

**“A project report**

**On**

**Housing - Price Prediction”**

**Submitted by**

**Himaja Ijjada**

**ACKNOWLEDGMENT**

I express my sincere gratitude to FlipRobo Technologies for giving me the opportunity to work on the Housing Price prediction projects using machine learning algorithms. I would also like to thank FlipRobo Technologies for providing me with the requisite datasets to work with. And I would like to express my gratitude to Mr. Mohd Kashif (SME FlipRobo) and Ms. Sapna Verma (SME FlipRobo) for being of a great help in completion of the project.

I acknowledge my indebtedness to the authors of papers titled: “House Price Prediction using a Machine Learning Model: A Survey of Literature” and “The impact of housing quality on house prices in eight capital cities, Australia” for providing me with valuable insights and knowledge of the dynamic relationships that exist in the economics of real estate and housing markets.

**CONTENTS**

* **Introduction**
* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken
* **Analytical problem framing**
* Mathematical/ Analytical Modeling of the Problem
* Data Sources and their formats
* Data Preprocessing Done
* Data Inputs- Logic- Output Relationships
* Assumptions
* Hardware and Software Requirements and Tools Used
* **Model/s Development and evaluation**
* Visualizations
* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration
* Interpretation of the Results
* **Conclusion**
* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**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 major contributors in the world’s economy. It has a huge chunk of market and there are various companies working in this domain. Data science becomes a very important tool to solve problems in this field 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 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?

**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 is crucial in understanding their mutually inclusive relationships with their pricing.

**Review of Literature**

Two research papers, namely: “House Price Prediction using a Machine Learning Model: A Survey of Literature” and “The impact of housing quality on house prices in eight capital cities, Australia” were reviewed and evaluated to gain insights into all the attributes that influence the price of house.

From studying the papers and analyzing the research work it, is learnt that locational attributes and structural attributes are prominent factors in predicting house prices. Studies suggest that there exists a close relationship between House pricing and locational attributes such as distance from the closest shopping center, train station, position offering views of hills or shore, the neighborhood in which the property is situated etc.

Structural attributes of the house like lot size, lot shape, quality and condition of the house, garage capacity, rooms, Lot frontage, number of bedrooms, bathrooms, overall finishing of the house etc. play a big role in influencing the house price. Neighborhood qualities can be included in deciding house price. Factors like efficiency of public education, community social status, and the socio-cultural demographics improve the worth of a property.

The demand side of the housing market is also a necessary component. Although population growth is widely known as a driver in housing demand, the key issue lies in the proportion of people with abundant financial resources.

Variables representing land value such as rents and material costs also demonstrate their influence in explaining house prices, which are positively related to housing prices. Multiple regression analysis models allow to ascertain price predictions by capturing independent and dependent variable data. In using multiple regression modelling techniques, we can describe changes brought to a dependent variable with changes in the independent variables.

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 that best fits their budget and will be of tremendous benefit, especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices and to acknowledge the best machine learning model to be used to conduct a study in this field.

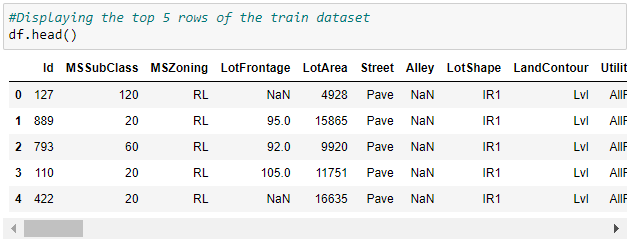
**Analytical problem framing**

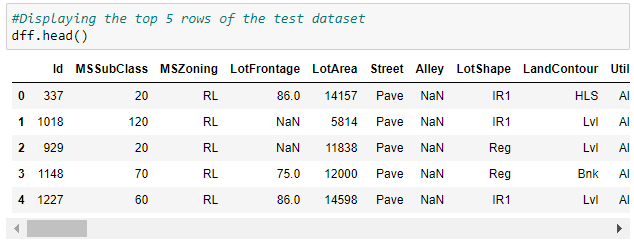
**Mathematical/ Analytical Modeling of the Problem**

Various Regression analysis techniques were used to build predictive models to understand the relationships that exist between Housing sales prices and various Housing property attributes. The Regression analysis models were used to predict the Sale price value for changes in Housing property attributes. Regression modelling techniques were used in this Problem since Sales Price data distribution is continuous in nature. In order to forecast house price, predictive models such as ridge regression Model, Random Forest Regression model, Decision tree Regression Model, Support Vector Machine Regression model, Extreme Gradient Boost Regression were used to describe how the values of Sale Price depended on the independent variables of various Housing property attributes.

**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 2 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).





Training Dataset contains 1168 entries and 81 variables, while Test Dataset contains 292 entries and 80 variables.

**Dataset Description**

The Independent Feature columns are as follows

* 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) livin g 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

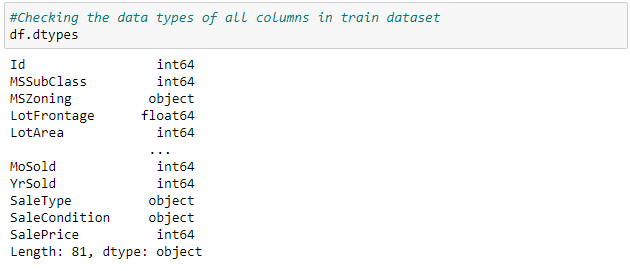
**Target Column:**

* SalePrice

**Data Preprocessing Done**

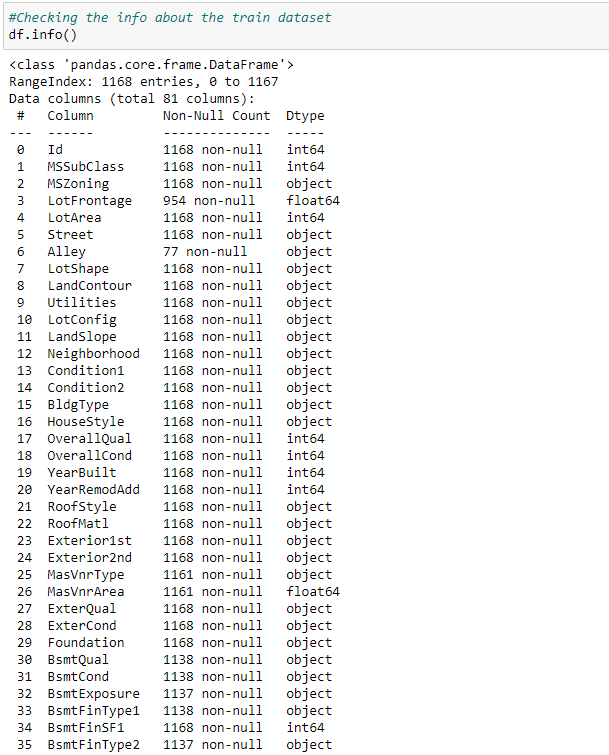
1. **Train Dataset**

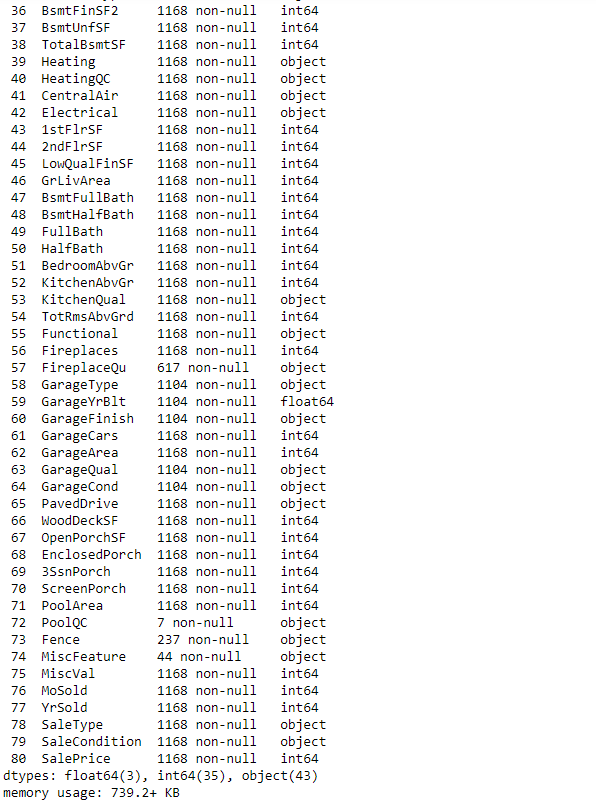
Checking the data types of all columns in train dataset



# Observations - In the train dataset we have int, object as well as flaot data types.

Checking the info about the train dataset



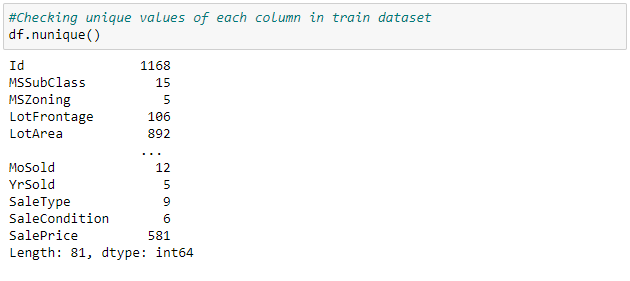


**Observations**

Above is the info about train dataset from which we can observe that

* There are some missing values in the dataset, which needs to be filled using imputation techniques.
* And in Alley, PoolQC, Fence and MiscFeature has more than 80% null values so we may drop these columns in the further steps

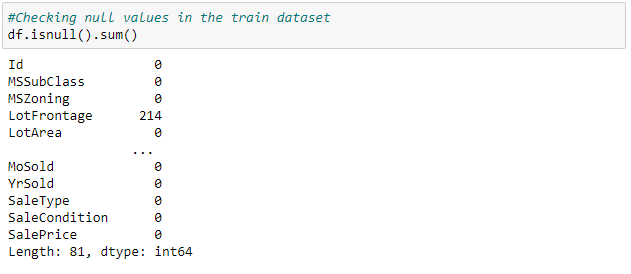
Checking unique values of each column in train dataset



# Observations

* In Id column the unique count is 1168 which means all the values in the column are unique and ID is the identity number given for perticular asset so this ID has no purpose in developing a model and training. Hence it can be dropped.
* In Utilities column unique value count is 1 which means all the entries are same this also has no purpose in model building so this column can also be dropped.

Checking null values in the train dataset



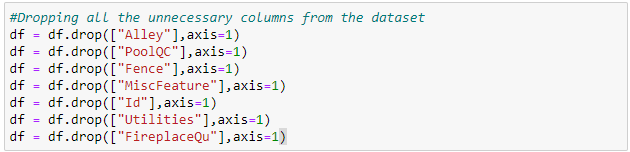
Visualizing null values in train dataset



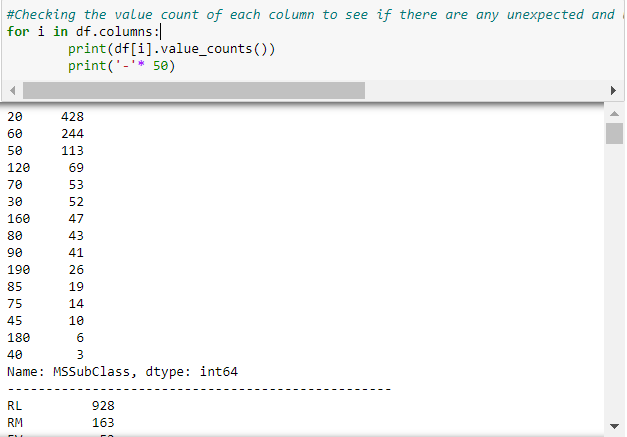
# Observations

* Those columns with more than 50% (584 entries) null values can be dropped to avoid unncessary issue for model development.
* Those columns with less than 50% (584 entries) null values can be treated using imputation techniques.
* The column FireplaceQu has 551 null values, which accounts to 50% approximately can also be dropped

Dropping all the unnecessary columns from the dataset



Checking the value count of each column to see if there are any unexpected and unwanted entries present in the column in train dataset.

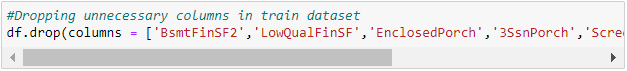


# Observations

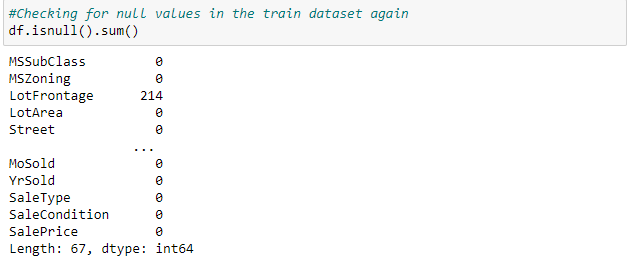
* There are no unnecessary or duplicate entries in any column of the train dataset.
* There are zero values as entries in some columns which are below 60%, hence they are acceptable and reasonable.
* But there are 85% zero values as entries in the following columns
  + BsmtFinSF2
  + LowQualFinSF
  + EnclosedPorch
  + 3SsnPorch
  + ScreenPorch
  + PoolArea
  + MiscVal

So lets drop these columns.

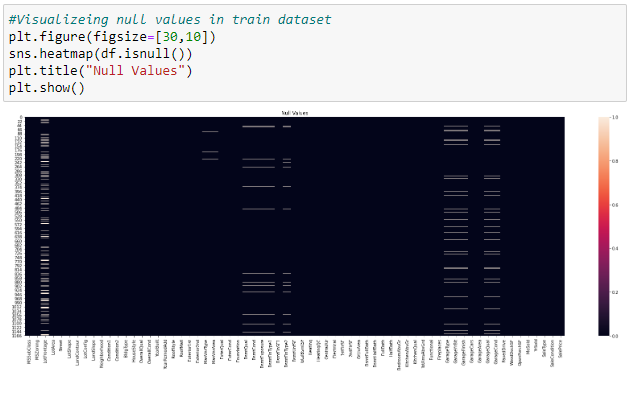
Dropping unnecessary columns in train dataset



Checking for null values in the train dataset again



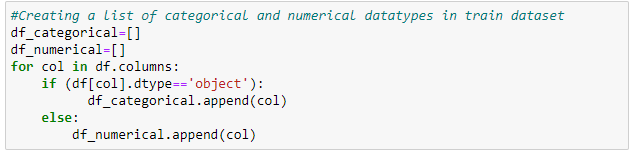
Visualizing null values in train dataset



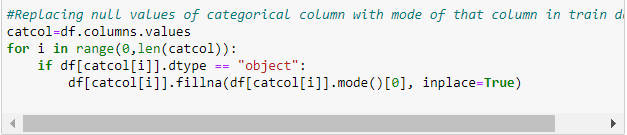
# Observations

We can observe that there are null values in most of the columns of the train dataset. We need to treat them using appropriate imputation techniquesand fill them with their respective values.

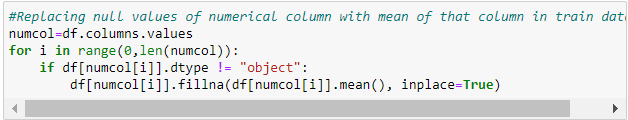
Creating a list of categorical and numerical datatypes in train dataset



Replacing null values of categorical column with mode of that column in train dataset

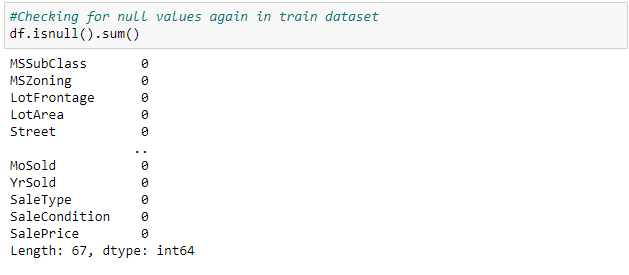


Replacing null values of numerical column with mean of that column in train dataset.



Replaced all the null values in Numerical columns and categorical coumns of the train dataset.

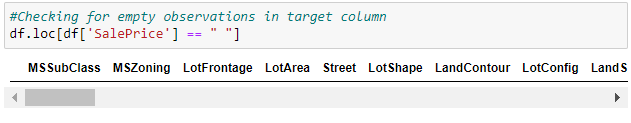
Checking for null values again in train dataset



Visualizing null values again after imputation in train dataset

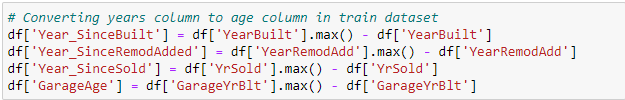


Checking for empty observations in target column

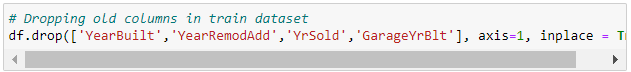


There are no empty observations in the target column of train dataset.

Converting years column to age column in train dataset



Dropping old columns in train dataset

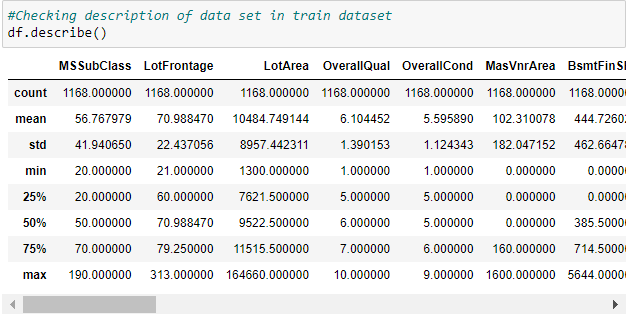


We have converted all the year columns to their respective age, as age helps us more than year Built in the dataset.

**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 analyzed 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 Sale Price. Whereas Age of House Remodeling age Garage age have a linear negative relationship and therefore increase in their values leads to a decrease in Sale Price.

Checking description of data set in train dataset



**Assumptions /Observations**

- Big difference between max value and 75% in SalePrice, MSSubClass, LotFrontage, LotArea, BsmtFinSF1, BsmtF inSF2, etc. indicates presence of outliers.

- A higher std than mean in columns:

- MasVnrArea

- BsmtFinSF1

- BsmtFinSF2

- WoodDeckSF

- OpenPorchSF, EnclosedPorch, 3SsnPorch etc. indicates presence of skewness.

- An Anomaly is displayed in the relationship between age of house and SalePrice. There is a general negative relationship between House age and Sale Price, ie. Increase in age leads to a decrease in SalePrice. However, houses built between 1880 and 1900 sold for the highest. The assumption made in this regard is that those houses were sold for the highest amount because of their antiquity value.

**Hardware and Software Requirements and Tools Used**

**Hardware Used:**

* Processor AMD Ryzen 9 5900HX(8 Cores 16 Logical Processors)
* Physical Memory: 16.0GB (3200MHz)
* GPU: Nvidia RTX 3060 (192 bits), 6GB DDR6 VRAM, 3840 CUDA cores.

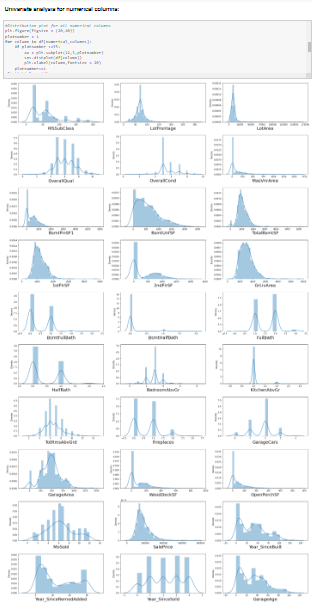
**Software Used:**

* Windows 10 Operating System
* Anaconda Package and Environment Manager: Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data- science packages suitable for Windows and provides a host of tools and environment for conducting Data Analytical and Scientific works. Anaconda provides all the necessary Python packages and libraries for Machine learning projects.
* Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
* Python3: It is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best languages used for Data Analytics and Data science projects/application. Python provides numerous libraries to deal with mathematics, statistics and scientific function.
* Python Libraries used:
* **Pandas:** For carrying out Data Analysis, Data Manipulation, and Data Cleaning etc.
* **Numpy:** For performing a variety of operations on the datasets.
* **matplotlib.pyplot, Seaborn**: For visualizing Data and various relationships between Feature and Label Columns
* **Scipy:** For performing operations on the datasets
* **Statsmodels:** For performing statistical analysis
* **sklearn** for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

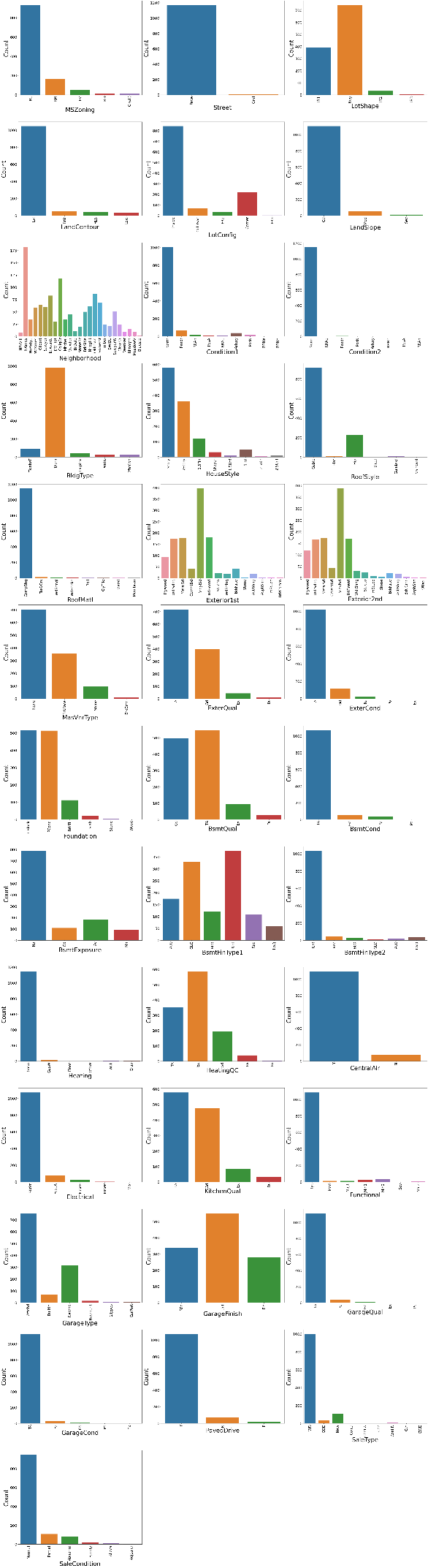
**Model/s Development and evaluation**

**Visualization**

## Univariate Analysis:

****

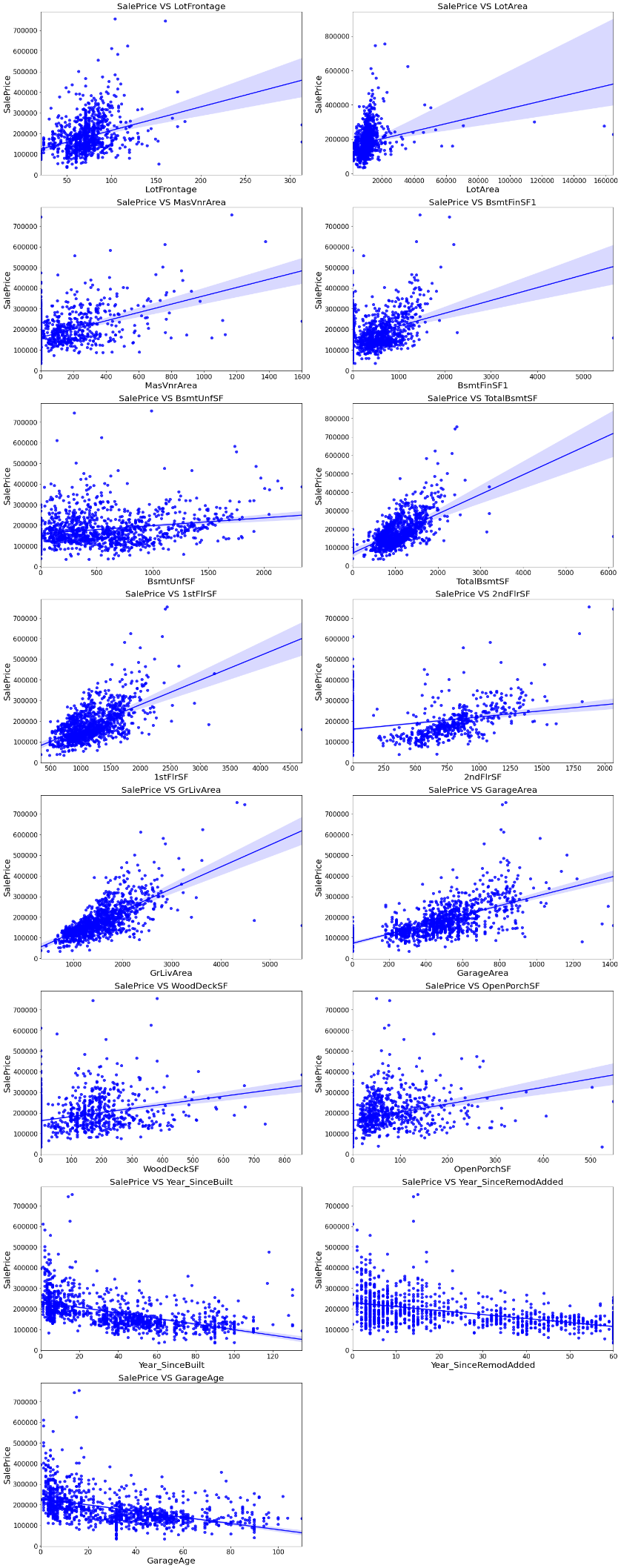
### Univariate analysis for categorical columns:

****

# Observations:

* It is found that Residential Low Density zoning has maximum count, for the feature general zoning classification of the sale(MSZoning).
* In Paved streets we can observe maximum count, for the feature Type of road access to property(Street).
* Regular shaped property has maximum count, for the feature General shape of property(LotShape).
* Near Flat/Level property has maximum count, for the feature Flatness of the property(LandContour).
* Inside lot configured property has maximum count, for the feature Lot configuration(LotConfig).
* Gentle sloped property has maximum count, for the feature Slope of property(LandSlope).
* If the property is located in North Ames then count is good compared to other locations, for the feature Physical locations within Ames city limits(Neighborhood).
* If the Proximity to various conditions-1 is normal then count is high for the feature Proximity to various conditions(Condition1).
* If the Proximity to various conditions-2 is normal then count is high for the feature Proximity to various conditions (if more than one is present)(Condition2).
* Single-family Detached dwelling has maximum count for the feature Type of dwelling(BldgType).
* One story dwelling housestyle has maximum count for the feature Style of dwelling(HouseStyle).
* For Gable roof style the count is high for the feature Type of roof(RoofStyle).
* For Standard (Composite) Shingle roof material the count is high for the feature Roof material(RoofMatl).
* For Vinyl Siding exterior-1 covering on house has maximum counts for the feature Exterior covering on house(Exterior1st).
* For Vinyl Siding exterior-2 covering on house has maximum counts for the feature Exterior covering on house (if more than one material)(Exterior2nd).
* For Masonry veneer type(MasVnrType) None has maximum count.
* For Typical/Average(TA) quality of the material on the exterior has maximum count, for the feature Evaluates the quality of the material on the exterior (ExterQual).
* For Typical/Average(TA) condition of the material on the exterior has maximum count for the feature Evaluates the present condition of the material on the exterior(ExterCond).
* For Cinder Block and Poured Contrete foundations the count is maximum for the feature Type of foundation(Foundation).
* For unfinished Rating of basement finished area-1 the count is maximum for the feature Rating of basement finished area(BsmtFinType1).
* For unfinished Rating of basement finished area-2 the count is maximum for the feature Rating of basement finished area (if multiple types)(BsmtFinType2).
* For Gas forced warm air furnace type of heating the count is maximum for the feature Type of heating(Heating).
* For Excellent Heating quality and condition the count is high for the feature Heating quality and condition(HeatingQC).
* For Central air conditioning-yes has maximum count for the feature Central air conditioning(CentralAir).
* For Standard Circuit Breakers & Romex Electrical system the count is high for the feature Electrical system(Electrical).
* For Typical/Average(TA) and good Kitchen quality the count is maximum for the feature Kitchen quality(KitchenQual).
* Typical Functionality has highest count for Home functionality (Assume typical unless deductions are warranted)(Functional).
* For good Fireplace quality the count is high for the feature Fireplace quality(FireplaceQu).
* If Garage location Attached to home then the count is high, for the feature Garage location(GarageType).
* For Unfinished Interior of the garage the count is maximum, for the feature Interior finish of the garage(GarageFinish).
* For Typical/Average(TA) Garage quality the count is high, for the feature Garage quality(GarageQual).
* For Typical/Average(TA) Garage condition the count is high, for the feature Garage condition(GarageCond).
* For Paved driveway the count is maximum, for the feature Paved driveway(PavedDrive).
* For Warranty Deed - Conventional type of sales the count is maximum, for the feature Type of sale(SaleType).
* For Normal sales condition the count is high, for the feature Condition of sale(SaleCondition).

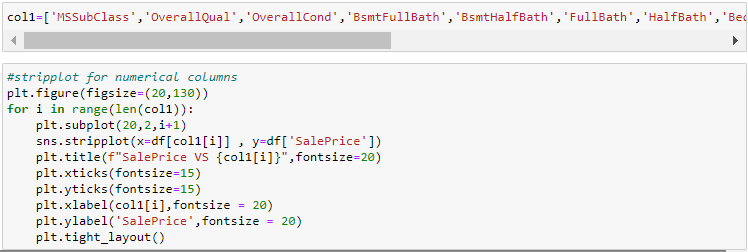
### Bivariate Analysis for numerical columns:

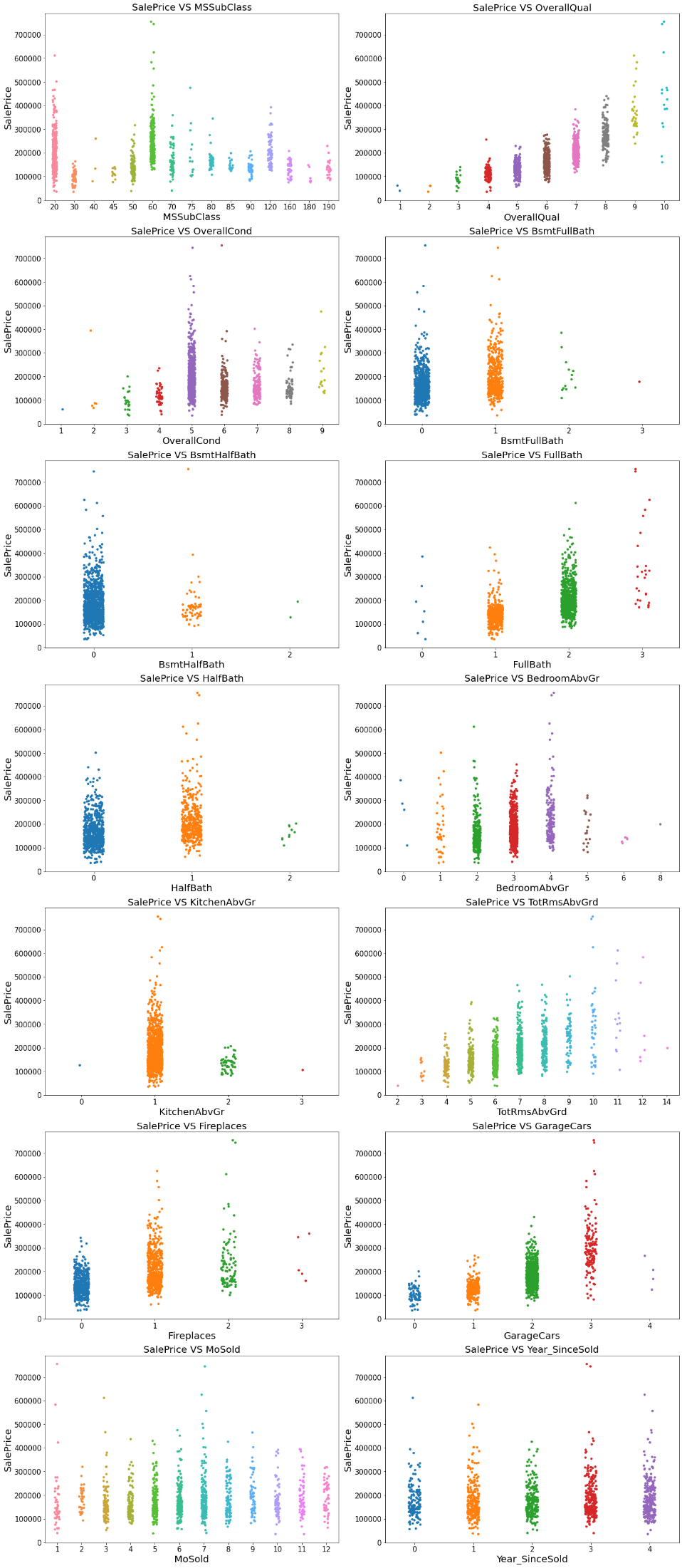
****

# Observations:

* As Linear feet of street connected to property(LotFrontage) is increseing sales is decreasing and the SalePrice is rangeing between 0-3 lakhs.
* As Lot size in square feet(LotArea) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Masonry veneer area in square feet(MasVnrArea) is increasing sales is decreasing and saleprice is rangeing between 0-4 lakhs.
* As Type 1 finished square feet(BsmtFinSF1) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Unfinished square feet of basement area(BsmtUnfSF) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs. There are some outliers also.
* As Total square feet of basement area(TotalBsmtSF) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As First Floor square feet(1stFlrSF) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Second floor square feet(2ndFlrSF) is increseing sales is increasing in the range 500-1000 and the saleprice is in between 0-4 lakhs.
* As Above grade (ground) living area square feet(GrLivArea) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Size of garage in square feet(GarageArea) is increseing sales is increseing and the saleprice is in between 0-4 lakhs.
* As Wood deck area in square feet(WoodDeckSF) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Open porch area in square feet(OpenPorchSF) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.
* As Year\_SinceBuilt is increseing sales is decreasing and the saleprice is high for newly built building and the sales price is in between 0-4 lakhs.
* As Since Remodel date (same as construction date if no remodeling or additions)(Year\_SinceRemodAdded) is increseing sales is decreasing and the saleprice is in between 1-4 lakhs.
* As Since Year garage was built(GarageAge) is increseing sales is decreasing and the saleprice is in between 0-4 lakhs.

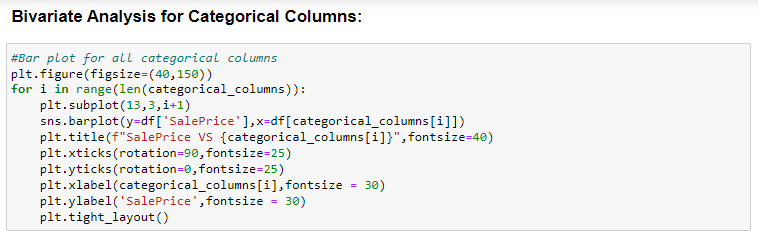
Plotting strip plot for the numerical columns

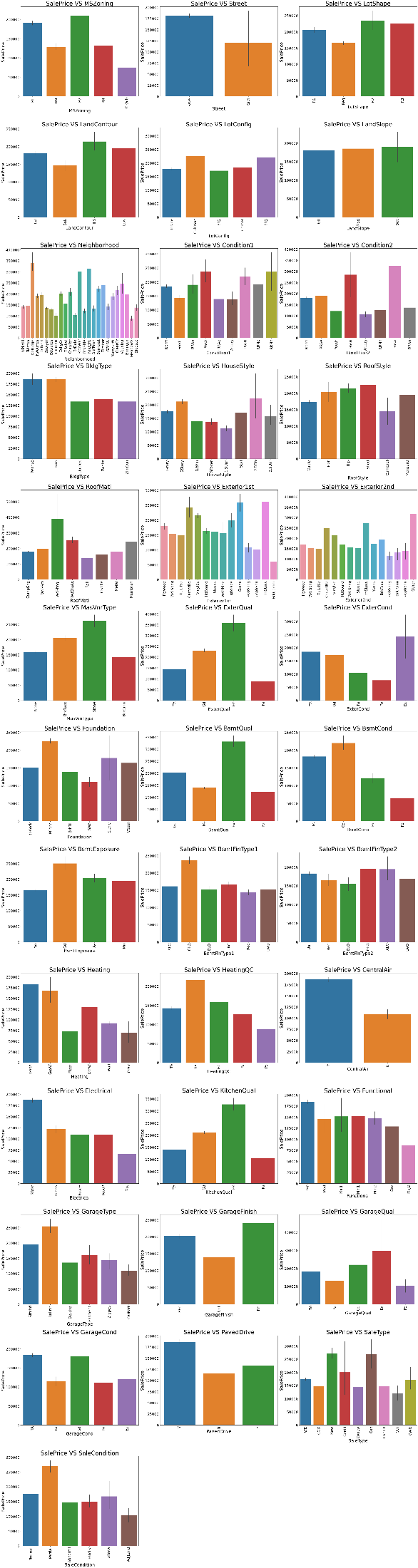
****

****

# Observations:

* For 1-STORY 1946 & NEWER ALL STYLES (20) and 2-STORY 1946 & NEWER (60) types of dwelling (MSSuubClass) the sales is good and SalePrice is also high.
* As Rates the overall material and finish of the house (OverallQual) is increasing linearly sales is also increasing And SalePrice is also increasing linearly.
* For 5(Average) overall condition of the house(OverallCond) the sales is high and SalePrice is also high.
* For 0 and 1 Basement full bathrooms(BsmtFullBath) the sales as well as SalePrice is high.
* For 0 Basement half bathrooms(BsmtHalfBath) the sales as well as SalePrice is high.
* For 1 and 2 Full bathrooms above grade(FullBath) the sales as well as SalePrice is high.
* For 0 and 1 Half baths above grade(HalfBath) the sales as well as SalePrice is high.
* For 2, 3 and 4 Bedrooms above grade (does NOT include basement bedrooms)(BedroomAbvGr) the sales as well as SalePrice is high.
* For 1 Kitchens above grade(KitchenAbvGr) the sales as well as SalePrice is high.
* For 4-9 Total rooms above grade (does not include bathrooms)(TotRmsAbvGrd) the sales as well as SalePrice is high.
* For 0 and 1 Number of fireplaces(Fireplaces) the sales as well as SalePrice is high.
* For 1 and 2 Size of garage in car capacity(GarageCars) the sales is high and for 3 Size of garage in car capacity(GarageCars) the SalePrice is high.
* In between april to august for Month Sold(MoSold) the sales is good with SalePrice.
* For all the Year\_SinceSold the salePrice and sales both are same.



****

# Observations:

* For Floating Village Residential(FV) and Residential Low Density(RL) zoning classification of the sale(MSZoning) the saleprice is high.
* For paved type of road access to property(Street) the SalePrice is high.
* For Slightly irregular(IR1), Moderately Irregular(IR2) and Irregular(IR3) shape of property(LotShape) the SalePrice is high.
* For Hillside - Significant slope from side to side(HLS) Flatness of the property(LandContour) the SalePrice is High.
* For Cul-de-sac(CulDSac) Lot configuration(LotConfig) the SalePrice is High.
* For all types of Slope of property(LandSlope) i.e.,Gentle slope(Gtl), Moderate Slope(Mod) and Severe Slope(Sev) the SalePrice is High.
* For Northridge(NoRidge) locations within Ames city limits(Neighborhood) the SalePrice is High.
* For Within 200' of North-South Railroad(RRNn), Adjacent to postive off-site feature(PosA) and Near positive off-site feature--park, greenbelt, etc.(PosN) Proximity to various conditions(Condition1) has the maximum SalePrice.
* For Adjacent to postive off-site feature(PosA) and Near positive off-site feature--park, greenbelt, etc.(PosN) Proximity to various conditions (if more than one is present)(Condition2) has maximum SalePrice.
* For Single-family Detached(1Fam) and Townhouse End Unit(TwnhsE) type of dwelling(BldgType) the SalePrice is high.
* For 2Story and Two and one-half story: 2nd level finished(2.5Fin) Style of dwelling(HouseStyle) the SalePrice is high.
* For Shed Type of roof(RoofStyle) the SalePrice is high.
* For Wood Shingles(WdShngl) Roof material(RoofMat1) the SalePrice is high.
* For Cement Board(CemntBd), Imitation Stucco(ImStucc) and Stone type of Exterior covering on house(Exterior1st) the SalePrice is high.
* For Cement Board(CemntBd), Imitation Stucco(ImStucc) and other Exterior covering on house (if more than one material)(Exterior2) has maximum SalePrice.
* For Stone Masonry veneer type(MasvnrType) the SalePrice is high.
* For Excellent(Ex) quality of the material on the exterior(ExterQual) the SalePrice is high.
* For Excellent(Ex) present condition of the material on the exterior(ExterCond) the SalePrice is high.
* For Poured Contrete(PConc) Type of foundation(Foundation) the SalePrice is high.
* For Excellent(100+ inches)(Ex) height of the basement(BsmtQual) the SalePrice is high.
* For Good(Gd) general condition of the basement(BsmtCond) the SalePrice is high.
* For Good Exposure(Gd) of walkout or garden level walls(BsmtExposure) has maximum SalePrice.
* For Good Living Quarters(GLQ) of basement finished area(BsmtFinType1) has maximum SalePrice.
* For Good Living Quarters(GLQ) and Average Living Quarters(ALQ) of basement finished area (if multiple types)(BsmtFinType2) has maximum SalePrice.
* For Gas forced warm air furnace(GasA) and Gas hot water or steam heat(GasW) Type of heating(Heating) has high SalePrice.
* For Excellent(Ex) Heating quality and condition(HeatingQC) the SalePriceis high.
* For building having Central air conditioning(CentralAir) the SalePrice is high.
* For Standard Circuit Breakers & Romex(Sbrkr) of Electrical system(Electrical) the SalePrice is Maximum.
* For Excellent(Ex) Kitchen quality(KitchenQual) the SalePrice is high.
* For Typical Functionality(Typ) type of Home functionality (Assume typical unless deductions are warranted)(Functional) the SalePrice is high.
* For Excellent - Exceptional Masonry Fireplace(Ex) of Fireplace quality(FireplaceQual) has highest SalePrice.
* For Built-In (Garage part of house - typically has room above garage)(BuiltIn) Garage location(GarageType) the SalePrice is maximum.
* For Completely finished(Fin) Interior of the garage(GarageFinish) the SalePrice is high.
* For Excellent(Ex) Garage quality(GarageQual) the SalePrice is high.
* For Typical/Average(TA) and Good(Gd) Garage condition(GarageCond) the SalePrice is high.
* For having Paved driveway(PavedDrive) the SalePriceis high.
* For Home just constructed and sold(New) and Contract 15% Down payment regular terms(Con) of type of sale(SaleType) has highest SalePrice.
* For Home was not completed when last assessed (associated with New Homes)(Partial) Condition of sale(SalesCondition) the SalePrice is maximum.

**Identification of possible problem-solving approaches (methods)**

# Checking for outliers:

****

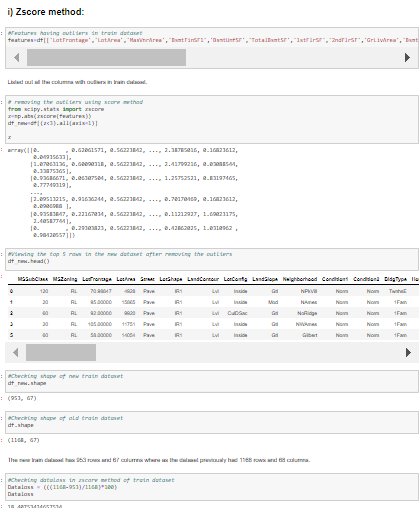
# Observations

The dataset has outliers present in the following columns -

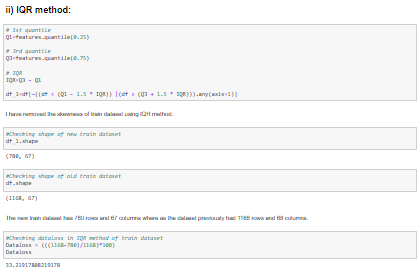
* MSSubClass
* LotFrontage
* LotArea
* OverallQual
* OverallCond
* MasVnrArea
* BsmtFinSF1
* BsmtUnfSF
* TotalBsmtSF
* 1stFlrSF
* 2ndFlrSF
* GrLivArea
* BsmtFullBath
* BsmtHalfBath
* BedroomAbvGr
* KitchenAbvGr
* TotRmsAbvGrd
* Fireplaces
* GarageCars
* GarageArea
* WoodDeckSF
* OpenPorchSF
* Year\_SinceBuilt
* GarageAge
* SalePrice

Since SalePrice is the target we need not remove outliers from this column.And some of the columns like MSSubClass, OverallQual and OverallCond are categorical so we need not remove outliers in those categorical columns.

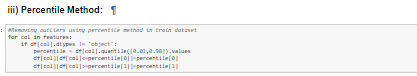
# Removing Outliers in train dataset:

****

Using Z-score method for trian dataset the data loss is more than 10% so let us have a look into IQR method to remove outliers.

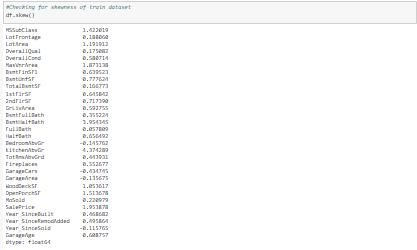
****

The new train dataset has 780 rows and 67 columns where as the dataset previously had 1168 rows and 68 columns. In IQR method of train dataset the data loss is more than 10% so let us have a look into percentile method to remove outliers.

****

We have successfully removed outliers in train dataset using percentile method.

## Checking for skewness:

****

# Observations

We can observe that the following columns have skewness present in train dataset

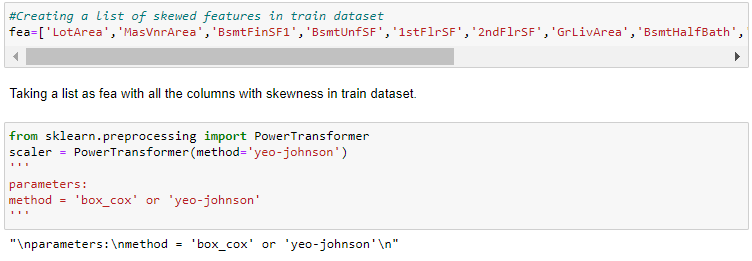
* MSSubClass
* LotArea
* OverallCond
* MasVnrArea
* BsmtFinSF1
* BsmtUnfSF
* 1stFlrSF
* 2ndFlrSF
* GrLivArea
* BsmtHalfBath
* HalfBath
* KitchenAbvGr
* Fireplaces
* WoodDeckSF
* OpenPorchSF
* SalePrice
* GarageAge

SalePrice is the target we need not remove skewness in this column.And MSSubClass and OverallCond are seems to be categorical so let us ignore these columns.

Dropping unnecessary column in train dataset



## Removing skewness using yeo-johnson method for train dataset:

****

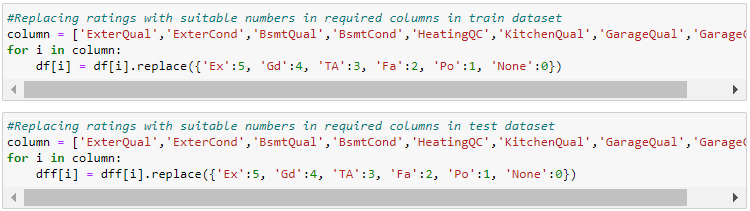
Using yeo\_johnson method we have removed the skewness in train dataset.

****

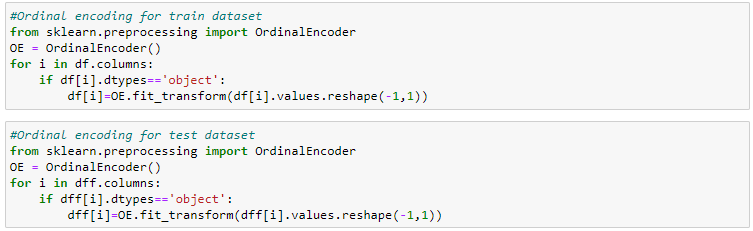
After removing skewness we have high skewness in BsmtHalfBath .let us drop this column from the train dataset

****

# Ordinal Encoding:

****

We have replaced all rating entries with required numbers.

****

We have encoded all my categorical columns in train and test datasets using Ordinal encoder.

# Correlation:

# 

Above are the correlations of all the pair of features of train dataset.To get better visualization on the correlation of features,let us plot it using heat map.

# Visualizing the correlation matrix by plotting heat map for train dataset

# 

# We can clearly observe a multicollinearity issue in some of the features of train dataset so we have to check VIF and Let us plot a bar graph to get better insight on targets correlation with other features.

# 

# Observations

# We can observe the correlation of the target with the other features in the train dataset

# - Here most of the features have positive correlation where as few of the features are negatively correlated.

# Separating features and label in train dataset:

# 

# We have separated the target and independent columns.

# Scaling the train data using standard scaler:

# 

# We have scaled the train data using standard scaler.

# 

# This is the train data of independent variables after scaling.

# Checking for multicolinearity issue in train dataset using VIF:

# 

# We can observe that GrLivArea has high VIF. We can drop this columnn to reduce the impact of multicollinearity in the dataset

# 

# We have dropped the GrLivArea column

# Checking VIF again

# 

# Although there are few columns in the dataset with high VIF than expected, we cannot drop them from the dataset without knowing their importance for the target/ dependent variable, so the multicolinearity issue is solved in train dataset upto some extent.

# Principle Component Analysis on train dataset

# 

# 60 components od the dataset explain around 95% variance in output/target

# Selecting Kbest Features

# 

# We can see the importance of each feature in the above output

# Dropping some least important columns from the dataset

# 

# We have dropped the above columns from the dataset

1. **Test Dataset**

# The EDA, Feature extraction, Feature engineering, Data cleaning, fixing multicollinearity has been done to the test data set in the similar process as we have done to the train dataset.

# Next I have checked for the shape of both the train data and the test data after cleaning and feature engineering

# 

# Both have equal no of columns so we can proceed further with model training

# Model building using train dataset:

# Testing of Identified Approaches (Algorithms)

## Finding Best Random State and Accuracy:

****

We got the best accuracy and random state.

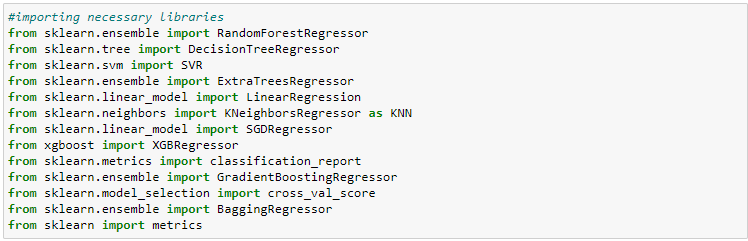
Creating train test split.



Created train test split.

**Run and Evaluate selected models**

## Regression Algorithms:



## i) RandomForestRegressor:



RandomForestRegressor is giving 90.16% r2\_score.

# ii) XGBRegressor:



XGBRegressor is giving me 89.05% r2\_score.

# iii) ExtraTreesRegressor:



ExtraTreesRegressor is giving me 90.77% r2\_score.

# iv) GradientBoostingRegressor:

# 

GradientBoostingRegressor is giving me 92.61% r2\_score.

# v) DecisionTreeRegressor:

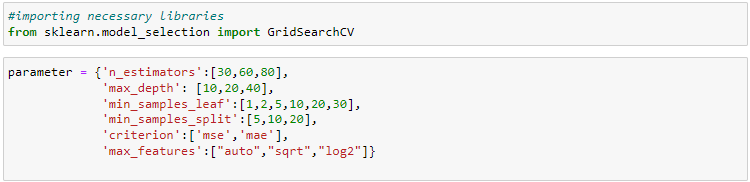
# 

DecisionTreeRegressor is giving me 73.76% r2\_score.

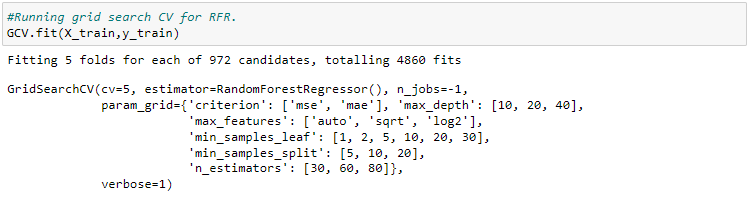
After observing the difference of model accuracy and cross validation score we can see RandomForestRegressor as the best model.

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

# Hyper parameter tunning for best model:



Giving RFR parameters and Running grid search CV for RFR



Tuning the model using GCV.

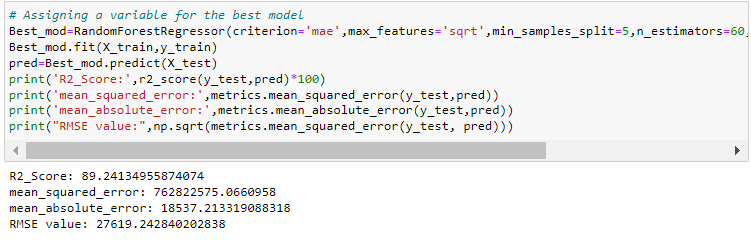
# Getting the best parameters

# 

Got the best parameters for RFR.

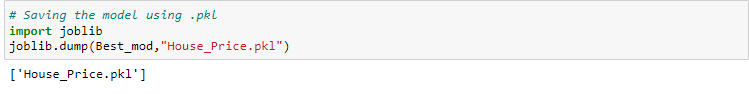
**Interpretation of the Results**

# Assigning a variable for the best model



This is the model r2\_score after tuning. We have 89.24% as r2\_score which is good to proceed with.

# Saving the model:

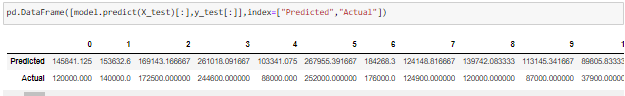


We have saved the model as House\_Price.Using .pkl

## Predicting House Price for test dataset using Saved model of train dataset

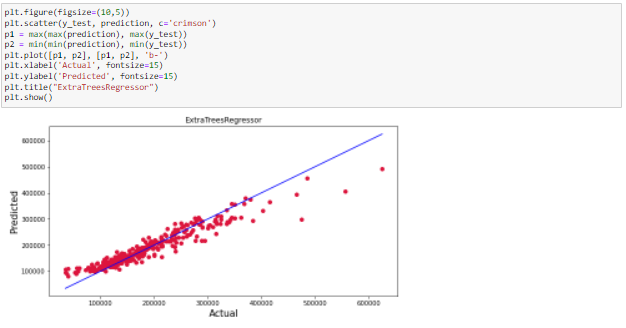
## 

## Creating a data frame of actual vs predicted values



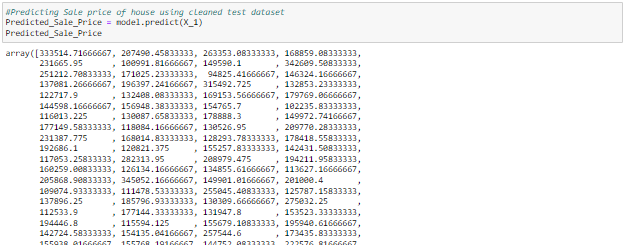
Above are the predicted values and the actual values.They look similar.

Visualizing the actual vs predicted values



Plotting Actual vs Predicted. To get better insights. Blue line is the actual line and red dots are the predicted values. They are similar apart from few exceptions

Predicting Sale price of house using cleaned test dataset



Making dataframe for predicted SalePrice



Saving the predictions to csv



We have saved the predicted values as csv file.

**Conclusion**

**Key Findings and Conclusions of the Study**

Based on the in-depth analysis of the Housing Project, The Exploratory analysis of the datasets, and the analysis of the Outputs of the models the following observations are made:

● Structural attributes of the house Structural attributes of the house like lot size, lot shape, quality and condition of the house, garage capacity, rooms, Lot frontage, number of bedrooms, bathrooms, overall finishing of the house etc play a big role in influencing the house price.

● Neighborhood qualities can be included in deciding house price.

● Various plots like Barplots, stripplots and distplots helped in visualising the Feature-label relationships which corroborated the importance of structural and locational attributes for estimating Sale Prices.

● Due to the Training dataset being very small, the outliers had to be retained for proper training of the models.

● Therefore, Random Forest Regressor, being robust to outliers and being indifferent to nonlinear features, performed well despite having to work on small dataset.

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

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove missing value and to replace null value and zero values with their respective mean, median or mode. This study is an exploratory attempt to use five machine learning algorithms in estimating housing prices, and then compare their results.

To conclude, the application of machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.

**Limitations of this work and Scope for Future Work**

* First drawback is the data leakage when we merge both train and test datasets.
* Followed by huge number of outliers and skewness which will reduce our model accuracy and efficiency.
* Also, we have tried best to deal with outliers, skewness, null values and zero values. So that we have achieved an accuracy of 90.13% even after dealing with these drawbacks.
* Also, this study will not cover all regression algorithms instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones.
* This model doesn't predict future prices of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process.
* While features that focus on structural and locational attributes of housing properties are crucial for estimating the Sale Price of Housing properties, they aren’t the only factors that influence the value in the housing market. Data on Demographics (Age, Income, Regional preferences of buyers, purpose of buying a property) is very important for understanding the Housing market. Interest Rates too impact the price and demand of houses. Economic cycles also influence Real Estate prices. Government Policies, Regulations, Legalizations are also important factors that may influence the sales of houses. The availability of data on above features would help build a predictive model that would more accurately understand the relationship between the features and target variable and yield more accurate predictions.
* To conclude, the application of machine learning in property research is still at an early stage. We hope this study can provide some methodological improvements and contributions to property calculation and presenting an alternative to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data including more features.

Thank you