Proyecto final: ¿qué explica el monto de la propina?

Durante el curso recorriste el ciclo completo de ciencia de datos con Gapminder: pregunta, datos, limpieza, exploración, visualización, modelo y comunicación. En este proyecto lo repites tú, sin guía, con otro dataset real. Se usa Python 3.11 o superior con pandas, matplotlib, seaborn, scikit-learn y plotly. El notebook `proyecto-propinas` contiene una solución de referencia completa: consúltalo al terminar, no antes.

Contexto

Un mesero registró cada propina que recibió durante unos meses de trabajo en un restaurante de EE. UU.: 244 mesas en total. Los datos se publicaron en una colección de casos de estadística para negocios (Bryant y Smith, 1995, *Practical Data Analysis: Case Studies in Business Statistics*) y hoy vienen incluidos en varias bibliotecas. Supón que el restaurante te pide entender de qué depende la propina.

Pregunta

¿Qué explica el monto de la propina? En concreto: ¿cuánto pesa el monto de la cuenta?, ¿influyen el día, el horario, el tamaño de la mesa, el sexo de quien paga o si hay fumadores?, y ¿qué tan bien se puede predecir la propina de una mesa nueva?

Datos

```
import plotly.express as px

tips = px.data.tips()
print(tips.shape)   # (244, 7)
```

Columna	Descripción
`total_bill`	Cuenta total, en dólares
`tip`	Propina, en dólares
`sex`	Sexo de quien paga: Female / Male
`smoker`	Si había fumadores en la mesa: Yes / No
`day`	Día: Thur, Fri, Sat, Sun
`time`	Horario: Lunch / Dinner
`size`	Personas en la mesa (1 a 6)

Requisitos mínimos

1. **Calidad de datos.** Forma, tipos, nulos y duplicados. Decide qué hacer con el duplicado y justifícalo.
2. **Variable derivada.** Crea `pct`, la propina como porcentaje de la cuenta, y explica por qué es útil.
3. **Exploración por grupo.** Tablas con `n`, propina media y porcentaje medio por `day`, `time`, `sex`, `smoker` y `size`.
4. **Visualización.** Al menos dos gráficas con título y unidades: propina contra cuenta y porcentaje por tamaño de mesa.
5. **Modelos.** Con `train_test_split(test_size=0.25, random_state=42)` y one-hot con `drop_first=True`: línea base (`DummyRegressor`), regresión lineal solo con la cuenta, con cuenta y tamaño, y con todas las variables. Reporta R^2 y MAE en prueba.
6. **Validación cruzada.** `cross_val_score` con `cv=5` para los modelos lineales; reporta la media y el rango de los pliegues.
7. **Conclusiones y limitaciones.** Responde la pregunta con cifras y enumera lo que los datos no permiten afirmar.

Cifras esperadas

Tu solución debe reproducir estos valores:

- 244 filas, 7 columnas, 0 nulos y 1 fila duplicada.
- Cuenta media 19.79 y propina media 3.00 dólares (mínimo 1.00, máximo 10.00).
- Porcentaje: media 16.08 y mediana 15.48; máximo 71.03 (cuenta 7.25, propina 5.15).
- Por día (n, propina, %): Thur 62 / 2.77 / 16.13; Fri 19 / 2.73 / 16.99; Sat 87 / 2.99 / 15.32; Sun 76 / 3.26 / 16.69.
- Dinner 176 / 3.10 / 15.95; Lunch 68 / 2.73 / 16.41. Female 87 / 2.83 / 16.65; Male 157 / 3.09 / 15.77. Fumadores: No 2.99, Yes 3.01.
- Porcentaje por tamaño de mesa: 1 → 21.73, 2 → 16.57, 3 → 15.22, 4 → 14.59, 5 → 14.15 y 6 → 15.62 (solo 4 mesas de una persona y 4 de seis).
- Correlación propina–cuenta 0.676; propina–tamaño 0.489.
- Línea base: MAE 1.038 (R^2 −0.079). Solo cuenta: coeficiente 0.1094, intercepto 0.862, R^2 0.44, MAE 0.652. Cuenta y tamaño: R^2 0.418, MAE 0.673. Todas: R^2 0.35, MAE 0.697.
- Validación cruzada: solo cuenta 0.457; cuenta y tamaño 0.466 (pliegues de 0.254 a 0.611); todas 0.421.
- Con cuenta y tamaño: una cuenta de 20 dólares en mesa de 2 predice 2.92; una de 40 en mesa de 4, 5.30.

La lectura que se espera

La cuenta explica la mayor parte de lo que se puede explicar: unos 11 centavos de propina por cada dólar de consumo. Sumar el tamaño de la mesa empata con el modelo de solo cuenta en validación cruzada (0.466 contra 0.457). Agregar día, horario, sexo y fumador no mejora la predicción: la empeora en prueba (0.35 contra 0.44) y en validación cruzada (0.421), porque con 244 mesas el modelo aprende ruido. Las diferencias entre grupos son pequeñas y descansan en pocos datos (19 mesas en viernes, 4 mesas de una persona). Un buen informe lo dice con esas cifras y no presenta diferencias de pocos centavos como hallazgos.

Extras opcionales

- **Otros modelos.** Prueba `RandomForestRegressor` u otro modelo no lineal y compáralo con la misma validación. Explica si la complejidad extra vale la pena.
- **Modelar el porcentaje.** Usa `pct` como objetivo y compara qué variables importan.
- **Gapminder por década.** Calcula la correlación entre `lifeExp` y `log10(gdpPercap)` para cada año del dataset. Valor esperado: 0.748 en 1952, máximo de 0.874 en 1987 y 0.809 en 2007. Explica, con lo visto en el curso, qué pudo hacer bajar la relación después de 1987.
- **Una gráfica interactiva** con plotly para explorar las propinas por grupo.

Entregables

1. **Un notebook** que corra de principio a fin sin errores, con la estructura: pregunta y datos, calidad, exploración, gráficas, modelos, validación cruzada, conclusiones.
2. **Una tabla de resultados** con cada modelo, su R^2 y MAE en prueba y su R^2 medio en validación cruzada.
3. **Un resumen** de media página para el gerente del restaurante: un mensaje principal, una gráfica, las cifras con su incertidumbre y las limitaciones (244 mesas, un mesero, un restaurante, correlación que no prueba causa).

Rúbrica (100 puntos)

- **Calidad de datos (10 puntos):** revisión completa y decisión justificada sobre el duplicado.
- **Exploración (20 puntos):** `pct` bien calculado, tablas por grupo con `n` y cifras esperadas.
- **Visualización (15 puntos):** gráficas adecuadas a la pregunta, con título, unidades y ejes honestos.
- **Modelos (25 puntos):** línea base, tres modelos lineales sin fuga y métricas correctas.

- **Validación (10 puntos):** validación cruzada con media y rango, usada para elegir el modelo.
- **Conclusiones y comunicación (20 puntos):** respuesta con cifras, limitaciones claras y resumen comprensible para alguien sin formación técnica.

Los extras no suman puntos, pero muestran si sabes distinguir una mejora real de un efecto del azar.