Trab Multi

February 4, 2021

1 Análise Multivariada e Aprendizado Não-Supervisionado

por Sidnei Gazola Junior, n°USP: 9378888

Professora orientadora: Cibele Russo

ICMC USP São Carlos.

Este trabalho visa responder as seguintes questões propostas:

Considere dados demográficos e econômicos de estados brasileiros, disponíveis, por exemplo, em https://www.ibge.gov.br/cidades-e-estados. Considerando dados (multivariados) que achar relevantes, desenvolva cada um dos itens abaixo, usando a linguagem de sua preferência.

1. Desenvolva as análises descritivas e exploratórias para os dados em questão. Interprete brevemente os resultados
2. Obtenha agrupamentos hierárquicos aglomerativos usando o dendrograma e justifique escolha do número de grupos. Utilize a distância e método de ligação que achar conveniente. Interprete brevemente os resultados
3. Obtenha agrupamentos não-hierárquicos utilizando o algoritmo de K-médias. Interprete brevemente os resultados.

[5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import AgglomerativeClustering, KMeans
from scipy.cluster import hierarchy
from pandasql import sqldf
```

1.1 Análises descritivas e exploratórias

[8]:
```
data = pd.read_csv("dadoses.csv", index_col=0,decimal= ",")
data.head()
```

[8]:
```
          Área Territorial - km² [2019]  População estimada - pessoas [2020]  \
UF [-]
Acre                          164123.964                                894470
```

Alagoas	27843.295	3351543
Amapá	142470.762	861773
Amazonas	1559167.889	4207714
Bahia	564760.427	14930634

UF [-]	Densidade demográfica - hab/km² [2010] \
Acre	4.47
Alagoas	112.33
Amapá	4.69
Amazonas	2.23
Bahia	24.82

UF [-]	Matrículas no ensino fundamental - matrículas [2018] \
Acre	157646
Alagoas	490587
Amapá	136185
Amazonas	705007
Bahia	2034711

UF [-]	IDH Índice de desenvolvimento humano [2010] \
Acre	0.663
Alagoas	0.631
Amapá	0.708
Amazonas	0.674
Bahia	0.660

UF [-]	Receitas realizadas - R$ (×1000) [2017] \
Acre	6.632883e+06
Alagoas	1.195044e+07
Amapá	5.396417e+06
Amazonas	1.732846e+07
Bahia	5.019100e+07

UF [-]	Despesas empenhadas - R$ (×1000) [2017] \
Acre	6.084417e+06
Alagoas	1.046063e+07
Amapá	4.224464e+06
Amazonas	1.532490e+07
Bahia	4.557016e+07

UF [-]	Rendimento mensal domiciliar per capita - R$ [2019] \

```
Acre                                                    890
Alagoas                                                 731
Amapá                                                   880
Amazonas                                                842
Bahia                                                   913

          Total de veículos - veículos [2018]
UF [-]
Acre                                    277831
Alagoas                                 834827
Amapá                                   195039
Amazonas                                883083
Bahia                                  4139107
```

[9]:
```
data.describe()
```

[9]:
```
       Área Territorial - km² [2019]  População estimada - pessoas [2020]  \
count                   2.700000e+01                          2.700000e+01
mean                    3.151961e+05                          7.842803e+06
std                     3.751197e+05                          9.316952e+06
min                     5.760783e+03                          6.311810e+05
25%                     7.609896e+04                          2.932272e+06
50%                     2.236445e+05                          4.064052e+06
75%                     3.349228e+05                          9.401862e+06
max                     1.559168e+06                          4.628933e+07

       Densidade demográfica - hab/km² [2010]  \
count                               27.000000
mean                                68.040741
std                                105.909690
min                                  2.010000
25%                                  6.325000
50%                                 33.410000
75%                                 71.475000
max                                444.660000

       Matrículas no ensino fundamental - matrículas [2018]  \
count                                       2.700000e+01
mean                                        1.006814e+06
std                                         1.079327e+06
min                                         9.658200e+04
25%                                         3.908680e+05
50%                                         5.562480e+05
75%                                         1.300333e+06
max                                         5.367614e+06

       IDH <span>Índice de desenvolvimento humano</span> [2010]  \
```

count	27.000000
mean	0.704519
std	0.049284
min	0.631000
25%	0.664000
50%	0.699000
75%	0.737500
max	0.824000

	Receitas realizadas - R$ (×1000) [2017] \
count	2.700000e+01
mean	3.571354e+07
std	4.590409e+07
min	4.419450e+06
25%	1.203733e+07
50%	1.968562e+07
75%	3.681568e+07
max	2.328225e+08

	Despesas empenhadas - R$ (×1000) [2017] \
count	2.700000e+01
mean	3.237044e+07
std	4.592375e+07
min	3.384684e+06
25%	9.875718e+06
50%	1.762717e+07
75%	2.945779e+07
max	2.319822e+08

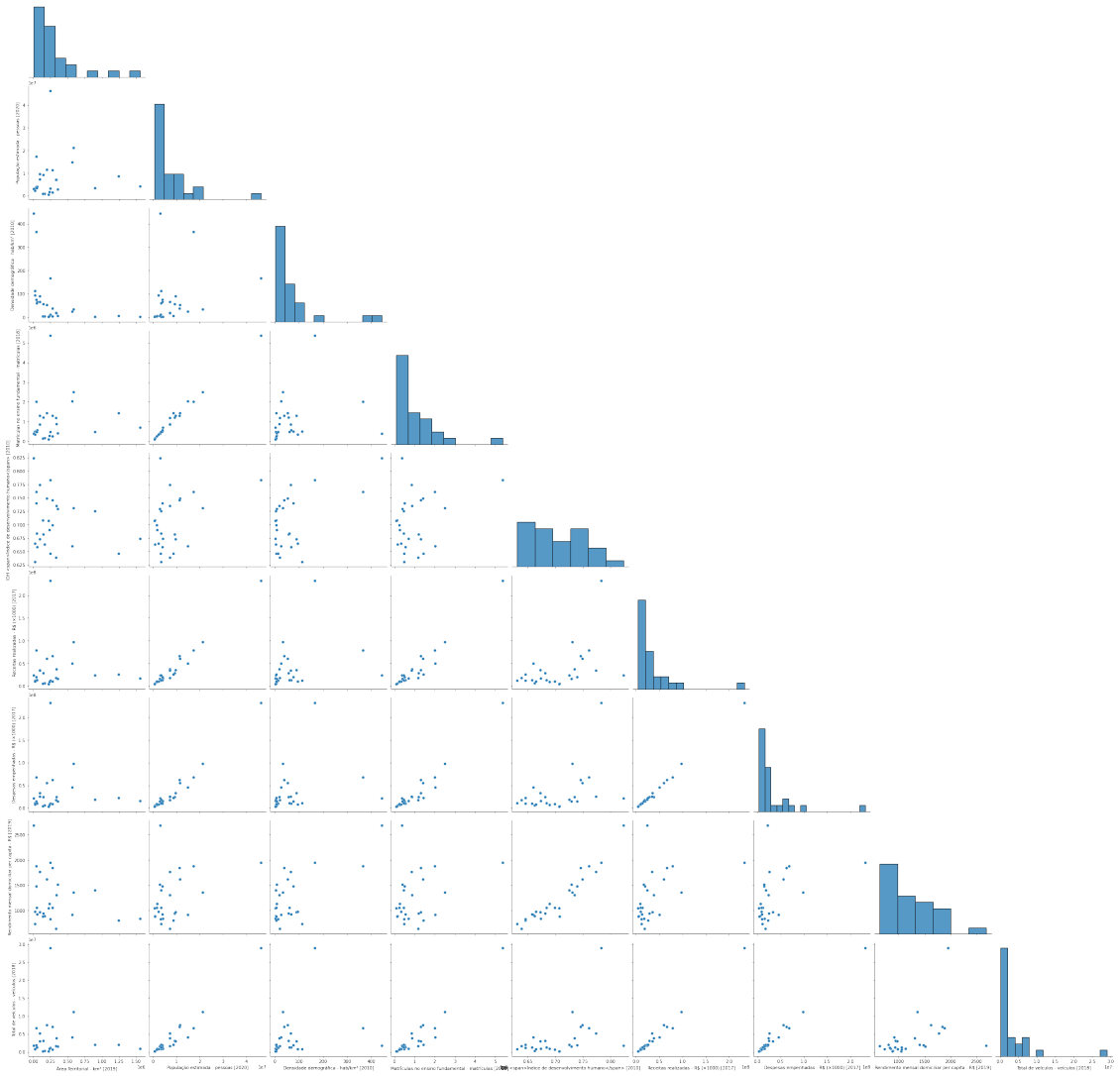
	Rendimento mensal domiciliar per capita - R$ [2019] \
count	27.000000
mean	1238.703704
std	476.693968
min	636.000000
25%	901.500000
50%	1056.000000
75%	1495.500000
max	2686.000000

	Total de veículos - veículos [2018]
count	2.700000e+01
mean	3.731354e+06
std	5.729603e+06
min	1.950390e+05
25%	9.340650e+05
50%	1.812473e+06
75%	4.024268e+06

```
max                                2.905775e+07
```

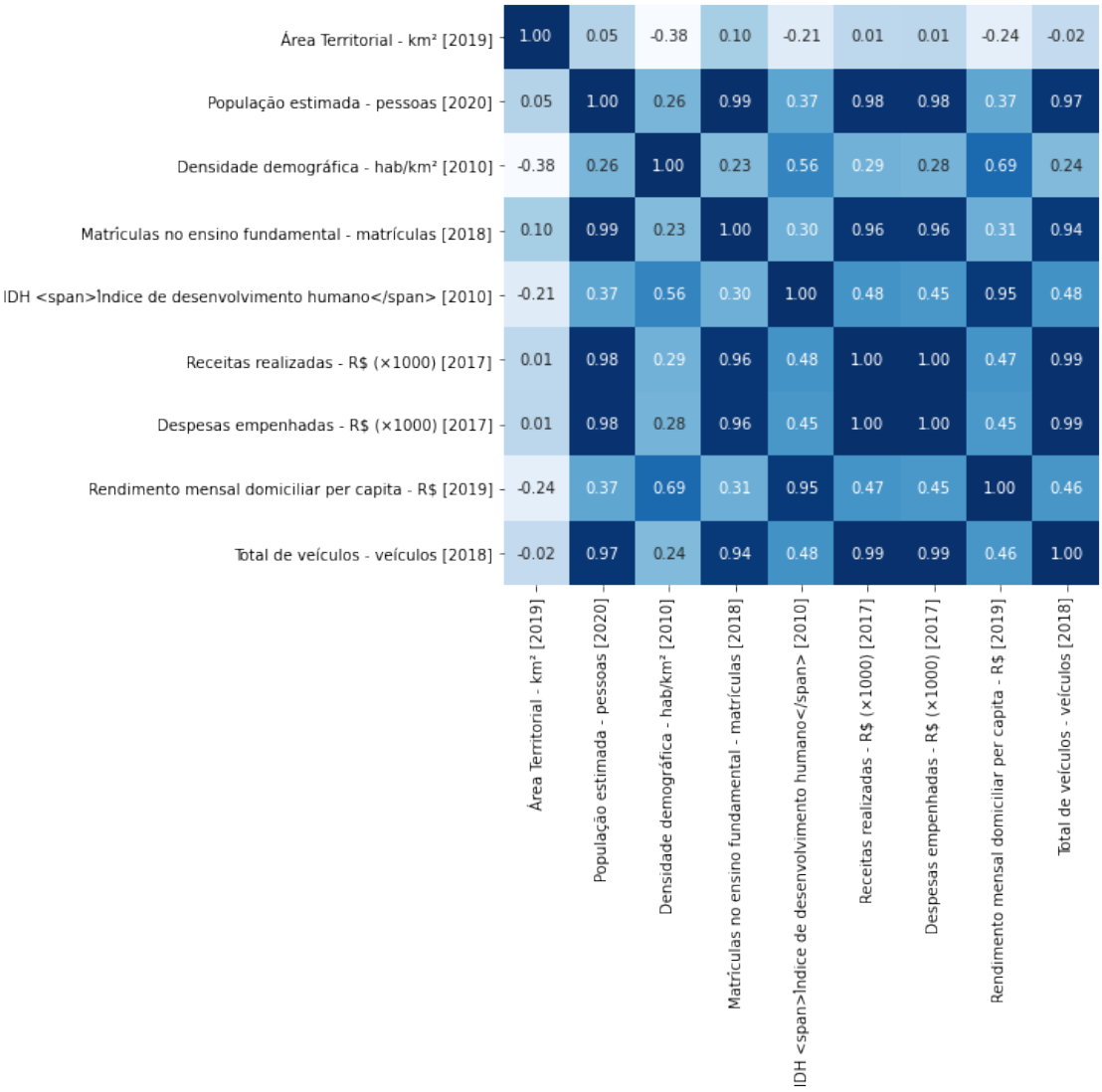
[10]:
```
plt.figure(figsize=(15, 15))
sns.pairplot(data,kind='scatter', diag_kind='hist', corner=True, height=4 )
```

[10]:
```
<seaborn.axisgrid.PairGrid at 0x1872bc77970>

<Figure size 1080x1080 with 0 Axes>
```



[11]:
```
plt.figure(figsize=(7, 7))
corr = np.corrcoef(data.values, rowvar=False)
sns.heatmap(corr, annot=True, cmap='Blues', fmt='.2f', cbar=False,␣
 ↪xticklabels=data.columns, yticklabels=data.columns)
```

[11]: <AxesSubplot:>



Com as análises descritivas e exploratórias, podemos perceber que tem algumas colunas que são bem correlacionadas, o que já era esperado, pois são dados economicos então é esperado por exemplo que o IDH seja relacionado com o rendimento familiar per capita.

1.2 Agrupamento hierárquico via dendrograma

[12]:
```
data_s = data.copy()
data_s.iloc[:, :] = StandardScaler().fit_transform(data)

data_s.head()
```

[12]:

UF [-]	Área Territorial - km² [2019]	População estimada - pessoas [2020]
Acre	-0.410402	-0.759980
Alagoas	-0.780622	-0.491235
Amapá	-0.469225	-0.763556
Amazonas	3.379371	-0.397591
Bahia	0.677966	0.775237

UF [-]	Densidade demográfica - hab/km² [2010]
Acre	-0.611669
Alagoas	0.426146
Amapá	-0.609553
Amazonas	-0.633222
Bahia	-0.415864

UF [-]	Matrículas no ensino fundamental - matrículas [2018]
Acre	-0.801744
Alagoas	-0.487397
Amapá	-0.822007
Amazonas	-0.284952
Bahia	0.970492

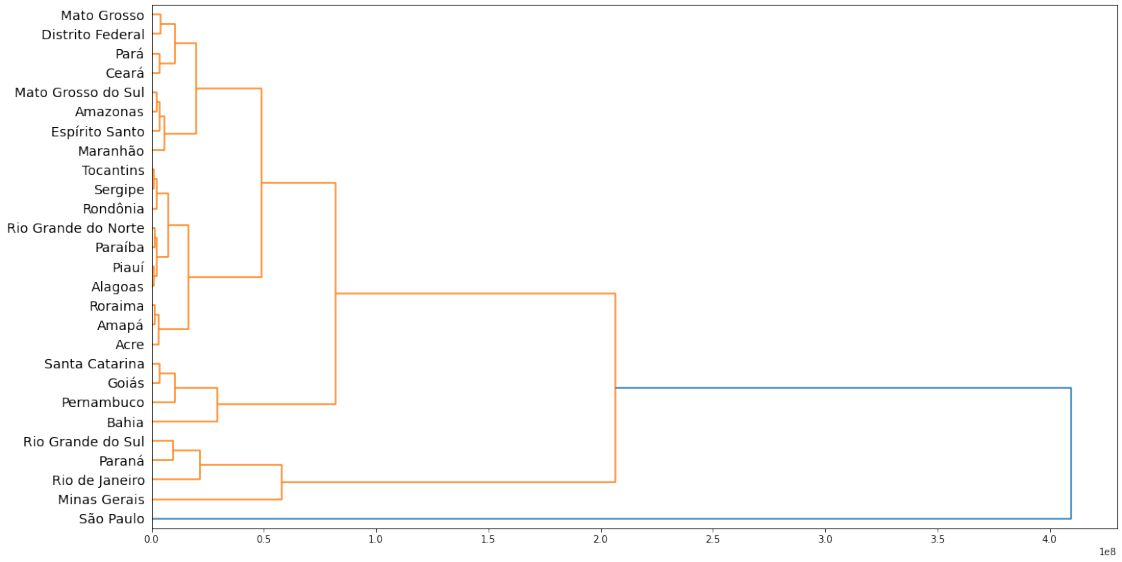
UF [-]	IDH Índice de desenvolvimento humano [2010]
Acre	-0.858489
Alagoas	-1.520160
Amapá	0.071987
Amazonas	-0.631039
Bahia	-0.920520

UF [-]	Receitas realizadas - R$ (×1000) [2017]
Acre	-0.645577
Alagoas	-0.527530
Amapá	-0.673026
Amazonas	-0.408140
Bahia	0.321393

UF [-]	Despesas empenhadas - R$ (×1000) [2017]
Acre	-0.583288
Alagoas	-0.486179
Amapá	-0.624560
Amazonas	-0.378241
Bahia	0.292902

```
          Rendimento mensal domiciliar per capita - R$ [2019]  \
UF [-]
Acre                                                -0.745439
Alagoas                                             -1.085340
Amapá                                               -0.766816
Amazonas                                            -0.848051
Bahia                                               -0.696271

          Total de veículos - veículos [2018]
UF [-]
Acre                                    -0.614233
Alagoas                                 -0.515167
Amapá                                   -0.628958
Amazonas                                -0.506585
Bahia                                    0.072522
```

[13]:
```
Z = hierarchy.linkage(data, 'ward')
plt.figure(figsize=(18, 10))
dn = hierarchy.dendrogram(Z, labels=list(data.index),␣
↪leaf_font_size=14,orientation='right')
```



[14]:
```
n_clusters = 3
cluster = AgglomerativeClustering(n_clusters=n_clusters, affinity='euclidean',␣
↪linkage='ward')
groups = cluster.fit_predict(data)

estados = list(data.index)
```

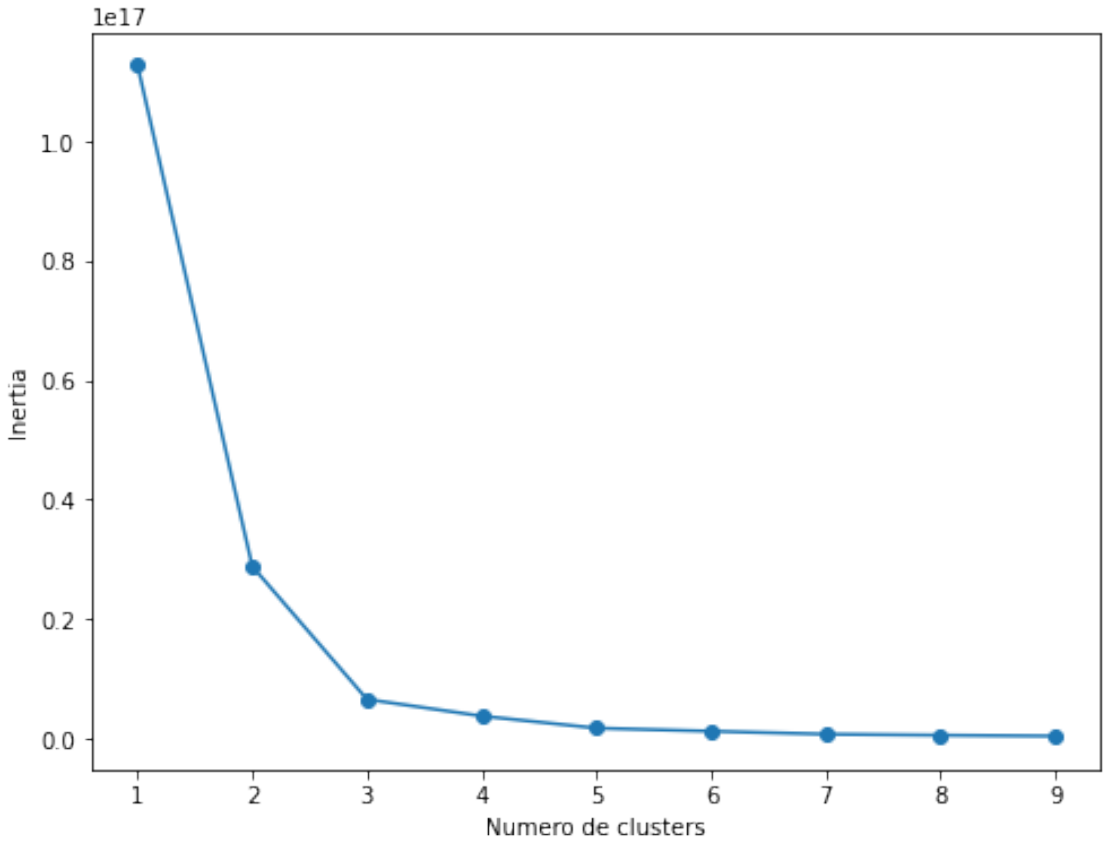
```
g_estados = {i: [] for i in range(n_clusters)}
for estados, group in zip(estados, groups):
    g_estados[group].append(estados)

for gp, est in g_estados.items():
    print(f'Cluster {gp}: {est}\n')
```

```
Cluster 0: ['Acre', 'Alagoas', 'Amapá', 'Amazonas', 'Bahia', 'Ceará', 'Distrito
Federal', 'Espírito Santo', 'Goiás', 'Maranhão', 'Mato Grosso', 'Mato Grosso do
Sul', 'Pará', 'Paraíba', 'Pernambuco', 'Piauí', 'Rio Grande do Norte',
'Rondônia', 'Roraima', 'Santa Catarina', 'Sergipe', 'Tocantins']

Cluster 1: ['São Paulo']

Cluster 2: ['Minas Gerais', 'Paraná', 'Rio de Janeiro', 'Rio Grande do Sul']
```

Para essa análise de agrupamento hierárquica via dendrograma foi utilizado a medida euclidiana para medir a distância e o método de ligação Ward, foi realizado um corte na distância 1 para definir o número de clusters. Com essa ánalise de cluster podemos notar uma forte relação geográfica com os dados económicos. Sendo os estados mais ao sudeste e sul sendo separados do resto dos outros estados, o que já era esperado pois é a região mais forte economicamente do país. Vale ressaltar que o estado de São Paulo ficou bem distante dos demais, o que também bate com a realidade, pois é o estado com a maior economia do país.

1.3 Agrupamento não-hierárquico (K-médias)

[15]:
```
inertias = []
for k in np.arange(1, 10):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(data)
    inertias.append(kmeans.inertia_)

# Plotando o gráfico do Elbow Method
plt.figure(figsize=(8, 6))
plt.plot(np.arange(1, 10), inertias, '-o')
plt.xlabel('Numero de clusters')
plt.ylabel('Inertia');
```



[16]:
```
n_clusters = 3
kmeans = KMeans(n_clusters=n_clusters)
data['cluster'] = kmeans.fit_predict(data)
data
```

[16]:

UF [-]	Área Territorial - km² [2019]	\
Acre	164123.964	
Alagoas	27843.295	
Amapá	142470.762	
Amazonas	1559167.889	
Bahia	564760.427	
Ceará	148894.441	
Distrito Federal	5760.783	
Espírito Santo	46074.447	
Goiás	340203.329	
Maranhão	329642.182	
Mato Grosso	903207.019	
Mato Grosso do Sul	357145.534	
Minas Gerais	586521.123	

Pará	1245870.798
Paraíba	56467.242
Paraná	199298.979
Pernambuco	98067.881
Piauí	251756.515
Rio de Janeiro	43750.427
Rio Grande do Norte	52809.602
Rio Grande do Sul	281707.156
Rondônia	237765.240
Roraima	223644.527
Santa Catarina	95730.684
São Paulo	248219.481
Sergipe	21925.424
Tocantins	277466.763

UF [-]	População estimada - pessoas [2020] \
Acre	894470
Alagoas	3351543
Amapá	861773
Amazonas	4207714
Bahia	14930634
Ceará	9187103
Distrito Federal	3055149
Espírito Santo	4064052
Goiás	7113540
Maranhão	7114598
Mato Grosso	3526220
Mato Grosso do Sul	2809394
Minas Gerais	21292666
Pará	8690745
Paraíba	4039277
Paraná	11516840
Pernambuco	9616621
Piauí	3281480
Rio de Janeiro	17366189
Rio Grande do Norte	3534165
Rio Grande do Sul	11422973
Rondônia	1796460
Roraima	631181
Santa Catarina	7252502
São Paulo	46289333
Sergipe	2318822
Tocantins	1590248

UF [-]	Densidade demográfica - hab/km^2 [2010] \

Acre	4.47
Alagoas	112.33
Amapá	4.69
Amazonas	2.23
Bahia	24.82
Ceará	56.76
Distrito Federal	444.66
Espírito Santo	76.25
Goiás	17.65
Maranhão	19.81
Mato Grosso	3.36
Mato Grosso do Sul	6.86
Minas Gerais	33.41
Pará	6.07
Paraíba	66.70
Paraná	52.40
Pernambuco	89.62
Piauí	12.40
Rio de Janeiro	365.23
Rio Grande do Norte	59.99
Rio Grande do Sul	37.96
Rondônia	6.58
Roraima	2.01
Santa Catarina	65.27
São Paulo	166.23
Sergipe	94.36
Tocantins	4.98

UF [-]	Matrículas no ensino fundamental - matrículas [2018] \
Acre	157646
Alagoas	490587
Amapá	136185
Amazonas	705007
Bahia	2034711
Ceará	1198116
Distrito Federal	377622
Espírito Santo	502059
Goiás	877593
Maranhão	1178949
Mato Grosso	471613
Mato Grosso do Sul	404114
Minas Gerais	2511483
Pará	1439788
Paraíba	556248
Paraná	1427218
Pernambuco	1301930

Piauí	480126
Rio de Janeiro	2003315
Rio Grande do Norte	467629
Rio Grande do Sul	1298736
Rondônia	269626
Roraima	96582
Santa Catarina	851993
São Paulo	5367614
Sergipe	331297
Tocantins	246183

	IDH Índice de desenvolvimento humano [2010]
\	
UF [-]	
Acre	0.663
Alagoas	0.631
Amapá	0.708
Amazonas	0.674
Bahia	0.660
Ceará	0.682
Distrito Federal	0.824
Espírito Santo	0.740
Goiás	0.735
Maranhão	0.639
Mato Grosso	0.725
Mato Grosso do Sul	0.729
Minas Gerais	0.731
Pará	0.646
Paraíba	0.658
Paraná	0.749
Pernambuco	0.673
Piauí	0.646
Rio de Janeiro	0.761
Rio Grande do Norte	0.684
Rio Grande do Sul	0.746
Rondônia	0.690
Roraima	0.707
Santa Catarina	0.774
São Paulo	0.783
Sergipe	0.665
Tocantins	0.699

	Receitas realizadas - R$ (×1000) [2017] \
UF [-]	
Acre	6.632883e+06
Alagoas	1.195044e+07
Amapá	5.396417e+06

Amazonas	1.732846e+07
Bahia	5.019100e+07
Ceará	2.842022e+07
Distrito Federal	2.381221e+07
Espírito Santo	1.968562e+07
Goiás	3.788534e+07
Maranhão	1.850326e+07
Mato Grosso	2.395853e+07
Mato Grosso do Sul	1.639666e+07
Minas Gerais	9.719982e+07
Pará	2.584945e+07
Paraíba	1.309701e+07
Paraná	6.016358e+07
Pernambuco	3.574603e+07
Piauí	1.212422e+07
Rio de Janeiro	7.848814e+07
Rio Grande do Norte	1.352755e+07
Rio Grande do Sul	6.639747e+07
Rondônia	9.122311e+06
Roraima	4.419450e+06
Santa Catarina	3.469677e+07
São Paulo	2.328225e+08
Sergipe	1.014505e+07
Tocantins	1.030510e+07

UF [-]	Despesas empenhadas - R$ (×1000) [2017] \
Acre	6.084417e+06
Alagoas	1.046063e+07
Amapá	4.224464e+06
Amazonas	1.532490e+07
Bahia	4.557016e+07
Ceará	2.460835e+07
Distrito Federal	2.199046e+07
Espírito Santo	1.439234e+07
Goiás	2.424838e+07
Maranhão	1.762717e+07
Mato Grosso	1.818736e+07
Mato Grosso do Sul	1.450692e+07
Minas Gerais	9.839167e+07
Pará	2.253347e+07
Paraíba	1.007470e+07
Paraná	5.553440e+07
Pernambuco	3.332049e+07
Piauí	9.676736e+06
Rio de Janeiro	6.796555e+07
Rio Grande do Norte	1.133096e+07

Rio Grande do Sul	6.247628e+07
Rondônia	7.085530e+06
Roraima	3.384684e+06
Santa Catarina	2.559510e+07
São Paulo	2.319822e+08
Sergipe	8.494927e+06
Tocantins	8.929456e+06

UF [-]	Rendimento mensal domiciliar per capita - R$ [2019] \
Acre	890
Alagoas	731
Amapá	880
Amazonas	842
Bahia	913
Ceará	942
Distrito Federal	2686
Espírito Santo	1477
Goiás	1306
Maranhão	636
Mato Grosso	1403
Mato Grosso do Sul	1514
Minas Gerais	1358
Pará	807
Paraíba	929
Paraná	1621
Pernambuco	970
Piauí	827
Rio de Janeiro	1882
Rio Grande do Norte	1057
Rio Grande do Sul	1843
Rondônia	1136
Roraima	1044
Santa Catarina	1769
São Paulo	1946
Sergipe	980
Tocantins	1056

UF [-]	Total de veículos - veículos [2018]	cluster
Acre	277831	0
Alagoas	834827	0
Amapá	195039	0
Amazonas	883083	0
Bahia	4139107	2
Ceará	3148369	0
Distrito Federal	1812473	0

Espírito Santo	1936862	0
Goiás	3909429	0
Maranhão	1696683	0
Mato Grosso	2080848	0
Mato Grosso do Sul	1583142	0
Minas Gerais	11191341	2
Pará	2013952	0
Paraíba	1293668	0
Paraná	7571122	2
Pernambuco	3010638	0
Piauí	1196192	0
Rio de Janeiro	6725822	2
Rio Grande do Norte	1290903	0
Rio Grande do Sul	7077972	2
Rondônia	985047	0
Roraima	219290	0
Santa Catarina	5152615	0
São Paulo	29057749	1
Sergipe	772380	0
Tocantins	690169	0

[17]:
```
estados = list(data.index)
groups = data['cluster']

g_estados = {i: [] for i in range(n_clusters)}

for estado, group in zip(estados, groups):
    g_estados[group].append(estado)

for gp, est in g_estados.items():
    print(f'Cluster {gp}: {est}\n')
```

```
Cluster 0: ['Acre', 'Alagoas', 'Amapá', 'Amazonas', 'Ceará', 'Distrito Federal',
'Espírito Santo', 'Goiás', 'Maranhão', 'Mato Grosso', 'Mato Grosso do Sul',
'Pará', 'Paraíba', 'Pernambuco', 'Piauí', 'Rio Grande do Norte', 'Rondônia',
'Roraima', 'Santa Catarina', 'Sergipe', 'Tocantins']

Cluster 1: ['São Paulo']

Cluster 2: ['Bahia', 'Minas Gerais', 'Paraná', 'Rio de Janeiro', 'Rio Grande do
Sul']
```

A análise de agrupamento não-hierárquica via k-médias, teve resultado bem semelhantes com a análise hieraquica, com o estado de São Paulo destoando dos demais, e logo após no segundo cluste os estados do sul e sudeste, a diferença nesse método foi o distrito federal, que é geográficamente distante, estar nesse mesmo cluster, mas como é um estado bem pequeno e que contém a capital do país é extremamente compreensível ele estar junto nesse cluster, pois a capital movimenta muito

dinheiro também. Para escolher o número de clusters foi utlizado o “método de cotovelo”.