

# GRUPO 1

## APRESENTAÇÃO DE: ANÁLISE DE CLUSTERS

CAIO RODRIGUES GOMES 11208012

RODRIGO DORNELES FERREIRA DE SOUZA 11295831

VITOR CAETANO DA SILVA 9276999

**Profª Ana Amélia**

**Monitores: Maurício e Rosa**

MQAM - ACH2036

# NOSSOS TÓPICOS DE HOJE

O que é a Clusterização?

Como usar a Clusterização?

O Dataset Escolhido

Como fazer a Análise de Cluster?

Identificação das Variáveis Utilizadas na Análise

Pergunta a ser respondida com a Análise

Resultados da Análise

Método hierárquico e não-hierárquico

Interpretação dos Resultados

Conclusão

# O QUE É A CLUSTERIZAÇÃO?

EXPLICANDO COMO FUNCIONA

Um **cluster** é um **subconjunto** de um dataset, no qual através de **características compartilhadas**, alguns dados foram **aglomerados** com um centro.

**O centro pode ser inicializado de forma aleatória e atualizado com a disposição dos dados, de forma que os dados mais próximos deste centro formam um cluster.**

## DESTA FORMA

A análise de clusters é um conjunto de **técnicas de interdependência** que faz o **agrupamento** de dados em **clusters** segundo sua **semelhança**.

## A análise desses clusters é algo exploratório:

- Clusters podem ser complexos (com mais de 3 dimensões);
- Os clusters dependem do método utilizado para defini-los;
- Os clusters mudam dependendo de quantas iterações foram feitas;

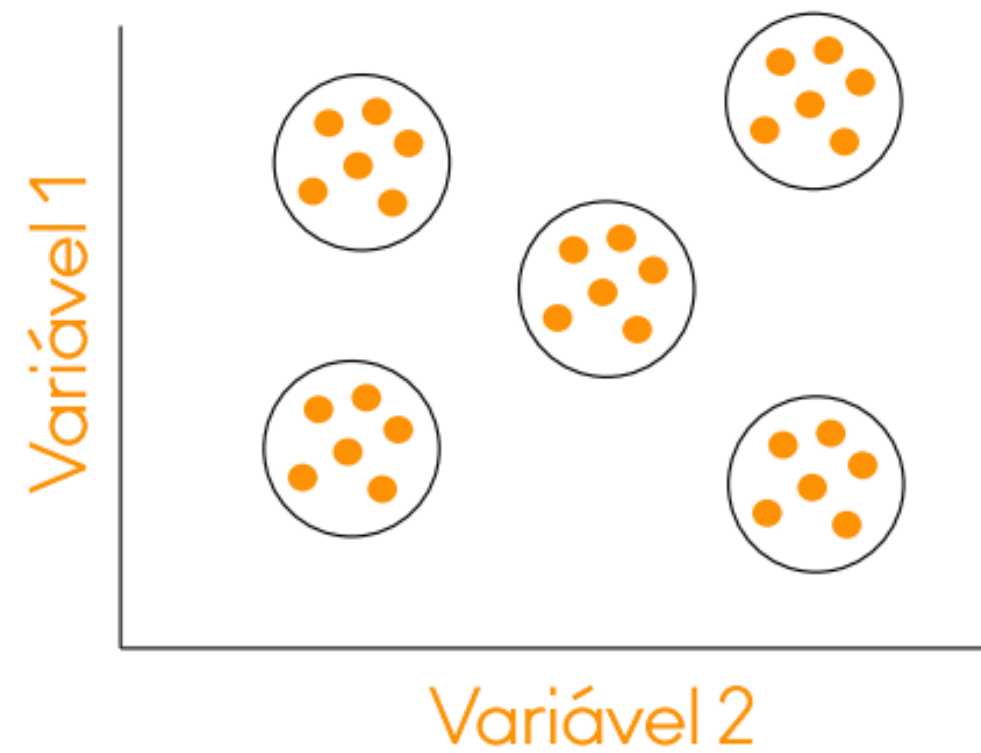
# REMOVER CERTAS VÁRIAVEIS

É também necessário eliminar as variáveis que não diferenciem significativamente os grupos. Em outras palavras, após os clusters serem formados, variáveis que não diferenciem os grupos não justificam as personas ou as ações sugeridas para cada cluster. Além disso, não se deve incluir variáveis indiscriminadamente porque a análise é sensível à inclusão de variáveis irrelevantes.

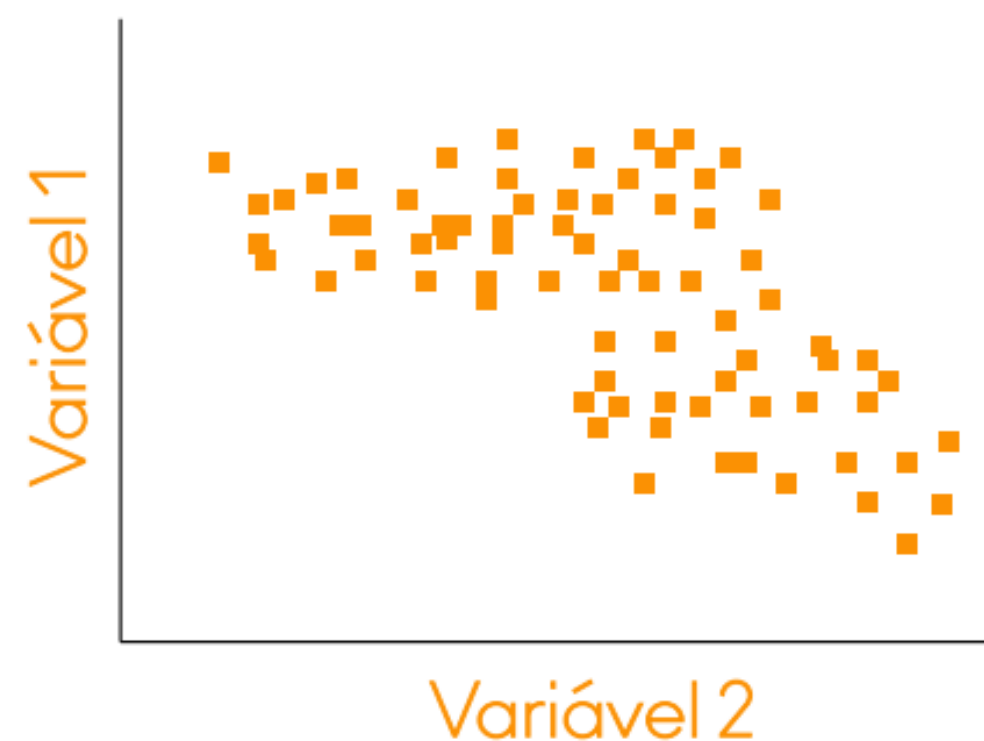


# EXEMPLO

○ IDEAL



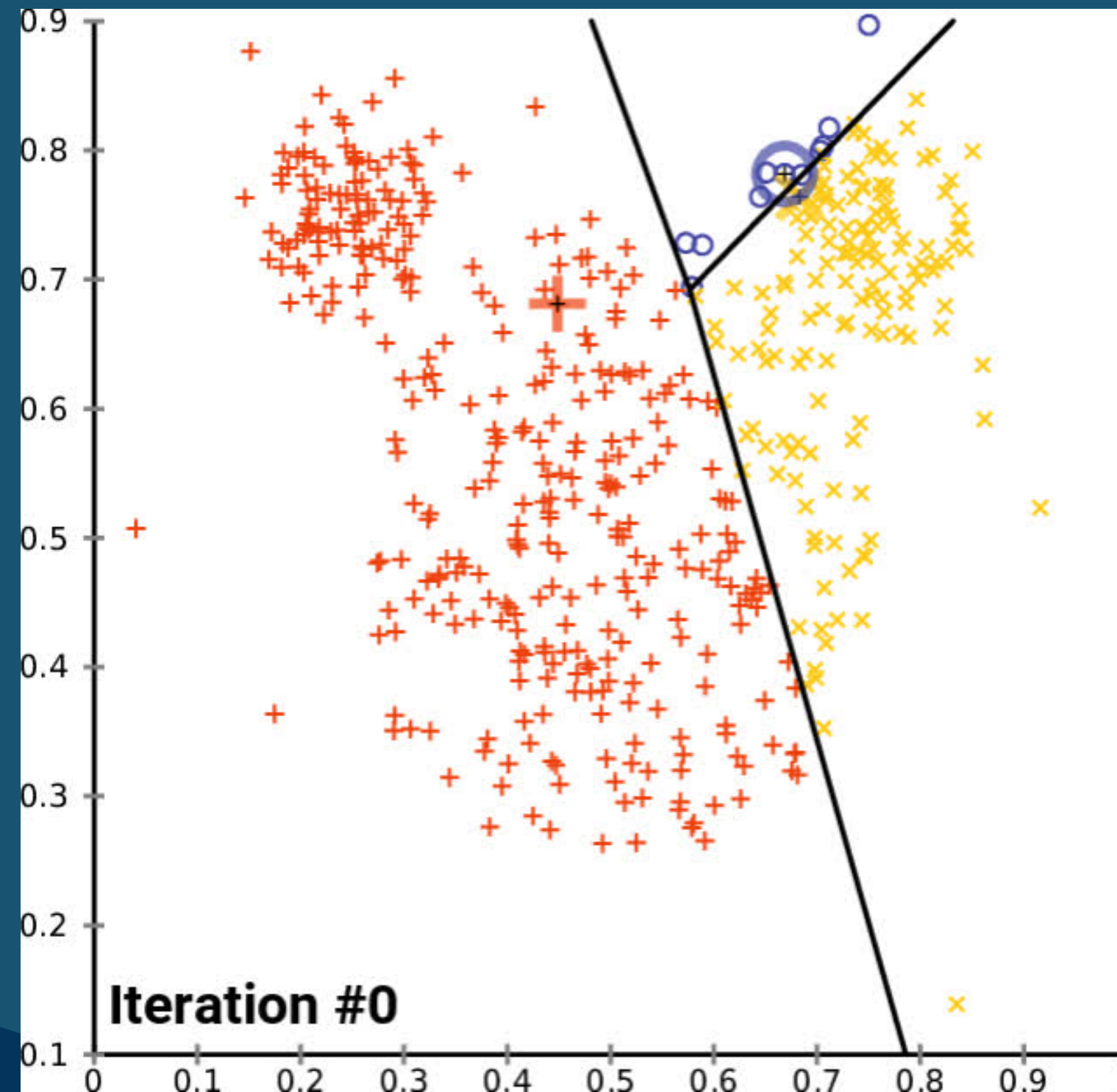
○ REAL



A **ANÁLISE DE CLUSTER** PEDE POR ALGUMAS PREMISAS PARA QUE POSSA SER EXECUTADA. EM PRIMEIRO LUGAR, É **PRECISO ELIMINAR AS VARIÁVEIS** QUE **NÃO REPRESENTAM POSSIBILIDADES DE ADERÊNCIA** COM A **ELIMINAÇÃO DE VARIÁVEIS QUE NÃO COMPÕEM FATORES**.

# EXEMPLO DE CONVERGENCIA

11



# MÉTODO HIERÁRQUICO

12

Em clustering hierarquico, o algoritmo funciona da seguinte forma:

- Cada **dado inicialmente é tratado como um cluster**. O número de dados e de clusters total sempre é K.
- Conforme a **distância euclidiana mínima** de clusters aumenta, **junte 2 clusters em 1**. Isso nos dá K-1 clusters.
- **Repita isso** até que tudo se torne um **enorme cluster**.
- **Usando dendrogramas** após o fim do algoritmo, podemos **dividir os clusters** conforme quisermos, dependendo do caso de uso.

# MÉTODO HIERÁRQUICO

13

**Vantagens:** Simples de ser implementado.

**Desvantagens:** Sensível a erros no começo (objetos agrupados errados no começo não podem ser consertados), além de não lidar bem com outliers.

# MÉTODO (KMEANS) NÃO-HIERÁRQUICO

- **Escolha K centróides** para existirem.
- Aleatoriamente **inicialize** os centróides.
- **Separe** os pontos em clusters dependendo da distância deles aos centróides. Ele se torna parte do cluster com o centróide mais próximo.
- Faça a **média dos pontos** dentro dos clusters, assinale o novo centróide à média.
- Pare quando os centróides pararem de se mexer, e nenhum novo ponto mudar de centróide (**convergência**).

# MÉTODO (KMEANS) NÃO-HIERÁRQUICO

**Vantagens:** Escalavel para grandes datasets e adaptável.

**Desvantagens:** Devido a aleatoriedade inicial, resultados podem não ser reobtidos (instável). Tem problemas se os clusters não tem uma forma esférica.



# DATASET: DADOS DE SAÚDE DA POPULAÇÃO DOS PAÍSES

16

## OS DADOS

Acompanhamento do *estado da saúde da população*: Fatores relacionados à imunização, Fatores de mortalidade, Fatores econômicos e Fatores sociais.  
(2017)



## LEVANTAMENTO

Estes dados foram coletados pela **Organização Mundial da Saúde** e pelas **Nações Unidas**, sendo raspados por Deeksha Russell e Duan Wang.

## TAMANHO

**2938** linhas e **22** colunas

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>



# AS VARIÁVEIS DO DATASET

NOME DA COLUNA	QUALITATIVA NOMINAL	QUALITATIVA ORDINAL	QUANTITATIVA CONTÍNUA	QUANTITATIVA DISCRETA
Expectativa de vida			X	
País	X			
Ano		X		
Status (Desenvolvido / Em desenvolvimento)	X			
Alcoolismo				X
Morte infantil (por 1000)				X
Morte adulta(por 1000)				X
Vacinação contra hepatite B (%)				X
IMC			X	
Vacinação contra pólio (%)				X
Vacinação contra difteria(%)				X
PIB per capita			X	
Anos na escola (média)				X

A EXPECTATIVA DE  
VIDA, EDUCAÇÃO E  
PIB DE UM PAÍS  
REALMENTE  
EXPLICAM A SUA  
CATEGORIA COMO  
DESENVOLVIDO?

Para que não haja duplicidade de países, nos restringiremos ao ano de 2015 no tratamento dos dados.

# GRÁFICOS

NOS SLIDES A SEGUIR

# CÓDIGO FONTE ()

21



```
dados1 <- aggregate(mydata[,c(4)], by= list(mydata$"Year"), FUN=mean, na.rm=TRUE)

b1 <- barplot(height=dados1[,c(2)],width=dados1[,c(1)], ylim=c(50, 80),main = "Média da expectativa de
vida por ano", xpd=FALSE, names = dados1[,c(1)])

text(x=b1, y=dados1[,c(2)]+2, labels=round(dados1[,c(2)],1), cex=0.8)


dados1 <- aggregate(mydata[,c(4)], by= list(mydata$"Schooling"), FUN=mean, na.rm=TRUE)

b1 <- barplot(height=dados1[,c(2)],width=dados1[,c(1)], ylim=c(50, 100),main = "Anos na escola",
xpd=FALSE, names = dados1[,c(1)])

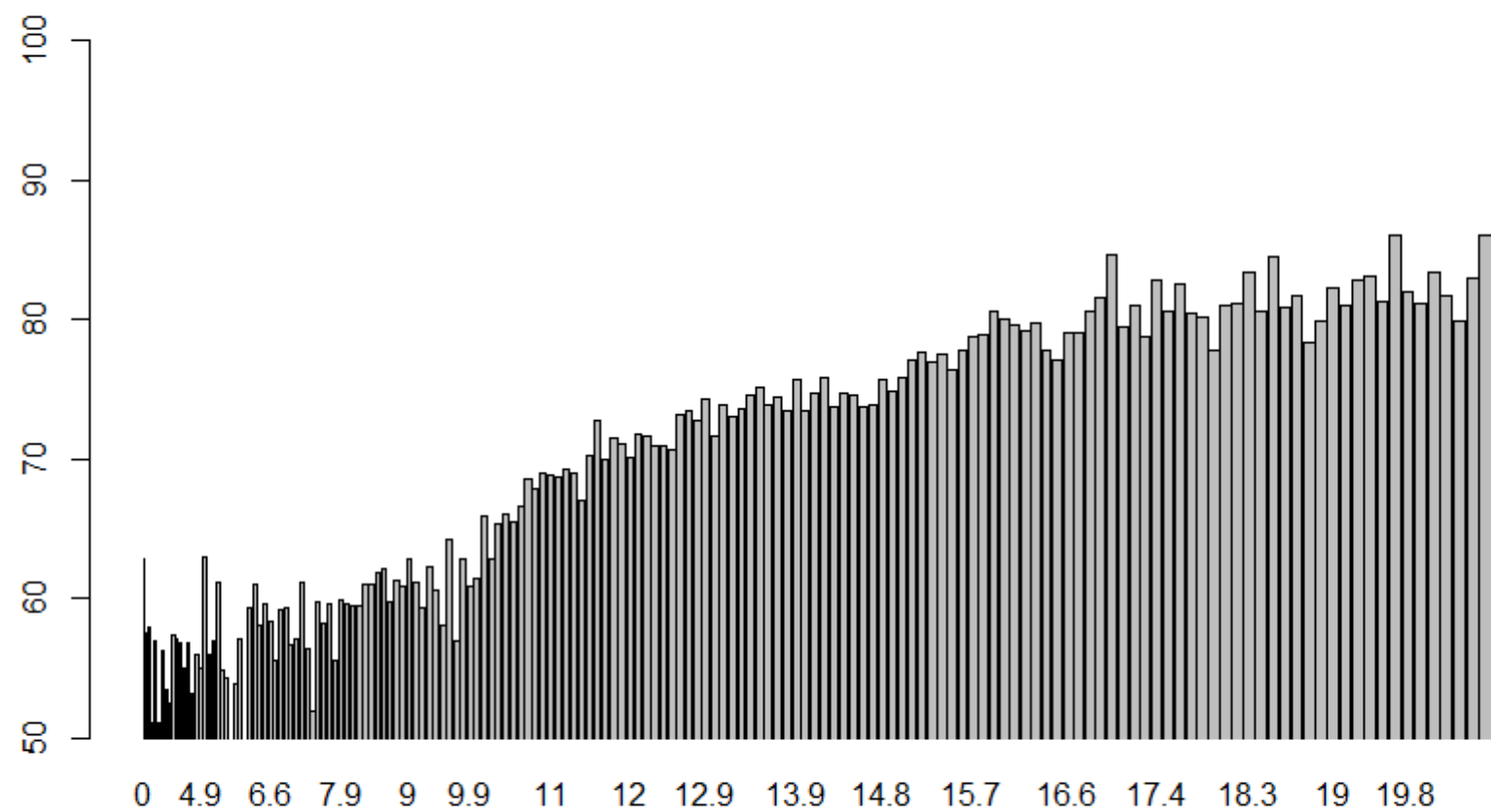
text(x=b1, y=dados1[,c(2)]+2, labels=round(dados1[,c(2)],1), cex=0.8)


dados1 <- aggregate(mydata[,c(4)], by= list(mydata$"Income.composition.of.resources"), FUN=mean,
na.rm=TRUE)

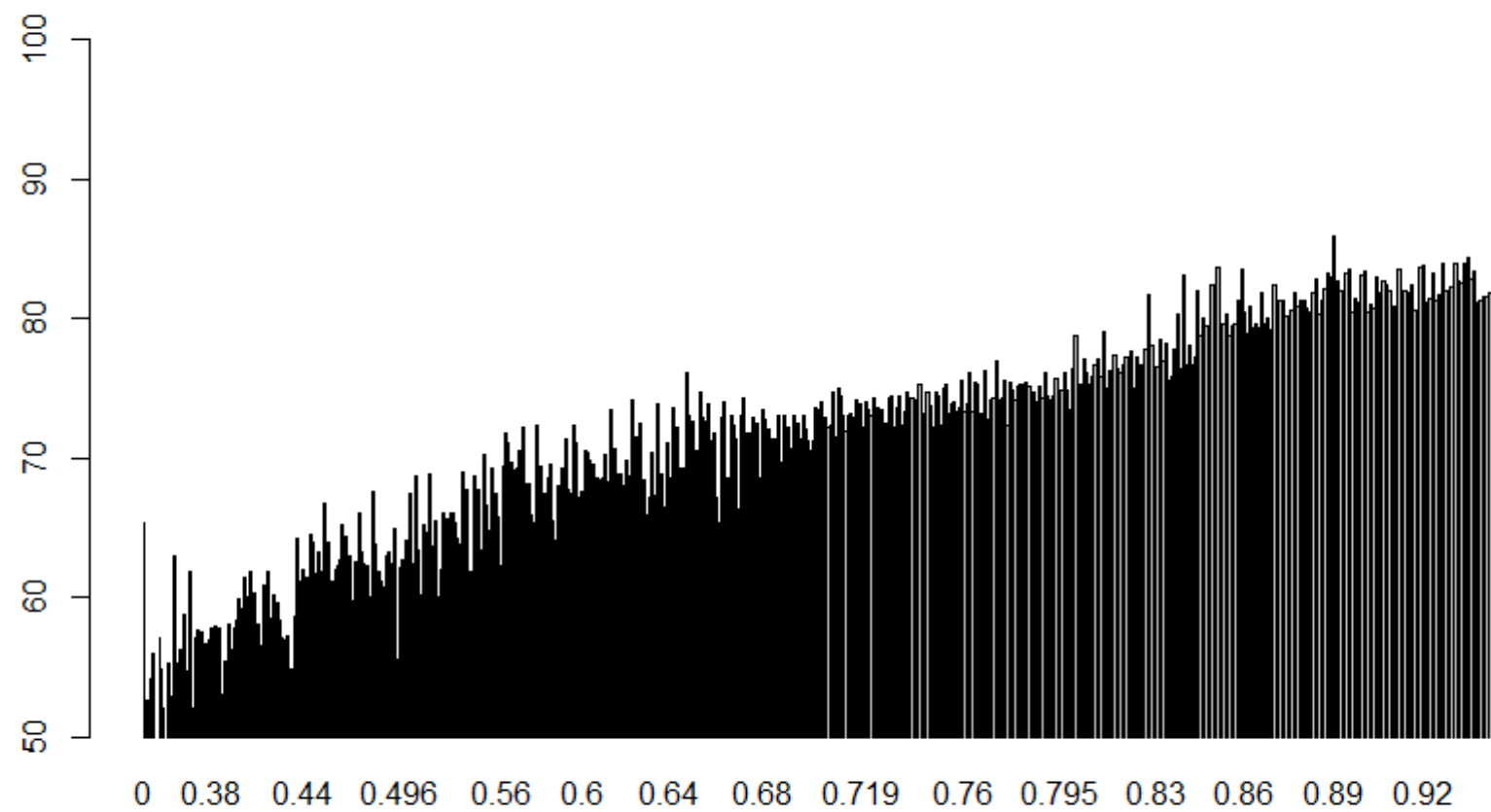
b1 <- barplot(height=dados1[,c(2)],width=dados1[,c(1)], ylim=c(50, 100),main = "PIB per Capita",
xpd=FALSE, names = dados1[,c(1)])

text(x=b1, y=dados1[,c(2)]+2, labels=round(dados1[,c(2)],1), cex=0.8)
```

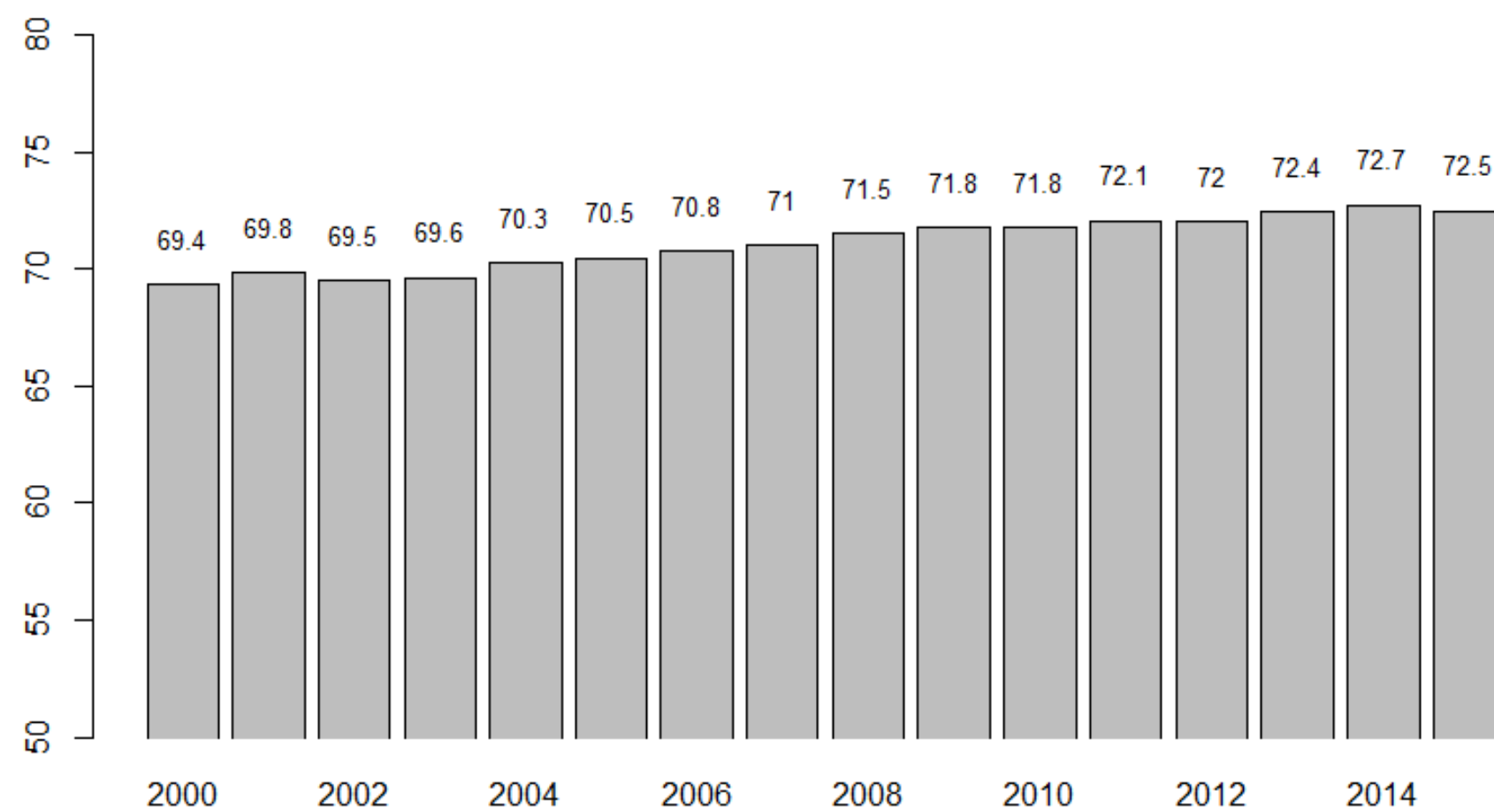
Anos na escola



PIB per Capita



Média da expectativa de vida por ano



# CÓDIGO FONTE ()

23



```
#filtrar para o ano mais recente
```

```
mydata <- mydata[mydata$Year == '2015',]  
dim(mydata)
```

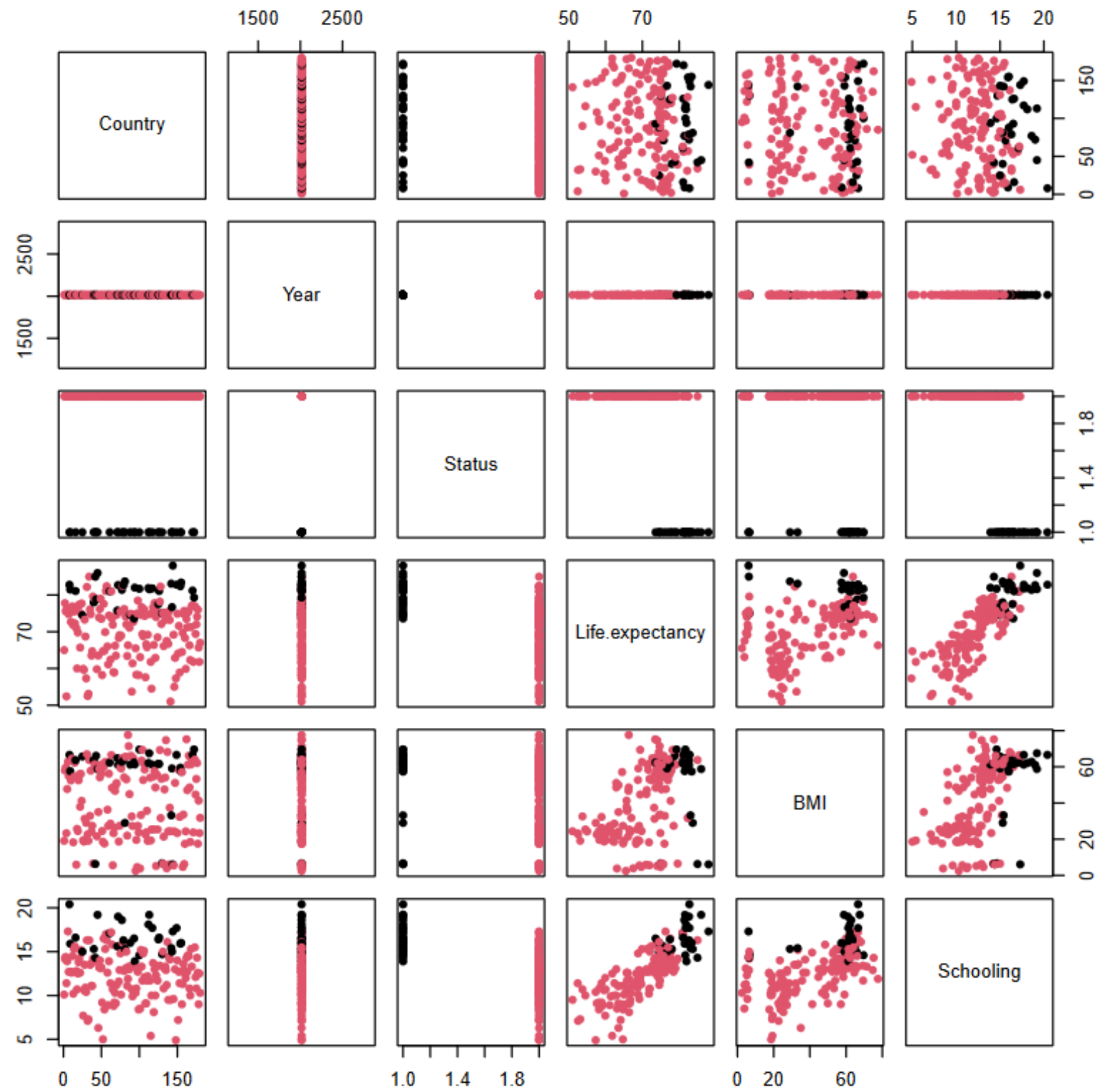
```
#-----  
-----
```

```
#subset dos dados
```

```
valores_relevantes <- mydata[,c(1,2,3,4,11,22)]  
valores_relevantes$Status <- as.factor(valores_relevantes$Status)
```

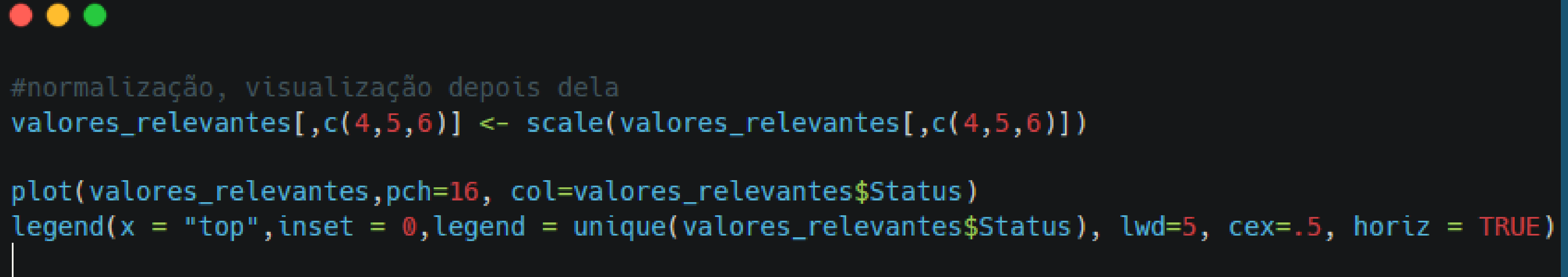
```
colnames(valores_relevantes)
```

```
plot(valores_relevantes, pch=16, col=valores_relevantes$Status)  
legend(x = "top", inset = 0, legend = unique(valores_relevantes$Status), lwd=5, cex=.5, horiz = TRUE)
```



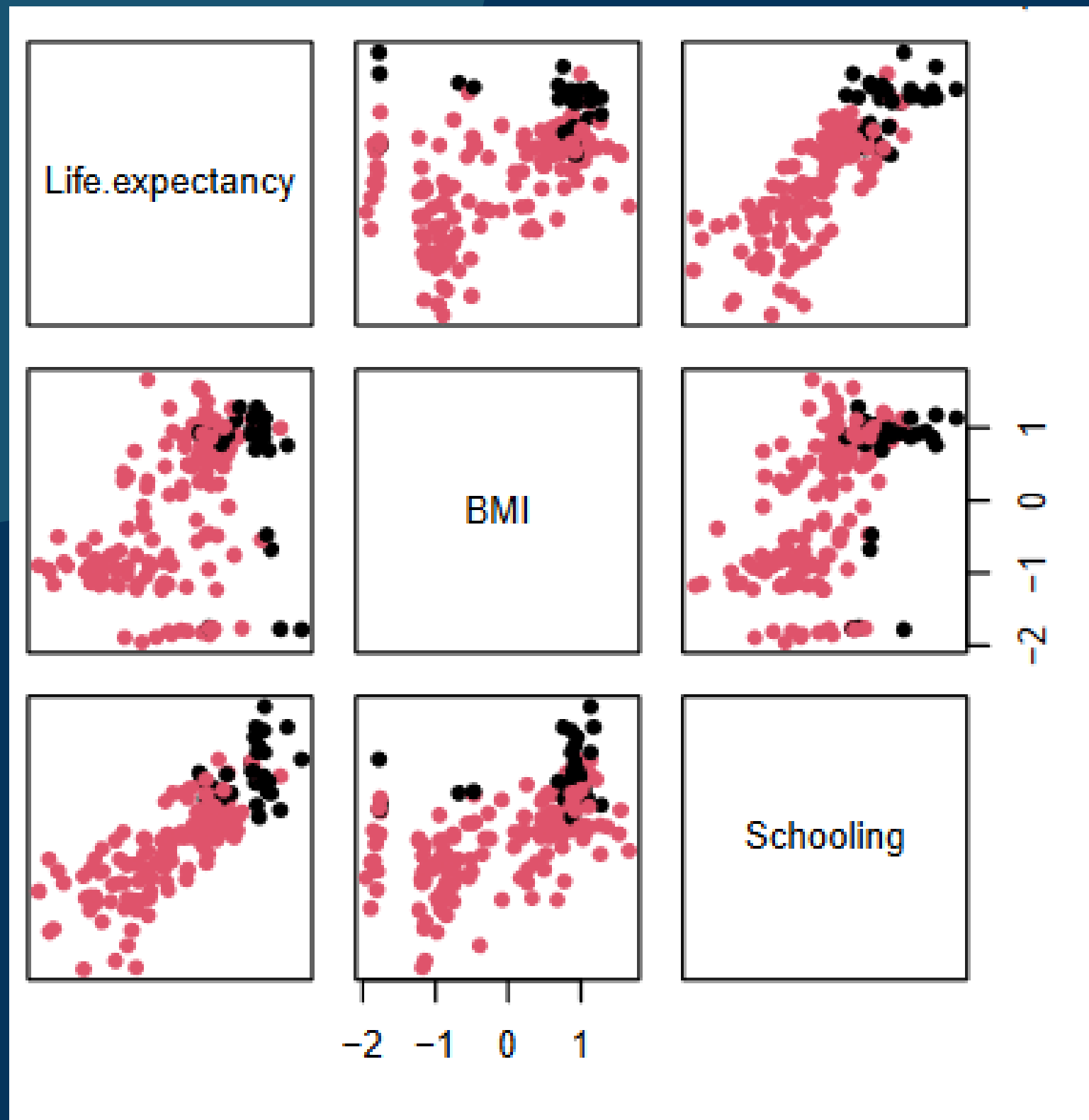


# NORMALIZAÇÃO



```
#normalização, visualização depois dela
valores_relevantes[,c(4,5,6)] <- scale(valores_relevantes[,c(4,5,6)])

plot(valores_relevantes,pch=16, col=valores_relevantes$Status)
legend(x = "top",inset = 0,legend = unique(valores_relevantes$Status), lwd=5, cex=.5, horiz = TRUE)
|
```



**Preto:**

Desenvolvido

**Vermelho:**

Subdesenvolvido

Para todos os próximos processos:

- A distância usada foi a euclidiana;
- Seeds para o K-Means não foram guardadas.

# MÉTODO HIERÁRQUICO

NOS SLIDES A SEGUIR

# APLICAÇÃO HIERÁRQUICA



```
#modelo hierarquico
```

```
numeros <- valores_relevantes[,c(4,5,6)]
```

```
rownames(numeros) <- valores_relevantes[,c(1)]
```

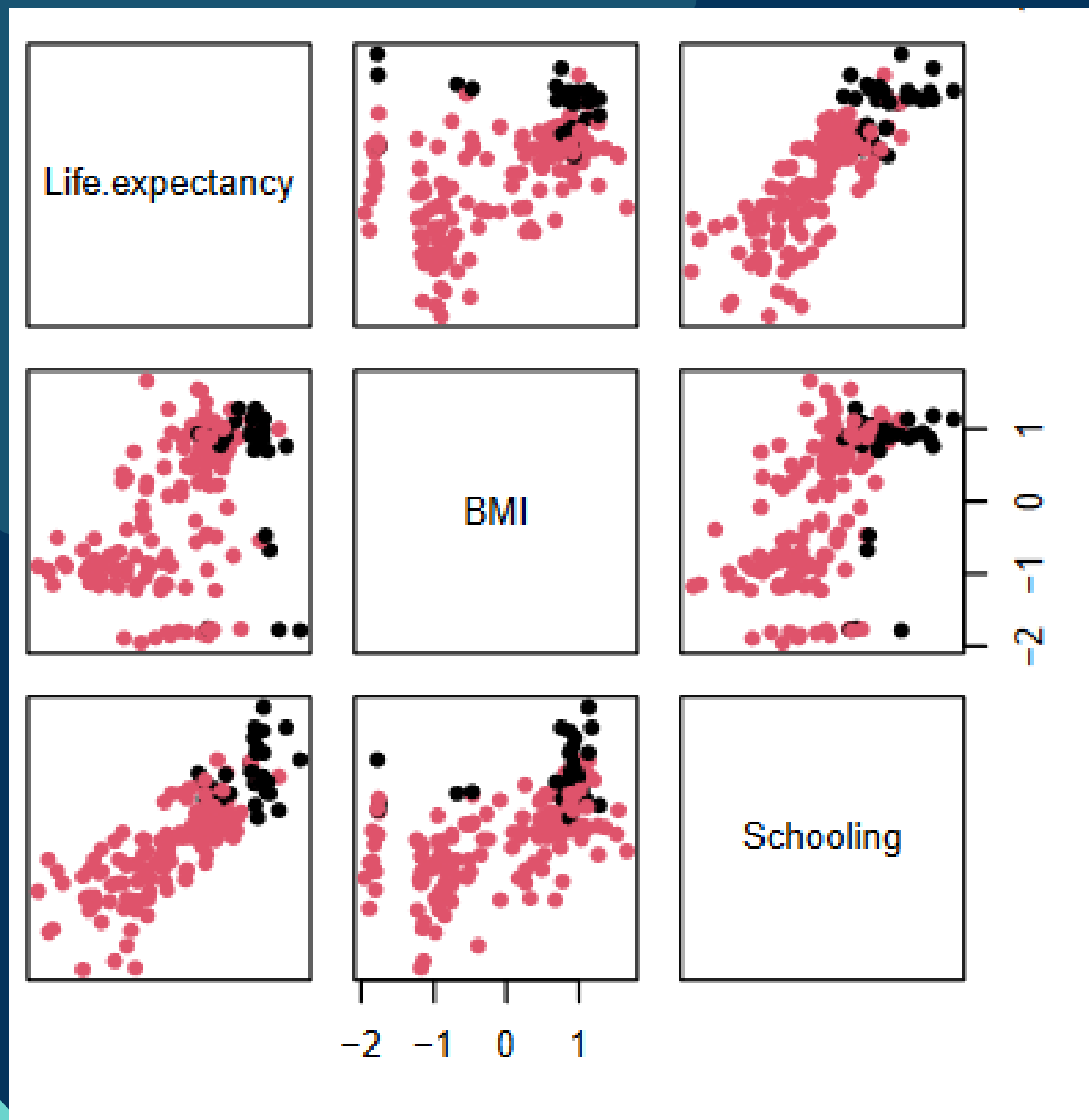
```
#método 'complete'
```

```
modelo <- hclust(dist(numeros))
```

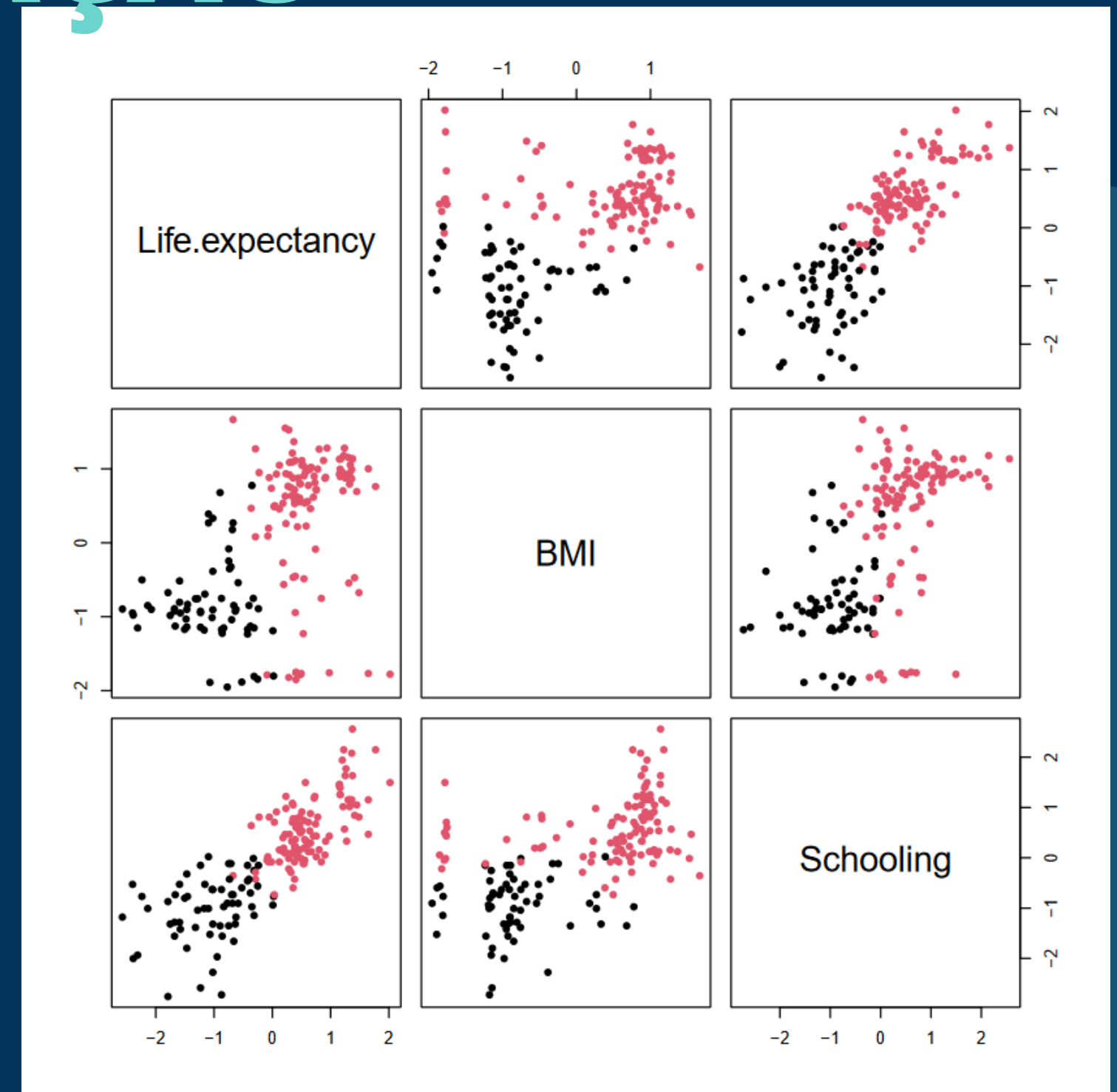
```
plot(modelo, cex=0.3)
```

```
plot(numeros, pch=16, col=cutree(modelo, 2))
```

# COMPARAÇÃO



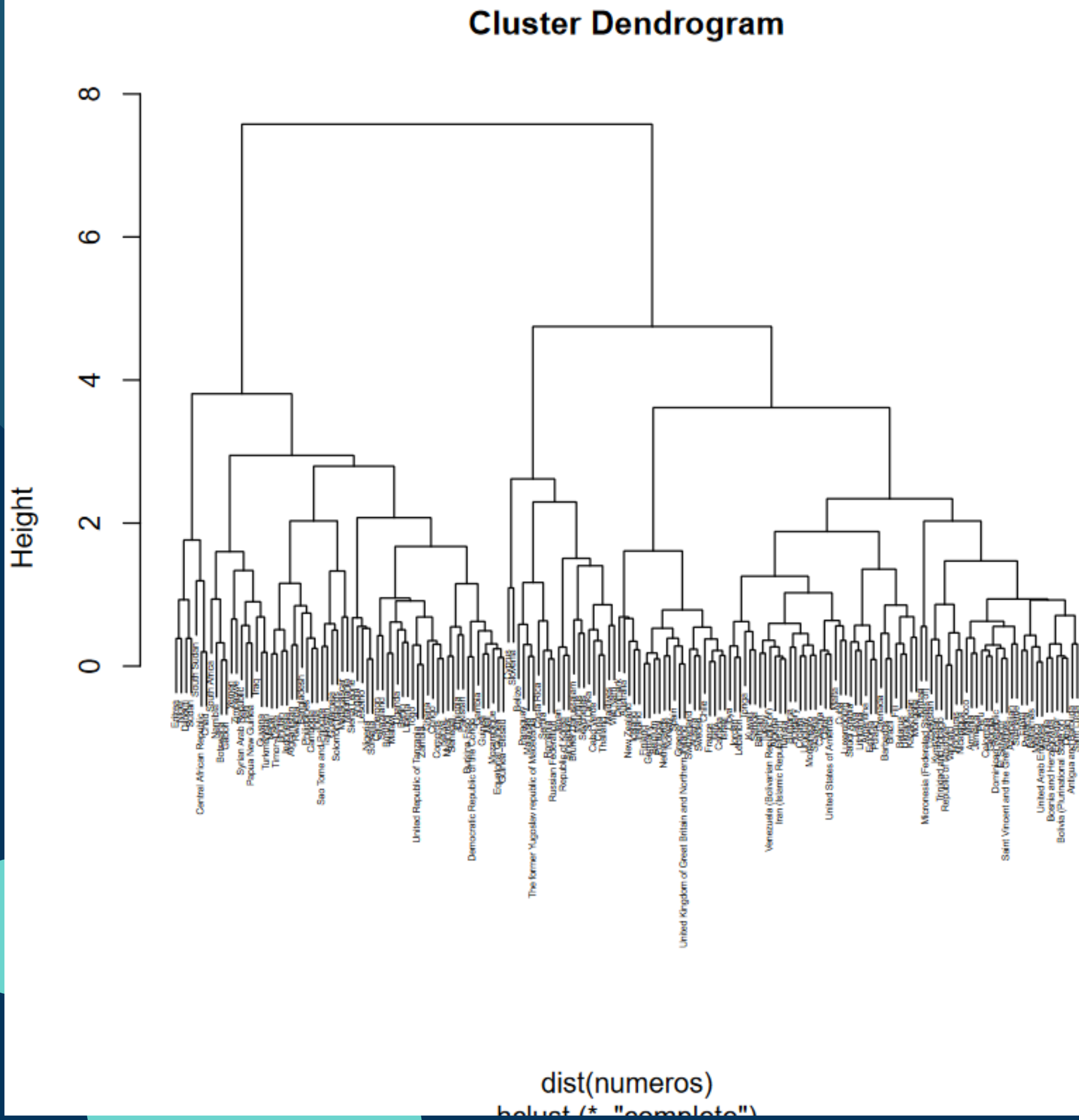
Detalhe: as  
cores  
invertem  
->

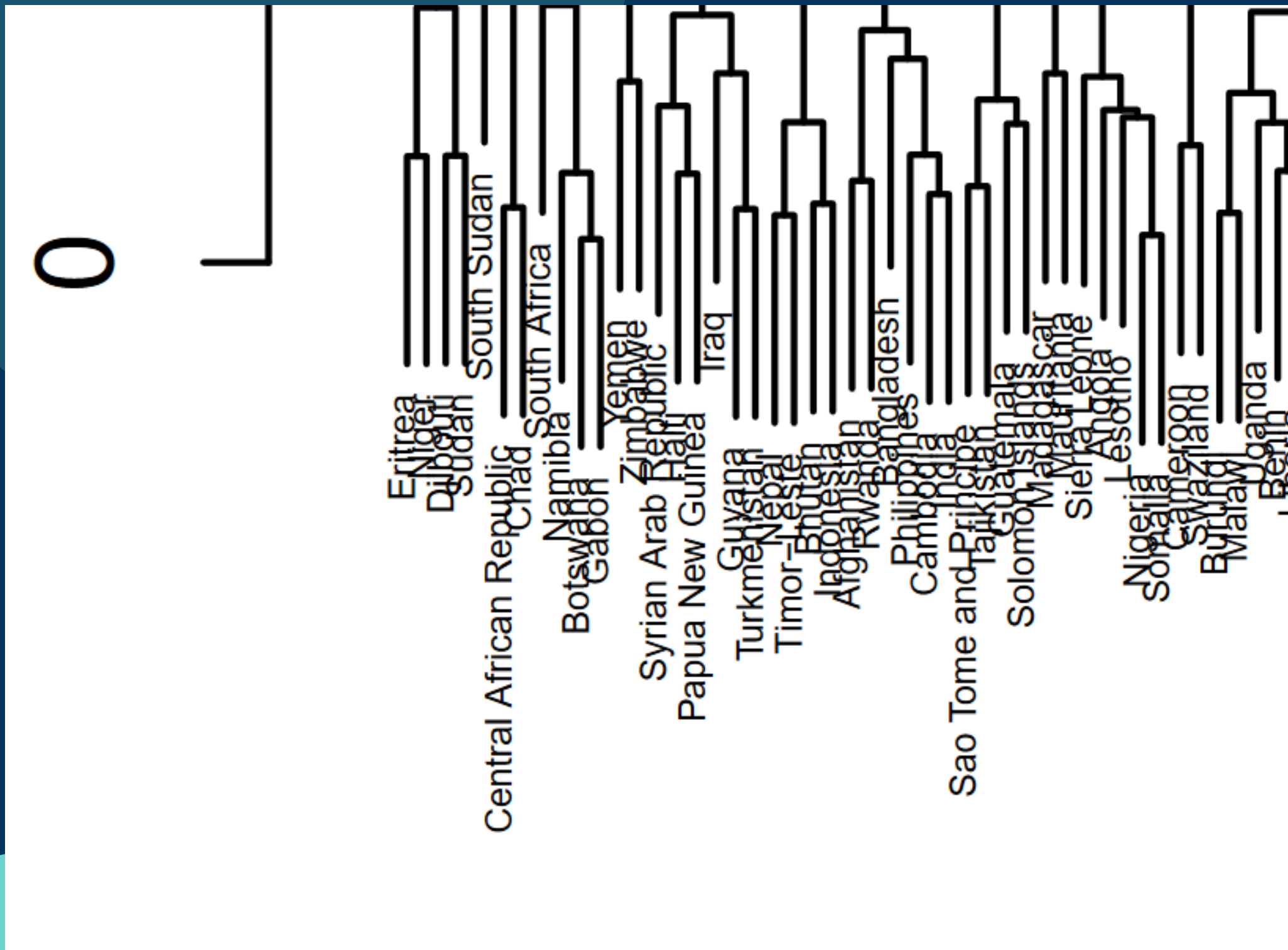


Classificação Real

Resultado  
Hierarquico

# DENDROGRAMA





**Como são +190  
países, é  
necessário  
bastante zoom  
para ver os  
nomes.**



# COMPARAÇÃO



```
# A semelhança entre os modelos  
# a ordem das cores padrão é 1-black, 2-red  
cm <- as.matrix(table(Actual = valores_relevantes[,c(3)], Predicted = cutree(modelo, 2)))  
  
#precisão invertida pois as cores estão invertidas  
accuracy <- 1 - (sum(diag(cm)) / sum(cm))  
  
print(accuracy)
```

Nos mede **54,44%** de semelhança entre os clusters e o  
real

```
[1] 0.5444444
```

# MÉTODO NÃO HIERÁRQUICO

NOS SLIDES A SEGUIR

# APLICAÇÃO K-MEANS



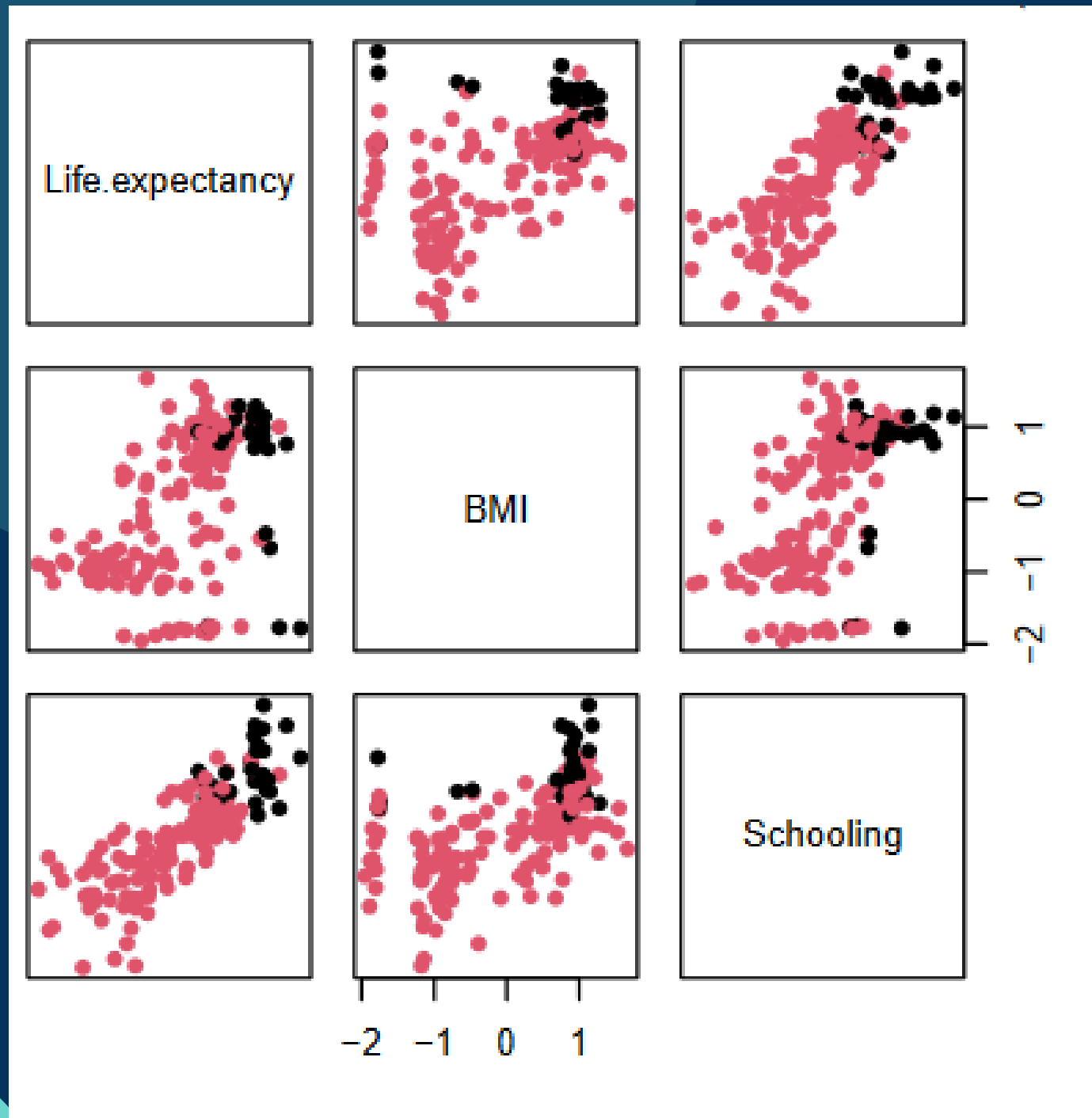
```
#agora fazendo o mesmo, com kmeans
```

```
modelo2 <- kmeans(na.omit(dist(numeros)), 2, algorithm="Hartigan-Wong")
```

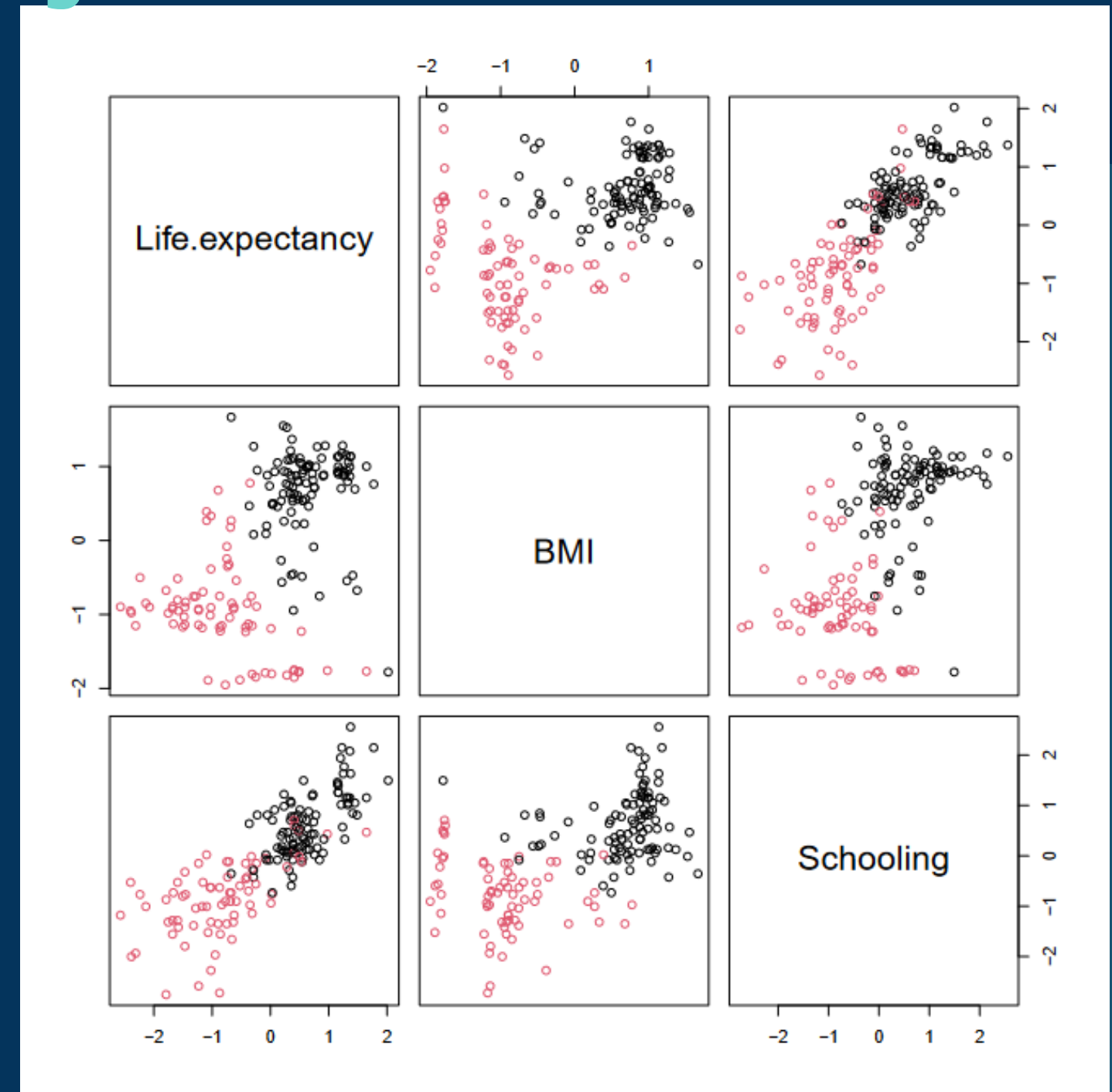
```
plot(numeros, col = modelo2$cluster)
```

```
points(modelo2$center, col=1:2, pch=8, cex=1)|
```

# COMPARAÇÃO



Classificação Real



Resultado do  
K-Means

# COMPARAÇÃO

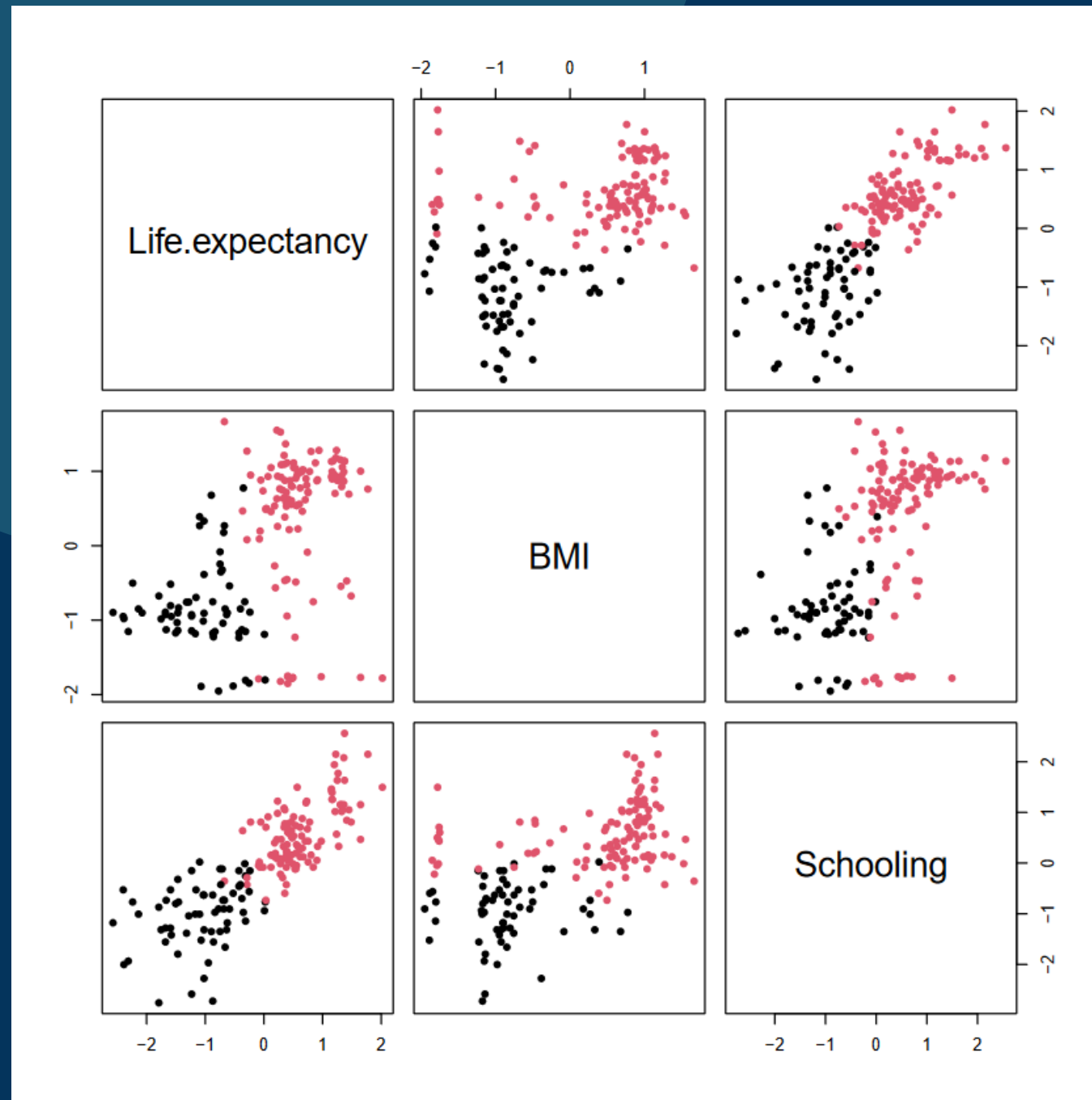


```
# semelhança entre os modelos v2  
cm <- as.matrix(table(Actual = valores_relevantes[,c(3)], Predicted = modelo2$cluster))  
  
#sem inversão, porém pode ser necessário dependendo de quando você roda o algoritmo  
accuracy <- sum(diag(cm)) / sum(cm)  
print(accuracy)
```

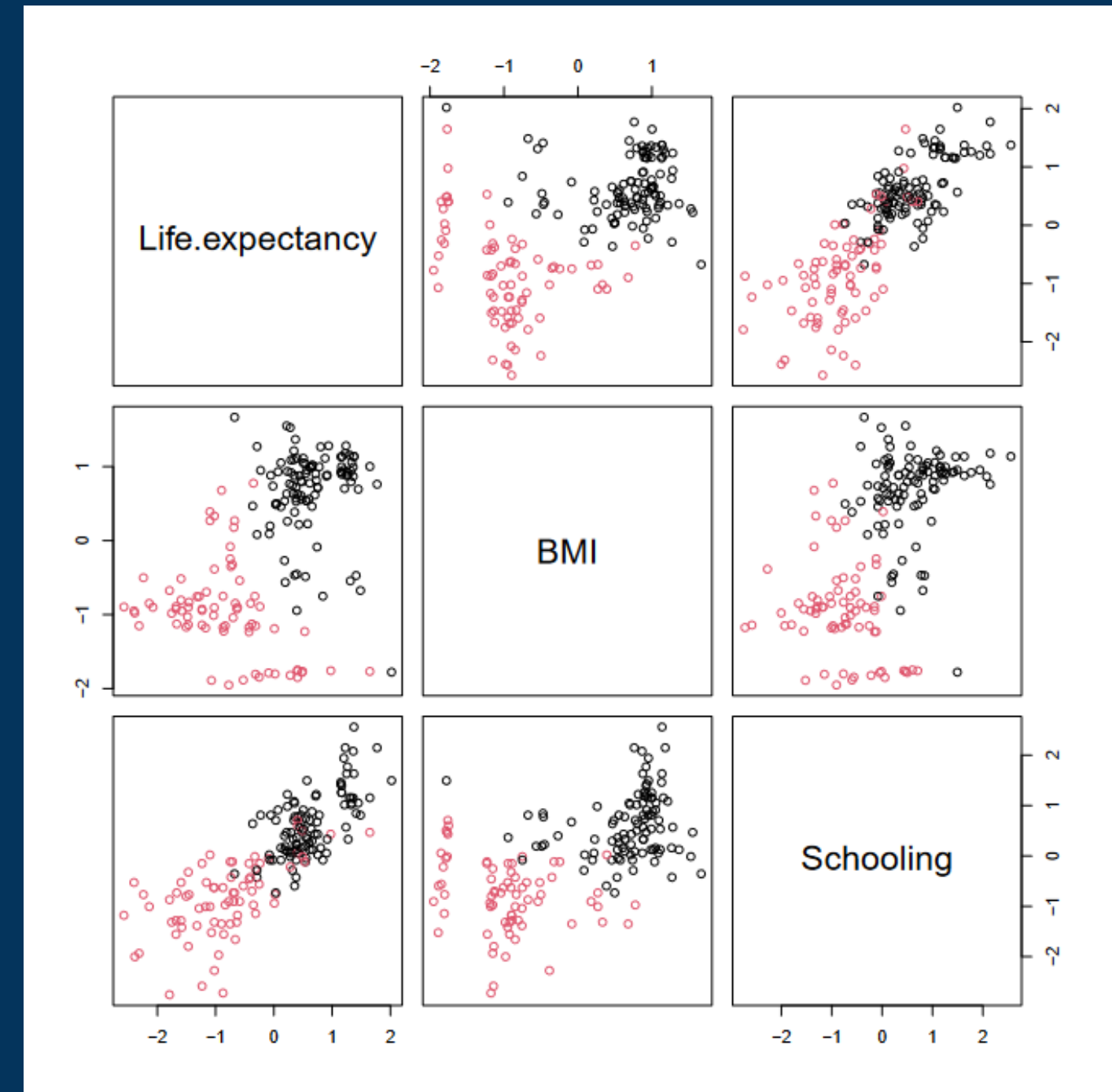
Nos mede **57,77%** de semelhança entre os clusters e o  
real

```
[1] 0.57777778
```

# COMPARAÇÃO ENTRE OS DOIS



Resultado Hierárquico



Resultado do K-Means

# COMPARAÇÃO ENTRE OS DOIS



```
# semelhança entre os modelos v3
```

```
cm <- as.matrix(table(Actual = cutree(modelo, 2), Predicted = modelo2$cluster))
```

```
#pode ser necessário inverter, pode não, devido a aleatoriedade do kmeans de escolher as cores.
```

```
accuracy <- sum(diag(cm)) / sum(cm)
```

```
print(accuracy)
```

Os modelos são **94,44%** iguais entre si  
(em uma dada execução)

```
[1] 0.9444444
```

# CONCLUSÕES

RESULTADOS DAS ANÁLISES E CONCLUSÕES DAS HIPÓTESES



# CONCLUIMOS QUE:

O agrupamento dos países nas classes de desenvolvido e subdesenvolvido é majoritariamente explicado pela expectativa de vida, educação e PIB.

Com dados normalizados, parecemos nos aproximar do estipulado, o que significa que a ONU implementa algo similar para sua própria classificação.



# REFERÊNCIAS BIBLIOGRÁFICAS

## LINKS

42

**GLEN NOAH**, DATA CLUSTER: DEFINITION, EXAMPLE, & CLUSTER ANALYSIS

[HTTPS://ANALYSTANSWERS.COM/DATA-CLUSTER-DEFINITION-EXAMPLE-CLUSTER-ANALYSIS/#DATA\\_CLUSTER\\_DEFINITION](https://ANALYSTANSWERS.COM/DATA-CLUSTER-DEFINITION-EXAMPLE-CLUSTER-ANALYSIS/#DATA_CLUSTER_DEFINITION), ACESSO EM 08/11/22

**WIKIPEDIA** ET AL. CLUSTER ANALYSIS

[HTTPS://EN.WIKIPEDIA.ORG/WIKI/CLUSTER\\_ANALYSIS](https://EN.WIKIPEDIA.ORG/WIKI/CLUSTER_ANALYSIS) ACESSO EM 08/11/22

**BONTHU HARIKA**, UNDERSTANDING KMEANS CLUSTERING FOR DATA SCIENCE BEGINNERS

[HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2021/08/KMEANS-CLUSTERING/](https://WWW.ANALYTICSVIDHYA.COM/BLOG/2021/08/KMEANS-CLUSTERING/) ACESSO EM 08/11/22

**ARAVIND CR.** EXPLORING CLUSTERING ALGORITHMS: EXPLANATION AND USE CASES

[HTTPS://NEPTUNE.AI/BLOG/CLUSTERING-ALGORITHMS](https://NEPTUNE.AI/BLOG/CLUSTERING-ALGORITHMS) ACESSO EM 08/11/22

**CIENCIA NA ILUMEO**, COMO FUNCIONA A ANÁLISE DE CLUSTER?

[HTTPS://ILUMEO.COM.BR/TODOS-POSTS/2019/04/03/COMO-FUNCIONA-A-ANALISE-DE-CLUSTER](https://ILUMEO.COM.BR/TODOS-POSTS/2019/04/03/COMO-FUNCIONA-A-ANALISE-DE-CLUSTER)  
ACESSO EM 08/11/22

**CHIRE WIKIPEDIA** K-MEANS CLUSTERING [HTTPS://EN.WIKIPEDIA.ORG/WIKI/K-](https://EN.WIKIPEDIA.ORG/WIKI/K-MEANS_CLUSTERING)

[MEANS\\_CLUSTERING#/MEDIA/FILE:K-MEANS\\_CONVERGENCE.GIF](https://EN.WIKIPEDIA.ORG/WIKI/K-MEANS_CLUSTERING#/MEDIA/FILE:K-MEANS_CONVERGENCE.GIF) ACESSO EM 08/11/22



**OBRIKADO!**