Análisis Exploratorio

de Datos

Sesión 3

Correlación de Variables: Análisis y Visualización

- La correlación es una medida estadística que describe el grado de relación entre dos variables. En ciencia de datos, comprender la correlación es esencial para identificar patrones, predecir comportamientos y tomar decisiones informadas.

- Es fundamental recordar que correlación no implica causalidad, lo que significa que una relación entre dos variables no necesariamente indica que una causa a la otra. Existen diferentes tipos de correlación: positiva (cuando ambas variables aumentan), negativa (cuando una aumenta y otra disminuye) y nula (sin relación clara).

Tablas de Contingencia

¿Qué son?

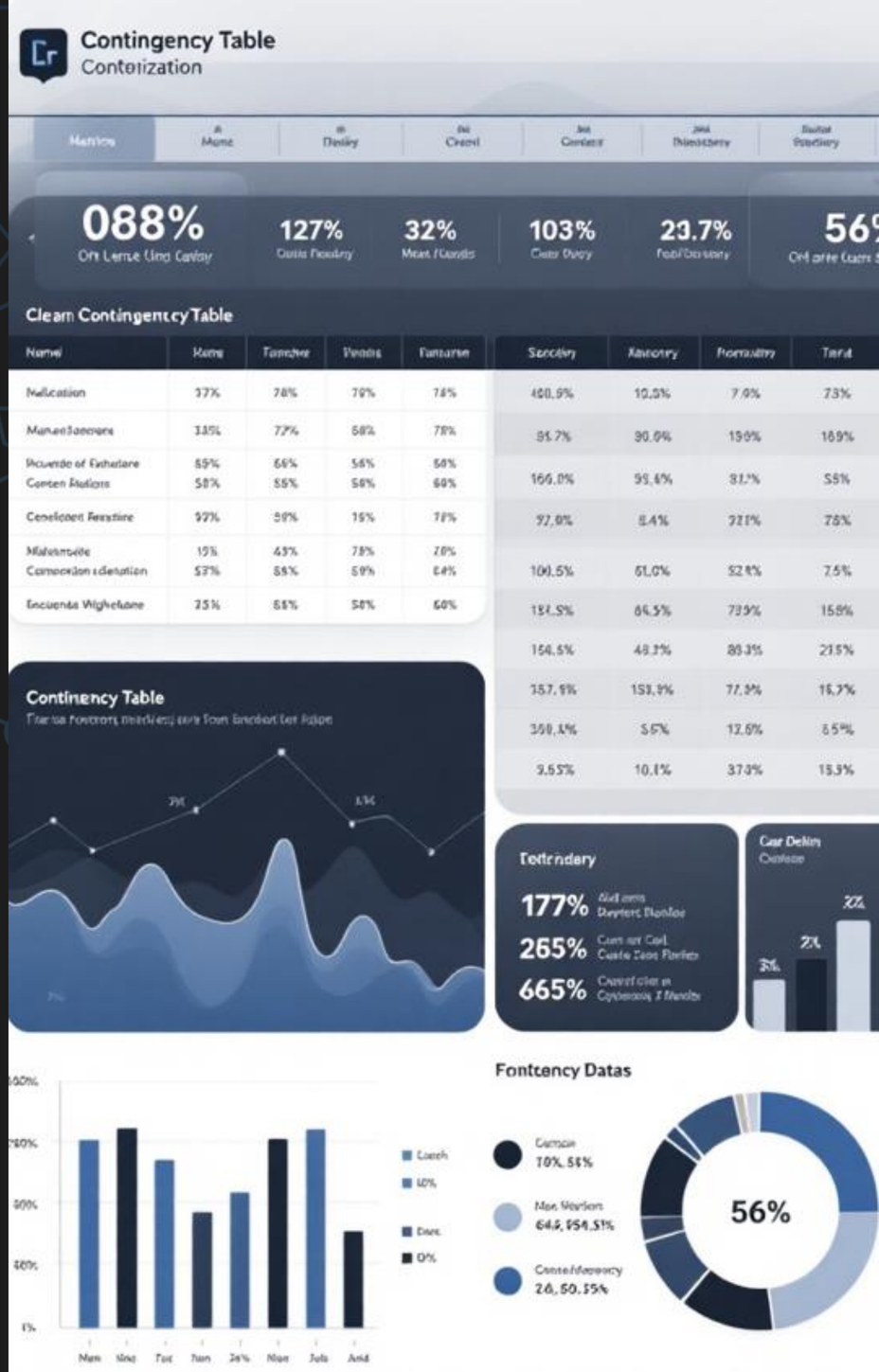
Una tabla de contingencia es una tabla de frecuencias utilizada para analizar la relación entre dos variables categóricas. Muestra cuántas veces ocurre cada combinación de categorías y permite evaluar si existe alguna asociación entre ellas.

¿Cuándo usarlas?

Para examinar la relación entre dos variables categóricas (por ejemplo, género y preferencia de producto), calcular probabilidades condicionales y frecuencias relativas, y construir una tabla de chi-cuadrado para evaluar independencia entre variables.

Interpretación

Las filas y columnas representan las categorías de cada variable. Cada celda muestra la frecuencia de esa combinación específica, permitiendo analizar patrones y asociaciones entre las variables categóricas.



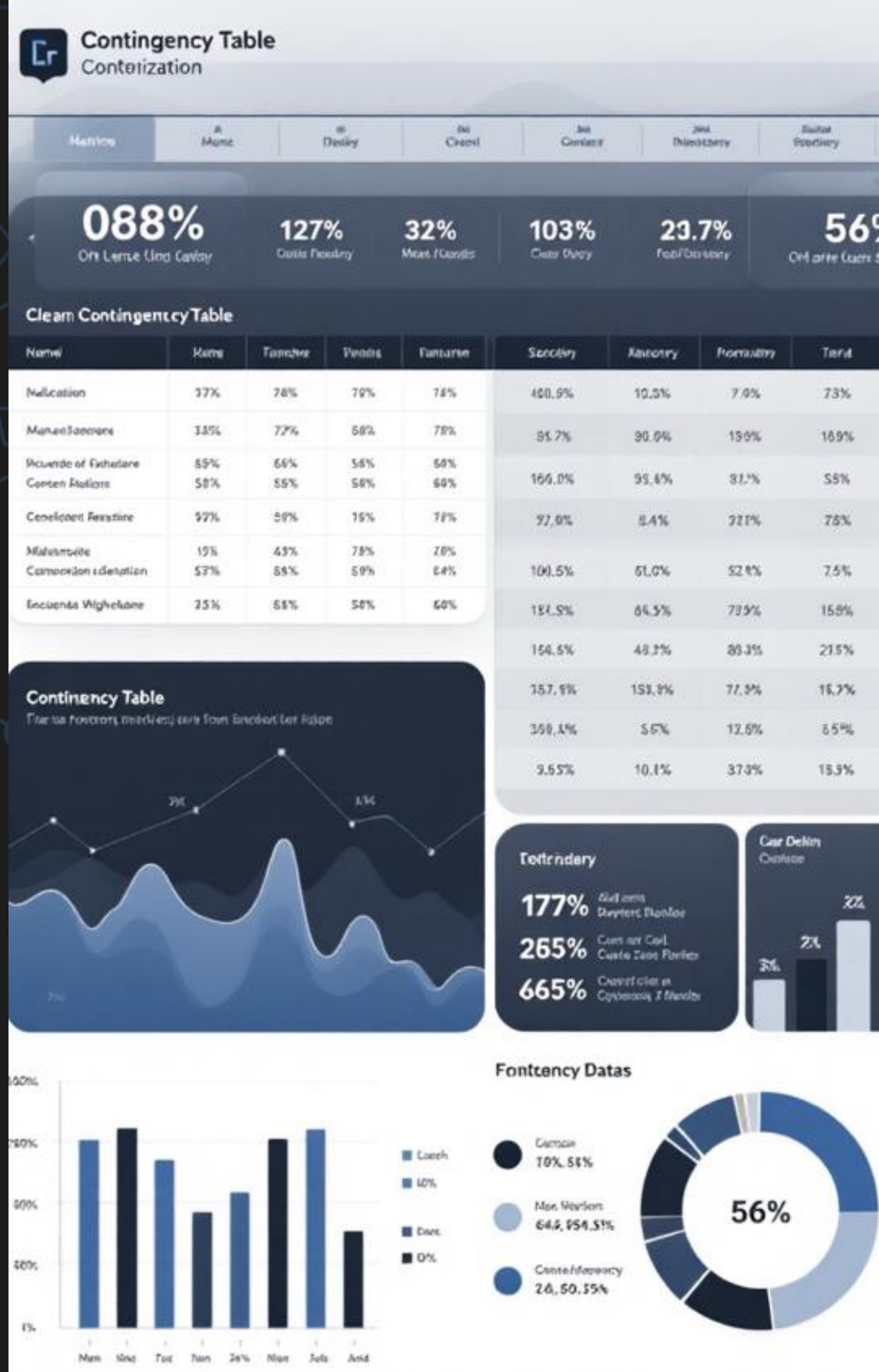
Ejemplo en Python

```python
import pandas as pd

# Crear un DataFrame con variables categóricas
df = pd.DataFrame({
    'Género': ['Masculino', 'Femenino', 'Femenino', 'Masculino'],
    'Preferencia': ['Deportes', 'Cine', 'Deportes', 'Cine']
})

# Crear tabla de contingencia
tabla_contingencia = pd.crosstab(df['Género'], df['Preferencia'])
print(tabla_contingencia)
```

📌Tabla de Contingencia Resultante:

La función pd.crosstab() cuenta la frecuencia de cada combinación de valores entre las columnas Género y Preferencia. El resultado es el siguiente:

GÉNERO	CINE	DEPORTES
Femenino	1	1
Masculino	1	1



Ejemplo en Python

Interpretación de la tabla de contingencia:

📌 Filas (Género):

- Femenino: Hay 1 persona de género femenino que prefiere Cine y 1 persona que prefiere Deportes.
- Masculino: Hay 1 persona de género masculino que prefiere Cine y 1 persona que prefiere Deportes.

📌 Columnas (Preferencia):

- Cine: En total, 2 personas prefieren Cine (1 femenino y 1 masculino).
- Deportes: En total, 2 personas prefieren Deportes (1 femenino y 1 masculino).

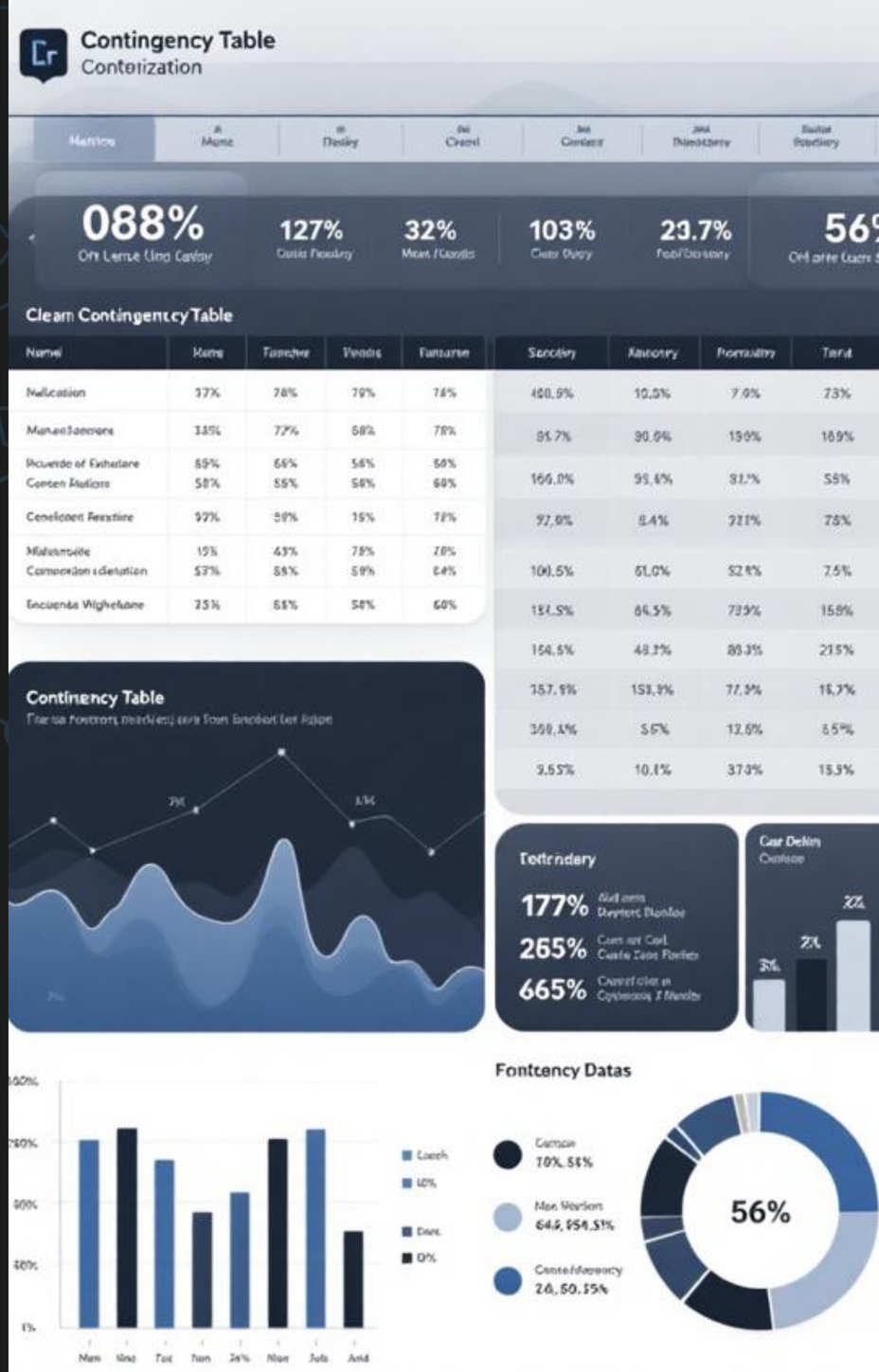


Gráfico Scatterplot

Definición

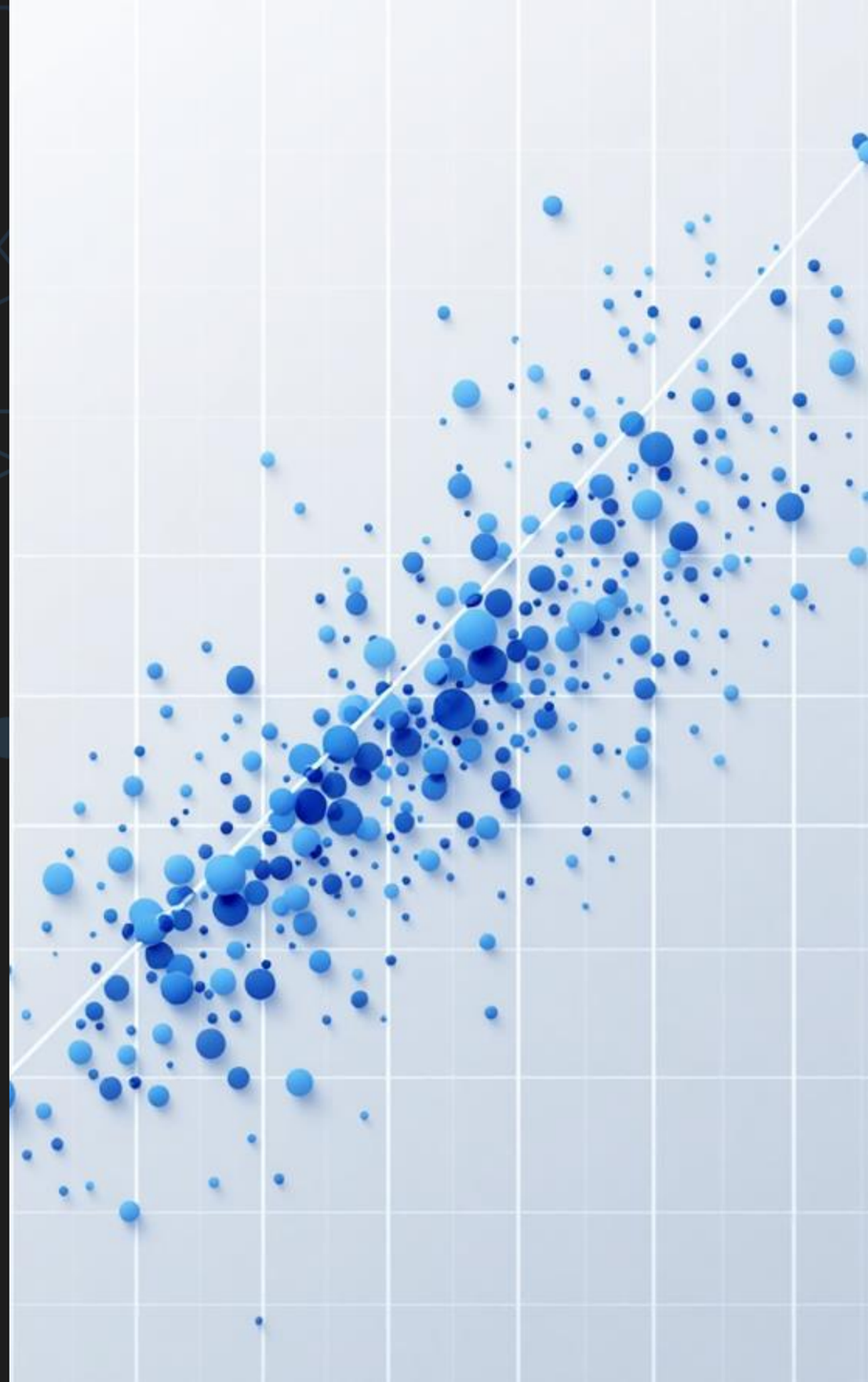
Un scatterplot (diagrama de dispersión) es una herramienta visual utilizada para examinar la relación entre dos variables numéricas. Cada punto en el gráfico representa una observación en el conjunto de datos.

Casos de uso

Es ideal para visualizar si existe una correlación entre dos variables numéricas, detectar outliers o patrones en los datos, y analizar la relación entre variables en problemas de regresión.

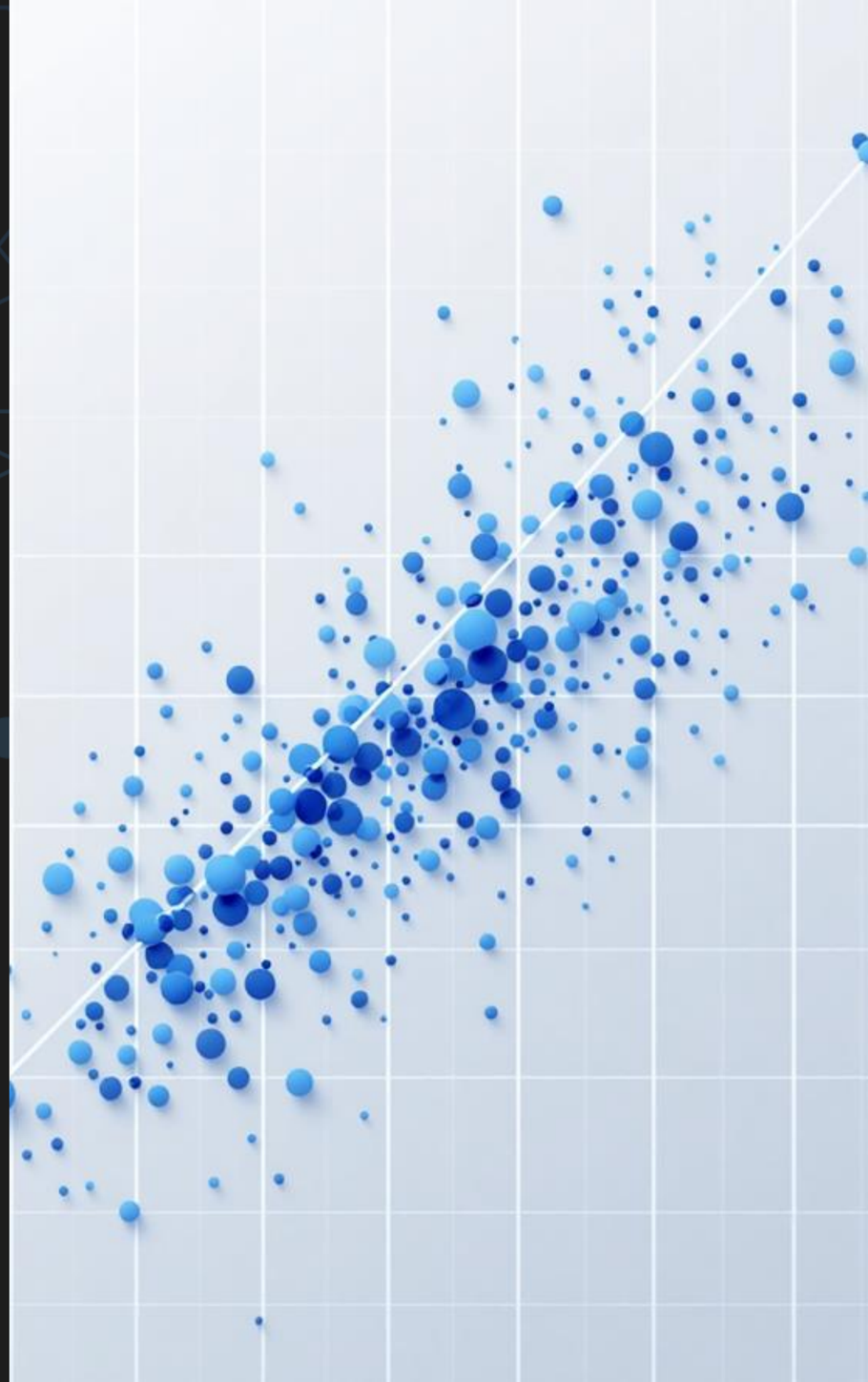
Interpretación

Si los puntos siguen una línea ascendente, hay una correlación positiva. Si forman una línea descendente, hay una correlación negativa. Si están dispersos sin una forma clara, no hay correlación.



Ejemplo en Python

```python
import matplotlib.pyplot as plt
import numpy as np

# Datos ficticios de ventas (cantidad de productos vendidos y precio promedio)
np.random.seed(42)
cantidad_vendida = np.random.randint(10, 100, 50)  # Número de productos vendidos
precio_promedio = cantidad_vendida * np.random.uniform(0.8, 1.2, 50)  # Precio con variación aleatoria

# Crear el scatterplot
plt.figure(figsize=(8, 5))
plt.scatter(cantidad_vendida, precio_promedio, color='blue', alpha=0.5)
plt.xlabel('Cantidad Vendida')
plt.ylabel('Precio Promedio')
plt.title('Relación entre Cantidad Vendida y Precio Promedio')
plt.grid(True)
plt.show()
```



📌Interpretación del scatterplot:

- Si los puntos siguen una línea ascendente, hay una correlación positiva (cuando una variable aumenta, la otra también).
- Si los puntos forman una línea descendente, hay una correlación negativa.
- Si los puntos están dispersos sin una forma clara, no hay correlación.

Coeficiente de Correlación de Pearson

Definición

El coeficiente de correlación de Pearson (r) es una medida estadística que cuantifica la relación lineal entre dos variables numéricas. Su rango va de -1 a 1, donde 1 indica correlación positiva perfecta, -1 correlación negativa perfecta, y 0 ausencia de correlación.

Cálculo

Se calcula mediante una fórmula que relaciona la covarianza de las variables con el producto de sus desviaciones estándar.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Ejemplo en Python:

Podemos calcular el coeficiente de correlación de Pearson en Python usando scipy.stats.pearsonr() o numpy.corrcoef().

Ejemplo 1: Usando SciPy

```python
import numpy as np
import scipy.stats as stats

# Datos ficticios: Ventas de un producto y presupuesto de marketing
presupuesto_marketing = np.array([1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500])
ventas = np.array([200, 270, 340, 410, 480, 550, 600, 660])

# Calcular el coeficiente de Pearson
coef, p_valor = stats.pearsonr(presupuesto_marketing, ventas)

print(f"Coeficiente de correlación de Pearson: {coef:.2f}")
print(f"P-valor: {p_valor:.4f}")
```

Ejemplo en Python:

Podemos calcular el coeficiente de correlación de Pearson en Python usando scipy.stats.pearsonr() o numpy.corrcoef().

Ejemplo 1: Usando NumPy

```
# Calcular correlación usando NumPy
corr_matrix = np.corrcoef(presupuesto_marketing, ventas)
print(f"Coeficiente de Pearson usando NumPy: {corr_matrix[0, 1]:.2f}")
```

Interpretación del Coeficiente de Pearson

Los valores de r se interpretan según la siguiente escala:

Rango de r	Interpretación
0.8 ≤ r ≤ 1.0	Correlación positiva fuerte
0.5 r 0.8	Correlación positiva moderada
0.2 r 0.5	Correlación positiva débil
-0.2 r 0.2	Correlación despreciable o nula
-0.5 r -0.2	Correlación negativa débil
-0.8 r -0.5	Correlación negativa moderada
-1.0 r -0.8	Correlación negativa fuerte

Si el p-valor es menor a 0.05, la correlación es estadísticamente significativa.

Consideraciones al Usar el Coeficiente de Pearson

Solo mide relaciones lineales

Si la relación entre las variables es no lineal (como una curva o parábola), el coeficiente de Pearson puede no detectarla correctamente o subestimar la fuerza de la relación. En estos casos, es mejor utilizar otras medidas como la correlación de Spearman.

Es sensible a outliers

Los valores extremos o atípicos pueden influir significativamente en el coeficiente y dar resultados engañosos. Es importante identificar y tratar adecuadamente los outliers antes de calcular la correlación.

No implica causalidad

Una alta correlación entre dos variables no significa que una cause la otra. Pueden estar relacionadas debido a un tercer factor o por pura coincidencia. Siempre se debe complementar con análisis adicionales.

Complementar con visualización

Un gráfico de dispersión (scatterplot) puede ayudar a interpretar mejor la relación entre las variables y detectar patrones que el coeficiente por sí solo no revela.

Causalidad vs. Correlación

1

Correlación

Mide la relación estadística entre dos variables, indicando si tienden a moverse juntas. No establece dirección del efecto y puede estar influenciada por terceros factores. Se comprueba mediante métodos estadísticos como el coeficiente de Pearson.

2

Diferencias Clave

La correlación no implica relación directa, mientras que la causalidad sí. La causalidad establece dirección del efecto mediante pruebas, mientras que la correlación no. En la causalidad se intentan descartar terceros factores, en la correlación pueden influir.

3

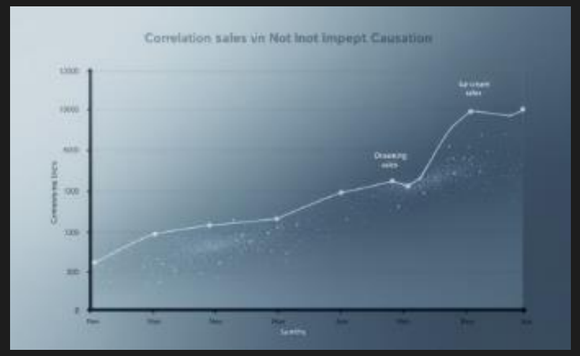
Causalidad

Ocurre cuando un cambio en una variable provoca directamente un cambio en otra. Requiere pruebas rigurosas y descartar otros factores influyentes. Se demuestra mediante experimentos controlados y estudios longitudinales.

Diferencias Clave entre Correlación y Causalidad

Característica	Correlación	Causalidad
Relación Directa	No necesariamente	Sí
Dirección del Efecto	No se puede determinar	Se establece mediante pruebas
Terceros Factores	Pueden influir	Se intentan descartar
Comprobación	Métodos estadísticos como Pearson	Experimentos y estudios longitudinales

Ejemplos de Confusión entre Correlación y Causalidad



Helados y Ahogamientos
Existe una correlación positiva entre las ventas de helado y los casos de ahogamiento. Esto no significa que comer helado cause ahogamientos, sino que ambos aumentan en verano debido a un tercer factor: el clima cálido que lleva a más personas a comprar helados y a nadar.

Cigüeñas y Nacimientos
En algunas regiones, se ha observado una correlación entre la población de cigüeñas y la tasa de natalidad. Esto no implica que las cigüeñas traigan bebés, sino que ambas variables pueden estar relacionadas con factores como la ruralidad o el desarrollo económico.

Felicidad y Longevidad
Aunque existe correlación entre felicidad y longevidad, no podemos afirmar que ser feliz cause directamente una vida más larga. Factores como el nivel socioeconómico, hábitos de salud o genética pueden influir en ambas variables.

Mejores Prácticas

1. **Toma de Decisiones Informadas**
   Basadas en análisis riguroso

2. **Combinación de Métodos**
   Estadísticos y experimentales

3. **Visualización de Datos**
   Complementar con gráficos

4. **Análisis Crítico**
   Cuestionar relaciones aparentes

5. **Conocimiento del Dominio**
   Contexto del problema

📌 La correlación es útil para identificar patrones, pero no implica causalidad. Para demostrar causalidad, se requieren experimentos, estudios longitudinales y modelos de regresión. En ciencia de datos, es crucial diferenciar entre ambas para evitar conclusiones erróneas y tomar decisiones informadas.

📌 Muchas veces, encontrar una correlación puede ser el primer paso para investigar una posible causalidad, pero nunca debe ser la única evidencia. El análisis riguroso y la combinación de métodos estadísticos con conocimiento del dominio son fundamentales para establecer relaciones causales válidas.

Preguntas

Sección de preguntas

Análisis Exploratorio

de Datos

Continúe con las actividades