CURSO: Análisis de Datos Reproducible con R

Introducción al Análisis Exploratorio de Datos

DSLAB

2017-2018

Índice

Introducción

El **análisis exploratorio de datos** (*exploratory data analysis* o EDA) es un conjunto de técnicas que permiten resumir las características más importantes de un conjunto de datos, normalmente con especial énfasis en el uso de métodos de visualización gráfica.

El término fue popularizado, entre otros, por el estadístico norteamericano John W. Tukey (1915-2000), a través de su libro *Exploratory Data Analysis* (1977) como método para descubrir informacion importante o valiosa contenida en los datos que no se pueda anticipar o no sea evidente. Esta aproximación contrasta con el **análisis confirmatorio de datos** (*confirmatory data analysis* o CDA) donde normalmente buscamos comprobar la validez de una hipótesis mediante la aplicación de técnicas estadísticas sobre los datos (*inferencia estadística*).

Aunque tradicionalmente la Estadística ha volcado su interés en el análisis confirmatorio de datos, las nuevas técnicas computacionales de tratamiento y visualización de datos han permitido grandes avances en el EDA. En particular, R nos proporciona una gran cantidad de herramientas para efectuar el EDA con diferentes tipos de datos.

Variables cualitativas y cuantitativas

En estadística, existe un debate intenso pero poco conocido sobre cómo debemos clasificar los diferentes tipos de variables y escalas de medida que podemos usar para caracterizar la realidad que vamos a analizar [Hand, 1996]. Por ejemplo, Stevens (1946, 1951) propuso una clasificación de escalas de medida desde el campo de las ciencias sociales y del comportamiento. Sin embargo, otros autores posteriores criticaron esta clasificación y propusieron otras alternativas, por ejemplo Mosteller y Tukey (1977) o Chrisman (1997).

Puesto que no existe un consenso bien definido sobre este tema, a continuación se presenta una clasificación

pragmática de tipos de variables, basada en la práctica de métodos y técnicas estadísticas comunmente aplicados. Su filosofía sigue, en cierto modo, la orientación propuesta por Mosteller y Tukey.

Variable cualitativa (también llamada **categórica** o **atributo**) es aquella que refleja una cualidad de la realidad o elemento observado y, por tanto, su valor no se representa mediante un número. Por ejemplo, la respuesta a una pregunta (`Sí` o `No`), el estado civil (`soltero`, `casado`, `divorciado`, `viudo`, etc.) o el tipo sanguíeno.

A su vez, las variables cualitativas pueden ser **dicótomas** o **polítomas**. Una variable cualitativa es **dicótoma** cuando solo puede tomar dos posibles valores, por ejemplo (`Sí` o `No`). Una variable cualitativa es **polítoma** cuando puede tomar valores en más de dos posibles categorías, por ejemplo el grupo sanguíneo o el estado civil.

Variable cuantitativa es aquella cuyo valor se indica con un número, puesto que corresponde a características que se representan mediante cantidades. Ejemplos serían la temperatura en ºC, el peso en kg., el tiempo en seg. o un recuento de elementos.

Las variables cuantitativas pueden, a su vez, ser **discretas** o **continuas**. Una variable cuantitativa **discreta** es aquella que solo puede tomar un numero finito o infinito numerable de valores (numéricos). El segundo ocurre cuando en un intervalo finito la variable solo puede tomar un número finito (limitado) de valores. Un ejemplo serían los números naturales, puesto que para un intervalo finito como [1,4] solo puede tomar los valores {1, 2, 3, 4} (y ningún valor entre estos). Por tanto, una variable que contenga un *recuento* siempre es un caso de variable cuantitativa discreta (su valor es 0 o un número entero positivo).

Una variable cuantitativa **continua** es la que puede tomar infinitos valores numéricos en un intervalo finito, por ejemplo los números reales en el intervalo [1,4]. Otros ejemplos serían magnitudes físicas comom peso, altura, tiempo, etc. En la práctica, al final siempre se impone un límite de precisión al medir una variable cuantitativa continua, porque no podemos medir una magnitud con infinitos decimales.

Además, existen algunas otras consideraciones que conviene tener en cuenta para caracterizar nuestras variables :

- Una variable **acotada** es aquella que solo toma valores dentro de un intervalo y no está definidas (matemáticamente) fuera del mismo. Un ejemplo es un recuento, que tiene un límite inferior de 0 (o hay algún elemento o no hay ninguno) o un porcentaje (acotado entre 0 y 100).
- Una variable *cuantitativa* ha sido **dicotomizada** cuando su escala cuantitativa se ha dividido en solo dos clases separadas por un valor de corte. Un ejemplo serían los valores de corte que se escogen en medicina para la presión sanguínea o el colesterol con el fin de diagnosticar hipertensión o hipercolesterolemia (el resultado es `Sí` o `No` dependiendo de si el valor medido cae por encima o por debajo del valor de corte entre los dos grupos). Una variable *cuantitativa* ha sido **politomizada** si se eligen varios valores de corte para separar su escala cuantitativa original en más de dos grupos.

En R podemos dividir una variable cuantitativa en intervalos mediante la función `cut`, o también mediante la función `Hmisc::cut2`, más potente y flexible que la primera.

```r
library(Hmisc)
disp_cut <- cut2(mtcars$disp, g = 3)
table(disp_cut)
```

```
## disp_cut
## [ 71.1,147) [146.7,304) [304.0,472]
##          11          11          10
```

Esta técnica (también conocida en inglés como *binning* o *bucketing*) se utiliza por ejemplo al representar el histograma, puesto que se divide la variable en intervalos para su representación gráfica.

Debemos tener precaución cuando usemos variables dicotomizadas o politomizadas, puesto que la elección de los puntos de corte puede desvirtuar los resultados de nuestro análisis. Como norma general, *siempre que sea posible* deberíamos utilizar *variables cuantitativas en su escala original* en nuestro análisis y evitar convertirlas en categóricas.

Escalas de medida

Las variables que utilizamos en técnicas y métodos de ciencia de datos se suelen clasificar mediante una taxonomía de **niveles o escalas de medida**, complementaria a la anterior, propuesta por Stanley S. Stevens (1946, 1951), según la escala y propiedades de los valores que usamos para medir:

- Escala **categórica**: Es aquella cuyos valores representan diferentes categorías, a cada una de las cuales se le asigna un identificador unívoco. Ejemplos de este tipo serían una variable que representa la respuesta a una pregunta (valores `Sí` o `No`) o una variable que representa diferentes colores (`Rojo`, `Verde`, `Azul`). Dentro de este tipo podemos distinguir entre variables *binarias* y *nominales*.
  - Escala *binarias*: Son variables categóricas que únicamente pueden tomar dos posibles valores. Una variable que codifica la respuesta a una pregunta como `Sí` o `No` es un ejemplo de esta clase. Otro ejemplo sería una variable que toma los valores (`0`, `1`), interpretados como dos niveles diferentes.
  - Escala *nominales*: Son variables categóricas que pueden tomar más de dos posibles valores. La variable que codifica diferentes colores (`Rojo`, `Verde`, `Azul`) es un ejemplo de esta clase.
- Escala **ordinal**: Es similar a la escala categórica, pero en este caso las categorías se pueden organizar en un ranking (o lista ordenada) porque existe un *orden lógico entre sus valores*. Un caso común sería los resultados de un examen (`Sobresaliente`, `Notable`, `Aprobado`, `Suspenso`), que se pueden organizar en un ranking desde el mejor hasta el peor (o al contrario, empezando por el `Suspenso` hasta el mejor nivel).
- Escala de **intervalo**: Aquella en la que intervalos de la misma longitud representan la misma diferencia en la propiedad que mide la variable. Un buen ejemplo sería la escala Celsius para medir la temperatura,

que se construye dividiendo el intervalo entre dos puntos de referencia (la temperatura de congelación y la de ebullición del agua en determinadas condiciones) en 100 intervalos de la misma longitud. De este modo, sabemos que la diferencia de temperatura al pasar de 15ºC a 20ºC es la misma que al pasar de 10ºC a 15ºC. Sin embargo (en contra de lo que la sabiduría popular pueda afirmar), no es correcto afirmar cuando estamos a 30ºC que "hace el doble de calor" que cuando estamos a 15ºC. Para convencernos, bastan pasar estos valores a escala Farenheit (otra escala de intervalo), y comprobar que 86ºF (30ºC) no es el doble de 59ºF (15ºC), aunque representan la misma diferencia de temperaturas.
En resumen, solo podemos medir *diferencias* entre dos valores de la escala de un modo consistente, pero los valores que toma la variable no dan información per sé, de forma absoluta. Otro ejemplo clásico de este tipo es una variable expresada en forma de **porcentaje**. Por ejemplo, dos países pueden tener una deuda equivalente al 4% de su PIB, lo que indica que tienen el mismo nivel de deuda en términos relativos (respecto a su PIB total, dividido en 100 partes iguales). Sin embargo, el 4% del PIB de EE.UU. puede representar una cantidad de dinero (en términos absolutos) mucho mayor que el 4% del PIB de Zimbabwe (para obtener ese dato y comparar en términos absolutos, habría que multiplicar por el PIB total de cada país).

- Escala de **proporción o razón** (*ratio*): Tienen las mismas propiedades que la escala de intervalo (se pueden definir) intervalos entre dos valores y compararlos de forma consistente), pero además definen un *cero absoluto* como valor de referencia, que representa la ausencia total del fenómeno o característica que estamos midiendo. Por tanto, en este caso sí se pueden hacer ratios con sentido entre dos medidas de la variable. Un ejemplo sería la escala Kelvin para la temperatura, en la que se define el 0ºK (*cero absoluto*) como la ausencia total de actividad térmica en las partículas de una materia. Por el contrario , en la escala Celsis el 0ºC no indica ausencia de temperatura, sino simplemente el punto que arbitrariamente se escogió como referencia (punto de congelación del agua). Otros ejemplos de varibles racionales son cualquiera de las magnitudes físicas medidas según el Sistema Internacional de Unidades.

Datos de sección cruzada, series temporales y paneles de datos

Otro criterio importante para caracterizar el tipo de datos que estamos utilizando es comprobar si existe algún tipo de **dependencia temporal** entre los valores recopilados para una variable o no. Según esto tenemos:

- **Datos de sección cruzada**: Aquellos que se obtienen al medir la misma variable para distintos casos en el mismo instante temporal. Un ejemplo serían los resultados de una encuesta (con variables cuantitativas y cualitativas) que han respondido 200 personas en un día concreto.
- **Series temporales**: Datos obtenidos para una variable ordenados según transcurren en el tiempo. Por ejemplo, la serie de valores de cotización en bolsa de las acciones de una compañía tomados cada día durante un mes, o el número total de ejemplares que una revista ha vendido cada mes entre el año 2000 y el 2014.
- **Paneles de datos**: Combinación de los dos tipos anteriores. Tenemos varias series temporales con valores de una variable, y cada serie corresponde a un valor distinto de otra variable. Por ejemplo, cuando

contamos con una serie de valores diarios de cotización en bolsa durante un mes de cada una de las empresas del Ibex 35.

La **abstracción de datos** (clase) que utilizamos para analizar datos de sección cruzada tanto en Python (con la biblioteca Pandas) como en R (por defecto) y otras plataformas computacionales como Spark es el **data frame**. Las series temporales suelen contar con su propia abstracción de datos en cada lenguaje. En Pandas tenemos la clase Series, en R hay varias clases para escoger (como `ts`, `POSIXlt`, `POSIXct`) o bibliotecas como `zoo` o `xts` (véase para más información la Task View para series temporales en CRAN.

Tipos de variables: conclusiones

Podemos encontrar en fuentes como Wikipedia intentos de crear una clasificación unificada de tipos de variables, atendiendo a varios de los criterios que hemos mencionado.

Desde un punto de vista pragmático para su aplicación en estudios, el científico de datos como mínimo debe saber *identificar variables cuantitativas y cualitativas o categóricas*, puesto que las técnicas descriptivas y de análisis estadístico que podemos aplicar a cada tipo son muy diferentes. En general, las variables categóricas restringen el repertorio de herramientas que podemos utilizar para su análisis.

Finalmente, conviene recordar que en muchos casos las variables categóricas se representan mediante valores numéricos para realizar los cálculos en un análisis. Eso no significa que se hayan convertido en variables cuantitativas, sino que simplemente es un modo conveniente de operar con esos valores matemáticamente (por ejemplo, `Sí` = 1 y `No` = 0). Sin embargo, hay que prestar atención cuando la escala de medida es **ordinal** (ranking) y no categórica, puesto que deberíamos preservar el orden natural de los valores de la variable al realizar y presentar el análisis. Por ejemplo, en un estudio demográfico no implica gran diferencia presentar primero los resultados para el grupo `hombres` y luego para `mujeres` o viceversa (la escala es binaria y no ordinal). Sin embargo, en un estudio medioambiental con posibles niveles de concentración de partículas {`Bajo`, `Medio`, `Alto`, `Extremo`} no sería muy lógico presentar comparativas y resultados de estos cuatro grupos en cualquier orden. Debemos respetar la ordenación que existe de forma natural en estos valores (de menor a mayor o de mayor a menor, pero siempre en el orden original).

Resumen numérico de datos

Como se ha indicado antes, las técnicas que podemos aplicar para descripción y análisis de datos son diferentes dependiendo del tipo de variable con el que estemos trabajando. A continuación presentamos la técnicas más frecuentes para resumir de forma numérica los valores de nuestras variables cuantitativas y cualitativas.

Paquetes y datasets

Antes de comenzar, el siguiente *chunk* reúne los diferentes paquetes y datasets utilizados en los ejemplos de las siguientes secciones, para facilitar la labor de comprobación de requisitos previos antes de ejecutar los ejemplos.

```
library(moments); library(gmodels); library(scales)
library(survival); library(zoo); library(quantmod); library(vcd); library(vcdExtra)
library(faraway); library(Hmisc); library(car); library(MASS); library(latticeExtra)
library(dplyr); library(tidyr); library(ggplot2)
library(gcookbook)
```

```
?CO2             # Tolerancia al frío de especies de cesped Echinochloa crus-galli
?UCBAdmissions   # Datos agregados solicitantes Univ. Berkeley 1973
?DaytonSurvey    # Datos de Agresti (2002); encuesta en Dayton (Ohio) sobre uso de
                 # drogas en alumnos senior de insitituto
?babyfood        # Estudio sobre enfermedades respiratorias en niños
?pima            # Estudio del NIDDKD sobre 768 mujeres indias Pima adultas
?Salaries        # Salario (9 meses) de profesores en EE.UU. en 2008-2009
?gala            # Estudio de diversidad de especies en las Islas Galápagos
?Ornstein        # Conexiones entre consejos directivos de grandes compañías en Canadá
?diamonds        # Precios y otros atributos de 54.000 diamantes
?austres         # Número trimestral de residentes en Australia (Marzo 1971 - Marzo 199
?mgus            # Datos estudio de gamopatía monoclonal
?Prestige        # Encuesta sobre prestigio de ocupaciones en Canadá
?Arthritis       # Datos sobre estudio de un nuevo tratamiento para artritis reumatoide
                 # (Kock y Edwards, 1988)
?countries       # Datos económicos y de salud del Banco Mundial sobre países (1960-201
"S&P 500"        # Datos del S&P 500 desde FRED (St. Louis Fed)
"ppg2008.csv"    # Estadísticas jugadores NBA (2008)
                 # http://datasets.flowingdata.com/ppg2008.csv
```

Resumen numérico de variables cualitativas

La forma más sencilla de resumir las variables cualitativas es mediante una tabla de distribución de frecuencias, llamada **tabla de contingencias**. La tabla muestra, para cada valor que tome una variable categórica, o para cada combinación de valores de dos o más variables categóricas, el número de casos que aparecen con dicho valor o combinación de valores.

```
# Ejemplo de creación de tabla con dataset CO2
# Para más información, ?CO2
```

```r
# La función `table` genera tablas de contigencias
# Aquí comprobamos que el experimento fue diseñado para que exista el mismo
# número de casos en cada celda o cruce de valores
with(CO2, table(Type, Treatment))
```

```
##              Treatment
## Type          nonchilled chilled
##   Quebec              21      21
##   Mississippi         21      21
```

```r
# Algunos datasets ya vienen en formato de tabla como `UCBAdmissions` o `Titanic`
# Cuando tenemos 3 o más categorías es más útil usar `ftable` que imprime la
# tabla de contigencias en formato compacto
ftable(UCBAdmissions)
```

```
##                  Dept   A   B   C   D   E   F
## Admit    Gender
## Admitted Male        512 353 120 138  53  22
##          Female       89  17 202 131  94  24
## Rejected Male        313 207 205 279 138 351
##          Female       19   8 391 244 299 317
```

Además, se pueden obtener datos adiconales como las frecuencias marginales (para cada variable, que corresponde a una dimensión de la tabla) con `margin.table` o las proporciones para cada categoría con `prop.table`:

```r
# El argumento `margin` controla la variable o dimensión que usamos para agregar
# Recuento por Admitidos/Excluidos
margin.table(UCBAdmissions, margin = 1)
```

```
## Admit
## Admitted Rejected
##     1755     2771
```

```r
# Recuento por Género
margin.table(UCBAdmissions, margin = 2)
```

```
## Gender
##   Male Female
##   2691   1835
```

```r
# Recuento por Departamento
margin.table(UCBAdmissions, margin = 3)
```

```
## Dept
```

```
##   A   B   C   D   E   F
## 933 585 918 792 584 714
```

La forma de utilizar `prop.table` es análoga, el argumento `margin` controla sobre que variable (margen) calculamos los porcentajes de cada cruce. Usamos `ftable` para imprimir la tabla en forma compacta.

```
# Porcentajes sobre el total de Admitidos/Excluidos
ftable(prop.table(UCBAdmissions, margin = 1))
```

```
##             Dept         A         B         C         D         E         F
## Admit    Gender
## Admitted Male      0.291737892 0.201139601 0.068376068 0.078632479 0.030199430 0.0125356
##          Female    0.050712251 0.009686610 0.115099715 0.074643875 0.053561254 0.013675214
## Rejected Male      0.112955612 0.074702274 0.073980512 0.100685673 0.049801516 0.126669O
##          Female    0.006856730 0.002887044 0.141104294 0.088054854 0.107903284 0.114399134
```

```
# Porcentajes sobre el total por Género
ftable(prop.table(UCBAdmissions, margin = 2))
```

```
##             Dept         A         B         C         D         E         F
## Admit    Gender
## Admitted Male      0.190263842 0.131178001 0.044593088 0.051282051 0.019695281 0.0081753
##          Female    0.048501362 0.009264305 0.110081744 0.071389646 0.051226158 0.013079019
## Rejected Male      0.116313638 0.076923077 0.076179859 0.103678930 0.051282051 0.1304347
##          Female    0.010354223 0.004359673 0.213079019 0.132970027 0.162942779 0.172752044
```

```
# Porcentajes sobre el total por Departamento
ftable(prop.table(UCBAdmissions, margin = 3))
```

```
##              Dept         A         B         C         D         E         F
## Admit    Gender
## Admitted Male      0.54876742 0.60341880 0.13071895 0.17424242 0.09075342 0.03081232
##          Female    0.09539121 0.02905983 0.22004357 0.16540404 0.16095890 0.03361345
## Rejected Male      0.33547696 0.35384615 0.22331155 0.35227273 0.23630137 0.49159664
##          Female    0.02036442 0.01367521 0.42592593 0.30808081 0.51198630 0.44397759
```

También podemos utilizar la función `xtabs` si preferimos utilizar fórmulas como argumento para expresar la organización de la tabla que deseamos:

```
# El resultado es similar al que vimos antes, comprobamos
# diseño balanceado con mismo número de muestras por celda
# (todas tienen 21 casos)
ftable(xtabs(~ Treatment + Type, data = CO2))
```

```
##            Type Quebec Mississippi
```

```
## Treatment
## nonchilled          21          21
## chilled             21          21
```

Existen métodos más sofisticados para agregar datos por variables o aplicar diversas funciones resumen en la agregación. La sección 2.5 de la viñeta Working with categorical data with R and the `vcd` and `vcdExtra` packages muestra un ejemplo más complejo para análisis categórico con diferentes variables categóricas (binarias y nominales).

Por último, la función `gmodels::CrossTable` (del paquete `gmodels`) muestra otra forma compacta de presentar la tabla de contingencias junto con las proporciones marginales:

```
# Instalar las bibliotecas `gmodels` y `vcdExtra` en caso necesario
library(gmodels)
library(vcdExtra)
with(DaytonSurvey, CrossTable(sex, race, format = 'SPSS'))
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  32
##
##              | race
##          sex |    white  |    other  | Row Total |
## -------------|-----------|-----------|-----------|
##       female |        8  |        8  |       16  |
##              |    0.000  |    0.000  |           |
##              |   50.000% |   50.000% |   50.000% |
##              |   50.000% |   50.000% |           |
##              |   25.000% |   25.000% |           |
## -------------|-----------|-----------|-----------|
##         male |        8  |        8  |       16  |
##              |    0.000  |    0.000  |           |
##              |   50.000% |   50.000% |   50.000% |
##              |   50.000% |   50.000% |           |
##              |   25.000% |   25.000% |           |
```

```
## -------------|-----------|-----------|-----------|
## Column Total |        16 |        16 |        32 |
##              |   50.000% |   50.000% |           |
## -------------|-----------|-----------|-----------|
##
##
```

Resumen numérico de variables cuantitativas

Medidas de posición (centralidad)

La función básica para obtener el resumen numérico de una variable cuantitativa en R es `summary`:

```
# Velocidad y distancia de detención de coches, data.frame `cars` (?cars)
# disponible por defecto en R (paquete `datasets`)
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

`summary` nos proporciona los siguientes valores resumen:

- Los valores **mínimo** y **máximo** que toma la variable.
- La **media** y la **mediana** de la serie de valores de variable.
- El **primer y tercel cuartil** de la serie de valores de la variable.

Respecto a la media, R nos muestra la media artimética utilizando todos los valores de la variable, sobre los que se aplica la función `mean`. Esta función (usada por separado) también admite argumentos adicionales para otros cálculos alternativos de la media. Una variante interesante es la **media recortada** (*trimmed mean*), que elimina simétricamente la fracción que indiquemos de las observaciones extremas por ambos lados, antes de calcular la media. Existen otras muchas variantes para el cálculo de la media (media geométrica, armónica, ponderada, windsorizada, etc.). Todas ellas pueden calcularse en R, aunque a veces a través de un paquete de CRAN adicional.

La media ofrece un resumen razonable de los valores de una variable cuando su distribución es simétrica y no existen muchos valores extremos. De lo contrario, su valor se altera rápidamente en presencia de casos atípicos. En esa situación, es preferible usar la mediana o una media recortada (eliminando suficientes atípicos de los extremos) que refleje mejor la zona central de la distribución.

Cuando aplicamos `summary` a un data frame que mezcla columnas con variables cuantitativas y cualitativas, nos presenta la versión adecuada para cada tipo de variable (recuento de casos para variables categóricas o resumen numérico para variables cuantitativas).

```
# Biblioteca que acompaña al libro "Linear Models with R" (2ª Ed.) de J. Faraway
# Instalar en caso necesario antes de ejecutar
# Dataset `babyfood`: desarrollo de enfermedades respiratorias en niños (?babyfood)
library(faraway)
summary(babyfood)
```

```
##     disease        nondisease      sex        food
##  Min.   :16.00   Min.   :111.0   Boy :3   Bottle:2
##  1st Qu.:22.00   1st Qu.:180.0   Girl:3   Breast:2
##  Median :39.00   Median :358.5            Suppl :2
##  Mean   :39.67   Mean   :306.0
##  3rd Qu.:47.75   3rd Qu.:420.0
##  Max.   :77.00   Max.   :447.0
```

`Hmisc::describe` es otra función un poco más completa para obtener resúmenes de variables (disponible en el paquete `Hmisc`). Incluye número total de casos en la variable (`n`), número de datos faltantes y de valores únicos, los cuantiles más importantes y los cinco valores mínimos y máximos de la variable. Para variables que se interpreten como recuento (`pregnant` en el siguiente ejemplo, de tipo `integer`) se incluye un resumen numérico de la frecuencia de aparición de cada valor. Finalmente, usando `latex(describe)` podemos obtener una versión de la salida lista para insertar en un documento LaTeX, incluyendo pequeños histogramas resumen para las variables cuantitativas (ver ejemplo).

```
# Instalar el paquete Hmisc previamente si es necesario
# Dataset `pima`: estudio en 768 mujeres indias pima viviendo cerca de Phoenix, EE.UU.
library(Hmisc)
describe(pima[,1:4])
```

```
## pima[, 1:4]
##
##  4  Variables      768  Observations
## ---------------------------------------------------------------------------
## pregnant
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      768        0       17    0.986    3.845    3.698        0        0
##      .25      .50      .75      .90      .95
##        1        3        6        9       10
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency    111   135   103    75    68    57    50    45    38    28
```

```
## Proportion 0.145 0.176 0.134 0.098 0.089 0.074 0.065 0.059 0.049 0.036
##
## Value         10    11    12    13    14    15    17
## Frequency     24    11     9    10     2     1     1
## Proportion 0.031 0.014 0.012 0.013 0.003 0.001 0.001
## ------------------------------------------------------------------------
## glucose
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      768        0      136        1    120.9    35.48     79.0     85.0
##      .25      .50      .75      .90      .95
##     99.0    117.0    140.2    167.0    181.0
##
## lowest :   0  44  56  57  61, highest: 195 196 197 198 199
## ------------------------------------------------------------------------
## diastolic
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      768        0       47    0.998    69.11    18.81     38.7     54.0
##      .25      .50      .75      .90      .95
##     62.0     72.0     80.0     88.0     90.0
##
## lowest :   0  24  30  38  40, highest: 106 108 110 114 122
## ------------------------------------------------------------------------
## triceps
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      768        0       51    0.974    20.54    17.93        0        0
##      .25      .50      .75      .90      .95
##        0       23       32       40       44
##
## lowest :  0  7  8 10 11, highest: 54 56 60 63 99
## ------------------------------------------------------------------------
```

Otras medidas de posición

Otros estadísticos de posición que podemos obtener para cualquier variable cuantitativa (continua o discreta) son los *cuantiles*. Tras ordenar los valores de una variable cuantitativa, el cuantil `q` es el valor para el cual el q% de los valores quedan por debajo del mismo. La mediana corresponde al q-50.

```
## Medidas de posición relativa
# Cuantiles, por defecto devuelve los cuartiles de la distribución de valores
quantile(pima$glucose)
```

```
##      0%     25%     50%     75%    100%
##    0.00   99.00 117.00 140.25 199.00
```

En ocasiones, nos referimos a los *cuartiles*, los *deciles* o a los *percentiles* de la distribución de valores de una variable. Una vez ordenados los valores de una variable, los cuartiles la dividen en cuatro partes iguales (cuantiles q-25, q-50 y q-75), los deciles son los valores que dividen la variable en 10 partes iguales (q-10, q-20, q-30 etc.), mientras que los percentiles son los valores que dividen la variable en 100 partes iguales (q-1, q-2, q-3 etc.).

Por último, es conveniente resaltar que existen muchas formas de calcular los cuantiles. La función `quantile` de R implementa hasta 9 métodos distintos (!!), siendo el método por defecto el que se empleaba también en el lenguaje de programación estadística S, del que deriva R (más detalles con `?var`).

Medidas de dispersión

Las medidas de dispersión indican lo alejados que están los valores de una variable cuantitativa con respecto a su posición central (normalmente la media o la mediana). Algunas medidas de dispersión importantes son:

- **Rango**: Diferencia entre los valores mínimo y máximo de la variable.
- **Varianza**: Mide la dispersión de los valores de la variable respecto a la media. Su cálculo el distinto dependiendo de si trabajamos con una muestra de una población, a partir de la cual queremos calcular un estimador de la varianza real de dicha población que llamamos *varianza muestral*, o usamos todos los datos de la población (si es posible) para calcular el valor real de la *varianza poblacional*. Es importante señalar que la función `var` en R (mostrada en el siguiente ejemplo) siempre calcula la *varianza muestral* y por defecto usa para calcularla el método de Pearson (ver fórmulas (1) y (2) a continuación).
  - En la *varianza muestral* s^2 la fórmula (1) se divide por $n-1$ (siendo n el número de elementos de la muestra) para que el estimador no esté sesgado (es decir, se acerque más al valor real de la varianza de la población). En el numerador usamos la media muestral $\bar{x}$, ya que en general la media poblacional μ se desconoce. Sin embargo, se puede demostrar que, en promedio, las observaciones x_i de la muestra tienden a estar más cerca de la media muestral $\bar{x}$ que de la media poblacional μ. Para compensar este efecto se divide por $n-1$, ya que dividiendo por n la varianza muestral s^2 sería una medida de dispersión, en promedio, consistentemente más pequeña que el verdadero valor de la varianza poblacional μ que intenta estimar.
    Otra forma de justificarlo es mediante el concepto estadístico de *grados de libertad*. Las n desviaciones respecto a la media $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$ en el numerador siempre suman cero (no así la suma de sus cuadrados). Así que si fijamos libremente los valores de $x-1$ de estas cantidades el valor de la restante siempre queda determinado de forma automática.
  - En la *varianza poblacional* σ^2 la fórmula (2) se divide por N (tamaño de la población), puesto que conocemos la media real de la población μ (no intentamos calcular un estimador).

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N} \tag{2}$$

- La **desviación típica** es la raíz cuadrada positiva de la varianza muestral o poblacional. La desviación típica muestral se denota usualmente por s y la poblacional normalmente por σ. Su utilidad radica en que permite medir la dispersión de los valores respecto a la media en las mismas unidades que la variable, por lo que su interpretación es más sencilla.
- El **rango intercuartílico** (IQR) se calcula como la diferencia entre el tercer cuartil (percentil 75) y el primer cuartil (percentil 25).
- El **coeficiente de variación** (CV) representa la desviación típica en unidades de la media, y se suele expresar en porcentaje. Por ejemplo, $CV = 60\%$ indica que el valor de la desviación típica es 0.6 veces la magnitud de la media.
- **Mediana de las desviaciones absolutas** (respecto a la mediana): Se calcula como `constant * cMedian(abs(x - center))`, donde $constant = 1.4826$ (es decir `1/qnorm(0.75)`), para intentar asegurar consistencia al comparar distintas variables, siempre que su distribución sea cercana a una Normal. Esta medida de dispersión es preferible cuando la mediana es mejor estadístico de centralidad que la media (por ejemplo, si hay muchos valores atípicos en la distribución). En R la función `mad` calcula en realidad la MADN (MAD normalizada por el percentil 75 de la distribución normal). Para más información ver `?mad`.

```
## Medidas de dispersión

# Rango
range(pima$glucose)
```

```
## [1]   0 199
```

```
# Varianza
# Siempre calcula varianza muestral, dividiendo por n-1
# Por defecto usa método de Pearson (method = "pearson"). Otros métodos disponibles so
# "kendall" y "spearman", ver ?var
var(pima$glucose)
```

```
## [1] 1022.248
```

```
# Desviación típica o estándar
sd(pima$glucose)
```

```
## [1] 31.97262
```

```
# Rango intercuartílico
IQR(pima$glucose)
```

```
## [1] 41.25
```

```
# Coeficiente de variación
# Desviación estándar medida en unidades de la media de la distribución
# y normalmente expresada en porcentaje
Cv <- function(mean, sd){
      (sd/mean)*100
      }
Cv(mean(pima$glucose), sd(pima$glucose))
```

```
## [1] 26.4467
```

```
# Mediana de las desviaciones absolutas (MAD)
mad(pima$glucose) # MADN
```

```
## [1] 29.652
```

Medidas de forma

El **coeficiente de asimetría** mide la falta de simetría en la distribución de valores de una variable. Una distribución de valores perfectamente simetrica tendrá un coeficiente de asimetría cero. Si el coeficiente de asimetría es positivo, entonces tenemos mucha densidad para valores bajos (a la izquierda) y una cola larga a la derecha, con muchos valores altos que se separan de la media. Si es negativo, entonces sucede al contrario: mucha densidad para valores altos (a la derecha) y cola larga a la izquierda, con muchos valores bajos que se separan de la media.

```
## Medidas de forma (de la distribución de valores)

# Coeficiente de asimetría (skewness)
# Instalar el paquete `moments` si fuese necesario
library(moments)
skewness(rbeta(n = 10000, shape1 = 2, shape2 = 5)) # izq. gráfico, coef. asimetría posi
```

```
## [1] 0.5785899
```

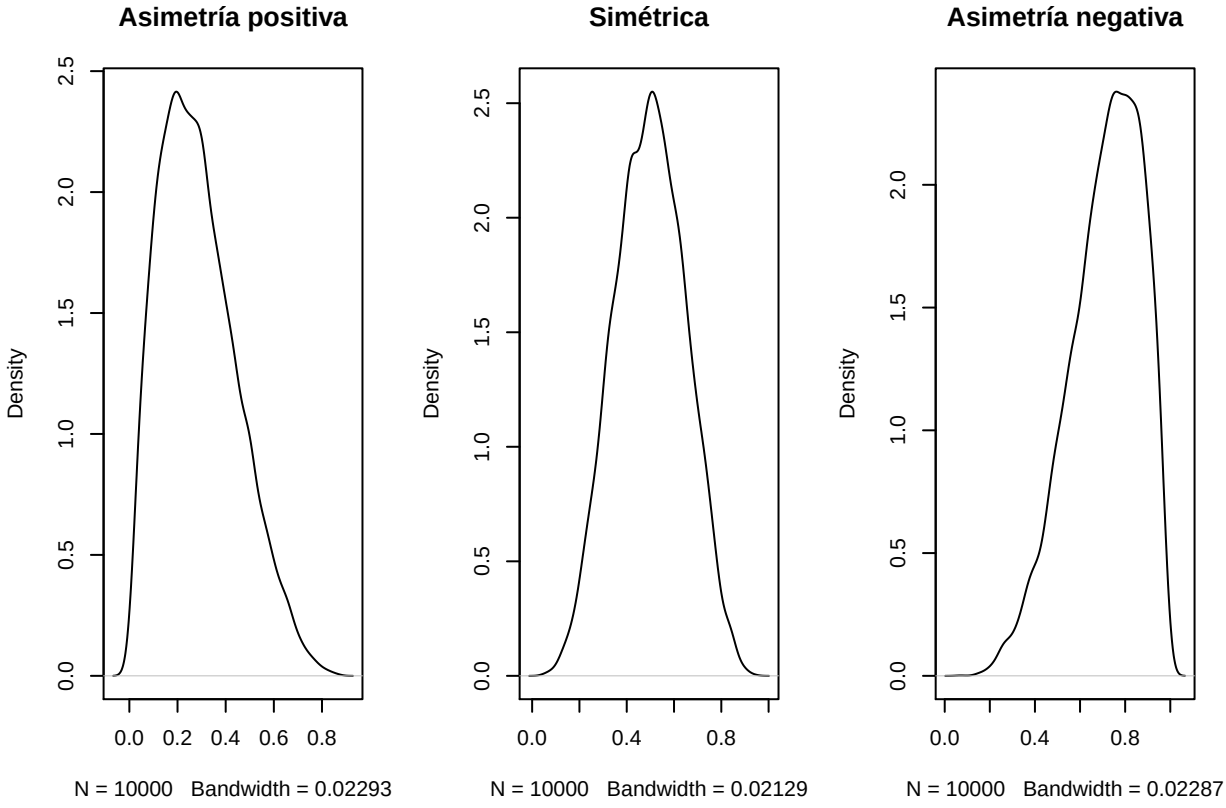
```r
skewness(rbeta(n = 10000, shape1 = 5, shape2 = 5)) # centro, teóricamente coef. asimetr
```

```
## [1] -0.02002136
```

```r
skewness(rbeta(n = 10000, shape1 = 5, shape2 = 2)) # der. gráfico, coef. asimetría nega
```

```
## [1] -0.5761046
```

```r
# KDEs (densidad de probablidad) de las dos distribuciones y una Normal (centro) para
# Usamos una distribución Beta para comparar
par(mfrow=c(1,3))
plot(density(rbeta(n = 10000, shape1 = 2, shape2 = 5)),
     main = "Asimetría positiva")
plot(density(rbeta(n = 10000, shape1 = 5, shape2 = 5)),
     main = "Simétrica")
plot(density(rbeta(n = 10000, shape1 = 5, shape2 = 2)),
     main = "Asimetría negativa")
```



El **coeficiente de apuntamiento** (curtosis) mide cómo de puntiaguda es la función de densidad de los valores de una variable, comparada con una distribución Normal. El valor teórico de la **curtosis** para una **distribución Normal** es $K = 3$, por lo que las comparaciones hay que hacerlas respecto a dicho valor de referencia. Según esto, podemos distinguir los siguientes tipos:

- Platycúrtica: la forma de la distribución es más aplanada que la de una Normal.
- Mesocúrtica: apuntamiento similar al de una distribución Normal.
- Leptocúrtica: la forma de la distribución es más puntiaguda que la de una Normal.

```
# Coeficiente de apuntamiento (curtosis)

kurtosis(rbeta(10000, shape1 = 2, shape2 = 2)) # Platicúrtica; Bernouilli, prob = 0.5
```

```
## [1] 2.14837
```

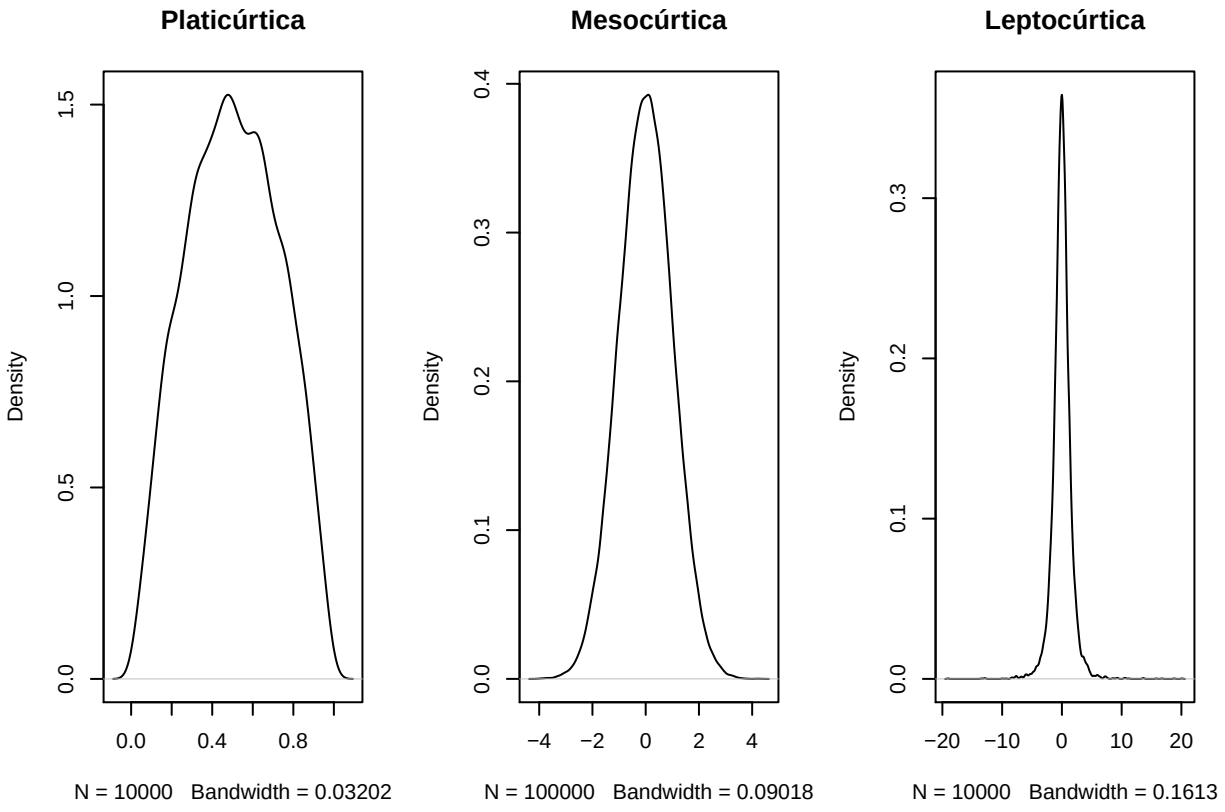
```
kurtosis(rnorm(100000)) # Mesocúrtica
```

```
## [1] 3.017879
```

```
kurtosis(rt(10000, df = 3))# Leptocúrtica
```

```
## [1] 276.4464
```

```
par(mfrow=c(1,3))
plot(density(rbeta(10000, shape1 = 2, shape2 = 2)), main = "Platicúrtica")
plot(density(rnorm(100000)), main = "Mesocúrtica")
plot(density(rt(10000, df = 3)), main = "Leptocúrtica")
```



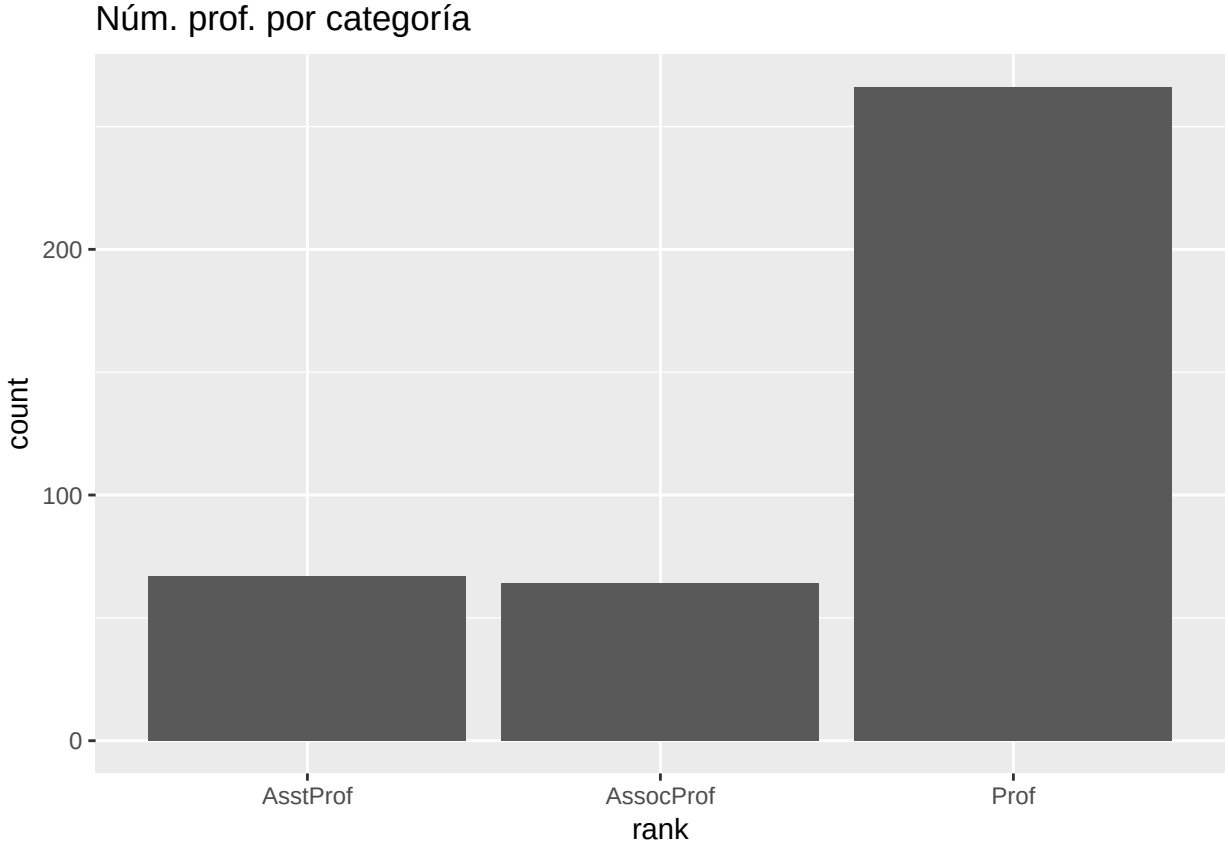
Visualización de datos para EDA

Al igual que ocurre con las técnias para obtener resúmenes numéricos de datos, las técnicas gráficas para descriptiva y exploración de datos son distintas dependiendo de si las variables que aparecen son cuantitativas o cualitativas. A continuación presentamos algunas de las herramientas más útiles.

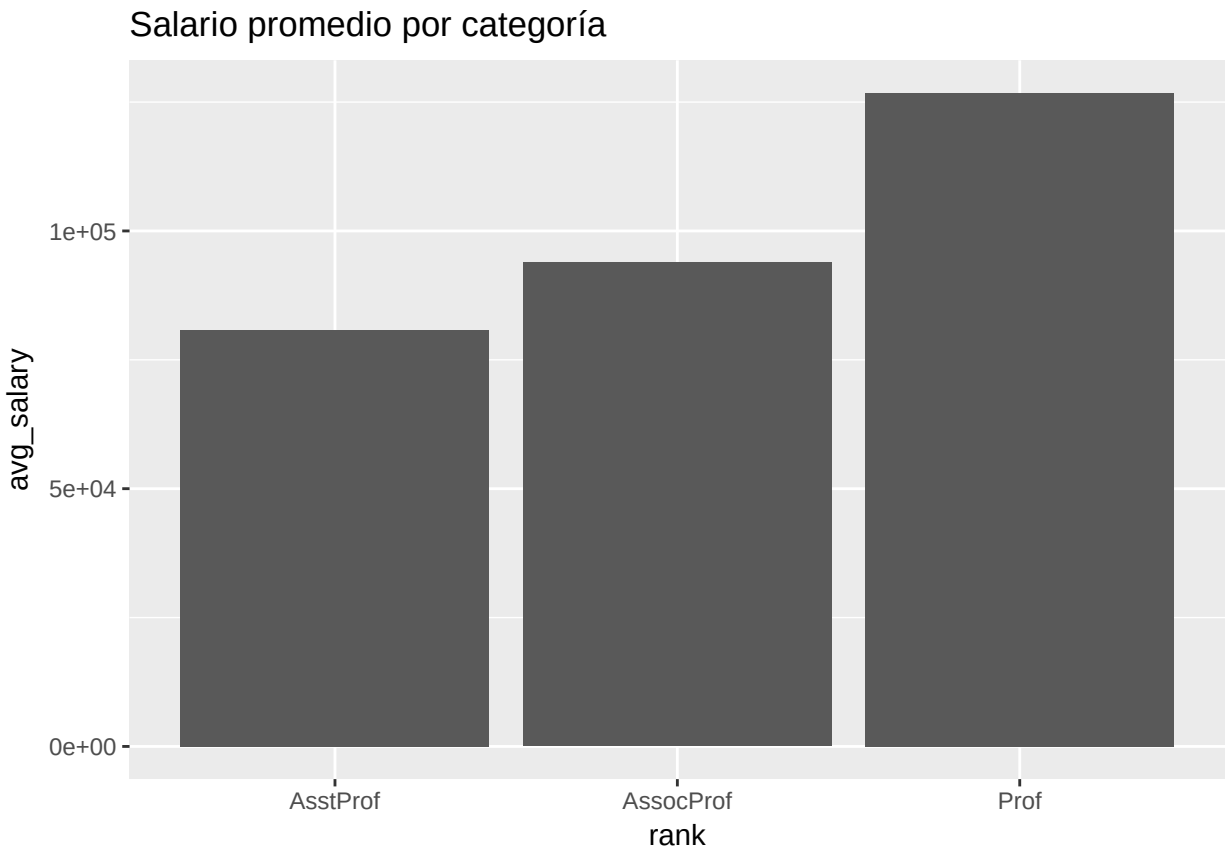
Gráficos para EDA con variables cualitativas individuales

En estadística descriptiva univariante, un gráfico muy frecuente para representación de variables cualitativas es el gráfico de barras. A continuación, mostramos un ejemplo con el conjunto de datos `Salaries` del paquete `car`.

```
# Gráfico de barras
# Instalar biblioteca `car` si es preciso
# `car::Salaries`: datos sobre salarios (9 meses) de profesores en EE.UU.
library(car)
library(dplyr)
library(ggplot2)
# Número de casos por categoría
ggplot(Salaries, aes(rank)) + geom_bar() + ggtitle("Núm. prof. por categoría")
```



```r
# Salario promedio por categoría
# Se puede hacer de otras formas, pero usando dplyr es bastante intuitivo
Salaries %>%
  group_by(rank) %>%
  summarise(avg_salary = mean(salary)) %>%
  ggplot(aes(x=rank, y=avg_salary)) + geom_bar(stat = "identity") +
  ggtitle("Salario promedio por categoría")
```



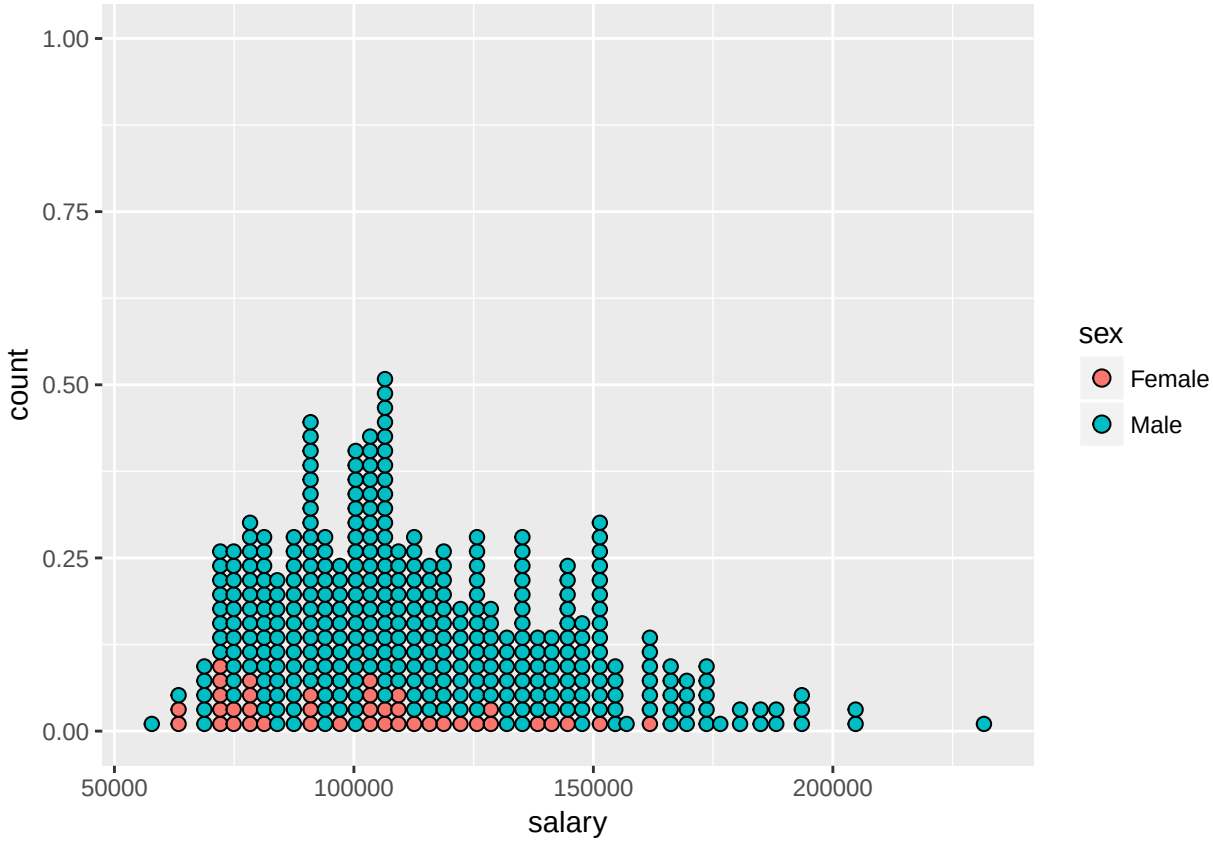
También frecuente encontrar gráficos de sector (en inglés *pie chart*) para representar estos datos. Es un tipo de gráfico muy utilizado en medios de comunicación e informes corporativos. Sin embargo, tiene varios problemas graves. El principal de ellos es que el ser humano suele tener bastantes problemas para percibir correctamente diferencias en sectores angulares, especialmente cuando su amplitud es muy similar.

No obstante, el gráfico de sector sí puede tener utilidad en el caso de que las diferencias entre sectores sean muy grandes, o para representación simbólica en infografía. En cualquier caso, desde el punto de vista de la estadística descriptiva expertos como Edward Tufte o Stephen Few llevan muchos años desaconsejando vivamente su utilización, a pesar de que como decíamos antes su uso siga proliferando en ciertos entornos.

Gráficos para EDA con variables cuantitativas individuales

El más sencillo de ellos es el **gráfico de puntos** (*dotplot*), propuesto por L. Wilkinson (1999).

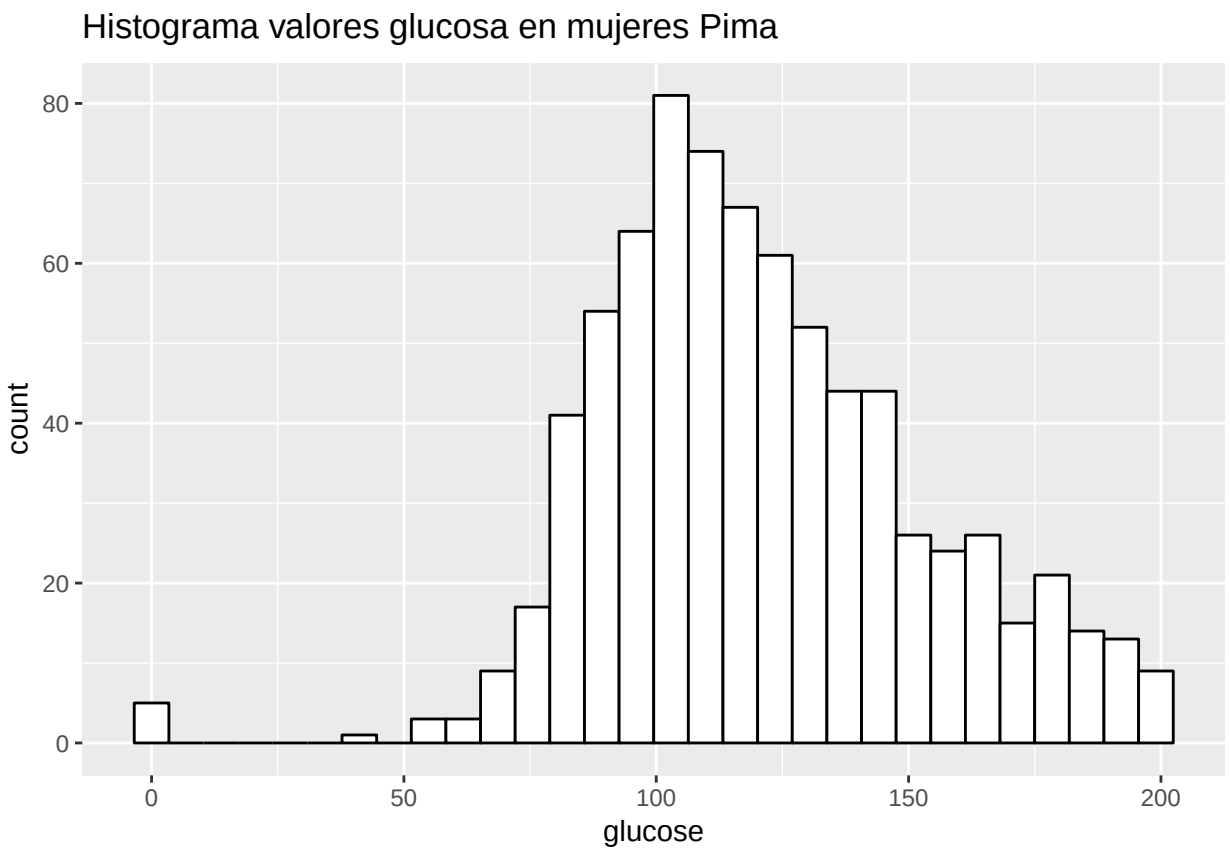
```
# Diagrama de puntos

## Wilkinson dotplot (como histograma pero con casos individuales)

ggplot(Salaries, aes(x = salary, fill = sex)) +
  geom_dotplot(binwidth = 3000, stackgroups = TRUE, binpositions="all")
```



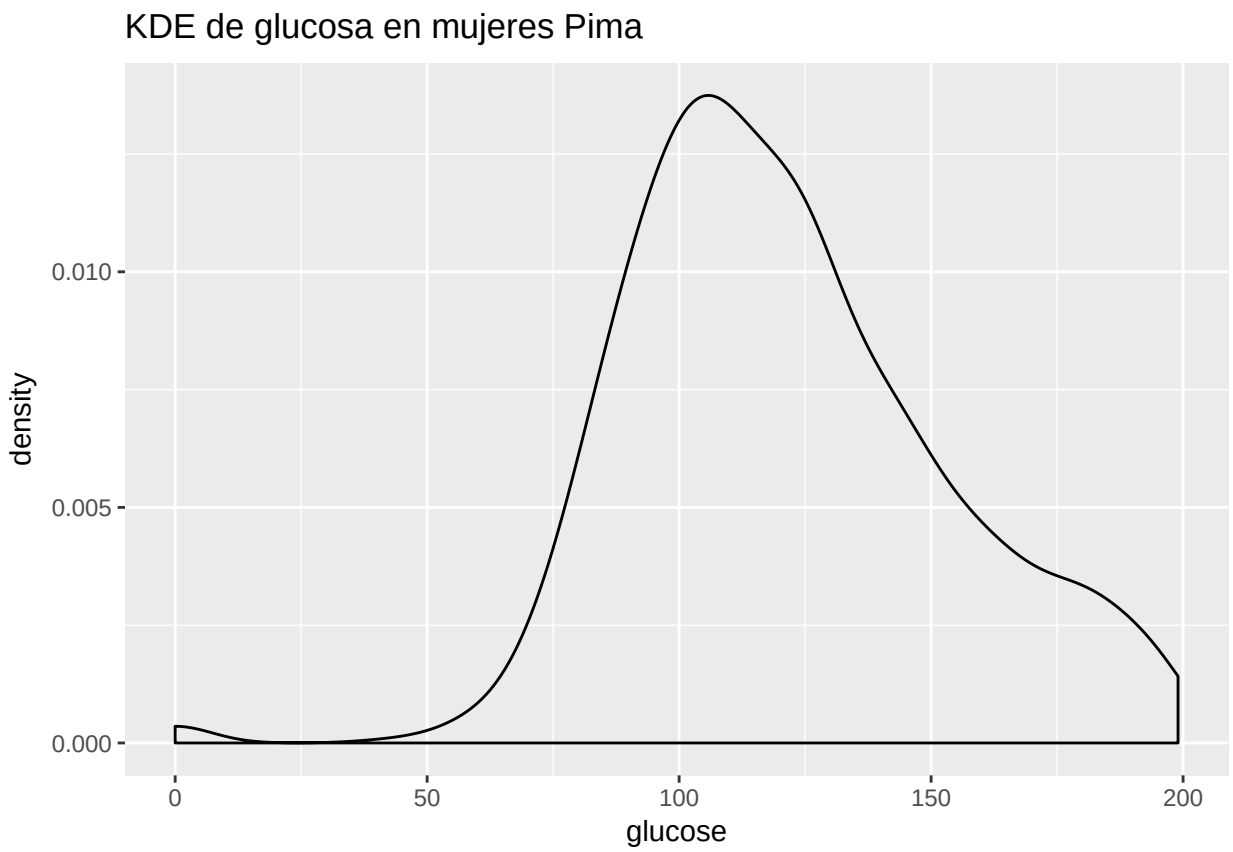
El **histograma** proporciona una representación de la distribución de frecuencias (o la densidad de probabilidad) de los valores de la variable. Por defecto, las funciones de R (como la función estándar `hist` o `ggplot2::geom_histogram`) eligen algoritmos para calcular el tamaño y número de los intervalos utilizados para representar el histograma. Sin embargo, podemos también configurar esa selección.

```
# Histograma
ggplot(pima, aes(x = glucose)) +
  geom_histogram(fill="white", colour="black") +
  ggtitle('Histograma valores glucosa en mujeres Pima')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Función de densidad de probabilidad (KDE)
ggplot(pima, aes(x = glucose)) +
geom_density() +
ggtitle('KDE de glucosa en mujeres Pima')
```

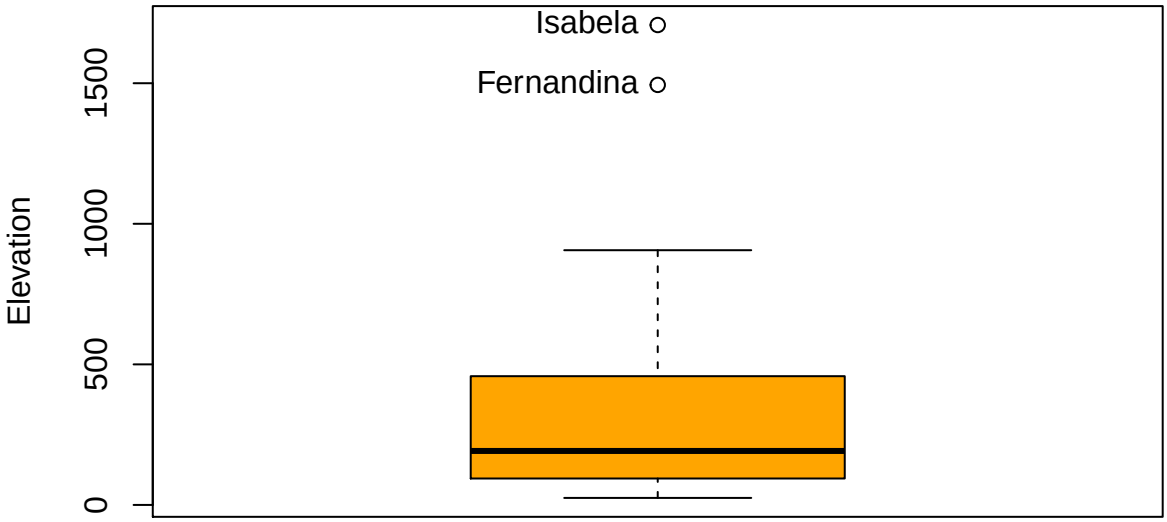


Otro gráfico muy utilizado para variables cuantitativas univariantes es el **boxplot**, también llamado *box-and-whisker plot* (diagrama de caja y bigotes). Es especialmente útil para detectar posibles datos atípicos en los valores de una variable, siempre que su distribución sea parecida a una Normal. El gráfico muestra:

- Una caja cuyos límites son el primer y el tercel cuartil de la distribución de valores.
- Una línea central, que marca la mediana.
- Los bigotes, que por defecto (en R) se extienden hasta 1.5 veces el valor del rango intercuartílico (IQR) por encima y por debajo de la caja.
- Puntos individuales, que quedan más allá del límite de los bigotes, marcan posibles datos atípicos.

En distribuciones muy asimétricas o con muchos valores extremos, muy diferentes a una Normal, aparecerán demasiados puntos más allá de los bigotes y no se podrán apreciar fácilmente los atípicos (demasiados puntos considerados como tales). En ese caso, es conveniente intentar una transformación de la variable antes de representar el boxplot.

```
# Boxplot
# Función `car::Boxplot` proporciona opciones avanzadas
# por ejemplo, identificación automática de atípicos en el gráfico (necesita nombres
# para cada fila/caso en el data frame)
# http://proquest.safaribooksonline.com/book/programming/r/9781449363086/6dot-summariz
Boxplot(~Elevation, data = gala, col= 'orange')
```

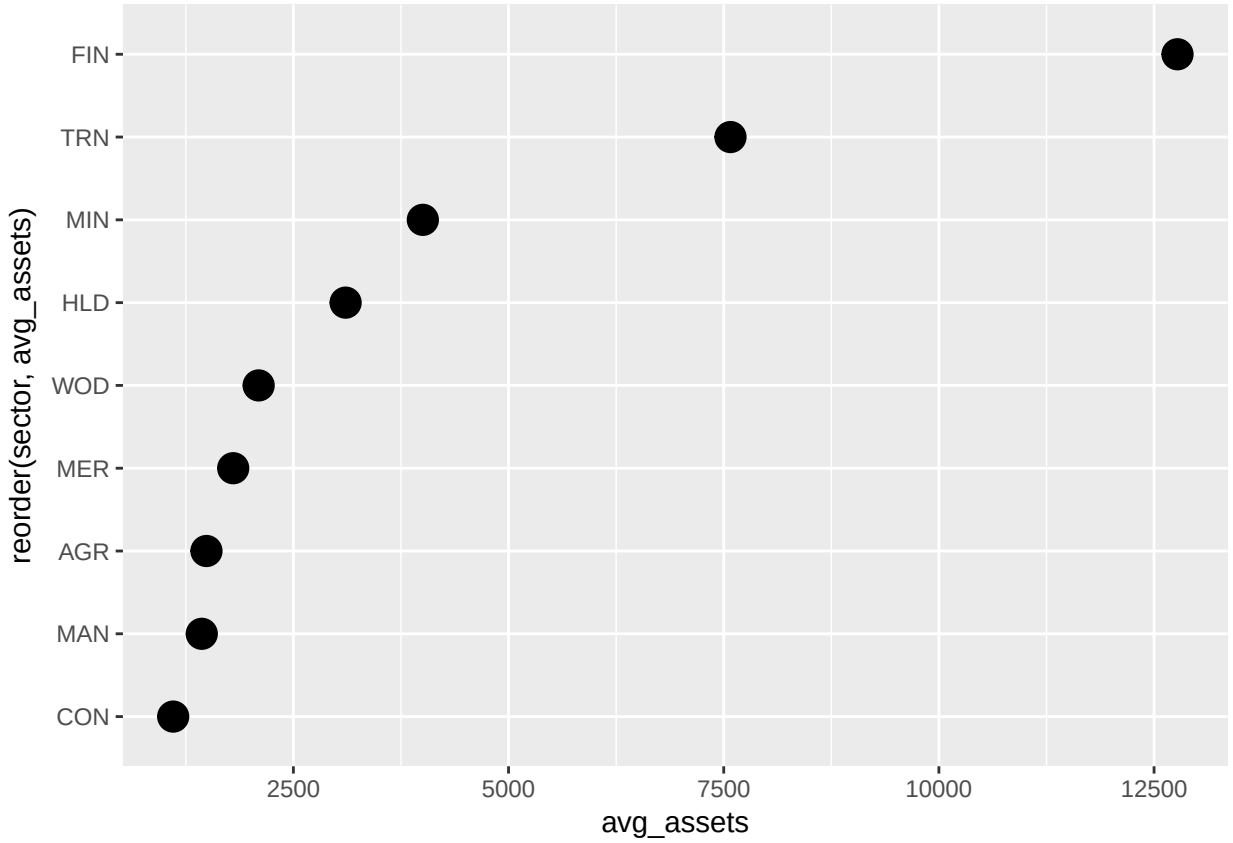



```
## [1] "Fernandina" "Isabela"
```

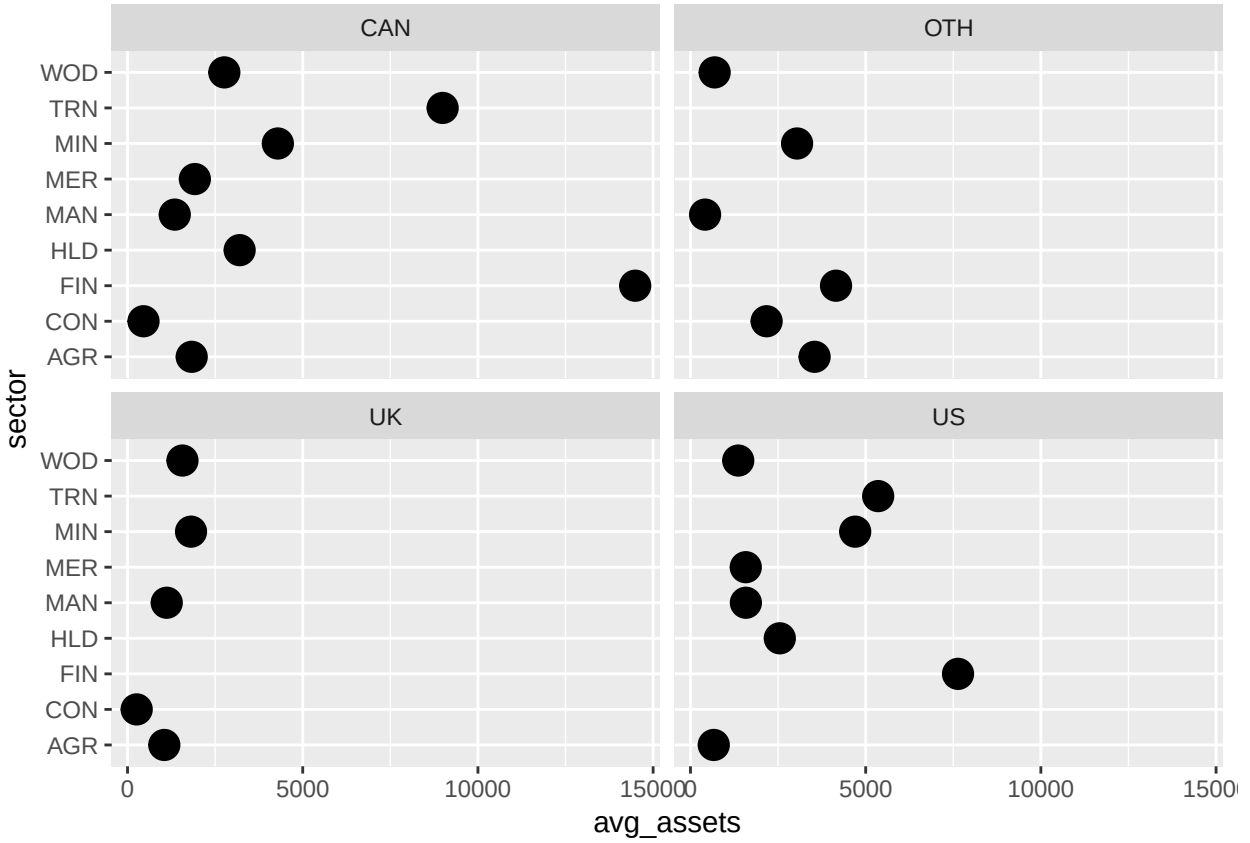
Gráficos para EDA multivariante

Una de las facetas más potentes e interesantes del EDA es la de poder representar gráficamente varias variables para poder explorar la existencia de posibles relaciones entre ellas. El primer ejemplo de este tipo de gráficos es una variante del diagrama de puntos (*dotplot*) propuesta por William. S. Cleveland como alternativa al diagrama de barras, para explorar posibles relaciones entre variables cuantitativas y factores. Además, se pueden representar varios paneles para incluir dos variables categóricas en el mismo gráfico.

```
# Diagrama de puntos (dotplot)

## Cleveland dotplot (variable cuantitativa frente a categórica)
## Media de activos por sector para todos los países
Ornstein %>%
  filter(sector != 'BNK') %>%
  group_by(sector) %>%
  summarize(avg_assets = mean(assets)) %>%
  ggplot(aes(x = avg_assets, y = reorder(sector, avg_assets))) +
  geom_point(size = 5)
```



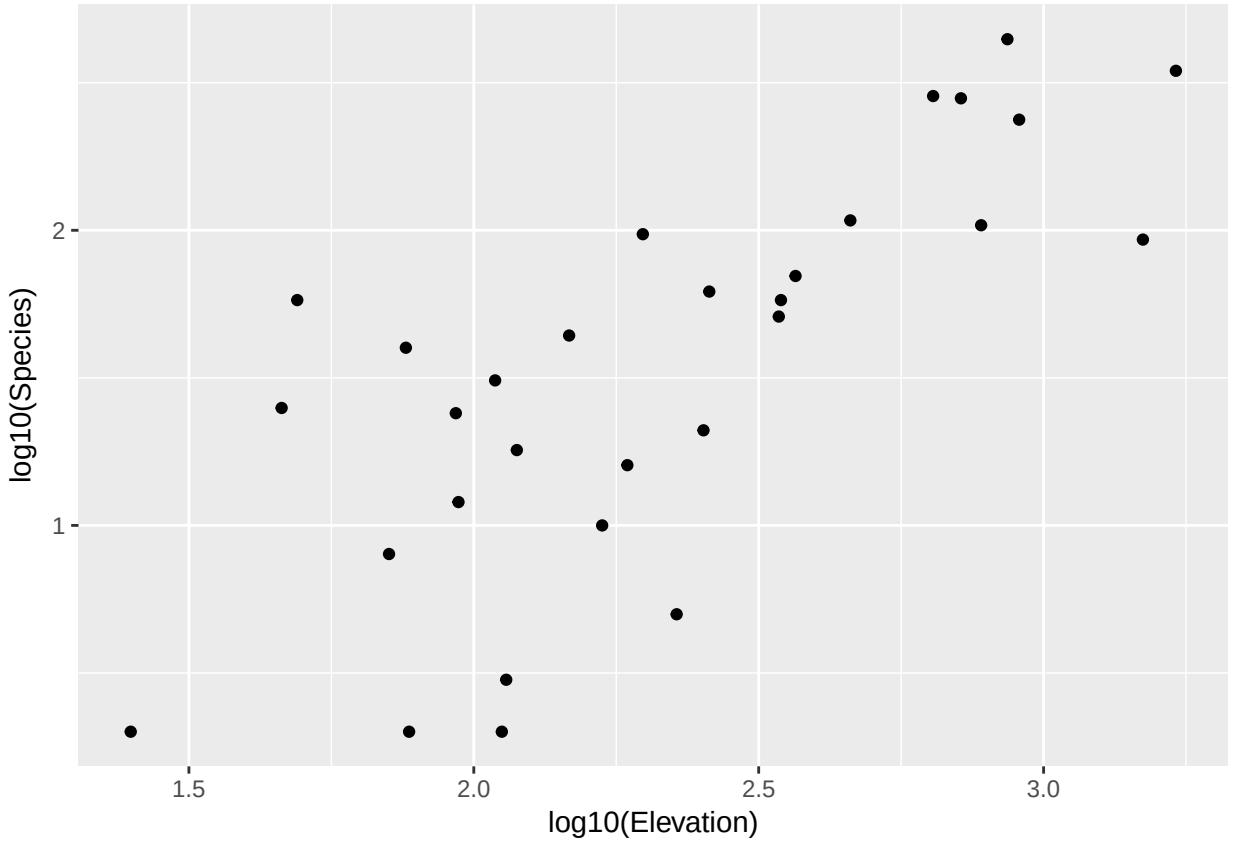

```
## Por paneles
Ornstein %>%
  filter(sector != 'BNK') %>%
  group_by(nation, sector) %>%
  summarize(avg_assets = mean(assets)) %>%
  ggplot(aes(x = avg_assets, y = sector)) +
  geom_point(size = 5) +
  facet_wrap(~ nation)
```



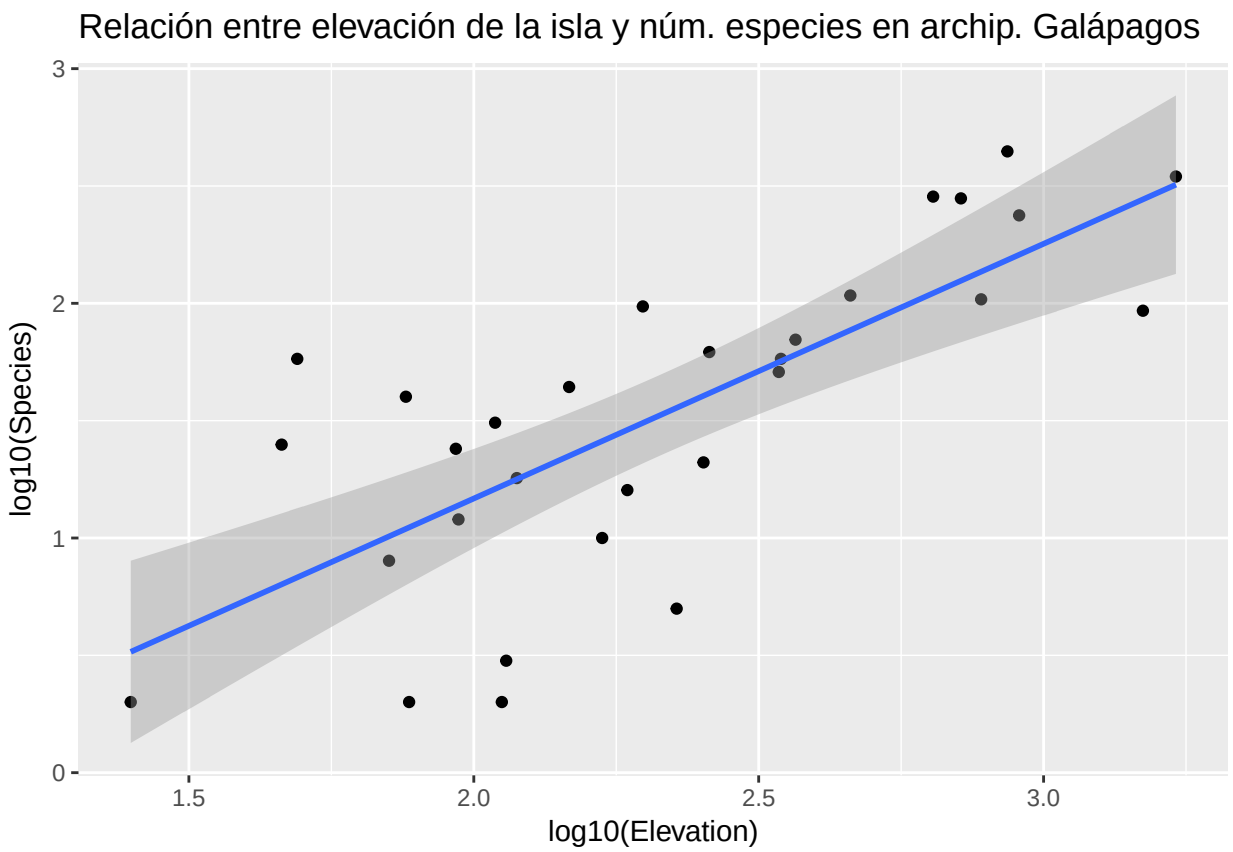
El **diagrama de dispersión** (en inglés *scatterplot*) es otro ejemplo importante de gráfico para explorar la relación entre dos variables cuantitativas. Además disponemos de muchas alternativas para explorar simultáneamente posibles relaciones con variables categóricas (varios paneles o uso de diferentes colores o tipo de caracter para los puntos de la misma categoría).

Además, es bastante común añadir al gráfico la respresentación de un modelo de ajuste para los datos, que pueda revelar en qué medida existe una relación entre las variables. Opciones típicas son una recta de regresión lineal, un modelo lineal polinómico o un ajuste no paramétrico local (para obtener una curva suave). A continuación se muestran algunos ejemplos.

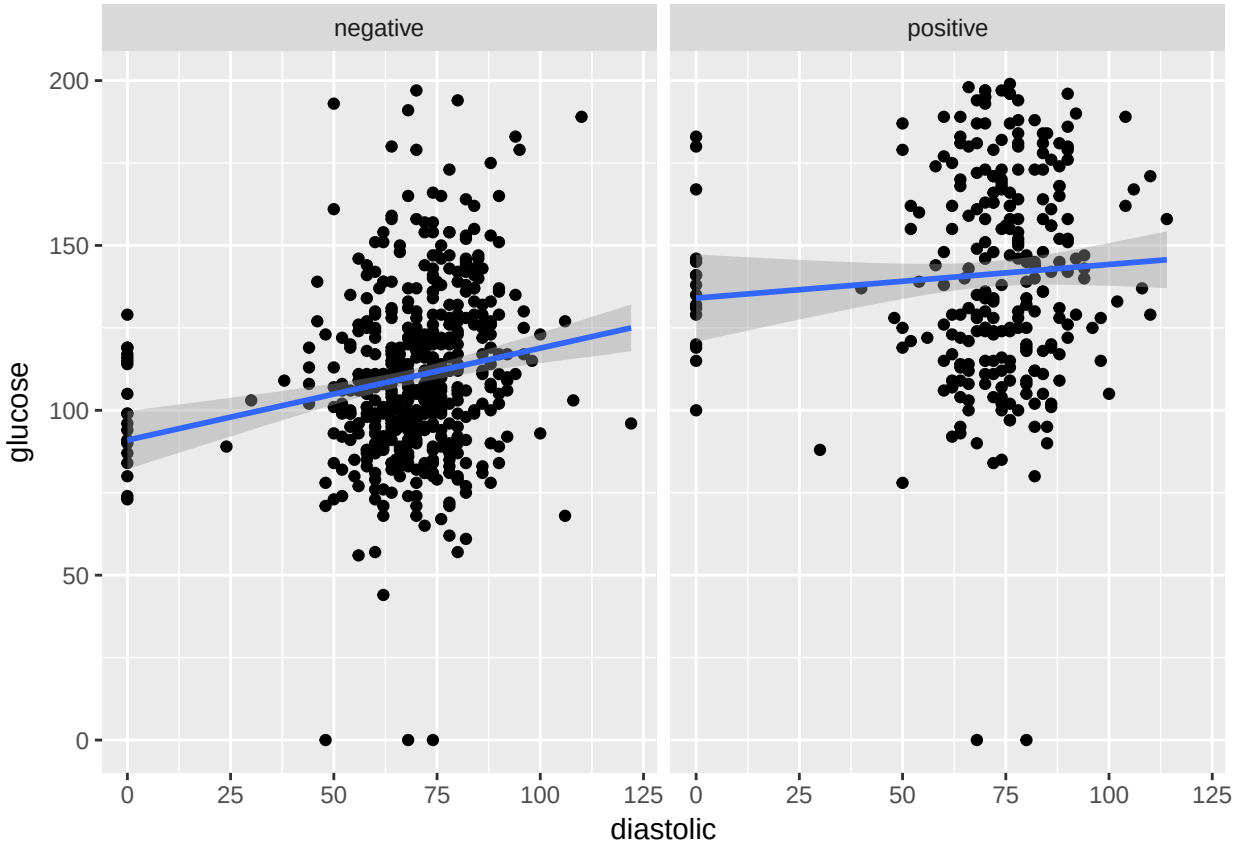
```
# Diagrama de dispersión (scatterplot)

## Individual (2 variables)
qplot(log10(Elevation), log10(Species), data = gala)
```




```
## Con ajustes (regresión, loess, etc.)
qplot(log10(Elevation), log10(Species), data = gala) +
  geom_smooth(method = "lm") +
  ggtitle('Relación entre elevación de la isla y núm. especies en archip. Galápagos')
```

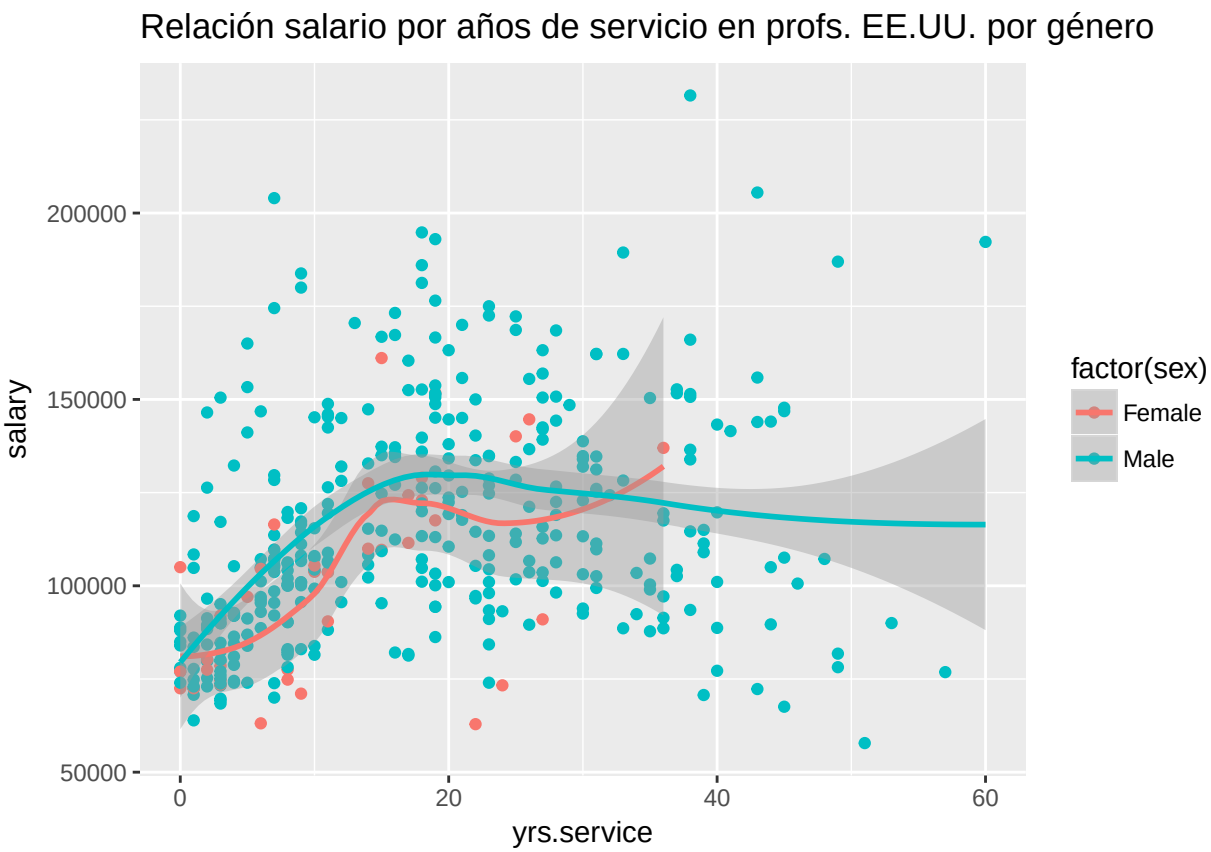



```
## Por paneles
pima %>%
  mutate(test = factor(test, labels = c('negative', 'positive'))) %>%
  ggplot(aes(x = diastolic, y = glucose)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ test)
```



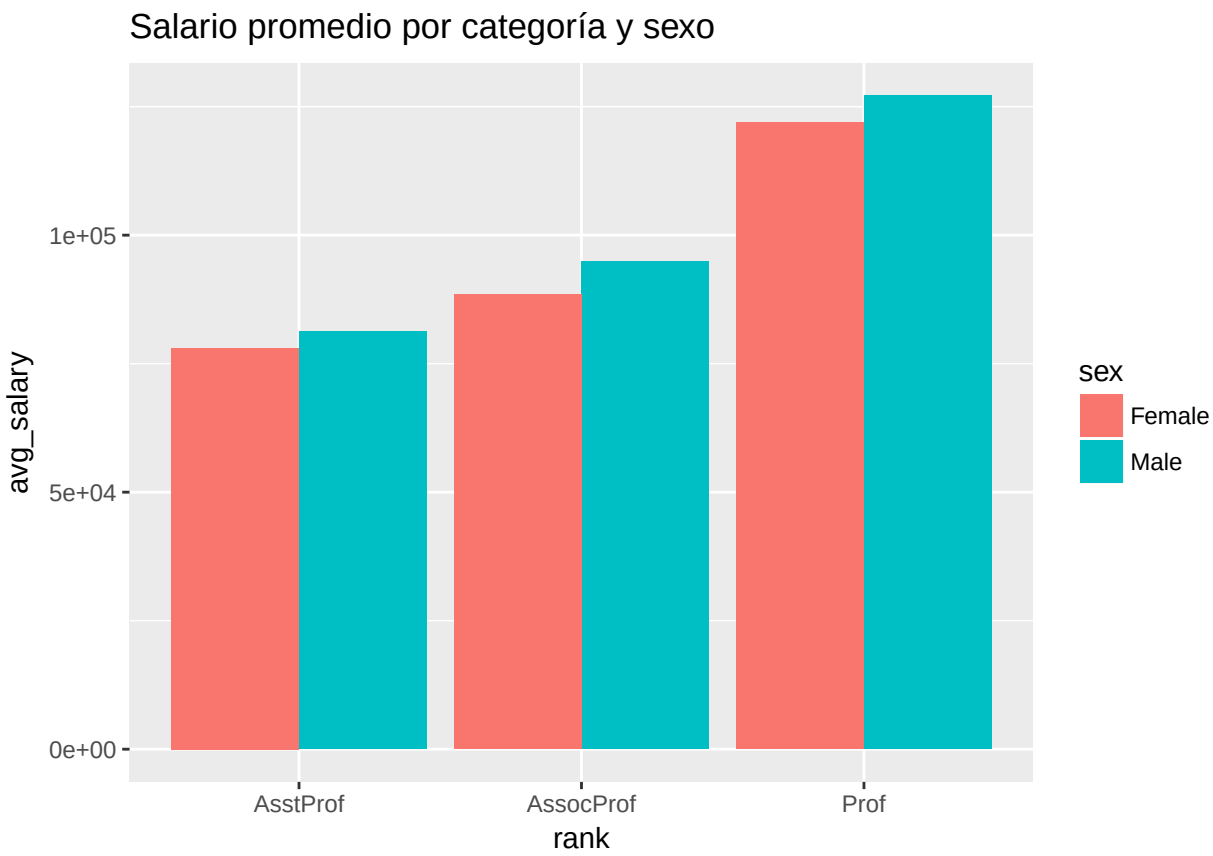

```
## Por colores o tipo de punto
qplot(yrs.service, salary, data = Salaries, colour = factor(sex)) +
  geom_smooth() +
  ggtitle('Relación salario por años de servicio en profs. EE.UU. por género')
```

```
## `geom_smooth()` using method = 'loess'
```

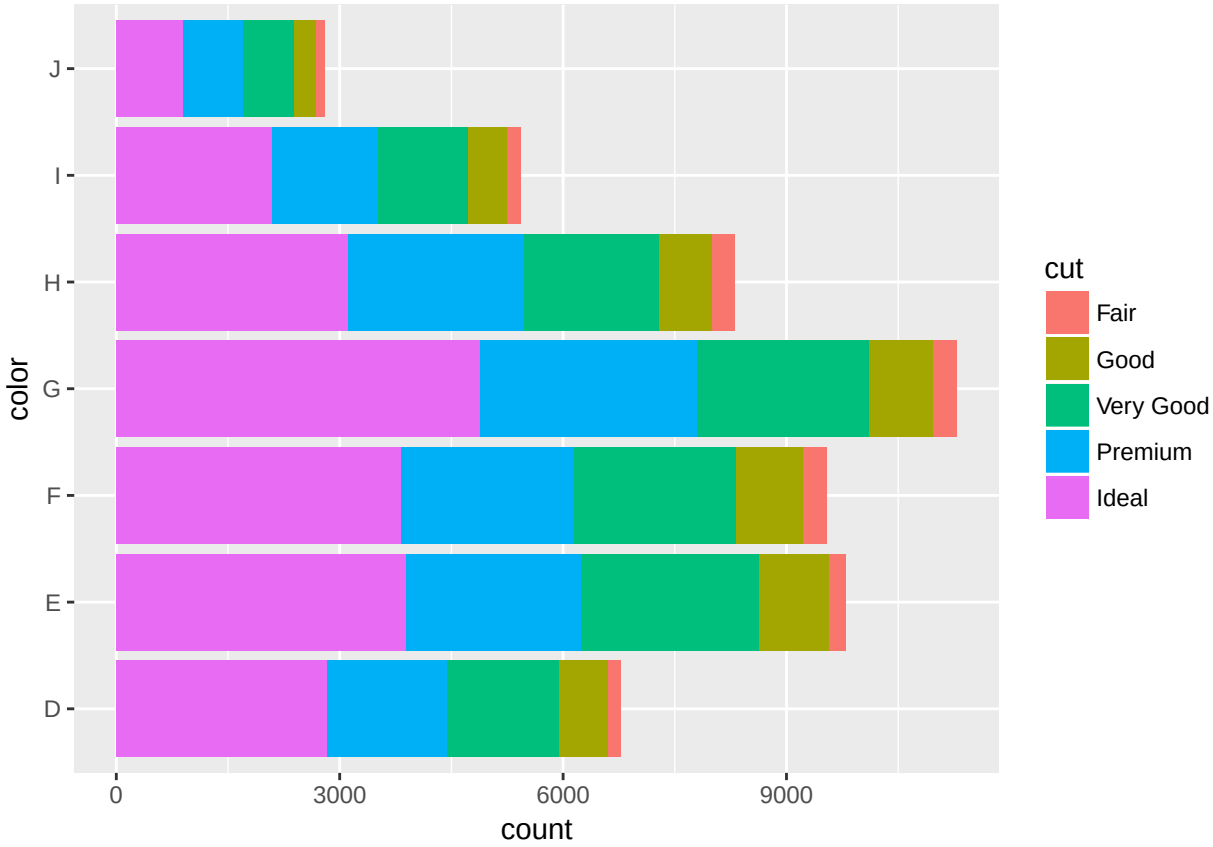


Análogamente, el **diagrama de barras** también puede modificarse para incluir variables cuantitativas y categóricas en el mismo gráfico, bien dividiendo la salida por paneles, por colores o apilando las barras. Las barras apiladas son muy útiles cuando queremos representar explícitamente la información de proporción sobre el total de elementos de una categoría (como alternativa al diagrama de sectores).

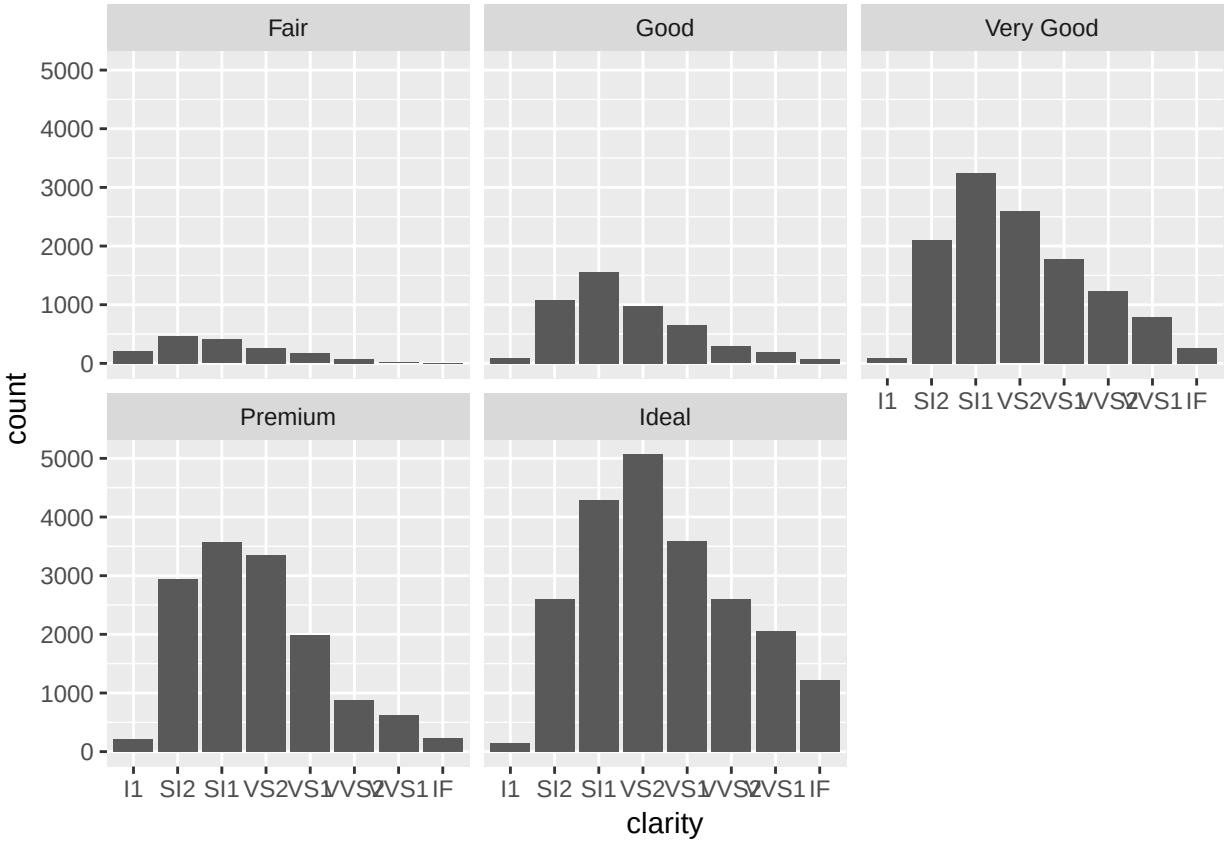
```
# Diagrama de barras (por paneles y colores)
# Por colores
Salaries %>%
  group_by(rank, sex) %>%
  summarise(avg_salary = mean(salary)) %>%
  ggplot(aes(x=rank, y=avg_salary, fill=sex)) + geom_bar(stat = "identity",
                                                         position = "dodge") +
  ggtitle("Salario promedio por categoría y sexo")
```




```
# Apilado
ggplot(diamonds, aes(color, fill=cut)) + geom_bar() + coord_flip()
```

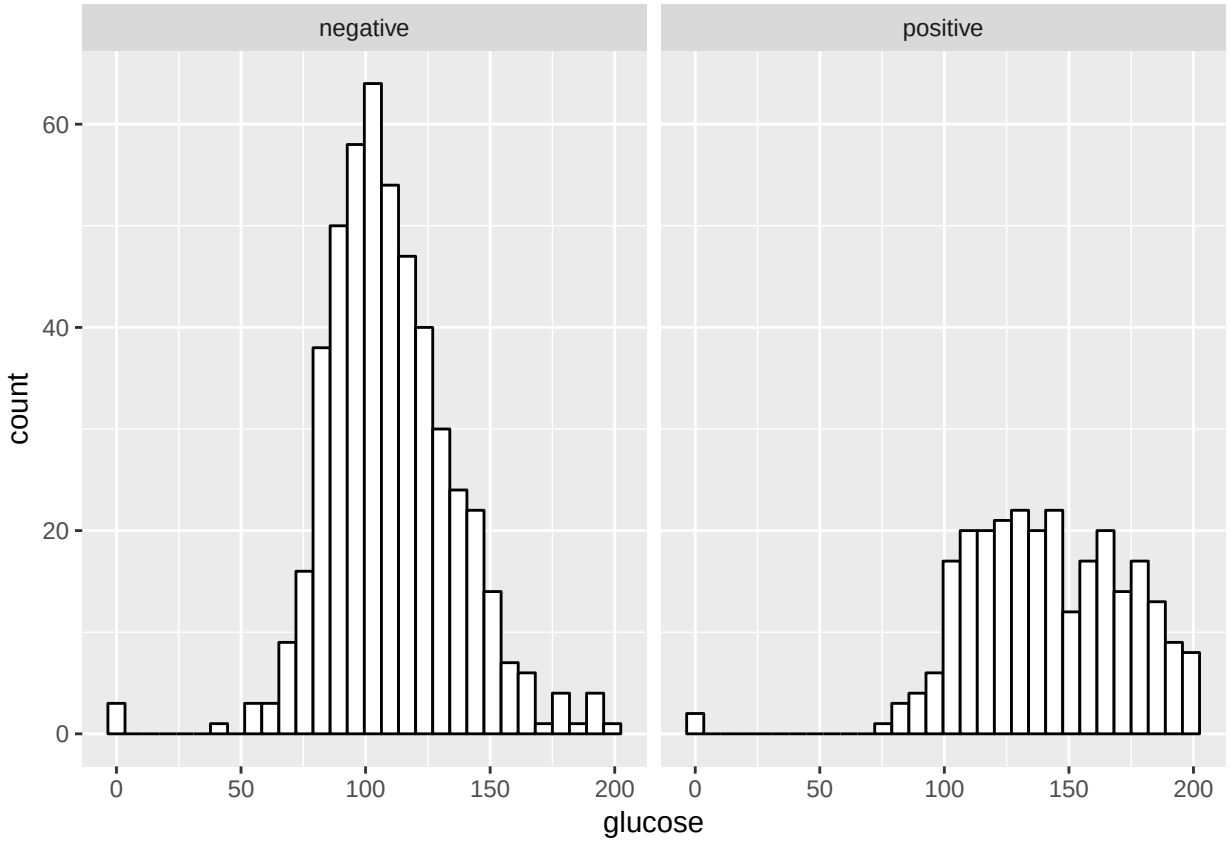



```
# Por paneles
ggplot(diamonds, aes(clarity)) + geom_bar() +
  facet_wrap(~ cut)
```

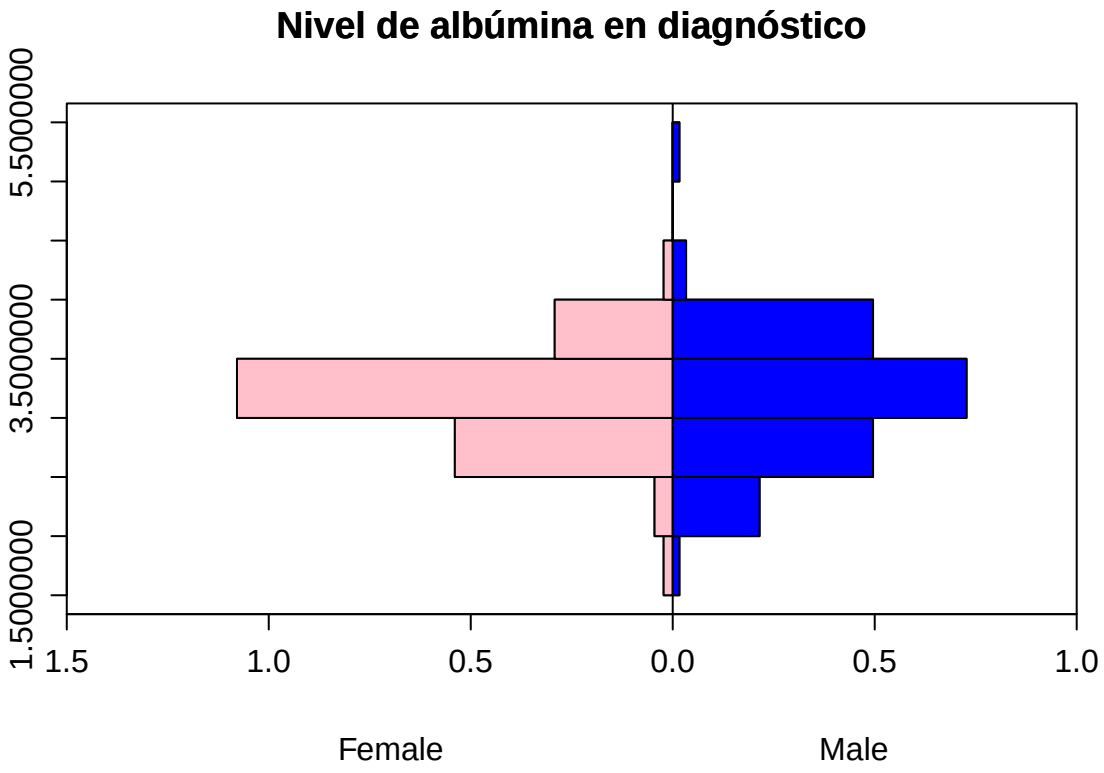


El **histograma** también se puede utilizar con mezcla de variables cuantitativas y categóricas, pero por su diseño solo es recomendable representar diferentes paneles con cada histograma, uno por cada nivel de la variable categórica a considerar. Un caso particular, frecuente en ciencias sociales y sobre todo en estudios demográficos, es el de los histogramas enfrentrados (que en esas áreas reciben el nombre de pirámides de población). Permiten comparar cara a cara los histogramas para las dos categorías de una variable dicótoma. En R podemos usar para este fin, entre otras alternativas, la función `Hmisc::histbackback`.

```
# Histogramas (recomendables solo por paneles)
pima %>%
  mutate(test = factor(test, labels = c('negative', 'positive'))) %>%
  ggplot(aes(x = glucose)) +
  geom_histogram(fill="white", colour="black") +
  facet_wrap(~ test)
```

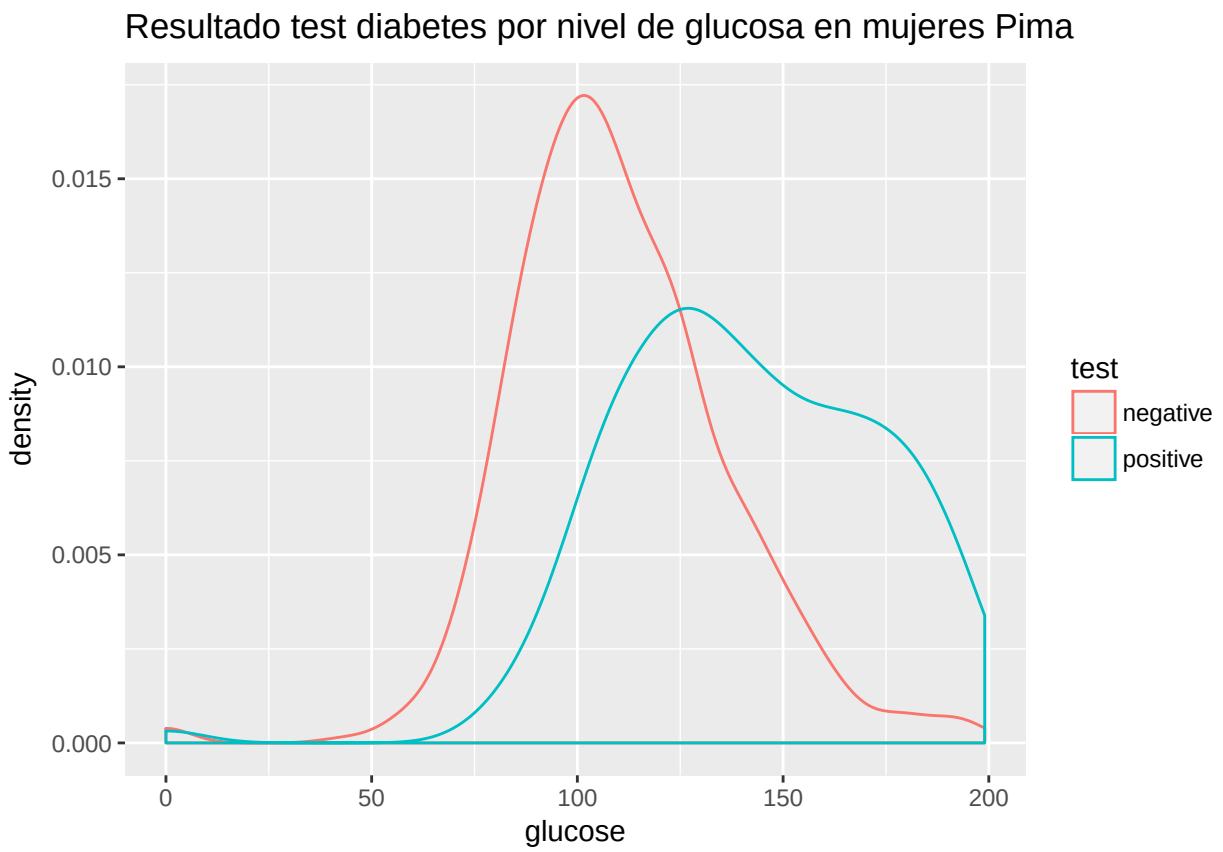



```r
# Histbackback (Hmisc)
# Similar a los gráficos de pirámide poblacional en demografía
# Importar biblioteca `survival` si es preciso para acceder a dataset `mgus`
library(survival)
out = histbackback(list(Female = mgus %>% filter(sex == "female") %>% filter(!is.na(alb
                        Male = mgus %>% filter(sex == "male") %>% filter(!is.na(alb)) %>
                   probability = TRUE, main = "Nivel de albúmina en diagnóstico")
# Colorear mitad izquierda y derecha del gráfico
barplot(-out$left, col="pink" , horiz=TRUE, space=0, add=TRUE, axes=FALSE)
barplot(out$right, col="blue", horiz=TRUE, space=0, add=TRUE, axes=FALSE)
```

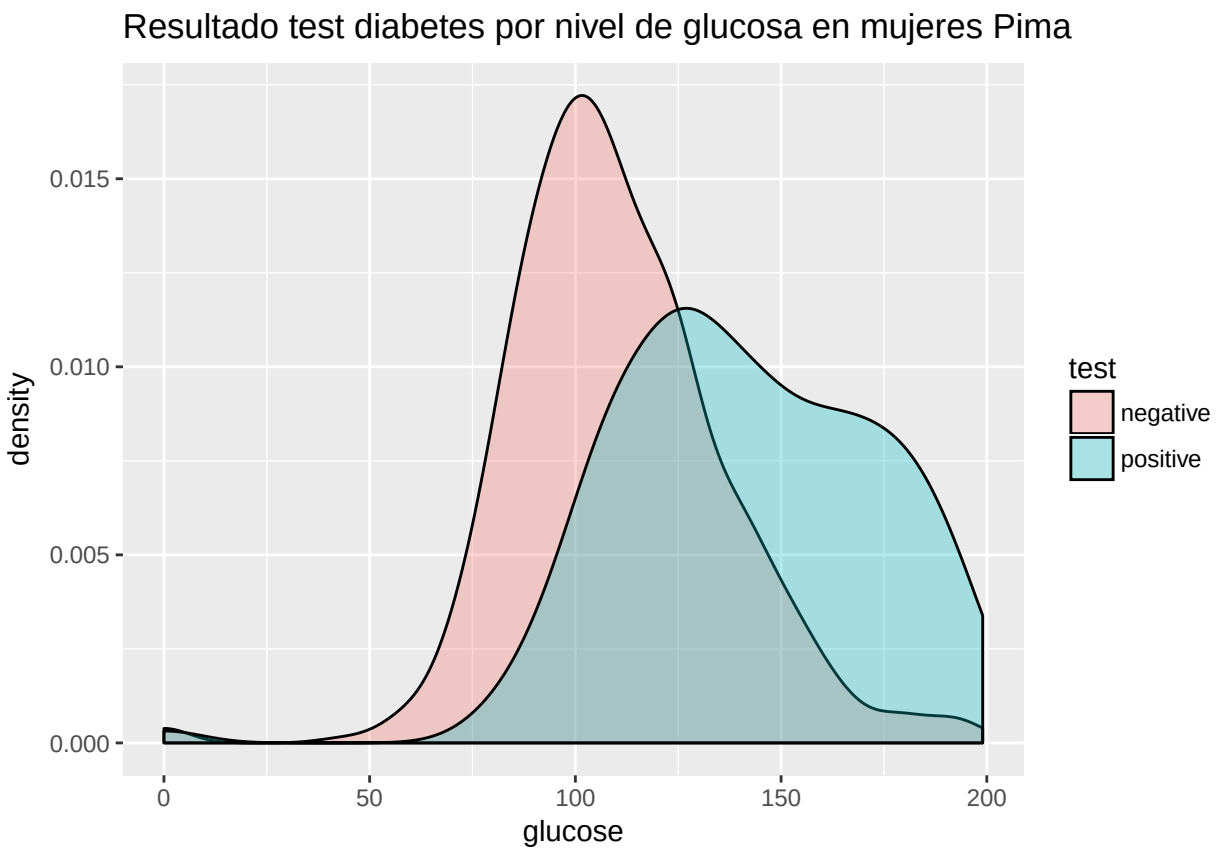


Por su diseño, los **diagramas de densidad de probabilidad** sí que permiten la representación simultánea de varias gráficas en un mismo panel. Si coloreamos el área bajo la curva, es conveniente usar cierto nivel de transparencia para mejorar la legibilidad del gráfico (argumento `alpha` en las funciones de geometría de `ggplot2`).

```
# Múltiples KDEs

pima %>%
  mutate(test = factor(test, labels = c('negative', 'positive'))) %>%
  ggplot(aes(x = glucose, colour = test)) +
  geom_density() +
  ggtitle('Resultado test diabetes por nivel de glucosa en mujeres Pima')
```

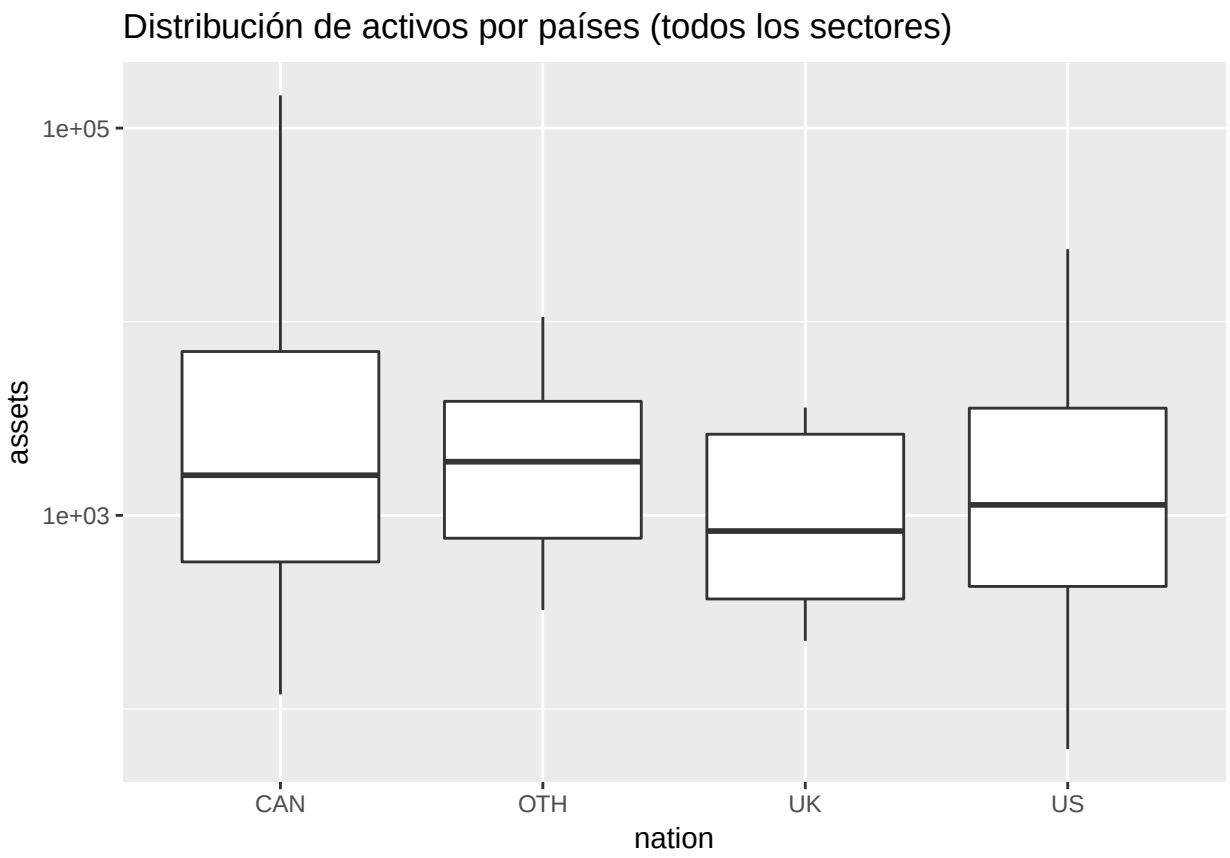



```
pima %>%
  mutate(test = factor(test, labels = c('negative', 'positive'))) %>%
  ggplot(aes(x = glucose, fill = test)) +
  geom_density(alpha = .3) +
  ggtitle('Resultado test diabetes por nivel de glucosa en mujeres Pima')
```

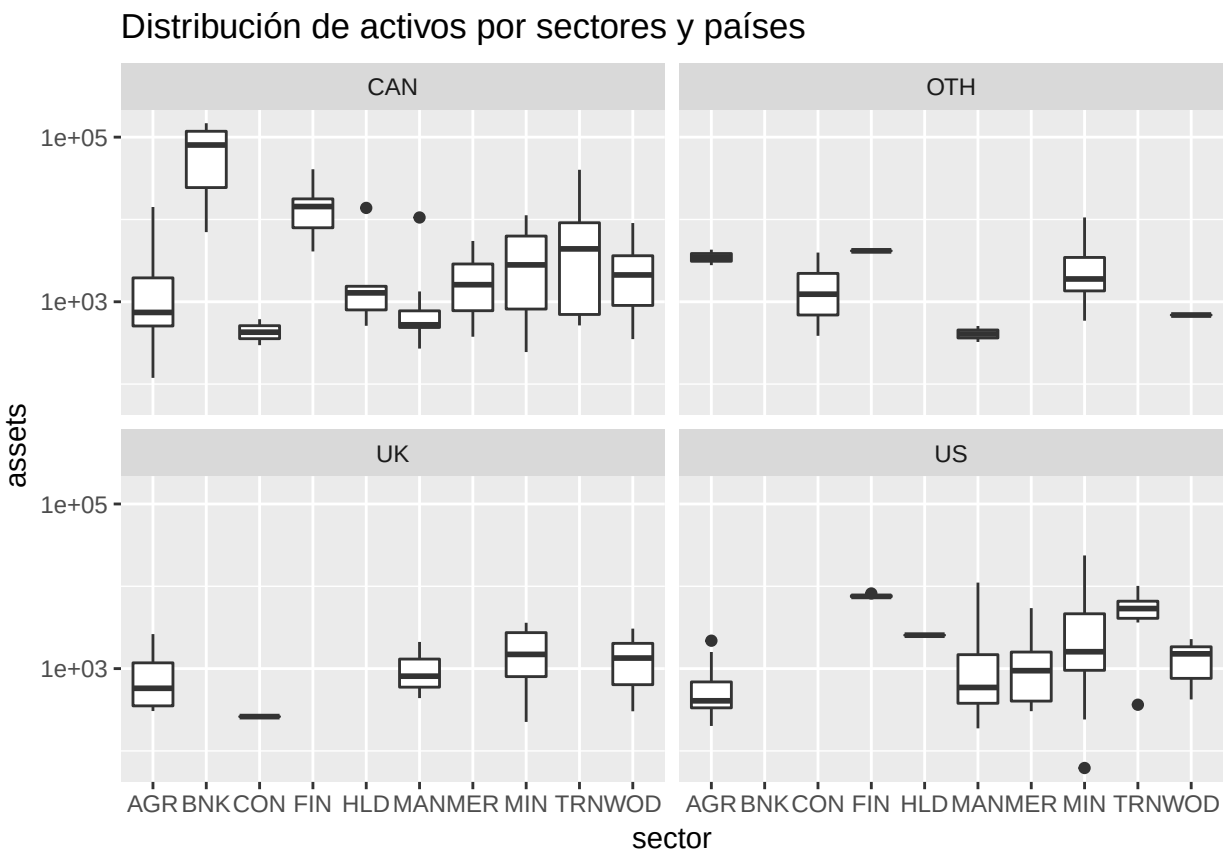


Por su diseño, el **boxplot** también resulta muy práctico para realizar comparativas entre variables cuantitativas y cualitativas. Puesto que uno de sus objetivos principales es ayudar a identificar atípicos, permite una comparación rápida de la distribución de valores entre varias categorías y cuántos posibles atípicos se identifican en cada una.

```
# Múltiples boxplots
# Transformamos la variable `assets` con log10 para tener una distribución
# de valores más simétrica
ggplot(Ornstein, aes(x = nation, y = assets)) + geom_boxplot() +
  scale_y_log10() + ggtitle('Distribución de activos por países (todos los sectores)')
```



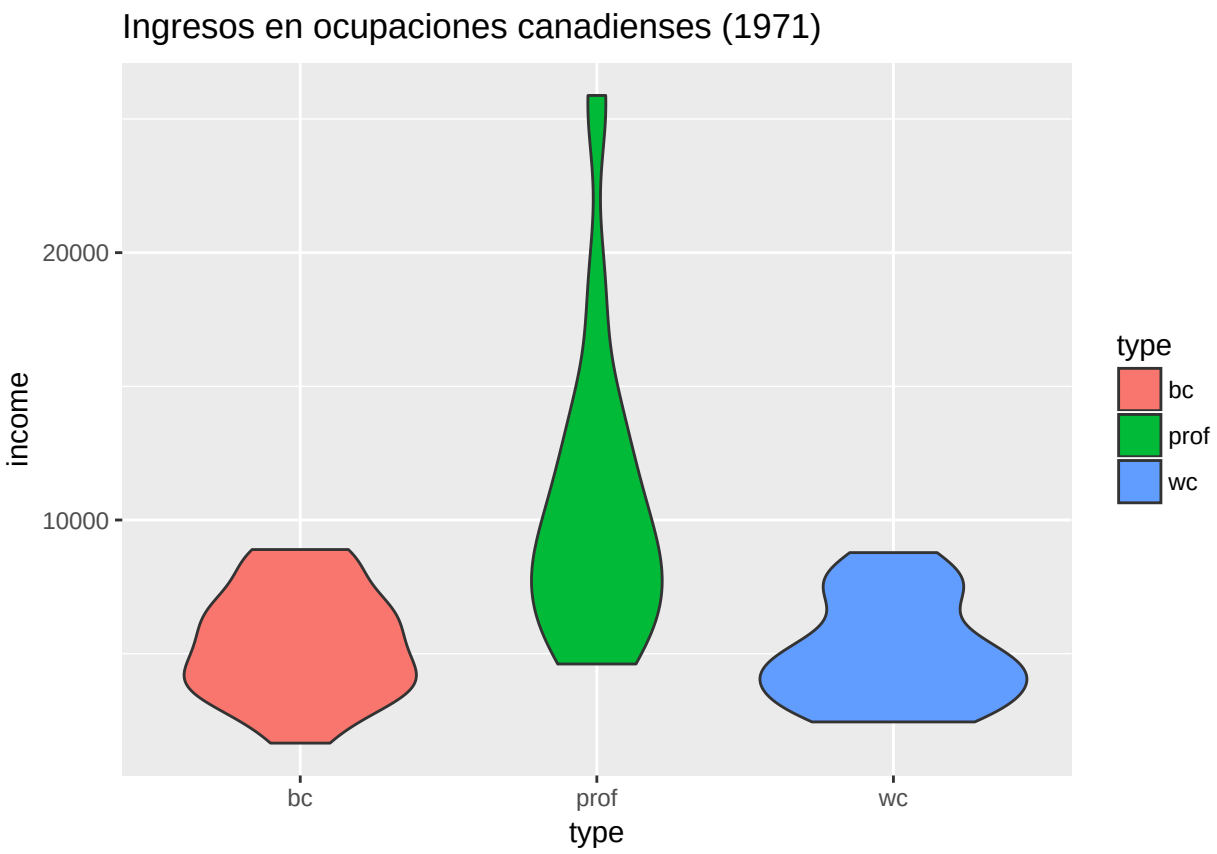

```
ggplot(Ornstein, aes(x = sector, y = assets)) + geom_boxplot() +
  facet_wrap(~ nation)+ scale_y_log10() +
  ggtitle('Distribución de activos por sectores y países')
```



Una variante un poco más sofisticada del boxplot es el **diagrama de violín**, donde se representa la función de densidad de probabilidad (KDE) sobre el gráfico que correspondería al boxplot de cada categoría. Para facilitar la comparación, en realidad se dibuja para cada categoría el KDE y su imagen simétrica.

```
# Gráfico de violín
pima %>%
  mutate(test = factor(test, labels = c('negative', 'positive'))) %>%
  ggplot(aes(x = test, y = glucose)) +
  geom_violin(aes(fill = test)) +
  ggtitle('Resultado test diabetes por nivel de glucosa en mujeres Pima')
```


```
ggplot(subset(Prestige, !is.na(type)), aes(x = type, y = income)) +
  geom_violin(aes(fill = type)) +
  ggtitle('Ingresos en ocupaciones canadienses (1971)')
```

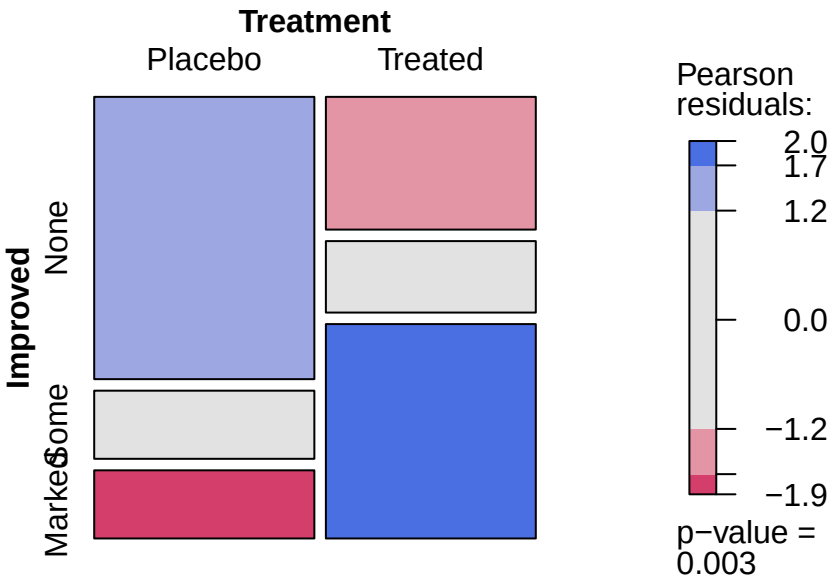


Gráficos específicos para variables categóricas

El **diagrama de mosaico** (en inglés *mosaicplot*) permite representar gráficamente una tabla de contigencias para dos o más variables categóricas. Adicionalmente, también se puede incluir información en el gráfico sobre modelos estadísticos que analicen las relaciones entre los diferentes grupos, como en el ejemplo siguiente. Podemos encontrar una introducción completa y muchos más ejemplos en este manual del paquete `vcdExtra`.

```
# Mosaicplot (vcd)
# Instalar paquete `vcd`, si es preciso
library(vcd)
art <- xtabs(~ Treatment + Improved, data = Arthritis)
mosaic(art, gp = shading_max, split_vertical = TRUE,
       main="Arthritis: [Treatment] [Improved]")
```

Arthritis: [Treatment] [Improved]



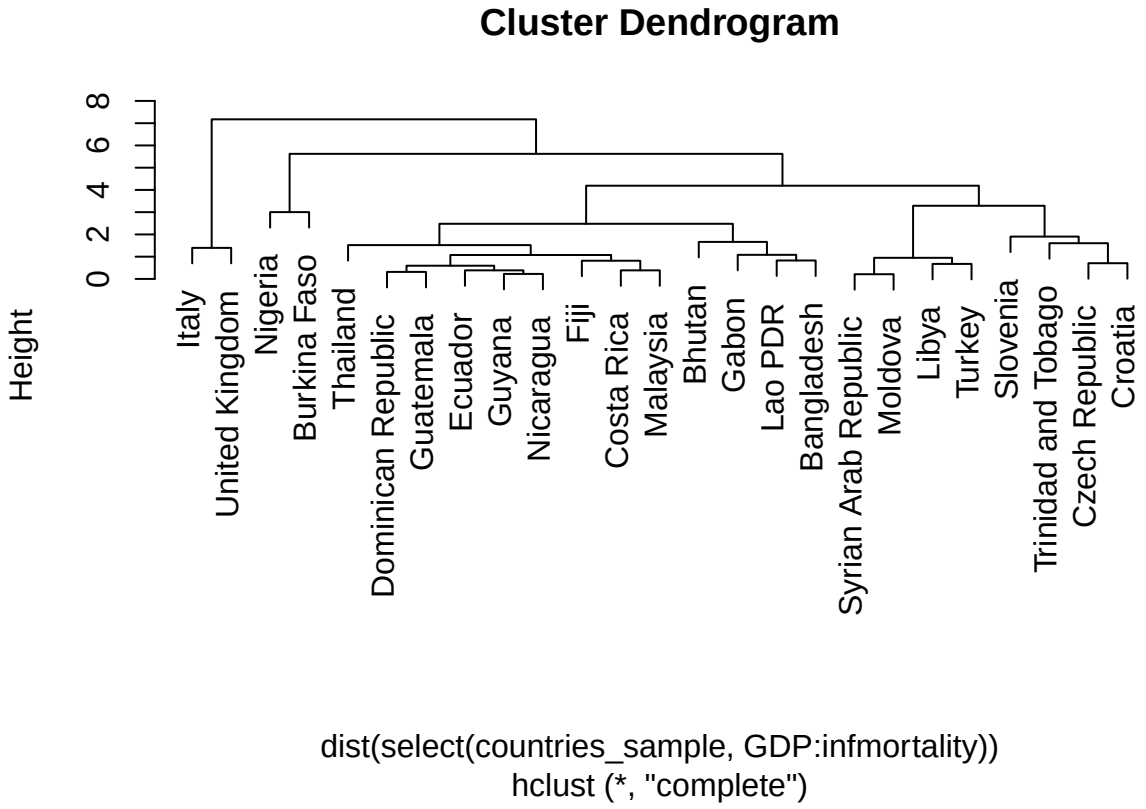

```
summary(art)
```

```
## Call: xtabs(formula = ~Treatment + Improved, data = Arthritis)
## Number of cases in table: 84
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 13.055, df = 2, p-value = 0.001463
```

Otra variante muy útil es el **dendograma**, que suele representarse como producto de la aplicación previa de un algoritmo de clasificación o *clustering*. El objetivo de este gráfico es hacer más evidente las similitudes entre diferentes casos que comparten características semejantes de las variables que los describen.

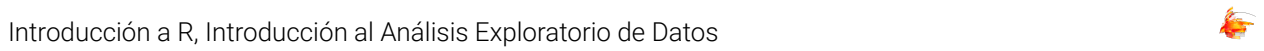
```
# Dendograma (clustering)
# Ejemplo de R Graphics Cookbook (O'Reilly, 2012) actualizado para utilizar
# `dplyr` y `tidyr`
# Instalar biblioteca `gcookbook` si fuese necesario
library(gcookbook)
# Dataset: `gcookbook::countries` datos económicos y de salud del
# Banco Mundial (1960-2010)
# Filtramos muestra de 25 casos completos (sin NAs) del año 2009
countries_sample <-
  countries %>%
    filter(Year == 2009) %>%
    filter(complete.cases(.)) %>%
```

```
    sample_n(25) %>%
    select(-Code, -Year) %>%
    mutate(GDP = scale(GDP), laborrate = scale(laborrate),
           healthexp = scale(healthexp), infmortality = scale(infmortality) )

hc <- hclust(dist(select(countries_sample, GDP:infmortality)))
# Mostrar el dendrograma
plot(hc, labels = countries_sample$Name)
```

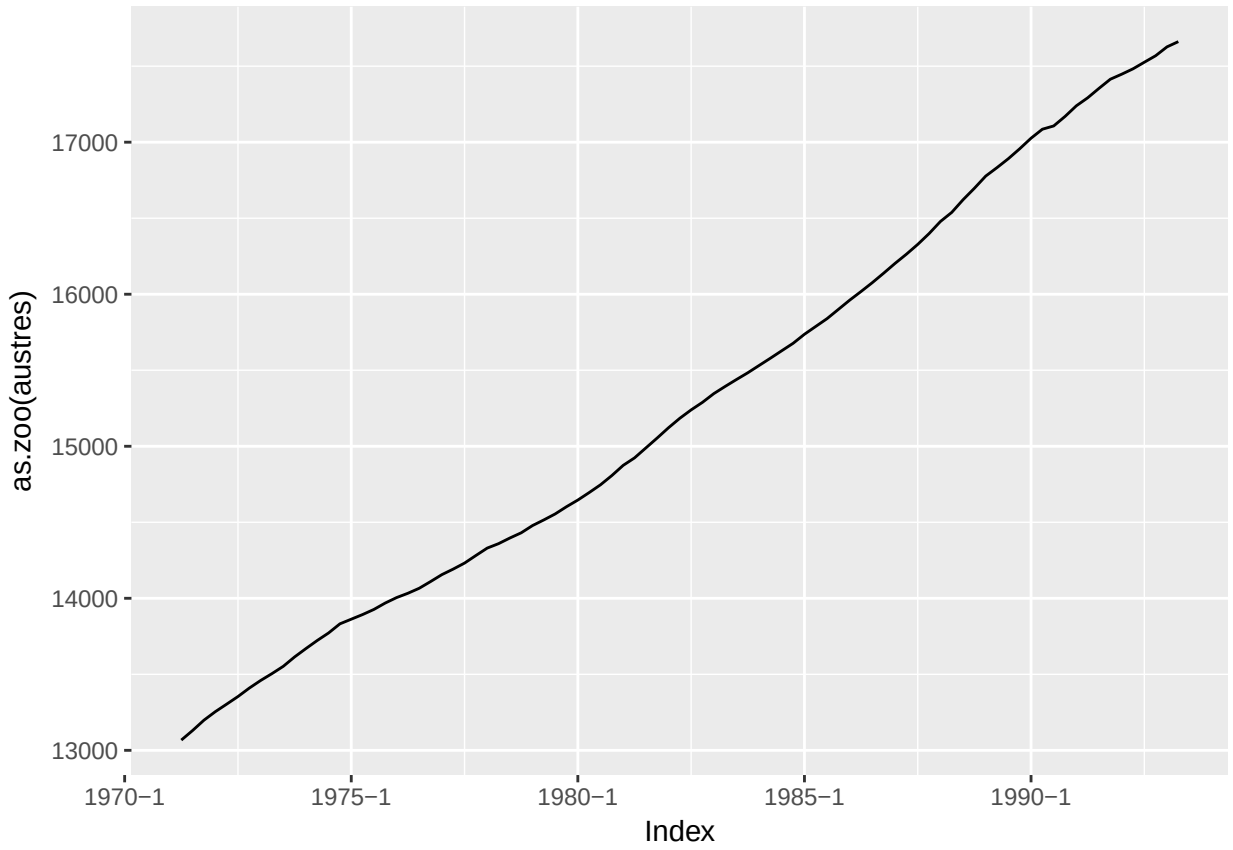


Representación gráfica de series temporales

Las **series temporales** merecen un capítulo aparte, puesto que los datos representados a lo largo del tiempo suelen tener características singulares, muy distintas a los datos de sección cruzada. Normalmente, se suelen representar con el tiempo en el eje horizontal y la variable medida en el eje vertical. La gráfica se suele repesentar mediante una línea que une las diferentes muestras tomadas en cada instante, haciendo así énfasis en la dependencia que existe entre el dato tomado en un instante y el siguiente. Aún así, existen muchas variantes y estilos de representación distintos. Además, también se pueden representar varias series sobrepuestas en el mismo gráfico, o en diferentes paneles (para cada categoría de una variable cualitativa). Con series sobrepuestas en el mismo gráfico buscamos facilitar la detección de comportamientos distintos en algún intervalo temporal. La representación por paneles favorece la comparación de formas y tendencias de las series completas.

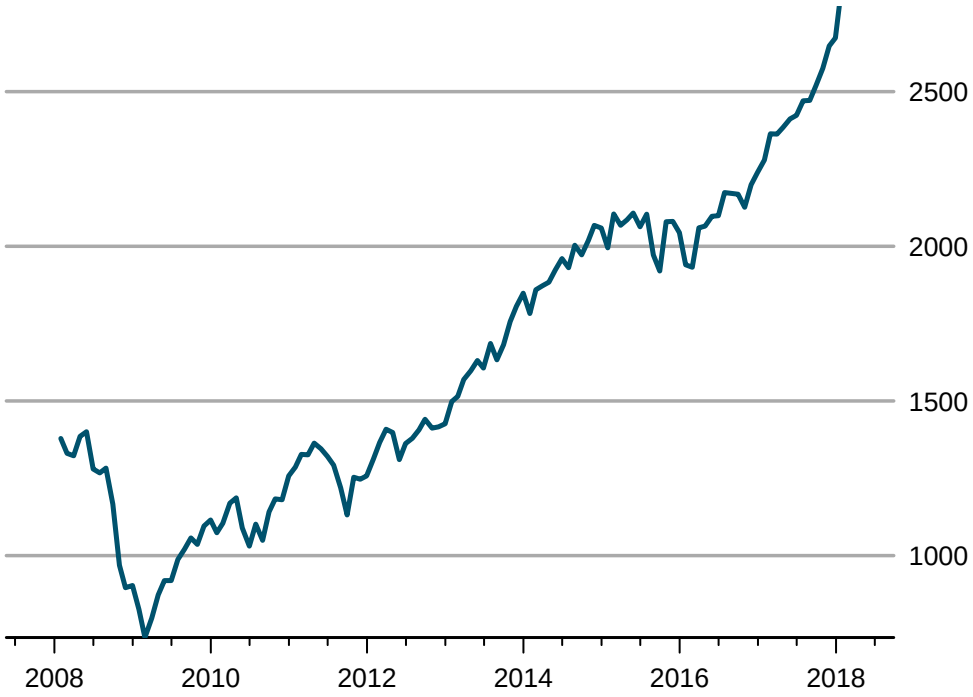


```r
# Diagrama de líneas (por paneles y colores)
# Normalmente se usan con temporales (ts, zoo, etc.)
# Instalar las bibliotecas `zoo`, `xts`, `quantmod`, `latticeExtra` si es necesario
library(scales) # si error "no se pudo encontrar la función 'pretty_breaks'"
library(zoo)
autoplot(as.zoo(austres), geom = "line") + scale_x_yearqtr()
```

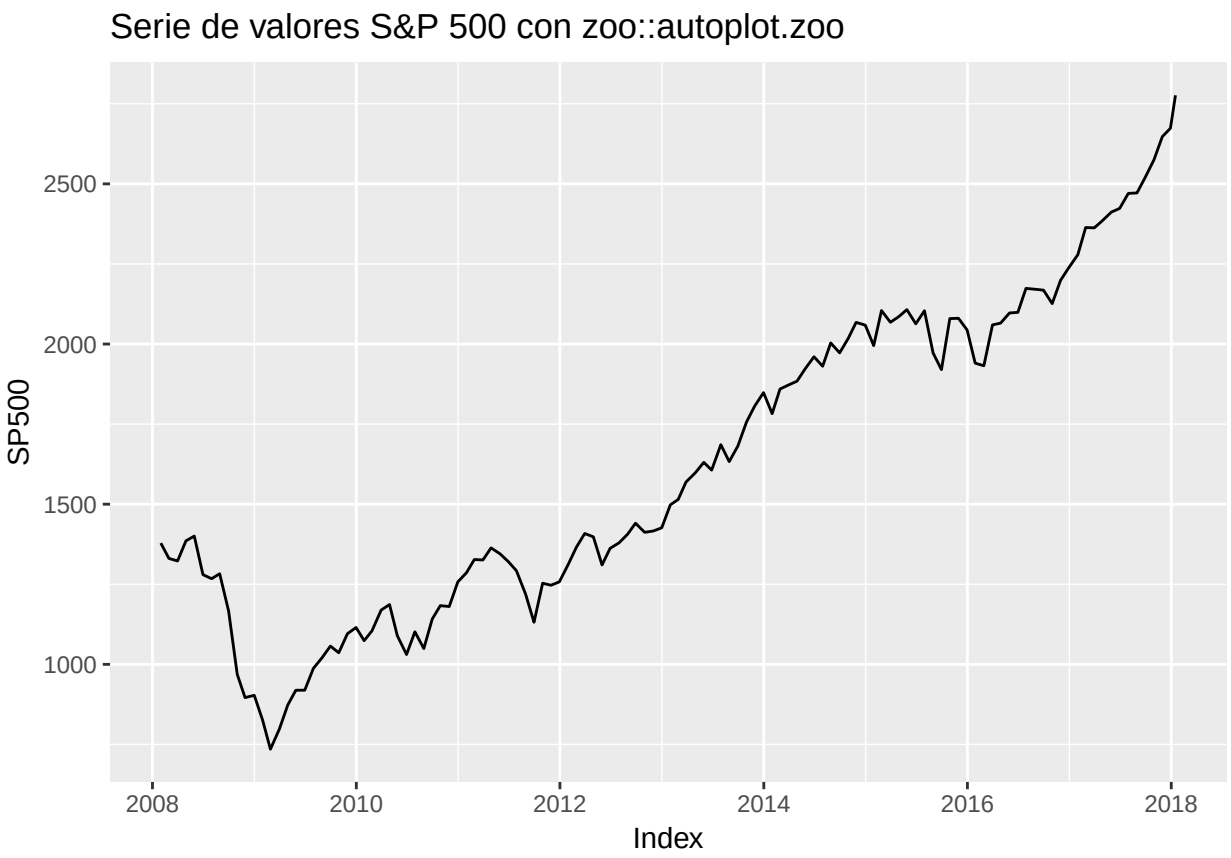



```r
library(quantmod)
library(latticeExtra)

## Obtener datos del S&P 500 desde FRED (St. Louis Fed)
sp500 <- na.omit(
  getSymbols(
    "SP500",
    src = "FRED",
    from = "2005-12-31",
    auto.assign = FALSE
  )
)
# Frecuencia de datos por meses
```

```
sp500.monthly <- sp500[endpoints(sp500, on ="months")]

## Gráfica estilo The Economist
asTheEconomist(
  xyplot(
    sp500.monthly,
    scales = list( y = list( rot = 0 ) ),
    main = "Serie de valores S&P 500 con lattice::xyplot.xts"
  )
)
```




```
## ggplot2
autoplot.zoo(sp500.monthly) +
  ggtitle("Serie de valores S&P 500 con zoo::autoplot.zoo")
```




```
## Muchos más ejemplos:
## https://timelyportfolio.github.io/rCharts_time_series/history.html
```

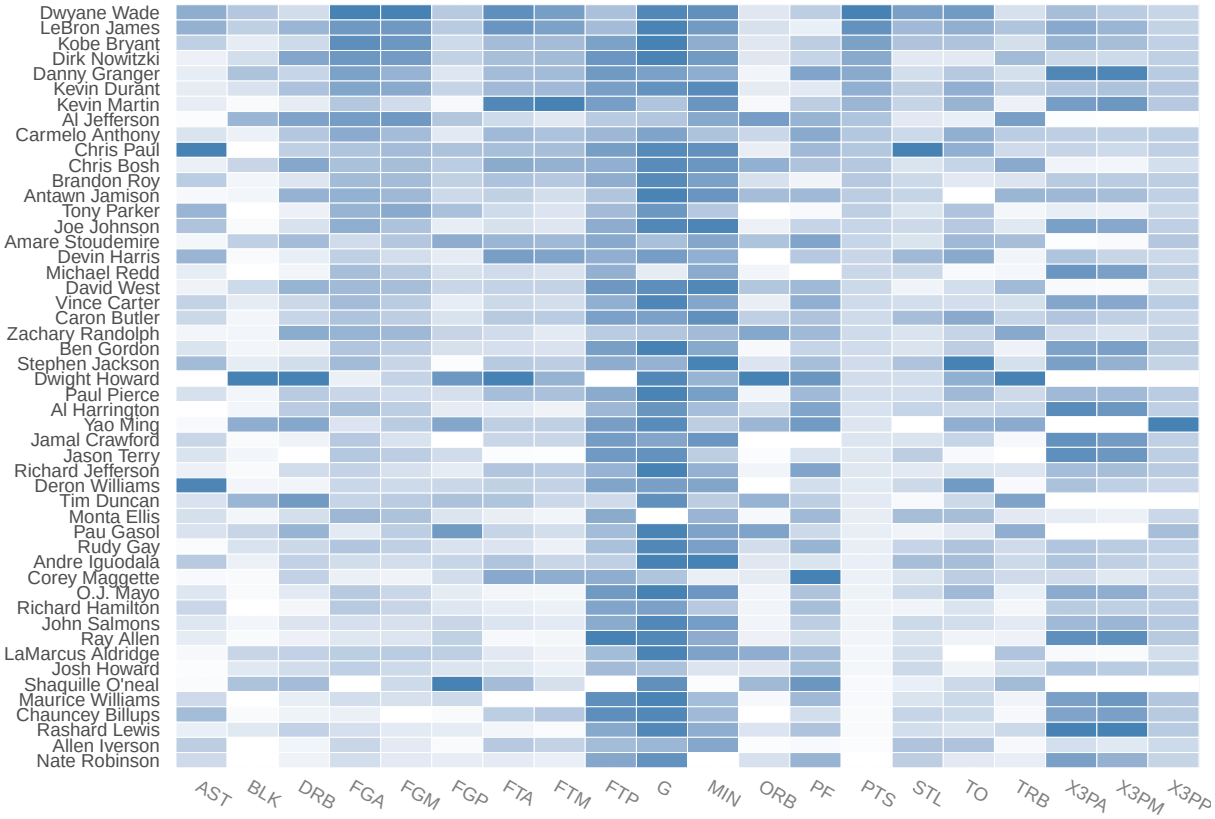
Gráficos avanzados

Existen muchos otros tipos de gráficos avanzados para representación y exploración de datos. Un buen ejemplo de la gran variedad disponible en R es el paquete `plotrix` que incluye una larga lista de diagramas y gráficos para cada ocasión.

Un tipo de gráfico avanzado muy útil es el **mapa de calor**. Permite representar mediante un gradiente de colores la magnitud que toma una variable cuantitativa en celdas de una malla 2D. Cada celda suele corresponder a la intersección de categorías de dos variables cualitativas, o de una variable cualitativa y una cuantitativa categorizada. El siguiente ejemplo muestra un caso de análisis de estadísticas de jugadores de la NBA, usando las nuevas bibliotecas `dplyr`y `tidyr`.

```
# Diagrama de calor (heatmap)
library(tidyr)
nba <- read.csv("http://datasets.flowingdata.com/ppg2008.csv")
base_size = 9
nba %>%
```

```
  mutate(Name = factor(Name, levels=nba[order(nba$PTS), "Name"])) %>%
  gather(variable, value, -Name) %>%
  group_by(variable) %>%
  mutate(rescale = rescale(value)) %>%
  ggplot(aes(variable, Name)) +
  geom_tile(aes(fill = rescale), colour = "white") +
  scale_fill_gradient(low = "white", high = "steelblue") +
  theme_grey(base_size = base_size) + labs(x = "", y = "") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  theme(legend.position = "none",
        axis.ticks = element_blank(),
        axis.text.x = element_text(size = base_size * 0.8, angle = 330,
                                   hjust = 0, colour = "grey50"))
```



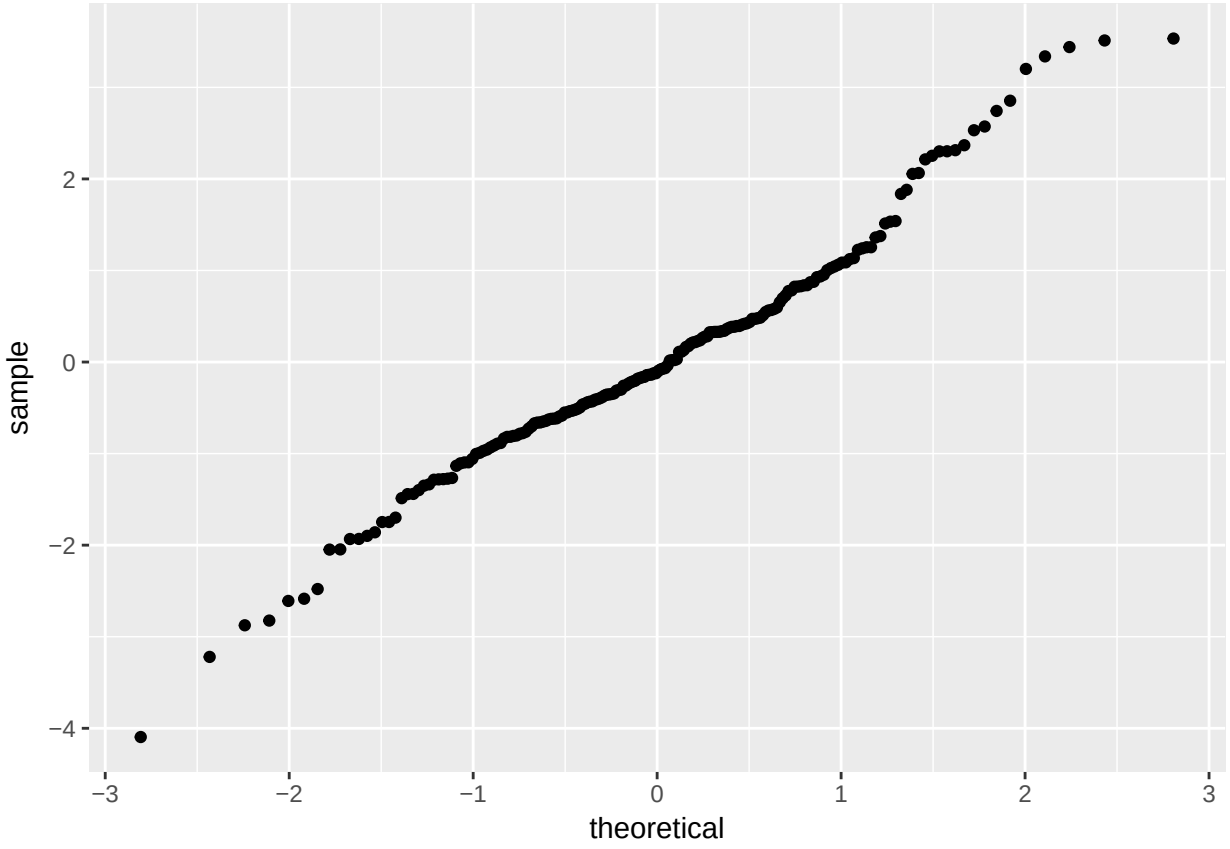
Otro diagrama avanzado que merece ser resaltado es el **diagrama cuantil-cuantil o gráfico Q-Q** (conocido en inglés como *Q-Q plot*). La idea es sencilla: representamos en el eje vertical los valores de los cuantiles de la distribución de nuestra variable; en el eje horizontal se representan los cuantiles de la distribución téorica contra la que queremos comparar nuestra variable empírica. El gráfico se dibuja como un diagrama de dispersión. Cuanto más se acerquen los puntos a la bisectriz del cuadrante (que se suele dibujar tambien como referencia), más se parece la distribución empírica de nuestra varaible a la distribución teórica y, por tanto, tendremos

argumentos (empíricos) para afirmar que nuestra variable sigue dicha distribución.

En el siguiente ejemplo, hacemos una prueba comparando una muestra aleatoria de 200 elementos de una distribución t de Student contra la distribución t de Student teórica. Como podemos ver, siempre existe cierta variabilidad (no sigue una línea recta perfecta), incluso con muestras extraídas exactamente del mismo tipo de distribución. Por este motivo, conviene antes entrenarnos examinando varios gráficos de este tipo antes de comparar la distribución empírica contra un tipo de distribución teórica candidata mediante el gráfico Q-Q .

```
y = rt(200, df = 5)
qplot(sample = y, stat="qq")
```

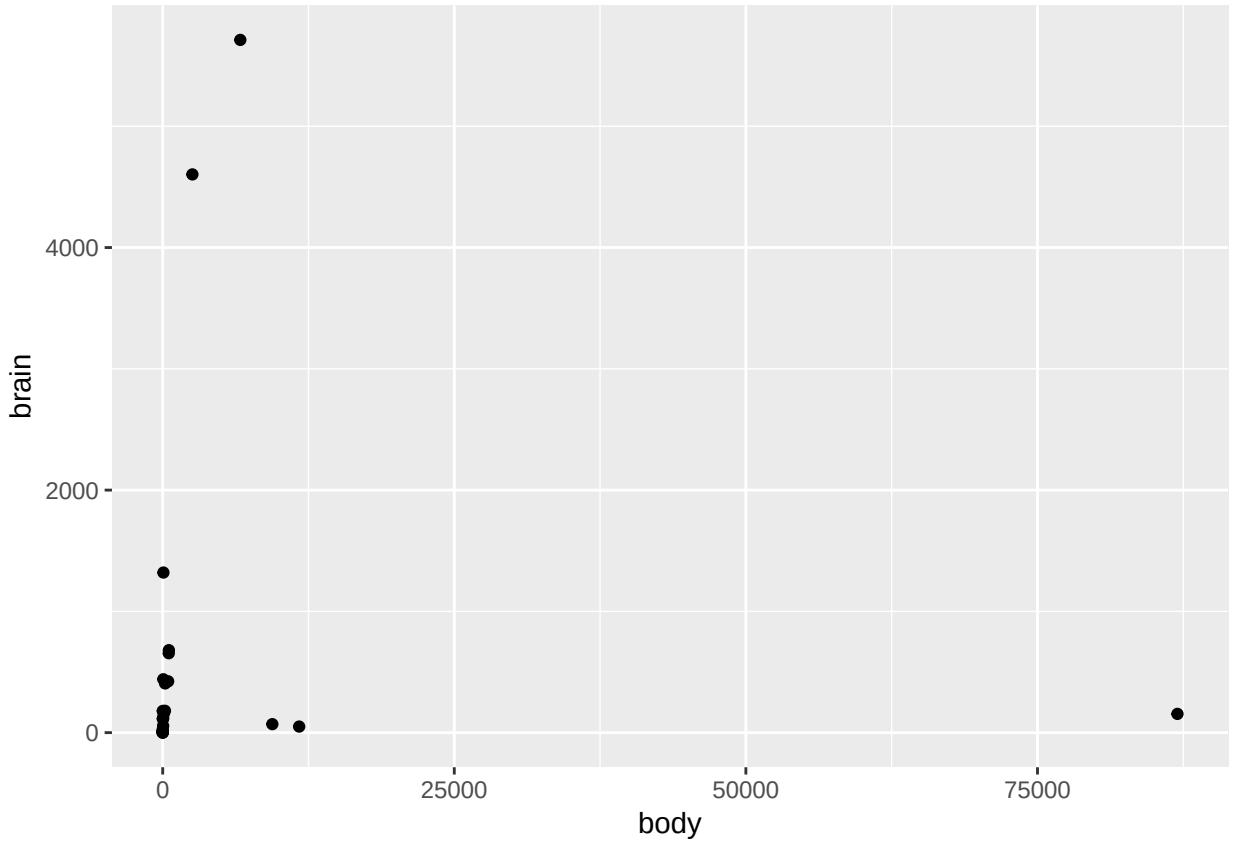
```
## Warning: `stat` is deprecated
```



Transformación de variables

Cuando realizamos un EDA, las variables que exploramos puede que no siempre tengan un rango de valores adecuado para su representación gráfica o para buscar relaciones entre varias variables. Como ejemplo, consideremos el siguiente ejemplo en el que representamos un diagrama de dispersión para explorar la relación entre el peso del cuerpo y el del cerebro de 28 animales terrestres:

```r
# Instalar paquete `MASS` si fuese necesario
library(MASS)
ggplot(Animals, aes(x = body, y = brain)) + geom_point()
```



Podemos detectar detalles interesantes, pero no mucho más. Por ejemplo, que hay especies con peso medio de su cerebro desproporcionadamente mayor que el de su cuerpo (esquina superior izquierda) y viceversa, algunas con cuerpos demasiado pesados en comparación con su cerebro (esquina inferior derecha). Si nos fijamos bien, tenemos un grave problema con la escala de la representación: hay demasiados puntos cerca del origen de coordenadas. La razón es que los valores atípicos que antes hemos mencionado están muy lejos de la mayoria de puntos, por lo que R tiene que elegir un rango de valores muy amplio en ambos ejes para poder incluir todos los puntos en el gráfico. Por contra, si hacemos zoom en alguna de las áreas para ver más detalle perdemos de vista los restantes puntos, por lo que no podríamos detectar una relación válida para todo el rango de valores de ambas variables.

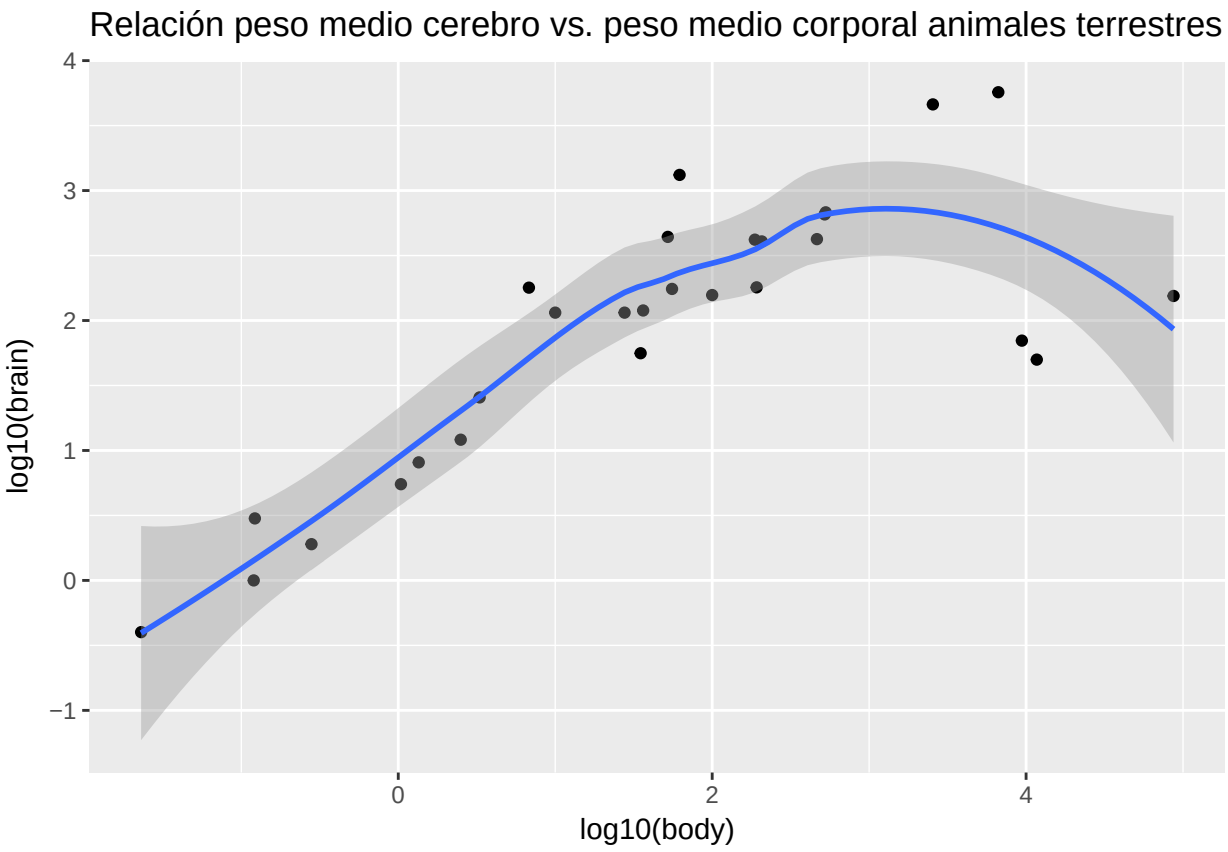
La solución es la **transformación de variables**. Aplicamos una función matemática (invertible, para luego recuperar si hace falta los valores originales) a los valores de la variable para transformar su rango y distribución de densidad a lo largo del mismo. Algunos objetivos suelen ser:

- Alejar altas concentraciones de valores cerca del cero, repartiéndolos de forma más distribuida por todo el rango (por ejemplo, lo que conseguimos aplicando logaritmos).
- Hacer la distribución de valores más simétrica.

- Conseguir rangos de valores comparables entre todas las variables (por ejemplo, con los llamados z-scores).

En nnuestro ejemplo, si aplicamos a en ambos ejes el `log10()` nos queda:

```
ggplot(Animals, aes(x = log10(body), y = log10(brain))) +
  geom_point() + geom_smooth(method = "loess") +
  ggtitle("Relación peso medio cerebro vs. peso medio corporal animales terrestres")
```



Las transformaciones suelen afectar a datos cuantitativos. En el caso de datos cualitativos, las únicas "transformaciones" posibles serían, por ejemplo, recodificar las variables, o subsumir varias categorías en una sola categoría más general.

Transformaciones para variables cuantitativas

En el caso de estar representando gráficos multivariantes como, por ejemplo, dos variables x e y, podemos elegir entre transformar una de las variables o las dos simultánemanete (usando la misma o distinta transformación). A continuación presentamos algunas de las transformaciones más usuales y útiles.

- **Logaritmo**:
- **Raíz cuadrada o cúbica**:
- **Inversa**:

- **Transformación general de potencias**: También llamada **transformación de Box-Cox**, ya que fue propuesta por Box y Cox (1964). Engloba a las anteriores mediante la siguiente fórmula general:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda}, \quad if\ \lambda \neq 0,$$
$$y(\lambda) = log(y), \quad if\ \lambda = 0.$$

Referencias

Datasets

Agresti, A. (2002). "Categorical Data Analysis"", 2nd Ed., New York: Wiley-Interscience, Table 9.1, p. 362.

G. Koch & S. Edwards (1988). "Clinical efficiency trials with categorical data". In K. E. Peace (ed.), Biopharmaceutical Statistics for Drug Development, 403–451. Marcel Dekker, New York.

L. Wilkinson (1999). "Dot plots". The American Statistician (American Statistical Association) 53 (3): 276–281. doi:10.2307/2686111. JSTOR 2686111.