Clase 036 — Matplotlib: line, scatter, bar, histogram, boxplot

Parte: 0 — Prerrequisitos · Fuente: VanderPlas, cap. 4 §§ 4.2–4.5 Simple Line/Scatter/Bar/Histogram Plots. · Duración estimada: 75 min.

Clase 036 — Matplotlib: line, scatter, bar, histogram, boxplot

> *Parte: 0 — Prerrequisitos · Fuente: VanderPlas, cap. 4 §§ 4.2–4.5 Simple Line/Scatter/Bar/Histogram Plots. · Duración estimada: 75 min.*

Objetivo

Que el alumno conozca los 5 plots básicos que cubren el 80% del trabajo de EDA, y sepa cuándo cada uno: line (tendencia temporal), scatter (relación dos variables), bar (categóricas), histogram (distribución), boxplot (5 estadísticos + outliers).

Resultados de aprendizaje

Al finalizar la clase, el alumno podrá:

1. Elegir el plot correcto según el tipo de variables (continua/categórica) y el objetivo.
2. Ajustar marker, color, linestyle, alpha para legibilidad.
3. Construir histogramas con bins adecuados (regla de Freedman-Diaconis o 'auto').
4. Interpretar boxplot: mediana, Q1/Q3, whiskers, outliers.
5. Combinar bar + error bars para mostrar incertidumbre.

Temas

#	Tema	Por qué importa
1	Line: tendencias y series temporales	El más fácil de leer mal.
2	Scatter: relación entre dos variables	Con c= y s= para 3ª/4ª dimensión.
3	Bar y barh: categóricas	Vertical vs horizontal.
4	Histogram: distribución de una continua	Bins importan.
5	Boxplot: distribución resumida + outliers	Cuando hay muchos grupos.
6	Errorbar y fill_between	Mostrar incertidumbre.

Definiciones y características

Line plot

: Une puntos con líneas. Implica continuidad/orden en X — solo úsalo cuando X tiene orden natural (tiempo, espacio, secuencia).

Scatter

: Puntos no conectados. Muestra relación entre 2 continuas. Con c= codificas 3ª dim (color), con s= 4ª (tamaño). Más de 4 dims sobrecarga.

Bar / barh

: Barras para categóricas. Vertical (bar) si etiquetas son cortas, horizontal (barh) si son largas o muchas.

Histogram

: Distribución de UNA continua. Bins importan: pocos esconden estructura, muchos generan ruido. bins='auto'

usa Freedman-Diaconis (buen default).

Boxplot

: Resumen de distribución: mediana (línea), Q1-Q3 (caja), whiskers (1.5×IQR), outliers (puntos). Útil para comparar muchos grupos rápido.

Errorbar

: Barra + línea vertical/horizontal indicando incertidumbre (std, IC95%). Sin esto, las barras mienten visualmente.

Dataset / recursos

Palmer Penguins. Sin descarga adicional.

Ejercicios

1. Line. Serie temporal de ventas mensuales (sintética). Anota máximo con flecha.

2. Scatter. body_mass vs bill_length, color por species. Adicionalmente: s= con flipper_length para tamaño.

3. Bar. Count por species, ordenado descendente. Vertical y horizontal — compara legibilidad.

4. Histogram. Distribución de body_mass con bins='auto' y bins=10. Compara.

5. Boxplot. body_mass por species: 3 cajas lado a lado. Identifica outliers.

Homework verificable

Notebook con penguins: (a) 5 plots básicos cada uno bien etiquetado; (b) scatter decorado con color y tamaño codificando 3 dimensiones; (c) bar con errorbars de std; (d) boxplot agrupado con interpretación de outliers.

Criterio de aceptación: Cada plot tiene título, labels, leyenda donde aplica. Bins justificados.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Line plot entre 2 categorías "A", "B"	Conectar categorías con línea engaña (sugi
Histograma con escala Y rara	Por default density=False (counts). Si qui
Boxplot todos iguales por outliers extremo	Outliers dominan visualmente; cajas quedan
Bar chart con colores random distrae	Sin codificación significativa, color = ru
Scatter con miles de puntos = blob negro	Overplotting. Fix: alpha=0.3, hexbin (plt.

Preguntas frecuentes

¿Pie chart cuándo?

Casi nunca. El ojo humano compara mal ángulos. Para proporciones: bar o stacked bar. Pie tolerable solo con 2-3 categorías y proporciones muy distintas.

¿Cuántos bins en un histograma?

bins='auto' (Freedman-Diaconis) es buen default. Si es estudios académicos: regla de Sturges (bins=int(np.log2(n)+1)). Experimenta con 10/30/50 si dudas.

¿Boxplot o violinplot?

Boxplot: rápido, 5 estadísticos, outliers claros. Violinplot: muestra distribución completa (multimodalidad). Para comparar 3-10 grupos, ambos OK. Para >10, boxplot gana en densidad.

¿Errorbars con std o con IC?

Std: dispersión natural de los datos. IC95% de la media: incertidumbre del estimador (más pequeño con N grande). Para inferencia, IC. Para describir, std.

¿Plot 3D buen idea?

Casi nunca. Oclusión + perspectiva engañan. 2D con color/tamaño suele comunicar mejor. Excepción: superficies analíticas z = f(x, y).

Referencias

- VanderPlas, cap. 4 §§ 4.2-4.5.
- matplotlib gallery

Siguiente clase

Clase 037 — Matplotlib: subplots y gridspec

Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
rng = np.random.default_rng(42)

# Penguins-like sintético
df = pd.DataFrame({
    'species'    : np.repeat(['Adelie', 'Chinstrap', 'Gentoo'], [50, 30, 40]),
    'body_mass'  : np.concatenate([rng.normal(3700, 400, 50), rng.normal(3700, 400, 30), rng.normal(5050, 500, 40)]),
    'bill_length': np.concatenate([rng.normal(39, 2, 50),     rng.normal(48, 3, 30),     rng.normal(48, 3, 40)]),
    'flipper'    : np.concatenate([rng.normal(190, 6, 50),    rng.normal(196, 7, 30),    rng.normal(217, 7, 40)]),
})
```

Archivos complementarios

- notebook.ipynb