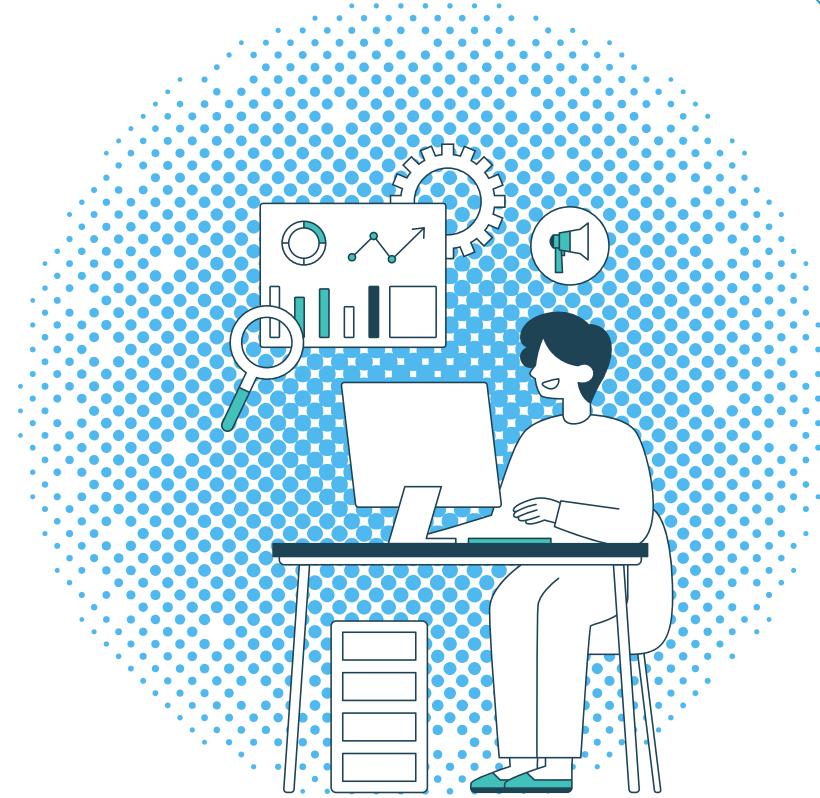


# Análisis de datos

Carrera de Especialización en  
Inteligencia Artificial



# Programa de la materia

## 1 Introducción al análisis de datos

Flujo de trabajo típico. Aplicaciones. Configuración del entorno. Introducción a las principales bibliotecas.

## 2 Análisis exploratorio de datos

EDA, estadística descriptiva y visualización.

## 3 Preprocesamiento de datos y Feature Engineering

Tipos de variables. Codificación. Manejo de datos faltantes. Creación de nuevos features.

## 4 Reducción de dimensionalidad

Métodos para reducir la dimensionalidad y técnicas para selección de features.

## 5 Pruebas estadísticas y validación de hipótesis

Pruebas de normalidad y de correlación. ANOVA y test de dependencia.

## 6 Taller práctico

Análisis de datos completo de un dataset. Buenas prácticas. Discusión grupal.

## 7 Presentación de trabajos finales

Exposición por grupos y devolución de los docentes.

## 8 Automatización del análisis de datos

Herramientas para automatizar el EDA y otras tareas del flujo de trabajo.

# Breve repaso de la clase anterior y espacio para consultas



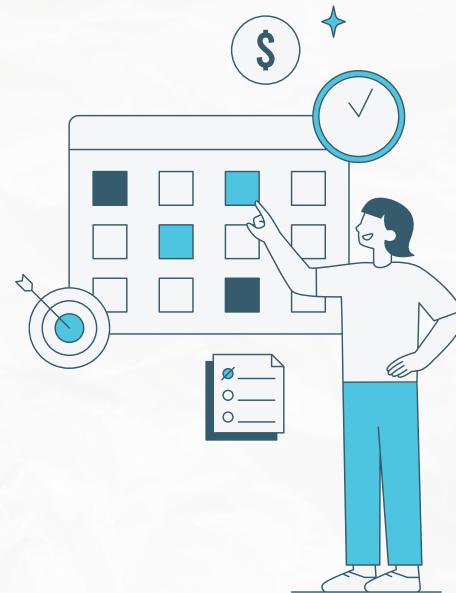


# ¿Qué es el análisis exploratorio de datos (EDA)?

- El análisis exploratorio sirve para resumir, visualizar y entender los datos.
- Ayuda a detectar patrones, anomalías, y descubrir información útil antes de desarrollar modelos de machine learning.

# Variables aleatorias y datasets

- Una variable aleatoria (v.a.) es una forma de mapear los resultados de algún fenómeno aleatorio a números.
- Constituyen una herramienta fundamental para modelar y analizar fenómenos que involucran incertidumbre o aleatoriedad.
- Las columnas de un dataset pueden considerarse como un conjunto de observaciones de una variable aleatoria, tomada de una población más grande.



# Ejemplos de v. aleatorias en el dataset del Titanic

Columna	Fenómeno Aleatorio
<b>Survived</b> (¿sobrevivió?)	Varía aleatoriamente según factores como la ubicación en el barco, el acceso a botes salvavidas, la clase social y decisiones personales durante el naufragio.
<b>Sex</b>	Varía aleatoriamente según la composición de quienes decidieron o pudieron viajar. Influenciada por factores sociales como roles de género o motivos del viaje.
<b>Age</b>	Variable en función de la población que decidió o pudo viajar en el Titanic. Condicionada por el año de nacimiento de las personas y las circunstancias que los llevaron a ese viaje.
<b>SibSp</b> (cantidad de hermanos o esposo/a)	Influenciada por decisiones familiares (ej., viajar juntos), tamaño de la familia y factores sociales de la época.
<b>Fare</b> (tarifa)	Variable según factores como la clase (primera, segunda, tercera), el puerto de embarque o decisiones individuales de compra.
<b>Embarked</b> (puerto de embarque)	Influenciada por el lugar de origen, itinerario personal y disponibilidad de pasajes desde cada puerto.

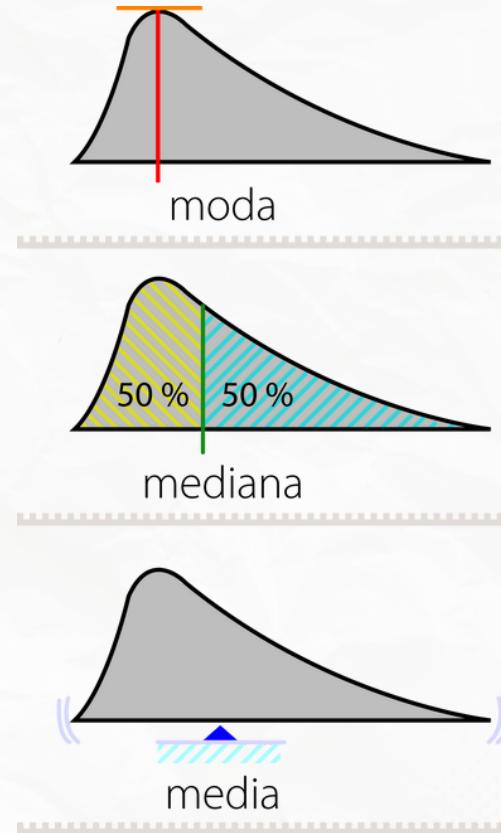
# Estadística descriptiva



- Las métricas descriptivas permiten estimar las características de una variable aleatoria utilizando una muestra de datos.
- La distribución de datos numéricos se puede caracterizar en términos de su tendencia central, su variabilidad (dispersión) y su forma.

# Tendencia central: media, mediana y moda

Son distintas formas de medir el “centro” en un conjunto numérico de datos.



# Media, mediana y moda: ejemplo

Ejemplo (notas de alumnos): {9, 9, 8, 9, 10, 5}

$$\text{Media} = \frac{9+9+8+9+10+5}{6} = 8,33$$

$$\text{Mediana} = \{5, 8, \underbrace{9, 9}_{9}, 9, 10\}$$

(Media de los dos valores del centro por ser cantidad par)

$$\text{Moda} = \underbrace{\{9, 9, 8, 9\}}_{9}, 10, 5\}$$

Puede o no haber moda.  
Puede haber más de una (bimodal, trimodal, multimodal).

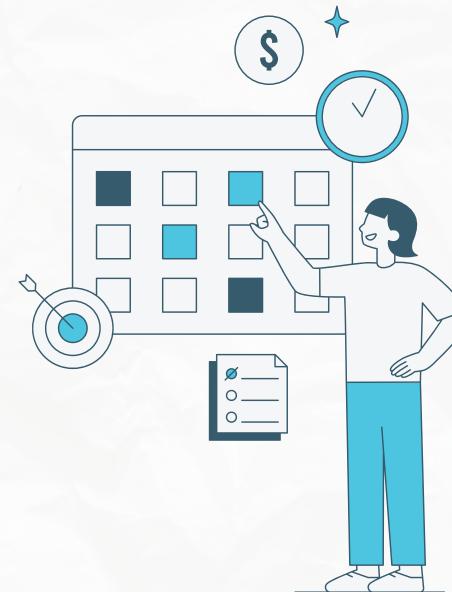
Representa la nota si todos se hubieran sacado lo mismo

La nota del medio. Una mitad de la clase sacó menos de 9 y la otra mitad, más que 9.

La nota más frecuente fue 9.

# Medidas de dispersión: varianza y desviación estándar

- Sirven para medir qué tan “desparramados” están los datos alrededor de la media.
- La desviación estándar suele ser más fácil de interpretar.



# Varianza y desviación est<sup>a</sup>ndar - ej.

$$C1 = \{-10, 0, 10, 20, 30\}$$



$$C2 = \{8, 9, 10, 11, 12\}$$

$$\bar{x}_1 = \frac{-10+0+10+20+30}{5} = 10$$

$$s_1^2 = \frac{(-10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2 + (30-10)^2}{4} = 250$$

$$\bar{x}_2 = \frac{8+9+10+11+12}{5} = 10$$

$$s_2^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{4} = 2,5$$

$$s_1 = \sqrt{s_1^2} = \sqrt{250} = 15,8$$

$$s_2 = \sqrt{s_2^2} = \sqrt{2,5} = 1,58$$

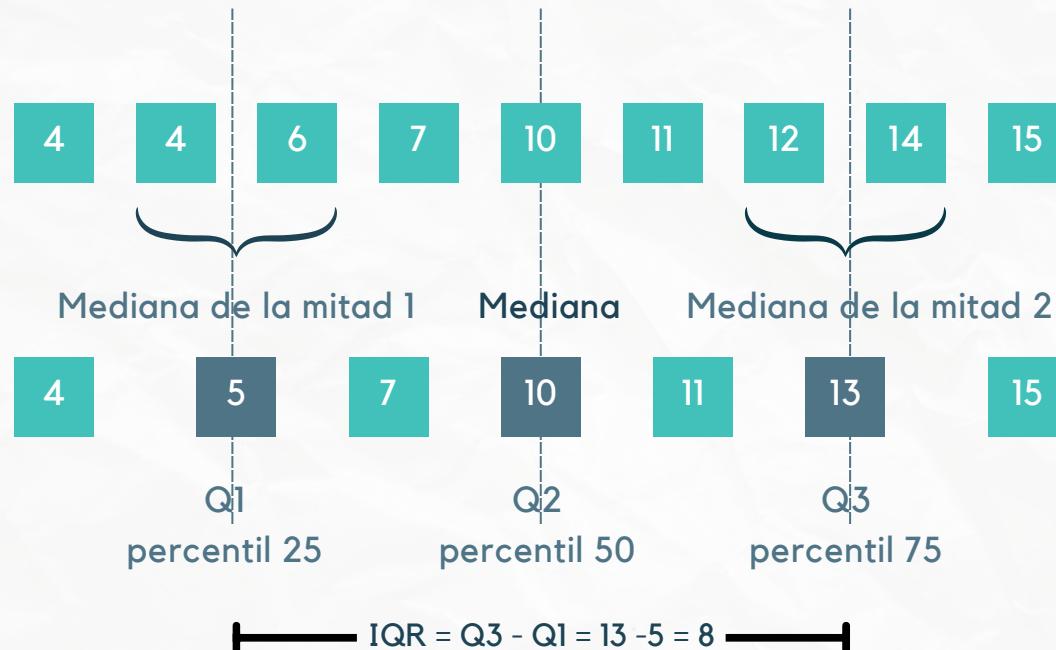
# Cuantiles

- Son valores que dividen un conjunto de datos ordenados en partes iguales.
- Se usan para analizar la distribución de los datos.
- Tipos de cuantiles más comunes:
  - **Cuartiles:**
    - Q1 (25%): El 25% de los datos son menores que este valor.
    - Q2 (50%): La mitad de los datos están por debajo.
    - Q3 (75%): El 75% de los datos son menores que este valor.
  - **Deciles:** dividen los datos en 10 partes iguales.
  - **Percentiles:** dividen los datos en 100 partes.



# Cuartiles y rango intercuartil (IQR)

Ejemplo: {4, 4, 10, 11, 15, 7, 14, 12, 6}



# Evaluación del IQR

- Qué mide: la dispersión del 50% central de los datos.

Convención para evaluarlo:

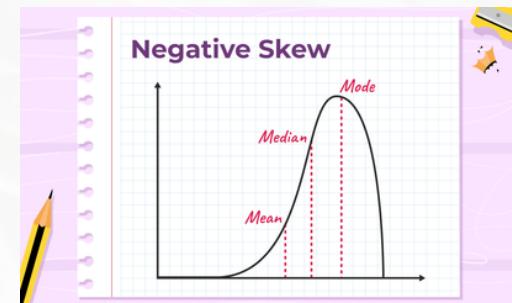
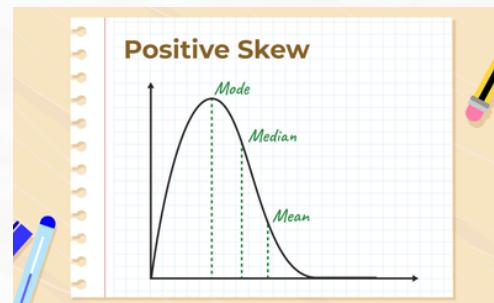
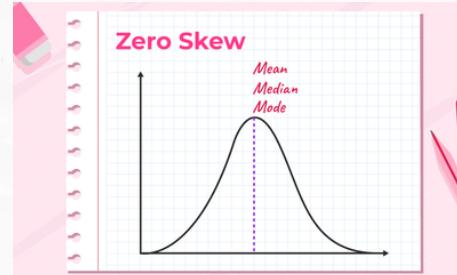
- Pequeño: centro concentrado
- Grande: centro disperso
- Potenciales outliers:
  - Valores menores a  $Q1 - 1,5 \cdot IQR$
  - Valores mayores a  $Q3 + 1,5 \cdot IQR$
- ¿Qué es pequeño o grande?
  - Si  $IQR < 10-20\%$  del rango total ( $\text{máx} - \text{mín}$ ) se considera pequeño



# Asimetría (Skewness)

- Es una medida de la simetría de la distribución de los datos con respecto al valor central.
- Los datos reales casi siempre presentan asimetría.
- En casos de asimetría, la media deja de ser una buena representación del centro.

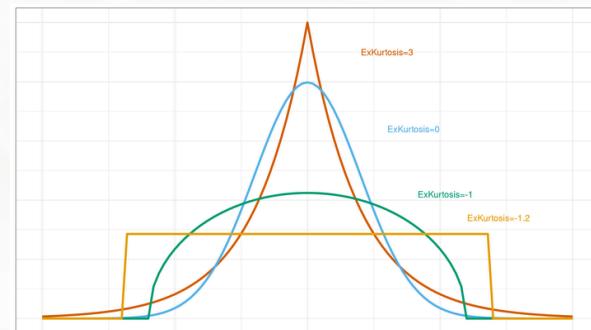
$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$



# Curtosis en exceso

$$\text{CurtosisExceso} = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s} \right)^4 - 3$$

- Indica cuántos datos hay en las “colas” de las distribución y (generalmente) que tan afilado o plano es el pico comparado con una Normal.
- La curtosis en exceso es la más utilizada. Para una distribución Normal estándar, la Curtosis en exceso es 0.
- Una curtosis en exceso positiva puede indicar muchos datos en las colas y, por ende, la presencia de outliers.



# Evaluación de asimetría y curtosis

- Tiene más sentido usarlas para analizar distribuciones continuas y unimodales

Convención para la asimetría:

- Entre -0,5 y 0,5 se considera aproximadamente simétrica.
- Entre -1 y +1 se considera moderadamente asimétrica.
- Menor que -1 y mayor que +1 se considera altamente asimétrica.

Convención para la curtosis:

- Cerca de 0: "mesocúrtica", se considera Normal.
- Mayor que 0: "leptocúrtica" (colas pesadas ¿outliers? - pico habitualmente más afilado, aunque no siempre)
- Menor que 0: "plasticúrtica" (colas menos pesadas - pico habitualmente más plano, pero no siempre)



# Ejemplos prácticos en Jupyter



# Medidas estadísticas descriptivas - resumen



## Tendencia central

1

Media: valor promedio.

2

Mediana: en un conjunto ordenado de números, es el valor que se encuentra en el medio.

3

Moda: valor más frecuente.

## Dispersión

4

Varianza: dispersión de los datos con respecto a la media. Promedia diferencias al cuadrado.

5

Desviación estándar: dispersión con respecto a la media pero en las mismas unidades que los datos.

6

Cuartiles y rango intercuartil (IQR): otra forma de cuantificar dispersión de los datos.

## Forma

7

Curtosis y skewness: describe la forma de la distribución (simetría y la presencia de colas más pronunciadas o aplanadas).