Predicting house sale prices using regression models

Zhi Yuan, Yu Kiat, Kimberly, Hang Feng, Yong Gui

1

Problem Statement

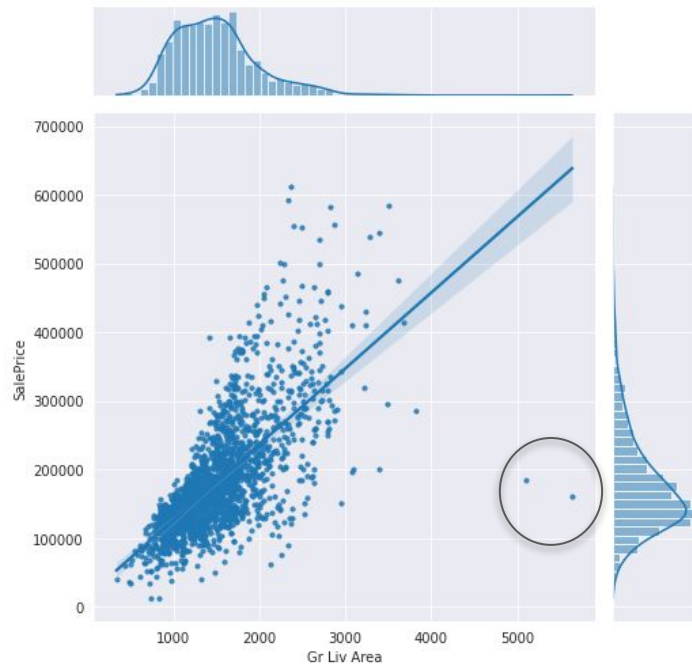Livable

Home Flipping Business

Living Area

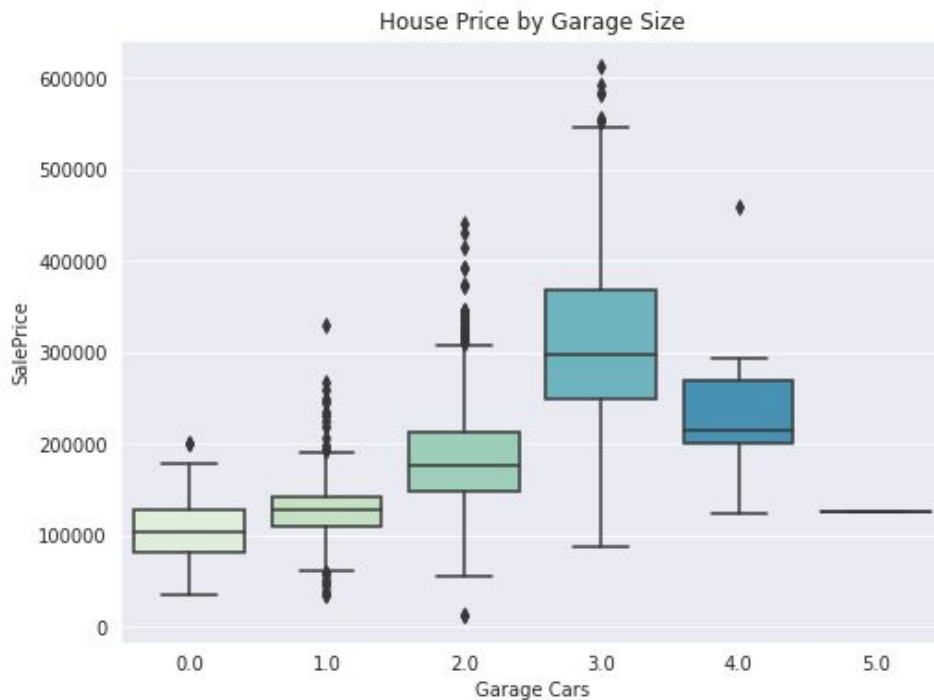Living area has a linear relationship with house price.

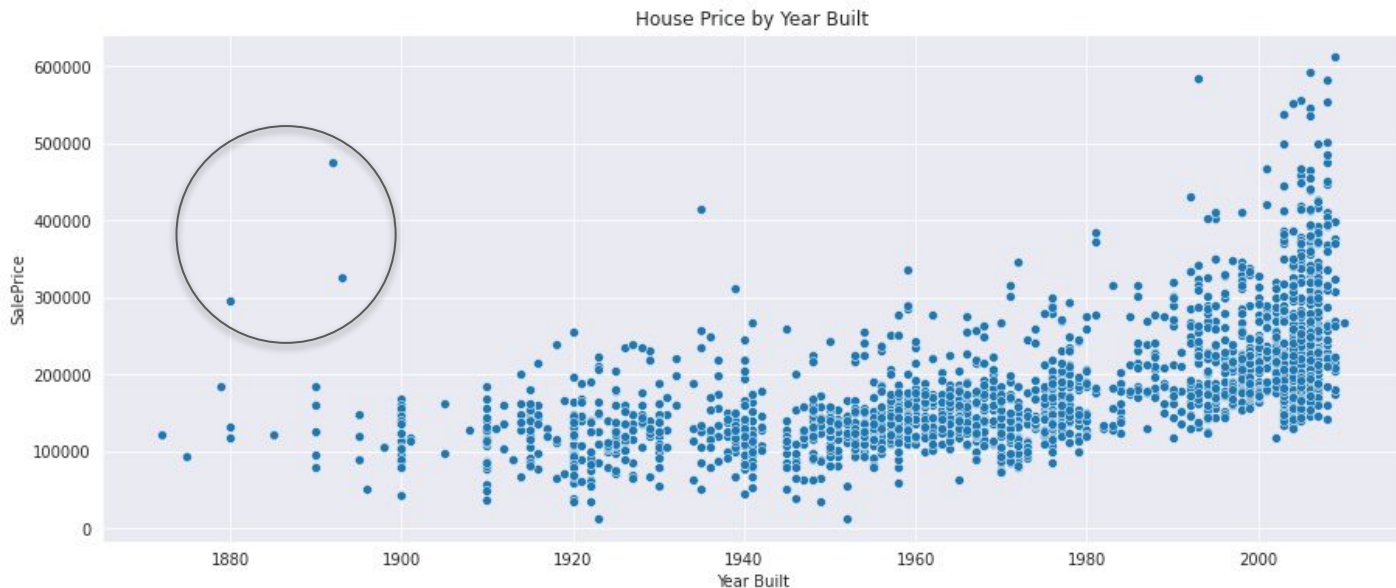# Garage Cars

Sale Price

Garage that can hold 4 cars or more < Garage that can hold 3 cars or less
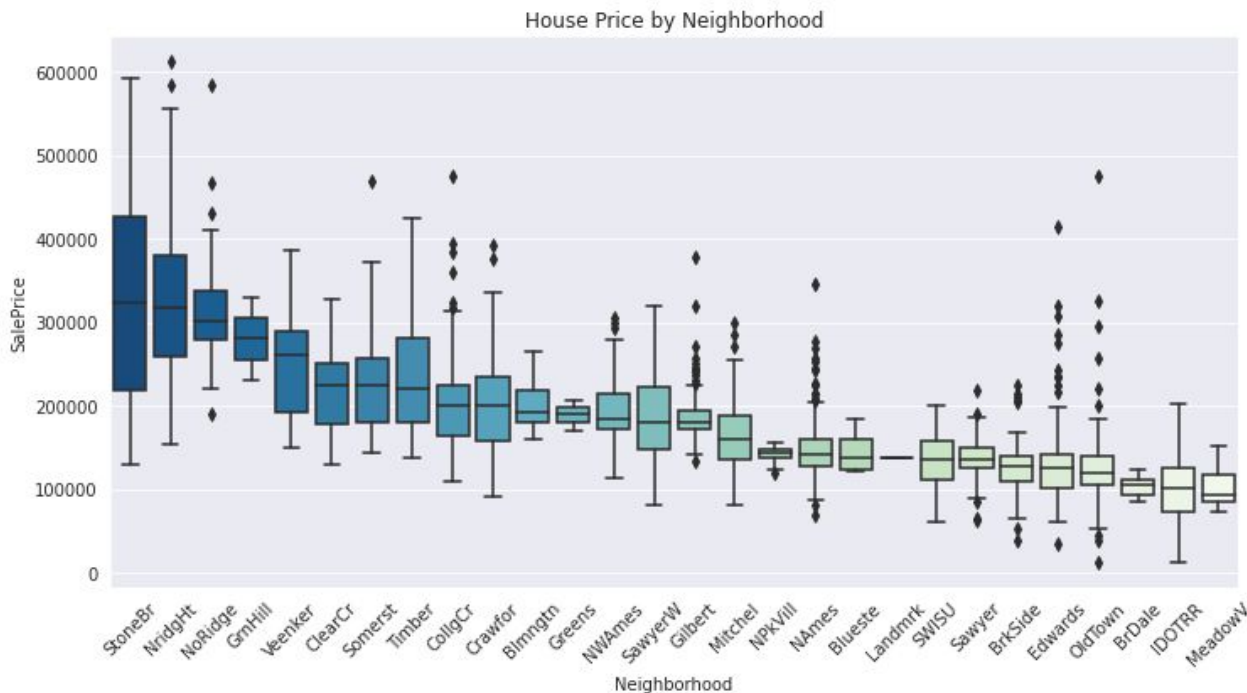

House Price by Garage Size

4

Year Built

The age of the house plays an important role in its price. There are several houses built before 1900 having a high price.

House Price by Year Built

Neighbourhood

House Price by Neighborhood

Most expensive neighbourhoods:

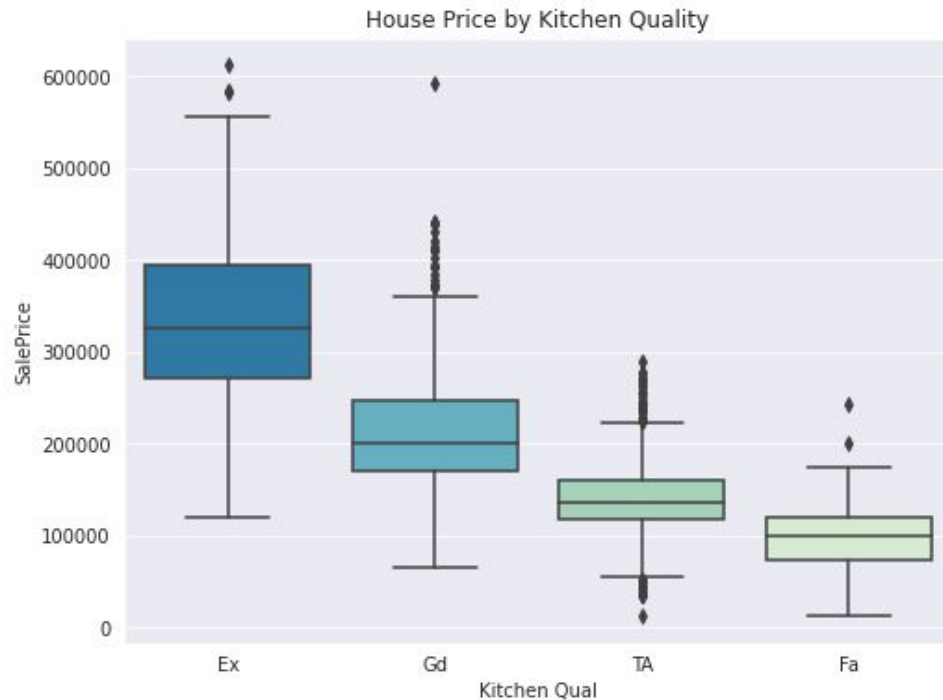1. Northridge Heights
2. Northridge
3. Stone Brook

3x difference in median sale price

Most affordable neighbourhoods:

1. Briardale
2. Iowa DOT and Rail Road
3. Meadow Village

Kitchen Quality

The average price difference between a house with a good kitchen and one with an excellent kitchen is about $120,000.

House Price by Kitchen Quality

House price distribution

Most of the house prices are between 100,000 and 200,000.

The distribution of SalePrice is right-skewed. The right tail would be treated like outliers when modelling
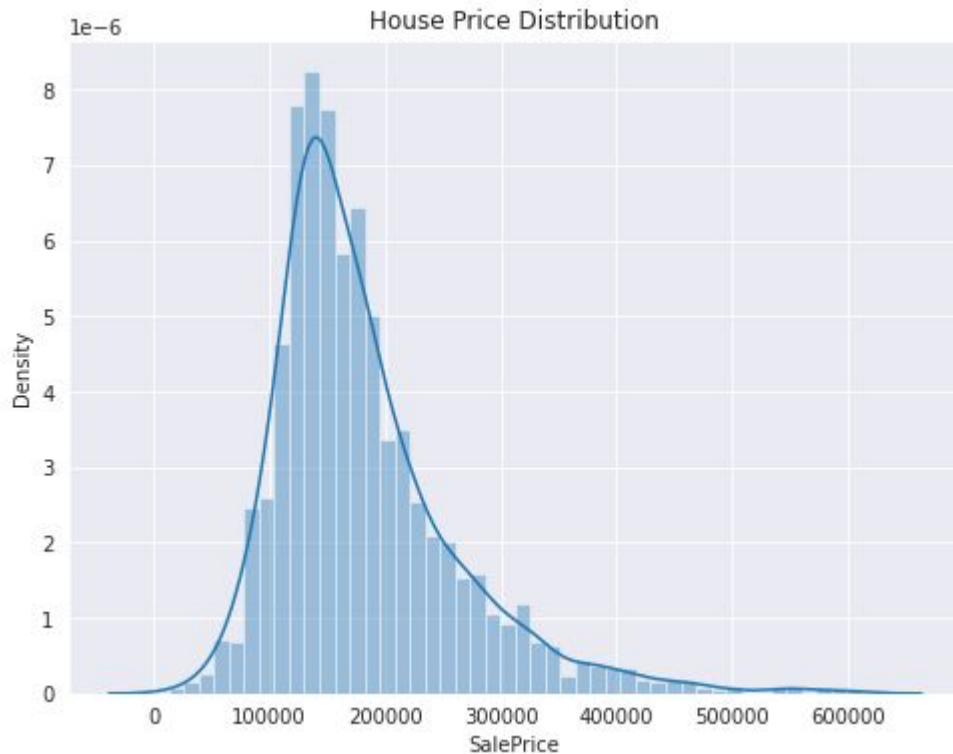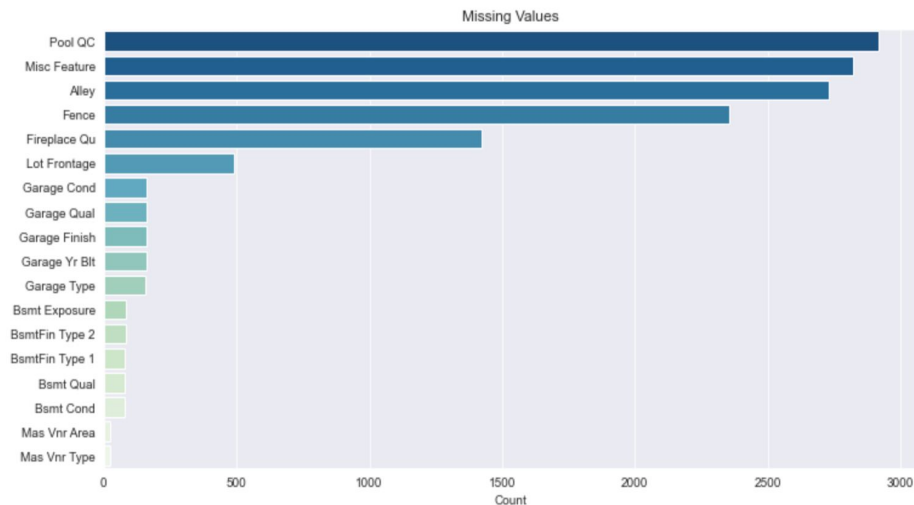
Dealing with Outliers

Regression models are sensitive to outliers, so they have to be removed.

The model may predict that houses with large above ground living area are cheaper if we do not remove the outliers

House Price vs. Living Area

9

Missing Values

Missing Values

Category 1:
Textual features where missing values mean a lack of the feature.
➡ Replaced with "None"

Category 2:
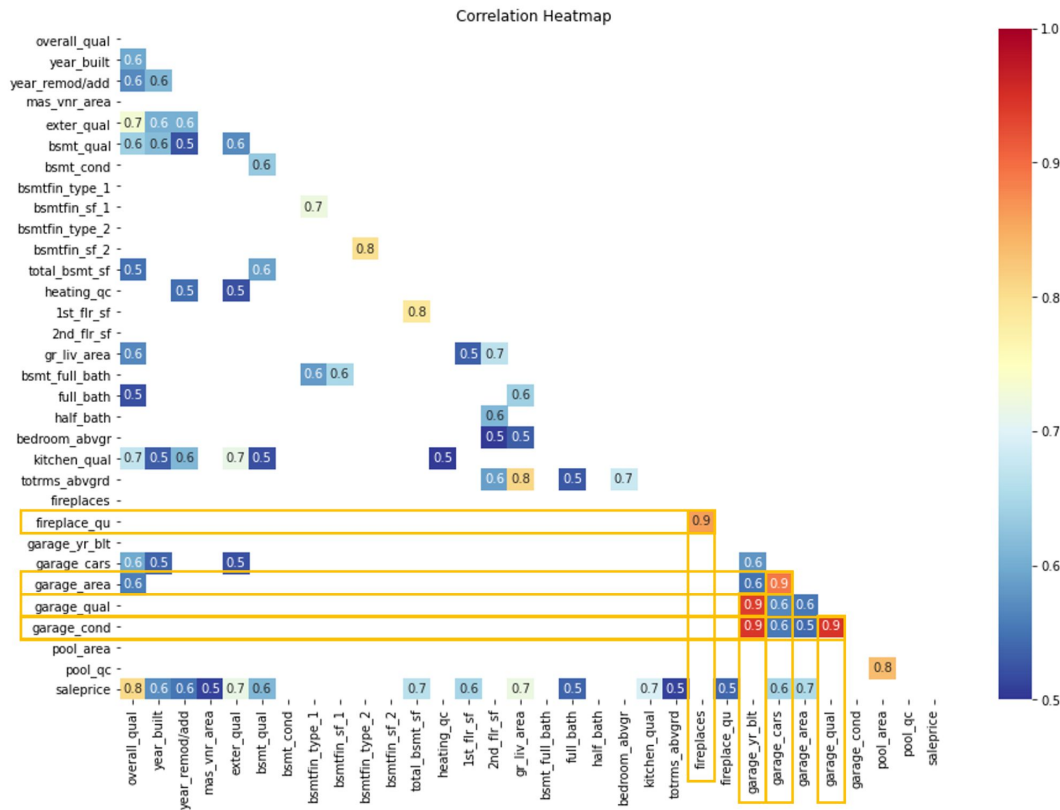Numerical features related to Category 1 items.
➡ Replaced with 0

Category 3:
Others
➡ Numerical: replaced with average
➡ Textual: replaced with most common value

Multicollinearity

Correlation Heatmap

There are several highly linearly related features.

Independent features should be *independent.*

Reduces precision of coefficient estimates.
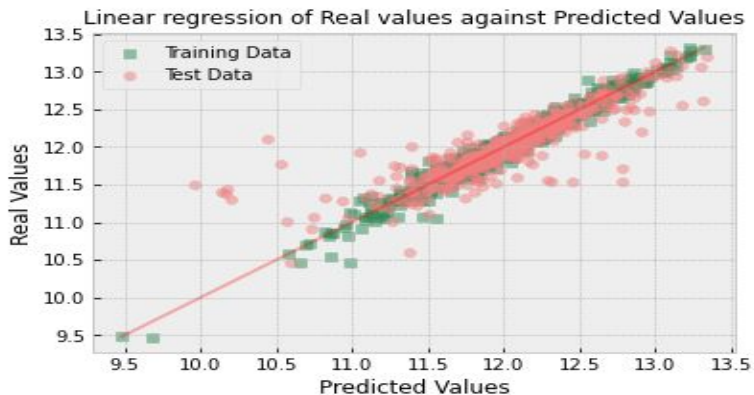
Feature Engineering
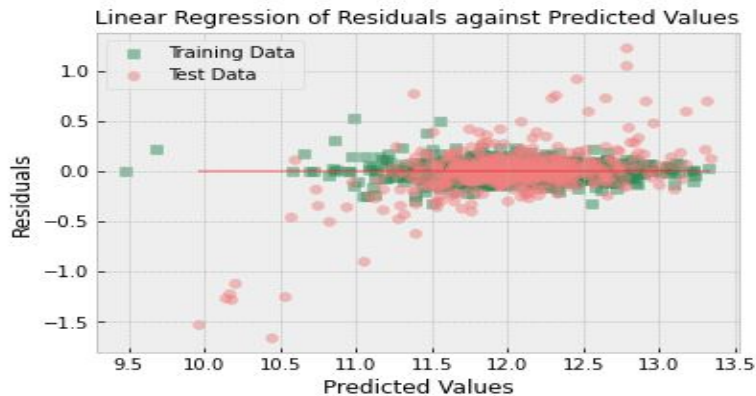
Combined features

E.g. Age, Size

New interaction features

E.g. Score = Quality x Condition or Quality x Quantity

The new features would produce multicollinearity but the modelling process would also help handle this issue.

Model Baseline - Linear Regression

Linear Regression of Residuals against Predicted Values

Linear regression of Real values against Predicted Values

R^2 Score:

- **Train Score (Lr):**
    - **0.65**

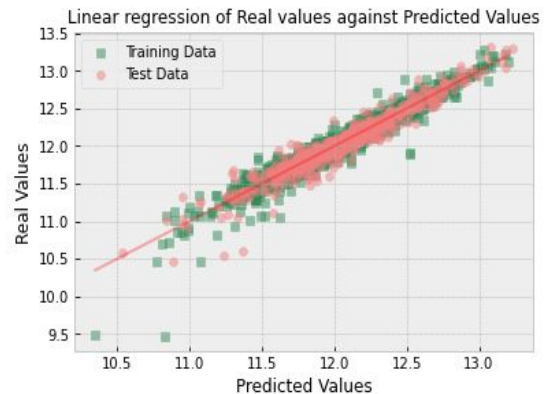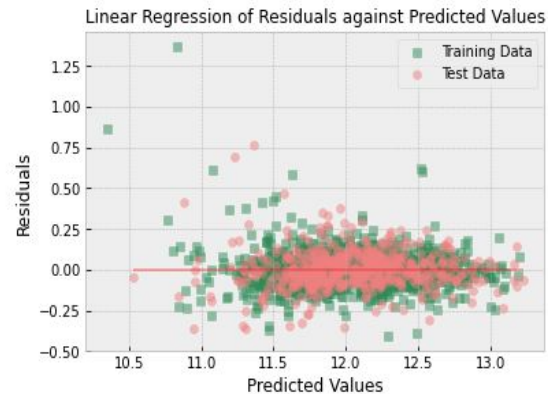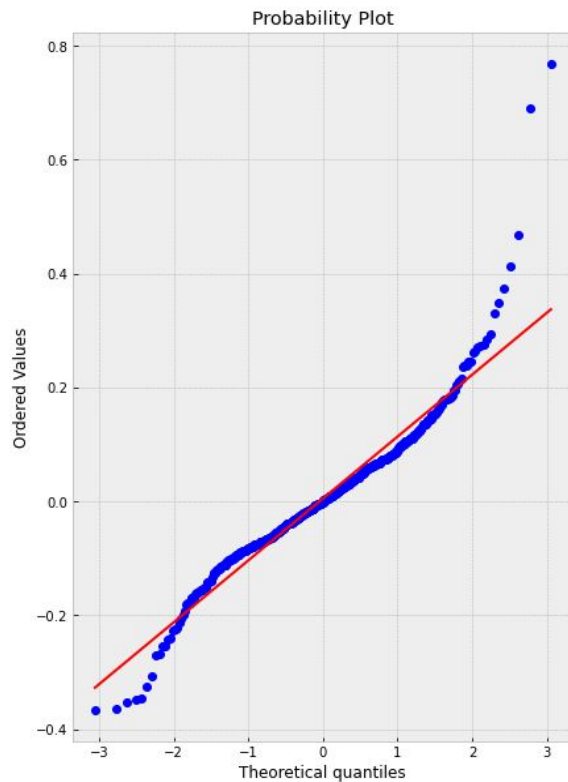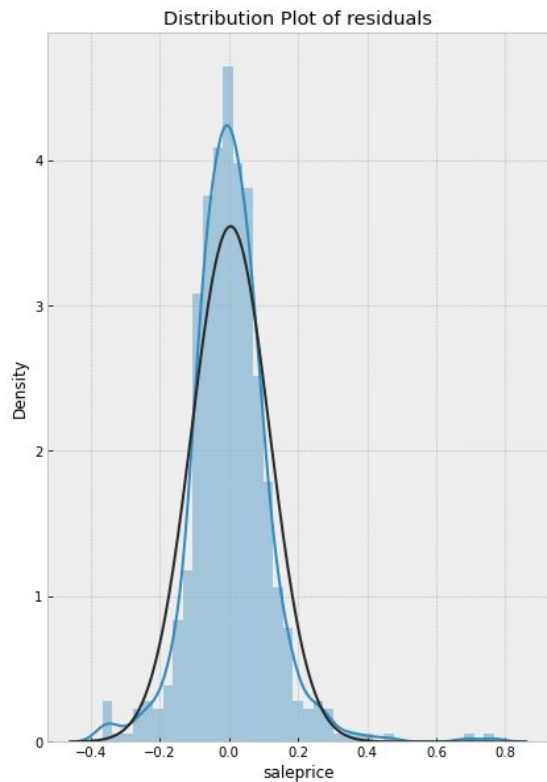- **Test Score (Lr):**
    - **-20**

RMSLE
- **Train Score (Lr):**
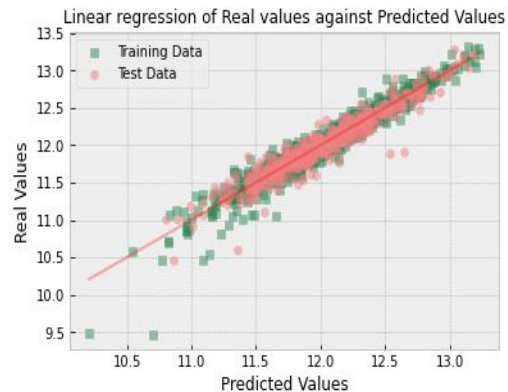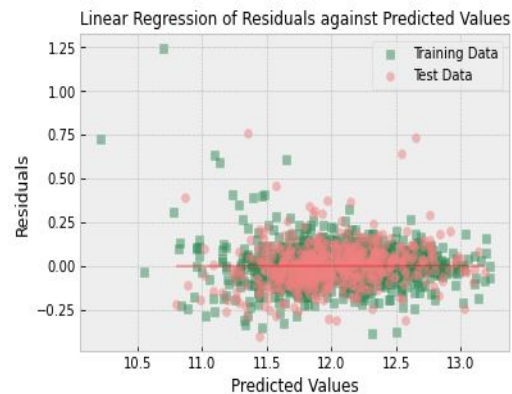   - **44,670.20**

- **Test Score (Lr):**
    - **$290,447.65**
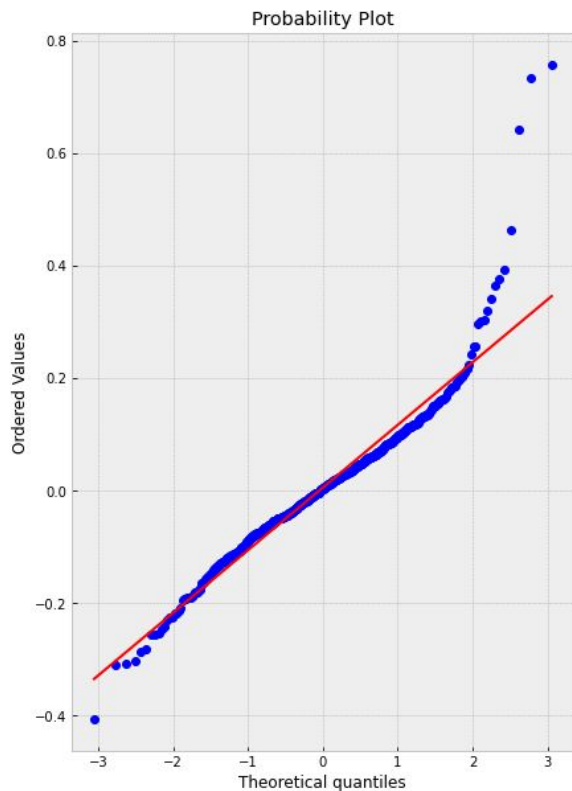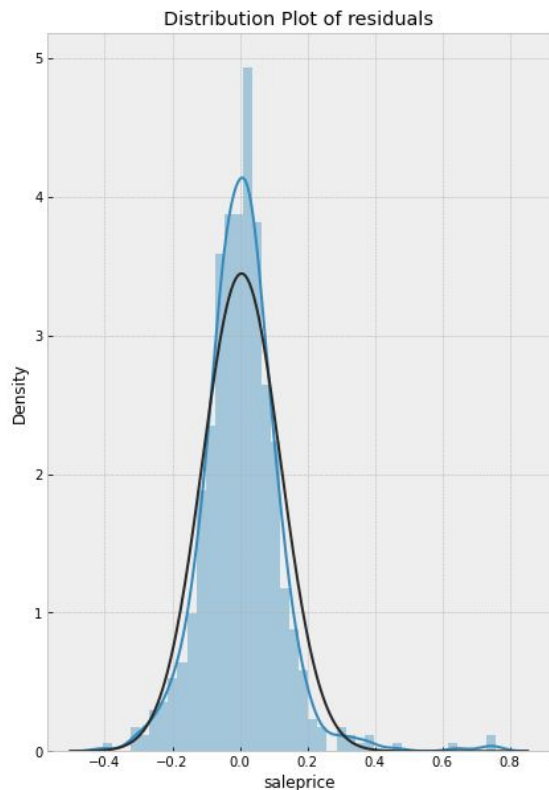
Modelling - Performance Summary

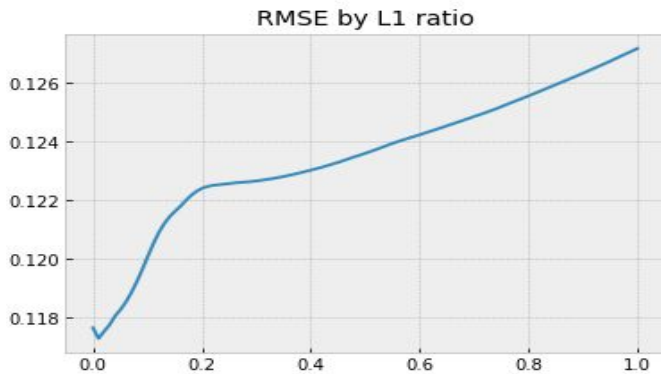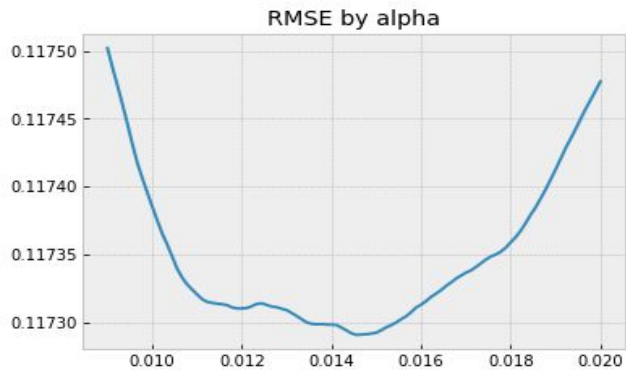| Model	Train R2	Test R2	Train RMSE	Test RMSE
Linear Regression	65.0%	-20.2%	44670.20	290447.65
Ridge Regression	92.0%	91.1%	20002.95	22804.35
Lasso Regression	91.7%	91.1%	21107.78	23870.26
ElasticNet	**92.0%**	**90.9%**	**19988.28**	**22764.66**

Model Performance - Distribution of residuals

Model Selection - ElasticNet

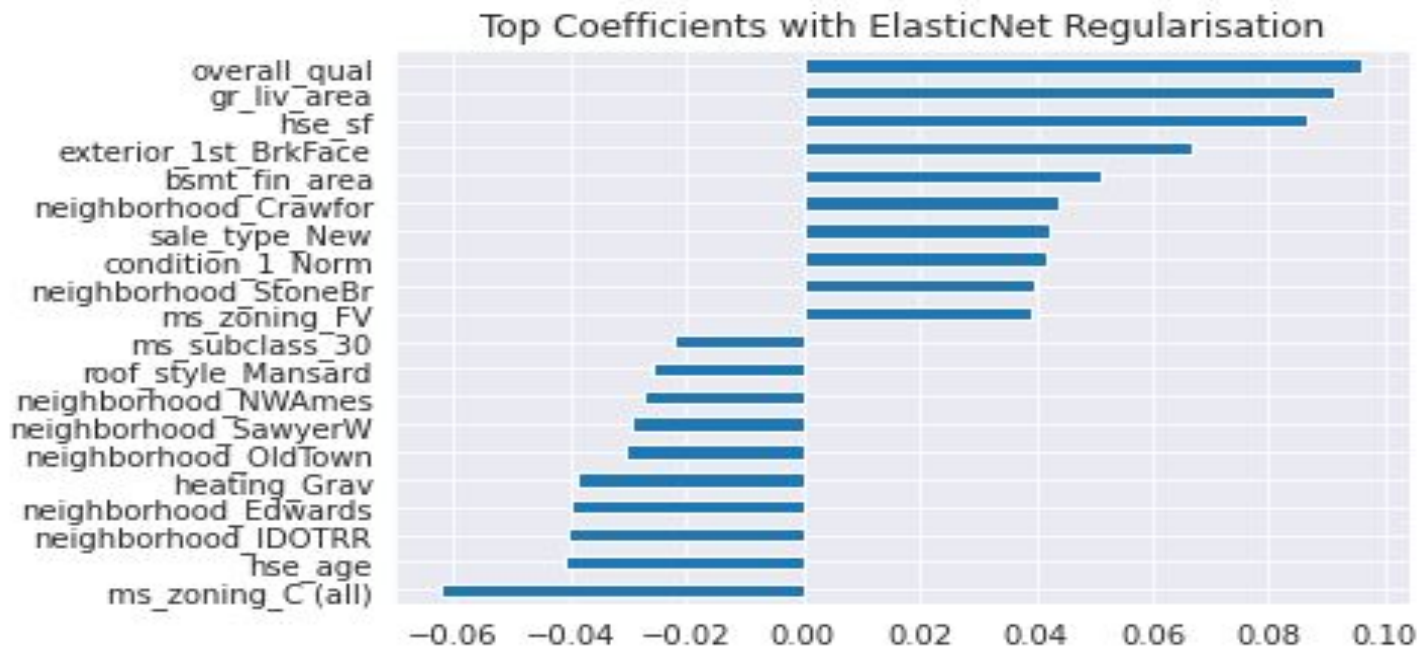# Model Tuning for RMSLE

- **Optimal Alpha:**
  - **0.015**

- **Optimal L1 ratio:**
  - **0.01**

Modelling - Features with Largest Coefficients

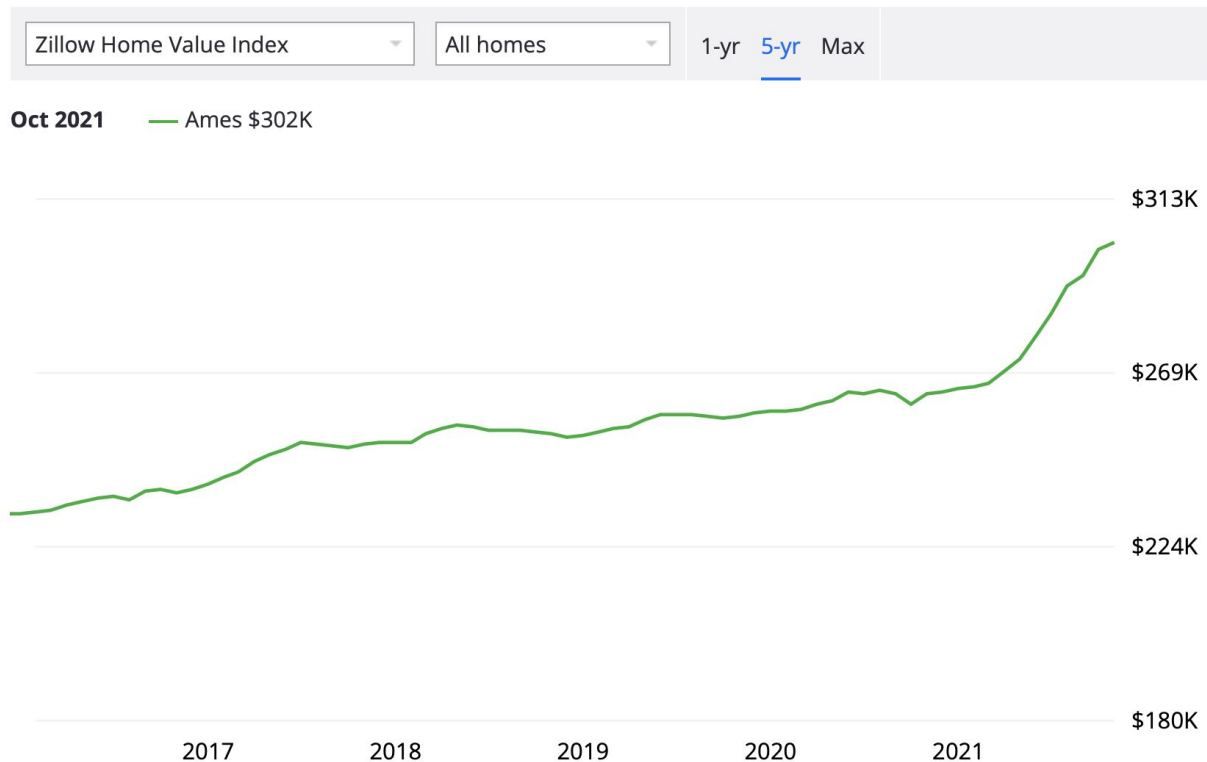Top Coefficients with ElasticNet Regularisation

18

Recommendation

- Avoid certain neighbourhoods
- Focus on aesthetic features that carry a premium
- Brick exterior

Next steps:

- We would require a companion analysis on labour and material costs on a case by case basis to predict profitability of the renovations.

External factors

Conclusion

- ElasticNet model performed the best due to the normality of the distribution
- Recommend frequently recording housing specs and sales price

Q&A