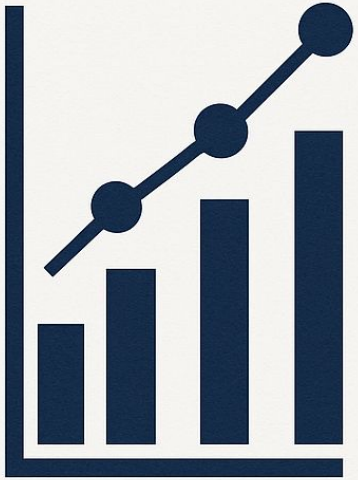
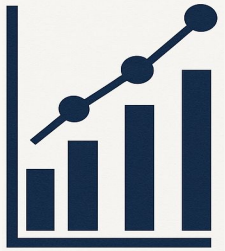


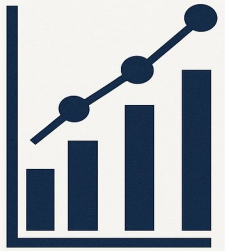
INTRODUCCIÓN AL ANÁLISIS DE DATOS UTN

¿QUÉ VEREMOS HOY?



1. Introducción a la Estadística Descriptiva
2. Medidas de Tendencia Central
3. Medidas de Dispersión
4. Correlación entre Variables
5. Análisis Descriptivo con Pandas y NumPy
6. Visualización de Datos

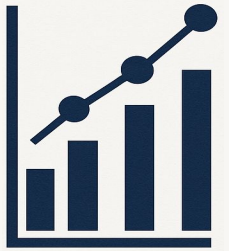
Introducción



La estadística descriptiva es una rama fundamental de la estadística que se enfoca en resumir, describir y organizar datos para facilitar su interpretación. Utilizando herramientas computacionales como Pandas y NumPy en Python, podemos realizar estos análisis de forma rápida y eficiente.

El objetivo es conocer las características esenciales de un conjunto de datos antes de aplicar modelos predictivos o técnicas de minería de datos.

¿Qué es el Análisis Descriptivo?

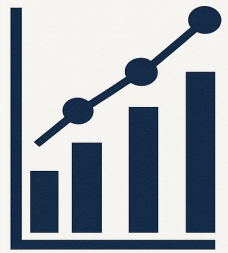


Resumen estadístico para entender y caracterizar datos.

Combina medidas de:

- Tendencia central (media, mediana, moda).
- Dispersión (varianza, desviación estándar, rango, cuartiles).
- Correlación (relación entre variables).

Esquema General del Análisis Descriptivo



El análisis estadístico descriptivo se basa en tres dimensiones principales:

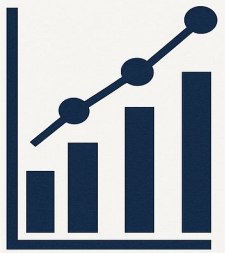
Tendencia Central: indica dónde se ubica el "centro" de los datos.

Dispersión: muestra cuánto se alejan los datos del centro.

Correlación: examina la relación entre dos variables distintas

Estas dimensiones permiten una comprensión integral de cualquier dataset.

Medidas de Tendencia Central



Las medidas de tendencia central buscan representar un conjunto de datos con un solo valor que sea indicativo de su comportamiento general.

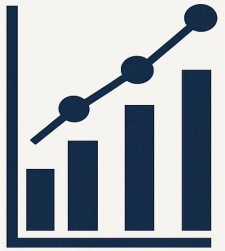
Media: promedio aritmético; es sensible a valores extremos

Mediana: valor central; más robusta ante outliers.

Moda: valor más frecuente; útil para variables categóricas.

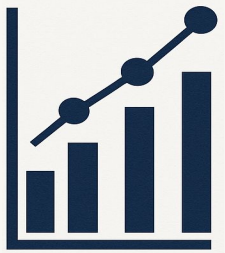
Estas medidas ayudan a resumir los datos en torno a un valor "típico".

Ejemplo en Python - Tendencia Central



```python
import numpy as np
from scipy import stats
ventas = [10, 12, 12, 15, 20, 25, 25, 30, 50]
media = np.mean(ventas)
mediana = np.median(ventas)
moda = stats.mode(ventas)[0][0]
```

Medidas de Dispersión



Estas medidas cuantifican cuánto varían los datos alrededor del centro:

Varianza: promedio de las desviaciones cuadráticas respecto a la media.

Desviación Estándar: raíz cuadrada de la varianza.

Cuartiles e IQR: dividen los datos en partes iguales y ayudan a detectar valores extremos.

Rango: diferencia entre el máximo y mínimo.

Una alta dispersión implica que los datos están muy dispersos del centro.

¿PREGUNTAS?