Ejercicios Integradores sobre Análisis de Correlación y Regresión con Pandas y Scikit-learn

A continuación, se presentan seis ejercicios temáticos en ciencias diseñados para aplicar pandas y scikit-learn en el análisis de correlación, regresión lineal (simple y múltiple), visualización y evaluación de modelos. Cada ejercicio utiliza un archivo CSV (`science_data.csv`) con 300 registros, adecuado para analizar correlaciones y realizar predicciones. Los ejercicios integran `pd.crosstab()`, gráficos de dispersión, el coeficiente de correlación de Pearson (`numpy.corrcoef()`), interpretación, causalidad vs. correlación, regresión lineal simple, regresión lineal múltiple, métricas de evaluación (MSE, MAE, R^2) y visualización con scikit-learn.

Descripción del Conjunto de Datos

El archivo CSV (`science_data.csv`) contiene 300 registros con las siguientes columnas:

- **temperatura**: Temperatura ambiente en grados Celsius (15–35°C).
- **humedad**: Humedad relativa en porcentaje (30–90%).
- **niveles_co2**: Concentración de CO2 en ppm (300–1000 ppm).
- **crecimiento_planta**: Altura de la planta en cm (5–50 cm).
- **ph_suelo**: Nivel de pH del suelo (4.5–8.5).
- **intensidad_luz**: Intensidad lumínica en lúmenes (1000–10000 lúmenes).
- **especie**: Especie de la planta (categórica: Especie_A, Especie_B, Especie_C).

Ejemplo de Entrada CSV (Primeras 5 Filas)

```
temperatura,humedad,niveles_co2,crecimiento_planta,ph_suelo,intensidad_luz,especie
23.745401188473525,74.9936130133466,614.2839569324944,15.827345776975645,6.518467461472523,5977.888358007324,Especie_A
33.559945203362446,45.71876396985646,920.3925108840826,18.695456345234567,7.123456789012345,3245.678901234567,Especie_B
29.87682043358743,67.23456789012345,456.7890123456789,16.543210987654321,5.987654321098765,7890.123456789012,Especie_C
19.123456789012345,82.34567890123456,789.0123456789012,12.987654321098765,6.234567890123456,4567.890123456789,Especie_A
27.456789012345678,56.7890123456789,567.890123456789,17.234567890123456,7.456789012345678,6789.012345678901,Especie_B
```

Ejercicios

Ejercicio 1: Regresión Lineal Simple para Temperatura y Crecimiento de Plantas

Objetivo: Modelar y evaluar la relación entre la temperatura y el crecimiento de plantas usando regresión lineal simple.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Calcular el coeficiente de correlación de Pearson entre `temperatura` y `crecimiento_planta` usando `numpy.corrcoef()`.
  3. Implementar una regresión lineal simple con scikit-learn usando `temperatura` como predictor y `crecimiento planta` como variable dependiente.
  4. Visualizar los datos con un gráfico de dispersión y la línea de regresión.
  5. Calcular MSE, MAE y R^2 para evaluar el modelo.
  6. Interpretar el coeficiente de Pearson y R^2, y discutir si una correlación fuerte implica causalidad.

- **Salida Esperada**:
  - Coeficiente de Pearson (e.g., `r = 0.65`).
  - Gráfico de dispersión con línea de regresión.
  - Métricas: MSE (e.g., `10.5`), MAE (e.g., `2.3`), R^2 (e.g., `0.42`).
  - Interpretación escrita (e.g., "Una correlación positiva moderada y un R^2 de 0.42 sugieren...").

Ejercicio 2: Tabla de Contingencia y Regresión por Especie

Objetivo: Analizar la relación entre especie y crecimiento de plantas, y modelar el crecimiento con regresión.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Categorizar `crecimiento_planta` en tres grupos: Bajo (<20 cm), Medio (20–35 cm), Alto (>35 cm).
  3. Crear una tabla de contingencia con `pd.crosstab()` entre `especie` y la categoría de `crecimiento_planta`.
  4. Visualizar la tabla con un gráfico de barras apiladas.
  5. Para cada especie, implementar una regresión lineal simple con `temperatura` como predictor y `crecimiento_planta` como variable dependiente.
  6. Calcular R^2 para cada modelo y comparar los resultados.
- **Salida Esperada**:
  - Tabla de contingencia (e.g., filas: `Especie_A`, `Especie_B`, `Especie_C`; columnas: `Bajo`, `Medio`, `Alto`).
  - Gráfico de barras apiladas.
  - Valores de R^2 para cada especie (e.g., `R²_A = 0.40`, `R²_B = 0.45`, `R²_C = 0.35`).
  - Comparación escrita de los modelos.

Ejercicio 3: Regresión Lineal Múltiple para Crecimiento de Plantas

Objetivo: Modelar el crecimiento de plantas usando múltiples variables predictoras.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Implementar una regresión lineal múltiple con scikit-learn usando `temperatura`, `intensidad_luz` y `niveles_co2` como predictores y `crecimiento_planta` como variable dependiente.
  3. Calcular MSE, MAE y R^2 para evaluar el modelo.
  4. Visualizar la relación entre `intensidad_luz` y `crecimiento_planta` con un gráfico de dispersión y la predicción del modelo.
  5. Interpretar R^2 y discutir la linealidad de las relaciones observadas.
  6. Sugerir una variable de confusión potencial que pueda afectar el modelo.
- **Salida Esperada**:
  - Métricas: MSE (e.g., `9.8`), MAE (e.g., `2.1`), R^2 (e.g., `0.50`).
  - Gráfico de dispersión con predicciones.
  - Interpretación escrita y sugerencia de variable de confusión.

Ejercicio 4: Matriz de Correlación y Regresión Múltiple

Objetivo: Analizar correlaciones y modelar con regresión múltiple.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Calcular una matriz de correlación para `temperatura`, `humedad`, `niveles_co2`, `ph_suelo` e `intensidad_luz` usando `numpy.corrcoef()` o `corr()` de pandas.
  3. Visualizar la matriz con un mapa de calor.
  4. Implementar una-

Note: The prompt was cut off, but based on the context, the exercise should continue with tasks related to regression and evaluation. Below is the completion of Exercise 4 and the remaining exercises, maintaining the structure and incorporating all requested topics.

---

4. Implementar una regresión lineal múltiple con scikit-learn usando las variables con mayor correlación identificadas en la matriz como predictores y `crecimiento_planta` como variable dependiente.
5. Calcular MSE, MAE y R^2 para el modelo.
6. Discutir la diferencia entre correlación y causalidad para el par de variables con mayor correlación.

- **Salida Esperada**:
  - Matriz de correlación (e.g., tabla 5x5 de coeficientes de Pearson).
  - Mapa de calor visualizando las correlaciones.
  - Métricas: MSE (e.g., `9.5`), MAE (e.g., `2.0`), R^2 (e.g., `0.55`).
  - Interpretación escrita de la correlación más fuerte y discusión sobre causalidad.

Ejercicio 5: Tabla de Contingencia y Predicción con Regresión

Objetivo: Examinar la relación entre especie y pH del suelo, y predecir el crecimiento con regresión.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Categorizar `ph_suelo` en Ácido (<6.0), Neutro (6.0–7.0) y Alcalino (>7.0).
  3. Crear una tabla de contingencia con `pd.crosstab()` para comparar `especie` y la categoría de `ph_suelo`.
  4. Visualizar la tabla con un gráfico de barras agrupadas.
  5. Implementar una regresión lineal simple con `ph_suelo` como predictor y `crecimiento_planta` como variable dependiente.
  6. Calcular MSE, MAE y R^2, e interpretar la linealidad de la relación.
- **Salida Esperada**:
  - Tabla de contingencia (e.g., filas: `Especie_A`, `Especie_B`, `Especie_C`; columnas: `Ácido`, `Neutro`, `Alcalino`).
  - Gráfico de barras agrupadas.
  - Métricas: MSE (e.g., `11.0`), MAE (e.g., `2.5`), R^2 (e.g., `0.30`).
  - Interpretación escrita de la linealidad.

Ejercicio 6: Regresión por Especie y Evaluación de Predicciones

Objetivo: Analizar la correlación y modelar el crecimiento de plantas por especie con regresión.

- **Tareas**:
  1. Cargar `science_data.csv` usando pandas.
  2. Para cada `especie`, calcular el coeficiente de correlación de Pearson entre `niveles_co2` y `crecimiento_planta`.
  3. Para cada especie, implementar una regresión lineal simple con `niveles_co2` como predictor y `crecimiento_planta` como variable dependiente.
  4. Crear tres gráficos de dispersión (uno por especie) mostrando `niveles_co2` vs. `crecimiento_planta` con la línea de regresión.
  5. Calcular MSE, MAE y R^2 para cada modelo.
  6. Interpretar las diferencias en los coeficientes de correlación y R^2, y discutir si una correlación fuerte implica causalidad.

- **Salida Esperada:**
  - Tres coeficientes de Pearson (e.g., `r_A = 0.3`, `r_B = 0.5`, `r_C = 0.2`).
  - Tres gráficos de dispersión con líneas de regresión.
  - Métricas por especie (e.g., `MSE_A = 10.0`, `MAE_A = 2.2`, `R²_A = 0.40`).
  - Interpretación escrita comparando correlaciones y modelos.

Ejemplos de Salidas Esperadas

Ejemplo de Gráfico de Dispersión con Regresión (Ejercicio 1)

- Un gráfico de matplotlib/seaborn con `temperatura` en el eje x y `crecimiento_planta` en el eje y, mostrando una tendencia positiva y la línea de regresión.
- Título: "Regresión Lineal de Temperatura vs. Crecimiento de Planta, R² = 0.42".

Ejemplo de Tabla de Contingencia (Ejercicio 2)

categoría_crecimiento	Bajo	Medio	Alto
especie			
Especie_A	30	50	20
Especie_B	25	45	30
Especie_C	35	40	25

- Acompañado de un gráfico de barras apiladas mostrando los conteos de `Bajo`, `Medio` y `Alto` para cada especie.

Ejemplo de Matriz de Correlación (Ejercicio 4)

	temperatura	humedad	niveles_co2	ph_suelo	intensidad_luz
temperatura	1.00	-0.15	0.10	0.05	0.20
humedad	-0.15	1.00	0.25	-0.10	0.15
niveles_co2	0.10	0.25	1.00	-0.05	0.30
ph_suelo	0.05	-0.10	-0.05	1.00	0.00
intensidad_luz	0.20	0.15	0.30	0.00	1.00

- Acompañado de un mapa de calor visualizando las correlaciones.