Project Name:- Predict Stock Market on Tesla Dataset By using Linear regression

In statistics, Linear regression is an approach for modelling the relationship between a scalar dependt variable y and one more explanatory variables (or independent variables)denoted X.The case of onr explanatory vaiables is called simple linear regression

What is Linear Regression

Here is the formal defintion "Linear regression is an approach for modelling the reletionship between a scalar dependent variables y and one or more explanatory variables (or independent variables)denoted x"

Let me explain the concept of regression in a very basic manner, so imagine that you run a company that builds cars and you want to understand how the change in prices of raw materials (let's say Steel) will affect the sales of the car. The general understanding is this, the rise in the price of steel will lead to a rise in the price of the car resulting in lesser demand and in turn lesser ales. But how do we quantify this? And how do we predict how much change in sales will happen based on the degree of change in steel price. That's when the regression comes

Linear regression is the analysis of two separate variables to define a single relationship and is a useful measure for technical and quantitative analysis in financial markets.

Plotting stock prices along a normal distribution-bell curve-can allow traders to see when a stock is overbought or oversold.

I Using linear regression, a trader can identify key price points-entry price, stop-loss price, and exit prices.

A stock's price and time period determine the system parameters for linear regression, making the method universally applicable. Stock market close price is an important piece of information that is very useful for every short-term trader. The close prices are very important, especially for swing traders and position traders.

In this case study we choose linear regression for our analysis. First, we divide the data into two parts of training and testing. Then we use the training section for starting analysis and defining the model.

About the dataset-

Opening price

The opening price is the value that each share has the opening price gives a good indication of where the stock will move during the day. Since the Stock exchange can be likened with an auction market i.e. buyers and sellers meet to make deals with the highest bidder, the

opening price does not have to be the same as the last day's closing price.

Closing Price

An adjusted closing price is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open. The adjusted closing price is often used when examining historical returns or performing a detailed analysis on historical returns

Volume

Volume is one of the most basic and beneficial concepts to understand when trading stocks. Volume is defined as, "the number of shares or contracts traded in a security or an entire market during a given period of time"

Name - Md Sharfe Alam

SID - 90749

About The Project-

ML01 Lab Project-

Name:- Predict Stock market on Tesla dataset

Objective : Best models for close price prediction of the price of Tesla Stocks

Steps and Tasks :

Step 1 : Reading and understanding of data

Step 2 : Data cleaning and Preparation

Step 3 : Visualizing the data

Step 4 : Deriving new features

Step 5 : Train-Test Split and feature scaling

Step 6 : Model Building

Step 7 : Accuracy and Evaluationn

Importing Some Required Libraries

In [1]:
```
# libraries for dataframes & array handling
import pandas as pd
import numpy as np

# libraries for data visualization
import seaborn as sns
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (7,4)
plt.rcParams['figure.dpi'] = 300
%matplotlib inline
sns.set_theme(style='darkgrid', palette='inferno')

# library to ignore warnings that arise during visualizations
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
# SKLEARN CLASSES & LIBRARIES

# importing train test split & GridSearchCV (for performing grid search on var
from sklearn.model_selection import train_test_split, GridSearchCV

# import StandarScaler for data Standardization
from sklearn.preprocessing import StandardScaler

# importing linear regression, polynomial regression, Regulariation classes (R
from sklearn.linear_model import LinearRegression, RidgeCV, LassoCV,ElasticNe
from sklearn.preprocessing import PolynomialFeatures

# importing model evaluation metrics
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score,
```

Step 1 : Reading and understanding of data

In [3]:
```
#Read The Csv File From my Location
data=pd.read_csv("C:\\Users\\mdsha\\Downloads\\archive (3)\TSLA.csv")
```

In [4]:
```
data.head(10)#checking out data
```

Out[4]:

	Date	Open	High	Low	Close	Adj Close	Volume
0	2010-06-29	3.800	5.000	3.508	4.778	4.778	93831500
1	2010-06-30	5.158	6.084	4.660	4.766	4.766	85935500
2	2010-07-01	5.000	5.184	4.054	4.392	4.392	41094000
3	2010-07-02	4.600	4.620	3.742	3.840	3.840	25699000
4	2010-07-06	4.000	4.000	3.166	3.222	3.222	34334500
5	2010-07-07	3.280	3.326	2.996	3.160	3.160	34608500
6	2010-07-08	3.228	3.504	3.114	3.492	3.492	38557000
7	2010-07-09	3.516	3.580	3.310	3.480	3.480	20253000
8	2010-07-12	3.590	3.614	3.400	3.410	3.410	11012500
9	2010-07-13	3.478	3.728	3.380	3.628	3.628	13400500

Step 2 : Data cleaning and Preparation

In [5]:
```
# checking shape of dataframe
data.shape
```

Out[5]: (2956, 7)

In [6]:
```
data.drop('Adj Close',axis=1,inplace=True)
```

In [7]:
```
data.head()
```

Out[7]:

	Date	Open	High	Low	Close	Volume
0	2010-06-29	3.800	5.000	3.508	4.778	93831500
1	2010-06-30	5.158	6.084	4.660	4.766	85935500
2	2010-07-01	5.000	5.184	4.054	4.392	41094000
3	2010-07-02	4.600	4.620	3.742	3.840	25699000
4	2010-07-06	4.000	4.000	3.166	3.222	34334500

In [8]:
```
# checking for null vlaues
data.isnull().sum()
```

Out[8]:
```
Date      0
Open      0
High      0
Low       0
Close     0
Volume    0
dtype: int64
```

In [9]:
```
#Check Missing values in dataset
data.isna().any()
```

Out[9]:
```
Date      False
Open      False
High      False
Low       False
Close     False
Volume    False
dtype: bool
```

In [10]:
```
# getting basic data info
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2956 entries, 0 to 2955
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Date    2956 non-null   object 
 1   Open    2956 non-null   float64
 2   High    2956 non-null   float64
 3   Low     2956 non-null   float64
 4   Close   2956 non-null   float64
 5   Volume  2956 non-null   int64  
dtypes: float64(4), int64(1), object(1)
memory usage: 138.7+ KB
```

In [11]:
```
# getting basic statistics for our data, as all the columns are of numerical co

data.describe()
```

Out[11]:

	Open	High	Low	Close	Volume
count	2956.000000	2956.000000	2956.000000	2956.000000	2.956000e+03
mean	138.691296	141.771603	135.425953	138.762183	3.131449e+07
std	250.044839	255.863239	243.774157	250.123115	2.798383e+07
min	3.228000	3.326000	2.996000	3.160000	5.925000e+05
25%	19.627000	20.402000	19.127500	19.615000	1.310288e+07
50%	46.656999	47.487001	45.820002	46.545000	2.488680e+07
75%	68.057001	69.357500	66.911501	68.103998	3.973875e+07
max	1234.410034	1243.489990	1217.000000	1229.910034	3.046940e+08

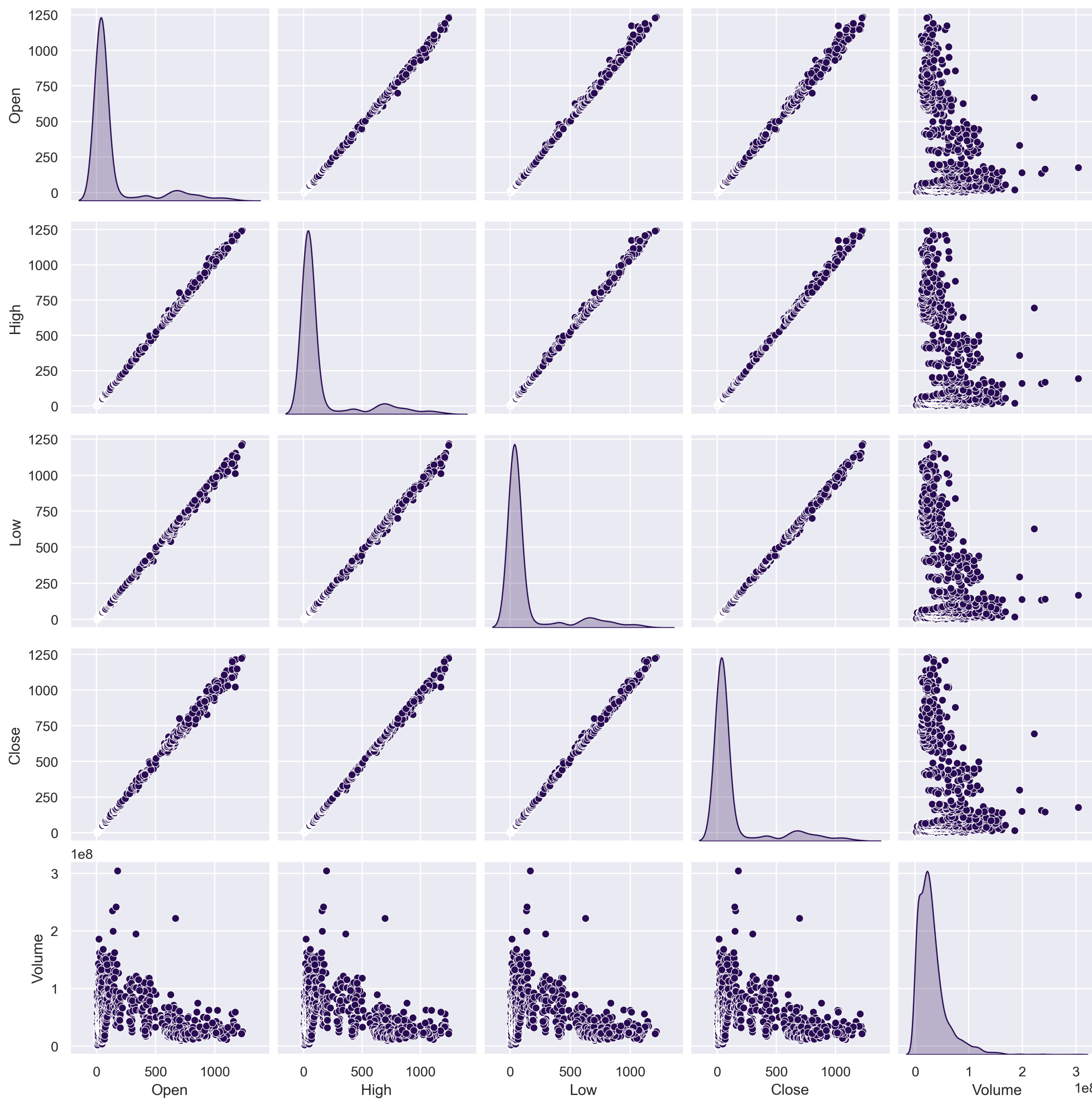
In [12]:
```
#Lenth of my Data
print(len(data))
```

```
2956
```

Step 3 : Visualizing the data

In [13]:
```
# creating a pair plot to visualize realtionship between all the columns at or
sns.pairplot(data, diag_kind='kde')
```

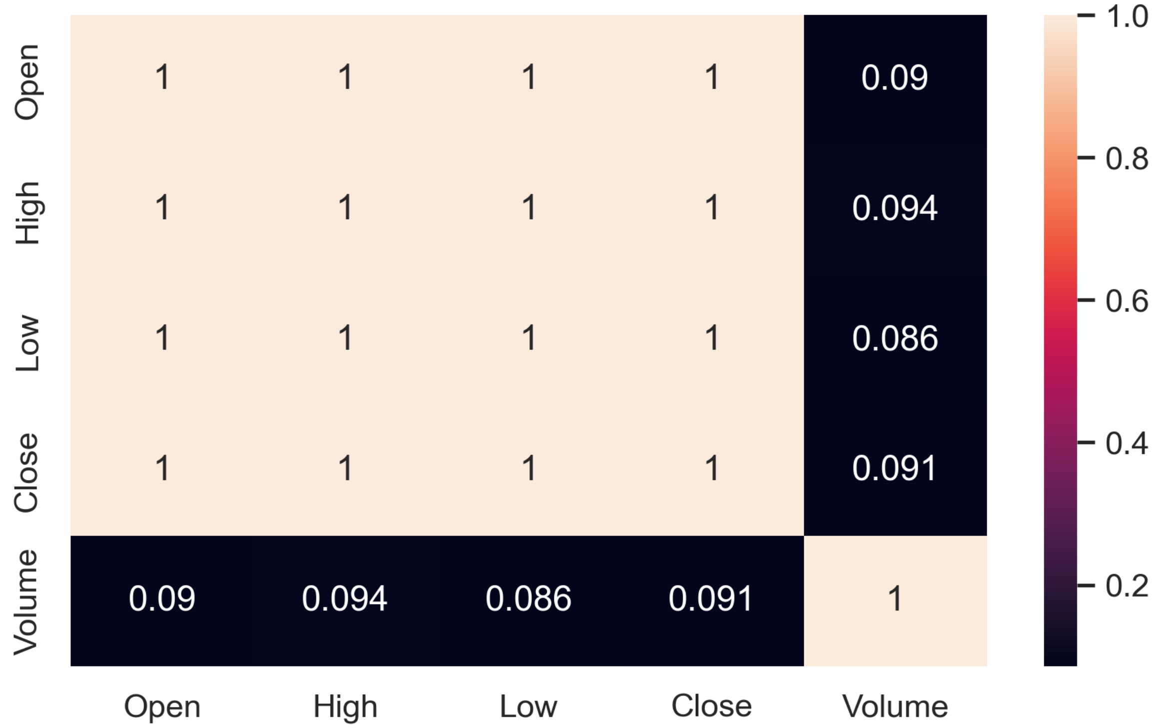
Out[13]: <seaborn.axisgrid.PairGrid at 0x2cdfd2df6d0>



In [14]:
```
# checking for correlation
print(data.corr())
sns.heatmap(data.corr(),annot=True)
```

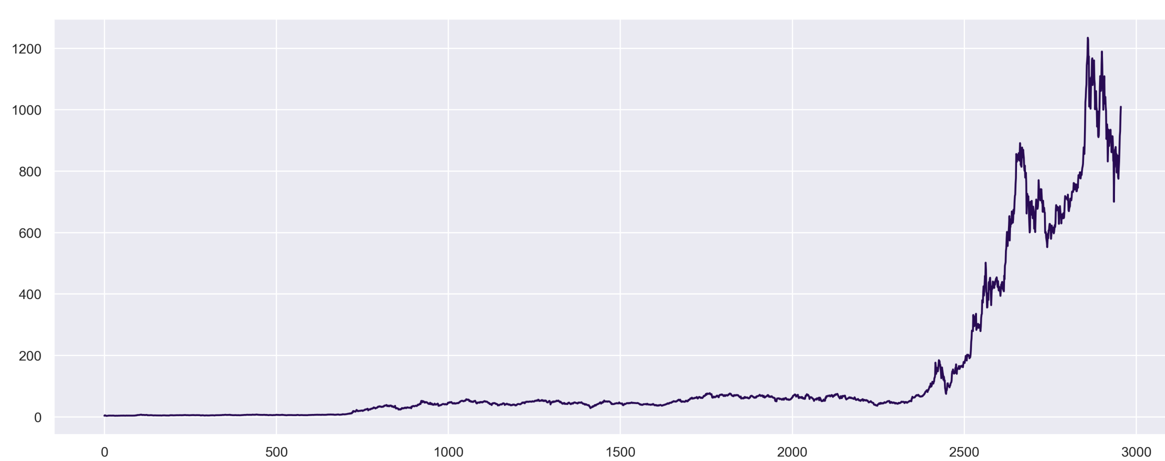
```
            Open      High       Low     Close    Volume
Open    1.000000  0.999726  0.999617  0.999247  0.089750
High    0.999726  1.000000  0.999595  0.999666  0.093625
Low     0.999617  0.999595  1.000000  0.999670  0.085906
Close   0.999247  0.999666  0.999670  1.000000  0.090602
Volume  0.089750  0.093625  0.085906  0.090602  1.000000
```

Out[14]:
```
<Axes: >
```



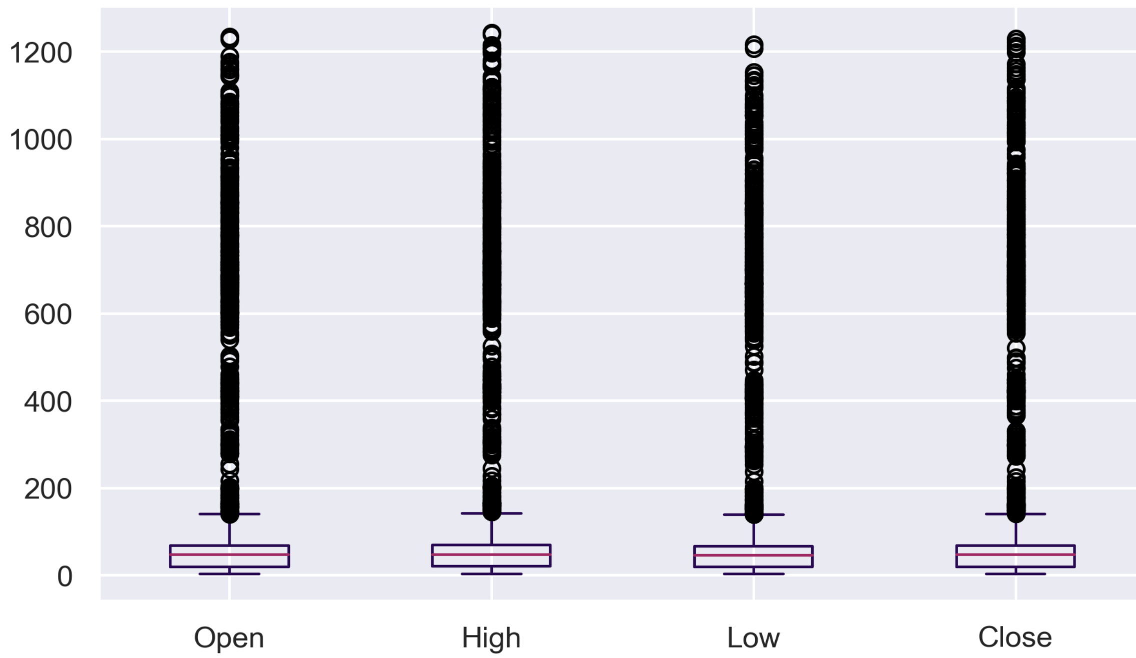
In [15]:
```
data['Open'].plot(figsize=(16,6))
```

Out[15]:
```
<Axes: >
```



In [16]:
```
data[['Open','High','Low','Close']].plot(kind='box') #box plot
```

Out[16]: <Axes: >



Step 4 : Train-Test Split and feature scaling¶

In [17]:
```
x=data[['Open','High','Low','Volume']]
y=data['Close']
```

In [18]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y, random_state=0)
```

In [19]:
```
x_train.shape
```

Out[19]: (2217, 4)

In [20]:
```
x_test.shape
```

Out[20]: (739, 4)

Step 5 : Model Building

Regression for Stock Price Prediction

In [21]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import confusion_matrix,accuracy_score
regressor=LinearRegression()
```

In [22]:
```
regressor.fit(x_train,y_train)
```

Out[22]:
```
LinearRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [23]:
```
print(regressor.coef_)
```

```
[-6.96214496e-01  9.34024620e-01  7.59984324e-01  6.49724984e-09]
```

In [24]:
```
print(regressor.intercept_)
```

```
-0.1439953680371957
```

In [25]:
```
predicted=regressor.predict(x_test)
```

In [26]:
```
print(x_test)
```

```
            Open        High         Low    Volume
1749   74.884003   75.374001   70.959999  86307000
643     6.832000    6.970000    6.784000   7183500
118     5.734000    5.994000    5.706000   3714500
252     5.558000    5.650000    5.534000   4446000
1311   50.220001   50.849998   49.933998  14454500
...          ...         ...         ...       ...
794    31.400000   32.459999   31.000000  64659500
2429  167.800003  172.699997  164.440002  75961000
517     7.000000    7.042000    6.476000  12846500
1943   63.299999   64.150002   61.933998  37421500
1912   70.199997   71.931999   69.725998  20988500

[739 rows x 4 columns]
```

In [27]:
```
predicted.shape
```

Out[27]:
```
(739,)
```

Step 6 : Deriving new features¶

In [28]:
```
Df=pd.DataFrame(y_test,predicted)
```

In [53]:
```
Df=pd.DataFrame({'Actual Price':y_test,'Predicted Price':predicted})
```

In [56]:
```
print(Df)
```

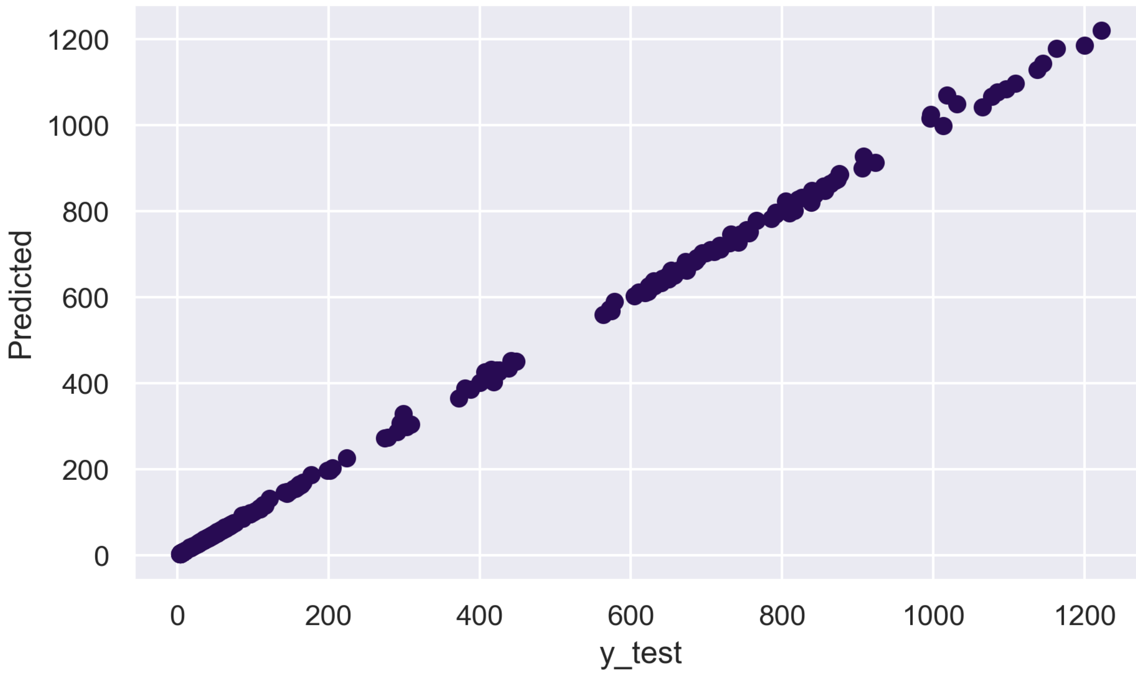
```
      Actual Price  Predicted Price
1749     71.463997        72.611094
643       6.876000         6.812025
118       5.920000         5.823059
252       5.622000         5.498324
1311     50.638000        50.430232
...            ...              ...
794      32.368000        32.292931
2429    166.757996       169.802621
517       6.670000         6.565030
1943     62.712002        63.015313
1912     69.849998        71.295040

[739 rows x 2 columns]
```

In [57]:
```
Df.head(10)
```

Out[57]:

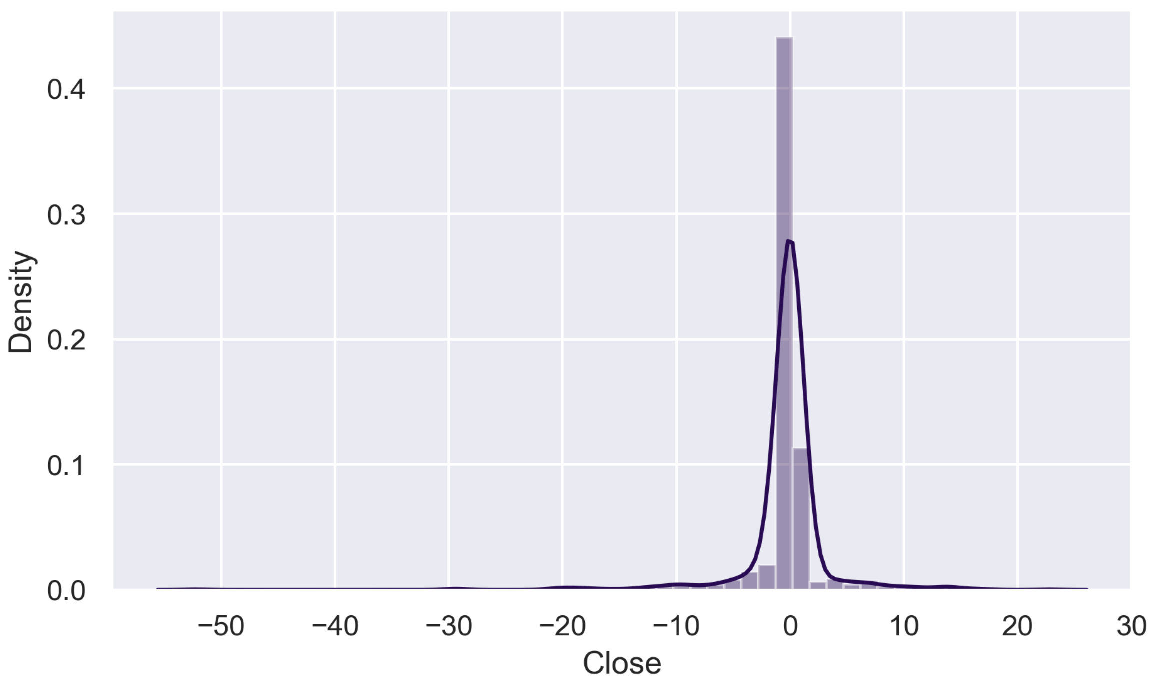
	Actual Price	Predicted Price
1749	71.463997	72.611094
643	6.876000	6.812025
118	5.920000	5.823059
252	5.622000	5.498324
1311	50.638000	50.430232
1083	45.270000	45.689660
719	11.158000	11.409259
1467	50.293999	50.309781
2407	113.912003	117.280422
1592	40.551998	40.493807

In [58]:
```
plt.scatter(y_test, predicted)
plt.ylabel('Predicted')
plt.xlabel('y_test')
plt.show()  # Don't forget to add this line to display the plot
```



In [59]:
```
sns.distplot((y_test-predicted))
```

Out[59]: <Axes: xlabel='Close', ylabel='Density'>



Step 7 : Accuracy and Evaluation¶

In [60]:
```
from sklearn.metrics import confusion_matrix,accuracy_score
```

In [61]:
```
regressor.score(x_test,y_test)
```

Out[61]: 0.9997322153411554

In [62]:
```
import math
# checking model performance by using various evaluation metrics
# (it compares actual y_test and predicted values)
```

In [63]:
```
regressor.score(x_test,predicted)
```

Out[63]: 1.0

In [67]:
```
mse = mean_squared_error(y_test,predicted)
```

In [68]:
```
print(mse)
```

17.8062465372882

In [73]:
```
cm = confusion_matrix(x_test,predicted)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[73], line 1
----> 1 cm = confusion_matrix(x_test,predicted)

File ~\AppData\Local\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:317, in confusion_matrix(y_true, y_pred, labels, sample_weight, normalize)
    232 def confusion_matrix(
    233     y_true, y_pred, *, labels=None, sample_weight=None, normalize=None
    234 ):
    235     """Compute confusion matrix to evaluate the accuracy of a classification.
    236 
    237     By definition a confusion matrix :math:`C` is such that :math:`C_{i, j}`
   (...)
    315     (0, 2, 1, 1)
    316     """
--> 317     y_type, y_true, y_pred = _check_targets(y_true, y_pred)
    318     if y_type not in ("binary", "multiclass"):
    319         raise ValueError("%s is not supported" % y_type)

File ~\AppData\Local\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:95, in _check_targets(y_true, y_pred)
     92     y_type = {"multiclass"}
     94 if len(y_type) > 1:
---> 95     raise ValueError(
     96         "Classification metrics can't handle a mix of {0} and {1} targets".format(
     97             type_true, type_pred
     98         )
     99     )
    101 # We can't have more than one value on y_type => The set is no more needed
    102 y_type = y_type.pop()

ValueError: Classification metrics can't handle a mix of continuous-multioutput and continuous targets
```

In [70]:
```
print("Mean Squared Error:",mean_squared_error(y_test,predicted))
```

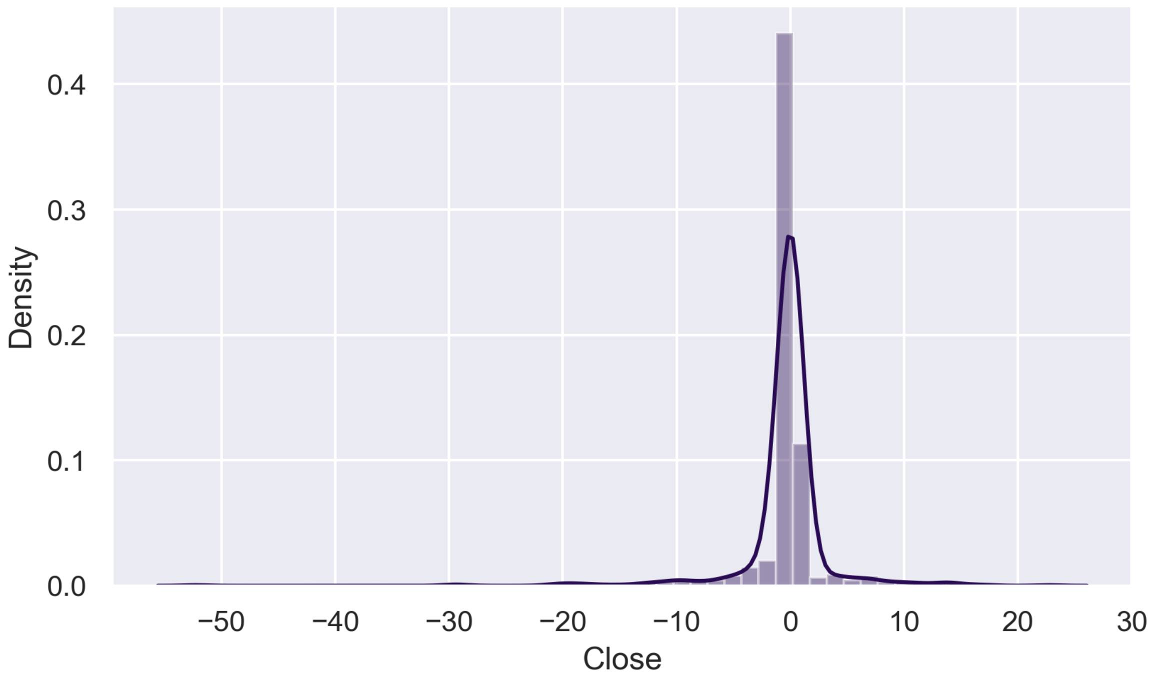
```
Mean Squared Error: 17.8062465372882
```

In [71]:
```
print("Root Mean Squared Error:",mean_squared_error(y_test,predicted))
```

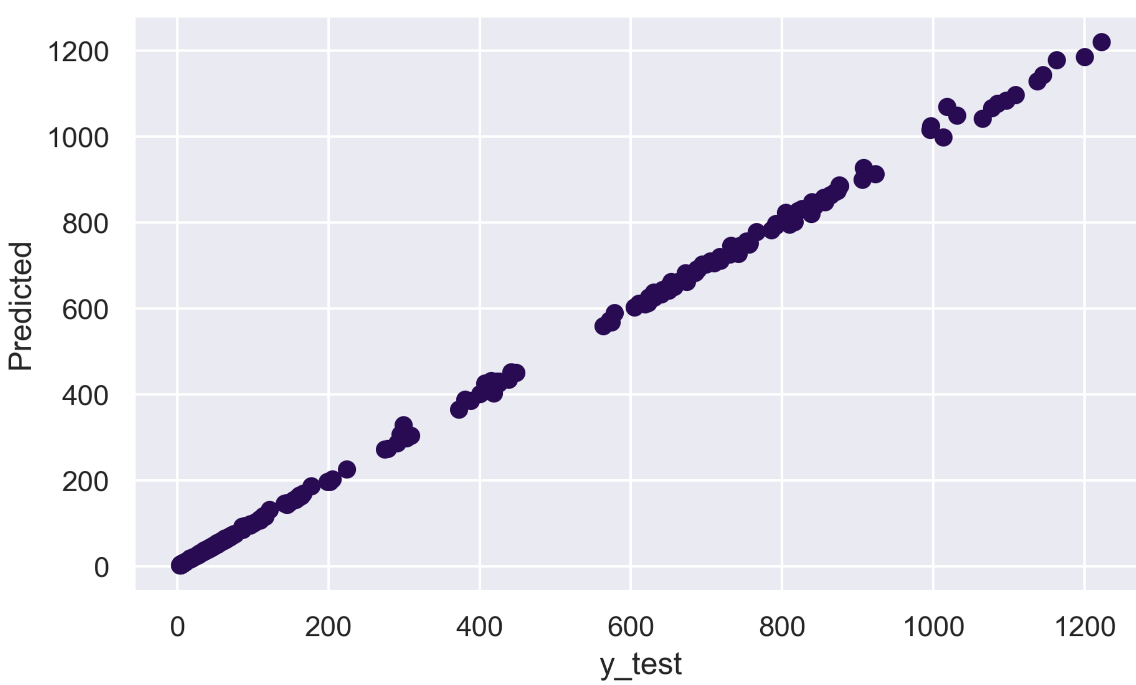
```
Root Mean Squared Error: 17.8062465372882
```

In [87]:
```
sns.distplot((y_test-predicted))
```

Out[87]: <Axes: xlabel='Close', ylabel='Density'>



In [88]:
```
plt.scatter(y_test, predicted)
plt.ylabel('Predicted')
plt.xlabel('y_test')
plt.show()  # Don't forget to add this line to display the plot
```



In [ ]:

In [ ]: