**Name of the project:-**

**HOUSING: PRICE PREDICTION**

**Submitted by:**

**Sujay Anil Bharmal.**

**ACKNOWLEDGMENT**

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

**Introduction:- Problem Definition**

Houses are one of the necessary needs of every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors to the world’s economy. It is a very large market and various companies are working in the domain. Data science comes as a very important tool to solve problems in the domain to help companies increase their overall revenue, and profits, improve their marketing strategies and focus on changing trends in house sales and purchases. Predictive modeling, Market mix modeling, and recommendation systems are some of the machine learning techniques used for achieving the business goals of housing companies. 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. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. This company wants to know:

**Business Goal:**

we are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**Data Analysis:-**

In this project, we have 80 inputs and 1 output that are available according to my information I see that this case is a supervised learning case.

First, we import all the necessary libraries. Like pandas, NumPy seaborn, and matplotlib, and then run that cell. After that, we read the all data and save it into the df folder. There are 1168 rows and 81 columns. There are 80 columns of the input and price is the output variable.

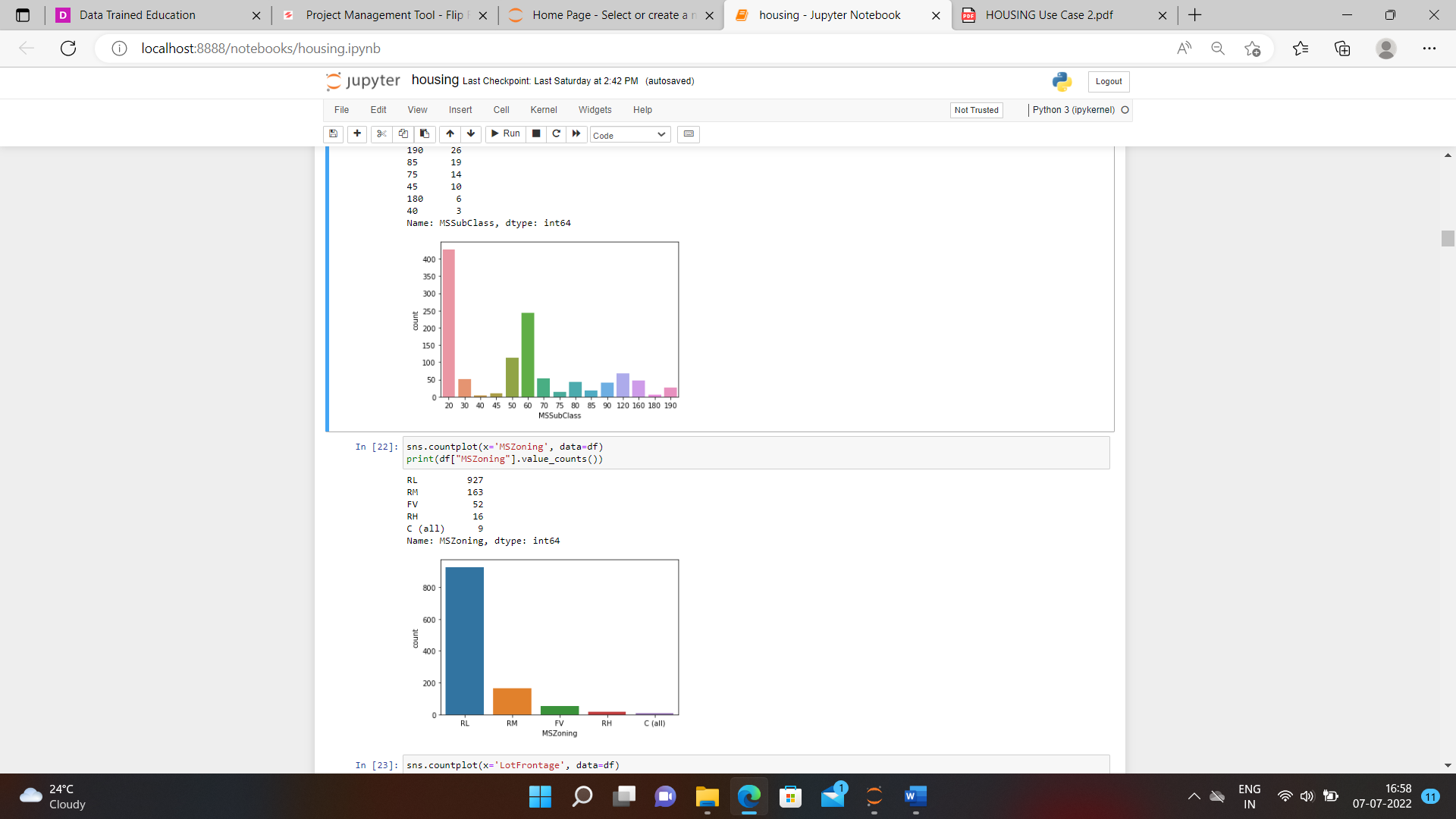
First, we see that there are in this dataset’s null values are present or not, we give df. IsNull().sum() code to run and we see that there is no null value present.

Then we check all column datatypes by using the df.info() code.

We see that in that data set there are lots of null values present so we clear that problem first, firstly we drop the columns there almost 60% of data are null values so we delete the 5 columns where their null values are more than the 60 %. Then use this method df=df.dropna(how='all',axis=1) after that we lastly used the filling method df=df.fillna(method='pad') and see that there are any null values are presents or not. Lastly, we use df=df.drop a() method.

**EDA process:-**

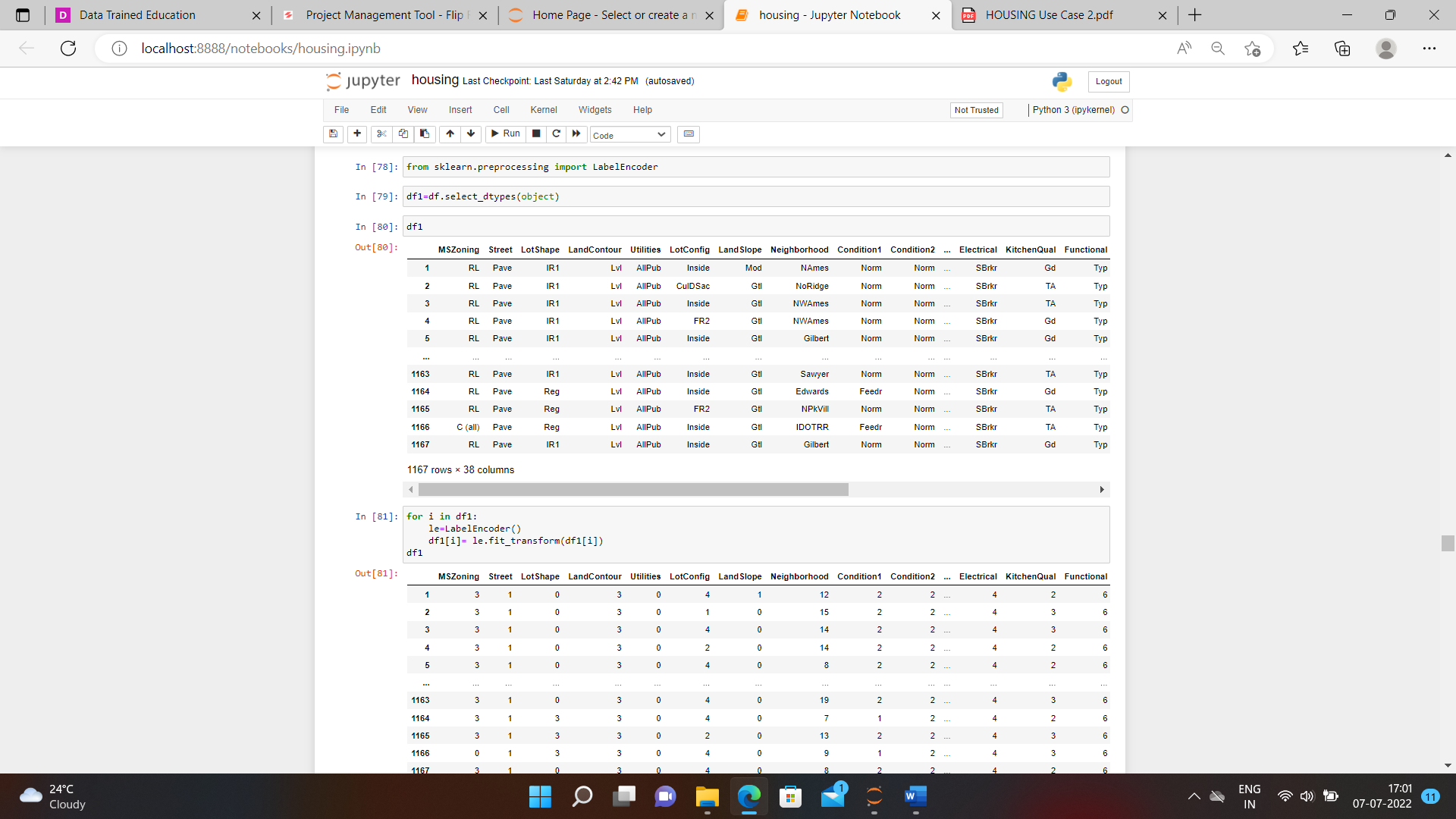
For the EDA process, we use the seaborn and matplotlib libraries,



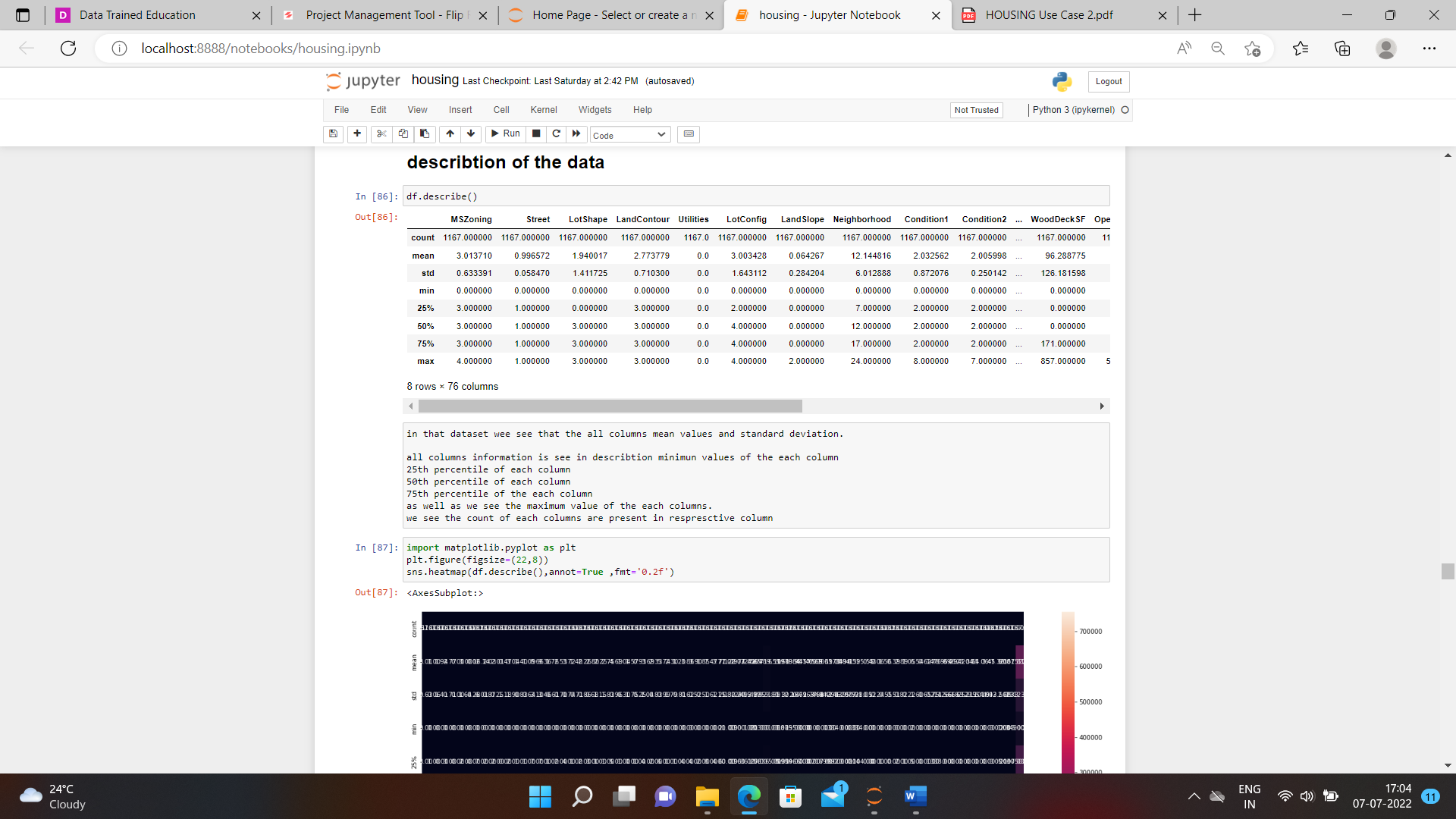
We see that in that photo we use the counterplot of all necessary columns.

Then we see that the data is in the categorical form so we convert it to numeric format because the machine only knows the numeric data.

Using the label encoder method we convert,

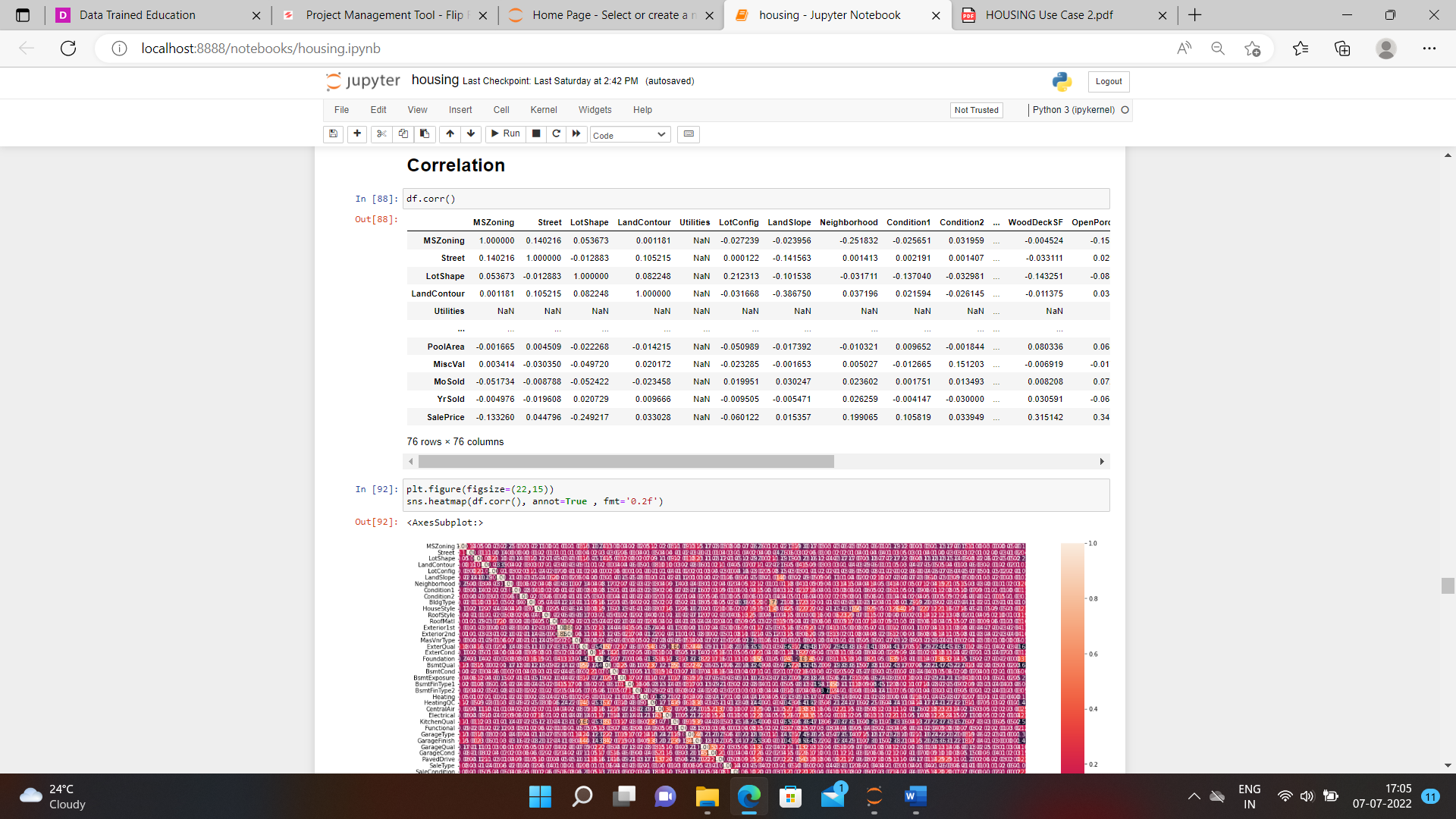


After that we go further step is called the description of the data.



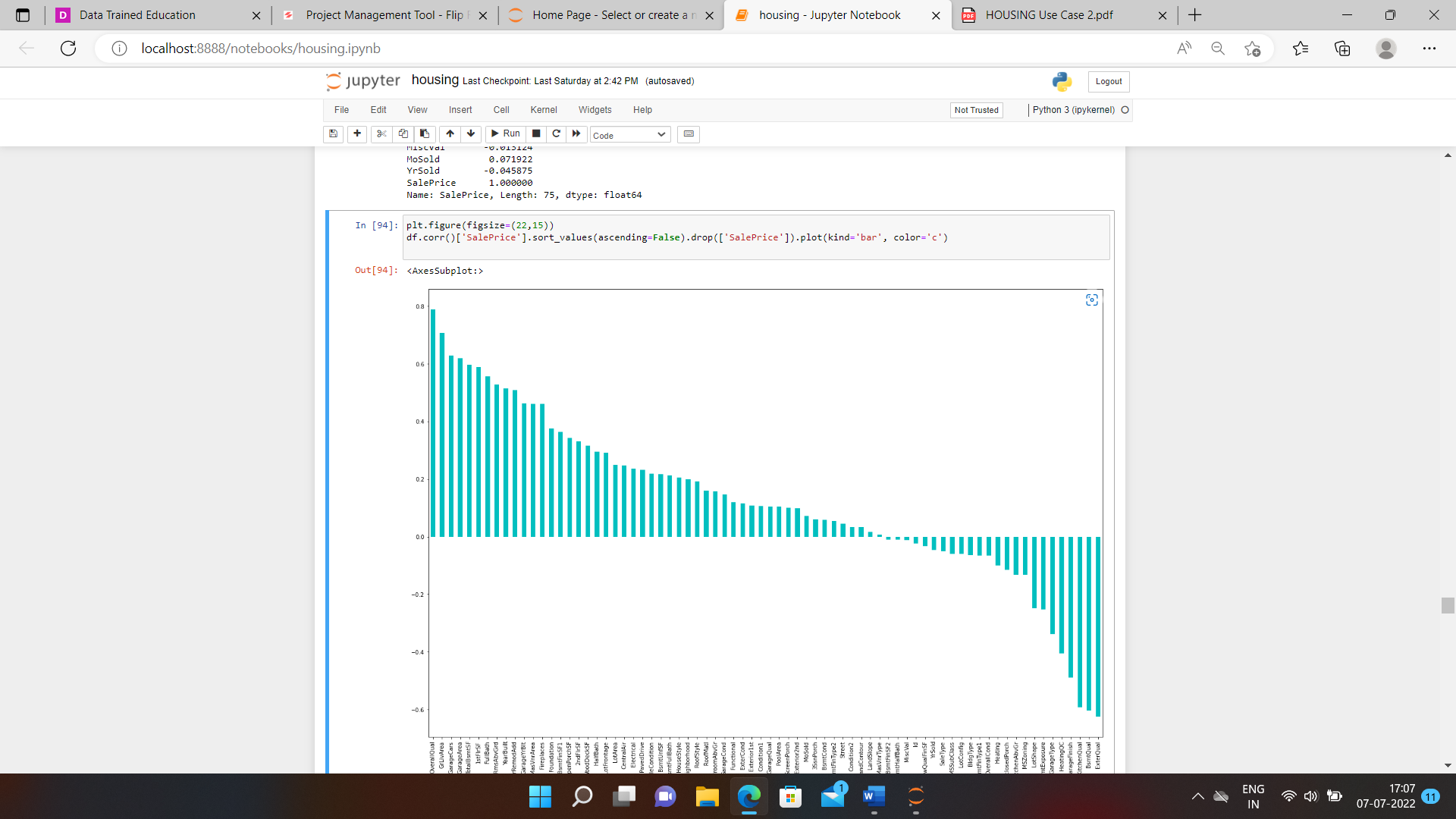
Then we come to the next step called describing of data and which we see using the df. describe() method. And also see the there count, mean, standard deviation, minimum value, 25 percentile, 50 percentile as well as 75 percentile, and maximum values of the data set.

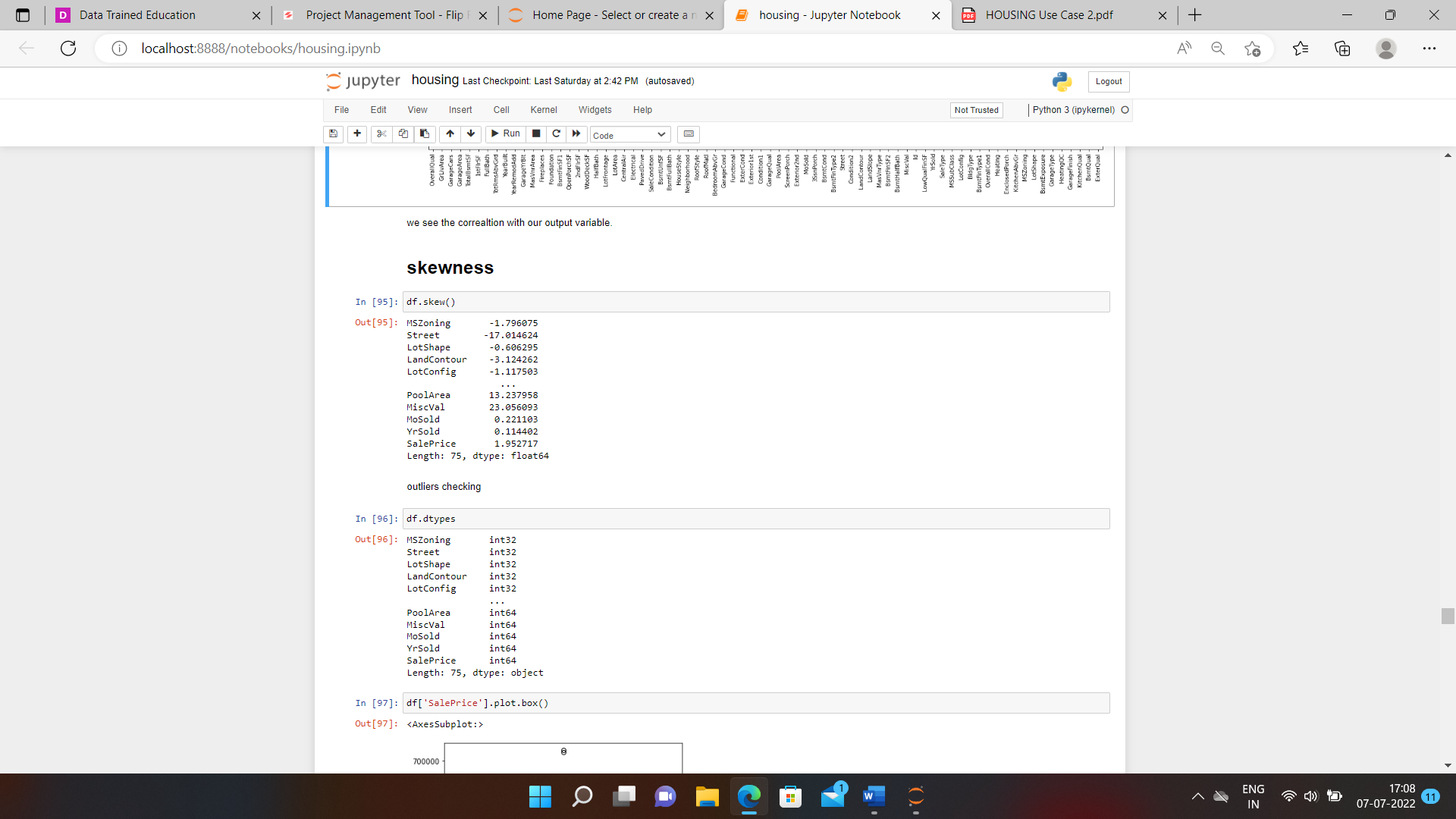
After that, we use the correlation function and see the co-relation of columns with each other. Use code df.corr(). And plot that correlation using the matplotlib.



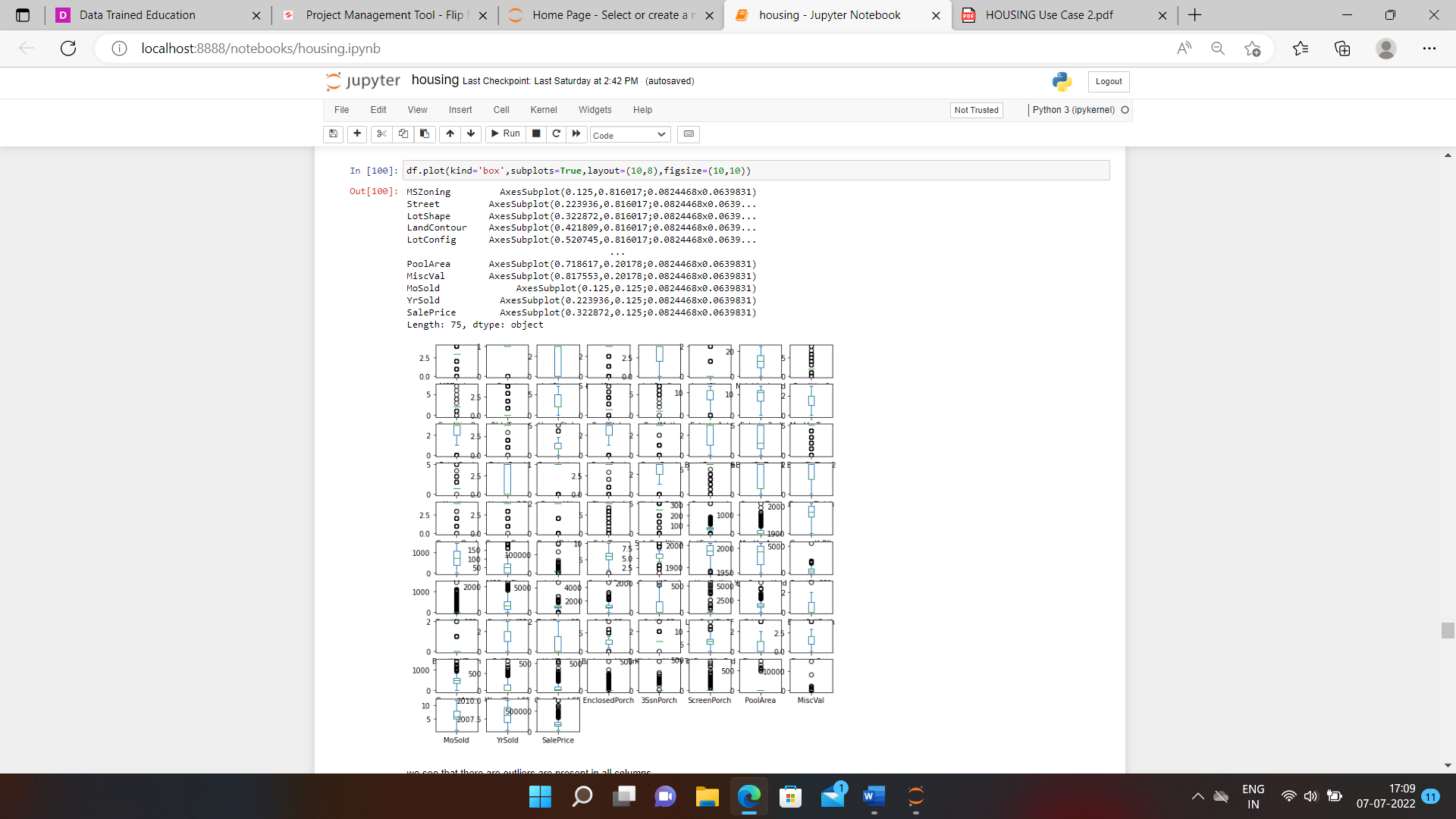
We see the correlation with each other.

After that, we go further skewness of the data so we use the df. skew() method to see the skewness of the dataset. And also see the distplot of all the columns.

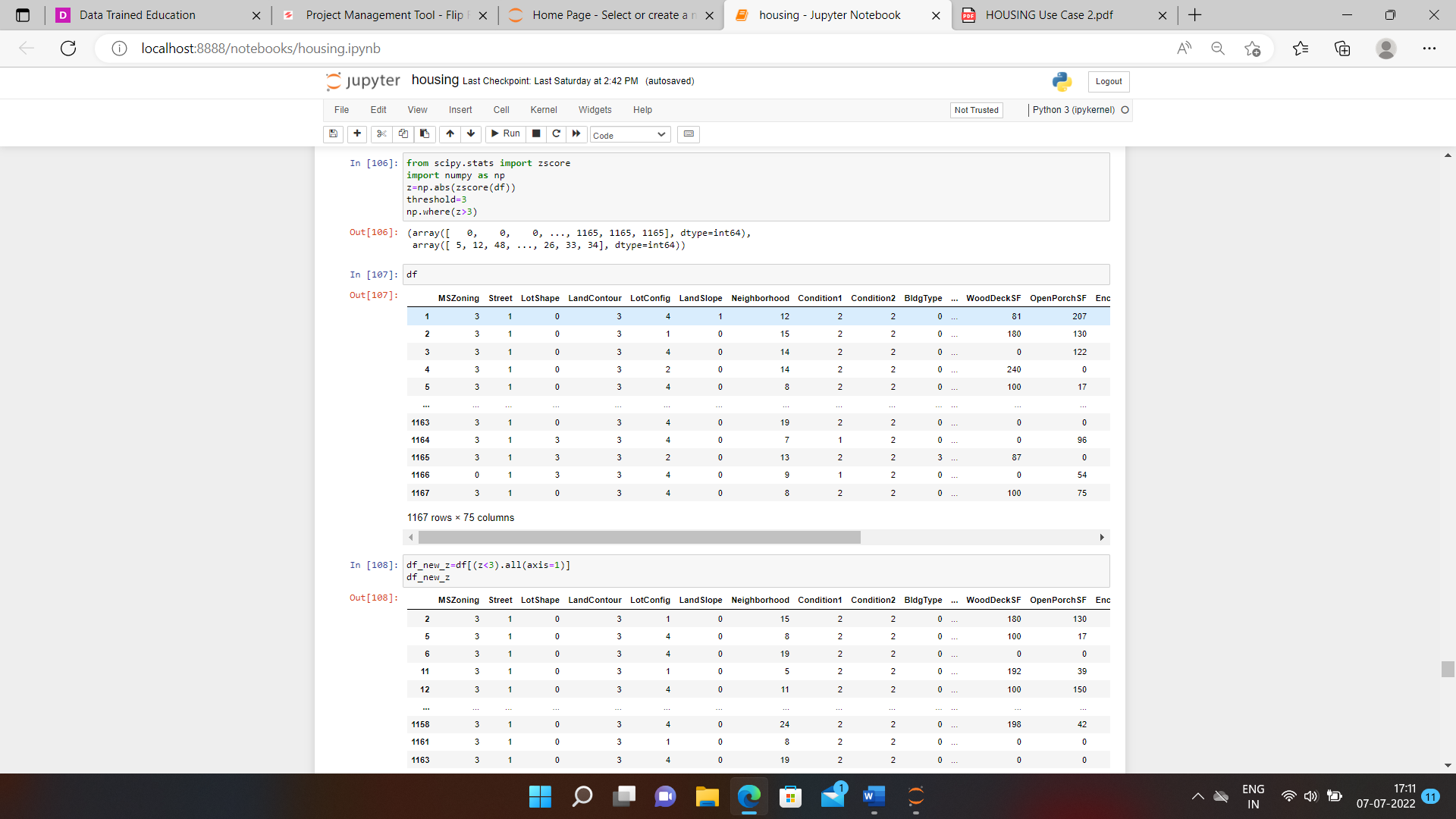




We see the skewness and using the distplot we visualize that, in this dataset, there are outliers present using the boxplot we see the outliers.



We see that there are lots of outliers present so using the scoring method we remove the outliers.



We see that we remove the outliers and our EDA process will be completed.

**Building Machine Learning Models:-**

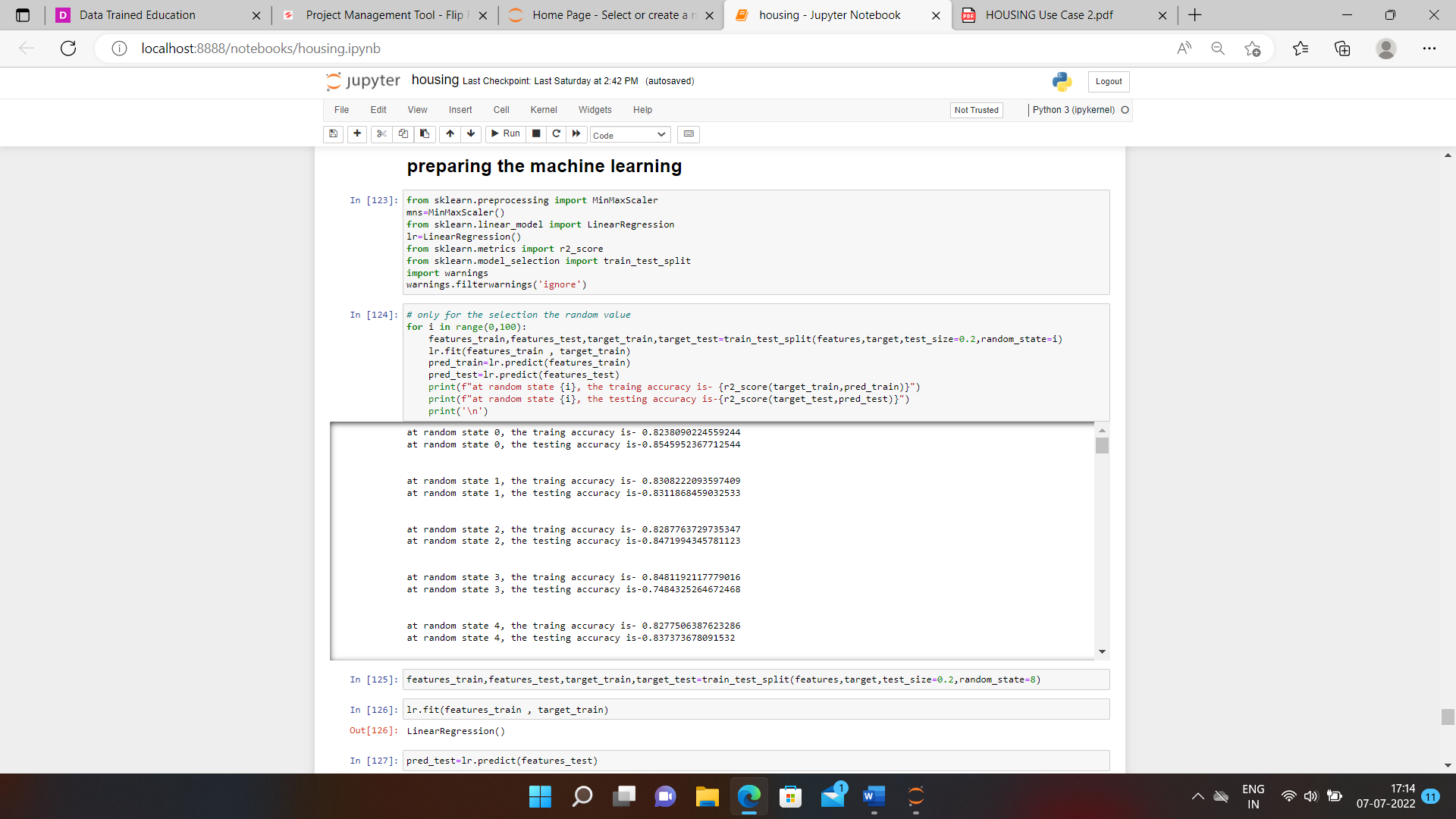
Firstly we separate the target and features in that problem the sales price is the target variable so we used that and there are also the new test data that also we used.

We see that sale prices are in continuous form value so this is the way that this is the linear regression method we apply.

Firstly we import all libraries are MinMaxScaler and this is used for all values are in one shape, then we import the LinearRegression which is used for the regression purpose.

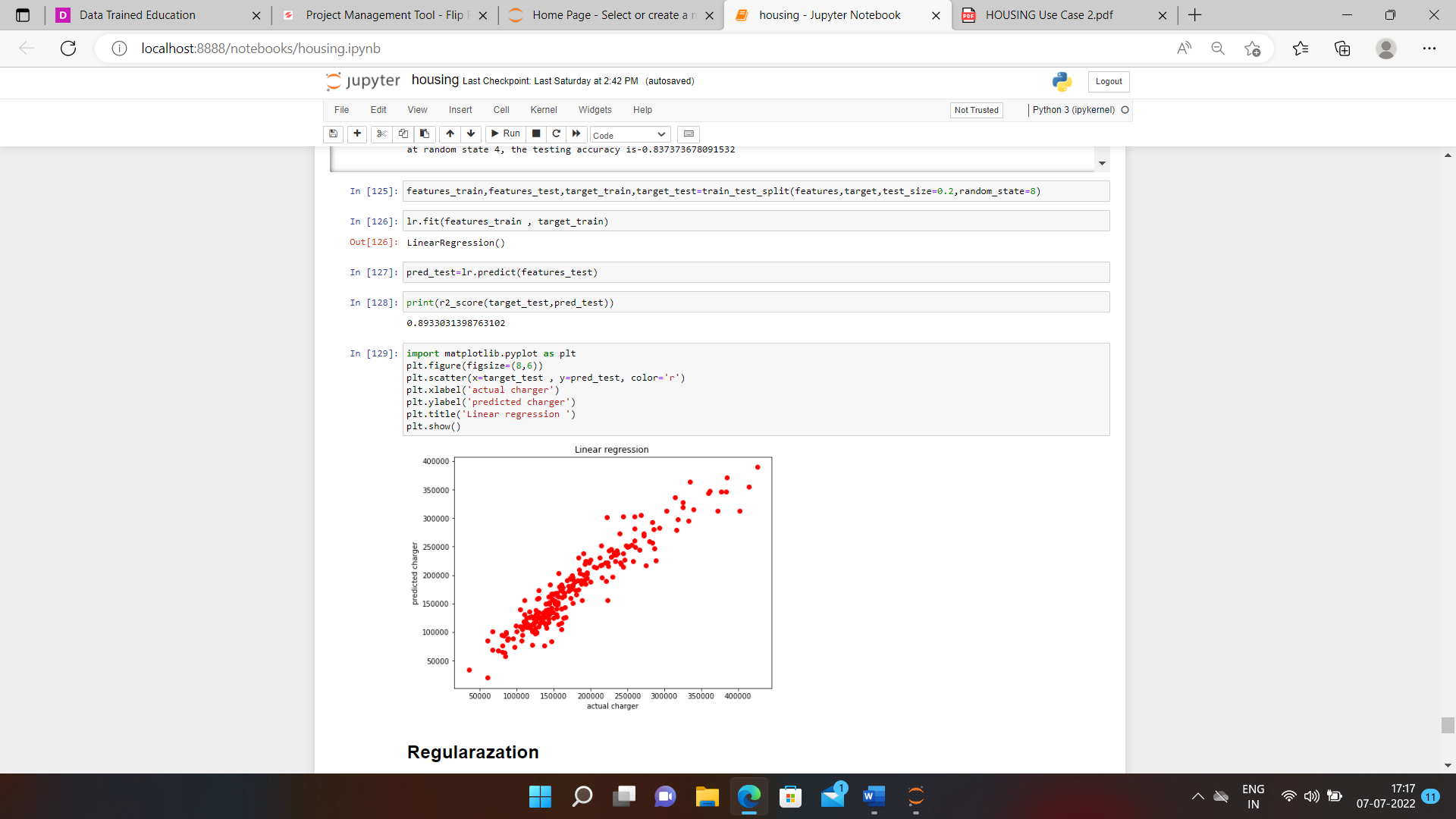
For accuracy score we import r2\_score

And splitting the data we use the train \_test\_split method.

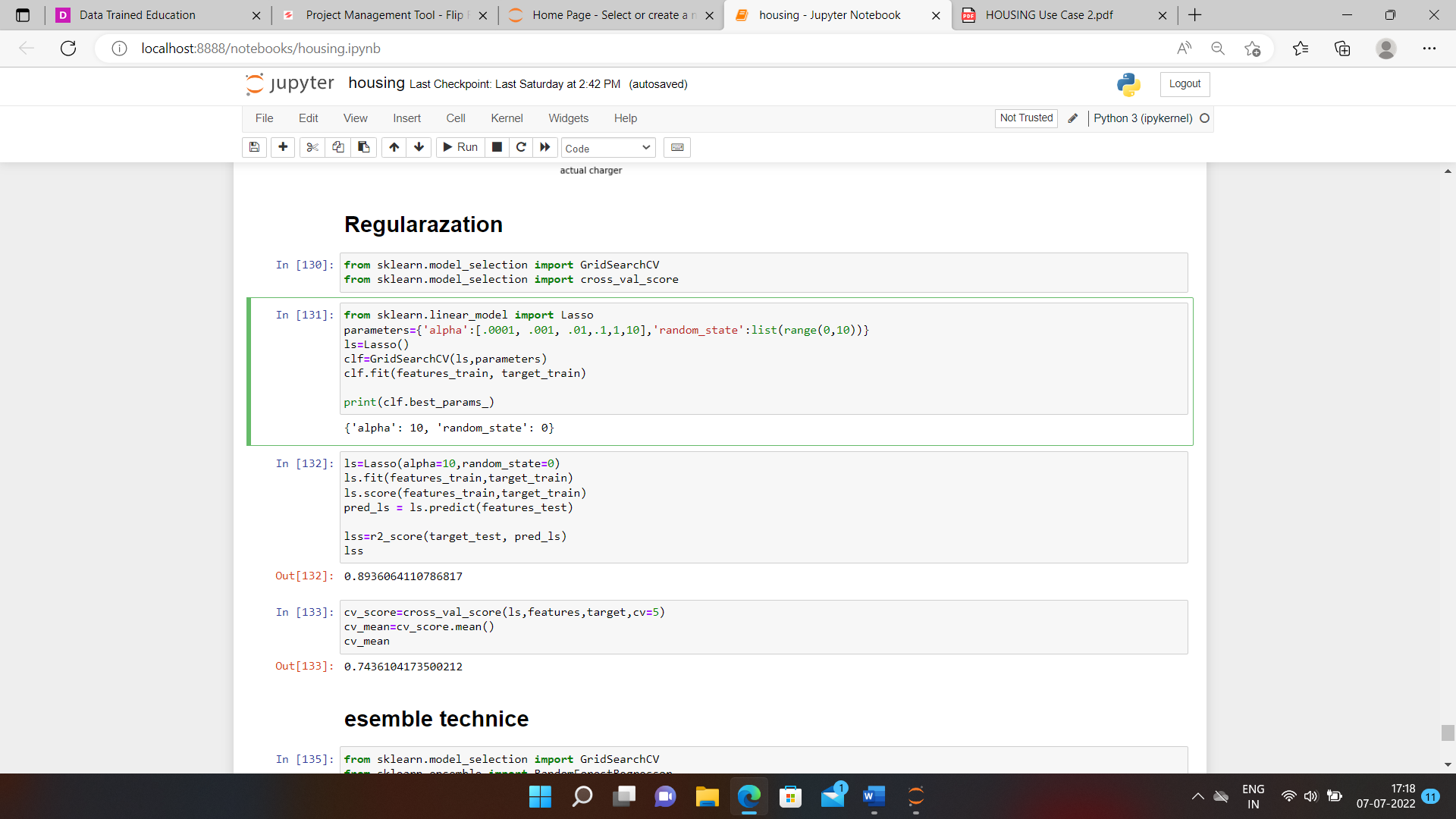


After the train test method, we see the accuracy score as well as the R2 score.

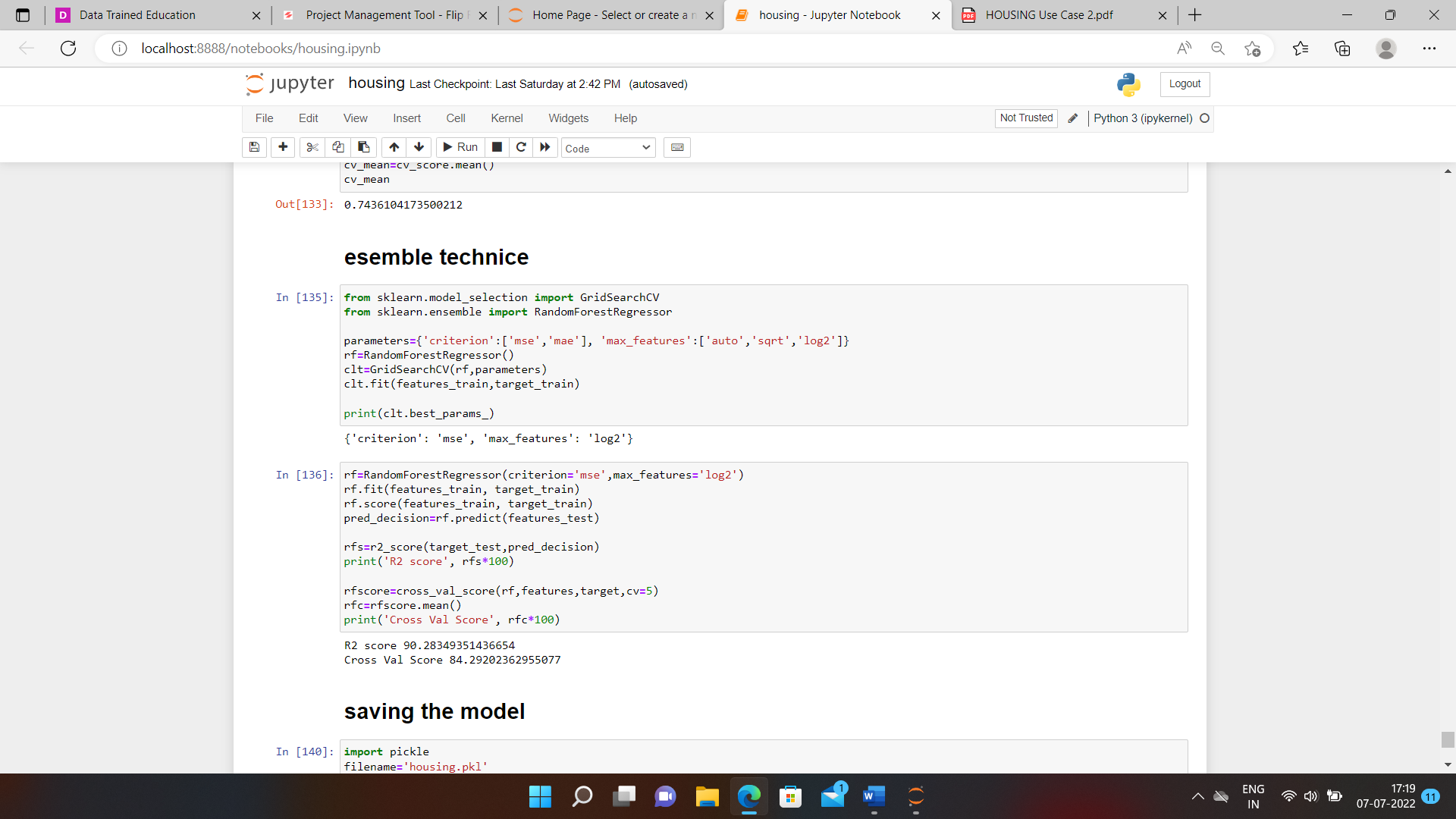
Then we see this result in graphical form.



After the result will see then we used the regularisation method for overlapping and gridsearchcv.



Using the regulation we see the best model then we used the ensemble technique.

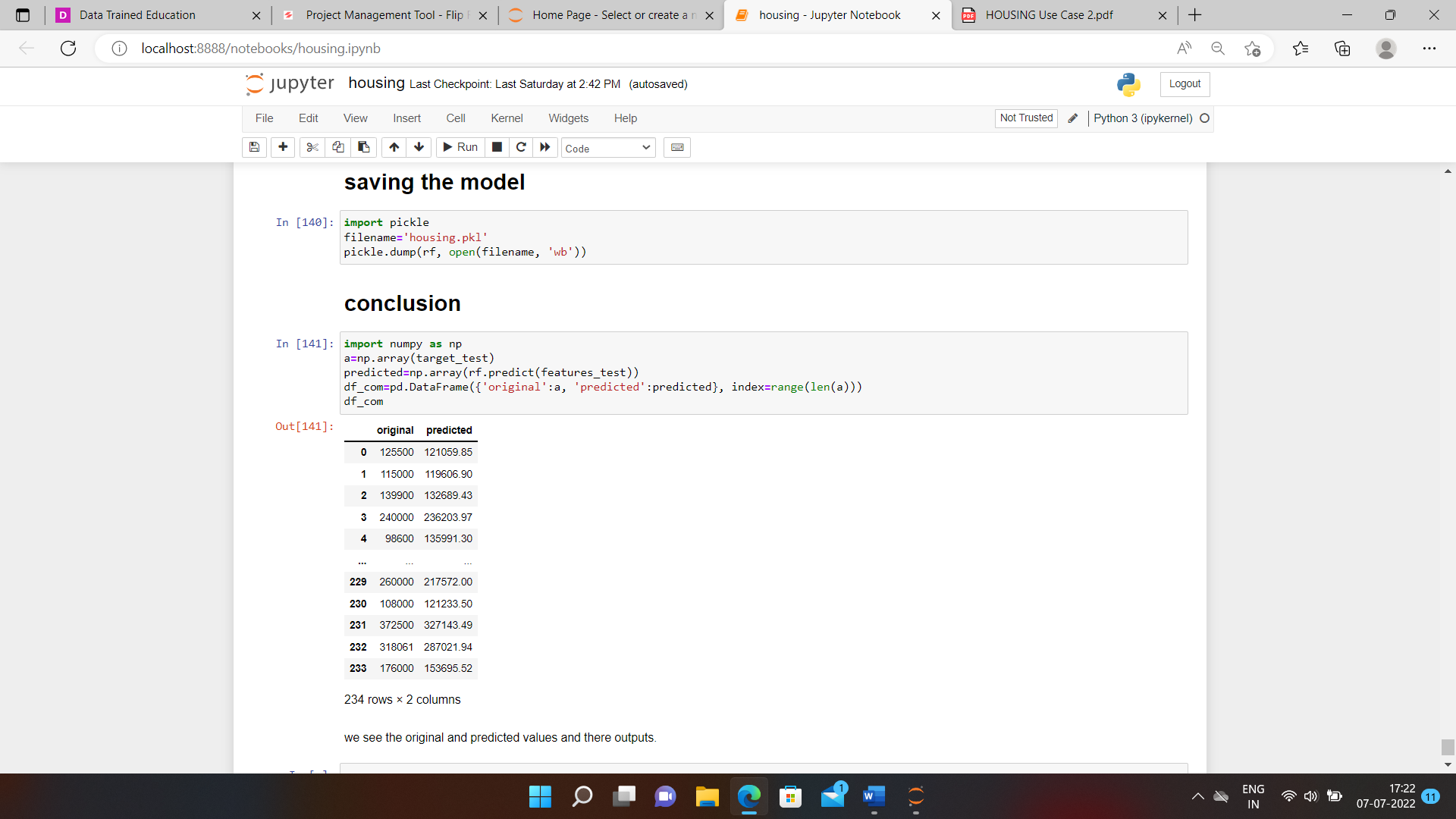


After that, we see the exact accuracy and cross-value score

R2 score of the model is 90%

Cross Val score is 84 %

And this is the end of the machine learning model then we save that model using the pickling method and lastly, the conclusion or the testing data answer is shown.



We see the predicted value as well as the actual value for that.