PROJECT 2

Predicting house sale prices using regression models

Zhi Yuan

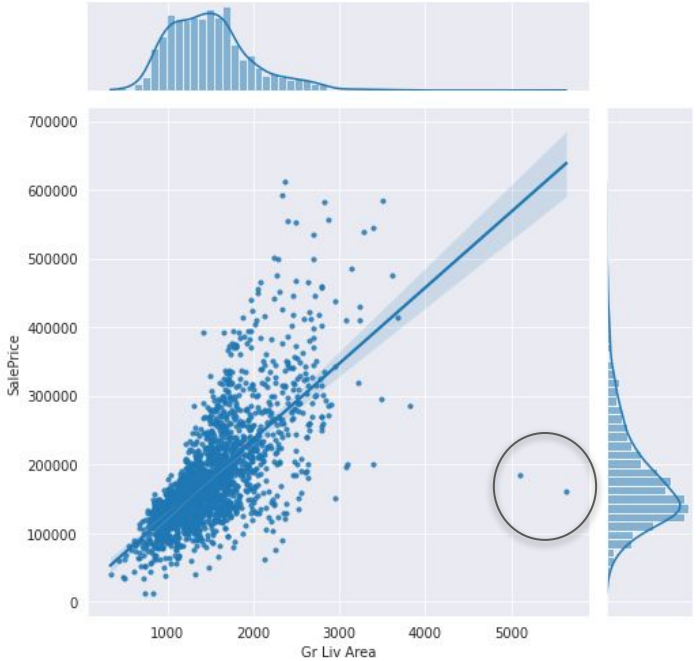
Problem Statement

Livable

Home Flipping Business

Living Area

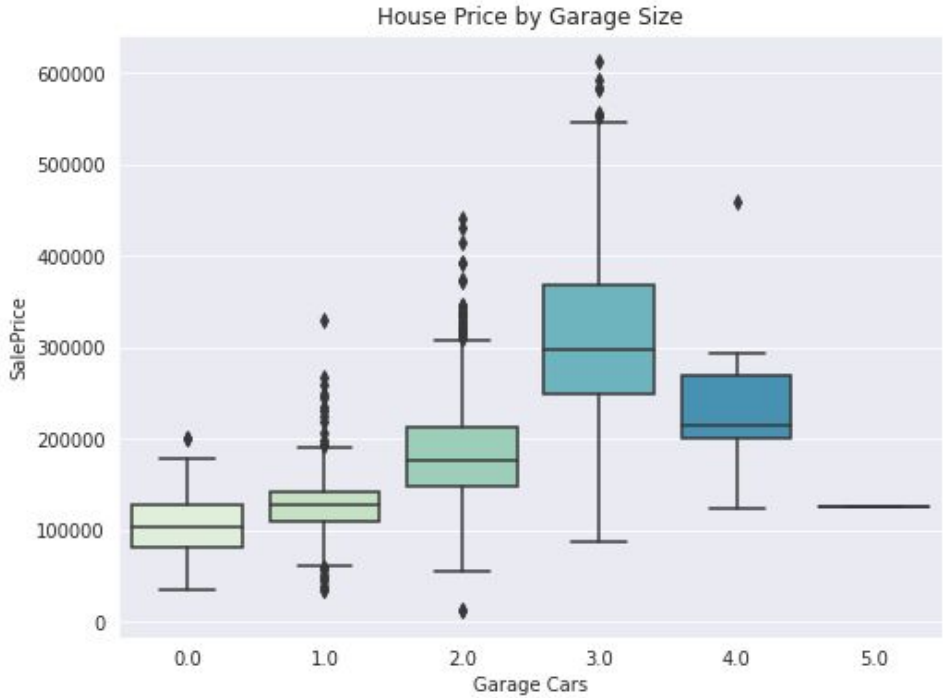
Living area has a linear relationship with house price.



Garage Cars

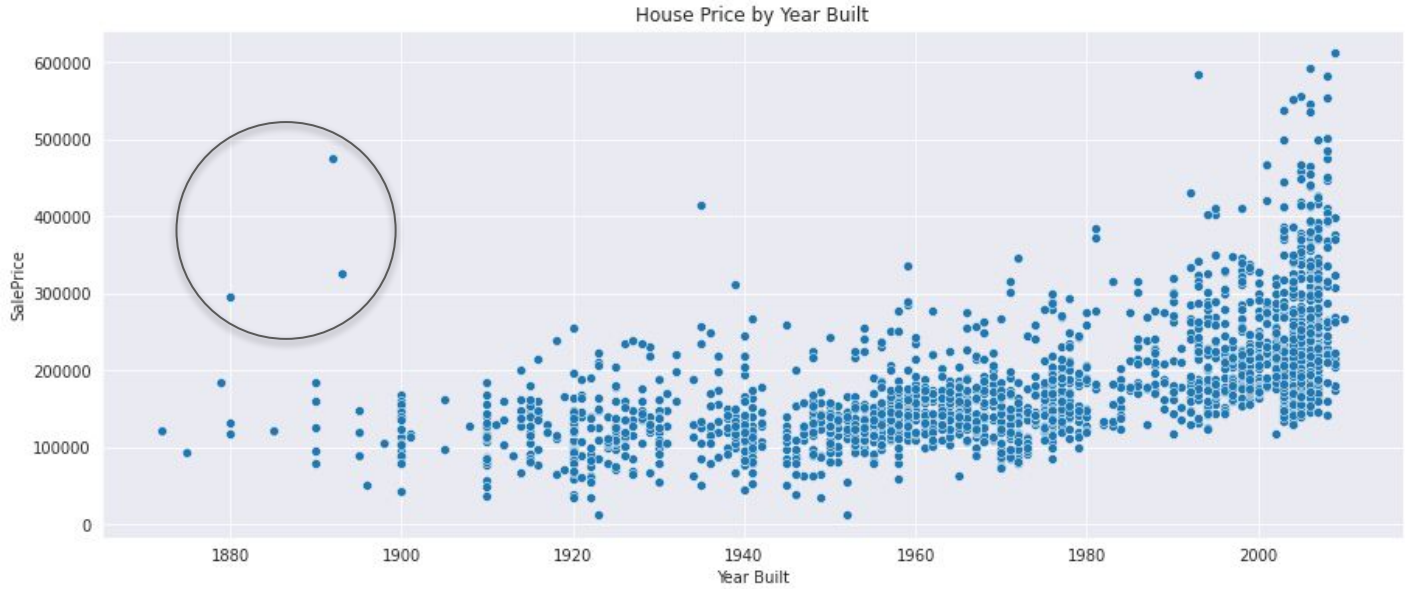
Sale Price

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

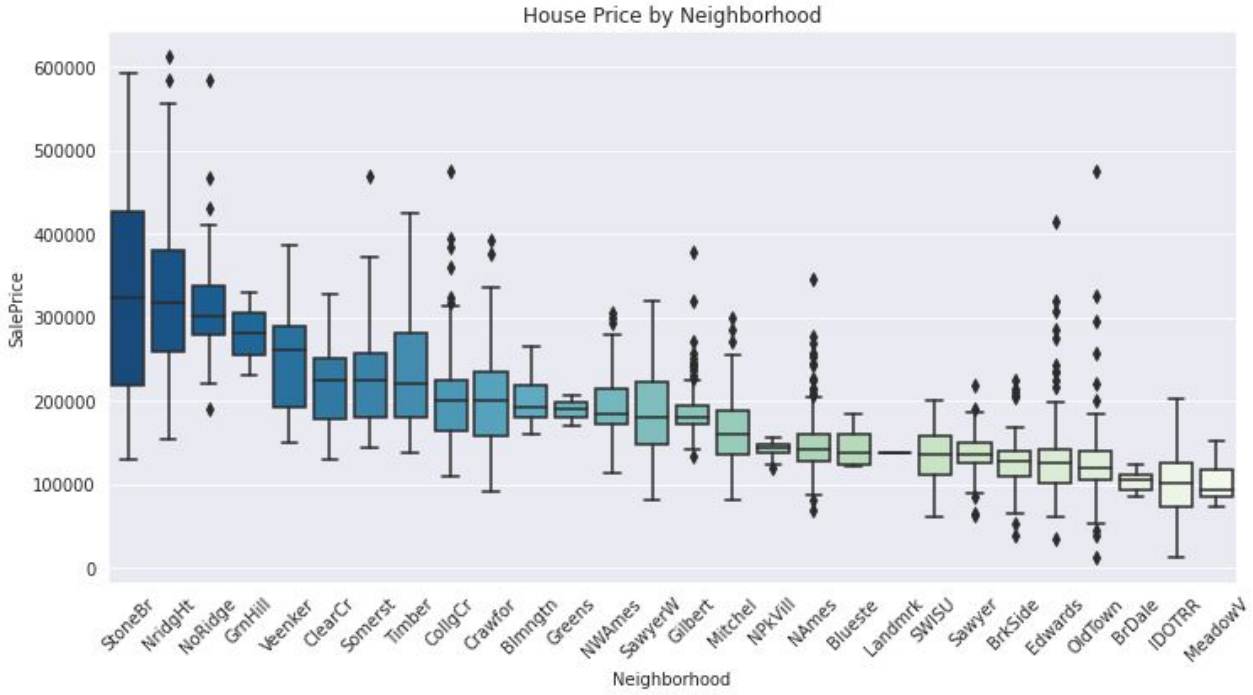


Year Built

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



Neighbourhood



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

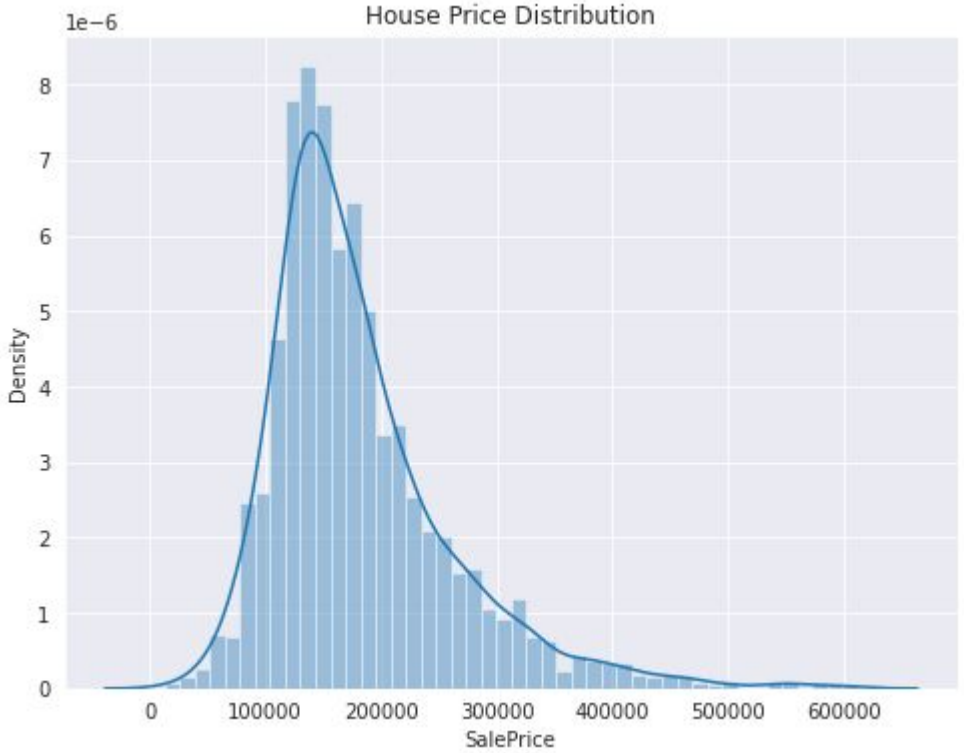
Kitchen Quality

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

House price 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

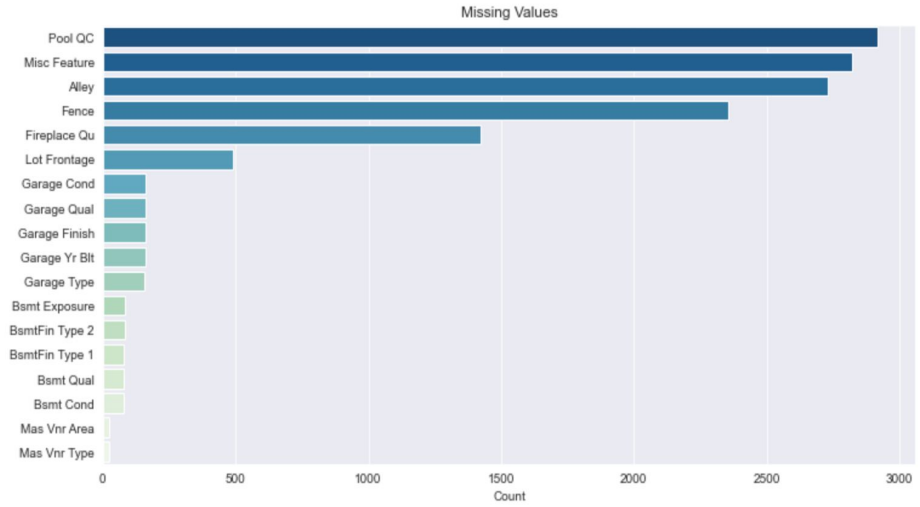


Dealing with Outliers

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

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.

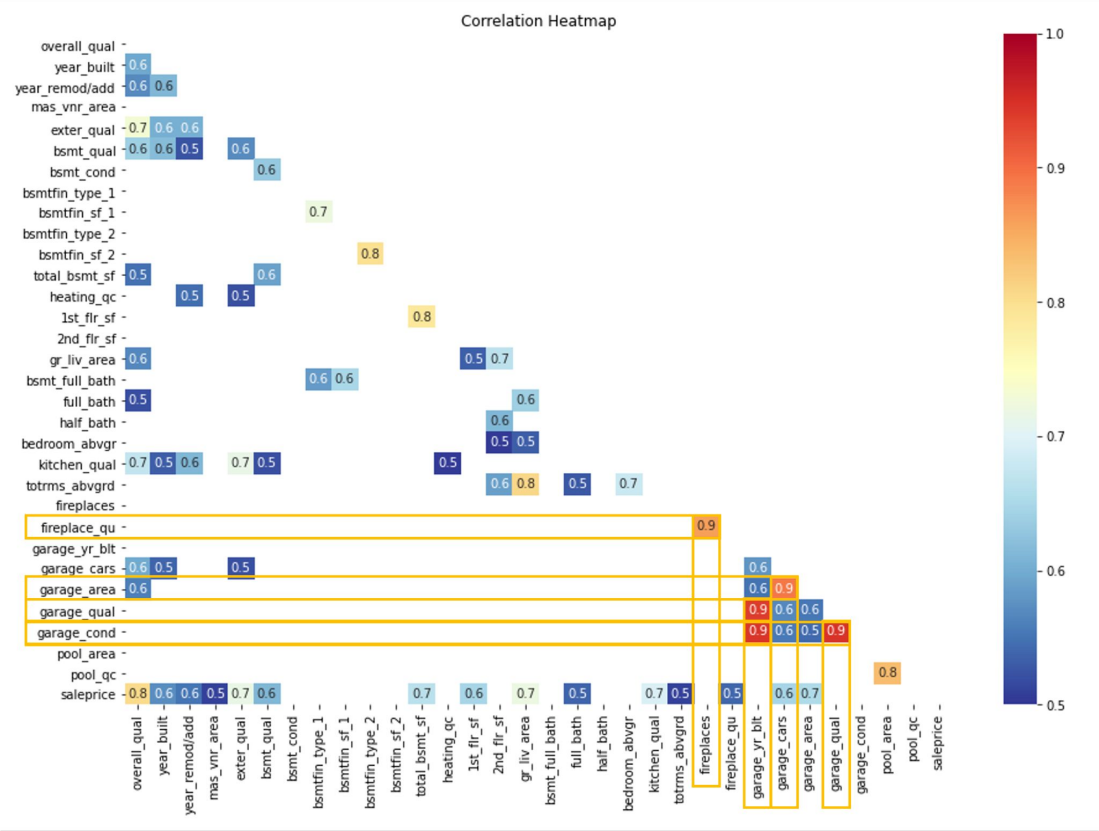
- Replaced with 0

Category 3:

Others

- Numerical: replaced with average
- Textual: replaced with most common value

Multicollinearity



There are several highly linearly related features.

Independent features should be *independent.*

Reduces precision of coefficient estimates.

Feature Engineering

Combined features

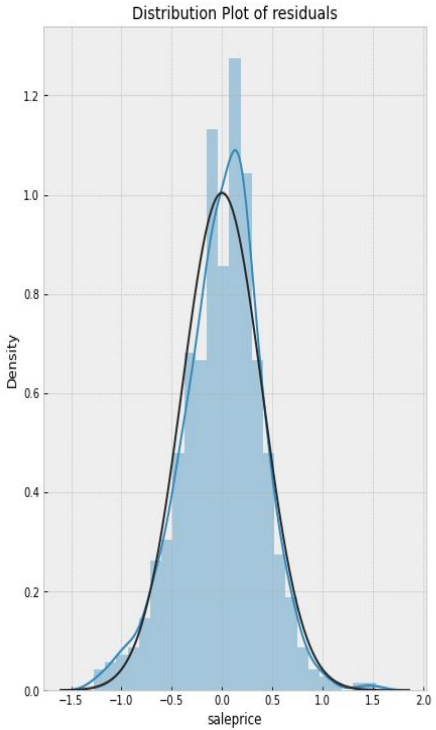
E.g. Age, Size

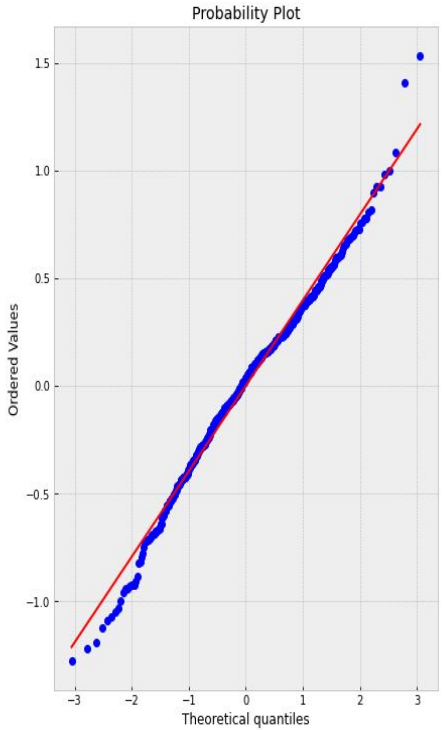
New interaction features

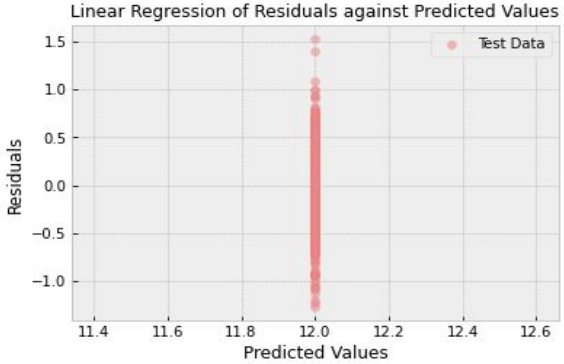
E.g. Score = Quality x Condition or Quality x Quantity

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

Baseline- mean

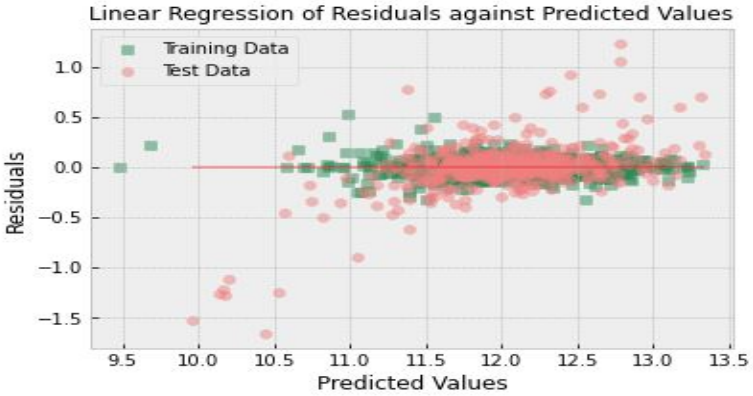


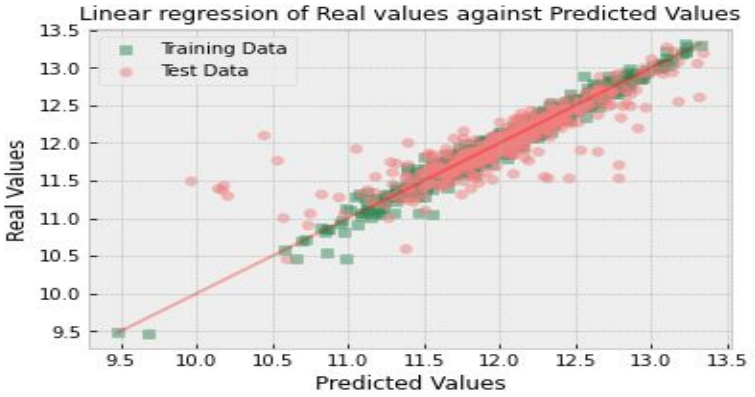




- **R^2 Score: 0**
- **RMSE: $78,000**

Model - Linear Regression





R^2 Score:

- **Train Score (Lr):**
  - **0.65**
- **Test Score (Lr):**
  - **-20**

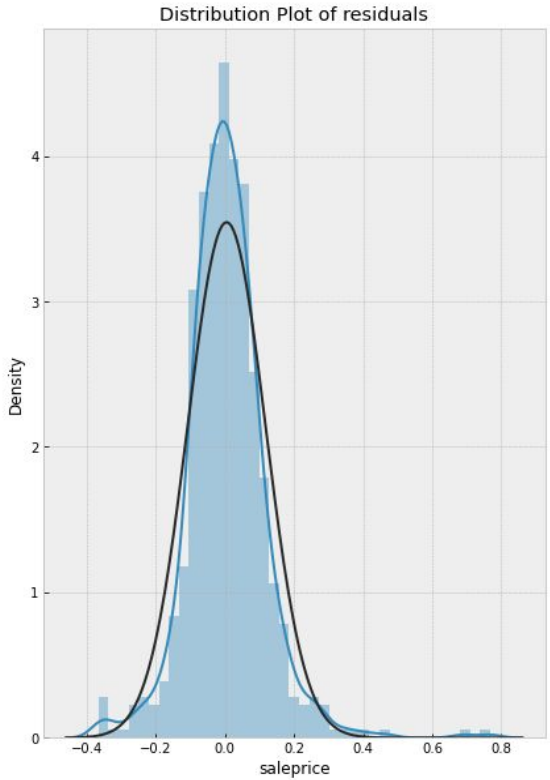
RMSE

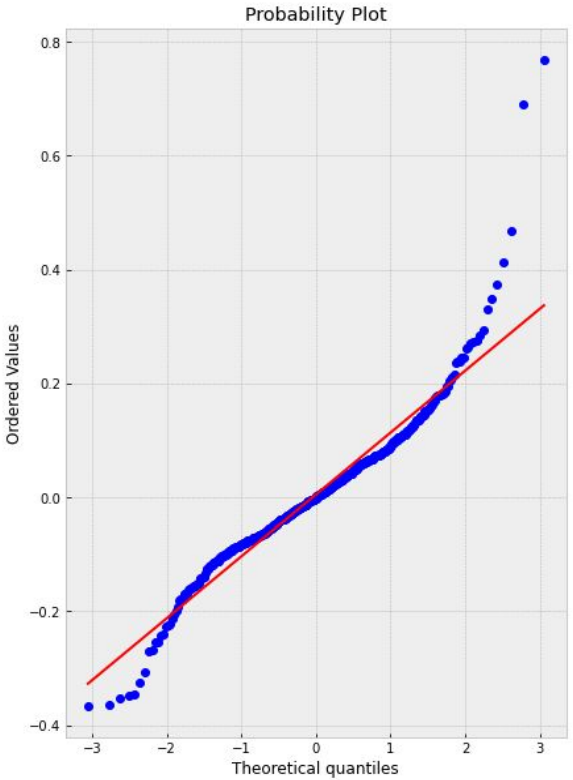
- **Train Score (Lr):**
  - **$44,670.20**
- **Test Score (Lr):**
  - **$290,447.65**

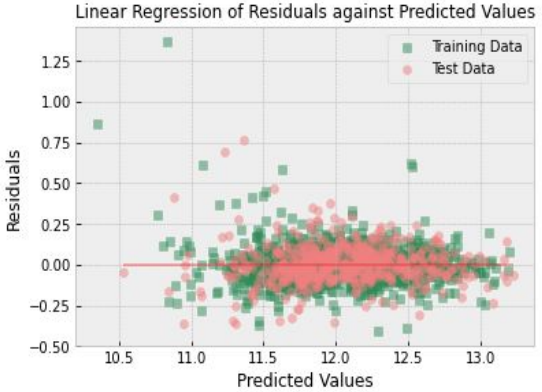
Modelling - Performance Summary

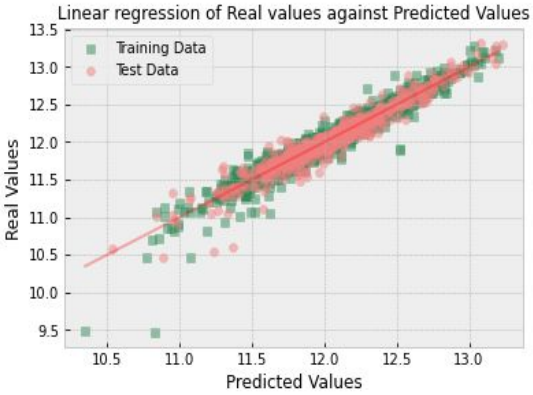
Model	Train R2	Test R2	Train RMSE	Test RMSE
Linear Regression	65.0%	-20.2%	44,670.20	290,447.65
Ridge Regression	92.0%	91.1%	20,002.95	22,804.35
Lasso Regression	91.7%	91.1%	21,107.78	23,870.26
ElasticNet	**92.0%**	**90.9%**	**19,988.28**	**22,764.66**

Model Performance - Distribution of residuals

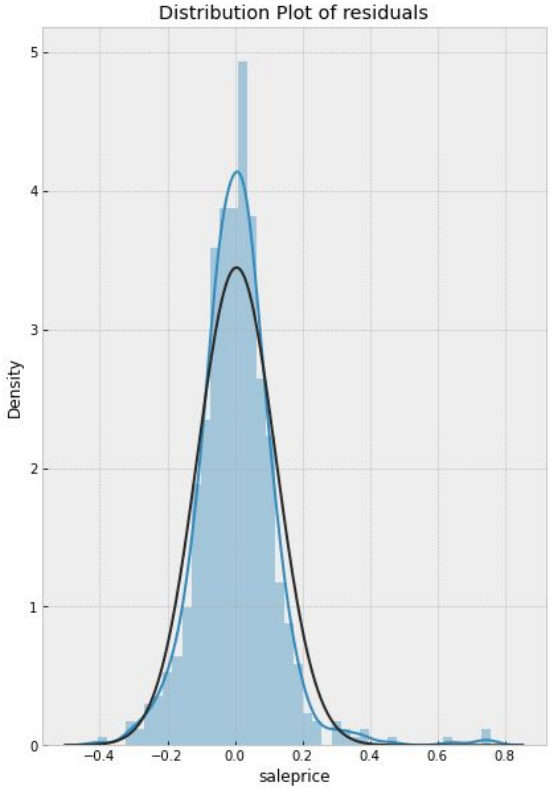


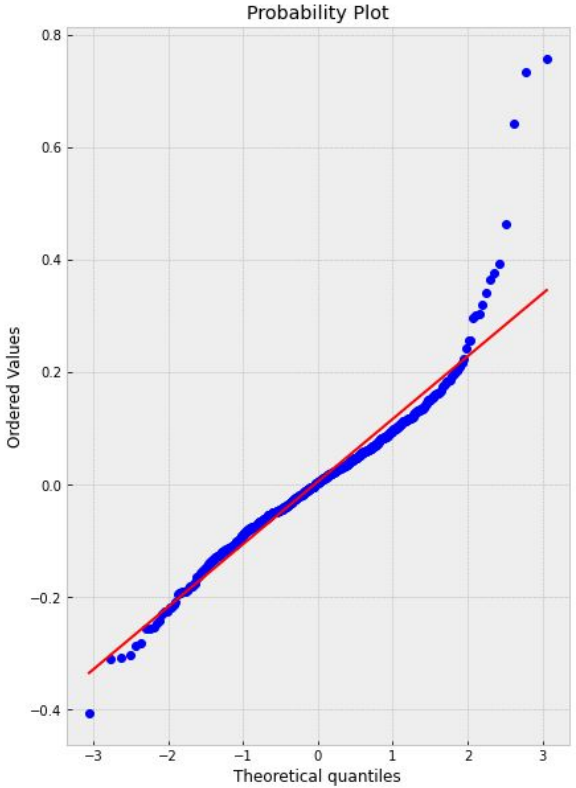


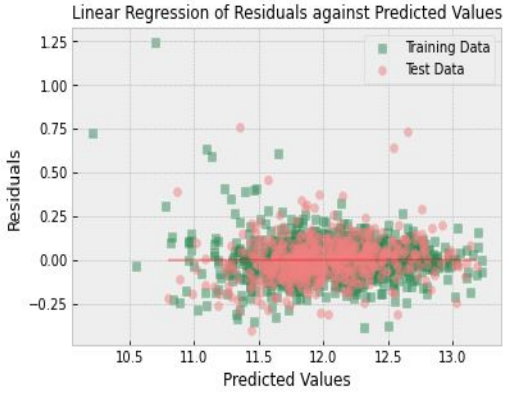


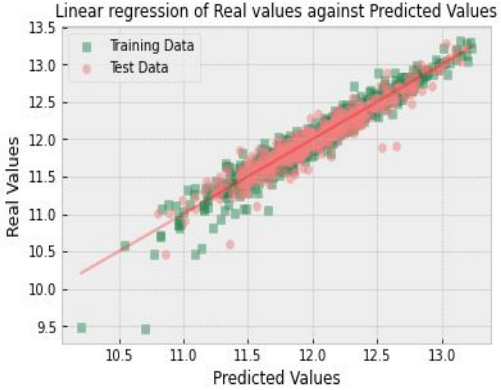


Model Selection - ElasticNet

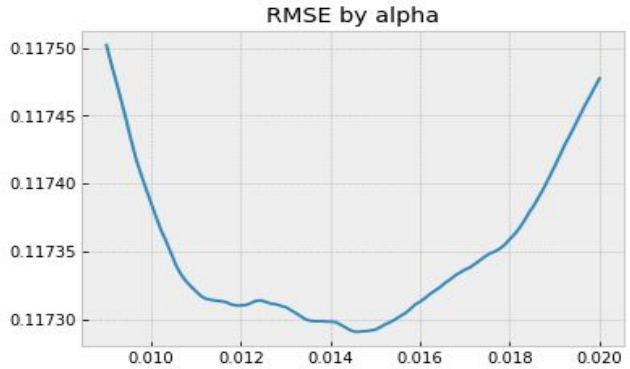


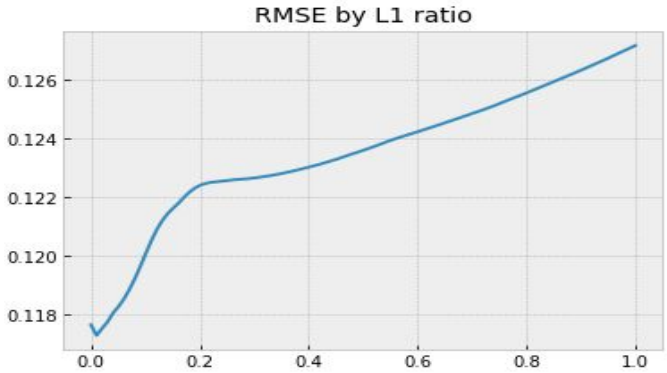






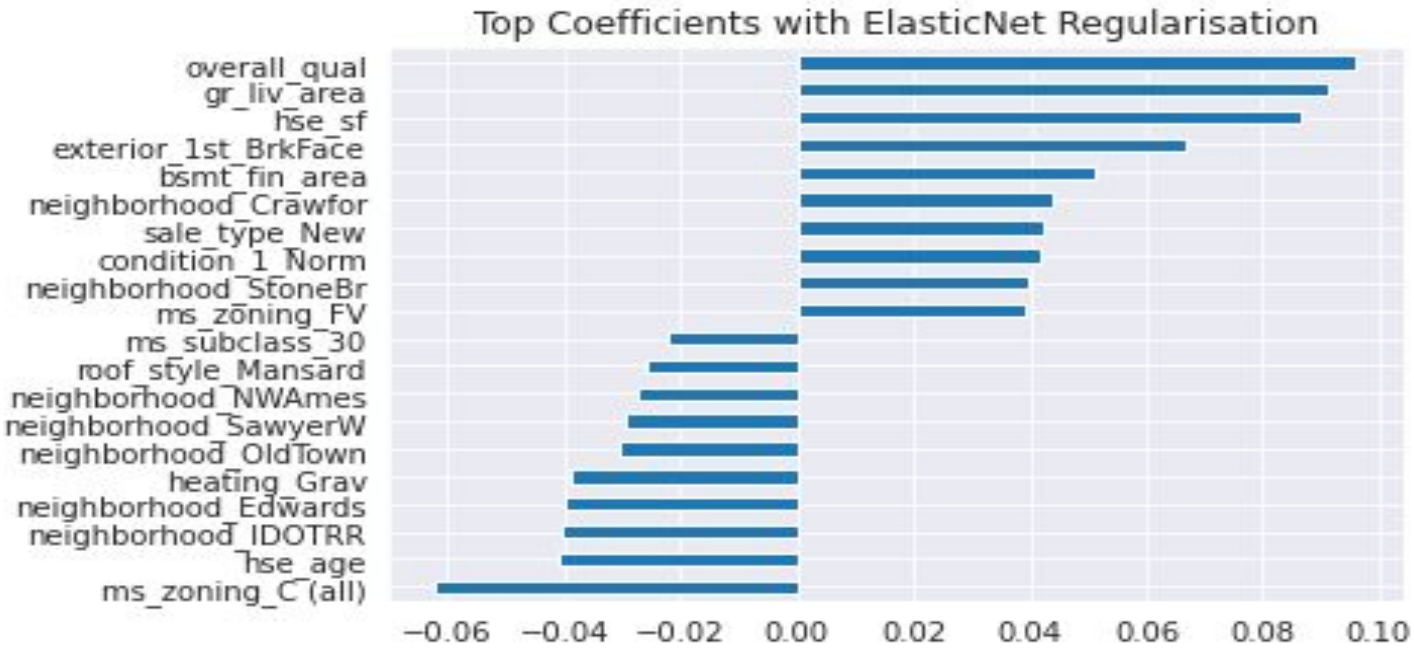
Model Tuning for RMSLE





- **Optimal Alpha:**
  - **0.015**
- **Optimal L1 ratio:**
  - **0.01**

Modelling - Features with Largest Coefficients



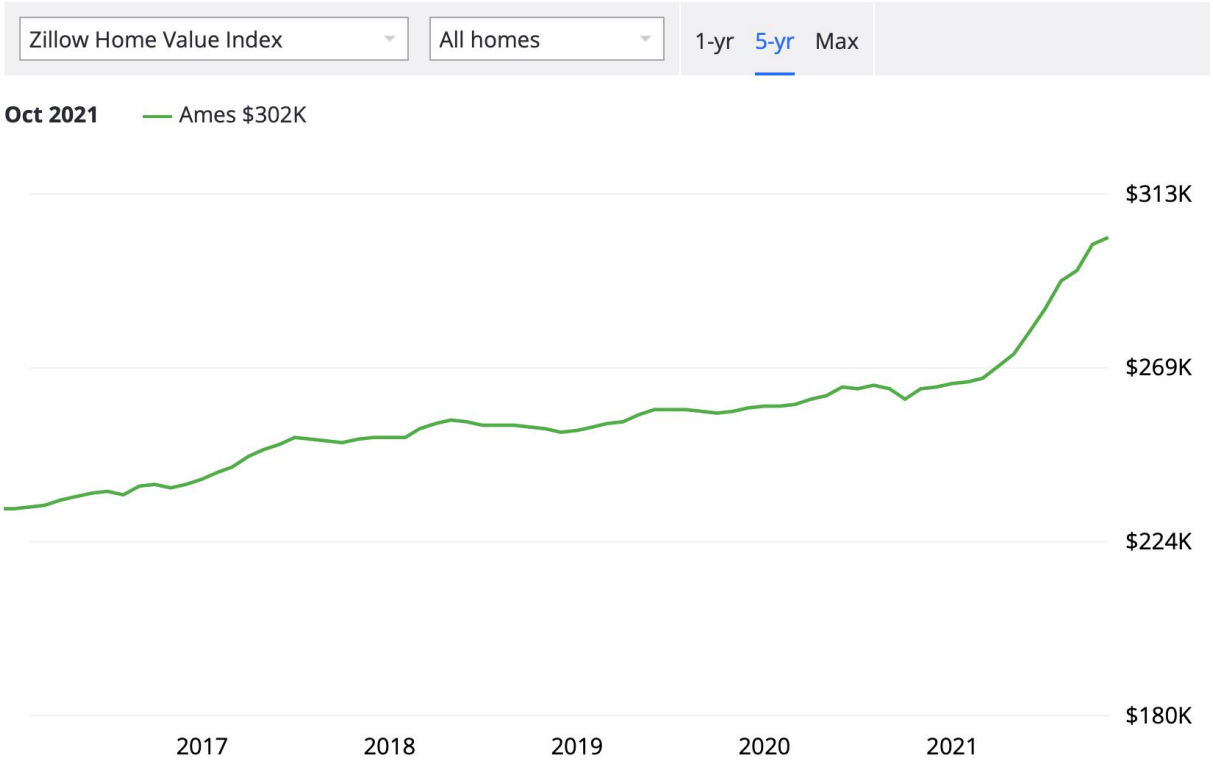
Recommendation

- 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

Annex - Missing Values

- Split columns with missing/null values into 2 groups:
  1. Categorical values where NA means no feature:
     - Pool QC, Misc Feature, Alley, Fence, Fireplace Qu, Garage Type, Garage Finish, Garage Qual, Garage Cond, Bsmt Qual, Bsmt Cond, Bsmt Exposure, BsmtFin Type 1, BsmtFin Type 2, Mas Vnr Type
       - Imputed with “None”
  2. Numerical values where NA means no feature:
     - Garage Area, Garage Cars, BsmtFin SF 1, BsmtFin SF 2, Bsmt Unf SF, Total Bsmt SF, Bsmt Full Bath, Bsmt Half Bath, Mas Vnr Area
       - Imputed with 0
  3. Others:
     - Lot Frontage
       - Imputed with the mean
     - Garage Yr Built
       - Imputed with Year Built
     - Functional, MS Zoning, Electrical, Kitchen Qual, Exterior 1, Exterior 2, Sale Type, Utilities
       - Imputed with the mode

Annex - Engineering New Features

1. Total square feet
   - Adding all the square feet of the floors
     - TotalBsmtSF + 1stFlrSF + 2ndFlrSF
2. Total bathrooms
   - Adding together the number of bathrooms
     - FullBath + BsmtFullBath + 0.5(HalfBath + BsmtHalfBath
3. Age of house
   - Number of years between built and sold
     - Yr Sold - Year Built
4. Remodeled or not?
   - Whether the house was remodelled
     - If Year Remod/Add = Year Built, 0 (i.e. False)
     - Else, 1 (i.e True)
5. New or not?
   - Whether the house is new
     - If Yr Sold = Year Built, 1 (i.e. True)
     - Else, 0 (i.e. False)

Plug and play for lambda/alpha