

Surprise Housing Price Prediction Project Use Case Report



Submitted by:

Sri Bharath Vasa

**ACKNOWLEDGMENT**

Thanks to Flip Robo and Datatrained for guiding and giving us projects in helping in learning the data science course. Thanks to my SME(Shubham Yadav) for guiding us in the projects.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1. <https://www.google.com/>
2. <https://www.youtube.com/>
3. <https://scikit-learn.org/stable/user_guide.html>
4. <https://github.com/>
5. <https://www.kaggle.com/>
6. <https://medium.com/>
7. <https://towardsdatascience.com/>
8. <https://www.analyticsvidhya.com/>

**INTRODUCTION**

* Business Problem Framing

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain.

Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

* Conceptual Background of the Domain Problem

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

1. Which variables are important to predict the price of a variable?
2. How do these variables describe the price of the house?

* Review of Literature

Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy houses to enter the market. The data set explains it is a regression problem as we need to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. Also, we have other independent features that would help to decide which all variables are important to predict the price of the variable and how do these variables describe the price of the house.

* Motivation for the Problem Undertaken

Our main objective of doing this project is to build a model to predict the house prices with the help of other supporting features. We are going to predict by using Machine Learning algorithms.

The sample data is provided to us from our client database. In order to improve the selection of customers, the client wants some predictions that could help them in further investment and improvement in selection of customers.

House Price Index is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price.

There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern themselves with the performance of individual models and neglect the less popular yet complex models.

As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the price of variables & also how do these variables describe the price of the house. This will help to determine the price of houses with the available independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the

conditional expectation of the dependent variable when the independent variables take on a given set of values.

Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

* Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 1168 rows and 81 columns. There are 2 data sets that are given. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. Size of training set: 1168 records.

2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 292 records.

* Data Pre-processing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre- process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

1. Loading the training dataset as a dataframe
2. Used pandas to set display I ensuring we do not see any truncated information
3. Checked the number of rows and columns present in our training dataset
4. Checked for missing data and the number of rows with null values
5. Verified the percentage of missing data in each column and decided to discard the ones that have more than 50% of null values
6. Dropped all the unwanted columns and duplicate data present in our dataframe
7. Separated categorical column names and numeric column names in separate list variables for ease in visualization
8. Checked the unique values information in each column to get a gist for categorical data
9. Performed imputation to fill missing data using mean on numeric data and mode for categorical data columns
10. Used Pandas Profiling during the visualization phase along with pie plot, count plot, scatter plot and the others
11. With the help of ordinal encoding technique converted all object datatype columns to numeric datatype
12. Thoroughly checked for outliers and skewness information
13. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
14. Separate feature and label data to ensure feature scaling is performed avoiding any kind of biasness
15. Checked for the best random state to be used on our Regression Machine Learning model pertaining to the feature importance details
16. Finally created a regression model function along with evaluation metrics to pass through various model formats

* Data Inputs- Logic- Output Relationships

When we loaded the training dataset, we had to go through various data pre-processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to logical part the domain expertise of understanding how real estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In Data Science community there is a saying “Garbage In Garbage Out” therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data how they were related to each other as well as our target label.

With the objective of predicting hosing sale prices accurately we had to make sure that a model was built that understood the customer priorities trending in the market imposing those norms when a relevant price tag was generated. I tried my best to retain as much data possible that was collected but I feel discarding columns that had lots of missing data was good. I did not want to impute data and then cause a biasness in the machine learning model from values that did not come from real people.

* State the set of assumptions (if any) related to the problem under consideration

The assumption part for me was relying strictly on the data provided to me and taking into consideration that the separate training and testing datasets were obtained from real people surveyed for their preferences and how reasonable a price for a house with various features inclining to them were.

* Hardware and Software Requirements and Tools Used

Hardware Used:

1. RAM: 12 GB
2. CPU: 11th Gen Intel(R) Core TM) i5-1135G7 @ 2.40GHz
3. GPU: intel iRISXe Graphics card

Software Used:

1. Programming language: Python
2. Distribution: Anaconda Navigator
3. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn and pandas\_profiling

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models.

For this project we need to predict the sale price of houses, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected Extra Trees Regressor as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of less amount of data.

In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.

Once we are able to get our desired final model, we can ensure to save that model before loading the testing data and start performing the data pre-processing as the training dataset and obtaining the predicted sale price values out of the Regression Machine Learning Model.

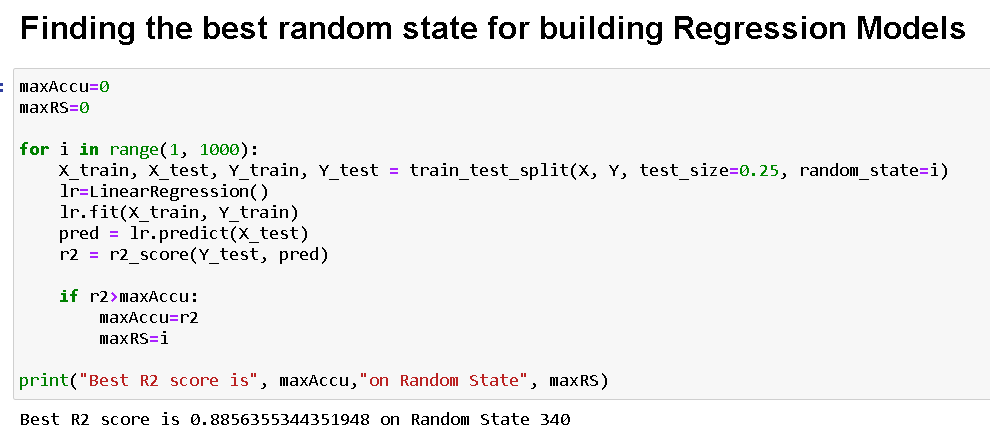
* Testing of Identified Approaches (Algorithms)

The algorithms used on training and test data are as follows:

1. Linear Regression Model
2. Ridge Regularization Regression Model
3. Lasso Regularization Regression Model
4. Support Vector Regression Model
5. Decision Tree Regression Model
6. Random Forest Regression Model
7. K Nearest Neighbours Regression Model
8. Gradient Boosting Regression Model
9. Ada Boost Regression Model
10. Extra Trees Regression Model

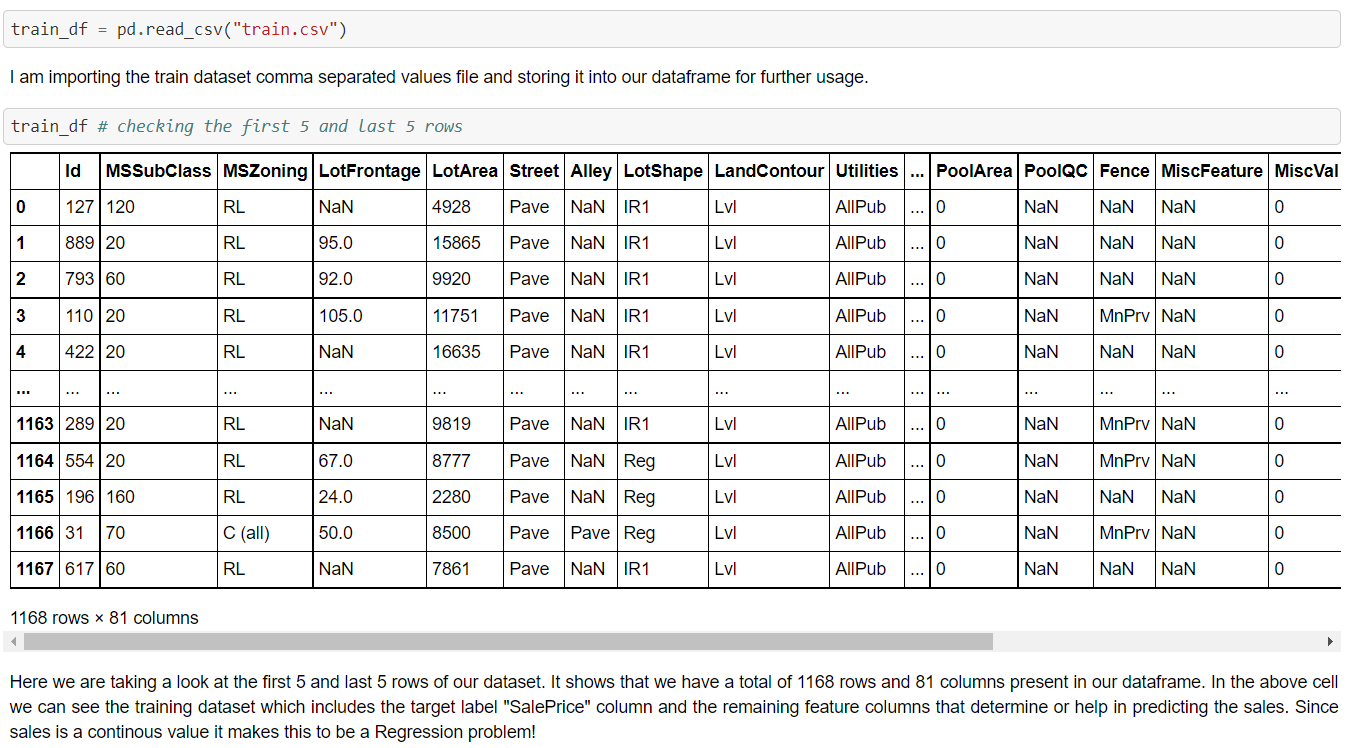
* Running and Evaluating Selected Models

Used the above said 10 models and tested for the best random state number among 1-1000 and defined a function to evaluate and train the model.

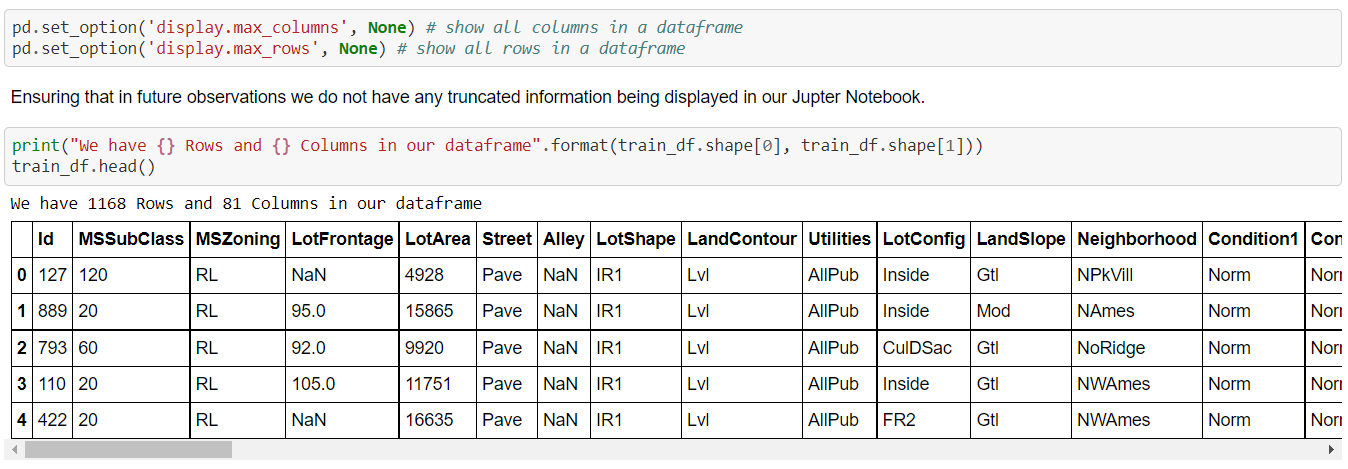


First, we start by importing the needed libraries and load the train dataset and look for the first and last data in the dataset.

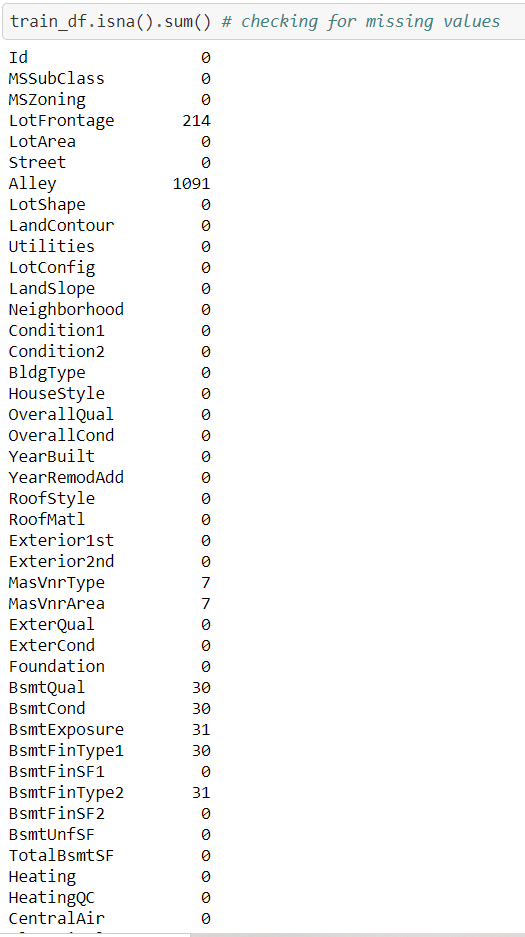


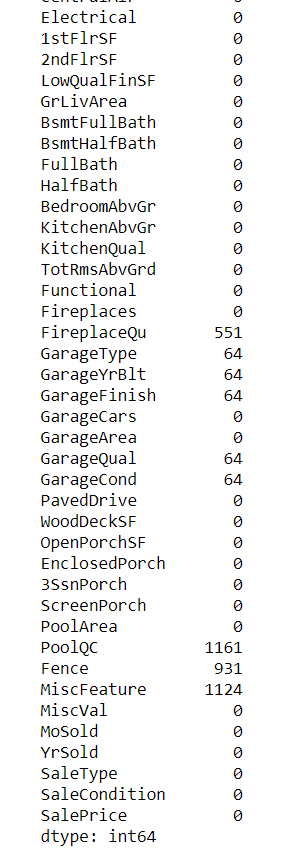


EDA (Exploratory Data Analysis)



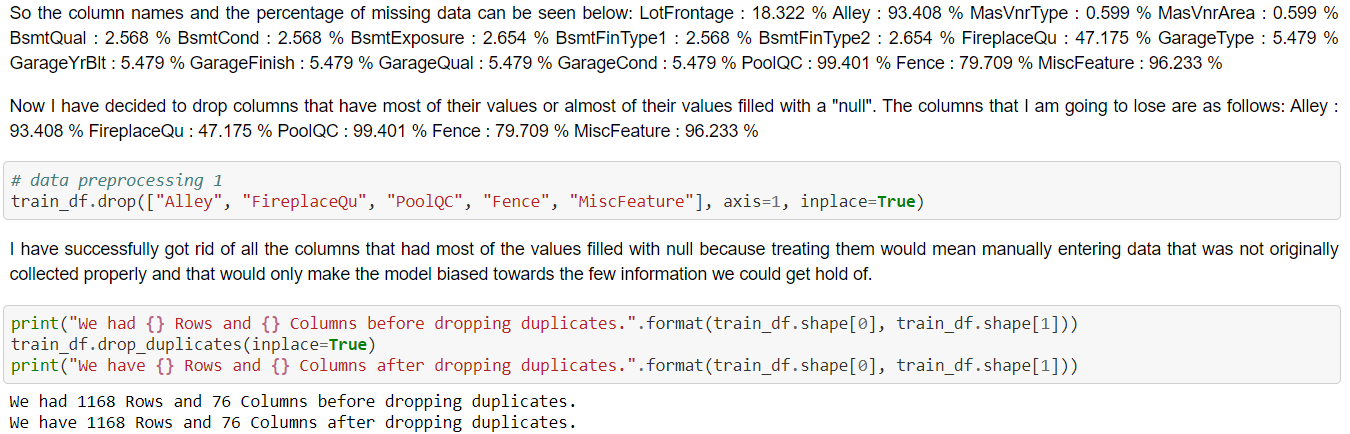
Then we are checking for the null values present in the train data.





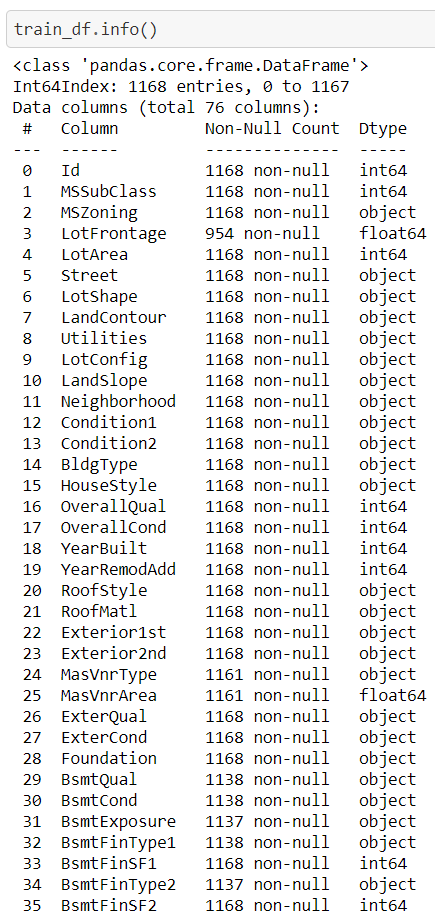
Using the isna() and sum functions together on our dataframe we can take a look at missing data information. It looks like we do have missing values present in few of our columns. However, we will check the percentage of missing information before we began treating them.

Then we are checking the percentage of missing data and dropping the non-required columns and checking for duplicity of data in the rows and columns.

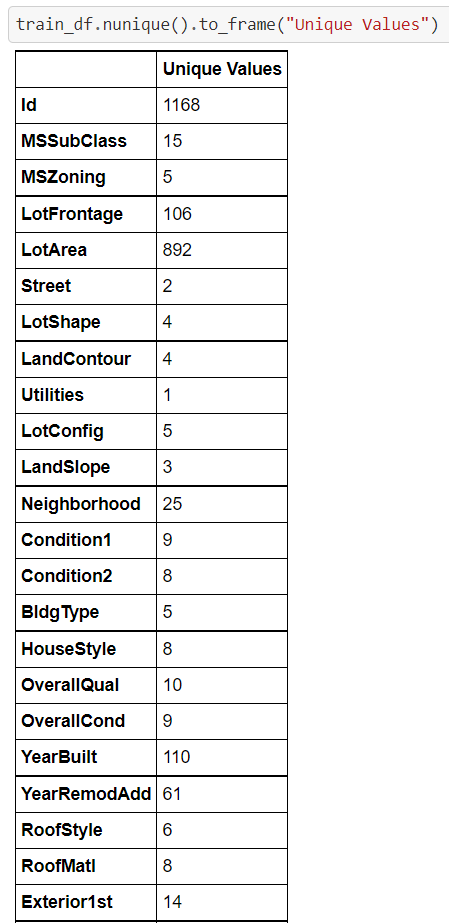


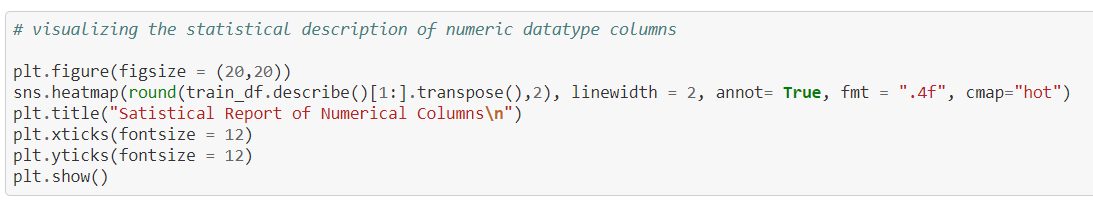
With the drop\_duplicates option I was trying to get rid of all the duplicate values present in our dataset. However, we can see that there are no duplicate data present in our dataset. I tried doing the same thing for dropping null values but we were losing lots of data.

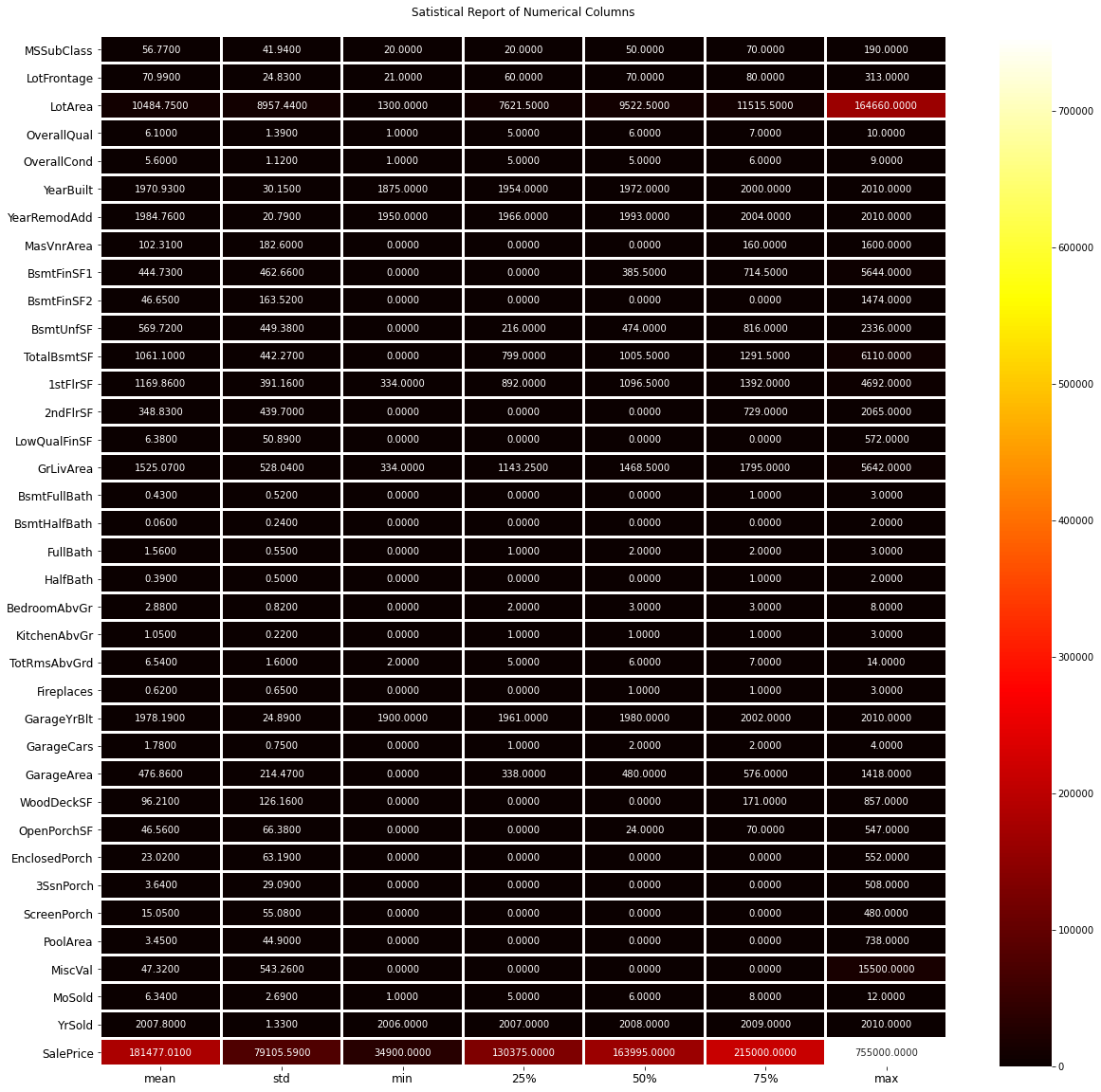
Further we are checking we are checking the datatype of each column and uniqueness of values in the dataset.



In the above cell we see that there are 3 columns with float datatype, 35 columns with integer datatype and 38 columns with object datatype.

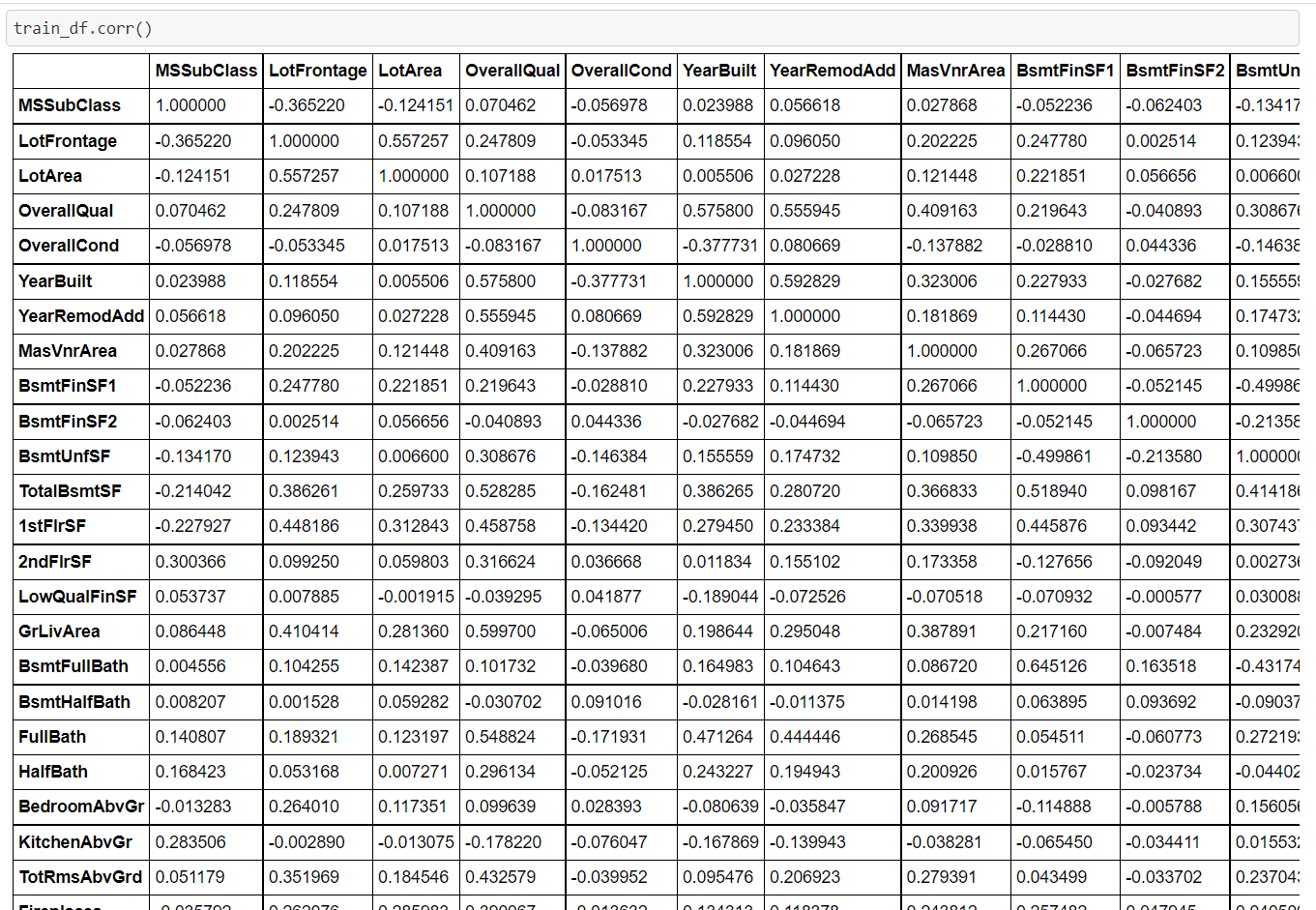






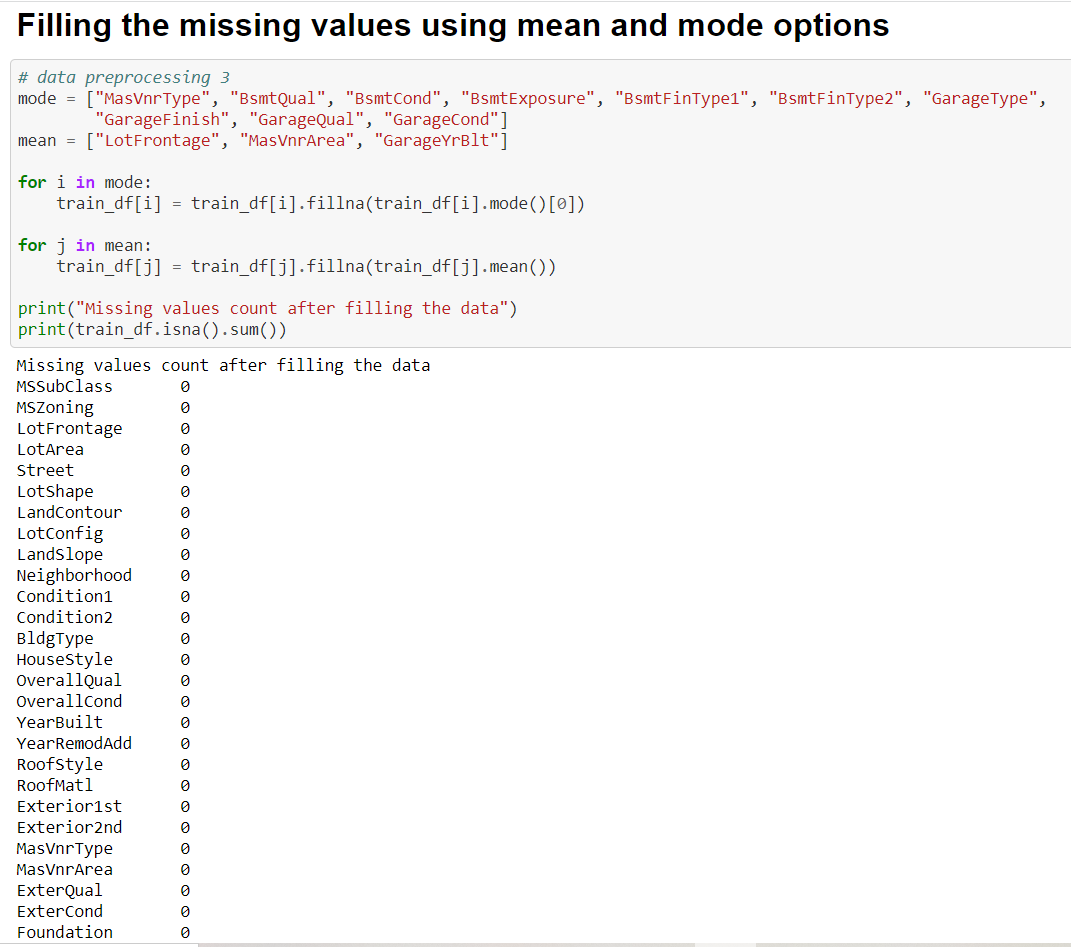
Using the above visualization on the describe method we are able to observe that our target label "SalePrice" has values that are higher than the other feature column details.

Checking the correlation values for all the numeric datatype columns.

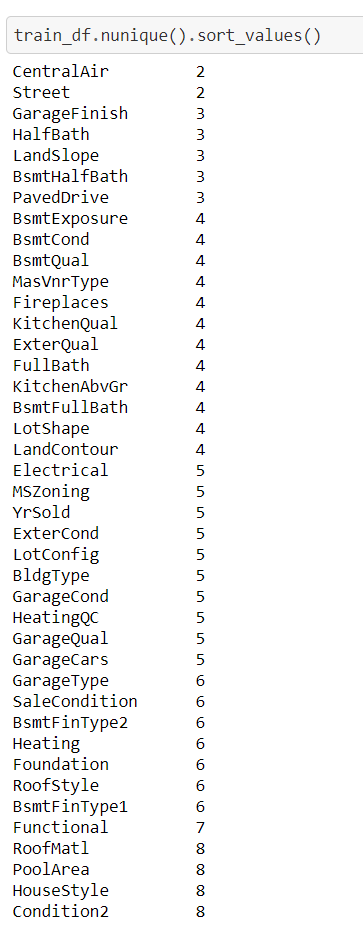


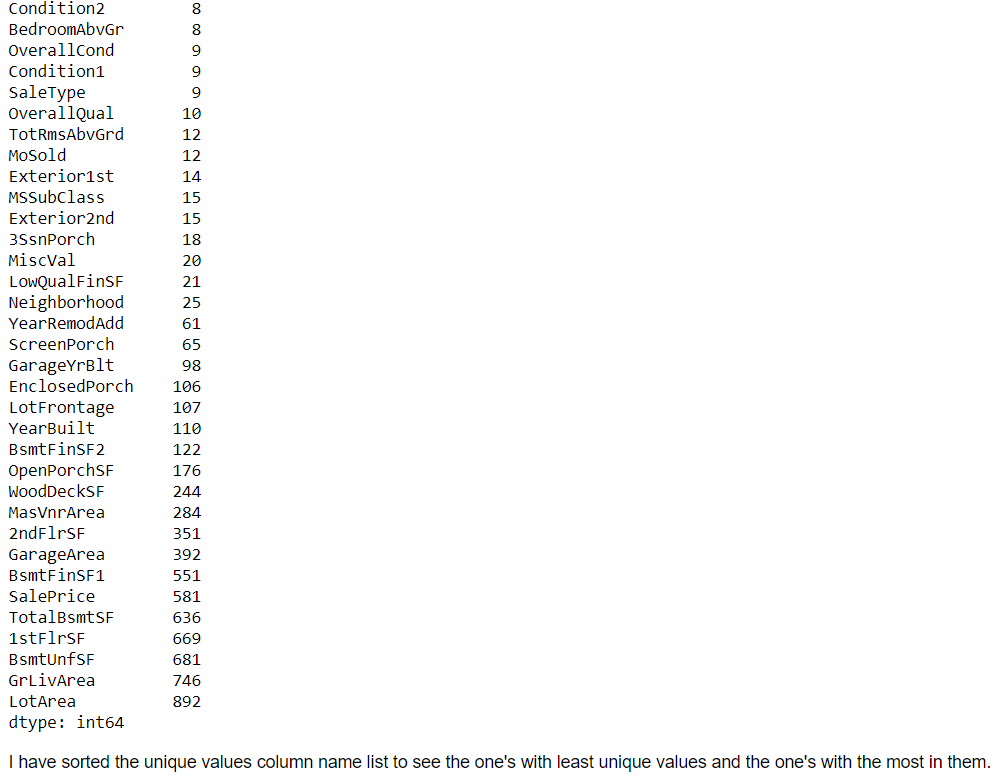
First, we checked for the presence of missing values and got it treated and using the below function we can see that there are no more null values present in the dataset.

Below, using the mean and mode method we can see that no null values present in the dataset.



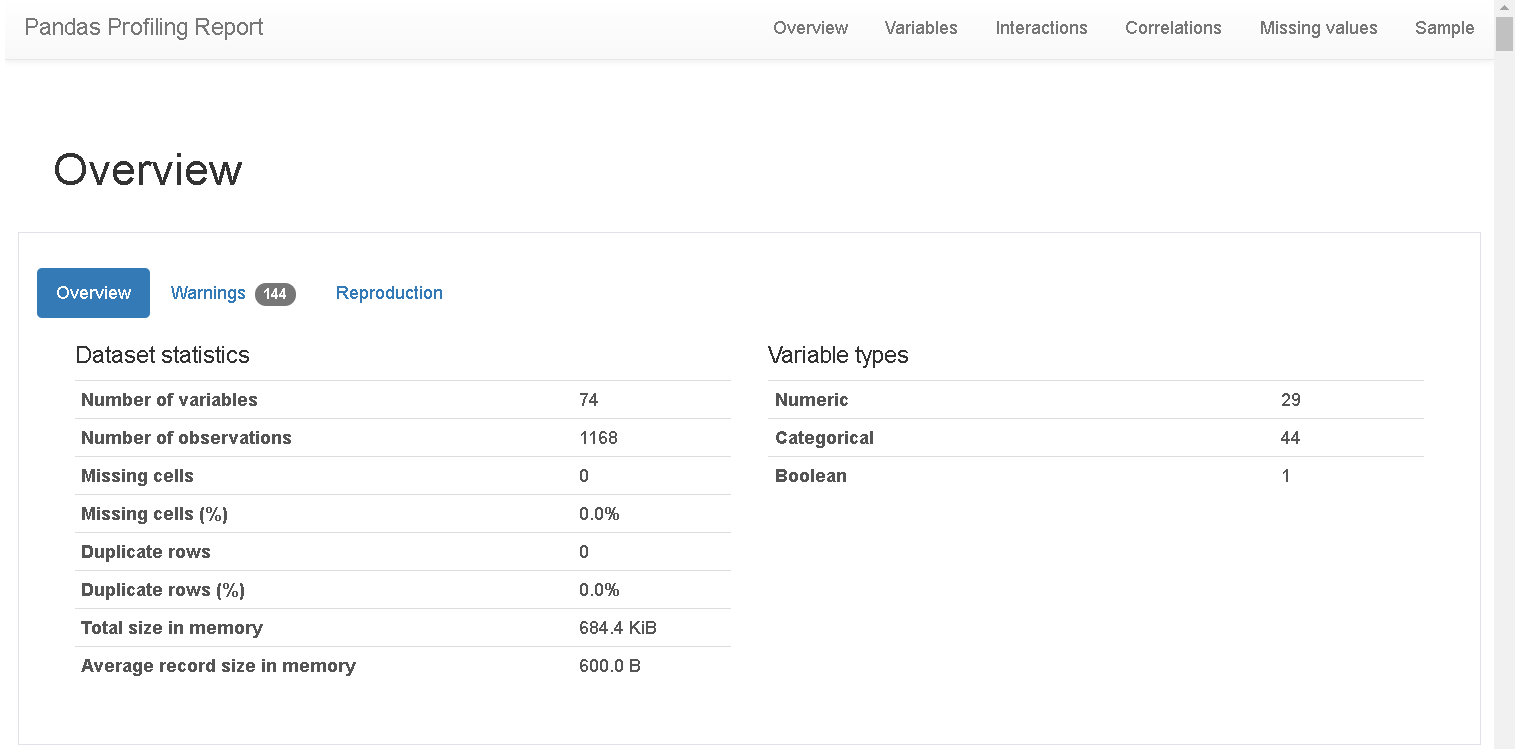
Visualization





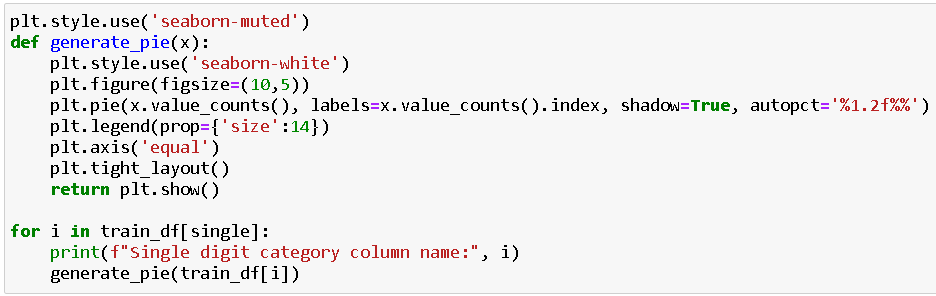
Then we are profiling the pandas and getting the output.



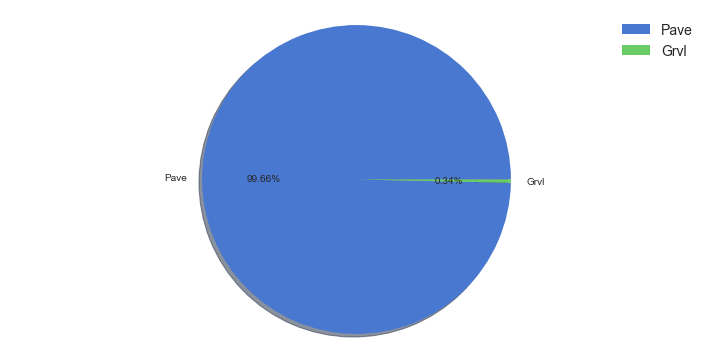


Then created Pie plots’, Count plots and Scatter plots to get further insight about the various variables featuring in our training dataset.

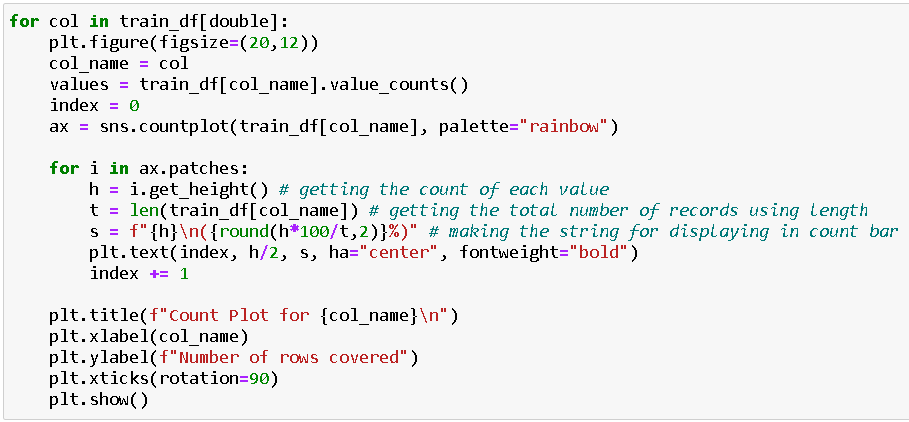
Code:



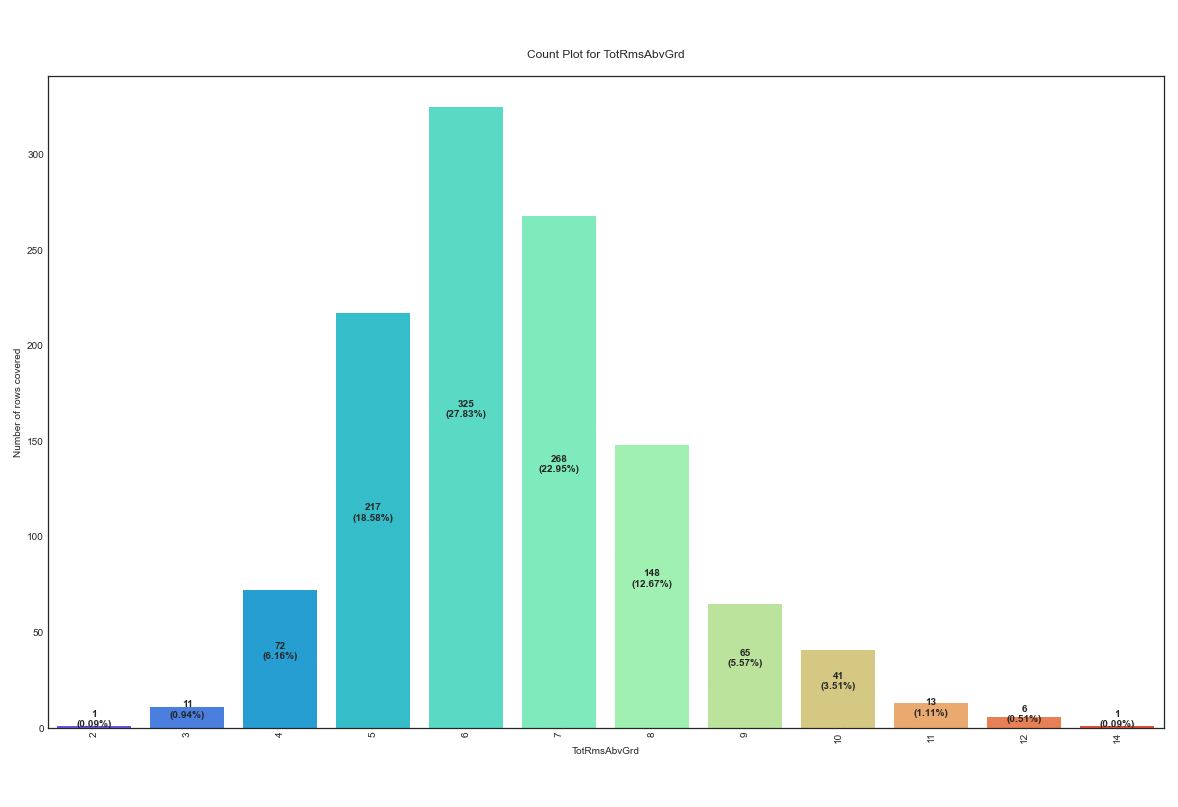
Output:



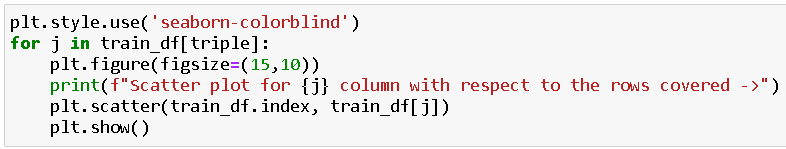
Code:



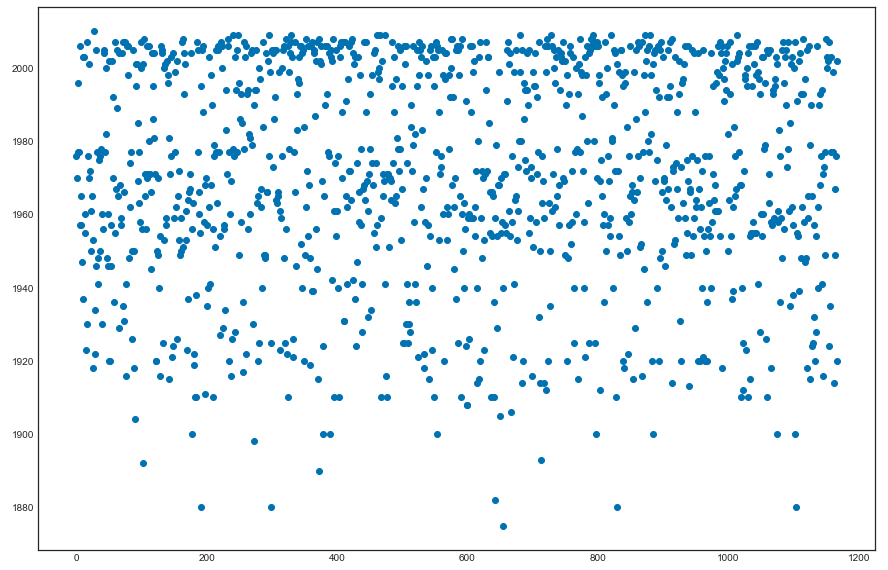
Output:



Code:



Output:



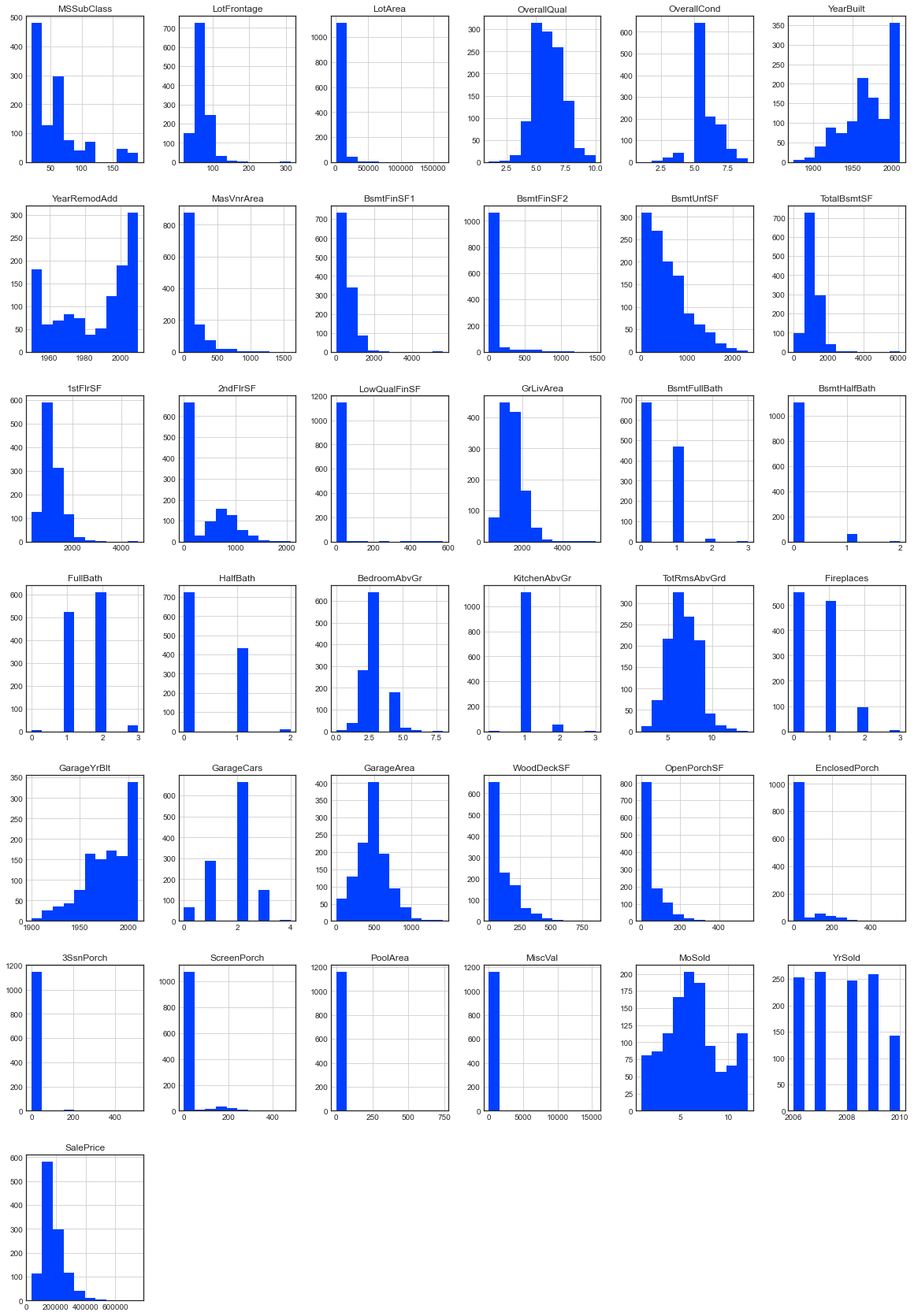
* Interpretation of the Results

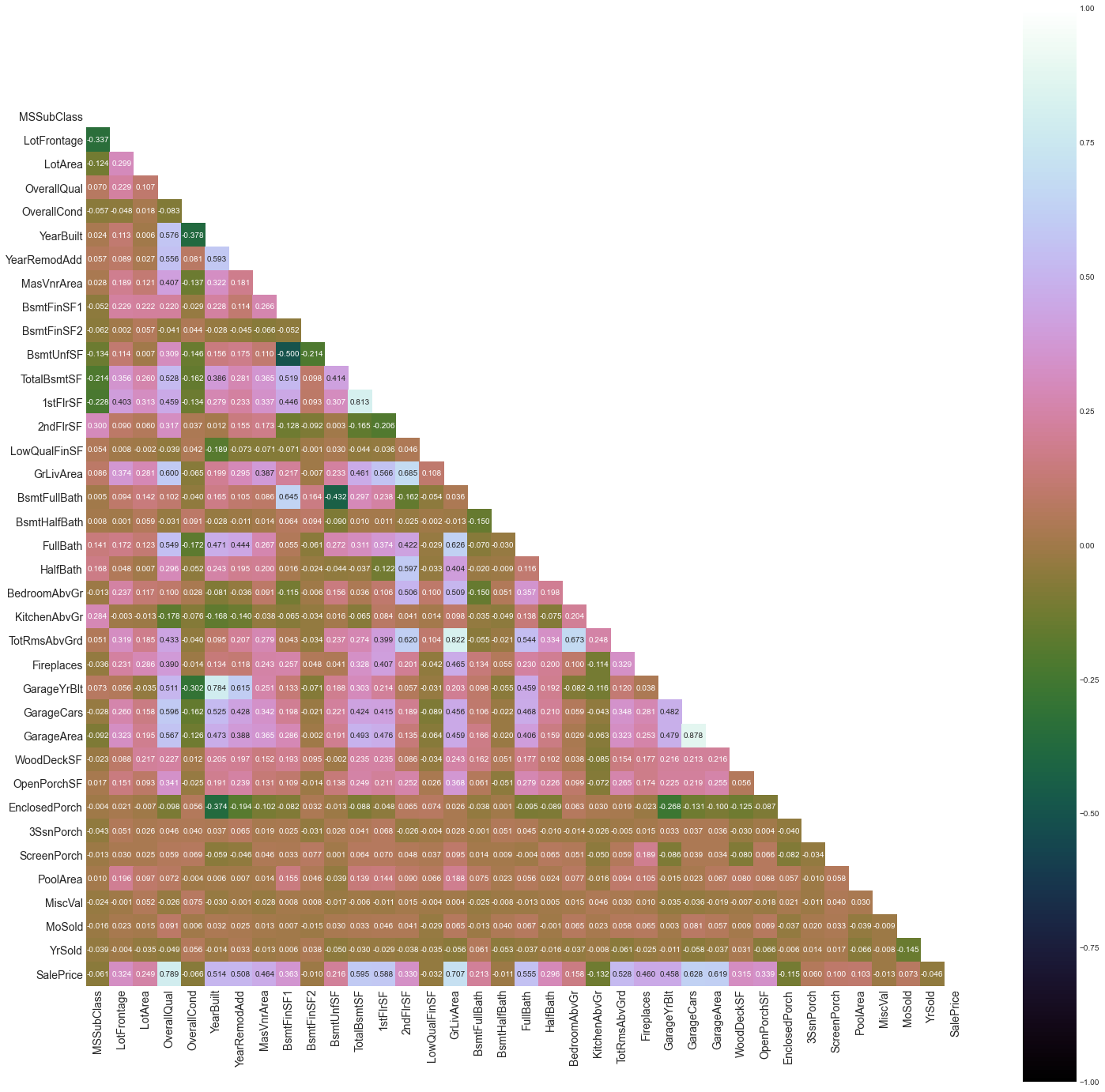
Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importance and to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot. I got to know the count of a particular category for each feature by using count plot and most importantly with predicted target value distribution as well as scatter plot helped me to select the best model.

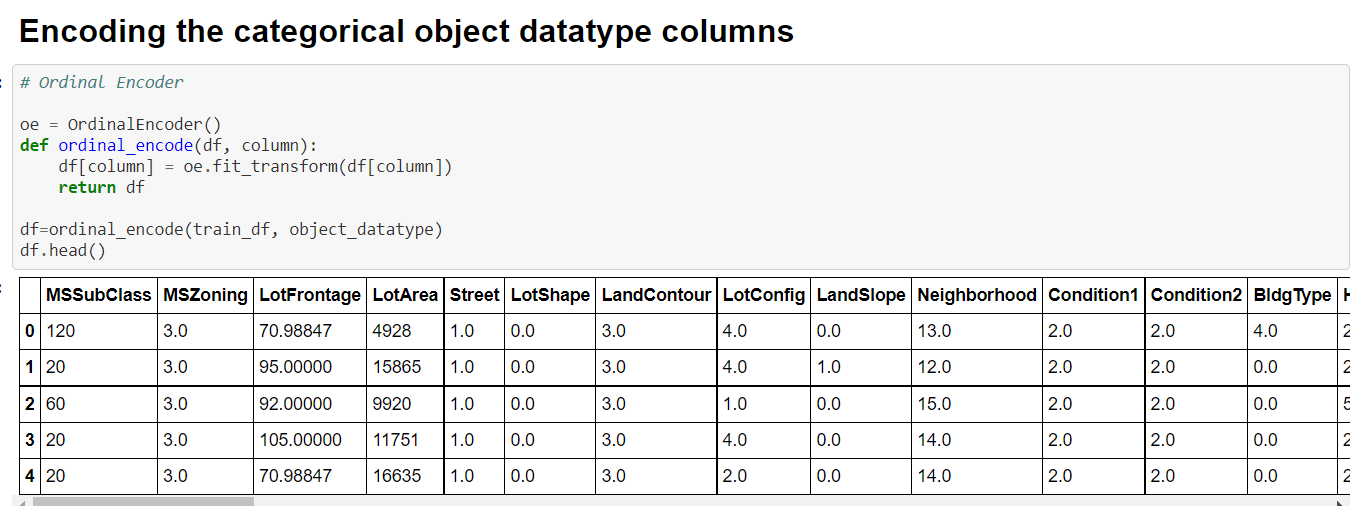
Pre-processing: Basically, before building the model the dataset should be cleaned and scaled by performing few steps. As I mentioned above in the pre-processing steps where all the important features are present in the dataset and ready for model building.

Model Creation: Now, after performing the train test split, I have x\_train, x\_test, y\_train & y\_test, which are required to build Machine learning models. I have built multiple regression models to get the best R2 score, MSE, RMSE & MAE out of all the models

Then we got Histogram, Heatmap for further insight on the price variance.

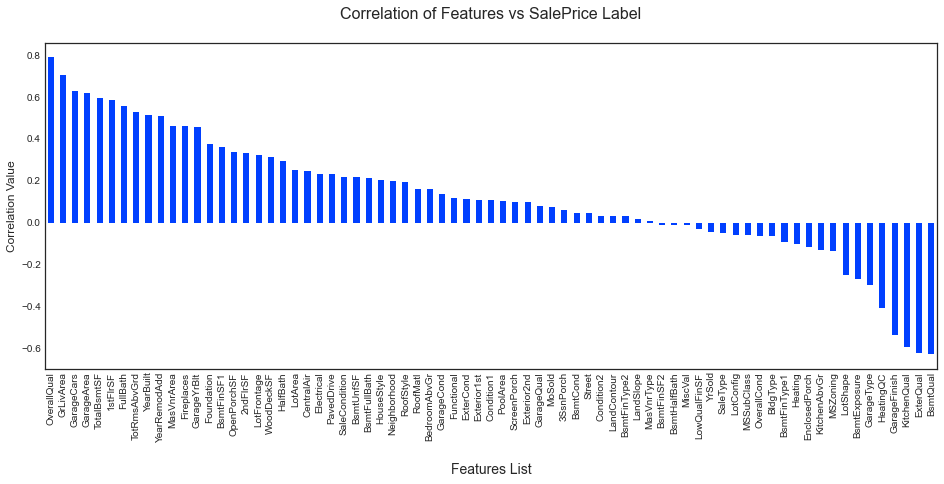


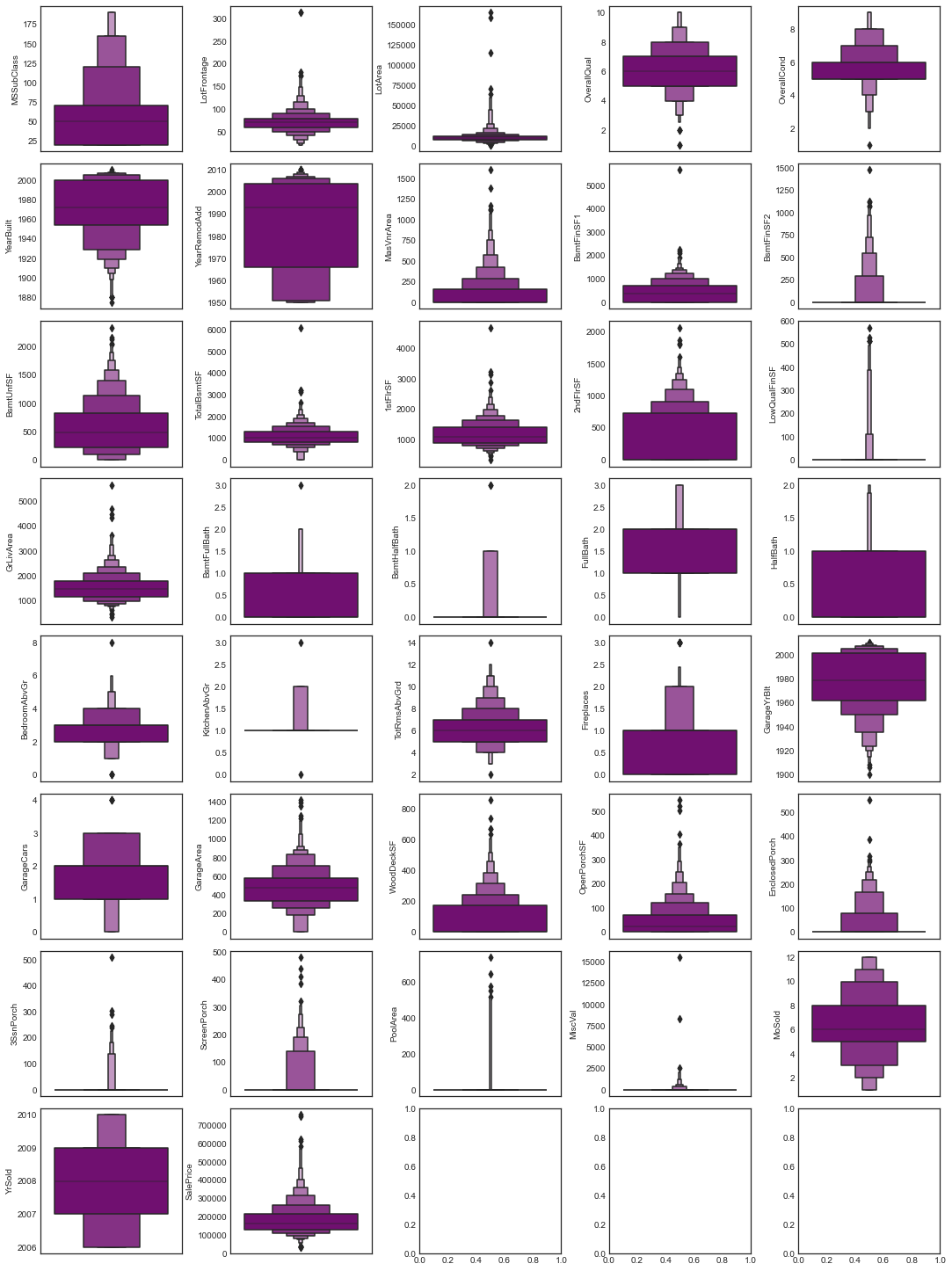


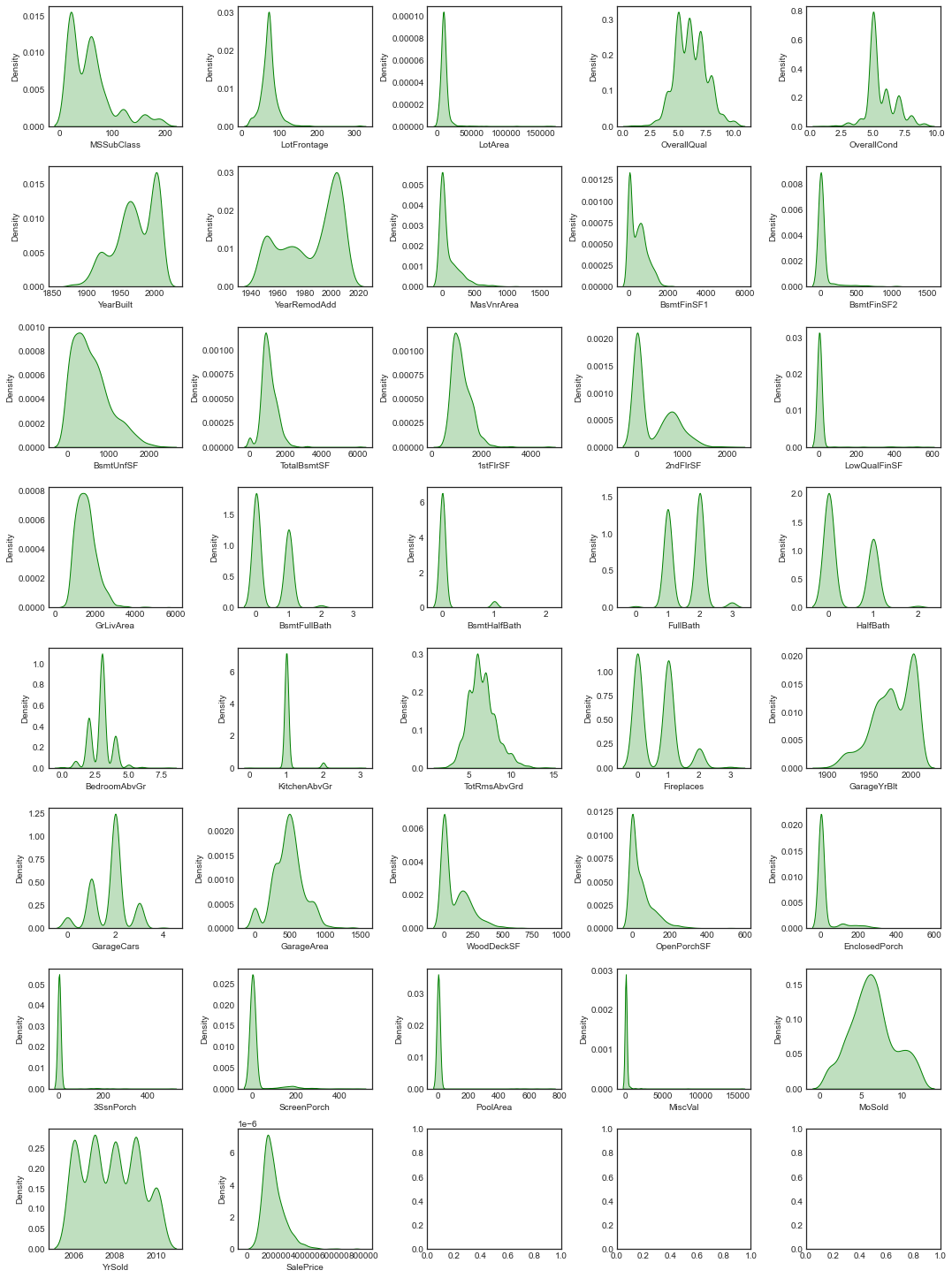


The heatmap gives us the correlation between positive and negative correlation of the dataset and the feature columns to convert the object datatype columns to numeric format.

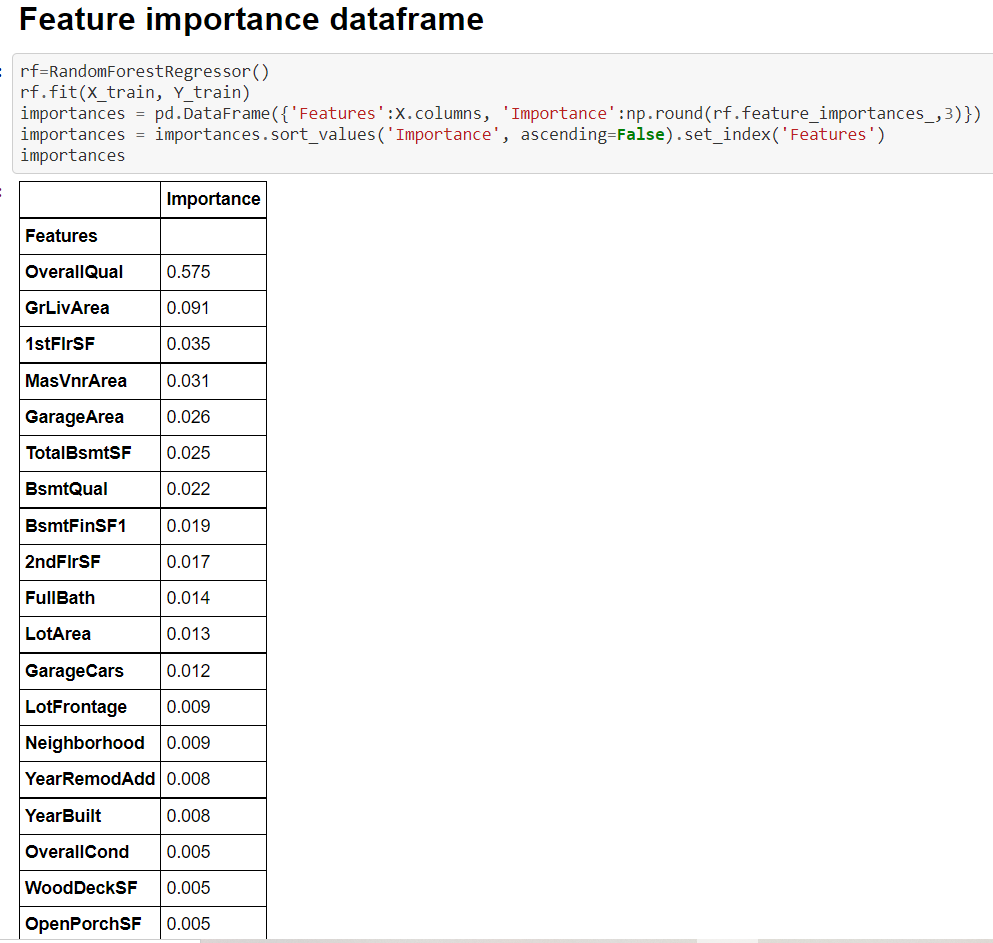
Further Bar Chart, Boxen Plot and Distribution plot are formed for our further price increase of housing.



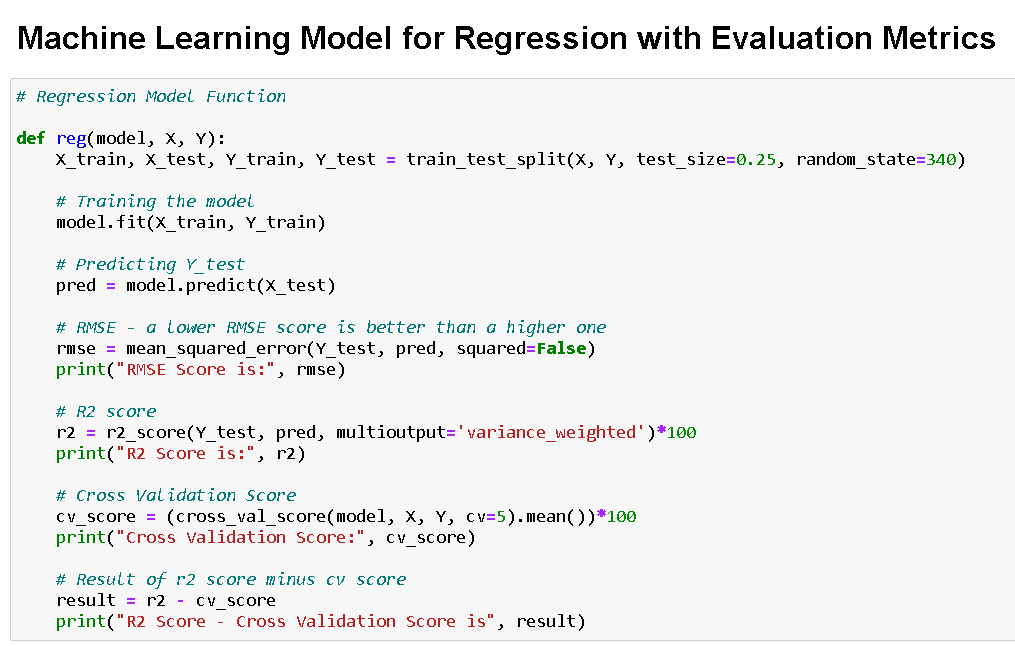




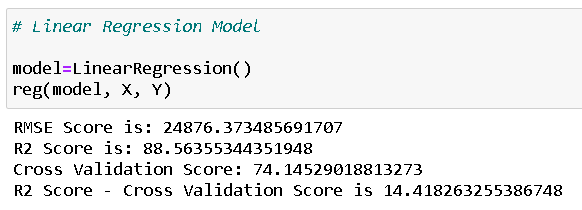
Post plotting of all the plots and X Y variable splitting we do



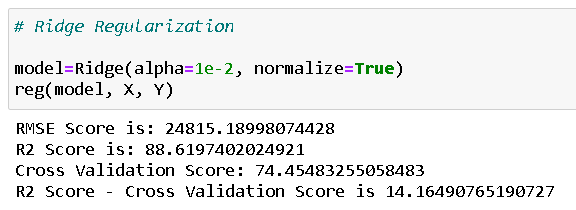
Regression Model Function:



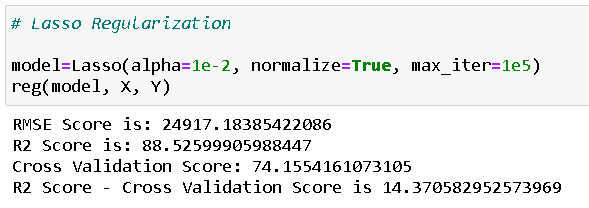
Linear Regression being tested:



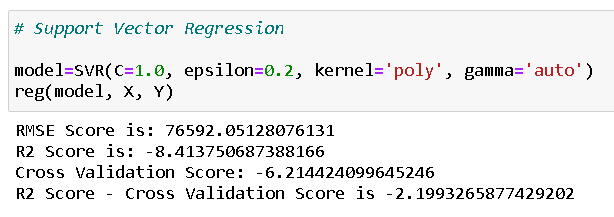
Ridge Regularization:



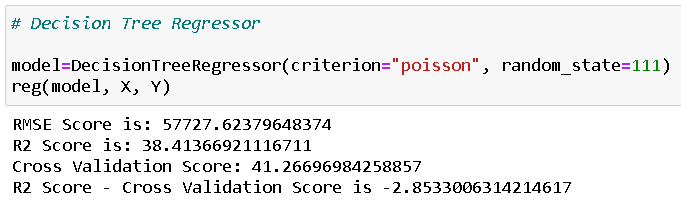
Lasso Regularization:



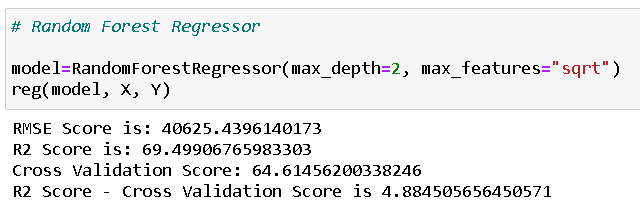
Support Vector Regressor:



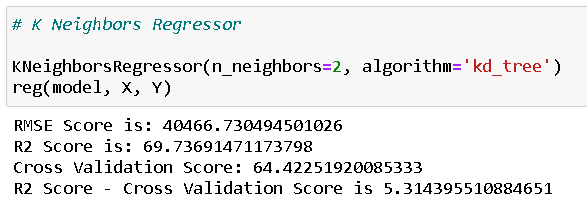
Decision Tree Regressor:



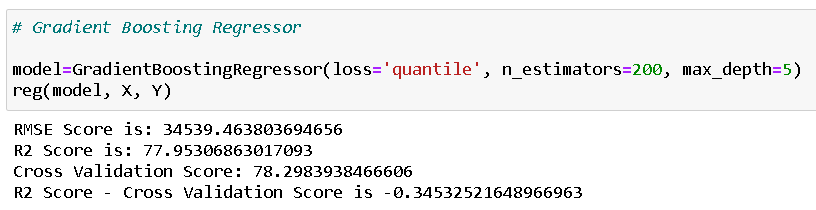
Random Forest Regressor:



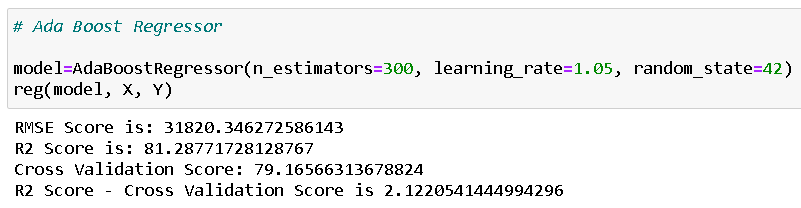
K Nearest Neighbour Regressor:



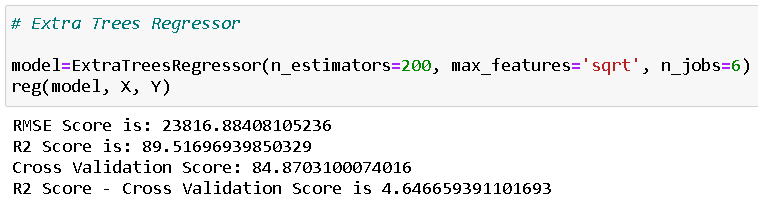
Gradient Boosting Regressor:



Ada Boost Regressor:

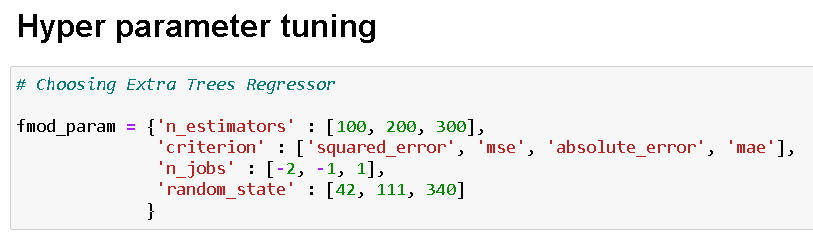


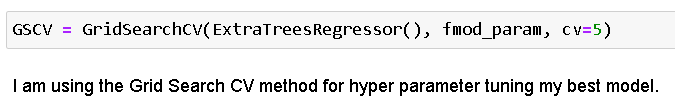
Extra Tree Regressor:

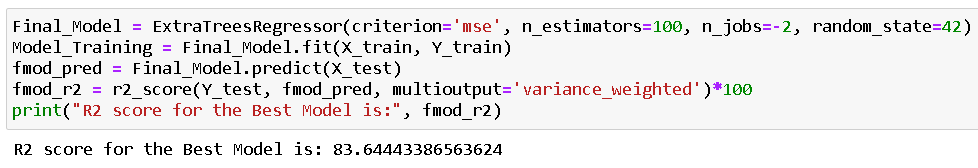


Post testing on the regression models we are hyper tuning the parameter.

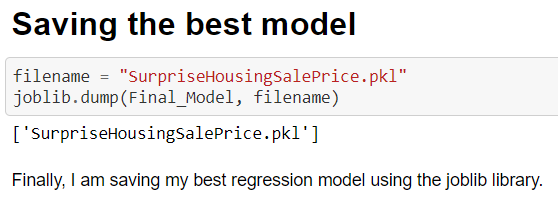
Code:



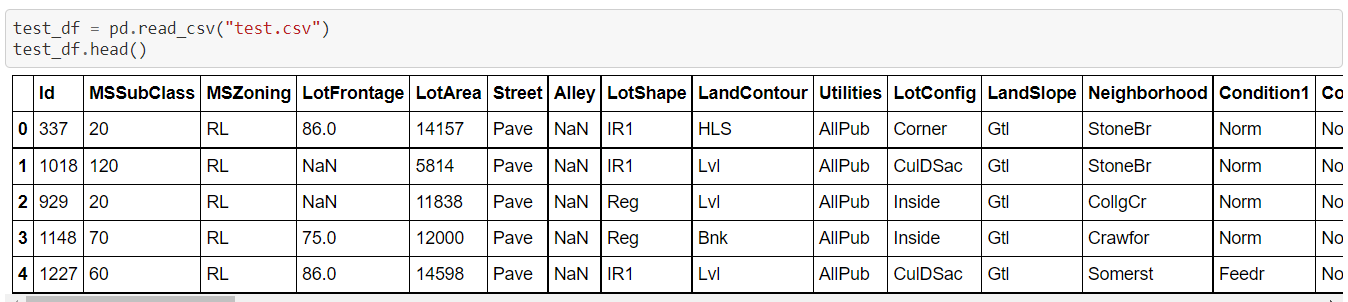




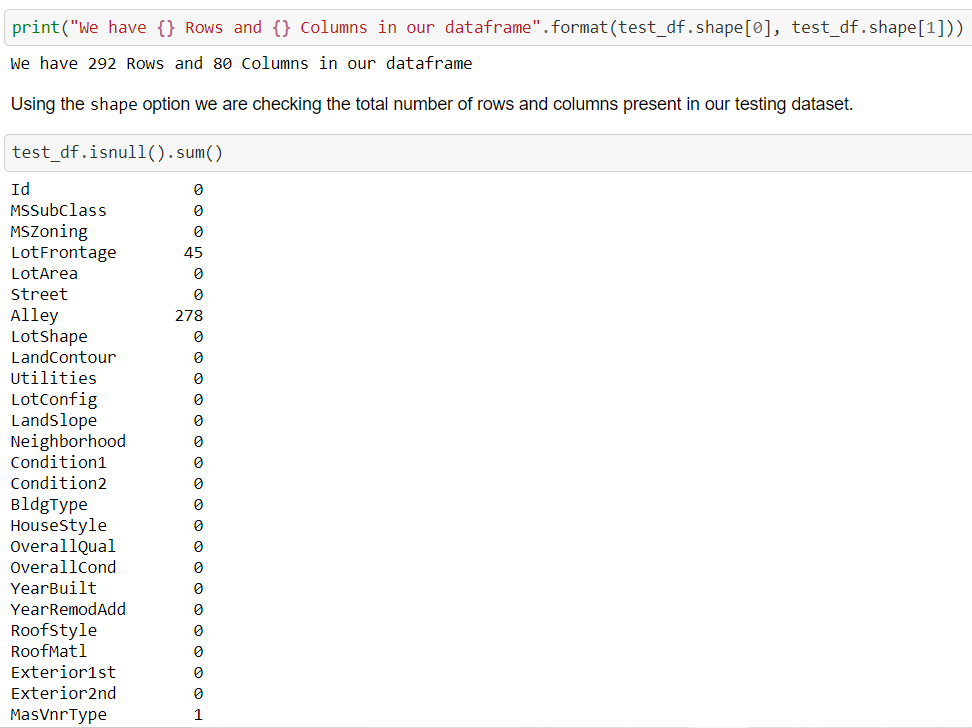
After getting the best R2 score of the model we are saving the model.



Then, we start testing the test dataset available to us.



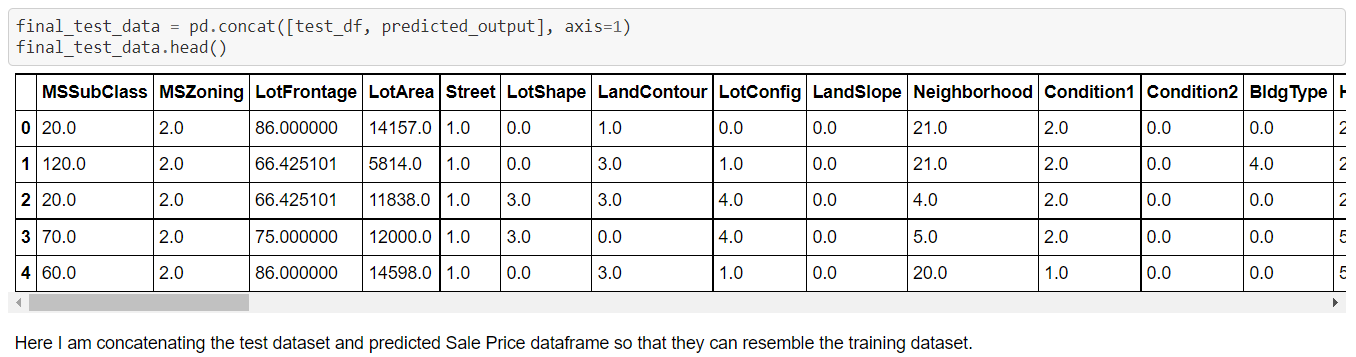
Then, we are checking the number of rows and columns and using isna() function to find null values present in the dataset or not.



After checking for the missing data percentage just like the train dataset we are doing for test data set too and treating the missing values and looking if null values are further present or not as below.



Further Predicting the sale price



Conclusion:

-> After getting an insight of this dataset, we were able to understand that the Housing prices are done on basis of different features.

-> First, we loaded the train dataset and did the EDA process and other pre-processing techniques like outlier and skewness check, handling the null values present, filling the missing data with mean and mode, visualizing the distribution of data, etc.

-> Then we did the model training, building the model and finding out the best model on the basis of different metrices scores we got like R2 score, Cross Validation score, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, etc.

-> We got Extra Trees Regressor as the best algorithm among all as it gave more r2\_score and cross\_val\_score. Then for finding out the best parameter and improving the scores, we performed Hyperparameter Tuning.

-> As the scores were not increased, we also tried using Ensemble Techniques like RandomForestRegressor, AdaBoostRegressor and GradientBoostingRegressor algorithms for boosting up our scores. Finally, we concluded that Extra Trees Regressor remained the best performing algorithm, although there were more errors in it and it had less RMSE compared to other algorithms. It gave an r2\_score of 89.51 and cross\_val\_score of 84.87 which is the highest scores among all.

-> We saved the model in a pickle with a filename in order to use whenever we require. -> We predicted the values obtained and saved it separately in a csv file.

-> Then we used the test dataset and performed all the pre-processing pipeline methods to it. -> After treating missing values, we loaded the saved model that we obtained and did the predictions over the test data and then saving the predictions separately in a csv file.

-> From this project, I learnt how to handle train and test data separately and how to predict the values from them. This will be useful while we are working in a real-time case study as we can get any new data from the client we work on and we can proceed our analysis by loading the best model we obtained and start working on the analysis of the new data we have.

-> The final result will be the predictions we get from the new data and saving it separately.

-> Overall, we can say that this dataset is good for predicting the Housing prices using regression analysis and Extra Trees Regressor is the best working algorithm model we obtained.

-> We can improve the data by adding more features that are positively correlated with the target variable, having less outliers, normally distributed values, etc.

-> Also, we can work upon many factors to originally improve the quality of our features before providing it as an input for our machine learning models

Learning Outcomes of the Study in respect of Data Science

The above study helps one to understand the business of real estate. How the price is changing across the properties. With the Study we can tell how multiple real estate amenities like swimming pool, garage, pavement and lawn size of Lot Area, and type of Building raise decides the cost. With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can project the price of the property.

Future Work

* The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.
* Make use of the available features and if they could be combined as binning features has shown that the data got improved.
* Training the datasets with different regression methods such as Elastic net regression that combines both L1 and L2 norms. In order to expand the comparison and check the performance.
* The correlation has shown the association in the local data. Thus, attempting to enhance the local data is required to make rich with features that vary and can provide a strong correlation relationship.
* The factors that have been studied in this study has a weak correlation with the sale price. Hence, by adding more factors to the local dataset that affect the house price, such as GDP, average income, and the population. In order to increase the number of factors that have an impact on house prices.
* The results of this study have shown that ANN is prone to overfitting. However, ANN still a strong algorithm that has a lot of options that could, with the right methods, provide a better prediction accuracy. ANN has a lot of possibilities that could lead to a different output. For instance, experimenting with the model when using combinations of layers and neurons over several iterations in order to find what fits the algorithm.
* ANN model was designed using feed-forward architecture. The model could make use of applying the proper back-propagation method to reduce the weight between neurons and give a better training performance resulting in better prediction accuracy.