DSIF3
Group 1
Project 2

Brandon
Cheng Kiat
MunYee
Thien Sean
Raymond

Problem Statement

As real estate analysts in Iowa, we are responsible for managing our organization's real estate holdings.

We are tasked with understanding real estate market trends and to minimize current and future real estate holding risks.

We will be conducting data analysis on the Iowa real estate market and determine which are the factors that affects property prices.

The purpose of this analysis is to better understand property prices and provide suitable insights to management and potential buyers of the organization's real estate.
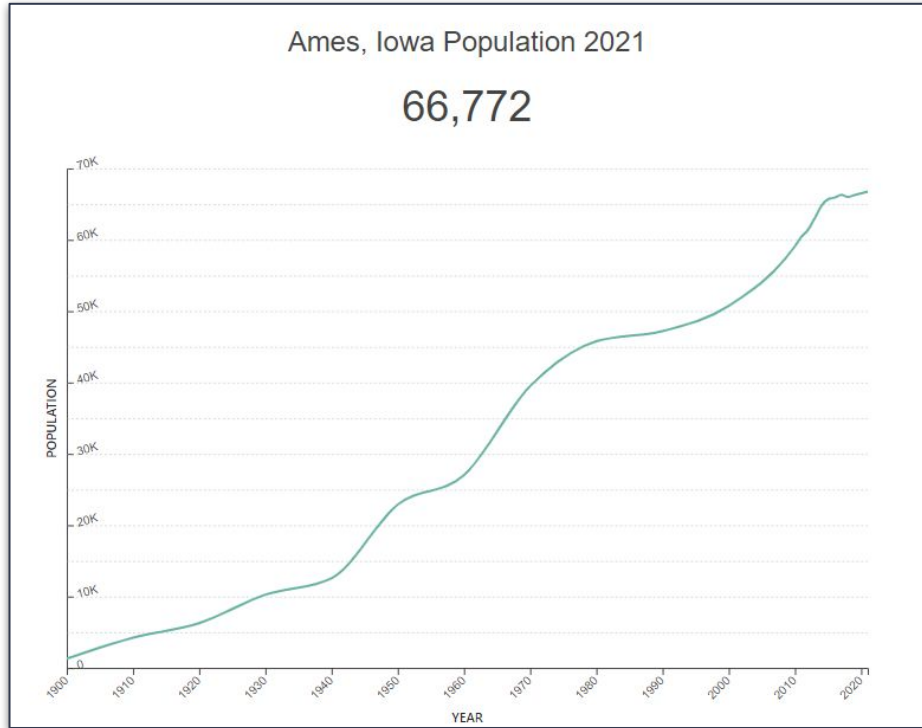
www.G1 propnet.com

Overview :

1. Purpose Statement
2. Background & Data
3. Methodology :
- EDA & data cleaning
- Feature engineering
- Data transformation
- Modeling
4. Conclusion & Recommendation
5. Q&A

Introduction

Ames, Iowa Population 2021

66,772

Ames, Iowa population have been steadily increasing year to year and this is a great opportunity in the housing sector to help families find the perfect home.

But what makes a property to be high in demand? What features fetch the best pricing during sales closing?

EDA & Data Cleaning

Checking for Missing Values

Divided the missing values into 3 groups (Basement Variables, Garage Variables & Individual Variables)

Variables	Number of Missing Values	Method of Resolving
Pool QC	2042	Leave it as NaN since NaN means no pool
Misc Feature	1986	Leave it as NaN since NaN means no misc feature
Alley	1911	Leave it as NaN since NaN means no alley access
Fence	1651	Leave it as NaN since NaN means no fence
Fireplace Qu	1000	Leave it as NaN since NaN means no fireplace
Lot Frontage	330	Impute 0 to the missing values as missing values in Lot Frontage occurs mainly in residential low density zones which means no lot frontage
Mas Vnr Type	22	Assign as None after value_counts showed 60% of the houses have None
Mas Vnr Area	22	Assign as 0 since the number of missing values is equal to Mas Vnr Type

Checking for Missing Values (Garage Variables)

Variables	Number of Missing Values
Garage Yr Blt	114
Garage Cond	114
Garage Qual	114
Garage Finish	114
Garage Type	113
Garage Cars	1
Garage Area	1

- The garage type has 1 less missing value as compare to the variables above it thus indicating there is only 1 missing value among the garage variables.

- Pulling out the entry tells that the garage type with missing values is detchd.

- We then impute the most common values for Detchd garages to fill the NaN.

	Garage Type	Garage Yr Blt	Garage Finish	Garage Cars	Garage Area	Garage Qual	Garage Cond
1712	Detchd	NaN	NaN	NaN	NaN	NaN	NaN

Checking for Missing Values (Basement Variables)

Variables	Number of Missing Values
Bsmt Exposure	58
BsmtFin Type 2	56
Bsmt Cond	55
Bsmt Qual	55
BsmtFin Type1	55

- Divided the basement variables into 2.
  - Categorical - Pink (Slide 8 & 9)
  - Numeric - Green (Slide 10)

- There is 1 more BsmtFin Type 2 and 3 more Bsmt Exposure.

- Examining Bsmt Exposure shows the basement were unfinished thus the exposure was not missing.

- Imputed No to these 3 missing values as No is the most common occurrence. Basement are normally not exposed unless there is construction.

	Bsmt Qual	Bsmt Cond	Bsmt Exposure	BsmtFin Type 1	BsmtFin SF 1	BsmtFin Type 2	BsmtFin SF 2	Bsmt Unf SF	Total Bsmt SF	Bsmt Full Bath	Bsmt Half Bath
1456	Gd	TA	NaN	Unf	0.0	Unf	0.0	725.0	725.0	0.0	0.0
1547	Gd	TA	NaN	Unf	0.0	Unf	0.0	1595.0	1595.0	0.0	0.0
1997	Gd	TA	NaN	Unf	0.0	Unf	0.0	936.0	936.0	0.0	0.0

Checking for Missing Values (Basement Variables)

Variables	Number of Missing Values
Bsmt Exposure	55
BsmtFin Type 2	56
Bsmt Cond	55
Bsmt Qual	55
BsmtFin Type1	55

- Updated table after Bsmt Exposure values are filled.

- There is now only 1 more BsmtFin Type 2

- Pulling out the entry shows the house has 2 basement and 1 of them is finished while the second is unfinished thus the type is missing.

- The missing value is change to Unf which stands for unfinished.

	Bsmt Qual	Bsmt Cond	Bsmt Exposure	BsmtFin Type 1	BsmtFin SF 1	BsmtFin Type 2	BsmtFin SF 2	Bsmt Unf SF	Total Bsmt SF	Bsmt Full Bath	Bsmt Half Bath
1147	Gd	TA	No	GLQ	1124.0	NaN	479.0	1603.0	3206.0	1.0	0.0

Checking for Missing Values (Basement Variables)

Variables	Number of Missing Values
Bsmt Qual	55
Bsmt Half Bath	2
Bsmt Full Bath	2
Total Bsmt SF	1
Bsmt Unf SF	1
Bsmt Unf SF 2	1
Bsmt SF 1	1

- Bringing in a numeric variable (Bsmt Qual) for comparison

- Pulling out the data shows both the houses do not have basement.

- Impute 0 to all missing values since the house has no basement.

	Bsmt Qual	BsmtFin SF 1	BsmtFin SF 2	Bsmt Unf SF	Total Bsmt SF	Bsmt Full Bath	Bsmt Half Bath
616	NaN	0.0	0.0	0.0	0.0	NaN	NaN
1327	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Distribution of Target Variable (SalePrice)

The target variable behaves as a normal distribution with mean about 150,000 and slightly skew to the right.

This indicates that there are some houses which are sold at extremely high prices thus creating this skew.

Distribution of Sales Price
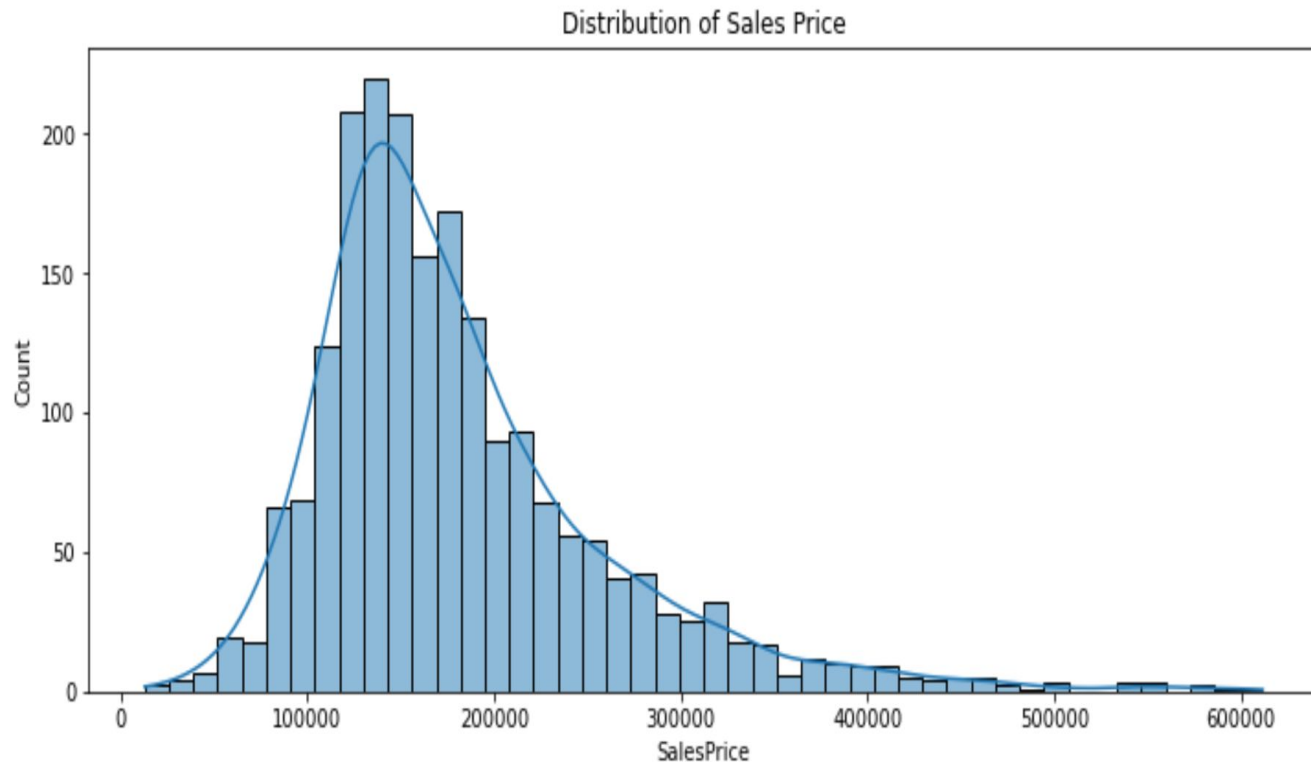
Outliers

Box Plot of Sales Price
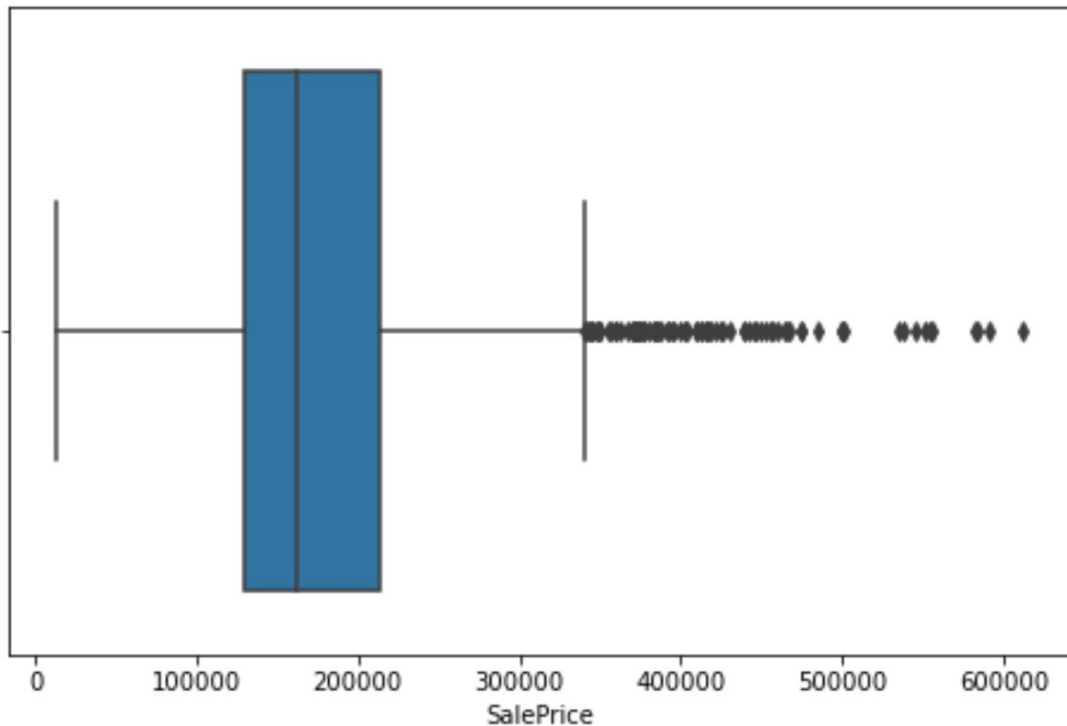
Although there are outliers, this does not necessary mean that those data are wrong.

It is just the price of some houses may fetch high prices due to their location, accessibility to facilities and amenities.

Leaving the outliers out will certainly improve the model's ability to predict prices of houses in the normal range.

However leaving them out will cause the model to fail in predicting prices of houses that fetch high prices.

The assessment matrix selected for this model is rmse and one outlier can cause a huge rise in rmse thus the outliers will be retained to train the model.

Summary Statistics

	Full Bath	Half Bath	Bedroom AbvGr	Kitchen AbvGr	TotRms AbvGrd	Fireplaces	Garage Yr Blt	Garage Cars	Garage Area	Wood Deck SF
count	2051.000000	2051.000000	2051.000000	2051.000000	2051.000000	2051.000000	1938.000000	2051.000000	2051.000000	2051.000000
mean	1.577279	0.371039	2.843491	1.042906	6.435885	0.590931	1978.682663	1.776694	473.441736	93.833740
std	0.549279	0.501043	0.826618	0.209790	1.560225	0.638516	25.458580	0.764367	216.132969	128.549416
min	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000	1895.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	2.000000	1.000000	5.000000	0.000000	1961.000000	1.000000	319.000000	0.000000
50%	2.000000	0.000000	3.000000	1.000000	6.000000	1.000000	1980.000000	2.000000	480.000000	0.000000
75%	2.000000	1.000000	3.000000	1.000000	7.000000	1.000000	2002.000000	2.000000	576.000000	168.000000
max	4.000000	2.000000	8.000000	3.000000	15.000000	4.000000	2207.000000	5.000000	1418.000000	1424.000000

Summary Statistics shows an error in the Garage Yr Built.

	Year Built	Year Remod/Add	Garage Yr Blt
1699	2006	2007	2207.0

Checking the max value for year house is built or remodel returns 2010.

Checking the entry reveals that likely is a wrong data entry and imputed the value 2007.

Pairplot for Numeric Variables
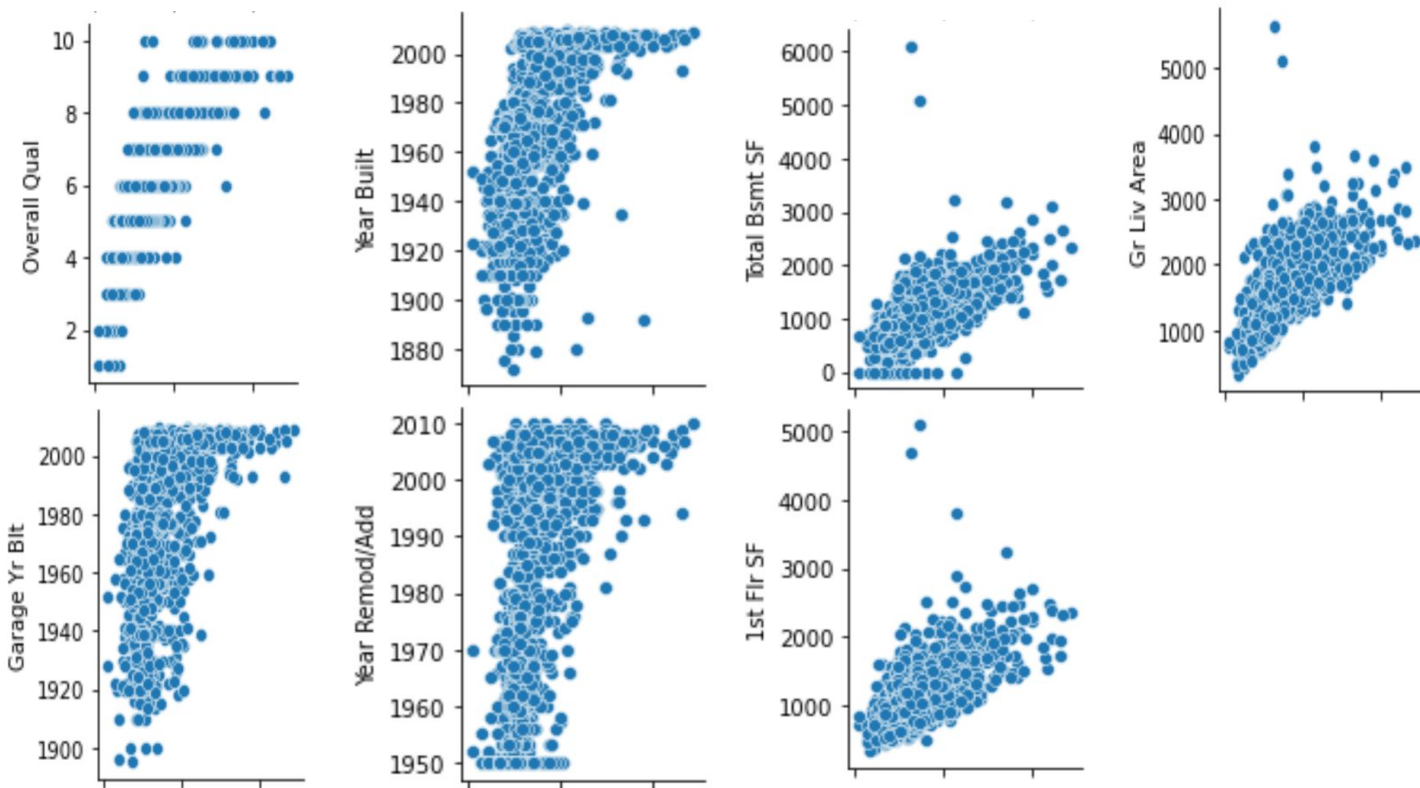
Using pairplot to plot numeric variables against SalePrice

Pairplot reveals 7 variables that are linearly related to SalePrice

Pairplot for other Variables
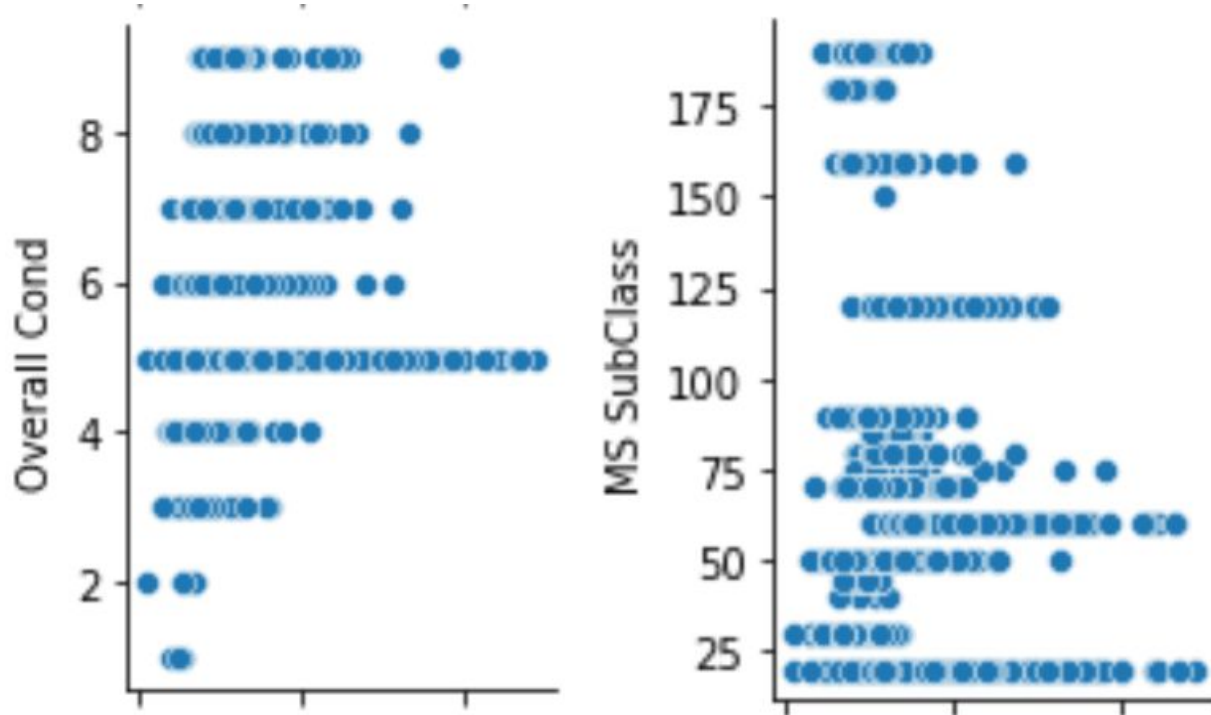
2 other variables of interest are overall cond and MS subClass.

Normally the higher the overall condition of the house, the higher will be the SalePrice, however this shows that other factors like location of amenities also play an important part in the SalePrice.

MS SubClass shows 2 particular housing type fetch high SalePrice.

Heat Map

Heat Map shows the variables on slide 14 all has high correlation with SalePrice that is at least higher than 0.5.

The 2 variables in slide 15 however has negative correlation.
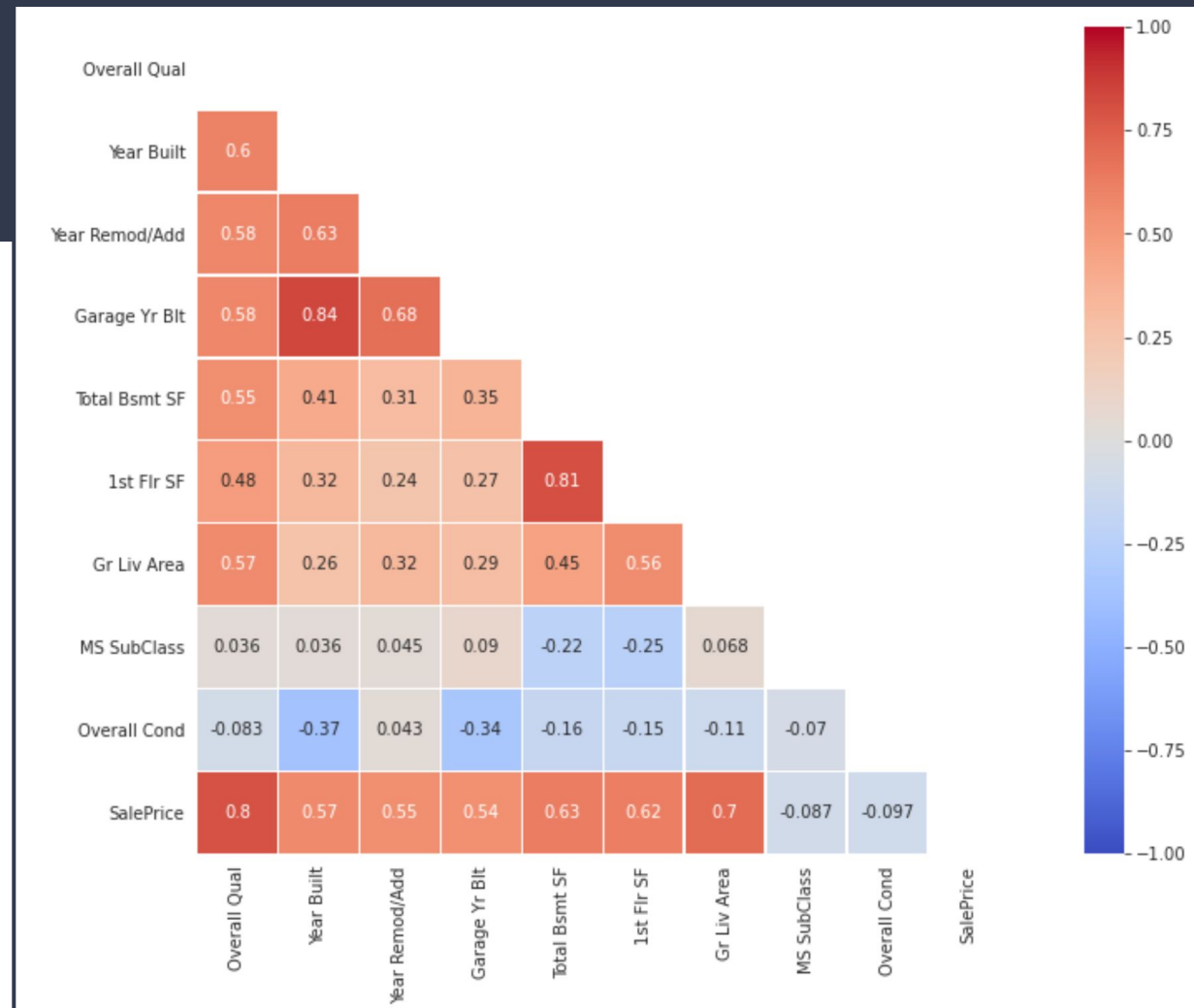
Feature Engineering

Data processing

Discrete & Nominal Data:

- All the categorical data are in string format unable to process.
- One hot encode to produce binary output.

Ordinal data:

- Note that not all ordinal features have the same scale.
- Remap all the data on the same performance scale of 1-to-5.

Data processing

- Imputation was done to resolve the Null values the data in the previous chapter.
- Next we will transform the sales price to a logarithm scale to fit it to a normal distribution.

House Sales Price Distribution

House Sales Price Distribution
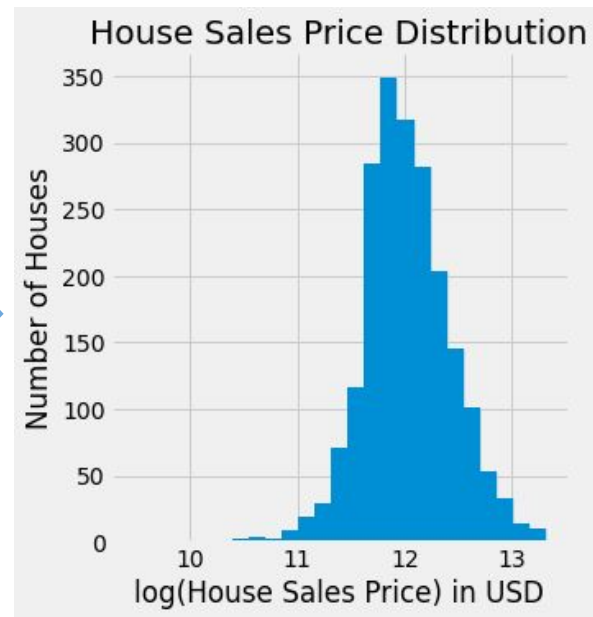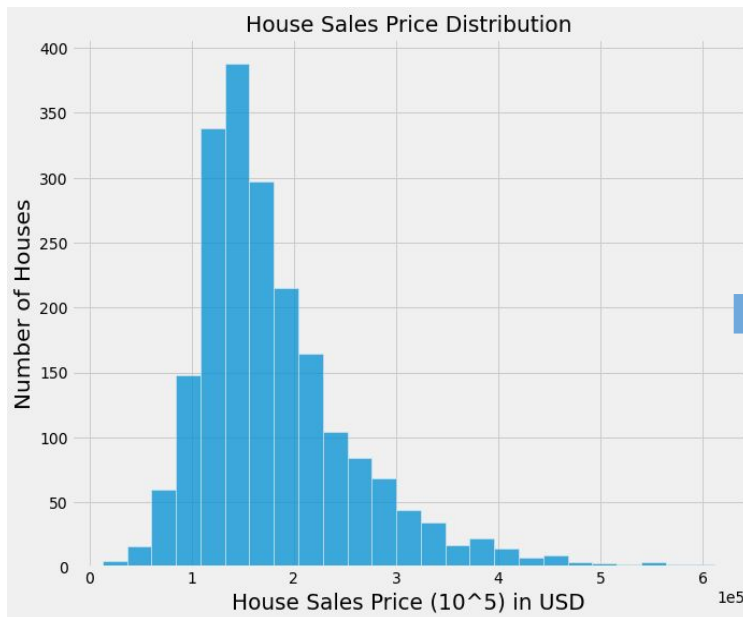
Top features impacting Sales Price in Ames, Iowa

Ranking	Features	Ridge Coefficient
1	Gr Liv Area	0.129777
2	Overall Qual	0.073051
3	Year Built	0.062722
4	Total Bsmt SF	0.054044
5	Year Remodel/Add	0.041420
6	Fireplace Qu	0.039419
7	Garage Area	0.037992
8	Kitchen Qual	0.031634

★ Key features that impacts the sales price in Iowa are mainly Ground living area, Overall quality, year built and etc.

Heatmap representation
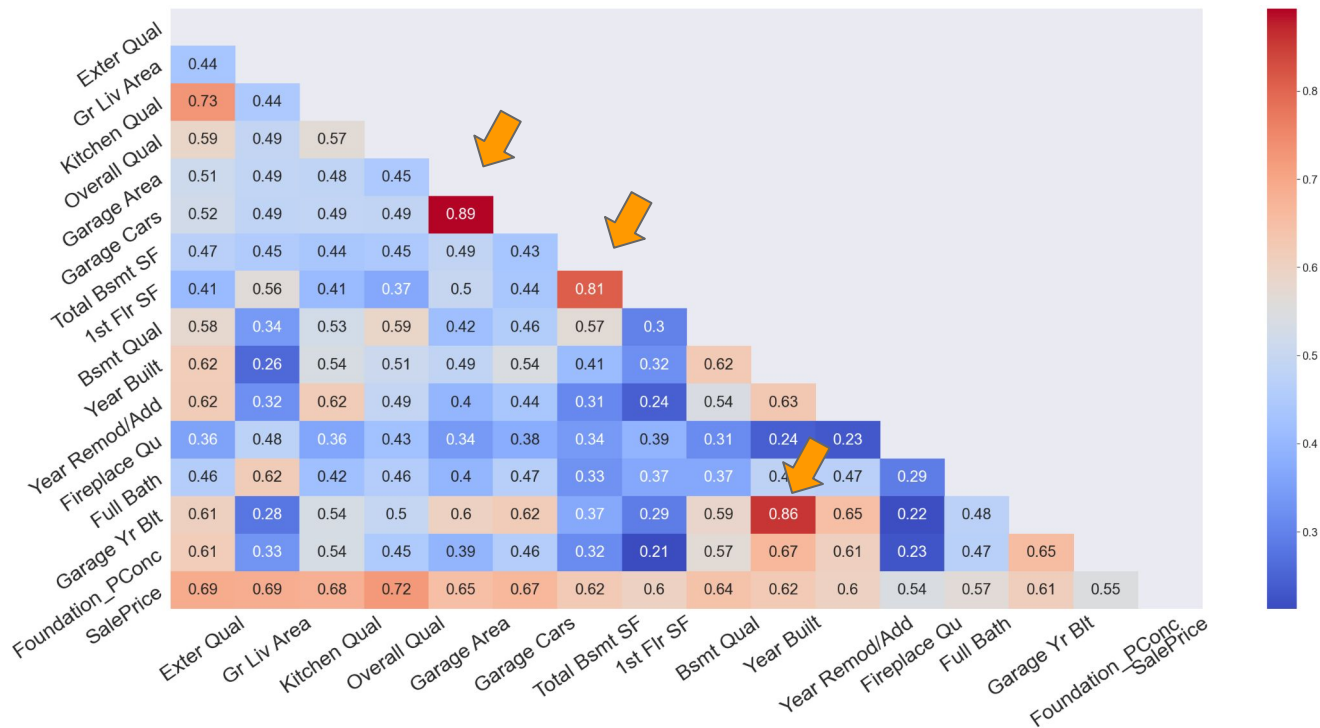
Highest pearson coefficient correlation to sales price, goes to Overall Quality,Ground Living Area and Kitchen quality.

From here, we can also see multicollinearity in some of the top features.
- Garage area and garage cars
- Total bsmt SF and 1st Flr SF
- Year built and Garage year built

Correlation of Gr Living area/Overall Quality to Sales Price

Iowa Real Estate SalesPrice vs Ground Living Area
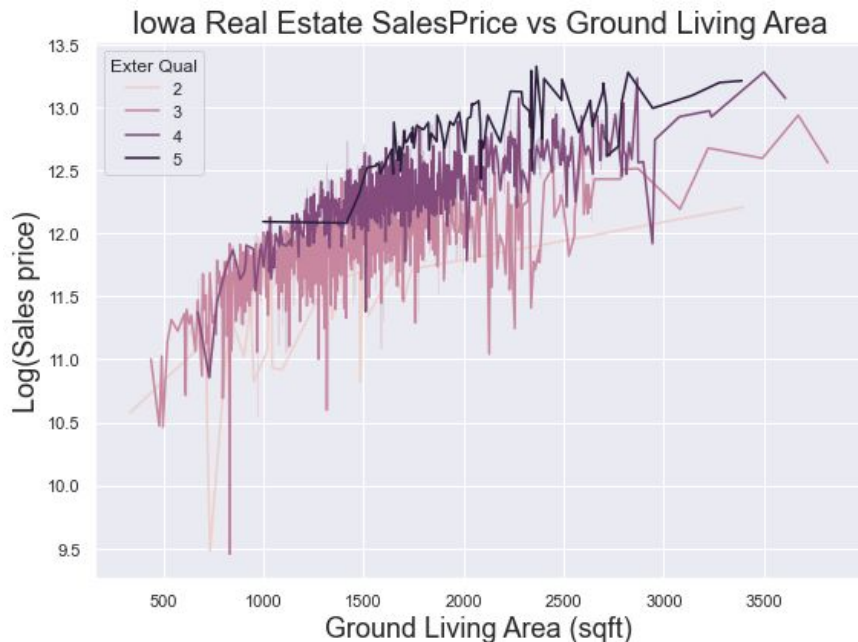
- From the lineplot, we can see an increasing trend for the log of salesprice as ground living area increases.
- For every 1k sqft increase of living area, we expect prices to shoot up by $160,921
- In fact, we can see on top of ground living area, external quality has a different baseline for saleprice.

Correlation of Gr Living area/Overall Quality to Sales Price

Iowa Real Estate SalesPrice vs Overal Quality
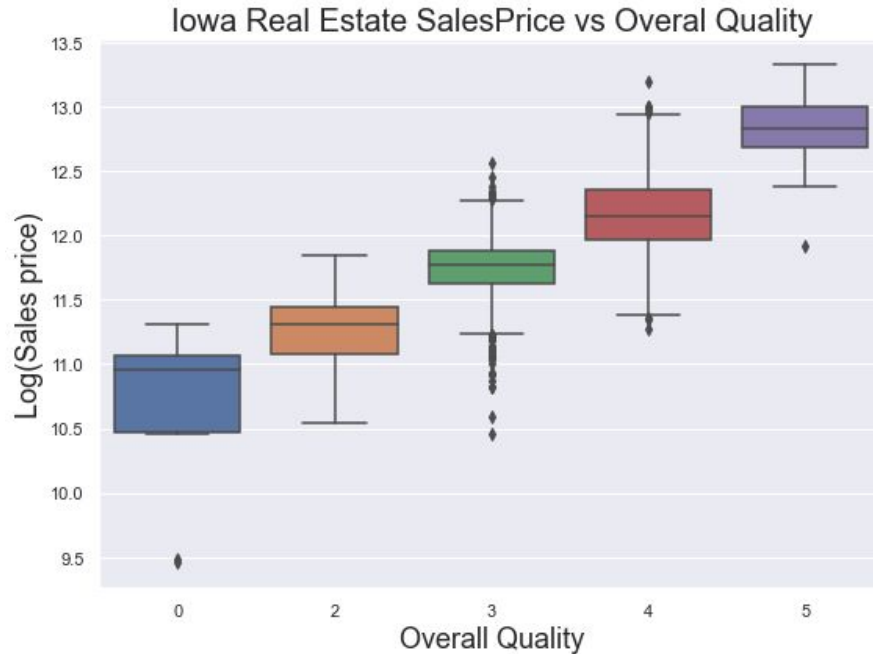
- The overall quality here rates the overall material and finish of the house.
- We can see that from a 5 point scale of 1-poor to 5-excellent, the log sales price is moving towards a positive increasing trend.

SalesPrice vs Year Built

Iowa Real Estate Sales Price vs Year Built

Iowa Real Estate Sales Price vs House Age
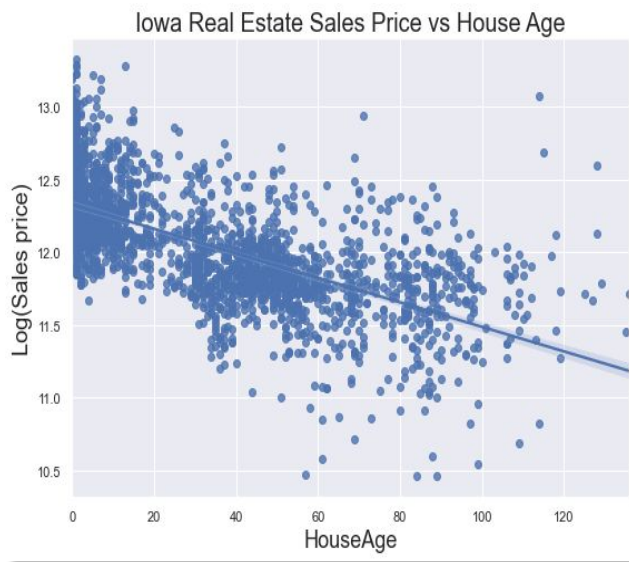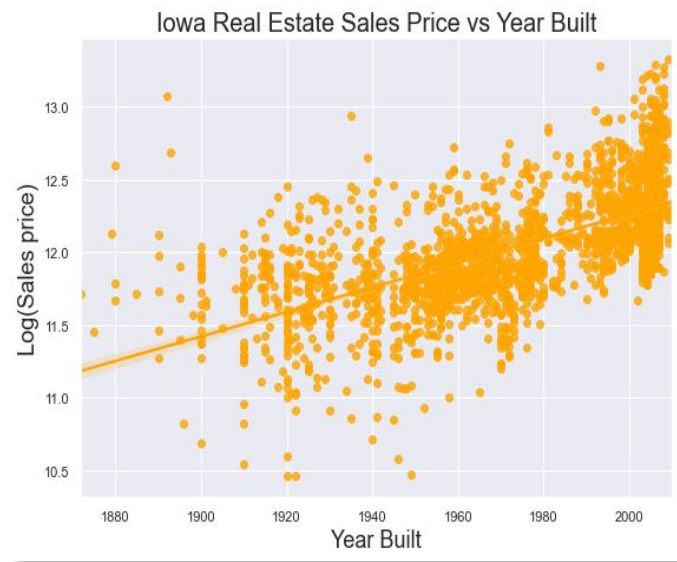
- **Newer houses** with higher built year also gets the highest sale prices.
- This feature would add more value if we remodel it to the **delta** of the **year sold & year built.**

Sales Price vs Total Basement SF
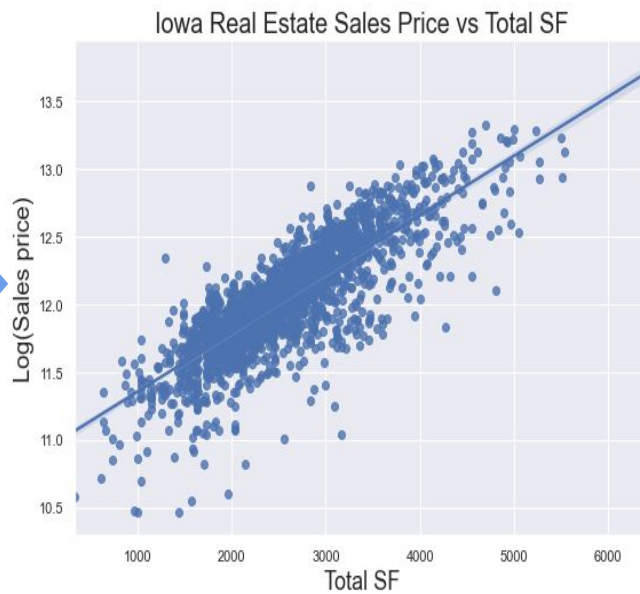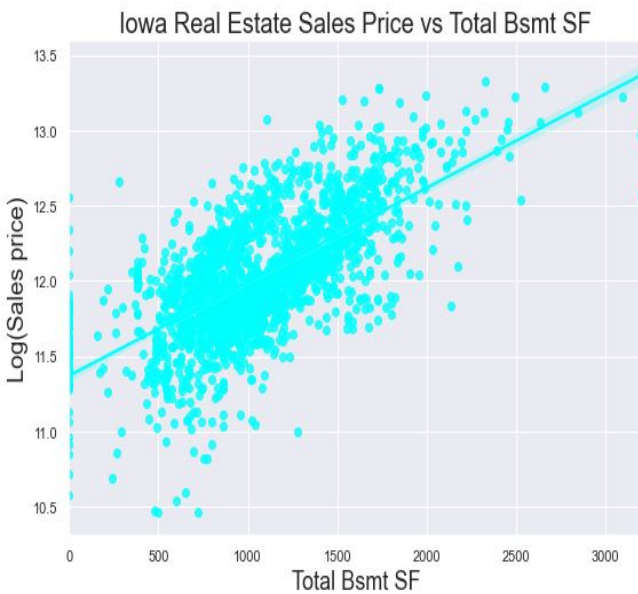
★ We also see a positive correlation for Total Bsmt SF to Sale Price.

★ Since buyers don't usually look at individual SF but total SF of the property.

★ Adding up all basement SF, 1st floor and 2nd floor will give us a bigger picture on the total SF.

Sales Price vs Year Remodelled/Add

Iowa Real Estate Sales Price vs Year Remodel/Add
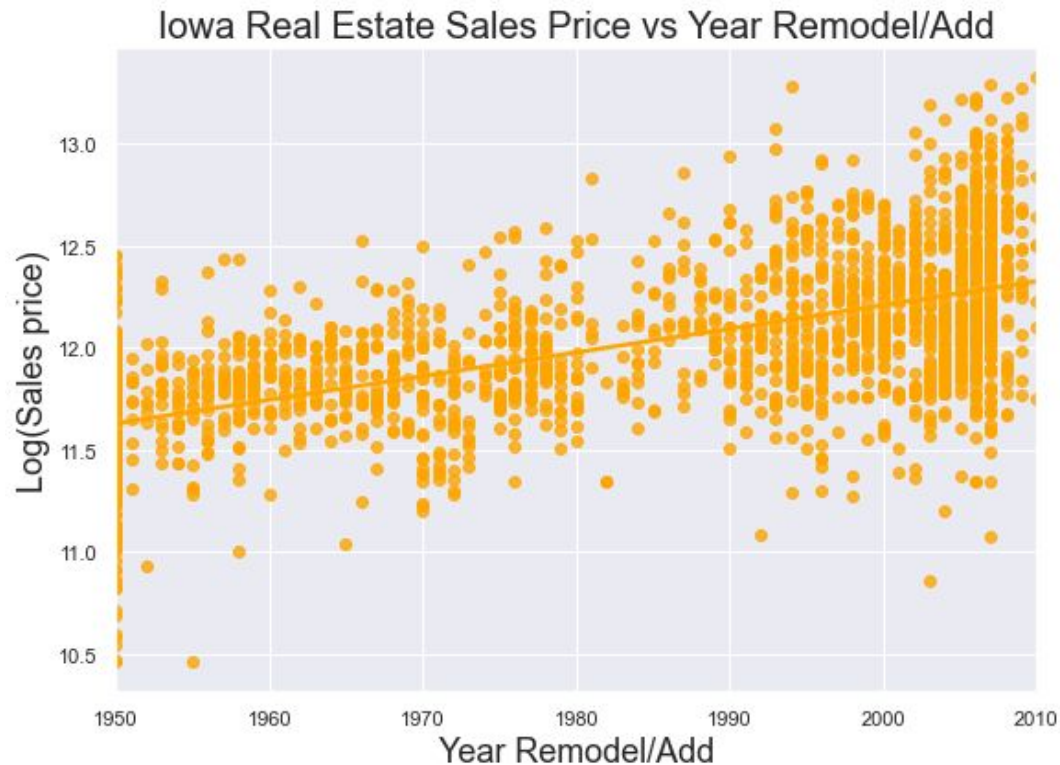
- The year remodelled or added has a positive correlation to log of sales price.

Sales Price vs Fireplace Quality

Iowa Real Estate SalesPrice vs Fireplace Quality
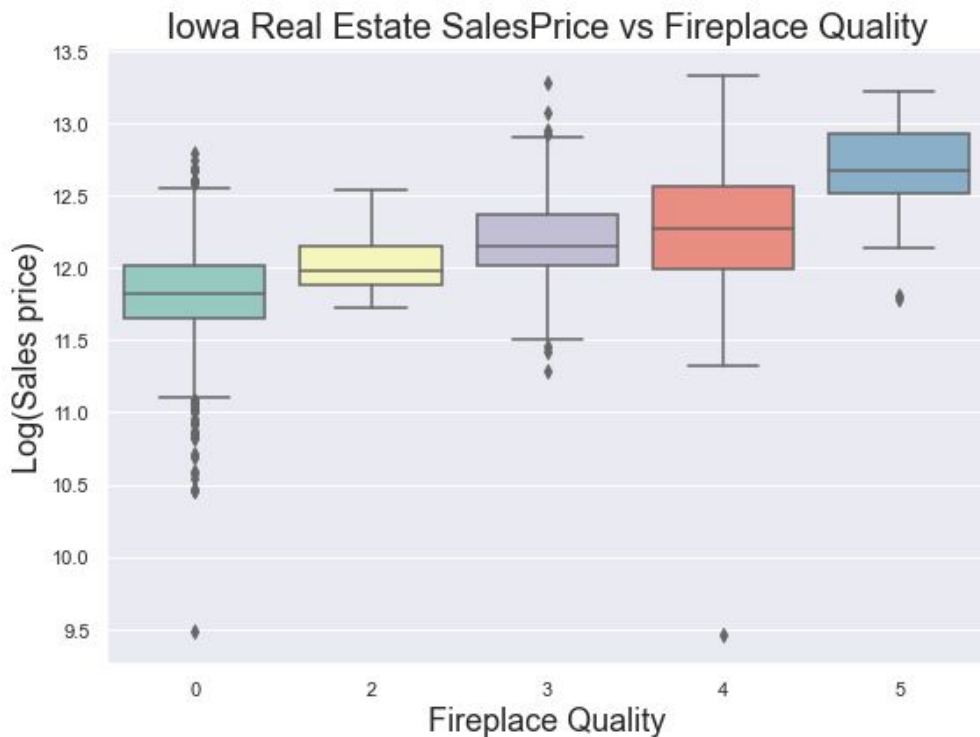
The data also suggests that fireplace quality have an influence on increasing the sales price of the property in Iowa

Sales Price vs Garage Area

Iowa Real Estate Sales Price vs Garage Area
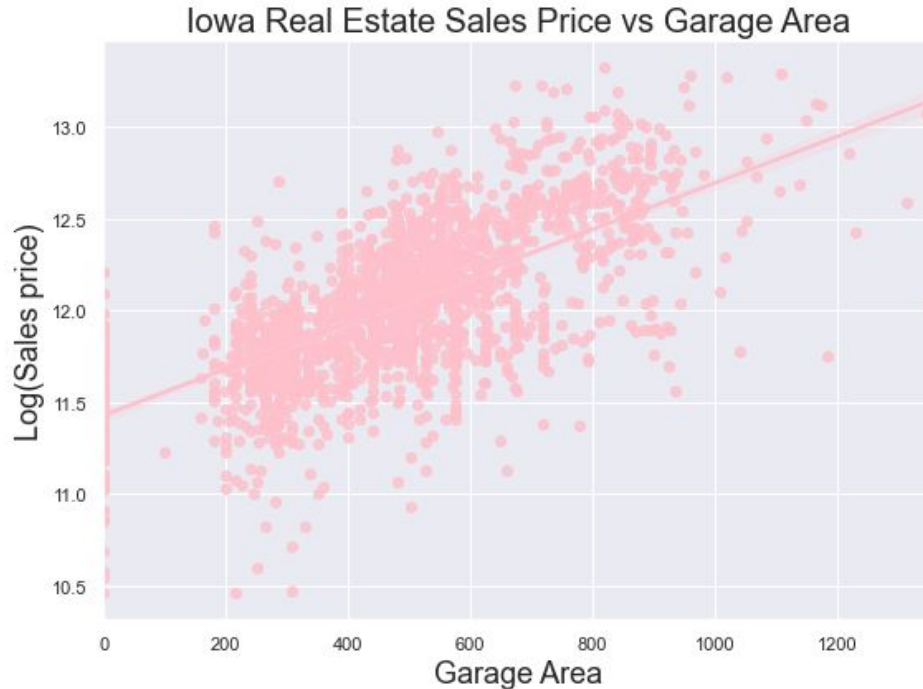
The sales price also increases with increase of garage area which matches that of overall ground living area as well.

Sales price vs Kitchen quality

Iowa Real Estate SalesPrice vs Kitchen Qual
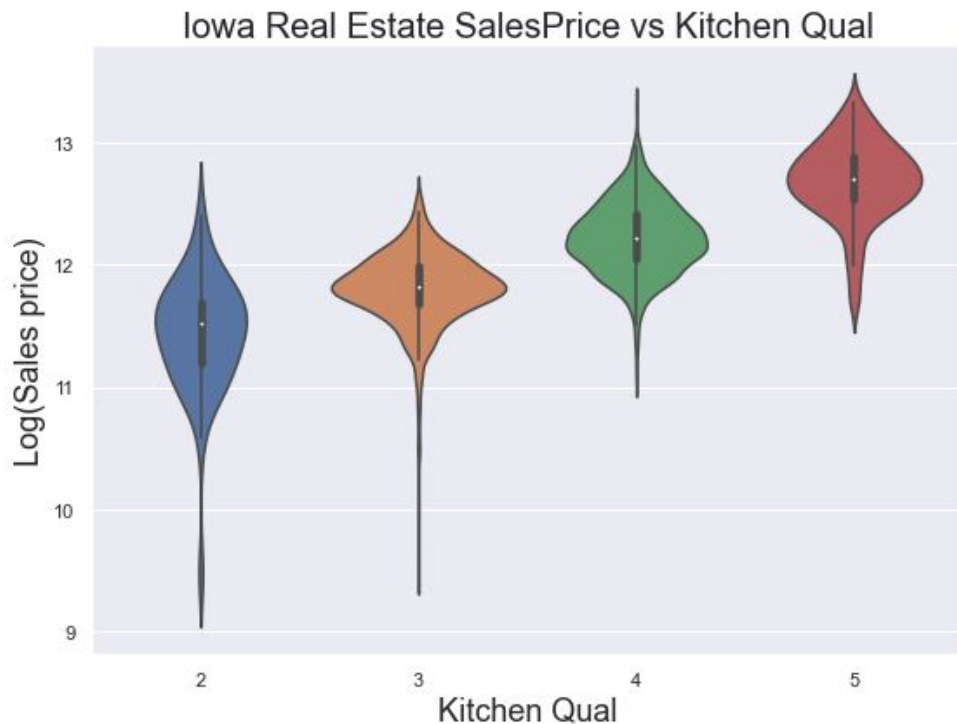
From the violinplot, we can see the requirement for better quality kitchen greatly increases the price of the real estate.

Model Selection

Models selection:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. ElasticNet

Scaling data (Before) – Standard Scaler
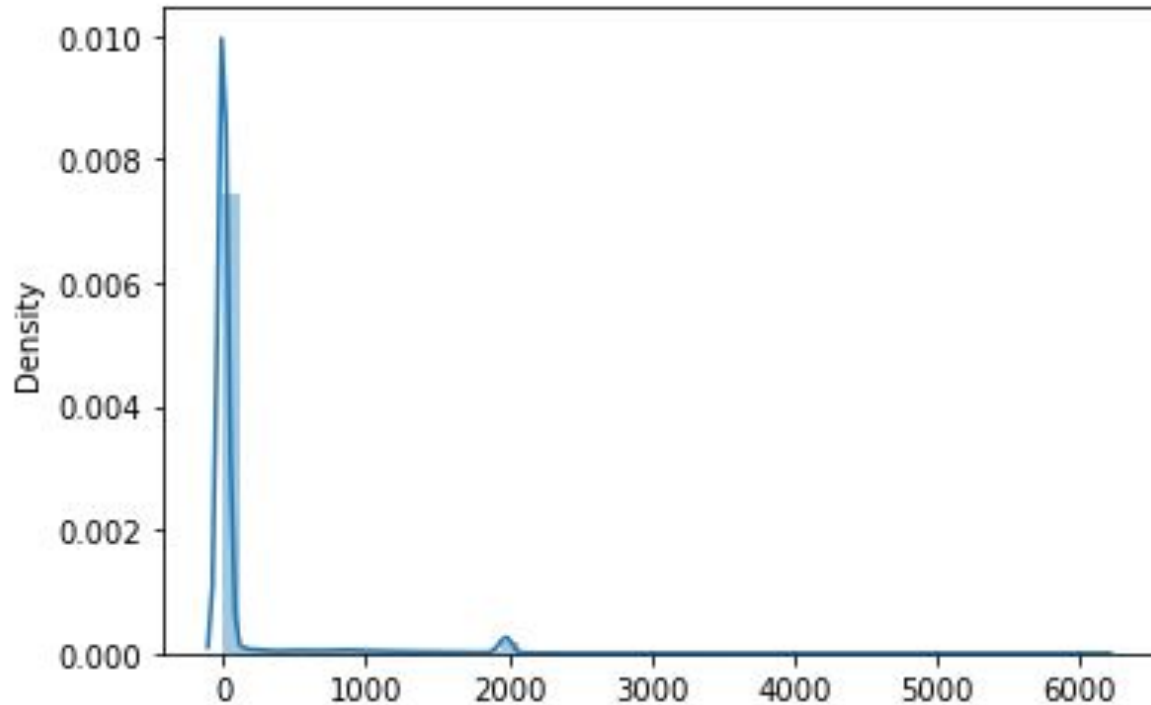
Distribution plot was plotted to represent data distribution.

Before scaling data

Scaling data (After) – Standard Scaler
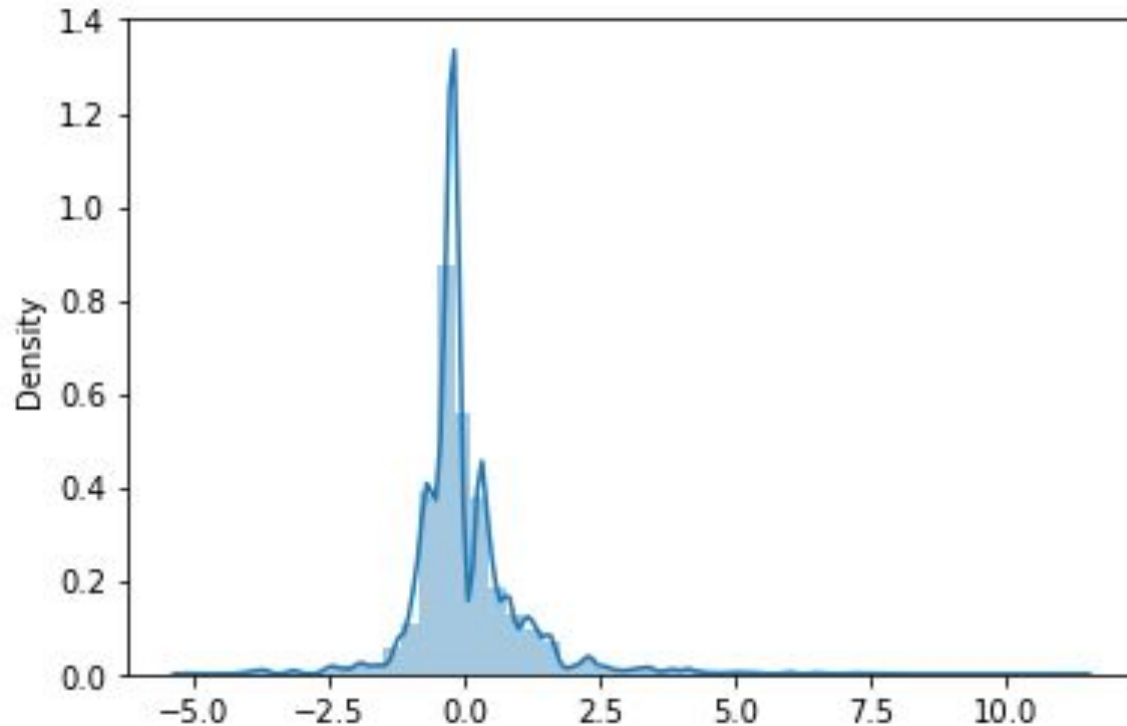
Distribution plot was plotted to represent data distribution.

After scaling data

Normally distributed

Ordinary Linear Regression
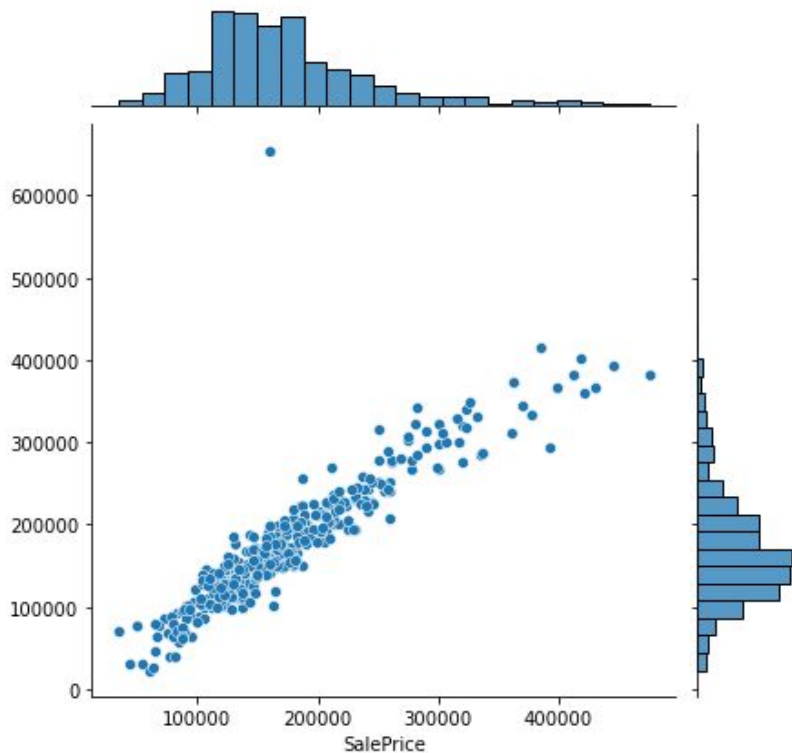
Using Ordinary Linear Regression,

1. Train score was 0.90
2. Test score was 0.79

This model showed evidence of overfitting.

Bias vs Variance – Avoiding overfitting

Lasso Regression
Ridge Regression

Regularization techniques were implemented to improve the model fit.

1. L1 Regularization
2. L2 Regularization

Lasso Regression

L1 Regularization

L1 Regularization Lasso Regression

L1 Regularization model was first used to attempt to penalize beta coefficients or zero them.

LassoCV was used to select the best parameters for a suitable bias-variance tradeoff.

Lasso: y_test vs y_pred jointplot
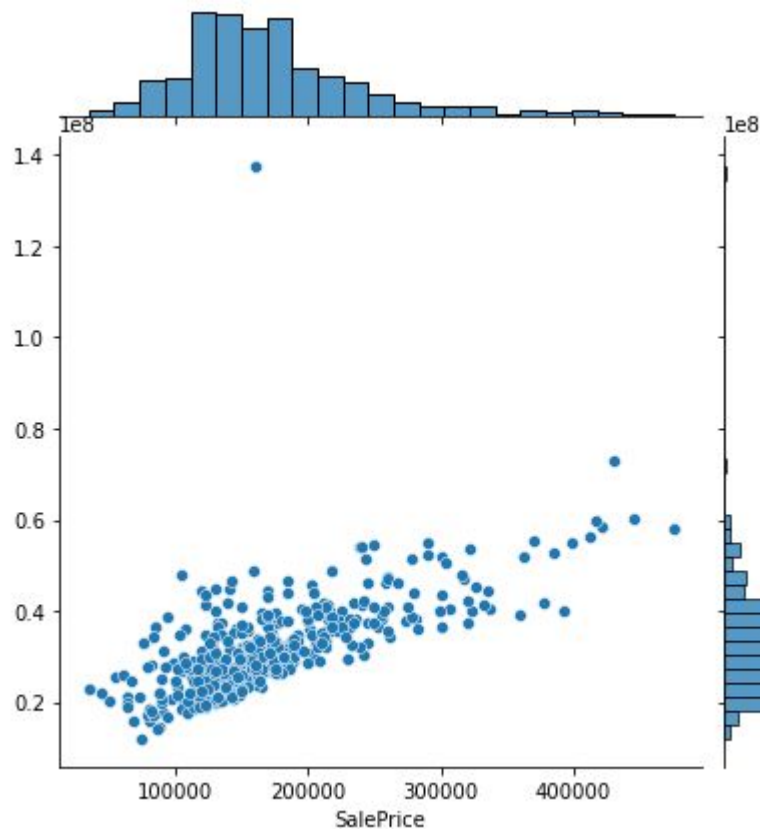
L1 Regularization
Lasso Regression

Scoring:
Lasso cross validation score (cv = 10) = 0.86
Lasso r2 test score = 0.77

Lasso: y_test vs y_pred jointplot
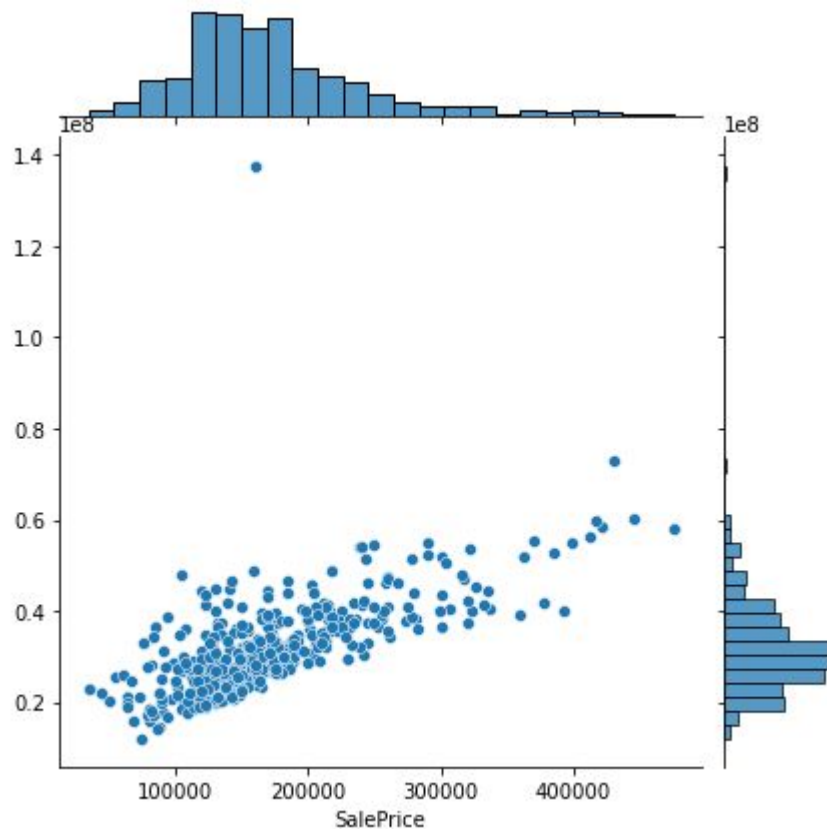
Ridge Regression

L2 Regularization

L2 Regularization Ridge Regression

L2 Regularization model was used to appropriately weigh the model's variables according to its importance.

RidgeCV was used to select the best parameters for a suitable bias-variance tradeoff.
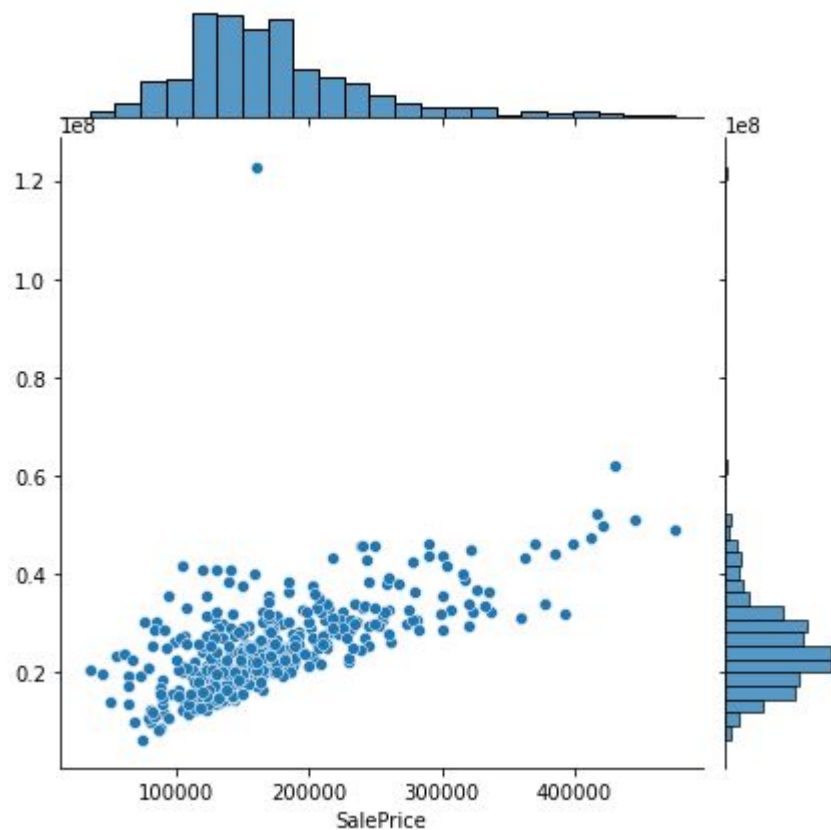

Ridge: y_test vs y_pred jointplot
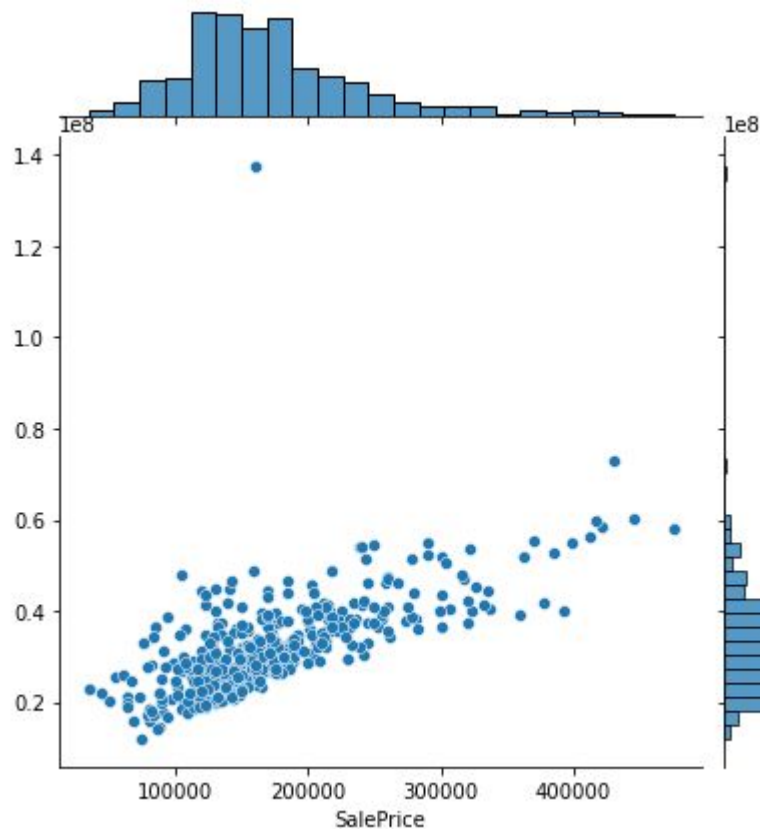
L2 Regularization Ridge Regression

Scoring:
Ridge cross validation score (cv = 10) = 0.86
Ridge r2 test score = 0.78
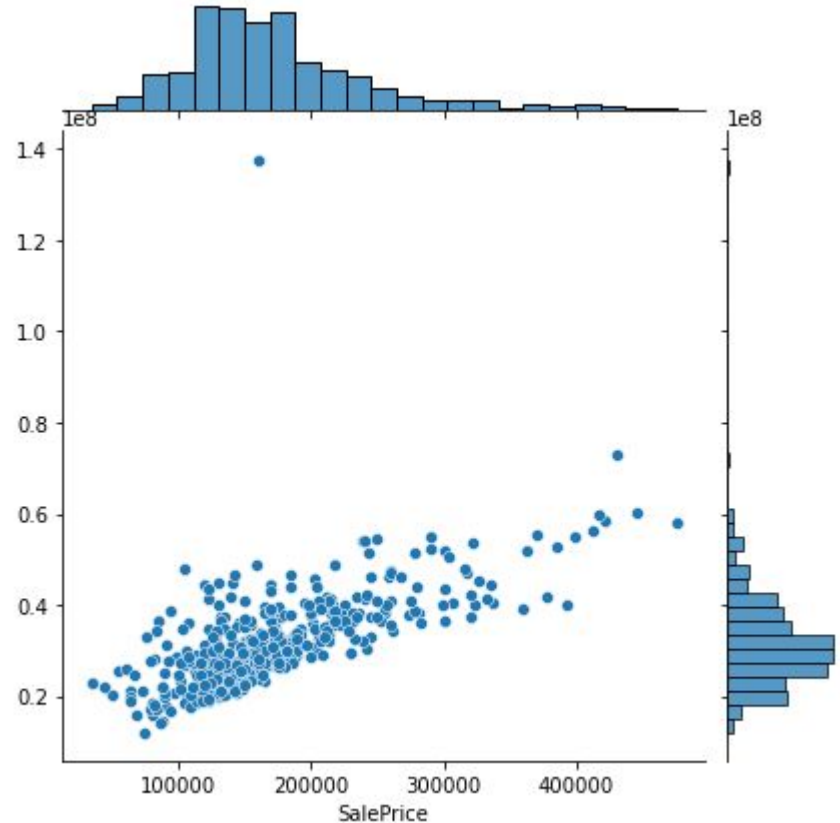

Ridge: y_test vs y_pred jointplot

ElasticNet

L1 & L2 Regularization

L1 & L2 Regularization ElasticNet Regression

L1 & L2 Regularization model; Used 2 different regularization techniques to achieve a better score

ElasticNetCV was used to select the best parameters for a suitable bias-variance tradeoff.

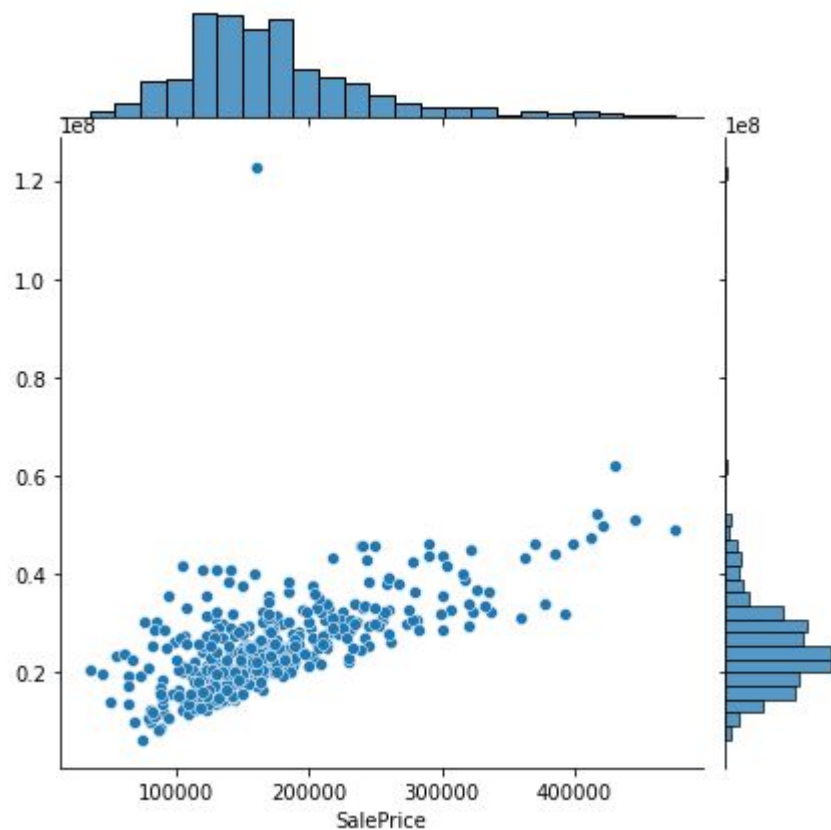ElasticNet: y_test vs y_pred jointplot

L1 & L2 Regularization ElasticNet Regression

Scoring:
ElasticNet cross validation score (cv = 10) = 0.86
ElasticNet r2 test score = 0.61

ElasticNet: y_test vs y_pred jointplot

Residuals Plot – Detecting outliers and abnormalities
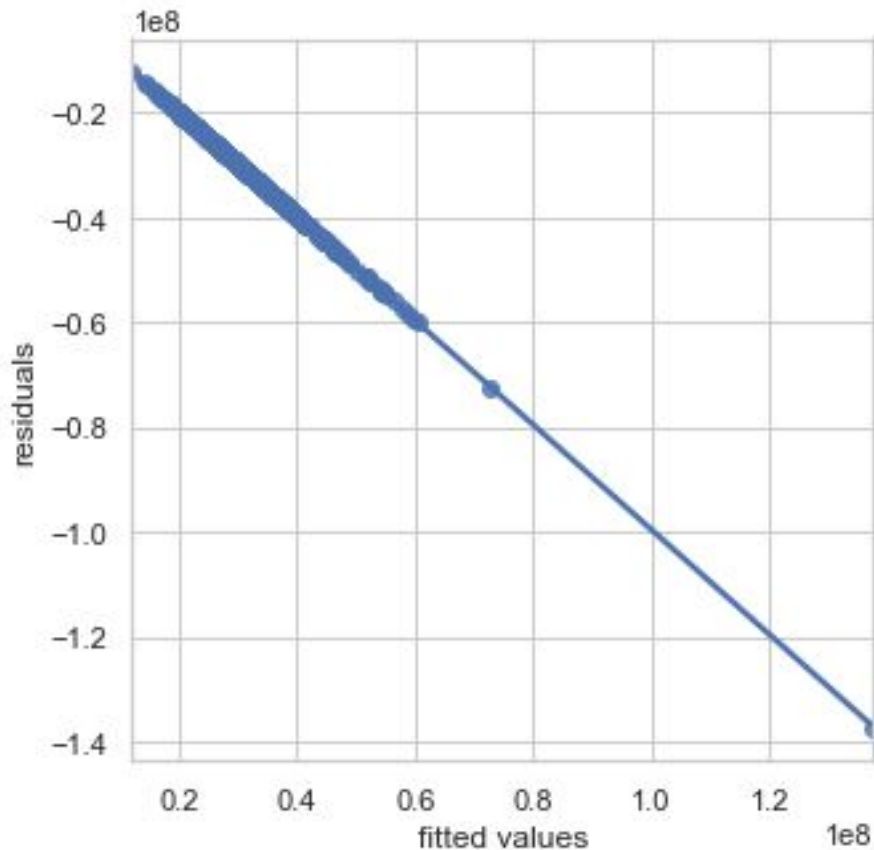
Lasso Residuals Plot

Lasso

y_pred vs (y_test - y_pred)

No abnormalities detected in the chart.

Equal distributed residuals

Residuals Plot

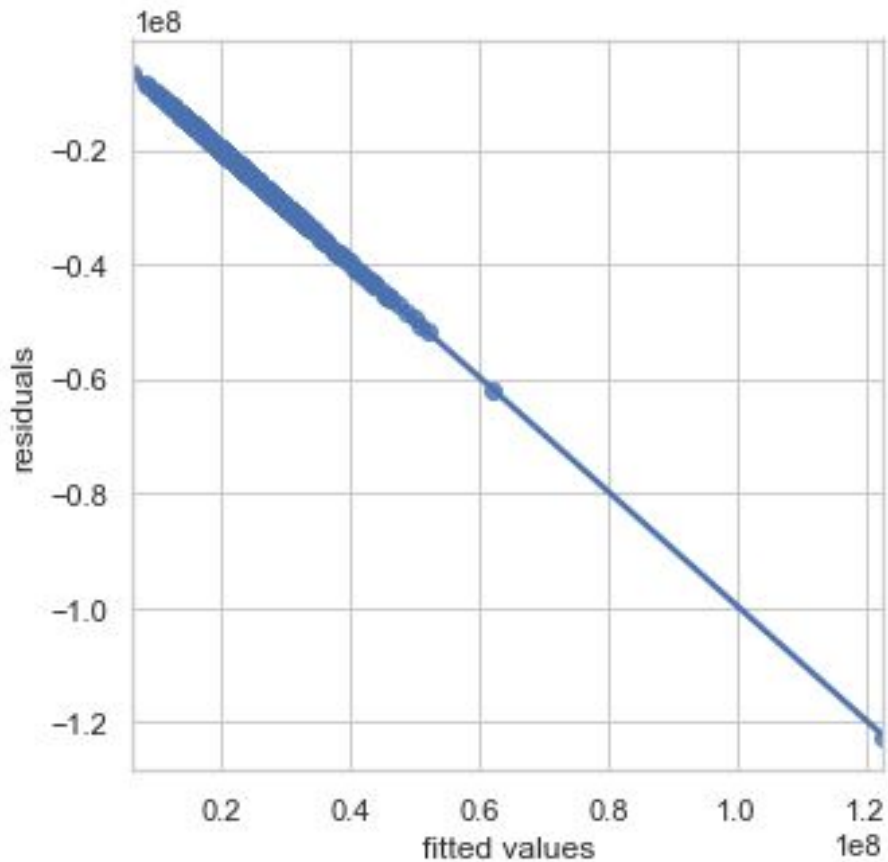Lasso: Residuals vs Fitted values

Ridge Residuals Plot

Ridge

y_pred vs (y_test - y_pred)

No abnormalities detected in the chart.

Equal distributed residuals

Residuals Plot

Ridge: Residuals vs Fitted values
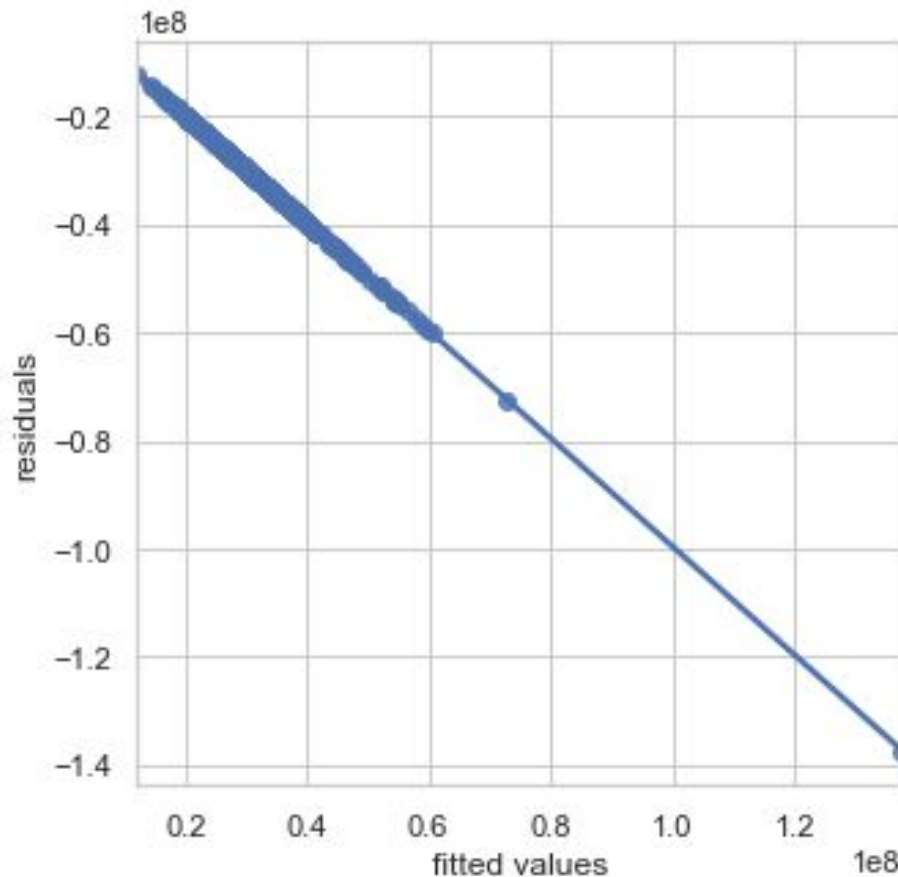
ElasticNet Residuals Plot

ElasticNet

y_pred vs (y_test - y_pred)

No abnormalities detected in the chart.

Equal distributed residuals

Residuals Plot

ElasticNet: Residuals vs Fitted values

Histogram of Residuals Plot – Detecting magnitude of outliers

Lasso Histogram of Residuals

Lasso

Low frequency of outliers of up to of:
-400,000 and +200,000

Histogram shows residuals
normally distributed around 0.

Histogram of Residuals

Lasso:

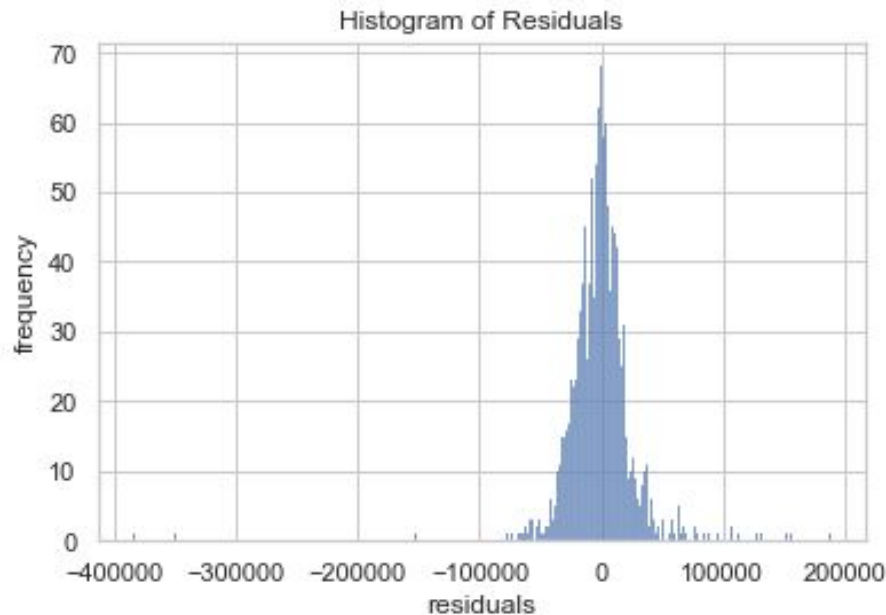Histogram of Residuals

Ridge Histogram of Residuals

Ridge

Low frequency of outliers of up to of:
-400,000 and +200,000

Histogram shows residuals
normally distributed around 0.

Histogram of Residuals

ElasticNet Histogram of Residuals

ElasticNet
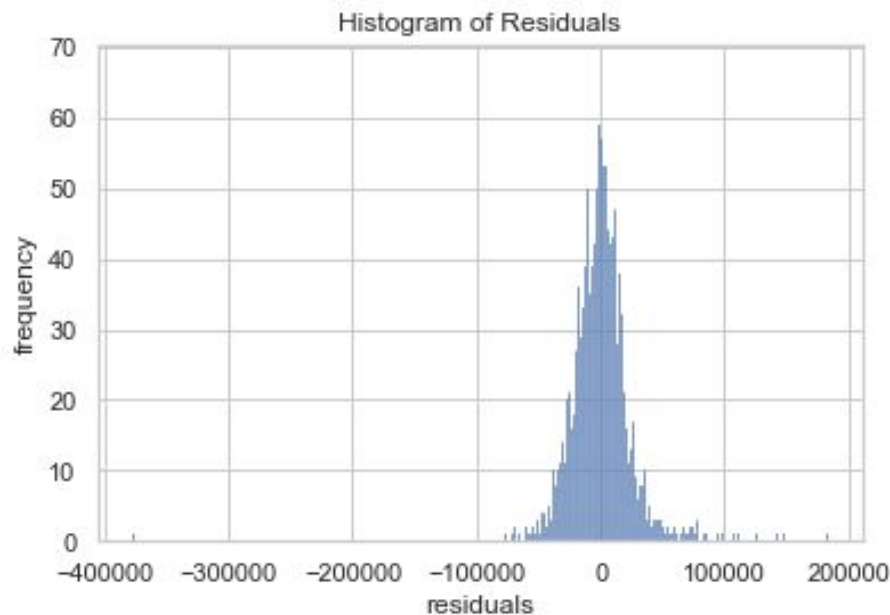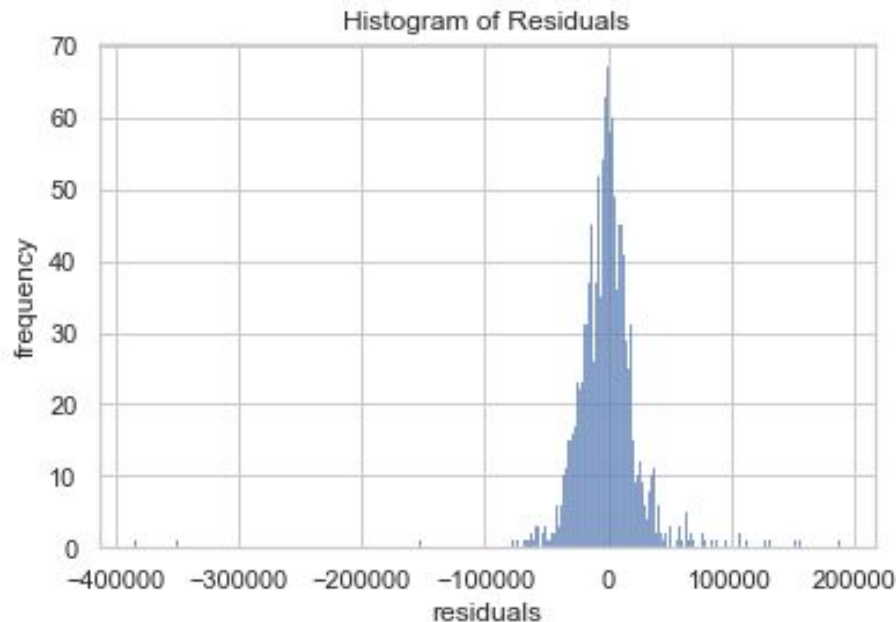
Low frequency of outliers of up to of:
-400,000 and +200,000

Histogram shows residuals
normally distributed around 0.

Histogram of Residuals

ElasticNet:

Histogram of Residuals

Model Selection for final sales price prediction

Model Selection – Ridge Regression

Lasso cross validation score = 0.8606

Lasso R2 test score = 0.771

Ridge cross validation score = 0.8601

Ridge R2 test score = 0.787

ElasticNet cross validation score = 0.8606
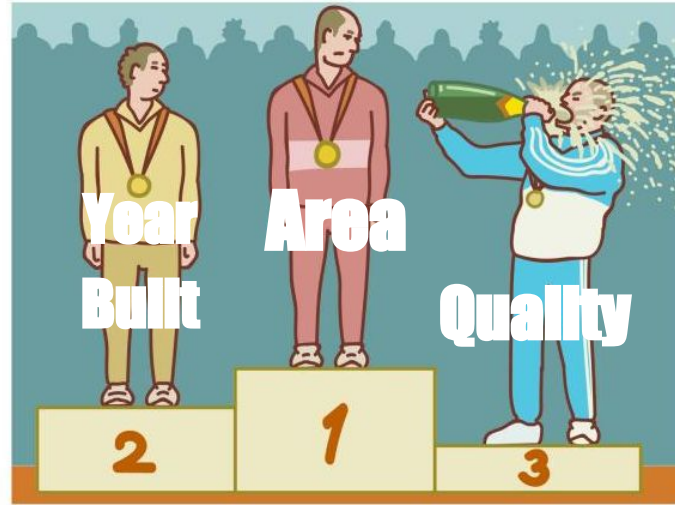
ElasticNet R2 test score = 0.612

Hence, Ridge Regression was selected as it had comparable cross validation score & a higher R2 score than other models.

Conclusion

Conclusion

Best features that affect house sale price

Recommendation for house owner

Ways to retain housing value:

1. Maintain quality of the house.
2. Schedule maintenance such as:
   - Repaint the house every 5 - 10 years.
   - Spring cleaning.
3. Use quality furnishing.
4. Remodel the house.

No strong correlation features :

- Swimming pool
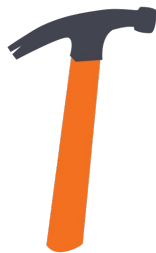- Porch
- Misc Value
- Month Sold

Recommendation for investors

House Flipping

Buy low quality houses at lower price

+

Renovation

=

Re-sell the house at higher value.

+

Profit

Housing Price analysis model

Ridge regression model.

- Ridge is suited the best for the housing price prediction.
- Highest R2 score = 0.787.
- Most suitable to be used at the same district or city area.

Model Improvement

Effective/Weighted age of the property.

The effective age is calculated by taking the percentage of the remodeling or modernization in relation to the whole.

For example:

70% house structure = 50 years,

25% house structure = 10 years

5% house structure = 5 years.

Effective age = 0.7*50 + 0.25*10 + 0.5*5 = **40 years**

Data required:

1. Percentage of remodelling.
2. Area remodelled.

References

Sources:

https://www.santacruzcountyaz.gov/208/Effective-Weighted-Age#:~:text=The%20effective%20age%20is%20calculated,0.50%20X%2040%20%3D%2020.0%20years

https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp

https://www.newrez.com/blog/home-buying-selling/5-things-that-influence-home-prices-in-2021/

https://www.investopedia.com/articles/mortages-real-estate/11/the-truth-about-the-real-estate-market.asp

https://www.whitehouse.gov/cea/written-materials/2021/09/09/housing-prices-and-inflation/

https://www.americanactionforum.org/insight/understanding-the-national-increase-in-house-prices/

https://www.opendoor.com/w/blog/factors-that-influence-home-value