

**Housing Project**

Submitted by:

Sourav Chaudhury

**ACKNOWLEDGMENT**

I would like to take this opportunity to show my gratitude to the resources mentioned below for helping me complete this project successfully.

* Flip Robo Technologies
* DataTrained Education
* TowardsDataScience
* AnalyticsVidya
* StackOverflow
* GitHub

**INTRODUCTION**

* **Business Problem Framing**

Today, most real estate data companies and professionals are finding the way how to efficiently market their listings. This market keeps on changing and if a particular technique has work for some-time doesn't guarantee that it will work in future. This requires constant evaluation, training and review of the work process of a Realtor. Experienced agents get most of their business from repeat clients and referrals from past clients. A new agent simply, by definition, does not have past clients. It takes years to build up a roster of past clients that will keep you busy with no advertising expense beyond a few postage stamps and nowadays fewer postage stamps because email campaigns, and twitter, Facebook, and even the old phone call. Accurate market analysis and predictions need correct statistics, which, in turn, can only be drawn from a well maintained, accurate data set. Collating this kind of a data set involves aggregating a billion documents, ranging from public records, valuations, pricing data, to mortgage performance information.

* **Conceptual Background of the Domain Problem**

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

* **Review of Literature**

Data is an essential ingredient of our present economy. People from different fields rely so much on it that it has pervaded their personal and social lives. The information derived from analysing data can alter the manner in which we undertake research, real estate research inclusive. Data science has brought about a lot of positive impact in the real estate profession, such as solving the challenges evolving in the profession. This is evident that data analytics has provided a great support for property development in many countries. It helps property companies to implement diversified investments by reminiscent of potential worth data. Digital personal data and revolutionary thinking changes create new opportunities for real estate companies to grow innovative investments in this age. Likewise, data science has been of use in forecasting house prices, as indicated by who derived the prices by analysing internet searches and media reports. Additionally, displayed the efficacy of data analytics in the creation of models. They performed an analysis of search trends on Google and used the result to create a house-prediction template or model which was more accurate than predictions from the National Association of Realtors by a significant margin of over 20%. It is apparent from the above that several research projects have been conducted on big data and real estate profession in developed and developing countries of the world. Moreover, most of the research findings are based on individual findings/opinions. This review therefore seeks to harmonize the various findings

In this project data science with machine learning algorithms are used to make a model which is capable of predicting sales prices of houses based of different variables provided. The data is provided by a US based housing company and they wanted to know which variables are more important for predicting house prices and which variables have more influence over the prices. A through data cleaning is done with mathematical and statistical analysis and descriptive statistics is laid down for a better understanding of the data. After this an extensive EDA (Exploratory Data Analysis) is done with pictorial representation of the data. Based on various factors the important variables are laid down which have a greater say in deciding the price of the houses. A part of data was trained with specific algorithms and then it was also tested with metrics to find the accuracy of the model. The data provided by the client was pre processed in ways to make it understandable by the model. Efficient machine learning libraries were used to complete the project with good accuracy.

* **Motivation for the Problem Undertaken**

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. I made a Machine Learning model in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

This company wants to know:

• Which variables are important to predict the price of properties.

• How do these variables describe the price of the house.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

We have received the data from the client and we needed to find out the variables that were affecting the prices of the houses. An extensive EDA (Exploratory Data Analysis) is performed to find out the variables that are more correlated to the target variable which is the sales prices of the properties. There were missing values in the dataset which were treated with certain strategies. I have used descriptive statistics in order to find the statistical elements of the data provided by the client. The descriptive statistics gave me certain insights and also helped me to pre conceptualize ways to deal with the problem.

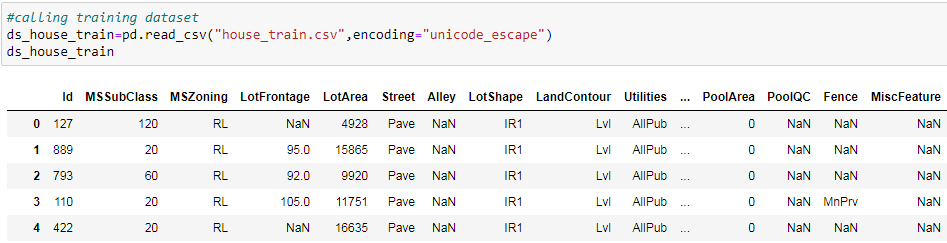
There were around 70 columns of input variable and one target variable that is the sales price of the houses. The training data was used to train the model and the testing data was used to predict the values. Al the training and testing data came in separate files, all the steps in pre-processing was first done to the training dataset and then the same applied to the testing dataset for consistency.

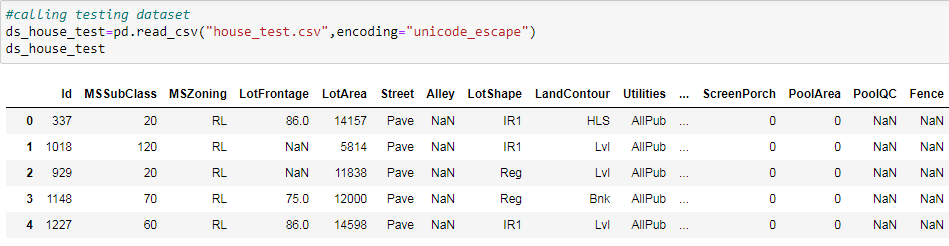
The data went through the pre processing steps which included replacing the missing values in the dataset, removing outliers, fixing skewness and finally quantifying the categorical values in order to feed them into the model. Later several algorithms were used to check which works best with the given dataset in given circumstances.

* **Data Sources and their formats**

The data was provided by the housing company . They have gathered the data from the sale of houses in Australia. Two datasets were provided by the company they are train.csv and test.csv. Both the files were in csv format . The train.csv is for training the model and the test.csv is to predict the house prices. While calling the csv in Jupyter notebook the encoding format was given as unicode escape. Unicode escape sequences convert a single character to the format of a 4-digit hexadecimal code point. There were null values in both the training and testing the dataset which are later treated with specific strategies.

Below are the snapshots of the training and testing dataset.





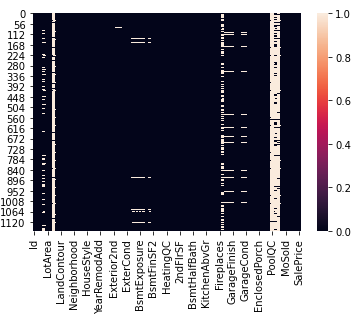
* Below are the summary snapshots of the training and testing dataset.

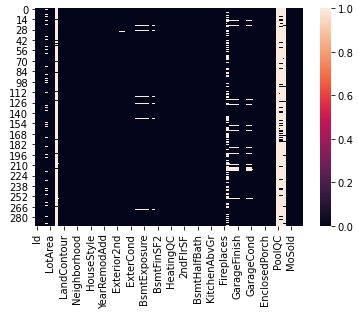
Capture6.PNG

* The above snapshot shows the training data is a combination of float, integer and object or categorical variable. There are total of 3 columns with float data types, 35 columns with integer data types and 43 columns with object or categorical data types.

Capture7.PNG

* The above snapshot shows the testing data is a combination of float, integer and object or categorical variable. There are total of 4 columns with float data types, 34 columns with integer data types and 42 columns with object or categorical data types.
* There were null values present in the dataset. Below are a snapshot of the null values of the training and testing dataset in a heatmap.





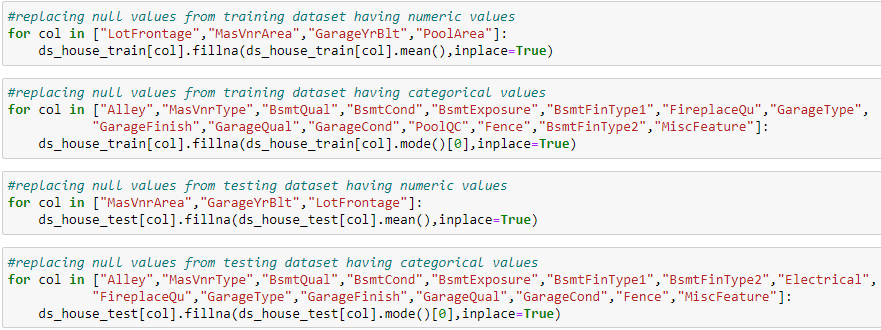
* **Data Preprocessing Done**

Data Preprocessing is one of the most important part of making any machine learning model. It involves processing the raw data in a way that it is easily understandable by the model for higher efficiency and efficacy.

* **Replacing Missing Values**

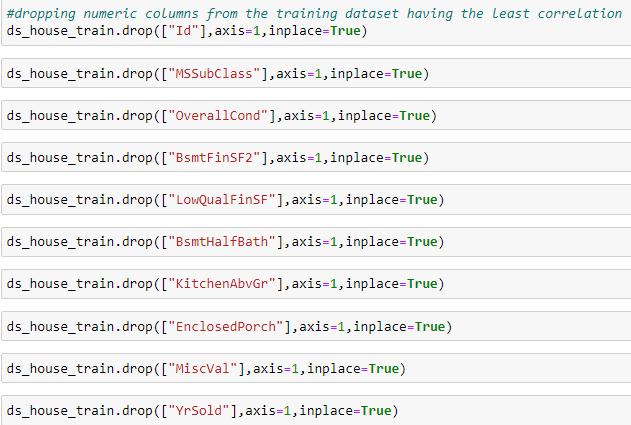
The data preprocessing started with replacing the missing values from the columns of both the training and testing dataset. Different strategies were used to replace null values with different data types. For columns having numeric data types mean or average value of the column is used and for categorical columns the mode or the most frequently occurring value in the column is used to replace the missing values.

Below are the codes used to replace the missing values for both the training and testing dataset.



* **Dropping Columns**

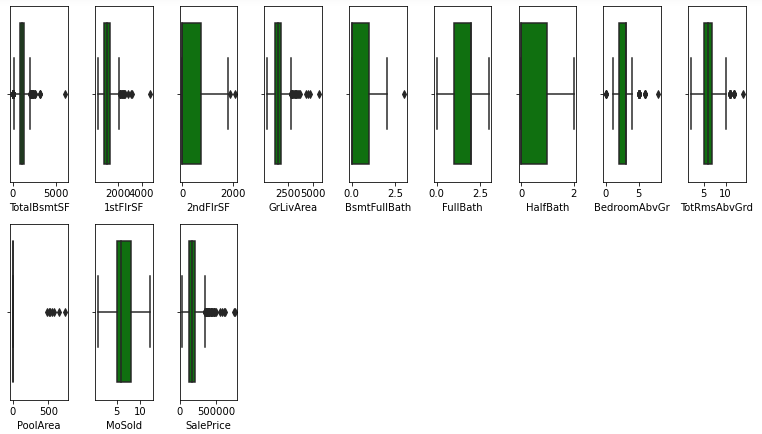
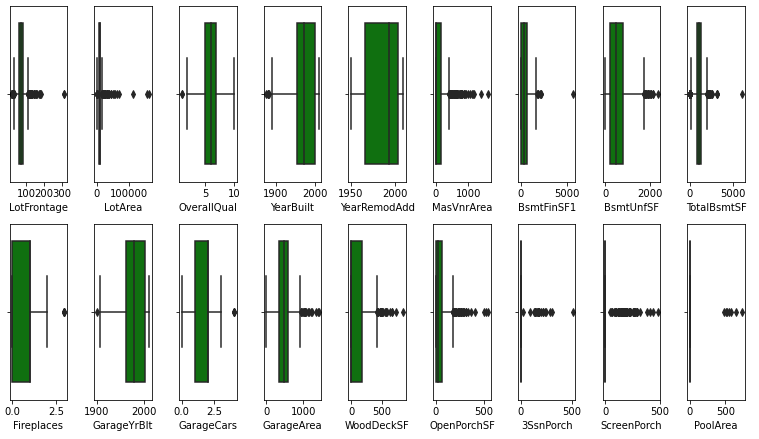
* Correlation graph is plotted considering the target variable i.e. the sales price and any variable that shows a negative or lesser correlation is with it will be dropped. I have already attached the correlation graph in the above descriptive statistic section and below are the codes to drop the variables showing a lesser correlation with the target variable.



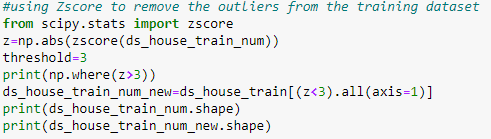
The above code is for dropping the columns in the training dataset and the same strategy will be applied to the testing dataset as well for consistency.

Below are the columns that are dropped from both the training and testing dataset.

* Type of dwelling involved in the sale.
* Rates the overall condition of the house
* Type 2 finished square feet
* Low quality finished square feet (all floors)
* Basement half bathrooms
* Kitchens above grade
* Enclosed porch area in square feet
* $Value of miscellaneous feature
* Year Sold (YYYY)
* **Plotting Outliers**
* Removing outliers are also an essential part of pre processing and making an efficient machine learning model. Present of outliers will have affect on variance, and standard deviation of a data distribution. In a data distribution, with extreme outliers, the distribution is skewed in the direction of the outliers which makes it difficult to analyze the data and will result in biased insights.
* Below are some of the numeric columns with outliers plotted in a box plot.

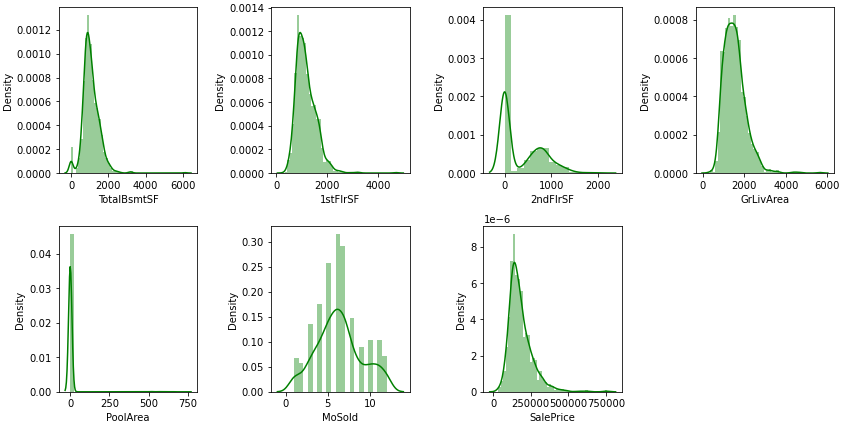
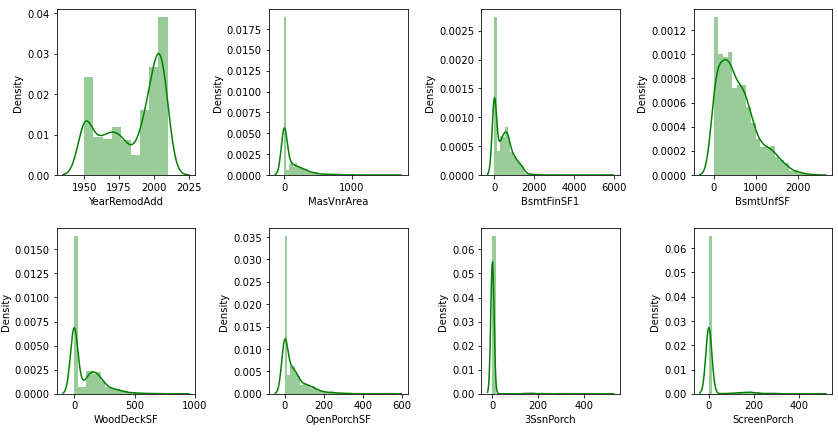
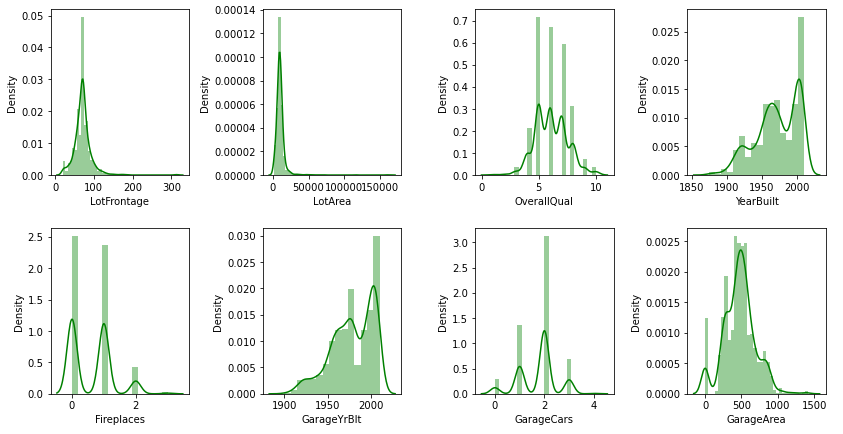


* **Removing Outliers**
* I have used Z score to remove the outliers. A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. Any points that will be above Z-score 3 will be considered as outliers.
* Below is the code to remove the outliers from the dataset.

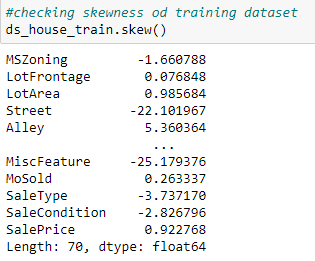


Capture15.PNG

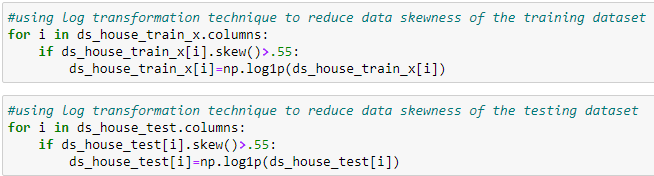
* The above shape shows there were 1168 data points before removing the outliers and after removing it came down to 980.
* Only numeric columns were considered while implementing Z score. The total number of numeric columns in the training dataset is 24 and total columns were 70.
* **Checking Skewness**
* Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. A data can be right skewed or left skewed. A easy graphical representation of checking skewness of data is by plotting them in a distribution graph and also listing down there skew values.
* Below are some of the data distribution of the columns in the dataset.



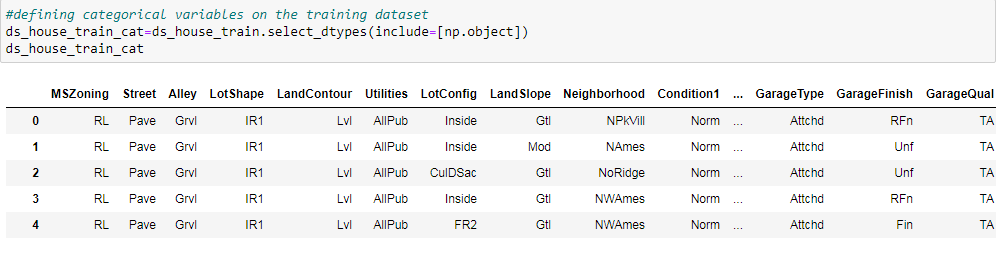
* Below is the code to describe skewness of data.



* Any columns throwing a value greater than + or - .55 is highly skewed an needs to be treated. Below are the codes to treat skewness of both the training and testing dataset.



* The above code log transforms any columns that shows a skewness greater than + or - .55 and does not manipulates the ones with values lesser than the threshold.
* **Label Encoding**
* Label Encoding is a technique to quantify the categorical values in a dataset. This is an essential step in constructing any ML model as the model only understands numeric values. First all the categorical columns are compiled in a dataset an then label encoding is applied in the dataset.
* Below is the code to extract all the categorical columns in a fresh dataset.



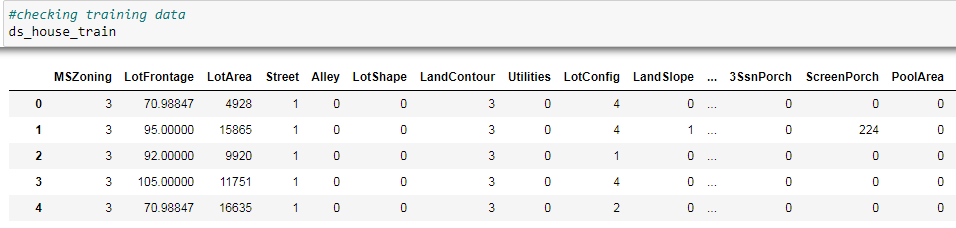
* Now we have a fresh dataset with just the categorical columns. We can now apply label encoding to quantify all the categorical columns.
* Describing all the categorical columns in both the training and testing dataset.



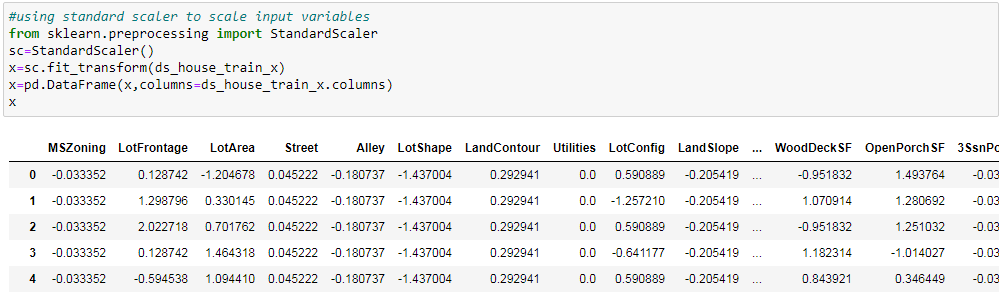
* **Applying Label Encoding**
* All the above columns will be quantified as all of them will be considered in the model. Below are the codes to quantifying the categorical elements of the dataset. The below code is just for the training dataset, the same applies to the testing dataset also.



* The above code is just for the training dataset, the same applies to the testing dataset also.
* Below is the new training dataset with quantified categorical columns.



* **Scaling the Input Variable**
* Feature scaling is a method used to normalize the range of independent variables or features of data. It is also known as data normalization.  Unscaled input variables can result in a slow or unstable learning process, whereas unscaled target variables on regression problems can result in exploding gradients causing the learning process to fail. For this particular model I will be using standard scaler. Below is the code for the same.



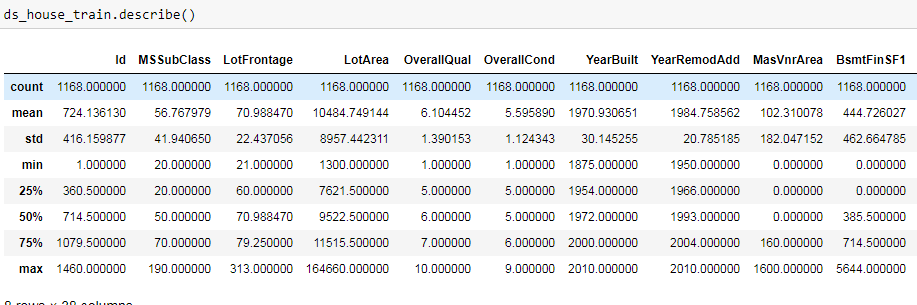
* **Hardware and Software Requirements and Tools Used**
* This project is done using Python 3.0, Ms Excel and Ms Word. The GUI (Graphical User Interface) used in this project is Jupyter Notebook. The coding part of the project is done in Jupyter Notebook while the training and testing data is received in CSV format and the project report is written in Ms Word.

**Model/s Development and Evaluation**

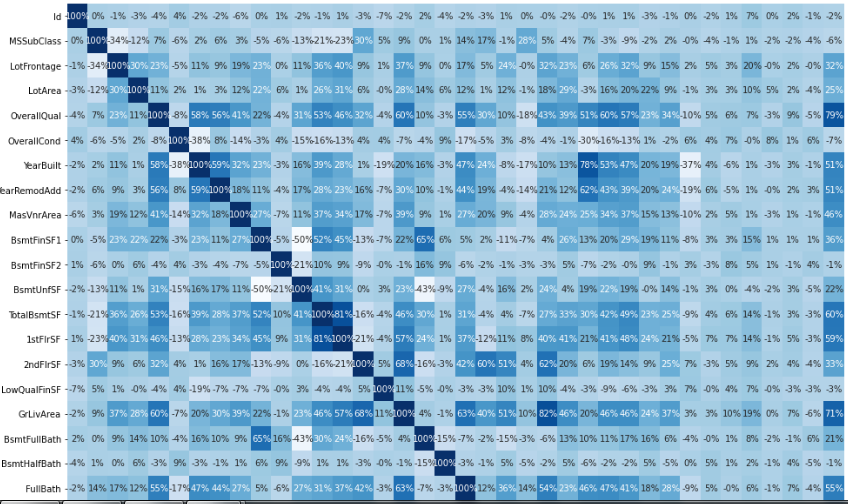
* Identification of possible problem-solving approaches (methods)
* **Descriptive Statistics**

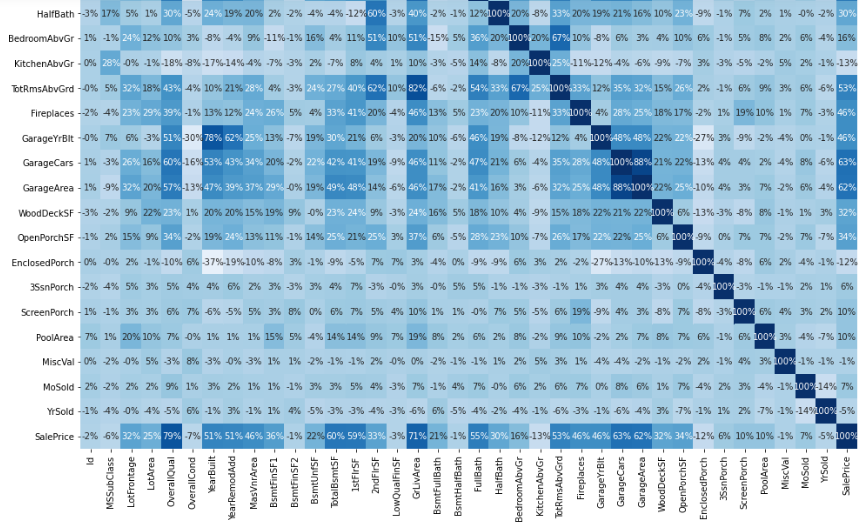
A descriptive statistic is a [summary statistic](https://en.wikipedia.org/wiki/Summary_statistic) that quantitatively describes or summarizes features from a collection of [information](https://en.wikipedia.org/wiki/Information). It is the process of using and analysing those statistics. Descriptive statistics is distinguished from [inferential statistics](https://en.wikipedia.org/wiki/Statistical_inference) by its aim to summarize a [sample](https://en.wikipedia.org/wiki/Sample_(statistics)), rather than use the data to learn about the [population](https://en.wikipedia.org/wiki/Statistical_population) that the sample of data is thought to represent.

* Various statistical measures that were laid down to understand the data in depth. There was a combination of categorical and numerical data in the dataset. Descriptive statistics were implemented in the numeric dataset to find out various insights. Measures such as count, mean, standard deviation, minimum, maximum and inter quartile ranges are found for each of the numeric variables.
* Below is the table that describes descriptive statistics.

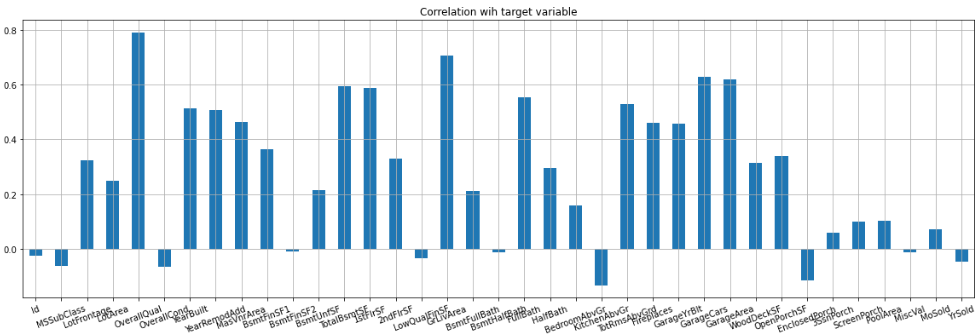


* Mean column describes the average value of the column. Std (Standard Deviation) describes how far the points are from the mean. Min and Max columns describes the minimum and maximum values in the column. The 25%, 50% and 75% are the inter quartile ranges. For the columns that have mean is greater than the 50% quartile range indicates that the data in the column is right skewed. For the columns that have mean is lesser than the 50% quartile range indicates that the data in the column is left skewed. Columns which have huge differences in between the maximum and 75% of the quartile ranges means they have outliers in them.
* **Correlation Matrix, Heatmap and Graph**
* A Correlation matrix is also plotted where we can see the variables and their co-linearity with other variables and most importantly with the target variable. Below is a partial snapshot of the correlation matrix in a heatmap showing linear relationship of variables with each other. The co relation percentage is also mentioned where it shows either the variable is positively correlated or negatively correlated with each other. The co linearity of one variable with the other is more, then the matrix gets darker in shade and vice-versa.

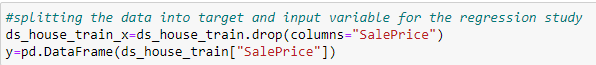




* Other than the correlation matrix, the other thing that is very useful in understanding co-linearity among variables is to plot a co-relation graph. The below co-relation graph shows the relation of all the numeric variables of the dataset with the target variable that is the sales price of the houses.

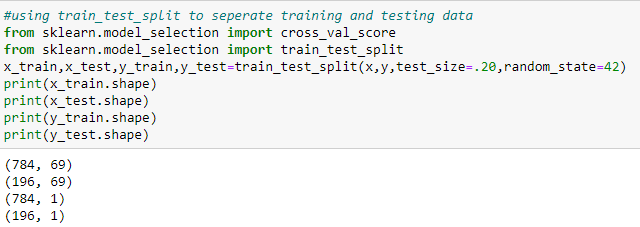


* The Y axis shows the percentage of co-linearity with the target variable Sales Price and X axis shows the name of the numeric variables in the dataset. The variables which shows lesser co-linearity will be not considered in making the model as considering them might reduce the efficiency of the model. The above diagram is for the training dataset. All the variables that shows negative correlation will be removed from both the training and testing dataset.
* **Separating Input and Target Variable**
* The target and input variables are separated from the training dataset. Below is the code to perform the same.

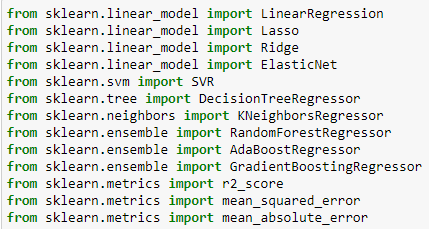


The input variables or the independent variables is considered as X and the target variable or the dependent variable is considered as Y.

* Testing of Identified Approaches (Algorithms)
* Separating training and testing data
* The entire training dataset is splitted into training and testing data. The ratio used for this study id 80-20%. 80% of the data is used to train the model the rest 20% is used for testing the model. Below is the code to do the same.

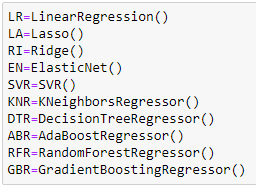


* Importing Regression Algorithms and Metrics
* Below are the codes to import the algorithms and metrics that were used in the following study. All the below algorithms and metrics are precisely for the regression study.



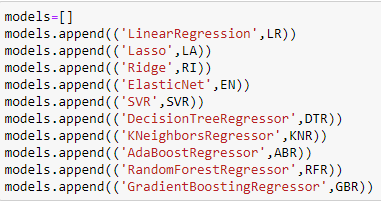
* Run and Evaluate selected models

Aliases are assigned to the algorithms. Below are the codes.



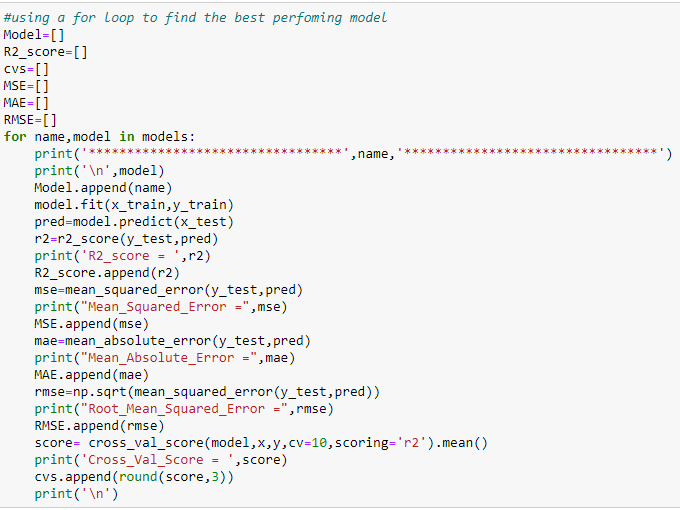
Creating Lists

A list is created for the algorithms so that we can put all of them in a loop for the training part. We are using the append function to test each algorithms with the dataset. Below is the code.



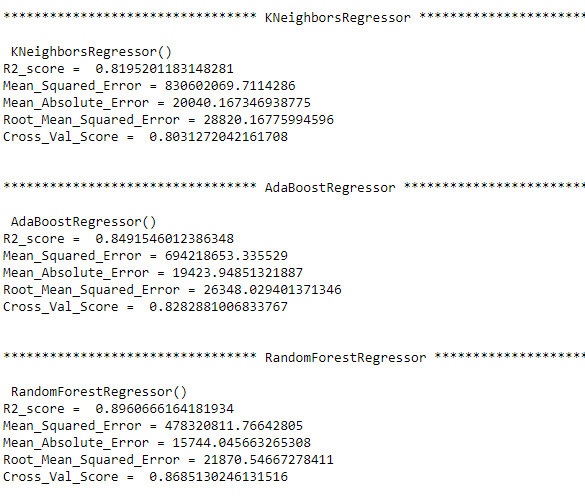
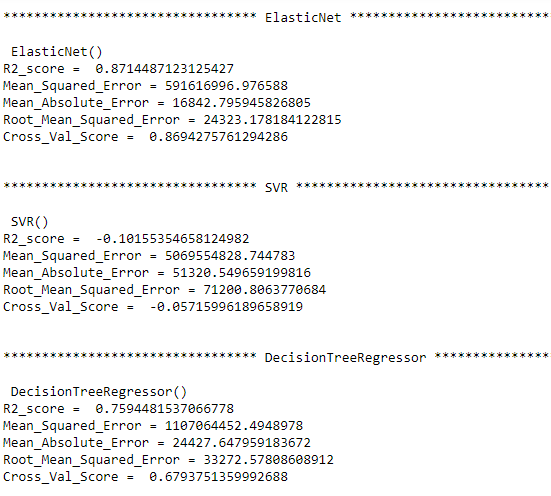
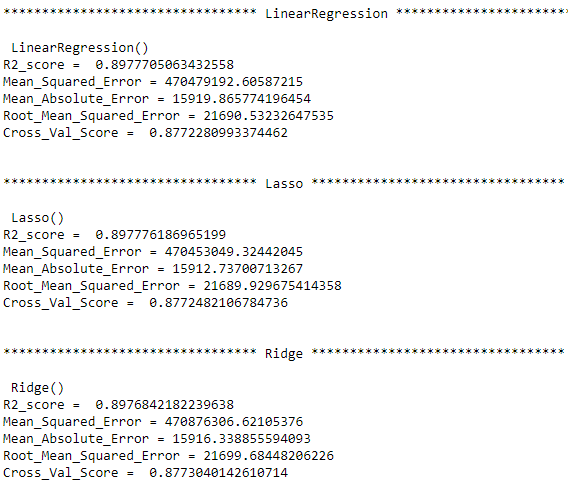
Machine learning

Now the most important part of the entire study, running the training and testing dataset through the different algorithms and using metrics to evaluate the models efficiency. I am using a loop function to test the dataset. Below is the code.

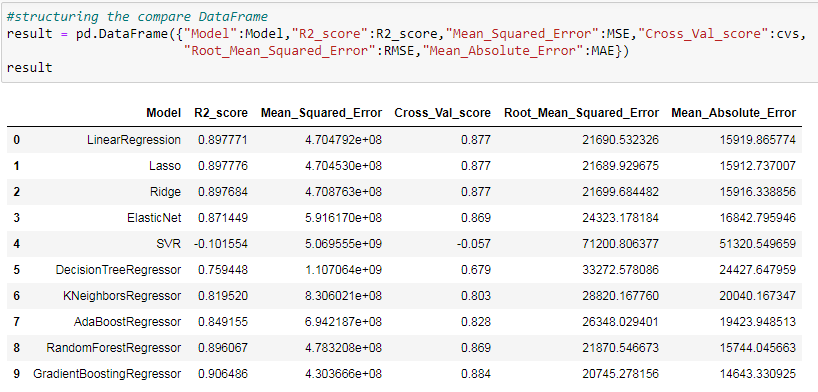


The above code shows different lists are created for the models and metrics and then a for loop is used to fit the train and test data in various algorithms and then the predicted value is compared with the test values to find the accuracy of models.

Below are the scores based on metrics of the algorithms after the training and testing the dataset.



Below is a comparison table to compare all the algorithms at a glance.



* Key Metrics for success in solving problem under consideration

Metrics are various parameters that are used to judge the efficiency and efficacy of a model. The metrics used to judge different parameters of this model are

* R2 Score
* Mean Squared Error
* Cross Validation Score
* Root Mean Squared Error
* Mean Absolute Error

R2 Score - "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Mean Squared Error - The **mean squared error** (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them.

Cross Validation Score - Cross-validation starts by shuffling the data to prevent any unintentional ordering errors and splitting it into k folds. Then k models are fit on the data training and test splits.

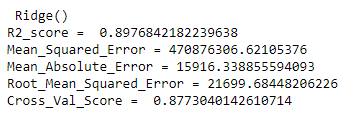
**Root Mean Square Error**(RMSE) - It is the [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [residuals](https://www.statisticshowto.com/residual/) or [prediction errors](https://www.statisticshowto.com/prediction-error-definition/). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

Mean Absolute Error- The mean absolute error of a model is the mean of the absolute values of the individual prediction errors on over all [instances](https://doi.org/10.1007/978-0-387-30164-8_406) in the [test set](https://doi.org/10.1007/978-0-387-30164-8_820). Each prediction error is the difference between the true value and the predicted value for the instance.

**Choosing the best Model**

Considering the algorithms scores in all the above parameters the Ridge algorithm has worked best with this dataset as the R2 score and Crossing Validation Score is the highest and Mean Squared Error, Mean Absolute Error and Root Mean Squared Error are the lowest.

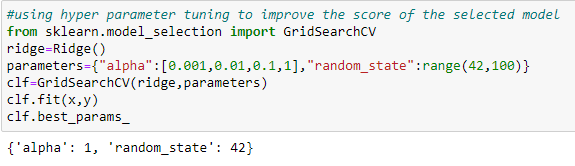
Below are the scores of the Ridge Algorithm



Hyper Parameter Tuning

In  a hyperparameter tuning the [parameter](https://en.wikipedia.org/wiki/Parameter)s are tuned to control the learning process of the model. Different datasets with specific algorithms requires different learning rate for the model to work more accurately. So it forms different combinations to get the one parameter with best result. We will be tuning the Ridge algorithm as it has the highest score.

Below is the code to perform hyper parameter tuning.



According to the above tuning the alpha or the learning rate 1 with random state 42 gives the model the highest accuracy. Now this parameters can be used while implementing the model.

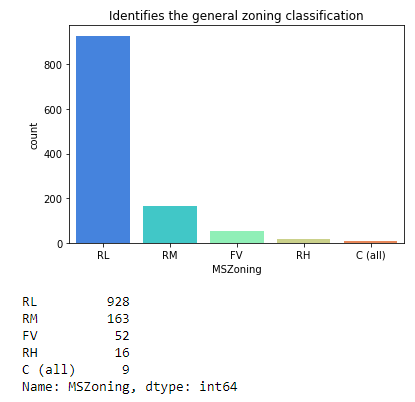
* Visualizations

Two kinds of analysis are done in order to visualize the findings of this study, they are Univariate Analysis and Bivariate Analysis. Univariate analysis is to describe the data in order to find out the patterns in the data. This is done by looking at the mean, mode, median, standard deviation, dispersion, etc. Bivariate analysis is used to find out if there is a relationship between two sets of values. It usually involves the [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)X and Y.

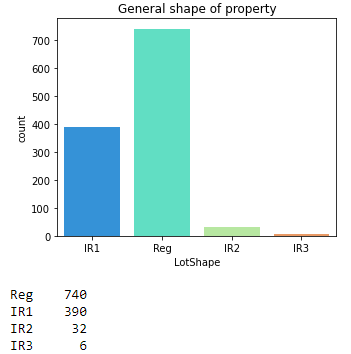
Below are the kinds of plots and graphs used to visualize the findings from the study.

* Bar Plot
* Count Plot
* Scatter Plot
* Box Plot
* Histogram
* **Univariate Analysis**

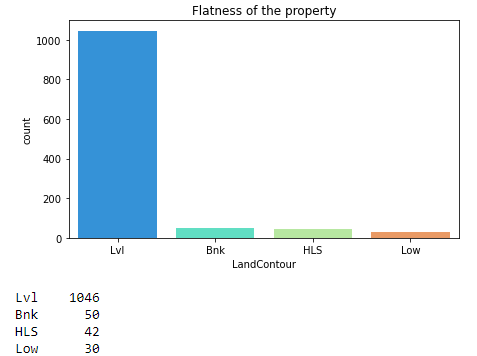
Below are some of the univariate data visualizations of the variables from the dataset.



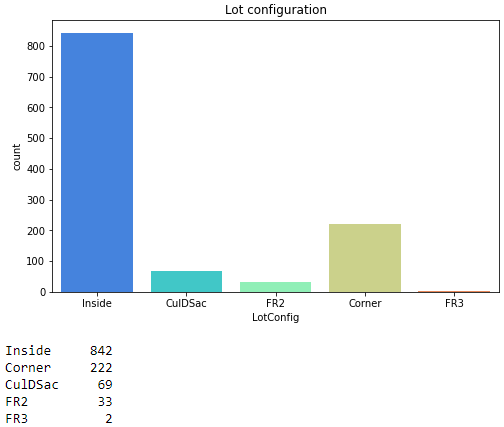
* The above count plot shows the number of houses based on General Zoning Classification (MSZoning), Most of the instances in the dataset were Residential Low Density houses and least being Commercial Houses. Below are the abbreviations
* C Commercial
* FV Floating Village Residential
* I Industrial
* RH Residential High Density
* RL Residential Low Density
* RM Residential Medium Density



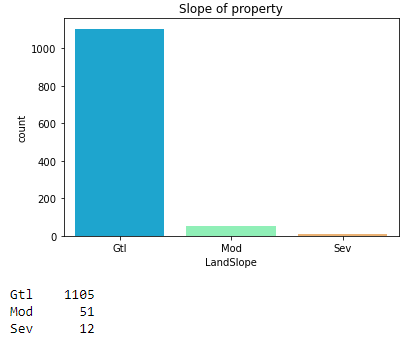
* The above countplot shows the general shape of the property. For most of the instances in the dataset, the shape of the property was regular and irregular shape being the least. Below are the abbreviations.
* Reg Regular
* IR1 Slightly irregular
* IR2 Moderately Irregular
* IR3 Irregular



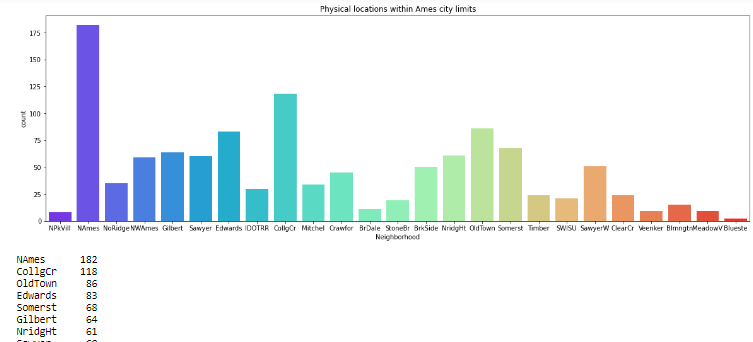
* The above countplot shows number of houses in the dataset based on Flatness of the Property. For most of the instances the property was near flat (Lvl) and number of depressed properties being the lowest.
* Lvl Near Flat/Level
* Bnk Banked - Quick and significant rise from street
* HLS Hillside - Significant slope from side to side
* Low Depression



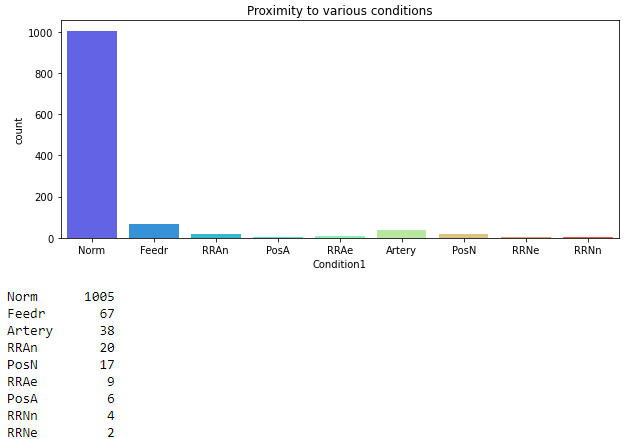
* The above countplot shows number of houses in the dataset based on Lot Configurations. For most of the instances the property was in the Inside lot and Frontage on 3 sides of property being the lowest. Below are the abbreviations
* Inside Inside lot
* Corner Corner lot
* CulDSac Cul-de-sac
* FR2 Frontage on 2 sides of property
* FR3 Frontage on 3 sides of property

****

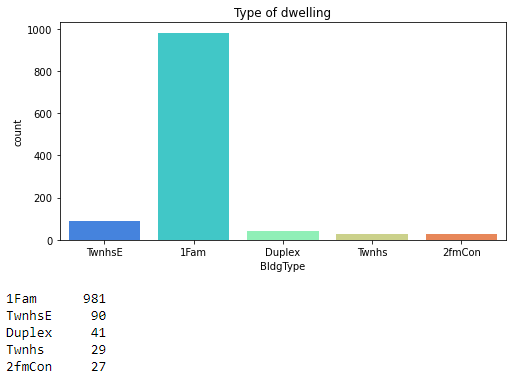
* The above countplot shows number of houses in the dataset based on Slope of Property. For most of the instances the property has a gentle slope with severe slope being the lowest. Below are the abbreviations
* Gtl Gentle slope
* Mod Moderate Slope
* Sev Severe Slope

****

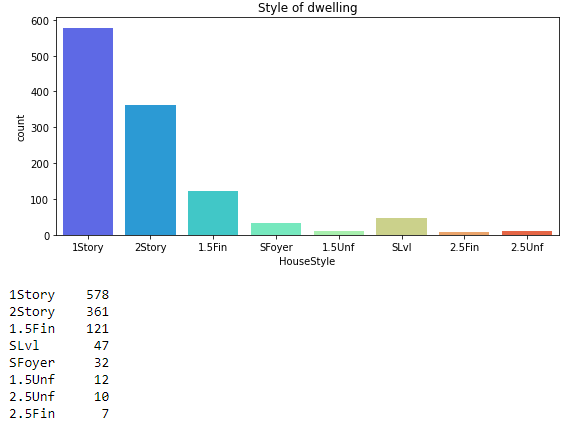
* The above countplot shows number of houses in the dataset based on location among areas in the city. For most of the instances the properties were in North Ames and properties blueste being the lowest. Below are the abbreviations
* Blmngtn Bloomington Heights
* Blueste Bluestem
* BrDale Briardale
* BrkSide Brookside
* ClearCr Clear Creek
* CollgCr College Creek
* Crawfor Crawford
* Edwards Edwards
* Gilbert Gilbert
* IDOTRR Iowa DOT and Rail Road
* MeadowV Meadow Village
* Mitchel Mitchell
* Names North Ames
* NoRidge Northridge
* NPkVill Northpark Villa
* NridgHt Northridge Heights
* NWAmes Northwest Ames

****

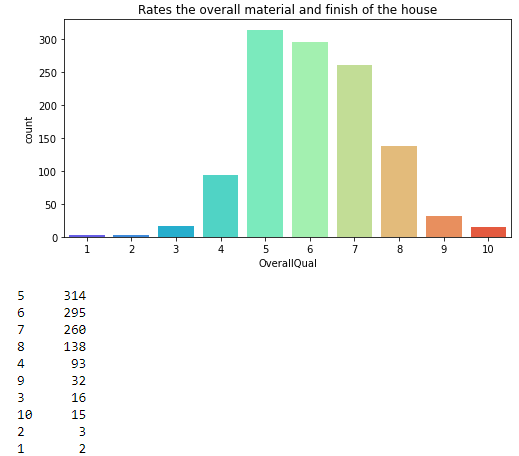
* The above countplot shows number of houses in the dataset based on proximity to various conditions. For most of the instances the properties were normal located with properties Within 200' of East-West Railroad being the lowest. Below are the abbreviations.
* Artery Adjacent to arterial street
* Feedr Adjacent to feeder street
* Norm Normal
* RRNn Within 200' of North-South Railroad
* RRAn Adjacent to North-South Railroad
* PosN Near positive off-site feature--park, greenbelt
* PosA Adjacent to postive off-site feature
* RRNe Within 200' of East-West Railroad
* RRAe Adjacent to East-West Railroad

****

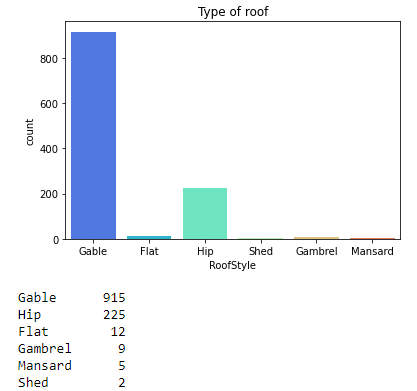
* The above countplot shows number of houses in the dataset based on their types of dwelling. For most of the instances the dwelling is Single Family Detached with Two-family Conversion being the lowest. Below are the abbreviations.
* 1Fam Single-family Detached
* 2FmCon Two-family Conversion
* Duplx Duplex
* TwnhsE Townhouse End Unit
* TwnhsI Townhouse Inside Unit

****

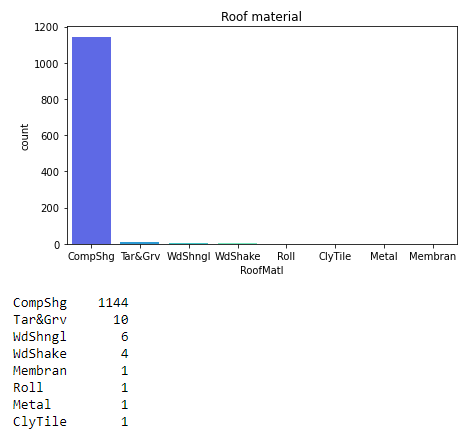
* The above countplot shows number of houses in the dataset based on their style of dwelling. For most of the instances the dwelling is One Story house with Two and one-half story: 2nd level finished being the lowest. Below are the abbreviations.
* 1Story One story
* 1.5Fin One and one-half story: 2nd level finished
* 1.5Unf One and one-half story: 2nd level unfinished
* 2Story Two story
* 2.5Fin Two and one-half story: 2nd level finished
* 2.5Unf Two and one-half story: 2nd level unfinished
* SFoyer Split Foyer
* SLvl Split Level

****

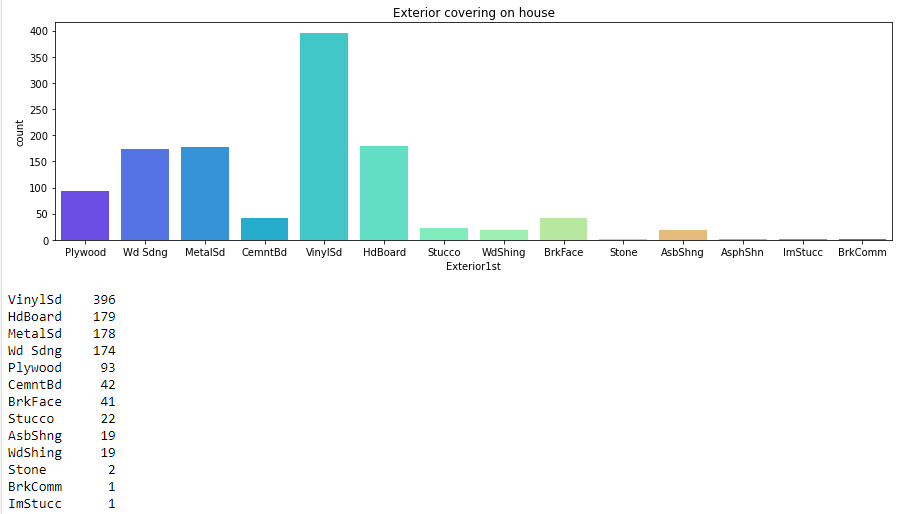
* The above countplot shows number of houses in the dataset based on overall material and finish of the house. For most of the instances the condition is rated as 5 with 1 being the lowest. Below are the abbreviations.
* 10 Very Excellent
* 9 Excellent
* 8 Very Good
* 7 Good
* 6 Above Average
* 5 Average
* 4 Below Average
* 3 Fair
* 2 Poor
* 1 Very Poor

****

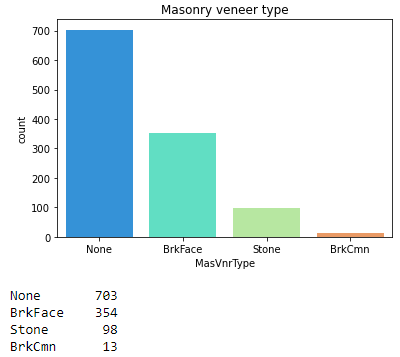
* The above countplot shows number of houses in the dataset based on type of roof. For most of the instances the house has Gable roof with Shed roofs being the lowest. Below are the abbreviations.
* Flat Flat
* Gable Gable
* Gambrel Gabrel (Barn)
* Hip Hip
* Mansard Mansard
* Shed Shed

****

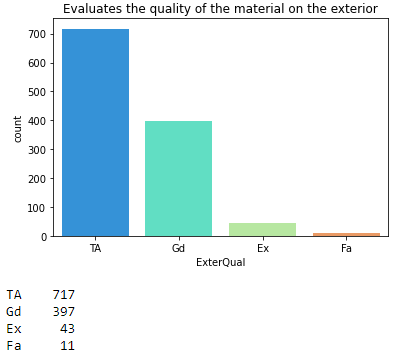
* The above countplot shows number of houses in the dataset based on the roof material. For most of the instances the house roof material made with Standard (Composite) Shingle and rest of the materials have negligible counts. Below are the abbreviations.
* ClyTile Clay or Tile
* CompShg Standard (Composite) Shingle
* Membran Membrane
* Metal Metal
* Roll Roll
* Tar&Grv Gravel & Tar
* WdShake Wood Shakes
* WdShngl Wood Shingles

****

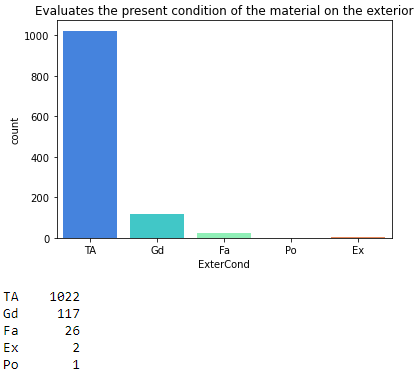
* The above countplot shows number of houses in the dataset based on the exterior covering of the house. For most of the instances the exterior covering of the house is Vinyl Siding with Imitation Stucco being the lowest. Below are the abbreviations.
* AsbShng Asbestos Shingles
* AsphShn Asphalt Shingles
* BrkComm Brick Common
* BrkFace Brick Face
* CBlock Cinder Block
* CemntBd Cement Board
* PreCast PreCast
* Stone Stone
* Stucco Stucco
* VinylSd Vinyl Siding
* WdShing Wood Shingles

****

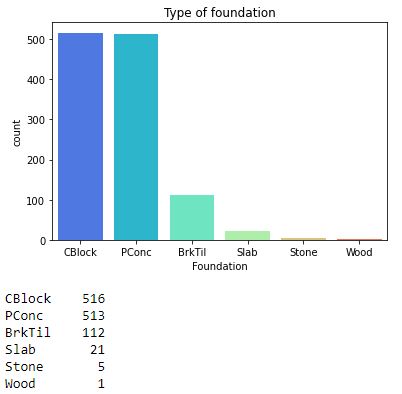
* The above countplot shows number of houses in the dataset based on the masonry veener type of the house. For most of the instances the house has Masonry Veneer type as none with Brick Common type being the lowest. Below are the abbreviations.
* BrkCmn Brick Common
* BrkFace Brick Face
* CBlock Cinder Block
* None None
* Stone Stone

****

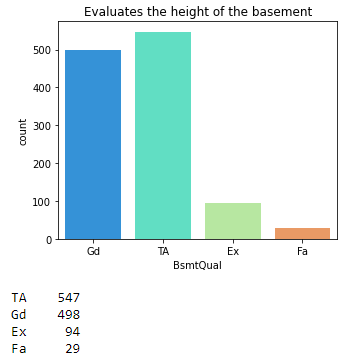
* The above countplot shows number of houses in the dataset based on the quality of material in the exterior. For most of the instances the quality of material is used is Average/Typical with Fair Quality used being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Average/Typical
* Fa Fair
* Po Poor

****

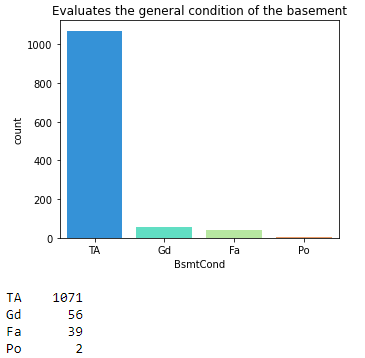
* The above countplot shows number of houses in the dataset based on the present condition of the material in the exterior. For most of the instances the present condition of the material is used is Average with Poor being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Average/Typical
* Fa Fair
* Po Poor

****

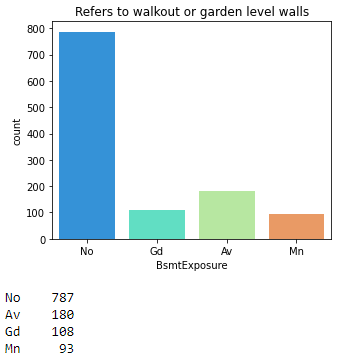
* The above countplot shows number of houses in the dataset based on the type of foundation. For most of the instances the type of foundation used is Poured Concrete with Wood foundations being the lowest. Below are the abbreviations.
* BrkTil Brick & Tile
* CBlock Cinder Block
* PConc Poured Contrete
* Slab Slab
* Stone Stone
* Wood Wood

****

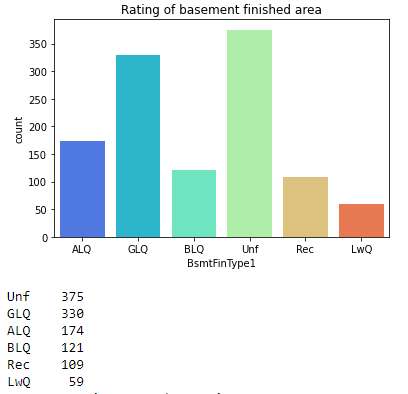
* The above countplot shows number of houses in the dataset based on the height of the basement. For most of the instances the type of foundation used is Typical (80-89 inches) with Fair (70-79 inches) basement height being the lowest. Below are the abbreviations.
* Ex Excellent (100+ inches)
* Gd Good (90-99 inches)
* TA Typical (80-89 inches)
* Fa Fair (70-79 inches)
* Po Poor (<70 inches
* NA No Basement

****

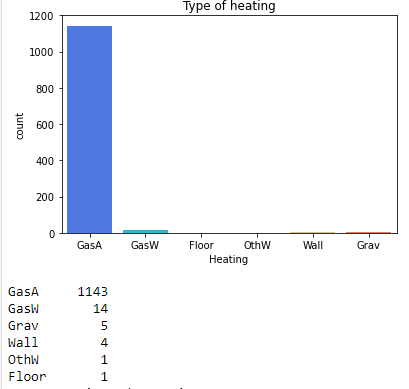
* The above countplot shows number of houses in the dataset based on general condition of the basement. For most of the instances the type of foundation used is Typical (80-89 inches) with Fair (70-79 inches) basement height being the lowest. Below are the abbreviations.
* Ex Excellent (100+ inches)
* Gd Good (90-99 inches)
* TA Typical (80-89 inches)
* Fa Fair (70-79 inches)
* Po Poor (<70 inches
* NA No Basement

****

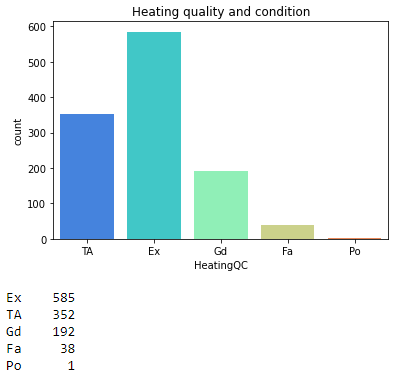
* The above countplot shows number of houses in the dataset based on walkouts or garden level walls. For most of the instances the house has no garden walls with minimum exposure being the lowest. Below are the abbreviations.
* Gd Good Exposure
* Av Average Exposure
* Mn Mimimum Exposure
* No No Exposure
* NA No Basement



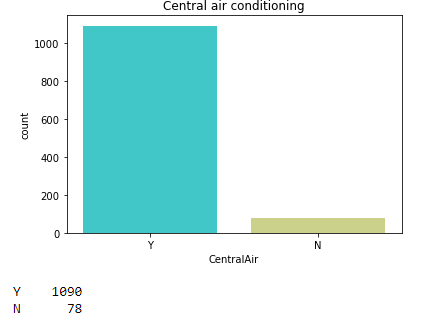
* The above countplot shows number of houses in the dataset based on quality of basement finished area . For most of the instances the house has unfinished basement area with low quality finish being the lowest. Below are the abbreviations.
* GLQ Good Living Quarters
* ALQ Average Living Quarters
* BLQ Below Average Living Quarters
* Rec Average Rec Room
* LwQ Low Quality
* Unf Unfinshed
* NA No Basement

****

* The above countplot shows number of houses in the dataset based on type of heating. For most of the instances the house has Gas forced warm air furnace with Gravity furnace being the lowest. Below are the abbreviations.
* Floor Floor Furnace
* GasA Gas forced warm air furnace
* GasW Gas hot water or steam heat
* Grav Gravity furnace
* OthW Hot water or steam heat other than gas
* Wall Wall furnace

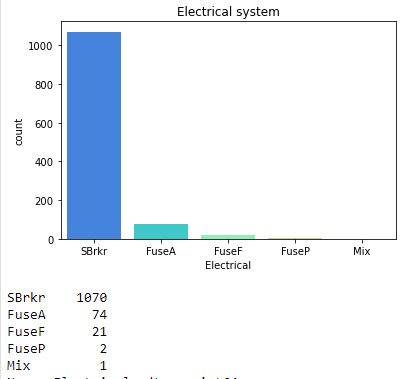
****

* The above countplot shows number of houses in the dataset based on Heating Quality and Condition. For most of the instances the house has Excellent heating quality with Poor heating quality being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Average/Typical
* Fa Fair
* Po Poor

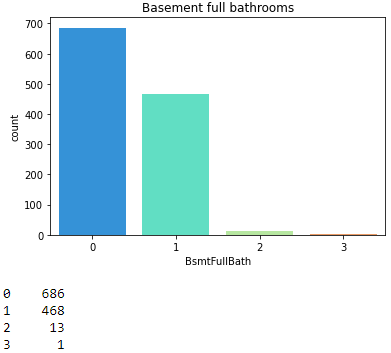
****

* The above countplot shows number of houses in the dataset that has central air condition or not. For most of the instances the houses have Central Air Condition with houses not having Central Air Conditioning being the lowest. Below are the abbreviations.

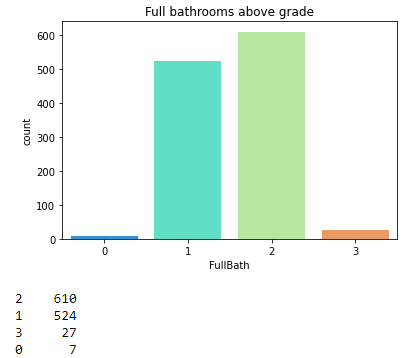
* N No
* Y Yes

****

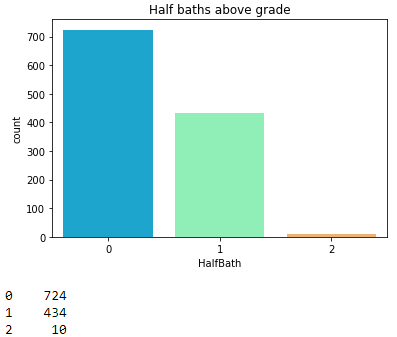
* The above countplot shows number of houses in the dataset based on their Electrical Systems. For most of the instances the house has houses have Standard Circuit Breakers & Romex with houses having Mixed Electrical Systems being the lowest. Below are the abbreviations.
* SBrkr Standard Circuit Breakers & Romex
* FuseA Fuse Box over 60 AMP and all Romex wiring (Averag
* FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
* FuseP 60 AMP Fuse Box and mostly knob & tube wiring
* Mix Mixed

****

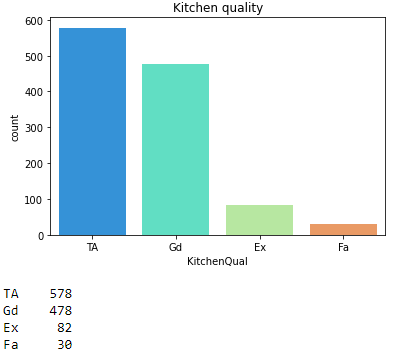
* The above countplot shows number of houses in the dataset based on the number of Basemen Full Bathrooms. For most of the instances the house has no Bathrooms in the Basement with houses having three Bathrooms being the lowest.

****

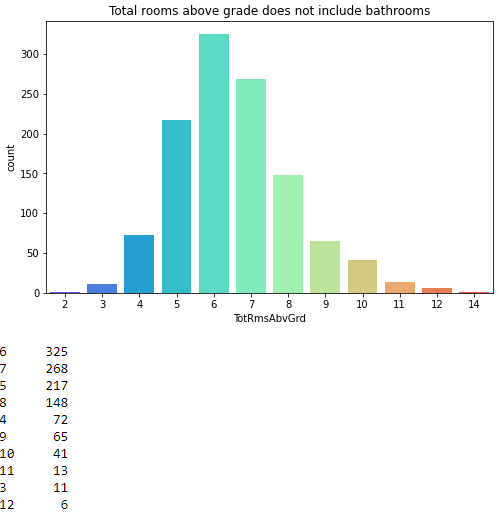
* The above countplot shows number of houses in the dataset based on the number of Full Bathrooms above grades. For most of the instances the house has two Full Bathrooms with houses having no Full Bathrooms being the lowest.

****

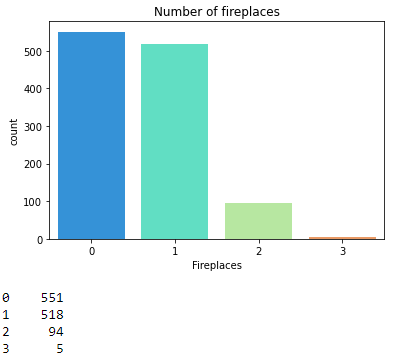
* The above countplot shows number of houses in the dataset based on the number of Half Bathrooms above grades. For most of the instances the house has no Half Bathrooms with houses having Two Half Bathrooms being the lowest.

****

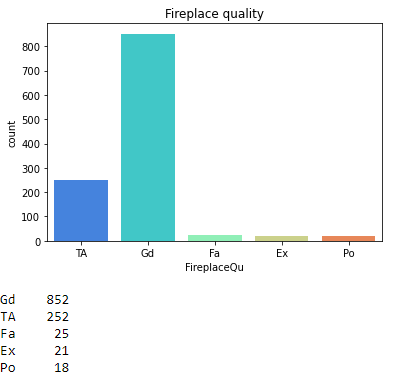
* The above countplot shows number of houses in the dataset based on the Kitchen Quality. For most of the instances the house has Typical or Average kitchen quality with houses having Fair kitchen quality being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Typical/Average
* Fa Fair
* Po Poor

****

* The above countplot shows number of houses in the dataset based on the number of total rooms in the houses. For most of the instances the house has six rooms with houses having two rooms being the lowest. Below are the abbreviations.

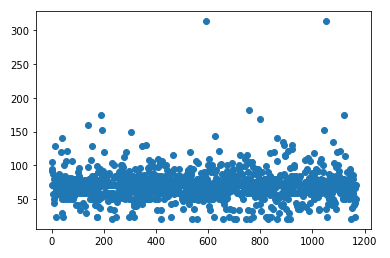
****

* The above countplot shows number of houses in the dataset based on the number of fireplaces in the houses. For most of the instances the house has no fireplaces with houses having five fireplaces being the lowest.

****

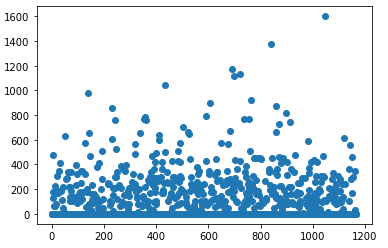
* The above countplot shows number of houses in the dataset based on the fireplace quality. For most of the instances the house has no fireplaces with houses having five fireplaces being the lowest. Below are the abbreviations
* Ex Excellent - Exceptional Masonry Fireplace
* Gd Good - Masonry Fireplace in main level
* TA Average
* Fa Fair - Prefabricated Fireplace in basement
* Po Poor - Ben Franklin Stove
* NA No Fireplace

**Plotting Linear feet of street connected to property**

****

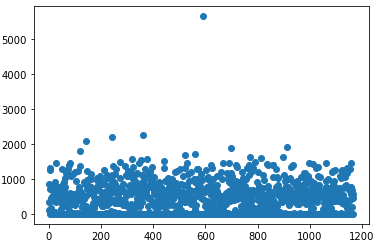
* The above scatter plot shows linear feet of Street connected to the property. From the above diagram we can see that for most of the instances the linear feet of the houses were in between 50 to 100 feet.

**Plotting Masonry veneer area in square feet**

****

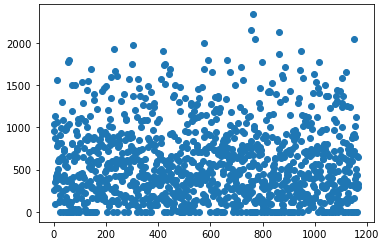
* The above scatter plot shows Masonry veneer area in square feet. From the above diagram we can see that for most of the instances the Masonry veneer area in square feet of the houses were in between 0 to 400 feet with some outliers.

**Plotting Type 1 finished square feet**

****

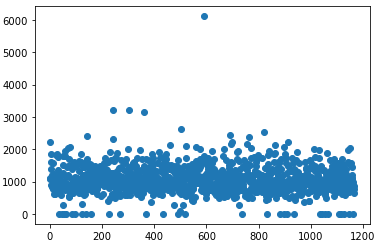
* The above scatter plot shows linear feet of Street connected to the property. From the above diagram we can see that for most of the instances the linear feet of the houses were in between 50 to 100 square feet, with some outliers.

**Plotting Unfinished square feet of basement area**



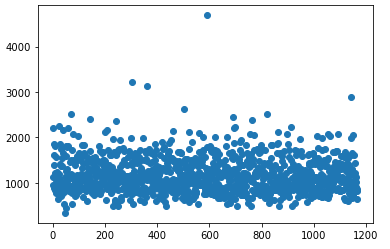
* The above scatter plot shows unfinished square feet of basement area of the houses. From the above diagram we can see that for most of the instances unfinished square feet were in between 0 to 800 square feet, and the density decreases going beyond 800 .

**Plotting Total square feet of basement area**

****

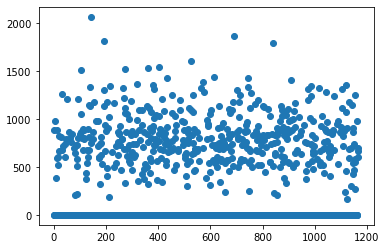
* The above scatter plot shows total square feet of basement area of the houses. From the above diagram we can see that for most of the instances total square feet of basement area were in between 500 to 2000 square feet, with some outliers displayed in the diagram .

**Plotting First Floor square feet**

****

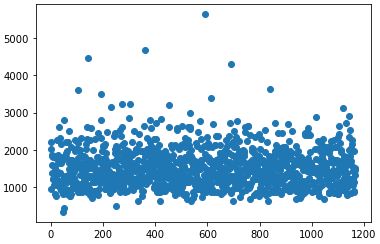
* The above scatter plot shows total square feet of first floor area of the houses. From the above diagram we can see that for most of the instances total square feet of the first floor area were in between 500 to 1800 square feet, with some outliers displayed in the diagram .

**Plotting Second floor square feet**



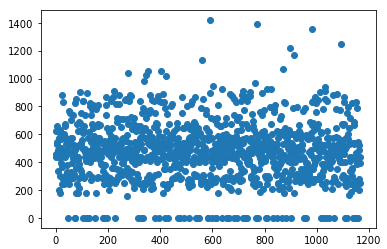
* The above scatter plot shows total square feet of second floor area of the houses. From the above diagram we can see that for most of the instances total square feet of the first floor area were in between 500 to 1000 square feet. The density of the instances reduces while it goes beyond 1000.

**Plotting Above grade (ground) living area square feet**

****

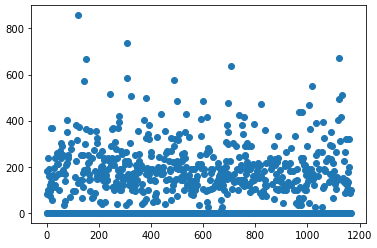
* The above scatter plot shows total square feet of living area of the houses. From the above diagram we can see that for most of the instances total square feet of the living area were in between 500 to 2000 square feet. The density of the instances reduces while it goes beyond 2000.

**Plotting Size of garage in square feet**

****

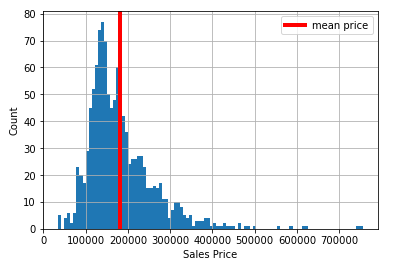
* The above scatter plot shows total square feet of garage of the houses. From the above diagram we can see that for most of the instances total square feet of the garage area were in between 200 to 850 square feet. The density of the instances reduces while it goes beyond 850.

**Plotting Wood deck area in square feet**

****

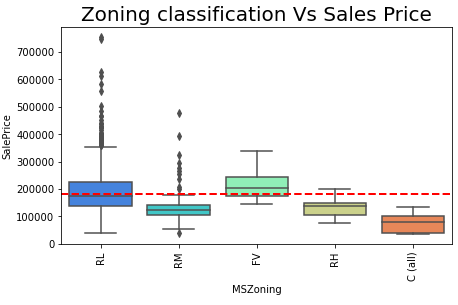
* The above scatter plot shows total square feet of wood deck area of the houses. From the above diagram we can see that for most of the instances total square feet of wood deck area were in between 50 to 300 square feet. The density of the instances reduces while it goes beyond 300.

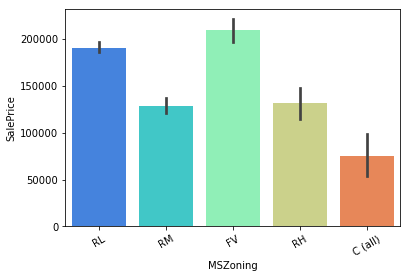
**Plotting the sales price of the houses in a histogram**

****

* The above histogram shows the price distribution of the houses in the dataset. From the above diagram we can see that the minimum price of the house is around 40000 USD and maximum price being 750000 USD. The mean or the average price of the houses is around 180000 USD.
* **Bivariate Analysis**

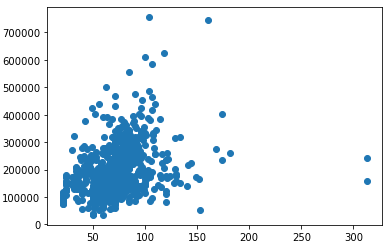
Below are some of the bivariate analysis. For most of the analysis causal relationship are shown with the target variable that is the sales price.





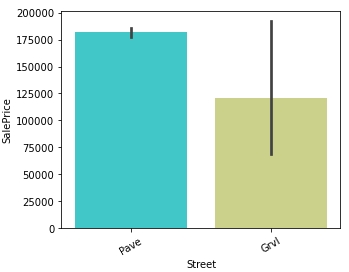
* The above box plot and the bar plot shows Sales Price based on Zoning Classification of houses. Sales price for Floating Village Residential houses are the maximum with Commercial houses being the lowest. Below are the abbreviations.
* A Agriculture
* C Commercial
* FV Floating Village Residential
* I Industrial
* RH Residential High Density
* RL Residential Low Density
* RP Residential Low Density Park
* RM Residential Medium Density

**Plotting Linear feet of street connected to property vs Sales Price**

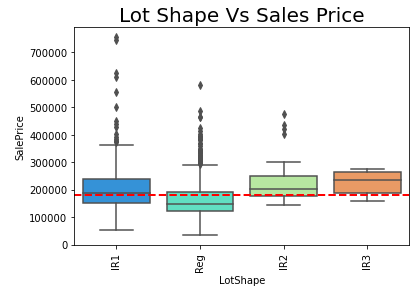
****

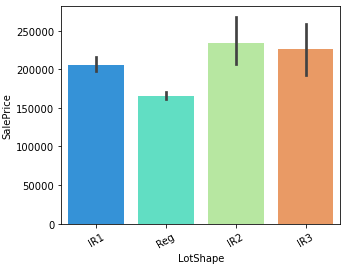
* The above scatter plot shows a somewhat linear relationship between the feet of street connected to the property to the sales price. Which the rise in linear feet of the houses the sales price also tends to increase.

**Plotting Type of road access to property vs Sales price**

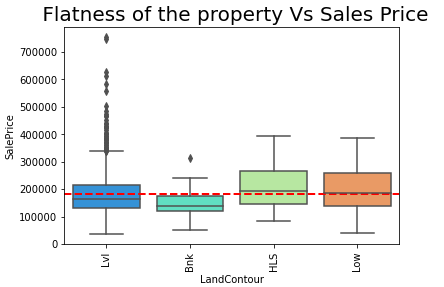
****

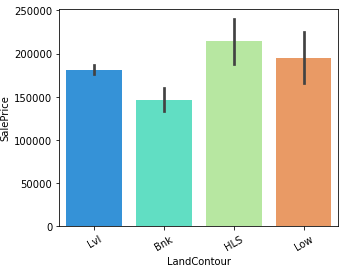
* The above bar plot shows type of road access to the property to that of the sales price. Paved road houses are costlier than gravel road houses.

****

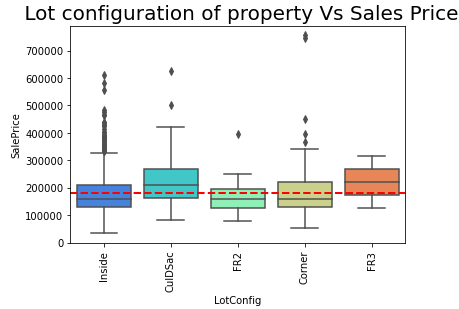
****

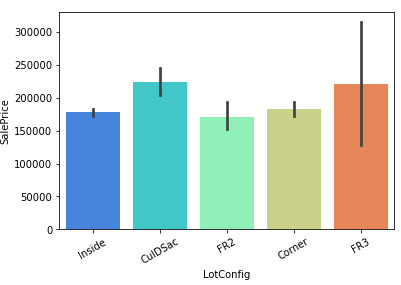
* The above box plot and the bar plot shows Sales Price based on Lot Shape of houses. Sales price for moderately irregular lot shapes are the highest with regular lot shape being the lowest. Below are the abbreviations.
* Reg Regular
* IR1 Slightly irregular
* IR2 Moderately Irregular
* IR3 Irregular

****

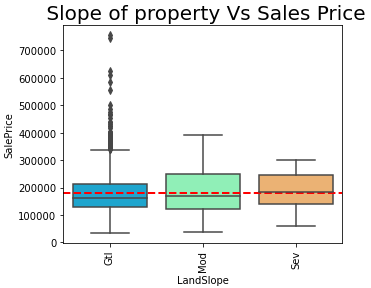
****

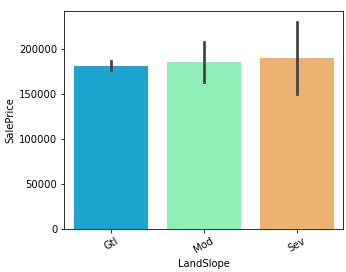
* The above box plot and the bar plot shows Sales Price based on Flatness of the Property. Sales price for Hillside - Significant slope from side to sides are the highest with Banked - Quick and significant rise from street grade to building being the lowest. Below are the abbreviations.
* Lvl Near Flat/Level
* Bnk Banked - Quick and significant rise from street
* HLS Hillside - Significant slope from side to side
* Low Depression

****

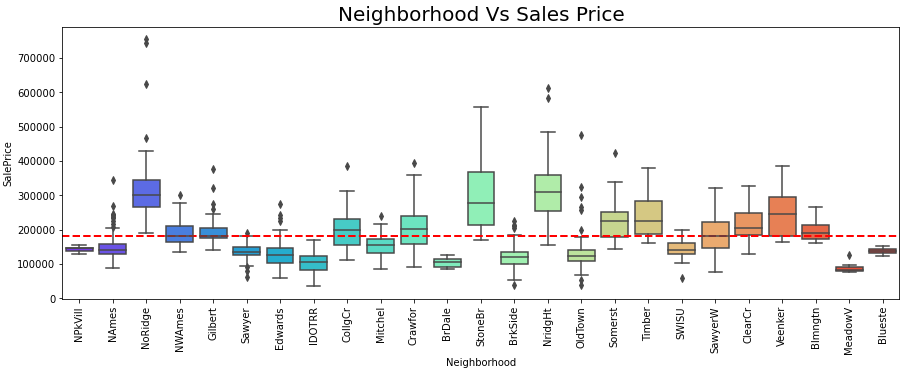
****

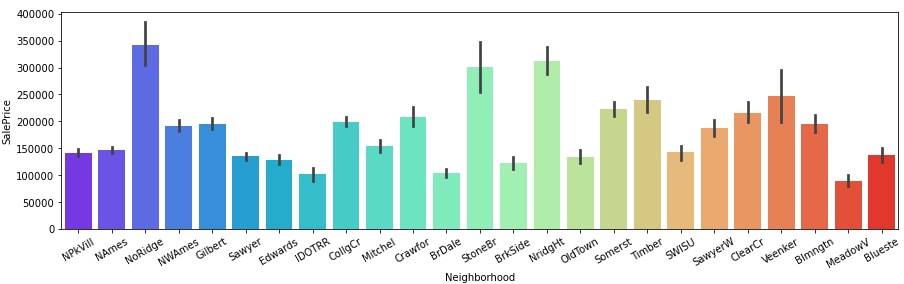
* The above box plot and the bar plot shows Sales Price based on Lot Configuration of the Property. Sales price for Cul-De-Sac properties are the highest with Frontage on 2 sides of property being the lowest. Below are the abbreviations.
* Inside Inside lot
* Corner Corner lot
* CulDSac Cul-de-sac
* FR2 Frontage on 2 sides of property
* FR3 Frontage on 3 sides of property

****

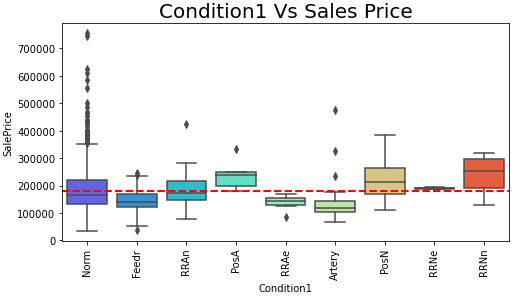
****

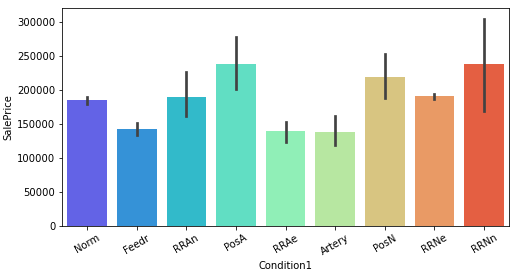
* The above box plot and the bar plot shows Sales Price based on Slope of the Property. Sales price for severely sloped properties are the highest with gentle sloped properties being the lowest. Below are the abbreviations.
* Gtl Gentle slope
* Mod Moderate Slope
* Sev Severe Slope

****

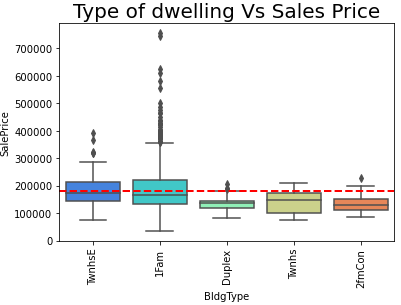
****

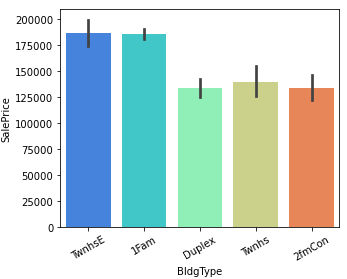
* The above box plot and the bar plot shows Sales Price based on Neighbourhood of the property. We can check the abbreviations and know how the price varies according to the neighbourhoods . Below are the abbreviations.
* Blmngtn Bloomington Heights
* Blueste Bluestem
* BrDale Briardale
* BrkSide Brookside
* ClearCr Clear Creek
* CollgCr College Creek
* Crawfor Crawford
* Edwards Edwards
* Gilbert Gilbert
* IDOTRR Iowa DOT and Rail Road
* MeadowV Meadow Village
* Mitchel Mitchell
* Names North Ames
* NoRidge Northridge
* NPkVill Northpark Villa
* NridgHt Northridge Heights
* NWAmes Northwest Ames
* OldTown Old Town
* SWISU South & West of Iowa State University
* Sawyer Sawyer
* SawyerW Sawyer West
* Somerst Somerset
* StoneBr Stone Brook
* Timber Timberland
* Veenker Veenker

****

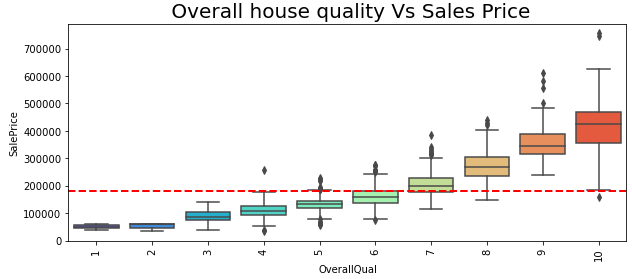
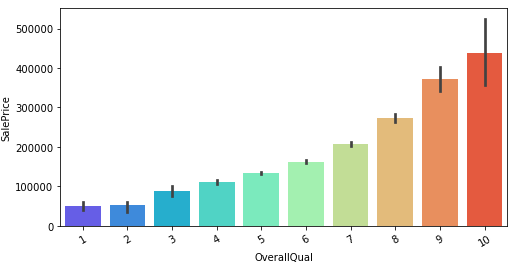
****

* The above box plot and the bar plot shows Sales Price based on Proximity to various conditions. We can check the abbreviations and know how the price varies according to the conditions . Below are the abbreviations.
* Artery Adjacent to arterial street
* Feedr Adjacent to feeder street
* Norm Normal
* RRNn Within 200' of North-South Railroad
* RRAn Adjacent to North-South Railroad
* PosN Near positive off-site feature--park, greenbelt, etc.
* PosA Adjacent to postive off-site feature
* RRNe Within 200' of East-West Railroad
* RRAe Adjacent to East-West Railroad

****

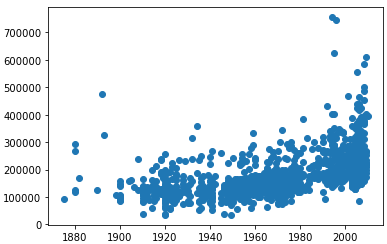
****

* The above box plot and the bar plot shows Sales Price based on type of dwellings. We can see that Sales Price for Townhouse End Units are the highest with Duplex properties being the lowest. Below are the abbreviations.
* 1Fam Single-family Detached
* 2FmCon Two-family Conversion
* Duplx Duplex
* TwnhsE Townhouse End Unit
* TwnhsI Townhouse Inside Unit

**** 

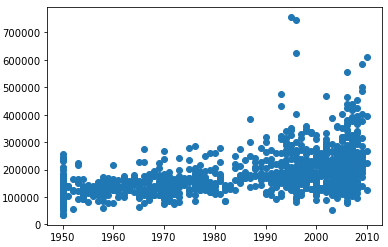
* The above box plot and the bar plot shows Sales Price based on overall quality of the house. We can see that Sales Price tends to increase with the overall ratings of the house with 10 being the highest sales price and 1 being the lowest.

**Plotting Year Built vs Sales Price**

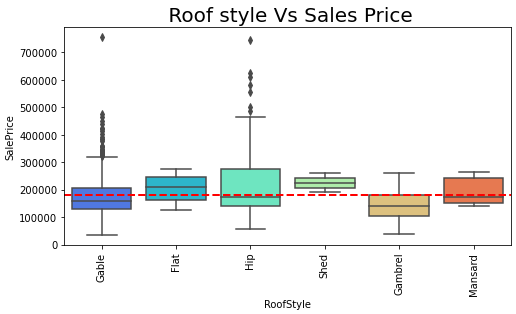
****

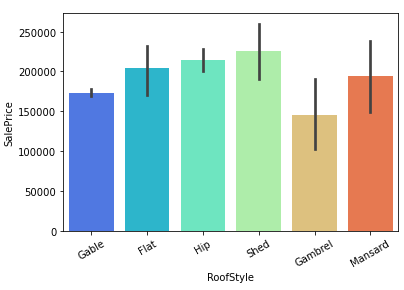
* The above scatter plot shows Sales prices plotted across year built of the houses. Price tends to increase in a gradual manner as the year increases. We can also see that for most of the instances the houses were built from 1950s to 2000 and beyond.

**Plotting Remodelling Year vs Sales Price**

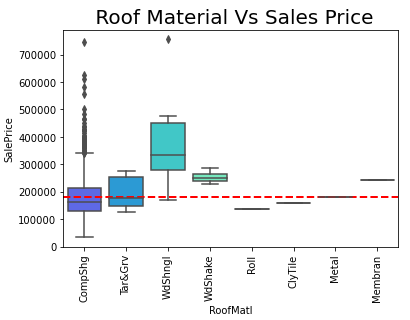


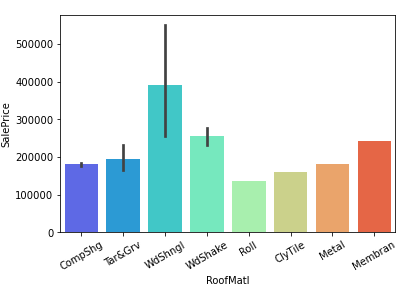
* The above scatter plot shows Sales prices plotted across remodelling date of the houses. Price tends to be almost same across all the years with a little increase in prices after the 2000s We can also see that for most of the instances the houses were remodelled after 2000s and beyond.



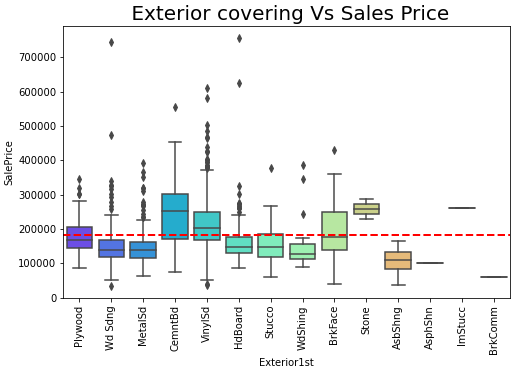


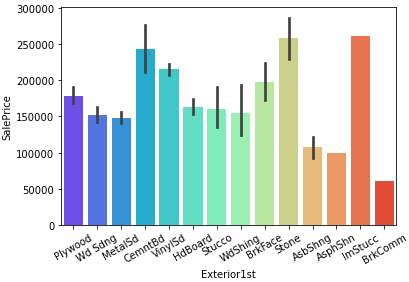
* The above box plot and the bar plot shows Sales Price based on roof styles of the houses. Sales price for shed roof properties are the highest with gabrel roof properties being the lowest. Below are the abbreviations.
* Flat Flat
* Gable Gable
* Gambrel Gabrel (Barn)
* Hip Hip
* Mansard Mansard
* Shed Shed



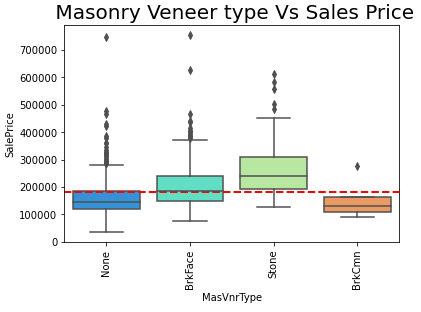


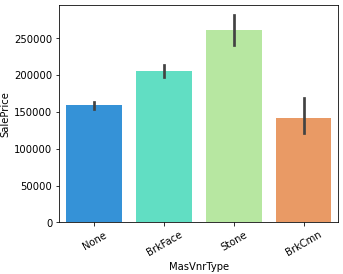
* The above box plot and the bar plot shows Sales Price based on roof material of the houses. Sales price for Wood Shingles roof material properties are the highest with roll roof material properties being the lowest. Below are the abbreviations.
* ClyTile Clay or Tile
* CompShg Standard (Composite) Shingle
* Membran Membrane
* Metal Metal
* Roll Roll
* Tar&Grv Gravel & Tar
* WdShake Wood Shakes
* WdShngl Wood Shingles

****

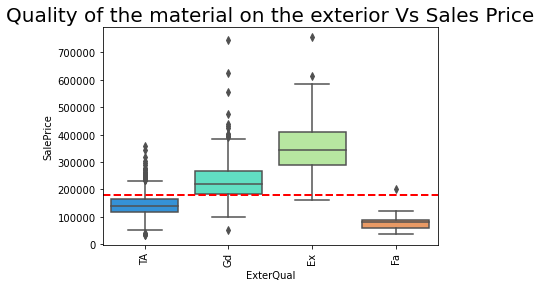
****

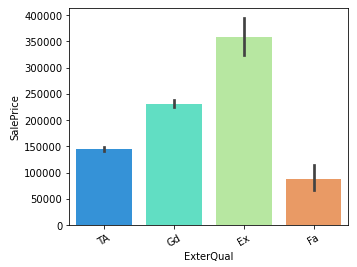
* The above box plot and the bar plot shows Sales Price based on exterior covering of the houses. Sales price for stone exterior covering are the highest with Brick Common exteriors being the lowest. Below are the abbreviations.
* AsbShng Asbestos Shingles
* AsphShn Asphalt Shingles
* BrkComm Brick Common
* BrkFace Brick Face
* CBlock Cinder Block
* CemntBd Cement Board
* HdBoard Hard Board
* ImStucc Imitation Stucco
* MetalSd Metal Siding
* Other Other

****

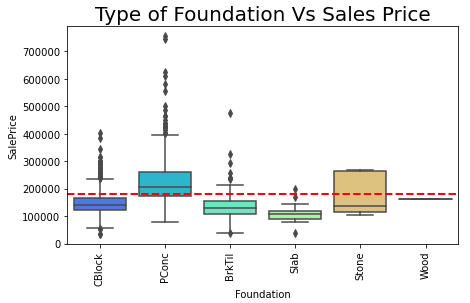
****

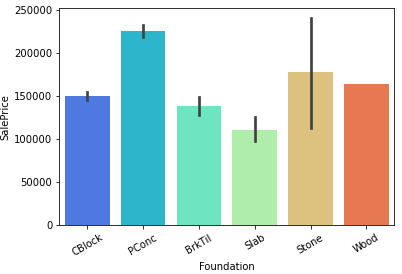
* The above box plot and the bar plot shows Sales Price based on mansory veneer type of the houses. Sales price for stone exterior covering are the highest with Brick Common exteriors being the lowest. Below are the abbreviations.
* BrkCmn Brick Common
* BrkFace Brick Face
* CBlock Cinder Block
* None None
* Stone Stone

****

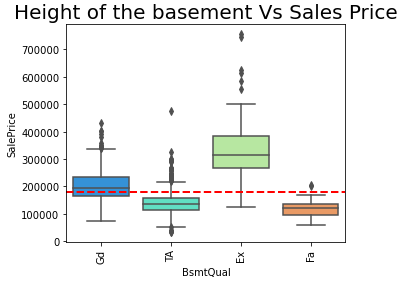
****

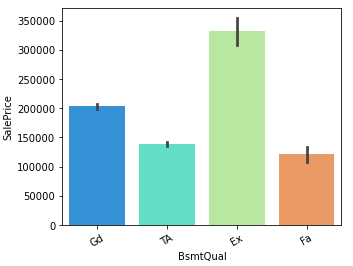
* The above box plot and the bar plot shows Sales Price based on exterior quality of the material of the houses. Sales price for excellent exterior quality are the highest with Fair quality exteriors being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Average/Typical
* Fa Fair
* Po Poor

****

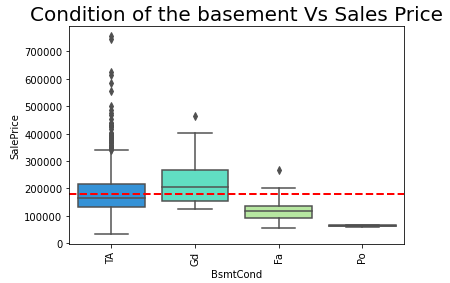
****

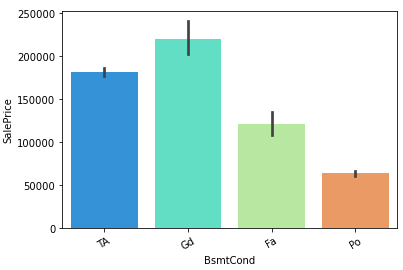
* The above box plot and the bar plot shows Sales Price based on exterior type of foundation of the houses. Sales price for poured concrete foundations are the highest with Slab foundations being the lowest. Below are the abbreviations.
* BrkTil Brick & Tile
* CBlock Cinder Block
* PConc Poured Concrete
* Slab Slab
* Stone Stone
* Wood Wood

****

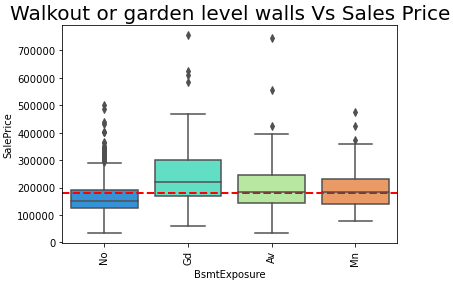
****

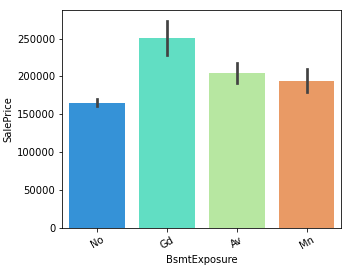
* The above box plot and the bar plot shows Sales Price based on height of the basement of the houses. Sales price for excellent house basement are the highest with fair height basements being the lowest. Below are the abbreviations.
* Ex Excellent (100+ inches)
* Gd Good (90-99 inches)
* TA Typical (80-89 inches)
* Fa Fair (70-79 inches)
* Po Poor (<70 inches
* NA No Basement

****

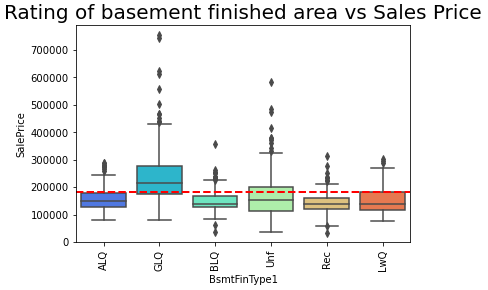
****

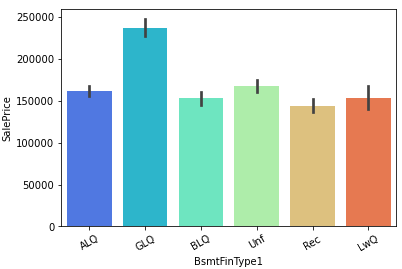
* The above box plot and the bar plot shows Sales Price based on condition of the basement of the houses. Sales price for good condition basement are the highest with poor basements condition being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Typical - slight dampness allowed
* Fa Fair - dampness or some cracking or settling
* Po Poor - Severe cracking, settling, or wetness
* NA No Basement

****

****

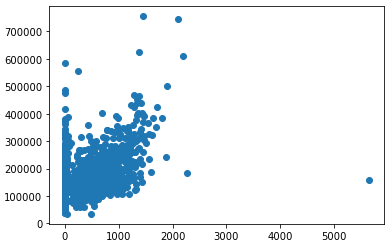
* The above box plot and the bar plot shows Sales Price based on walkouts or garden level walls of the houses. Sales price for good basement exposure houses are the highest with no basements exposures being the lowest. Below are the abbreviations.
* Gd Good Exposure
* Av Average Exposure
* Mn Mimimum Exposure
* No No Exposure
* NA No Basement

****

****

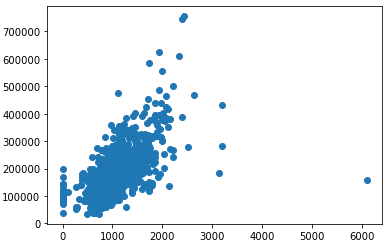
* The above box plot and the bar plot shows Sales Price based on rating of basement finished area of the houses. Sales price for good living quarter houses are the highest with Below Average Living Quarters being the lowest. Below are the abbreviations.
* GLQ Good Living Quarters
* ALQ Average Living Quarters
* BLQ Below Average Living Quarters
* Rec Average Rec Room
* LwQ Low Quality
* Unf Unfinshed
* NA No Basement

**Type 1 finished square feet vs Sales Price**

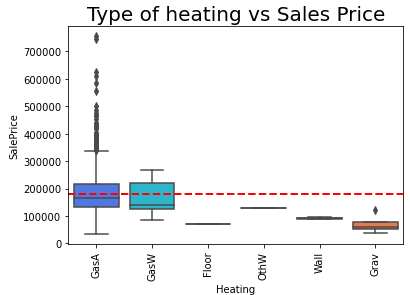
****

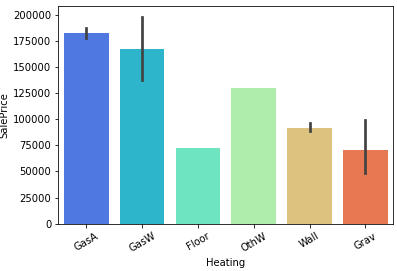
* The above scatter plot shows Sales prices plotted across type 1 finished square feet of the houses. Price tends to increase with the increase in type 1 square feet of the house. We can also see that for most of the instances the square feet ranges from 0 to 1500.

**Total square feet of basement area vs Sales Price**

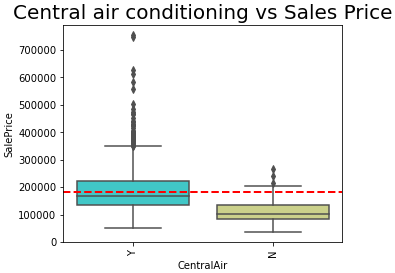
****

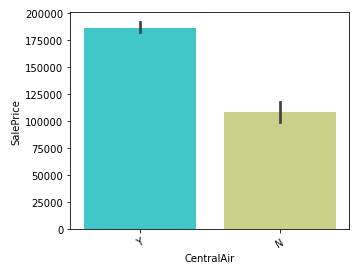
* The above scatter plot shows Sales prices plotted across type square feet of the basement area. Price tends to increase with the increase in square feet of the basement area. We can also see that for most of the instances the square feet ranges from 0 to 2000.

****

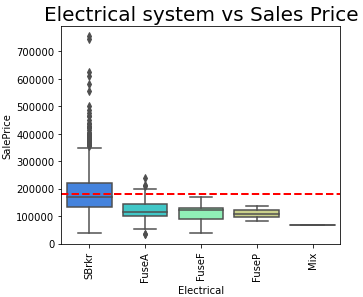
****

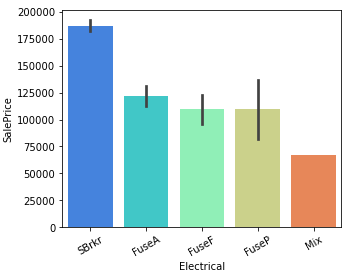
* The above box plot and the bar plot shows Sales Price based on type of heating of the houses. Sales price for Gas forced warm air furnace are the highest with floor furnace type of heating being the lowest. Below are the abbreviations.
* Floor Floor Furnace
* GasA Gas forced warm air furnace
* GasW Gas hot water or steam heat
* Grav Gravity furnace
* OthW Hot water or steam heat other than gas
* Wall Wall furnace

****

****

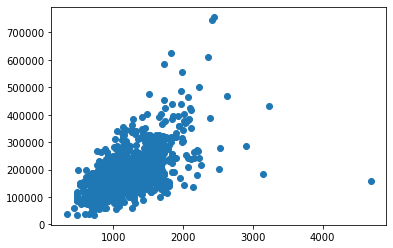
* The above box plot and the bar plot shows Sales Price based on central air conditioning of the houses. Sales price for houses with air conditioning are more than houses without it.

****

****

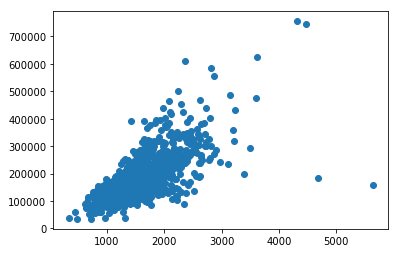
* The above box plot and the bar plot shows Sales Price based on Electrical system of the houses. Sales price for Standard Circuit Breakers & Romex are the highest with mixed electrical system being the lowest. Below are the abbreviations.
* SBrkr Standard Circuit Breakers & Romex
* FuseA Fuse Box over 60 AMP and all Romex wiring
* FuseF 60 AMP Fuse Box and mostly Romex wiring
* FuseP 60 AMP Fuse Box and mostly knob & tube wiring
* Mix Mixed

**First Floor square feet vs Sales Price**

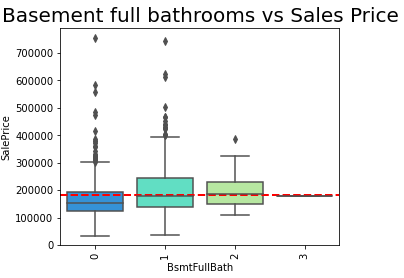
****

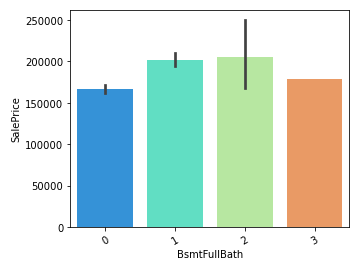
* The above scatter plot shows Sales prices plotted across first floor square feet of the houses. Price tends to increase with the increase in first floor square feet of the house. We can also see that for most of the instances the square feet ranges from 100 to 2000.

S**econd floor square feet vs Sales Price**

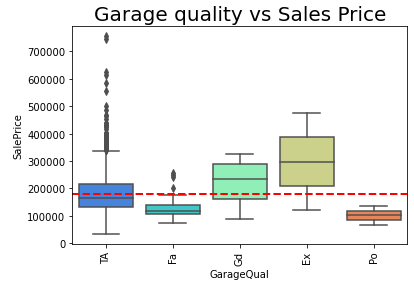
****

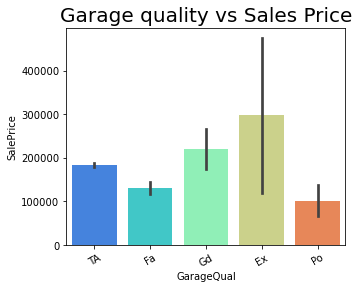
* The above scatter plot shows Sales prices plotted across second floor square feet of the houses. Price tends to increase with the increase in second floor square feet of the house. We can also see that for most of the instances the square feet ranges from 100 to 3000.

****

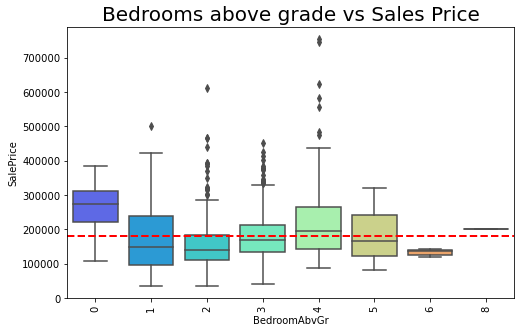
****

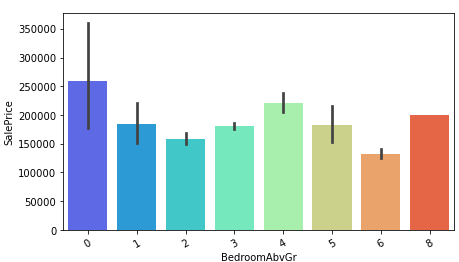
* The above box plot and the bar plot shows Sales Price based on number full bathrooms in the basement of the houses. Sales price of houses with two bathrooms are the highest with houses with no bathrooms being the lowest.

****

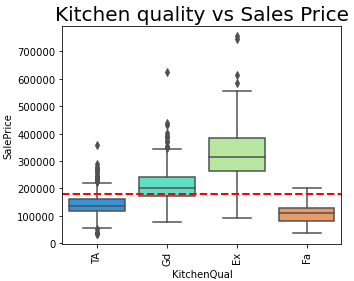
****

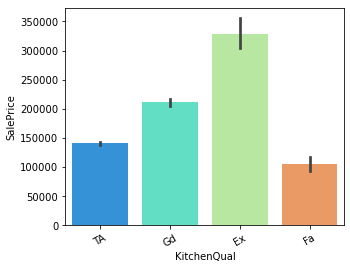
* The above box plot and the bar plot shows Sales Price based on garage quality of the houses. Sales price of houses with excellent garage quality are the highest with houses with poor garage quality being the lowest.
* Ex Excellent
* Gd Good
* TA Typical/Average
* Fa Fair
* Po Poor
* NA No Garage

****

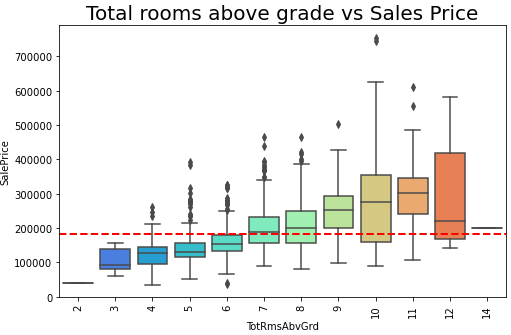
****

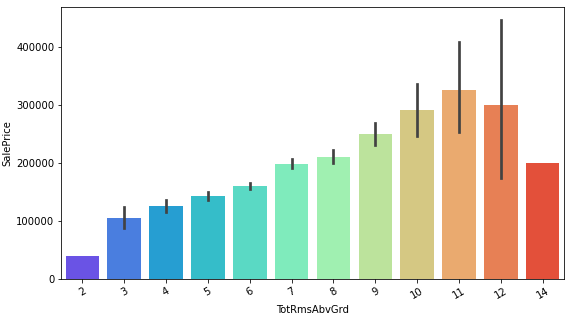
* The above box plot and the bar plot shows Sales Price based on number of bedrooms of the houses. Sales price of houses with 4 bedrooms are the highest with houses with 2 bedrooms being the lowest.

****

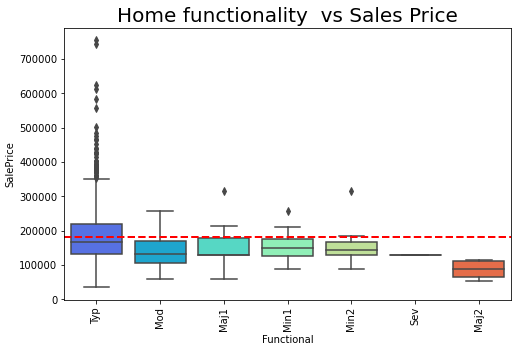
****

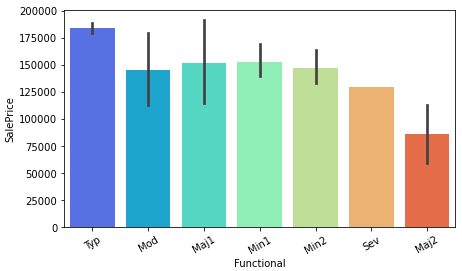
* The above box plot and the bar plot shows Sales Price based on number of kitchen quality of the houses. Sales price of houses with excellent kitchen quality are the highest with houses with fair kitchen quality being the lowest. Below are the abbreviations.
* Ex Excellent
* Gd Good
* TA Typical/Average
* Fa Fair
* Po Poor

****

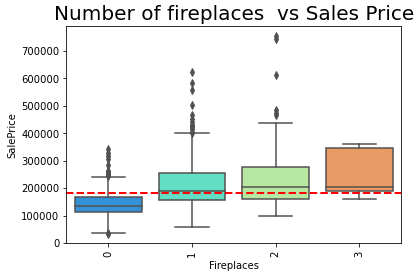
****

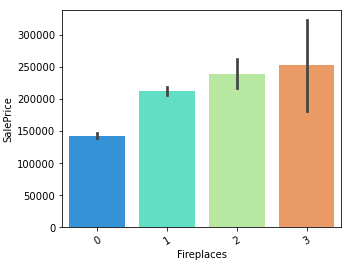
* The above box plot and the bar plot shows Sales Price based on total number of rooms in the houses. Sales price of houses with 11 rooms are the highest with houses 2 rooms being the lowest.

****

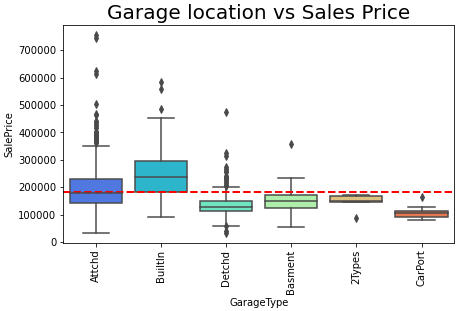
****

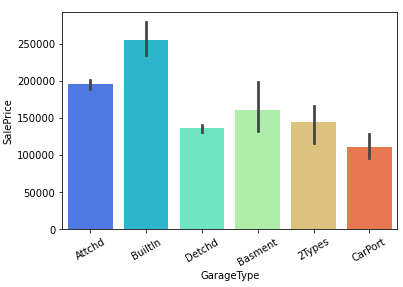
* The above box plot and the bar plot shows Sales Price based on total number home functionality of the houses. Sales price of Typical Functionality houses are the highest with Major Deductions 2 houses being the lowest. Below are the abbreviations.
* Typ Typical Functionality
* Min1 Minor Deductions 1
* Min2 Minor Deductions 2
* Mod Moderate Deductions
* Maj1 Major Deductions 1
* Maj2 Major Deductions 2
* Sev Severely Damaged
* Sal Salvage only

****

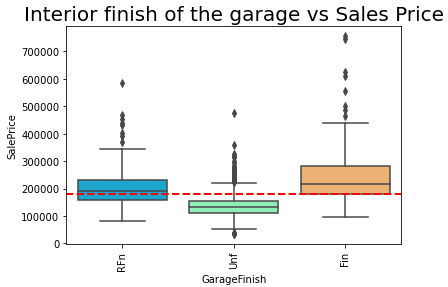
****

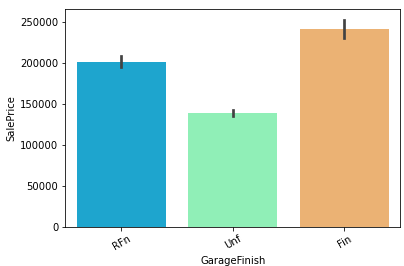
* The above box plot and the bar plot shows Sales Price based on number of fireplaces of the houses. Sales price of houses with 3 fireplaces are the highest with houses having no fireplaces being the lowest.

****

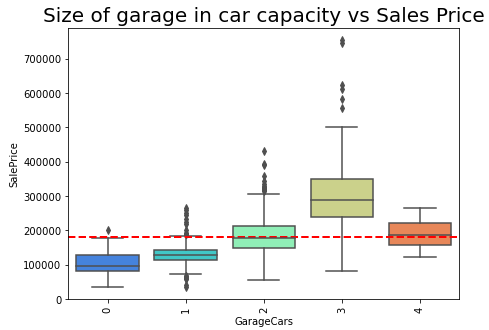
****

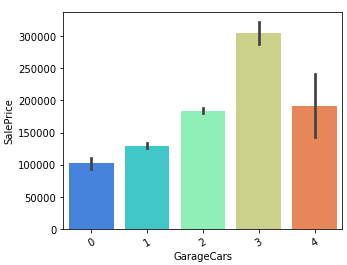
* The above box plot and the bar plot shows Sales Price based on garage location of the houses. Sales price of houses Built-In garages are the highest with houses having detached garages being the lowest. Below are the abbreviations.
* 2Types More than one type of garage
* Attchd Attached to home
* Basment Basement Garage
* BuiltIn Built-In
* CarPort Car Port
* Detchd Detached from home
* NA No Garag

****

****

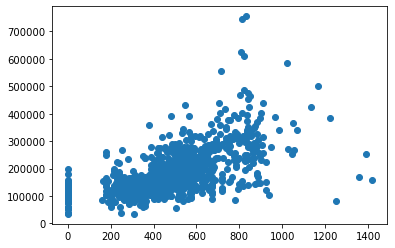
* The above box plot and the bar plot shows Sales Price based on interior finish of the garage of the houses. Sales price of houses with finished garages are the highest with houses having unfinished garages being the lowest. Below are the abbreviations.
* Fin Finished
* RFn Rough Finished
* Unf Unfinished
* NA No Garage

****

****

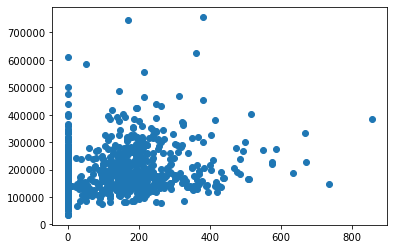
* The above box plot and the bar plot shows Sales Price based on number of car capacity of the garage of the houses. Sales price of houses with garages that has capacity of 3 cars are the highest with houses having garage capacity of 1 car being the lowest.

**Size of garage in square feet vs Sales Price**

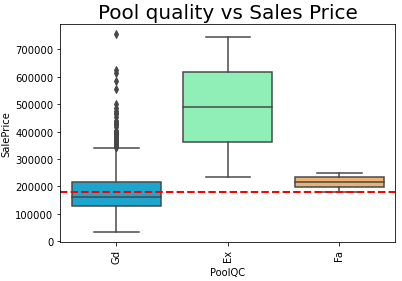
****

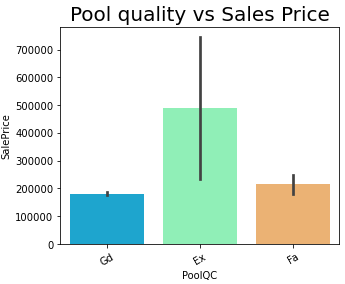
* The above scatter plot shows Sales prices plotted across size of garage in square feet. Price tends to increase with the increase in square feet of the garage. We can also see that for most of the instances the square feet ranges from 200 to 900.

**Wood deck area in square feet vs Sales Price**

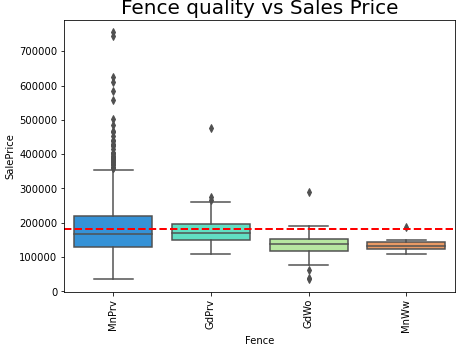
****

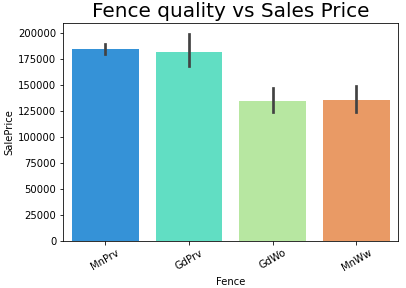
* The above scatter plot shows Sales prices plotted across wood deck area in square feet. Price tends to increase slowly with the increase in square feet of the wood deck area. We can also see that for most of the instances the square feet ranges from 200 and the density decreases going beyond.

****

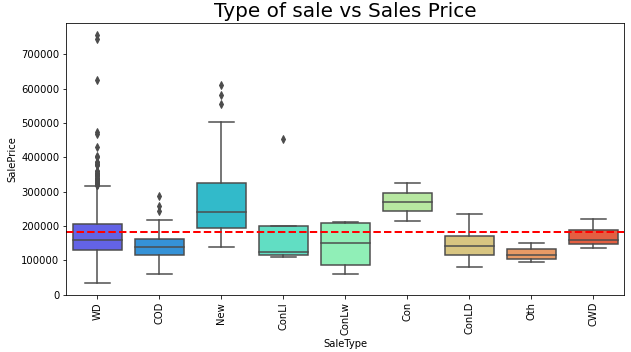
****

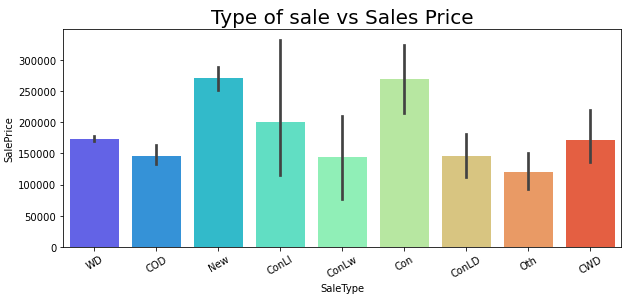
* The above box plot and the bar plot shows Sales Price based on pool quality of the houses. Sales price of houses with excellent pool quality are the highest with houses having fair pool quality being the lowest.
* Ex Excellent
* Gd Good
* TA Average/Typical
* Fa Fair
* NA No Pool

****

****

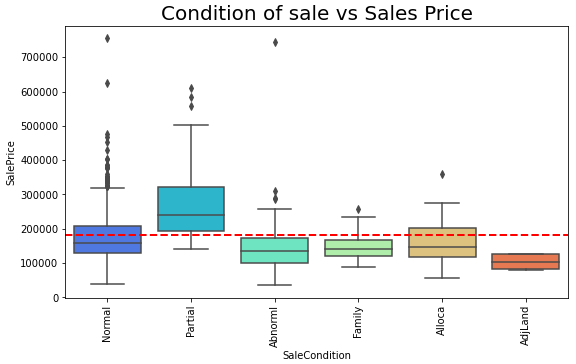
* The above box plot and the bar plot shows Sales Price based on fence quality of the houses. Sales price of houses with minimum privacy fences are the highest with houses having good wood fence being the lowest.
* GdPrv Good Privacy
* MnPrv Minimum Privacy
* GdWo Good Wood
* MnWw Minimum Wood/Wire
* NA No Fence

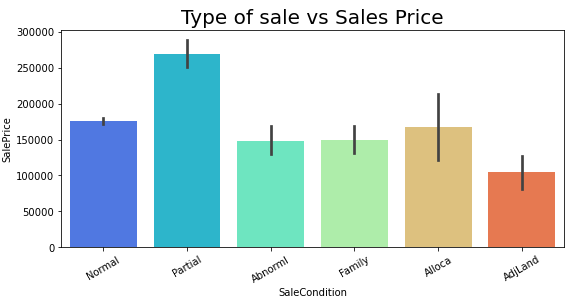
****

****

* The above box plot and the bar plot shows Sales Price based on type of sale of the houses. Sales price of houses just constructed and sold are the highest with houses having Contract Low Down payment and low interest being the lowest. Below are the abbreviations.

* WD Warranty Deed - Conventional
* CWD Warranty Deed - Cash
* VWD Warranty Deed - VA Loan
* New Home just constructed and sold
* COD Court Officer Deed/Estate
* Con Contract 15% Down payment regular terms
* ConLw Contract Low Down payment and low interest





* The above box plot and the bar plot shows Sales Price based on type of sale of the houses. Sales price of houses was not completed when last assessed (associated with New Homes) are the highest with houses having Abnormal Sale - trade, foreclosure, short sale being the lowest. Below are the abbreviations.
* Normal Normal Sale
* Abnorml Abnormal Sale - trade, foreclosure, short sale
* AdjLand Adjoining Land Purchase
* Alloca Allocation - two linked properties with separate deeds
* Family Sale between family members
* Partial Home was not completed when last assessed.
* Interpretation of the Results

Interpretation from pre-processing

The pre-processing started with finding the Null values from both the training and testing dataset. Then the Null values are replaced using various strategies. I used mean strategy to replace the numeric variables and used mode strategy to replace categorical variables. There were total of 19 columns with missing values and all were replaced using these strategies. A correlation graph was made which showed the correlation co efficient of all the numeric variables with the target variable. Variables which showed lesser correlation with the target variable are then dropped from the dataset. Total of 10 columns that has shown lesser or negative correlation had been dropped from the dataset. A separate dataset is made by extracting the categorical variables in the dataset and then those are quantified using label encoding techniques. The numeric variables are plotted in a box plot to find outliers and then z score is used to remove any data points that is beyond the threshold. The threshold here is 3. Variable which were skewed are fixed by using log transformation. The threshold for skewness is kept as .55. Then after separating the target and input variable, the input variables are scales using standard scaler.

Interpretation from Modelling

The training and testing data is separated using train test split. The data set is run through the algorithms. The ridge worked best with this dataset.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The finding from the study is mostly generated from doing an extensive EDA. Using analysis like univariate analysis and bivariate analysis made a huge difference in terms of finding out the insights. Univariate analysis gave an insight about the general dataset and the way the primary research was conducted. It gave the count of different variables those were considered while collecting the data while bivariate analysis gave a relation of the variables with the problem statement and gave the variables those are more important in determining the price of the houses.

* Learning Outcomes of the Study in respect of Data Science

The power of data visualization is something that is well found out in this study. Due to extensive EDA the relationship of different variables are well concluded. We could see from the visualizations which variables have a greater say in determining the prices and what alters might help the company to get a better price for their homes. Also in these kind of regression study the classic regression algorithms works best. Here in this case also the algorithms like linear regression, lasso, ridge worked very fine.