Experiment 3

Aim: Perform Data Modeling on dataset.

Theory:

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.**

In this experiment, we partitioned a dataset of vehicle recall data into a training set (75%) and a test set (25%) and validated this partitioning using a two-sample Z-test. The dataset consists of 28,671 records with features such as manufacturer, recall components, and the number of potentially affected vehicles.

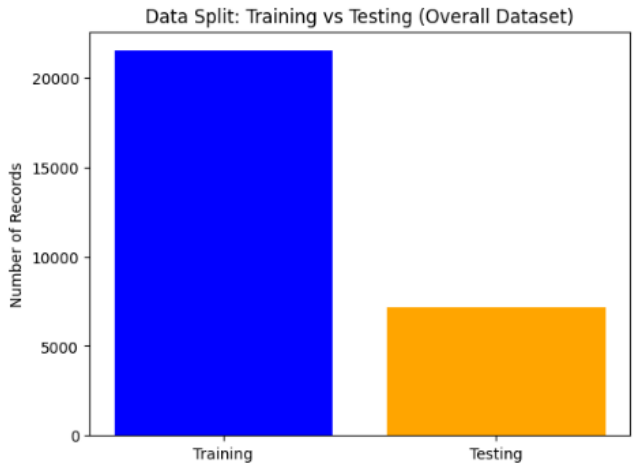
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import scipy.stats as stats

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))
```

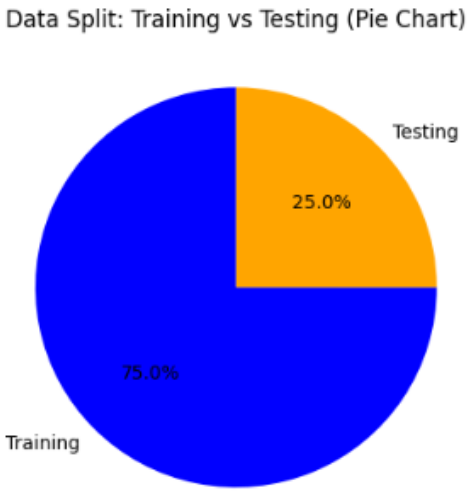
b. **Use a bar graph and other relevant graph to confirm your proportions.**

To confirm the proportions of the data split into training and test sets, we have used a bar graph and a pie chart:

Bar Graph: It displays the number of records in the training and test sets, allowing us to visually confirm the 75%-25% split. The x-axis represents the two sets, and the y-axis shows their respective record counts.

Pie Chart: This chart visually illustrates the percentage split between the training (75%) and test (25%) sets, providing an easy confirmation of the partition.





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

We used the train_test_split method to split the dataset into training and test sets. The partition was visually confirmed through a bar plot, showing the expected 75%-25% split, with 21,503 records in the training set and 7,168 in the test 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.

To validate the partitioning, we performed a two-sample Z-test manually to compare the "Potentially Affected" values between the training and test datasets. The Z-statistic was calculated based on the means, standard deviations, and sizes of both samples. We then compared the calculated p-value to the significance level of 0.05.

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.

Conclusion:

The partitioning of the dataset into training and test sets was validated successfully. The Z-test confirmed that there was no significant difference between the datasets, making the partition reliable for further analysis.