Ciencias de Datos 2025 - Trabajo Práctico 1

ANÁLISIS DE DATOS CON PANDAS

En este ejercicio deberá analizar un conjunto de datos utilizando el paquete pandas de python.

- **1.** Acceso a datos

  - Busque en Kaggle la base de datos Chocolate Sales Data.
  - Descargue el archivo y guárdelo en un lugar adecuado en su Drive de la cuenta institucional de la UNC.
  - Cree una instancia de Google Colab, en donde resolverá los ejercicios y la cual deberá compartir para ser evaluado.
  - Verifique que el archivo existe en la ubicación que corresponda desde Colab.
  - Lea el archivo y guárdelo en un DataFrame de pandas.
  - Calcule la cantidad de filas y de columnas presentes en el archivo.

- **2.** Exploración inicial de datos

  - Identifique qué tipo de dato se encuentra en cada columna.
  - Para cada variable numérica, encuentre los valores mínimos y máximos y las medias muestrales.
  - Para las variables categóricas, encuentre la cantidad de valores únicos.
  - Determine si existen datos faltantes y en ese caso contar cuántos datos faltan en cada columna
  - Calcule el precio por caja y agregue este dato como columna en el DataFrame

- **3.** Realice las siguientes visualizaciones:

  - gráfico de barras del ingreso por ventas según el tipo de producto
  - histograma del ingreso por ventas

- serie temporal de la cantidad de ventas por día.
- gráfico de dispersión de los ingresos por ventas y la cantidad de cajas vendidas

▶ **4.** Identificación de datos peculiares

- Identifique el país con más ventas de chocolate
- Identifique el producto más popular de cada país
- Calcule la cantidad de vendedores por cada país

▶ **5.** Análisis de datos

Responda las siguientes preguntas, justificando sus respuestas en base a los datos:

- ¿Las ventas de chocolate dependen de la época del año? Analice el país con mayor cantidad de ventas.
- ¿La cantidad de ventas depende del precio de la caja?

Reglas de decisión

▶ **1.** En el problema unidimensional con dos categorías se decide por ω_1 si $x > \theta$, y caso contrario se decide por ω_2.

Calcule la probabilidad de error esperado teniendo en cuenta que:

- La variable X condicional a ω_1 tiene distribución normal con media 3 y varianza 3. Es decir,

$$P(x|\omega_1) \sim N(3,3)$$

- La variable X condicional a ω_2 tiene distribución normal con media 1 y varianza 4.

$$P(x|\omega_1) \sim N(1,4)$$

- Se toma $\theta = 2$
- $P(\omega_1) = 0{,}4$

Explique el resultado gráficamente a partir de una visualización del problema que incluya las densidades de probabilidad $P(x|\omega_1)$ y $P(x|\omega_2)$.