

**HOUSING: PRICE PREDICTION**

**Submitted by:**

Ankit Soran

ACKNOWLEDGEMENT

The internship opportunity, I have with FlipRobo Technologies is a great chance for learning and professional development. I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills acknowledge in the best possible way.

I would like to extend my appreciation and thanks for the mentors from Data Trained and professionals from FlipRobo Technologies who had extended their help and support.

**References:**

<https://studygyaan.com/data-science-ml/linear-regression-machine-learning-project-for-house-price-prediction>

<https://medium.com/codex/house-price-prediction-with-machine-learning-in-python-cf9df744f7ff>

<https://thecleverprogrammer.com/2020/12/29/house-price-prediction-with-python/>

<https://www.analyticsvidhya.com/blog/category/machine-learning/>

****

**Problem Statement:**

Houses are one of the necessary need 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. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in 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. 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 in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

**Business Goal:**

You 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.

**Technical Requirements:**

• Data contains 1460 entries each having 81 variables.

• Data contains Null values. You need to treat them using the domain knowledge and your own understanding.

• Extensive EDA has to be performed to gain relationships of important variable and price.

• Data contains numerical as well as categorical variable. You need to handle them accordingly.

• You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.

• You need to find important features which affect the price positively or negatively.

• Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file.

The “Data file.csv” and “Data description.txt” are enclosed with this file

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Data Sources and their formats
* Data Pre-processing Done
* Data Inputs- Logic- Output Relationships
* State the set of assumptions (if any) related to the problem under consideration

1. Which variables should we dummify?
2. How will we handle missing values? Do we drop the row, impute using the mean, or use kNN?
3. Do any of the features look skewed? Should we transform them using Box-Cox, log, sqrt?
4. Quick correlation plots. Variables with high correlation to prices should be prioritized as predictors
5. How to deal with outliers?
6. How to deal with multi-collinearity?

* Hardware and Software Requirements and Tools Used

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration
* Visualizations
* Interpretation of the Results

**Analytical Problem Framing:**

House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. While there are people who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood.

According to [1], housing markets have a positive impact on a country's currency, which is an important national economy scale. Homeowners will purchase goods such as furniture and household equipment for their home, and homebuilders or contractors will purchase raw material to build houses to satisfy house demand, which is an indication of the economic wave effect created by the new house supply. Besides that, consumers have capital to make a large investment, and the construction industry is in good condition can be seen through a country's high level of house supply.

According to [2], numerous international organizations and human rights have emphasized house importance. House is profoundly rooted in the economic, financial, and political structure of each country.

Nevertheless, [3] reported that the fluctuation of house prices has always been an issue for house owners, buildings and real estate, besides [4] stated that house has become unaffordable as there is substantial price growth in several countries in the housing sector. Residents' quality of life as well as national economy depends on the potential house price increase. Ultimately, this issue will affect investors who are making their house as an investment. An increase in house demand occurs each year, indirectly causing house price increases every year.

The problem arises when there are numerous variables such as location and property demand that may influence the house price, thus most stakeholders including buyers and developers, house builders and the real estate industry would like to know the exact attributes or the accurate factors influencing the house price to help investors make decisions and help house builders set the house price. House price prediction can be done by using a multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, and more. There are many benefits that home buyers, property investors, and house builders can reap from the house-price model. This model will provide a lot of information and knowledge to home buyers, property investors and house builders, such as the valuation of house prices in the present market, which will help them determine house prices. Meanwhile, this model can help potential buyers decide the characteristics of a house they want according to their budget [5]. Previous studies focused on analysing the attributes that affect house price and predicting house price based on the model of machine learning separately. However, this article combines such a both predicting house price and attributes together.

MATHEMATICAL MODELLING OF PROBLEM:

**Mathematical modelling is simply the method of implementing statistical analysis to a dataset where a Statistical Model is a mathematical representation of observed data.**

While analysing the data, there are an array of statistical models we can choose to utilize.

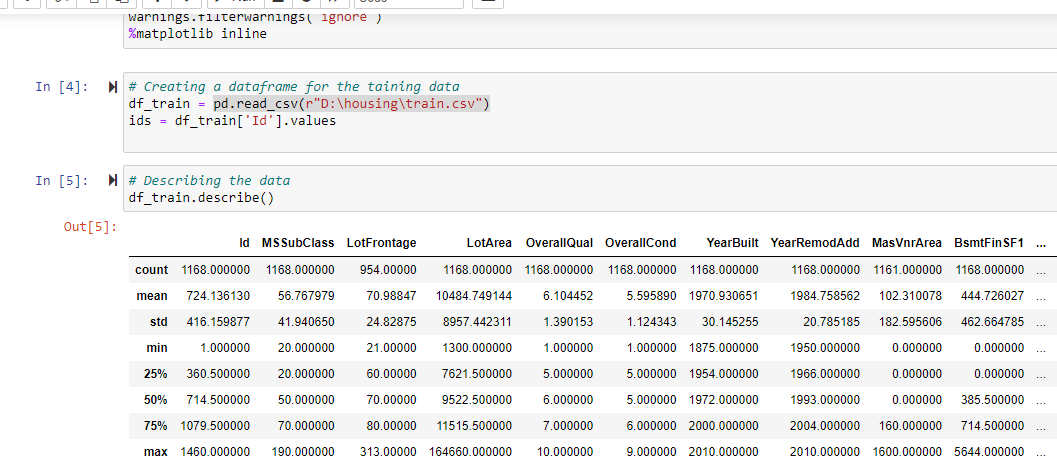
For the given project, we need to predict house price.

This is a classification problem. There are wide varieties of classification models like Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor etc.

**DATA SOURCE AND FORMAT:**

The data has been provided by client in a comma separated Values (.csv) format.

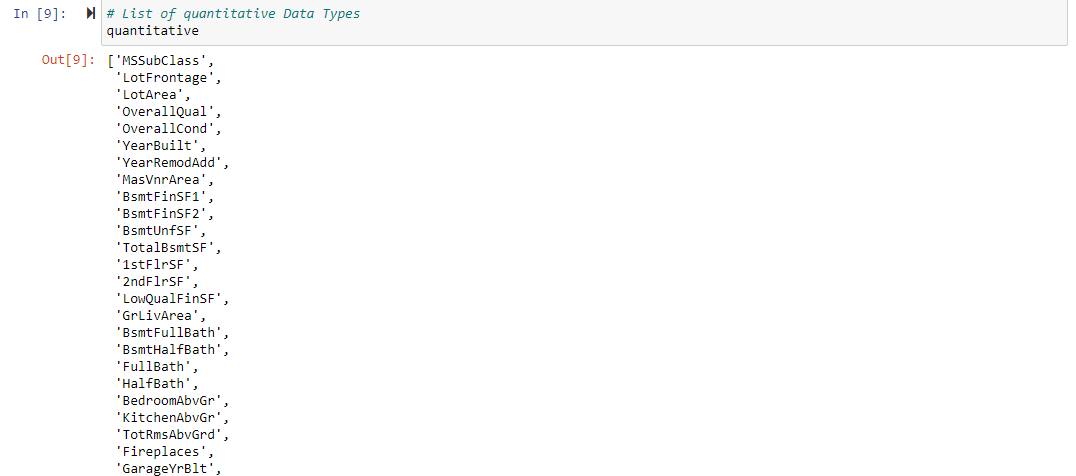
The data will be loaded into Pandas Data frame.



After doing basic exploratory data analysis (EDA) with the above Data Frame like:

* Checking Info
* Checking shape (1168, 81)
* Identifying features.
* Removing unwanted columns.
* Checking Quantitative features
* Checking qualitative features



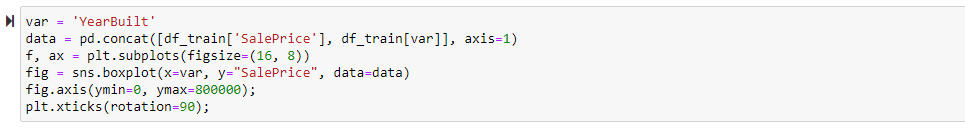


**Moving forward with Data frame we observed that there is a huge no of missing data in data frame.**

**Data Inputs- Logic- Output Relationships:**

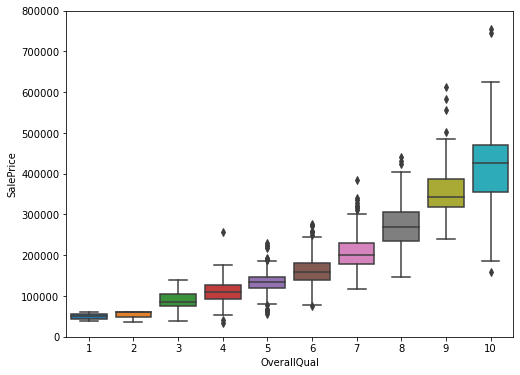
****

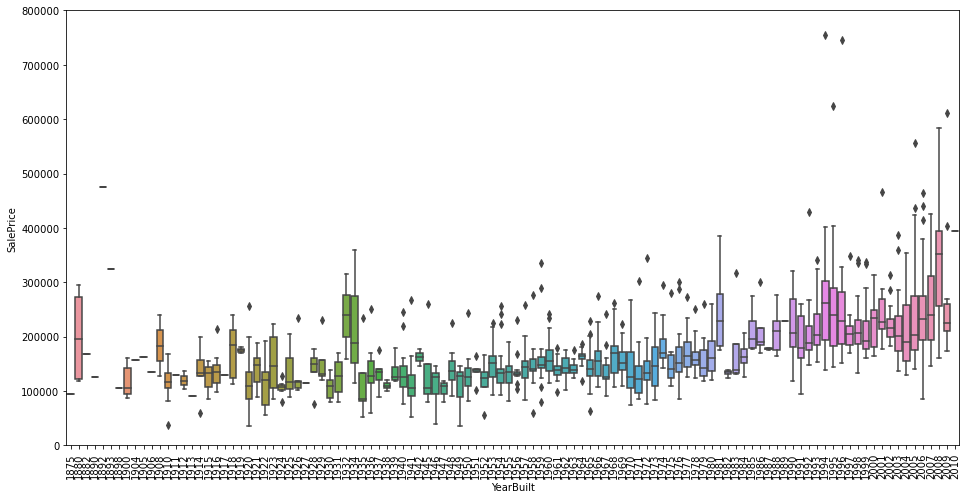
****

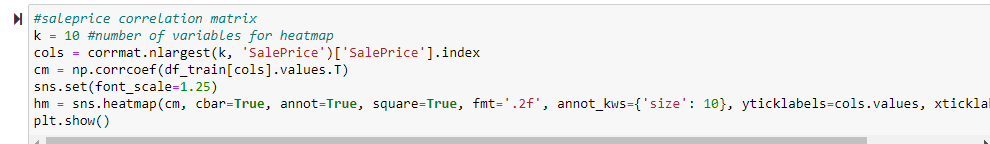
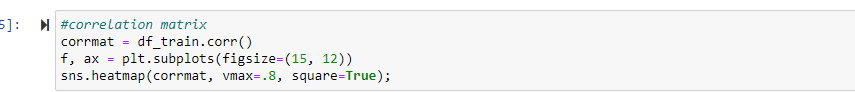
****

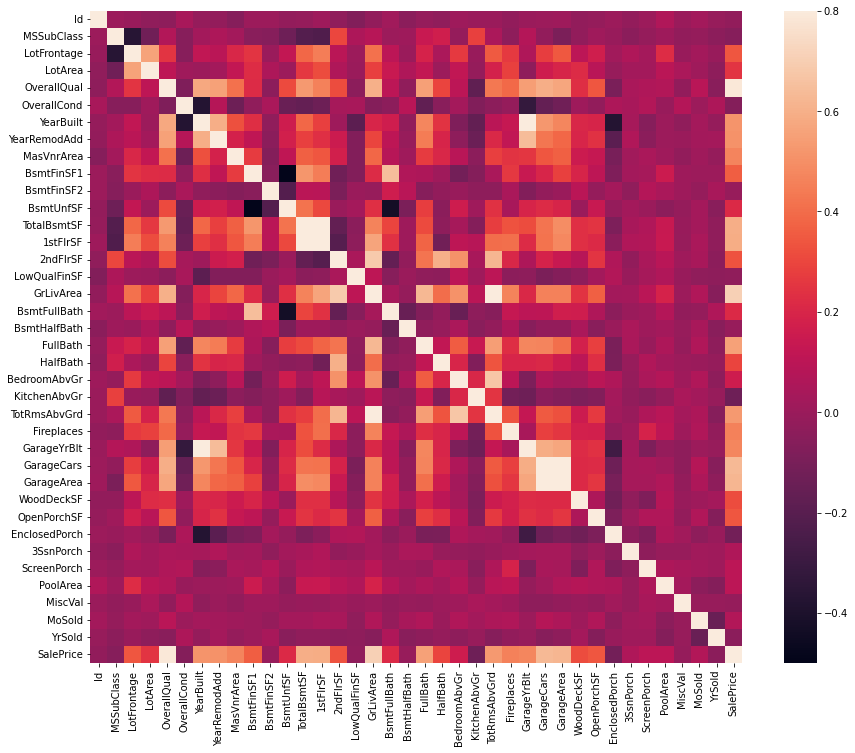
GrLivArea' and 'TotalBsmtSF' seem to be linearly related with 'Sale Price'

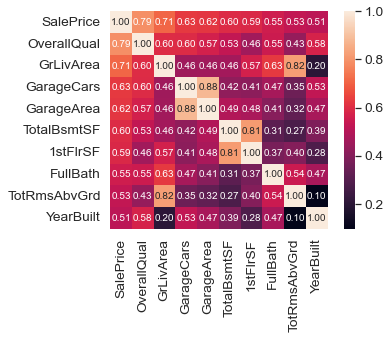
'OverallQual' and 'Year Built' also seem to be related with 'Sale Price'. The relationship seems to be stronger in the case of 'OverallQual', where the box plot shows how sales prices increase with the overall quality.









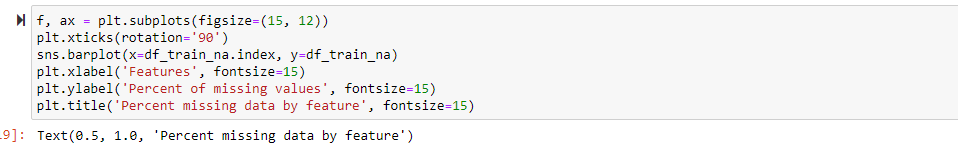
****

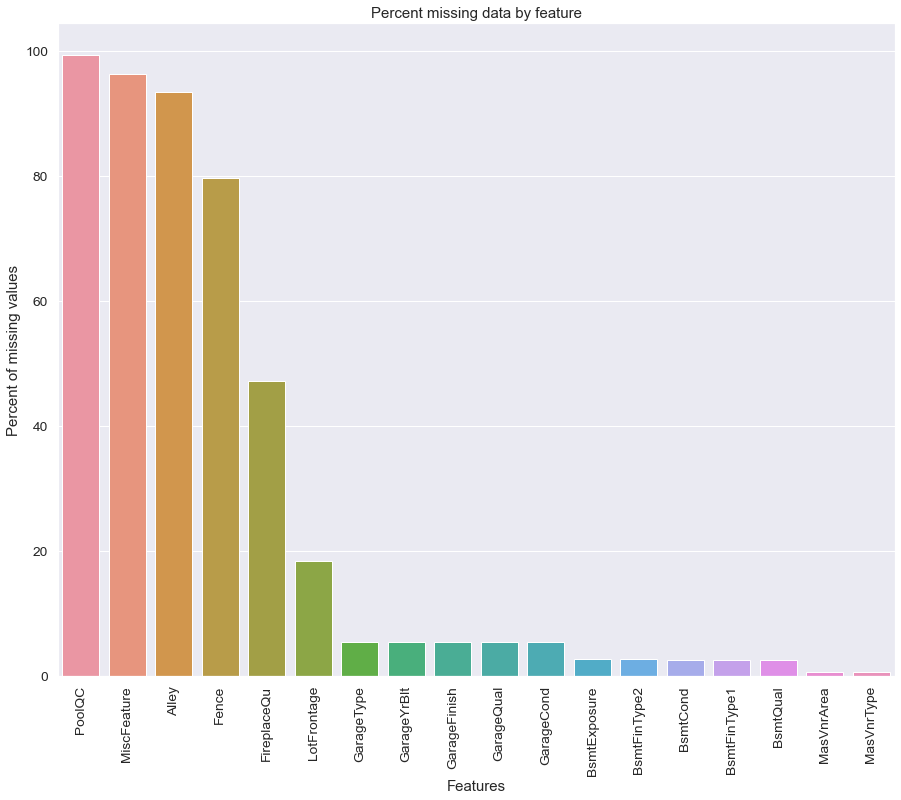
* 'OverallQual', 'GrLivArea' and 'TotalBsmtSF' are strongly correlated with 'Sale Price'
* 'Garage Cars' and 'Garage Area' are also some of the most strongly correlated variables
* 'TotalBsmtSF' and '1stFloor' also seem to be twin brothers.
* 'TotRmsAbvGrd' and 'GrLivArea' are twin brothers again
* 'Year Built' is slightly correlated with 'Sale Price'

**Features Highly Correlated with sale price:**

****



****

****



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

**Hardware required**: -

* 1. Processor — core i5 and above
* 2. RAM — 8 GB or above
* 3. SSD — 250GB or above

**Software/s required**: -

* 1.Anaconda

**Model/s Development and Evaluation:**

Linear regression:

Simple linear regression statistical method allows us to summarize and study the relationship between two continuous quantities variables.

• One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.

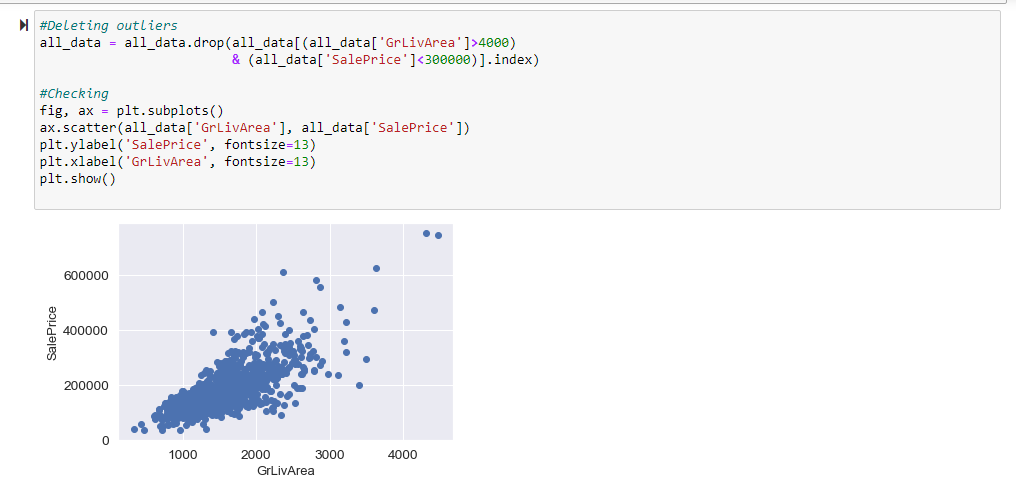
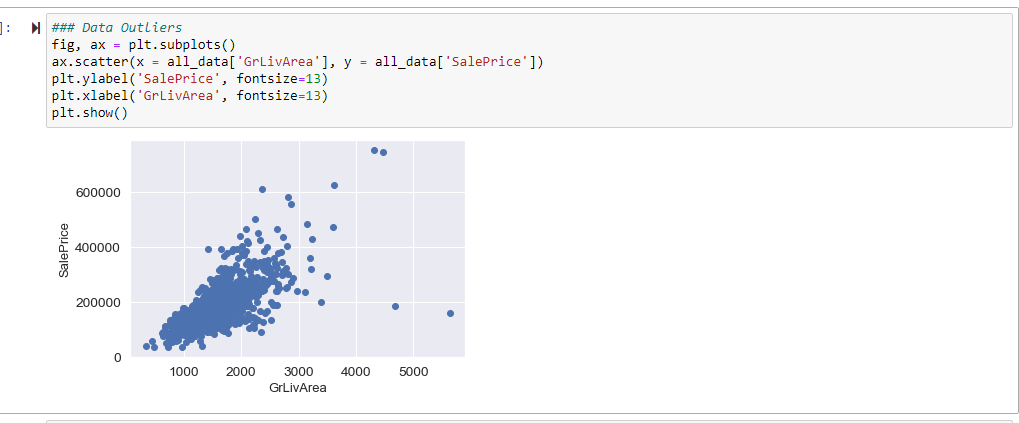
• The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

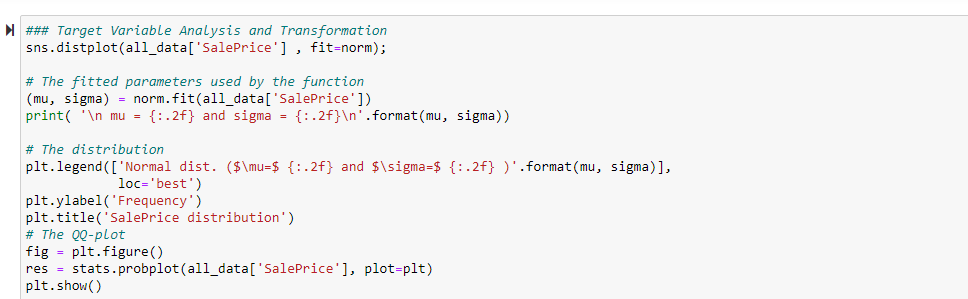
Multiple Regression:

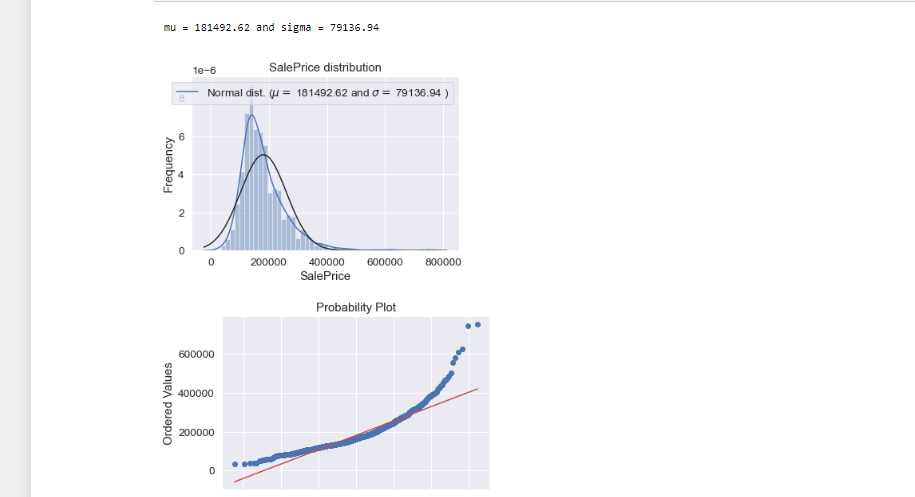
Analysis multiple regression analysis is used to check whether there is a statistically noteworthy association the middle of sets of variables.

****

Importing data & checking for outliers

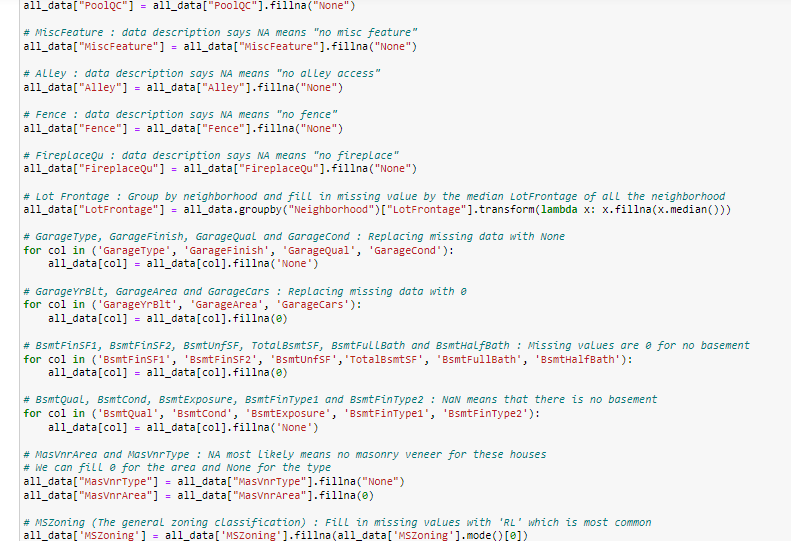


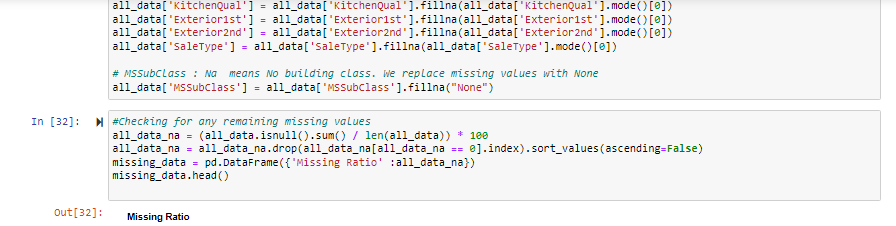


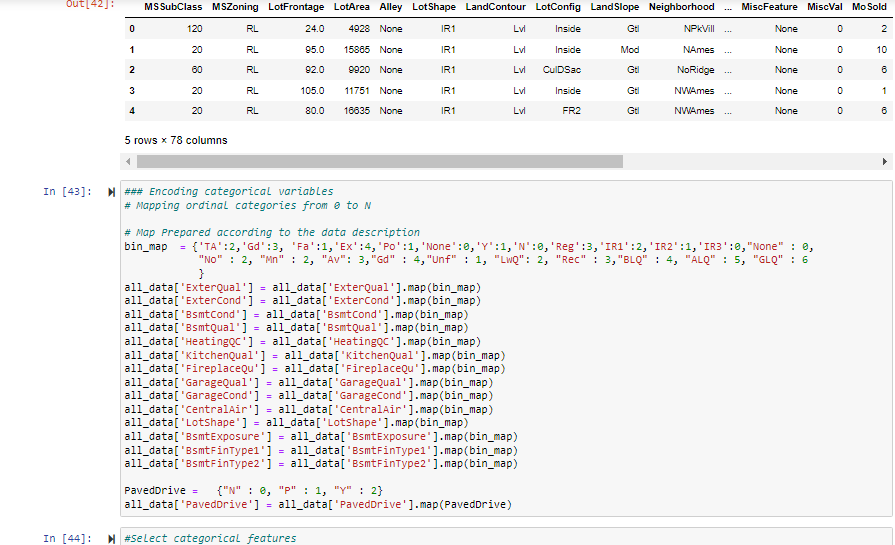


The target variable is right skewed. For linear models it’s better to have normally distributed data, so we to transform this variable and make it more normally distributed.

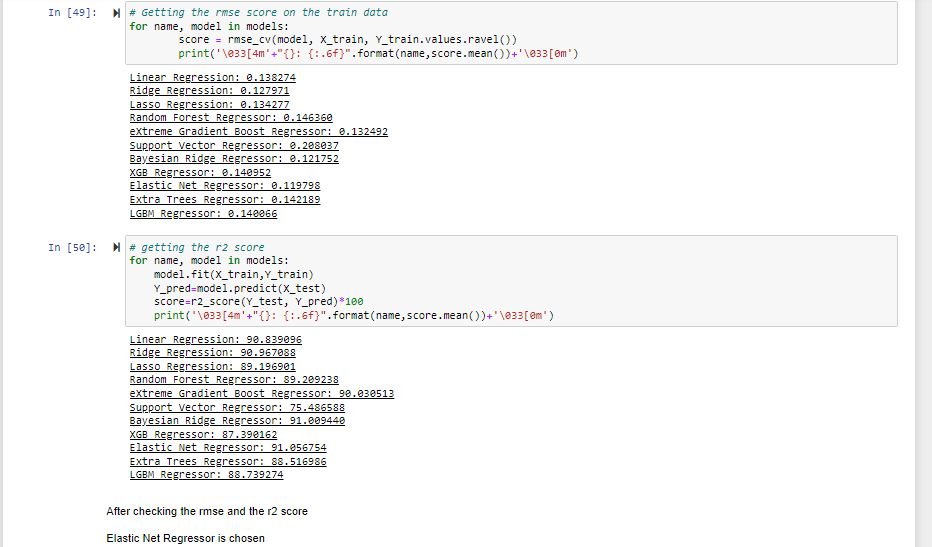
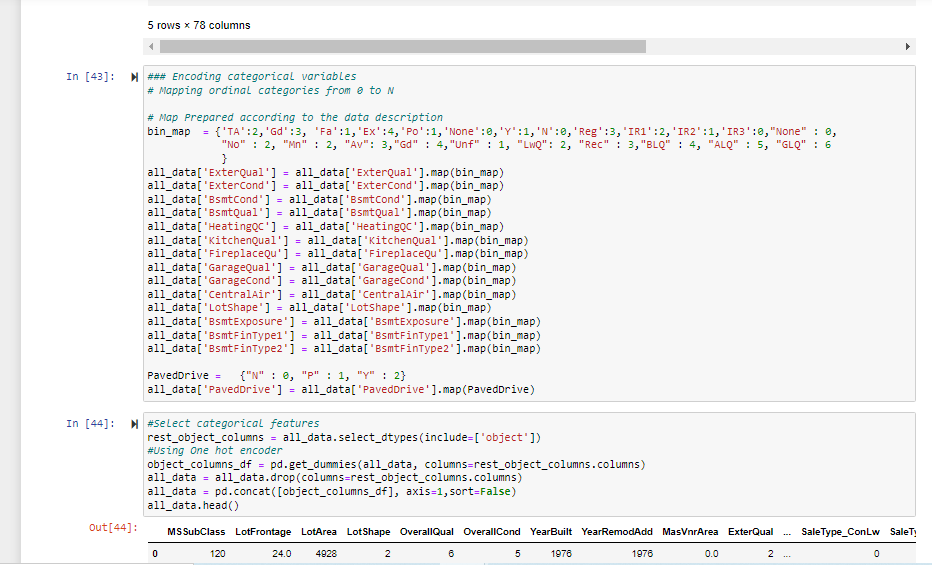
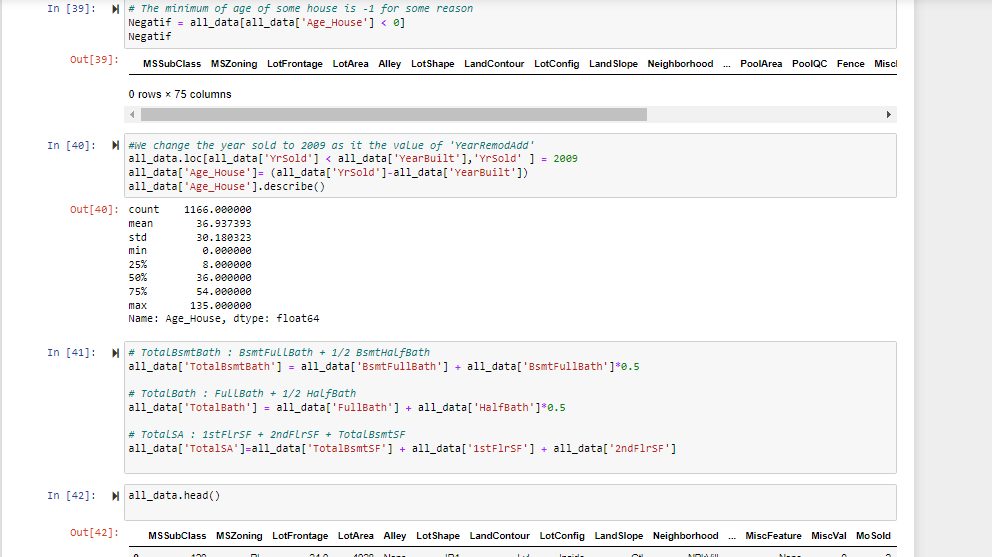
**Treating missing values:**

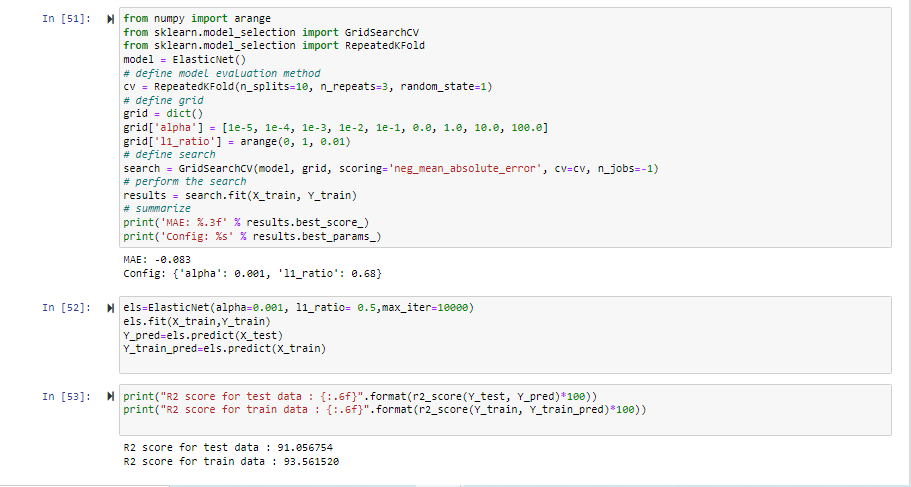
****











The Model does not seems to be over fitted or under fitted

The accuracy score for the regression model is 91%

Interpretation of the Results:

We have managed out how to prepare a model that gives users for a novel best approach with take a gander at future lodging value predictions. A few relapse strategies have been investigated Furthermore compared, when arriving during a prediction strategy In light of regression support. Straight former imply works bring been utilized within our model, something like that that future value predictions will have a tendency towards All the more sensible values. We concocted an approach with use similarly as considerably information as time permits for our prediction system, by adopting those ideas from claiming gradient boosting. In spite of Hosting generated all the attempting provision that met our introductory requirements, there are Different upgrades that could be produced later on. These incorporate upgrades we didn't settle on because of constrained duration of the time. A real worry for the prediction framework may be the stacking period. Moreover, our data set takes more than one day should prepare. As opposed performing the computations sequentially, we might utilize various processors and parallel the computations involved, which might possibly decrease the preparation time Furthermore prediction period. Include All the more functionalities under the model, we can give choices for client with select a district alternately locale should produce those high temperature maps, as opposed to entering in the list.

**Thank you**