Employee analysis

Thiago Silva 01/05/2021

This dataset is a fictitious base of a company's staff.

The objective is to apply the K-means clustering algorithm to separate the individuals in the database into different groups based on their characteristics.

1. About the database

Our dataset has information on 220 individuals and their salary information, years of experience, position in the company and gender.

2. Importing the files

Loading the database and adjusting the column names.

```
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. Exploring the database

Checking for missing values in the database.

```
#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
```

In this case, we do not have any missing values in our database, and no treatment is required for this purpose.

Exploring the set to understand its dimensions.

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

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

We have 220 rows and 6 columns.

Identifying sex proportions distributed in the base.

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

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

Thus, assuming that the categorical variable 1 is male and variable 0 is female, we can see that the base is distributed in 65.91% of male individuals and 34.09% of female individuals.

We can understand a little more of the basis by plotting correlations to identify variables that have correlation.

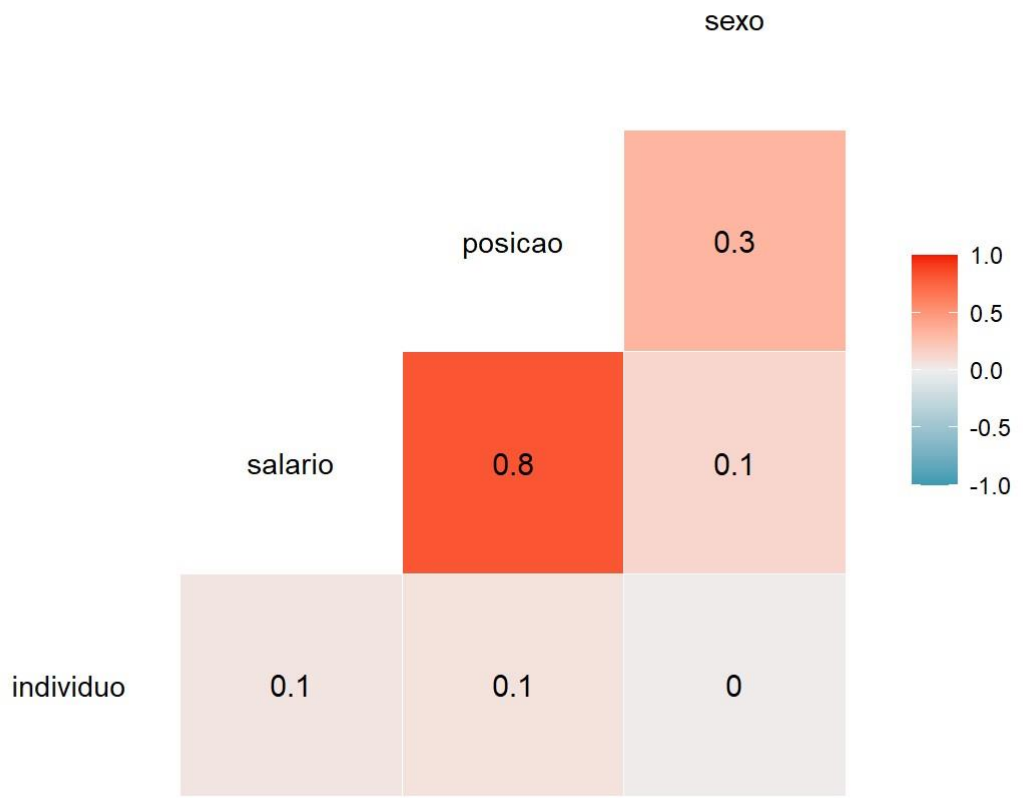
```
#Visualizando correlacoes
library(GGally)
```

```
## 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
```



It is noted that there is a correlation between the position that the individual has in the company and the salary he receives, which makes sense given that as your position evolves, you are expected to receive a higher salary.

4. Transforming the database

To start the construction of the K-means algorithm, we will exclude the ID variable “individuals” which will not be useful to us.

```
#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)
```

Checking the data type of the columns to know if we need to do some kind of treatment.

```
#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 ...
```

It is noted that the variables position, years of experience and sex are categorical variables and were identified as numerical, in this case it will be necessary to perform a treatment on these in order to transform them to the factor type.

Another important point is that the variable “years of experience” has the decimal places separated by a comma, in this case it will be necessary to replace the comma with a period, according to the code below:

```
#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. Creating the k-means algorithm

So that the model is not biased due to different magnitudes, we will perform a data normalization through the scale function.

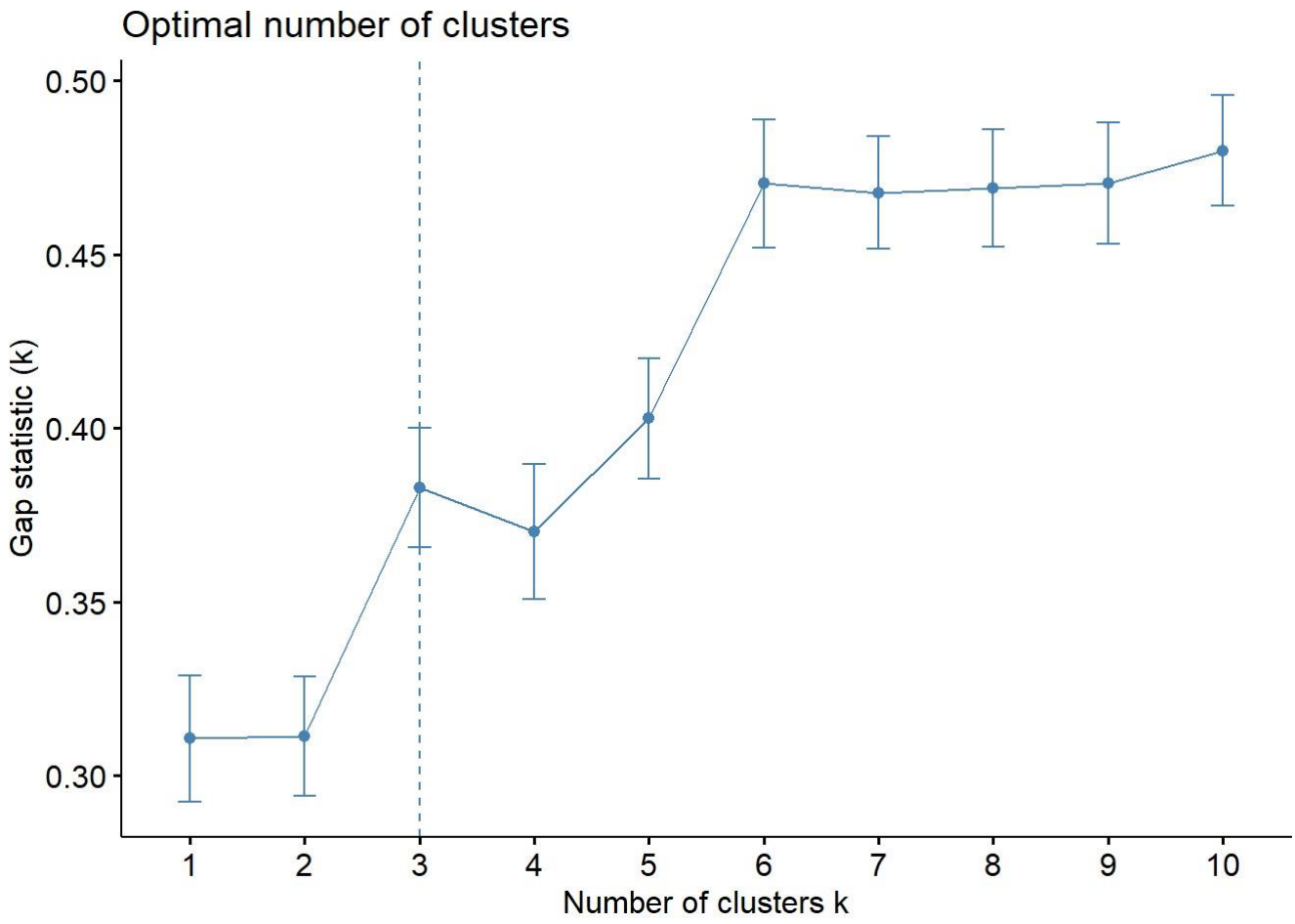
```
#NORMALIZANDO DADOS E ARMAZENANDO NA VARIAVEL "DADOS"
dados<- scale(BDclusters[,c(1:4)])
```

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")
```



As we see explicitly above, we will work with 3 different groups.

We start the prediction by directing the k-means algorithm to separate the set into 3 different groups.

```
library(caret)
```

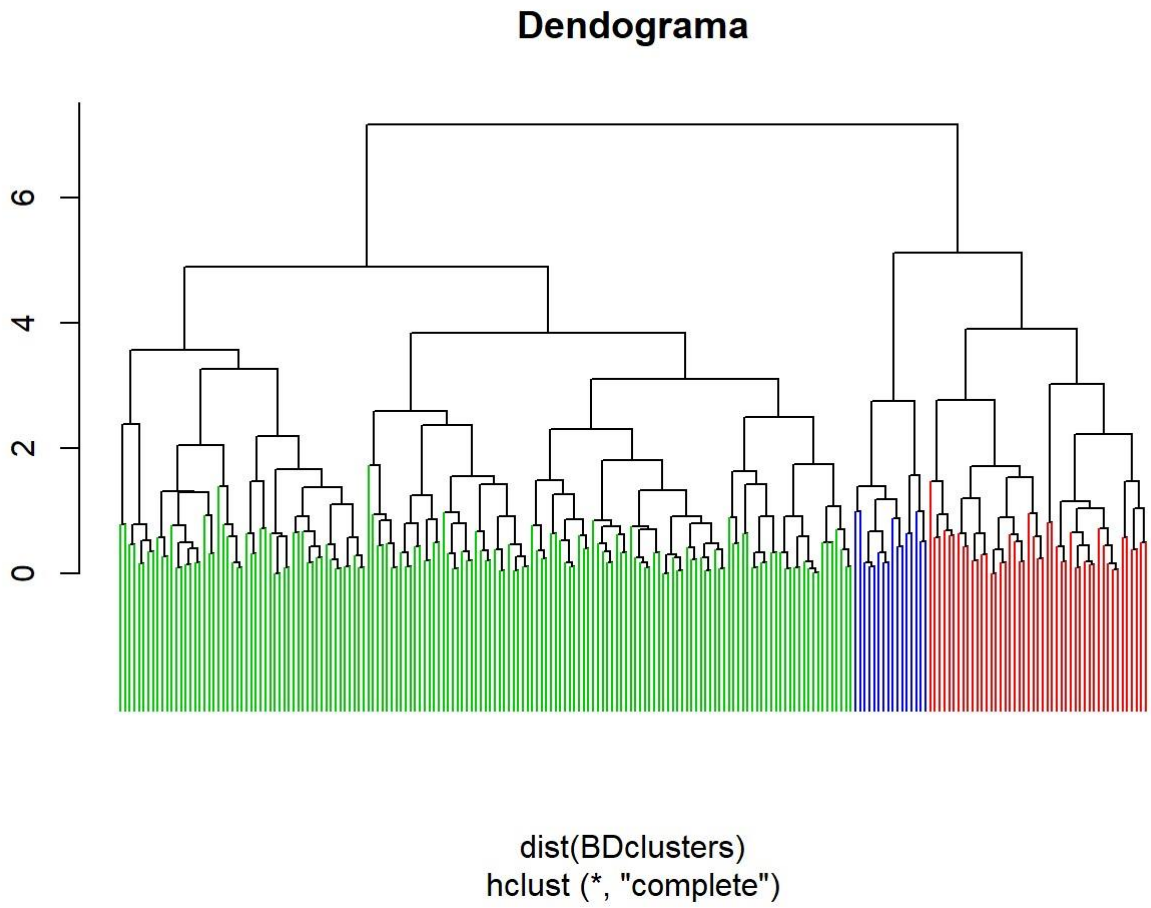
```
## Loading required package: lattice
```

```
#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, centers3)
```

6. Exploring groups

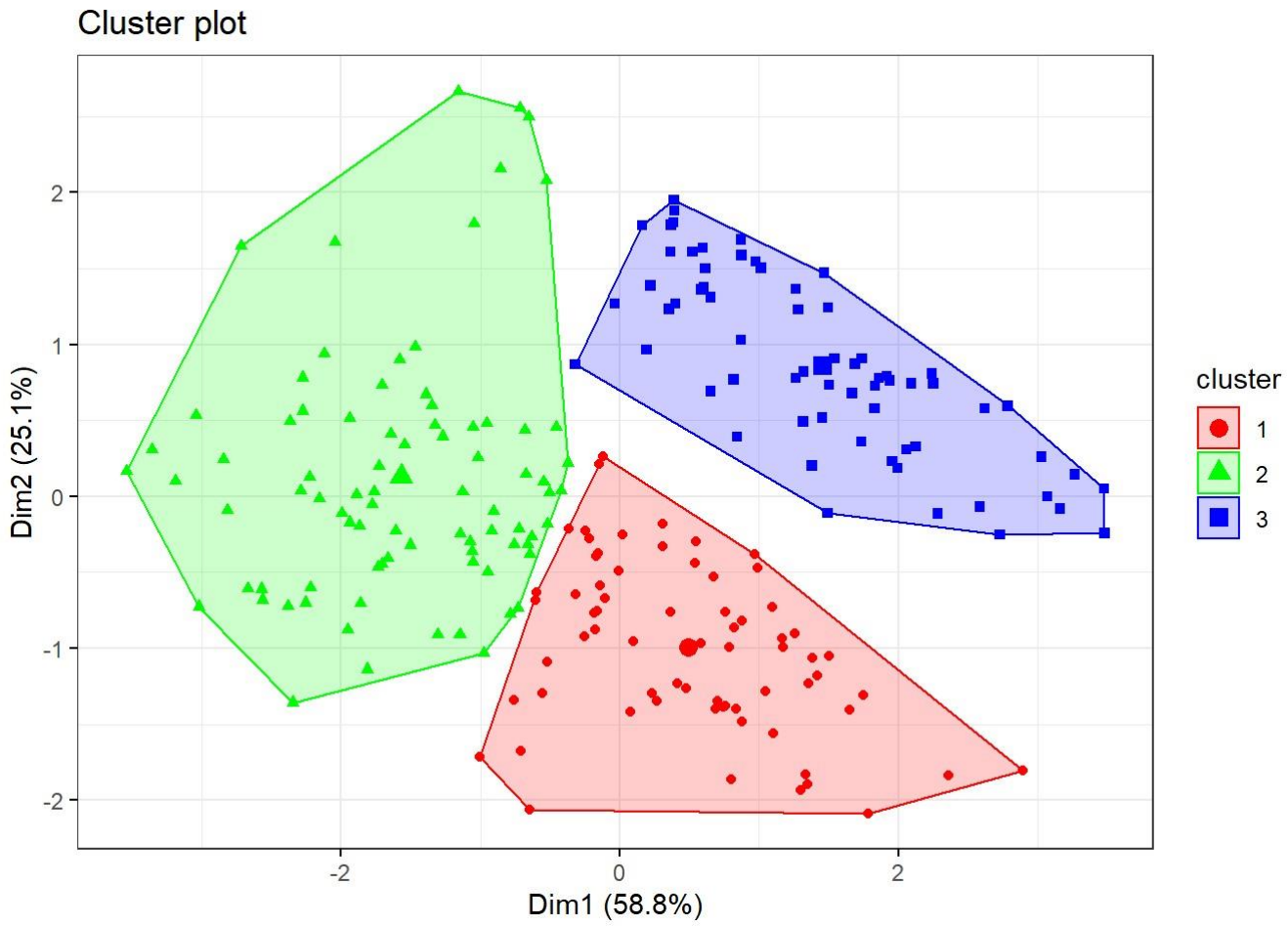
After separating the groups, we can visualize the dendrogram with the groups separated by 3 different colors.

```
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)
```



As an auxiliary view, we have:

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



We can see some statistics of the clusters, such as the clusters, the centroids and how many individuals were allocated to each 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
```

The next step will be to include our predictions in the original base so that we can explore these clusters.

```
#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
```

We can visualize in % the proportion of the dataset in each group.

```
#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
```

In order to get a clearer view, we rounded the years of experience variable to 0 decimal places, in addition to replacing the comma with the dot in the decimal place format this time in our initial dataset.

```
#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)
```

We also made the individual, position, sex and cluster variables to be of the factor type.

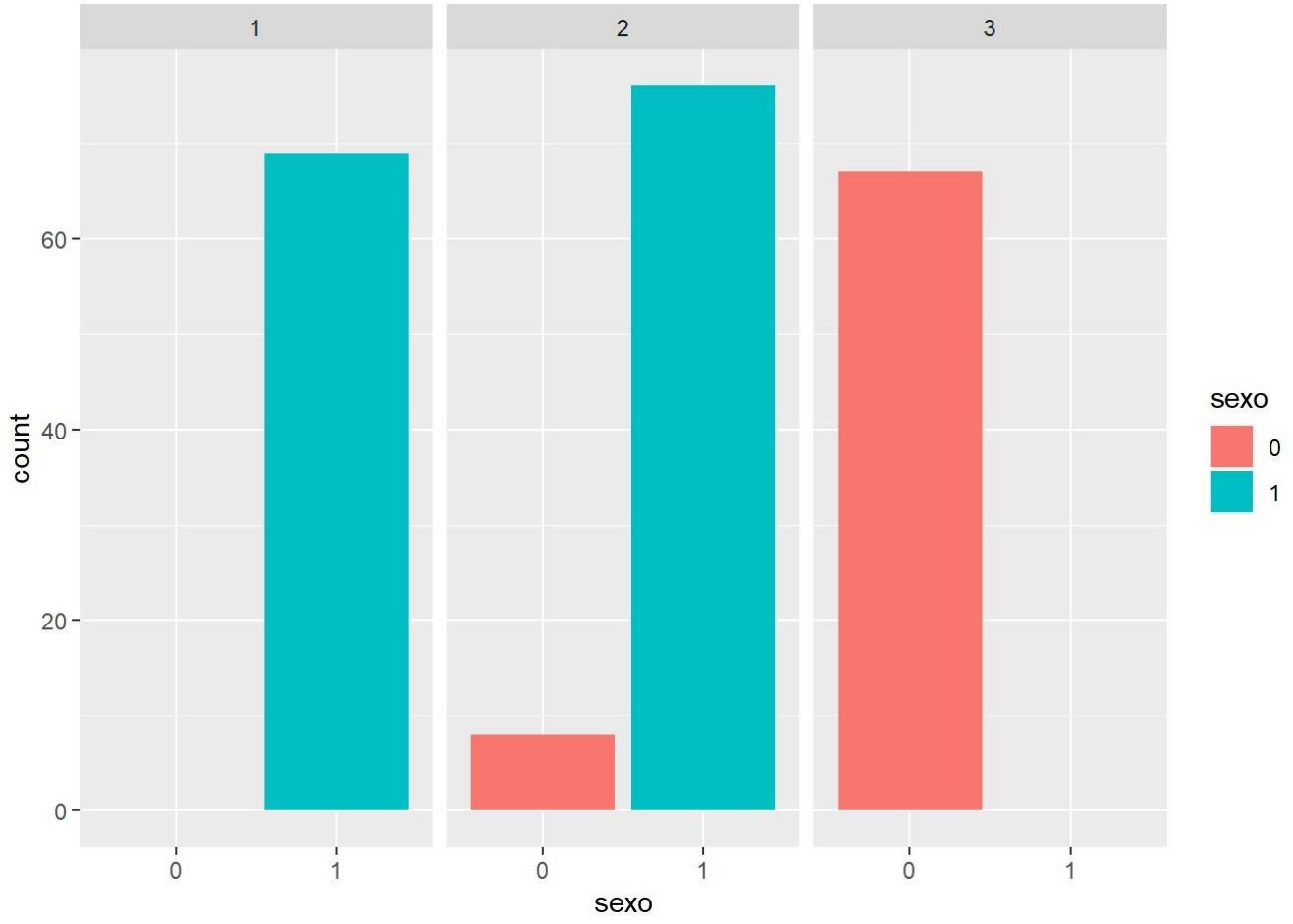
```
#Alterando tipos das variaveis
str(BD)
```

```
## '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)
```

We begin a deeper exploratory analysis in order to understand what the groups have in common.

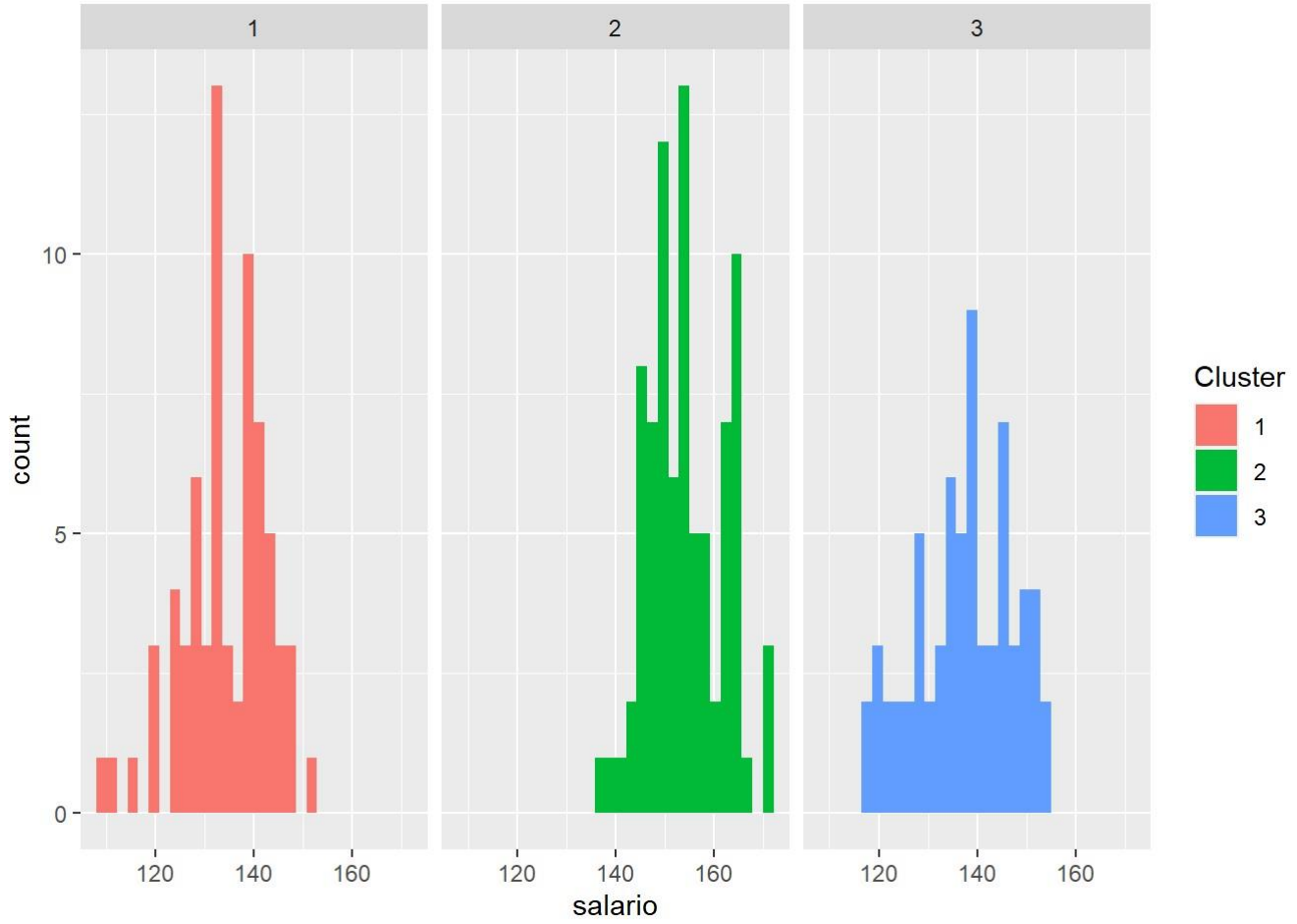
```
#Plotando grupos
library(ggplot2)

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



We can see that group 1 is composed only of male individuals. Group 2 has around 92% of its composition male individuals and group 3 is composed only of female individuals.

Plotting the salary in the form of a histogram.

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



Note that group 1 has the individuals with the lowest salary despite having a high concentration of salaries that permeate the average salary point of our data set. Group 2 has the individuals with the highest salary and group 3 has the individuals with the lowest salary variance, given that they receive neither the lowest nor the highest salary in our database and are slightly concentrated at the midpoint of the scale. of wages.

We can explore a little more of the salary variable:

```
#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))
```

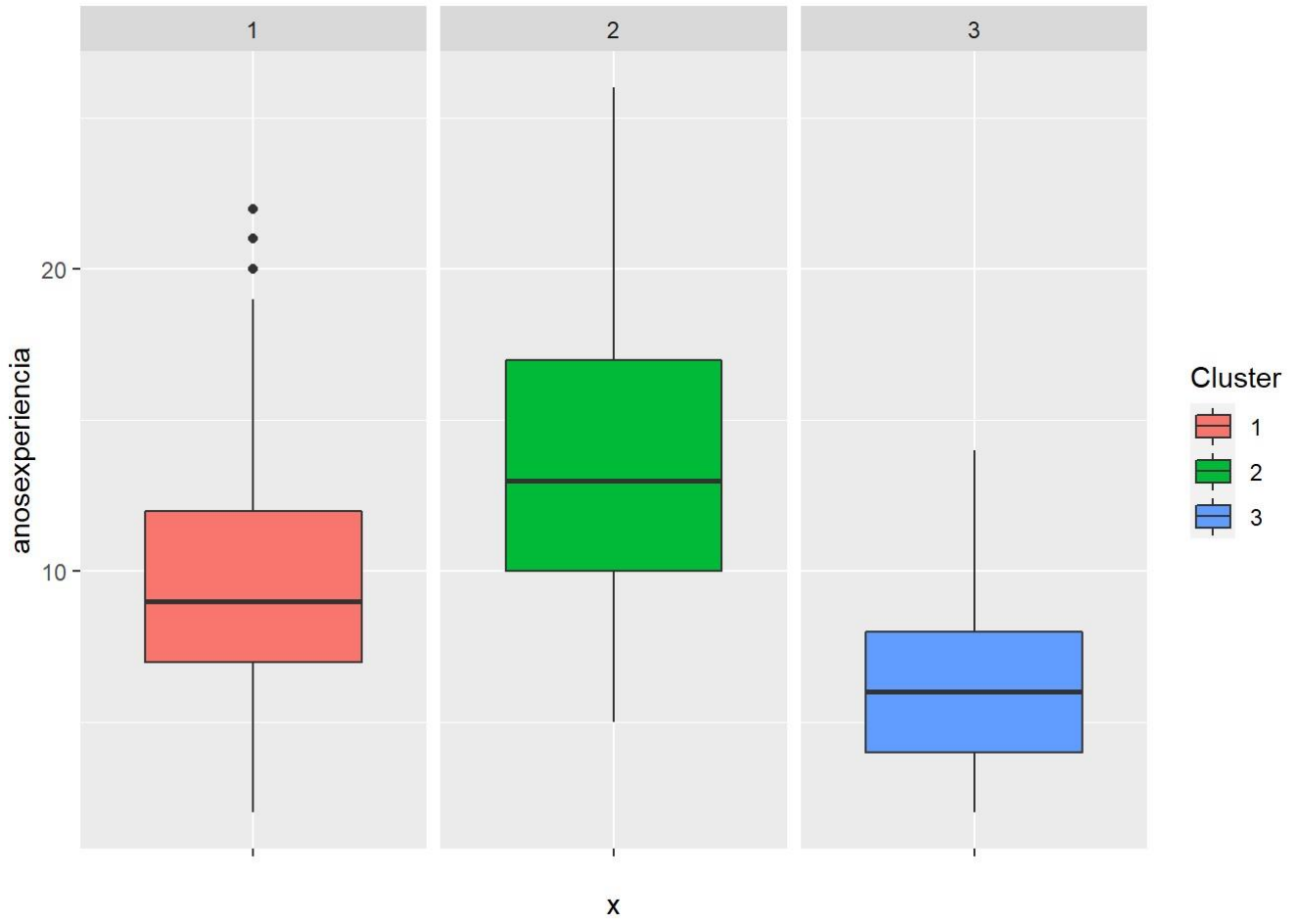
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
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
```

Group 1 has a minimum salary of 110, an average of 134 and a maximum salary of 152.
Group 2 has a minimum salary of 136, an average of 154 and a maximum salary of 172.
Group 3 has a minimum salary of 118, an average of 137 and a maximum salary of 153.

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



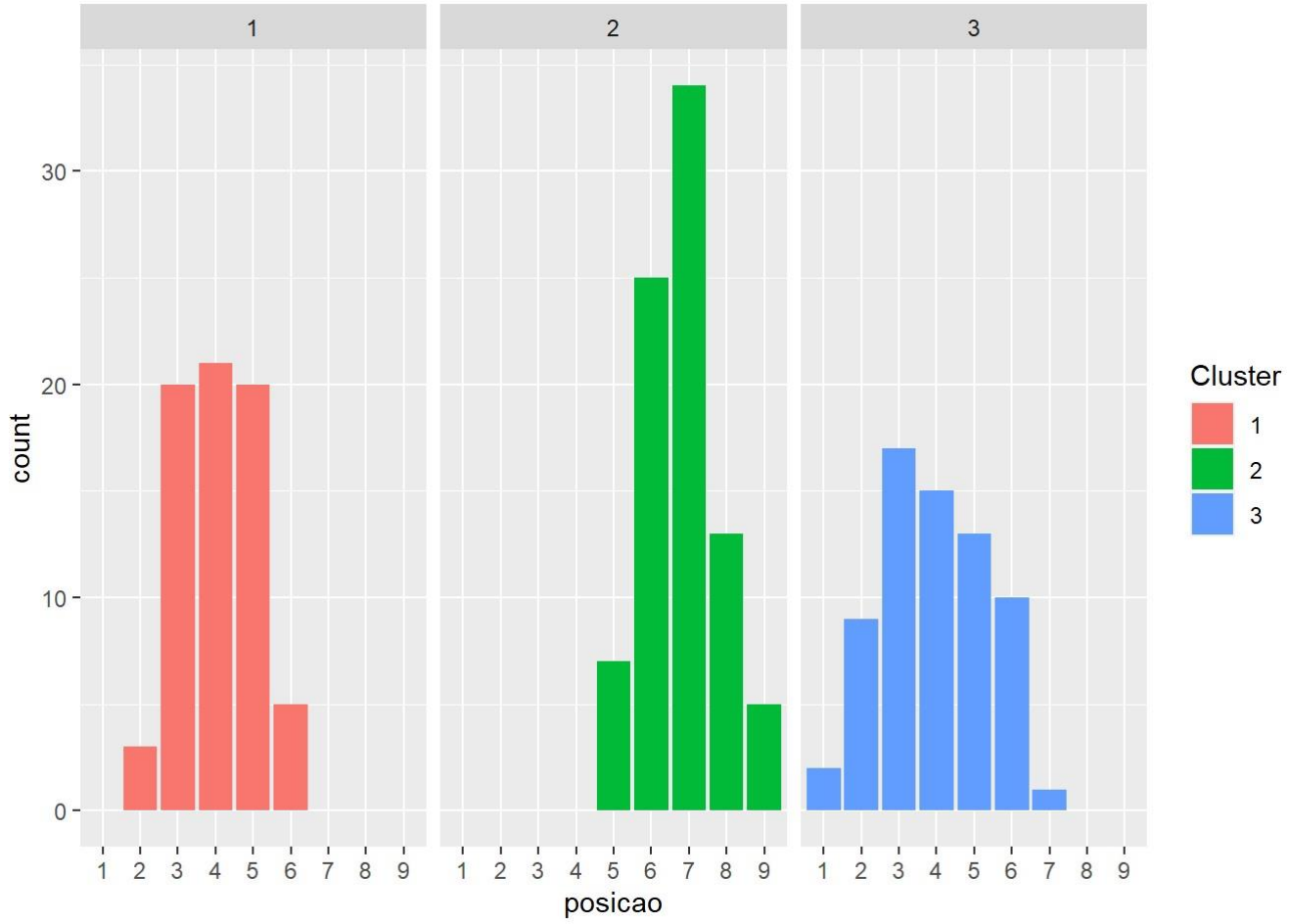
From this visualization, we can observe groups at different levels of seniority where group 3 has individuals at the beginning of their careers, as it is the group with less experience. Group 1 with individuals who are

probably transitioning from junior to full seniority, individuals who already have a little more experience. Group 2, individuals who are probably the references of their areas with more years of experience.

It is also noted that group 1 is the set that has more outliers, which are probably people who did not have qualified access to education and may have stagnated at an initial seniority.

We can confirm these assumptions with the chart below that shows the number of individuals by position in the company segmented by groups.

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



7. Conclusion

Finally, it is concluded that the Ajax dataset has 3 different groups of individuals, namely:

Group 1: Only male individuals with an average of around 8 years of experience who have starting seniority and starting salaries consistent with their position.

Group 2: Comprising more than 90% of male individuals, it is the group that has the highest seniority, most years of experience, occupies the highest positions and has the highest salary.

Group 3: Comprised only of women, the group has the lowest salary variance, is the group with the fewest years of experience and has the largest distribution of individuals in positions of full seniority.