3-James, a public health data analyst at a health organization, is investigating the relationship between average glucose levels and the incidence of diabetes. Using a dataset containing medical and health information from individuals, James aims to determine whether individuals with high average glucose levels have a higher likelihood of developing diabetes compared to those with low glucose levels.

```
import pandas as pd
import numpy as np
from scipy.stats import ttest_ind
import matplotlib.pyplot as plt


data = pd.read_csv('./Q 3.csv')

# View first few rows of the dataset
data.head()

# Then select numerical columns for analysis
numerical_data = data.select_dtypes(include=np.number)
numerical_data.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
# Load the dataset
data_url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
columns = [
    "Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin",
    "BMI", "DiabetesPedigreeFunction", "Age", "Outcome"
]
data = pd.read_csv(data_url, header=None, names=columns)

# Divide dataset into high and low glucose groups
median_glucose = data["Glucose"].median()
high_glucose = data[data["Glucose"] > median_glucose]
low_glucose = data[data["Glucose"] <= median_glucose]

# Calculate proportions of diabetes diagnoses in each group
high_diabetes_proportion = high_glucose["Outcome"].mean()
low_diabetes_proportion = low_glucose["Outcome"].mean()

# Perform T-Test to compare mean glucose levels
with_diabetes = data[data["Outcome"] == 1]["Glucose"]
without_diabetes = data[data["Outcome"] == 0]["Glucose"]
t_stat, p_value = ttest_ind(with_diabetes, without_diabetes)

# Print results
print("High Glucose Group - Diabetes Proportion:", high_diabetes_proportion)
print("Low Glucose Group - Diabetes Proportion:", low_diabetes_proportion)
print("T-Statistic:", t_stat)
print("P-Value:", p_value)

if p_value < 0.05:
    print("\nConclusion: Glucose levels are significantly associated with diabetes incidence.")
else:
    print("\nConclusion: Glucose levels are not significantly associated with diabetes incidence.")

# Visualization
# Bar plot for diabetes proportions
plt.bar(["High Glucose", "Low Glucose"], [high_diabetes_proportion, low_diabetes_proportion],
        color=["red", "blue"], alpha=0.7)
plt.title("Proportion of Diabetes Cases by Glucose Group")
plt.ylabel("Proportion of Diabetes Cases")
plt.show()

# Histogram for glucose levels
```

```
plt.hist(with_diabetes, bins=20, alpha=0.7, label="With Diabetes", color="red")
plt.hist(without_diabetes, bins=20, alpha=0.7, label="Without Diabetes", color="blue")
plt.title("Glucose Levels Distribution")
plt.xlabel("Glucose Level")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

```
High Glucose Group - Diabetes Proportion: 0.5411140583554377
Low Glucose Group - Diabetes Proportion: 0.1636828644501279
T-Statistic: 14.600060005973894
P-Value: 8.935431645289912e-43

Conclusion: Glucose levels are significantly associated with diabetes incidence.
```