

HOUSE PRICE PREDICTION PROJECT

Submitted by:

MOHAMMED ZAHED N

**ACKNOWLEDGMENT**

Below are some of the sources from which code snippets have been helpful during the project completion

**References:**

1. <https://www.scikitlearn.org>

2. <https://www.askpython.com>

3. <https://www.stackoverflow.com>

4. <https://www.geeksforgeeks.org>

**INTRODUCTION**

House Price:

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. 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.

* Motivation for the Problem Undertaken

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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

Many Statistical models were used also mathematical models necessary, Some of the used models are listed below.

Statistical Model:

1. Logarithmic Transformation
2. Random Forest Regressor
3. Decision Tree Regressor
4. Support Vector Regressor
5. K Neighbors Regressor
6. Gradient Boosting Regressor

* Data Sources and their formats

Most of the Code snippets and their formats are taken from below site

<https://www.geeksforgeeks.org>

* Data Pre-processing Done

1. Checking Dimensionality

2. Checking Missing Values

3. Cleaning Data

4. Renaming columns for better understanding

5. Dropping insignificant features with less meaningful data.

* Hardware and Software Requirements and Tools Used

List of hardware and software requirements along with the tools, libraries and packages used.

**Softwares:**

1. Ms Word: Ms Word for documentation purpose

2. Ms Excel: To view Dataset and to perform basic subtraction on columns for comparison

3. Ms PowerPoint: To display detailed analysis of the project and its screenshots for a presentation.

4. Jupiter Notebook (Anaconda): To run python code for building suitable model for prediction.

**Hardware:**

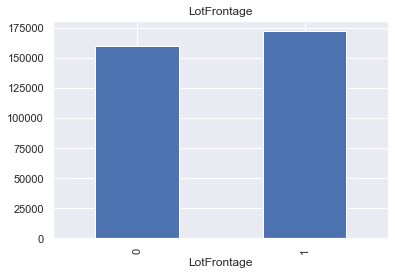
OS: Windows 10

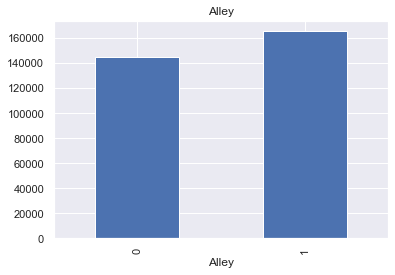
RAM: 4GB

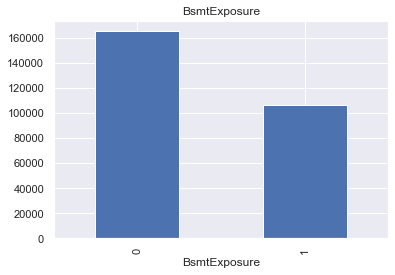
HDD: 1TB

* Visualizations

1. Code Snippet to relation between the missing values and the dependent variable in the dataset.

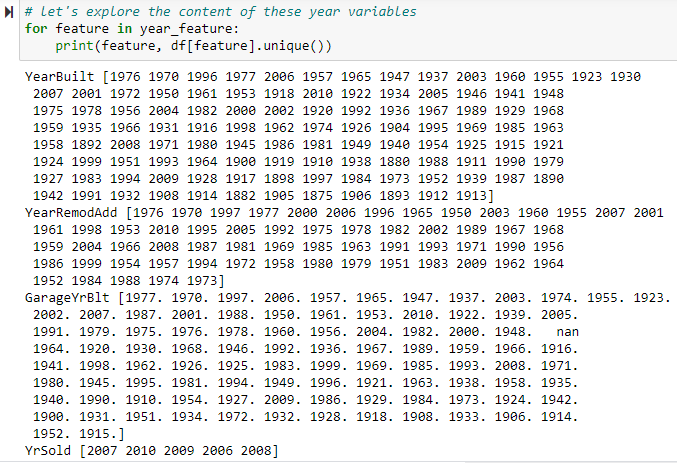




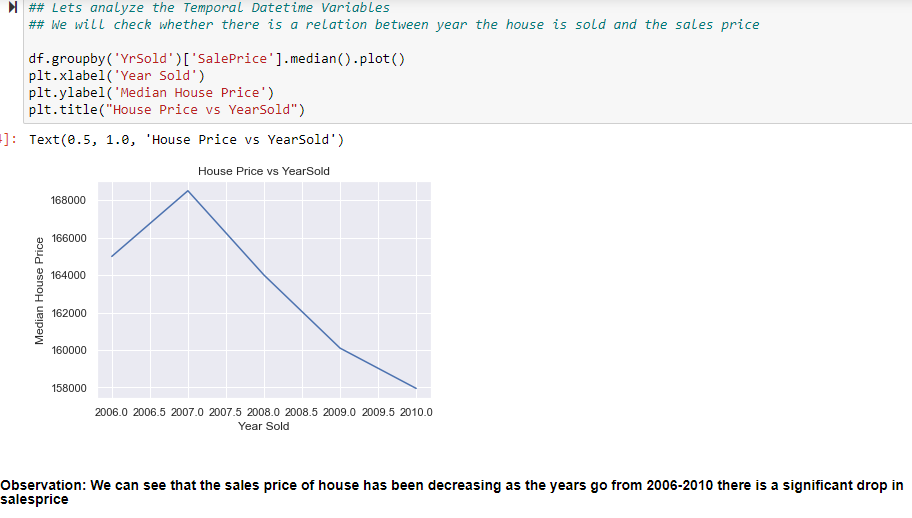


Observation: These above plot shows that there is some relation & We need to replace these nan values with something meaningful, because here With the relation between the missing values and the dependent variable is clearly visible

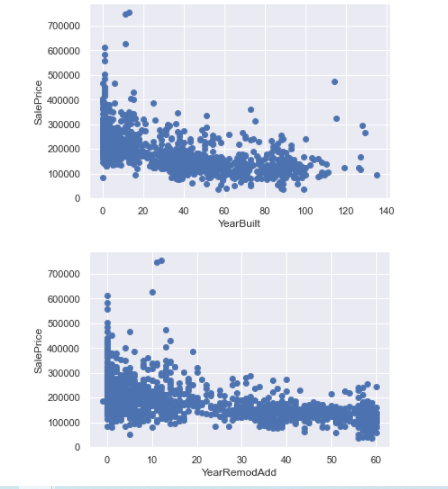
# list of variables that contain year information



1. This is a visualisation for the sales price columns to see how the price is changed over the years.

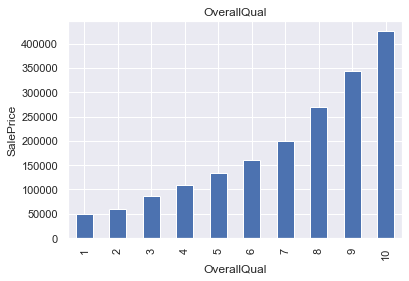


1. Scatter plot: Visualisation to see how the sale price data is distributed among the different year data.



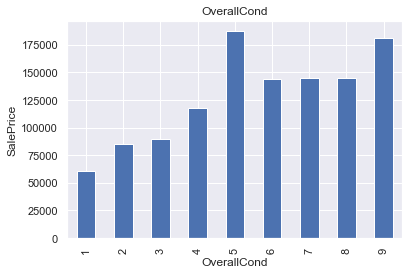
Observation: We can see that recently built modified houses have high price whereas the old ones are of low price.

1. Barplot: Visualisation to see how the Sales price is related to one of the discrete data Overall quality.

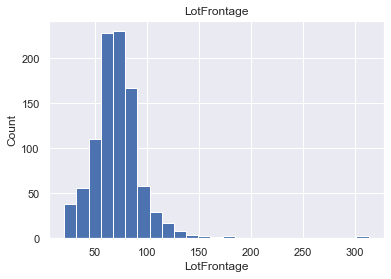


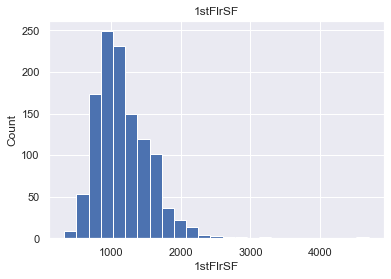
Observation: We can see that Sales Price of a house is Directly Proportional to the Overall Quality as the overall quality increases the price of the house increases.

Barplot: Visualisation to see how the Sales price is related to one of the discrete data Overall Condition.

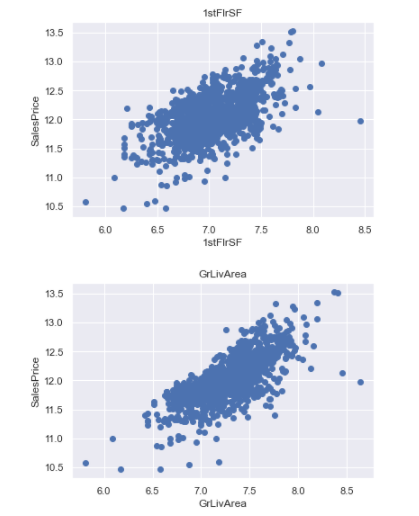


Some of the visualisations of the histogram for various columns consisting of skewed data in it.



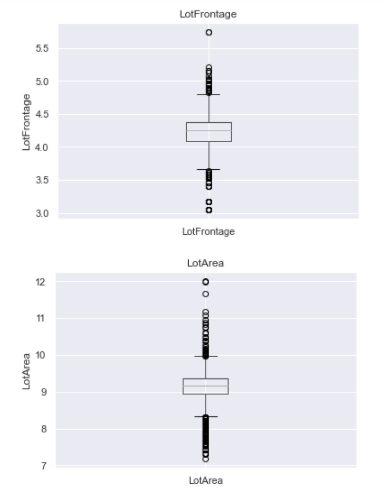


Scatter Plot: Visualisation to see how the Sales price is related to one of the numerical data 1stFirSF & GrLiveArea.



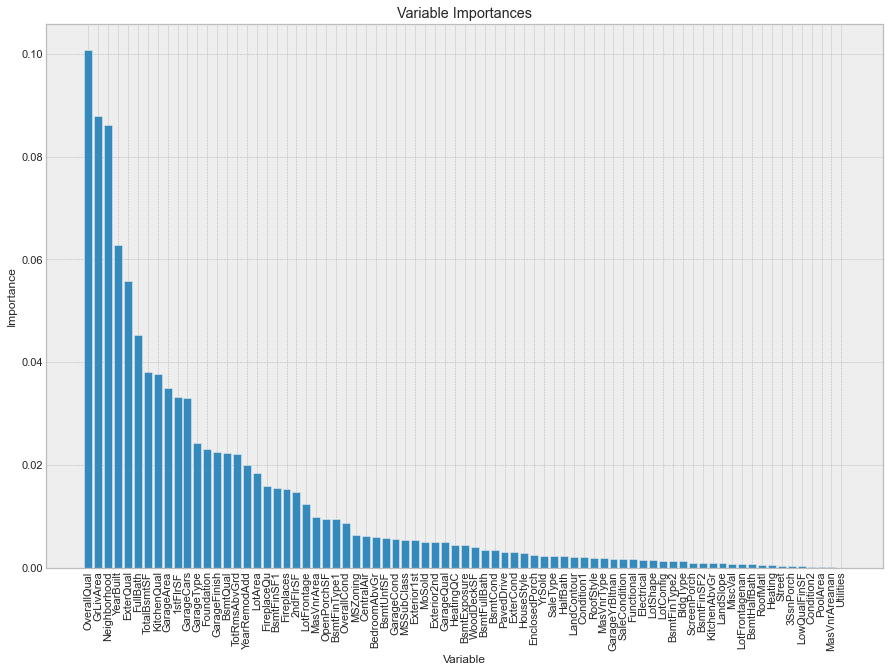
Observation: We can see that data sales price is normally distributed.

Boxplot: To see if there are any outlier in the dataset we have plotted the below



Observation: We can see outliers presence which is why the histogram data was skewed.

Feature Importance: This Below figure shows the Feature importance of how much value does each feature contribute to predict Sales Price.



Observation: A future aim may be to cut the less relevant features (lets say we can drop some columns in terms of importance), estimate a new model and compare it with the old ones.

I reckon it would lose predictive power, but on the other hand it would improve in terms of training speed.

In conclusion, these are my final considerations on the model:

\* The best feature to reliably predict the House price is Type of OverallQual. After OverallQual there is a group of 10 features including: 1 0.088005 GrLivArea

\* 2 Neighborhood 0.086088

\* 3 YearBuilt 0.062831

\* 4 ExterQual 0.055818

\* 5 FullBath 0.045347

\* 6 TotalBsmtSF 0.037992

\* 7 KitchenQual 0.037619

\* 8 GarageArea 0.034932

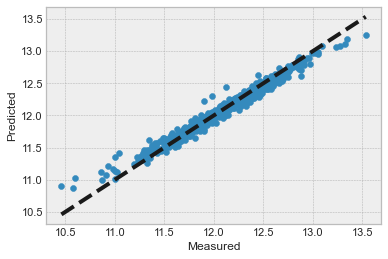
\* 9 1stFlrSF 0.033212

\* 10 GarageCars 0.033007

\* 11 GarageType 0.024202

\* Utilities is the least relevant feature in this cluster.

**CONCLUSION**

****

* Key Findings and Conclusions of the Study

We can see that the actual and predicted house price are aligned with each other Normally distributed which means the model is accurate and predicted values are considerable.

**Final Conclusion:**

Columns like **OverallQual, GrLivArea, Neighborhood, YearBuilt, ExterQual, FullBath, TotalBsmtSF, KitchenQual, GarageArea, 1stFlrSF, GarageCars & GarageType** have effect on House Price.

The Actual vs Predicted plot clearly indicates the predicted values are almost linear hence performance of model is considerably Good

Random Forest Regressor model predicts the House price more accurately than Decision Tree.