DATA ANALYSIS

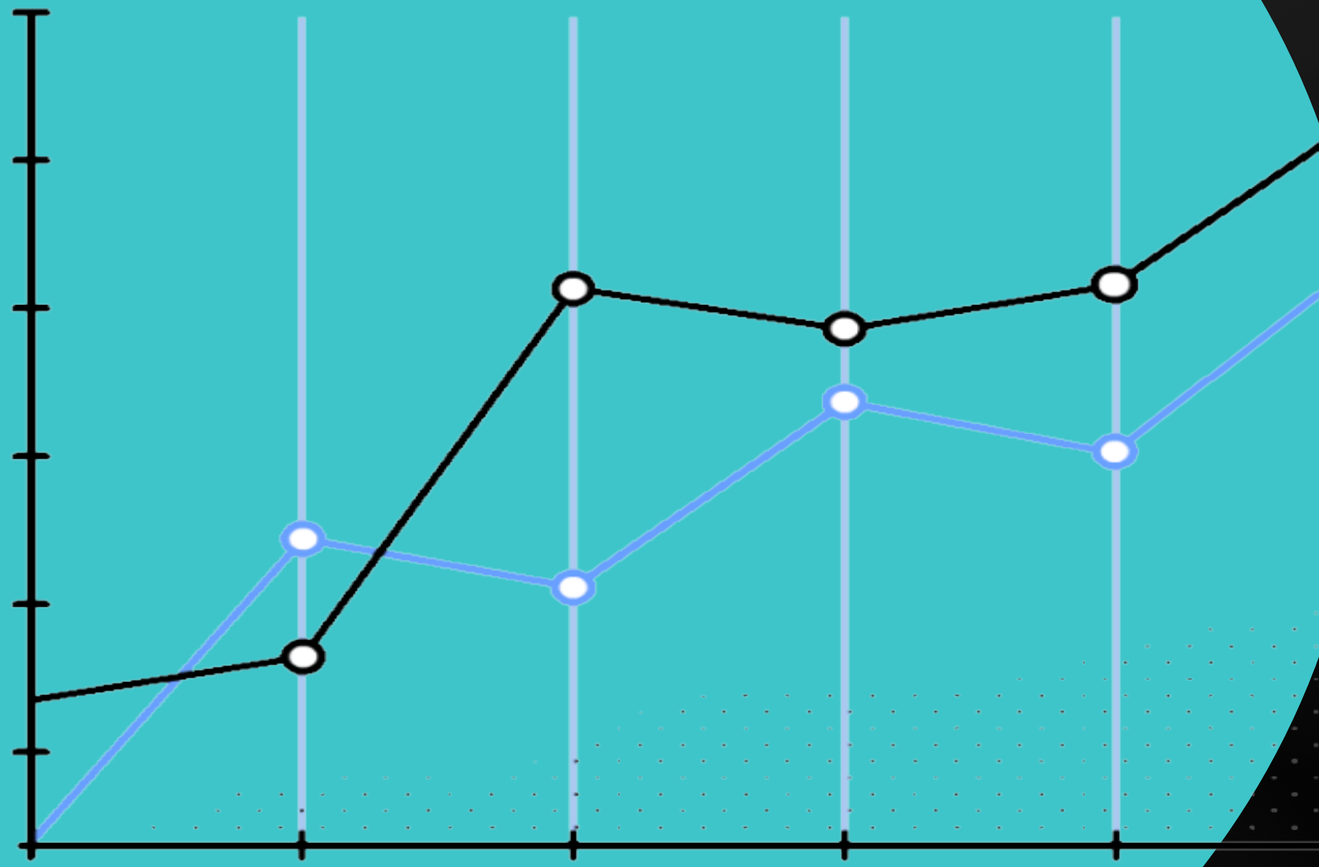
Patients with Diabetes

What is Data Analysis?

Understanding the Basics

Data analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

It involves examining datasets to draw insights that inform decision-making, improve efficiency, and predict future outcomes across various industries.



THE DATA ANALYSIS PROCESS

Data Collection

Data Collection is the systematic process of gathering raw data from various sources to be further analyzed and interpreted.

Data Analysis

Data analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

Cleaning

Data cleansing is the process of identifying and correcting errors, inconsistencies, and inaccuracies within a dataset to ensure its accuracy and reliability for analysis.

Interpretation and Reporting

Interpretation involves attaching meaning and significance to the analyzed data, explaining descriptive patterns, and looking for relationships and linkages among descriptive dimensions

GRAPH USED IN DATA ANALYSIS

- Histogram Plot
- Bar Plot
- Scatter Plot
- Box Plot
- Heat Map

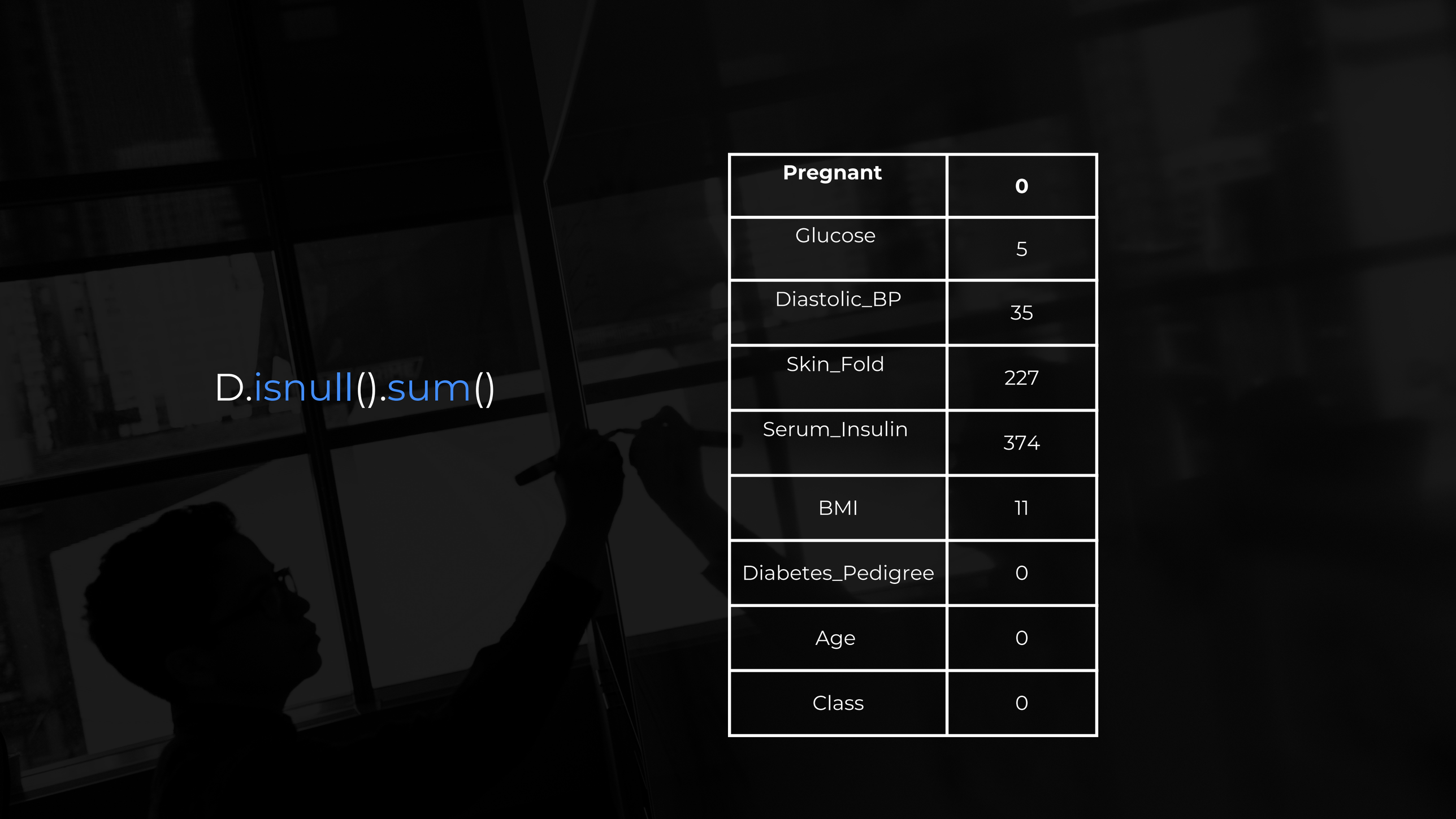
DATA SUMMARY

DATAFRAME	VALUES
Number of rows	768
Number of columns	9

COLUMN TYPES	COUNTS
float64	6
int32	3

COLUMN'S NAME

- GLUCOSE
- PREGNANT
- DIASTOLIC BP
- SKIN FOLD
- SERUM INSULIN
- BMI
- DIABETES PEDIGREE
- AGE
- CLASS



D.isnull().sum()

Pregnant	**0**
Glucose	5
Diastolic_BP	35
Skin_Fold	227
Serum_Insulin	374
BMI	11
Diabetes_Pedigree	0
Age	0
Class	0

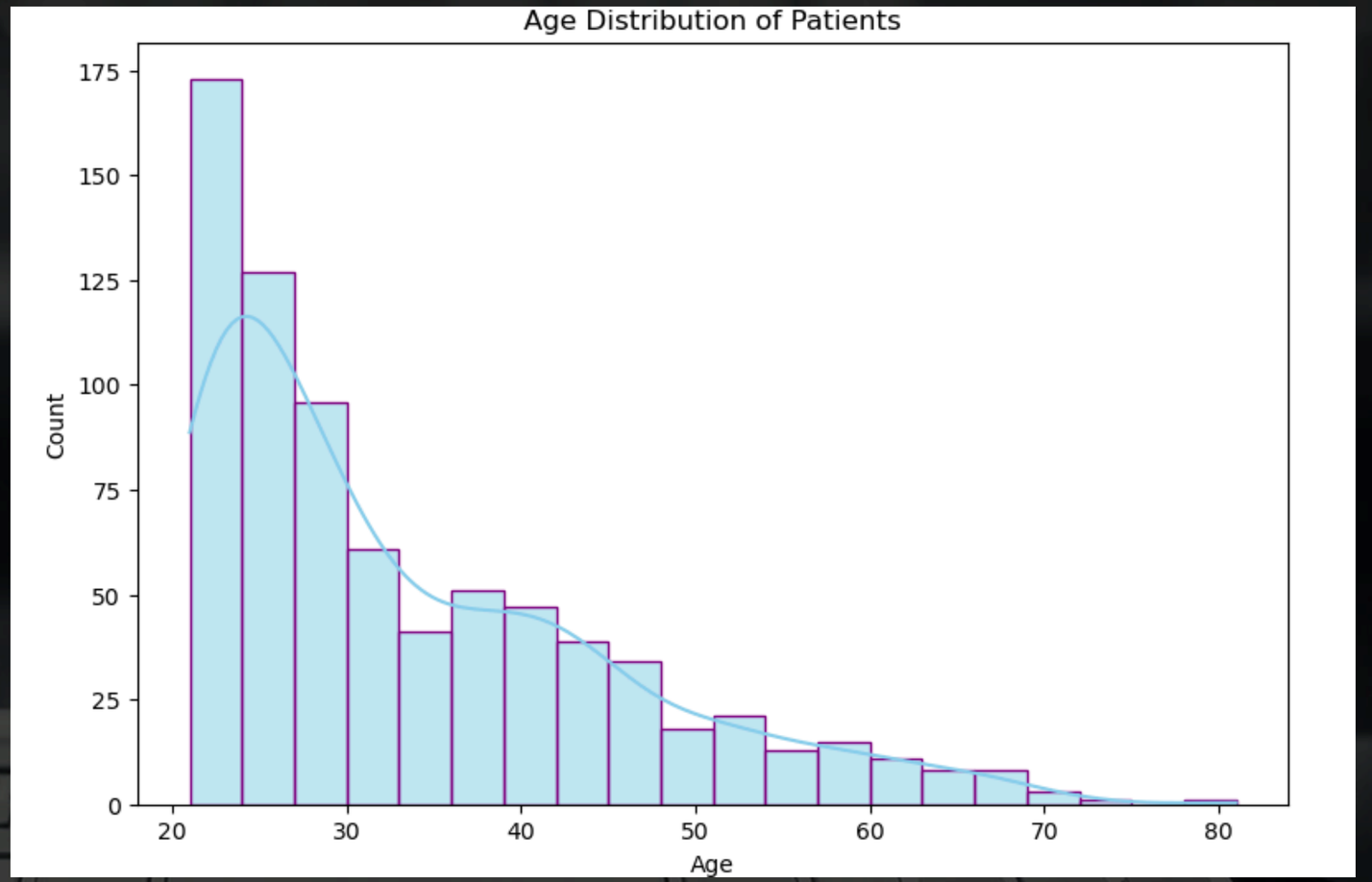
Q1. How many missing values are there in each column, and what is the best way to handle them?

```
# Checking for missing values
    missing_values = D.isnull().sum()
    missing_percentage = (D.isnull().sum() / len(D)) * 100

# Combine results into a DataFrame
    missing_df = pd.DataFrame({"Missing Values": missing_values, "Percentage (%)": missing_percentage})
    print(missing_df)
```

	Missing Values	Percentage (%)
Pregnant	0	0.000000
Glucose	5	0.651042
Diastolic_BP	35	4.557292
Skin_Fold	227	29.557292
Serum_Insulin	374	48.697917
BMI	11	1.432292
Diabetes_Pedigree	0	0.000000
Age	0	0.000000
Class	0	0.000000

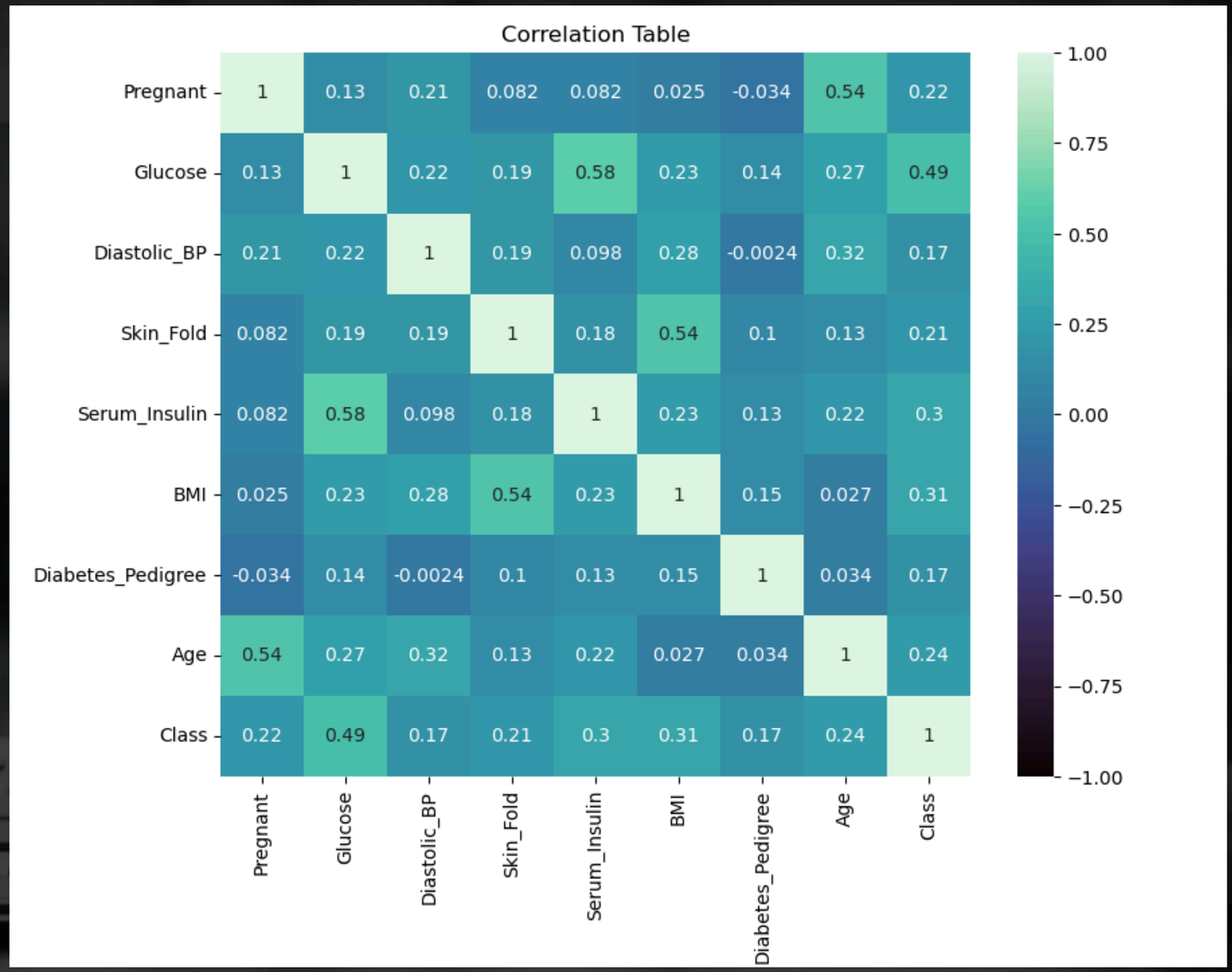
Q2. What is the age distribution of patients in this dataset?

```
plt.figure(figsize=(9,6))
sns.histplot(D["Age"],bins=20,kde=True,color="skyblue",edgecolor="purple")
plt.xlabel("Age")
plt.ylabel("Count")
plt.title("Age Distribution of Patients")
```



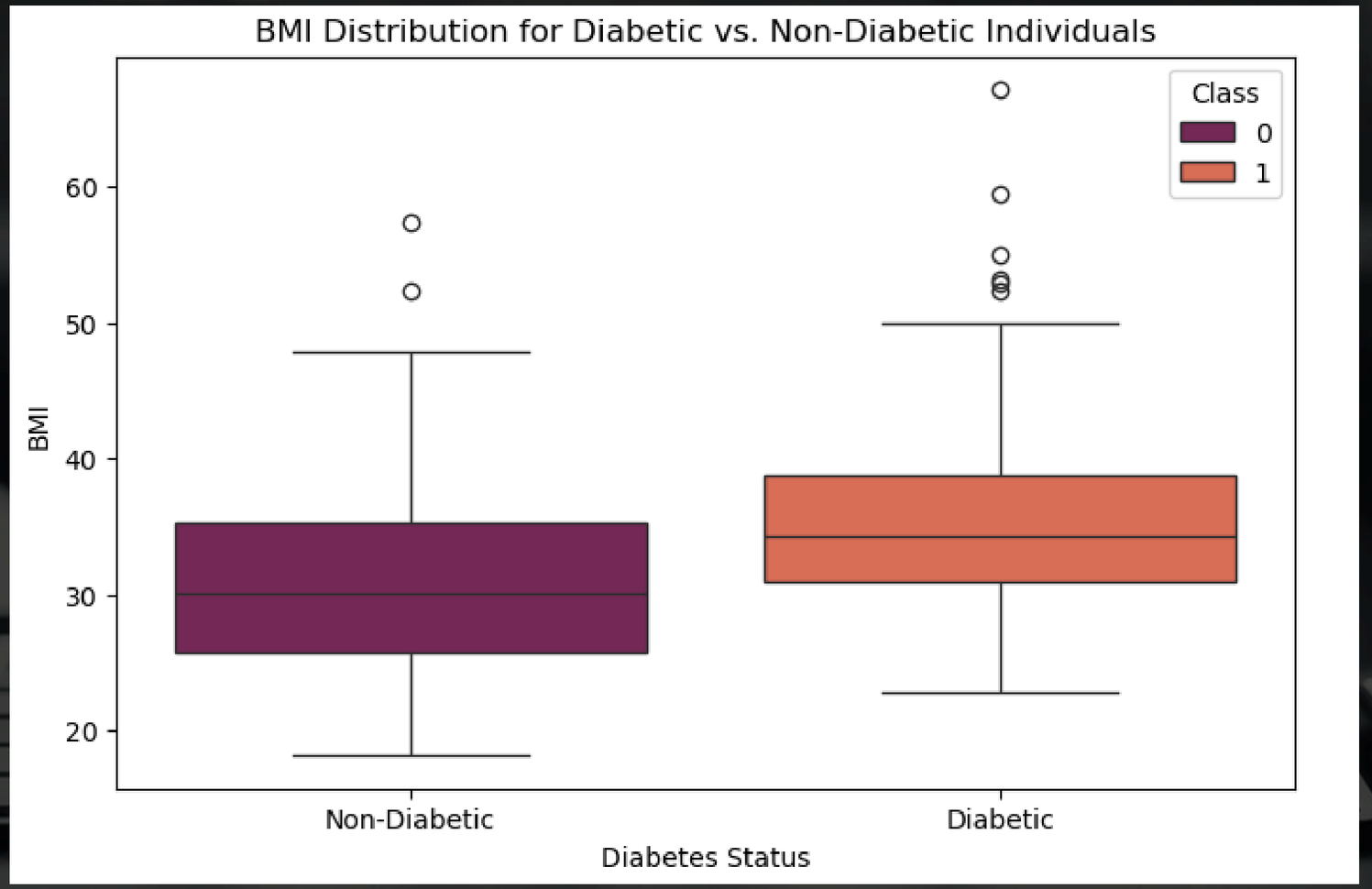
Q3. What is the correlation between Glucose, BMI, and Diabetes occurrence?

```
plt.figure(figsize=(9,7))
sns.heatmap(D.corr(),annot=True,cmap="mako",vmax=1,vmin=-1)
plt.title("Correlation Table")
plt.show()
```



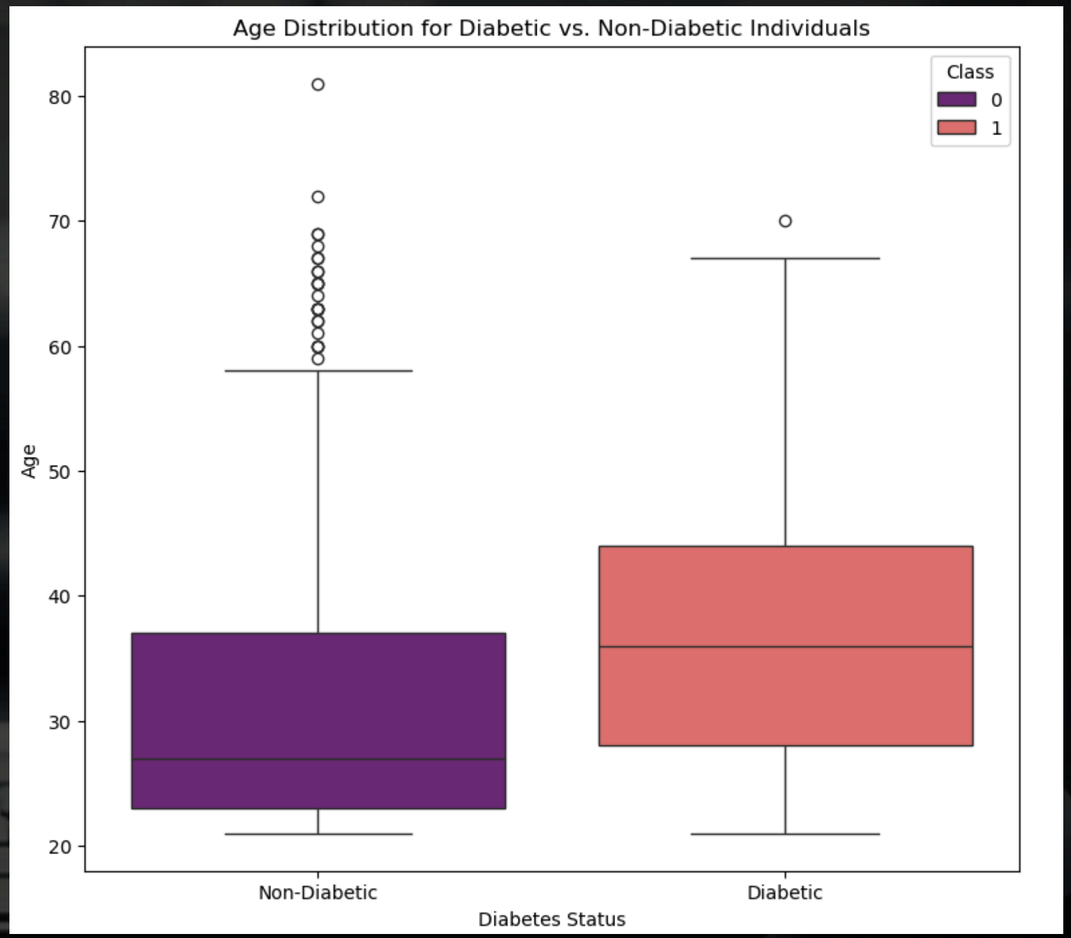
Q4. How does BMI vary among diabetic and non-diabetic individuals?

```
plt.figure(figsize=(8,5))
sns.boxplot(x=D["Class"], y=D["BMI"], hue=D["Class"], palette="rocket", legend=True)
plt.xticks([0,1], ["Non-Diabetic", "Diabetic"])
plt.title("BMI Distribution for Diabetic vs. Non-Diabetic Individuals")
plt.xlabel("Diabetes Status")
plt.ylabel("BMI")
plt.show()
```



Q5. Does age influence the likelihood of diabetes?

```
plt.figure(figsize=(9,8))
sns.boxplot(x=D["Class"],y=D["Age"],hue=D["Class"],palette="magma",legend=True)
plt.xticks([0,1],["Non-Diabetic", "Diabetic"])
plt.xlabel("Diabetes Status")
plt.ylabel("Age")
plt.title("Age Distribution for Diabetic vs. Non-Diabetic Individuals")
plt.show()
```



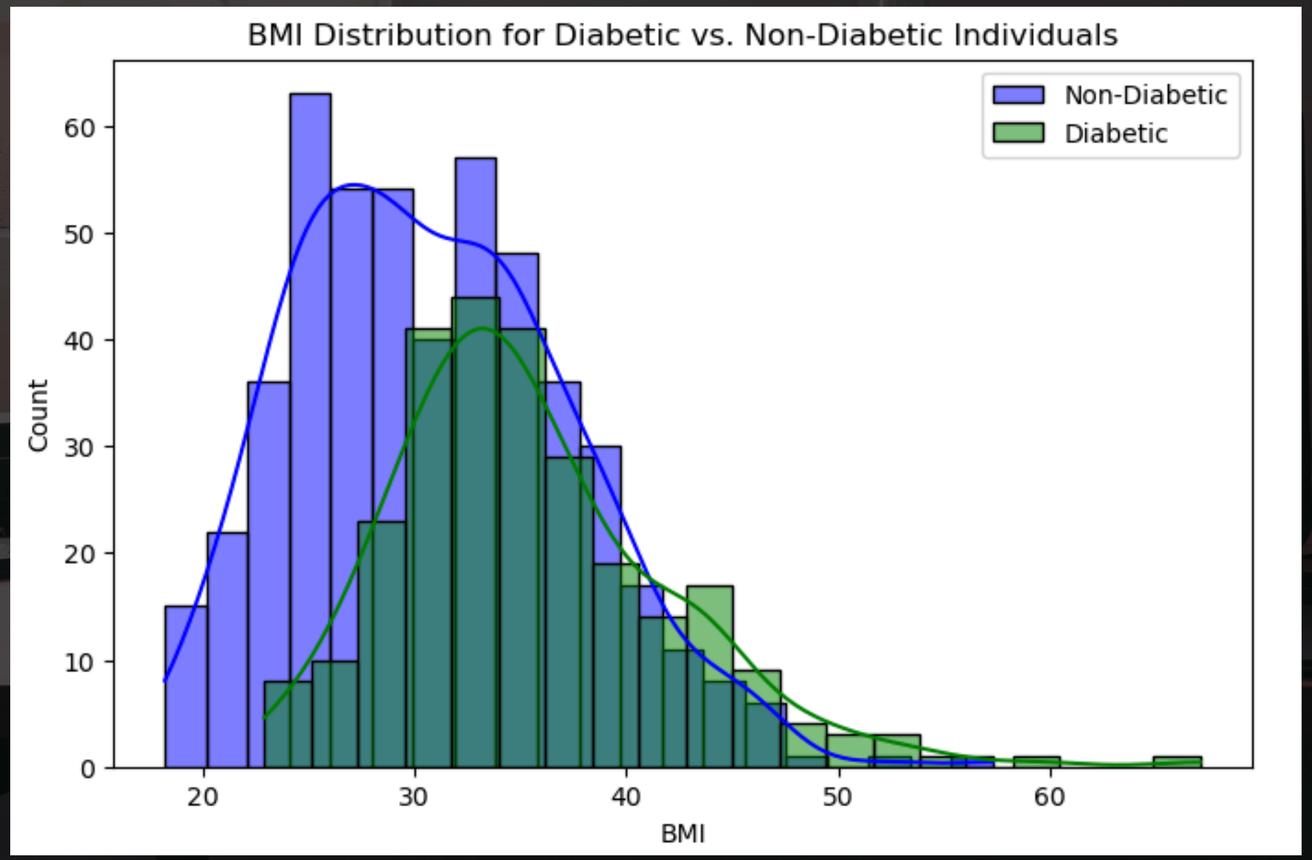
Q6. What is the average glucose level for diabetic vs. non-diabetic

```
D.groupby("Class")["Glucose"].mean()
```

```
Class
0    110.710121
1    142.165573
Name: Glucose, dtype: float64
```

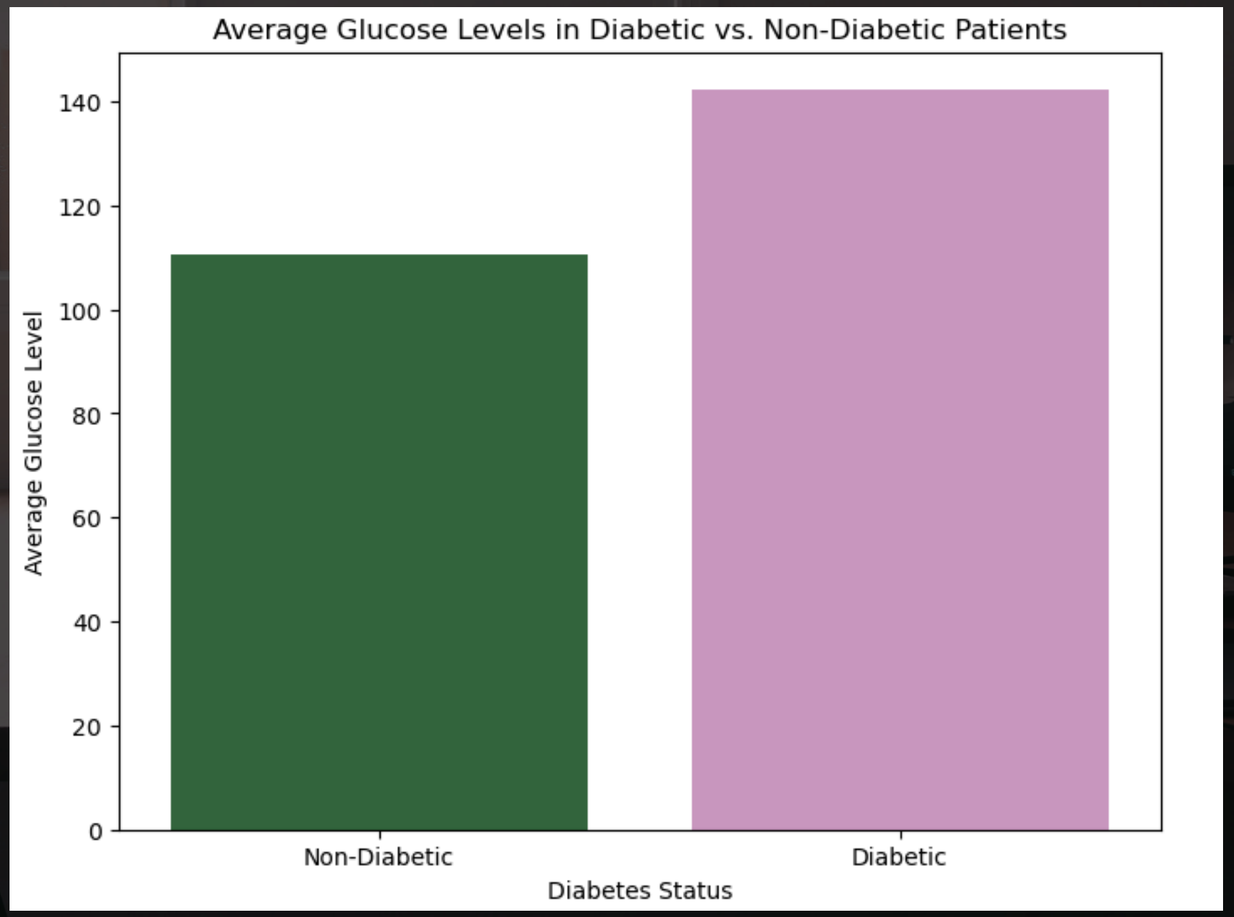
Q7. What are the most common BMI ranges for diabetic and non-diabetic individuals?

```
plt.figure(figsize=(8,5))
sns.histplot(D[D["Class"]==0]["BMI"],bins=20,color="blue",label="Non-Diabetic",kde=True)
sns.histplot(D[D["Class"]==1]["BMI"],bins=20,color="green",label="Diabetic",kde=True)
plt.legend()
plt.title("BMI Distribution for Diabetic vs. Non-Diabetic Individuals")
plt.xlabel("BMI")
plt.ylabel("Count")
plt.show()
```



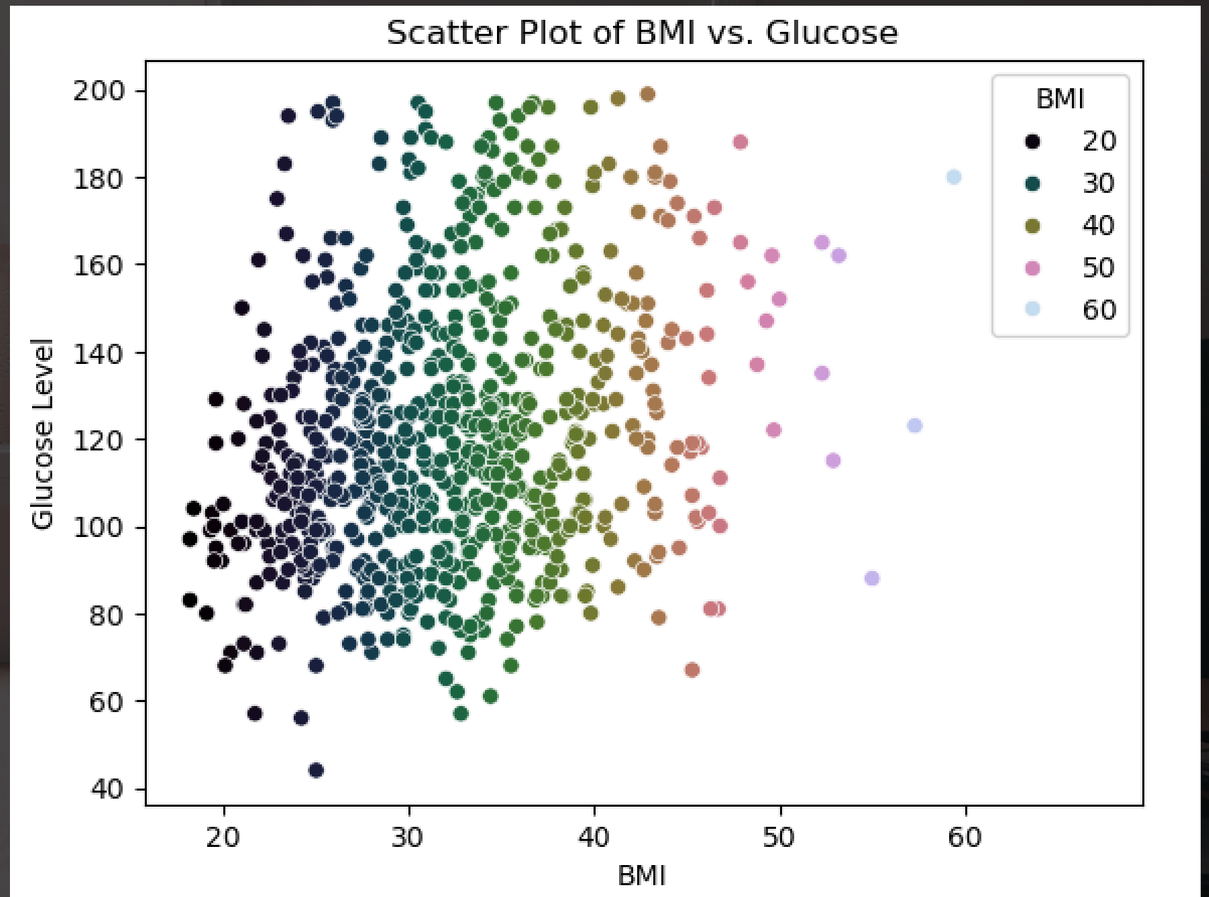
Q8. How do glucose levels vary between diabetic and non-diabetic patients?

```
plt.figure(figsize=(9, 8))
glucose_means = D.groupby("Class")["Glucose"].mean()
sns.barplot(x=glucose_means.index, y=glucose_means.values, palette="cubehelix")
plt.title("Average Glucose Levels in Diabetic vs. Non-Diabetic Patients")
plt.xticks([0, 1], ["Non-Diabetic", "Diabetic"])
plt.xlabel("Diabetes Status")
plt.ylabel("Average Glucose Level")
plt.show()
```



Q9. What is the correlation between BMI and Glucose levels?

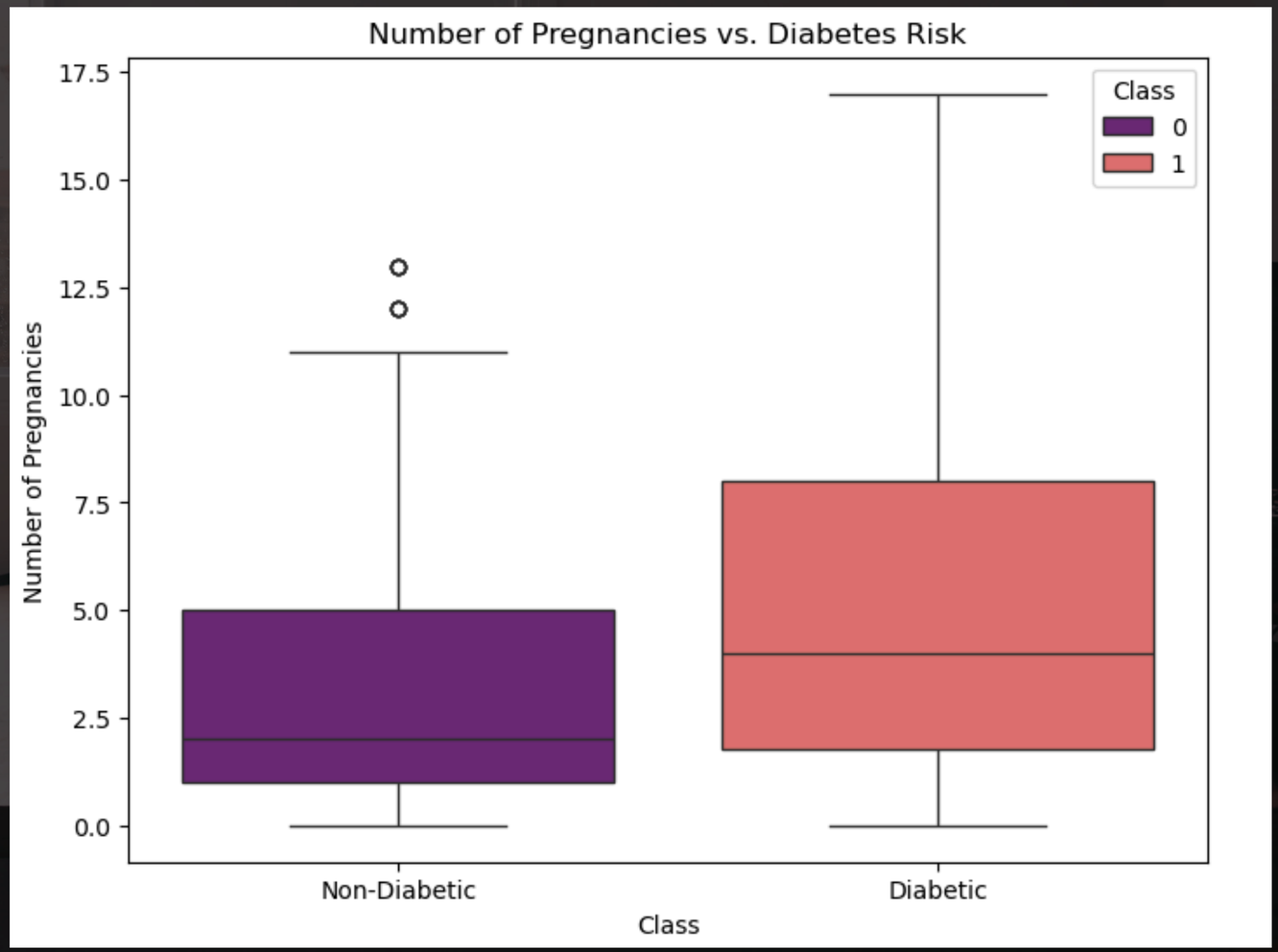
```
sns.scatterplot(x=D["BMI"], y=D["Glucose"],hue=D["BMI"],palette="cubehelix")
plt.title("Scatter Plot of BMI vs. Glucose")
plt.xlabel("BMI")
plt.ylabel("Glucose Level")
plt.show()

D[["BMI", "Glucose"]].corr()
```



	BMI	Glucose
BMI	1.00000	0.23353
Glucose	0.23353	1.00000

Q10. Does pregnancy affect diabetes risk?

```
plt.figure(figsize=(8,6))
sns.boxplot(x=D["Class"], y=D["Pregnant"], hue=D["Class"],palette="magma",legend=True)
plt.xticks([0,1],["Non-Diabetic", "Diabetic"])
plt.xlabel("Class")
plt.ylabel("Number of Pregnancies")
plt.title("Number of Pregnancies vs. Diabetes Risk")
plt.show()
```



CONCLUSION

This project highlights the importance of leveraging data visualization tools for informed decision-making.