

HOUSE PRICE PREDICTION

Submitted by:

VIKAS OJHA

**ACKNOWLEDGMENT**

First of all, I would like to thank Flip Robo Technologies for giving me an opportunity to work on this project. I would also like to thank our Institute DataTrained for giving an excellent platform to learn Machine Learning using Jupyter Notebook as a part of Data Scientist course.

This project has been completed under the guidance of our SME Mr. Shubham Yadav who helped us in completing the project.

This project is based on data collected from telecom Industry by Microfinance Institution (MFI) for providing loans to the needy persons which is to be repay back to the company within given time frame. The person who is not able to repay back are Defaulters, which we need to identify by using Machine Learning Techniques.

Thanks

Vikas Ojha

**INTRODUCTION**

* Business Problem Framing

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main objective of our project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price

With a large number of unstructured resources and documents, the Real estate industry has become a highly competitive business. The data science process in such an industry provides an advantage to the developers by processing those data, forecasting future trends and thus assisting them to make favorable knowledge- driven decisions.

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, and then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

There are three factors that influence the price of a house which include physical conditions, concept and location. There are several approaches that can be used to determine the price of the house, one of them is the prediction analysis. We use various regression techniques in this pathway.

* Conceptual Background of the Domain Problem

The real estate market is one of the most crucial components of any national economy. Hence, observations of the real estate market and accurate predictions of real estate prices are helpful for real estate buyers and sellers as well as economic specialists. However, real estate forecasting is a complicated and difficult task owing to many direct and indirect factors that inevitably influence the accuracy of predictions.

In general, factors influencing real estate prices could be quantitative or qualitative. The quantitative factors possibly include macroeconomic factors, business cycles, and real estate attributes. The macroeconomic factors contain unemployment rates, share index, current account of a country, industrial production, and gross domestic product. Attributes of real estate, for example, includes past sale prices, land area, years of constructions, floor space, surface area, number of floors and building conditions, etc. The qualitative factors refer to subject preferences of decision makers, such as views, building styles, and living environment.

The real estate sector is a sector whose reach is vast; it is therefore affected by many social, political and economic factors which mean that there is a huge amount of complex data available. It is through analyzing and understanding this data that models can be created which aim to replicate the changes in the sector and evolve in order to anticipate what we might see in the future.

* Review of Literature

Real estate price prediction is crucial for the establishment of real estate policies and can help real estate owners and agents make informative decisions. The aim of this study is to employ actual transaction data and machine learning models to predict prices of houses. The actual transaction data contain attributes and transaction prices of real estate that respectively serve as independent variables and dependent variables for machine learning models.

Real estate is one of the most fast-paced and emerging industries today. Nowadays everyone wants to be the owner of their house rather than live on rent. Therefore, people are very cautious in searching for the most suitable house. Different people have different budgets and so are their desires. This project draws attention to the house rate predictions based on different objectives like financial status and expectations of non-house holders.

Housing is the utmost need of any individual; it can be either bought or rented. With time more and more people are drawn towards buying their own houses. Each has a different set of budgets and priorities for their house. People also surf the internet to search the house of their choice. For fulfilling their needs, machine learning engineers from across the world with data scientists are working to predict precise results to the shareholders and customers.

The models use variables regarding house facilities like interior square feet of the property, number of bedrooms, number of bathrooms, total number of rooms, the quality score assigned for rooms based on buyer reviews, the quality score assigned for bathroom based on buyer reviews, the quality score assigned for bedroom based on buyer reviews, the overall quality score assigned for the property, the sale condition of the house and the type of building.

* Motivation for the Problem Undertaken

House Price prediction is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality.

The House Price prediction model is very essential in filling the information gap and improves Real Estate efficiency. The goal of the paper is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analysing previous market trends and price ranges, and also upcoming developments future prices will be predicted. This model will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

Nowadays, e-education and e-learning is highly influenced. Everything is shifting from manual to automated systems. The objective of this project is to predict the house prices so as to minimize the problems faced by the customer. The present method is that the customer approaches a real estate agent to manage his/her investments and suggest suitable estates for his investments. But this method is risky as the agent might predict wrong estates and thus leading to loss of the customer’s investments.

In this project, the main focus is on developing a model which not only predicts the sale price of properties for a customer according to his\her interests, but also recognizes the most preferred location of real estate and other utilities and features in any given area.

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 markets which are 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?

We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. Real estate developers can then accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Thus, the machine learning-based model is a substantial and feasible way to forecast real estate prices, and can provide relatively competitive and satisfactory results.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We have been given a dataset which contains 2 different files, train and test. We will analyse the model using train dataset and prepare best model. This model will be used on test dataset to predict the outcome.

After analysing the dataset, it was observed that it is a Regression problem, where our end goal is to predict the Prices of House based on given data. I will be dividing my train dataset into Training and Testing parts. A Regression Model will be built and trained using the Training data and the Test data will be used to predict the outcomes. This will be compared with available test results to find how well the model has performed.

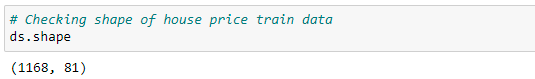
The ‘r2’ score will be used to determine the best model among, Linear Regression, Decision Tree Regressor, Random Forest Regressor, KNeighbors Regressor, AdaBoost Regressor, SVR, Lasso, Ridge, Elastic Net and Gradient Boosting Regressor

The best results were obtained using Random Forest Regressor model.

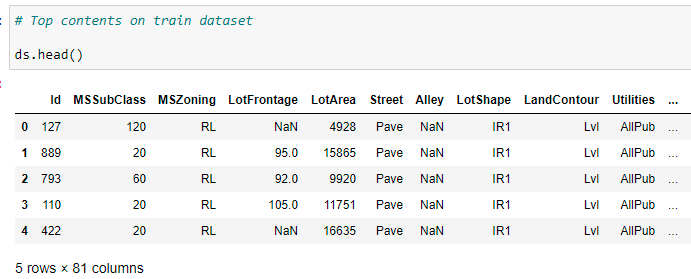
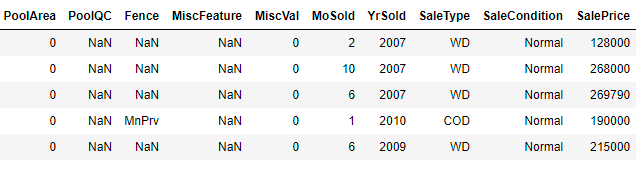
* Data Sources and their formats

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.

As per provided data, there were 1168 rows and 81 columns.



Here are the top data of our dataset.



The last column Sale Price is our Target variable.

Data Description:

MSSubClass: Identifies the type of dwelling involved in the sale.

* 20 1-story 1946 & newer all styles 75 2-1/2 story all ages
* 30 1-story 1945 & older 80 split or multi-level
* 40 1-story w/finished attic all ages 85 split foyer
* 45 1-1/2 story - unfinished all ages 90 duplex - all styles and ages
* 50 1-1/2 story finished all ages 150 1-1/2 story pud - all ages
* 60 2-story 1946 & newer 160 2-story pud - 1946 & newer
* 70 2-story 1945 & older 180 pud - multilevel - incl split lev/foyer
* 120 1-story pud (planned unit development) - 1946 & newer
* 190 2 family conversion - all styles and ages

MSZoning: Identifies the general zoning classification of the sale.

* A: Agriculture
* C: Commercial
* FV: Floating Village Residential
* I: Industrial
* RH: Residential High Density
* RL: Residential Low Density
* RP: Residential Low Density Park
* RM: Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

* Grvl Gravel
* Pave Paved

Alley: Type of alley access to property

* Grvl: Gravel
* Pave: Paved
* NA: No alley access

LotShape: General shape of property

* Reg: Regular
* IR1: Slightly irregular
* IR2: Moderately Irregular
* IR3: Irregular

LandContour: Flatness of the property

* Lvl: Near Flat/Level
* Bnk: Banked - Quick and significant rise from street grade to building
* HLS Hillside - Significant slope from side to side Low: Depression

Utilities: Type of utilities available

* AllPub: All public Utilities (E, G, W & S)
* NoSewr: Electricity, Gas, and Water (Septic Tank)
* NoSeWa: Electricity and Gas Only ELO: Electricity only

LotConfig: Lot configuration

* Inside: Inside lot
* Corner: Corner lot
* CulDSac: Cul-de-sac
* FR2: Frontage on 2 sides of property
* FR3 Frontage on 3 sides of property

LandSlope: Slope of property

* Gtl: Gentle slope
* Mod: Moderate Slope
* Sev: Severe Slope

Neighborhood: Physical locations within Ames city limits

* Blmngtn: Bloomington Heights
* Blueste: Bluestem
* BrDale: Briardale BrkSide: Brookside
* ClearCr: Clear Creek
* CollgCr: College Creek
* Crawfor: Crawford Edwards:Edwards
* Gilbert: Gilbert
* IDOTRR: Iowa DOT and Rail Road
* MeadowV: Meadow Village
* Mitchel:Mitchell
* Names: North Ames
* NoRidge: Northridge
* NPkVill: Northpark Villa
* NridgHt: Northridge Heights
* NWAmes: Northwest Ames
* OldTown: Old Town
* SWISU: South & West of Iowa State University
* Sawyer: Sawyer
* SawyerW: Sawyer West
* Somerst: Somerset
* StoneBr: Stone Brook
* Timber: Timberland Veenker: Veenker

Condition1: Proximity to various conditions

* Artery: Adjacent to arterial street
* Feedr: Adjacent to feeder street
* Norm: Normal
* RRNn: Within 200' of North-South Railroad
* RRAn: Adjacent to North-South Railroad
* PosN: Near positive off-site feature--park, greenbelt, etc.
* PosA: Adjacent to positive off-site feature
* RRNe: Within 200' of East-West Railroad
* RRAe: Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

* Artery: Adjacent to arterial street
* Feedr: Adjacent to feeder street Norm: Normal
* RRNn: Within 200' of North-South Railroad
* RRAn: Adjacent to North-South Railroad
* PosN: Near positive off-site feature--park, greenbelt, etc. PosA: Adjacent to postive off-site feature
* RRNe: Within 200' of East-West Railroad
* RRAe: Adjacent to East-West Railroad

BldgType: Type of dwelling

* 1Fam: Single-family Detached
* 2FmCon: Two-family Conversion; originally built as one-family dwelling
* Duplx: Duplex
* TwnhsE: Townhouse End Unit
* TwnhsI: Townhouse Inside Unit

HouseStyle: Style of dwelling

* 1Story: One story
* 1.5Fin: One and one-half story: 2nd level finished
* 1.5Unf: One and one-half story: 2nd level unfinished
* 2Story: Two story
* 2.5Fin: Two and one-half story: 2nd level finished
* 2.5Unf: Two and one-half story: 2nd level unfinished SFoyer Split Foyer SLvl Split Level

OverallQual: Rates the overall material and finish of the house

* 10 Very Excellent
* 9 Excellent
* 8 Very Good
* 7 Good
* 6 Above Average
* 5 Average
* 4 Below Average
* 3 Fair
* 2 Poor
* 1 Very Poor

OverallCond: Rates the overall condition of the house

* 10 Very Excellent
* 9 Excellent
* 8 Very Good
* 7 Good
* 6 Above Average
* 5 Average
* 4 Below Average
* 3 Fair
* 2 Poor
* 1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

* Flat: Flat
* Gable: Gable
* Gambrel: Gabrel (Barn)
* Hip: Hip
* Mansard: Mansard
* Shed: Shed

RoofMatl: Roof material

* ClyTile: Clay or Tile
* CompShg: Standard (Composite) Shingle
* Membran: Membrane
* Metal: Metal
* Roll: Roll
* Tar&Grv: Gravel & Tar
* WdShake: Wood Shakes
* WdShngl: Wood Shingles

Exterior1st: Exterior covering on house

* AsbShng: Asbestos Shingles
* AsphShn: Asphalt Shingles
* BrkComm: Brick Common
* BrkFace: Brick Face
* CBlock: Cinder Block
* CemntBd Cement Board
* HdBoard: Hard Board
* ImStucc: Imitation Stucco
* MetalSd: Metal Siding
* Other: Other
* Plywood: Plywood
* PreCast: PreCast
* Stone: Stone
* Stucco: Stucco
* VinylSd: Vinyl Siding
* Wd Sdng: Wood Siding
* WdShing: Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

* AsbShng: Asbestos Shingles
* AsphShn: Asphalt Shingles
* BrkComm: Brick Common
* BrkFace: Brick Face
* CBlock: Cinder Block
* CemntBd Cement Board
* HdBoard: Hard Board
* ImStucc: Imitation Stucco
* MetalSd: Metal Siding
* Other: Other
* Plywood: Plywood
* PreCast: PreCast
* Stone: Stone
* Stucco: Stucco
* VinylSd: Vinyl Siding
* Wd Sdng: Wood Siding
* WdShing: Wood Shingles

MasVnrType: Masonry veneer type

* BrkCmn: Brick Common
* BrkFace: Brick Face
* CBlock: Cinder Block
* None: None
* Stone: Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

* Ex: Excellent
* Gd: Good
* TA: Average/Typical
* Fa: Fair
* Po: Poor

ExterCond: Evaluates the present condition of the material on the exterior

* Ex: Excellent
* Gd: Good
* TA: Average/Typical
* Fa: Fair
* Po: Poor

Foundation: Type of foundation

* BrkTil: Brick & Tile
* CBlock: Cinder Block
* PConc: Poured Contrete
* Slab: Slab
* Stone: Stone
* Wood: Wood

BsmtQual: Evaluates the height of the basement

* Ex: Excellent (100+ inches)
* Gd: Good (90-99 inches)
* TA: Typical (80-89 inches)
* Fa: Fair (70-79 inches)
* Po: Poor (<70 inches)
* NA: No Basement

BsmtCond: Evaluates the general condition of the basement

* Ex: Excellent (100+ inches)
* Gd: Good (90-99 inches)
* TA: Typical (80-89 inches)
* Fa: Fair (70-79 inches)
* Po: Poor (<70 inches)
* NA: No Basement

BsmtExposure: Refers to walkout or garden level walls

* Gd: Good Exposure
* Av: Average Exposure (split levels or foyers typically score average or above)
* Mn: Mimimum Exposure
* No: No Exposure
* NA: No Basement

BsmtFinType1: Rating of basement finished area

* GLQ: Good Living Quarters
* ALQ: Average Living Quarters
* BLQ: Below Average Living Quarters
* Rec: Average Rec Room
* LwQ: Low Quality
* Unf: Unfinshed
* NA: No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

* GLQ: Good Living Quarters
* ALQ: Average Living Quarters
* BLQ: Below Average Living Quarters
* Rec: Average Rec Room
* LwQ: Low Quality
* Unf: Unfinshed
* NA: No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

* Floor: Floor Furnace
* GasA: Gas forced warm air furnace
* GasW: Gas hot water or steam heat
* GravGravity: furnace
* OthW: Hot water or steam heat other than gas
* Wall: Wall furnace

HeatingQC: Heating quality and condition

* Ex: Excellent
* Gd: Good
* TA: Average/Typical
* Fa: Fair
* Po: Poor

CentralAir: Central air conditioning

* N: No
* Y: Yes

Electrical: Electrical system

* SBrkr: Standard Circuit Breakers & Romex
* FuseA: Fuse Box over 60 AMP and all Romex wiring (Average)FuseF: 60 AMP Fuse Box and mostly Romex wiring (Fair)
* FuseP: 60 AMP Fuse Box and mostly knob & tube wiring (poor)
* Mix: Mixed

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

* Ex: Excellent
* Gd: Good
* TA: Typical/Average
* Fa: Fair
* Po: Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

* Functional: Home functionality (Assume typical unless deductions are warranted)
* Typ: Typical Functionality
* Min1: Minor Deductions 1
* Min2: Minor Deductions 2
* Mod: Moderate Deductions
* Maj1: Major Deductions 1
* Maj2: Major Deductions 2
* Sev: Severely Damaged
* Sal: Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

* Ex: Excellent - Exceptional Masonry Fireplace
* Gd: Good - Masonry Fireplace in main level
* TA: Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
* Fa: Fair - Prefabricated Fireplace in basement
* Po: Poor - Ben Franklin Stove
* NA: No Fireplace

GarageType: Garage location

* 2Types: More than one type of garage
* Attchd: Attached to home Basment Basement Garage
* BuiltIn: Built-In (Garage part of house - typically has room above garage)
* CarPort: Car Port Detchd Detached from home
* NA: No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

* Fin: Finished
* RFn: Rough Finished
* Unf: Unfinished
* NA: No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

* Ex:Excellent
* Gd:Good
* TA: Typical/Average
* Fa: Fair
* Po: Poor
* NA: No Garage

GarageCond: Garage condition

* Ex: Excellent
* Gd: Good
* TA: Typical/Average
* Fa: Fair
* Po: Poor
* NA: No Garage

PavedDrive: Paved driveway

* Y: Paved
* P: Partial Pavement
* N: Dirt/Gravel

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

* Ex: Excellent
* Gd: Good
* TA: Average/Typical
* Fa: Fair
* NA: No Pool

Fence: Fence quality

* GdPrv: Good Privacy
* MnPrv: Minimum Privacy
* GdWo: Good Wood
* MnWw: Minimum Wood/Wire
* NA: No Fence

MiscFeature: Miscellaneous feature not covered in other categories

* Elev: Elevator
* Gar2: 2nd Garage (if not described in garage section)
* Othr: Other
* Shed: Shed (over 100 SF)
* TenC: Tennis Court
* NA: None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

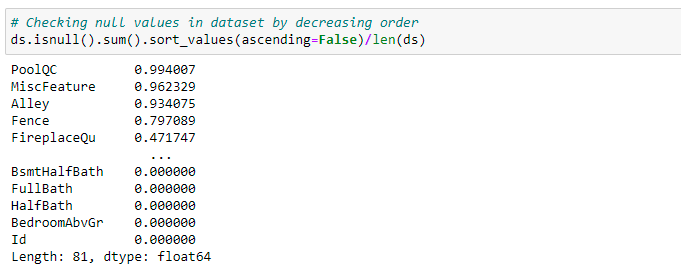
SaleType: Type of sale

* WD: Warranty Deed - Conventional
* CWD: Warranty Deed - Cash
* VWD: Warranty Deed - VA Loan
* New: Home just constructed and sold
* COD: Court Officer Deed/Estate
* Con: Contract 15% Down payment regular terms
* ConLw: Contract Low Down payment and low interest
* ConLI: Contract Low Interest
* ConLD: Contract Low Down
* Oth: Other

SaleCondition: Condition of sale

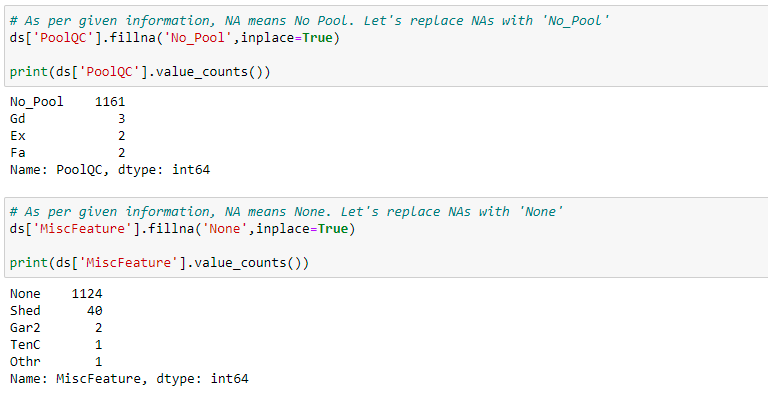
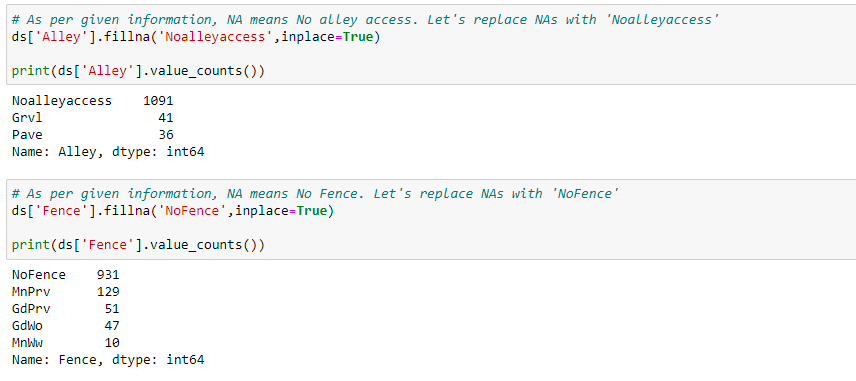
* Normal: Normal Sale
* Abnorml: Abnormal Sale - trade, foreclosure, short sale
* AdjLand: Adjoining Land Purchase
* Alloca: Allocation - two linked properties with separate deeds, typically condo with a garage unit
* Family: Sale between family members
* Partial: Home was not completed when last assessed (associated with New Homes)

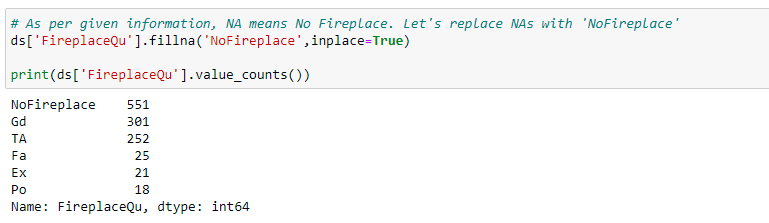
**DATA PREPROCESSING DONE**

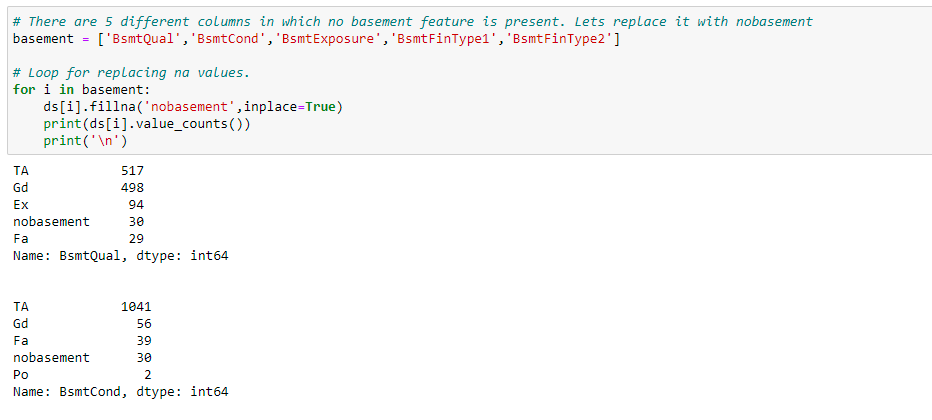


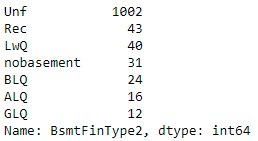
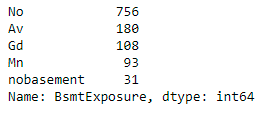
We can see from above that there are many missing values in dataset. Initially I will replace few nan values with data given in description file.

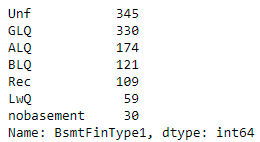
1. **REPLACING NAN VALUES**

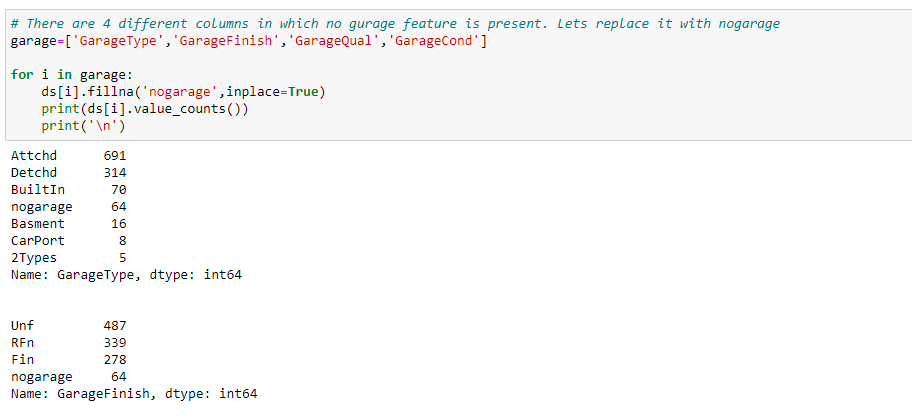
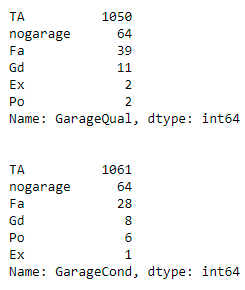






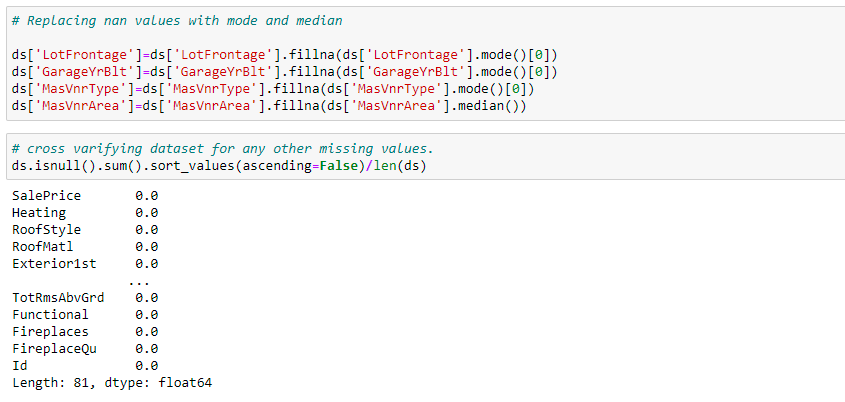






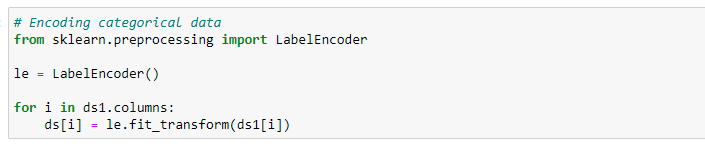
At present we have replaced few missing values with correct data as per description file provided.

Now l has checked other columns with missing values and I’ll replace them with suitable data.



Now there are no missing values in our train dataset.

1. **LABEL ENCODING**



Here I’ve converted categorical variables to numeric form with the help of Lavel Encoder.

1. **SKEWNESS REDUCTION**



Here I’ve removed the skewness from dataset.

1. **STANDARD SCALER**

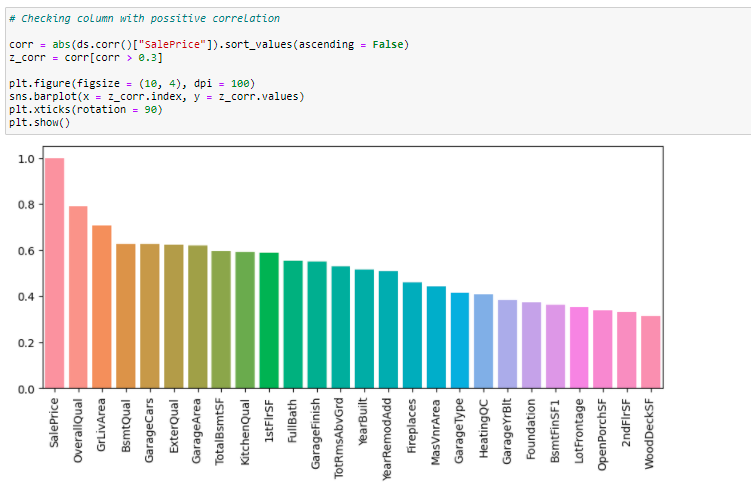
To bring our dataset to a uniform scale, I’ve scaled the x data (Independent Features) of dataset using Standard Scaler.



* Data Inputs- Logic- Output Relationships

In order to gain insights of the relationship between inputs and the target variable(output) I have studied the correlation using pandas function.

The features with correlation of more than 0.3 are shown below.



Following observations were drawn: -

1) These features or columns are highly correlated with target Sale Price:

* OverallQual
* YearBuilt
* YearRemodAdd
* MasVerArea
* Foundation
* BsmtFinSF1
* TotoalBsmtSF
* 1stFlrSF
* 2ndFlrSF
* GrLivArea
* FullBath
* TotRmsAbvGrd
* Fireplaces
* GarageYrBlt
* GarageCars
* GarageArea
* OpenPorchSF

2) Following features or columns are negatively correlated with target Sale Price:

* LotShape
* ExterQual
* BsmtQual
* BsmtExposure
* HeatingQC
* KitchenQual
* GarageType
* GarageFinish
* Hardware and Software Requirements and Tools Used

Open-source web-application used for programming:

1. Anaconda 2020.07
2. Jupyter Notebook(6.1.4)

Python Libraries / Packages used were:

1. Pandas: Pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language. I have used pandas to import the csv. All data analysis has been done using the pandas and numpy libraries. The data characteristics have been studied using pandas functions like df.shape(), df .dtypes, df.columns etc.
2. NumPy: NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. It can be utilised to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic.
3. Matplotlib: library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
4. Seaborn: Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. All the visualizations are built using the seaborn library. Alias used for seaborn is sns. sns.boxplot(), sns.heatmap(), sns.distplot(), sns.scatterplot(), heatmap are few of the libraries used.
5. SciPy: SciPy, a scientific library for Python is an open-source library for mathematics, science and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The main reason for building the SciPy library is that, it should work with NumPy arrays. In this particular project scipy functions such as scipy.stats is used. Zscores are also obtained via scipy.stats library
6. Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, and clustering and dimensionality reduction via a consistence interface in Python. All the Machine Learning regression algorithms have been imported from the sklearn package. The evaluation metrics, RMSE, MSE,MAE functions are also imported from same.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

One of the most crucial steps of building a machine learning model is data exploration. It is done to gain various insights of our dataset and figure out the errors. Some of the common errors in datasets are missing values, outliers and skewness Dataset is first cleaned, before predicting the accuracy.

I have used the train dataset to make the algorithm learn the data behaviour and make predictions for the test dataset.

The very first step includes identifying the type of problem i.e., whether it is a classification problem or a regression problem. As our target variable “Sale Price” is continuous in nature our problem is regression problem

The very first thought that shall strike us upon looking at the dataset is what features might affect our prediction either positively or negatively. Hence it becomes vital to observe correlation between all the input variables and the target. The corr() function has helped us determine the features of utmost importance.

After selection of the most relatable features the very next aspect is to reduce the asymmetricity in the data using skewness reduction or outlier detection and treatment.

As most of the data was categorical in nature and there were outliers in our dataset but It was observed that after removing them approx. 7% of data was lost. So, I’ve decided to proceed by keeping outliers.

In our project, 80% of the readings have been selected for the train set and the other 20% for the test set. The values for dependent variables are determined by applying certain algorithms on the training set. Then, a training set algorithm was put in the testing data for only independent variables. The obtained dependent variables and original values are compared and the one with the minimum error is selected.

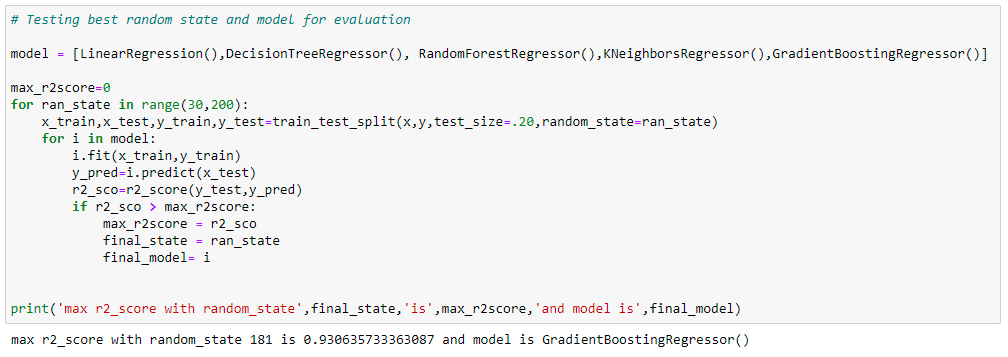
The sale price depends on all the features combined rather than on just two features. This study will further show that the sale price of a house depends on various features combined and not just a few correlated features.

Testing of Identified Approaches (Algorithms)

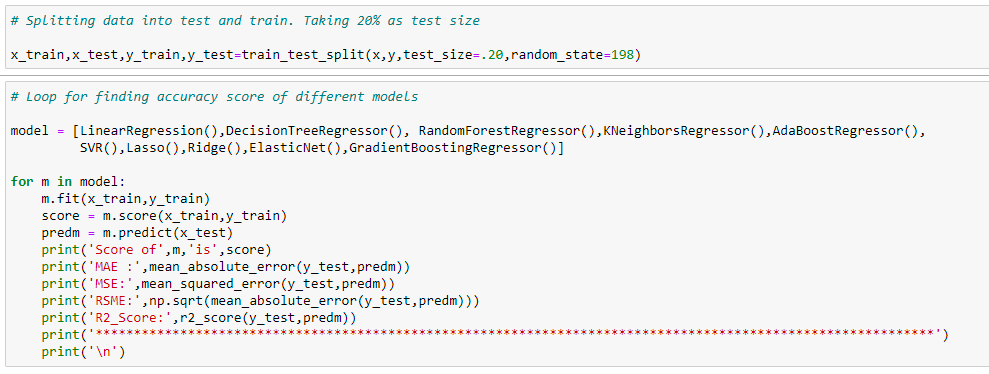
Following algorithms were used to identify the best model that predicts best to the given dataset: -

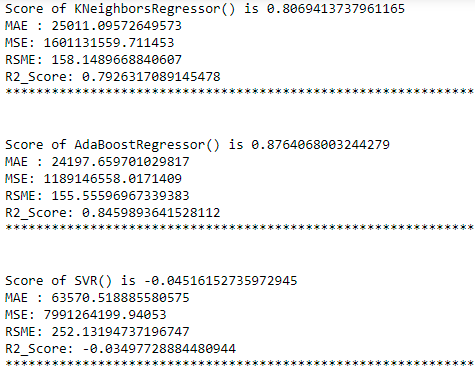
* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor
* KNeighborsRegressor
* AdaBoostRegressor
* SVR
* Lasso
* Ridge
* ElasticNet
* Gradient Boosting Regressor.
* Run and Evaluate selected models

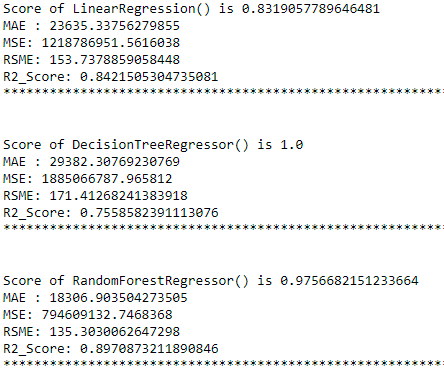
Initially I’ll find out the best random state and best score for the given dataset. Then it will be applied to different models to gain the performance of different models.

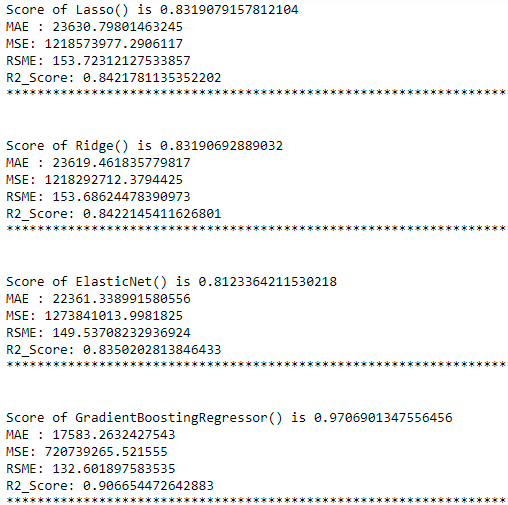


It can be observed that the r2 score of 93% was achieved with random state of 198. Then I’ve divided the dataset into train and test. I’ve chosen 80% of data for training and 20% for testing.



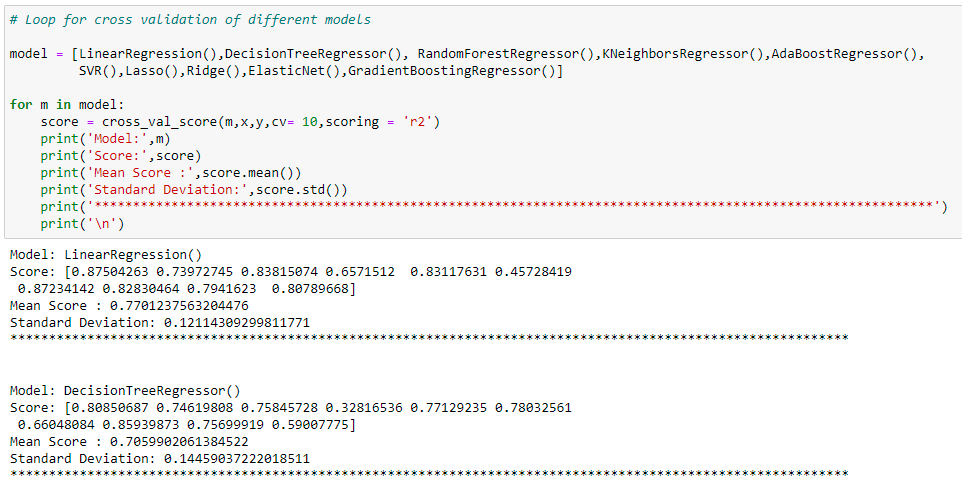
After applying different models following results were found:

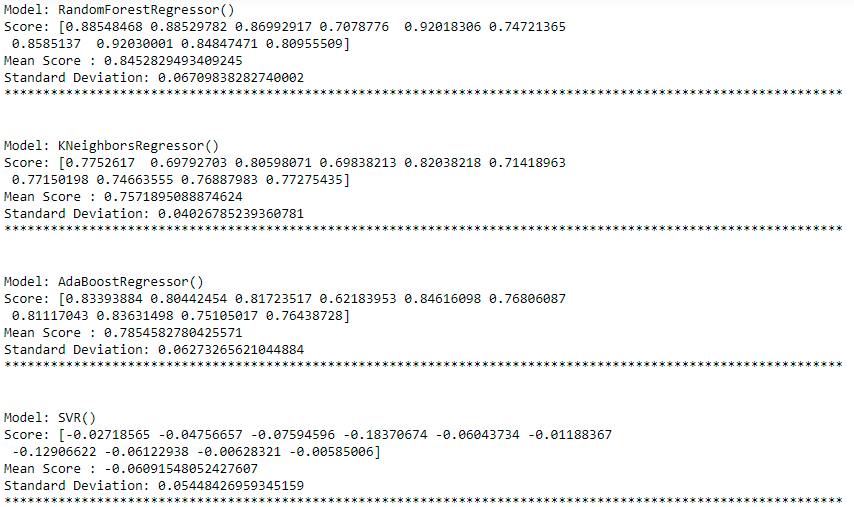


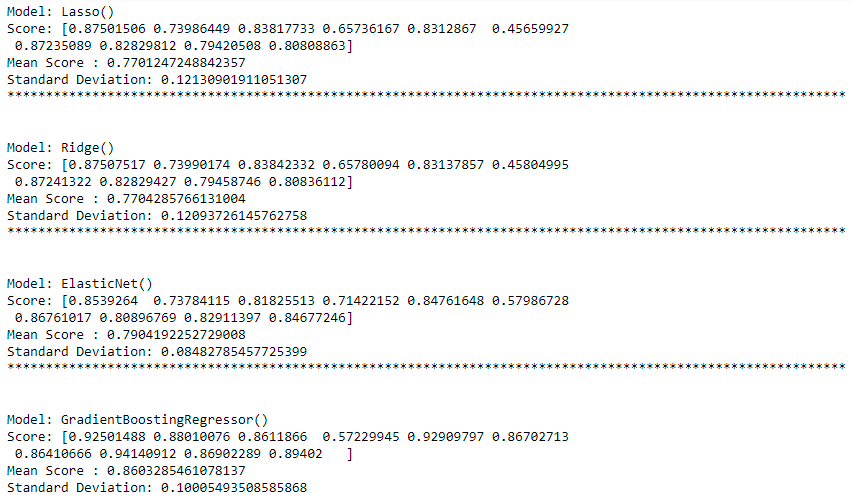


Observations:

1. It was observed that Random Forest Regressor is giving the best score of 97.57%.
2. Let’s run cross validation for different models. If the model is overfitted, the score of the model will reduce. If not, the model is good.





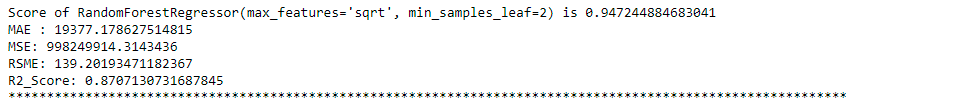


Observations:

Random forest regressor is performing well in cross validation. So, We can say that our model is not overfitted.

Let’s hyper tune the model to check if we can improve the score further.



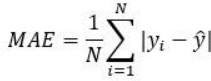
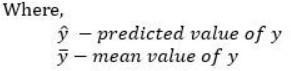


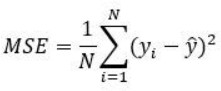
Observations:

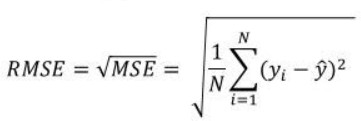
* Earlier the Score of Random Forest Regressor was 98% and R2\_Score was 91.98%.
* After Hypertuning the model, the score was reduced to 96% and R2\_Score was 91.12%.
* We can now proceed further to save our best model.
* Key Metrics for success in solving problem under consideration

The objective of Regression is to find a line that minimizes the prediction error of all the data points. The essential step in any machine learning model is to evaluate the accuracy of the model. The Mean Squared Error, mean absolute error, Root Mean Squared Error, and R-Squared or Coefficient of determination metrics are used to evaluate the performance of the model in regression analysis.

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.



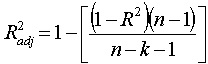
Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

https://miro.medium.com/max/274/1*X0_3mtDXwuhd3dl88xR4yA.png

Adjusted R squared is a modified version of R square, and it is adjusted for the number of independent variables in the model, and it will always be less than or equal to R². In the formula below n is the number of observations in the data and k is the number of the independent variables in the data.



Differences among these evaluation metrics

Mean Squared Error (MSE) and Root Mean Square Error penalizes the large prediction errors vi-a-vis Mean Absolute Error (MAE). However, RMSE is widely used than MSE to evaluate the performance of the regression model with other random models as it has the same units as the dependent variable (Y-axis).

MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE. Therefore, in many models, RMSE is used as a default metric for calculating Loss Function despite being harder to interpret than MAE.

MAE is more robust to data with outliers.

The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable.

R Squared & Adjusted R Squared is used for explaining how well the independent variable in the linear regression model explains the variability in the dependent variable. R Squared value always increases with the addition of the independent variables which might lead to the addition of the redundant variables in our model. However, the adjusted R-squared solves this problem.

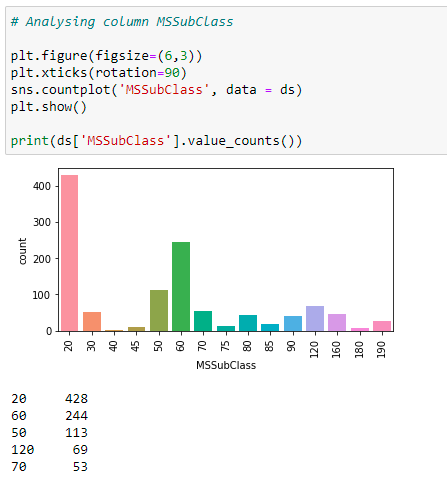
Adjusted R squared takes into account the number of predictor variables, and it is used to determine the number of independent variables in our model. The value of Adjusted R squared decreases if the increase in the R square by the additional variable isn’t significant enough.

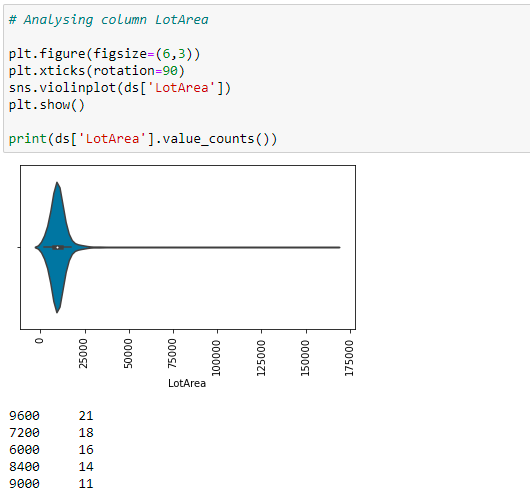
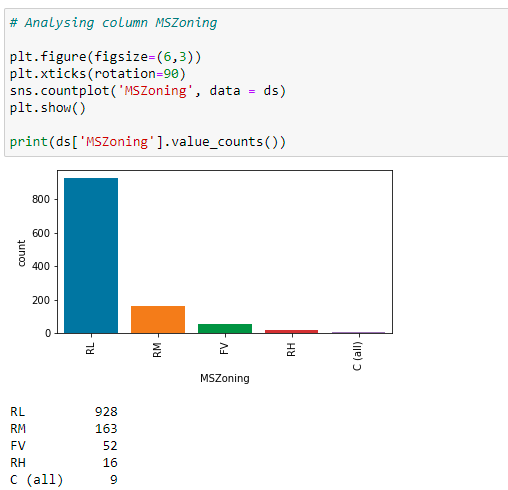
For comparing the accuracy among different linear regression models, RMSE is a better choice than R Squared. Therefore, if comparing the prediction accuracy among different linear regression (LR) models then RMSE is a better option as it is simple to calculate and differentiable. However, if your dataset has outliers then choose MAE over RMSE.

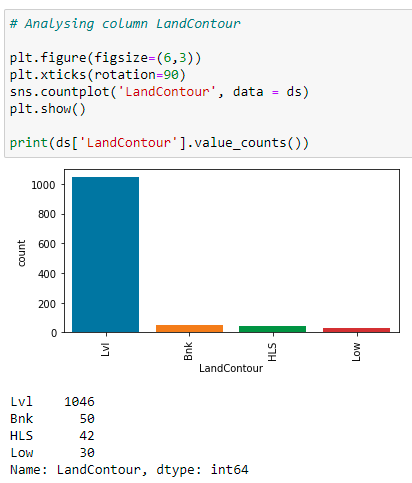
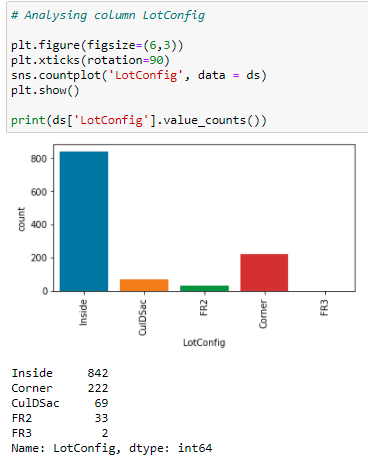
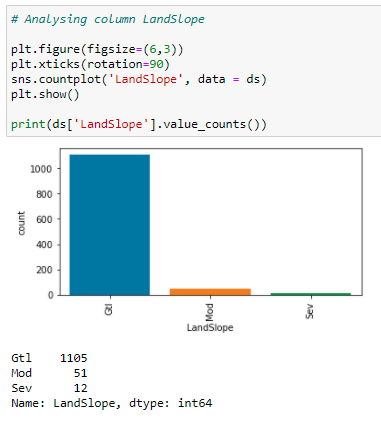
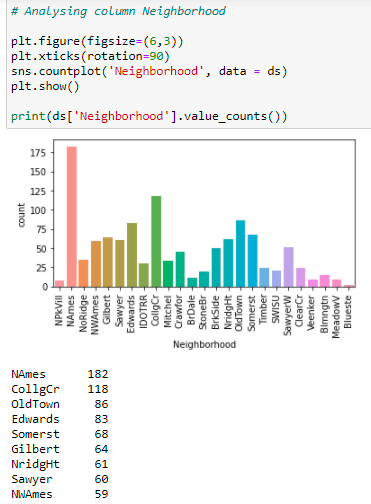
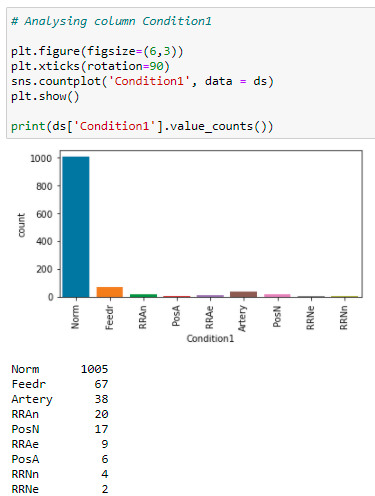
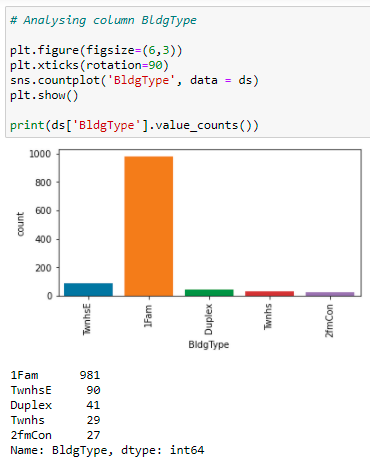
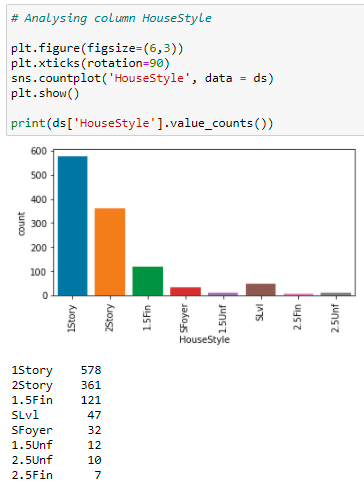
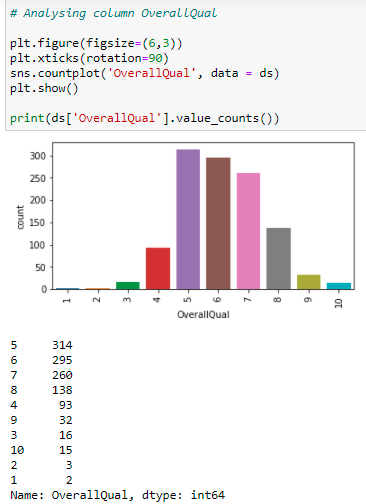
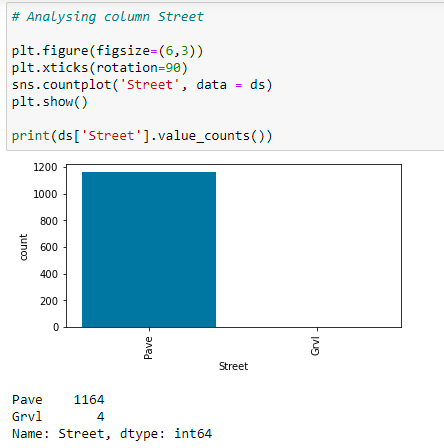
Besides, the number of predictor variables in a linear regression model is determined by adjusted R squared, and choose RMSE over adjusted R squared if you care about evaluating prediction accuracy among different LR models.

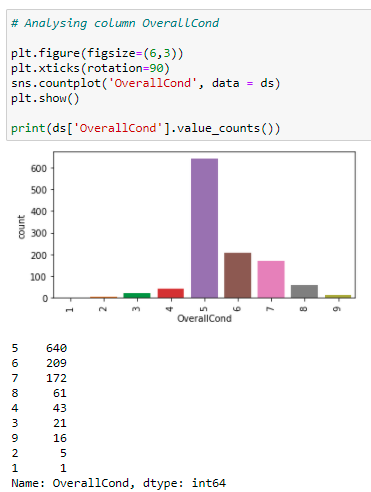
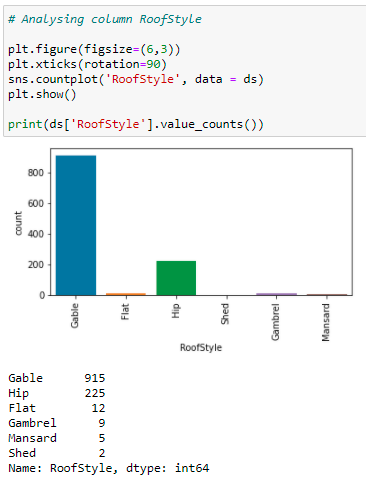
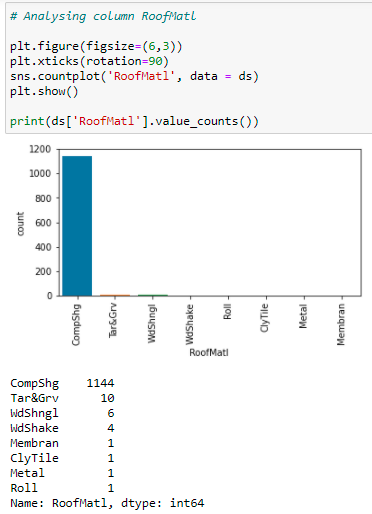
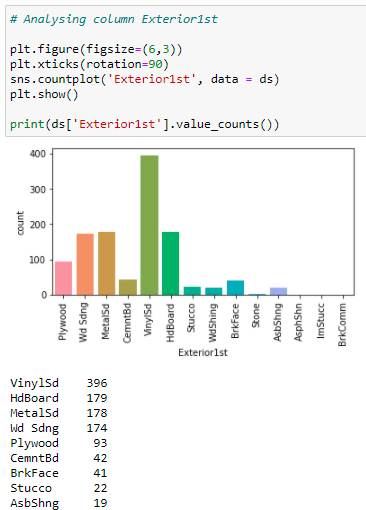
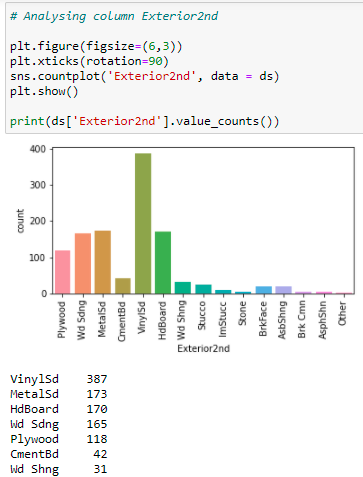
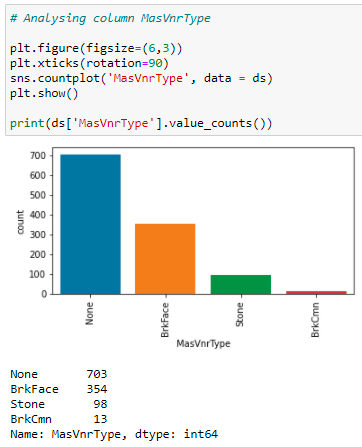
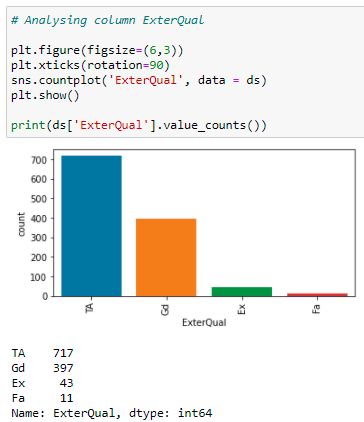
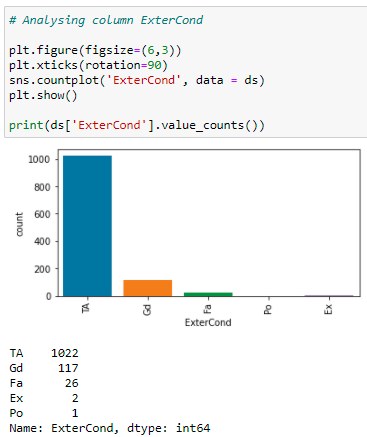
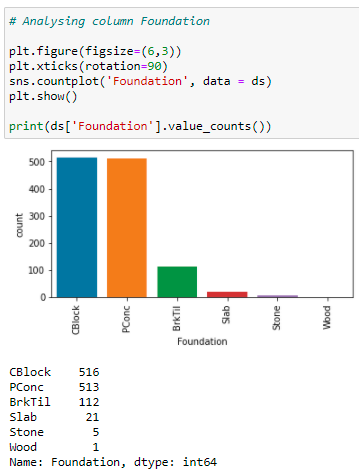
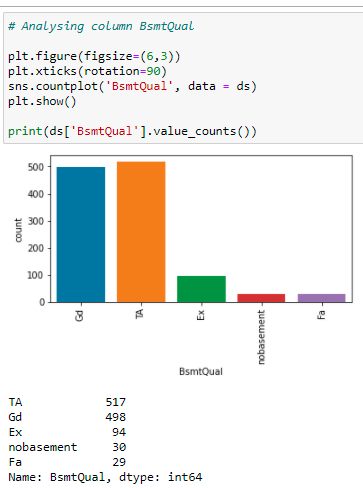
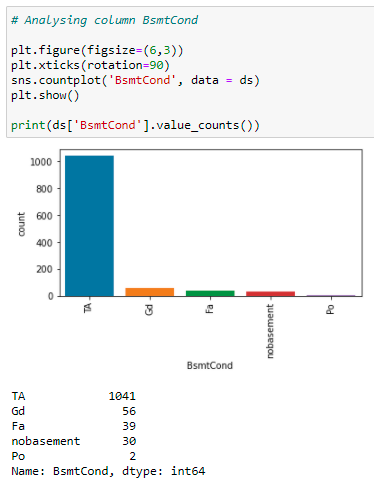
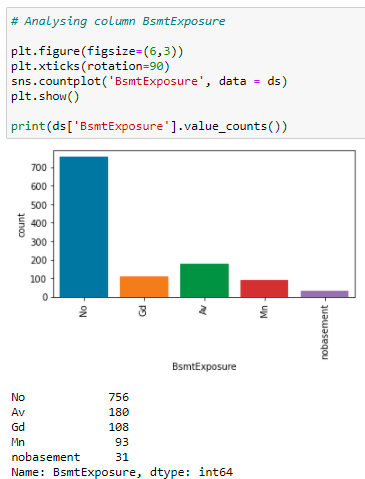
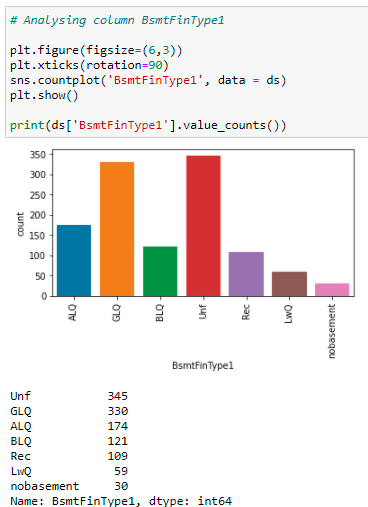
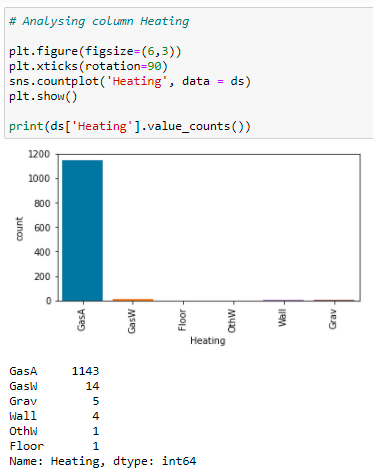
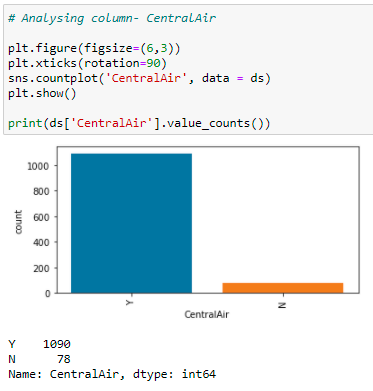
* Visualizations

In Visualization, Univariate and Bivariate Analysis were performed. I’ve also plotted the skewness plot and boxplot to detect the outliers.

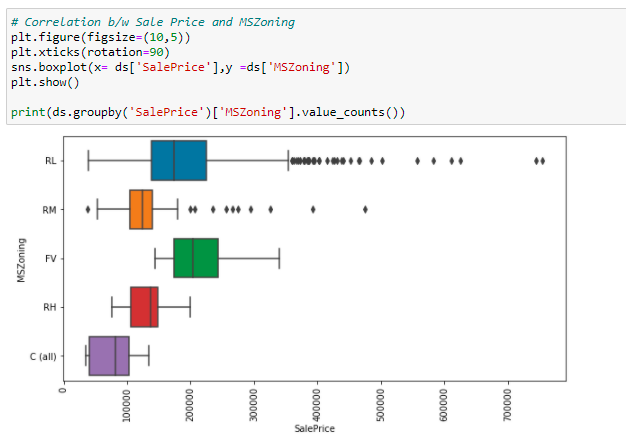
**Univariate Analysis**



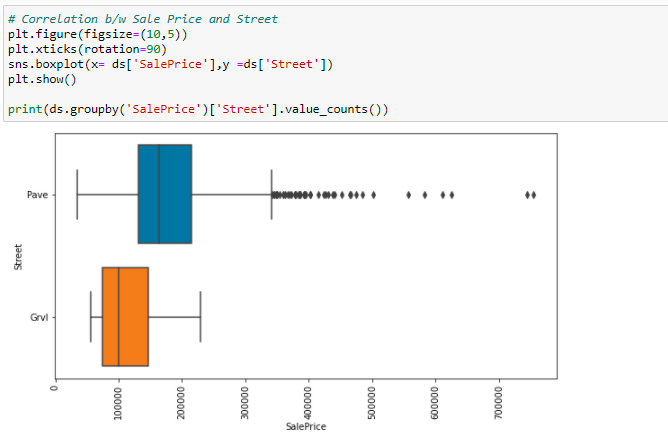
****

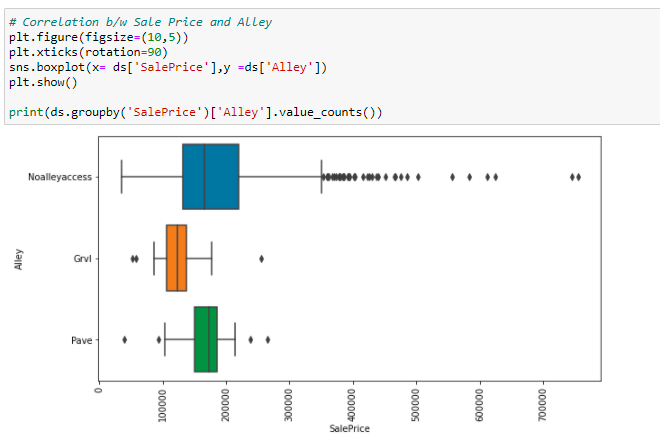


**Bivariate Analysis**

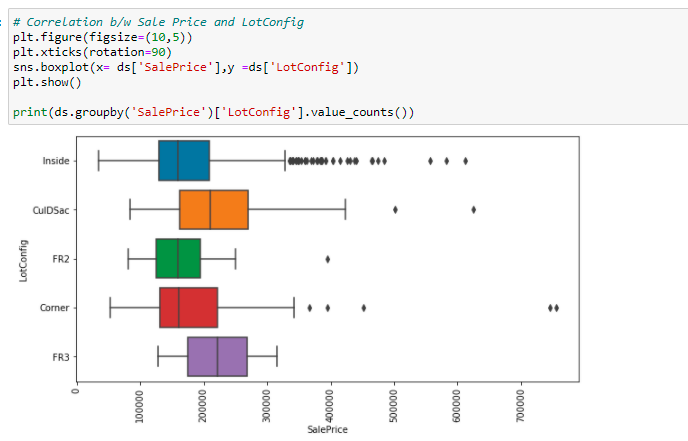
****

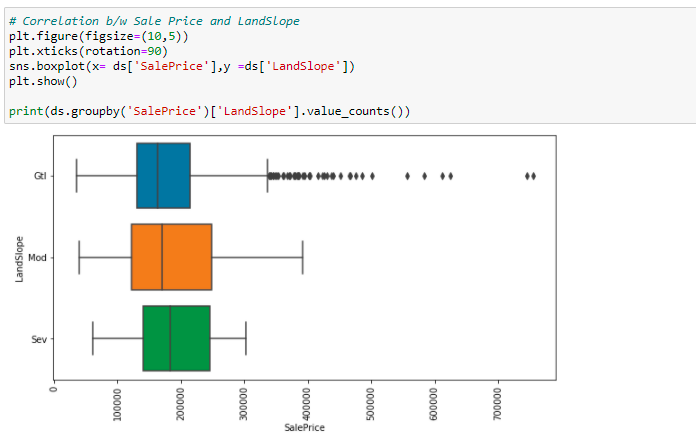
* Interpretation of the Results

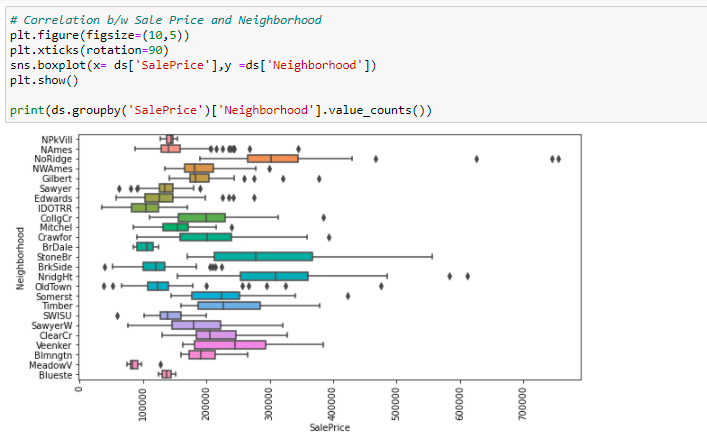




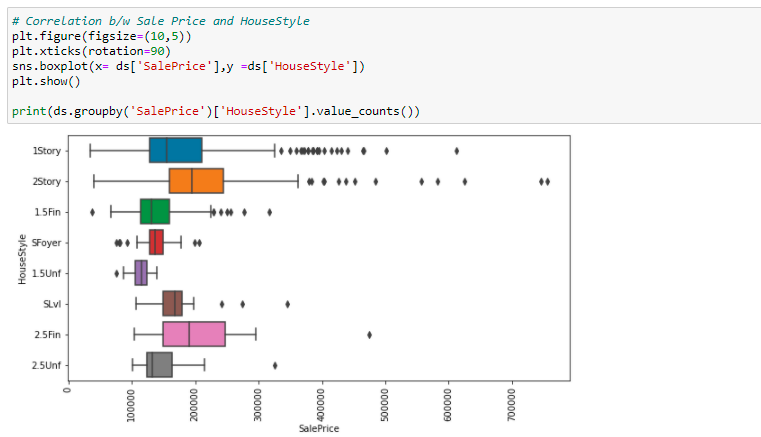


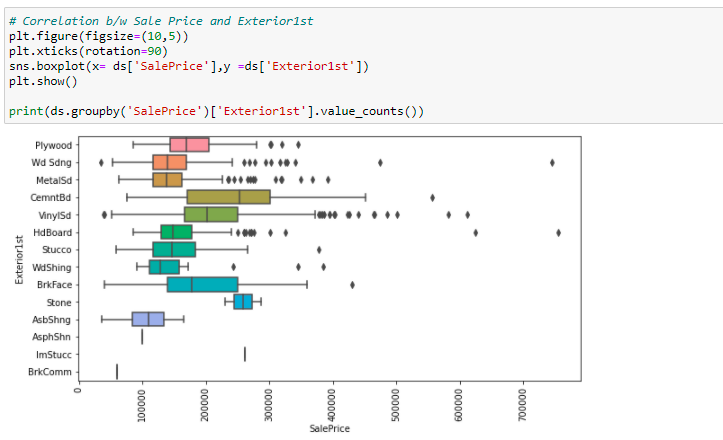




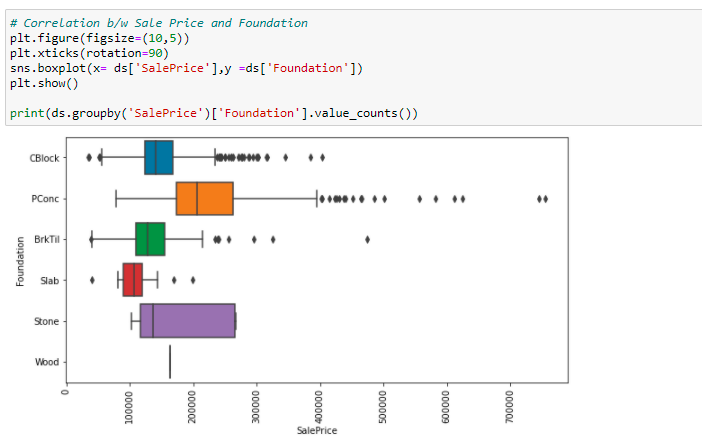


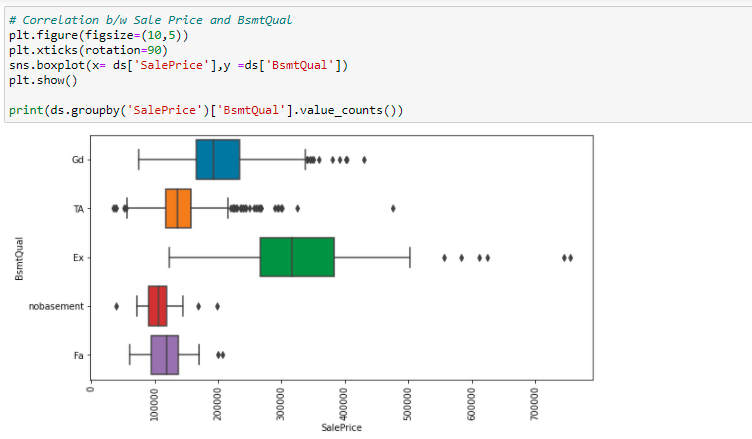


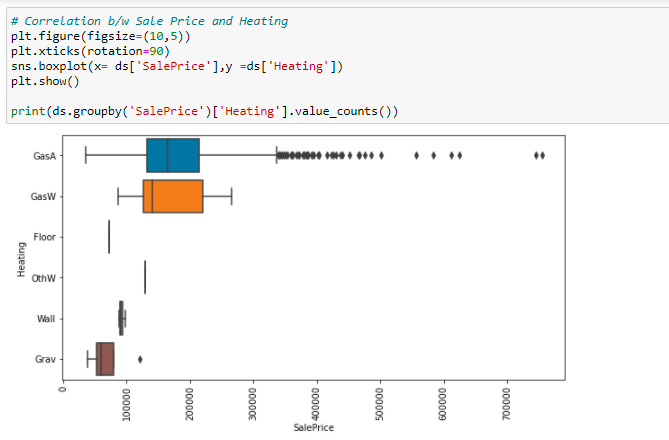


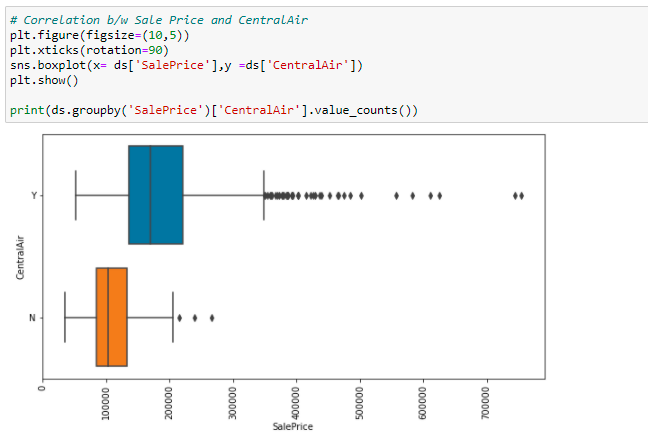


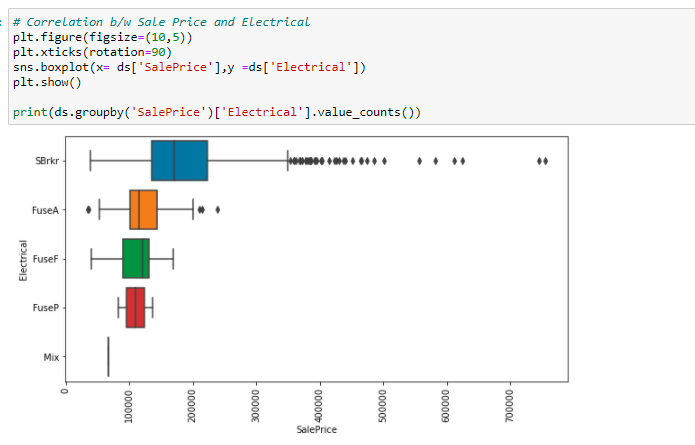












* **Interpretation of the Results**

After performing exploratory data analysis, the following inferences can be drawn:

The numerical data was not symmetrical and needed cleaning and pre-processing

There were a few null values in the dataset which were filled using mean and median and assigning different class to categorical Nan values.

Out of 80 features we have built a model using only 55 features which were most relevant to the target.

**CONCLUSION**

Real estate is a rapidly growing business. Each year, more and more people are buying houses. People take into consideration several features before buying a house.

House price prediction predicts house pricing based on different features. Real estate prices vary due to a wide variety of attributes. The machine learning-based model is a substantial and feasible way to forecast real estate prices, and can provide relatively competitive and satisfactory results.

In this project, 55 features are used to predict the sale price. Many valuation techniques like Mean Squared Error, Root Mean Square Error, mean absolute Error and R-Squared Score are used on different machine learning algorithms, which are Linear Regression, KNN Regression, Decision Trees Regression, and Random Forest Regression. After comparing all the algorithms, it is concluded that Random Forest Regression gives the best result. This helps the buyers to predict the price of the housing more accurately.

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Features that greatly contribute to house price predictions were locality (neighbourhood), above grade living area, parking capacity of the properties, Basement area and area above grade, facilities like electricity, water, gas septic tanks, fire place, number of full and half baths, the year of construction, Overall quality of the house including exterior and material quality.

Few features like MS Zone, Miscellaneous features, land slope, month in which property was sold, and etc. did not make any significant impact on the prediction of sale prices.

Thus, Machine learning models, including linear regression, KNN regression, decision tree regression and random forest regression, were used in this investigation to forecast house prices.

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

There is a huge pressure on real estate industry to unlock the potential of data science and incorporate machine learning evidences-based approaches in their work flows. Property valuation is an imprecise science. Individual appraisers and valuers bring their own experience, metrics and skills to a job. Consistency is difficult with UK and Australian-based studies suggesting valuations between two professionals can differ by up to 40%.

Perhaps a well-trained machine could perform this task in place of a human, with greater consistency and accuracy. Thanks to machine learning and data analytics, real estate professionals and investors are now able to make more accurate property assessments than ever before.

In the near future, data science will have an important role to play, as it will be able to not only improve a business strategy but also improve the way and quality of our lives. Data science which works in tandem with Artificial Intelligence (AI) will be able to analyse behaviour, interests and preferences in order to propose the ideal apartment for each client. This will mean that clients interested in a property will be able to visit it on their smartphone, projecting themselves into what it would be like to live there, by eliminating a wall or changing the colours for example.

A successful data-driven approach can yield powerful insights. In one example, an application combining a large database of traditional and non-traditional data was used to forecast the three-year rent per square foot for multifamily buildings in Seattle. These machine-learning models predicted rents with an accuracy rate that exceeded 90 per cent.

Adopting data analytics and AI into existing processes can be particularly valuable Technologies such as AI and machine learning can use data to improve efficiency by identifying patterns and opportunities, predicting future scenarios, and even automating certain tasks

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

To date, most developers and investors continue to make heuristic, or instinct-based, decisions rather than informed ones. Many are unaware of the wide variety of datasets, lack the analytical capabilities to generate insights, and/or are resistant to change. Some may be unsure where to start; others may not know which new skills and capabilities should be added to begin. Here are some of the key barriers:

Lack of awareness about new datasets and analytical techniques

For future work, diverse data types such as comments of real estate attributes, prices from social media, images from Google maps, and economic indicators are possible sources added as inputs for machine learning models to improve forecasting accuracy.One of the major future scopes is adding estate database of more cities which will provide the user to explore more estates and reach an accurate decision. More factors like recession that affect the house prices shall be added. In-depth details of every property will provide ample details of a desired estate. This will help the system to run on a larger level.

* **References:**

1. Fik T J, Ling D C and Mulligan G F 2003 Modeling spatial variation in housing prices: a variable interaction approach. Real Estate Economics, 31(4), 623-46.

2. Varma A, Sarma A, Doshi S and Nair R 2018 House Price Prediction Using Machine Learning and Neural Networks. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) 1936-39. IEEE.

3. Bhagat N, Mohokar A and Mane S 2016 House price forecasting using data mining. International Journal of Computer Applications, 152(2), 23-26.

4. Amri S and Tularam G A 2012 Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. Journal of Mathematics and Statistics, 8(4), 419-434.

5. Vineeth N, Ayyappa M and Bharathi B 2018 House Price Prediction Using Machine Learning Algorithms. In International Conference on Soft Computing Systems, 425-33.

6. Alfiyatin A N, Febrita R E, Taufiq H and Mahmudy W F 2017 Modeling house price prediction using regression analysis and particle swarm optimization. International Journal of Advanced Computer Science and Applications, 8.

7. Afonso B, Melo L, Oliveira W, Sousa S and Berton L 2019 Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. In Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional 389-400. SBC.

8. McKinsey & Company 2018|Article Getting ahead of the market: How big data is transforming real estate