

HOUSE PRICE PREDICTION

Submitted by:

ANJALI WADHWANI

**ACKNOWLEDGMENT**

<https://matplotlib.org/3.3.3/users/index.html>

<https://seaborn.pydata.org/tutorial.html>

<https://pandas.pydata.org/docs/user_guide/index.html>

KAGGLE

**INTRODUCTION**

* Business Problem Framing

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. 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: • Which variables are important to predict the price of variable? • How do these variables describe the price of the house?

Columns and details

The dataset contains 81 columns

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

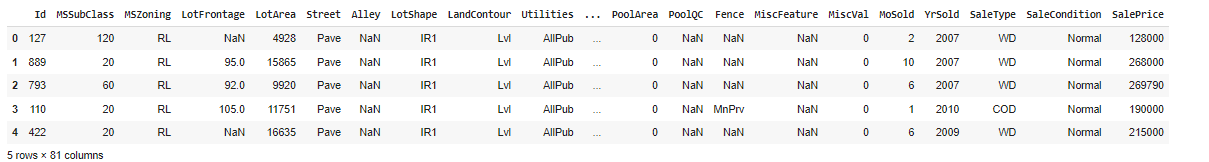
SaleCondition: Condition of sale

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This is a regression model as we are suppose to predict house price on sale .Sale price is in integer value (in dollars) which makes it a regression model.

* Data Sources and their formats

****

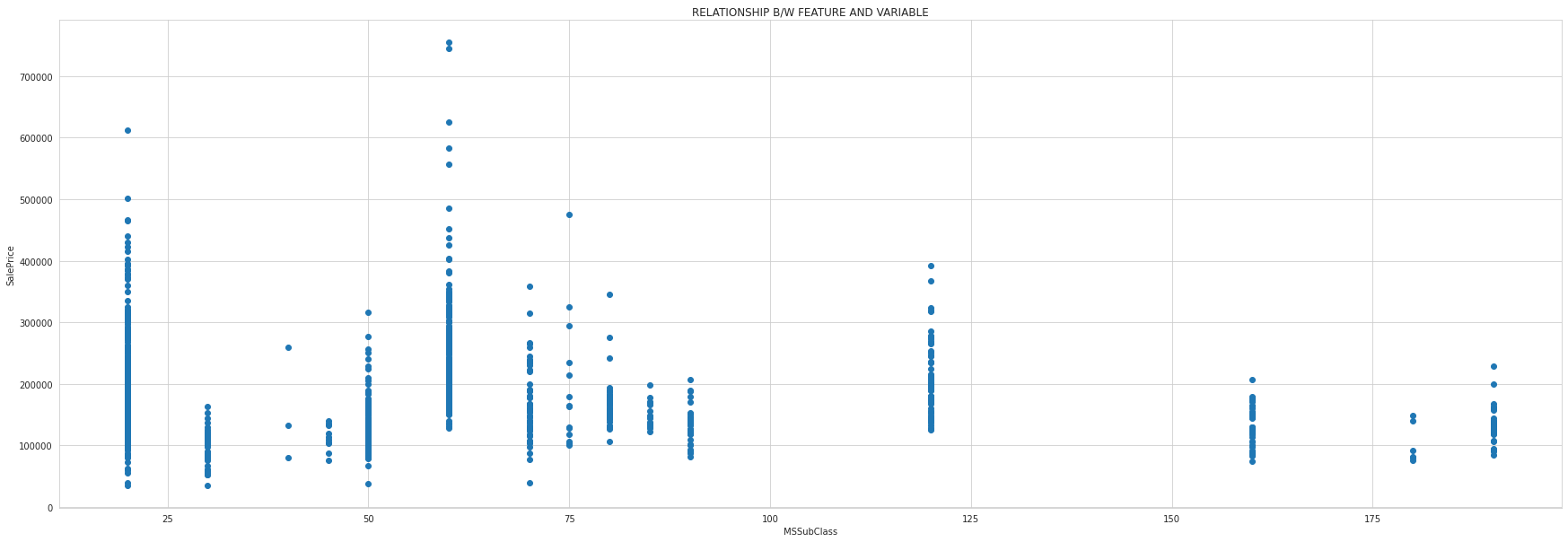
* Data Preprocessing Done

The dataset contains null values.

Filling in the null values with mean (numerical values) and with mode(categorical columns).

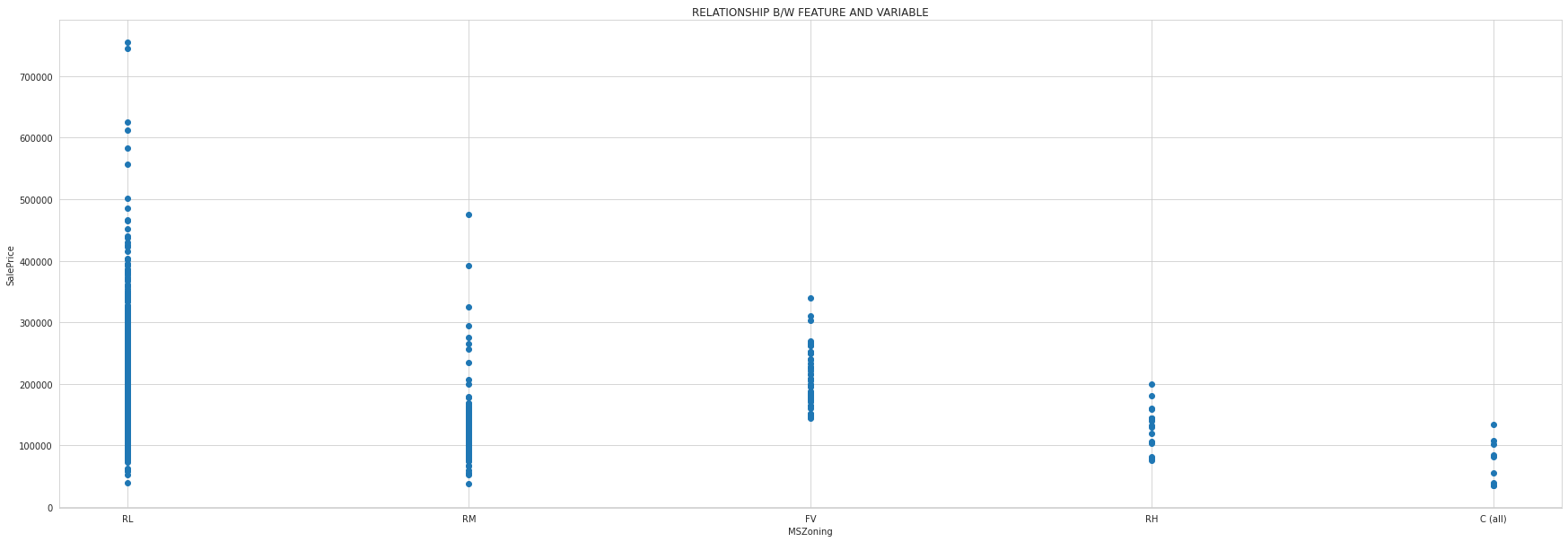
Dropped miscellaneous feature column because it had only 77 filled values rest all were null. Filling it with mode have made the data biased.



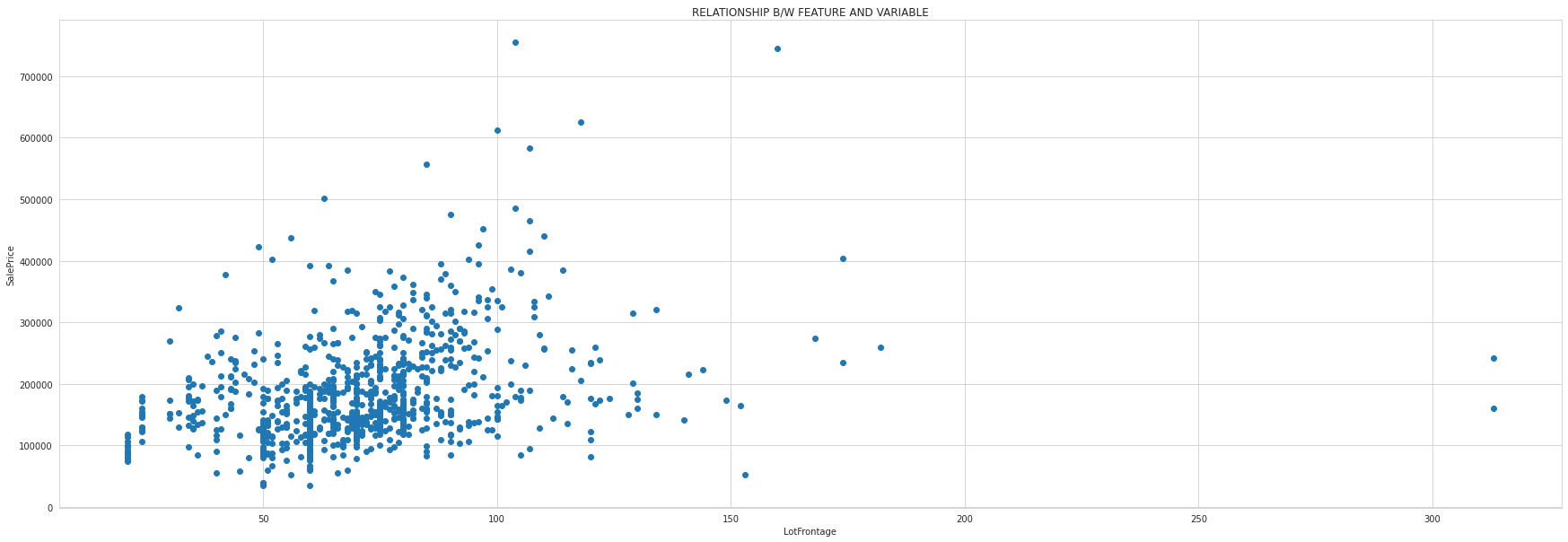
* ****EDA(RELATIONSHIP BETWEEN FEATURES AND VARIABLE(I.E. SALE PRICE) some examples:-

Subclasses below 100 have higher prices.

2.

****

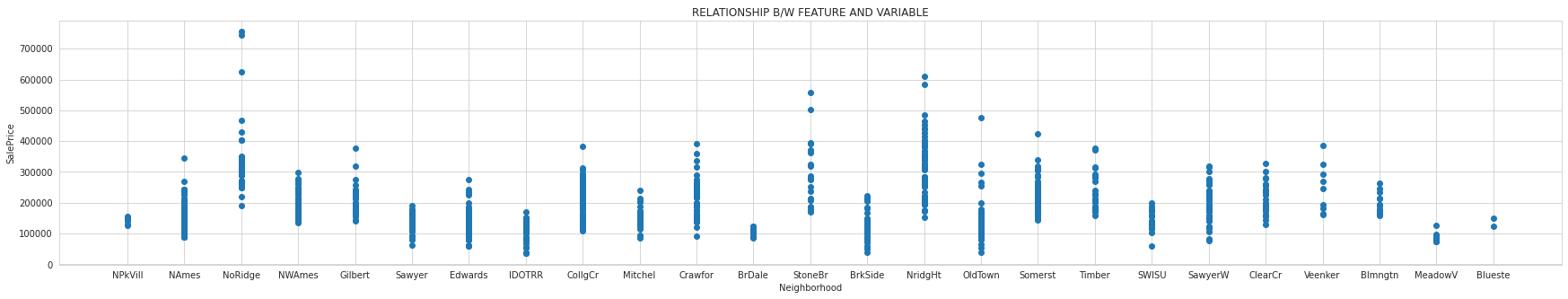
Residential areas with low density have higher prices.

3.

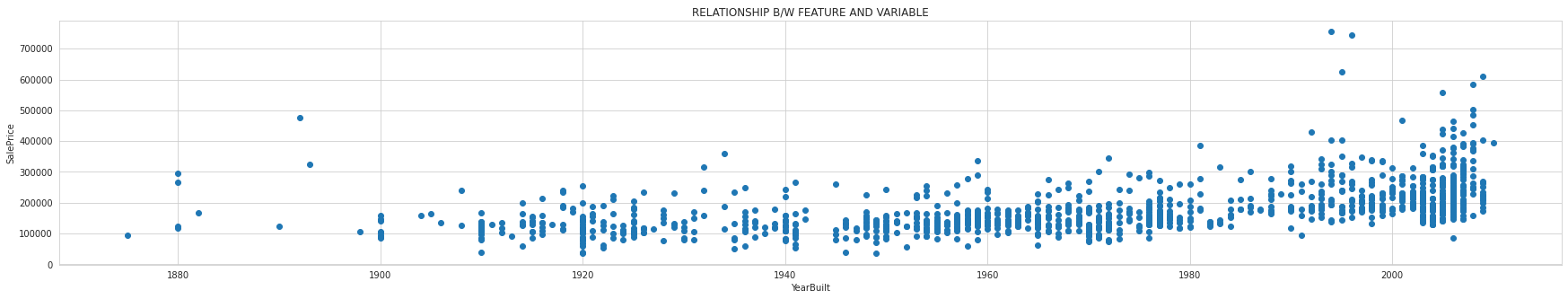
Plots which have lot area more than 100sq.metre have higher sale prices.

4. 

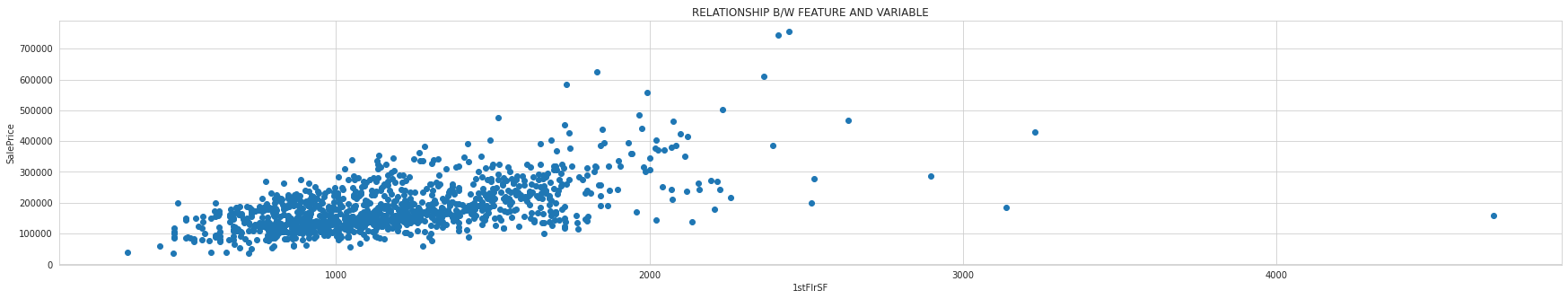
Paved street houses have high sale price.

5.

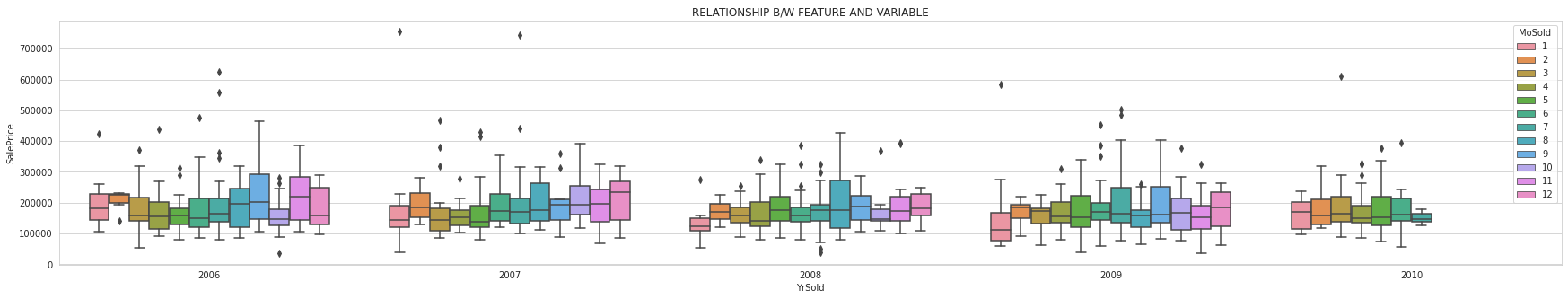
Notridge plot have high values.

6.

The newer construction has high sale value.

7.

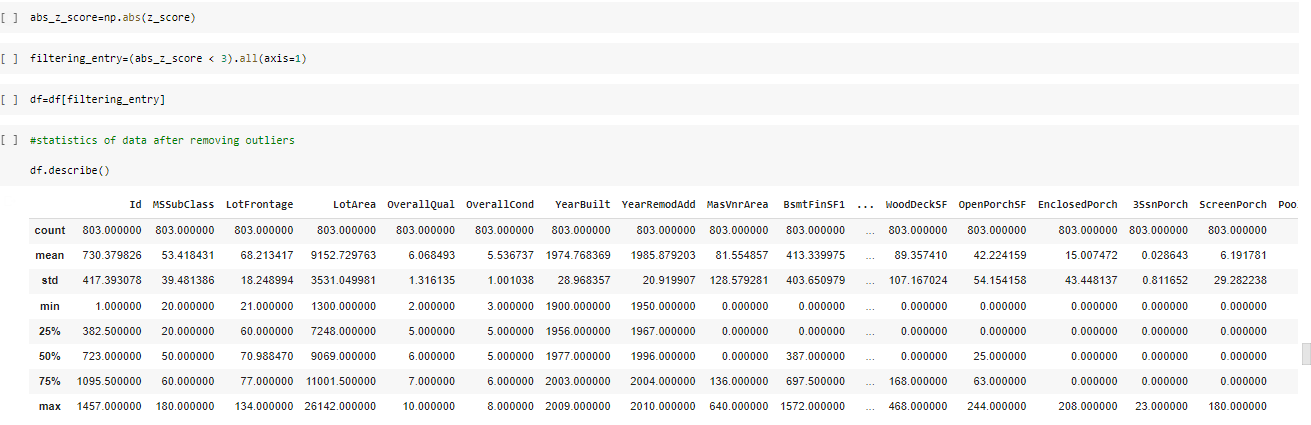
More the surface area higher is the sale price.

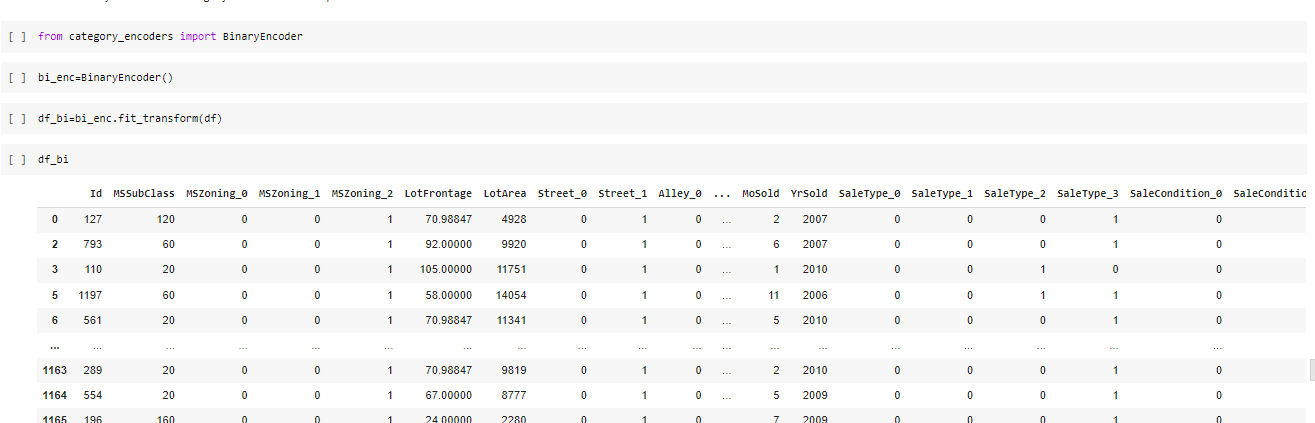
8.

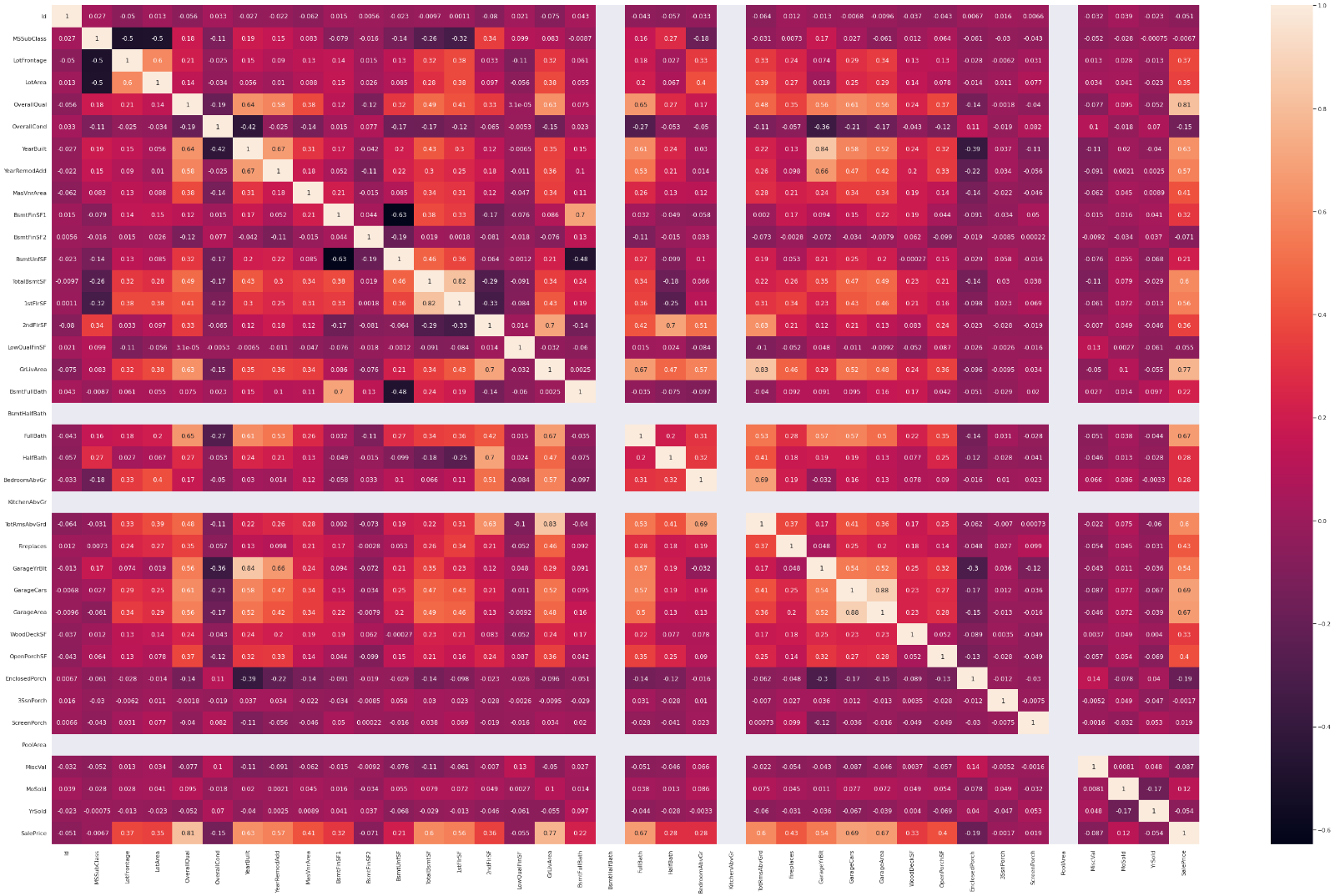
2010 HOUSES ARE LESS ON SALE.

**Model/s Development and Evaluation**

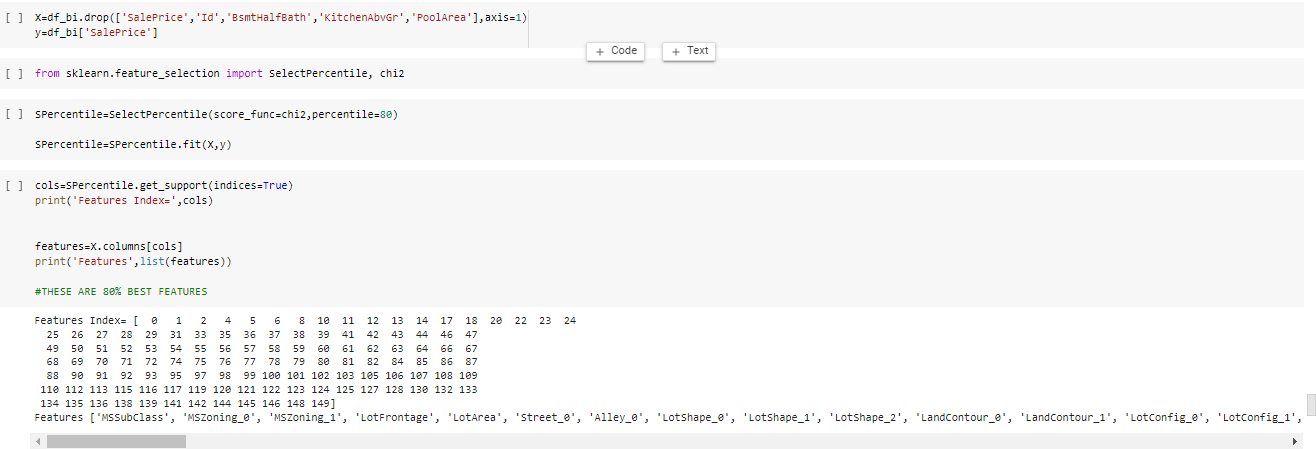
* Identification of possible problem-solving approaches (methods)

With the help of z-score removed outliers.(zscore<3)

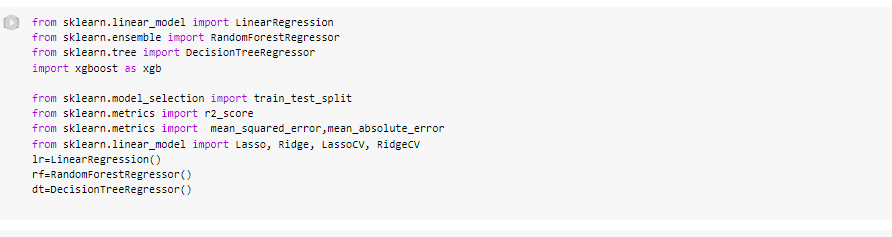
used binary coders to encode data as column had more than 5 categories each.Get\_dummies would have created many columns stressing the model training.

Checked correlation and dropped columns with zero correlation.

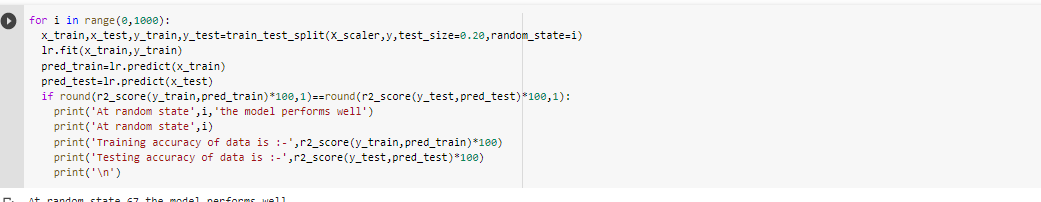
Using select percentile chose best 80% features to train model.

and then standardised the data.

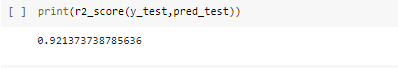
* Testing of Identified Approaches (Algorithms)

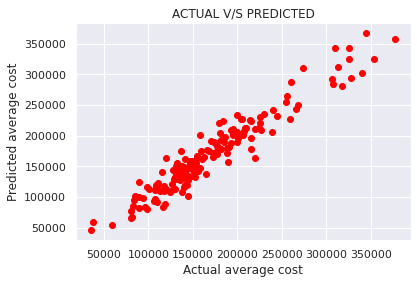


* Run and Evaluate selected models

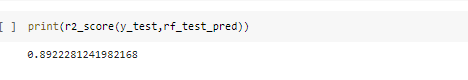


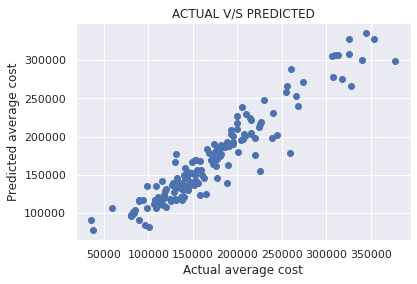
Linear regression





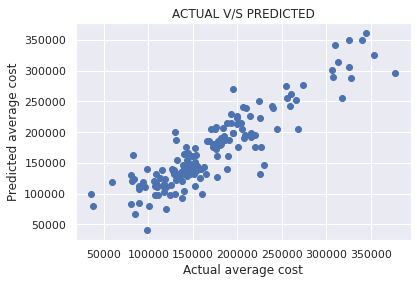
Random forest regressor



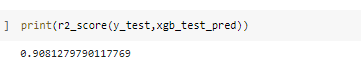


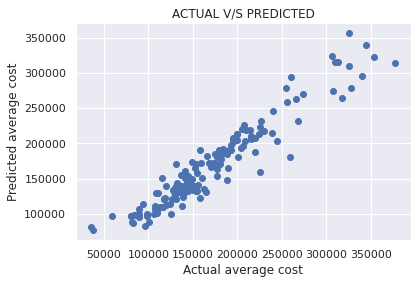
Decision tree regressor





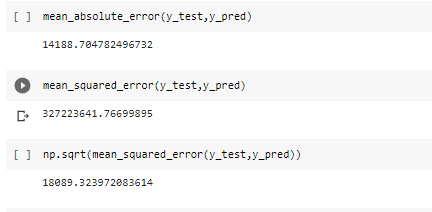
Gradient boosting regressor



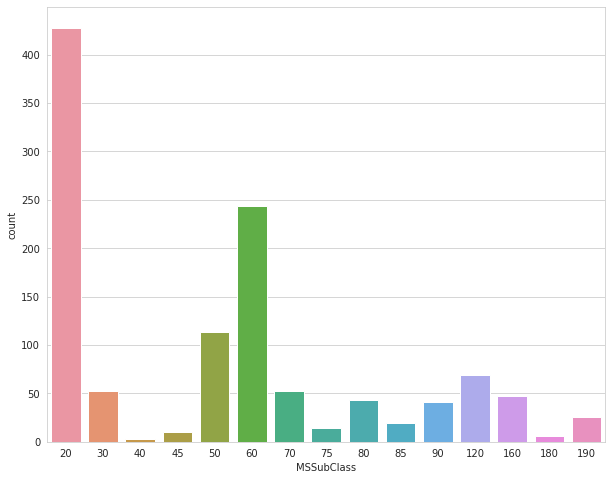


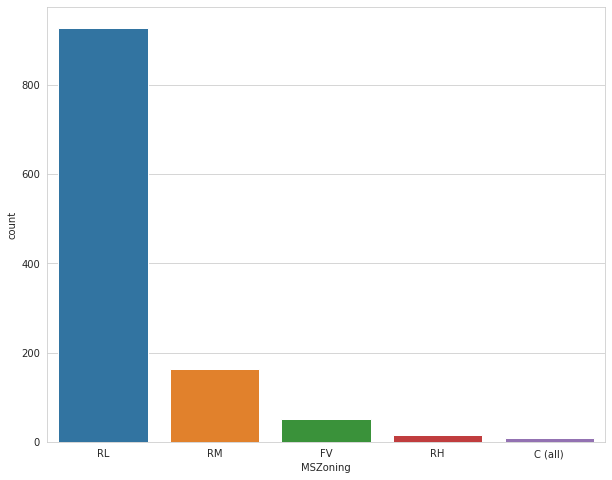
* Key Metrics for success in solving problem under consideration

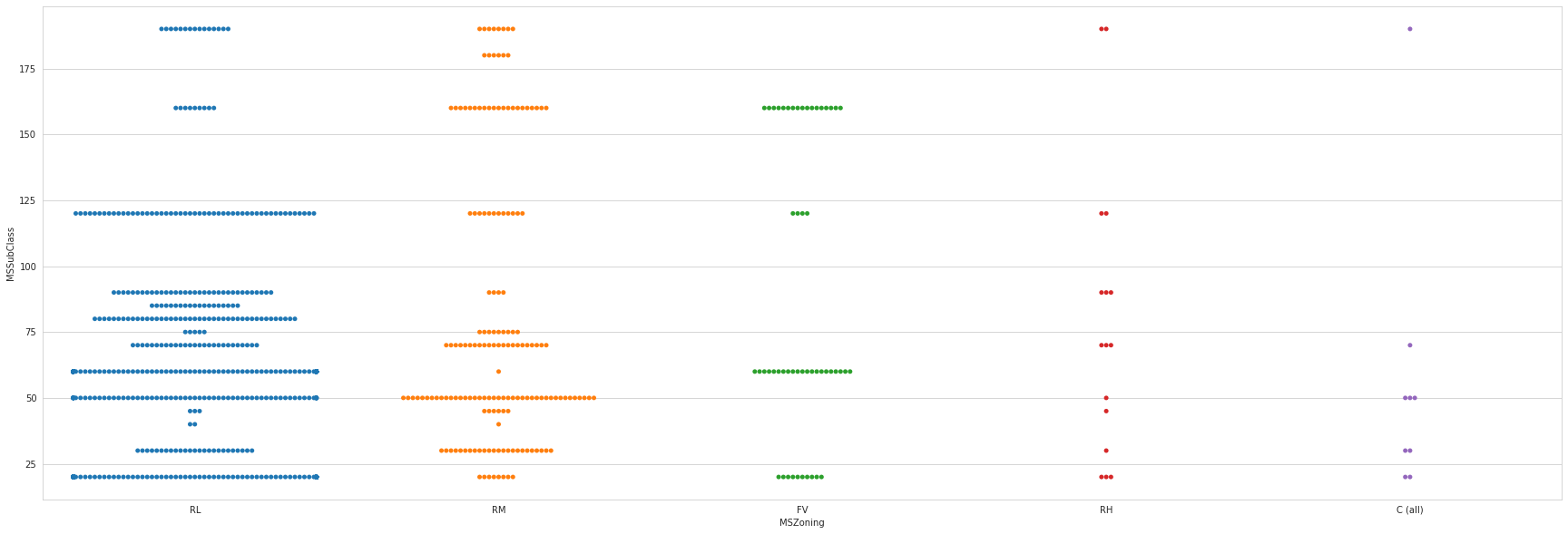
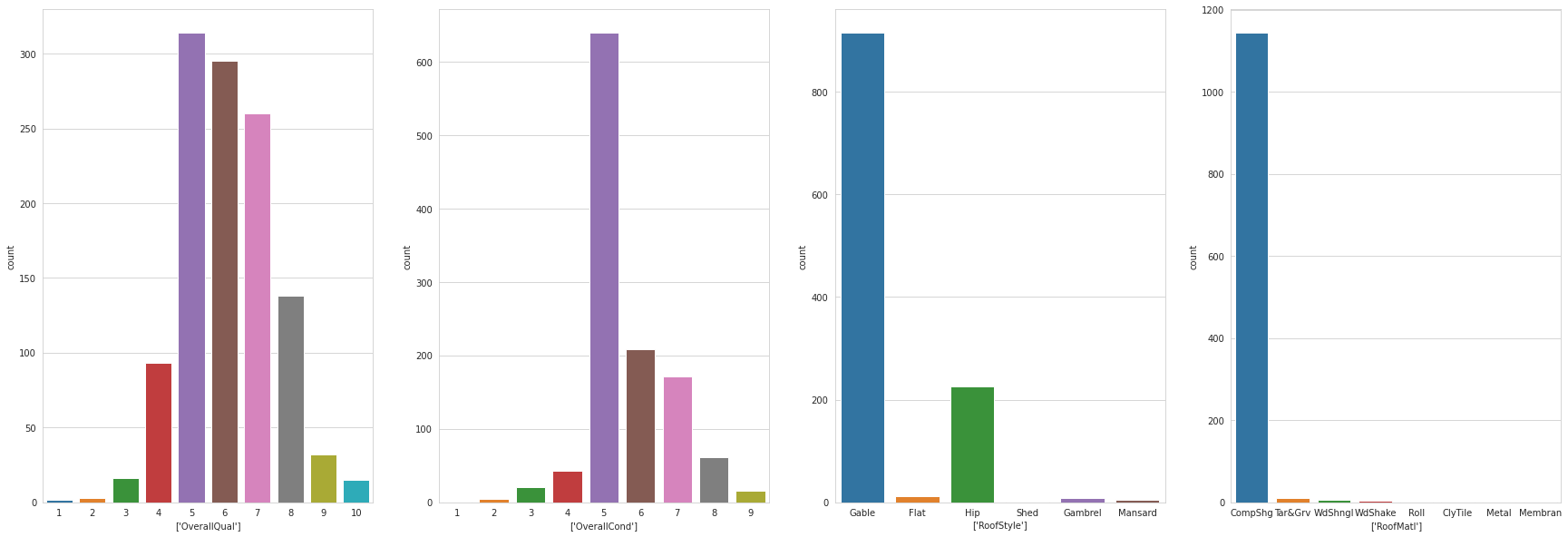
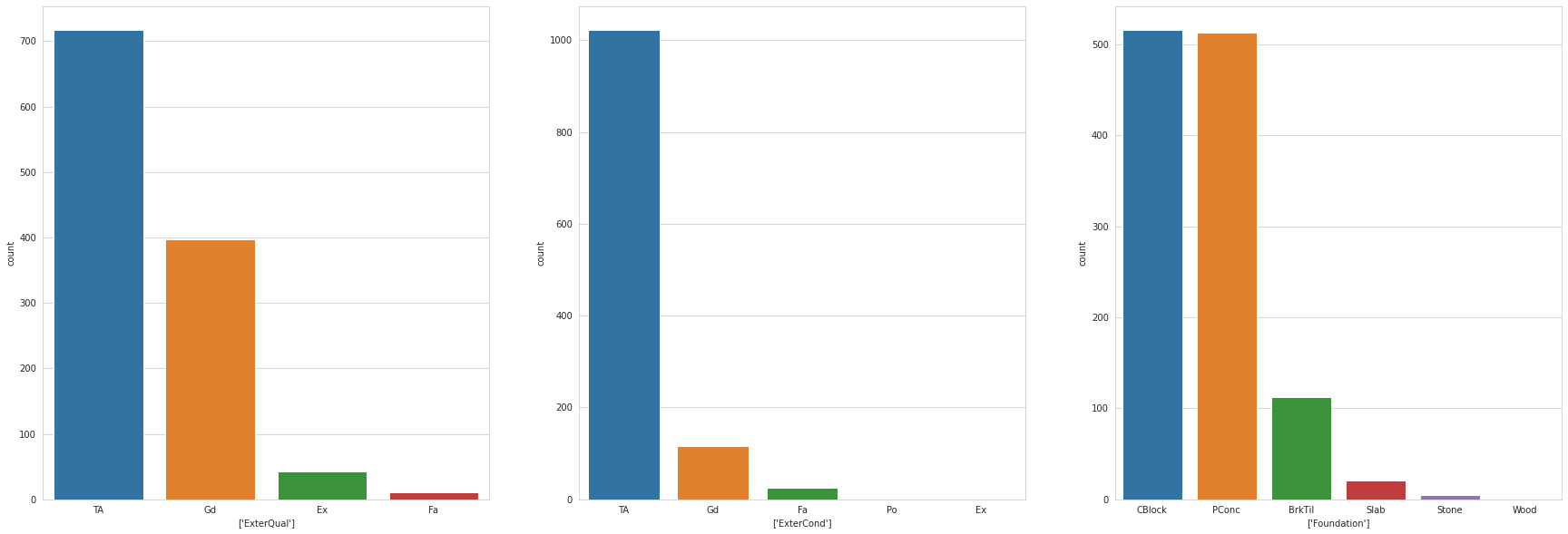
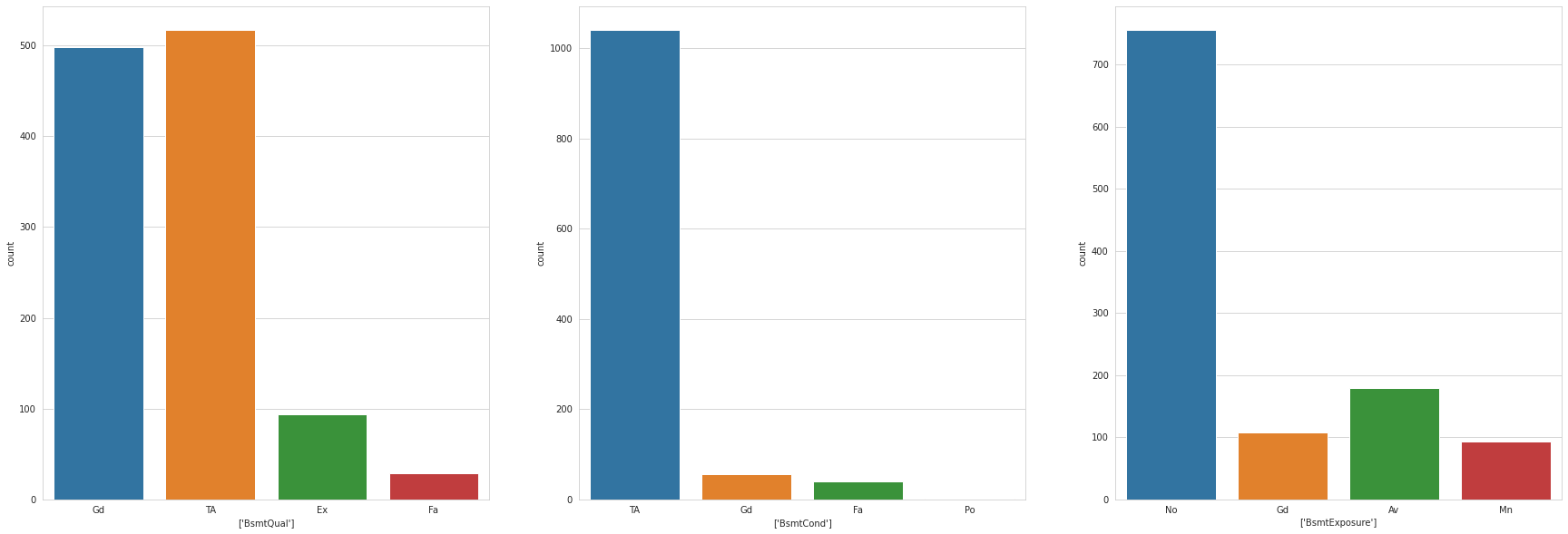
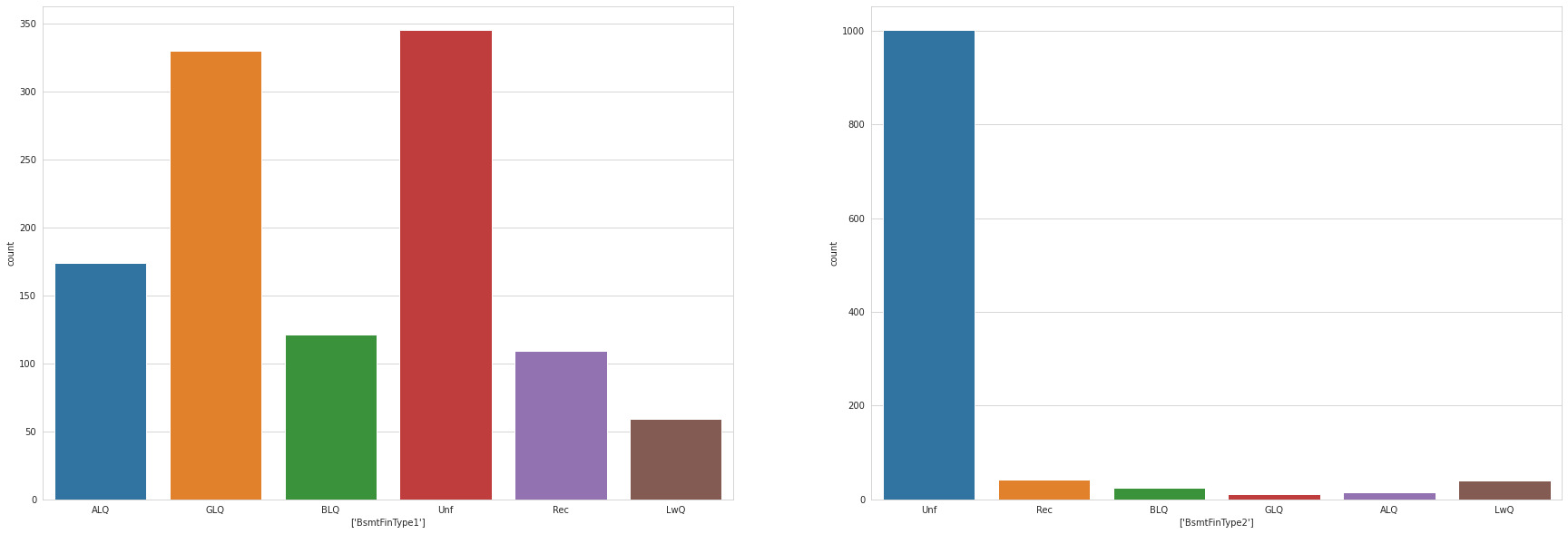
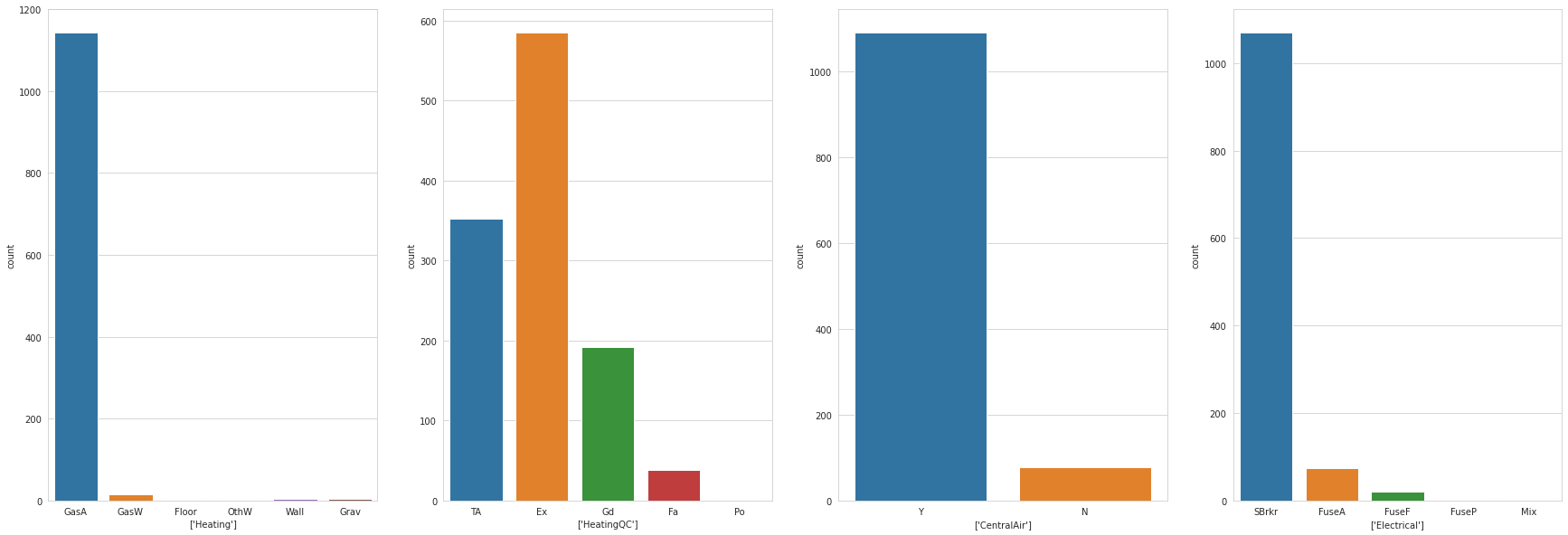
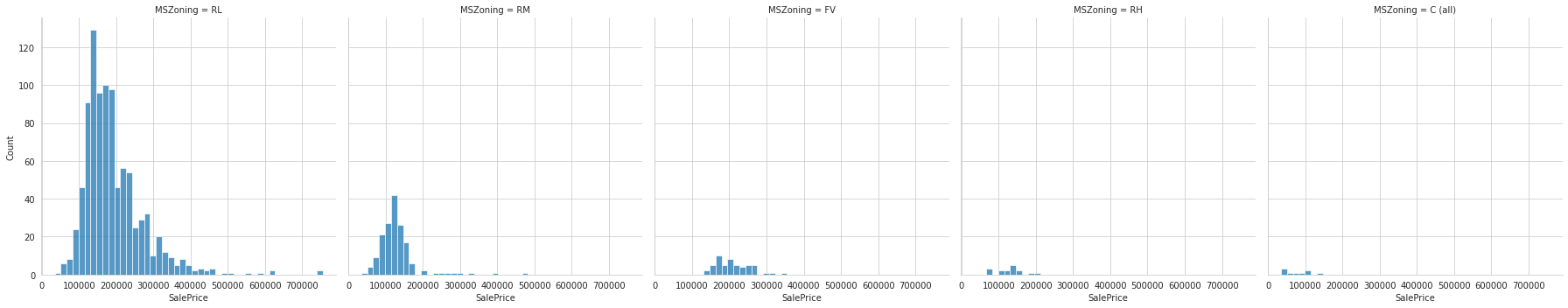
Mae and Mse metrics

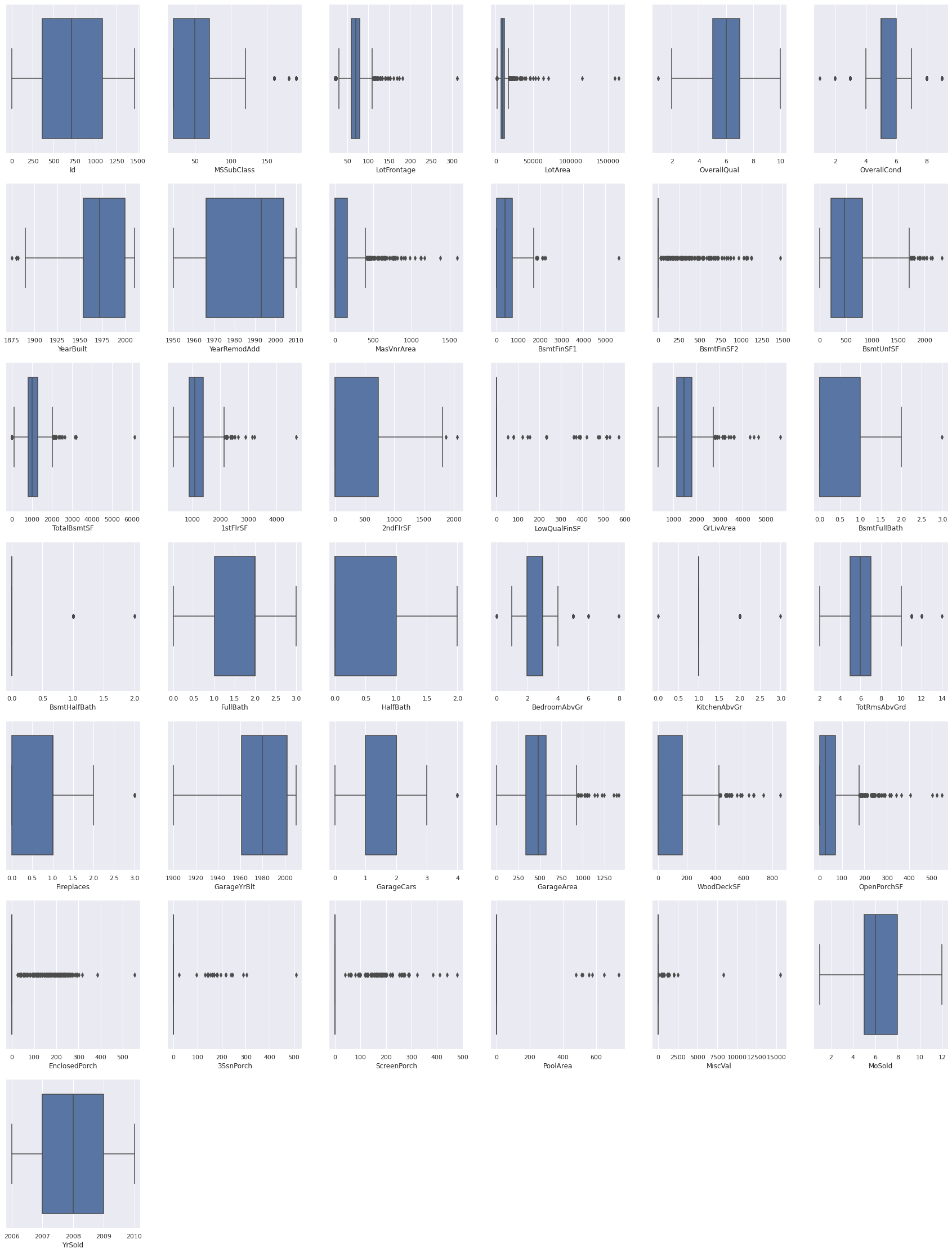


* Visualizations







* Interpretation of the Results

We have choosen linear regression model with 92% accuracy.