**House Price Prediction Using Machine Learning Algorithms**

*By Kunal Chand*

**1. Problem Definition.**

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors to the world’s economy. It is a very large market and various companies are 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, improve their marketing strategies and focus on changing trends in house sales and purchases. Predictive modeling, Market mix modeling, 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 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 a 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.

**2. Data Analysis**.

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can analyze data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are, and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with

Which they are working. We divided the data 8:2 for Training and Testing purposes respectively.

**3. EDA Concluding Remark.**

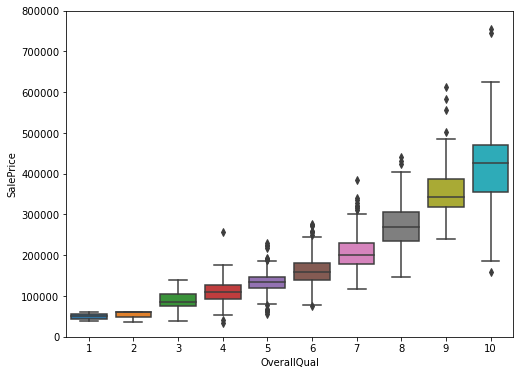
As for any basic model building, we have to understand the type of target variable; the data of the target variable is continued or classified.

Data Analysis is always the difficult part, for better understanding different kinds of bar plots, distribution plots are created with the target Column for finding the insights of the dataset we have.

Analytical Modeling always starts with the target variable we have, and in that case, our target variable is Sales Price, for that, we create some box plots with the target variable to understand which feature columns help to learn the model best and which feature columns reduce the accuracy of the model.

And after finding the relation and correlation with the target variable we choose either Regression Model or Classification Model. Here in this problem, our target feature column continues so we build our Machine Learning model on Regression.

Visualization of the Attrition with gender column.

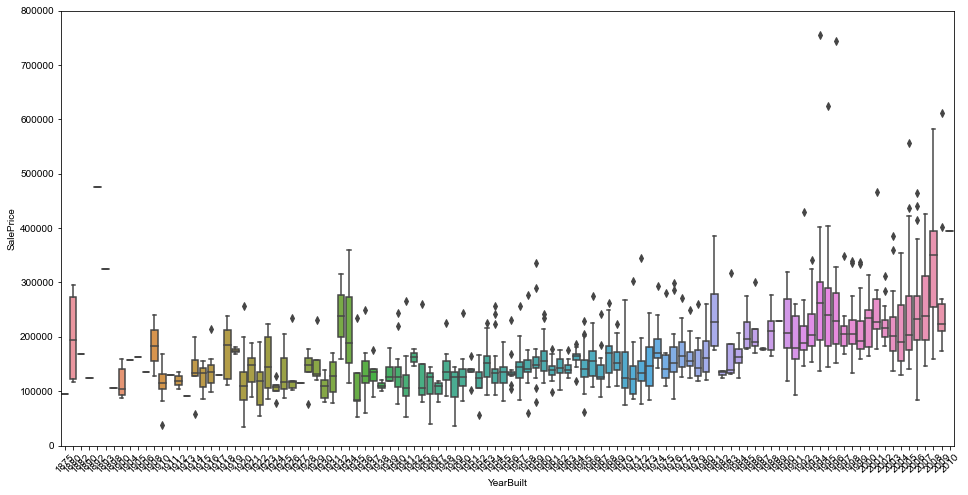


Observations:

1. As from the above observations, Overall quality increases the Sale price of the House.

2. Overall quality will depend on other features as well.

3. As the Score increases Sales price also increases.

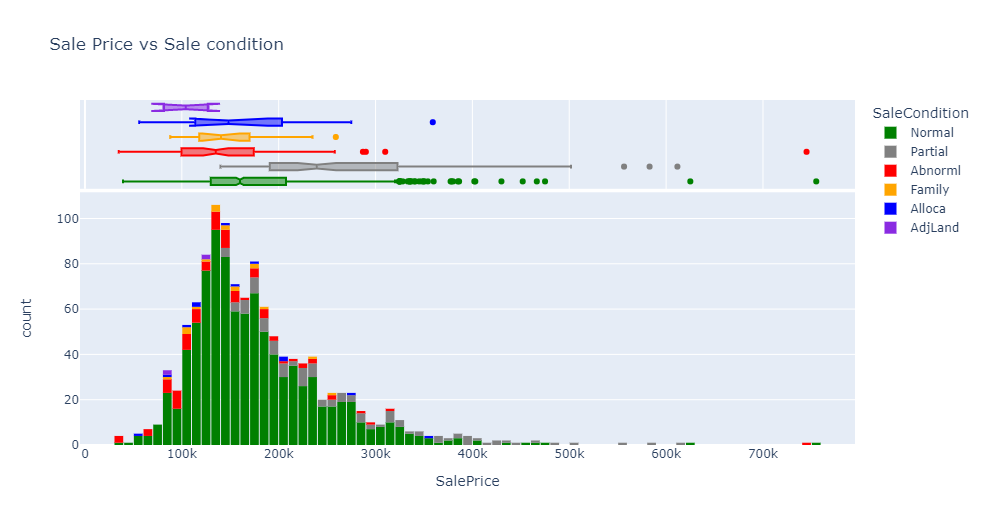
****

Observations:

1. As from the above observations, Overall quality increases the Sale price of the House.

2. Overall quality will depend on other features as well.

3. As the Score increases Sales price also increases.

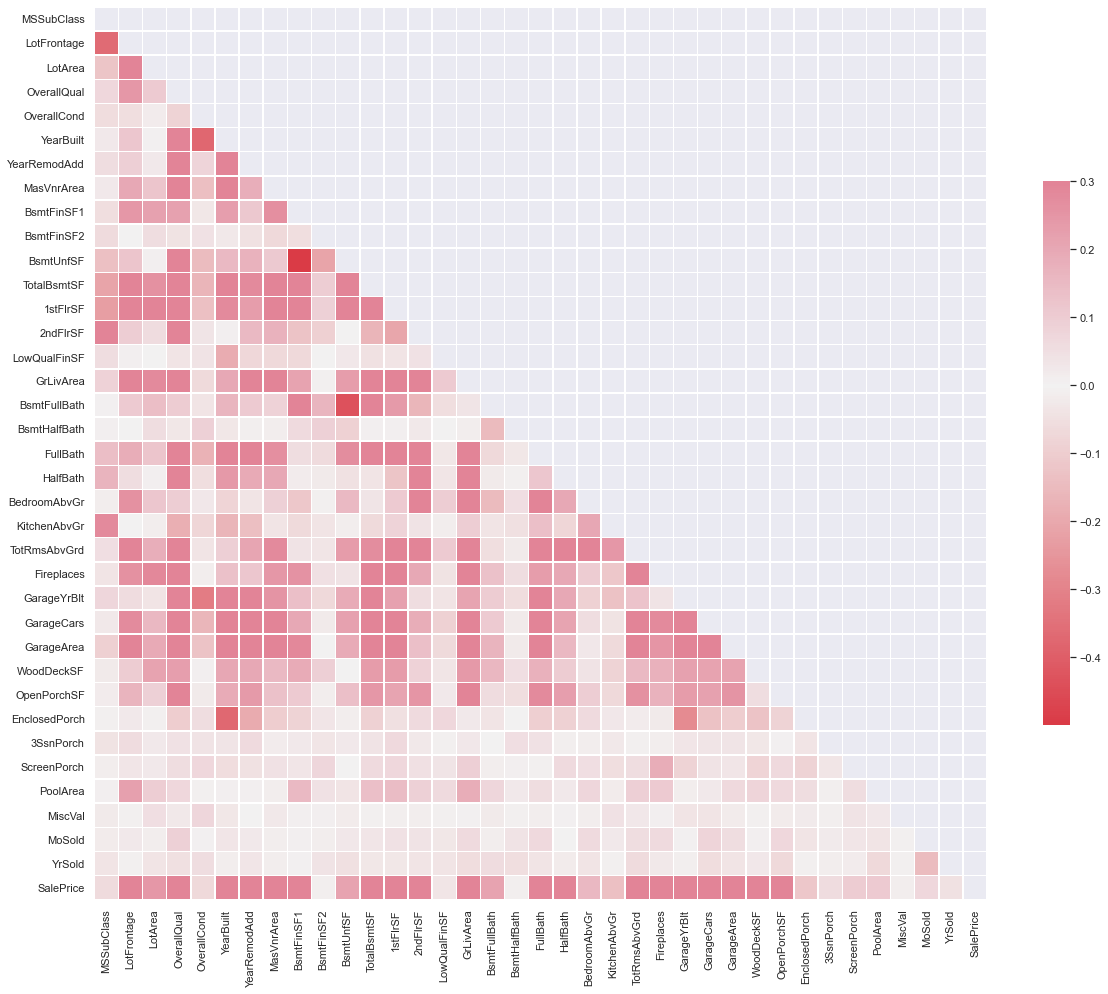


Observations:

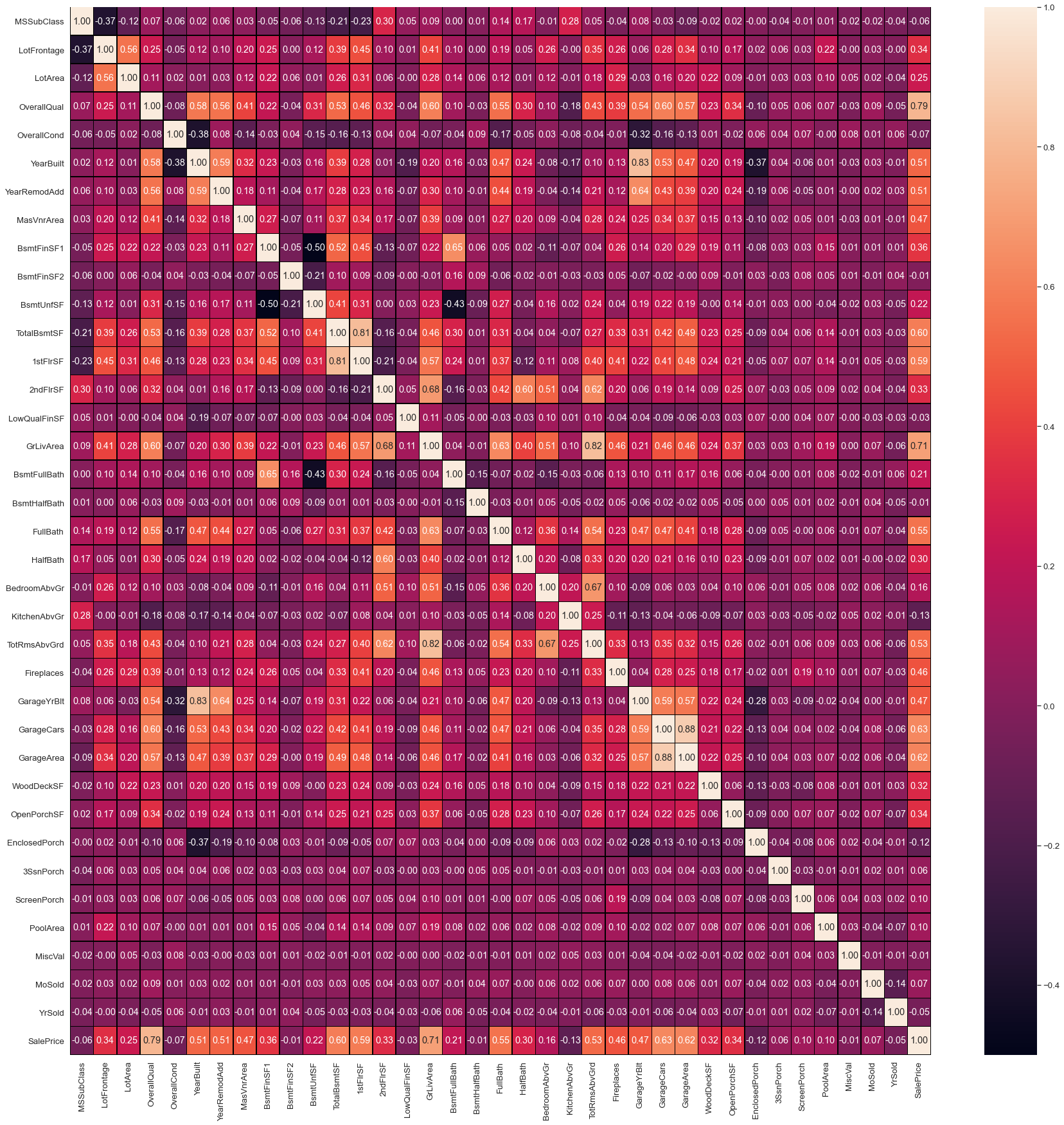
1. As from above observations and plotting we can easily see the Sales Conditions increases the Sales Pricing.

2. Abnormal condition increases the sales Prices as expected.

**CORRELATION BETWEEN THE COLUMNS:**



Heat Maps Plotting



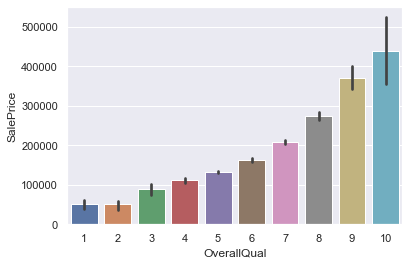
Top Correlated feature columns



observations:

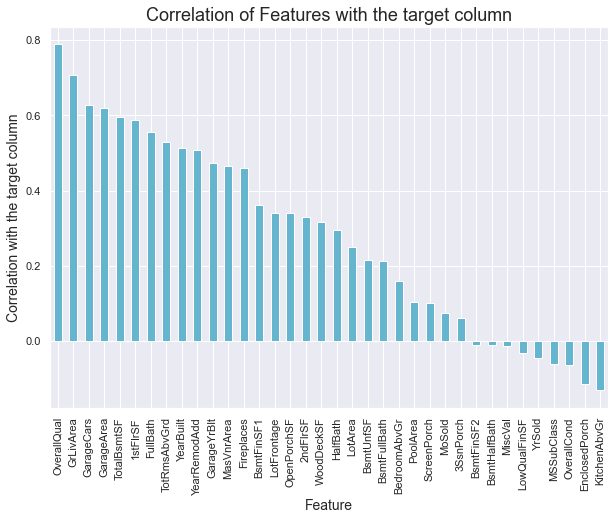
1. Most of the features are correlated with each other like Garage Cars and Garage Areas.

2.OverallQual is highly correlated with target feature SalePrice 0.79 can you see. we'll see how it affected the sale price in the below graph.



Observations:

Overall quality increases the sales prices of the house.



observations:

1. Overall quality, Ground live area, garage cars are highly

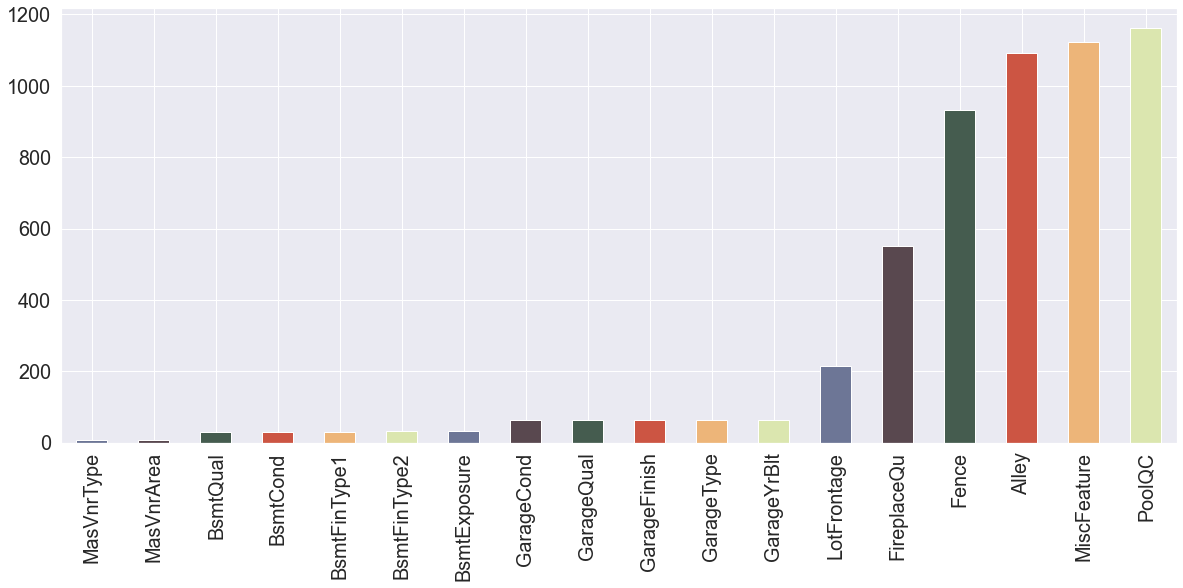
correlated with target variables and impact more on

predicting Sales price.

2. kitchen above ground and enclosed porch are

negatively impacting on prediction of Sales prices.

**Missing Values**

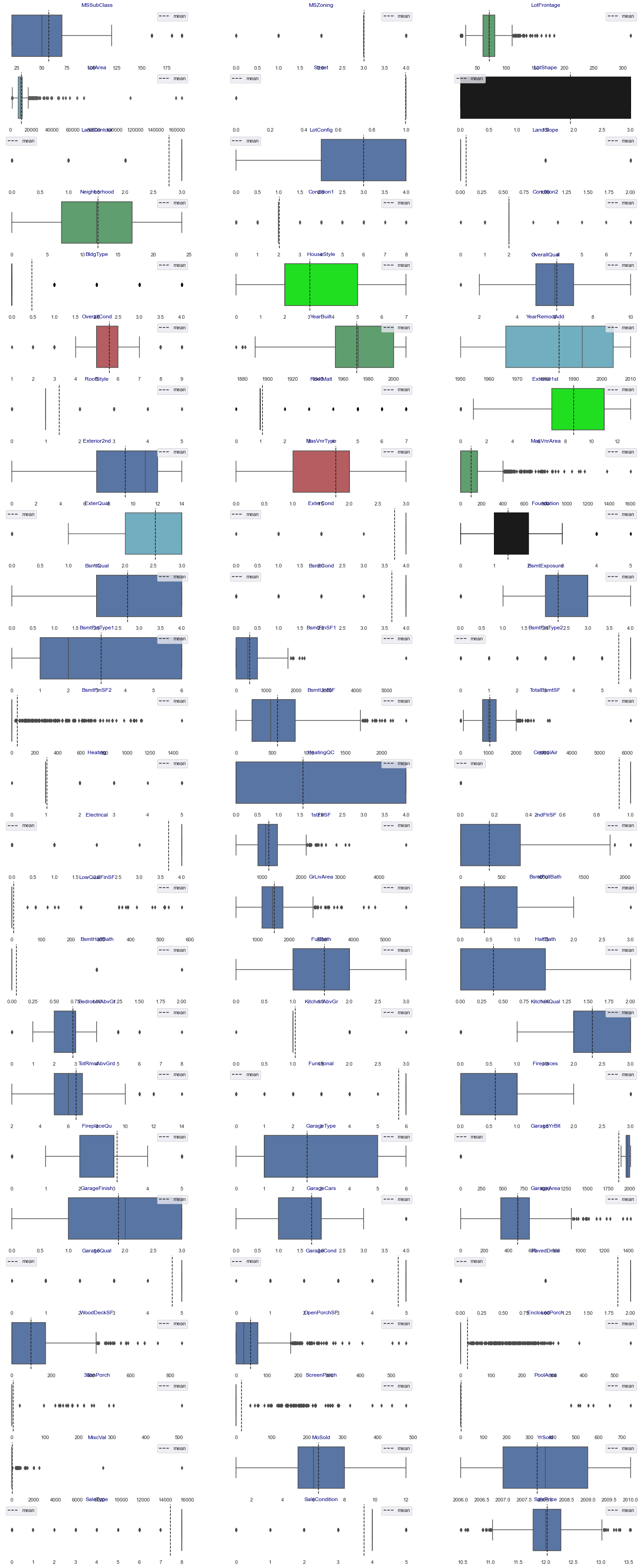


Observations:

Poolqc, Miscfeatures, alley, fence having most missing values.

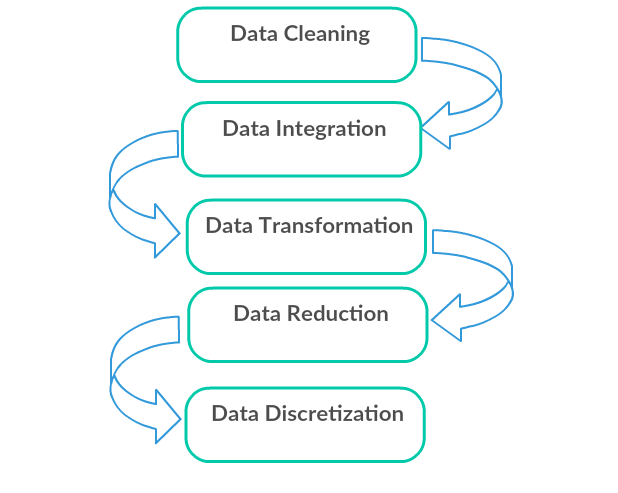
So, we drop these feature columns.

**Detecting outliers:**



**4. Pre-Processing Pipeline.**

Data pre-processing can refer to the manipulation or dropping of data before it is used to ensure or enhance performance, and is an important step in the data mining process.



1. Data Cleaning: First we clean the data which have no use in prediction like the ID column, and then we drop the data which has a high no of missing percentages.
2. Data Integration: then we do some EDA process for finding out the meaning full insights of the data.
3. Data transformation is the process of changing the format, structure, or values of data; we use a labeled encoder for coding the object data into integer data.
4. Data Reduction: it is the process of finding the most correlated columns, and combining them because the machine does not understand which feature columns impact the most on accuracy.
5. Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy, using box plots is makes a clear understanding of the data.

**5. Building Machine Learning Models**.

After analyzing the dataset, I observe that many of the feature columns are object type so first, we have to convert them in integer or float type so that machine interpret the data and for that we do label encode all the feature column.

After label encoding, we find that many feature columns have Nan values so we use mean and median for filling that missing data,

Then find the correlation between the columns with target columns and delete the non-related feature columns.

We observe that the target column is skewed so we remove the skewness of the target column because normal data gives better results when we make the M.L model.

The target column is continuous type so we start work on Regression models building.

* Testing of Identified Approaches (Algorithms)

1. Linear Regression
2. Regurgitation:

Lasso & Ridge Regression

1. Ensemble techniques

Decision Tree Regression

Random forest Regression

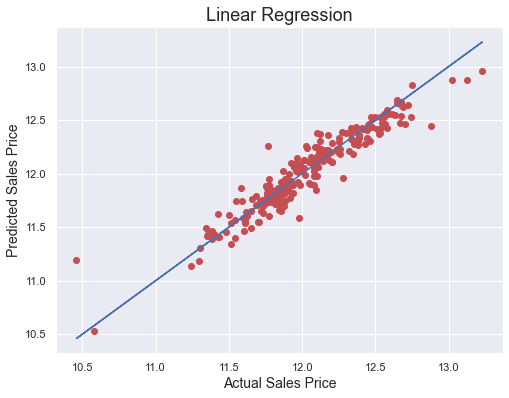
1. Gradient Boosting Regression
2. Support vector machine
3. K-nearest Neighbour Regression

**Linear 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.



**Observations:**

1. This Linear Regression Performs with 90% accuracy for predicting house prices.
2. We use the best-fit line and we can easily see that most of the price points are fall on the line.

Decision Tree Regression

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g Normal, Abnormal, and Family), each representing values for the attribute tested. Leaf node (e.g., present or not-present) represents a decision on the numerical target. The topmost decision node in a tree that corresponds to the best predictor is called the **root node**. Decision trees can handle both categorical and numerical data.



**Observations:**

1. This Decision Tree Regression Performs with 76% accuracy for predicting house prices.
2. After predicting and plotting the predicted data on the best fit line we observe that DTR is not so accurate.

**Random Forest Regression Model**

1. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

2. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

3. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.



Observations:

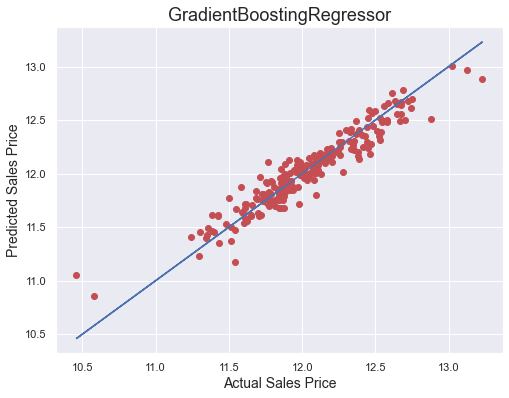
1. RFR performs well but not that well.

2. CV at 6 is giving good results.

3. After predicting and plotting the predicted data on the best fit line we observe that RFR is not so accurate.

# Gradient Boosting Regression

# Gradient boosting is a machine learning technique for regression, classification, and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is a weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

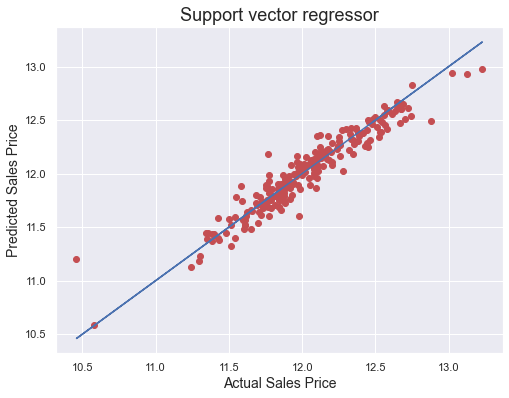


Observations:

1. GBR performs better and gives optimum results.
2. After predicting and plotting the predicted data on the best fit line we observe that GBR is accurate.

# Support Vector Regression

# SVMs or Support Vector Machines are one of the most popular and widely used algorithms for dealing with classification problems in machine learning. However, the use of SVMs in regression is not very well documented. This algorithm acknowledges the presence of non-linearity in the data and provides a proficient prediction model.



Observations:

1. SVR performs better and gives optimum results.

2. After predicting and plotting the predicted data on the best fit line we observe that SVR is accurate.

3. But when we observe that cv is not better than GBR.

# K-nearest Neighbors Regression

KNN regression is a non-parametric method that, intuitively, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood*.* The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error.

While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.



Observations:

1. KNN performs not well and gives no proper results.

2. After predicting and plotting the predicted data on the best fit line we observe that KNN is far behind from remaining algorithms.

**6. Concluding Remarks.**

So, our Aim is achieved as we have successfully ticked all our parameters as mentioned in our Aim Column. It is seen over time as the most effective attribute in predicting the Attrition and that the Gradient Boosting Regression is the most effective model for our Dataset with an R2 score is 0.9055.

The best model is Gradient Boosting Regression. Since the difference between the percentages score of cross-validation and R2\_score is optimum.

At CV: - 9

R2 Score: 90.55288176914567

Cross Val Score: 88.64708889146074

That's it! We reached the end of our exercise.

Throughout this kernel, we put in practice many of the strategies for predicting the prices of the house. We philosophized about the variables, we analyzed 'Sale Price' alone and with the most correlated variables, we dealt with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed categorical variables into dummy variables. That's a lot of work that Python helped us make easier**.**

**Thank you**