



Fundamentos de Data Science con Python

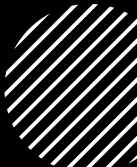
Sesión 2: Análisis Exploratorio de Datos (EDA)

Dr. Julio Lopez-Nunez

Diciembre, 2025



Objetivos de la sesión.



Comprender qué es el Análisis Exploratorio de Datos (EDA).



Detectar problemas comunes en los datos:
valores faltantes,
duplicados, tipos
incorrectos.



Aplicar técnicas de limpieza básica en Python con pandas.



Generar visualizaciones iniciales con matplotlib y seaborn. Y, desarrollar criterios para interpretar patrones y tendencias.

¿Qué es EDA?



Etapa inicial del ciclo de Data Science.



Objetivos principales:



Explorar estructura y calidad de los datos.



Detectar patrones, anomalías y relaciones.



Generar hipótesis iniciales.



Enfoque:



“dejar que los datos hablen” antes de aplicar modelos.

Limpieza de datos.



Problemas frecuentes

- I. Valores nulos (NaN).
- II. Duplicados.
- III. Tipos de datos inconsistentes (ej.: número como texto).

Herramientas pandas

`df.isnull().sum()` → detectar faltantes.
`df.dropna()` → eliminar.
`df.fillna(valor)` → imputar.
`df.duplicated()` y `df.drop_duplicates()`.
`df.astype(tipo)` para conversión.

Medidas estadísticas.



Tendencia central: media, mediana, moda.



Dispersión: varianza, desviación estándar, percentiles.



Biblioteca pandas:

.describe() .mean() .median() .mode().



Reflexión:

¿Qué nos dice realmente cada medida sobre nuestros datos?

Visualización de datos.



Matplotlib: Gráficos de barras, Gráficos de líneas, Histogramas.

Seaborn: sns.histplot(),
sns.boxplot(), sns.scatterplot()

Buenas prácticas:

Usar títulos, etiquetas,
escalas claras.

Evitar gráficos engañosos.

Actividad práctica.



Detectar valores faltantes y duplicados.



Corregirlos (drop o fill).



Calcular medidas de tendencia central y dispersión.

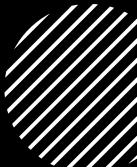


Generar Gráficas:

Distribución, Comparación, Datos Atípicos.



Discusión grupal.



Preguntas guía:



- ¿Qué patrones detectaron en el dataset?
- ¿Qué carrera presenta mayor variabilidad en notas?
- ¿Qué decisiones se podrían tomar a partir de estas visualizaciones?



“If you torture the data long enough, it will confess.”

(Atribuido a Ronald Coase, Premio Nobel de Economía)



¿Qué significa para ti “exprimir los datos”?



¿Dónde está el límite entre analizar y manipular datos?

¿Cómo podemos enseñar a nuestros estudiantes un uso ético de la información?

Cierre de la sesión 2.



Definición de EDA.



Medidas estadísticas básicas.



Visualización con **matplotlib** y **seaborn**.

Limpieza de datos en pandas.