

ANALYSIS OF AMES HOUSING

BACKGROUND

I am a working for an Auction House. I have a list of houses that I intend to put up for auction in Ames. I collected the past housing dataset in Ames to research and eventually set a minimum Sale Price of the houses for the auction.

OBJECTIVE

To seek approval to set the baseline price based on the predicted price for the list of houses

DATA IMPORTING

- *There are 2 datasets namely test.csv and train.csv*
- *The price column is the target variable*
- *The rest of the variables are the features*

- *There are missing data in both train and test dataset on the numeric fields*
- *I replaced the null values with 0 in both train and test dataset*

DATA IMPORTING

- *Cleaning of data in raw dataset*

Fields	Final Field	Remarks
Bsmt Full Bath, Bsmt Half Bath, Full Bath, Half Bath,	Total Baths	Consolidate all the baths data into one field
Total Bsmt SF, Gr Liv Area, Wood Deck SF, Open Porch SF, Enclosed Porch, 3Ssn Porch, Screen Porch	totalsqft	Sum up all the area in a house
Garage Finish, Garage Yr Blt, Garage Type	has_finished_garage, Detached Garage	transformed the categorical 'Garage Finish' column into 'Has Finished Garage' and 'Has Detached Garage' using one-hot encoding to indicate finished vs unfinished garages. All remaining garage columns were dropped
Year Remod/Add, Year Built	is_remodelled	I introduced a new field to indicate that the house has been remodelled

DATA IMPORTING

- *Cleaning of data in raw dataset*

Fields	Final Field	Remarks
BsmtFin Type 2, BsmtFin Type 1	Has Finished Basement	There were multiple categorical columns describing type of basement finish, so I converted these into a one-hot encoded 'Has Finished Basement' column
Pool QC	has_pool	'Pool QC' changed to 'Has Pool'
Fence	has_fence	'Fence' changed to 'Has Fence'
Paved Drive	has_paved_drive	'Paved Drive' changed to 'Has Paved Drive'
Central Air	has_central_air	'Central Air' changed to 'Has Central Air'

DATA IMPORTING

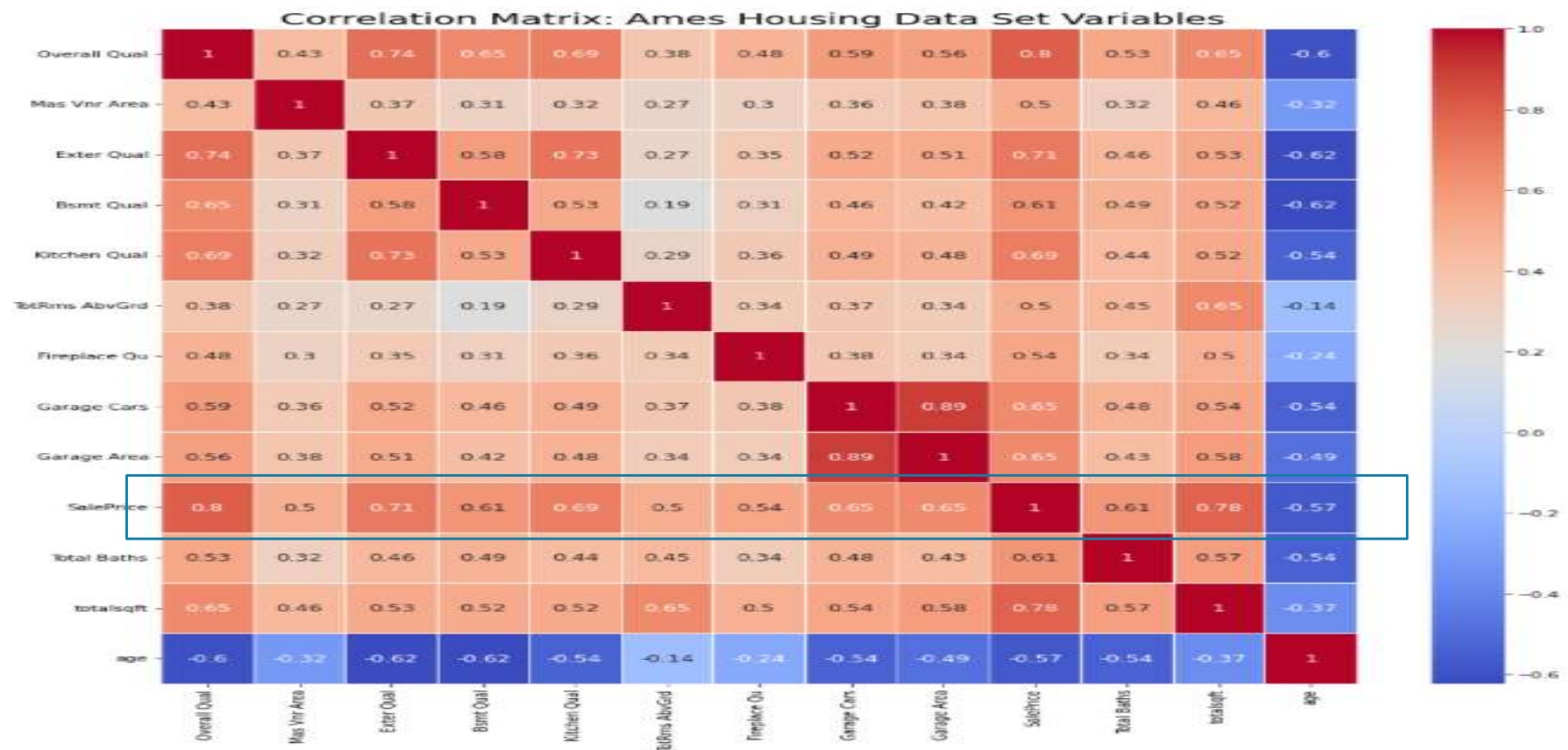
- ***Cleaning of data in raw dataset***

Fields	Final Field	Remarks
Lot Config, MS Zoning, Misc Feature, Neighbourhood, House Style, Bldg Type		A number of categorical variables were transformed into dummy column
Bsmt Qual, Bsmt Cond, Fireplace Qu, Heating QC, Garage Qual, Garage Cond, Exter Qual, Exter Cond, Kitchen Qual, Exter Qual, Exter Cond		Quality rankings - transforming categorical variables to ordinal ones
'Year Built'	age	Turning the 'Year Built' column into 'Age'

- ***Drop all the non numerical fields***

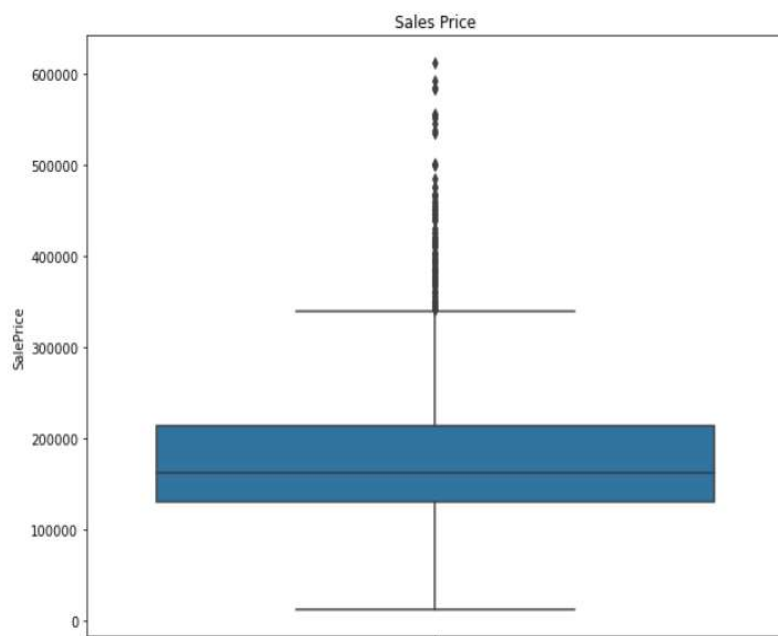
CORRELATION OF PREDICTORS

- These are the set of predictors with absolute correlation > 0.5 with respect to SalePrice



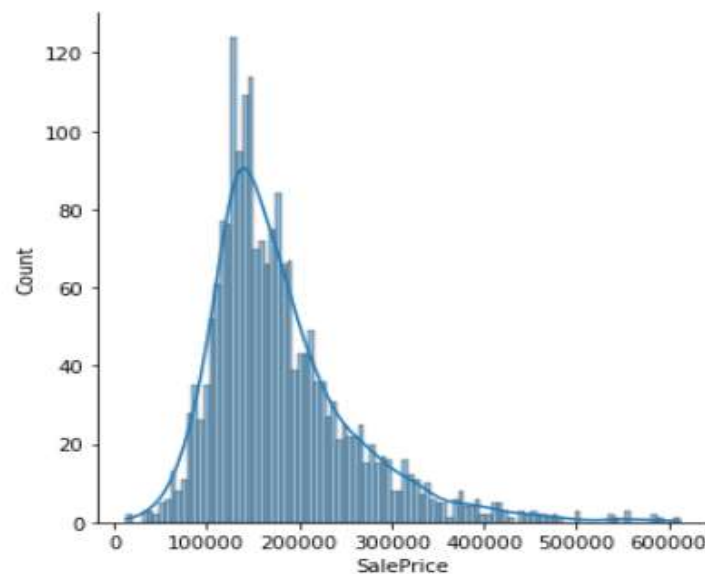
EDA – SALE PRICE

- Everything looks good here, with no missing values. A tail extends to high prices. There seemed to have many outlier points observed at higher price range. I will log the price data in order to get a narrow range with a normal distribution



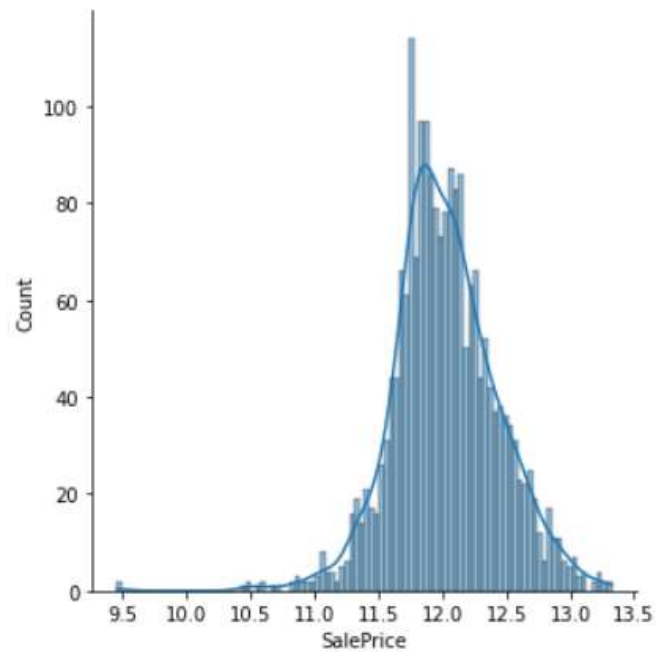
Median price: \$162500.00
Mean price: \$181469.70

Median price: \$162500.00
Mean price: \$181469.70



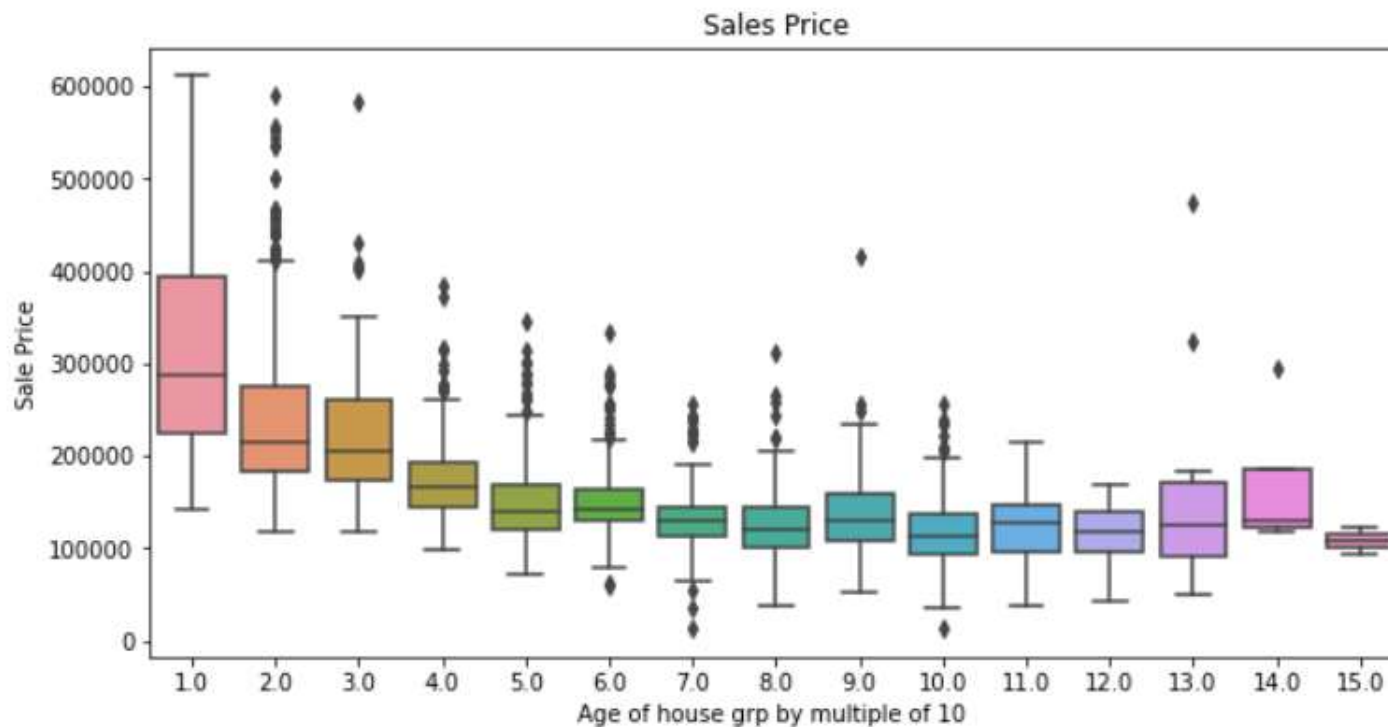
EDA — SALE PRICE

- After I perform a logarithmic transformation of the Sale Price, it made highly skewed distributions less skewed



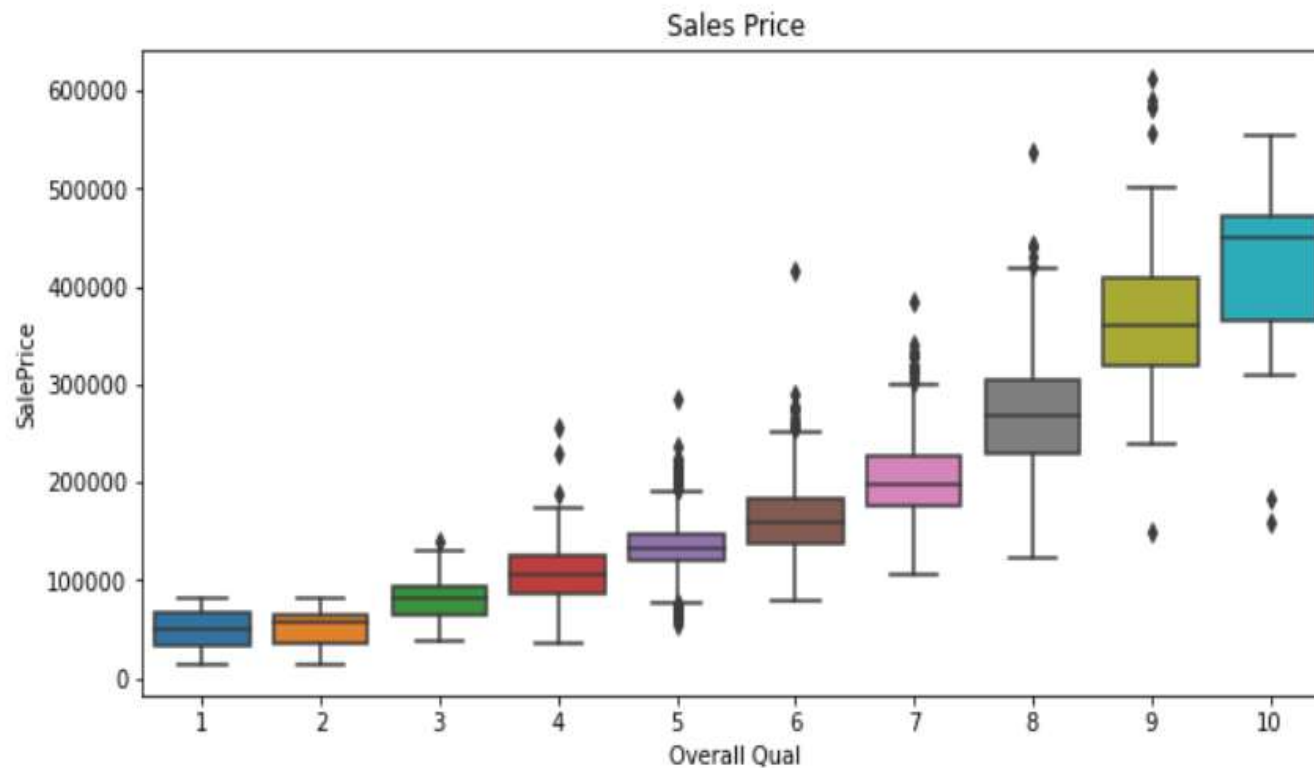
EDA — SALE PRICE WRT TO AGE OF HOUSE

I defined the age of the building as the year of sale minus the year of construction. New houses have a price premium that declines as they age even by 10 to 20 years. After a while the effect of age plateaus off, only to come back for very old houses.



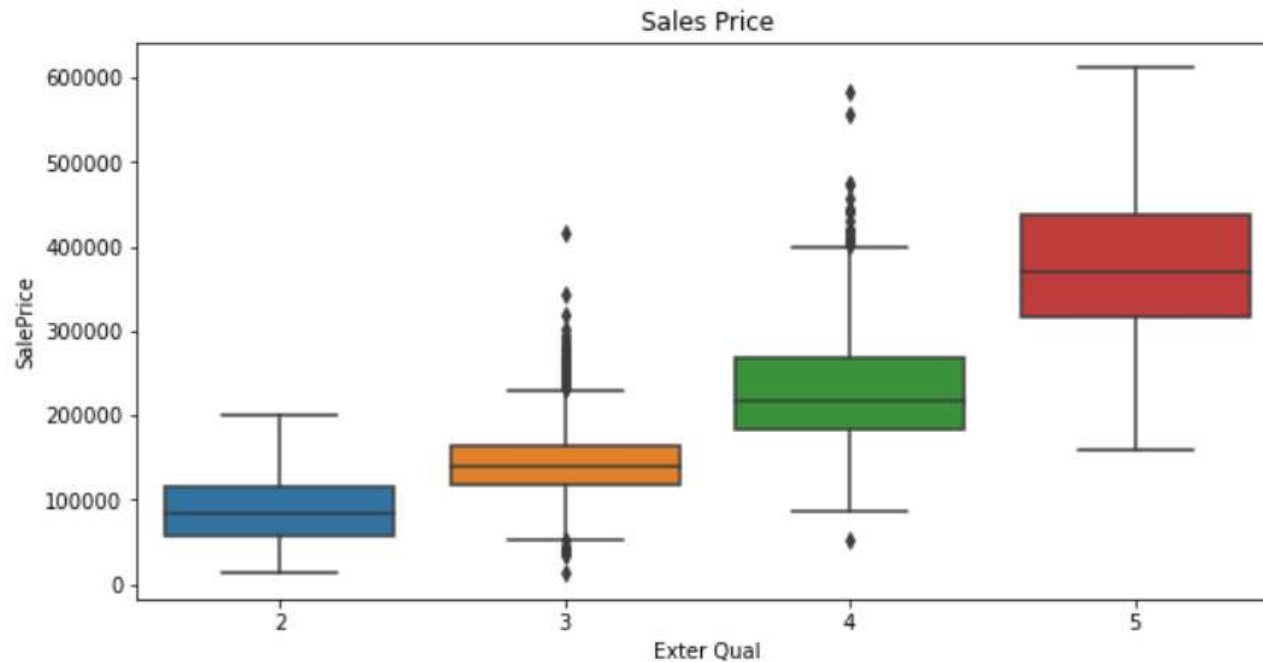
EDA — SALE PRICE WRT OVERALL QUALITY

Overall Quality of the houses from one to ten. It turns out they are great predictors of sale price, with higher quality and commanding higher prices



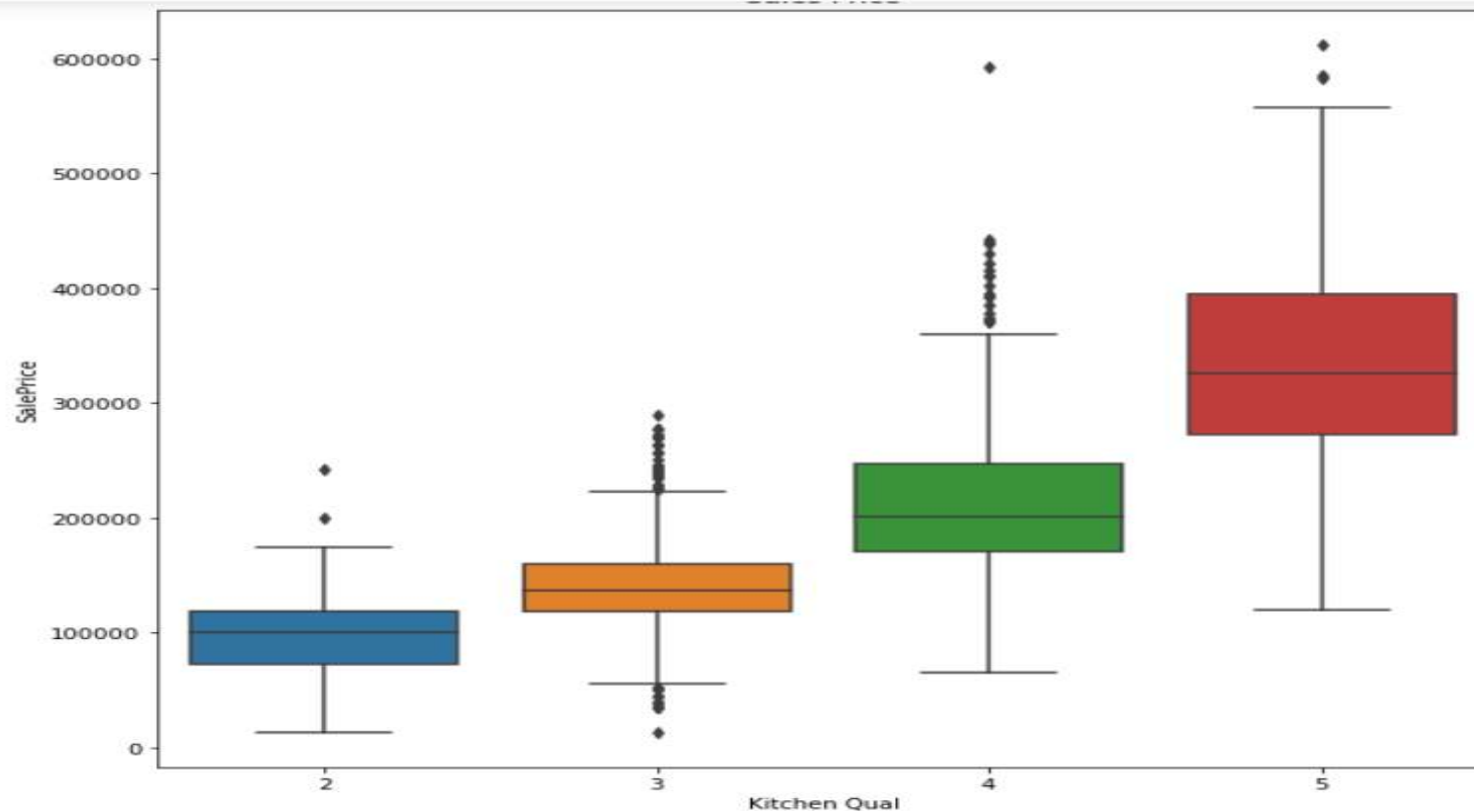
EDA — SALE PRICE WRT EXTERNAL QUALITY

External Quality of the houses from one to ten. It turns out they are great predictors of sale price, with higher quality and commanding higher prices



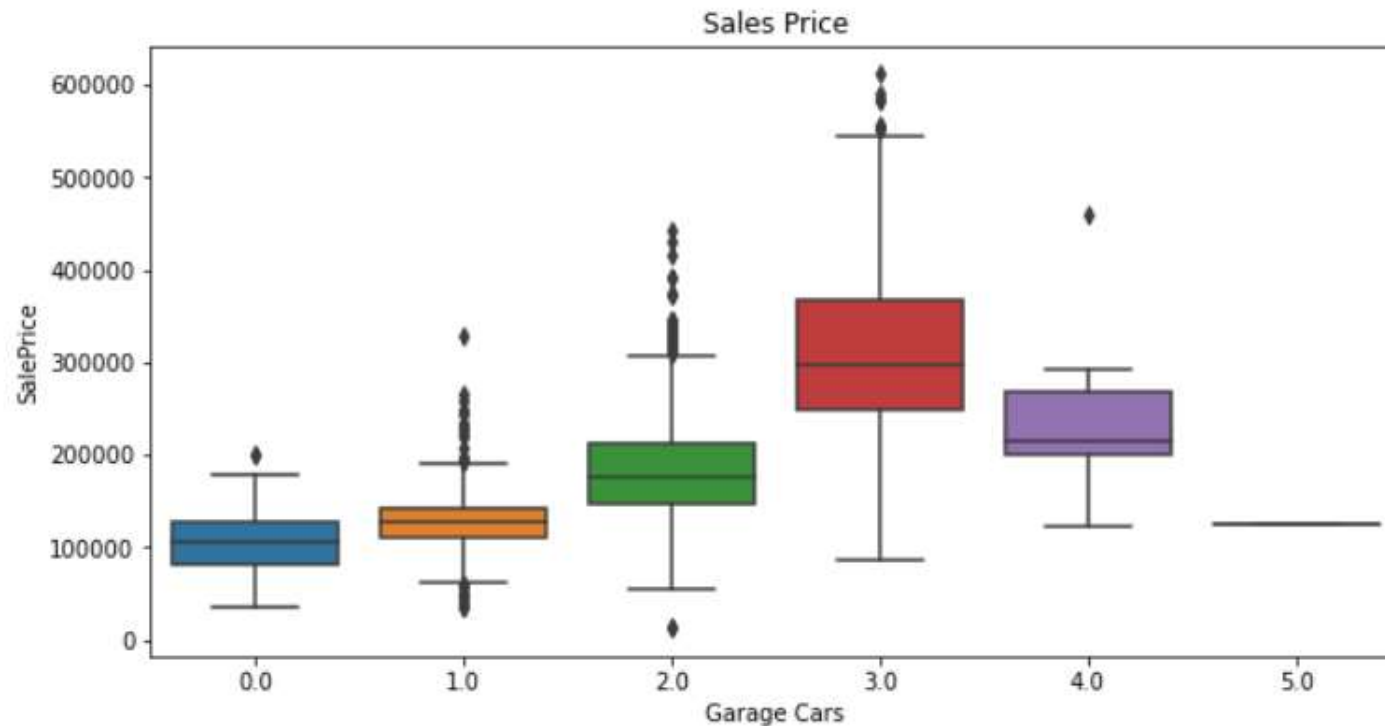
EDA — SALE PRICE WRT KITCHEN QUALITY

Kitchen Quality of the houses from one to ten. It turns out they are great predictors of sale price, with higher quality and commanding higher prices



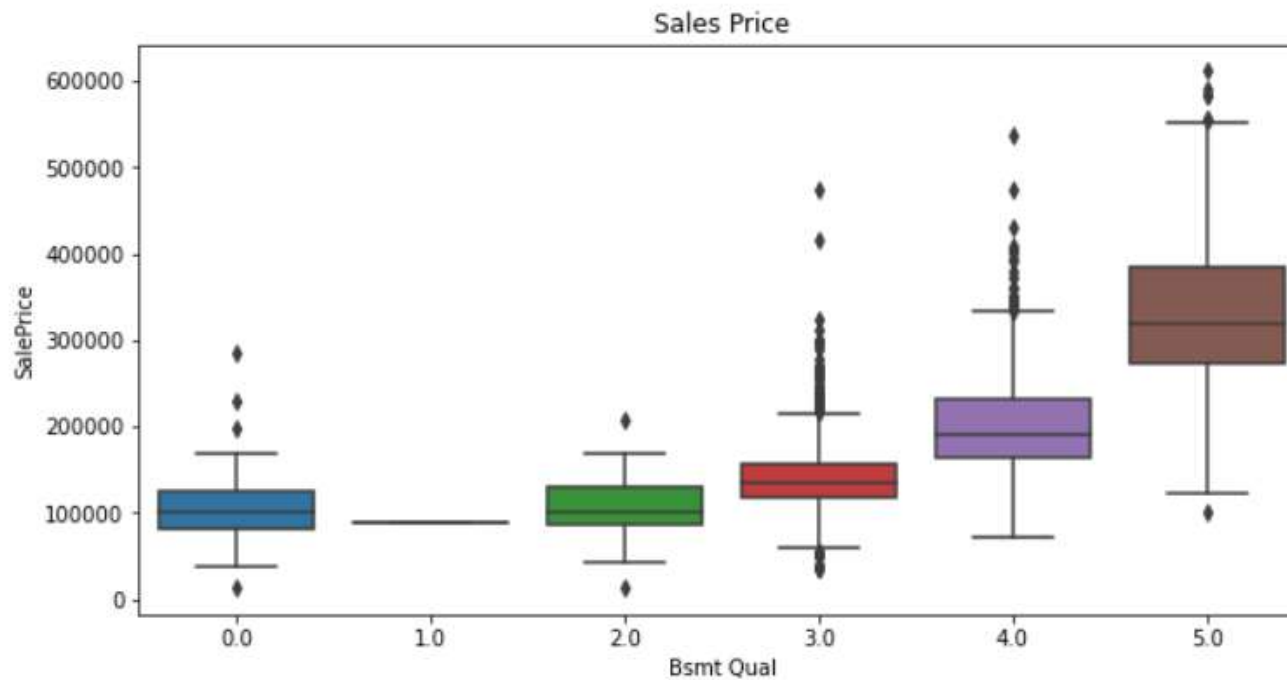
EDA — SALE PRICE WRT GARAGE CARS

Sale Price increases with the increase in garage cars. Once it reaches 4 Garage cars, the price decreases



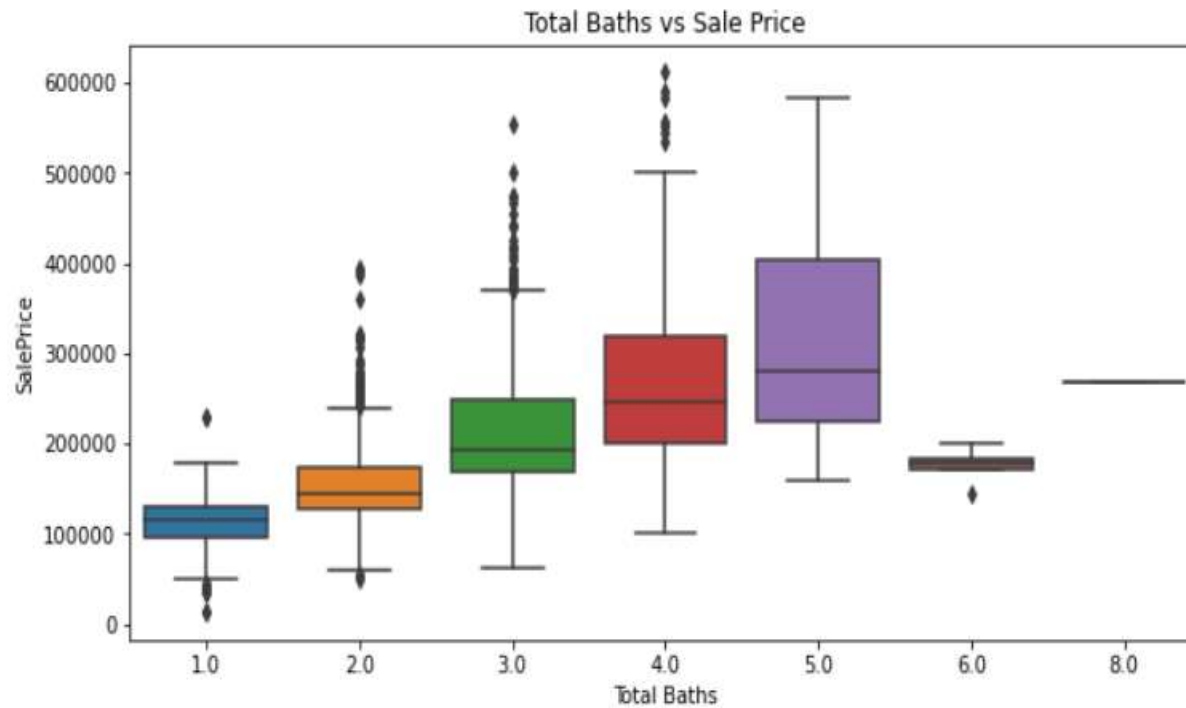
EDA — SALE PRICE WRT BASEMENT QUALITY

Sale Price increases with the increase in basement quality.



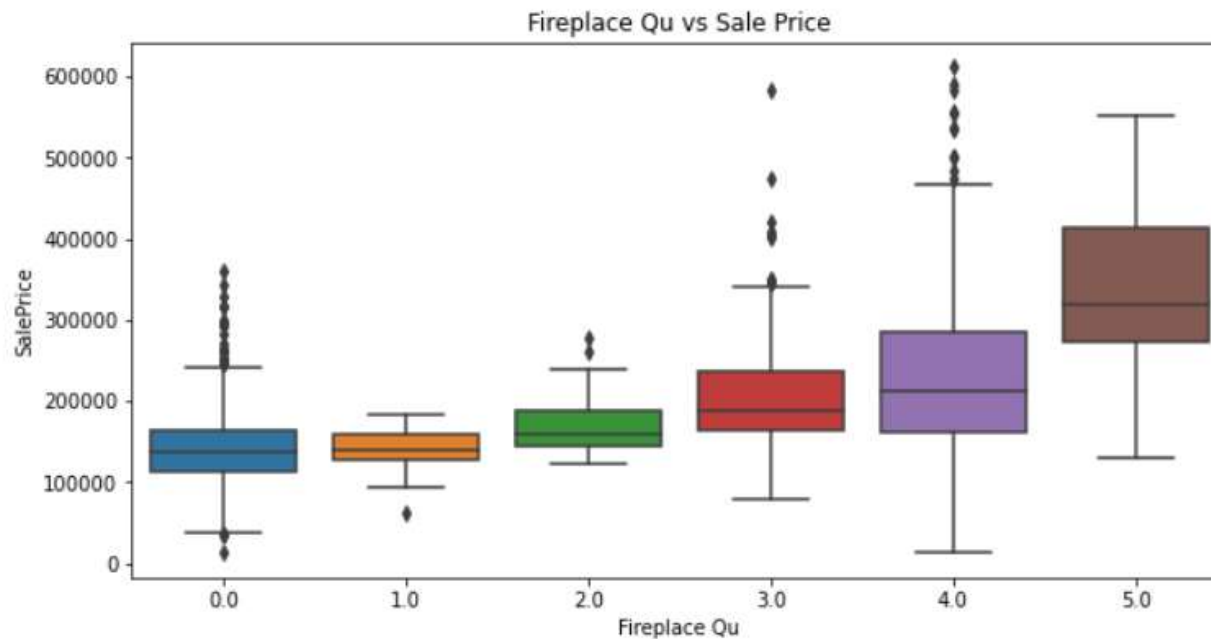
EDA — SALE PRICE WRT NO OF BATHS

- Sale Price increases with the increase in no of baths. When it is beyond 5, it will not have effect on the price



EDA — SALE PRICE WRT FIREPLACE QUALITY

- Sale Price increases with the increase in Fireplace Quality.



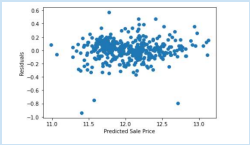

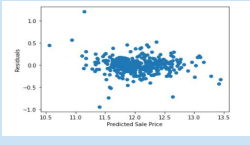
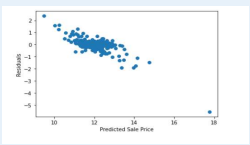
FEATURE ENGINEERING

P Value of 'Exter Qual', 'Mas Vnr Area' are more than 0.05 which is not significant. I decide to remove these features from the mode

	coef	std err	t	P> t	[0.025	0.975]
const	10.6232	0.040	264.599	0.000	10.544	10.702
Overall Qual	0.0831	0.005	16.961	0.000	0.074	0.093
totalsqft	0.0001	7.84e-06	15.896	0.000	0.000	0.000
Exter Qual	0.0104	0.011	0.952	0.341	-0.011	0.032
Kitchen Qual	0.0626	0.009	7.205	0.000	0.046	0.080
Garage Area	0.0001	4.05e-05	2.512	0.012	2.23e-05	0.000
Garage Cars	0.0383	0.012	3.297	0.001	0.016	0.061
Bsmt Qual	0.0182	0.006	2.953	0.003	0.006	0.030
Total Baths	0.0415	0.005	7.595	0.000	0.031	0.052
Fireplace Qu	0.0227	0.002	9.251	0.000	0.018	0.028
TotRms AbvGrd	0.0091	0.003	2.705	0.007	0.003	0.016
Mas Vnr Area	-1.937e-05	2.46e-05	-0.789	0.430	-6.75e-05	2.88e-05
age	-0.0016	0.000	-8.457	0.000	-0.002	-0.001
Omnibus:	1620.406	Durbin-Watson:	2.005			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	90658.993			
Skew:	-3.244	Prob(JB):	0.00			
Kurtosis:	34.918	Cond. No.	3.21e+04			

TRAIN/SCORE/EVALUATE OF MODELS

I trained and tested 4 models namely Linear Regression, Ridge Regression, Lasso Regression and ElasticNet Regression with varying Degree. Below is the Linear Regression Result

Model	Degree	Residue vs Predicted Plot	Remark
Linear Regression	1		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Linear Regression	2		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Linear Regression	3		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Linear Regression	4		The plot are not evenly distributed and they have an outlier as they have a clear shape to them. So, we can conclude that this model is not ideal.


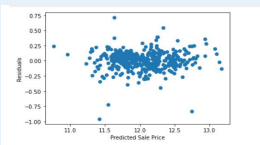
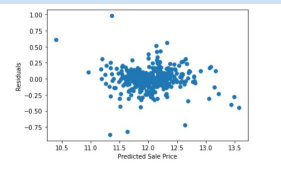

TRAIN/SCORE/EVALUATE OF MODELS

Linear Regression Result

Model	Degree	MSE (train)	MSE (test)	CVS	Remark
Linear Regression	1	0.02871836725337317	0.02429025756115978	0.03025966189638025	
Linear Regression	2	0.02028496	0.02406107	0.024583811306377802	Best among Linear Regression
Linear Regression	3	0.01461285	0.03120501	0.06199709808065891	
Linear Regression	4	0.00807620	0.21694482	2.544365455471457	

TRAIN/SCORE/EVALUATE OF MODELS

Ridge Regression

Model	Degree	Residue vs Predicted Plot	Remark
Ridge Regression	1		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Ridge Regression	2		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Ridge Regression	3		The points in the plot are not so symmetrically distributed. Model starts to deteriorate.
Ridge Regression	4		The plot is not evenly distributed and they have an outlier. There is clear shape of the plot.


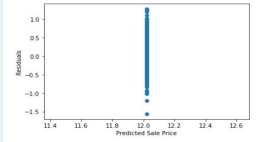
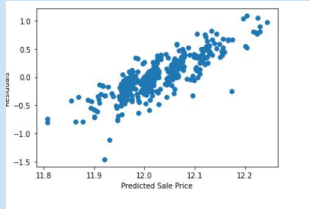
TRAIN/SCORE/EVALUATE OF MODELS

Ridge Regression Result

Model	Degree	MSE (train)	MSE (test)	CVS	Remark
Ridge Regression	1	0.028831218507909623	0.024593535056232775	0.030068272885232138	
Ridge Regression	2	0.020517549721110532	0.023634145039232196	0.023915266254815767	Best Model with lowest CVS among Ridge Regression
Ridge Regression	3	0.01780170233268262	0.027336867181414117	0.029844168853918728	
Ridge Regression	4	0.030844580485791646	0.04520048742066716	0.04875207798515581	Residue plot not ideal

TRAIN/SCORE/EVALUATE OF MODELS

Lasso and ElasticNet Regression Result

Model	Degree	Residue vs Predicted Plot	Remark
Lasso Regression	1		The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot
Lasso Regression	2		There is a clear shape of the plot. This model is not desirable
ElasticNet Regression	1		The plot has an upwards trend and points are not symmetrically distributed. This model is not desirable

TRAIN/SCORE/EVALUATE OF MODELS

Lasso and ElasticNet Regression Result

Model	Degree	MSE (train)	MSE (test)	CVS	Remark
Lasso Regression	1	0.02872769121745429	0.024330515185106642	0.030251245451425525	
Lasso Regression	2	0.16898931570251494	0.16958113581952464	0.1691760789228832	Residue plot not ideal Model gets worse
ElasticNet Regression	1	0.1218681398553341	0.12175688412319963	0.12232122843040724	Residue plot not ideal Model is not ideal

TRAIN/SCORE/EVALUATE OF MODELS

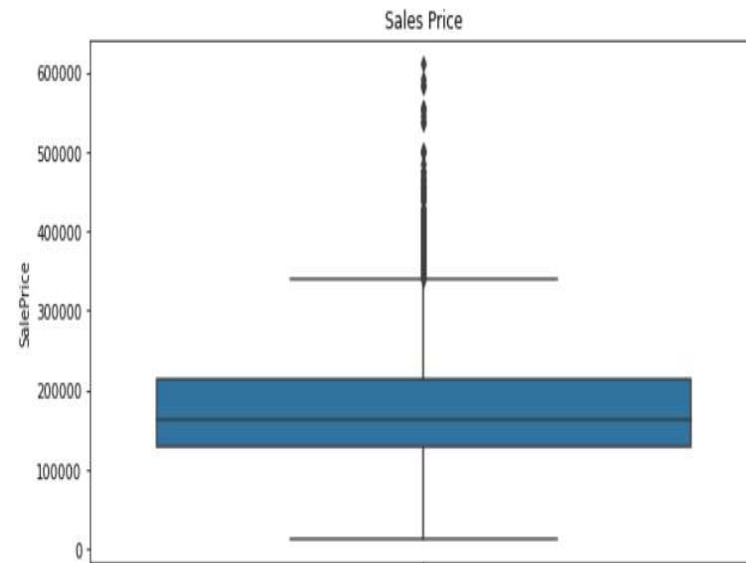
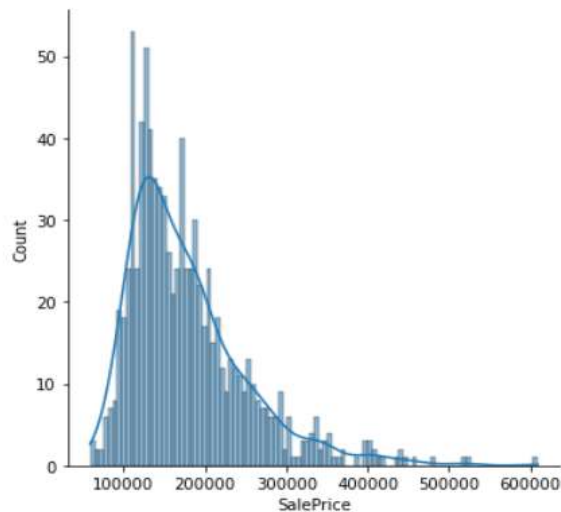
Shortlisting of the more desirable models

Model	Degree	MSE (train)	MSE (test)	CVS	Remark
Linear Regression	2	0.02028496	0.02406107	0.024583811306377802	Best among Linear Regression
Ridge Regression	2	0.020517549721110532	0.023634145039232196	0.023915266254815767	Best Model with lowest CVS
Lasso Regression	1	0.02872769121745429	0.024330515185106642	0.030251245451425525	

CONCLUSION

I selected the Ridge Regression with Degree 2 $\alpha=65.79332246575679$ with the lowest CVS. The mean price of list of houses is \$161,019.62 and the median price is \$177,687.50

Median price: \$161019.62
Mean price: \$177687.50



CONCLUSION

The mean price of list of houses is \$161,019.62.

The predictors based on this price are

- Fireplace quality is at least 3
- No of baths is at least 2
- Garage Cars is at least 2
- External Quality is at least 3
- Kitchen Quality is at least 3
- Overall Quality is at least 3

To fetch a higher price, we can improve the External Quality, Kitchen Quality and Overall Quality, Fireplace Quality of the house

Approval to set the predicted price as the baseline price for the houses