

PFA HOUSING PROJECT

Submitted by:

Anurag Shrivastav

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to the sources Medium, TowardsDataScience, StackOverflow, Krish Naik’s youtube channel which helped me to accomplish this project.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain Problem

The company is looking at prospective properties to buy houses to enter the market. We 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.

* Review of Literature

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.

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.

We 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.

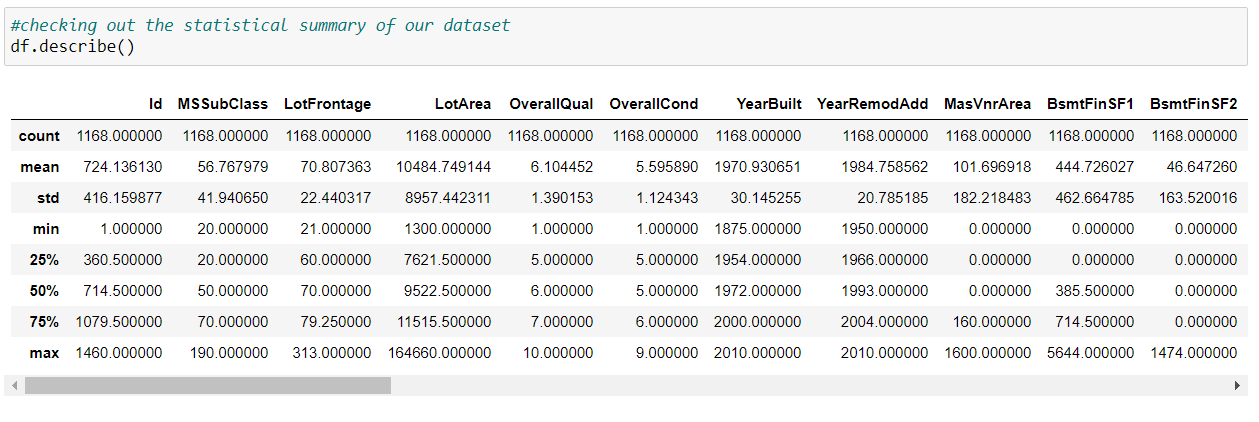
* Motivation for the Problem Undertaken

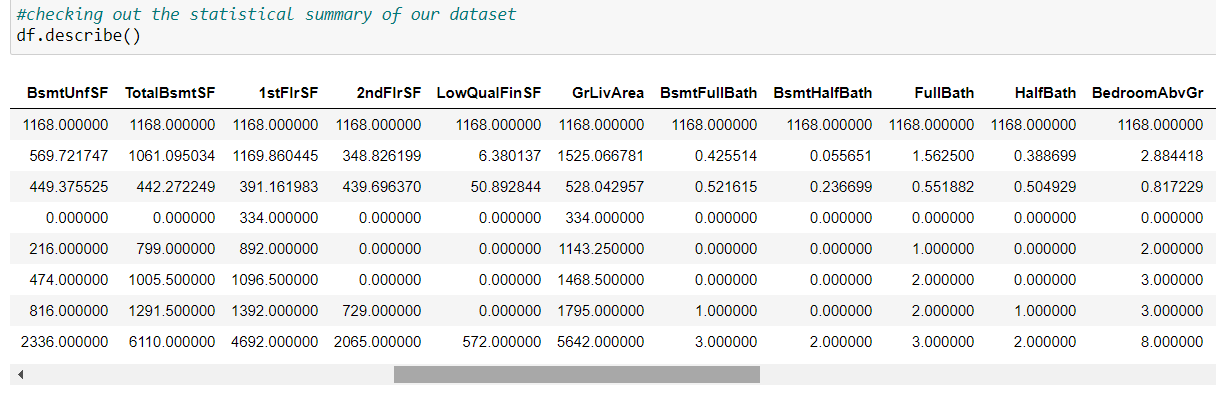
Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science 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. Our problem is related to one such housing company.

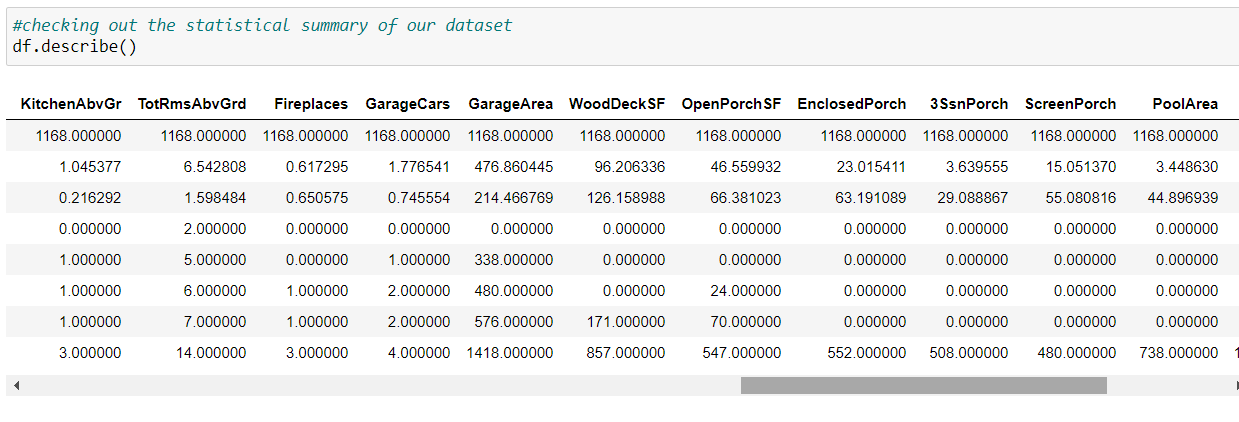
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We first look into the statistics of data shown in fig 1.







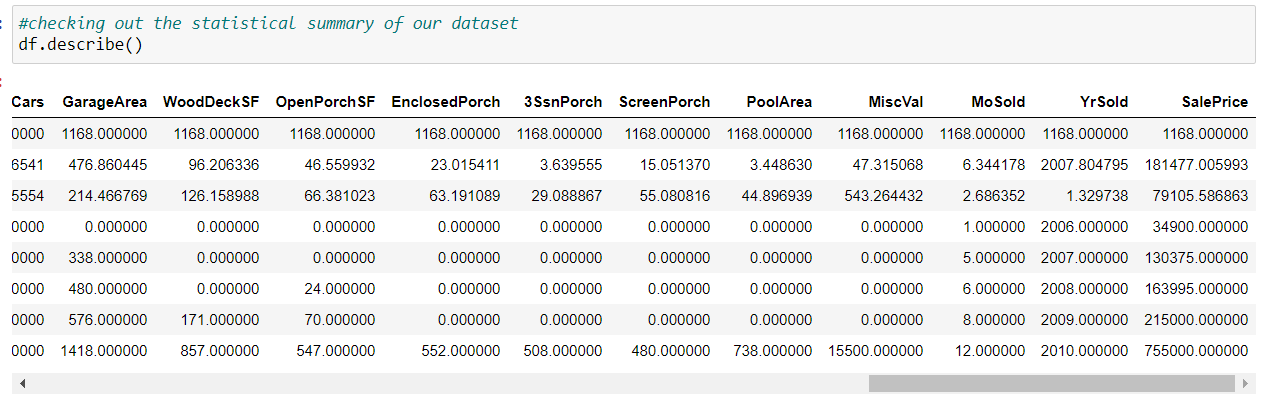


Fig 1 Statastical decription of data

From this statastical analysis we make some of the interpretations that,

1. Maximum standard deviation of 8957.44 is observed in LotArea column.

2. Maximum SalePrice of a house observed is 755000 and minimum is 34900.

3. In the columns Id, MSSubclass, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfsF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, HalfBath, TotRmsAbvGrd, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, Miscval, salePrice mean is considerably greater than median so the columns are positively skewed.

4. In the columns FullBath, BedroomAbvGr, Fireplaces, Garagecars, GarageArea, YrSold Median is greater than mean so the columns are negatively skewed.

5. In the columns Id, MSSubClass, LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtHalfBath, BedroomAbvGr, ToRmsAbvGrd, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, SalePrice there is considerable difference between the 75 percentile and maximum so outliers are present.

* Data Sources and their formats

The variable features of this problem statement are,

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

The data types of features are shown in fig 4,

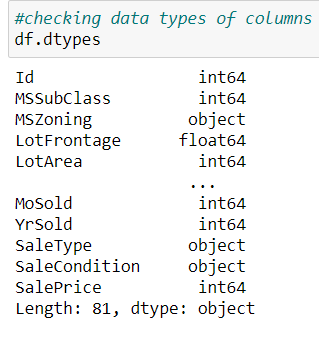


Fig 4 Data types of features

* Data Preprocessing Done

We first done data cleaning. We first looked percentage of values missing in columns then we imputed missing values .

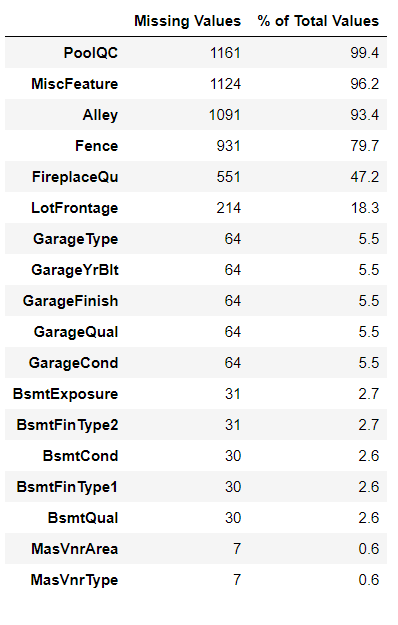


Fig 5 Missing values

We then explored categorical variables as shown in fig 6.

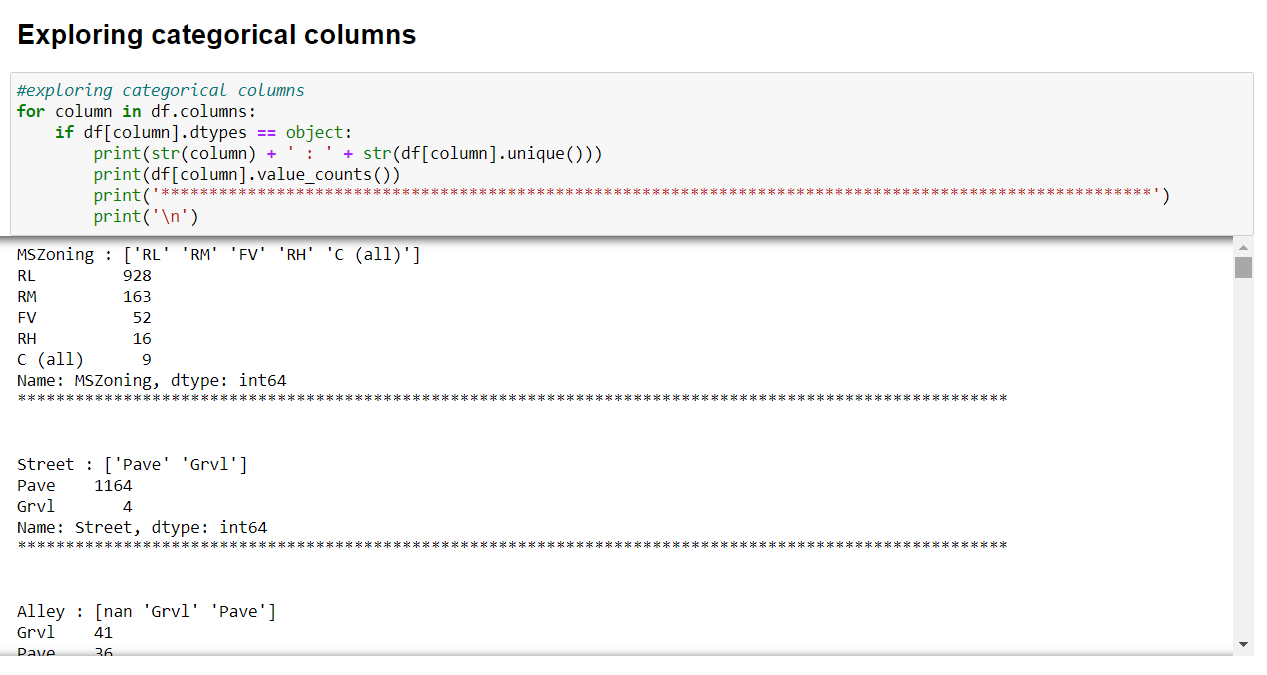


Fig 6 Exploring categorical variables

We observed that there is only one unique value present in Utilities so will be dropping this column. Then we encoded all the categorical columns into numerical colums using dummy variables.

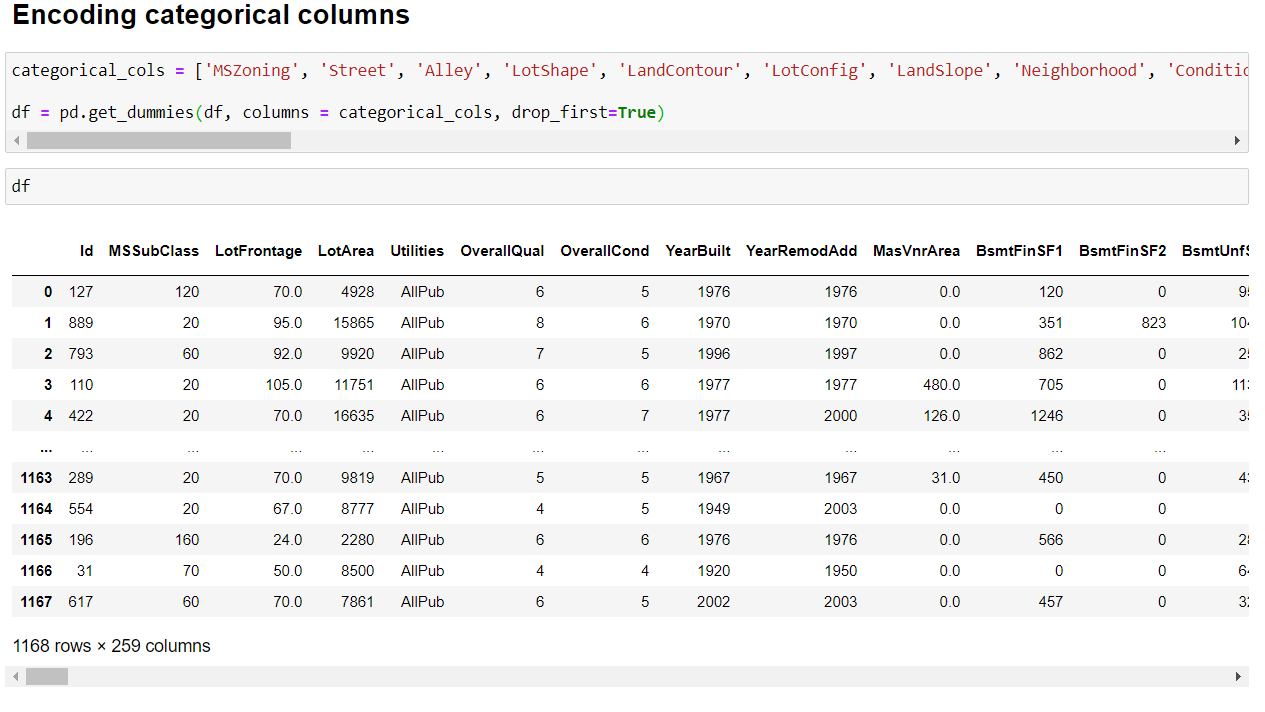


Fig 7 Encoding categorical columns

Then we checked the correlation with the help of heatmap as shown in fig 8,

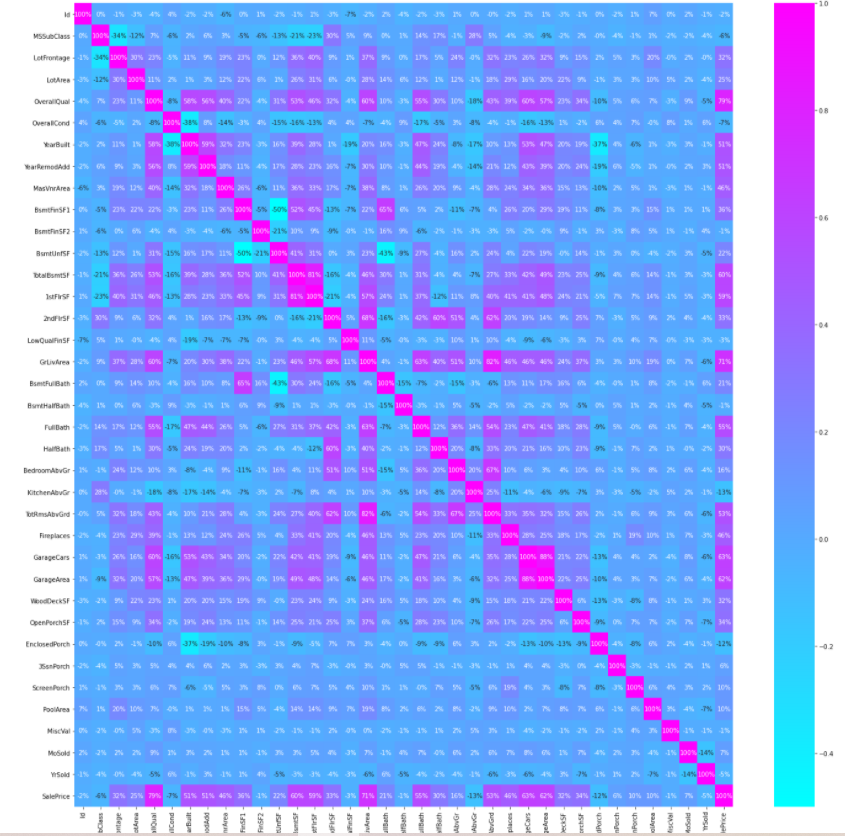


Fig 8 Heatmap of correlation

While checking the heatmap of correlation we observed that,

1. SalePrice is highly positively correlated with the columns OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, 1stFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, GarageCars, GarageArea.

2. SalePrice is negatively correlated with OverallCond, KitchenAbvGr, Encloseporch, YrSold.

3. We observe multicollinearity in between columns so we will be using Principal Component Analysis(PCA).

4. No correlation has been observed between the column Id and other columns so we will be dropping this column.

* Data Inputs- Logic- Output Relationships

Here we check the correlation between all our feature variables with target variable label as shown in fig 10.

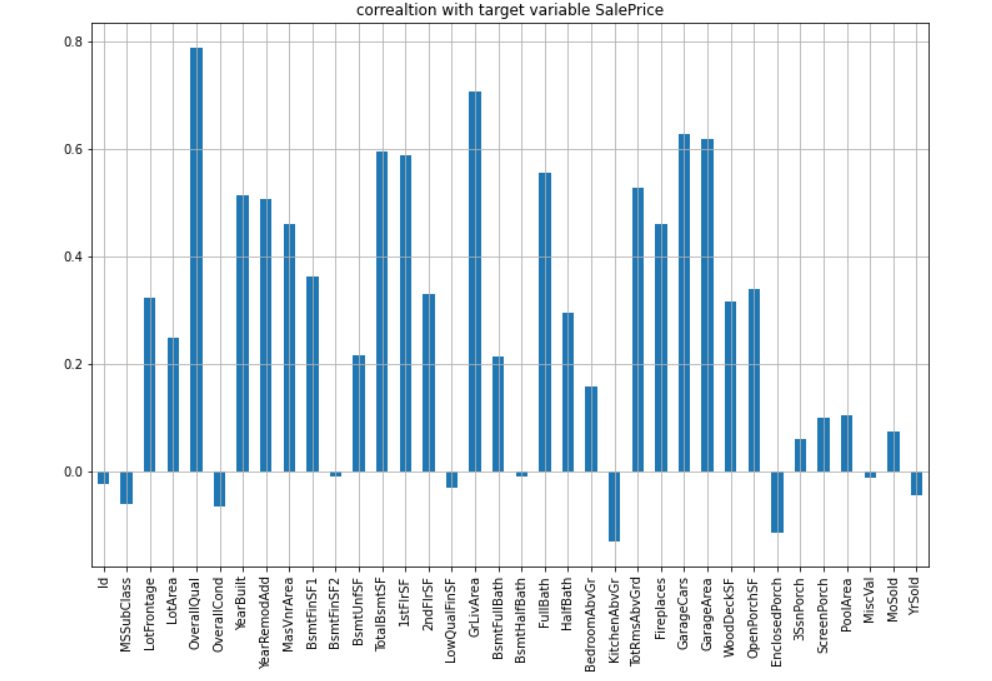


Fig 10 correlation with target variable label

1. The column OverallQual is most positively correlated with SalePrice.

2. The column KitchenAbvGrd is most negatively correlated with SalePrice.

* Set of assumptions related to the problem under consideration

By looking into the target vaariable label we assumed that it was a Regression type of problem.

We observed multicollinearity in between columns so we assumed that we will be using Principal Component Analysis (PCA).

We also observed that only one single unique value was present in Utilities column so we assumed that we will be dropping this columns.

* Hardware and Software Requirements and Tools Used

This project was done on laptop with i5 processor with quad cores and eight threads with 8gb of ram and latest GeForce GTX 1650 GPU on Anaconda, jupyter notebook.

The tools, libraries and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy stats, sklearn.decomposition pca, sklearn standardscaler, GridSearchCV, joblib.

Through pandas library we loaded our csv file ‘Data file’ into dataframe and performed data manipulation and analysis.

With the help of numpy we worked with arrays.

With the help of matplotlib and seaborn we did plot various graphs and figures and done data visualization.

With scipy stats we treated outliers through winsorization technique.

With sklearn.decomposition’s pca package we reduced the number of feature variables from 256 to 100 by plotting scrre plot with their Eigenvalues and chose the number of columns on the basis of their nodes.

With sklearn’s standardscaler package we scaled all the feature variables onto single scale.

Through GridSearchCV we were able to find the right parameters for hyperparameter tuning.

Through joblib we saved our model in csv format.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We first converted all our categorical variables to numeric variables with the help of dummy variables to checkout and dropped the columns which we felt were unnecessary.

We observed skewness in data so we tried to remove the skewness through treating outliers with winsorization technique.

The data was improper scaled so we scaled the feature variables on a single scale using sklearn’s StandardScaler package.

There were too many (256) feature variables in the data so we reduced it to 100 with the help of Principal Component Analysis(PCA) by plotting Eigenvalues and taking the number of nodes as our number of feature variables.

* Testing of Identified Approaches (Algorithms)

The algorithms we used for the training and testing are as follows:-

* + Linear Regression
  + Lasso
  + Ridge
  + Elastic Net
  + SVR
  + KNeighbors Regressor
  + Decision Tree Regressor
  + Random Forest Regressor
  + Ada Boost Regressor
  + Gradient Boosting Regressor

* Run and Evaluate selected models

The algorithms we used are shown in fig 11,

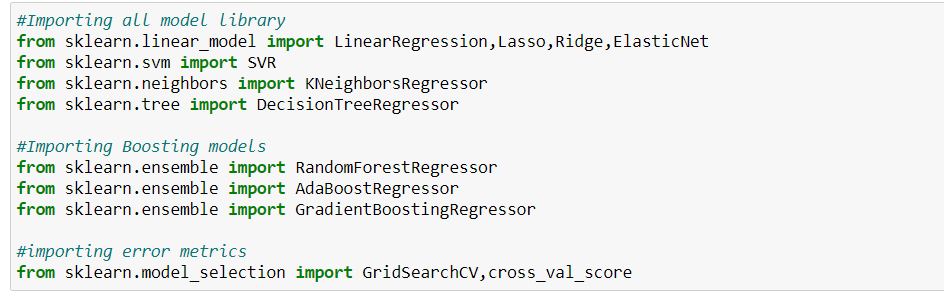
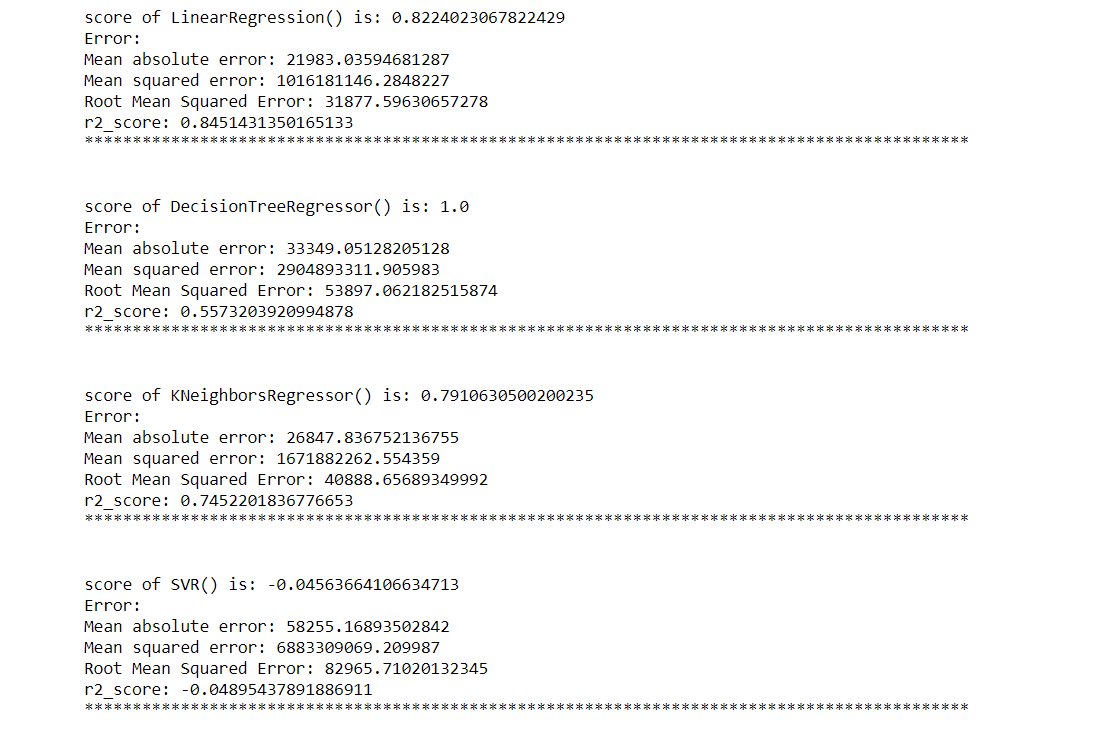
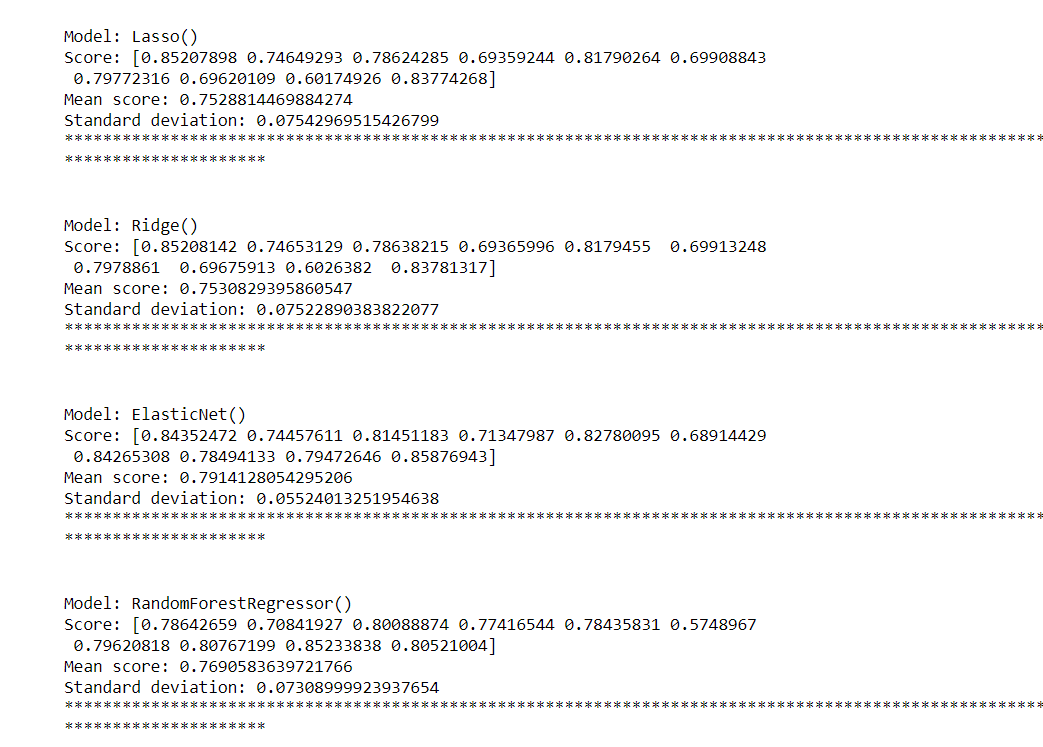


Fig 11 Algorithms used

The results observed over different evaluation metrics are shown in fig 12,





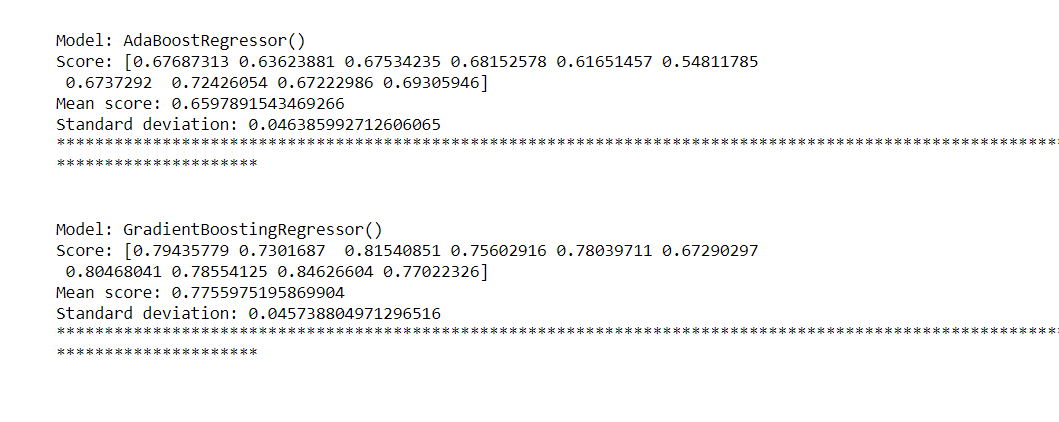


Fig 12 Results observed

* Key Metrics for success in solving problem under consideration

we used the metric Root Mean Squared Error by selecting the Ridge Regressor model which was giving us best(minimum) RMSE score.

* Visualizations

ViolinPlot of SalePrice :-

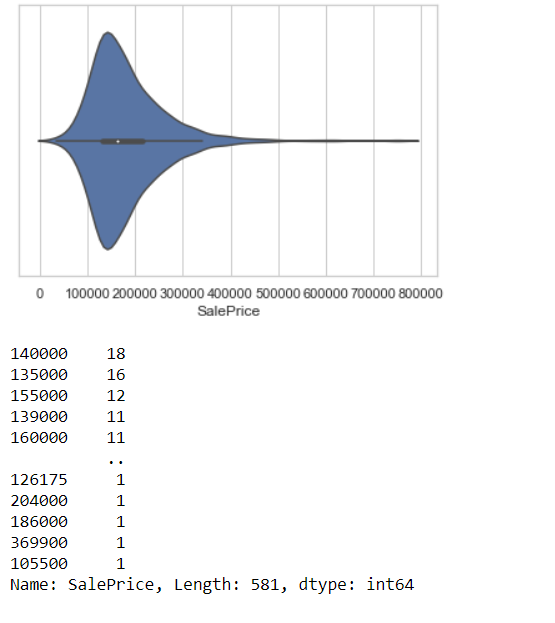


Fig 13 VioilinPlot of SalePrice

Observation:

Maximum number of SalePrice lies between 140000 and 230000.

Countplot of MSZoning:-

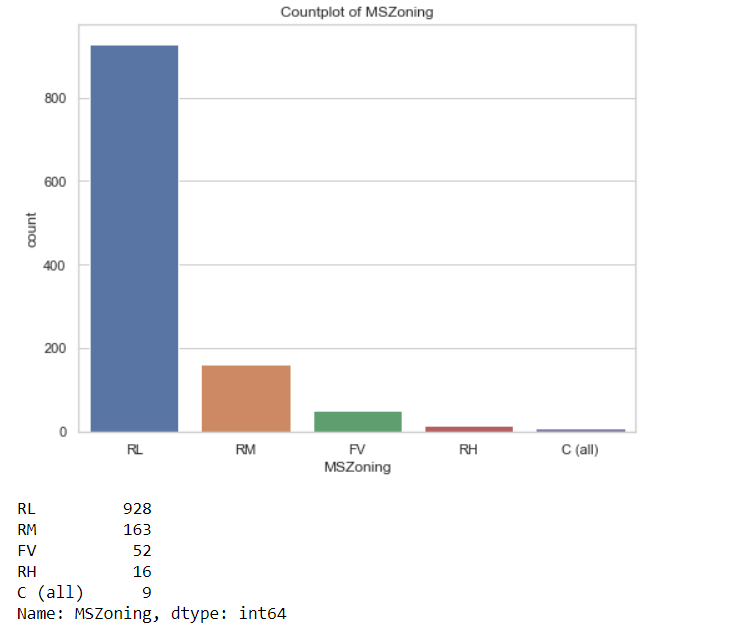


Fig 14 Countplot of MSZoning

Observation:

Maximum, 928 number of MSZoning are RL.

Checking column LotShape with SalePrice:-

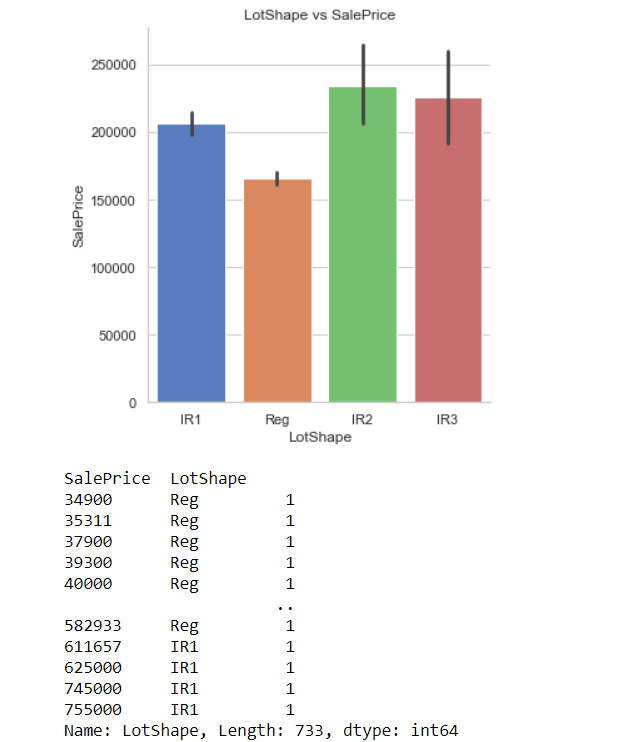


Fig 15 LotShape vs SalePrice

Observation:

SalePrice is maximum with IR2 LotShape.

Checking column Neighborhood with SalePrice:-

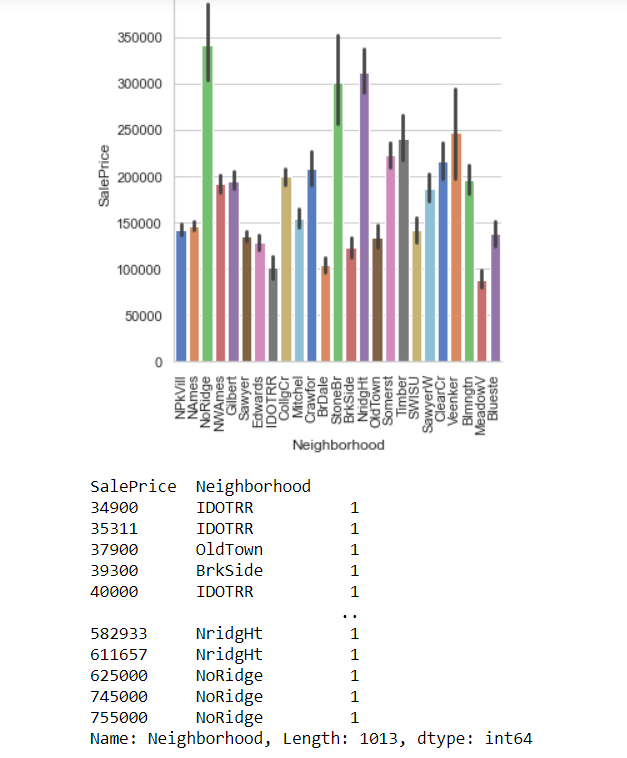


Fig 16 Neighborhood vs SalePrice

Observation:

SalePrice is maximum with NoRidge Neighborhood.

Checking the column HouseStyle with SalePrice:-

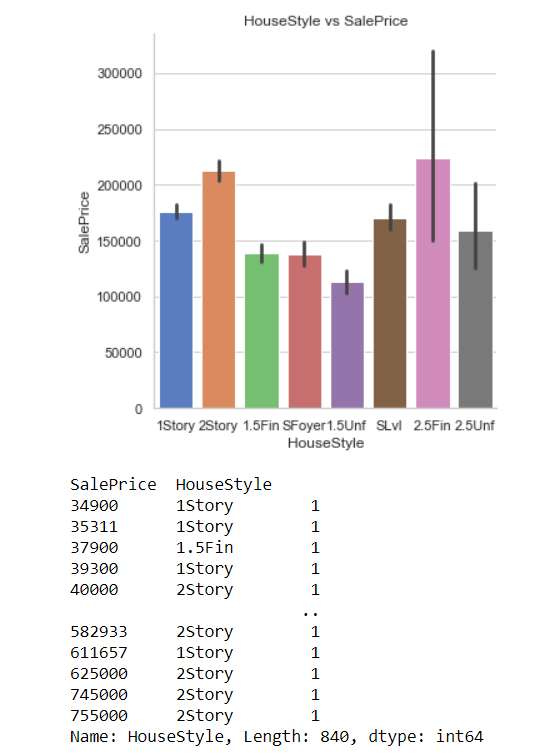


Fig 17 HouseStyle vs SalePrice

Observation:

SalePrice is maximum with 2.5Fin HouseStyle.

Checking KitchenQual and CentralAir with SalePrice:-

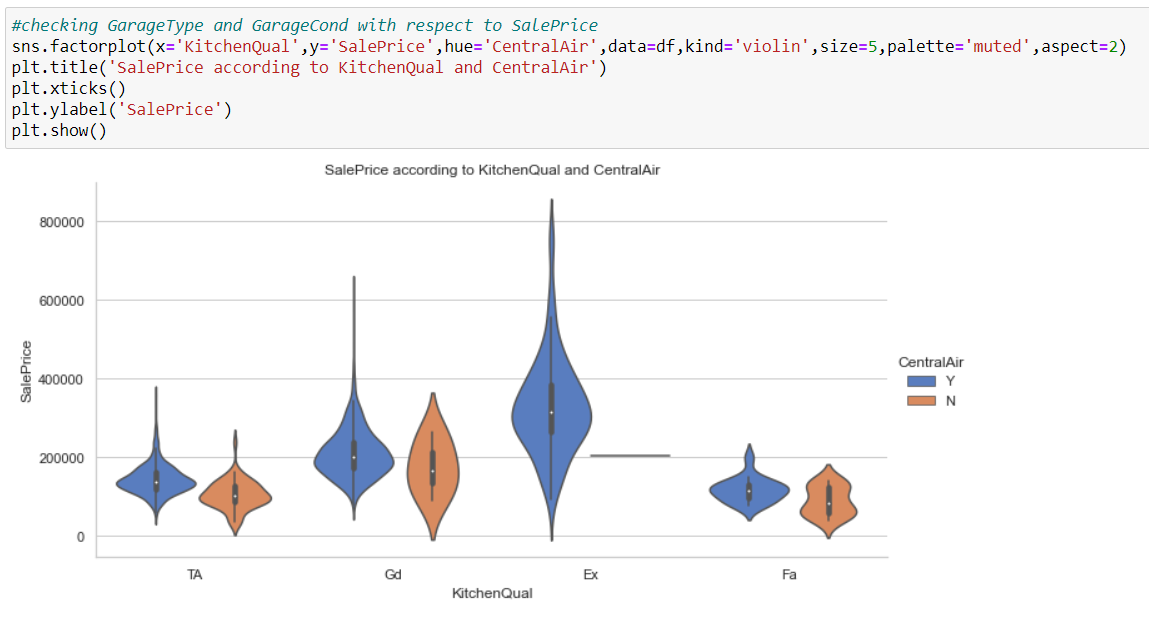


Fig 18 ViolinPlot between KitchenQual and CentralAir with respect to SalePrice

Observation:

SalePrice is maximum with Ex kitchenQual and CentralAir.

Checking SalePrice, OverallQual, YearBuilt, GrLivArea, GarageCars:-

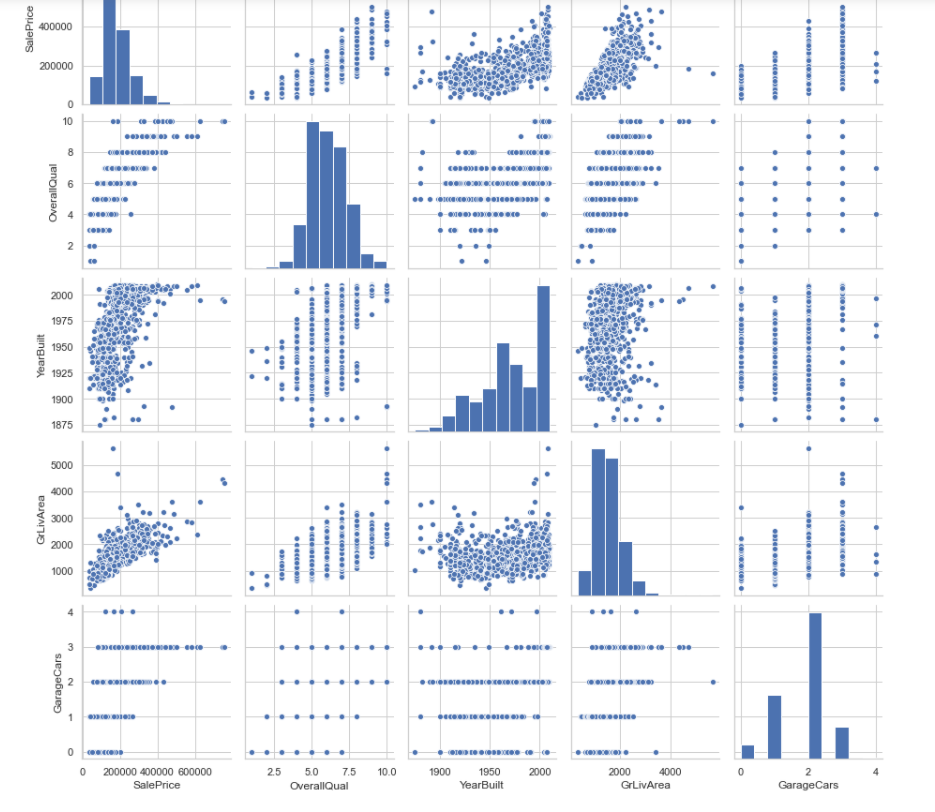


Fig 19 pairplot

Observation:

SalePrice is highly positively correlated with GrLivArea and OverallQual.

* Interpretation of the Results

From the visualization we interpreted that the target variable SalePrice was highly positively correlated with the columns GrLivArea, YearBuilt, OverallQual, GarageCars, GarageArea.

From the preprocessing we interpreted that data was improper scaled.

From the modeling we interpreted that after hyperparameter tuning Ridge Regressor works best with respect to our model with minimum RMSE of 31806 as shown in fig 20

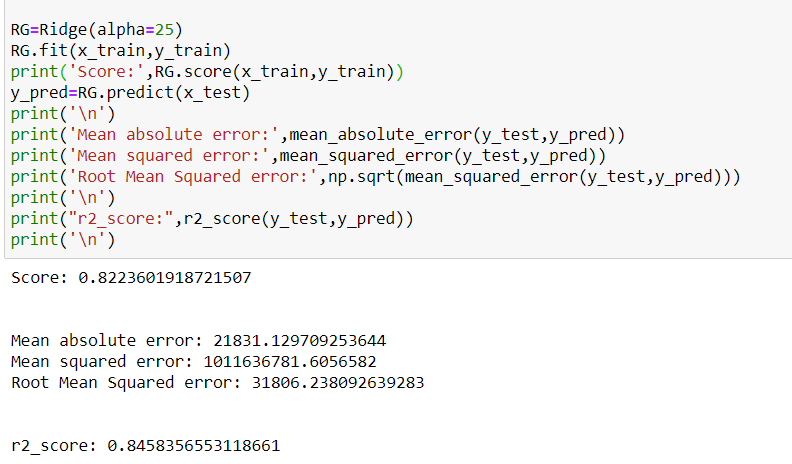


Fig 20 score of Ridge after Hyperparameter tuning.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project we have tried to show how the house prices vary and what are the factors related to the changing of house prices.The best(minimum) RMSE score was achieved using the best parameters of Ridge Regressor through GridSearchCV though Lasso Regressor model performed well too.

* Learning Outcomes of the Study in respect of Data Science

This project has demonstrated the importance of sampling effectively, modelling and predicting data.

Through different powerful tools of visualization we were able to analyse and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The few challenges while working on this project where:-

* + Improper scaling
  + Too many features
  + Missing values
  + Skewed data due to outliers

The data was improper scaled so we scaled it to a single scale using sklearns’s package StandardScaler.

There were too many(256) features present in the data so we applied Principal Component Analysis(PCA) and found out the Eigenvalues and on the basis of number of nodes we were able able to reduce our features upto 90 columns.

There were lot of missing values present in different columns which we imputed on the basis of our understanding.

The columns were skewed due to presence of outliers which we handled through winsorization technique.

* Limitations of this work and Scope for Future Work

While we couldn’t reach out goal of minimum RMSE in house price prediction without letting the model to overfit, we did end up creating a system that can with enough time and data get very close to that goal. As with any project there is room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and vesatility to the project.