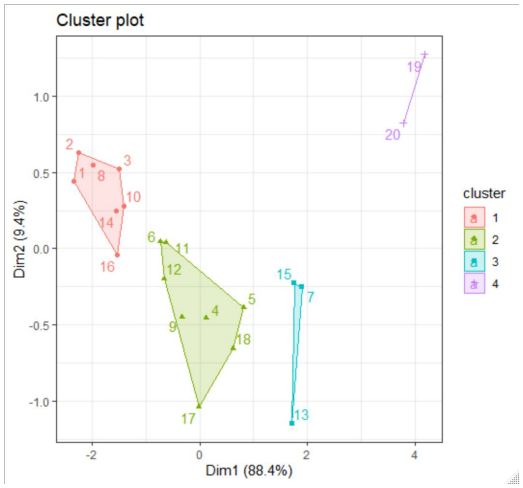
Ejercicio 1.

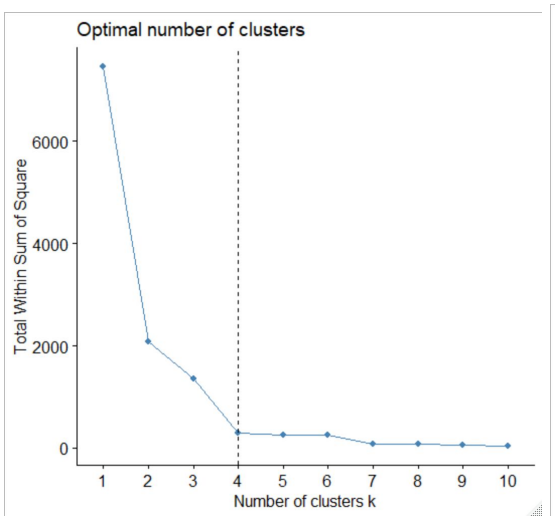
Primero calculo la matriz de distancias euclidianas para visualizar los datos. Uso la distancia euclidiana porque los datos son **numéricos** y "la norma" de los vectores representa una buena distancia para el análisis. Elijo el método **hc jerárquico** para poder **visualizar los datos antes de elegir el k**.
Con la misma calculo las correlaciones con los 4 métodos de cluster jerárquico (complete, average, ward, single). Todas me dan en razón de (0.90,0.93). Elijo Average por 0.9278.
Realizo cluster jerárquico con hc y datos en scale y obtengo el siguiente dendograma, elijo k=4

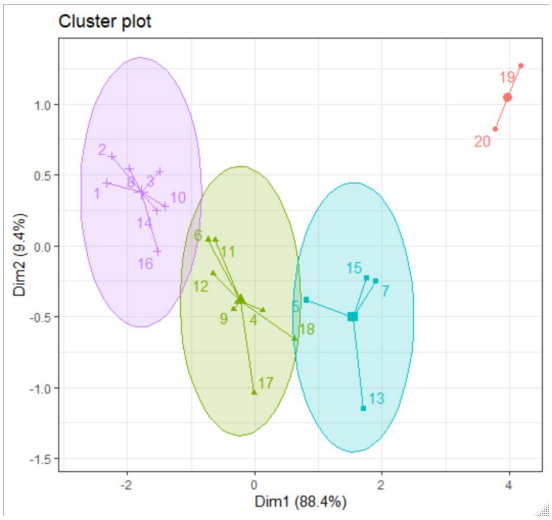


Siendo los grupos:

1 CABALLO	1 HAMSTER	1 RENO	1 FOCA
2 BURRO	2 RATA	2 CONEJO	2 DELFÍN
3 CEBRA	3 OVEJA	3 CIERVO	
4 MULA	4 CERDO		
5 CAMELLO	5 BÚFALO		
6 LLAMA	6 ZORRO		
7 BISONTE	7 GATO		
	8 PERRO		

Pruebo tambien con k-means para k=4, y de yapa verificamos que k=4 es óptimo





Ahora veamos los vectores medios (usando HC jerárquico, ya que ambos dieron muy parecidos pero el resultado final agrupado me gusto mas):

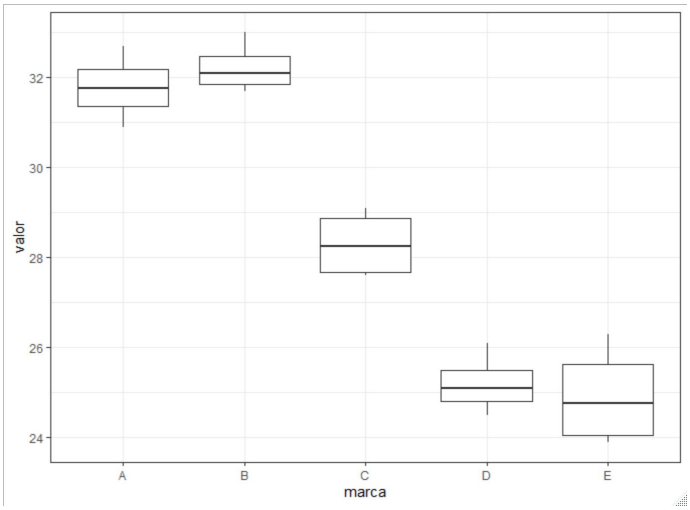
Agua Proteínas Grasa Lactosa

88.242857 3.071429 2.471429 5.714286
    Agua Proteínas   Grasa   Lactosa
 80.1000   7.6500   7.6125   3.9250
    Agua Proteínas   Grasa   Lactosa
67.333333 11.133333 17.700000 2.333333
    Agua Proteínas   Grasa   Lactosa
   45.65   10.15   38.45   0.45

Vemos como describen muy bien las diferencias de cada cluster ya que se alejan las medias en todas las dimensiones. Si comparamos el 1 y el 4 vemos que la diferencia es abismal en todas las variables.

Ejercicio 2.

Lo primero que hago luego de importar la data es dibujar el biplot para ver cómo se distribuyen los datos.



Se puede apreciar que las formas son parecidas (La mediana se acerca a la media) pero están distribuidas en tres grupos.

Aplico Shapiro-Wilk para testear normalidad de las marcas:

p-value = 0.9916 p-value = 0.6135 p-value = 0.216 p-value = 0.85 p-value = 0.4546

No rechazo normalidad en **ninguna** de las marcas.

Pruebo homocedasticidad con **Bartlett** y **Levene**

Bartlett: Bartlett's K-squared = 1.4337, df = 4, **p-value = 0.8383**

Levene: Test Statistic = 1.2337, **p-value = 0.3384**

No rechazo en ninguno de los dos casos, así que se **cumplen normalidad y homocedasticidad**, puedo asumir los supuestos de anova. Luego **aplicando anova**:

            Df Sum Sq Mean Sq F value   Pr(>F)
marca   4 193.24   48.31   74.53 1.03e-09 ***
Residuals   15   9.72   0.65

El p-valor es muy chico, así que **rechazo H0** y asumo que existen medias distintas.

Aplicando el test de **Tukey**:

      diff   lwr   upr   p adj
B-A 0.450 -1.307903 2.207903 0.9294510
C-A -3.475 -5.232903 -1.717097 0.0001688
D-A -6.575 -8.332903 -4.817097 0.0000001
E-A -6.850 -8.607903 -5.092097 0.0000000
C-B -3.925 -5.682903 -2.167097 0.0000435
D-B -7.025 -8.782903 -5.267097 0.0000000
E-B -7.300 -9.057903 -5.542097 0.0000000

D-C -3.100 -4.857903 -1.342097 0.0005542
E-C -3.375 -5.132903 -1.617097 0.0002306
E-D -0.275 -2.032903 1.482903 0.9877881

Los p-valores de las "**t**" rechazan en todas menos en B con A y en E con D con un **IC de 95%**. Siendo A-B cerveza rubia, D-E cerveza negra, y **C que es negra pero rechaza contra todos**.

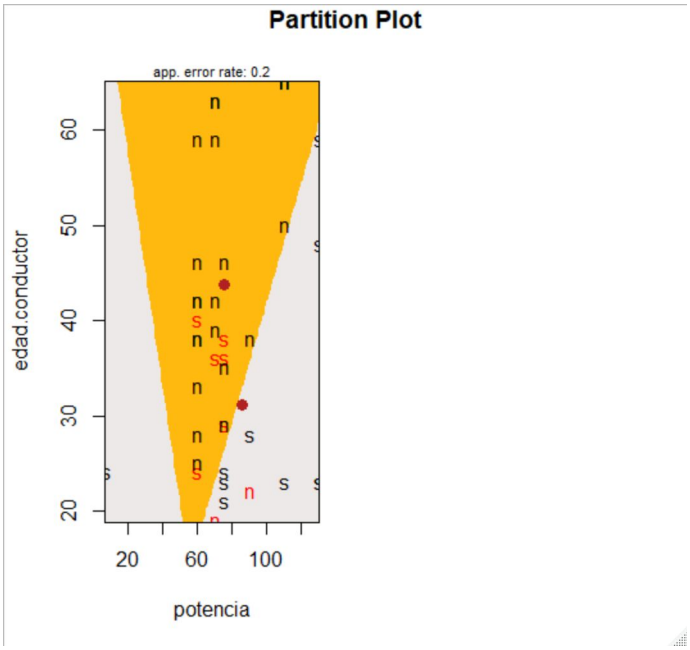
Ejercicio 3.

Lo primero que hice fue sacar la columna id que no era pertinente para el analisis, hacer factor la columna grave (Para que solo pueda tener 2 valores), y luego calcular los vectores de medias para ambas clases:

```
> autos.si.mean
   antigüedad edad.conductor      potencia
       6.4375       31.1875       85.4375
> autos.no.mean
   antigüedad edad.conductor      potencia
     5.208333      43.791667      75.208333
```

Luego el test de Hotelling arroja: **P-value: 0.0009905** Rechazando H0. Veamos los demas
M-shapiro: **p-value = 0.001458** Rechaza
Box's M: **p-value = 0.001754** Rechaza

Para evaluar las técnicas utilizaremos la **tasa de error ingenuo** y el grafico **partimat**.
Probamos LDA: La tasa de error ingenuo me dio **22.4** (Altisimo)
Probamos QDA: La tasa de error ingenuo me dio **17.5** (Mejoró un poco pero sigue siendo alto)



Viendo la distribución de los datos en **partiplot**, un método robusto no lo soluciona porque no están mal clasificados por outliers sino por intersección (Todos los valores superpuestos que se ven en el medio). La única manera de solucionar esto es con una SVM logrando una transformación (Que queda fuera del scope de este ejercicio, ya que no son linealmente separables).