School of Computing, Engineering and Digital Technologies

Department of Computing and Games

Teesside University

Middlesbrough TS1 3BA

**HOUSE PRICE PREDICTION**

**USING MACHINE LEARNING ALGORITHAMS**

Analysis, Design and Implementation Report

Submitted in partial requirements for the Degree of MSc DATA SCIENCE

**DATE: 09/05/2023**

**NAVEEN KUMAR PALLABOTHULA**

**STUDENT ID: W9543507**

**Supervisor: Karayaneva, Yordanka**

# ACKNOWLEDGEMENTS

The final masters project titled “**HOUSE PRICE PREDICTION USING MACHINE LEARNING ALGORITHAMS**”. I would like to thanks for the continuous guidance from my parents was provide during the entire course time and mainly during the completion of this final project.

Mainly, I would like to acknowledge my supervisor **Karayaneva , Yordanka.** She has given priceless support to complete the final master project. Without the guidance and support of her this project would never become reality. I heartly extend my thanks to people who supported and give guidance for completing the dissertation.

Table of Contents

[ACKNOWLEDGEMENTS ii](#_Toc134539997)

[Table of Contents iii](#_Toc134539998)

[FIGURE TABLE iv](#_Toc134539999)

[ABSTRACT v](#_Toc134540000)

[INTRODUCTION 1](#_Toc134540001)

[LITERATURE REVIEW 1](#_Toc134540002)

[METHODOLOGY 3](#_Toc134540003)

[1.BUSINESS PROBLEM 3](#_Toc134540004)

[2.DATA COLLECTION 4](#_Toc134540005)

[OBJECTIVES 4](#_Toc134540006)

[3.TOOLS SETUP FOR PROJECT: 4](#_Toc134540007)

[4.LOADING DATASET 4](#_Toc134540008)

[5.DATA PRE-PROCESSING: 5](#_Toc134540009)

[6.MAKE DATA READY FOR TRAINING 23](#_Toc134540010)

[7.MACHINE LEARNING MODELs 24](#_Toc134540011)

[1.SUPPORT VECTOR REGRESSION 24](#_Toc134540012)

[2.RANDOM FOREST 25](#_Toc134540013)

[3.RIDGE REGRESSION 26](#_Toc134540014)

[4.GRADIENT BOOST REGRESSION 26](#_Toc134540015)

[5.LINEAR REGRESSION 26](#_Toc134540016)

[6.LASSO REGRESSION: 27](#_Toc134540017)

[8.METRICS 27](#_Toc134540018)

[MEAN SQUARED ERROR (MSE): 28](#_Toc134540019)

[ROOT MEAN SQURED ERROR (RMSE): 28](#_Toc134540020)

[R2 SQUARED ERROR 28](#_Toc134540021)

[MEAN ABSOLUTE ERROR (MAE) 29](#_Toc134540022)

[9. RESULTS AND EVALUATION: 29](#_Toc134540023)

[10. CONCLUSION 32](#_Toc134540024)

[11.FUTURE WORK 32](#_Toc134540025)

[12. REFERENCES 32](#_Toc134540026)

[APPENDIX 33](#_Toc134540027)

# FIGURE TABLE

[Figure 1 CORRELATION MATRIX 12](#_Toc134492153)

[Figure 2 NEW CORRELATION MATRIX 13](#_Toc134492154)

[Figure 3 PAIR-WISE SCATTER PLOT 14](#_Toc134492155)

[Figure 4 CONCAT OF SALESPRICE AND OVERALLQUAL 15](#_Toc134492156)

[Figure 5 CONCAT OF SALES PRICE AND YR BUILT 17](#_Toc134492157)

[Figure 6 CONCAT OF SALEPRICE AND 1ST FLOOR SF 18](#_Toc134492158)

[Figure 7 CONCST OF SALEPRICE AND TOTAL BSMT SF 19](#_Toc134492159)

[Figure 8 CODE FOR UNIVARIATE OUTLIER ANLYSIS TO TARGET VARIABLE 20](#_Toc134492160)

[Figure 9 MULTIVARIATE OUTLIERS 21](#_Toc134492161)

[Figure 10 MULTIVARIATE OUTLIERS SALEPRICE AND 1ST FLR SF 21](#_Toc134492162)

[Figure 11 CODE FOR DELETING OUTLIERS 22](#_Toc134492163)

[Figure 12 DELETING OUTLIERS OF SALEPRICE AND 1ST FLR SF 22](#_Toc134492164)

[Figure 13 CODE TO CHECK THE NORMAL PROBABILITY OF THE SALEPRICE 23](#_Toc134492165)

[Figure 14 HISTOGRAM PLOT OF SALEPRICE 23](#_Toc134492166)

[Figure 15 PROBABILITY PLOT O SALEPRICE 24](#_Toc134492167)

[Figure 16 CODE TO SKEWNESS THE SALEPRICE 24](#_Toc134492168)

[Figure 17 LOG TRANSFORMATION TO TARGET VARIABLE 25](#_Toc134492169)

[Figure 18 SKEWNESS OF SALEPRICE 25](#_Toc134492170)

[Figure 19 CODE 25](#_Toc134492171)

[Figure 20 VISUALISATION AFTER THE LOG TRANSFORMATION 26](#_Toc134492172)

[Figure 21 CODE TO FIND MISSING VALUES 27](#_Toc134492173)

[Figure 22 TRAINING AND TESTING THE DATA 30](#_Toc134492174)

# ABSTRACT

The accurate prediction of house price prediction plays an important role for buying or selling the house by the individual or to the organisation. It helps for the real estate, finance companies and House planning construction agencies to predict we are using the AMES HOUSING DATSET.

In this study the nonlinear relationships are done between the features and the saleprice. Detecting the outliers and making treatment to remove outliers, handled with the missing data and doing advanced feature engineering.

In this study of house price prediction, we approached a systematic route to use regression models to predict accuracy. The models are linear regression model, support vector machine, Ridge regression, Lasso Regression, Gradient boost Regression and Random Forest and the model is evaluated by the regression metrics like Mean squared error, Root mean squared error, R2 Square metric and mean absolute error.

We have found that by using base model accuracy we are predicting the accurate accuracy value by LASSO REGRESSION with **0.0447** accuracy value and the SUPPORT VECTOR REGRESSION with **0.929** accuracy value. By using MSE metrics we found Gradient boost regression model gives the low mse score **0.01789** and predicted good accuracy, By using RMSE metrics we found Gradient boost regression model gives the low rmse score 0.12and predicted good accuracy, By using R2 square metric we found Ridge Regression model gives the low R2 score **0.796** and predicted good accuracy, by using MAE metric we found Gradient boost regression model gives the low mae score **0.08** and predicted good accuracy.

# INTRODUCTION

House price prediction is a demanding in real estate business because it helps to buy a house, sell a house, and smart money in properties. It can help buyers to buy a house and sellers to sell house at a good price, bargaining for better deals, and avoid underselling or overpaying to the properties. Accurate house price prediction can provide good rate in the present market. The prediction can be done by considering these three factors including place, physical conditions and concept. House price prediction using machine learning is important research thought for that it can help people to make to exact strategies to buy the house. Additionally, house rice prediction will help more to the real estate business industry to invest money in properties, analysing the market trends, and helps to give some guidance to their customers.it gives accurate valuable price to the properties.

Now a days the real estate organisation is doing the market strategies to make the house price high. Because of the real estate agencies and organisation, the house price prediction are obstacles by the some of the features. Some organisations are investing more in the houses and selling the houses in very high prices. (S-J-LIN et al., 1999). Predicting the house price accurately will help the individual to buy or sell or invest in the house property. By predicting the accurate house price so that it helps to government to control the house prices fluctuation. The impact of the house price prediction is drastic change from the house to house or area to area, it mostly depends on the square feet of the house. sales price is linear to square ft of the house if the square foot of the house increases sales price will be increase automatically. Some of the researchers are dividing the datset into multiple datsets and comparing the features each other and predicting some of the regression models to predict. |How ever we used the dataset the accuracy will same for most of the cases.(Adai et al., 1996)

# LITERATURE REVIEW

literature review is based on the published papers and articles with open access articles, and a peer-reviewed publications from IEEE Xplore, Scopus, ACM DIGITAL LIBRARY and Google Scholar. In any country, it is very common for a people to do research on the house they need before they purchase.

House price prediction is trending topic in the 21st century because of increasing in real estate business and growing importance in property investment. Some of the Researchers developed accurate models to find the house price prediction. In these literature review, we will discuss some of the best research studies which have been conducted on the house price prediction. The authors said by using, Gradient Boost regression, random forest regression, Support vector Regression they have predicated the accuracy of the house price (A Varma , A Sarma, S Doshi and R Nair et al.., 2018).

According to (Z. ZHANG etal., 2021) to solve the house price prediction he used the random forest to get the prediction and to solve the problems and also he used to other machine learning models to predict the price. He can clearly says random forest regression is used to achieve the best results and accuracy for house price prediction. He has done some data pre-processing techniques the he applies the random forest regressor to predict the accuracy. He also found that the house price prediction is also depends on the population near to that area, education like schools or colleges or universities, malls, etc., We can use random forest regressor for the house price prediction to Ames Housing Dataset.

According to (ZiyueYan, Lu zong et al.,2020) to solve the house price prediction they have used beijing dataset to find the house price prediction. They had both developed the machine learning algorithms and the hedonic regression. The author has don some exploratory data analysis and feature engineering to the dataset and then h used the machine learning models that are lasso regression, linear regression, random forest regression, xgboost and other machine learning models. they also proved that XGBoost regressor is getting the most accuracy. From these article I am using some of the machine learning algorithms in my project. We can use lasso regression, ridge regression, random forest regression to our house price prediction to predict the accuracy.

The authors (A. P. Singh, K. Rastogi and S. Rajpoot et al., 2021) had developed the five machine learning algorithms for the designing the system model. The author has done some feature engineering, data pre-processing, fitting the model. They have used the Decision tree, support vector machine, KNN, Random Forest, linear regression to predict 90% accuracy. the authors are very clear that these models are best for the accuracy. So, from this paper I am some models to add up in my paper for the predictions. I am trying to use the models not the content. The authors says that the accuracy value mostly depends on the features taking from the dataset and how we are splitting the dataset.

By carefully observing into the research papers most of the papers are discussing about the regression models. Regression models give the best accuracy than any other model predictions or classifications. The author(Y.FANG, T.LI, H.ZHAO et al., 2022) in the research paper discussed about the prediction of the house price by random forest model. He correlated the features in the dataset to see most correlated features in order prediction. In order to achieve the best house price prediction, he created a model which it predicts the price. And then he compared the accuracy by the other regressions model like support vector regression, decision tree, he also evaluated the models by using root mean squared error metric to goodness it fit. From these articles I would like use support vector regression model and to evaluate the prediction I want to use RMSE metric.

The author (B. Almaslukh et al., 2020) in these articles explained about the house price prediction. The author is trying to predict the house prices for estates agencies and the organisation. The author says that predicting the house price using the predictive model is most challenge thing in the present generation. to solve this problem, he has taken the benchmark dataset to solve the house price prediction. The author used gradient boost regression model to predict the house price prediction and he used to metrics to evaluate the accuracy. He found that gradient boost regressor is getting the lowest result when compared to others. I would like to add gradient boost regressor model in my project to check the prediction accuracy.

The author(Z. Li etal., 2021) in this paper studied about the house price prediction based on machine learning methods. He used the data from the Kaggle website. The author used some prediction models like linear regression, ridge regression, lasso regression for the prediction models. The author is used correlated matrix to see more correlated variables in the dataset and by thar he predicted the accuracy. The author says that gradient boost regression gives best accuracy after evaluating with metrices. The author wrote the paper with good content. From these I would like to use gradient boost regression model to predict the better house price ,the prediction will accurate.

The author(N.T.M SAGALA et al., 2022) builds a house price prediction using the linear regression model. In these paper the the author u has done the data pre-processing, modelling and evaluation the base models. The author used regression models to find the accuracy of the predictive model. The author used the linear regression model for the house price prediction. The author evaluated with RMSE metric with value of 0.034 .linear regression model gives a best accuracy . but the author can use some other models to predict the accuracy of the house price prediction. From this paper I would like to use the linear regression model to get best accuracy result

# METHODOLOGY

This project has been explained and proved through theoretical research and practical implementation of the machine learning algorithms used to find the prediction and to present accuracy, rmse, r2, mae, mse metrics. The theoretical part based on the peer-reviewed articles and giving answers to the research questions. The practical way of proving the code is based on the data pre-processing, feature selection, train and testing the data, base models and evaluating by metrics following by the cross-validation method to predict the machine learning models

Here are some of the steps which usually involved in building a machine learning algorithm for house price prediction.

* **Data cleaning:** In this step we have to collect the data which is suitable for the house price prediction. We can find the data set in the public sources like Kaggle, GitHub etc.,
* **Data cleaning and Data pre-processing:** The dataset contains some missing values or incorrect data that needs to be cleaned. The data should be transformed or re arranged in some way to use it for the machine learning algorithms.
* **Feature Engineering:** we have to identify the important features which are used to predict the price of the house by using some features like number of the rooms, size of the house, area, location of the house, age of house, how many floors, kitchen, Bedroom, Toilet etc., and make data ready for the train and test
* **Model selection:** selecting some of the machine learning algorithms to find the solution for this problem. some of the common algorithms used for regressions are linear regression, Lasso Regression, ridge regression decision trees, and Random Forest, support vector regression, Gradient Boost Regression.
* **Model Training:** Model Training is nothing but the training of the model after the data set cleaning and data pre-processing. These step helps to find the relationship between the features of the house you want to buy or sell and the price of the house.
* **Model Evaluation:**  Evaluating the performance of the machine learning model by using some of the technique like validation or cross-validation.

These are steps which involved for house price prediction, machine learning is best tool for house price prediction. These helps to buy the house or sell the house or the clients who are interested to invest money in the properties.

# 1.BUSINESS PROBLEM

Data science solves the real-time business problems by utilising the data to construct the models or algorithms and write the programs that help to solve to every single problems. Data science solves the real time business problems by utilizing combination models of math, stats and using different techniques in computer science to get concrete thoughts.

**PROBLEM STATEMENT AND GOING TO PREDICT:**

Predicting the price of the house by using multiple features of the house like: number of the rooms, size of the house, area, location of the house, age of house, how many floors, type of the house like furnished or partially furnished or unfurnished, kitchen, Bedroom, Toilet etc.,

By using these features, we can easily predict the price of the house by using some machine learning algorithms by using some of the data science technique we can solve the problem and can give the best prediction for house price.

# 2.DATA COLLECTION

The data I am using in this project was collected from the Kaggle. whereas Kaggle is the most trusted data science community.it gives the best resource for the project. I am using the “Ames housing Dataset”. This dataset is similar to the California Housing data set. I have downloaded the dataset from the Kaggle and it is a comma separated value (CSV) file.

It is a public dataset.it consists of records ranging from 2006-2010. This dataset consists of 82 columns or features including SalePrice , year built, plot area, locality, Yr sold, etc., The list of all features are available at [appendix]. The dataset consists of both numerical data and the categorical data. The numerical data consists of the number of the rows a, number of toilets, number o floors etc., whereas the categorical data consists of the neighbourhood etc.,

The dataset is useful for the feature engineering to make the data into train and test the dataset to splitting. It also useful for the finding the trends and the opportunity to find the patterns. The dataset is useful to find the research objectives., overall, it is very specious data for the machine learning experts, data scientist and data analyst to learn how to find the regression models to find the house price prediction. By analysing each variable in the dataset to find the machine learning models for the regression problems.

## OBJECTIVES

1. Which algorithm perform better and gives the most accurate results.
2. Predicting house price which helps for market analysis.
3. Predictive by Machine learning models helps organisational and individual person for investment decisions.
4. Helps for the property valuation and real estate valuation.
5. Accurate price predictions help for risk assessment for real estate agency and individuals’ lenders.
6. We can compare easily with other properties based on their characteristics.

# 3.TOOLS SETUP FOR PROJECT:

The Programming tools which are required for the software developers to complete Machine Learning project, which is used to create the program, debugging the program, maintaining the program and support other applications.

The specific tools which are required to complete this machine learning project are: ANACONDA NAVIGATOR, JUPYTER NOTEBOOKS.

# 4.LOADING DATASET

In machine learning, loading a Dataset is a process of importing a file into a program. The steps which involved to load the dataset is firstly we must choose the dataset which is suitable for project. we have to download the dataset, then we have to load the dataset. After loading the dataset, we have to prepare the dataset by cleaning and transforming the dataset.

# 5.DATA PRE-PROCESSING:

Firstly, we must load the dataset by using the (pd.read\_csv) syntax. This syntax will read the dataset.

df.head(10) – this method will display the first 10 rows of the dataset. This is used to check quickly the contents and the Features of the dataset. And gives an idea about the dataset.

df.tail(10) – This method will display the last 10 rows of the dataset. This method is used to know about the features of the dataset quickly.

df.dtypes – this method calls all the datatypes in the dataset.

df.isnull().any() – this function will check whether there are any null values in the dataset. If there are any null values, it shows TRUE. If there are no null values, it shows FALSE.

df.shape – This method gives the dimensions of the dataset. It means it gives the number of rows and number of columns. The first element in the tuple is number of rows and second element in the tuple is number of columns of the dataset.

df.describe() – this method gives the statistical information of the dataset. The Statistical information consists of the count, mean value, standard deviation count, minimum value, Maximum value, 25% value, 50% value and 75% value of the dataset.

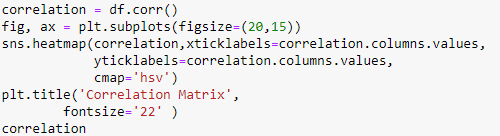
df['SalePrice'].describe() – this method gives the statistical information of the sale price column. The statistical information or the summary consists of the mean value, count, minimum value, Standard deviation, 25% value, 50% value,75% value and maximum value. It also gives the datatype of the sales price.

**Correlation matrix:**

Correlation matrix is a table which shows the correlation between the variables in the dataset. The correlation done with the method corr(). This method in the pandas gives the correlation between the columns in the dataset. The place where the column labels and row label are the same then it is equal to 1.

The correlation is mostly used in the regression techniques, this correlation is used to find the relationships between columns in the dataset and to know which column is mostly correlated

To do this correlation matrix we have to import the libraries required for it and imported the required libraries at the beginning of the code.



The correlation is done by using the heatmap.

Corr() – is method calls the correlation between the columns.

Cmap- it is parameter which gives color to the heatmap.

Correlation is a variable where the values of correlation matrix are stored.

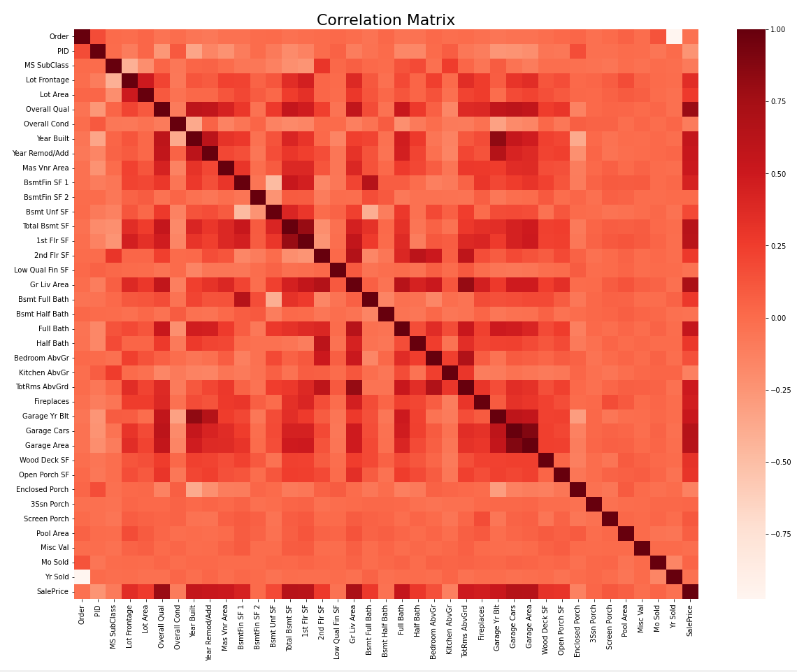


Figure 1 correlation matrix

The above figure is the correlation matrix which represents in the form of heatmap. The red colour in the figure indicates where column labels and row labels are same.

**Correlation between the saleprice:**



A new correlation has been done with sales price and create the new columns. Firstly, it Selects the **sales price** column from the correlational matrix and creates the Boolean which is true for all the correlation variables with the sales price which are greater than 0.5 and gives the list of the new columns. Then we get the new list with positive correlation between the **saleprice**. This list is used for the regression model. This list helps for the house price prediction.

**New Correlation matrix of sales price:**

A new correlation has been done from the new list which has been created in the previous line of code. Now it calculates the correlation between the new list from the above code and it stores the data in the ‘new\_correlation’ variable, and now we create the new corelation matrix by using heatmap. The title also given to the heatmap as ‘new correlation matrix’. A ‘cmap’ parameter is used to set colour to heatmap.

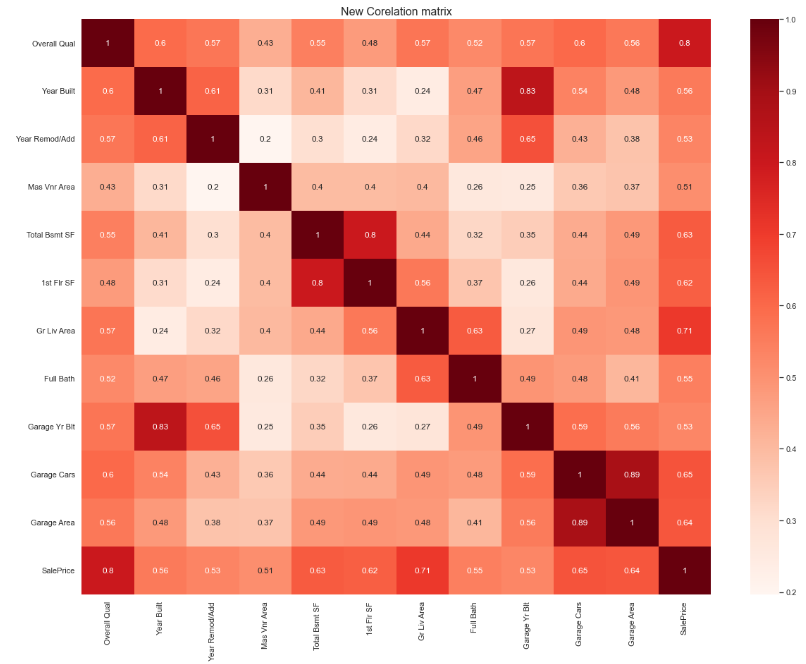
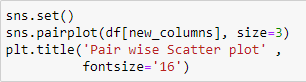


Figure 2 NEW CORRELATION MATRIX

Above Figure, is the new correlation matrix of the sales correlated columns. From the above figure we can observe that the value 1 represents the place where the columns labels and the row labels are the same. So, it is showing as the as 1. The correlation coefficients are colour coded according to the colour map. The ‘cmap’ is a parameter which specified for colour to the heatmap.

**Pair wise scatter plot:**

Now we create the pair wise scatter plot matrix to see which variables are to see which variables are having the correlation coefficient greater than 0.5 with the **saelprice** column. For the individual plot we are giving size by using the method.



The scatterplot has been done with above code by sing new columns list and size parameter gives size to individual plot in the visual.

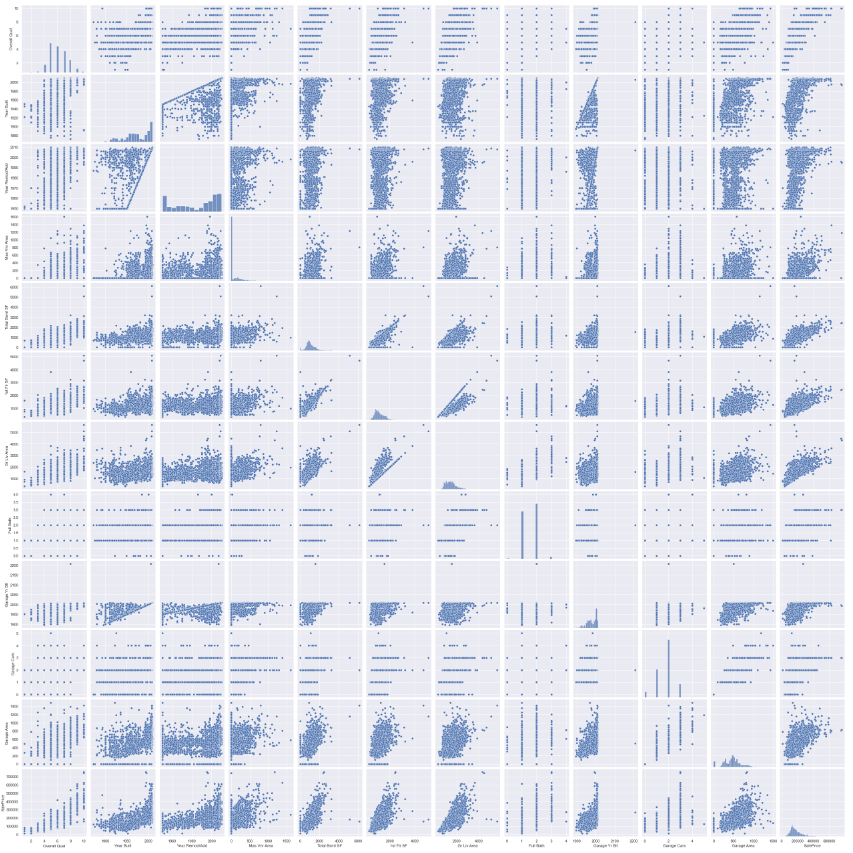
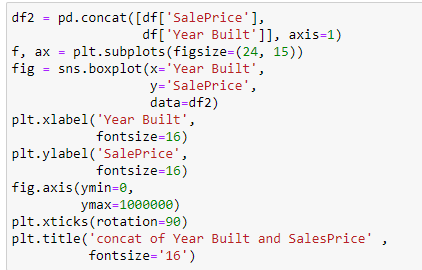


Figure 3 PAIR-WISE SCATTER PLOT

From the above figure is the scattering plot of pairwise columns. The scatter plot gives clear understanding of these variables behaved with all the other variables and with sales price.

RELATIONSHIP OF CATEGORICAL FETURES WITH SALESPRICE AND OVERALLQUAL:

Now in this we must create new data frame to store the new relationship values. The relationship shown in the form of boxplot. The concatenation has been done between the **sales price** and **overall Qual** columns.



A new dataframe is created by concatenating two variables of **salesprice** and **Overall Qual**. create new figure to show the relationship in the form of boxplot and giving figure size by using ‘figsize’ parameter. And labelling the x-axis and y-axis, limits to the plot, and title parameter is used to give title to the plot.

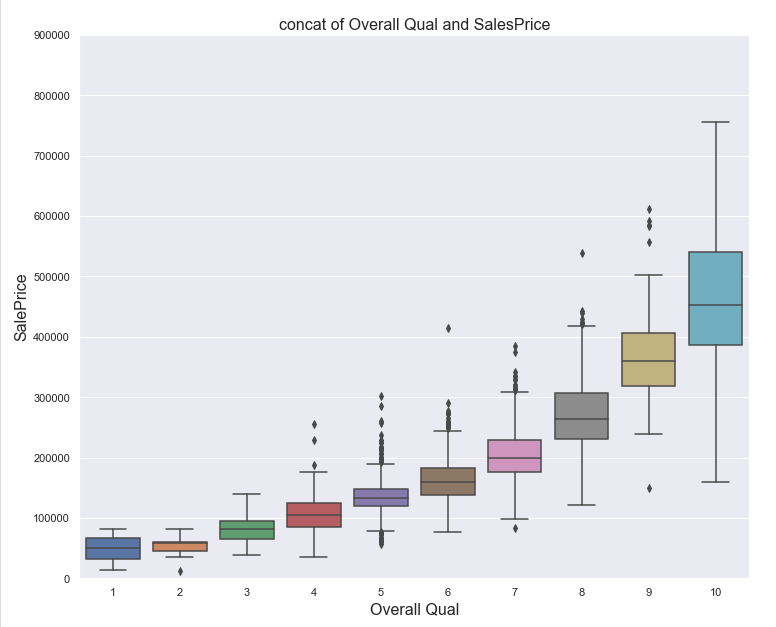
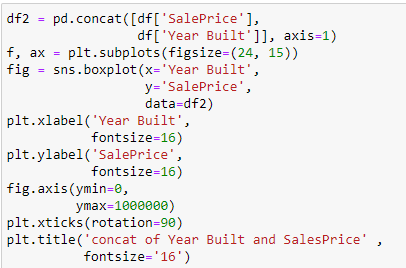


Figure 4 CONCAT OF SALESPRICE AND OVERALLQUAL

From the above figure we can say that I the **overall qual** increases then **sales price** increase, if the **overall qual** is decreasing the **salesprice** will decrease. So that we can say the **sales price** depends on the **overall qual**. It is like a vice-versa.

**REALTIONSHIP OF CATEGORICAL FEATURES WITH SALEPRICE AND YEARBUILT:**

A new dataframe which is created is used to store the categorical data of concatenating the **salesprice** and the **yr built**. The result plot is shown in the boxplot.



A new dataframe df2 is created by concatenating the **salesprice** and the **Year built.** Now create the new figure and use ‘figsize’ parameter to give size to the figure. Creating boxplot to see the relationship of categorical features of **salesprice** and **yr built** . X and Y parameters are used to use the X axis and the Y-axis. Label parameter is used to labelling in the X and Y axis. And fontsize parameter is used to give the size to font. And giving limits to X-axis and Y-axis, rotation parameter is used for readability of the labels on the X-axis. Title parameter is used to title the figure.

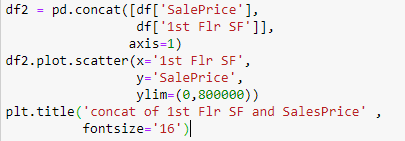
Figure 5 concat of sales price and Yr built

From the above figure we can say that the sales price is not dependent on the yr of built. In some cases, the year of built is not a very good significant. In some cases, **Yr built** will depends on the sales price.

Overall qual is mostly related to the sales price than the yr built. The sales are mostly depending on the overall quality. if the quality of houses is good then the sales price will be high. We can say by seeing the box plots of figure 4 and figure5.

**RELATIONSHIP OF NUMERICAL FEATURES BETWEEN THE SALEPRICE AND 1ST FLOOR SF:**

We can create a new dataframe by concatenating the sales price and 1st floor sf. A new plot is created to better understanding the relationships with saleprice. We can see the relationships in the form of visualisation. So that we can understand the data very clearly. When we are dealing with data, we have to more careful about it. We must observe every bit of the data very clearly and carefully.



The above code creates the scatter plot of numerical relationship of the saleprice and 1st floor SF. After creating new dataframe we must concat the saleprice and 1st floor SF of the previous dataframe. A scatter plot is created by using plot.scatter() method. X and Y parameters are the variables to plot on x- axis and Y- axis, ylim parameters is used to give the limit to y axis, and title parameter is used to add the title to the plot.

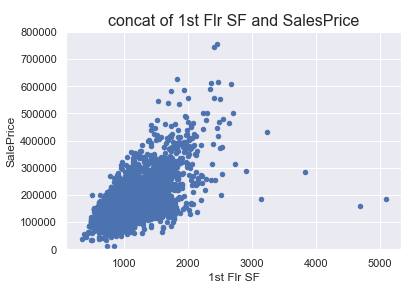
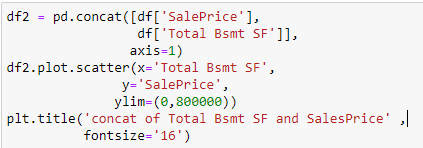


Figure 6 concat of saleprice and 1st floor SF

The above graph is a scatter plot, from that graph we can say that if the 1st floor SF increases the sales price will be increase. It applicable in most of the cases. Although sometimes if the 1st floor SF is high but the saleprice low. It depends because of other features. By seeing above graph and visualisation we can say that 1st floor SF is more, saleprice will be high.

**RELATIONSHIP OF NUMERICAL FEATURES BETWEEN THE SALEPRICE AND TOTAL BSMT SF:**

A new dataframe is created by concatenating the saleprice and Total Bsmt SF. A scatter plot is created to see the numerical features of relationship between the saleprice and total bsmt SF.



The above code creates the scatter plot which shows the relationship between the saleprice and Total Bsmt SF. By creating new dataframe and concat the two numerical Features that are saleprice and TOTAL BSMT SF. The X and Y parameters used for variables to plot on the X-axis and Y-axis, and ylim parameter is used to limit for the y axis. And title parameter is used to add the title the plot.

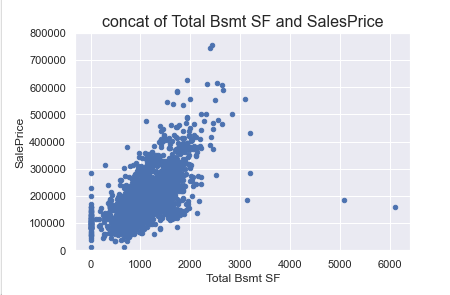


Figure 7 CONCST OF SALEPRICE AND TOTAL BSMT SF

From the above figure we can say that Total Bsmt SF is more, than the sales price will be increases. The salesprice and the total Bsmt SF are vice versa. It does not increase like linear. It increases like the slope. By visually we can say that saleprice is dependent on Total Bsmt SF.

**OUTLIERS:**

Outliers are the data values which are very high or very low. Because of the outliers the accuracy may change sometimes. Outliers may be happened by different types like type error, wrong data, Measurements errors, value errors etc.,

Visualisations are great thing to visualise the data in the form of graph. So, we can see clearly, and we can understand if there are any values which are so far from the crowd. We must remove the outliers if the values are very far away from the group. Because of the outliers we will mislead by it. So, to get good accuracy we must remove the outliers.

From the figure7 we can say that two data points are away from the group. These may happen because of the value error, measurement error. Sometimes the accuracy may be different. We must remove that outliers to get the good accuracy.

Outliers are divided into different categories, but we will some of the outliers to solve the outlier’s problem

1.univariate outliers:

Univariate outliers are done to single variable. Sale price is the single variable we are using univariate analysis. We are doing univariate outliers to the sales price to finds the outliers in the sales price. We will standardise the data apply a method to get the sales price.

To get that we are using standardScaler() - it is a pre-processing technique used in machine learning to standardise the dataset. The standard scaler usually works as the subtraction of the mean value from the original value and then divided by the standard deviation. The mathematical equation can be written is follows

Where,

Z= standardized value or normalized value

MEAN VALUE= MEAN VALUE OF THE FEATURE

ORIGINAL VALUE= ORIGINAL VALUE OF THE FEATURE

MEAN= MEAN VALUE OF THE FEATURE

STANDARD DEVIATION= STANDARD DEVIATIONOF THE FEATURE

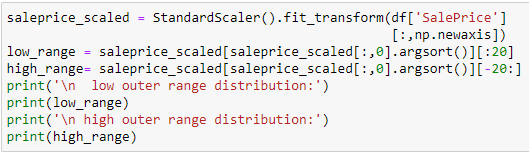
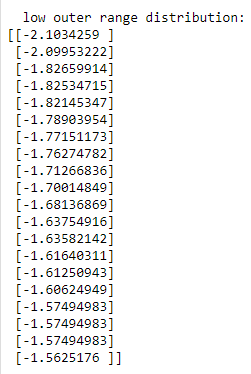
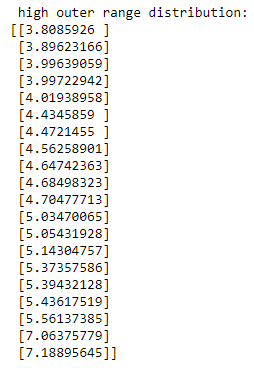


Figure 8 CDE FOR UNIVARIATE OUTLIER ANLYSIS TO TARGET VARIABLE

the code **standardscaler()** method is used to standardise the data and then new axis I created that contains the standardised value. **argsort()**  will sort the values in the high range or the lower range. And we are printing 20 values of higher range and the lower range.

Above figures are the higher and the lower range values distribution. We can see that the lower range distribution mostly consists of the negative values whereas the higher outer range distribution as the positive values. These is about the univariate outlier analysis.

2.mulitivariate outliers:

If the outliers are doing between the two variables are bivariate outliers. If any values are strange, then we consider it has a outliers and then we have to delete the outliers. We are analysing between the two variables that are saleprice and 1st flr SF to check the relation ship between them to detect the outliers.

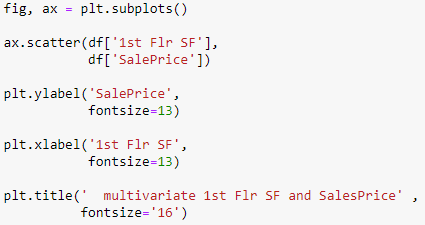


Figure 9 multivariate outliers

The code demonstrates the multivariate outliers between the saleprice and 1st flr SF by creating the scatter plot to visualise the plot. **Scatter** function is to display the scatter plot. xlabel, ylabel functions are used to display the variables on X and Y axis. **Title** function is used to give title to the plot.

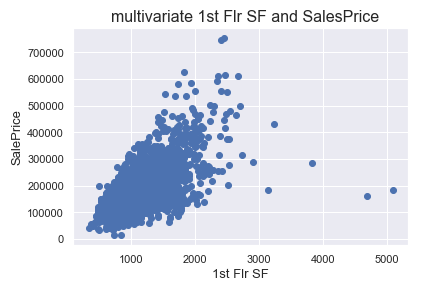


Figure 10 Multivariate outliers saleprice and 1st flr SF

From the above figure we can se that the values which are away from the group. We call them as the outliers. We have found that outliers and we have delete that outliers to get good accuracy. The two values of 1st flr SF seems to be strange and they are deviating from the group. We can see that two values are above the figure I think they are some special case. We should not delete that one.

**DELETING OUTLIERS:**

Deleting outliers is a method to deleting the values which are away from the group. We are deleting the outliers to get the good accuracy. We have to outliers with clear ratio to get the good accuracy.

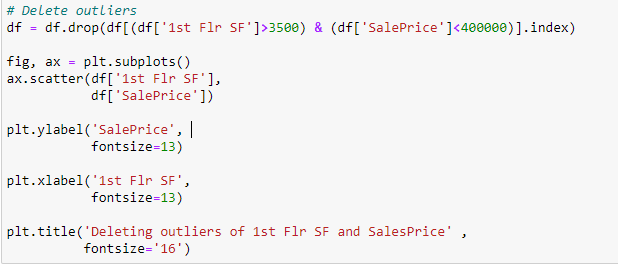


Figure 11 code for deleting Outliers

In this code, we are deleting outliers by using drop and index attribute from were we have to drop the values from the dataframe. After removing the removing the outliers we are plotting the scatter plot of 1st flr SF and sale price by doing visualize relationship.

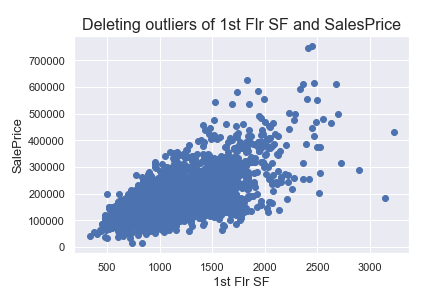


Figure 12 Deleting outliers of saleprice and 1st flr SF

From the above figure we can see that the outliers are deleted from the visualisation. By doing this relationship we can predict the exact accuracy. Removing outliers is not a good thing or safe method. although, we are deleting outliers which are very far from the group.

**HISTOGRAM AND NORMAL PROBABILITY TO SALES PRICE:**

Now we must check the normal probability to the target variable. Saleprice is the target variable in the dataset. We must apply a histogram plot and normal probability plot to the saleprice to check the values are following in the straight line or not.

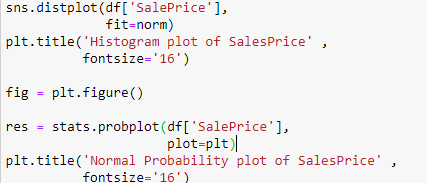


Figure 13 code to check the normal probability of the saleprice

In the above code we are plotting the histogram by using **displot** function to plot the histogram of saleprice variable. **Stats.probplot** function is used to create the normal probability plot of the saleprice variable.

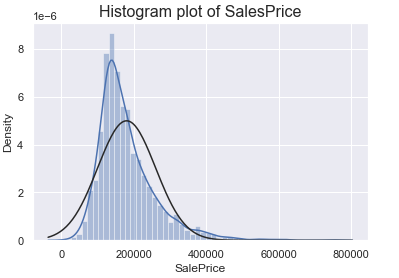


Figure 14Histogram plot of saleprice

From the above figure we can say that the actual curve and the normal distribution is lot change .so that we have skewness the data to get good accuracy. The target variable is the saleprice.

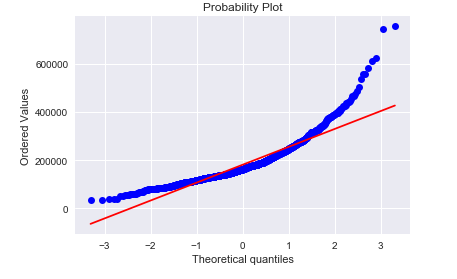


Figure 15 Probability plot o saleprice

From the above figure we can see the distribution of the target variable that is saleprice values are right skewed. The data is not normally distributed we actual know that linear regression models love the normal distribution t. so, we have to make that to normal to do that we have to skew the target variable that is saleprice.

**SKEWNESS TO TARGET VARIABLE:**

We are applying the skewness to target variable saleprice to make it to a normal distribution.

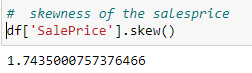


Figure 16 code to skewness the saleprice

If the value is positive the curve is right skewed, if the value is negative, it means the curve is left skewed. from the above code the skewness is applied to the target variable sale price. We got the positive value. we already know that the curve is right skewed.

**LOG TRANSFORMATION TO SALE PRICE:**

we are applying log transformations to the target variables to skewness. If the value ids positive it right skewed, if the value is negative, it means it is left skewed. By using log transformation, it makes zero. The log transformation makes the variable to suitable for all the models.

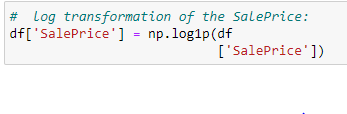


Figure 17 log transformation to target variable

In the above code log1p function is used to do log transformation to target variable that is sale price. By applying log transformation, the saleprice and make the variable to logarithm plus one.

After the log transformation we must check the skewness of the saleprice. we have to use skew() function to check the skewness of the saleprice. We apply log transformation to make right skewed to normal distribution.

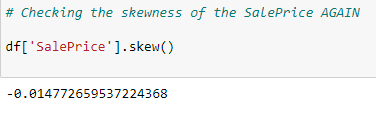


Figure 18 skewness of saleprice

From the above code we can see that the skewness is checking again to the target variable after the log transformation to the sale price is right skewed too normal.

**TRANSFORMED HISTOGRAM AND NORMAL PROBABILITY PLOT AFTER THE LOG TRANSFORMATION:**

After doing the log transformation to the target variable We have to plot the histogram and normal probability.

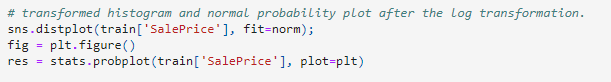


Figure 19 code

Code for the transformed histogram and normal probability plot to the log transformed . By distplot function we will plot the histogram.

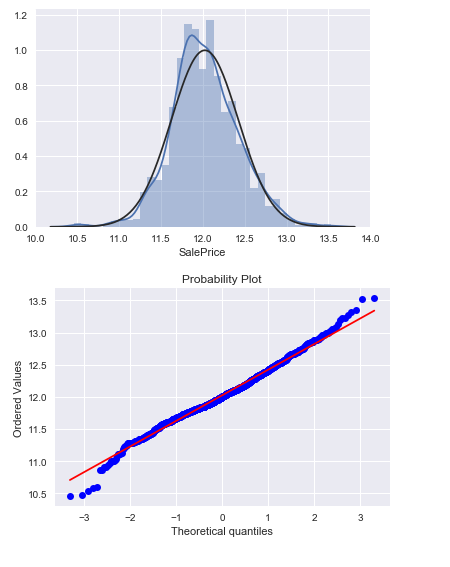


Figure 20 visualisation after the log transformation

From the above figure we can say that the values are normal distributed. The skewness is in straight line. From the histogram plot and probability plot the values are in normal distribution. now we can to more models to find the accuracy. Because of these we will get the good accuracy.

**FINDING MISSING VALUES FROM THE DATASET:**

Now we have to find the missing from the dataset, we have to check the percentage of missing values in each column of the dataset. Finding missing values is on the important step in the data pre-processing. If we find the missing values we can predict the accuracy and the dataset will be completely good for using the models. Many models and the algorithms cannot handle the missing values in the dataset.

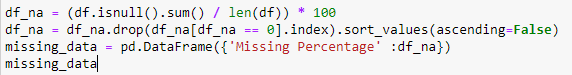
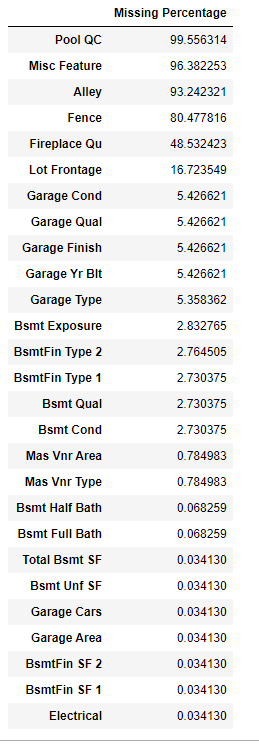


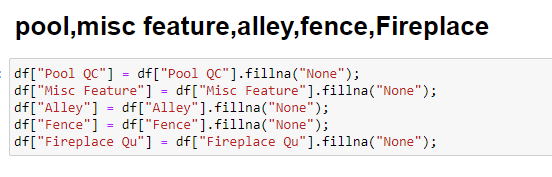
Figure 21 Code to find missing values

From the above code we can calculate the missing values percentage from the dataset from each column. Then it drops the columns which have the missing values in the dataset. Isnull() function is to find the null values.

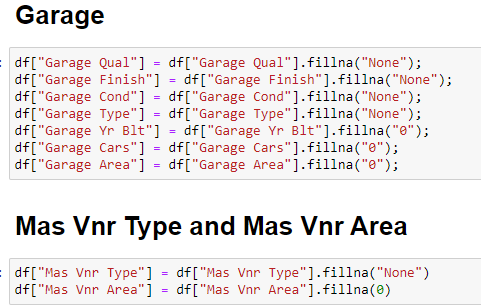


From the figure we can see that all the missing value columns from the dataset has been printed with missing values and the values are arranged in the ascending order.

Now we have all the missing values from all the specific columns of the dataset. now we have to target the missing values from each and we have to replace the values to **none** in each column. So that we have to use **fillna()** method to replace the missing values from the columns to none **.**



From the above code the missing values in the dataset are making it none by using the **fillna()** function. These attribute makes the null values to none so that it is useful to make the machine models easily. We must make all the null values **none or 0.**  By doing that we can easily launch the machine learning models for prediction.



From the figure we can see that the missing values in the dataset are filling with none and 0. In the above we can see that that missing values in the Garage Qual , Garages Finish, Garage cond, garage Type AND Mas Vnr Type are filling with the NONE and Garage Yr Built, Garage Cars, Garage Area an Mas Vnr Area are filled with 0. By doing we can make predictions and we get good accuracy.



From the above figure a code has written to make the missing values to **none or 0.** From the figure we can se that the some of the features like BsmtFin Type 1, Bsmt Qual, Bsmt Exposure, BsmtFin Type 2, Bsmt Cond are the features which are filling with **none.** The Features like Bsmt Half Bath, Bsmt Unf SF, Total Bsmt SF, BsmtFin SF 1, Bsmt Full Bath, BsmtFin SF 2, Electrical are the features are having the missing so they are filling with value **0.**  So that we can remove the missing values and w can predict the models and predict accuracy. The numerical values are making to **0** and categorical values are making to **none.**

After removing all the missing data, we have rechecked the data is any missing values are there in the dataset or not. We have to check every row and display all rows. After doing all these we have to split the data for train and test and do some regression predictions.

# 6.MAKE DATA READY FOR TRAINING

Splitting the dataset is common thing in machine learning to make the data into train data and test data. The purpose of splitting the dataset to train and test to evaluate the machine learning models. The main reason of spitting the dataset for the model evaluation, preventing overfitting of the data, hyperparameter tuning of the data by doing the machine learning model and also avoiding the data leakages. Usually,

#we will divided the date into three categories to find the machine learning models. They are training data, testing data, and validating data. Training data is used to train the model and testing data is used to test the model performance, and the validation is used to test the hyperparameter models. we can also avoid the data leakages. For this process we will divide the data into 50% and 50%. 50% is for the training the data and 50% for testing the data.

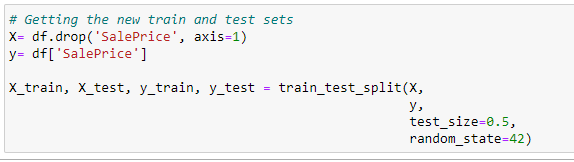


Figure 22 Training and Testing the data

The code makes the data to training and testing the data. **train\_test\_split** function is used to split the data. X represents the feature matrix and Y represents the target variable. That function divides the data into train and split for 50% 50%. now we will have four new variables **X\_train**  is used as the training features**, X\_test**  is used as the testing Features**, Y\_train**  is used as the training target**, Y\_test**  is used as the testing target. These are the variables used for the training and testing the machine learning modes.  **Random\_state** parameter is used to divide the data for reproducibility of the split. X and Y are the variables that represents the X axis and Y axis.

After splitting data we are ready to do the machine learning models to predict the data and also to find the accuracy, root mean squared error, mean squared error, R2 squared, mean absolute error.

# 7.MACHINE LEARNING MODELs

I have used liner regression, ridge regression, support vector regression, lasso regression, random forest regression and gradient boost regression. These models are mostly used for the house price prediction. In this project I am using these models to predict which model gives the best house price prediction. This prediction is useful to people who are looking to buy the house or sell the house.

## 1.SUPPORT VECTOR REGRESSION

Support vector regression model is a type regression algorithm that uses same rules as the support vector machine for the classification type of problems. It is used for the regression analysis and also it takes cares about the linear data as well as the nonlinear data and it is a powerful algorithm prediction model for the regression type of the analysis….(10)

The basic rule of support vector machine is used to convert the input data into the high dimensional feature space. The transformation is done by using the kernel function in the code.

The support vector regression model can be represented as

……(10)

Where

Y= y is the predicted valuable of the dependent variable from the dataset

W= weight

A= a denotes the input features of the independent variable

b= it is intercept bias term.

The main objective of the support vector machine is used to reduce the errors. Support vector

regressors are used to handle the nonlinear data by using the various kernel function to predict the

accuracy. These kernel function main duty is used to convert the input data into high dimensional

feature data.

## 2.RANDOM FOREST

Random forest is a classification and regression process it is used to solve the classification problems and regression problems in the machine learning. Random forest classification is a supervised learning. Random forest has more advantages than the other machine learning models . [P.F. SMITH.,2013]

Fig 2. - 
Random Forest
….(12)

Figure 1 RANDOM FOREST

From the above we can say that in the random forest regression, we can train the multiple decision trees at a time. And each decision tree will predict the results separately, and they will be averaged by sum of the results by number of the result and the final output will be new accuracy value for the random forest regression.

The random forest regression model equation

…….(12)

Where

Ycap= predicted value

N= total numbers of the decision trees random forest technique

F(x)= single decision tree result

Randomly select the data from the dataset, and also select some of the features from the dataset and builds the decision tree and after building the n decision trees the results are n predictions will be get. We have to do average for the sum of the results divide by the number o results. There here you go the new predicted value.

## 3.RIDGE REGRESSION

RIDGE Regression is mathematical method, which is used for the regression analysis, especially when they are dealing with high correlation values among the predictor values. Ridge regression is a organised version of the linear regression.

In the mathematical form ridge regression is called as the minimize of residual sum of squares is added to the Ridge Penalty.

The formula of the ridge regression is written as follows

+ lambda\*….. ……. …..(13)

Where,

Ycap=predicted value

sum of the squared errors

Lambda= parameter

Β= error value

Ridged regressions are mostly used in the regression type of problems to predict the accuracy and evaluated by the evaluation metrics. Ridge regression helps the model to reduce the risk of overfitting problems. This is one of the valuable tools in the regression problems

## 4.GRADIENT BOOST REGRESSION

Gradient boosting regressor is one of the machine learning algorithms and it is used for the regression problems. We can predict the house price regression using the machine algorithm.it collects the weak models and combine all the models and predict the result.

The gradient boost regressor calculates the difference between the present prediction value and the target value. The equation for Gradient boost regressor can be written as follows

……(7)(8)

Where

F(x)= model selection

α=learning rate of the model

updated predicting value of the equation.

Gradient boost technique is most power model to predict the accuracy for these type of regression problems. Most of the authors will use the model in the paper for the prediction. We can also enhances the model by using regression ,model to the features of the dataset.

## 5.LINEAR REGRESSION

Linear regression model is mostly used regression model for the prediction. Linear regression is a statistical procedure is used to correlation between the one dependent variable which is usually denoted as Y and one or more independent variable which is usually denoted as X. The main aim of the linear regression model is to find the best relationship model between the dependent variable and independent variable. Linear regression model is represented by using a straight-line equation.

The straight-line equation is following below.

…….(14)

Where:

* Y = Dependent Variable
* =Independent variable
* = intercept variable
* = Regression coefficients.
* = is the error term

we are using linear regression to train the data, by using the dataset. These regression model can be used to find the accuracy by using evaluation metrics RMSE,RSE,R2 AND MAE.

## 6.LASSO REGRESSION:

lasso regression is one of the regression model used for the regression techniques to predict the accuracy of the data. Lasso regression can be called as the regularized algorithm that helps to find irrelevant parameters and then it eliminates the value and then it helps to select the model and regularize the model. The statistical form of the lasso regression can be written in the form of

The lasso regression model can be written in this type of equation s follows.

…………..(13)

Where

Ycap= predicted value for the model

Lambda = lambda is parameter to regularize the model

absolute value

error

Lasso regression is most useful regression model for the house price prediction. It is mostly used for the variable selection it is one of the models used for the regression model to get accuracy.

# 8.METRICS

For regression problems, there are 4 most used common metrics they are Mean squared error, root mean squared error, R-Squared and mean absolute error these are four metrics used in these projects. These metrices are used to analyse the performance of the model used for the prediction. We must select the best metrics used to check the performance of the model. Here are some of the most frequently used metrics they are:

## MEAN SQUARED ERROR (MSE):

MEAN SQUARED ERROR metrics which commonly called as MSE. Mean squared error is nothing but the calculating the average of the squared difference between the predicted value and the original value. the equation of the mean squared error is as shown below.

….. (6)

Where,

n = n is the number of samples

yi= yi is the original value

= is the original value

which model gives the lowest value of MSE is making the accurate predictions. MSE is useful metric for the regression models. And it is useful metric for evaluating the regression models.

## ROOT MEAN SQURED ERROR (RMSE):

ROOT MEAN SQUARED ERROR metrics which commonly called as RMSE. Which is used to measure the deviation of the observed values and the true values. these is one of the best regression metrics which is used for the regression problems.

The equation of the root mean squared error is as shown below

……….(6)(10)

Where

RMSE = ROOT MEAN SQUARED ERROR

N= number of the observations in the datapoint

fi= represents the value of the target variable

yi= represents the predicted value of the target variable.

RMSE is used to measure of the prediction errors in the same datapoints o the target variable. when we are using RMSE metrics to regression model. Which model gives the low predicted value is considered as the best predative model for RMSE. By removing the outliers only, we have to predict the metrics to get best accuracy value. RMSE is sensitive for the outliers. sometimes the accuracy will be changed because of the outliers. So, we have to be concentrated when we are doing the RMSE.

## R2 SQUARED ERROR

R2 squared error is a statistical method and it is subtracted from the sum of the squared residuals divided by total sum of the squares by 1. Therefore, the equation can be written as

………(10)

WHERE,

= SQUARED ERROR

= it is the sum of the squared difference between the predicted value of the independent variable and the actual value

= it is the sum of the squared difference between the actual value and the mean of the dependent variable

Which model gives the low predicted value is considered as the best predictive model for R2. For most of the prediction r squared is use for the evaluation the model prediction.

## MEAN ABSOLUTE ERROR (MAE)

Mean absolute error is the regression metric used for prediction of the regression model. The statistical formal

……….(12)

Where

MAE= MEAN ABSOLUTE ERROR

N= NUMBER OF THE OBSERVATIONS VALUES FROM THE DATSET.

Y= ACTUAL VALUE OF THE DEPENDENT VALUE

YCAP= YCAP IS THE PREDICTED VALUE FROM THE DEPENDENT VARIABLE

Which model gives the low predicted value is considered as the best predictive model for MAE. The mean problem with MAE is will not able to tell the errors. Either it is not able to tell the model trends are overestimated or underestimate the actual values from the dataset. It is most used regression metrics to predict the regression model and to find the performance of the model.

# 9. RESULTS AND EVALUATION:

This study is about the house price prediction using the machine learning models to predict the houses prices for sale, or buy. To get that we are using variables from the ames housing dataset to predict the which models gives the best prediction model to buy the house. It is mostly used for the people who want to buy the house individually, and it also helps to the real estate agency and the organisation. And it gives the market analysis and also help to check the risk assessment for buying o investment in the property. It is also useful for the government for making policies and the rules. We are predicting by using the machine learning models after the test and training the data, we have used the 6 models and 4 metrics to predict which model is best. The models which we used for the prediction are support vector regression, random forest regression, ridge regression, gradient boost regression, linear regression and Lasso regression model.

Now we will see the output of each base model in the tabular form.

**BASE MODEL TABLE:**

|  |  |  |
| --- | --- | --- |
| S.NO | BASE MODELS | SCORE |
| 1 | SUPPORT VECTOR REGRESSION MODEL | 0.929 |
| 2 | RANDOM FOREST | 0.982 |
| 3 | RIDGE REGRESSION | 0.999 |
| 4 | GRADIENT BOOST REGRESSION | 0.971 |
| 5 | LINEAR REGRESSION | 0.999 |
| 6 | LASSO REGRESSION | 0.047 |

From the above table we can say that **LASSO REGRESSION** model gives the best accuracy value to predict the house price prediction in the base models. The model which gives the lowest value of the base model is the best model to predict the house price prediction. The score for the **LASSO REGGRESSION MODEL** is **0.447.** Next best accuracy vale of the base model is **SUPPORT VECTOR REGRESSION** model it is also giving the best accuracy value for predicting the accuracy of the house price prediction. The next values come by support vector regression is **0.929.**  The models which we used in this project are best models because the score remaining models very nearby. But LASSO REGRESSION MODEL makes the prediction score low. And the next comes to SUPPORT VECTOR REGRESSION. So we can say that LASSO REGRESSION AND SUPPOT VECTOR REGRESSION MODEL are the best base models to predict the house price prediction.

But we have to check the results by the checking with metrics. The metrics which we are used in the project are mean absolute error, root mean squared eror,R2 square, mean absolute error to all the 6 models to get the scores of each value.

**MSE SCORE TABLE:**

|  |  |  |
| --- | --- | --- |
| S.NO | MODELS | MSE SCORE |
| 1 | SUPPORT VECTOR REGRESSION | 0.1538 |
| 2 | RANDOM FOREST | 0.0225 |
| 3 | RIDGE REGRESSION | 0.0339 |
| 4 | GRADIENT BOOST REGRESSION | 0.0179 |
| 5 | LINEAR REGRESSION | 6.0238 |
| 6 | LASO REGRESSION | 0.0397 |

The above table is the MSE scores of all the models and we are analysing which model gives the lowest value of MSE is making the accurate predictions. From the above table we can see that Gradient boost Regression model gives the lowest value. In that case Gradient Boost Regression model gives the lowest value in mse. It means Gradient boost regression model gives the accurate predictions in mean squared error with **0.0179.**

**RMSE SCORE TABLE:**

|  |  |  |
| --- | --- | --- |
| S.NO | MODELS | RMSE SCORE |
| 1 | SUPPORT VECTOR REGRESSION | 0.40 |
| 2 | RANDOM FOREST | 0.15 |
| 3 | RIDGE REGRESSION | 0.18 |
| 4 | GRADIENT BOOST REGRESSION | 0.12 |
| 5 | LINEAR REGRESSION | 0.37 |

The above table is the RMSE scores of all the models and we are analysing which model gives the lowest value of RMSE is making the accurate predictions .from the above table we can see that Gradient boost Regression model gives the lowest value. In that case we can say Gradient boost regression model gives the lowest value in rmse. It means Gradient boost regression model gives the accurate predictions in mean squared error with **0.12**

**R2 SCORE TABLE:**

|  |  |  |
| --- | --- | --- |
| S.NO | MODELS | R2 SCORE |
| 1 | SUPPORT VECTOR REGRESSION | -2.744 |
| 2 | RANDOM FOREST | 0.868 |
| 3 | RIDGE REGRESSION | 0.796 |
| 4 | GRADIENT BOOST REGRESSION | 0.890 |
| 5 | LINEAR REGRESSION | -3.617 |
| 6 | LASO REGRESSION | 0.796 |

The above table is the R2 scores of all the models and we are analysing which model gives the lowest value of R2 is making the accurate predictions. From the above table we can see that Ridge Regression model gives the lowest value. In that case we can say ridge regression model gives the lowest value in R2. It means Ridge regression model gives the accurate predictions in R squared with **0.796.**

**MAE SCORE TABLE:**

|  |  |  |
| --- | --- | --- |
| S.NO | MODELS | MAE SCORE |
| 1 | SUPPORT VECTOR REGRESSION | 0.316 |
| 2 | RANDOM FOREST | 0.316 |
| 3 | RIDGE REGRESSION | 0.131 |
| 4 | GRADIENT BOOST REGRESSION | 0.087 |
| 5 | LINEAR REGRESSION | 0.371 |
| 6 | LASO REGRESSION | 0.137 |

The above table is the MAE scores of all the models, and we are analysing which model gives the lowest value of MAE is making the accurate predictions. from the above table we can see that Gradient boost Regression model gives the lowest value. In that case we can say gradient boost regression model gives the lowest value in MAE. It means Gradient boost regression model gives the accurate predictions in mean absolute error with **0.08**

# 10. CONCLUSION

The study used the ames housing dataset for the house price prediction. In this study we have included the data pre-processing, future engineering, nonlinear relationships between the variables, detecting the outliers and removed outliers, and handles with missing data and made the missing data to none or 0.

Then we have used some of the regression models for predicting accuracy. The models that are used in the project are support vector regression model, Random Forest regression model, Ridge regression, Gradient Boost Regression, Linear Regression, lasso regression are the base models used for the prediction. From that models we can say that **LASSO REGRESSION** and the **SUPPORT VECTOR REGRESSION** model gives the best accuracy.

Furthermore, we have used 4 evaluation metrics for predicting the accuracy of the above predicting model. By using **MEAN SUARED ERROR(MSE)** metrics we found Gradient boost regression model gives the low mse score **0.01789** and predicted good accuracy, By using **ROOT MEAN SQUARED EROR(RMSE)** metrics we found Gradient boost regression model gives the low rmse score 0.12and predicted good accuracy, By using **R2 square(R2)** metric we found Ridge Regression model gives the low R2 score **0.796** and predicted good accuracy, by using **MEAN ABSOLUTE EROOR(MAE)** metric we found Gradient boost regression model gives the low mae score **0.08** and predicted good accuracy.

From the evaluations metrics we can say that Gradient boost regression is predicting the accuracy for **MSE,RMSE,R2.**

# 11.FUTURE WORK

We can use remaining regression models and the other metrics to predict the best accuracy and use cross validation to see the model fitting. Sometimes for cross validation we need big machine to run it. We can also predict the house price by using the other features in the dataset to get the accuracy by analysing deeply we can discover new insights and helps to improves the accuracy of predicting the house price prediction.

# 12. REFERENCES

[1]. J. Mu, F. Wu and A. Zhang, "Housing Value Forecasting Based on Machine Learning Methods", Abstract and Applied Analysis, pp. 7, 2014.

[2]. A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[3]. P. F. Smith, S. Ganesh and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience", J. Neurosci. Methods, vol. 220, no. 1, pp. 85-91, 2013.

[4]. D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.

[5]. C.-C. Lin, S.-J. Lin et al., “An estimation of elasticities of consumption demand and investment demand for owner-occupied housing in taiwan: a two-period model,” International Real Estate Review, vol. 2, no. 1,pp. 110–125, 1999.

[6]. Z. Zhang, "Decision Trees for Objective House Price Prediction," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 280-283, doi: 10.1109/MLBDBI54094.2021.00059.

[7]. Yan, Z. and Zong, L. (2020). Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms. *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International* Conference *on Big Data and Artificial Intelligence*.

[8]. A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi: 10.1109/ICAC3N53548.2021.9725552.

[9]. Adair Alastair S., Berry Jim N., and McGreal W. Stanley. 1996. Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* 13, 1 (1996), 67–83

[10]. Y. Fang, T. Li and H. Zhao, "Random Forest Model for the House Price Forecasting," 2022 14th International Conference on Computer Research and Development (ICCRD), Shenzhen, China, 2022, pp. 140-143, doi: 10.1109/ICCRD54409.2022.9730190.

[11]. B. Almaslukh, "A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems," 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taipei, Taiwan, 2020, pp. 217-222, doi: 10.1109/TAAI51410.2020.00047.

[12]. S. A. Septianingrum, M. Alfian Dzikri, M. A. Soeleman, P. Pujiono and M. Muslih, "Performance Analysis of Multiple Linear Regression and Random Forest for an Estimate of the Price of a House," 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2022, pp. 415-418, doi: 10.1109/iSemantic55962.2022.9920454.

[13]. Z. Li, "Prediction of House Price Index Based on Machine Learning Methods," 2021 2nd International Conference on Computing and Data Science (CDS), Stanford, CA, USA, 2021, pp. 472-476, doi: 10.1109/CDS52072.2021.00087.

[14]. N. T. M. Sagala and L. H. Cendriawan, "House Price Prediction Using Linier Regression," 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, 2022, pp. 1-5, doi: 10.1109/ICCED56140.2022.10010684.

# APPENDIX

A list of all features in the public data placed in Ames Housing Dataset

|  |  |  |
| --- | --- | --- |
| ID | FEATURE | DESCRIPTION |
| 1 | id | Identifies an entry |
| 2 | MS Sub Class | Identifies the type of dwelling involved in the sale |
| 3 | MS Zoning | Identifies the general zoning classification of the sale |
| 4 | Lot Frontage | Linear feet of street connected to property |
| 5 | Lot Area | Type of road access to property |
| 6 | Street | Type of alley access to property |
| 7 | Alley | General shape of property |
| 8 | Lot Shape | Flatness of the property |
| 9 | Land Contour | LandContour |
| 10 | Utilities | Type of utilities available |
| 11 | Lot Config | Lot configuration |
| 12 | Land Slope | Slope of property |
| 13 | Neighbor hood | Physical locations within city limits |
| 14 | Condition 1 | Proximity to various conditions |
| 15 | Condition 2 | Proximity to various conditions (if more than one is present) |
| 16 | Bldg Type | Type of dwelling |
| 17 | House Style | Style of dwelling |
| 18 | Overall Qual | Rates the overall material and finish of the house |
| 19 | Overall Cond | Rates the overall condition of the house |
| 20 | Year Built | Original construction date |
| 21 | Year Remod Add | Remodel date (same as construction date if no remodelling or additions) |
| 22 | Roof Style | Type of roof |
| 23 | Roof Matl | Roof material |
| 24 | Exterior 1 st | Exterior covering on house |
| 25 | Exterior 2 nd | Exterior covering on house (if more than one material) |
| 26 | Mas Vnr Type | Masonry veneer type |
| 27 | Mas Vnr Area | Masonry veneer area in square feet |
| 28 | Exter Qual | Evaluates the quality of the material on the exterior |
| 29 | Exter Cond | Evaluates the present condition of the material on the exterior |
| 30 | Foundation | Type of foundation |
| 31 | Bsmt Qual | Evaluates the height of the basement |
| 32 | Bsmt Cond | Evaluates the general condition of the basement |
| 33 | Bsmt Exposure | Basement exposure |
| 34 | Bsmt Fin Type1 | Refers to walkout or garden level walls |
| 35 | Bsmt Fin SF1 | Type 1 finished square feet |
| 36 | Bsmt Fin Type2 | Rating of basement finished area (if multiple types) |
| 37 | Bsmt Fin SF2 | Type 2 finished square feet |
| 38 | Bsmt Un fSF | Unfinished square feet of basement area |
| 39 | Total Bsmt SF | ---------- |
| 40 | Heating | Type of heating |
| 41 | Heating QC | Heating quality and condition |
| 42 | Central Air | Central air conditioning |
| 43 | Electrical | Electrical system |
| 44 | 1st Flr SF | First Floor square feet |
| 45 | 2nd Flr SF | Second floor square feet |
| 46 | Low Qual Fin SF | Low quality finished square feet (all floors) |
| 47 | Gr Liv Area | Above grade (ground) living area square feet |
| 48 | Bsmt Full Bath | Basement full bathrooms |
| 49 | Bsmt Half Bath | Basement half bathrooms |
| 50 | Full Bath | Full bathrooms above grade |
| 51 | Half Bath | Half baths above grade |
| 52 | Bedroom Abv Gr | Bedrooms above grade (does NOT include basement bedrooms) |
| 53 | Kitchen Abv Gr | Kitchens above grade |
| 54 | Kitchen Qual | Kitchen quality |
| 55 | Tot Rms Abv Grd | Total rooms above grade (does not include bathrooms) |
| 56 | Functional | Home functionality (Assume typical unless deductions are warranted) |
| 57 | Fire places | Number of fireplaces |
| 58 | Fireplace Qu | Fireplace quality |
| 59 | Garage Type | Garage location |
| 60 | Garage Yr Blt | Year garage was built |
| 61 | Garage Finish | Interior finish of the garage |
| 62 | Garage Cars | Size of garage in car capacity |
| 63 | Garage Area | Size of garage in square feet |
| 64 | Garage Qua | Garage quality |
| 65 | Garage Cond | Garage condition |
| 66 | Paved Drive | Paved driveway |
| 67 | Wood Deck SF | Wood deck area in square feet |
| 68 | Open Porh SF | Open porch area in square feet |
| 69 | Enclosed Porch | Enclosed porch area in square feet |
| 70 | 3Ssn Porch | Three season porch area in square feet |
| 71 | Screen Porch | Screen porch area in square feet |
| 72 | Pool Area | Pool area in square feet |
| 73 | Pool QC | Pool quality |
| 74 | Fence | Fence quality |
| 75 | Misc Feature | Miscellaneous feature not covered in other categories |
| 76 | Misc Va | Value of miscellaneous feature |
| 77 | Mo Sold | Month Sold (MM) |
| 78 | Yr Sold | Year Sold (YYYY) |
| 79 | Sale Type | Type of sale |
| 80 | Sale Condition | Condition of sale |
| 81 | Sale Price | Transaction price |