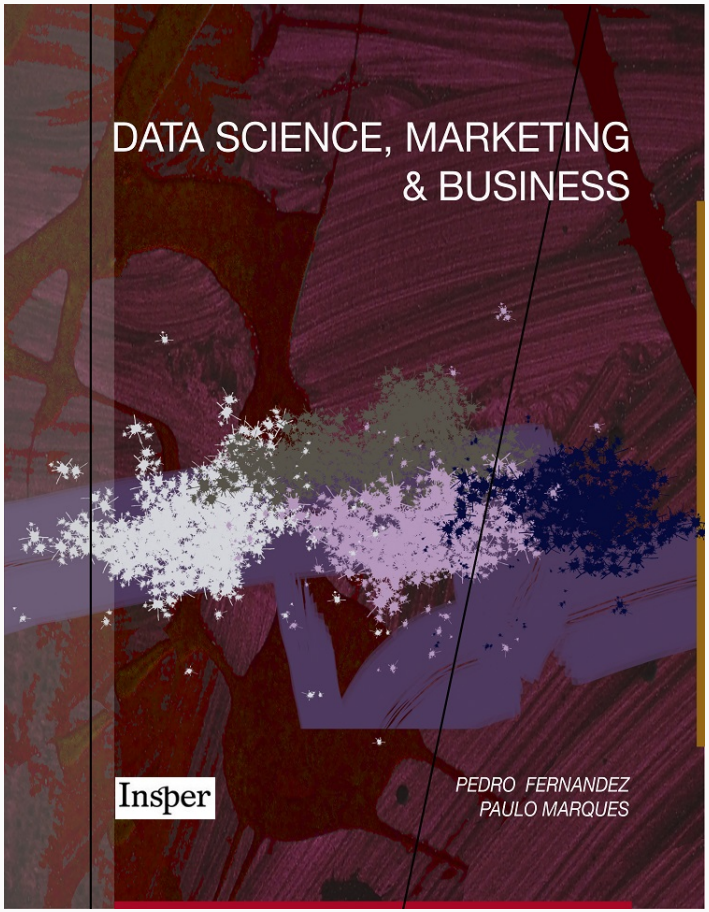
Uma Aplicação de MDS ao Mercado Automotivo

Tiago Mendonça
Insper
tiagoms.com



O livro pode ser acessado em datascience.insper.edu.br.

Um pouco da história do professor Pedro Fernandez e do livro pode ser acessada nesse link.

Tive o prazer de colaborar, em conjunto com os professores Paulo Marques e Hedibert F. Lopes, na produção do livro do professor Pedro Fernandez.

Introdução

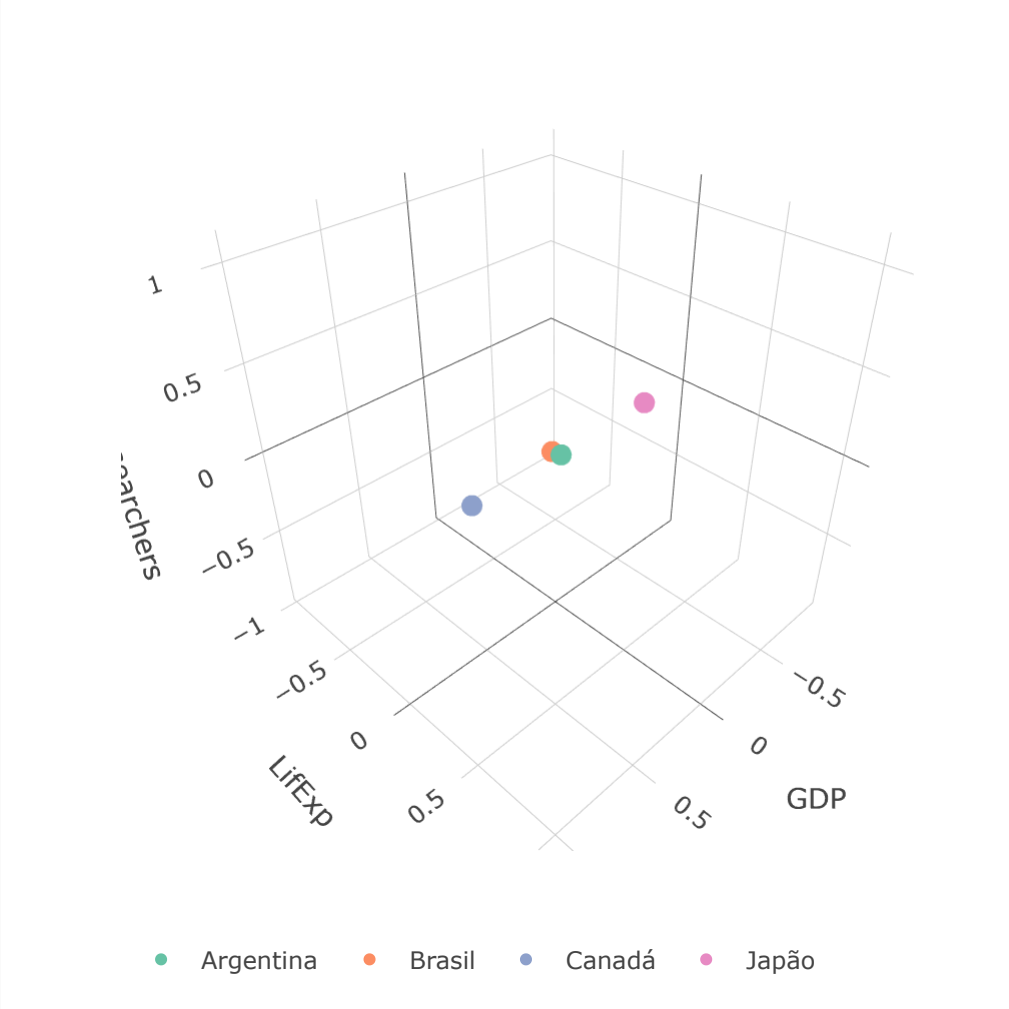
Motivação

Dados do The World Bank

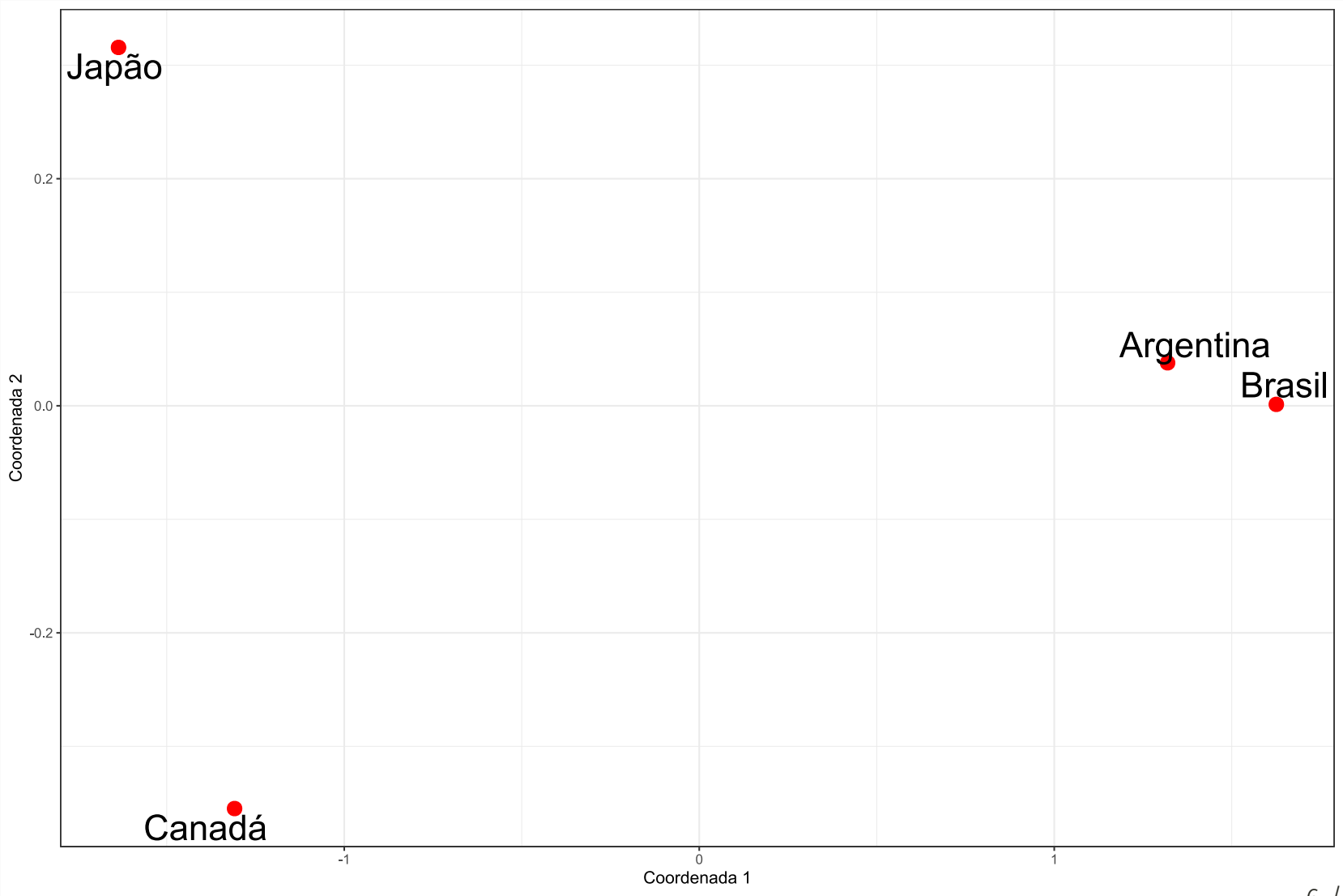
Pais	GDP	LifExp	Researchers
Argentina	11652.6	76	1233
Brasil	8920.8	75	881
Canadá	46210.5	82	4275
Japão	39286.7	84	5305

GDP per capita (current US$) / Life expectancy at birth, total (years) / Researchers in R&D (per million people)

Motivação


1
0.5
0
−0.5
−1
−0.5
0
0.5
LifExp
−0.5
0
0.5
GDP
Argentina
Brasil
Canadá
Japão

Motivação


Japão
Argentina
Brasil
Canadá
0.2
0.0
-0.2
Coordenada 2
-1
0
1
Coordenada 1

Motivação

Distância dos pontos em $\mathbb{R}^3$

	Argentina	**Brasil**	**Canadá**	**Japão**
Argentina	0.000	0.312	2.657	2.968
Brasil	0.312	0.000	2.956	3.277
Canadá	2.657	2.956	0.000	0.746
Japão	2.968	3.277	0.746	0.000

Distäncia dos pontos pelo **MDS** (em $\mathbb{R}^2$)

	Argentina	**Brasil**	**Canadá**	**Japão**
Argentina	0.000	0.308	2.657	2.968
Brasil	0.308	0.000	2.956	3.276
Canadá	2.657	2.956	0.000	0.746
Japão	2.968	3.276	0.746	0.000

Resumindo

- Queremos representar observações de alta dimensão num número reduzido de dimensões.
- Podemos fazer isso com base nos dados brutos ou diretamente de uma matriz de dissimilaridades.

Aplicação ao Mercado Automotivo

Introdução

Consumidores (n = 150) recebem 30 papeletas com modelos de carros e agrupam os carros considerados equivalentes (substituíveis no momento da compra).

Este processo define uma medida de dissimilaridade por:

$$d_{ij} = \frac{n - \text{quantas vezes } i \text{ e } j \text{ foram alocados no mesmo grupo}}{n}$$

Assim,

- $d_{ij} = 0$ quando os modelos i e j são alocados dentro do mesmo grupo por todos os consumidores,
- $d_{ij} = 1$ se i e j não forem alocados dentro do mesmo grupo por nenhum consumidor.

Exemplo

id	Ka	Uno	Gol	Fox	Fit
Cliente 1	2	2	1	1	
Cliente 2	2	2	1	1	1
Cliente 3			1	1	
Cliente 4			1	1	1
Cliente 5	2	2	1	1	

$$d_{ij} = \frac{n - \text{quantas vezes } i \text{ e } j \text{ foram alocados no mesmo grupo}}{n}$$

$$d_{\text{Fox,Gol}} = \frac{5-5}{5} = 0$$

$$d_{\text{Fit,Gol}} = \frac{5-2}{5} = 0.6$$

$$d_{\text{Fit,Uno}} = \frac{5-0}{5} = 1$$

Leitura dos dados

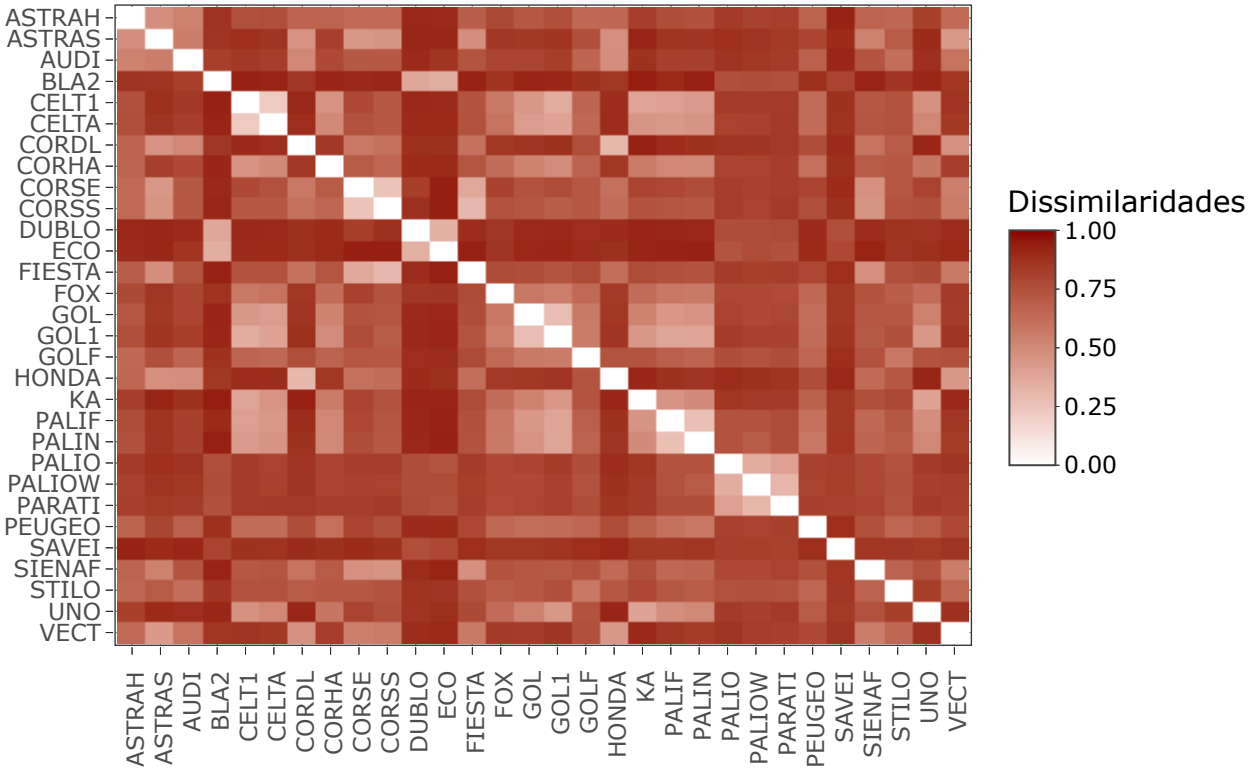
```
dissim ← read.csv("matriz_dissimilaridades.csv", row.names = 1)

dissim %>%
  round(2) %>%
  head() %>%
  kable(format = 'html', align = c(rep("c", ncol(dissim))))
```

	UNO	KA	VECT	CELT1	GOL	GOLF	CORSS	GOL1	ECO	BLA2	PALIN	SAVEI	HONDA	FOX
UNO	0.00	0.39	0.88	0.46	0.53	0.74	0.75	0.44	0.88	0.91	0.51	0.84	0.92	0.63
KA	0.39	0.00	0.90	0.38	0.53	0.73	0.74	0.44	0.92	0.94	0.49	0.85	0.91	0.63
VECT	0.88	0.90	0.00	0.87	0.83	0.75	0.56	0.86	0.90	0.85	0.83	0.86	0.44	0.84
CELT1	0.46	0.38	0.87	0.00	0.44	0.66	0.71	0.35	0.90	0.93	0.42	0.87	0.89	0.57
GOL	0.53	0.53	0.83	0.44	0.00	0.57	0.73	0.28	0.91	0.91	0.46	0.85	0.85	0.58
GOLF	0.74	0.73	0.75	0.66	0.57	0.00	0.71	0.56	0.89	0.87	0.66	0.89	0.73	0.64

Matriz de Dissimilaridades

```
ggplotly(
  dissim %>%
    mutate("mod1" = rownames(dissim)) %>%
    pivot_longer(-mod1, names_to = "mod2", values_to = "dissim") %>%
    ggplot(aes(mod1, mod2, fill = dissim)) +
      geom_tile() + xlab("") + ylab("") +
      scale_fill_gradient(low = "white", high = "red4",
                          limit = c(0,1), name = "Dissimilaridades") +
      theme(axis.text.x = element_text(angle = 90))
)
```

Matriz de Dissimilaridades

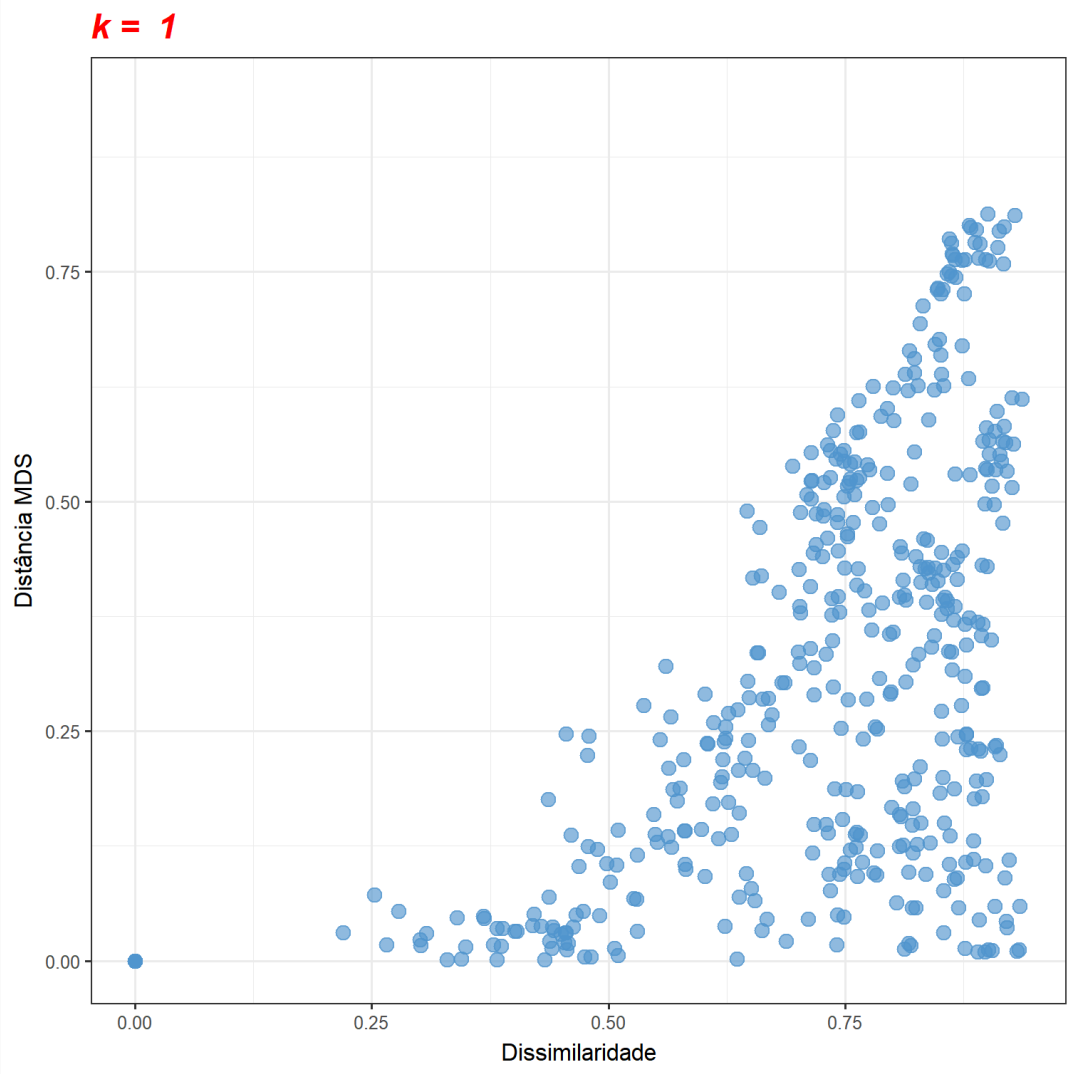


MDS

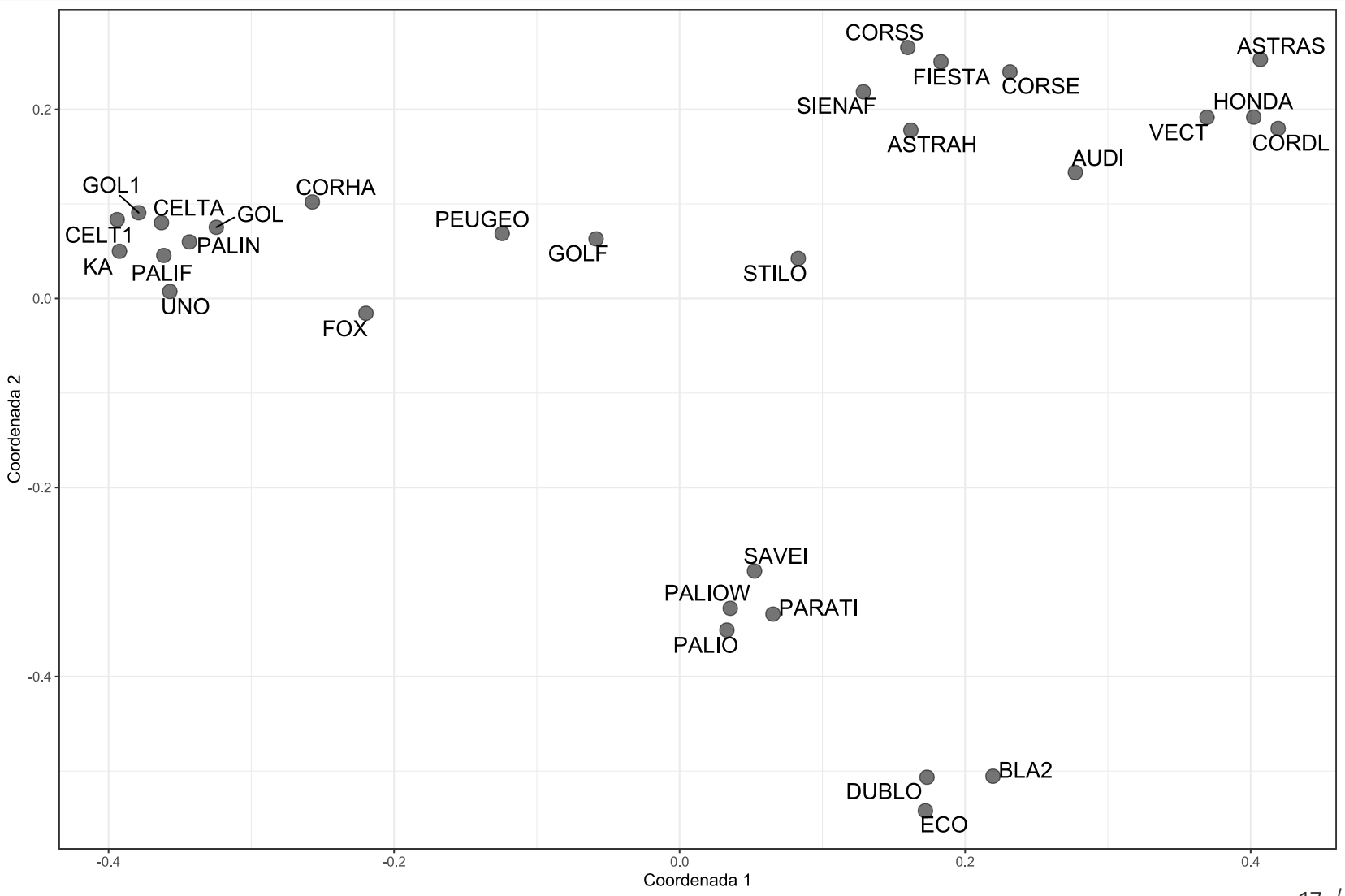
```
cars_mds ← cmdscale(dissim, k = 2, eig = TRUE)

head(cars_mds$points)
```

```
##              [,1]        [,2]
## UNO   -0.35708628 0.007426957
## KA    -0.39238141 0.049970533
## VECT   0.36941030 0.191605507
## CELT1 -0.39393623 0.083466421
## GOL   -0.32451871 0.075342140
## GOLF  -0.05852281 0.063068691
```

Simulação


k = 1
Distância MDS
0.00
0.25
0.50
0.75
Dissimilaridade
0.00
0.25
0.50
0.75

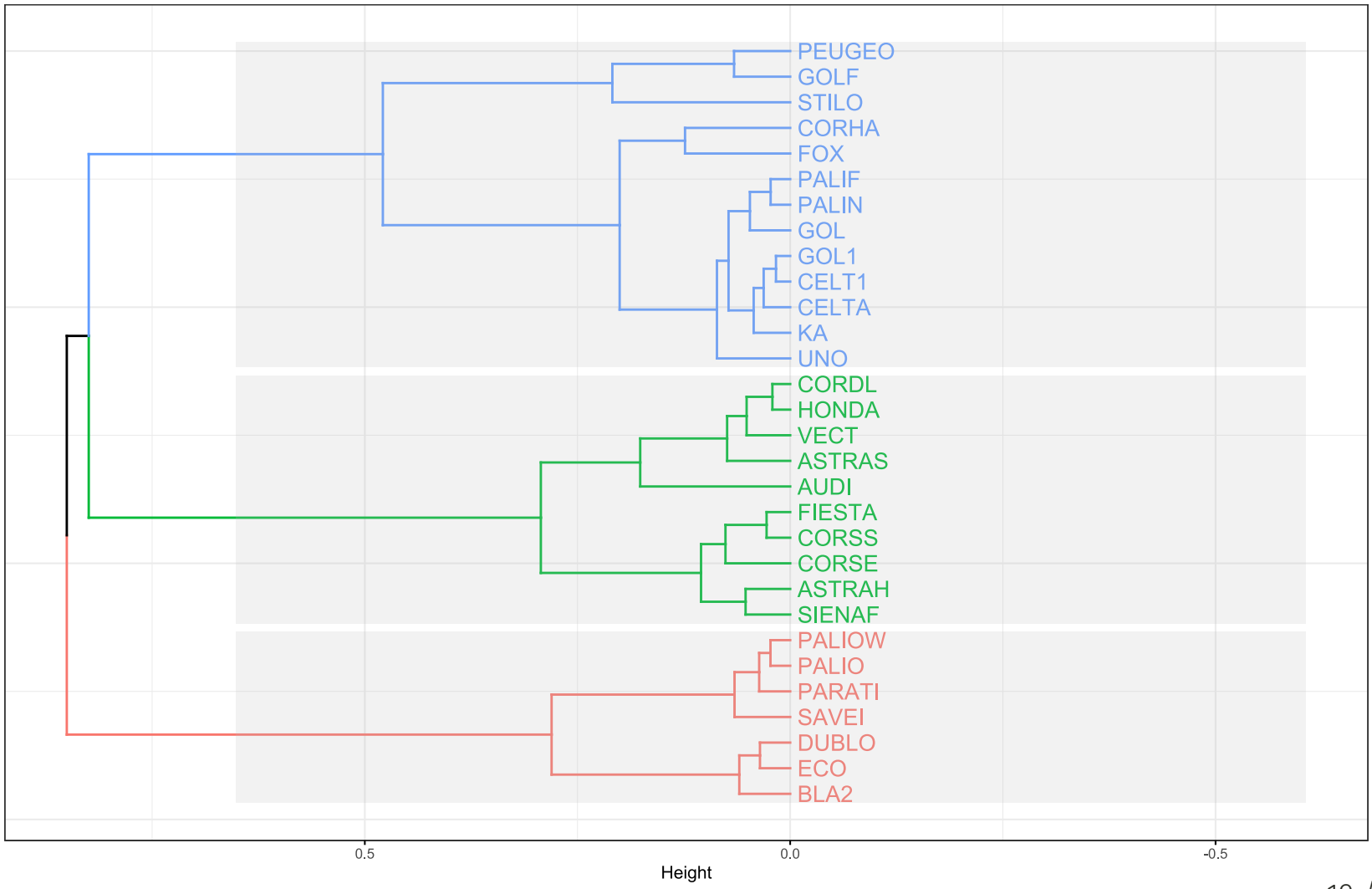
Projeção


CORSS
ASTRAS
FIESTA
CORSE
SIENAF
HONDA
VECT
CORDL
ASTRAH
AUDI
GOL1
CORHA
CELTA
GOL
PEUGEO
CELT1
PALIN
GOLF
KA
PALIF
STILO
UNO
FOX
SAVEI
PALIOW
PARATI
PALIO
BLA2
DUBLO
ECO
0.2
0.0
-0.2
-0.4
Coordenada 2
-0.4
-0.2
0.0
0.2
0.4
Coordenada 1

Cluster

```
carsHC ← hclust(dist(cars_mds$points))

library(factoextra)

fviz_dend(carsHC,
          k = 3,
          cex = 1,
          horiz = TRUE,
          rect = TRUE, rect_fill = TRUE,
          color_labels_by_k = TRUE,
          main = "",
          ggtheme = theme_bw(),
          )
```

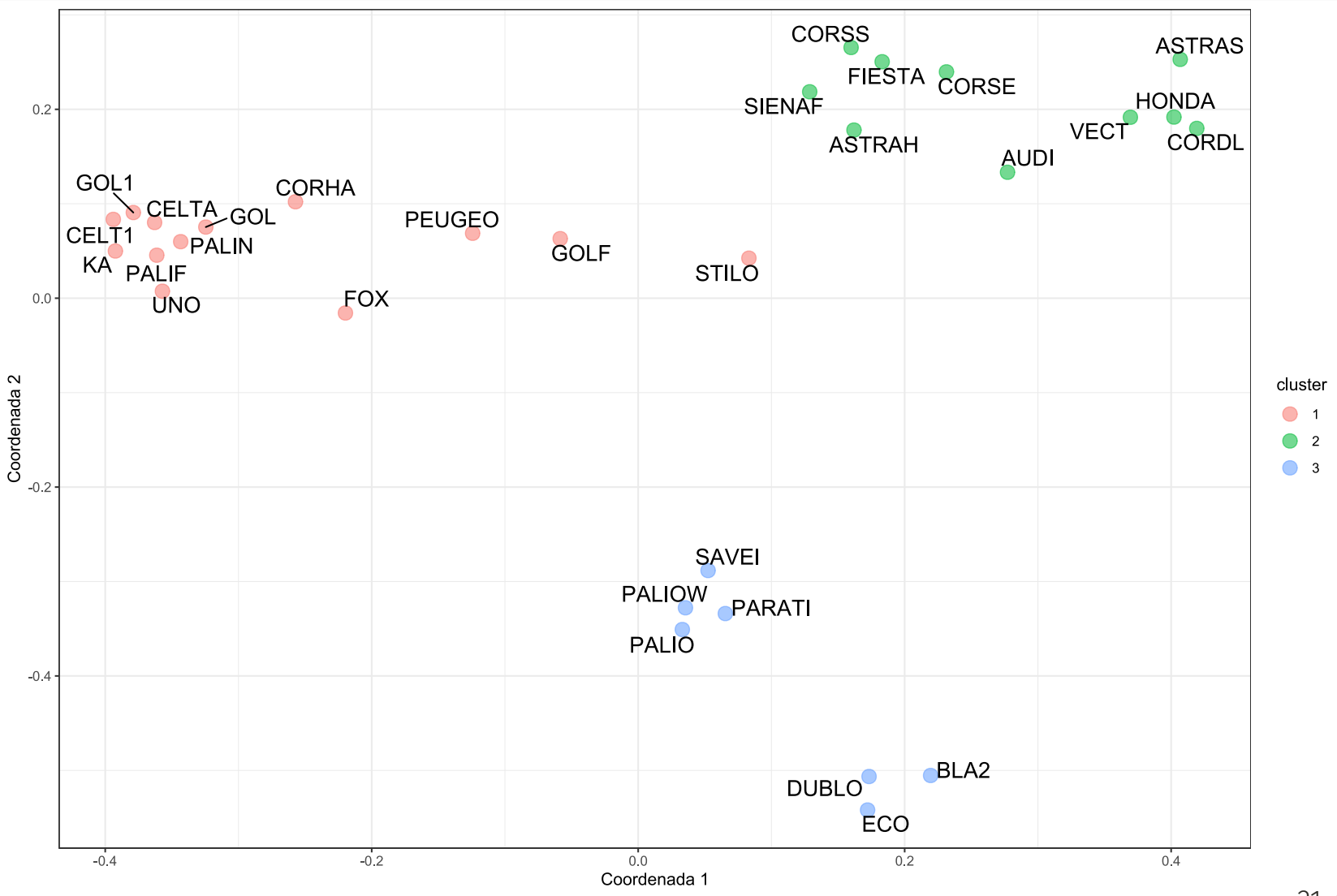
Cluster


PEUGEO
GOLF
STILO
CORHA
FOX
PALIF
PALIN
GOL
GOL1
CELT1
CELTA
KA
UNO
CORDL
HONDA
VECT
ASTRAS
AUDI
FIESTA
CORSS
CORSE
ASTRAH
SIENAF
PALIOW
PALIO
PARATI
SAVEI
DUBLO
ECO
BLA2
0.5
0.0
-0.5
Height

Cluster

```r
library(ggrepel)

clusters ← cutree(carsHC, k = 3)

df ← data.frame(mod = rownames(cars_mds$points),
                coord1 = cars_mds$points[,1],
                coord2 = cars_mds$points[,2],
                cluster = factor(clusters))

ggplot(df, aes(x = coord1, y = coord2, label = rownames(df))) +
  geom_point(aes(color = cluster), alpha = 0.55, size = 4) +
  xlab("Coordenada 1") + ylab("Coordenada 2") +
  geom_text_repel(size = 5) +
  theme_bw()
```

Cluster


CORSS
ASTRAS
FIESTA
CORSE
SIENAF
HONDA
VECT
CORDL
ASTRAH
AUDI
GOL1
CORHA
CELTA
GOL
PEUGEO
CELT1
PALIN
GOLF
KA
PALIF
STILO
UNO
FOX
SAVEI
PALIOW
PARATI
PALIO
BLA2
DUBLO
ECO
0.2
0.0
-0.2
-0.4
Coordenada 2
-0.4
-0.2
0.0
0.2
0.4
Coordenada 1
cluster
1
2
3

Aplicação

Os clusters formados podem ser utilizados de diversas maneiras:

- carros dentro de um mesmo cluster indicam **concorrentes diretos**, e esta informação pode ser utilizada por um dos players do mercado no monitoramento das atividades das empresas concorrentes.
- **sistemas de recomendação** simples também podem ser construídos. Por exemplo, usuários que navegaram na Internet fazendo buscas pelo modelo “Saveiro”, poderiam ser expostos a anúncios de modelos do mesmo cluster (“Palio”, “Parati” etc).

Mercado de Scotch Whisky

Mercado de Whisky

```
library(bayesm)

data(Scotch)

Scotch %>%
  kable(format = "html", align = c(rep("c", 21)))
```

Chivas.Regal	Dewar.s.White.Label	Johnnie.Walker.Black.Label	J...B	Johnnie.Walker.Red.Label
1	0	0	0	1
0	0	1	0	0
0	0	0	0	0
1	0	1	0	1
1	0	1	0	0
0	0	0	0	0
0	0	0	0	0
0	1	0	1	1

Mercado de Whisky

Quando cada atributo assume apenas os valores 0 e 1, podemos definir a **distância de Jaccard** da seguinte forma:

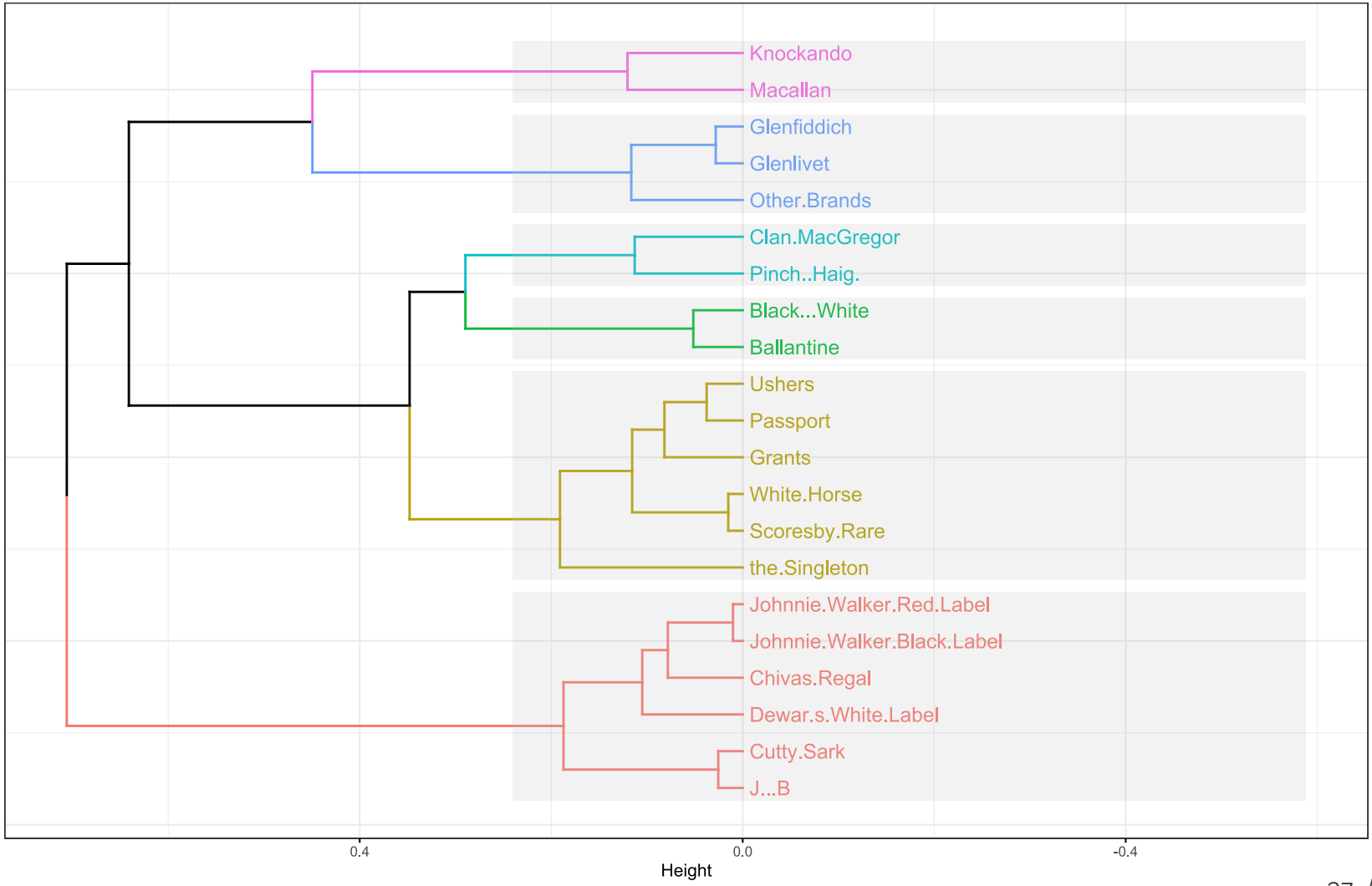
$$d_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{\sum_{m=1}^{p} x_{Am} x_{Bm}}{\sum_{m=1}^{p} x_{Am} + \sum_{m=1}^{p} x_{Bm} - \sum_{m=1}^{p} x_{Am} x_{Bm}}.$$

```
jaccard ← matrix(rep(0, ncol(Scotch)^2), ncol = ncol(Scotch))

dimnames(jaccard) ← list(colnames(Scotch), colnames(Scotch))

for (i in 1:(ncol(Scotch) - 1)){
  for (j in (i + 1):ncol(Scotch)){
    jaccard[i, j] ← 1 - sum(Scotch[,i]*Scotch[,j]) /
                    (sum(Scotch[,i]) + sum(Scotch[,j]) - sum(Scotch[,i]*Scotch[,j]))
    jaccard[j, i] ← jaccard[i, j]
  }
}
```

Mercado de Whisky

```
mds ← cmdscale(jaccard, eig = TRUE)

hc ← hclust(dist(mds$points))

library(factoextra)

fviz_dend(hc,
          k = 6,
          cex = 0.9,
          horiz = TRUE,
          rect = TRUE, rect_fill = TRUE,
          color_labels_by_k = TRUE,
          main = "",
          ggtheme = theme_bw()
          )
```

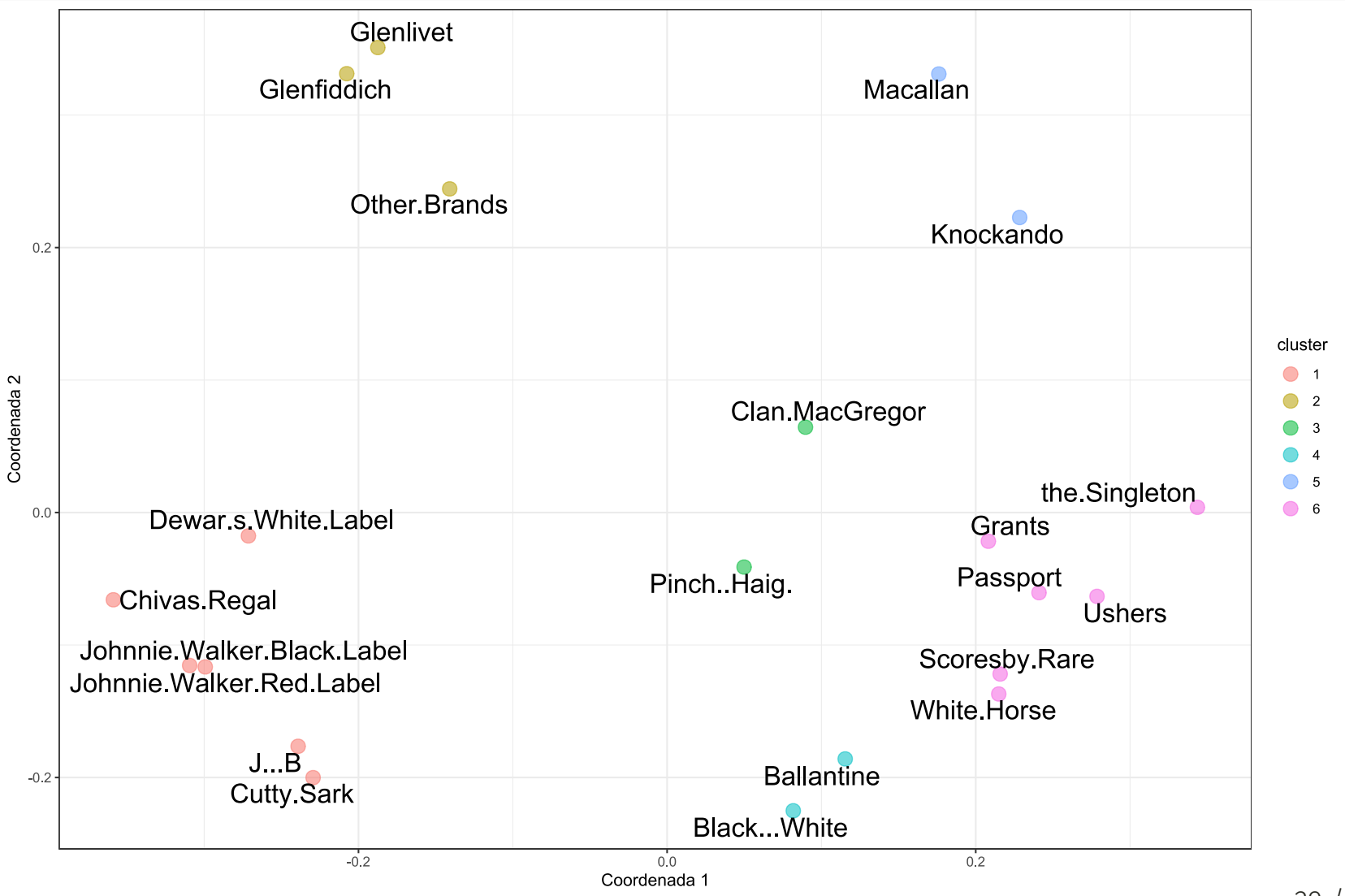
Mercado de Whisky



Mercado de Whisky

```
clusters ← cutree(hc, k = 6)

df ← data.frame(wishky = rownames(mds$points),
                coord1 = mds$points[,1],
                coord2 = mds$points[,2],
                cluster = factor(clusters))

ggplot(df, aes(x = coord1, y = coord2, label = rownames(df))) +
    geom_point(aes(color = cluster), alpha = 0.55, size = 4) +
    xlab("Coordenada 1") + ylab("Coordenada 2") +
    geom_text_repel() + theme_bw()
```

Mercado de Whisky



Mercado de Whisky

Os clusters obtidos revelam a estrutura de concorrência no mercado de Scotch whisky que pode ser utilizara para:

- formação de estoques e pedidos de compra
- organização de prateleiras em pontos de venda
- sistemas de recomendação.

Obrigado!

Mais exemplos podem ser encontrados no livro Data Science, Marketing & Business

Contato e redes sociais em **tiagoms.com**

Apresentação criada no pacote xaringan.

Um agradecimento especial aos **organizadores** e **patrocinadores** do evento!