

R COMMANDER

25/03/2021

ANÁLISIS COMPLETO DE UN CONJUNTO DE DATOS

> library(faraway)

> data(pima) → Información acerca de 768 mujeres de raza indígena al sur de EE.UU.

> pima

pregnant glucose diastolic triceps insulin bmi diabetes age test

1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0...

> summary(pima)

pregnant glucose diastolic triceps

Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00

1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 62.00 1st Qu.: 0.00

Median : 3.000 Median : 117.0 Median : 72.00 Median : 23.00

Mean : 3.845 Mean : 120.9 Mean : 69.11 Mean : 20.54

3rd Qu.: 6.000 3rd Qu.: 140.2 3rd Qu.: 80.00 3rd Qu.: 32.00

Max. : 17.000 Max. : 199.0 Max. : 122.00 Max. : 99.00

insulin bmi diabetes age test

Min. : 0.0 Min. : 0.00 Min. : 0.0780 Min. : 21.00 Min. : 0.000

1st Qu.: 0.0 1st Qu.: 27.30 1st Qu.: 0.2437 1st Qu.: 24.00 1st Qu.: 0.000

Median : 30.5 Median : 32.00 Median : 0.3725 Median : 29.00 Median : 0.000

Mean : 79.8 Mean : 31.99 Mean : 0.4719 Mean : 33.24 Mean : 0.349

3rd Qu.: 127.2 3rd Qu.: 36.60 3rd Qu.: 0.6262 3rd Qu.: 41.00 3rd Qu.: 1.000

Max. : 846.0 Max. : 67.10 Max. : 2.4200 Max. : 81.00 Max. : 1.000

> sort(pima\$diastolic) → Ordenar valores

[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

[19] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24

[37] 30 30 38 40 44 44 44 44 46 46 48 48 48 48 48 50 50 50

[55] 50 50 50 50 50 50 50 50 50 50 50 52 52 52 52 52 52 52 ...

> pima\$diastolic[pima\$diastolic==0]<-NA → Quitar 0 en los valores para evitar que modifiquen los datos.

> sort(pima\$diastolic) → Se comprueba que se ha realizado correctamente.

```
[1] 24 30 30 38 40 44 44 44 44 46 46 48 48 48 48 48 50 50
```

```
[19] 50 50 50 50 50 50 50 50 50 ...
```

> pima\$glucose[pima\$glucose==0]<-NA

> pima\$triceps[pima\$triceps==0]<-NA

> pima\$insulin[pima\$insulin==0]<-NA

> pima\$bmi[pima\$bmi==0]<-NA → Solo las variables que presentan datos que carezcan de sentido.

> pima\$test<-factor(pima\$test) → Convertir variables categóricas en factores para que no interactúa en summary

> summary(pima\$test)

```
0 1
```

```
500 268
```

> summary(pima) → Comprobamos modificaciones realizadas

```
pregnant    glucose    diastolic    triceps
```

```
Min. : 0.000 Min. : 44.0 Min. : 24.00 Min. : 7.00
```

```
1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 64.00 1st Qu.: 22.00
```

```
Median : 3.000 Median :117.0 Median : 72.00 Median :29.00
```

```
Mean : 3.845 Mean :121.7 Mean : 72.41 Mean :29.15
```

```
3rd Qu.: 6.000 3rd Qu.:141.0 3rd Qu.: 80.00 3rd Qu.:36.00
```

```
Max. :17.000 Max. :199.0 Max. :122.00 Max. :99.00
```

```
NA's :5 NA's :35 NA's :227
```

```
insulin     bmi     diabetes     age     test
```

```
Min. : 14.00 Min. :18.20 Min. :0.0780 Min. :21.00 0:500
```

```
1st Qu.: 76.25 1st Qu.:27.50 1st Qu.:0.2437 1st Qu.:24.00 1:268
```

```
Median :125.00 Median :32.30 Median :0.3725 Median :29.00
```

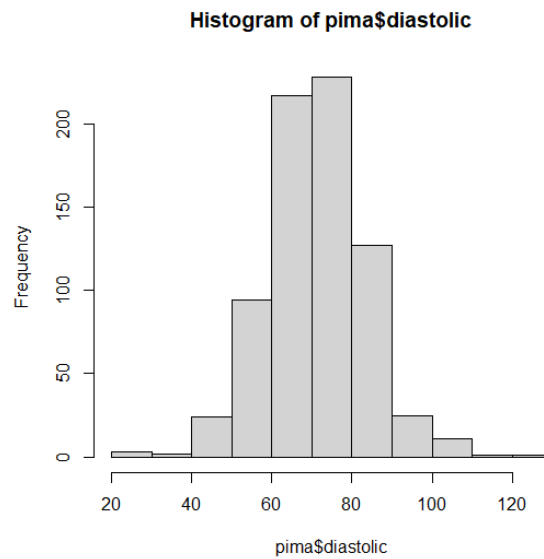
```
Mean :155.55 Mean :32.46 Mean :0.4719 Mean :33.24
```

```
3rd Qu.:190.00 3rd Qu.:36.60 3rd Qu.:0.6262 3rd Qu.:41.00
```

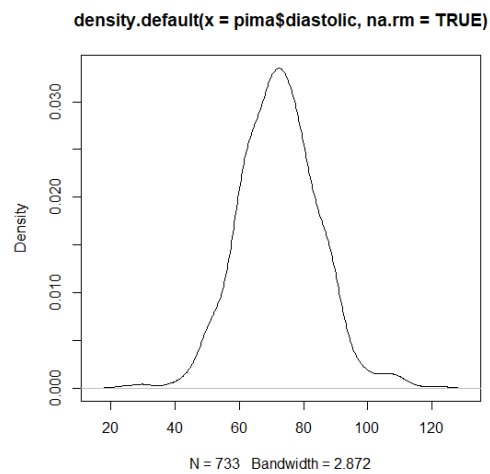
```
Max. :846.00 Max. :67.10 Max. :2.4200 Max. :81.00
```

```
NA's :374 NA's :11
```

```
> hist(pima$diastolic)
```



```
> plot(density(pima$diastolic,na.rm=TRUE)) → “na.rm” que ponga todos los datos de la distribución.
```



```
> plot(sort(pima$diastolic),pch=".") → ‘pch=“.”’ para emplear . como formas de puntos
```

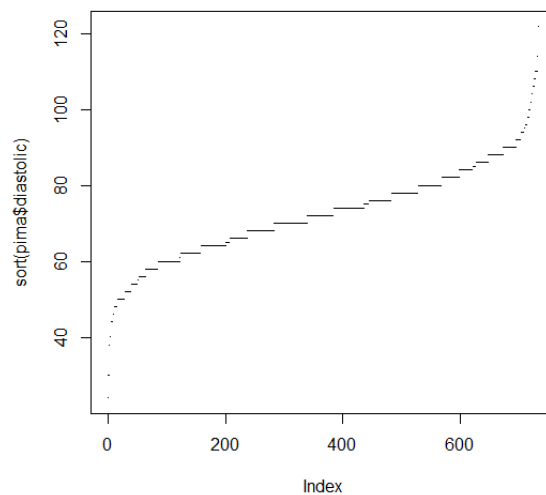
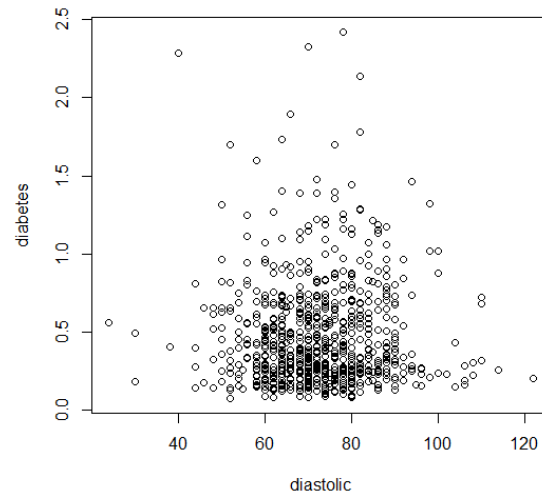
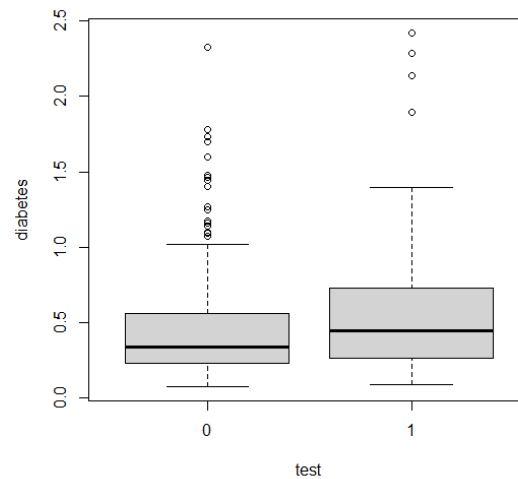


DIAGRAMA DE DENSIDAD DE KERNEL

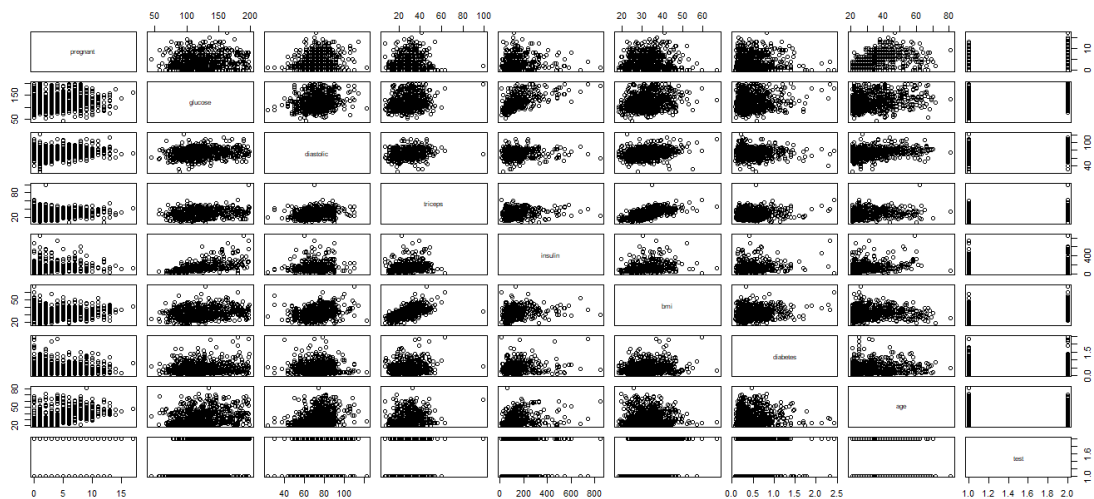
> plot(diabetes~diastolic,data=pima)



> plot(diabetes~test,data=pima)

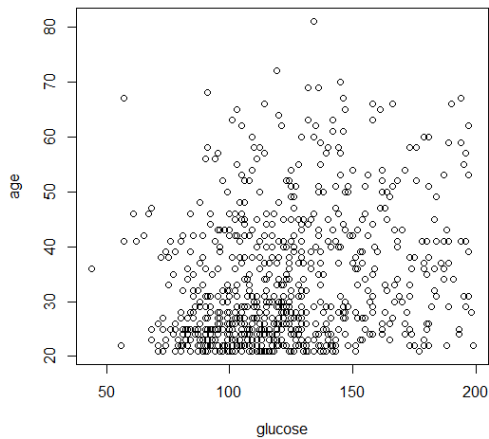


> pairs(pima) → Análisis por pares de relación. Es simétrico por lo que basta con mirar uno de los triángulos. Para evaluar que pares merecen especial atención.



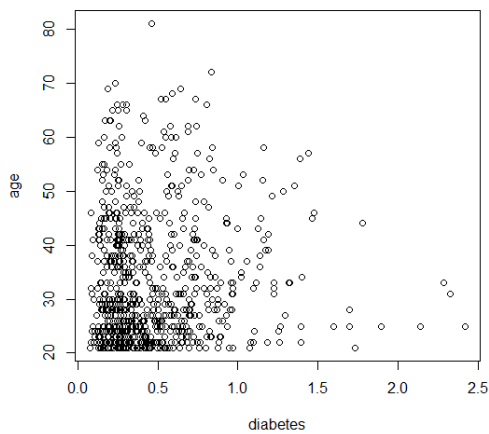
ANÁLISIS PORMENORIZADO Y VERBAL DE CADA PAR DE VARIABLES RESEÑABLES

```
> plot(age~glucose, data=pima)
```



Se concluye mayoría de la población con glucosa adecuada. Además, se intuye población joven o alta mortalidad ya que no hay mayores con glucosa elevada.

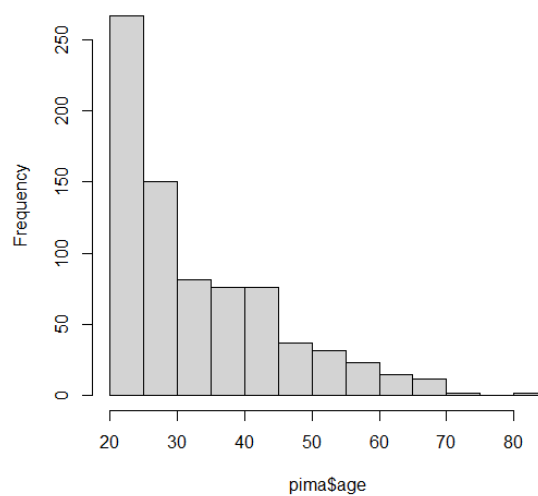
```
> plot(age~diabetes, data=pima)
```



Se entiende población sana con casos aislados de personas jóvenes con diabetes. Mismas conclusiones sobre esperanza de vida que antes

```
> hist(pima$age) → Indica muestra muy joven. Está sesgada.
```

Histogram of pima\$age



CONVERTIR y ANALIZAR TEXTOS

> toyota<-read.csv(file.choose(), header=TRUE) → Archivos Csv

> toyota

ï..Price.Age.KM.FuelType.HP.MetColor.Automatic.CC.Doors.Weight

```
1      13500;23;46986;Diesel;90;1;0;2000;3;1165
2      13750;23;72937;Diesel;90;1;0;2000;3;1165
3      13950;24;41711;Diesel;90;1;0;2000;3;1165
4      14950;26;48000;Diesel;90;0;0;2000;3;1165 ...
```

> toyota<-read.csv(file.choose(),sep="\t", header=TRUE) → Incluyendo sep ="\t" para indicar que los textos se separan por tabulaciones, sino “,” o “;”

> toyota

Price Age KM FuelType HP MetColor Automatic CC Doors Weight

```
1 13500 23 46986 Diesel 90 1 0 2000 3 1165
2 13750 23 72937 Diesel 90 1 0 2000 3 1165
3 13950 24 41711 Diesel 90 1 0 2000 3 1165 ...
```

> summary(toyota)

Price Age KM FuelType

Min. :4350 Min. :1.00 Min. : 1 Length:1436

1st Qu.: 8450 1st Qu.:44.00 1st Qu.: 43000 Class :character

Median : 9900 Median :61.00 Median : 63390 Mode :character

Mean :10731 Mean :55.95 Mean : 68533

3rd Qu.:11950 3rd Qu.:70.00 3rd Qu.: 87021

Max. :32500 Max. :80.00 Max. :243000

HP MetColor Automatic CC Doors Weight

Min. : 69.0 Min. :0.0000 Min. :0.00000 Min. :1300 Min. :2.000 Min. :1000

1st Qu.: 90.0 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:1400 1st Qu.:3.000 1st Qu.:1040

Median :110.0 Median :1.0000 Median :0.00000 Median :1600 Median :4.000 Median :1070

Mean :101.5 Mean :0.6748 Mean :0.05571 Mean :1567 Mean :4.033 Mean :1072

3rd Qu.:110.0 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:1600 3rd Qu.:5.000 3rd Qu.:1085

Max. :192.0 Max. :1.0000 Max. :1.00000 Max. :2000 Max. :5.000 Max. :1615

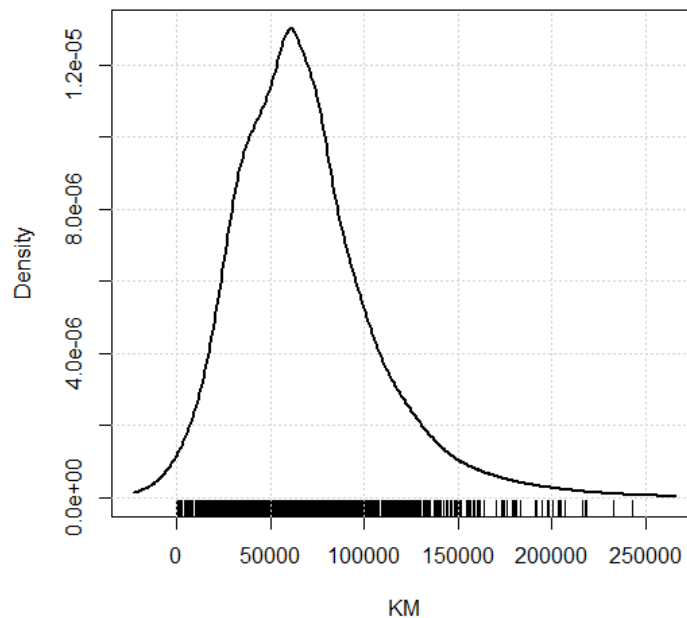
```
> library(RcmdrMisc)
```

```
>
```

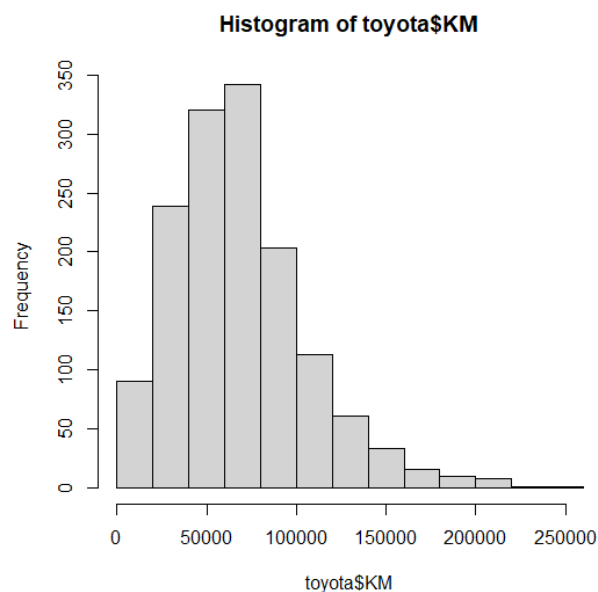
```
numSummary(toyota[,c("KM","Price","Weight"),drop=FALSE],statistics=c("mean","sd","IQR","quantiles"),quantiles=c(0,0.25,0.75,1)) → Colocando num justo delante de summary te lee únicamente las variables numéricas seleccionadas.
```

	mean	sd	IQR	0%	25%	75%	100%	n
KM	68533.26	37506.44	8877.44	0	20750	87000	243000	1436
Price	10730.82	3626.96	458	0	4350	8450	11950.00	1436
Weight	1072.46	52.64	112	45.00	1000	1040	1085.00	1436

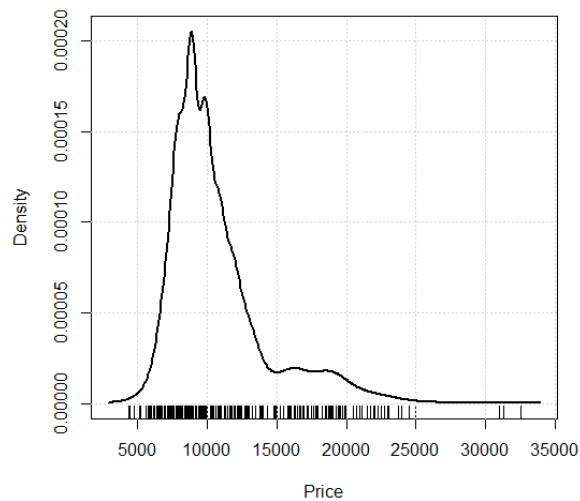
```
> densityPlot(~KM, data=toyota,bw=bw.SJ,adjust=1,kernel=dnorm, method="adaptive")
```



```
> hist(toyota$KM)
```

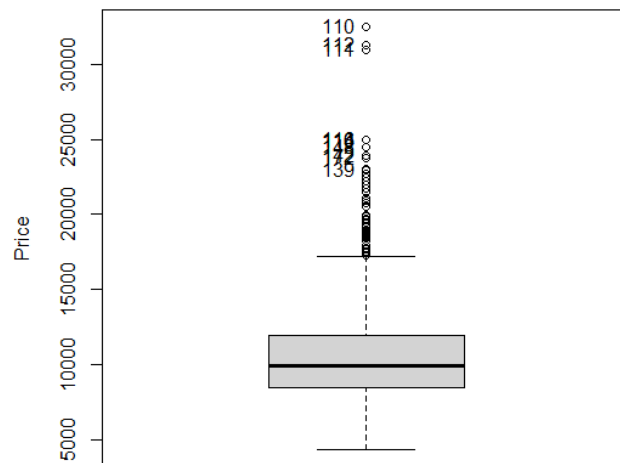


```
> densityPlot(~Price, data=toyota,bw=bw.SJ,adjust=1,kernel=dnorm, method="adaptive")
```



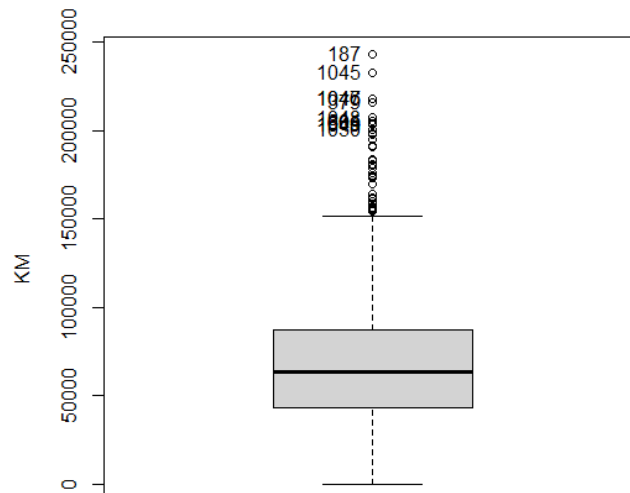
> Boxplot(~Price, data=toyota,id=list(method="y")) → 'method="y"' sacar los outlier gráficamente y en el código. No se obtiene el valor de los 'outlier' sino la posición.

```
[1] "110" "112" "111" "116" "113" "114" "148" "142" "172" "139"
```



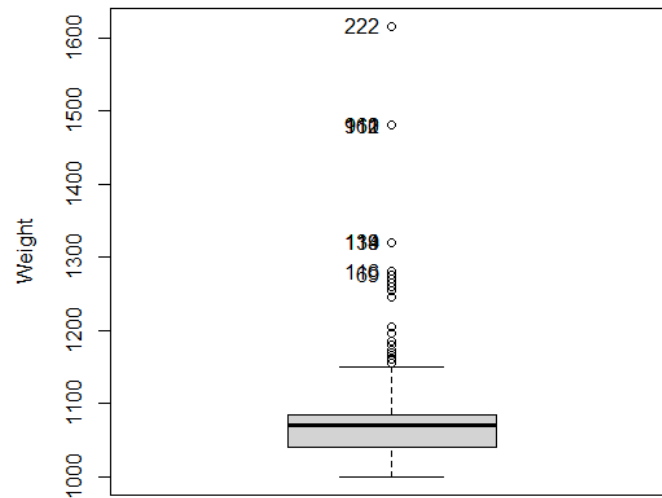
```
> Boxplot(~KM, data=toyota,id=list(method="y"))
```

```
[1] "187" "1045" "1046" "1047" "379" "1048" "604" "605" "1049" "1050"
```




```
> Boxplot(~Weight, data=toyota,id=list(method="y"))
```

```
[1] "222" "110" "111" "112" "961" "113" "114" "139" "116" "69"
```



Si se repiten en todas o la gran mayoría de estas plot sería recomendable limpiar la line y eliminar los outliers

- RECODIFICAR VARIABLES (Renombrar)

```
> library(RcmdrMisc)
```

```
> toyota<-within(toyota,{
```

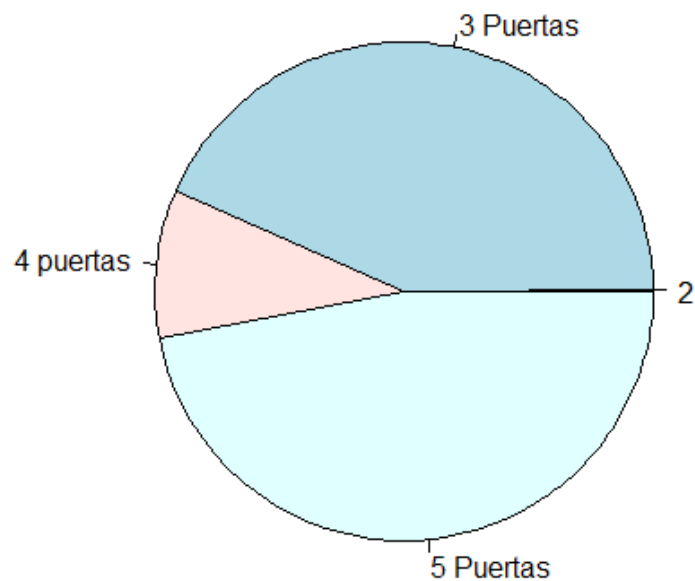
```
+ variable<-Recode(Doors,'3="3 Puertas";4="4 puertas";5="5 Puertas";;;;',as.factor=TRUE))
```

```
> table(toyota$Doors)
```

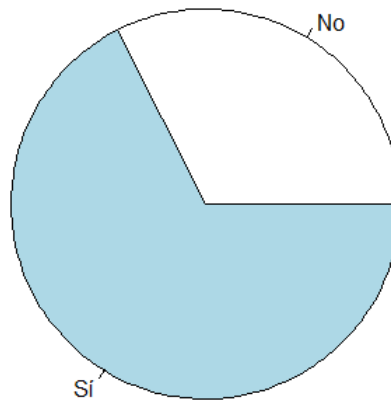
```
2 3 4 5
```

```
2 622 138 674
```

```
> pie(table(toyota$variable))
```

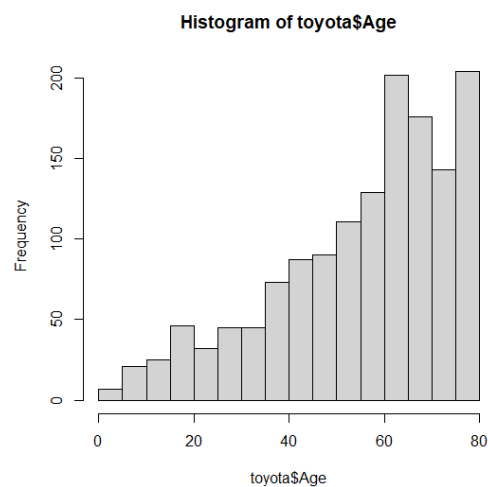


```
> toyota<-within(toyota,{
+ metalizado<-Recode(MetColor,'0="No";1="Si";;;;;;;',as.factor=TRUE))
> pie(table(toyota$metalizado))
```

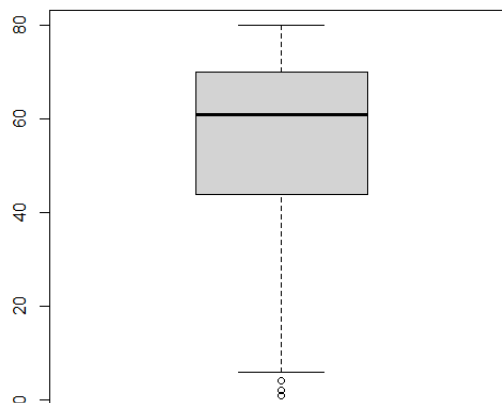


- Respecto de los ; son sitios cambiados (antes 3) si el número es menor habrá que aumentar el número de ; para que se machaquen.
- Tantos ; como el mayor salto que halla de opciones.

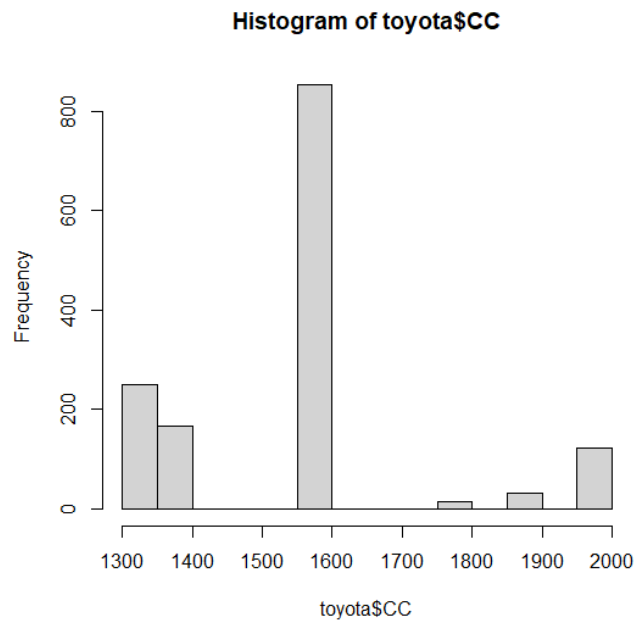
```
> hist(toyota$Age)
```



```
> boxplot(toyota$Age)
```



```
> hist(toyota$CC)
```



```
> toyota<-within(toyota,{
```

```
+ automatico<-Recode(Automatic,'0="No";1="Sí";;;;;;;',as.factor=TRUE))} → Se utilizará más adelante
```

- ASOCIAR VARIABLES LOCALES

```
> local({
```

```
+ .Table<-xtabs(~automatico+metalizado,data=toyota)
```

```
+ cat("\nFrequency table:\n")
```

```
+ #Para ordenar datos
```

```
+ print(.Table)
```

```
+ cat("\nColumn percentages:\n")
```

```
+ print(colPercents(.Table))
```

```
+ .Test<-chisq.test(.Table, correct=FALSE)
```

```
+ print(.Test)
```

```
+ })
```

Frequency table:

metalizado

automatico No Sí

No 438 918

Sí 29 51

Column percentages:

	metalizado	
automatico	No	Sí
No	93.8	94.7
Sí	6.2	5.3
Total	100.0	100.0
Count	467.0	969.0

Pearson's Chi-squared test>

#Contraste de Pearson

data: .Table

X-squared = 0.53686, df = 1, p-value = 0.4637 → Indica que no es significativo (Debería ser muy pequeño)

```
> local({  
+ .Table<-xtabs(~metalizado+variable,data=toyota)  
+ cat("\nFrequency table:\n")  
+ print(.Table)  
+ cat("\nColumn percentages:\n")  
+ print(colPercents(.Table))  
+ .Test<-chisq.test(.Table, correct=FALSE)  
+ print(.Test)  
+ cat("\nChi-square components:\n") → División por componentes  
+ print(round(.Test$residuals^2,2))  
+ })
```

Frequency table:

	variable			
metalizado	2	3 Puertas	4 puertas	5 Puertas
No	1	232	39	195
Sí	1	390	99	479

Column percentages:

		variable			
metalizado		2	3 Puertas	4 puertas	5 Puertas
No	50	37.3	28.3	28.9	
Sí	50	62.7	71.7	71.1	
Total	100	100.0	100.0	100.0	
Count	2	622.0	138.0	674.0	

Pearson's Chi-squared test

data: .Table

X-squared = 11.847, df = 3, p-value = 0.007925

nChi-square components:

		variable			
metalizado		2	3 Puertas	4 puertas	5 Puertas
No	0.19	4.37	0.77	2.67	
Sí	0.09	2.10	0.37	1.29	

Warning message:

In chisq.test(.Table, correct = FALSE) :

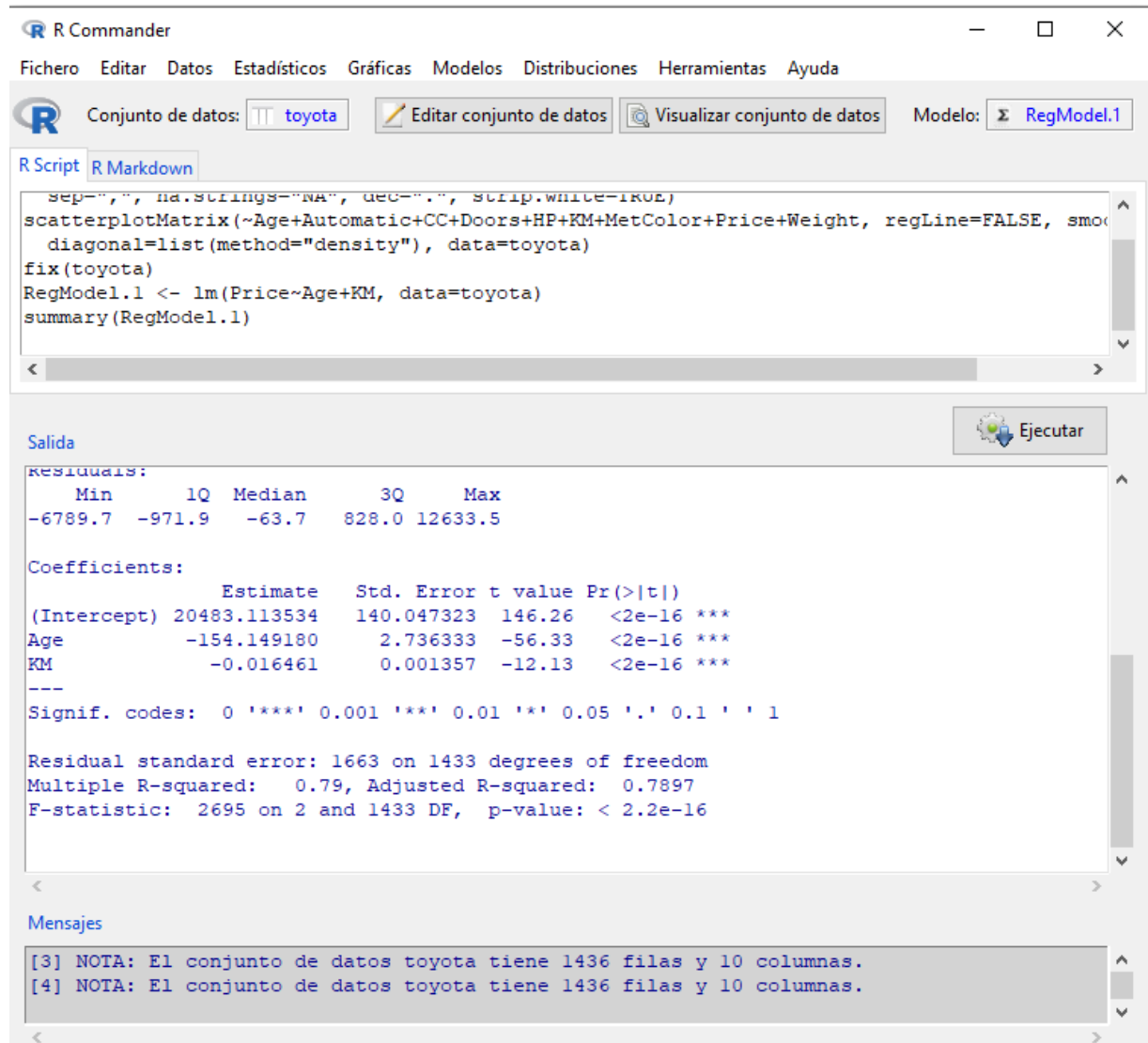
Chi-squared approximation may be incorrect

> pairs(toyota)

Error in pairs.default(toyota) : non-numeric argument to 'pairs'

No funciona → Desde Rcmdr

Datos → Importar datos → Desde archivo de texto/portapapeles → Seleccionar comas como elementos de separación y seleccionar archivo .txt → Matriz de diagramas de dispersión → Seleccionar las variables que se quieren comparar y aceptar.



The screenshot shows the R Commander window with the following components:

- Menu Bar:** Fichero, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, Ayuda.
- Toolbar:** Conjunto de datos: ; Editar conjunto de datos; Visualizar conjunto de datos; Modelo:
- R Script Panel:**

```
sep="," , na.strings="NA", dec=".", strip.white=TRUE,
scatterplotMatrix(~Age+Automatic+CC+Doors+HP+KM+MetColor+Price+Weight, regLine=FALSE, smoo
diagonal=list(method="density"), data=toyota)
fix(toyota)
RegModel.1 <- lm(Price~Age+KM, data=toyota)
summary(RegModel.1)
```
- Salida Panel:**

Residuals:

Min	1Q	Median	3Q	Max
-6789.7	-971.9	-63.7	828.0	12633.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20483.113534	140.047323	146.26	<2e-16 ***
Age	-154.149180	2.736333	-56.33	<2e-16 ***
KM	-0.016461	0.001357	-12.13	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1663 on 1433 degrees of freedom
Multiple R-squared: 0.79, Adjusted R-squared: 0.7897
F-statistic: 2695 on 2 and 1433 DF, p-value: < 2.2e-16
- Mensajes Panel:**

[3] NOTA: El conjunto de datos toyota tiene 1436 filas y 10 columnas.
[4] NOTA: El conjunto de datos toyota tiene 1436 filas y 10 columnas.

