



ESCUELA POLITÉCNICA NACIONAL

Facultad de Ingeniería en Sistemas



CLASE 2:

Limpieza y preparación de datos:

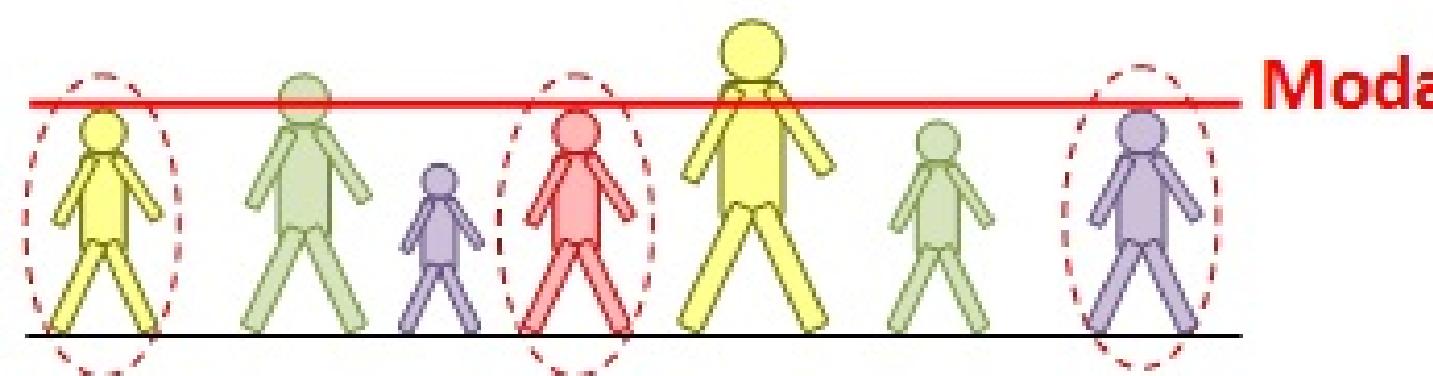
1. Análisis Exploratorio de Datos (EDA)
2. Limpieza e Imputación de Datos
3. Ingeniería y Transformación de Características.
4. Selección de Características.



CONCEPTOS PREVIOS

Medidas de tendencia central

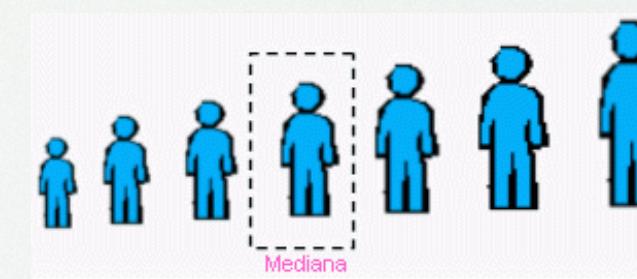
Moda



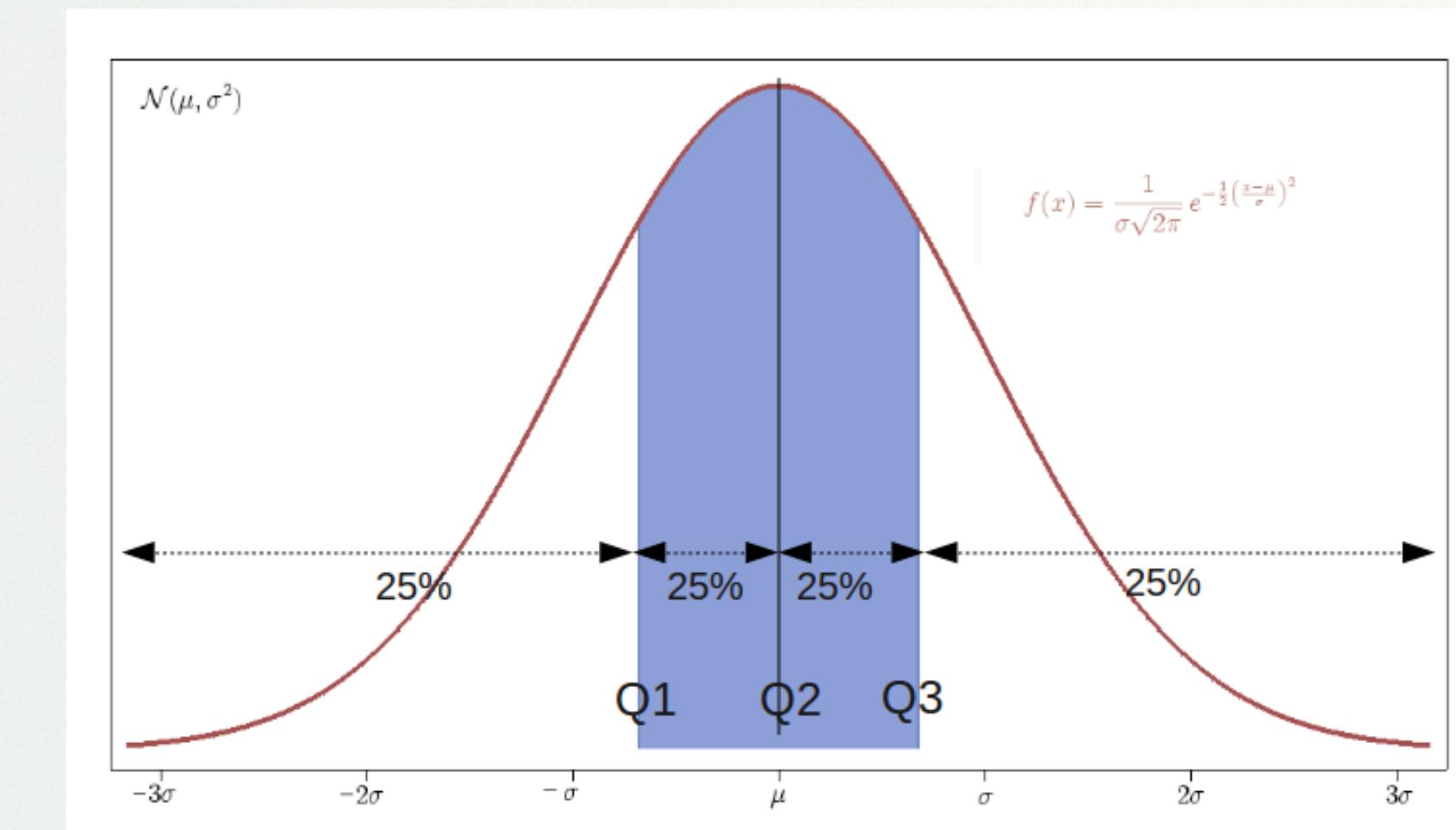
Media

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_n}{N}$$

Mediana



Cuartiles

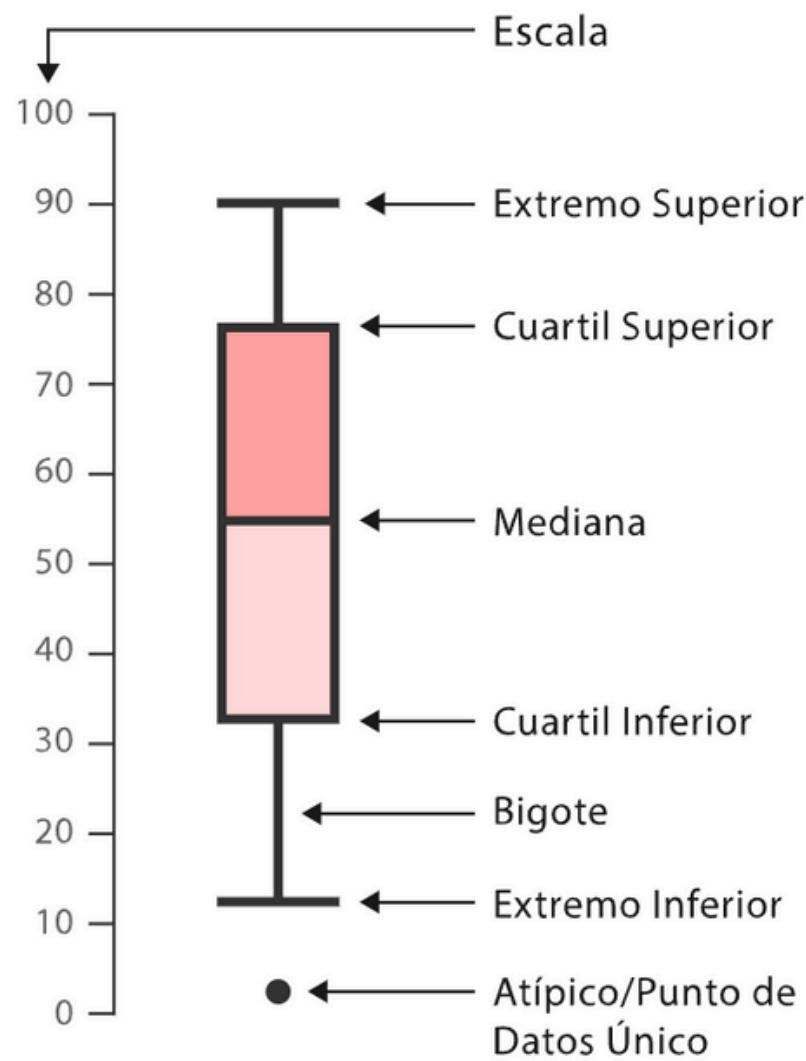


Desviación estándar

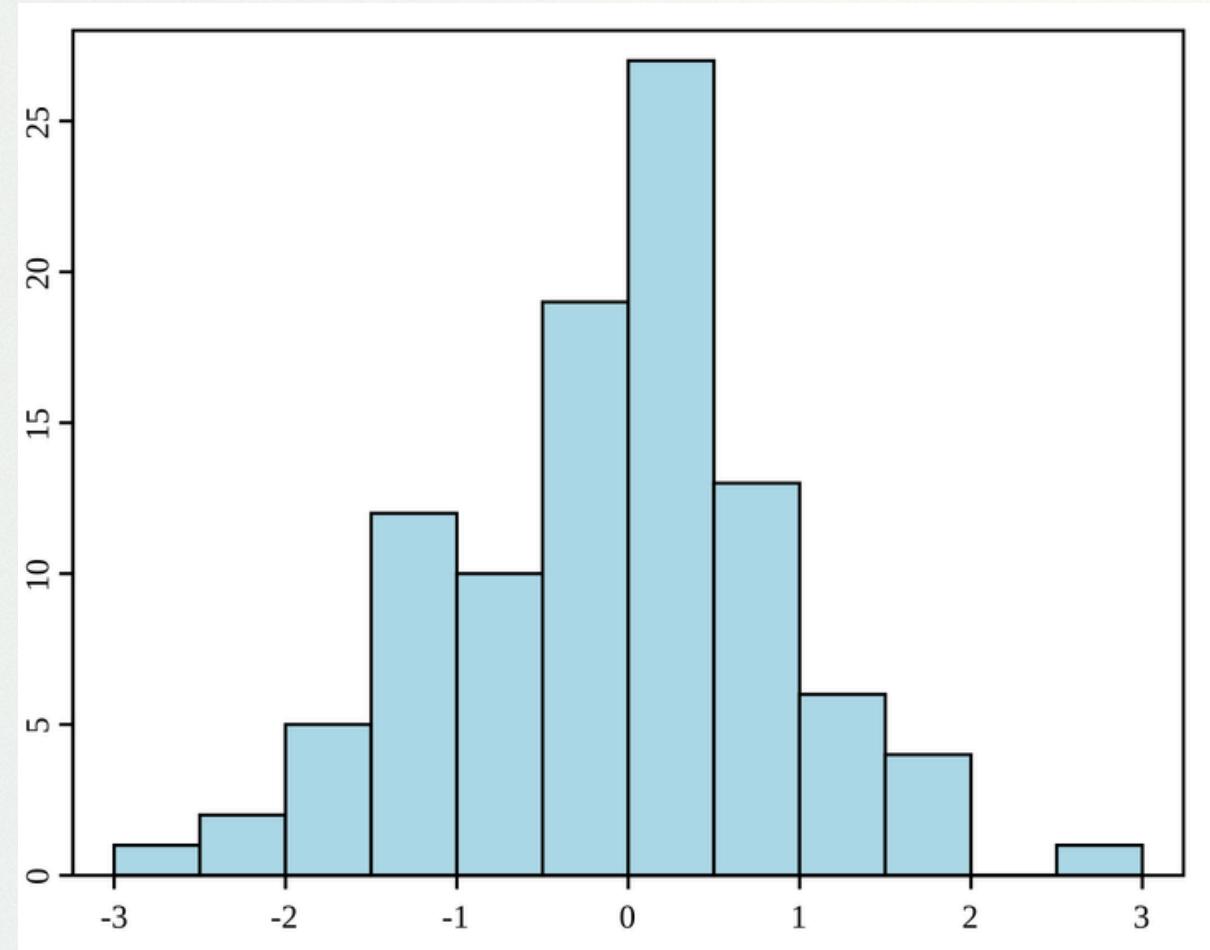


CONCEPTOS PREVIOS

■ Diagrama de cajas



■ Histograma





ANÁLISIS EXPLORATORIO DE DATOS

 Ver tipos de datos y valores no nulos.

 Entender la naturaleza y significado de nuestros datos.

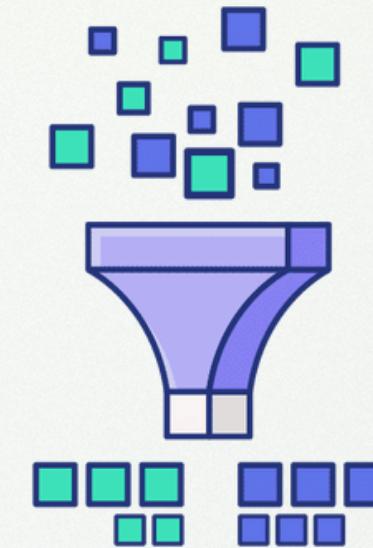
 Análisis estadístico inicial





LIMPIEZA E IMPUTACIÓN DE DATOS

"Garbage In, Garbage Out" (Si entra basura, sale basura).



Tratamiento de Valores Atípicos (Outliers)

- Son valores extremos que pueden distorsionar el modelo.
- Se pueden eliminar, o "acotar"



Manejo de Valores Nulos (Imputación):

- Numéricos: Rellenar con la media (si no hay atípicos) o la mediana (más robusto a atípicos).
- Categóricos: Rellenar con la moda (el valor más común).



Corrección de Tipos de Datos:

Asegurar que las columnas numéricas sean int o float y no object.



FEATURE ENGINEERING



Escalado de Características Numéricas

Necesario para algoritmos sensibles a la escala (Redes Neuronales, SVM, KNN).

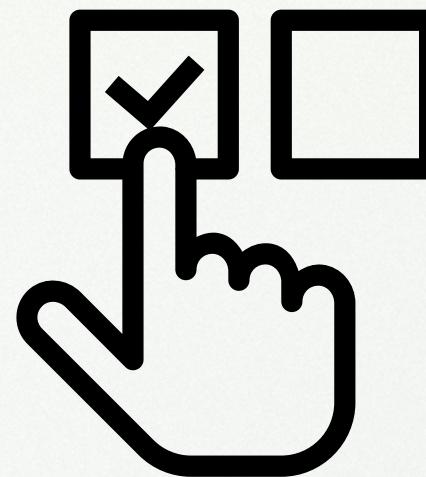


Creación de Nuevas Características

- Combinar variables para crear nuevas (ej. ingreso / miembros_familia).
- Extraer información de fechas (ej. día de la semana, mes).



SELECCIÓN DE CARACTERÍSTICAS



Análisis de Correlación

Buscamos Alta Correlación entre Features:

Buscamos Alta Correlación con el Target:



Reducción de dimensionalidad (PCA)

Reducir el espacio de características conservando la mayor parte de la varianza de los datos.