In [4]:
```
# Paso 1: Importar librerías básicas
import pandas as pd
import numpy as np
```

In [5]:
```
# Cargar la base de datos
import kagglehub

# Download latest version
path = kagglehub.dataset_download("uciml/pima-indians-diabetes-data

print("Path to dataset files:", path)
```

```
Using Colab cache for faster access to the 'pima-indians-diabetes-da
tabase' dataset.
Path to dataset files: /kaggle/input/pima-indians-diabetes-database
```

In [9]:
```
# Crear un Dataframe del Archivo
df = pd.read_csv(f"{path}/diabetes.csv")
```

In [11]:
```
# Columnas y primeras filas
df.head()
```

Out[11]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabe
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 [13]:
```
# Cantidad de datos
df.shape
```

Out[13]: (768, 9)

In [14]:
```
#Tipo de datos
df.info()
```

```
<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
```

In [16]:
```
#Valores nulos
df.isnull().sum()
```

Out[16]:

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

dtype: int64

Descripción de variables seleccionadas

Glucose: Variable cuantitativa continua. Representa el nivel de glucosa en la sangre.

BMI (Body Mass Index): Variable cuantitativa continua. Indica la relación entre el peso y la estatura de la persona.

Age: Variable cuantitativa discreta. Representa la edad del paciente en años.

Outcome: Variable categórica binaria. Indica si el paciente presenta diabetes (1) o no (0).

In [17]:
```
# Seleccionamos las variables de interés
variables = ['Glucose', 'BMI', 'Age', 'Outcome']

# Calculamos estadísticas descriptivas básicas
estadisticas = df[variables].describe().T  # Transponer para que se
estadisticas['mediana'] = df[variables].median()
estadisticas
```

Out[17]:

	count	mean	std	min	25%	50%	75%	max	me
Glucose	768.0	120.894531	31.972618	0.0	99.0	117.0	140.25	199.0	
BMI	768.0	31.992578	7.884160	0.0	27.3	32.0	36.60	67.1	
Age	768.0	33.240885	11.760232	21.0	24.0	29.0	41.00	81.0	
Outcome	768.0	0.348958	0.476951	0.0	0.0	0.0	1.00	1.0	

En general, los datos muestran que hay mucha diferencia en los niveles de glucosa y en el IMC entre los pacientes.
La mayoría son adultos jóvenes con una tendencia a tener sobrepeso, lo cual puede aumentar el riesgo de diabetes.
Alrededor del 35% de las personas del estudio tienen un resultado positivo a diabetes, lo que muestra una presencia importante de la enfermedad en este grupo.

In [19]:
```
# 1. Promedio de glucosa en personas con y sin diabetes
df.groupby('Outcome')['Glucose'].mean()
```

Out[19]:

Outcome	Glucose
0	109.980000
1	141.257463

dtype: float64

Las personas con diabetes tienen más glucosa (141) que las que no (109).

In [20]:
```
# 2. Promedio de BMI según si tienen diabetes o no
df.groupby('Outcome')['BMI'].mean()
```

Out[20]:

Outcome	BMI
0	30.304200
1	35.142537

dtype: float64

Su BMI también es más alto (35) comparado con los que no tienen diabetes (30).

In [22]:
```
# 3. Promedio de edad de las personas que tiene dianbtes y las que
df.groupby('Outcome')['Age'].mean()
```

Out[22]:

Outcome	Age
0	31.190000
1	37.067164

dtype: float64

Las personas con diabetes tienen una edad promedio de 37 años, y las que no tienen diabetes promedian 31 años.