

# Housing Price Prediction

**Submitted By:**

**Sowmya Vedam**

**ACKNOWLEDGMENT**

This includes all the references, research papers, data sources, professionals and other resources that helped in completion of the project.

I’ve referred the DATATRAINED material for different concepts used for solving regression problems.

Also I’ve gone through some sample projects and solutions using google website.

# INTRODUCTION

House is a fundamental necessity for all human beings and so real estate sector plays a major role in the world’s economy. It is a very large market and so various companies are working in this domain. Here, **Data Science** has a huge role in this sector for the companies to channelize their efforts, detect any problems in their functioning, and in turn increase the overall revenue.

This problem is related to a housing company “**Surprise Housing**” situated in US. This company wants to enter into Australian market. This company uses data analytics to purchase houses at prices less than the actual prevailing rates and flip them at a higher price.

Housing prices depend upon so many variables such has the location of the house, its total area, garage properties, lawn area details etc. Here we try to build a model to understand how the prices vary with the variables. It helps in understanding the price dynamics of the new market.

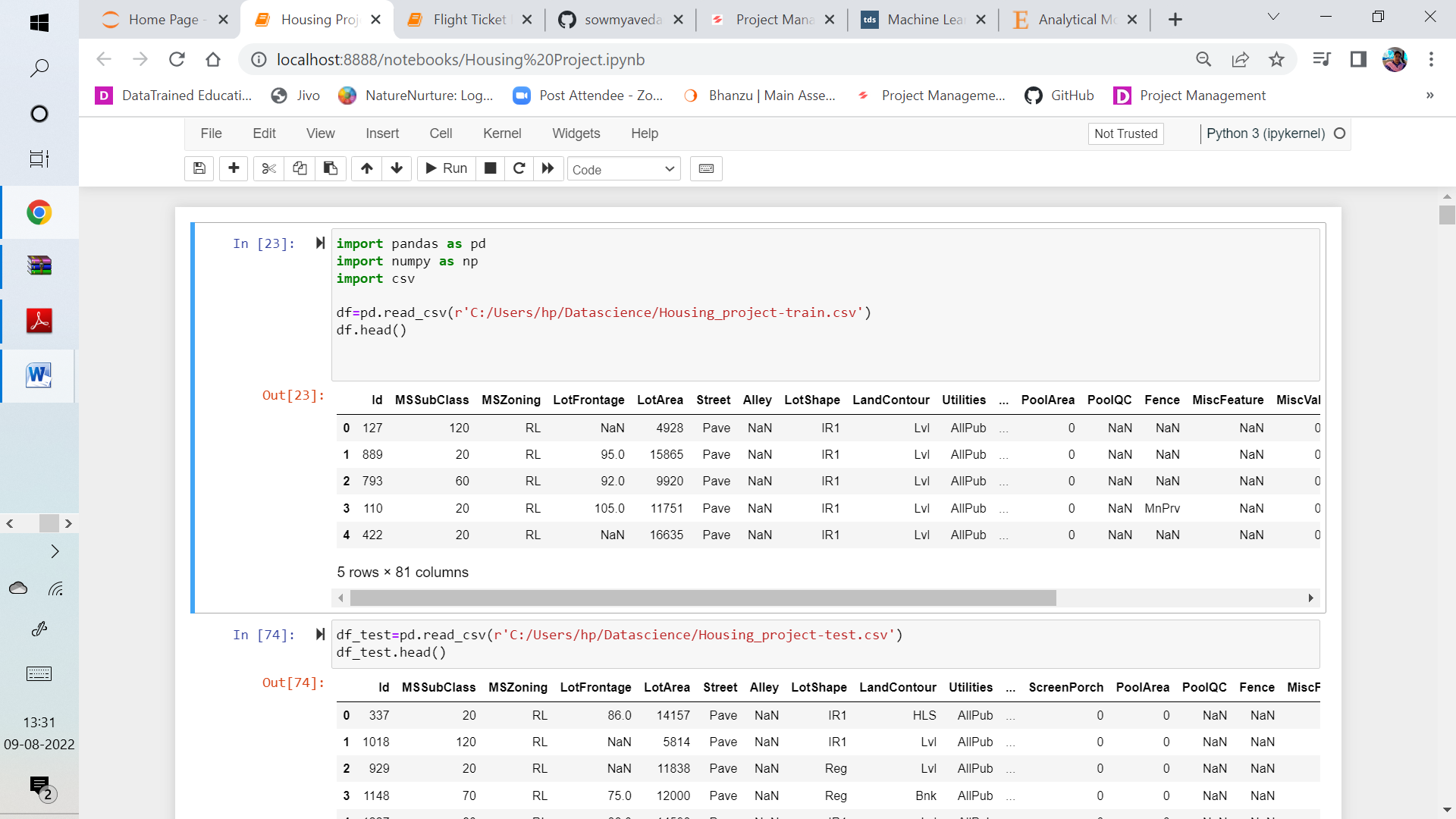
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This is a dynamic dataset where the Sale Price of the houses constantly varies depending on the variables in consideration. The Sale Price has a linear relation with almost all the variables and so it is a regression problem. We have to build a linear regression model in machine learning to predict the house prices.

* Data Sources and their formats

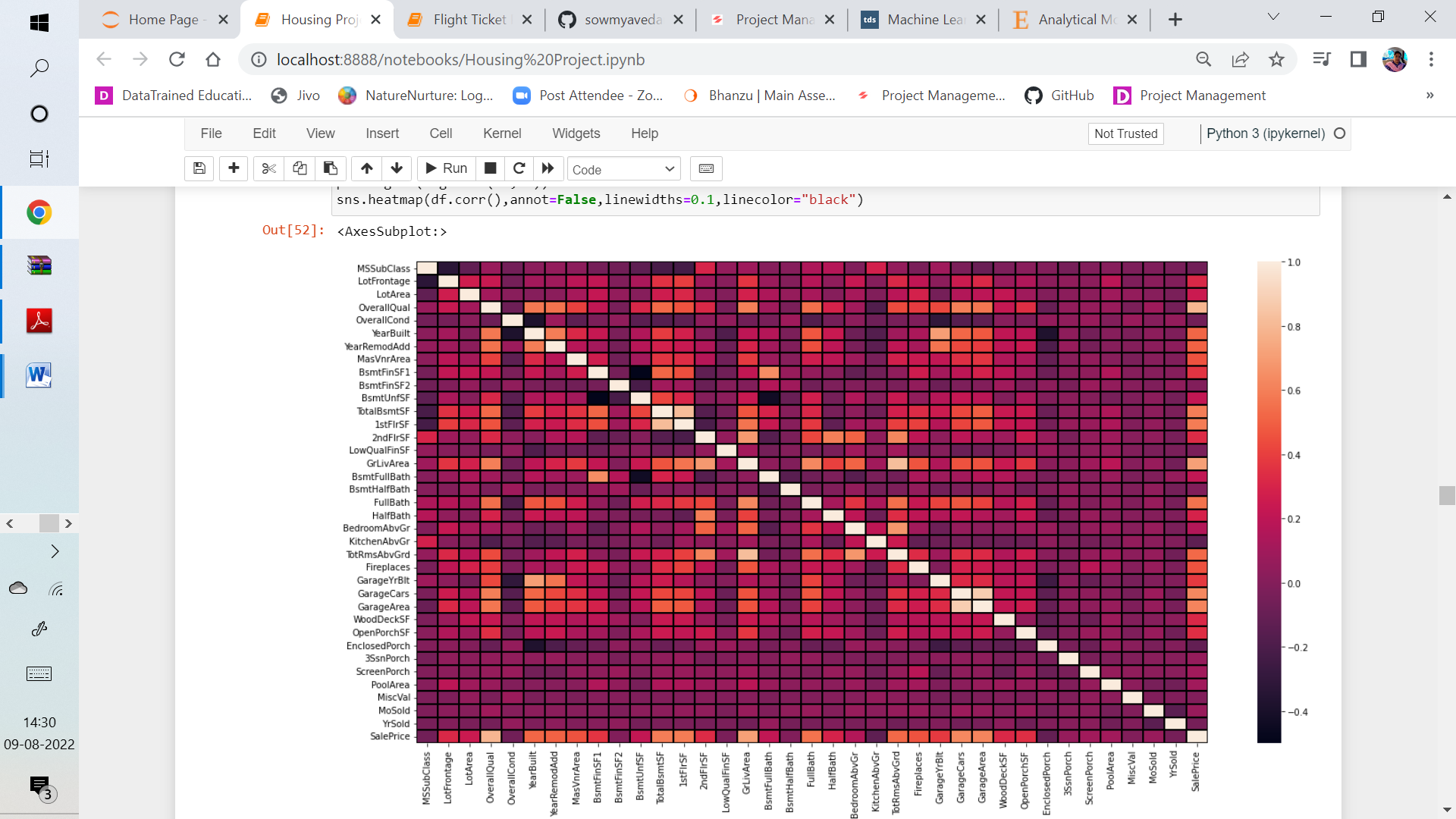
The data source of the project is given in **CSV** file format. It has a train and test data set provided separately. A model is built for prediction and the test data is provided so observe the prediction patterns.



* Data Pre-processing

The data preprocessing is done in a step by step process. First, the null values are checked for, and are replaced with the mode of the individual columns. “**ID”** column is removed, as it will not contribute to the output variable.

Then the data visualizations are done to view the relationship of the output or dependent variable. This is the correlation map.



* Data Inputs- Logic- Output Relationships

**Univariate Analysis:** All the categorical, nominal and ordinal data columns are plotted using count plots to view the data distribution.

All the continuous data columns here, it is the Output Variable, are plotted using distribution plots.

**Bivariate** **Analysis**: Here we can observe the relation or response of the dependent variable with the different input variables available in the data.

**Label Encoding** is done for the **object** type data columns so that the columns take part in the model building.

Using **PCA(Principal Component Analysis),** feature selection is done, where most of the features which have an impact on the output variable are selected.

* Hardware and Software Requirements and Tools Used

Coding is done using the Jupyter notebook from the Anaconda Package.

Software used to build the solution: Python, Numpy, Pandas.

**Model/s Development and Evaluation**

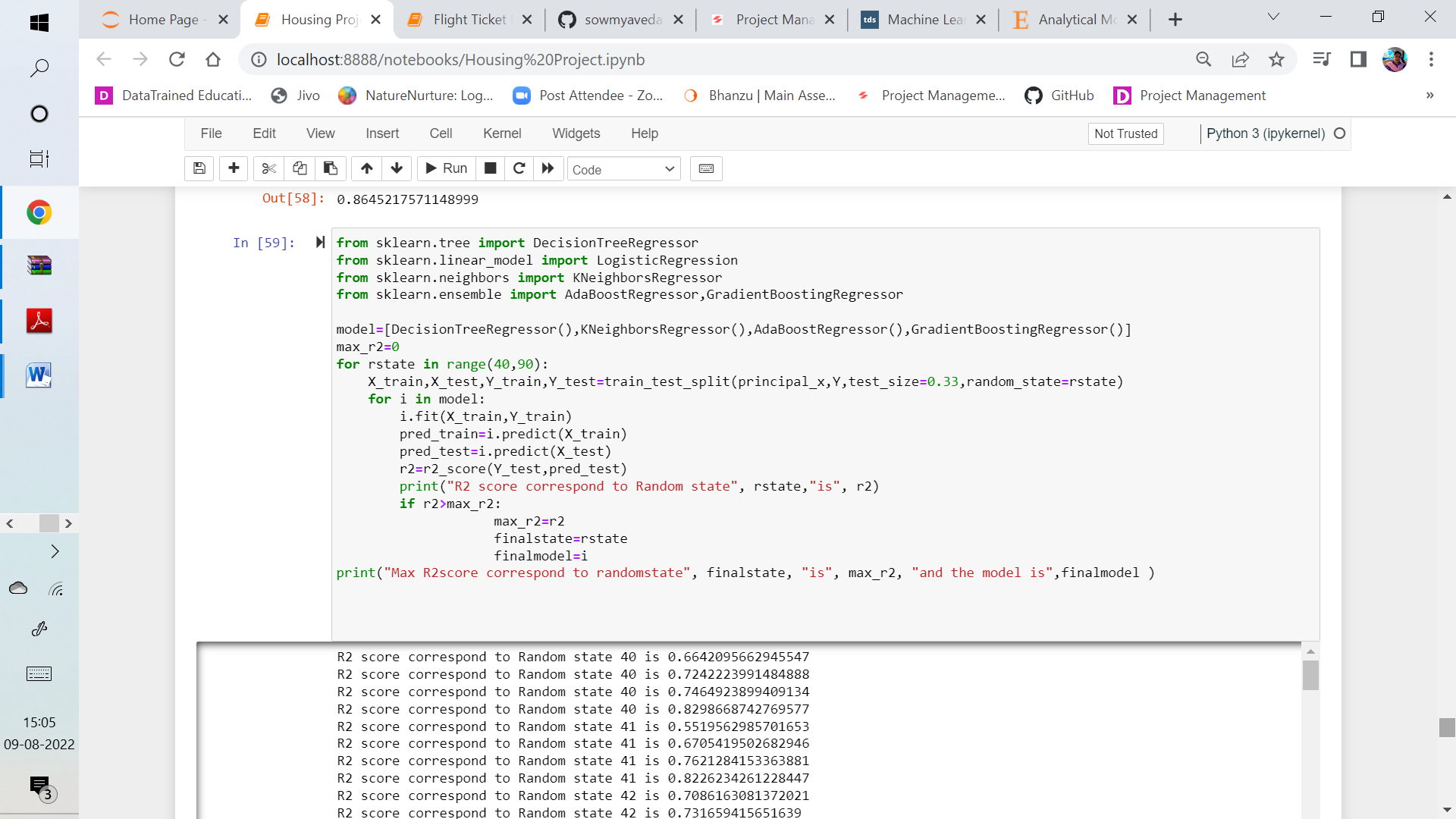
* Testing of Identified Approaches (Algorithms)

All the algorithms used for the training and testing:

* Linear Regression
* Gradient Boosting Regressor
* AdaBoost Regressor
* Decision Tree Regressor
* KNeighbours Regressor

Accuracy of each model is identified using the r2 score calculated for the model. Gradient Boosting Regressor has emerged with the highest accuracy score of 89%.

* Run and Evaluate selected models

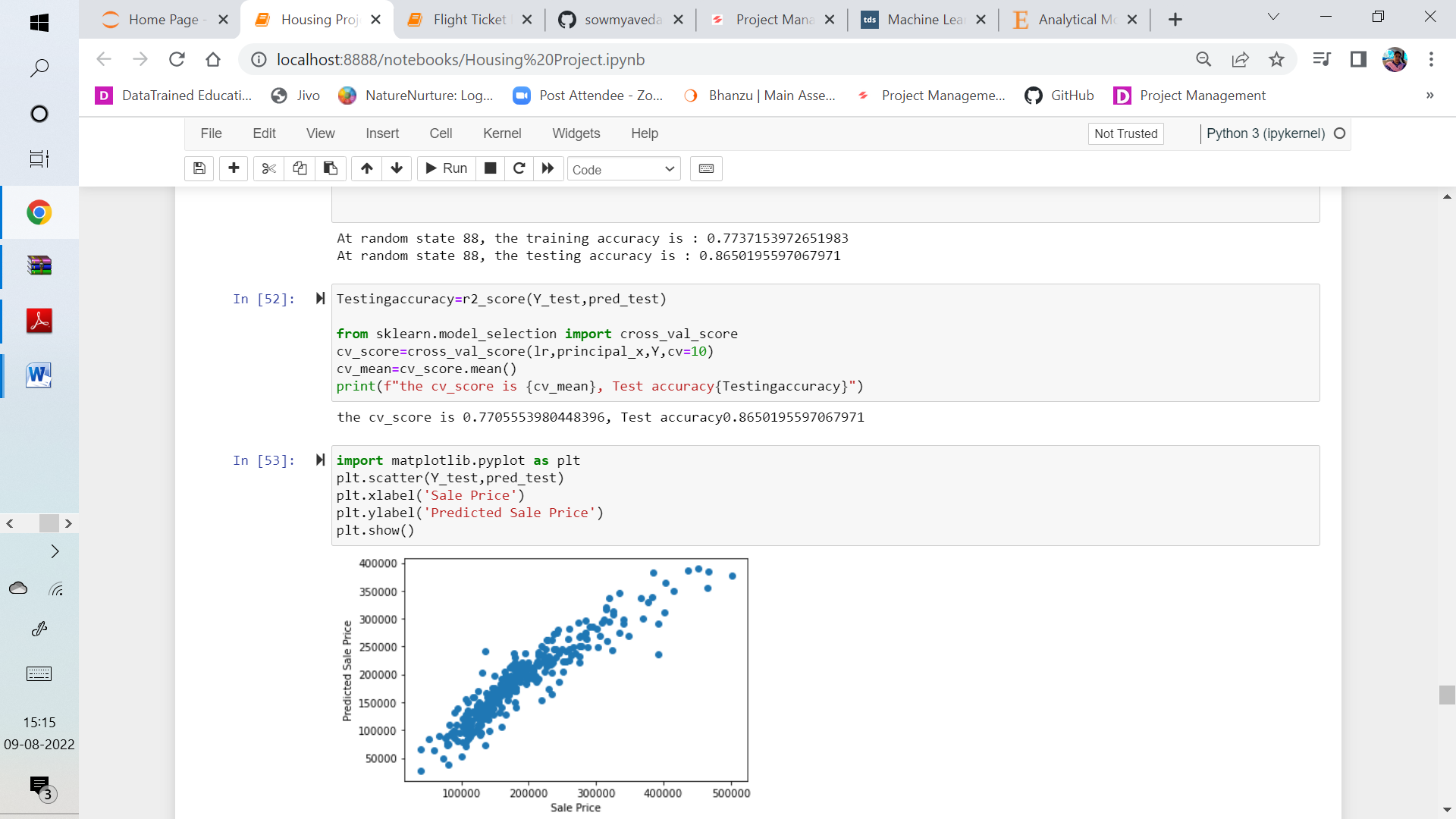


Over 89 % accuracy is achieved using Gradient Boosting Regression. Cross Validation of the model is done. Also Hyper-parameter tuning is done, to avoid over-fitting problem, if any. The model performed well and an accuracy of about 87% is achieved.

* Key Metrics for success in solving problem under consideration

**R2 score** is the key metric used for evaluating a machine learning model for this regression model.

* Interpretation of the Results



**CONCLUSION**

With the Data Visualization techniques in pandas, seaborn, etc, we could identify the data distribution of the different data columns and identify any correlation in between the features. With the Feature Selection techniques we could select the features that most effect the output variable.

Thereby, the accuracy score is obtained for each model and the model which gives the best score is being selected in building a model.

This model is saved as a pickle, by importing the pickle library.