



**HOUSING: PRICE PREDICTION**

**Submitted by:**

**Mrs. CHAITALI M. NAKADE**

**ACKNOWLEDGMENT**

* Towards data science – A medium publication sharing concepts, ideas and codes
* James Le (May 14, 2018) - FIFA World Cup 2018: A Data-Driven Approach to deal Team Line-Ups
* Laree Pertold -Tree Hozz (18th February, 2020)
* Niklas Donges ([May 14, 2018) - Predicting the Survival of Titanic Passengers](https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8?source=post_page-----30870ccc7e8--------------------------------)

### **Article Resources**

* **Notebook and Data:** [GitHub](https://github.com/LearnDataSci/article-resources/tree/master/Housing%20Price%20Index%20Regression), Jupyter Notebook
* **Libraries:** numpy, pandas, sklearn, scipy, joblib, matplotlib, seaborn, statsmodels

**INTRODUCTION** 

In this report we will discuss the complete information about model building of “Housing: price prediction” through linear regression/Random Forest regressor/Decision tree regressor and find the best model from that.

* Business Problem:

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy.

Data science plays very important role to solve problems in this field to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

For the same purpose, the company has collected a data set from the sale of houses in Australia. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company we wants to know:

1. Which variables are important to predict the price of variable?
2. How do these variables describe the price of the house?

* Conceptual Background of the Domain Problem

Matching to your opinion in below image which house is most expensive?

After analysing the below image, we find that left orange one is the most expensive because we can see yellow house is hugest



and latest but there are so many factors which affects the prices of houses like location, locality, zone, building type and many more. Still, you might have a unlike answer after reading this report. Regarding that factors we will discuss in detail during model building.

* Review of Literature

As we all know the population of world is increasing day by day and the demand for residence remains unmet as per the requirement because lack of houses present in the market. So, many purchasers are ready to buy the houses but houses are not ready to possession resulting those buyers are continuously requesting to outside estate agent. On top of that, buyers having to pay more are borrowing more.

It works on simple economics, as demand increases the price is automatically goes high. Some reason why it is happening,

* A lack of accessible land
* Population goes high
* Low government investment
* And may more
* Motivation for the Problem Undertaken

The**connection between house prices and the budget** is an important motivating factor for predicting house prices.

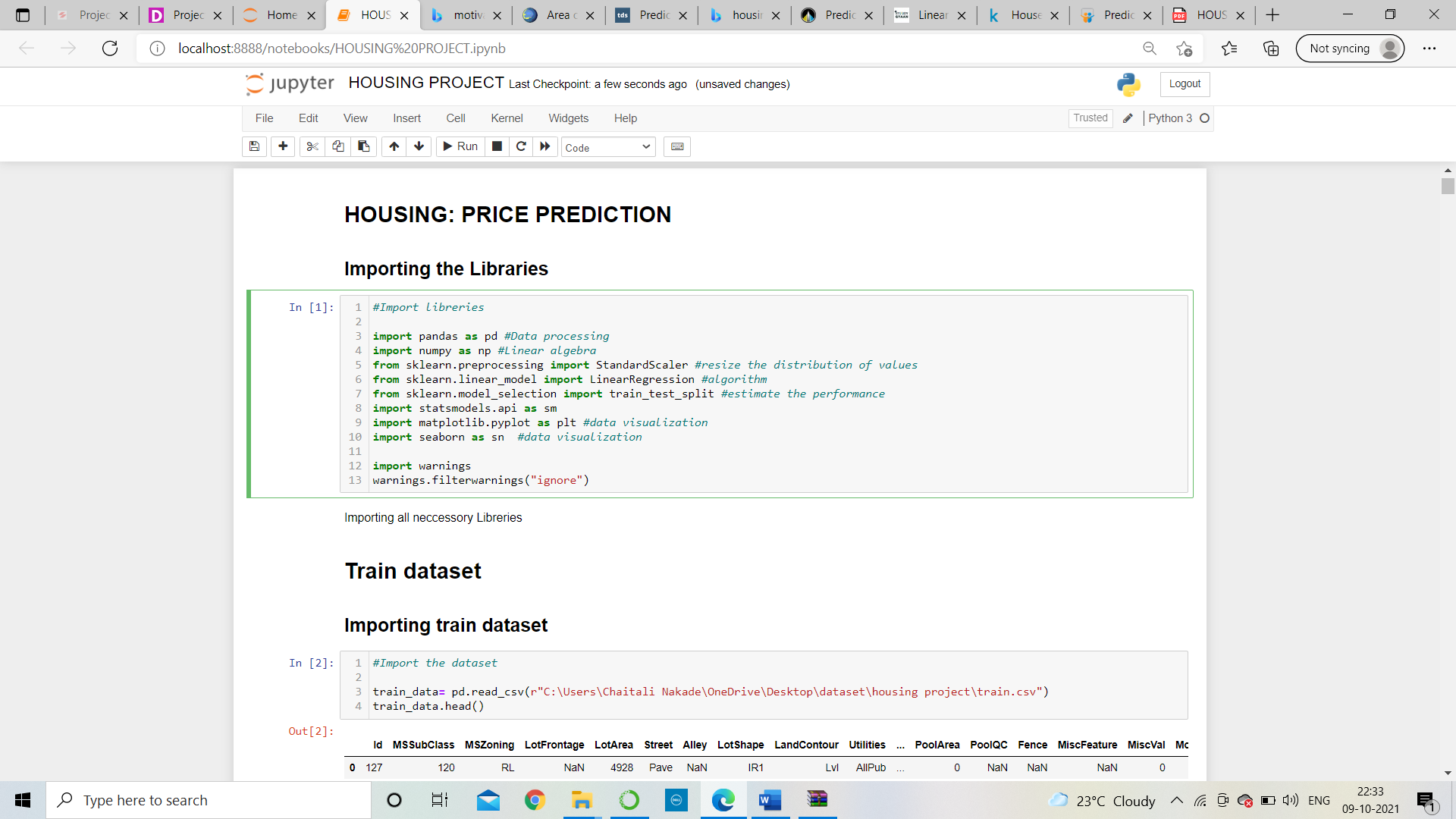
My main moto to make this project is to predict the house prices in and those people who is having less financial status and will get good houses in less price or we can say with their budget.

**Analytical Problem Framing**

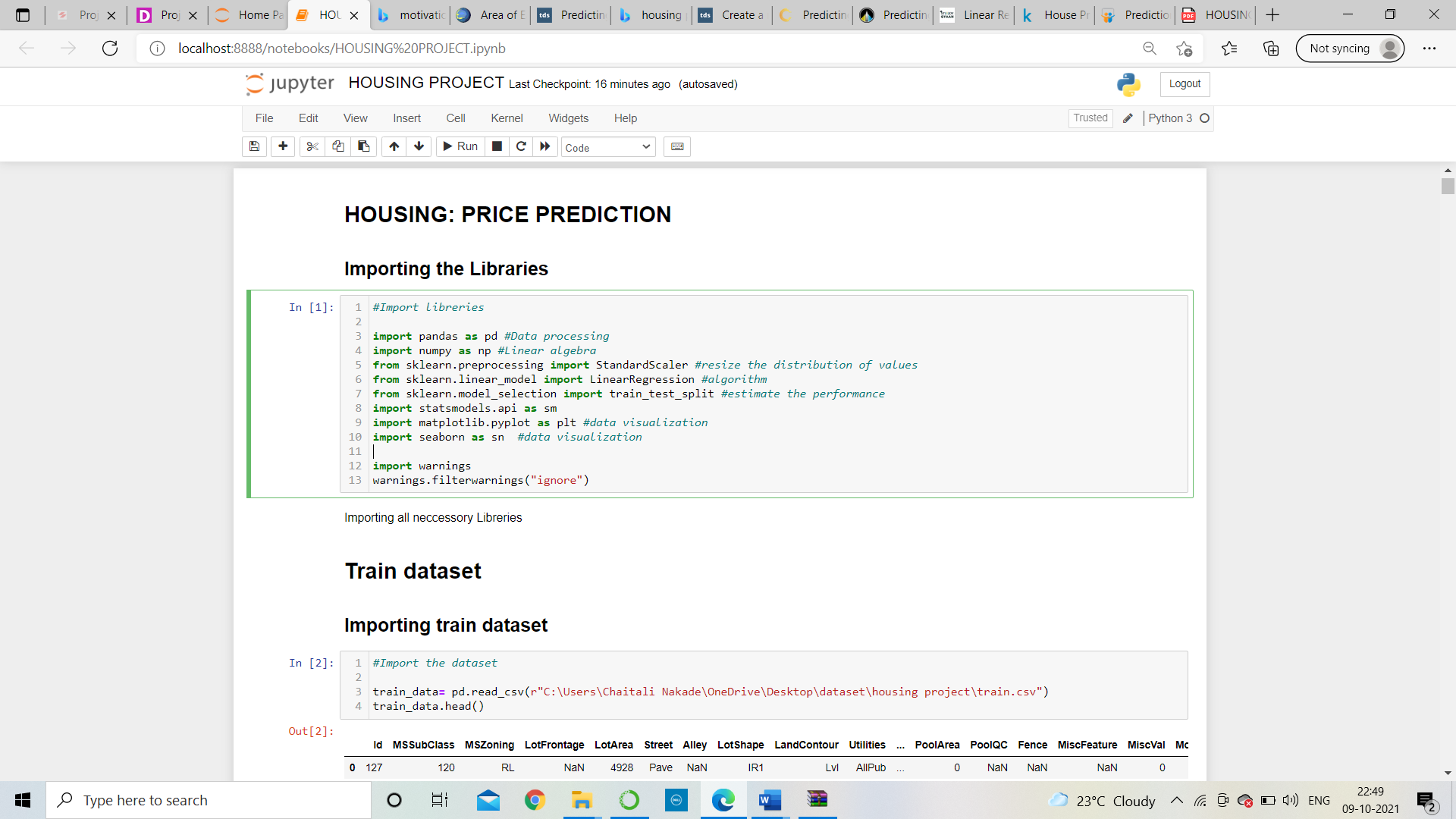
* Mathematical/ Analytical Modelling of the Problem

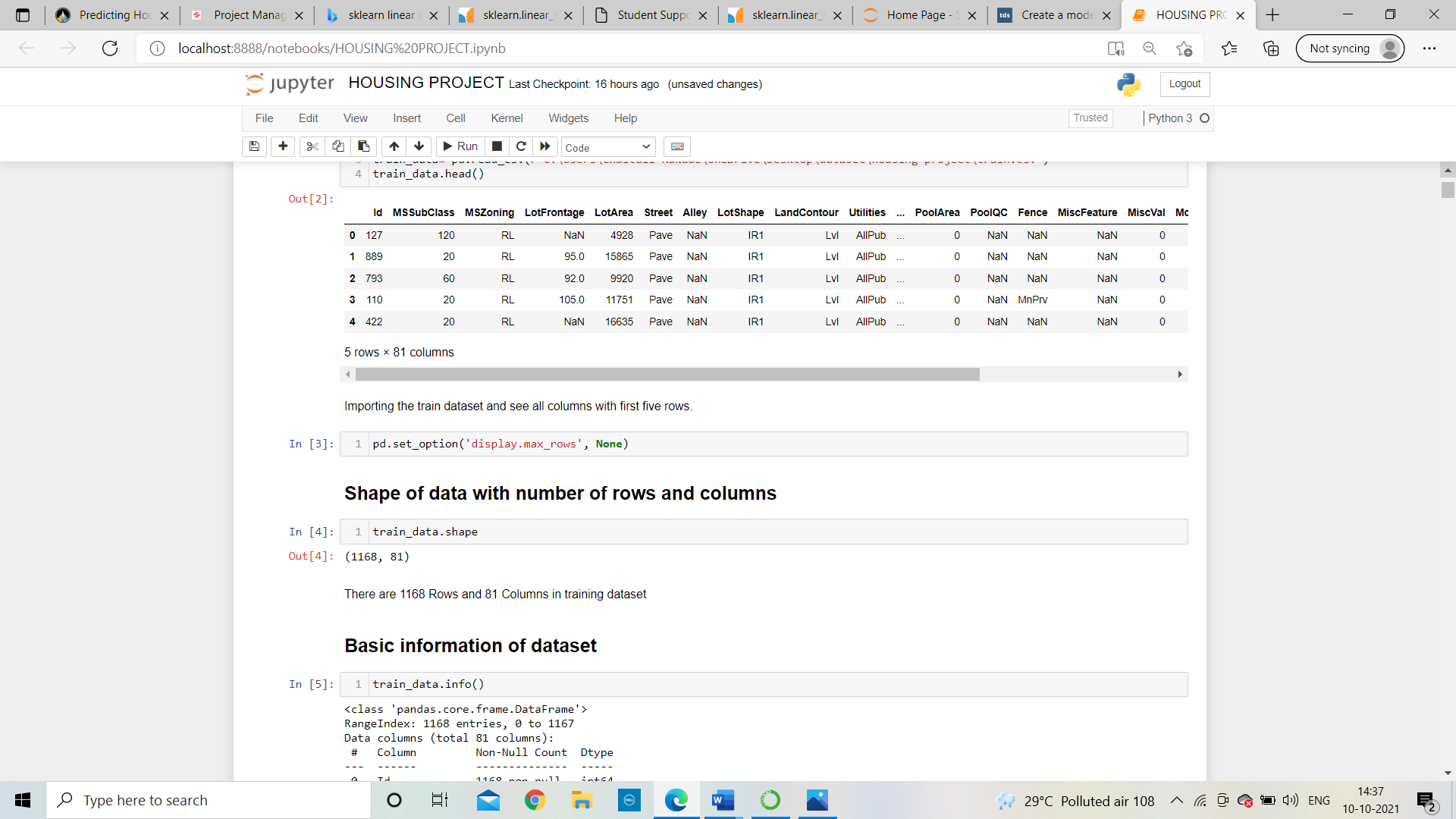
Now we will work on model building, we are using machine learning techniques to predict house prices. Here we are having two dataset and that are train and test dataset. The train dataset contains 1168 rows 81 columns and test dataset contain 292 rows and 80 columns.

Now work with dataset with importing necessary libraries.



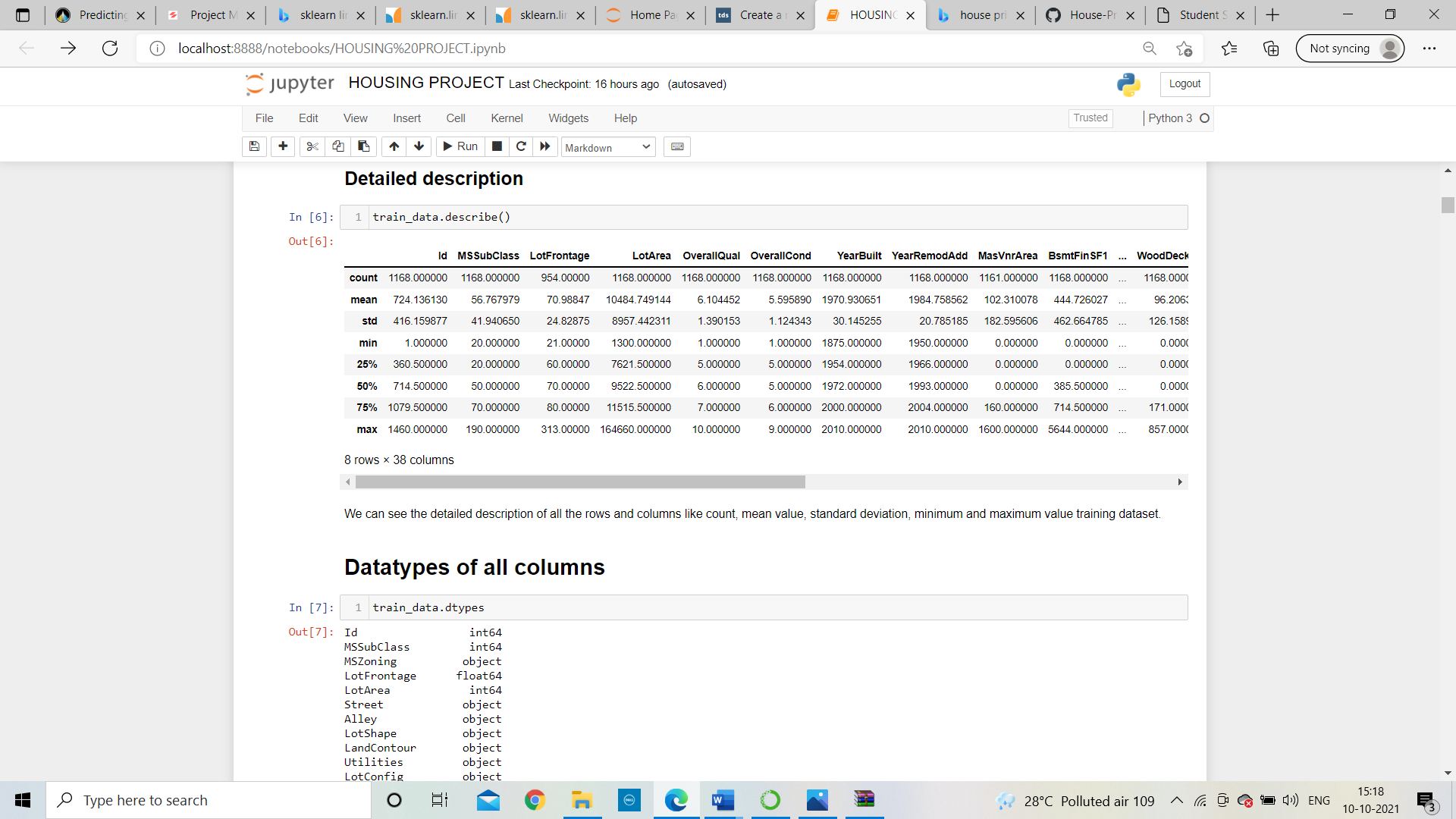
All these libraries are used for model building and data processing. Use of particular libraries are mentioned over there.





The complete training dataset is imported in variable name train\_data and we can see there are some columns are need to alter them into numeric data because machine learning will not work with object data type but before that need to analyse the data.

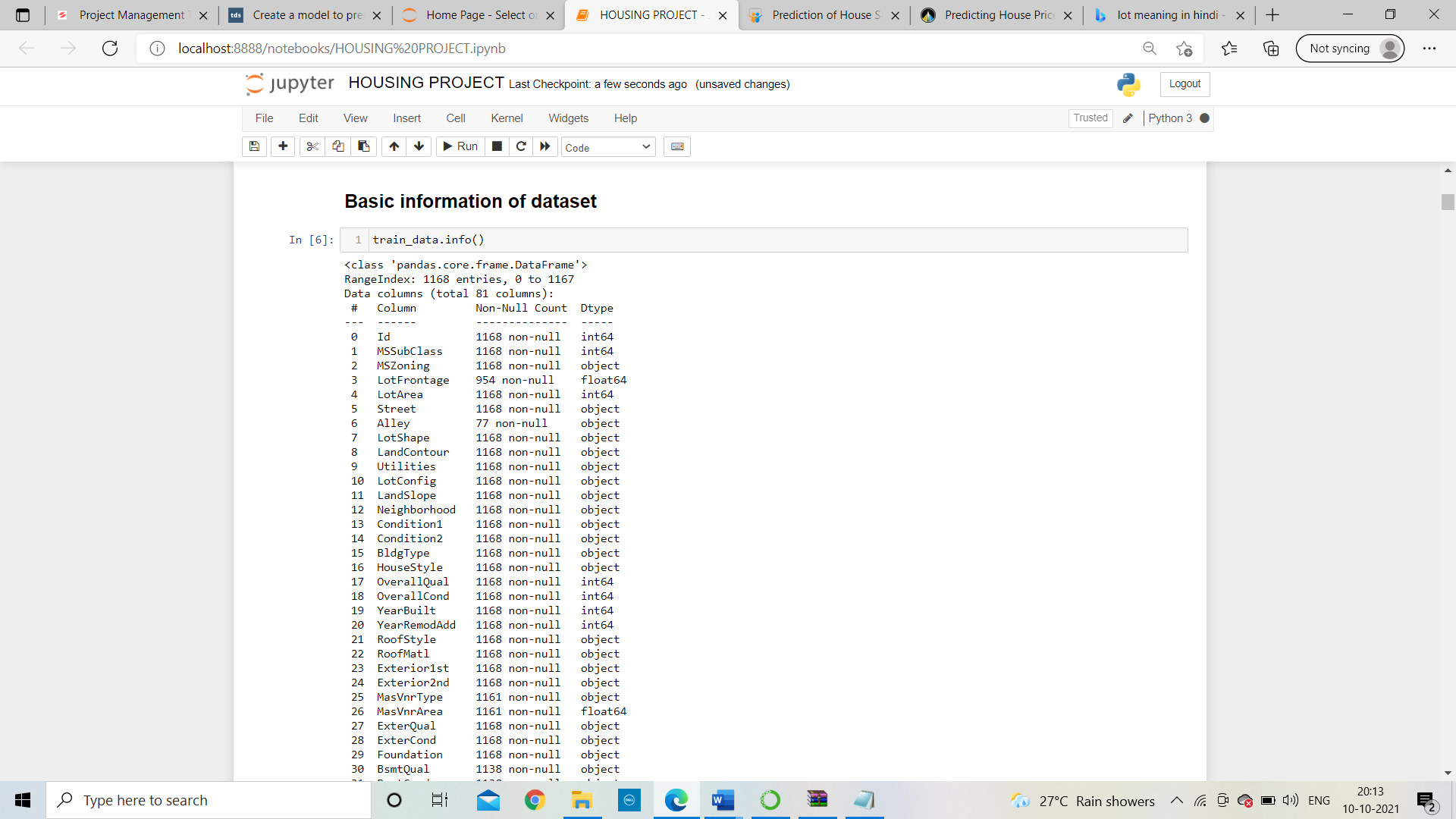
* Data Sources and their formats

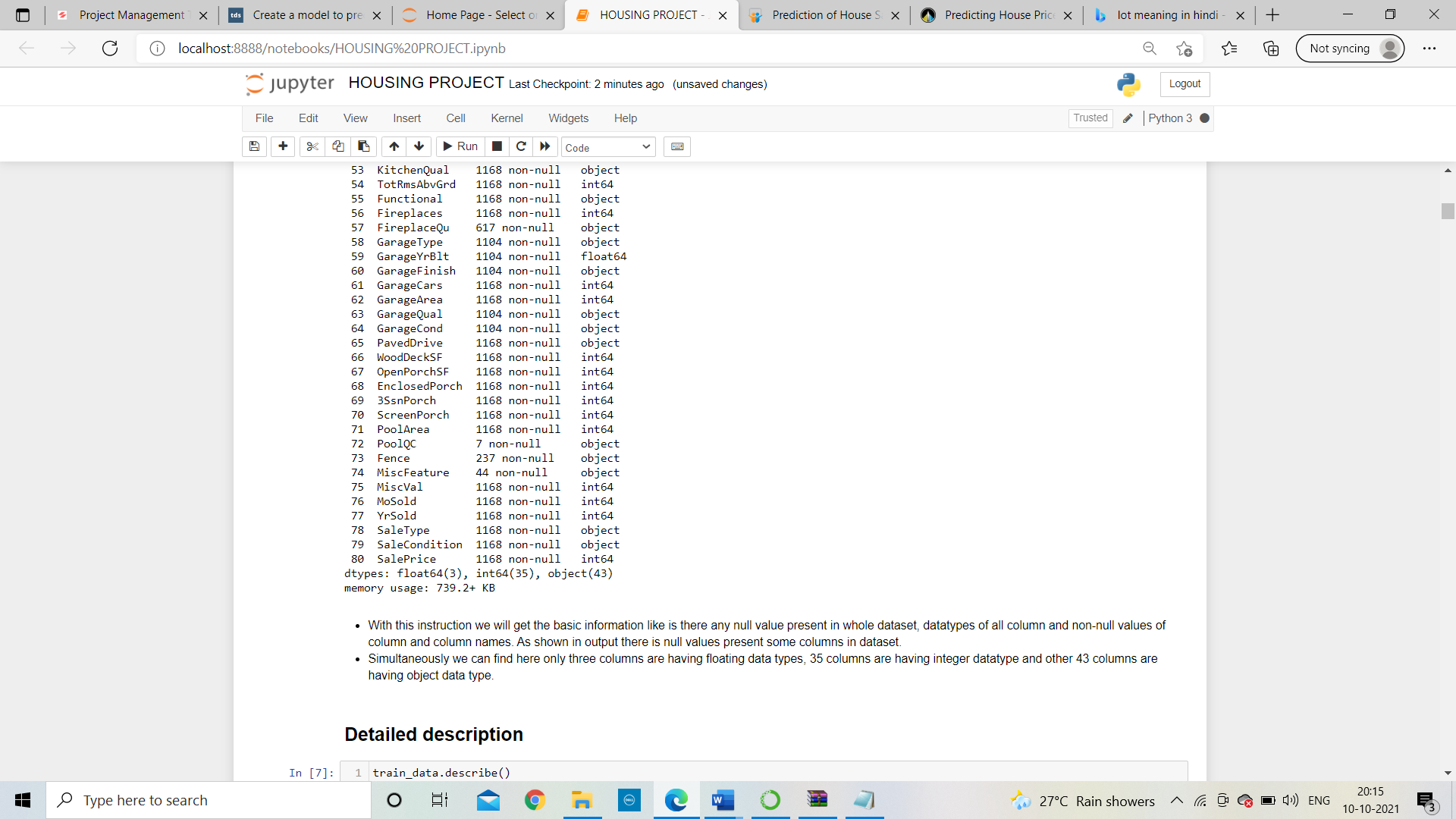


Express more about the dataset

* 1. The dataset has a quality column in which Rates the overall material and finishing of the house where some houses are having superior quality 10 as a maximum value and some in very poor quality with 1.
  2. Year built shows the older house is 1875 years old whereas new one is 2010 years old.

Similarly, we can conclude so many things by just observing at the describe function.

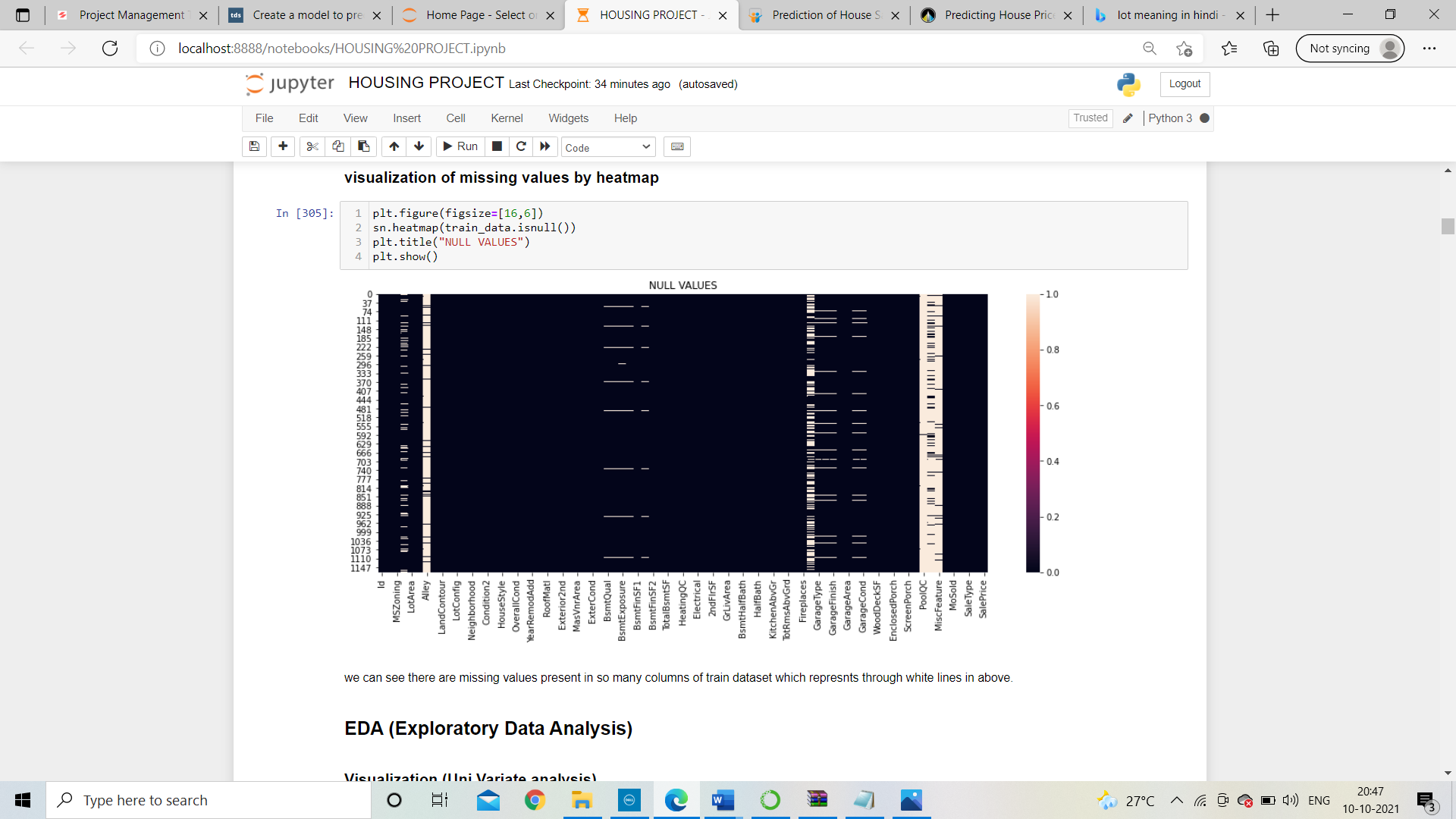




* There are 1168 Rows and 81 Columns in training dataset
* We can see some columns are having null values also.
* We can find here only three columns are having floating data types, 35 columns are having integer datatype and other 43 columns are having object data type.
* Data Inputs- Logic- Output Relationships

Visualization of data:

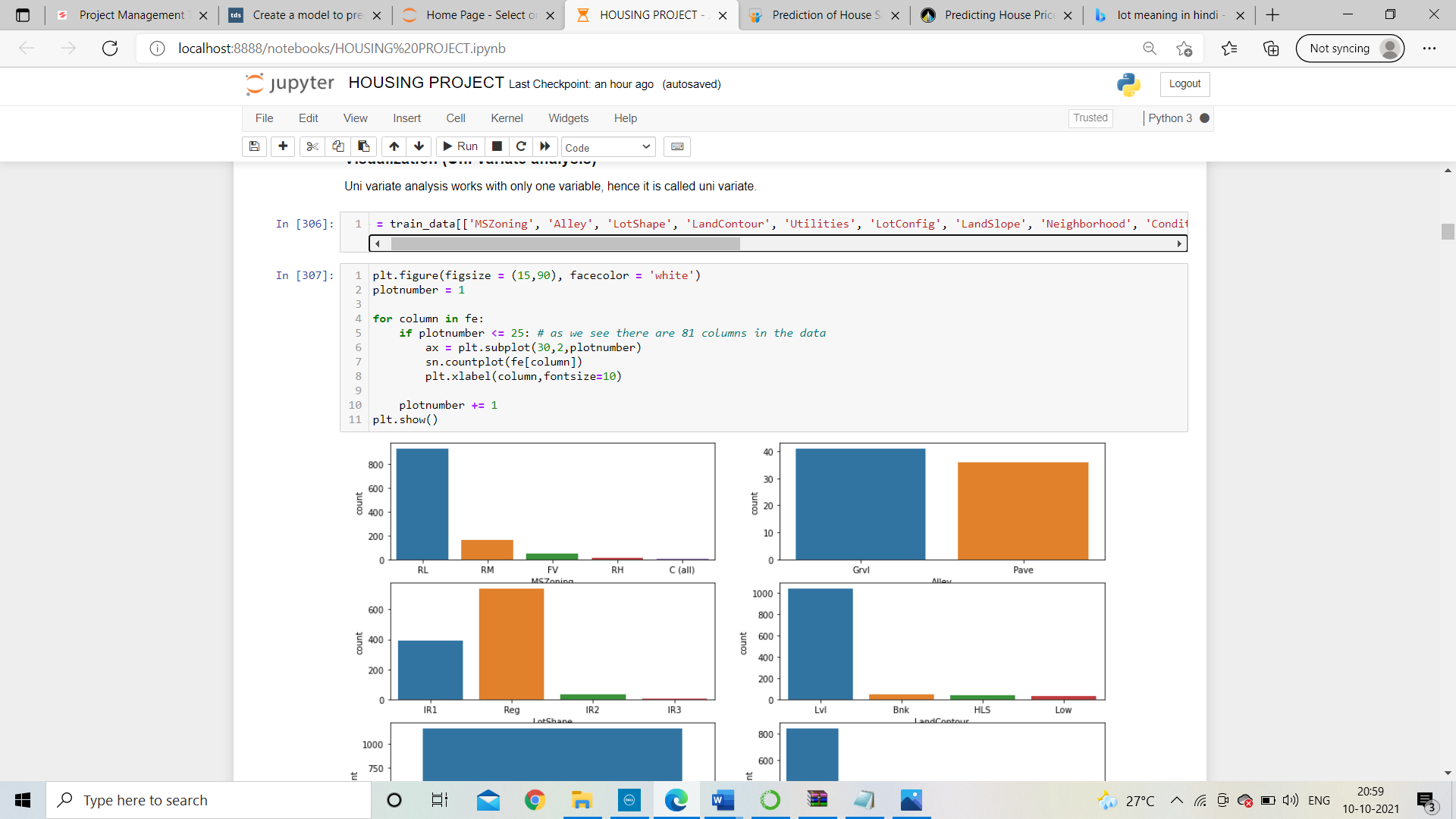
1. Null values



The null value represents with white line shown in above image. poolQC, miscfeature, alley and fireplaces columns are having too much of misssing data.

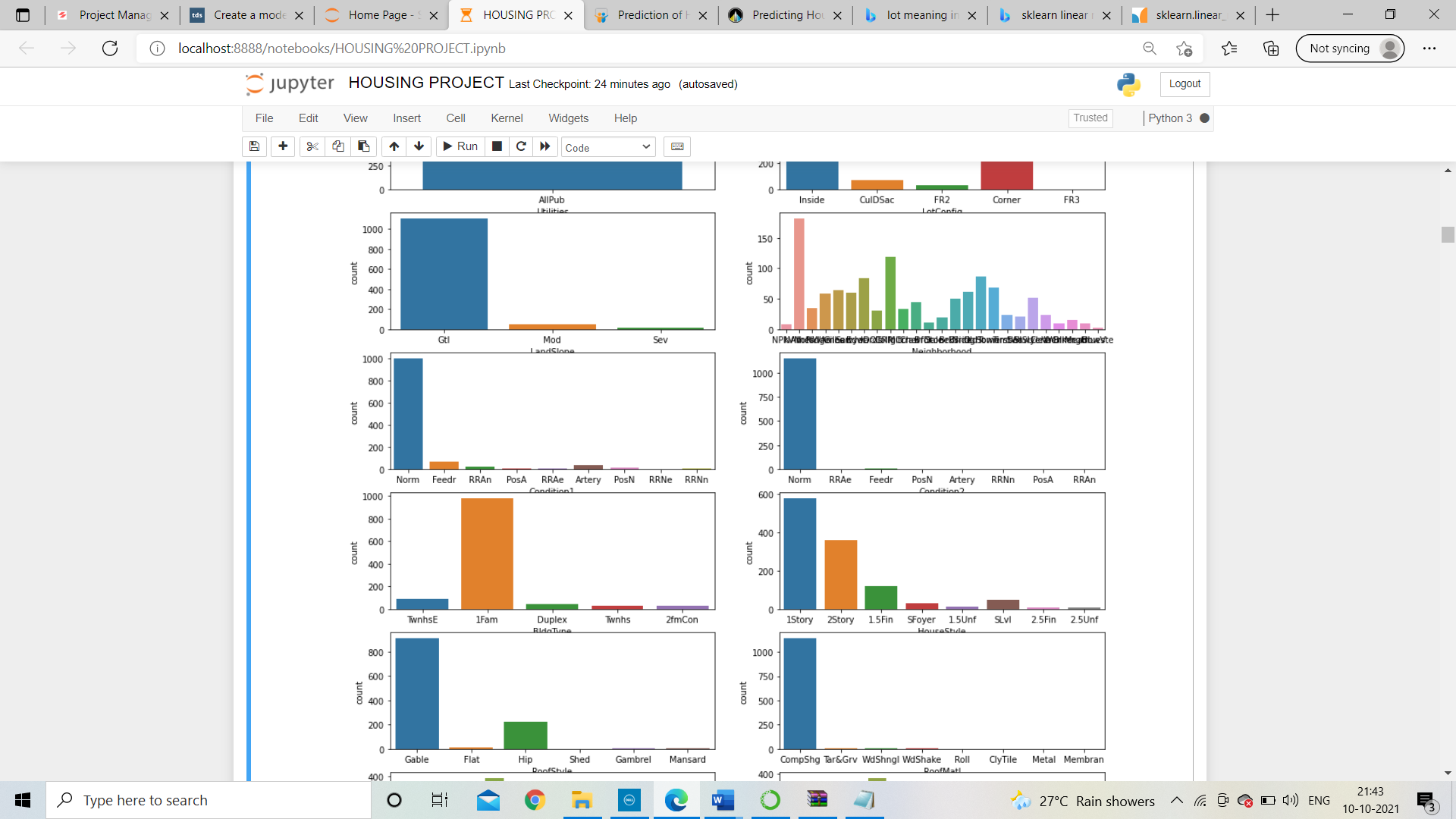
Univariate analysis:

* In which zone most commonly houses are sold?



As we can see from the visualization, Residential in Low-Density houses are most commonly sold. So how is it beneficial? For a builder having this data, He can make a new construction in Residential with Low-Density to fascinate more buyers.

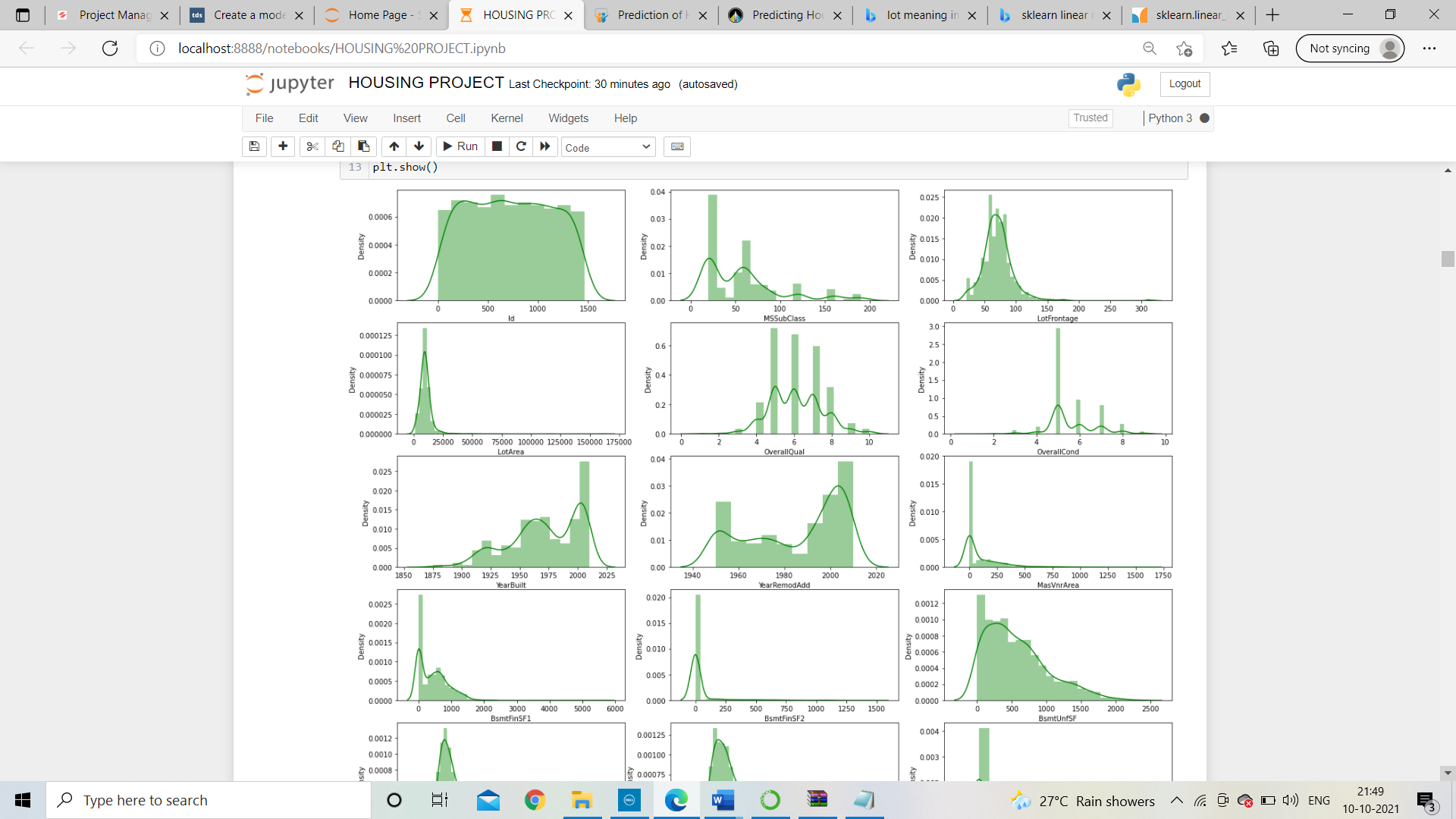
* Most commonly houses are sold as building type?



Single-family Detached houses are most commonly sold. So, as I said single family building should be first preference of builders.

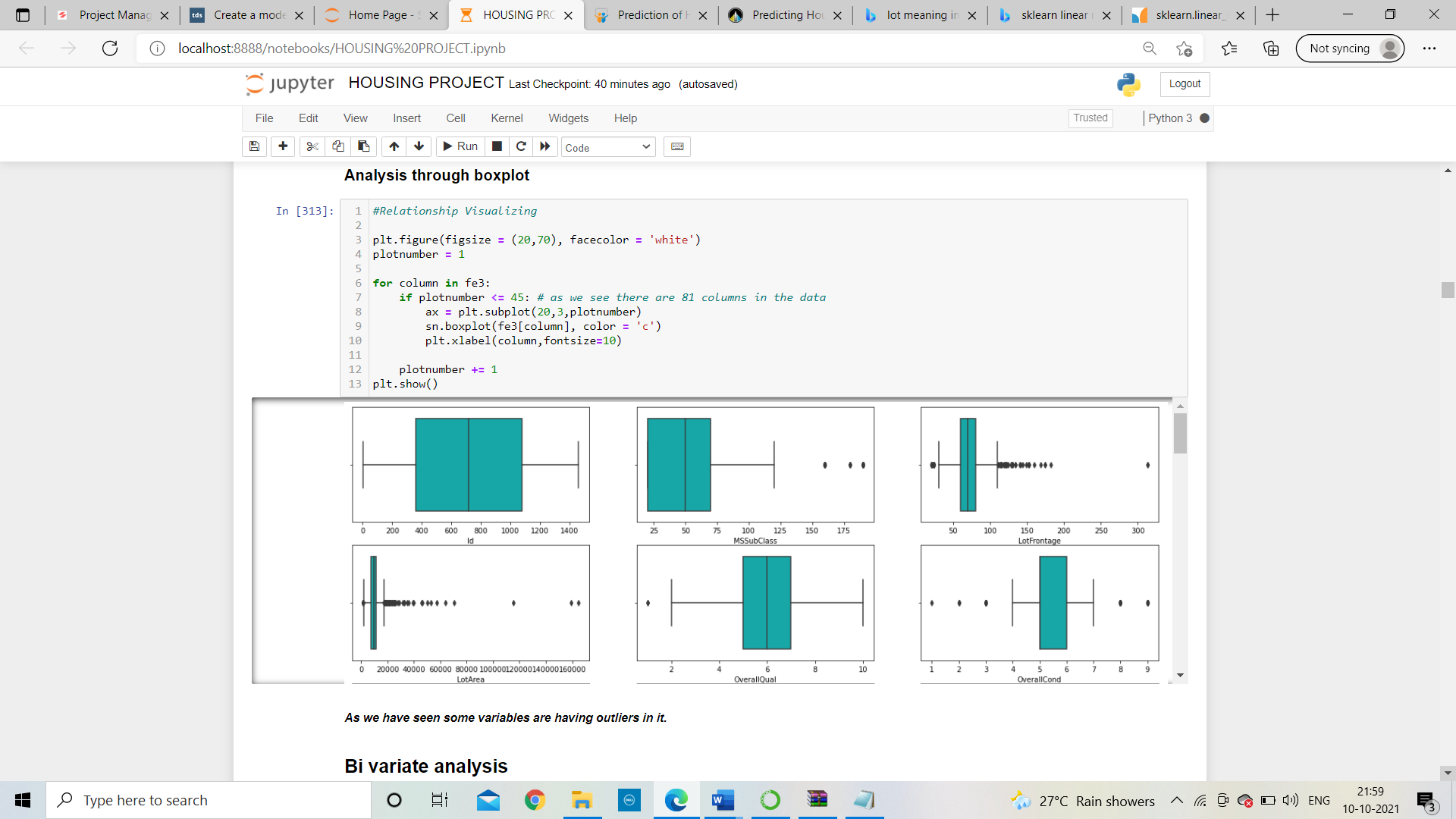
* Most important factor for sale price?

Depends on quality of houses the houses are sold



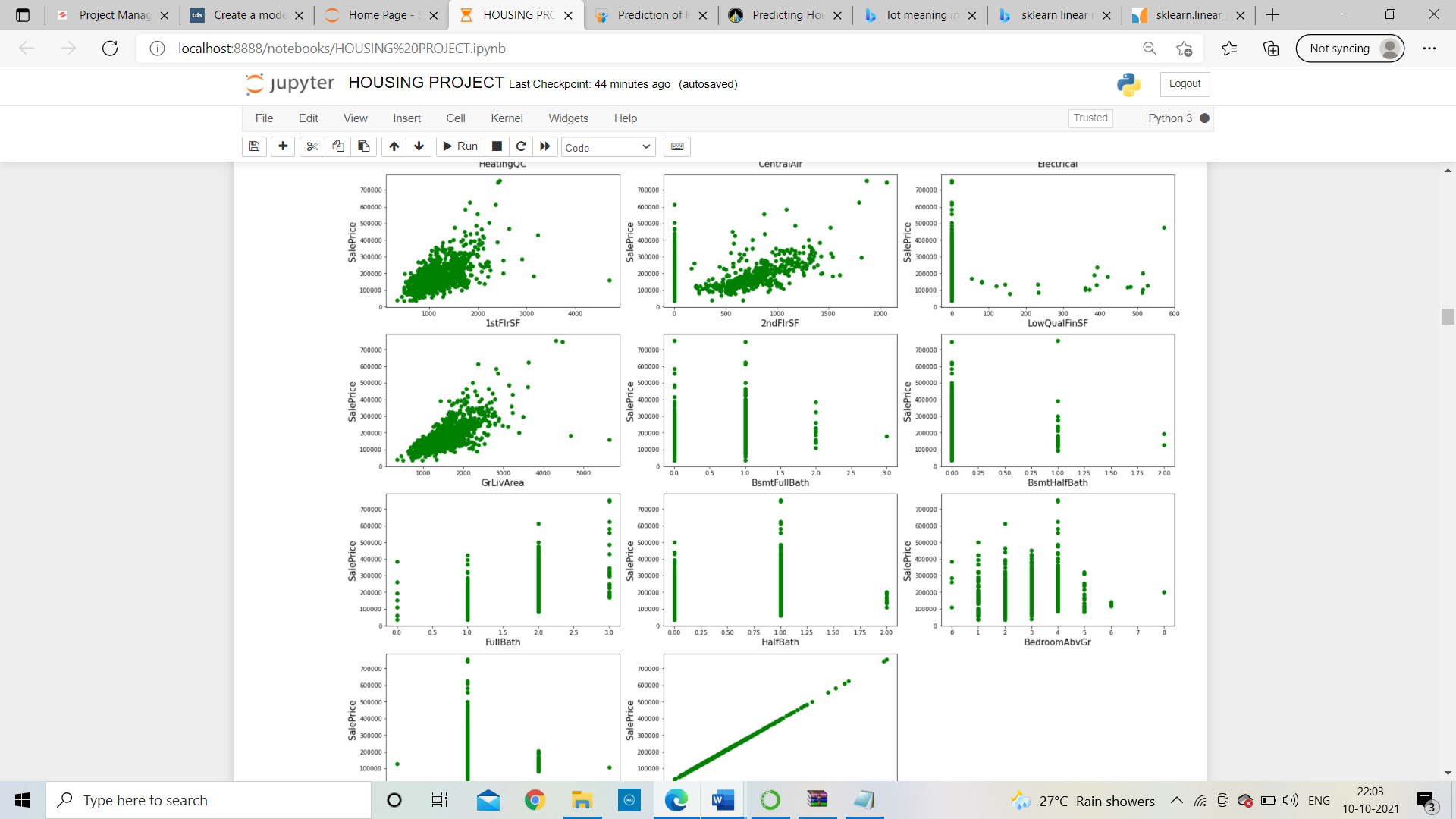
After visualization we can see overall quality of houses is pick with 5 rating followed by 6 and 7 most commonly sold. But rating with 3 and 10 are least soled houses.

* Check is there outliers are present in dataset?



Some columns are having outliers which we need to treat that with z score technique which removes the outliers.

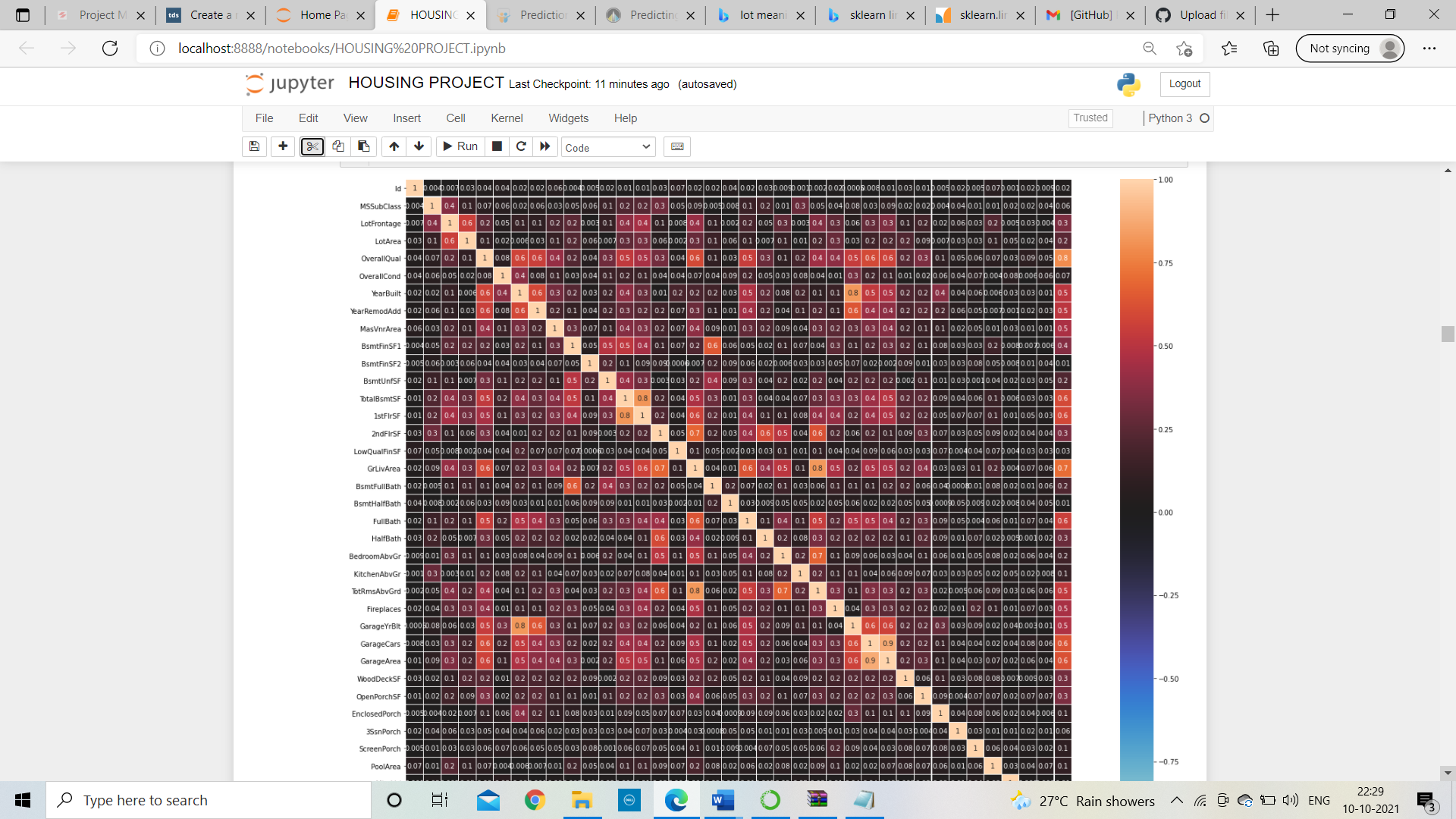
* Scatter plot to check the relationship with sale price?



Shown in above for First Floor square feet is 1000 to 2000 square feet house sale price is near about 10lakh to 40 lakh and some are more expensive also.

Similarly, we can also analyse the data for 2nd floor square feet and Garage living area.

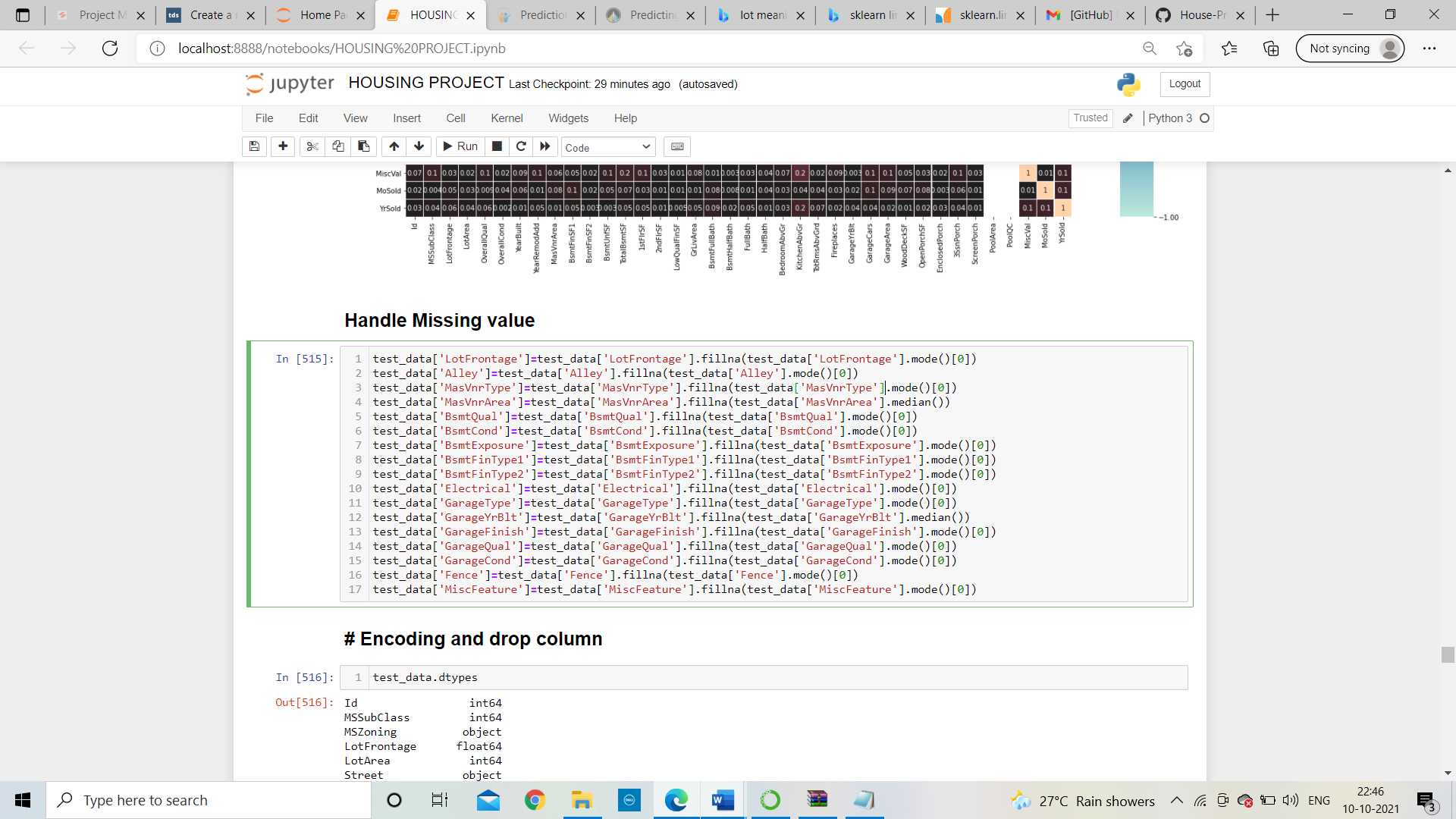
* Correlation with saleprice?



In above plot we can found the most important factor i. e. overall quality of houses.

# As shown in heatmap overallQual is highly correlated with target variable i. e. SalePricee. Hence, it is an important variable to predict the saleprice. Through this variable the overall quality of house can be described

* Data Pre-processing Done
  1. Work on missing value



With mean/median and mode we can solve the problem of missing values. If the data is categorical than use mode method and if column is continuous in nature than put mean/median method.

* 1. Drop unnecessary columns

Columns which are having poor correlation with target variable will be deleted or we can drop that column from the dataset to avoid overfitting problem.

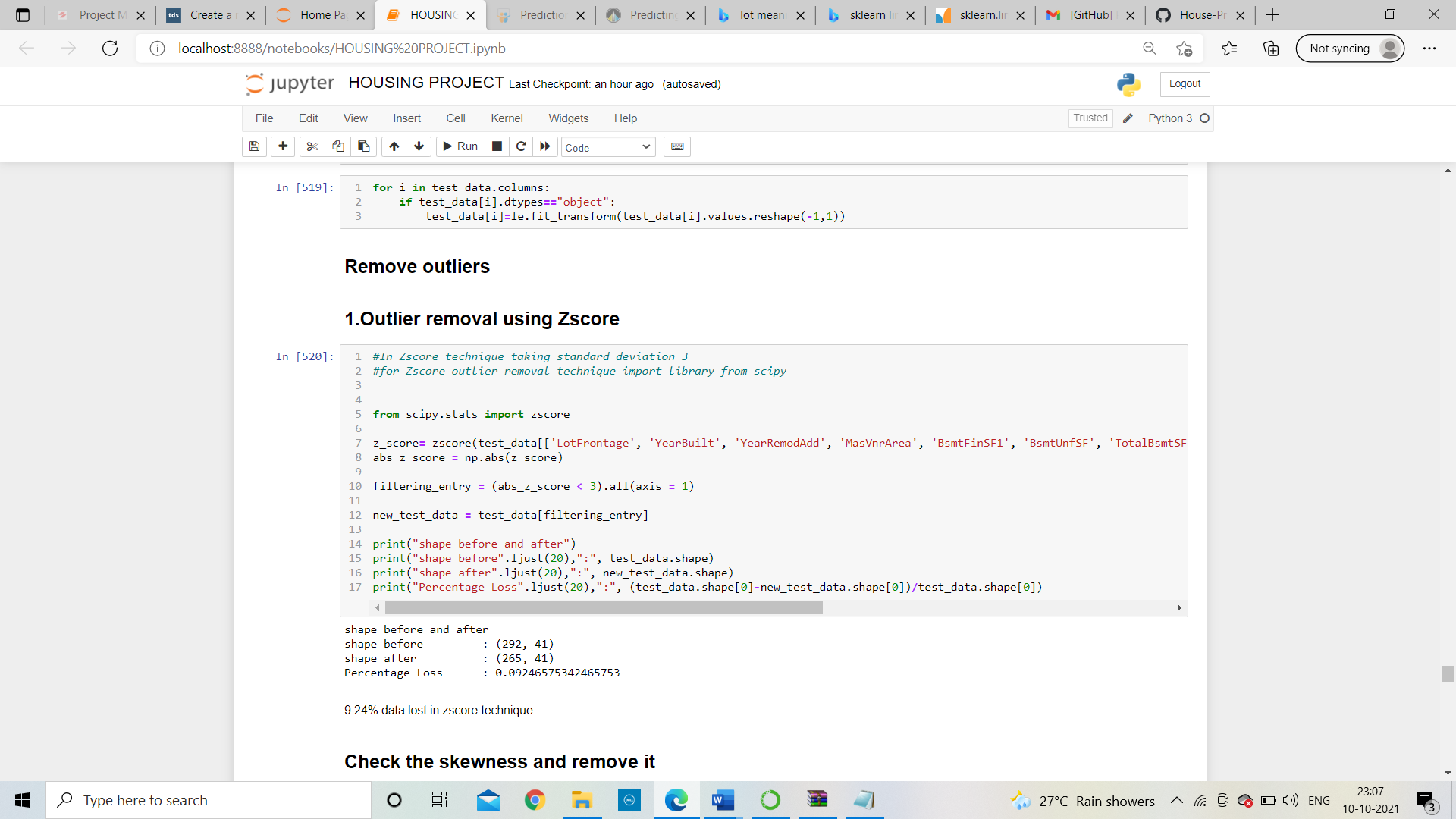
* 1. Encode the columns which are having object data type

Encoding data is very important for model because data will not work with object data type. In this project I will encode the object columns with label encoder.

Label encoder is work on alphabetical order and encode the column.

* 1. Remove outliers using zscore technique

We can use any technique to remove the outliers like z score technique or Inter Quartile range method but our first preference should be less data loss. We have to go with that technique in which less data will be lost.



Shown above 9.24% of data will loss after applying Z-score technique

* 1. Remove skewness

For continuous columns which are having skewness value greater than +0.5 to -0.5 then it is important to skewness from that.

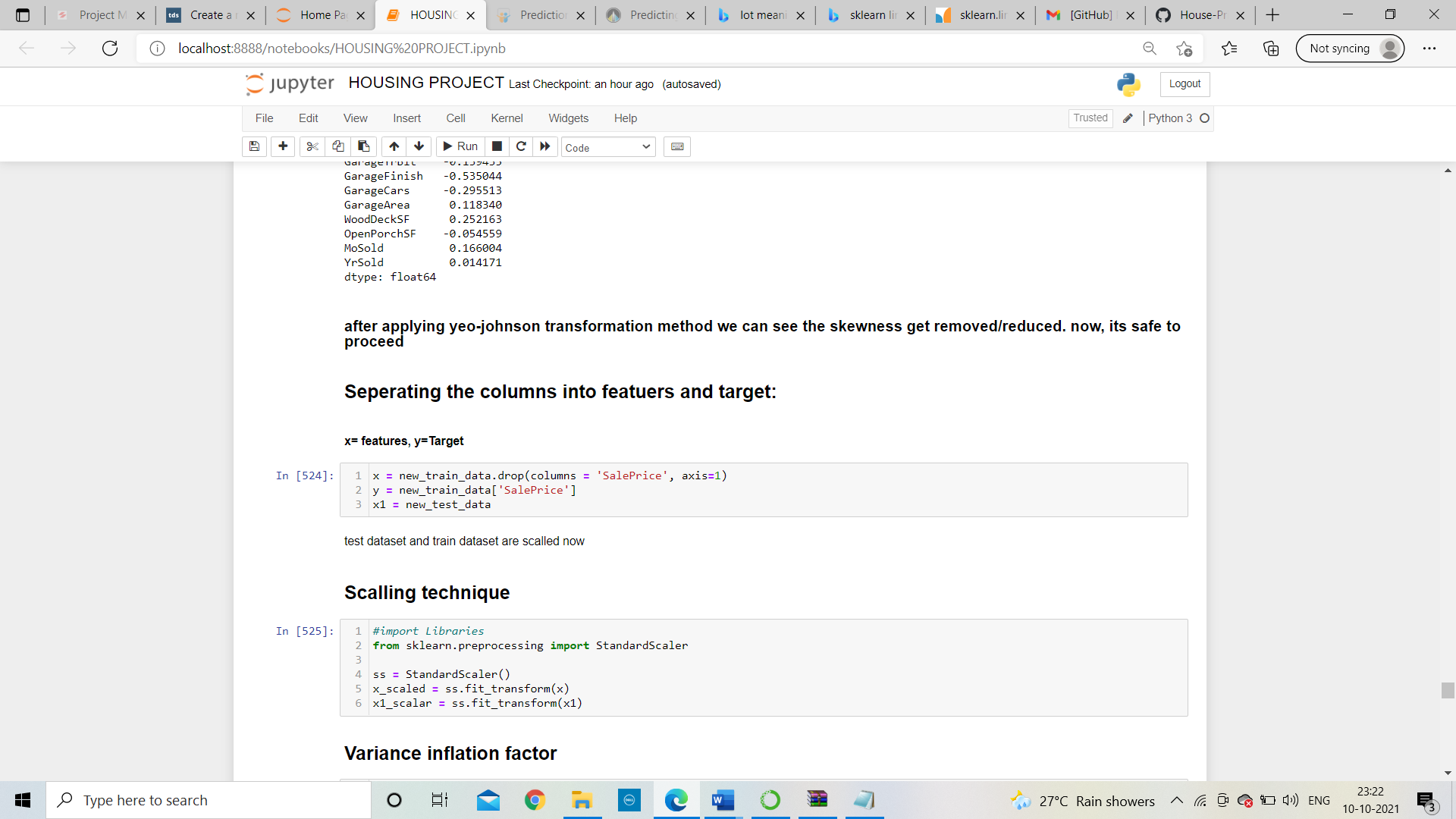
* Hardware and Software Requirements and Tools Used
  1. **Notebook and Data:** [GitHub](https://github.com/LearnDataSci/article-resources/tree/master/Housing%20Price%20Index%20Regression), Jupyter Notebook
  2. **Libraries:** numpy, pandas, sklearn, scipy, joblib, matplotlib, seaborn, statsmodels

**Above all steps we need to done with test dataset also.**

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Now we will train several Machine Learning models and compare their results. We need to use the predictions on the training set to compare the algorithms with each other. Later on, we will use cross validation. After that I will compare the difference of accuracy score and cross validation score of all models. The model which is having minimum difference will be consider as a best model and further procedures will done on that model. Following are the basic steps of model building.

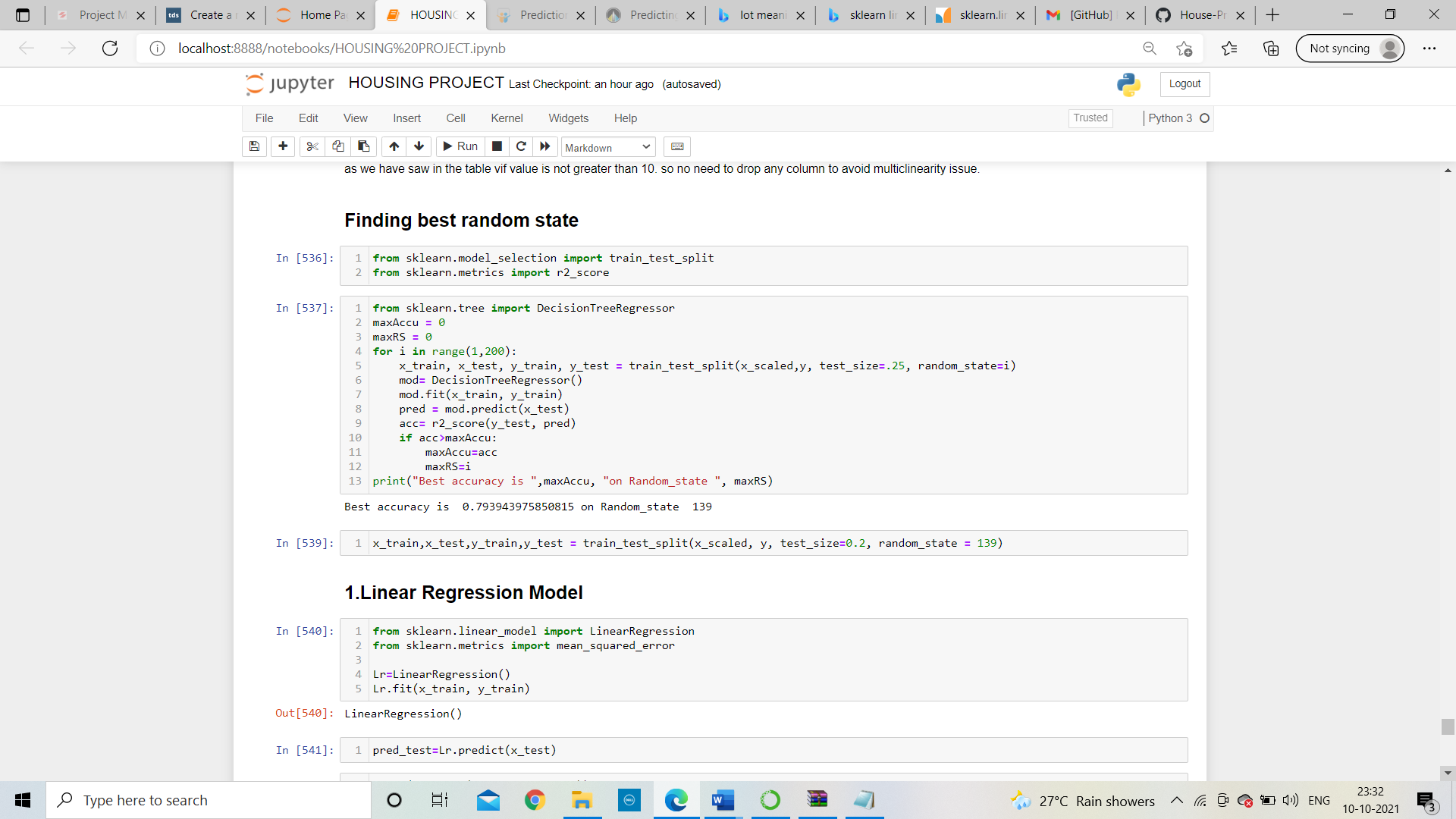


The values of input variable are scaled

Variance inflation factor is use to find the amount of multicollinearity in independent variables. If the variation inflation factor is more than 10 for any column, than there will be multicollinearity problem will occur and then we should remove the highest value vif columns. Rest of columns vif value will automatically decreases. If the vif value of any column is greater than 10 we should drop that column to avoid multicollinearity.

* Testing of Identified Approaches (Algorithms)
* Find best random state:

For work on any new data our model needs to train first. So, I divide my datasets into some percentage of training set and some percentage of the test sets. Therefore, its necessity of model to shuffling the training and test set randomly otherwise it will always select from starting. Random state is use to avoid biasness in data prediction.

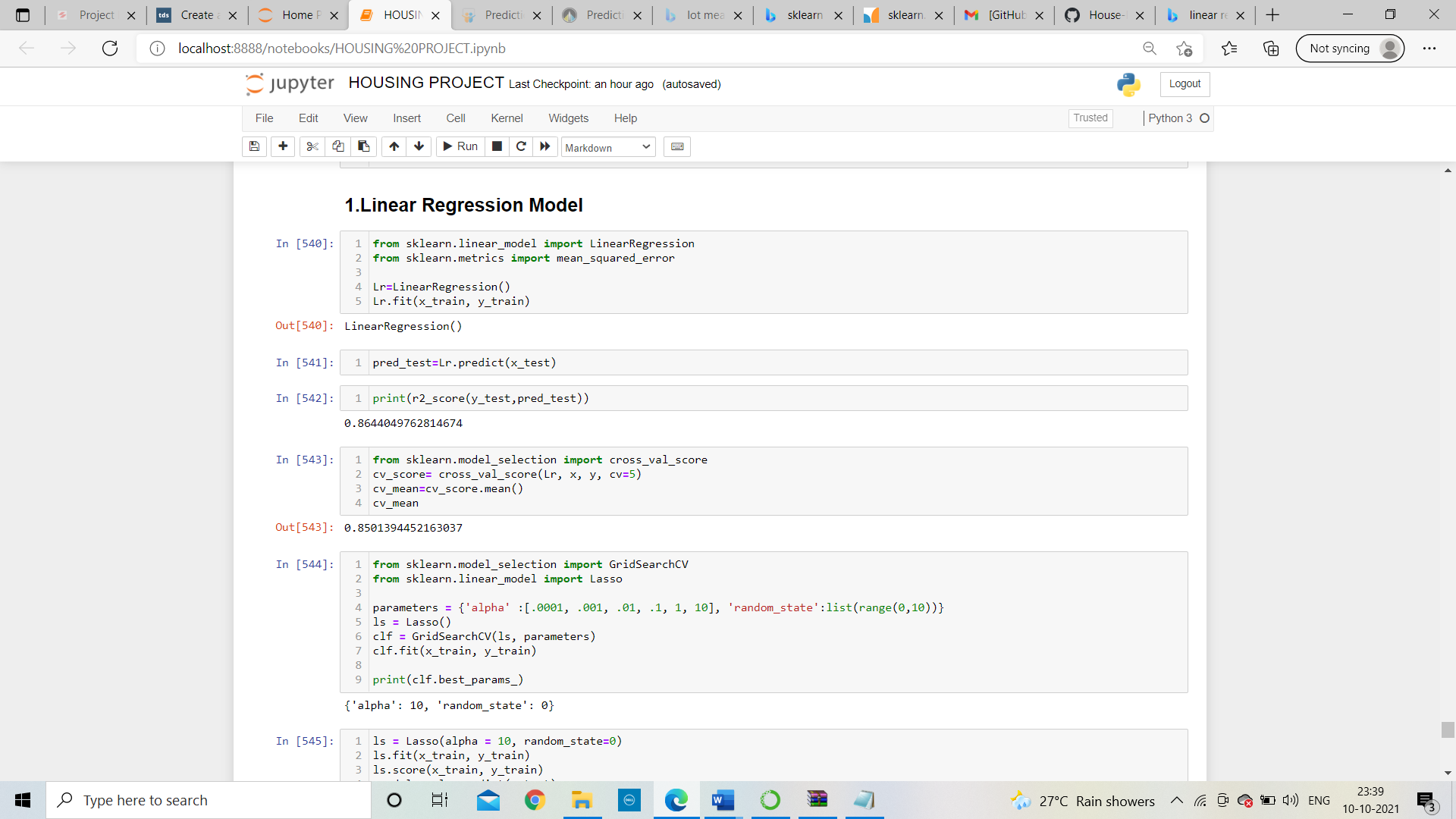


Now, splitting our dataset by train test split with random state: 139

* Run and evaluate selected models

1. Linear regression model

Linear regression is an algorithm used to expect, or visualize, a relationship between two different features/variables. In linear regression tasks, there are two kinds of variables being examined: the dependent variable and the independent variable.



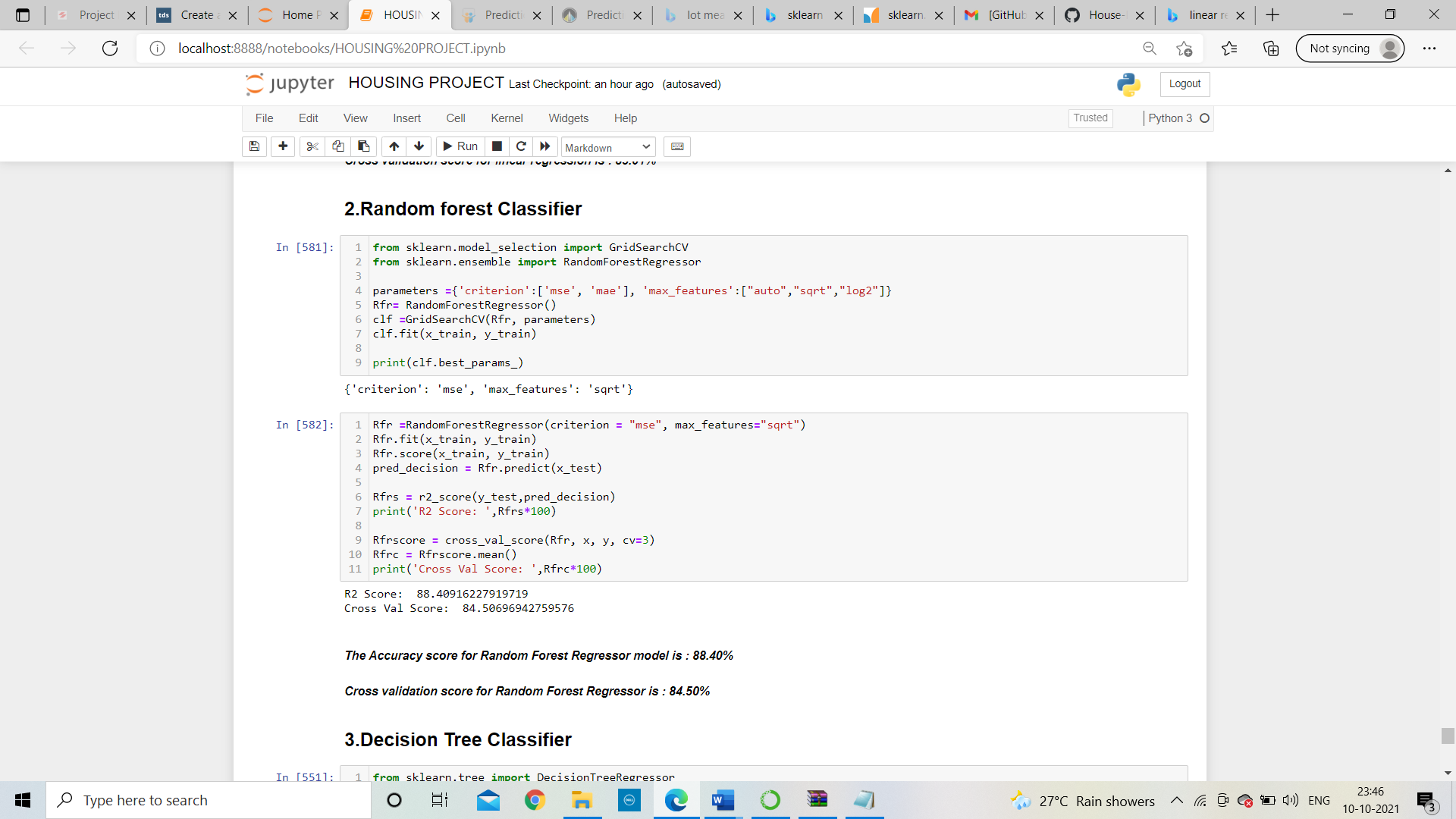
After applying above linear regression method, the output of r2 score is: **86.44%**

And cross validation score is**: 85.01%**

The cross-validation score is used to find weather the model is having overfitting problem or not. If r2 score of models is near to cross validation score than we can conclude that our model is not overfitted else if the value is having too much difference than the model is overfitted. Let see the output of cross validation score of linear regression.

1. Random forest regressor:

Random forest regressor is a type of supervised machine learning algorithm.



It is used in both classification as well as regression model. It made up with decision tree. It gives the single result which is made with combination of multiple decision tree output. In the similar way of logistic regression, I will apply the random forest algorithm and find the accuracy score and cross validation score also.

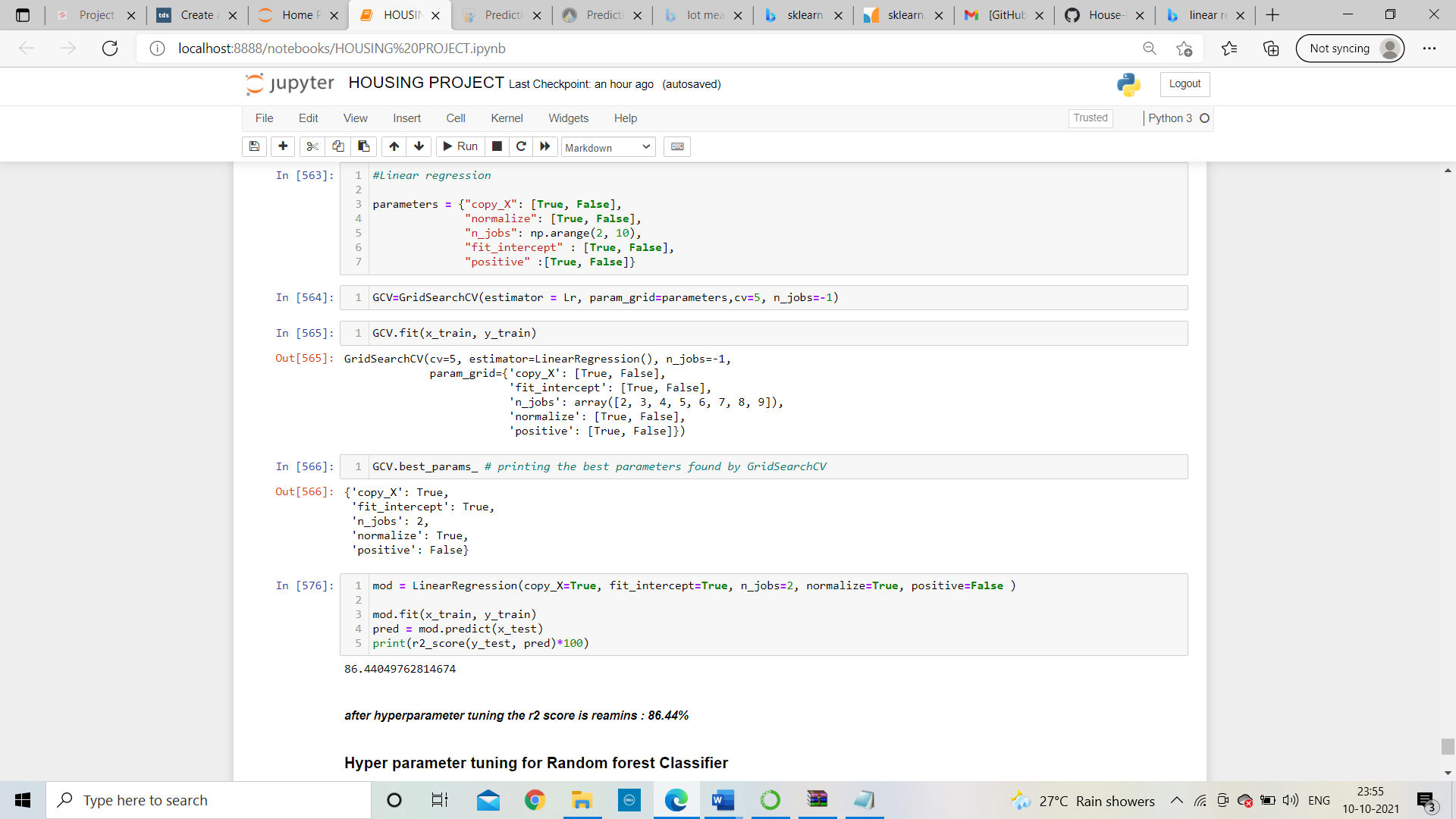
* Interpretation of the Results
* **Comparison of all models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr.  No. | Algorithms | Accuracy score  (a) | Cross  -Validation Score(b) | Difference  (a-b) |
| 1. | Linear Regression | 86.44% | 85.01% | 1.43 % |
| 2. | Random forest regressor | 88.40% | 84.50% | 3.8 % |

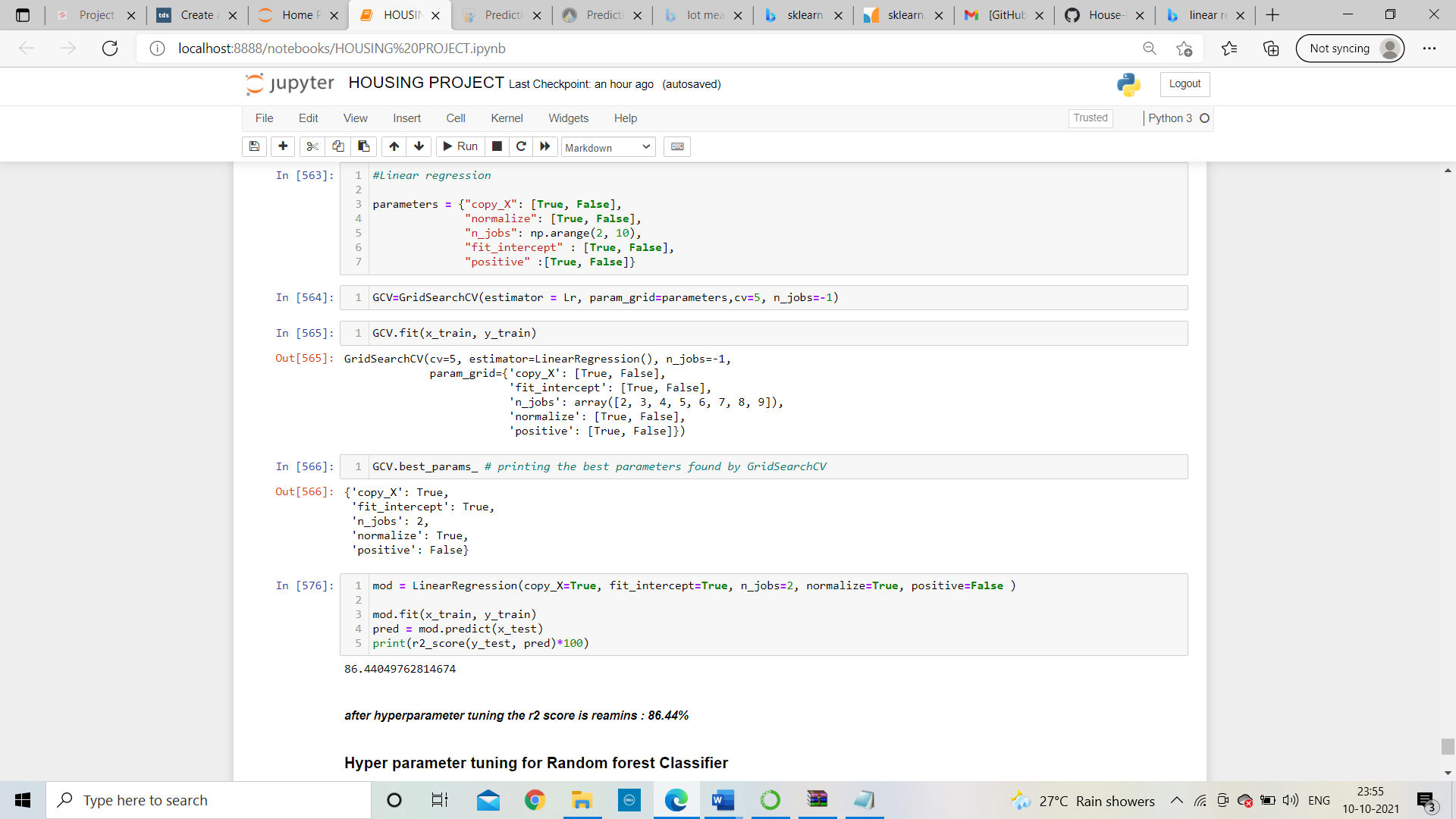
as shown in above table Linear regression is having minimum difference, so Linear regression is a best model. Now, I will work on best model with hyper parameter tunning.

**Hyper Parameter tuning:**

Grid search cross validation (GCV) in hyperparameter tuning will help us to finding the best hyperparameter and tune it with training data set and give best r2 score:



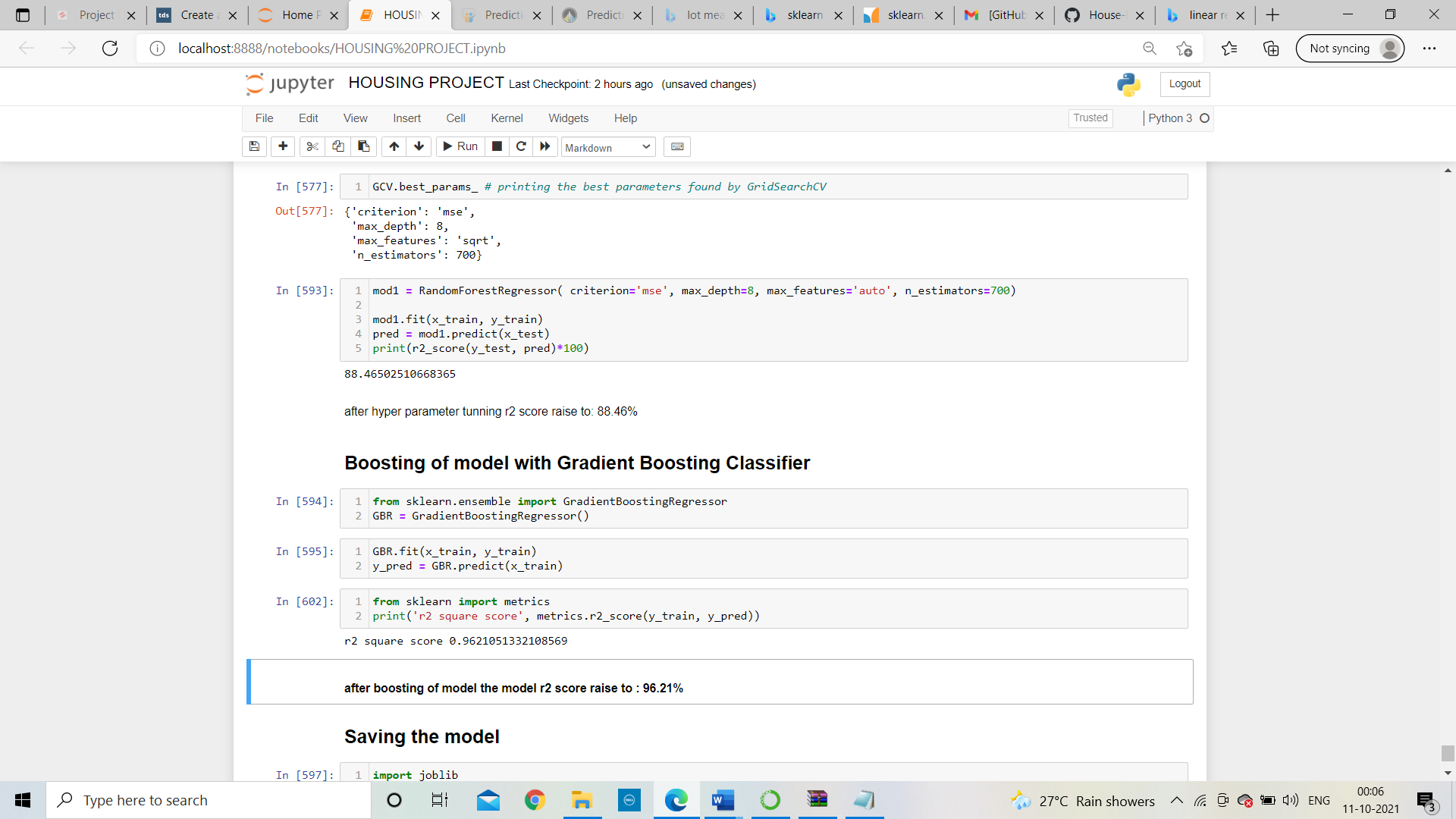
In above figure we can see the default hyperparameter which I was tunned with linear regression with GCV and then find the best parameters, now fit it in to model and find the r2 score;



##### *After hyperparameter tuning the r2 score is remains: 86.44%*

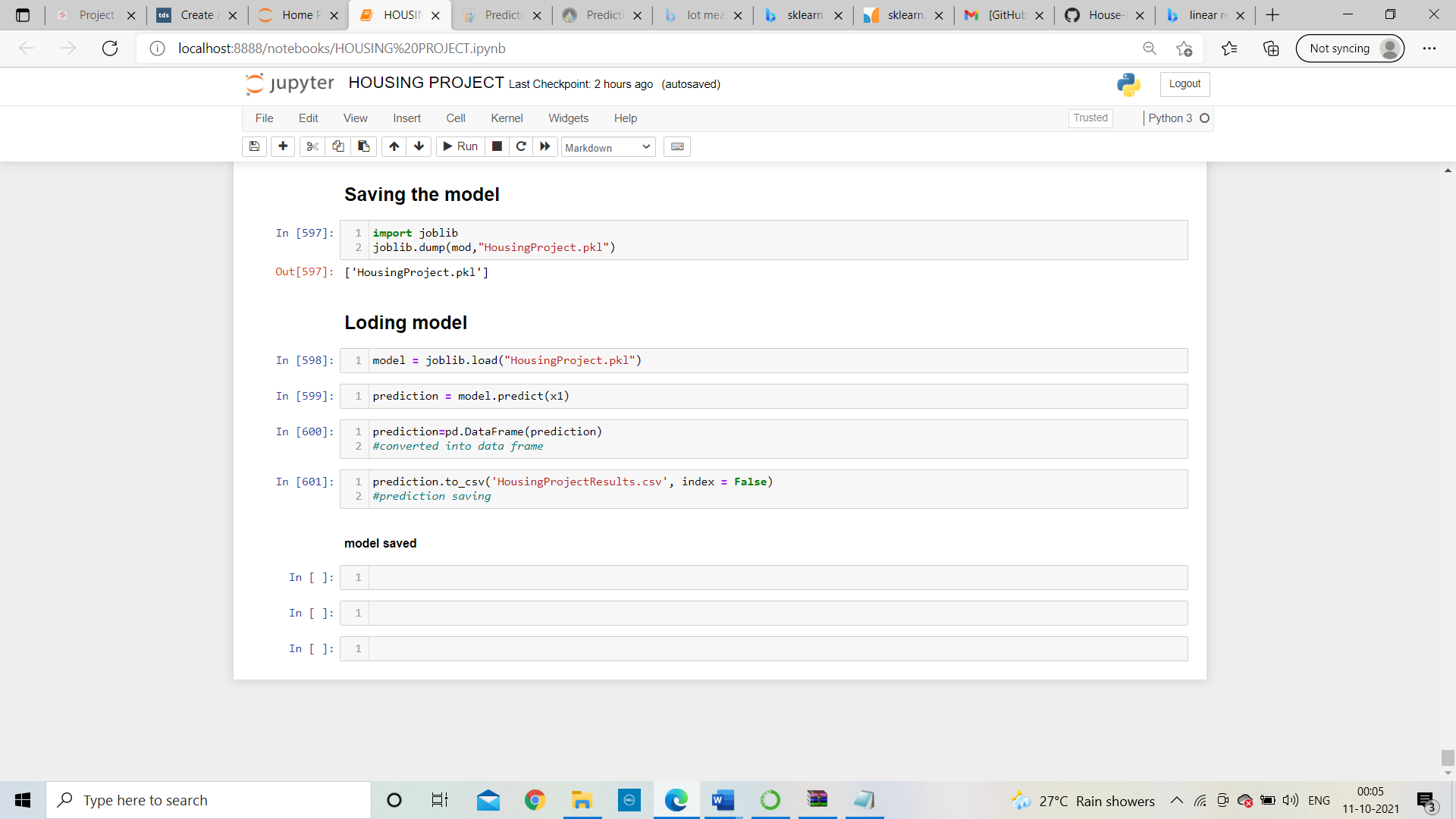
**Boosting of model: -**

With the help of Gradient Boosting Regressor, I will boost my model and then my r2 score raise to:



* **Saving and loading the model:**

I save this model with name “HousingProject.pkl”



Here we need to take test dataset for prediction and I had saved the result as name: “HousingProjectResults.csv”

**CONCLUSION**

* Key Findings and Conclusions of the Study

# After observation on data and as shown in diagram overall Quality is highly correlated with target variable i. e. Sale Price. Hence, it is an important variable to predict the sale price. Through this variable the overall quality of house can be described.

# 

# As we see if the quality of house increases than selling price will also increase.

# We fit our training data into the gradient boosting model and check for accuracy

# We got an accuracy of****96.21%**** which is amazing.

# 

# In above image shows overall quality is positively correlate with sale price but evaluates the height of the basement will negatively correlate with target.

* Learning Outcomes of the Study in respect of Data Science

As I mentioned on each and every step which method will used and why like in missing data we will use mean, median and mode method

For removing outliers, I used z-score technique because in IQR technique too much data will be lost.

In skewness removal I had used yeo\_johnson method because it works with positive as well as negative data also.

Best model is linear regression, r2 score and cross validation score is having small difference compare to other algorithms.

* Limitations of this work and Scope for Future Work

As I mentioned above as the population increases simultaneously demand of houses will also increase. So, this project is having huge scope according to future perspective but there are some limitations also present in it similarly the prices will go high as the demand increases.