Estadística Descriptiva - Diabetes Dataset

Estadística Descriptiva Análisis de Datos en Python

Miembros del equipo: A00839729 | Josué Santiago Aguiñaga Cázres | ITC A00842711 | César Pecero Lara | Negocios A00227413 | Andrés Martínez Ramos | IID

In [1]:
```
# Cargar Datos
import pandas as pd

df = pd.read_csv("diabetes.csv")
df.head()
```

Out[1]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeF
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

In [2]:
```
# Ver cantidad de datos y variables
print("Número de objetos (filas):", len(df))
print("Número de variables (columnas):", df.shape[1])
print("Columnas:", list(df.columns))
```

```
Número de objetos (filas): 768
Número de variables (columnas): 9
Columnas: ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
```

In [3]:
```
# Revisar información general
df.info()
df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000

In [4]:
```
# Verificar valores nulos
df.isnull().sum()
```

Out[4]:
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Descripción de Variables

- **Pregnancies:** Cuantitativa discreta → número de embarazos.
- **Glucose:** Cuantitativa continua → concentración de glucosa.
- **BloodPressure:** Cuantitativa continua → presión diastólica (mm Hg).
- **SkinThickness:** Cuantitativa continua → grosor del pliegue cutáneo (mm).
- **Insulin:** Cuantitativa continua → insulina sérica (mu U/ml).
- **BMI:** Cuantitativa continua → índice de masa corporal.

- **DiabetesPedigreeFunction:** Cuantitativa continua → función de historial familiar.
- **Age:** Cuantitativa discreta → edad del paciente (años).
- **Outcome:** Cualitativa nominal binaria → 1 = positivo a diabetes, 0 = negativo.

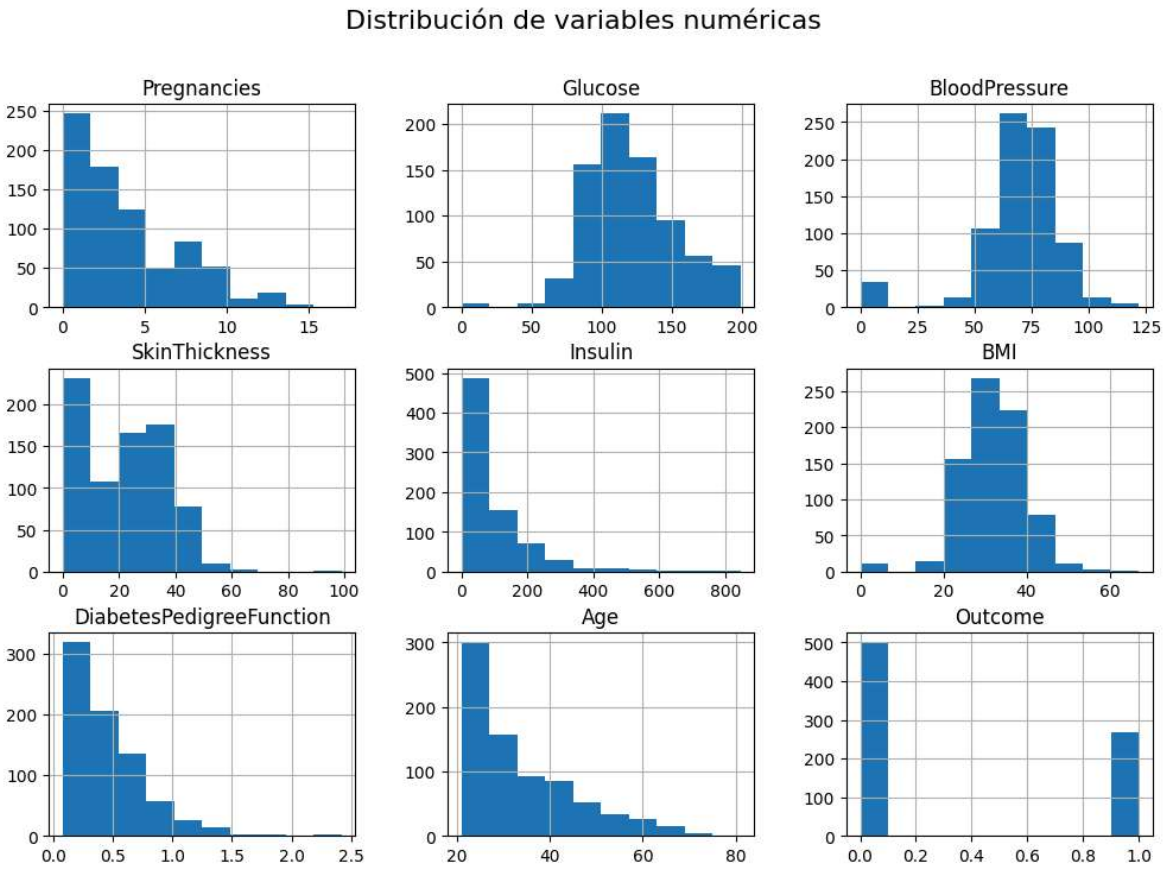
Análisis de variables (exploración individual)

In [5]:
```
# Rango, min, max, media y mediana
for col in df.columns:
    if df[col].dtype != 'object':
        print(f"📊 {col}")
        print(f"  Min: {df[col].min()}, Max: {df[col].max()}")
        print(f"  Media: {df[col].mean():.2f}, Mediana: {df[col].median():.2f}")
        print("-"*40)
```

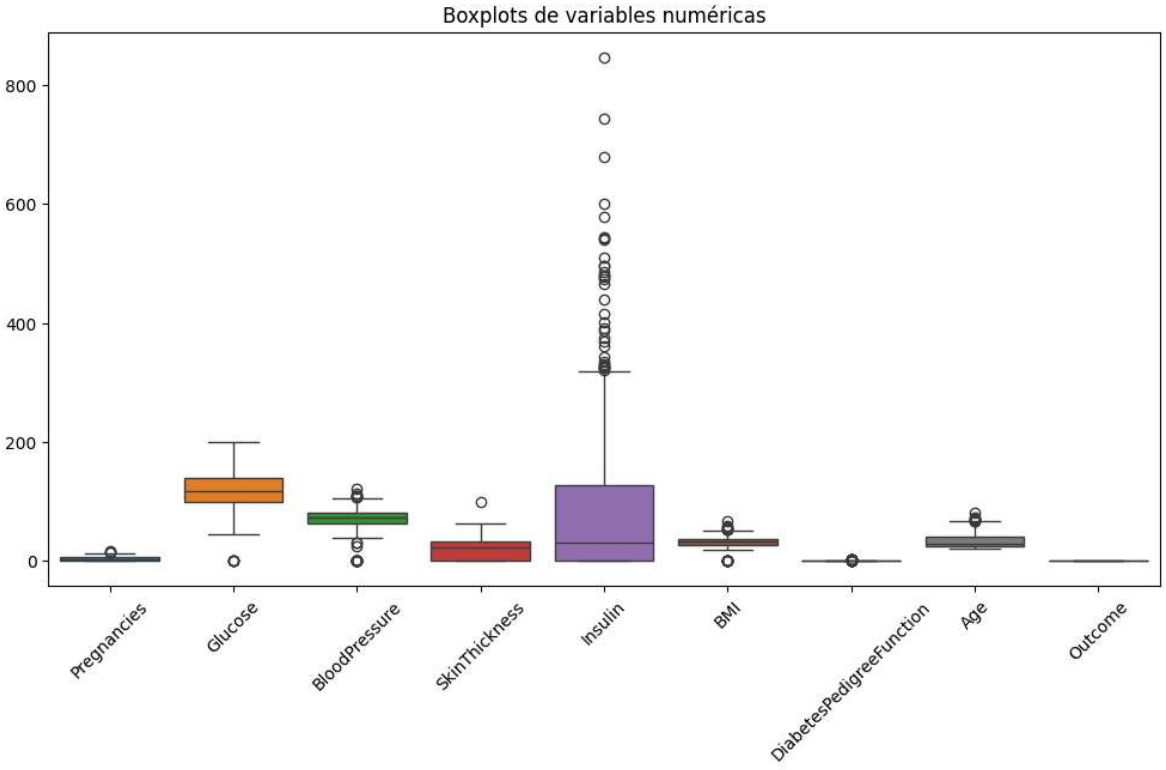
```
📊 Pregnancies
  Min: 0, Max: 17
  Media: 3.85, Mediana: 3.00
----------------------------------------
📊 Glucose
  Min: 0, Max: 199
  Media: 120.89, Mediana: 117.00
----------------------------------------
📊 BloodPressure
  Min: 0, Max: 122
  Media: 69.11, Mediana: 72.00
----------------------------------------
📊 SkinThickness
  Min: 0, Max: 99
  Media: 20.54, Mediana: 23.00
----------------------------------------
📊 Insulin
  Min: 0, Max: 846
  Media: 79.80, Mediana: 30.50
----------------------------------------
📊 BMI
  Min: 0.0, Max: 67.1
  Media: 31.99, Mediana: 32.00
----------------------------------------
📊 DiabetesPedigreeFunction
  Min: 0.078, Max: 2.42
  Media: 0.47, Mediana: 0.37
----------------------------------------
📊 Age
  Min: 21, Max: 81
  Media: 33.24, Mediana: 29.00
----------------------------------------
📊 Outcome
  Min: 0, Max: 1
  Media: 0.35, Mediana: 0.00
----------------------------------------
```

In [6]:
```
# Andres Martinez Ramos - A00227413
# Histogramas
import matplotlib.pyplot as plt
import seaborn as sns
```

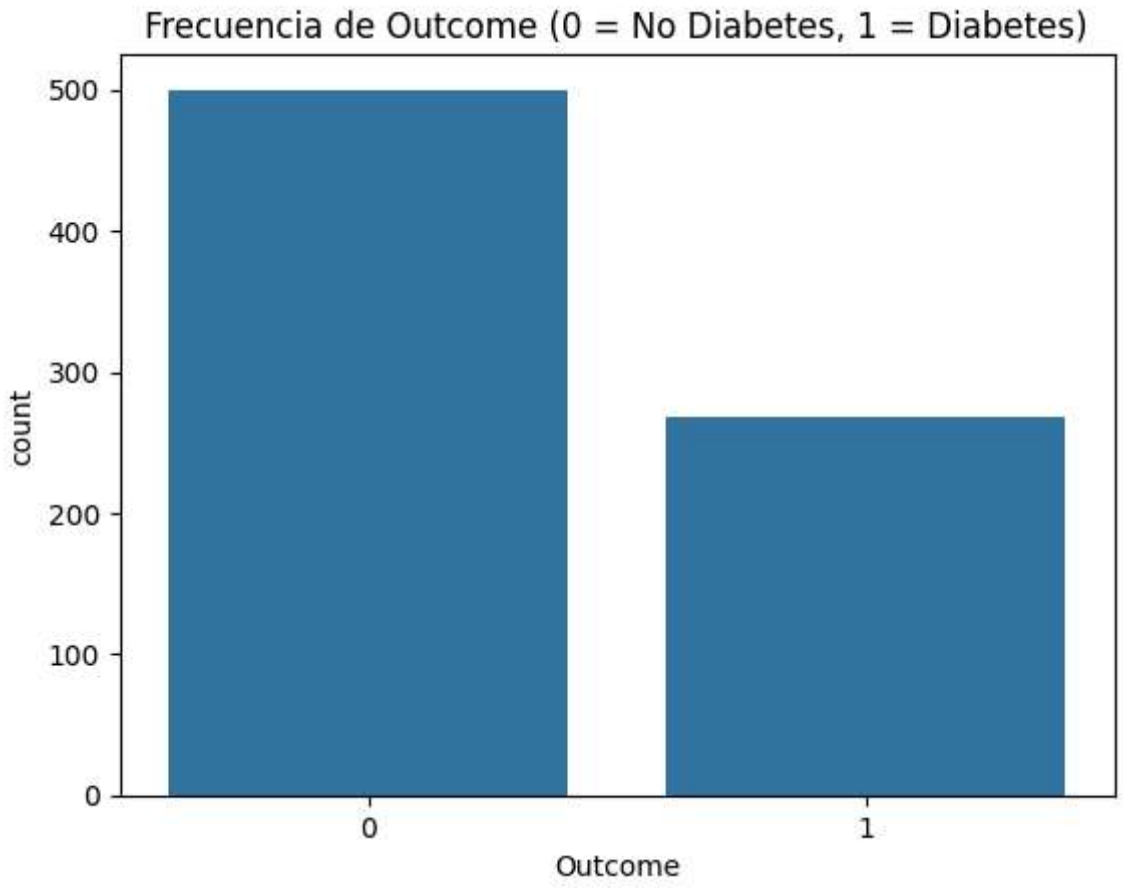
```
df.hist(figsize=(12,8))
plt.suptitle("Distribución de variables numéricas", fontsize=16)
plt.show()
```



In [7]:
```
# Andres Martinez Ramos - A00227413
# Boxplots
plt.figure(figsize=(12,6))
sns.boxplot(data=df)
plt.title("Boxplots de variables numéricas")
plt.xticks(rotation=45)
plt.show()
```



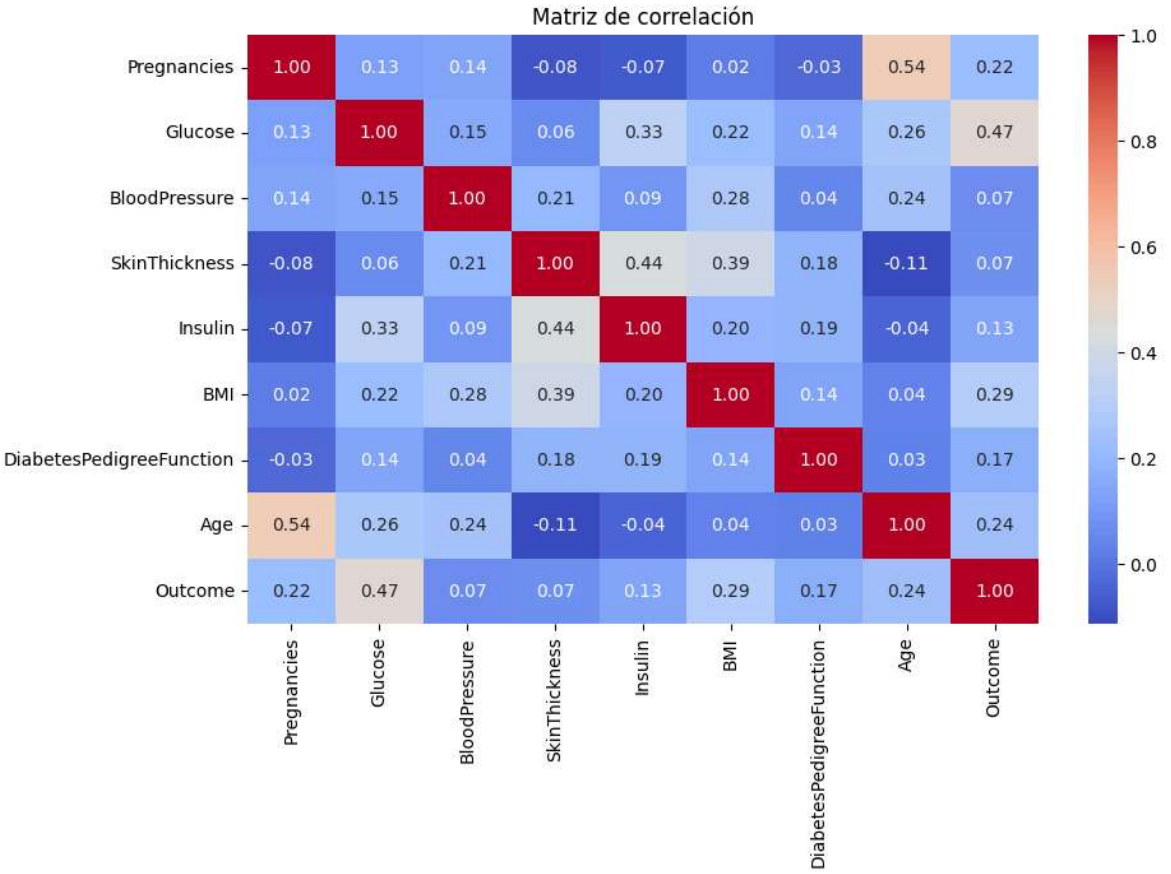
In [8]:
```
# César Pecero Lara - A00842711
# Analizar relación con Outcome
sns.countplot(x="Outcome", data=df)
plt.title("Frecuencia de Outcome (0 = No Diabetes, 1 = Diabetes)")
plt.show()

# promedios por grupo de Outcome
df.groupby("Outcome").mean().round(2)
```



Out[8]:

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesI
0	3.30	109.98	68.18	19.66	68.79	30.30	
1	4.87	141.26	70.82	22.16	100.34	35.14	

In [9]:
```
# Comprobación de correlaciones
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Matriz de correlación")
plt.show()
```



Conclusión

Conclusiones

- El dataset tiene 768 registros y 9 variables.
- La variable **Outcome** está balanceada de forma moderada (mayoría 0, menor cantidad 1).
- Variables como **Glucose** y **BMI** tienden a valores altos en pacientes con Outcome=1.
- Hay variables con ceros que representan datos faltantes (BloodPressure, SkinThickness, Insulin, BMI).
- Según la matriz de correlación, **Glucose** y **BMI** se correlacionan más con **Outcome.**