# PRODUCT DEMAND PREDICTION WITH MACHINE LEARNING

## Phase 3 : Document submission

**Name : Sudharson R**
**NMid : au723721205045**

## DATA PREPROCESSING

### IMPORT LIBRARY:

Import the required libraries such as PANDAS , NUMPY, MATPLOTLIB, and Sklearn

**NumPy :** It is used for mathematical and numerical functions

**PANDAS :** It is a analysis tool used for data manipulation

**MATPLOTLIB :** It is used for data visualization

```
#Importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
```

**SKLEARN :** To Implement Machine Learning model and Statistical Modeling

### LOADING THE DATA SET :

```
#Loading Dataset
dataset = pd.read_csv("PoductDemand.csv")
```

Identify the shape of the dataset...

```
#Shape of the dataset
print("Shape:",dataset.shape)
```

```
Shape: (150150, 5)
```

The dataset contains 150150 rows and 5 columns

To check the head and tail of the dataset use the following

- dataset.head(contains the no. of rows)

- dataset.tail(contains the no. of rows)

This shows the first 6 rows of the dataset.

```
dataset.head(6)
```

| | ID | Store ID | Total Price | Base Price | Units Sold |
|---|---|---|---|---|---|
| 0 | 1 | 8091 | 99.0375 | 111.8625 | 20 |
| 1 | 2 | 8091 | 99.0375 | 99.0375 | 28 |
| 2 | 3 | 8091 | 133.9500 | 133.9500 | 19 |
| 3 | 4 | 8091 | 133.9500 | 133.9500 | 44 |
| 4 | 5 | 8091 | 141.0750 | 141.0750 | 52 |
| 5 | 9 | 8091 | 227.2875 | 227.2875 | 18 |

This shows the last 6 rows of the dataset

```
dataset.tail(6)
```

| | ID | Store ID | Total Price | Base Price | Units Sold |
|---|---|---|---|---|---|
| 150144 | 212637 | 9984 | 239.4000 | 239.4000 | 23 |
| 150145 | 212638 | 9984 | 235.8375 | 235.8375 | 38 |
| 150146 | 212639 | 9984 | 235.8375 | 235.8375 | 30 |
| 150147 | 212642 | 9984 | 357.6750 | 483.7875 | 31 |
| 150148 | 212643 | 9984 | 141.7875 | 191.6625 | 12 |
| 150149 | 212644 | 9984 | 234.4125 | 234.4125 | 15 |

We have the data of Product ID , Store ID , Total Price , Base Price , Units Sold in our dataset.There are 5 columns and 150150 rows.

## Handle Missing Data:

Missing Data can affect the performance of your machine learning models, and bias the conclusions derived from all the statistical analysis on the data.

So, it is necessary to handle the missing data before training our Machine Learning model.

Missing data can be handled in following ways:

- Deleting rows

  The row with the missing data can be removed if it have 70 – 75% of missing values

- Replacing with Mean / Median / Mode

  i) We can calculate the mean, median or mode of the feature and replace it with the missing values

  ii) Replacing with the above three approximations are a statistical approach of

handling the missing values

Number of missing value in our dataset

```
#before
dataset.isnull().sum()
```

```
ID            0
Store ID      0
Total Price   1
Base Price    0
Units Sold    0
dtype: int64
```

It seems to be 1 missing value in the ‘Total Price’ of our dataset

We have to remove the missing value using the method called dropna() method

```
dataset.dropna(subset=["Total Price"],how='all',inplace=True)
```

```
#after
dataset.isnull().sum()
```

```
ID            0
Store ID      0
Total Price   0
Base Price    0
Units Sold    0
dtype: int64
```

Now there is no missing value in our dataset. so we can use this dataset for the training of machine learning model

## Splitting the Dataset:

The dataset should be split into Train data and Test data for the following reasons

- To estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications
- To measure the accuracy of our model

Before splitting our dataset we should separate the Dependent variable and Independent variable

In our dataset the Dependent variable are ‘Total Price’ , ‘Base Price’ , ‘Units Sold’ and the Independent variable is ‘Product ID’.

Dependent Variable - The dependent variable is the variable about which predictions or explanations are being sought.

Independent Variable - Independent variables (also referred to as Features) are the input for the prediction in our model which is not dependent on other variable.

```
#Seperate the Dapendent Variable and Independent Variable
X = dataset.iloc[:,2:]
X
```

| | Total Price | Base Price | Units Sold |
|---|---|---|---|
| 0 | 99.0375 | 111.8625 | 20 |
| 1 | 99.0375 | 99.0375 | 28 |
| 2 | 133.9500 | 133.9500 | 19 |
| 3 | 133.9500 | 133.9500 | 44 |
| 4 | 141.0750 | 141.0750 | 52 |
| ... | ... | ... | ... |
| 150145 | 235.8375 | 235.8375 | 38 |
| 150146 | 235.8375 | 235.8375 | 30 |
| 150147 | 357.6750 | 483.7875 | 31 |
| 150148 | 141.7875 | 191.6625 | 12 |
| 150149 | 234.4125 | 234.4125 | 15 |

150149 rows × 3 columns

Let us take Y as Dependent Variable . It includes all the value of Product ID or ID.

```
Y = dataset.iloc[:,0]
Y
```

```
0              1
1              2
2              3
3              4
4              5
            ...
150145    212638
150146    212639
150147    212642
150148    212643
150149    212644
Name: ID, Length: 150149, dtype: int64
```

Let us split the dataset into Train and Test set.

**Visualization**

●Scatterplot Visualization

The following contains the visualization of dataset using Scatterplot

```
total_price = data['Total Price']
units_sold = data['Units Sold']

# Create a scatter plot
plt.scatter(total_price, units_sold,color='green',alpha=0.3)
plt.xlabel('Total Price')
plt.ylabel('Units Sold')
plt.title('Total Price vs. Units Sold')

# Display the plot
plt.show()
```

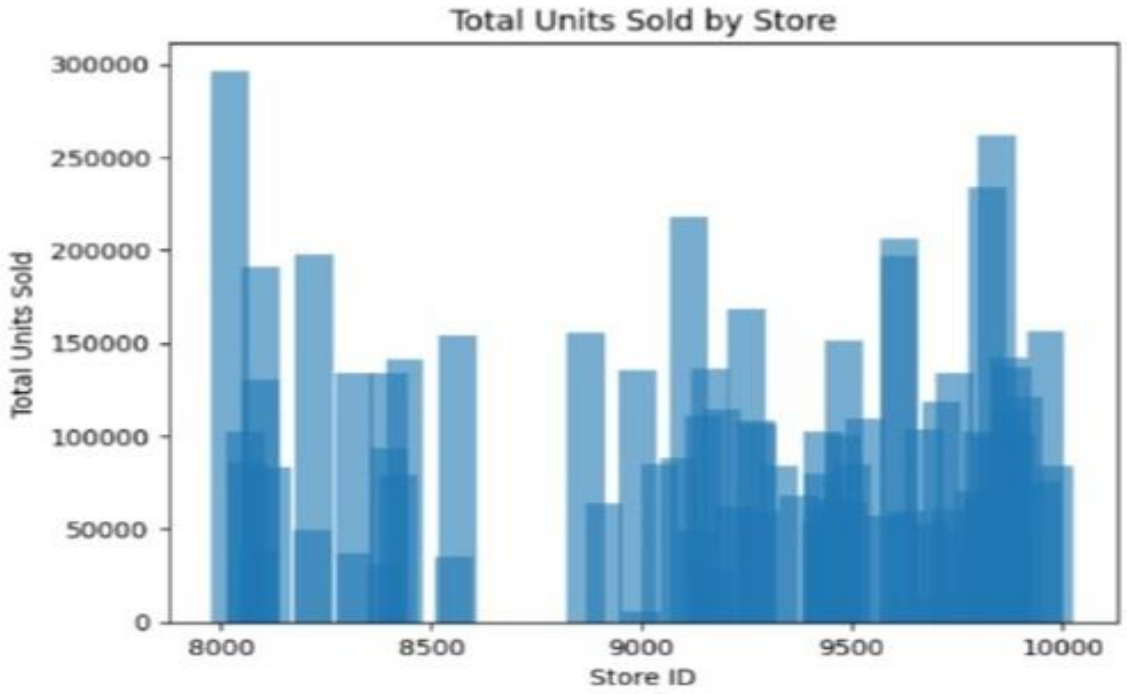

●Barchart Visualization

The following contains the visualization of dataset using Barchart

```
# Group the data by Store ID and calculate the sum of Units Sold for each store
units_sold_by_store = data.groupby('Store ID')['Units Sold'].sum()

# Create a bar chart
plt.bar(units_sold_by_store.index, units_sold_by_store.values,width=90,alpha=0.6)
plt.xlabel('Store ID')
plt.ylabel('Total Units Sold')
plt.title('Total Units Sold by Store')

# Display the plot
plt.show()
```




## Conclusion

Thus the preprocessing of product demand prediction with machine learning for_the_dataset https://www.kaggle.com/datasets/chakradharmattapalli/product-demand-prediction-with-machine-learning is concluded