Análise exploratória inicial

---

1) Estrutura dos dados

Atributo	Valor
Linhas (observações)	**5**
Colunas	**11**
Coluna identificada como granularidade/região	**Região**
Colunas numéricas detectadas (lista)	`População_2024`, `Percentual_Pop_Nacional`, `Área_km2`, `Densidade_Demografica`, `Taxa_Crescimento_2010_2022`, `Estados`, `Municípios`, `PIB_Per_Capita_2022`, `Idade_Mediana_Estimada`, `Taxa_Urbanização`

Resumo qualitativo:

- Não foram detectados valores ausentes no conjunto (0 missing).
- Cada linha representa uma das 5 regiões oficiais do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste, Sul).
- Com apenas 5 linhas, as medidas de dispersão entre regiões são sensíveis a valores extremos (outliers) — trate-as como descritivas e exploratórias.

(Tabelas detalhadas — cabeçalho, tipos das colunas, estatísticas completas e matriz de correlação — foram geradas e enviadas como tabelas interativas no ambiente; pode consultá-las diretamente.)

---

2) As 3 principais características que distinguem as regiões

Para identificar as variáveis que mais distinguem as regiões calculei a variabilidade entre as médias regionais (variância das médias por região). As **top 3** variáveis são:

1. **População_2024** — maior diferença absoluta entre as regiões no número de habitantes (ex.: Sudeste com média ~88.6M vs Sul ~31.1M etc.).
2. **Área_km2** — diferenças substanciais de extensão territorial (Norte apresenta área muito superior às demais).
3. **PIB_Per_Capita_2022** — variação marcante entre regiões no PIB per capita.

Para clareza, segue a tabela com as médias por região para essas 3 variáveis (médias calculadas a partir do dataset):

Região	População_2024	Área_km2	PIB_Per_Capita_2022
Centro-Oeste	17,071,595	1,606,371.5	42,680
Nordeste	57,112,096	1,554,256.8	18,950
Norte	18,669,345	**3,853,676.9**	21,870

Região	População_2024	Área_km2	PIB_Per_Capita_2022
Sudeste	**88,617,693**	924,511.3	**45,280**
Sul	31,113,021	576,743.5	39,340

Interpretação sucinta:

- **População_2024**: Sudeste concentra a maior população; Nordeste também é numericamente grande; diferenças são expressivas.
- **Área_km2**: Norte é muito maior territorialmente (região Amazônica), explicando alta variância.
- **PIB per capita**: Sudeste e Sul lideram, Nordeste tem menor PIB per capita médio — isso destaca desigualdades econômicas regionais.

---

3) Estatísticas descritivas básicas (média, mediana, desvio padrão)

Apresentei estatísticas para as variáveis numéricas principais. Abaixo os valores arredondados:

Variável	Média	Mediana	Desvio padrão
População_2024	42,516,750	31,113,021	30,350,126
Área_km2	1,703,112	1,554,256.8	1,277,827
Densidade_Demografica	40.38	36.70	36.85
PIB_Per_Capita_2022	33,624	39,340	12,288.5
Idade_Mediana_Estimada	32.50	32.10	3.65
Taxa_Urbanização	82.62 (%)	84.90 (%)	8.956
Taxa_Crescimento_2010_2022	8.622 (%)	9.31 (%)	4.87

Comentários:

- A **média da população** é puxada para cima pelo Sudeste; a mediana (31.1M) mostra que metade das regiões tem população abaixo desse valor.
- **Área_km2** tem grande desvio padrão, refletindo a enorme área do Norte.
- **PIB per capita** mostra que, embora a média seja ~33.6k, a mediana está acima da média, indicando assimetria nos valores regionais (Sudeste e Sul elevam a mediana).

---

4) Outliers e valores interessantes detectados

Método: regra IQR (valores fora de `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]`).

Principais achados:

- **Área_km2**: 1 outlier — **Região: Norte**, valor = **3,853,676.90 km²**. Justificativa plausível: a região Norte inclui vastas áreas amazônicas; esse valor é consistente com a realidade geográfica.

- **Taxa_Crescimento_2010_2022**: 1 outlier — **Região: Centro-Oeste**, valor = **15.87 (%)**. Isso é notável e merece verificação (pode refletir forte crescimento populacional ou mudança metodológica/estimativa).
- Todas as demais variáveis numéricas não apresentaram outliers pelo critério IQR no dataset atual.

Outros pontos de interesse:

- **Nenhum valor ausente** encontrado — facilita análises posteriores.
- O fato de existirem somente 5 observações (uma por região) torna cada outlier especialmente influente nas estatísticas agregadas — recomendo validar as entradas atípicas contra fontes oficiais (IBGE, dados estaduais) antes de tirar conclusões.

---

5) Cinco perguntas analíticas recomendadas (próximas análises)

Com base nesse dataset e nos achados, sugiro as seguintes perguntas que podemos responder / explorar:

1. **Como a urbanização (Taxa_Urbanização) se relaciona com PIB per capita por região?** — investigar correlação e direções de causa sugeridas.
2. **Quais regiões apresentaram maior crescimento relativo entre 2010 e 2022 e quais fatores (densidade, urbanização) podem explicar esse crescimento?** — explorar regressões simples e comparação temporal (se tivermos séries).
3. **Existe uma relação entre densidade demográfica e idade mediana?** — testar hipótese se regiões mais densas têm idade mediana diferente.
4. **Como a participação percentual da população nacional (`Percentual_Pop_Nacional`) evolui se compararmos com o PIB per capita — há regiões sub-representadas economicamente?** — análise de "participação populacional vs. participação econômica".
5. **Se agrupássemos municípios/estados por faixa de PIB per capita e densidade, como se distribuiriam os indicadores sociais (idade mediana, taxa de urbanização)?** — clustering / segmentação para políticas públicas.

Observações finais e próximos passos sugeridos

Mestre, recomendo as seguintes ações operacionais imediatas (posso executar ao seu comando):

- **Validar as observações atípicas** (Norte — área; Centro-Oeste — taxa de crescimento) com fontes oficiais (IBGE ou bases estaduais).
- **Visualizações**: mapas coropléticos (por região), boxplots por variável e gráficos de barras das médias regionais (ajudam muito na comunicação).

- **Análises de correlação/associação**: matriz de correlação entre variáveis numéricas e pequenas regressões exploratórias.
- **Se houver dados temporais (ex.: séries históricas por região)**, rodar análise de tendência para crescimento populacional e PIB.

Posso proceder imediatamente com qualquer um dos itens acima — por exemplo, gerar gráficos (mapa ou barras), validar valores atípicos ou construir um pequeno relatório em PDF/PowerPoint. Indique qual desses caminhos prefere que eu execute em seguida.

Coloquei tabelas interativas (cabeçalho, dtypes, estatísticas completas, médias por região, matriz de correlação, resumo de outliers e amostra das linhas com outliers) no ambiente para consulta direta. Quer que eu gere agora gráficos ilustrativos (barras/matrizes/boxplots) ou valide os outliers com fontes externas?