ICP5 Report

Harika Gurram

Class ID : 17

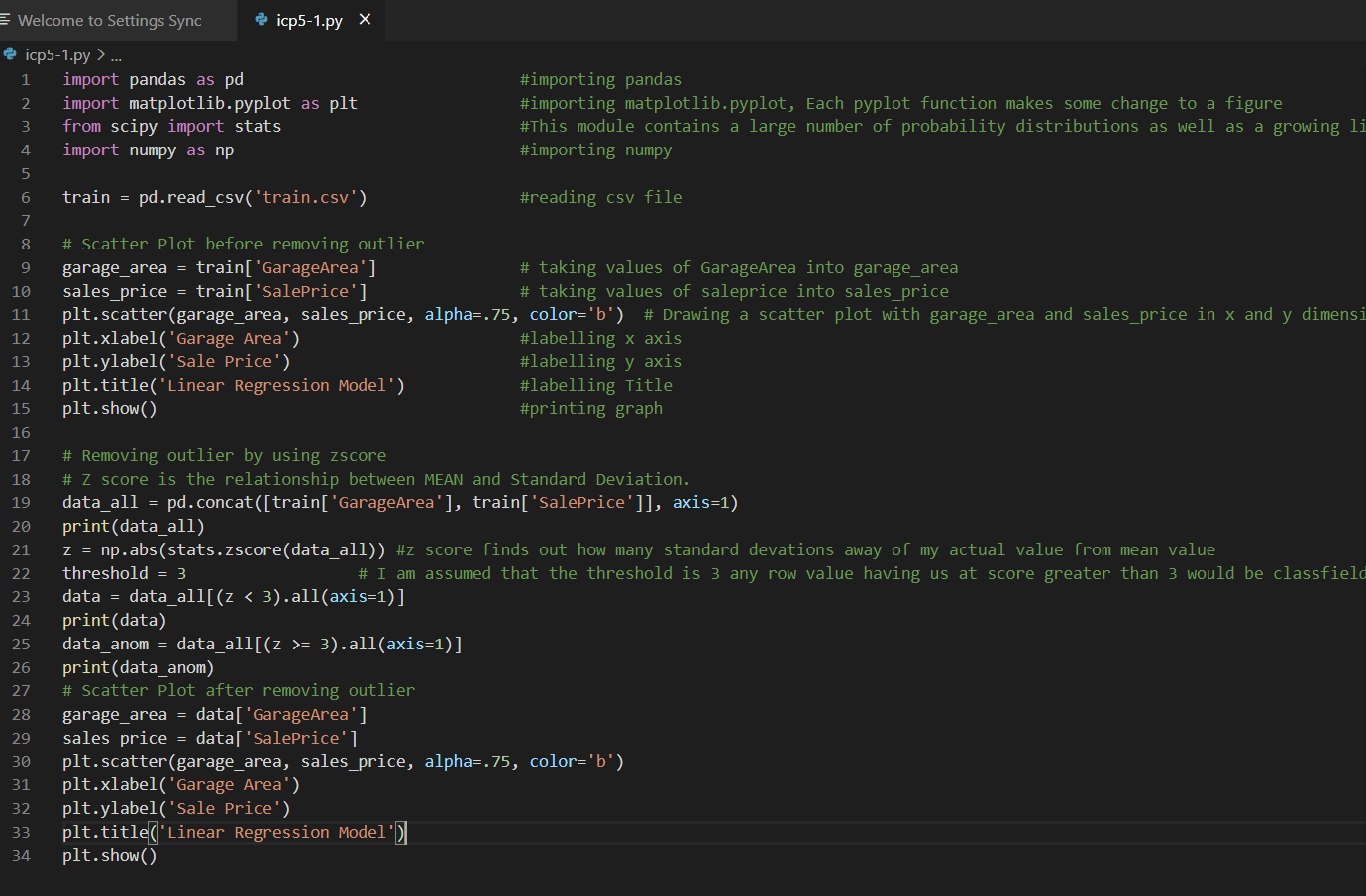
**Question 1) Delete all the outlierdata for the GarageArea field (for the same data set in the use case: House Prices).**

**for this task you need to plot GaurageArea field and SalePrice in scatter plot, then check which numbers are anomalies.**

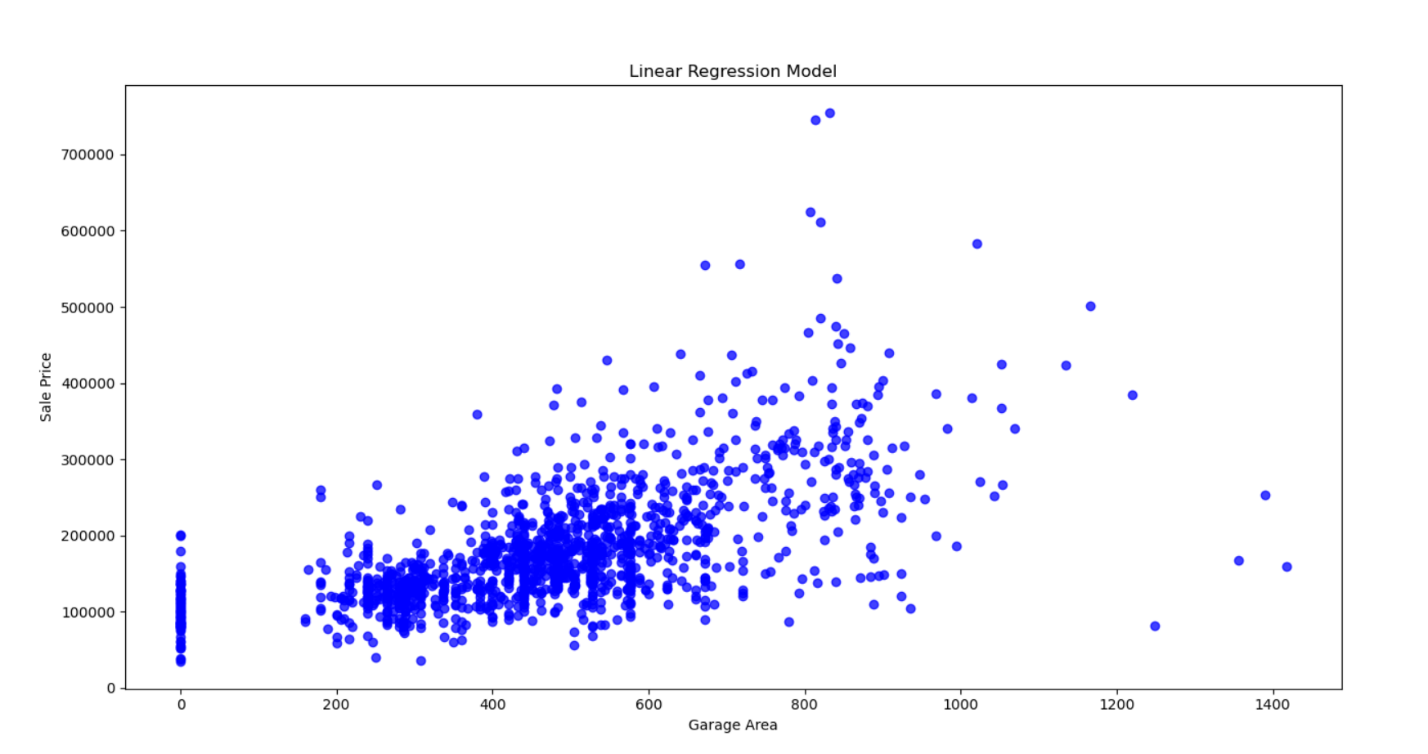
**Code:**

I have used Z-Score function to find out Outlier Values in dataset. 99% of values have a Z-score between -3 and 3 i.e., three standard deviations from the mean. So I have taken threshold value as 3. And, plotting the data where Z < 3.

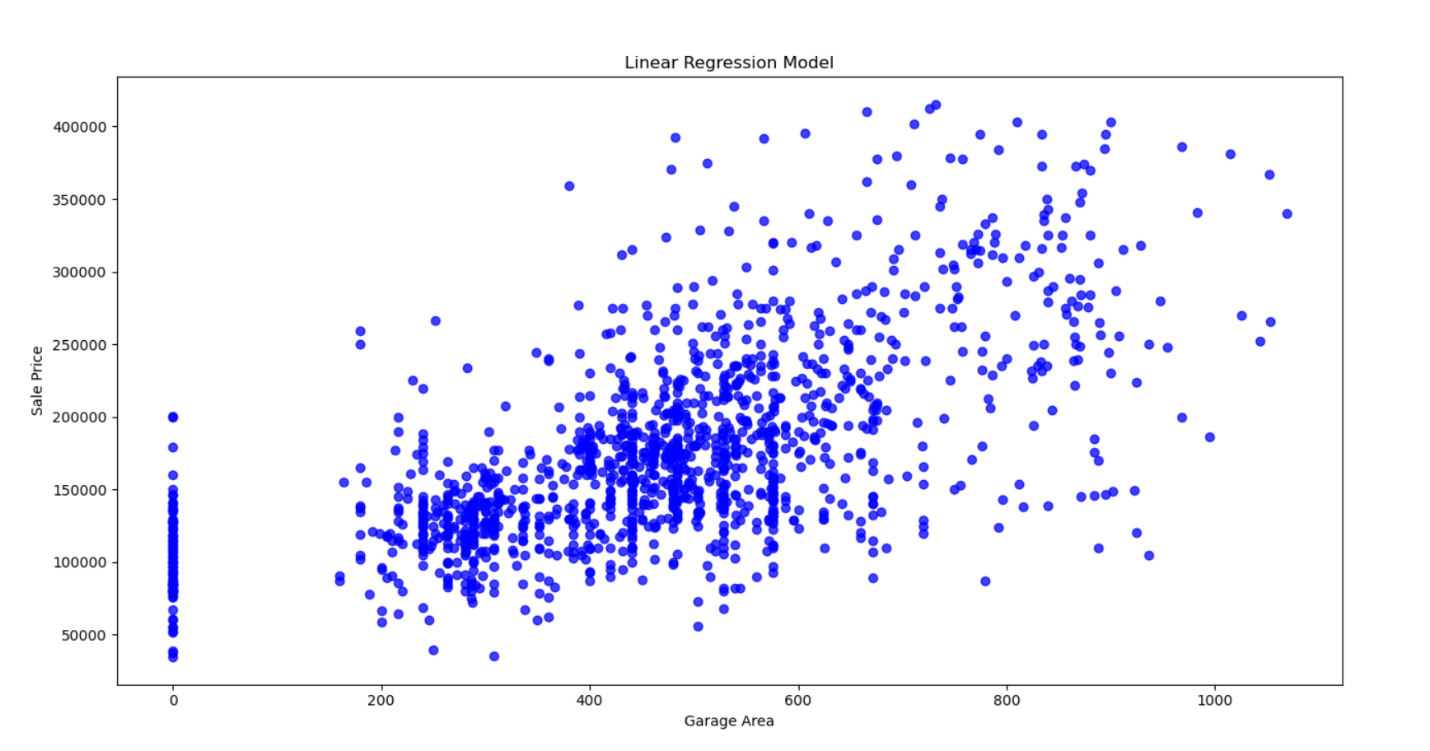
By using matplotlib.pyplot I am plotting a scatter plot with garbage area on x – axis and sales prices on y axis, and got the below outputs as shown in below screenshots.



**Output: Before removing outlier values.**



**Output : After removing outlier values.**



**Question 2) Create Multiple Regression for the “wine quality” dataset. In this data set “quality” is the target label. Evaluate the model using RMSE and R2**

**You need to delete the null values in the data set. You need to find the top 3 most correlated features to the target label(quality)**

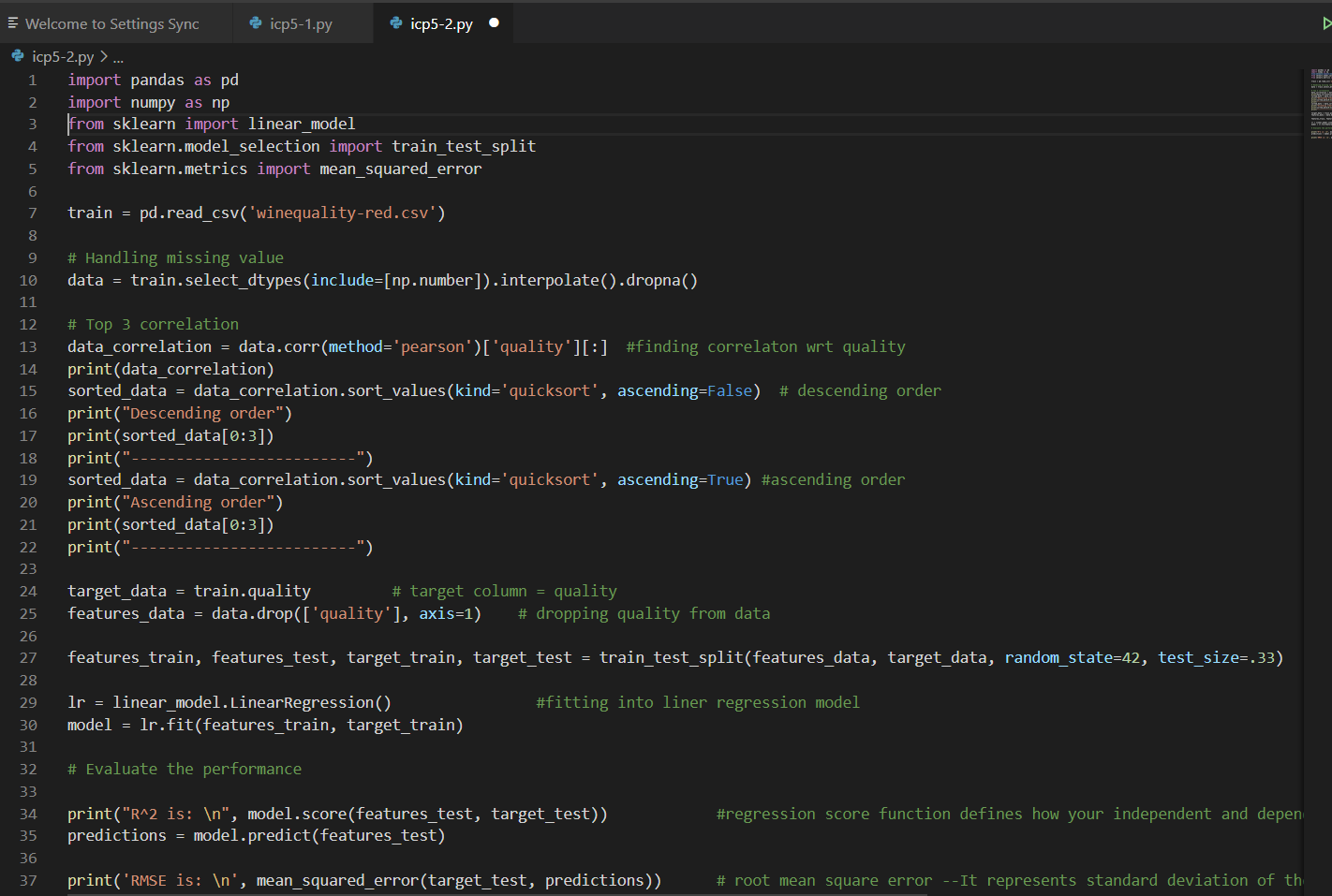
**Code:** Here I am finding the correlation of quality with respective to all column fields and printing the top 3 and least 3 of correlated values.

And, taken 67% data for training and 33% data for testing and fit into linear regression model and found out R2 and RMSE values.

R2 = regression score function defines how your independent and dependent varibales are dependent on each other

RSME = ROOT MEAN SQUARE ERROR

It represents standard deviation of the differences between predicted values and observed values



**Output:**

