Analise de funcionários

Thiago Silva

01/05/2021

Este conjunto de dados é uma base ficticia do quadro de funcionários de uma empresa.

O objetivo é aplicar o algorítimo de agrupamento K-means para separar os individuos da base de dados em diferentes grupos com base nas suas caracteristicas.

1.Sobre a base de dados

O nosso conjunto de dados possui informações de 220 individuos e suas informações de salário, anos de experiência, posição na empresa e sexo.

2. Importando os arquivos

Carregando a base de dados e ajustando o nome das colunas.

```
BD<- read.csv(file.choose(), na.strings = "", sep = ";")

colnames(BD)
```

```
## [1] "ï..indivÃ.duo"   "salario"         "posicao"         "anosexperiencia"
## [5] "sexo"
```

```
colnames(BD) <- c("individuo", "salario", "posicao", "anosexperiencia", "sexo")
head(BD)
```

```
##   individuo salario posicao anosexperiencia sexo
## 1         1     148       7            16,7    1
## 2         2     165       7             6,7    1
## 3         3     145       5            14,8    1
## 4         4     139       7            13,9    0
## 5         5     142       6             6,4    0
## 6         6     144       5             9,1    1
```

3. Explorando a base de dados

Checando se há valores faltantes na base de dados.

```
#CHECANDO PORCENTAGEM DE DADOS FALTANTES DE CADA VARIAVEL
NAs<- round(colSums(is.na(BD))*100/nrow(BD),2)
NAs
```

```
##       individuo         salario         posicao anosexperiencia           sexo
##               0               0               0               0              0
```

```
#CHECANDO SE EXISTE ALGUM DADO FALTANTE NA BASE
anyNA(BD)
```

```
## [1] FALSE
```

Neste caso, não possuimos nenhum valor faltante em nossa base de dados, não sendo necessário nenhum tipo de tratamento para este fim.

Explorando o conjunto para entender suas dimensões.

```
#CHECANDO DIMENSAO DO CONJUNTO DE DADOS
dim(BD)
```

```
## [1] 220   5
```

Temos 220 linhas e 6 colunas.

Identificando proporção de sexo distribuidos na base.

```
#Explorando dados
propsexo <- round(table(BD$sexo)*100/nrow(BD),2)
propsexo
```

```
##
##     0     1
## 34.09 65.91
```

Com isso, adotando que a variavel categórica 1 seja o sexo masculino e a variavel 0 o sexo feminino, podemos ver que a base está distribuida em 65,91% de individuos do sexo masculino e 34,09% de individuos do sexo feminino.

Podemos entender um pouco mais da base plotando correlações para identificarmos variaveis que possuem correlação.

```
#Visualizando correlacoes
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggcorr(BD, label=T)
```

```
## Warning in ggcorr(BD, label = T): data in column(s) 'anosexperiencia' are not
## numeric and were ignored
```

Nota-se que existe correlação da posição que o individuo tem na empresa com o salário que ele recebe, o que faz sentido dado que conforme a seu cargo evolui espera-se que você receba um salário maior.

4. Transformando a base de dados

Para darmos inicio a construção do algoritmo K-means excluiremos a variavel ID “individuos” que não nos será útil.

```
#Removendo variavel ID
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
BDclusters <- BD %>% select(-1)
```

Verificando o tipo de dado das colunas para sabermos se precisamos fazer algum tipo de tratamento.

```
#Verficando o tipo das colunas
str(BDclusters)
```

```
## 'data.frame':    220 obs. of  4 variables:
##  $ salario        : int  148 165 145 139 142 144 128 143 157 150 ...
##  $ posicao        : int  7 7 5 7 6 5 3 6 7 7 ...
##  $ anosexperiencia: Factor w/ 127 levels "1,7","10","10,1",..: 43 98 33 29 95 120 115 47 2
3 62 ...
##  $ sexo           : int  1 1 1 0 0 1 0 1 1 1 ...
```

Nota-se que as variaveis posicao, anos de experiencia e sexo são variaveis categóricas e foram identificadas como númericas, neste caso será necessário realizar um tratamento nestas, afim de transforma-las para o tipo fator.

Outro ponto importante é que a variavel “anos experiencia” está com as casas decimais separadas por virgula, neste caso será necessário substituir a virgula por ponto, conforme código abaixo:

```
#Substituindo virgula por ponto na variavel anosexperiencia e formatando para tipo numerico
anosexperiencia <- sapply(BDclusters, function(x) any(grepl(",", x)))
BDclusters$anosexperiencia <- sapply(BDclusters[,c("anosexperiencia")], function(x) as.numeri
c(sub(",", ".", x)))
#Verficando o tipo das colunas
str(BDclusters)
```

```
## 'data.frame':    220 obs. of  4 variables:
##  $ salario        : int  148 165 145 139 142 144 128 143 157 150 ...
##  $ posicao        : int  7 7 5 7 6 5 3 6 7 7 ...
##  $ anosexperiencia: num  16.7 6.7 14.8 13.9 6.4 9.1 8.5 18.2 13 21.6 ...
##  $ sexo           : int  1 1 1 0 0 1 0 1 1 1 ...
```

5. Criando o algoritmo k-means

Para que o modelo não seja viesado por conta de grandezas diferentes, realizaremos uma normalização de dados atráves da função scale.

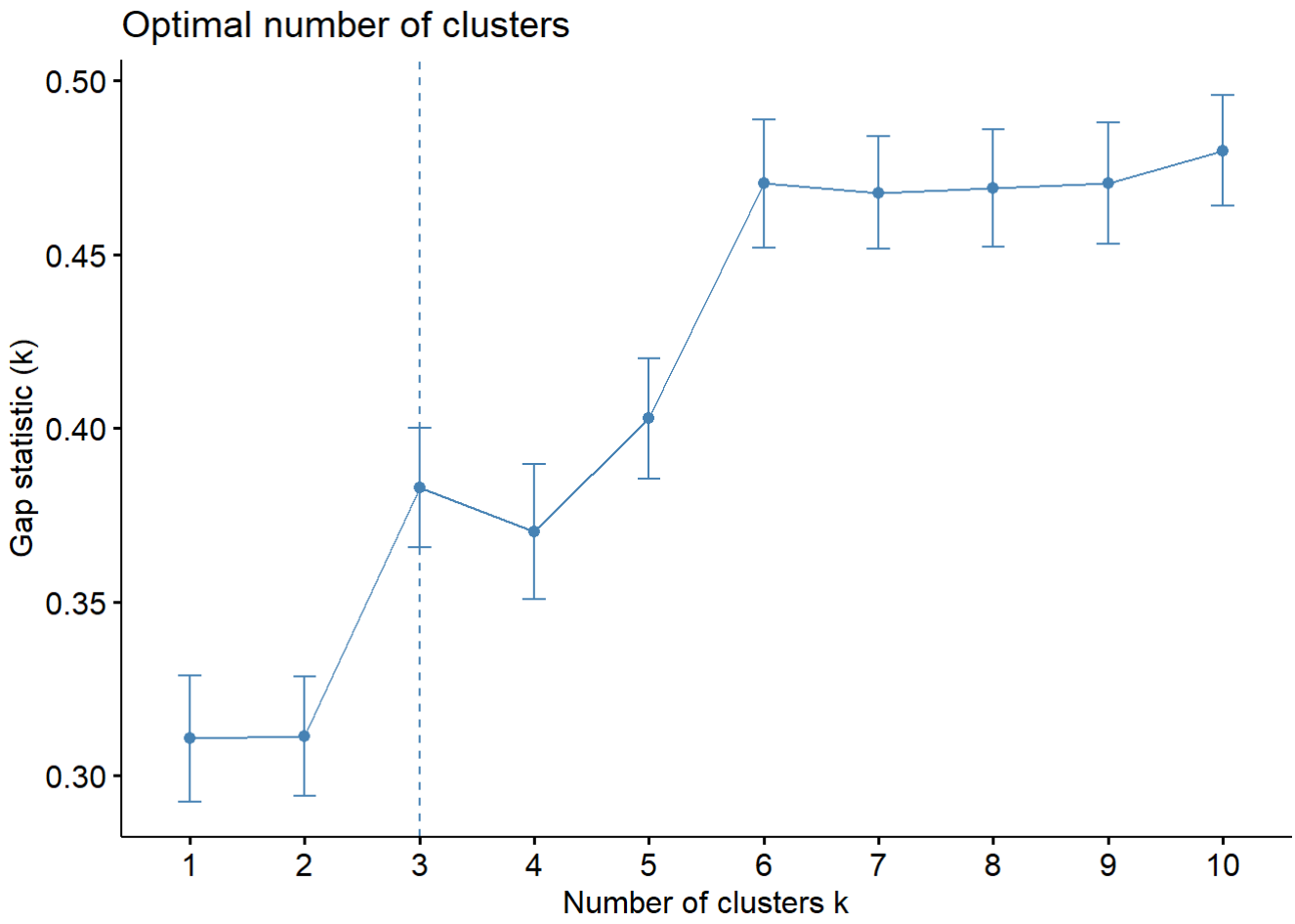
```
#NORMALIZANDO DADOS E ARMAZENANDO NA VARIAVEL "DADOS"
dados<- scale(BDclusters[,c(1:4)])
```

Através do método “Cotovelo” identificaremos a quantidade ideal de grupos para esse conjunto de dados.

```
#ENCONTRANDO O NUMERO IDEAL DE CLUSTERS
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
#EXECUTANDO ALGORITMO KMEANS E METODO GAP_STAT PARA ENCONTRAR O NUMERO IDEAL DE CLUSTERS
fviz_nbclust(dados, kmeans, method= "gap_stat")
```



Como vemos explicitamente acima, trabalharemos com 3 grupos diferentes.

Damos inicio a predição orientando o algoritmo k-means a separar o conjunto em 3 grupos diferentes.

```
library(caret)
```

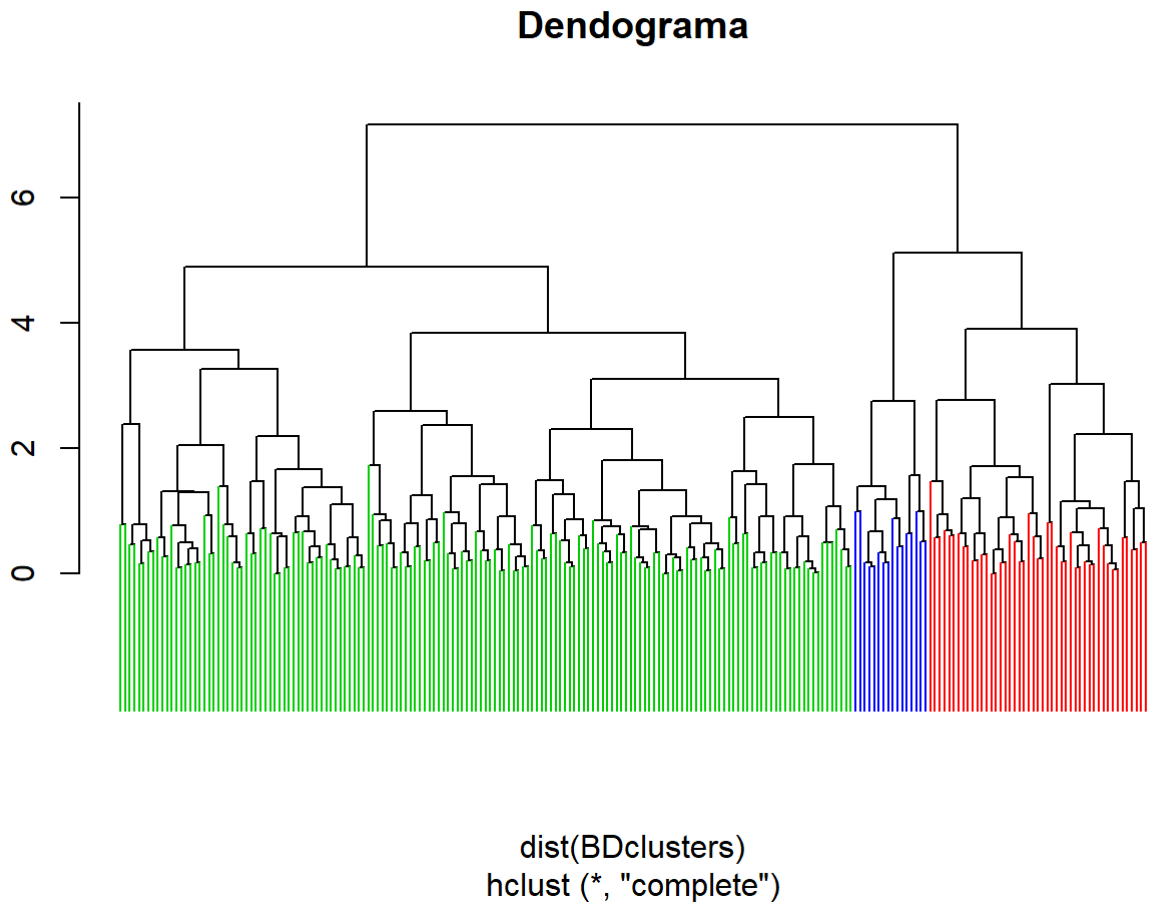
```
## Loading required package: lattice
```

```
#REALIZANDO PREDICAO COM ALGORITMO DE CLUSTERIZACAO E UTILIZANDO METODO DE NORMALIZACAO "SCAL
E"
BDclusters <- predict(preProcess(BDclusters, method ="scale") ,BDclusters)
#COMANDO PARA GARANTIR QUE O LEITOR CHEGUE AO MESMO RESULTADO
set.seed(1)
clusters<- kmeans(BDclusters, centers=3)
```

6. Explorando grupos

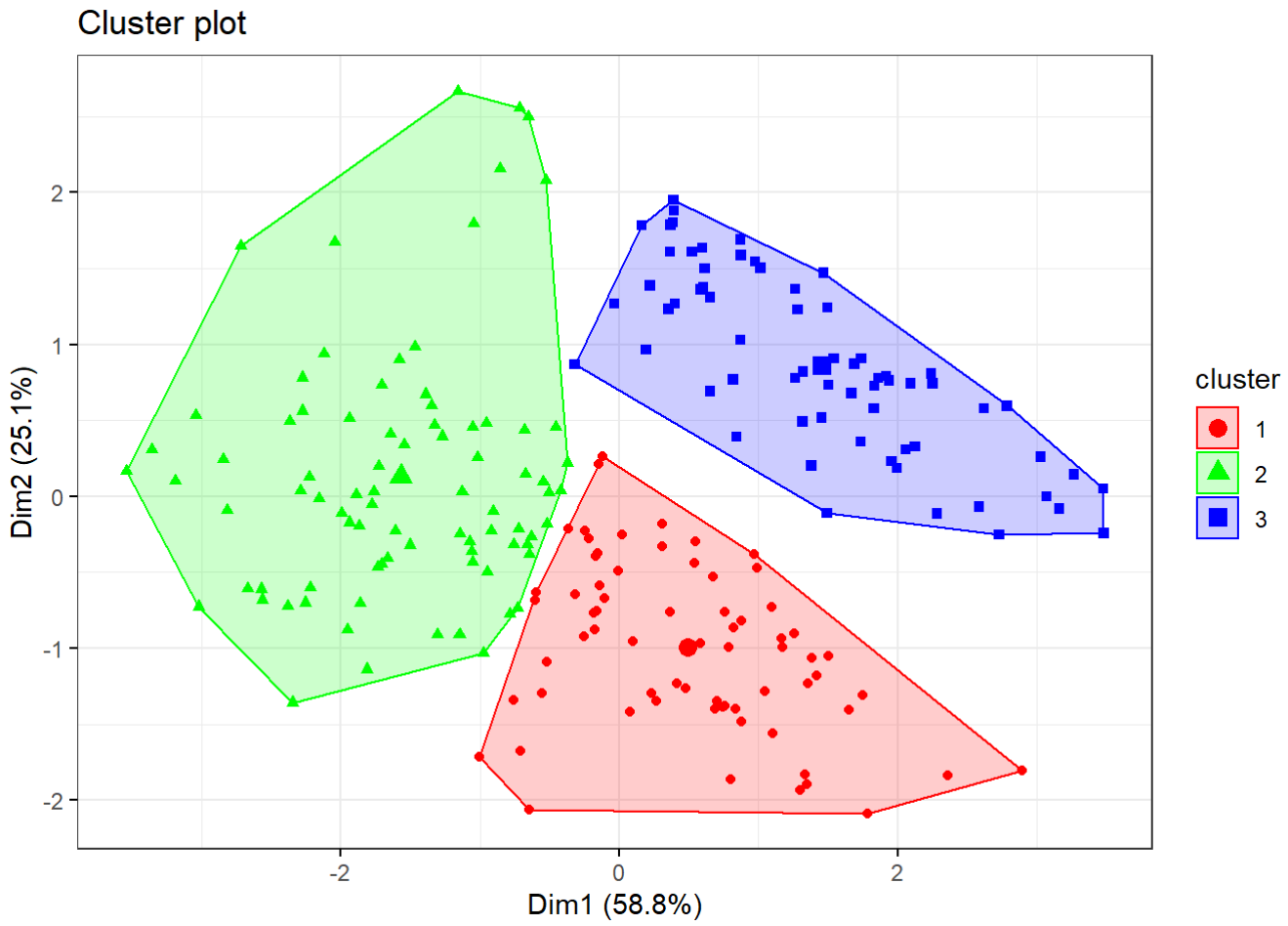
Após separação dos grupos, podemos visualizar o dendograma com os grupos separados por 3 cores diferentes.

```
dendograma <- hclust(dist(BDclusters))
#plot(dendograma)
#Dividindo o dendograma em 3 grupos com cores diferentes
y = cutree(dendograma, 3)
library(sparcl)
ColorDendrogram(dendograma, y = y, labels = names(y), main = "Dendograma", branchlength = 80)
```



Como visualização auxiliar, temos:

```
#Visualizando grupos
fviz_cluster(clusters, data = BDclusters,
             palette = c("red", "green", "blue"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
)
```



Podemos ver algumas estatisticas dos agrupamentos, como os clusteres os centróides e quantos individiuos foram alocados em cada cluster.

```
#CLUSTERS
clusters$cluster
```

```
##   [1] 2 2 2 3 3 1 3 2 2 2 1 3 1 2 3 2 1 2 1 2 2 2 3 2 1 3 3 1 2 1 1 2 3 2 3 3 3
##  [38] 1 2 2 2 2 2 3 1 1 3 3 2 1 2 3 1 1 3 3 1 1 2 1 3 1 2 1 3 3 1 1 1 2 2 3 2 1
##  [75] 3 2 3 2 1 1 3 1 3 3 2 2 2 3 3 3 1 2 2 2 3 2 1 3 3 1 3 2 1 2 2 1 1 1 2 1 3
## [112] 3 1 1 3 2 2 1 1 1 1 1 3 2 2 3 1 2 3 2 1 2 1 2 1 3 1 3 3 2 1 2 3 1 3 3 2 2
## [149] 3 2 2 3 1 3 3 2 3 3 2 3 1 2 2 2 1 2 2 3 3 1 1 2 2 1 2 3 2 2 3 3 1 3 1 1 2
## [186] 1 2 1 2 1 2 2 2 2 3 1 2 3 3 2 2 1 2 3 2 1 3 1 2 3 1 2 2 3 3 1 1 2 1 2
```

```
#CENTROS DOS CLUSTERS
clusters$centers
```

```
##    salario  posicao anosexperiencia     sexo
## 1 10.71517 2.269808        1.902795 2.104838
## 2 12.31883 3.808876        2.673347 1.904378
## 3 10.98853 2.195639        1.286458 0.000000
```

```
#TAMANHO DOS CLUSTERS
clusters$size
```

```
## [1] 69 84 67
```

O proximo passo será incluir nossa predições na base original para que possamos explorar estes agrupamentos.

```
#INSERINDO CLUSTERS NO DATASET E VISUALIZANDO A CLASSIFICACAO EM TABELA
BD$Cluster <- clusters$cluster
#head(BDclusters)
head(BD)
```

```
##   individuo salario posicao anosexperiencia sexo Cluster
## 1         1     148       7            16,7    1       2
## 2         2     165       7             6,7    1       2
## 3         3     145       5            14,8    1       2
## 4         4     139       7            13,9    0       3
## 5         5     142       6             6,4    0       3
## 6         6     144       5             9,1    1       1
```

Podemos visualizar em % a proporção do conjunto de dados em cada grupo.

```
#Proporcao da quantidade de individuos por grupo
proporcaocluster <- round(table(BD$Cluster)*100/nrow(BD),2)
proporcaocluster
```

```
##
##     1     2     3
## 31.36 38.18 30.45
```

Afim de conseguirmos uma visualização mais clara, arredondamos a variavel de anos de experiencia para 0 casas decimais, além da substituição da virgula pelo ponto na formatação de casa decimal desta vez no nosso conjunto de dados inicial.

```
#substituindo ponto por virgula
anosexperiencia <- sapply(BD, function(x) any(grepl(",", x)))
BD$anosexperiencia <- sapply(BD[,c("anosexperiencia")], function(x) as.numeric(sub(",", ".",
 x)))
#Verficando o tipo das colunas
str(BD)
```

```
## 'data.frame':    220 obs. of  6 variables:
##  $ individuo      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ salario        : int  148 165 145 139 142 144 128 143 157 150 ...
##  $ posicao        : int  7 7 5 7 6 5 3 6 7 7 ...
##  $ anosexperiencia: num  16.7 6.7 14.8 13.9 6.4 9.1 8.5 18.2 13 21.6 ...
##  $ sexo           : int  1 1 1 0 0 1 0 1 1 1 ...
##  $ Cluster        : int  2 2 2 3 3 1 3 2 2 2 ...
```

```
BD$anosexperiencia <- round(BD$anosexperiencia,0)
```

Tornamos tambem as variaveis individuo, posicao, sexo e cluster como sendo do tipo fator.

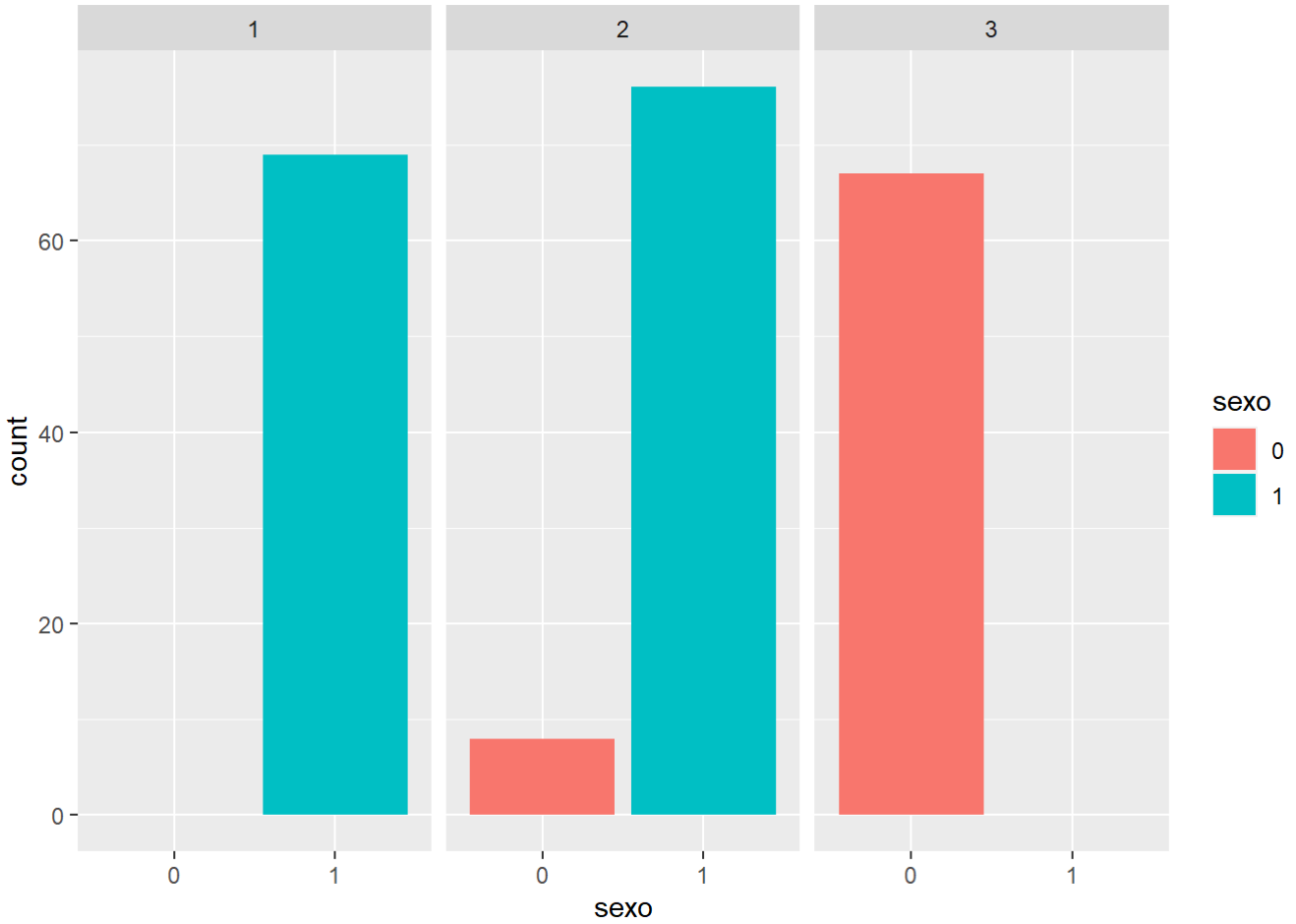
```
#Alterando tipos das variaveis
str(BD)
```

```
## 'data.frame':    220 obs. of  6 variables:
##  $ individuo      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ salario        : int  148 165 145 139 142 144 128 143 157 150 ...
##  $ posicao        : int  7 7 5 7 6 5 3 6 7 7 ...
##  $ anosexperiencia: num  17 7 15 14 6 9 8 18 13 22 ...
##  $ sexo           : int  1 1 1 0 0 1 0 1 1 1 ...
##  $ Cluster        : int  2 2 2 3 3 1 3 2 2 2 ...
```

```
BD$individuo <- as.factor(BD$individuo)
BD$posicao <- as.factor(BD$posicao)
BD$sexo <- as.factor(BD$sexo)
BD$Cluster <- as.factor(BD$Cluster)
```

Damos inicio a uma análise exploratória mais profunda afim de entendermos o que os grupos possuem em comum.

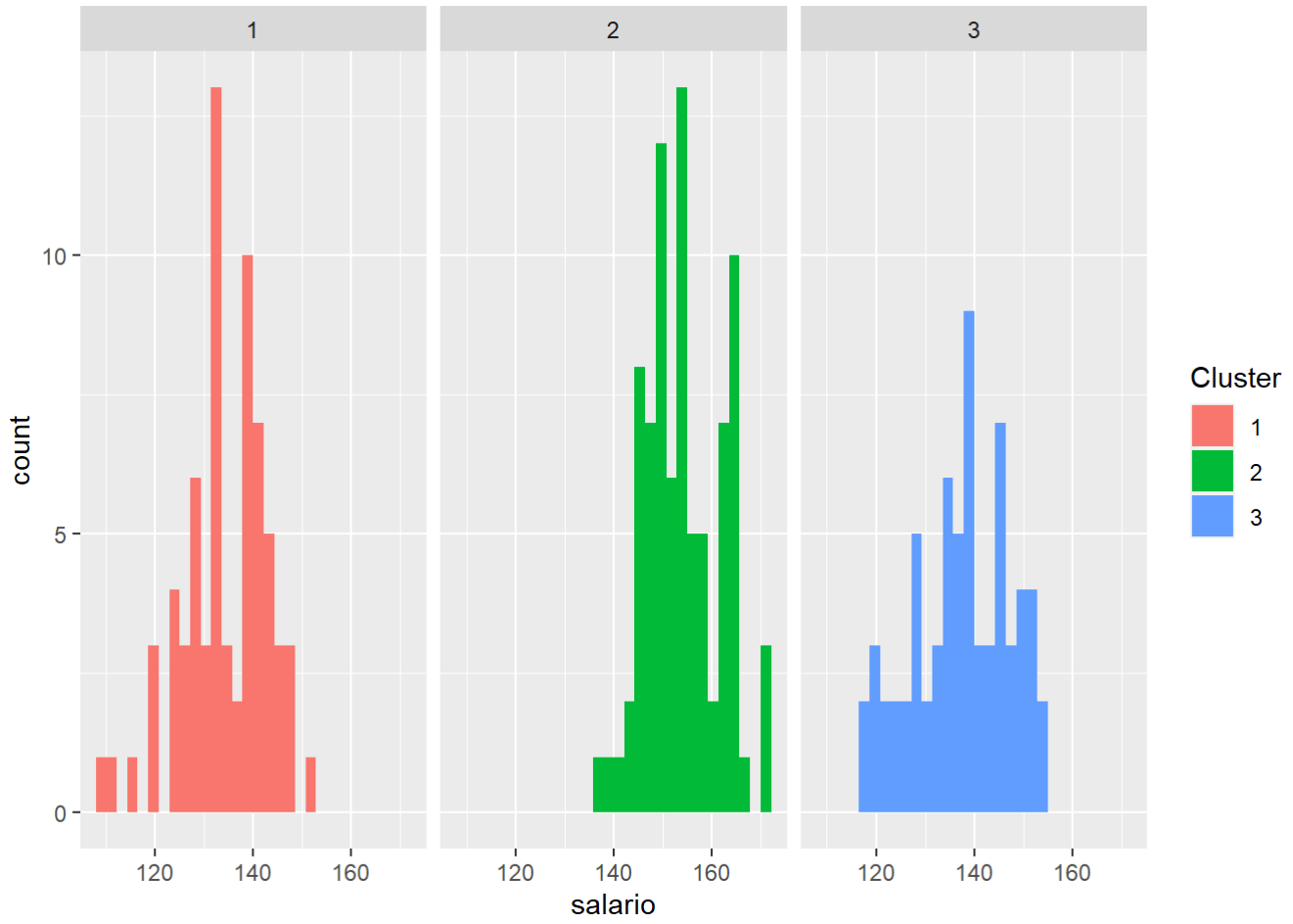
```
#Plotando grupos
library(ggplot2)

#Sexo barras
ggplot(BD) +
  aes(x = sexo, fill = sexo) +
  geom_bar() +
  scale_fill_hue() +
  theme_gray() +
  facet_wrap(vars(Cluster))
```



Podemos notar que o grupo 1 é composto apenas por individuos do sexo masculino. O grupo 2 possui em torno de 92% da sua composição individuos do sexo masculino e o grupo 3 é comppsto apenas por individuos do sexo feminino.

Plotando o salario em forma de histograma.

```
#salario histogram
ggplot(BD) +
  aes(x = salario, fill = Cluster) +
  geom_histogram(bins = 30L) +
  scale_fill_hue() +
  theme_gray() +
  facet_wrap(vars(Cluster))
```



Nota-se que o grupo 1 possui os individuos com o salário mais baixo apesar de ter alta concentração de salários que permeiam o ponto médio de salario do nosso conjunto de dados. O grupo 2 possui os individuos com maior salario e o grupo 3 possui os individuos com menor variancia de salário, haja visto que estes não recebem nem o menor e nem o maior salário da nossa base de dados e estão levemente concentrado no ponto médio da escala de salarios.

Podemos explorar um pouco mais da variavel salario:

```
#Explorando os salarios dos grupos
mediasalario <- BD %>% group_by(Cluster) %>% summarise(media.salario= mean(salario))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
minimosalario <- BD %>% group_by(Cluster) %>% summarise(minimo.salario= min(salario))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
maxsalario <- BD %>% group_by(Cluster) %>% summarise(max.salario= max(salario))
```

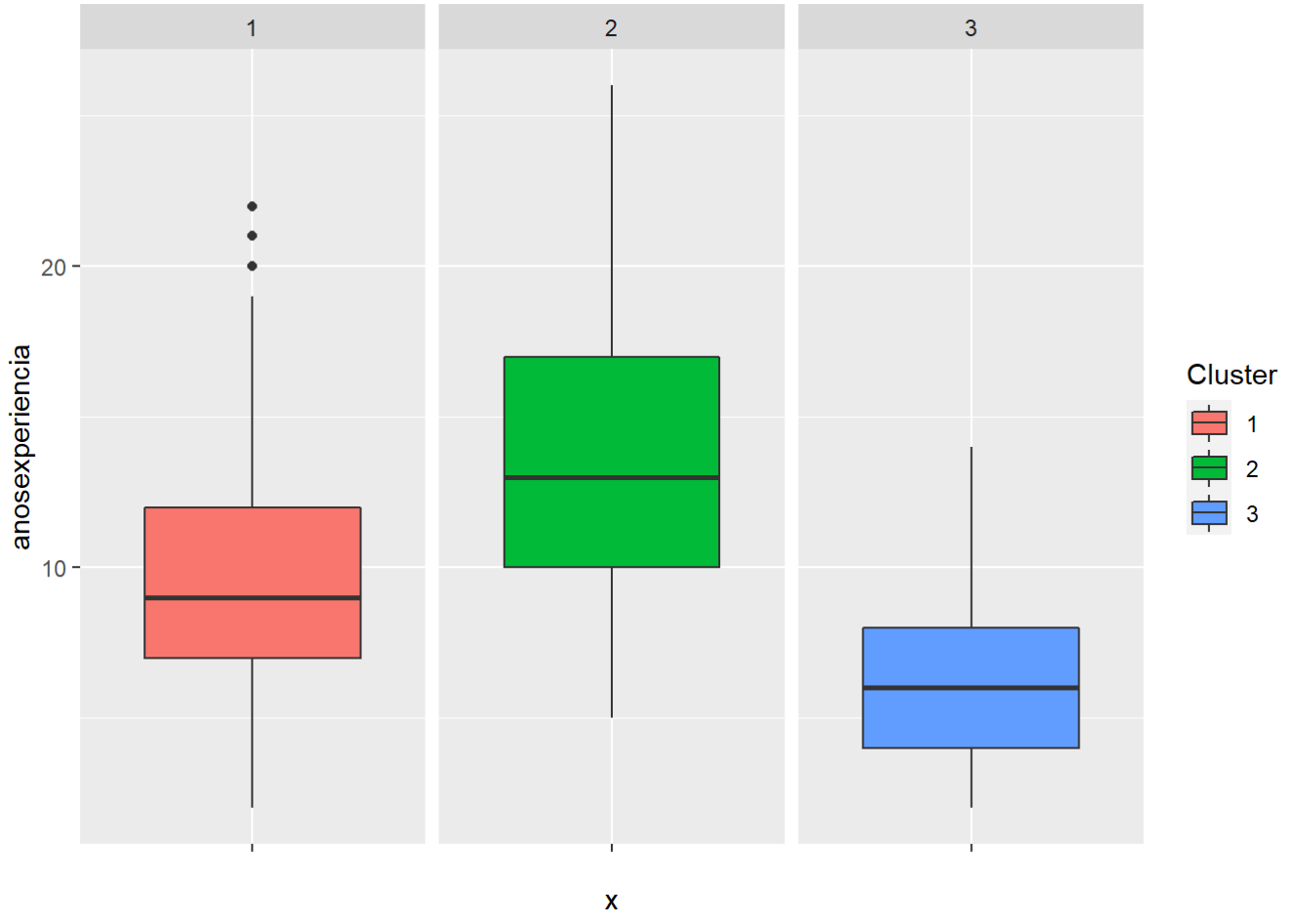
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
cbind(minimosalario, mediasalario, maxsalario)
```

```
##   Cluster minimo.salario Cluster media.salario Cluster max.salario
## 1       1            110       1      134.1594       1         152
## 2       2            136       2      154.2381       2         172
## 3       3            118       3      137.5821       3         153
```

O grupo 1 possui um salario minimo de 110, uma media de 134 e um salario maximo de 152.
O grupo 2 possui um salario minimo de 136, uma media de 154 e um salario maximo de 172.
O grupo 3 possui um salario minimo de 118, uma media de 137 e um salario maximo de 153.

```
#Anos experiencia boxplot
ggplot(BD) +
  aes(x = "", y = anosexperiencia, fill = Cluster) +
  geom_boxplot() +
  scale_fill_hue() +
  theme_gray() +
  facet_wrap(vars(Cluster))
```

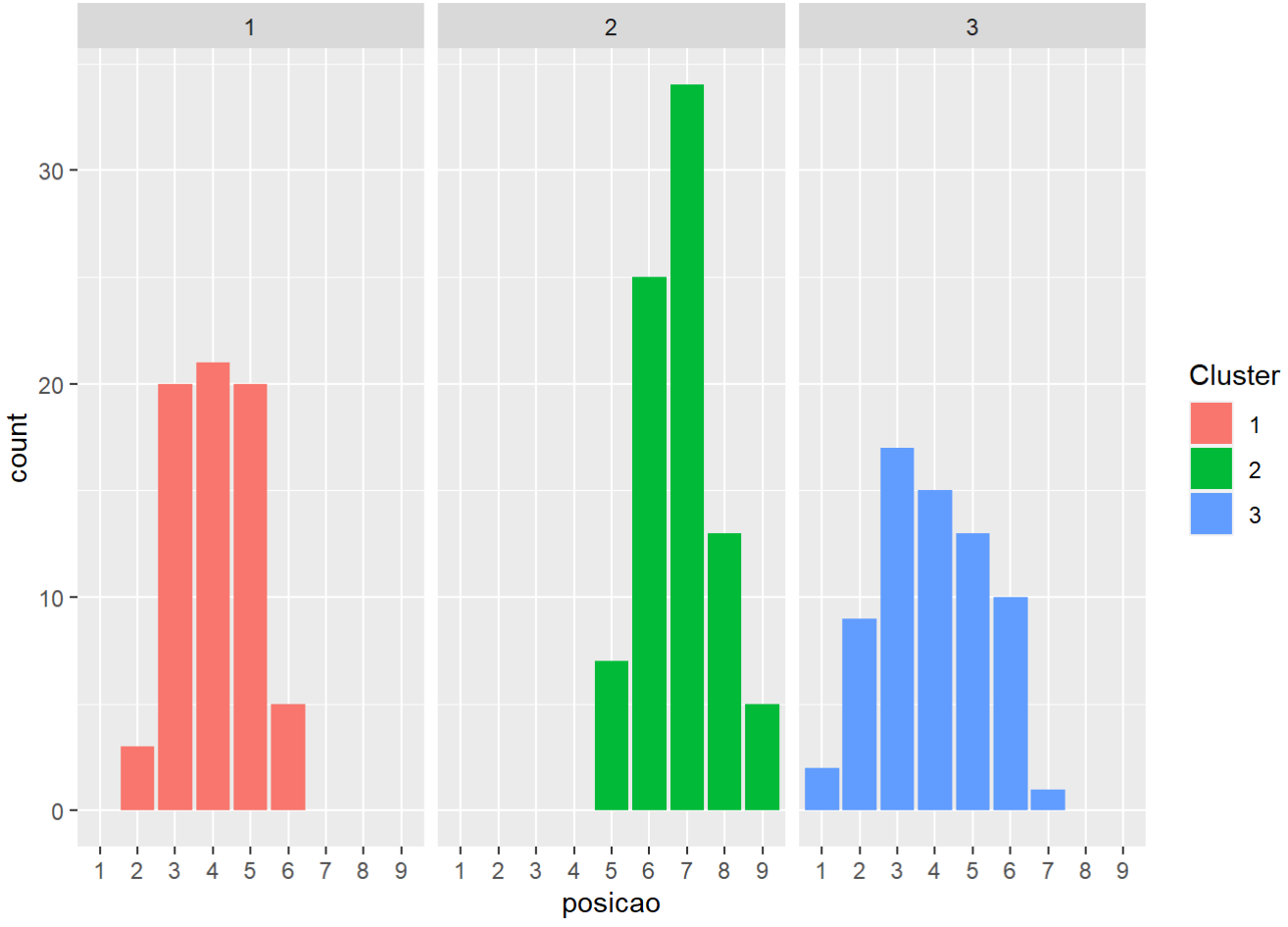


A partir desta visualização, podemos observar os grupos em diferentes niveis de senioridade onde o grupo 3 possui os individuos em inicio de carreira, haja visto ser o grupo com menos tempo de experiencia. O grupo 1 com individuos que provavelmente estão em transição de senioridade de junior para pleno, individuos que ja possuem um pouco mais de experiencia.O grupo 2, individuos que provavelmente são as referências das suas áreas com mais anos de experiencia.

Nota-se também que o grupo 1 é o conjunto que possui mais outliers, onde provavelmente sejam pessoas que não tiveram acesso qualificado a educação e podem ter estagnado em uma senioridade inicial.

Podemos confirmar essas premissas com o gráfico abaixo que mostra a quantidade de individuos por posição na empresa segmentado por grupos.

```
#Posicao barras
ggplot(BD) +
  aes(x = posicao, fill = Cluster) +
  geom_bar() +
  scale_fill_hue() +
  theme_gray() +
  facet_wrap(vars(Cluster))
```



7. Conclusão

Por fim, conclui-se que o conjunto de dados Ajax possui 3 grupos diferentes de individuos, sendo eles:

Grupo 1: Apenas individuos do sexo masculino com uma média em torno de 8 anos de experiencia que possuem senioridades iniciais e salarios iniciais condizentes com sua posição.

Grupo 2: Composto por mais de 90% de individuos do sexo masculino, é o conjunto que possui a maior senioridade, mais anos de experiência, ocupa as posições mais altas e possuem o maior salario.

Grupo 3: Composto apenas por mulheres, o conjunto possui a menor variancia de salarios, é o grupo com menos anos de experiência e que possui maior distribuição de individuos nas posições de senioridade pleno.