Projeto Cluster

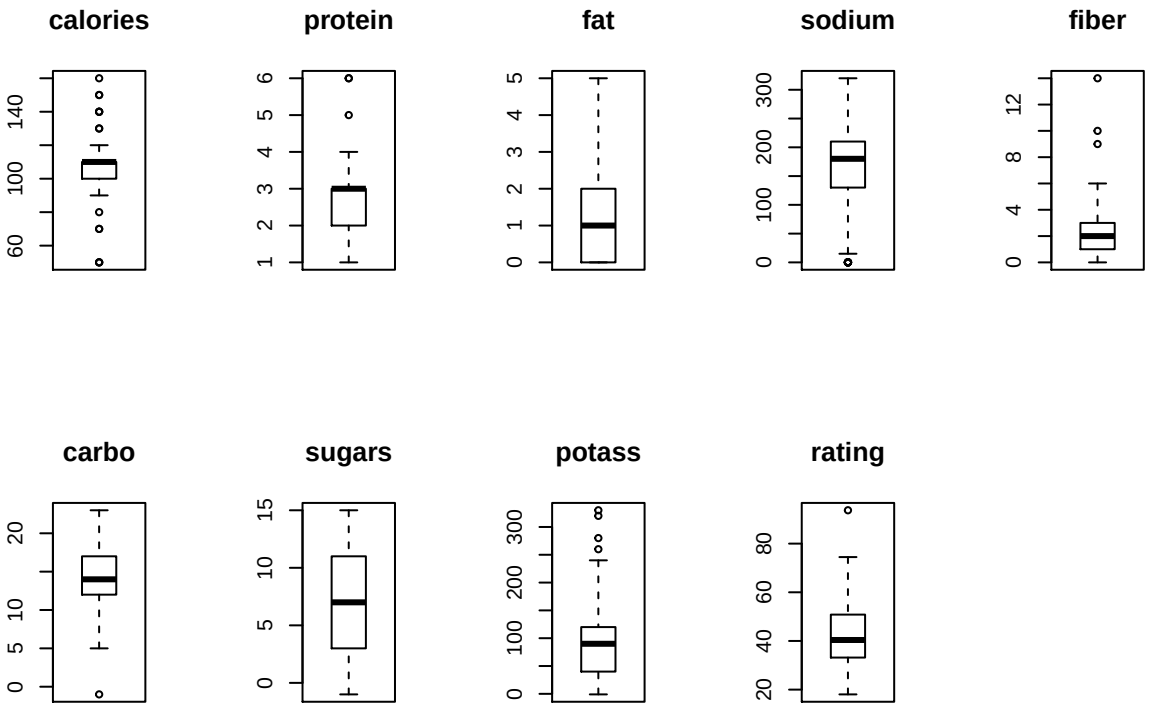
Carregando dados

```
# Carregar Dados
dados <- read_excel("dados.xlsx", sheet = "CEREALS_2")

# Escolher seed
set.seed(18)

# Selecionando Colunas
ce = dados[2:12]

# Transformando chr em factor para calcular distâncias com daisy
ce$type=as.factor(ce$type)
ce$mfr = as.factor(ce$mfr)
```

Explorando os dados

Distribuição dos dados

```
par(mfrow=c(2,5))
for (i in colnames(ce[3:11])){
  boxplot(ce[i], main = i)
}
```



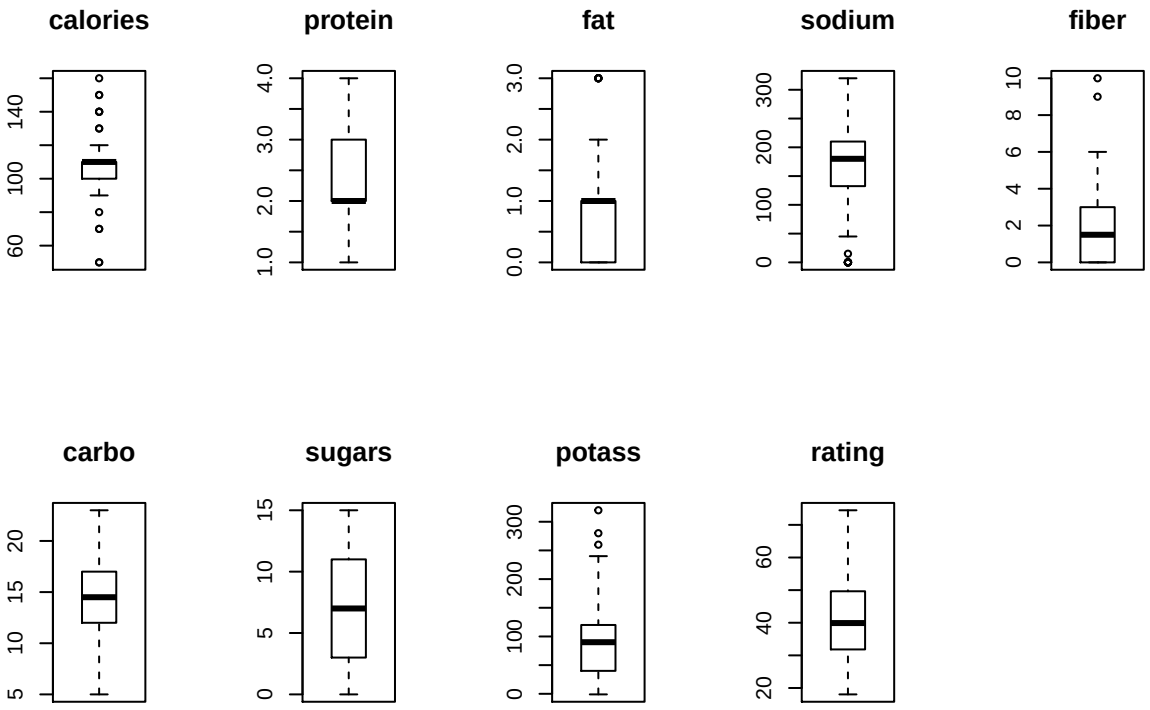
Removendo os outliers

```
ce_out = subset(ce,
       # Regras para outliers
        ce$rating < 80 &
        ce$fiber < 12 &
        ce$carbo > 0 &
        ce$fat < 5 &
        ce$protein < 5
        )

# Outliers removidos
print(dim(ce)[1] - dim(ce_out)[1])
```

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

Dados após remoção dos outliers

```
par(mfrow=c(2,5))
for (i in colnames(ce_out[3:11])){
  boxplot(ce_out[i], main = i)
}
```



Correlacao entre as variaveis

```
# Calcular correlacao
correl=cor(ce_out[3:11]) # Selecionando apenas quantitativos
round(correl,digits=3)
```

```
##          calories protein    fat sodium  fiber  carbo sugars potass rating
## calories    1.000   0.091  0.535  0.334 -0.128  0.226  0.560  0.067 -0.647
## protein     0.091   1.000  0.249 -0.110  0.651 -0.020 -0.199  0.680  0.465
## fat         0.535   0.249  1.000  0.122  0.105 -0.282  0.333  0.245 -0.439
## sodium      0.334  -0.110  0.122  1.000 -0.057  0.257  0.106  0.000 -0.489
## fiber      -0.128   0.651  0.105 -0.057  1.000 -0.336 -0.047  0.907  0.472
## carbo       0.226  -0.020 -0.282  0.257 -0.336  1.000 -0.534 -0.323  0.179
## sugars      0.560  -0.199  0.333  0.106 -0.047 -0.534  1.000  0.105 -0.767
## potass      0.067   0.680  0.245  0.000  0.907 -0.323  0.105  1.000  0.277
## rating     -0.647   0.465 -0.439 -0.489  0.472  0.179 -0.767  0.277  1.000
```

Removendo variaveis com alta correlacao

```
# Remover variaveis
ce_novo=ce_out[,-c(
  10 # Remover Potassio
  )]
```

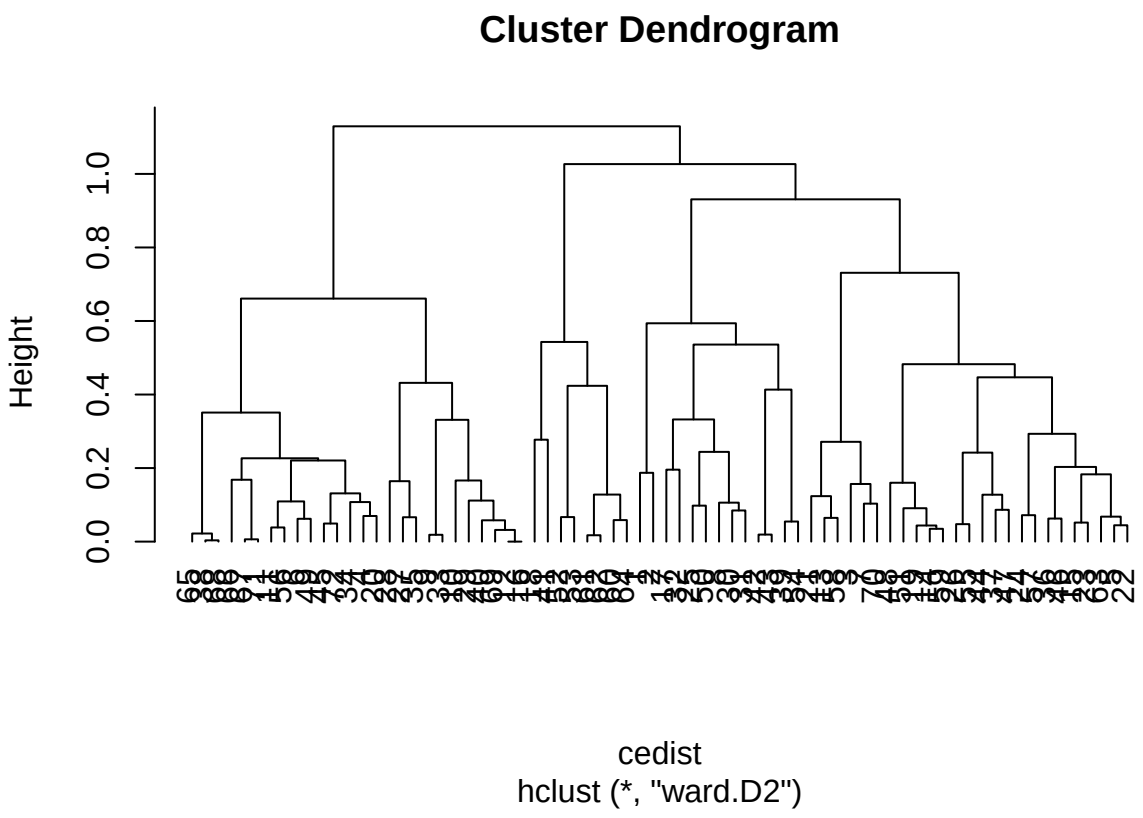
Selecionando Drivers

```
ce_novo.drivers = ce_novo[,-c(10)]
colnames(ce_novo.drivers)
```

```
## [1] "mfr"      "type"     "calories" "protein"  "fat"      "sodium"   "fiber"
## [8] "carbo"    "sugars"
```

Dendograma

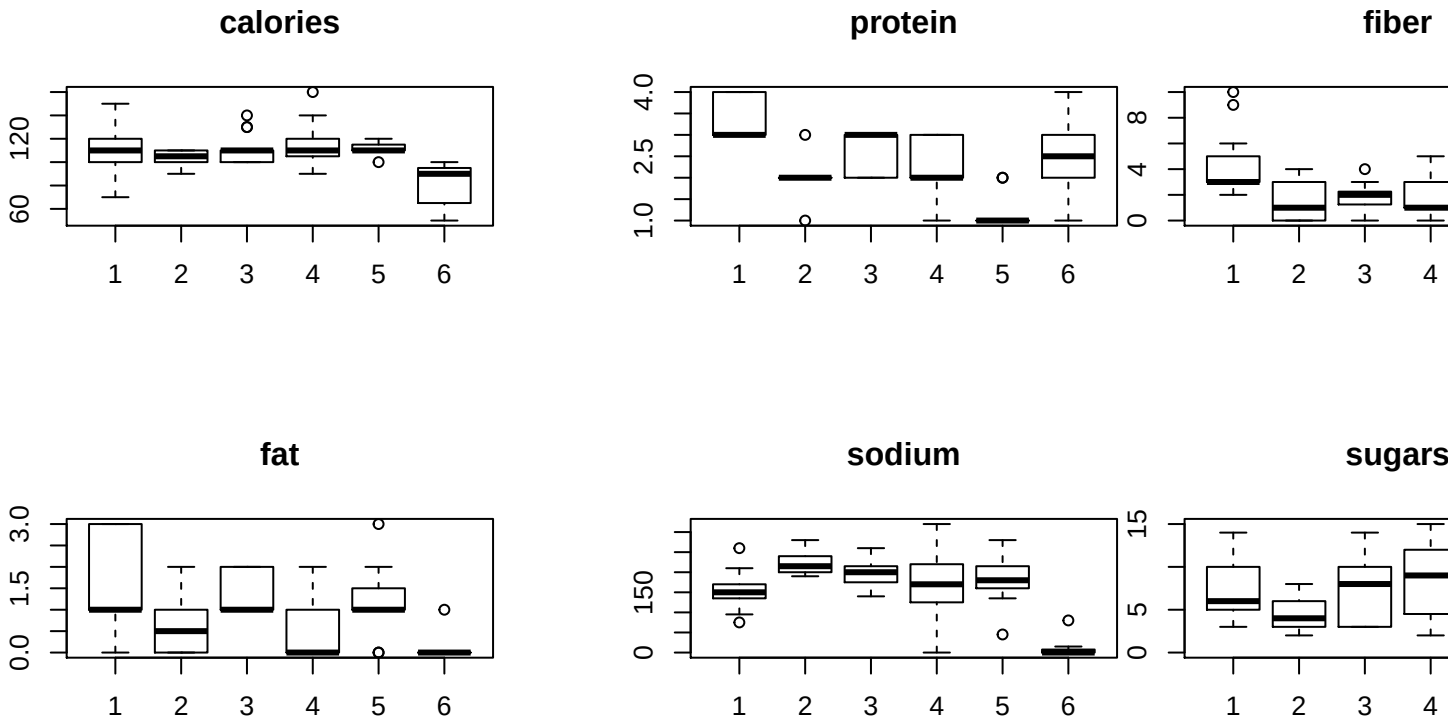
```
ce_dend = ce_novo

# Matriz de distancias
cedist = daisy(ce_novo.drivers)

# Fazer cluster
hcb = hclust(cedist, method = 'ward.D2')

# Coluna com cluster
ce_dend$cluster = as.character(cutree(hcb,
       6  # Numero de clusters
       ))

# Plotar Cluster
plot(hcb, hang = -1)
```



Visualizacao dos resultados

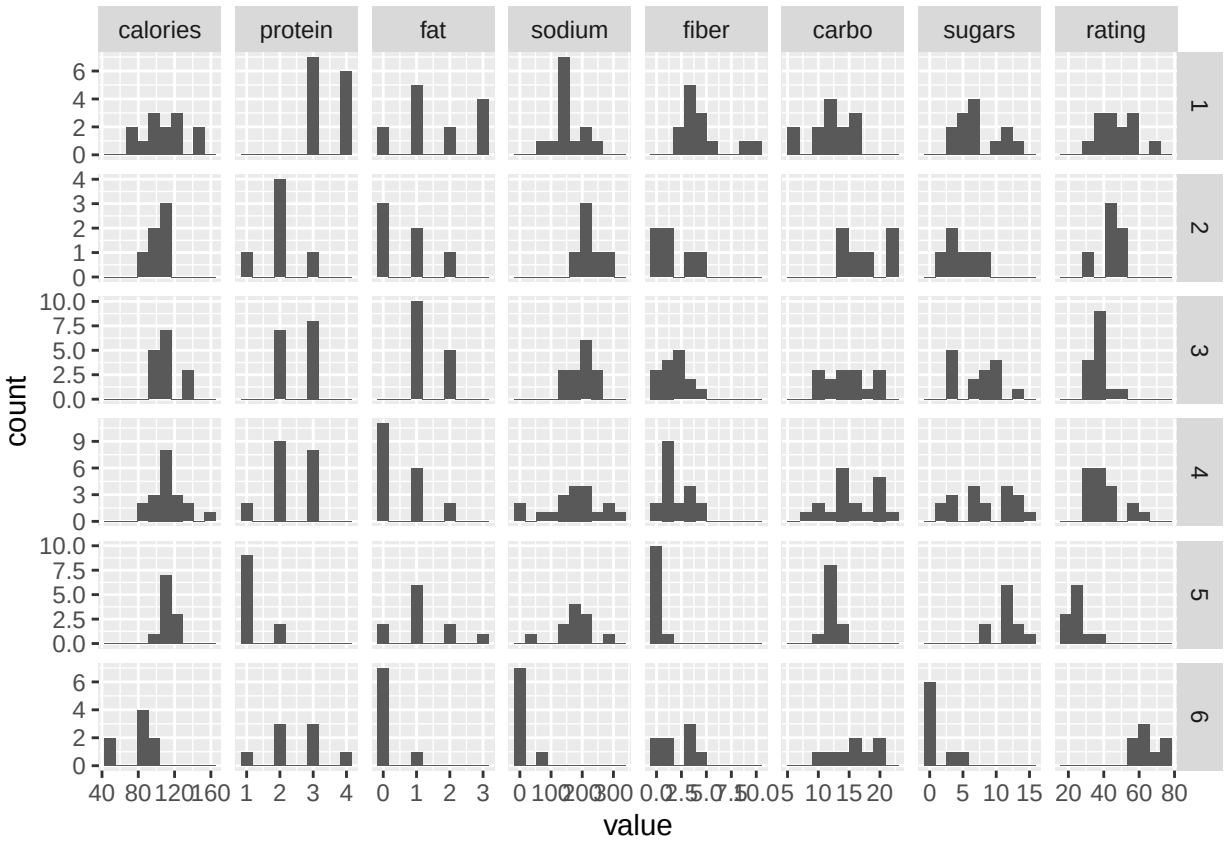
Variaveis Quant

```
par(mfrow=c(2,2))
v = ce_dend[3:10]
n = 1
for (i in v){
 boxplot(i~ce_dend$cluster, main = colnames(v)[n], xlab = NULL, ylab = NULL)
  n = n + 1
}
```



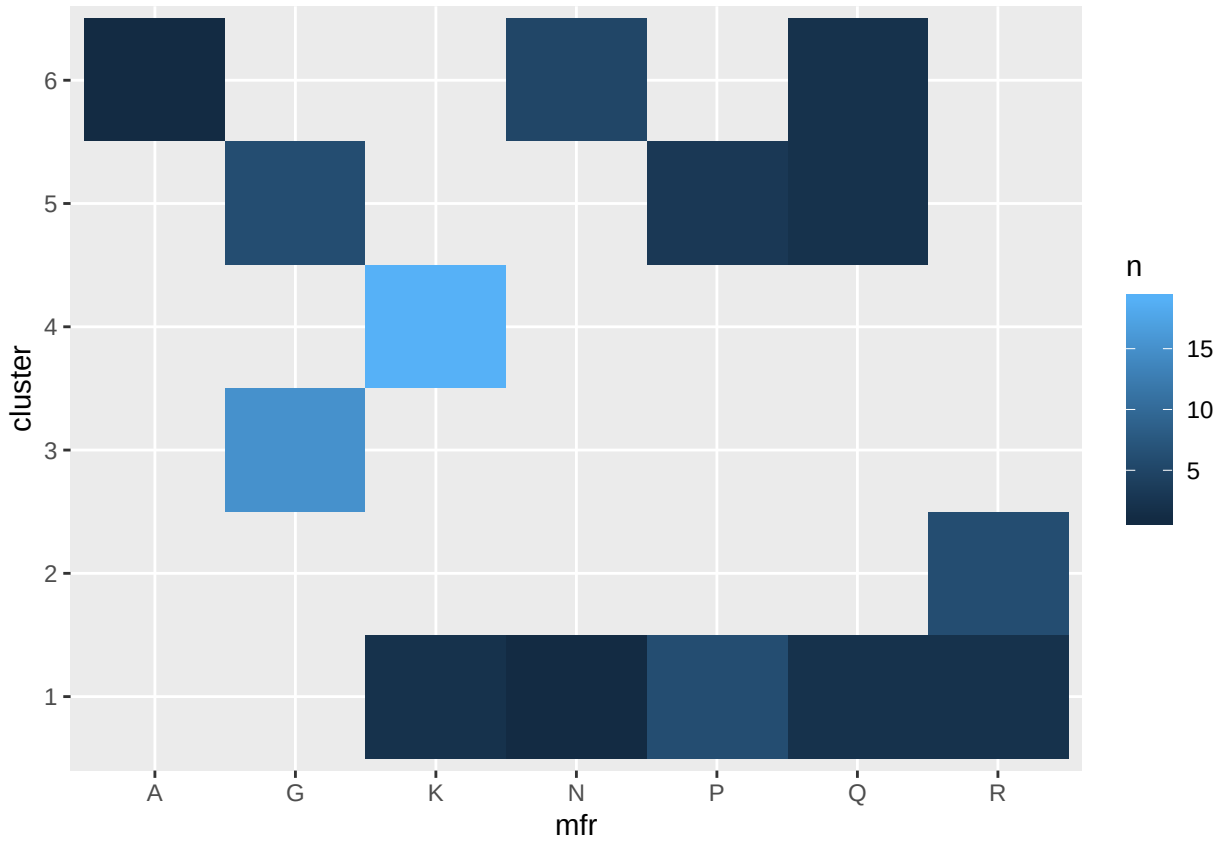
Variaveis Hist

```
df = melt(ce_dend[3:11], id.vars = "cluster", variable.name = 'series' )

ggplot(df, aes(x = value)) + geom_histogram(bins = 10) +
  facet_grid(cluster ~ series, scales = "free")
```

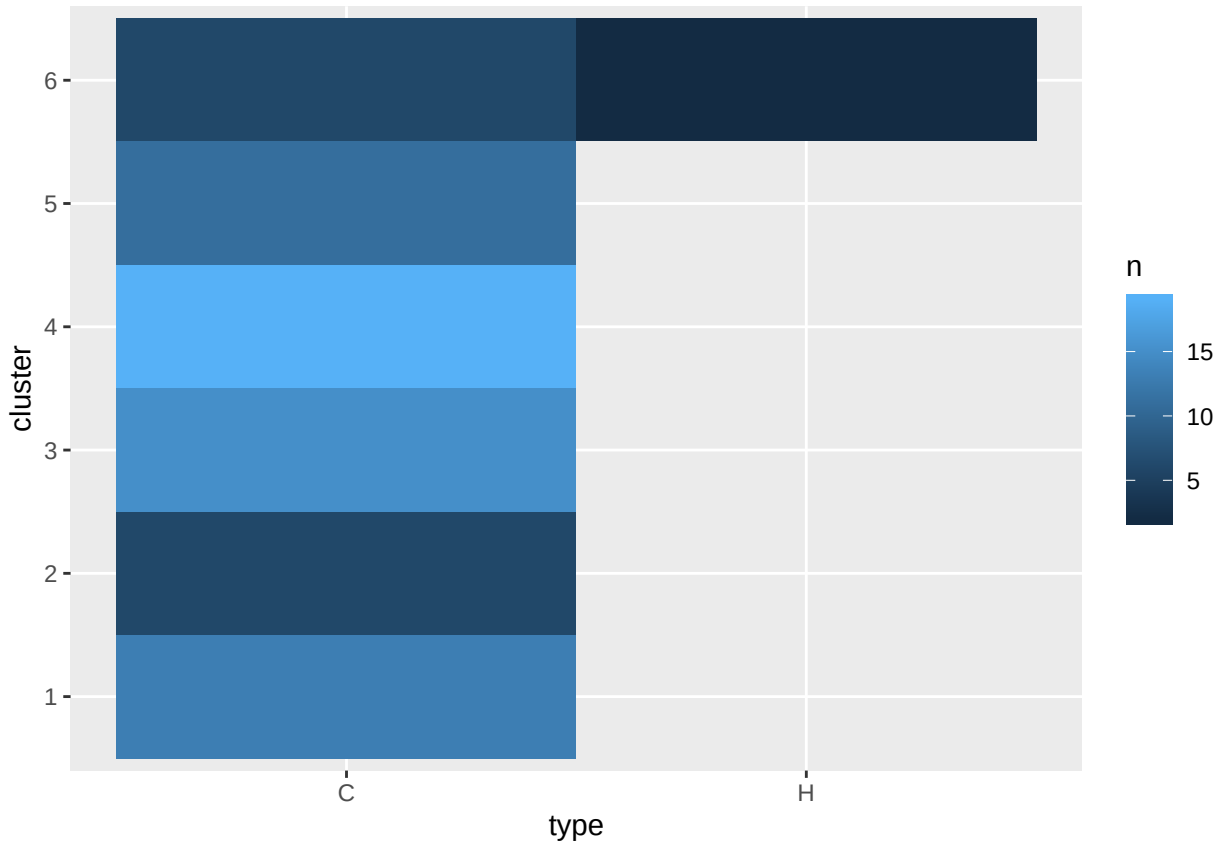


Variaveis Quali

```
d = group_by(ce_dend,cluster, mfr) %>% summarise(n = n())
ggplot(data = d, aes(mfr, cluster)) +
  geom_raster(aes(fill = n))
```

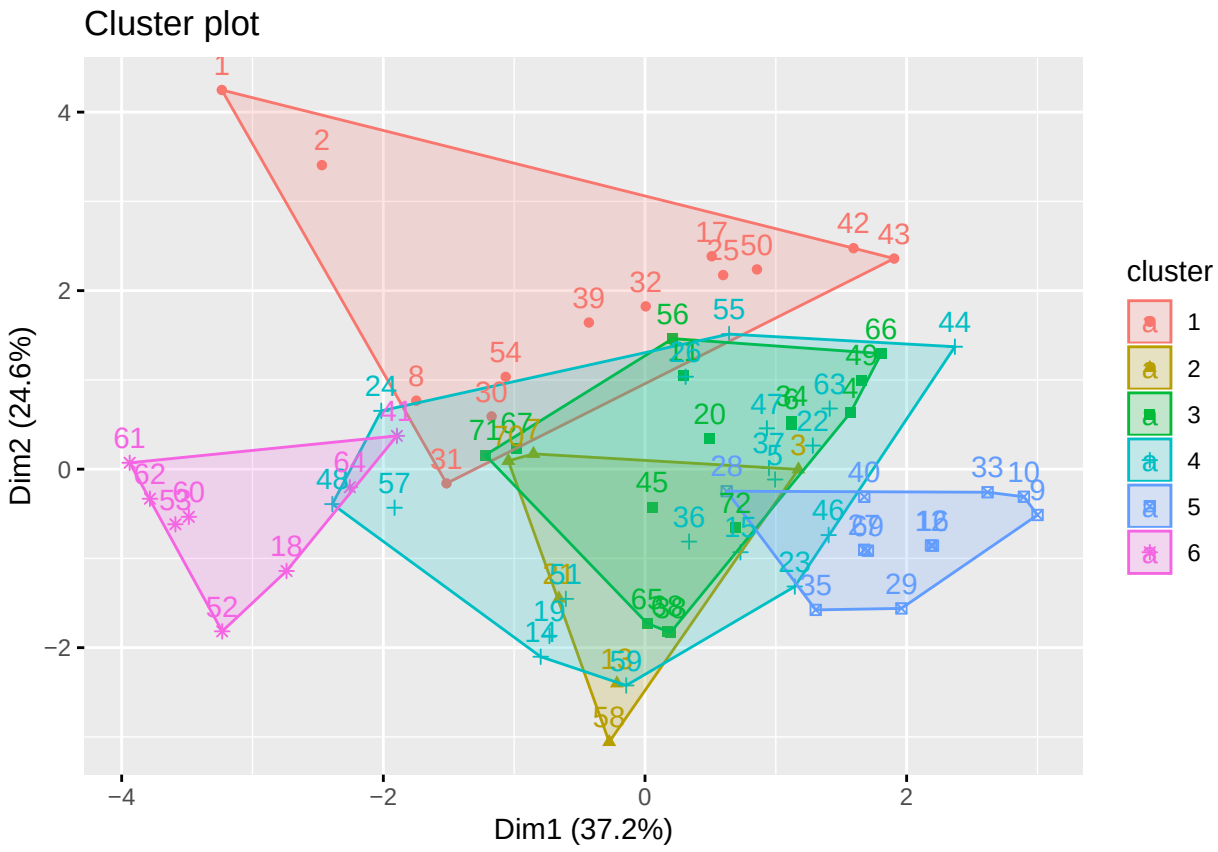



```
d = group_by(ce_dend,cluster, type) %>% summarise(n = n())
ggplot(data = d, aes(type, cluster)) +
  geom_raster(aes(fill = n))
```



Representacao PCA

```
fviz_cluster(list(data = ce_dend[3:10], cluster = ce_dend$cluster), show.clust.cent = F)
```



k-moid

```
ce_pamk = ce_novo

# Matriz de distancias
cedist = daisy(ce_novo.drivers)

# Realizar kmoid
kk=pamk(cedist, krange = 2:6, diss = T, critout = T )
```

```
## 2  clusters  0.1777012
## 3  clusters  0.2064856
## 4  clusters  0.25711
## 5  clusters  0.2298035
## 6  clusters  0.2470171
```

```
# Numero ideal de clusters
kk$nc
```

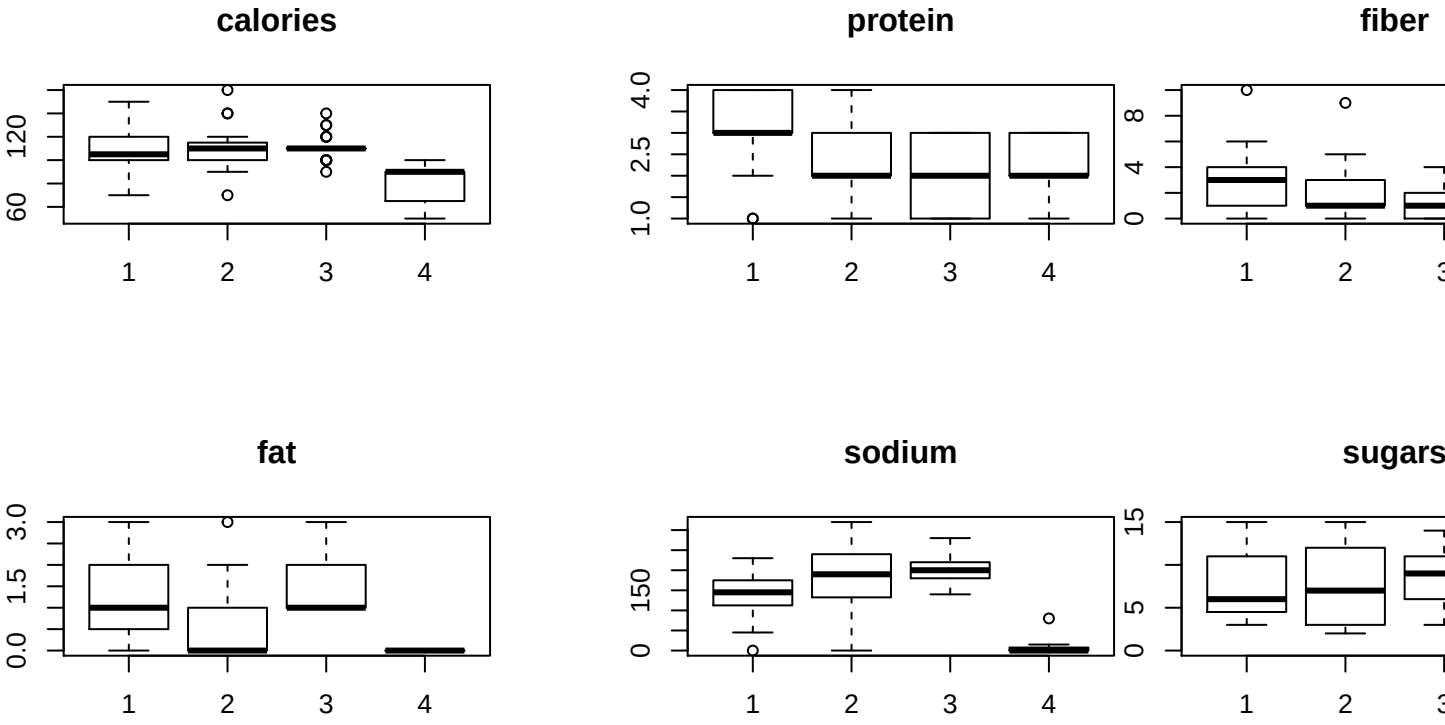
```
## [1] 4
```

```
# Passar cluster para dataframe
ce_pamk$cluster = kk$pamobject$clustering
```

Visualizacao dos resultados

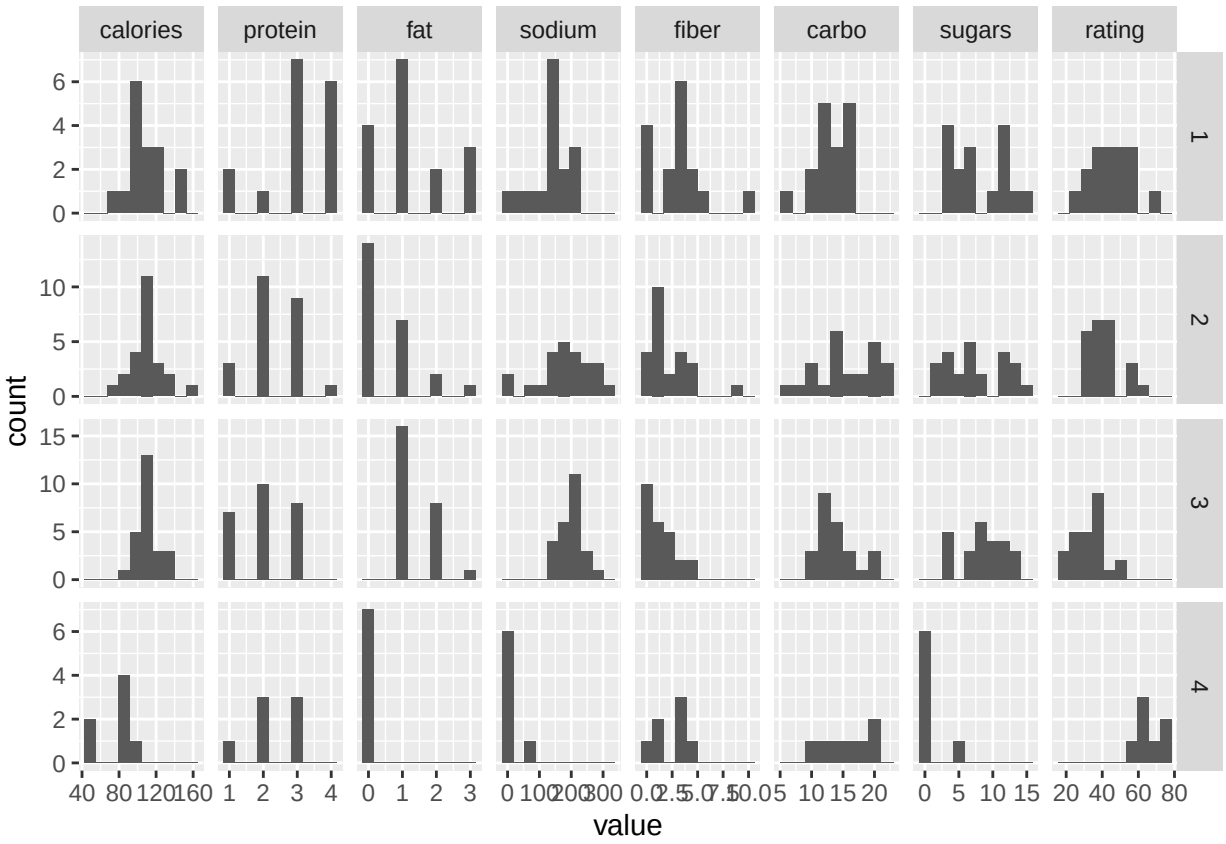
Variaveis Quant

```
par(mfrow=c(2,2))
v = ce_pamk[3:10]
n = 1
for (i in v){
 boxplot(i~ce_pamk$cluster, main = colnames(v)[n], xlab = NULL, ylab = NULL)
  n = n + 1
}
```



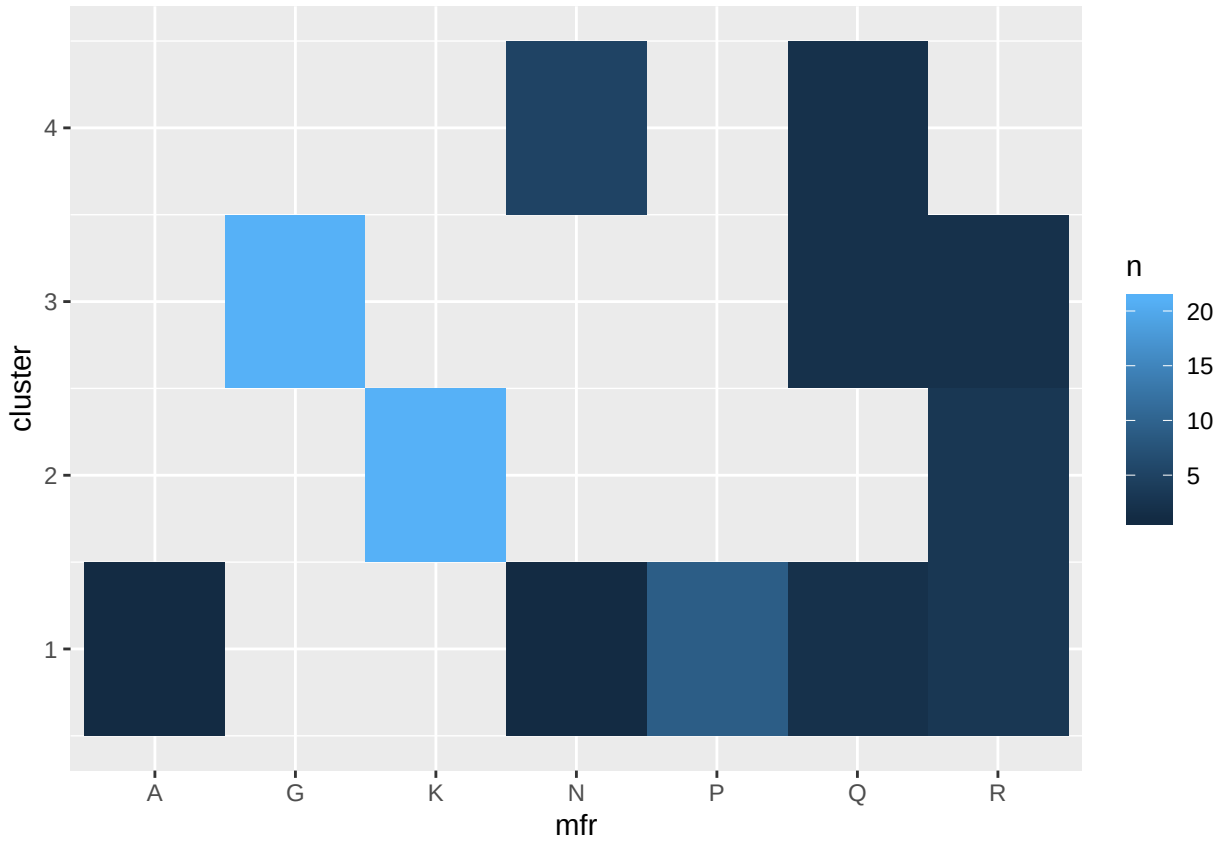
Variaveis Hist

```
df = melt(ce_pamk[3:11], id.vars = "cluster", variable.name = 'series' )

ggplot(df, aes(x = value)) + geom_histogram(bins = 10) +
  facet_grid(cluster ~ series, scales = "free")
```

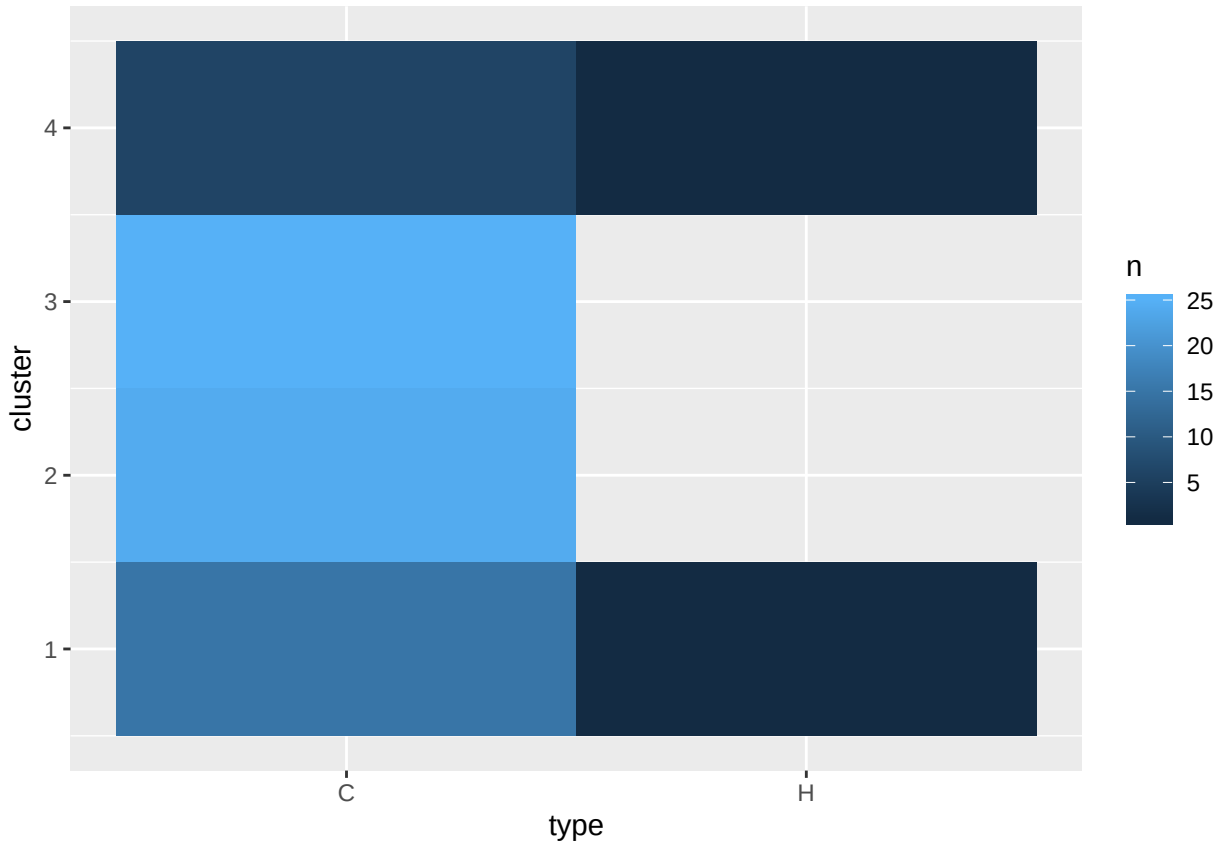


Variaveis Quali

```
d = group_by(ce_pamk,cluster, mfr) %>% summarise(n = n())
ggplot(data = d, aes(mfr, cluster)) +
  geom_raster(aes(fill = n))
```

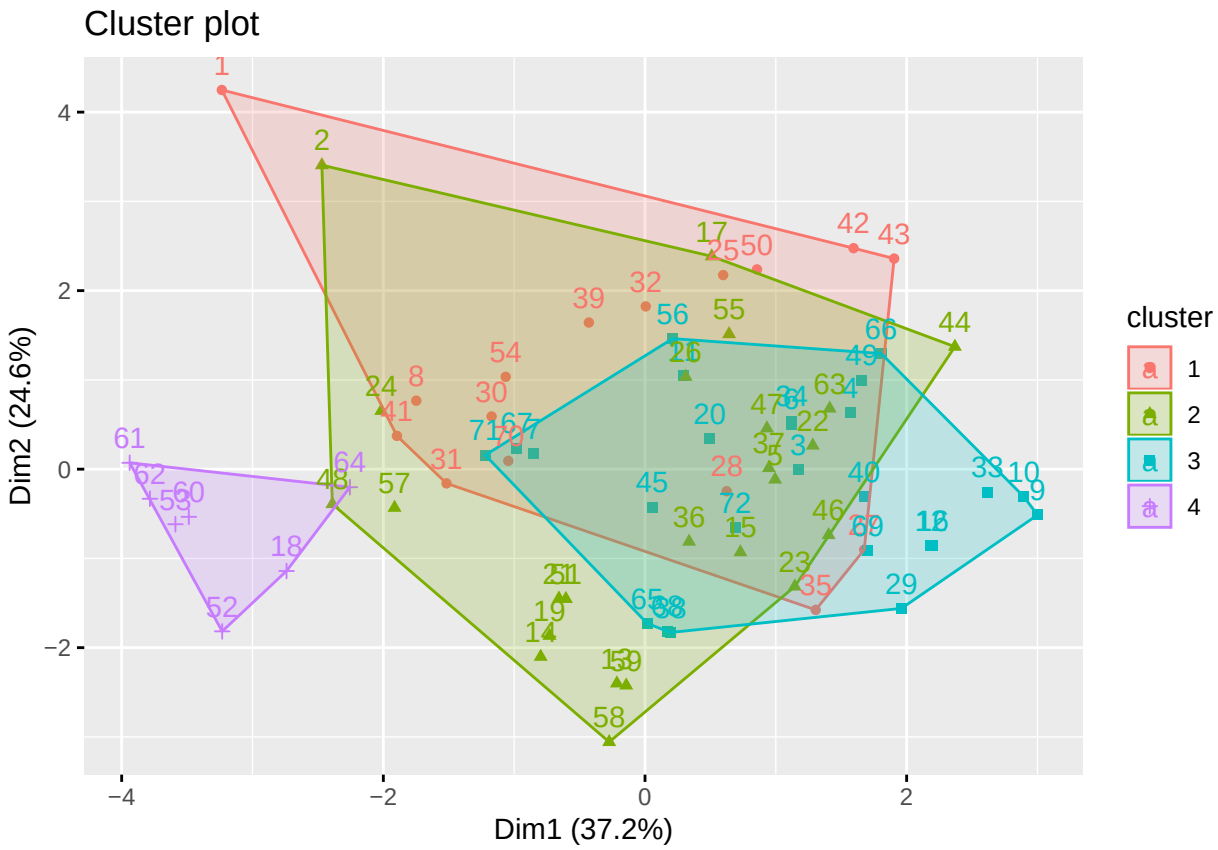



```
d = group_by(ce_pamk,cluster, type) %>% summarise(n = n())
ggplot(data = d, aes(type, cluster)) +
  geom_raster(aes(fill = n))
```



Representacao PCA

```
fviz_cluster(list(data = ce_pamk[3:10], cluster = ce_pamk$cluster), show.clust.cent = F)
```


Cluster plot
Dim2 (24.6%)
Dim1 (37.2%)
cluster
1
2
3
4