Case - Data Scientist Plusoft

William de Oliveira Nery

Março/2022

Agenda:

1 - Contexto:

2 - Desafio:

3 - Desenvolvimento da Solução:

4 - Conclusão e Demonstração:

5 - Próximos passos:

1 - Contexto:

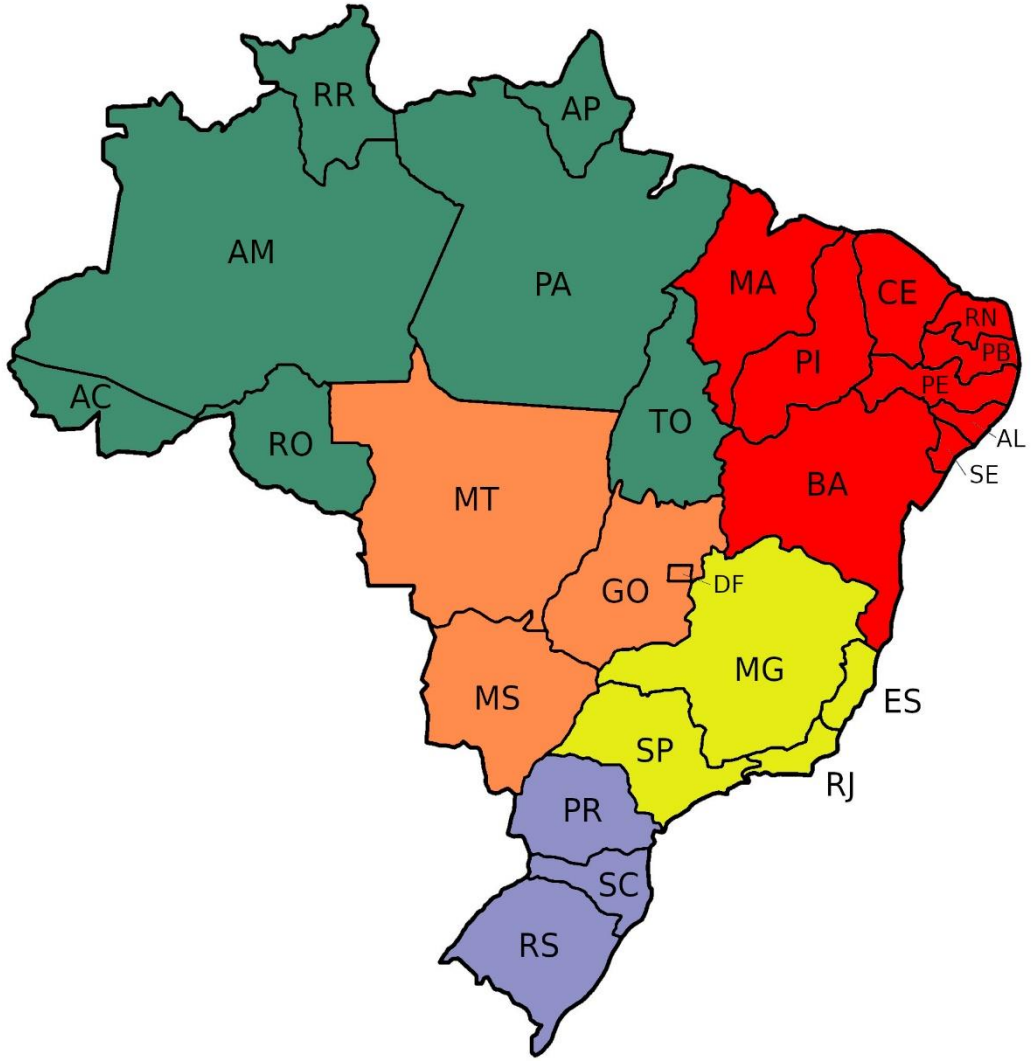
Elaborar estratégia de entrada de multinacional varejista do ramo de supermercados no mercado brasileiro, com base nos dados disponíveis.

2 - Desafio:

- Classificação dos municípios brasileiros em grupos.

- Classificar a entrada de um novo município entre os grupos já criados.

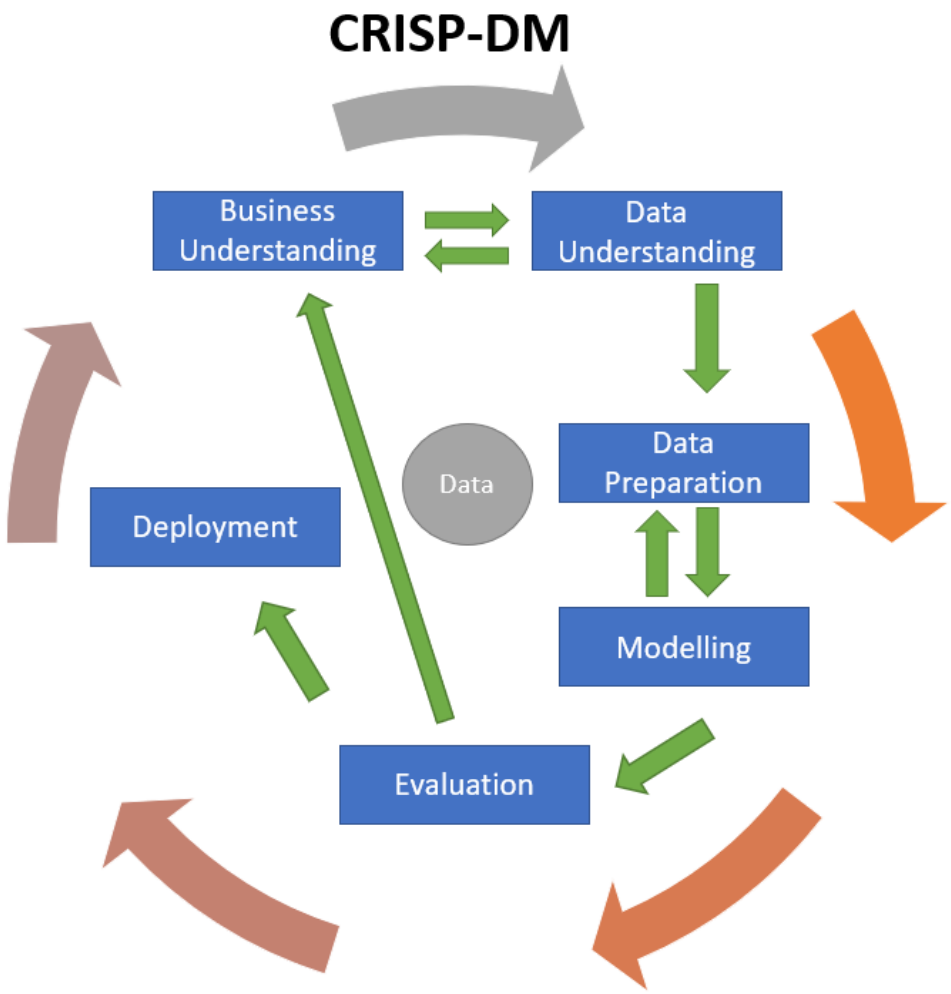
- Quais grupos de municípios devem ser a porta de entrada no país e porque.



3- Desenvolvimento da Solução:

Metodologia CRISP-DM

- 3.1: Análise preliminar da base de dados
- 3.2: Criação de features
- 3.3: Utilização de algoritmo de Machine Learning para identificação dos grupos
- 3.4: Análise exploratória dos dados
- 3.5: Preparação dos dados
- 3.6: Seleção de Features
- 3.7: Aplicação do algoritmo de ML
- 3.8: Deploy do Modelo



3.1 Análise preliminar da base de dados:

Características da base de dados:

- Quantidade de municípios: 5507
- 25 features relacionadas a:
  - Natalidade
  - Alfabetização
  - População
  - Desenvolvimento sócio-econômico

	A	B	C	D	E	F
1	Código	Município	Área (km²)	Densidade demográfica, 2000	Distância à capital (km)	Esperança de vida ao nascer, 2000
2	355030	São Paulo	1528,5	6808,1	0	70,66
3	330455	Rio de Janeiro	1264,2	4627,9	0	70,26
4	292740	Salvador	709,5	3440,3	0	69,64
5	310620	Belo Horizonte	331,9	6718	0	70,52
6	230440	Fortaleza	313,8	6814	0	69,63
7	530010	Brasília	5822,1	350,9	0	70,37
8	410690	Curitiba	430,9	3682,8	0	71,57
9	261160	Recife	218,7	6501,8	0	68,62
10	130260	Manaus	11458,5	122,5	0	67,65
11	431490	Porto Alegre	496,1	2741,2	0	71,48
12	150140	Belém	1070,1	1196	0	70,5
13	520870	Goiânia	743	1467,8	0	70,06
14	351880	Guarulhos	317,9	3369,9	14,11985556	69,27
15	350950	Campinas	797,6	1213,5	83,5023372	72,22
16	330350	Nova Iguaçu	559,4	1636,3	29,58690144	67,99
17	330490	São Gonçalo	251,3	3540,9	17,88699208	69,51
18	211130	São Luís	831,7	1043,3	0	69,19
19	270430	Maceió	512,8	1553,9	0	65,03
20	330170	Duque de Caxias	465,7	1655,3	16,80624065	67,49
21	221100	Teresina	1679,8	425,2	0	69,06
22	240810	Natal	169,9	4175,5	0	68,78
23	354870	São Bernardo do Campo	407,1	1720,5	17,76101981	69,93

3.2 Criação de Features:

Criação de 4 Features:

- Receita da população disponível pra gasto com mercado: gasto médio mensal do brasileiro com itens adquiridos em supermercados: 21,81%

- Latitude e Longitude dos municípios(1 coluna para cada) : para ver impactos com relação a posição geográfica

- Crescimento Populacional de 1991 a 200: para avaliar em quais aspectos um alto crescimento ou um baixo crescimento podem afetar

TABELA 1
Evolução do índice de peso no consumo final das famílias, por setor
(Em %)

Setor	2000	2005	2009
Alimentos *in natura*	7,29	7,73	7,56
Alimentos industrializados	12,40	13,19	12,85
Vestuário	6,62	6,06	6,40
Combustíveis	5,44	5,45	5,24
Produtos farmacêuticos	2,41	2,67	2,64
Perfumaria, sabões e artigos de limpeza	2,12	2,13	2,22

```
## Alimentos in natura: 7,29% + Alimentos industrializados: 12,40% + Sabões e artigos de limpeza: 2,12 = 21,81%
df2['receita_pop_mercado_2000']  = (df2['População total, 2000'] * df2['Renda per Capita, 2000'] ) * 0.22
```

```
df2['cresc_popu_1991_2000'] = ( df2['População total, 2000'] - df2['População total, 1991']) / df2['População total, 1991']
```

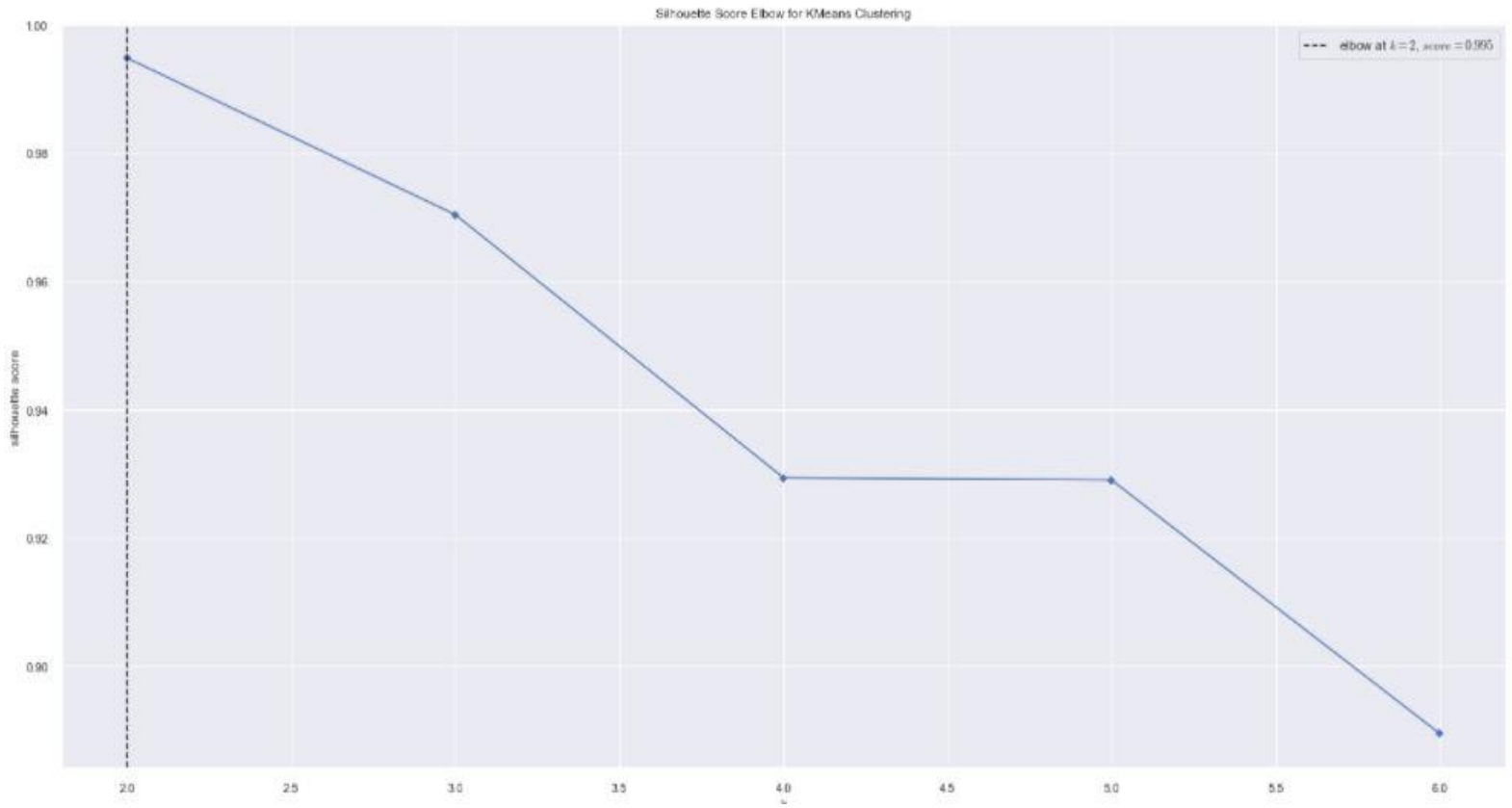
3.3 Algoritmo de ML para clusterização preliminar:

Algoritmo utilizado: KMeans

Métrica: Silhouette Score

Para K = 6 -> SS = 0.87.

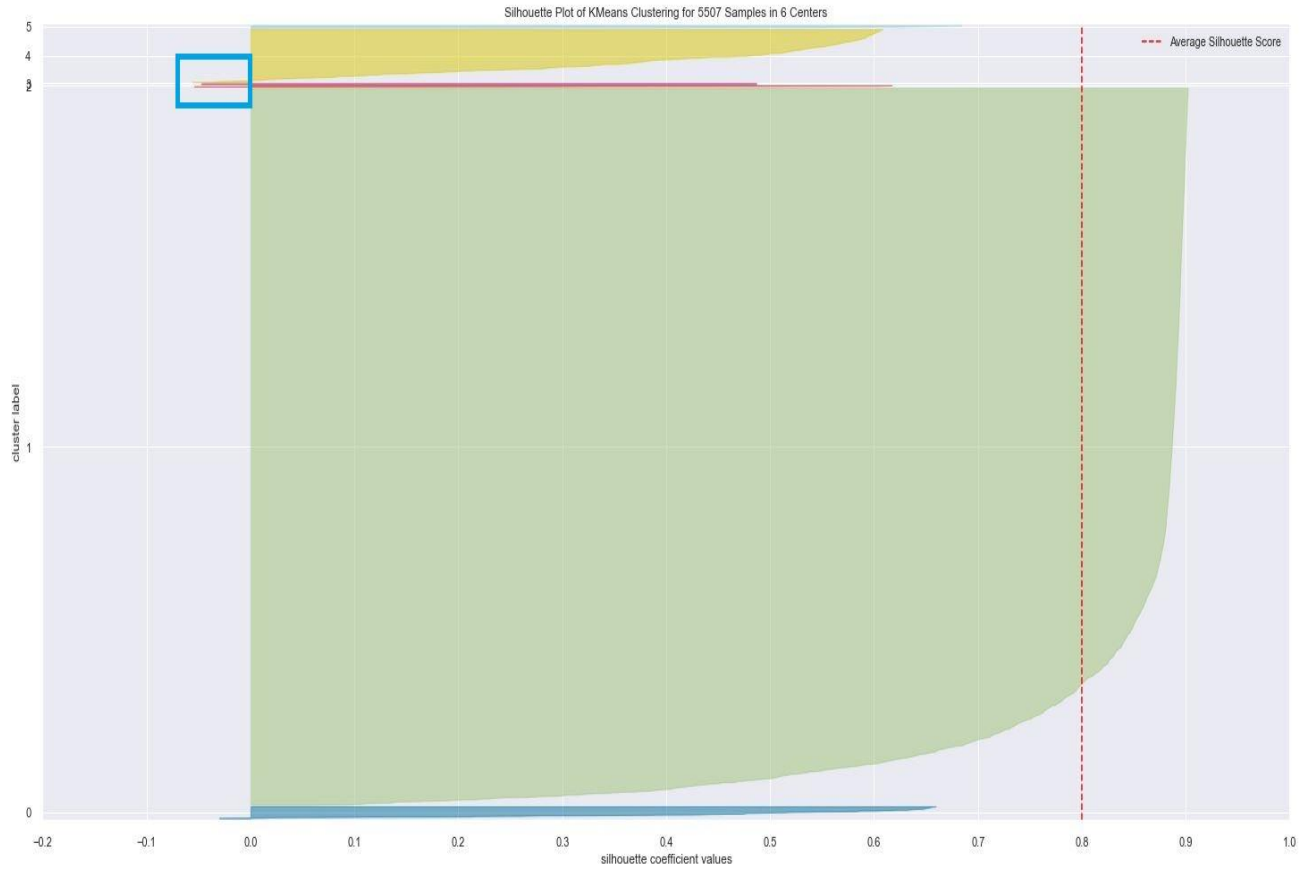
Como para k = 6, ainda temos uma alta acurácia, utilizaremos esse valor para a separação dos grupos.



3.3 Algoritmo de ML para clusterização preliminar:

Análise dos cluster: Silhouette Score

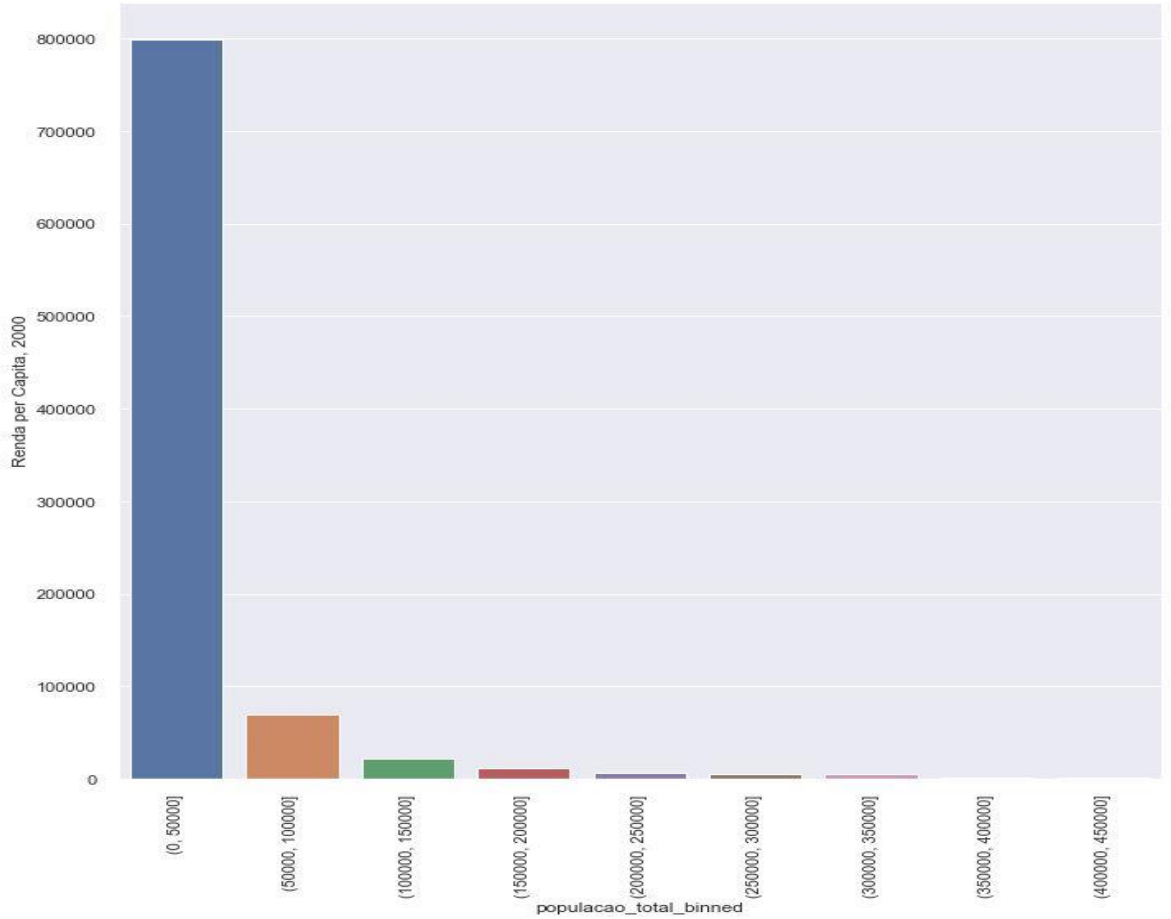
- Um dos grupos possui uma grande quantidade de municípios (cluster verde)
- Pelo ‘Average Silhouette Score’, também podemos observar que nesse cluster é aonde temos elementos dos grupos mais definidos.
- Mesmo para os grupos pequenos, podemos observar que alguns pontos ainda não foram completamente isolados dos outros grupos (retângulo azul).



3.4 Análise exploratória dos dados(EDA):

Teste de Hipóteses:

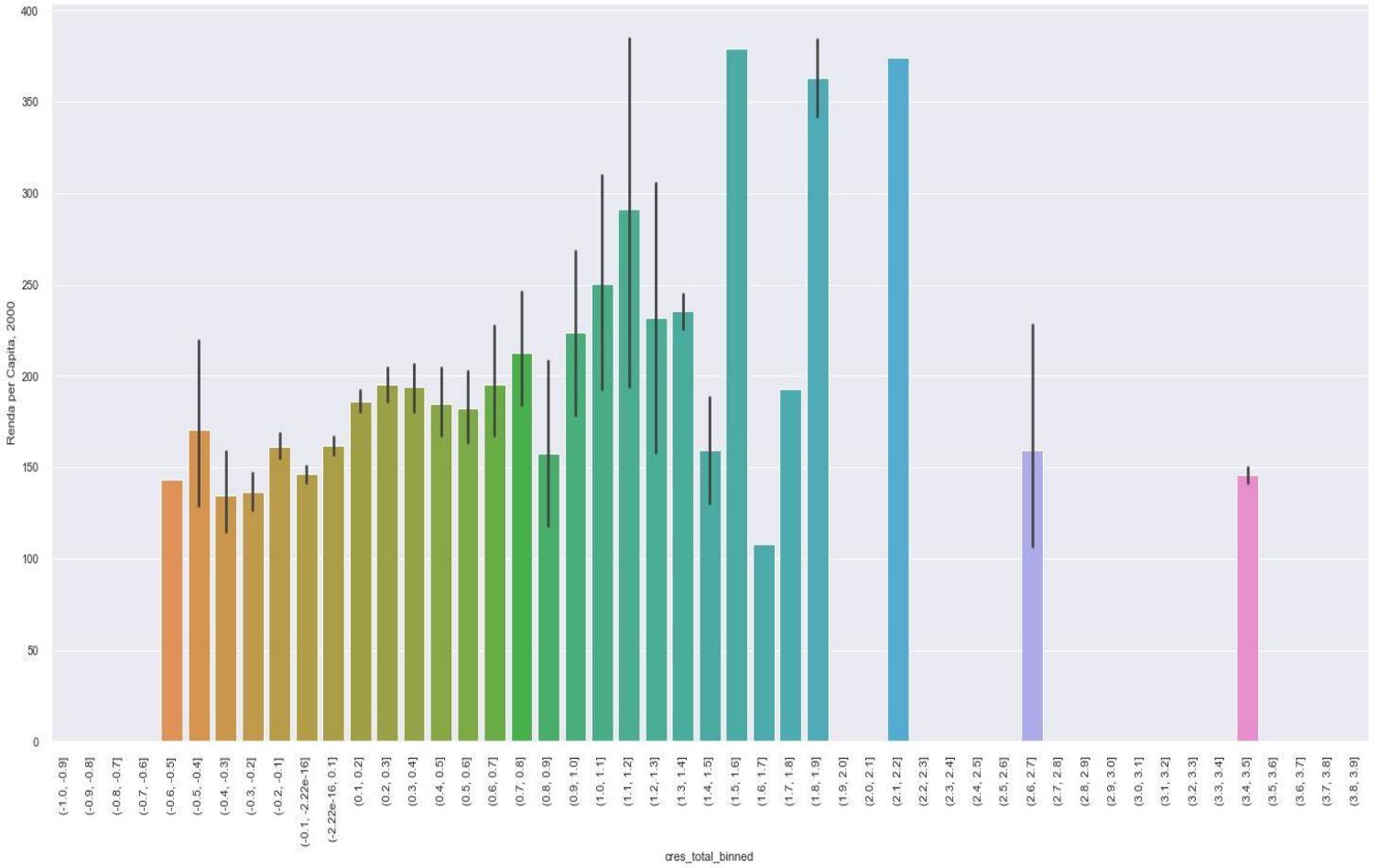
- H1: A renda per Capita é bem distribuída pela população total.
- H1 – Falso: a Renda per Capita está muito concentrada em uma pequena parcela da população.



3.4 Análise exploratória dos dados(EDA):

Teste de Hipóteses:

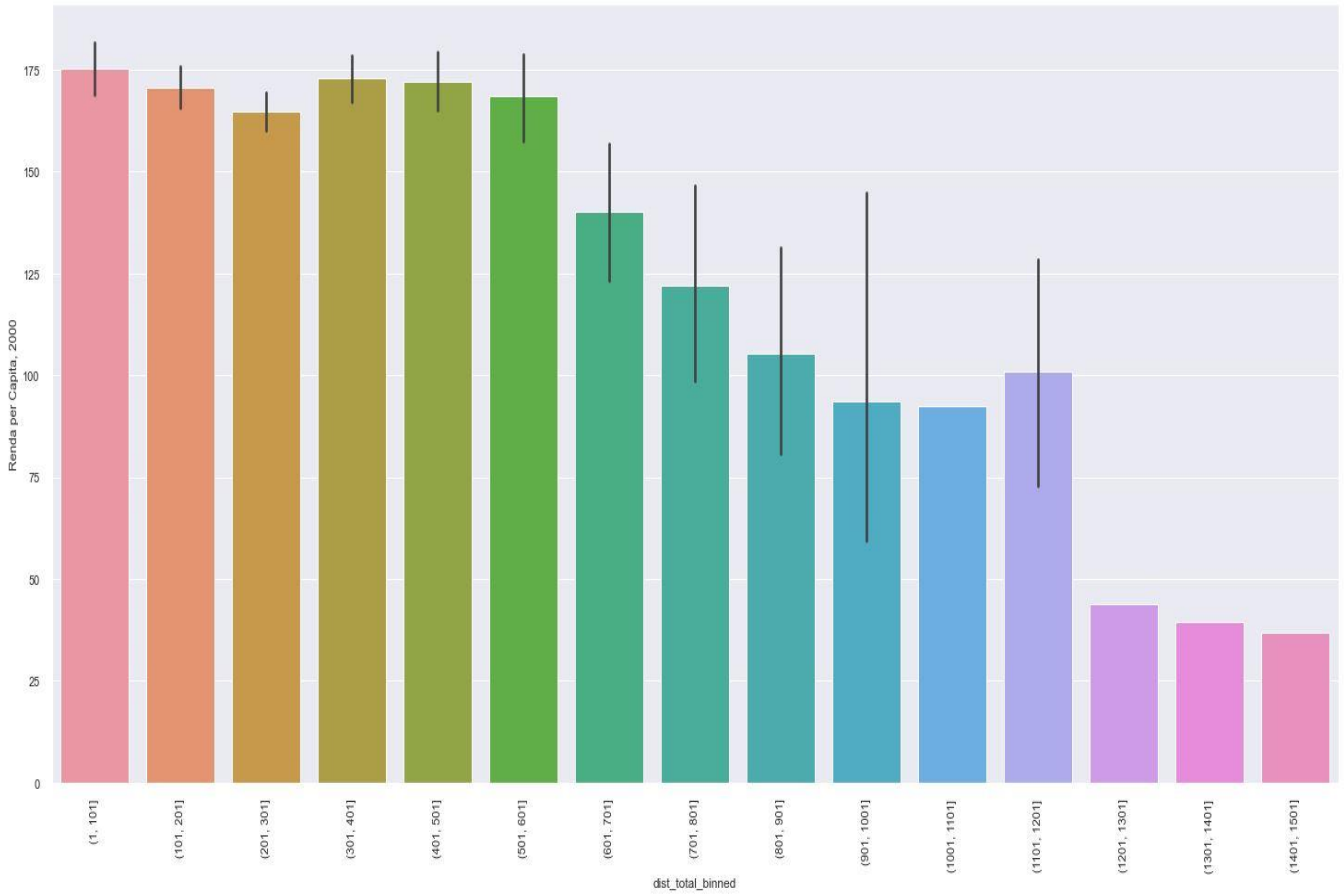
- H2: Municípios com maior crescimento populacional(1991-2000), apresentam melhor Renda per Capita.
- H2 - Falso: alguns municípios que apresentaram queda no crescimento(1991-2000), tem Renda per Capita maior do que municípios que obtiveram crescimento.



3.4 Análise exploratória dos dados(EDA):

Teste de Hipóteses:

- H3: Municípios mais distantes da capital possuem Renda per Capita menor.
- H3 - Verdadeiro: quanto maior a distância em relação a capital, a Renda per Capita tende a ser menor.

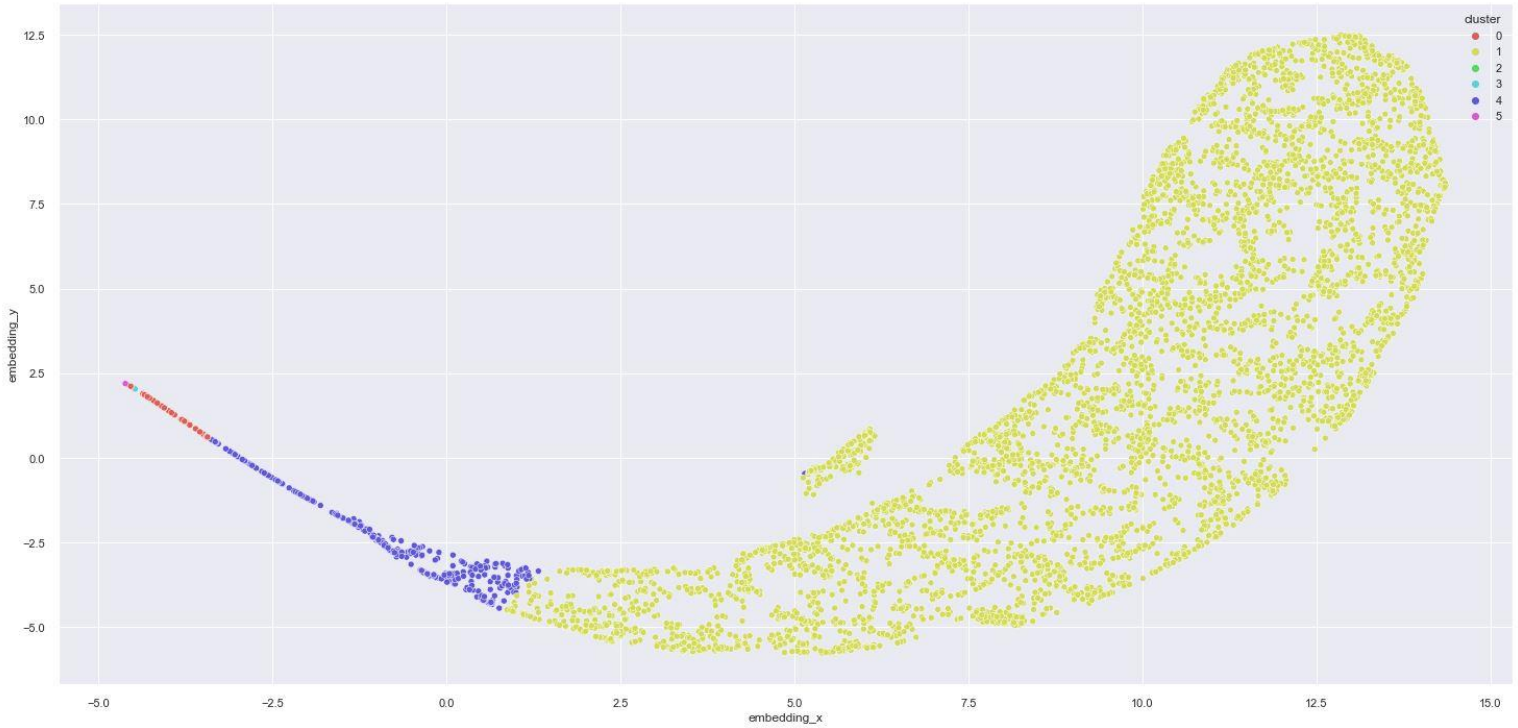


3.5 Preparação dos dados:

- Ao utilizar técnicas de normalização (StandardScale, MinMaxScaler) a projeção dos dados após aplicar o algoritmo fica melhor, porém nas áreas de interseção dos grupos, não fica muito claro as premissas de separação.

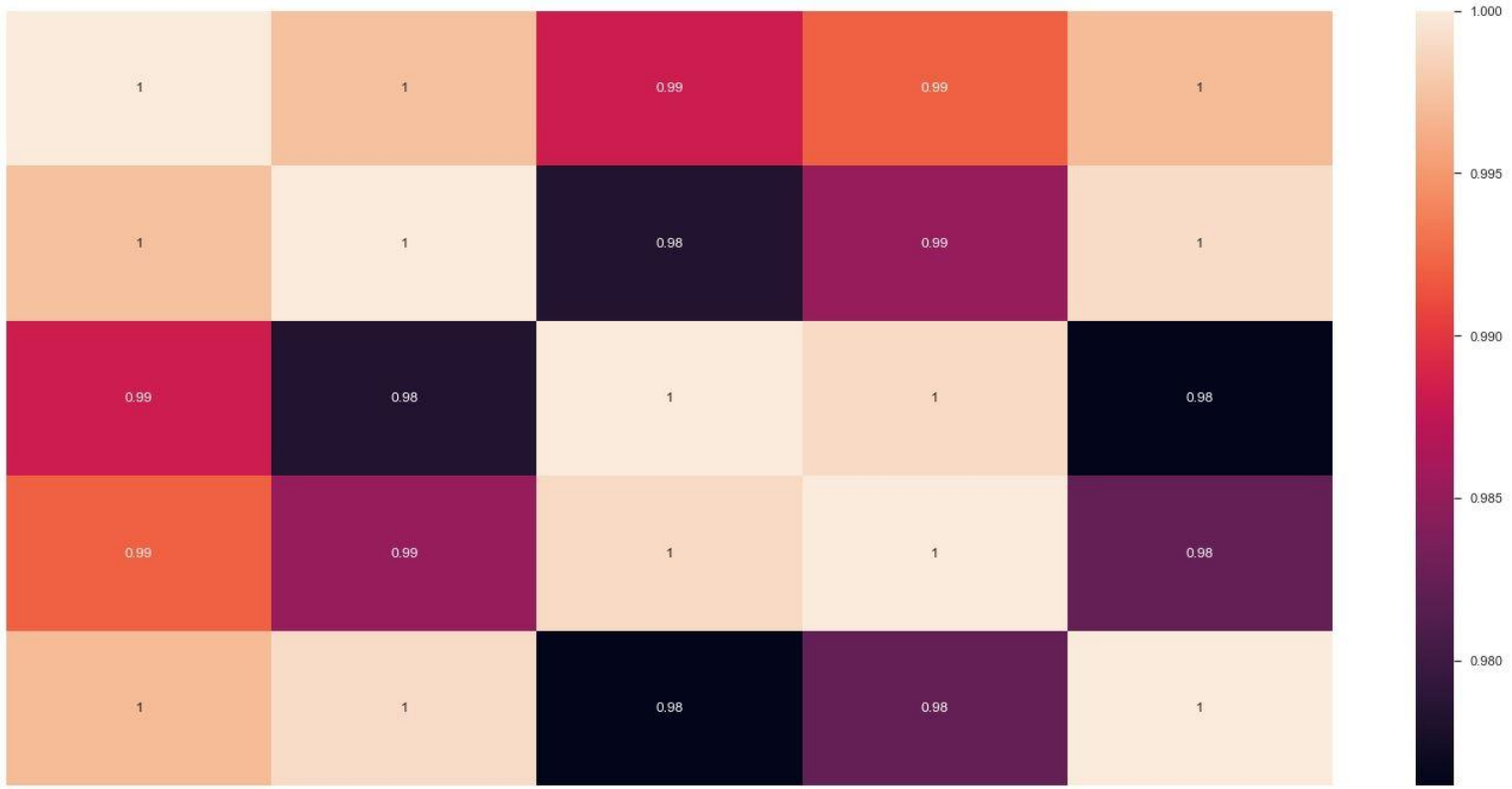
- A padronização será aplicada na próxima rodada do CRISP-DM.

- O agrupamento dos clusters na imagem ao lado, só foi possível utilizando uma biblioteca de projeção de dados de alta dimensionalidade, chamada “umap”.



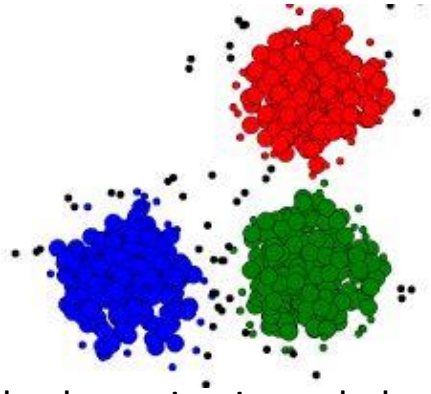
3.6 Seleção de features:

- Foram retiradas as 5 features relacionadas a população. Elas apresentam forte correlação entre si (>= 0.98%), e a seção 3.2, já foi criada uma feature visando avaliar crescimento populacional.
- Features retiradas:
- 'População de 25 anos ou mais de idade, 1991',
- 'População de 25 anos ou mais de idade, 2000',
- 'População de 65 anos ou mais de idade, 1991',
- 'População de 65 anos ou mais de idade, 2000',
- 'População total, 1991'



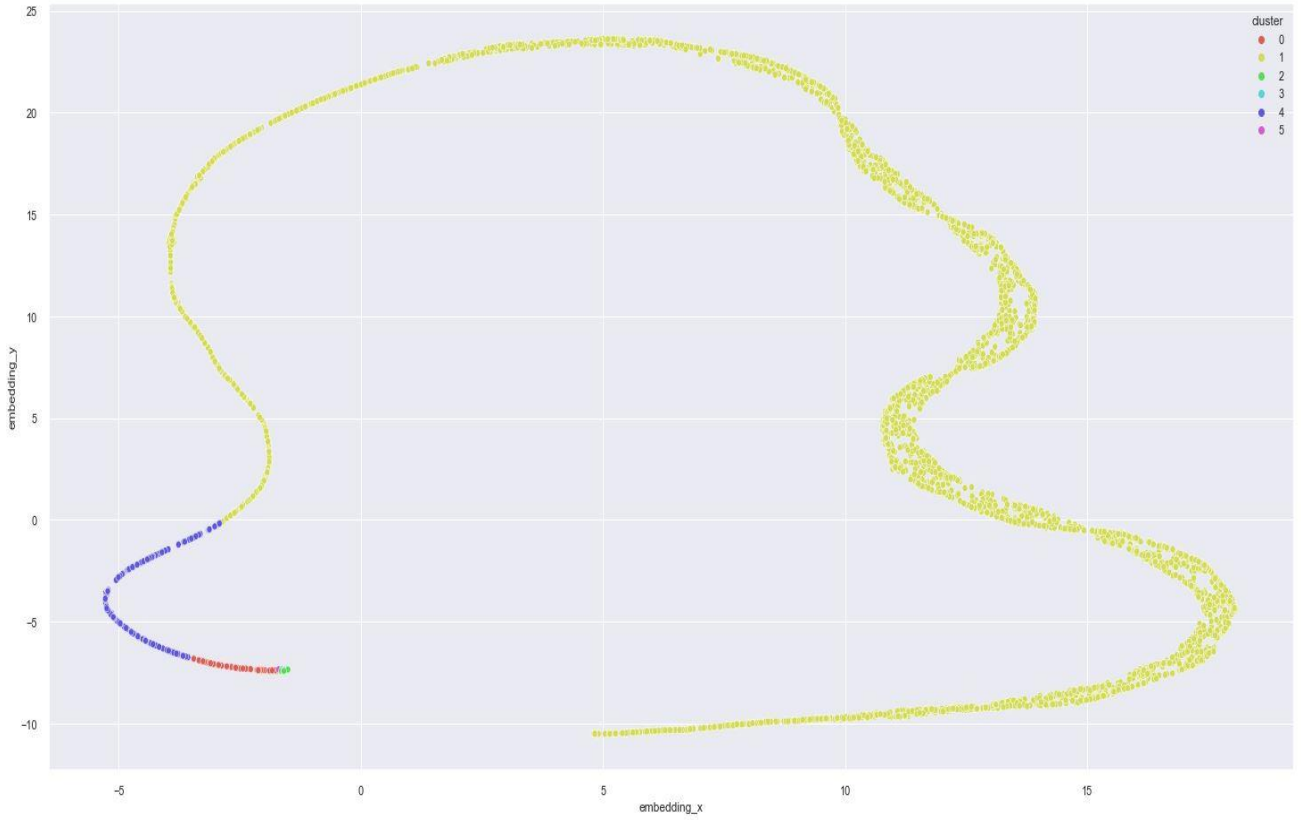
3.7 Aplicação do Algoritmo de Machnie Learning:

- Algoritmo utilizado: KMeans
- Expectativa ao rodar um modelo de clusterização:



- Como desde a primeira rodada para achar o valor de K, a base de dados já estava limpa, o KMeans continua sendo uma boa opção.

Realidade:



3.8 Deploy do Modelo:

- O modelo foi disponibilizado em produção, via Heroku. Segue link de acesso:
- https://plusoft-app.herokuapp.com

4.0 Conclusão e Demonstração

- A estratégia de entrada no mercado brasileiro deve seguir a seguinte ordem:

- 1° Grupo: 5
- 2° Grupo: 3
- 3° Grupo: 2
- 4° Grupo: 0
- 5° Grupo: 4
- 6° Grupo: 1

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

Premissas adotadas

Junção de 3 fatores:

- Atingir o máximo de pessoas em uma quantidade menor de municípios.
- Renda per Capita mais elevada, para maiores gastos em itens de supermercado.
- IDHM alto, visando manter o negócio no longo prazo.

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

1° Grupo de entrada: Grupo 5

- Municípios : São Paulo (SP) e Rio de Janeiro(RJ)
- Com a entrada apenas nesses 2 munícipios, atingimos um total de 9,59% da população total.
- A 2° maior receita/mês para compras disponível dos grupos.
- Maior IDHM médio dos grupos.

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

2° Grupo de entrada: Grupo 3

- Municípios : Com a entrada nesses 9 munícipios , atingimos um total de 9,01% da população total.
- Em média R$186,576 milhões/mês por município de receita prevista para itens de supermercado.
- 2° maior IDHM médio dos grupos.

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

3° Grupo de entrada: Grupo 2

- Municípios : Com a entrada nesses 30 munícipios , atingimos um total de 11,25% da população total.
- Em média R$54,374 milhões/mês por município de receita prevista para itens de supermercado.
- 3° maior IDHM médio dos grupos.

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

Para os demais grupos: 0, 1 e 4

- Como o número de municípios começa a aumentar muito, a estratégia será:
- Posicionar as lojas em municípios estratégicos, seguindo as premissas adotadas e não tão distantes da respectiva capital.

	cluster	numero_de_municipios	População total, 2000	perc_populacao, 2000	receita_pop_mercado_2000	IDHM_medio
0	0	67	18691596	11.008061	1.365543e+09	0.803343
1	1	5135	72808636	42.879265	2.562848e+09	0.692341
2	2	30	19118581	11.259526	1.631235e+09	0.818600
3	3	9	15307394	9.014999	1.680808e+09	0.830667
4	4	264	27580807	16.243193	1.714149e+09	0.789943
5	5	2	16292156	9.594956	2.169294e+09	0.841500

5 Próximos Passos:

- Estratégia de realocação das lojas dos grupos 0, 1 e 4.
- Padronização das features.
- Rodar algoritmo “Boruta” para melhor seleção de features.
- Tunagem-Fina de hiperparâmetros.

Obrigado !

Referências:

- http://repositorio.ipea.gov.br/bitstream/11058/6779/1/TD_2209.pdf
- https://servicodados.ibge.gov.br/api/docs/localidades