

HOUSE PRICE PREDICTION

Submitted by:

Amritesh Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shubam Yadav as well as Flip Robo. Technologies who gave me the opportunity to do this project on HOUSE PRICE PREDICTION & also helping me to gain in-depth knowledge of Machine Learning and DataScience to derive insights for organizational goals or meet business needs.

Also, I have utilized a few external resources that helped me to complete this project. All the external resources that were used in creating this project are listed below:

<https://stackoverflow.com/questions>

<https://medium.com/>

<https://www.kaggle.com/>

<https://www.geeksforgeeks.org/>

<https://www.codegrepper.com/>

<https://www.analyticsvidhya.com/>

<https://towardsdatascience.com/>

<https://github.com/>

**INTRODUCTION**

Business Problem Framing

**Problem Overview**

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The 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

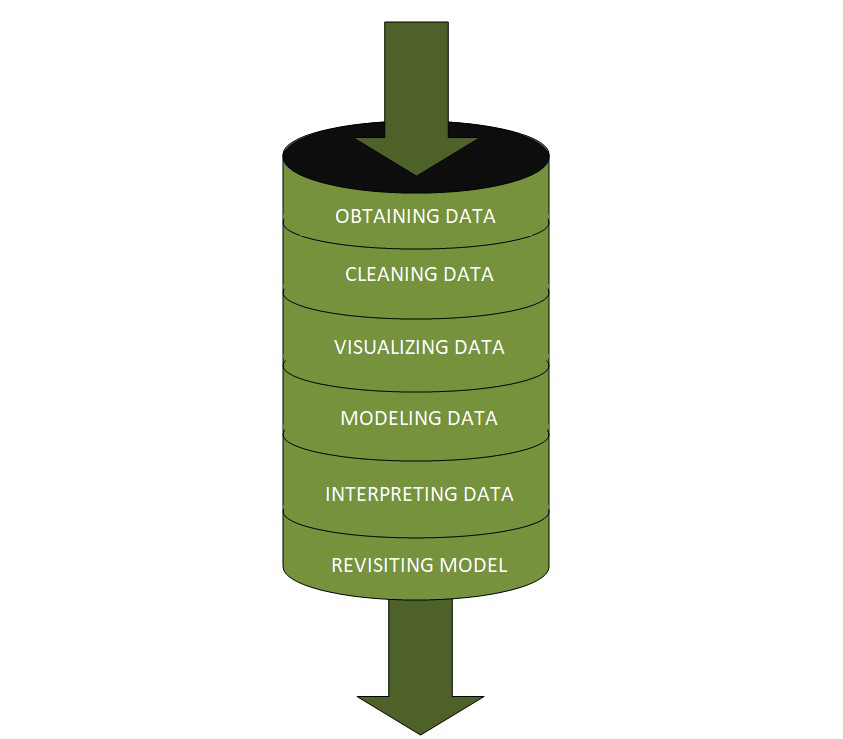
**MACHINE LEARNING AND DATA SCIENCE FOR BUSINESS:**

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn from experience, make predictions and gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data science will increase, requires to assist in the identification of the most relevant business questions and subsequently the data to answer them. Following are the ways Data science can add value to Business :

* Empowering management and officers to make better decision
* Directing actions based on trends—which in turn help to define goals
* Challenging the staff to adopt best practices and focus on issues that matter
* Identifying opportunities
* Decision making with quantifiable, data-driven evidence
* Testing these decisions
* Identification and refining of target audiences

**DATASCIENCE PIPELINE:**

The data science pipeline is a collection of connected tasks that aims at delivering an insightful data science product or service to the business organization. The responsibilities include collecting, cleaning, exploring, modeling, interpreting the data, and other processes of the launching of the product. This final product can be used for to achieve Business Goals.



**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

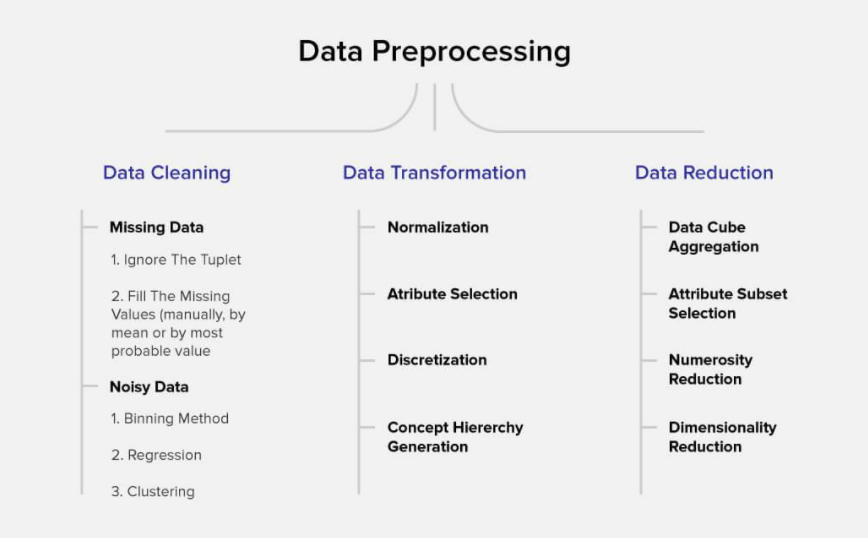
Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions

**TYPES OF EXPLORATORY DATA ANALYSIS:**

* Univariate Non-graphical
* Multivariate Non-graphical
* Univariate graphical
* Multivariate graphical

**DATA PRE-PROCESSING & FEATURE ENGINEERING:**

Preprocessing simply refers to perform series of operations to transform or change data. It is transformation applied to our data before feeding it to algorithm. When creating a machine learning project, and doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.



Data pre-processing is a very vital input to machine learning models, It is to prepare the raw data & make it suitable for efficient machine learning model. These are the methods of data preprocessing and we are going to use the required ones in our project.

**FEATURE ENGINEERING:**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model.

Feature engineering**, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.** It can produce new features for both supervised and unsupervised learning, with the goal of**simplifying and speeding up data transformations**while also**enhancing model accuracy.**

**Feature Engineering Techniques for Machine Learning**

* **Imputation**
* **Handling Outliers**
* **Log Transform**
* **One-hot encoding/Label Encoding**
* **Scaling**

**Data Transformation:**

**Label Encoding:**

**As we mentioned above in library installation,** Label Encoder is used to encode labels by assigning them numbers. It is used to encode single or multiple columns. Thus, if the feature is color with values such as [‘white’, ‘red’, ‘black’, ‘blue’]., using Label Encoder may encode color string label as [0, 1, 2, 3]

**Handling Outliers:**

The most important phase in Feature Engineering is handling outliers because it ensures that our model is trained on accurate data which leads to accurate models. An outlier may occur due to the variability in the data. It may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution). We have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Below are some of the techniques of detecting outliers

* Boxplots
* Z-score

## Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library, Let us import the package.

**SKEWNESS REMOVAL-(POWER-TRANSFORM):**

Key step prior to initiating Machine learning models, optimizing, scaling the data to provide it as a input to start the modelling.

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. The log transform is a specific example of a family of transformations known as power transforms. The power\_transform library present in the Sklearn. Pre-processing package.

**MINMAX SCALER:**

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

Before scaling we have to train test split the data.since we have to do skewness removal and scaling only on input data.

**TRAIN TEST SPLIT:**

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

Review of Literature

**ABSTRACT:**

In this House Price Prediction Project, we are going to predict the price of the houses in Australia with collection records of the sale of houses in Australia. We are doing this prediction for the US-based housing company named Surprise Housing.This company has decided to enter the Australian market to purchase houses at a price below their actual values and flip them at a higher price.So our detailed analysis, Machine Learning Model predictions done can be used for specific Business Requirements, Challenges and Improvements of the d Surprise Housing Company.

Motivation for the Problem Undertaken

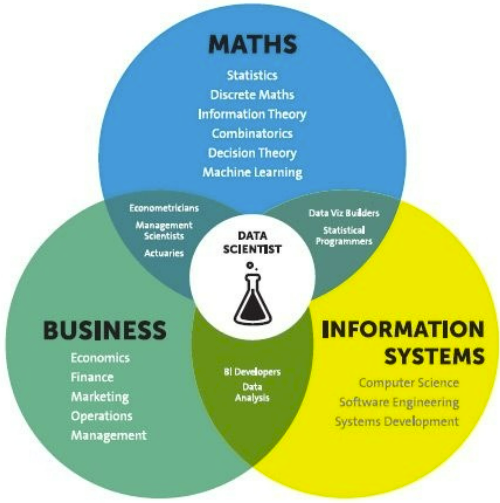
**Business Goal:**

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. They can 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.

**Analytical Problem Framing**

Mathematical/ Statistical /Analytical Modeling of the Problem

Mathematics, Statistics and Analytics are three of the most important concepts of Data Science. Data Science revolves around these three fields and draws their concepts to operate on the data.we will explore its practical usages in this field.  So let’s first explore how much these three are required for data science.



**Mathematical Modelling**

Mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Machine Learning is a field that focuses on computers having the ability to learn/operate without being programmed to do so.

Mathematics is playing an essential role in the latest technologies like Machine Learning, Artificial Intelligence, Data Science and Deep Learning, etc., It is because every algorithm built in the latest technologies has a mathematical function behind it and aid in identifying patterns.

The understanding of various notions of Statistics and Probability Theory are key for the implementation of such algorithms in data science. Notions include: Regression, Maximum Likelihood Estimation, the understanding of distributions (Binomial, Bernoulli, Gaussian (Normal)) and Bayes’ Theorem.

The main reason for a greater significance of mathematics is because of its various concepts like: –

· Linear Algebra

· Probability

· Calculus

· Statistics

**Linear Algebra & Calculus**

Deep learning requires us to understand linear algebra & calculus, to understand how it works, for example forward propagation, backward propagation, parameters setting etc. For linear algebra, there are matrix operations (plus, minus, times, divide), scalar product, dot product, eigen-vectors and eigenvalues.

It is a branch of Mathematics for studying systems of equations. it can be one, two, and multi-dimensional equations. it helps us to solve numerical data or relations between two or more variables by establishing relations or equations between them. for example,

here' one basic algebraic equation:

y = a + bx + cx2

linear-algebra has a wide range of applications such as statics and matrices calculations, linear regression equations, descriptive statistics, graphic image vectors, Fourier series, graphs, and network establishment.

machine-learning algorithms like linear regression, logistic regression uses linear algebra to solve our target variables with given inputs/attributes or feature vectors given in the data set.

**Calculus**

 Calculus is used essentially in optimization techniques. Using calculus, you can carry out mathematical modeling of artificial neural networks and also increase their accuracy and performance. For calculus, the data scientist need to understand various differentiation (to second-order derivative), integration, partial differentiation.

**Differential Calculus**

  Differential Calculus studies the rate at which the quantities change. Derivates are most widely used for finding the maxima and minima of the functions. Derivates are used in optimization techniques where we have to find the minima in order to minimize the error function.

**Integral Calculus**

It is the mathematical study of the accumulation of quantities and for finding the area under the curve. Integrals are further divided into definite integrals and indefinite integrals.

**Probability**

The probability theory is very much helpful for making the prediction and Estimation.With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

Probability is a very important mathematical concept for data science, used in validating hypothesis, bayes theorem and interpreting outputs in machine learning.

Bases on these we try to estimate various events, and the likelihood of the outcome. sometimes we wat graphical representations of probable outcomes which we call probability density functions or density curves.

Concepts of probability help us estimate expected value from given variables, to solve confusion matrix in classification algorithms, information entropy, evidence of particular attributes in naive Bayes classification, and even in statistics for hypothesis testings.

**Statistics**

A statistical model is a mathematical representation (or mathematical model) of observed data. When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.

So the areas in statistics are simple statistics like measurement of centrality, distributions and different probability distributions (Weibull, Poisson etc), Baye’s Theorem

statistics is divided into two –

* Descriptive Statistics
* Inferential Statistics

#### **Descriptive Statistics**

Descriptive Statistics or summary statistics is used for describing the data. It deals with the quantitative summarization of data. This summarization is performed through graphs or numerical representations.

### Descriptive Statistics:

1) Mean, Median, Mode

2) IQR, percentiles

3) Std deviation and Variance

4) Normal Distribution

5) Z-statistics and T-statistics

6) correlation and linear regression

**Inferential Statistics**

It is the procedure of inferring or concluding from the data. Through inferential statistics, we make a conclusion about the larger population by running several tests and deductions from the smaller sample.

### Inferential Statistics:

1) Sampling distributions

2) confidence interval

3) chi-square test

4) Advanced regression

5) ANOVA

The mathematical concepts noted above are key in understanding/implementing the following Machine Learning techniques.

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### **Regression Models**

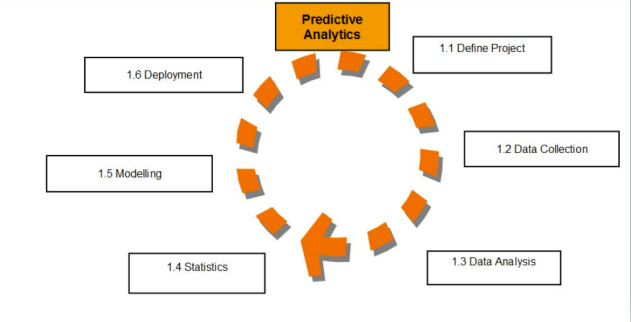
Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

### **Classification Models**

**Classification** is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

**Analytical Models:**

An analytical model estimates or classifies data values by essentially drawing a line through data points. When applied to new data or records, a model can predict outcomes based on historical patterns.



.  An analytical model is quantitative in nature, and used to answer a specific question or make a specific design decision. Different analytical models are used to address different aspects of the system, such as its performance, reliability, or mass properties.Data analysis comes with the fundamental types of data analytics encounter in data science: Descriptive, Diagnostic, Predictive, and Prescriptive.

* Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.
* Descriptive analysis is often used when reviewing any past or present data. This is because raw data is difficult to consume and interpret, while the metrics offered by descriptive analysis are much more focused.
* The example of descriptive statistics or analytics is to calculate the mean, median mode, standard deviation, and similar kinds of statistical calculation on finance or sales data.
* Diagnostic analytics takes it a step further to uncover the reasoning behind certain results. Diagnostic analytics is usually performed using such techniques as data discovery, drill-down, data mining, and different type of bivariant data analysis like  correlations.etc.,
* Predictive Analytics is a **statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors**. Predictive Analytics can take both past and current data and offer predictions of what could happen in the future.
* Predictive models typically utilize variability in data to make the correct prediction and more variability of ingredient data that shows the relationship with what is possible to predict that united together into a prediction or valid score.
* Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, algorithms, and machine learning to make predictions and then suggests decision options to take advantage of the predictions. Prescriptive means (optimization and simulation).

Data Sources and their formats

**Technical Requirements**:

• Data contains 1460 entries each having 81 variables.

• Data contains Null values. We treated them using the domain knowledge and our own understanding.

• Extensive EDA has been performed to gain relationships of important variable and price.

• Data contains numerical as well as categorical variable. We handled them accordingly.

• We built Machine Learning models, applied regularization and determined the optimal values of Hyper Parameters.

• We found important features which affect the price positively or negatively.

• Two datasets are provided (test.csv, train.csv). We had done training on train.csv dataset and prediction on test.csv file.

The datasets are enclosed in notebook file

The dataset is provided to us by FlipRobo Technologies.And the dataset is in .csv file format.

**Data Description:**

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

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

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 postive 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

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

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

Grav Gravity 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

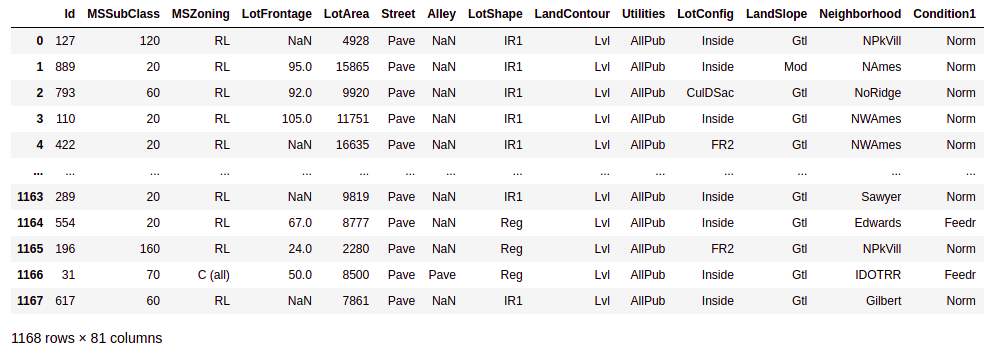
AdjLand Adjoining Land Purchase

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 ACQUISITION**



**FEATURE DESCRIPTION:**

Here Id,MSSubClass,LotFrontage,LotArea,Neighborhood,Condition1,Condition2,HouseStyle,OverallQua,OverallCond,YearBuilt,YearRemodAdd,Exterior1st,Exterior2nd,MasVnrArea,BsmtFinSF1,BsmtFinType2,BsmtFinSF2,BsmtUnfSF,TotalBsmtSF,1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea,BedroomAbvGr,TotRmsAbvGrd,Functional,GarageType,GarageYrBlt,GarageArea,WoodDeckSF,OpenPorchSF,EnclosedPorch,3SsnPorch,ScreenPorch,PoolArea,MiscVal,MoSold,SaleType are all the columns which has categorical ordinal data type.

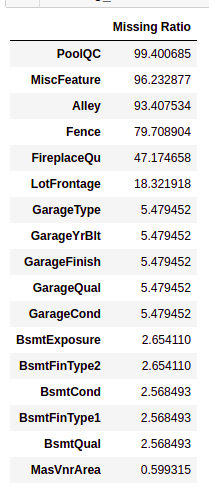
MSZoning,Street,Alley,LotShape,LandContour,Utilities,LotConfig,LandSlope,BldgType,RoofStyle,RoofMatl,MasVnrType,ExterQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,Heating,HeatingQC,CentralAir,Electrical,BsmtFullBath,BsmtHalfBath, FullBath,HalfBath,KitchenAbvGr,KitchenQual,Fireplaces,FireplaceQu,GarageFinish,GarageCars,GarageQual,GarageCond,PavedDrive, PoolQC,Fence,MiscFeature,YrSold,SaleCondition are all the columns which has categorical nominal data type.

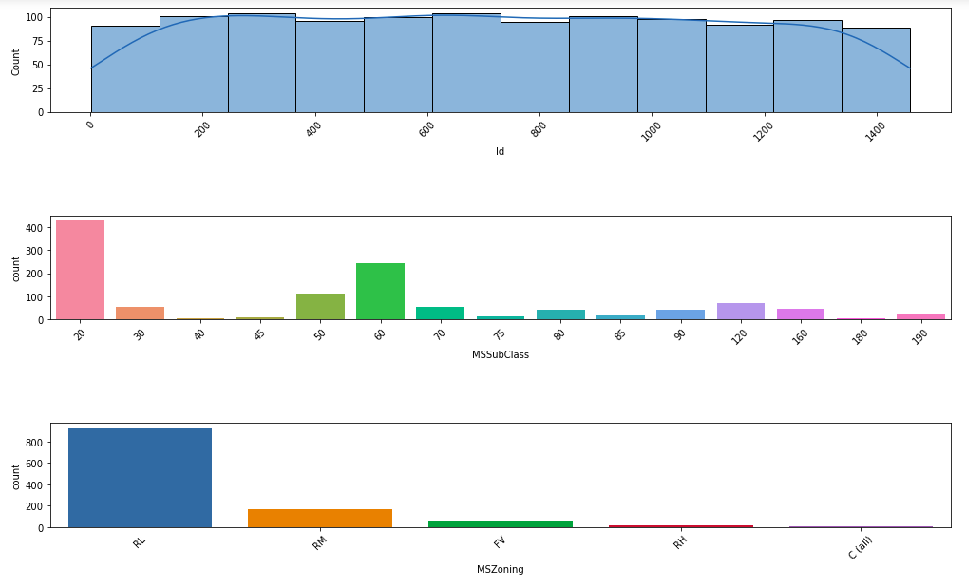
Here Our Target column is the SalePrice column which is we are going to predict.Hence our problem is the Regression problem.

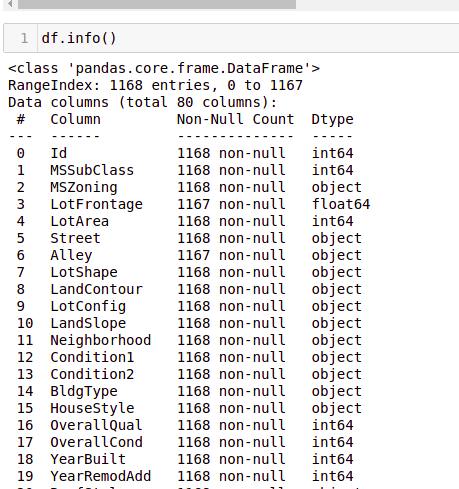


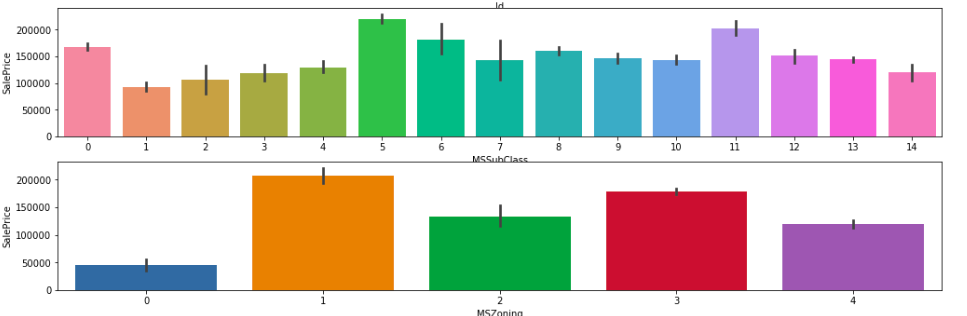
Data Preprocessing Done

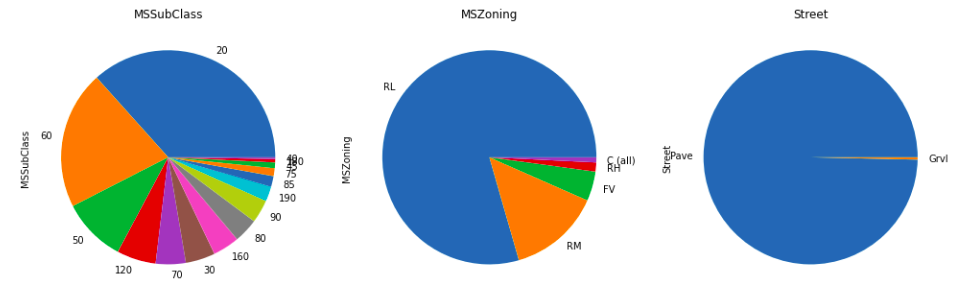


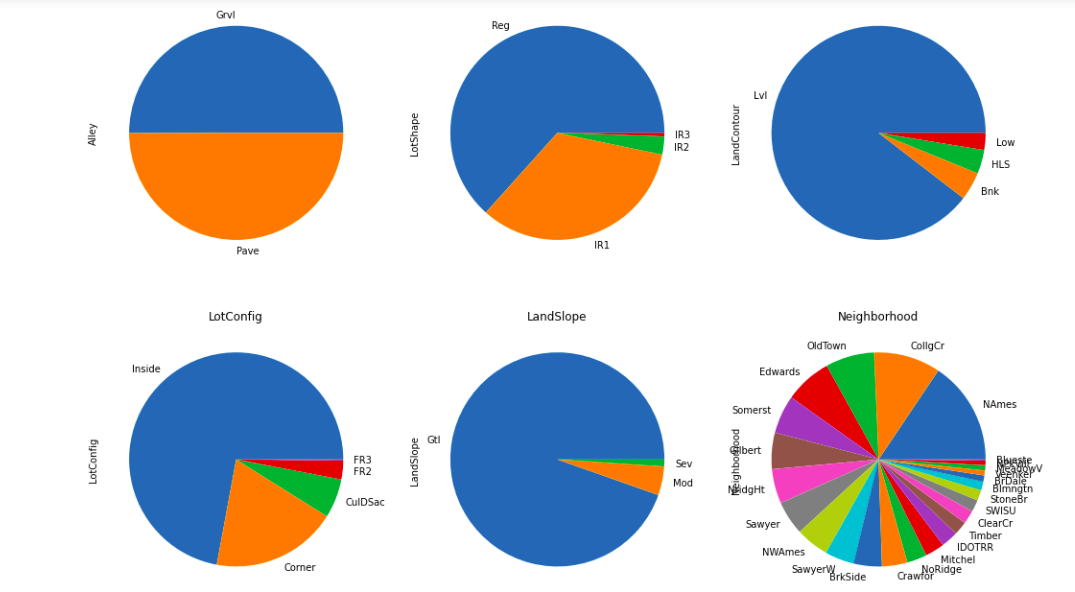


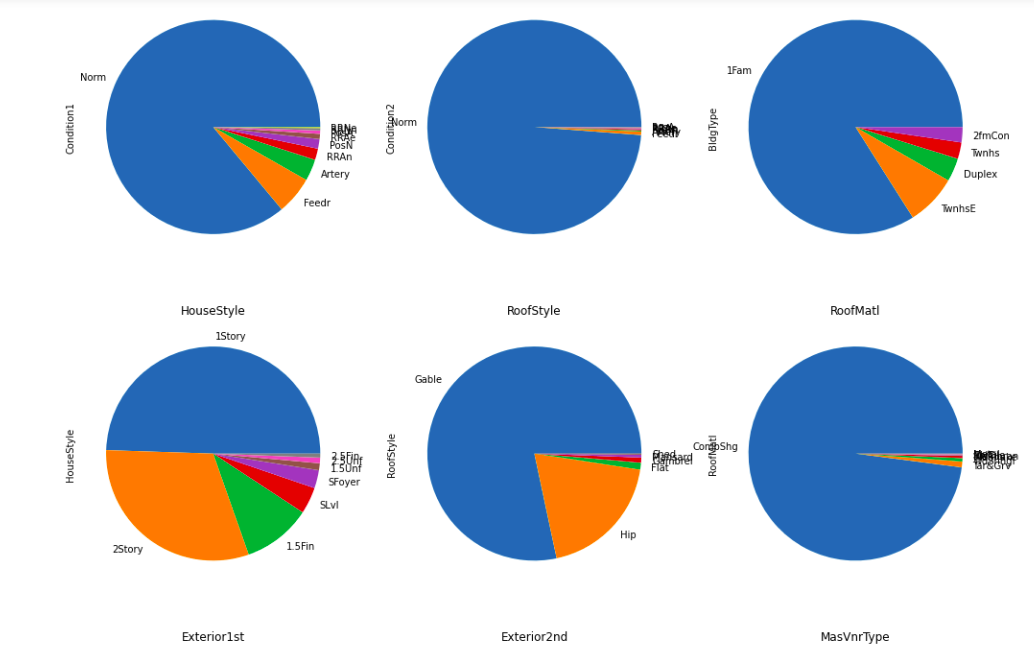


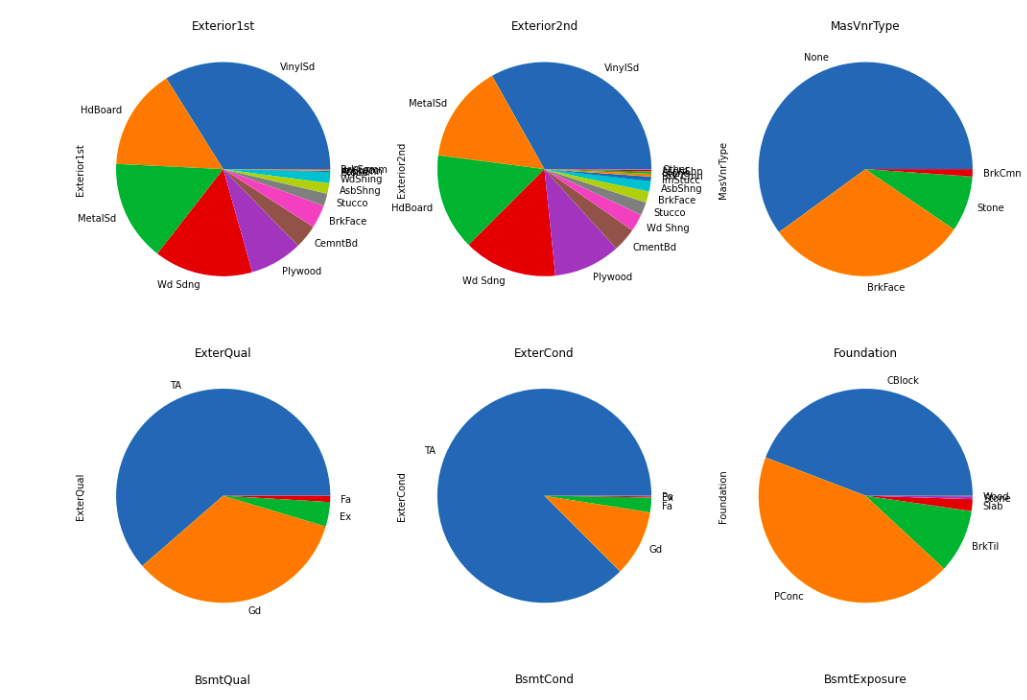


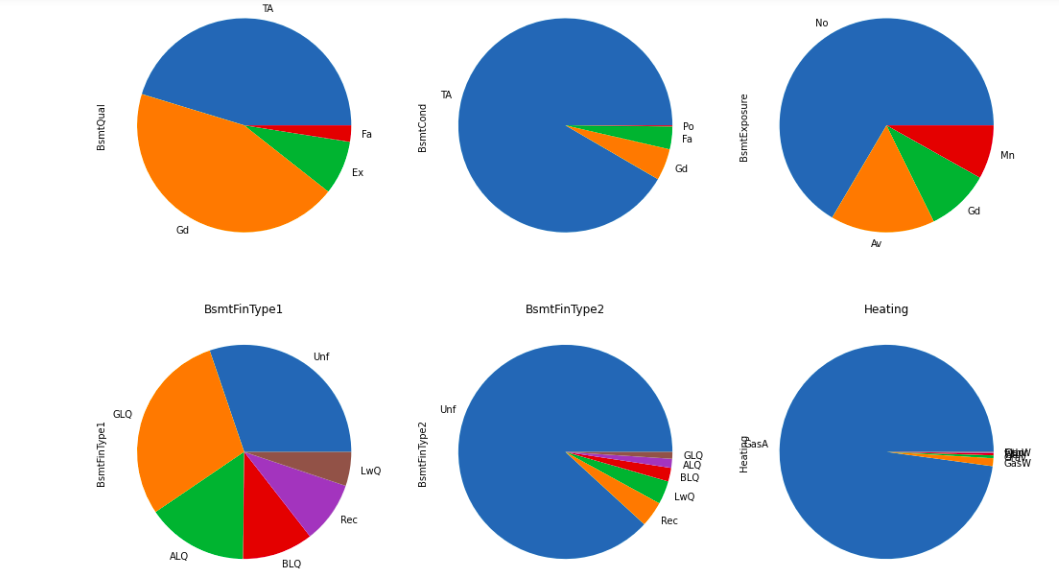
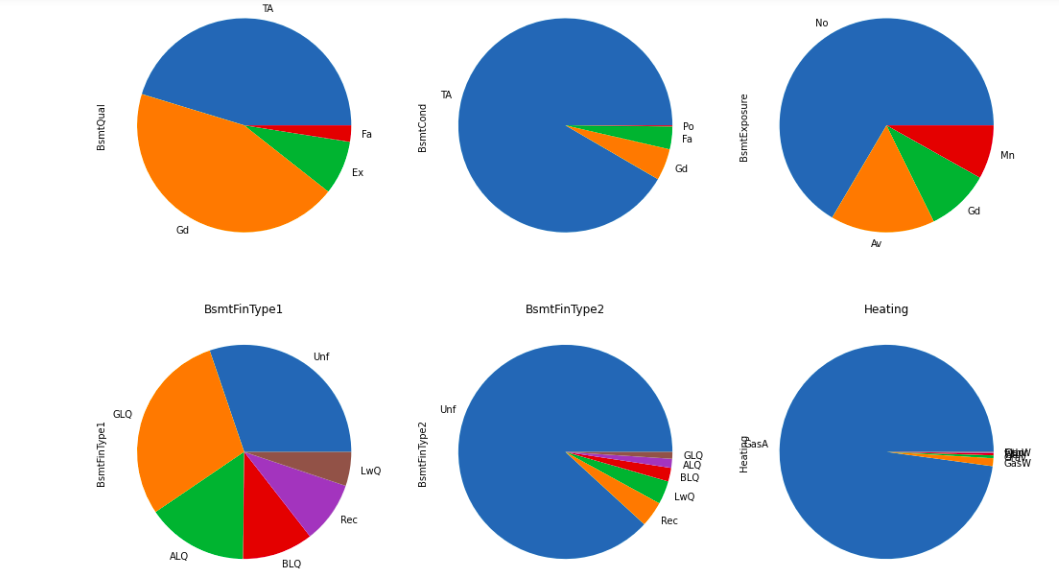












Relationships

1. If y represents the **dependent variable and x** the independent variable, this relationship is described as the regression of y on x. The relationship can be represented by a simple equation called the regression equation.
2. Linear regression **attempts to model the relationship between two variables by fitting a linear equation to observed data**. ... A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable.
3. The output variable (also called dependent variable, or regressand) is assumed to be a linear function of the input variables (also called **independent variables**, or regressors) and of an unobservable error term that adds noise to the linear relationship between inputs and outputs.
4. To do this a line is created that best fits a set of data pairs. The value of y is derived through the value of x, reflects their correlation
5. The RESPONSE of a system is linear **when the output is directly proportional to the input**, that is, any change in the input produces a proportional change in the output. When plotted on a graph, a straight line results.
6. **Correlation** describes the relationship between two sets of data.
7. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.
8. In multiple regression, there are multiple independent variables that enable us to estimate the dependent variable y.

Multiple regression equation is derived by:

Y = a + b1\*1 + b2\*2 + b3\*3……………. bk\*k

Here, y is an independent variables whereas  b1, b2 and bk

1. Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data.
2. In Multiple Linear Regression, a Residual is the Difference Between Estimated Dependent Variables and Actual Dependent Variables. Multiple linear regression assumes that the remaining variables’ error is similar at each point of the linear model. This is known as homoscedasticity. When the data analysis is done, the standard residuals against the predicted values are plotted to determine if the points are properly distributed across independent variables’ values. Larger residuals indicate that the regression line is a poor fit for the data, i.e. the actual data points do not fall close to the regression line. Smaller residuals indicate that the regression line fits the data better, i.e. the actual data points fall close to the regression line.

The [coefficient of determination](https://www.investopedia.com/terms/c/coefficient-of-determination.asp) (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

Multiple linear regression (MLR) is used to determine a mathematical relationship among several random variables.

In a multiple linear regression, the model calculates the line of best fit that minimizes the variances of each of the variables included as it relates to the dependent variable. Because it fits a line, it is a linear model.

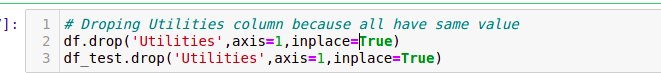
The relationship can also be non-linear, and the dependent and independent variables will not follow a straight line. Linear and non-linear regression are used to track a response using two or more variables.

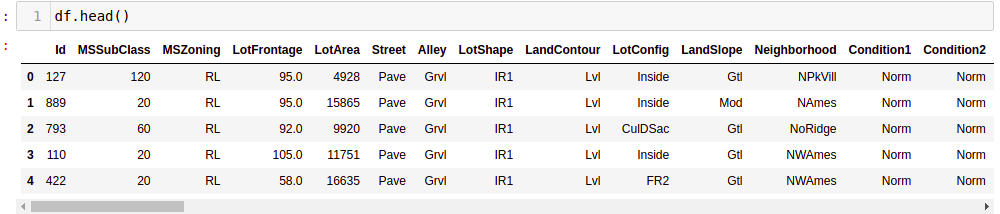
Assuming a linear relation in population, mean of Y for given X equals α+βX i.e. the "population regression line". If Y = a + bX is the estimated line, then the fitted Ŷi = a + bXi is called the fitted (or predicted) value, and Yi Ŷi is called the residual.

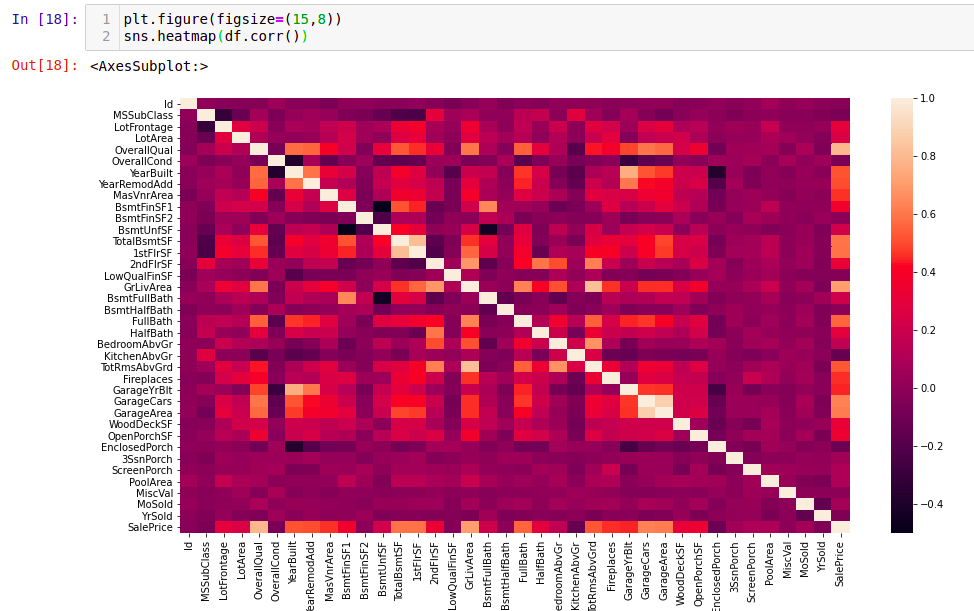
The regression coefficient of y on x is represented by b yx and x on y as b xy. Both of the regression coefficients must have the same sign. If b yx is positive, bxy will also be positive and it is true for vice versa. If one regression coefficient is greater than unity, then others will be lesser than unity.

For a bivariate data (Xi, Yi), the relationship may be Y depends on X or X depends on Y. If Y depends on X then the regression line is Y on X. Y is dependent variable and X is independent variable. If X depends on Y, then regression line is X on Y and X is dependent variable and Y is independent variable.

State the set of assumptions (if any) related to the problem under consideration







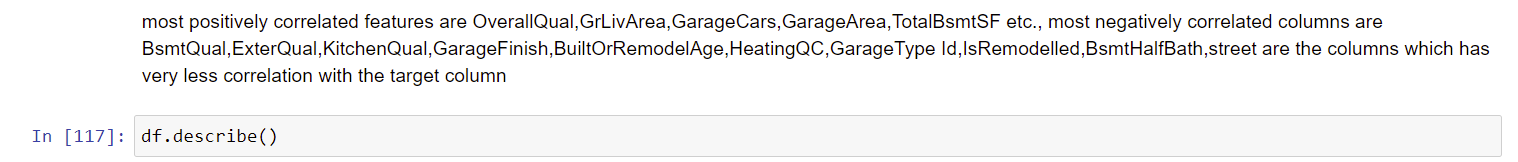
**Correlation with Heatmap:**

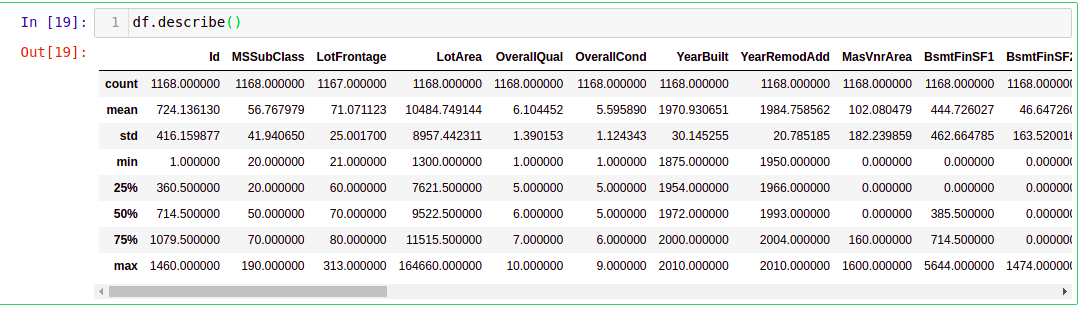
The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect [negative correlation](https://www.investopedia.com/terms/n/negative-correlation.asp), while a correlation of 1.0 shows a perfect [positive correlation](https://www.investopedia.com/terms/p/positive-correlation.asp). A correlation of 0.0 shows no linear relationship between the movement of the two variables.Correlation statistics can be used in finance and investing. Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure**is used. Such matrix is called as **correlation matrix.**

Correlation heatmap is graphical representation of **correlation matrix**representing correlation between different variables.

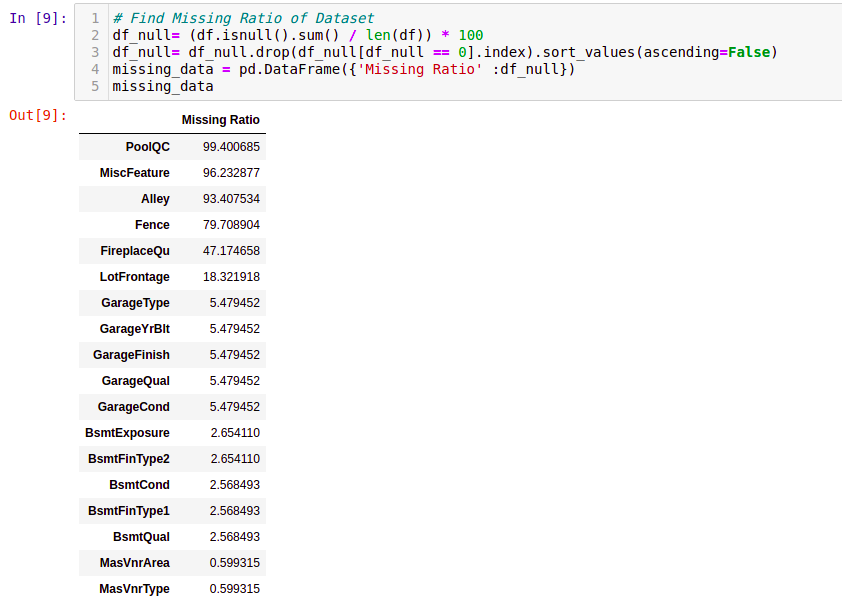
**For to do feature selection and make feature ready for the model building.we check correlation of variables using heatmap.And describe method for the census data set.**





**Correlation model:**

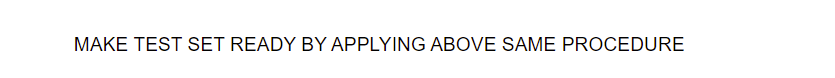
Graph depicts clearly the positive and negative correlation of each variables with target column, justifies the outcome outlined in Multivariate analysis, that higher the education higher the gain & vice-versa



Drop ID column since it makes not much correlation with the target column.

Drop street column since it has high vif

Drop GarageCond column since it has high vif value



Hardware and Software Requirements and Tools Used

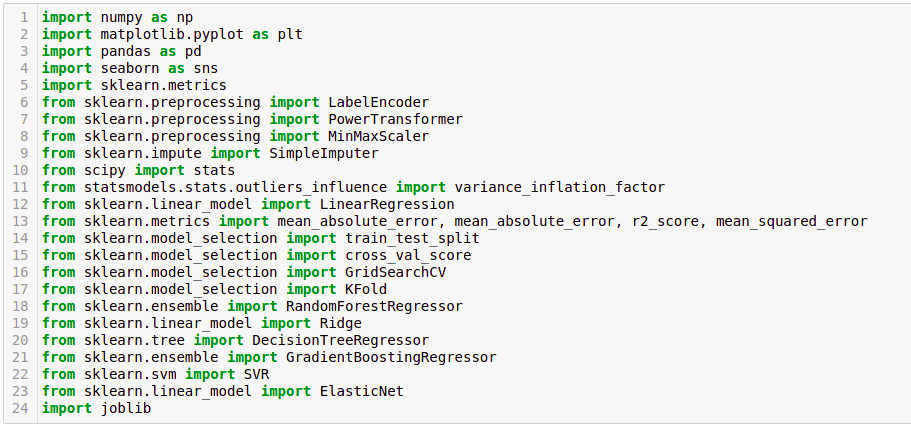
**HARDWARE & Software Tools, Libraries and Packages Used:**

Hardware :Intel i7,RAM 16GB used.

Software: Jupyter Notebook (Anaconda 3)

Language: Python

Libraries:

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Sklean
6. Scipy
7. Statsmodels
8. Pip-Package install Manager

|  |  |  |
| --- | --- | --- |
| **Category** | **Tool** | **Function** |
| Data loading and analysis | Import pandas as pd | Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. Pandas is built on top of numpy. So, numpy gets some superpower with pandas. It offers data structures and operations for manipulating numerical tables and time series. |
| Import numpy as np | NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.It has Quantile method too for removing outliers. It is the fundamental package for scientific computing with Python |
| Data visualization | Import matplotlib.pyplot as plt | Matplotlib is a plotting library used for data visualization. |
| Import seaborn as sns | Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib |
| Scikit Learn Preprocessing Libraries | Sklearn.preprocessing | Package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.Has power transformer to remove skewness. In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate.  It has MinMaxScaler to scale the data. |
| Sklearn.preprocessing import LabelEncoder | Label Encoding in Python can be implemented using the Sklearn Library. Sklearn furnishes a very effective method for encoding the categories of categorical features into numeric values. Label encoder encodes labels with credit between 0 and n-1 classes where n is the number of diverse labels. |
| Import statistics | Import statsmodels.api as sm | From scipy import stats This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. This library provides a number of common functions and types useful in statistics. It focus on high performance, numerical robustness, and use of good algorithms |

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library and we can import this library.

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

Used to check the error and score of the model.The error tells how much is the difference between actual and predicted result.The score tells the accuracy of the model.

Grid search is used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyper-parameters.

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Various Models of Machine learning tools can be used in to do various iterations & select the optimal model for problem solving and predictions.

We can import the necessary evaluation metrics library and other libraries from sklearn.linear\_model

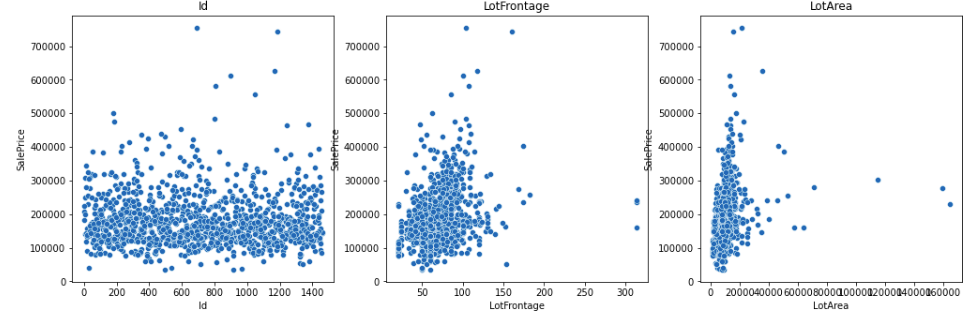
And sklearn.metrics.

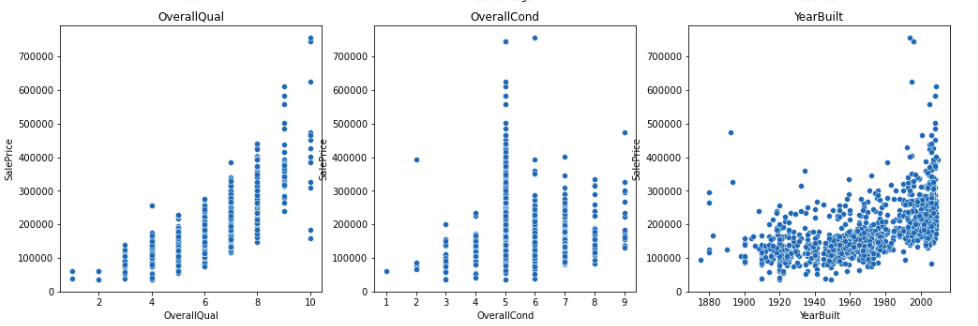
 Ensemble learning is a technique in machine learning which takes the help of several base models and combines their output to produce an optimized model. This type of machine learning algorithm helps in improving the overall performance of the model.

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

'LotFrontage','LotArea','TotalBsmtSF','GrLivArea','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3SsnPorch','ScreenPorch','PoolArea','SalePrice' .These are the columns having outliers which is shown in box plot.It seems all area columns which is continuous having outliers.





'LotFrontage','LotArea','TotalBsmtSF','GrLivArea','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3SsnPorch','ScreenPorch',

'PoolArea','SalePrice'.These columns has skewness too because of outliers.



Testing of Identified Approaches (Algorithms)

These are all the Algorithms used for Model Building and Prediction.We did Hyper Parameter Tuning with these algorithms using the GridSearchCV.

**Ridge Regressor**

**LinearRegression**

**RandomForestRegressor**

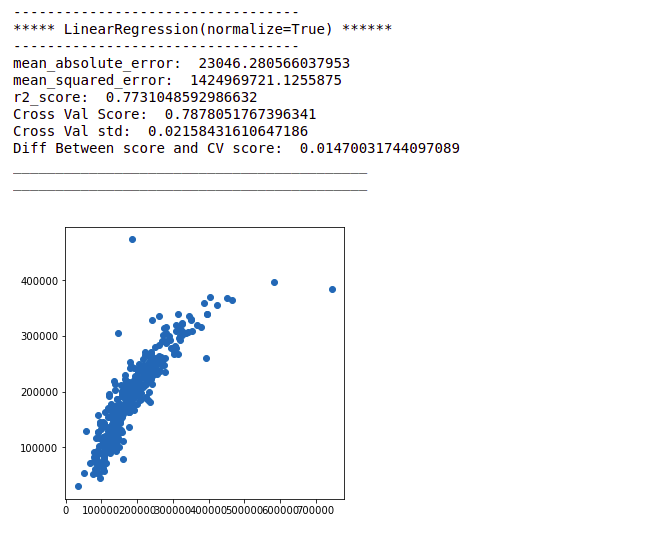
**ElasticNet Regression**

**These algorithms has been used for both Training and Testing purpose and got evaluated with r2 score.And also the predicted result got Evaluated with Key Metrics.**

Run and Evaluate selected models

Linear Regression:

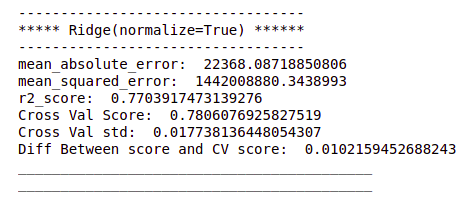
* Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
* In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (y) and an independent variable (x).

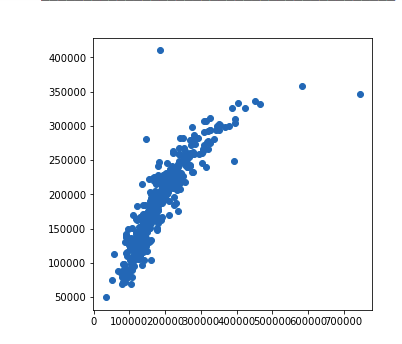


Ridge Regression:

* The Ridge regression is a technique which is specialized to analyze multiple regression data which is multicollinearity in nature.
* Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

The biggest benefit of ridge regression is its ability to produce a lower test mean squared error (MSE) compared to least squares regression when multicollinearity is present.





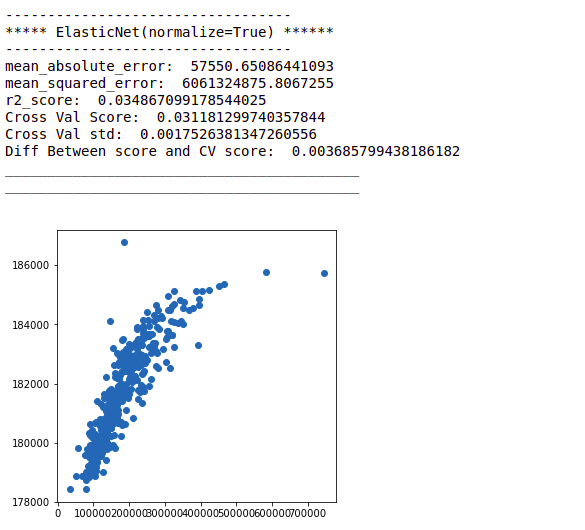
ELASTIC NET

**Elastic net** is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions.

In this tutorial, you will discover how to develop Elastic Net regularized regression in Python.

After completing this tutorial, you will know:

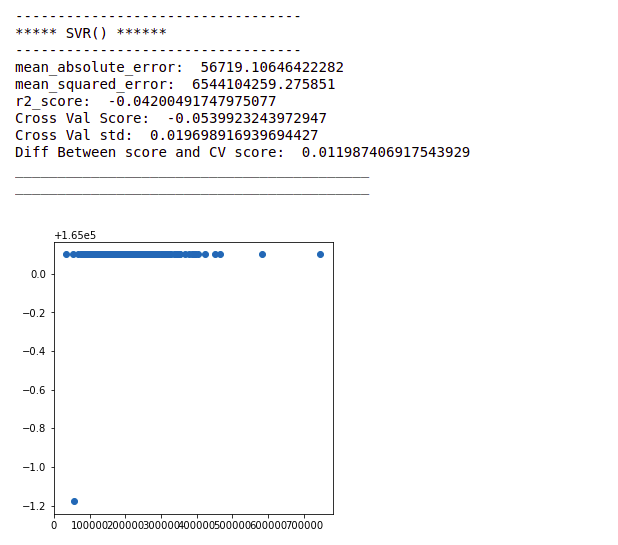
* Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.
* How to evaluate an Elastic Net model and use a final model to make predictions for new data.
* How to configure the Elastic Net model for a new dataset via grid search and automatically.



SUPPORT VECTOR REGRESSOR

Support Vector Machine is a discriminative algorithm that tries to find the optimal hyperplane that distinctly classifies the data points in N-dimensional space(N - the number of features). In a two-dimensional space, a hyperplane is a line that optimally divides the data points into two different classes. In a higher-dimensional space, the hyperplane would have a different shape rather than a line.

 Support Vector Regression (SVR) is quite different than other Regression models. It uses the Support Vector Machine (SVM, a classification algorithm) algorithm to predict a continuous variable.



These are all the algorithms used and it described here with the snapshot of their code and the results observed over different evaluation metrics are also mentioned.

The evaluation metrics used here is r2 score.

R2 score:

It is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).” Another definition is “(total variance explained by model) / total variance.” So if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all.

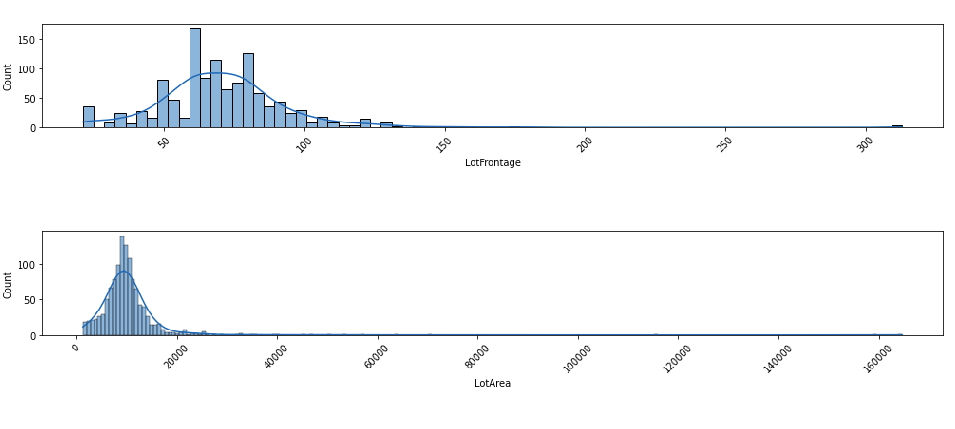
Key Metrics for success in solving problem under consideration

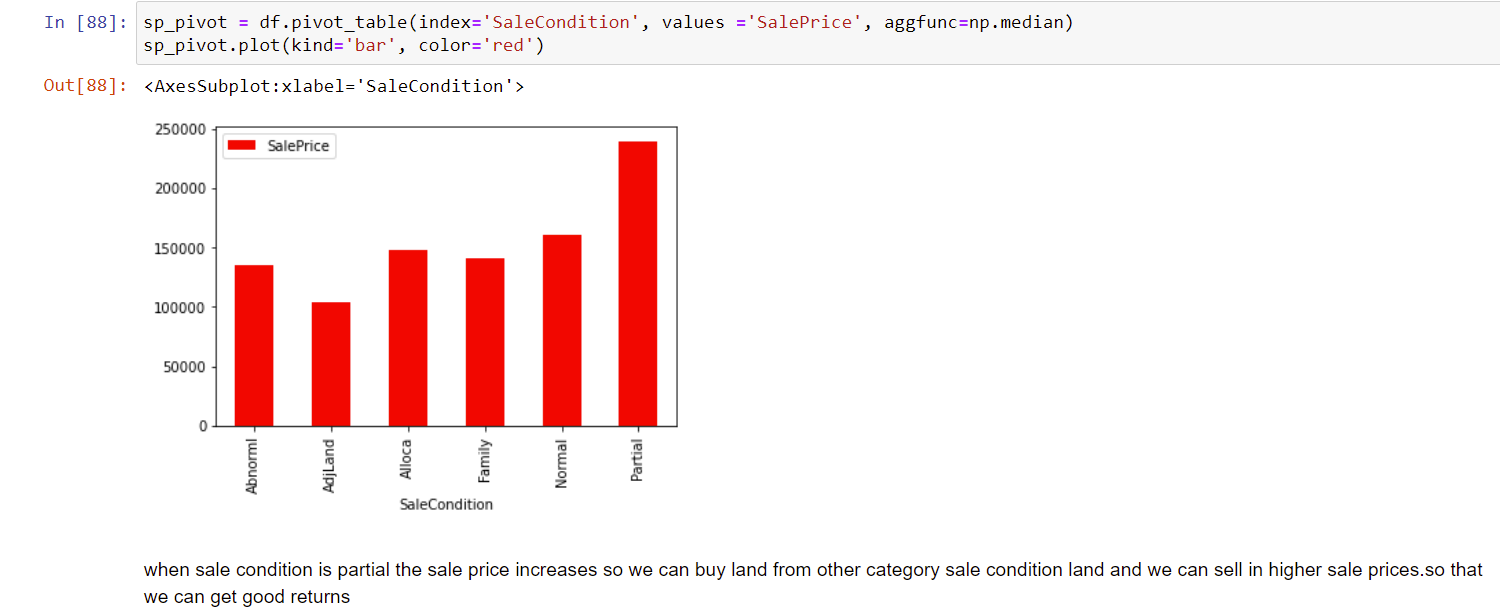
Error addresses exactly this and summarizes on average how close predictions were to their expected values. There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE) and Mean Absolute Error(MAE).

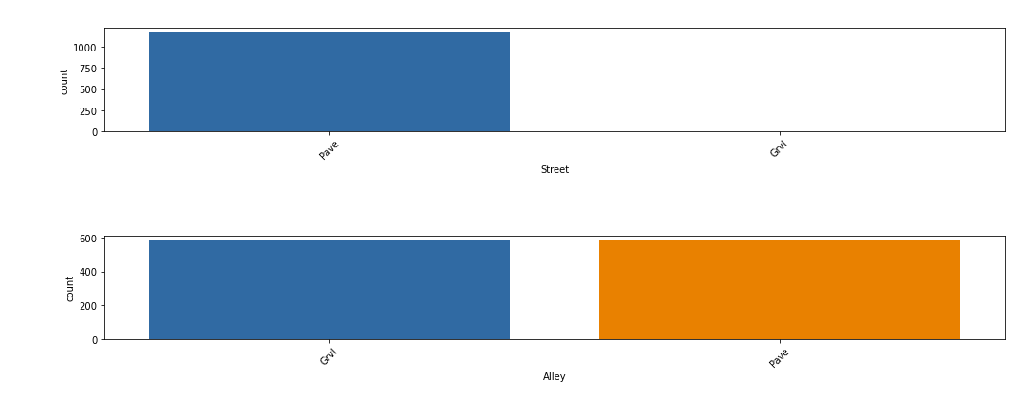
We got minimum errors in Ridge Regression Model of our project.

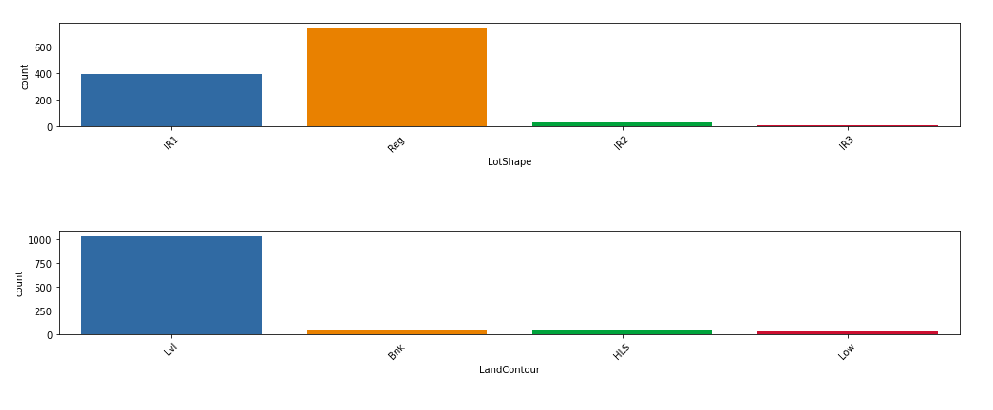
Because it handles multicollinearity with features well and gives good accuracy with less errors.so we used this metrics.

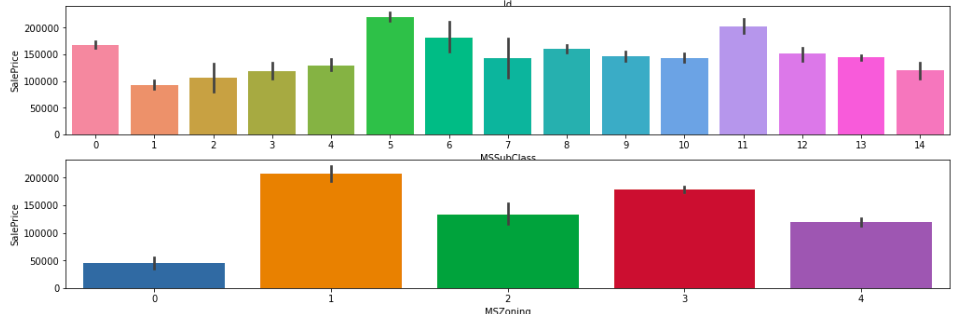
Visualizations

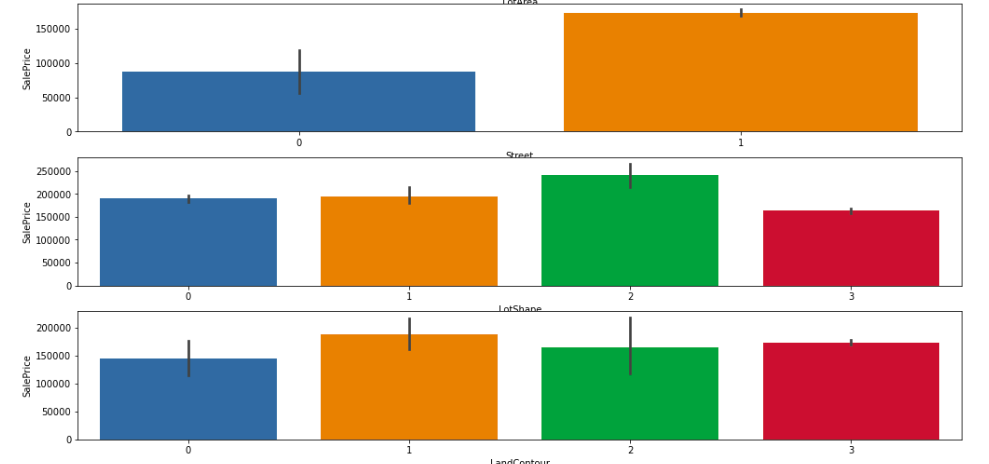


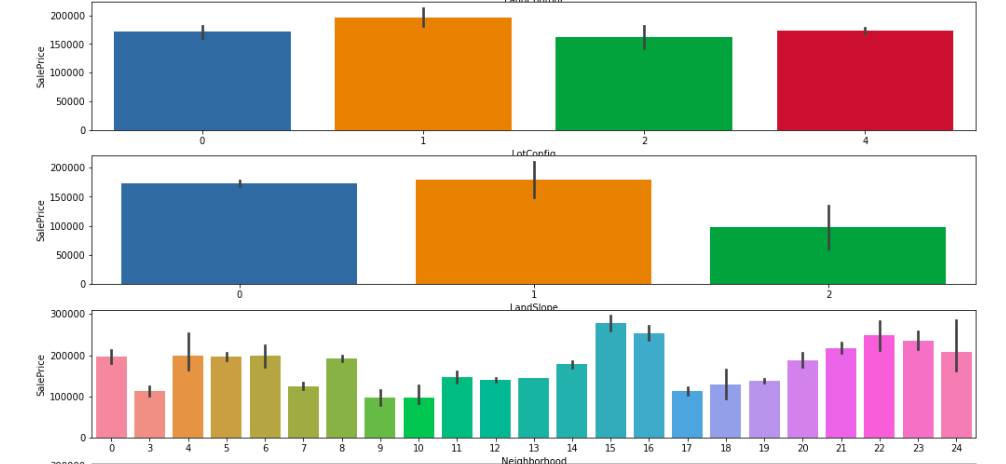




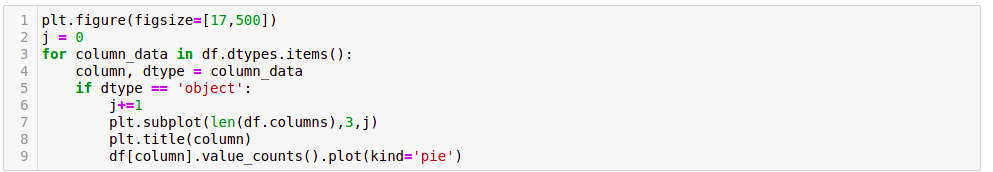




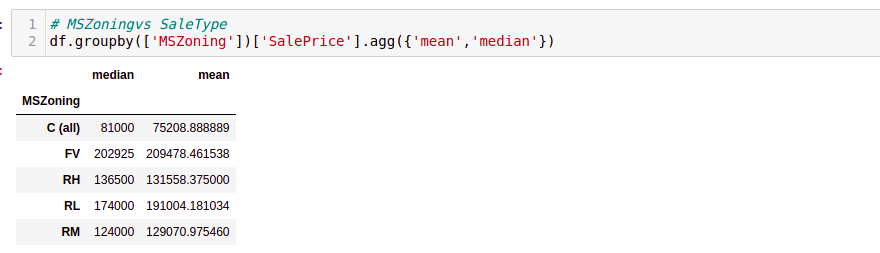




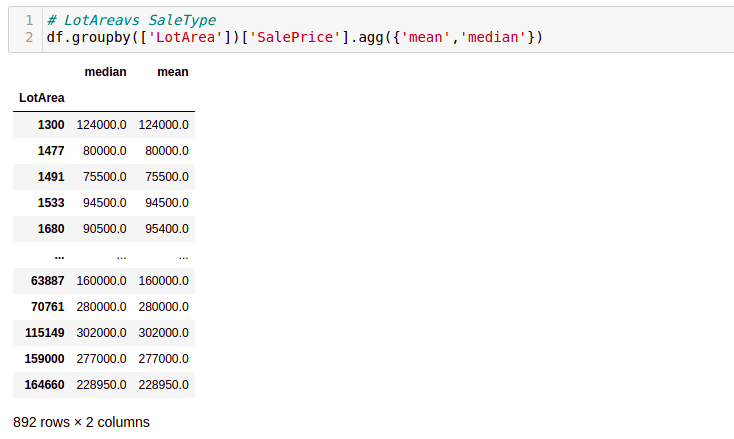
Interpretation of the Results







MODEL SAVING:



Inferences:

shows houses with zone,condition,age,sale type and price details

the mssubclass type 20,60,50 which are 1farm category only most available

the building type 1farm and housestyle 1story and 2 story are most available

the housestyle 1story and 2story which is in RL zone are most available

the housestyle 1story and 2story with condition normal are most available

when overallcondition is 9 which is is excellent then sale price will be increased

when sale condition is partial then sale price also increase

when overallquality very excellent then sale price also increases and will be high

mssubclass 20 are higher in numbers and has higher sum amount

mssubclass 20 are higher in numbers and has higher amount of selling price too

These are the most correlated features.The price of the house mostly depends on these features.

**CONCLUSION**

Key Findings and Conclusions of the Study

Saleprice of the house mostly depends on overall condition,salecondition,house zones and types,garage area,lot area,totalroom above ground,age of the house,remodelled or not,fireplace,Full Bath etc.,

These are the variables are important to predict the price of variable.

These variables describes the price of the house because these depends on environment,people’s living style,and needs of Australia.

All Houses are fixed with price depends on the above all prescribed variables only.

Learning Outcomes of the Study in respect of Data Science

From the above models,Ridge Regressor performs well,Because Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

This is our Best Fit Model.So we save this model for our analysis

Because to get out of this issue of without completely removing some predictor variables from the model is to use a method known as ridge regression, which instead seeks to minimize the following: where j ranges from 1 to p and λ ≥ 0. This second term in the equation is known as a shrinkage penalty.

To get out of this issue without completely removing some predictor variables from the model is to use a method known as ridge regression, which instead seeks to minimize the following: where j ranges from 1 to p and λ ≥ 0. This second term in the equation is known as a shrinkage penalty.

Thus this Ridge Regression Model performs well.so we saved this model.

Limitations of this work and Scope for Future Work

 The biggest drawback of ridge regression is its inability to perform variable selection since it includes all predictor variables in the final model.

Since some predictors will get shrunken very close to zero, this can make it hard to interpret the results of the model.

Handling of multicollinearity with more accuracy can improve the model performance to great extent and can predict any kind of similar datasets with more accurate results.