**“Python machine learning project - HOUSING: PRICE PREDICTION”**

Submitted by:

Shubham Aggarwal

**ACKNOWLEDGMENT**

i am thankful to all my teachers and mentors who guided me and helped me enough for completion of this project.

**INTRODUCTION**

Business Problem Framing- 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.

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 - 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.

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.

**Independent Variables :**

MSSubClass'

'MSZoning',

'LotFrontage'

'LotArea',

'Street',

'LotShape',

'LandContour',

'Utilities',

'LotConfig',

'LandSlope',

'Neighborhood',

'Condition1',

'Condition2',

'BldgType',

'HouseStyle',

'OverallQual',

'OverallCond',

'YearBuilt',

'YearRemodAdd',

'RoofStyle',

'RoofMatl',

'Exterior1st',

'Exterior2nd',

'MasVnrType',

'MasVnrArea',

'ExterQual',

'ExterCond',

'Foundation',

'BsmtQual',

'BsmtCond',

'BsmtExposure',

'BsmtFinType1',

'BsmtFinSF1',

'BsmtUnfSF',

'TotalBsmtSF',

'Heating',

'HeatingQC',

'CentralAir'

, 'Electrical',

'1stFlrSF',

'2ndFlrSF',

'LowQualFinSF',

'GrLivArea',

'BsmtFullBath',

'BsmtHalfBath',

'FullBath',

'HalfBath',

'BedroomAbvGr', '

KitchenAbvGr',

'KitchenQual',

'TotRmsAbvGrd',

'Functional',

'Fireplaces',

'FireplaceQu',

'GarageType',

'GarageYrBlt',

'GarageFinish',

'GarageCars',

'GarageArea',

'GarageQual',

'GarageCond',

'PavedDrive',

'WoodDeckSF',

'OpenPorchSF',

'EnclosedPorch',

'3SsnPorch',

'ScreenPorch',

'PoolArea',

'MiscVal',

'MoSold',

'YrSold',

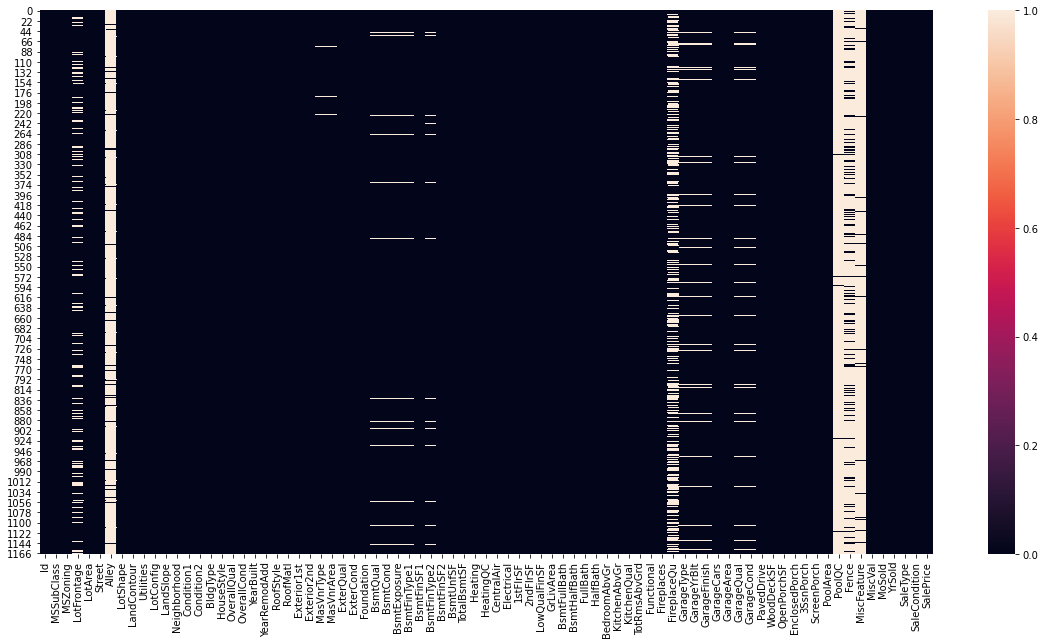
'SaleType',

'SaleCondition',

**Dependent Variable (Target Variable):**

Sale price

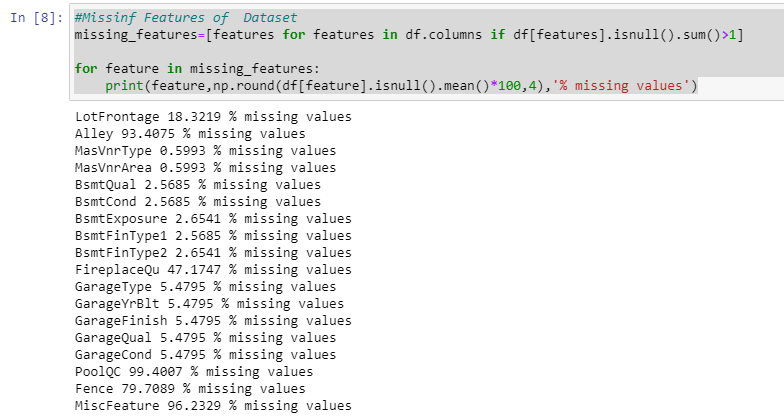
**Missing values:**



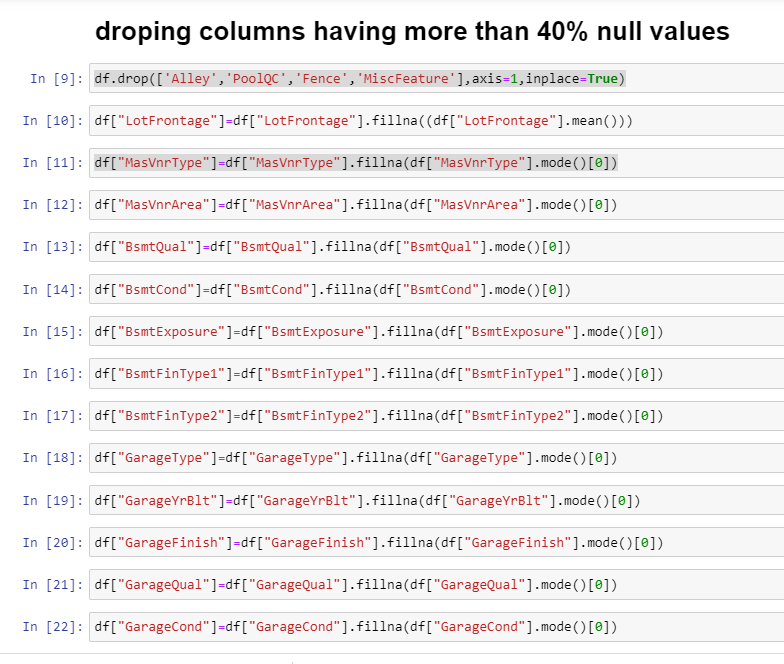
These are a few null values presented by white lines in the heatmap.

Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in a real-life scenarios. Missing Data can also refer to as NA(Not Available) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed

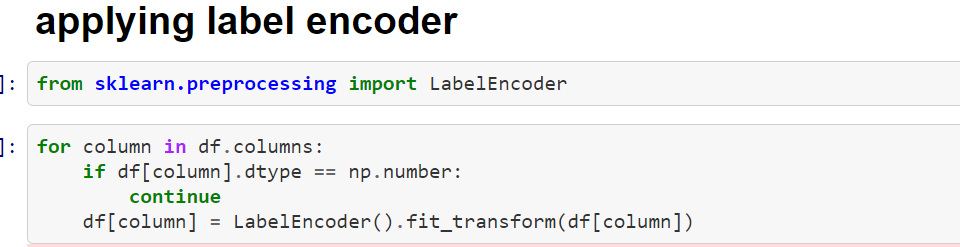
To fill the null values, we have used mean values for the float and for categorical data, we have used mode.



Filling of null values:



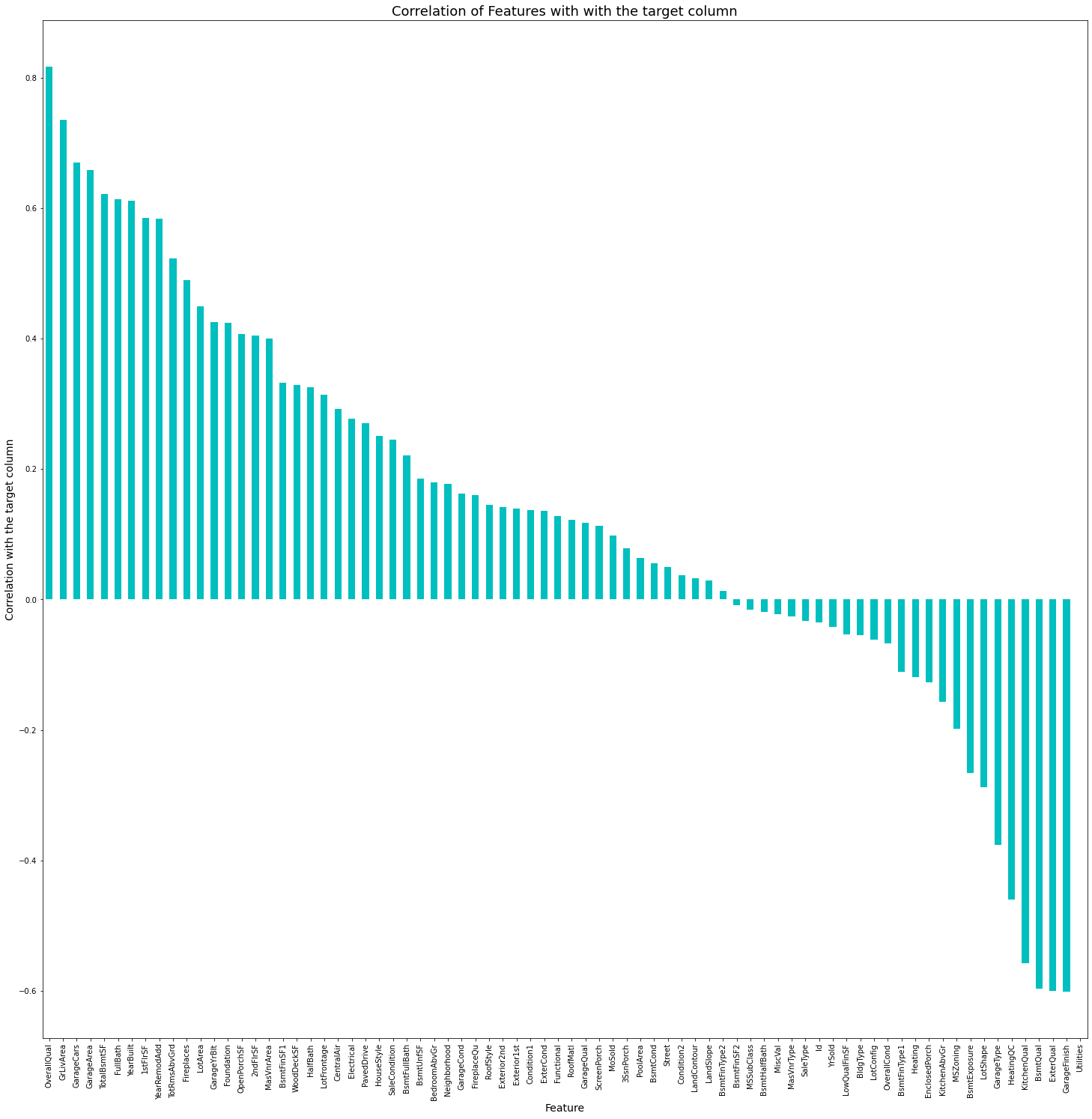
**Label encoder**



**Label Encoding**

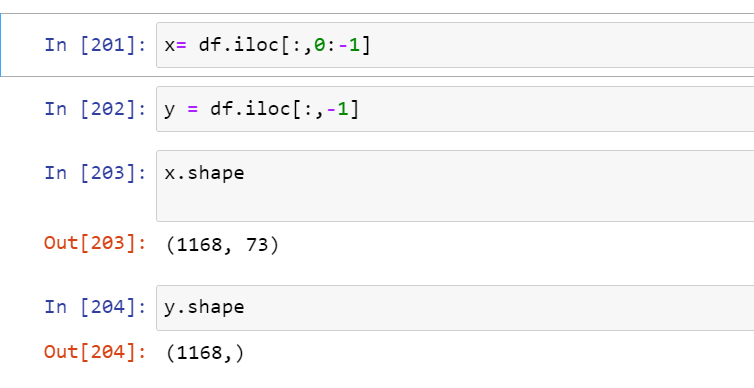
It refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Correlation:



Checking the correlation with target column ( sale price )

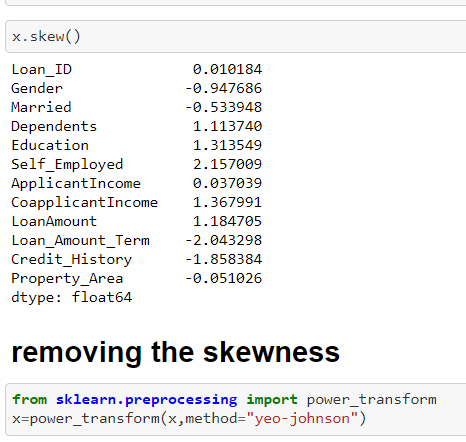
Assigning the variable



**Skewness**

Skewness is **a measure of the asymmetry of data distribution**. Skewness is an asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

Removing the skewness



The skewness is removed from the X named variable not from the Y variable and it was removed using the power transform library which is presented In the sklearn it is important to remove the skewness as it is the part of data cleaning and it also improved the accuracy level of machine learning model

Standard scaler

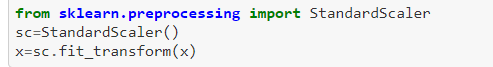
Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.

The standard scaling is calculated as:

z = (x - u) / s

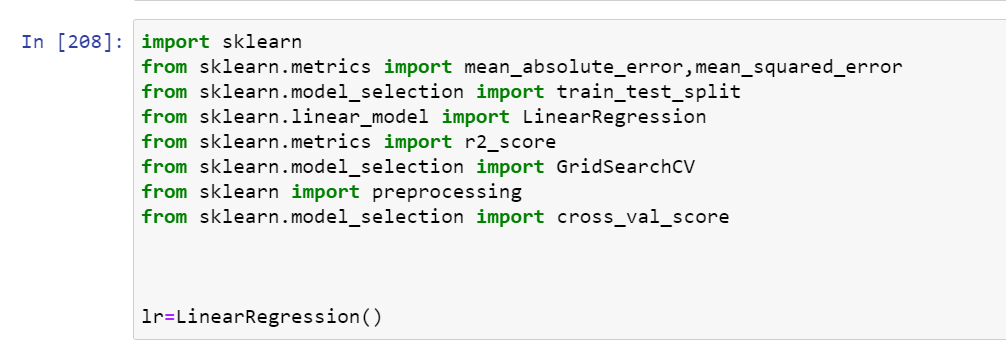
Where :

* z is scaled data.
* x is to be scaled data.
* u is the mean of the training samples
* s is the standard deviation of the training samples

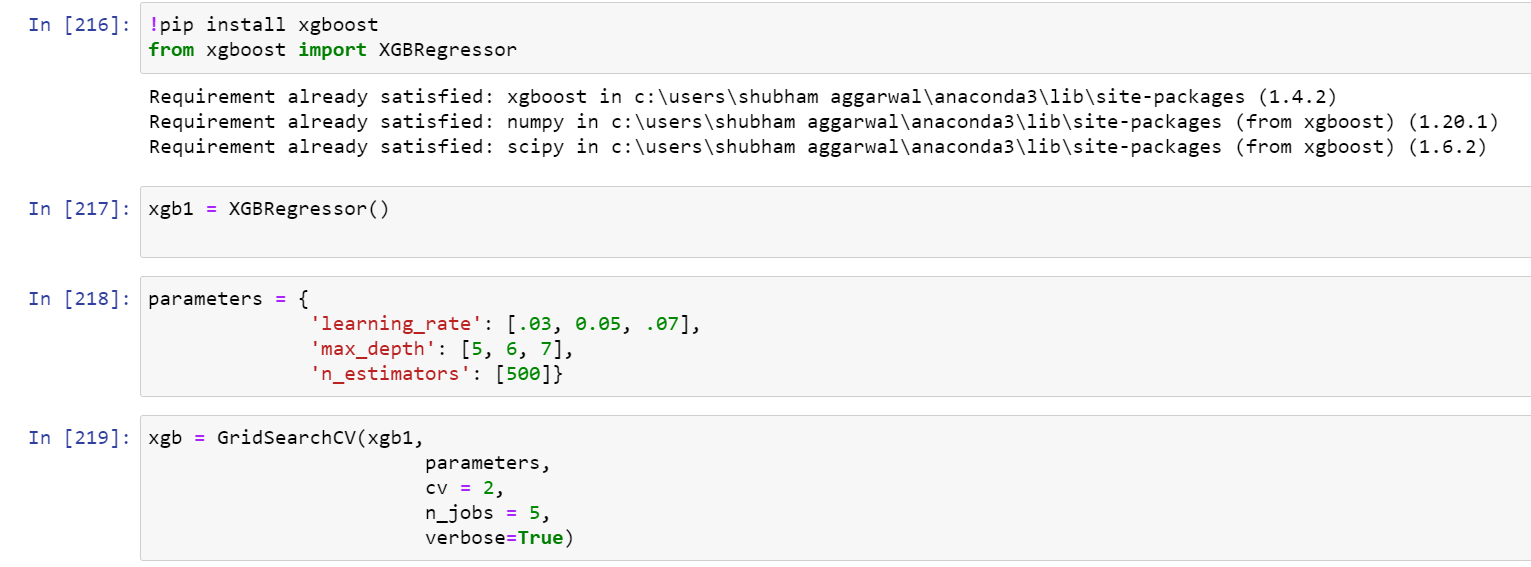


There is a library called standard scaler in sklearn by using this we standardized our data which was present in the X variables

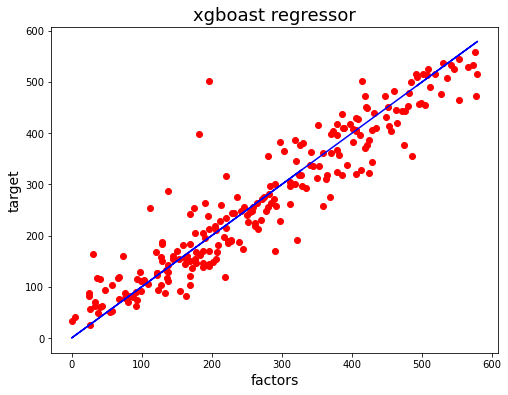
Importing the various libraries:



Here, xgboost was performing the best and was giving 89.05% accuracy.



XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a **gradient boosting framework**. In prediction problems involving unstructured data (images, text, etc.) ... A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.



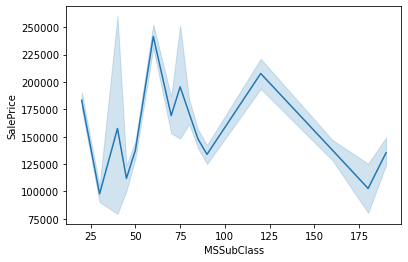
**Grid search CV**

gridSearchCV is a function that comes in Scikit-learn’s (or SK-learn) model\_selection package.So an important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

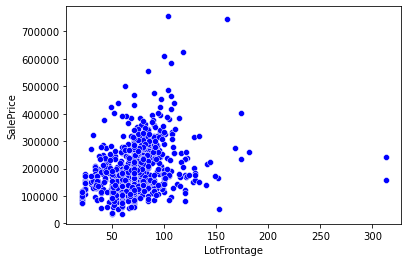
Conclusion from EDA:

Following steps should be taken care of:

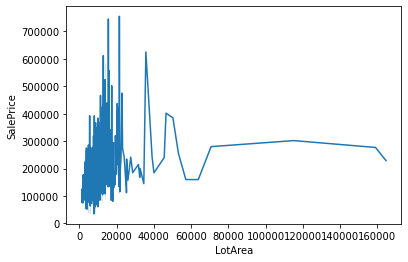
From visualization we came to know that:



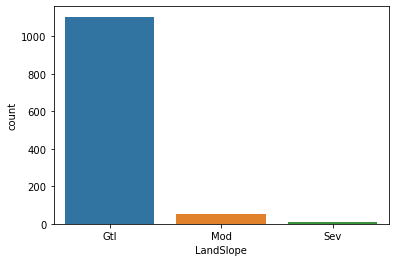
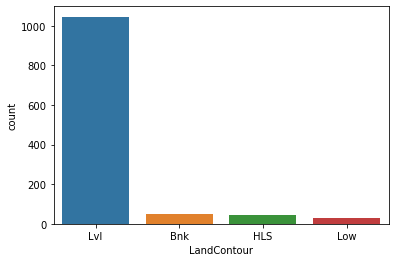
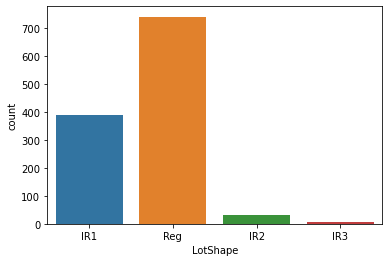
1946 & NEWER -STORY 1945 & OLDER were in the demand but later as per the trend 2-STORY 1946 & NEWER 2-STORY 1945 & OLDER (its sales are increasing the most )



Mostly lot sizes were between 50- 100 but sometimes price increases with increase in lot frontage.



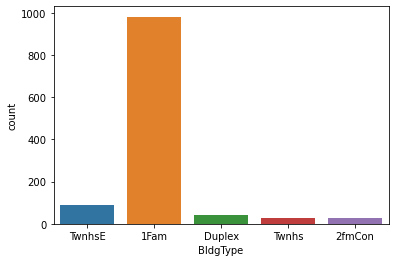
initially the price were high but with the increase in lot areas, prices are constant



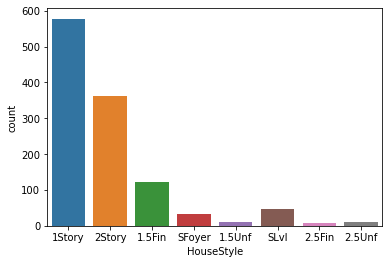
People mostly prefer the shape of property to be regular followed by slightly irregular.

Mostly properties are demanded at flat areas and with that demand of all the public utilities.

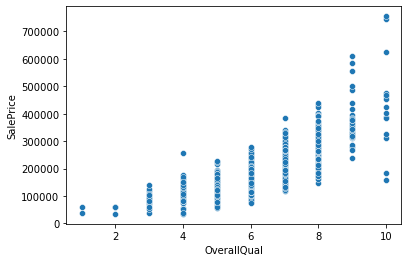
Also, gentle slopes are preferred.



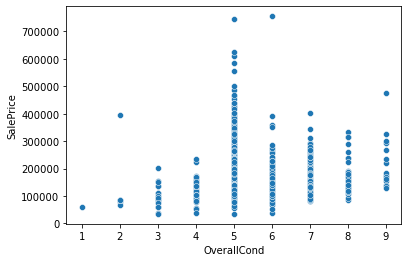
Single family prefer detached type of dwelling property



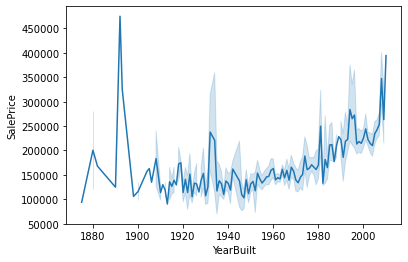
Single story are mostly preferred followed by 2 story.



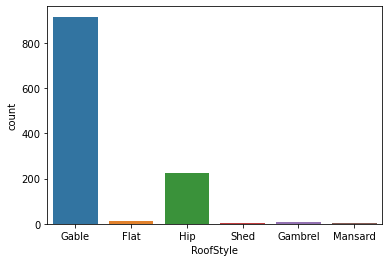
Quality increases with the increase in price.



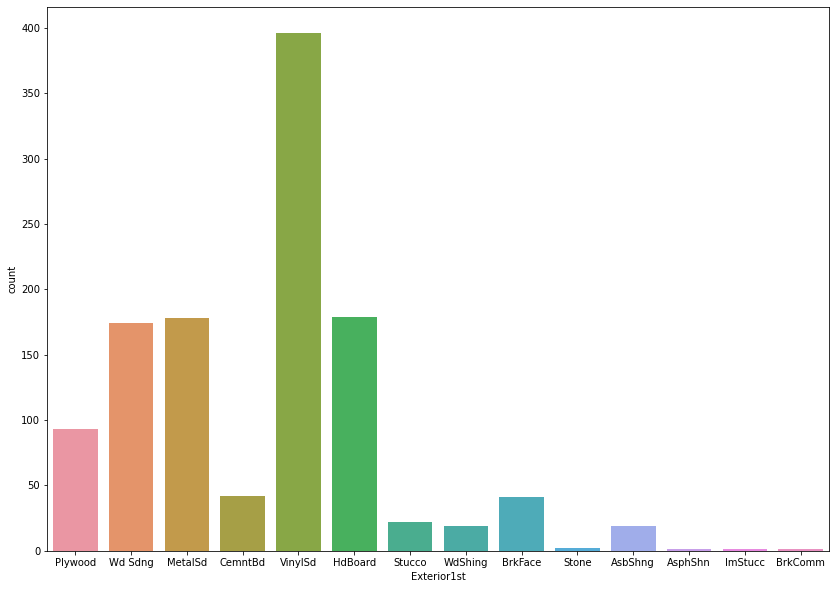
There is not much impact of price on overall condition.



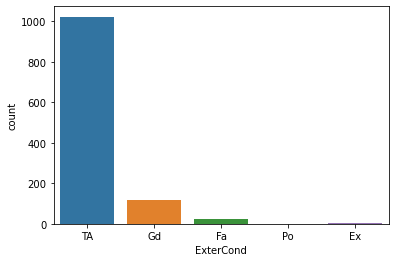
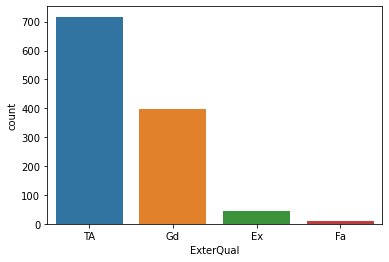
The prices of the house were low between 1920- 1980 but after 2000 prices are expected to increase sharply. Also people prefers remodelling with the passing time.



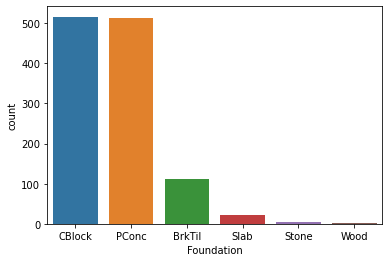
Gable type of roof is preferred with vinylsiding



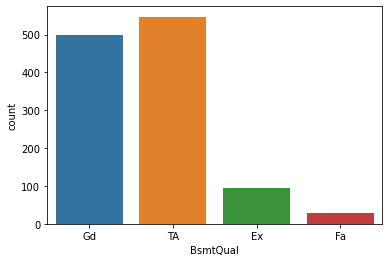
Earlier with the increase in Masonry veneer, the prices were slightly high but their is a sudden fall in the prices.



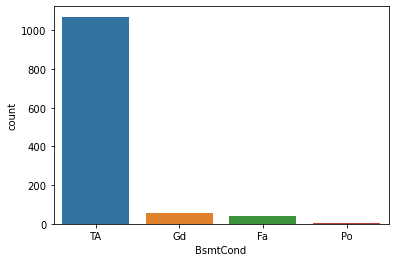
People prefer average exterior quality with average material



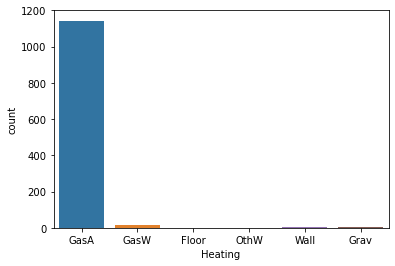
Poured contrete and brick and tiles are mostly used.



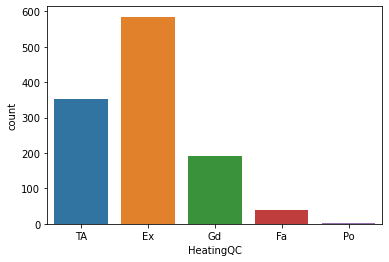
Height of the basement is 80-89 inches then preferred 90-99 inches



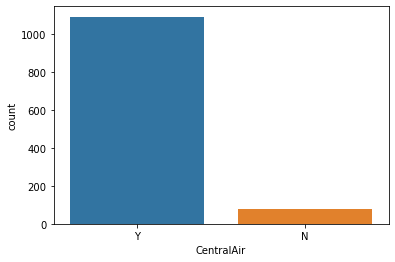
General condition of basement is slightly damped and it is mostly unfinished.

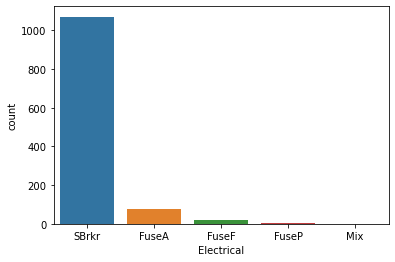


Gas forced warm air furnace id mostly used for heating

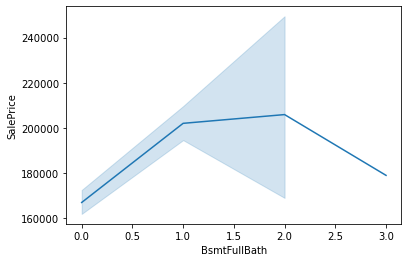


Heating quality and condition is excellent followed by average.

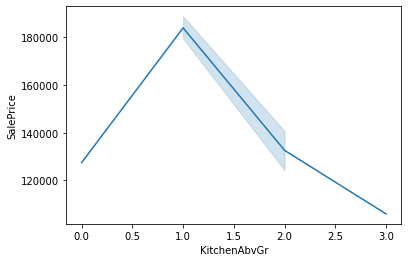




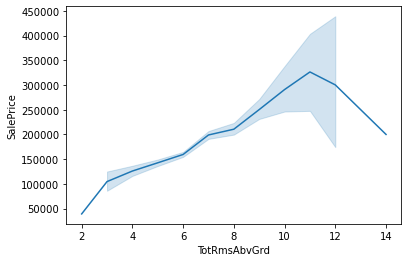
It mostly has a condition of central air conditioning and electrical system is mostly standard circuit.



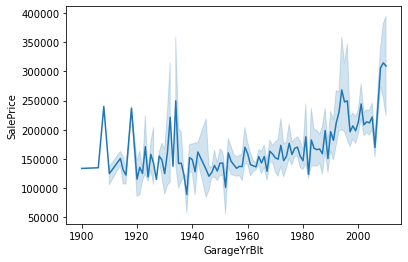
Basement full bathroom increases with decreasing in sales price and same is the case with basement half bathroom



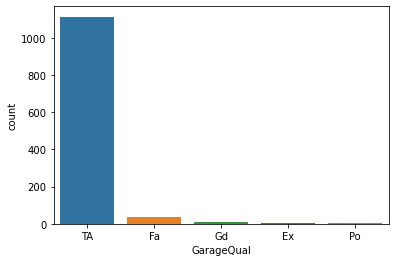
Kitchen grade is increasing with decreasing price.



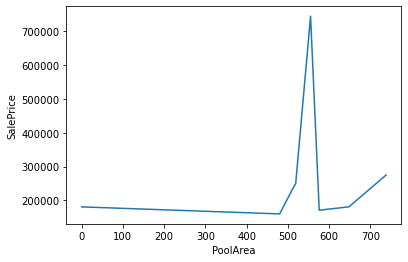
Total rooms are above grade is increasing with decreasing price.



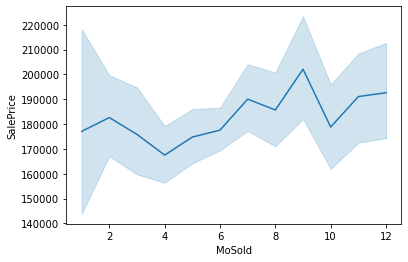
Trend of building garage is increase with the increase in size



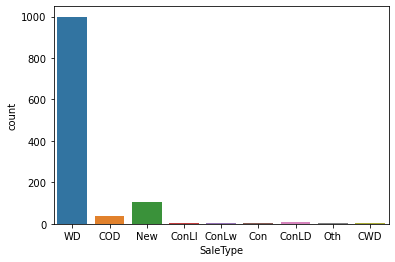
Garage quality is average



Pool area in sq ft increases with increasing price.



Maximum sales are between august and october.



Maximum sales are with warranty deed with normal condition sales.