

Housing Project: Price Prediction

Submitted by:

Akriti Kakkar

**ACKNOWLEDGMENT**

I have referred to data trained study material and python documentation for this model development.

**INTRODUCTION**

* Business Problem Framing
  + **Housing price prediction is a regression based problem statement to:**
    1. **Which variables are important to predict the price of variable?**
    2. **How do these variables describe the price of the house?**

**Data Science is an evolving and new topic industry-wide. Real Estate is one of the beneficiaries of this development. Solution to these problem statements and its dominance in real estate sector can reap in a lot of profits by segmenting customer needs in a few minutes and making personalized recommendations possible in the click of a button.**

* Conceptual Background of the Domain Problem
  + 1. A US BASED HOUSING COMPANY, SURPRISING HOUSING, WANTS TO EXPAND ITS BUSINESS TO AUSTRALIA. IT HAS PROVIDED 80 FEATURES FOR 1460 HOUSES AND WANTS A PREDICTIVE MODEL TO PREDICT PRICE OF HOUSES. SO THAT THEY CAN BUY LOW PRICED PROPERTIES AND SELL HIGH PRICED PROPERTIES. IN ADDITION, IT WANTS A COMPLETE MARKETING MIX MODELLING TO UNDERSTAND HOW DIFFERENT FACTORS IMPACT PRICES.
    2. OBJECTIVE: To help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.
* Review of Literature

|  |  |
| --- | --- |
| * **We will understand the data by drawing anomalies and understanding their buying habits:** * **1. There are 43 categorical columns, containing housing data, example, 1 storey housing, 2 storey housing, etcetera. Normal sale condition is dominant, appearing, 80.91% times.** * **2. Most of the lot shape are regular.** * **3. Paved streets are the most popular.** * **4. Most of the zoning classification of sales are Residential Low Density.** * **Q&A Answered** * **How many houses are being surveyed: 1168** * **How many sale condition are observed: 6** * **What is the most popular fence quality: Minimum Privacy**   **EDA Steps Involve:**   * **HEAD VIEW OF DATA** * **TAIL VIEW OF DATA** * **SAMPLE VIEW OF DATA** * **GROUPBY EXPLORATION** * **DESCRIPTIVE STATISTICS** * **SCATTER PLOTS** * **CORRELATION ANALYSIS** * **BOX PLOTS EXPLORATION** * **DESCRIPTIVE STATISTICS** * **DISTRIBUTION PLOTS** |  |

* + 1. **Based on above analysis:**
    2. **1. There are many outliers in the data.**
    3. **2. Strong multicolliearity features are important for prediction because there f test and p value are acceptable. This means that the amount of multicollineaity is insignificant and removing the feature will impact the model much.**
    4. **3. Extereme leptokurtic and right skewed features are also relatively significant based on f test and p test.**
    5. **Hence, as a solution, feature scaling will do a better job in explaining the dependent variable than removing whole columns.**
    6. **After passing through vif test to remove multicollinearity, only 33 features seem to be low bias with seemingly strong explanatory power.**
    7. **As part of data handling, I have closely analyzed features with high outliers (by analyzing box plots, dist plots, variable plot and scatter plots).**
    8. **I have removed features with multicollinearity by analyzing correlation, correlation heatmaps and variance inflation factor.**
    9. **I have done ANOVA testing, wherever, applicable to weigh importance against bias.Hence, the model can be expected to be low variance and low bias model**
* Motivation for the Problem Undertaken
  + 1. OBJECTIVE: To help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

****

1. **Data Extraction: Using read\_ csv function of pandas library to read the data in tabulated format and analyze it.**
2. **Data Cleaning for missing values detection and its handiling.**
3. **Feature Engineering for encoding object format data and deriving more features.**
4. **EDA For data visualization and biasness detection:**

* **HEAD VIEW OF DATA**
* **TAIL VIEW OF DATA**
* **SAMPLE VIEW OF DATA**
* **GROUPBY EXPLORATION**
* **DESCRIPTIVE STATISTICS**
* **SCATTER PLOTS**
* **CORRELATION ANALYSIS**
* **BOX PLOTS EXPLORATION**
* **DESCRIPTIVE STATISTICS**
* **DISTRIBUTION PLOTS**

1. **VIF Test for multicollinearity reduction.**
2. **Power Transformation for standard scaling and outliers transformation.**
3. **Data PreProcessing for data transformation, scaling and vectorization.**
4. **Feature Selection (Ensemble Methods): ANOVA Test, p value, ftest, constant threshold filter to classify features based on relevance and biasness and select the most relevant features.**
5. **Model Development, Evaluation And Selection (Ensemble Methods and Grid Search CV) to do best hyper parameter tuning and develop low bias and low variance with right fit and minimal difference between test metrics and train metrics.**

* 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

* Data Preprocessing Done
* **Data Pre Processing**
  + **POWER TRANSFORM**
  + **MIN MAX SCALING**
  + **VECTORIZATION**
  + **TRAIN TEST SPLIT**

**The data was further used for feature selection.**

**Assumption made:**

* **Acceptable Skewness Range Is +/-0.65**
* **Acceptable VIF Score Is Than 6**
* **Acceptable P Value Is Less Than 0.05**
* **Variance Threshold Is 0.01**
* Data Inputs- Logic- Output Relationships
* **Data Inputs include, Index(['LandSlope\_encoded', 'OpenPorchSF', 'Fireplaces', 'LotShape\_encoded',  
         'HeatingQC\_encoded', 'BsmtFinSF1\_pct\_change', 'MasVnrArea',  
         'WoodDeckSF', 'GarageFinish\_encoded\_pct\_change', 'HalfBath'],  
        dtype='object')**

**Data Input Type: float; Min Max Scaling in the range of 0 to 1.**

**Impact On Output: Lot Shape\_encoded; HeatingQC\_encoded and GarageFinish\_encoded\_pct\_change are negatively correlated with labeland others are positively correlated with label.**

**'LandSlope\_encoded':** 0.015484795080526005**,**

**'OpenPorchSF':** 0.33949955918549074**,**

**'Fireplaces':** 0.459610550802869**,**

**'LotShape\_encoded':** -0.24817105697155653**,  
       'HeatingQC\_encoded':** -0.4066035594011184**,**

**'BsmtFinSF1\_pct\_change':** 0.03927441750012172**,**

**'MasVnrArea':** 0.46120570017748197**,  
       'WoodDeckSF':** 0.31544416227339683**,**

**'GarageFinish\_encoded\_pct\_change':** -0.23583840095816014**,**

**'HalfBath':** 0.29559237431380403

* State the set of assumptions (if any) related to the problem under consideration
  + Acceptable Skewness Range Is +/-0.65
  + Acceptable VIF Score Is Than 6
  + Acceptable P Value Is Less Than 0.05
  + Variance Threshold Is 0.01
* Hardware and Software Requirements and Tools Used

**Installation Of Anaconda Library.**

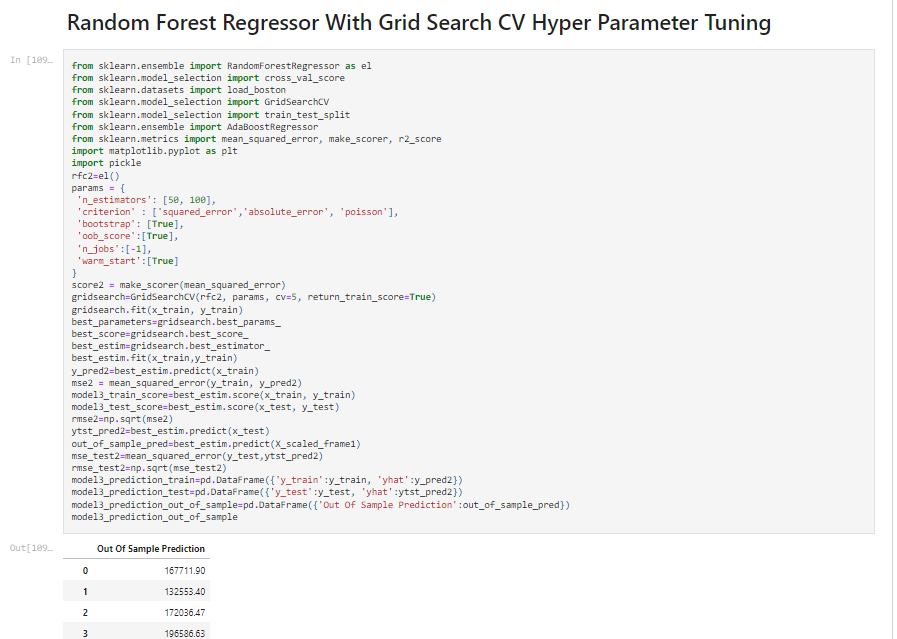
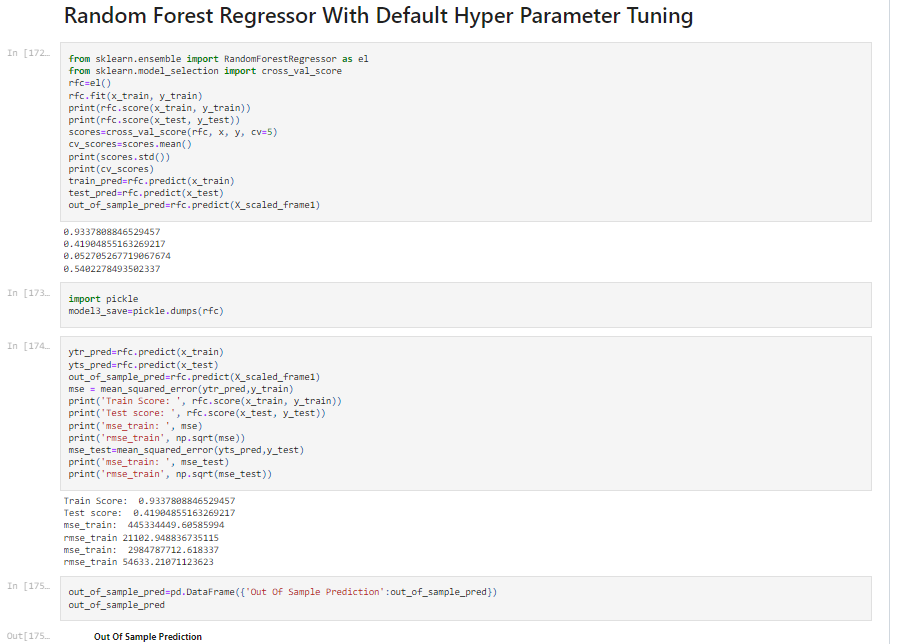
**Required Installations:**

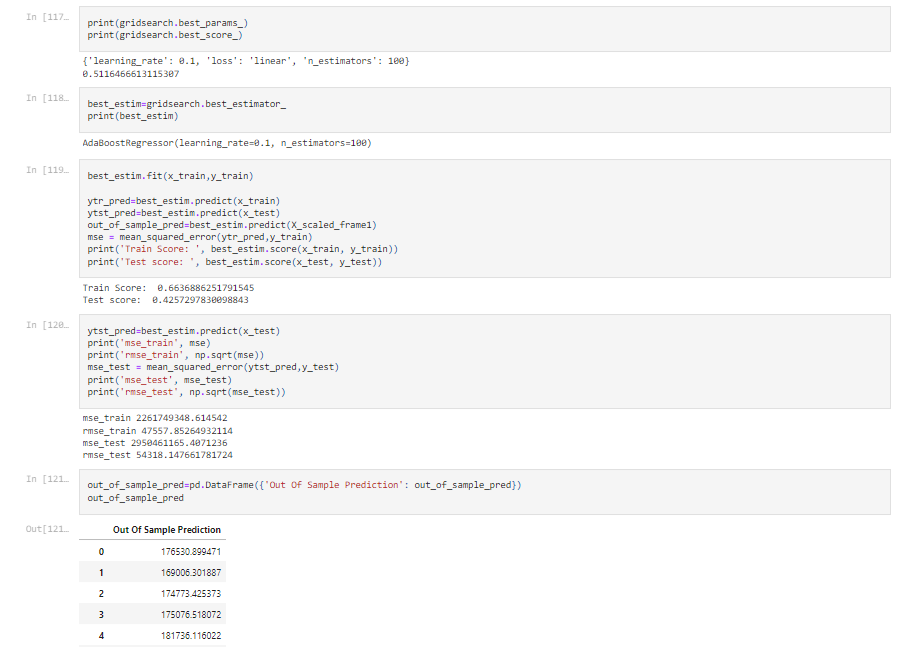
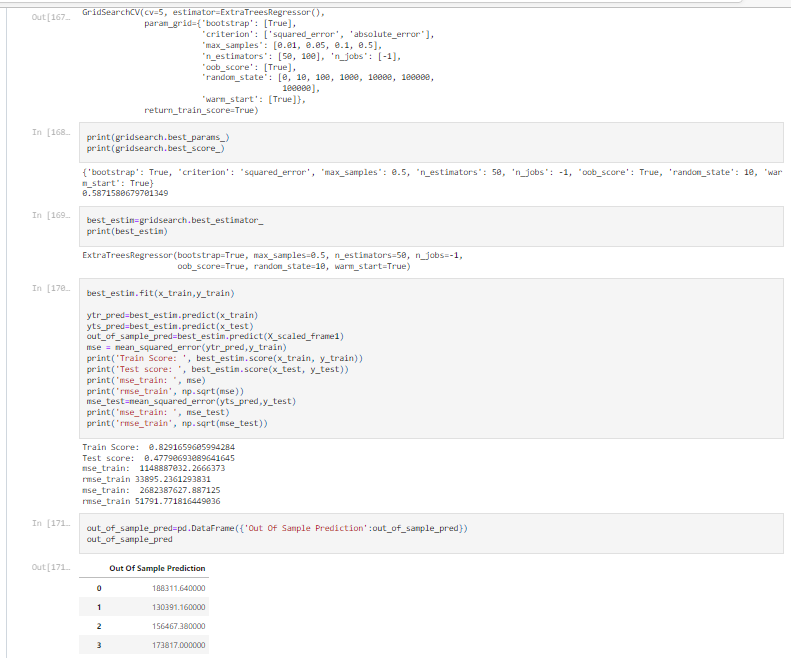
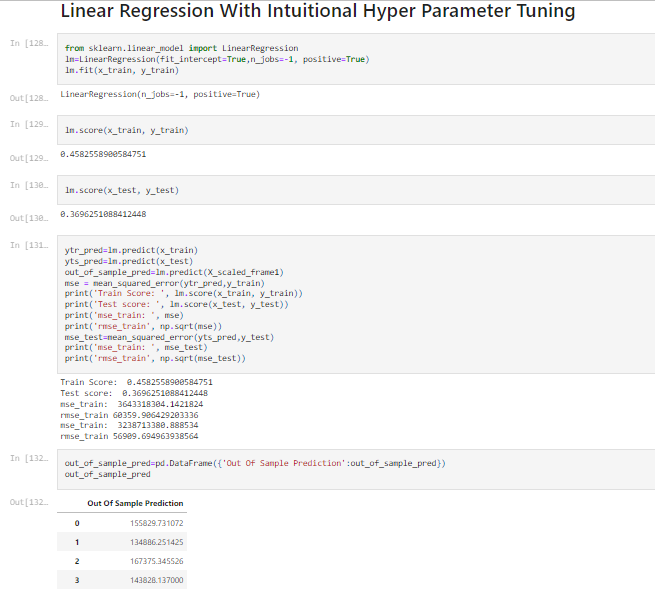
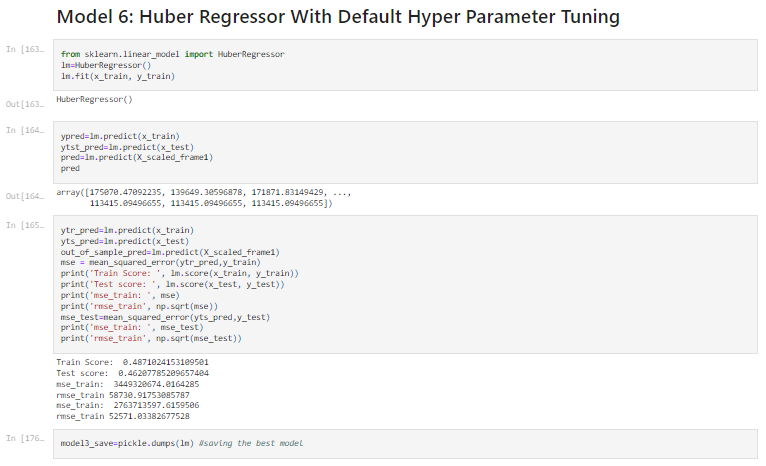
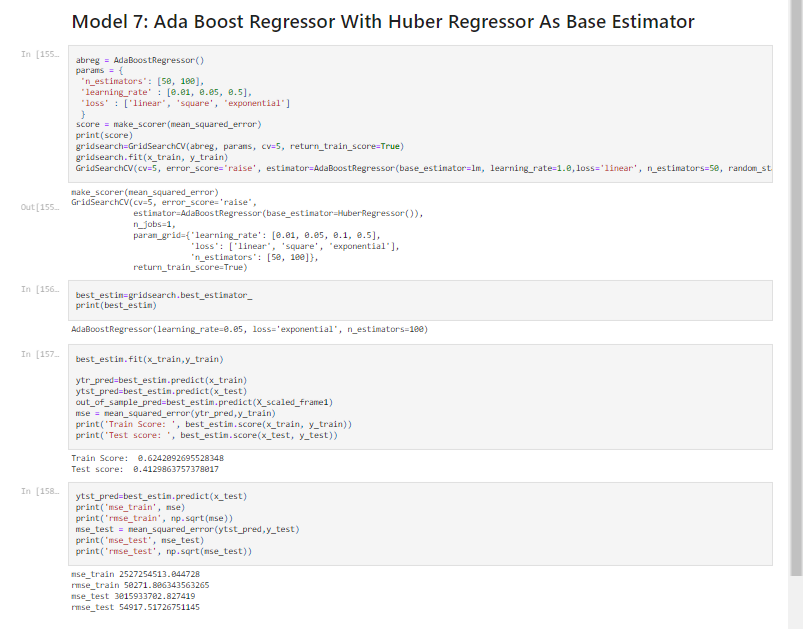
* **Pandas (Within environment)**
* **Numpy (Within environment)**
* **Seaborn (Within environment)**
* **Matplotlib (Within environment)**
* **Cufflinks**
* **Plotly Express**
* **Sklearn**

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* 
  + **Total Models = 7**
  + **Selection Reasoning Of Models 1 to 7**
  + **The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.**
  + **Two families of ensemble methods are usually distinguished:**
  + **In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.**
  + **Examples: Bagging methods, Forests of randomized trees, etcetera**
  + **By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.**
  + **Examples: AdaBoost, Gradient Tree Boosting,etcetera**
  + **The use case assigned revolves around uneven label points hence there is high probality of not achieving a good fit. Hence, I have tried different ensembele techniques that can lower variance and bias and help achieve good scores. (Just as a reminder, I have already applied variance threshold of 0.01 to ensure that risk of models is low).**
  + **The theories in the above two cells explain why I have chosen Model 1, Model 2, Model 3, Model 4, Model 5, Model 6 and Model7**
* Testing of Identified Approaches (Algorithms)

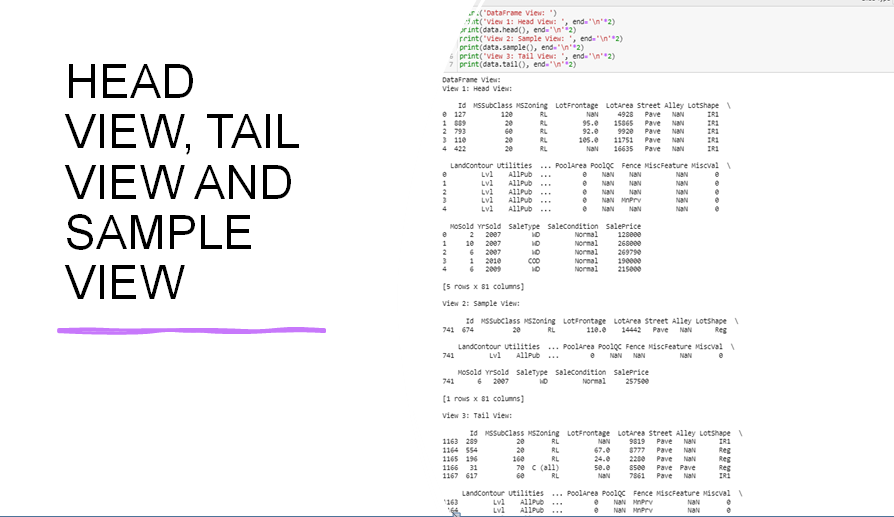
Total Models = 7

* **Model 1:** **Random Forest Regressor With Grid Search CV Hyper Parameter Tuning**
* **Model 2:** **Random Forest Regressor With Default Hyper Parameter Tuning**
* **Model 3:** **Ada Boost Regressor And Random Forest Regressor With Grid Search CV Hyper Parameter Tuning and Ada Boost Boosting**
* **Model 4: Extra Trees Regressor With Grid Search CV Hyper Parameter Tuning**
* **Model 5:** **Linear Regression With Intuitional Hyper Parameter Tuning**
* **Model 6: Huber Regressor With Default Hyper Parameter Tuning**
* **Model 7: Ada Boost Regressor With Huber Regressor As Base Estimator**
* Run and Evaluate selected models
* **Model 1:** **Random Forest Regressor With Grid Search CV Hyper Parameter Tuning**
  + MSE Train:  472351089.3783724
  + RMSE Train:  21733.6395796556
  + MSE Test:  2941145189.730843
  + RMSE Train:  54232.32605864184
  + 
* **Model 2:** **Random Forest Regressor With Default Hyper Parameter Tuning**
  + mse\_train:  445334449.60585994  
    rmse\_train 21102.948836735115  
    mse\_train:  2984787712.618337  
    rmse\_train 54633.21071123623
  + 

* **Model 3:** **Ada Boost Regressor And Random Forest Regressor With Grid Search CV Hyper Parameter Tuning and Ada Boost Boosting**
  + mse\_train 2261749348.614542  
    rmse\_train 47557.85264932114  
    mse\_test 2950461165.4071236  
    rmse\_test 54318.147661781724
  + 
* **Model 4: Extra Trees Regressor With Grid Search CV Hyper Parameter Tuning**
  + mse\_train:  1148887032.2666373  
    rmse\_train 33895.2361293831  
    mse\_train:  2682387627.887125  
    rmse\_train 51791.771816449036
  + 
* **Model 5:** **Linear Regression With Intuitional Hyper Parameter Tuning**
  + mse\_train:  3643318304.1421824  
    rmse\_train 60359.906429203336  
    mse\_train:  3238713380.888534  
    rmse\_train 56909.694963938564
  + 
* **Model 6: Huber Regressor With Default Hyper Parameter Tuning**
  + mse\_train:  3449320674.0164285  
    rmse\_train 58730.91753085787  
    mse\_train:  2763713597.6159506  
    rmse\_train 52571.03382677528
  + 
* **Model 7: Ada Boost Regressor With Huber Regressor As Base Estimator**
  + mse\_train 2527254513.044728  
    rmse\_train 50271.806343563265  
    mse\_test 3015933702.827419  
    rmse\_test 54917.51726751145
  + 
* Key Metrics for success in solving problem under consideration
  + - 1. Power Transform: To remove outliers from extremely spread out data.
      2. VIF Scores: To reduce multicollinearity from a highly biased dataset.
      3. Ensemble Methods: To remove over fitting in a complex dataset and finding maximum explanatory power .
* Visualizations

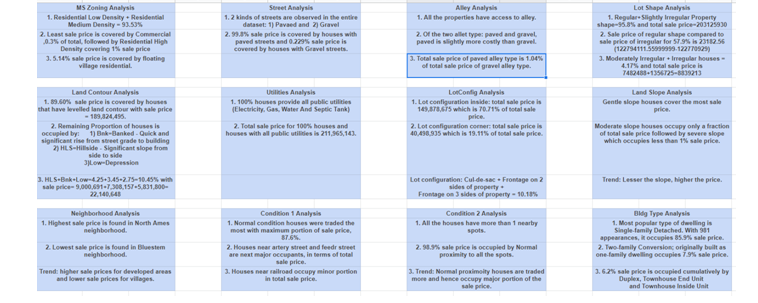
**EDA Steps Involve:**

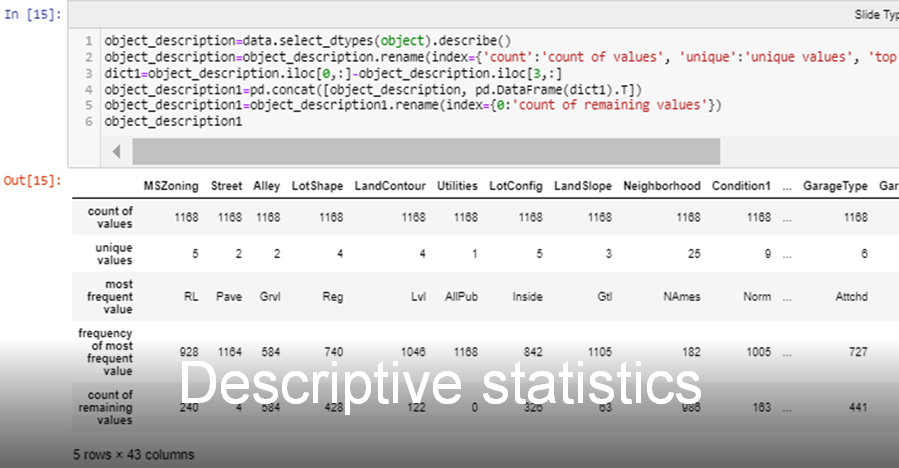
* **HEAD VIEW OF DATA**
* **TAIL VIEW OF DATA**
* **SAMPLE VIEW OF DATA**
* **GROUPBY EXPLORATION**
* **DESCRIPTIVE STATISTICS**
* **SCATTER PLOTS**
* **CORRELATION ANALYSIS**
* **BOX PLOTS EXPLORATION**
* **DESCRIPTIVE STATISTICS**
* **DISTRIBUTION PLOTS**



* **# View 1: Analysis**
* **1. Data alongside represents first five houses of 1168 houses in train dataset.**
* **2. Data seems complex and incomplete due to 81 columns and many missing values.**
* **3.  Upper limit of sale price of first five houses is 268000 and lower limit is 128000**
* **4. All the five houses have been sold in normal condition with mode lying in the month of junefor the years 2007 and 2009 respectively.**
* **5. Only one house provides fencing facility and all the five houses provide pub utility**
* **6. Street type is pavement for all the five houses**
* **7. Saletype for 1 house is court office deed and the other four houses is warranty deed conventional.**
* **8. There are three varities of housing seen in mssubclass with 3 dwellings identified as 20 : 1 story 1946 and newer all styles , 1 dwelling identified as 60 : 2 story 1946 and newer and 1 dwelling identified as 120 : 1 story planned unit developement 1946 & newer.**
* **9. The  general shape of all the five properties is IR1 slightly irregular and the flatness of all the property ois LV1 which is near flat/level.**
* **10. The identified general zoning classification is RL : Residential Low Density for all the five properties.**
* # View 2: Analysis
* 1. It shows a random house at index 741.It is 1 storey 1846 and newer all styles, regular shape house that provides all public utilities and is sold by warranty deed for 257500.
* # View 3: Analysis
* 1. It represents last 5 rows. In which all are warranty deeds, sold in normal condition.







* **We will understand the data by drawing anomalies and understanding their buying habits:**
* **1. There are 43 categorical columns, containing housing data, example, 1 storey housing, 2 storey housing, etcetera. Normal sale condition is dominant, appearing, 80.91% times.**
* **2. Most of the lot shape are regular.**
* **3. Paved streets are the most popular.**
* **4. Most of the zoning classification of sales are Residential Low Density.**
* **Q&A Answered**
* **How many houses are being surveyed: 1168**
* **How many sale condition are observed: 6**
* **What is the most popular fence quality: Minimum Privacy**
* **# Anomalies Detected**
* **## Highest sale price is 755000 and is obtained for a property built in 1994 in 2 storey 1946 and newer.**
* **## Maximum lot area is 164660 and lot area of maximum sale price property is 21535.**
* **## Overall quality of this house Very Excellent and overall Condition is above average.**
* **## Most significant sale price pct change is 6.5.**
* **# Q&A Answered**
* **## What is the range of Open Porch SF: 0 to 600.**
* **## What is the range of Garage Cars: 0 to 4.**
* **## What is the range of Total Price Above Ground: 0 to 15.**
* **## What is the range of Half Bath: 0 to 2.**

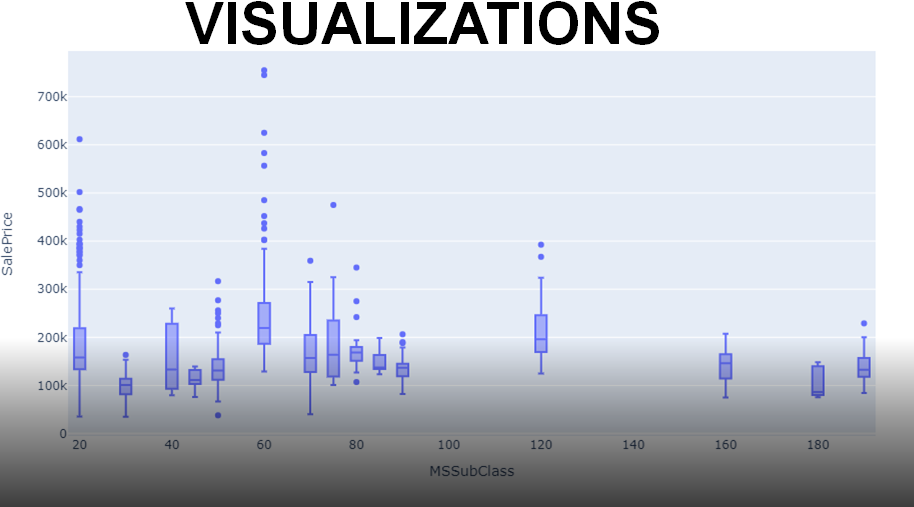


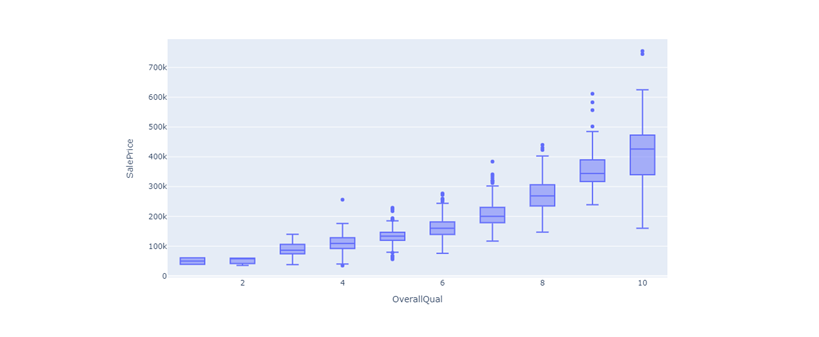
* **# The above iplot represents correlation of features with label. Correlation of features with label is of high relevance. The stronger the relationship of label with axis, the more accurate the prediction. In the above line graph:**
* **## 1. Ignoring sale price pct change, highest correlation with label is observed to be 78.92%, that is shared with Overall Quality. Minimum correlation is -0.4%, shared with BsmtHalfBath\_pct\_change.**
* **## 2. Weak Positive To Strong Positive Relationship Is Found With 110 Features, including these:**
* **BsmtHalfBath\_pct\_change    0.004155282515700739**
* **ExterCond\_encoded\_pct\_change    0.004999542764231268**
* **Fence\_encoded\_pct\_change    0.00618819242998486**
* **MasVnrType\_encoded    0.006763415444002462**
* **BsmtCond\_encoded\_pct\_change    0.011509937574556038**
* **MasVnrType\_encoded\_pct\_change    0.011631168977197338**
* **LandSlope\_encoded    0.015484795080526005**
* **BsmtFinSF2\_pct\_change    0.01694247841896264**
* **## 3. Weak Negative to Strong Negative Relationship Is Found With Following Features:**
* **ExterQual\_encoded    -0.624820046916615**
* **BsmtQual\_encoded    -0.6074933757146362**
* **KitchenQual\_encoded    -0.5924675972943289**
* **GarageFinish\_encoded    -0.48745288300973294**
* **HeatingQC\_encoded    -0.4066035594011184**
* **ExterQual\_encoded\_pct\_change    -0.4037123724436022**
* **## 4. Correlation Of Label with itself is of no relevance**

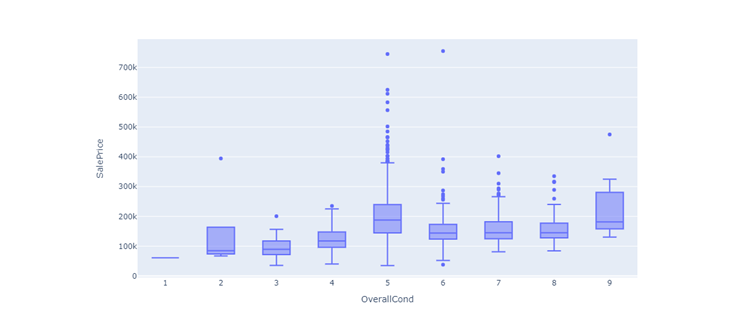


* **1. Label has weak positive and strong positive relationship with these features:**
* **['LotFrontage', 'LotArea', 'OverallQual', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'BsmtFullBath', 'FullBath', 'HalfBath', 'BedroomAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MoSold', 'SalePrice', 'Street\_encoded', 'Alley\_encoded', 'LandContour\_encoded', 'LandSlope\_encoded', 'Neighborhood\_encoded', 'Condition1\_encoded', 'Condition2\_encoded', 'HouseStyle\_encoded', 'RoofStyle\_encoded', 'RoofMatl\_encoded', 'Exterior1st\_encoded', 'Exterior2nd\_encoded', 'MasVnrType\_encoded', 'ExterCond\_encoded', 'Foundation\_encoded', 'BsmtCond\_encoded', 'BsmtFinType2\_encoded', 'CentralAir\_encoded', 'Electrical\_encoded', 'Functional\_encoded', 'GarageQual\_encoded', 'GarageCond\_encoded', 'PavedDrive\_encoded', 'PoolQC\_encoded', 'MiscFeature\_encoded', 'SaleCondition\_encoded', 'LotFrontage\_pct\_change', 'LotArea\_pct\_change', 'OverallQual\_pct\_change', 'YearBuilt\_pct\_change', 'YearRemodAdd\_pct\_change', 'MasVnrArea\_pct\_change', 'BsmtFinSF1\_pct\_change', 'BsmtFinSF2\_pct\_change', 'BsmtUnfSF\_pct\_change', 'TotalBsmtSF\_pct\_change', '1stFlrSF\_pct\_change', '2ndFlrSF\_pct\_change', 'LowQualFinSF\_pct\_change', 'GrLivArea\_pct\_change', 'BsmtFullBath\_pct\_change', 'BsmtHalfBath\_pct\_change', 'FullBath\_pct\_change', 'HalfBath\_pct\_change', 'BedroomAbvGr\_pct\_change', 'TotRmsAbvGrd\_pct\_change', 'Fireplaces\_pct\_change', 'GarageYrBlt\_pct\_change', 'GarageCars\_pct\_change', 'GarageArea\_pct\_change', 'WoodDeckSF\_pct\_change', 'OpenPorchSF\_pct\_change', '3SsnPorch\_pct\_change', 'PoolArea\_pct\_change', 'MiscVal\_pct\_change', 'MoSold\_pct\_change', 'SalePrice\_pct\_change', 'Street\_encoded\_pct\_change', 'Neighborhood\_encoded\_pct\_change', 'Condition1\_encoded\_pct\_change', 'Condition2\_encoded\_pct\_change', 'BldgType\_encoded\_pct\_change', 'HouseStyle\_encoded\_pct\_change', 'RoofStyle\_encoded\_pct\_change', 'RoofMatl\_encoded\_pct\_change', 'Exterior1st\_encoded\_pct\_change', 'Exterior2nd\_encoded\_pct\_change', 'MasVnrType\_encoded\_pct\_change', 'ExterCond\_encoded\_pct\_change', 'Foundation\_encoded\_pct\_change', 'BsmtCond\_encoded\_pct\_change', 'CentralAir\_encoded\_pct\_change', 'Electrical\_encoded\_pct\_change', 'Functional\_encoded\_pct\_change', 'GarageQual\_encoded\_pct\_change', 'GarageCond\_encoded\_pct\_change', 'PavedDrive\_encoded\_pct\_change', 'PoolQC\_encoded\_pct\_change', 'Fence\_encoded\_pct\_change', 'SaleCondition\_encoded\_pct\_change']**
* **2. Label has weak negative to strong negative relationship with these features in this heatmap:**
* **['Id', 'MSSubClass', 'OverallCond', 'BsmtFinSF2', 'LowQualFinSF', 'BsmtHalfBath', 'KitchenAbvGr', 'EnclosedPorch', 'MiscVal', 'YrSold', 'MSZoning\_encoded', 'LotShape\_encoded', 'LotConfig\_encoded', 'BldgType\_encoded', 'ExterQual\_encoded', 'BsmtQual\_encoded', 'BsmtExposure\_encoded', 'BsmtFinType1\_encoded', 'Heating\_encoded', 'HeatingQC\_encoded', 'KitchenQual\_encoded', 'FireplaceQu\_encoded', 'GarageType\_encoded', 'GarageFinish\_encoded', 'Fence\_encoded', 'SaleType\_encoded', 'Id\_pct\_change', 'MSSubClass\_pct\_change', 'OverallCond\_pct\_change', 'KitchenAbvGr\_pct\_change', 'EnclosedPorch\_pct\_change', 'ScreenPorch\_pct\_change', 'YrSold\_pct\_change', 'MSZoning\_encoded\_pct\_change', 'Alley\_encoded\_pct\_change', 'LotShape\_encoded\_pct\_change', 'LandContour\_encoded\_pct\_change', 'LotConfig\_encoded\_pct\_change', 'LandSlope\_encoded\_pct\_change', 'ExterQual\_encoded\_pct\_change', 'BsmtQual\_encoded\_pct\_change', 'BsmtExposure\_encoded\_pct\_change', 'BsmtFinType1\_encoded\_pct\_change', 'BsmtFinType2\_encoded\_pct\_change', 'Heating\_encoded\_pct\_change', 'HeatingQC\_encoded\_pct\_change', 'KitchenQual\_encoded\_pct\_change', 'FireplaceQu\_encoded\_pct\_change', 'GarageType\_encoded\_pct\_change', 'GarageFinish\_encoded\_pct\_change', 'MiscFeature\_encoded\_pct\_change', 'SaleType\_encoded\_pct\_change']**
* **3. Strong Multicollinearity is detecte in most of the features, including features like:**
* Overall Quality  
    
  Half Bath, and many more.  
    
  We will explore this deeply in vif section of data analysis.
* **4. Multicollinearity seems to be moderate among all other data points.**
* **Q&A Answered**
* 1. Most useful feature is: Overall Quallity.  
    
      2. Features that can cause bias are: Overall Quality, Half Bath and many more.  
    
      3. Explanatory Power of all variable: Weak to strong (due to weak, moderate and strong correlation with label).
* **Overall strong correlation among dataset. There seems some multicollinearity due to presence of Overall Quality, Half Bath, etcetera. We will do further eda before arriving at a conclusion to delete these columns.**
* **The most useful feature is Overall Quality, apart from this, there seems modest correlation of features with label.**
* **As a conclusion, we will do further cause and effect analysis based on skewness and distribution plots to conclude if we want to remove multicollinarity pairs:**
* i. Half Bath  
    
  ii. MasVnrArea  
    
  iii. Total Bsmt SF, etcetera

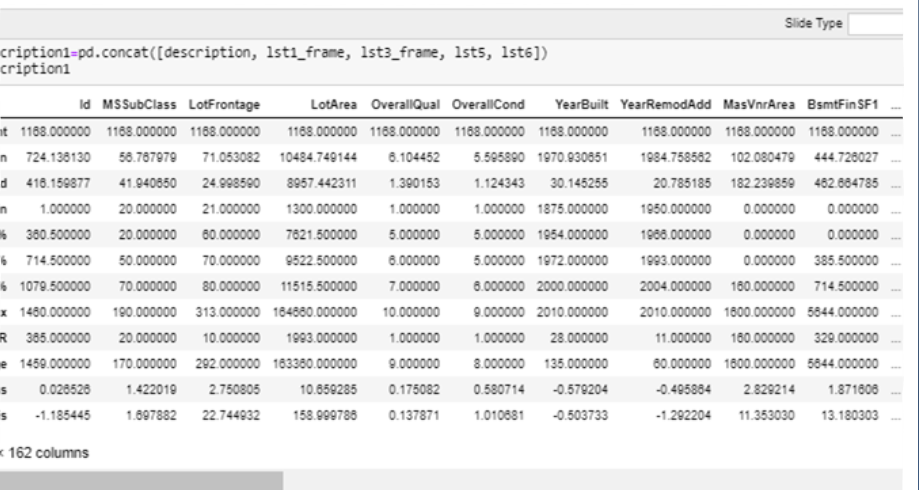
Box Plots (I have included few visualizations so that file can be uploaded in github repository and is not forbidden because of too large size, In jupyter notebook, all visualizations can be seen clearly).

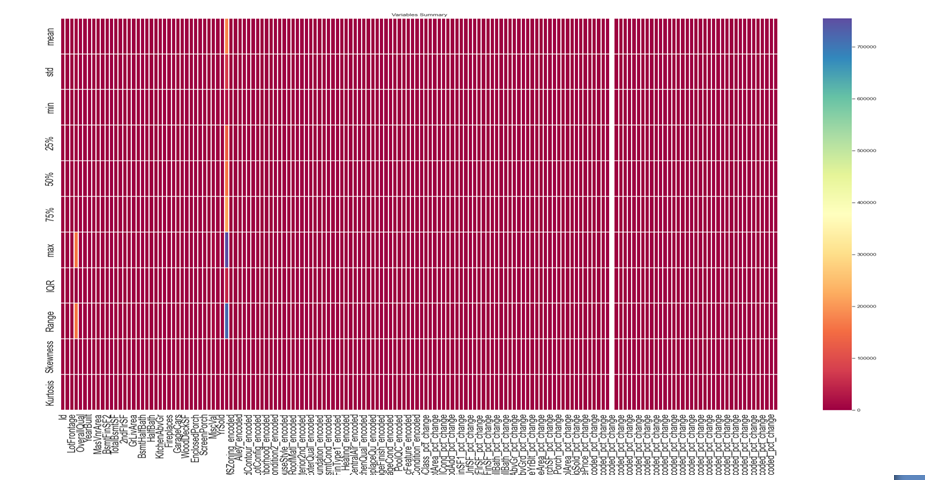






* 1. The data has many outliers.
* 2. To name a few:
* MS Sub Class
* Lot Frontage
* Lot Area
* **By studying box plots and correlation analysis, we have found that outliers are present in all the columns, hence, I am removing outliers by data points. We will do further eda with vif, dist plots and descriptive statistics to finally select a set of features.**





* **Mean = sum of values/count of values**
* **std = sqrt(((value - mean of distribution)\*\*2 / number of values ))**
* **3 quartile are measures of variance, calculated to spot the placeholder value, it returns index of the produced value. Step 1: sort the dataset  
  Step2:  
  i) Lower Quartile (Q1: 25% distribution) = ((number of values+1)/4)th Term  
  ii) Middle Quartile (Q2: 50% distribution) = ((number of values +1)/2)th Term  
  Also, know as median (central value).  
  iii) Upper Quartile (Q3: 75% distribution) = ¾(number of values + 1)th Term  
  iv) IQR = Upper Quartile - Lowe Quartile**
* **Range = Maximum Value - Minimum Value**
* **Skewness = (sumation(value - mean of distribution)3)/((number of values - 1) \* std3)**
* **Kurtosis = number of values \* ((sumation(value - mean of distribution)4) / std4)**
* **1. For values that are scaled upto 1, mean is mostly around 0 and standard deviation is comparatively low. Hence, making the data much more acceptable by algorithms to process it more accurately.**
* **2. The entire dataset ranges from -1 till 755000.**
* **3. Skewness is 0 and within +/- 0.65 for:**
* ['Id', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd',  
         'BsmtFullBath', 'FullBath', 'BedroomAbvGr', 'TotRmsAbvGrd',  
         'GarageYrBlt', 'GarageCars', 'GarageArea', 'MoSold', 'YrSold',  
         'Alley\_encoded', 'LotShape\_encoded', 'Utilities\_encoded',  
         'Neighborhood\_encoded', 'HouseStyle\_encoded', 'Exterior1st\_encoded',  
         'Exterior2nd\_encoded', 'MasVnrType\_encoded', 'Foundation\_encoded',  
         'BsmtFinType1\_encoded', 'HeatingQC\_encoded', 'FireplaceQu\_encoded',  
         'GarageFinish\_encoded', 'PoolQC\_encoded', 'Fence\_encoded',  
         'YearBuilt\_pct\_change', 'YearRemodAdd\_pct\_change',  
         '2ndFlrSF\_pct\_change', 'Fireplaces\_pct\_change',  
         'GarageYrBlt\_pct\_change', 'YrSold\_pct\_change',  
         'Street\_encoded\_pct\_change', 'Alley\_encoded\_pct\_change',  
         'LotShape\_encoded\_pct\_change', 'Utilities\_encoded\_pct\_change',  
         'Foundation\_encoded\_pct\_change', 'BsmtQual\_encoded\_pct\_change',  
         'HeatingQC\_encoded\_pct\_change', 'CentralAir\_encoded\_pct\_change',  
         'KitchenQual\_encoded\_pct\_change', 'GarageFinish\_encoded\_pct\_change',  
         'PavedDrive\_encoded\_pct\_change', 'PoolQC\_encoded\_pct\_change',  
         'Fence\_encoded\_pct\_change']
* **4. Acceptable skewness is +/- 0.65 and skewness for bell shaped curve should be 0.**
* ['Id', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd',  
     'BsmtFullBath', 'FullBath', 'BedroomAbvGr', 'TotRmsAbvGrd',  
     'GarageYrBlt', 'GarageCars', 'GarageArea', 'MoSold', 'YrSold',  
     'Alley\_encoded', 'LotShape\_encoded', 'Utilities\_encoded',  
     'Neighborhood\_encoded', 'HouseStyle\_encoded', 'Exterior1st\_encoded',  
     'Exterior2nd\_encoded', 'MasVnrType\_encoded', 'Foundation\_encoded',  
     'BsmtFinType1\_encoded', 'HeatingQC\_encoded', 'FireplaceQu\_encoded',  
     'GarageFinish\_encoded', 'PoolQC\_encoded', 'Fence\_encoded',  
     'YearBuilt\_pct\_change', 'YearRemodAdd\_pct\_change',  
     '2ndFlrSF\_pct\_change', 'Fireplaces\_pct\_change',  
     'GarageYrBlt\_pct\_change', 'YrSold\_pct\_change',  
     'Street\_encoded\_pct\_change', 'Alley\_encoded\_pct\_change',  
     'LotShape\_encoded\_pct\_change', 'Utilities\_encoded\_pct\_change',  
     'Foundation\_encoded\_pct\_change', 'BsmtQual\_encoded\_pct\_change',  
     'HeatingQC\_encoded\_pct\_change', 'CentralAir\_encoded\_pct\_change',  
     'KitchenQual\_encoded\_pct\_change', 'GarageFinish\_encoded\_pct\_change',  
     'PavedDrive\_encoded\_pct\_change', 'PoolQC\_encoded\_pct\_change',  
     'Fence\_encoded\_pct\_change']
* **5. Kutosis is upto 3 for most dataset, indicating platykurtic curves:**
* ['Id', 'MSSubClass', 'OverallQual', 'OverallCond', 'YearBuilt',  
     'YearRemodAdd', 'BsmtUnfSF', '2ndFlrSF', 'BsmtFullBath', 'FullBath',  
     'HalfBath', 'BedroomAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt',  
     'GarageCars', 'GarageArea', 'WoodDeckSF', 'MoSold', 'YrSold',  
     'Alley\_encoded', 'LotShape\_encoded', 'Utilities\_encoded',  
     'LotConfig\_encoded', 'Neighborhood\_encoded', 'HouseStyle\_encoded',  
     'RoofStyle\_encoded', 'Exterior1st\_encoded', 'Exterior2nd\_encoded',  
     'MasVnrType\_encoded', 'Foundation\_encoded', 'BsmtQual\_encoded',  
     'BsmtExposure\_encoded', 'BsmtFinType1\_encoded', 'HeatingQC\_encoded',  
     'KitchenQual\_encoded', 'FireplaceQu\_encoded', 'GarageType\_encoded',  
     'GarageFinish\_encoded', 'PoolQC\_encoded', 'Fence\_encoded',  
     'YearBuilt\_pct\_change', 'YearRemodAdd\_pct\_change',  
     '2ndFlrSF\_pct\_change', 'BsmtFullBath\_pct\_change', 'FullBath\_pct\_change',  
     'HalfBath\_pct\_change', 'TotRmsAbvGrd\_pct\_change',  
     'Fireplaces\_pct\_change', 'GarageYrBlt\_pct\_change',  
     'GarageCars\_pct\_change', '3SsnPorch\_pct\_change', 'YrSold\_pct\_change',  
     'Alley\_encoded\_pct\_change', 'LotShape\_encoded\_pct\_change',  
     'Utilities\_encoded\_pct\_change', 'LotConfig\_encoded\_pct\_change',  
     'BldgType\_encoded\_pct\_change', 'HouseStyle\_encoded\_pct\_change',  
     'RoofStyle\_encoded\_pct\_change', 'MasVnrType\_encoded\_pct\_change',  
     'Foundation\_encoded\_pct\_change', 'BsmtQual\_encoded\_pct\_change',  
     'BsmtExposure\_encoded\_pct\_change', 'BsmtFinType1\_encoded\_pct\_change',  
     'HeatingQC\_encoded\_pct\_change', 'KitchenQual\_encoded\_pct\_change',  
     'GarageType\_encoded\_pct\_change', 'GarageFinish\_encoded\_pct\_change',  
     'PoolQC\_encoded\_pct\_change', 'Fence\_encoded\_pct\_change']
* **6. Kurtosis is greater than 3, indicating, leptokurtic curve:**
* ['LotFrontage', 'LotArea', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2',  
     'TotalBsmtSF', '1stFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtHalfBath',  
     'KitchenAbvGr', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch',  
     'ScreenPorch', 'PoolArea', 'MiscVal', 'SalePrice', 'MSZoning\_encoded',  
     'Street\_encoded', 'LandContour\_encoded', 'LandSlope\_encoded',  
     'Condition1\_encoded', 'Condition2\_encoded', 'BldgType\_encoded',  
     'RoofMatl\_encoded', 'ExterQual\_encoded', 'ExterCond\_encoded',  
     'BsmtCond\_encoded', 'BsmtFinType2\_encoded', 'Heating\_encoded',  
     'CentralAir\_encoded', 'Electrical\_encoded', 'Functional\_encoded',  
     'GarageQual\_encoded', 'GarageCond\_encoded', 'PavedDrive\_encoded',  
     'MiscFeature\_encoded', 'SaleType\_encoded', 'SaleCondition\_encoded',  
     'Id\_pct\_change', 'MSSubClass\_pct\_change', 'LotFrontage\_pct\_change',  
     'LotArea\_pct\_change', 'OverallQual\_pct\_change',  
     'OverallCond\_pct\_change', 'MasVnrArea\_pct\_change',  
     'BsmtFinSF1\_pct\_change', 'BsmtFinSF2\_pct\_change',  
     'BsmtUnfSF\_pct\_change', 'TotalBsmtSF\_pct\_change', '1stFlrSF\_pct\_change',  
     'LowQualFinSF\_pct\_change', 'GrLivArea\_pct\_change',  
     'BsmtHalfBath\_pct\_change', 'BedroomAbvGr\_pct\_change',  
     'KitchenAbvGr\_pct\_change', 'GarageArea\_pct\_change',  
     'WoodDeckSF\_pct\_change', 'OpenPorchSF\_pct\_change',  
     'EnclosedPorch\_pct\_change', 'ScreenPorch\_pct\_change',  
     'PoolArea\_pct\_change', 'MiscVal\_pct\_change', 'MoSold\_pct\_change',  
     'SalePrice\_pct\_change', 'MSZoning\_encoded\_pct\_change',  
     'Street\_encoded\_pct\_change', 'LandContour\_encoded\_pct\_change',  
     'LandSlope\_encoded\_pct\_change', 'Neighborhood\_encoded\_pct\_change',  
     'Condition1\_encoded\_pct\_change', 'Condition2\_encoded\_pct\_change',  
     'RoofMatl\_encoded\_pct\_change', 'Exterior1st\_encoded\_pct\_change',  
     'Exterior2nd\_encoded\_pct\_change', 'ExterQual\_encoded\_pct\_change',  
     'ExterCond\_encoded\_pct\_change', 'BsmtCond\_encoded\_pct\_change',  
     'BsmtFinType2\_encoded\_pct\_change', 'Heating\_encoded\_pct\_change',  
     'CentralAir\_encoded\_pct\_change', 'Electrical\_encoded\_pct\_change',  
     'Functional\_encoded\_pct\_change', 'FireplaceQu\_encoded\_pct\_change',  
     'GarageQual\_encoded\_pct\_change', 'GarageCond\_encoded\_pct\_change',  
     'PavedDrive\_encoded\_pct\_change', 'MiscFeature\_encoded\_pct\_change',  
     'SaleType\_encoded\_pct\_change', 'SaleCondition\_encoded\_pct\_change']
* **7. Kurtosis for bell shaped curve should be 3.**

Chart, diagram

Description automatically generated

**Observations**

**1. Acceptable skewness is +/-0.65 and Right skewness for bell shaped curve is 0**

* **2. Acceptable and Outliers Prone Left skewness is observed in many columns, including:**
* YearBuilt                          -0.579204  
      YearRemodAdd                       -0.495864  
      GarageYrBlt                        -0.645078  
      GarageCars                         -0.358556  
      MSZoning\_encoded                   -1.796785  
      Street\_encoded                    -17.021969  
      LotShape\_encoded                   -0.603775
* **3. Acceptable And Outliers Prone Right Skewness is observed in most of the columns, including :**
* Id    0.026526032012241022  
      MSSubClass    1.422018988135284  
      LotFrontage    2.75080497659666  
      LotArea    10.659284548299626  
      OverallQual    0.1750824992845271
* Interpretation of the Results
* **## Based On EDA done above in two parts, I will do ftest and pvalue test on these seemingly weak indicators based primarily on skewness, kurtosis and multicollinearity:**
* **Misc Val (Skewness & Kurtosis)**
* **Street Encoded (Skewness & Kurtosis)**
* **Total Bsmt SF (Multicollinearity, Skewness & Kurtosis)**
* **Most of the biased features have strong explanatory power in terms of correlation with feature, skewness or kurtosis and hence can be filtered in ensemble method of feature selection by p value and constant variance threshold.**

**ANOVA Test On Selected Features**

* **Ftest score should be greater than 1 and p value should be less than 0.05, to determine to keep these features for further analysis**

Graphical user interface, text, application, email

Description automatically generated

* **Based on above analysis:**
* **1. There are many outliers in the data.**
* **2. Strong multicolliearity features are important for prediction because there f test and p value are acceptable. This means that the amount of multicollineaity is insignificant and removing the feature will impact the model much.**
* **3. Extereme leptokurtic and right skewed features are also relatively significant based on f test and p test.**
* **Hence, as a solution, feature scaling will do a better job in explaining the dependent variable than removing whole columns.**

Outliers Transformation

Table

Description automatically generated

Min Max Scaler Transformation

Table

Description automatically generated

Variance Inflation Factor

**After passing through vif test to remove multicollinearity, only 33 features seem to be low bias with seemingly strong explanatory power**

**Concluding Points:**

* **As part of data handling, I have closely analyzed features with high outliers (by analyzing box plots, dist plots, variable plot and scatter plots).**
* **I have removed features with multicollinearity by analyzing correlation, correlation heatmaps and variance inflation factor.**
* **I have done ANOVA testing, wherever, applicable to weigh importance against bias.Hence, the model can be expected to be low variance and low bias model**

**CONCLUSION**

* Key Findings and Conclusions of the Study
* **Based on above analysis:**
* **1. There are many outliers in the data.**
* **2. Strong multicolliearity features are important for prediction because there f test and p value are acceptable. This means that the amount of multicollineaity is insignificant and removing the feature will impact the model much.**
* **3. Extereme leptokurtic and right skewed features are also relatively significant based on f test and p test.**
* **Hence, as a solution, feature scaling will do a better job in explaining the dependent variable than removing whole columns.**

Outliers Transformation

Table

Description automatically generated

Min Max Scaler Transformation

Table

Description automatically generated

Variance Inflation Factor

**After passing through vif test to remove multicollinearity, only 33 features seem to be low bias with seemingly strong explanatory power**

**Concluding Points:**

* **As part of data handling, I have closely analyzed features with high outliers (by analyzing box plots, dist plots, variable plot and scatter plots).**
* **I have removed features with multicollinearity by analyzing correlation, correlation heatmaps and variance inflation factor.**
* **I have done ANOVA testing, wherever, applicable to weigh importance against bias.Hence, the model can be expected to be low variance and low bias model**
* Learning Outcomes of the Study in respect of Data Science

Visualizations and data cleaning convert a whole complex and messy dataset into insightful and interesting representation, which make it easier to reach the core of the problem and solve it.

The best model is Huber Regressor With Default Hyper Parameter tuning, the most challenging part in models development process was to reduce overfitting and that is why I have applied ensemble methods on base estimators... Huber Regressor provided the best framework to reduce overfitting.

* Limitations of this work and Scope for Future Work

Further optimatization can be obtained by applying deep learning solutions. Since, it requires very high RAM capacity, it could not be displayed in jupyter notebook... I would like to update Google Colab Notebook for future projects, if acceptable... That can help me to submit a completely optimized model.