Instituto de Matemática e Estatística
Universidade de São Paulo

Lista 2

MAE0330 - Análise Multivariada de Dados

Profª Lucia Pereira Barroso

Bruno Groper Morbin - *nºUSP 11809875*

Luigi Pavarini de Lima - *nºUSP 11844642*

São Paulo
14 setembro, 2022

Parte 1: Conjunto de dados `carros`

Primeiramente serão analisados os dados encontrados no arquivo `carros.xls` que possuí os rótulos dos modelos de carro observados, além das seguintes variáveis:

- cilindrada: em cm^3
- desempenho: velocidade máxima, em km/h
- consumo: consumo urbano em km/l de gasolina
- autonomia: em km (gasolina)
- potencia: potência máxima em cv a aproximadamente 6000 rpm
- aceleracao: de 0 a 100 km, em segundos

O objetivo desta análise descritiva será utilizar o algoritmo hierárquico para agrupar os modelos de carros em diferentes grupos com características semelhantes.

Análise Drescritiva

Relembrando as Faces de Chernoff da lista passada, em que a dupla concordou acerca do número de grupos formados (3 grupos foram formados).

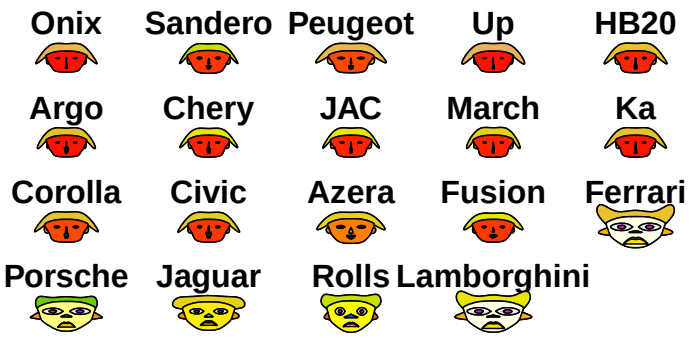

Figura 1: Faces de Chernoff

Seguiremos essa avaliação com mais ferramentas, buscando interpretar os resultados afim de validar o nossa escolha na lista 1.

Com isso então, chegamos a conlusão que após a padronização por média e desvio padrão a medida de parecença que melhor traduz as unidades avaliadas é a euclidiana e por isso a adotamos.

Métodos e Algoritmos

Em seguida, para métodos hierárquicos vamos avaliar 3 algortimos diferentes (decididos a partir de uma pré-avaliação mais qualitativa) afim de eleger apenas 1 e adotá-lo para os passos seguintes da análise. São eles: Vizinhos mais próximos(“complete”), Vizinhos mais distantes(“single”) e Ward (“Ward.D2”).

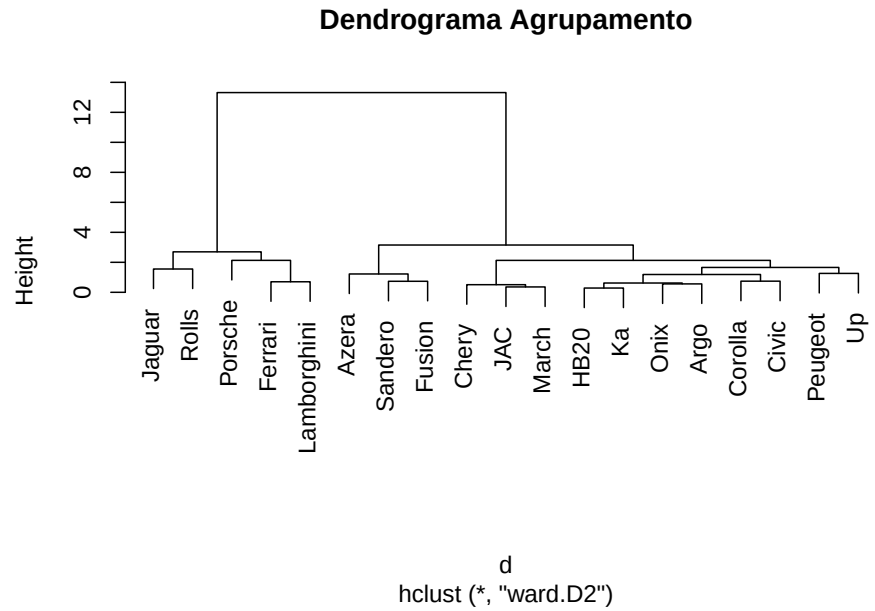

Figura 2: Dendrograma Ward

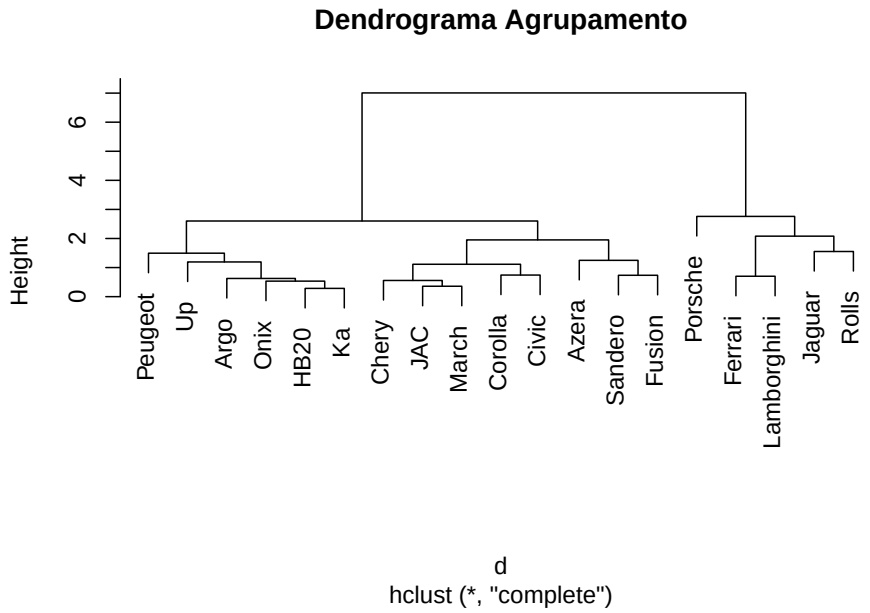

Figura 3: Dendrograma Vizinhos mais prox.

Notemos que para os dois primeiros métodos apresentados, a formação de grupos tem uma “evolução” bastante rápida onde o que mais se percebe é que há uma diferença bastante forte entre dois grupos. No entando, gostariamos de considerar 2 ou 3 grupos.

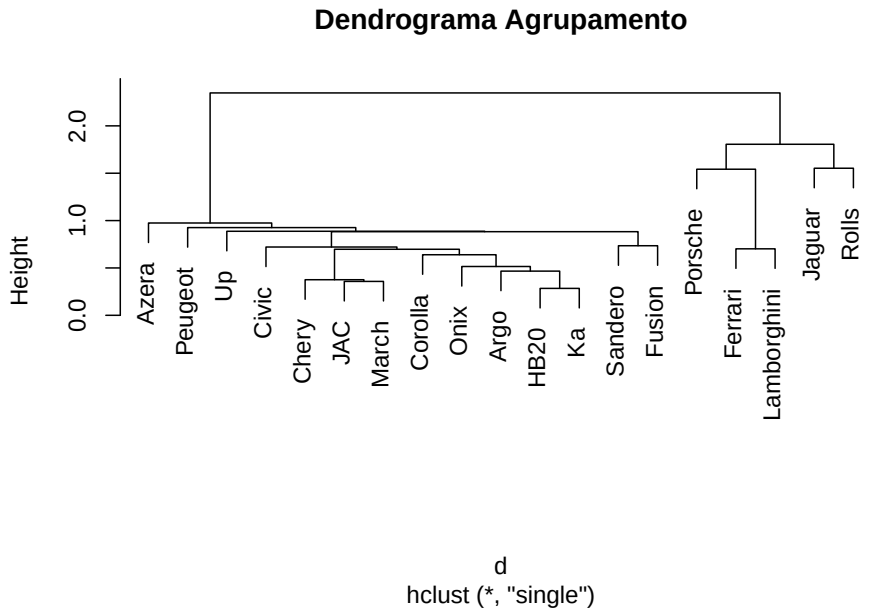

Figura 4: Dendrograma Vizinhos mais dist.

Neste caso, já notamos uma evolução que condiz melhor com a interpretação que tivemos a priori. A formação de 3 ou 4 grupos ainda parece razoável apesar da diferença ainda marcante entre dois grandes grupos. Decidimos então, seguirmos com um Cutree de 3 grupos como mostra a figura seguinte.

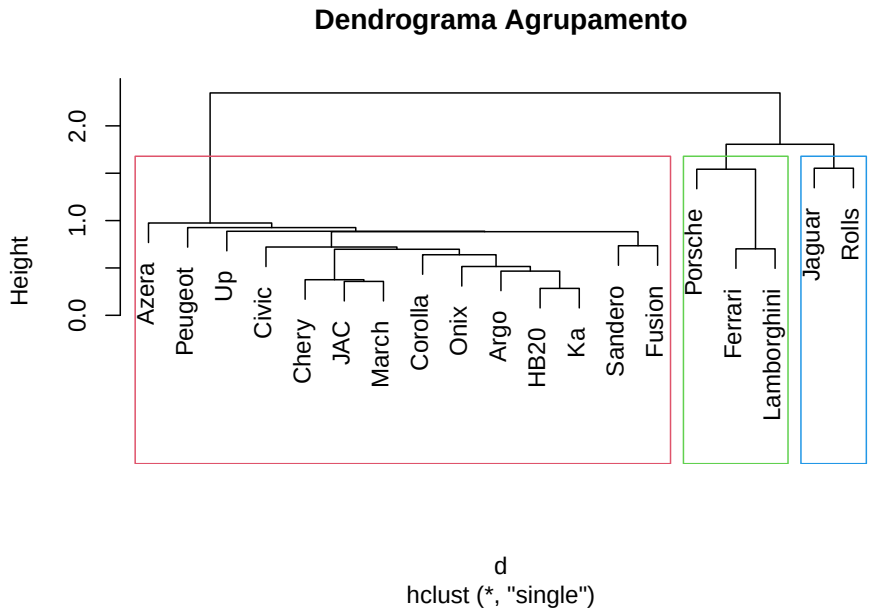

Figura 5: Dendrograma Vizinhos mais dist. (k=3)

Validação do Agrupamento

Além disso, obtivemos a seguinte correlação cofenética: 0.9219, notemos que é um coeficiente de correlação muito próximo de 1, indicando que houve uma perda de informação mínima já que a distância do dendrogama (recuperada) é diretamente proporcional à distância original dos pontos. Nesse sentido, seguimos então para o gráfico de silhueta.

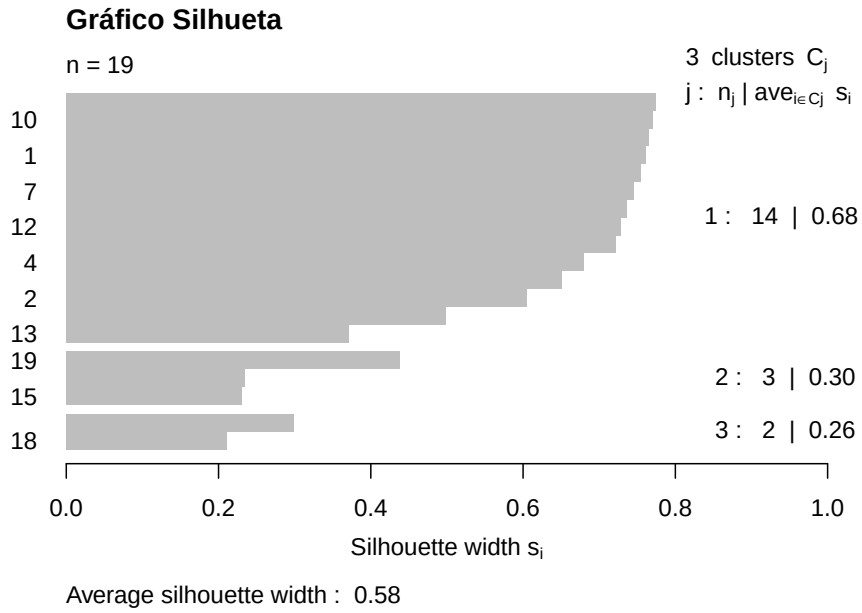

Figura 6: Gráfico de Silhueta

Notemos que na formação do primeiro grupo apenas duas marcas de carro tiveram comprimento de silhueta menor de 0.5, o que forma geral demonstra que as marcas daquele grupo foram bem alocadas. Já nos dois outros grupos formados, acreditamos que pela quantidade de elementos, ficamos meio sensíveis a qualquer diferença entre as marcas de carro, entretanto apesar dos resultados distantes de 1 eles não são negativos garantindo assim, mesmo que minimamente, uma boa alocação das marcas nos agrupamentos.

Característica dos Grupos

```
$‘1‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    999    1498    1594    1685    1994    2999

$‘2‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5733    6114    6496    6242    6497    6498

$‘3‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5000    5398    5796    5796    6194    6592

$‘1‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  181.0   184.8   190.5   196.1   200.2   235.0

$‘2‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    330     335     340     340     345     350

$‘3‘
```

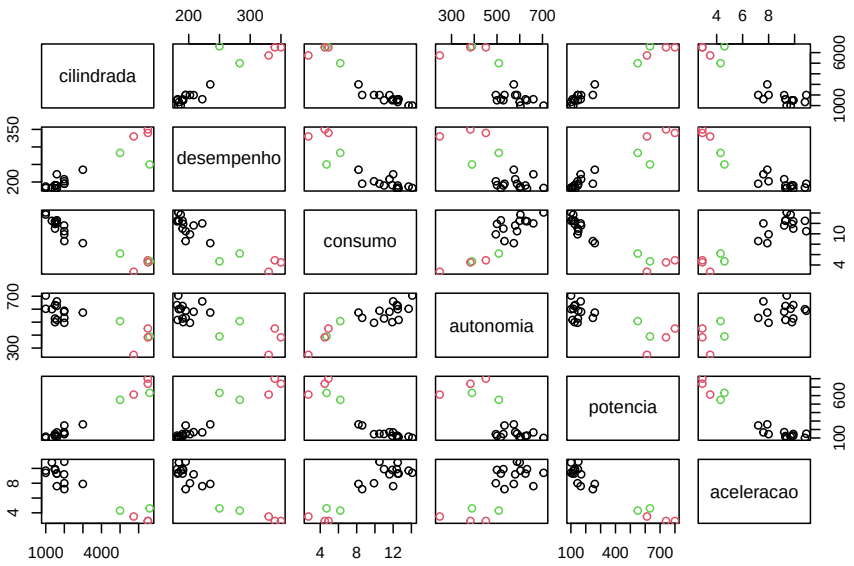

Figura 7: Painel de Pontos

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  250.0   258.2   266.5   266.5   274.8   283.0

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.20   10.62   11.95   11.54   12.50   14.10

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.700   3.600   4.500   4.033   4.700   4.900

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.700   5.075   5.450   5.450   5.825   6.200

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  495.0   529.2   584.0   581.4   619.5   705.0

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  248.0   315.5   383.0   360.7   417.0   451.0

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  390.0   419.5   449.0   449.0   478.5   508.0

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  101.0   117.8   136.5   149.4   162.0   261.0

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  612.0   676.0   740.0   717.3   770.0   800.0

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  550.0   570.5   591.0   591.0   611.5   632.0

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.200   8.300   9.350   9.207   9.875  10.900

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.9     2.9     2.9     3.1     3.2     3.5

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   4.375   4.450   4.450   4.525   4.600
```

Notemos que 2 a 2 as variáveis, quando separamos os grupos por cor, vemos que de forma geral a escolha da alocação para cada marca de carro foi bem coerente com nossas primeiras suspeitas ao usar o método do "Vizinho mais distante".Onde de fato, num painel como esse, conseguimos avaliar que há um critério de separação por cores que conseguimos detectar "no olho" depois de usar uma série de ferramentas quantitativas. Além disso vemos que para cilindrada vemos que as médias dos grupos 2 e 3 (que abrigam os carros de luxo) são bem maiores que as do grupo 1. No quesito desempenho há uma separação bem definida dos grupos que é validada também pela descritiva das variáveis o que também ocorre em Consumo, entretanto enquanto o grupo 2 tem maior média em desempenho e menor em Consumo, o grupo 1 tem a menor média de desempenho e a maior de consumo. Assim como em consumo, ocorre de forma semelhante em autonomia. Já em Potencia, os grupos 2 e 3 estão mais próximos em valores do que o grupo 1, onde o 2 tem a maior potência e o 1 a menor. Por fim, em Aceleração ocorre novamento que os valores do grupo 2 e 3 estão mais proximos do que do 1, mas diferente da variável passada o grupo 2 tem a menor média e o 1 a maior.

Parte 2: Conjunto de dados `iris`

Nesta segunda parte do relatório, serão analisados os dados do arquivo `iris.xls` que contém informações sobre espécies de flores, composto por 4 colunas (variáveis) quantitativas e 1 coluna categórica - nome da espécie - que será utilizada para comparar o agrupamento obtido a partir do algoritmo k-médias.

Tem-se o seguinte dicionário das colunas:

- Especie: Nome da espécie
- CS: Comprimento da sépala
- LS: Largura da sépala
- CP: Comprimento da pétala
- LP: Largura da pétala

Análise Descritiva

`Estrutura das variáveis:`

```
------------------------------------------------------------------------

tibble [150 x 5] (S3: tbl_df/tbl/data.frame)
 $ Especie: chr [1:150] "setosa" "setosa" "setosa" "setosa" ...
 $ CS     : num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ LS     : num [1:150] 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ CP     : num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ LP     : num [1:150] 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...

Medidas Resumo das variáveis:
------------------------------------------------------------------------

       CS              LS              CP              LP       
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  

 SD     :0.828   SD     :0.436   SD     :1.765   SD     :0.762  
```

É possível observar que as variáveis possuem pequena diferença em sua ordem de grandeza, sendo mais evidente essa observação quando analisadas os desvios padrão (SD - Standard Deviation) na tabela de medidas resumo. Dessa forma, entende-se que é coerente padronizar as variáveis para prosseguir na análise.

Algoritmo K-médias

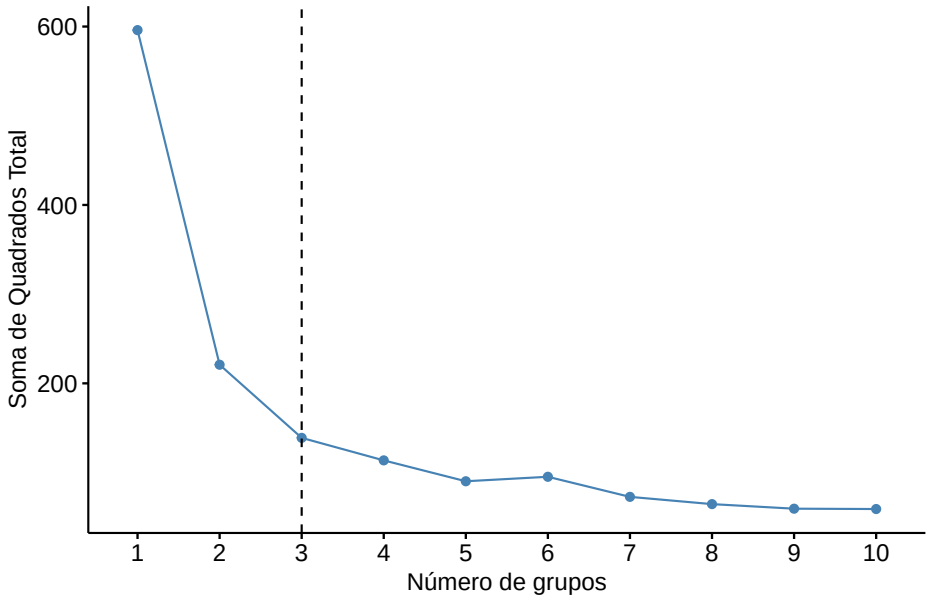

Figura 8: Linha de diferença da variabilidade pelo número de grupos

Analisando a diferença da variabilidade conforme o número de grupos, nota-se que a partir de 4 grupos a diferença entre as variabilidades apresenta-se muito pequena. Sendo assim, escolhe agrupar em apenas 3 classes.

Avaliação do agrupamento

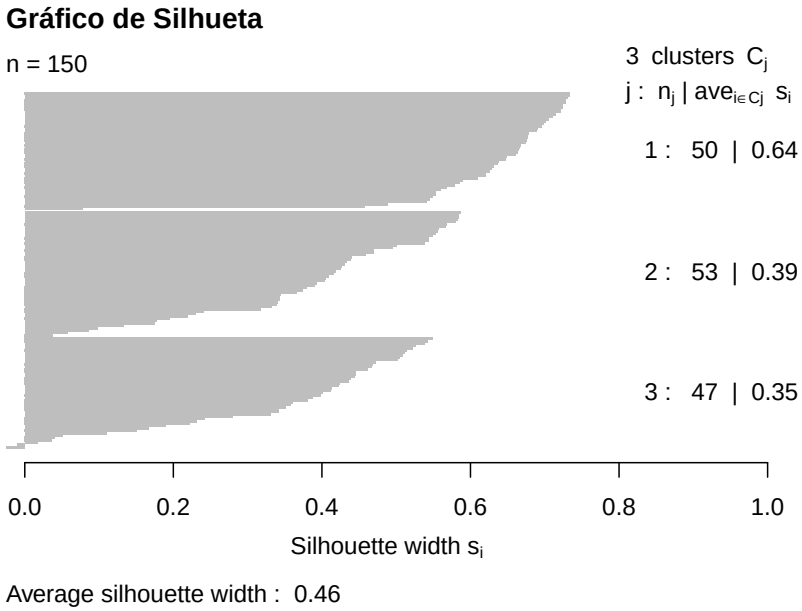

Figura 9: Gráfico de Silhueta para o algoritmo k-médias no conjunto de dados iris

```
Média de cada variável por cada grupo:
------------------------------------------------------------------------

           CS          LS         CP         LP
1 -1.01119138  0.85041372 -1.3006301 -1.2507035
2 -0.05005221 -0.88042696  0.3465767  0.2805873
3  1.13217737  0.08812645  0.9928284  1.0141287

Soma de Quadrados Total (Variabilidade) dos grupos:
------------------------------------------------------------------------
        1       2       3

47.35062 44.08754 47.45019

Tamanho dos grupos:
------------------------------------------------------------------------
1   2   3

50 53 47

Tabela de confundimento:
------------------------------------------------------------------------


             1  2  3
  setosa     50  0  0
  versicolor  0 39 11
  virginica   0 14 36
```

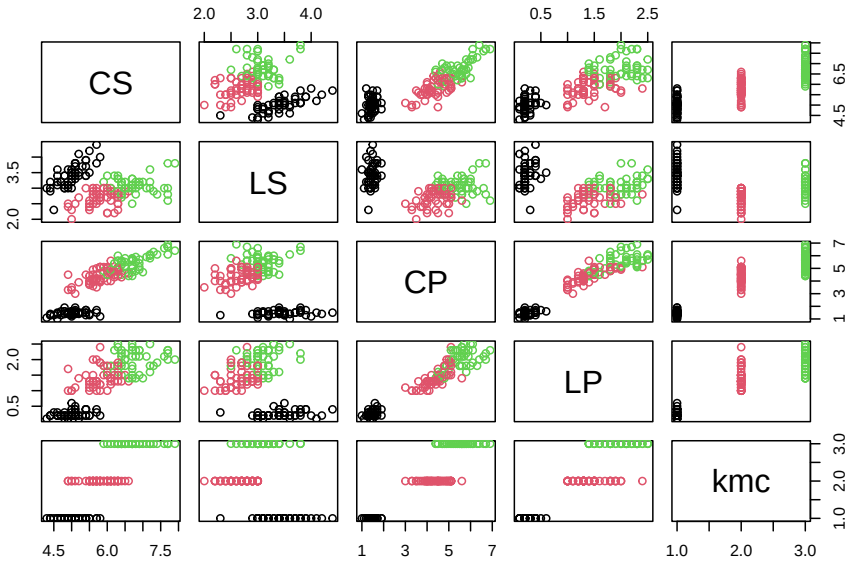

Figura 10: Gráfico de dispersão das variáveis combinadas 2 a 2 com grupos destacados pelas diferentes cores

Observando a tabela de confundimento, nota-se que o algoritmo k-médias escolheu "erroneamente" 25 observações relacionadas a suas respectivas espécies. Destaca-se o grupo 1 que coincidiu perfeitamente com as flores de espécie "setosa".

Características dos grupos

Pela Figura 9, e pela tabela de médias dos grupos, nota-se que o grupo 1 (setosa) possui os menores valores de CS (comprimento da sépala), enquanto que o grupo 3 (majoritariamente composto pela espécie virginica) apresentou os maiores valores desta variável. Já para a variável `LS`, o grupo 1 apresentou os maiores valores enquanto que o grupo 2 apresentou os menores. Em relação à `CP` (comprimento da pédala), os grupos 1 e 3 se constrastam novamente sendo o primeiro apresentando os menores valores e o segundo, os maiores relativos a esta variável. Para a largura da pédala, a diferença é destacada entre os grupos 1 e 3, na mesma ordem da variável `CP`.

Apêndice

Código

```
library(readxl)
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
carros <- read_excel("carros.xls")
iris <- read_excel("iris.xls")
carros1 <- carros[c("cilindrada","desempenho","consumo","autonomia","potencia","aceleracao")]
carrosp <- scale(carros1)

library(aplpack)

### plotando chernoff pra justificar os proximos passos
```

```
faces(carros1, labels=carros$Nome)

d<-dist(carrosp,method = "euclidean")

library(cluster)
hc1<-hclust(d,method="ward.D2")
plot(hc1, labels = carros$Nome, main = "Dendrograma Agrupamento")

hc2<-hclust(d,method="complete")
plot(hc2, labels = carros$Nome, main = "Dendrograma Agrupamento")

hc3<-hclust(d,method = "single")
plot(hc3, labels = carros$Nome, main = "Dendrograma Agrupamento")

plot(hc3, labels = carros$Nome, main = "Dendrograma Agrupamento")
rect.hclust(hc3 , k = 3, border = 2:6)
abline(h = 3, col = 'red')

d.coph <- cophenetic(hc3)
x <- cor(d, d.coph)

grupo <- cutree(hc3,3)
silhuetac <- silhouette(grupo,d)
plot(silhuetac,main = "Gráfico Silhueta")

dados <- data.frame(carros, grupo)
library(lattice)
# todos os grupos em um mesmo painel com cores diferentes
plot(carros1, col=dados$grupo)

tapply(dados$cilindrada, dados$grupo, summary)
tapply(dados$desempenho, dados$grupo, summary)
tapply(dados$consumo, dados$grupo, summary)
tapply(dados$autonomia, dados$grupo, summary)
tapply(dados$potencia, dados$grupo, summary)
tapply(dados$aceleracao, dados$grupo, summary)

cat(paste0("Estrutura das variáveis:\n","--------------------------\
           ------------------------------------------"))
str(iris)

# Selecionando as colunas quantitativas
iris1 <- iris[,2:5]

# Mostrando medidas resumo das variáveis
cat(paste0("Medidas Resumo das variáveis:\n","--------------------------------\
----------------------------"))
summary(iris1)
for (i in c(1,2,3,4)){
  cat(paste0(" SD     :", format(sd(unlist(iris1[,i])),digits=3,nsmall = 3))," ")
}

irisp <- scale(iris1)
```

```
library(lattice)

dt<-dist(irisp,method = "euclidean")

irisd <-as.data.frame(irisp)

library(factoextra)

fviz_nbclust(irisd, kmeans, method="wss")+
  geom_vline(xintercept=3, linetype=2)+
  labs( y = "Soma de Quadrados Total", x="Número de grupos", title="")


km<-kmeans(irisp,3)
kmc <- km$cluster

dados1 <- data.frame(iris1, kmc)
dados1


silhuetak <- silhouette(kmc, dt)
silhuetak
plot(silhuetak, main="Gráfico de Silhueta")

plot(dados1, col = kmc)

cat(paste0("Média de cada variável por cada grupo:\n","------------------------------\
------------------------------------------"))
km$centers

cat(paste0("Soma de Quadrados Total (Variabilidade) dos grupos:\n","----------------------\
-----------------------------------------------\n","\t\t1\t\t2\t\t3"))
cat(km$withinss)

cat(paste0("Tamanho dos grupos:\n","--------------------------------------\
           ------------------------------\n","1   2   3"))
cat(km$size)

irisd <- data.frame(iris, kmc)

cat(paste0("Tabela de confundimento:\n","----------------------------------\
           ----------------------------------"))
table(irisd$Especie, irisd$kmc)
```