FLIP ROBO

HOUSING: PRICE PREDICTION PROJECT

Submitted by:

ANJANA.P

ACKNOWLEDGMENT

I would like to express my appreciation to team Fliprobo for giving such a realistic data for analysis, with a full-length description of the project. My mentor Ms. Khushboo Garg has helped me in many stages of this project where I was stuck with problems. I use this opportunity to thank her for helping me at the right time without any delay.

I also thank DataTrained academy team for their wonderful classes which helped me to make this project complete .Also, this project made me search for a lot of data's in several webpages and sites, that helped me to rectify my doubts and, I was able to study more about data analysis.

INTRODUCTION

Business Problem Framing

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.

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. We need to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

Conceptual Background of the Domain Problem

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.

- ### Technical Requirements:
  - Data contains 1460 entries each having 81 variables.
  - Data contains Null values. We 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. We need to handle them accordingly.
  - We have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
  - We need to find important features which affect the price positively or negatively.
  - Two datasets are being provided to us (test.csv, train.csv). We will train on train.csv dataset and predict on test.csv file.

First the basic details of houses should be known .We should be aware of the features which are important for a house before buying it.Basic needs of a house like kitchen bathroom,garage,car parking ,fire place etc should have a minimum quality.Also the year of renovation or construction is also important.They are negatively affected.That is as age increases house price decreases.

- ## Review of Literature

First of all the data is saved in a csv file. Since train and test datas are given separate we should combine them before starting EDA analysis.Then its shape, datatypes, column value counts are all checked to get an outline of the data collected. Null values and correlation between the columns are checked using heatmap. Univariate and bivariate analysis are done for more clarification .skewness of the input columns are checked and resolved. PCA is done to rectify multicollinearity problem and standardscaler is used to resolve the unrealistic values. z score method is used, and outliers are replaced with the median values, so as to lose only very less data. Then train test split is done and checked for the best model and a model with high accuracy score and also relatively high cross validation score is selected as the best model.

- ## Motivation for the Problem Undertaken

The main objective behind doing this project is to make an understanding of the houses and also real estate business that are widely accepted nowadays as a leading business. They also focus primarily quality of the houses,materials used to build and surrounding details etc.Hope this analysis may help housing companies like Surprise housing and also real estate business to acquire more profit.

ANALYTICAL PROBLEM FRAMING

Mathematical/ Analytical Modeling of the Problem

In the describe function we have checked mean, std.deviation, minimum, maximum, 25 percentile, 50 percentile, 75 percentile of each attribute columns.

Mean is the average, median is the central value and mode is the frequency.
Percentile is the value below which the percentage of data falls.

We also use evaluation matrix like mean squared error, mean absolute error,r2 score etc. These can be expressed in mathematical formulas as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Data Sources and their formats

FlipRobo technologies have provided this dataset for detailed analysis which was collected from a US based housing company named surprise housing.. The data collected was in two excel sheets,which are train data and test data with a very detailed description of each columns with it. The data is converted into csv file and loaded in Jupiter notebook first. There are 82 columns and 1460 rows in this dataset. The data was in integer, float and object datatypes. The output column name is ‘saleprice’, which is filled with linear price values. Thus, it is understood that Linear regression ,lasso ridge and elastic net methods should be used for the model prediction.

Data Preprocessing Done

First the important libraries for preprocessing and also csv files for analysis is imported. Then the train and test datas are combined for the EDA analysis .Anew column names source is added to understand whether the data belongs to train or test.Train data has source value 1 and test data has source value 0.Shape and datatypes of columns are checked. There are 1460 rows and 82 columns in our dataset. There are object, integer and float datatypes. The object datatypes should be converted to integer datatypes for the analysis. Unnecessary columns like index and columns with majority null values like pool qc and misc feature are dropped as they provide no necessary information for our analysis. Null values are checked and should be cleared if there is any.

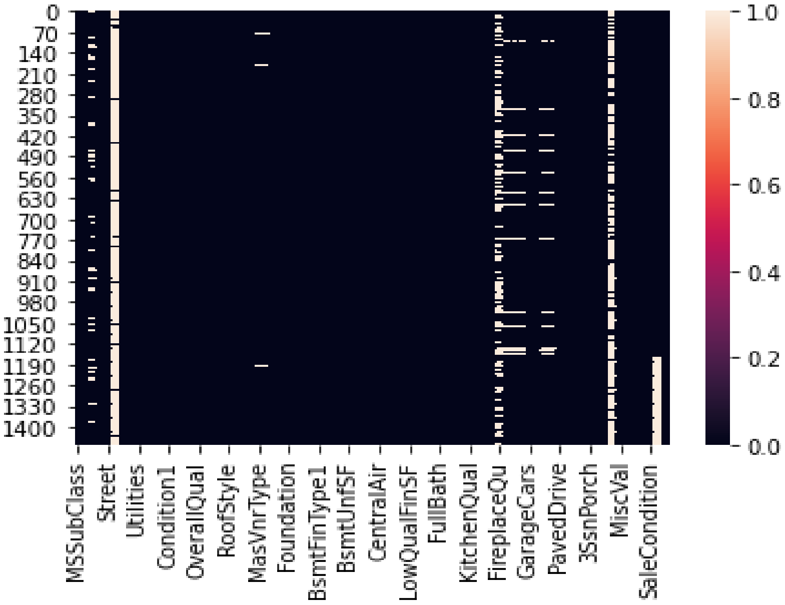


Figure 1:Heat map

The interrupted lines in figure 1 shows the presence of null values.Null values are treated using Simple Imputer.Null values in object datatype columns are filled using most frequent data and null values in integer datatype columns are treated with mean of that column.

Label encoder is used to convert object datatype columns into integer datatypes.This step should be done before the describe function.

Describe functions provide information with minimum,maximum,mean,std.deviation,25th percentile,50th percentile 75th percentile of each column. There is large difference between 75% and maximum for many columns, this represents the presents of outliers. So we must treat this unrealistic datas and outliers for a better results and predicting the best model.

Data Inputs- Logic- Output Relationships

The relation between each column and also relationship of each column with the output column can be represented graphically using a heatmap.From the heatmap we can understand that sale price is having a strong positive relation with overalquality, total basement area, above ground living area and garage area, garage cars etc.

Sale price is having a strong negative relation with MSsubclass, overall condition, basementfin type2,low qual finsf, kitchen above grade, enclosedporch, basementhalfbath and year sold. we can see highly correlated columns other than label column. The highly related columns shows the presence of multicollinearity between feature columns. Multicollinearity can be identified by various methods such as correlation heatmap, pairplot, Variance inflation factor etc. Multicollinearity may lead to low accuracy results and also may affect our model prediction. So multicollinearity must be treated. They can be treated using PCA (principle component analysis) method. Skewness is checked to check whether the data is skewed or not. Highly skewed datas shows the presence of outliers. The threshold value of skewness ranges from -0.55 and +0.55. Skewness can be resolved using log transformation, boxcox method or yeo-johnson method . Here I used yeo-johnson method because it does not strictly needs the input variables to be positive.

Zscore is calculated and instead of removing the outliers I filled outlier values with the median of each column so that we won’t lose much data.

Assumptions under consideration

Sale Price is having a very positive correlation with overall quality, Grliv area, garage cars, garage area, total basements SF, 1st floor SF, full bath ,total rooms above ground ,year built, year remod add. That means customers give more importance to how old the house is total rooms ,car space and garage area.If these values are in an average range then house will get sold easily without much loss.

Sale Price is having a very poor relation with heating QC, garage finish, kitchen quality, basement quality, exterior quality etc. So, we can assume that customers more focus on the area of car space, kitchen, bedrooms, basement etc and not on their finish or material used to built or their overall quality.

Thus we get an idea about the behaviour of customers.They may rebuild or renovate them for their personel uses or they may resell them or may have plans to give it for rent.

Hardware and Software Requirements and Tools Used

I used intel core i3 processor, 4GB RAM and 64 bit operating system as hardware and windows 10, MS excel, MS word and python 3 Jupyter notebook as software for the completion of this project. In jupyter notebook various libraries are also used. They include pandas, numpy, matplotlib , seaborn , imblearn and sklearn.

MODEL/s DEVELOPMENT AND EVALUATION

Identification of possible problem-solving approaches

The major problems we dealt with this dataset are:

i. The train and test datas were given separately,so it have to be combined first using concat function. Then a new column named source is added to understand whether the data belongs to train or test.

ii. There were a lot of null values which have to be treated first.They are treated using simple imputer function.Null values in columns with object datatypes are converted using the most_frequent datas of the column.Null values in columns with integer datatypes are treated with mean of that column.

iii. Label encoder is used to convert object datatype columns to integer datatypes.

iv. It was really hard to find the correlation between columns using heatmap and also with correlation matrix since the data was too long

v. Unrealistic data-We dealt with unrealistic datas in this dataset. They are solved using standard scaler. standard scaler standardises the features by removing the mean and scaling to unit variance

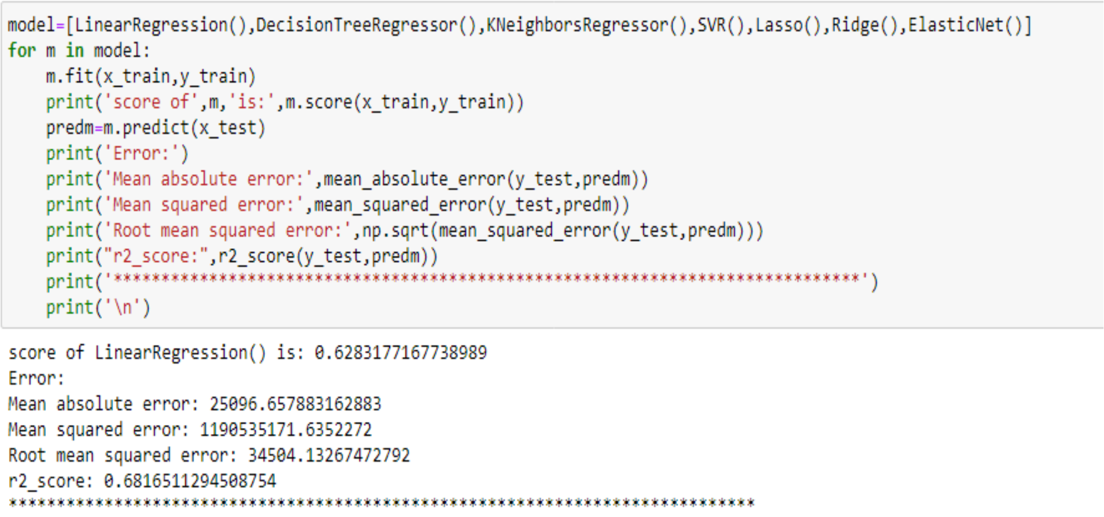
vi. Multicollinearity- multicollinearity refers to the collinearity between the features. Multicollinearity occurs when our model includes multiple factors that are correlated with each others other than with label. It makes more difficult to predict the correct model and also affects the accuracy. They are treated using PCA (principle component analysis) method. This algorithm reduces the no. of columns by removing highly correlated feature columns

Testing of Identified Approaches (Algorithms)

After the train test split of the data many models like linear regression, decision tree regressor, KNeighbor regressor, SVR, Lasso, Ridge elastic net, Random forest regressor, and ada boost regressor are checked. Random forest regressor gives the best r2 score, so the cross-validation score of random forest is checked.

Run and Evaluate selected models

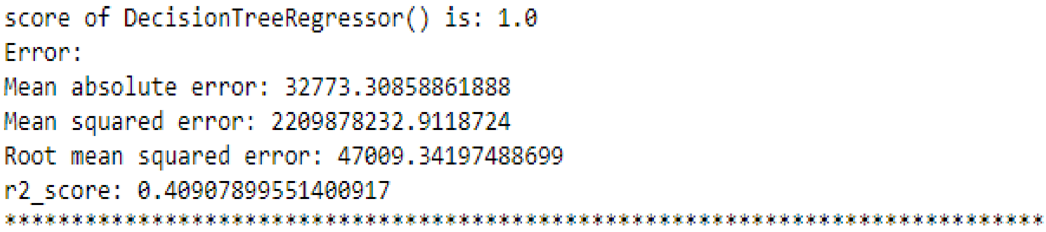
1) Linear regression- Linear regression is a model that assumes linear relation between input variables and single output variable.



```
model=[LinearRegression(),DecisionTreeRegressor(),KNeighborsRegressor(),SVR(),Lasso(),Ridge(),ElasticNet()]
for m in model:
    m.fit(x_train,y_train)
    print('score of',m,'is:',m.score(x_train,y_train))
    predm=m.predict(x_test)
    print('Error:')
    print('Mean absolute error:',mean_absolute_error(y_test,predm))
    print('Mean squared error:',mean_squared_error(y_test,predm))
    print('Root mean squared error:',np.sqrt(mean_squared_error(y_test,predm)))
    print("r2_score:",r2_score(y_test,predm))
    print('*****************************************************************************')
    print('\n')
```

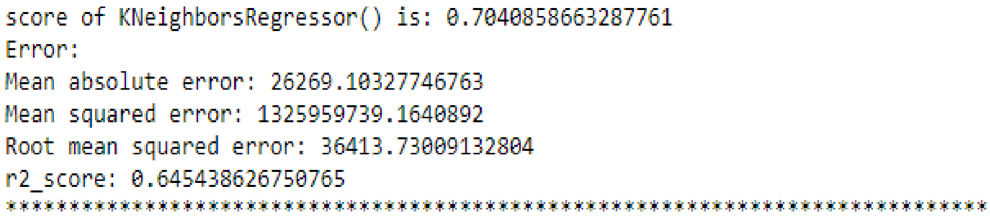
```
score of LinearRegression() is: 0.6283177167738989
Error:
Mean absolute error: 25096.657883162883
Mean squared error: 1190535171.6352272
Root mean squared error: 34504.13267472792
r2_score: 0.6816511294508754
*****************************************************************************
```

2) Decision tree regressor- Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuos output.



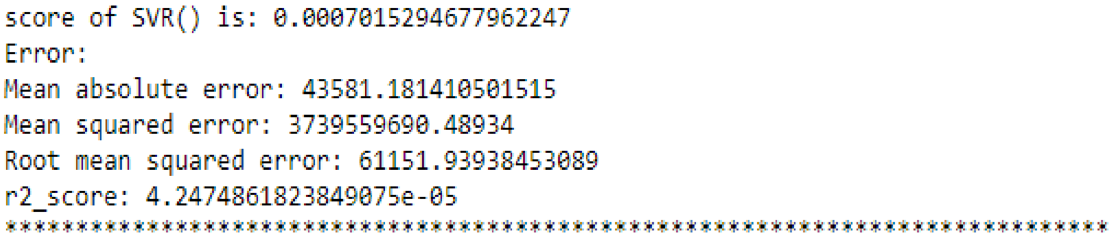
```
score of DecisionTreeRegressor() is: 1.0
Error:
Mean absolute error: 32773.30858861888
Mean squared error: 2209878232.9118724
Root mean squared error: 47009.34197488699
r2_score: 0.40907899551400917
*****************************************************************************
```

3) KNeighbor regressor- It is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure



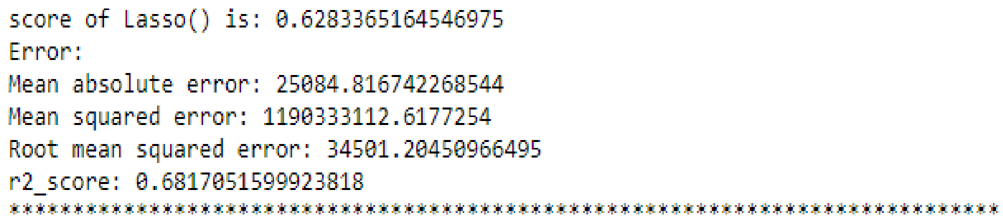
```
score of KNeighborsRegressor() is: 0.7040858663287761
Error:
Mean absolute error: 26269.10327746763
Mean squared error: 1325959739.1640892
Root mean squared error: 36413.73009132804
r2_score: 0.645438626750765
*****************************************************************************
```

4) SVR-support vector regressor is a statistical method that examines the linear relationship between two continouse variables



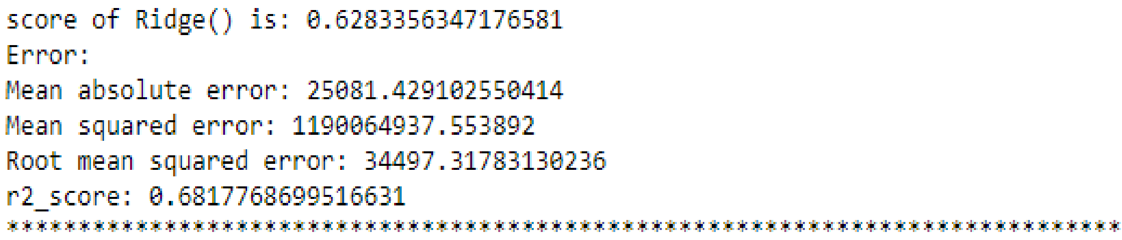
```
score of SVR() is: 0.0007015294677962247
Error:
Mean absolute error: 43581.181410501515
Mean squared error: 3739559690.48934
Root mean squared error: 61151.93938453089
r2_score: 4.247486182384907 5e-05
*****************************************************************************
```

5) Lasso-It is a regularization technique used over regression methods for more accurate prediction.



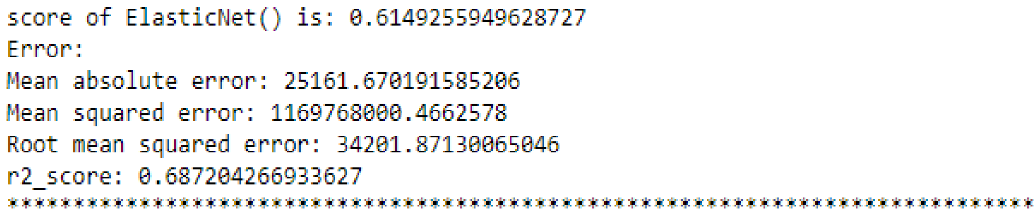
```
score of Lasso() is: 0.6283365164546975
Error:
Mean absolute error: 25084.816742268544
Mean squared error: 1190333112.6177254
Root mean squared error: 34501.20450966495
r2_score: 0.6817051599923818
******************************************************************************
```

6) Ridge-It is a technique for analyzing mutltiple regression data that suffer from multicollinearity



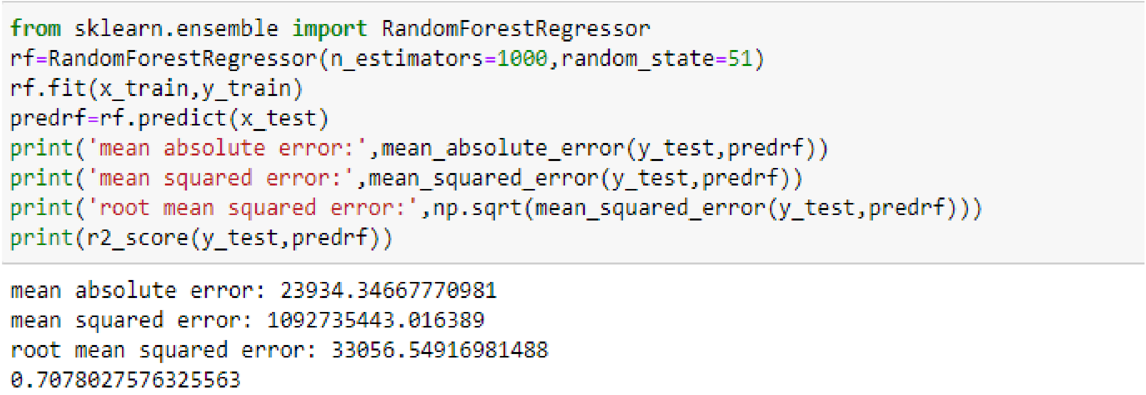
```
score of Ridge() is: 0.6283356347176581
Error:
Mean absolute error: 25081.429102550414
Mean squared error: 1190064937.553892
Root mean squared error: 34497.31783130236
r2_score: 0.6817768699516631
******************************************************************************
```

7) Elastic net-It combines the power of both lasso and ridge regression into one algorithm



```
score of ElasticNet() is: 0.6149255949628727
Error:
Mean absolute error: 25161.670191585206
Mean squared error: 1169768000.4662578
Root mean squared error: 34201.87130065046
r2_score: 0.687204266933627
******************************************************************************
```

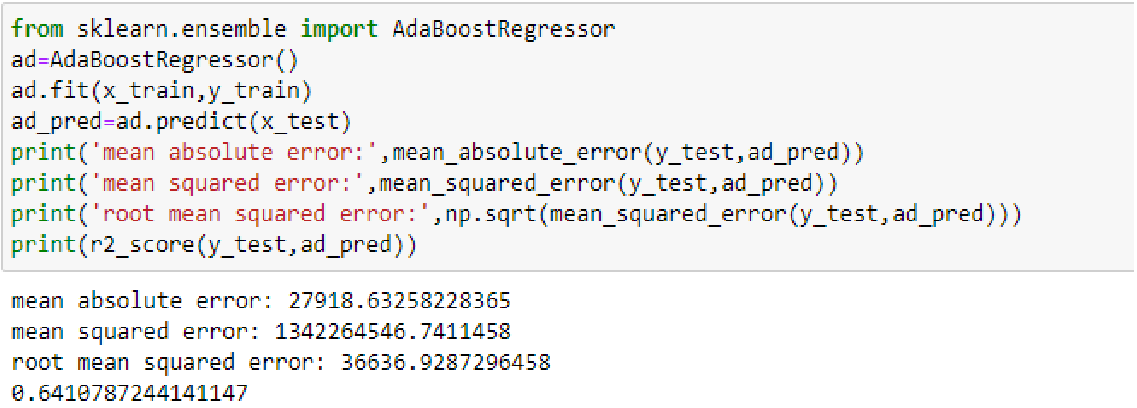
8) Random Forest regressor-They are ensemble technique capable of performing both regression and classification task with the use of multiple decision trees and a technique called bagging



```
from sklearn.ensemble import RandomForestRegressor
rf=RandomForestRegressor(n_estimators=1000,random_state=51)
rf.fit(x_train,y_train)
predrf=rf.predict(x_test)
print('mean absolute error:',mean_absolute_error(y_test,predrf))
print('mean squared error:',mean_squared_error(y_test,predrf))
print('root mean squared error:',np.sqrt(mean_squared_error(y_test,predrf)))
print(r2_score(y_test,predrf))
```

```
mean absolute error: 23934.34667770981
mean squared error: 1092735443.016389
root mean squared error: 33056.54916981488
0.7078027576325563
```

9) Ada Boost Regressor-It is a meta estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset.

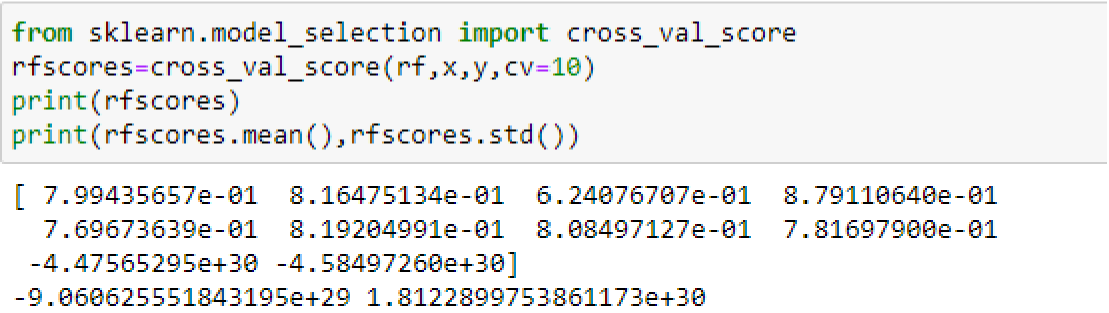


```
from sklearn.ensemble import AdaBoostRegressor
ad=AdaBoostRegressor()
ad.fit(x_train,y_train)
ad_pred=ad.predict(x_test)
print('mean absolute error:',mean_absolute_error(y_test,ad_pred))
print('mean squared error:',mean_squared_error(y_test,ad_pred))
print('root mean squared error:',np.sqrt(mean_squared_error(y_test,ad_pred)))
print(r2_score(y_test,ad_pred))
```

```
mean absolute error: 27918.63258228365
mean squared error: 1342264546.7411458
root mean squared error: 36636.9287296458
0.6410787244141147
```

Key Metrics for success in solving problem under consideration

Random forest regressor is giving a maximum r2 score of 0.70,so its cross validation score is checked .



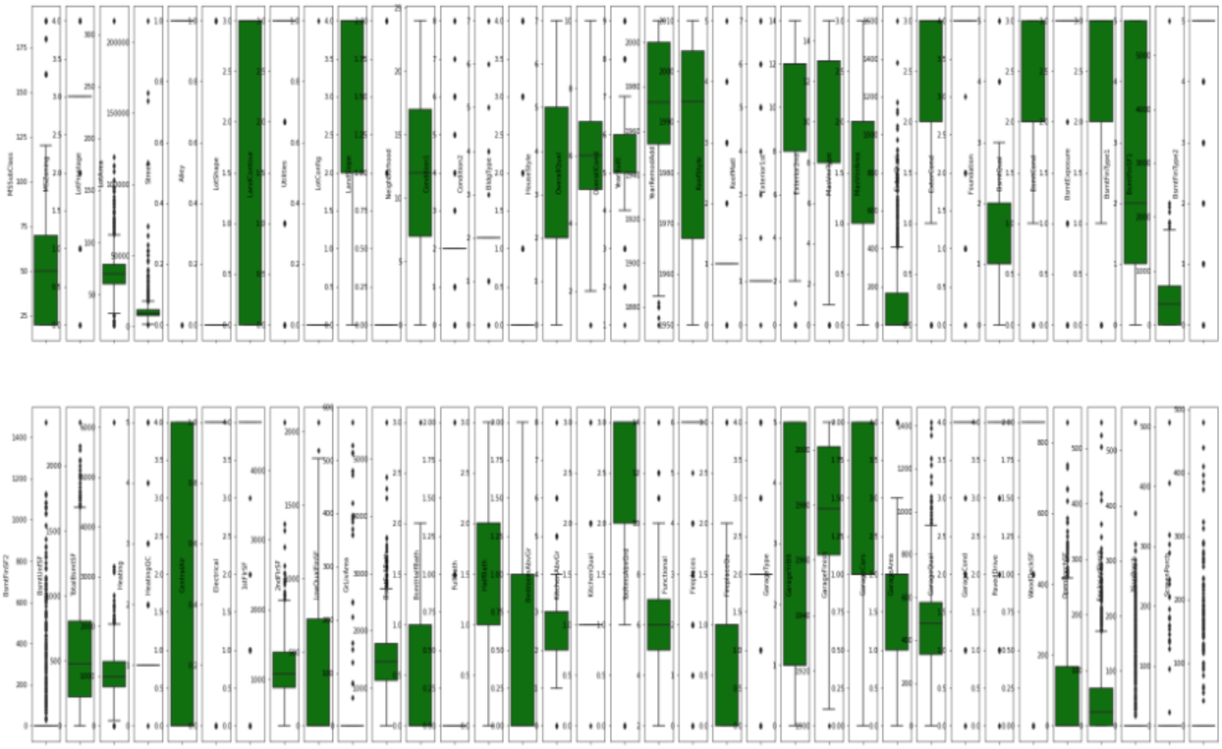
```
from sklearn.model_selection import cross_val_score
rfscores=cross_val_score(rf,x,y,cv=10)
print(rfscores)
print(rfscores.mean(),rfscores.std())
```

```
[ 7.99435657e-01  8.16475134e-01  6.24076707e-01  8.79110640e-01
  7.69673639e-01  8.19204991e-01  8.08497127e-01  7.81697900e-01
 -4.47565295e+30 -4.58497260e+30]
-9.060625551843195e+29 1.8122899753861173e+30
```

Since the random forest regressor gives the best result,it is saved as my best model in pickle format.

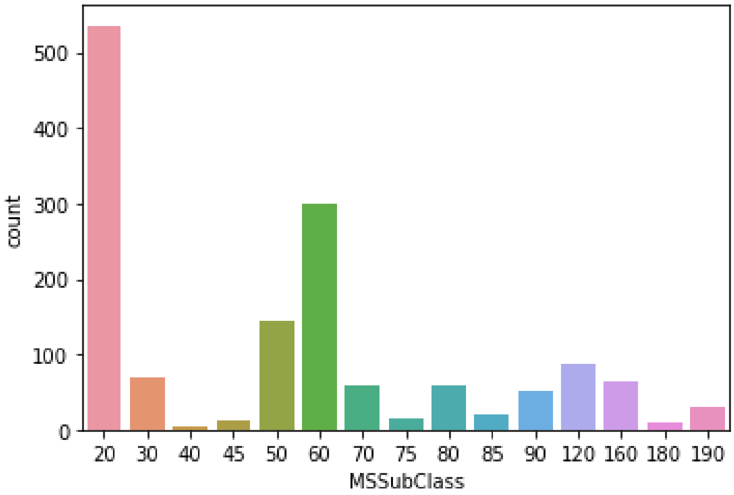
VISUALIZATIONS

1) Boxplot- Boxplots are the best methods to check for the presence of outliers

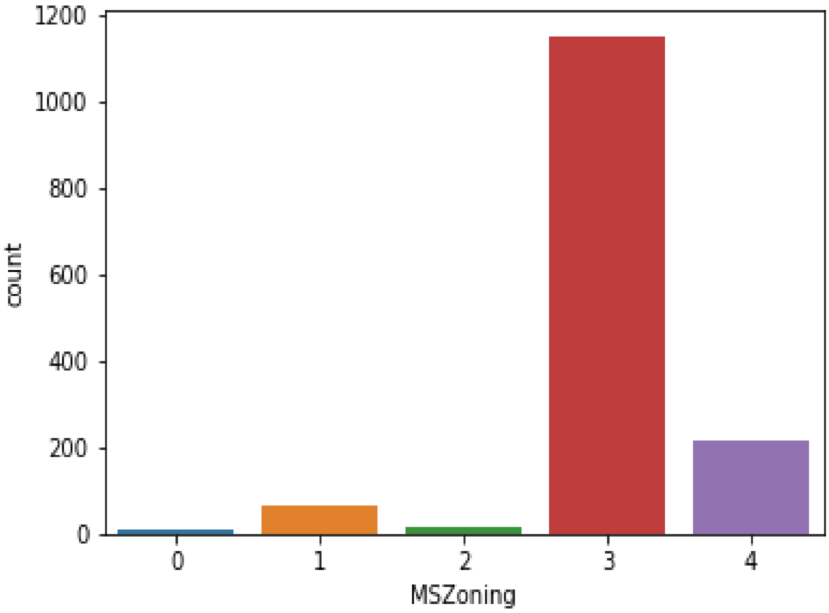


Here the black dots above and below the green coloured boxes represents outliers.

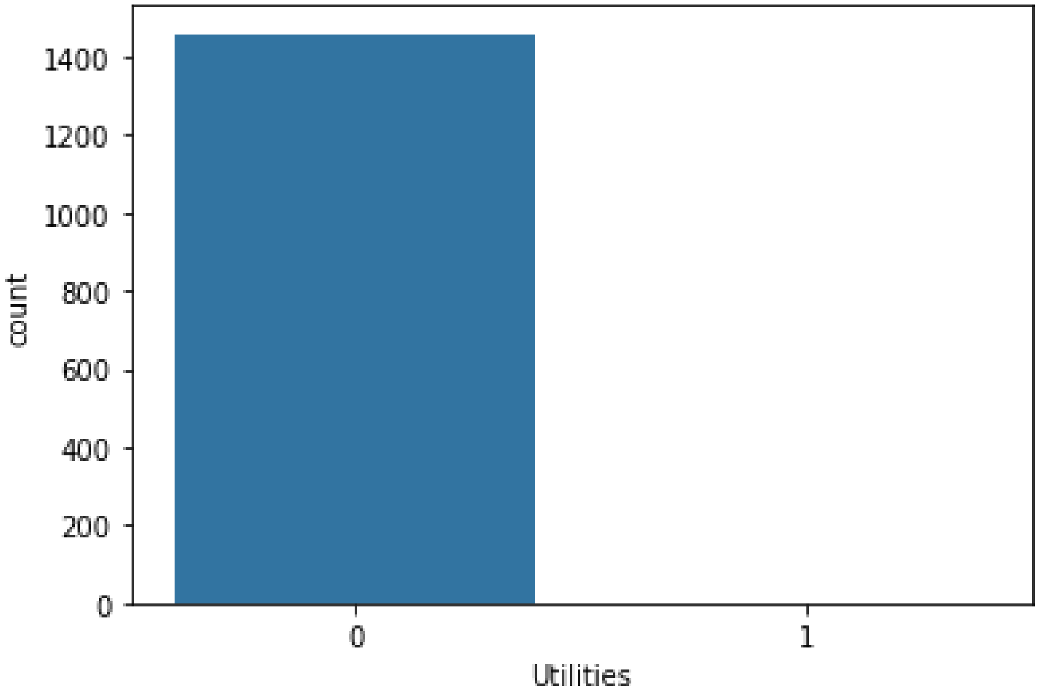
2) Univariate Analysis



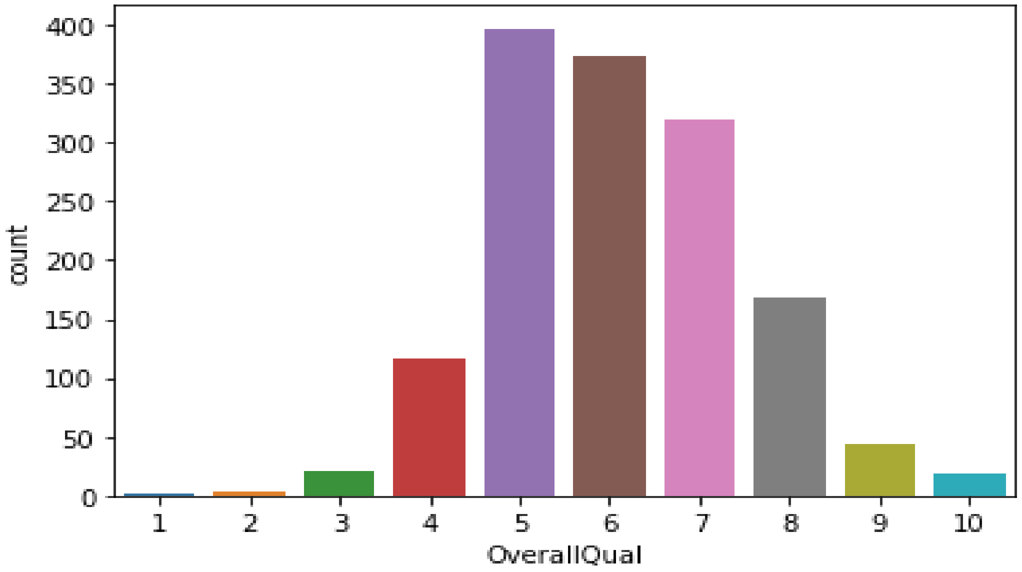
Most of the houses (above 500) comes under 20 category,ie,- 1story 1946 and newer all styles. nearly 300 houses comes under 60- 2story 1946 and newer and all others comes below 200 count least count is for 40- 1story finished attic all ages.



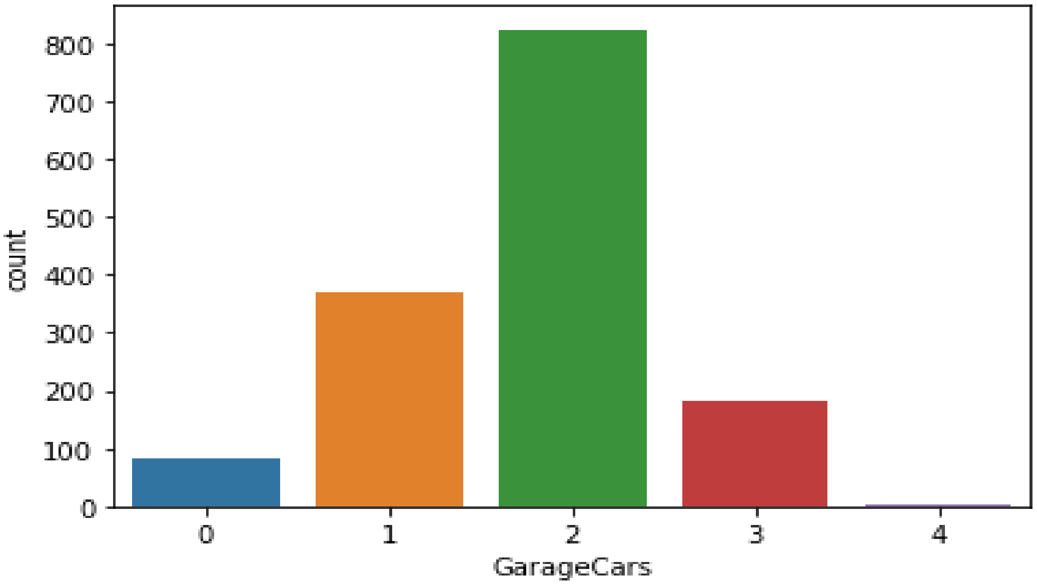
Here majority (nearly 1200 counts) comes under relatively low density category.Nearly 200 houses are in relatively medium density area.



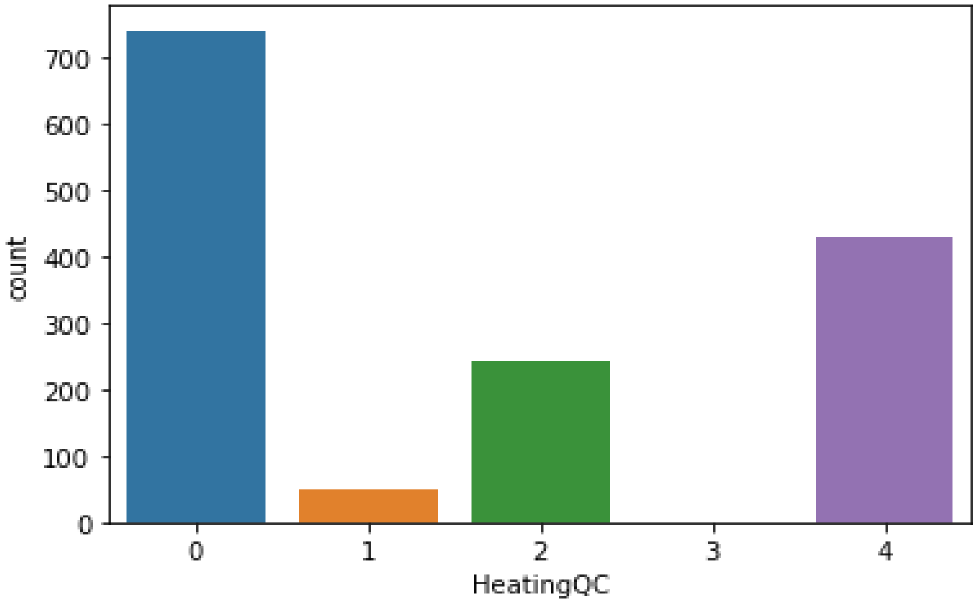
All public utilities are available in all houses



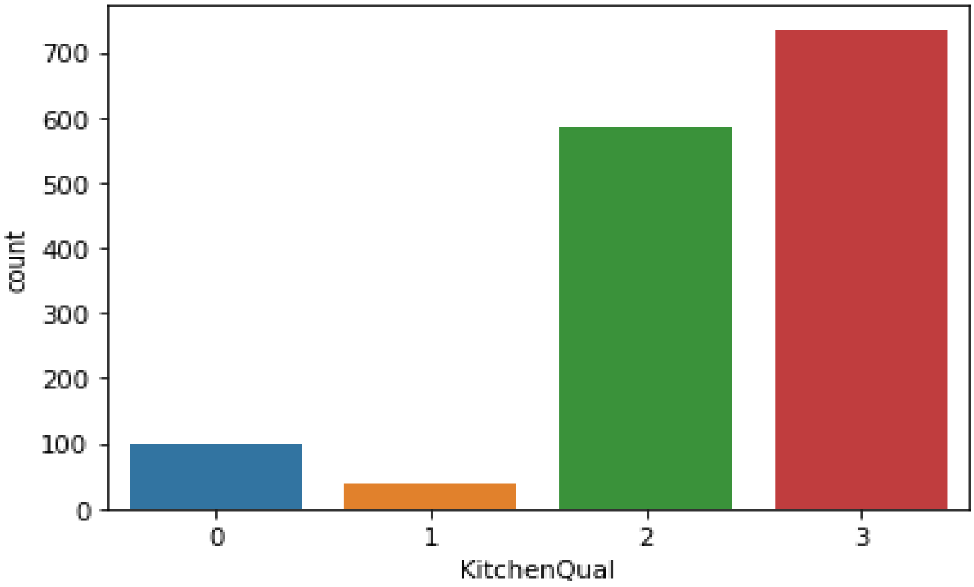
Most of the houses comes under categories 5-average, 6-above average, 7-good, 8-very good and only few comes under.
1-very poor, 2-poor, 3-fair, 4-below average



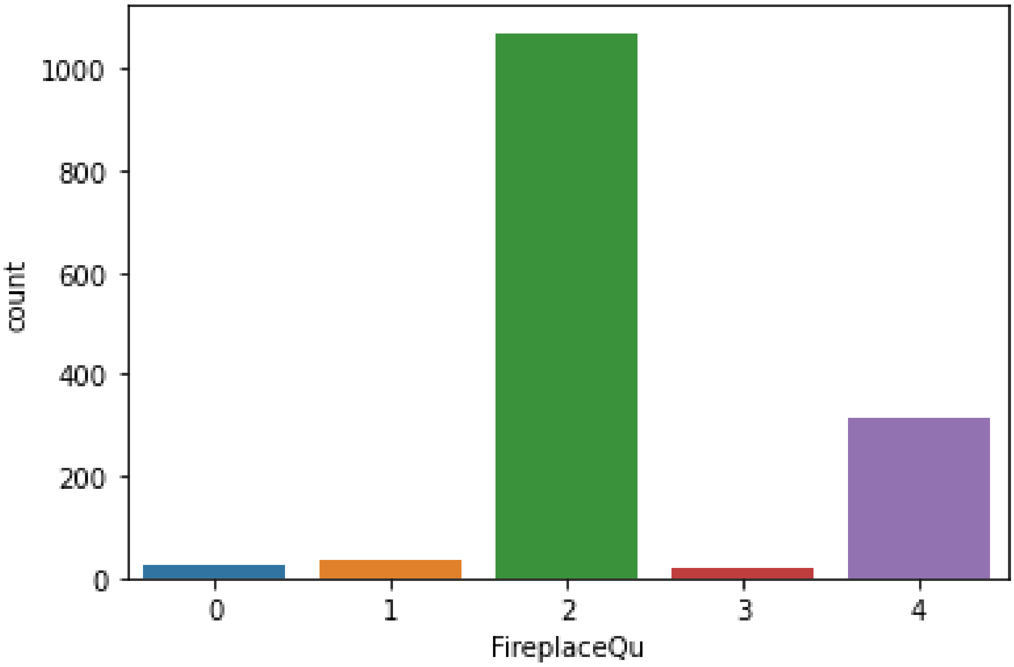
Most of the houses have an average car capacity garage



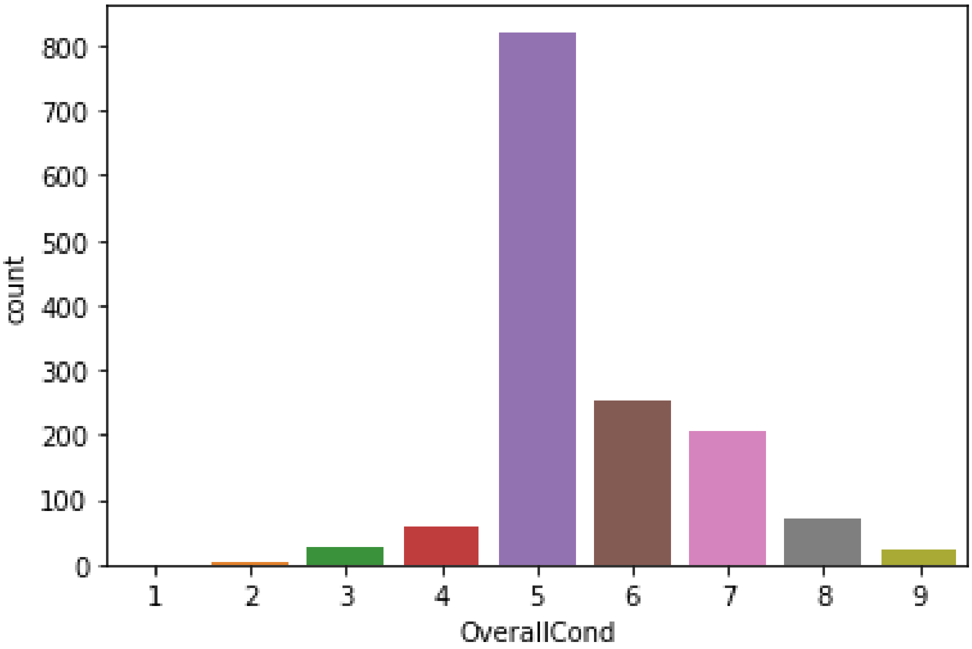
Above 700 houses have excellent heating quality and nearly 500 houses have an average heating quality.



Around 700 houses have an average kitchen quality and nearly 600 have a good kitchen quality

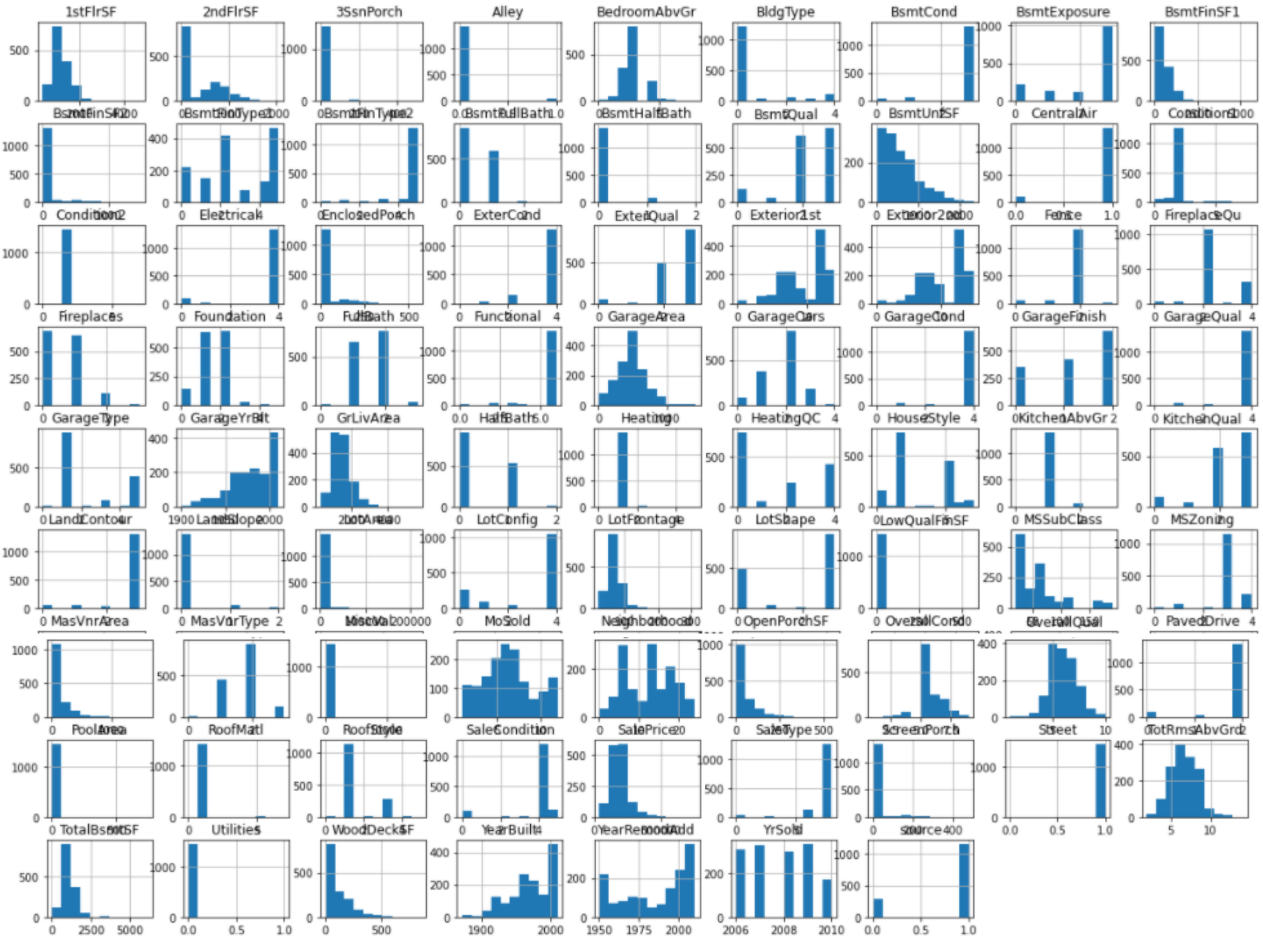


Above 1000 houses have an average fire quality place



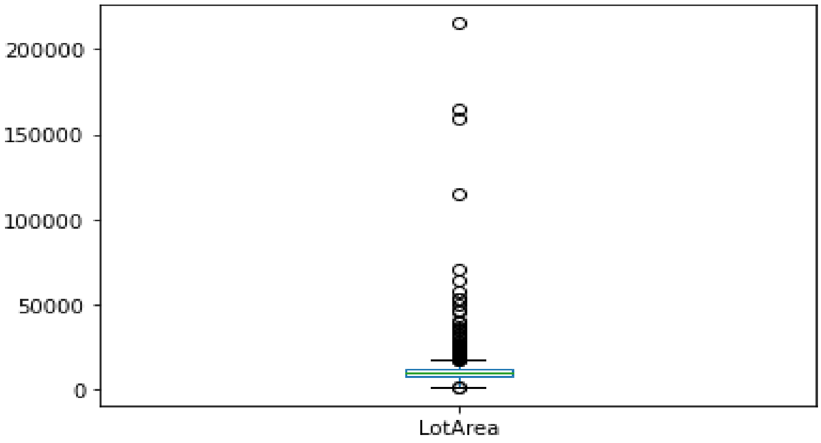
Nearly 800 houses have average overall condition .around 250 have above average condition. Nearly 200 have good condition and below 100 houses have an excellent overall condition.

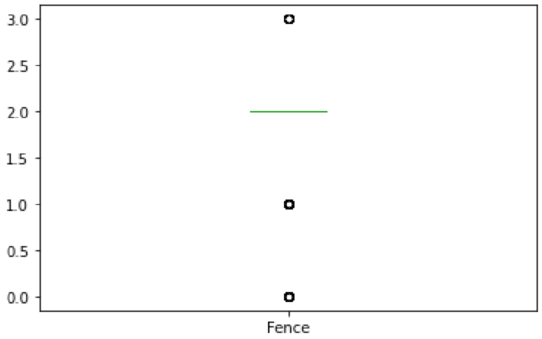
3) Bivariate Analysis

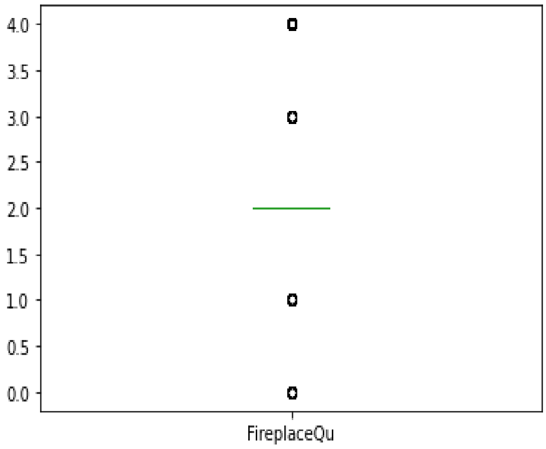


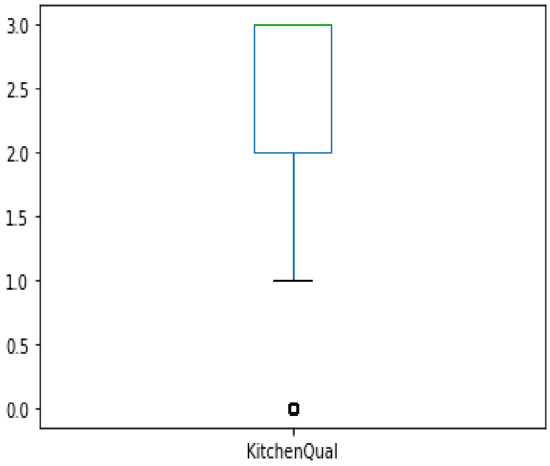
This is a histogram showing number and distribution of each feature columns.

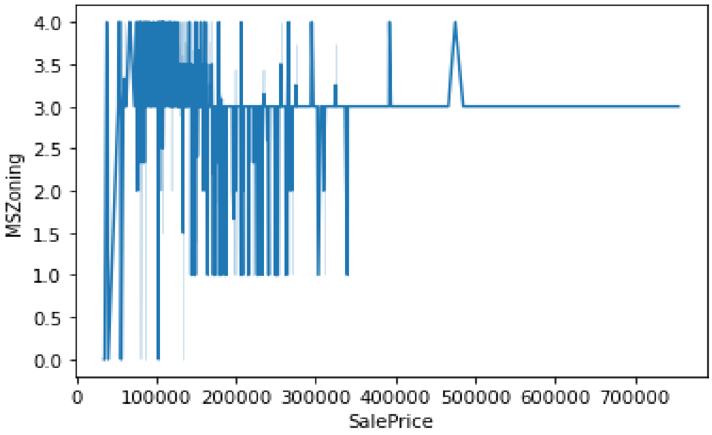
Boxplots are the best methods to detect the presence of outliers. Below are some of the boxplots to see the presence of outliers in our datasets.

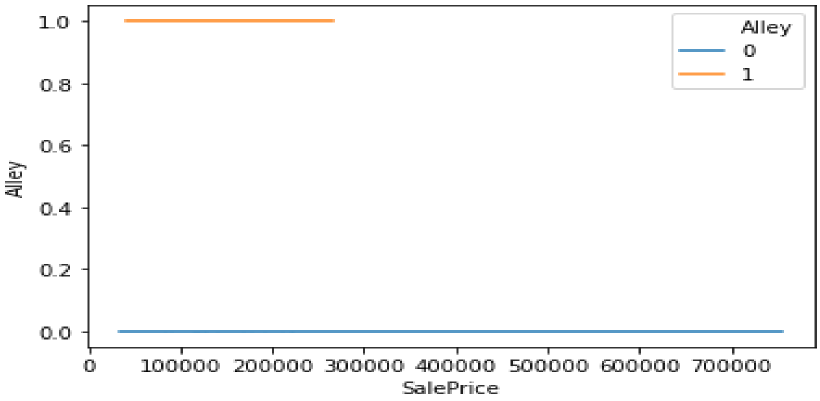



3.0
2.5
2.0
1.5
1.0
0.5
0.0
Fence

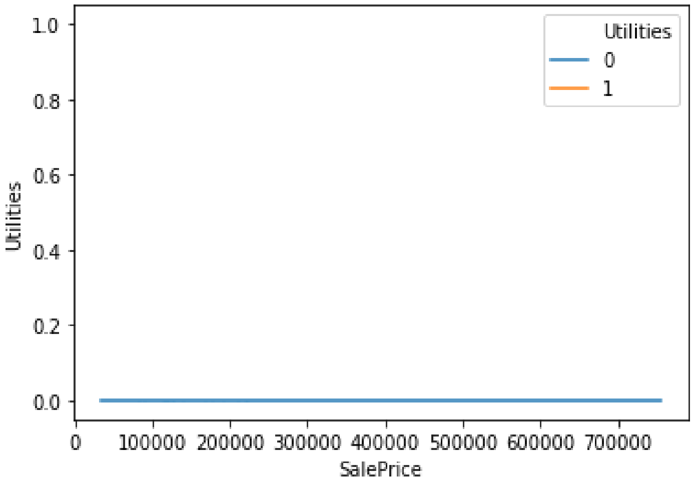

4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
FireplaceQu

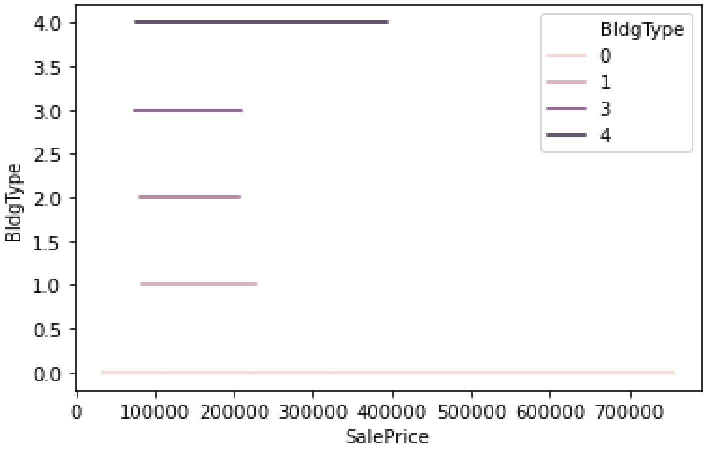

3.0
2.5
2.0
1.5
1.0
0.5
0.0
KitchenQual


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
MSZoning
0
100000
200000
300000
400000
500000
600000
700000
SalePrice

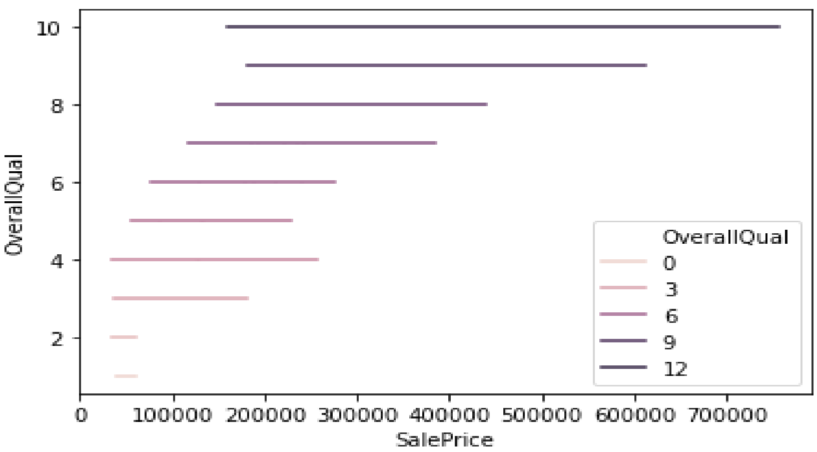


Saleprice is higher for alley 0 and less for alley 1.Sale price reaches above 7 lakhs if alley is 0 and sale price maximum for alley 1 is around 3 lakhs

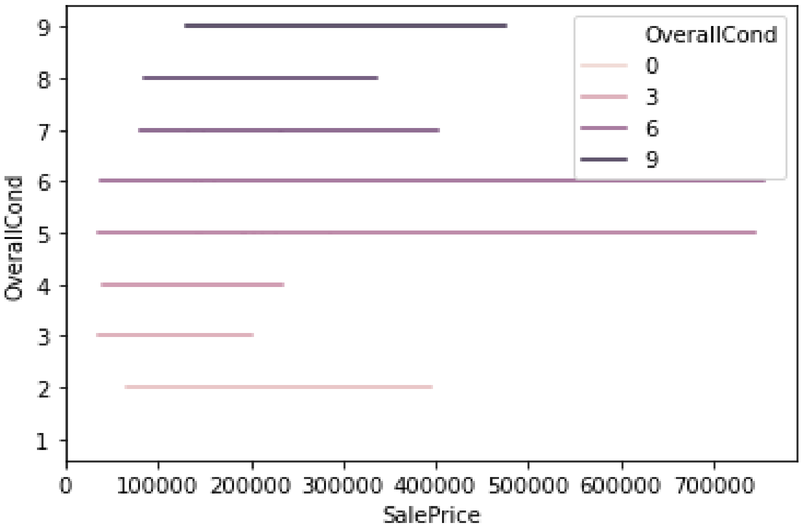




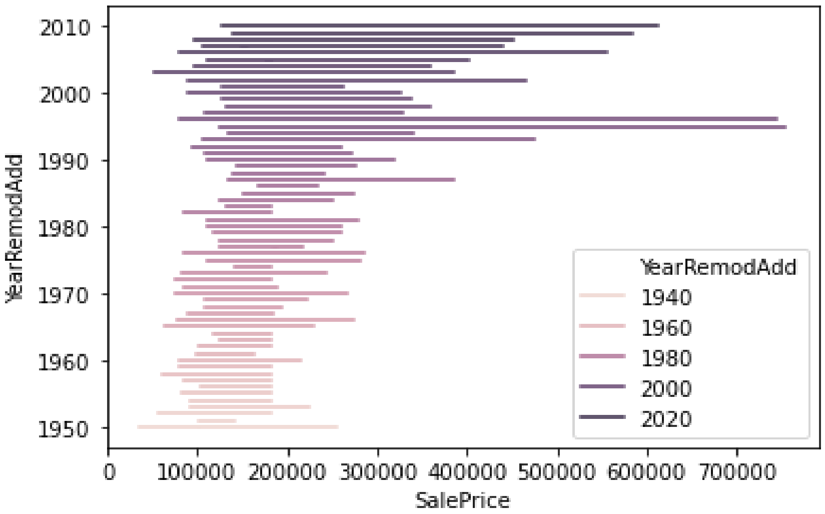
Sale price varies with building type .It varies with single family,two family,townhouse inside and townhouse outside.



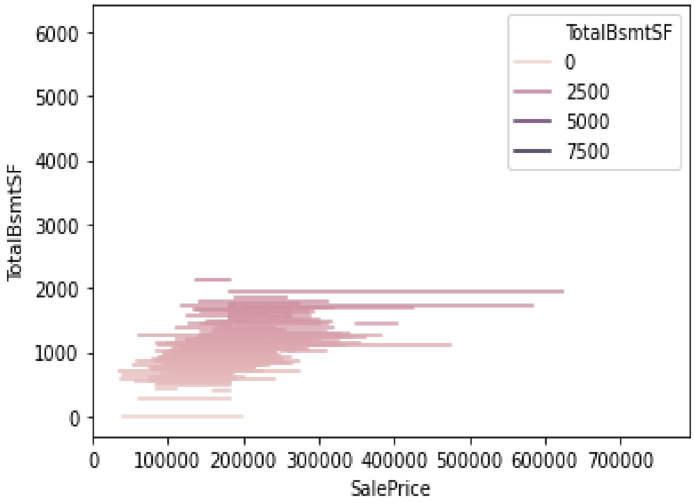
OverallQual
SalePrice
OverallQual
0
3
6
9
12



OverallCond
SalePrice
OverallCond
0
3
6
9



YearRemodAdd
SalePrice
YearRemodAdd
1940
1960
1980
2000
2020



Interpretation of the Results

From this analysis we can understand that sale price depends directly on many of the features.Some of them are kitchen area,basement area,garage area,car parking capacity ,no.of rooms etc..And also many of the features doesnot effect the sale price .They are kitchen quality,overall quality,fire place quality ,year built etc..This means that customers mainly look for the space and comfortness for their future use and renovation.So they may be buying the houses for their personal use,so that they can renovate as per their wish ,if the house is spacious.Also demand is high for two family houses,town area ,good street or road access etc.

Or they may be planning to resell the houses,or give them for rent or other uses otherthan personel.That may be the reason they don’t focus on quality of material and finish.

CONCLUSION

Key Findings and Conclusions of the Study

There are a lot of datas available in this dataset, Many of them make sense and some of them are not so important or they don't make any sense to our analysis. They can be identified from our correlation map. Also we used PCA technique to solve the problems created by this .From the whole analysis we understood that customers look for houses built not so early.They need more area .They are ready for renovation .They don't focus much on quality of materials used and also on finish.

But they look for good background like town area,paved road access, utilities available etc.
Thus we can earn more profit in future if we concentrate more on town area and a good society with spacious houses,and also good road access and availability of emergency services.

Learning Outcomes of the Study in respect of Data Science

I have tried many models like linear regression,decision tree regressor, K neighbour regressor, lasso,ridge,elastic net,random forest and ada boost regressor.From these Random forest regressor gives a maximum r2 score when compared to others.So cross validation score of Random forest is checked and I selected it as my best model.So Random forest model is saved in pickle format.

Limitations of this work and Scope for Future Work

Data visualisation took a lot of time and so was unable to visualise each column and check relation between the columns as planned.

There were also a lot of problems during each phase of the project, they were all resolved by searching webpages and also with the help of mentor and datatrained support team.

Some of the url's which made possible for me to do this project are listed in Reference section.

REFERENCE

1. https://cxl.com/blog/outliers/

2. https://www.google.com/search?q=multicollinearity+in+regression&oq=muticollinearity&aqs=chrome.2.69i57j0i10i433j0i10l6.7595j0j7&sourceid=chrome&ie=UTF-8

3. https://www.google.com/search?sxsrf=ALeKk02HmEy5i9YItlkcqpojW8oOavLcIg%3A1613537096613&ei=SJ8sYJD3JKyc4-EPkJS4oAI&q=yeo+johnson+transformation&oq=YEO&gs_lcp=Cgdnd3Mtd2l6EAEYATIECCMQJzIFCAAQkQIyBwgAELEDEEMyCgguELEDEIMBEEMyBwguELEDEEMyBAgAEEMyBwguELEDEEMyBwguELEDEEMyBQguELEDMgUIABCxAzoICAAQsQMQgwE6AggAULabT1ibpE9gsMFPaABwAngAgAGXAogB-AWSAQUwLjMuMZgBAKABAaoBB2d3cy13aXrAAQE&sclient=gws-wiz