**FOLHA DE RESPOSTAS**

**Análise multivariada de dados para tomada de decisões**

Com base nos conteúdos aprendidos no curso, desenvolva os exercícios a seguir:

**PARTE 1 – REGRESSÃO LOGÍSTICA**

Analisando a idade mínima para aposentadoria no Brasil

Fonte: IPEA, Nota Técnica sobre Reforma da Previdência e Mercado de Trabalho. Abril de 2017.

Essa Nota Técnica do IPEA pretendeu demonstrar se há, no Brasil, uma maior probabilidade de desemprego para pessoas na faixa etária de 16 a 54 anos vis-à-vis aquelas de 55 a 64 anos. Para tanto, estimou-se uma regressão logística binária em que os desempregados eram 1 e os ocupados eram 0. As variáveis independentes, também binárias, foram:

a) sexo, sendo 0 para homens e 1 para mulheres;

b) idade, sendo 0 para pessoas de 16 a 54 anos e 1 para aquelas de 55 a 64 anos;

c) posição da família, sendo 0 para chefes e cônjuges e 1 para outras posições;

d) região do país, sendo 0 para sul e centro-oeste e 1 para demais regiões do país;

e) educação, sendo 0 para ensino superior e 1 para demais formações.

Os resultados da referida regressão encontram-se a seguir.

Regressão Logística Binária - Probabilidade de Desemprego Brasil - PNAD/IBGE de 2015

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variável** | **B** | **S.E** | **Wald** | **df** | **Sig.** | **Exp(B)** |
| Sexo | 0,586 | 0,001 | 699759,207 | 1 | 0,000 | 1,797 |
| Idade | -0,813 | 0,002 | 232421,205 | 1 | 0,000 | 0,444 |
| Posição na Família | 1,118 | 0,001 | 2514602,206 | 1 | 0,000 | 3,057 |
| Escolaridade | 0,882 | 0,001 | 481047,737 | 1 | 0,000 | 2,415 |
| UF/Região | 0,376 | 0,001 | 172036,452 | 1 | 0,000 | 1,456 |
| Constante | -3,948 | 0,002 | 6505749,103 | 1 | 0,000 | 0,019 |

Fonte: Elaboração a partir dos micro dados da PNAD/IBGE de 2015

Responda aos seguintes questionamentos:

1. Todas as variáveis independentes foram estatisticamente significantes ao nível de 5%?

**R** – Sim, todas elas estão abaixo de 5% do p-value, isto é, dentro de uma significância acima ou igual a 95%.

1. Qual é a interpretação do sinal de cada uma das variáveis independentes?

**R** – Para a variável sexo que está com o valor positivo, a mulher tem mais probabilidade de estar desempregada do que os homens, de acordo com este modelo.

Para idade com o coeficiente negativo, pessoas entre 16 a 54 anos tem mais chances de ficar desempregados do que pessoas entre 55 a 64 anos.

Já para posição na família, o coeficiente positivo nos diz que os chefes e cônjuges tem menos chances de ficar desempregados do que de outras posições.

No caso de escolaridade, o valor positivo indica que as pessoas com outras formações não sendo formação superior tende a ter mais chances de ficar desempregado, do que os formados, isto é, as pessoas com formação superior tendem a estar menos desempregada.

De acordo com o modelo, o valor positivo da variável indica que as outras regiões do país têm mais probabilidade de terem desempregados, do que as regiões sul e centro-oeste.

1. Qual seria o cenário encontrado (conjunto de variáveis independentes) de maior probabilidade de desemprego?

**R** – As mulheres entre 16 a 54 anos, possivelmente não casadas e não sendo chefe de família, com baixa formação escolar, no caso, sem formação superior e em grande parte da região brasileira, que contempla: sudeste, norte e nordeste tem maior probabilidade de desemprego.

**PARTE 2 – ANÁLISE DE CONGLOMERADOS**

Determinado gestor da área de RH ouviu falar, em um noticiário da internet, sobre a técnica

K-Means (análise não hierárquica). Ele deseja avaliar a performance dos seus funcionários com base nas seguintes variáveis: pontualidade, relacionamento e resultados.

Como ele faria a realização dessa técnica no Software R? Digite os comandos necessários.

R – Foi realizado a análise do arquivo de Excel **Desempenho UN4.xlsx** e nela continha linhas em branco do qual foi ajustada dentro do R. Os procedimentos serão descritos abaixo:

Base de dados Original:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pessoas | Pontualidade | Produtividade | Lideranca | Relacionamento |
|  |  |  |  |  |
| 1 | 9 | 8 | 5 | 6 |
|  |  |  |  |  |
| 2 | 8 | 9 | 4 | 4.5 |
|  |  |  |  |  |
| 3 | 9.5 | 8 | 5.5 | 4 |
|  |  |  |  |  |
| 4 | 7 | 6 | 3 | 4.5 |
|  |  |  |  |  |
| 5 | 6 | 5 | 2 | 2 |
|  |  |  |  |  |
| 6 | 3 | 4 | 2 | 1 |
|  |  |  |  |  |
| 7 | 4 | 5 | 2 | 2 |
|  |  |  |  |  |
| 8 | 5 | 6 | 1 | 2.5 |
|  |  |  |  |  |
| 9 | 8 | 7 | 4 | 3.5 |
|  |  |  |  |  |
| 10 | 7 | 7 | 3 | 1.5 |
|  |  |  |  |  |
| 11 | 3 | 4 | 3 | 2.5 |
|  |  |  |  |  |
| 12 | 6 | 7 | 2 | 1 |
|  |  |  |  |  |
| 13 | 4 | 6 | 1 | 2.5 |
|  |  |  |  |  |
| 14 | 9.5 | 8 | 5.5 | 6.5 |
|  |  |  |  |  |
| 15 | 7 | 6 | 3 | 2.5 |
|  |  |  |  |  |
| 16 | 6 | 7 | 2 | 1.5 |
|  |  |  |  |  |
| 17 | 10 | 8 | 6 | 5 |
|  |  |  |  |  |
| 18 | 9 | 9 | 5 | 5 |
|  |  |  |  |  |
| 19 | 3 | 4.5 | 1 | 1.5 |
|  |  |  |  |  |
| 20 | 4 | 5.5 | 1 | 2.5 |
|  |  |  |  |  |
| 21 | 8 | 8.5 | 4 | 6.5 |

Ações dentro do R Studios:

# cluster analysis

# import dataset

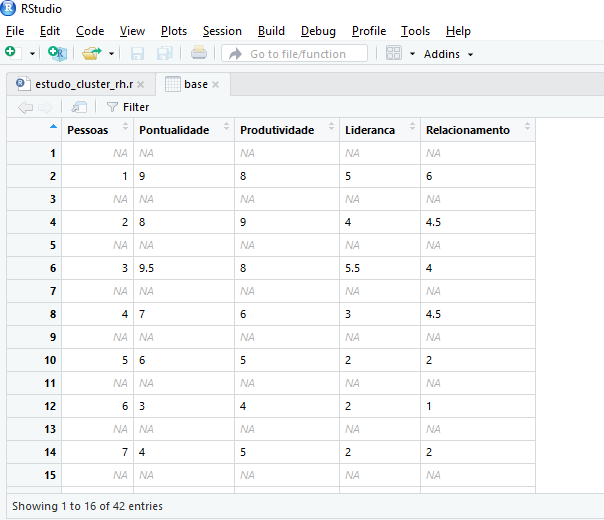
install.packages('pvclust')

library(cluster)

library(readxl)

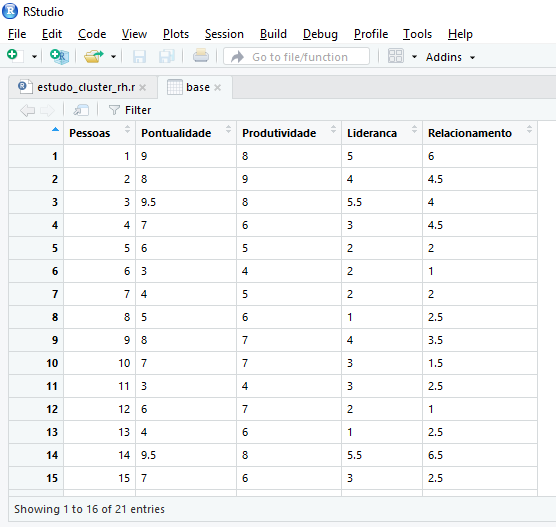
#Base carregada

base <- read\_excel("D:/cursos/Análise multivariada de dados para tomada de decisões/Desempenho UN4.xlsx")



# Retirada de todas as linhas NA

base <- base[complete.cases(base),]



# values of data set (mydata)

mydata <- base[, 2:5]

# Elimina variável não esperada no escopo da análise

mydata$Lideranca <- NULL

# Houve um comportamento não esperado que a função

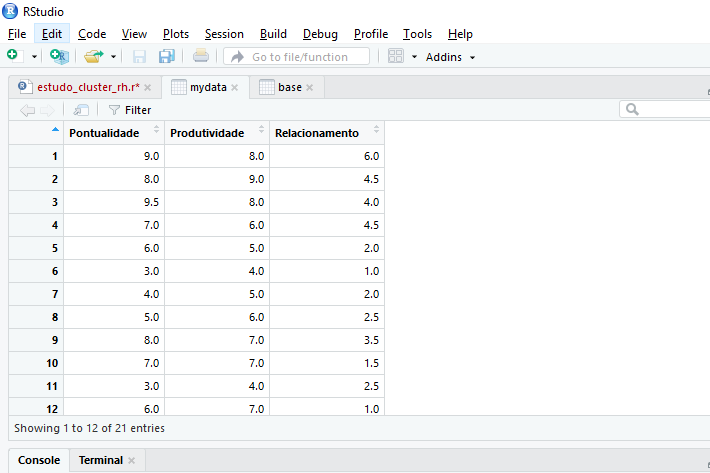
# scale() deixou de funcionar conforme o esperado, tive que transformar

# em numérico para poder realizar a padronização dos dados.

mydata$Pontualidade <- as.numeric(mydata$Pontualidade)

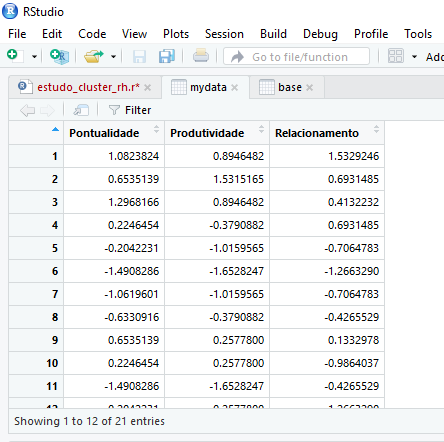
mydata$Produtividade <- as.numeric(mydata$Produtividade)

mydata$Relacionamento <- as.numeric(mydata$Relacionamento)



#scaled matrix (zscores)

mydata <- scale(mydata)



# Distância calculada para todas as variáveis

# Distance matrix by Euclidean

d <- dist(mydata, method = 'euclidean');d

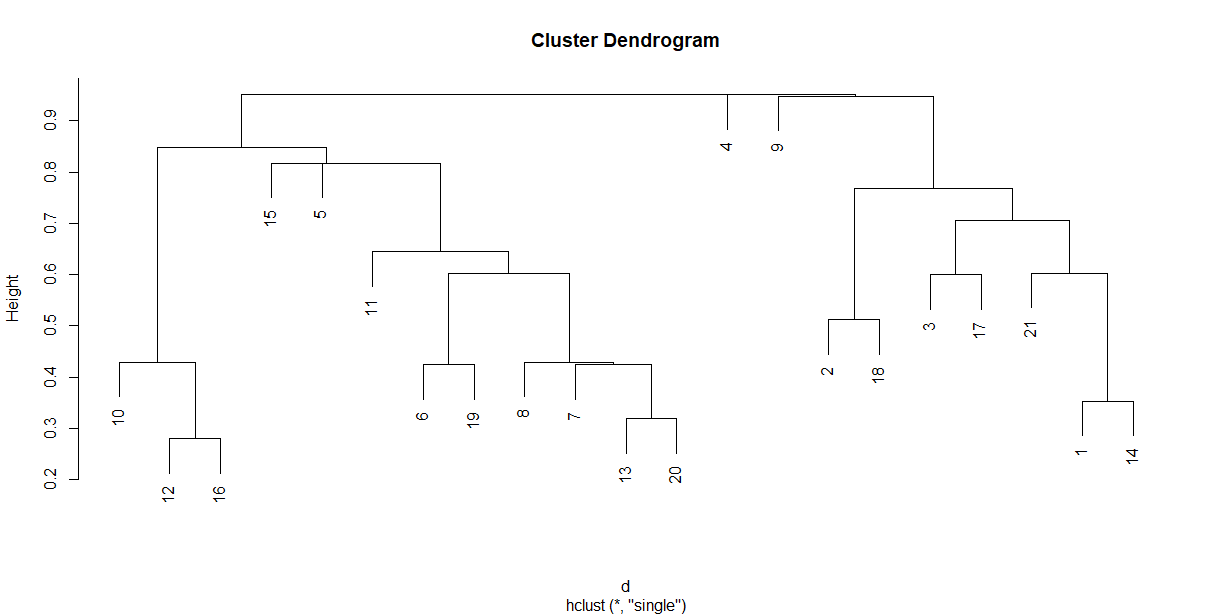
# Hierarchial Method - Single

# Calcula os clusters hierárquico de acordo com a distância definida

fit <- hclust(d, method = 'single');fit

# plotando um gráfico dendrogram

plot(fit)



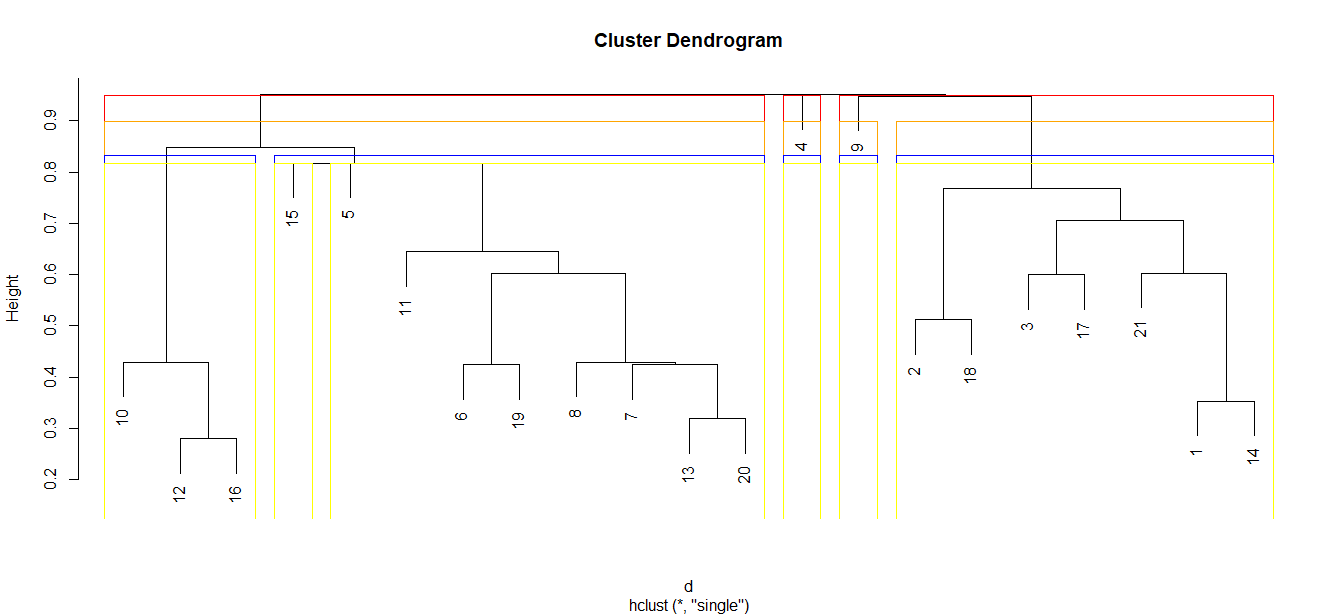
# Definindo visualmente os grupos

rect.hclust(fit, k = 3, border = 'red')

rect.hclust(fit, k = 4, border = 'orange')

rect.hclust(fit, k = 5, border = 'blue')

rect.hclust(fit, k = 6, border = 'yellow')



# A quantidade de cluster ideal para este modelo é **3 em vermelho** de acordo com o dendrograma analisado.

# k-means com 3 clusters - validation and interpretation

# K-means realiza cluster não hierárquicos

fit <- kmeans(mydata, 3);fit

Resultado da implementação do cluster com uma acuracidade de 81,8%. Podemos observar que o grupo 3 tem médias de performance mais alta em relação aos outros grupos. Enquanto o grupo 1 tem a pior performance esperada.

K-means clustering with 3 clusters of sizes 6, 8, 7

Cluster means:

Pontualidade Produtividade Relacionamento

1 -1.27639431 -1.1221012 -0.7064783

2 0.01021115 -0.1402626 -0.4965343

3 1.08238238 1.1221012 1.1730206

Clustering vector:

[1] 3 3 3 2 2 1 1 2 2 2 1 2 1 3 2 2 3 3 1 1 3

Within cluster sum of squares by cluster:

[1] 2.254762 5.613706 3.030028

(between\_SS / total\_SS = 81.8 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"

[6] "betweenss" "size" "iter" "ifault"