Unidad 3 - Análisis exploratorio de datos: Medidas de Resumen

Fundamentos de ciencia de datos

Ejercicio N°1

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, el cual también fue registrado en el campo “tiempo de tratamiento”.

1. Importe los datasets al entorno de trabajo y realice una descripción general de los mismos que incluya el tipo y rango de datos que componen cada columna y el número de registros.
2. Calcule la media, la moda, la mediana, la varianza, la desviación estándar y la MAD del campo “aporte_calorico_kcal”.
3. Para el mismo campo, calcule Q1, Q3 y el rango intercuartil. Luego calcule los percentiles del 25, 50 y 75%. Estos últimos valores, ¿coinciden con alguno/s

calculado/s previamente?

4. Represente la distribución de los valores observados de la variable “aporte_calorico_kcal” a través de un histograma y grafique sobre el mismo la moda, la media y la mediana.
5. Represente la distribución de los valores observados de la variable “aporte_calorico_kcal” a través de un boxplot. Identifique en el gráfico la mediana, el primer y el tercer cuartil y el rango intercuartil. ¿Cómo caracterizaría a la distribución en relación a sus características de simetría?
6. En función de los gráficos realizados en los puntos anteriores, ¿existe alguna observación que pueda ser considerada como atípica? En caso afirmativo, elimine dichas observaciones en el dataset y vuelva a realizar el gráfico del ítem anterior. ¿Detecta algún cambio?
7. Calcule la media y la mediana de la variable “aporte_calorico_kcal” según el tipo de alimento.
8. Realice un boxplot para representar la distribución de los aportes calóricos según el tipo de alimento. ¿Cuál es la categoría de alimentos que parece aportar menos calorías? ¿Qué categoría de alimentos aporta valores calóricos más variables y cuál menos variables? ¿Qué medida descriptiva le aporta información para responder a estas últimas preguntas?
9. Utilizando los datos de los/las pacientes agregue una columna al dataset en la que se calcule la variación del peso corporal para cada paciente (peso final - peso inicial) y represente los valores observados en función del género a través de un boxplot. ¿En que género considera que se obtuvieron los mejores resultados para el tratamiento? ¿Existen valores atípicos en la distribución de la variación de peso corporal para alguno de los dos géneros?

Ejercicio N°2

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.

1. Importe el dataset al entorno de trabajo y realice cualquier tipo de limpieza y adecuación del mismo que considere necesaria para su posterior análisis,

incluyendo manejo de valores faltantes y de datos duplicados y/o potencialmente erróneos.

2. Realice una tabla en la que se presenten, para las variables densidad y pH, las siguientes medidas: media, mediana, desvío estándar y rango intercuartil.
3. Represente la distribución de la variable contenido de alcohol (”alcohol). En función del gráfico realizado, ¿cuál de las siguientes medidas de posición o centralidad (media aritmética/mediana) le parece más adecuada para describir a esta variable?
4. Realice un gráfico que le permita visualizar 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°3

Importe y explore el conjunto de datos `titanic.csv`

1. Realice una descripción general del conjunto de datos que incluya la descripción de la información brindada por cada columna, el tipo de datos que contiene cada una y el número de registros.
2. Realice cualquier tipo de limpieza y adecuación del dataset que considere necesaria para su posterior análisis, incluyendo manejo de valores faltantes y de datos duplicados y/o potencialmente erróneos.
3. Calcule la media, la mediana y la desviación estándar de la edad de los/las pasajeros/as que murieron y sobrevivieron para cada clase. Realice un boxplot que muestre la distribución de edades para cada grupo (murieron/sobrevivieron) dentro de cada clase. ¿En qué clase las edades de las personas que sobrevivieron fueron más variables? ¿Cuál fue la edad de la persona más joven que sobrevivió en tercera clase?.
4. Represente gráficamente la distribución de los precios de los pasajes en función de la clase del pasajero y calcule el promedio, la moda, la mediana, la desviación estándar y el rango intercuartil del precio del pasaje para cada grupo. ¿En qué clase los precios de pasaje presentaron una mayor variabilidad?
5. a. ¿Qué medida resumen calcularía si quisiera conocer aquel valor que representa el precio que sólo el 25% de los pasajeros superaron a la hora de comprar su boleto?

b. Identifique cuáles fueron los pasajeros que pagaron un pasaje igual o más caro que el valor calculado en el ítem anterior. Construya una tabla en la que se informen los nombres de estas personas, el número total de personas vinculadas a ellas que se encontraban en el barco y la ciudad en la que embarcaron.

c. En base a la tabla construida en el ítem anterior, ¿con cuántos acompañantes, en promedio, viajaban estos pasajeros? ¿En qué puerto embarcó la mayoría de ellos?

6. Construya una tabla en la que se resuma la distribución de pasajeros del Titanic en función de la clase en la que viajaron. La misma debe contener la siguiente información (en distintas columnas): cantidad de pasajeros/as que viajaron en cada clase y porcentajes en relación al total. ¿A qué clase pertenecía la mayoría de los pasajeros del Titanic?
7. Construya una tabla de contingencia cruzando las variables “survived” y “Pclass”. ¿Qué proporción de personas de cada clase sobrevivieron al naufragio del Titanic? Represente gráficamente esta información en un gráfico de barras.
8. a. Categorice la variable edad en los siguientes grupos etarios: 0-18 años, 19-35 años, 36-56 años y >57 años.

   b. Construya un gráfico de barras que muestre la cantidad de personas de género masculino y femenino que sobrevivieron y murieron según el rango etario definido anteriormente.

Ejercicio N°4

1. Vuelva a cargar el dataset `alimentos.csv` y construya una tabla de frecuencias que resuma la cantidad de alimentos que aportan vitamina A, B y C y que incluya la frecuencia relativa de alimentos que pertenecen a cada grupo. ¿Qué proporción de alimentos aportan vitamina A?
2. Construya un gráfico de barras para representar la información anterior, en el que en el eje “y” se represente el porcentaje de alimentos y en el eje “x” el tipo de vitamina que aporta.

Ejercicio N°5

Utilizando el set de datos winequality-red.csv

1. Construye la matriz de covarianza entre todas las variables numéricas y grafícala.
2. Encuentra los 5 pares de variables que tienen la mayor covarianza positiva y negativa en el set de datos, grafíquelas y describa en palabras como se relacionan.
3. Construye la matriz de correlación de Pearson entre todas las variables numéricas y grafícala.
4. Encuentra los 5 pares de variables que tienen la mayor correlación positiva en el set de datos, grafique la relación y describa en palabras como se relacionan. Repita la operación, pero con los 5 pares de variables que tienen la mayor correlación negativa. ¿Las variables con la mayor covarianza (positiva o negativa) coinciden con las de mayor correlación? ¿Pueden existir dos variables que tengan un alto índice de covarianza pero no estén correlacionadas entre sí?
5. Calcule la matriz de correlación de Spearman y compárela con la matriz de Pearson construida en el punto d). ¿Qué puede concluir a cerca de la forma en que se correlacionan las variables? ¿Qué información podría aportar al análisis construir la matriz de correlación de Spearman?

Ejercicio N°6

Utilizando el dataset “calidad_producto.csv”

1. Genere la matriz de correlación y calcule el coeficiente de correlación de Pearson entre las dos variables.
2. Calcule el coeficiente de correlación de Spearman, ¿son similares los resultados?
3. Realice un gráfico de dispersión de un variable vs. la otra, observe los puntos. Busque y remueva valores atípicos en el dataset y vuelva a calcular ambos coeficientes. ¿que observa?

Ejercicio N°7

Utilice el dataset “estación_meteorologica.csv”

1. Calcule y grafique las matrices de correlación y covarianza.
2. Identifique las variables más correlacionadas y grafíquelas una vs. la otra en un gráfico de dispersión. Luego grafique cada una a lo largo del tiempo (para todas

las fechas).

3. Identifique las variables menos correlacionadas y grafíquelas una vs. la otra en un gráfico de dispersión. Luego grafique cada una a lo largo del tiempo (para todas las fechas).

¿Qué puede concluir de los gráficos del punto 2 y 3?