SEMINAR REPORT ON

**“STOCK MARKET PREDICTION USING LINEAR REGRESSION”**

BY

**Mr.Shantanu Badwe**

**Mr.Swaraj Buchude**

**Mr.Parth Jamdade**

**Mr.Nileet Salve**

Under The Guidance of

**Dr. S. N. Zaware**

**Head of the Department Computer Engineering**



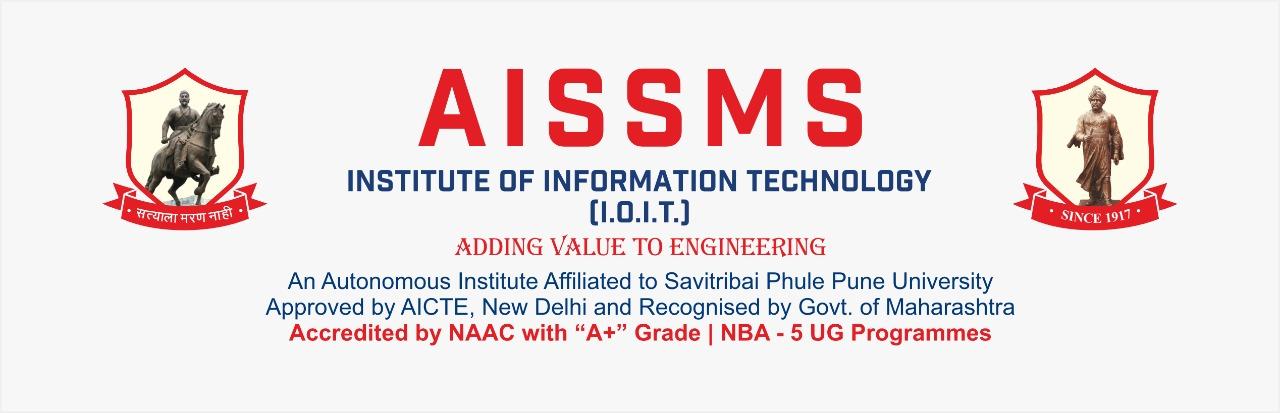
**DEPARTMENT OF COMPUTER ENGINEERING**

**ALL INDIA SHRI SHIVAJI MEMORIAL SOCIETY’S INSTITUTE OF INFORMATION TECHNOLOGY PUNE 411041**



**SAVITRIBAI PHULE PUNE UNIVERSITY**

2022-2023



**CERTIFICATE**

This is to certify that Mr. Swaraj Buchude, Mr. Shantanu Badwe, Mr. Nileet Salve, Mr. Partth Jamdade from Third Year IInd Shift Computer Engineering has successfully completed their seminar work titled

**“STOCK MARKET PREDICTION USING LINEAR REGRESSION”**

at All India Shri Shivaji Memorial Society’s Institute of Information Technology, Pune in the partial fulfilment of the Bachelor's Degree in Computer Engineering

Mrs. S. N. Zaware Mrs.Prajwal Gaikwad

Internal Guide

Project Coordinator

Seal/Stamp of the college

Place: Pune

Date:

**ABSTRACT**

Stock market prediction is the model of determining future values of a company’s stock prices. It helps people who have a great extent in investing their money in stocks and to achieve higher profits. It has been a great mystery for peoples to predict the stock prices as it depends on many factors of a company profile. Stock market keeps varying day by day .In this paper, a regression model is developed to predict the stock values of a company using regression. Analysing such a huge market will prove beneficial to all investors of the system. An application which focuses on the patterns generated in this stock trade over the period of time , and extraction the beneficial information from those patterns to predict the future behaviour of the stock market is necessary. An Application representing the graphical view of stocks values in visual form for user or investors to invest in a particular stock for higher profit is a key requirement. Many Researchers are analysing the data to predict the stock prices but all of them has it’s own short coming. In this Model ,We proposed the application of Machine Learning using Python to predict Stock prices. The algorithm can be used for training set of market data collected for the period of any days.

**ACKNOWLEDGMENT**

Apart from the project members' efforts, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We would like to show our greatest appreciation to our guide Dr.S.N. Zaware . We cannot thank her enough for her tremendous support and help. Her support and continuous guidance was exceptionally helpful to us in solving our queries.

We would also like to express our gratitude to Dr. S.N. Zaware for her exceptional support during the course of the mini-project. We take this as an opportunity to thank all those who were involved in helping and guiding us throughout the project by direct or indirect means.

Mr.Shantanu Badwe

Mr.Swaraj Buchude

Mr.Parth Jamdade

Mr.Nileet Salve

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No**. | **Topic** | **Pg. No.** |
| 1. | Supervised Learning. | 7 |
| 2. | Regression. | 8 |
| 2.1 | Linear Regression. | 8 |
| 3. | Data Description. | 10 |
| 4. | Train and Test. | 11 |
| 5. | Tools and Packages used | 12 |
| 6. | System Architecture | 13 |
| 7. | Source Code | 14 |
| 8. | Result | 19 |
| 9. | Conclusion | 20 |

**List of Figure**

2.1 Linear Regression Equation

2.2 Linear Regression Graph

5.1 Mean Squared Mean

7.0 System Architecture.

1. **Supervised Learning**

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

Supervised learning is classified into two categories of algorithms:

Classification: A classification problem is when the output variable is a category, such as “Red” or “blue” ,“disease” or “no disease”.

Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Here we have used Linear Regression

Supervised learning deals with or learns with “labelled” data. This implies that some data is already tagged with the correct answer.

**2.0 Linear Regression**

**Regression:**

Regression is essentially an applied mathematics approach to seek out the link between variables. In machine learning, this is often wont to predict the end result of an incident supporting the link between variables obtained from the data-set. Statistical regression is one kind regression employed in Machine Learning. Any statistical regression model is drawn as,

Output = coefficient1 + coefficient2 \* input.

For example, take into account prediction of weight supported height.

Using statistical regression,

Weight = a + b \* height

Having found a relationship between these two variables, the coefficients a and b are often detected, supporting that the load for a given height are often expected.

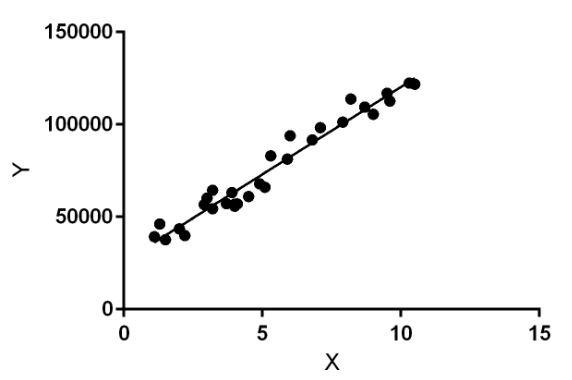
**2.1 Linear Regression:**

The most basic machine learning algorithmic rule that may be enforced on this information is regression. The regression model returns associate degree equation that determines the link between the independent variables and therefore the variable. The equation for regression may be written as:

2.1 Linear Regression Equation

Here, x1, x2,….xn represent the freelance variables whereas the coefficients θ1, θ2, …. θn represent the weights..

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.



2.2 Linear Regression Graph

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model

**3.0 Data Description.**

Opening price:

The opening price is the value that each share has when the S&P 500 stock exchange opens for trading. 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.

Highest/lowest price of the day

The highest and the lowest price of the day are taken the day before and gives an indication of how much the shares usually move during a day and how this in the end will affect the closing price. It also shows the general cyclical movement for each share.

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 the given period of time

market during a given period of time.”

**4.0 Train and Test:**

**Train:**

The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model*.* Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task. The type of training data that we provide to the model is highly responsible for the model's accuracy and prediction ability. It means that the better the quality of the training data, the better will be the performance of the model. Training data is approximately more than or equal to 60% of the total data for an ML project.

**Test:**

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset. However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

**Need for splitting the Data Set:**

Splitting the dataset into train and test sets is one of the important parts of data pre-processing, as by doing so, we can improve the performance of our model and hence give better predictability. We can understand it as if we train our model with a training set and then test it with a completely different test dataset, and then our model will not be able to understand the correlations between the features.

In this way, we can easily evaluate the performance of our model. Such as, if it performs well with the training data, but does not perform well with the test dataset, then it is estimated that the model may be overfitted.

**5.Accuracy**

**5.1 Mean Squared Error**

The Mean squared error (MSE) represents the error of the estimator or predictive model created based on the given set of observations in the sample. Intuitively, the MSE is used to measure the quality of the model based on the predictions made on the entire training dataset vis-a-vis the true label/output value. In other words, it can be used to represent the cost associated with the predictions or the loss incurred in the predictions. And, the squared loss (difference between true & predicted value) is advantageous because they exaggerate the difference between the true value and the predicted value. The lesser the MSE, the better the regression model is. When the linear regression model is trained using a given set of observations, the model with the least mean sum of squares error (MSE) is selected as the best model. The Python or R packages select the best-fit model as the model with the lowest MSE or lowest RMSE when training the linear regression models.

5.1 Mean Squared Mean

**5.2 R-Squared.**

R-Squared is the ratio of the sum of squares regression (SSR) and the sum of squares total (SST). Sum of Squares Regression (SSR) represents the total variation of all the predicted values found on the regression line or plane from the mean value of all the values of response variables. The sum of squares total (SST) represents the total variation of actual values from the mean value of all the values of response variables. R-squared value is used to measure the **goodness of fit or best-fit line.** The greater the value of R-Squared, the better is the regression model as most of the variation of actual values from the mean value get explained by the regression model.R-Squared is also termed as the **coefficient of determination.**

**6. TOOLS/PACKAGES AND PLATFORM USED**

In the proposed system python is used as the programming language in which several packages are imported such as

1. Numpy

2. Panda

3. Scikit

4. Matplotlib

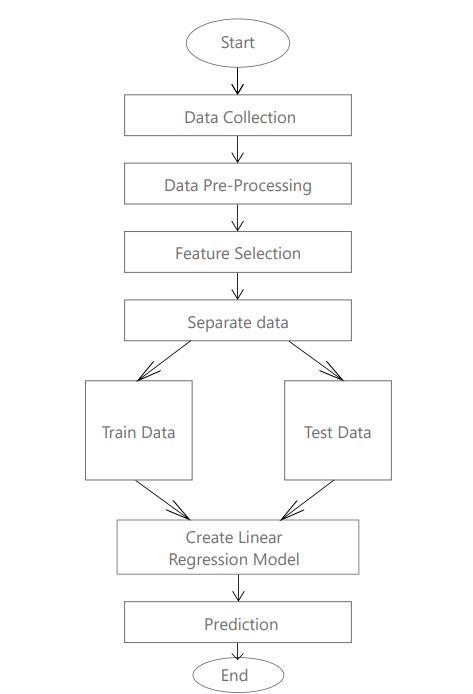
Here matplotlib is used for plotting the graph.

Numpy is used for algebraic calculations.

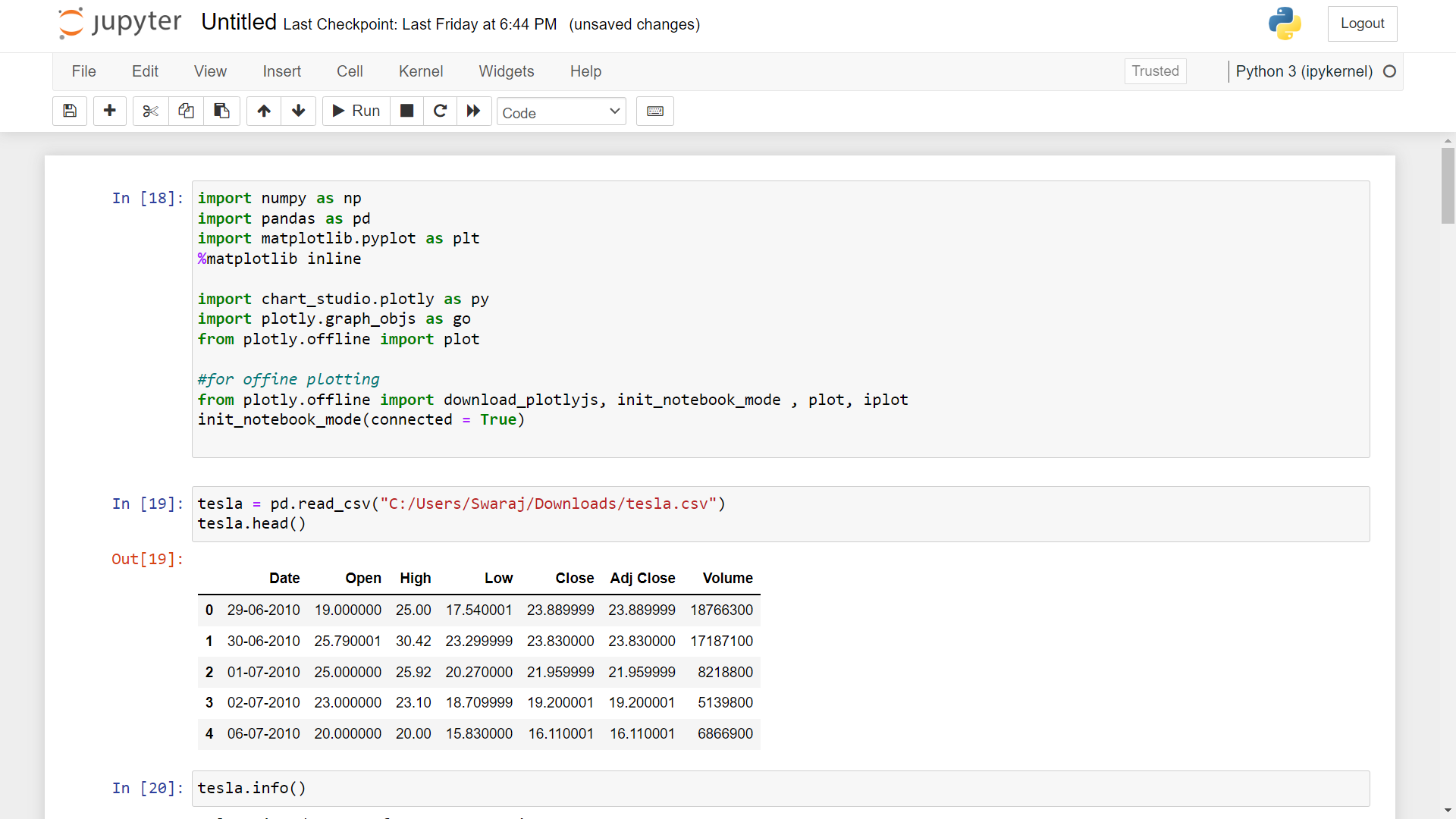
Pandas is a powerful data analytics and data manipulation tool.

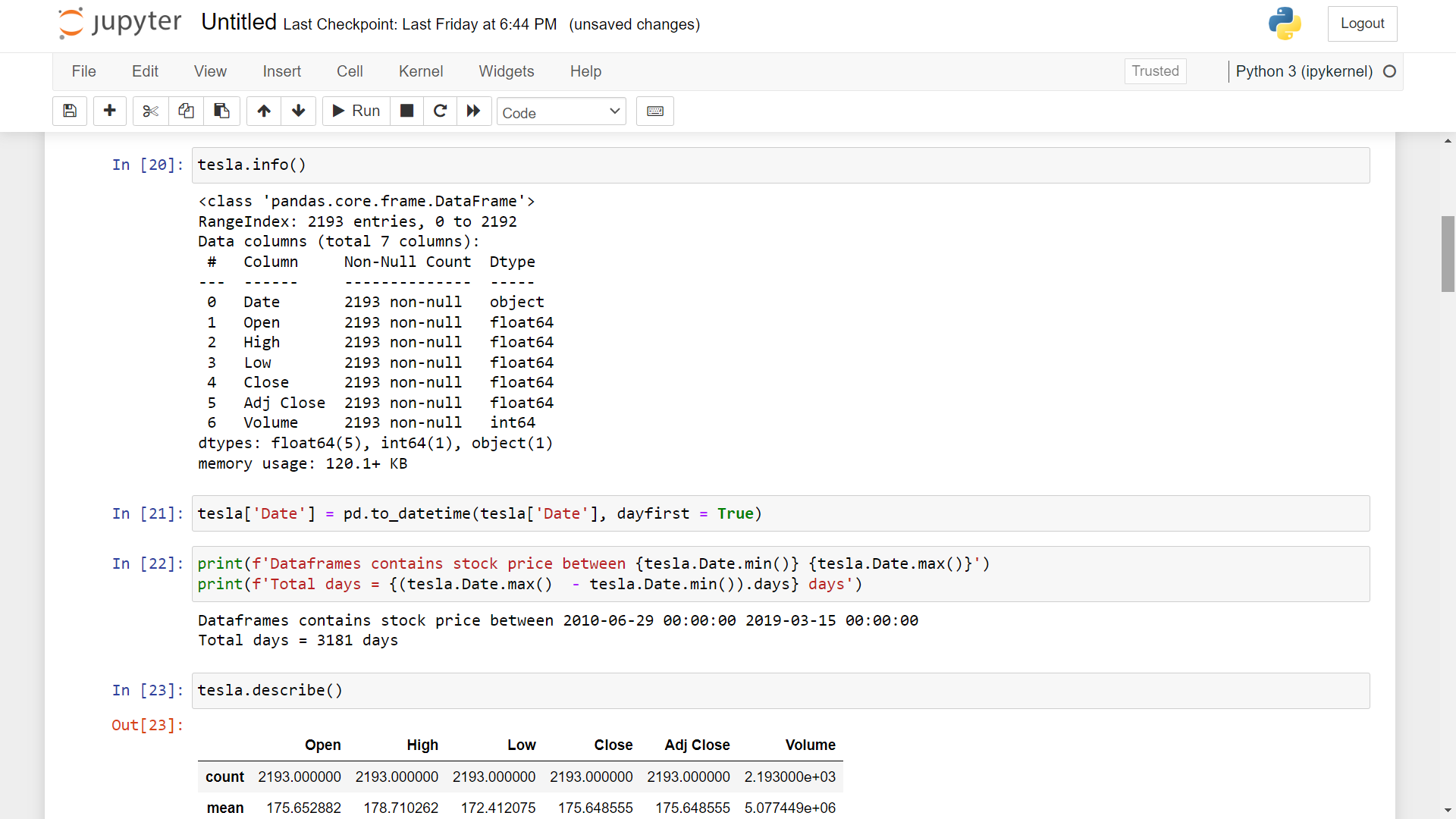
Scikit learn is used to import the linear regression and for splitting the data.

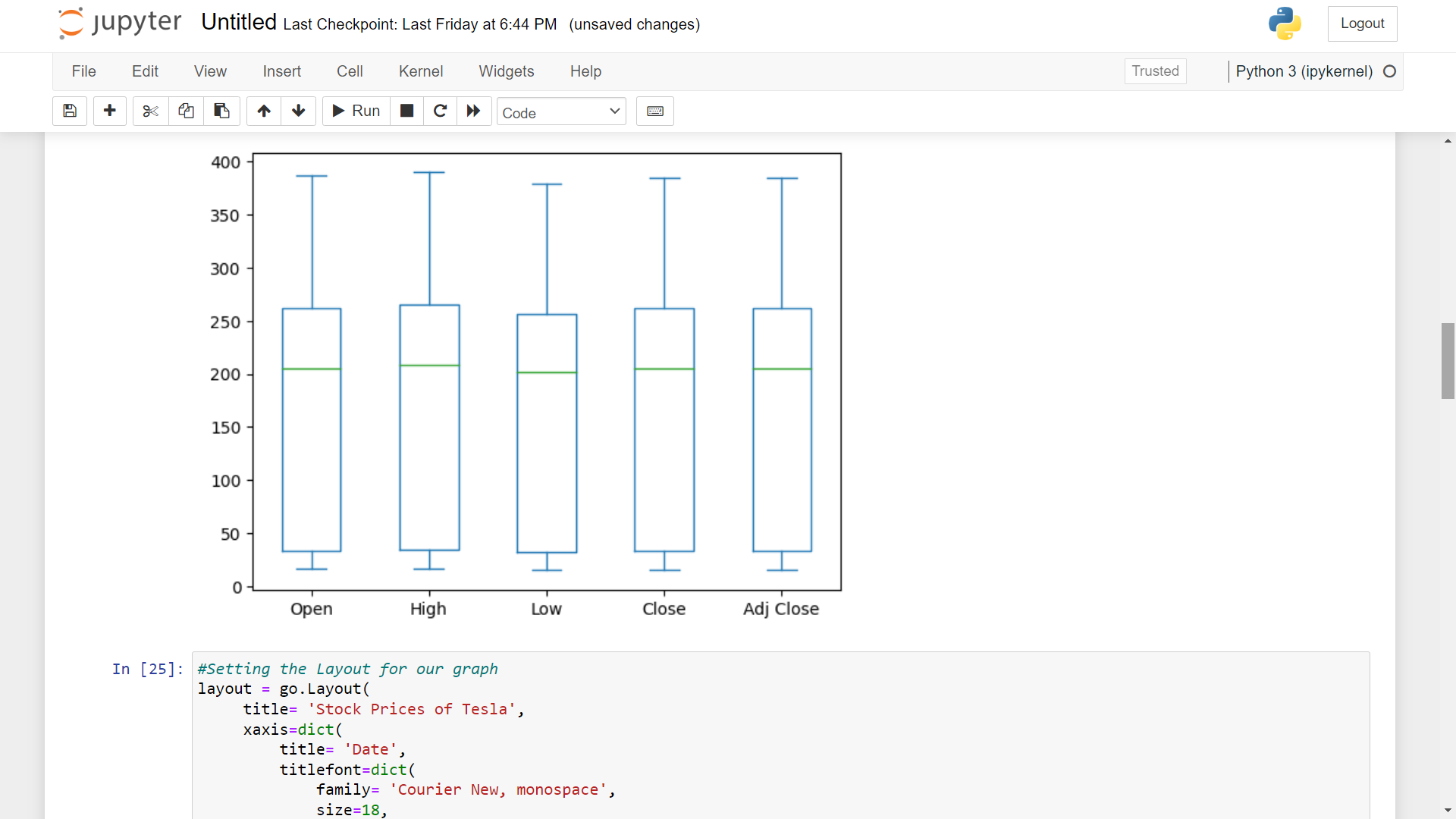
**7. SYSTEM ARCHITECTURE:**

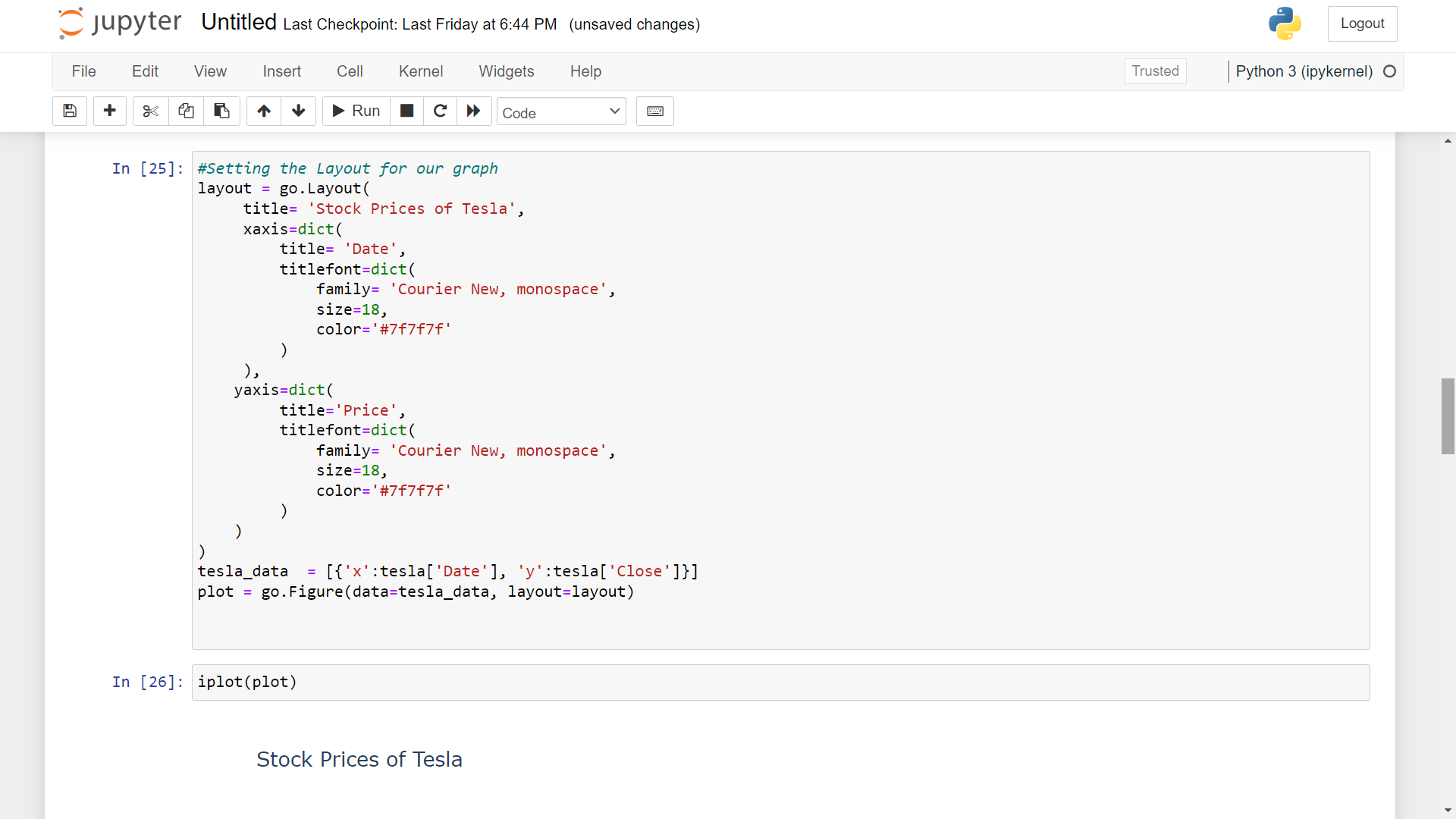
****

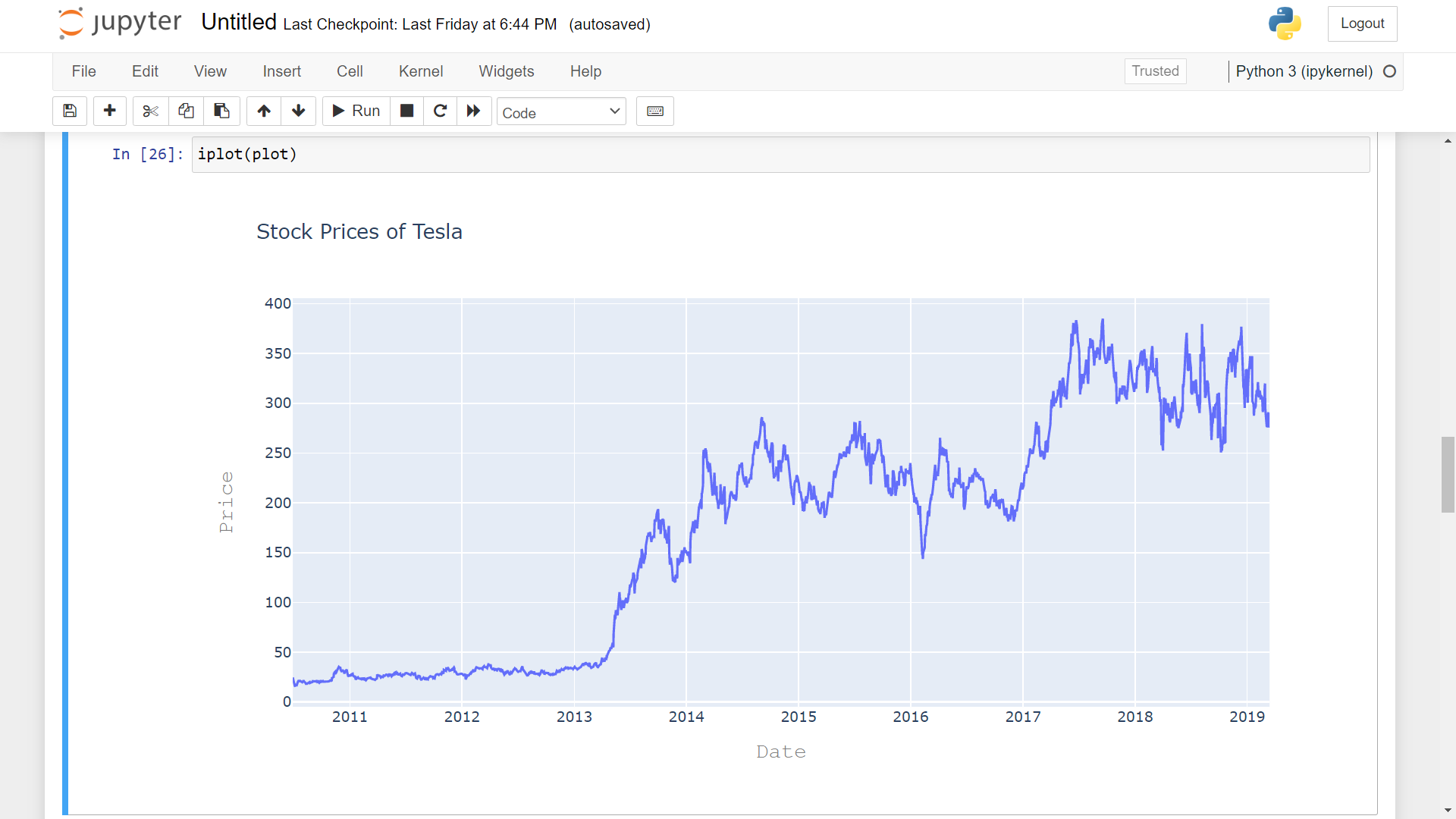
**8. SOURCE CODE :**

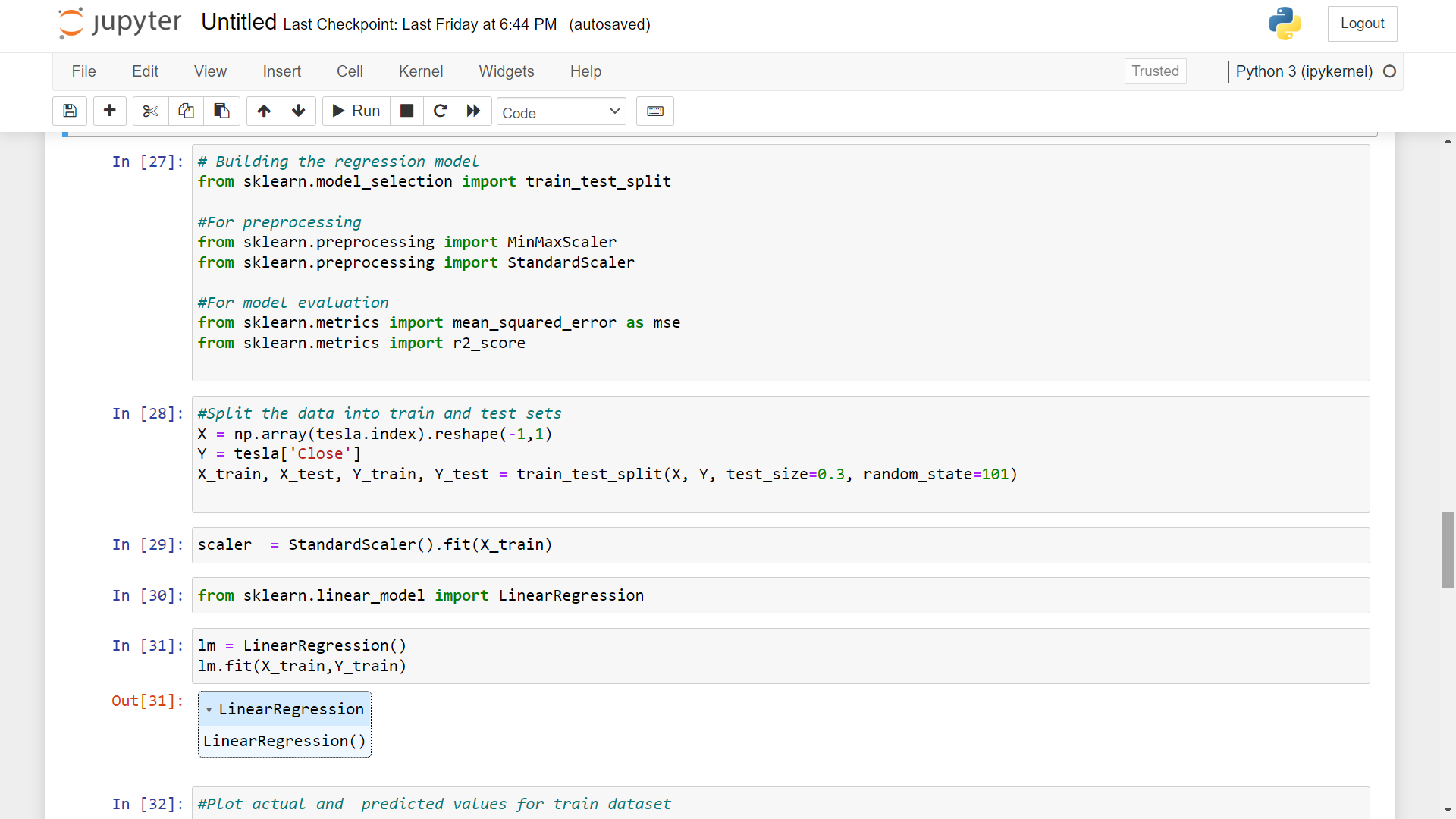
****

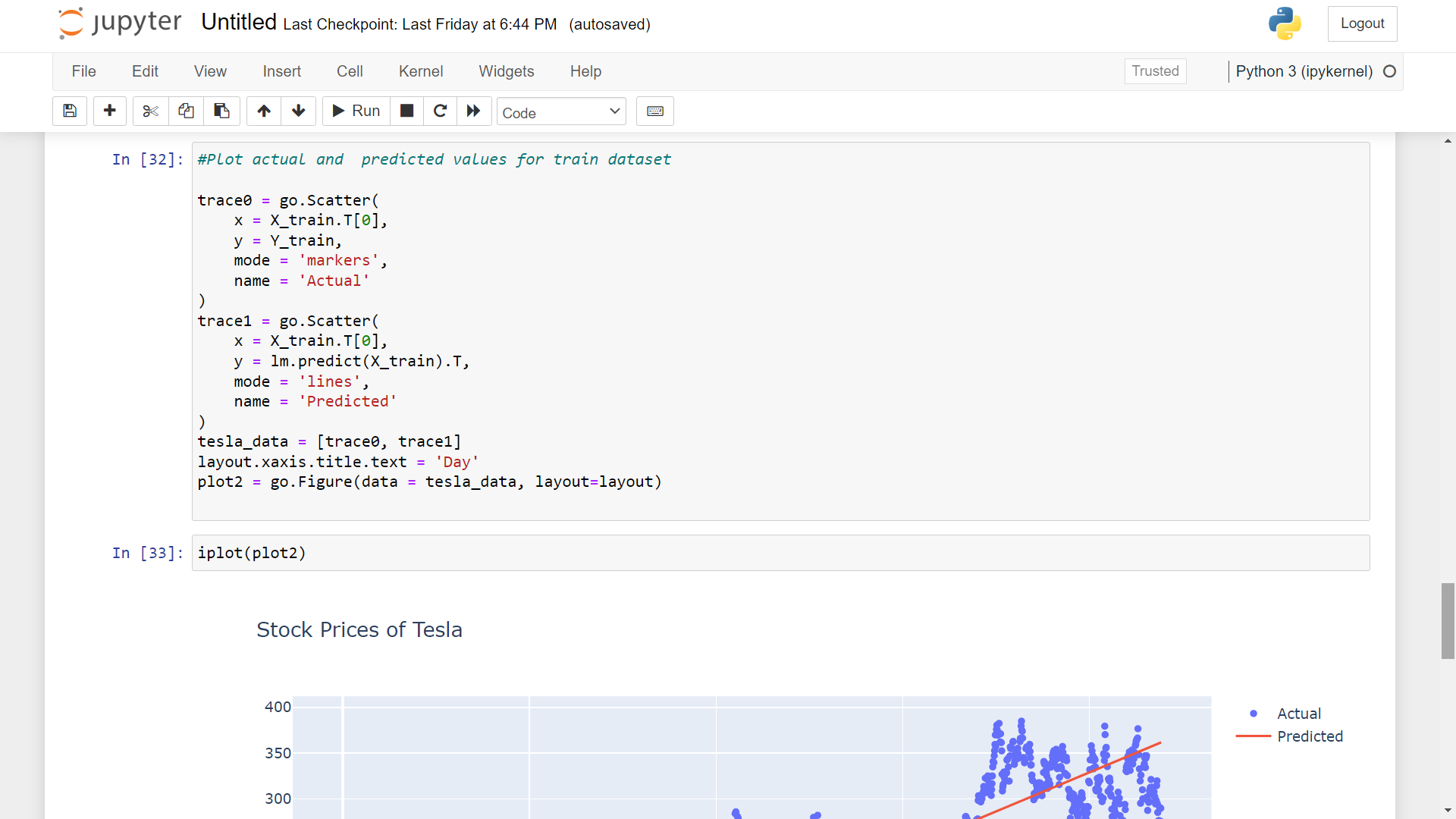
****

****

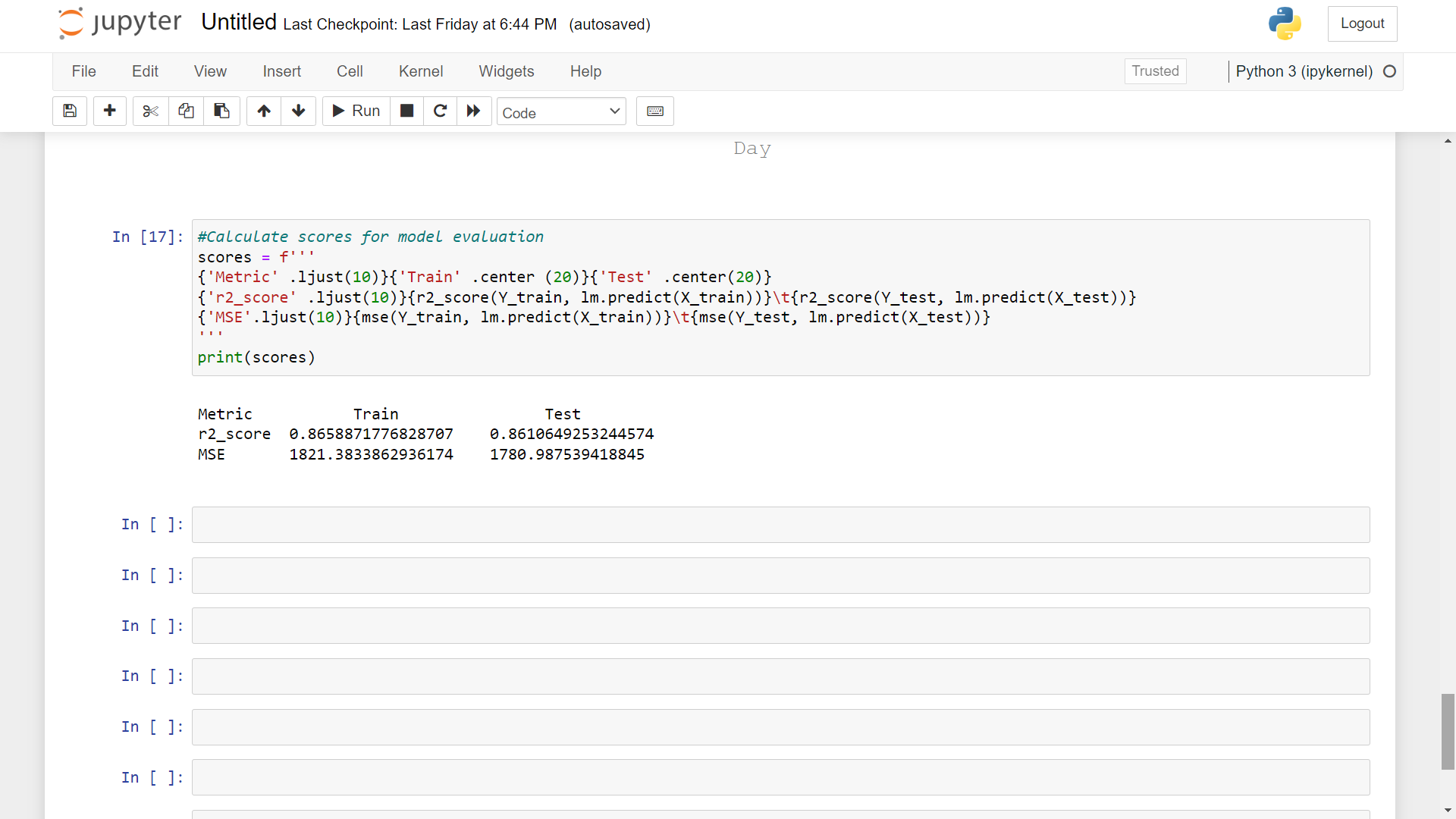
****

****

****

****

****

****

**9. OUTPUT:**

**R2\_score :**

Train - 0.8658871776828707

Test - 0.8610649253244574

**MSE:**

Train - 1821.3833862936174

Test - 1780.987539418845

**10. Conclusion :**

Thus, we implemented Linear Regression Algorithm to predict the stock price of Tesla using Supervised Learning. The R-Squared for train dataset came out to be 0.8658871776828707

And for the Test dataset the R-squared came out to be - 0.8610649253244574.