Report On

**Housing Prices Prediction**

Machine Learning with Python Mini Project

Submitted by:

**Vithan A Rai |** 18030141CSE064

**Sachin Rampur |** 18030141CSE068

**Abhishek Kuntappalavar |** 18030141CSE069

Under the Guidance of (Faculty Guide)

**Prof. RADHA R**

in Semester VI of the Bachelor of Technology

Batch - 2018-2022



**ABSTRACT**

This project is essentially used to predict the housing prices based on the data in the ‘Boston Housing’ data set. We use various Python libraries to fit, analyze and predict the outcomes desired to us (in this case the housing prices). We go through the concepts of linear regression and multilinear regression to get the predicted values. We check the linearity between each independent variable and dependent variable present in the data set by plotting them. We then use specific Python libraries and their functions to perform Multilinear regression on the data set, which is then used to predict the price of a house accordingly. At the end we use a special function to get the summary of the entire process. This model has been created to predict the prices based on a very small data set containing less than five hundred elements in it. It may not be ideal for a much larger data set, as such large sets of data may contain large number of outliers to which a primitive model like this is very sensitive and may lead to various errors.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapters** | **Topics** | **Page** |
| **1** | Introduction | 4 |
| **2** | Project profile | 5 |
| **3** | Observation / Analysis | 6-11 |
|  | Conclusion | 12 |
|  | References | 13 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig no** | **Topic** | **Page** |
| 1.1 | Overview of the Dataset | 6 |
| 1.2 | Statistics of the dataset | 7 |
| 1.3 | Covariance matrix | 8 |
| 1.4 | MEDV Vs RM plot | 9 |
| 1.5 | MEDV Vs PTRATIO plot | 9 |
| 1.6 | MEDV Vs LSTAT plot | 10 |
| 1.7 | Intercept and Coefficients | 10 |
| 1.8 | Final output | 11 |

**CHAPTER 1**

**1.1 INTRODUCTION**

The rapid growth in the size of data is tremendous. And we now have an opportunity of analyzing this data and bringing out various desired outcomes in any field feasible to the human mind. With the amount of data available to us now, we can apply various statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data, find hidden patterns and relationships in the data. The application of such statistical algorithms that was not possible in the past is made possible by the rapid advancements in the field of science & technology and by the people who are willing to work on this huge amount of data using their vast skill sets.

We use the method of predictive analysis on various subjects such as predicting stock trends, fraud detection, recommendations on various platforms, customer specific advertisements and so on. In this project, we will develop and evaluate the performance of the predictive power of a model trained and tested on data collected from houses in Boston's suburbs. Once we get a good fit, we will use this model to predict the monetary value of a house located in Boston’s area. You can find the entire code, data set used in the project and the linear regression plots on this link : [https://github.com/Vithanrai/MLP-project](https://github.com/Vithanrai/MLP-project%20)

**CHAPTER 2**

**2.1 PROJECT PROFILE**

**2.1.1 OBJECTIVES OF THE STUDY**

The objective of this project is to understand the concept of Multilinear regression and to implement it in a simple machine learning model using Python and libraries like sklearn and pandas.

**2.1.2 PREREQUISITES**

The software requirements for this project are

* Python
* A suitable Python IDE (jupyter is recommended)
* Python libraries -> sklearn, pandas, numpy and matplotlib.
* A suitable data set (in our case housing prices dataset).

**CHAPTER 3**

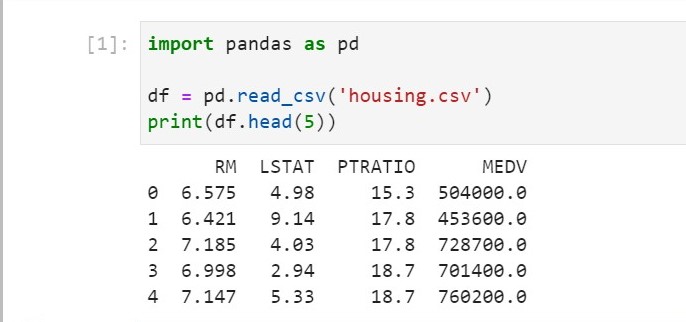
**3.1 OBSERVATION /ANALYSIS**

**3.1.1 LOADING DATA**

The project code begins with loading the data from the csv file ‘housing.csv’ into the data frame. We print the data frame and check if the csv file has been loaded correctly. The csv file contains 4 variables and 489 data points. Among the 4 variables, 3 are independent variables and 1 is a dependent variable. The 3 independent variables are as such ['RM', 'PTRATIO', 'LSTAT'] , and the dependent variable is ['MEDV'].

* RM: This is the average number of rooms per dwelling.
* PTRATIO: This is the pupil-teacher ratio by town.
* LSTAT: This is the percentage lower status of the population.
* MEDV: This is the median value of owner-occupied homes in $21000s

This is an overview of the dataset, with its original features:

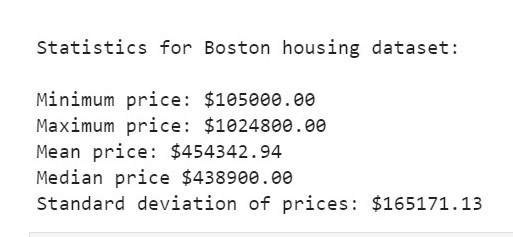


*Fig 1.1 overview of dataset*

* The features ‘RM’, ‘LSTAT’ and ‘PTRATIO’, give us quantitative information about each datapoint.
* The target variable, ‘MEDV’, will be the variable we seek to predict. We will store it in prices.

**3.1.2 DATA EXPLORATION**

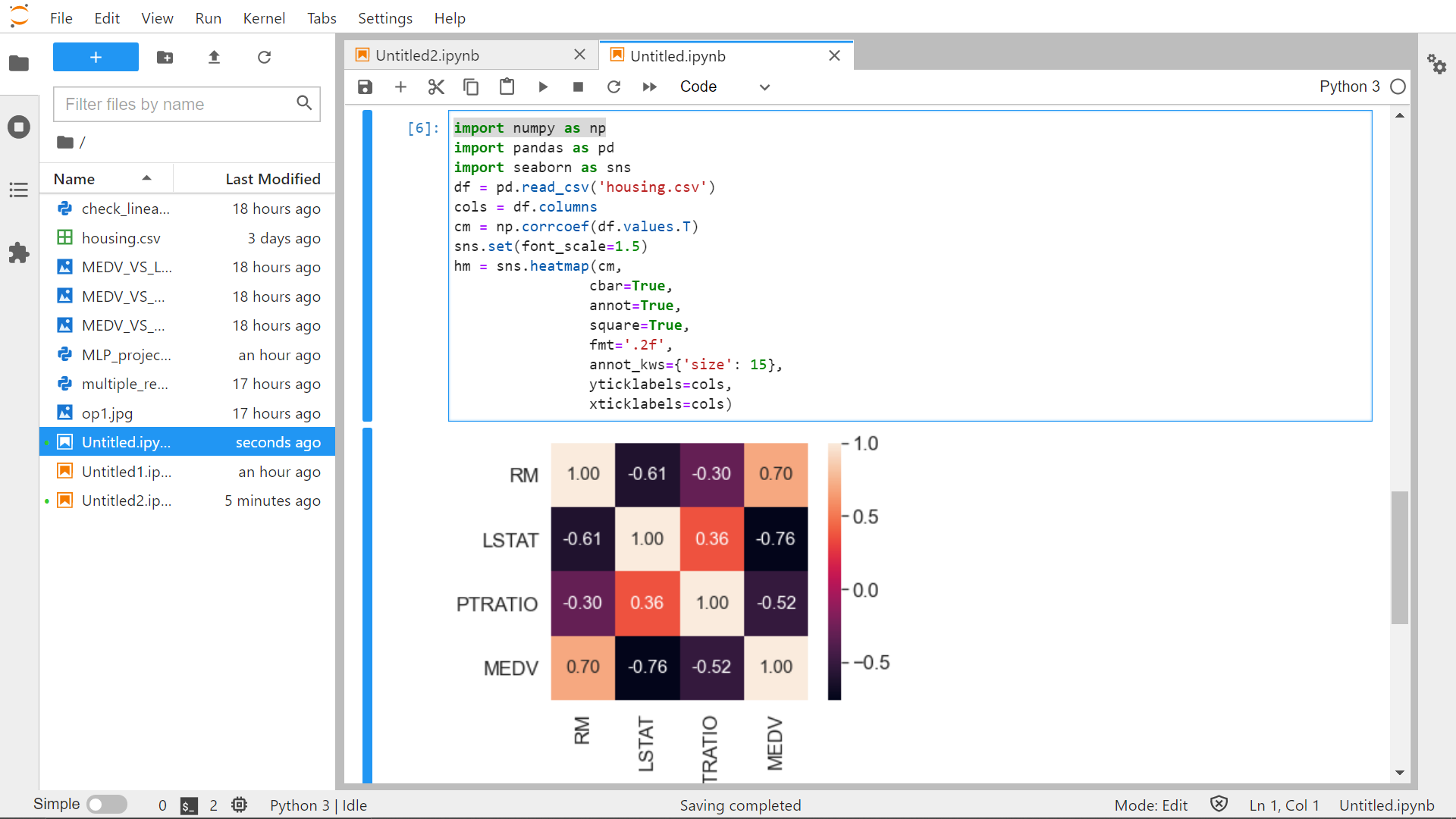
In the first section of the project, we will make an exploratory analysis of the dataset and provide some observations. By calculating statistics of the data we get this output.



*Fig 1.2 statistics of the dataset*

**3.1.3 CORRELATION MATRIX**

We are going to create now a correlation matrix to quantify and summarize the relationships between the variables. This correlation matrix is closely related with covariance matrix, in fact it is a rescaled version of the covariance matrix, computed from standardize features. The image shown below is the correlation matrix of the data set.



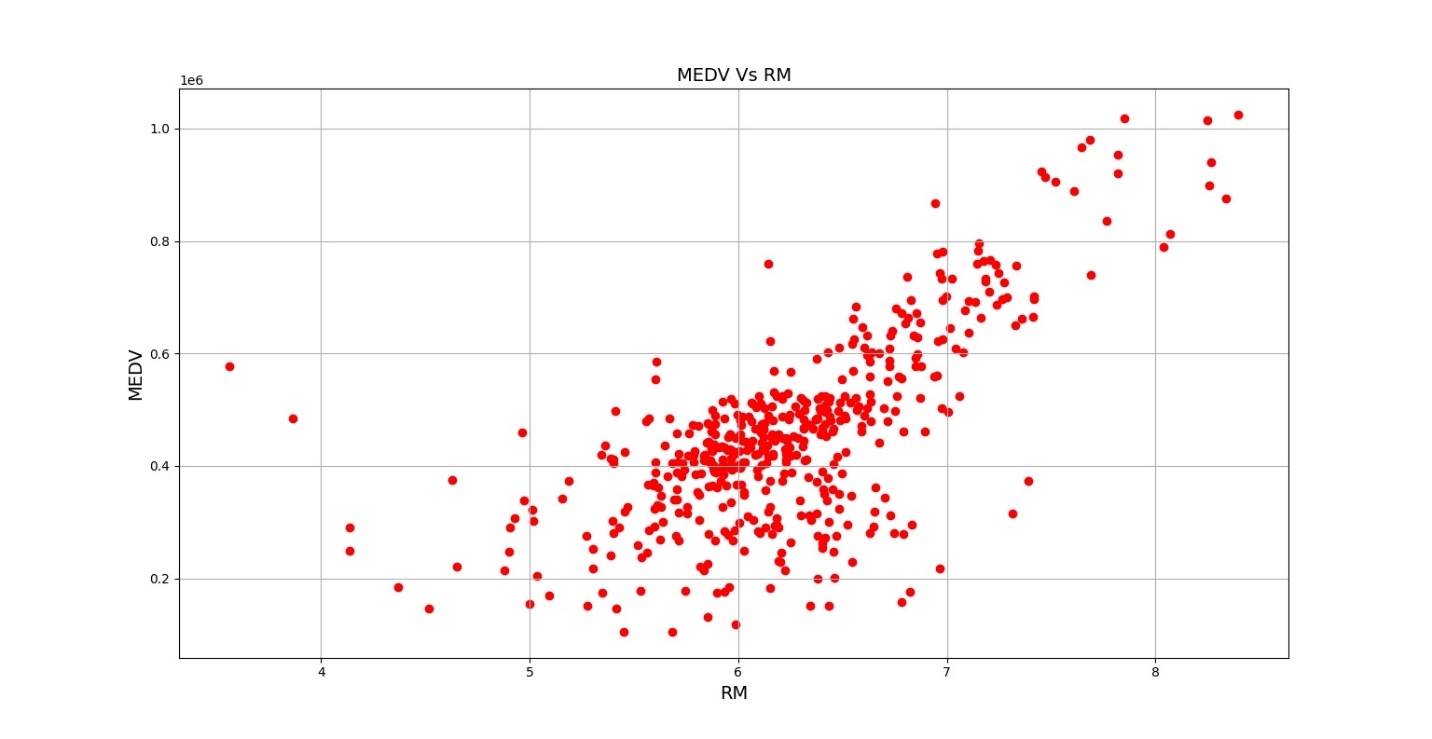
*Fig 1.3 covariance matrix*

**3.1.4 CHECKING FOR LINEARITY**

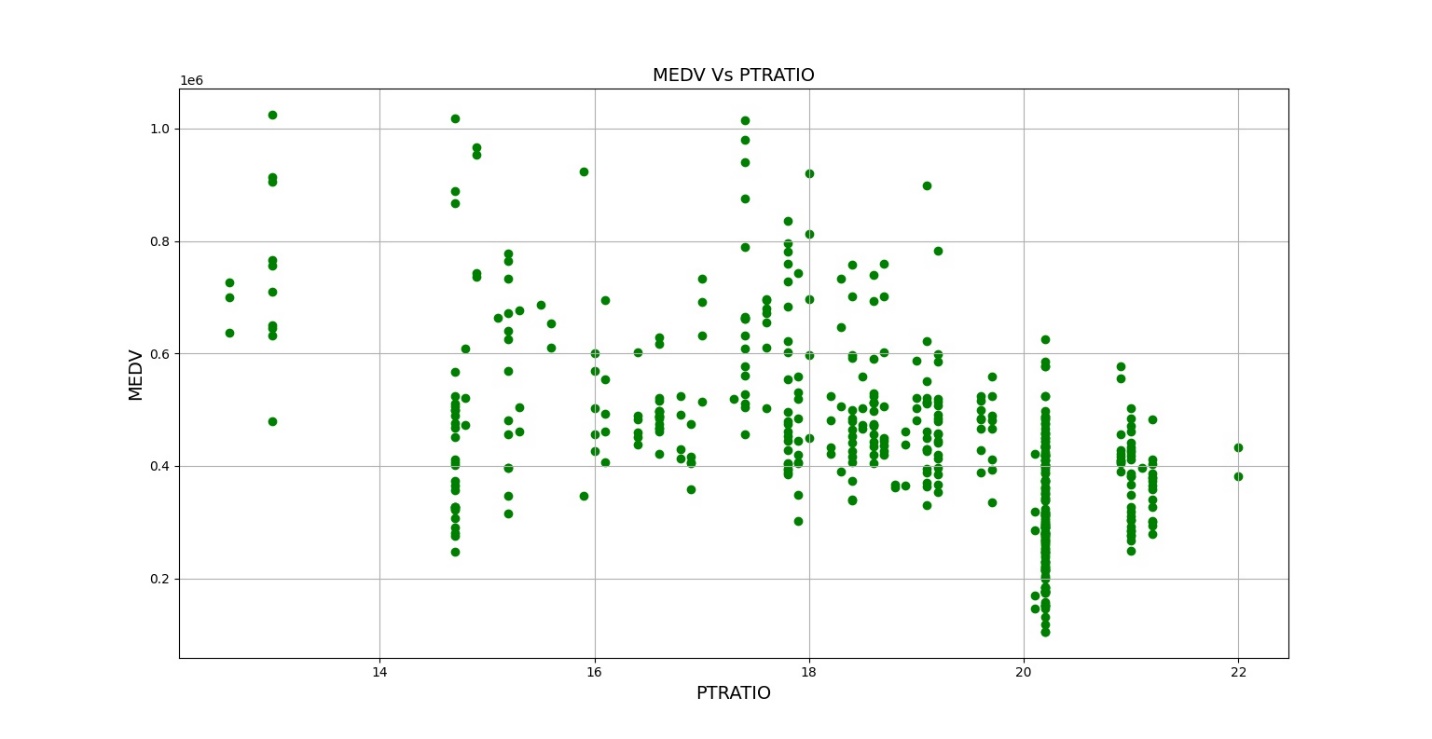
Before we execute a linear regression model, it is recommended to validate that certain assumptions are met. Hence we have to check that a linear relationship exists between the dependent variable and the independent variable/s. In our example we will check that a linear relationship exists between the:

* MEDV(dependent variable) and RM(independent variable)
* MEDV(dependent variable) and PTRATIO(independent variable)
* MEDV(dependent variable) and LSTAT(independent variable)

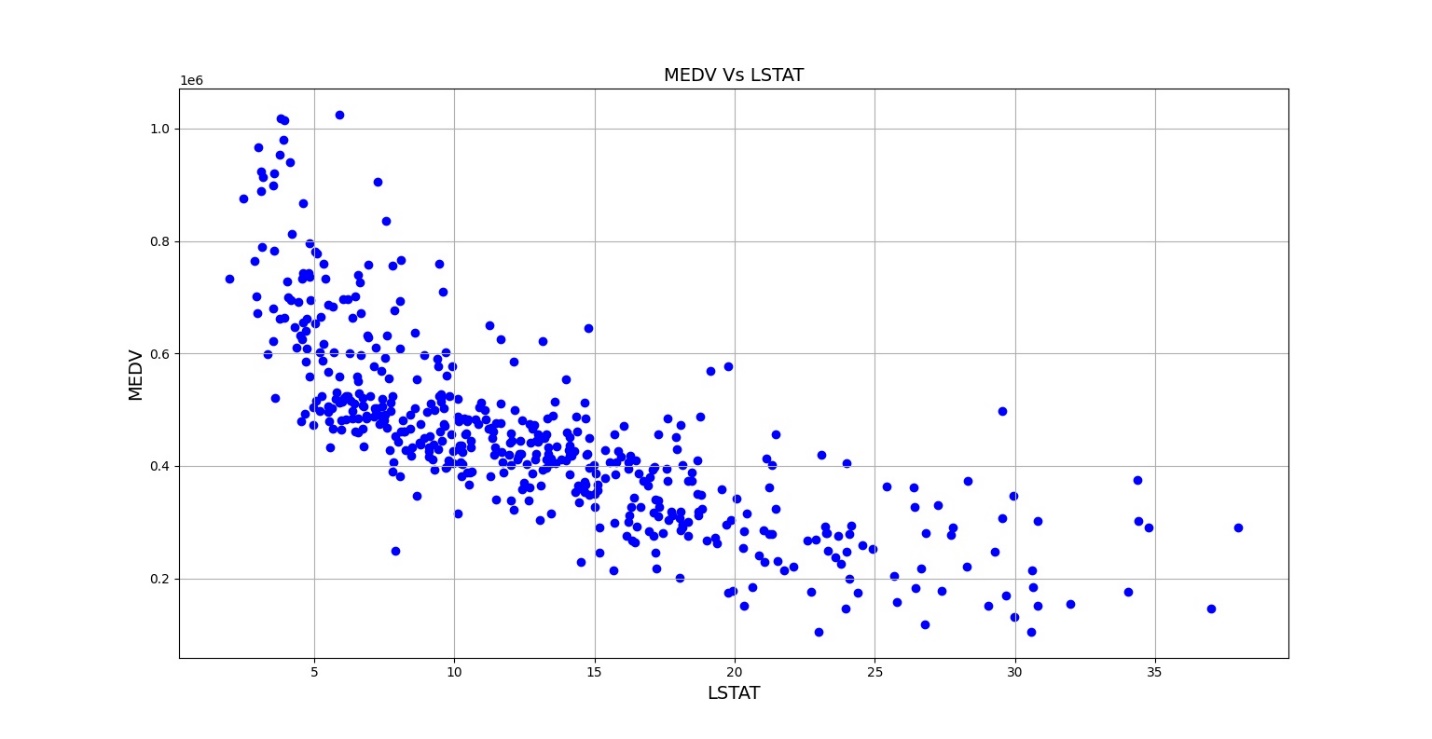
To perform a linearity check we will use scatter diagrams utilizing the *matplotlib* library.



*Fig 1.4 MEDV vs RM plot*



*Fig 1.5 MEDV vs PTRATIO*



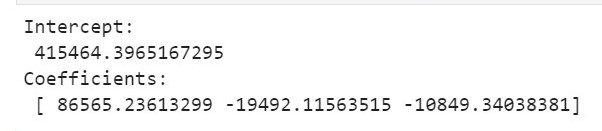
*Fig 1.6 MEDV vs LSTAT*

We notice that a linear relationship exists between the dependent and independent variables.

Next, we are going to perform the multilinear regression using Python.

**3.1.5 PERFORMING THE MULTILINEAR REGRESSION**

Once we have added the data into Python we will use the sklearn library to get the regression results. After running the python code we get the ‘Intercept’ and ‘Coefficient’ values as shown in the figure below.



*Fig 1.7 Intercept and Coefficents*

We can use the ‘Intercept’ and ‘Coefficients’ to build the multilinear regression equation as follows:

House\_Price = (Intercept) + (RM\_coeff)\*X1 + (PTRATIO\_coeff)\*X2+ (LSTAT\_coeff)\*X3

After inserting the corresponding values we get the below equation:

House\_Price = (415464.3965167295) +

(86565.23613299)\*X1 +

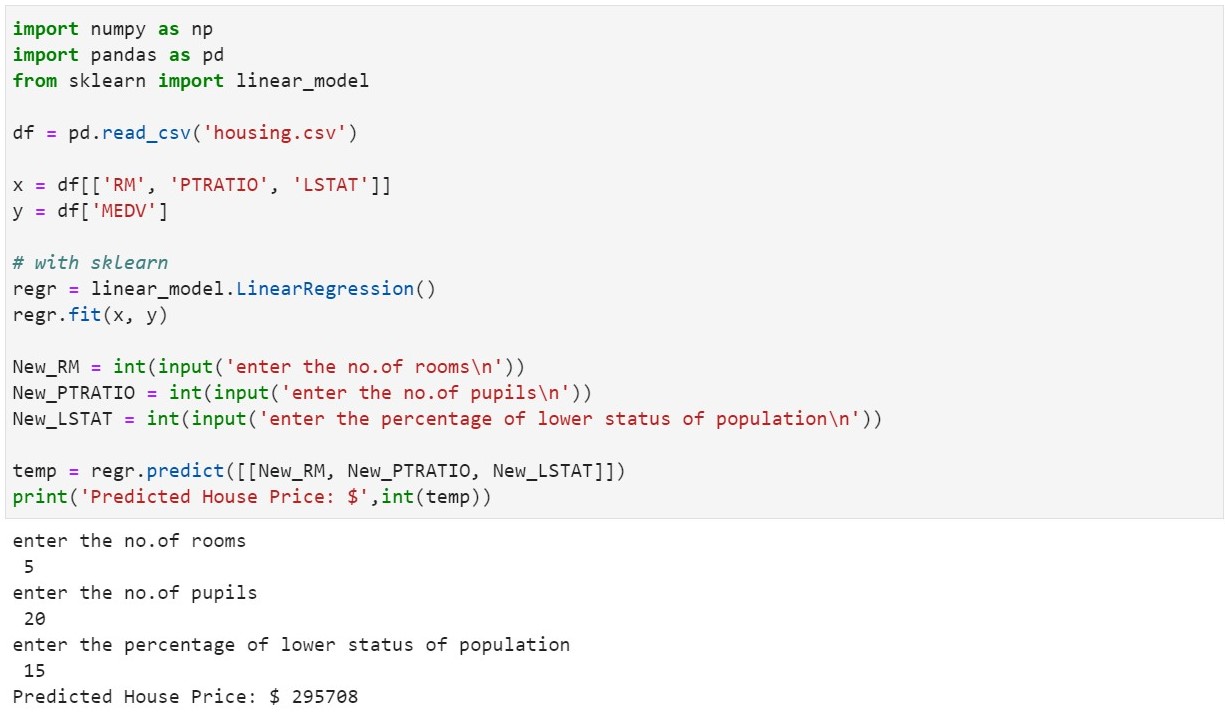
(-19492.11563515)\*X2+

(-10849.34038381)\*X3

After we get the above values, we see a prompt message which asks the user to input the

1. Number of rooms
2. Number of pupils and
3. Percentage of lower status of population

The below figure demonstrates the same



*Fig 1.8 Final Output*

**CONCLUSION**

We have successfully implemented the Multivariate regression and developed this project for predicting the housing prices based on the ‘Boston Housing Dataset’.

We have been successful in our attempt to predict the housing prices accurately. This model can be used to by a realtor as an everyday tool to help in his/her profession. Tools like these are being developed everyday by hundreds throughout the world. Such tools are the proof that machine learning is indeed the next big thing.

**REFERENCES**

<https://datatofish.com/multiple-linear-regression-python/>

<https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d>

<https://github.com/rromanss23/Machine_Leaning_Engineer_Udacity_NanoDegree/blob/master/projects/boston_housing/housing.csv>

Textbook:

Joel Grus, (2019). Data Science from Scratch. O’ReillyR