CASE DATA SCIENCE IBOPE

Candidato: Juan CR Soto Sotelo

1. Segmentar os municípios brasileiros para a estratégia de entrada de uma multinacional varejista
2. Determinar os municípios que deveriam ser a porta de entrada para empresa
3. Elaborar um modelo de classificação para o cálculo da probabilidade de um determinado município pertencer a um dos grupos criados.

Dados : Informações Municipais

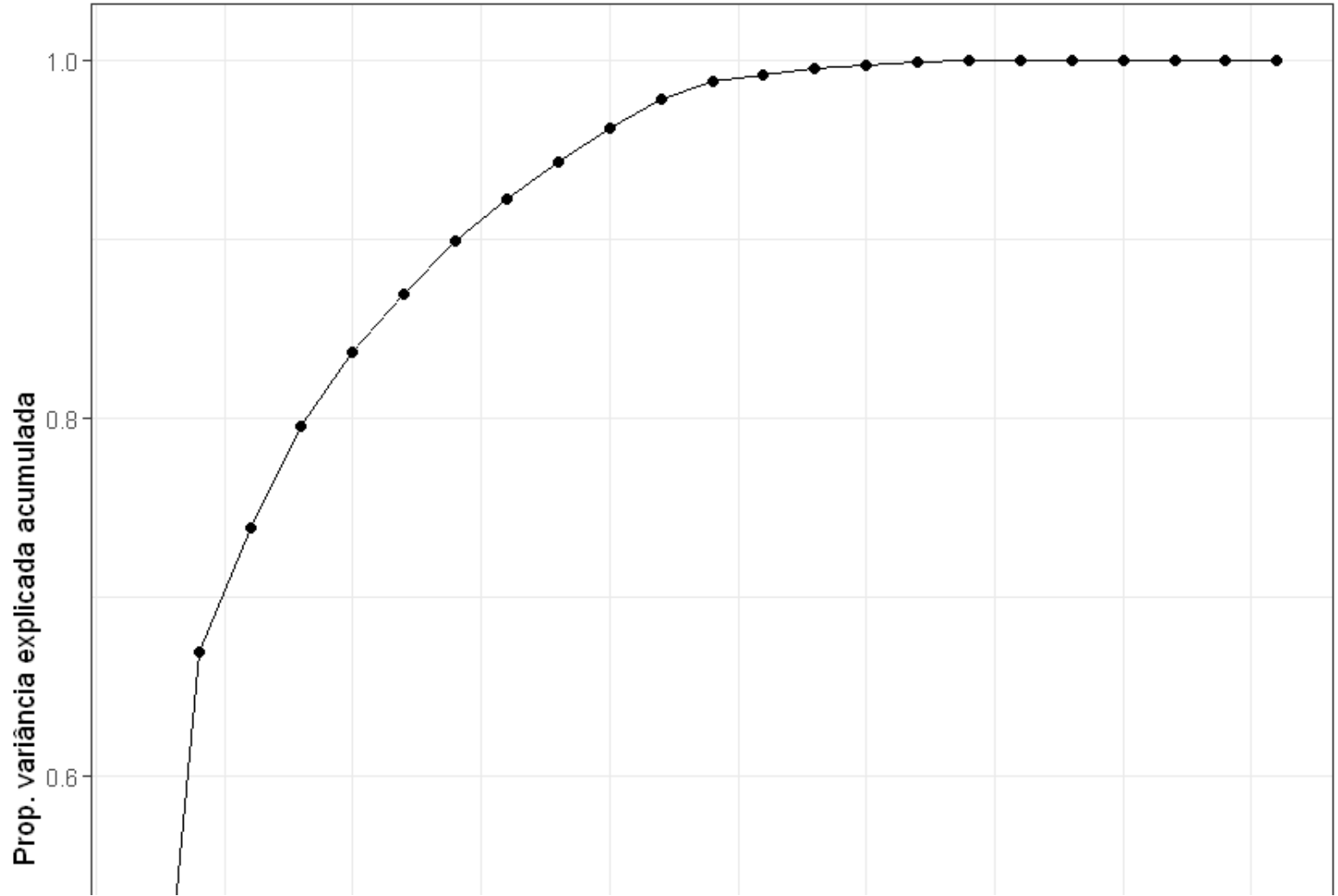
In [6]:
```
#LISTA DAS VARIAVEIS
print(names(Case_Data_Science_IBOPE))
```

```
 [1] "Código"
 [2] "Município"
 [3] "Área (km²)"
 [4] "Densidade demográfica, 2000"
 [5] "Distância à capital (km)"
 [6] "Esperança de vida ao nascer, 2000"
 [7] "Mortalidade até um ano de idade, 2000"
 [8] "Taxa de fecundidade total, 2000"
 [9] "Percentual de pessoas de 25 anos ou mais analfabetas, 2000"
[10] "Renda per Capita, 2000"
[11] "Índice de Gini, 2000"
[12] "Intensidade da indigência, 2000"
[13] "Intensidade da pobreza, 2000"
[14] "Índice de Desenvolvimento Humano Municipal, 2000"
[15] "Taxa bruta de freqüência à escola, 2000"
[16] "Taxa de alfabetização, 2000"
[17] "Média de anos de estudo das pessoas de 25 anos ou mais de idade, 2000"
[18] "População de 25 anos ou mais de idade, 1991"
[19] "População de 25 anos ou mais de idade, 2000"
[20] "População de 65 anos ou mais de idade, 1991"
[21] "População de 65 anos ou mais de idade, 2000"
[22] "População total, 1991"
[23] "População total, 2000"
[24] "População urbana, 2000"
[25] "População rural, 2000"
```

1. 5507 municípios e 25 variáveis demográficas
2. Variáveis com diferentes magnitude e unidades de medida,
3. Nenhum registro vazio (missing).
4. Passo prévio : redução da dimensionalidade -> 9 componentes
5. Clusters : 5 agrupamentos

In [38]:

```
library(ggplot2)
ggplot(data = data.frame(prop_varianza_acum, pc = 1:23),
       aes(x = pc, y = prop_varianza_acum)) +
  geom_point() +
  geom_line() +
  theme_bw() +
  labs(x = "Componente principal",
       y = "Prop. variância explicada acumulada")
```



In [44]:
```
Case_Data_Science_IBOPE2[Case_Data_Science_IBOPE2$segmento==2,c(2,26)]
```

	Município	segmento
103	Altamira (PA)	2
496	Barcelos (AM)	2
3414	Oriximiná (PA)	2
4603	São Félix do Xingu (PA)	2
4627	São Gabriel da Cachoeira (AM)	2
5090	Tapauá (AM)	2

In [45]:
```
Case_Data_Science_IBOPE2[Case_Data_Science_IBOPE2$segmento==3,c(2,26)]
```

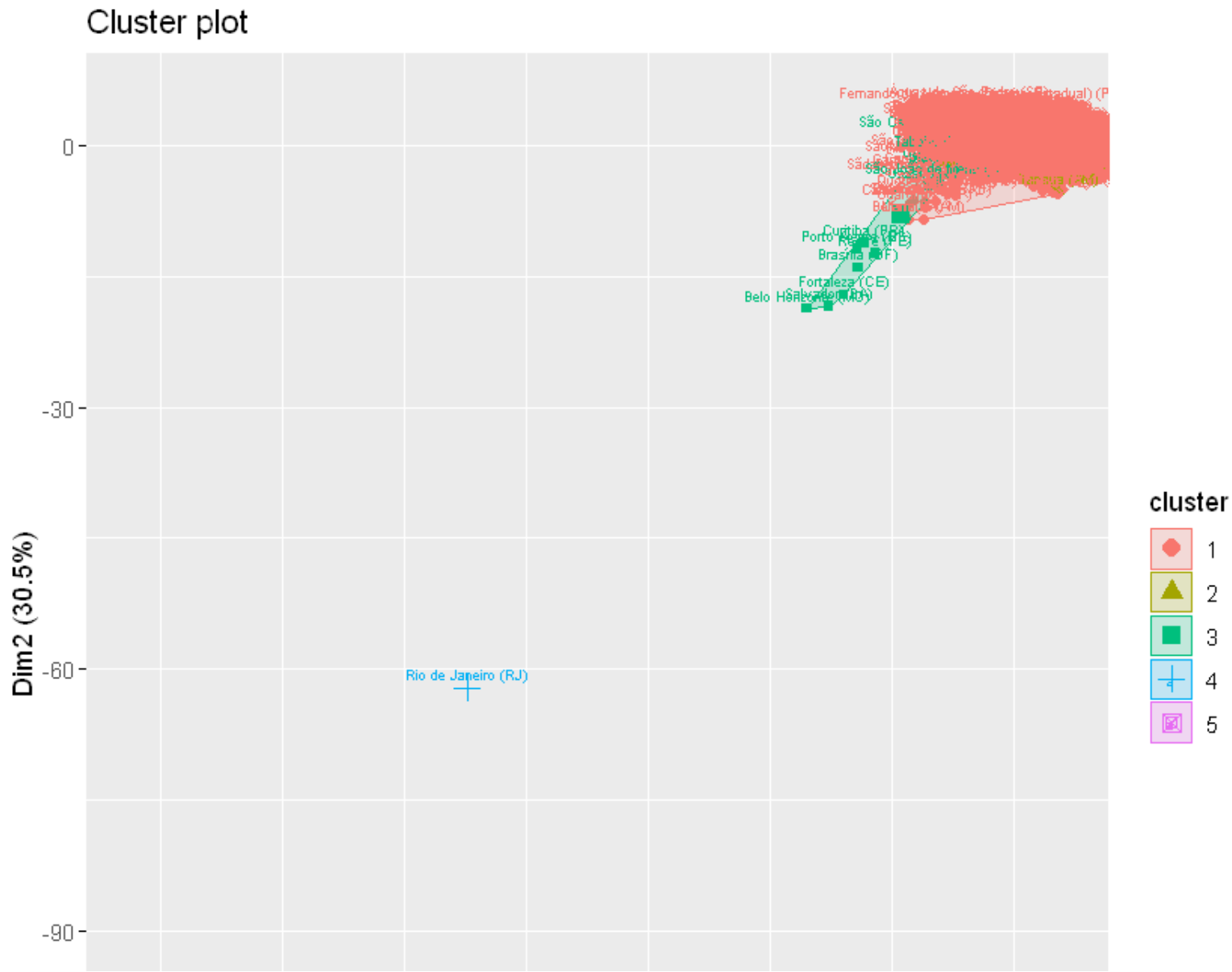
	Município	segmento
585	Belo Horizonte (MG)	3
743	Brasília (DF)	3
1075	Carapicuíba (SP)	3
1488	Curitiba (PR)	3
1530	Diadema (SP)	3
1808	Fortaleza (CE)	3
3226	Nilópolis (RJ)	3
3398	Olinda (PE)	3
3423	Osasco (SP)	3
3897	Porto Alegre (RS)	3
4060	Recife (PE)	3
4280	Salvador (BA)	3
4571	São Caetano do Sul (SP)	3
4674	São João de Meriti (RJ)	3
5053	Taboão da Serra (SP)	3

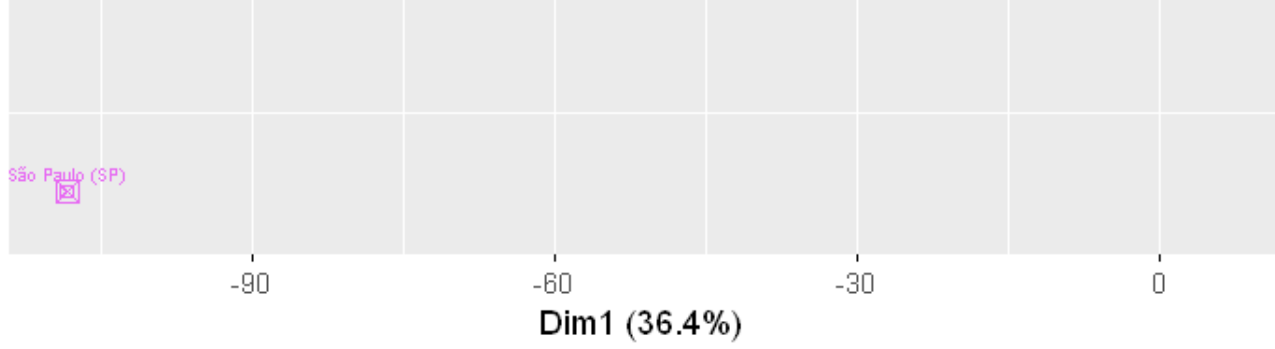
In [49]:
```
Case_Data_Science_IBOPE2[Case_Data_Science_IBOPE2$segmento==4,c(2,26)]
Case_Data_Science_IBOPE2[Case_Data_Science_IBOPE2$segmento==5,c(2,26)]
```

	Município	segmento
Rio de Janeiro (RJ)	Rio de Janeiro (RJ)	4

	Município	segmento
São Paulo (SP)	São Paulo (SP)	5

In [50]:
```
row.names(Case_Data_Science_IBOPE2) <- Case_Data_Science_IBOPE2$Município
fviz_cluster(list(data = Case_Data_Science_IBOPE2[, -c(1:2,26)], cluster = segmento),labelsize = 6)
```




São Paulo (SP)
-90
-60
-30
0
Dim1 (36.4%)

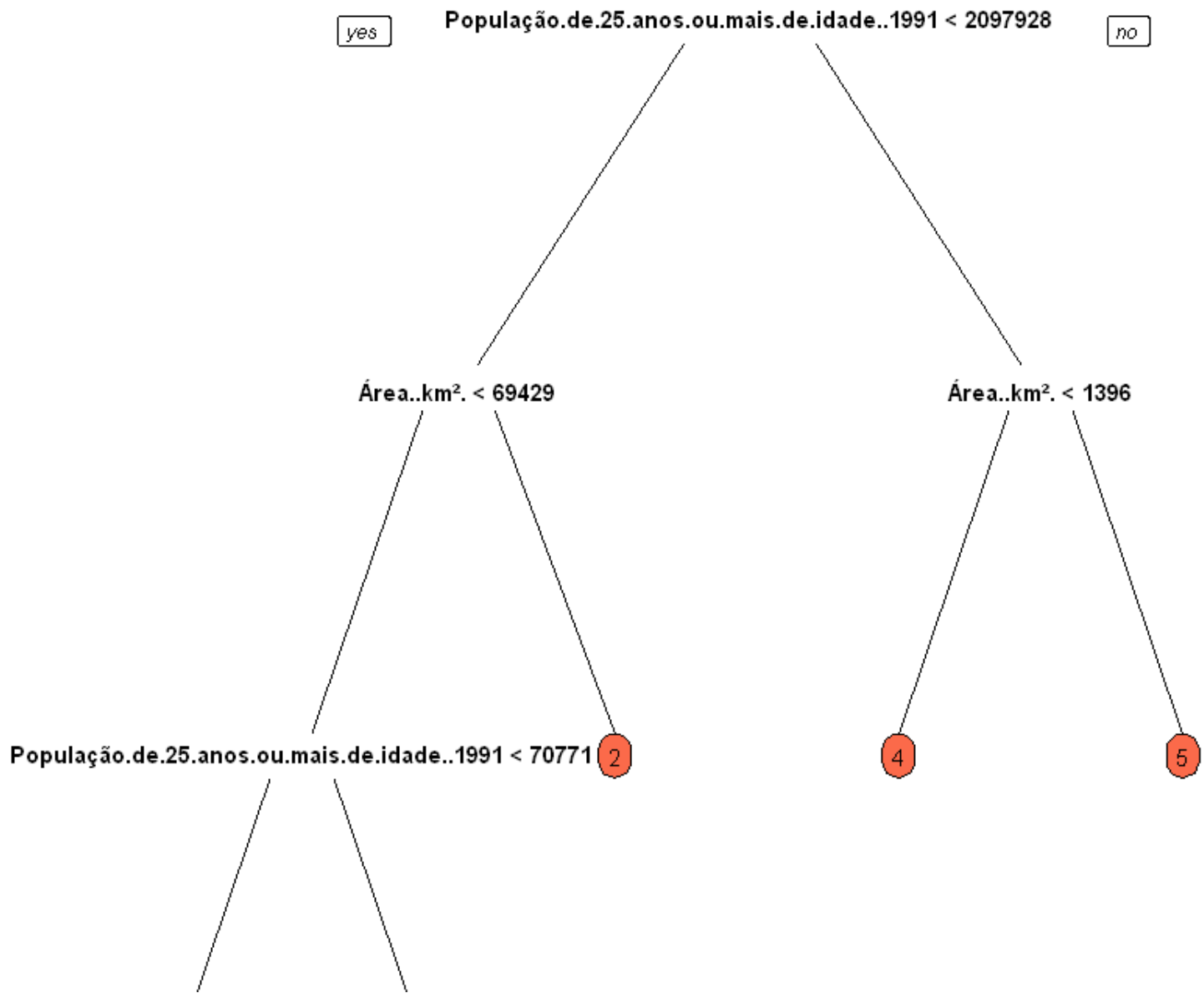
Comentário

Os municípios dos grupos 3,4 e 5 devem ser a porta de entrada. por serem os grupos melhor definidos e com melhores indicadores demográficos

Classificação

- Método usado : Árvore de decisão
- Em geral a árvore de decisão consegue reproduzir a segmentação dos municipios

In [53]:
```
prp(dtree_fit$finalModel,cex = 0.75,box.palette = "Reds",digits=-3, varlen=-50)
```




1
3

In [54]:
```
#Classificação na base inteira
test_pred <- predict(dtree_fit, newdata = Case_Data_Science_IBOPE2[, -c(1,2)])
confusionMatrix(test_pred, Case_Data_Science_IBOPE2$segmento )  #check acuracidade
```

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4    5
         1 5384    0    0    0    0
         2    3    6    0    0    0
         3   97    0   15    0    0
         4    0    0    0    1    0
         5    0    0    0    0    1

Overall Statistics

               Accuracy : 0.9818
                 95% CI : (0.978, 0.9852)
    No Information Rate : 0.9958
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3112

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3  Class: 4  Class: 5
Sensitivity            0.9818 1.000000 1.000000 1.0000000 1.0000000
Specificity            1.0000 0.999455 0.982338 1.0000000 1.0000000
Pos Pred Value         1.0000 0.666667 0.133929 1.0000000 1.0000000
Neg Pred Value         0.1870 1.000000 1.000000 1.0000000 1.0000000
Prevalence             0.9958 0.001090 0.002724 0.0001816 0.0001816
Detection Rate         0.9777 0.001090 0.002724 0.0001816 0.0001816
Detection Prevalence   0.9777 0.001634 0.020338 0.0001816 0.0001816
Balanced Accuracy      0.9909 0.999727 0.991169 1.0000000 1.0000000
```

In [58]:
```
!jupyter nbconvert Juan_Apresentacao_IBOPE.ipynb --to slides --post serve --template out
put_toggle
```

```
Error in parse(text = x, srcfile = src): <text>:1:10: unexpected symbol
1: !jupyter nbconvert
             ^
Traceback:
```