Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise de Cluster e Discriminante

Douglas de Paula Nestlehner
Hélio Mota Ezequiel

São Carlos
2022

Sumário

Capítulo 1

Análise de Cluster

Considerando os dados padronizados dos gastos de 25 municípios do estado de São Paulo:

Admin	Educ	Saude	Urban	Munic
-1.02	-0.36	-1.12	-1.04	Adamantina
1.39	0.64	1.35	0.28	Aracatuba
0.32	0.82	1.38	0.22	Araraquara
-0.80	-0.91	-0.34	-0.52	Assis
-1.04	-0.58	-0.55	-0.31	Avare
1.62	0.55	-0.21	0.36	Barretos
1.71	1.36	0.78	1.88	Bauru
0.49	-0.80	-0.76	-0.44	Botucatu
0.80	0.09	-0.06	0.01	Braganca_Pta
1.49	0.31	-0.13	-1.25	Caraguatatuba
-0.10	0.38	-0.29	-0.46	Catanduva
0.58	1.51	2.46	1.67	Franca
-0.76	-0.33	-0.14	0.43	Guaratingueta
-0.37	-0.12	-0.20	-0.25	Itapetininga
-0.86	0.61	-0.36	-0.76	Itapeva
-1.11	-1.71	-1.10	-1.03	Jales
-0.86	-0.27	-0.78	-0.21	Jau
0.98	2.38	1.46	2.58	Limeira
2.61	0.67	2.78	0.58	Marilia
0.02	0.59	1.75	1.38	Pres_Prudente
-1.47	-1.21	-0.89	-0.75	Registro
1.08	0.34	0.19	0.45	Rio_Claro
-0.80	-0.53	-0.51	-0.75	SJoao_Boa_Vista
-0.69	-1.62	-1.06	-0.64	Tupa
-0.60	-1.14	-0.89	-0.86	Votuporanga

1. Realizar um agrupamento pelo método de linkagem que for sorteado para o grupo;
2. Realizar o agrupamento pelo método das k-médias;
3. Indicar o agrupamento das k-médias no mapa, verificando se há um possível agrupamento espacial. Pode usar o SAS ou o R.

1.1 Análise descritiva

Antes de realizar o agrupamento, realizamos uma breve análise descritiva nos dados, no intuito de verificar alguma inconsistência.

Na Tabela 1.1 representamos algumas medidas descritivas, para as variáveis quantitativas.

Tabela 1.1: Medidas descritivas.

Admin	Educ	Saude	Urban
Min. :-1.4707	Min. :-1.71290	Min. :-1.1179	Min. :-1.24750
1st Qu.:-0.8028	1st Qu.:-0.57870	1st Qu.:-0.7639	1st Qu.:-0.75060
Median :-0.1034	Median : 0.09180	Median :-0.2117	Median :-0.25330
Mean : 0.1044	Mean : 0.02692	Mean : 0.1112	Mean : 0.02322
3rd Qu.: 0.9819	3rd Qu.: 0.61320	3rd Qu.: 0.7840	3rd Qu.: 0.42820
Max. : 2.6099	Max. : 2.38120	Max. : 2.7797	Max. : 2.58210

Em geral, as variáveis se encontram na mesma escala, e aparentam comportamento semelhante, também construímos os boxplot de cada variável, representado na Figura 1.1.

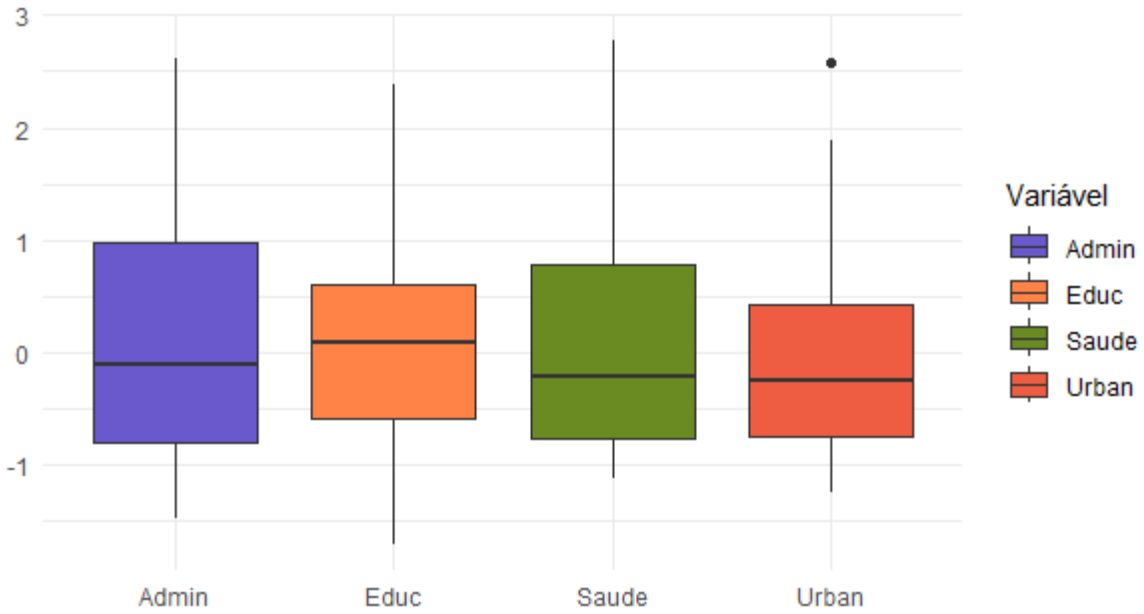

Figura 1.1: Boxplot de cada variável quantitativa.

Assim como observado na Tabela 1.1, notamos que os dados não apresentam muitos outliers, e estão na mesma escala, não tendo a necessidade de padronizar.

Observamos também a presença de outilier na variavel “Urban” o que poderia impactar na análise, pois o método de linkagem completa (a ser utilizado) é sensível a outliers,

entretanto, por termos apenas um caso, decidimos por prosseguir considerando no estudo.

1.2 Agrupamento linkagem Completa

Em seguida, calculamos a matriz de distancias entre as observações no contexto multivariado, considerando a distancia *Euclidiana* para o calculo. Na tabela 1.2, representamos as distancias calculadas para as cinco primeiras observações.

Tabela 1.2: Distancias das cinco primeiras observações.

	Adamantina	Aracatuba	Araraquara	Assis	Avare
Adamantina	0.00	3.83	3.32	1.11	0.95
Aracatuba	3.83	0.00	1.09	3.27	3.36
Araraquara	3.32	1.09	0.00	2.79	2.79
Assis	1.11	3.27	2.79	0.00	0.50
Avare	0.95	3.36	2.79	0.50	0.00

Com as distancias calculadas, realizamos o agrupamento pelo método hierárquico aglomerativo de linkagem completa. Na Figura 1.2 representamos pelo dendrograma o agrupamento obtido.

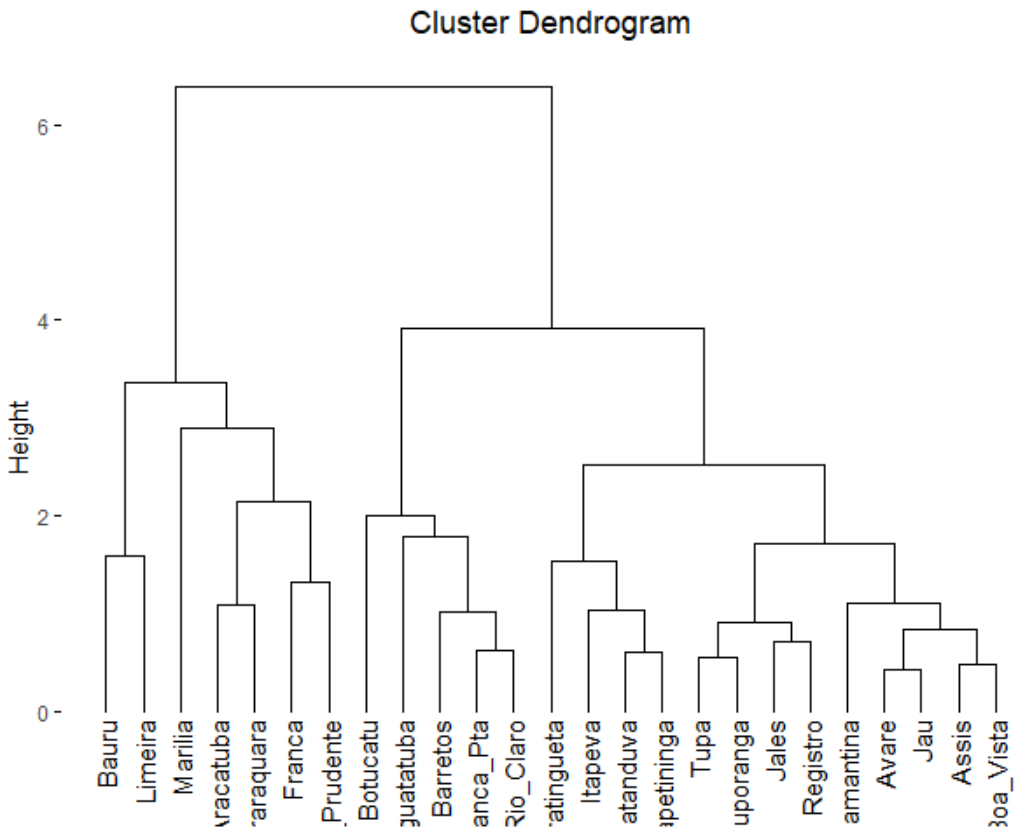

Figura 1.2: Dendrograma.

Para definir o numero de cluster a ser escolhido, utilizamos inicialmente o método que considera a altura entre as junções dos grupos. Também levamos em consideração o critério proposto por Mojena (1977).

$$\text{Corte} = \bar{h} + 1.25 * s_h,$$

em que $\bar{h}$ é a média das alturas, e s_h o desvio padrão das alturas.

Na Figura 1.3 representamos as estimativas das alturas, em vermelho o corte estimado pelo critério de Mojena.

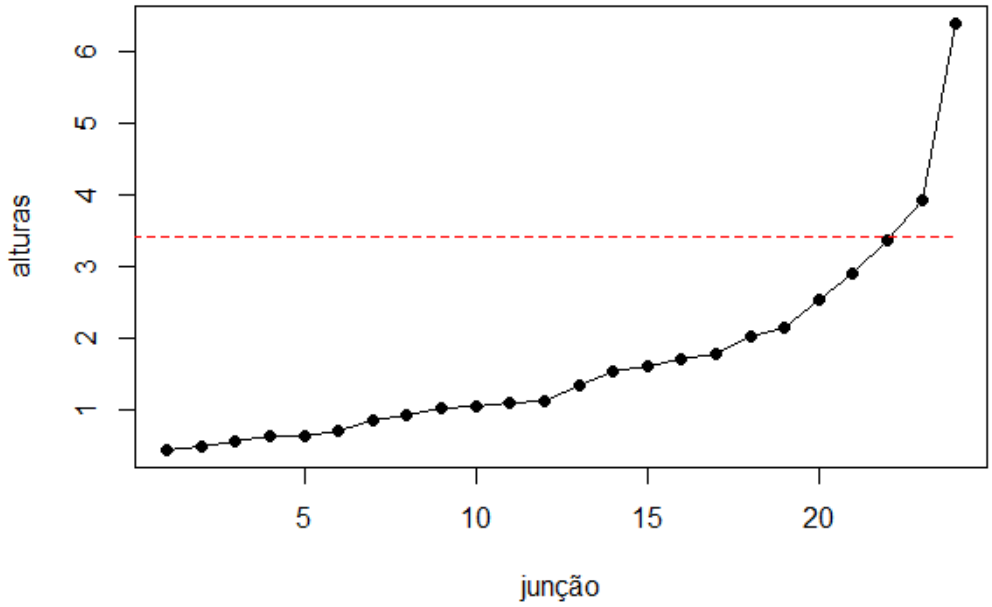

Figura 1.3: Gráfico das alturas.

Pela Figura 1.3, o corte escolhido seria algo próximo de 4.5, entretanto pelo critério de Mojena, o corte estimado foi de 3.40. Para continuidade da análise, avaliamos pela figura 1.2 a quantidade de clusters que seriam escolhidos no intervalo [3.40, 4.5], e definimos com melhor escolha, um numero total de 3 cluster.

Desse modo, no dendrograma representado na Figura 1.4 representamos os clusters obtidos.

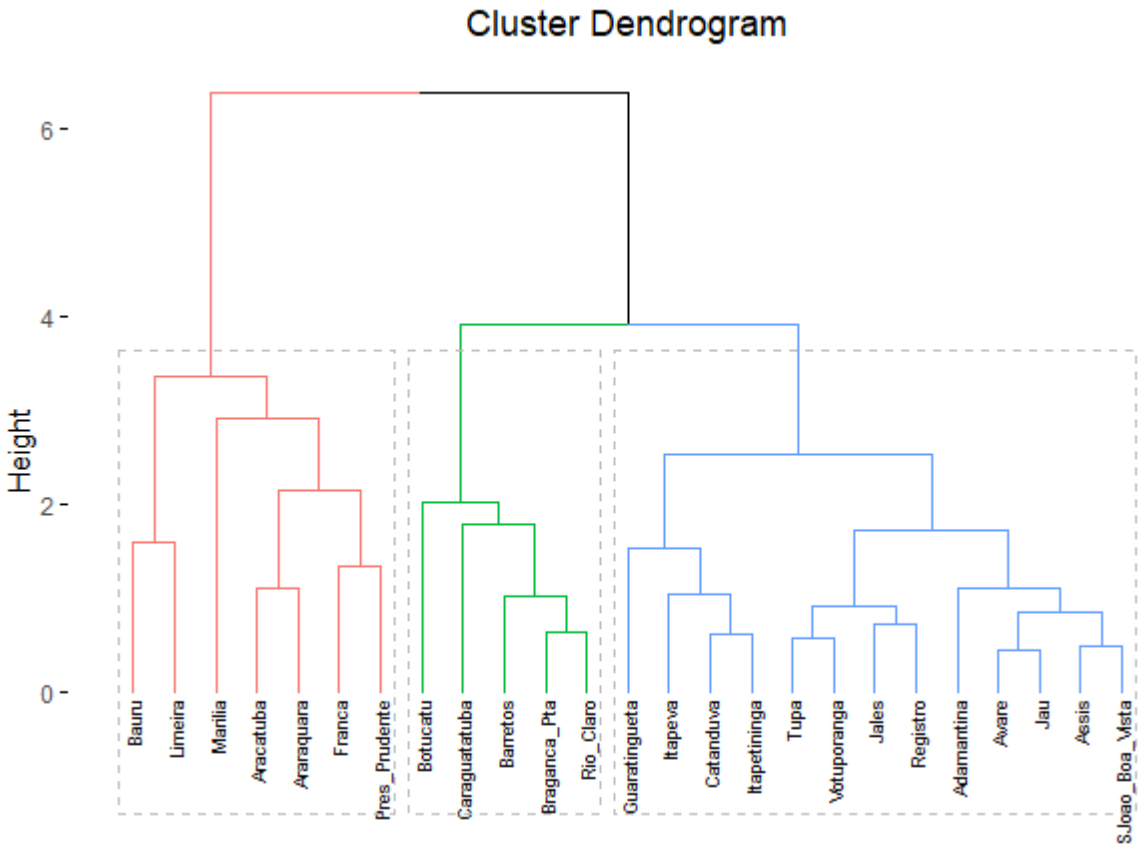

Figura 1.4: Dendrograma com k = 3.

1.3 Agrupamento K-médias

O método de clusterização K-means é um algoritmo de aprendizado não supervisionado utilizado para agrupar objetos ou dados em clusters ou grupos com base em suas características similares. O objetivo do algoritmo é agrupar os dados de forma que a variação dentro dos clusters seja minimizada e a variação entre os clusters seja maximizada.

O processo de clusterização K-means começa com a definição do número de clusters, k, que o usuário deseja identificar. Em seguida, o algoritmo seleciona aleatoriamente k objetos como centroides iniciais para cada cluster. A partir daí, cada objeto é atribuído ao cluster cujo centróide está mais próximo. Essa etapa é chamada de "atribuição de cluster".

Depois disso, os centroides de cada cluster são recalculados, movendo-os para a média dos objetos que pertencem a esse cluster. Isso é chamado de "atualização do centróide". O processo de atribuição de cluster e atualização do centróide é repetido até que não haja mais mudanças nos objetos atribuídos aos clusters.

O resultado final do algoritmo de clusterização K-means é um conjunto de k clusters,

cada um contendo objetos que são mais semelhantes entre si do que aos objetos nos outros clusters.

Realizando o agrupamento pelo método das k-médias para o nosso caso considerando os dados padronizados dos gastos de 25 municípios do estado de São Paulo, o primeiro passo que iremos realizar na análise é determinar o número ótimo de clusters.

Para nos auxiliarmos, iremos utilizar o método do cotovelo. Ele envolve a execução do algoritmo K-means várias vezes, com diferentes valores de K, e plotar a soma dos erros quadráticos (SSE) em relação ao número de clusters. O gráfico resultante mostrará uma curva em forma de cotovelo, e o ponto de inflexão no cotovelo é geralmente considerado o número ideal de clusters.

Utilizando o auxilio de um pacote estatístico chamado factoextra no R, utilizamos uma função e obtemos o seguinte gráfico de cotovelo:

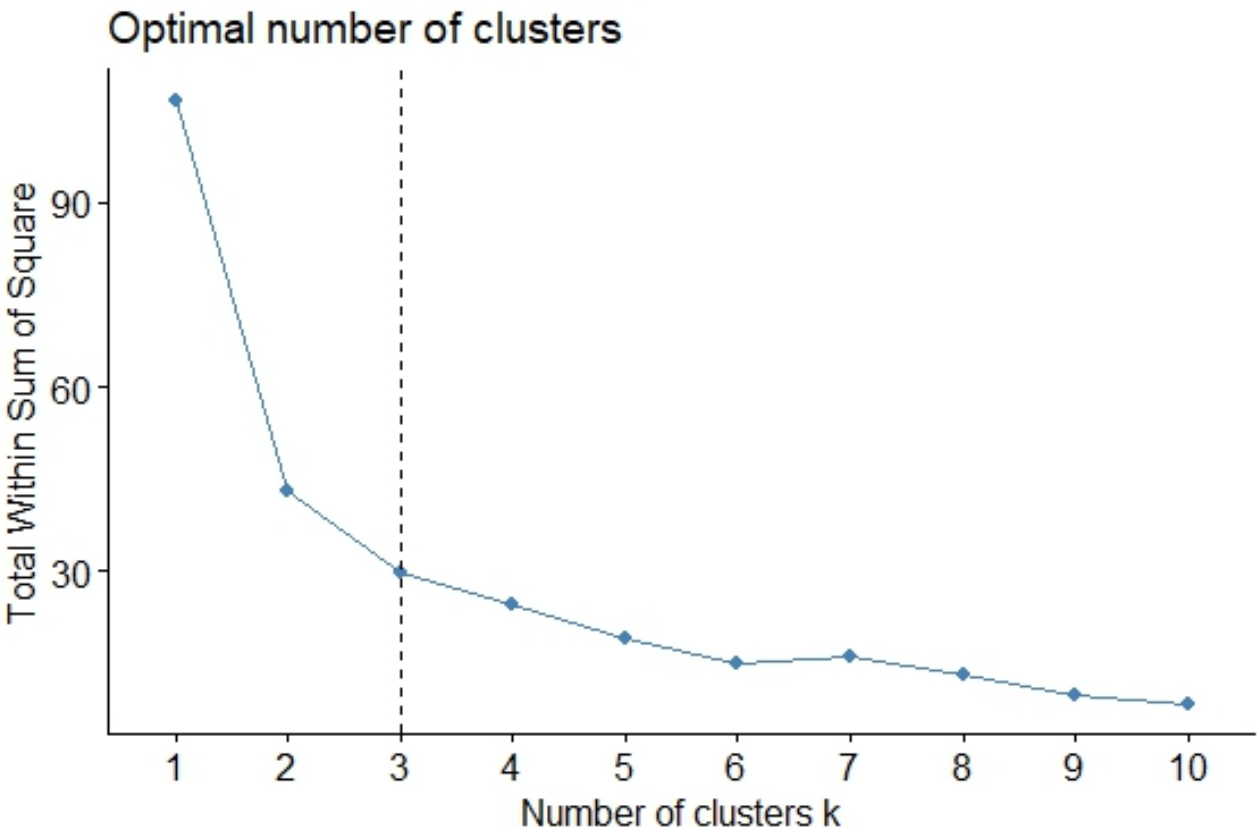

Figura 1.5: Gráfico de Cotovelo

Desta forma, utilizando a noção da soma dos quadrados é possível verificar que o número ótimo de clusters é 3. Isto porque novos clusters acima de 3 possuem baixo ganho para aumentar a diferenciação dos demais.

Em seguida, realizamos o agrupamento pelo método das k-medias com o número de

cluster sendo 3 com o auxilio da função kmeans do R e obtemos o seguinte resultado a seguir:

```
K-means clustering with 3 clusters of sizes 13, 6, 5

Cluster means:
      Admin       Educ      Saude       Urban
1 -0.690600 -0.6333154 -0.6042385 -0.50495385
2  1.117167  0.4588667  0.4198667  0.01463333
3  1.180100  1.3025400  1.8466800  1.61912000

Clustering vector:
      Aracatuba      Araraquara           Assis           Avare
              2               2               1               1
       Barretos           Bauru        Botucatu    Braganca_Pta
              2               3               1               2
  Caraguatatuba       Catanduva          Franca   Guaratingueta
              2               1               3               1
   Itapetininga         Itapeva           Jales             Jau
              1               1               1               1
        Limeira         Marilia   Pres_Prudente        Registro
              3               3               3               1
      Rio_Claro SJoao_Boa_Vista            Tupa     Votuporanga
              2               1               1               1

Within cluster sum of squares by cluster:
[1] 11.72486  6.36522 10.85718
 (between_SS / total_SS =  71.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Como resultado temos que o cluster 1 é composto por 13 cidade, o cluster 2 por 6 cidades e o cluster 3 por 5 cidades. E também temos as médias dos centros por clusters e a classificação de cada cidade dentro do respectivo cluster.

	Cluster	Admin	Educ	Saude	Urban
1	1	-0.6906	-0.6333	-0.6042	-0.5050
2	2	1.1171	0.4588	0.4198	0.0146
3	3	1.1801	1.3025	1.8466	1.6191

Tabela 1.3: Média do cluster

Com o cálculo da média podemos analisar através da tabela 1.3 que as cidades com

gastos administrativos médios mais altos encontram-se no cluster 3, junto com aqueles com maior quantidade de gastos em educação, saúde e urbanização, o que caracteriza cidades mais desenvolvidas. Já no cluster 1, temos as cidades com gastos administrativos médios negativos, juntamente com Educação, Saúde e Urbanização também negativos, o que nos indica cidades menos desenvolvidas. No cluster 2, temos o gasto médio administrativo como 1.12, Educação com 0.46, Saúde com 0.42 e Urbanização bem próximo de zero, com 0.01, o que poderíamos caracterizar como cidades em desenvolvimento.

Para uma melhor visualização dos agrupamentos, podemos plotar um gráfico a partir dos dados originais e os clusters encontrados utilizando a técnica de componentes principais, como podemos ver na imagem a seguir:

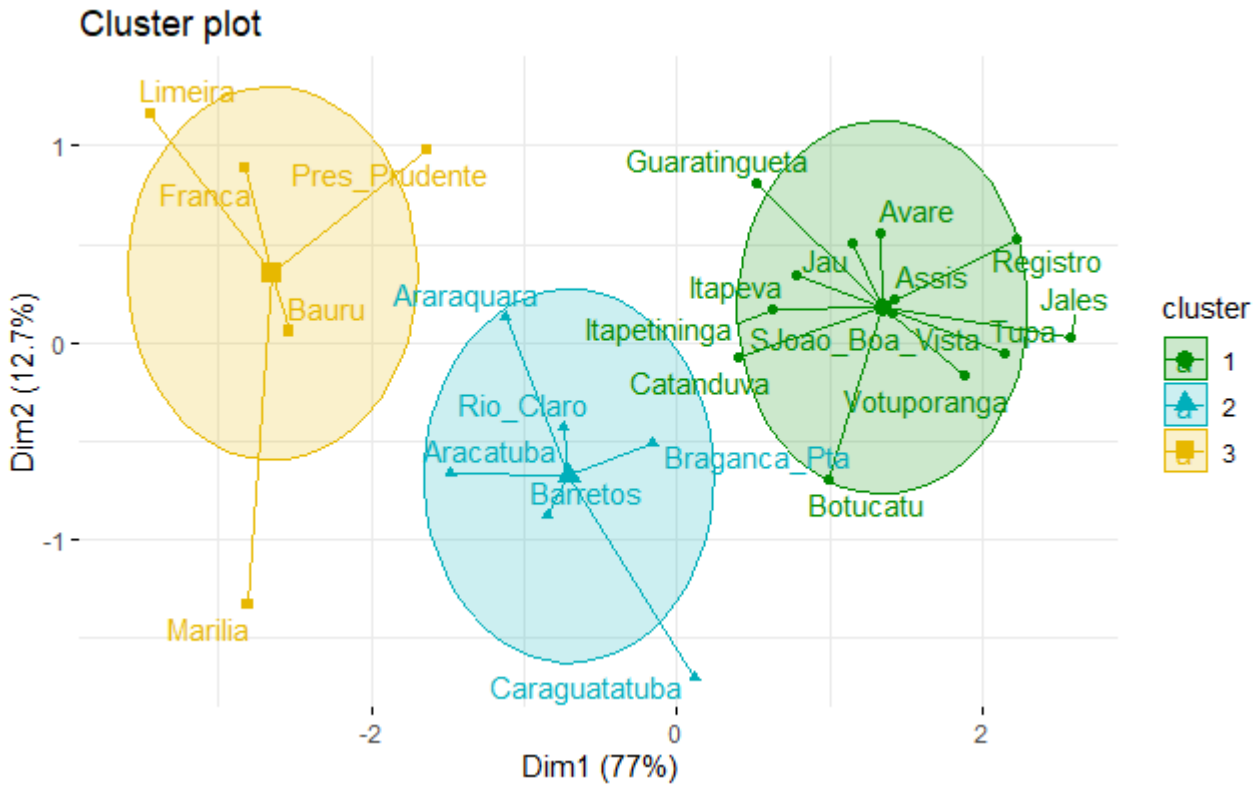

Figura 1.6: Agrupamento pelo método k-means.

Na Figura 1.6, podemos observar os agrupamentos realizados pelo k-means, permitindo observar municípios que foram classificados em um grupo, mas também estão próximos de outro, como exemplo: a cidade de Marília, que foi classificada como grupo 3, mas também está próxima do grupo 2, podendo assim ser uma inconsistência.

1.4 Mapa

Indicando o agrupamento obtido das k-médias no mapa do estado de São Paulo, temos o seguinte resultado a seguir:

```
> km.clust$cluster
     Adamantina        Aracatuba      Araraquara           Assis
              1                2               2               1
          Avare        Barretos           Bauru        Botucatu
              1                2               3               1
   Braganca_Pta   Caraguatatuba       Catanduva         Franca
              2                2               1               3
  Guaratingueta    Itapetininga         Itapeva          Jales
              1                2               1               1
            Jau         Limeira         Marilia   Pres_Prudente
              1                3               3               3
       Registro       Rio_Claro SJoao_Boa_Vista           Tupa
              1                2               1               1
    Votuporanga
              1
```

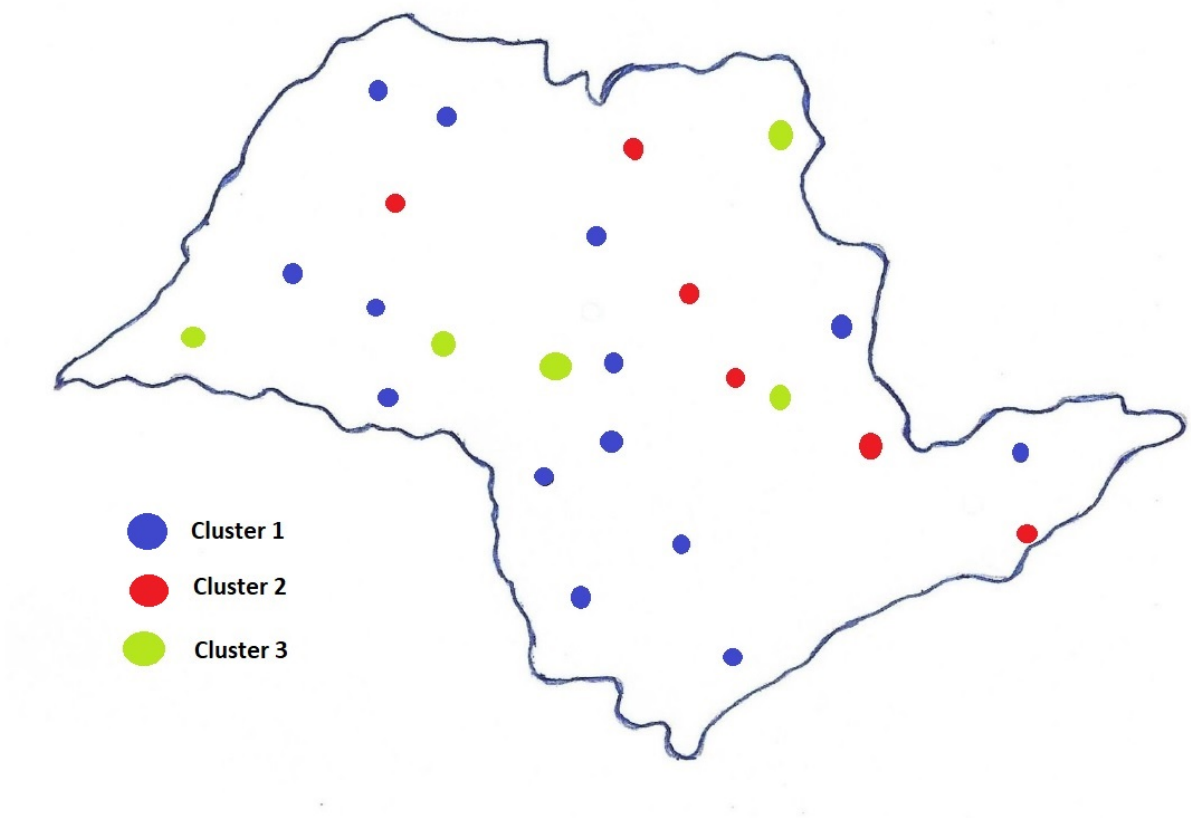

Figura 1.7: Mapa de São Paulo com os Clusters

Na figura 1.7 temos o mapa do estado de São Paulo com as localizações das cidades coloridas com os 3 cluster obtidos a partir do método de agrupamento do k-médias do item anterior. Podemos verificar que não temos nenhum indício que os agrupamentos obtidos tem correlação com a localização das cidades, portanto, concluímos que não há um possível agrupamento espacial.

Capítulo 2

Análise discriminante

Considerando os agrupamentos dos dados dos gastos de municípios do estado de São Paulo, formados pelo método das k-médias no capitulo anterior, obtivemos os seguintes grupos para os municípios analisados:

Grupo	Municipio	Admin	Educ	Saude	Urban
1	Assis	-0.80	-0.91	-0.34	-0.52
1	Avare	-1.04	-0.58	-0.55	-0.31
1	Botucatu	0.49	-0.80	-0.76	-0.44
1	Catanduva	-0.10	0.38	-0.29	-0.46
1	Guaratingueta	-0.76	-0.33	-0.14	0.43
1	Itapetininga	-0.37	-0.12	-0.20	-0.25
1	Itapeva	-0.86	0.61	-0.36	-0.76
1	Jales	-1.11	-1.71	-1.10	-1.03
1	Jau	-0.86	-0.27	-0.78	-0.21
1	Registro	-1.47	-1.21	-0.89	-0.75
1	SJoao_Boa_Vista	-0.80	-0.53	-0.51	-0.75
1	Tupa	-0.69	-1.62	-1.06	-0.64
1	Votuporanga	-0.60	-1.14	-0.89	-0.86
2	Aracatuba	1.39	0.64	1.35	0.28
2	Araraquara	0.32	0.82	1.38	0.22
2	Barretos	1.62	0.55	-0.21	0.36
2	Braganca_Pta	0.80	0.09	-0.06	0.01
2	Caraguatatuba	1.49	0.31	-0.13	-1.25
2	Rio_Claro	1.08	0.34	0.19	0.45
3	Bauru	1.71	1.36	0.78	1.88
3	Franca	0.58	1.51	2.46	1.67
3	Limeira	0.98	2.38	1.46	2.58
3	Marilia	2.61	0.67	2.78	0.58
3	Pres_Prudente	0.02	0.59	1.75	1.38

Contudo, nesse capitulo iremos:

- Caracterizar os grupos formados na análise de cluster, descrevendo-os segundo as

variáveis utilizadas no agrupamento;

- Realizar uma análise discriminante linear (AD) para classificar os municípios, considerando probabilidades a priori que achar conveniente (indicar quais foram as prioris) e descrever os grupos formados pela AD, apresentando os coeficientes da classificação linear, separados por grupo;

- Calcular a APER, os erros de classificação por categoria, o erro global e a taxa média de erros

2.1 Análise descritiva (Clusters)

No intuito de caracterizar cada grupo encontrado na análise de cluster, realizamos uma breve análise descritiva considerando as variáveis utilizadas no agrupamento.

Na Figura 2.1, representamos os box-plots obtidos para cada variável, considerando os grupos obtidos.

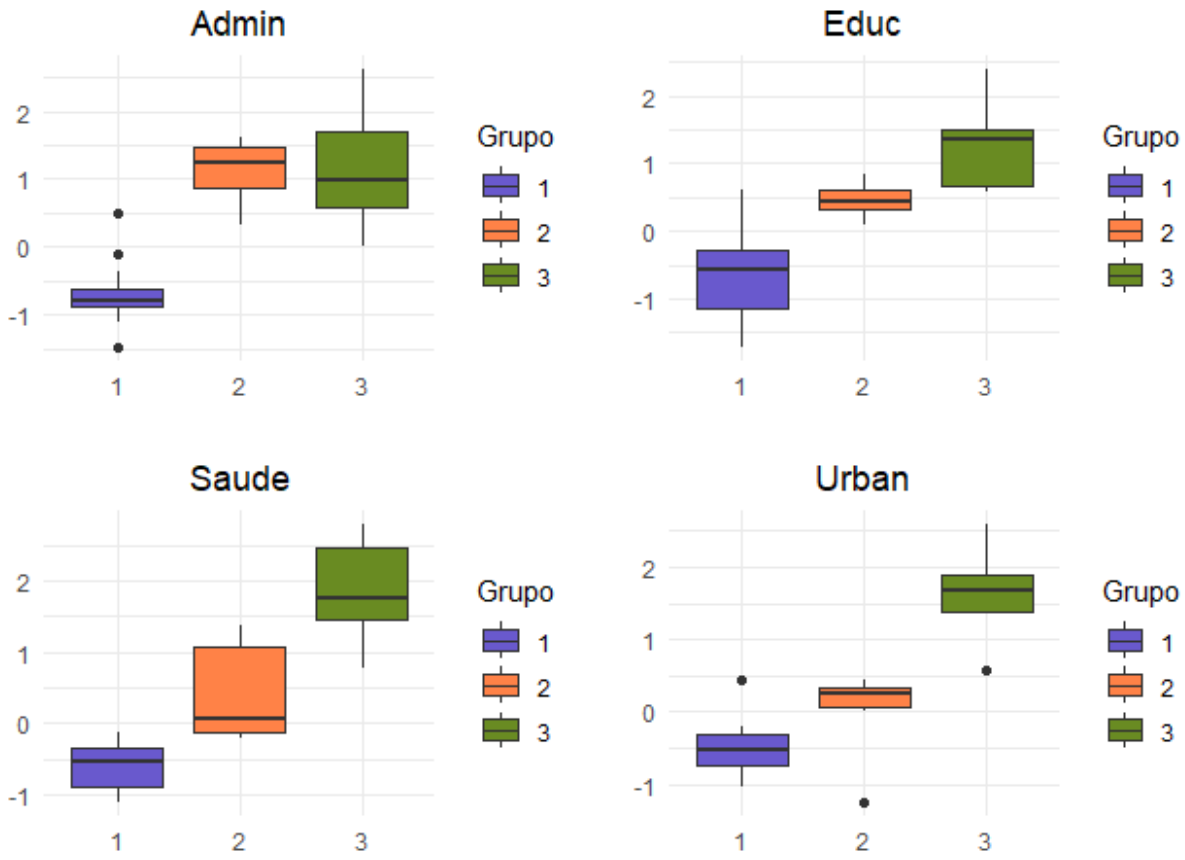

Figura 2.1: Boxplot de cada variável, considerando o agrupamento.

Vamos caracterizar cada um desses grupos com base nas variáveis utilizadas no agrupamento, com o auxilio do boxplot e da tabela com os grupos formados a partir do k-means:

Podemos notar que o **Grupo 1** é composto por 13 municípios, que apresentam gastos relativamente baixos em todas as áreas analisadas. Em particular, eles gastam menos em Administração, Educação, Saúde e Urbanismo do que a maioria dos outros municípios. Portanto, o grupo 1 pode ser caracterizado como sendo de baixo investimento público ou como **"Municípios menos desenvolvidos"**.

Já o **Grupo 2** é composto por seis municípios, que apresentam gastos mais elevados em áreas como Administração, Educação e Saúde, quando comparados com o grupo 1. Além disso, eles gastam menos em Urbanismo em relação a outros grupos. Esse grupo pode ser caracterizado como sendo de médio investimento em serviços públicos e com maior enfoque em áreas sociais como Educação e Saúde ou como **"Municípios em desenvolvimento"**

Por fim, o **Grupo 3** é composto por cinco municípios, que apresentam os maiores gastos em todas as áreas analisadas. Esses municípios são os que mais investem em Administração, Educação, Saúde e Urbanismo em relação aos outros grupos. Esse grupo pode ser caracterizado como sendo de alto investimento público e com foco em todas as áreas ou pode ser caracterizado como **"Municípios desenvolvidos"**, pois todas as variáveis são altas nesse grupo, trazendo indícios que são cidades com maior desenvolvimento.

Também plotamos um gráfico de dispersão e correlação para as variáveis analisadas, e caracterizando os pontos conforme os grupos formados.

Na Figura 2.2, observando a parte inferior do gráfico, notamos que existe um certo padrão de respostas dependendo de cada grupo, para praticamente todas as combinações de variáveis, o que é um forte indicio que as observações estão bem agrupadas.

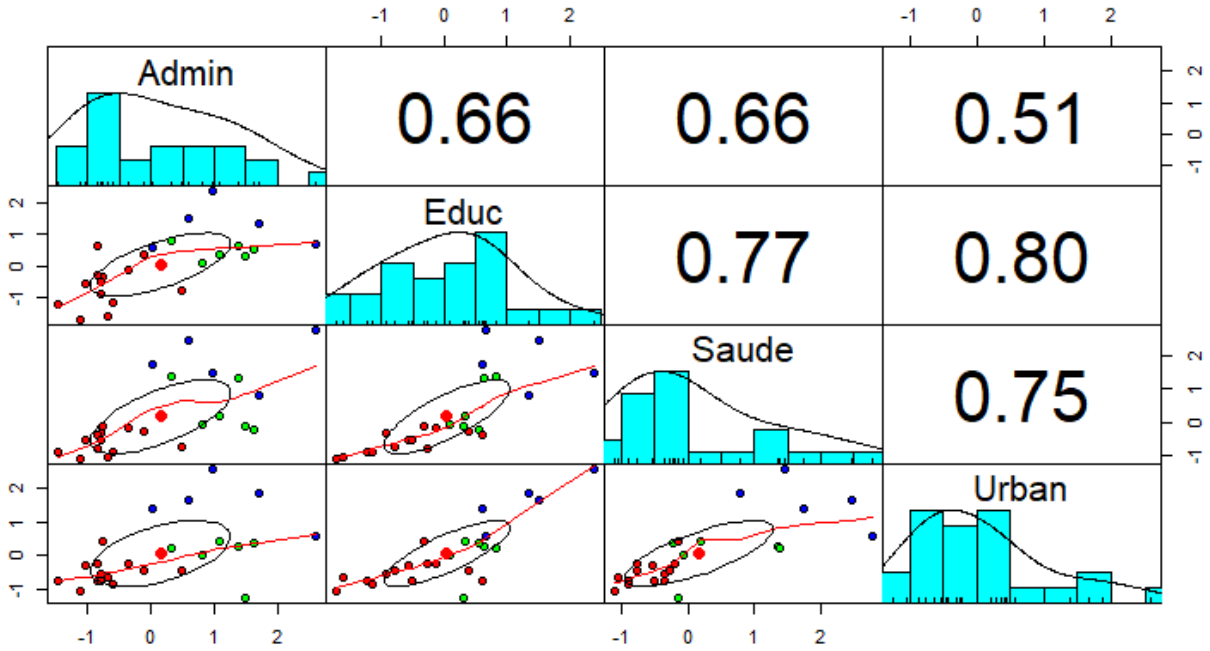

Figura 2.2: Gráfico de dispersão e correlação, pontos em vermelho = Grupo 1; pontos em verde = Grupo 2; e pontos em azul = Grupo 3.

2.2 Análise Discriminante Linear

Nessa seção iremos conduzir uma análise discriminante linear (AD) para classificar os municípios com base nas variáveis "Admin", "Educ", "Saude" e "Urban".

Para realizar a análise, utilizamos o software R, e a função "lda()" do pacote "factoextra".

Iniciando o ajuste, removemos as informações de dois municípios (Aracatuba e Votuporanga), no intuito de realizar as classificações com base nos resultados da AD. Com isso, definimos as probabilidades a prioris como sendo a proporção de observações de cada grupo, as estimativas estão representadas na Tabela 2.1.

Tabela 2.1: Probabilidades a prioris

Grupos	1	2	3
Prioris	0.5454	0.2273	0.2273

Desse modo, representamos na Tabela 2.2 as estimativas dos coeficientes das funções discriminantes obtidas no ajuste,

Na Figura 2.3, representamos as estimativas para cada município e suas respectivas classificações.

Tabela 2.2: Coeficientes das funções discriminantes

	LD1	LD2
Admin	-0.9138	-1.0666
Educ	0.2552	-0.7596
Saude	-1.3050	0.6252
Urban	-1.3604	1.0291

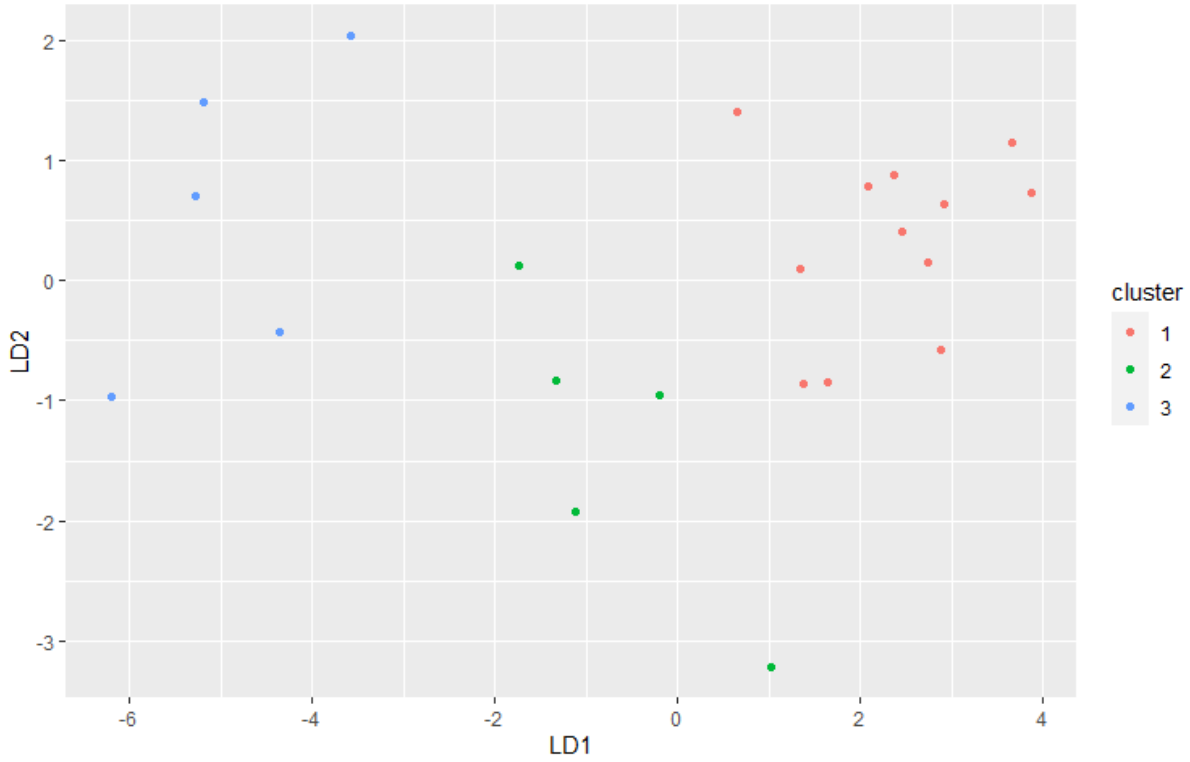

Figura 2.3: .

De modo geral, podemos observar que os pontos apresentam um padrão para cada grupo, tendo uma boa distinção de cada grupo.

Na Tabela 2.3, representamos o desempenho da classificação, comparamos os resultados reais com as observações classificadas pela AD.

Tabela 2.3:

	Cluster		
Classificação	1	2	3
1	12	0	0
2	0	5	0
3	0	0	5

Notamos que a classificação acertou 100% dos resultados, indicio que o ajuste foi bem sucedido, e que os grupos são bem discriminados.

Afim de verificar o desempenho da AD para novos municípios, fizemos a classificação

das cidades Aracatuba e Votuporanga que não foram consideradas no ajuste.

A função "predict()" do R permite realizar novas classificações de forma automática, entretanto, achamos interessante mostrar o processo de classificação. Portanto, calculamos os valores LD1 e LD2, conforme os coeficientes estimados e apresentados na Tabela 2.2, para cada cidade.

$$\begin{aligned} \text{LD1 (Aracatuba)} &= -0.9138.\text{Admin} + 0.2552 * \text{Educ} - 1.3050 * \text{Saude} - 1.3604 * \text{Urban} \\ &= -2.8199 \end{aligned}$$

$$\begin{aligned} \text{LD2 (Aracatuba)} &= -1.0666 * \text{Admin} - 0.7596 * \text{Educ} + 0.6252 * \text{Saude} + 1.0291 * \text{Urban} \\ &= -0.8380 \end{aligned}$$

Com os pontos estimados, podemos adicionar no gráfico representado na Figura 2.3 para tornar a classificação mais clara/visual. Também realizamos o mesmo processo para o município de Votuporanga, obtendo o ponto (3.0303, 0.0622).

Contudo representamos na Figura 2.4, as classificações para os novos municípios, também traçamos as retas que representam os critérios de classificação.

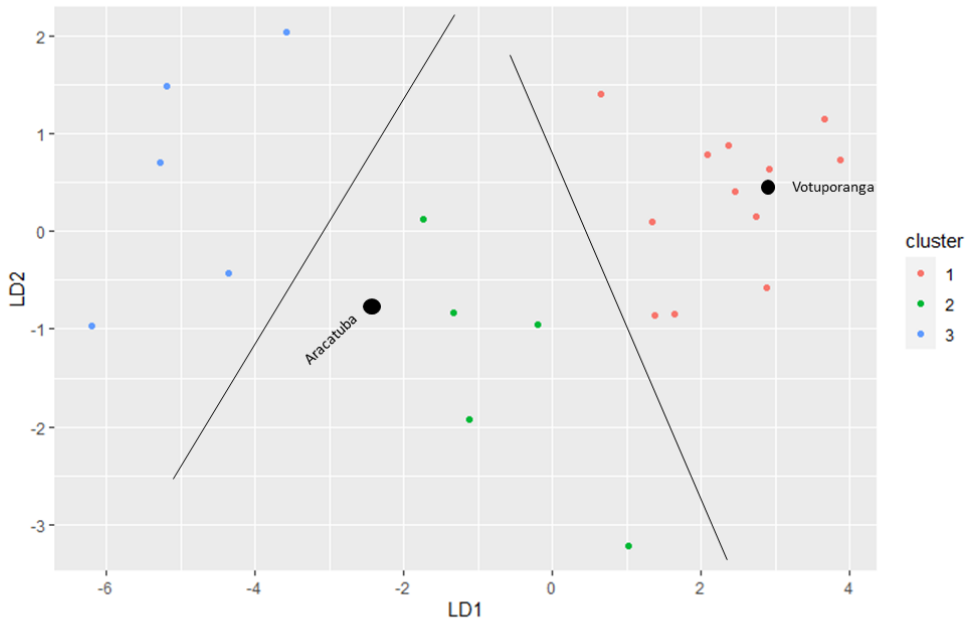

Figura 2.4: Classificações das cidades de Aracatuba e Votuporanga.

Aracatuba foi classificada como grupo 2, o que realmente é apresentada no cluster. Já Votuporanga foi classificada como grupo 1, estando em linha também com o real cluster.