

House Price Prediction

Submitted by:

Debanti Roy

**ACKNOWLEDGMENT**

I, Debanti Roy, would like to convey my sincere gratitude to DataTrained Academy and Flip Robo Technologies for giving me this opportunity to do this project. I would like to thank all mentors and SME’s for extending their support all through the process which helped me complete this project.

**INTRODUCTION**

Business Problem Framing

According to a 2016 Gallup Poll[1], real estate was rated the best long-term investment – well ahead of gold, stocks and mutual funds, savings accounts/CDs and bonds. Hence, in order to make an investment in the real estate market, one needs to have a good understanding of the market trends. Analysing the real estate data can help in better understanding what the demands of the market is.

Conceptual Background of the Domain Problem

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.

Review of Literature

Predicting House prices are an important aspect of the real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in Australia to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Australia. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.

Motivation for the Problem Undertaken

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 company is looking at prospective properties to buy houses to enter the market. Thus, the analysis can help them understand the actual value of the prospective properties and decide whether to invest in them or not.

The company will also get to understand the following:

* Which variables are important to predict the price of variable?
* How do these variables describe the price of the house?

**Analytical Problem Framing**

Mathematical/ Analytical Modeling of the Problem

The data provided to us in the problem statement is unsupervised data. The problem statement contains both utilitarian value and hedonic values. Thus, I have performed both univariate and bivariate analysis to analyse these values using different plots like strip plot, distribution plot, box plot and violin plot. .

In this project I have also done various mathematical and statistical analysis such as describing the statistical summary of the columns in which I found that the Since the dataset contains object data type, I used label encoding method to convert the object data into numerical data. Checked for correlation between the features and visualized it using heatmap.

**Data Sources and their formats**

The data was collected from the US real estate market. Results indicate the price of a property and which factors can led to the increase or decrease in the price. The dataset is provided to us by Flip Robo which is in CSV format in two files “test.csv” and “train.csv”. The data contains a total of 1460 entries each having 81 variables, test data has 292 rows and 80 columns and train dataset contains 1168 rows and 81 columns.

I have used train data to train my model and test data to make predictions. The data contains several object columns which I have treated using different encoding techniques. I have also dropped certain unimportant columns and removed null values. While describing the data I found the presence of skewness and outliers. Since most of the columns are categorical so I haven’t removed outliers and skewness.

The data is provided

Data Preprocessing Done

* Firstly, I have imported the necessary libraries and dataset.

****

* Checked the dimension of the dataset, data types of the features.
* Checked the information of the data frame using info() and checked the number of unique values present in the dataset.
* I have checked the null values and replaced them using imputation techniques.
* Checked the value count of each column. By checking value count function, I found some of the columns contains duplicate and unwanted entries, so I have replaced them using appropriate values.
* Visualized each feature using seaborn and matplotlib libraries by plotting strip plot, distribution plot, box plot and violin plot.
* Encoded the data.
* Checked the skewness and outliers. Since most the columns are categorical, so I haven’t removed skewness and outliers.
* Checked the correlation between the features and visualised it using heat map.
* Described the data using describe().

Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

* State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Hardware required:

* + Processor: core i5 or above
  + RAM: 8 GB or above
  + ROM/SSD: 250 GB or above

Software required:

* + Anaconda 3- language used Python 3
  + Microsoft Excel Libraries: The important libraries that I have used for this project are below:

*import numpy as np*

It is defined as a Python package used for performing various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

*import pandas as pd*

Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. The data-frame term is coming from Pandas only.

*import matplotlib.pyplot as plt and import seaborn as sns*

Matplotlib and Seaborn acts as the backbone of data visualization through Python.

Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statical interferences and plotting 2D graphs of arrays.

Seaborn: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy. It is built on the roof of Matplotlib and is considered as a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

*from sklearn.preprocessing import LabelEncoder*

There are several encoding techniques like Label Encoder, OneHotEncoder, Ordinal Encoder.

In this project I have used OneHotEncoder technique to convert categorical data or object type data into numerical data.

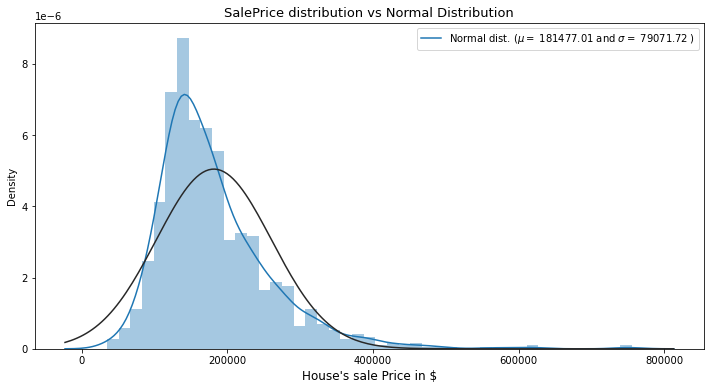
**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods):

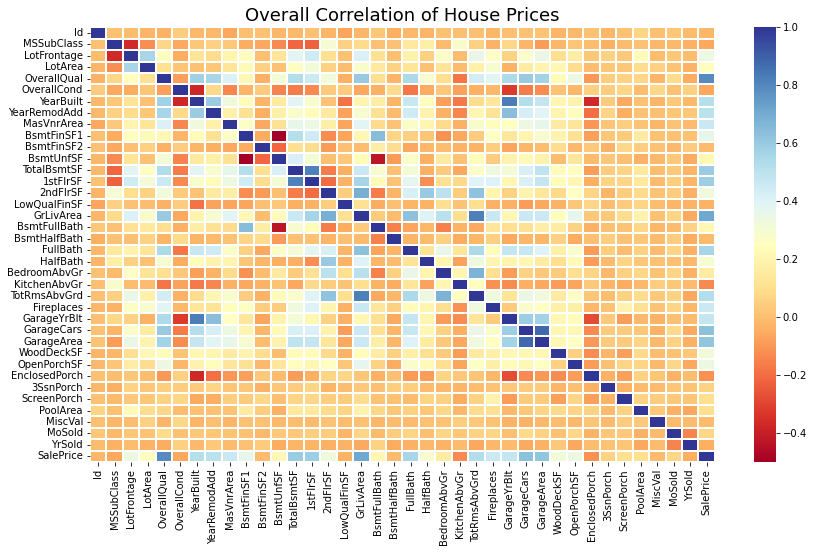
* I have used “.replace()” function to replace unwanted entries in the columns.
* Used “OneHotEncoder” method to encode the categorical features into numerical.
* Described the statistical details of the features using “.describe()” method.
* Used “Pearson’s method” to check the correlation between the features.
* Performed both univariate and bivariate analysis using seaborn and matplotlib.

Visualizations

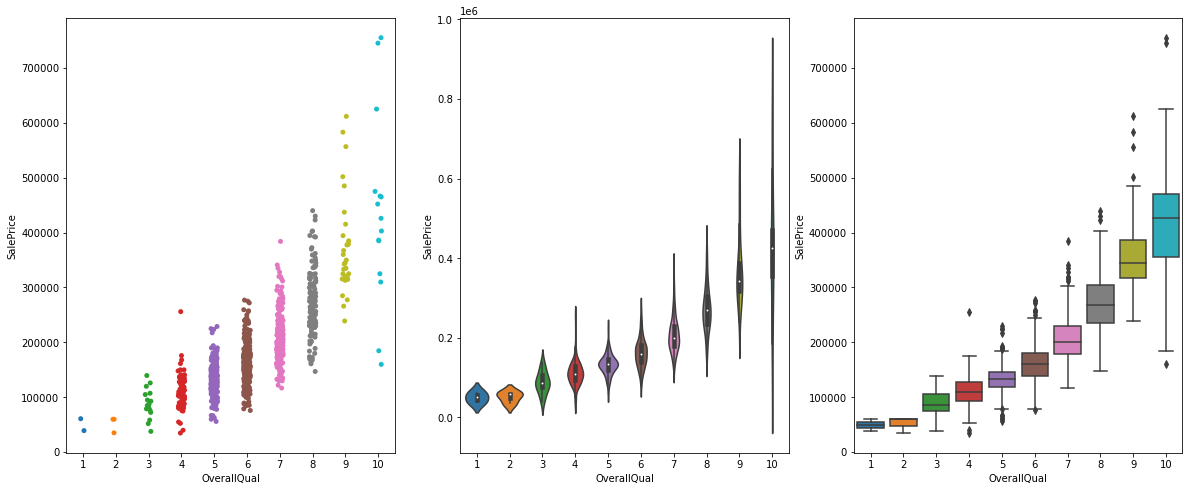
To understand any kind of data it is important to perform Exploratory data analysis (EDA). This is a combination of visualizations and statistical analysis (uni, bi, and multivariate) that helps us to better understand the data we are working with and to gain insight into their relationships. So, let's explore our target variable and how the other features influence it.



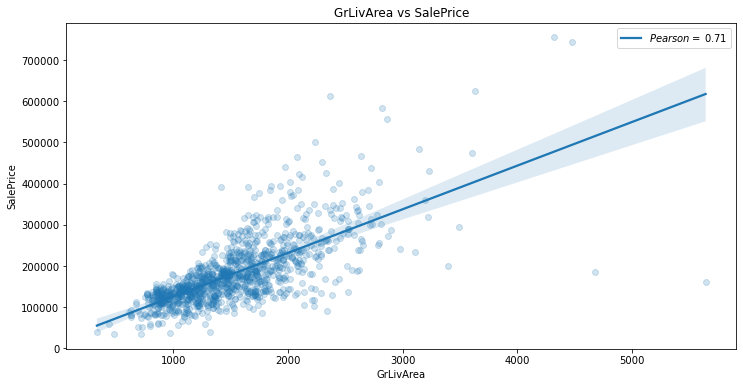
We can clearly see the plot is right-skewed. Despite that, let's leave it like that for now, we'll deal with that later in the notebook.

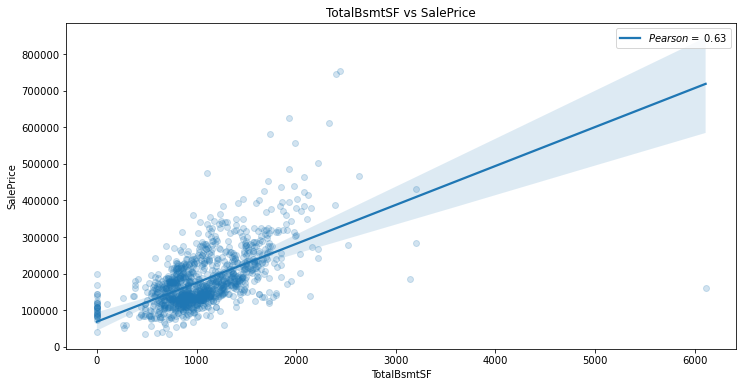


The correlation matrix is the best way to see all the numerical correlation between features. Here we can understand which features correlates most with our target variable we can investigate them more in depth.

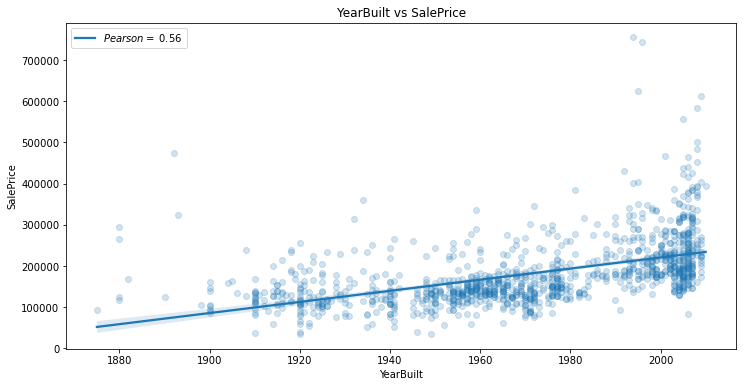


The plots help us understand the relation between SalePrice, OverallQual and TotRmsAbvGrd.

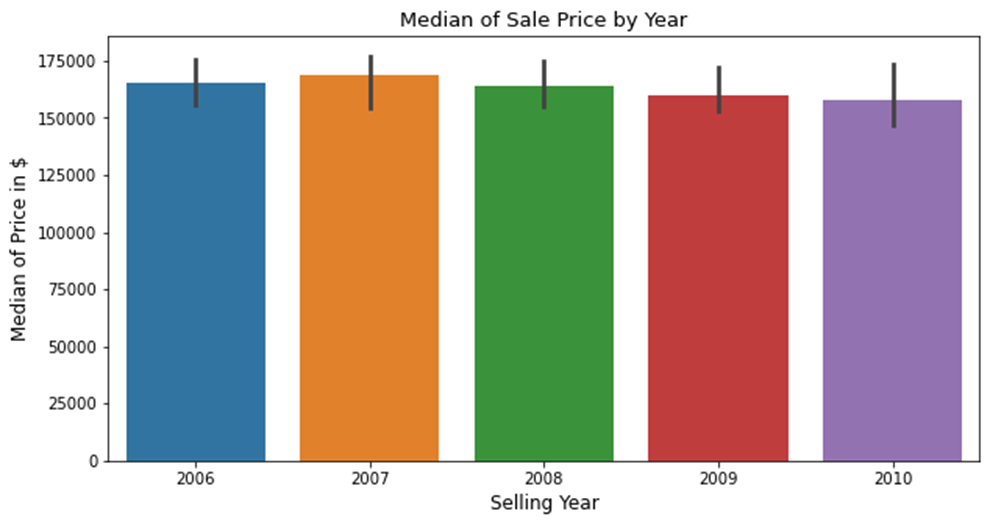




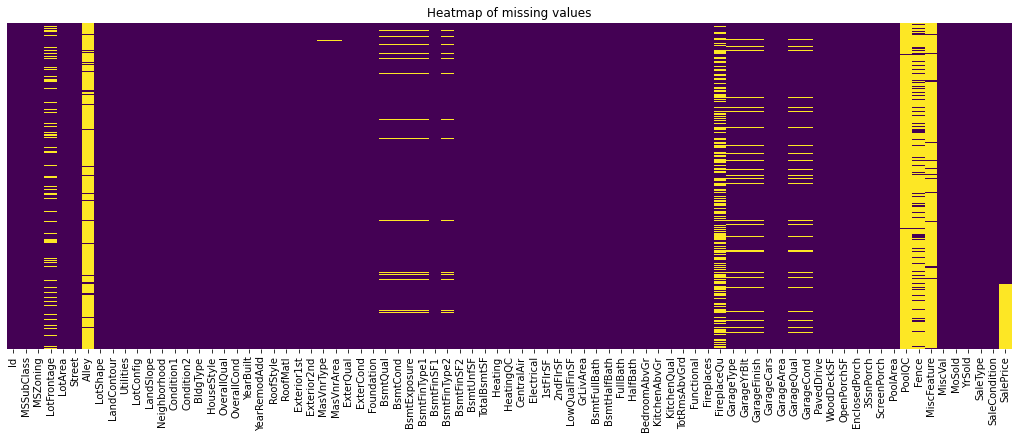
These plots help us understand how Sale Price was affected due to GrLivArea, TotalBsmtSF.



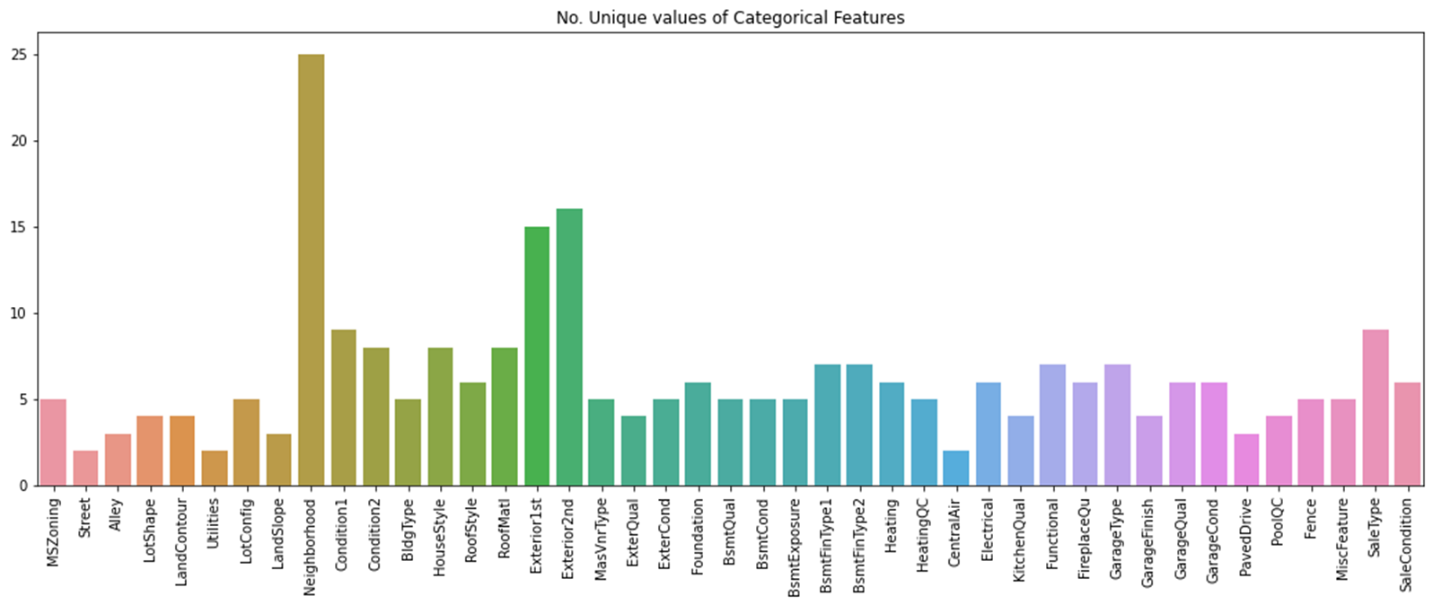
We can see that YearBuilt has an undeniable relation with the SalePrice.



A box plot describes the Median of Sale Price by Year.



We have used heatmap to spot missing values.



This bar plot gives us an idea of the unique values in the categorical features.

Interpretation of the Results

The results that were interpreted from the visualization are as follows:

* The survey reveals that TotalRmsAbvGrd had greater impact than OverallQual while comparing with SalePrice.
* The data also reveals that GrLivArea is more preferred than TotalBsmtSF. There is a constant relation between YearBuilt and SalePrice.

**CONCLUSION**

This model can 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.

Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.