DSI GROUP - Project 2

By: Jude Darren
Tan Kar Gim
Andrea Wong

TABLE OF CONTENTS

Problem Statement

Our **real estate agency** in **Ames, Iowa,** is working on a solution to help our in-house real estate agents in comparative market analysis to determine a fair and competitive offering price. With our model, we plan to support our in-house real estate agents in the following areas:

- offer **proprietary estimate of a property's value** based on the **key features** of the property
- a useful reference point in **assessing the fairness of a home's price**

Methodology

Define Objectives
Define the problem to be solved and create a clear objective

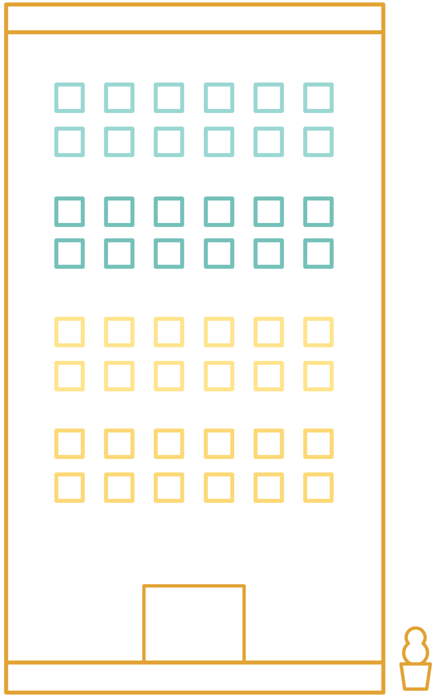
Collect, Prepare & Manipulate data
Collect and prepare the data to be used for modelling

Data Exploration & Analysis
Find significant patterns and trends using statistical methods and visualisations

Modelling
Build, fit and validate model

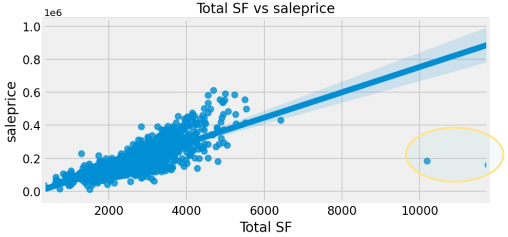
Model Optimization
Assessing model performance and make changes to improve model

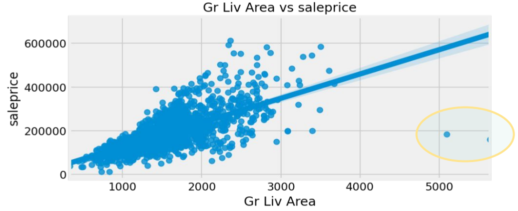
Data Preparation



Outliers

Gr Liv Area > 5000
Total SF > 10000



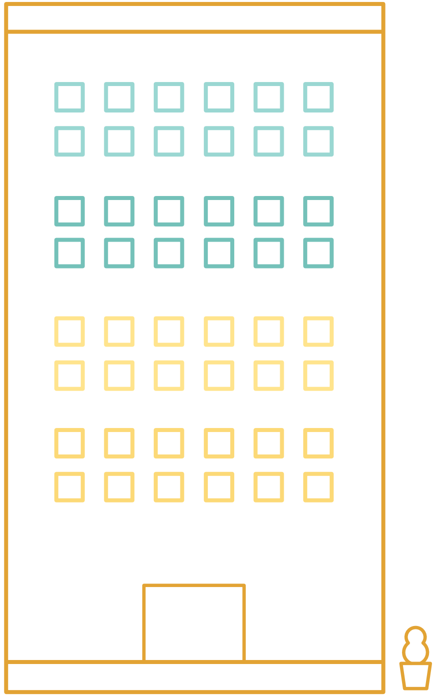


Removal of Rows

0.15% of samples have incomplete data. Rows with 1 or 2 missing values will removed

features	null_values	decision
BsmtFin SF 2	1	Remove row
Bsmt Unf SF	1	Remove row
Total Bsmt SF	1	Remove row
Bsmt Full Bath	2	Remove row
Bsmt Half Bath	2	Remove row

Data Preparation



Remove Columns

Lot Frontage - 16.1% missing

features	null_values	decision
Lot_Frontage	330	Drop column

One-hot encoding

Categorical variables will be hot encoded before modelling

Kitchen Qual
Gd
TA
Fa
Ex

→

Kitchen Qual_Gd	Kitchen Qual_TA	Kitchen Qual_Fa	Kitchen Qual_Ex
1	0	0	0
0	1	0	0
0	0	0	1
1	0	0	0

EDA - Dataset breakdown

Housing Dataset

80 features characteristics of a property

Categorical Features (52.5%)

Represented as ‘strings’ or ‘categories’ and are finite in number

Numerical Features (47.5%)

Integers and floats are the most common and widely used numeric data types.

01

Histograms

show the no of observations in each category

- Skew of data

02

Boxplots

Visualize the distribution of data through quartiles

03

Heatmap

Shows multicollinearity between features through colour

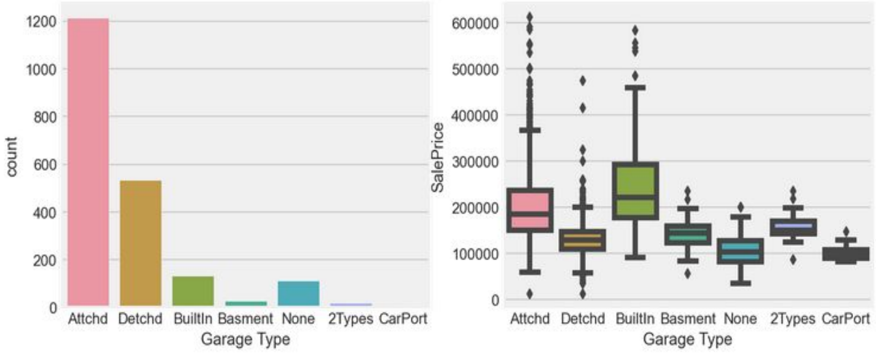
04

Scatterplot

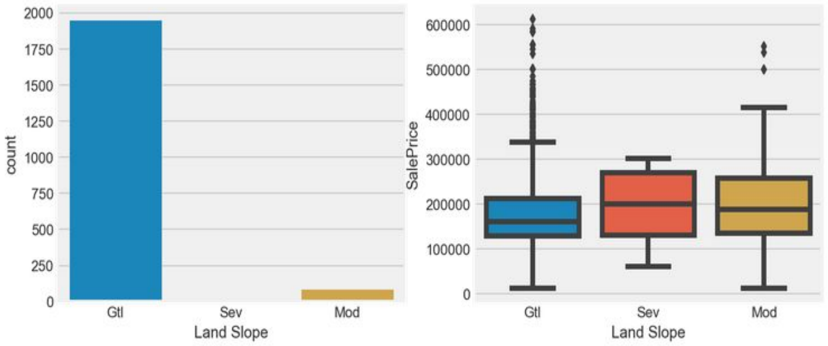
Plotting of data points on a horizontal and a vertical axis show relationships between two numeric variables

Which categorical features will be useful?

Useful Feature



Less Useful Feature

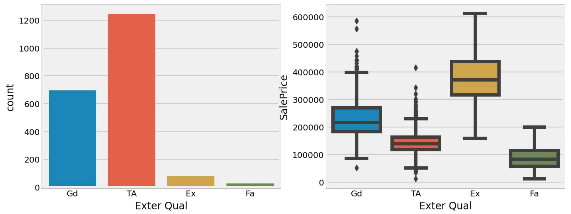


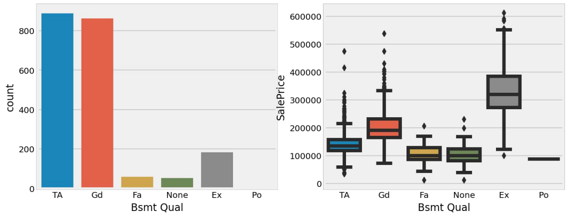
Key properties of useful categorical features:

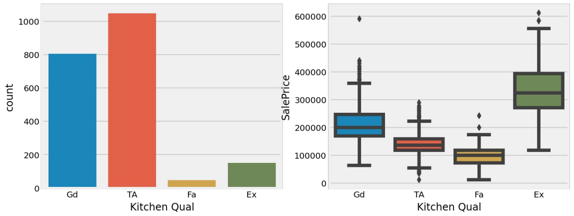
1. High variance
2. Significant difference median

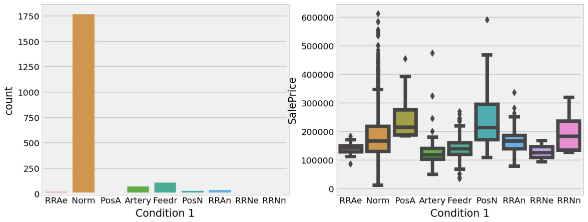
Which categorical features will be useful?

Useful Feature

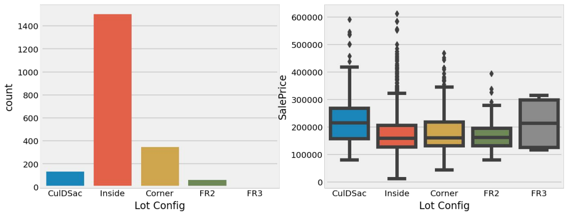


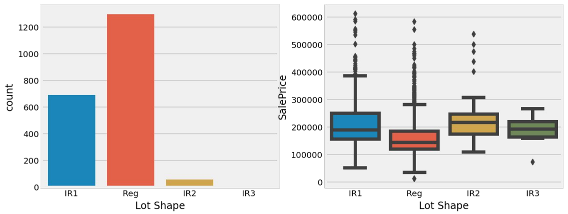






Less Useful Feature





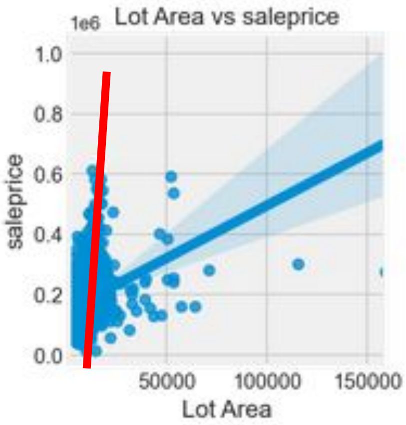
Categorical features considered -initial observations

features	reasons for considering this variable
Neighborhood	The boxplots were able to show that the prices in different neighborhoods vary. The median housing price is much higher in certain neighborhoods as compared to others. Furthermore, the barcharts also show that all the neighborhoods were represented in training data.
Condition 1	While most of the selections in the training data selected 'Norm', it was noticed in the boxplots that when other selections were chosen for example, 'PosN', it can affect the median pricing of the house.
Condition 2	Same as condition 1, while most of the selections in the training data selected 'Norm', it was noticed in the boxplots that when other selections were chosen for example, 'PosN' a nd 'PosA', it can have a positive effect in increasing the median pricing of the house.
Kitchen Qual	While most of the kitchens fall under 'Gd' and 'TA', the kitchens under 'Ex' can shift the sales prices much higher. The 25th percentile of the 'Ex' kitchens is already higher than the 75th percentile of 'Gd' kitchens.
Fireplace Qu	Through the boxplots, it is observed that different selections under fireplace quality can have a positive impact on the sale price of housing even though more of the houses in the training data do not have fireplaces.
Garage Type	Built in garages generally have higher sale prices as compared to other selections according to the box plots.
Garage Qual	Even though most of the counts for garage quality fall under 'typical/ average', the quality of the garage strongly affects the sales prices as seen in the box plots.
Pool QC	The boxplots show that the pool quality has an impact on the sale prices. Even though not all houses have pools, the median prices of the houses are much higher for pools with good or excellent ratings.

Which numeric features will be useful?

Useful Feature

Useful Feature (With Outlier)

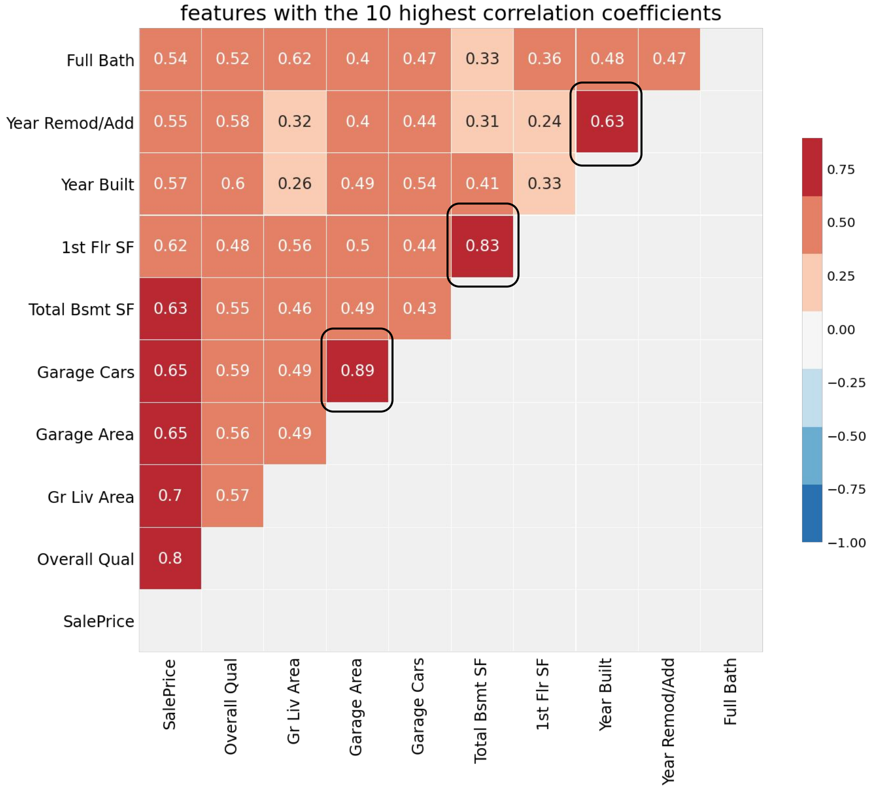


Not Useful Feature

Key properties of useful numeric features:

1. Highly linear
2. No outliers (Linear models are sensitive to outliers)

EDA - Multicollinearity

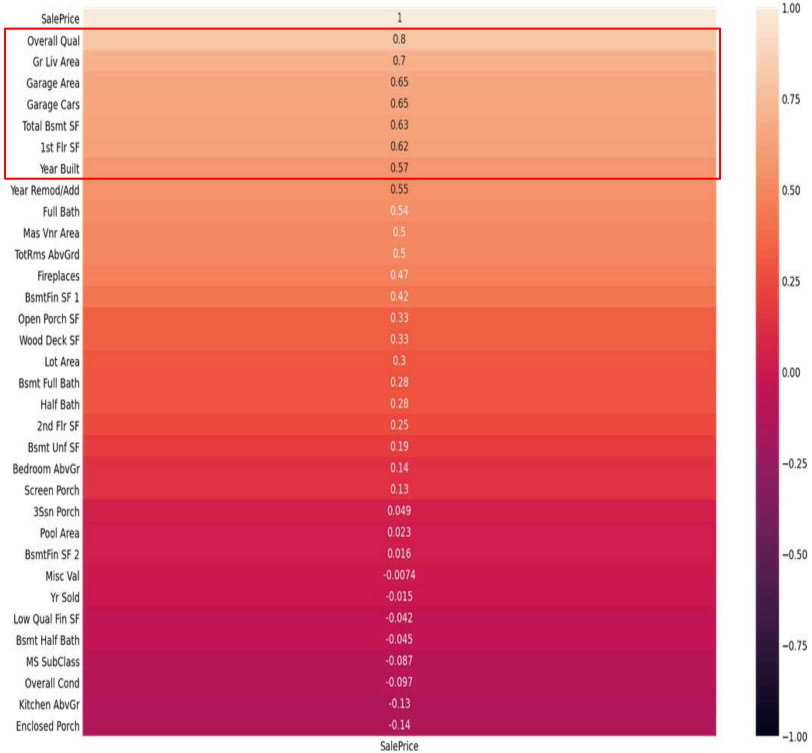


Occurrence of **high intercorrelations** is observed among two or more independent variables

1. Year remod/Add & Year Built
2. 1st Flr SF & Total Bsmt SF
3. Garage Cars & Garage Area

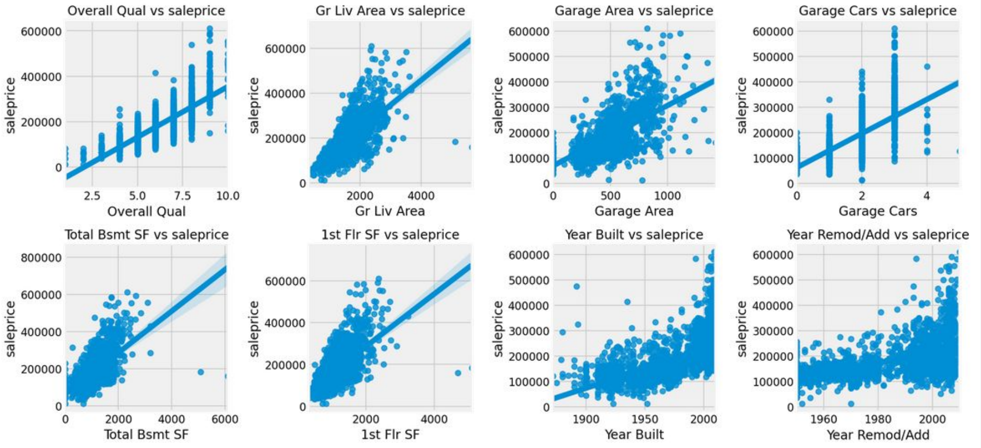
Multicollinearity **generates high variance of the estimated coefficients** and hence, the coefficient estimates corresponding to those interrelated explanatory variables **will not be accurate** in giving us the actual picture. They can become very sensitive to small changes in the model.

EDA - Feature Engineering (Initial)

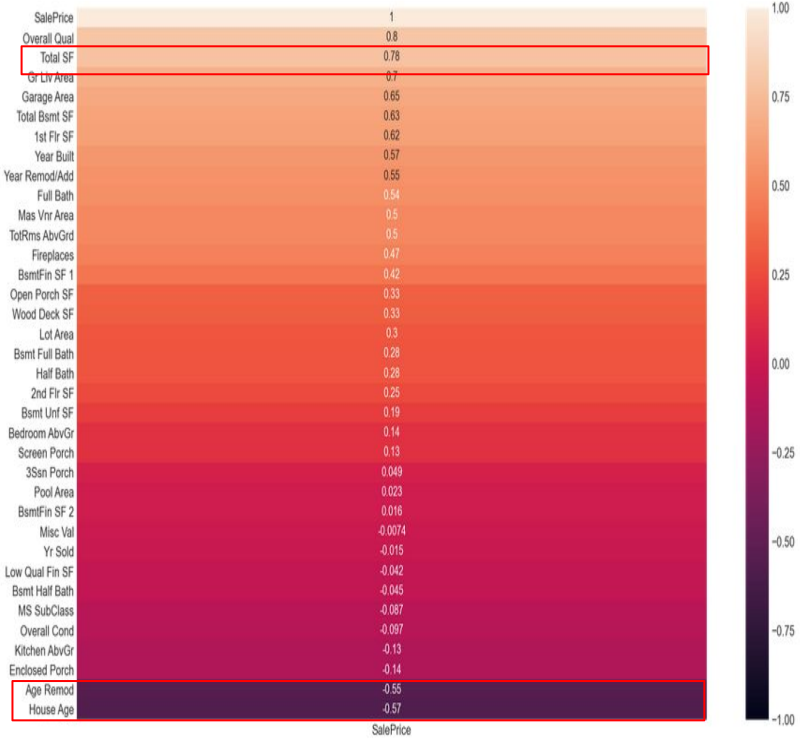


Top 5 Features (correlation with SalePrice) :

1. Overall Qual
2. Gr Liv Area
3. Garage Area
4. Garage Cars
5. Total Bsmt SF



EDA - Feature Engineering



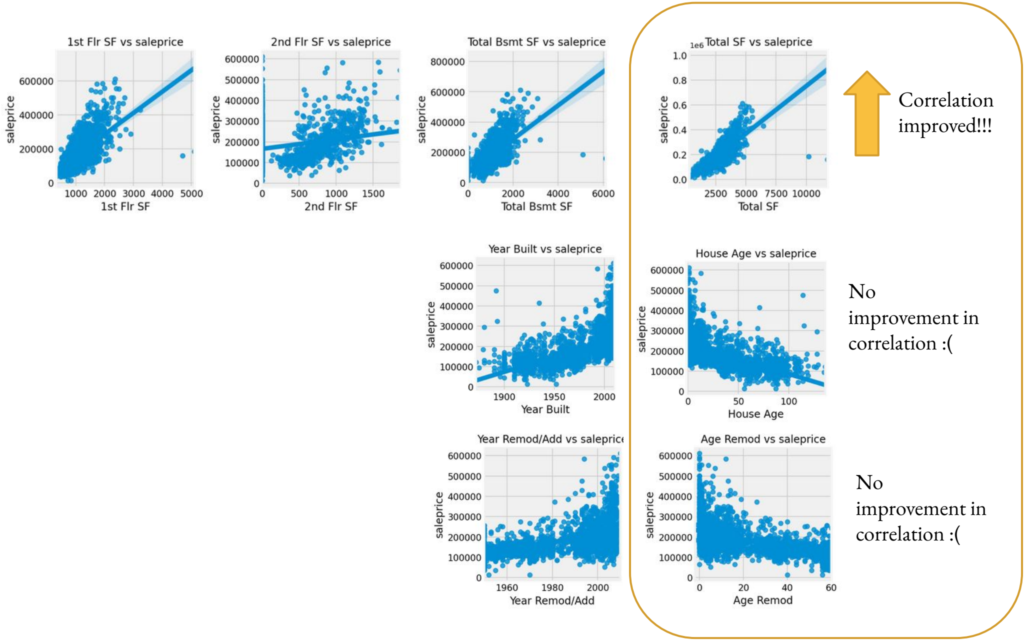
Transformed Columns with multicollinearity

Total SF = Total Basement SF + 1st Floor SF + 2nd Floor SF

House Age = Year Sold - Year Built

Age Remod = Year Sold - Year Remodelled/Add

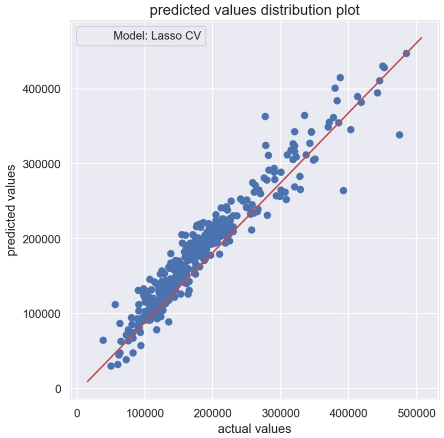
Before **After**



Initial Modelling

1. Train-test split (Test size = 0.2)
2. Features with low variance were removed for dimensionality reduction as a feature with low variance cannot explain much of the variance in Sale Price

	Linear Regression	RidgeCV	LassoCV
RMSE	22,387.65	21,773.30	21,320.51
R^2	0.918	0.923	0.926



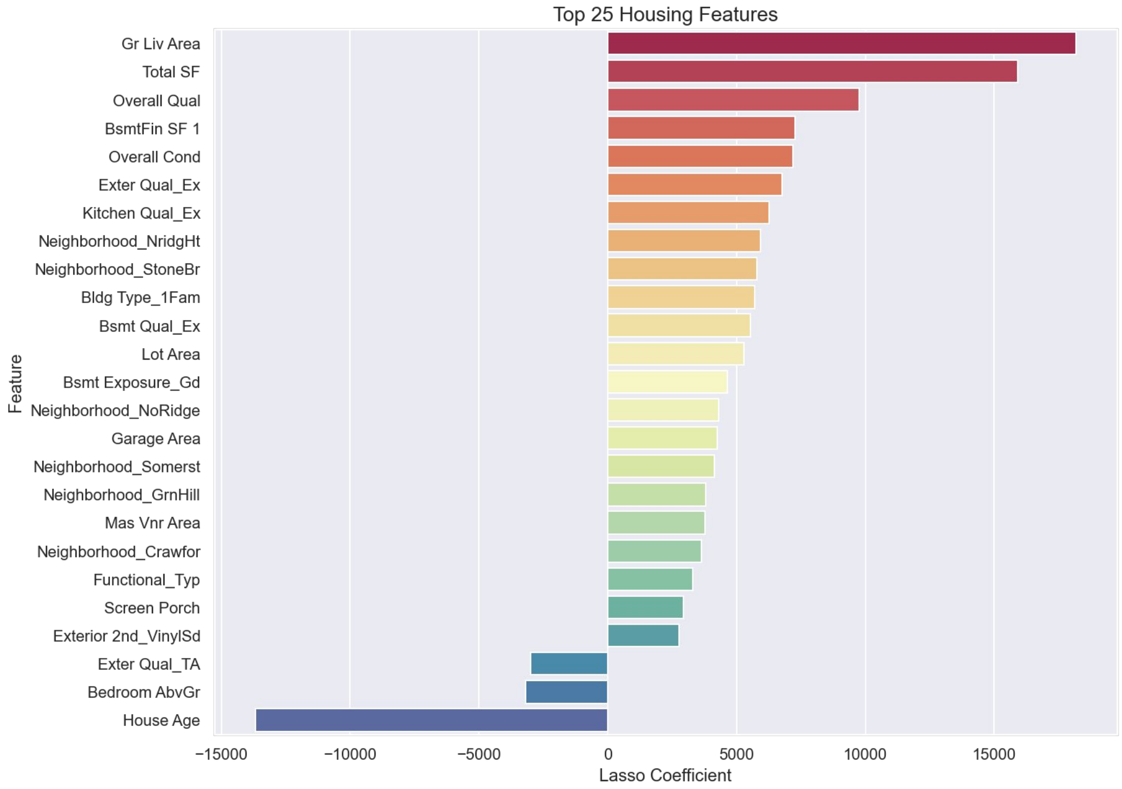
Conclusion

The Lasso Regression Model was the best model to test my training data because it was able to manage well unknown data according to the $R2$ score. The scatter plot of the predicted saleprice and actual sale price has a generally strong linear relationship except for two outliers.

Top 25 features - LassoCV Model

The top features fall under these key categories

1. Square feet
2. Quality /Condition
3. Building Type
4. Location
5. Exterior Features
6. House Amenities



Model Tuning & Optimization

Improved the model through the following methods

1. Hyperparameter tuning - Adjusted the alpha to the optimal alpha and adjusted max_iter
2. Top 25 features identified within the lasso model was used to rerun the model

	Initial LassoCV	LassoCV Hyperparameter tuning	LassoCV Top 25 features
RMSE	21,320.51	21317.64	21554.33
R^2	0.9257389	0.9257590	0.9241011

Recommendations

Sellers

Identify key features to upgrade to obtain a higher selling price.

Government

Monitor market price for properties to identify specific trends in home ownership impacted by sales price.

Buyers

Sellers to review the sales price and features if present to offer a fair value.

Bank

Provide a fair value basis on specific features of the home against market value for borrowers

Recommendations

Developers

Optimize new development timing, identify specific features which may allow price segmentation and buyer profile to maximize value

Investors

Optimize their holdings and regularly assess conditions that lead them to divest or capture value

Property Agents

Allows greater visibility in terms of price and features and allow them to manage customer expectations in terms of fair price and future trends.

Construction

Increase product portfolio which requires the manpower and materials required upgrading to specific features

References

Cock, Dean. "AmesHousing.txt" JSE
http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

How big data is transforming real estate analytics
https://www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate

Questions?

Conclusion

The lasso model was the best performing model in terms of RMSE, R2.

The key features that are most important in predicting sale prices s are affecting the house prices:

1) Square feet Area
2) Condition
3) Age
4) Location

Modelling

Evaluate Regression Model

- train-test split

- cross-validation / grid searching for hyperparameters

- strong exploratory data analysis to question correlatio
and relationship across predictive variables