Unidad 4 - Análisis exploratorio de datos: Visualizaciones

Fundamentos de ciencia de datos

Ejercicio N° 1

El set de datos `viajes_tup.xlsx` contiene información sobre el número de viajes mensuales registrados en el Transporte Urbano de Pasajeros (TUP) de la ciudad de Rosario entre los años 2015 y 2021.

1. Realice una tabla que resuma el total de viajes realizados por año y represente gráficamente dicha información. ¿Cuál fue el año en el que se registró la mayor cantidad de viajes en el TUP?
2. Construya un gráfico en el que se represente la evolución del número de viajes registrados en el TUP a lo largo de los meses para los años 2019 y 2020. Comente brevemente lo observado.

Ejercicio N° 2

Utilizando el dataset `partos2022.txt` , el cual contiene información sobre los partos atendidos en el 2022 en el Hospital Roque Sáenz Peña (HRSP) y la Maternidad Martin (MR), efectores municipales de la ciudad:

1. Indique los meses en los que se registró la mayor y la menor cantidad de partos atendidos. ¿Qué porcentajes del total de partos atendidos en el año

representan?

2. Represente gráficamente la distribución del número de partos atendidos en el 2022 según el efector. ¿Qué puede decir acerca de la institución en la que tuvieron lugar los partos?

3. Realice un gráfico que permita comparar la distribución del peso de los recién nacidos entre las distintas categorías de la edad gestacional. ¿Qué observa?

4. a) Realice una descripción general de la variable rango etario de la madre, incluyendo tipo de variable, valores que toma, distribución en la muestra y presencia de datos faltantes.

   b) Recategorice la variable `rango_edad_mama` de la siguiente manera: 10-19 años, 20-29 años, 30-39 años y 40 años o más.

   c) Construya un gráfico de barras lado a lado que ilustre la distribución general del tipo de parto según el rango etario de la madre, en el que los porcentajes de cada categoría se encuentren calculados **sobre el total general de partos atendidos para los que se cuenta con información sobre la edad de la madre (n = 4577)**.

   d) Construya un gráfico de barras lado a lado que muestre la distribución del tipo de parto según el rango etario de la madre, en el que los porcentajes de cada categoría se encuentren calculados **sobre el total de partos atendidos para cada uno de estos grupos etarios**.

   e) Compare los gráficos realizados en los ítems b y c. ¿Qué tipo de información brinda cada uno?

   f) ¿Cuál de los gráficos anteriores le permite analizar si la edad de la madre influye en la probabilidad de recurrir a una cesárea como método de parto? ¿Qué observa?

Ejercicio N° 3

El dataset `iris.csv` contiene información sobre 150 flores de iris de tres especies diferentes: *setosa*, *versicolor* y *virginica*. Para cada flor, se midieron cuatro características: longitud y ancho del sépalo (la parte que rodea y protege el capullo de la flor) y longitud y ancho del pétalo (la parte coloreada de la flor).

1. Construya una tabla que contenga, para cada una de las cuatro variables cuantitativas, las siguientes medidas descriptivas: media aritmética, desvío estándar, mediana, rango intercuartílico y los valores mínimo y máximo.

2. Construya un gráfico que le permita visualizar la distribución de los valores observados del ancho de sépalo. A partir del gráfico realizado, ¿qué puede decir acerca de la simetría de la distribución?
3. Realice un gráfico que permita comparar la distribución del largo del pétalo de las flores entre las distintas especies. Comente brevemente lo observado.
4. a) Construya un gráfico que le permita analizar la relación general que existe entre las variables ancho y largo del pétalo. ¿Qué observa?

   b) Modifique el gráfico realizado en el ítem anterior de tal manera que le permita analizar si la relación general entre el ancho y el largo del pétalo se mantiene según la especie. Comente brevemente lo observado.
5. a) Construya una matriz de gráficos que le permitan estudiar la asociación que existe entre todos los pares de variables cuantitativas del dataset. Genere la matriz de correlación lineal de Pearson y represéntela gráficamente.

   b) A partir de lo realizado en el ítem anterior, caracterice el grado de asociación entre los distintos pares de variables de interés, incluyendo tipo, fuerza y dirección y analizando la correspondencia entre los valores calculados y lo observado gráficamente.

Ejercicio N° 4

El dataset `registro_temperatura365d_smn.csv` contiene las temperaturas máximas y mínimas registradas diariamente entre el 04/05/2022 y el 03/05/2023 en todas las estaciones meteorológicas de superficie pertenecientes al Servicio Meteorológico Nacional.

1. a) Construya un gráfico que le permita comparar las distribuciones de temperaturas mínimas y máximas diarias entre los últimos 12 meses (mayo 2022 a abril 2023) registradas en la estación del Aeropuerto Rosario (”ROSARIO AERO”). Describa brevemente lo observado.

   b) ¿Cuál fue el mes del último año con la mayor temperatura máxima mediana y cuál el que presentó la menor temperatura mínima mediana? Informe ambos valores e interprételos.
2. Realice nuevamente el ítem 1a) con los datos correspondientes a la estación meteorológica localizada en la Base Marambio de la Antártida Argentina. Compare los dos gráficos y comente las diferencias que encuentra en las distribuciones de las temperaturas registradas en ambas estaciones.