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

Ejercicio Lectura de clase:

De las transparencias de clase

1. Supongamos que usamos las siguientes líneas de comando

```
pp<- c(1,0,0,1,1,0)
qq<- c("female","female","male","male","female","male")
```

   ¿Qué tipo de variable resultan pp y qq? ¿Son tipo factor? En caso negativo, ¿Cómo podrían definirse como factor?

2. ¿Cómo se calcula el primer cuartil muestral?

3. ¿Qué aspecto de los datos resume la media muestral? ¿y el desvío muestral?

4. ¿Cómo se calcula la mediana muestral? ¿Existe alguna relación entre la media y la mediana muestrales?

5. Supongamos que $x_1, \ldots, x_n$ son n datos, desarrolle las siguientes propuestas.

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

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

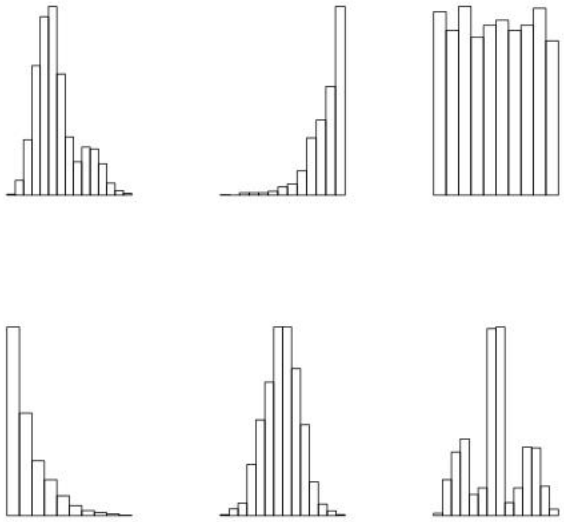
6. ¿En qué difieren un histograma y un gráfico de barras?

7. ¿Qué miramos en un histograma?

8. ¿Cómo caracterizaría estos histogramas?

9. ¿Cuáles son los pasos para hacer un boxplot?

10. ¿Qué miramos en un boxplot?



Ejercicio 1

Airquality

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

11. ¿Cuántas observaciones tiene el conjunto de datos? ¿Cuántas variables?
12. ¿Cuáles son los nombres de las variables?
13. ¿De qué tipo es cada variable?
14. ¿Qué variables tienen datos faltantes?
15. ¿Cuántas observaciones corresponden cada mes?
16. Realizar un histograma de las variables `Wind` y `Temp`. ¿De qué tipo de distribución se trata en cada caso?
17. Realizar un boxplot para las variables `Wind` y `Temp`. ¿Identifica outliers? En caso afirmativo, ¿a qué día y mes corresponden?
18. Realizar un diagrama de dispersión para `Wind` vs. `Temp`. ¿Hay alguna tendencia o asociación?

Ejercicio 2

Titanic

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.

En el presente práctico se trabajará con el conjunto de datos titanic, que figura en el archivo titanic.csv. El 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 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)

Los datos contienen 1028 pasajeros y algunas variables contienen respuestas faltantes.

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

20. Leer el conjunto de titanic.csv teniendo en cuenta que en la primera línea del archivo figura el nombre de las variables y el tipo de separación de los datos y asígnelo al data.frame titanic. Hint: experimentar con `read.csv("titanic.csv",header=T,sep="\t")`.

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

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

23. Establecer el número de variables y de casos.

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

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

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

27. Calcular la chance de sobrevivir por sexo.

28. 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.

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

30. 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.