plusoft

Analytics e Data Science

Fevereiro, 2023

01

Objetivo e Premissas
[5 min]

02

Análise Exploratória
[10 min]

03

Segmentação
[10 min]

04

Classificação
[10 min]

05

Materiais e Dúvidas
[? min]

OBJETIVO

O QUE ESPERAR DESTA APRESENTAÇÃO?

Solução do case como um teste para o processo de seleção.

Transmitir **didática** na apresentação **para não especialistas** quantitativos.

Justificar técnicas para fins de **validação de conhecimentos** nos modelos, mas não buscar o estado da arte.

PREMISSAS

1. **Validação dos Dados**: não houve validação dos dados fornecidos, apenas seu uso AS-IS.

2. **Enriquecimento**: dados do IBGE foram utilizados, cruzando com os dados de municípios fornecidos para possibilitar análises geográficas.

3. **Questionamento estratégico**: como não há uma profundidade no questionamento estratégico, a análise visa meramente uma sugestão com a visão técnica do analista.

01

Objetivo e Premissas

02

Análise Exploratória
[10 min]

03

Segmentação

04

Classificação

05

Materiais e Dúvidas

ANÁLISE EXPLORATÓRIA

BIG NUMBERS

O QUE TEMOS?

5567

Municípios no Brasil em 2010, conforme base do IBGE.

23

Variáveis com dados

0

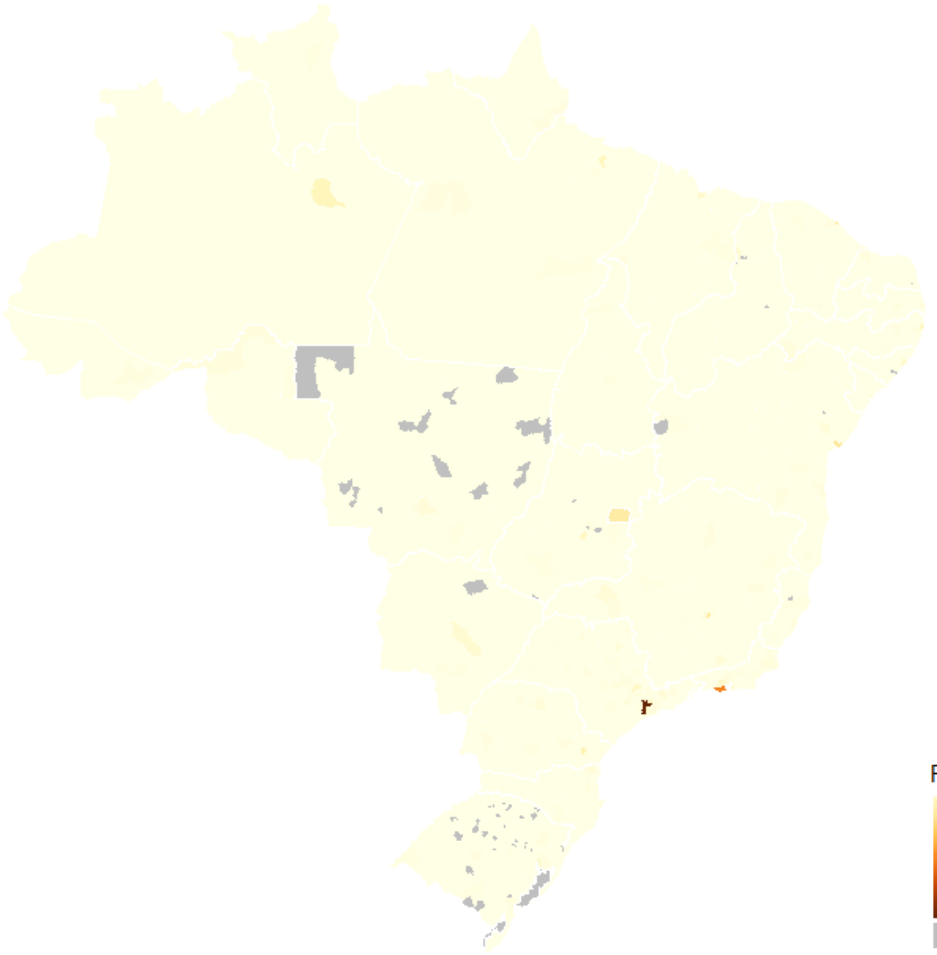
Variáveis com dados faltantes

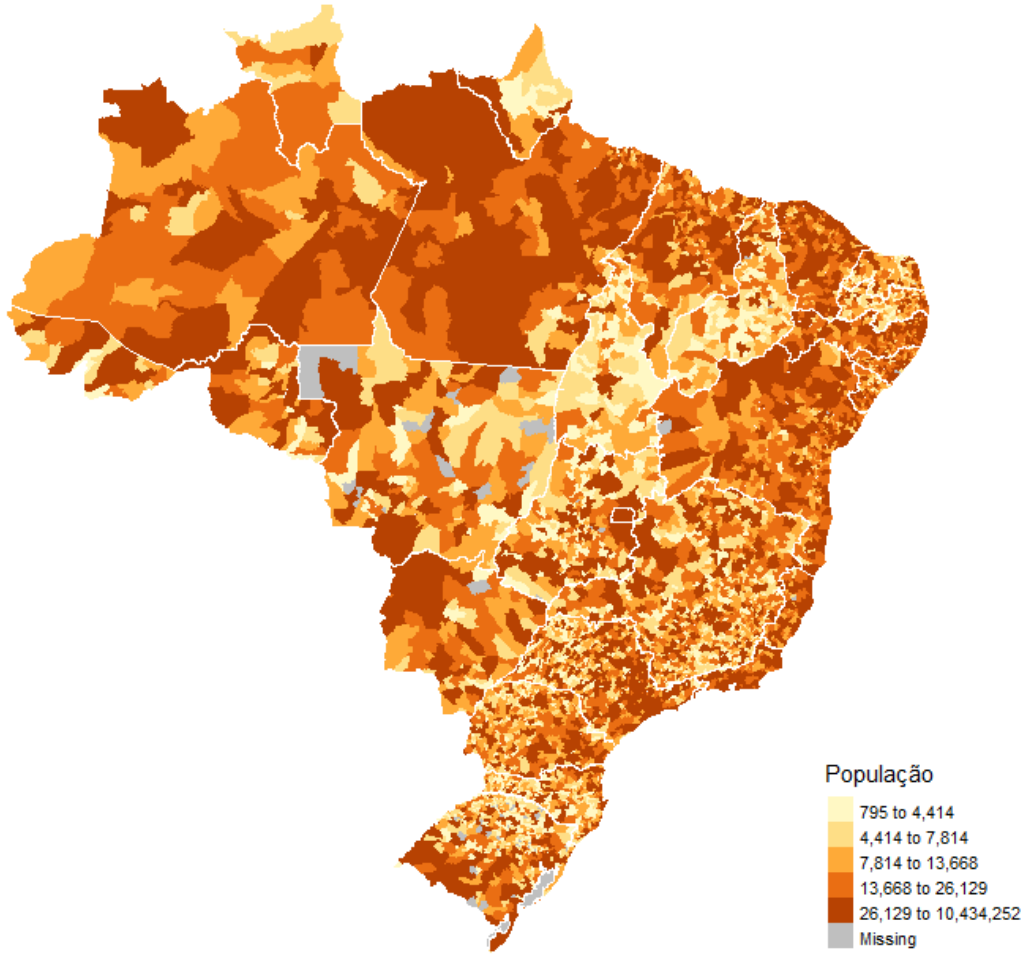
5507

Municípios com dados para análise na base da Plusoft.

ANÁLISE EXPLORATÓRIA

POPULAÇÃO

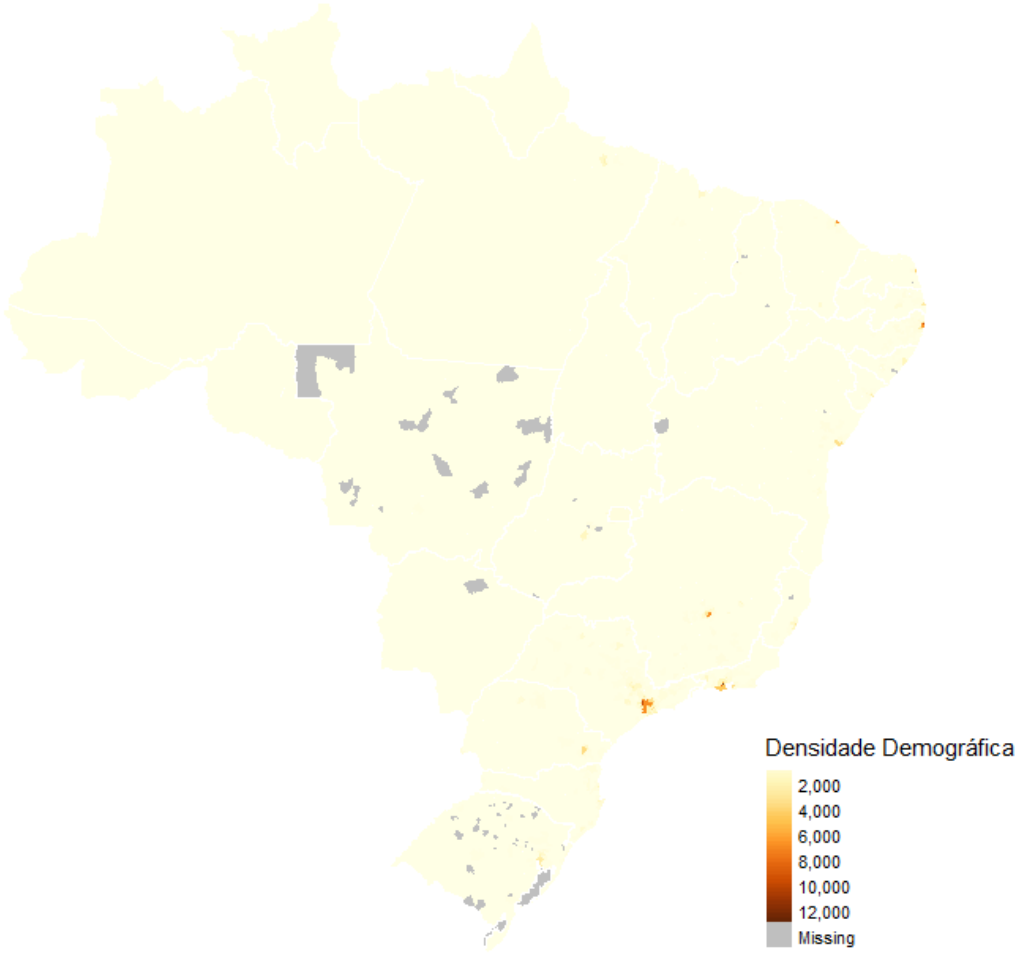


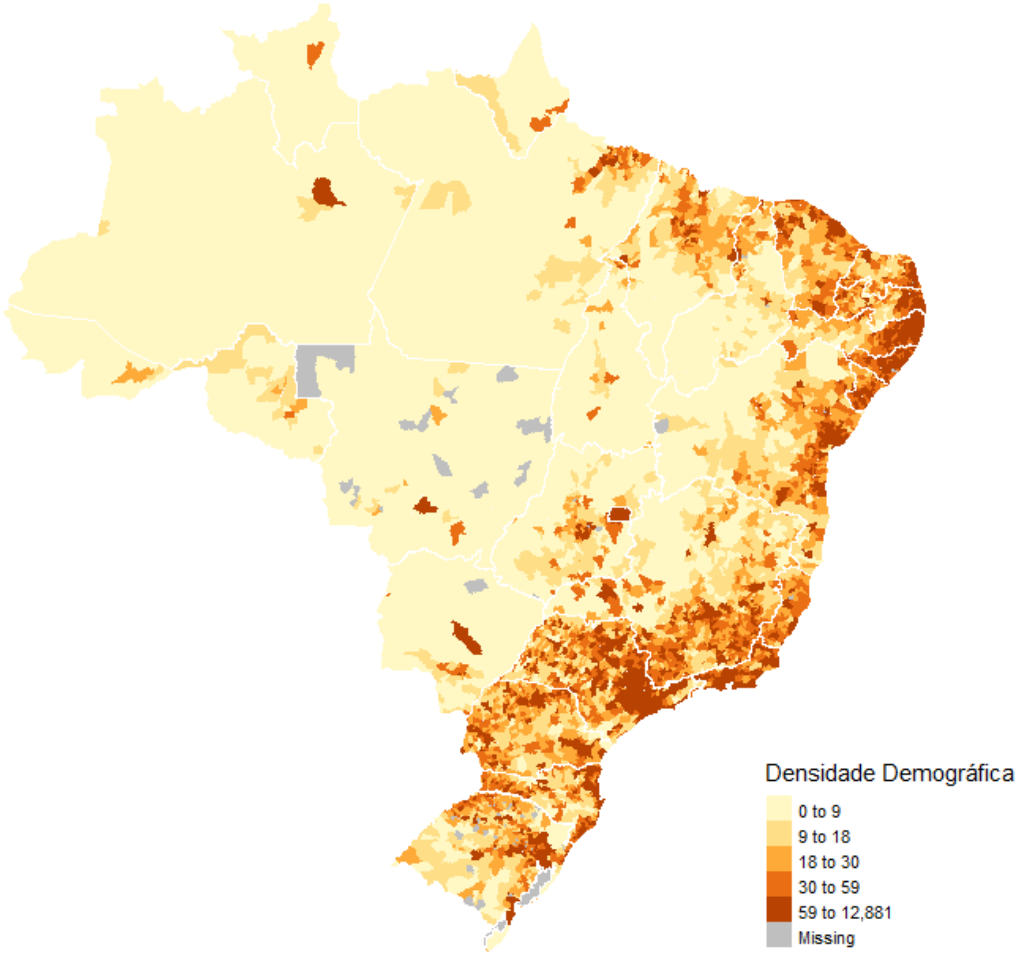


ANÁLISE EXPLORATÓRIA

DENSIDADE DEMOGRÁFICA

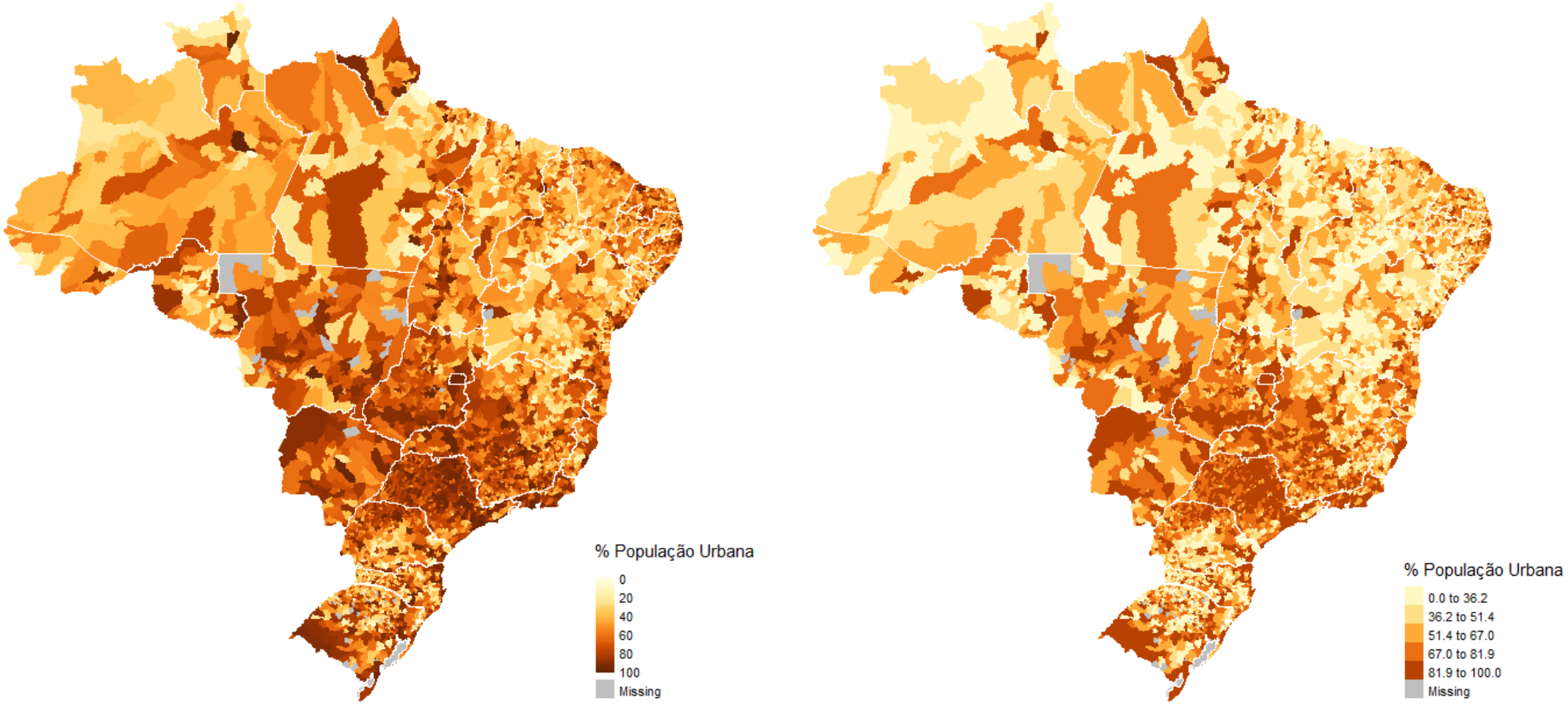
QUANTO MAIOR O VALOR, MAIOR A CONCENTRAÇÃO DE GENTE.





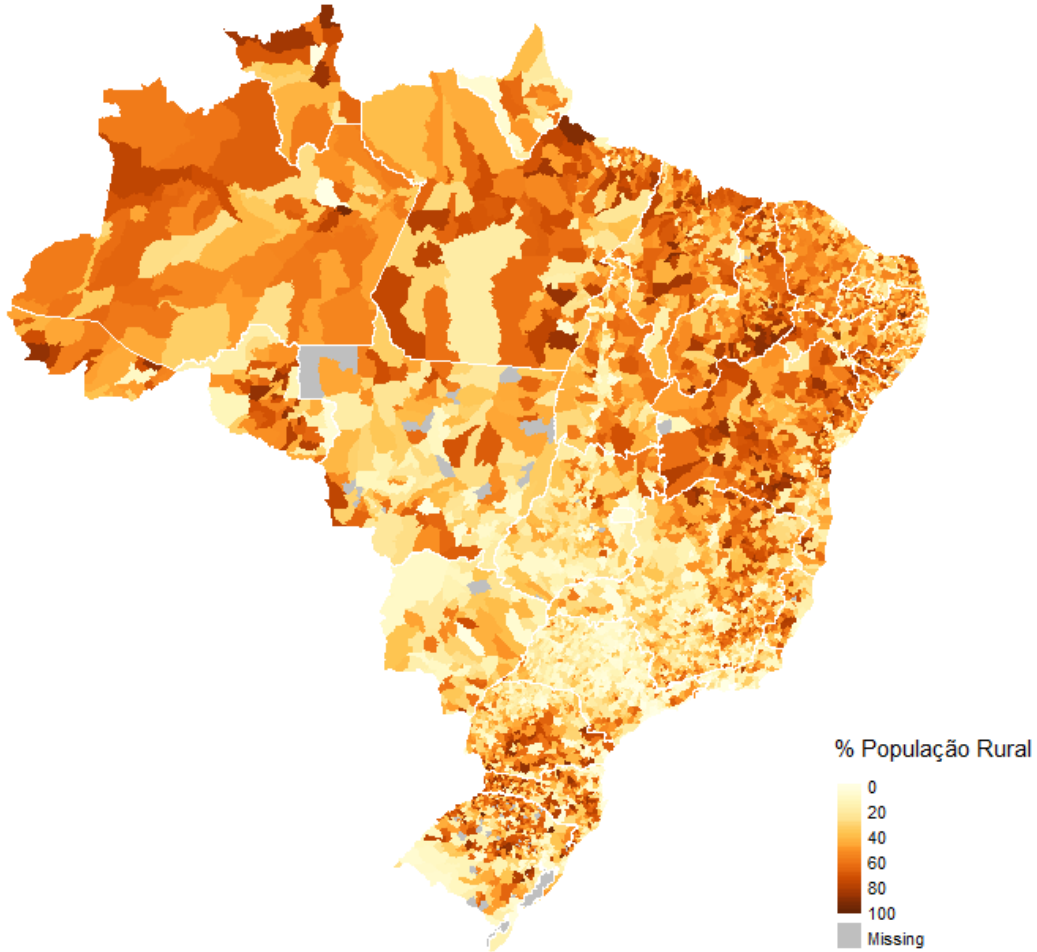
ANÁLISE EXPLORATÓRIA

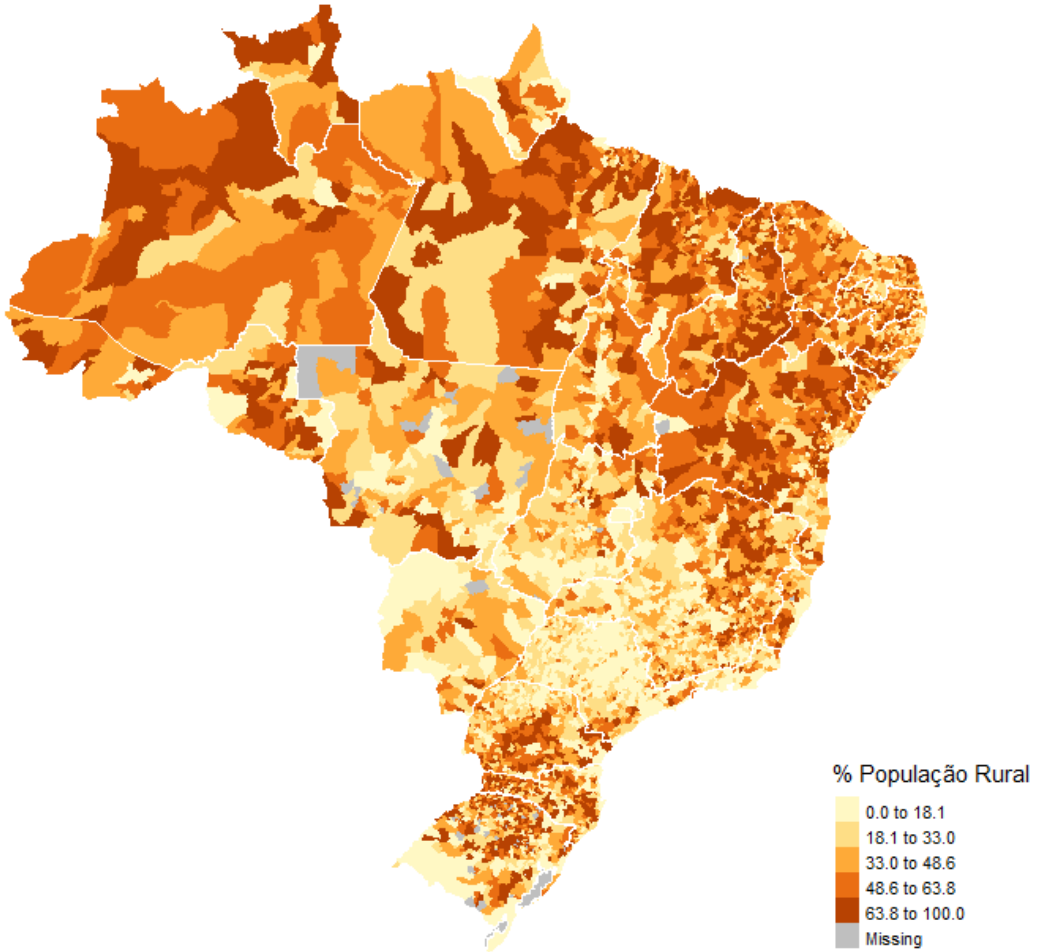
% POPULAÇÃO URBANA



ANÁLISE EXPLORATÓRIA

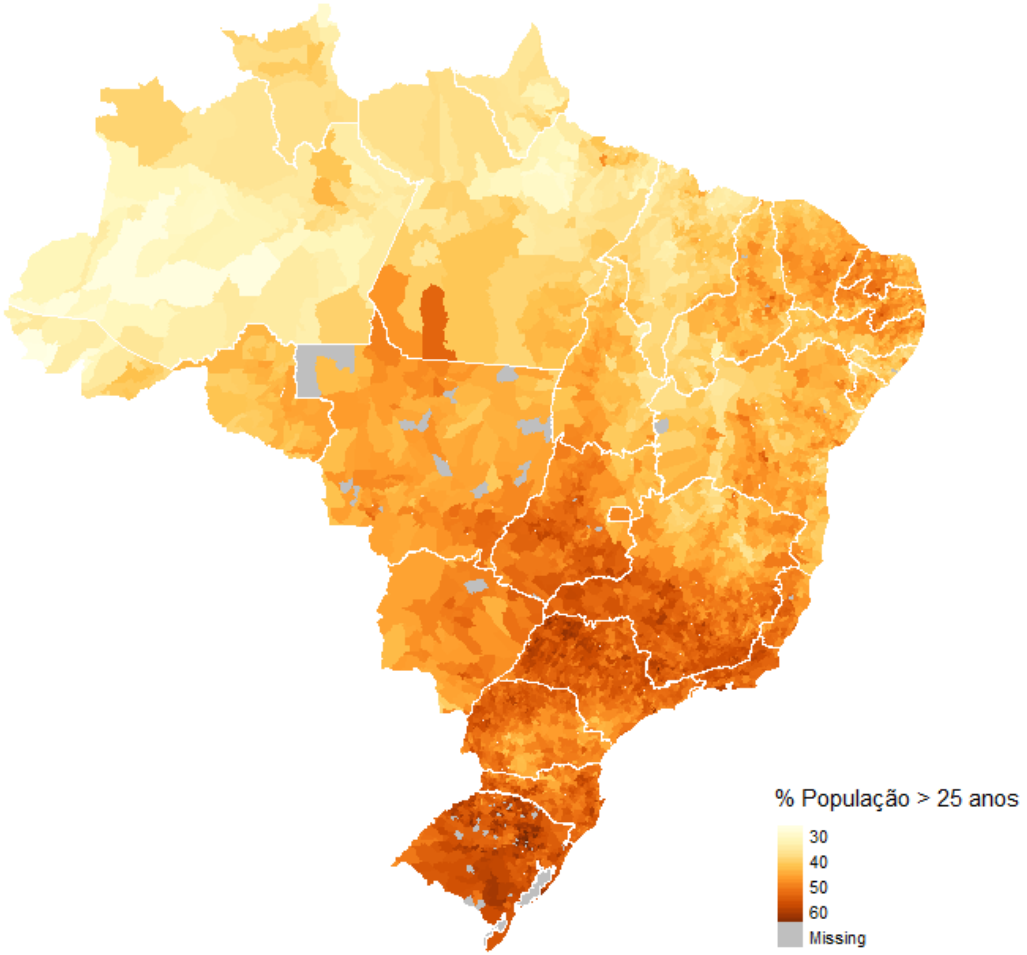
% POPULAÇÃO RURAL

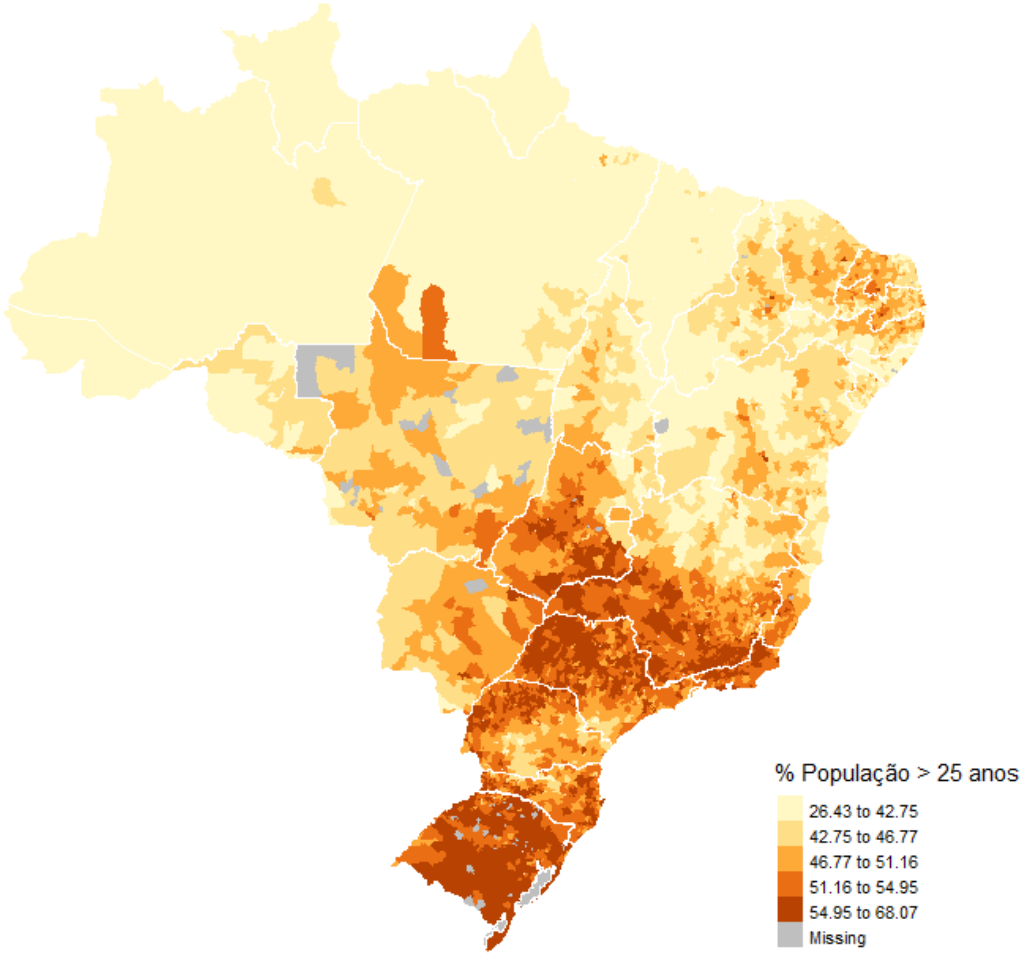




ANÁLISE EXPLORATÓRIA

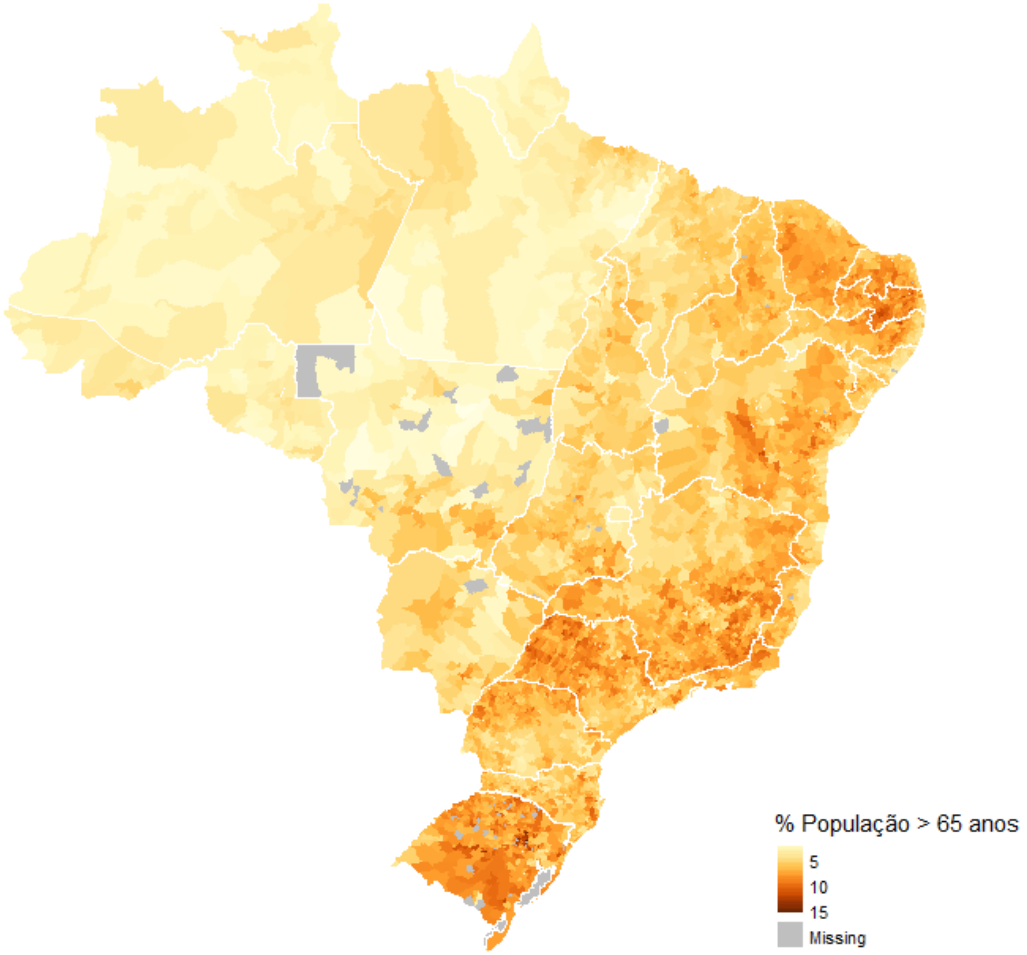
% POPULAÇÃO > 25 ANOS

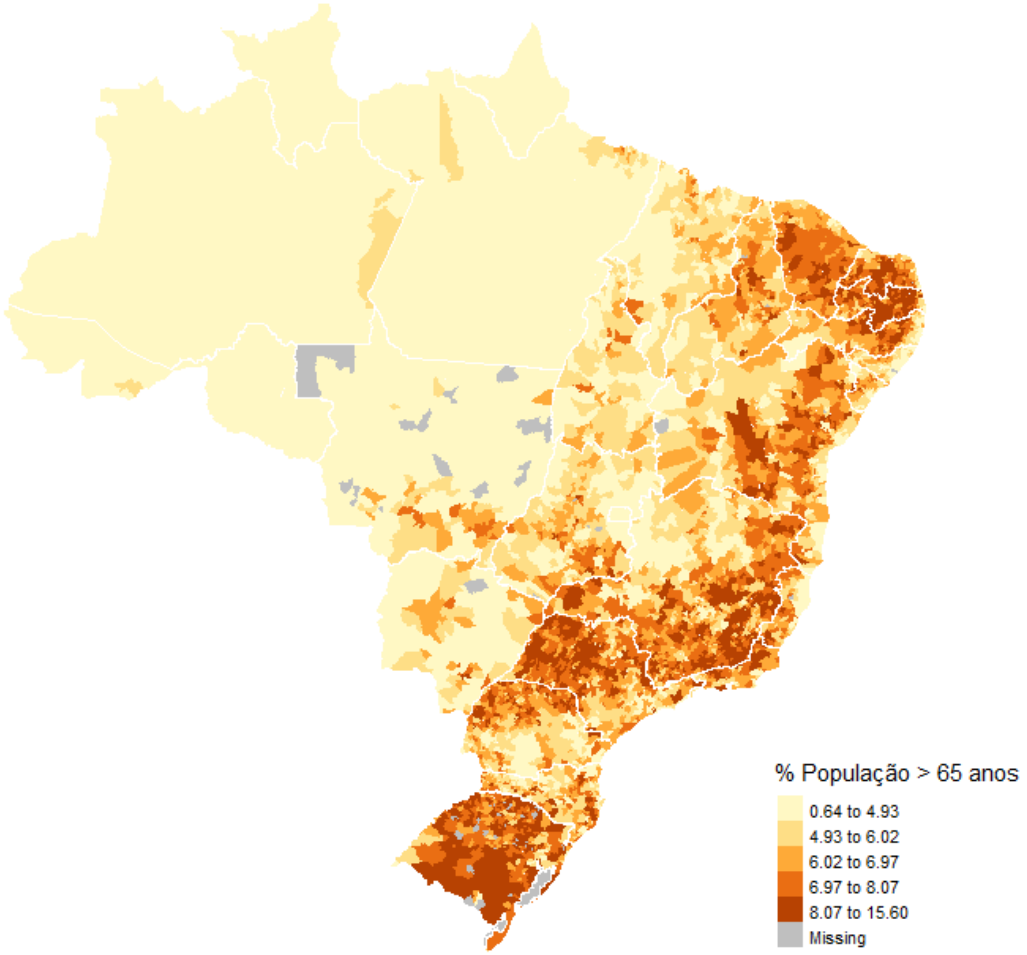




ANÁLISE EXPLORATÓRIA

% POPULAÇÃO > 65 ANOS

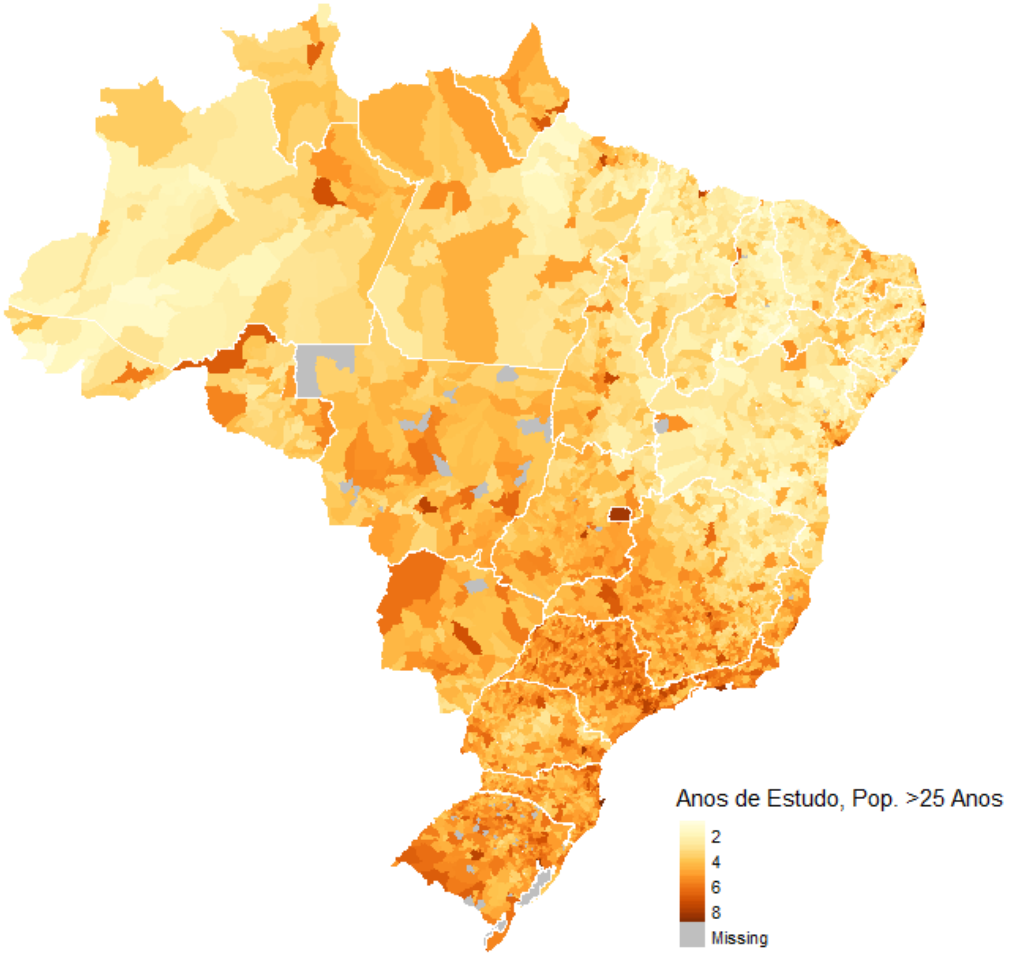


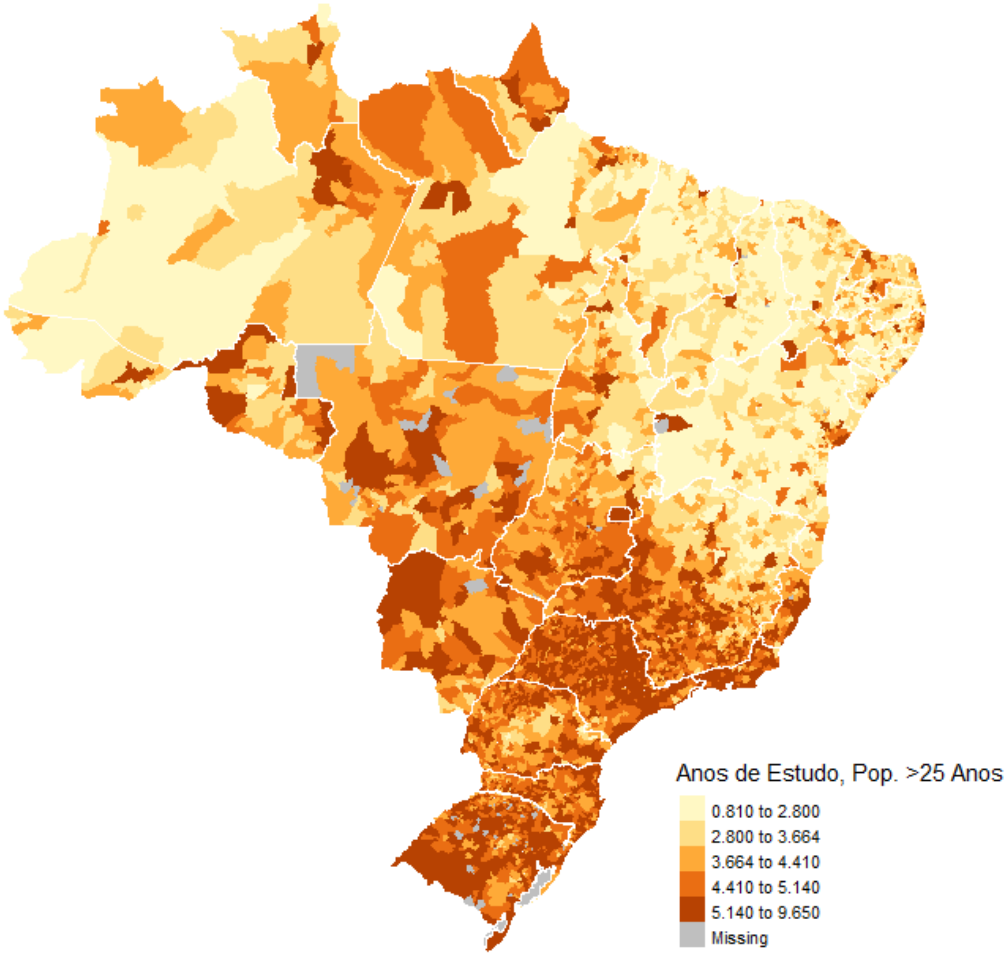


ANÁLISE EXPLORATÓRIA

MÉDIA DE ANOS DE ESTUDOS

QUANTO MAIOR O VALOR, MAIS A POPULAÇÃO > 25 ANOS ESTUDOU.

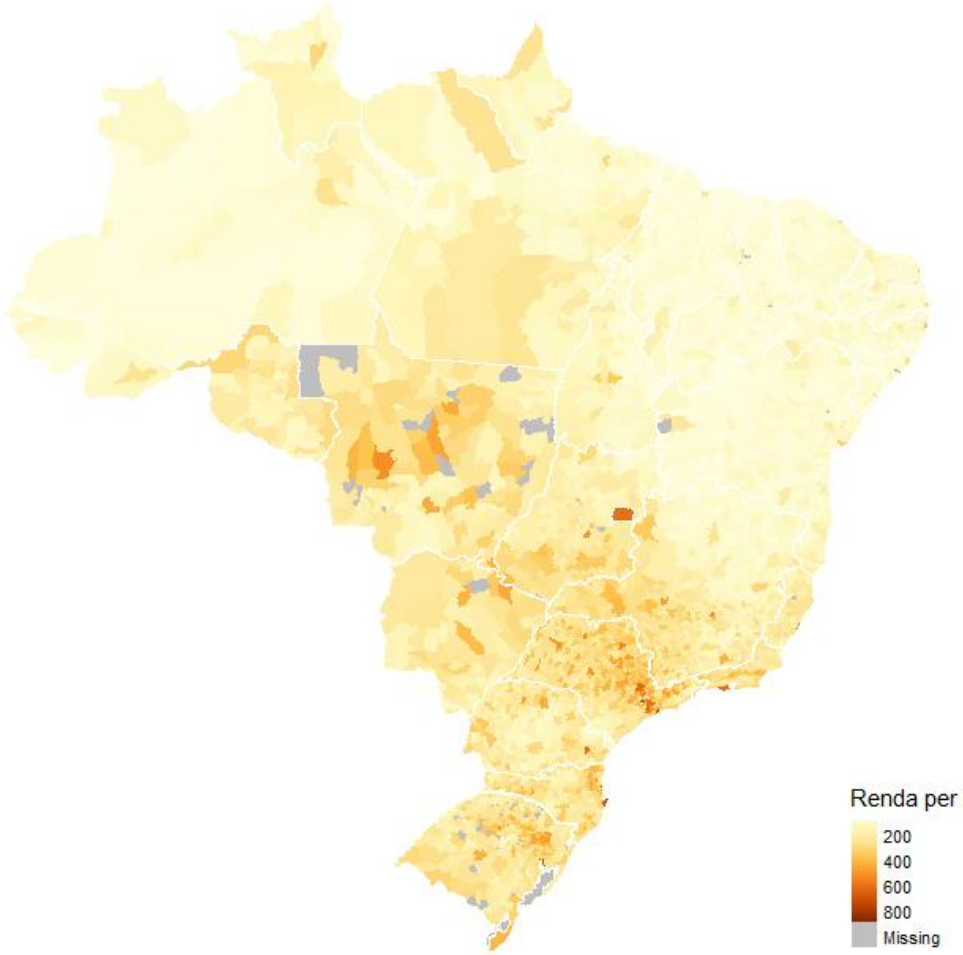


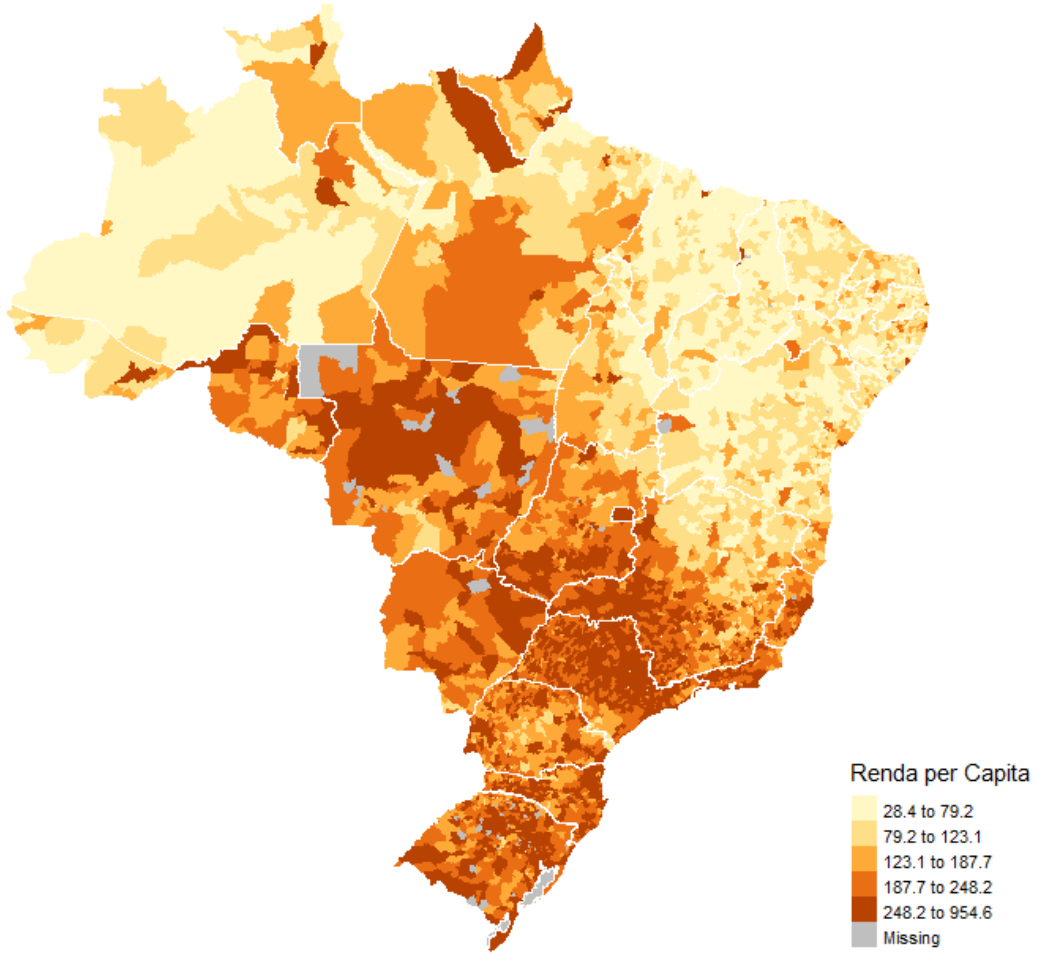


ANÁLISE EXPLORATÓRIA

RENDA PER CAPITA

QUANTO MAIOR O VALOR, MAIOR A RENDA MÉDIA.

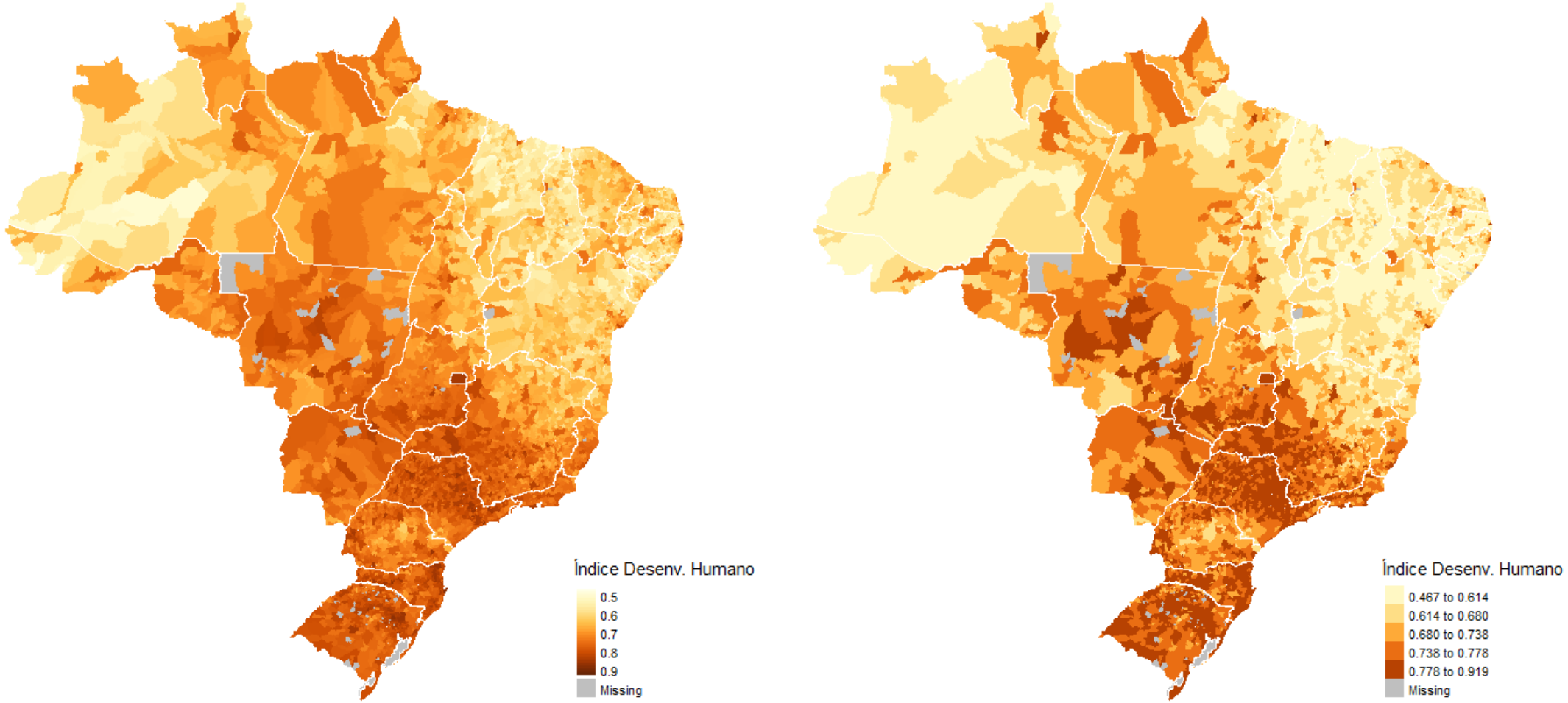




ANÁLISE EXPLORATÓRIA

ÍNDICE DE DESENV. HUMANO

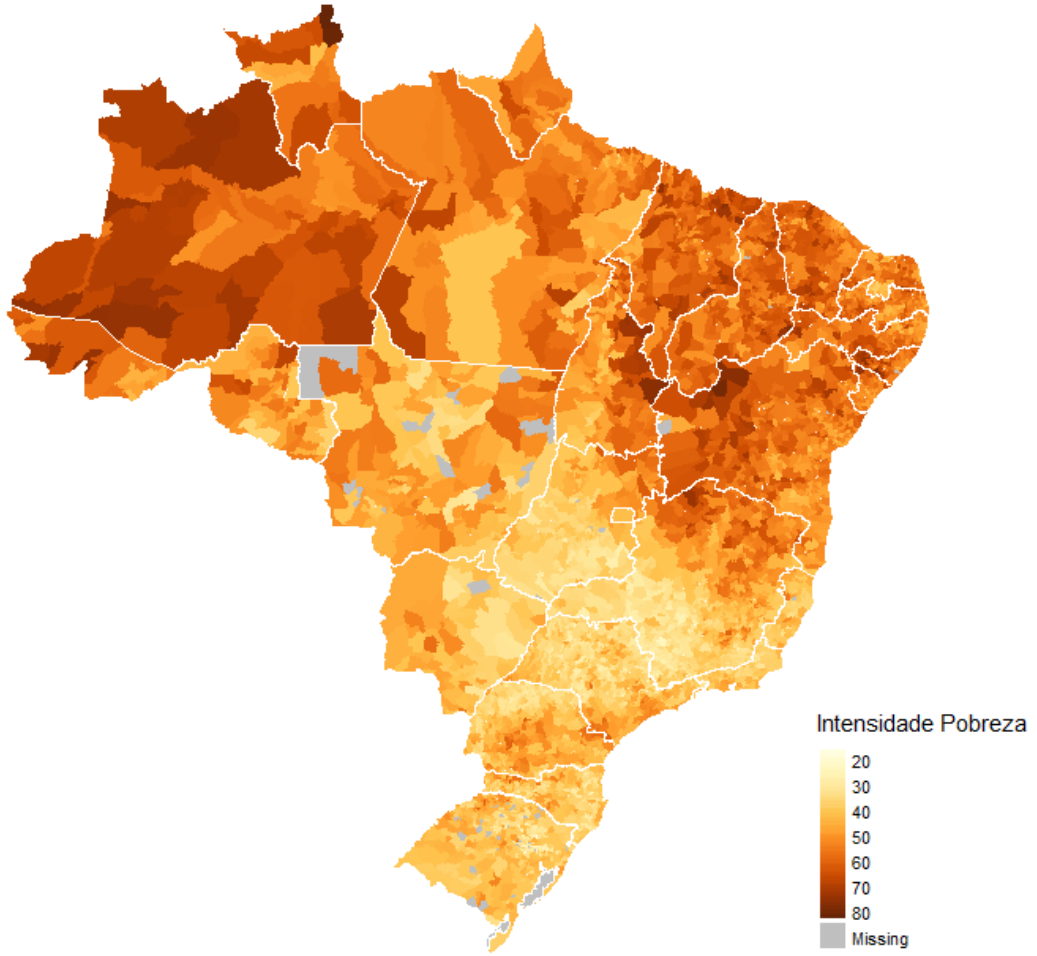
QUANTO MAIOR O ÍNDICE, MAIS DESENVOLVIDO.

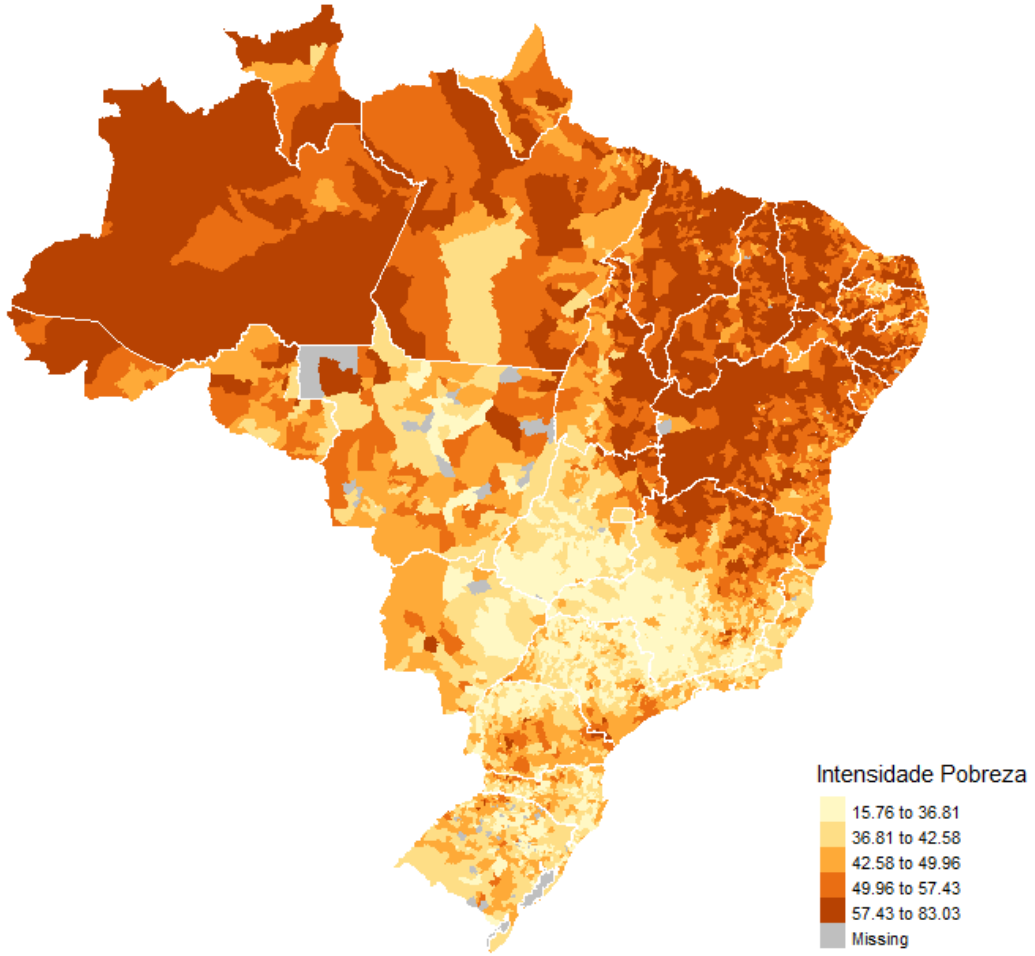


ANÁLISE EXPLORATÓRIA

INTENSIDADE DA POBREZA

QUANTO MAIOR O VALOR, MAIS INTENSA A POBREZA

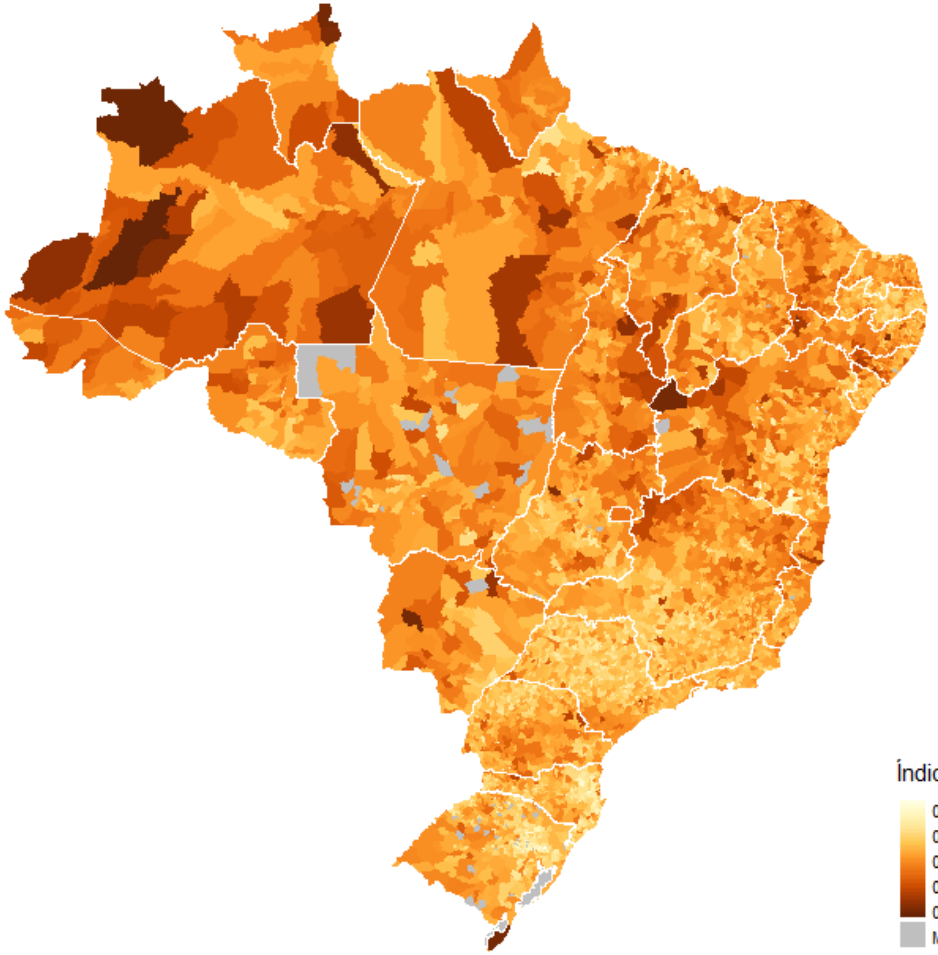


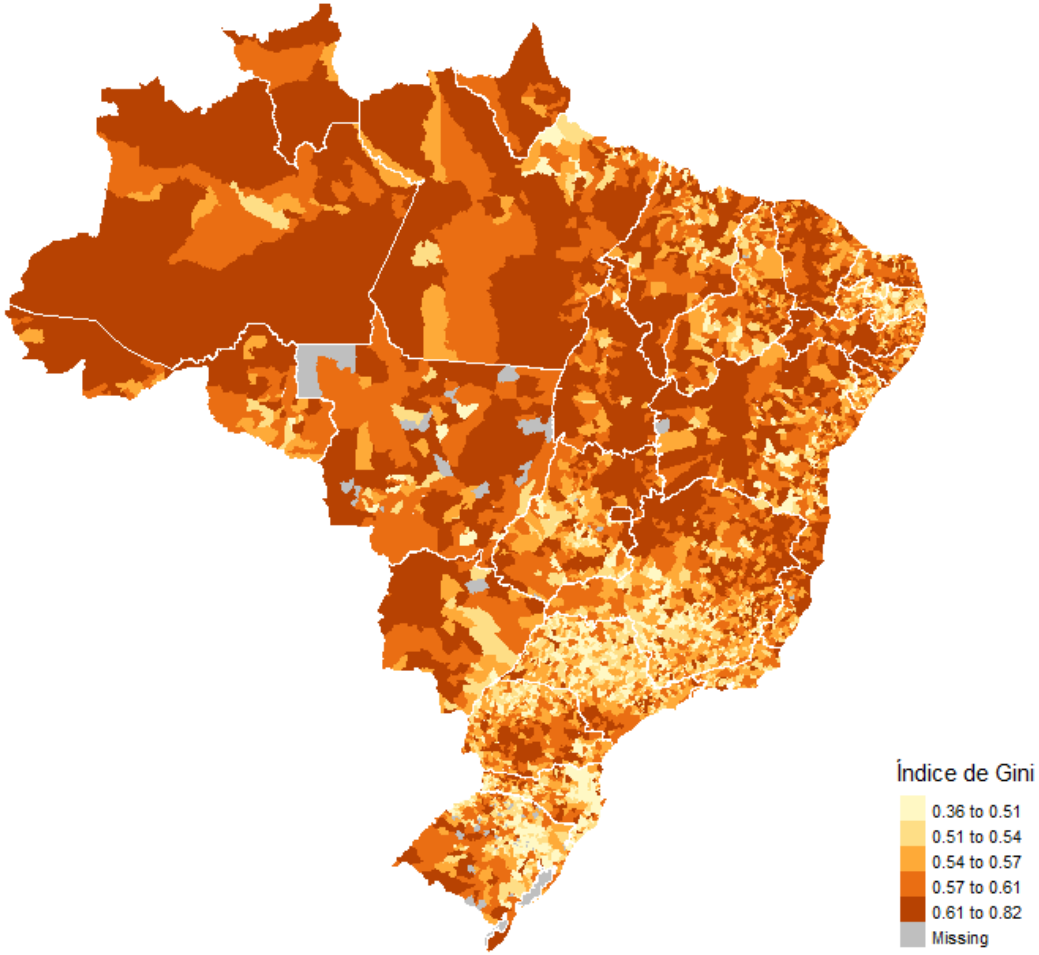


ANÁLISE EXPLORATÓRIA

ÍNDICE DE GINI

QUANTO MAIOR O ÍNDICE, MAIOR A DESIGUALDADE





01
Objetivo e Premissas

02
Análise Exploratória

03
Segmentação
[10 min]

04
Classificação

05
Materiais e Dúvidas

SEGMENTAÇÃO

CLUSTERIZAÇÃO

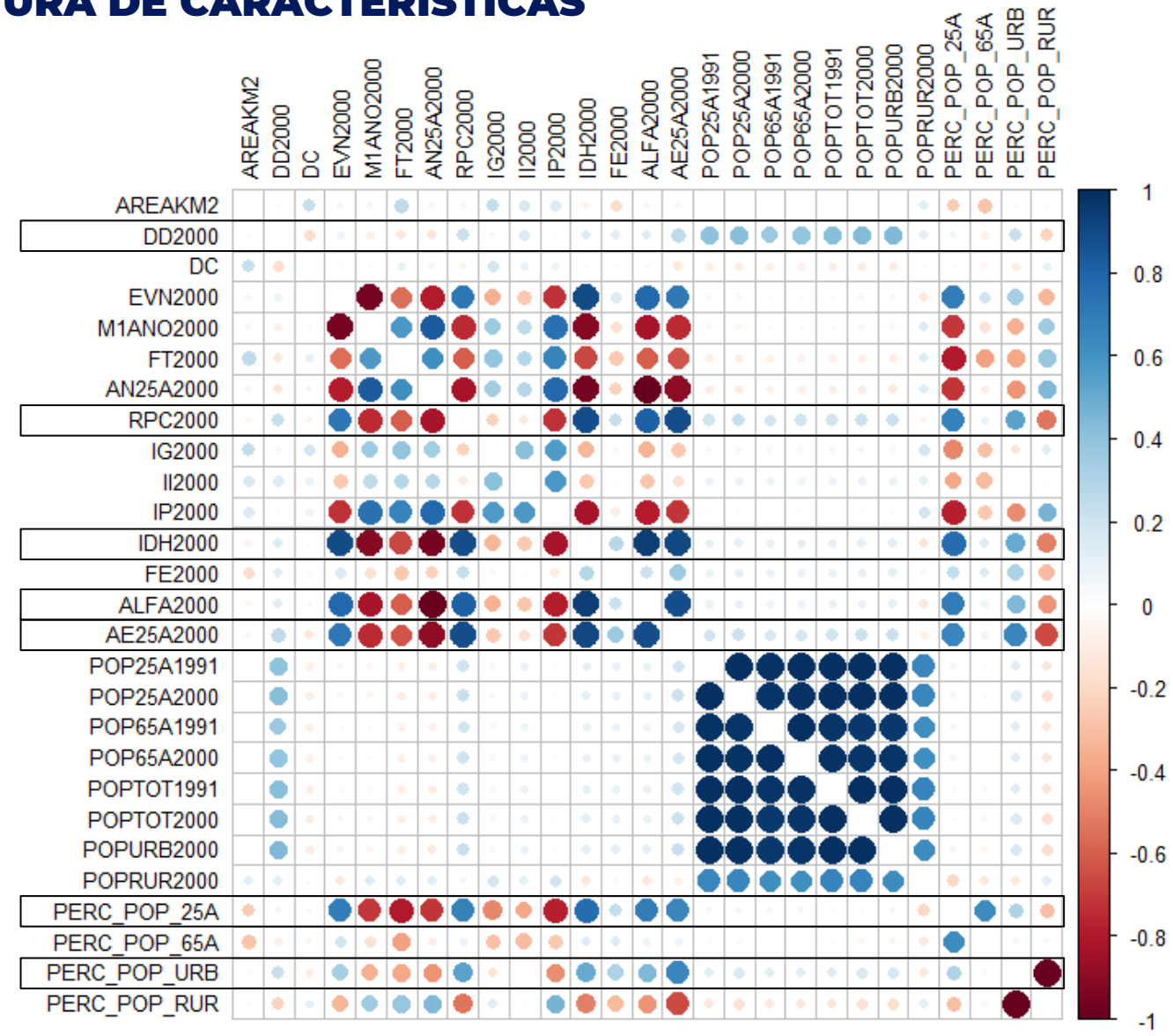
O QUE FIZEMOS?

Aplicar técnica para separar os municípios conforme os dados existentes, buscando descobrir características, diferenciações e encontrar sugestões para o negócio.

SEGMENTAÇÃO

CORRELAÇÕES

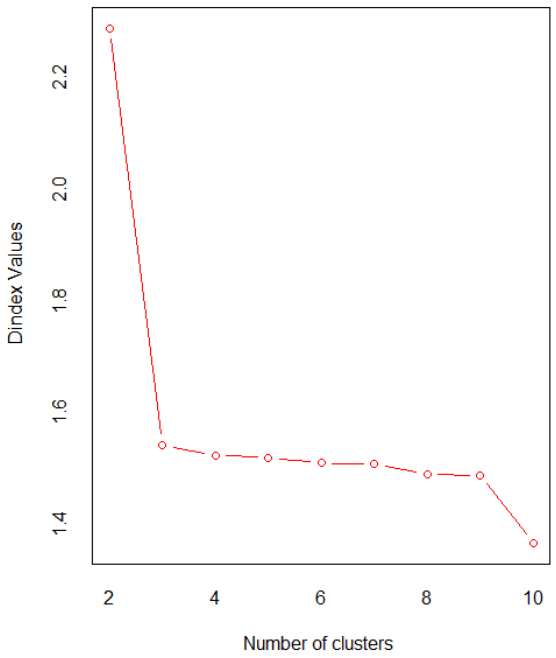
SELEÇÃO PARA CAPTURA DE CARACTERÍSTICAS

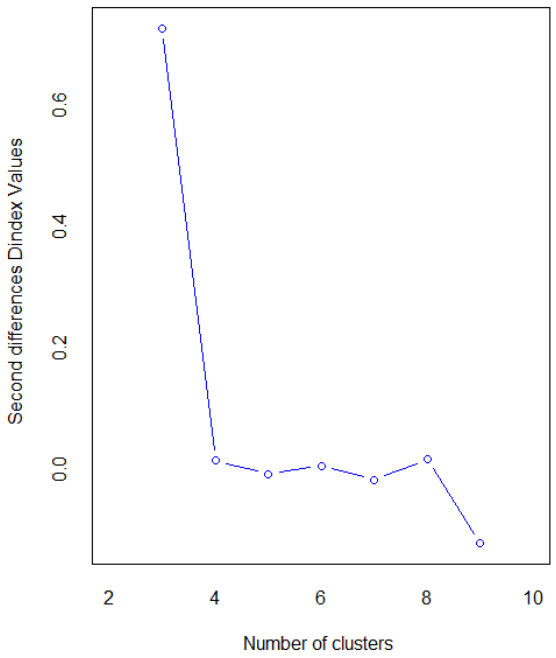


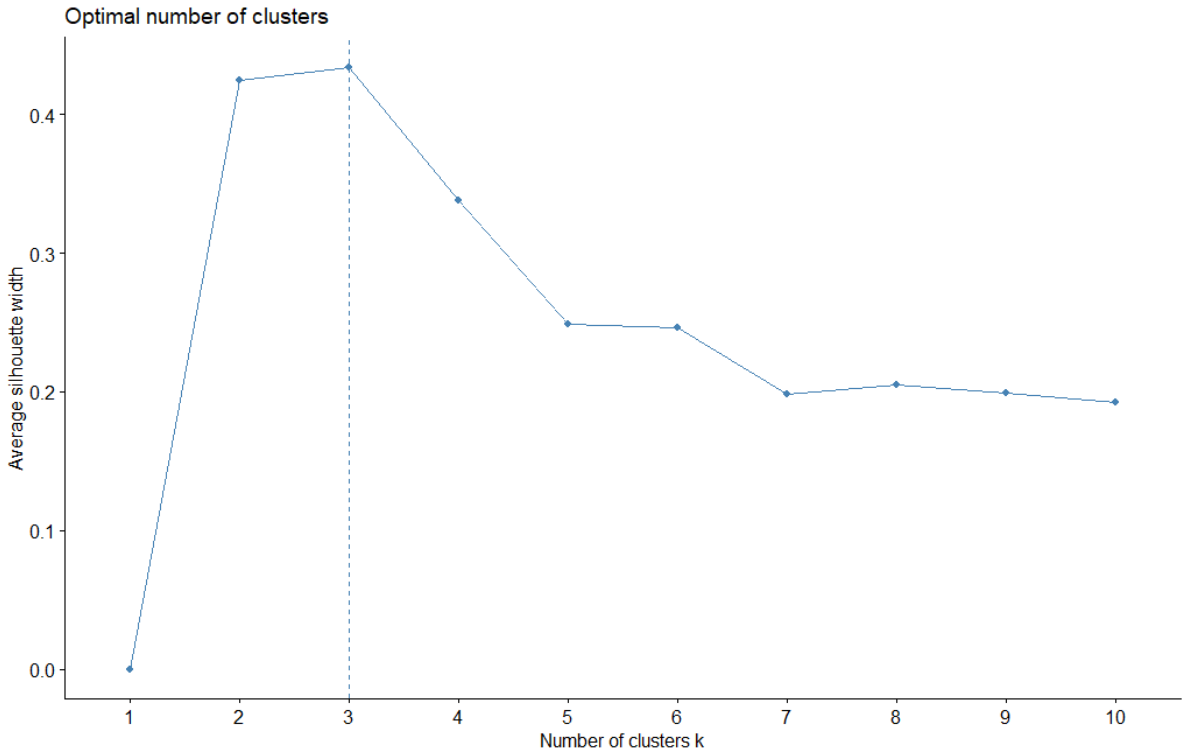
SEGMENTAÇÃO

CLUSTERIZAÇÃO

POSSIBILIDADES DE AGRUPAMENTOS



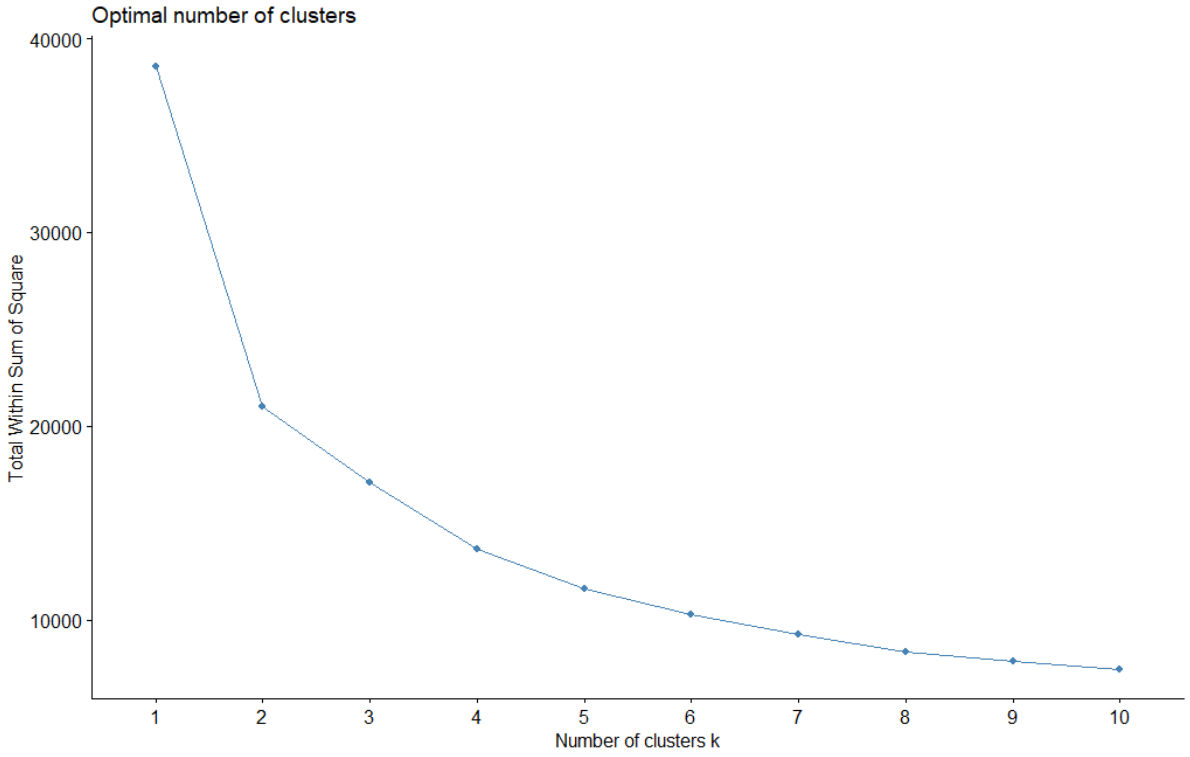


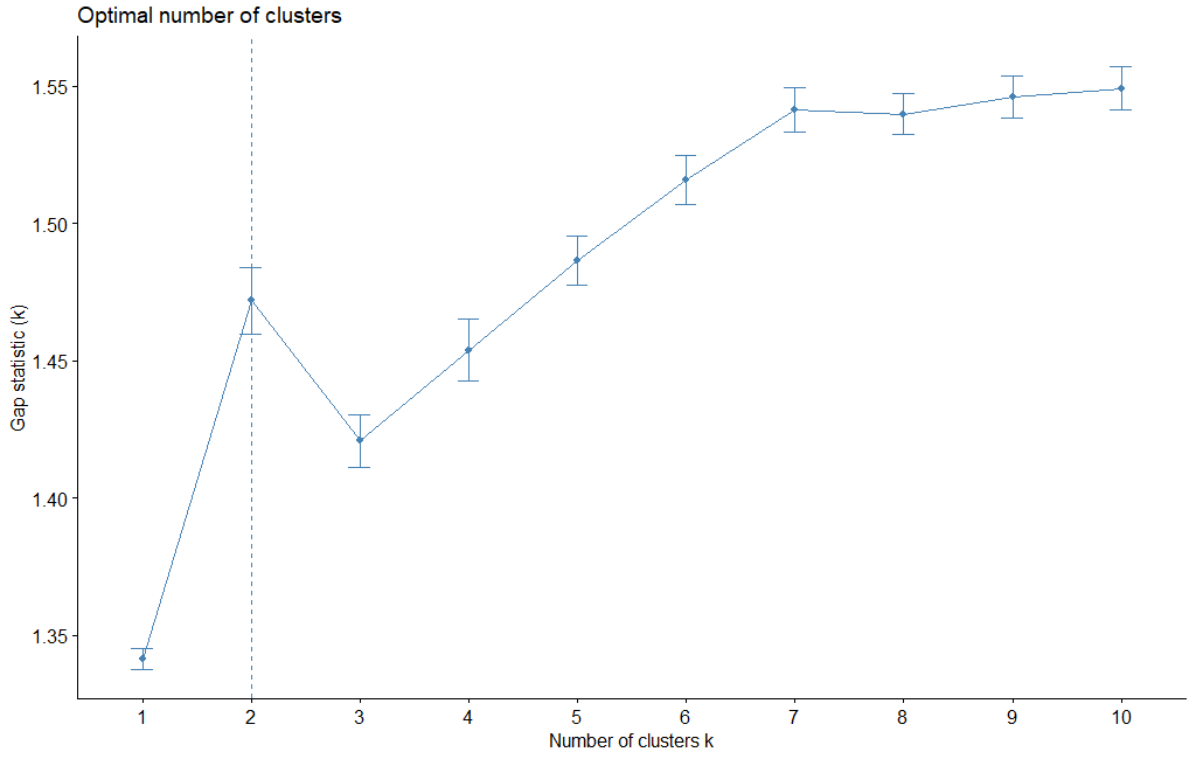


SEGMENTAÇÃO

CLUSTERIZAÇÃO

POSSIBILIDADES DE AGRUPAMENTOS

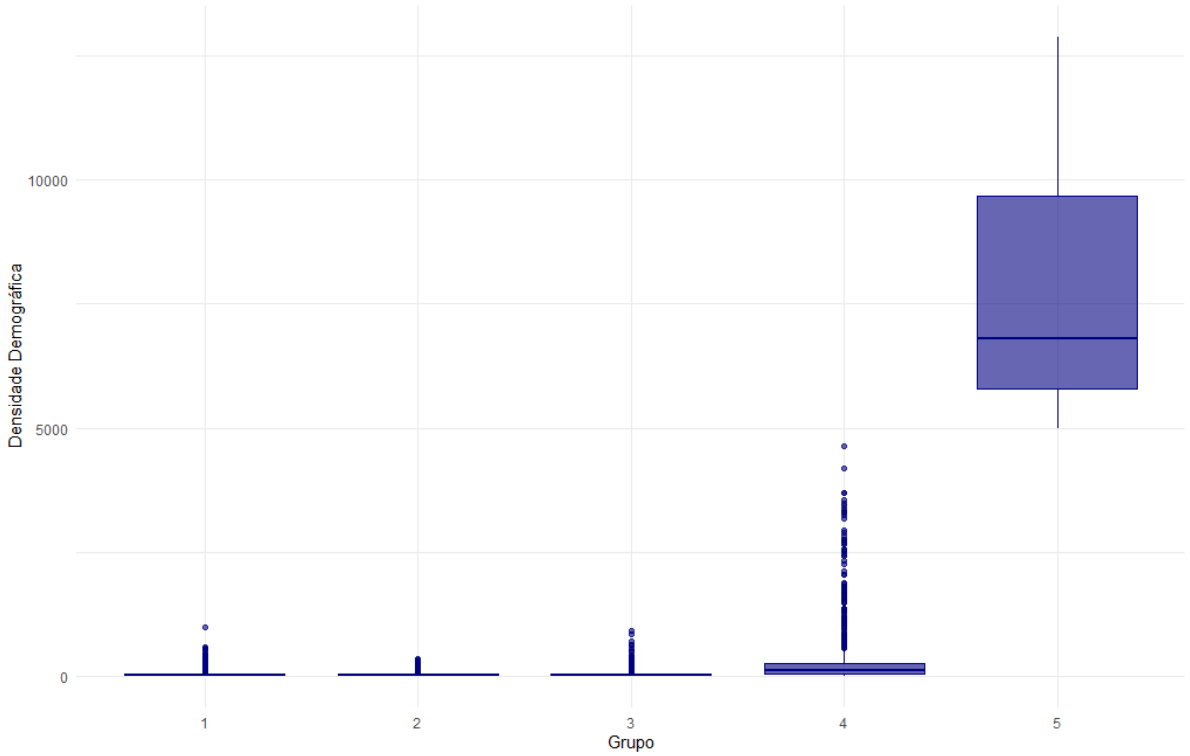


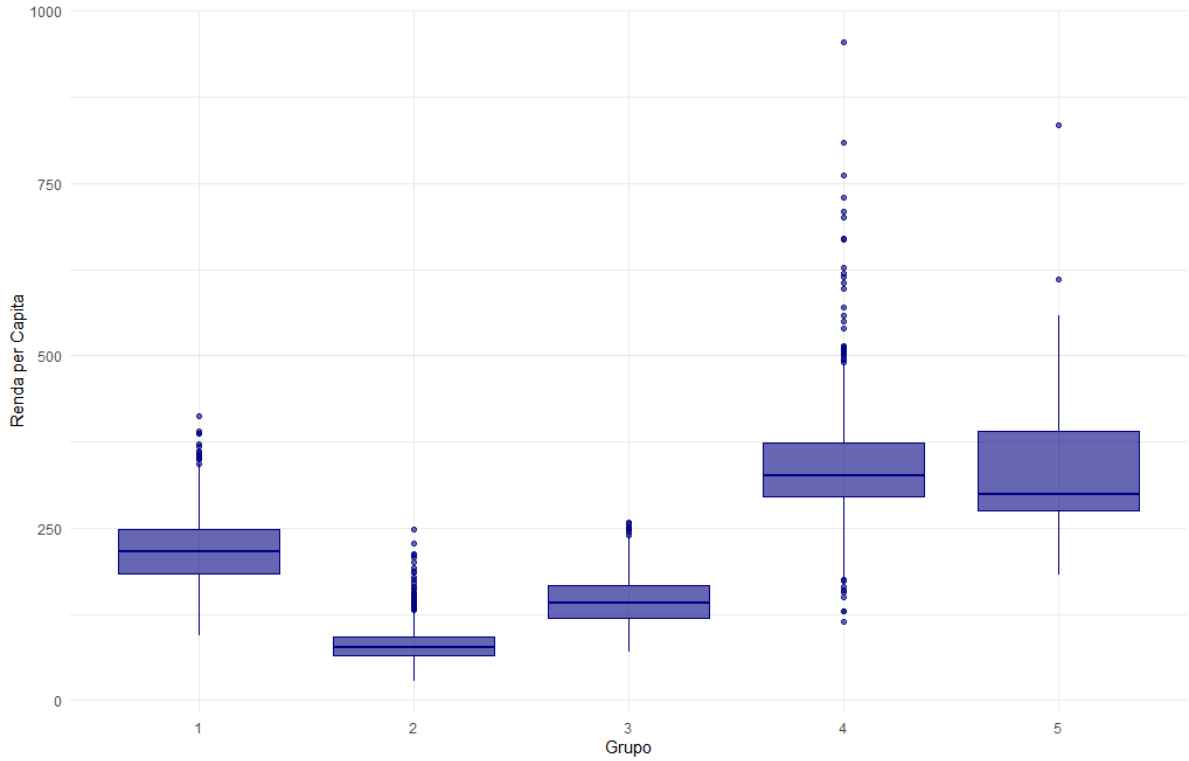


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

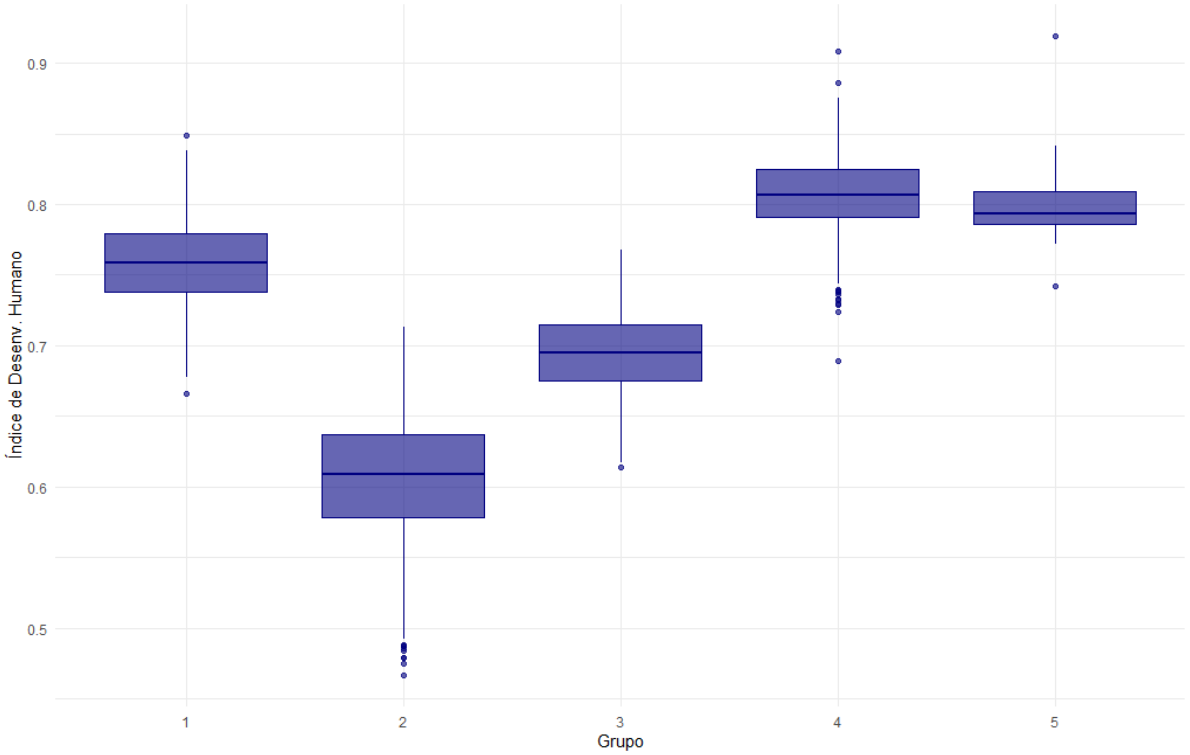


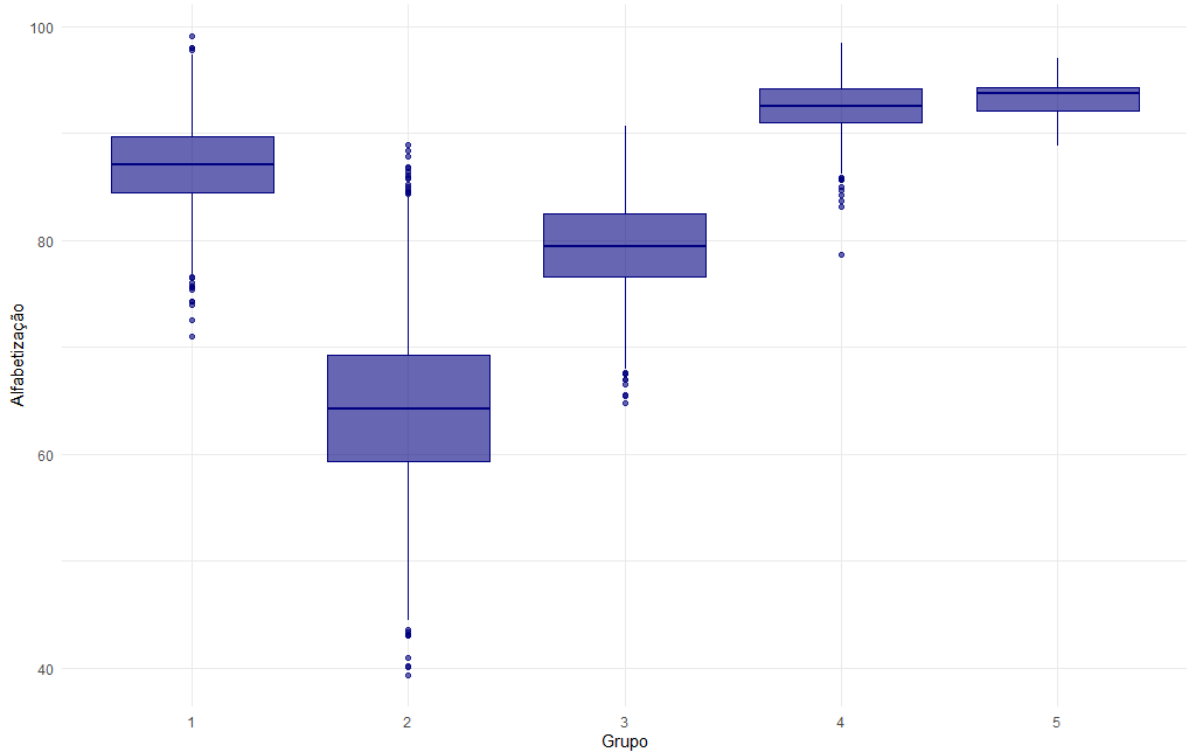


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

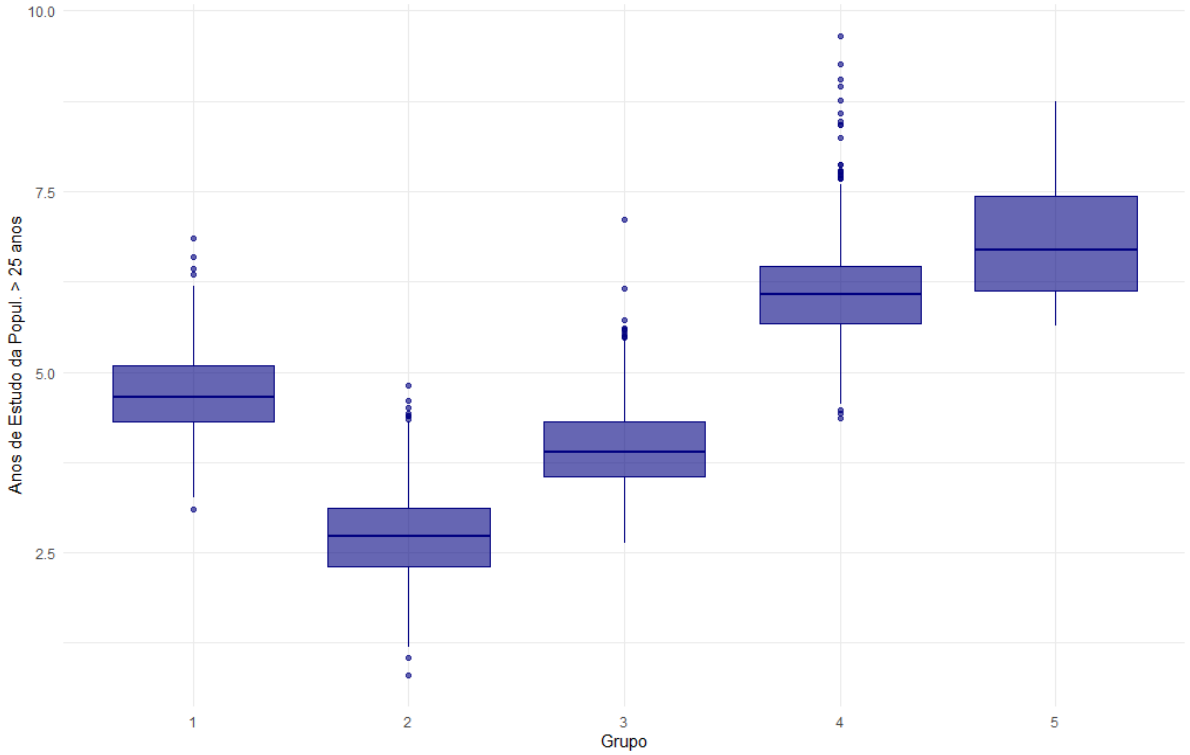


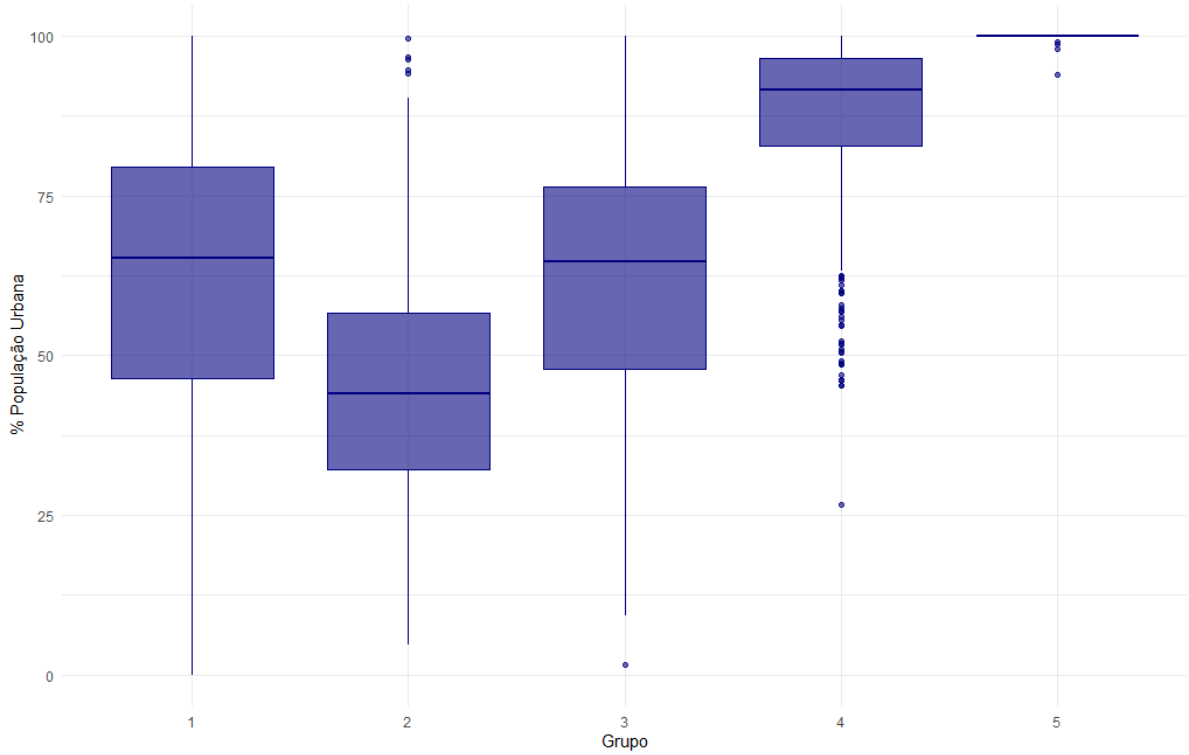


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

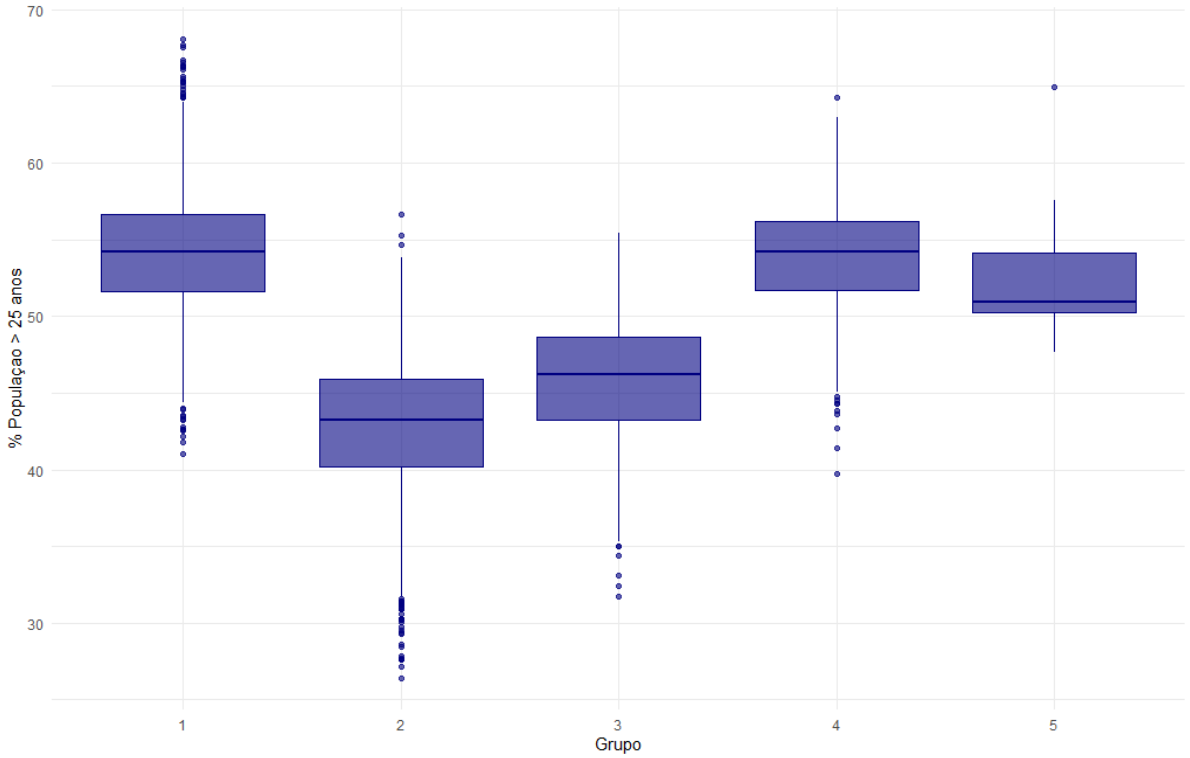


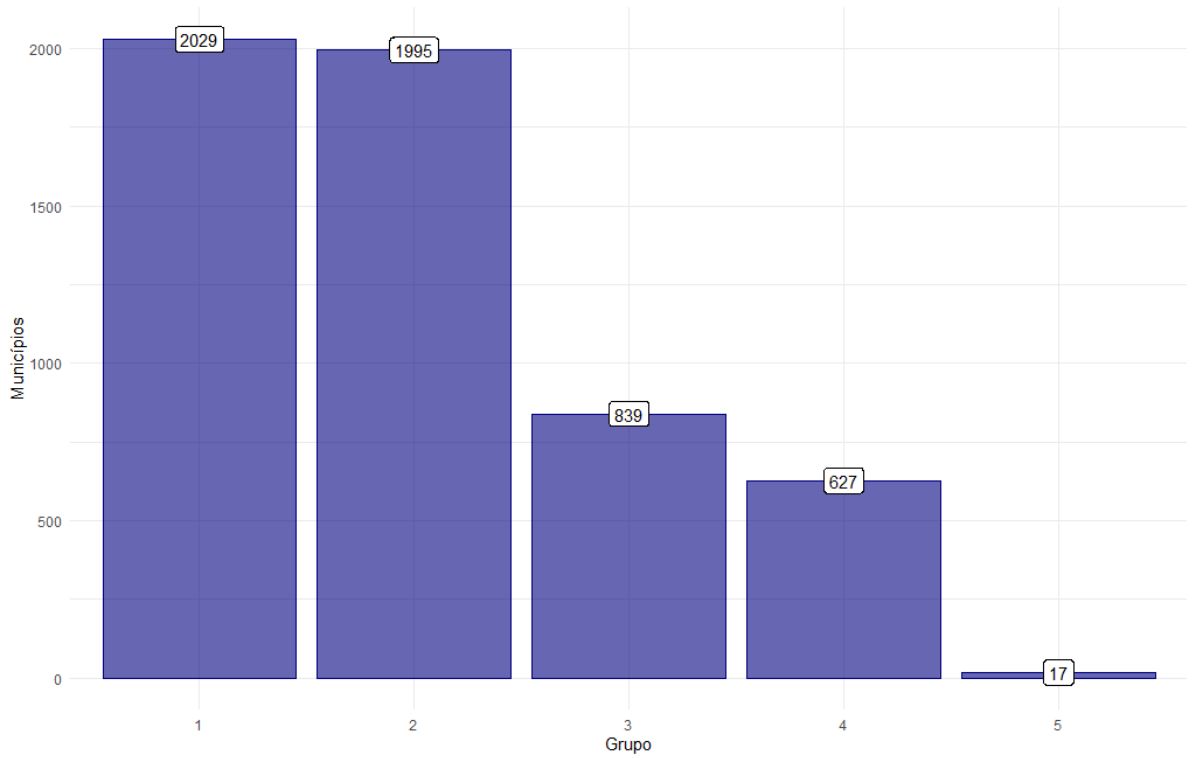


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO



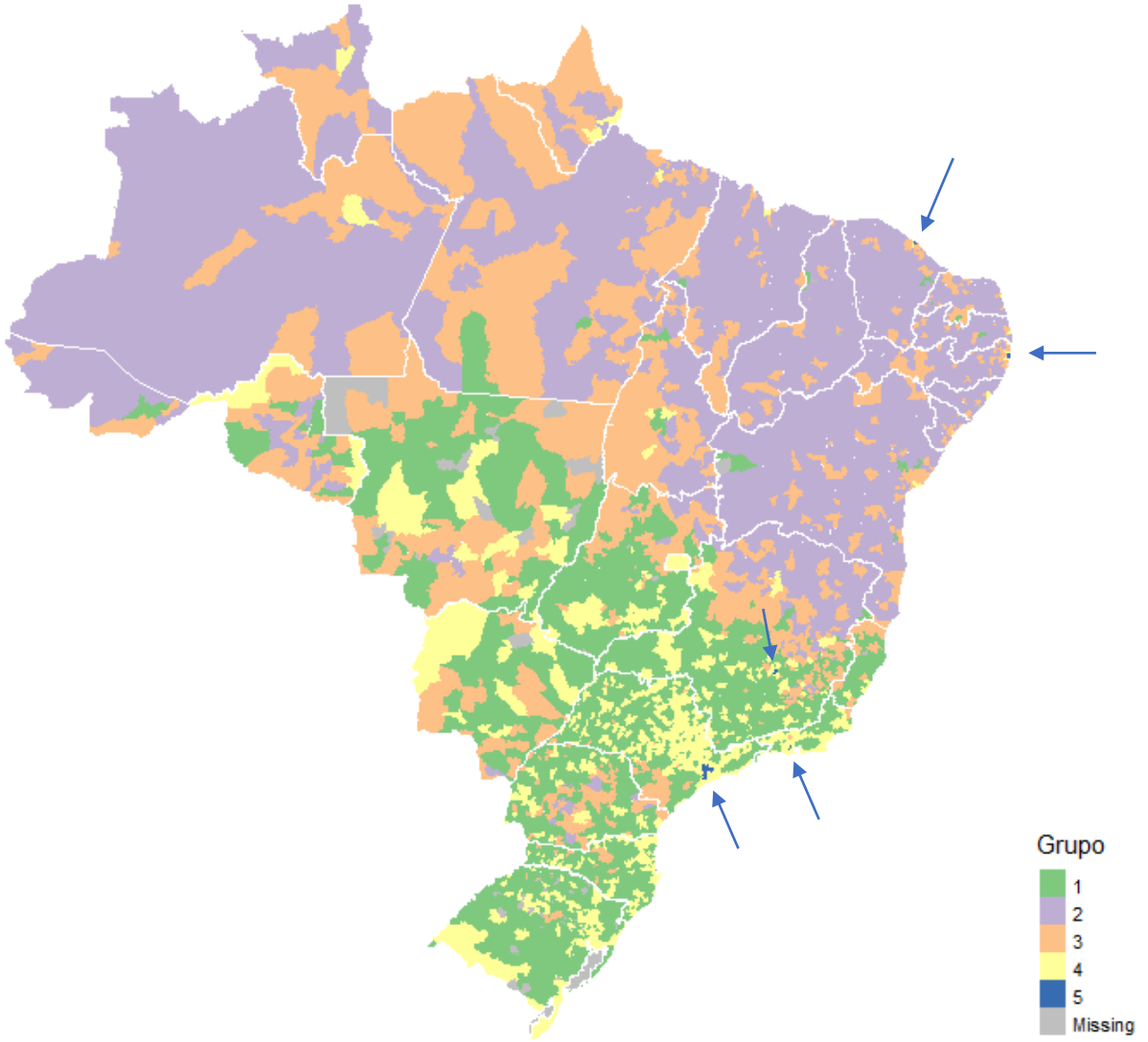


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)

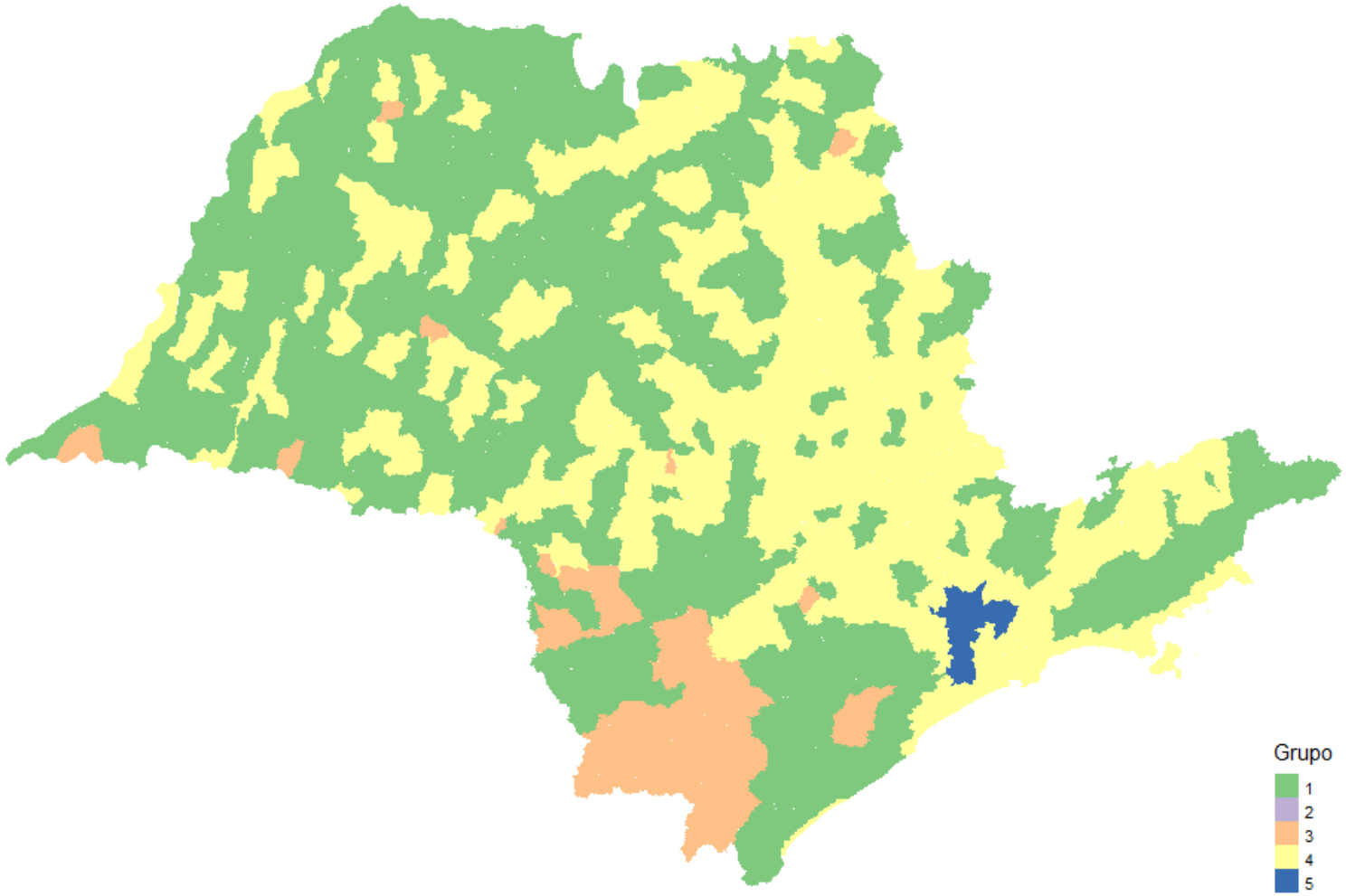


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)

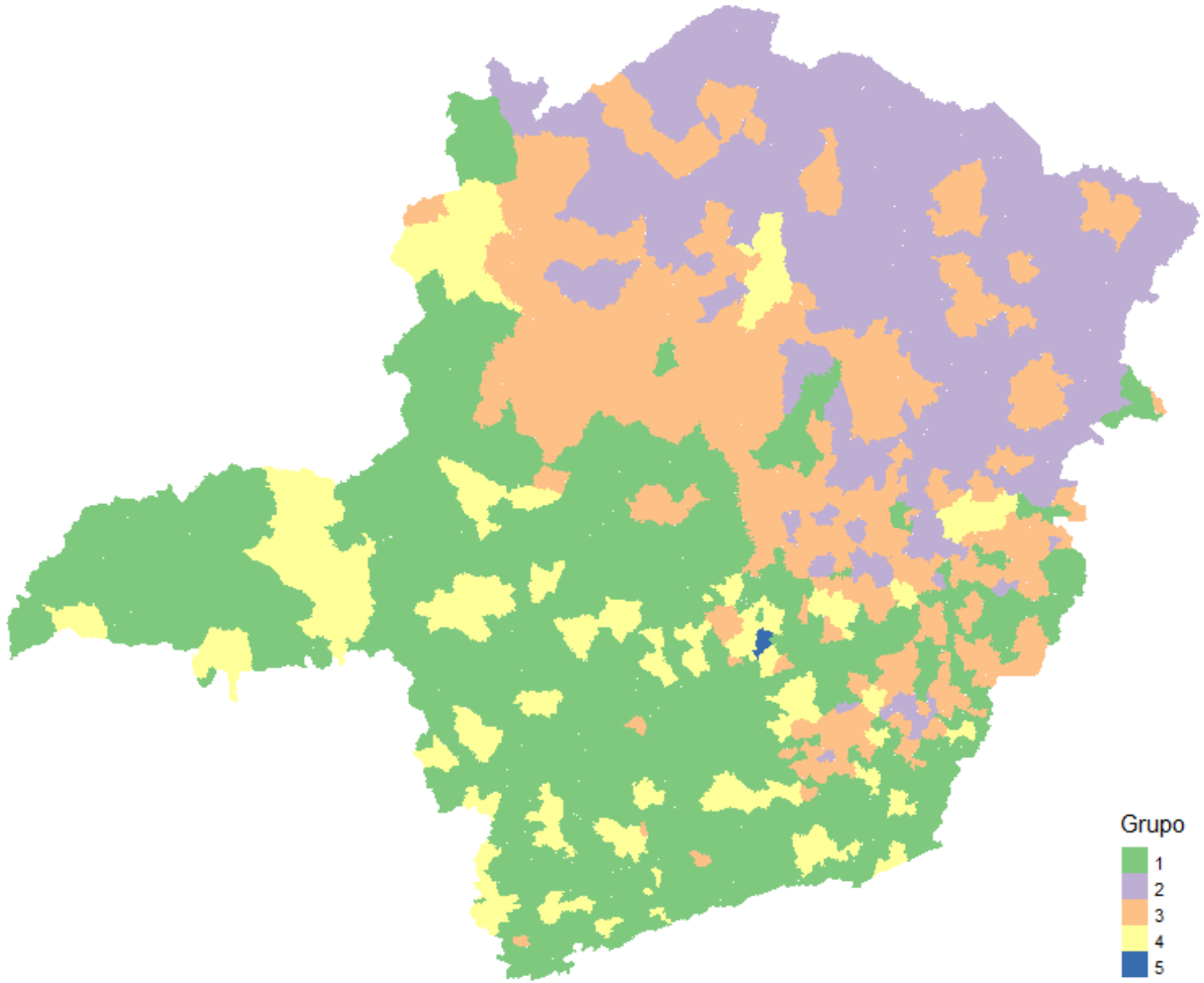


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)

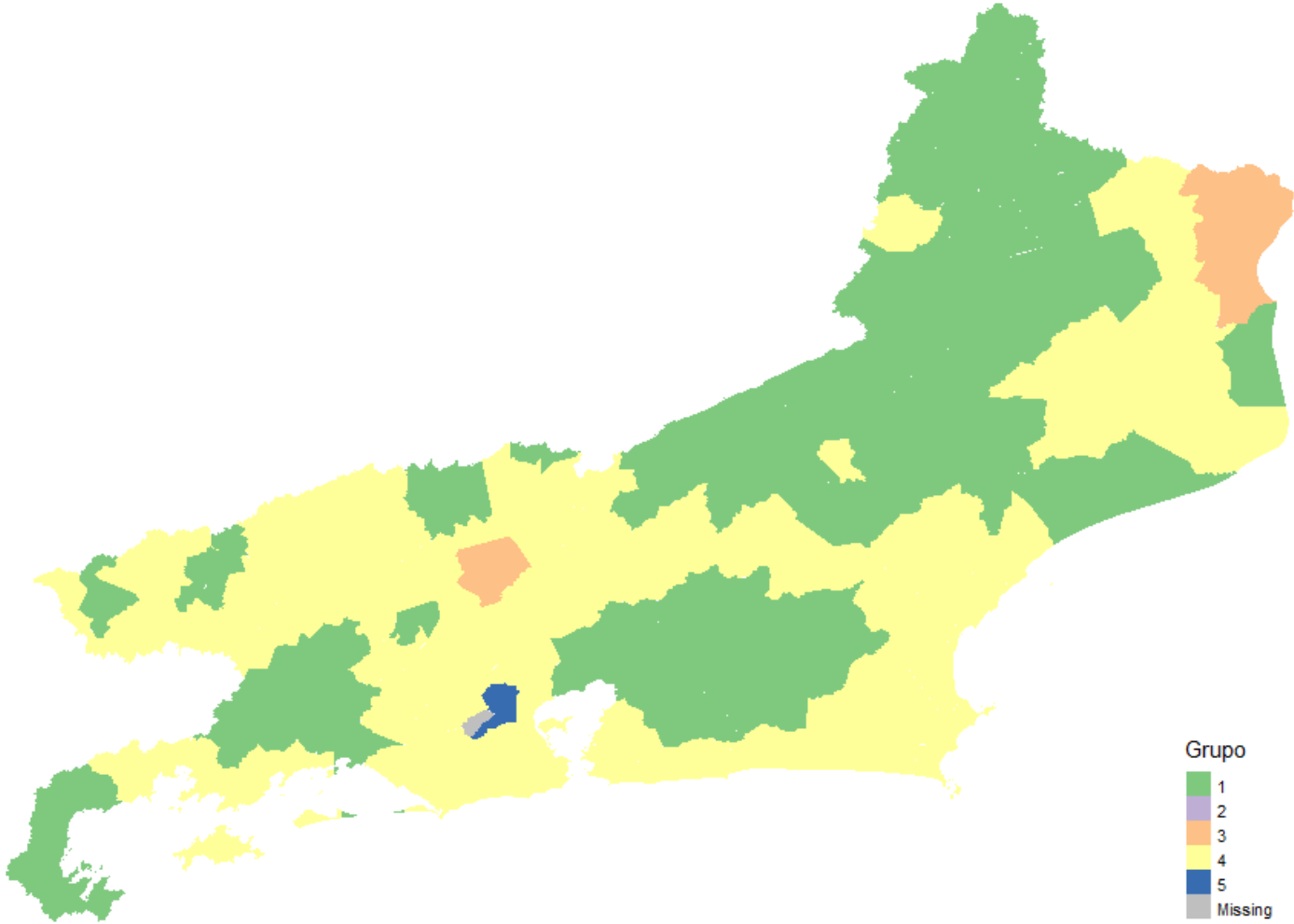


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)

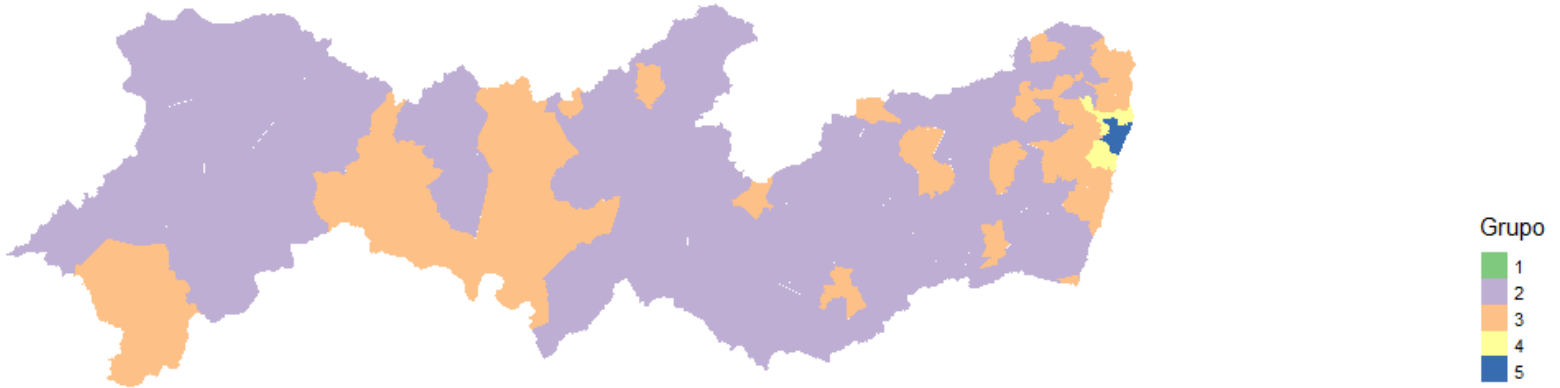


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)

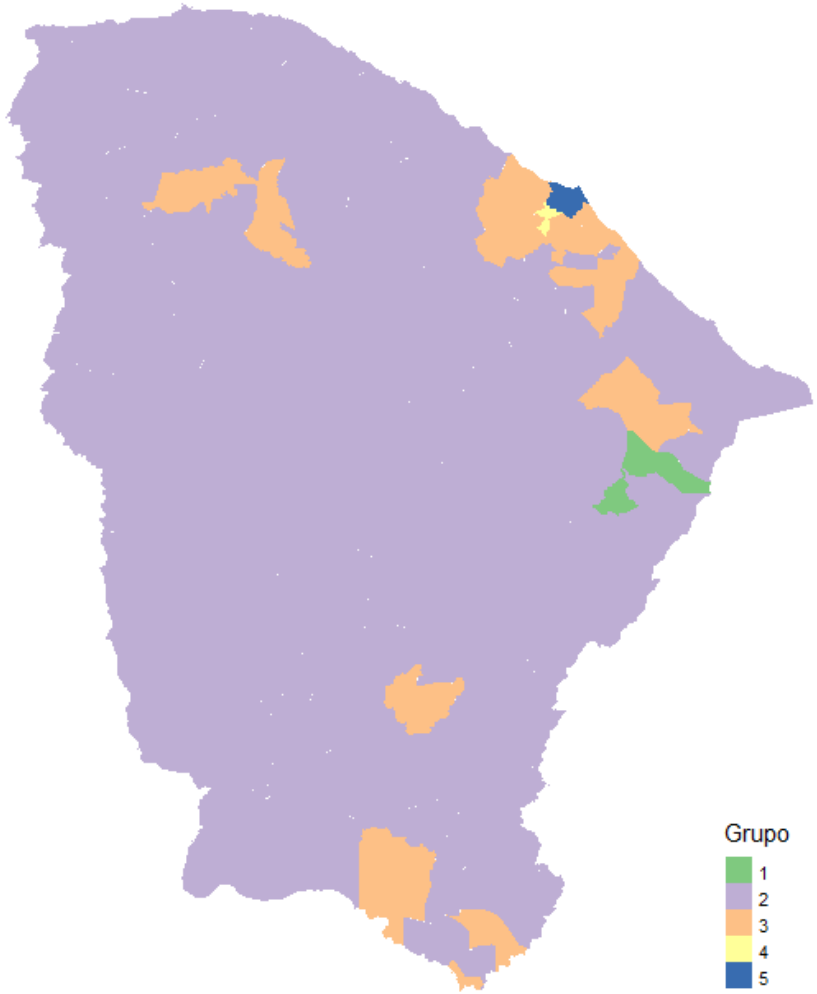


SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

	CODIGO	MUNICIPIO
1	330045	Belford Roxo (RJ)
2	310620	Belo Horizonte (MG)
3	351060	Carapicuíba (SP)
4	351380	Diadema (SP)
5	351570	Ferraz de Vasconcelos (SP)
6	230440	Fortaleza (CE)
7	352500	Jandira (SP)
8	352940	Mauá (SP)
9	330320	Nilópolis (RJ)
10	260960	Olinda (PE)
11	353440	Osasco (SP)
12	353980	Poá (SP)
13	261160	Recife (PE)
14	354880	São Caetano do Sul (SP)
15	330510	São João de Meriti (RJ)
16	355030	São Paulo (SP)
17	355280	Taboão da Serra (SP)



SEGMENTAÇÃO

CLUSTERIZAÇÃO COM 5 GRUPOS

ANÁLISE DO AGRUPAMENTO

Problemas e Melhorias para continuidade

- Problema aparente na predominância da Densidade Demográfica para a separabilidade.
- Importante mais experimentos com e sem a variável.
- Métricas estatísticas distinguiram/criticaram pouco os agrupamentos de 3 a 7.
- Aplicar técnica para capturar características nas variáveis (redução de dimensões).
- Discutir mais o negócio para priorizar características mais aderentes e variáveis.

01
Objetivo e Premissas

02
Análise Exploratória

03
Segmentação

04
Classificação
[10 min]

05
Materiais e Dúvidas

CLASSIFICAÇÃO

PREDIÇÃO DE UM GRUPO

O QUE FIZEMOS?

Criação de um modelo que possibilite indicar a probabilidade de um município pertencer a um grupo.

CLASSIFICAÇÃO

PREDIÇÃO DE UM GRUPO

PREMISSAS DO PRIMEIRO EXPERIMENTO

Utilização das mesmas variáveis da clusterização, sem tratamentos

- Densidade Demográfica.
- Renda per Capita.
- Índice de Desenvolvimento Humano.
- Taxa de Alfabetização.
- Anos de Estudos (População > 25 anos).
- Percentual da População > 25 anos.
- Percentual da População Urbana.

CLASSIFICAÇÃO

PREDIÇÃO DE UM GRUPO

PREMISSAS DO PRIMEIRO EXPERIMENTO

Estratificação usando:

- Divisão de **80% para treino** e 20% para teste.
- Estratificação **aleatória simples** (holdout sampling).

CLASSIFICAÇÃO

PREDIÇÃO DE UM GRUPO

PREMISSAS DO PRIMEIRO EXPERIMENTO

Modelo baseado em Árvore de Classificação (Classification Tree)

- Escolha **do algoritmo C5.0**. como primeiro experimento e benchmark.
- Possibilita **predição dos Grupos e Probabilidade** de cada grupo.
- Método básico de **boosting** para impulsionar desempenho sem exigir poder muito poder computacional, utilizando no **máximo 20 interações**.

CLASSIFICAÇÃO

RESULTADOS

ONDE CHEGAMOS?

1. **Acurácia**: **(VP+VN)/(P+N).** Proporção de valores corretos, sem considerar o que é negativo ou positivo.
2. **Sensibilidade**: **(VP)/(VP+FN).** Proporção de verdadeiros positivos.
3. **Especificidade: (VN)/(FP+VN).** Proporção de verdadeiros negativos.
4. **Eficiência**: **(SENS+ESPECIF)/2.** Média aritmética entre especificidade e sensibilidade, que, normalmente, caminham em direções opostas.
5. **Precisão**: **(VP)/(VP+FP).** Proporção de acerto do modelo de predição.

Observação: É fundamental prestar atenção para o desbalanceamento entre as classes. A classe com menor proporção tende a apresentar piores taxas de classificação. Em casos como esse, avaliar metodologias para modelagem de eventos raros.

Confusion Matrix and Statistics

```
          Reference
Prediction   1   2   3   4   5
         1 387   0   6   6   0
         2   1 386  14   0   0
         3  11  10 149   0   0
         4   3   0   0 125   0
         5   0   0   0   0   4

Overall Statistics

               Accuracy : 0.9537
                 95% CI : (0.9396, 0.9654)
    No Information Rate : 0.3648
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9339

 Mcnemar's Test P-Value : NA
```

Statistics by Class:

```
                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.9627   0.9747   0.8817   0.9542  1.00000
Specificity            0.9829   0.9788   0.9775   0.9969  1.00000
Pos Pred Value         0.9699   0.9626   0.8765   0.9766  1.00000
Neg Pred Value         0.9787   0.9857   0.9785   0.9938  1.00000
Balanced Accuracy      0.9728   0.9768   0.9296   0.9756  1.00000
```

01
Objetivo e Premissas

02
Análise Exploratória

03
Segmentação

04
Classificação

05
Materiais e Dúvidas
[? min]

MATERIAIS E DÚVIDAS

DOCUMENTAÇÕES

GitHub da Solução

https://github.com/Vilson-Ferreira/Plusoft-Analise.git

Apresentação

Resultados/Apresentacao.pdf

plusoft

Analytics e Data Science

Fevereiro, 2023