

HOUSING: PRICE PREDICTION

Submitted by:

HIMANSHU SHARMA

**ACKNOWLEDGMENT**

One of the pleasant aspects of preparing a project report is the opportunity to thank those who have contributed to make this project possible.

We are extremely thankful to Mrs.Astha Mishra, whose active interest in the project & insight helped us to formulate, redefine implement our approach to the project.

We are also thankful to our institute & Other seen unseen hands which have given us direct & indirect help in completion of this project.

* **Himanshu Sharma**

**INTRODUCTION**

* **Business Problem Framing**

This project is related to house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for 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.

* **Conceptual Background of the Domain Problem**

Conceptual Background in this dataset, i.e.We will use Statistics and machine learning algorithms to solve the dataset. Since the target value is continuous , we will use the regression model algorithms in it. To visualize the data, we will use matplotlib and seaborn library, which based on python language.

* **Review of Literature**

Review of Literature is basically related to comprehensive summary of dataset as well as descriptions of relations between input variables and output variable. In given dataset following variables are described which tell we about relation between dependent variable to independent variable.

* **Motivation for the Problem Undertaken**

Motivation for the problem undertaken because project scenario basically 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.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

In this section we understand and describe the mathematical, statistical and analytics modelling done during this project along with the proper justification. First of all for better understanding about dataset and given attributes information we use **houseDF.info()** command, which tell me, The total number of attributes, what is name of attributes, datatype of attributes and how many Non-null values are present in dataset.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1460 entries, 0 to 1459

Data columns (total 81 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Id 1460 non-null int64

1 MSSubClass 1460 non-null int64

2 MSZoning 1460 non-null object

3 LotFrontage 1201 non-null float64

4 LotArea 1460 non-null int64

5 Street 1460 non-null object

6 Alley 91 non-null object

7 LotShape 1460 non-null object

8 LandContour 1460 non-null object

9 Utilities 1460 non-null object

10 LotConfig 1460 non-null object

11 LandSlope 1460 non-null object

12 Neighborhood 1460 non-null object

13 Condition1 1460 non-null object

14 Condition2 1460 non-null object

15 BldgType 1460 non-null object

16 HouseStyle 1460 non-null object

17 OverallQual 1460 non-null int64

18 OverallCond 1460 non-null int64

19 YearBuilt 1460 non-null int64

20 YearRemodAdd 1460 non-null int64

21 RoofStyle 1460 non-null object

22 RoofMatl 1460 non-null object

23 Exterior1st 1460 non-null object

24 Exterior2nd 1460 non-null object

25 MasVnrType 1452 non-null object

26 MasVnrArea 1452 non-null float64

27 ExterQual 1460 non-null object

28 ExterCond 1460 non-null object

29 Foundation 1460 non-null object

30 BsmtQual 1423 non-null object

31 BsmtCond 1423 non-null object

32 BsmtExposure 1422 non-null object

33 BsmtFinType1 1423 non-null object

34 BsmtFinSF1 1460 non-null int64

35 BsmtFinType2 1422 non-null object

36 BsmtFinSF2 1460 non-null int64

37 BsmtUnfSF 1460 non-null int64

38 TotalBsmtSF 1460 non-null int64

39 Heating 1460 non-null object

40 HeatingQC 1460 non-null object

41 CentralAir 1460 non-null object

42 Electrical 1459 non-null object

43 1stFlrSF 1460 non-null int64

44 2ndFlrSF 1460 non-null int64

45 LowQualFinSF 1460 non-null int64

46 GrLivArea 1460 non-null int64

47 BsmtFullBath 1460 non-null int64

48 BsmtHalfBath 1460 non-null int64

49 FullBath 1460 non-null int64

50 HalfBath 1460 non-null int64

51 BedroomAbvGr 1460 non-null int64

52 KitchenAbvGr 1460 non-null int64

53 KitchenQual 1460 non-null object

54 TotRmsAbvGrd 1460 non-null int64

55 Functional 1460 non-null object

56 Fireplaces 1460 non-null int64

57 FireplaceQu 770 non-null object

58 GarageType 1379 non-null object

59 GarageYrBlt 1379 non-null float64

60 GarageFinish 1379 non-null object

61 GarageCars 1460 non-null int64

62 GarageArea 1460 non-null int64

63 GarageQual 1379 non-null object

64 GarageCond 1379 non-null object

65 PavedDrive 1460 non-null object

66 WoodDeckSF 1460 non-null int64

67 OpenPorchSF 1460 non-null int64

68 EnclosedPorch 1460 non-null int64

69 3SsnPorch 1460 non-null int64

70 ScreenPorch 1460 non-null int64

71 PoolArea 1460 non-null int64

72 PoolQC 7 non-null object

73 Fence 281 non-null object

74 MiscFeature 54 non-null object

75 MiscVal 1460 non-null int64

76 MoSold 1460 non-null int64

77 YrSold 1460 non-null int64

78 SaleType 1460 non-null object

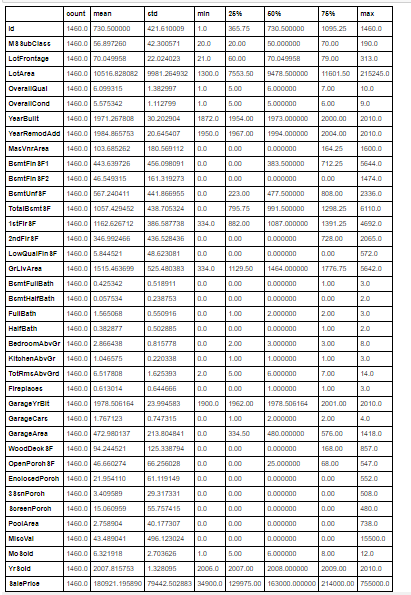
79 SaleCondition 1460 non-null object

80 SalePrice 1460 non-null int64

dtypes: float64(3), int64(35), object(43)

memory usage: 924.0+ KB

After understanding the dataset values, we take the statistical descriptions of dataset using **houseDF.describe().T** command in python, which tells the following statistical descriptions:



After know about statistical description we move forward in the way of finding co-relation of variables, then move to data-cleaning, then move to visualized the data set and their relationship. After this we move to data pre-processing and modelling as well as testing the accuracy of model.

* **Data Sources and their formats**

Machine learning algorithms are almost always optimized for raw, detailed source data. Thus, the data environment must provision large quantities of raw data for discovery-oriented analytics practices such as data exploration, data mining, statistics, and machine learning.

Tabular data for machine learning is typically found is .csv files. Csv files are text-based files containing comma separated values (csv). Csv files are popular for ML as they are easy to view/debug and easy to read/write from programs (no compression/indexing).

* **Data Preprocessing Done**

Data pre-processing in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In simple words, data pre-processing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

**Steps in Data Pre-processing in Machine Learning:**

* **Acquire the dataset**

To build and develop Machine Learning models, we must first acquire the relevant dataset. We use download data using given link through the company.

### Import all the crucial libraries

### The core Python libraries used for this data pre-processing in Machine Learning are:

### 

* **Import the dataset**

In this step, you need to import the dataset/s that you have gathered for the ML project at hand.



* **Identifying and handling the missing values**

In data pre-processing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data.

In this dataset, there is no null value, but some faulty data is present.

* **Encoding the categorical data**

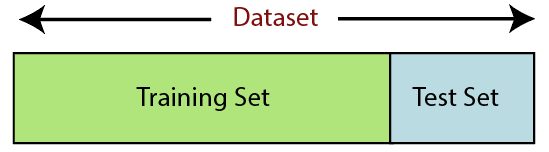
Machine Learning models are primarily based on mathematical equations. Thus, you can intuitively understand that keeping the categorical data in the equation will cause certain issues since you would only need numbers in the equations.

you can use the LabelEncoder() class from the sci-kit learn library. The code will be as follows –

****

* **Splitting the dataset**

Every dataset for Machine Learning model must be split into two separate sets – training set and test set.

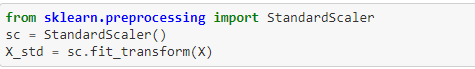


In the same way, we using below code:

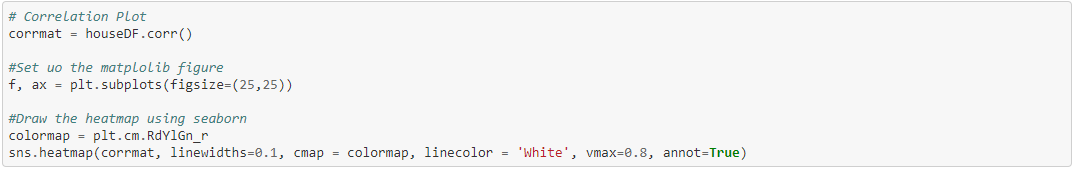


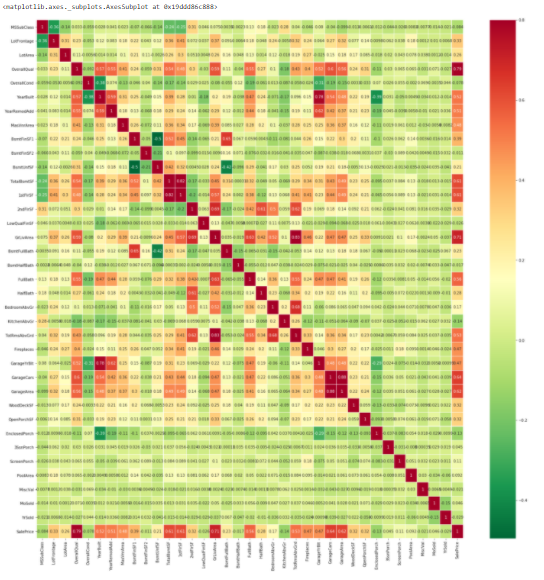
* **Feature scaling**

Feature scaling marks the end of the data pre-processing in Machine Learning. It is a method to standardize the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds.



* Data Inputs- Logic- Output Relationships

****

****

* **Hardware and Software Requirements and Tools Used**

In this project dataset is too large for processing or modelling, that’s why we use good hardware configuration like as above 4GB RAM, above or equal core i3 processor and also need good storage HDD. In way of software, we use any operating system which support python language for coding.

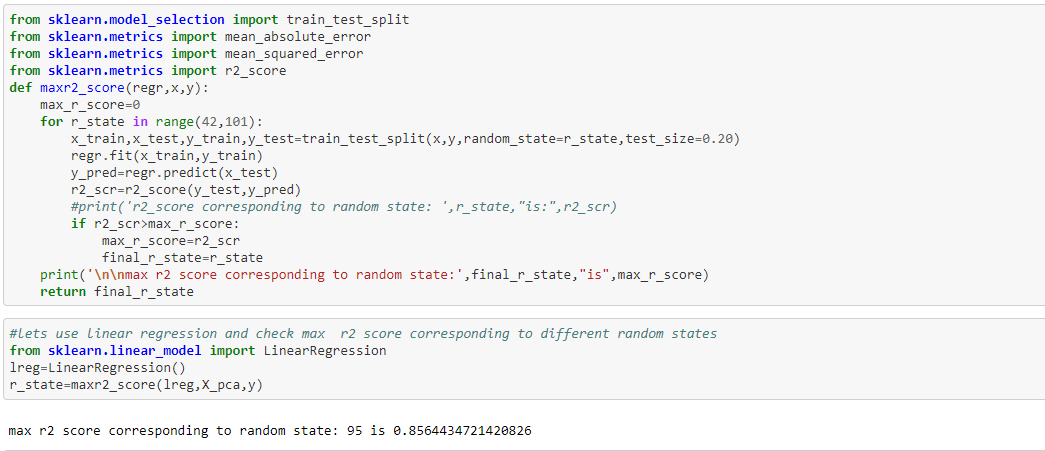
**Model/s Development and Evaluation**

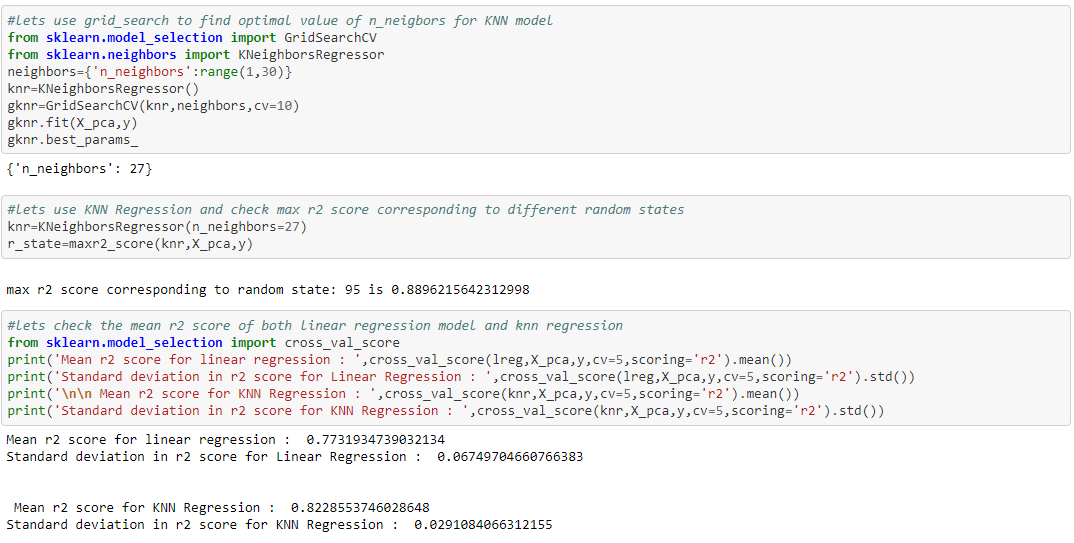
* **Identification of possible problem-solving approaches (methods)**

In this dataset we build the model will be a good way for the

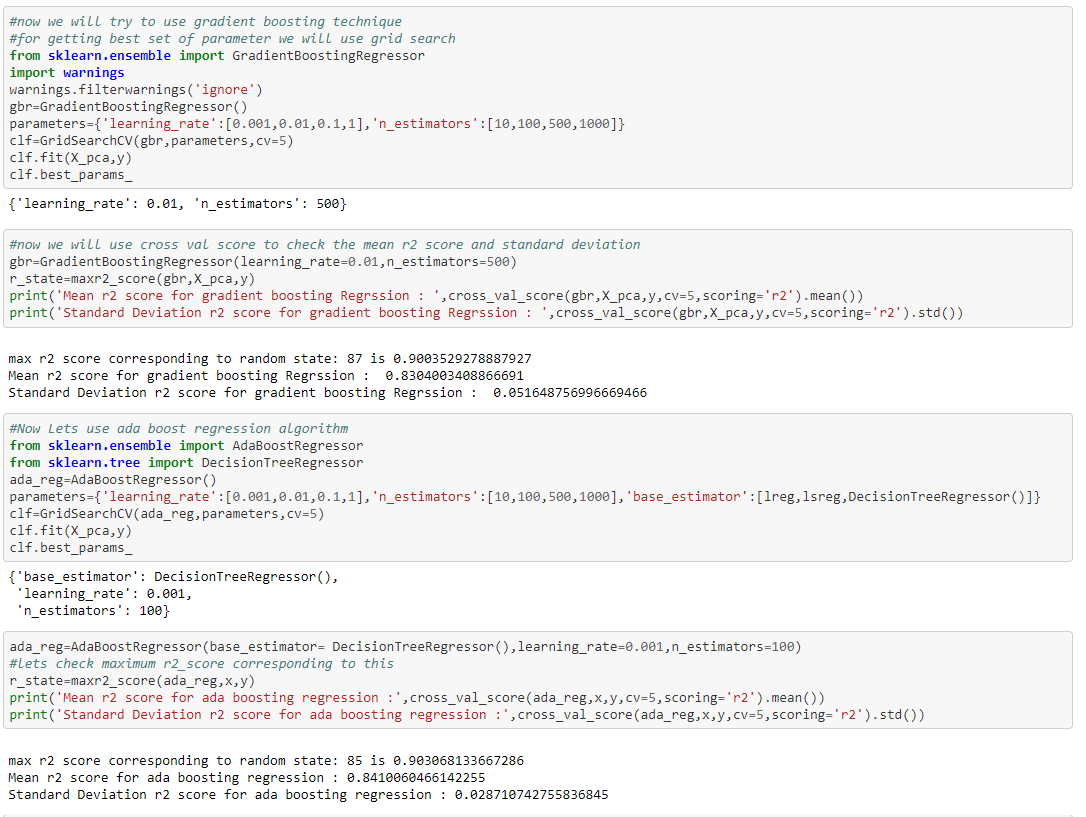
management to understand the pricing dynamics of a new market.

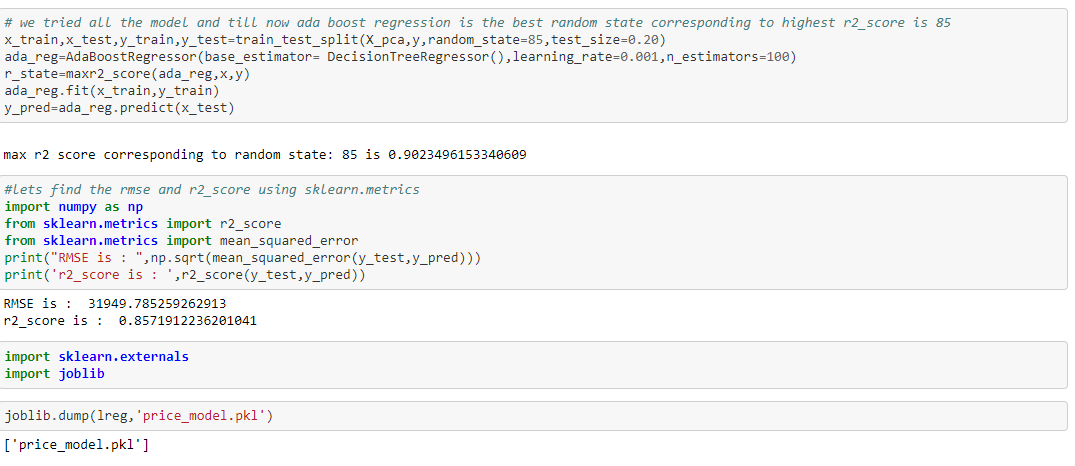
* **Testing of Identified Approaches (Algorithms)**

****

****

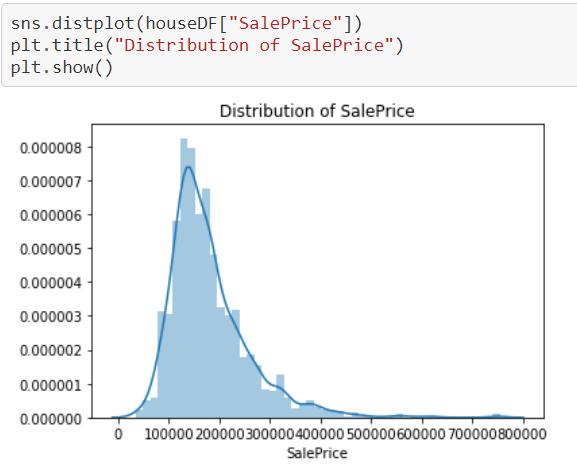
****

****

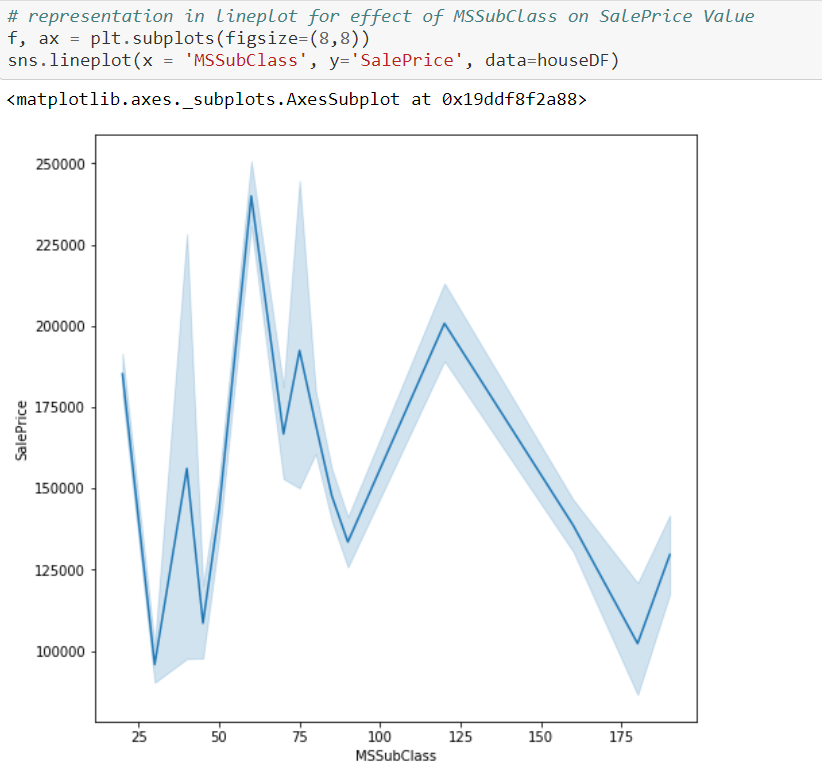
****

* **Visualizations**

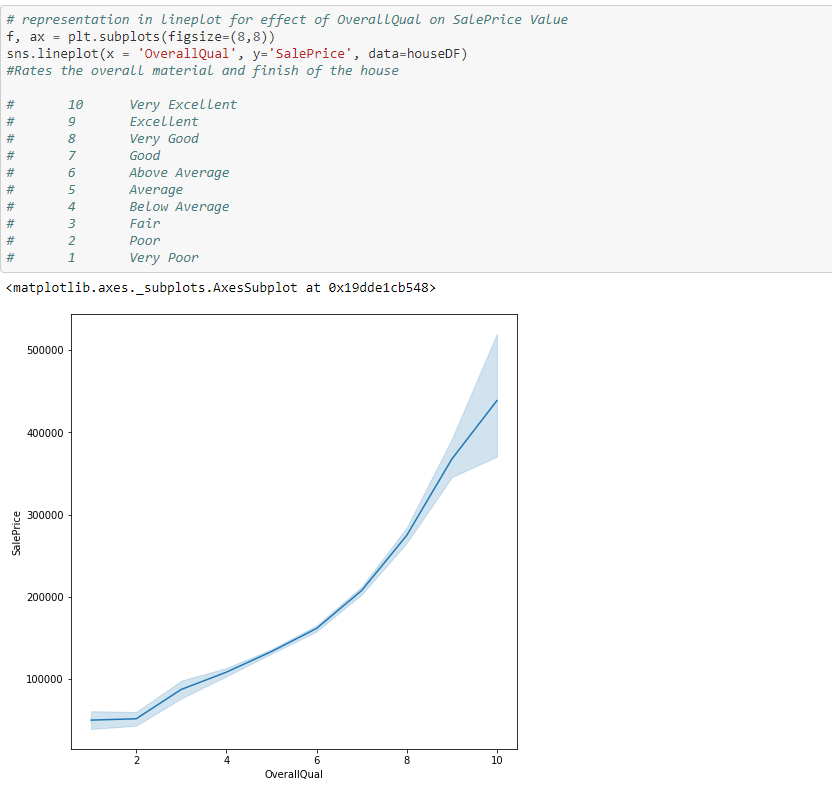
1. **Visualize the Distribution of SalePrice**

****

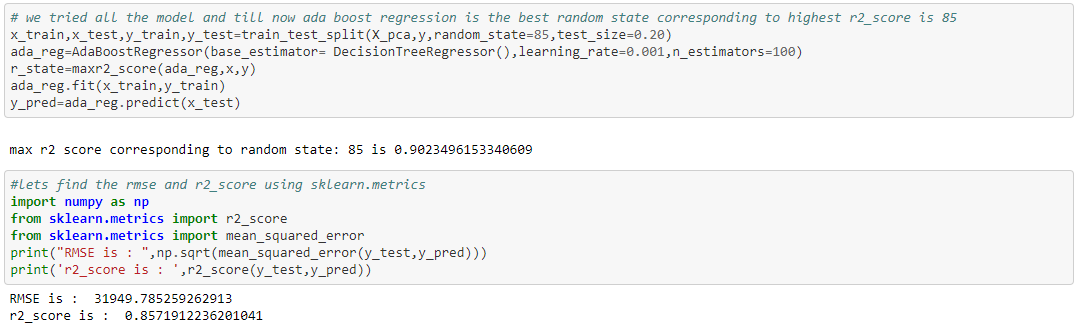
1. **Visualize effect of MSSubClass on SalePrice**

****

1. **Visualize the Effect of OverallQual on SalePrice**

****

* **Interpretation of the Results**

****

**CONCLUSION**

After analysing data, visualization and modelling, we come to the conclusion that using the ada boost regression is the best random state corresponding to highest r2\_score is 85.