

```
In [ ]: ACTIVIDAD = "Visualización y Análisis de Datos – Diabetes"
ALUMNA = "Fabiola Ochoa A01752754"
print(ACTIVIDAD, "-", Fabiola)
```

```
In [2]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

df.head()
```

```
Out[2]:
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunc
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 [3]: # Ver dimensiones, tipos y valores nulos
print("Dimensiones (filas, columnas):", df.shape)
df.info()
df.isnull().sum()
```

```
Dimensiones (filas, columnas): (768, 9)
<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      0
         Glucose          0
         BloodPressure    0
         SkinThickness    0
         Insulin          0
         BMI              0
         DiabetesPedigreeFunction  0
         Age              0
         Outcome          0
         dtype: int64
```

```
In [4]: df.describe()
```

```
Out[4]:
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
count	768.000000	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	0.471471
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331335
min	0.000000	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	0.167000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.331000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.471000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	0.627000

```
In [5]: variables = ["Pregnancies", "DiabetesPedigreeFunction", "Outcome"]
         df_sel = df[variables]
         df_sel.head()
```

```
Out[5]:
```

	Pregnancies	DiabetesPedigreeFunction	Outcome
0	6	0.627	1
1	1	0.351	0
2	8	0.672	1
3	1	0.167	0
4	0	2.288	1

```
In [6]: sns.set(style="whitegrid")

         # Diabetes Pedigree Function por Outcome
         sns.barplot(data=df, x='Outcome', y='DiabetesPedigreeFunction', palette='coolwarm')
         plt.title('Promedio de Diabetes Pedigree Function por Diagnóstico (Outcome)')
         plt.show()

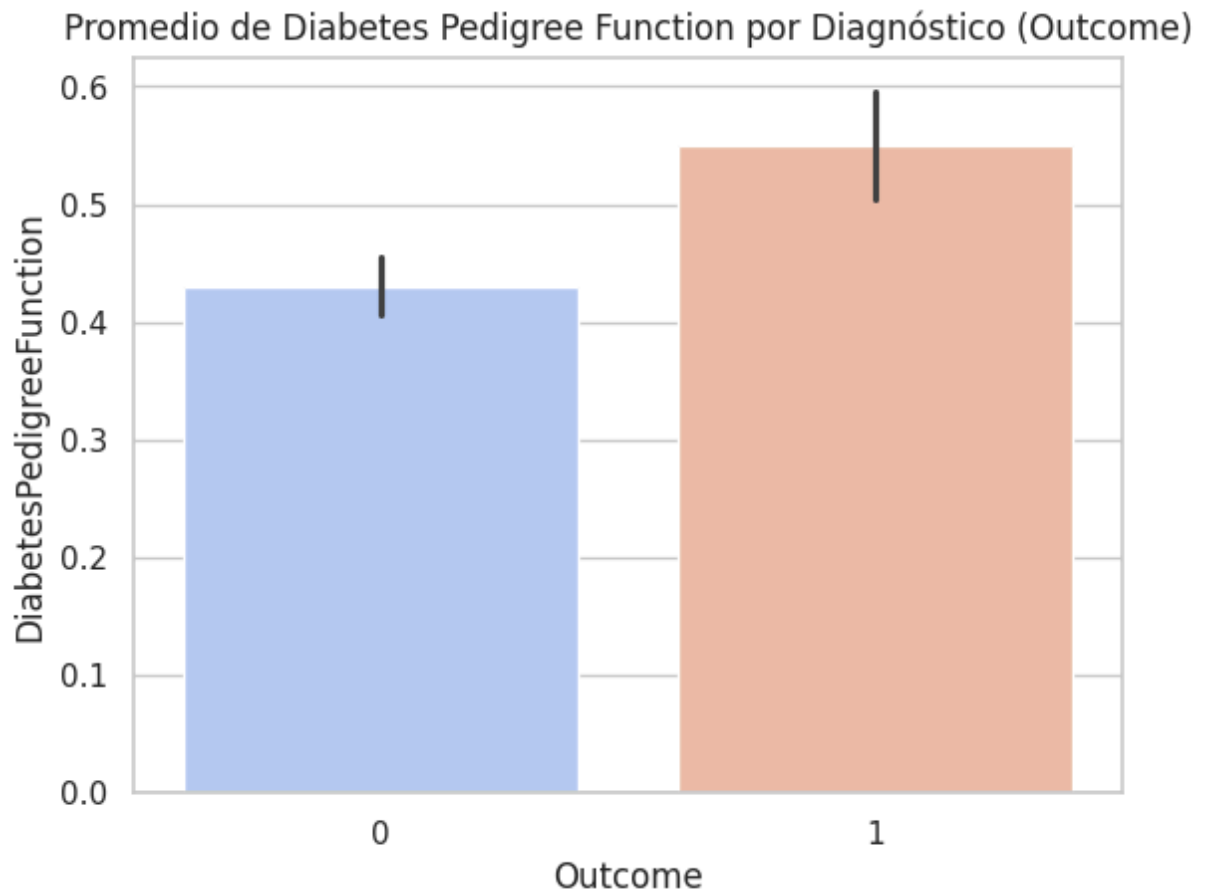
         # Pregnancies por Outcome
```

```
sns.barplot(data=df, x='Outcome', y='Pregnancies', palette='viridis')
plt.title('Promedio de Embarazos por Diagnóstico (Outcome)')
plt.show()
```

/tmp/ipykernel_9149/3848688924.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

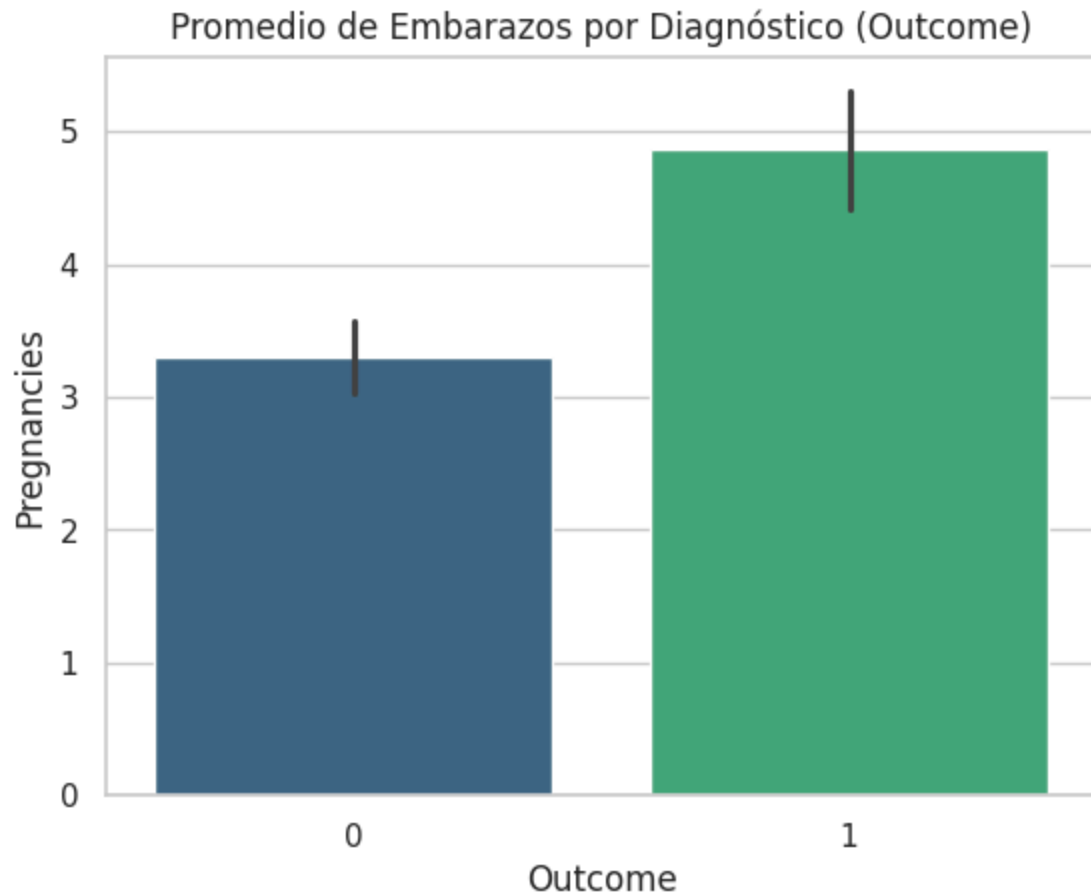
```
sns.barplot(data=df, x='Outcome', y='DiabetesPedigreeFunction', palette='coolwarm')
```



/tmp/ipykernel_9149/3848688924.py:9: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(data=df, x='Outcome', y='Pregnancies', palette='viridis')
```

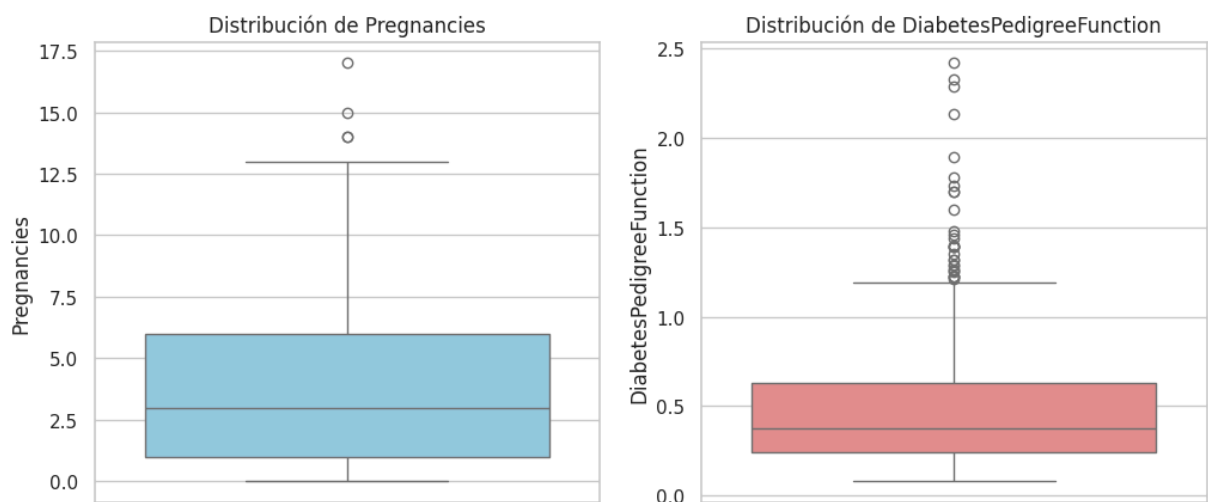


```
In [13]: fig, axes = plt.subplots(1, 2, figsize=(12,5))

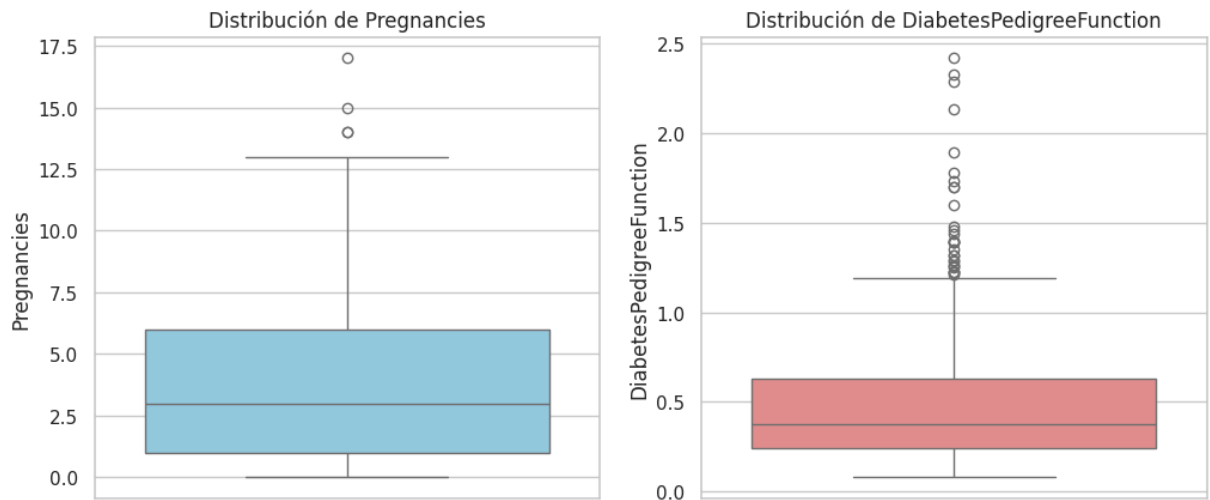
sns.boxplot(data=df, y='Pregnancies', ax=axes[0], color='skyblue')
axes[0].set_title('Distribución de Pregnancies')

sns.boxplot(data=df, y='DiabetesPedigreeFunction', ax=axes[1], color='lightcoral')
axes[1].set_title('Distribución de DiabetesPedigreeFunction')

plt.show()
```



```
In [8]:
```

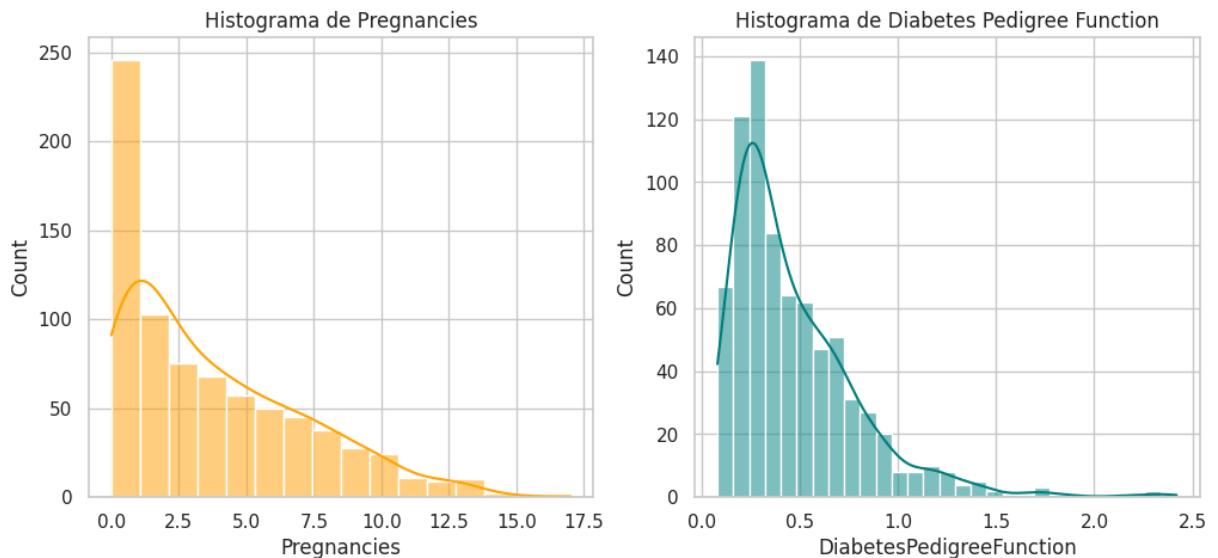


```
In [9]: fig, axes = plt.subplots(1, 2, figsize=(12,5))

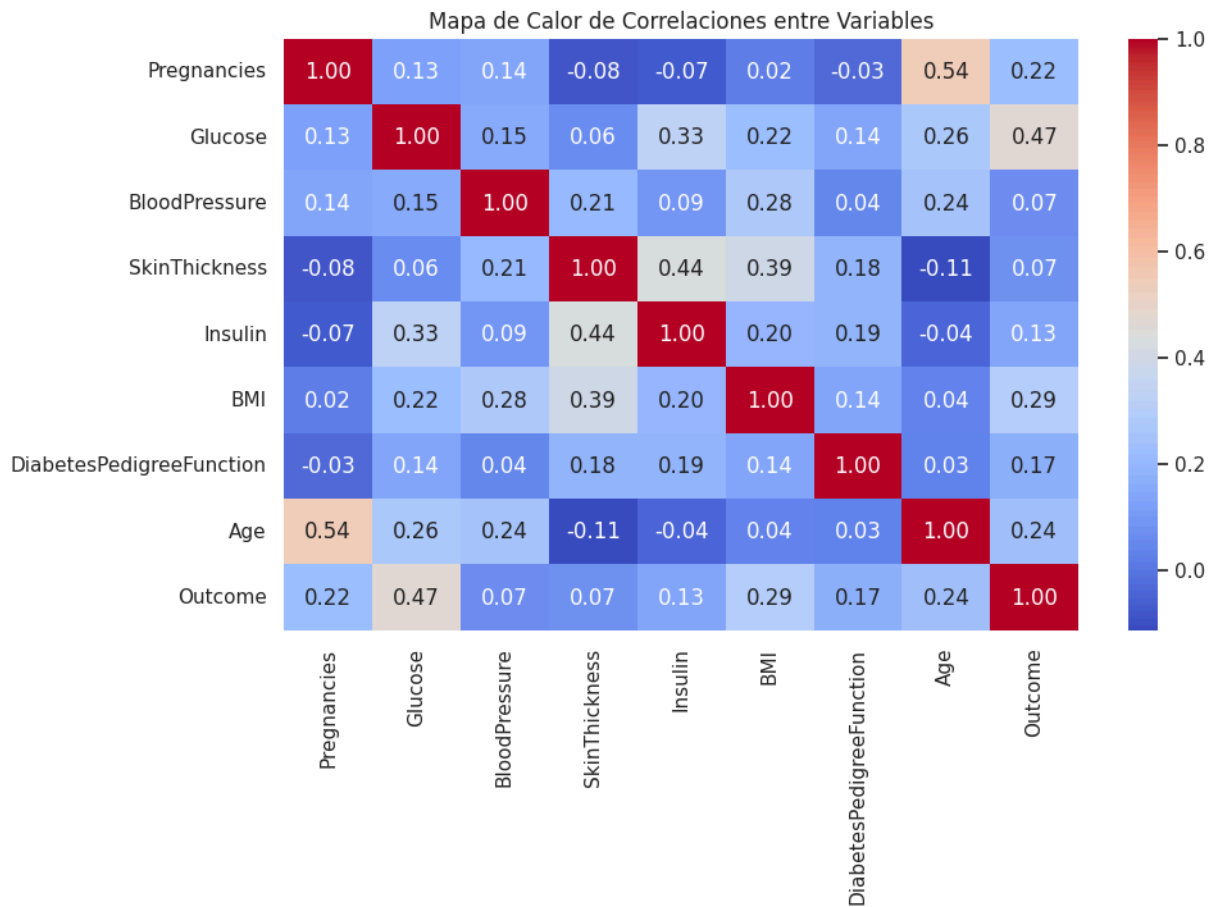
sns.histplot(df['Pregnancies'], kde=True, color='orange', ax=axes[0])
axes[0].set_title('Histograma de Pregnancies')

sns.histplot(df['DiabetesPedigreeFunction'], kde=True, color='teal', ax=axes[1])
axes[1].set_title('Histograma de Diabetes Pedigree Function')

plt.show()
```



```
In [10]: plt.figure(figsize=(10,6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Mapa de Calor de Correlaciones entre Variables')
plt.show()
```



```
In [11]: #Promedio de DiabetesPedigreeFunction por Outcome
print("Promedio de DPF por Outcome:")
print(df.groupby('Outcome')['DiabetesPedigreeFunction'].mean(), "\n")

#Promedio de embarazos por Outcome
print("Promedio de Pregnancies por Outcome:")
print(df.groupby('Outcome')['Pregnancies'].mean(), "\n")

#Casos con muchos embarazos y alto DPF
filtro = df[(df['Pregnancies'] > 10) & (df['DiabetesPedigreeFunction'] > 1)]
print("Casos con >10 embarazos y DPF >1:", len(filtro))
filtro.head()
```


```
Promedio de DPF por Outcome:
Outcome
0    0.429734
1    0.550500
Name: DiabetesPedigreeFunction, dtype: float64
```

```
Promedio de Pregnancies por Outcome:
Outcome
0    3.298000
1    4.865672
Name: Pregnancies, dtype: float64
```

```
Casos con >10 embarazos y DPF >1: 2
```

Out[11]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu
259	11	155	76	28	150	33.3	
744	13	153	88	37	140	40.6	



In [14]: *## Conclusiones del Análisis de Visualización*

****1. ¿Hay alguna variable que no aporta información?****
 Todas las variables que analizamos aportan información relevante aunque, *DiabetesP

****2. Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?****
 Ninguna de las tres, ya que *Pregnancies* y *DiabetesPedigreeFunction* son indicado

****3. Si comparas el rango de las variables (min-max), ¿todas están en rangos simila**
 No,
 - *Pregnancies*: 0-17
 - *DiabetesPedigreeFunction*: 0.08-2.42
 - *Outcome*: 0 o 1
 Es decir, las escalas son diferentes y sería necesario normalizarlas antes de compa

****4. ¿Existen variables que tengan datos atípicos?****
 Sí. *Pregnancies* tiene valores muy altos (más de 10 embarazos) y *DiabetesPedigree

****5. ¿Existe correlación alta entre variables?****
 La correlación es positiva moderada entre *Pregnancies* y *Outcome*,
 lo que indica que un mayor número de embarazos podría asociarse con un mayor riesgo
 La correlación entre *DiabetesPedigreeFunction* y *Outcome* es débil pero también p

Cell In[14], line 3

****1. ¿Hay alguna variable que no aporta información?****

^

SyntaxError: invalid character '¿' (U+00BF)

In []: