HOUSE PRICE PREDICTION ANALYSIS

JAGADEESH K

HARIKA

SRI HARI

SUMIT

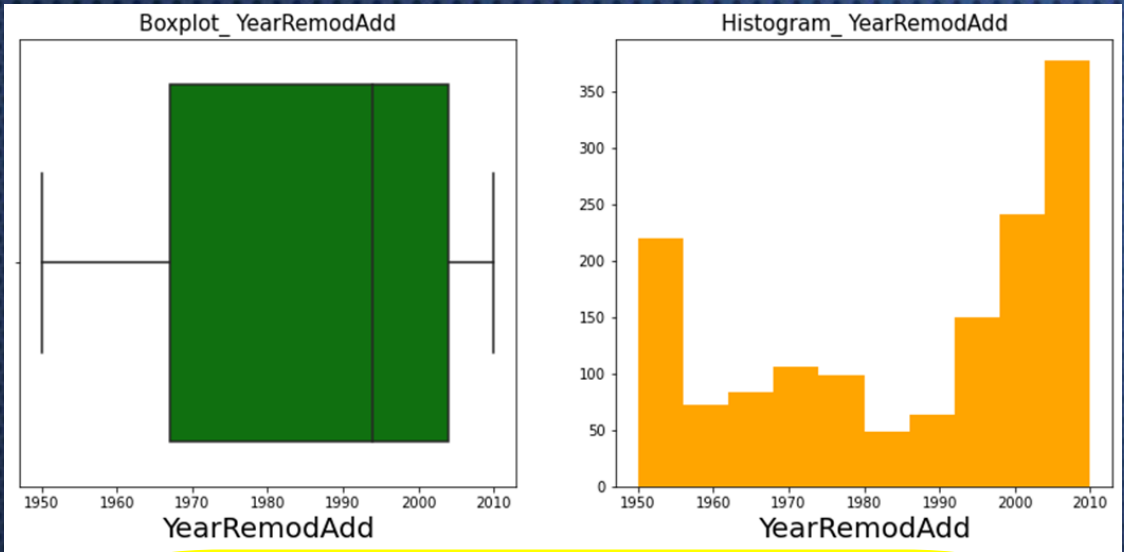
DATASET INFORMATION & PROCESS

81

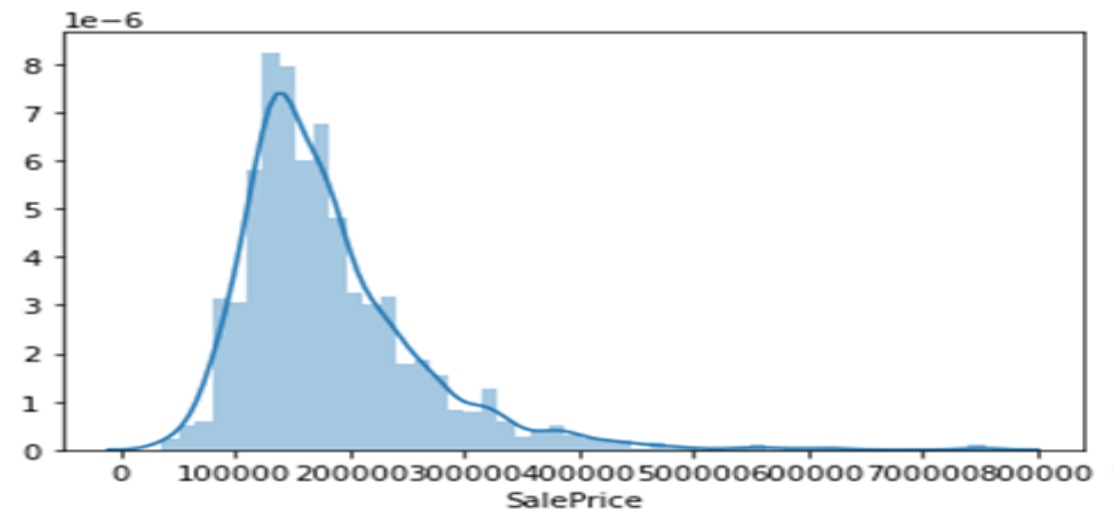
ATTRIBUTES

1. Shape of the Dataset : 1460(Rows) / 81(Attributes).
2. Response Variable : Sales Price attribute
3. Dropped the attributes, If it is having more than 70% of the data as NA's (Alley, Fence).
4. Dropped the attribute Id, It won't reflect any impact on the Response Variable.
5. Treated the attributes having Null values as per their mean, median and mode.
6. Treated the attributes having Outliers with their Inter Quartile Range.
7. Visualization done with different plots.
8. We Built the Train and Testing sets with 0.8:0.2 range.
9. We Built the Linear Regression Model in the dataset and implemented the model accuracy By Backward Elimination Process, Error Plots and selected the best 8 attributes .
10. After implementing the Linear Regression model we applied the PCA (Principal Component Analysis on those best 8 attributes.

BASIC VISUALIZATION ON THE DATASET



Plot for Checking Outliers of one attribute

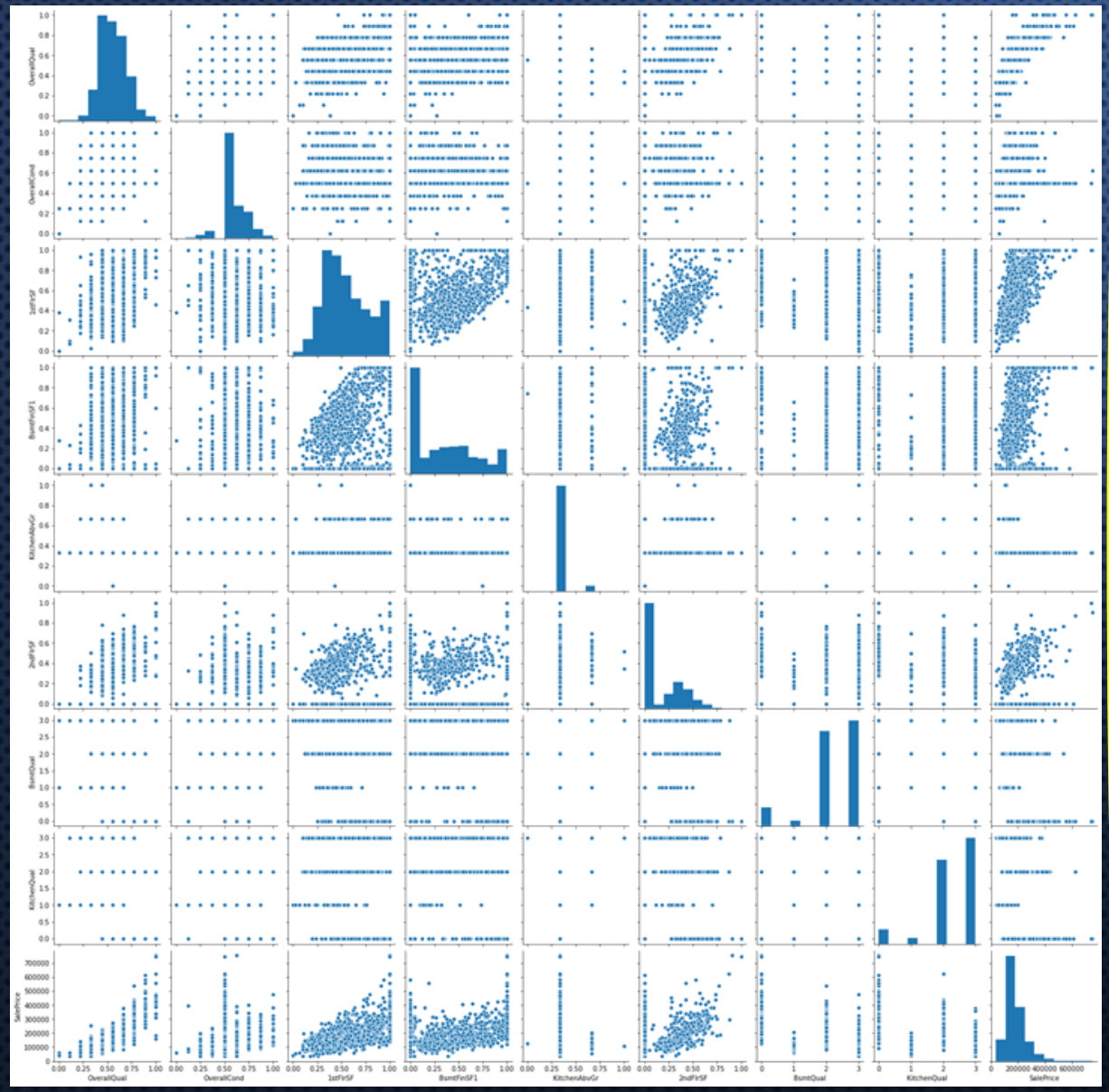


	OverallQual	OverallCond	1stFlrSF	BsmtFinSF1	KitchenAbvGr	2ndFlrSF	BsmtQual	KitchenQual	SalePrice
OverallQual	1.00	-0.09	0.46	0.21	-0.18	0.30	-0.60	-0.56	0.79
OverallCond	-0.09	1.00	-0.15	-0.04	-0.09	0.03	0.23	0.07	-0.08
1stFlrSF	0.46	-0.15	1.00	0.37	0.08	-0.23	-0.34	-0.34	0.60
BsmtFinSF1	0.21	-0.04	0.37	1.00	-0.09	-0.16	-0.22	-0.15	0.37
KitchenAbvGr	-0.18	-0.09	0.08	-0.09	1.00	0.06	0.12	0.12	-0.14
2ndFlrSF	0.30	0.03	-0.23	-0.16	0.06	1.00	-0.12	-0.14	0.32
BsmtQual	-0.60	0.23	-0.34	-0.22	0.12	-0.12	1.00	0.51	-0.62
KitchenQual	-0.56	0.07	-0.34	-0.15	0.12	-0.14	0.51	1.00	-0.59
SalePrice	0.79	-0.08	0.60	0.37	-0.14	0.32	-0.62	-0.59	1.00

Plot for Checking Co-Relation between the Final attributes

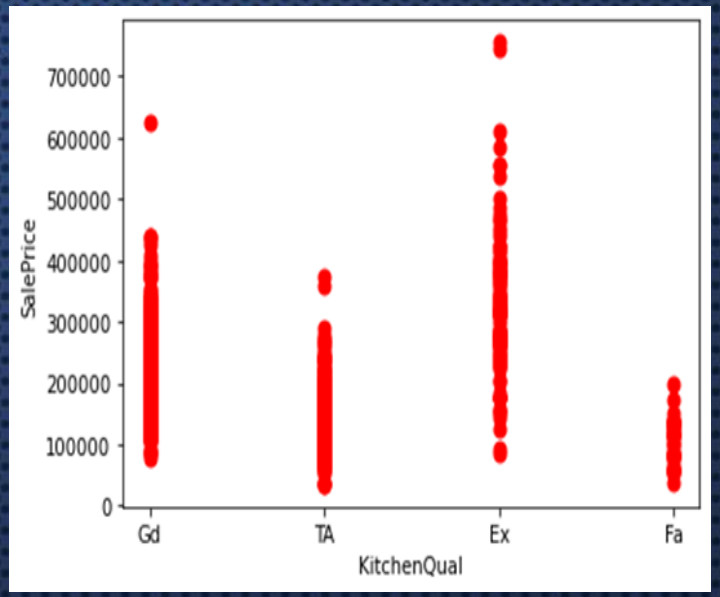
Plot for Checking Normalcy of Response attribute

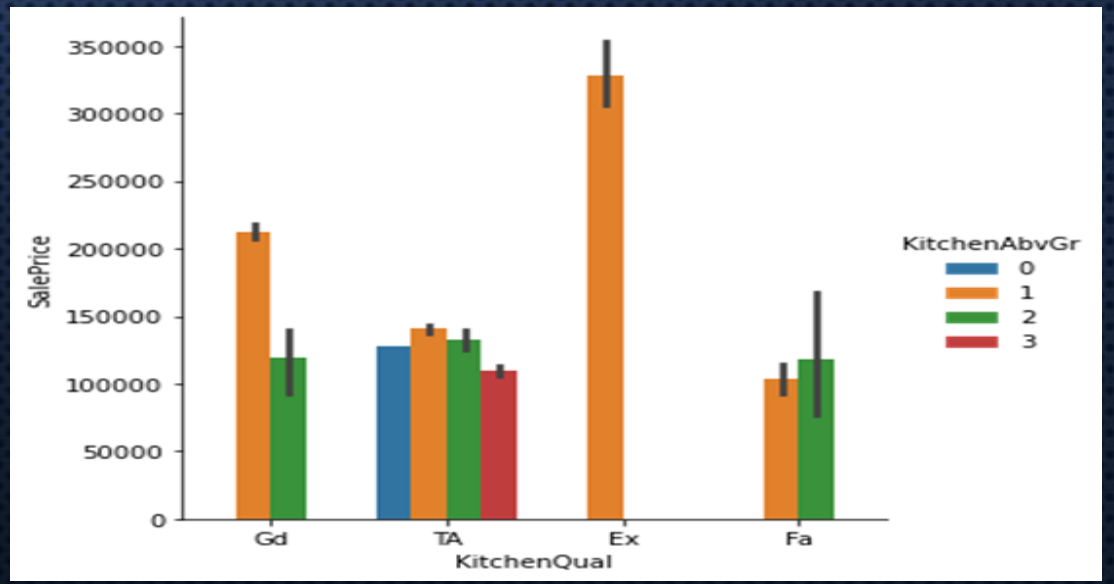
BASIC VISUALIZATION ON THE DATASET



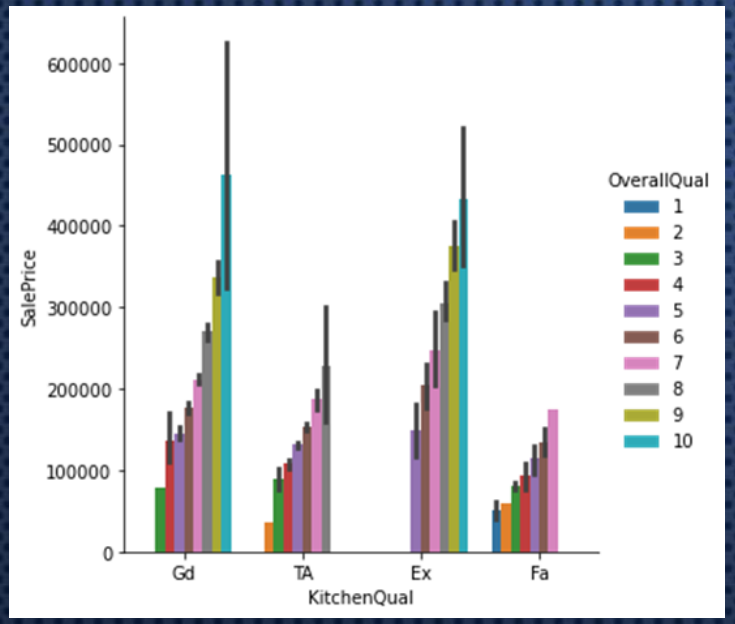
PAIR PLOT

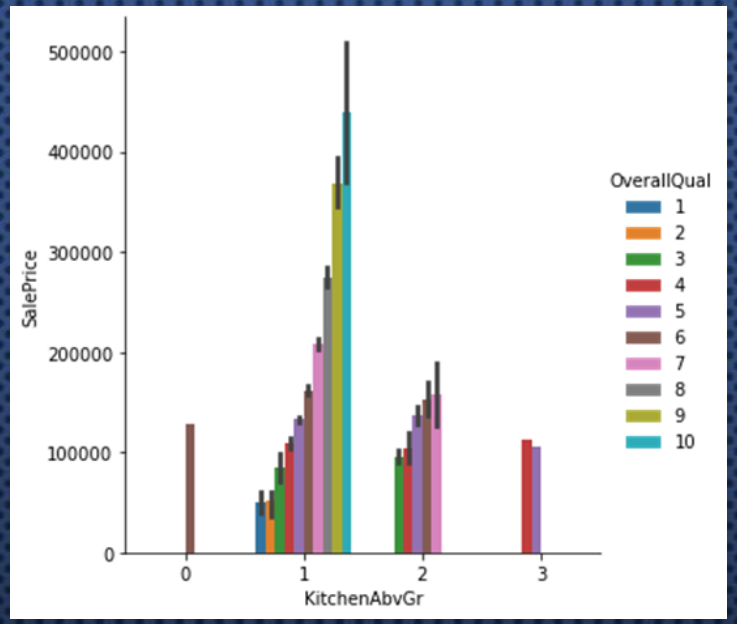
Plot for Checking Relation between the Kitchen Quality and Sales Price with respect to Kitchen above ground

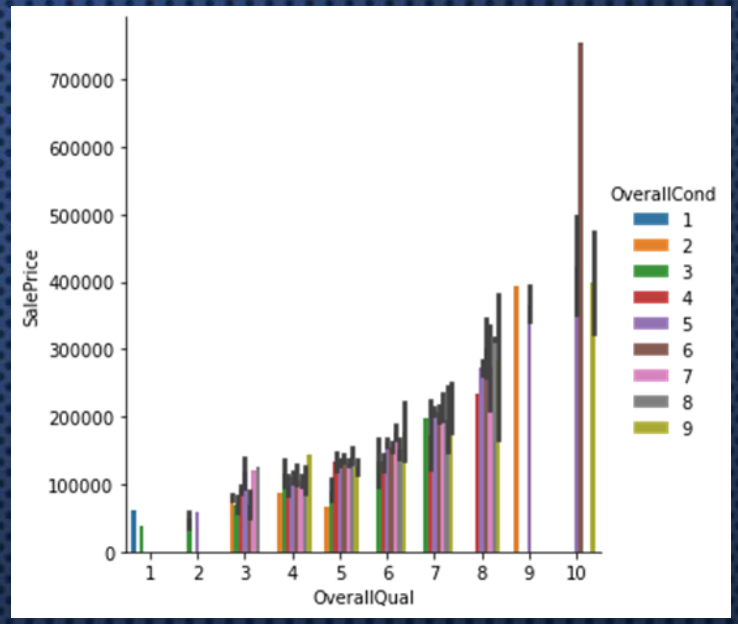




BASIC VISUALIZATION ON THE DATASET





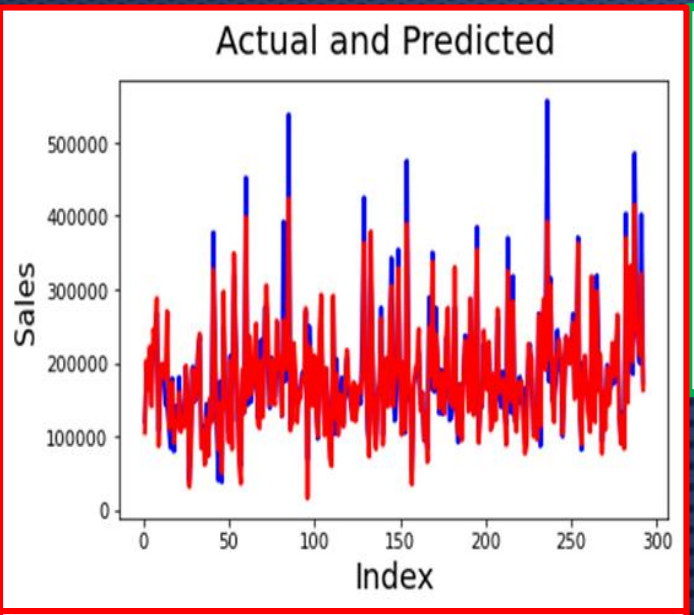


The chart is plotted between Sales Price and Kitchen Quality with respect to Overall Quality of the house.

The chart is plotted between Sales Price and Kitchen above ground area with respect to Overall Quality of the house.

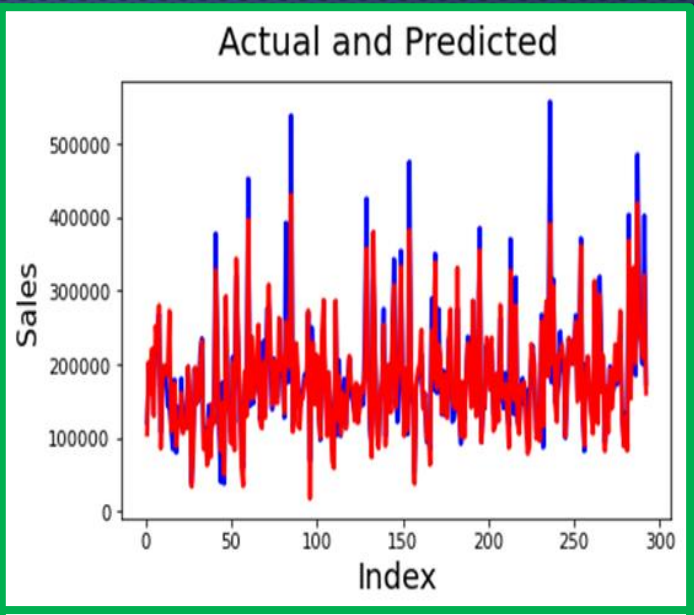
The chart is plotted between Sales Price and Overall Quality of house with respect to Overall condition of the house.

LINEAR REGRESSION MODEL ON THE DATASET



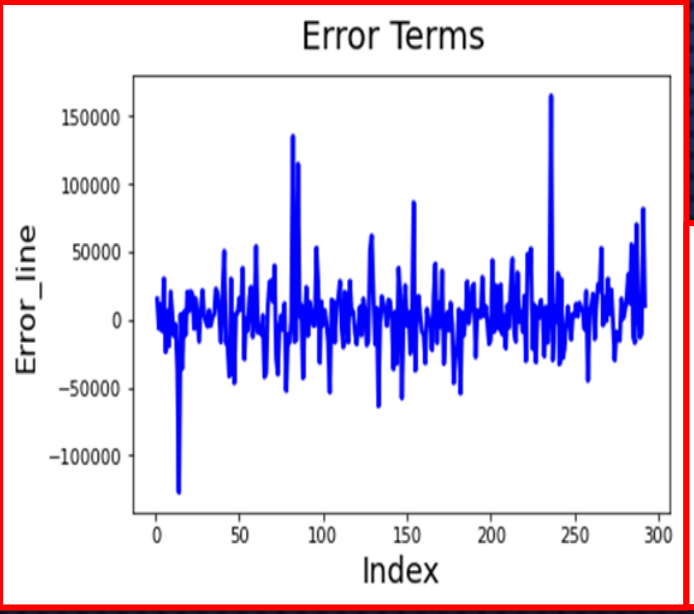
OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.856
Model:	OLS	Adj. R-squared:	0.846
Method:	Least Squares	F-statistic:	87.93
Date:	Sun, 24 Jan 2021	Prob (F-statistic):	0.00
Time:	15:47:25	Log-Likelihood:	-13700.
No. Observations:	1168	AIC:	2.755e+04
Df Residuals:	1093	BIC:	2.793e+04
Df Model:	74		
Covariance Type:	nonrobust		



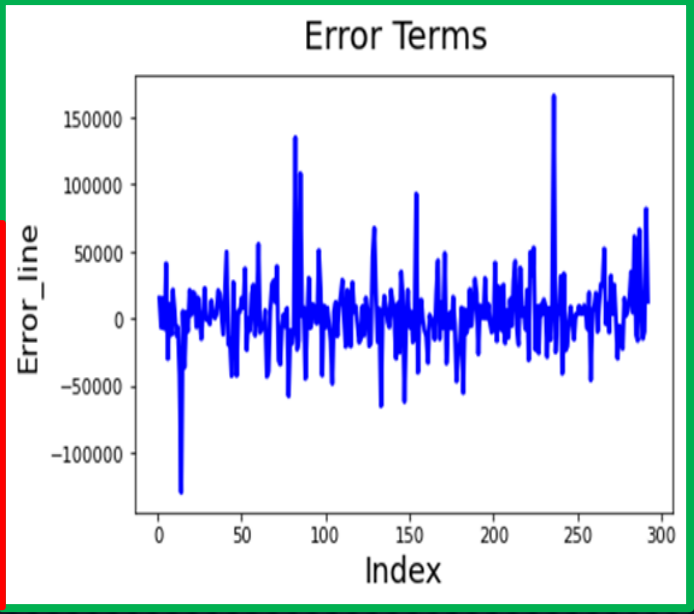
MODEL : - 1

MODEL : - 2

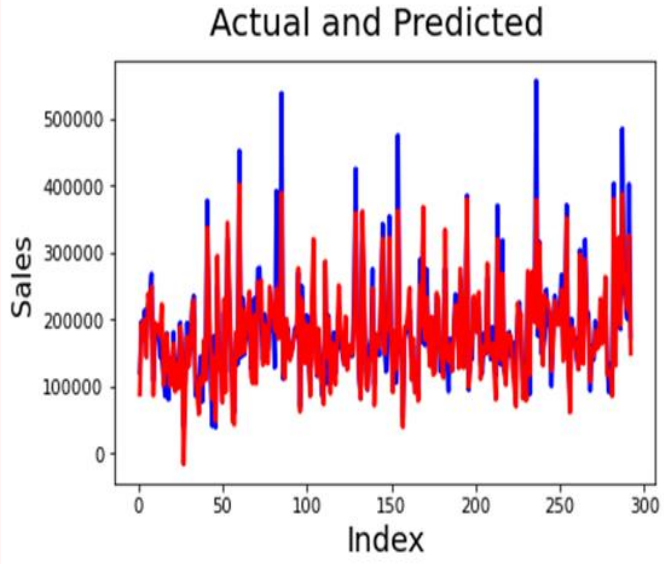


OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.854
Model:	OLS	Adj. R-squared:	0.845
Method:	Least Squares	F-statistic:	94.66
Date:	Sun, 24 Jan 2021	Prob (F-statistic):	0.00
Time:	15:47:33	Log-Likelihood:	-13708.
No. Observations:	1168	AIC:	2.755e+04
Df Residuals:	1099	BIC:	2.790e+04
Df Model:	68		
Covariance Type:	nonrobust		

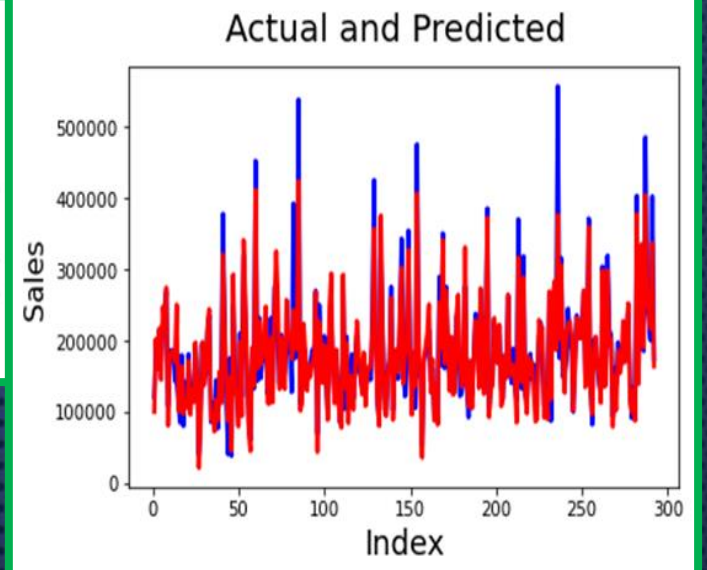


LINEAR REGRESSION MODEL ON THE DATASET

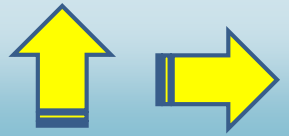


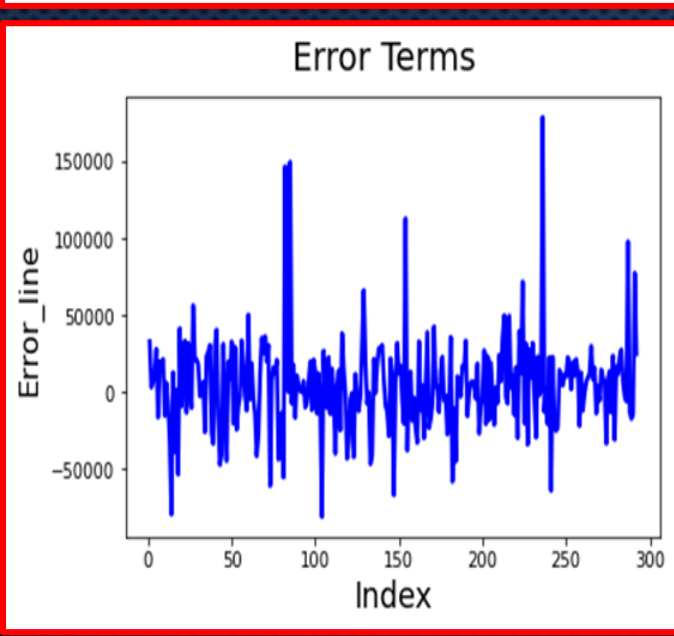
OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	258.8
Date:	Sun, 24 Jan 2021	Prob (F-statistic):	0.00
Time:	15:47:35	Log-Likelihood:	-13766.
No. Observations:	1168	AIC:	2.758e+04
Df Residuals:	1144	BIC:	2.770e+04
Df Model:	23		
Covariance Type:	nonrobust		



MODEL : - 3

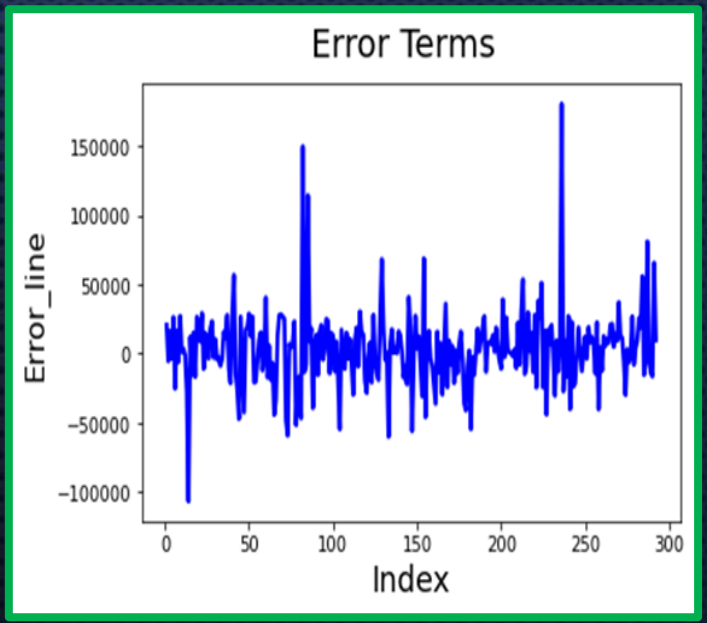




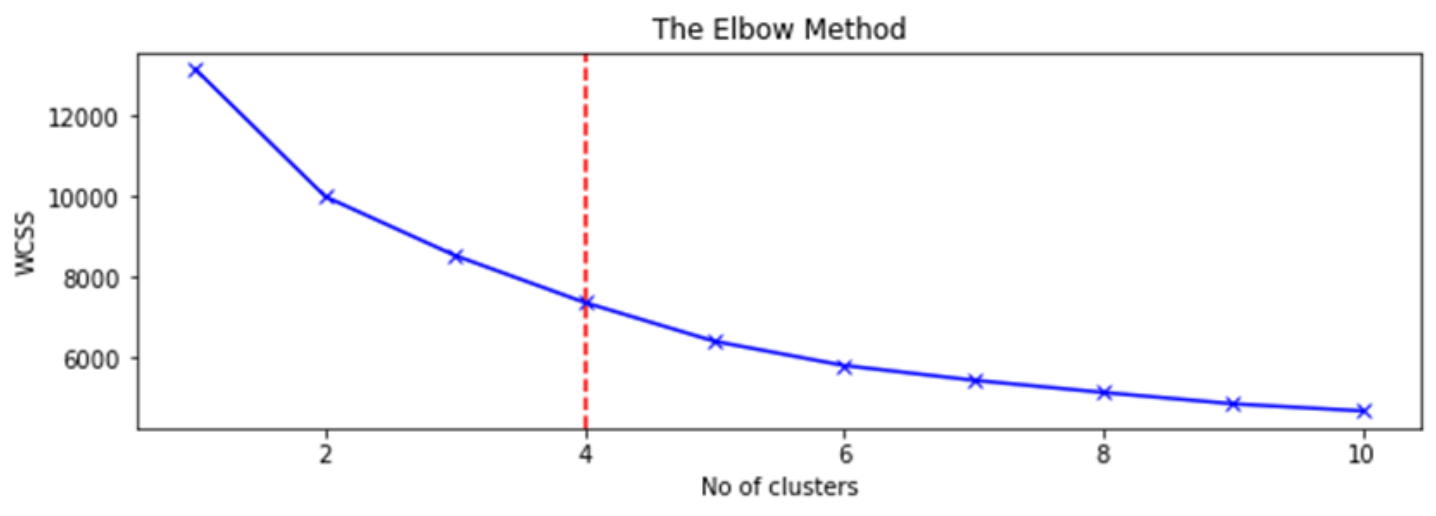
MODEL : - 7

OLS Regression Results

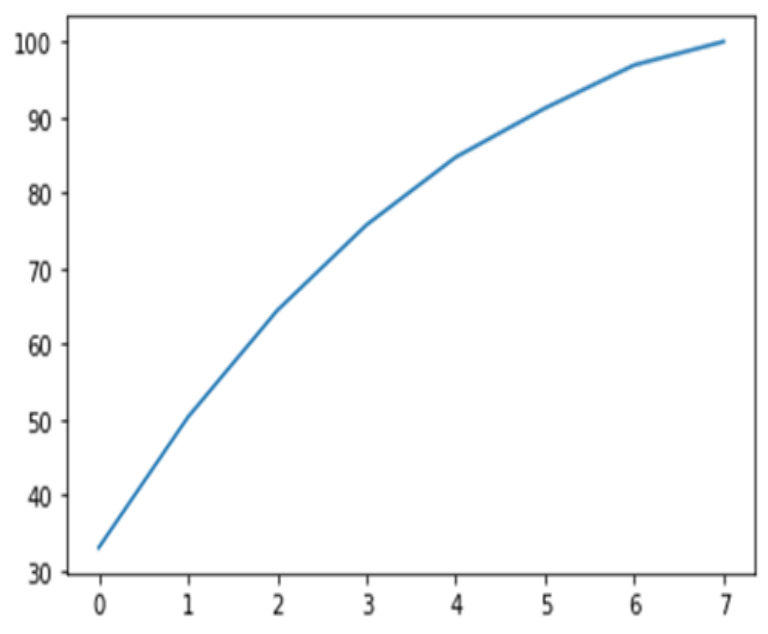
Dep. Variable:	SalePrice	R-squared:	0.807
Model:	OLS	Adj. R-squared:	0.806
Method:	Least Squares	F-statistic:	606.5
Date:	Sun, 24 Jan 2021	Prob (F-statistic):	0.00
Time:	15:47:42	Log-Likelihood:	-13871.
No. Observations:	1168	AIC:	2.776e+04
Df Residuals:	1159	BIC:	2.781e+04
Df Model:	8		
Covariance Type:	nonrobust		



CLUSTERING AND PCA ON THE DATASET



ELBOW PLOT (OPTIMIZATION PLOT) FOR SELECTING THE K VALUE FOR CLUSTERING



PCA MODEL APPLIED AND FOUND MORE THAN 75 % ON 5 ATTRIBUTES.

APPLIED LINEAR REGRESSION ON ATTRIBUTES SELECTED FROM PCA.

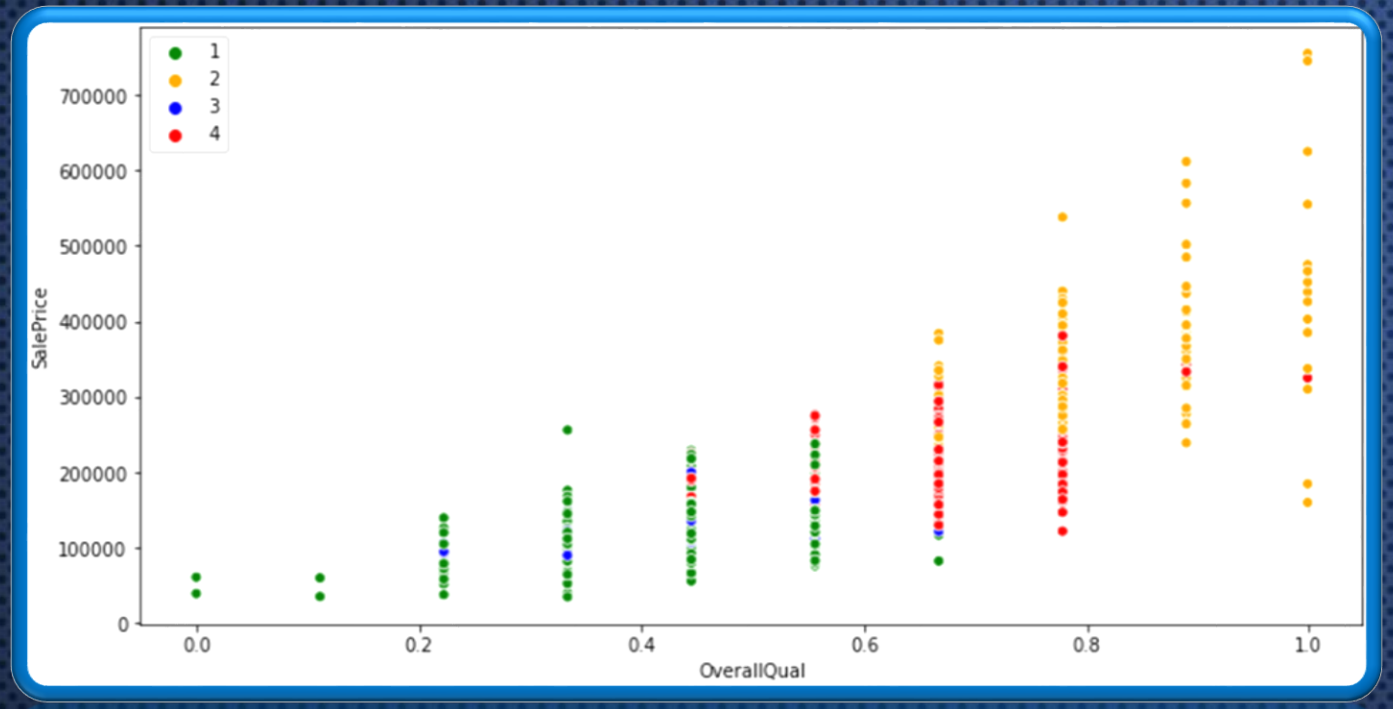
PCA - A -- 84.7 %

LR - A – 74.1 %

@ 5 Attributes

OLS Regression Results

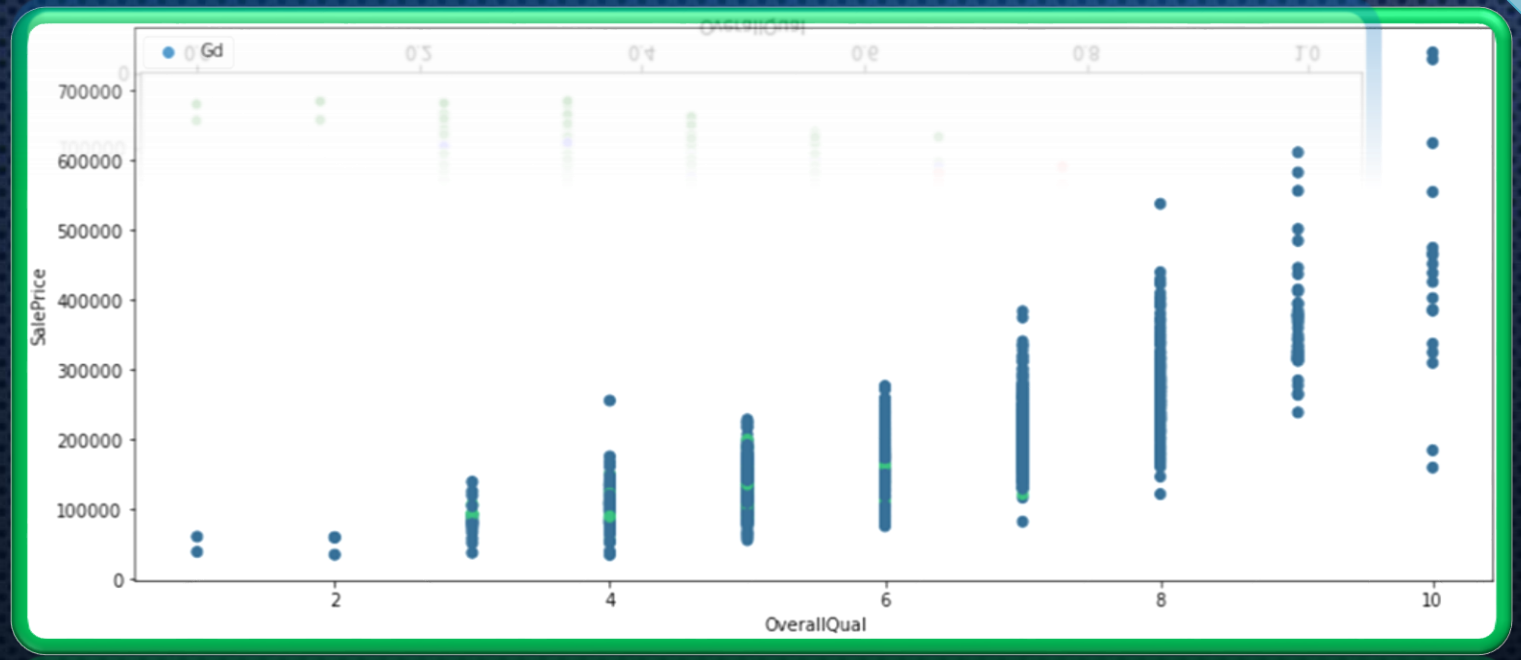
Dep. Variable:	SalePrice	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.741
Method:	Least Squares	F-statistic:	667.9
Date:	Sun, 24 Jan 2021	Prob (F-statistic):	0.00
Time:	16:49:58	Log-Likelihood:	-14041.
No. Observations:	1168	AIC:	2.809e+04
Df Residuals:	1162	BIC:	2.813e+04
Df Model:	5		
Covariance Type:	nonrobust		

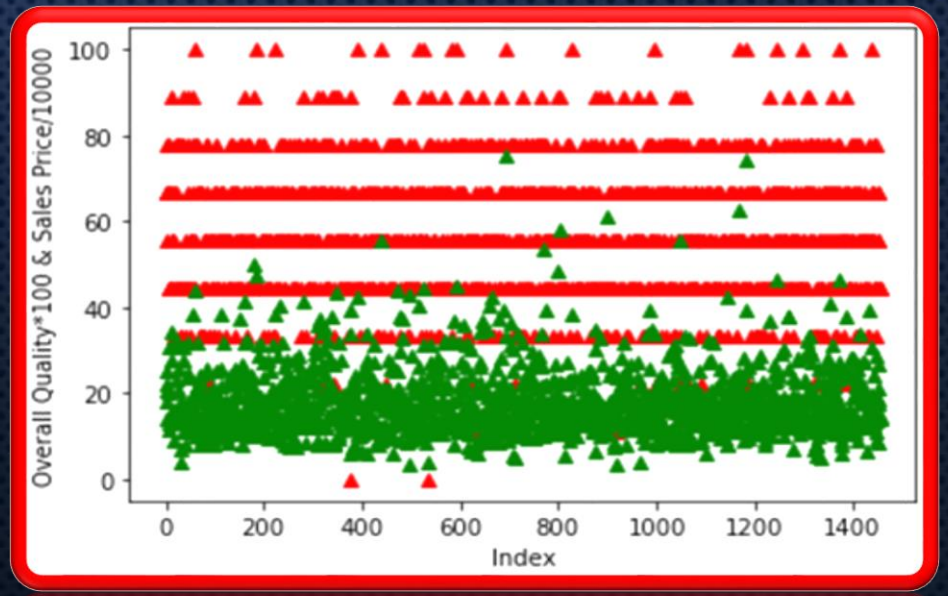


Clustering Plot on K value as 4

Dot Plot – There is a increase in manner in increase in Quality there is a increase in Sales Price.

Scatter plot for Overall Quality and Sales Price





HAPPY LEARNING