualizações de dados e modelos de Machine Learning.

Flexibilidade

Este dataset oferece uma base sólida para pesquisadores e analistas explorarem as complexas interações entre fatores socioeconômicos e resultados eleitorais no Brasil.

A conclusão deste processo fornece uma base valiosa para a compreensão do cenário político-socioeconômico brasileiro de 2022.

Pré-processamento de Dados: Taxa de Analfabetismo

Análise da Taxa de Analfabetismo no Brasil

Esta etapa detalha o processo de coleta, extração e análise de dados sobre a taxa de analfabetismo por Unidade Federativa no Brasil, desde a importação de bibliotecas até a criação e salvamento de um DataFrame, destacando as etapas cruciais para a obtenção de insights valiosos.



Preparação do Ambiente e Definição de Variáveis

Importação de Bibliotecas Essenciais

Para iniciar o projeto de web scraping e análise de dados, importam-se as bibliotecas necessárias: **requests**, para fazer requisições HTTP; **BeautifulSoup**, para parsing HTML; **pandas**, para manipulação de dados; e **pathlib**, para gerenciar caminhos de arquivos.

```
import requestsfrom bs4 import BeautifulSoupimport pandas as pdfrom pathlib import Path
```

Declaração de URL e Caminho de Saída

Define-se a URL da fonte de dados, que é o site do IPEAData, e o caminho para o arquivo CSV de saída. Isso garante que os dados sejam baixados da fonte correta e salvos no local desejado.

```
URL =  
"https://www.ipeadata.gov.br/ExibeSerieR.aspx?stub=1&  
serid=2096726409&MIND"OUT_CSV =  
Path("taxa_analfabetismo_2022_por_uf.csv")
```



Pré-processamento de Dados: Taxa de Analfabetismo

Realizando a Requisição HTTP

A primeira etapa para obter os dados é fazer uma requisição HTTP para a URL especificada. Utiliza-se a biblioteca `requests` a fim de acessar o conteúdo da página web, simulando um navegador para evitar bloqueios.

Requisição GET

Envia-se uma requisição GET para a URL do IPEAData, incluindo um cabeçalho 'User-Agent' para imitar um navegador web.
Isso permite acessar o HTML da página.

Conteúdo HTML

O texto retornado pela requisição é o conteúdo HTML completo da página, que é posteriormente analisado para extrair os dados relevantes.

Pré-processamento de Dados: Taxa de Analfabetismo

Extração dos Dados Principais

Após obter o HTML da página, utiliza-se a biblioteca **BeautifulSoup** para *parsear* o conteúdo e localizar os elementos da tabela que contêm os dados de analfabetismo.

Parsing com BeautifulSoup

O HTML é processado pelo **BeautifulSoup** usando o parser "lxml", que é eficiente para grandes documentos HTML.

Isso cria um objeto que permite navegar pela estrutura da página.

Seleção das Linhas da Tabela

São selecionadas as linhas da tabela que contêm os dados de interesse, identificadas pelo ID "grd DXMainTable" e pelo padrão "grd DXDataRow" para as linhas de dados.

```
soup = BeautifulSoup(html, "lxml")rows = soup.select("table#grd DXMainTable tr[id^=grd DXDataRow]")
```

Conversão para Lista de Dicionários

As linhas da tabela extraídas são iteradas para converter cada conjunto de dados em um formato estruturado, como uma lista de dicionários.

Este passo é crucial para a posterior criação de um DataFrame.

Iteração e Limpeza

Cada linha da tabela é percorrida, e o texto de cada célula (td) é extraído, removendo espaços em branco e caracteres indesejados como "\xa0".

Estruturação dos Dados

Os valores extraídos são atribuídos a variáveis como sigla, estado e os valores dos anos (1991, 2000, 2010, 2022), preparando-os para serem adicionados aos registros.

Tratamento de Erros

É importante notar que um erro de "ValueError: too many values to unpack" pode ocorrer se o número de colunas esperadas não corresponder ao número de colunas encontradas. Isso exige depuração na extração.

Criação do DataFrame

Com os dados estruturados em uma lista de dicionários, o próximo passo é criar um DataFrame do pandas. Este DataFrame é a estrutura central para análise e manipulação dos dados.

Construção do DataFrame

A lista de dicionários é passada para o construtor `pd.DataFrame()` para criar o DataFrame. Cada dicionário se torna uma linha, e as chaves dos dicionários se tornam as colunas.

Definição do Índice

A coluna "UF" (Unidade Federativa) é definida como o índice do DataFrame. Isso facilita a consulta e manipulação dos dados por estado.

```
df = pd.DataFrame(records).set_index("UF")
```

Pré-processamento de Dados: Taxa de Analfabetismo

Salvando os Dados Processados

Após a criação e organização do DataFrame, é fundamental salvar os dados para uso futuro. Isso garante que o trabalho de extração e processamento não precise ser repetido.



Seleção de Colunas

Selecionam-se apenas as colunas "Estado" e "2022" para o arquivo de saída, focando nos dados mais relevantes para a análise atual.

Ordenação dos Dados

Os dados são ordenados pela coluna "2022" (taxa de analfabetismo), permitindo uma visualização rápida dos estados com as menores ou maiores taxas.



Exportação para CSV

O DataFrame é exportado para um arquivo CSV chamado "raw/analfabetismo_uf_2022.csv", garantindo que os dados estejam disponíveis em um formato padrão.



Análise do Coeficiente de Gini no Brasil

Esta etapa detalha o processo de coleta e análise de dados do Coeficiente de Gini para os estados brasileiros, abrangendo o período de 2012 a 2024.

Exploraram-se os passos de web scraping, extração de dados e estruturação para análise, culminando na visualização das disparidades regionais.

A análise do Coeficiente de Gini por estado revela as disparidades na distribuição de renda no Brasil, fornecendo insights valiosos para políticas públicas.

8ZI. ZAdZa Hl.

A sesdla youjumies, polised otase pecasss and toal vohitilly anccome to the praller, poveras, anlldior boting, sites arached imurcipom cart.

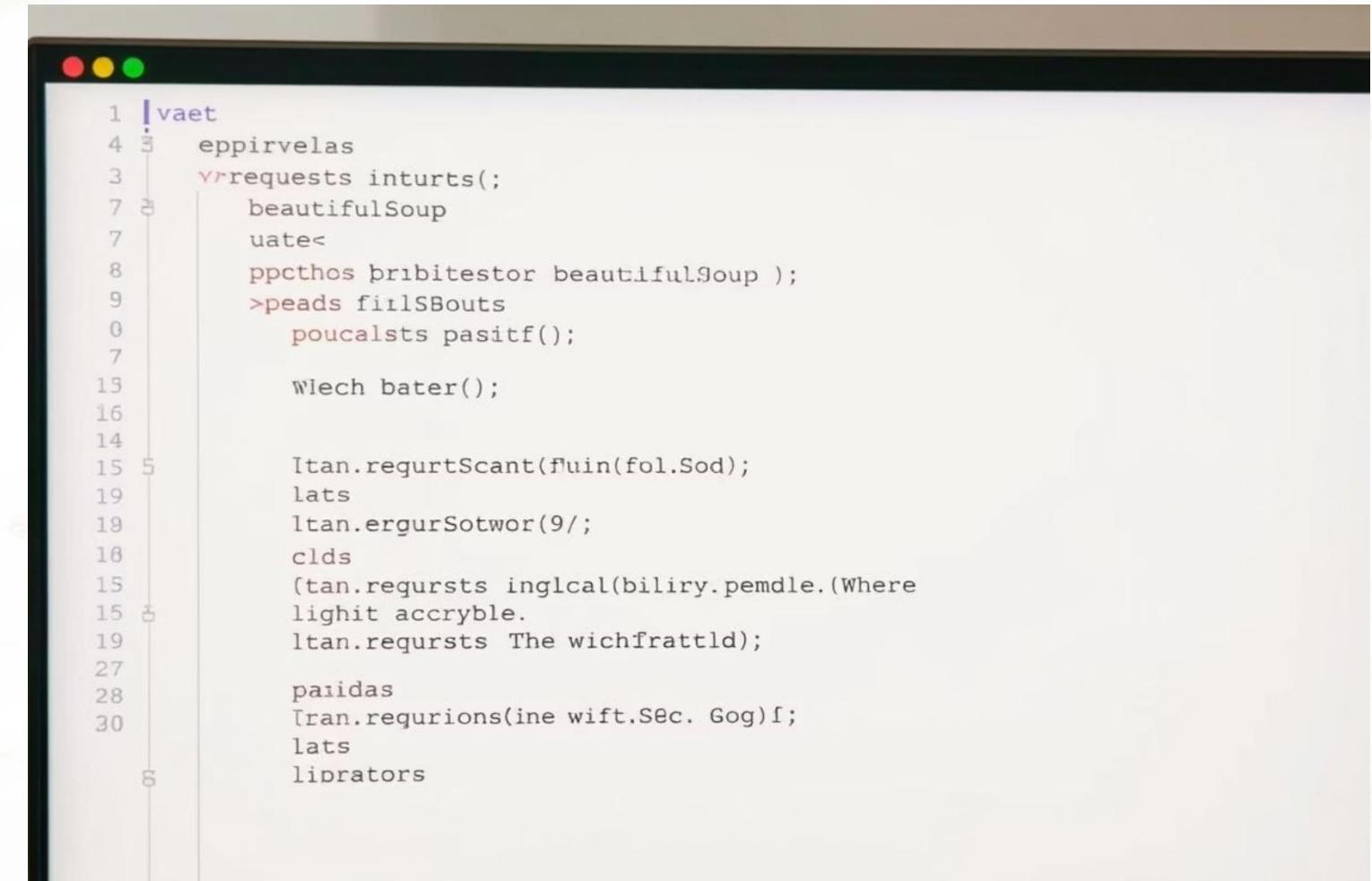


Preparação do Ambiente e Definição de Variáveis

A primeira etapa do projeto envolve a importação das bibliotecas necessárias para a requisição HTTP, parsing de HTML e manipulação de dados. Utilizamos `requests` para obter o conteúdo da web, `BeautifulSoup` para analisar o HTML e `pandas` para gerenciar os dados em formato de `DataFrame`.

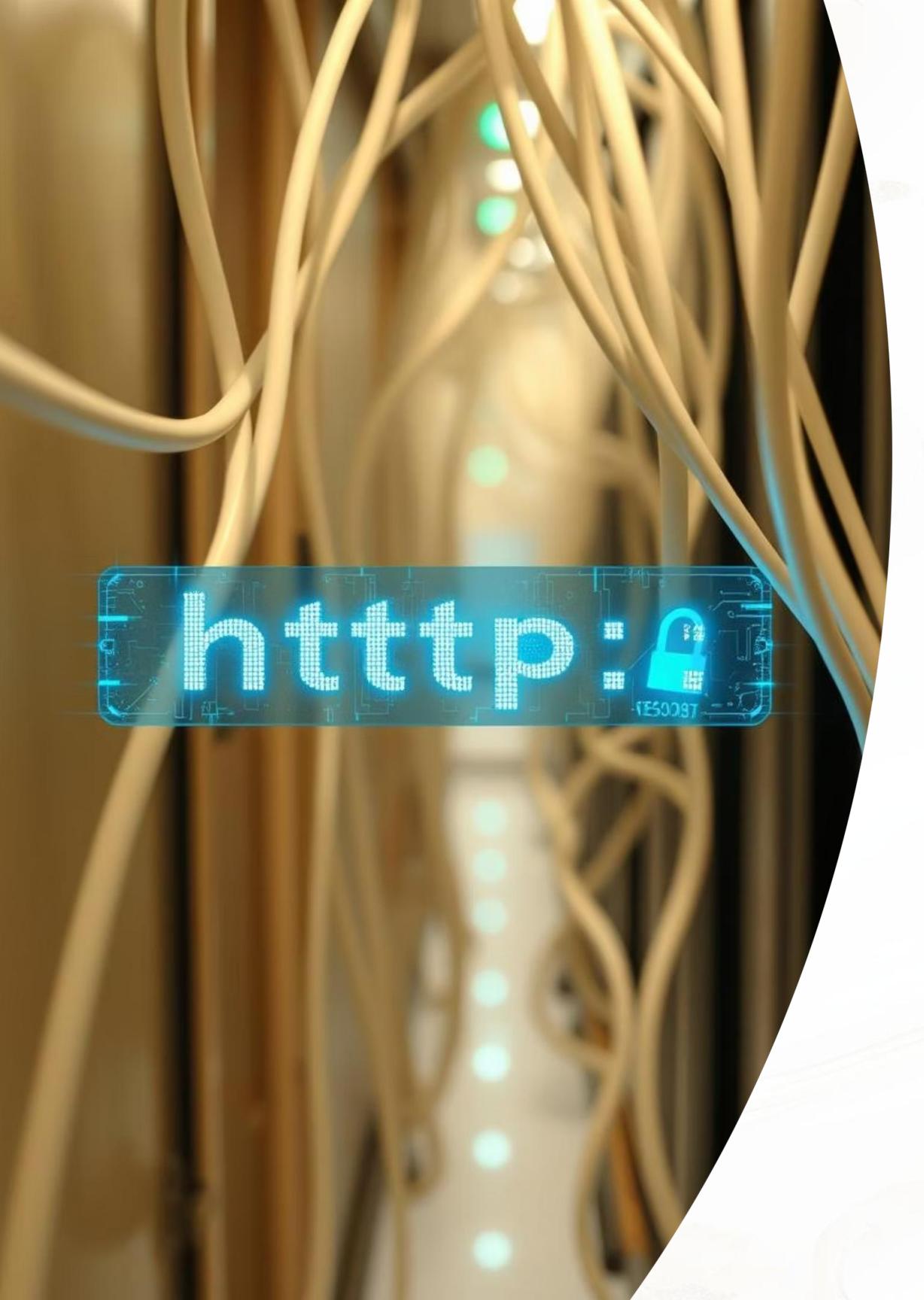
Além disso, definem-se a URL da fonte de dados (IPEADATA) e o caminho para o arquivo CSV de saída, garantindo que os dados extraídos sejam armazenados de forma organizada para análises futuras.

```
import requestsfrom bs4 import BeautifulSoupimport pandas as pdfrom pathlib import PathURL = "https://www.ipeadata.gov.br/ExibeSerieR.aspx?MAXDATA=2025&MIN DATA=2012&TN"OUT_CSV = Path("indice_de_gini.csv")
```



The screenshot shows a terminal window with a dark theme. The code is displayed in white text. The code imports requests, BeautifulSoup, and pandas. It defines a URL for IPEADATA and sets OUT_CSV to a path for a CSV file named 'indice_de_gini.csv'.

```
1 | vaet
4 |     eppirvelas
3 |     vrequests inturts();
7 |     beautifulSoup
7 |     uate<
8 |     ppcthos pribitestor beautifulSoup );
9 |     >peads firSBouts
0 |     poucalsts pasitf();
7 |
13 |
16 |
14 |
15 |     ltan.regurtScant(fuin(fol.Sod);
19 |     lats
19 |     ltan.ergurSotwor(9/;
18 |     clds
15 |     (tan.reqrsts inglcal(biliry.pemdle.(Where
15 |     5 light accryble.
19 |     ltan.reqrsts The wichfrattld);
27 |
28 |
30 |     paiddas
|     Tran.requorions(ine wift.Sec. Gog)f;
|     lats
|     librators
```



Pré-processamento de Dados: Coeficiente de Gini

Realizando a Requisição HTTP

Obtenção do HTML

Para acessar os dados do Coeficiente de Gini, realiza-se uma requisição HTTP GET à URL do IPEADATA.

É crucial incluir um cabeçalho `User-Agent` para simular um navegador e evitar bloqueios por parte do servidor, garantindo o acesso ao conteúdo da página.

Conteúdo da Página

O resultado da requisição é o código HTML completo da página. Este conteúdo será a base para a próxima etapa, onde utilizaremos o BeautifulSoup para navegar e extrair as informações relevantes, focando na tabela que contém os dados do Gini.

Extração dos Dados Principais da Tabela

Com o HTML em mãos, o próximo passo é *parseá-lo* usando BeautifulSoup para localizar a tabela que contém os dados do Coeficiente de Gini. Identifica-se a tabela pelo seu ID específico (grd_DXMainTable).

Após a seleção da tabela, extraem-se os cabeçalhos das colunas para identificar os anos e, em seguida, faz-se iteração sobre as linhas de dados para coletar as informações de cada estado.

Este processo é fundamental para estruturar os dados brutos para a conversão em um formato mais utilizável.



```
soup = BeautifulSoup(html, "lxml")table =
soup.select_one("table#grd_DXMainTable")header_cells =
table.select("tr#grd_DXHeadersRow0 td")[2:]anos =
[td.get_text(strip=True) for td in header_cells]rows =
table.select("tr[id^=grd_DXDataRow]")
```

Conversão dos Dados para Lista de Dicionários

Após a extração das linhas da tabela, converte-se cada linha em um dicionário, onde as chaves são os nomes das colunas (UF, Estado e os anos) e os valores são os dados correspondentes.

É crucial limpar e formatar os valores numéricos, substituindo vírgulas por pontos para garantir a correta conversão para tipo float.

Este formato de lista de dicionários é ideal para a criação de um DataFrame Pandas, facilitando a manipulação e análise posterior dos dados do Coeficiente de Gini por estado e ano.

Pré-processamento de Dados: Coeficiente de Gini

Estruturação dos Dados em DataFrame Pandas

Com a lista de dicionários pronta, o passo final é criar um DataFrame Pandas.

Define-se a coluna 'UF' como índice para facilitar a consulta e manipulação dos dados por estado.

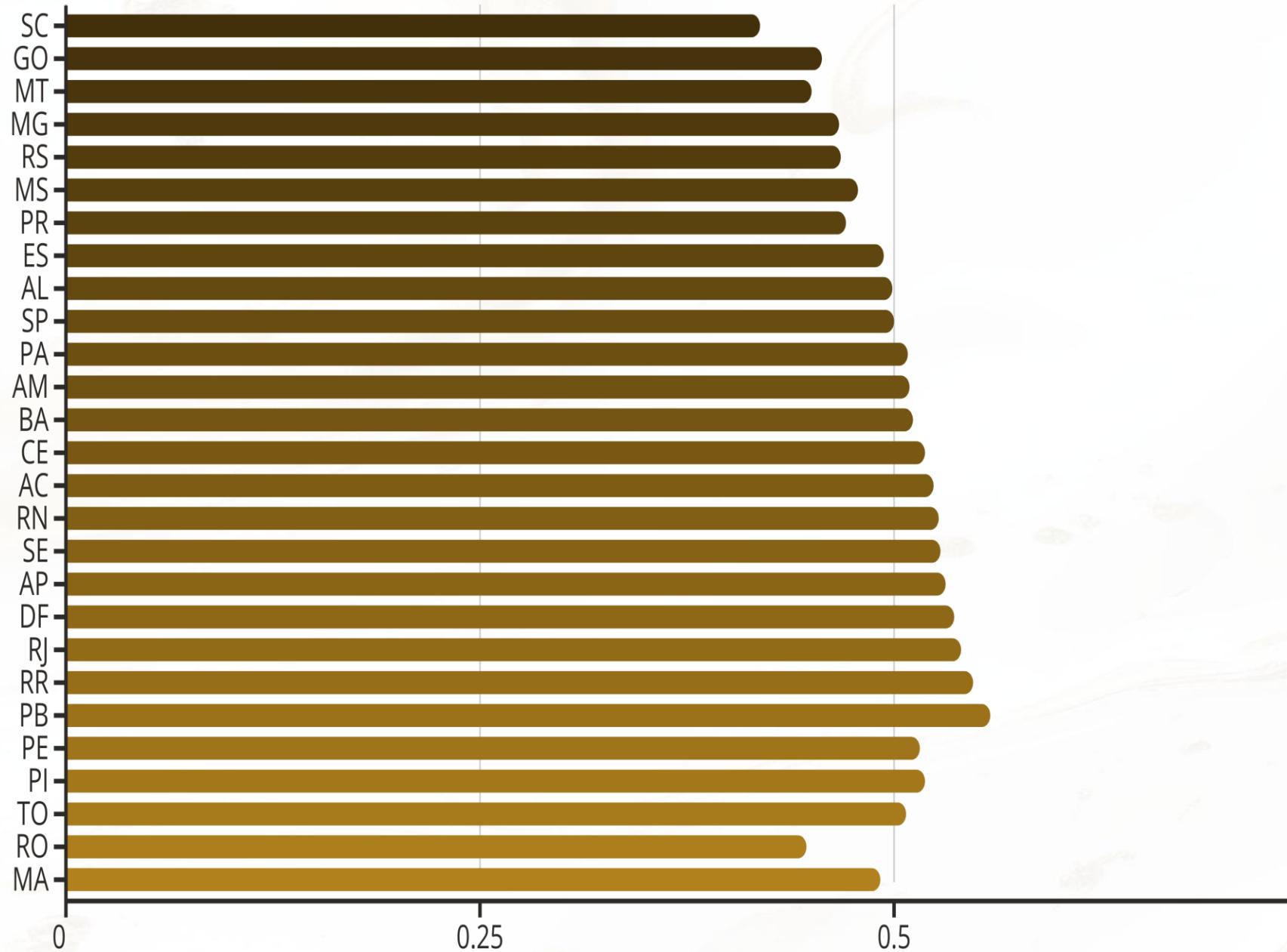
O DataFrame oferece uma estrutura tabular robusta, permitindo operações eficientes de filtragem, ordenação e análise estatística.

A visualização inicial do DataFrame (`df.head()`) confirma que os dados foram carregados corretamente, com as colunas de anos e os valores do Coeficiente de Gini prontos para serem explorados.

Data Frame									
1	Inteer	Colum A	Column D	yoiris C	Strings	Colum C	Column	Values	
1	9010	1340	1000	5707	200	2000	280		
2	0046	1158	1000	5000	300	3900	200		
3	9540	1300	1000	5032	180	3800	278		
3	8600	1100	1000	2067	180	2207	242		
4	1409	1340	1000	1000	160	2700	255		
5	1240	1139	1000	5532	300	3000	347		
5	2500	1489	1000	5982	160	3500	517		
16	5495	2180	4000	5052	190	3900	569		
11	5600	2380	2500	5000	150	3000	180		
25	3450	2455	5700	6637	180	3700	300		
13	3460	2389	2000	5502	180	3300	200		
20	5440	2338	2000	9000	188	3700	180		
35	5340	1330	3000	5000	160	3000	180		
31	3180	1359	3000	6002	180	3700	200		
32	3760	1100	3000	5251	100	3000	200		
31	3550	1399	3000	5002	100	3100	300		

```
df =  
pd.DataFrame(records).set_index("UF")print(df.head())
```

Análise do Coeficiente de Gini por Estado (2022)



Esta tabela e gráfico de barras horizontais apresentam o Coeficiente de Gini para cada estado brasileiro em 2022, ordenados do menor para o maior.

O Coeficiente de Gini é uma medida de desigualdade de renda, onde valores mais próximos de 0 indicam menor desigualdade e valores mais próximos de 1 indicam maior desigualdade.

Santa Catarina (SC) se destaca com o menor índice, enquanto Paraíba (PB) e Roraima (RR) apresentam índices mais elevados, refletindo as disparidades regionais na distribuição de renda.

Análise Exploratória de Dados: Composição Percentual de Votos

Análise da Composição Percentual de Votos por Bloco Ideológico no Brasil (2022)

Detalha-se nesta etapa a distribuição média de votos por bloco ideológico nos 27 estados brasileiros em 2022. Utiliza-se o dataset consolidado "uf_votos_ideologia_socioeco_2022.csv" para calcular e visualizar as médias percentuais de votos para cada ideologia.

Faz-se a análise através de tabelas e gráficos, oferecendo uma compreensão clara das tendências políticas no país e sobre a composição ideológica do eleitorado brasileiro.



Análise Exploratória de Dados: Votos por Ideologia

Metodologia e Carregamento de Dados

A análise começa com o carregamento do dataset "uf_votos_ideologia_socioeco_2022.csv".

Este arquivo contém informações cruciais sobre o percentual de votos por ideologia em cada Unidade Federativa (UF), além de dados socioeconômicos relevantes.

Utilizam-se bibliotecas como Pandas para manipulação de dados e Matplotlib para visualização.

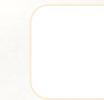
A detecção das colunas de percentual de votos por ideologia é um passo fundamental para garantir a precisão dos cálculos subsequentes.



Colunas de Voto Percentual Detectadas

Para a análise, identificam-se as colunas que representam o percentual de votos para cada bloco ideológico.

As colunas detectadas são:



Centro



Centro-Direita



Centro-Esquerda



Direita



Esquerda

Essa categorização permite uma análise detalhada da distribuição ideológica dos votos em todo o território nacional.

Média de Votos por Ideologia nas Unidades Federativas

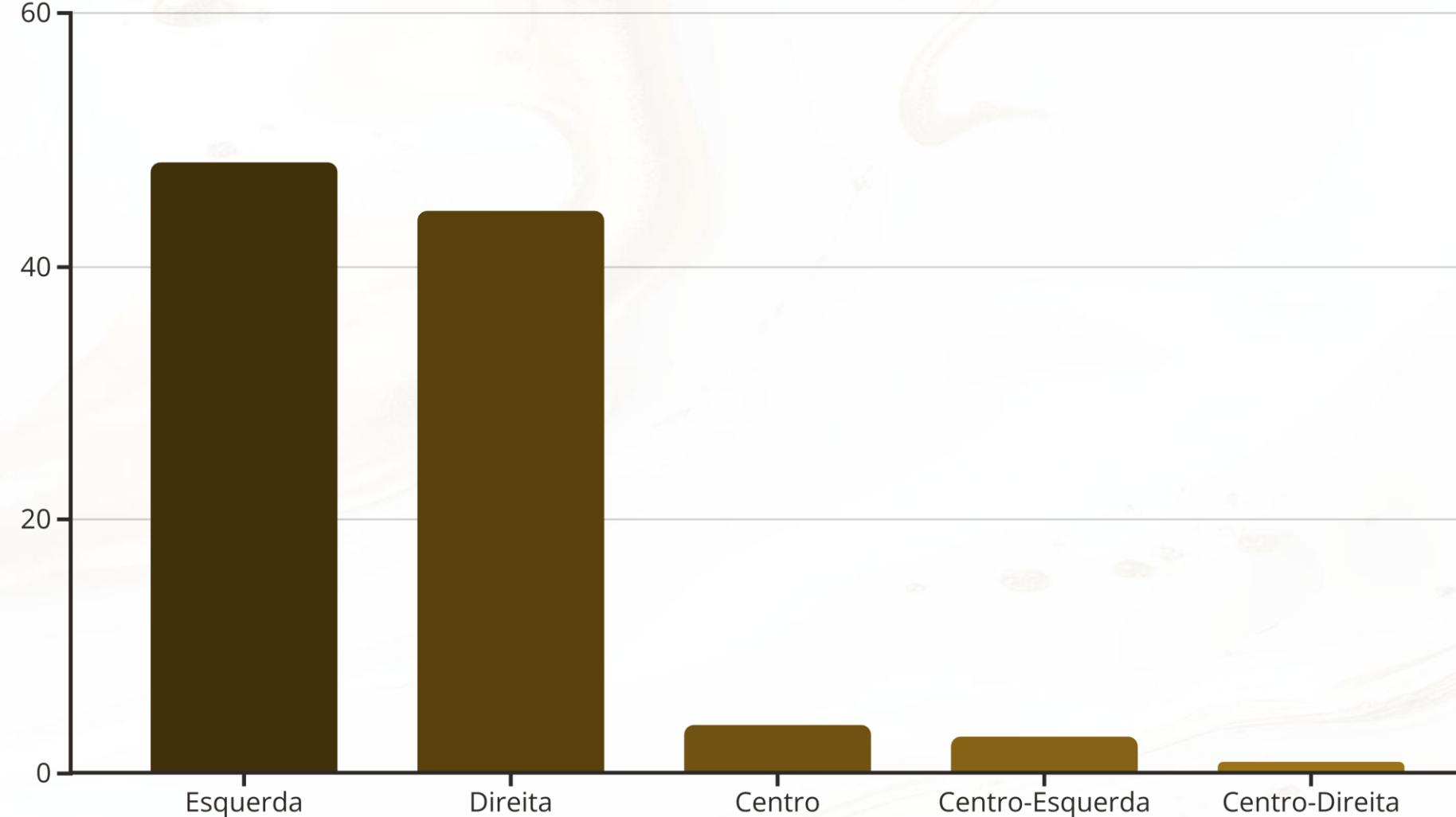
Calcula-se a média do percentual de votos para cada bloco ideológico entre as 27 Unidades Federativas (UF). Os resultados, ordenados de forma decrescente, são apresentados na tabela abaixo:

Esquerda	48.18
Direita	44.37
Centro	3.79
Centro-Esquerda	2.84
Centro-Direita	0.83

Estes dados revelam as tendências predominantes na média nacional de votos por ideologia, reforçando o clima de polarização reportado pela imprensa à época entre Esquerda e Direita.

Análise Exploratória de Dados: Votos por Ideologia

Gráfico de Barras: Média Nacional por Ideologia

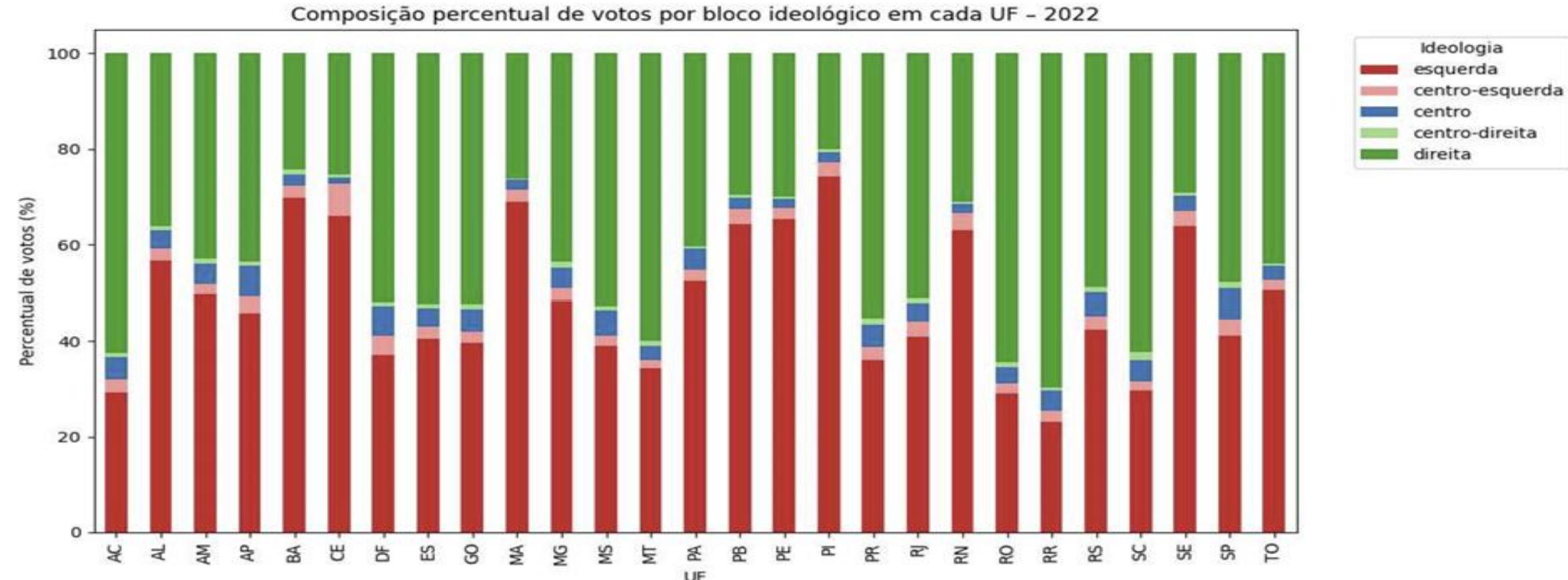


Este gráfico expõe a média nacional de votos por bloco ideológico em 2022, destacando a predominância da esquerda e direita, e a menor representatividade dos blocos de centro.

Análise Exploratória de Dados: Votos por UF

Análise da Distribuição de Votos por Unidade Federativa

Este gráfico de barras empilhadas ilustra a composição percentual de votos por bloco ideológico em cada Unidade Federativa (UF). Ele oferece uma visão detalhada de como as diferentes ideologias se distribuem regionalmente, permitindo identificar padrões e particularidades em cada estado.



A paleta de cores fixa para cada ideologia facilita a comparação visual entre as UFs, revelando a diversidade do cenário político brasileiro.

Claras Compreensões Regionais e Tendências

A análise da composição por UF revela que, embora existam tendências nacionais, cada estado possui suas próprias nuances ideológicas.

Alguns estados mostram uma forte inclinação para a esquerda (ex.: BA, MA, PI), enquanto outros se inclinam mais para a direita (ex.: AC, RO, RR, SC).

Os blocos de centro, centro-esquerda e centro-direita, embora com menor percentual médio, desempenham papéis importantes em certas regiões, influenciando o equilíbrio político local (ex: DF, RJ, SP).



Correlação de Pearson: Análise Dinâmica de Colunas Numéricas

Esta apresentação aborda a detecção dinâmica de colunas numéricas para o cálculo da Correlação de Pearson, resolvendo o erro KeyError em DataFrames. Exploram-se o carregamento de dados, a verificação de colunas, a conversão segura de tipos e a visualização das correlações.

Visão Geral do Projeto

1

Carregamento e Verificação

O processo inicia com o carregamento do dataset "uf_votos_ideologia_socioeco_2022.csv" e a verificação das colunas-alvo existentes.

2

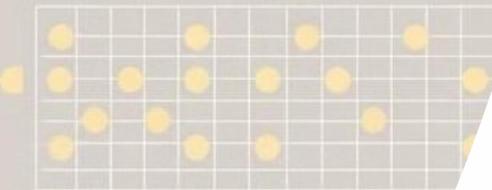
Conversão Segura de Dados

Strings com vírgula são convertidas para float de forma segura, garantindo a integridade dos dados numéricos.

3

Cálculo e Visualização

As correlações de Pearson são calculadas e visualizadas para insights sobre as relações entre as variáveis.



Preparação do Ambiente e Carregamento de Dados

Para iniciar a análise, importam-se as bibliotecas necessárias: pandas para manipulação de dados, matplotlib para visualização e pathlib para gerenciamento de caminhos de arquivo. O dataset é carregado e as primeiras linhas são exibidas para uma visão inicial.

```
import pandas as pdimport matplotlib.pyplot as pltfrom pathlib import Pathcsv_path =  
Path('data_lake/processed/uf_votos_ideologia_socioeco_2022.csv')df = pd.read_csv(csv_path)print('Colunas  
encontradas:', list(df.columns))df.head()
```

Colunas encontradas: ['UF', 'PIB_milhoes', 'PIB_per_capita', 'IDHM_2021', 'Esperanca_vida_anos', 'Mort_infantil_pmil', 'Gini', 'Analfabetismo_perc', 'centro', 'centro-direita', 'centro-esquerda', 'direita', 'esquerda', 'total_votos', 'centro_perc', 'centro-direita_perc', 'centro-esquerda_perc', 'direita_perc', 'esquerda_perc']

Detecção e Filtragem de Colunas Numéricas

Define-se uma lista de colunas numéricas alvo e filtram-se apenas aquelas que realmente existem no DataFrame. Isso garante que a análise seja realizada apenas com dados válidos, evitando erros de KeyError.

```
target_numeric = [  
    'PIB_per_capita', 'PIB_milhoes', 'IDHM_2021', 'Gini',  
    'Esperanca_vida_anos', 'Mort_infantil_pmil', 'Analfabetismo_perc',  
    'esquerda_perc', 'centro-esquerda_perc', 'centro_perc', 'centro-  
    direita_perc', 'direita_perc']numeric_cols = [c for c in  
target_numeric if c in df.columns]print('Colunas numéricas  
detectadas:', numeric_cols)
```

Colunas numéricas detectadas: ['PIB_per_capita', 'PIB_milhoes', 'IDHM_2021', 'Gini', 'Esperanca_vida_anos', 'Mort_infantil_pmil', 'Analfabetismo_perc', 'esquerda_perc', 'centro-esquerda_perc', 'centro_perc', 'centro-direita_perc', 'direita_perc']

MA	CI	Acn	NE	F	CL
28,13905	12800		113600	8000	113575
20/13642	2560		2565	2080	2630
13,13550	2450		2671	3180	3250
13,15350	3550		2375	2500	3230
10,13550	4600		2278	4670	2870
13,19250	4600		2290	2600	2350
13,15250	6630		2290	4680	3895
21,13590	6630		2590	2790	2990
12,19540	4550		2770	2760	2590
12,19950	5000		1600	2490	1500
13,13490	3400		2370	2770	2790
12,13500	3740		2720	2630	2630
13,13500	3750		1800	2930	2930
13,13250	3660		1500	2350	2350
12,18250	2900		1570	3989	2830
2,13350	9530		1600	2850	2850
1,18550	8630		2077		
15350	8900		3675		
1550	6930				
150	6930				
0	9530				
	9360				
	4330				
	9930				
	9570				



Conversão de Dados e Criação de Variáveis de Voto

Para garantir a correta manipulação dos dados, convertem-se strings com vírgulas para o formato float, removendo também o símbolo de porcentagem. Em seguida, criam-se as variáveis 'left_vote' e 'right_vote' combinando as porcentagens de votos de esquerda/centro-esquerda e direita/centro-direita, respectivamente.

```
for col in numeric_cols:    df[col] =  
(df[col].astype(str).str.replace(',', '.',  
regex=False)                  .str.replace('%',  
'', regex=False))    df[col] = pd.to_numeric(df[col],  
errors='coerce')df[numeric_cols].dtypes
```

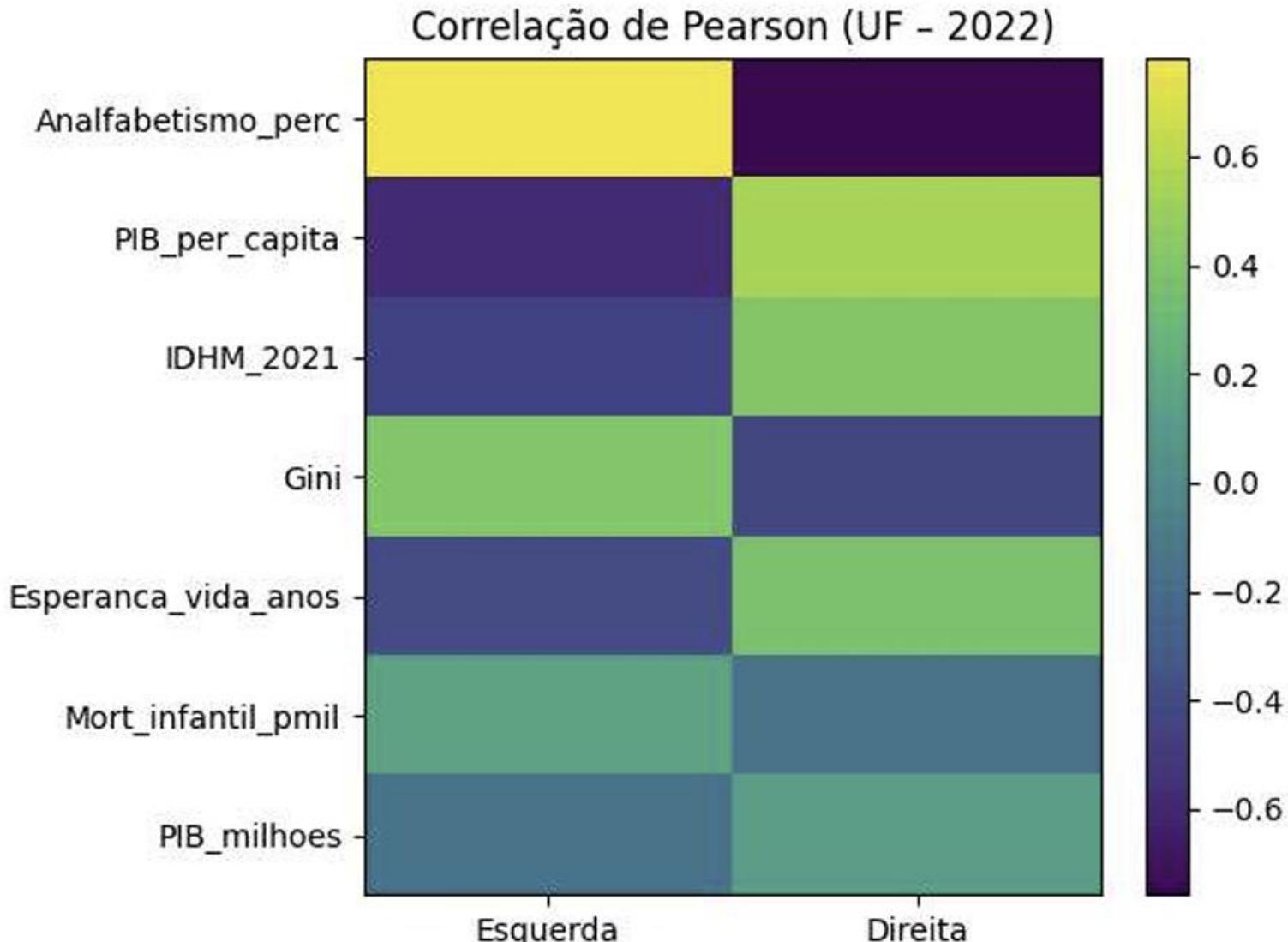
```
df['left_vote'] = df.get('esquerda_perc', 0) +  
df.get('centro-esquerda_perc', 0)df['right_vote'] =  
df.get('direita_perc', 0) + df.get('centro-  
direita_perc', 0)
```

Cálculo das Correlações de Pearson

Calculam-se as correlações de Pearson entre as variáveis socioeconômicas e as novas variáveis de voto ('left_vote' e 'right_vote'). Isso permite identificar o grau e a direção da relação linear entre esses fatores. Os resultados são organizados em um DataFrame para facilitar a análise.

variavel	corr_direita	corr_esquerda
Analfabetismo_perc	-0.757748	0.779423
PIB_per_capita	0.541687	-0.563596
IDHM_2021	0.425039	-0.441600
Gini	-0.416709	0.406385
Esperanca_vida_anos	0.380523	-0.401964
Mort_infantil_pmil	-0.165495	0.171547
PIB_milhoes	0.120746	-0.150453

Visualização das Correlações com Mapa de Calor



Para uma compreensão visual das correlações, segue-se o mapa de calor gerado.

Este gráfico colorido permite identificar rapidamente as relações mais fortes e mais fracas entre as variáveis socioeconômicas e as tendências de voto.

A escala de cores indica a intensidade e a direção da correlação.

O mapa de calor exposto ilustra as correlações de Pearson entre as variáveis socioeconômicas e as porcentagens de voto para esquerda e direita.

Observa-se que o Analfabetismo_perc tem uma forte correlação negativa com o voto de direita e positiva com o voto de esquerda, enquanto o PIB_per_capita mostra o oposto.



Análise Exploratória de Dados: Indicadores Socioeconômicos e Tendência de Votos

Resultados e Conclusões

Este notebook demonstrou um fluxo de trabalho robusto para análise de correlação, desde o carregamento seguro de dados até a visualização. A detecção dinâmica de colunas e a conversão de tipos são cruciais para evitar erros e garantir a precisão da análise.

Principais Insights

Identificam-se correlações significativas entre fatores socioeconômicos e tendências de voto:

- **Analfabetismo:** forte relação entre analfabetismo e voto de esquerda.
- **PIB per capita e IDHM 2021:** relação moderada com ideologia da direita

Predominância Ideológica

Esquerda e direita dominam o cenário político brasileiro, com o centro e suas variações desempenhando papéis menores na média nacional.

Diversidade Regional

A distribuição de votos varia significativamente entre as UFs, refletindo a complexidade e a diversidade do eleitorado brasileiro.

Aplicações

Os métodos apresentados são aplicáveis a diversas análises de dados, garantindo robustez e confiabilidade em projetos futuros.

Esta análise fornece uma base sólida para entender a dinâmica política do Brasil em 2022 e serve como ponto de partida para investigações mais profundadas.

Organização e entrega técnica

Conteúdo do Repositório

Este repositório consolida indicadores socioeconômicos de cada Unidade da Federação (UF) em 2022 e cruza esses dados com o resultado de votos nominais válidos do 1º turno das eleições de 2022, agrupando os partidos em blocos ideológicos.

Link do Github: https://github.com/frpbotero/Eleicoes_2022

Uma análise aprofundada das relações entre fatores sociais e econômicos e o comportamento eleitoral no Brasil.

Dados Socioeconômicos

O arquivo `indicadores_uf_2022.csv` contém dados de PIB, PIB per capita, IDHM, Gini, expectativa de vida, mortalidade infantil, analfabetismo e IPCA por UF em 2022.

Dados de Votação

O arquivo `votacao_partido_munzona_2022_BR.csv` detalha os votos nominais válidos por partido, zona e município, conforme dados do TSE.

Dataset Integrado

O `uf_votos_ideologia_socieco_2022.csv` é o dataset final, pronto para análise, gerado a partir da integração dos dados socioeconômicos e de votação.



Organização e entrega técnica

Estrutura de Blocos Ideológicos

Para a análise, os partidos políticos foram agrupados em cinco blocos ideológicos principais, facilitando a correlação com os indicadores socioeconômicos.



Esquerda	PT, PCB, UP, PSTU
Centro-Esquerda	PDT, PSB, REDE, PV
Centro	MDB, PSD, PODE, CIDADANIA, AVANTE
Centro-Direita	UNIÃO, PSDB, PP, NOVO, SOLIDARIEDADE
Direita	PL, PTB, DC, PATRIOTA, PRTB

Pré-requisitos Técnicos

Ambiente Python

É necessário ter Python versão 3.9 ou superior instalado, juntamente com os pacotes **pandas**, **matplotlib** e **jupyter** (ou VS Code) para a execução dos notebooks.

Ollama Server

O **Ollama server** deve estar instalado e o modelo local **deepseek-r1:1.5b** precisa ser baixado para a geração de metadados.

Instalação de Dependências

Todos os pacotes necessários podem ser instalados via **pip install -r requirements.txt**, garantindo as versões corretas das bibliotecas.

Organização e entrega técnica

Executando o Projeto: Passo a Passo

Geração do Dataset Integrado

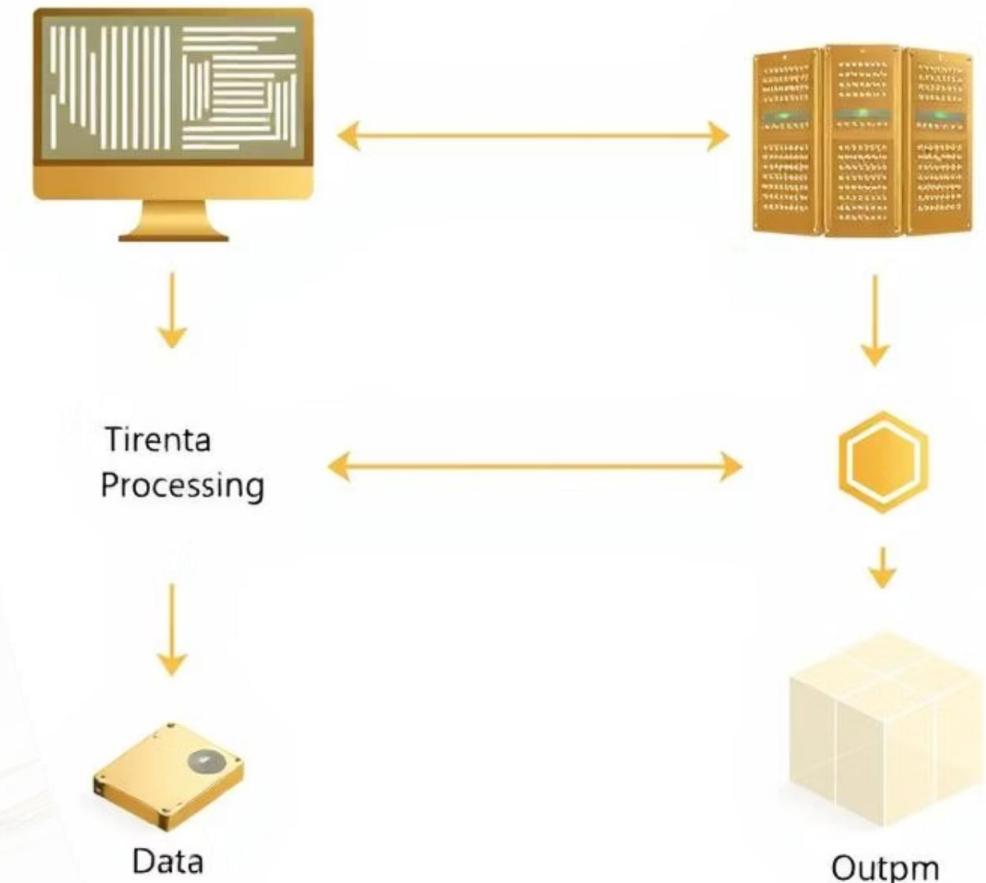
Execute o comando `jupyter nbconvert --to notebook --execute notebooks/process_uf_socio_votos.ipynb` para criar o dataset unificado.

Exploração de Correlações

Abra e execute o notebook `notebooks/correlacao_pearson_dynamic_cols.ipynb` para calcular as correlações de Pearson.

Visualização de Médias Ideológicas

Abra e execute `notebooks/ideologia_media_votos_notebook.ipynb` para explorar as médias de votos por bloco ideológico e gerar gráficos.



Organização e entrega técnica

Geração Opcional de Metadados

Para uma compreensão mais aprofundada dos dados brutos, é possível gerar metadados detalhados para cada arquivo CSV e um sumário geral.

Primeiro, inicie o Ollama server com `ollama serve &` (caso não esteja ativo). Em seguida, execute o script Python `scripts/extract_metadata.py`.

A saída serão arquivos JSON localizados em `data_lake/metadata/`, incluindo um arquivo por CSV e o sumário `all_files_summary_metadata.json`.

