**House Price Prediction Project**

**Abstract:**

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are some factors that influence the price of a house which include number ofbedrooms,number of bathrooms,conditions,sqft\_living,sqft\_lot and etc. This research aims to predict house prices in INDIA based on REGRESSION.

**Keywords:**

Kingcountry House Price Prediction ,Pandas, NumPy, Matplotlib, Exploratory Data Analysis (EDA), Data Cleaning,Outlier detection, Dimensionality Reduction, Data Visualization,Multiple linear Regression,Polynomial Regression,Random forest etc.

**Objective:**

The objective was to forecast the price of a specific house based on market pricing while accounting for various "features”.

**Applications:**

1.Real Estate

2.Own House Dream

**Description:**

By observing the data, we can know that the price is dependent on various features like bedrooms, bathrooms, sqft\_living, sqft\_lot, Year built etc. The price is also dependent on the location of the house where it is present. The other features like waterfront, view are less dependent on the price. Of all the records, there are some missing values, which we have to remove so this helps us creating better model.

**Dataset:**

There are 21,613 observation and 21 variables in our dataset. Let's take a closer look at what each variable name represents:

|  |  |
| --- | --- |
| Variable | Description |
| Id | Unique identifier of the house |
| Date | Date of sale |
| Price | Sell price |
| Bedrooms | Number of bed rooms |
| Bathrooms | Number of bathrooms. Noninteger values exist due to "1/2 bathrooms" and "3/4 bathrooms" |
| Sqft\_live | Size of interior space in square fee |
| Sqft\_long | Size of land lot in square feet |
| Floors | Number of floors. Noninteger values exist due to "half floor" architecture |
| Waterfront | '1' if property has a waterfront, '0' if not |
| View | An index from 0 to 4 of how good the property's view is |
| Condition | Condition of the house, ranked from 1 to 5, 5 being the greatest condition |
| Grade | Classification by construction material and worksmanship quality. Numeric scale with higher numbers being better. For more information see the King County glossary |
| Sqft\_above | Square feet above ground |
| Sqft\_below | Square feet below ground |
| Yr\_built | Year bulit |
| Yr\_renov | Year renovated. '0' if never renovated |
| Zipcode | 5 digit zip code |
| Long | Longitude |
| Lat | Latitude |
| Squft\_liv15 | Average size of interior space for closest 15 houses, in square feet |
| Squft\_lot15 | Average size of land lot for closest 15 houses, in square feet |

**Libraries:-**

**Pandas:**

Pandas is mainly used for data analysis and associated manipulation of tabular data in DataFrames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel.

**Numpy:**

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of highlevel mathematical functions to operate on these arrays.

**Matplotlib:**

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython etc.

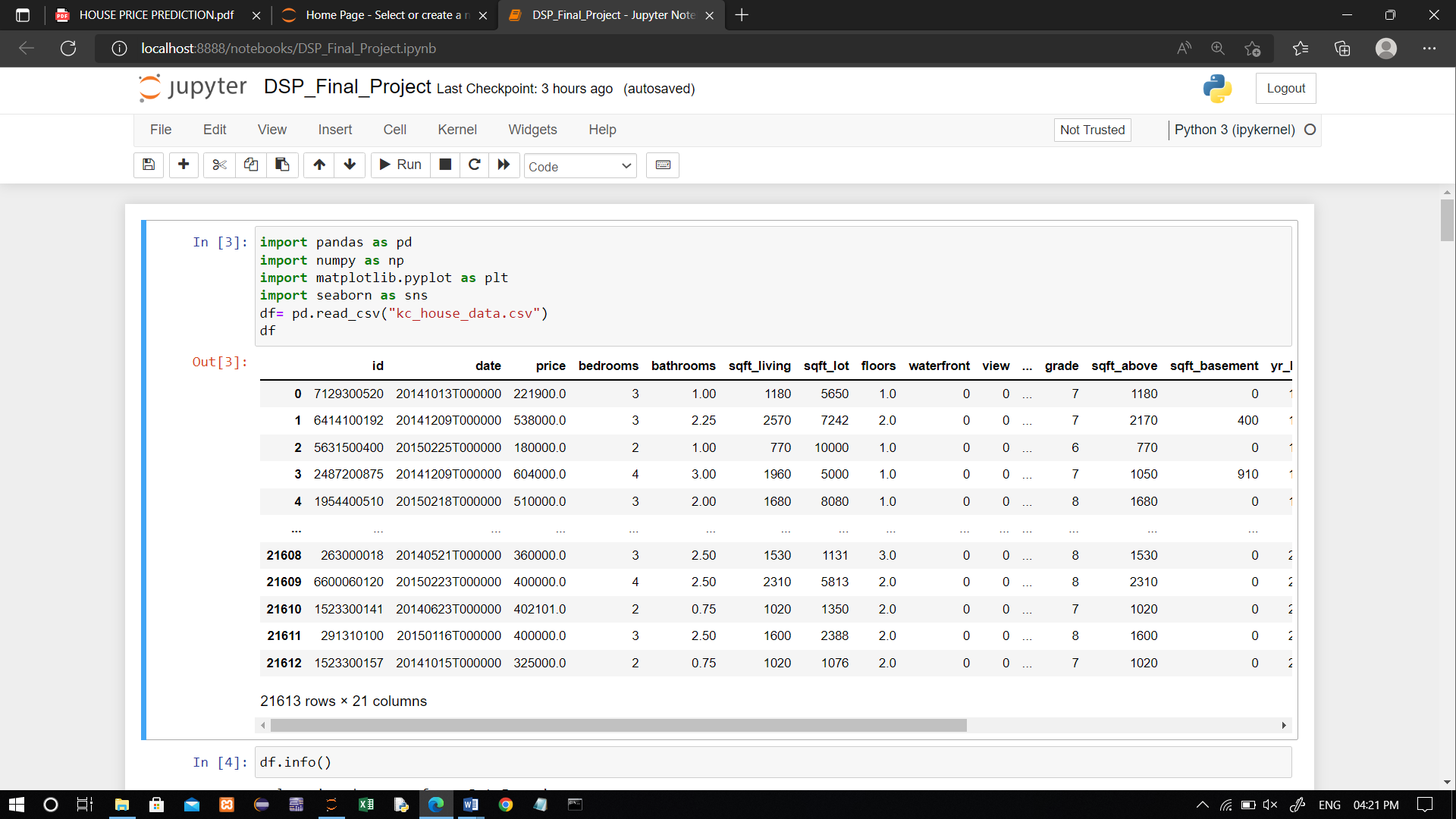
**Seaborn:**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**SKlearn:**

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

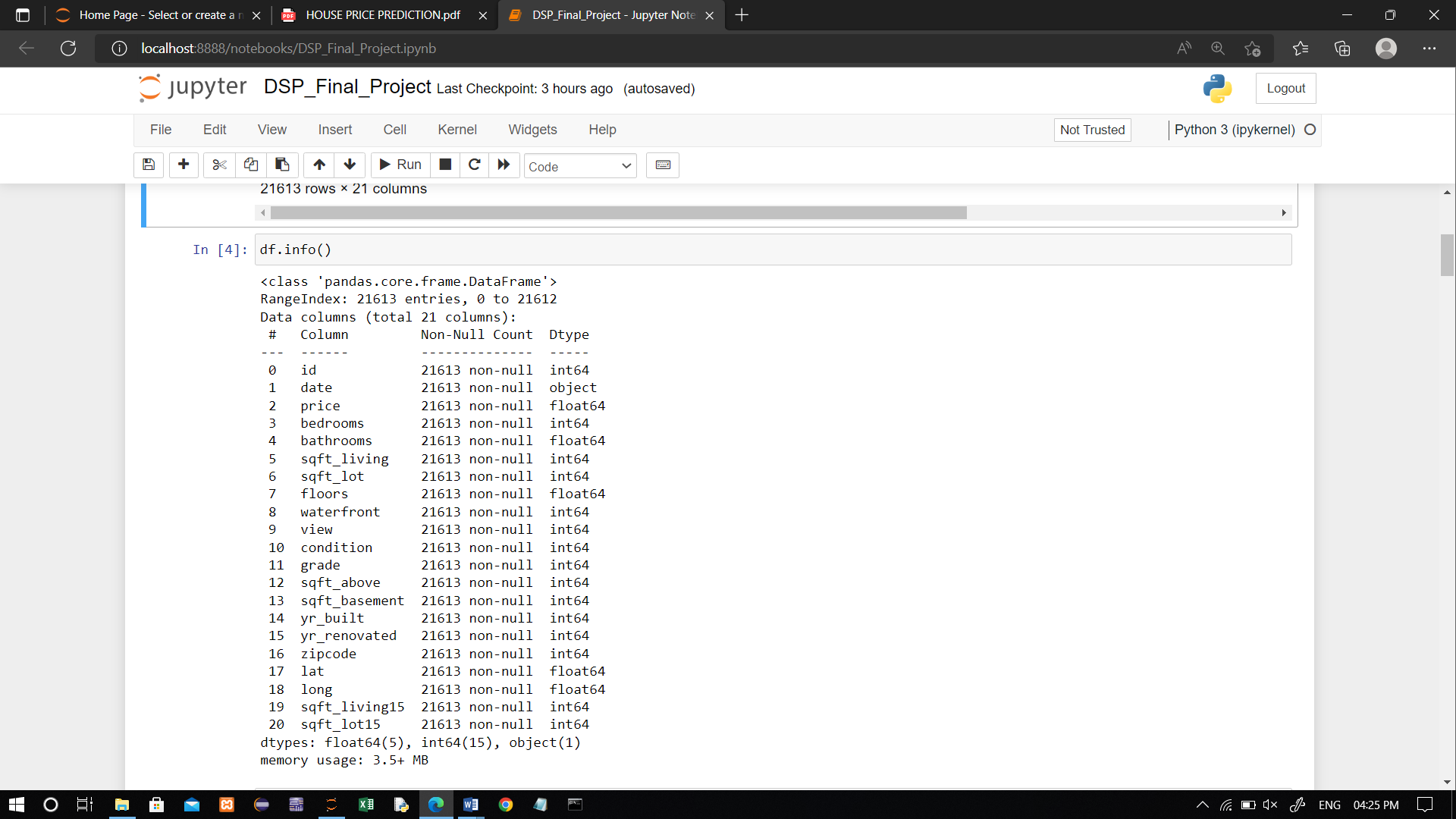
**Importing Dataset:**



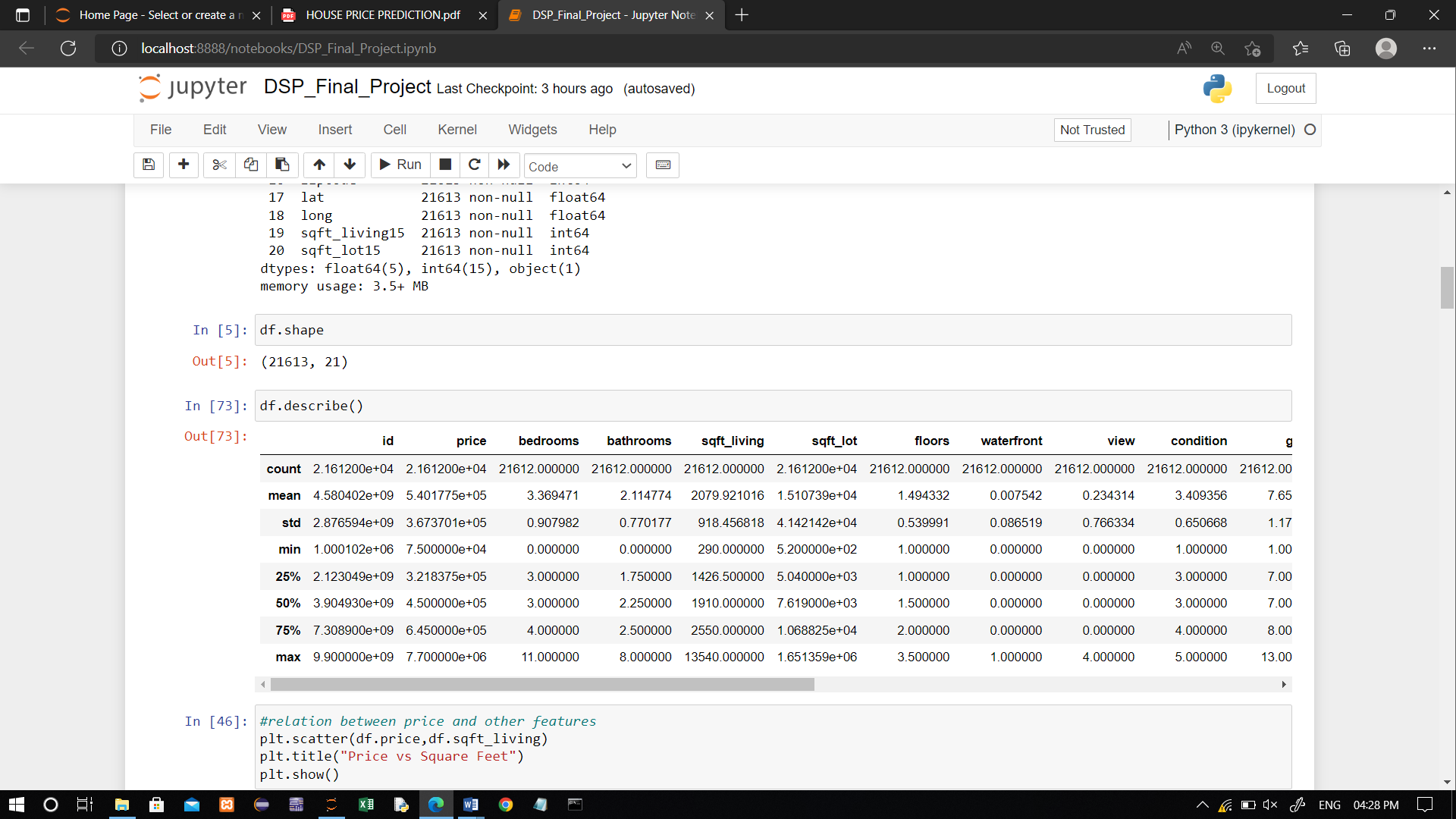
**Data Analysis:**

Data Analysis. Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

.info()

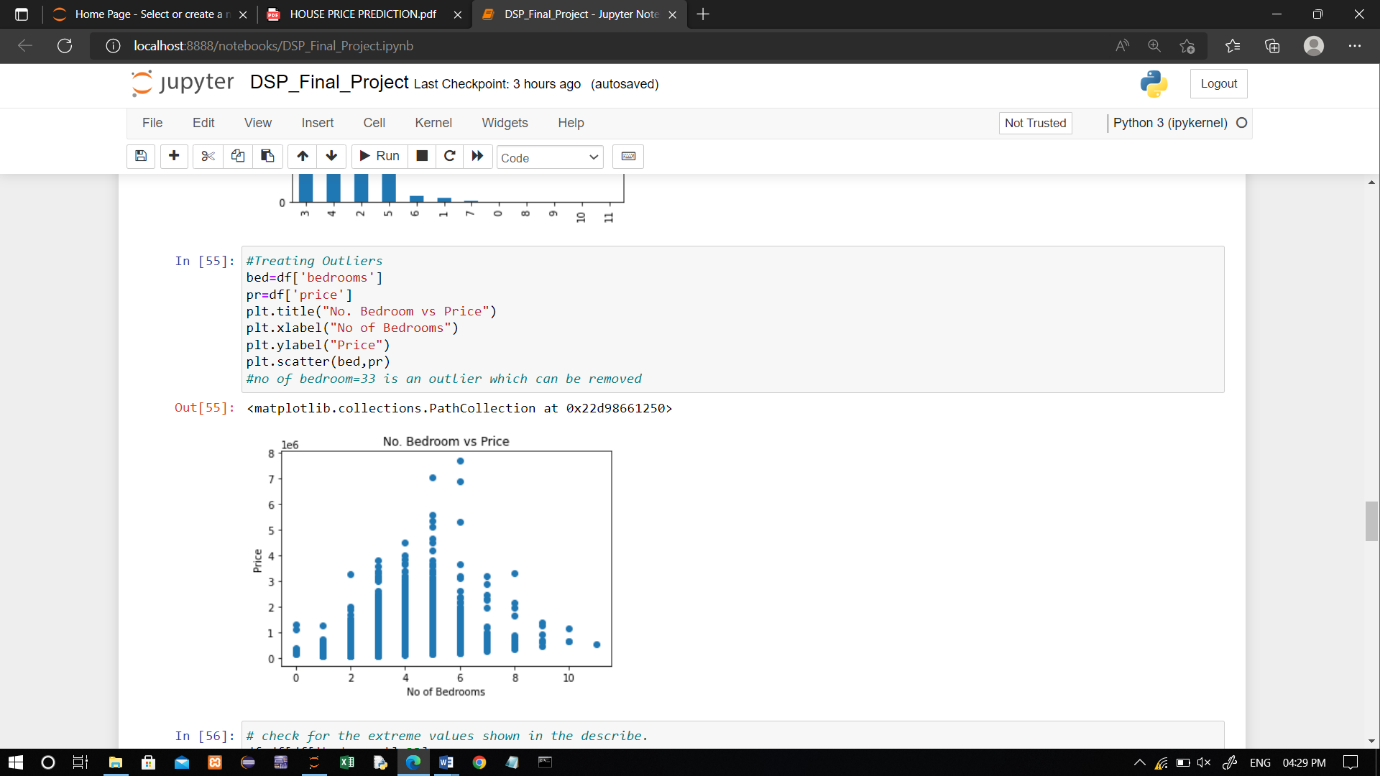


**Check shape of the dataset and use describe function:**



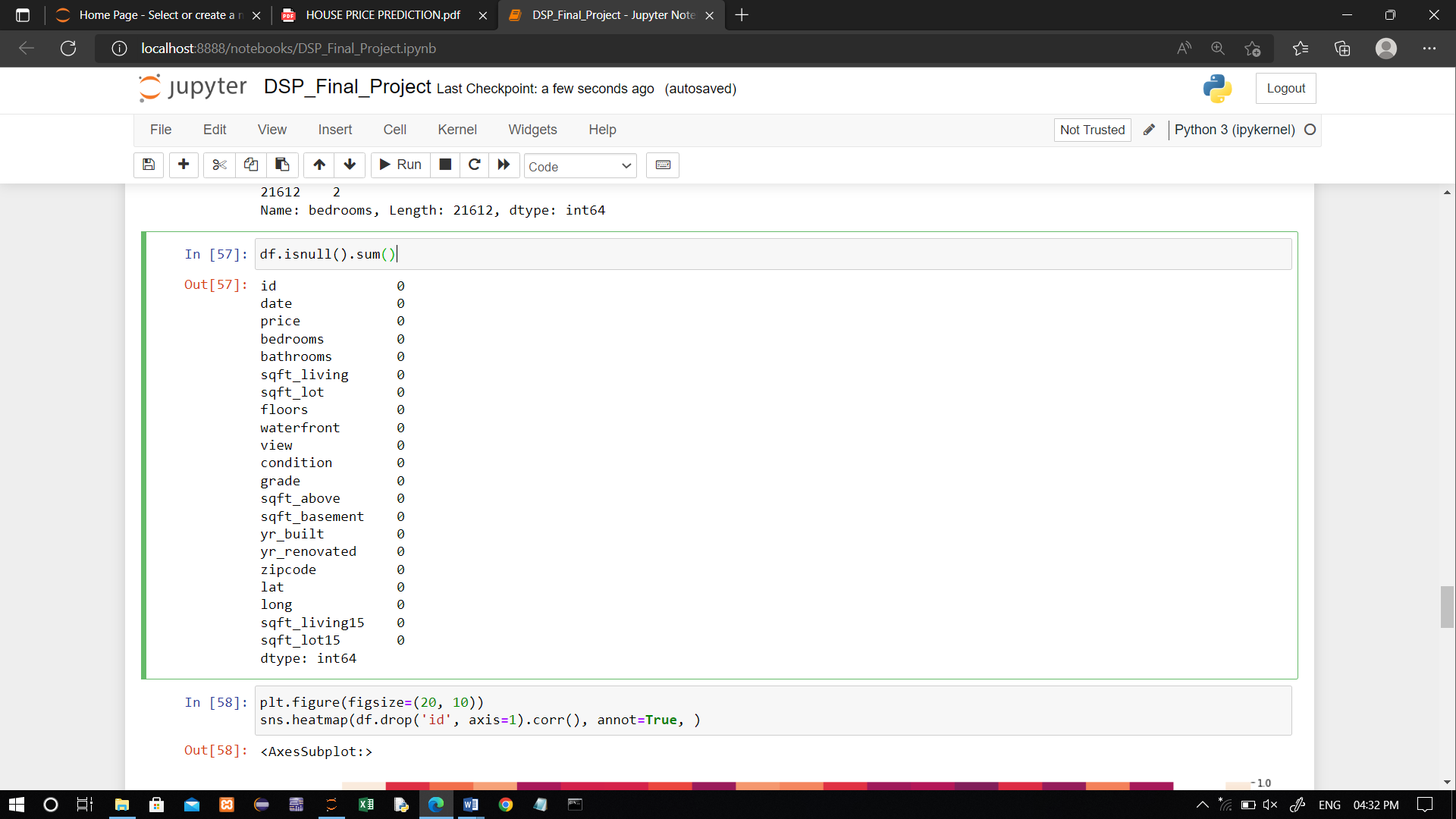
**Outliers:**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population



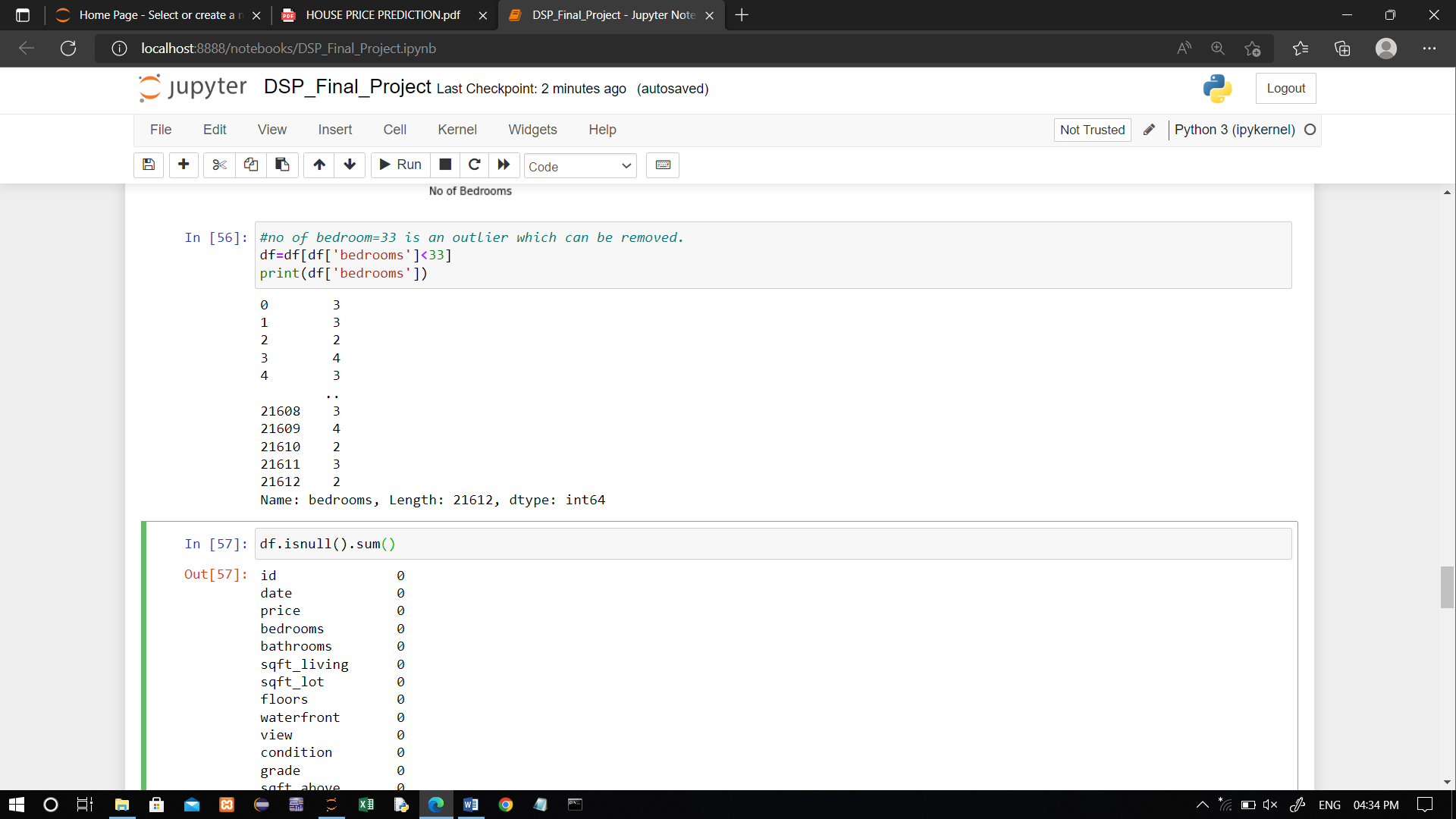
**Check null values:**

A null value in a relational database is used when the value in a column is unknown or missing. A null is neither an empty string (for character or datetime data types) nor a zero value (for numeric data types).

.

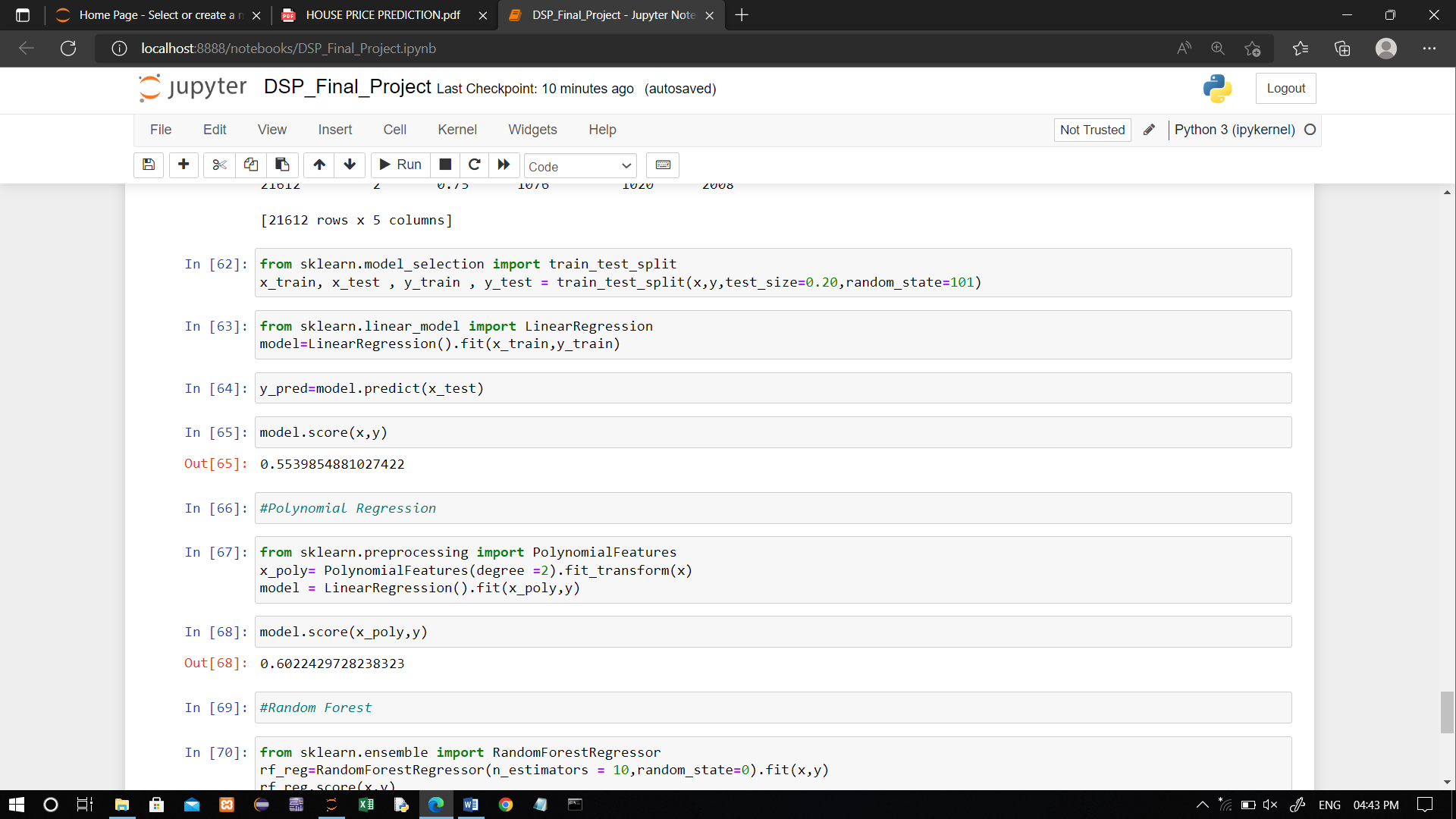
**Data preparation:**

**Removing outliers:**



**Spliting data into training and testing:**

Here we will discuss how to split a dataset into Train and Test sets in python.The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based algorithms/Applications.This method is a fast and easy procedure to preform such that we can compare our own machine learning model results to machine results.by default,the test set is split into 20% of actual data and the training set is split into 80% of the actual data.



**Model Training:**

Here I used 3 types of regression.

Those are

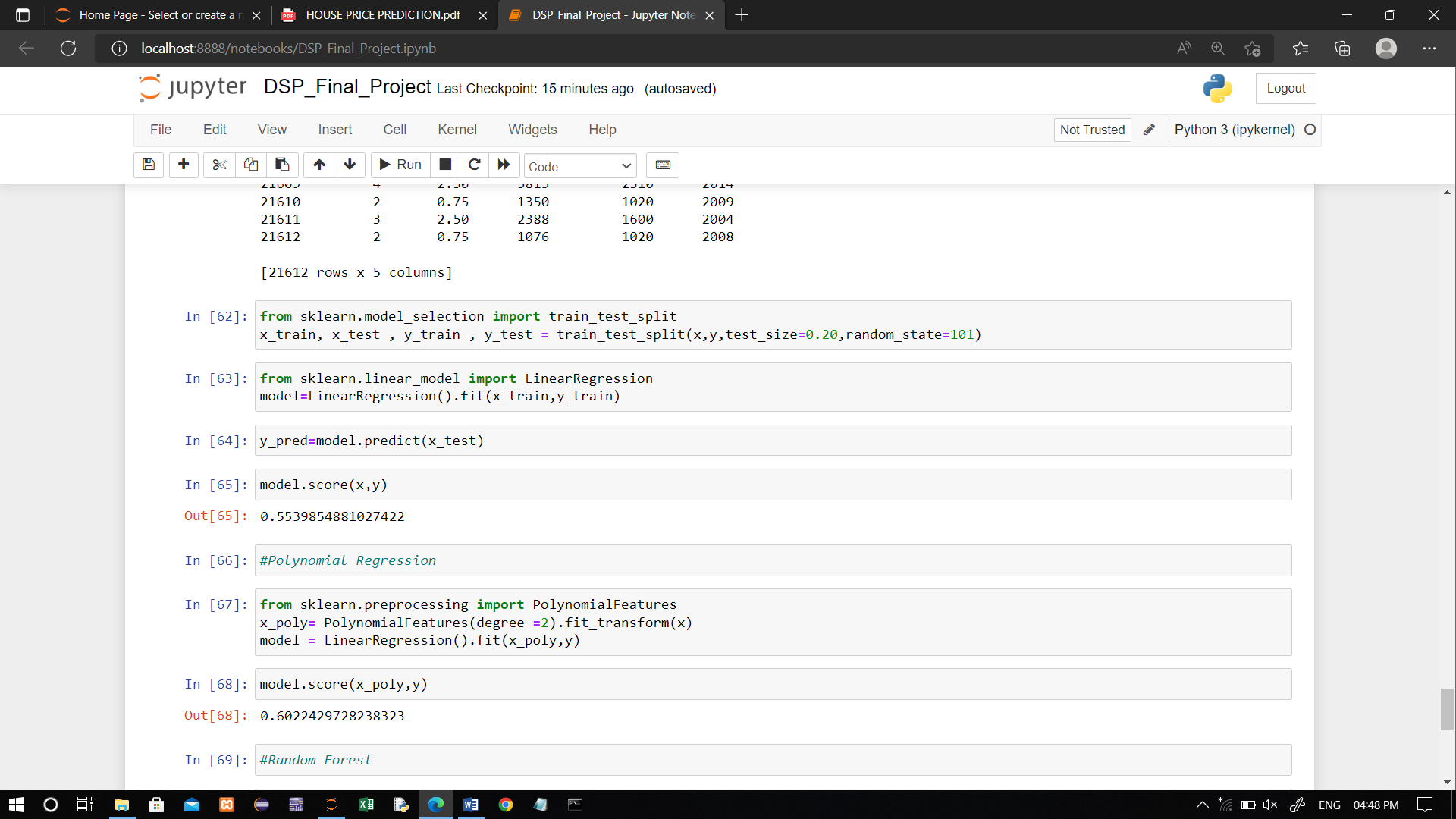
* Multiple
* Polynomial
* Random Forest

**Multiple Linear Regression:**

In Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be various cases in which the response variable is affected by more than one predictor variable; for such cases, the Multiple Linear Regression algorithm is used.

Moreover, Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. We can define it as:

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.



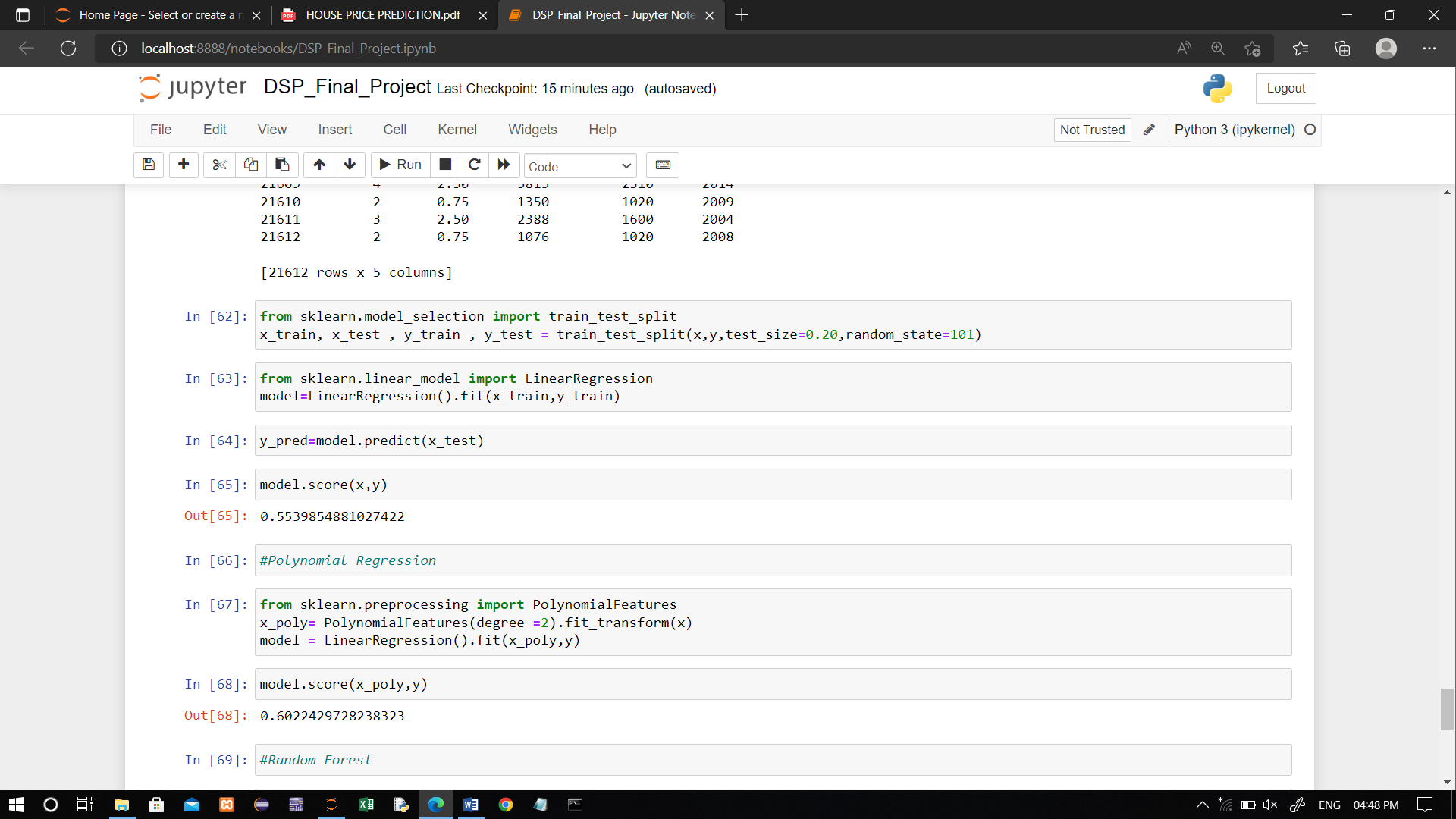
**Polynomial Regression:**

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

y= b0+b1x1+ b2x1^2+ b2x1^3+...... bnx1^n

It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

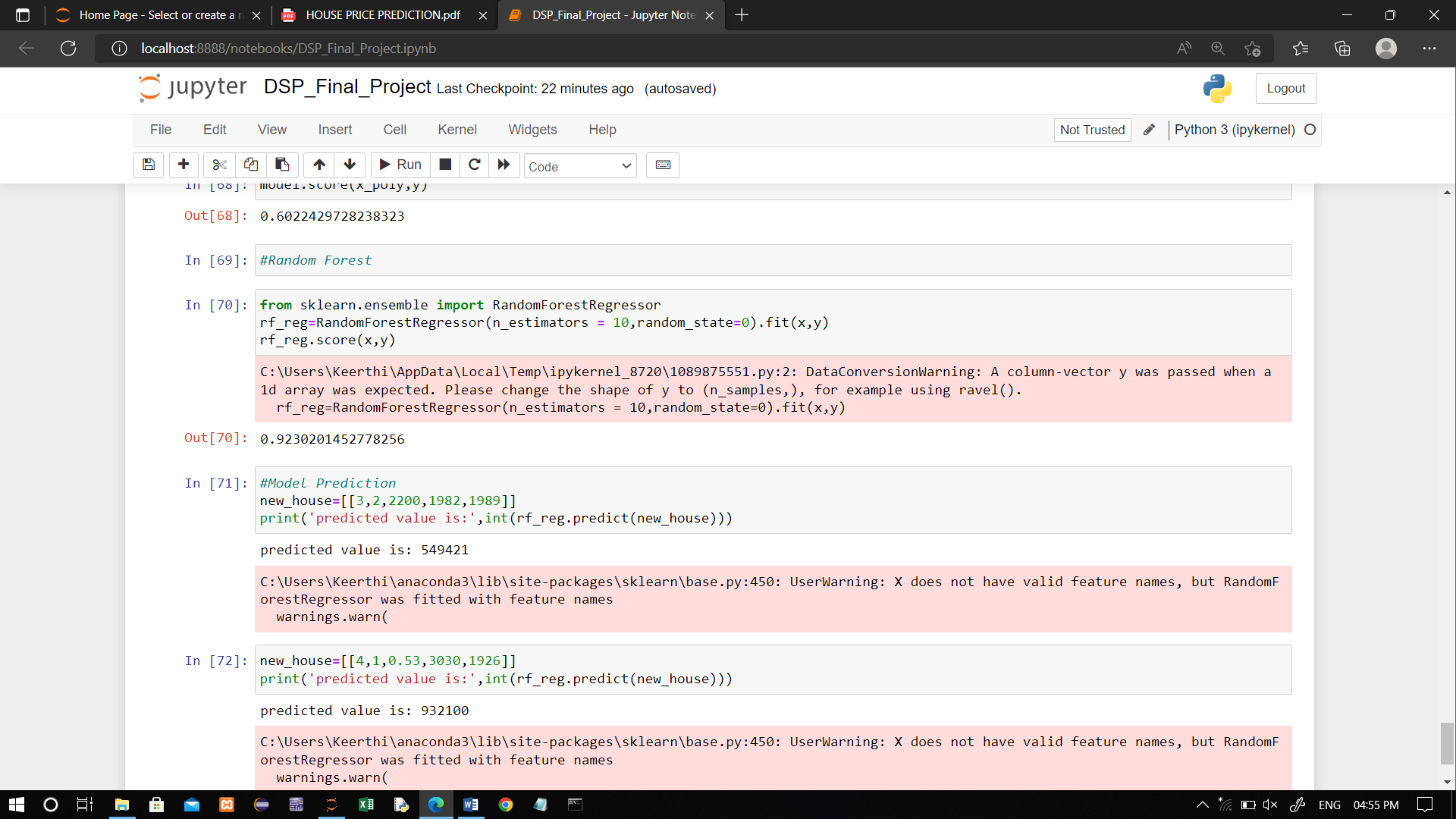
* It is a linear model with some modification in order to increase the accuracy.
* The dataset used in Polynomial regression for training is of non-linear nature.
* It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.



**Random forest regression:**

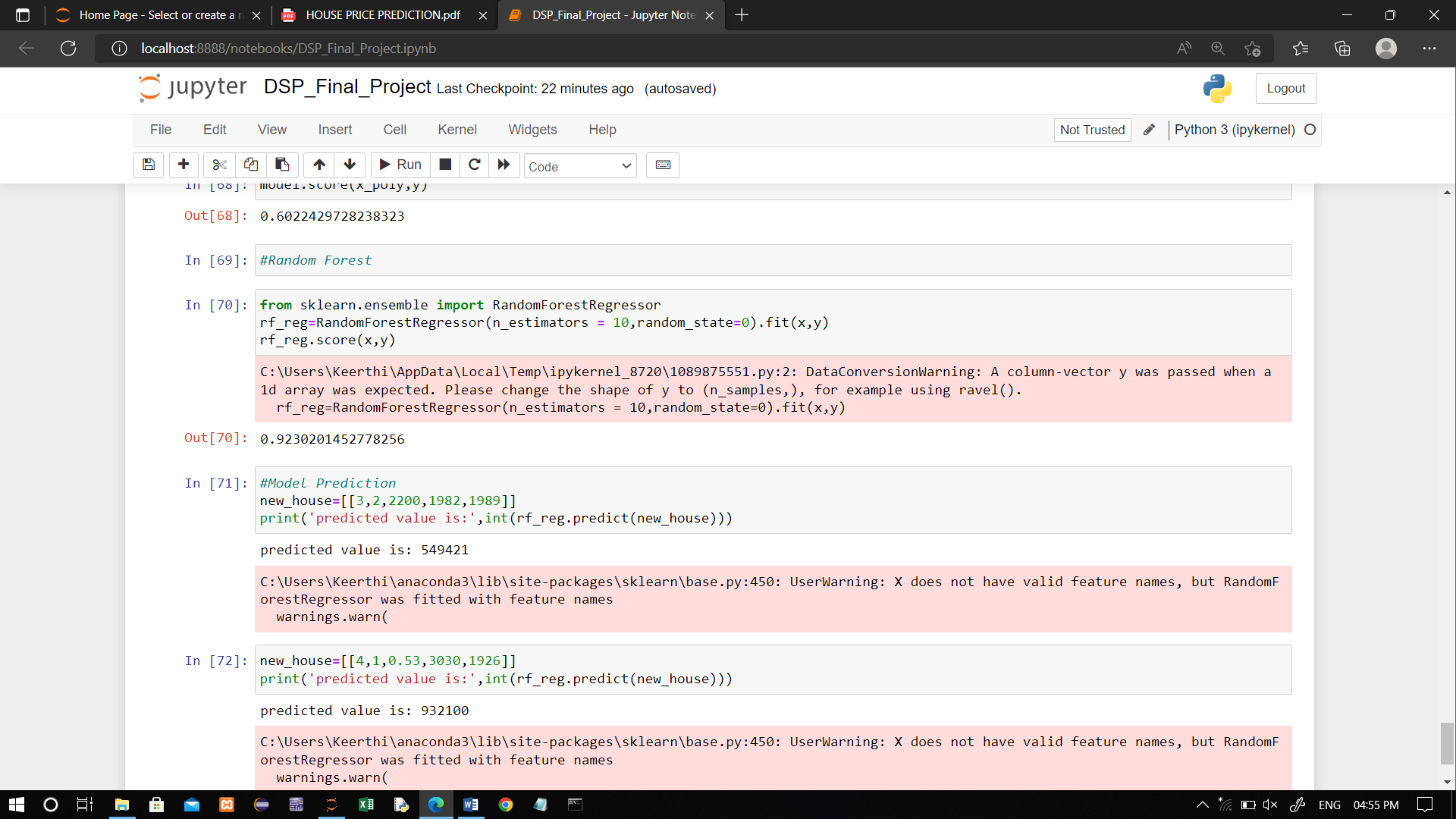
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single modelThe greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest Code🡪



**Model prediction:**

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.



The actual house rate is 6,99,000 but the model predicts the rate of house is 6,69,175.

**Resources and References:**

* Kaggle website for collecting the dataset.
* Any standard textbook or some blogs for getting a concrete idea on regressions.
* Python documentations on visualization tools and libraries for working with datasets.

**Gratitude:**

It feels so fruitful at the semester for learning a subjects at its high peak possible.Moreover,having a project done on the basis of all the theoretical and practical knowledge that we gained throughout the semester from you is a thing that we should be greatful for.Hereby,submitting the project we,from the bottom of our heart,want to convey our immense gratitude for guiding us all through Data Science and even other aspects of life. Thank you sir;

**Subject Faculty:**

Asst.Prof.N.Sesha kumar,Department of Computer Science and Engineering,RGUKT IIIT Srikakulam,Ap

**Done by**

D.Vijayabhanu\_S180308\_CSE-3D

K.Pravalika\_S180514\_CSE-3D