

Análisis de MTCARS

Obtener datos:

```
> library (datasets)      # necesario sólo si usas Rcmdr
> ? mtcars                 # descripción
> M <- mtcars              # para ahorrar pulsaciones
```

Las variables de este dataframe son:

mpg : millas por galón

cyl : número de cilindros

dis : cubicaje en pulgadas cúbicas (desplazamiento de los cilindros)

hp : potencia en caballos de vapor

drat : razón entre ejes (número de revoluciones de la transmisión por cada vuelta de rueda)

wt : peso en miles de libras

qsec : segundos para recorrer un cuarto de milla

vs : tipo de cilindros (en V; en línea o Serie)

am : tipo de transmisión (automática; manual)

gear : número de marchas hacia adelante

carb : número de carburadores

1. Transformar millas por galón en litros por 100 km

Consideraremos:

- 1 milla $\simeq$ 1,609 km
- 1 galón $\simeq$ 3,785 litros

Cadena de trasformaciones:

```
> km.galon      <-      M$mpg * 1.609          # kilómetros por galón
> km.litro      <-      km.galon / 3.785        # kilómetros por litro
> litros.km     <-      1 / km.litro            # litros por kilómetro
> litros.100km  <-      litros.km * 100         # litros por cada 100 km
```

o bien directamente

```
> M$L100k <- 100 * 3.785 / (1.61 * M$mpg)
```

Siempre conviene comprobar los resultados:

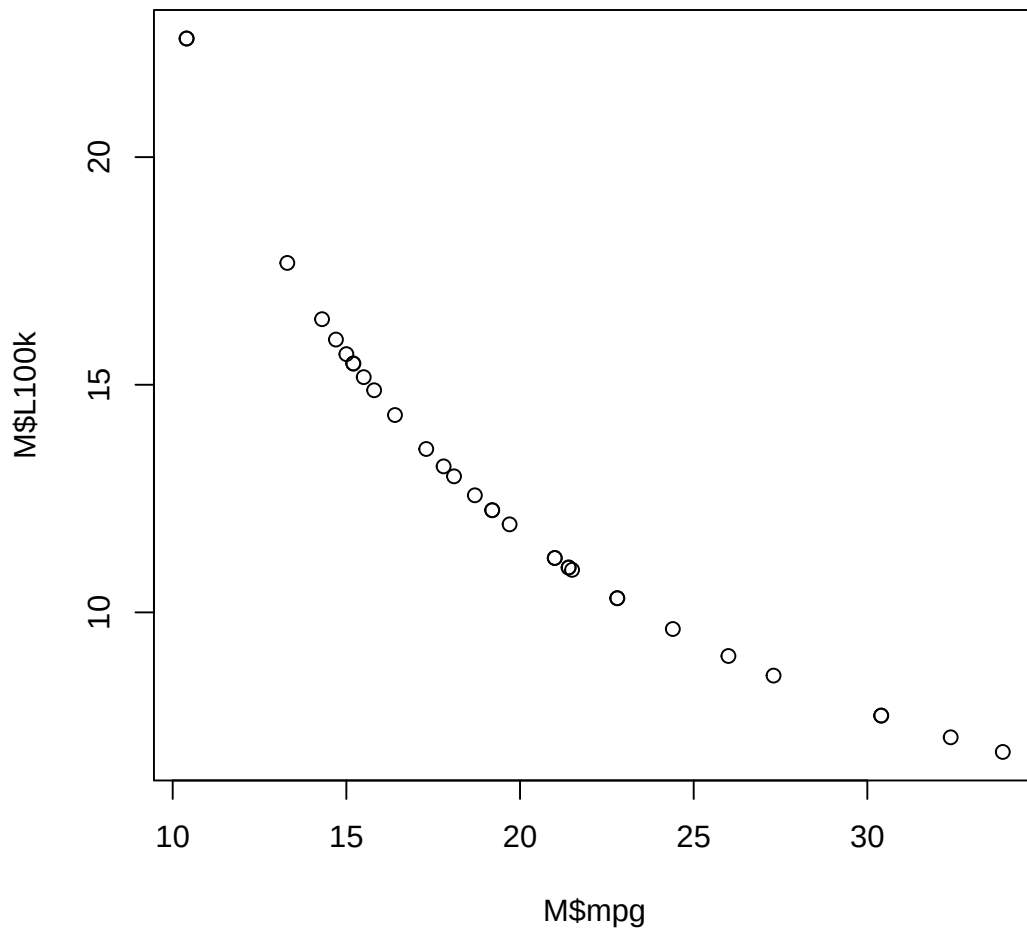
```
> head      (M[, c("mpg", "L100k")])
```

	mpg	L100k
Mazda RX4	21.0	11.19491
Mazda RX4 Wag	21.0	11.19491
Datsun 710	22.8	10.31110
Hornet 4 Drive	21.4	10.98566
Hornet Sportabout	18.7	12.57183
Valiant	18.1	12.98857

```
> summary (M[, c("mpg", "L100k")])
```

mpg	L100k
Min. :10.40	Min. : 6.935
1st Qu.:15.43	1st Qu.:10.311
Median :19.20	Median :12.244
Mean :20.09	Mean :12.748
3rd Qu.:22.80	3rd Qu.:15.242
Max. :33.90	Max. :22.605

```
> plot (M$mpg, M$L100k)
```



2. Transformar “segundos hasta 1/4 de milla” en “segundos de 0 a 100 km/h”

Suponiendo que se parte del reposo y que el movimiento es uniformemente acelerado, las fórmulas que relacionan el espacio recorrido, la velocidad alcanzada y el tiempo empleado son:

$$\begin{aligned}e(t) &= \frac{1}{2} a t^2 \\v(t) &= a t\end{aligned}$$

Aplicaremos estas expresiones a los datos de `mtcars`. Debemos tener cuidado sobre qué unidades usamos en las fórmulas:

1. espacio

$$\begin{aligned}\frac{1}{4} \text{ de milla} &= \frac{1,609}{4} \text{ km} = 0,402 \text{ km} = e(\text{qsec}) = \frac{1}{2} a \text{qsec}^2 \\ \Rightarrow a &= \frac{2 \times 0,402 \text{ km}}{\text{qsec}^2} = \frac{0,804 \text{ km}}{\text{qsec}^2}\end{aligned}$$

Esa aceleración a está expresada en km/s^2 (kilómetros por segundo cada segundo), por lo que vamos a expresar la velocidad en km/s (kilómetros por segundo) en vez de km/h .

2. velocidad

$$\begin{aligned}100 \text{ km/h} &= \frac{100}{3600} \text{ km/s} = \frac{1}{36} \text{ km/s} = \frac{1 \text{ km}}{36 \text{ s}} = v(t) = a t = \frac{0,804 \text{ km}}{\text{qsec}^2} t \\ \Rightarrow t &= \frac{1 \text{ km} \cdot \text{qsec}^2}{0,804 \text{ km} \cdot 36 \text{ s}} = \frac{\text{qsec}^2}{28,944 \text{ s}}\end{aligned}$$

Cálculo y comprobación:

```
> M$T100k <- M$qsec^2 / 28.944      # tiempo en segundos de 0 a 100 km/h
> head (M[, c("qsec", "T100k")])
```

	qsec	T100k
Mazda RX4	16.46	9.360544
Mazda RX4 Wag	17.02	10.008306
Datsun 710	18.61	11.965592
Hornet 4 Drive	19.44	13.056716
Hornet Sportabout	17.02	10.008306
Valiant	20.22	14.125498

```
> summary (M[, c("qsec", "T100k")])
```

	qsec	T100k
Min.	:14.50	Min. : 7.264
1st Qu.:	16.89	1st Qu.: 9.859
Median	:17.71	Median :10.837
Mean	:17.85	Mean :11.114
3rd Qu.:	18.90	3rd Qu.:12.341
Max.	:22.90	Max. :18.118

Un promedio de once segundos en pasar de 0 a 100 es razonable.

Transformar otras variables

Ya que hemos hecho la transformación de `mpg` y `qsec` a unidades europeas, haremos lo mismo con

disp (cubicaje): pulgadas cúbicas $\rightarrow$ centímetros cúbicos

wt (peso): miles de libras $\rightarrow$ toneladas

```
> M$cubi <- M$disp * 2.54^3
> M$peso <- M$wt * 0.4536
> ## dataframe con sólo variables en unidades europeas:
> M1 <- M [, c ("L100k", "cyl", "cubi", "hp", "drat", "peso",
+              "T100k", "vs", "am", "gear", "carb")]
> round (head (M1), 2)
```

	L100k	cyl	cubi	hp	drat	peso	T100k	vs	am	gear	carb
Mazda RX4	11.19	6	2621.93	110	3.90	1.19	9.36	0	1	4	4
Mazda RX4 Wag	11.19	6	2621.93	110	3.90	1.30	10.01	0	1	4	4
Datsun 710	10.31	4	1769.80	93	3.85	1.05	11.97	1	1	4	1
Hornet 4 Drive	10.99	6	4227.86	110	3.08	1.46	13.06	1	0	3	1
Hornet Sportabout	12.57	8	5899.34	175	3.15	1.56	10.01	0	0	3	2
Valiant	12.99	6	3687.09	105	2.76	1.57	14.13	1	0	3	1

3. Aplicar análisis de componentes principales

Aunque todas las variables son numéricas, un vistazo a la documentación de `mtcars` nos indica que dos de ellas, `am` (automático/manual) y `vs` (cilindros en V / en serie o línea), son cualitativas. Las excluirémos en principio del análisis.

3.1. ¿Matriz de covarianzas o de correlaciones?

Analizando las variables que quedan:

```
> head (M1 [, c("L100k", "cyl", "cubi", "hp", "drat",  
+              "peso", "T100k", "gear", "carb")])
```

	L100k	cyl	cubi	hp	drat	peso	T100k	gear	carb
Mazda RX4	11.19491	6	2621.930	110	3.90	1.188432	9.360544	4	4
Mazda RX4 Wag	11.19491	6	2621.930	110	3.90	1.304100	10.008306	4	4
Datsun 710	10.31110	4	1769.803	93	3.85	1.052352	11.965592	4	1
Hornet 4 Drive	10.98566	6	4227.863	110	3.08	1.458324	13.056716	3	1
Hornet Sportabout	12.57183	8	5899.343	175	3.15	1.560384	10.008306	3	2
Valiant	12.98857	6	3687.089	105	2.76	1.569456	14.125498	3	1

Las variables son de naturaleza muy distinta, en unidades que poco tienen que ver unas con otras, y las magnitudes de sus valores son completamente diferentes.

Por ello, realizaremos el análisis usando la matriz de correlaciones:

```
> acp <- princomp (~ L100k + cyl + cubi + hp + drat +  
+                   peso + T100k + gear + carb,  
+                   data = M1,  
+                   cor = TRUE)
```

3.2. ¿Cuántas componentes principales?

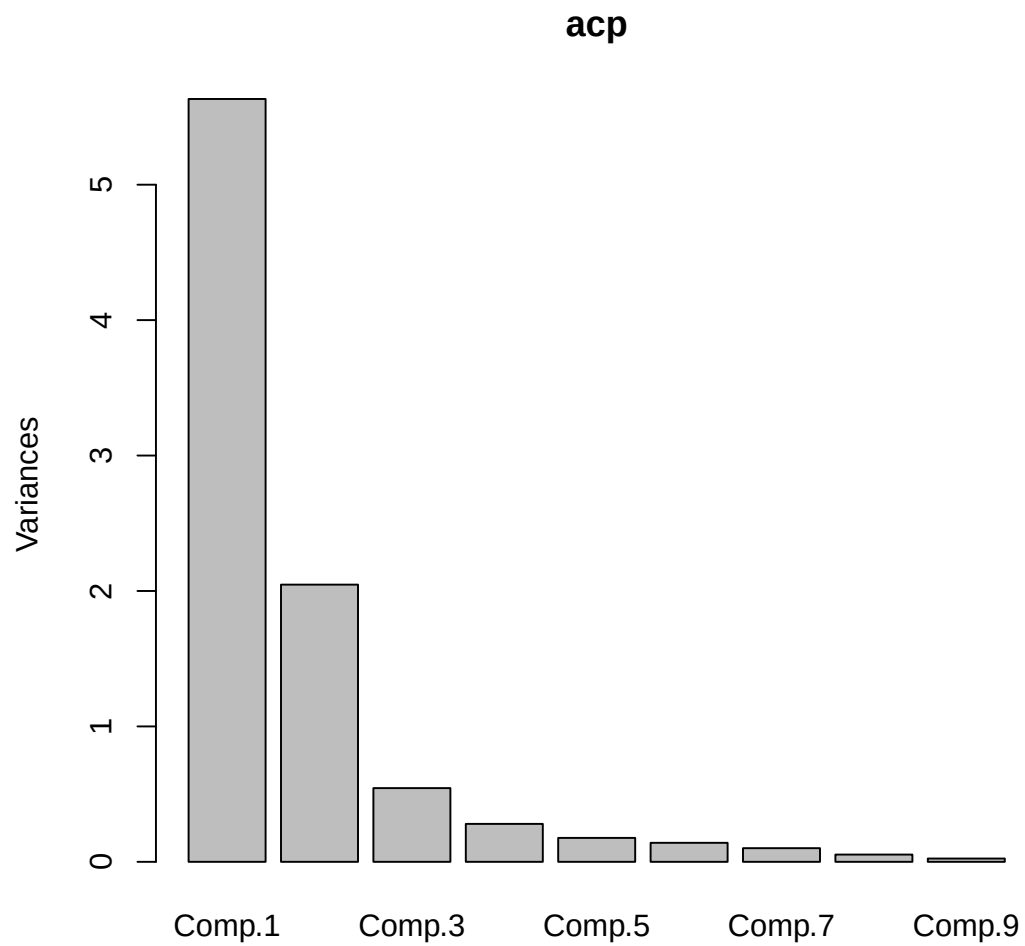
```
> summary (acp)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.3733559	1.4306692	0.73784598	0.5292828	0.42039691
Proportion of Variance	0.6258687	0.2274238	0.06049074	0.0311267	0.01963706
Cumulative Proportion	0.6258687	0.8532925	0.91378328	0.9449100	0.96454704

	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	0.37460274	0.31748552	0.231552844	0.155998830
Proportion of Variance	0.01559191	0.01119967	0.005957413	0.002703959
Cumulative Proportion	0.98013895	0.99133863	0.997296041	1.000000000

```
> plot (acp)
```



- Con una componente no se llega al 70%
- Con dos componentes se sobrepasa el 85%
- La tercera componente explica mucho menos que la segunda.
- Se aprecia en el gráfico de barras la forma del “codo” que sugiere quedarse con dos componentes.

3.3. Significado de cada componente

```
> names (acp)
```

```
[1] "sdev"      "loadings" "center"    "scale"     "n.obs"     "scores"    "call"
```

```
> acp $ loadings [, 1:2]
```

	Comp.1	Comp.2
L100k	0.3901074	0.03423539
cyl	0.4014256	-0.02238517
cubi	0.4014380	0.08525626
hp	0.3661841	-0.27442208
drat	-0.3111216	-0.33924026
peso	0.3771095	0.16834845
T100k	-0.2208211	0.47633254
gear	-0.2124568	-0.55089400
carb	0.2419139	-0.49188280

```
> round (sort (acp $ loadings [, 1]), 3)
```

drat	T100k	gear	carb	hp	peso	L100k	cyl	cubi
-0.311	-0.221	-0.212	0.242	0.366	0.377	0.390	0.401	0.401

```
> round (sort (acp $ loadings [, 2]), 3)
```

gear	carb	drat	hp	cyl	L100k	cubi	peso	T100k
-0.551	-0.492	-0.339	-0.274	-0.022	0.034	0.085	0.168	0.476

```
> head (acp$scores[,1:2])
```

	Comp.1	Comp.2
Mazda RX4	-0.7834578	-1.1930277
Mazda RX4 Wag	-0.7481317	-1.0092614
Datsun 710	-2.3982122	0.3250471
Hornet 4 Drive	-0.2968024	1.9872478
Hornet Sportabout	1.5136855	0.8127920
Valiant	-0.0482835	2.4813268

```
> names (head (sort (acp $ scores [, 1])))
```

```
[1] "Honda Civic"      "Toyota Corolla" "Fiat 128"        "Fiat X1-9"  
[5] "Porsche 914-2"   "Lotus Europa"
```

```
> names (tail (sort (acp $ scores [, 1])))
```

```
[1] "Camaro Z28"        "Duster 360"        "Maserati Bora"  
[4] "Chrysler Imperial" "Cadillac Fleetwood" "Lincoln Continental"
```

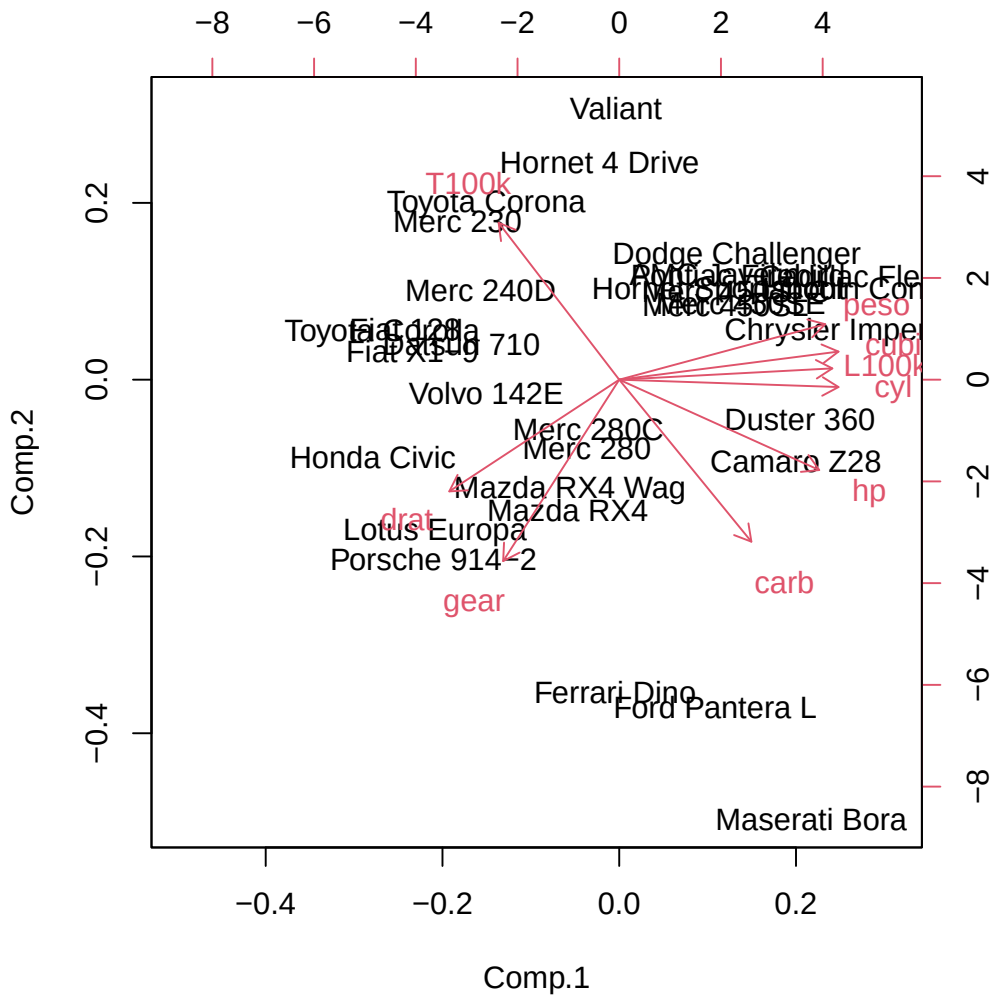
```
> names (head (sort (acp $ scores [, 2])))
```

```
[1] "Maserati Bora"    "Ford Pantera L" "Ferrari Dino"    "Porsche 914-2"  
[5] "Lotus Europa"    "Mazda RX4"
```

```
> names (tail (sort (acp $ scores [, 2])))
```

```
[1] "AMC Javelin"      "Dodge Challenger" "Merc 230"        "Toyota Corona"  
[5] "Hornet 4 Drive"   "Valiant"
```

```
> biplot (acp)
```



Primera componente

Parece ser el eje de “tamaño en general” por:

- Tiene coeficientes positivos para mayoría de variables que intuitivamente tienen relación directa con el tamaño.
- Tiene coeficiente negativo con T100k.

Matizaciones:

- Llama la atención el coef. negativo de **gear**.
- El de **drat** es difícil de interpretar; en principio,
 - **drat** pequeño = coche que responde rápidamente;
 - **drat** grande = coche que ahorra combustible;

Segunda componente

Coches con puntuaciones altas en esta componente tienen

- poca **hp**,
- poca aceleración (mucho T100k),
- pocas marchas y pocos carburadores.

Este eje podría etiquetarse como “estilo, línea o carácter del coche”.

Separaría coches con “carácter deportivo” de coches con “carácter familiar”.

3.4. Comportamiento de las CC.PP. según...

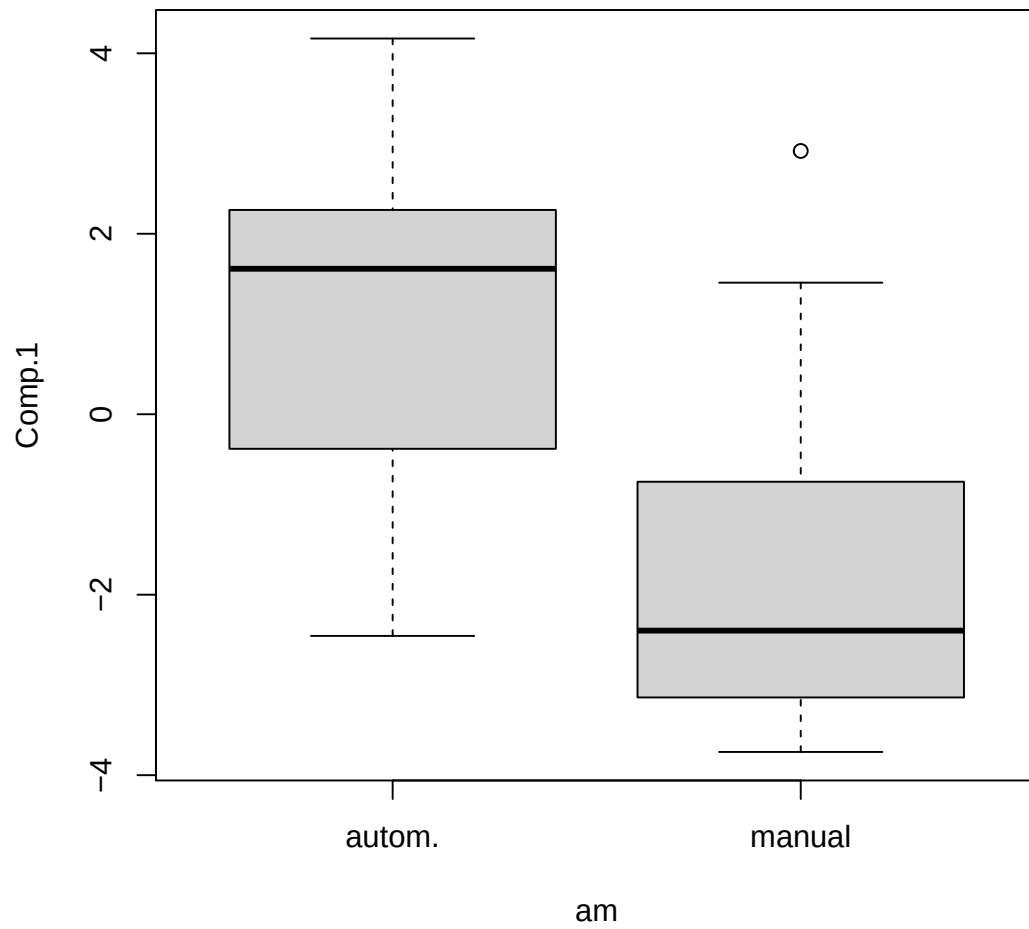
Añadimos las componentes a los datos:

```
> M1 <- cbind (M1, acp$scores[,1:2])  
> round (cor (M1), 1)
```

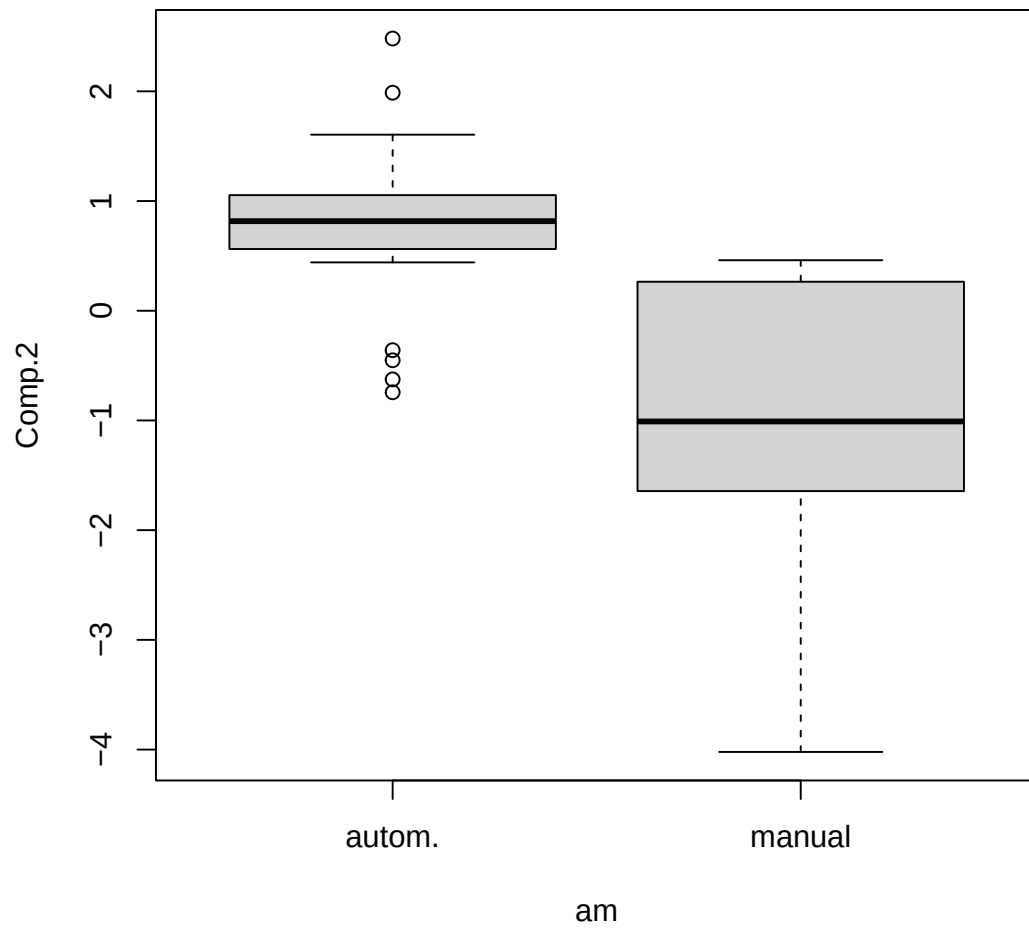
	L100k	cyl	cubi	hp	drat	peso	T100k	vs	am	gear	carb	Comp.1	Comp.2
L100k	1.0	0.8	0.9	0.8	-0.6	0.9	-0.4	-0.6	-0.5	-0.5	0.5	0.9	0.0
cyl	0.8	1.0	0.9	0.8	-0.7	0.8	-0.6	-0.8	-0.5	-0.5	0.5	1.0	0.0
cubi	0.9	0.9	1.0	0.8	-0.7	0.9	-0.4	-0.7	-0.6	-0.6	0.4	1.0	0.1
hp	0.8	0.8	0.8	1.0	-0.4	0.7	-0.7	-0.7	-0.2	-0.1	0.7	0.9	-0.4
drat	-0.6	-0.7	-0.7	-0.4	1.0	-0.7	0.1	0.4	0.7	0.7	-0.1	-0.7	-0.5
peso	0.9	0.8	0.9	0.7	-0.7	1.0	-0.2	-0.6	-0.7	-0.6	0.4	0.9	0.2
T100k	-0.4	-0.6	-0.4	-0.7	0.1	-0.2	1.0	0.7	-0.2	-0.2	-0.6	-0.5	0.7
vs	-0.6	-0.8	-0.7	-0.7	0.4	-0.6	0.7	1.0	0.2	0.2	-0.6	-0.8	0.3
am	-0.5	-0.5	-0.6	-0.2	0.7	-0.7	-0.2	0.2	1.0	0.8	0.1	-0.6	-0.6
gear	-0.5	-0.5	-0.6	-0.1	0.7	-0.6	-0.2	0.2	0.8	1.0	0.3	-0.5	-0.8
carb	0.5	0.5	0.4	0.7	-0.1	0.4	-0.6	-0.6	0.1	0.3	1.0	0.6	-0.7
Comp.1	0.9	1.0	1.0	0.9	-0.7	0.9	-0.5	-0.8	-0.6	-0.5	0.6	1.0	0.0
Comp.2	0.0	0.0	0.1	-0.4	-0.5	0.2	0.7	0.3	-0.6	-0.8	-0.7	0.0	1.0

3.4.1. ...tipo de transmisión: automático o manual

```
> M1$am <- factor (M1$am, labels = c ("autom.", "manual"))  
> boxplot (Comp.1 ~ am, M1)
```



```
> boxplot (Comp.2 ~ am, M1)
```

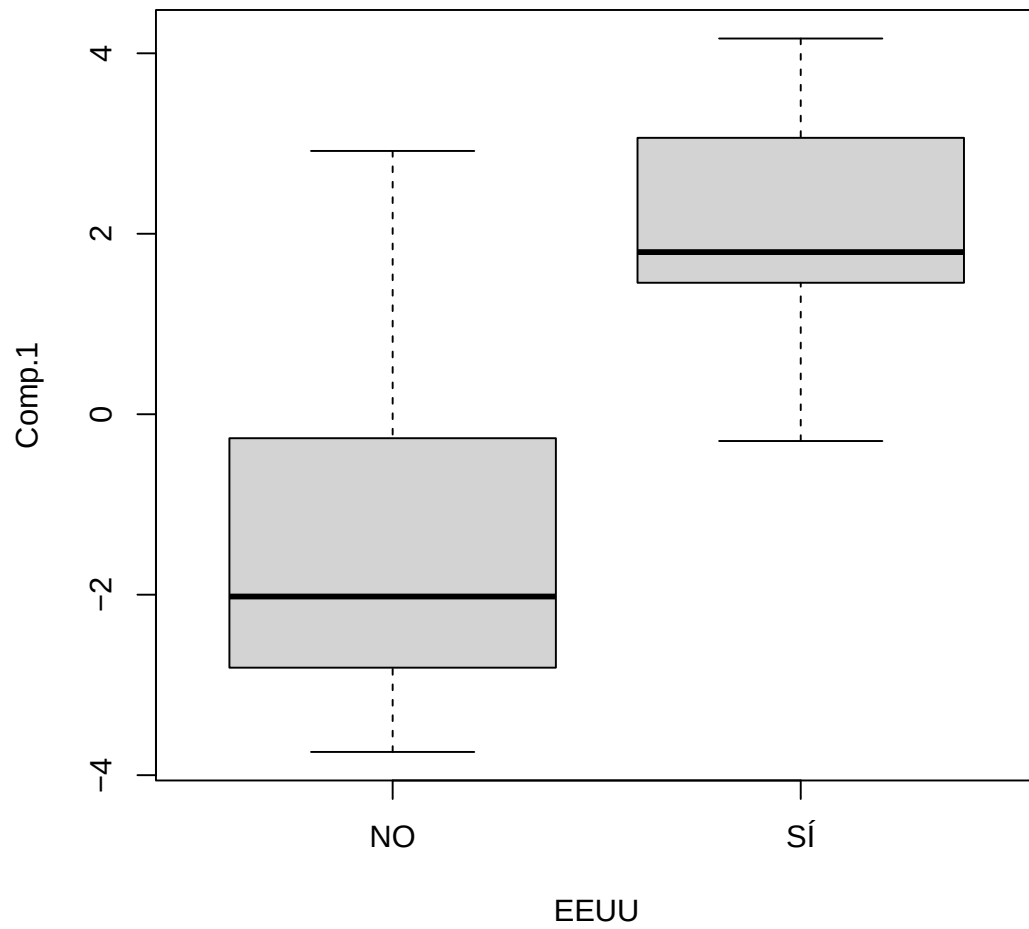


Hay una separación bastante nítida entre grupos en ambas variables:

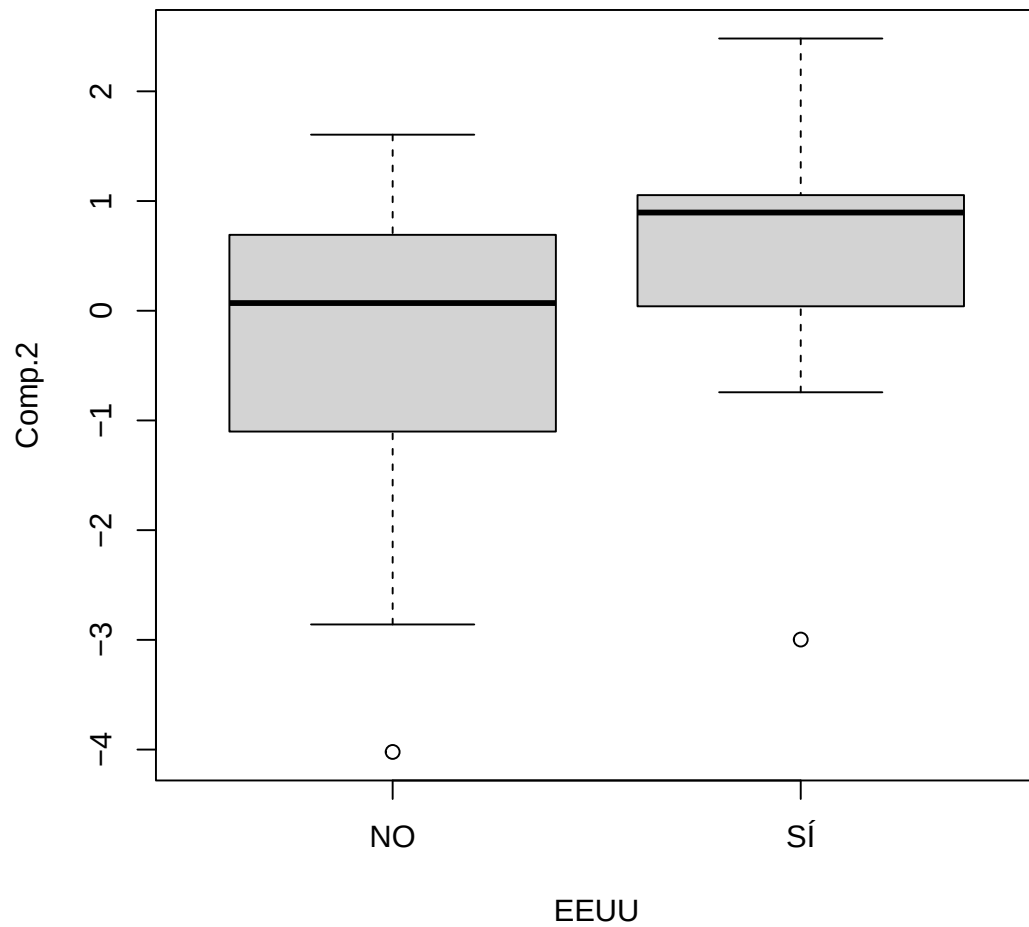
- En la primera componente, “automático” está asociado a coche “grande”.
- En la segunda, “automático” está asociado a coche “familiar”, y “manual” a “deportivo”.

3.4.2. ...lugar de fabricación: EE.UU. o extranjero

```
> M1$pais <- NA
> M1 [grep ("^Mazda", rownames(M1)), "pais"] <- "Japón"
> M1 [grep ("^Datsun", rownames(M1)), "pais"] <- "Japón"
> M1 [grep ("^Hornet", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Valiant", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Duster", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Merc", rownames(M1)), "pais"] <- "Alemania"
> M1 [grep ("^Cadillac", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Lincoln", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Chrysler", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Fiat", rownames(M1)), "pais"] <- "Italia"
> M1 [grep ("^Honda", rownames(M1)), "pais"] <- "Japón"
> M1 [grep ("^Toyota", rownames(M1)), "pais"] <- "Japón"
> M1 [grep ("^Dodge", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^AMC", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Camaro", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Pontiac", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Porsche", rownames(M1)), "pais"] <- "Alemania"
> M1 [grep ("^Lotus", rownames(M1)), "pais"] <- "R.U."
> M1 [grep ("^Ford", rownames(M1)), "pais"] <- "EE.UU."
> M1 [grep ("^Ferrari", rownames(M1)), "pais"] <- "Italia"
> M1 [grep ("^Maserati", rownames(M1)), "pais"] <- "Italia"
> M1 [grep ("^Volvo", rownames(M1)), "pais"] <- "Suecia"
> M1$EEUU <- factor (M1$pais == "EE.UU.", labels = c ("NO", "Sí"))
> boxplot (Comp.1 ~ EEUU, M1)
```

```
> boxplot (Comp.2 ~ EEUU, M1)
```



Ahora la separación entre los grupos no es nítida, pero se conserva en cierto grado:

- En la primera componente, “estadounidense” está asociado a coche “grande”.
- En la segunda, “estadounidense” está algo asociado a coche “familiar”, y los “deportivos” corresponderían a europeos y japoneses.

La coincidencia de conclusiones sugiere que la variable `am` tiene cierta relación con la variable `EEUU`. Calculemos una tabla de contingencia:

```
> table (M1 [, c ("am", "EEUU")])
```

	EEUU	
am	NO	SÍ
autom.	8	11
manual	12	1

Entre los carros de transmisión manual sólo hay uno estadounidense; entre los automáticos, algo más de la mitad. Por tanto, hay cierta relación entre ambas variables pero no es alta.