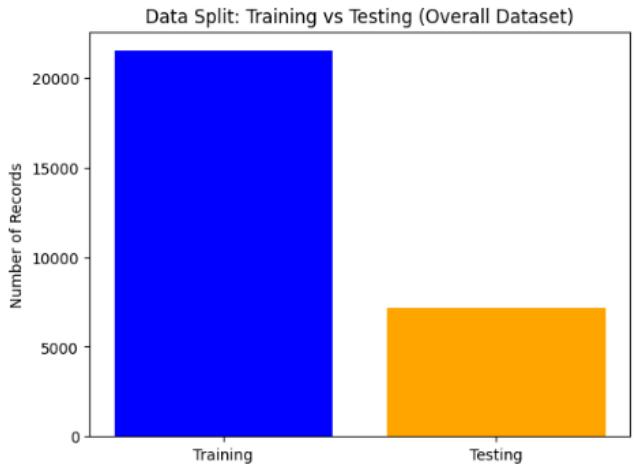
Experiment 3

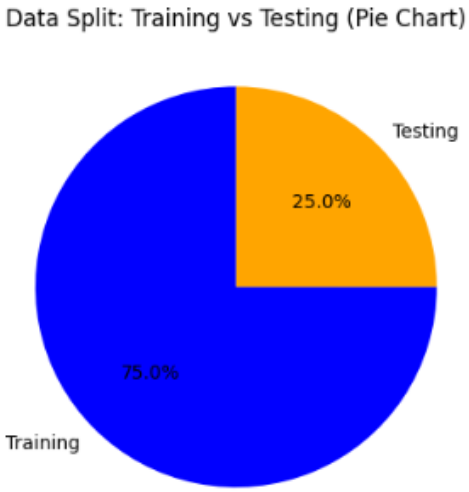
a. Partition the data set, for example 75% of the records are included in the training data set and 25% are included in the test data set.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import scipy.stats as stats

df=pd.read_csv('Recalls_Data.csv')
train_data, test_data = train_test_split(df, test_size=0.25, random_state=42)
labels = ['Training', 'Testing']
sizes = [len(train_data), len(test_data)]
plt.bar(labels, sizes, color=['blue', 'orange'])
plt.title("Data Split: Training vs Testing (Overall Dataset)")
plt.ylabel("Number of Records")
plt.show()
plt.pie(sizes, labels=labels, autopct='%1.1f%%', colors=['blue', 'orange'], startangle=90)
plt.title("Data Split: Training vs Testing (Pie Chart)")
plt.show()
print("Total records in the training data set:", len(train_data))
print("Total records in the testing data set:", len(test_data))
```

b. Use a bar graph and other relevant graph to confirm your proportions.





c. Identify the total number of records in the training data set.

```
Total records in the training data set: 21503
Total records in the testing data set: 7168
```

d. Validate partition by performing a two-sample Z-test.

The Z-test showed that there was no significant difference between the training and test datasets, as the p-value was greater than 0.05. This confirmed that the partitioning process was unbiased.