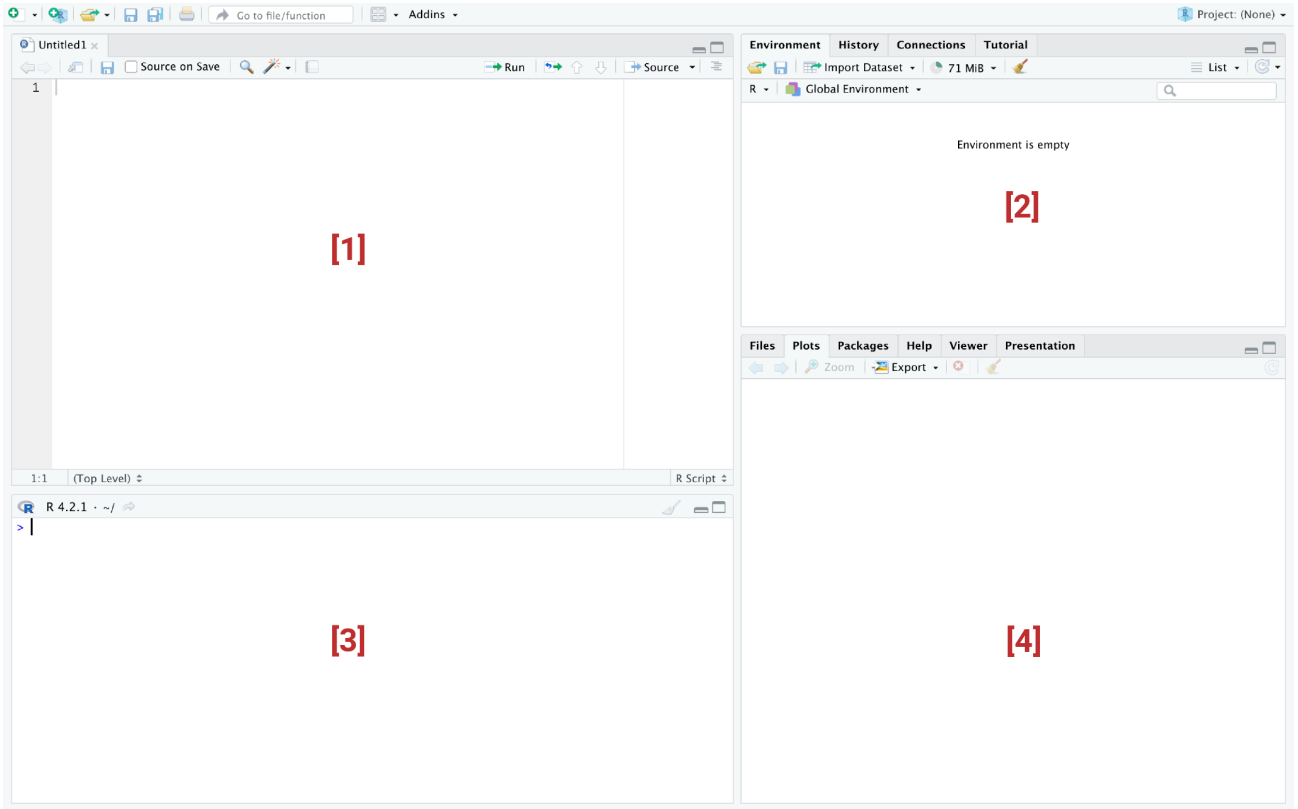
Introducción a la Ciencia de Datos - 1C 2023

Guía de Trabajos Prácticos Nº 1

Les proponemos comenzar a analizar un *set* de datos (conjunto de datos) a partir de una serie de consignas. Para esto vamos a utilizar el lenguaje de programación R, utilizando **RStudio** como entorno de desarrollo integrado (IDE). Si es que no saben de qué se trata, este último término lo verán en detalle en la materia **Programación I**. Por ahora quedémonos con la idea de que **RStudio** es el *programa* que vamos a usar para usar R.

Conociendo R y RStudio

1. Abran **RStudio** en la computadora que estén trabajando. Deberían encontrarse con una pantalla como la siguiente:



donde pueden ver 4 ventanas, además de la barra de opciones en la parte superior:

[1] es el editor de sintaxis: se trata del lugar donde escribimos el código para posteriormente ejecutarlo. Al escribir allí no sucederá nada, a no ser que se apriete algún botón para ejecutar los comandos o la tecla ctrl+enter. Si cuando abren RStudio esta ventana no les aparece, es porque tienen que crear un nuevo *script* con el botón de arriba a la izquierda en la barra de herramientas (el que tiene una hoja con un +), abrir uno con el típico botón de abrir, o usar el atajo Ctrl+Shift+N.

[2] es el "entorno de trabajo" del programa: en este lugar se muestra el conjunto de datos y los "objetos" (resultados, variables, etc.) que se almacenan al ejecutar diferentes análisis. Cuando aparezcan cosas ahí, pueden cliquearlas para explorarlas.

[3] es la consola. Allí el software ejecuta las operaciones realizadas desde el editor de sintaxis. También se puede escribir y ejecutar código desde allí directamente. *(OJO: no se guarda lo aquí ejecutado en el* script *que exporten).*

[4] esta zona tiene varias subpestañas: (i) la pestaña *files* es un explorador de archivos como el del sistema; (ii) la pestaña *plots* permite visualizar los gráficos que se generen y visualizar los previos; (iii) la pestaña packages permite ver los paquetes descargados y guardados en el disco duro así como gestionar su instalación o actualización; (iv) la pestaña *help* permite acceder a la página oficial del software con recursos para el programa: manuales para el usuario, cursos *on line*, información general, descarga de paquetes, información de los paquetes instalados, etc.

La idea es que en esta clase vayan familiarizándose con el uso del lenguaje de análisis de datos R utilizando **RStudio** como interfaz.

Análisis del dataset `iris`

2. En RStudio abran el *script* iris.R. que pueden descargar desde el Campus de la Materia. En ese *script* van a encontrar el código que vamos a usar en la clase de hoy. Vamos a ir repasando qué es lo que hace cada línea del código y ustedes van a tener que responder preguntas a partir de lo que vayan obteniendo.
   Para no correr todo el código de una vez, la idea es que **copien cada comando que vayan necesitando desde el editor de sintaxis y lo peguen en la consola para ejecutarlo.** También pueden seleccionar el texto y ejecutarlo en la consola directamente con el atajo **Ctrl+Enter**.

   Esta dinámica va a permitir que ustedes vayan entendiendo qué hace cada comando y que puedan utilizarlo en el futuro (por ejemplo, para la Entrega 1 😉).

3. En primer lugar, ejecutar (de ahora en más, cada vez que digamos "ejecutar" nos referiremos a copiar, pegar y correr -apretar enter- en la ventana de la consola o usar el atajo Ctrl+Enter en la venta del editor) las líneas

   ```
   library(tidyverse)
   ```

   para cargar la librería que les van a permitir, entre otras cosas, hacer gráficos con el comando `ggplot`. En general, los comandos para cargar librerías se ubican arriba de todo en los *scripts*.

4. En la clase de hoy vamos a trabajar con un dataset muy famoso llamado `iris`. Dado que este conjunto de datos está incorporado en R, se lo puede cargar en el *data frame* "`data`" utilizando el siguiente comando

   `data <- iris`.

   Los *data frames* son estructuras de datos de dos dimensiones que pueden contener datos de diferentes tipos. Es la estructura de datos más usada para realizar análisis de datos. Ejecutando el comando

   `view(data)`

   se puede activar un visor de datos tipo hoja de cálculo (como tabla de datos). Allí van a poder ver los datos ordenados en filas y columnas, donde cada una de estas últimas tendrá su *header* (encabezado), indicando el nombre de la variable que representa. Otra forma de acceder a esta presentación del conjunto de datos es haciendo clic en el nombre de la variable en el "entorno de trabajo".

   Este comando nos va a servir para poder explorar el *dataset* y entender de qué va. Sin embargo, es probable que esto no sea suficiente.

5. Ejecuten el comando

   `help(iris)`

   para obtener información sobre el dataset. Esta aparecerá en la ventana **[4]** de RStudio. Utilicen esta información, junto con la que encuentren con la herramienta más importante que tienen y tendrán por el resto de su vida: **Google**. Jamás subestimen una buena búsqueda en google. Busquen la información necesaria para entender qué representa cada columna del dataset (https://letmegooglethat.com/?q=Sepalo). Esta es una tarea que tendrán que realizar cada vez que se enfrenten a un nuevo dataset.

6. Responder las siguientes preguntas sobre el dataset en este formulario: shorturl.at/muHY4.
   a. ¿Cuántas filas hay?
   b. ¿Cuántas columnas hay?
   c. ¿Qué representa cada fila? Es decir, ¿cuáles son las **unidades** de este dataset? Recuerden que para contestar esta pregunta correctamente es necesario ser lo más específicx posible.
   d. ¿Qué variables son categóricas?
   e. ¿Qué variables son numéricas?
   f. Estas últimas, ¿son enteros o puntos flotantes?

7. ¿Cuántas y cuáles son las especies de flores presentes en el dataset? Para responder esta pregunta pueden aplicar el comando `unique` a la columna Species. Este comando devuelve una lista con

los **elementos únicos** del vector en cuestión. Para elegir una variable de un dado *data frame* hay varias opciones, algunas son:

```
data$Species
data["Species"]
```

Gráfico de dispersión o Scatter plot

Los gráficos de dispersión se utilizan para identificar y mostrar relaciones entre dos variables numéricas. Utiliza puntos para representar los valores de dos variables numéricas de un conjunto de datos. La posición de cada punto en los ejes horizontal y vertical indica los valores de un punto de datos individual. Las relaciones entre las variables pueden describirse de muchas maneras: positivas o negativas, fuertes o débiles, lineales o no lineales.

8. En el próximo punto vamos a pedirles que realicen un gráfico de dispersión (*scatter plot*) del Ancho del Sépalo (Sepal.Width) vs. el Largo del Sépalo (Sepal.Length). Pero antes, piensen qué tipo de relación esperan encontrar entre estas variables y hagan un gráfico a mano alzada en un papel.
9. Ahora sí, con el comando

   ```
   ggplot(data, aes(x=Sepal.Length, y=Sepal.Width)) +
   geom_point() +
   xlab("Largo del Sépalo [cm]") +
   ylab("Ancho del Sépalo [cm]")
   ```

   realicen el gráfico y compárenlo con sus expectativas. ¿Se parece a lo que esperaban? Si la respuesta es no, ¿en qué difiere? ¿por qué puede ser?

   Respecto al código que usamos en este caso, lo primero que tuvimos que indicarle al comando `ggplot` es el *data frame* (`data`) y con al comando `aes` le indicamos qué columna queremos usar como la variable x y cuál otra como la variable y. Con el comando `geom_point()` indicamos el tipo de gráfico que queremos realizar (gráfico "de puntos") y por último `xlab` e `ylab` nos permiten ponerle nombre a los ejes. Noten que en este caso no sólo estamos dándole un nombre al eje sino que estamos indicando en qué unidades se está midiendo la magnitud en cuestión.
10. ¿Cuántas variables necesitaron para hacer este gráfico? (Para responder en el formulario)
11. ¿Qué representa cada uno de los puntos del gráfico? (Para responder en el formulario)
12. ¿Qué información brinda este gráfico? ¿Se observa alguna relación clara entre ambas variables? (Para responder en el formulario)

Una modificación común del gráfico de dispersión es la adición de una **tercera variable**. Los valores de la tercera variable pueden codificarse, por ejemplo, modificando la forma de los puntos. En el caso

de una tercera variable que indique valores **categóricos** la codificación más común es mediante el color de los puntos. Dar a cada punto un tono distinto permite mostrar la pertenencia a cada uno de los grupos.

13. Agreguemos al gráfico anterior la información de una **variable categórica**. La única variable categórica en este *dataset* es la especie. Para ello ejecutar el código

```
ggplot(data, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) +
geom_point() +
xlab("Largo del Sépalo [cm]") +
ylab("Ancho del Sépalo [cm]")
```

    que como ven es igual al anterior, excepto porque le agregamos el argumento `color` al comando `aes`, indicando qué columna queremos que tenga en cuenta para colorear los puntos.

14. Discutan en grupo si alguna de las conclusiones del punto 13 cambiaron, o si surge alguna nueva. Elijan las opciones que consideren correctas en el Formulario.

15. Repitan los puntos del 8 al 14 pero ahora utilizando las variables Ancho del Pétalo (Petal.Width) vs. el Largo del Pétalo (x=Petal.Length). Para realizar los gráficos de dispersión correspondientes, van a tener que modificar el código del punto 9. Allí, deberán cambiar cuáles son las columnas que quieren usar ahora como variable x y como variable y.

Para trabajar en casa

Antes de empezar: contestá el Cuestionario de la Clase 1 antes de la clase virtual de esta semana.

Instalación de R y RStudio

Pueden seguir los pasos del siguiente link para instalar tanto R como RStudio en sus computadoras.

https://bookdown.org/jboscomendoza/r-principiantes4/instalacion.html

Si usan Ubuntu este script para Ubuntu 22 y este para Ubuntu 18 ya lo instala con los paquetes necesarios.

Si tienen problemas de instalación, pueden utilizar "Colab", una herramienta disponible con su drive. Para ello, usen este enlace y les abrirá un entorno de ejecución de R en la nube ☘.

Para tener en cuenta antes de ponerse a trabajar

Los paquetes que vamos usando en R hay que instalarlos. Por ejemplo, para lo que hicimos en esta guía necesitamos usar el paquete tidyverse que ya habíamos instalado en las compus que usaron en los Laboratorios. Si ese no fuera el caso, antes de eso es necesario instalarlo corriendo en la consola el comando:

```
install.packages("tidyverse")
```

Análisis dataset insurance

Les proponemos analizar por su cuenta el dataset `insurance` que van a poder encontrar en el campus de la materia como el archivo insurance.csv. Este *dataset* brinda información sobre el precio del seguro médico de varias personas, junto con otros atributos. La fuente del dataset es la siguiente: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download. Allí van a poder encontrar información para entender cómo se formó el dataset y qué representa cada una de las columnas. Si tienen problemas con el idioma, pueden usar algún traductor.

En el archivo insurance.r van a encontrar el comando que pueden utilizar para cargar en un *data frame* la información que contiene el archivo csv. Dentro de las comillas deberán copiar la ruta de acceso del archivo, algo como:

/Users/user/Downloads/insurance.csv.

(Si usan windows tengan en cuenta esto.)

Una vez cargado, deberán usar los comandos que aprendieron en esta Guía para responder las siguientes preguntas por su cuenta:

1. ¿Cuántas variables tiene el dataset? ¿Cuántos datos? ¿Qué variables son categóricas y cuáles son numéricas?
2. ¿Qué variables se podrían usar para hacer un *scatter plot*? ¿Cómo se imaginan que va a resultar el gráfico en cada caso? Hagan el esquema en lápiz y papel antes de graficar con la computadora.

La elección de qué variable ubicar en cada eje dependerá **de lo que ustedes quieran mostrar y explicar**. Sin embargo, de materias como Matemática y/o Análisis Matemático sabemos que cuando realizamos el gráfico de una función entre dos variables (y = f(x)) se ubica en el eje x a la variable independiente mientras que en el eje y se ubica la dependiente. Cuando realizamos un gráfico de dispersión entre dos variables de un conjunto de datos es posible que no sepamos qué variable es la que modifica a la otra, pero sí podemos llegar a intuirlo y esa es una buena forma de decidir qué variable poner en cada eje.

3. Realizar un gráfico de dispersión entre las variables `age` y `charges`. Según lo planteado en el párrafo anterior, elegir qué variable poner en cada eje. Para esto, utilicen el comando del punto 10 de la Guía 1. No olviden modificar el nombre de los ejes e indicar las unidades en caso de ser necesario.
4. ¿Qué tipo de relación hay entre ambas variables? ¿Es lo que esperaban?
5. Sumen ahora una "tercera dimensión" a este gráfico, como hicimos con la especie en el caso del *dataset* iris. Elijan esa variable de tal manera que esta revele nuevas relaciones que antes no podíamos ver. Para ello **prueben** con cada una de las variables categóricas del dataset.
6. Entendamos si nuestra producción está lista:
    a. ¿Qué sería necesario agregar o cambiar para que el gráfico se pueda presentar a alguien que no conozca el dataset? ¿Podrían, por ejemplo, presentárselo algún familiar o amigx que esté con ustedes mientras terminan el trabajo, sin explicarles previamente nada del dataset? Si no es así, ¿qué le cambiarían? Para realizar estos cambios pueden valerse del comando

    `+ labs(title = "Escribir acá el título", subtitle = "Escribir acá el subtítulo")`

    sumándolo al comando ggplot como se hace con el `xlab` e `ylab`.

    b. ¡Pongamos a prueba el punto anterior! Muéstrenle el gráfico a alguna persona, puede ser física o virtualmente. Registren la experiencia: ¿qué preguntas y devoluciones recibieron? ¿Hay alguna forma de resolver esas inquietudes en el mismo gráfico? Si es así, ¡háganlo!

Entrega 1

Armen un .pdf con la producción que realizaron (*scatter plot* + tercera dimensión) junto con un breve párrafo explicando lo que se ve en ella y las conclusiones que obtienen a partir del análisis. Suban el documento al campus de la materia en la sección de "Entrega N° 1".

Tengan en cuenta que en la solapa donde aparecen los gráficos en **RStudio** van a encontrar una opción que dice "Export" que va a permitirles copiar el gráfico para agregarlo al documento.