

Aim: Perform Data Modeling.

- 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.
- Use a bar graph and other relevant graphs to confirm your proportions.
- Identify the total number of records in the training data set.
- Validate partition by performing a two-sample Z-test.

Steps:

Step 1: 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.

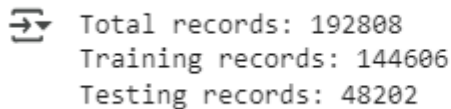
We split the dataset into two parts: 75% for training (to teach the model) and 25% for testing (to check its accuracy). This ensures the model learns from most data while being tested on unseen data.

Code:

```
from sklearn.model_selection import train_test_split

train_df, test_df = train_test_split(df, test_size=0.25, random_state=42)
print(f"Total records: {len(df)}")
print(f"Training records: {len(train_df)}")
print(f"Testing records: {len(test_df)}")
```

Result:



```
⇒ Total records: 192808
   Training records: 144606
   Testing records: 48202
```

Step 2: Use a bar graph and other relevant graphs to confirm your proportions.

Graphs help us confirm that the data is split correctly. We use bar and pie charts to show the proportion of training and testing data clearly.

Code:

```
import matplotlib.pyplot as plt
import seaborn as sns

total = len(df)
sizes = [len(train_df) / total * 100, len(test_df) / total * 100]
labels = ['Training Set', 'Testing Set']

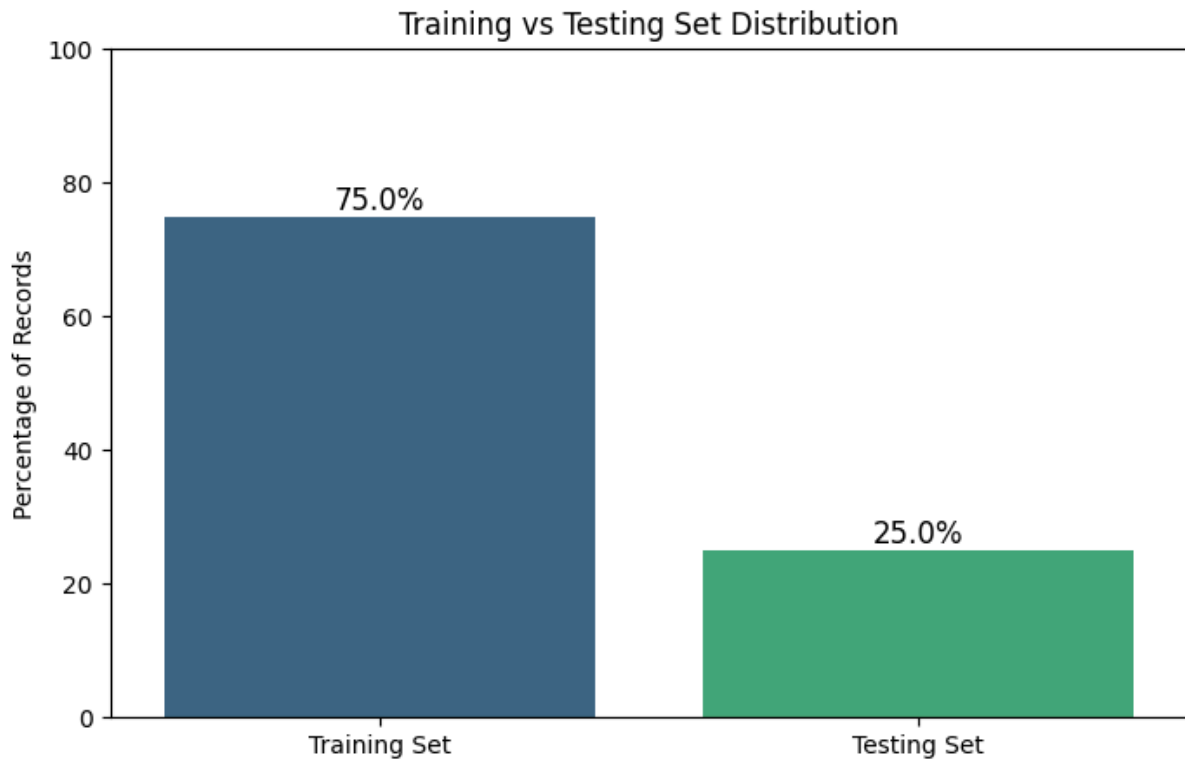
plt.figure(figsize=(8, 5))
sns.barplot(x=labels, y=sizes, palette="viridis")

for i, v in enumerate(sizes):
    plt.text(i, v + 1, f"{v:.1f}%", ha='center', fontsize=12)

plt.ylabel("Percentage of Records")
```

```
plt.title("Training vs Testing Set Distribution")
plt.ylim(0, 100) # Ensure y-axis goes from 0 to 100%
plt.show()
```

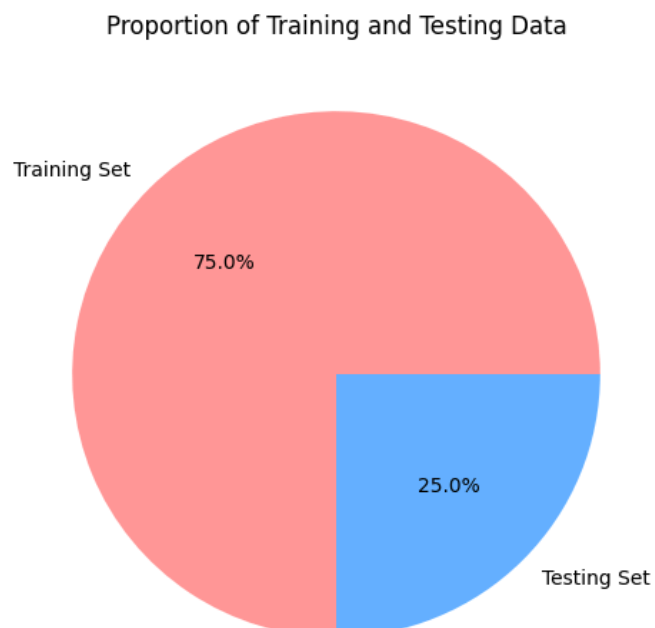
Result:



Code:

```
plt.figure(figsize=(6,6))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', colors=['#ff9999','#66b3ff'])
plt.title("Proportion of Training and Testing Data")
plt.show()
```

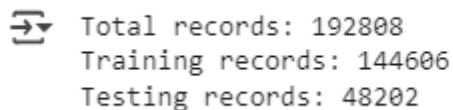
Result:



Step 3: Identify the total number of records in the training data set.**Code:**

```
from sklearn.model_selection import train_test_split

train_df, test_df = train_test_split(df, test_size=0.25, random_state=42)
print(f"Total records: {len(df)}")
print(f"Training records: {len(train_df)}")
print(f"Testing records: {len(test_df)}")
```

Result:

```
➞ Total records: 192808
   Training records: 144606
   Testing records: 48202
```

Step 4: Validate partition by performing a two-sample Z-test.

A two-sample Z-test checks if the training and testing datasets have similar characteristics. It compares the mean values to ensure the split is balanced and doesn't introduce bias.

Code:

```
import numpy as np
from scipy.stats import norm

train_values = train_df["Data_Value"]
test_values = test_df["Data_Value"]

mean_train = np.mean(train_values)
mean_test = np.mean(test_values)
std_train = np.std(train_values, ddof=1)
std_test = np.std(test_values, ddof=1)

n_train = len(train_values)
n_test = len(test_values)

z_score = (mean_train - mean_test) / np.sqrt((std_train**2 / n_train) +
                                              (std_test**2 / n_test))

p_value = 2 * (1 - norm.cdf(abs(z_score)))

print(f"Z-score: {z_score:.4f}")
print(f"P-value: {p_value:.4f}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The means are significantly different.")
else:
    print("Fail to reject the null hypothesis: No significant difference in means.")
```

Result:



Z-score: -1.0861

P-value: 0.2774

Fail to reject the null hypothesis: No significant difference in means.

Conclusion :

In this experiment, we successfully partitioned the dataset into training and testing sets, ensuring a balanced distribution. We used graphs to visualize the split and performed a Z-test to confirm there is no significant difference in data characteristics. The results showed that the partitioning is valid, meaning the dataset is well-prepared for modeling.