UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

MÉTODOS ESTADÍSTICOS
LECTURA 2 & LAB 1 Y 2
PRIMERA ENTREGA

AITOR VENTURA DELGADO

GRADO EN INGENIERÍA INFORMÁTICA

25 DE OCTUBRE DE 2019

ESCUELA DE INGENIERÍA INFORMÁTICA

ÍNDICE

LECTURA 2: CUESTIÓN 1

Los datos siguientes representan el peso en kilos de una muestra de 80 personas de la Escuela de Ingeniería Informática.

50	73	73	68	67	74	73	67	71	º79
74	74	77	74	71	80	72	74	77	75
71	73	75	76	77	71	81	68	66	73
91	75	89	77	93	57	66	83	86	90
55	77	78	91	82	83	87	96	85	88
101	97	80	73	76	80	89	76	78	99
80	85	84	72	65	69	79	84	92	83
86	76	80	81	74	73	72	79	55	66

a) Obténgase una distribución de datos en intervalos de amplitud del 5% de la distribución, construir una tabla de frecuencias absolutas y relativas y definir cada representante de la clase.

```
> x <- c(50,74,71,91,55,101,80,86,73,74,73,75,77,97,85,76,73,77,75,89,78,
+        80,84,80,68,74,76,77,91,73,72,81,67,71,77,93,82,76,65,74,74,80,
+        71,57,83,80,69,73,73,72,81,66,87,80,79,72,67,74,68,83,96,76,84,
+        79,71,77,66,86,85,78,92,55,79,75,73,90,88,99,83,66)
> library(fdth)
> dist <- fdt(x,breaks="Sturges")
> dist
 Class limits  f   rf rf(%) cf  cf(%)
  [49.5,56.1)  3 0.04  3.75  3   3.75
  [56.1,62.6)  1 0.01  1.25  4   5.00
  [62.6,69.2)  9 0.11 11.25 13  16.25
  [69.2,75.8) 23 0.29 28.75 36  45.00
  [75.8,82.3) 23 0.29 28.75 59  73.75
  [82.3,88.9) 11 0.14 13.75 70  87.50
  [88.9,95.4)  6 0.08  7.50 76  95.00
   [95.4,102)  4 0.05  5.00 80 100.00
```

Siendo *f* la frecuencia absoluta, *rf* la frecuencia relativa, *rf(%)* la frecuencia relativa porcentual, *cf* la frecuencia acumulada, y *cf(%)* la frecuencia acumulada porcentual.

b) Calcular la media muestral y la desviación estándar muestral.

Conocemos que la desviación estándar se define como la raíz de la varianza.

```
> mean(x)
[1] 77.225
> var(x)
[1] 92.10063
> desviacion <- sqrt(var(x))
> desviacion
[1] 9.596907
```

c) Encontrar la mediana, los cuartiles y el rango intercuartílico

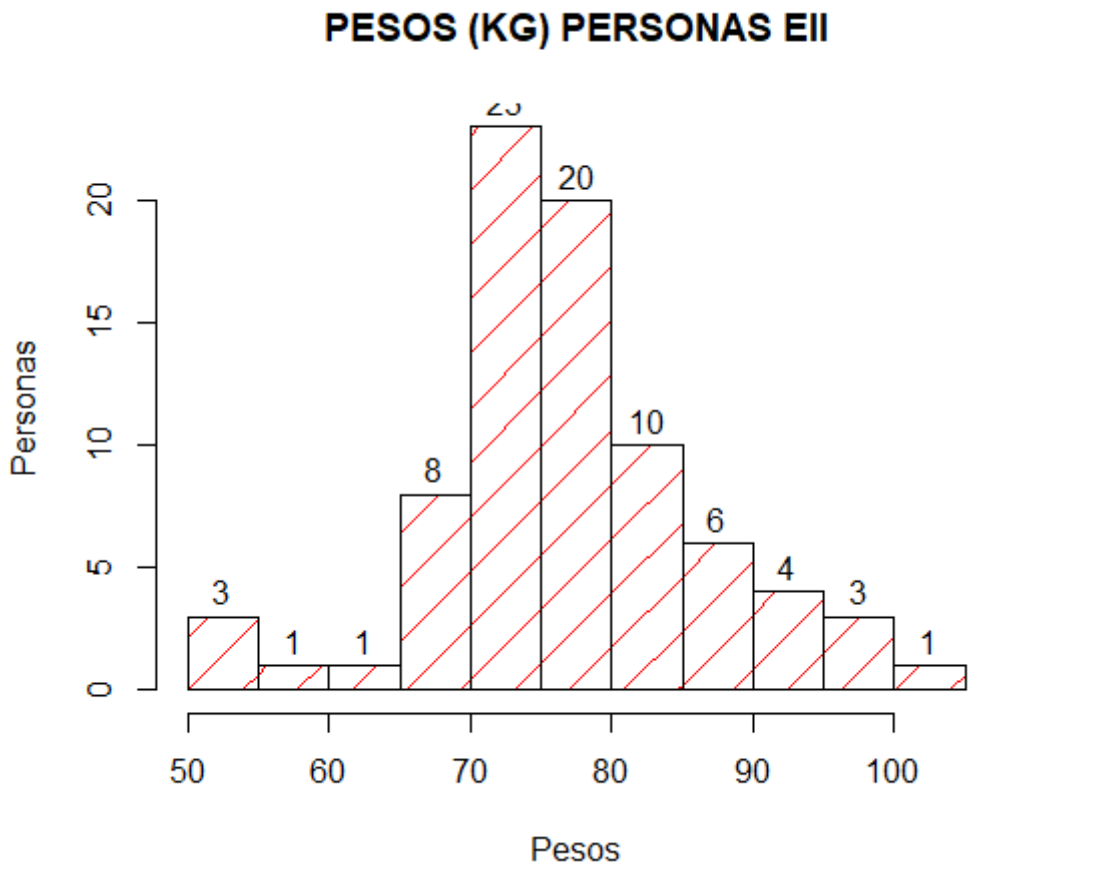
La mediana se calcula poniendo todos los datos en orden, y quedándonos justo en la mitad. Los cuartiles se calculan a partir de tablas de función y fórmulas. El rango intercuartílico es la diferencia entre los percentiles 75 y 25.

```
> median(x)
[1] 76.5
> sort(x)
 [1]  50  55  55  57  65  66  66  66  67  67  68  68  69  71  71  71  71
[18]  72  72  72  73  73  73  73  73  73  73  74  74  74  74  74  74  75
[35]  75  75  76  76  76  76  77  77  77  77  77  78  78  79  79  79  80
[52]  80  80  80  80  80  81  81  82  83  83  83  84  84  85  85  86  86
[69]  87  88  89  90  91  91  92  93  96  97  99 101
> quantile(x)
    0%    25%    50%    75%   100%
 50.00  72.75  76.50  83.00 101.00
```

d) Elaborar un histograma con los datos.

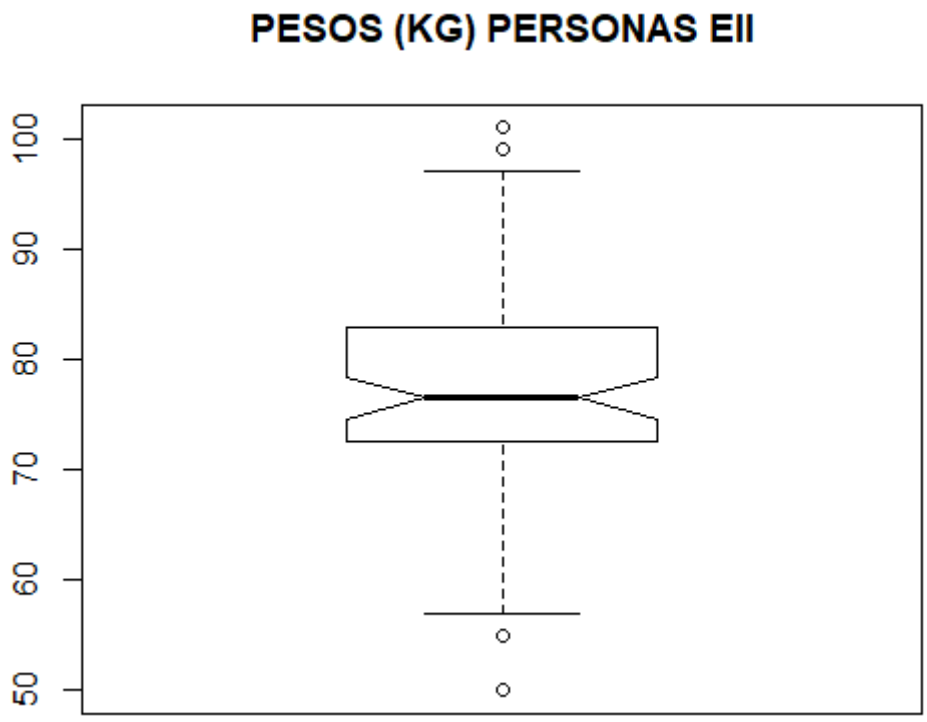
Un histograma sirve para ver gráficamente datos.

```
> c <- c(50,55,60,65,70,75,80,85,90,95,100,105)
> hist(x, breaks = c, xlab = "Pesos", ylab = "Personas", main =
+      "PESOS (KG) PERSONAS EII", col = "red", density = 5.0, border =
+      "black", labels = TRUE)
```



e) Construir un diagrama de caja y mostrar los casos atípicos y otros elementos relevantes del mismo.

```
boxplot(x, main="PESOS (KG) PERSONAS EII", notch = TRUE)
```



De esta manera observamos que los casos más frecuentes son aquellas personas cuyo peso se encuentra entre los 70 y los 90 kilogramos, mientras que los casos menos frecuentes son aquellos que pertenecen al rango de los 50 y al rango de los 100 kilogramos.

LECTURA 2: CUESTIÓN 3

Los datos siguientes se corresponden con las causas más frecuentes de suspenso o abandono de la asignatura de Métodos Estadísticos.

- **Falta de motivación por la asignatura (10)**
- **Escasa base matemática para comprender los conceptos (25).**
- **Horario del semestre (15).**
- **Carga de trabajo del curso donde se ubica la asignatura excesiva (37)**
- **Laboratorios deficientes (15)**
- **Prácticas muy laboriosas (8)**
- **Poco tiempo para el trabajo de curso (28)**
- **Explicaciones en clases teóricas no satisfactorias (10)**
- **Otras causas (6)**

a) Construir un diagrama de Pareto y evaluar los porcentajes de causas que se pueden explicar por categorías.

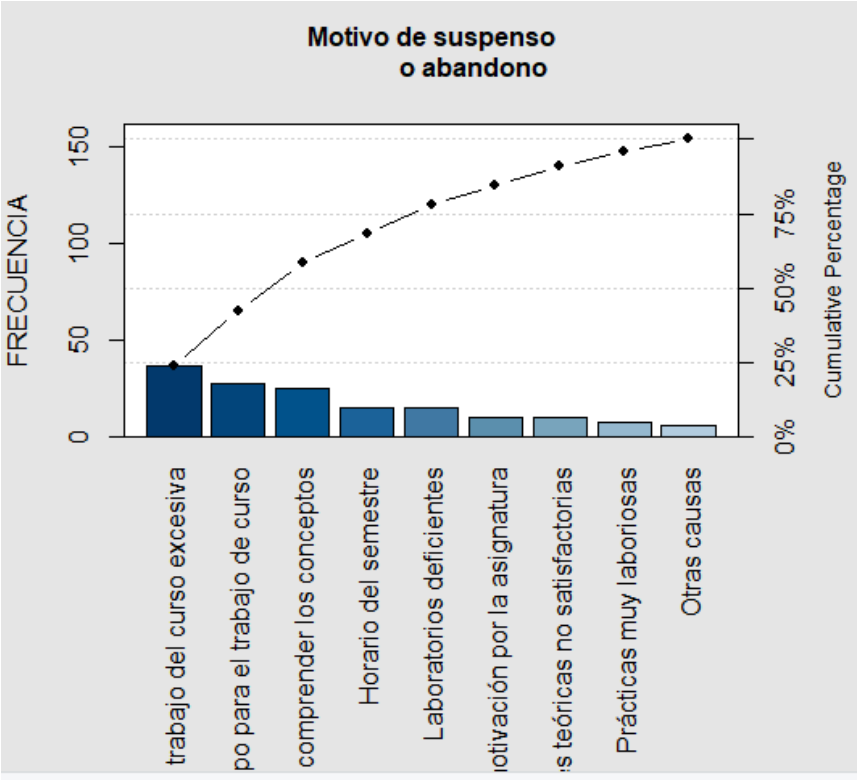
```
> library(qcc)
> Y <- c(10,25,15,37,15,8,28,10,6)
> names(Y) <- c("Falta de motivación por la asignatura",
+               "Escasa base matemática para comprender los conceptos",
+               "Horario del semestre",
+               "Carga de trabajo del curso excesiva",
+               "Laboratorios deficientes",
+               "Prácticas muy laboriosas",
+               "Poco tiempo para el trabajo de curso",
+               "Explicaciones en clases teóricas no satisfactorias",
+               "Otras causas")
> pareto.chart(Y, ylab = "FRECUENCIA", main = "Motivo de suspenso
+              o abandono")

Pareto chart analysis for Y
                                                    Frequency
  Carga de trabajo del curso excesiva                37.000000
  Poco tiempo para el trabajo de curso               28.000000
  Escasa base matemática para comprender los conceptos 25.000000
  Horario del semestre                               15.000000
  Laboratorios deficientes                           15.000000
  Falta de motivación por la asignatura              10.000000
  Explicaciones en clases teóricas no satisfactorias 10.000000
  Prácticas muy laboriosas                            8.000000
  Otras causas                                        6.000000
```

```
Pareto chart analysis for Y
                                                    Cum.Freq.
  Carga de trabajo del curso excesiva                37.000000
  Poco tiempo para el trabajo de curso               65.000000
  Escasa base matemática para comprender los conceptos  90.000000
  Horario del semestre                              105.000000
  Laboratorios deficientes                          120.000000
  Falta de motivación por la asignatura             130.000000
  Explicaciones en clases teóricas no satisfactorias 140.000000
  Prácticas muy laboriosas                          148.000000
  Otras causas                                      154.000000

Pareto chart analysis for Y
                                                    Percentage
  Carga de trabajo del curso excesiva                24.025974
  Poco tiempo para el trabajo de curso               18.181818
  Escasa base matemática para comprender los conceptos  16.233766
  Horario del semestre                                9.740260
  Laboratorios deficientes                            9.740260
  Falta de motivación por la asignatura               6.493506
  Explicaciones en clases teóricas no satisfactorias  6.493506
  Prácticas muy laboriosas                            5.194805
  Otras causas                                        3.896104

Pareto chart analysis for Y
                                                    Cum.Percent.
  Carga de trabajo del curso excesiva                  24.025974
  Poco tiempo para el trabajo de curso                 42.207792
  Escasa base matemática para comprender los conceptos  58.441558
  Horario del semestre                                 68.181818
  Laboratorios deficientes                             77.922078
  Falta de motivación por la asignatura                84.415584
  Explicaciones en clases teóricas no satisfactorias   90.909091
  Prácticas muy laboriosas                             96.103896
  Otras causas                                        100.000000
```



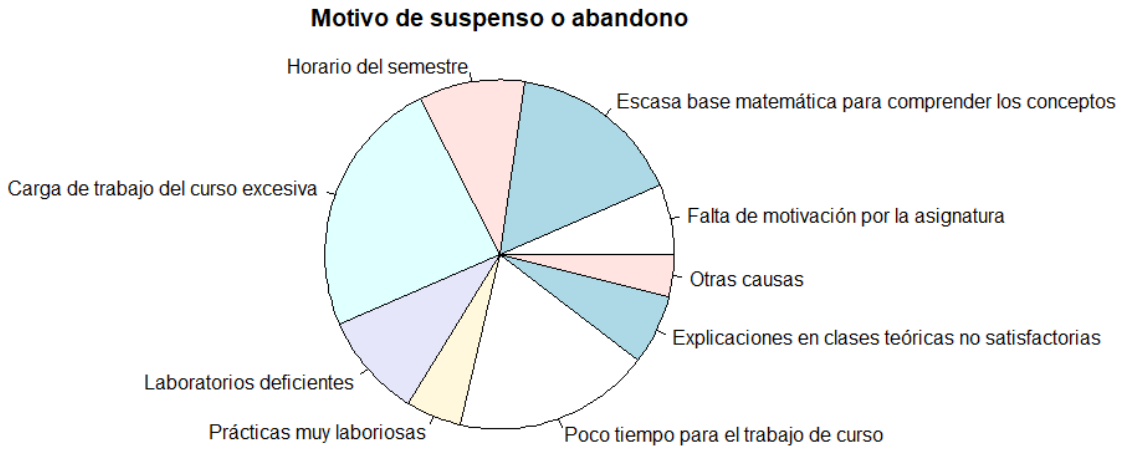
b) Analizar gráficamente el problema y establecer conclusiones y recomendaciones para el/la profesor/a, el departamento, y el centro.

Hemos de diferenciar las causas por el suspenso de la asignatura o el abandono de esta. Las referentes a causas por alumnado, y las referentes a causas por profesores.

Las referentes a causa de alumnado son la falta de motivación por la asignatura, las explicaciones en clases teóricas no satisfactorias, las prácticas muy laboriosas y la escasa base matemática para comprender los conceptos. A modo de solución, simplemente podemos recomendar que el alumnado tenga la oportunidad de preguntar sus dudas en tutorías, o se intente ayudar por sí mismo.

Las referentes a causa del profesorado, y/o la administración, encontramos el horario del semestre, la carga de trabajo del curso donde se ubica la asignatura excesiva, los laboratorios deficientes, el poco tiempo para el trabajo de curso y las explicaciones en clases teóricas no satisfactorias. Aquí podemos recomendar una mejor organización del horario de clases, así como intentar variar las explicaciones de forma más amena y/o acompañándolas de ejemplos visuales y prácticos para la captación de conocimientos.

```
> pie(Y,labels = names(Y), edges = 200, radius = 1, col = NULL,
+     main = "Motivo de suspenso o abandono")
```



LECTURA 3: CUESTIÓN 1

La probabilidad en una empresa de distribución de que el pedido de un cliente no se envía a tiempo es de 0.05. Un cliente hace tres pedidos, separados entre sí bastante tiempo. Se pide:

a) Razonar sobre la independencia de los sucesos.

Para comprobar si estos sucesos son independientes debemos comprar que $P(A|B) = P(B)$ ó $P(B|A) = P(A)$.

b) ¿Cuál es la probabilidad de que todos los pedidos se envíen a tiempo?

P(A'∩B'∩C') = 0.950*0.950*0.950 = 0.8574

c) ¿Cuál es la probabilidad de que exactamente un pedido no se envíe a tiempo?

P((A∩B'∩C')∪(A'∩B∩C')∪(A'∩B'∩C)) = 0.135

d) ¿Cuál es la probabilidad de que dos o más pedidos no se envíen a tiempo?

P((A∩B∩C')∪(A∩B'∩C)∪(A'∩B∩C)∪(A∩B∩C)) = 8.906*10^-7

LECTURA 3: CUESTIÓN 2

Un lote contiene 15 monitores de "MediaMark" y 25 monitores del "Corte Inglés" con las mismas características. Se seleccionan al azar monitores sin reemplazo del lote de 40. Sea A el evento de que el primer monitor sea de "MediaMark" y B el evento de que el segundo monitor sea del "Corte Inglés". Determinar:

a) P(A)

P(A) = 15/40 = 3/8 = 0.375

b) P(B|A)

P(B|A) = P(A∩B)/P(A) = 25/39 = 0.641

c) P(B∩A)

P(B∩A) = P(B|A)*P(A) = 0.641*0.375 = 0.24

d) P(B∪A)

P(B∪A) = P(A) + P(B) – P(A∩B) = 0.76

Supóngase que se seleccionan 3 monitores al azar sin reemplazo del lote de 40, donde C denota el evento de que el tercer monitor es de "MediaMark". Determinar:

e) P(A∩B∩C)

P(A∩B∩C) = 0.089

f) P(A∩B∩C')

P(A∩B∩C') = 0.152

LECTURA 3: CUESTIÓN 3

La cadena de supermercados MERCADONA compra jamones elaborados en las empresas "Jamones del Sur S.A" y "Embutidos Extremaduras S.L", comprando a la primera 4.5 veces más de lo que le compra a la segunda. Se sabe que el 6% de los jamones de "Jamones del Sur S.A" y que el 12% de los jamones de "Embutidos Extremadura S.L" llegan en mal estado.

a) Calcular el porcentaje de jamones que MERCADONA compra en mal estado.

Para calcular el porcentaje de jamones que MERCADONA compra en mal estado debemos de realizar la operación 18.18*0.21 + 81.81 * 0.06 = 7.0902.

b) Calcular la cantidad de jamones que hay que comprar a cada empresa si se quiere tener un total de 1350 jamones en buen estado.

Tenemos que multiplicar el número de jamones en buen estado por el porcentaje de jamones que MERCADONA compra en mal estado, así, obtenemos:

1350 * 0.0709 = 95.715, luego el número de jamones que hay que comprar será 1350+96 = 1446, y por tanto la ecuación resultará 4.5X + X = 1446. Notar que este es el número de jamones que hay que comprar a Embutidos Extremadura S.L.

LECTURA 3: CUESTIÓN 4

En un municipio de Tejeda el 59.5% de la población es femenina. Se sabe también que el 27% de los hombres y el 47.5% de las mujeres están en paro. Si se selecciona al azar una persona y resulta estar parada. ¿Cuál es la probabilidad de que sea una mujer?

La probabilidad de que sea una mujer sabiendo que está en paro será de

28.2625 / 39.1975 = 0.721 = 72.1%.

LAB 1 & 2: EJERCICIO 1

Analizar con el comando search() los paquetes presentes en el entorno de trabajo. Con library(help=package), seleccionar el paquete datasets, y, dentro de los distintos conjuntos de datos, visualizar en la consola los contenidos de varios de ellos con distintas características (tipos de variables, series, etc.).

a) Analizar cómo están estructurados los datos para familiarizarse con ellos.

```
> search()
 [1] ".GlobalEnv"        "package:qcc"       "package:fdth"
 [4] "tools:rstudio"     "package:stats"     "package:graphics"
 [7] "package:grDevices" "package:utils"     "package:datasets"
[10] "package:methods"   "Autoloads"         "package:base"
> library(help="datasets")
> titanic <- data.frame(Titanic)
> head(titanic)
  Class    Sex   Age Survived Freq
1   1st   Male Child       No    0
2   2nd   Male Child       No    0
3   3rd   Male Child       No   35
4  Crew   Male Child       No    0
5   1st Female Child       No    0
6   2nd Female Child       No    0
> tail(titanic)
   Class    Sex   Age Survived Freq
27   3rd   Male Adult      Yes   75
28  Crew   Male Adult      Yes  192
29   1st Female Adult      Yes  140
30   2nd Female Adult      Yes   80
31   3rd Female Adult      Yes   76
32  Crew Female Adult      Yes   20
```

b) Distinguir claramente en su contenido aquellos que contengan factores y vectores.

Los factores son la clase, el sexo, la edad, el valor de si han supervivido. El vector es la frecuencia.

c) Visualizar y direccionar su contenido y realizar algunos cálculos sencillos sobre el mismo.

```
> mean(Nile)
[1] 919.35
> median(Nile)
[1] 893.5
```

d) Generar, utilizando R Markdown, un report de laboratorio que recoja la sesión y explicar en él los resultados que se han obtenido. Utilizar aquellos trozos de código R empotrados (code chunks) con sintaxis knitr que se consideren necesarios para este fin.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

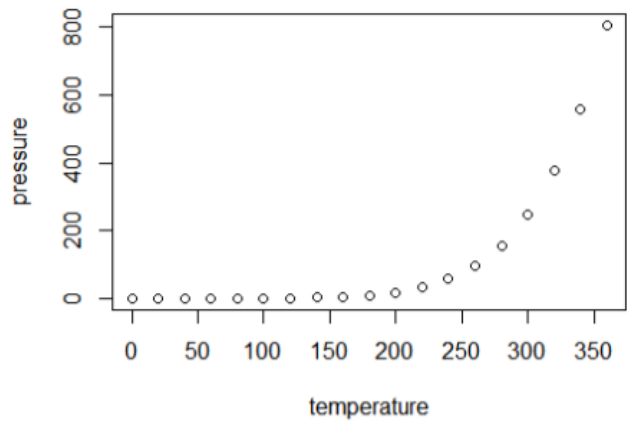
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

LAB 1 & 2: EJERCICIO 2

El Data Set "MplsStops" de la librería carData contiene datos de incidencias de personas implicadas en actuaciones policiales por el Departamento de Policía de Minneapolis en 2017.

a) Analizar su contenido y visualizar los factores y vectores.

```
> library(carData)
> library(knitr)
> datos <- MplsStops
> str(datos)
'data.frame':   51920 obs. of  14 variables:
 $ idNum         : Factor w/ 61212 levels "16-395258","16-395296",..: 6823 6824
 6825 6826 6827 6828 6829 6830 6831 6832 ...
 $ date          : POSIXct, format: "2017-01-01 00:00:42" "2017-01-01 00:03:07"
 ...
 $ problem       : Factor w/ 2 levels "suspicious","traffic": 1 1 2 1 2 2 1 2 2
 2 ...
 $ MDC           : Factor w/ 2 levels "MDC","other": 1 1 1 1 1 1 1 1 1 1 ...
 $ citationIssued: Factor w/ 2 levels "NO","YES": NA NA NA NA NA NA NA NA NA NA
 ...
 $ personSearch  : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
 $ vehicleSearch : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
 $ preRace       : Factor w/ 8 levels "Black","White",..: 3 3 3 3 3 3 3 3 3 3
 ...
 $ race          : Factor w/ 8 levels "Black","White",..: 3 3 2 4 2 4 1 7 2 1
 ...
 $ gender        : Factor w/ 3 levels "Female","Male",..: 3 2 1 2 1 2 2 1 2 2
 ...
 $ lat           : num  45 45 44.9 44.9 45 ...
 $ long          : num  -93.2 -93.3 -93.3 -93.3 -93.3 ...
 $ policePrecinct: int  1 1 5 5 1 1 1 2 2 4 ...
 $ neighborhood  : Factor w/ 87 levels "Armatage","Audubon Park",..: 11 20 84 84
 20 20 20 51 59 28 ...
```

```
> attach(datos)
> datos_s <- subset(datos[problem=="traffic",],select=c(race,gender,
+                                                       neighborhood))
> detach(datos)
> attach(datos_s)
> str(datos_s)
'data.frame':   26098 obs. of  3 variables:
 $ race        : Factor w/ 8 levels "Black","White",..: 2 2 4 7 2 1 NA 1 1 2 ...
 $ gender      : Factor w/ 3 levels "Female","Male",..: 1 1 2 1 2 2 NA 2 2 2 ...
 $ neighborhood: Factor w/ 87 levels "Armatage","Audubon Park",..: 84 20 20 51 5
9 28 11 71 20 44 ...
> head(datos_s)
            race gender               neighborhood
6825       White Female                  Whittier
6827       White Female             Downtown West
6828 East African   Male             Downtown West
6830       Other Female              Marcy Holmes
6831       White   Male Nicollet Island - East Bank
6832       Black   Male                    Folwell
> tail(datos_s)
             race  gender     neighborhood
60827 East African    Male  Powderhorn Park
60830        Black    Male         Whittier
60834        Black  Female     Marcy Holmes
60836        Black    Male St. Anthony East
60837        White    Male     Marcy Holmes
60838      Unknown Unknown  Lowry Hill East
> kable(datos_s[1:10,])
```

	race	gender	neighborhood
6825	White	Female	Whittier
6827	White	Female	Downtown West
6828	East African	Male	Downtown West
6830	Other	Female	Marcy Holmes
6831	White	Male	Nicollet Island - East Bank
6832	Black	Male	Folwell
6834	NA	NA	Cedar Riverside
6835	Black	Male	St. Anthony East
6836	Black	Male	Downtown West
6841	White	Male	Logan Park

Diferenciamos a los factores como el sexo, la raza, el sitio donde viven, y a los vectores como el número de accidentes.

b) Explicar el uso del comando subset() y emplearlo para obtener un subconjunto de este data set que contenga los vectores race, gender, y neighborhood para el caso de actuaciones derivadas de accidentes de tráfico.

El comando subset() lo utilizamos para seleccionar datos dentro de otro set de datos. Es decir, es una selección de los datos que nos interesan.

```
> subset(datos, select=c(race,gender,neighborhood))
                race  gender               neighborhood
6823         Unknown Unknown          Cedar Riverside
6824         Unknown    Male             Downtown West
6825           White  Female                  Whittier
6826    East African    Male                  Whittier
6827           White  Female             Downtown West
6828    East African    Male             Downtown West
6829           Black    Male             Downtown West
```

c) Utilizando el comando ftable() analizar los diferentes porcentajes de accidentes de tráfico según raza y género.

```
> ftable(race,gender)
                gender Female Male Unknown
race
Black                   2607 7119      30
White                   3225 4928       7
Unknown                  118  256     381
East African             374 1166       5
Latino                   318 1004       5
Native American          162  177       0
Other                    202  602       7
Asian                    194  312       0
```

Para analizar bien las diferentes posibilidades, volveremos a entablar la raza por un lado y el género por otro. Así:

```
> table(race)
race
          Black           White         Unknown    East African
           9762            8167             756            1545
         Latino Native American           Other           Asian
           1327             339             811             507
```

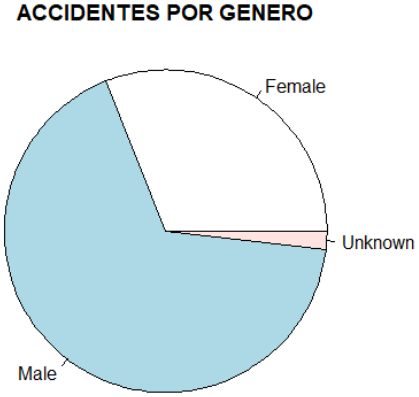
```
> table(gender)
gender
 Female    Male Unknown
   7200   15564     435
```

Observamos que las mujeres tienen muchos menos accidente de tráfico que los hombres de manera general, observando que, independientemente de la raza, los hombres tienen más del doble de accidentes que las mujeres.

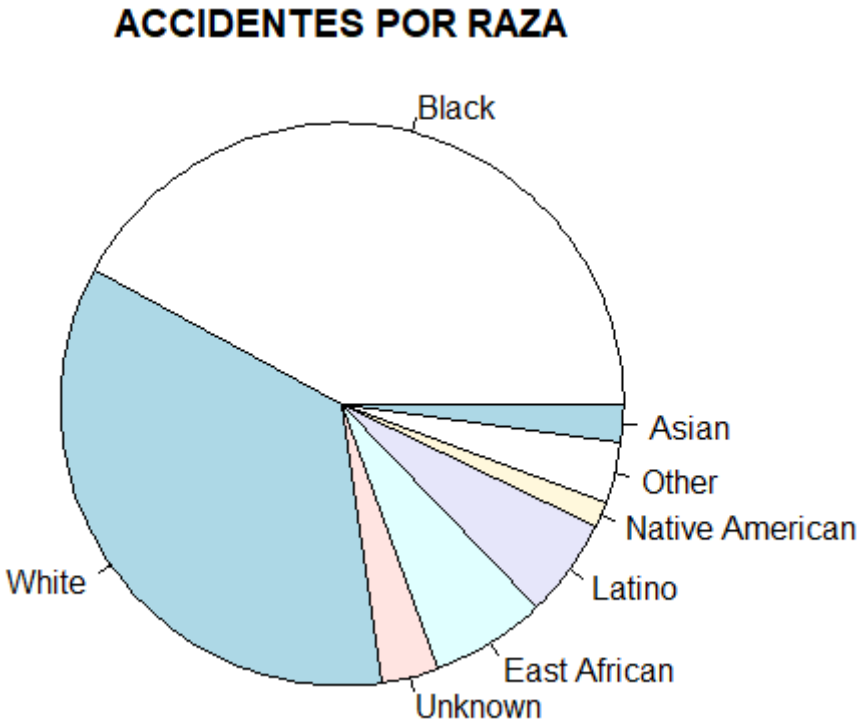
Observamos que, en las razas, las personas negras tienen más accidentes de tráfico en Minneapolis. Esto se debe a que esta ciudad está mayoritariamente constituida por personas de raza negra.

d) Visualizar con el comando gráfico pie() los resultados del apartado anterior.

```
> pie(ftable(gender), labels=levels(gender), radius = 1,
+     main = "ACCIDENTES POR GENERO")
```



```
> pie(ftable(race), labels=levels(race), radius = 1,
+     main = "ACCIDENTES POR RAZA")
```



e) Encontrar en qué zona de Minneapolis se registraron más accidentes.

```
> sitios <- ftable(neighborhood);sitios
neighborhood Armatage Audubon Park Bancroft Beltrami Bottineau Bryant Bryn - Maw
r Camden Industrial CARAG Cedar - Isles - Dean Cedar Riverside Central Cleveland
 Columbia Park Como Cooper Corcoran Diamond Lake Downtown East Downtown West Eas
t Harriet East Isles East Phillips ECCO Elliot Park Ericsson Field Folwell Fulto
n Hale Harrison Hawthorne Hiawatha Holland Howe Humboldt Industrial Area Jordan
 Keewaydin Kenny Kenwood King Field Lind - Bohanon Linden Hills Logan Park Longf
ellow Loring Park Lowry Hill Lowry Hill East Lyndale Lynnhurst Marcy Holmes Mars
hall Terrace McKinley Mid - City Industrial Midtown Phillips Minnehaha Morris Pa
rk Near - North Nicollet Island - East Bank North Loop Northeast Park Northrop P
age Phillips West Powderhorn Park Prospect Park - East River Road Regina Seward
 Sheridan Shingle Creek St. Anthony East St. Anthony West Standish Steven's Squa
re - Loring Heights Sumner - Glenwood Tangletown University of Minnesota Ventura
 Village Victory Waite Park Webber - Camden Wenonah West Calhoun Whittier Willar
d - Hay Windom Windom Park
```

```
> barriodt <- data.frame(ftable(sitios))
> kable(barriodt)
```

neighborhood	Freq
Armatage	12
Audubon Park	348
Bancroft	21
Beltrami	158
Bottineau	281
Bryant	19
Bryn - Mawr	47
Camden Industrial	22
CARAG	325
Cedar - Isles - Dean	99
Cedar Riverside	262
Central	304
Cleveland	150
Columbia Park	87
Como	314
Cooper	56
Corcoran	140
Diamond Lake	34
Downtown East	119
Downtown West	1071
East Harriet	89

```
> max(sitios)
[1] 1977
> which.max(sitios)
[1] 84
> which.max(barriodt$Freq)
[1] 84
```

El resultado de la función es el número de accidentes, y su posición es el numero 84. Luego hace referencia a:

whittier	1977

LAB 1 & 2: EJERCICIO 3

Utilizar el Data Set "Davis" de la librería carData, que proporciona los datos de hombres y mujeres que realizan ejercicio regularmente de peso y altura, tanto medidos como comunicados por los/las afectados/as. El Data Set contiene datos no disponibles (NA/S). Analizar la estructura de los datos correspondientes y:

a) Estudiar y aplicar posibles soluciones para los NA/S.

```
> library(carData)
> datos <- na.omit(Davis)
> attach(datos)
> names(datos)
[1] "sex"    "weight" "height" "repwt"  "repht"
```

Para solucionar los errores de NA/S simplemente decimos que los omita, así no nos preocupamos por este problema. Esto se hace indicando lo que omitir como vemos en la línea dos del código.

b) Encontrar las variaciones de altura y peso reales en función del género. Calcular las medias, medianas y desviación estándar correspondientes.

```
> summary(datos)
 sex        weight          height          repwt           repht
 F:99   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
 M:82   1st Qu.: 56.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:161.0
        Median : 63.0   Median :169.0   Median : 63.00   Median :168.0
        Mean   : 66.3   Mean   :170.2   Mean   : 65.68   Mean   :168.7
        3rd Qu.: 75.0   3rd Qu.:178.0   3rd Qu.: 74.00   3rd Qu.:175.0
        Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
```

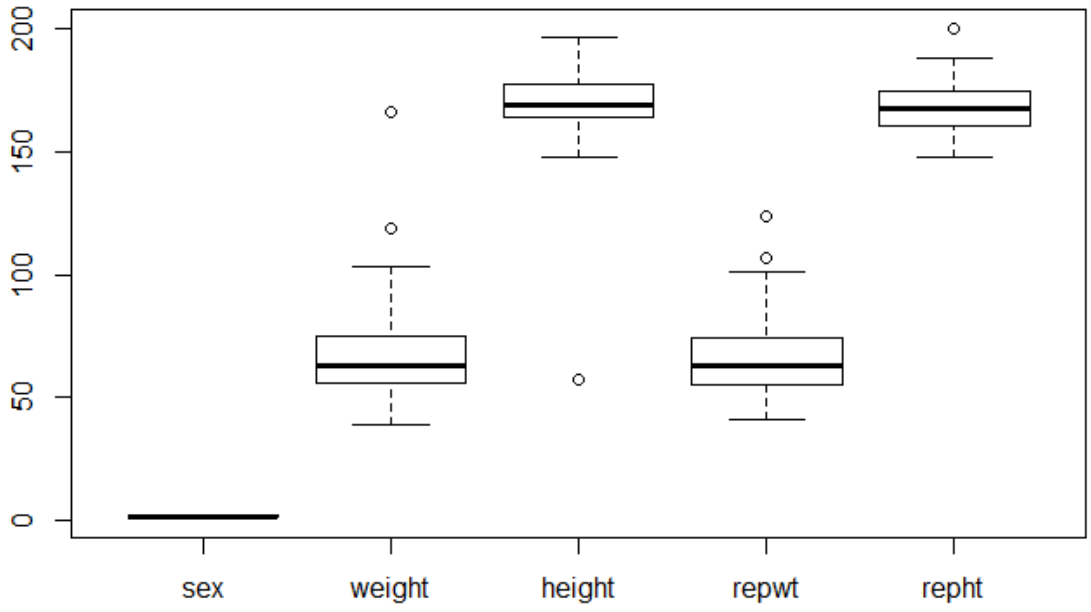
Siendo weight el peso, height la altura, repwt el peso reportado, y repht la altura reportada.

c) Analizar las variaciones de altura y peso comunicadas en función del género. Calcular las medias, medianas y desviación estándar correspondientes.

```
> summary(datos)
 sex        weight          height          repwt           repht
 F:99   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
 M:82   1st Qu.: 56.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:161.0
        Median : 63.0   Median :169.0   Median : 63.00   Median :168.0
        Mean   : 66.3   Mean   :170.2   Mean   : 65.68   Mean   :168.7
        3rd Qu.: 75.0   3rd Qu.:178.0   3rd Qu.: 74.00   3rd Qu.:175.0
        Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
```

Siendo weight el peso, height la altura, repwt el peso reportado, y repht la altura reportada.

d) Visualizar gráficamente, utilizando boxplot(), una comparativa de los datos de peso medido y peso declarado por un lado y de la altura medida y la altura declarada por otro. Establecer justificadamente las conclusiones.



e) Encontrar si hay diferencias significativas entre lo medido y declarado según el género y analizar las posibles formar de corregirla.

Observamos que no hay tanta diferencia entre los valores reales y los medidos. Hay variaciones mínimas entre ambas, siendo casi inapreciables pero que todavía deben ser mencionadas. Así, vemos una diferencia en la media de ambos casos de un kilo y de dos centímetros.

LAB 1 & 2: EJERCICIO 5

Leer el fichero "cosas.txt" que incluye el precio medio de viviendas en miles de euros por localizaciones en España. Generar un vector "Precios" a partir de los datos indicados en el fichero. Realizar a continuación las siguientes operaciones:

```
A<-rank(Precios)
B<- sort(Precios)
C<- order(Precios)
Comparativa<-data.frame(Precios,A,B,C)
Comparativa
```

Explicar la diferencia entre las diferentes columnas que resultan en cada caso y obtener las casas de precio medio superior a 190.000€.

```
> casas <- read.table("casas.txt", header = TRUE, sep = "\t")
> attach(casas)
> precios <- (Precio)
> A <- rank(precios)
> A
 [1] 12.0 10.0  5.0  6.0  7.0  2.0 11.0  8.5  1.0  3.0  8.5  4.0
> B <- sort(precios)
> B
 [1]  95 101 117 121 157 162 164 188 188 201 211 325
> C <- order(precios)
> C
 [1]  9  6 10 12  3  4  5  8 11  2  7  1
> Comparativa <- data.frame(precios,A,B,C)
> Comparativa
   precios    A   B  C
1      325 12.0  95  9
2      201 10.0 101  6
3      157  5.0 117 10
4      162  6.0 121 12
5      164  7.0 157  3
6      101  2.0 162  4
7      211 11.0 164  5
8      188  8.5 188  8
9       95  1.0 188 11
10     117  3.0 201  2
11     188  8.5 211  7
12     121  4.0 325  1
```

```
> preciosO <- B
> localO <- Localizacion[order(Precio)]
> casasord <- data.frame(localO, preciosO)
> kable(casasord)
```

localO	preciosO
Cadiz	95
Zaragoza	101
Albacete	117
Lanzarote	121
Barcelona	157
Castellon	162
Badalona	164
Teruel	188
Tenerife	188
Salamanca	201
Malaga	211
Madrid	325

```
> kable(casasord[precios>190,])
```

	localO	preciosO
1	Cadiz	95
2	Zaragoza	101
7	Badalona	164

La variable A indica la posición del valor en el vector ordenado. La variable B tendrá asignada en ella los diferentes precios ordenados de menor a mayor. La variable C nos señala la posición que ocupa cada localización en el vector inicial.

LAB 1 & 2: EJERCICIO 6

El fichero "Accidentes_1969_1984_UK.txt" contiene datos de series temporales referidas a conductores fallecidos o con lesiones graves en UK entre los años 1966 y 1984. En enero de 1983 entró en vigor la ley que obliga a la utilización del cinturón de seguridad. Entre otras variables se dispone de las siguientes:

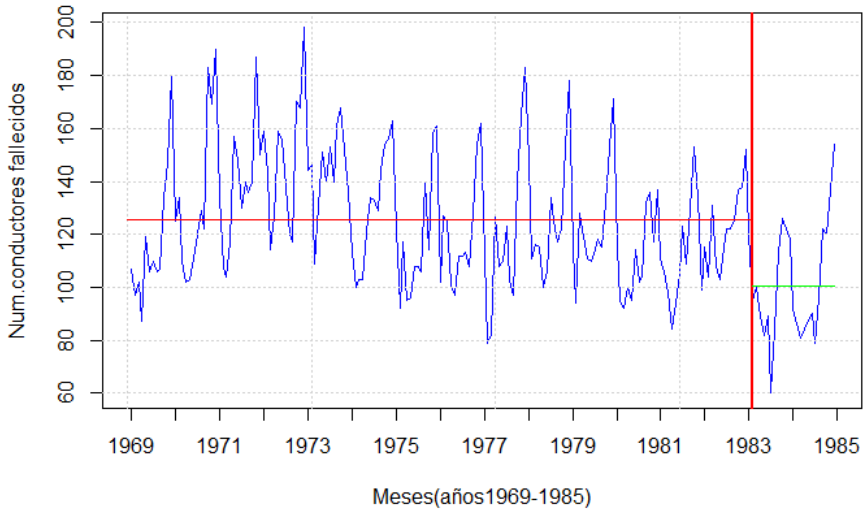
- *DriversKilled* : conductores de automóvil muertos.
- *front*: Pasajeros asientos delanteros muertos o gravemente heridos.
- *rear*: Pasajeros asientos delanteros muertos o gravemente heridos.
- *VanKilled*: número de conductores de furgonetas
- *law*: vigencia (0/1) de obligatoriedad del cinturón

a) Analizar la serie temporal de fallecidos en accidentes, encontrar sus zonas de máximo valor y visualizar el efecto de entrada en vigor de la ley.

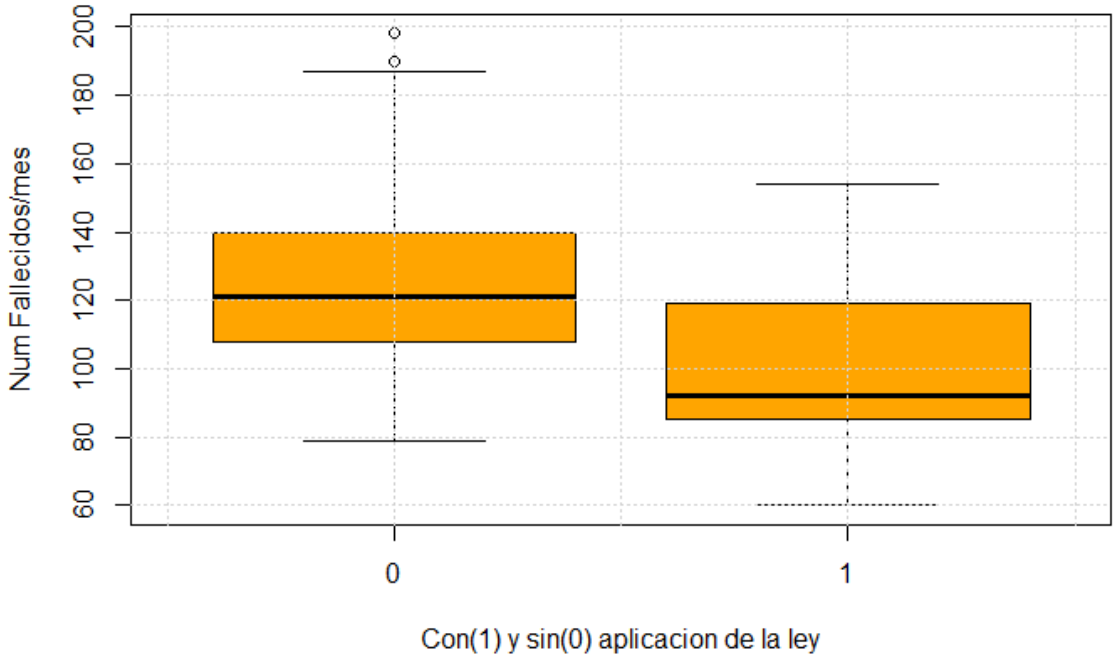
```
> library(knitr)
> help("read.table")
> datos_acc<-read.table("Accidentes_1969_1984_UK.txt",header=T, sep =",", dec =
  ".")
> attach(datos_acc)
> summary(datos_acc)
 DriversKilled      drivers         front           rear
 Min.   : 60.0   Min.   :1057   Min.   : 426.0   Min.   :224.0
 1st Qu.:104.8   1st Qu.:1462   1st Qu.: 715.5   1st Qu.:344.8
 Median :118.5   Median :1631   Median : 828.5   Median :401.5
 Mean   :122.8   Mean   :1670   Mean   : 837.2   Mean   :401.2
 3rd Qu.:138.0   3rd Qu.:1851   3rd Qu.: 950.8   3rd Qu.:456.2
 Max.   :198.0   Max.   :2654   Max.   :1299.0   Max.   :646.0
      kms          PetrolPrice        VanKilled          law
 Min.   : 7685   Min.   :0.08118   Min.   : 2.000   Min.   :0.0000
 1st Qu.:12685   1st Qu.:0.09258   1st Qu.: 6.000   1st Qu.:0.0000
 Median :14987   Median :0.10448   Median : 8.000   Median :0.0000
 Mean   :14994   Mean   :0.10362   Mean   : 9.057   Mean   :0.1198
 3rd Qu.:17203   3rd Qu.:0.11406   3rd Qu.:12.000   3rd Qu.:0.0000
 Max.   :21626   Max.   :0.13303   Max.   :17.000   Max.   :1.0000
> kable(datos_acc[1:10,])#datos del 1 al 10
```

DriversKilled	drivers	front	rear	kms	PetrolPrice	VanKilled	law
107	1687	867	269	9059	0.1029718	12	0
97	1508	825	265	7685	0.1023630	6	0
102	1507	806	319	9963	0.1020625	12	0
87	1385	814	407	10955	0.1008733	8	0
119	1632	991	454	11823	0.1010197	10	0
106	1511	945	427	12391	0.1005812	13	0
110	1559	1004	522	13460	0.1037740	11	0
106	1630	1091	536	14055	0.1040764	6	0
107	1579	958	405	12106	0.1037740	10	0
134	1653	850	437	11372	0.1030264	16	0

```
> plot(1:length(DriversKilled),DriversKilled, xaxt="n",
+      type = "l", col="blue",xlab="Meses(años1969-1985)",
+      ylab="Num.conductores fallecidos")
> years<-seq(1969,1985,1)
> axis(1,at=seq(1,length(DriversKilled)+12,12),
+      labels<-as.character(years))#convertir en caracter years
> grid()
> Febrero_83<-(1983-1969)*12+2#año de aplicacion de la ley
> abline(v=Febrero_83,col="red",lwd=2)#recta del año de aplicacion de la ley
> media_no_ley<-mean(DriversKilled[1:Febrero_83])
> media_ley<-mean(DriversKilled[Febrero_83:length(DriversKilled)])
> lines(c(0,Febrero_83),c(media_no_ley,media_no_ley),col="red")
> lines(c(Febrero_83,192),c(media_ley,media_ley),col="green")
```



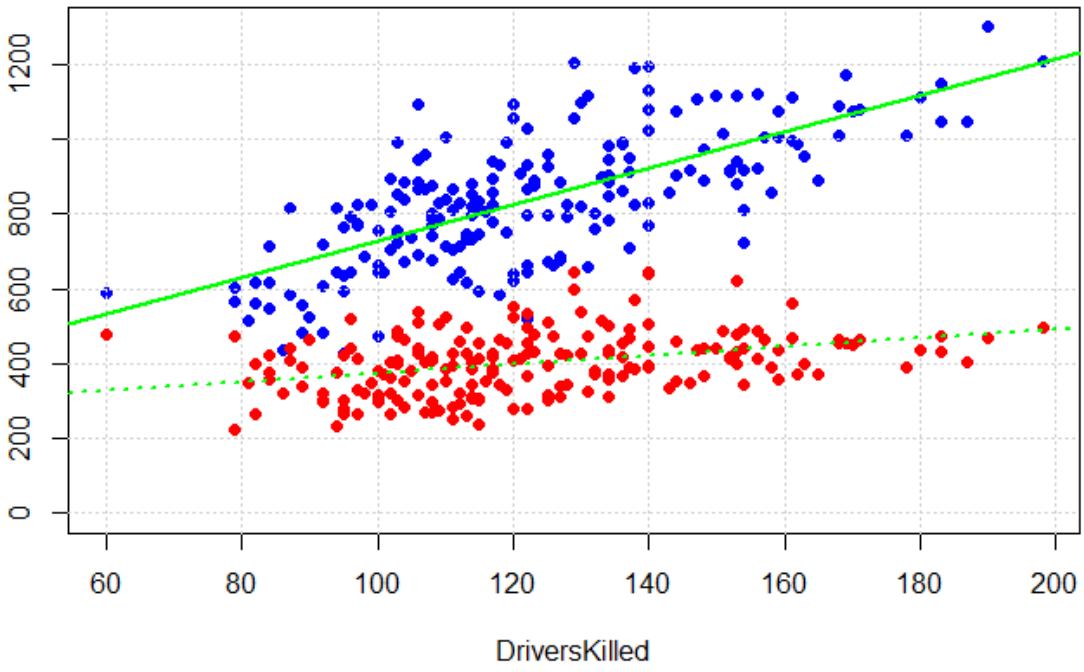

```
> boxplot(DriversKilled~law, xlab="Con(1) y sin(0) aplicacion de la ley",
+         ylab="Num Fallecidos/mes",col="orange")
> grid()
```



Vemos claramente cuando ha entrado la aplicación de la ley y asimismo cuán rápido han bajado los ratios de accidentes.

b) Analizar las relaciones existentes entre los conductores fallecidos y las víctimas según estuvieran en los asientos delanteros o traseros. Explicar y estudiar en detalle el alcance de las suposiciones establecidos en los posibles modelos.

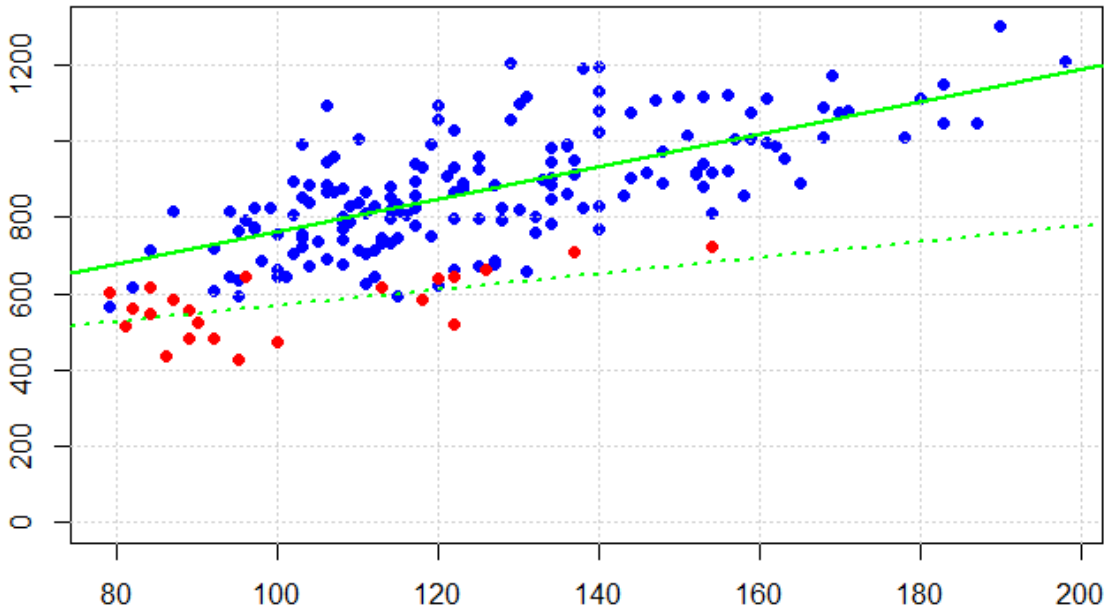
```
> plot(DriversKilled,front,pch=19,col="blue",
+      ylim = c(0,max(front)))
> grid()
> modelo1<-lm(front~DriversKilled)
> abline(modelo1,col="green",lwd=2)
> points(DriversKilled,rear,pch=19,col="red")
> modelo2<-lm(rear~DriversKilled)
> abline(modelo2,col="green",lwd=2,lty=3)
```



Observamos, como es lógico, que en casos de accidente aquellos que se sientan en el asiento delantero son mucho más probables a morir, mientras que los que se sientan en el asiento trasero son mucho más probables a sobrevivir. Esto lo observamos en la gráfica, donde los puntos azules nos indican las muertes de los que se sientan delante y los rojos de los que se sientan detrás.

c) Analizar y evaluar el efecto que tienen las furgonetas ligeras (tipo Van) en el conjunto de accidentes mortales antes y después de la aplicación de la ley. Justificar las respuestas.

```
> plot(DriversKilled[law==0],front[law==0],pch=19,col="blue",
+      ylim = c(0,max(front[law==0])))
> grid()
> modelo1<-lm(front[law==0]~DriversKilled[law==0])
> abline(modelo1,col="green",lwd=2)
> points(DriversKilled[law==1],front[law==1],pch=19,col="red")
> modelo2<-lm(front[law==1]~DriversKilled[law==1])
> abline(modelo2,col="green",lwd=2,lty=3)
```



Observamos en este caso que hay muchos menos accidentes una vez aplicada la ley en el caso de las furgonetas tipo Van. Los puntos azules indican las muertes cuando no existía la ley, mientras que los rojos las muertes una vez empezó a regularse.

LAB 1 & 2: EJERCICIO 7

El fichero "Ventas_Provincio.txt" contiene datos de ventas en euros de una empresa productora de cereales a distintas provincias españolas durante el año 2012. Se desea realizar un análisis de estos datos para valorar los procesos. Se pide:

a) Cantidades totales y las medias anuales de ventas por provincias.

```
> library(ggplot2)
> library(knitr)
> dvp<- read.table("Ventas_Provincia.txt", header = T, dec = ".", sep = ",")
> n<-length(dvp$Total_Ventas)
> s_p_12<- aggregate(dvp$Total_Ventas~dvp$Provincia, dvp, sum)
> kable(s_p_12)
```

dvp$Provincia	dvp$Total_Ventas
Albacete	728212.56
Alicante	99064.40
Almeria	450594.81
Asturias	429942.21
Avila	207869.08
Badajoz	440368.13
Barcelona	416216.34
Caceres	368265.55
Gerona	161298.07
Huelva	29392.34
Jaen	190400.45
Madrid	169592.82
Oviedo	26204.42

```
> s_m_12<- aggregate(dvp$Total_Ventas~dvp$Provincia, dvp, mean)
> kable(s_m_12)
```

dvp$Provincia	dvp$Total_Ventas
Albacete	60684.380
Alicante	9005.855
Almeria	37549.568
Asturias	35828.517
Avila	17322.423
Badajoz	36697.344
Barcelona	34684.695
Caceres	30688.796
Gerona	13441.506
Huelva	2449.362
Jaen	15866.704
Madrid	14132.735
Oviedo	3275.552

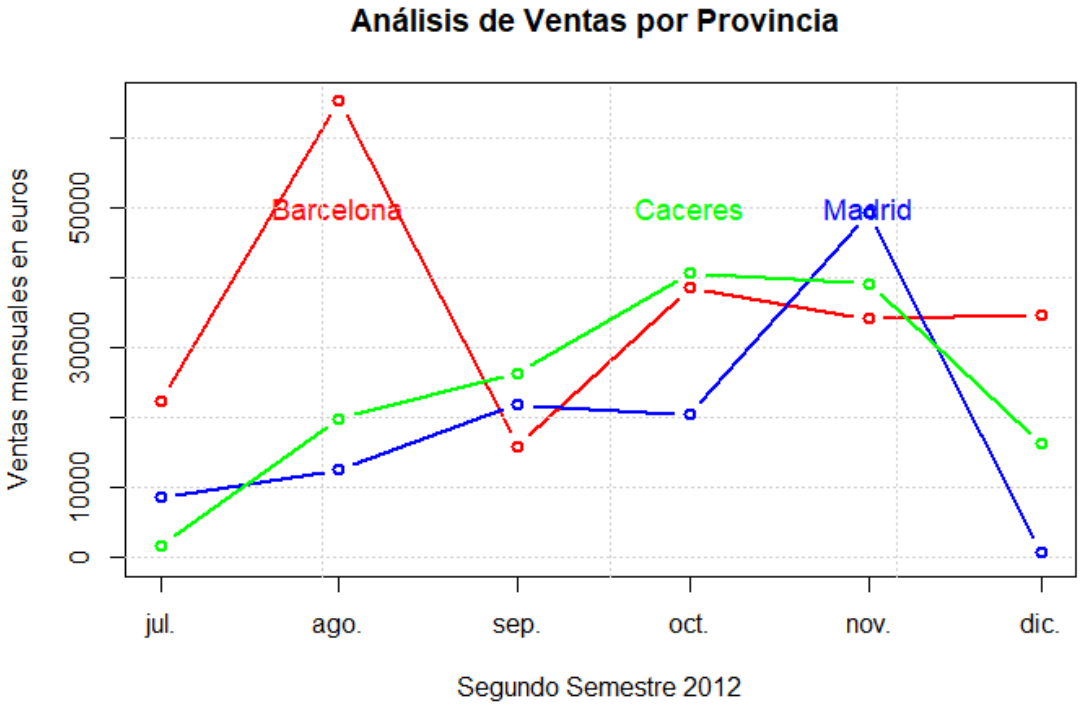
b) Provincia en la que más se vende y en la que menos.

```
> s_p_12$`dvp$Provincia`[which.max(s_p_12$`dvp$Total_Ventas`)]
[1] Albacete
13 Levels: Albacete Alicante Almeria Asturias Avila Badajoz ... Oviedo
> s_p_12$`dvp$Provincia`[which.min(s_p_12$`dvp$Total_Ventas`)]
[1] Oviedo
13 Levels: Albacete Alicante Almeria Asturias Avila Badajoz ... Oviedo
```

Albacete es la provincia que más vende y Oviedo la que menos.

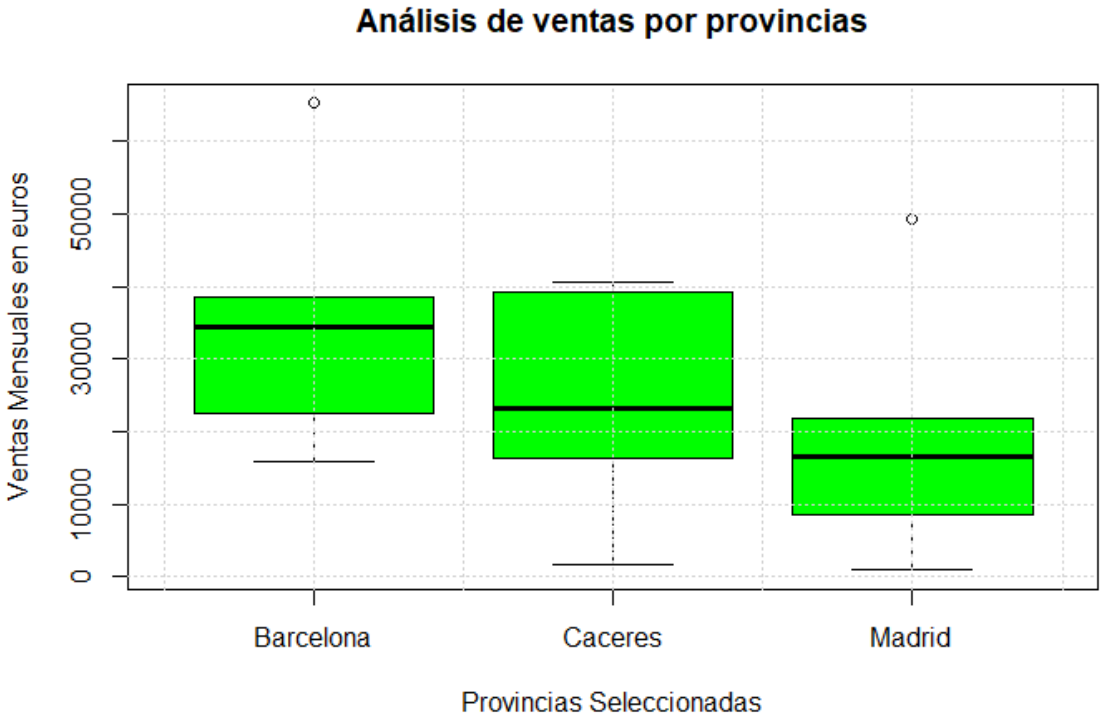
c) Estudiar la evolución de las ventas de las provincias de Cáceres, Madrid y Barcelona en el segundo semestre de 2012.

```
> Year_Mes_Semestre<- unique(year_mes[year_mes>"2012-06-01"])
> plot(Year_Mes_Semestre, Total_Ventas[Provincia=="Barcelona"], type = "b", col
 = "red", lwd = 2, ylim = c(0, max(Total_Ventas)), ylab = "Ventas mensuales en e
uros", xlab = "Segundo Semestre 2012", main = "Análisis de Ventas por Provincia
 ")
> text(Year_Mes_Semestre[2], 50000, labels = "Barcelona", col = "red", cex = 1.1
0)
> grid()
> points(Year_Mes_Semestre, Total_Ventas[Provincia=="Madrid"], type = "b", col =
 "blue", lwd=2)
> text(Year_Mes_Semestre[5], 50000, labels = "Madrid", col = "blue", cex = 1.10)
> points(Year_Mes_Semestre, Total_Ventas[Provincia=="Caceres"], type = "b", col
 = "green", lwd=2)
> text(Year_Mes_Semestre[4], 50000, labels = "Caceres", col = "green", cex = 1.1
0)
> provincias_estudio_semestre<-factor(provincias_estudio$Provincia, levels = uni
que(provincias_estudio$Provincia))
```



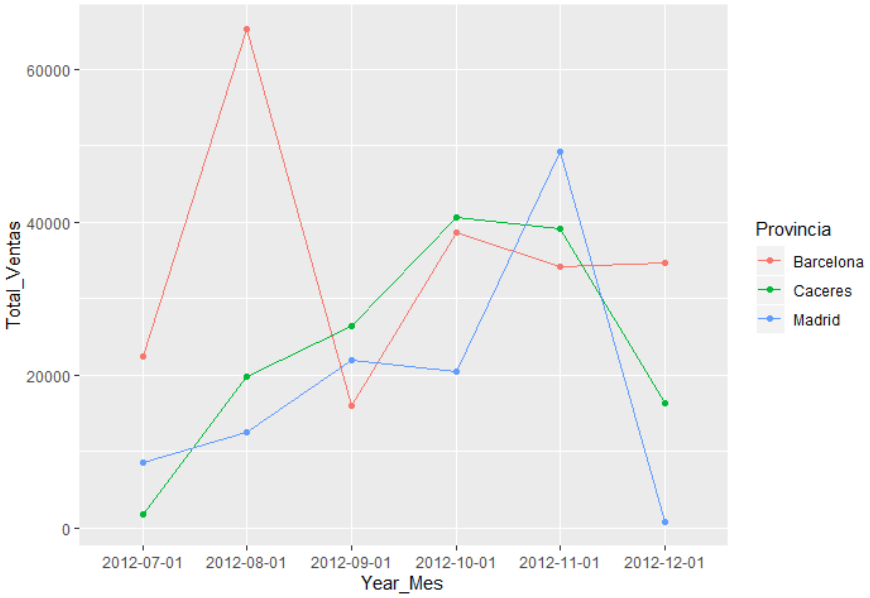
d) Utilizando los comandos gráficos de base de R, visualizar la evolución temporal de los datos del apartado c).

```
> boxplot(provincias_estudio$Total_Ventas~provincias_estudio_semestre, col= "gre
en", ylab = "Ventas Mensuales en euros", xlab="Provincias Seleccionadas", main=
 "Análisis de ventas por provincias")
> grid()
```



e) Alternativamente, utilizando ggplot2() realizar una visualización de la evolución mensual de los datos del apartado c), tanto absolutos como relativos al total de ventas de la empresa. Explicar las distintas soluciones adoptadas.

```
> g<- ggplot(data = provincias_estudio, mapping = aes(x=Year_Mes, y=Total_Venta
s, group=Provincia, colour= Provincia))
> g1<- g+ geom_point()+ geom_line()
> g1
```



f) Realizar cambios en la estética, la escala y el tema en el apartado e). Explicar las ventajas y diferencias en cada caso.

```
> g2<- g1+ xlab ("Segundo semestre 2012")+ ylab ("Ventas mensuales en euros")+ g
gtitle("Análisis de ventas por provincia")
> g2
```