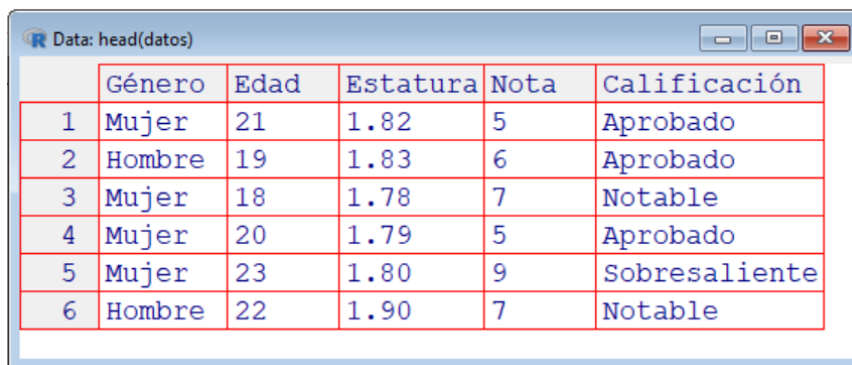


Actividad 1

Introducción al análisis de datos con R

Cargar el fichero de datos ejercicio1.txt, el cual contiene información de un conjunto de estudiantes y cuyas primeras seis filas se muestran a continuación:



The screenshot shows an R console window titled "Data: head(datos)". It displays the first six rows of a data frame with the following columns: Género, Edad, Estatura, Nota, and Calificación.

	Género	Edad	Estatura	Nota	Calificación
1	Mujer	21	1.82	5	Aprobado
2	Hombre	19	1.83	6	Aprobado
3	Mujer	18	1.78	7	Notable
4	Mujer	20	1.79	5	Aprobado
5	Mujer	23	1.80	9	Sobresaliente
6	Hombre	22	1.90	7	Notable

- 1) Mostrar las 10 últimas observaciones.
- 2) ¿Cuál es la estructura de los datos? Indicar dimensión y tipo de variables.
- 3) Calcular la media de las variables univariantes, esto es, de cada columna (en aquellas que se pueda).
- 4) Crear un nuevo *data frame* formado únicamente por los alumnos suspensos. ¿Qué dimensión tiene? Guardarlo y exportarlo en formato .txt o .csv
- 5) Para el *dataset* completo obtener el valor más frecuente, o moda de cada distribución, para las variables 'Edad', 'Estatura' y 'Nota'. ¿Hay alguna bimodal?
- 6) Reordenar el *data frame* en función de la variable 'Nota', de menor a mayor, y mostrar las seis primeras filas.

- 7) Realizar una tabla de frecuencias absolutas y otra de frecuencias relativas para la variable '*Calificación*'. Almacenar las tablas anteriores en dos variables llamadas '*absolutas*' y '*relativas*'.
- 8) Representar la variable '*Calificación*' mediante un diagrama de barras y un diagrama de sectores. Incluir un título adecuado para cada gráfico y colorear las barras y los sectores de colores diferentes.
- 9) Para la variable '*Edad*', realizar un histograma y un diagrama de caja considerando la opción *range* = 1.5. Incluir un título apropiado para cada gráfico. ¿Existe algún valor atípico en esta variable? Reduce el valor del argumento *range* hasta 0.5. ¿Aparece algún atípico? ¿A qué observación corresponde?
- 10) Realizar un resumen de la variable '*Nota*' con el comando *summary*. Comprobar que las medidas que proporciona *summary* coinciden con las medidas calculadas de forma individual usando su función específica.
- 11) Calcular la estatura media de los estudiantes y proporcionar, al menos, dos medidas que indiquen la dispersión de esta variable.
- 12) ¿Qué variable es más homogénea: la '*Edad*' o la '*Estatura*'? Para determinar la homogeneidad de una variable, esto es, la representatividad de su media, se calcula el *Coefficiente de Variación de Pearson* definido como el cociente entre la desviación típica y la media de la variable.
$$CV(x) = \frac{sd(x)}{E(x)}$$
- 13) Obtener la asimetría y curtosis de las variables. ¿Puede asegurarse que las variables siguen una distribución normal? ¿Y la variable multivariante?
- 14) ¿Existe alguna correlación entre la edad y la estatura? ¿Y entre el sexo y la nota?
- 15) Crea dos *dataframes*, uno formado sólo por mujeres cuya nota sea superior o igual a 5 y otro formado sólo por hombres con el mismo criterio.
- 16) Calcular la nota media por género empleando la función *tapply()*.
- 17) ¿Existe algún atípico multivariante? Represéntalos en 3D.

Responder de forma breve las siguientes preguntas

1. ¿Para qué sirve la estadística?
2. Una misma variable, ¿puede ser cuantitativa y/o cualitativa?
3. Una misma variable, ¿puede ser tratada como cuantitativa y/o cualitativa, en un estudio estadístico?
4. ¿Una variable cuantitativa siempre da más información que una cualitativa?
5. ¿Un histograma es lo mismo que un diagrama de barras?
6. ¿Un gráfico vale más que mil palabras?
7. ¿Cuantos más gráficos tenga un trabajo mejor evaluado será?
8. Un trabajo en el cual sólo aparece la media, sin ninguna medida de dispersión, ¿es un trabajo sin rigor científico?
9. ¿En qué casos debemos utilizar la media y en cuáles no?
10. ¿Puede ser más útil la mediana que la media?
11. Si la medida de dispersión nos dice que la media es adecuada, ¿debemos incluir siempre, en la publicación (trabajo, proyecto), la media en lugar de la mediana?
12. Unas publicaciones presentan $\text{media} \pm \text{desviación típica}$, y otras, $\text{media} \pm \text{error estándar}$. ¿Cuál debo utilizar? ¿Cuál es mejor?
13. ¿Sería correcto escribir en una publicación (trabajo, proyecto) la mediana acompañada de la desviación típica?
14. En general, en los *box-plot* se representa la caja con una raya dentro y unos bigotes. Pero a veces aparece un rombo pequeño dentro. ¿Qué significa?