Lista 2 - MAE0330

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

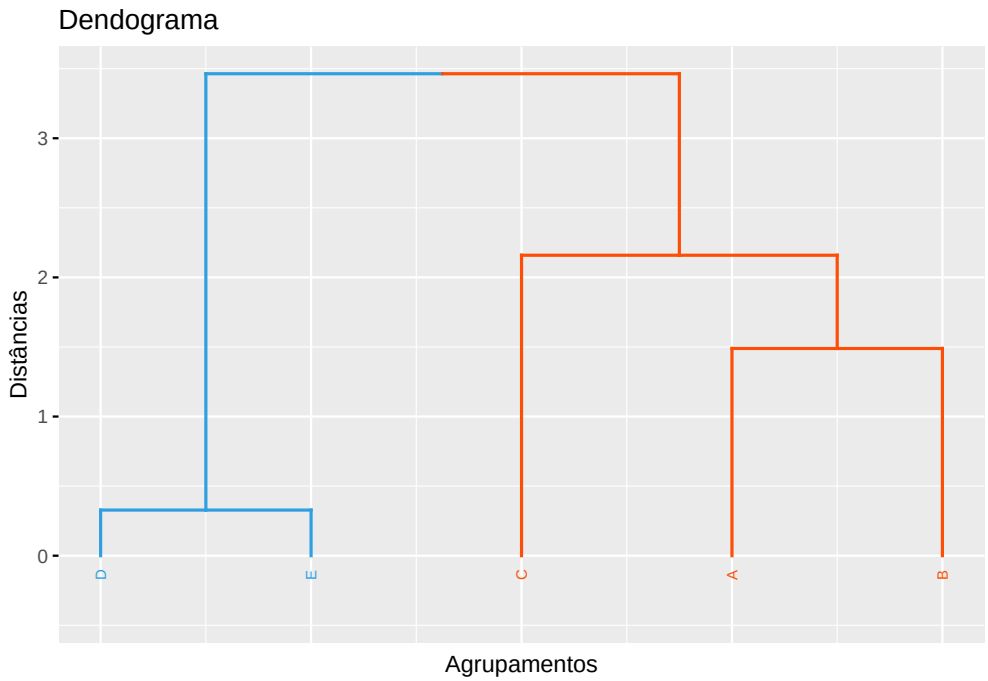
Exercício 2

Considere os dados de cinco unidades amostrais apresentados na tabela a seguir:

Item	X_1	X_2
A	2	0
B	5	2
C	1	4
D	8	4
E	7	4

(b) Refaça o agrupamento utilizando o método de Ward.

Resolução



Assim podemos notar que as unidades amostrais D e E formam um grupo e A, B e C formam outro grupo.

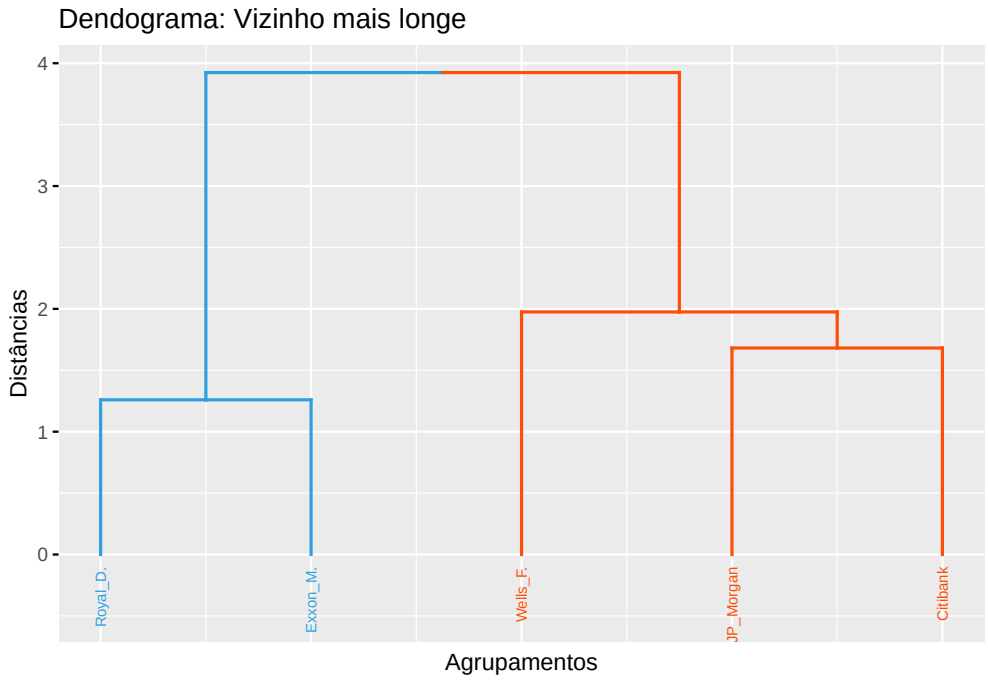
Exercício 3

A matriz abaixo corresponde a matriz de correlação entre as ações de 5 empresas:

$$\begin{array}{ccccccc} & \text{JP Morgan} & \text{Citibank} & \text{Wells Fargo} & \text{Royal DutchShell} & \text{Exxon Mobil} & \\ \begin{pmatrix} \\ \\ \\ \\ \\ \end{pmatrix} & \begin{matrix} 1 \\ 0,63 \\ 0,51 \\ 0,12 \\ 0,16 \end{matrix} & \begin{matrix} \\ 1 \\ 0,57 \\ 0,32 \\ 0,21 \end{matrix} & \begin{matrix} \\ \\ 1 \\ 0,18 \\ 0,15 \end{matrix} & \begin{matrix} \\ \\ \\ 1 \\ 0,68 \end{matrix} & \begin{matrix} \\ \\ \\ \\ 1 \end{matrix} & \begin{pmatrix} \\ \\ \\ \\ \\ \end{pmatrix} \end{array} \begin{matrix} \\ \text{JP Morgan} \\ \text{Citibank} \\ \text{Wells Fargo} \\ \text{Royal DutchShell} \\ \text{Exxon Mobil} \end{matrix}$$

Considerando as correlações amostrais como medidas de similaridade, agrupe as empresas utilizando o método de agrupamento hierárquico do vizinho mais longe. Construa o correspondente dendrograma e discuta os resultados.

Resolução

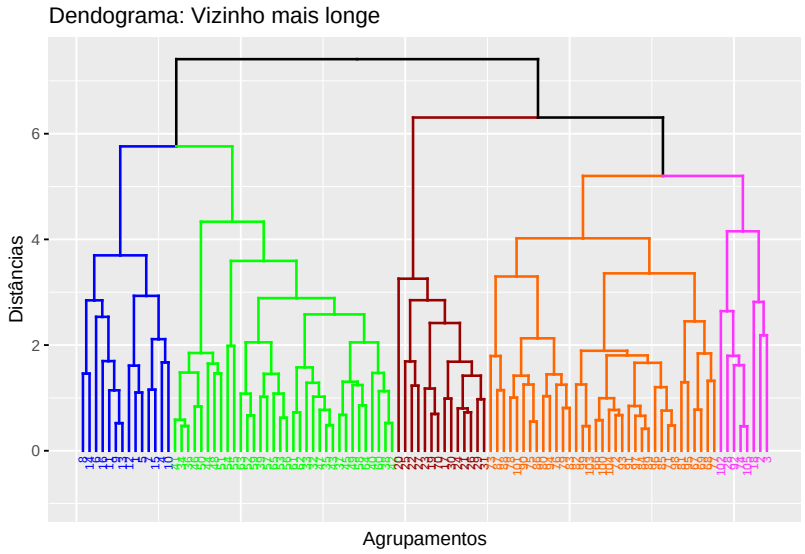


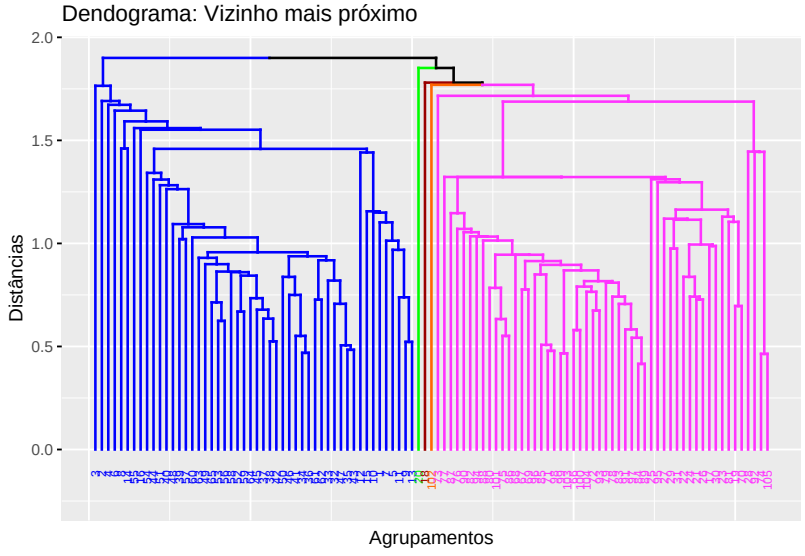
Primeiramente observando a matriz de correlação as empresas JP Morgan, Wells Fargo e Citibank apresentaram uma correlação maior que 0.5 enquanto as empresas Royal Dutch Shell e Exxon Mobil apresentam uma correlação de 0.68, com uma breve pesquisa na internet podemos notar que os agrupamentos que podem contidos no dendograma das empresas JP Morgan, Wells Fargo e Citibank como um grupo 1 e Royal Dutch Shell e Exxon Mobil como grupo 2, fazem sentindo pois o grupo 1 são bancos e o grupo 2 são petrolíferas.

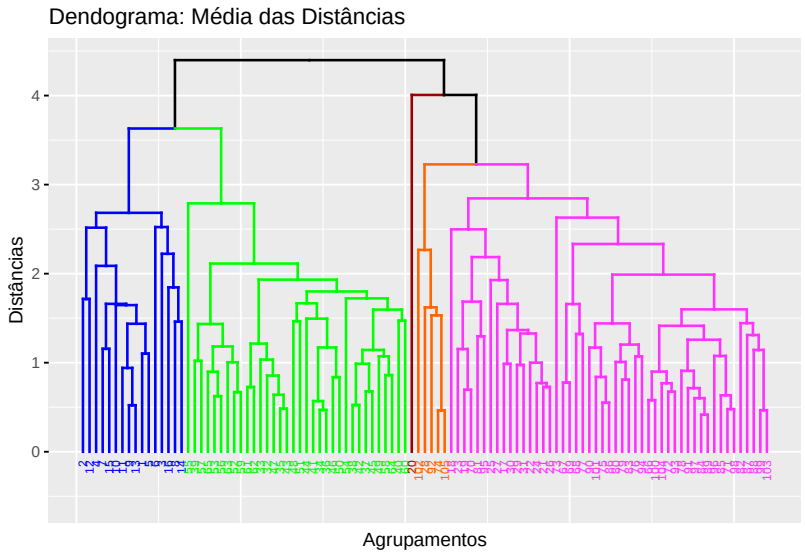
Exercício 4

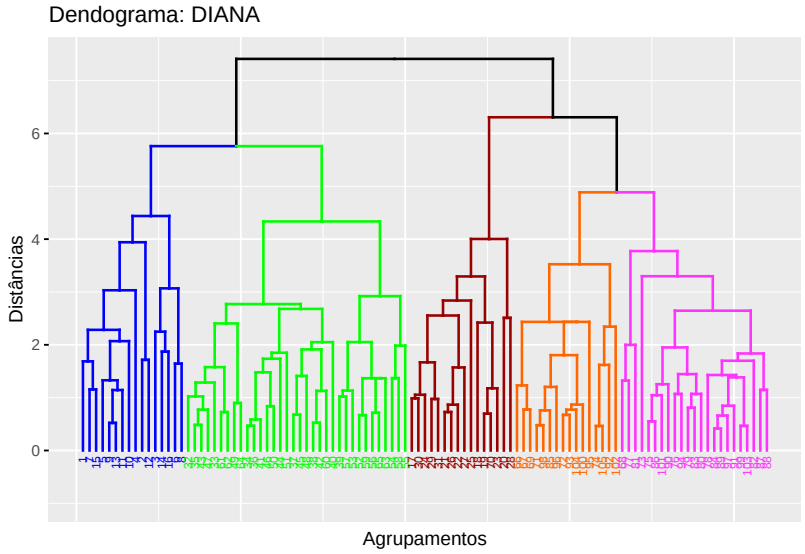
Os dados no arquivo **primate.scapulae.txt** são referentes a medidas feitas na escápula de cinco diferentes gêneros de primatas Hominoidea (Hylobates, Pong, Pan, Gorilla e Homo). As medidas estão nas variáveis AD.BD, AD.CD, EA.CD, Dx.CD, SH.ACR, EAD, β e γ. As cinco primeiras medidas são índices e as três últimas são ângulos. O ângulo não está disponível para os primatas Homo e, portanto, não deve ser usado na análise. Cuidado na leitura dos dados no formato texto, pois as medidas faltantes não estão representadas por **NA**. Com auxílio computacional, considerando apenas as 7 medidas das escápulas disponíveis, faça o agrupamento dos dados utilizando os métodos do vizinho mais próximo, vizinho mais longe, média das distâncias e também com o método hierárquico divisivo. Obtenha os correspondentes dendrogramas e compare os resultados do agrupamento com 5 grupos. Compare os agrupamentos com a classificação correta dos primatas (que não foi usada para a obtenção dos grupos). Para isso, você pode calcular as taxas de classificação incorreta e correta. Discuta os resultados.

Resolução

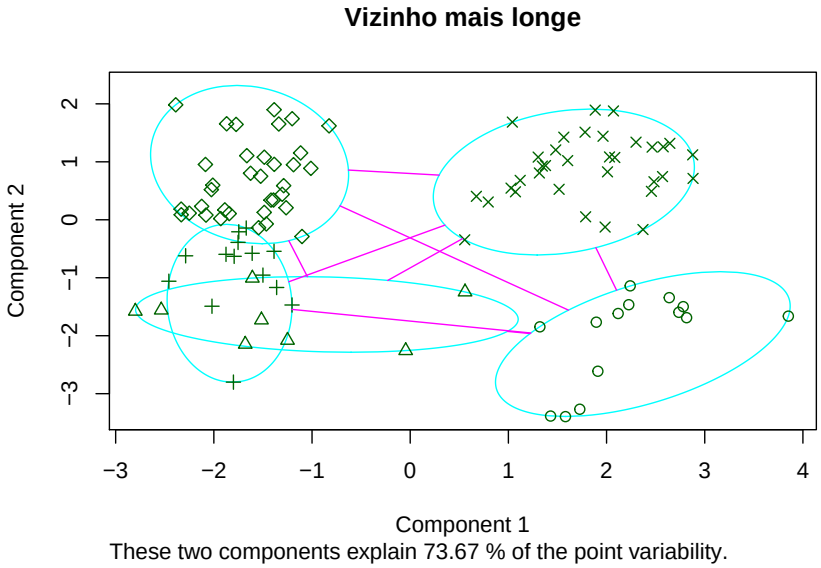






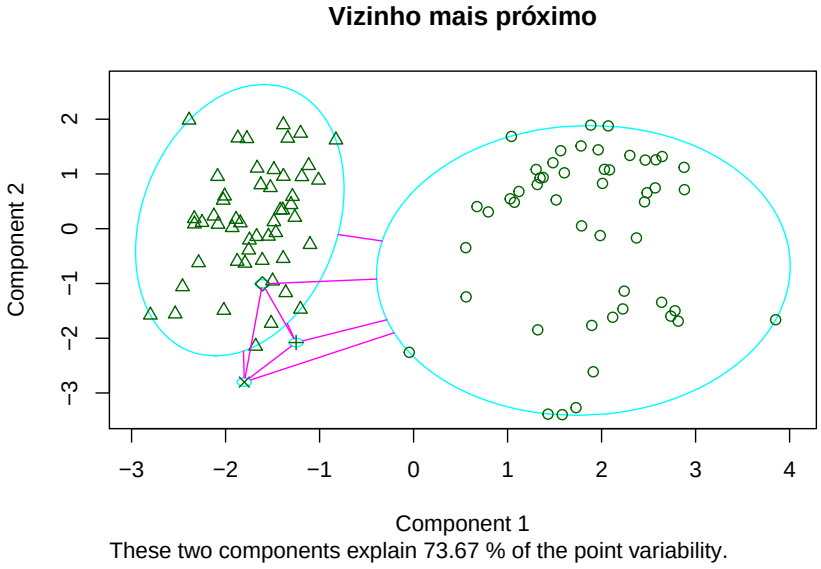


A fim de obter uma vizualização para as classificações, reduzimos a dimensionalidade do dados utilizando componentes principais, e apenas com 2 componentes nós temos 3.67 % da variância total explicada pelas componentes. Nos métodos do vizinho mais longe e DIANA, não observa-se grupos com apenas uma observação como é visto nos métodos do vizinho mais próximo e da média das distâncias.



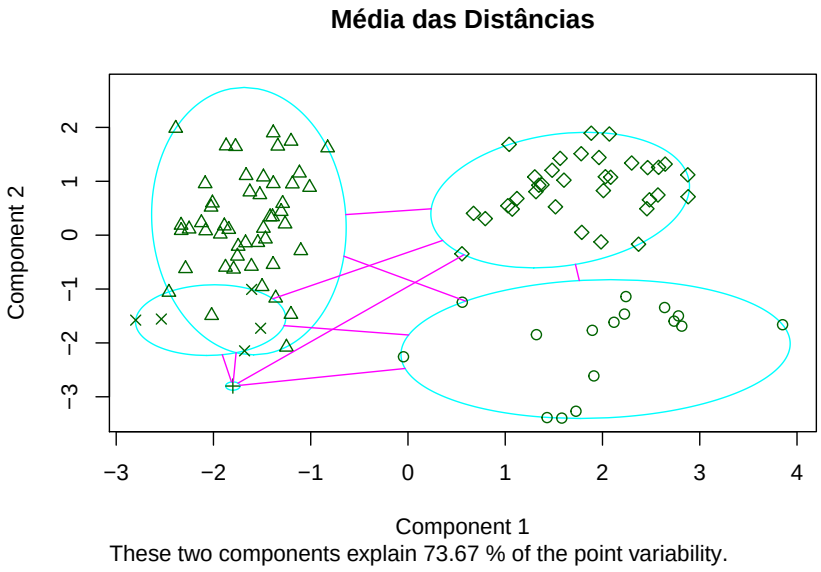
Taxa de acertos método Vizinho mais longe (%):

```
## [1] 61.9
```



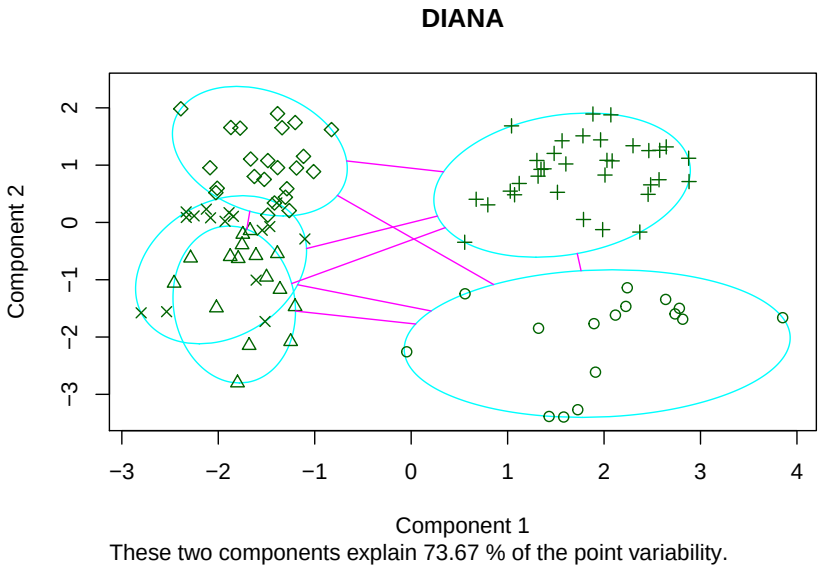
Taxa de acertos método Vizinho mais próximo (%):

```
## [1] 28.6
```



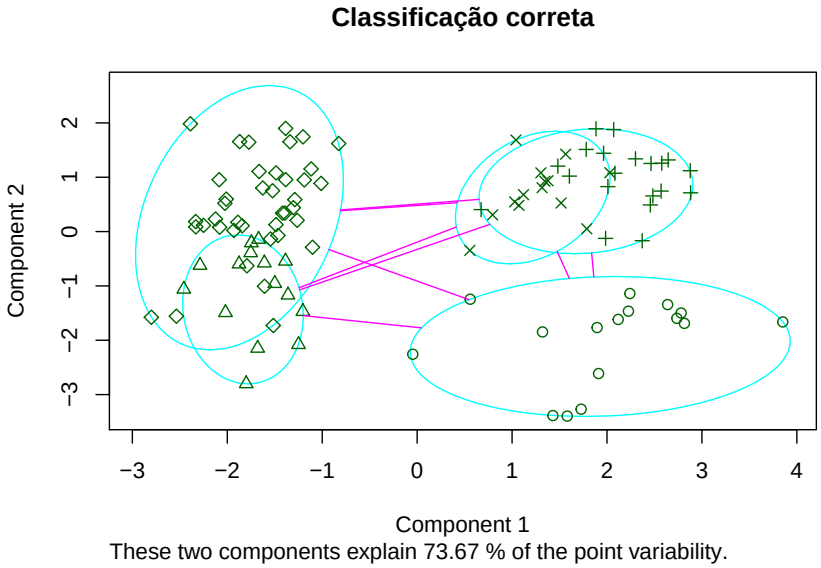
Taxa de acertos método Média das Distâncias (%):

```
## [1] 27.6
```



Taxa de acertos método DIANA (%):

```
## [1] 70.5
```



Pelos gráficos e taxas de classificação obtidos, no método do vizinho mais longe há um grupo com observações próximas de outros grupos, causando um confundimento, entretanto sua taxa foi a segunda maior. Para o segundo método há um grupo muito grande e três grupos com poucas observações. No método da média das distâncias, aparentemente há dois grupos muito parecidos se confundindo e um grupo com apenas uma observação. Os dois métodos anteriores apresentaram as piores taxas de classificação. O último método obteve a maior taxa de classificação e os grupos parecem distintos com exceção de 2 grupos que estão bem próximos. Comparando com a classificação correta é possível ver que nenhum método conseguiu identificar que existia dois grupos próximos na coordenada (2,1).

Exercício 5

Considere a seguinte matriz de variância amostral das variávies X_1 e X_2 :

$$S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

(a) Obtenha as componentes principais de S, bem como a variância explicada por cada componente.

Resolução

```
## Importance of components:
##                         Comp.1 Comp.2
## Standard deviation         1.5      0
## Proportion of Variance     1.0      0
## Cumulative Proportion      1.0      1
##
## Loadings:
##      Comp.1 Comp.2
## [1,]  1
## [2,]        -1
```

Como podemos observar toda a variância pode ser explicada apenas com a componete 1

(b) Obtenha a matriz de correlação dos dados a partir de S. Obtenha as componentes principais com base na matriz de correlação e também a variância explicada por cada componente. Compare os resultados com o item anterior.

Resolução

Matriz de correlação:

$$C = \begin{bmatrix} 1 & 0.632 \\ 0.632 & 1 \end{bmatrix}$$

```
## Importance of components:
##                           Comp.1 Comp.2
## Standard deviation     0.2598932      0
## Proportion of Variance 1.0000000      0
## Cumulative Proportion  1.0000000      1
##
## Loadings:
##      Comp.1 Comp.2
## [1,]  0.707  0.707
## [2,] -0.707  0.707
```

Comparando com o item anterior, pode-se observar que toda a variância pode ser explicada apenas com a primeira componete nos dois casos. Entretanto no primeiro caso a primeira componente era a primeira variável, por outro lado com a matriz de correlação a primeira compotente é a oposição entre a primeira e a segunda variável.

Exercício 6

Os dados das variáveis X_1 (vendas) e X_2 (lucro) das 10 maiores empresas no mundo em 2000 estão disponíveis no exercício 1.4 do livro do Johnson. O vetor de médias resultante e a matriz de variância são dados por

$$\bar{\mathbf{x}} = \begin{bmatrix} 155,60 \\ 14,70 \end{bmatrix} \text{ e } S = \begin{bmatrix} 7476,45 & 303,62 \\ 303,62 & 26,19 \end{bmatrix}.$$

(a) Obtenha as componentes principais de S e suas variâncias.

Resolução

```
## Importance of components:
##                           Comp.1 Comp.2
## Standard deviation     3589.097      0
## Proportion of Variance    1.000      0
## Cumulative Proportion     1.000      1
##
## Loadings:
```

```
##      Comp.1 Comp.2
## [1,]  0.999
## [2,]        -0.999
```

(b) Calcule a proporção da variância explicada pela primeira componente principal. Obtenha as correlações entre a primeira componente principal e as variáveis originais. Com base nos coeficientes e nas correlações, interprete a primeira componente principal.

Resolução

A proporção da variância explicada pela primeira componente principal é de:

```
## [1] 0.9981557
```

As correlações entre a primeira componente principal e as variáveis originais são:

Table 1: Correlção var. orginais e comp.

Correlação
-0.9999985
0.0295484

Com essas informações, conclui-se que a primeira componente explica a maioria da variabilidade total (99,82%) e essa é mais correlacionada com a primeira variável, logo a primeira variável é suficiente para a interpretação do problema.

(c) Refaça os itens (a) e (b) utilizando a matriz de correlção dos dados. Compare os resultados.

Resolução

```
## Importance of components:
##                            Comp.1 Comp.2
## Standard deviation     0.2219301      0
## Proportion of Variance 1.0000000      0
## Cumulative Proportion  1.0000000      1
##
## Loadings:
##      Comp.1 Comp.2
## [1,]  0.707  0.707
## [2,] -0.707  0.707
```

A proporção da variância explicada pela primeira componente principal é de:

```
## [1] 0.8430717
```

As correlações entre a primeira componente principal e as variáveis originais são:

Table 2: Correlção var. orginais e comp.

Correlação
0.9181894
-0.3961417

Comparando com os itens anteriores, assumimos que o primeiro componente é a oposição entre a primeira e a segunda variável, este possui 84,31% da variabilidade total dos dados. A componente ainda está mais correlacionada com a primeira variável mas a correlação da segunda variável é maior que no item b.

Exercício 7

Os dados no arquivo **T8-6.DAT** são referentes a recordes nacionais masculinos de corrida para diversos países. As colunas são referentes aos tempos recordes nas seguintes modalidades, respectivamente:

- 100 m (segundos);
- 200 m (segundos);
- 400 m (segundos);
- 800 m (minutos);
- 1500 m (minutos);
- 5000 m (minutos);
- 10.000 m (minutos);
- Maratona (minutos).

(a) Obtenha a matriz de correlação dos dados e seus autovalores e autovetores.

Resolução

Seja $\lambda_1, ..., \lambda_8$ os autovalores e $e_1, ..., e_8$ os autovalores da matriz de correlação dos dados, assim:

Table 3: Matrix de Correlações

	100m	200m	400m	800m	1500m	5000m	10000m	Maratona
100m	1.0000000	0.9147554	0.8041147	0.7119388	0.7657919	0.7398803	0.7147921	0.6764873
200m	0.9147554	1.0000000	0.8449159	0.7969162	0.7950871	0.7613028	0.7479519	0.7211157
400m	0.8041147	0.8449159	1.0000000	0.7677488	0.7715522	0.7796929	0.7657481	0.7126823
800m	0.7119388	0.7969162	0.7677488	1.0000000	0.8957609	0.8606959	0.8431074	0.8069657
1500m	0.7657919	0.7950871	0.7715522	0.8957609	1.0000000	0.9165224	0.9013380	0.8777788
5000m	0.7398803	0.7613028	0.7796929	0.8606959	0.9165224	1.0000000	0.9882324	0.9441466
10000m	0.7147921	0.7479519	0.7657481	0.8431074	0.9013380	0.9882324	1.0000000	0.9541630
Maratona	0.6764873	0.7211157	0.7126823	0.8069657	0.8777788	0.9441466	0.9541630	1.0000000

Table 4: Autovalores

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
6.703	0.638	0.228	0.206	0.098	0.071	0.047	0.01

Table 5: Autovetores

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
-0.332	-0.529	-0.344	0.381	0.300	-0.362	0.348	-0.066
-0.346	-0.470	0.004	0.217	-0.541	0.349	-0.440	0.061
-0.339	-0.345	0.067	-0.851	0.133	0.077	0.114	-0.003
-0.353	0.089	0.783	0.134	-0.227	-0.341	0.259	-0.039
-0.366	0.154	0.244	0.233	0.652	0.530	-0.147	-0.040
-0.370	0.295	-0.183	-0.055	0.072	-0.359	-0.328	0.706
-0.366	0.334	-0.244	-0.087	-0.061	-0.273	-0.351	-0.697
-0.354	0.387	-0.335	0.018	-0.338	0.375	0.594	0.069

(b) Obtenha as duas primeiras componentes principais das variáveis padronizadas. Obtenha também as correlações entre as variáveis originais e as duas primeiras componentes principais, além da variabilidade (acumulada) explicada por cada componente.

Resolução

Primeiramente é obtido os autovalores e autovetores, em seguida as duas primeiras componentes principais das variáveis padronizadas:

Table 6: Autovalores

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
6.703	0.638	0.228	0.206	0.098	0.071	0.047	0.01

Table 7: Autovetores

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
-0.332	-0.529	0.344	0.381	-0.300	-0.362	0.348	-0.066
-0.346	-0.470	-0.004	0.217	0.541	0.349	-0.440	0.061
-0.339	-0.345	-0.067	-0.851	-0.133	0.077	0.114	-0.003
-0.353	0.089	-0.783	0.134	0.227	-0.341	0.259	-0.039
-0.366	0.154	-0.244	0.233	-0.652	0.530	-0.147	-0.040
-0.370	0.295	0.183	-0.055	-0.072	-0.359	-0.328	0.706
-0.366	0.334	0.244	-0.087	0.061	-0.273	-0.351	-0.697
-0.354	0.387	0.335	0.018	0.338	0.375	0.594	0.069

Table 8: Componetes Principais

Com. 1	Comp. 2
-0.332	-0.529
-0.346	-0.470
-0.339	-0.345
-0.353	0.089
-0.366	0.154
-0.370	0.295
-0.366	0.334
-0.354	0.387

Obtemos então as correlações entre as duas primeiras componentes principais, respectivamente, e as variáveis originais:

Table 9: Correlção var. orginais e comp. Principais

Correlação
-0.9999985
0.0295484

Por fim obtemos a variabilidade (acumulada) explicada por cada componente:

Table 10: Variabilidade Explicada (acumulada) por cada comp.

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
0.8379112	0.9177125	0.9461531	0.9718842	0.9840814	0.9929174	0.9987851	1

(c) Interprete as duas primeiras componentes obtidas no item anterior.

Resolução

Pelo item anterior, a primeira componente está associada à todas as variaveis e a segunda ao contraste entre as variáveis (100m, 200m, 400m) e (1500m, 5000m, 10000m, Maratona), ou seja a diferença entre corridas pequenas e grandes. Essas duas componentes representam 91,77% da variabilidade total.

(d) Ordene os países com base nos escores obtidos para a primeira componente principal e discuta os resultados.

Resolução

	comp1	dat.Pais
54	-85.63052	U.S.A.
29	-85.81863	Kenya
17	-86.91273	France
19	-86.94300	GreatBritain
28	-87.06855	Japan

	comp1	dat.Pais
6	-87.06938	Brazil
4	-87.18382	Belgium
2	-87.22922	Australia
35	-87.28480	Mexico
43	-87.32580	Portugal
27	-87.33011	Italy
48	-87.50262	Spain
18	-87.58127	Germany
37	-88.16301	Netherlands
42	-88.19532	Poland
45	-88.23536	Russia
7	-88.40680	Canada
30	-88.44795	Korea,South
38	-88.54557	NewZealand
50	-88.55869	Switzerland
25	-88.68066	Ireland
9	-88.77554	China
39	-88.93878	Norway
14	-88.99931	Denmark
1	-89.03377	Argentina
49	-89.16637	Sweden
16	-89.28152	Finland
20	-89.67693	Greece
22	-89.68289	Hungary
13	-89.73418	CzechRepublic
3	-89.80797	Austria
10	-89.83993	Columbia
44	-89.92071	Romania
8	-89.92317	Chile
53	-89.97282	Turkey
31	-90.15981	Korea,North
23	-90.28064	India
12	-91.31959	CostaRica
26	-91.51816	Israel
32	-91.91272	Luxembourg
51	-91.92754	Taiwan
21	-92.02638	Guatemala
41	-93.25636	Philippines
52	-93.68469	Thailand
24	-93.79810	Indonesia
34	-94.34089	Mauritius
36	-95.03512	Myanmar(Burma)
15	-95.90629	DominicanRepub
5	-96.20086	Bermuda
47	-97.05492	Singapore
33	-97.30645	Malaysia
40	-98.14549	PapuaNewGuinea
46	-106.18833	Samoa
11	-110.88311	CookIslands

Pelo escores obtidos, o país com menor tempo nas provas pela primeira componente foi os EUA e os piores tempos são de Samoa e das Ilhas Cook, as quais aparecem com um valor muito diferente dos demais países.

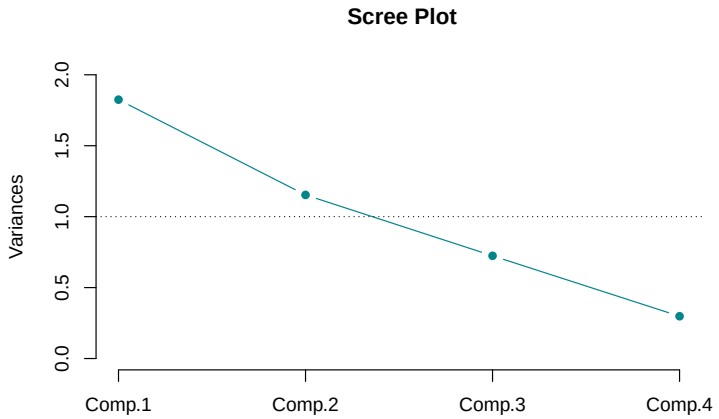
Exercício 8

O arquivo **oleo.csv** apresenta um sumário de consumo de petróleo em alguns países no ano de 2011. O consumo é apresentado em consumo diário (em m^3) e em consumo per capita (em m^3) e a tabela contém informação sobre a razão produção/consumo. Faça uma análise de componentes principais com os dados. Obtenha o *screeplot* e o *biplot*. Interprete o biplot, destacando possíveis agrupamentos de nações.

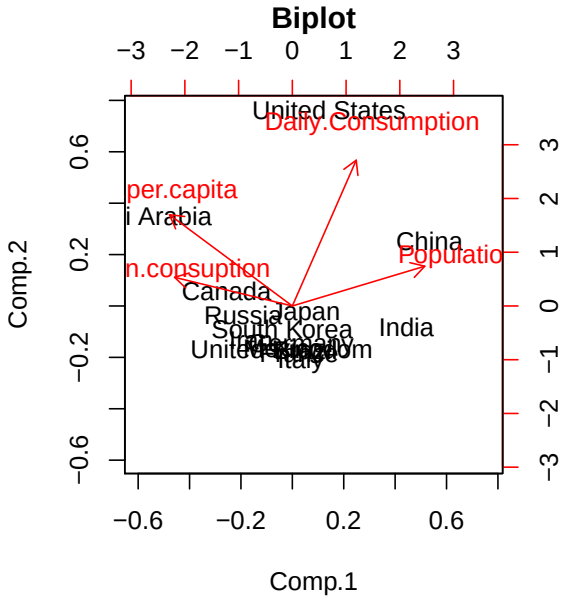
Resolução

```
##                       Daily.Consumption Population Annual.per.capita
## Daily.Consumption             1.0000000  0.3735493         0.1346379
## Population                    0.3735493  1.0000000        -0.4497889
## Annual.per.capita             0.1346379 -0.4497889         1.0000000
## production.consuption        -0.1676843 -0.2557144         0.4274717
##                       production.consuption
## Daily.Consumption                -0.1676843
## Population                       -0.2557144
## Annual.per.capita                 0.4274717
## production.consuption             1.0000000

## Importance of components:
##                            Comp.1    Comp.2    Comp.3    Comp.4
## Standard deviation     1.3507962 1.0738871 0.8509422 0.5459062
## Proportion of Variance 0.4561626 0.2883084 0.1810257 0.0745034
## Cumulative Proportion  0.4561626 0.7444709 0.9254966 1.0000000
##
## Loadings:
##                       Comp.1 Comp.2 Comp.3 Comp.4
## Daily.Consumption      0.284  0.815  0.157  0.480
## Population             0.589  0.222 -0.553 -0.546
## Annual.per.capita     -0.546  0.510  0.237 -0.621
## production.consuption -0.524  0.160 -0.784  0.294
```



Segundo o critério de Kaiser que considera que o número de componentes principais tem autovalores maiores que 1, assim observando o Scree plot acima, duas componentes principais é suficiente para explicar a maior parte da variavilidade dos dados.



Observando o biplot acima, podemos notar que países como EUA e China tem grandes populações e consumo de petróleo, Arábia Saudita tem grande consumo per capita e que os outros países se concentram pela razão produção/consumo e cosumo per capita.