TRABAJO FIN DE MÁSTER

Una Aplicación con Shiny: EDE (Estadística Descriptiva Educativa)

Máster en Estadística Aplicada para la Ciencia de Datos con R Software

AUTOR: José María Arroyo Sánchez

DIRECTOR: Juan Luis López

FECHA: 22 de Julio 2022

Máxima Formación S.L.

Resumen del Trabajo Fin de Máster (TFM)

El objetivo de este TFM es realizar una aplicación interactiva educativa mediante el paquete *Shiny* del software *R* para llevar a cabo la primera fase del análisis estadístico de cualquier estudio: la Estadística Descriptiva. En la primera versión de dicha aplicación, al seleccionar conjuntos de datos de *R* se proporciona toda la información referente a las medidas y gráficos estadísticos básicos en la Estadística, tanto desde un punto de vista univariante, como bivariante y multivariante, sin necesidad de que el usuario tenga conocimientos ni de Estadística ni del lenguaje de programación *R*.

Dicha aplicación se ha denominado **EDE** (Estadística Descriptiva Educativa).

EDE constará con seis pestañas totalmente operativas:

1. Pestaña 1 (Inicio)
2. Pestaña 2 (Estadística univariante)
3. Pestaña 3 (Estadística bivariante)
4. Pestaña 4 (Análisis correlación)
5. Pestaña 5 (Regresión lineal simple)
6. Pestaña 6 (Estadística multivariante)

Agradecimientos

En primer lugar me gustaría expresar mis agradecimientos al tutor de este proyecto, Juan Luis López, por el interés, rigor y profesionalidad que ha mostrado a lo largo del proceso. Su apoyo y dedicación han sido de gran ayuda para lograr el éxito de este trabajo.

Asimismo, quiero dar las gracias a la Doctora en Ciencias Matemáticas Mónica López Ratón por su colaboración y por haberme despertado el interés por un mundo tan fascinante como el del análisis de datos con aplicaciones Shiny.

Por último y el más importante, quiero agradecer la paciencia y el apoyo incondicional de mi familia, en especial a mi mujer Marta Carreño y a mis hijas Marta y María.

Índice

1.- Introducción

En la realización de cualquier estudio, una vez establecido el objetivo, definidas las hipótesis de investigación y realizada la recogida de información sobre las variables correspondientes, es necesaria una primera fase que consiste en el análisis descriptivo de los datos o de la información obtenida.

Dado el desconocimiento en Estadística de muchos profesionales e investigadores, es de crucial interés la implementación de software y aplicaciones que sean amigables para los usuarios de todos los ámbitos y que les permitan obtener fácilmente los resultados que desean, aplicados a cada situación. En esta línea, el **objetivo** de este **TFM** es la realización de una aplicación o herramienta interactiva para la Educación mediante el paquete *Shiny* del software estadístico *R* que permita el análisis descriptivo de un conjunto de datos. En dicha aplicación, al seleccionar conjuntos de datos de *R* se proporciona toda la información más relevante acerca de las medidas y gráficos estadísticos sin necesidad de que el usuario tenga conocimientos ni de Estadística ni del lenguaje de programación *R*.

2.- Material y métodos

2.1. Introducción a Shiny

Para que la compresión de dicha aplicación sea más fácil de entender se explica brevemente el funcionamiento del paquete Shiny. De esta forma si algún usuario quiere crear su propia App también le resultará más ágil y cómodo.

2.1.1. Introducción a SHINY

Shiny es un paquete de R que permite construir aplicaciones web interactivas a partir de los scripts de R.

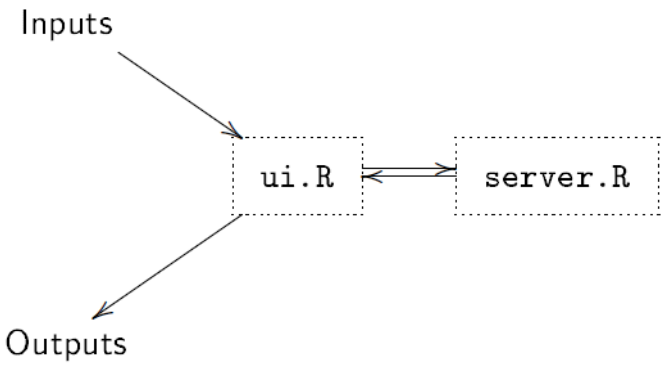
La **interactividad de estas aplicaciones** permite manipular los datos sin tener que manipular el código. De hecho, en la naturaleza de Shiny subyace un concepto aún más fuerte: la reactividad.

Las **aplicaciones web creadas con Shiny** pueden estar enfocadas a numerosos ámbitos: investigación, profesional o, por supuesto, la docencia. Estas aplicaciones pueden abrirse desde el propio ordenador, una tablet o incluso el móvil.

2.1.2. Crear un proyecto

Una **app de Shiny** consta (al menos) de dos archivos:

- un script para la interfaz del usuario, **(user-interface, ui.R)**, que recibe los inputs y muestra los outputs.
- un script para los cálculos **(server.R)**, que realiza los cálculos necesarios.

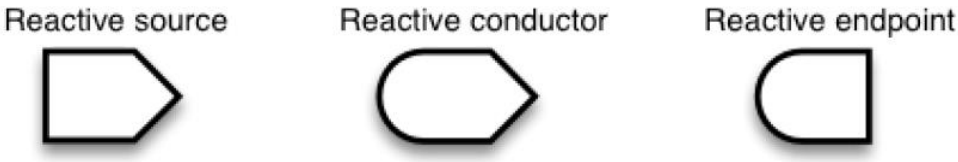


2.1.3. Interactividad

Como se ha indicado anteriormente, la principal característica de las aplicaciones creadas con Shiny es su **interactividad** (permiten manipular los datos sin manejar el código). Esto tiene que ver con la programación reactiva en la que cada modificación por parte del usuario “renueva” todo el proceso.

En Shiny, hay tres tipos de objetos relacionados con la programación reactiva:

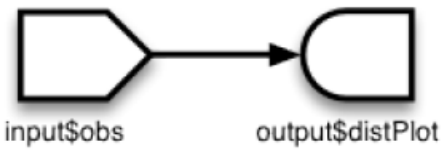
- **reactive sources (fuentes reactivas):** los inputs que se introducen en ui.R y se envían a server.R.
- **reactive conductors (conductores reactivos):** una transformación de los inputs que se usa en **server.R.**
- **reactive endpoints (puntos finales reactivos):** los outputs obtenidos en server.R y que se envían a **ui.R.**



2.1.4. Esquema simple

El **esquema más simple** toma como entrada una fuente reactiva y, a partir de ella, da como resultado un punto final reactivo:

Por ejemplo, supongamos que queremos mostrar histogramas de conjuntos de datos generados por una distribución normal. Pedimos al usuario que indique un tamaño muestral **input$obs** y se obtendrá el histograma **output$distPlot:**



2.1.5. Esquema simple ui.R

```
library(shiny)

shinyUI(fluidPage(

titlePanel("Hello Shiny!"),

sidebarLayout(
  sidebarPanel(
    sliderInput("obs",
                "Number of observations:",
                min = 1,
                max = 1000,
                value = 500)
  ),

  mainPanel(
    plotOutput("distPlot")
  ))))
```

2.1.6. Esquema simple server.R

```
library(shiny)

shinyServer(function(input, output) {

output$distPlot <- renderPlot({
  dist <- rnorm(input$obs)
  hist(dist)
})

})
```

2.1.7. Inputs

Los **inputs** pueden ser números o valores y parámetros lógicos. Más generalmente, Shiny dispone de los siguientes inputs que se incorporan mediante lo que se denominan **widgets** (sus nombres acaban generalmente con Input):

Tipo de **input**	Uso
`actionButton`	Action Button
`checkboxGroupInput`	A group of check boxes
`checkboxInput`	A single check box
`dateInput`	A calendar to aid date selection
`dateRangeInput`	A pair of calendars for selecting a date range
`fileInput`	A file upload control wizard
`helpText`	Help text that can be added to an input form
`numericInput`	A field to enter numbers
`radioButtons`	A set of radio buttons
`selectInput`	A box with choices to select from
`sliderInput`	A slider bar
`submitButton`	A submit button
`textInput`	A field to enter text

Una vez que se ha incluido un **widget en ui.R** (en cuya sintaxis hay que indicar el nombre que se le da a la variable que representa), se utilizará como **input$nombre** dentro de **server.R.**

2.1.8. De input a output

Los inputs anteriores que se introducen en **ui.R** se envían a **server.R** y se utilizan para obtener los **outputs**. Las operaciones que se realizan en server.R con los inputs y que dan como resultado los outputs, son de tipo reactivo (sus nombres empiezan por **render** y acaban dependiendo del tipo de objeto que devuelven):

Tipo de **objeto** que se obtiene	Uso
`renderImage`	images (saved as a link to a source file)
`renderPlot`	plots
`renderPrint`	any printed output
`renderTable`	data frame, matrix, other table like structures
`renderText`	character strings
`renderUI`	a Shiny tag object or HTML

En las funciones **render**, aparecerán como argumentos los **input$nombre** que hayamos introducido en **ui.R.**

Las funciones render se asignan a objetos del tipo **output$nombre.**

2.1.9. Outputs

Los resultados que se han obtenido con el proceso anterior se devuelven a **ui.R** que deberá mostrarlos (o no). Según el tipo de **output**, debe indicarse en **ui.R** utilizando las siguientes opciones:

Tipo de **output**	Significado
`htmlOutput`	raw HTML
`imageOutput`	image
`plotOutput`	plot
`tableOutput`	table
`textOutput`	text
`uiOutput`	raw HTML
`verbatimTextOutput`	text

Las **funciones Output** necesitan como argumento el "nombre" del output$nombre: **plotOutput("nombre").**

En server.R	⇝	En ui.R	⇝	Tipo de objeto
`renderImage`	⇝	`imageOutput`	⇝	Imagen
`renderPlot`	⇝	`plotOutput`	⇝	Gráfico
`renderTable`	⇝	`tableOutput`	⇝	Tabla
`renderText`	⇝	`textOutput`	⇝	Texto
`renderText`	⇝	`htmlOutput`	⇝	HTML
`renderText`	⇝	`verbatimTextOutput`	⇝	Texto

La aplicación creada se ha denominado **EDE** (**Estadística Descriptiva Educativa**) y puesto que el fin de dicha aplicación es la realización del análisis descriptivo de un conjunto de datos, el primer paso es la definición de todos los descriptivos, medidas estadísticas y gráficos, y la forma de obtenerlos con el software *R*, es decir, a partir de qué funciones o paquetes específicos se pueden calcular. Por eso además del paquete *shiny* también fueron necesarios otros paquetes de *R*.

2.2. Pestañas de EDE

2.2.1. Introducción a EDE

Para poder lograr el objetivo planteado, la creación de la aplicación interactiva, se han utilizado el software estadístico ***R***, el entorno de programación ***RStudio*** y el paquete ***Shiny*** de *R* que permite la construcción de aplicaciones web interactivas a partir de los scripts de *R*.

En los respectivos Anexos del trabajo, se explican con detalle todos estos aspectos, para que el lector o investigador que esté más interesado en conocer los detalles más técnicos, pueda consultarlos, en referencia a qué es la Estadística Descriptiva, en qué consiste, cuales son las medidas estadísticas y gráficos empleados (definición y utilidad) y como obtenerlos en *R*. Además, también se explica brevemente el manejo del programa estadístico *R* y del entorno *RStudio*, todos ellos fundamentales en este trabajo.

Para la ilustración de la aplicación EDE, se han utilizado los conjuntos de datos de muestra en *R cars*, *pressure*, *rock*, *iris* y *mtcars*.

- ¿Cómo se ha desarrollado la aplicación EDE?

El desarrollo y estructura de la aplicación EDE se ha realizado de la siguiente forma:

EDE consta de seis pestañas totalmente operativas:

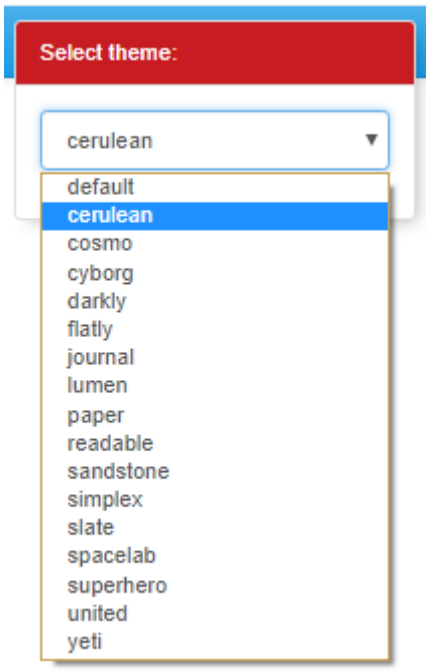
1. Pestaña 1 (Inicio)
2. Pestaña 2 (Estadística univariante)
3. Pestaña 3 (Estadística bivariante)
4. Pestaña 4 (Análisis correlación)
5. Pestaña 5 (Regresión lineal simple)
6. Pestaña 6 (Estadística multivariante)

A continuación se describe cada una de las pestañas anteriores:

2.2.2. Pestaña 1: Inicio

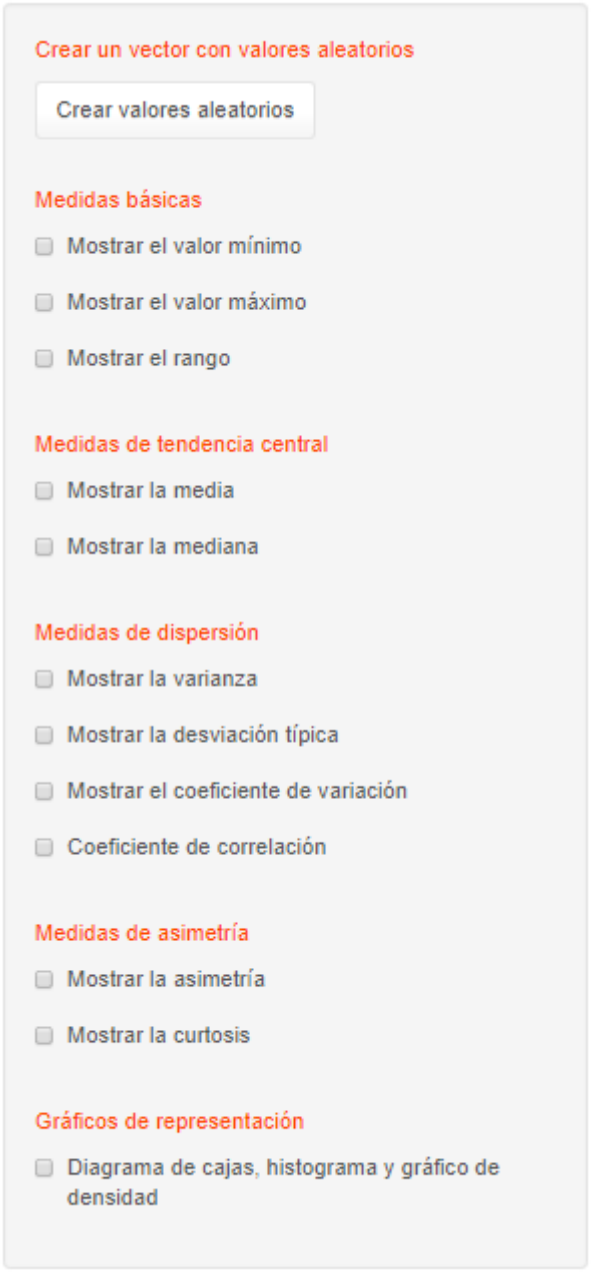
La **pestaña Inicio** muestra los motivos que me han llevado a realizar la aplicación así como un pequeño resumen sobre la Estadística Descriptiva.

En esta pestaña merece la pena resaltar el combo que se ha creado, un combo especial para que el usuario final pueda elegir el tema (o paleta de colores) con el que quiera trabajar. Por defecto aparece el tema default.



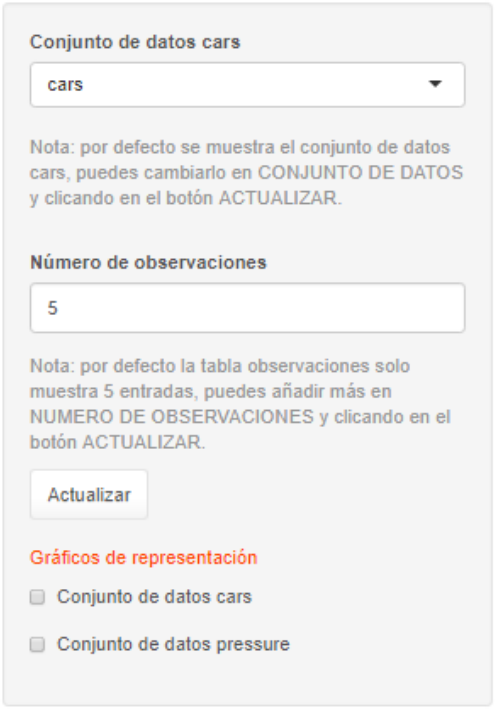
2.2.3. Pestaña 2: Estadística univariante

La **pestaña Estadística univariante** muestra todos los descriptivos básicos de la Estadística Descriptiva en un checklist (para poder seleccionar el o los que se desee):



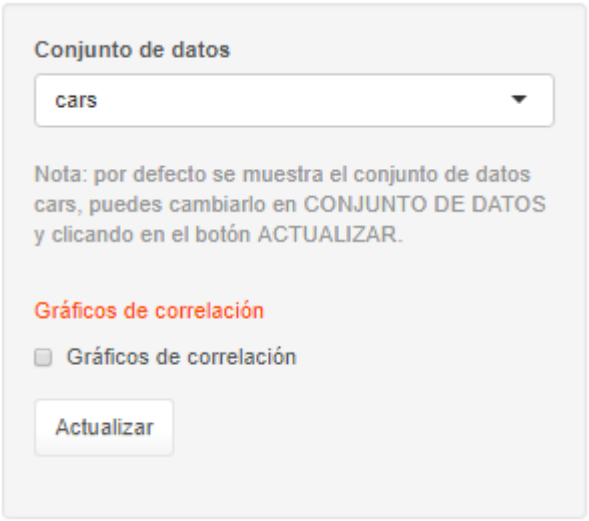
2.2.4. Pestaña 3: Estadística bivariante

La **pestaña Estadística bivariante** muestra los conjuntos de datos de muestra en *R cars* y *pressure*, así como el número de observaciones (por defecto 5 aunque este número se puede modificar). También muestra los gráficos de representación para el conjunto de datos en un checklist:



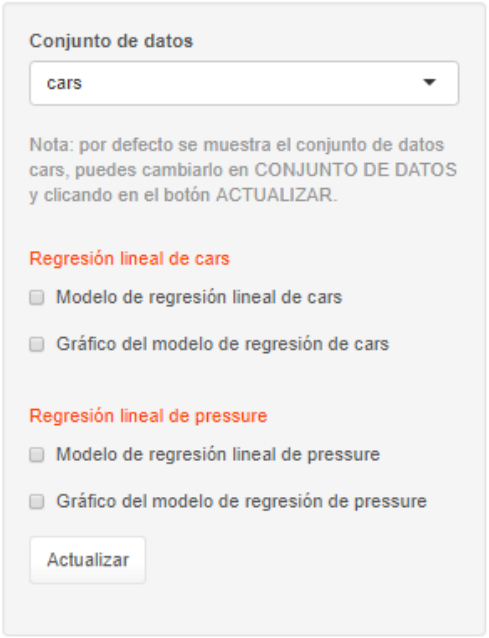
2.2.5. Pestaña 4: Análisis de correlación

En la **pestaña Análisis correlación** se continúa trabajando con los mismos conjuntos de datos del apartado anterior (*cars* y *pressure*). Tenemos un checklist para poder cambiar entre estos dos conjuntos de datos y para poder visualizar los gráficos de correlación:



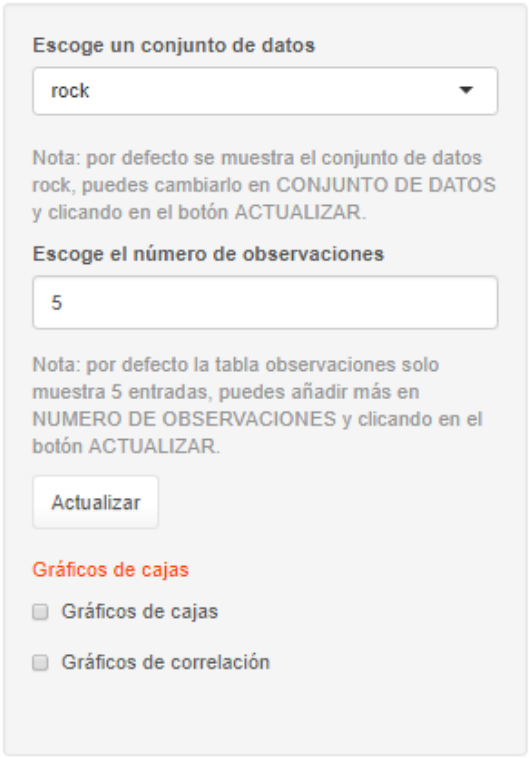
2.2.6. Pestaña 5: Regresión lineal simple

En la **pestaña Regresión lineal simple** se siguen considerando los mismos conjuntos de datos del apartado anterior (*cars* y *pressure*). Tenemos un checklist para poder cambiar entre estos dos conjuntos de datos y para poder visualizar tanto las regresiones lineales como los gráficos de éstas:



2.2.7. Pestaña 6: Estadística multivariante

La **pestaña Estadística multivariante** muestra los conjuntos de datos de muestra *rock*, *iris* y *mtcars*, así como el número de observaciones (por defecto 5 aunque este número se puede modificar). También se pueden realizar los gráficos de representación para el conjunto de datos correspondiente en un checklist:



3.- Resultados

A continuación se presentan los resultados que muestra la aplicación según cada una de las pestañas introducidas:

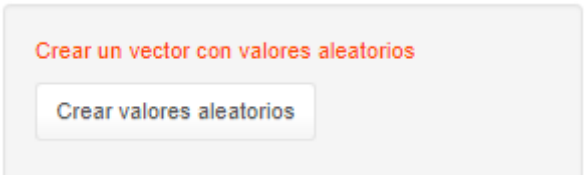
3.1. Pestaña 2: Estadística univariante

En la **pestaña 2 Estadística univariante** al marcar en el botón “Crear valores aleatorios”, se crea un vector aleatorio de 50 datos comprendidos entre los valores 0 y 100. Además, estos datos también se pueden repetir:

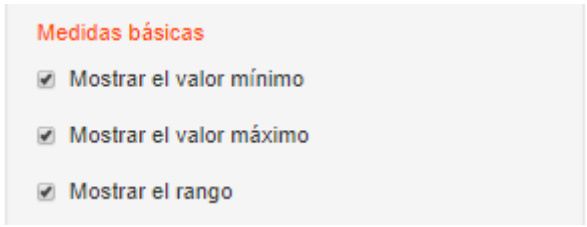
Aquí tenemos el vector de 50 valores creado de forma aleatoria entre los números 0 y 100.

85 40 85 40 19 39 26 94 56 21 8 99 75 19 73 79 85 47 11 9 37 26 25 77 14 48 5 54 12 48 96 9 98 0 61 33 12 83 54 10 36 71 48 12 7 68 22 92 94 65

Este vector se puede variar tantas veces como el usuario quiera haciendo clic en el botón “Crear valores aleatorios”:

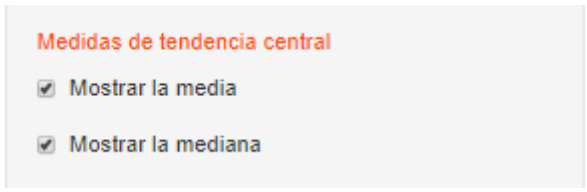


Una vez creado el vector y clicando en el checklist correspondiente se pueden visualizar los resultados de los siguientes estadísticos:



Medidas básicas

Valor mínimo	Valor máximo	Rango
6	99	93

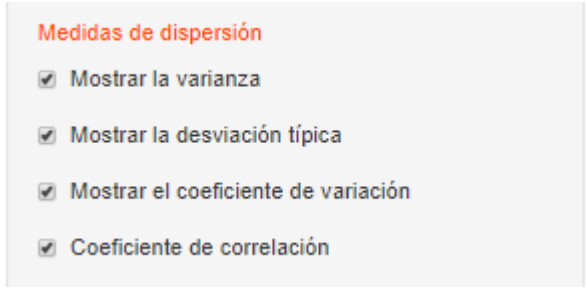


Medidas de tendencia central

1.- La media es la suma de todas las observaciones divididas entre el número de observaciones de los datos.

2.- La mediana es valor que divide un conjunto de observaciones, ordenadas de menor a mayor, en dos partes con el mismo número de observaciones.

Valor de la media	Valor de la mediana
53.280	52.500



Medidas de dispersión

1.- La varianza es la media aritmética de las desviaciones al cuadrado de los valores de la variable con respecto a su media. La varianza es siempre positiva y cuanto mayor sea su valor mayor será la dispersión de los datos.

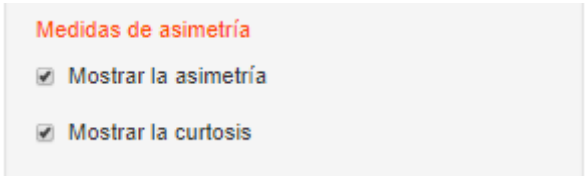
2.- La desviación típica o estándar es la raíz cuadrada positiva de la varianza. La desviación típica se usa más que la varianza, ya que está expresada en las mismas unidades que la variable, mientras que la varianza está expressada en unidades cuadradas.

Valor de la varianza	Valor de la desviación típica
821.675	28.665

3.- El coeficiente de variación es la división entre la desviación típica de una muestra y su media.

4.- El coeficiente de correlación es la medida que cuantifica la intensidad de la relación lineal entre dos variables, toma los valores entre -1 y 1. En nuestro caso, como solo tenemos una variable será siempre 1.

Valor del coeficiente de variación	Valor del coeficiente de correlación
0.538	1.000



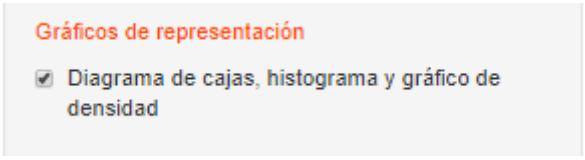
Medidas de asimetría

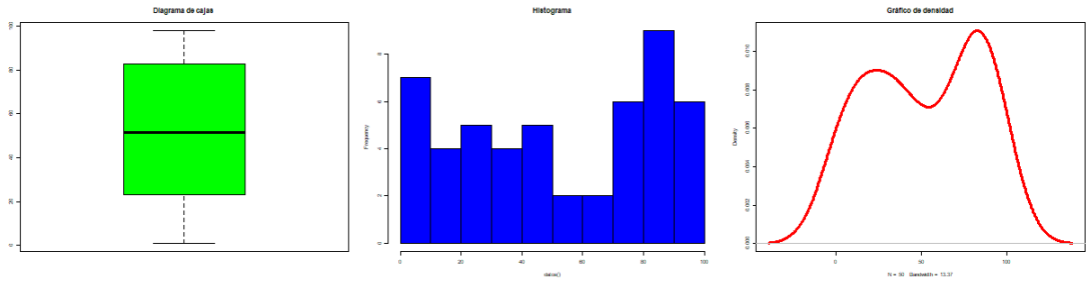
1.-La asimetría es la medida que indica la simetría de la distribución de una variable respecto a la media.

2.-La curtosis (o apuntamiento) mide a su vez cómo de apuntada o achatada es la distribución mirando la cantidad de elementos cercanos al valor central.

Valor de la asimetría	Valor de la curtosis
0.020	1.839

Y de los siguientes gráficos:





3.2. Pestaña 3: Estadística bivariante

Al entrar en la **pestaña Estadística bivariante** se muestra por defecto un resumen (“summary”) del conjunto de datos *cars* y las primeras 5 observaciones de este conjunto:

En este apartado se trabajará con dos conjuntos de datos de R, cars y pressure, los cuales tienen 2 variables.

1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

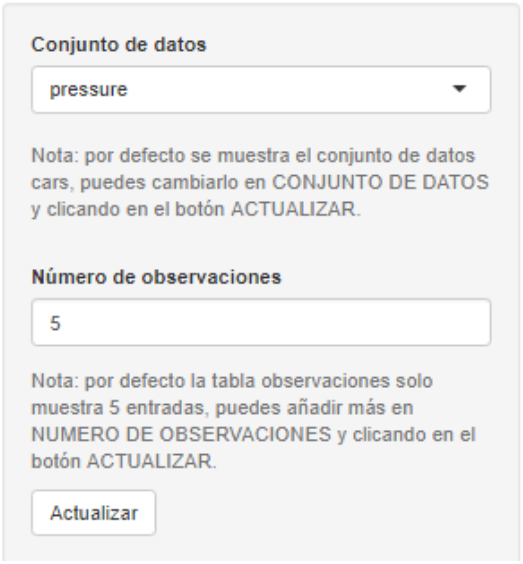
```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:

Observaciones

speed	dist
4.00	2.00
4.00	10.00
7.00	4.00
7.00	22.00
8.00	16.00

Aunque se puede cambiar el conjunto de datos haciendo clic en el combo y actualizando:



Y de esta forma podemos visualizar el “summary” del conjunto de datos pressure y sus primeras 5 observaciones:

1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

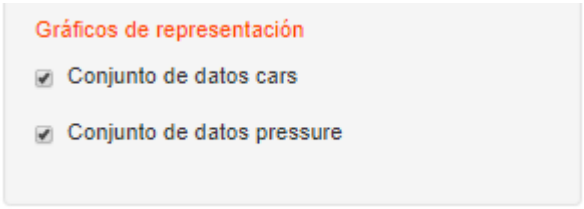
```
  temperature      pressure      
 Min.   :  0   Min.   :  0.0002  
 1st Qu.: 90   1st Qu.:  0.1800  
 Median :180   Median :  8.8000  
 Mean   :180   Mean   :124.3367  
 3rd Qu.:270   3rd Qu.:126.5000  
 Max.   :360   Max.   :806.0000  
```

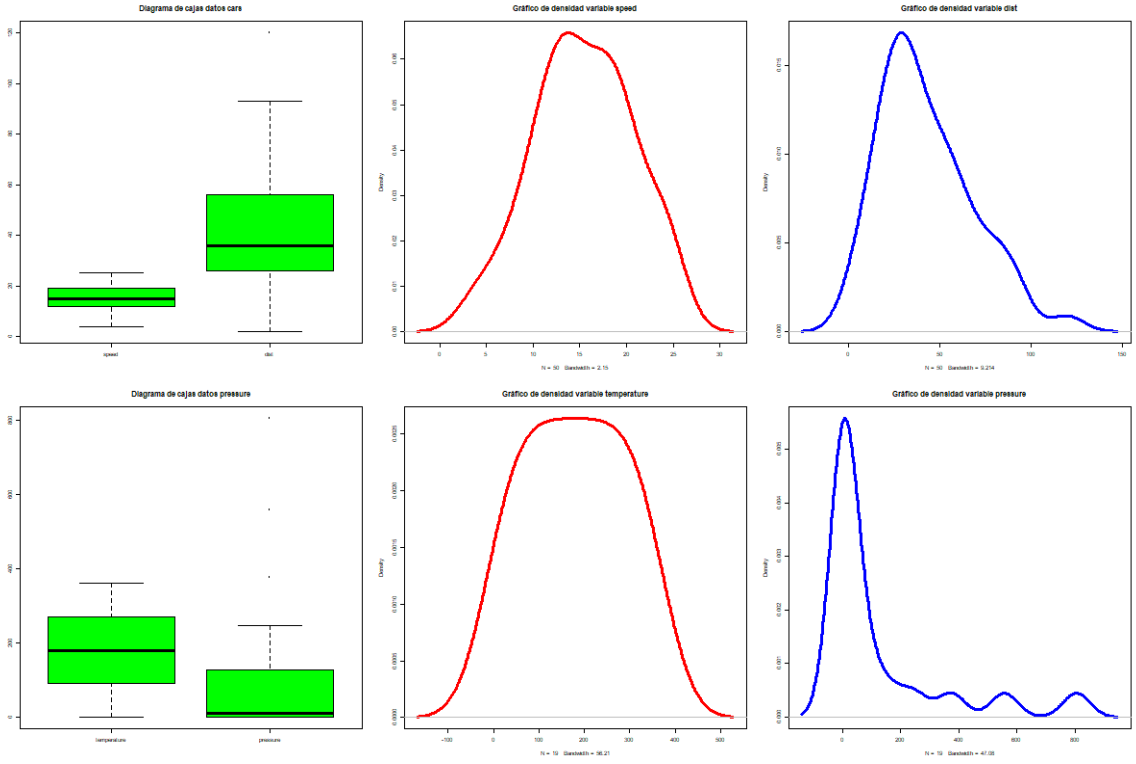
2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:

Observaciones

temperature	pressure
0.00	0.00
20.00	0.00
40.00	0.01
60.00	0.03
80.00	0.09

Al hacer clic en el checklist de "*Conjunto de datos cars*" o "*Conjunto de datos pressure*" nos presenta los principales gráficos estadísticos para dos variables (gráfico de cajas y diagramas de densidad):





3.3. Pestaña 4: Análisis de correlación

Al entrar en esta pestaña se muestra por defecto un "summary" del conjunto de datos *cars* y se hace un breve resumen del significado del término correlación:

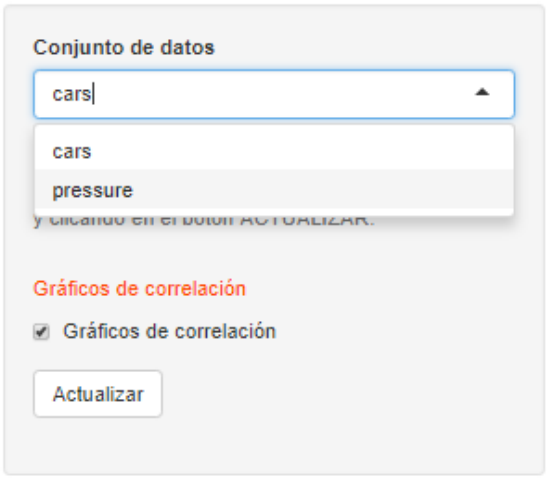
1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

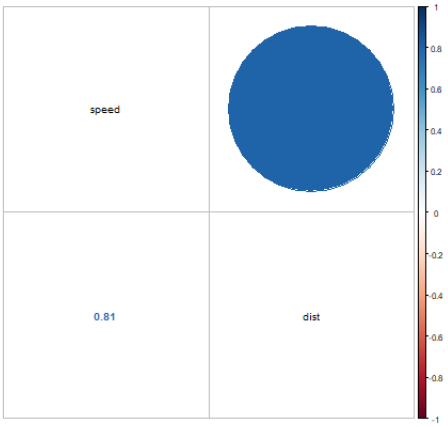
```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

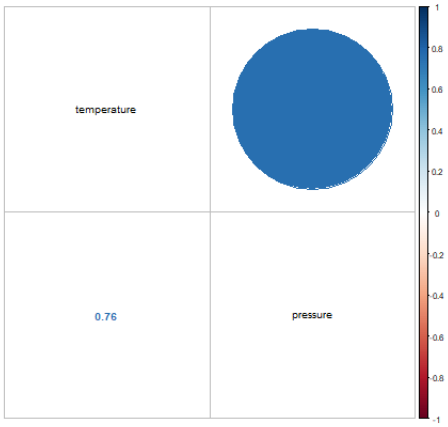
La correlación es un tipo de asociación entre dos variables numéricas, específicamente evalúa la tendencia (creciente o decreciente) en los datos.
1.- La correlación nos permite medir el signo y magnitud de la tendencia entre dos variables.
2.- El signo nos indica la dirección de la relación, como hemos visto en el diagrama de dispersión.
3.- La magnitud nos indica la fuerza de la relación, y toma valores entre -1 a 1.

Al hacer clic en el checklist de "*Gráficos de correlación*" nos muestra los gráficos de correlación de estos dos conjuntos de datos:

- Gráfico de correlación del conjunto de datos cars.
- Gráfico de correlación del conjunto de datos pressure.







3.4. Pestaña 5: Regresión lineal simple

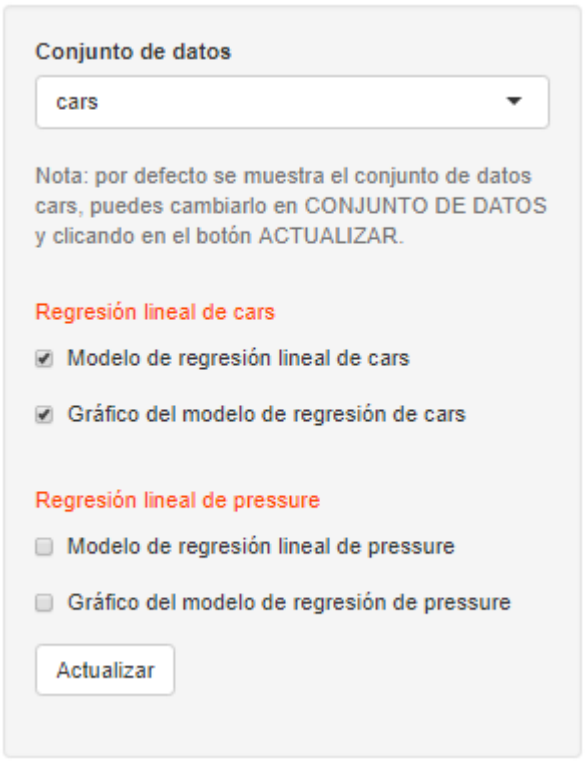
De modo similar a la pestaña anterior, al entrar en esta pestaña se muestra por defecto un resumen estadístico ("summary") de los datos *cars* y se hace un breve resumen del significado del término regresión lineal:

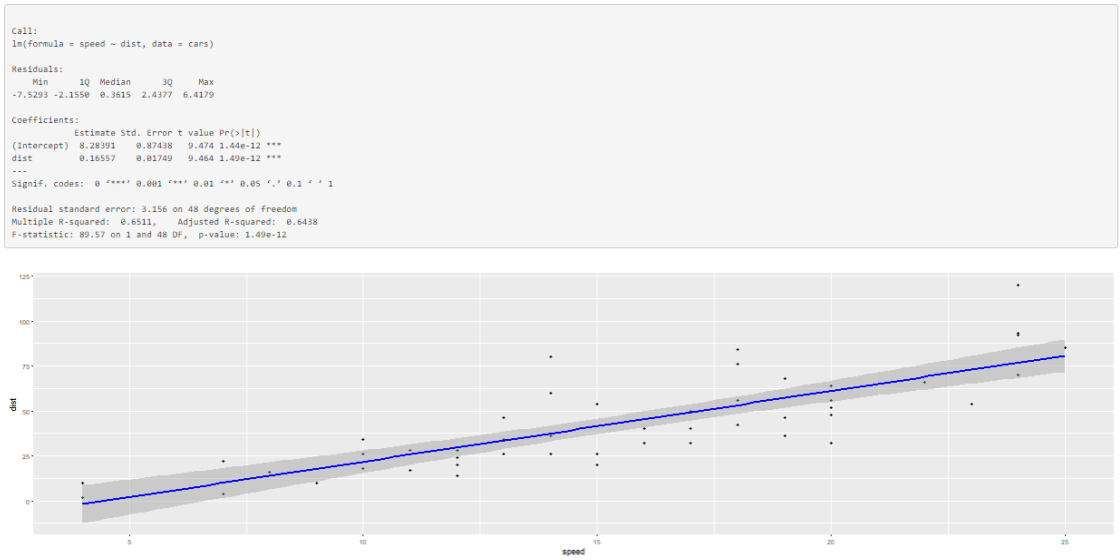
1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

```
     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
```

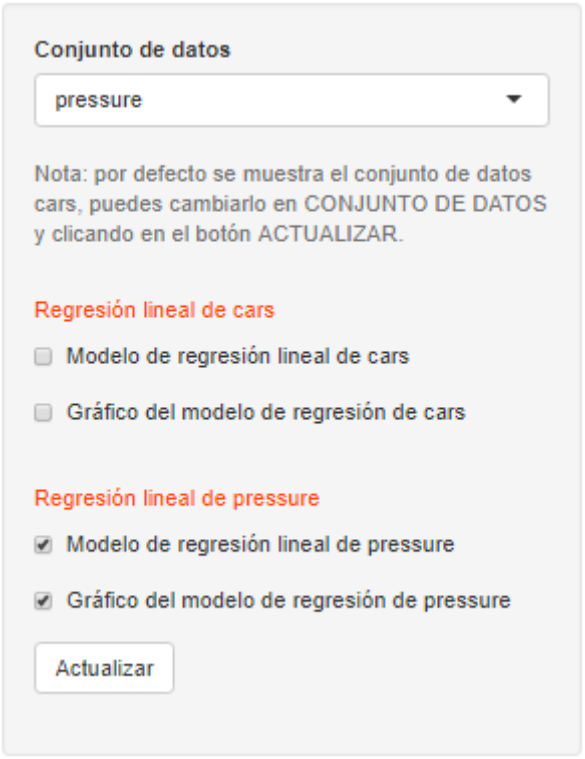
2.- El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente y un conjunto de variables independientes X1,..., Xn.

Al clicar en el checklist de "*Regresión lineal de cars*" nos muestra el análisis de regresión lineal de los datos así como el gráfico generado por ella:





De forma análoga para el otro conjunto de datos, cuando se hace clic en el checklist de “*Regresión lineal de pressure*” nos muestra tanto el análisis de regresión lineal de este conjunto de datos como el gráfico generado por ella:

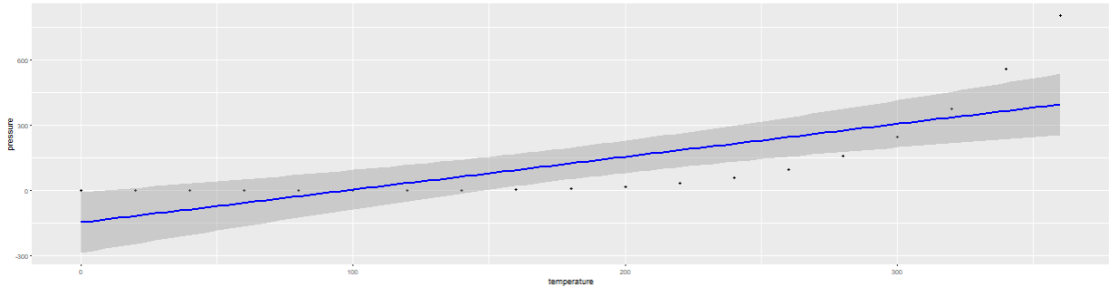


```
Call:
lm(formula = temperature ~ pressure, data = pressure)

Residuals:
     Min       1Q   Median       3Q      Max
-132.791  -62.813    6.507   67.033   90.759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 132.79072   19.94314   6.658 4.03e-06 ***
pressure      0.37969    0.07929   4.788 0.000171 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.56 on 17 degrees of freedom
Multiple R-squared:  0.5742,    Adjusted R-squared:  0.5492
F-statistic: 22.93 on 1 and 17 DF,  p-value: 0.000171
```



3.5. Pestaña 6: Estadística multivariante

Al ir a la **pestaña Estadística multivariante** por defecto se presenta un “summary” del conjunto de datos *rock* y sus primeras 5 observaciones, aunque se puede cambiar el conjunto de datos haciendo clic en el combo y actualizando:

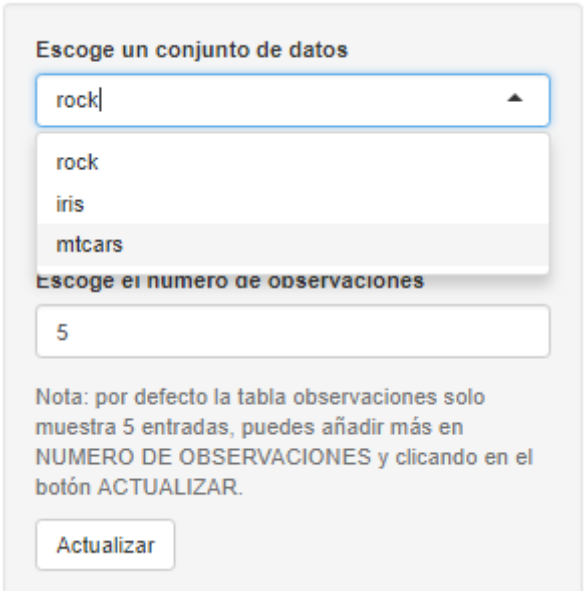
1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

```
      area            peri            shape             perm
 Min.   : 1016   Min.   : 308.6   Min.   :0.09033   Min.   :   6.30
 1st Qu.: 5305   1st Qu.:1414.9   1st Qu.:0.16226   1st Qu.:  76.45
 Median : 7487   Median :2536.2   Median :0.19886   Median : 130.50
 Mean   : 7188   Mean   :2682.2   Mean   :0.21811   Mean   : 415.45
 3rd Qu.: 8870   3rd Qu.:3989.5   3rd Qu.:0.26267   3rd Qu.: 777.50
 Max.   :12212   Max.   :4864.2   Max.   :0.46413   Max.   :1300.00
```

2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:

area	peri	shape	perm
4990	2791.90	0.09	6.30
7002	3892.60	0.15	6.30
7558	3930.66	0.18	6.30
7352	3869.32	0.12	6.30
7943	3948.54	0.12	17.10

Aunque se puede cambiar el conjunto de datos haciendo clic en el combo y actualizando:



Y de esta forma podemos visualizar el “summary” del conjunto de datos iris y sus primeras 5 observaciones:

1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.10	3.50	1.40	0.20	setosa
4.90	3.00	1.40	0.20	setosa
4.70	3.20	1.30	0.20	setosa
4.60	3.10	1.50	0.20	setosa
5.00	3.60	1.40	0.20	setosa

Y de igual forma con el conjunto de datos mtcars:

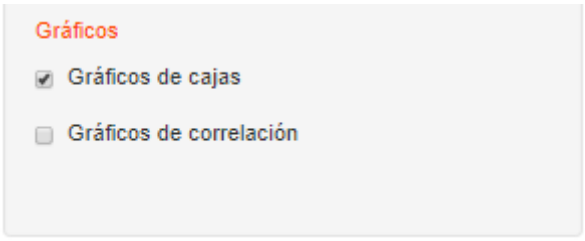
1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:

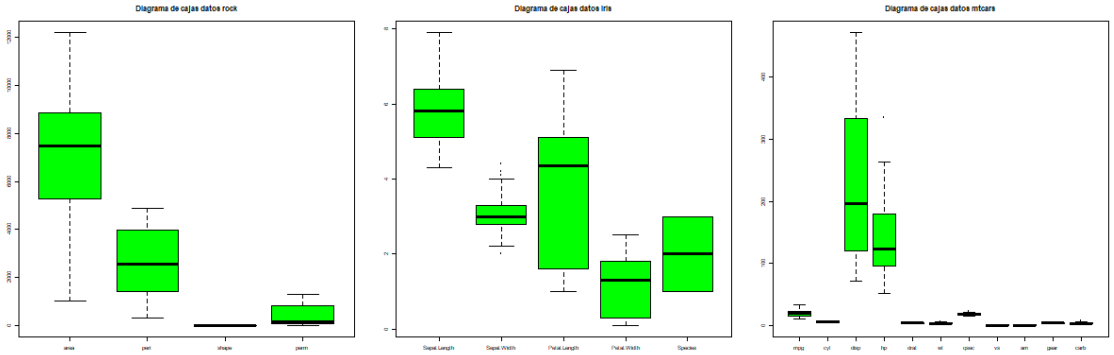
```
      mpg             cyl             disp             hp             drat             wt             qsec             vs              am              gear            carb
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71   Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:

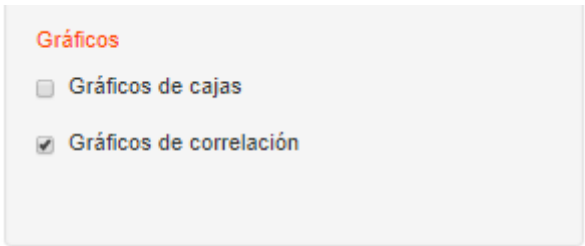
mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00

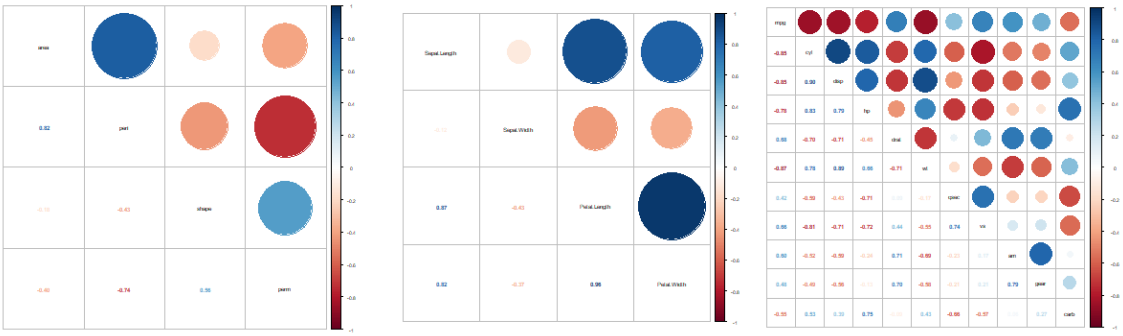
Al hacer clic en el checklist de “*Gráficos de cajas*” nos muestra estos gráficos para los tres conjuntos de datos considerados:





Y cuando clicamos en el checklist de “*Gráficos de correlación*” se obtienen estos tipos de gráficos para los tres conjuntos de datos analizados:





3.6. Archivos y alojamiento web de EDE

La aplicación web **EDE 1.0** se encuentran alojados en la siguiente URL:

https://arrocar.shinyapps.io/EDE_1/

Aunque toda la programación de **EDE 1.0** está en los anexos 6.18 y 6.19 también se ha decidido insertarlos aquí por si se quieren descargar directamente. En concreto son los archivos **ui.R** y **server.R**:

ui.R

server.R

4. Discusión y conclusiones

El objetivo de este proyecto era desarrollar una aplicación en la cual una persona que se quiera adentrar en el mundo del estudio y análisis de los datos tuviera una herramienta donde comprender de forma fácil y escueta los principales estadísticos y gráficos de la Estadística Descriptiva sin "perderse" en innumerables fórmulas y funciones del lenguaje de programación *R*.

Para poder lograr este objetivo lo primero que he tenido que hacer es estudiar y comprender la forma en la que actúa e interactúa *Shiny*. Para ello ha sido fundamental leer a conciencia el manual de Shiny: https://shiny.rstudio.com/.

Los resultados obtenidos al realizar la **aplicación EDE** han sido satisfactorios ya que se ha logrado explicar en esta aplicación todos las medidas y gráficos estadísticos básicos para el análisis de un conjunto de datos, es decir, un usuario sin experiencia y utilizando **EDE** puede comprender de forma fácil e intuitiva la Estadística Descriptiva.

A pesar del éxito que espero que tenga **EDE** una vez publicada en GitHub y dándola a conocer, tengo que reconocer que tiene algunas limitaciones o debilidades, las cuales quiero solucionar en un futuro. Éstas son:

- **Estadística univariante:** el usuario debe poder introducir sus propios valores para poder hacer pruebas y análisis estadísticos y no ceñirse únicamente a los conjuntos de datos que he establecido.
- **Estadística bivariante:** el usuario debe poder subir sus propios archivos en .csv. o .txt para poder realizar sus propias pruebas estadísticas con conjuntos de datos diferentes a los establecidos.
- **Estadística multivariante:** el usuario debe poder seleccionar las diferentes variables de los conjuntos de datos para poder realizar gráficos de densidad, dispersión, etc…

Dicho esto la aplicación **EDE 1.0** debe actualizarse a la versión **EDE 2.0** con los cambios mencionados anteriormente para que el proyecto sea completo.

5.- Referencias bibliográficas y Webgrafía

[1] Comtois, Domingo. 2018. "Introducción a las herramientas de resumen". https://cran.r-project.org/web/packages/summarytools/vignettes/Introduction.html .

[2] Elusa, Paula. 2009. "¿EXISTE VIDA MÁS ALLÁ DEL SPSS? DESCUBRE R." Revista Psicothema 21 (4): 652–55. http://www.psicothema.com/psicothema.asp?id=3686 .

[3] Field, Andy, Jeremy Miles y Zoe Field. 2012. Descubriendo Estadísticas Usando R . Edición: 1. Londres; Thousand Oaks, California: SAGE Publications Ltd.

[4] Grolemund, Garret. 2014. "Introducción a R Markdown". https://rmarkdown.rstudio.com/articles_intro.html .

[5] Ritchey, Ferris J. 2008. Estadística Para Las Ciencias Sociales. McGraw-Hill Interamericana de España SL

[6] Wickham, Hadley. 2017. "Tidyverse: instale y cargue fácilmente el 'Tidyverse'". Proyecto CRAN-R. https://CRAN.R-project.org/package=tidyverse .

[7] Wilkinson, Leland, D. Wills y D. Rope. 2005. La gramática de los gráficos. Edición: 2ª edición. 2005. Nueva York: Springer.

[8] Taller, Investigación Reproducible. 2016. "Escribir publicaciones con R." Escribiendo Publicaciones con R . http://www.geo.uzh.ch/microsite/reproducible_research/post/rr-r-publication/ .

[9] **https://shiny.rstudio.com/**

[10] **https://shiny.rstudio.com/gallery/**

[11] **https://personal.ua.es/es/julio-mulero/**

[12] **http://shiny.dmat.ua.es:3838/apps/shinyest/T3_unidimensional/**

[13] **https://www.universoformulas.com/estadistica/descriptiva/**

[14] **https://r-charts.com/es/**

6.- Anexos

6.1.- Anexo 1: Estadística Descriptiva

¿Qué es la Estadística Descriptiva?

La **Estadística Descriptiva** es la rama de la estadística que recolecta, analiza y caracteriza un conjunto de datos (por ejemplo, peso de la población, beneficios diarios de una empresa, temperatura mensual,…) con el objetivo de **describir** las características y comportamientos de este conjunto mediante **medidas de resumen, tablas y/o gráficos**.

6.2.- Anexo 2: Variables estadísticas

Una **variable estadística** es el conjunto de valores que puede tomar cierta característica de la población sobre la que se realiza el estudio estadístico y sobre la que es posible su medición. Estas variables pueden ser: la edad, el peso, las notas de un examen, los ingresos mensuales, las horas de sueño de un paciente en una semana, el precio medio del alquiler en las viviendas de un barrio de una ciudad, etc.

Las **variables estadísticas** se pueden clasificar por diferentes criterios. Según su medición existen dos tipos de variables:

- **Cualitativa** (o categórica): son las variables que pueden tomar como valores cualidades o categorías.
    - *Sexo (hombre, mujer). Nominal*
    - *Salud (buena, regular, mala). Ordinal*

- **Cuantitativas** (o numérica): variables que toman valores numéricos.
    - *Número de casas (1, 2,…). Discreta.*
    - *Temperatura (12.5; 24.3; 35.2,…). Continua.*

6.3.- Anexo 3: Medidas de posición central

Las **medidas de tendencia central** (o de centralización) son medidas que tienden a localizar en qué punto se encuentra la **parte central** de un **conjunto ordenado** de datos de una variable cuantitativa.

6.3.1. Media

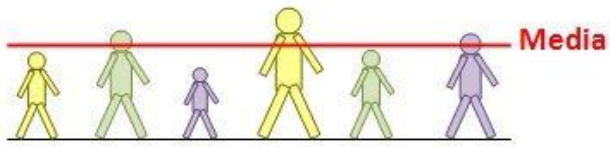
Definimos **media** (también llamada **promedio** o **media aritmética**) de un conjunto de datos (X_1, X_2,…, X_N) al valor característico de una serie de datos resultado de la suma de todas las observaciones dividido por el número total de datos.

$$Media(X) = \overline{x} = \frac{\sum_{i=1}^{N} X_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de observaciones

Es decir:

$$Media(X) = \overline{x} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$



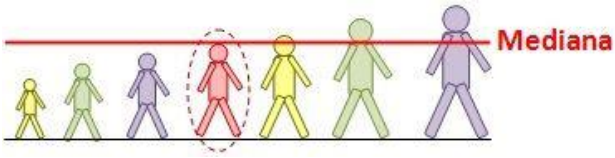
Visto desde un punto de vista más conceptual, la **media aritmética** es el centro de los datos en el sentido numérico, ya que intenta equilibrarlos por exceso y por defecto. Es decir, si sumamos todas las diferencias de los datos a la media da cero.

$$\sum_{i=1}^{N}(x_i - \overline{x}) = 0$$

La media no es representativa con datos atípicos y en ese caso es mejor utilizar la mediana, porque la media no siempre es el valor central, pero la mediana sí.

6.3.2. Mediana

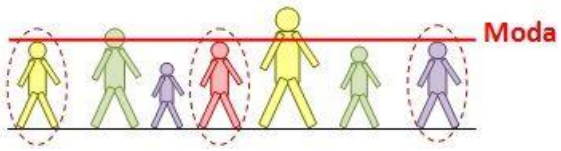
La **mediana Me(X)** es el elemento de un conjunto de datos ordenados (X_1, X_2, …, X_N) que **deja a izquierda y derecha la mitad de los valores**.



Es decir, una vez que se ordenan los datos, la **mediana** es el valor del conjunto tal que el 50% de los elementos son menores o iguales y el otro 50%, mayores o iguales.

6.3.3. Moda

La **moda Mo(X)** es el valor más repetido del conjunto de datos, es decir, el valor cuya frecuencia relativa es mayor. En un conjunto puede haber más de una moda, al contrario de lo que sucede con la media y la mediana.



6.4.- Anexo 4: Medidas de posición no central

Las **medidas de posición no central** (o **medidas de tendencia no central**) permiten conocer **puntos característicos** de una serie de valores, que no necesariamente tienen que ser centrales. La intención de estas medidas es **dividir el conjunto** de observaciones en grupos con el mismo número de valores en cada uno.

6.4.1. Cuartiles

Los **cuartiles** son los **tres valores** que dividen una serie de datos ordenada en cuatro partes iguales. El **primer cuartil (Q1)** deja a la izquierda el 25% de los datos. El **segundo (Q2)** deja a izquierda y derecha el 50% y coincide con la mediana. El **tercero (Q3)** deja a la derecha el 25% de valores. Los tres cuartiles son:

$$Cuartil_1 = Q_1 = X_{((N+1)/4)}$$
$$Cuartil_2 = Q_2 = Mediana(X) = X_{((N+1)/2)}$$
$$Cuartil_3 = Q_3 = X_{(3(N+1)/4)}$$

siendo $(X_1, X_2, ..., X_N)$ la serie de datos ordenada

6.4.2. Percentiles

Los **percentiles** P_i son los **99 puntos** que dividen una serie de datos ordenada en **100 partes iguales**, es decir, que contienen el mismo número de elementos cada una. El percentil 50 es la mediana.

Sea (X_1, X_2,…, X_N) una muestra de *N* elementos. El **percentil P_i** es:

$$P_i = X_{((N+1)\cdot i)/100}$$

siendo N el número de elementos del conjunto y $0 \leq i \leq 99$

Donde P_i es la posición del percentil buscado en la serie ordenada de datos.

Los **percentiles** están pensados para conjuntos de elementos de más de cien elementos.

6.5.- Anexo 5: Medidas de dispersión

Las **medidas de dispersión** o **medidas de variabilidad** muestran la **variabilidad** de un conjunto de datos, indicando la mayor o menor concentración de datos respecto a las medidas de centralización.

6.5.1. Rango

El **rango R** o recorrido estadístico es la diferencia entre el valor máximo y el mínimo de un conjunto de elementos.

$$Rango = (Max) - (Min)$$

6.5.2. Rango intercuartílico

El **rango intercuartílico IQR** (o **rango intercuartil**) es una estimación estadística de la dispersión de una distribución de datos. Consiste en la diferencia entre el **tercer** y el **primer** cuartil. Mediante esta medida se eliminan los valores extremadamente alejados. El rango intercuartílico es altamente recomendable cuando la medida de tendencia central utilizada es la mediana (ya que este estadístico es insensible a posibles irregularidades en los extremos).

$$IQR = Q_3 - Q_1$$

En una distribución, encontramos la mitad de los datos, el 50 %, ubicados dentro del rango intercuartílico.

Conforme aumente el **IQR**, indicará que la dispersión será mayor.

6.5.3. Varianza

La **varianza (S^2)** mide la **dispersión** de los datos de una muestra respecto a la media, calculando la media de los cuadrados de las distancias de todos los datos.

$$S_X^2 = \frac{\Sigma_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$ la media

Al elevar las diferencias al cuadrado se garantiza que las diferencias absolutas respecto a la media no se anulan entre sí. Además, resaltan los valores alejados.

6.5.4. Desviación típica

La **desviación típica** es la medida de dispersión (S) asociada a la media. Mide el promedio de las desviaciones de los datos respecto a la media en las mismas unidades de los datos.

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{N-1}}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos

El cuadrado de la desviación típica es la varianza.

6.5.5. Desviación media

La **desviación media** es la media de los valores absolutos de la diferencia de cada valor de la distribución con la media aritmética.

Su fórmula es:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{N}|X_i - \bar{x}|}{N}$$

Cuando los datos están agrupados en frecuencias:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{N}|X_i - \bar{x}| \cdot n_i}{N}$$

n_i, frecuencia del dato x_i

De forma similar se definiría la media, la varianza y la desviación típica cuando los datos se agrupan en frecuencias.

La desviación media es igual o menor que la desviación estándar:

$$D_{\bar{x}} \leq \sigma$$

No confundir la desviación media con la desviación absoluta de un dato respecto a la media:

$$D_{Mi} = |X_i - \bar{x}|$$

6.5.6. Coeficiente de variación de Pearson

El **coeficiente de variación de Pearson (r)** mide la **variación de los datos** respecto a la media, sin tener en cuenta las unidades en la que están.

$$r = \frac{S_X}{|\bar{x}|}$$

siendo S_X la desviación típica y $\bar{x}$ la media del conjunto de observaciones $(X_1, X_2, ..., X_N)$ y $\bar{x} \neq 0$

El **coeficiente de variación** toma valores entre **0 y 1**. Si el coeficiente es próximo al 0, significa que existe poca variabilidad en los datos y es una muestra muy compacta. En cambio, si tienden a 1 es una muestra muy dispersa y la media pierde confiabilidad. De hecho, cuando el coeficiente de variación supera el 30% (0,3) se dice que la media es poco representativa.

Para interpretar fácilmente el coeficiente, podemos multiplicarlo por cien para tenerlo en tanto por cien.

El coeficiente de variación es muy utilizado para comparar la dispersión de varias distribuciones, de modo que la tenga el menor coeficiente de variación tendrá menos dispersión.

6.6.- Anexo 6: Asimetría y curtosis

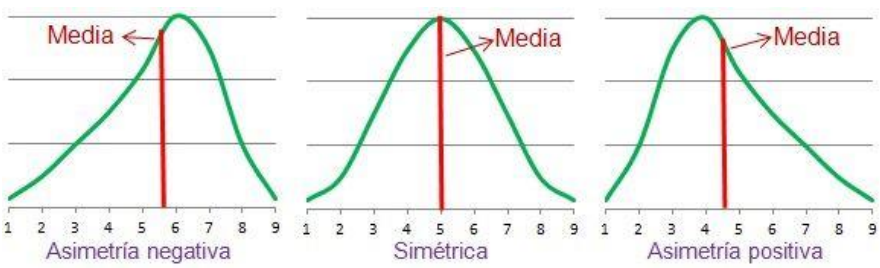
La **asimetría y curtosis** informan sobre la forma de la distribución de una variable. Estas medidas permiten saber las características de su asimetría y homogeneidad sin necesidad de representar los datos gráficamente.

6.6.1. Asimetría

La **asimetría** es la medida que indica la simetría de la distribución de una variable respecto a la media aritmética, sin necesidad de hacer la representación gráfica. Los coeficientes de asimetría indican si hay el mismo número de elementos a izquierda y derecha de la media.

Existen **tres tipos** de curva de distribución según su **asimetría**:

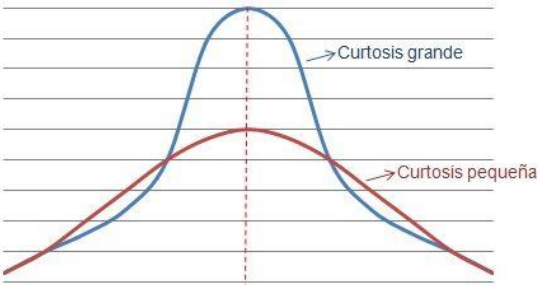
- **Asimetría negativa**: la cola de la distribución se alarga para valores inferiores a la media. En este caso, el coeficiente de asimetría es negativo
- **Simétrica**: hay el mismo número de elementos a izquierda y derecha de la media. En este caso, coinciden la media, la mediana y la moda. La distribución se adapta a la forma de la campana de Gauss, o distribución normal que se corresponde con un coeficiente de asimetría igual a cero
- **Asimetría positiva**: la cola de la distribución se alarga para valores superiores a la media que proporciona un coeficiente de asimetría positivo.



6.6.2. Curtosis

La **curtosis** (o **apuntamiento**) es una medida de forma que mide cuán escarpada o achatada está una curva o distribución.

Este coeficiente indica la cantidad de datos que hay cercanos a la media, de manera que a **mayor grado de curtosis, más escarpada** (o apuntada) será la forma de la curva.



La **curtosis** se mide promediando la cuarta potencia de la diferencia entre cada elemento del conjunto y la media, dividido entre la desviación típica elevada también a la cuarta potencia. Sea el conjunto $X=(X_1, X_2,\ldots, X_N)$, entonces el **coeficiente de curtosis** será:

$$g_2 = \frac{\sum_{i=1}^{N}(X_i - \bar{x})^4 \cdot n_i}{N S_x^4}$$

n_i la frecuencia absoluta de x_i

6.7.- Anexo 7: Frecuencias

La **frecuencia** es una medida que sirve para comparar la aparición de un elemento X_i en un conjunto de elementos (X_1, X_2,…, X_N). Mediante tablas de distribuciones de frecuencia se puede presentar organizadamente el recuento de datos.

Las **frecuencias** de cada elemento se pueden expresar tanto como frecuencias **absolutas** (número total de apariciones) como frecuencias **relativas** (proporción de apariciones).

6.7.1. Frecuencia absoluta

La **frecuencia absoluta** (n_i) de un valor X_i es el número de veces que el valor está en el conjunto (X_1, X_2,…, X_N).

La **suma** de las **frecuencias absolutas** de todos los elementos diferentes del conjunto debe ser el número total de sujetos *N*. Si el conjunto tiene *k* valores (o categorías) diferentes, entonces:

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k = N$$

6.7.2. Frecuencia absoluta acumulada

La **frecuencia absoluta acumulada** (N_i) de un valor X_i del conjunto (X_1, X_2,…, X_N) es la suma de las frecuencias absolutas de los valores menores o iguales a X_i, es decir:

$$N_i = n_1 + n_2 + \ldots + n_i$$

6.7.3. Frecuencia relativa

La **frecuencia relativa** **(f_i)** de un valor X_i es la **proporción** de valores iguales a X_i en el conjunto de datos (X_1, X_2,…, X_N). Es decir, la frecuencia relativa es la frecuencia absoluta dividida por el número total de elementos *N*:

$$f_i = \frac{n_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de datos
y n_i el total de valores igual a X_i

Las **frecuencias relativas** son valores entre 0 y 1, $0 \leq f_i \leq 1$. La suma de las frecuencias relativas de todos los sujetos da 1. Supongamos que en el conjunto tenemos *k* valores (o categorías) diferentes, entonces:

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \ldots + f_k = 1$$

Si se multiplica la frecuencia relativa por cien se obtiene el **porcentaje** (tanto por cien %).

6.7.4. Frecuencia relativa acumulada

Definimos la **frecuencia relativa acumulada** **(F_i)** de un valor X_i como la **proporción** de valores iguales o menores a X_i en el conjunto de datos (X_1, X_2,…, X_N). Es decir, la frecuencia relativa acumulada es la frecuencia absoluta acumulada dividida por el número total de sujetos *N*:

$$F_i = \frac{N_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de datos
y N_i el total de valores igual o menor a X_i

La **frecuencia relativa acumulada** de cada valor siempre es mayor que la frecuencia relativa. De hecho, la frecuencia relativa acumulada de un elemento es la suma de las frecuencias relativas de los elementos menores o iguales a él, es decir:

$$F_i = f_1 + f_2 + \ldots + f_i$$

6.8.- Anexo 8: Gráficos

Un **gráfico** (o **gráfica**) es el recurso de representar los datos numéricos por medio de líneas, diagramas, dibujos, etc. La representación gráfica es un importante suplemento al análisis y estudio estadístico. Permite visualizar de forma gráfica la distribución de los datos.

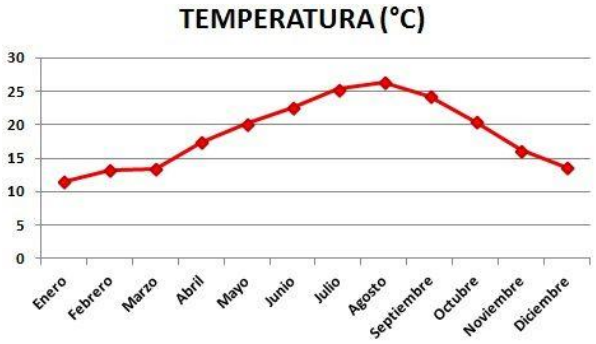
Existen muchas clases de gráficas dependiendo principalmente del tipo de variables de estudio, objetivo del análisis, … Se pueden destacar los siguientes **tipos**:

6.8.1. Gráfico lineal

El **gráfico lineal** (gráfico de líneas o **diagrama lineal**) se compone de una serie de datos representados por puntos, unidos por segmentos lineales.

Mediante este gráfico se puede comprobar rápidamente el cambio de tendencia de los datos.

El **diagrama lineal** se suele utilizar con variables cuantitativas, para ver su comportamiento en el transcurso del tiempo. Por ejemplo, en las **series temporales** mensuales, anuales, trimestrales, etc.



6.8.2. Diagramas

Un **diagrama** es un tipo de representación gráfica que sirve para representar un conjunto de datos.

Existen diferentes **tipos de diagramas**, entre los que se pueden destacar los siguientes:

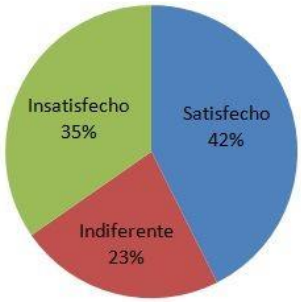
6.8.2.1. Diagrama de barras

El **diagrama de barras** es un gráfico que se utiliza para representar datos de variables cualitativas o cuantitativas discretas. Está formado por **barras** rectangulares cuya altura es proporcional a la frecuencia de cada uno de los valores de la variable.

6.8.2.2. Diagrama circular

El **diagrama circular** (también llamado **diagrama de sectores** o **diagrama de pastel o de tarta**) sirve, al igual que el diagrama de barras, para representar variables cualitativas o cuantitativas discretas. Se utiliza para representar la proporción de elementos de cada uno de los valores de la variable.

Consiste en partir el círculo en porciones proporcionales a la frecuencia relativa. Entiéndase como porción la parte del círculo que representa a cada valor que toma la variable.

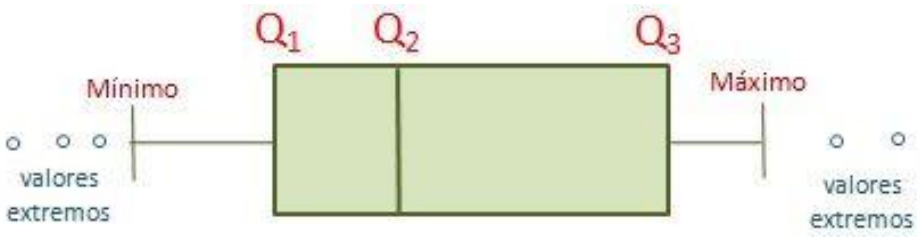


Es un gráfico útil cuando la variable tiene pocos valores o categorías diferentes

6.8.2.3. Diagrama de caja

El **diagrama de caja** es un gráfico utilizado para representar una variable cuantitativa (variable numérica). Este gráfico es una herramienta que permite visualizar, a través de los cuartiles, cómo es la distribución, su grado de asimetría, los valores extremos, la posición de la mediana, etc. Se compone de:

- Un rectángulo (*caja*) delimitado por el **primer y tercer cuartil** (Q_1 y Q_3). Dentro de la caja una línea indica dónde se encuentra la mediana (segundo cuartil Q_2)
- Dos *brazos* o *bigotes*, uno que empieza en el primer cuartil y acaba en el **mínimo**, y otro que empieza en el tercer cuartil y acaba en el **máximo**.
- Los **datos atípicos** (o valores extremos) que son los valores distintos que no cumplen ciertos requisitos de heterogeneidad de los datos.



6.8.2.4. Diagrama de tallo y hojas

El **diagrama de tallo y hojas** ("Stem-and-Leaf Diagram") es un semigráfico que permite presentar la distribución de una variable cuantitativa. Consiste en separar cada dato en el último dígito (que se denomina **hoja**) y las cifras delanteras restantes (que forman el **tallo**).

Es especialmente útil para conjuntos de datos de tamaño medio (**entre 20 y 50 elementos**) y que sus datos no se agrupan alrededor de un único tallo. Con él podemos hacernos la idea de qué distribución tienen los datos, la asimetría, etc.

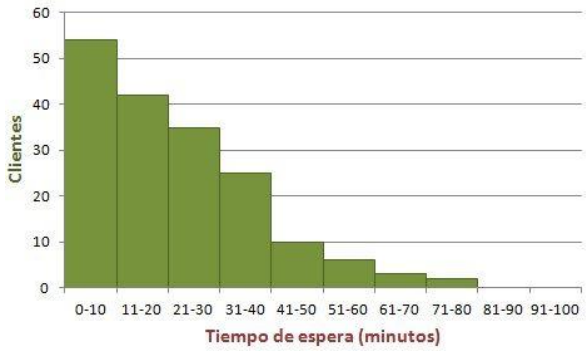
Tallo	Hoja
4	4 5 9
5	0 2 3 3 4 4 6 7 7 7 8
6	1 2 2 3 4 7 8 9
7	0 1 1 2 3 4 4 5 6 6 8 9
8	0 1 3 5

6.8.3. Histograma

Un **histograma** es una representación gráfica de **datos agrupados** mediante intervalos. Los datos provienen de una variable cuantitativa continua. Gracias a él se puede hacer uno rápidamente una idea de la distribución de los datos o muestra.

También se puede emplear para variables cualitativas ordinales, siendo necesario que el número de datos sea elevado.

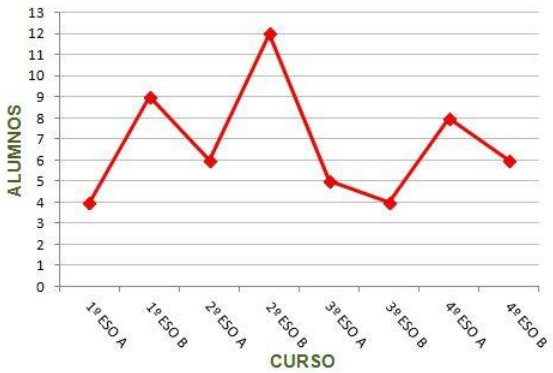
Un **histograma** es un conjunto de **rectángulos** que representan las frecuencias absolutas de cada uno de los intervalos. Los **intervalos** abarcan todo el conjunto sin solaparse, de manera que un elemento está solo en un intervalo.



6.8.4. Polígono de frecuencias

El **polígono de frecuencias** es un gráfico que permite la rápida visualización de las frecuencias de cada una de las categorías de la variable de estudio.

Normalmente se utiliza el polígono de frecuencias con frecuencias absolutas, pero también se utiliza con frecuencias relativas, porcentajes o frecuencias acumuladas.



6.9.- Anexo 9: R y RStudio

R es un lenguaje de programación de código abierto orientado a objetos desarrollado para el análisis estadístico de datos, usado principalmente en el ámbito de la investigación matemática y machine learning, minería de datos, etc.… Es multiplataforma, por lo que se puede usar en cualquier sistema operativo de escritorio.

Por su parte, ***RStudio*** es un entorno de desarrollo remoto, que se instala comúnmente en un servidor Linux y que permite manejar y ejecutar proyectos en R de manera remota, sin tener que instalar nada en el ordenador del usuario.

6.9.1. Funcionalidades principales de RStudio

RStudio ofrece todas las herramientas que podemos esperar de un **IDE moderno**, como coloreado de sintaxis, ayudas para completado y formateado de código. Ofrece además una plataforma de ejecución para los programas escritos en R, de modo que se pueden poner en marcha de manera cómoda, online y sin salir de la propia aplicación.

El **entorno de desarrollo** integra diversas herramientas adicionales dentro del espacio de trabajo, como la documentación del lenguaje R, sistemas de control de versiones (Git y otros), la gestión de proyectos y visualización de datos, así como un depurador que permite localizar y corregir errores en el código fácilmente. Además, se puede extender por medio de **packages** adicionales en función de las necesidades de los profesionales. Todo ello funciona en el navegador y por tanto es accesible desde cualquier lugar, simplemente disponiendo de un acceso a Internet, lo que permite el trabajo en remoto y la disponibilidad de las herramientas de análisis de datos, así como cualquiera de los archivos usados, desde cualquier lugar.

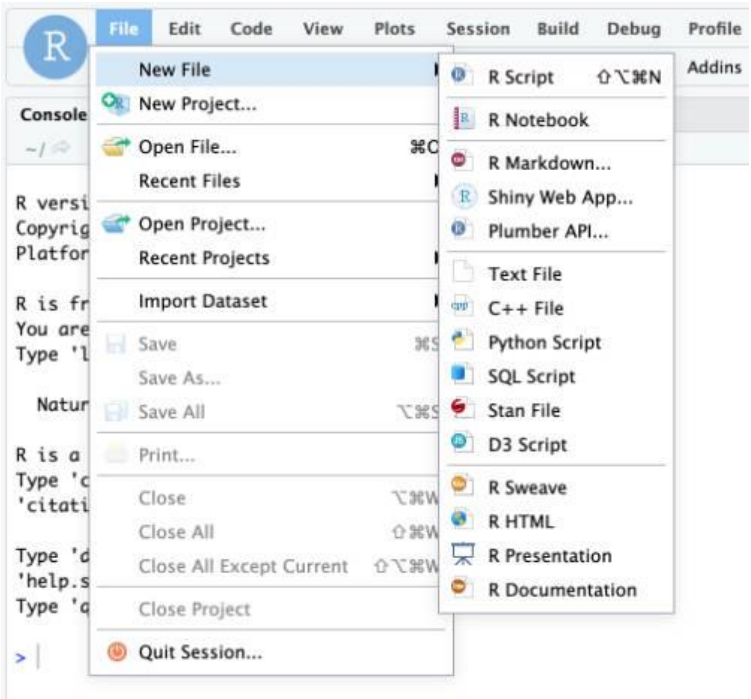
Dentro de la categoría del **machine learning**, ofrece un entorno de desarrollo completamente accesible desde el navegador desde el que puedes fácilmente

desarrollar y depurar código y organizar tus documentos en proyectos. La interfaz es una plataforma para análisis y cálculo para proyectos con grandes cantidades de datos o con funciones matemáticas complejas.

6.9.2. Primeros pasos con RStudio

RStudio permite administrar proyectos en los que se trabaja con múltiples tipos de archivos de código, entre los que encontramos R scripts, Documentos R Markdown, archivos HTML o TeX, y muchos otros.

Para comenzar vamos a localizar el menú **«File»**. Desde aquí se pueden crear archivos de **«R»** así como proyectos. Si lo deseamos podemos abrir varios archivos a la vez, que podremos seleccionar por medio de unas pestañas dentro de la interfaz del programa.



6.10.- Anexo 10: Tipos de objetos en R

En ***R*** existen varios tipos de objectos que permiten que el usuario pueda almacenar la información para realizar procedimientos estadísticos y gráficos.

Los principales objetos en R son vectores, matrices, arreglos, marcos de datos y listas. A continuación se presentan las características de estos objetos y la forma para crearlos.

6.10.1. Variable

Las **variables** sirven para almacenar un valor o valores que luego vamos a utilizar en algún procedimiento.

Para hacer la asignación de un valor o valores a alguna variable se utiliza el **operador <-** entre el valor y el nombre de la variable.

6.10.2. Vector

Los **vectores** son objetos ordenados en los cuales se puede almacenar información de tipo numérico (variable cuantitativa), alfanumérico (variable cualitativa) o lógico (TRUE o FALSE), pero no mezclas de éstos. Una de las funciones (o la función principal) de R para crear un vector es **c()** y significa concatenar; dentro de los paréntesis de esta función se ubica la información a almacenar. Una vez construido el vector se acostumbra a etiquetarlo con un nombre corto y representativo de la información que almacena, la asignación se hace por medio del operador **<-** entre el nombre que le damos al vector y su contenido.

6.10.3. Matriz

Las **matrices** son objetos rectangulares de filas y columnas con información numérica, alfanumérica o lógica. Para construir una matriz se usa la función **matrix( ).** Por ejemplo, para crear una matriz de 4 filas y 5 columnas (de dimensión 4×5) con los primeros 20 números positivos se escribe el código siguiente en la consola:

```
mimatriz <- matrix(data=1:20, nrow=4, ncol=5, byrow=FALSE)
```

El argumento **data** de la función sirve para indicar los datos que se van a almacenar en la matriz, los argumentos **nrow** y **ncol** sirven para definir la dimensión de la matriz, indicando el número de filas y el número de columnas, respectivamente, y por último el argumento **byrow** sirve para indicar si la información contenida en data se debe ingresar por filas o no. Para observar lo que quedó almacenado en el objeto **mimatriz** se escribe en la consola el nombre del objeto seguido de la tecla enter o intro.

6.10.4. Array

Un **array** o **arreglo** es una matriz de varias dimensiones con información numérica, alfanumérica o lógica. Para construir una arreglo se usa la función **array( )**. Por ejemplo, para crear un array de 3×4×2 con las primeras 24 letras minúsculas del alfabeto se escribe el siguiente código:

```
miarray <- array(data=letters[1:24], dim=c(3, 4, 2))
```

El argumento **data** de la función sirve para indicar los datos que se van a almacenar en el array y el argumento **dim** sirve para indicar las dimensiones del array. Para observar lo que quedó almacenado en el objeto **miarray** se escribe en la consola lo siguiente:

```
miarray
```

```
## , , 1
##
##      [,1] [,2] [,3] [,4]
## [1,] "a"  "d"  "g"  "j"
## [2,] "b"  "e"  "h"  "k"
## [3,] "c"  "f"  "i"  "l"
##
## , , 2
##
##      [,1] [,2] [,3] [,4]
## [1,] "m"  "p"  "s"  "v"
## [2,] "n"  "q"  "t"  "w"
## [3,] "o"  "r"  "u"  "x"
```

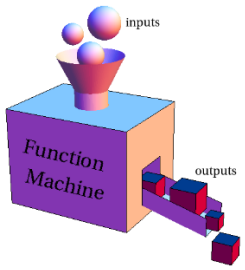
6.10.5. Data Frame

El **data frame** es uno de los objetos más utilizados porque permite agrupar vectores con información de diferente tipo (numérica, alfanumérica o lógica) en un mismo objeto, la única restricción es que los vectores deben tener la misma longitud. Para crear un marco de datos o data frame se usa la función **data.frame( )**. Como ejemplo vamos a crear un marco de datos con los vectores edad, deporte y comic_fav que ya fueran definidos previamente en la consola:

```
mimarco <- data.frame(edad, deporte, comic_fav)
```

6.11.- Anexo 11: Funciones básicas de R

En la figura de abajo se muestra una ilustración de lo que es una función o máquina general. Hay unas **entradas** (inputs) que luego son procesadas dentro de la caja para generar unas **salidas** (outputs). Un ejemplo de una función o máquina muy común en nuestras casas es la licuadora. Si a una licuadora le ingresamos leche, fresas, azúcar y hielo, el resultado será un delicioso jugo de fresa.



Las **funciones en *R*** se caracterizan por un nombre corto que dé una idea de lo que hace la función. Los elementos que se le pueden ingresar (inputs) a la función se llaman **parámetros** o **argumentos** y se ubican dentro de paréntesis; el cuerpo de la función se ubica dentro de llaves y es ahí donde se procesan los inputs para convertirlos en outputs. A continuación se muestra la estructura general de una función.

```
nombre_de_funcion(parametro1, parametro2, ...) {
  tareas internas
  tareas internas
  tareas internas
  salida
}
```

Cuando usamos una función sólo debemos escribir bien el nombre e ingresar correctamente los parámetros de la función, mientras que el cuerpo de la función ni lo vemos ni lo debemos modificar. A continuación se presenta un ejemplo de cómo usar la función **mean()** para calcular un promedio, en este caso de las notas de un estudiante:

```
notas <- c(4.0, 1.3, 3.8, 2.0)  # Notas de un estudiante
mean(notas)
```

```
## [1] 2.775
```

6.11.1. Operadores de asignación

En ***R*** se puede hacer asignación de varias formas. A continuación se presentan los operadores disponibles para tal fin.

- <- : este es el operador de asignación a izquierda, es el más usado y recomendado.
- -> : este es el operador de asignación a derecha, no es frecuente su uso.
- = : el símbolo igual sirve para hacer asignaciones pero NO se recomienda usarlo.
- <<- : este es un operador de asignación global y sólo debe ser usado por usuarios avanzados.

6.11.2. Operaciones básicas

En ***R*** se pueden hacer diversas operaciones usando operadores binarios. Este tipo de operadores se denominan binarios porque actúan entre dos objetos. A continuación se presenta el listado de tales operadores:

- + : operador binario para sumar.
- - : operador binario para restar.
- * : operador binario para multiplicar.
- / : operador binario para dividir.
- ^ : operador binario para potencia.
- %/% : operador binario para obtener el cociente en una división (número entero).
- %% : operador binario para obtener el residuo o resto en una división.

6.11.3. Pruebas lógicas

En ***R*** se puede verificar si un objeto cumple una condición dada. Para ello están las siguientes pruebas usuales:

- < : para saber si un número o valor es menor que otro.
- > : para saber si un número o valor es mayor que otro.
- == : para saber si un número o valor es igual que otro.
- <= : para saber si un número o valor es menor o igual que otro.
- >= : para saber si un número o valor es mayor o igual que otro.

6.11.4. Operadores lógicos

En el programa ***R*** están disponibles los operadores lógicos negación, conjunción y disyunción. A continuación se muestra el listado de los operadores lógicos entre los elementos (objetos) x e y:

```
!x  # Negación de x
x & y  # Conjunción entre x e y
x && y
x | y  # Disyunción entre x e y
x || y
xor(x, y)
```

6.11.5. Funciones sobre vectores

En **R** podemos destacar las siguientes funciones básicas sobre vectores numéricos:

- **min:** para obtener el mínimo de un vector.
- **max:** para obtener el máximo de un vector.
- **length:** para determinar la longitud de un vector.
- **range:** para obtener el rango de valores de un vector, entrega el mínimo y máximo.
- **sum:** entrega la suma de todos los elementos del vector.
- **prod:** multiplica todos los elementos del vector.
- **which.min:** nos entrega la posición en donde está el valor mínimo del vector.
- **which.max:** nos da la posición del valor máximo del vector.
- **rev:** invierte un vector.
- **rev:** ordena los elementos de un vector.

6.11.6. Funciones matemáticas

Otras funciones matemáticas básicas muy utilizadas, y concretamente en Estadística, son:

- **sin**
- **cos**
- **tan**

Funciones trigonométricas

- **log**
- **logb**
- **log10**

Funciones logarítmicas

- **exp:** función exponencial.
- **sqrt:** función de radicales.
- **abs:** Función de valor absoluto.

6.12.- Anexo 12: Medidas de posición central en R

6.12.1. Media

Para calcular la media de una variable cuantitativa se usa la **función mean**. Los argumentos básicos de la función mean son dos y se muestran a continuación.

```
mean(x, na.rm = FALSE)
```

El parámetro *x* indica la variable de interés para la cual se quiere calcular la media y el parámetro *na.rm* es un valor lógico que en caso de ser TRUE, significa que se deben remover las observaciones con NA, el valor por defecto para este parámetro es FALSE.

6.12.2. Mediana

Para calcular la mediana de una variable cuantitativa se usa la función median. Los argumentos básicos de la **función median** son dos y se muestran a continuación.

```
median(x, na.rm = FALSE)
```

El parámetro *x* indica la variable de interés para la cual se quiere calcular la mediana y el parámetro na.rm es un valor lógico que en caso de ser TRUE, significa que se deben remover las observaciones con NA, donde el valor por defecto para este parámetro es FALSE.

6.12.3. Moda

La moda de una variable cuantitativa corresponde al valor o valores que más se repiten, una forma sencilla de encontrar la moda es construir una tabla de frecuencias y observar los valores con mayor frecuencia.

Se construye la tabla con la **función table** y se crea por ejemplo el objeto tabla para almacenarla.

```
tabla <- table(datos$edad)
tabla
```

```
##
## 19 20 21 22 23 24 25 26 28 29 30 32 33 35 37 40 43 45 51 55 65
##  1  1  1  3  2  1  5  3  2  1  2  1  1  2  3  1  2  1  1  1  1
```

Para observar los valores con mayor frecuencia de la tabla se puede ordenar la tabla usando la **función sort** de la siguiente manera:

```
sort(tabla, decreasing=TRUE)
```

```
##
## 25 22 26 37 23 28 30 35 43 19 20 21 24 29 32 33 40 45 51 55 65
##  5  3  3  3  2  2  2  2  2  1  1  1  1  1  1  1  1  1  1  1  1
```

De esta manera se ve fácilmente que la variable edad es unimodal con moda el valor de 25 años.

6.13.- Anexo 13: Medidas de posición no central en R

6.13.1. Cuartiles

Para obtener cualquier cuantil (cuartiles, deciles y percentiles) se usa la **función quantile**. Los argumentos básicos de la función quantile son tres y se muestran a continuación.

```
quantile(x, probs, na.rm = FALSE)
```

El parámetro *x* indica la variable de interés para la cual se quieren calcular los cuantiles, el parámetro *probs* sirve para definir los cuantiles de interés y el parámetro na.rm es un valor lógico que en caso de ser TRUE, significa que se deben remover las observaciones con NA, con el valor por defecto para este parámetro es FALSE.

6.14.- Anexo 14: Medidas de dispersión en R

6.14.1. Rango

Para calcular el rango de una variable cuantitativa se usa la **función range**. Los argumentos básicos de la función range son dos y se muestran abajo.

```
range(x, na.rm = FALSE)
```

En el parámetro *x* se indica la variable de interés para la cual se quiere calcular el rango, mientras que el parámetro *na.rm* es un valor lógico que en caso de ser TRUE, significa que se deben remover las observaciones con NA, donde el valor por defecto para este parámetro es FALSE.

La función *range* entrega el valor mínimo y máximo de la variable que se ingresó. Para obtener el valor del rango se debe restar del valor máximo el valor mínimo.

6.14.2. Varianza

Para calcular la varianza muestral de una variable cuantitativa se usa la **función var**. Los argumentos básicos de la función var son dos que se explican a continuación:

```
var(x, na.rm = FALSE)
```

El parámetro *x* indica la variable de interés para la cual se quiere calcular la varianza muestral, y el parámetro *na.rm* es un valor lógico que en caso de ser TRUE, significa que se deben remover las observaciones con NA, con el valor por defecto para este parámetro igual a FALSE.

6.14.3. Desviación estándar

Para calcular en *R* la desviación muestral de una variable cuantitativa se usa la **función sd**. Los argumentos básicos de la función sd son dos y se muestran a continuación.

```
sd(x, na.rm = FALSE)
```

El parámetro *x* guarda la variable de interés para la cual se quiere calcular la desviación estándar muestral y el parámetro na.rm es un valor lógico que en

caso de ser TRUE, significa que se deben remover las observaciones con NA (valor por defecto es FALSE).

6.14.4. Coeficiente de variación

El coeficiente de variación se define como **CV=s/x** y es muy sencillo obtenerlo. La **función coef_var** mostrada abajo permite calcularlo.

```
coef_var <- function(x, na.rm = FALSE) {
  sd(x, na.rm=na.rm) / mean(x, na.rm=na.rm)
}
```

6.15.- Anexo 15: Medidas de correlación en R

La **función cor** permite calcular el coeficiente de correlación de Pearson, de Kendall o de Spearman para dos variables cuantitativas. La estructura de la función es la siguiente:

```
cor(x, y, use="everything",
    method=c("pearson", "kendall", "spearman"))
```

Los parámetros de la función son:

- **x, y:** vectores cuantitativos entre los que se desea hallar la correlación.
- **use:** parámetro que indica lo que se debe hacer cuando se presenten registros NA en alguno de los vectores. Las diferentes posibilidades son: everything, all.obs, complete.obs, na.or.complete y pairwise.complete.obs, y el valor por defecto es everything.
- **method:** tipo de coeficiente de correlación a calcular. Por defecto es pearson que indica el coeficiente de correlación lineal de Pearson, pero otros valores posibles son kendall y spearman.

6.16.- Anexo 16: Tablas de frecuencias en R

6.16.1. Tabla de frecuencias

La **función table** sirve para construir tablas de frecuencia de una vía, donde su estructura es la siguiente:

```
table(..., exclude, useNA)
```

Los parámetros de la función son:

- **...** : espacio para ubicar los nombres de los objetos (variables o vectores) para los cuales se quiere construir la tabla.
- **exclude:** vector con los niveles a remover de la tabla. Si exclude=NULL implica que se desean ver los NA, lo que equivale a useNA = 'always'.
- **useNA:** instrucción de la acción que se desea realizar con los NA. Hay tres posibles valores para este parámetro: 'no' si no se desean usar, 'ifany' y 'always' si se desean incluir.

6.16.2. Tabla de frecuencias relativas

La **función prop.table** se utiliza para crear tablas de frecuencias relativas a partir de tablas de frecuencias absolutas, donde la estructura de la función se muestra a continuación.

```
prop.table(x, margin=NULL)
```

- **x:** tabla de frecuencia.
- **margin:** valor de 1 si se desean proporciones por filas, 2 si se desean por columnas, y NULL si se desean frecuencias globales. Por defecto el valor es NULL.

6.16.3. Función *addmargins*

Esta función se puede utilizar para agregar los totales por filas o por columnas a una tabla de frecuencias absolutas o relativas. La estructura de la función es la siguiente:

```
addmargins(A, margin)
```

- **A:** tabla de frecuencia.
- **margin:** valor 1 si se desean proporciones por columnas, 2 si se desean por filas y NULL si se desean frecuencias globales.

6.16.4. Función *hist*

Construir tablas de frecuencias para variables cuantitativas es necesario en muchos procedimientos estadísticos. La **función hist** sirve para obtener este tipo de tablas. La estructura de dicha función es la siguiente:

```
hist(x, breaks='Sturges', include.lowest=TRUE, right=TRUE,
     plot=FALSE)
```

Los parámetros de la función son:

- **x:** vector numérico.
- **breaks:** vector con los límites de los intervalos. Si no se especifica se usa la regla de Sturges para definir el número de intervalos y el ancho de los mismos.
- **include.lowest:** valor lógico. Si es TRUE, una observación x_i que coincida con un límite de intervalo será ubicada en el intervalo izquierdo, mientras que si es FALSE será incluida en el intervalo a la derecha.
- **right:** valor lógico. Si es igual a TRUE, los intervalos serán cerrados a la derecha de la forma (liminf, limsup]. Por el contrario, si es FALSE, serán abiertos a la derecha.

- **plot:** valor lógico. Con FALSE sólo se obtiene la tabla de frecuencias mientras que con TRUE se obtiene la representación gráfica llamada histograma.

6.17.- Anexo 17: Gráficos básicos en R

Esta sección es una introducción a los gráficos básicos en R orientada a la inspección visual y rápida de conjuntos de datos, que es fundamental en todo proceso de análisis y, particularmente, en sus fases iniciales.

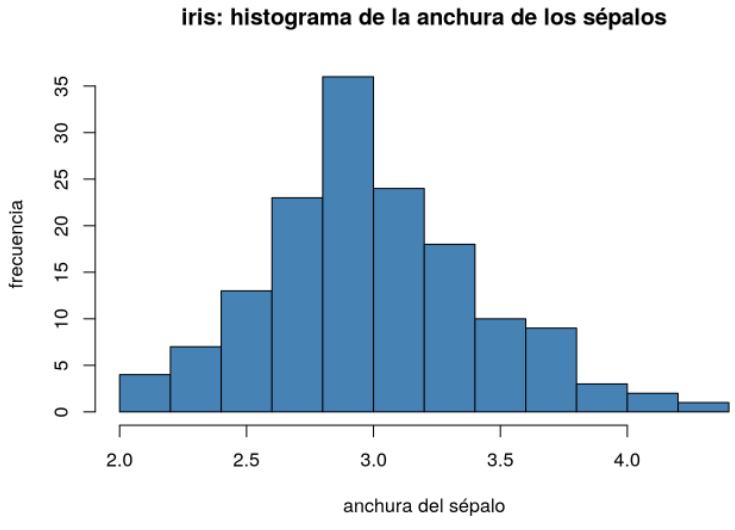
En particular, se explicará la manera de representar:

- Una variable continua
- Una variable categórica
- La relación entre dos variables continuas
- La relación entre una variable continua y otra categórica

6.17.1. Histograma

La manera más rápida (y recomendada) de hacerse una idea de la distribución de los datos de una variable numérica es usando **histogramas.** En R, para representar el histograma se puede hacer de la siguiente forma:

```
hist(iris$Sepal.Width, main = "iris: histograma de la anchura de los sépalos",
     xlab = "anchura del sépalo", ylab = "frecuencia",
     col = "steelblue")
```



Los argumentos **main, xlab, ylab** y **col** se pueden aplicar también a otros gráficos que veremos a continuación.

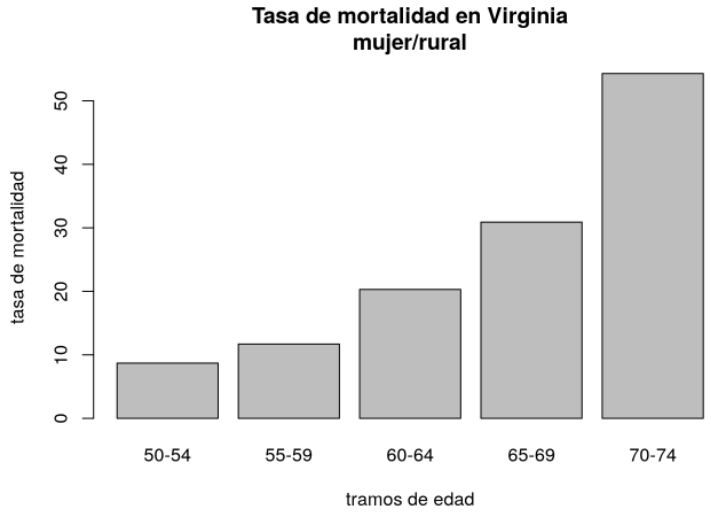
6.17.2. Diagrama de barras

En las tablas suelen coexistir variables continuas y categóricas. Por ejemplo, es interesante conocer la distribución (o frecuencia) de cada una de las categorías. Para eso se suelen usar los diagramas de barras que en particular, se pueden obtener con la **función barplot** de R.

Esta función no muestra directamente las frecuencias de una variable categórica. Es necesario calcular previamente dichas frecuencias, para lo cual usaremos la **función table.**

```
barplot(table(iris$Species))
```

```
barplot(VADeaths[, 2], xlab = "tramos de edad", ylab = "tasa de mortalidad",
        main = "Tasa de mortalidad en Virginia\nmujer/rural")
```

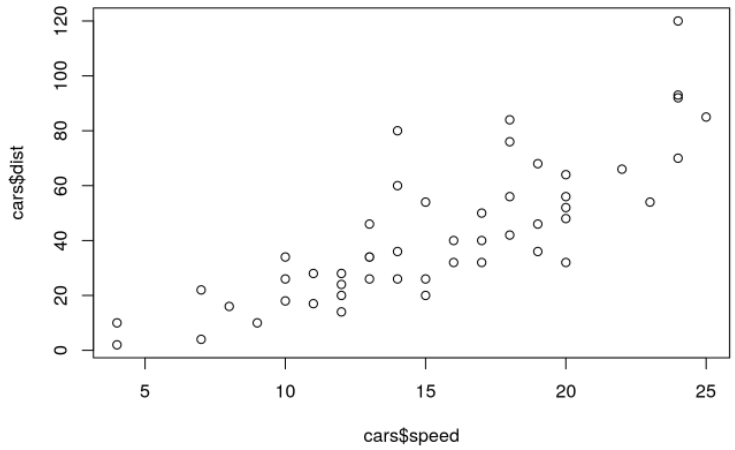


6.17.3. Gráfico de dispersión

Los aspectos más interesantes de los datos se revelan no examinando las variables independientemente sino en relación con otras. Los **gráficos de dispersión** muestran la relación entre dos variables numéricas. Se pueden

realizar con la función plot introduciendo los nombres de las dos variables a representar:

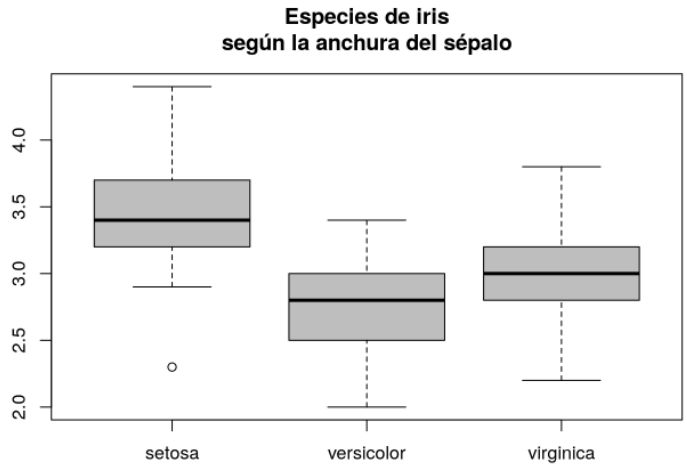
```
plot(cars$speed, cars$dist)
```



6.17.4. Boxplot

Los diagramas de cajas (**boxplots**) estudian la distribución de una variable continua en función de una variable categórica. Están relacionados con los histogramas porque resumen la distribución de una variable continua. Para ello utilizan una representación todavía más esquemática que la de un histograma: una caja y unos segmentos que acotan las regiones donde la variable continua concentra el grueso de las observaciones, como ya se vio cuando se explicaron los diagramas de caja.

```
boxplot(iris$Sepal.Width ~ iris$Species, col = "gray",
        main = "Especies de iris\nsegún la anchura del sépalo")
```



La notación **y ~x** es muy común en R y significa que **y en función de x**; por ejemplo, para la representación en un diagrama de cajas.

6.18.- Anexo 18: Programación ui.R

La **programación del archivo ui.R** es la siguiente:

```
#####################################################################
# APLICACION SHINY ESTADISTICA DESCRIPTIVA EDUCATIVA (EDE)
# JOSE MARIA ARROYO SANCHEZ
# VERSION 1.0
#####################################################################
library(shiny)
library(shinythemes)
library(moments) #Asimetría y curtosis
library(ggplot2)
library(psych) #Gráfico de correlación
library(corrplot)

ui = tagList(
  shinythemes::themeSelector(),
  navbarPage("EDE",

#####################################################################
    # PAGINA INICIO #
#####################################################################
        tabPanel("Inicio",
        sidebarPanel(
```

```
      width = 2,
      p(img(src="logo.png", height=200, width=200),align="center"),
      p("Aplicación realizada por",a("José María Arroyo",href="mailto:arrocar@gmail.com")," con el paquete Shiny del software R.", align="center"),
      br(),
      p("Versión 1.0", align="center", style="color:orangered"),
    ),
    mainPanel(
      tabsetPanel(
        tabPanel(
          h2("BIENVENIDOS A EDE (ESTADÍSTICA DESCRIPTIVA EDUCATIVA)", style="color:orangered"),
          br(),

          h4("Esta aplicación ha sido creada con el propósito de ayudar a entender la estadística descriptiva a aquellas personas que empiezan su aventura en el mundo del análisis de datos.", style="color:orangered"),
          br(),
          h3("Estadística descriptiva", style="color:royalblue"),
          br(),
          h4("El término “estadística descriptiva” se refiere al análisis, el resumen y la presentación de los resultados relacionados con un conjunto de datos derivados de una muestra o de toda la población."),
          h4("La estadística descriptiva comprende tres categorías principales:"),
          br(),
```

h3("1.- Medidas de tendencia central", style="color:royalblue"),

h4("Se refiere al resumen descriptivo de un conjunto de datos utilizando un único valor que refleja el centro de la distribución de los datos."),

h4("Las medidas de tendencia central también se conocen como medidas de localización central. La media y la mediana son consideradas las principales medidas de tendencia central."),

br(),

h4("1.- Media: es la medida de tendencia central más popular, es el valor medio en un conjunto de datos.", style="color:mediumseagreen"),

h4("2.- Mediana: se refiere a la puntuación media de un conjunto de datos en orden ascendente.", style="color:mediumseagreen"),

br(),

h3("2.- Medidas de variabilidad", style="color:royalblue"),

h4("Una medida de variabilidad es una estadística de resumen que refleja el grado de dispersión de una muestra."),

h4("Las medidas de variabilidad determinan la distancia que los puntos de datos parecen tener con respecto al centro."),

br(),

h4("1.- Rango: representa el grado de dispersión o la distancia entre los valores más altos y más bajos dentro de un conjunto de datos.", style="color:mediumseagreen"),

h4("2.- Desviación estándar: proporciona una idea de la distancia o la diferencia entre un valor de un conjunto de datos y el valor medio del mismo conjunto de datos.", style="color:mediumseagreen"),

h4("3.- Varianza: refleja el grado de dispersión y es esencialmente una media de las desviaciones al cuadrado.", style="color:mediumseagreen"),

br(),

h3("3.- Gráficos de representación", style="color:royalblue"),

```
        h4("Cuando se hace un estudio estadístico se obtiene una gran cantidad de datos numéricos. Para tener una información clara y rápida de lo obtenido en el estudio se han creado los gráficos de representación."),

        br(),

        h4("1.- Gráfico de cajas: es una caja rectangular, donde los lados más largos muestran el recorrido de los datos. Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana.", style="color:mediumseagreen"),

        h4("2.- Histograma: es la representación gráfica en forma de barras, que simboliza la distribución de un conjunto de datos.", style="color:mediumseagreen"),

        h4("3.- Diagrama de densidad: visualiza la distribución de datos en un intervalo continuo. Este gráfico es una variación de un histograma que usa el suavizado para trazar valores, permitiendo distribuciones más suaves .", style="color:mediumseagreen"),

          )

        )

      )

    ),

##########################################################################

   # ESTADISTICA UNIVARIANTE #

##########################################################################

 tabPanel("Estadística univariante",

       sidebarPanel(width = 2,

        p("Crear un vector con valores aleatorios", style="color:orangered"),

        p(actionButton("action", "Crear valores aleatorios", align = "right")),

        br(),
```

```
        p("Medidas básicas", style="color:orangered"),
        checkboxInput("minimo", "Mostrar el valor mínimo", FALSE),
        checkboxInput("maximo", "Mostrar el valor máximo", FALSE),
        checkboxInput("rango", "Mostrar el rango", FALSE),
        br(),
        p("Medidas de tendencia central", style="color:orangered"),
        checkboxInput("media", "Mostrar la media", FALSE),
        checkboxInput("mediana", "Mostrar la mediana", FALSE),
        br(),
        p("Medidas de dispersión", style="color:orangered"),
        checkboxInput("varianza", "Mostrar la varianza", FALSE),
        checkboxInput("sd", "Mostrar la desviación típica", FALSE),
        checkboxInput("coefvar", "Mostrar el coeficiente de variación", FALSE),
        checkboxInput("coefcor", "Coeficiente de correlación", FALSE),
        br(),
        p("Medidas de asimetría", style="color:orangered"),
        checkboxInput("asimetria", "Mostrar la asimetría", FALSE),
        checkboxInput("curtosis", "Mostrar la curtosis", FALSE),
        br(),
        p("Gráficos de representación", style="color:orangered"),
        checkboxInput("graficos", "Diagrama de cajas, histograma y gráfico de
densidad", FALSE),
    ),
      mainPanel(
        tabsetPanel(
```

```
tabPanel(
  h2("ESTADÍSTICA UNIVARIANTE", style="color:orangered"),
  br(),
  h4("Aquí tenemos el vector de 50 valores creado de forma aleatoria entre los números 0 y 100."),
  br(),
  h4(verbatimTextOutput ("datos")),
  br(),
  h4("Medidas básicas", style="color:orangered"),
  fluidRow(
    column(4, htmlOutput("minimo")),
    column(4, htmlOutput("maximo")),
    column(4, htmlOutput("rango"))
  ),
  h4("Medidas de tendencia central", style="color:orangered"),
  h4("1.- La media es la suma de todas las observaciones divididas entre el número de observaciones de los datos.", style="color:royalblue"),
  h4("2.- La mediana es valor que divide un conjunto de observaciones, ordenadas de menor a mayor, en dos partes con el mismo número de observaciones.", style="color:royalblue"),

 fluidRow(
    column(4, htmlOutput("media")),
    column(4, htmlOutput("mediana"))
  ),
  h4("Medidas de dispersión", style="color:orangered"),
```

h4("1.- La varianza es la media aritmética de las desviaciones al cuadrado de los valores de la variable con respecto a su media. La varianza es siempre positiva y cuanto mayor sea su valor mayor será la dispersión de los datos.", style="color:royalblue"),

h4("2.- La desviación típica o estándar es la raíz cuadrada positiva de la varianza. La desviación típica se usa más que la varianza, ya que está expresada en las mismas unidades que la variable, mientras que la varianza está expressada en unidades cuadradas.", style="color:royalblue"),

fluidRow(

column(4, htmlOutput("varianza")),

column(4, htmlOutput("sd"))

),

h4("3.- El coeficiente de variación es la división entre la desviación típica de una muestra y su media.", style="color:royalblue"),

h4("4.- El coeficiente de correlación es la medida que cuantifica la intensidad de la relación lineal entre dos variables, toma los valores entre -1 y 1. En nuestro caso, como solo tenemos una variable será siempre 1.", style="color:royalblue"),

fluidRow(

column(4, htmlOutput("coefvar")),

column(4, htmlOutput("coefcor"))

),

h4("Medidas de asimetría", style="color:orangered"),

h4("1.-La asimetría es la medida que indica la simetría de la distribución de una variable respecto a la media.", style="color:royalblue"),

h4("2.-La curtosis (o apuntamiento) mide a su vez cómo de apuntada o achatada es la distribución mirando la cantidad de elementos cercanos al valor central.", style="color:royalblue"),

```
        fluidRow(
          column(4, htmlOutput("asimetria")),
          column(4, htmlOutput("curtosis"))
        ),
        plotOutput("graficos"),
      )
      )
    ),
  ),
##########################################################################
  # ESTADISTICA BIVARIANTE #
##########################################################################
  tabPanel("Estadística bivariante",
    sidebarPanel(width = 2,
      selectInput(inputId = "dataset", label = "Conjunto de datos", choices = c("cars", "pressure")),
      helpText("Nota: por defecto se muestra el conjunto de datos cars, puedes cambiarlo en CONJUNTO DE DATOS y clicando en el botón ACTUALIZAR."),
      br(),
      numericInput(inputId = "obs", label = "Número de observaciones", value = 5),
      helpText("Nota: por defecto la tabla observaciones solo muestra 5 entradas, puedes añadir más en NUMERO DE OBSERVACIONES y clicando en el botón ACTUALIZAR."),
      actionButton("update", "Actualizar"),
      br(),
```

```
    br(),
    p("Gráficos de representación", style="color:orangered"),
    checkboxInput("graficos2", "Conjunto de datos cars", FALSE),
    checkboxInput("graficos3", "Conjunto de datos pressure", FALSE),
  ),
  mainPanel(
    tabsetPanel(
      tabPanel(
        h2("ESTADÍSTICA BIVARIANTE", style="color:orangered"),
        br(),
        h4("En este apartado se trabajará con dos conjuntos de datos de R, cars y pressure, los cuales tienen 2 variables."),
        br(),
        h4("1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:", style="color:royalblue"),
        br(),
        h4(verbatimTextOutput("summary")),
        br(),
        h4("2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:", style="color:royalblue"),
        br(),
        h4("Observaciones"),
        tableOutput("view"),
        br(),
        plotOutput("graficos2", height = 500, width = 1500),
```

```
        plotOutput("graficos3", height = 500, width = 1500),
      )
      )
    )
  ),
#####################################################################
  # ANALISIS DE CORRELACION #
#####################################################################
  tabPanel("Análisis correlación",
           sidebarPanel(width = 2,
                        selectInput("dataset3", label = "Conjunto de datos", choices = c("cars", "pressure")),
                        helpText("Nota: por defecto se muestra el conjunto de datos cars, puedes cambiarlo en CONJUNTO DE DATOS y clicando en el botón ACTUALIZAR."),
                        br(),
                        p("Gráficos de correlación", style="color:orangered"),
                        checkboxInput("graficos6", "Gráficos de correlación", FALSE),
                        actionButton("update3", "Actualizar"),
                        br(),
                        br(),
           ),
           mainPanel(
             tabsetPanel(
               tabPanel(
```

```
h2("ANÁLISIS CORRELACIÓN", style="color:orangered"),

br(),

h4("Para aprender que es un análisis de correlación utilizaremos el conjunto de datos cars, el cual tiene 2 variables numéricas."),

br(),

h4("1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:", style="color:royalblue"),

br(),

h4(verbatimTextOutput("summary3")),

br(),

h4("La correlación es un tipo de asociación entre dos variables numéricas, específicamente evalúa la tendencia (creciente o decreciente) en los datos.", style="color:royalblue"),

h4("1.- La correlación nos permite medir el signo y magnitud de la tendencia entre dos variables.", style="color:royalblue"),

h4("2.- El signo nos indica la dirección de la relación, como hemos visto en el diagrama de dispersión.", style="color:royalblue"),

h4("3.- La magnitud nos indica la fuerza de la relación, y toma valores entre -1 a 1.", style="color:royalblue"),

br(),

plotOutput("graficos6", height = 500, width = 1500),

)

)

)

),
```

```
#####################################################################
  # REGRESION LINAL SIMPLE #
#####################################################################
  tabPanel("Regresión lineal simple",
      sidebarPanel(width = 2,
          selectInput("dataset4", label = "Conjunto de datos", choices = c("cars", "pressure")),
          helpText("Nota: por defecto se muestra el conjunto de datos cars, puedes cambiarlo en CONJUNTO DE DATOS y clicando en el botón ACTUALIZAR."),
          br(),
          p("Regresión lineal de cars", style="color:orangered"),
          checkboxInput("regresion", "Modelo de regresión lineal de cars", FALSE),
          checkboxInput("modelo", "Gráfico del modelo de regresión de cars", FALSE),
          br(),
          p("Regresión lineal de pressure", style="color:orangered"),
          checkboxInput("regresion2", "Modelo de regresión lineal de pressure", FALSE),
          checkboxInput("modelo2", "Gráfico del modelo de regresión de pressure", FALSE),
          actionButton("update4", "Actualizar"),
          br(),
          br(),
      ),
```

```
mainPanel(
 tabsetPanel(
  tabPanel(
   h2("REGRESIÓN LINEAL SIMPLE", style="color:orangered"),
   br(),

   h4("Para aprender que es un análisis de regresión lineal simple utilizaremos los conjuntos de datos cars y pressure, los cuales tienen 2 variables numéricas."),
   br(),
   h4("1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:", style="color:royalblue"),
   br(),
   h4(verbatimTextOutput("summary4")),
   br(),
   h4("2.- El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente y un conjunto de variables independientes X1,..., Xn.", style="color:royalblue"),
   br(),
   h4(verbatimTextOutput("regresion")),
   br(),
   plotOutput("modelo"),
   h4(verbatimTextOutput("regresion2")),
   br(),
   plotOutput("modelo2"),
 )
```

```
            )
          )
    ),
##########################################################################
   # ESTADISTICA MULTIVARIANTE #
##########################################################################
       tabPanel("Estadística multivariante",
      sidebarPanel(width = 2,
       selectInput("dataset2",
                   label = "Escoge un conjunto de datos",
                   choices = c("rock", "iris", "mtcars")),
       helpText("Nota: por defecto se muestra el conjunto de datos rock, puedes cambiarlo en CONJUNTO DE DATOS y clicando en el botón ACTUALIZAR."),
       numericInput("obs2",
                    label = "Escoge el número de observaciones",
                    value = 5),
              helpText("Nota: por defecto la tabla observaciones solo muestra 5 entradas, puedes añadir más en NUMERO DE OBSERVACIONES y clicando en el botón ACTUALIZAR."),
       actionButton("update2", "Actualizar"),
       br(),
       br(),
       p("Gráficos de cajas", style="color:orangered"),
       checkboxInput("graficos4", "Gráficos de cajas", FALSE),
       checkboxInput("graficos5", "Gráficos de correlación", FALSE),
```

```
    br(),
  ),
  mainPanel(
    tabsetPanel(
      tabPanel(
        h2("ESTADÍSTICA MULTIVARIANTE", style="color:orangered"),
        br(),

        h4("En este apartado se trabajará con tres de los conjuntos de archivos más famosos de R, rock (tiene 4 variables), iris (tiene 5 variables) y mtcars (tiene 11 variables)."),
        br(),
        h4("1.- El resumen estadístico del conjunto de datos seleccionado es el siguiente:", style="color:royalblue"),
        br(),
        h4(verbatimTextOutput("summary2")),
        br(),
        h4("2.- Aquí podemos visualizar las primeras 5 filas del conjunto de datos seleccionado, aunque es posible cambiar este número:", style="color:royalblue"),
        br(),
        h5(tableOutput("view2")),
        br(),
        plotOutput("graficos4", height = 500, width = 1500),
        plotOutput("graficos5", height = 500, width = 1500),
      )
```

)

)

)

)

)

6.19.- Anexo 19: Programación server.R

La **programación del archivo server.R** es la siguiente:

```
#####################################################################
# APLICACION SHINY ESTADISTICA DESCRIPTIVA EDUCATIVA (EDE)
# JOSE MARIA ARROYO SANCHEZ
# VERSION 1.0
#####################################################################
library(shiny)
library(shinythemes)
library(moments) #Asimetría y curtosis
library(ggplot2)
library(psych) #Gráfico de correlación
library(corrplot)
server = function(input, output) {
 ###################################################################
  # ESTADISTICA UNIVARIANTE #

#####################################################################
   datos <- reactive ({
    input$action
    isolate ({
      return (sample (0:100, 50, TRUE))
      return (x)
    })
```

```
})


output$caption <- renderText({ input$caption })
output$datos <- renderText ({ datos () })
output$minimo <- renderTable({
  if(input$minimo == T){
    minimo <- as.data.frame(c(min(datos())))
    colnames(minimo) <- c("Valor mínimo")
    rownames(minimo) <- c("minimo")
    minimo
  }
}, digits=3)
  output$maximo <- renderTable({
  if(input$maximo == T){
    maximo <- as.data.frame(c(max(datos())))
    colnames(maximo) <- c("Valor máximo")
    rownames(maximo) <- c("maximo")
    maximo
  }
}, digits=3)
  output$rango <- renderTable({
  if(input$rango == T){
    rango <- as.data.frame(c(max(datos()-min(datos()))))
    colnames(rango) <- c("Rango")
```

```
    rownames(rango) <- c("rango")
    rango
  }
}, digits=3)
output$media <- renderTable({
  if(input$media == T){
    media <- as.data.frame(c(mean(datos())))
    colnames(media) <- c("Valor de la media")
    rownames(media) <- c("media")
    media
  }
}, digits=3)
 output$mediana <- renderTable({
  if(input$mediana == T){
    mediana <- as.data.frame(c(median(datos())))
    colnames(mediana) <- c("Valor de la mediana")
    rownames(mediana) <- c("mediana")
    mediana
  }
}, digits=3)
output$varianza <- renderTable({
  if(input$varianza == T){
    varianza <- as.data.frame(c(var(datos())))
    colnames(varianza) <- c("Valor de la varianza")
```

```
    rownames(varianza) <- c("varianza")
    varianza
  }
}, digits=3)
output$sd <- renderTable({
  if(input$sd == T){
    sd <- as.data.frame(c(sd(datos())))
    colnames(sd) <- c("Valor de la desviación típica")
    rownames(sd) <- c("sd")
    sd
  }
}, digits=3)
output$coefvar <- renderTable({
  if(input$coefvar == T){
    coefvar <- as.data.frame(c(sd(datos()/mean(datos()))))
    colnames(coefvar) <- c("Valor del coeficiente de variación")
    rownames(coefvar) <- c("coefvar")
    coefvar
  }
}, digits=3)
output$coefcor <- renderTable({
  if(input$coefcor == T){
    coefcor <- as.data.frame(c(cor(datos(), datos())))
    colnames(coefcor) <- c("Valor del coeficiente de correlación")
```

```
    rownames(coefcor) <- c("coefcor")
    coefcor
  }
}, digits=3)

output$asimetria <- renderTable({
  if(input$asimetria==T){
    asimetria <- as.data.frame(c(skewness(datos())))
    colnames(asimetria) <- c("Valor de la asimetría")
    rownames(asimetria) <- c("asimetria")
    asimetria
  }
}, digits=3)
output$curtosis <- renderTable({
  if(input$curtosis == T){
    curtosis <- as.data.frame(c(kurtosis(datos())))
    colnames(curtosis) <- c("Valor de la curtosis")
    rownames(curtosis) <- c("curtosis")
    curtosis
  }
}, digits=3)
output$graficos <-renderPlot({
  if(input$graficos == T){
  par(mfrow=c(1,3))
```

```
  cajas <- boxplot(datos(), main = "Diagrama de cajas", col = "green")
  cajas
      histograma <- hist(datos(), main = "Histograma", col = "blue")
      histograma
      densidad <- plot(density(datos()), main = "Gráfico de densidad", col = "red", lwd = 3)
      densidad
    }
  })
##########################################################################
  # ESTADISTICA BIVARIANTE #
##########################################################################
  datasetInput <- eventReactive(input$update, {
    switch(input$dataset,
           "cars" = cars,
           "pressure" = pressure)
  }, ignoreNULL = FALSE)
  output$caption <- renderText({ input$caption })
  output$summary <- renderPrint({
    dataset <- datasetInput()
    summary(dataset)
  })
  output$view <- renderTable({
    head(datasetInput(), n = isolate(input$obs))
  })
```

```
 output$graficos2 <-renderPlot({
   if(input$graficos2 == T){
     par(mfrow=c(1,3))
     cajas2 <- boxplot(cars, main = "Diagrama de cajas datos cars", col = "green")
     cajas2

     densidad2 <- plot(density(cars$speed), main = "Gráfico de densidad variable speed", col = "red", lwd = 3)
     densidad2
     densidad3 <- plot(density(cars$dist), main = "Gráfico de densidad variable dist", col = "blue", lwd = 3)
     densidad3
  }
})
 output$graficos3 <-renderPlot({
   if(input$graficos3 == T){
     par(mfrow=c(1,3))
     cajas3 <- boxplot(pressure, main = "Diagrama de cajas datos pressure", col = "green")
     cajas3
     densidad4 <- plot(density(pressure$temperature), main = "Gráfico de densidad variable temperature", col = "red", lwd = 3)
     densidad4
     densidad5 <- plot(density(pressure$pressure), main = "Gráfico de densidad variable pressure", col = "blue", lwd = 3)
```

```
    densidad5
   }
  })

##########################################################################
 # ANALISIS CORRELACION #
##########################################################################
 datasetInput3 <- eventReactive(input$update3, {
  switch(input$dataset3,"cars" = cars, "pressure" = pressure)
 }, ignoreNULL = FALSE)
 output$caption3 <- renderText({ input$caption3 })
 output$summary3 <- renderPrint({
  dataset3 <- datasetInput3()
  summary(dataset3)
 })
 output$graficos6 <-renderPlot({
  if(input$graficos6 == T){
   par(mfrow=c(1,2))
   correlacion4 <- corrplot.mixed(cor(cars), lower = "number", upper = "circle",
tl.col = "black")
   correlacion4
   correlacion5 <-corrplot.mixed(cor(pressure), lower = "number", upper =
"circle", tl.col = "black")
   correlacion5
  }
```

```
})

#####################################################################
 # REGRESION LINEAL SIMPLE #
#####################################################################
 datasetInput4 <- eventReactive(input$update4, {
   switch(input$dataset4,"cars" = cars, "pressure" = pressure)
 }, ignoreNULL = FALSE)
 output$caption4 <- renderText({ input$caption4 })

 output$summary4 <- renderPrint({
   dataset4 <- datasetInput4()
   summary(dataset4)
 })
 output$modelo <- renderPlot({
   if(input$modelo == T){
     modelo <- ggplot(cars, aes(x=speed, y=dist)) +
       geom_point() +
       geom_smooth(method='lm', formula=y~x, col='blue')
     modelo
   }
 })
 output$regresion <- renderPrint({
   if(input$regresion == T){
     dataset <- datasetInput()
```

```
    regresion <- lm(speed ~ dist, data=cars)
    summary(regresion)
  }
})
output$modelo2 <- renderPlot({
  if(input$modelo2 == T){
    modelo2 <- ggplot(pressure, aes(x=temperature, y=pressure)) +
      geom_point() +
      geom_smooth(method='lm', formula=y~x, col='blue')
    modelo2
  }
})
output$regresion2 <- renderPrint({
  if(input$regresion2 == T){
    dataset <- datasetInput()
    regresion2 <- lm(temperature ~ pressure, data=pressure)
    summary(regresion2)
  }
})

#####################################################################
 # ESTADISTICA MULTIVARIANTE #
#####################################################################
  datasetInput2 <- eventReactive(input$update2, {
   switch(input$dataset2,
```

```
        "rock" = rock,
        "iris" = iris,
        "mtcars" = mtcars)
  }, ignoreNULL = FALSE)
  output$caption2 <- renderText({ input$caption2 })
  output$summary2 <- renderPrint({
    dataset2 <- datasetInput2()
    summary(dataset2)
  })

  output$view2 <- renderTable({
    head(datasetInput2(), n = isolate(input$obs2))
  })
  output$graficos4 <-renderPlot({
    if(input$graficos4 == T){
      par(mfrow=c(1,3))
      cajas4 <- boxplot(rock, main = "Diagrama de cajas datos rock", col = "green")
      cajas4
      cajas5 <- boxplot(iris, main = "Diagrama de cajas datos iris", col = "green")
      cajas5
      cajas6 <- boxplot(mtcars, main = "Diagrama de cajas datos mtcars", col = "green")
      cajas6
    }
```

```
  })
  output$graficos5 <-renderPlot({
    if(input$graficos5 == T){
      par(mfrow=c(1,3))
      correlacion <- corrplot.mixed(cor(rock), lower = "number", upper = "circle", tl.col = "black")
      correlacion
      correlacion2 <- corrplot.mixed(cor(iris[1:4]), lower = "number", upper = "circle", tl.col = "black")
      correlacion2
      correlacion3 <- corrplot.mixed(cor(mtcars), lower = "number", upper = "circle", tl.col = "black")
      correlacion3
    }
  })
}
```