# 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

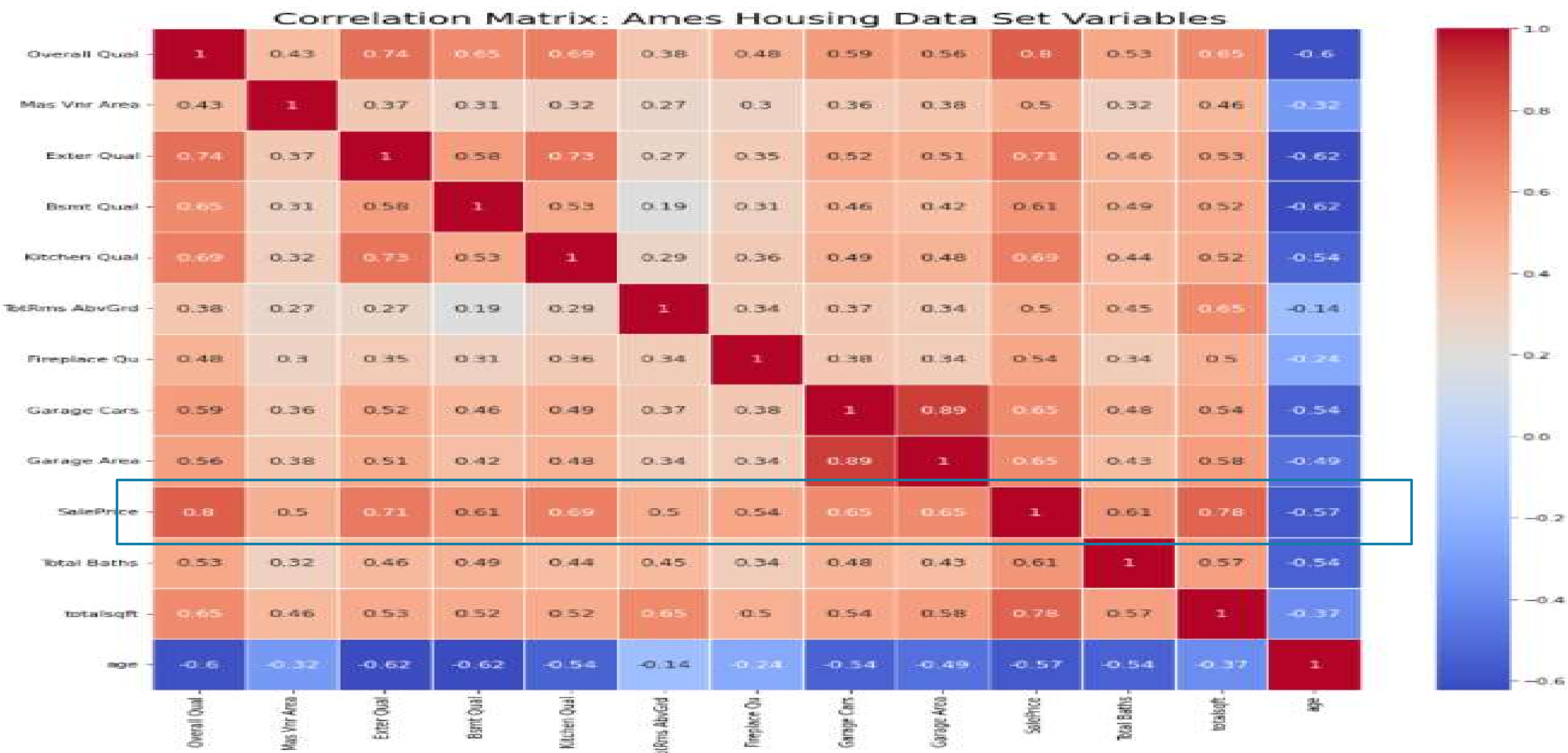
- ***Cleaning of data in raw dataset***

| 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 vairables to ordinal ones |
| 'Year Built' | age | Turning the 'Year Built' column into 'Age' |

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

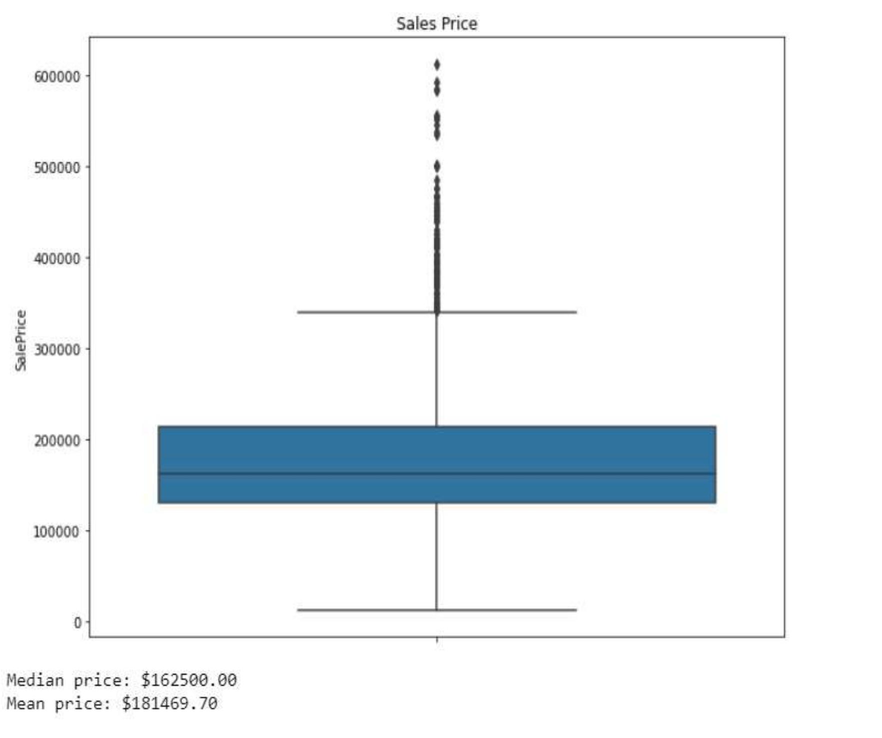
# CORRELATION OF PREDICTORS

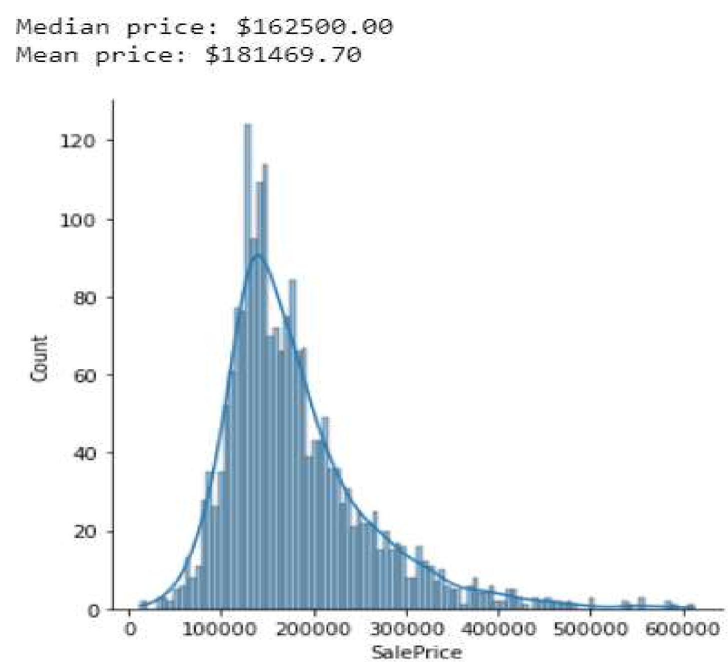
- ***These are the set of predictors with absolute correlation> 0.5 with respect to SalePrice***



# EDA – SALE PRICE

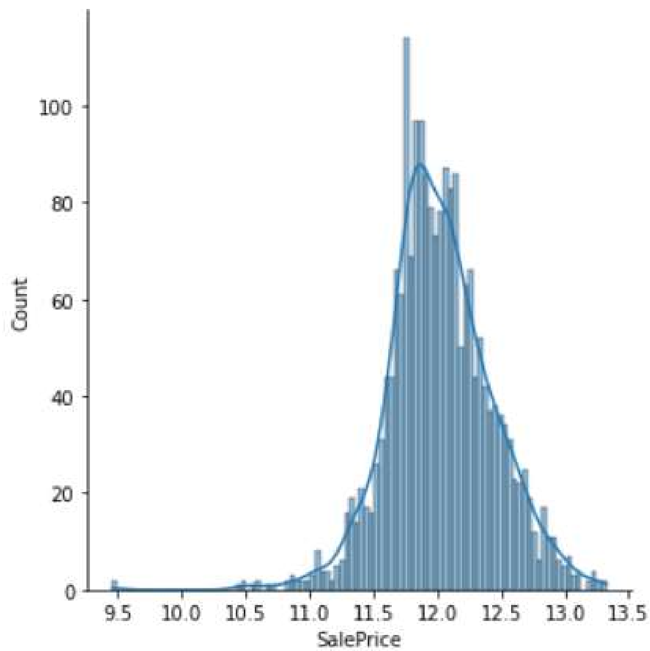
- Everything looks good here, with no missing values. A tail extends to high prices. There seemed to have many outlier points observed at higher price range. I will log the price data in order to get a narrow range with a normal distribution



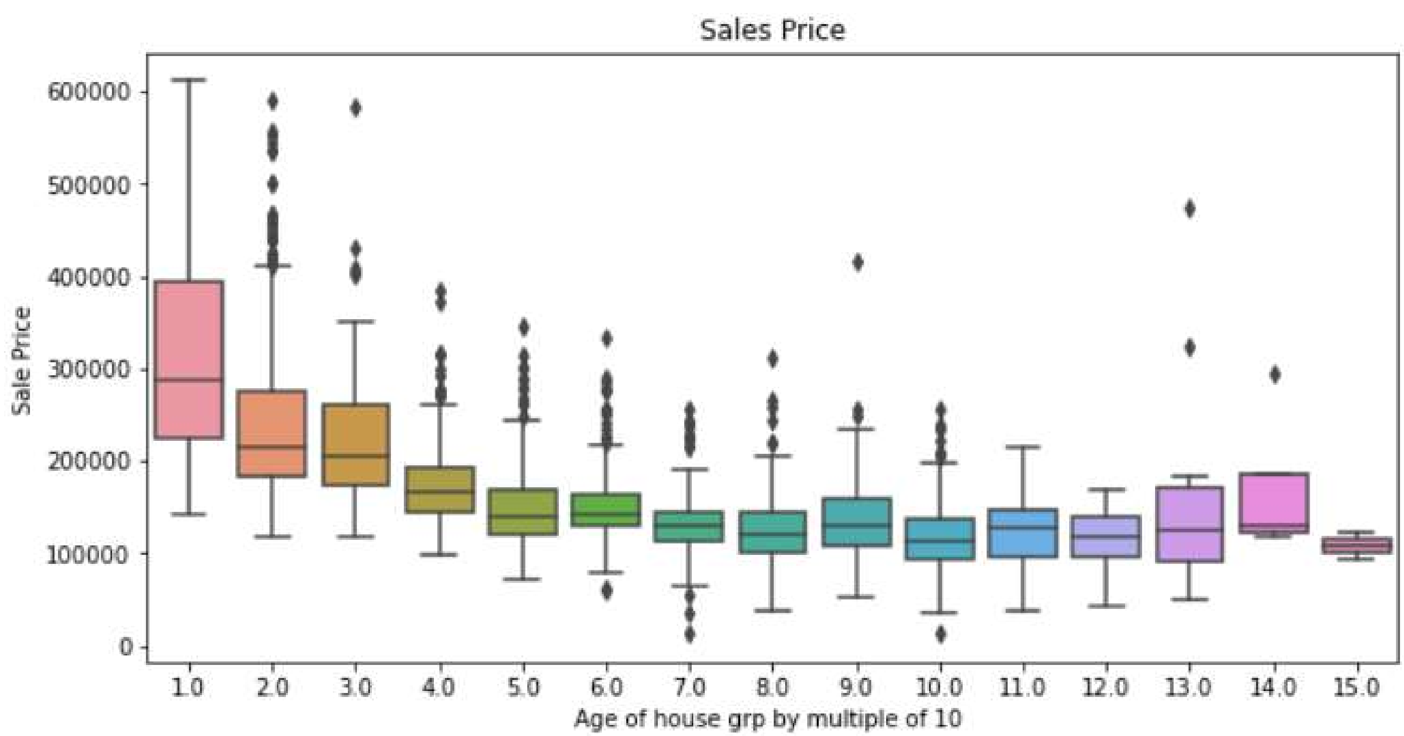
# EDA – SALE PRICE

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



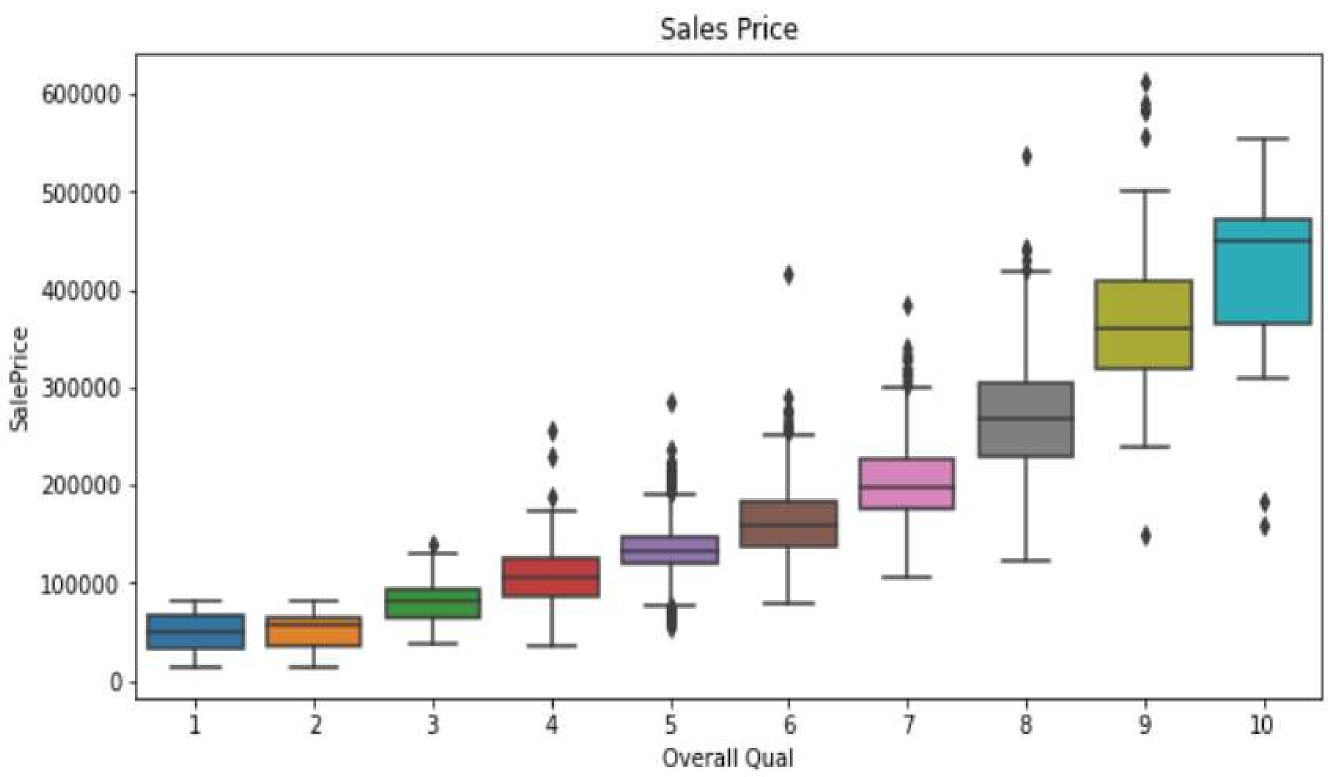
# EDA – SALE PRICE WRT TO AGE OF HOUSE

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.



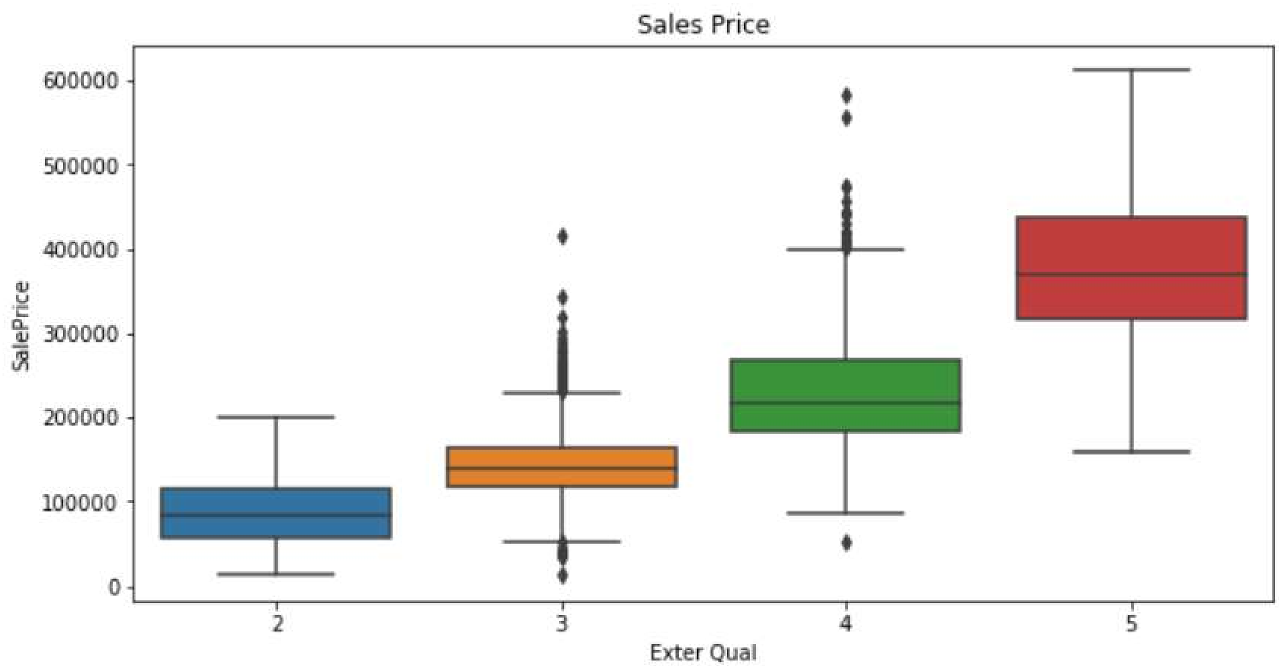
# EDA – SALE PRICE WRT OVERALL QUALITY

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



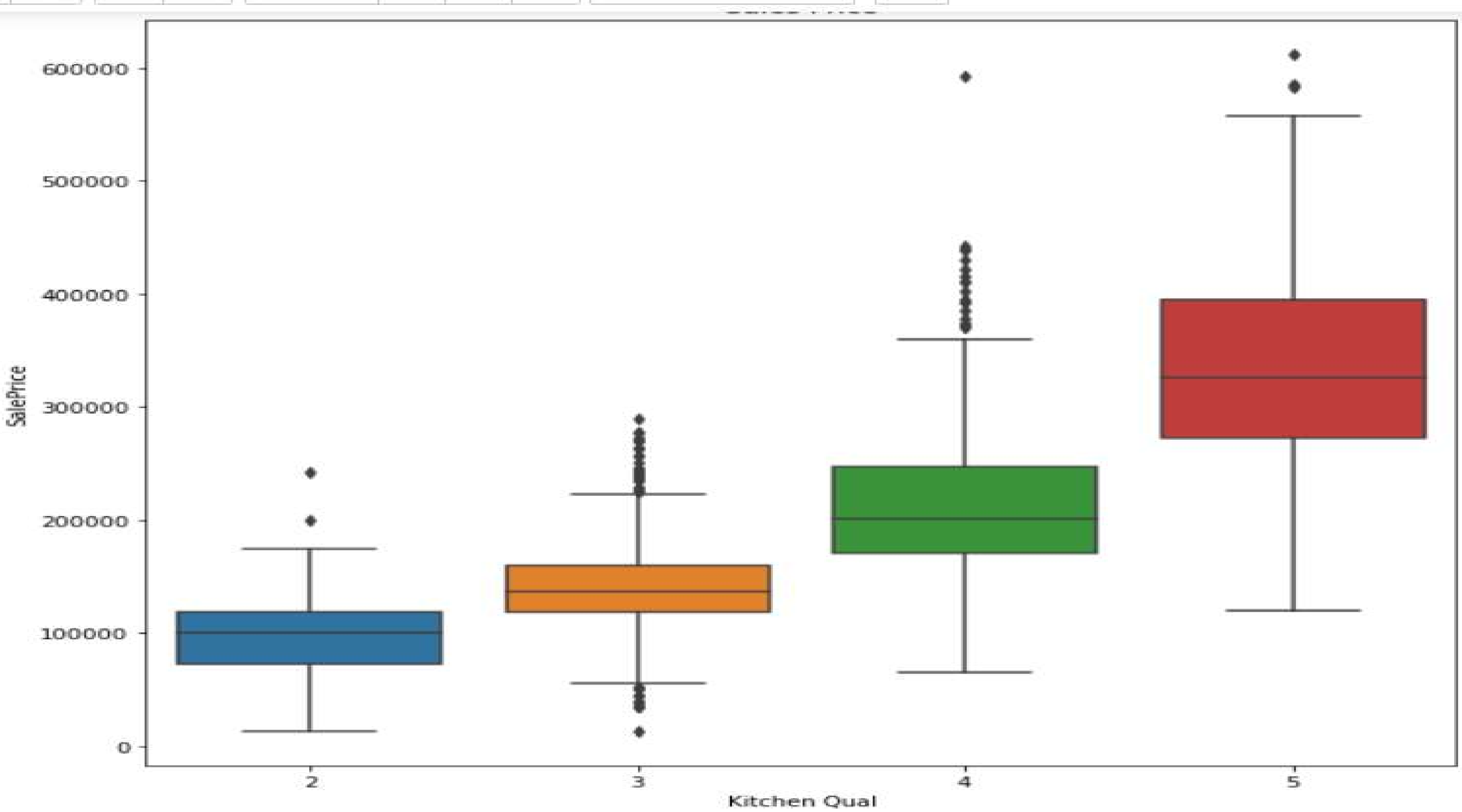
# EDA – SALE PRICE WRT EXTERNAL QUALITY

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



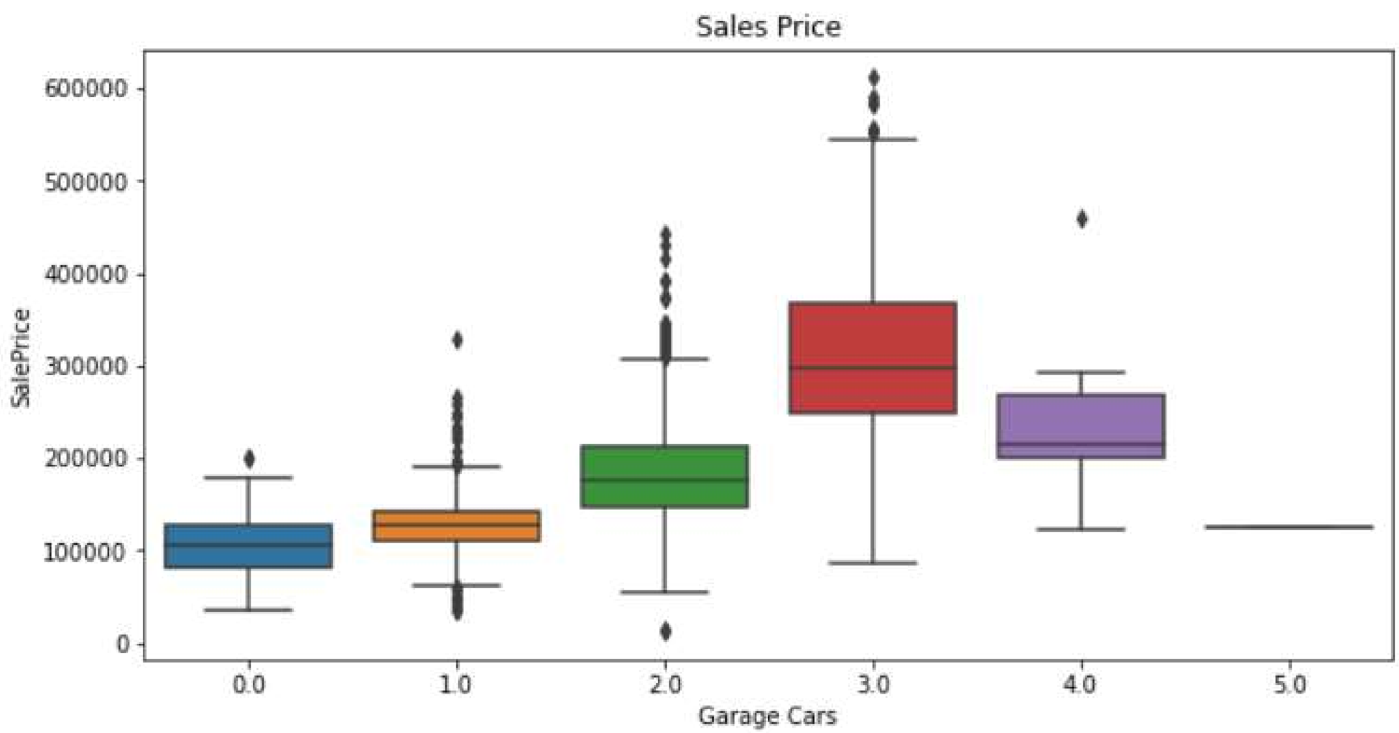
# EDA – SALE PRICE WRT KITCHEN QUALITY

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



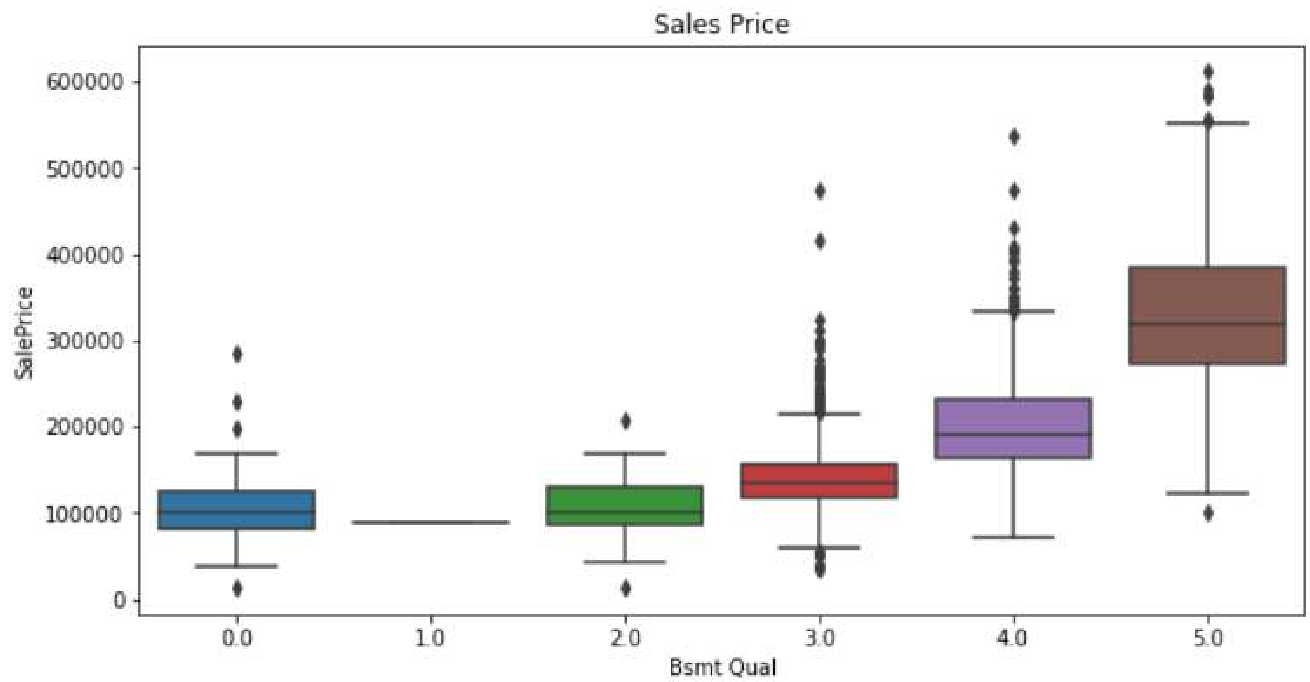
# EDA – SALE PRICE WRT GARAGE CARS

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



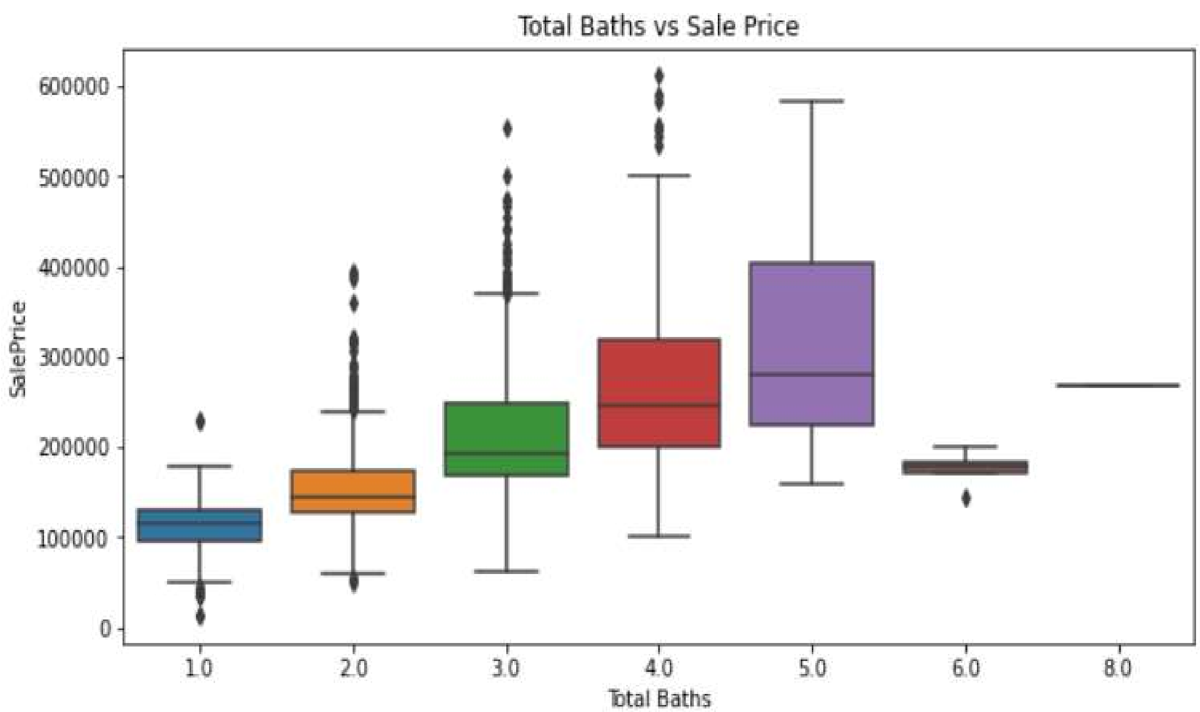
# EDA – SALE PRICE WRT BASEMENT QUALITY

Sale Price increases with the increase in basement quality.



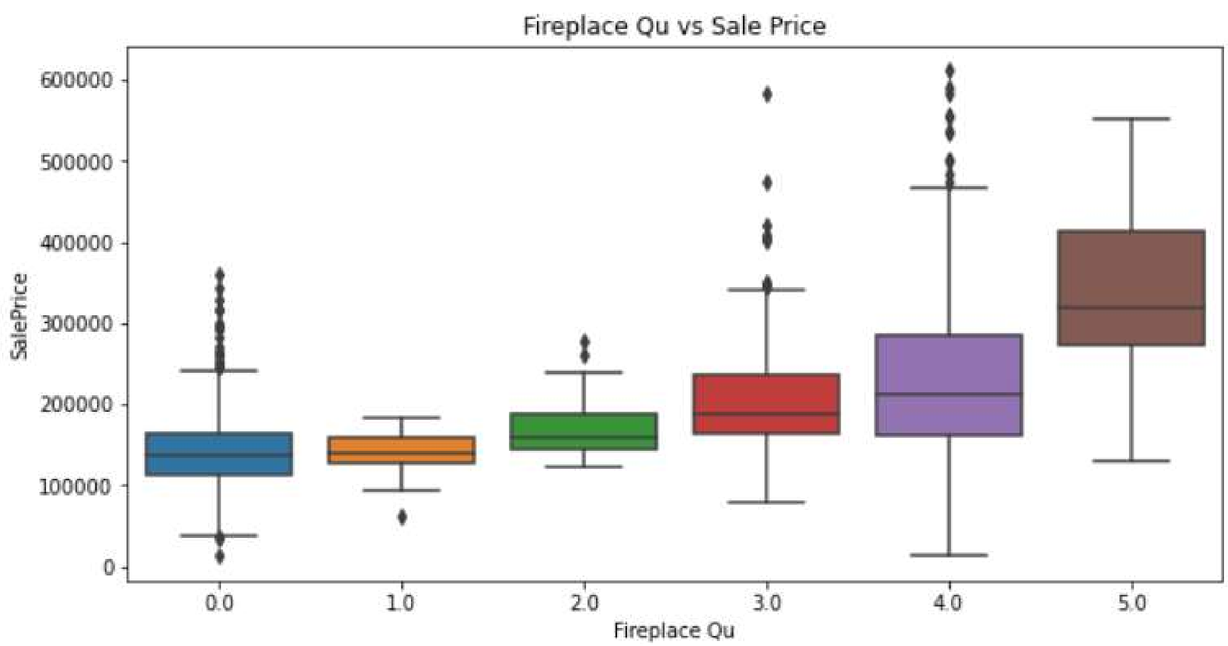
# EDA – SALE PRICE WRT NO OF BATHS

•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 | 10.6469 | 0.037 | 284.024 | 0.000 | 10.573 | 10.720 |
| Overall Qual | 0.0775 | 0.005 | 17.027 | 0.000 | 0.069 | 0.086 |
| totalsqft | 0.0002 | 7.48e-06 | 24.442 | 0.000 | 0.000 | 0.000 |
| Exter Qual | 0.0150 | 0.010 | 1.489 | 0.137 | -0.005 | 0.035 |
| Kitchen Qual | 0.0590 | 0.008 | 7.359 | 0.000 | 0.043 | 0.075 |
| Garage Cars | 0.0486 | 0.006 | 7.947 | 0.000 | 0.037 | 0.061 |
| Bsmt Qual | 0.0062 | 0.006 | 1.099 | 0.272 | -0.005 | 0.017 |
| Total Baths | 0.0356 | 0.005 | 7.056 | 0.000 | 0.026 | 0.045 |
| Fireplace Qu | 0.0180 | 0.002 | 7.931 | 0.000 | 0.014 | 0.022 |
| TotRms AbvGrd | 0.0017 | 0.003 | 0.565 | 0.572 | -0.004 | 0.008 |
| Mas Vnr Area | 1.352e-05 | 2.27e-05 | 0.595 | 0.552 | -3.1e-05 | 5.8e-05 |
| age | -0.0018 | 0.000 | -10.194 | 0.000 | -0.002 | -0.001 |

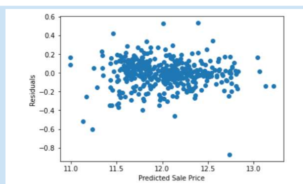
| | | | |
|---|---|---|---|
| Omnibus: | 1163.713 | Durbin-Watson: | 2.026 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 32250.681 |
| Skew: | -2.146 | Prob(JB): | 0.00 |
| Kurtosis: | 21.970 | Cond. No. | 3.17e+04 |

# TRAIN/SCORE/EVALUATE OF MODELS

I set a baseline model

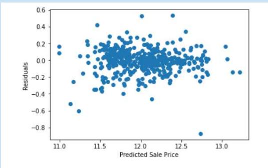
Baseline model used variables: Overall Qual, totalsqft, Exter Qual, Kitchen Qual, Garage Cars, Bsmt Qual, Total Baths, Fireplace Qu, TotRms AbvGrd, Mas Vnr Area, age

| Model | Degree | MSE (train) | MSE (test) | CVS |
|---|---|---|---|---|
| Linear Regression | 1 | 0.023862165669055902 | 0.02267001237644348 | 0.024425979399376 67 |

| 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 |

# TRAIN/SCORE/EVALUATE OF MODELS

I trained and tested 4 models
Model 1 used variables: Overall Qual, totalsqft, Kitchen Qual, Garage Cars,Total Baths, Fireplace Qu, age
I tested using Linear Regression, Ridge Regression and Lasso Regression

| 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 |
| Ridge Regression | 1 |  | The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot |
| Lasso Regression | 1 |  | The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot |

# TRAIN/SCORE/EVALUATE OF MODELS

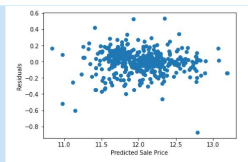
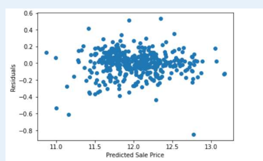
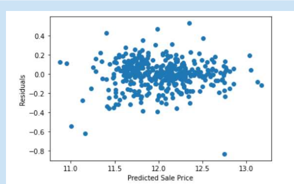
Model 1 Scores

| Model | Degree | MSE (train) | MSE (test) | CVS | Remark |
|---|---|---|---|---|---|
| Linear Regression | 1 | 0.023926949832131658 | 0.022589880655978015 | 0.024279238164378304 | |
| Ridge Regression | 1 | 0.02392817964988823 | 0.02258433059028586 | 0.02427708430597085 | Best among Model 1 |
| Lasso Regression | 1 | 0.02392882548841778 | 0.02258909916851307 | 0.024281318674783132 | |

# TRAIN/SCORE/EVALUATE OF MODELS

Model 2 used variables: Overall Qual, totalsqft, Kitchen Qual, Garage Cars,Total Baths, Fireplace Qu, age but with polynomial Degree 2

I tested using Linear Regression, Ridge Regression and Lasso Regression

| Model | Degree | Residue vs Predicted Plot | Remark |
|---|---|---|---|
| Linear Regression | 2 |  | 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 |
| Lasso Regression | 2 |  | The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot |

# TRAIN/SCORE/EVALUATE OF MODELS

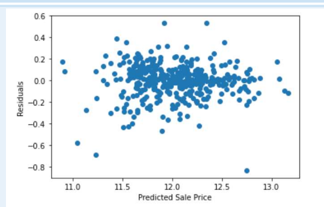
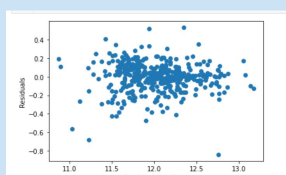
Model 2 Scores

| Model | Degree | MSE (train) | MSE (test) | CVS | Remark |
|---|---|---|---|---|---|
| Linear Regression | 2 | 0.02192103 | 0.02200216 | 0.024460194133659883 | |
| Ridge Regression | 2 | 0.022023794923807127 | 0.021761534481688397 | 0.024241644882198758 | |
| Lasso Regression | 2 | 0.022539032281693936 | 0.021854153072195 67 | 0.023998887746481 44 | Best among model 2 |

# TRAIN/SCORE/EVALUATE OF MODELS

Model 3 used variables: Overall Qual, totalsqft, Kitchen Qual, Garage Cars,Total Baths, Fireplace Qu but with polynomial Degree 2

I tested using Linear Regression, Ridge Regression and Lasso Regression

| Model | Degree | Residue vs Predicted Plot | Remark |
|---|---|---|---|
| Linear Regression | 2 |   | 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 |
| Lasso Regression | 2 |   | The points in the plot are pretty symmetrically distributed tending to cluster towards the middle of the plot |

# TRAIN/SCORE/EVALUATE OF MODELS

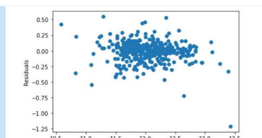
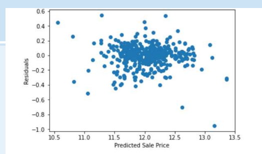
Model 3 Scores

| Model | Degree | MSE (train) | MSE (test) | CVS | Remark |
|---|---|---|---|---|---|
| Linear Regression | 2 | 0.02319557 | 0.02419273 | 0.025086988660436578 | |
| Ridge Regression | 2 | 00.023268444541707595 | 0.024117539956324784 | 0.024938207000677818 | |
| Lasso Regression | 2 | 0.02333066978598400 2 | 0.024226242692108962 | 0.024909124969940515 | Best among model 3 |

# TRAIN/SCORE/EVALUATE OF MODELS

Model 4 used variables: Overall Qual, totalsqft, Kitchen Qual, Garage Cars,Total Baths, Fireplace Qu, age but with polynomial Degree 3

I tested using Linear Regression, Ridge Regression

| Model | Degree | Residue vs Predicted Plot | Remark |
|---|---|---|---|
| Linear Regression | 3 |  | 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 pretty symmetrically distributed tending to cluster towards the middle of the plot |

# TRAIN/SCORE/EVALUATE OF MODELS

Model 2 Scores

| Model | Degree | MSE (train) | MSE (test) | CVS | Remark |
|---|---|---|---|---|---|
| Linear Regression | 3 | 0.01851985 | 0.02644404 | 0.029948109127820262 | Overfitting as MSE Train and MSE Test deviates |
| Ridge Regression | 3 | 0.019229721417379997 | 0.024144085657243339 | 0.026347308615236743 | Overfitting as MSE Train and MSE Test deviates |

# TRAIN/SCORE/EVALUATE OF MODELS

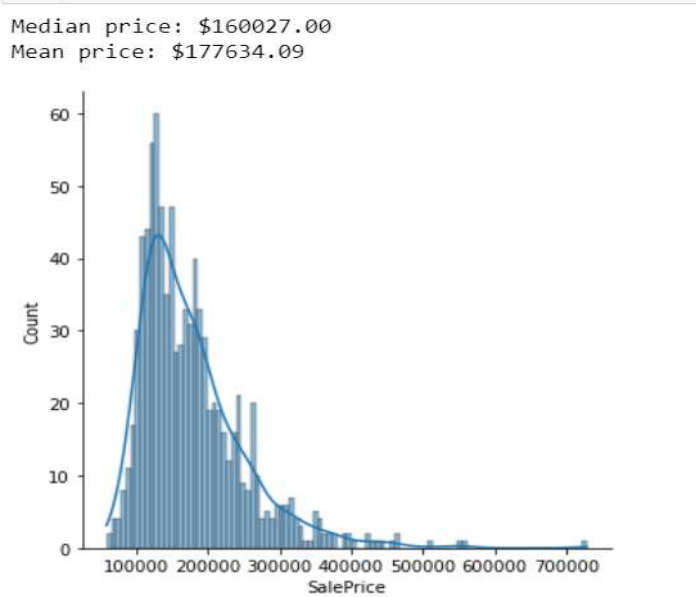
Shortlisting of the more desirable models

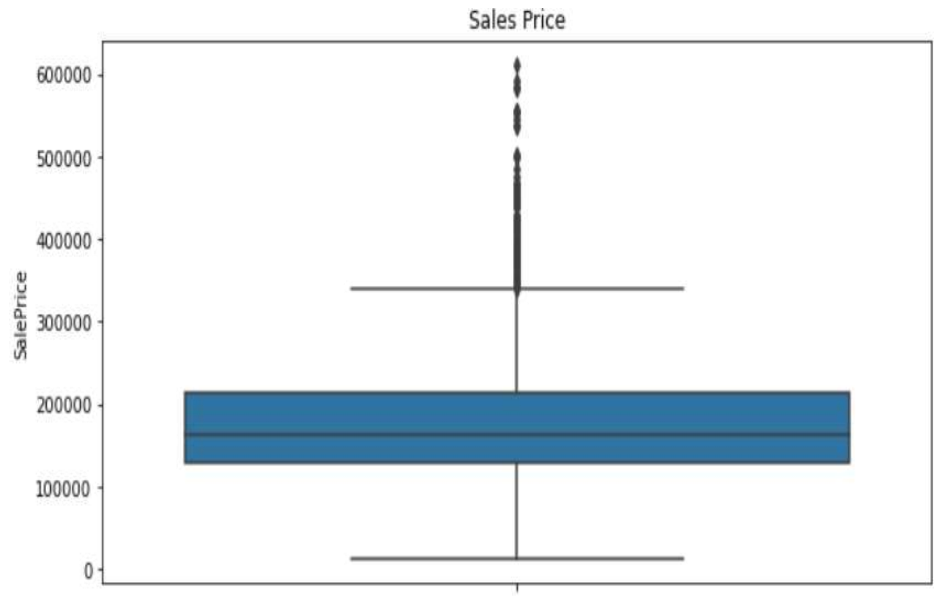
| Model | Degree | MSE (train) | MSE (test) | CVS | Remark |
|---|---|---|---|---|---|
| Model 1 - Ridge Regression | 1 | 0.0239281796498882 | 0.0225843305902858 | 0.0242770843059708 | |
| Model 2 - Lasso Regression | 2 | 0.0225390322816939 | 0.0218541530721956 | 0.0239988877464814 | Selected this as best model |
| Model 3 -Lasso Regression | 2 | 0.0233306697859840 | 0.0242262426921089 | 0.0249091249699405 | |

- I selected Lasso Regression with alpha=0.0037649358067924675 Degree 2 as the best model based on the MSE and CVS score. On top of it, I validated the residue plot against predicted price. The plot is random and even distributed. As for the actual value versus predicted value, it is linear and acceptable.
- Although this is more complex model, areas of improvement is explained in layman term to the customers. There is no concern in using the complexed mode

# CONCLUSION

I selected the Lasso Regression with Degree 2 alpha=0.0037649358067924675 with the lowest CVS.  The mean price of list of houses is $160,027 and the median price is $177,634



# 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