

Housing: Price Prediction

Submitted by:

Pramodh.P

**ACKNOWLEDGMENT**

References used that helped in completing the project Online Research

Data already provided , Data description provided

Site names refered for understanding of the steps and completed process:

Gate

Google

Medium

Kaggle Competition

**INTRODUCTION**

Introduction The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry. Notably, this has been done in Zillow’s Zestimate [4] and Kaggle’s competitions on housing prices [2]. In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties for a US-based housing company named **Surprise Housing**

* Business Problem Framing

Predict the price of the house with the provided variables

And understand how the prices are affected with the variable based on which we can come up with strategy and target areas that will yield high return.Using the model we can understand the market and come with the required strategy

**Analytical Problem Framing**

The main steps in our research were the following.

• Exploratory Data Analysis (EDA). By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

• Feature Selection

• Modeling We apply Liner regression for prediction of the housing prices.

**A pseudo code of understanding of the steps :**

Step 1: Import the data

Step2: Find the missing values in the data set as the cleaning process is done and separate the data

Step3: separate the data to numerical and categorical

Step4:”Visualization “

selection of feature by heatmap or matrix correlation.

Step5: splitting of the dataset into two part training data and the testing data

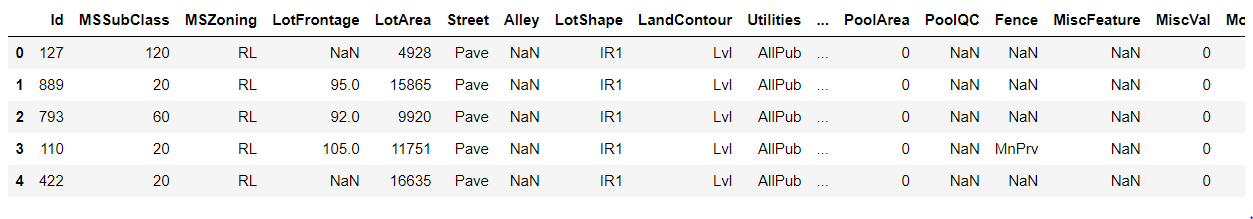
Step6: apply or fit the regression technique on training data and test it with testing data

Step7: Compare the accuracy result.

**Data Set**

**Importing data sets provided**

Example of the data



**Splitting the data**

We are splitting the features into two

1; Numerical Data 2. Categorical Data

**Cleaning The data**

Removing or Dropping unnecessary columns

Finding Nan

**Separating the Target Feature**

In our case the SALE PRICE

**Data Visualization**

Data visualization is the graphical representation of information and data. By

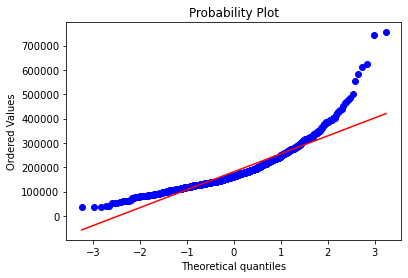
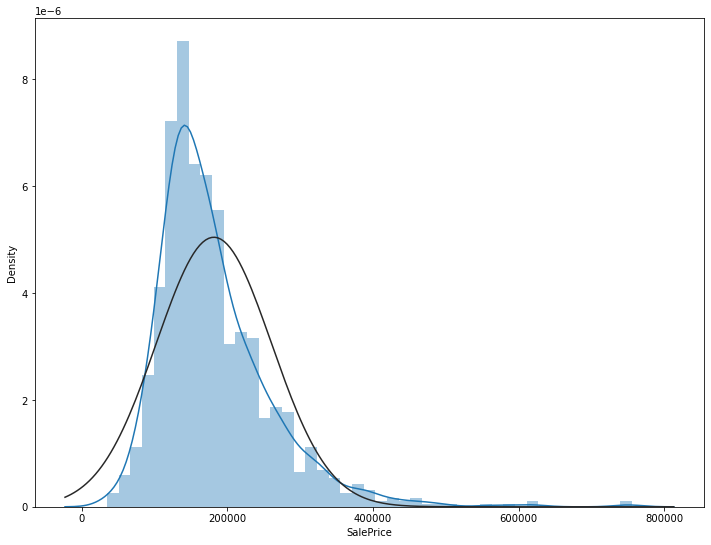
using visual elements like charts, graphs, and maps, data visualization tools provide an

accessible way to see and understand trends, outliers, and patterns in data. In the

world of Big Data, data visualization tools and technologies are essential to analyse

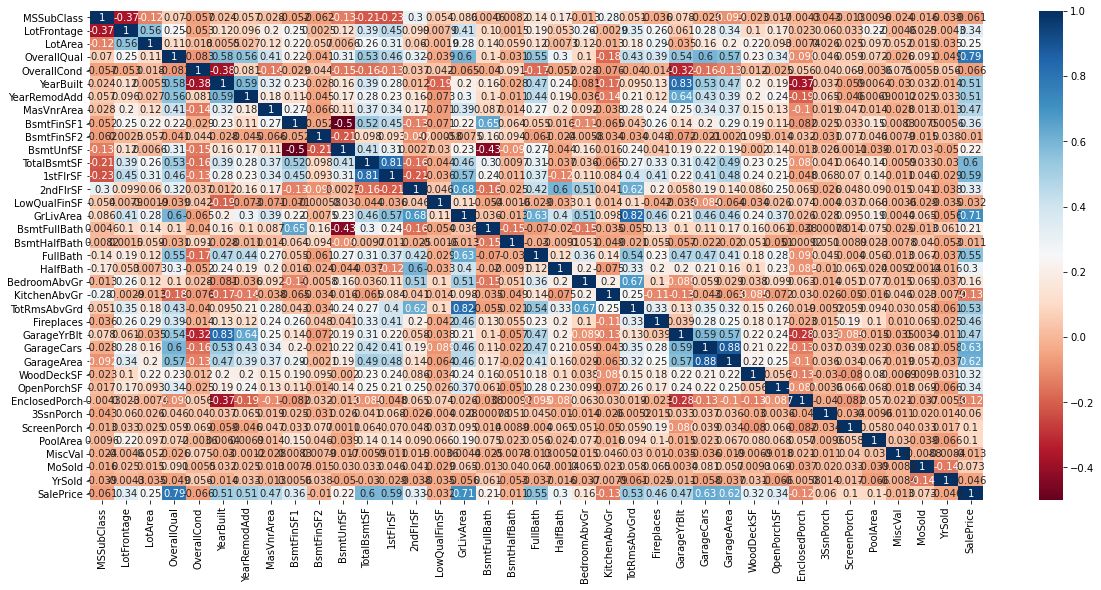
massive amounts of information and make data-driven decisions

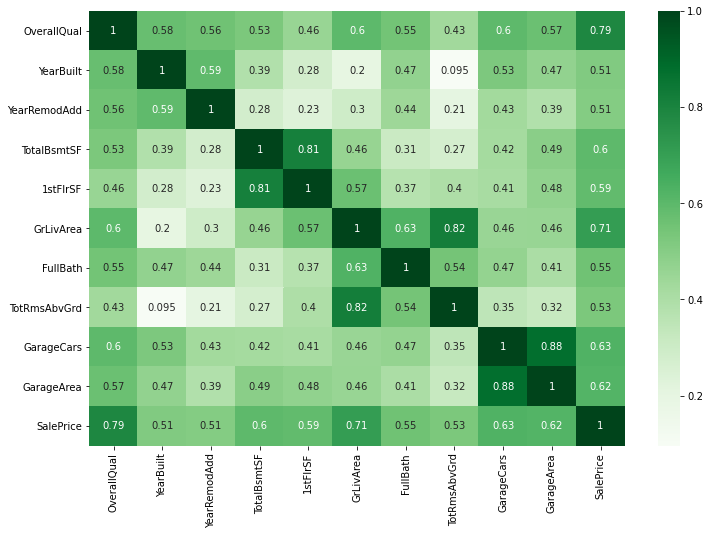
**Distribution Of target variable - Sale Price using Seaborn distplot and Probability Plot**



**We will use heatmap to see the Correlation between variables**

**Find the highest and plot using a heatmap**



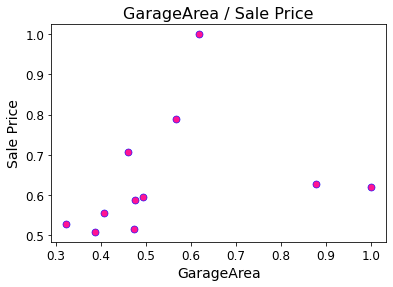
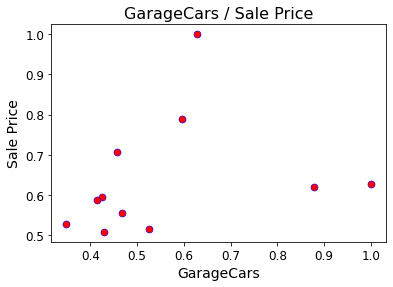
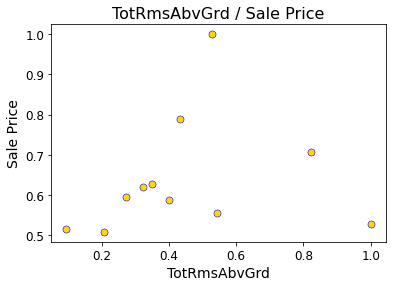
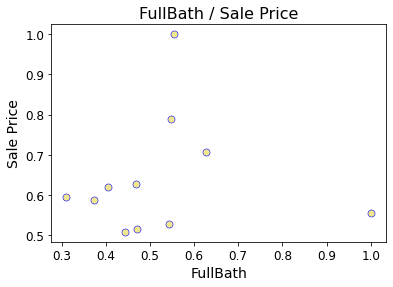
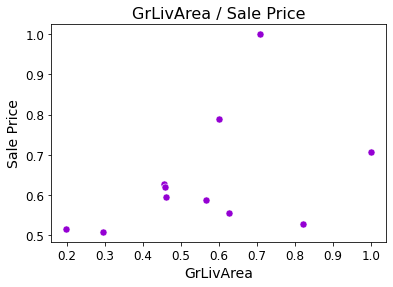
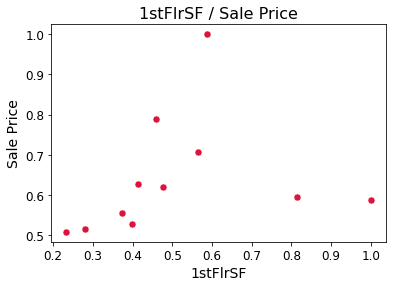
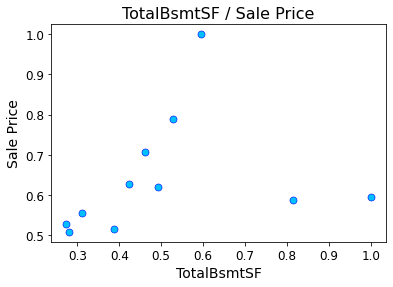
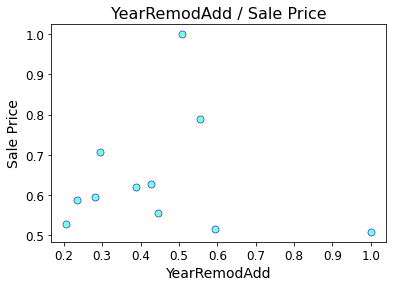
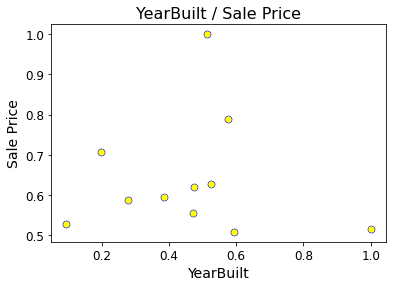
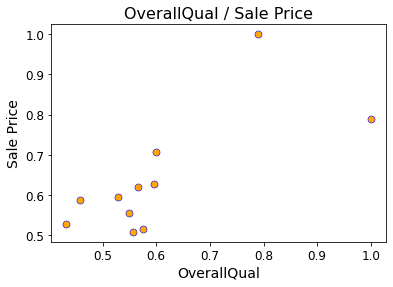


Visualize top variables with highest correlation.

After extracting the top features we will plot a graphbto visually see

I have used scatter plot for this.

Graphs below:

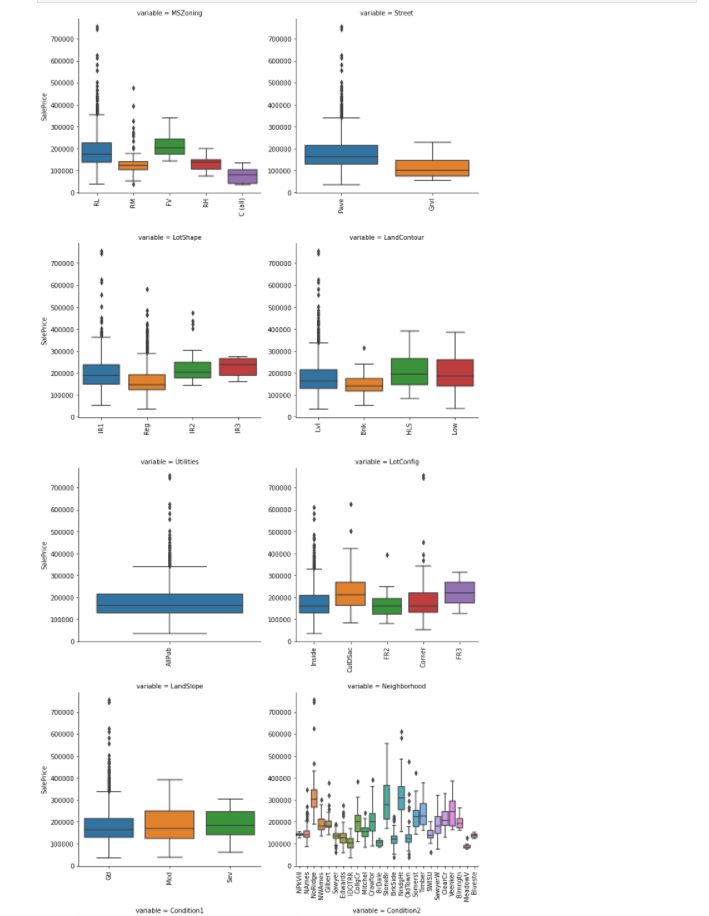


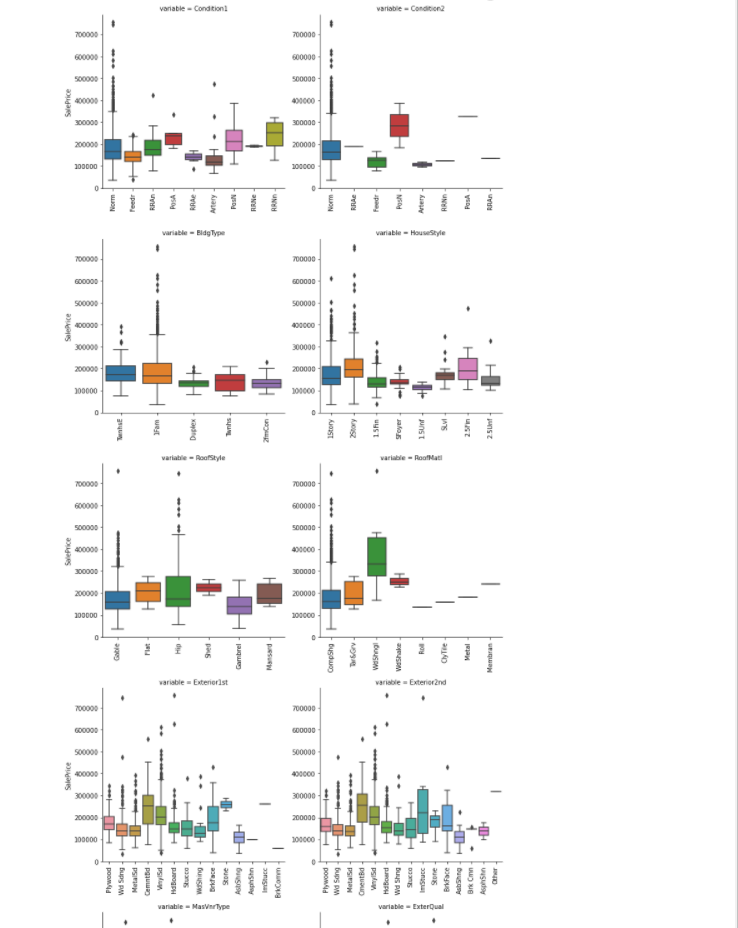
**Categorical Data**

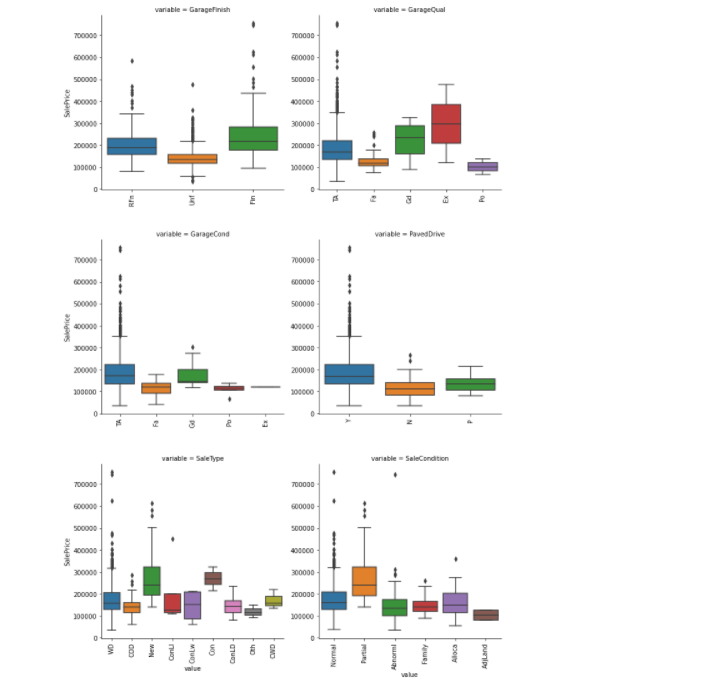
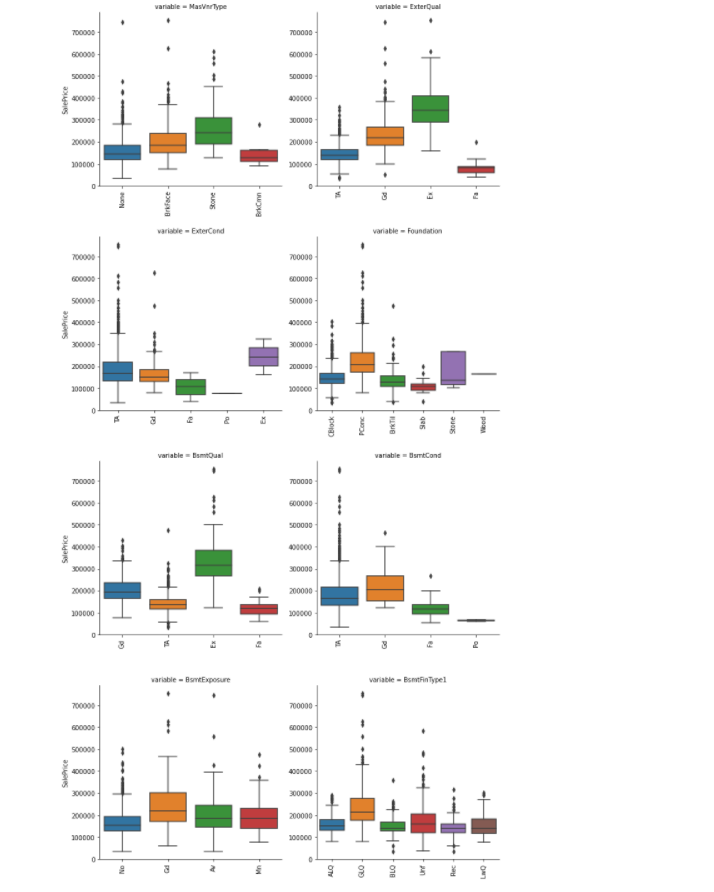
Step 1 : Checking for null or missing values and dropping those with more than 50%**.**

**Visualizing Categorical Data:**

Using Box plot to visualize the data against target feature and find the effect of the feature on the target I ours Sale Price.







**Preparing The data**

Filling the empty values in our data set

First method used is to find the mean column and fill the empty with the mean

Fill zero where ever appropriate based on the feature

**Training the model**

**First we split the data into training and test**

Regression Model

• Linear Regression is a machine learning algorithm based on supervised learning.

• It performs a regression task. Regression models a target prediction value based on

independent variables.

• It is mostly used for finding out the relationship between variables and forecasting.

**CONCLUSION**

**By analyzing the data provide we have created a mode for predication of house price based on various factors**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

* 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.