Práctica 7: Gráficos de variables estadísticas

En esta práctica nuestro objetivo es realizar un análisis descriptivo y exploratorio básico de un conjunto de datos. Para ello utilizaremos las funciones gráficas de R que hemos visto en la primera parte del Tema 4, y alguna más que introduciremos en este guión. Verás que muchos de los gráficos que se muestran a continuación admiten mejoras relativas a la apariencia (colores, márgenes, títulos etc.). Para no desviar la atención, esta tarea la dejaremos en un segundo plano durante la sesión de prácticas, centrándonos en los aspectos más importantes. No obstante muchas de estas posibles mejoras pueden ser un buen ejercicio para conocer en más profundidad las funciones. En este sentido se dejan como tareas propuestas.

1. Análisis de datos de empleados

Comenzamos analizando los datos almacenados en el fichero *Employee.txt* en PRADO. Se trata de un conjunto de datos hipotético que corresponden a 473 empleados/as de un banco para los que se han recogido las siguientes variables:

Variables	Descripción
salary	Salario actual (dólares)
age	Edad
edu	Nivel educativo (años)
startsal	Salario inicial (dólares)
jobtime	Número de meses desde que fue contratado/a
prevexp	Experiencia previa anterior al contrato (meses)
minority	Clasificación étnica (*min*, *no_ min*)
gender	Sexo (*f*, *m*)
jobcat	Categoría laboral (*clerical*, *custodial*, *manager*)

Previa inspección del fichero (desde por ejemplo el bloc de notas), carga los datos en R. Hazlo de modo que se almacenen en un objeto de tipo data frame, con nombre `employee`, en el que las variables *gender*, *jobcat* y *minority* se consideren como factores. Una vez almacenados renombra los niveles del factor `gender` como `female` y `male`.

Para simplificar el código que aparece a continuación coloca el data frame que has creado en la lista de búsqueda de R escribiendo:

```
> attach(employee)
```

recuerda que con esto ya podrás referirte en el código a las columnas del data frame directamente por su nombre (sin necesidad de escribir delante `employee$`).

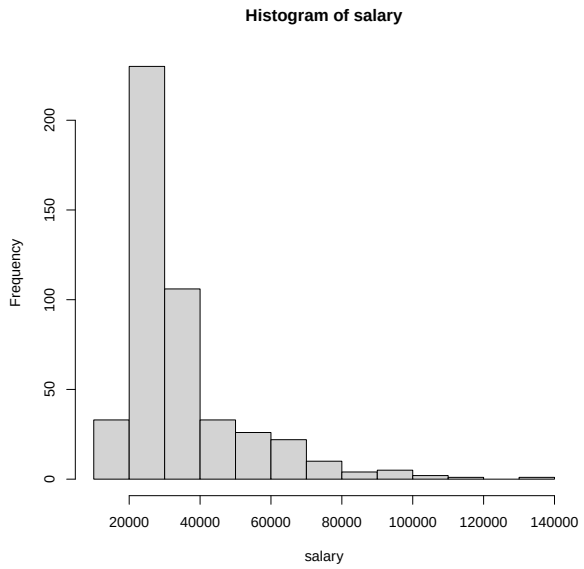
1.1. Variables cuantitativas: resúmenes y gráficos

En los datos podemos identificar como variables cuantitativas: `salary`, `startsal`, `age`, `jobtime`, `prevexp` y `edu`[1]. Por su naturaleza podríamos considerar las dos primeras como continuas y resto como discretas. No obstante hemos de ser conscientes de que todas se han medido de una forma dicretizada (números enteros), hecho que tendremos que tener en cuenta a la hora de leer e interpretar los resultados.

Variable `salary`

Comenzamos analizando la variable `salary`. Un histograma es quizá la forma más básica de representar datos de tipo continuo. La función `hist` nos permite obtener este tipo de gráfico. Su uso básico en este caso sería:

```
> hist(salary)
```



Esto nos muestra un histograma con las opciones por defecto de la función. Entre ellas vemos que se definen 13 intervalos para `salary`, todos de la misma amplitud, que se muestran en el eje de horizontal, y asociado a cada intervalo se representa un rectángulo con base definida por los límites del intervalo y altura la frecuencia absoluta correspondiente al intervalo. Los valores usados para el gráfico se pueden visualizar asignando el resultado de la función a un objeto e inspeccionando su contenido:

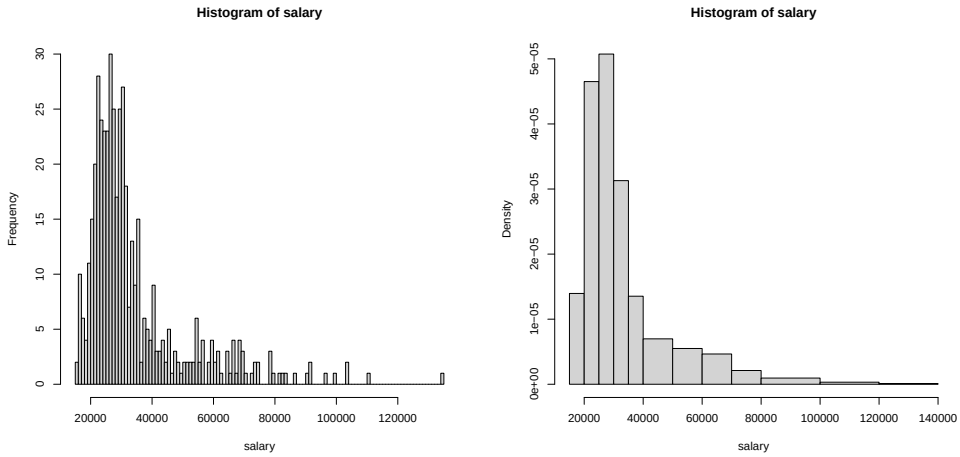
```
> res<-hist(salary,plot=FALSE)
> res
```

[1] Esta última también podría considerarse como un factor ordinal.

```
$breaks
 [1]  10000  20000  30000  40000  50000  60000  70000  80000  90000 100000 110000 12000
[13] 130000 140000

$counts
 [1]  33 230 106  33  26  22  10   4   5   2   1   0   1

$density
 [1] 6.976744e-06 4.862579e-05 2.241015e-05 6.976744e-06 5.496829e-06 4.651163e-06
 [7] 2.114165e-06 8.456660e-07 1.057082e-06 4.228330e-07 2.114165e-07 0.000000e+00
[13] 2.114165e-07

$mids
 [1]  15000  25000  35000  45000  55000  65000  75000  85000  95000 105000 115000 12500
[13] 135000

$xname
[1] "salary"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

A continuación vamos a hacer dos modificaciones de este histograma. La primera incrementando el número de intervalos a 100 y la segunda definiendo intervalos de distinta amplitud (más pequeña para los primeros, algo más grandes para los centrales y mayor para los últimos). Para ello modificamos el argumento `breaks`:

```
> hist(salary,breaks=100)
> # puntos de corte para intervalos con distinta amplitud
> x1 <-seq(15000,40000,by=5000)
> x2 <-seq(50000,80000,by=10000)
> x3 <-seq(100000,140000,by=20000)
> hist(salary,breaks=c(x1,x2,x3))
```



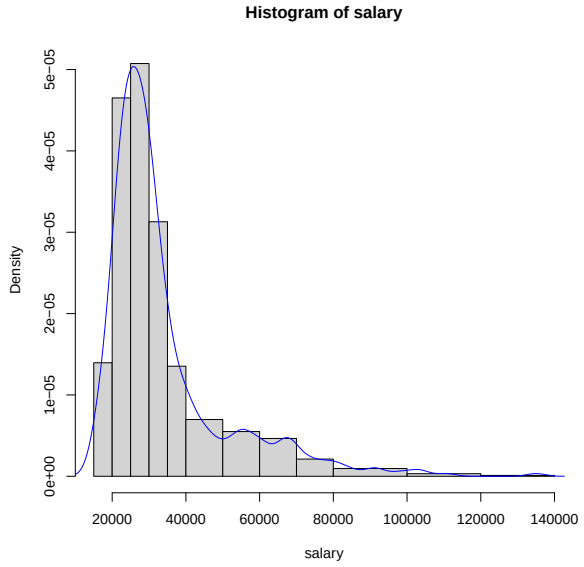
Observa que cuando construimos el gráfico con intervalos de distinta amplitud se representa la densidad de frecuencia (frecuencia dividida por amplitud) en lugar de la frecuencia absoluta. En el caso de amplitudes iguales también se podría construir el histograma representando densidades de frecuencia[2], si bien el gráfico sería equivalente al de frecuencias absolutas.

Un histograma nos da una representación de la distribución de frecuencias y consiste además en un estimador (no paramétrico) de la función de densidad de la variable. Una versión suavizada de este estimador se puede obtener usando la función `density`[3], que podemos superponer al histograma usando la función `lines`:

```
> hist(salary,breaks=c(x1,x2,x3))
> lines(density(salary),col='blue')
```

[2]Para ello solo tendrías que incluir el argumento `freq=FALSE`.

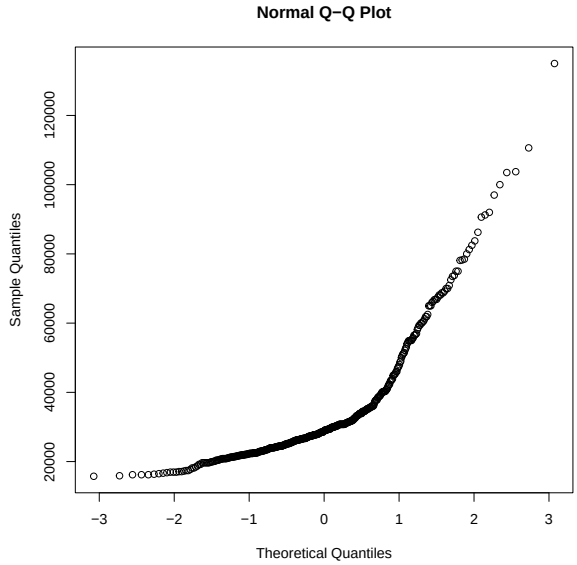
[3]Esta función calcula lo que se conoce como estimador tipo núcleo de la densidad (*kernel density estimator*). De este tipo de estimadores nos ocuparemos en el Tema 7.



Sobre gráfico anterior dibuja también la función de densidad de una Normal cuya media y desviación típica sean las de los datos de `salary`. ¿Te parece que este podría ser un buen modelo de probabilidad para describir estos datos?

El modelo de probabilidad Normal se asume como hipótesis en gran parte de los métodos de la Estadística clásica. Los histogramas y los gráficos de densidad que hemos construido antes nos dan una primera idea de la validez de dicha hipótesis. No obstante un gráfico más adecuado es el denominado gráfico probabilístico normal, que es un caso particular de los gráficos cuantil-cuantil (QQ-plot). Este tipo de gráfico lo podemos obtener con la función `qqnorm` y representa los cuantiles de la muestra de datos frente a los cuantiles de la distribución Normal. Para los datos de salarios sería:

```
> qqnorm(salary)
```



Si la distribución Normal fuera el modelo de probabilidad desde que se generaron los datos entonces deberíamos observar que los puntos están aproximadamente alineados[4]. Este no parece ser el caso de nuestros datos sino que los puntos muestran una forma claramente curvilínea. Para confirmar esto podemos plantear un contraste de hipótesis del tipo: H_0 :la variable salario sigue una distribución Normal; H_1 :sigue otro tipo de distribución. Algunas posibles soluciones a este problema nos las dan los contrastes de Kolmogorov-Smirnov y Shapiro-Wilks, que podemos obtener usando la función `ks.norm` y `shapiro.test`, respectivamente, como sigue:

```
> ks.test(salary,pnorm,mean=mean(salary),sd=sd(salary))

Warning in ks.test(salary, pnorm, mean = mean(salary), sd = sd(salary)): ties
should not be present for the Kolmogorov-Smirnov test


One-sample Kolmogorov-Smirnov test

data:  salary
D = 0.20857, p-value < 2.2e-16
alternative hypothesis: two-sided

> shapiro.test(salary)


Shapiro-Wilk normality test
```

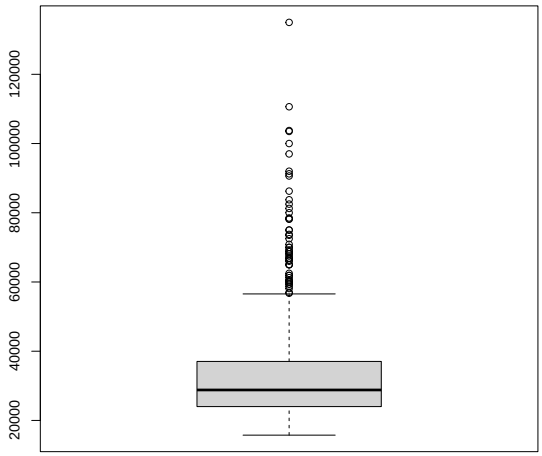
[4]Usando la función `qqline`, esto es, escribiendo `qqline(salary)`, puedes superponer una línea al gráfico para una mejor apreciación visual de la posible forma lineal.

```
data:  salary
W = 0.77045, p-value < 2.2e-16
```

El primer test nos muestra una advertencia indicando que en `salary` hay *ties*, esto es, valores repetidos. Esto no tiene mucho sentido si la distribución es continua, no obstante en nuestro caso está asociado a que las observaciones se han recogido discretizadas. Ignorando esta advertencia, el resultado de ambos test es el mismo: se rechaza H_0. Esto lo podemos concluir observando el p-valor (`p-value`) resultante, que en ambos casos es muy pequeño, por debajo de `2.2e-16`. Por tanto no sería adecuado modelizar estos datos con una Normal.[5]

Otro gráfico muy habitual y útil para explorar la distribución de variables de tipo continuo son los comúnmente denominados diagramas de "cajas con bigotes" (*boxplots*). La función `boxplot` nos permite construir este tipo de gráficos. De nuevo comenzamos haciendo un uso básico de la función:

```
> boxplot(salary)
```



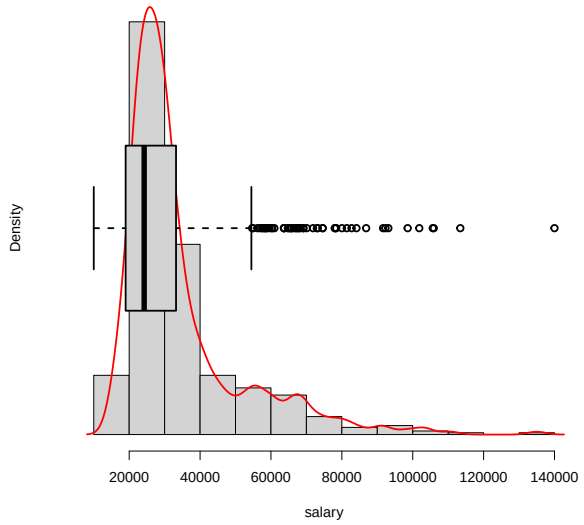
La caja se extiende desde el primer hasta el tercer cuartil y la línea que divide la caja se sitúa en la mediana. Los "bigotes" representan el rango de la variable, salvo que haya datos anómalos (*outliers*), como ocurre en estos datos, en cuyo caso estos se representan (por defecto) como puntos. Podemos comprobar los valores mostrados en el gráfico imprimiendo un resumen numérico de la variable:

[5]Esto es bastante habitual con datos correspondientes a salarios, es común observar distribuciones que, a diferencia de la Normal, son asimétricas con una cola pronunciada a la derecha asociada a que hay unos pocos individuos que reciben un salario notablemente mayor que el resto. Para representar este tipo de distribuciones pueden ser más adecuadas densidades de tipo log-Normal o Gamma.

```
> summary(salary)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15750   24000   28800   34418   37050  135000
```

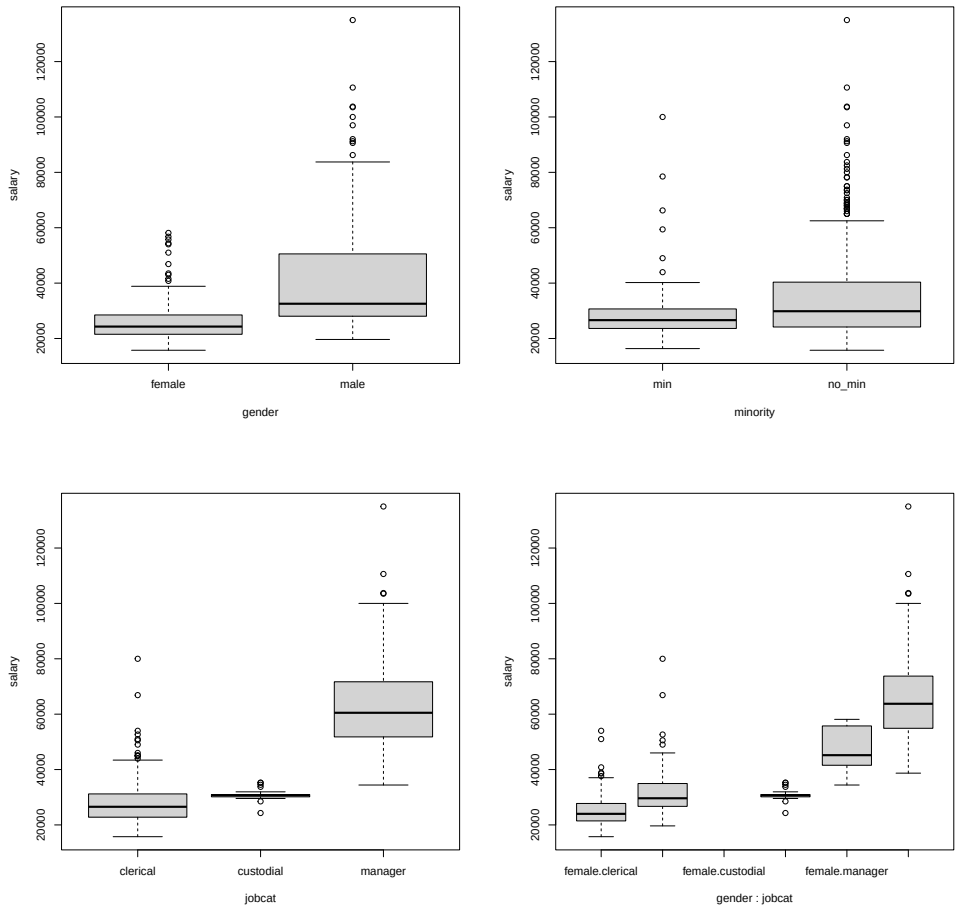
Observando un gráfico de cajas también podemos identificar posibles desviaciones respecto a un modelo Normal. Por ejemplo podemos construir el siguiente gráfico que superpone histograma, densidad suavizada y gráfico de cajas:

```
> hist(salary,probability=TRUE,main="",axes=FALSE)
> axis(1)
> lines(density(salary),col='red',lwd=2)
> par(new=TRUE)  ## Para que el próximo gráfico se superponga al anterior
> boxplot(salary,horizontal=TRUE, axes=FALSE,lwd=2)
```



Por otro lado podemos utilizar un gráfico de cajas para comparar la distribución del salario entre hombres y mujeres, así como entre los que corresponden o no a una minoría étnica. También en relación a la categoría profesional. Podemos obtener algunos gráficos adecuados para estos objetivos como sigue:

```
> boxplot(salary~gender)
> boxplot(salary~minority)
> boxplot(salary~jobcat)
> # A continuación salario con una doble clasificación
> boxplot(salary~gender*jobcat)
```



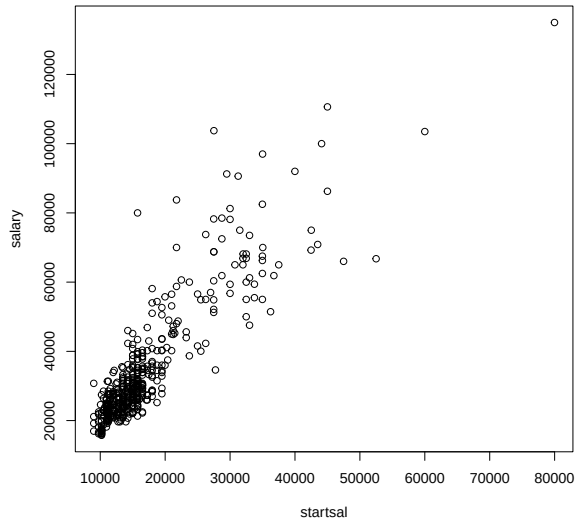
Observa la gran discrepancia entre los salarios de hombres y mujeres, incluso teniendo ambos la misma categoría profesional (como nos muestra el último gráfico donde se representa `salary` cruzando los niveles de `gender` y `jobcat`).

Realiza un análisis similar con las variables `startsal` y `age`.

Análisis conjunto de dos variables: `salary` y `startsal`

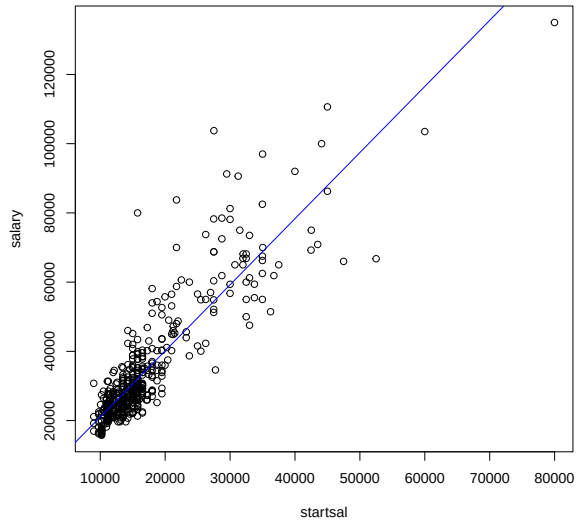
Un diagrama de dispersión permite visualizar la posible relación que existe entre dos variables cuantitativas. En nuestros datos podría pensarse en que esto ocurra por ejemplo con las variables salario y salario inicial. Para visualizar esta relación utilizamos la función `plot`:

```
> plot(startsal,salary)
```



El gráfico resultante nos muestra que parece existir una fuerte relación lineal entre ambas variables. Esta relación la podemos describir a través de un modelo de regresión lineal simple ($salary = \beta_0 + \beta_1 * startsal + \epsilon$). El ajuste del modelo a los datos lo podemos obtener usando la función `lm` y superponerlo al gráfico anterior usando la función `abline`:

```
> mod<-lm(salary~startsal)
> mod


Call:
lm(formula = salary ~ startsal)

Coefficients:
(Intercept)     startsal
    1929.52         1.91

> plot(startsal,salary)
> abline(mod,col='blue')
```



Observa que la pendiente de la recta β_1, que a partir de los datos se estima en 1.91 nos dice que un incremento de una unidad (1 dólar) en `startsal` supone un incremento en media de `salary` de aproximadamente 1.91 unidades[6].

Realiza un estudio similar que permita descubrir una posible relación entre: (i) las variables `salary` y `age`; y (ii) las variables `salary` y `edu`. ¿Qué puedes interpretar de los gráficos?

[6]El error de esta estimación lo podemos calcular a partir del resultado que nos da la función `lm`. Esto junto con un profundización en este tipo de modelos y la función `lm` lo dejamos para la próxima sesión.

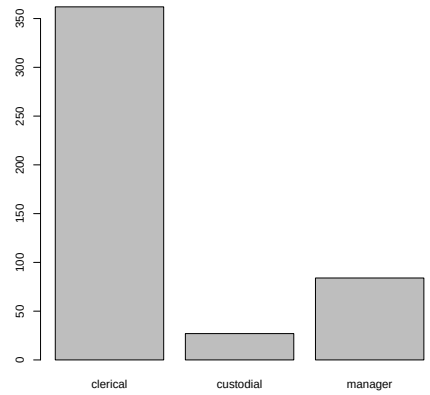
1.2. Variables cualitativas: tablas de frecuencias y gráficos

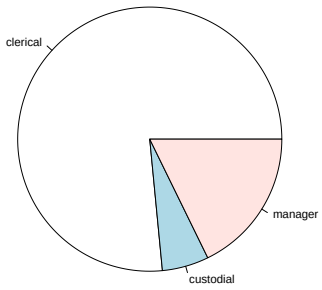
Para resumir y describir de forma numérica la información relativa a variables cualitativas de tipo factor (como son `gender`, `jobcat` y `minority`) podemos construir tablas de frecuencias (absolutas o relativas) unidimensionales. Para ello disponemos de las funciones `table` y `prop.table`, para frecuencias absolutas y relativas, respectivamente. Para la variable `jobcat` las obtenemos como sigue:

```
> tab<-table(jobcat)
> tab # frecuencias absolutas

jobcat
 clerical custodial   manager
      362        27        84

> tab.fi<-prop.table(tab)
> tab.fi # frecuencias relativas

jobcat
  clerical  custodial    manager
0.76532770 0.05708245 0.17758985

> # construimos una tabla clásica de frecuencias
> data.frame(tab,Freq.rel=as.numeric(tab.fi))

     jobcat Freq   Freq.rel
1  clerical  362 0.76532770
2 custodial   27 0.05708245
3   manager   84 0.17758985
```

Las funciones `barplot` y `pie` permiten realizar diagramas de barras y de sectores, respectivamente. Para la variable `catlab` escribiríamos:

```
> barplot(tab)
> pie(tab)
```





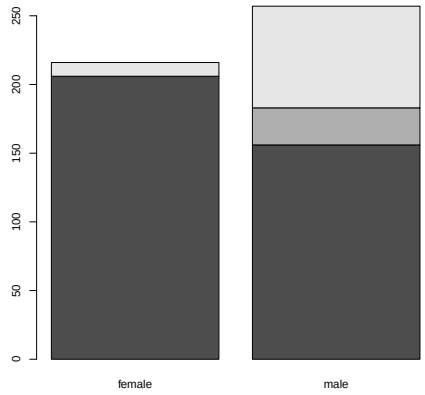
Realiza un análisis similar para los otros dos factores (`gender` y `minority`).

El estudio anterior también se puede hacer considerando conjuntamente dos (o incluso los tres factores). Por ejemplo podemos construir una tabla de frecuencias cruzada, o tabla de contingencia, que nos indique las frecuencias correspondientes a los factores `jobcat` y `gender` conjuntamente.

```
> tab2<-table(jobcat,gender)
> tab2

           gender
jobcat      female male
  clerical     206  156
  custodial      0   27
  manager       10   74

> # Y podemos añadir las sumas por filas y columnas
> addmargins(tab2)

           gender
jobcat      female male Sum
  clerical     206  156 362
  custodial      0   27  27
  manager       10   74  84
  Sum          216  257 473
```

Este tipo de tablas de contingencia se pueden visualizar por ejemplo utilizando diagramas de barras agrupadas o apiladas. Para ello usamos la función `barplot`. Su uso básico requiere proporcionar como argumento la tabla de contingencia (con los factores `jobcat` y `gender`) que hemos almacenado en el objeto `tab2`:

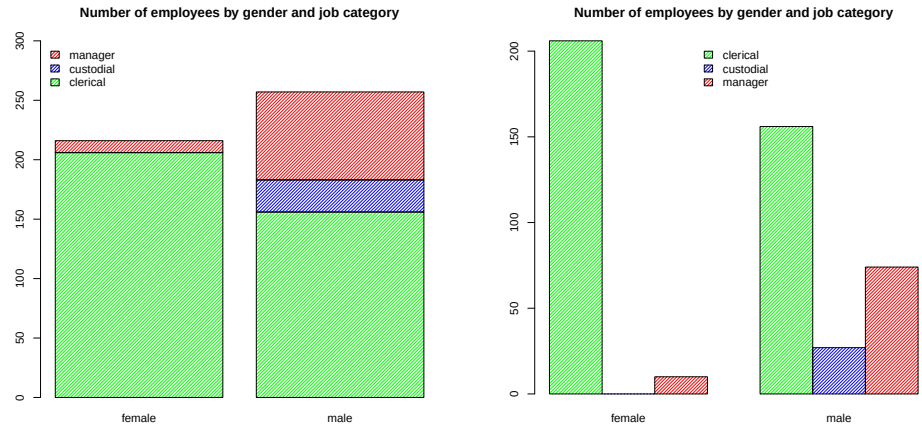
```
> barplot(tab2)
```



Observa que el gráfico resultante es un diagrama de barras apiladas. Muestra una barra para cada nivel del segundo factor de clasificación (`gender`), y a su vez cada barra se divide en porciones de acuerdo a los niveles del primer factor de clasificación (`jobcat`)[7]. Si bien el gráfico es correcto su apariencia admite muchas mejoras, una esencial es incluir alguna leyenda que permita identificar las porciones en las barras apiladas. A continuación incluimos este leyenda a la vez que realizamos algunas posibles mejoras:

```
> barplot(tab2,legend.text=TRUE, args.legend=list(x='topleft',bty='n'),
+         ylim=c(0,300), density=30,col=c('green','blue','red'),
+         main='Number of employees by gender and job category')
> barplot(tab2,legend.text=TRUE, args.legend=list(x='top',bty='n'),
+         density=30,col=c('green','blue','red'),
+         main='Number of employees by gender and job category',
+         beside=TRUE)
```

[7] El orden de los factores en el gráfico se puede cambiar si se desea pasando como argumento la tabla traspuesta, esto es, `barplot(t(tab2))`



Construye una tabla de contingencia que muestre la clasificación de los individuos según `jobcat` y `minority`. Representa un diagrama de barras que muestre dicha clasificación.

2. Ejercicio propuesto

El data frame `airquality` del paquete *datasets* contiene datos relativos a la calidad del aire en Nueva York. A partir de dichos datos escribe el código que permita resolver las siguientes tareas:

1. Construir un histograma del contaminante `Ozone` considerando intervalos de amplitud 10.

2. Superponer al histograma anterior la función de densidad de una distribución Normal cuyos parámetros media y desviación típica sean las de los datos `Ozone`. ¿Te parece que estos datos se podrían modelizar mediante ese modelo de probabilidad?

3. Construir un gráfico de normalidad para la variable `Ozone`. ¿Qué te indica el gráfico? Confirmar el resultado con un contraste de hipótesis.

4. Construir un diagrama de cajas del contaminante `Ozone`. ¿Qué puedes interpretar del gráfico?

5. Construir un diagrama de cajas múltiple del contaminante `Ozone` que permita comparar sus valores en los meses de mayo, junio, julio, agosto y septiembre. ¿Qué puedes interpretar del gráfico?

6. Construir dos diagramas de dispersión que nos permita visualizar la posible relación entre: (i) la velocidad del viento, `Wind`, y el contaminante `Ozone`; y (ii) la temperatura, `Temp`, y `Ozone`. ¿Qué puedes interpretar de los gráficos?