Análisis Descriptivo de Datos con Python

Objetivo:

Realizar un análisis descriptivo de un conjunto de datos utilizando pandas, numpy, y matplotlib. El análisis descriptivo incluirá medidas de tendencia central, dispersión y visualización de los datos.

Datos:

Utilizaremos un conjunto de datos ficticio de calificaciones de estudiantes en diferentes asignaturas.

Pasos:

1. **Cargar las librerías necesarias y los datos**:
    - pandas para la manipulación de datos.
    - numpy para cálculos numéricos.
    - matplotlib y seaborn para visualización.
2. **Descripción general de los datos**:
    - Visualización de las primeras filas del conjunto de datos.
    - Resumen estadístico de las variables numéricas.
    - Información sobre el tipo de datos y valores nulos.
3. **Medidas de tendencia central y dispersión**:
    - Cálculo de la media, mediana y moda.
    - Cálculo de la desviación estándar, varianza, mínimo y máximo.
4. **Visualización de los datos**:
    - Histogramas y boxplots para visualizar la distribución de las calificaciones.
    - Diagramas de dispersión para explorar relaciones entre las asignaturas.

```python
# Paso 1: Cargar las librerías necesarias y los datos
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Crear un conjunto de datos ficticio
data = {
    'Matematicas': np.random.randint(50, 100, size=100),
    'Fisica': np.random.randint(40, 95, size=100),
    'Quimica': np.random.randint(45, 90, size=100),
    'Biologia': np.random.randint(50, 100, size=100)
}

df = pd.DataFrame(data)
```

```python
# Paso 2: Descripción general de los datos
print("Primeras filas del conjunto de datos:")
print(df.head())

print("\nResumen estadístico de las variables numéricas:")
print(df.describe())

print("\nInformación sobre el tipo de datos y valores nulos:")
print(df.info())

# Paso 3: Medidas de tendencia central y dispersión
print("\nMedidas de tendencia central:")
print(f"Media:\n{df.mean()}")
print(f"Mediana:\n{df.median()}")
print(f"Moda:\n{df.mode().iloc[0]}")

print("\nMedidas de dispersión:")
print(f"Desviación estándar:\n{df.std()}")
print(f"Varianza:\n{df.var()}")
print(f"Mínimo:\n{df.min()}")
print(f"Máximo:\n{df.max()}")
```

```python
# Paso 4: Visualización de los datos
# Histograma
plt.figure(figsize=(12, 8))
df.hist(bins=10, edgecolor='black', grid=False, figsize=(10, 8))
plt.suptitle('Histogramas de Calificaciones', size=16)
plt.show()

# Boxplot
plt.figure(figsize=(12, 8))
sns.boxplot(data=df)
plt.title('Boxplots de Calificaciones', size=16)
plt.show()

# Diagrama de dispersión
plt.figure(figsize=(12, 8))
sns.pairplot(df)
plt.suptitle('Diagramas de Dispersión entre Asignaturas', size=16)
plt.show()
```

Descripción de los pasos:

1. **Cargar librerías y datos**:
    - Importamos las librerías necesarias.
    - Creamos un DataFrame con datos ficticios.
2. **Descripción general de los datos**:
    - Utilizamos head() para ver las primeras filas del DataFrame.
    - Utilizamos describe() para obtener un resumen estadístico.
    - Utilizamos info() para obtener información sobre el tipo de datos y valores nulos.
3. **Medidas de tendencia central y dispersión**:
    - Calculamos la media, mediana y moda de cada columna.
    - Calculamos la desviación estándar, varianza, mínimo y máximo de cada columna.
4. **Visualización de los datos**:
    - Creamos histogramas para visualizar la distribución de las calificaciones.
    - Creamos boxplots para visualizar la dispersión y detectar posibles valores atípicos.
    - Creamos diagramas de dispersión para explorar relaciones entre las calificaciones de diferentes asignaturas.