Instrucciones: El objetivo de esta guía es repasar algunos conceptos relativos a análisis exploratorio de datos. ¡Buena suerte!

Ejercicio 0

Supongamos que $x_1, \ldots, x_n$ son n datos:

I. Probar que $\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - a)^2$.

II. Sea $f(a) = \sum_{i=1}^{n}(x_i - a)^2$. Probar que $f(a)$ alcanza su mínimo en $a = \overline{x}$.

III. Probar que $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

Ejercicio 1

Airquality

Considerar el conjunto de datos `airquality`, incluido en la librería `datasets`. Pida ayuda con el comando `?airquality` para obtener información.

1. ¿Cuántas observaciones tiene el conjunto de datos? ¿Cuántas variables?
2. ¿Cuáles son los nombres de las variables?
3. ¿De qué tipo es cada variable?
4. ¿Qué variables tienen datos faltantes?
5. ¿Cuántas observaciones corresponden a cada mes?
6. Realizar un histograma de densidad de las variables `Wind` y `Temp`.
   a) ¿De qué tipo de distribución se trata en cada caso?
   b) aproximadamente, ¿qué proporción de datos tienen valor de `Wind` entre 65 y 70?
   c) aproximadamente, ¿cuántos datos tienen valor de `Wind` entre 65 y 70?
7. Realizar un boxplot para las variables `Wind` y `Temp`. ¿Identifica outliers? En caso afirmativo, ¿a qué día y mes corresponden?
8. Realizar un diagrama de dispersión para `Wind` vs. `Temp`. ¿Hay alguna tendencia o asociación?

Ejercicio 2

Titanic (este conjunto de datos está disponible para descargar en el campus)

El RMS Titanic fue en su momento el mayor barco de pasajeros del mundo, hundiéndose en su viaje inaugural de Southampton a Nueva York en el año 1912. En el evento fallecieron 1514 de las 2223 personas que iban a bordo, entre tripulación y pasajeros.

Este conjunto de datos es un clásico de las competencias de *Machine Learning*, donde se busca determinar un mecanismo de clasificación que, en función de diversos datos (variables) de cada pasajero, prediga si el pasajero sobrevivió o no a la catástrofe. Las variables del conjunto de datos son:

(a) survival: supervivencia (0 No, 1 Sí).

(b) pclass: clase del pasajero (1,2 o 3).

(c) name: nombre del pasajero (texto).

(d) sex: sexo del pasajero (`male`, `female`).

(e) age: edad del pasajero.

(f) sibsp: cantidad de hermanos y cónyuges (totalizado) embarcados (número entero).

(g) parch: cantidad de padres e hijos (totalizado) embarcados (número entero).

(h) ticket: código del boleto (texto).

(i) fare: tarifa del pasaje (número real).

(j) embarked: puerto de embarque (S= Southampton, Q=Queenstown, C = Cherbourg)

de las cuales, algunas contienen respuestas faltantes.

9. Borrar todos los objetos existentes en el entorno de trabajo y establecer directorio de trabajo.

10. Leer el conjunto de datos, titanic.csv, teniendo en cuenta que en la primera línea del archivo figura el nombre de las variables y cuál es el tipo de separación de los datos. Asignarlo al data.frame `titanic` utilizando el comando `read.csv("titanic.csv",header=T,sep="\t")`.

11. Inspeccionar los primeros casos del archivo y los últimos.

12. Abrir con el editor al data.frame e inspeccionar el archivo.

13. Establecer el número de variables y de casos.

14. Inspeccionar los nombres de las variables de titanic e identificar de qué tipo de variable se trata cada una de ellas.

15. Determinar la proporción de sobrevivientes por clase de cabina.

16. ¿Cree que la clase de cabina del pasajero está asociada con su supervivencia?

17. Calcular la proporción de sobrevivir por sexo.

18. Estudiar la distribución de las tarifas. ¿Qué observa? ¿Parece razonable suponer que la variable tarifa tenga distribución normal? ¿Puede decidir de antemano quién es más grande si la media o la mediana? Calcular la media y la mediana.

19. Estudiar la relación entre **tarifa** y **clase** y por otro lado entre **edad** y **clase**.

20. Algunos dicen que las últimas horas a bordo del Titanic estuvieron marcadas por la guerra de clases, otros sostienen que estuvieron caracterizadas por la caballerosidad de los varones. En su opinión y basándose en estos datos ¿fue la guerra de clases, la caballerosidad de los varones o una combinación de ambas lo que caracterizó las últimas horas del Titanic?

La idea de los puntos anteriores era explorar los datos usando boxplots, histogramas, scatterplots. Si le faltó usar alguno de estos gráficos revise.