Stock Price Prediction Using Machine Learning

By

*21781A33C4 (R Pavithra)*

*21781A33D0 (S Soniya)*

*21781A33D1 (Sanjana K)*

*21781A33E9 (V Naga Shanthi)*

*Guided by*

*Prof. Ms Raasha*

A Dissertation Submitted to

SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY, An Autonomous Institution affiliated to ‘JNTU Ananthapur’ in Partial Fulfilment of the Bachelor of Technology (***Computer Engineering***) with Specialization in ***Artificial Intelligence and Machine Learning.***

*May 2024*

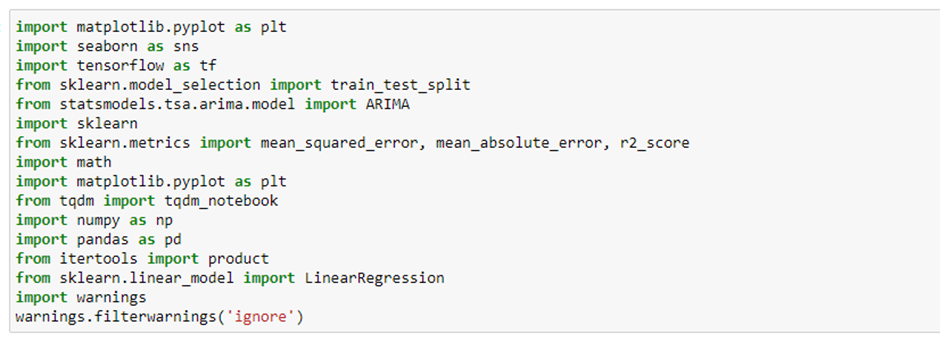


**SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY**

**R.V.S. Nagar Tirupathi Road, Andhra Pradesh– 517127**

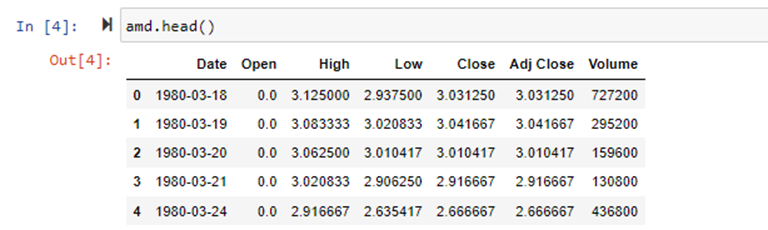
**DATA PREPROCESSING:**

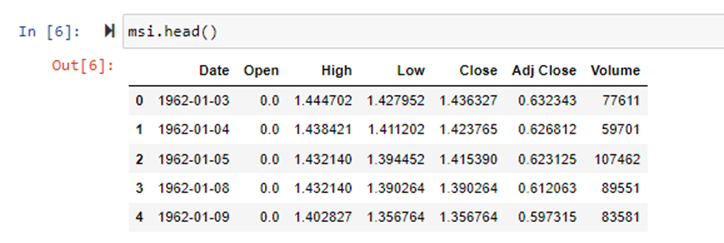
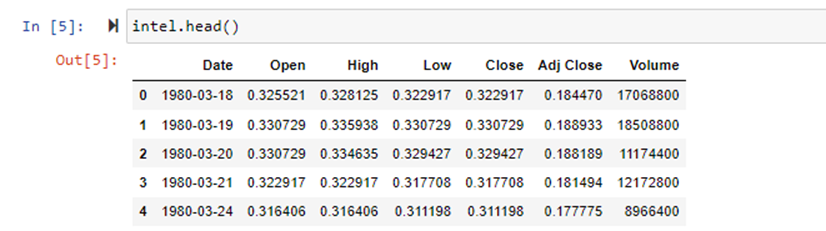
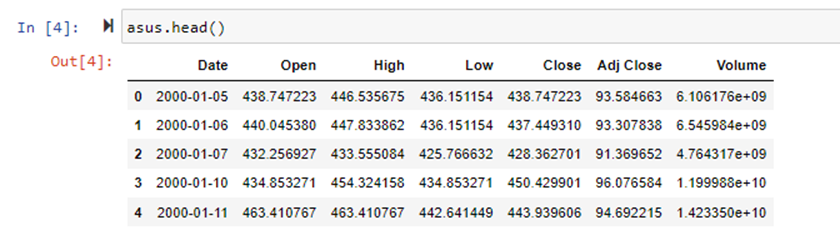
### ***Importing The Libraries***

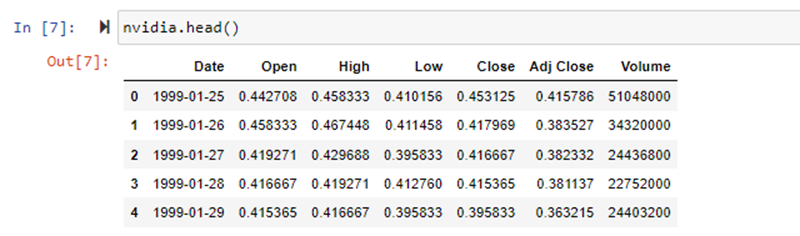


**Read The Dataset**

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.   
 In pandas, we have a function called read\_csv() to read the dataset. As a parameter, we have to give the directory of the CSV file.  
   
  


We can print the first 5 rows of the datasets using the .head() method as shown in the below screenshots.  
   






### ***Data Preparation***

### *As we have understood how the data is, let's pre-process the collected data.*

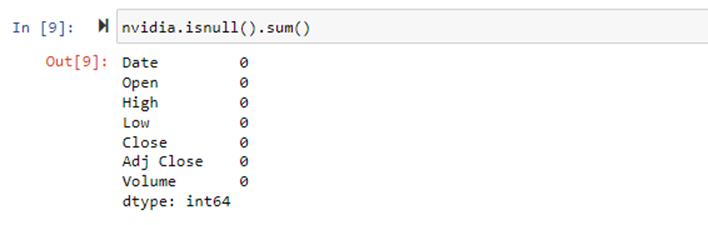
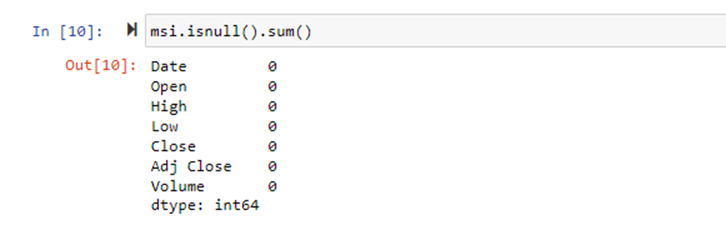
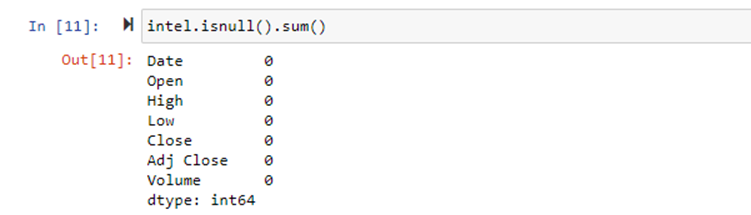
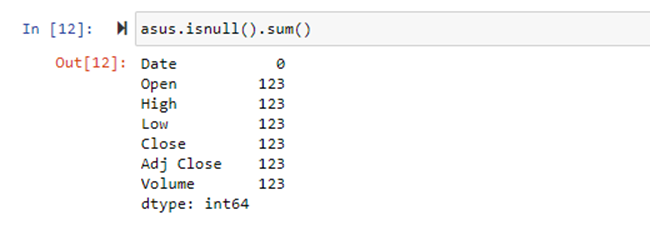
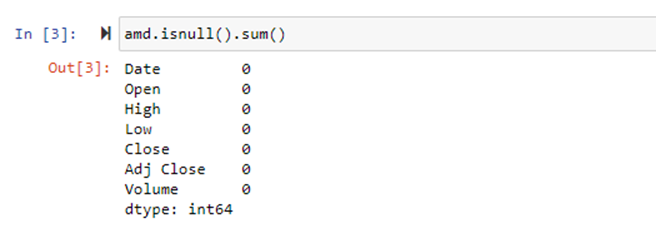
The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

* Checking for missing values
* Data manipulation
* Resampling the data
* Merging and splitting data into test and train variables

Note: These are the general steps of pre-processing the data before using it for training models. Depending on the condition of your dataset, you may or may not have to go through all these steps.

### ***Checking For Missing Values***

For checking the null values, df.isnull() function is used. To sum those null values we use .sum() function. From the below image we found that there are no null values present in our dataset. So we can skip handling the missing values step.



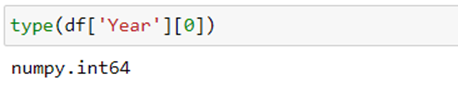
### ***Data Manipulation***

Since we found null values in the dataset related to ASUS company we are going to drop the null values but we can also fill those null values using mean/median/mode of the respective volumns.

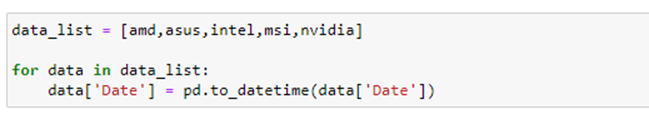


To convert the date strings to time stamps we are going to perform below steps.

First let’s check the type of data of the column “Year” using the type() function.



We will convert the data of “Year” column using the pandas “to\_datetime” function. We can change the ‘Date’ columns of all the datasets using a for loop to reduce Lines of code..



### ***Resampling The Data***

Now we are going to add new columns to all the loaded datasets so that when we merge our data it will be easier to find which rows of data belong to which company and also so that our model will be able to differentiate the rows of data based on the value of the new column named “Company”. The new column “Company” will be a categorical value ranging from 0 to 4 where each number denotes a company as shown below. We are going to process the datasets as mentioned above using a for loop.

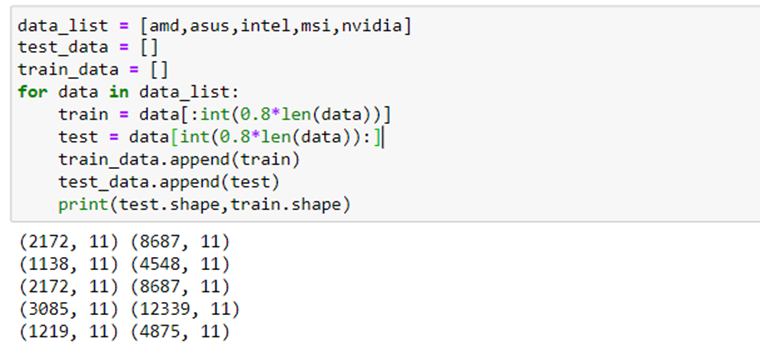
We will also three new columns of data namely “Year”, “Day” and “Month” which are derived from the “Date” column. We will use these three columns for training not the “Date” columns but we use the “Date” column for analysis in later stages.



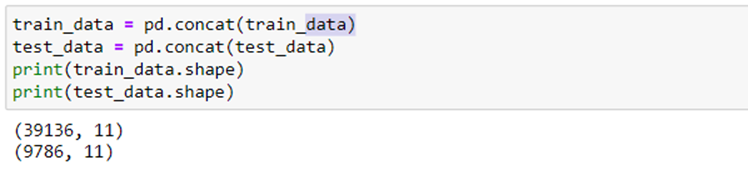
### ***Merging And Splitting Data Into Test And Train Variables***

Now we are going to merge our datasets, simultaneously we are also going to split the data to test and train variables using the below piece of code with a train to test ratio of 80 to 20.

First we are going to take 2 lists of train and test in which we are going to append train and test splits of every dataset into these 2 lists. In the output each line represents the data of companies in order as stored in the data\_list in below code.



Next we are going to merge the data of each variable of train\_data and test\_data using the pandas.concat() function as shown below.

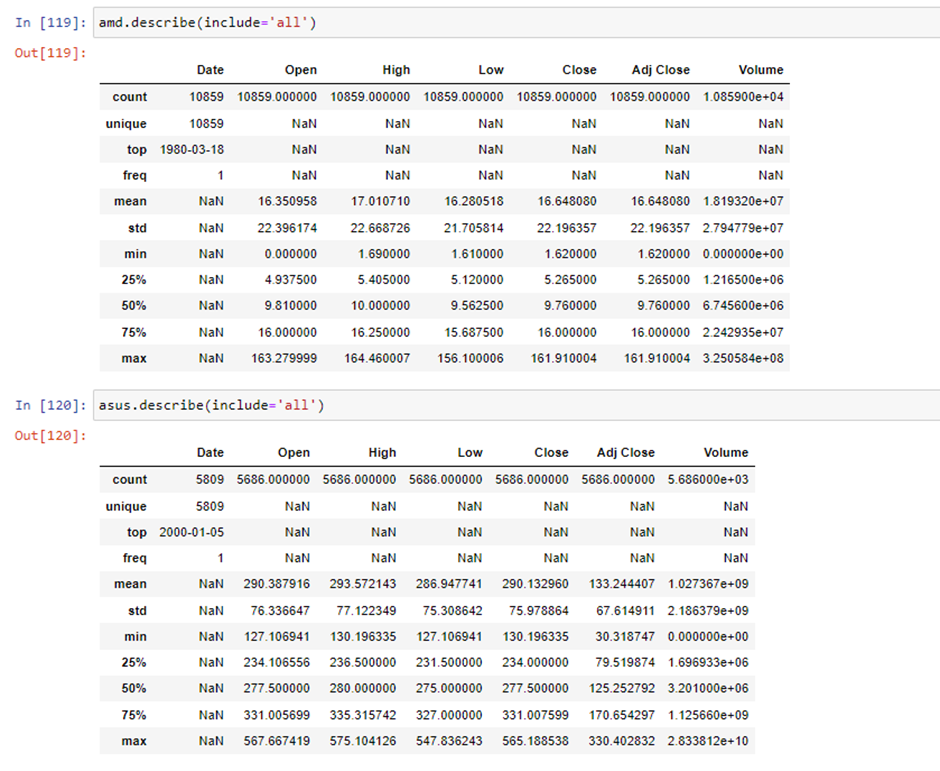


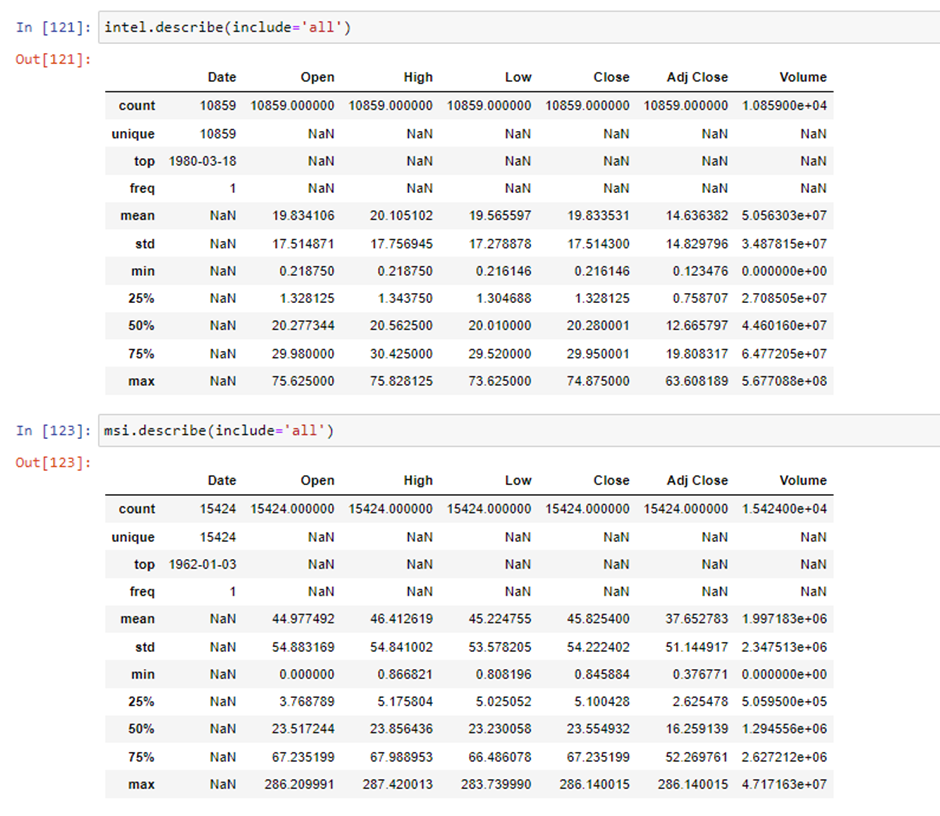
Finally we are going to split the train\_data and test\_data variables into x\_train, y\_train, x\_test and y\_test using the below piece of code.

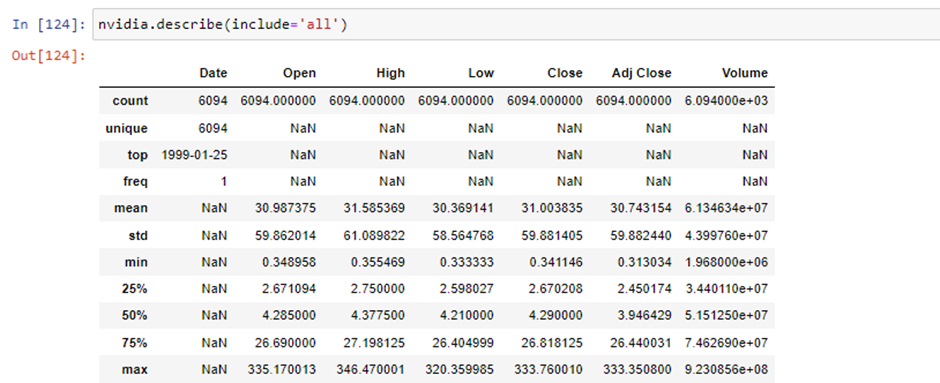
### ***Exploratory Data Analysis***

### ***Descriptive Statistical***

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.





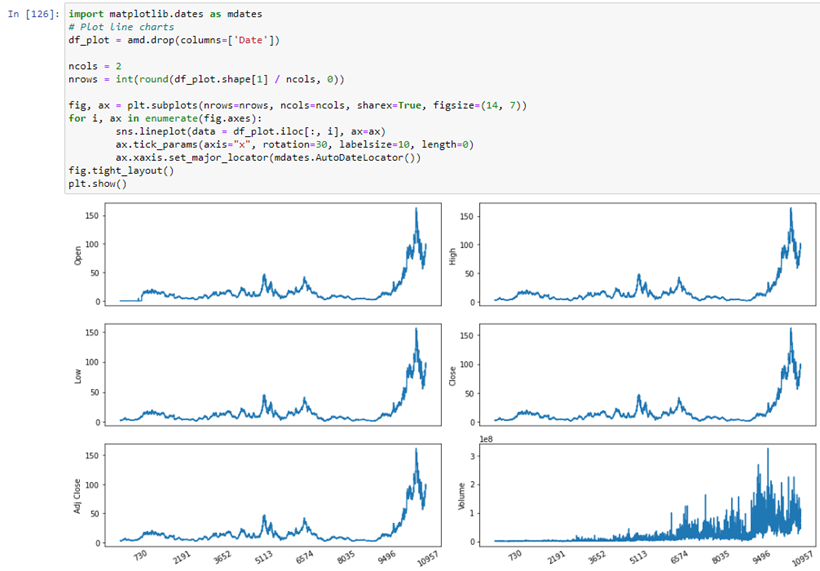


### ***Visual Analysis***

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

***Univariate analysis***   
 In simple words, univariate analysis is understanding the data with single feature. Here we don’t have much need to perform univariate analysis to understand the data as most of the columns provided are continuous.

***Bivariate analysis***  To find the relation between two features we use bivariate analysis.



In the line plots visualised below following observations can be observed:

* There is sudden increase in the prices of stocks of AMD in the recent years and the price of the stock didn’t retain for a long period but it soon fell to a lower price.
* The volume of the stocks being traded have increased tremendously in the recent years.

Similar to this, analysis on other companies can be done.