**A3 - INTELIGÊNCIA ARTIFICIAL**

**PROFA. MSC EDMILA MONTEZANI**

O banco de dados chamado AJAXX possui as seguintes informações:

1. Nome do indivíduo, que na verdade está enumerado de 1 até 220

2. Salário

3. Posição (grade – usado em RH – para estabelecer faixas salariais)

4. Anos de experiência na função

5. Sexo

|  |  |
| --- | --- |
| **Alunas** | **RA** |
| Andressa Emily Rabêlo Pereira | 823213904 |
| Camila Loranne de França Martins | 825118841 |
| Mylena Soares Rocha | 824144075 |

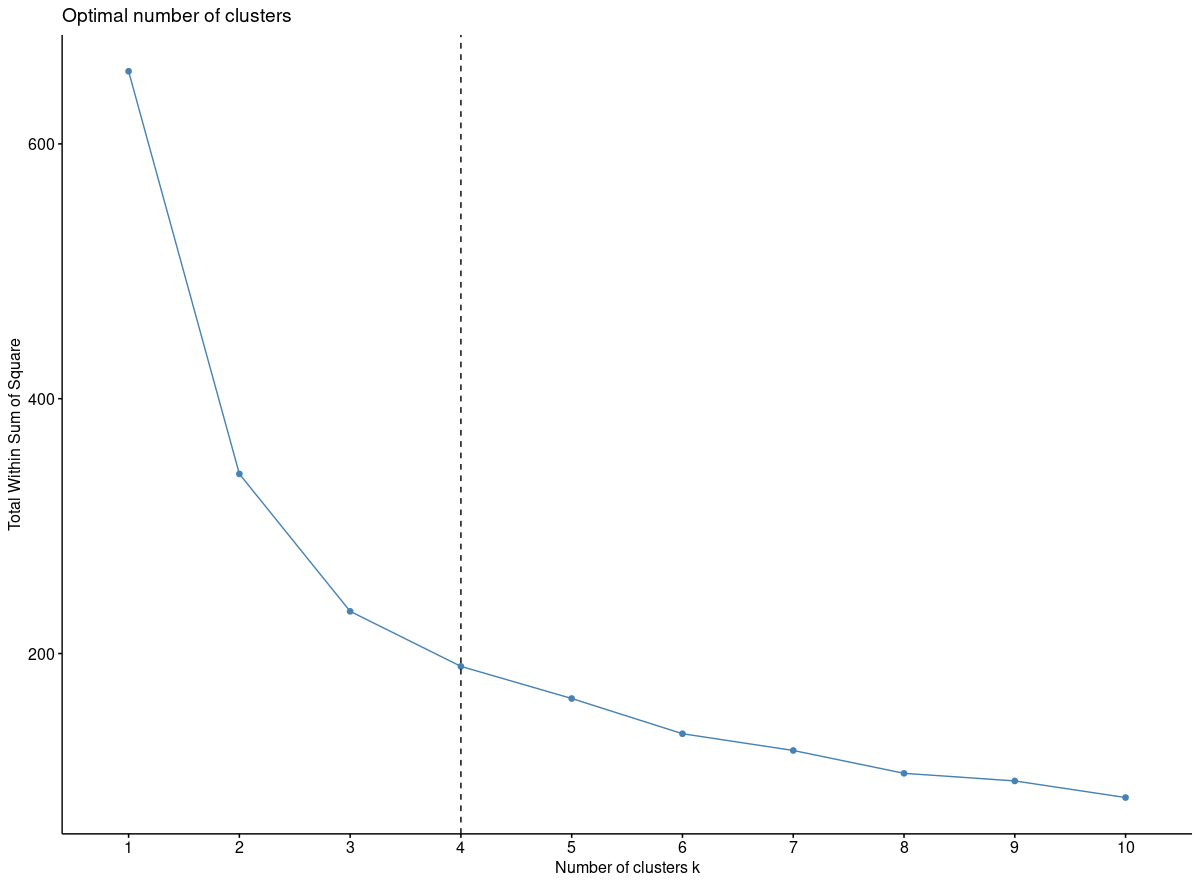
**A sua atividade A3 deverá ser:**

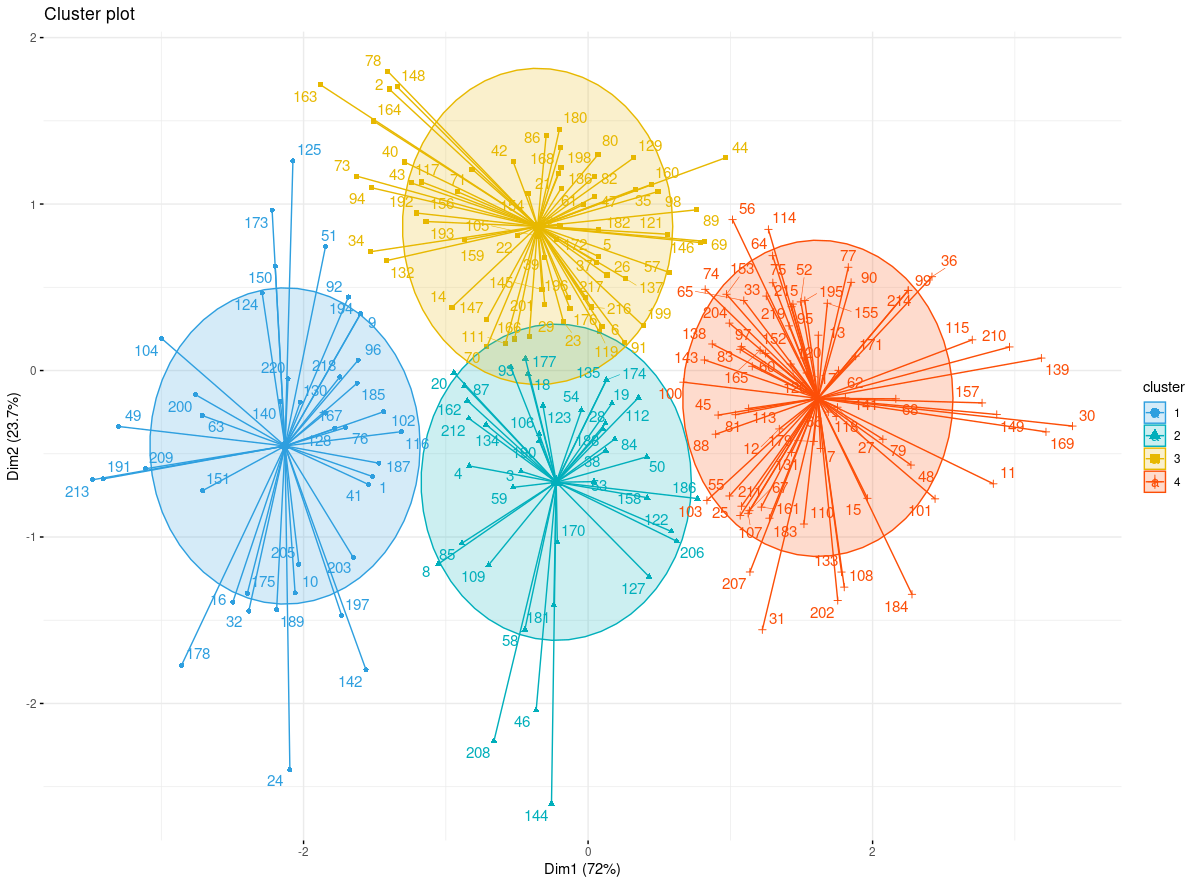
Uma análise de cluster, considerando todas as variáveis e sendo capaz de responder as seguintes perguntas:

1. **Existem quantos clusters diferentes para esse exemplo?**

**R:** Existem 4 clusters.

*Com base na análise do gráfico, determinou-se que 4 clusters é a quantidade ideal.*

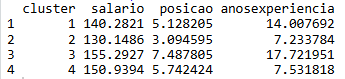




1. **Qual a média salarial de cada cluster?**

**R:** Foi realizada uma clusterização com 4 grupos usando as variáveis: salário, posição e anos de experiência.

**Médias por Cluster:**



* **Cluster 1:** Profissionais com média de 14 anos de experiência, posição intermediária (5,13) e salário médio de 140,28.

*Representa um grupo experiente, mas com remuneração e cargo medianos.*

* **Cluster 2:** (Menor média em todos os aspectos) 7,23 anos de experiência, posição 3,08 e salário de 130,15.

*Perfil de profissionais iniciantes ou em cargos mais baixos.*

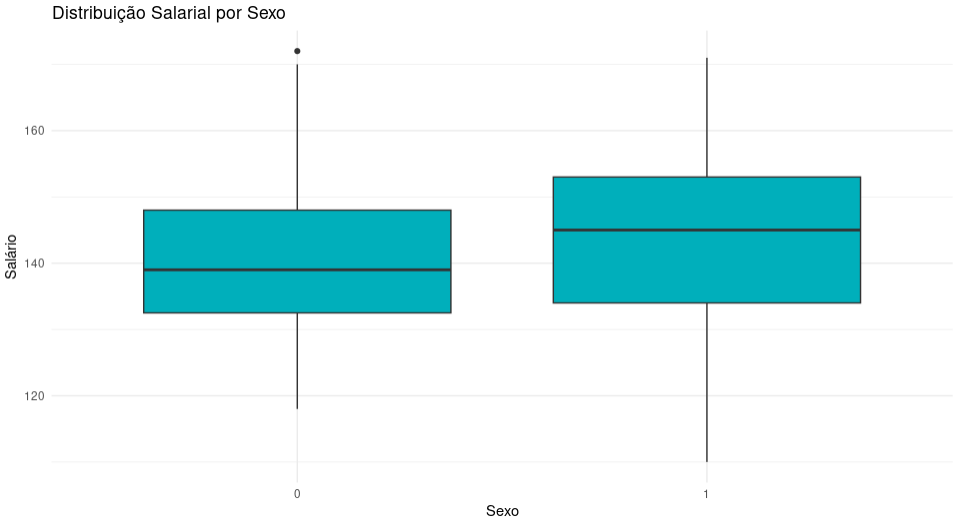
* **Cluster 3:** Grupo mais sênior, com 17,72 anos de experiência, posição elevada (7,49) e salário mais alto (155,29).
* **Cluster 4:** Apesar de ter pouca experiência (7,53 anos), apresenta posição relativamente alta (5,74) e salário de 150,94.

*Pode indicar rápida ascensão ou valorização por outros fatores.*

| **Cluster** | **Média Salarial** |
| --- | --- |
| (3) | R$155.29 |
| (1) | R$140.28 |
| (4) | R$150.94 |
| (2) | R$130.15 |

1. **Existe alguma relação entre o sexo e o salário desses indivíduos? Analise e Explique.**

**R:** Foi realizada uma análise comparativa de salário por sexo



A análise sugere que existe uma **diferença leve na média salarial entre os sexos**, com o grupo 1 (ex.: homens) recebendo um pouco mais em média do que o grupo 0 (ex.: mulheres).

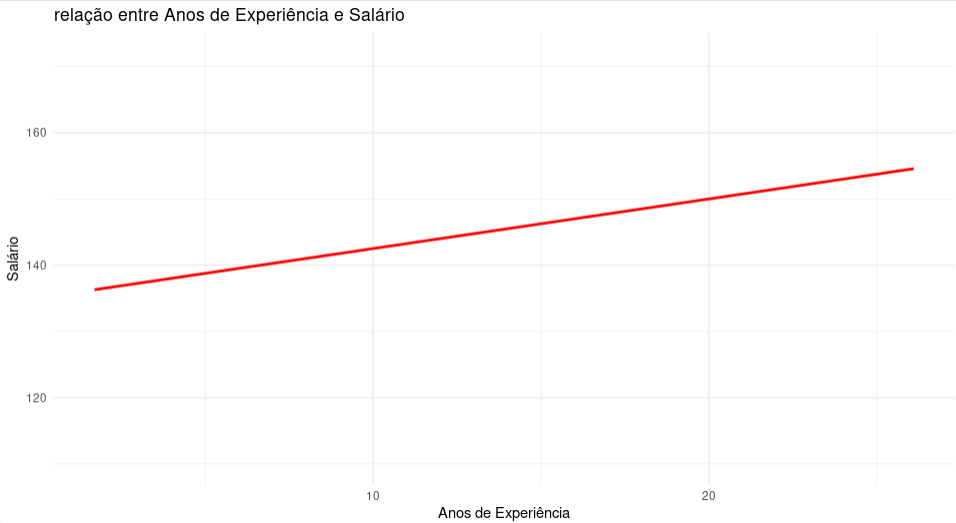
| **Sexo** | **Média Salarial por Sexo** |
| --- | --- |
| 0 | R$140.47 |
| 1 | R$144.11 |

1. **Se analisarmos os anos de experiência na função, é possível afirmar que quanto maior o tempo na função, maior o salário? Analise e Explique.**

**R:** Foi calculada a correlação de Pearson:

*Valor da correlação: r = 0.31 (correlação positiva – moderada)*

Existe uma correlação positiva moderada, o que significa que, de modo geral, quanto mais experiência o indivíduo possui, maior tende a ser seu salário. Entretanto, essa relação não é muito forte. De modo a sugerir que outros fatores além da experiência também influenciam significativamente os salários dos indivíduos.



**Código R**

**# Instalar pacotes**

install.packages("ggplot2", dependencies = TRUE)

install.packages("factoextra", dependencies = TRUE)

**# Carregar bibliotecas**

library(factoextra)

library(ggplot2)

library(readxl)

**# 1. Leitura do arquivo Excel**

EXER\_A3 <- read\_excel("EXER\_A3.xlsx")

**# 2. Padronização dos dados**

df <- scale(EXER\_A3)

head(df, n=3)

**# 3. Reduzindo para variáveis selecionadas**

df\_reduced <- scale(EXER\_A3[, c("salario", "posicao", "anosexperiencia")])

**# PERGUNTA A:**

**# 4. Gráfico do cotovelo com dados reduzidos**

fviz\_nbclust(df\_reduced, kmeans, method = "wss") +

geom\_vline(xintercept = 4, linetype = 2)

**# 5. Clusterização k-means**

set.seed(123)

km.res = kmeans(df\_reduced, 4, nstart = 25)

print(km.res)

**# 6. Adicionando cluster ao dataset original**

EXER\_A3\_clustered <- cbind(EXER\_A3, cluster = km.res$cluster)

head(EXER\_A3\_clustered)

km.res$centers

**# PERGUNTA B:**

**# 7. Médias reais por cluster (com os dados originais)**

aggregate(EXER\_A3[, c("salario", "posicao", "anosexperiencia")],

by = list(cluster = km.res$cluster),

mean)

**# PERGUNTA C:**

**#Média salarial por sexo**

aggregate(salario ~ sexo, data = EXER\_A3, mean)

**#Visualização**

ggplot(EXER\_A3, aes(x = as.factor(sexo), y = salario)) +

geom\_boxplot(fill = "#00AFBB") +

labs(x = "Sexo", y = "Salário", title = "Distribuição Salarial por Sexo") +

theme\_minimal()

**# 8. Visualização gráfica dos clusters**

library(ggplot2)

library(factoextra)

fviz\_cluster(km.res, data = df\_reduced,

palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),

ellipse.type = "euclid",

star.plot = TRUE,

repel = TRUE,

ggtheme = theme\_minimal()

)

**#Calculo da matriz de distâncias entre os indivíduos - distância euclidiana, que é a distância “reta” entre dois pontos no espaço.**

dista=dist(df, method="euclidean")

**#Mostra as distâncias entre os 3 primeiros indivíduos.**

as.matrix(dista)[1:3,1:3]

**#Realiza a clusterização hierárquica com o método ward.D**

dista.hc=hclust(d=dista, method="ward.D")

**#representação gráfica:**

library("factoextra")

fviz\_dend(dista.hc, k = 4, rect = TRUE, cex = 0.5)

**# Teste Estatistico:**

t.test(salario ~ sexo, data = EXER\_A3)

**# PERGUNTA D:**

**# Correlação salário x Anos de experiência:**

**# Calcula o coeficiente de correlação de Pearson entre os anos de experiência e o salário.**

correlacao <- cor(EXER\_A3$salario, EXER\_A3$anosexperiencia)

print(paste("Correlação entre anos de experiência e salário", round(correlacao, 2)))

**#Gráfico de dispersão com linha de tendência:**

ggplot(EXER\_A3, aes(x = anosexperiencia, y = salario))+

geom\_point(color = "00AFBB") +

geom\_smooth(method ="lm", se = FALSE, color = "red")+

labs(x = "Anos de Experiência", y = "Salário",

title = "relação entre Anos de Experiência e Salário")+

theme\_minimal()