**PRODUCT DEMAND PREDICTION WITH MACHINE LEARNING**

**TeamLeader: Santhoshkumar K**

**Nm Id: au723721243044**

**DATA PREPROCESSING**

**IMPORT LIBRARY:**

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

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

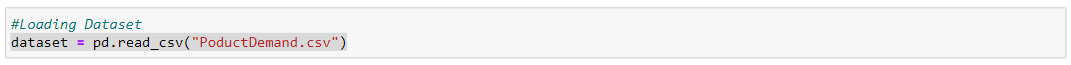
**PANDAS :** It is a analysis tool used for data manipulation

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

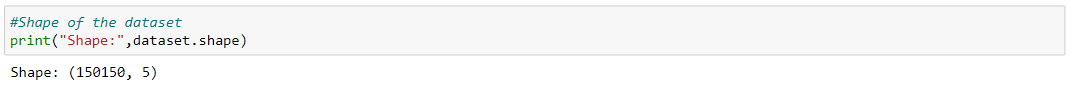


**SKLEARN :**  To Implement Machine Learning model and Statistical Modeling

**LOADING THE DATA SET :**



Identify the shape of the dataset...

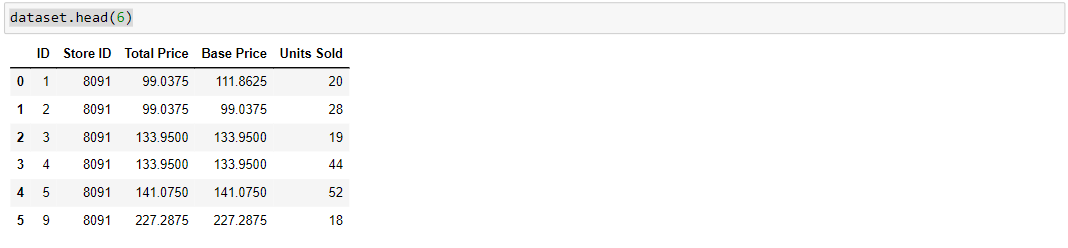


The dataset contains 150150 rows and 5 columns

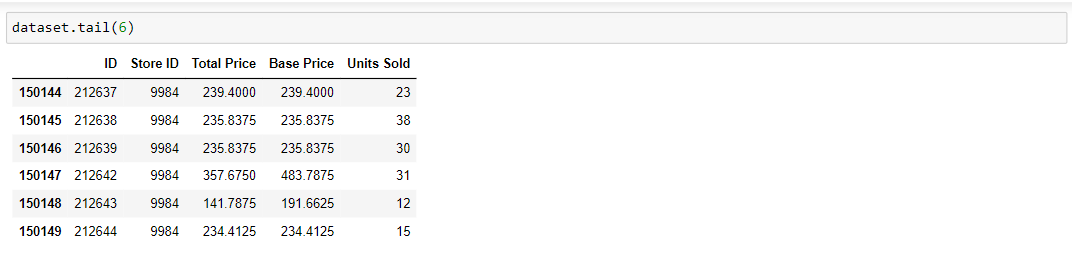
To check the head and tail of the dataset use the following

* dataset.head(contains the no. of rows)
* dataset.tail(contains the no. of rows)

This shows the first 6 rows of the dataset.



This shows the last 6 rows of the dataset



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

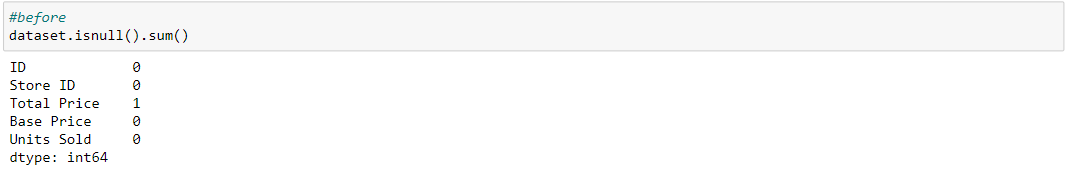
The row with the missing data can be removed if it have 70 – 75% of missing values

* 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

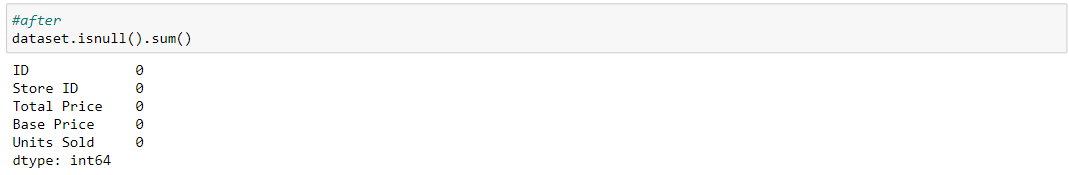
handling the missing values

Number of missing value in our dataset

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





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

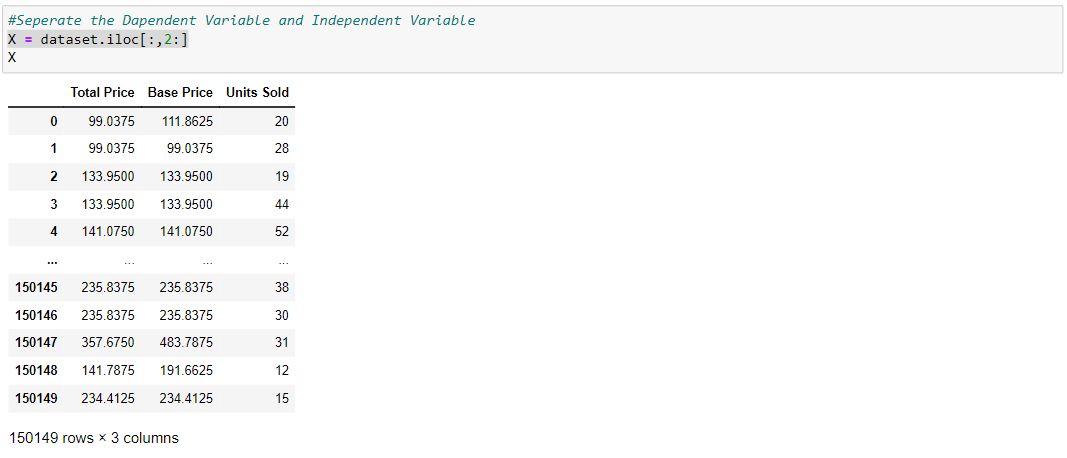
* To estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications
* To measure the accuracy of our model

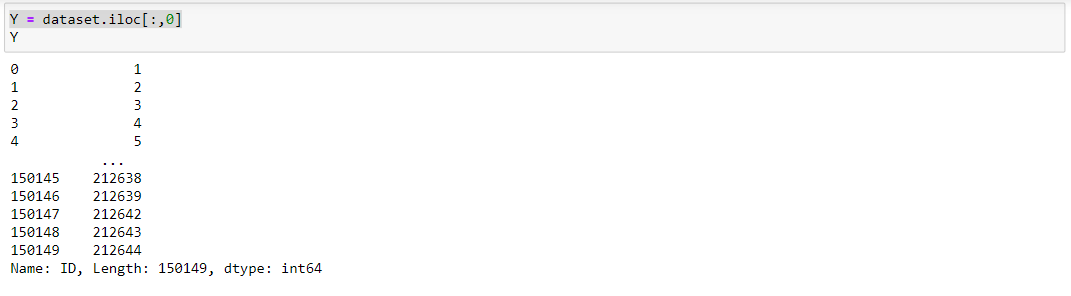
Before splitting our dataset we should separate the Dependent variable and Independent variable

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.



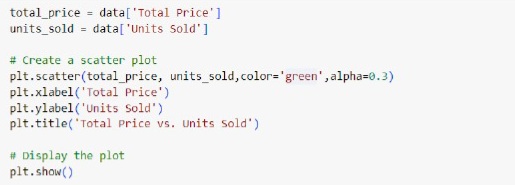
Let us take Y as Dependent Variable . It includes all the value of Product ID or ID.

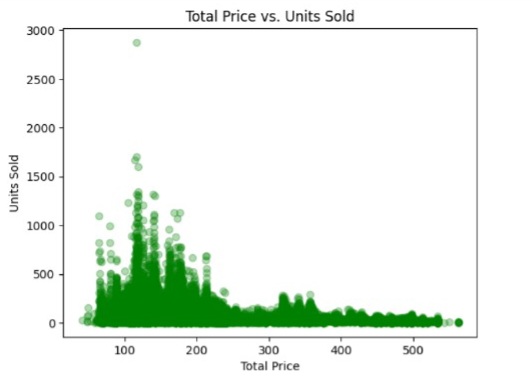
Let us split the dataset into Train and Test set.

**Visualization**

●Scatterplot Visualization

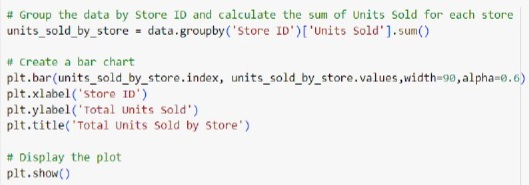
The following contains the visualization of dataset using Scatterplot

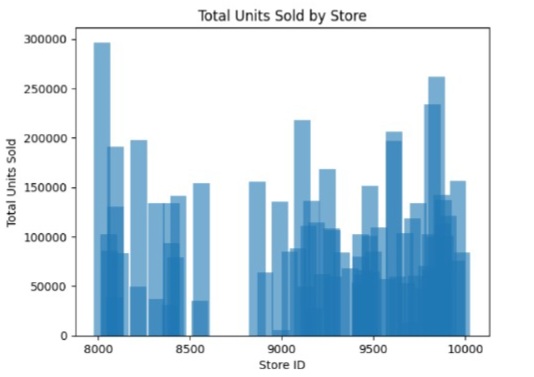




●Barchart Visualization

The following contains the visualization of dataset using Barchart





**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