

PFA Housing Project

Submitted by:

ROHIT KATTEWAR

ACKNOWLEDGEMENT

I would like to thank and express my sincere gratitude to Flip Robo Technologies for giving me the opportunity to work on this project named ‘PFA Housing Project’ using Machine Learning algorithms.

Primarily, I would like to thank to the author of the paper titled: “Housing Price Prediction Project” for providing me invaluable knowledge and insights in the form of dataset. A survey of literature and the impact of housing quality on house prices in eight capital cities depicts the dynamic relationship that exist in the economics of the estate and housing markets.

Finally, I will thank my mentors, under whose guidance I learned a lot about Machine Learning, Natural Language Processing and much more.

INTRODUCTION

* **Business Problem Framing**

A United States of America 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. 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. 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.

* **Conceptual Background of the Domain Problem**

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Hedonic Characteristics of Housing Price: A Hedonic approach is preferred for predicting the sale prices in the housing market because the market displays resilience, flexibility and spatial fixity. Housing Attributes: Studying the structural, locational, and economic attributes of housing properties are crucial in understanding their mutually inclusive relationships with the pricing.

* **Review of the Literature**

Two research papers, namely: “House Price Prediction using aMachine Learning Model: A Survey of Literature” and “Theimpact of housing quality on house prices in eight capital cities,Australia” were reviewed and evaluated to gain insights into allthe attributes that influence the price of house. In this research, various models were built in which the house Sale Price is projected as separate and dependent variable while locational, structural and various other attributes of housing properties were treated as independent variables. Therefore, the house price is set as a target or dependency variable, while other attributes are set as independent variables to determine the main variables by identifying the correlation coefficient of each attribute.

* **Motivation for the Problem Undertaken**

There is a steady rise in house demand with every passing year, and consequently the house prices are rising every year. The problem arises when there are numerous variables such as location and property demand that influence the pricing. Therefore, buyers, sellers, developers and the real estate industry are keen to know the most important factors influencing the house price to help investors make sound decisions and help house builders set the optimal house price. There are many benefits that home buyers, property investors, and house builders can reap from the house-price model. This model aims to serve as a repository of such information and gainful insights to home buyers, property investors and house builders, that will help them determine best house prices. This model can be useful for potential buyers in deciding the characteristics of a house they want.

ANALYTICAL PROBLEM FRAMING

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

In order to forecast house price, predictive models such as Random Forest Regression model, Decision tree Regression Model, Support Vector Machine Regression model, Gradient Boost Regression were used to describe how the values of Sale Price depended on the independent variables of various Housing property attributes. Regression modelling techniques were used in this Problem since Sales Price data distribution is continuous in nature.

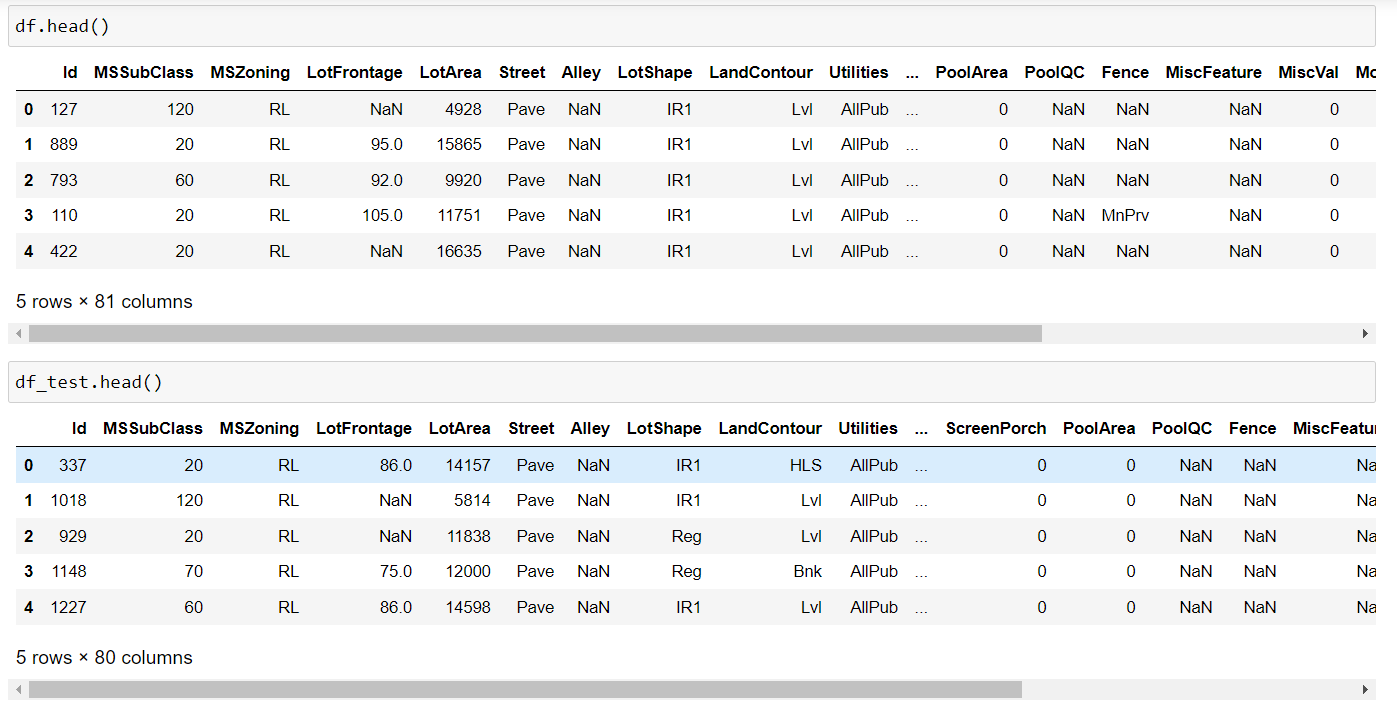
* **Data Sources and their Formats:**

The dataset was compiled by a US-based housing company

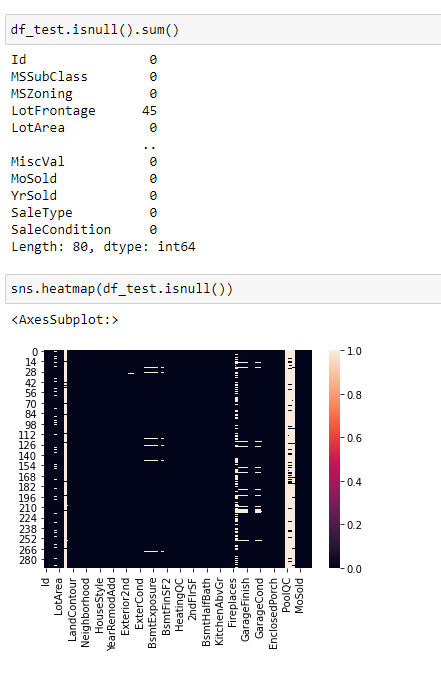
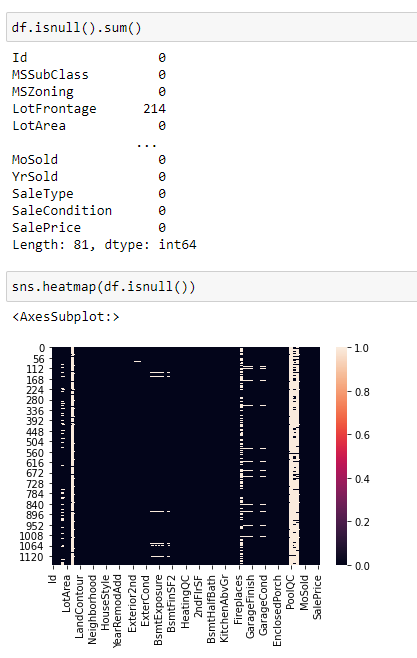
named Surprise Housing. The company has collected a data

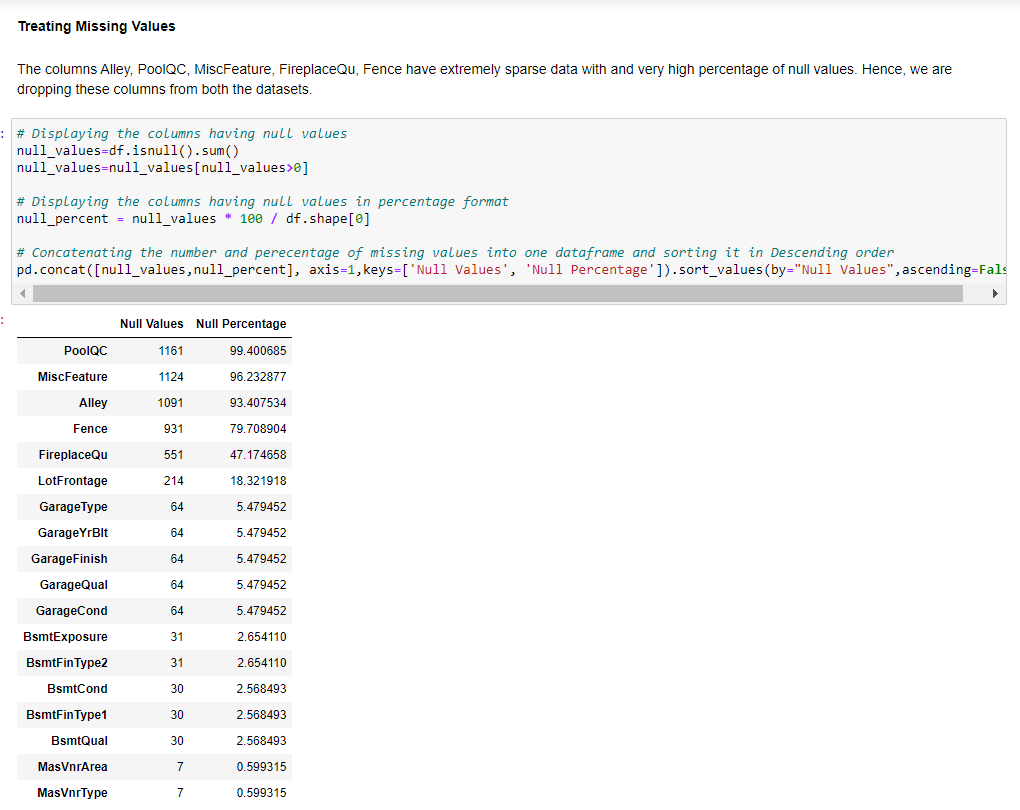
set from the sale of houses in Australia. The dataset was

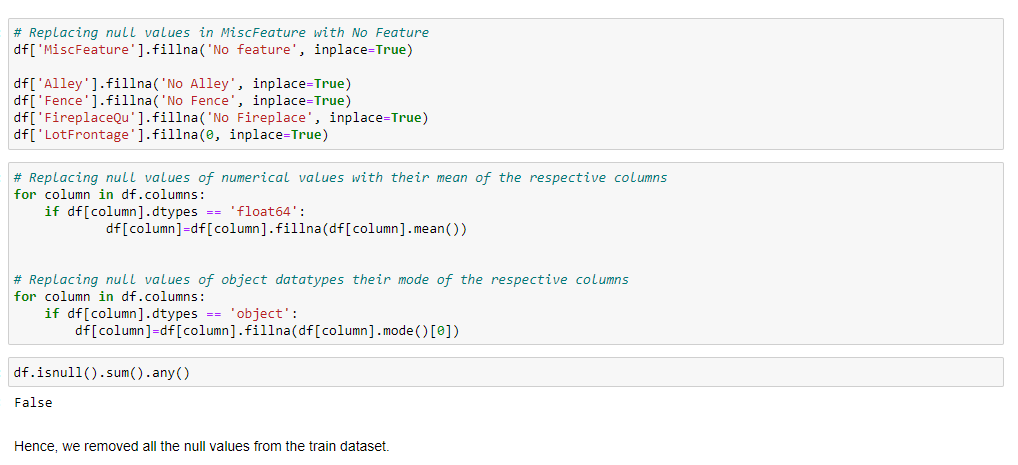
made available in .csv file format. There are two datasets: One for training the predictive machine learning models and the second one to be used by the models for predicting the SalePrice (target variable).



* **Dataset Description:**
* MSSubClass: Identifies the type of dwelling involved in the sale.
* MSZoning: Identifies the general zoning classification of the sale.
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access to property
* Alley: Type of alley access to property
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to various conditions
* Condition2: Proximity to various conditions (if more than one is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Rates the overall material and finish of the house
* OverallCond: Rates the overall condition of the house
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Evaluates the quality of the material on the exterior
* ExterCond: Evaluates the present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Evaluates the height of the basement
* BsmtCond: Evaluates the general condition of the basement
* BsmtExposure: Refers to walkout or garden level walls
* BsmtFinType1: Rating of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Rating of basement finished area (if multiple types)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 1stFlrSF: First Floor square feet
* 2ndFlrSF: Second floor square feet
* LowQualFinSF: Low quality finished square feet (all floors)
* GrLivArea: Above grade (ground) living area square feet
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
* Kitchen: Kitchens above grade
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality (Assume typical
* unless deductions are warranted)
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* WoodDeckSF: Wood deck area in square feet
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* PoolQC: Pool quality
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold (MM)
* YrSold: Year Sold (YYYY)
* SaleType: Type of sale
* SaleCondition: Condition of sale
* **Data Pre-processing Done**







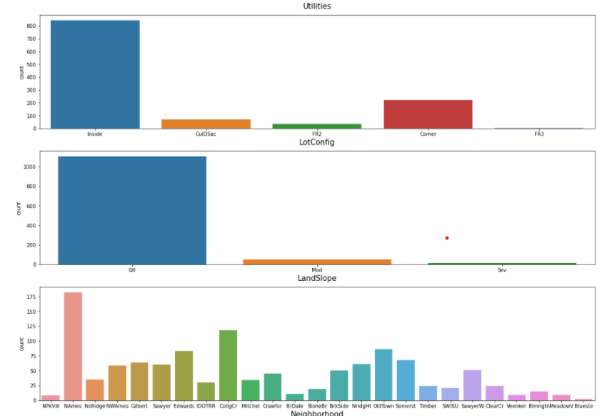
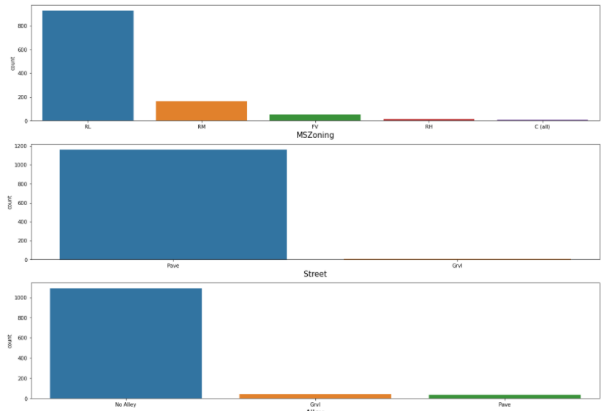
Plotting a heatmap of null values revealed that in both training and testing datasets, Columns titled: Alley, PoolQC, MiscFeature, FireplaceQu, Fence have extremely sparse data with overwhelmingly high percentage of null values.

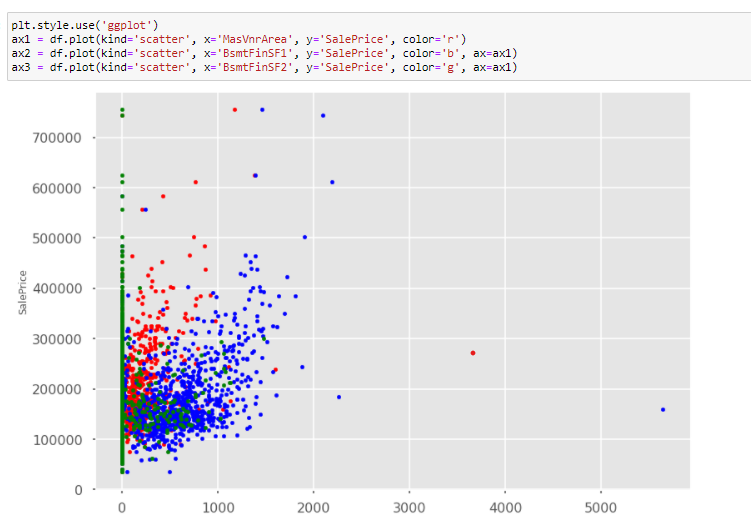
* **Data Inputs- Logic- Output Relationships:**

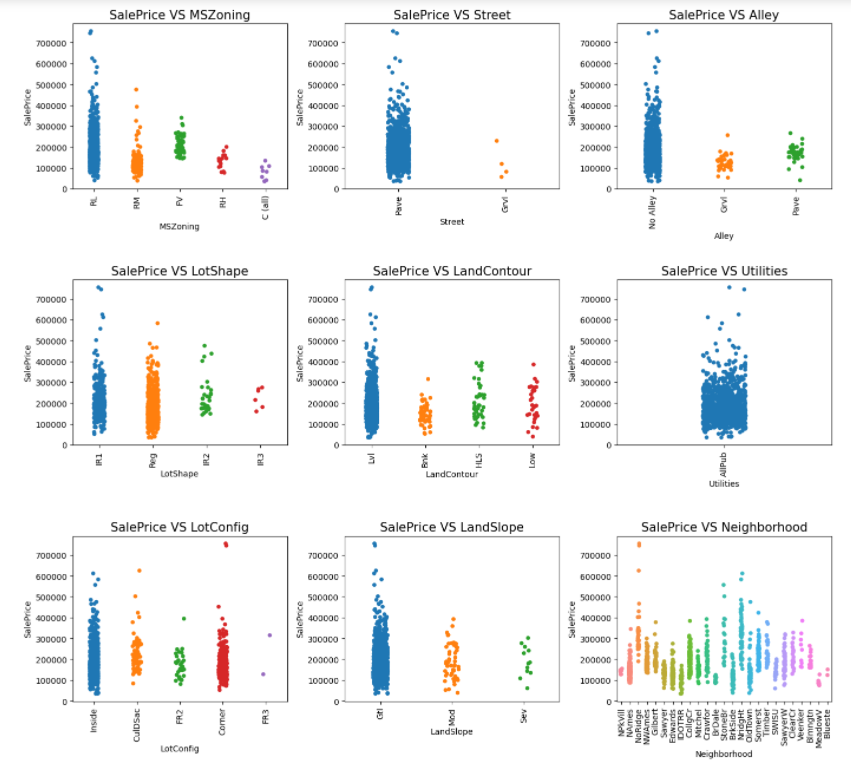
The Datasets consist mainly of object data type variables and a few float and int data type variables. The relationships between the independent variables and dependent variable were analysed. Features like Lot area, Lot Frontage, Overall Quality, Overall Condition, Basement Finishing, Total Basement Surface Area, first and 2nd Floor square feet, Garage capacity, Total rooms have a positive linear relationship, therefore increase in their values leads to increase in SalePrice. Whereas Age of Housem Remodellling age Garrage age have a linear negative relationship and therefore increase in their values leads to a decrease in SalePrice.

* **Exploratory Data Analysis:**

**Visualizations:** Bar plots, Count plots, Box plots, Swarm plots, Scatter plots were used to visualise the data of all the columns and their relationships with Target variable.







From the graphs above the following observations are made:

1. Most of the houses belongs to Residential Low-Density zone and many houses from this zone are having higher prices than other zones.

2. Almost all houses are having paved streets and few are having gravel streets.

3. More number of houses are having General shape of property slightly irregular or regular. Few of them are having irregular shape.

4. SalePrice vs LandContour plot shows most of the houses are nearly Flat/Leveled.

5. Nearly all houses has kind of utilities.

6. More number of lots are inside or at corners.

7. Most of the houses are having gentle slopes.

8. Houses with sever slopes are having slightly lower prices.

9. Houses located in Northridge are having more prices compared to other locations.

10. SalePrice vs Condition1 shows most number of houses bearing normal conditions.

11. SalePrice vs Condition2 shows most houses having normal conditions and very few with other conditions.

12. Most houses are Single-family detached and are having higher sale prices than other categories.

13. SalePrice vs HouseStyle shows the houses which are having style of dweling 1-story and 2-story are having higher prices than other types.

14. Many houses are having roof style with gable and hip. and very less number of houses are having shed.

15. many houses are having roof material as standard (Composite) Shingle and houses with roof material as Standard (Composite) Shingle and Wood Shingles are having higher prices.

16. Most houses are having Vinyl Siding as 1st and 2nd covering on the house and are also having higher prices, houses with hard board and cement shilding are also having higher prices.

17. The houses with four Masonry veneer types that are, Brick Common, Brick Face, Cinder Block, Stone. Houses with Brick Common are having lower price.

18. The prices of houses are higher when material used for exterior are good or excellent.

19. Many houses are having cinder block and Poured Concrete foundation and very less houses are having wood foundations, houses with Poured Concrete foundations are having higher prices.

20. Basement quality is mostly average or good and the houses with excellent basement quality are having more prices.

21. Basement exposure is not strongly related to the sale price.

22. Most of the houses are having Heating type as Gas forced warm air furnace and Sale price of houses are higher whenever the quality of heating is excellent.

23. Most houses are having central air conditioning and are having more prices than that of houses which are without air conditioning.

24. Most of the houses with Standard Circuit Breakers & Romex electrical system and are having higher sale prices as well. Very a smaller number of houses are with Mixed type of electrical systems.

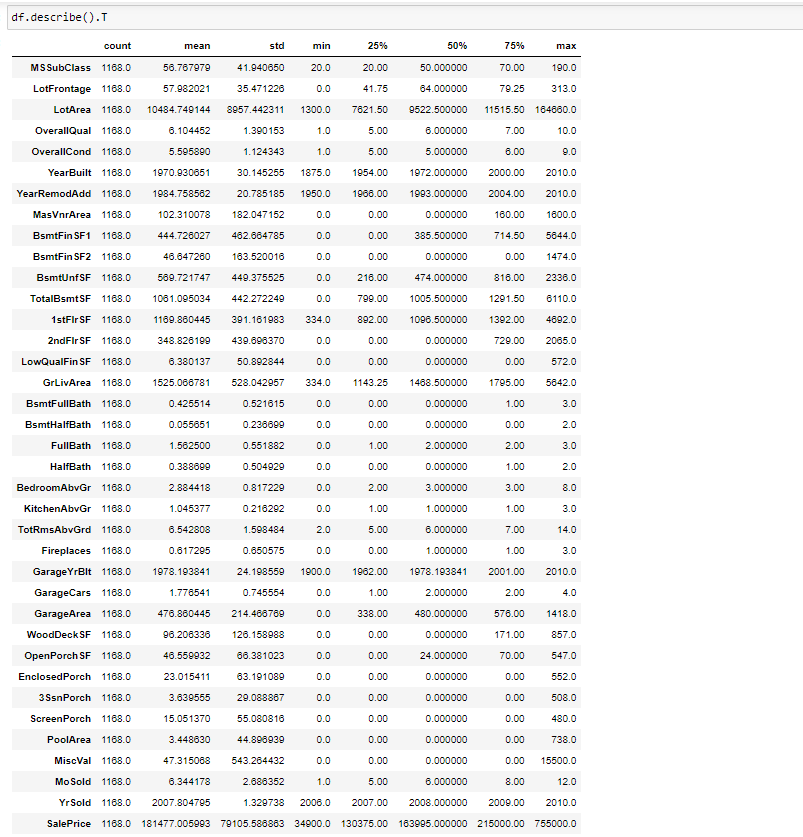
25. Most houses are with good and average kitchen quality, houses are having higher prices when kitchen quality is excellent.

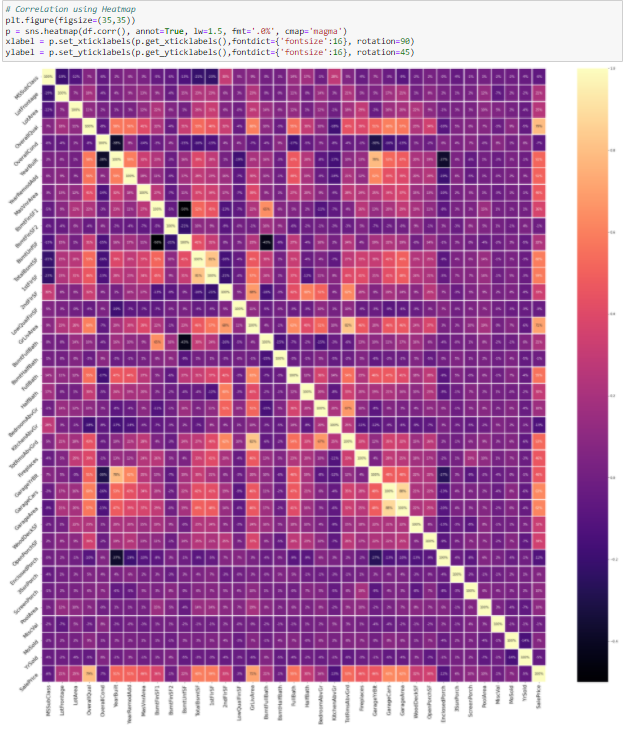
26. In very rare cases fire place are prefabricated fireplace in basement and ben franklin Stove and these houses are having lower prices.

27. In most of the cases garage is attached to the house only. The houses with attached garage or Built-In the house prices are higher.

28. Many houses are having Sale type of Warranty Deed Conventional and just constructed and sold and are having higher prices.

Correlation:





OverallQual, GrLiveArea, GarageCars, GarageArea, TotalBsmtSF, 1st

FlrSF, Full Bath, TotRmsAbvGrd, MasVnrArea, FirePlaces have the

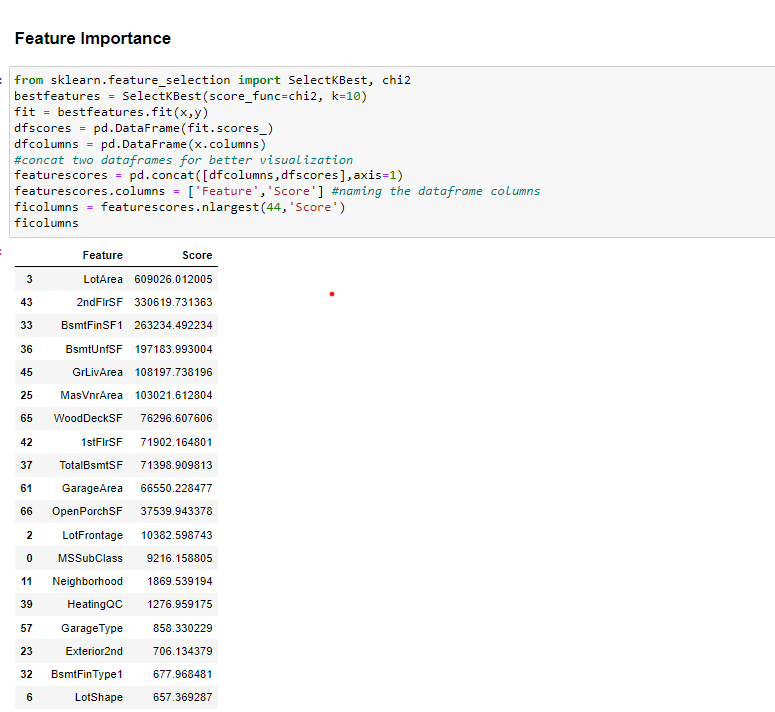
strongest positive correlation with SalePrice while

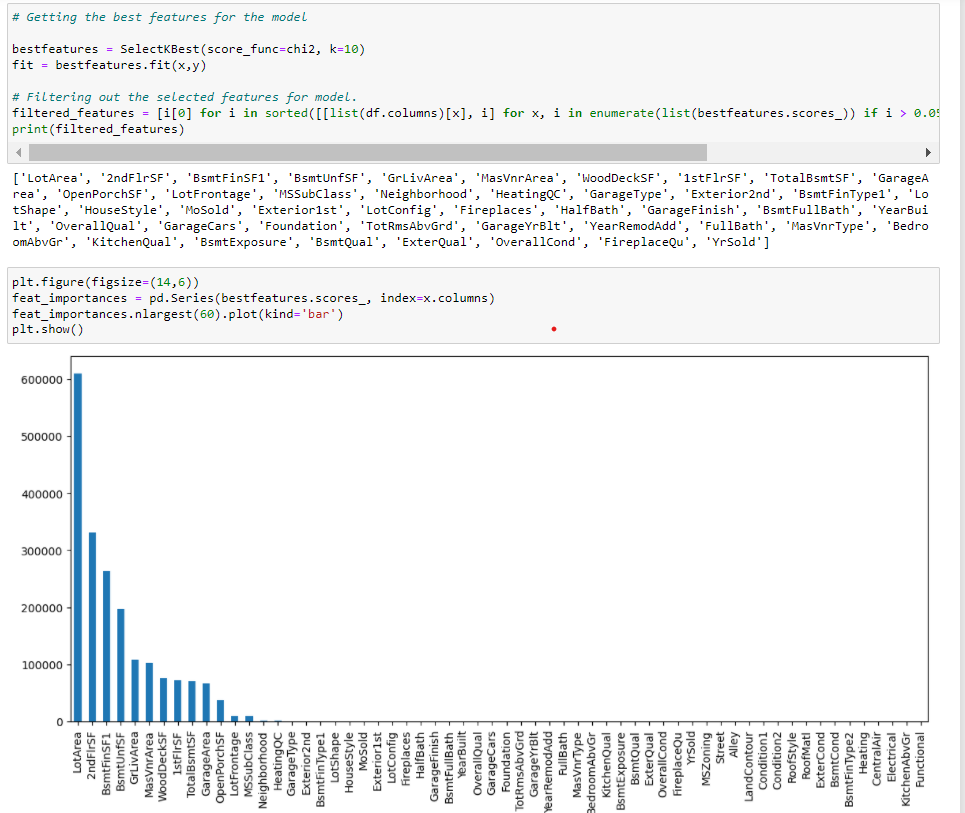
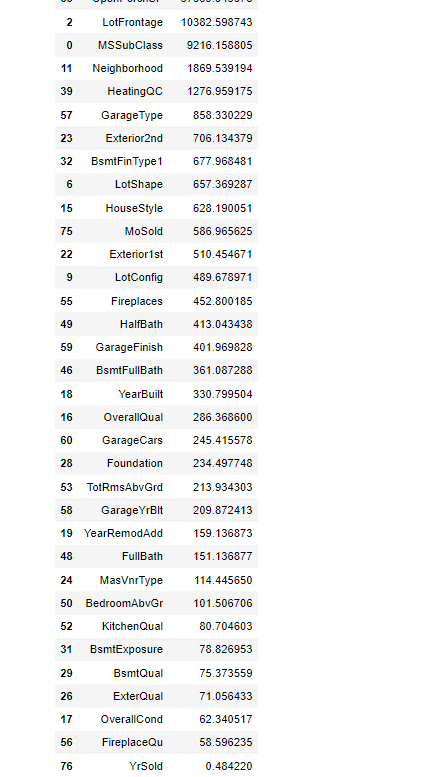
BsmtQual, ExterQual, KitchenQual, GarageFinish, House\_age, Remod\_age, HeatingQC, Garage\_age have the strongest negative

correlation with SalePrice.

* **Model Building:**

**Feature Importance:**

****

****

**Identification of possible problem-solving approaches (methods)** The whole problem-solving approach includes the following steps:

* **Problem Framing:** It includes understanding the problem that is whether the problem is of regression or classification. The present project is of Classification type.
* **Data Understanding:** Data understanding means having an intimate grasp of both the distributions of variables and the relationships between variables. It also includes summary statistics and data visualization.
* **Data Cleaning:** The process of identifying and repairing issues with the data is termed as data cleaning. Statistical methods are used for data cleaning. Some of them are outlier detection and imputation. There are many outliers in the present dataset, so we have imputed outliers with the upper bridge values. We can also remove outliers which may lead to loss of valuable data.
* **Data Selection:** The process of reducing the scope of data to those elements that are most useful for making predictions is called data selection.
* **Data Preparation:** It includes the data to identify the features to be selected and removed. In the present dataset, we have set “Id” column as index and included only 44 features. Before passing the data into the model, we should convert all the categorical data into numerical. So, we Label encoder and convert. In addition to this, the skewness is mitigated by “log transformation” method.
* **Model Evaluation:** Model evaluation consists of identifying input and output variables and splitting the dataset into train and test datasets. The output variable is “label” and the remaining features are input variables. In this project, we have split into 80:20 train and test respectively.
* **Model Configuration:** Hyperparameter tuning the models will get the best fit parameters of each and every model. In this project we use GridSearchCV for knowing the best fit parameters. This can be seen detail in the further sections with a snapshot of it.
* **Model Selection:** The process of selecting one method as the solution is called model selection. It includes the regression model performance metrics.

**Feature Selection:**

Before passing the features into the model, we select some features which is of high importance. The feature selection is shown in below figure.

**Testing of Identified Approaches (Algorithms):**

The Machine Learning Algorithms used in this project for training and testing to predict the prices of the houses are namely:

* Linear Regression
* Lasso Regression
* Ridge Regression
* Decision Tree Regressor

In addition to the above algorithms, ensembling techniques are also use. They are:

* Random Forest Regressor
* Gradient Boosting Regressor

**Linear Regression:**

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, it refers to the method as multiple linear regression.

**Lasso Regression:**

Lasso regression is a type linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models. This particular type of regression is well-suited for models showing high levels of multicollinearity.

**Ridge Regression:**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

**Decision Tree Regressor:**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

**Random Forest Regressor:**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

**Gradient Boosting Regressor:**

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

**Run and Evaluate selected models:**

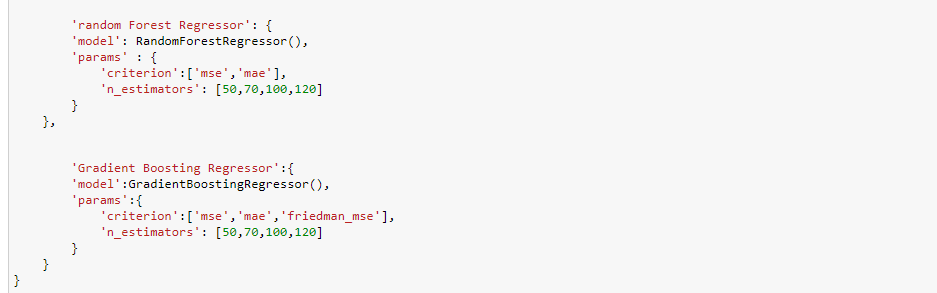
The above-mentioned algorithms have been run in the jupyter notebook and the performance metrics are found as shown in further below figures:

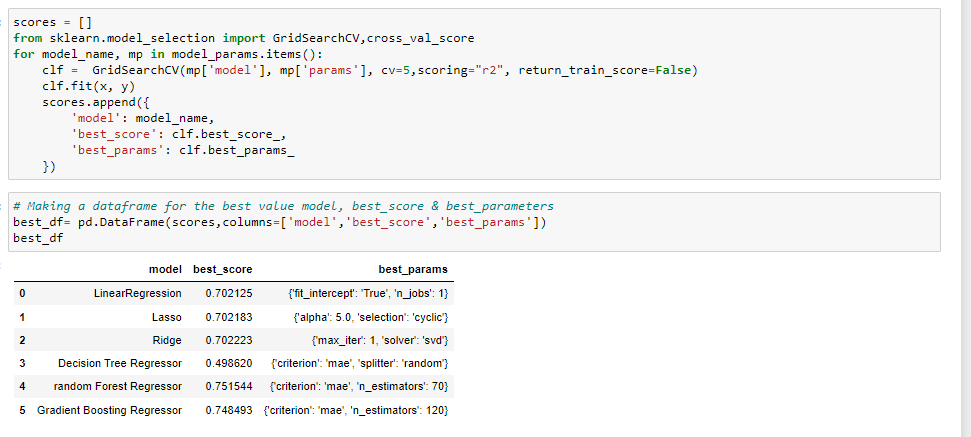
To check whether the model is overfitting/underfitting GridSearchCV is used and cross validated the models as shown in below figures.



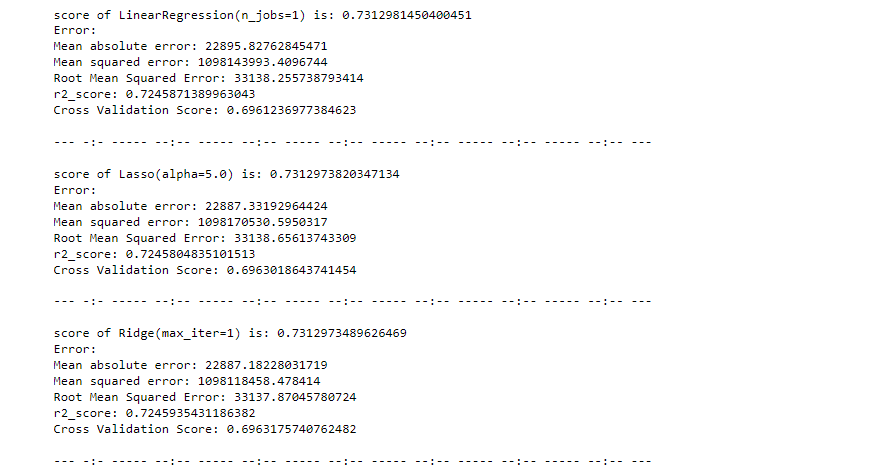
Gradient Boosting Regressor gave the best R2 score with random state of 72.

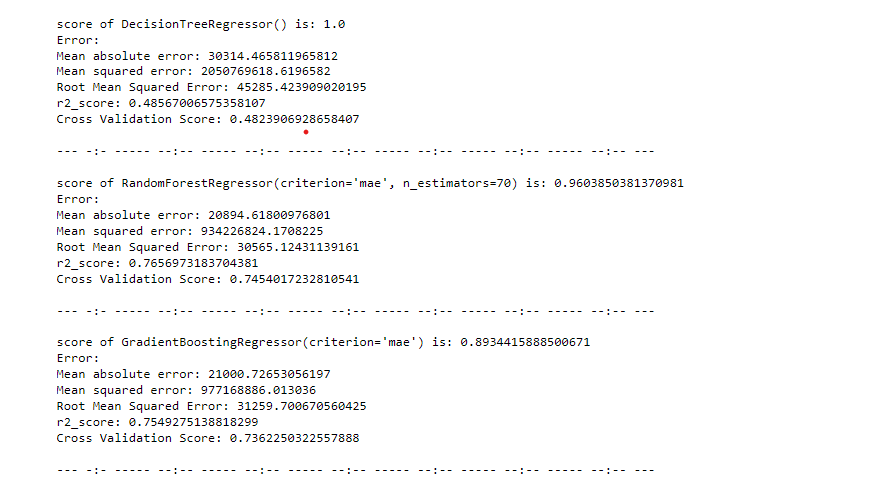






After passing the best parameters into the models we got the following results:





From the figure, we can say that Gradient Boosting Regressor works best.

**Key Metrics for success in solving problem under consideration:**

Evaluating a model is a major part of building an effective machine learning model. The performance metrics of a Regression model are R2 score, mean absolute error, mean squared error, Root Mean Squared Error. Among these R2 score should be in between 0 & 1.

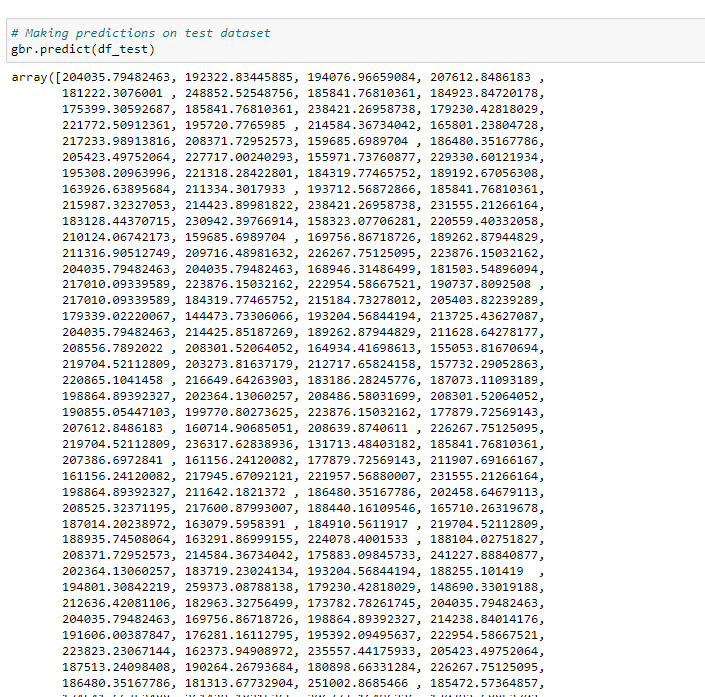
**Train Test Split the data:**



From the figure, we can say that Gradient Boosting Regressor works best. The performance metrics obtained are good.

All the data pre-processing steps are done in test dataset which is similar to train dataset and is executed in a single cell.

The predictions made on the test dataset are shown in below figure: 



**Conclusion:**

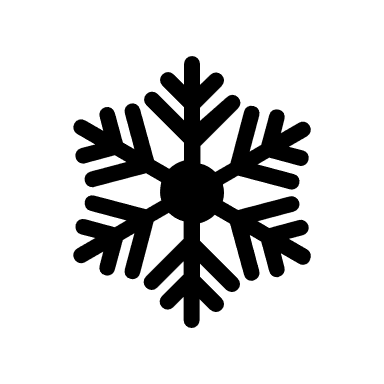
**Key Findings and Conclusions of the Study:**

The goal is to achieve the system which will reduce the human effort to find a house having reasonable price. The proposed system House Price Prediction model approximately tries to achieve the same one. We have managed out how to prepare a model that gives users for a best approach with future lodging value predictions. Proposed system focused on predict the house price according to the area and machine learning methods are used. The experimental results showed that this technique that is used while developing system will give best prediction of house price.

**Learning Outcomes of the Study in respect of Data Science:**

Data Visualization made the project problem easy to understand easy and every feature. While doing the research of the problem, we have clearly identified and imputed the columns with missing values. All the Data pre-processing steps made the problem easier to clean the data. New treatments of outliers are learnt. I have always dropped the outliers but, in this project imputed with other values without losing data.

Five Algorithms are used, in which “Gradient Boosting Regressor” has the best performance metrics with R2 score. We couldn’t reach predicting the house price with R2 score of 1 in practical is the limitation of this work.

----- --:-- ----- ----- --:-- -----  ----- --:-- ---------- --:-- -----