Experiment 3

Aim -

Perform data Data Modeling.

Todo -

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

Dataset -

The dataset contains data about customers' purchases during the Black Friday sale. This dataset was taken from Kaggle. The dataset has 550k rows and 12 columns. The various columns of the dataset are age, marital status, gender, total purchase amount, and many other features.

Theory -

Data partitioning:

In many large-scale solutions, data is divided into partitions that can be managed and accessed separately. Partitioning can improve scalability, reduce contention, and optimize performance. It can also provide a mechanism for dividing data by usage pattern. For example, you can archive older data in cheaper data storage. However, the partitioning strategy must be chosen carefully to maximize the benefits while minimizing adverse effects.

Partitioning a data set is splitting the data into two, sometimes three smaller data sets. These are called Training, Validation, and Test. This technique is best practice when creating a predictive model but is only possible when working with enough data. Test data sets are less common due to the volume of data required. If a predictive model is created to fit a specific data set, it is possible to create a highly predictive model. To ensure that this model will predict new data well, it should be tested on a different sample of data to see how accurate it is. Data partitioning is used to split the original data set before the model is created so that there is ‘new’ data available to assess the model.

The three data sets that can be used are described:

● **Training:** The subset of data used to explore the data's characteristics and create a model.

● **Validation:** Data that remains unseen when building the model. It is used to tune the model parameter estimates.

● **Test:** A data set that can measure overall model performance and compare the performance between different candidate models.

Why partition data?

● Improve scalability:

When you scale up a single database system, it will eventually reach a physical hardware limit. If you divide data across multiple partitions, each hosted on a separate server, you can scale out the system almost indefinitely.

● Improve performance:

Data access operations on each partition take place over a smaller volume of data. Correctly done, partitioning can make your system more efficient.

● Improve security:

In some cases, you can separate sensitive and nonsensitive data into different partitions and apply different security controls to the sensitive data.

● Provide operational flexibility:

Partitioning offers many opportunities for fine-tuning operations, maximizing administrative efficiency, and minimizing cost

● Match the data store to the pattern of use:

Partitioning allows each partition to be deployed on a different type of data store, based on cost and the built-in features that the data store offers. For example, large binary data can be stored in blob storage, while more structured data can be held in a document database.

● Improve availability:

Separating data across multiple servers avoids a single point of failure. If one instance fails, only the data in that partition is unavailable. Operations on other partitions can continue.

Hypothesis testing

Hypothesis testing is an act in statistics whereby an assumption regarding a population parameter is tested. The methodology to be used depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

1. **Null Hypothesis:**

The null hypothesis is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value. We either reject or fail to reject the null hypothesis. The null hypothesis is denoted by H0.

2. **Alternate Hypothesis:**

The alternative hypothesis is the statement that the parameter has a value that is different from the claimed value. It is denoted by HA.

3. **Level of significance:**

It means the degree of significance in which we accept or reject the null hypothesis. Since in most of the experiments 100% accuracy is not possible for accepting or rejecting a hypothesis, we, therefore, select a level of significance. It is denoted by alpha ($\propto$).

How Hypothesis Testing Works?

In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to try two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population means the return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

All hypotheses are tested using a four-step process:

i) The first step is for the analyst to state the two hypotheses so that only one can be right.

ii) The next step is to formulate an analysis plan, which outlines how the data will be evaluated.

iii) The third step is to carry out the plan and physically analyze the sample data.

iv) The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

Results -

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.

```
from sklearn.model_selection import train_test_split
print("Number of rows and columns = "+str(df.shape))
```

```
Number of rows and columns = (550068, 13)
```

In [41]:
```
y = df['Purchase']
y.head()
```

Out[41]:
```
0    0.348992
1    0.634181
2    0.058875
3    0.043634
4    0.332248
Name: Purchase, dtype: float64
```

In [42]:
```
x =  df.drop('Purchase',axis=1)
x.head()
```

Out[42]:

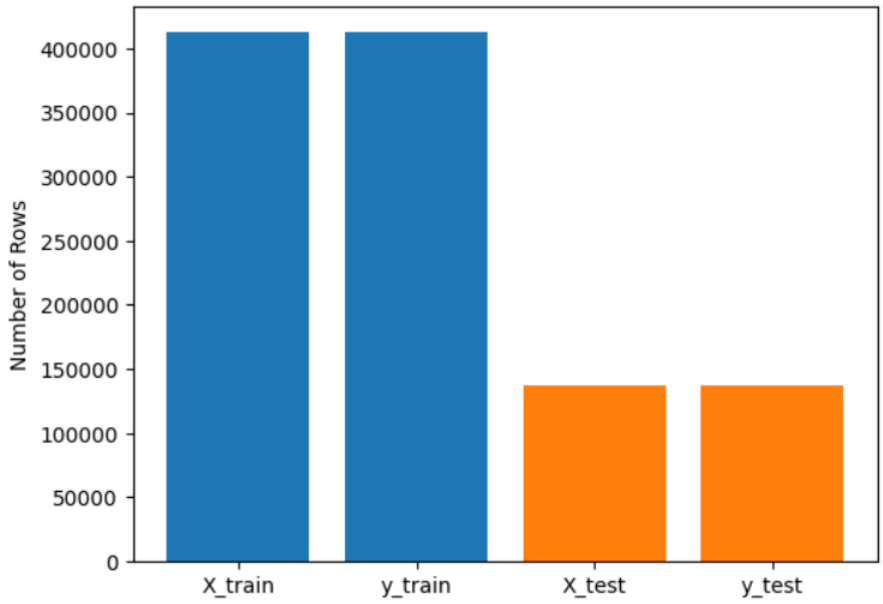
	Product_ID	Age	Occupation	Stay_In_Current_City_Years	Product_Category_1	Married	Not_Married	Female	Male	City_Category_A	City_Category_B	City_
0	P00069042	0-17	10	2	3	1	0	1	0	1	0	
1	P00248942	0-17	10	2	1	1	0	1	0	1	0	
2	P00087842	0-17	10	2	12	1	0	1	0	1	0	
3	P00085442	0-17	10	2	12	1	0	1	0	1	0	
4	P00285442	55+	16	4+	8	1	0	0	1	0	0	

In [43]:
```
X_train, X_test, y_train, y_test = train_test_split(x, y, train_size=0.75, random_state=42)
```

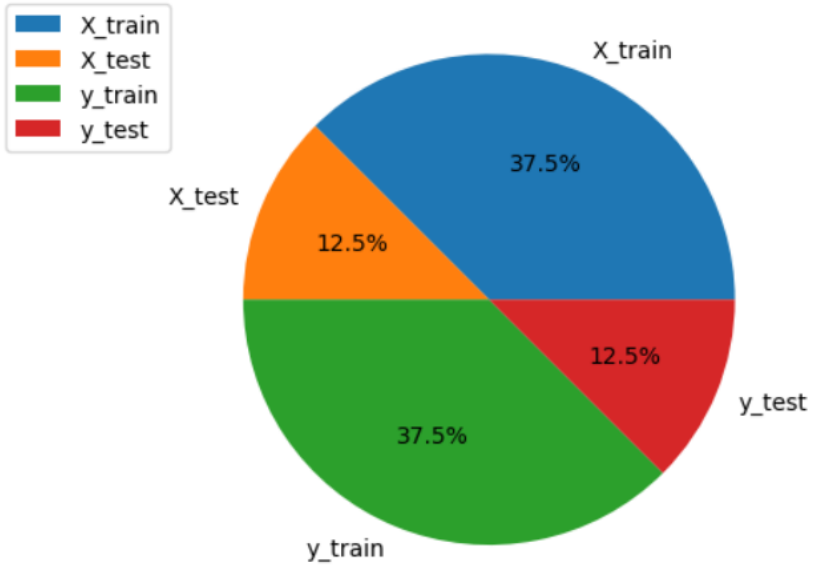
2. Use a bar graph and other relevant graphs to confirm your proportions.

```
In [44]: plt.bar(height = [X_train.shape[0],y_train.shape[0],], x=['X_train','y_train'])
         plt.bar(height = [X_test.shape[0],y_test.shape[0]], x=['X_test','y_test'])
         plt.ylabel('Number of Rows')
```

```
Out[44]: Text(0, 0.5, 'Number of Rows')
```



```
In [45]: plt.pie([X_train.shape[0],X_test.shape[0],y_train.shape[0],y_test.shape[0]],
                 labels=['X_train','X_test','y_train','y_test'], autopct='%1.1f%%')
         plt.legend(bbox_to_anchor=(0,1))
         plt.show()
```



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

```
In [46]: print("X train size: "+str(X_train.shape))
         print("X test size: "+str(X_test.shape))
         print("y train size: "+str(y_train.shape))
         print("y test size: "+str(y_test.shape))
```

```
X train size: (412551, 12)
X test size: (137517, 12)
y train size: (412551,)
y test size: (137517,)
```

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

In [47]:
```
from statsmodels.stats.weightstats import ztest
```

In [48]:
```
val1, val2 = ztest(y_train, y_test, value=0)
print(val1)
print(val2)
```

```
-0.11981714872728104
0.9046279970888207
```

In [49]:
```
if(val2>=0.05):
    print("Null hypothesis accepted.")
else:
    print("Alternative hypothesis accepted.")
```

```
Null hypothesis accepted.
```

Conclusion-

We have successfully performed data modeling. We split our data into training and testing data. Also, we used different visualization and z-test for the analysis of the data.