



Sociedad Mexicana
de Materiales A.C.

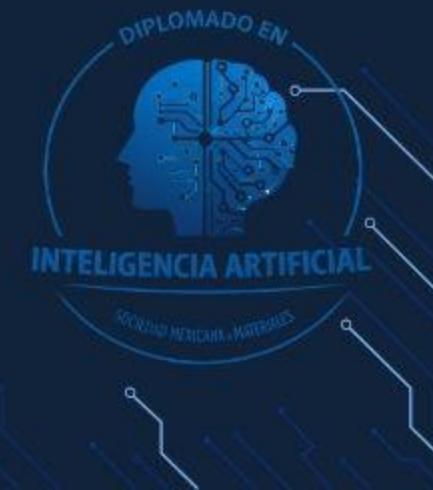
TRONCO COMÚN

Módulo 1:

Introducción a la Minería de Datos

Dr. Irvin Hussein López Nava

Centro de Investigación Científica y de Educación Superior de Ensenada
Facultad de Ciencias, Universidad Autónoma de Baja California

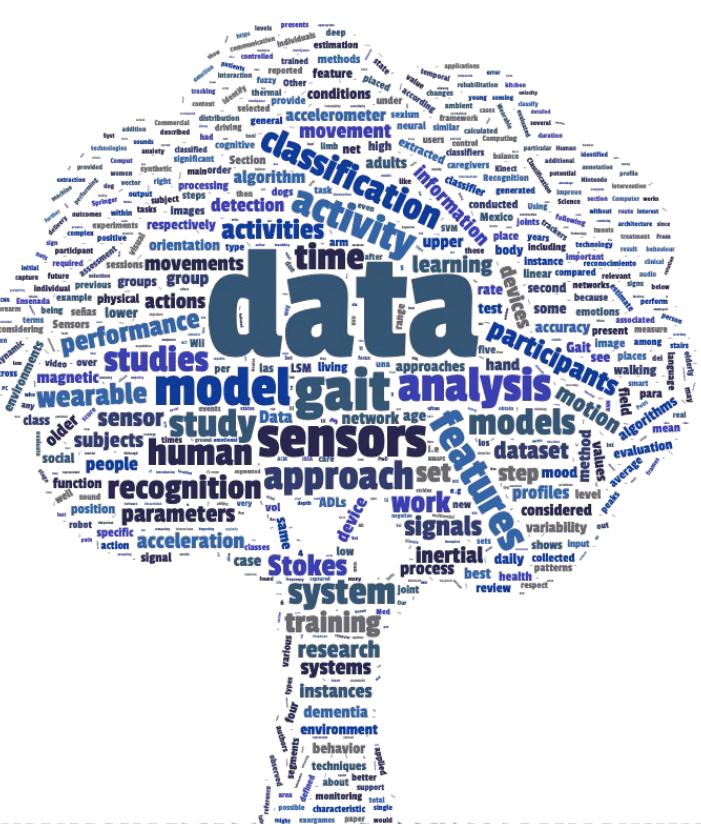




ACERCA DE MÍ

Investigación

desde 2008



Docencia

desde 2006



ACERCA DE MÍ



Laboratorio de Ciencia de Datos y Aprendizaje Automático



TEMARIO

Martes 10/feb

Introducción (1 hora)

- ¿Qué es la minería de datos?
- Conceptos básicos
- Visualización

Ejercicios (2 horas)

- Visualización de datos genéricos
- Exploración de datos reales

Limpieza (1 hora)

- Preprocesamiento
- Reducción de dimensionalidad
- Selección de atributos
- Balanceo de clases

Ejercicios (2 horas)

- Limpieza
- Reducción
- Selección
- Balanceo

Martes 17/feb

Evaluación (1 hora)

- Partición de datos
- Métricas de rendimiento

Ejercicios (1 hora)

- Validación de modelo simple
- Análisis de resultados

Relación con el diplomado

El **diplomado** busca:

- Brindar conocimientos multidisciplinarios en IA.
- Desarrollar competencias teóricas y prácticas.
- Enfocarse en aplicaciones en ciencia de materiales y procesos industriales.

La **minería de datos** contribuye directamente a:

- Identificación y análisis de problemas complejos.
- Preparación y estructuración de datos reales.
- Fundamento para modelos de aprendizaje automático y profundo.

INTRODUCCIÓN Y FUNDAMENTOS

Introducción al módulo

Módulo del tronco común

- Introducción a la minería de datos
- Aprendizaje de máquina
- Aprendizaje profundo

8 horas (teoría + prácticas)

Propósito del módulo:

Establecer las bases conceptuales y metodológicas de la minería de datos como componente fundamental de la inteligencia artificial aplicada.



Alcance de esta sesión

En esta **primera sección** se abordarán:

- Qué se entiende por minería de datos dentro del contexto de la IA.
- Cuál es su objetivo fundamental en entornos científicos e industriales.
- Cómo se estructura el proceso general de minería de datos.
- Cuáles son los conceptos básicos necesarios para el resto del módulo.
- El papel de la visualización exploratoria como primer contacto con los datos.

Esta sesión establece el **marco conceptual** común para todos los participantes.

Qué sí y qué no se cubre aquí

Qué sí se cubre en esta sección

- Definiciones formales y marco conceptual.
- Tipos de datos y tareas principales.
- Rol de la exploración visual.

Qué no se cubre aún**

- Algoritmos específicos.
- Implementación detallada en Python.
- Evaluación de modelos y métricas.

** Estos aspectos se desarrollarán progresivamente en los siguientes bloques del módulo y del diplomado.

1.1 Conceptos básicos

¿Qué es la minería de datos?

Pregunta fundamental:

¿Qué entendemos por **minería de datos**
en el contexto de la inteligencia artificial?

La **minería de datos** surge como respuesta a un problema central:

- La disponibilidad masiva de datos no implica automáticamente conocimiento.
- Es necesario un proceso sistemático para extraer información útil.



Definición formal

La **minería de datos** puede definirse como:

“El proceso de extraer patrones y conocimientos útiles
a partir de grandes volúmenes de datos”

—Aggarwal, Data Mining: The Textbook, 2015

Elementos clave de la definición:

- Proceso, no evento puntual.
- Extracción, no simple consulta.
- Patrones y conocimiento, no solo estadísticas.
- Grandes volúmenes de datos.

Proceso, no algoritmo

La **minería de datos** **no** se reduce a aplicar un modelo o algoritmo.

Incluye múltiples etapas:

Preparación

Análisis

Interpretación

En particular, Aggarwal enfatiza que:

- El preprocessamiento suele ser más crítico que el modelo mismo.
- Un mal tratamiento de los datos invalida cualquier resultado posterior.

Objetivo fundamental

Extraer **patrones significativos, relaciones relevantes y estructuras latentes** a partir de los **datos**, que permitan:

- Comprender fenómenos complejos.
- Apoyar la toma de decisiones.
- Generar hipótesis científicas.
- Automatizar procesos analíticos.

El **objetivo** no es “predecir por predecir”,
sino **comprender y modelar**.



Tipos de conocimiento que se buscan

La minería de datos puede revelar:

Patrones

- Regularidades frecuentes o recurrentes.

Relaciones

- Dependencias entre variables.

Estructuras

- Agrupamientos o jerarquías no evidentes.

Anomalías

- Comportamientos atípicos o raros.

Estos resultados no siempre son evidentes a simple inspección.

MD dentro del ecosistema de IA

La minería de datos:

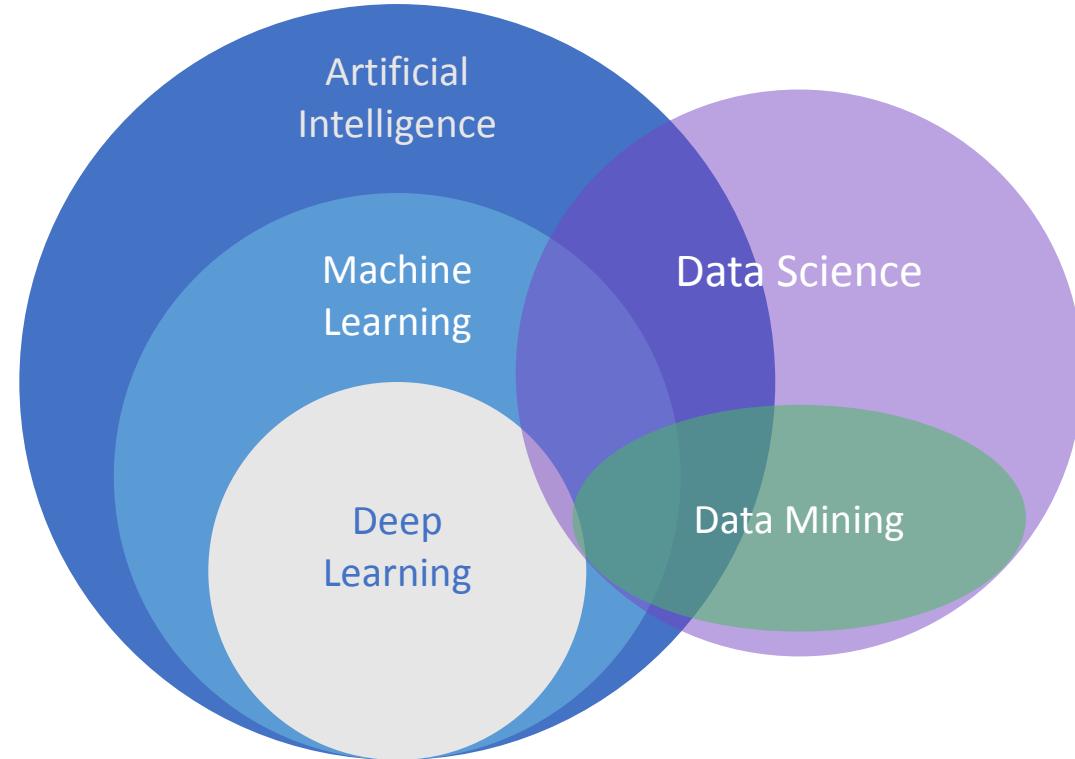
- Se apoya en estadística.
- Utiliza algoritmos de machine learning.
- Alimenta modelos de deep learning.

Pero se distingue porque:

- Pone énfasis en el proceso completo.
- Integra análisis exploratorio, modelado e interpretación.

En este diplomado, la **minería de datos** funciona como el puente entre los datos crudos y los modelos avanzados de IA.





Ciclo de vida

De acuerdo con **Aggarwal**, la **minería de datos** debe entenderse como un proceso compuesto por fases interdependientes, no como una secuencia lineal rígida.

Características clave:

- Iterativo.
- Dependiente del dominio.
- Guiado por los datos y los resultados intermedios.



Ciclo de vida



Fase 1: Recolección de datos

En esta fase se determina:

- Qué datos se utilizarán.
- De dónde provienen.
- En qué formato están disponibles.

Aspectos críticos:

- Heterogeneidad de fuentes.
- Calidad inicial de los datos.
- Sesgos de adquisición.

Nota: Errores en esta fase se propagan a todo el proceso.

Fase 2: Preprocesamiento

La fase más crítica del proceso!!

Aggarwal enfatiza que el **preprocesamiento** suele consumir:

- La mayor parte del tiempo.
- El mayor esfuerzo conceptual.

Incluye:

- Limpieza de datos.
- Manejo de valores faltantes.
- Normalización y escalamiento.
- Transformación y selección de atributos.

Un modelo sofisticado no compensa datos mal preparados.



Fase 3: Procesamiento analítico

En esta fase se aplican las **tareas** fundamentales, tales como:

- Clasificación.
- Regresión.
- Agrupamiento.
- Detección de anomalías.

Puntos clave:

- La elección del método depende del problema.
- No existe un algoritmo universalmente óptimo.
- El resultado debe interpretarse en contexto.



Retroalimentación y refinamiento

Los resultados obtenidos:

- Se evalúan.
- Se interpretan.
- Se utilizan para refinar etapas previas.

Esto puede implicar:

- Cambiar atributos.
- Ajustar el preprocesamiento.
- Replantear la tarea analítica.

La minería de datos es un proceso de refinamiento progresivo del conocimiento.



1.2 Conceptos básicos

Datos, instancias y atributos

En minería de datos, un **dataset** se compone de:

- **Instancias** (registros, ejemplos, observaciones)
- **Atributos** (variables, características, features)

Formalmente:

- Cada instancia se representa como un **vector de atributos**.
- El conjunto de datos puede verse como una tabla multidimensional.

La correcta interpretación de los atributos es clave para todo el proceso analítico.

¿Qué es un atributo?

Un **atributo** es una propiedad medible o descriptiva de una instancia, que puede representar:

- Una medición física.
- Una categoría.
- Un conteo.
- Un descriptor simbólico.

Ejemplos:

- Temperatura, presión, composición.
- Tipo de material.
- Fase cristalina.
- Presencia / ausencia de una característica.

Atributos nominales

Atributos categóricos sin orden

Características

- Toman valores de un conjunto finito.
No existe un orden intrínseco entre los valores.

Ejemplos

- Tipo de material: {metal, cerámico, polímero}
Método de síntesis: {sol-gel, CVD, sputtering}

Implicaciones

- No se pueden aplicar operaciones aritméticas.
Requieren representaciones especiales para modelado.



Atributos ordinales

Atributos categóricos con orden

Características

- Existe un orden natural entre los valores.
Las diferencias entre niveles no son necesariamente cuantificables.

Ejemplos

- Nivel de pureza: {bajo < medio < alto}
Clasificación cualitativa de calidad.

Implicaciones

- El orden importa.
La distancia entre categorías no es uniforme.



Atributos discretos

Variables cuantitativas contables

Características

- Toman valores enteros.
Representan conteos.

Ejemplos

- Número de defectos.
Número de capas.
Número de iteraciones experimentales.

Implicaciones

- Se pueden usar operaciones aritméticas.
Suelen modelarse como variables numéricas.



Atributos continuos

Variables cuantitativas reales

Características

- Pueden tomar cualquier valor dentro de un intervalo.
Alta resolución y variabilidad.

Ejemplos

- Temperatura (K).
Energía (eV).
Tiempo (s).

Implicaciones

- Requieren normalización o escalamiento.
Sensibles a ruido y outliers.



Datos mixtos

Datasets reales = combinación de tipos

Un mismo **dataset** suele contener:

- Atributos nominales.
- Ordinales.
- Discretos.
- Continuos.

Consecuencias:

- La elección del algoritmo depende del tipo de atributos.
- Las métricas de similitud cambian.
- El preprocessamiento es obligatorio.

No existe una representación universal óptima para todos los tipos de datos.



Importancia del tipo de atributo

El **tipo** de atributo determina:

- Qué transformaciones son válidas.
- Qué métricas pueden usarse.
- Qué algoritmos son apropiados.
- Cómo se visualizan los datos.

Errores comunes:

- Tratar atributos nominales como numéricos.
- Ignorar escalas distintas.
- Mezclar tipos sin preprocessamiento.

Muchos errores en **minería de datos** no son algorítmicos, sino conceptuales.



¿Qué se hace con los datos?

En **minería de datos**, las **tareas** analíticas fundamentales se agrupan en tres grandes categorías:

- Regresión
- Clasificación
- Agrupamiento

Estas tareas:

- Definen el tipo de problema.
- Determinan el tipo de salida.
- Condicionan los métodos y métricas que se utilizan.



Regresión

○ Predicción de valores numéricos

Se utiliza cuando

- La variable objetivo es continua.
Se desea estimar un valor real a partir de los atributos.

Ejemplos

- Predicción de propiedades físicas.
Estimación de rendimiento, energía o tiempo.
Modelado de relaciones funcionales entre variables.

Características

- Aprendizaje supervisado.
Evaluación mediante error.



Clasificación

○ Asignación de etiquetas discretas

Se utiliza cuando

- La variable objetivo es categórica.
Cada instancia pertenece a una o más clases.

Ejemplos

- Clasificación de materiales.
Identificación de fases.
Diagnóstico, detección o categorización.

Características

- Aprendizaje supervisado.
Evaluación basada en aciertos y errores.



Agrupamiento

○ Descubrimiento de estructura sin etiquetas

Se utiliza cuando

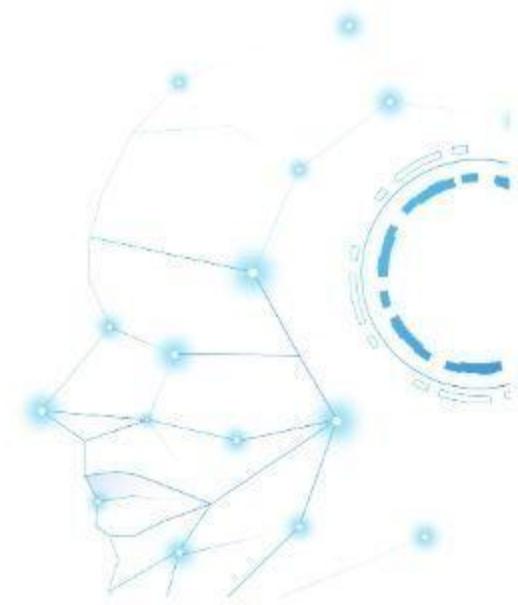
- No existen etiquetas conocidas.
Se desea descubrir estructura interna en los datos.

Ejemplos

- Segmentación de conjuntos experimentales.
Identificación de tipos de comportamiento.
Exploración inicial de datos desconocidos.

Características

- Aprendizaje no supervisado.
Interpretación dependiente del dominio.



Relación entre las tareas

El agrupamiento o *clustering* puede usarse como:

- Exploración previa a la clasificación.
- Generación de hipótesis.

La regresión puede verse como:

- Generalización de la clasificación binaria.
- Modelado continuo de fenómenos discretizados.

La clasificación puede apoyarse en:

- Resultados de clustering.
- Reducción de dimensionalidad.

En la práctica, los flujos de análisis combinan múltiples tareas.



Ejemplos conceptuales

Formulación correcta del problema

Un mismo **conjunto de datos** puede dar lugar a distintas tareas:

- Regresión
 - “¿Cuál es el valor esperado de esta propiedad?”
- Clasificación
 - “¿A qué categoría pertenece esta instancia?”
- Agrupamiento
 - “¿Existen patrones naturales en los datos?”



Clasificación: más de un solo caso

Aunque suele hablarse de “clasiﬁcación” como un problema único, en realidad existen distintas formulaciones, dependiendo de:

- El número de clases.
- La relación entre clases e instancias.
- La estructura de las etiquetas.

Estas diferencias no son triviales:

- Cambian los modelos.
- Cambian las métricas.
- Cambia la interpretación de resultados.



Clasificación binaria

- Dos clases posibles

Definición

- Cada instancia pertenece a **una** de dos clases.

Ejemplos

- Acceptable / no aceptable.
Presencia / ausencia de una propiedad.
Normal / anómalo.

Características

- Es el caso más simple.
Muchas métricas se definen originalmente para este escenario.
Sirve como base para casos más complejos.

Clasificación multiclase

- Más de dos clases, una sola por instancia

Definición

- Cada instancia pertenece a **una y solo una** clase, de entre varias posibles.

Ejemplos

- Tipo de material (A, B, C, ...).
Fase cristalina.
Categorías experimentales mutuamente excluyentes.

Características

- Generaliza la clasificación binaria.
Requiere estrategias específicas en algunos modelos.
La confusión entre clases es más rica y compleja.

Clasificación multiclase

- Múltiples etiquetas por instancia

Definición

- Una instancia puede pertenecer simultáneamente a **varias** clases.

Ejemplos

- Un material con múltiples propiedades funcionales.
Un sistema con varios mecanismos activos.
Etiquetado no excluyente.

Características

- Las clases no son mutuamente excluyentes.
La salida es un conjunto de etiquetas, no una sola.
Las métricas tradicionales deben adaptarse.

Clasificación multi-instancias

○ Instancias compuestas

Definición

- Cada ejemplo está formado por un conjunto de instancias internas.
La etiqueta se asigna al conjunto completo.

Ejemplos

- Un experimento compuesto por múltiples mediciones.
Un objeto descrito por múltiples regiones o muestras.

Características

- La relación etiqueta–dato es indirecta.
Requiere modelos especializados.

Errores conceptuales frecuentes

Problemas comunes:

- Tratar multi-etiqueta como multiclasificación.
- Forzar problemas complejos a esquemas binarios.
- Evaluar con métricas incorrectas.

Consecuencias:

- Resultados engañosos.
- Interpretaciones erróneas.
- Comparaciones inválidas entre modelos.
- Antes de elegir un algoritmo, hay que formular correctamente el problema.



1.3 Visualización de datos

¿Por qué visualizar antes de modelar?

La visualización es una herramienta analítica, no decorativa, y permite:

- Detectar errores de captura.
- Identificar valores atípicos.
- Evaluar distribuciones.
- Intuir relaciones entre variables.

Aggarwal subraya que muchas fallas de modelado:

- Se detectan visualmente antes de aparecer en métricas.



De datos a gráficos

Entonces, la **visualización** de datos consiste en:

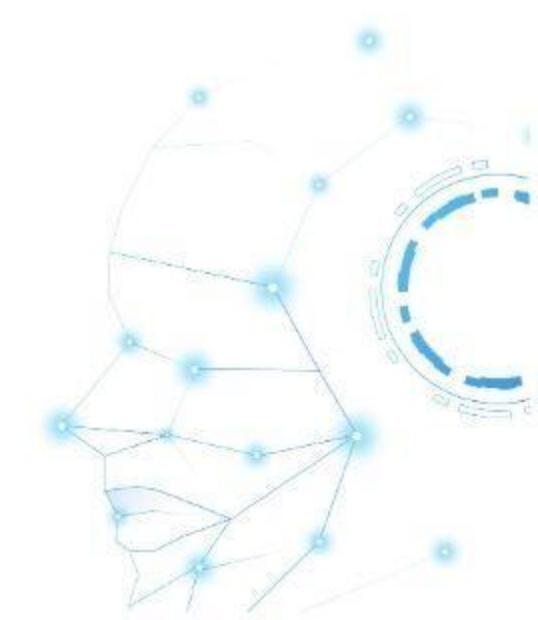
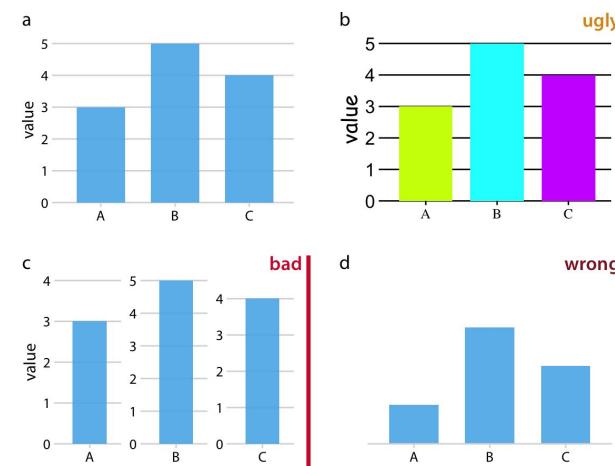
Mapear valores de datos a atributos visuales perceptibles

No es:

- Dibujar gráficos “bonitos”.
- Decorar resultados.

Es:

- Traducir información numérica o categórica a señales visuales que el sistema perceptual humano puede interpretar.

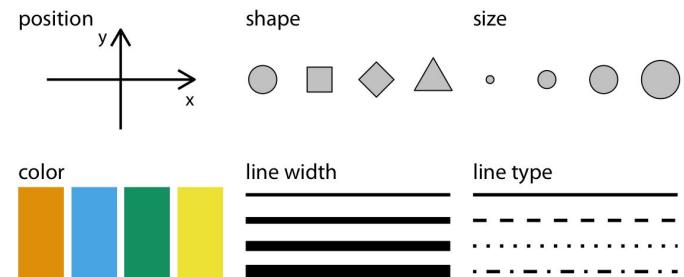


Atributos visuales

¿Qué se puede controlar visualmente?

Wilke identifica atributos visuales fundamentales:

- Posición (en ejes x, y)
- Longitud
- Área
- Color
- Forma



Principio clave:

- No todos los atributos son igual de efectivos.
- La posición es el canal más preciso para comparación cuantitativa.
- Elegir mal el atributo visual distorsiona la interpretación.



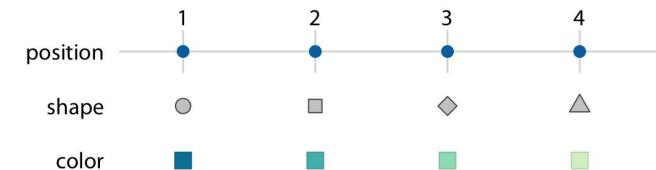
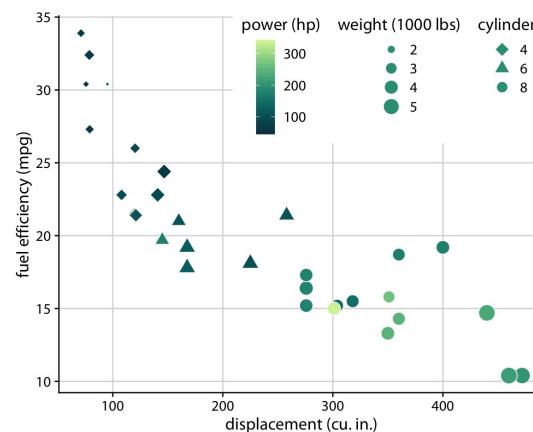
Escalas: del dato a la percepción

Una escala define cómo un valor numérico o categórico se transforma en:

- Posición.
- Tamaño.
- Intensidad de color.

Tipos de escalas :

- Lineales.
- Categóricas.
- Secuenciales (color).
- Divergentes (color).



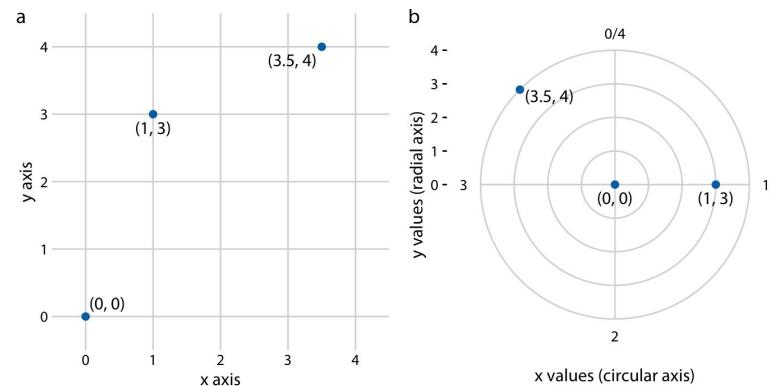
La escala no es neutra: influye directamente en lo que se percibe.

Sistemas de coordenadas y color

Decisiones básicas pero críticas

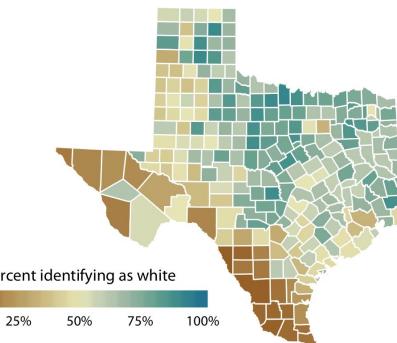
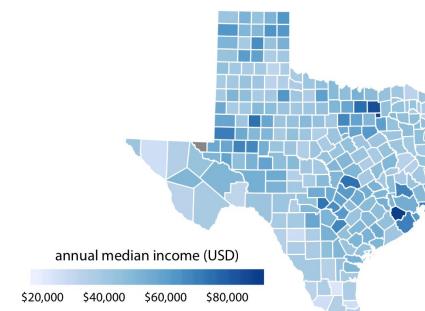
Coordenadas

- El sistema cartesiano es el estándar para exploración.
- Facilita comparación y detección de patrones.



El color puede usarse para:

- Distinguir categorías.
- Representar valores.
- Resaltar elementos.



El mal uso del **color** introduce sesgos perceptuales.



¿Qué son las cantidades?

Las **cantidades** representan:

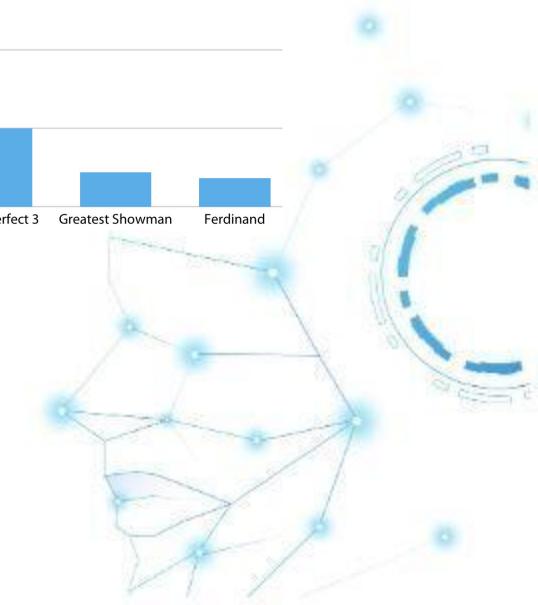
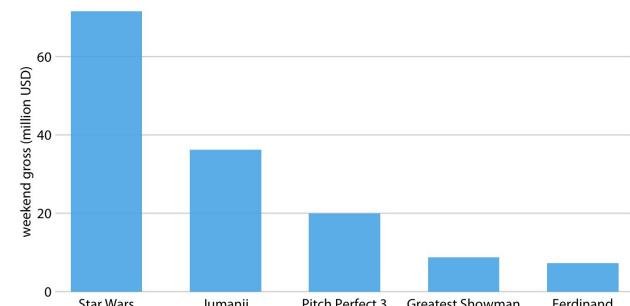
- Magnitudes absolutas.
- Conteos.
- Intensidades agregadas.

Preguntas típicas:

- ¿Cuánto hay de cada cosa?
- ¿Qué categoría tiene mayor o menor magnitud?
- ¿Cómo se comparan varios valores?

La comparación directa es el objetivo principal.

| Rank | Title | Weekend gross |
|------|--------------------------------|---------------|
| 1 | Star Wars: The Last Jedi | \$71,565,498 |
| 2 | Jumanji: Welcome to the Jungle | \$36,169,328 |
| 3 | Pitch Perfect 3 | \$19,928,525 |
| 4 | The Greatest Showman | \$8,805,843 |
| 5 | Ferdinand | \$7,316,746 |



Barras: el estándar

Uso

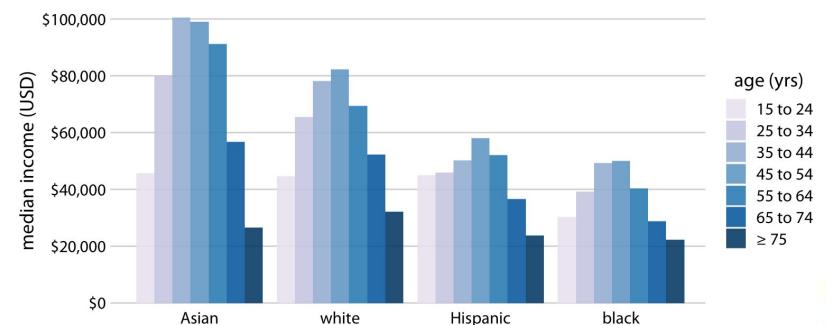
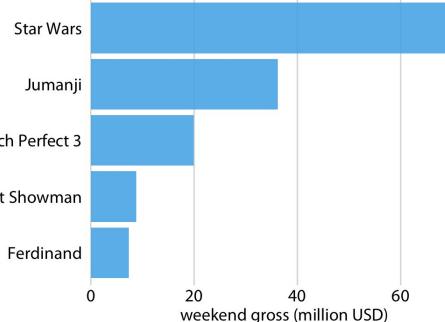
- Comparar magnitudes entre categorías.

Fortalezas

- Claridad inmediata.
Fácil comparación cuando el eje comienza en cero.

Advertencias

- El eje debe empezar en cero para evitar distorsión.
Barras apiladas dificultan comparaciones precisas.
El exceso de categorías reduce legibilidad.



Gráficos de puntos

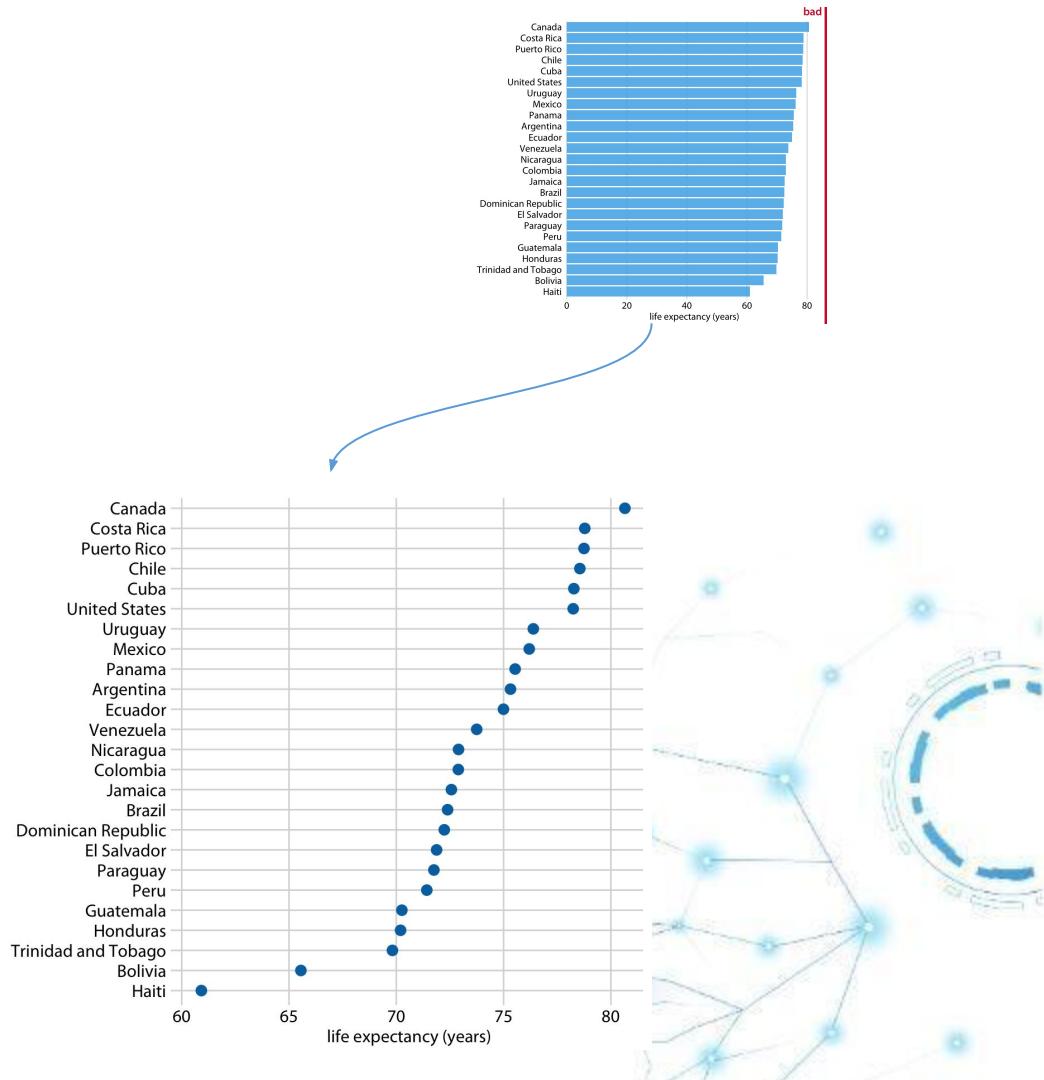
Ventajas frente a barras:

- Menos “tinta” innecesaria.
Comparación más precisa de valores cercanos.
Mejor escalabilidad con muchas categorías.

Fortalezas

- Exploración rápida.
Comparaciones finas entre cantidades.

Nota: Preferir gráficos de puntos cuando la **precisión comparativa** es prioritaria.



Mapas de calor

Puede representar

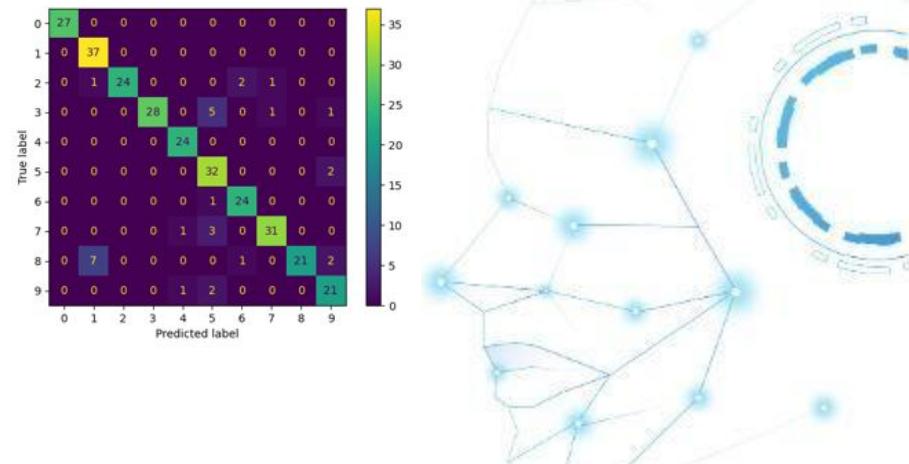
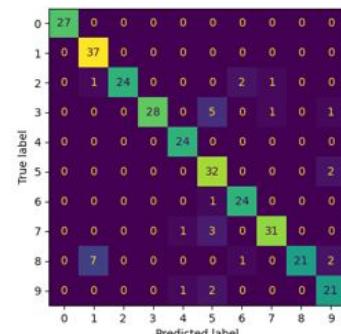
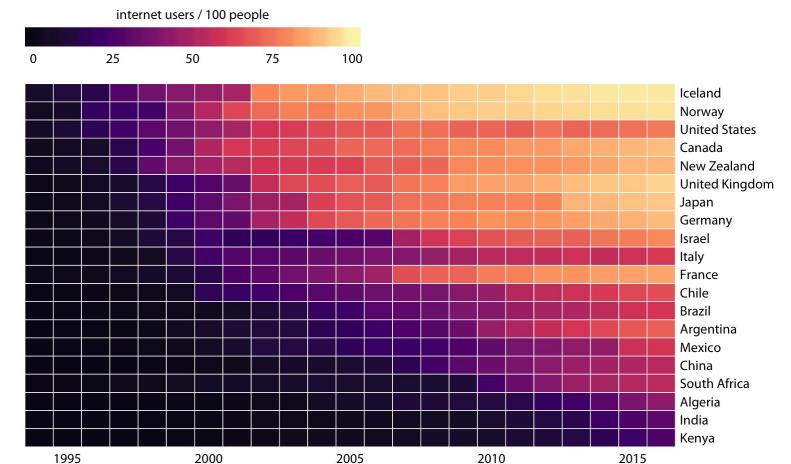
- Frecuencias.
- Conteos bivariados.
- Intensidad promedio por combinación de variables.

Uso

- Detectar concentraciones.
- Identificar patrones globales.
- Resumir grandes tablas.

Advertencia

- El significado depende críticamente de la escala de color.



¿Por qué visualizar distribuciones?

Una **distribución** describe cómo se reparten los valores de una variable.

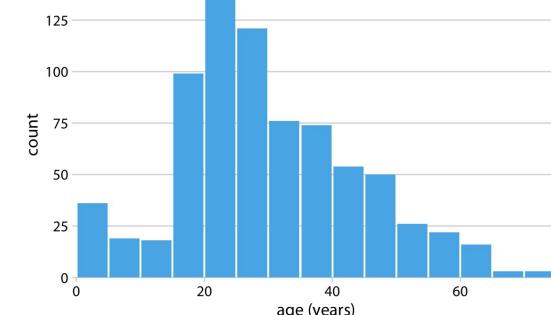
Visualizar distribuciones permite:

- Entender forma y dispersión.
- Detectar asimetría y colas largas.
- Identificar multimodalidad.
- Localizar valores atípicos.

Pregunta central:

¿Cómo están distribuidos los datos,
no solo cuál es su promedio?

| Age range | Count | Age range | Count | Age range | Count |
|-----------|-------|-----------|-------|-----------|-------|
| 0-5 | 36 | 31-35 | 76 | 61-65 | 16 |
| 6-10 | 19 | 36-40 | 74 | 66-70 | 3 |
| 11-15 | 18 | 41-45 | 54 | 71-75 | 3 |
| 16-20 | 99 | 46-50 | 50 | | |
| 21-25 | 139 | 51-55 | 26 | | |
| 26-30 | 121 | 56-60 | 22 | | |



Histogramas

Función

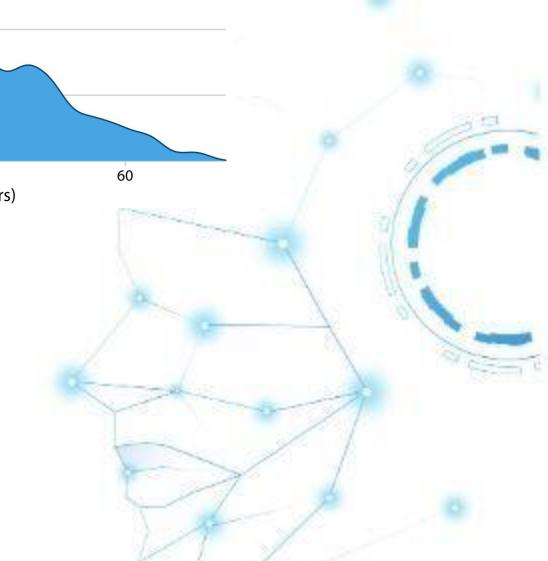
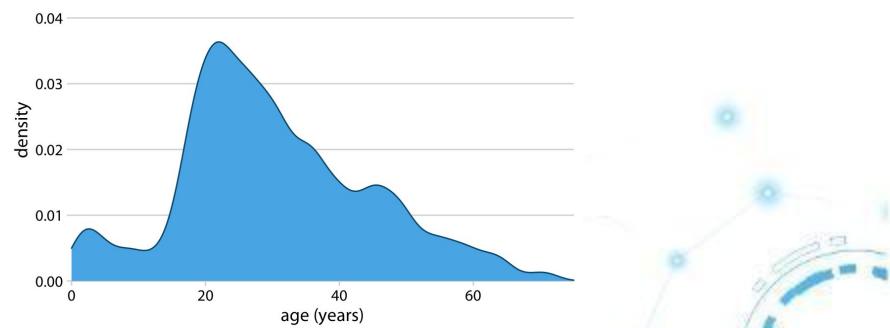
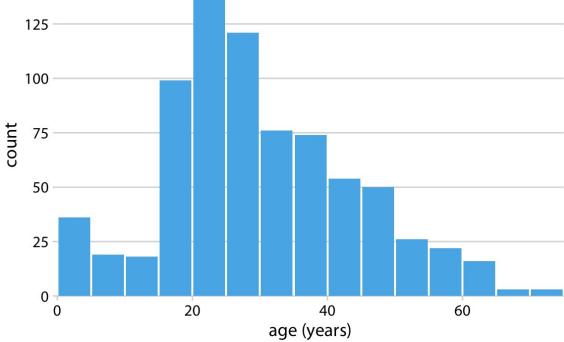
- Agrupan valores en intervalos (bins).
Revelan forma global de la distribución.

Curvas de densidad

- Representación suavizada.
Facilitan comparación entre distribuciones.

Advertencias

- El número de bins afecta la interpretación.
Densidades pueden ocultar detalles locales.



Comparación de múltiples distribuciones

¿Cómo comparar grupos?

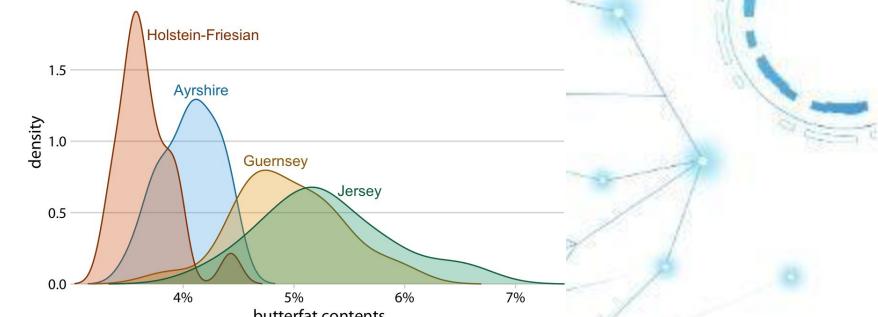
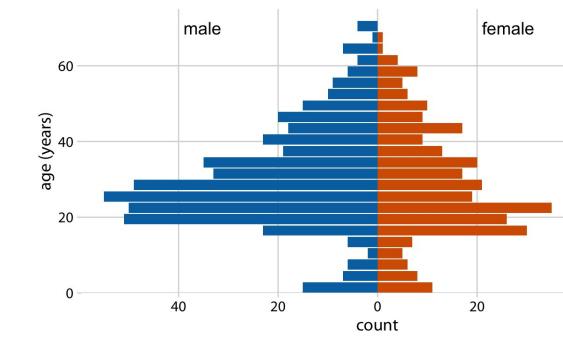
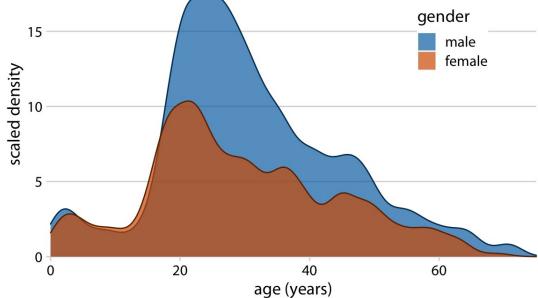
- Histogramas superpuestos (con cautela).
- Densidades múltiples.
- Distribuciones alineadas a un eje común.

Objetivo:

- Comparar forma, no solo centro.
- Detectar diferencias sistemáticas entre grupos.

Problema común:

Saturación visual cuando hay muchos grupos.



Diagramas de cajas y bigotes

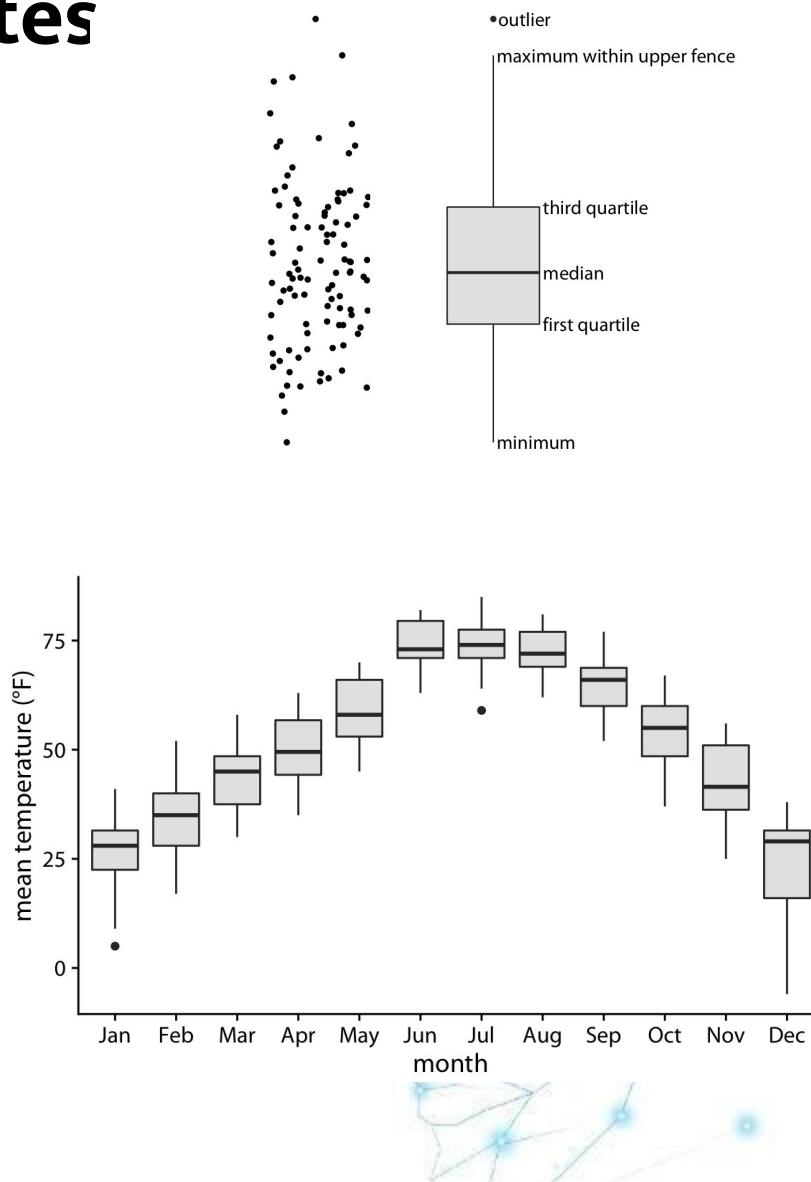
Elementos

- Mediana.
- Cuartiles.
- Rango intercuartílico.
- Outliers.

Ventajas:

- Comparación directa entre muchos grupos.
- Resistencia a valores extremos.

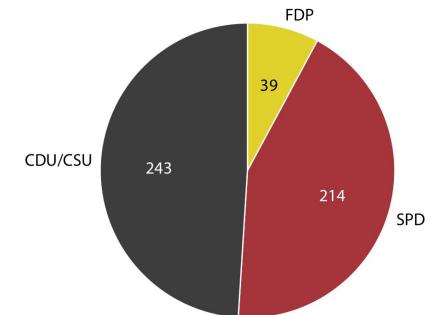
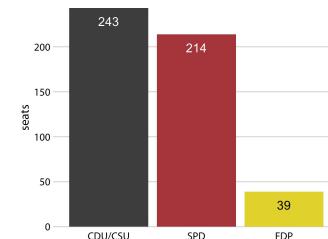
Ideales para exploración comparativa, no para detalle fino.



¿Qué son las proporciones?

Las **proporciones** describen:

- Cómo se reparte un total entre categorías.
- Cómo cambia la composición entre grupos.



Pregunta central:

¿Qué fracción del todo corresponde a cada categoría?

A diferencia de cantidades:

El énfasis está en el **reparto relativo**, no en la magnitud absoluta.

Comparación de proporciones entre grupos

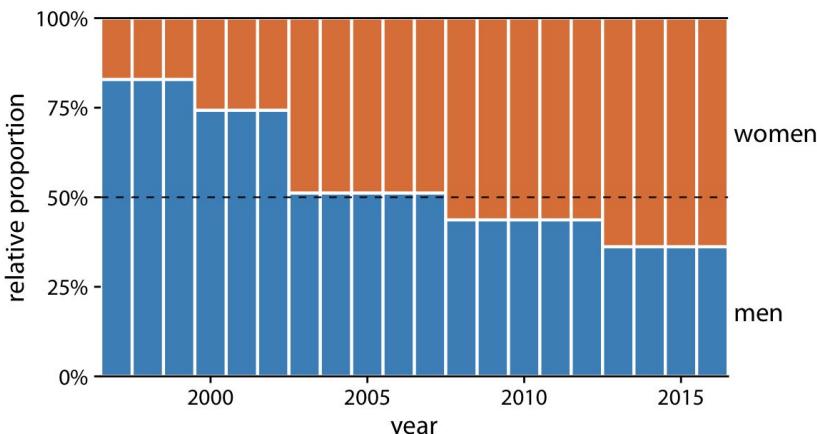
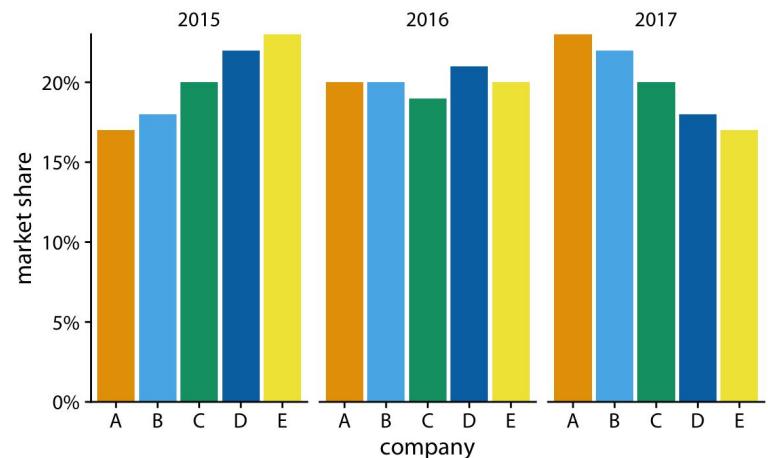
Barras lado a lado

- Facilitan comparación directa entre categorías.
Recomendadas para exploración.

Barras apiladas

- Útiles para ver composición global.
Dificultan comparar segmentos internos.

Para explorar diferencias, cuando **comparar** es más importante que mostrar el total.



Proporciones anidadas

Cuando las proporciones tienen niveles:

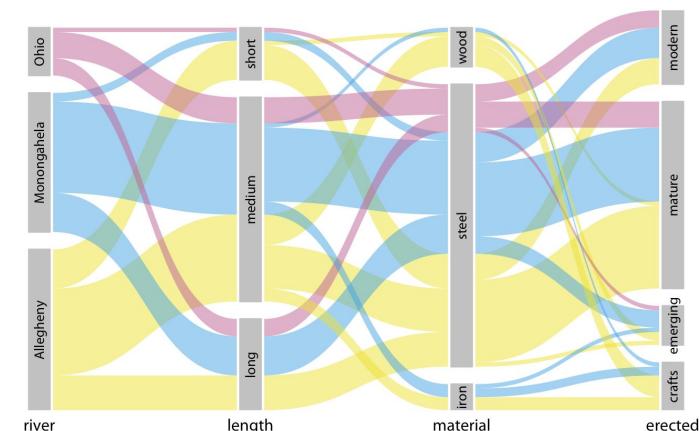
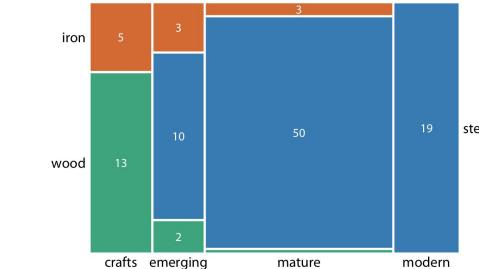
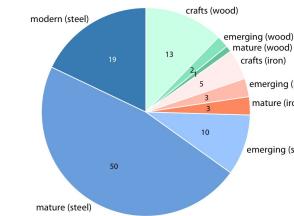
- Categorías dentro de categorías.
Subgrupos dentro de grupos.

Visualizaciones:

- Nested pies and mosaic plots.
Treemaps.
Parallel sets.

Uso

- Detectar cambios de composición.
Identificar desbalanceos estructurales.



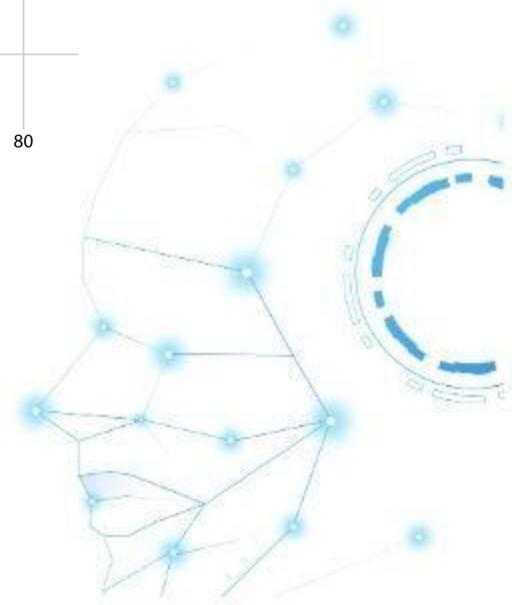
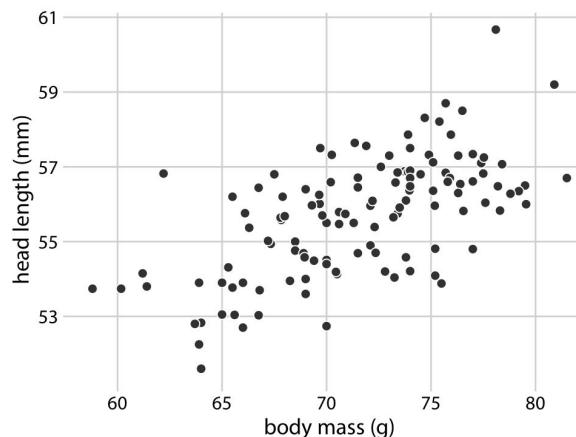
¿Qué es una asociación?

Una **asociación** describe cómo dos o más variables cuantitativas varían conjuntamente.

Preguntas típicas:

- ¿Existe relación entre dos variables?
- ¿La relación es lineal o no lineal?
- ¿Hay grupos o estructuras latentes?
- ¿Existen variables redundantes?

Visualizar asociaciones es clave antes de modelar.



Diagramas de dispersión

Uso

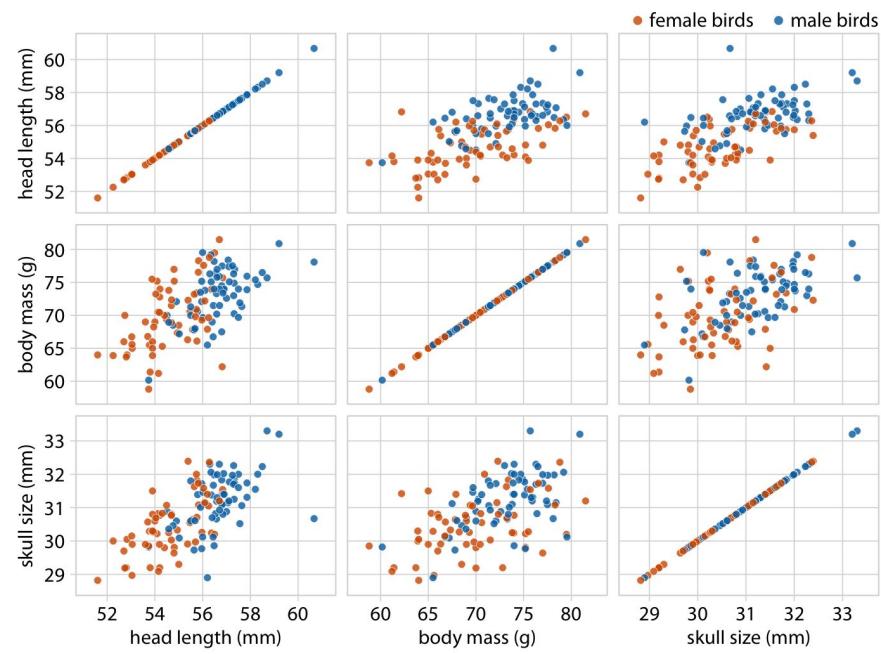
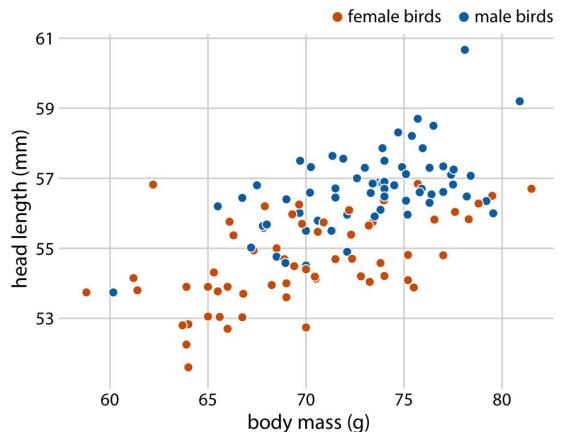
- Relación bivariada entre variables cuantitativas.

Permiten observar

- Tendencias (positiva, negativa, nula).
No linealidades.
Agrupamientos.
Outliers estructurales.

Advertencias

- La sobreposición de puntos puede ocultar patrones.
Transparencia o muestreo ayudan en datasets grandes.



Correlogramas y relaciones múltiples

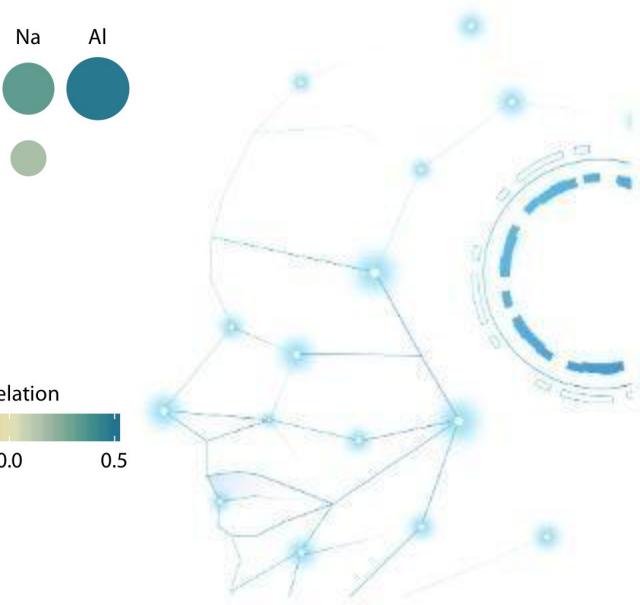
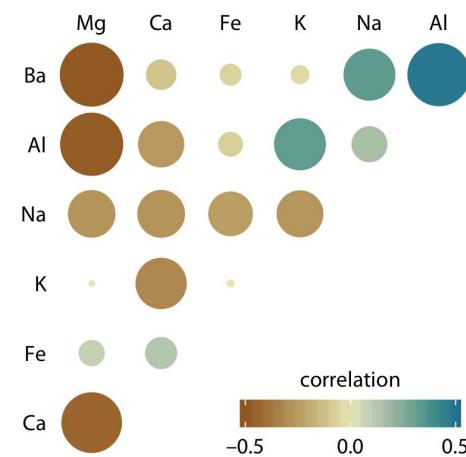
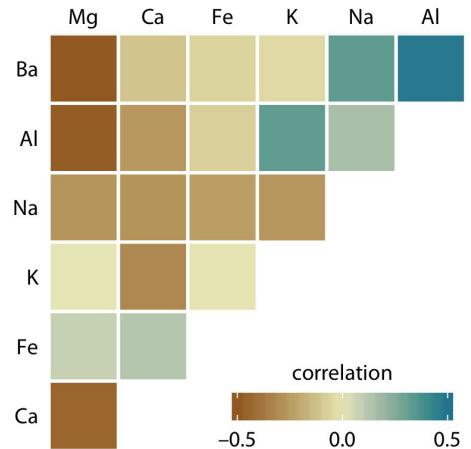
Uso

- Resumen global de asociaciones.
Identificación rápida de redundancia.

Permiten observar

- Selección de atributos.
Detección de multicolinealidad.
Guía para reducción de dimensionalidad.

Nota: Correlación visual \neq causalidad.



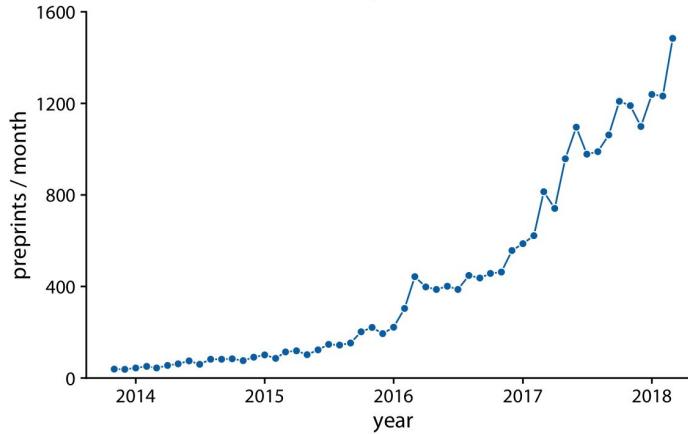
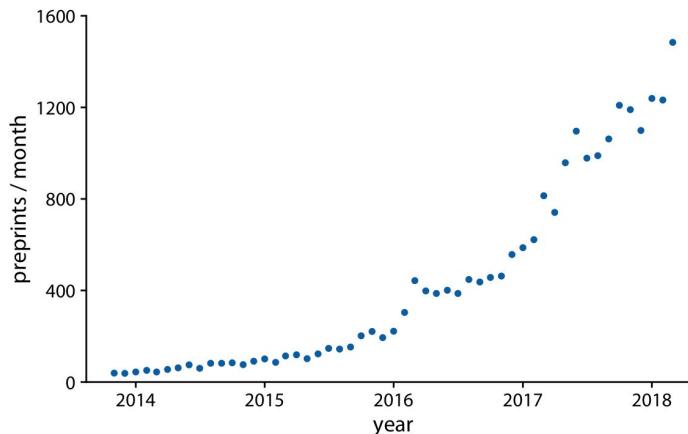
Variables ordenadas

Una variable ordenada es aquella cuyos valores:

- Poseen un orden natural.
- Representan una progresión lógica o experimental.
- Definen un eje independiente sobre el cual se observa la respuesta del sistema.

Ejemplos típicos:

- Tiempo (t).
- Dosis o concentración.
- Número de iteración o ciclo.
- Distancia o posición secuencial.



Series individuales y múltiples

Series individuales

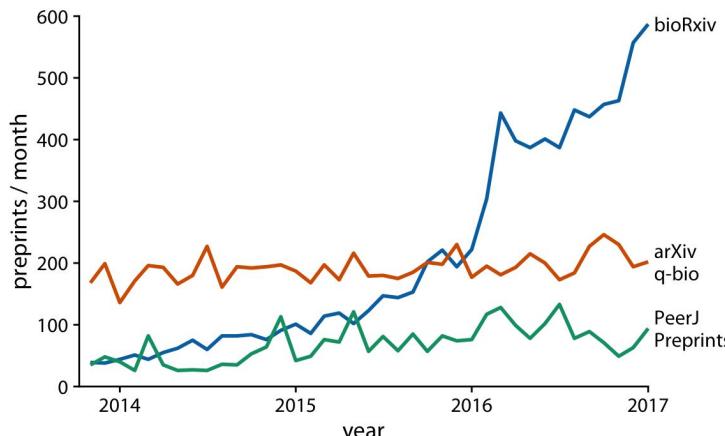
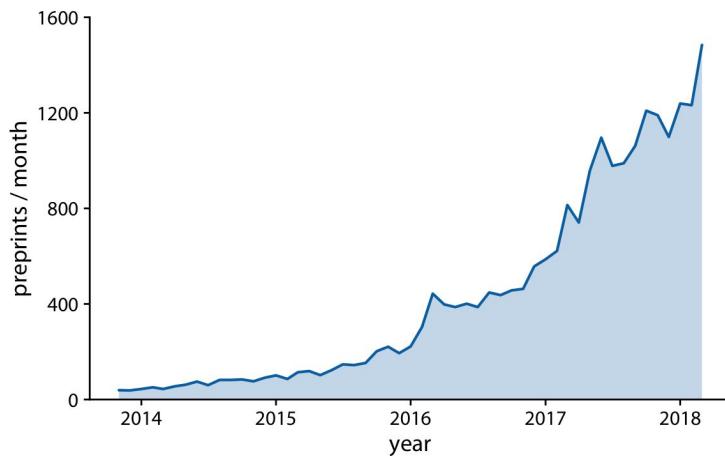
- Identifican tendencias.
Detectan rupturas.

Series múltiples

- Comparan comportamientos entre grupos.
Revelan patrones comunes o divergentes.

Uso exploratorio

- No modelar aún.
Detectar estructura y ruido.



Visualización en Python

Como ya vimos, en **minería de datos**, la **visualización** cumple un rol analítico, no decorativo.

En **Python**, este rol se articula mediante un ecosistema de herramientas complementarias.

- Comprensión del conjunto de datos.
- Validación de supuestos.
- Detección temprana de problemas.
- No se aborda como “presentación de resultados”.

La herramienta correcta depende del tipo de pregunta, no del estilo gráfico.





Pandas proporciona la estructura conceptual del análisis:

- Datos tabulares como DataFrame.
- Atributos como variables explícitas.
- Observaciones como unidades analíticas.

Funciones relevantes para exploración

- **Inspección estructural:**
dimensiones, tipos de datos, valores únicos.
- **Estadística descriptiva:**
tendencia central, dispersión, rangos.
- **Agrupación y agregación:**
preparación natural para visualización.

matplotlib

Matplotlib es la capa gráfica de bajo nivel:

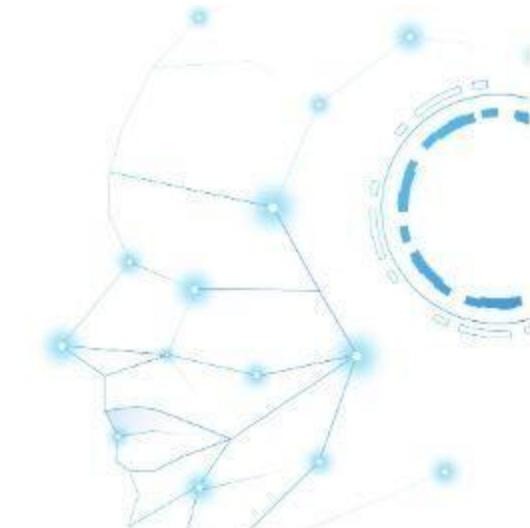
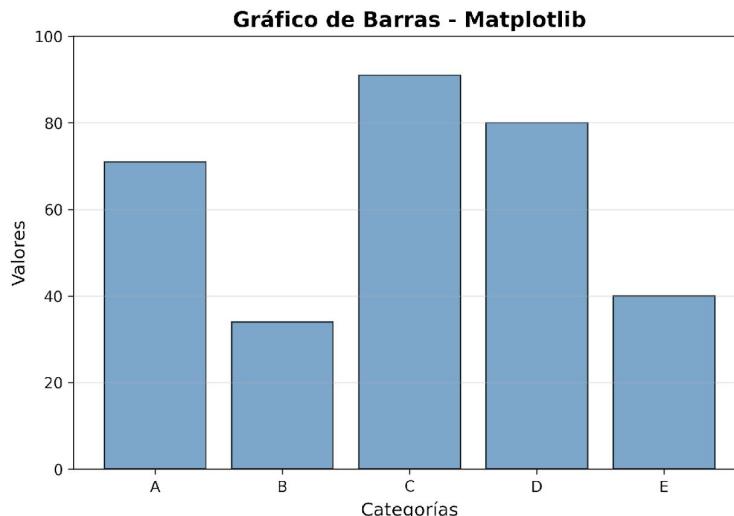
- Define el lenguaje visual base en **Python**.
- Control explícito de todos los elementos del gráfico.

Conceptos estructurales

- Figura vs. ejes.
- Escalas y transformaciones.
- Anotaciones, etiquetas y referencias visuales.

Implicación práctica

- Alta flexibilidad.
- Mayor responsabilidad del analista en decisiones visuales.
- Mensaje clave



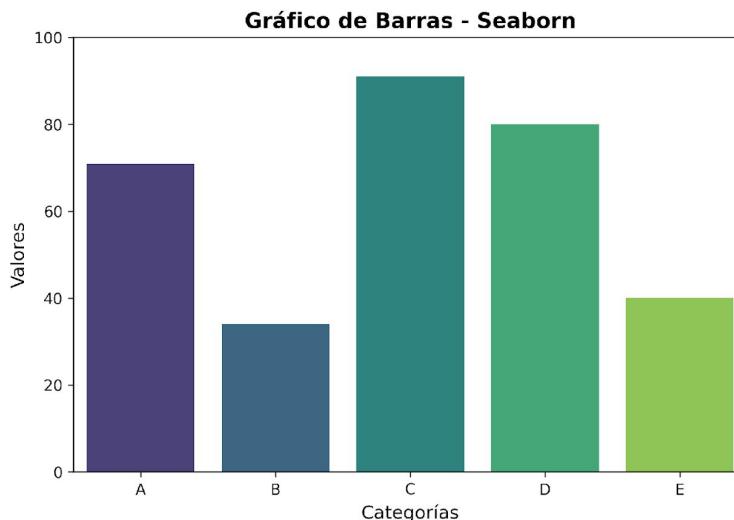


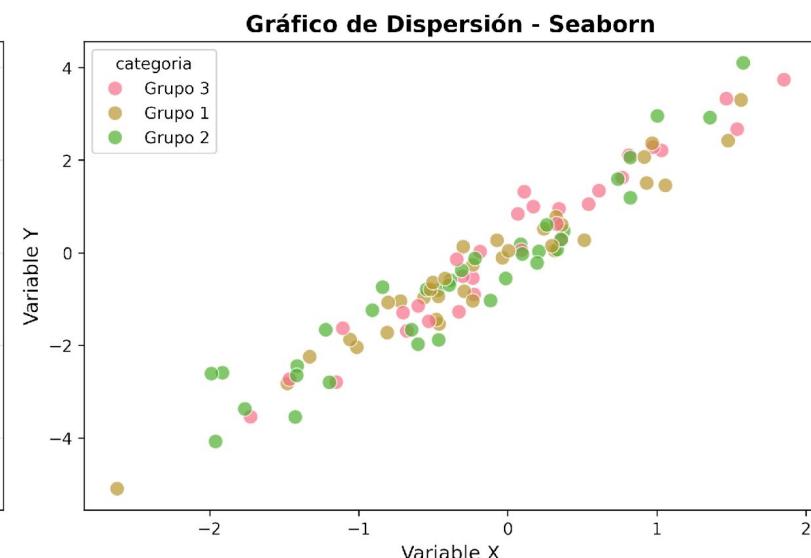
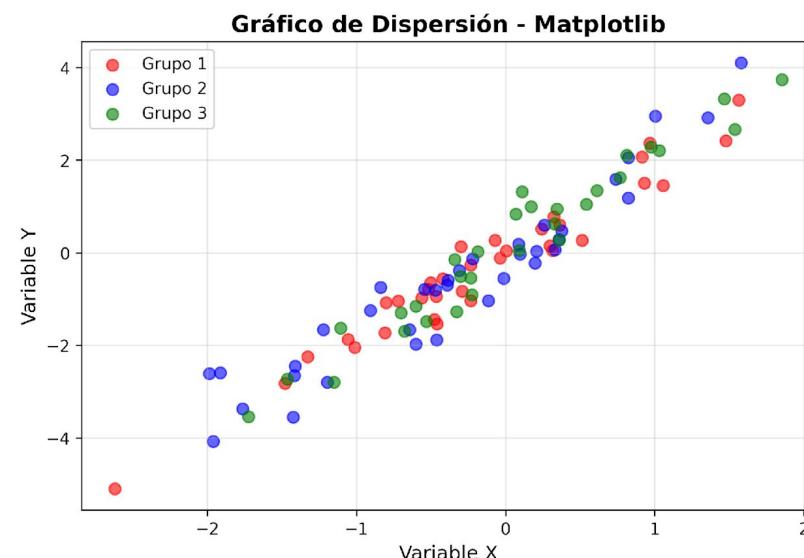
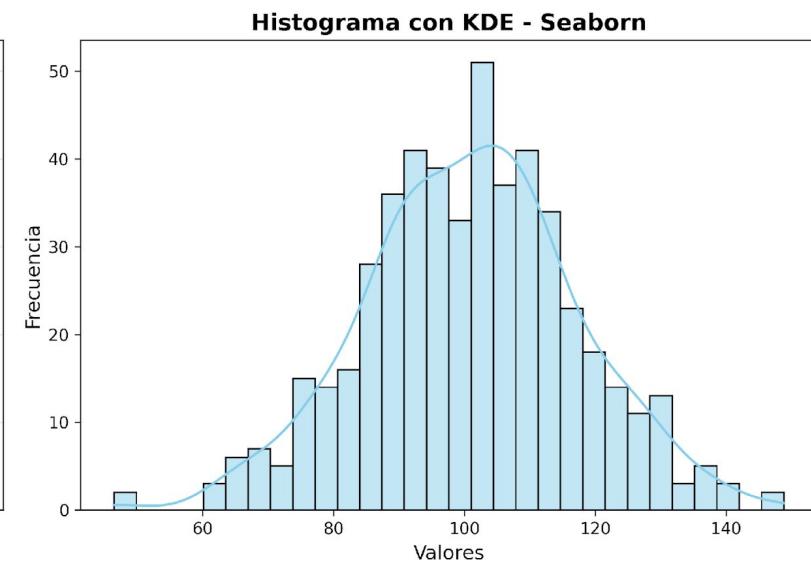
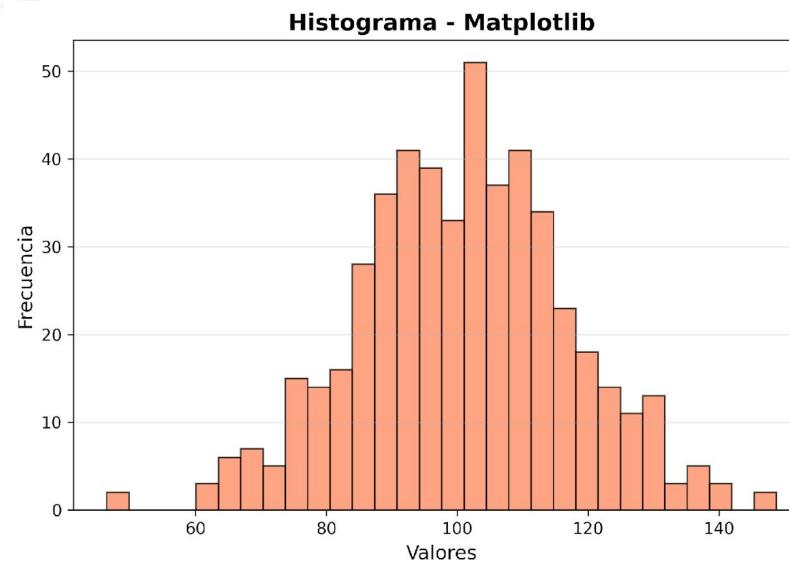
Seaborn introduce una capa semántica:

- Integra estadística y visualización.
- Automatiza elecciones gráficas comunes.

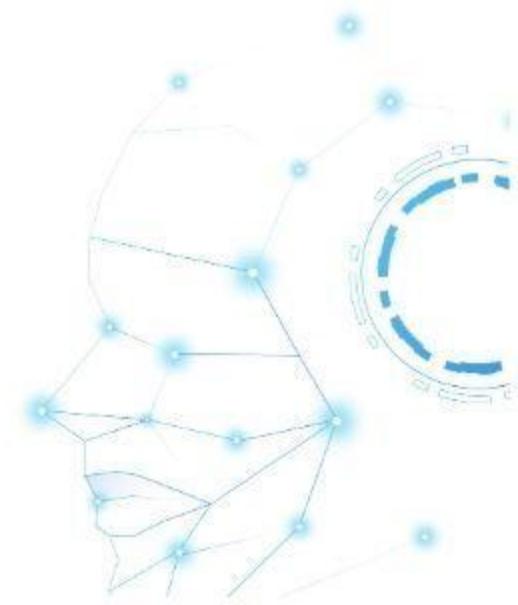
Ventajas para exploración

- **Gráficos alineados con:**
distribuciones,
comparaciones,
asociaciones.
- **Uso explícito del tipo de variable:**
categórica vs. numérica.

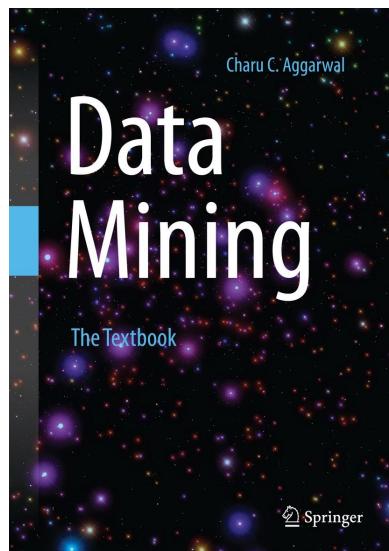




¿Qué se construyó en esta sección?



Bibliografía

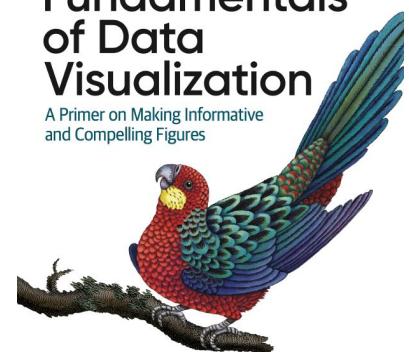


Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1, No. 3). New York: springer.

O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures



Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.

<https://clauswilke.com/dataviz/>