2-Investigate the Relationship Between Physical Activity and Obesity Levels Mark, a data scientist, is exploring the relationship between physical activity and obesity levels using a synthetic dataset containing data on eating habits, physical activity, and obesity indicators. Mark aims to determine whether individuals with higher physical activity levels have significantly different obesity levels compared to those with lower activity levels.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import norm


data = pd.read_csv('./Q 2.csv')

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

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

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
0	21.0	1.62	64.0	2.0	3.0	2.0	0.0	1.0
1	21.0	1.52	56.0	3.0	3.0	3.0	3.0	0.0
2	23.0	1.80	77.0	2.0	3.0	2.0	2.0	1.0
3	27.0	1.80	87.0	3.0	3.0	2.0	2.0	0.0
4	22.0	1.78	89.8	2.0	1.0	2.0	0.0	0.0

```
# Map NObeyesdad categories to numerical values
obesity_mapping = {
    "Insufficient_Weight": 1,
    "Normal_Weight": 2,
    "Overweight_Level_I": 3,
    "Overweight_Level_II": 4,
    "Obesity_Type_I": 5,
    "Obesity_Type_II": 6,
    "Obesity_Type_III": 7
}
data["Obesity_Score"] = data["NObeyesdad"].map(obesity_mapping)

# Then Group individuals based on CALC
high_activity = data[data["CALC"].isin(["Frequently", "Always"])]
low_activity = data[data["CALC"].isin(["no", "Sometimes"])]

# Calculate mean and standard deviation for both groups
high_mean = high_activity["Obesity_Score"].mean()
high_std = high_activity["Obesity_Score"].std()

low_mean = low_activity["Obesity_Score"].mean()
low_std = low_activity["Obesity_Score"].std()

# Perform Z-Test
n_high = len(high_activity)
n_low = len(low_activity)

# standard errors are calculating
se = np.sqrt((high_std**2 / n_high) + (low_std**2 / n_low))

# Z-Statistic
z_stat = (high_mean - low_mean) / se

# P-Value
p_value = 2 * (1 - norm.cdf(abs(z_stat)))

# Print results
print("High Activity Mean Obesity Level:", high_mean)
print("High Activity Standard Deviation:", high_std)
```

```
print("Low Activity Mean Obesity Level:", low_mean)
print("Low Activity Standard Deviation:", low_std)
print("Z-Statistic:", z_stat)
print("P-Value:", p_value)

if p_value < 0.05:
    print("\nConclusion: Significant difference in obesity levels between high and low physical activity groups.")
else:
    print("\nConclusion: No significant difference in obesity levels between high and low physical activity groups.")

# Visualization
# Bar chart for mean obesity levels
plt.bar(["High Activity", "Low Activity"], [high_mean, low_mean], color=["green", "blue"], alpha=0.7)
plt.title("Average Obesity Levels by Physical Activity Group")
plt.ylabel("Average Obesity Score")
plt.show()

# Histogram for obesity level distribution
plt.hist(high_activity["Obesity_Score"], bins=7, alpha=0.7, label="High Activity", color="green")
plt.hist(low_activity["Obesity_Score"], bins=7, alpha=0.7, label="Low Activity", color="blue")
plt.title("Obesity Level Distribution by Activity Group")
plt.xlabel("Obesity Score")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

```
High Activity Mean Obesity Level: 3.4507042253521125
High Activity Standard Deviation: 1.204618877421001
Low Activity Mean Obesity Level: 4.135294117647059
Low Activity Standard Deviation: 2.003021114921595
Z-Statistic: -4.573615122246789
P-Value: 4.793800836289108e-06

Conclusion: Significant difference in obesity levels between high and low physical activity groups.
```