ANÁLISE DA PRESENÇA FEMININA NOS ÚLTIMOS ANOS DO ENSINO MÉDIO DE 2019 NOS MUNICÍPIOS DO RN

João Pedro de Souza e Silva
Luiza Lorena Toscano de Medeiros
Thatiana Jéssica da Silva Ribeiro
Volney Lourenço dos Santos Oliveira

Recapitulando...

Objetivo da pesquisa: aplicar os conhecimentos aprendidos durante a disciplina de Ciência de Dados para fazer o agrupamento e traçar o perfil de algum problema real.

Nas primeiras discussões, decidiu-se explorar os dados referentes as turmas dos últimos anos do ensino médio, tendo um foco na presença feminina (do ponto de vista de sexo biológico).

Brainstorming : Perguntas

01. Presença feminina vs masculina nos últimos anos do ensino médio

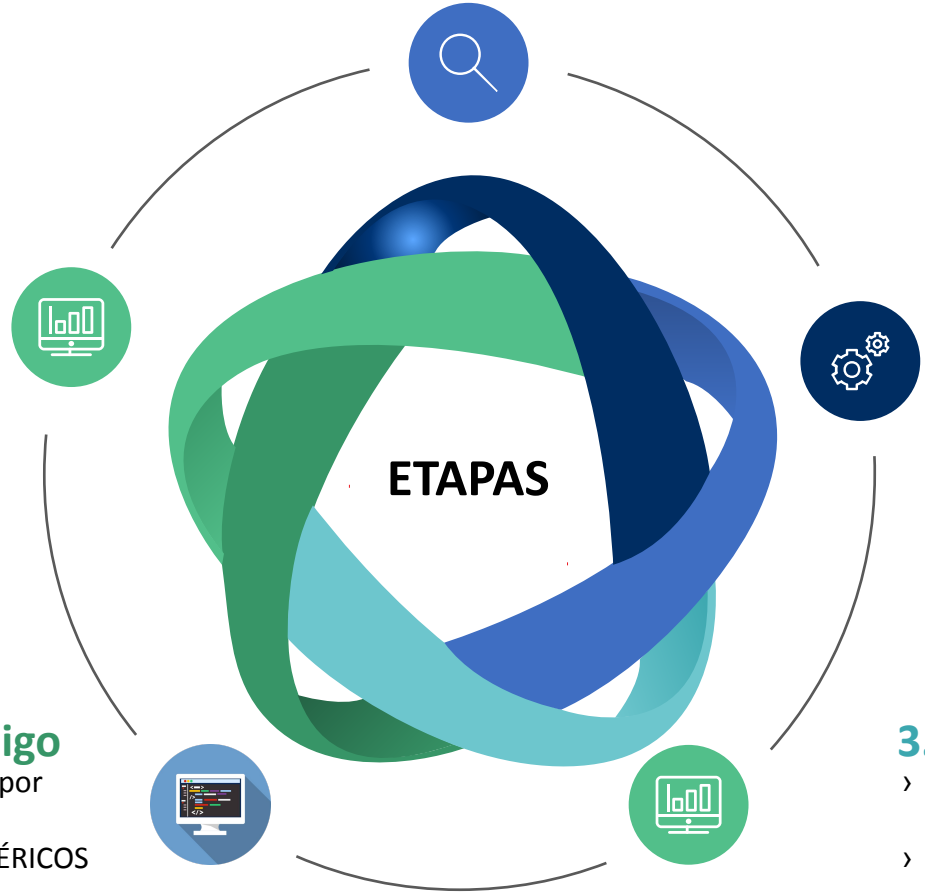
02. Perfil das mulheres nos últimos anos do ensino médio

03. Para onde direcionar recursos para melhorar a educação dessas mulheres?

04. Intensificar a política de “cotas raciais”

1. Proposta de estudo inicial

› *Brainstorm* inicial
› Escolha de *datasets:* dados de matriculas no ensino médio, no ano de 2019, na região nordeste (INEP)



2.Ajuste nos dados

› Importação do *dataset*
› Tratamento de dados inválidos ou inexistentes
› Filtragem de dados: sexo biologico, etapa do ensino médio, RN

3.Análise de dados

› Inicialmente, abordagem com foco nas ESTUDANTES
› Discussões do porquê a ideia inicial não funcionaria
› Novo *brainstorm*
› Mudança na abordagem: foco nos MUNICÍPIOS

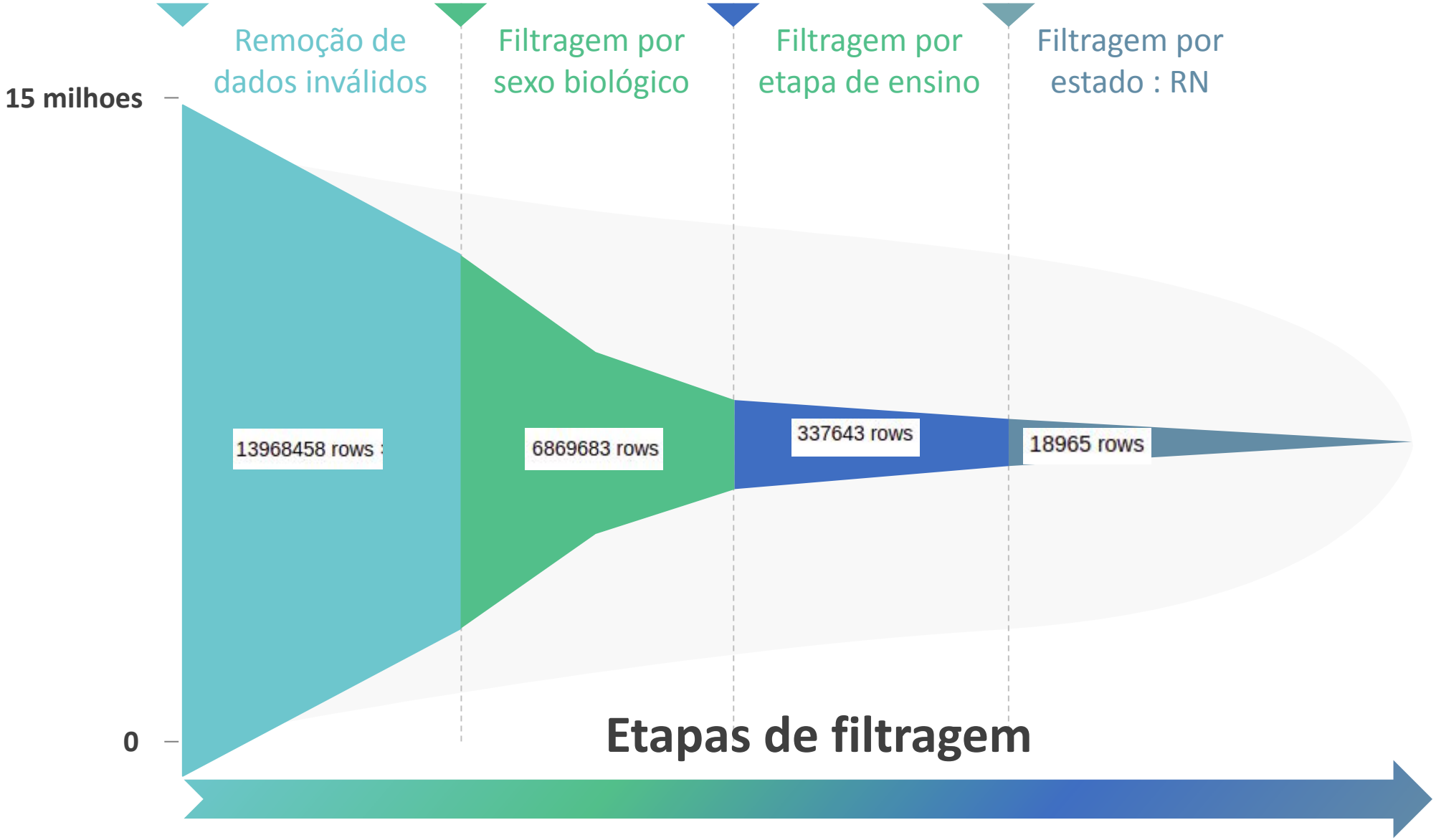
4.Refactoring do código

› Foco na agregação de dados por MUNICÍPIOS
› Ênfase em gerar dados NUMÉRICOS percentuais
› Divisão em duas análises:
    Com 2 características
    Com 3 características (esta vai ser mostrada)

5.Análise final dos dados

› Agrupamento baseado em MUNICÍPIOS - Kmeans
› Analise estatística dos *clusters*
› Discussões sobre perfil das estudantes para cada *cluster*

Ajustes no *dataset*

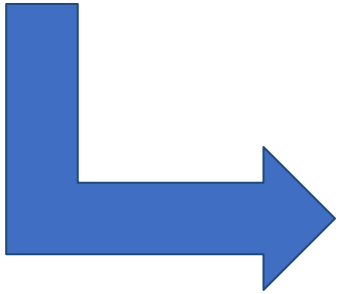


Características selecionadas e dificuldade encontrada

Out[12]:

	TP_COR_RACA	TP_ZONA_RESIDENCIAL	CO_MUNICIPIO	TP_DEPENDENCIA
6448835	0	2	2404408	2
6448883	3	1	2410405	2
6448903	1	1	2408003	2
6448939	1	1	2408102	2
6448940	3	1	2408102	2
...	...	...	...	...
7485738	1	2	2414159	2
7486028	3	1	2409506	2
7486055	3	1	2407104	2
7486086	2	1	2408102	2
7486597	1	1	2408102	1

18965 rows × 4 columns



Out[14]:

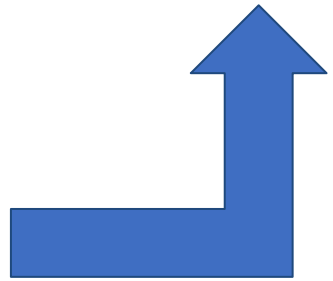
	Código do Município	Nome do Município
0	2400109	Acari
1	2400208	Açu
2	2400307	Afonso Bezerra
3	2400406	Água Nova
4	2400505	Alexandria
...	...	...
162	2414704	Várzea
163	2414753	Venha-Ver
164	2414803	Vera Cruz
165	2414902	Viçosa
166	2415008	Vila Flor

167 rows × 2 columns

Out[19]:

	NOME_MUNICIPIO	COUNT
0	Acari	25
1	Afonso Bezerra	38
2	Alexandria	40
3	Almino Afonso	12
4	Alto do Rodrigues	83
...	...	...
161	Vera Cruz	89
162	Vila Flor	16
163	Viçosa	4
164	Várzea	26
165	Água Nova	31

166 rows × 2 columns



Etapas de processamento dos dados

01

Conversão dos dados “categóricos” para porcentagem

Os dados numéricos que descrevem características categóricas foram processados de forma que fosse possível fazer uma melhor análise estatística dos dados baseada em porcentagem

02

Preparação dos dados para aplicar o *Kmeans*

O *dataframe* é convertido para *array* (formato de input para o algoritmo do *Kmeans*). É realizada também a escolha do número de *clusters*, baseada no “*Elbow Method*”.

03

Aplicação do *Kmeans* e exportação dos dados

Como trabalhamos com todas as dimensões (12), não foi gerada a representação gráfica. Para visualizar o resultado do agrupamento é necessário exportar o arquivo para um csv

04

Análise estatística dos *clusters*

Calculo de média, desvio padrão, máximos e mínimos para cada característica, buscando entender melhor o agrupamento. Elaboração de *boxplots*.

1. Conversão dos dados “categóricos” para porcentagem

Out[25]:

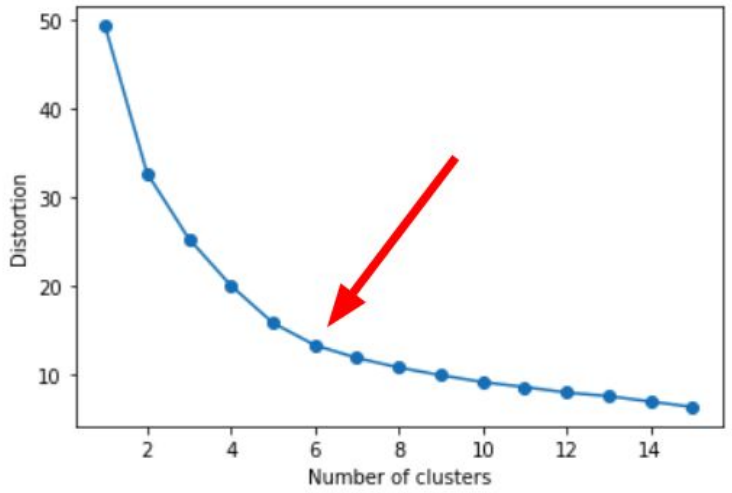
	NOME_MUNICIPIO	COUNT	TP_ZONA_RESIDENCIAL_1	TP_ZONA_RESIDENCIAL_2	TP_DEPENDENCIA_1	TP_DEPENDENCIA_2	TP_DEPENDENCIA_3	TP_DEPENDENCIA_4	TP_COR_RACA_0	TP_COR_RACA_1	TP_COR_RACA_2	TP_COR_RACA_3	TP_COR_RACA_4	TP_COR_RACA_5
0	Acari	25	0.920000	0.080000	0.0	1.0	0.0	0.0	0.800000	0.080000	0.000000	0.120000	0.0	0.0
1	Afonso Bezerra	38	0.526316	0.473684	0.0	1.0	0.0	0.0	0.000000	0.289474	0.026316	0.684211	0.0	0.0
2	Alexandria	40	0.675000	0.325000	0.0	1.0	0.0	0.0	0.575000	0.175000	0.000000	0.250000	0.0	0.0
3	Almino Afonso	12	0.916667	0.083333	0.0	1.0	0.0	0.0	0.000000	0.750000	0.000000	0.250000	0.0	0.0
4	Alto do Rodrigues	83	0.638554	0.361446	0.0	1.0	0.0	0.0	0.385542	0.180723	0.012048	0.421687	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
161	Vera Cruz	89	0.775281	0.224719	0.0	1.0	0.0	0.0	0.258427	0.213483	0.033708	0.494382	0.0	0.0
162	Vila Flor	16	0.812500	0.187500	0.0	1.0	0.0	0.0	0.562500	0.250000	0.000000	0.187500	0.0	0.0
163	Viçosa	4	1.000000	0.000000	0.0	1.0	0.0	0.0	0.000000	0.250000	0.000000	0.750000	0.0	0.0
164	Várzea	26	0.884615	0.115385	0.0	1.0	0.0	0.0	0.153846	0.153846	0.000000	0.692308	0.0	0.0
165	Água Nova	31	0.709677	0.290323	0.0	1.0	0.0	0.0	0.677419	0.225806	0.000000	0.096774	0.0	0.0

2. Preparação dos dados para aplicar *Kmeans*

In [69]:
```
np_zona_dep_cor = df_zona_dep_cor.to_numpy(dtype=np.float32)
print(np_zona_dep_cor)
```

In [70]:
```
# calculate distortion for a range of number of cluster
distortions = []
N_C = 16
for i in range(1, N_C):
    km = KMeans(
        n_clusters=i, init='random',
        n_init=20, max_iter=2000,
        tol=1e-04, random_state=0
    )
    km.fit(np_zona_dep_cor)
    distortions.append(km.inertia_)

# plot
plt.plot(range(1, N_C), distortions, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Distortion')
plt.show()
```



3.Aplicação do *Kmeans* e exportação dos dados

In [71]:
```
km_zona_dep_cor = KMeans(
    n_clusters=6, init='random',
    n_init=20, max_iter=2000,
    tol=1e-04, random_state=0
)

zona_dep_cor_clusters = km_zona_dep_cor.fit_predict(np_zona_dep_cor)
zona_dep_cor_clusters_list = list(zona_dep_cor_clusters)
```

Out[93]:

	NOME_MUNICIPIO	CLUSTER	COUNT	TP_ZONA_RESIDENCIAL_1	TP_ZONA_RESIDENCIAL_2	TP_DEPENDENCIA_1	TP_DEPENDENCIA_2	T
155	Timbaúba dos Batistas	0	5	0.800000	0.200000	0.000000	1.000000	
69	Lagoa de Velhos	0	7	0.714286	0.285714	0.000000	1.000000	
104	Pilões	0	13	0.769231	0.230769	0.000000	1.000000	
38	Fernando Pedroza	0	14	0.857143	0.142857	0.000000	1.000000	
162	Vila Flor	0	16	0.812500	0.187500	0.000000	1.000000	
...	...	...	...	...	...	...	...	
32	Doutor Severiano	5	53	0.396226	0.603774	0.000000	1.000000	
46	Guamaré	5	74	0.418919	0.581081	0.000000	1.000000	
65	Lagoa Nova	5	92	0.510870	0.489130	0.000000	1.000000	
23	Caraúbas	5	95	0.547368	0.452632	0.000000	1.000000	
26	Ceará-Mirim	5	573	0.485166	0.514834	0.193717	0.790576	

166 rows × 15 columns

In [94]:
```
df_zona_dep_cor2.to_csv("/home/thaty/cien_dadosthati/ApresentacaoFinal/data_clusters2.csv")
```

4. Análise estatística dos *clusters*

```
In [77]: df_preview_analysis2 = df_zona_dep_cor2.copy().drop(["NOME_MUNICIPIO"], axis=1)
         df_analysis2 = df_preview_analysis2.groupby(["CLUSTER"]).agg(["mean", "std", "max", "min"])
```

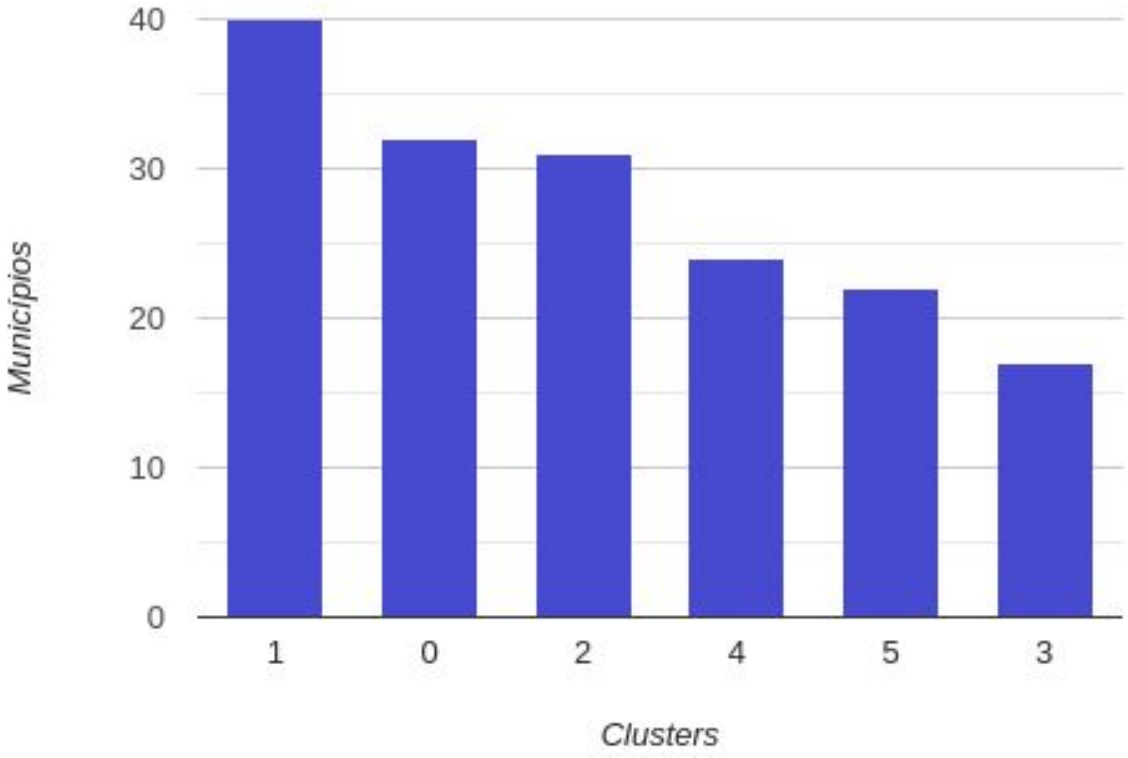
Out[79]:

	TP_ZONA_RESIDENCIAL_1				TP_ZONA_RESIDENCIAL_2				TP_DEPENDENCIA_1		...	TP_COR_RACA_3		
	mean	std	max	min	mean	std	max	min	mean	std	...	max	min	mean
CLUSTER														
0	0.771886	0.097582	1.000000	0.621622	0.228114	0.097582	0.378378	0.000000	0.000000	0.000000	...	0.428571	0.000000	0.001182
1	0.831354	0.118528	1.000000	0.631579	0.168646	0.118528	0.368421	0.000000	0.003926	0.024828	...	0.906977	0.434783	0.004812
2	0.485849	0.121243	0.628571	0.208955	0.514151	0.121243	0.791045	0.371429	0.000000	0.000000	...	0.400000	0.058824	0.000424
3	0.859679	0.095466	0.991051	0.639344	0.140321	0.095466	0.360656	0.008949	0.405251	0.178319	...	0.634615	0.246753	0.004004
4	0.789084	0.148560	1.000000	0.500000	0.210916	0.148560	0.500000	0.000000	0.000000	0.000000	...	0.414634	0.000000	0.000000
5	0.413839	0.092858	0.547368	0.204545	0.586161	0.092858	0.795455	0.452632	0.008805	0.041301	...	0.857143	0.301887	0.002307

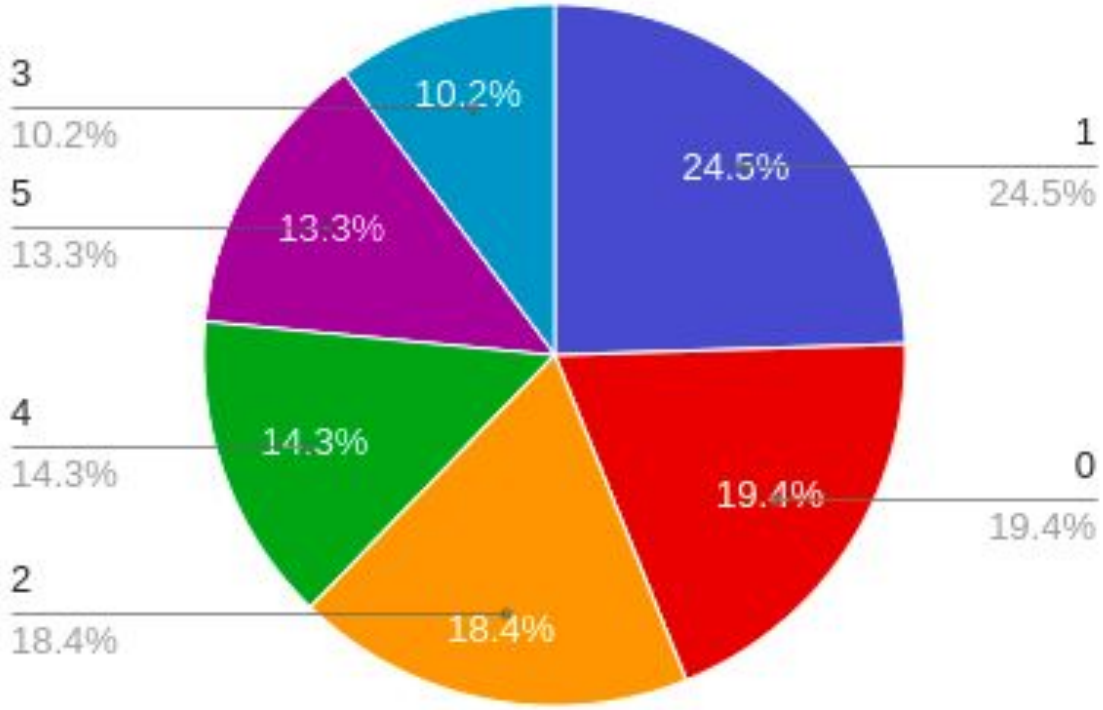
6 rows × 48 columns

Resultados e discussões

Clusters



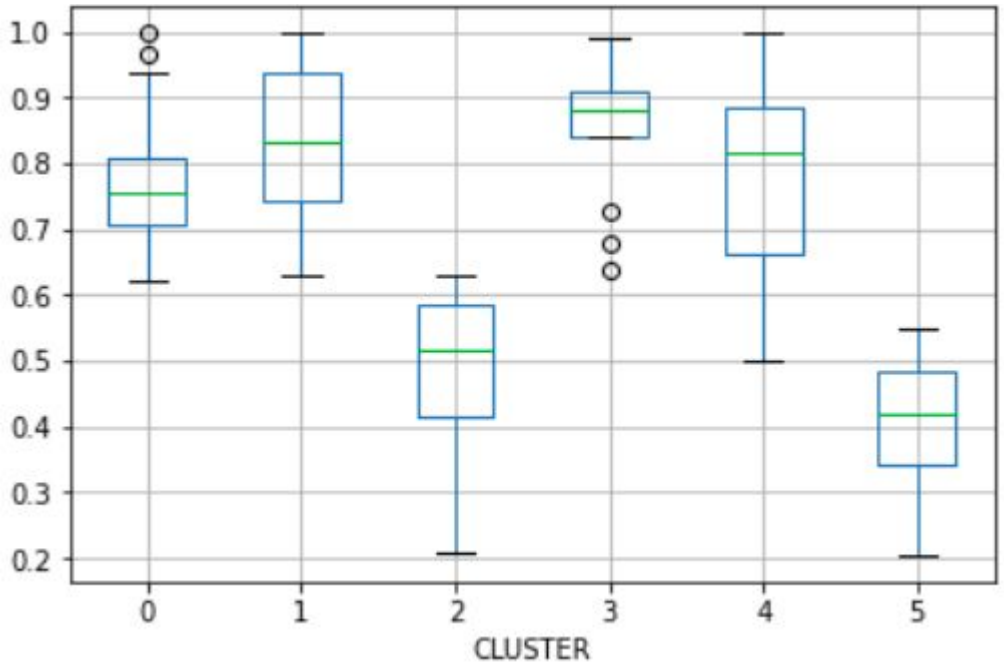
Total de municipios : 166



Análise estatística

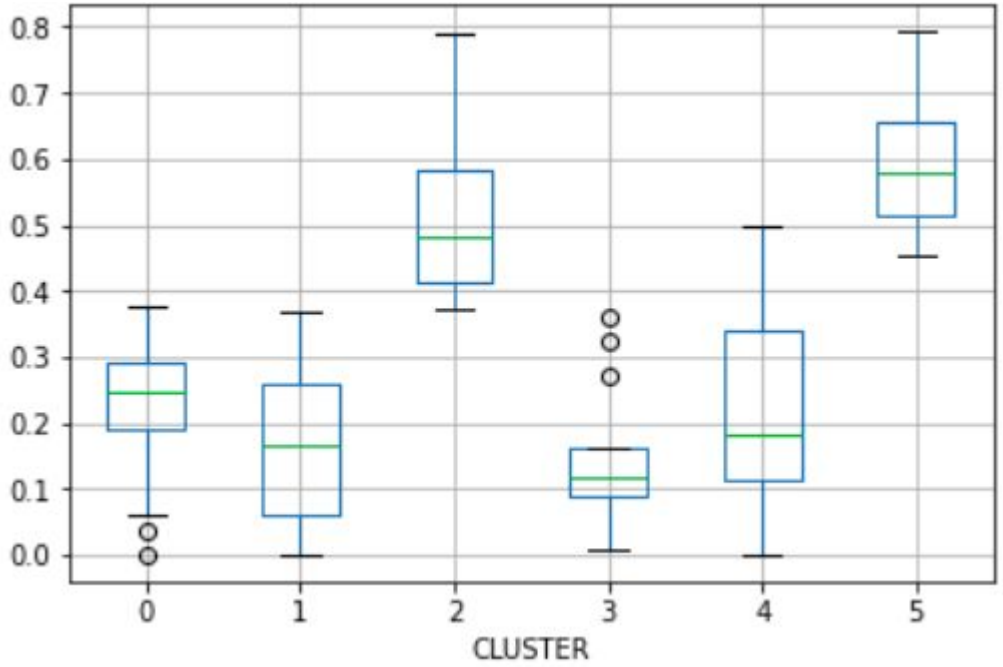
- Zona residencial

Zona urbana



Clusters mais similares: 1 e 4 (mediana similar, porém valores mínimos são distintos)
Cluster mais diferente: 5
Clusters com outliers p/ valores máximos: 0
Clusters com outliers p/ valores mínimos: 3

Zona rural

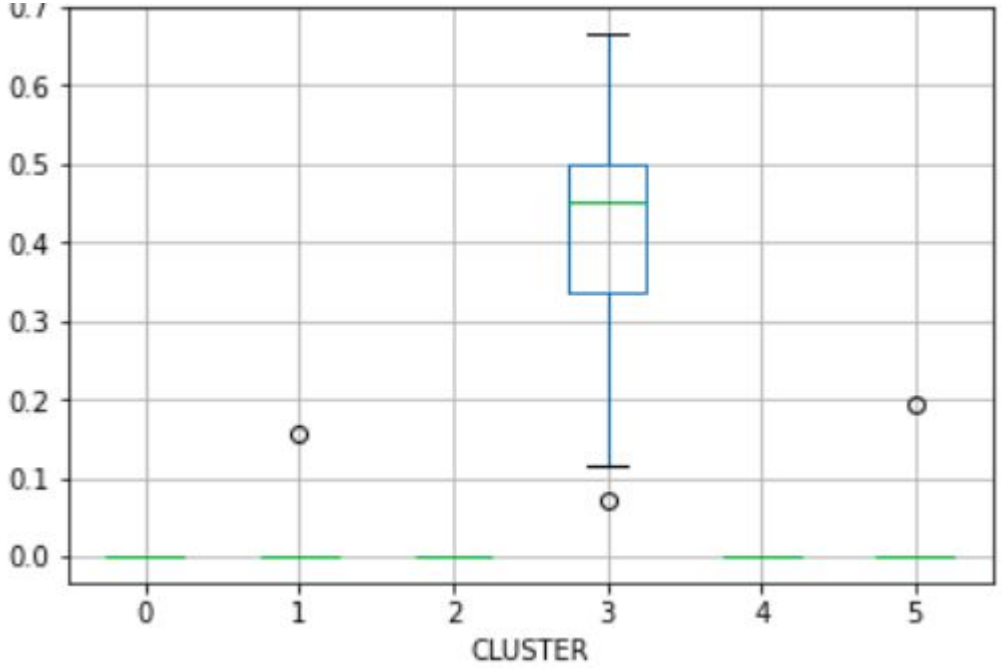


Clusters mais similares: 1 e 4 (mediana similar, porém valores máximos são distintos)
Cluster mais diferente: 5
Clusters com outliers p/ valores máximos: 3
Clusters com outliers p/ valores mínimos: 0

Análise estatística

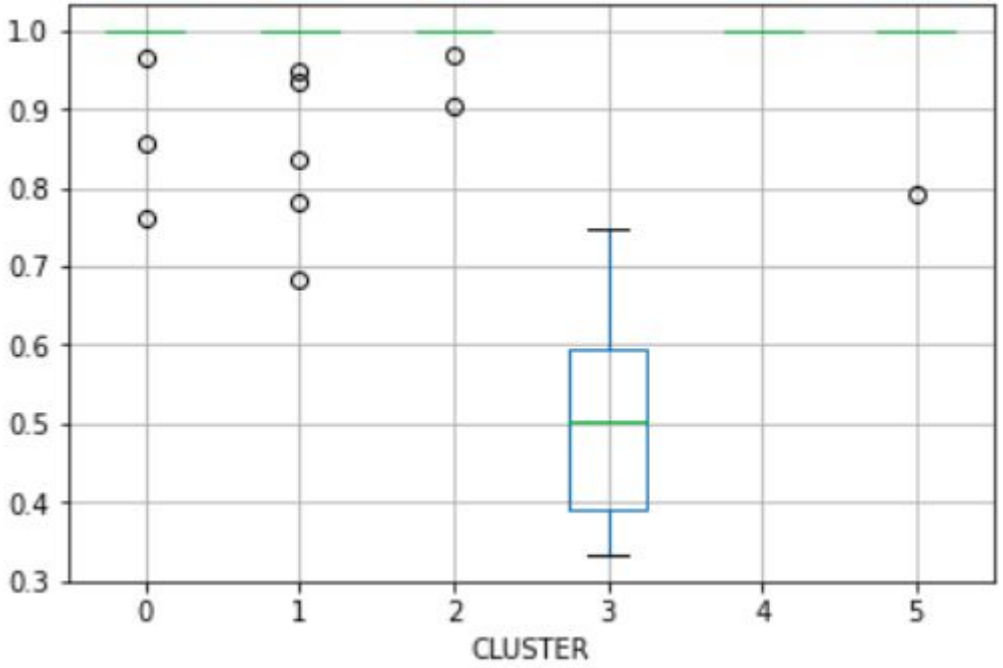
- Dependência da escola

Federal



Somente o *cluster* 3 apresenta maior distribuição com relação ao tipo de escola (onde está cidades como Natal, Mossoró e Parnamirim). Os clusters 1 e 5 tem alguns *outliers* com escolas federais, que são as cidades de : Macaíba e Ceará Mirim

Estadual

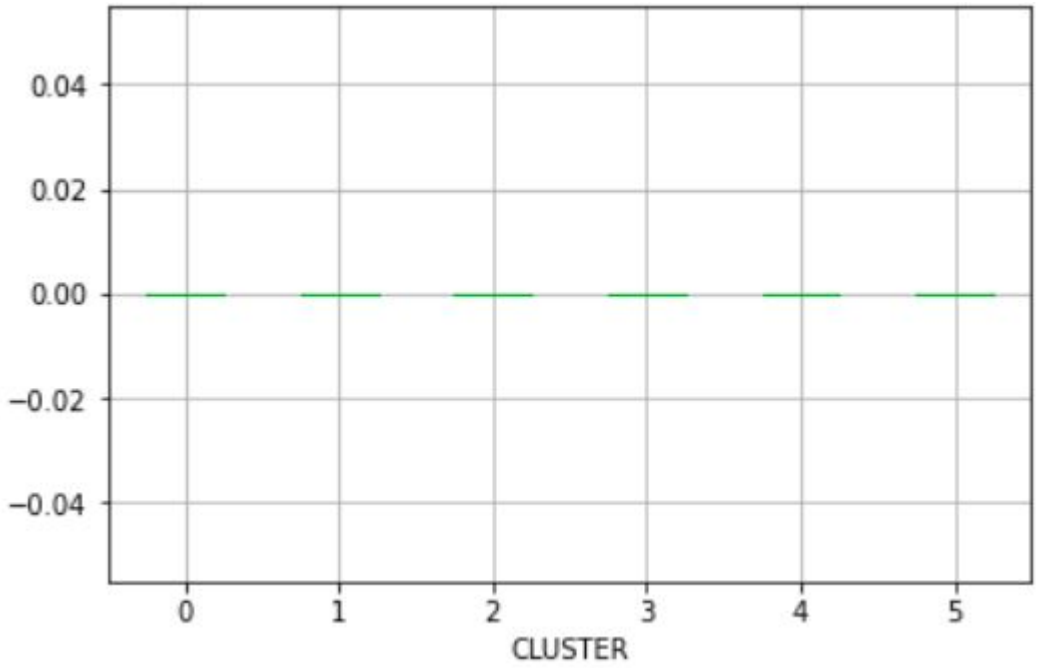


O *cluster* 4 é composto totalmente por escolas estaduais, por isso não há uma distribuição no boxplot.
Ja para os *clusters* 0, 1, 2 e 5, há cidades nas quais há escolas diferentes de estaduais, porém não muitas. (ex 5: somente Ceará Mirim)
Somente o *cluster* 3 apresenta uma maior distribuição de escolas.

Análise estatística

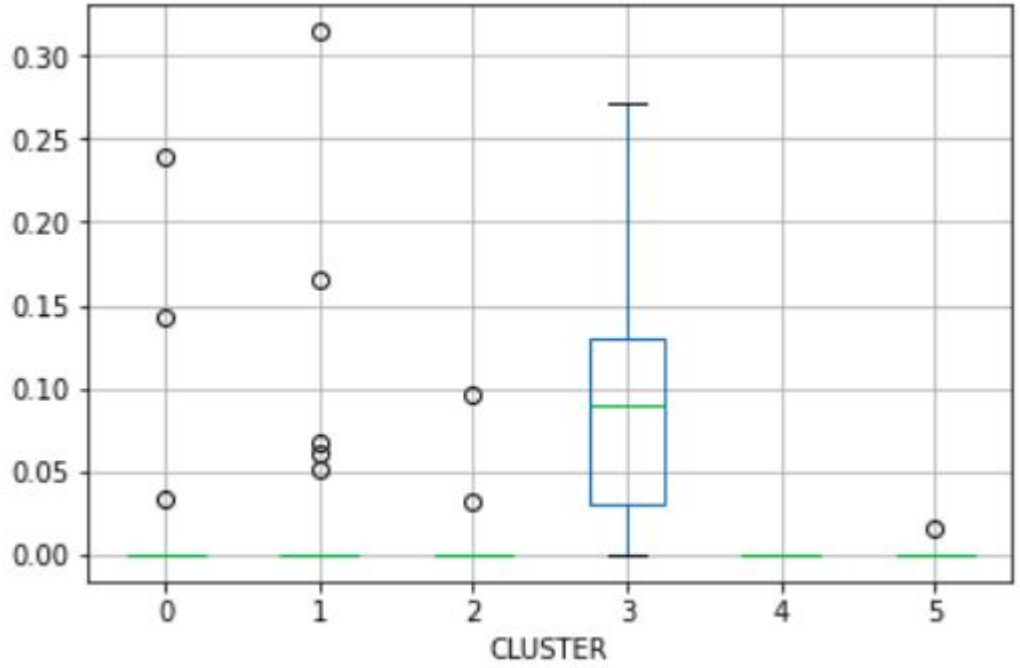
- Dependência da escola

Municipal



Nenhuma amostra do nosso *database* apresentou escolas municipais.

Privada

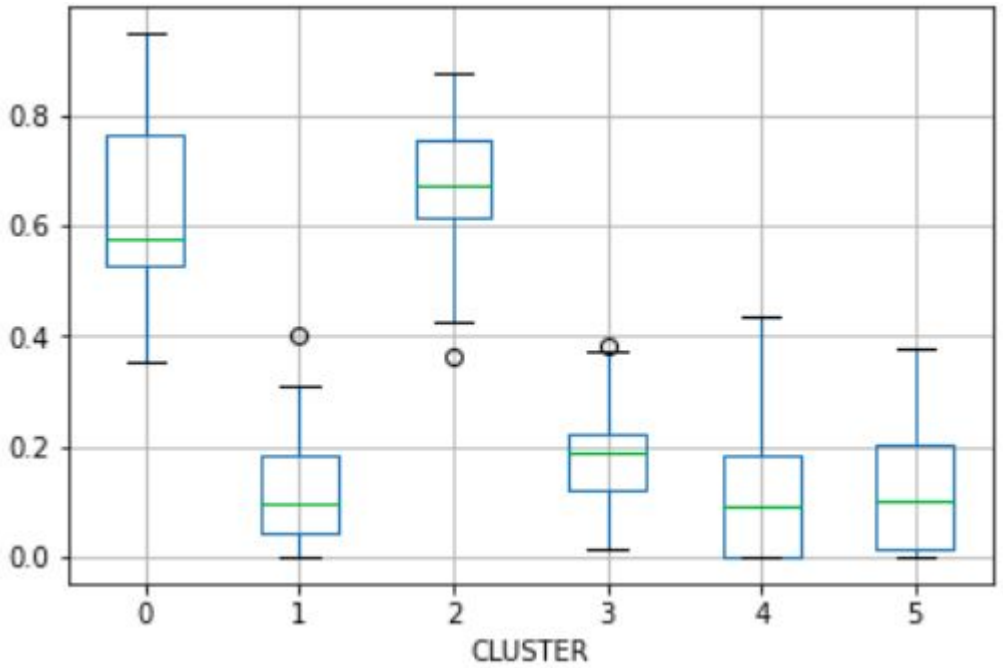


Percebe-se a tendência de distribuição de tipo de escolas somente no *cluster* 3 (onde está cidades como Natal, Mossoró e Parnamirim). No *cluster* 4 não há escolas privadas e nos outros há poucas (ex 5: também em Ceará Mirim, 2: somente Santo Antônio e Goianinha)

Análise estatística

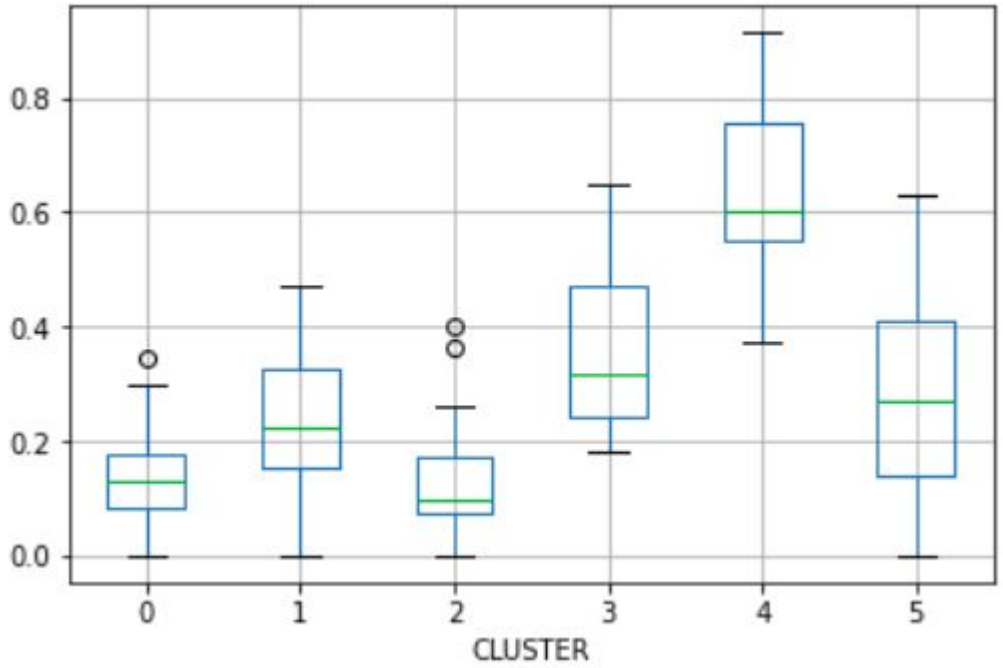
- Etnia

Não declarada



Clusters mais similares: 1, 4 e 5
Cluster com maior dificuldade de fazer a declaração da etnia: 0 e 2
Clusters com outliers p/ valores máximos: 1 e 3
Clusters com outliers p/ valores mínimos: 2

Branca

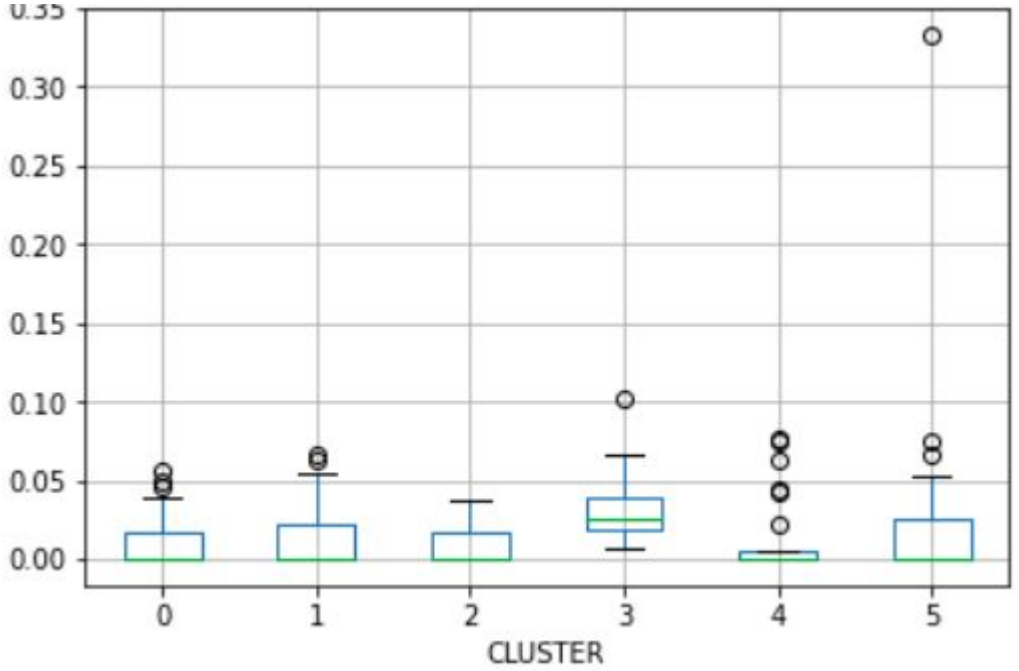


Cluster onde a maioria se declara branca: 4
Cluster onde a minoria se declara branca: 2 e 0
Clusters com *outliers* p/ valores máximos: 0 e 2 (Ex: 0 na cidade de Jardim de Piranhas, 2: em Antonio Martins e Olho D'agua do Borges)

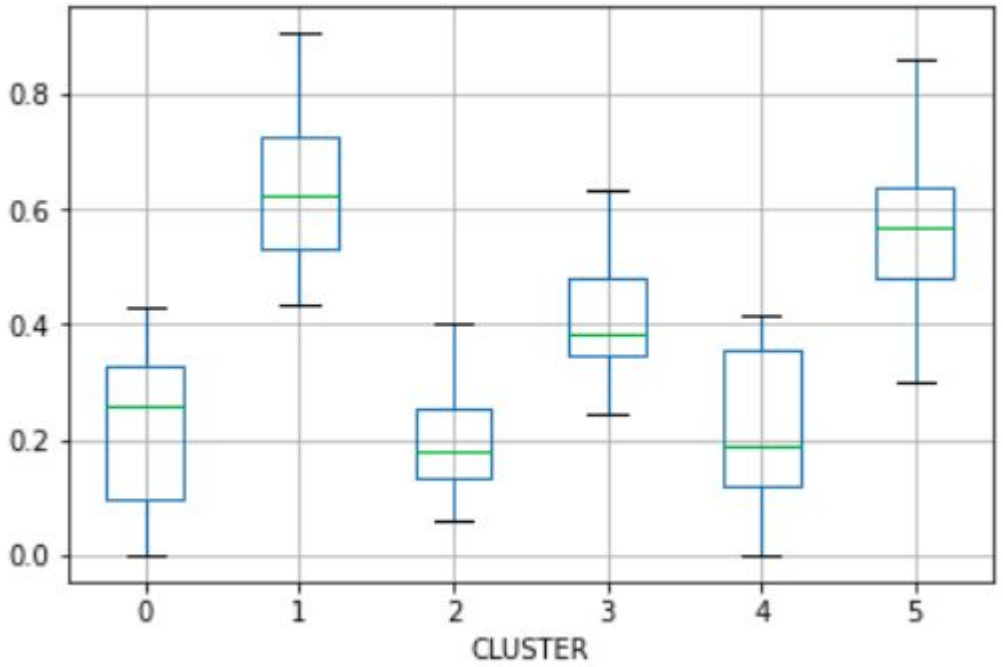
Análise estatística

- Etnia

Negra



Parda



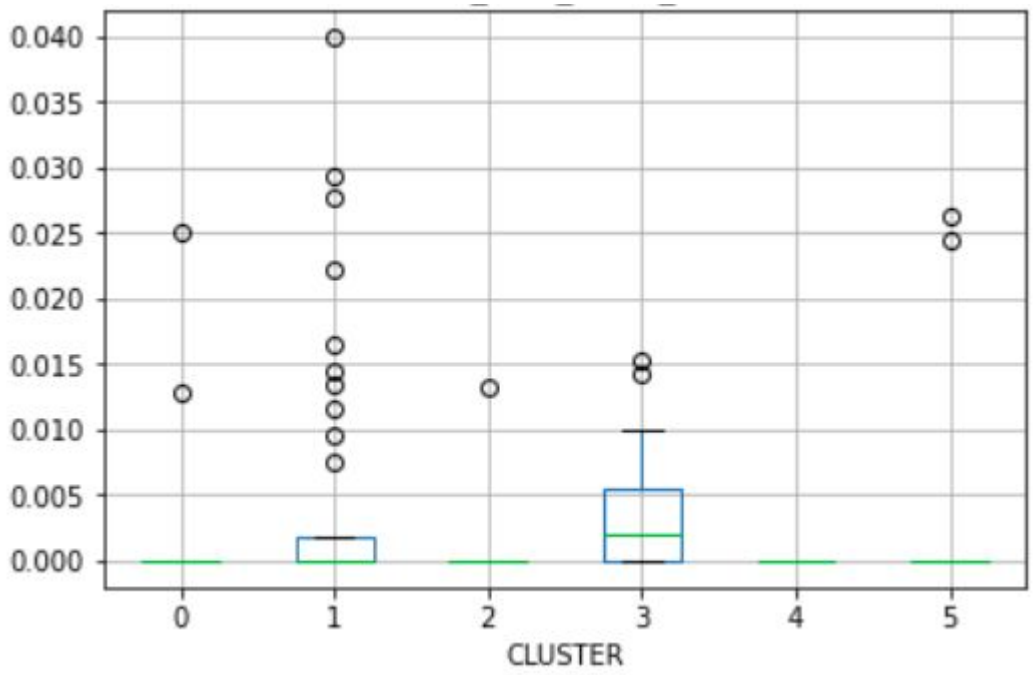
De modo geral, há uma baixíssima representatividade de etnia negra no estado no cenário dos últimos anos do ensino médio entre as mulheres.
Há algumas cidades que há *outliers* em todos os *clusters*, exceto no 2. (ex: 5 na cidade de Bodó)

Bastante estudantes se identificam com a etnia Parda no estado.
Clusters com maior representatividade Parda: 1 e 5
A distribuição é mais uniforme em todos os *clusters*, não aparecem *outliers.*

Análise estatística

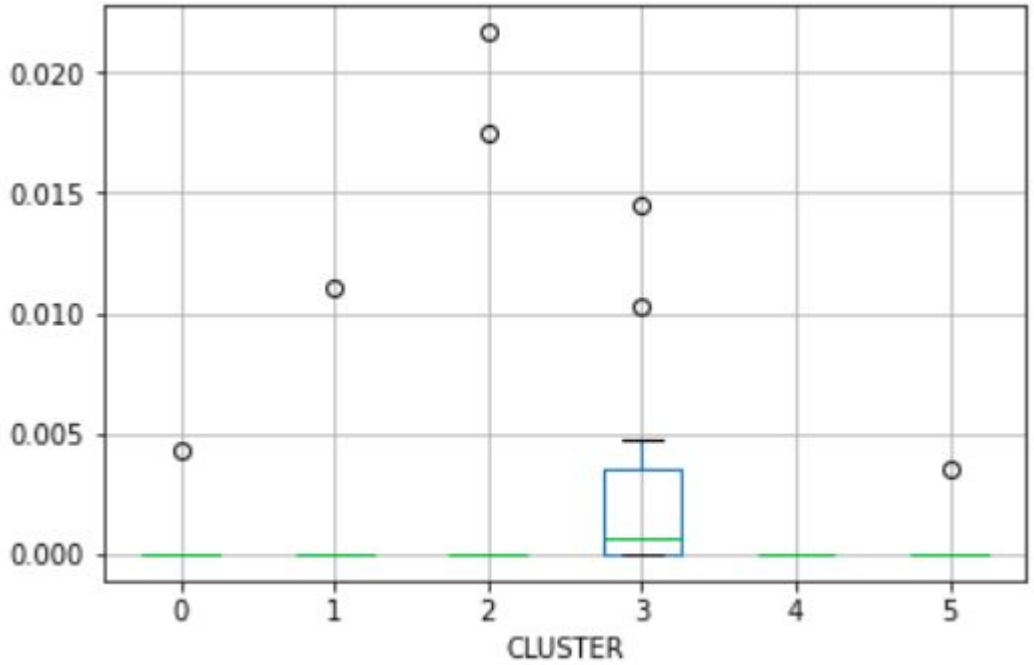
- Etnia

Amarela



De modo geral, há uma baixíssima representatividade de etnia amarela.
No cluster 4 não há nenhuma cidade onde alguém se declarou como etnia amarela.
Há algumas cidades que são *outliers* em todos os *clusters* restantes, mas mesmo assim com baixo valor. (ex: 1 na cidade de Lagoa d'anta)

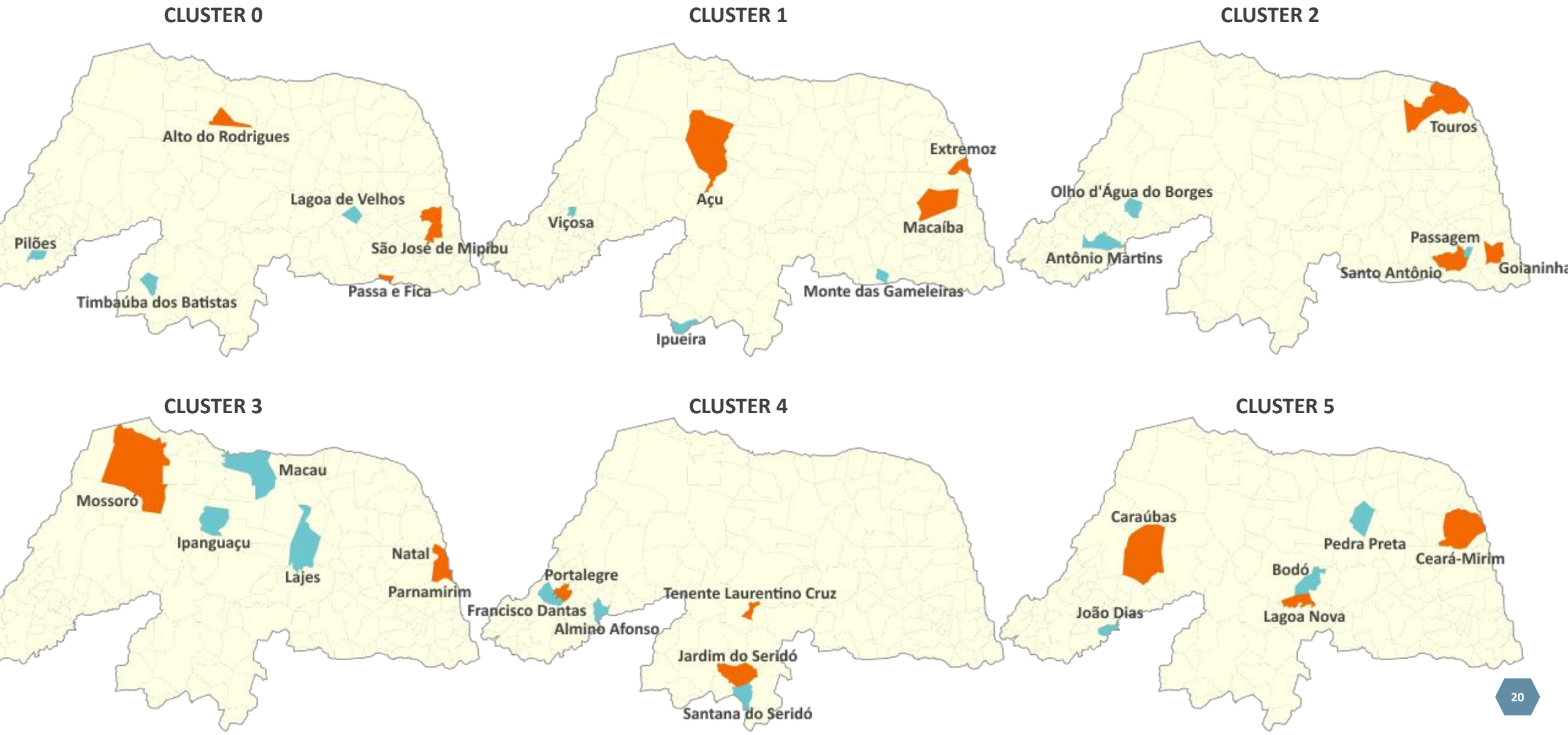
Indígena



De modo geral, há uma baixíssima representatividade de etnia indígena.
No cluster 4 não há nenhuma cidade onde alguém se declarou como etnia indigena.
Há algumas cidades que são *outliers* em todos os *clusters* restantes, mas mesmo assim com baixo valor. (ex: 2 na cidade de Sao Tome e São Miguel do Gostoso)

Entendendo melhor os *clusters*

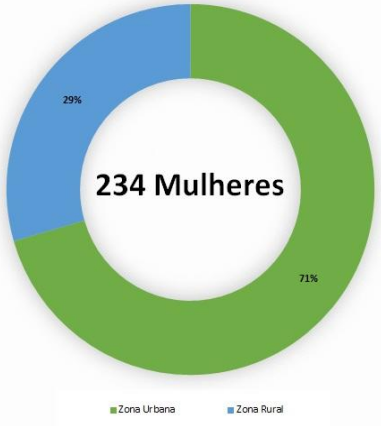
- Cidades do cluster com a maior quantidade de estudantes mulheres
- Cidades do cluster com a menor quantidade de estudantes mulheres



Entendendo melhor os *clusters (cidades com maior quantidade)*

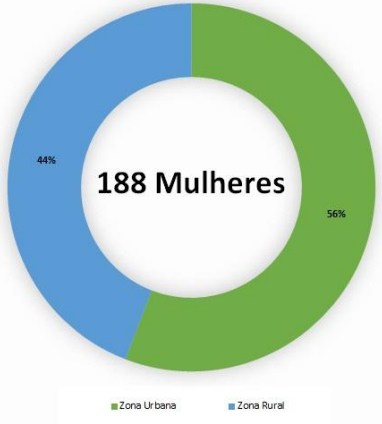
- Zona Residencial

São José do Mipibu



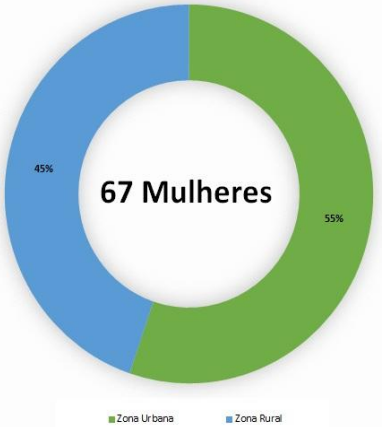
Cluster 0

Goianinha



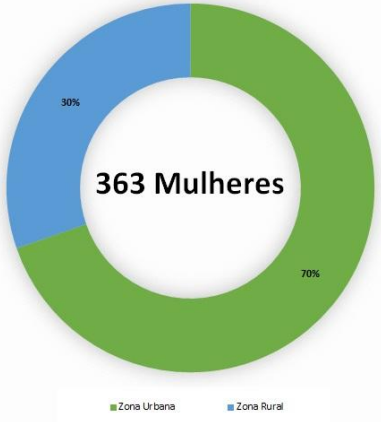
Cluster 2

Tenente Laurentino Cruz



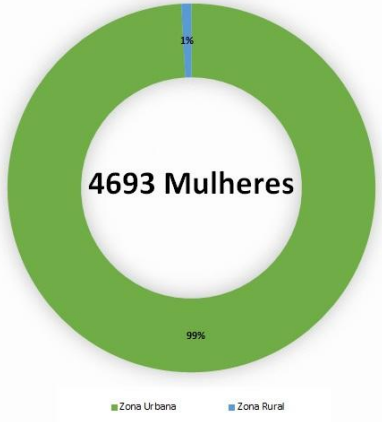
Cluster 4

Macaíba



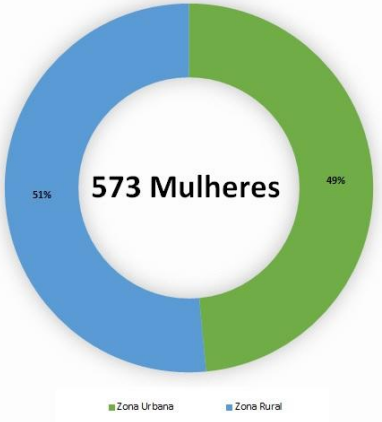
Cluster 1

Natal



Cluster 3

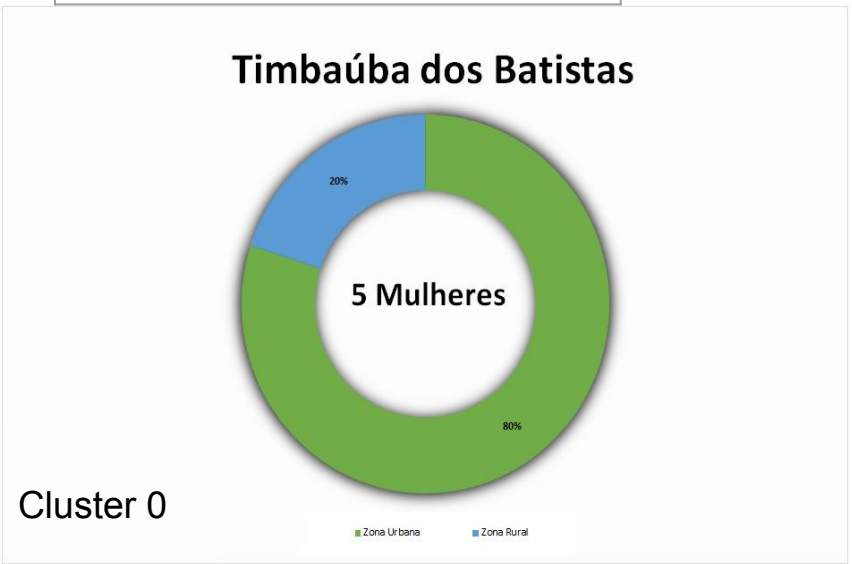
Ceará-Mirim



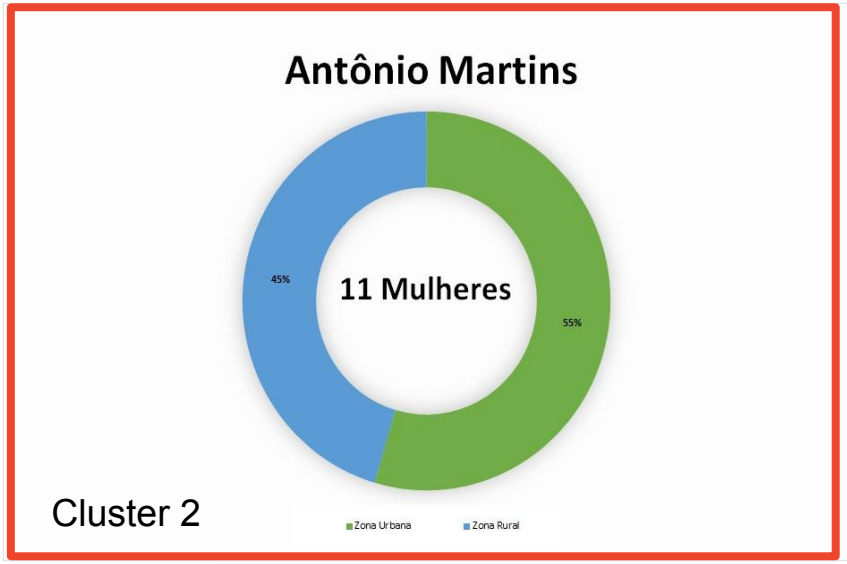
Cluster 5

Entendendo melhor os *clusters (cidades com menor quantidade)*

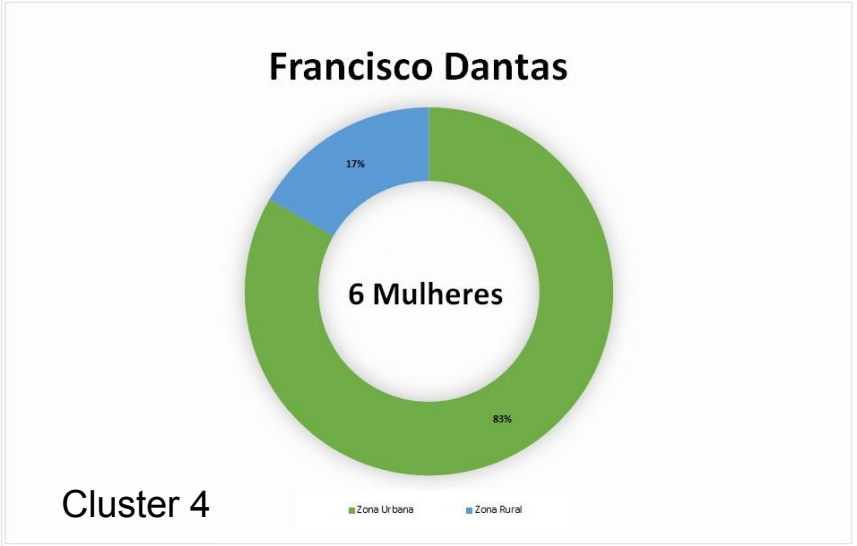
- Zona Residencial



Cluster 0

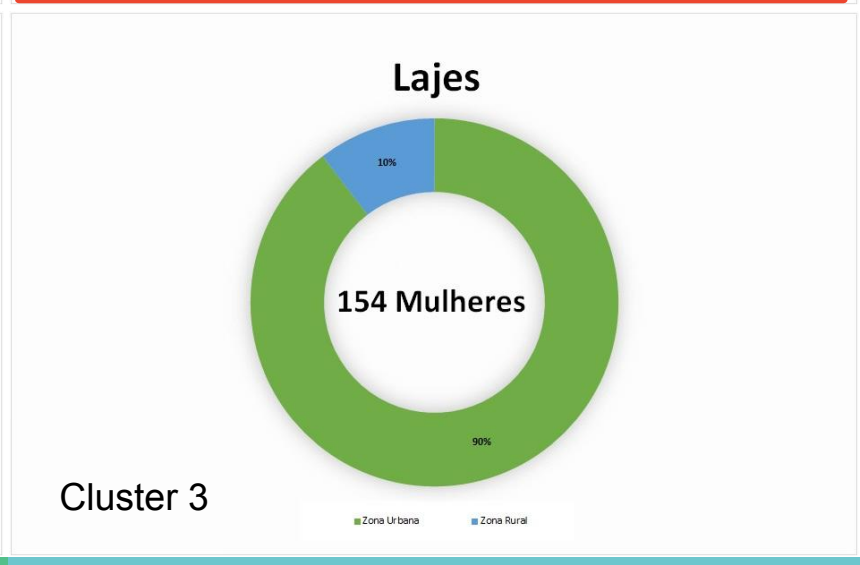


Cluster 2



Cluster 4

Cluster 1



Cluster 3

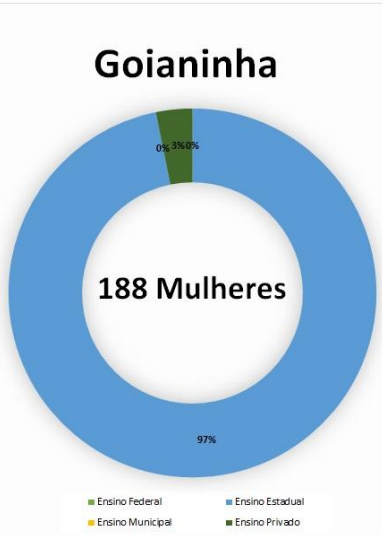
Cluster 5

Entendendo melhor os *clusters (cidades com maior quantidade)*

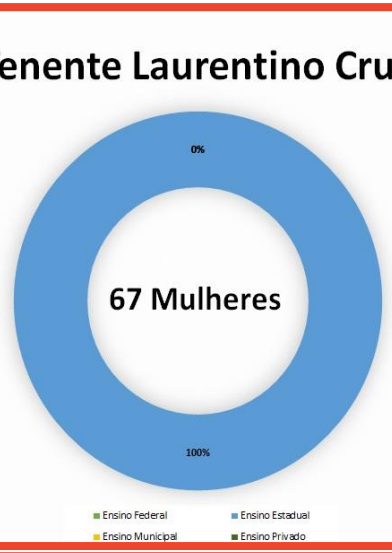
- Dependência da escola

Cluster 0

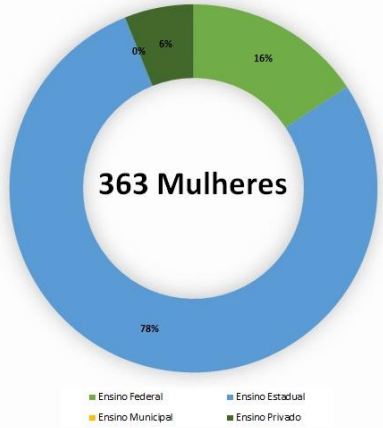
Cluster 2



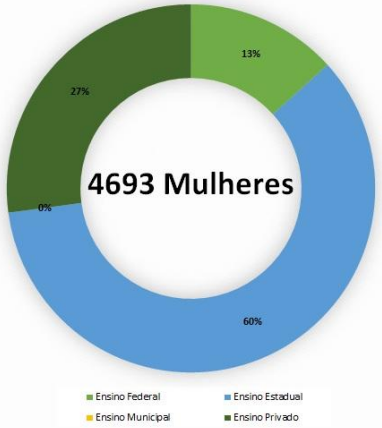
Cluster 4



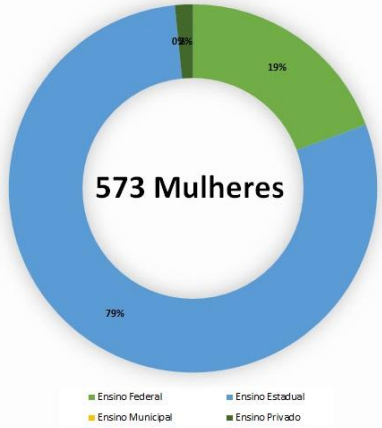
Cluster 1



Cluster 3



Cluster 5



Entendendo melhor os *clusters (cidades com menor quantidade)*

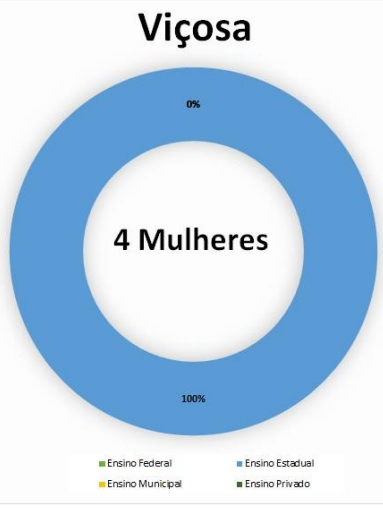
- Dependência da escola

Cluster 0

Cluster 2

Cluster 4

Cluster 1



Cluster 3

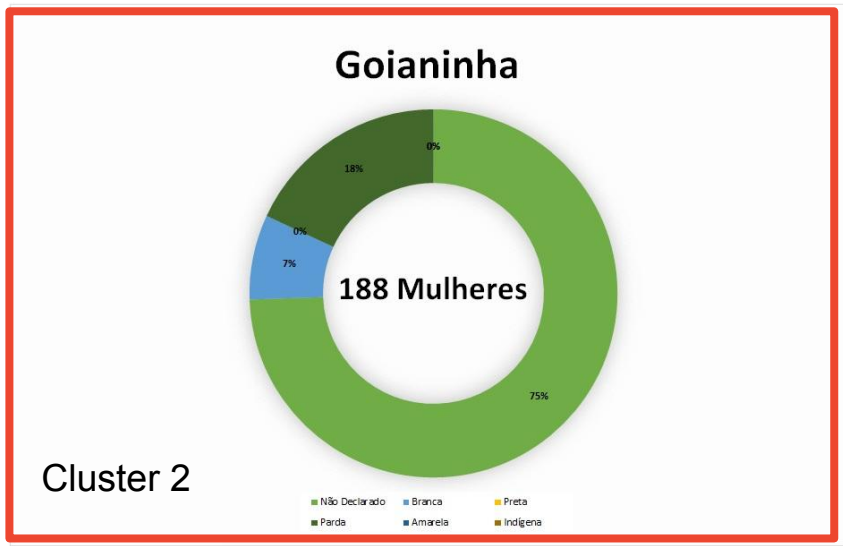
Cluster 5

Entendendo melhor os *clusters (cidades com maior quantidade)*

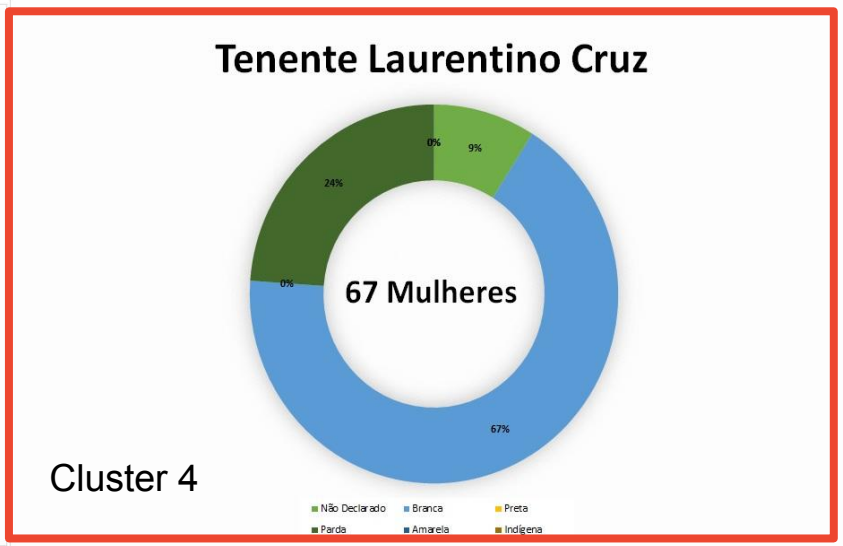
- Etnia

Cluster 0

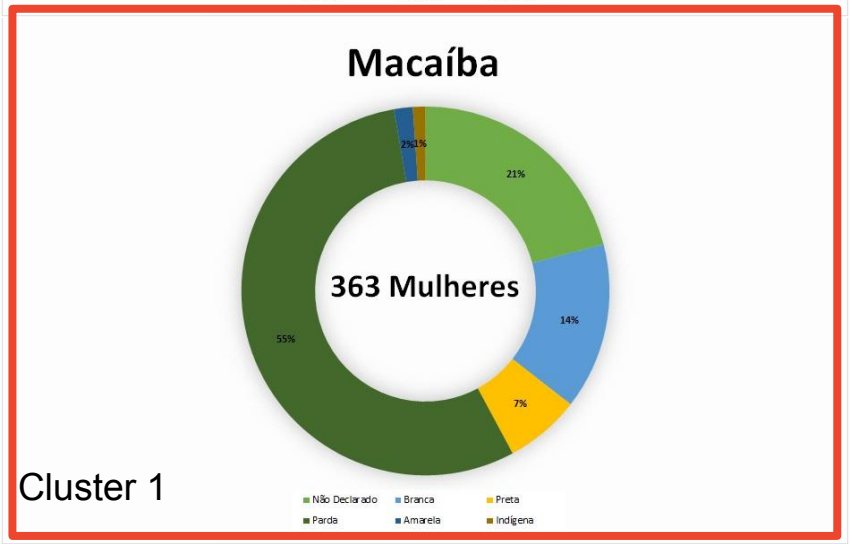
Cluster 2



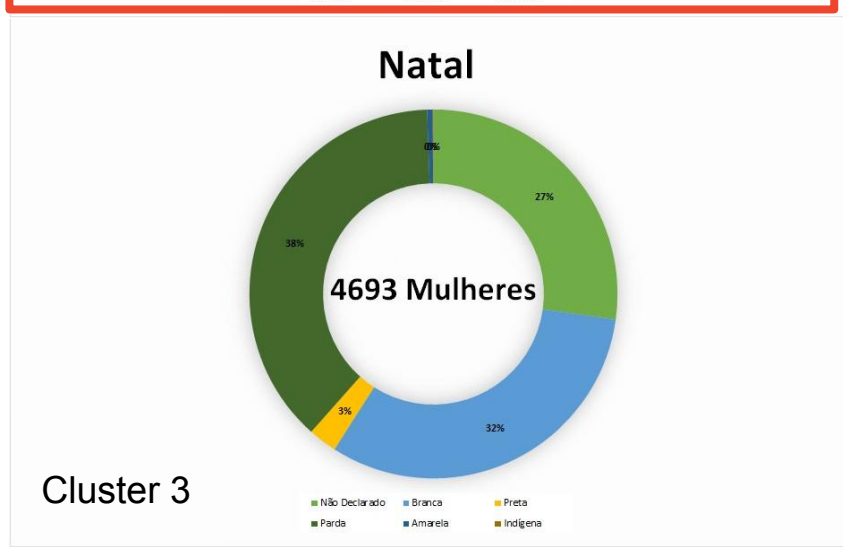
Cluster 4



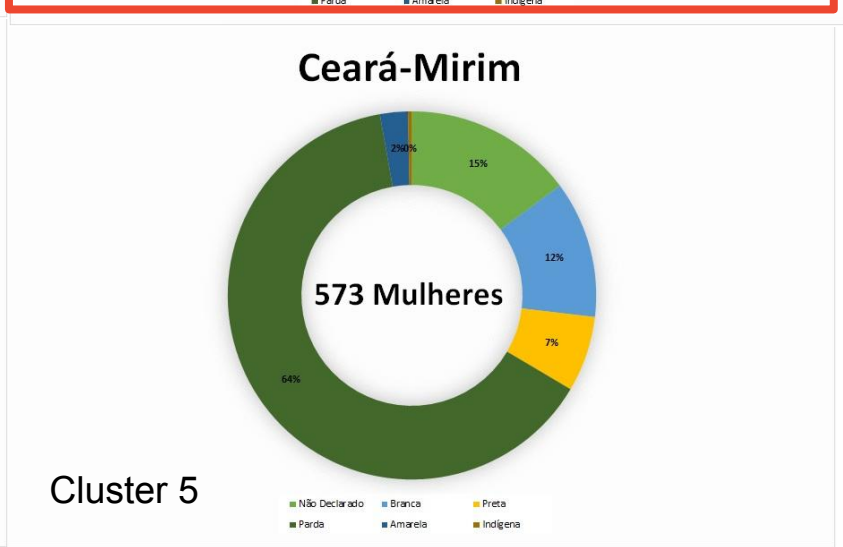
Cluster 1



Cluster 3



Cluster 5



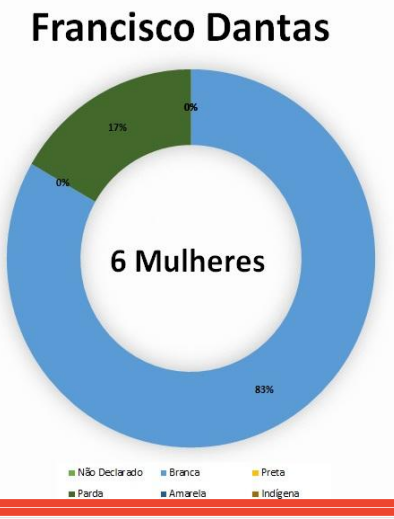
Entendendo melhor os *clusters (cidades com menor quantidade)*

- Etnia

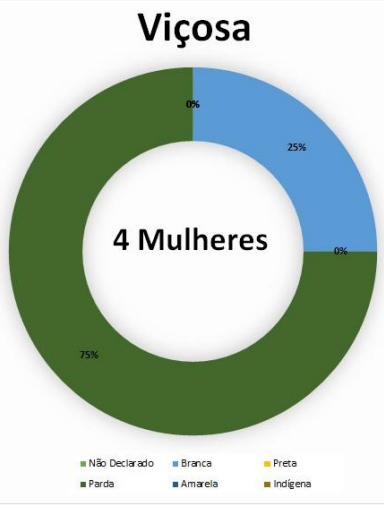
Cluster 0

Cluster 2

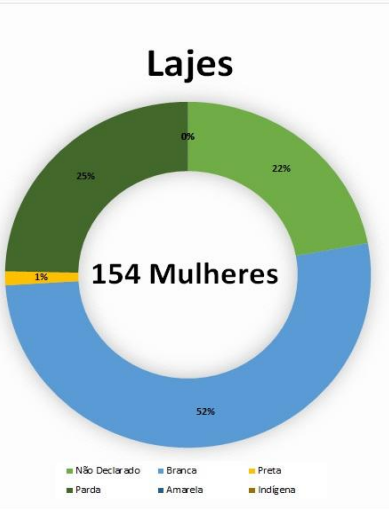
Cluster 4



Cluster 1



Cluster 3



Cluster 5

Entendendo melhor os *clusters*

PIB (em milhões) das 3 maiores cidades de cada cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1ª Cidade	São José de Mipibu (815 mi)	Macaíba (1.679 mi)	Goianinha (410 mi)	Natal (24.856 mi)	Tenente Laurentino Cruz (131 mi)	Ceará-Mirim (899 mi)
2ª Cidade	Alto do Rodrigues (488 mi)	Açu (1.242 mi)	Touros (680 mi)	Mossoró (6.926 mi)	Portalegre (64 mi)	Caraúbas (377 mi)
3ª Cidade	Passa e Fica (123 mi)	Extremoz (452 mi)	Santo Antônio (239 mi)	Parnamirim (5.595 mi)	Jardim do Seridó (178 mi)	Lagoa Nova (324 mi)

PIB (em milhões) da menor cidade de cada cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cidade	Timbaúba dos Batistas (35 mi)	Viçosa (19 mi)	Antônio Martins (58 mi)	Lajes (119 mi)	Francisco Dantas (29 mi)	João Dias (24 mi)

Conclusões

01. Presença feminina vs masculina nos últimos anos do ensino médio

Para o estado do RN: mulheres nos ultímos anos correspondem a 54,21 % e homens 45,79 %

02. Perfil das mulheres nos últimos anos do ensino médio

A maioria se declara branca? negra? indigena? Parda ou na realidade nem se declara
Moram em zona rural? urbana? Urbana
Frequentam escolas federais? estaduais? municipais? privadas? Estaduais

03. Para onde direcionar recursos para melhorar a educação dessas mulheres?

Levar institutos federais à regiões mais “interiorizadas”
Incentivos fiscais para implantação de escolas privadas também nessas regiões

04. Intensificar a política de “cotas raciais”

Adicionar mais vagas por cotas em institutos federais?
Incentivos fiscais para implantação de cotas também em escolas privadas?

Lei de cotas no IFRN

***Segundo o último censo IBGE, 5.24% da população do RN se declara preta, 52.48% parda e 0.08% indígena, totalizando 57.8% de PPI no estado.**

Passos futuros sugeridos

- Traçar também um perfil dos estudantes do sexo masculino e fazer uma comparação com as do sexo feminino que foi o objeto deste estudo
- Analisar como esse perfil que foi identificado neste estudo tem se comportado ao longo dos anos, utilizando *datasets* de anos anteriores ou posteriores ao aqui utilizado
- Utilizando-se um *dataset* a partir de 2020, seria possível analisar se a pandemia afetou o perfil que foi analisado por este estudo
- Traçar paralelos entre este estudo, e algum estudo posterior que conseguisse analisar a progressão dessas estudantes, como por exemplo, utilizando um *dataset* relativo aos dados do SISU, e também o ingresso nas universidades

GitHub https://github.com/thatianajessica/CienciaDeDados

Dados gráficos


Obrigado