

HOUSING PRICE PREDICTION PROJECT

Submitted by:

JASMINE FERNANDES

**ACKNOWLEDGMENT**

Following are the acknowledgement sources:

My lecture notes and codes

You tube

**INTRODUCTION**

* Business Problem Framing

Real estate is not a very straight forward industry, especially when entering a new market in a new country. The challenge is to purchase houses at a price below their actual values and resell them at a higher price. Before investing a large sum of capital, the aim is to perform data analysis on the collected data set from the sale of houses in Australia to make an informed decision related to the housing prices.

* Conceptual Background of the Domain Problem

The company is finding perspective properties to enter the market. The aim is to predict the actual value of the prospective properties and decide whether to invest in them or not.

* Motivation for the Problem Undertaken

The motive behind this analysis to pick the right properties for investing In order to do so, the objective is to create an effective machine learning model that is able to accurately estimate the price of the house based on given features given and use these insights for investing. The purpose of this the report is to find the best method to predict with real time factors that impact the housing industry.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

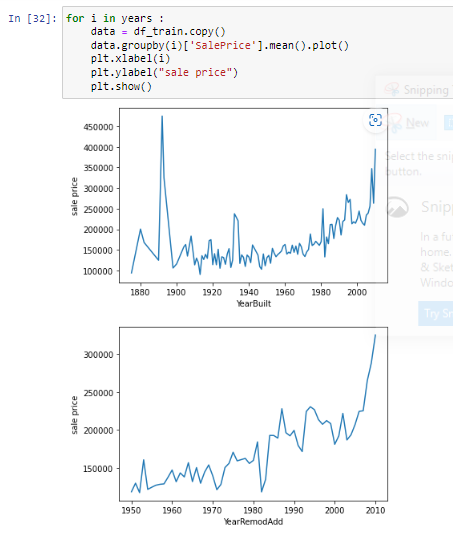
Linear regression modelling is used as we are dealing with a continuous variable. Various techniques to build models are used: Random Forest, Linear regression, Decision Tee and Adaboost. The main aim of using Linear Regression model find the best fit linear line such that the error is minimized. Error is the difference between the actual value and Predicted value.

* Data Sources and their formats

Following formats are used:

* Scatter plots
* Correlation plot
* Skewness table
* Used histogram to view distribution of all the numeric variables

All data related to ‘year’ is separated to find their correlation with Sales price



* Data Preprocessing Done

Following steps are used for data cleaning:

1. Dropped the ‘cust id’ from the test data and train data as it does not impact the outcome. It is an unnecessary detail.
2. Searched for Nan values with isna().sum() function. LotFrontage has imbalanced data
3. Searched for duplicate columns, but few columns had many missing values and data would be lost(data set is as it is small). Hence other methods were used
4. Finding if there are any unique values in object data types and separating object type columns
5. Added numerical values to the object type columns for modelling
6. Viewing distribution with histogram for numerical columns
7. Separating float and int64 values and checking for any missing value
8. #GarageYrBlt , LotFrontage have missing values, we added mode and mean method
9. Checked for any missing data pattern
10. Segregated all year columns - 'YearBuilt', 'YearRemodAdd', 'GarageYrBlt', 'YrSold' and tried to vew their relationship with Sales price
11. Used correlation matrix to find correlations

* Data Inputs- Logic- Output Relationships
* With the following code , I established relations between all object type data and target column – ‘Sales Price’:

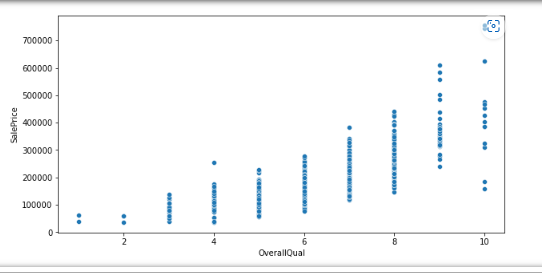
for col in df\_train.columns:

if df\_train[col].dtype != 'object':

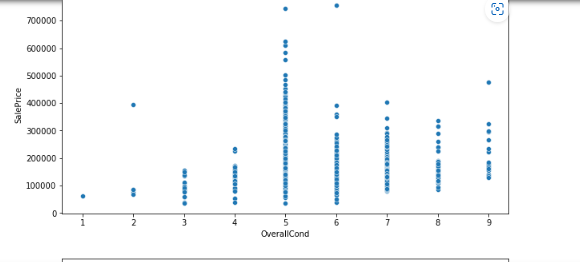
plt.figure(figsize=(12,6))

sns.scatterplot(x=col,y='SalePrice',data=df\_train)

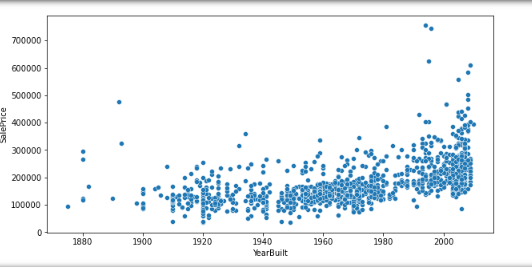
* Sales Price and OverallQual(Rates the overall material and finish of the house) : Sales price is highest for 9-10(10 stands for excellent) –



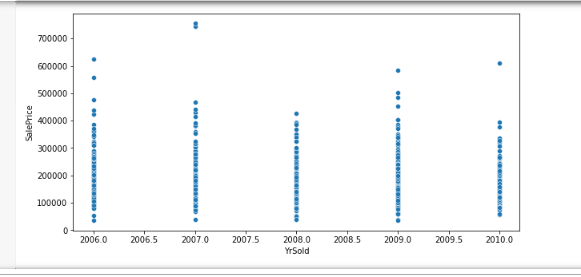
* Sales Price vs OverallCond(Rates the overall condition of the house): The sales price is highest for 5, which is average



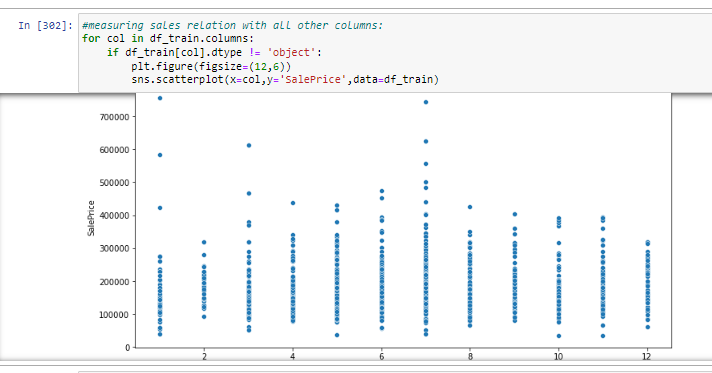
* Sales Price vs YearBuilt(Original construction date): The year 1997-98 has the highest sales price



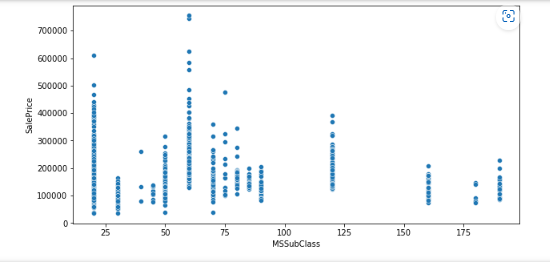
* Sales Price vs YearRemodAdd: Remodel date (same as construction date if no remodeling or additions) - The year 1997-98 has the highest sales price
* Sales Price vs YrSold(Year Sold (YYYY): - The year 2007 has the highest sales price



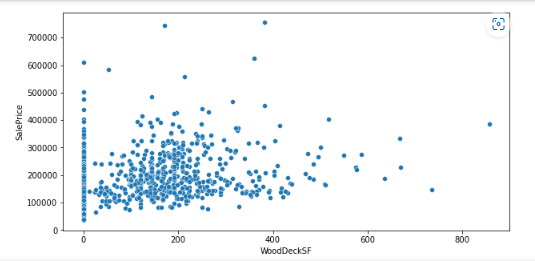
* Sales Price vs Month Sold (MM)): - The month July, 7th month has the highest sales price



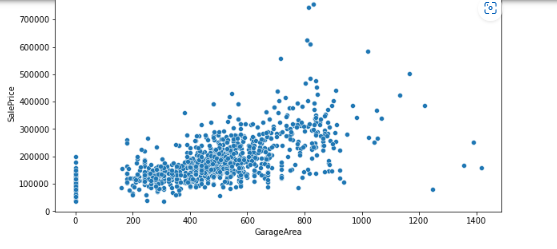
* Sales Price vs MSSubClass (Identifies the type of dwelling involved in the sale.)): - 60 has the highest sales price. 60 stands for 2-STORY 1946 & NEWER

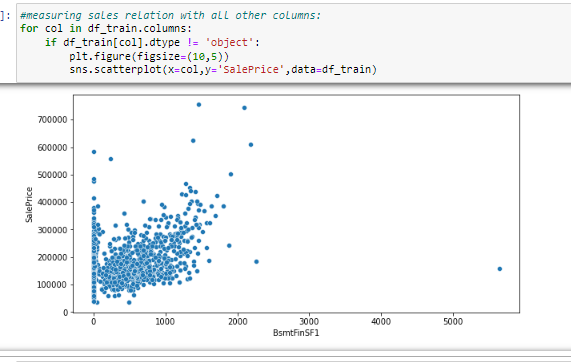


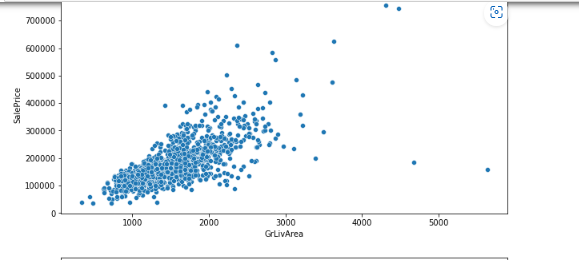
* Sales Price vs **WoodDeckSF**(**Wood deck area in square feet)**: - 380-400 sq ft has the highest sales price



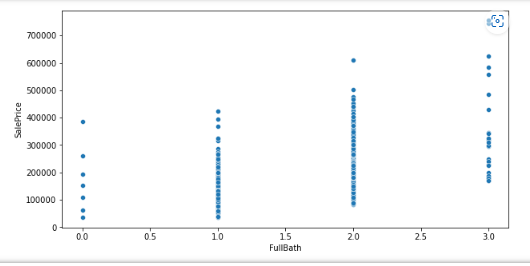
* Sales Price vs Garage area: 800-820 sq ft has the highest sales price



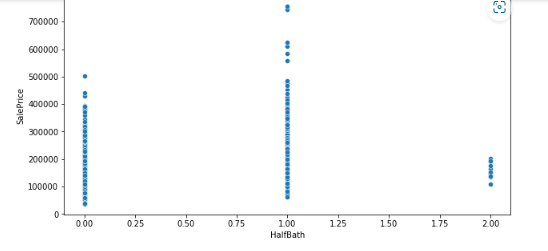
* Sales Price vs TotalBsmtSF-Total square feet of basement area: majority of the basement area available is between **0-1000 sqft** 
* Sales Price vs GrLivArea(Above grade (ground) living area square feet )- majority of the basement area available is between **1000-2000 sqft and costs between 100000-300000**



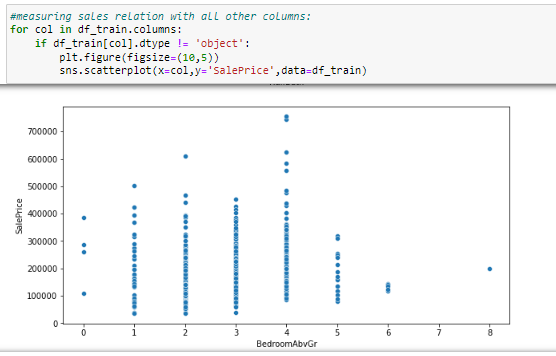
* Sales Price vs FullBath(Basement full bathrooms)-The category 3- full bath basement bathrooms is sold for the highest price



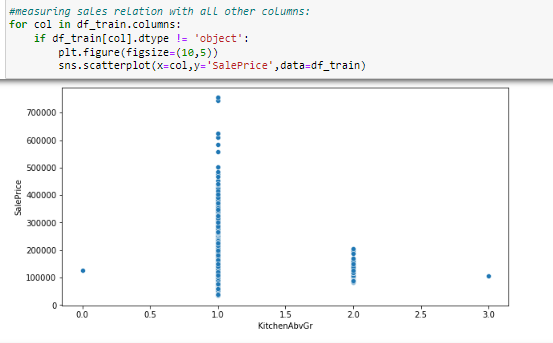
* Sales Price vs HalfBath(Basement half bathrooms)-The ‘1’ category half bath basement bathrooms is sold for the highest price



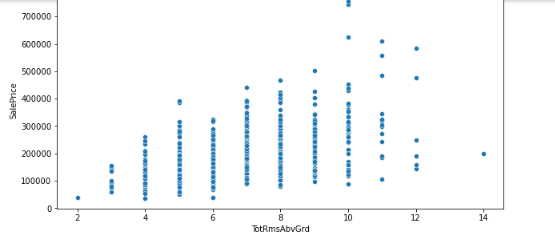
* Sales Price vs BedroomAbvGr:The ‘4’ category is sold for the highest price



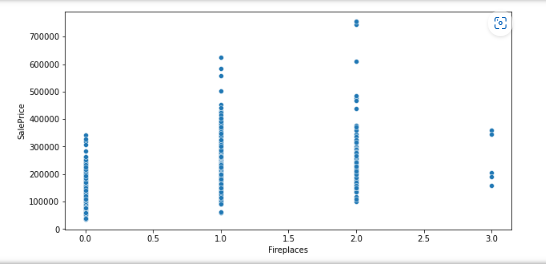
* Sales Price vs KitchenAbvGr:The ‘1’ category is sold for the highest sales price



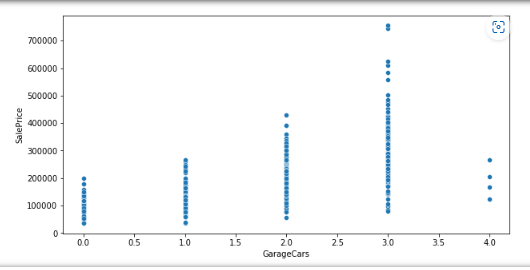
* Sales Price vs TotRmsAbvGr:The ‘10’ category is sold for the highest sales price



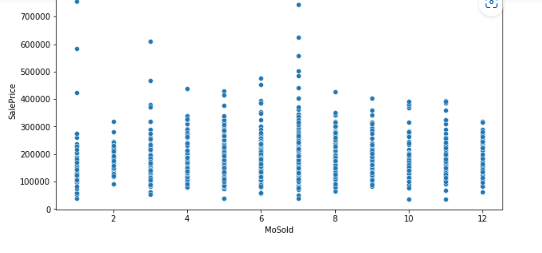
* Sales Price vs Fireplaces: The ‘2’ category is sold for the highest sales price



* Sales Price vs GarageCars: The ‘2’ category is sold for the highest sales price

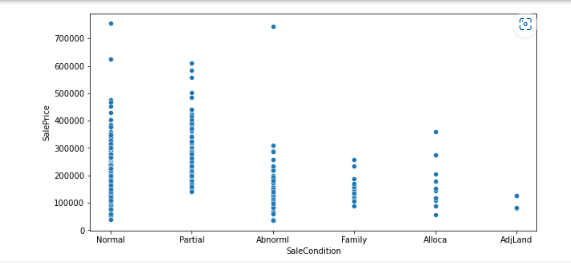


* Sales Price vs MoSold(month sold): The ‘7th ’ category month has the highest sales price

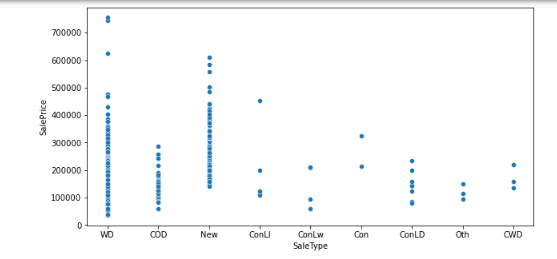


**RELATIONS BETWEEN NUMERICAL COLUMNS & SALES PRICE**

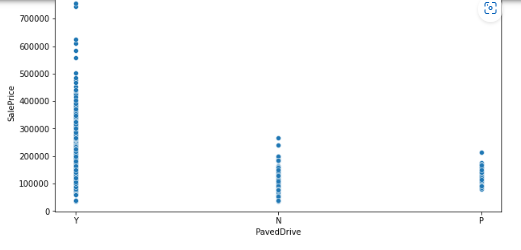
1. Sales Price vs Sales Condition: The normal and partial category fetches the highest sales price



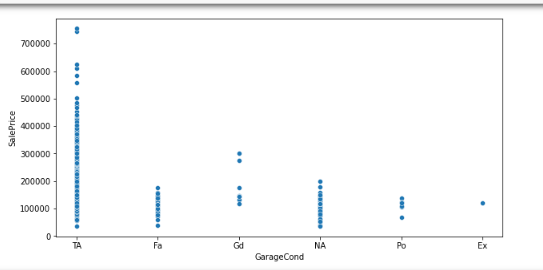
1. Sales Price vs Sales Type: WD(Warranty Deed – Conventional) fetches the highest sales price



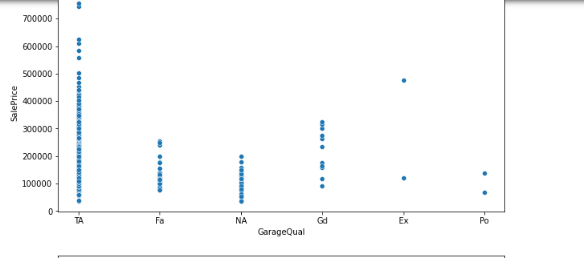
1. Sales Price vs Paved Drive: Paved fetches the highest sales price



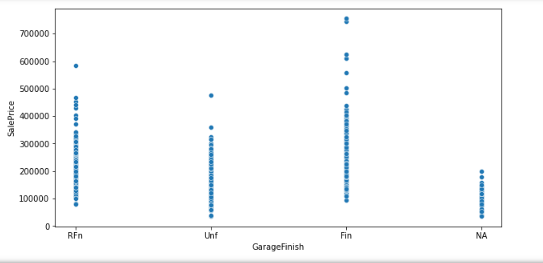
1. Sales Price vs GarageCond: TA fetches the highest sales price



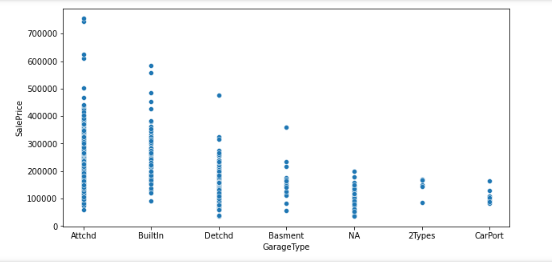
1. Sales Price vs GarageQual: TA fetches the highest sales price



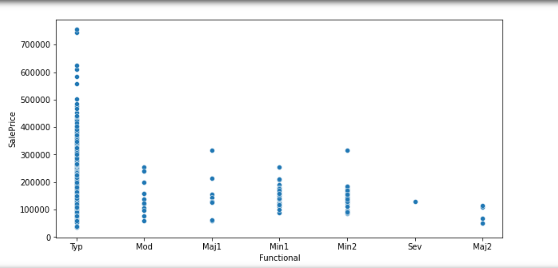
1. Sales Price vs GarageFinish: Fin fetches the highest sales price

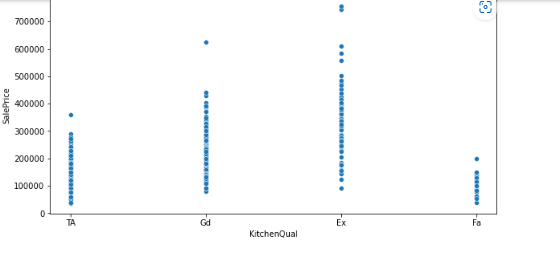


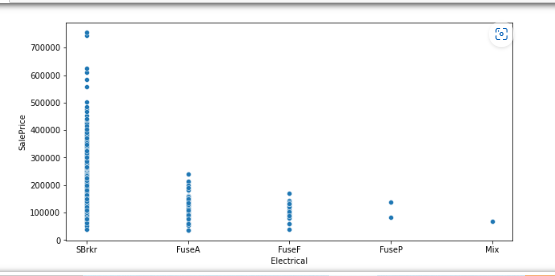
1. Sales Price vs GarageType: Attchd fetches the highest sales price

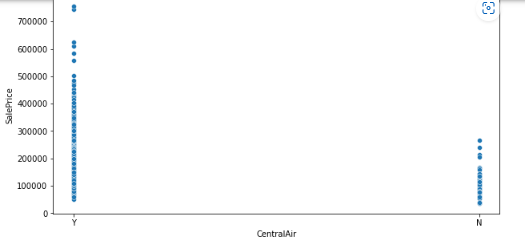


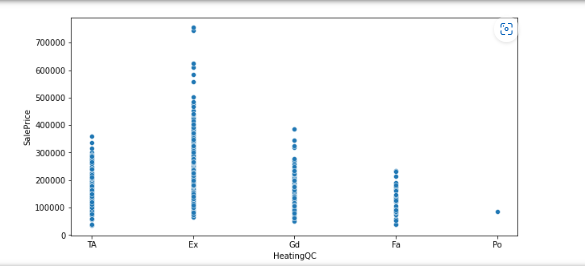
1. Sales Price vs Functional : Typ fetches the highest sales price

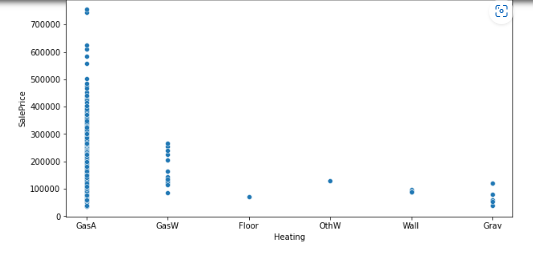


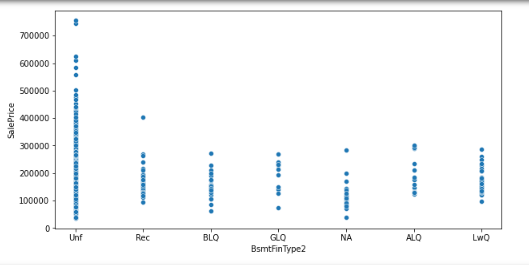


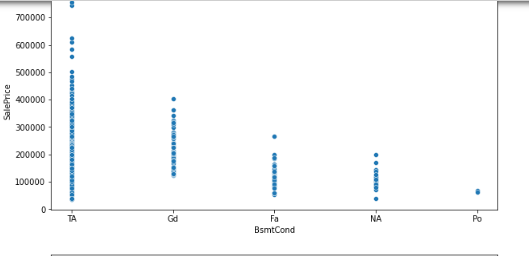


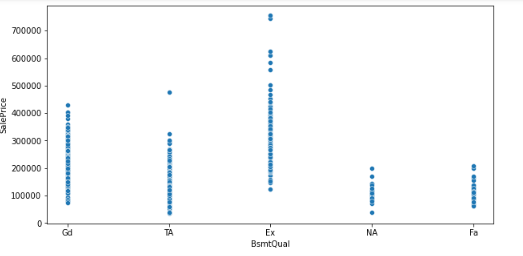


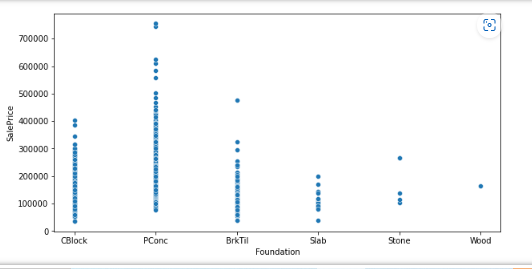


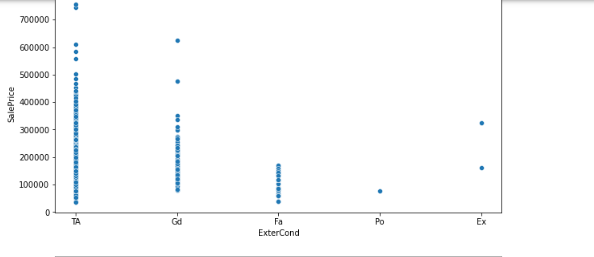


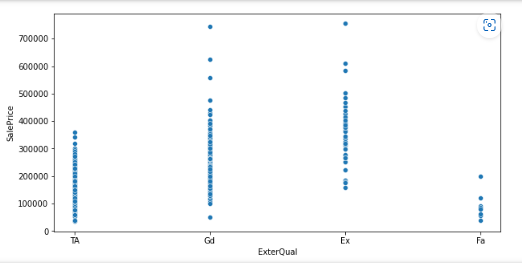


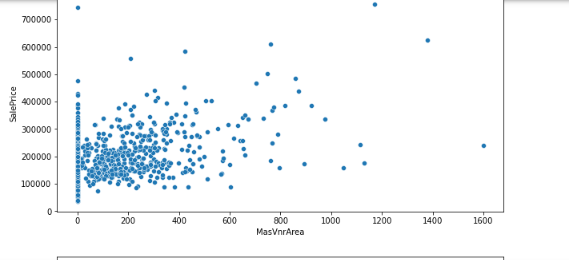


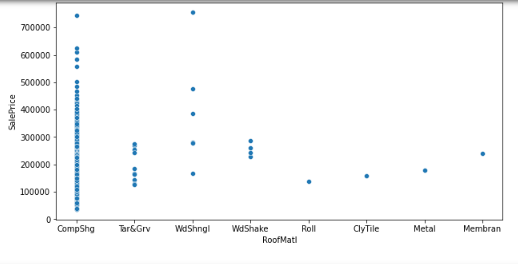


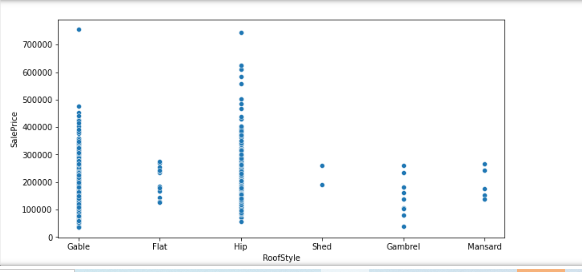


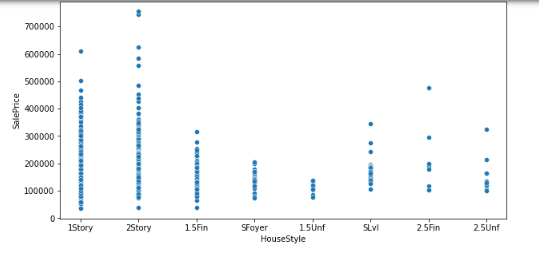


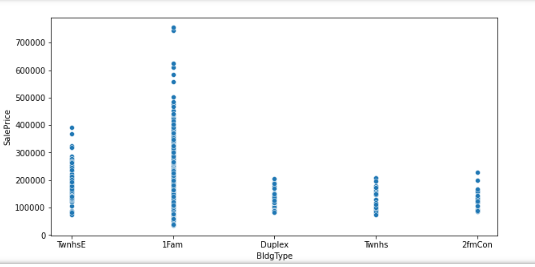


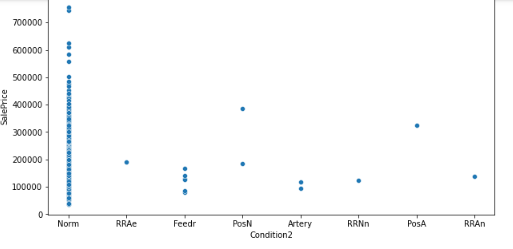


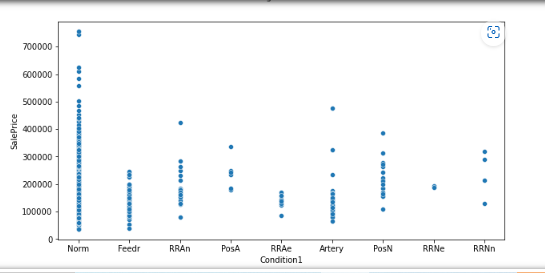


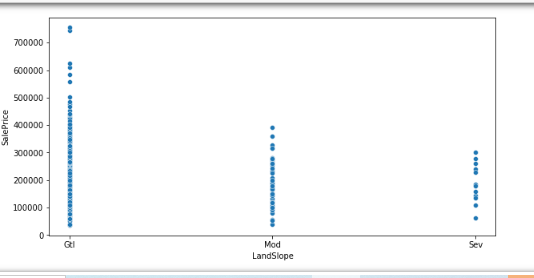


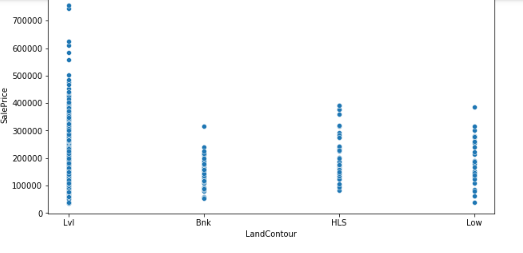


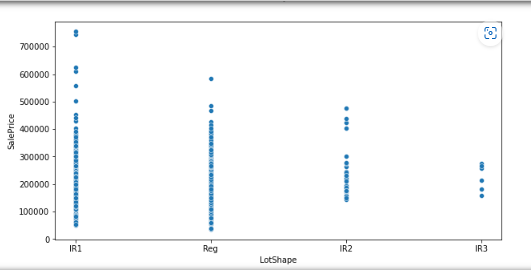


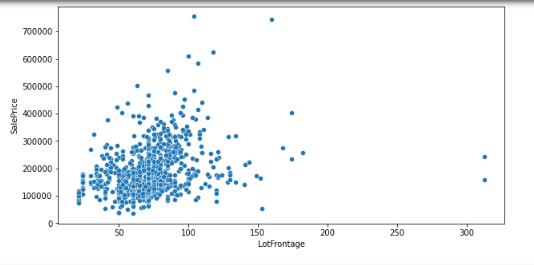




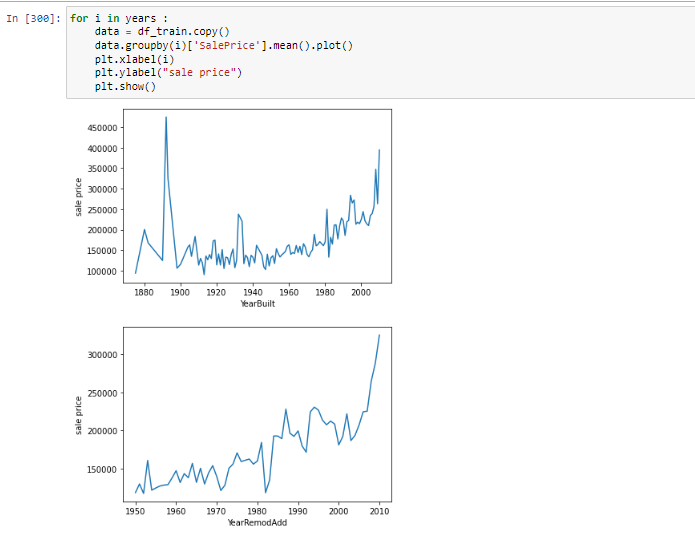


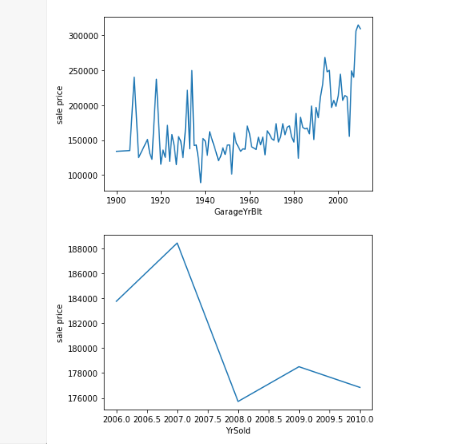
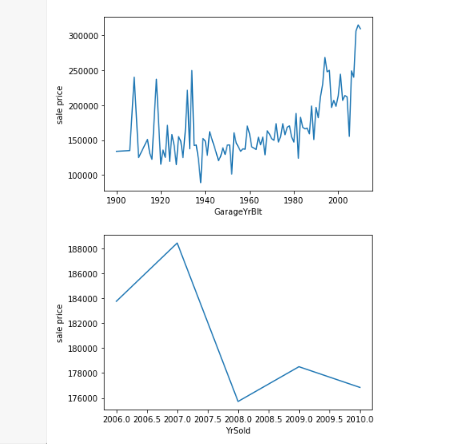






**RELATION ESTABLISHED WITH YEARS COLUMNS AND SALES PRICE**



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I used the chi and percentile method to decrease the number of features. I have taken only 80% of features

**Following algorithms used for the training and testing:**

Random Forest

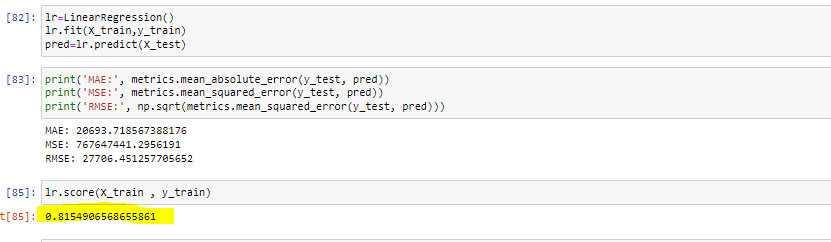
Decision Tree

Ada Boost

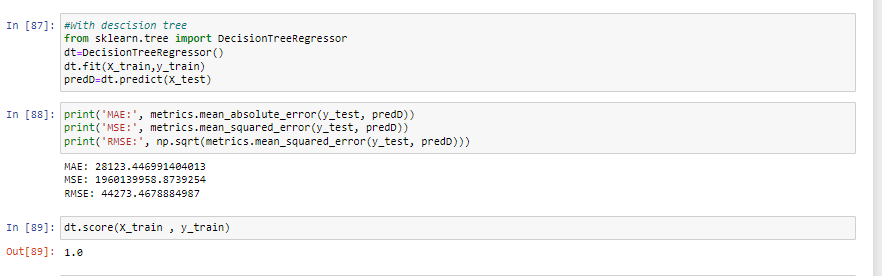
Linear regression

* Run and Evaluate selected models

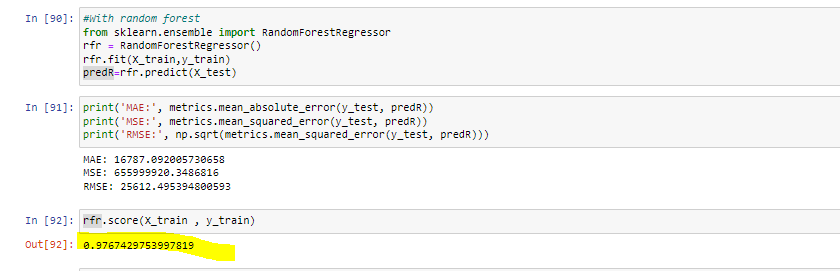
1. Linear regression: 82%



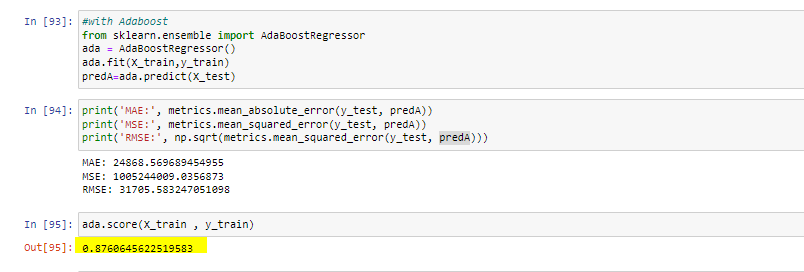
1. Decision Tree



1. Random Forest – 97%

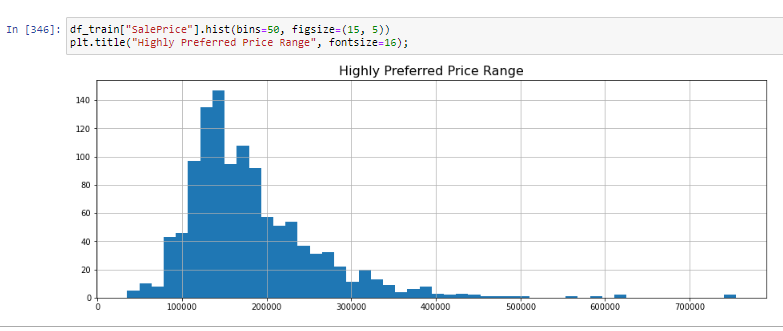


1. Adaboost-87%



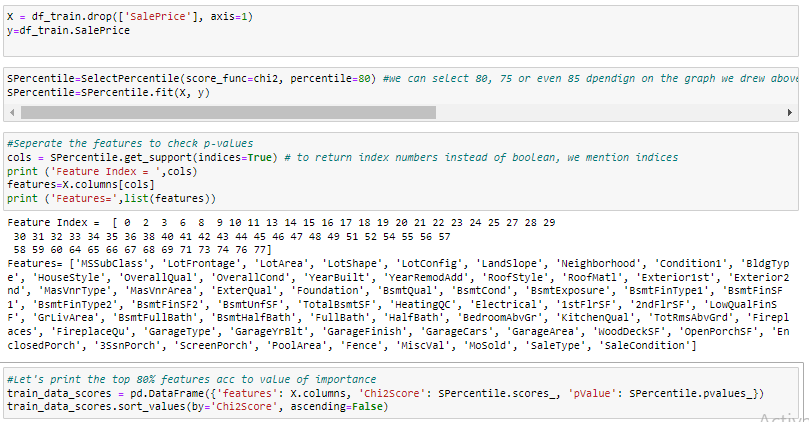
Key Metrics for success in solving problem under consideration

* The df\_train.skew was used to find the features that were skewed
* Some of them were:
* MiscVal 22.996908
* PoolArea 13.203388
* Condition2 11.479618
* LotArea 10.714330
* Heating 10.072518
* ...
* GarageCond -3.559980
* SaleType -3.664995
* Functional -4.005071
* PoolQC -15.855023
* Street -16.970487
* The highly preferred price is between 100000-200000



* According to the corr matrix following were highly correlated:

There is a high correlation between the following features as per correlation matrix:

* BldgType and MSSubClass
* overallQual n SalePrice
* SalePrice and GrLivArea
* GrLivArea n OverallQual
* TotRmsAbvGrd vs GrLivArea
* GarageCars vs GarageArea
* GrLivArea vs SalePrice
* GrLivArea vs totrmsabvgrd
* TotalBsmtSF vs 1stFlrSF
* Fllwong are the 80% of the features selected for modelling via chi score and percentile method
* 

**CONCLUSION**

Key Findings and Conclusions of the Study

Random Forest is the best method. Hyper tuning is not used as models already have a high accuracy

