Tarea 2 - Visualización de Datos

Instructor: Jesús Adolfo Mejía de Dios

1 de septiembre de 2022

Instrucciones: Realice las siguientes actividades en equipos con máximo cuatro integrantes.

Fecha de Entrega: La fecha límite de entrega de esta tarea es el viernes 9 de septiembre de 2022 antes de la 1PM (hora local). Por cada 24 horas de retraso, la tarea perderá un 20% de su valor original. Note que hay actividades opcionales con puntos extra. Los puntos extra sólo se tomarán en cuenta si todos los demás ejercicios se realizaron.

Modo de entrega: La tarea se deberá enviar al correo `jesusmejded@gmail.com` con el asunto "**Tarea 2: Visualización de datos**" e indicar en el cuerpo del correo los nombres y matrículas de cada integrante del equipo.

Actividad 1

Lea el siguiente artículo sobre la creación de *cheat sheets* (hojas de trucos o infografías) para técnicas de visualización de datos.

> Wang, Z., Sundin, L., Murray-Rust, D., & Bach, B. (2020, April). Cheat sheets for data visualization techniques. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Realice una *cheat sheet* (o infografía) de **al menos dos** de los siguientes conceptos:

a. (2 puntos) Gráfica de Cajas.
b. (2 puntos) Gráfica de Dispersión.
c. (2 puntos) Gráfica de Violín.
d. (2 puntos) Histograma.

Los principales puntos a evaluar en esta actividad son la originalidad y la utilidad de la hoja de trucos. De esta actividad se podrán obtener como **máximo 5 puntos** aunque realice los cuatro productos. Dichos productos lo deberá anexar en formato PDF, JPG, PNG o similares (un archivo por cada *cheat sheet*).

Fuente de información sugerida:

- Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *The VLDB Journal*, 29(1), 93-117.
- Telea, A. C. (2014). Data visualization: principles and practice. *CRC Press*.

Actividad 2

Realizar una libreta (`notebook`) en Pluto que contenga lo siguiente:

1. (1 punto) Agregue un botón que permita abrir un archivo CSV, el cual sea leído como un `DataFrame`.
2. (1 punto) Calcule las medidas de tendencia central y de dispersión que considere relevantes.
3. (1 punto) Describa con sus propias palabras lo que están indicando dichos valores numéricos y que puedan aportar a la explicación de los datos.
4. (1 punto) Realice una gráficas de caja por cada variable numérica del conjunto de datos leído.
5. (1 punto) Interprete y describa los resultados mostrados en las gráficas de cajas generadas.
6. (1 punto) Responda las siguientes preguntas: ¿Hay datos faltantes? ¿Hay datos atípicos en alguna variable? ¿Qué técnica de pre-procesamiento sugeriría y por qué?
7. (1 punto extra) Genere además otro tipo de gráficas que considere importante para comunicar mejor las características de los datos. Describa la interpretación de dichas gráficas generadas.

El conjunto de datos a utilizar en esta actividad es conocido como *Pima Indian Diabetes* y lo puede descargar desde cursos.ooptim.com/visualizaciondedatos

Para esta **Actividad 2** debe descargar el `notebook`, ya sea en formato `JL` o formato `HTML`, y anexarlo cuando entregue esta tarea.