

data set link <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('/content/diabetes.csv')
df.head()
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
0	6	148	72	35	0
1	1	85	66	29	0
2	8	183	64	0	0
3	1	89	66	23	94
4	0	137	40	35	168

Double-click (or enter) to edit

```
print("Dataset Info:")
df.info()

print("\nSummary Statistics:")
df.describe()
```



Dataset 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

Summary Statistics:

	Pregnancies	Glucose	BloodPressure	SkinThickness	
count	768.000000	768.000000	768.000000	768.000000	768
mean	3.845052	120.894531	69.105469	20.536458	768
std	3.369578	31.972618	19.355807	15.952218	1161
min	0.000000	0.000000	0.000000	0.000000	0
25%	1.000000	99.000000	62.000000	0.000000	0
50%	3.000000	117.000000	72.000000	23.000000	3
75%	6.000000	140.250000	80.000000	32.000000	12
max	17.000000	199.000000	122.000000	99.000000	84

```

cols_with_zero = ['Glucose', 'BloodPressure', 'SkinThickness']
df[cols_with_zero] = df[cols_with_zero].replace(0, np.nan)

print("Missing Values per Column:")
print(df.isnull().sum())

for col in df.columns:
    df[col].fillna(df[col].median(), inplace=True)

print("\nMissing Values After Cleaning:")
print(df.isnull().sum())

```

```

Missing Values per Column:
Pregnancies          0
Glucose              5
BloodPressure        35
SkinThickness       227
Insulin             374
BMI                 11
DiabetesPedigreeFunction  0
Age                 0
Outcome             0
dtype: int64

```

```

Missing Values After Cleaning:
Pregnancies          0
Glucose              0
BloodPressure        0
SkinThickness        0
Insulin              0
BMI                  0
DiabetesPedigreeFunction  0
Age                  0
Outcome              0
dtype: int64

```

/tmp/ipython-input-79454907.py:10: FutureWarning: A value is
The behavior will change in pandas 3.0. This inplace method w

For example, when doing 'df[col].method(value, inplace=True)'

```
df[col].fillna(df[col].median(), inplace=True)
```

```
def handle_outliers(df, col):
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    df[col] = np.where(df[col] < lower, lower, np.where(df[col] > upper, upper, df[col]))

num_cols = df.select_dtypes(include=['float64', 'int64']).columns

for col in num_cols:
    handle_outliers(df, col)

print("Outliers handled successfully.")
```

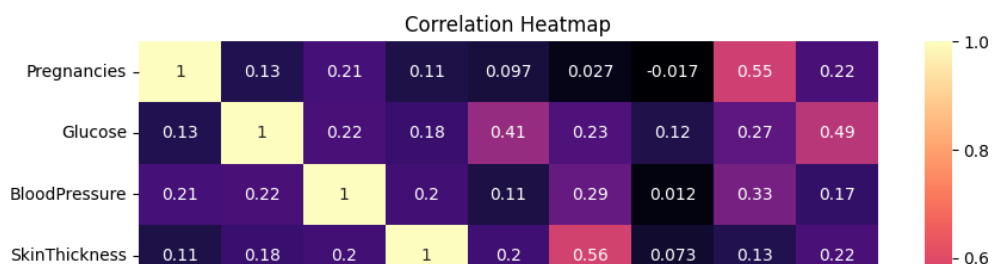
Outliers handled successfully.

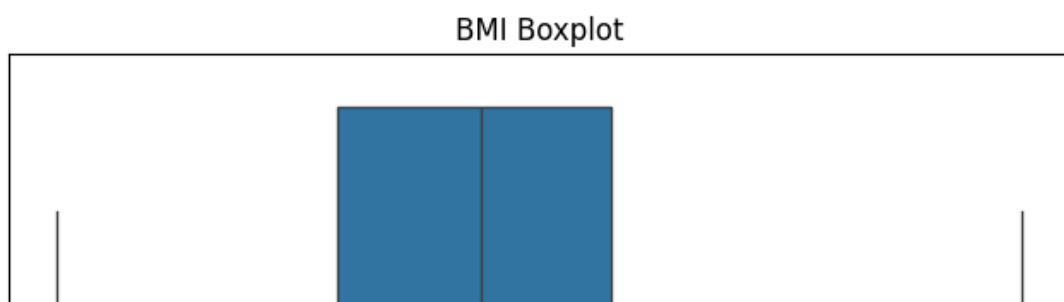
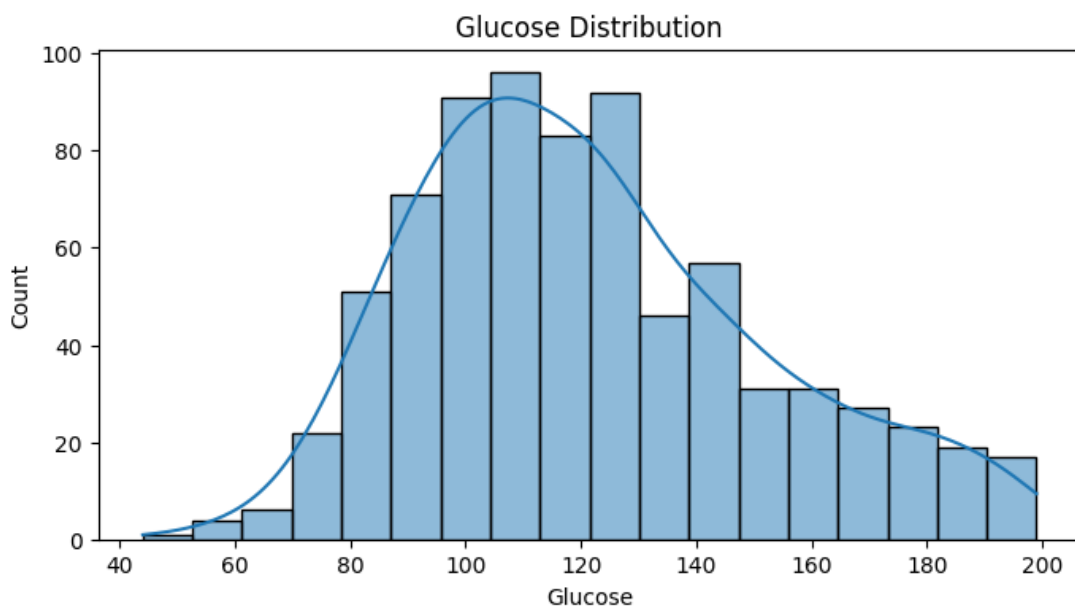
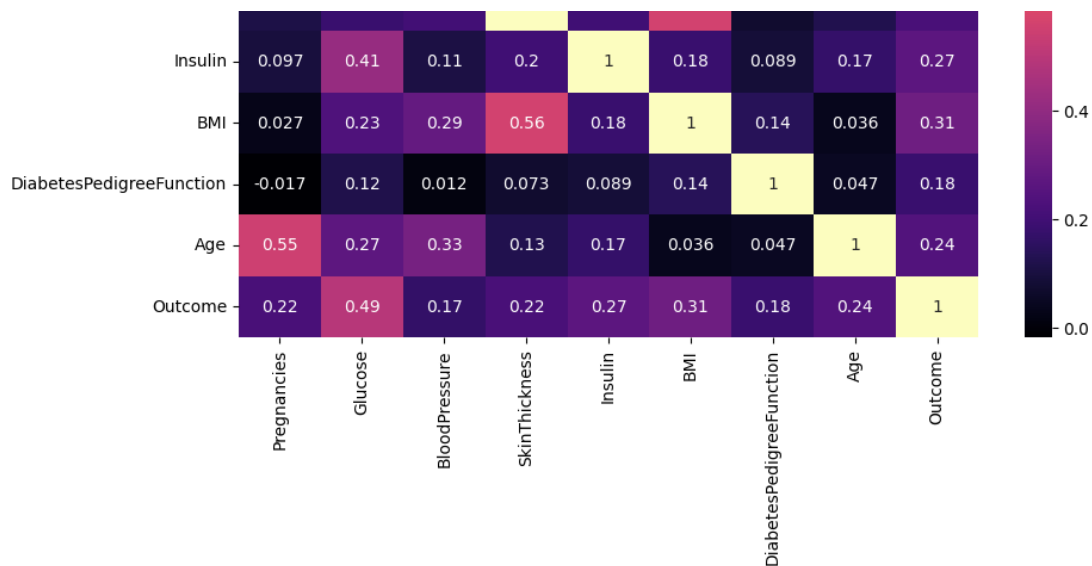
```
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='magma')
plt.title('Correlation Heatmap')
plt.show()
```

```
plt.figure(figsize=(8,4))
sns.histplot(df['Glucose'], kde=True)
plt.title('Glucose Distribution')
plt.show()
```

```
plt.figure(figsize=(8,4))
sns.boxplot(x=df['BMI'])
plt.title('BMI Boxplot')
plt.show()
```

```
plt.figure(figsize=(6,4))
sns.countplot(x='Outcome', data=df, palette='pastel')
plt.title('Outcome Distribution (0 = No Diabetes, 1 = Diabetes)')
plt.show()
```





```
df.to_csv('/content/cleaned_diabetes_dataset.csv', index=False)  
print("Cleaned dataset saved successfully.")
```

Cleaned dataset saved successfully.