Práctica N° 3 - Análisis exploratorio de datos

Medidas de Resumen

Ejercicio N° 1

El dataset **`winequality-red.csv`** contiene un conjunto de variables relacionadas con propiedades fisicoquímicas que fueron determinadas sobre una serie de vinos de una misma variedad, así como un puntaje asignado en cada caso por un panel de enólogos en sesiones de cata. Importe el dataset al entorno de trabajo y realice cualquier tipo de limpieza y adecuación del mismo que considere necesaria para su posterior análisis.

1. Clasifique las variables del dataset en cualitativas, cuantitativas discretas y cuantitativas continuas.
2. El 25% de los vinos del dataset tiene un contenido de alcohol superior a… ¿qué valor?
3. Realice una tabla en la que se presenten, para las variables densidad y pH, **únicamente** las siguientes medidas descriptivas: media, mediana, desvío estándar y rango intercuartil. A continuación, responda a las siguientes preguntas **sin realizar ningún gráfico**:

- ¿Cómo describiría ambas distribuciones en relación a sus características de simetría?
- ¿Cuál de los dos conjuntos de observaciones (densidad o pH) presenta mayor variabilidad?

4. Represente la distribución de las observaciones de la variable contenido de alcohol (`alcohol`) a través de un boxplot. *Sugerencia:* utilice la función `sns.boxplot()` de la librería `seaborn`. Basándose en el gráfico, ¿cuál de las siguientes medidas de posición o centralidad (media aritmética/mediana) le parece más adecuada para describir a esta variable?
5. Realice una tabla de frecuencias para resumir la distribución de los vinos del dataset en función del puntaje asignado según su calidad (`quality`).

- ¿Cuál de los puntajes fue recibido por una mayor cantidad de vinos?
- ¿Qué porcentaje de los vinos de la muestra recibieron la calificación más baja?

Ejercicio N° 2

El dataset **`alimentos.csv`** fue elaborado por una clínica de nutrición que suministró a sus pacientes una lista de alimentos permitidos con sus respectivos contenidos calóricos. También se detalló el tipo de alimento del que se trataba (fruta, verdura, etc.) y el tipo de vitamina que aportaba cada uno (A, B o C).

Por otra parte, la nutricionista a cargo del estudio lleva una planilla de control de la evolución de 50 pacientes (**`pacientes.csv`**) en la que registra la edad, el sexo, la altura, el peso inicial y el peso final de cada uno de ellos luego de seguir un plan de dieta por una cierta cantidad de tiempo, información que también fue registrada en el campo “tiempo de tratamiento”.

1. Importe ambos datasets al entorno de trabajo y realice cualquier tarea de limpieza y/o adecuación de los mismos que considere necesaria.
2. En relación al campo `aporte_calorico_kcal` informe las medidas descriptivas que le brinden información sobre los siguientes aspectos:

- Las kcal que aportan, en promedio, los alimentos que forman parte del dataset.
- Aquel valor de aporte calórico tal que el 50% de los alimentos del dataset presentan aportes calóricos menores o iguales a él.
- El rango en el que se encuentra el 50% central de las observaciones.
- El o los valores que se presentan con mayor frecuencia entre las observaciones.

3. Represente la distribución de las observaciones de la variable `aporte_calorico_kcal` a través de un boxplot.

- Identifique en el gráfico la mediana, el primer y el tercer cuartil.
- ¿Cómo caracterizaría a la distribución en relación a sus características de simetría?
- En función a lo observado, ¿qué par de medidas de centralidad/posición (media aritmética - mediana) y de dispersión (rango intercuartil - rango - desviación estándar) le parece más adecuada para describir a este conjunto?
- ¿Existe alguna observación que pueda ser considerada como atípica? En caso de respuesta afirmativa, ¿cuántas observaciones recibirían esta calificación?

4. ¿Qué tipo de alimento presenta la mayor mediana de aporte calórico?
5. Realice un boxplot múltiple para representar la distribución de los aportes calóricos de alimentos de los siguientes tipos: frutas, verduras y alimentos elaborados.

- ¿Qué tipo de alimentos presenta valores calóricos más variables y cuál menos variables?
- ¿Qué medida descriptiva utilizó para responder a estas últimas preguntas?

6. Utilizando los datos de los/las pacientes, genere una variable que corresponda a la variación de peso para cada paciente a lo largo del tratamiento (`peso_final_kg` - `peso_inicial_kg`).

- Represente la distribución de los valores observados de la variable "diferencia de peso" para ambos sexos a través de un boxplot múltiple.
- ¿Qué medida descriptiva utilizaría para comparar los resultados del tratamiento entre personas de ambos sexos? En función de su respuesta, ¿las personas de qué sexo obtuvieron los mejores resultados para el tratamiento?

Ejercicio N° 3

Teniendo en cuenta la variable `altura_m` que se encuentra en el dataset **`pacientes.csv`** trabajado en el Ejercicio anterior, genere una tabla de frecuencias en la que las observaciones se encuentren segmentadas en subintervalos de 10 cm de amplitud que estén "cerrados por izquierda", es decir, que tengan la forma **[extremo_inferior, extremo_superior)**.

La tabla de frecuencias generada deberá contener columnas en las que se especifiquen las frecuencias absolutas, relativas y relativas acumuladas correspondientes a cada subintervalo.

¿Qué porcentaje de las personas del dataset tienen una altura **menor a 1.8 m**?

Ejercicio N° 4

El dataset `wine_quality.xlsx` contiene información acerca del puntaje que un panel de enólogos asignó a una serie de 76 vinos de tipo *Pinot Noir*. Las cualidades evaluadas incluyeron algunas propiedades organolépticas como claridad (*clarity*), aroma (*aroma*), cuerpo (*body*) y sabor (*flavor*) y una valoración de la calidad general del vino (*quality*). Adicionalmente, se recabó información sobre el grado de envejecimiento (*aging*) de cada uno de los productos evaluados, la cual se encuentra en el dataset `wine_aging.csv`.

1. Importe ambos datasets y realice cualquier tarea de limpieza y adecuación de los mismos que considere necesaria para su posterior análisis.
2. ¿Cuál es el tipo de vino (crianza/reserva/gran reserva) que presenta la mayor mediana para el sabor?
3. Construya la matriz de covariancia de las distintas variables cuantitativas que componen el dataset y comente qué tipo de información le aporta acerca de la relación entre los distintos pares de variables cuantitativas del dataset.
4. Construya la matriz de correlación de las distintas variables cuantitativas que componen el dataset. En base al mismo, identifique la/s variable/s que se encuentran más fuertemente correlacionadas e informe e interprete la medida de asociación lineal correspondiente
5. Elija el par de variables que identificó en el ítem anterior como aquellas que se encuentran más fuertemente correlacionadas linealmente y realice un gráfico que le permita visualizar la relación general que existe entre las mismas.

Ejercicio N° 5

Utilizando el dataset **`calidad_producto.csv`**, que contiene dos variables registradas en el área de control de calidad de una industria:

- `desviacion_largo`: desviación del largo del producto respecto a un valor estándar o deseado.
- `indice_calidad_producto`: índice o puntuación que se construye a partir de una serie de aspectos relacionados a la calidad general del producto.

1. Calcule los coeficientes de correlación de Pearson y Spearman entre ambas variables. Interprete los valores obtenidos en relación al tipo de información que le brinda cada uno acerca del grado de asociación entre las variables.
2. Construya un gráfico que le permita visualizar la relación general que existe entre las variables analizadas. ¿Qué observa?
3. Calcule nuevamente ambos coeficientes sin tomar en cuenta los registros que incluyan observaciones potencialmente atípicas. ¿Cómo resultan los valores obtenidos en comparación con los calculados en el ítem 1?
4. En función de las características de la relación entre ambas variables que se observan gráficamente, ¿cuál de las dos métricas informaría para describir en forma cuantitativa el grado de asociación entre ellas?