Big Data, Machine Learning & Business Intelligence

Lab: Programación en Lenguaje R

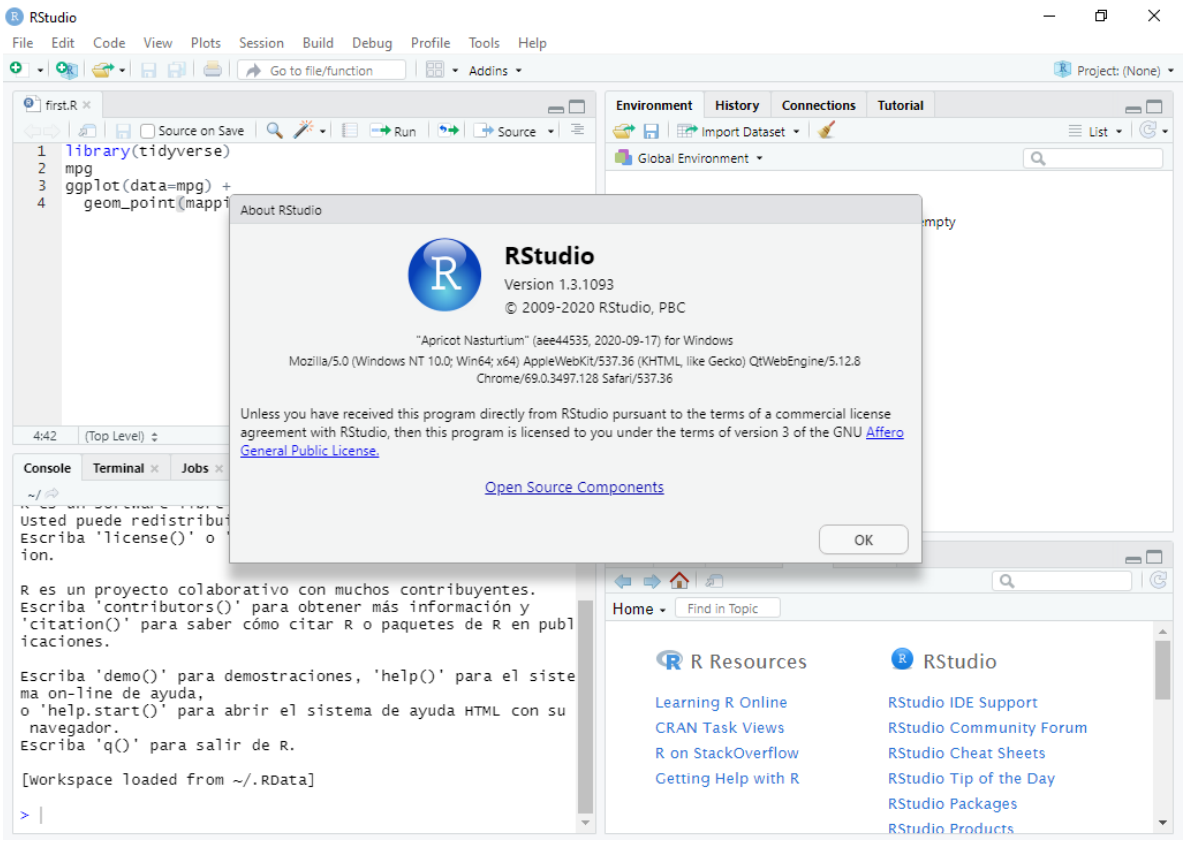
Objetivos

Después de realizar este laboratorio el participante podrá:

- Programar aplicaciones con Lenguaje R

Procedimiento

Realiza los siguientes ejercicios en Lenguaje R y R Studio.



1. Un conductor europeo de viaje por Estados Unidos apunta las millas recorridas por su coche cada vez que llena el depósito de gasolina. La relación de las últimas 6 veces que lo ha llenado es la siguiente:

   ```
   65311 65624 65908 66219 66499 66821 67145 67447
   ```

   (a) Crea una variable llamada "millas" que contenga los datos anotados. Crea una nueva variable "kms" y asígnale el valor de "millas" transformado en kms (un km son 1.609 millas).

   (b) ¿Qúe resultado produce la función `diff` aplicada sobre los datos anteriores?

   (c) ¿Que funciones son adecuadas para resumir estos datos?

2. Una persona dispone de un contrato de "pago mínimo" telefónico, con el que cuenta poder controlar sus gastos. A pesar de esto cada mes tiene que pagar una cantidad adicional, que finalmente decide revisar. En el último año estas cantidades, en euros, han sido las siguientes:

   ```
   46 33 39 37 46 30 48 32 49 35 30 48
   ```

   (a) ¿Cuanto le ha costado la factura del último año? ¿Cuanto ha pagado en promedio cada mes?

   (b) ¿Cuales son las cantidades mínimas y máximas pagadas? En que mes se realizó cada pago?

   (c) ¿Cuantos meses pagó mas de 40 euros? ¿Que porcentaje del gasto total representa esta cantidad?

3. ¿Cual será el resultado de ejectutar las siguientes instrucciones?:

   ```
   > x = c(1,3,5,7,9)
   > y = c(2,3,5,7,11,13)

   x+1
   y*2
   length(x) and length(y)
   x + y
   sum(x>5) and sum(x[x>5])
   sum(x>5 | x< 3) # ``|'' se lee 'o', ``&'' se lee 'y'
   y[3]
   y[-3]
   y[x] (¿Que significa NA?)
   y[y>=7]
   ```

Datos univariantes

1. Con los datos siguientes: `60 85 72 59 37 75 93  7 98 63 41 90  5 17 97}`

   (a) Haz un diagrama de tallo y hojas

   (b) Obten resúmenes numéricos de los datos.

   (c) ¿Que diferencias hay entre `summary(x)` y `fivenum(x)`?

2. `R`permite generar datos aleatorios con gran facilidad mediante instrucciones específicas que empiezan con "r" (`rnorm, rpois, rbinom,...`).

   (a) Genera 100 valores de una distribución normal con `rnorm(100)`.

   (b) Realiza un histograma de los valores. Repite el proceso un par de veces. ¿Que observas?

   (c) Realiza un resumen numérico de los datos.

3. De forma similar al ejercicio genera 30 valores de una distribución binomial de parámetros (n=5 y p=1).

   (a) Representa los resultados con un diagrama de barras o de pastel.

   (b) Realiza un resumen numérico de los datos y compáralo con el del ejercicio anterior. ¿Que deberías hacer para obtener un resumen similar?

4. Carga (o instala primero y luego carga)el paquete `UsingR` [1]

   (a) ¿Cuantos conjuntos de datos de trabajo contiene el paquete?

   (b) Representa gráficamente los datos contenidos en los conjuntos de datos ("datasets") `bumpers, firstchi, math` con un histograma y/o un boxplot.

   (c) Estima visualmente la medias, medianas y desviaciones estándar de cada conjunto de datos y a continuación calcula los valores anteriores con las funciones adecuadas. ¿Qué gráfico resulta de mayor ayuda para la aproximación?

5. El numero de fallos en los 23 primeros intentos de puesta en orbita de un satélite fue:

   `0 1 0 NA 0 0 0 0 0 1 1 1 0 0 3 0 0 0 0 0 2 0 1`

   (NA significa "not available" – se ha perdido el dato).

   (a) Representa gráficamente estos datos. Qué representación es más adecuada una diagrama de tallo y hojas o un diagrama de barras?

   (b) Tabula los datos y calcula el número medio de errores (Puedes tener que probar con `mean(x,na.rm=TRUE)` o `x[!is.na(x)]` para prescindir de los valores faltantes.

6. El conjunto de datos `brightness` contiene información sobre el brillo de 963 estrellas.

   (a) Representa estos datos mediante un histograma y un gráfico de densidad superpuesto.

   (b) Representa gráficamente estos datos mediante un diagrama de caja (boxplot). ¿Dirias que los datos presentan "outliers"? Cual es el segundo menor outlier?

   (c) Deseamos conservar los datos que de ninguna forma puedan ser considerados atípicos. Crea una nueva variable denominada `brightness.sin` que contenga tan sólo los valores que se encuentren por encima de la primera bisagra y por debajo de la cuarta.

Datos bivariantes

1. En una encuesta en la que se evalua el funcionamiento de un curso se han recogido las siguientes respuestas de 10 estudiantes a tres preguntas $P1$, $P2$ y $P3$:

Estudiante	P1	P2	P3
1	3	5	1
2	3	2	3
3	3	5	1
4	4	5	1
5	3	2	1
6	4	2	3
7	3	5	1
8	4	5	1
9	3	4	1
10	4	2	1

   (a) Entra los datos mediante `c()` `scan()`, `read.table()` o `data.entry()`.

   (b) Tabula los resultados de cada pregunta por separado.

   (c) Realiza tablas de contingencia cruzadas para cada pregunta, de 2 en 2 y las 3 a la vez.

   (d) Haz un diagrama de barras apiladas de las preguntas 2 y 3.

   (e) Haz un diagrama de barras con las tres preguntas simultanemente.

2. El paquete *MASS* contiene la base de datos `UScereal` con información relativa a desayunos con cereales.

   (a) ¿Cual es el tipo de datos de cada variable?

   (b) Utiliza los datos de cereales para investigar algunas asociaciones entre sus variables:

      i. La relación entre `manufacturer` y `shelf`.
      ii. La relación entre `fat` y `vitamins`.
      iii. La relación entre `fat` y `shelf`.
      iv. La relación entre `carbohydrates` y `sugars`.
      v. La relación entre `fibre` y `manufacturer`.
      vi. La relación entre `sodium` y `sugars`.

3. El conjunto de datos `mammals` contiene datos sobre la relación entre peso corporal y peso del cerebro.

   (a) ¿Cual es la correlación lineal entre estas variables?

   (b) Representa los datos mediante la instrucción `plot`

   (c) Transforma los datos mediante la función `log` y repite el estudio. ¿Cómo cambian los resultados?

4. Enlaza la base de datos *emisions* del paquete *UsingR*.

   (a) Estudia la relación entre las variables GDP (`Gross Domestic Product`), `perCapita` (pues eso) y `CO2` (Emisiones de CO2) de cada pais.

   (b) Construye un modelo de regresión para predecir las emisiones de CO2 a partir de cada una de las variables.

   (c) Identifica los outliers y prueba de ajustar el modelo de nuevo sin ellos.