Real Estate House Price Prediction

6th September 2021

Table of Contents

Engagement Background

A Leading Real Estate MNC is expanding it’s foot-print in a competitive market and hence needs a **Robust Model** to be developed and recommended for predicting the Sale price of a house given a set of independent variables / predictors.

The Organization has engaged the services of Sai Science Pte Ltd for the same.

Data Science Process

	Identify Business Problem	Explore Data	Model & Validate	Communicate	Implement
Key Tasks Performed in each Stage	➢ *Ask Business Question* ➢ *Identify stakeholders* ➢ *Define Objective, strategy, Value & Budget* ➢ *Establish Success Criteria* ➢ *Translate Business Question to a Data Question*	➢ *Identify Data sources / points* ➢ *Confirm Tools & Software to be used* ➢ *Load Cleanse & Transform data where required* ➢ *Explore Data – Analysis & Visualization* ➢ *Ascertain Quality, completeness & ability to answer questions / Identify patterns & Trends*	➢ *Feature Engineering* ➢ *Model Development* ➢ *Evaluation* ➢ *Select Model, apply & Validate*	➢ *Communicate to stakeholders to obtain Business Buy-in* ➢ *Present Findings & Recommendations*	➢ *Develop end-to-end solution* ➢ *Build, test, deploy and monitor*
Outcome from each Stage	➢ *Project Charter – (Business Background, Scope, Personnel, Metrics / Objectives* ➢ *Project Plan*	➢ *Detailed Specifications – Profiling, Cleansing, Transformation etc.* ➢ *Build / Code components to perform step 1* ➢ *Cleaned / Tidy data* ➢ *Data summary Report*	➢ *Feature Engineering – Methods explored & Used* ➢ *Model*	➢ *Powerpoint Presentation* ➢ *Model (Jupyter Notebook)*	➢ *Deploy model to production upon acceptance* ➢ *Monitor and refine based on actual data.*

Iterative

Identify Business Problem → Explore Data → Model & Validate → Communicate → Implement

Identify Business Problem

A **<u>Leading Real Estate MNC</u>** is expanding it's foot-print in a competitive market and hence needs a Robust Model to predict the Sale price of house based on a number of independent variables. The Organization has engaged the services of Sai Science Pte Ltd and needs the following done

- Based on the data provided, an **Exploratory Data Analysis**
- To demonstrate a few methods to do **Feature selection / Extraction**
- To **build various models** with one of the Feature Selection methods demonstrated
- **Recommend the best Model**

Identify Business Problem – Stakeholders

Key Client Stakeholders	Vendor Stakeholders
Client Engagement Director	Engagement Director
Client Project Manager	Project Manager
Client BA / SME	Lead Data Scientist
Business Sponsor – Head of New Business	Data Architect
Technology Sponsor – Head of Technology	Developers

Key Assumptions

- The client team would make themselves available to clarify any questions on the data set. (2 sessions of 2 hours each have been planned to tackle such questions)

- If there is a change in the # of Features, the model needs to be re-trained & validated.

- As discussed, and agreed upfront, this is a **3-weeks** engagement

- The data set is complete and a significant representation.

- The output will be the Model **(Jupyter notebook) & a Powerpoint Presentation**

Understanding the data

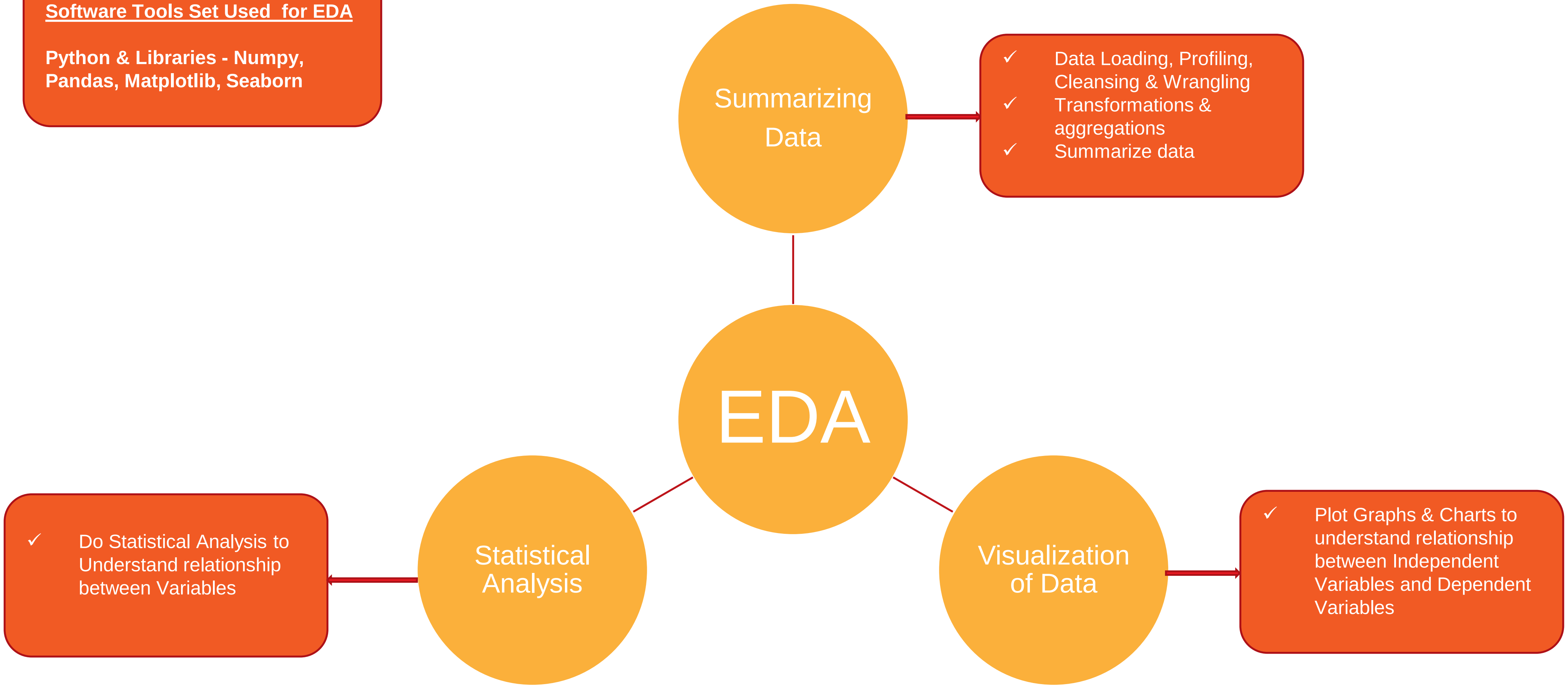
- ✓ **Data Source –** CSV file (Comma separated Values)
- ✓ **# of records –** 1460
- ✓ **# of Features / Variables –** 81
- ✓ **Type of Data – Real Estate –** To predict the Sale Price of house based on a set of independent variables / Predictors

*****House Price Data Description*** **–** *Link to the Metadata of the Dataset provided*

Explore Data – Key Components of EDA Considered

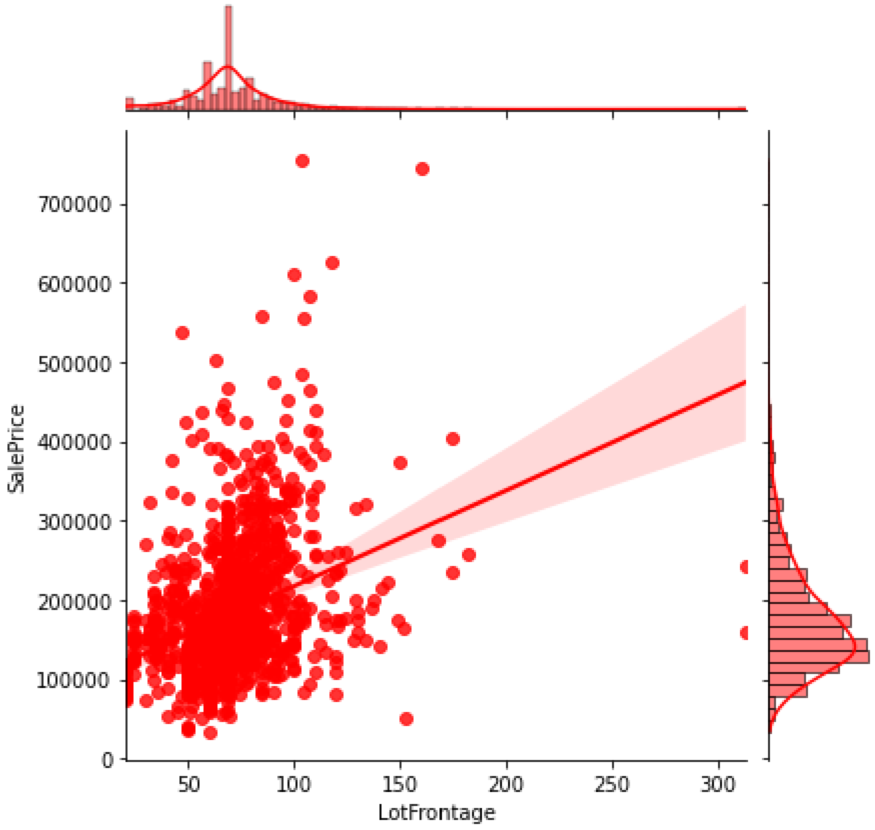
Software Tools Set Used for EDA

Python & Libraries - Numpy, Pandas, Matplotlib, Seaborn



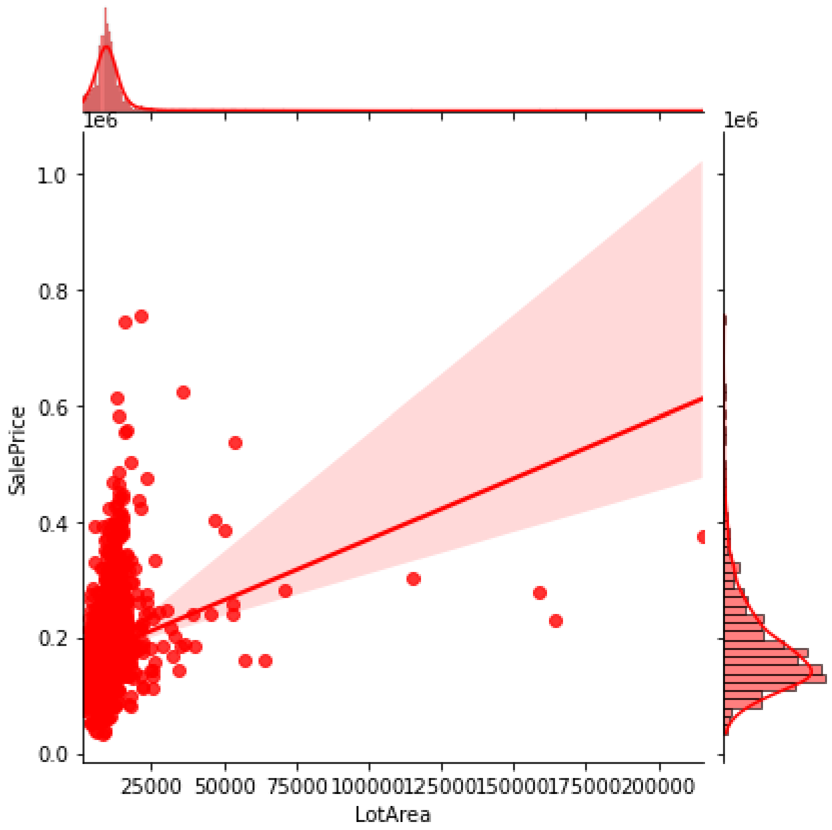
Explore Data – Key Numerical Data (Visualization)

Linear Feet of street connected to property Vs Sale Price



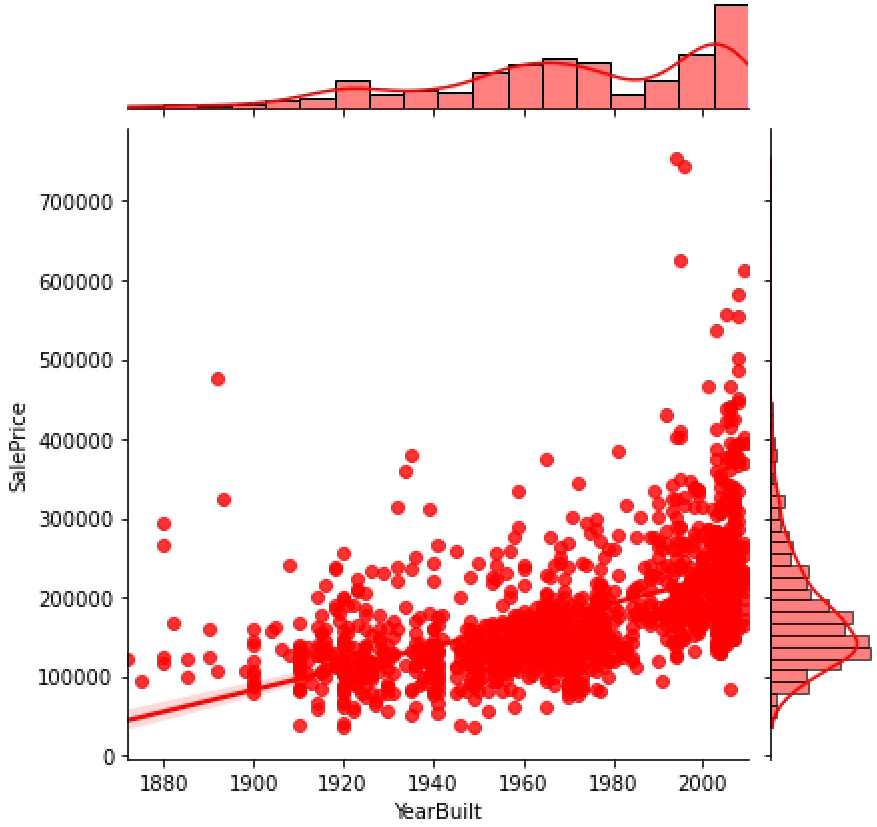
- ➢ **Sale Price Increases with Area**

Lot size in square feet Vs Sale Price



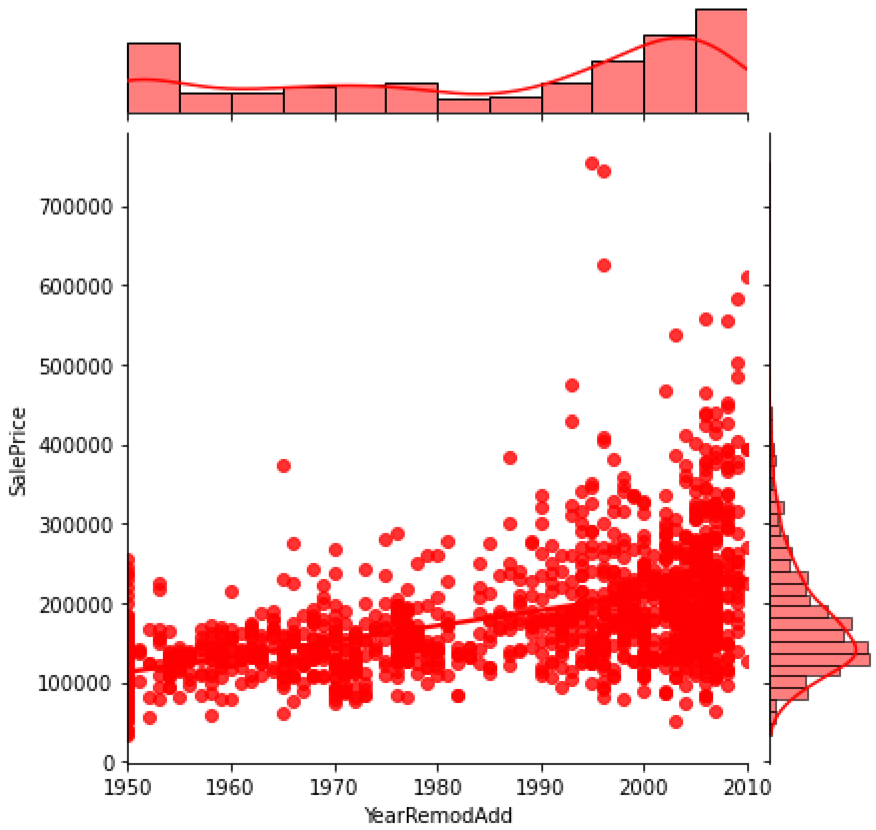
- ➢ **Sale Price Increases with Area**

Year built Vs Sale Price



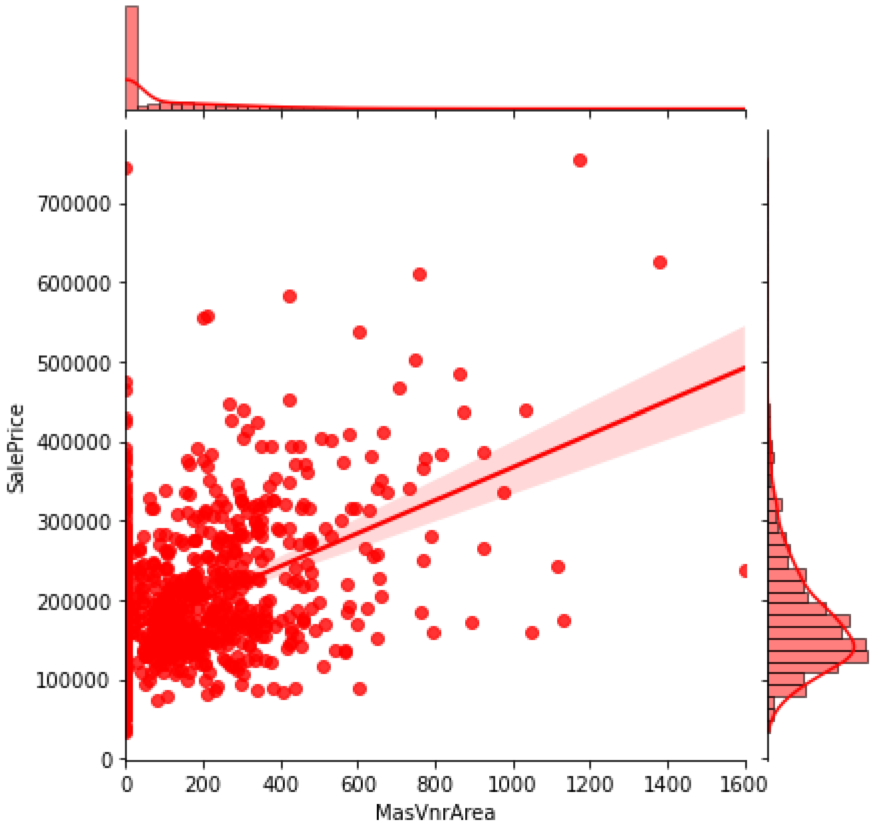
- **Sale Price Increases for newer houses**

Year Re-modelled Vs Sale Price



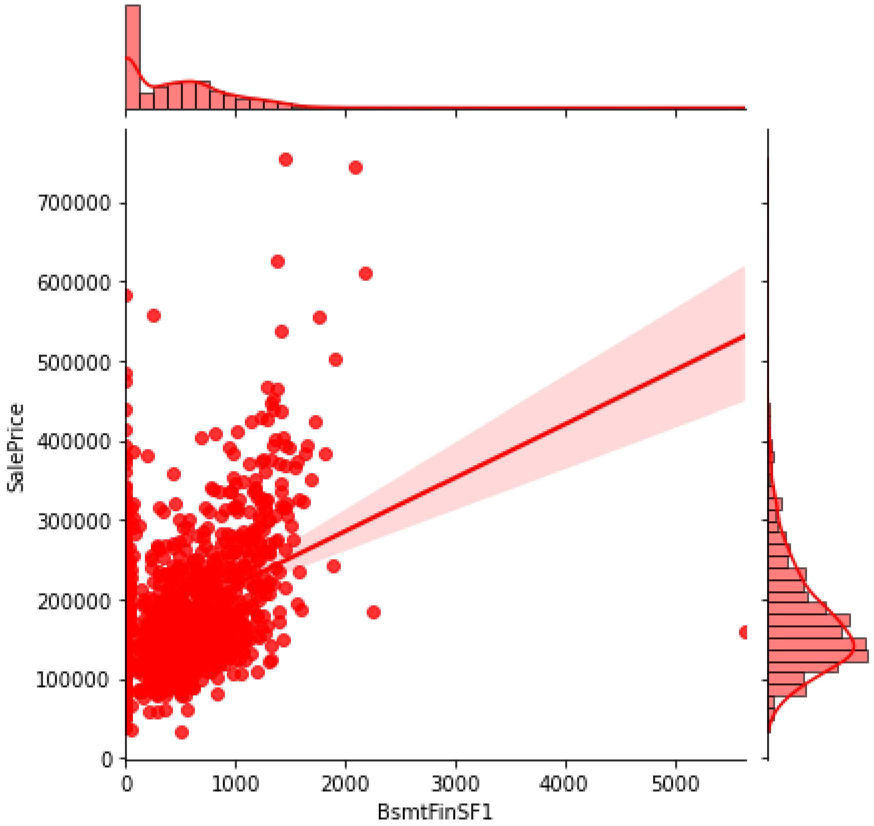
- Sale Price Increases for newer houses

Masonry Veneer Area in Sqft Vs Sale Price



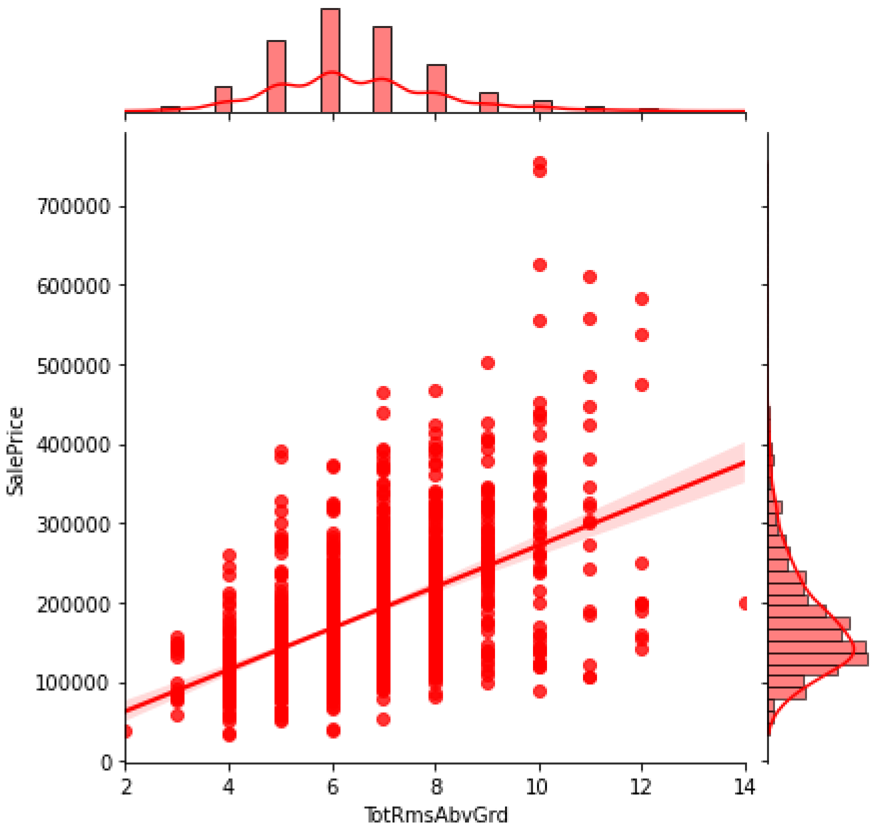
- ➢ **Sale Price Increases with area**

Type1 Finished Area in Sqft Vs Sale Price



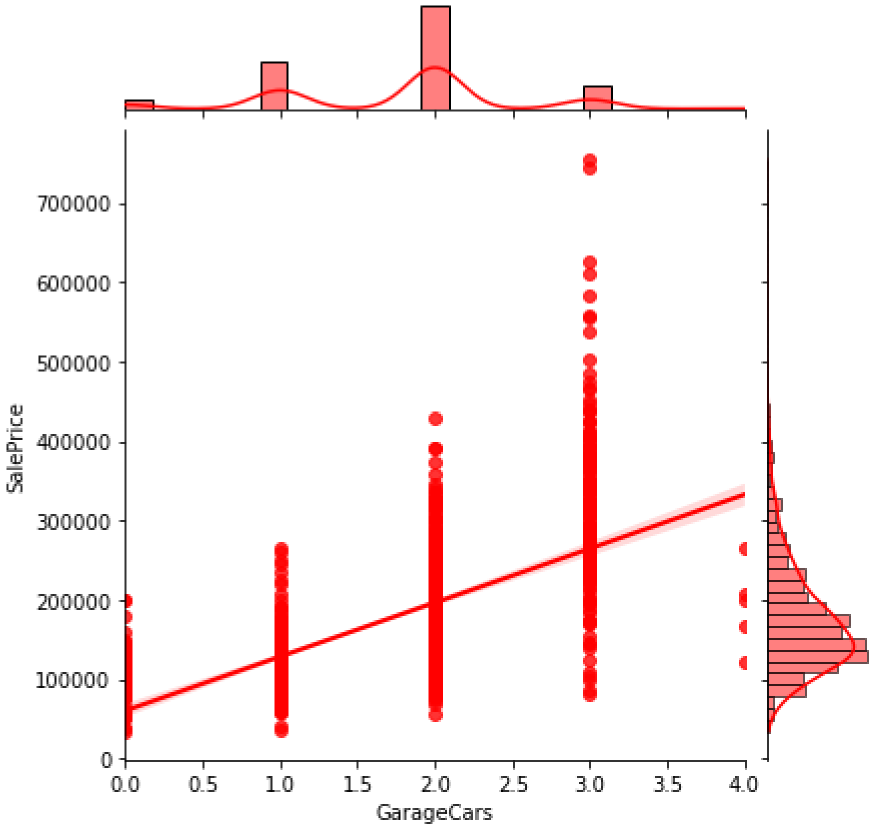
- Sale Price Increases with area

Total rooms above grade Vs Sale Price



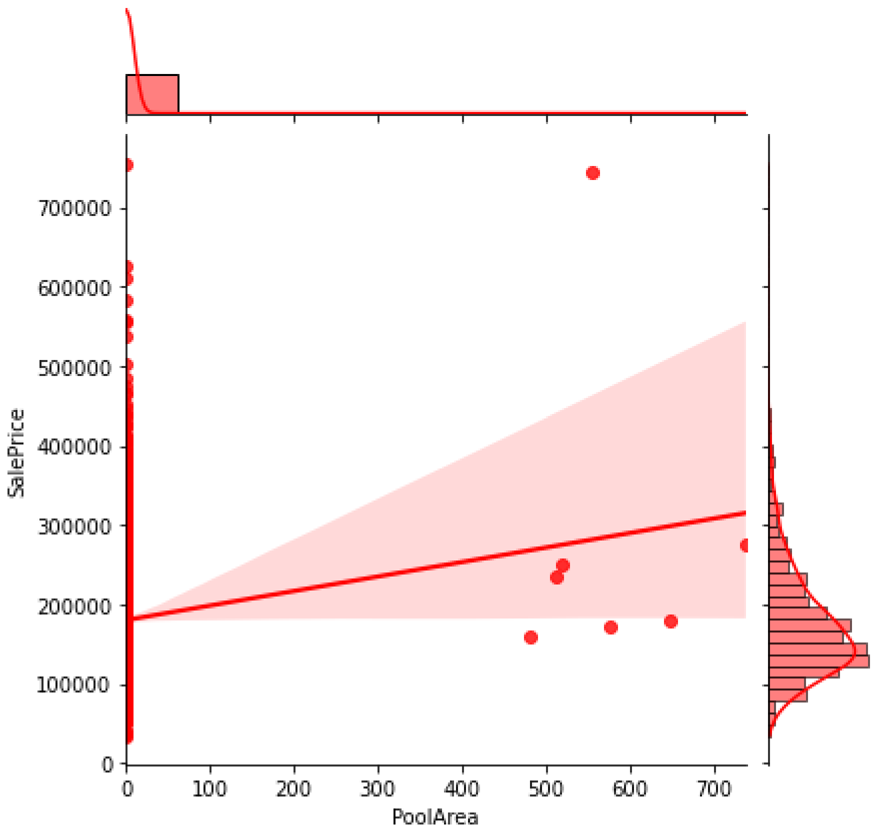
- ➢ **Sale Price Increases as the # of rooms above grade increases**

Size of Garage in Car Capacity Vs Sale Price



- **Sale Price Increases as the size of Garage increases**

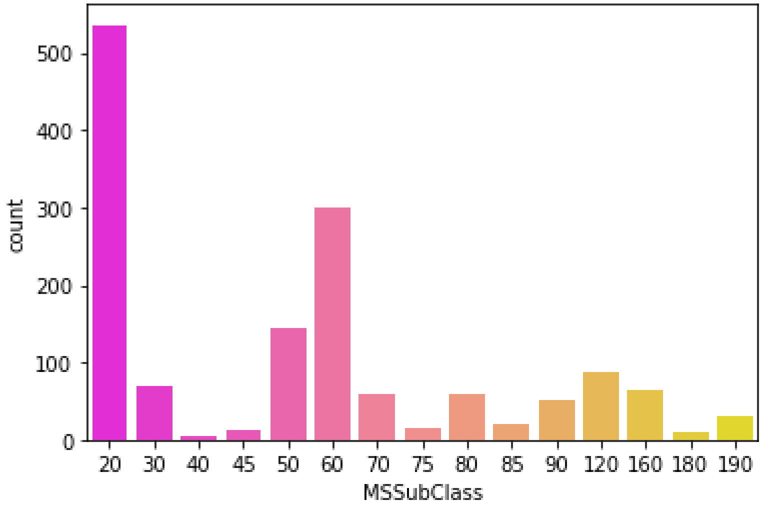
Pool Area Vs Sale Price



- ➢ **Sale Price increases with pool area**

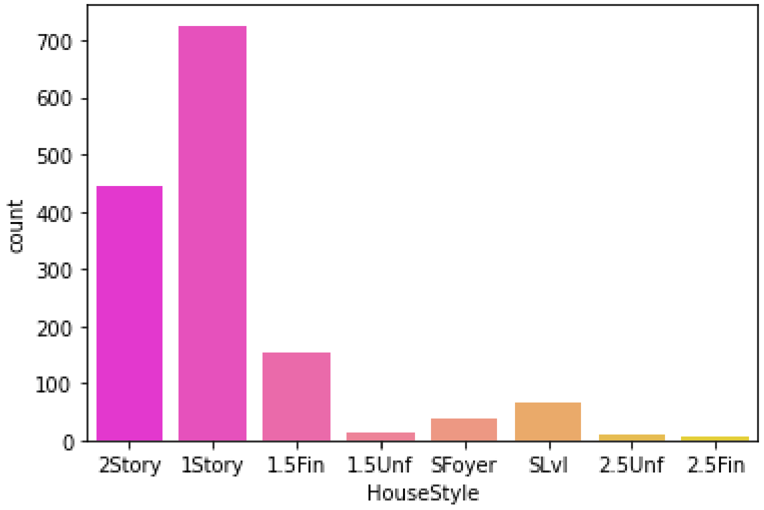
Explore Data – Key Categorical Data (Visualization)

Countplot for MSSubClass – Identifies type of dwelling involved in Sale



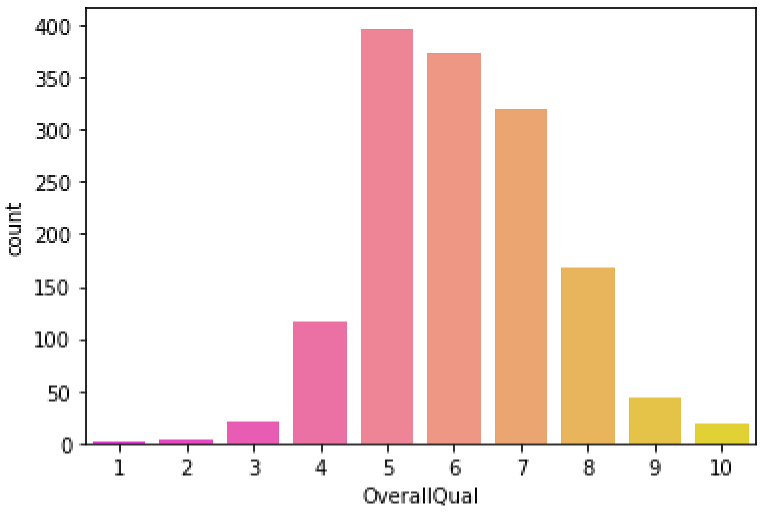
- 20 1-STORY 1946 & NEWER ALL STYLES
- 30 1-STORY 1945 & OLDER
- 40 1-STORY W/FINISHED ATTIC ALL AGES
- 45 1-1/2 STORY - UNFINISHED ALL AGES
- 50 1-1/2 STORY FINISHED ALL AGES
- 60 2-STORY 1946 & NEWER
- 70 2-STORY 1945 & OLDER
- 75 2-1/2 STORY ALL AGES
- 80 SPLIT OR MULTI-LEVEL
- 85 SPLIT FOYER
- 90 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

Countplot for 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

Countplot for OverallQual – Rates the Overall material & Finish of house



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

Model Approach

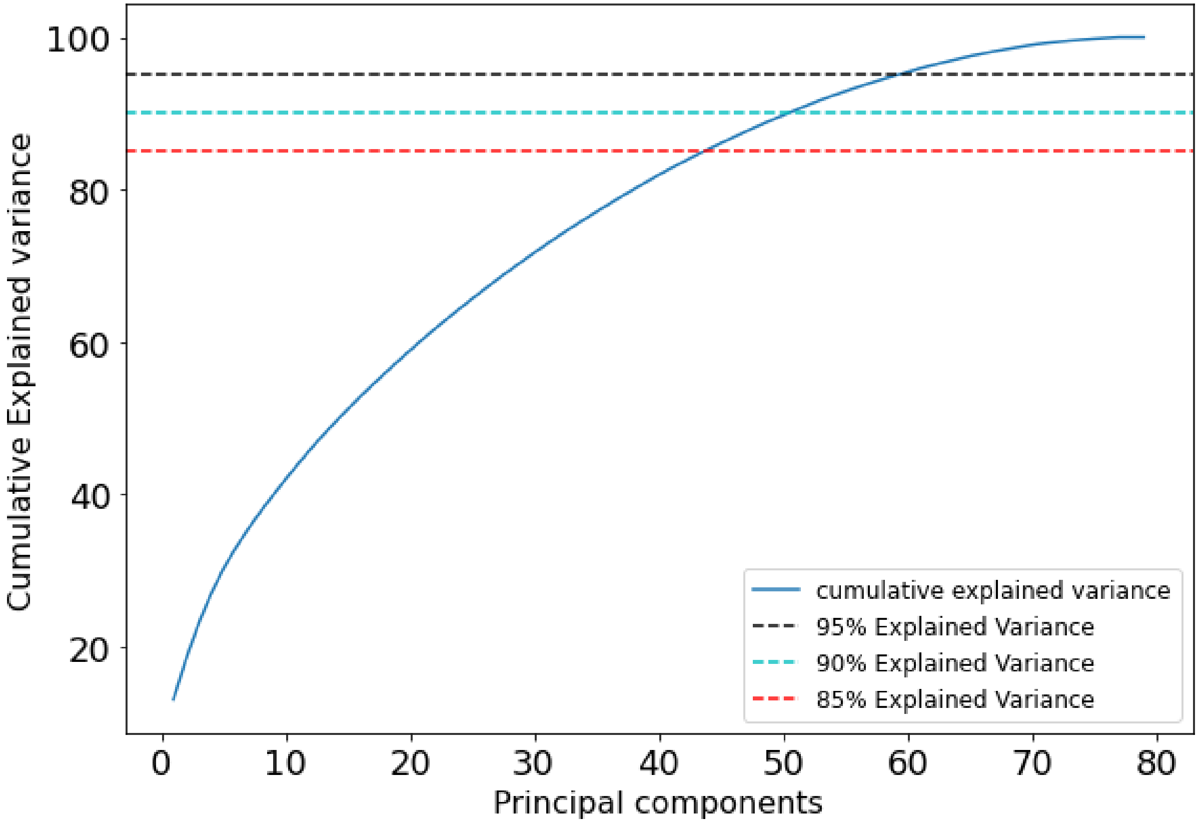
1 Feature Engineering → 2 Build Model → 3 Evaluate & Recommend

Train_Test_Split – 0.3

- Principal Component Analysis
- Correlation Plots
- Multicollinearity - VIF
- Forward Feature selection
- **Backward Elimination p-value**
  (Used in current Model)

- Linear Regression
- Lasso Regression
- Ridge regression
- Elastic Net Regression

Feature Engineering – Method 1 - PCA



Number of Features before PCA : **79**
Number of Features after PCA : **51**

Method 1 - Principal Component Analysis (PCA) is a common **feature extraction method** in data science.

This is a feature extraction method where **we create new features known as Principal Components**, and these features are not present in our original feature set but retains the maximum variance of the original dataset. These features are not interpretable .

Given the current dataset, as shown in the graph, the original number of Features are reduced from **79** to **51** after PCA is applied.

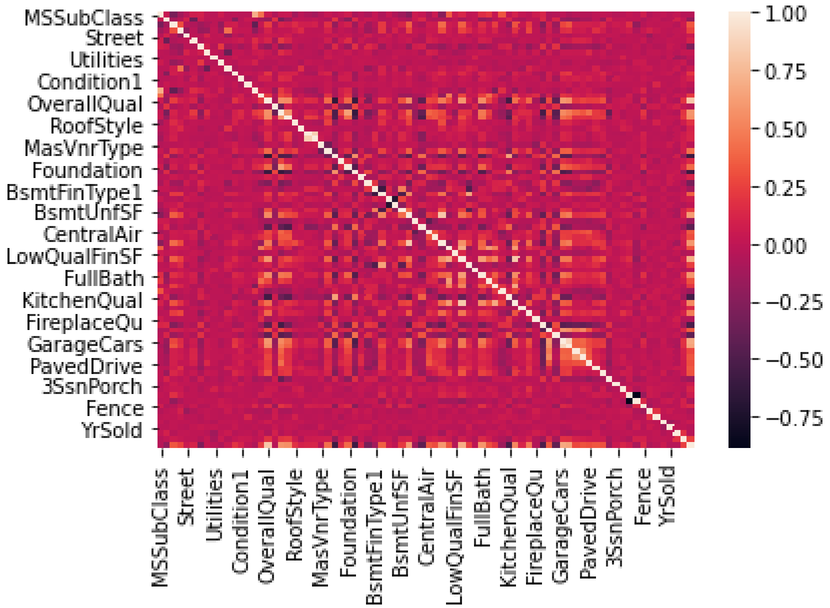
Key Parameters used in the algorithm
n_components = .90 means that scikit-learn will choose the minimum number
of principal components such that 90% of the variance is retained.

As shown in the graph, 51 Features correspond to 90% of explained Variance

Feature Engineering – Method 2 – Correlation of Target against Predictors

Method 2 – This used the correlation of the Target variable with the Predictors

In this case, we have selected only those Features, that have a correlation greater than 20% with the Target Variable.



ExterQual	OpenPorchSF
BsmtQual	GarageCond
KitchenQual	WoodDeckSF
GarageType	LotFrontage
HeatingQC	GarageYrBlt
GarageFinish	CentralAir
BsmtExposure	BsmtFinSF1
LotShape	Foundation
Neighborhood	MasVnrArea
BedroomAbvGr	Fireplaces
HouseStyle	TotRmsAbvGrd
BsmtUnfSF	YearRemodAdd
BsmtFullBath	YearBuilt
SaleCondition	FullBath
LotArea	1stFlrSF
GarageQual	TotalBsmtSF
Electrical	GarageArea
PavedDrive	GarageCars
HalfBath	GrLivArea
2ndFlrSF	OverallQual

Feature Engineering – Method 3 – Using VIF

Method 3 – Finding the Multicollinearity between the Predictors using VIF. A high VIF is indicative of collinearity.

VIF is a direct measure of how much variance of the coefficient is being inflated due to multicollinearity. The cut-off is taken as 5 or10 depending on the scenario as in a VIF > 5 or > 10 indicates a strong multicollinearity. **(Below Sample O/p from Jupyter Notebook)**

	vif
MSSubClass	5.877637
MSZoning	1.508787
LotFrontage	1.758253
LotArea	1.639578
Street	1.183140
...	...
MiscVal	1.113369
MoSold	1.077520
YrSold	1.096966
SaleType	1.163168
SaleCondition	1.212918

Feature Engineering – Method 4 – Forward Feature selection

<u>Method 4</u> – Forward Feature selection

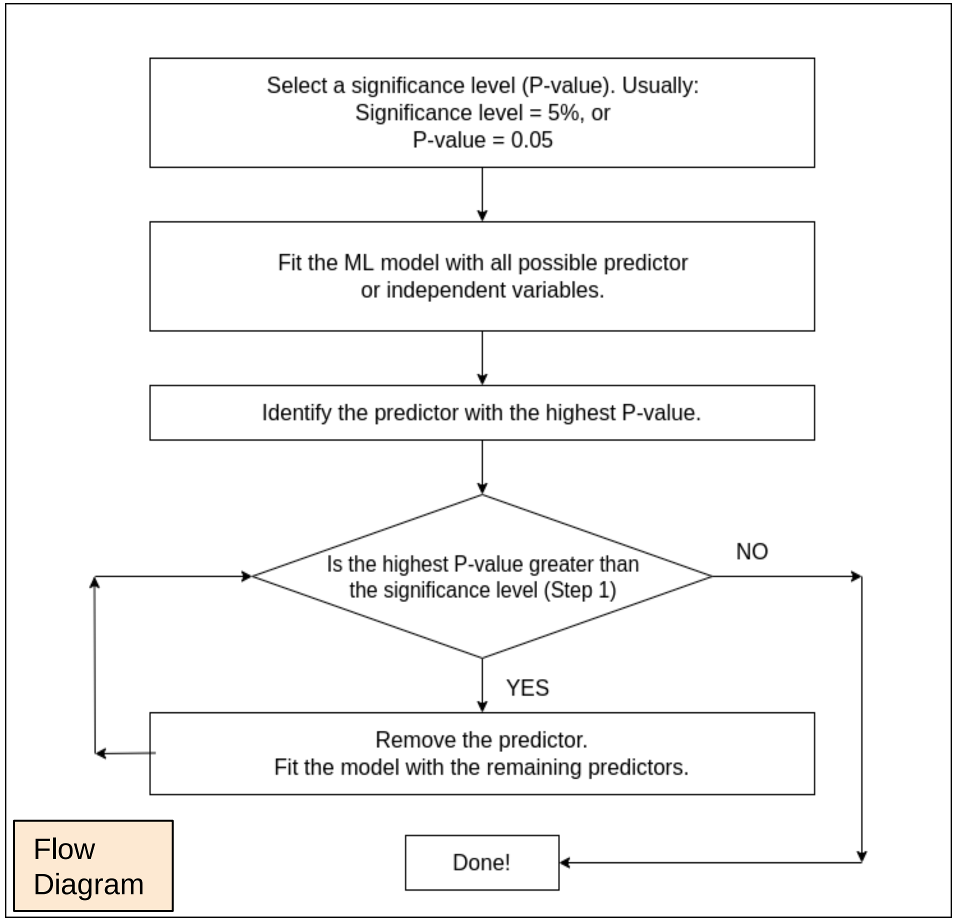
- Train Model using each feature individually and check the performance
- Choose the variable that gives the best performance
- Repeat the process and add one variable at a time
- Variable producing the highest improvement is retained
- Repeat the process till the point there is no significant improvement in the Model Performance

<u>For the Dataset, we have below is the resultant set of features obtained using Forward Feature Selection:</u>

<u>Resulting features:</u>
OverallQual, GrLivArea, YearBuilt, OverallCond, GarageCars, TotalBsmtSF, Fireplaces, BsmtFullBath, BldgType, SaleCondition, KitchenQual, CentralAir, LotArea, ScreenPorch, BsmtFinType1, HeatingQC, WoodDeckSF, PoolArea, Functional, FullBath, Condition2, Street, MSZoning, BsmtQual, PavedDrive, LotShape, FireplaceQu, GarageType, YrSold, YearRemodAdd, TotRmsAbvGrd, EnclosedPorch, LandSlope, HouseStyle, BsmtExposure, ExterCond, MiscFeature, LandContour, PoolQC, Alley, Neighborhood, 3SsnPorch, LotFrontage, RoofMatl, RoofStyle, HalfBath, 1stFlrSF, KitchenAbvGr, Foundation, BsmtCond, BsmtHalfBath, Utilities, SaleType

Feature Engineering – Method 5 – Backward Feature Elimination

Method 5 – Backward Feature Elimination **(This is the approach used for this Model)**



Feature Engineering – Method 5 – Backward Feature Elimination - Results

Method 5 – Backward Feature Elimination **(This is the approach used for this Model)**

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.869
Model:	OLS	**Adj. R-squared:**	0.866
Method:	Least Squares	**F-statistic:**	338.1
Date:	Sun, 29 Aug 2021	**Prob (F-statistic):**	0.00
Time:	11:54:17	**Log-Likelihood:**	750.62
No. Observations:	1460	**AIC:**	-1443.
Df Residuals:	1431	**BIC:**	-1290.
Df Model:	28		
Covariance Type:	nonrobus		

28 Variables are selected based on this approach.

Model Recommendation

1 Feature Engineering

2 Build Model

3 Evaluate & Recommend

Train_Test_Split – 0.3

- Principal Component Analysis
- Correlation Plots
- Multicollinearity - VIF
- Forward Feature selection
- **Backward Elimination P-value** (Used in current Model)

- **Linear Regression* (0.8749)**
- Lasso Regression (0.8735)
- Ridge regression (0.8741)
- Elastic Net Regression (0.8718)

- **Recommended is**
- **Linear Regression - Score of 0.87494**

Thank You