Identifying ‘Undervalued’ Properties

Modelling historical sales-data on fixed and renovate-able property features to identify future buying opportunities

Company Mission

Provide an **exceptional end-to-end** residential real-estate service

The problem

- While moving toward mission, company needs to identify undervalued residential stock on market
- Objective to create value in these properties for re-selling at profit
- Need reliable method to identify ‘undervalued’ stock

The solution

- Use machine learning using property feature and historical sale data to create an algorithm that reliably estimates the value of residential houses based on:
  - Fixed features
  - Renovate-able features

The data

- **Ames Housing dataset**
  - 1461 properties in the Ames, Iowa region
  - 80 property features
    - Fixed (e.g. size of lot, slope of lot, general living area etc.,)
    - Renovate-able (e.g. overall condition, kitchen quality etc.,)
    - Sale Price
      - Year of sale
      - Sale price

Data Cleaning

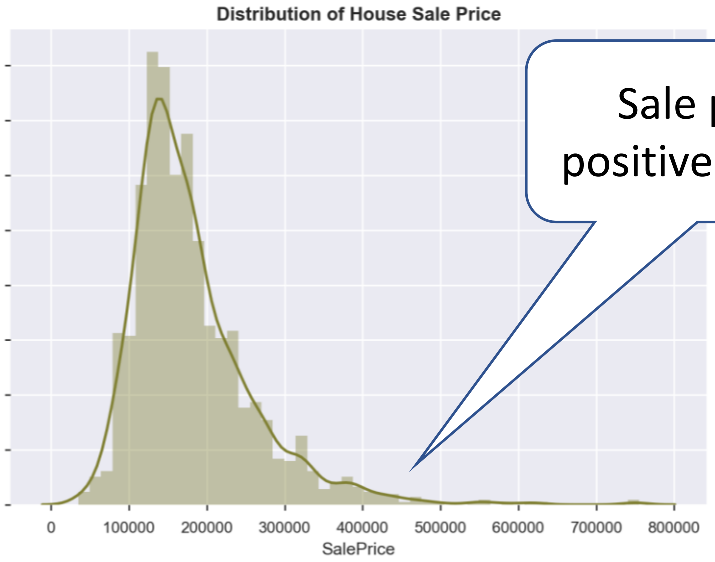
- **Null values**
  - LotFrontage 259 → imputed with median
  - Street → alley to No alley access
- **Transformations**
  - LotFrontage → log transformation
  - LotArea → log transformation
  - GrLivArea → log transformation
- **Feature Creation**
  - Ratio bed to bath
  - Total full baths
  - Total half baths
- **Recoding**
  - Condition1 → reduced number of classes
  - ExterQual → to quasi-interval variable

Fixed Features

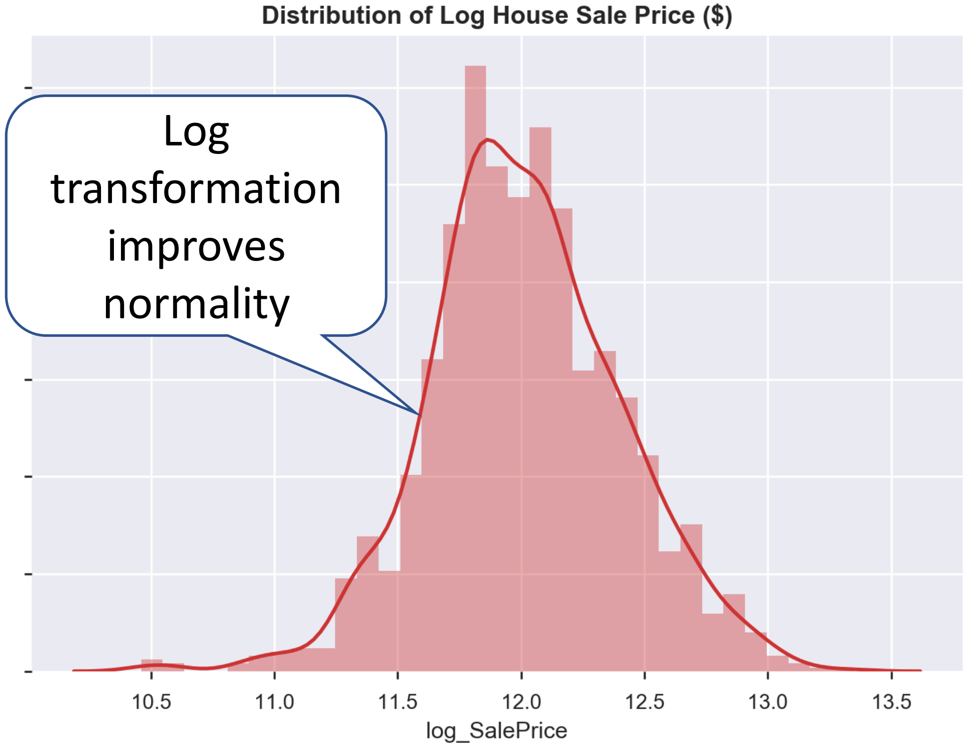
Median Sale price was **$162,000**. Sale price showed a positive skew. Therefore, a log transformation was chosen as the target.

$162,000

Median property price



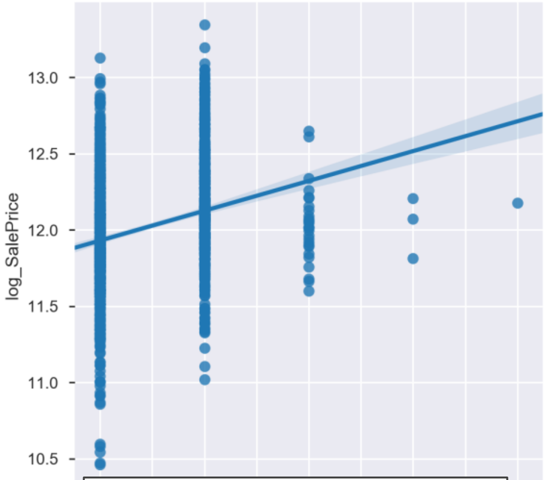
Sale price is positively skewed



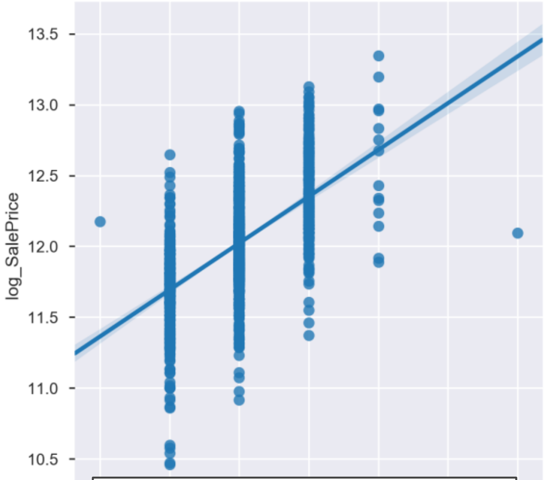
Log transformation improves normality

Target is the log_SalePrice

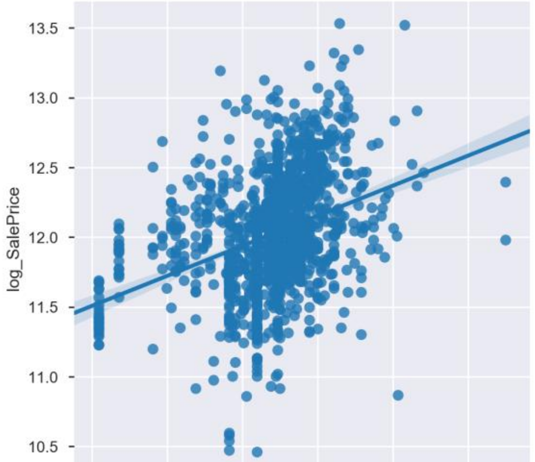
Investigation of variables showed generally expected relationships to the (log) sale price



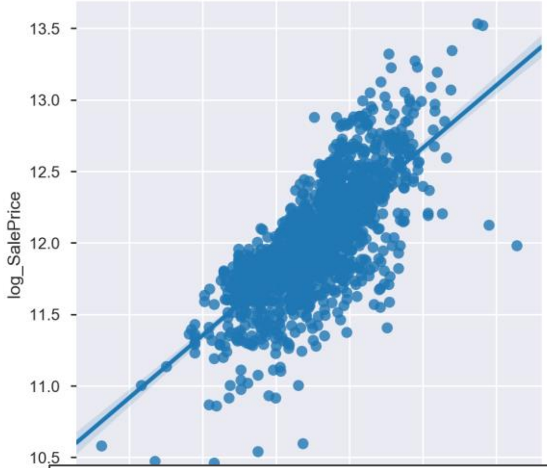
Total Half Baths



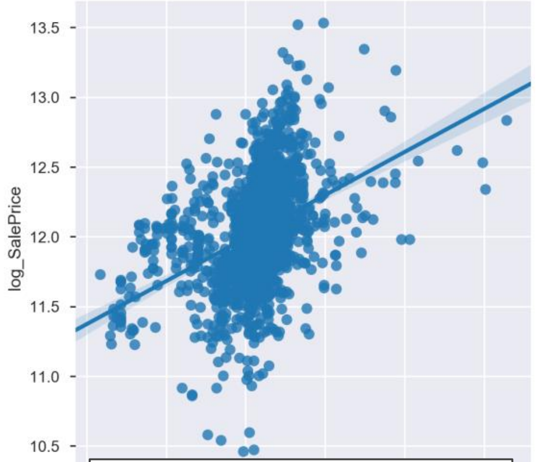
Total Full Baths



Log Lot Frontage

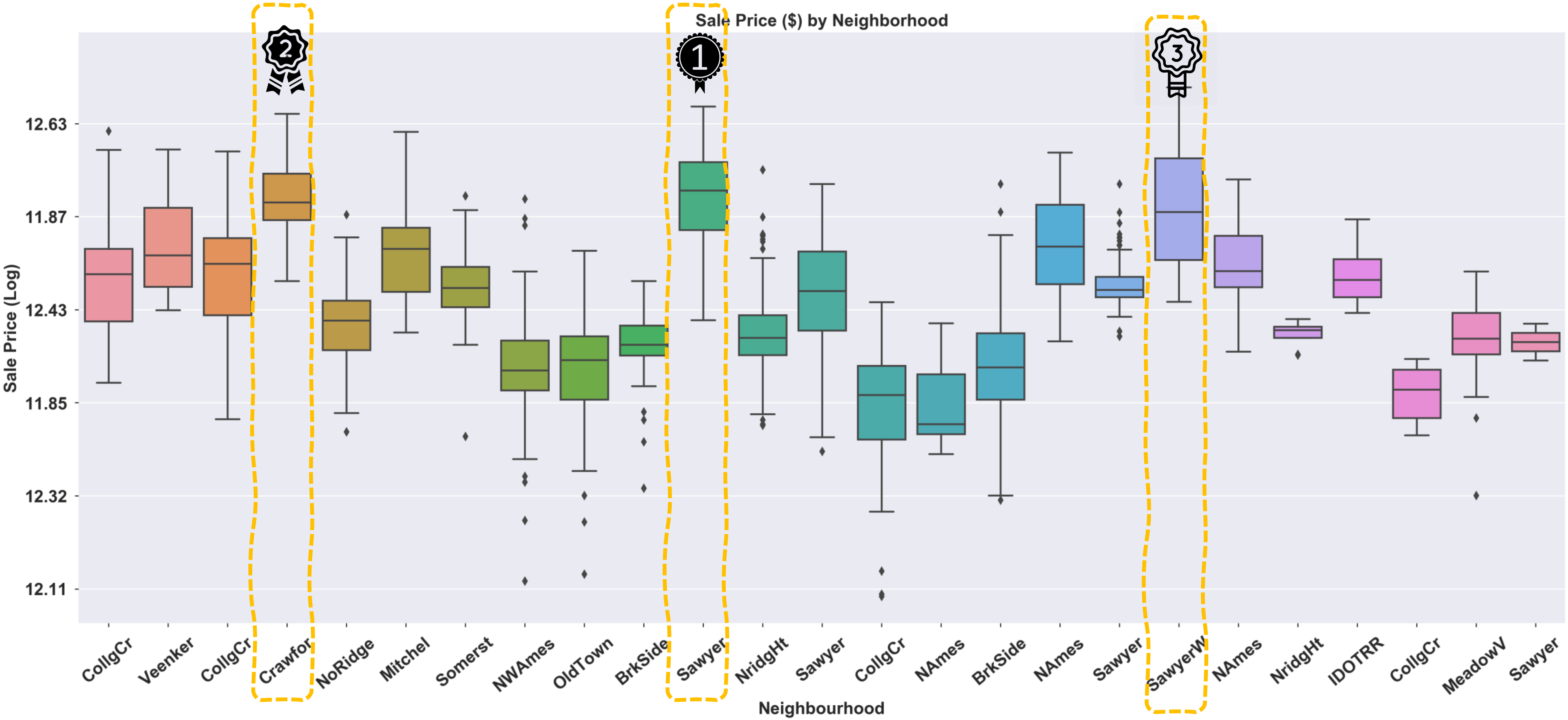


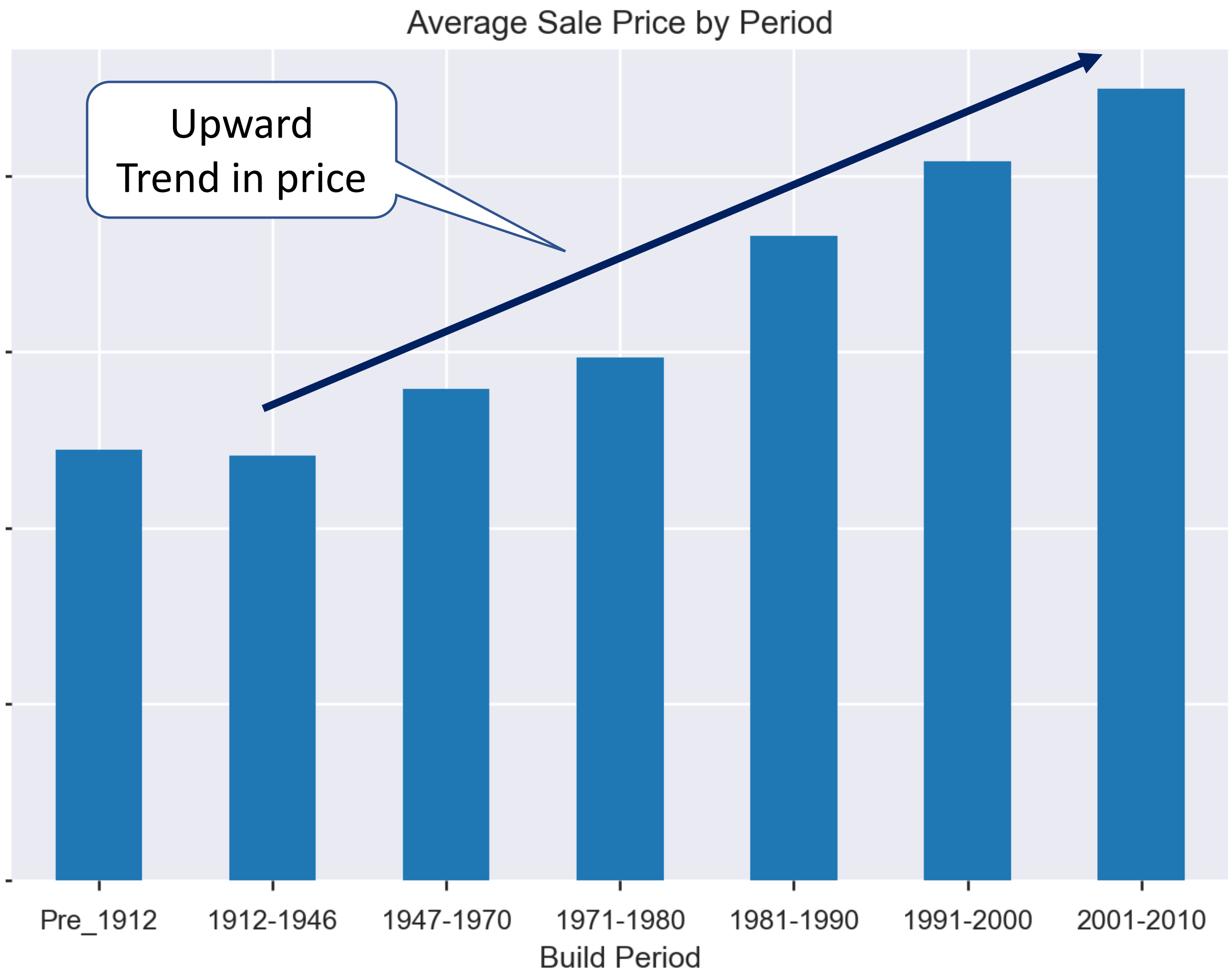
Log General Living Area



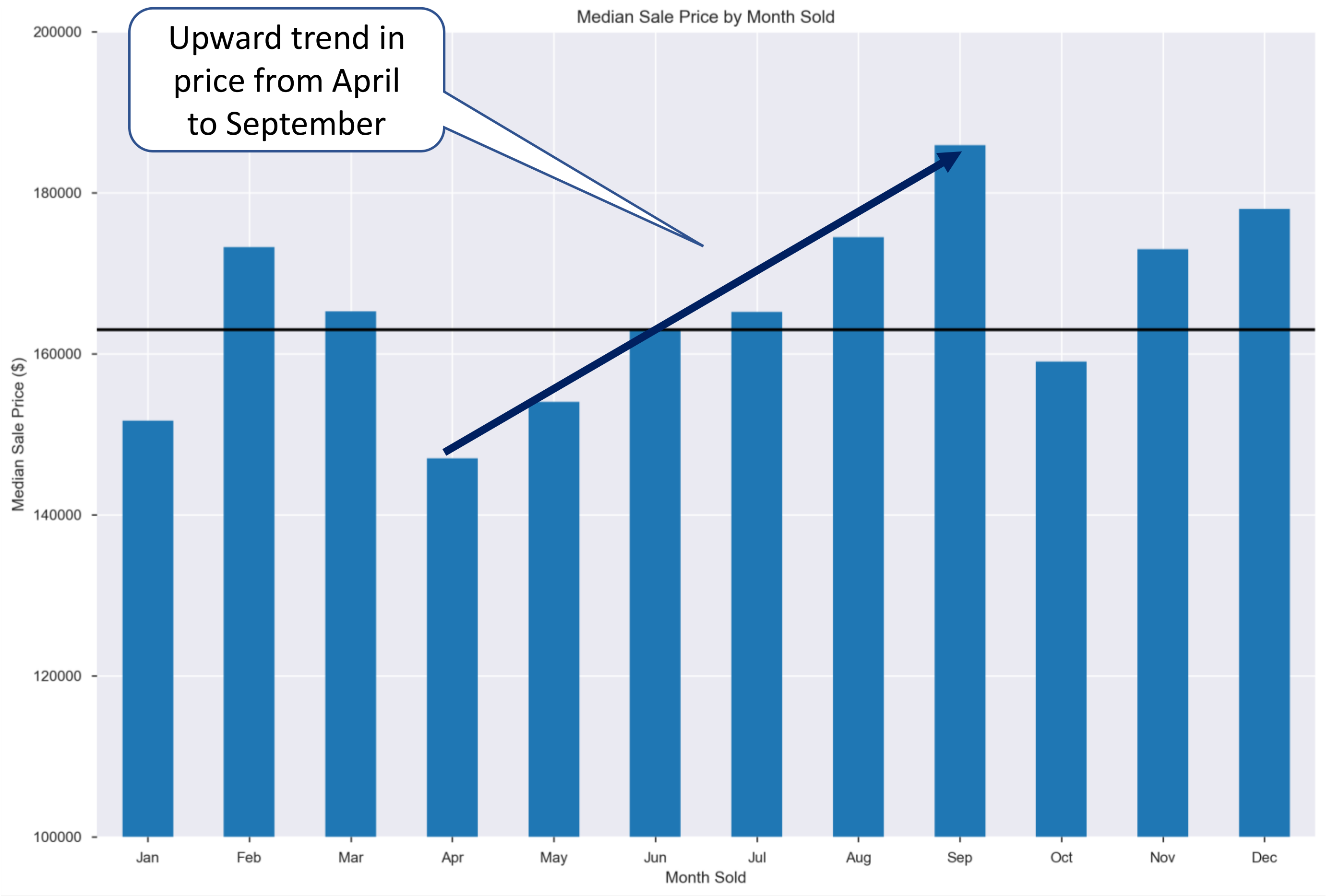
Log Living Area

(Log) sale price as expected varies across neighbourhoods, with more sought after neighbourhoods showing higher prices.





Properties built in later periods tend to have **higher sales** price. This likely represents more recent properties being designed for modern requirements around living and lower maintenance costs



There is also evidence of **seasonality** in sale price with higher sale prices generally recorded in the latter half of the year, peaking in September

Three fixed model tests were tested with similar results yielded from Linear and Lasso Regression. Linear regression showed better overall performance on cross-validation and predicting 2010 prices

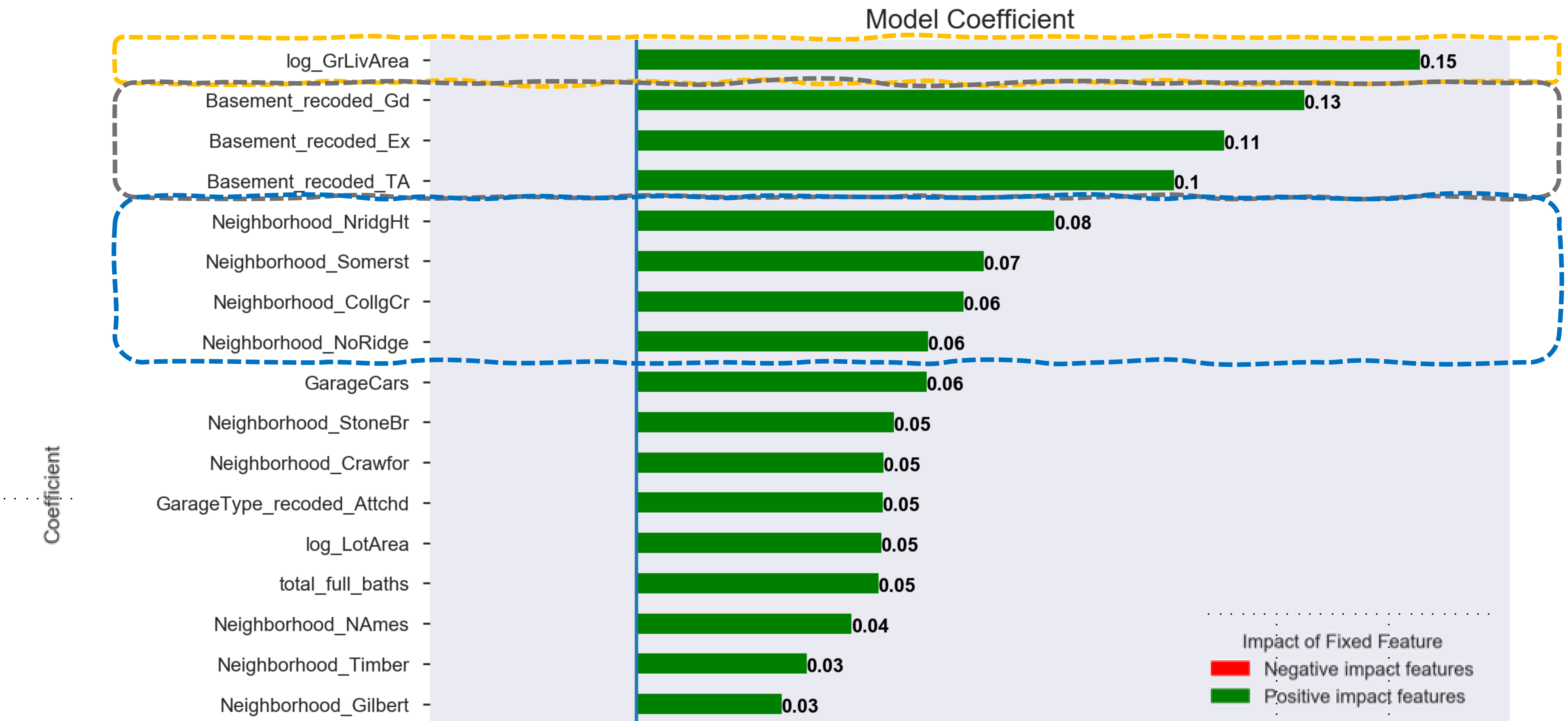
Cross-Validated approach includes all data

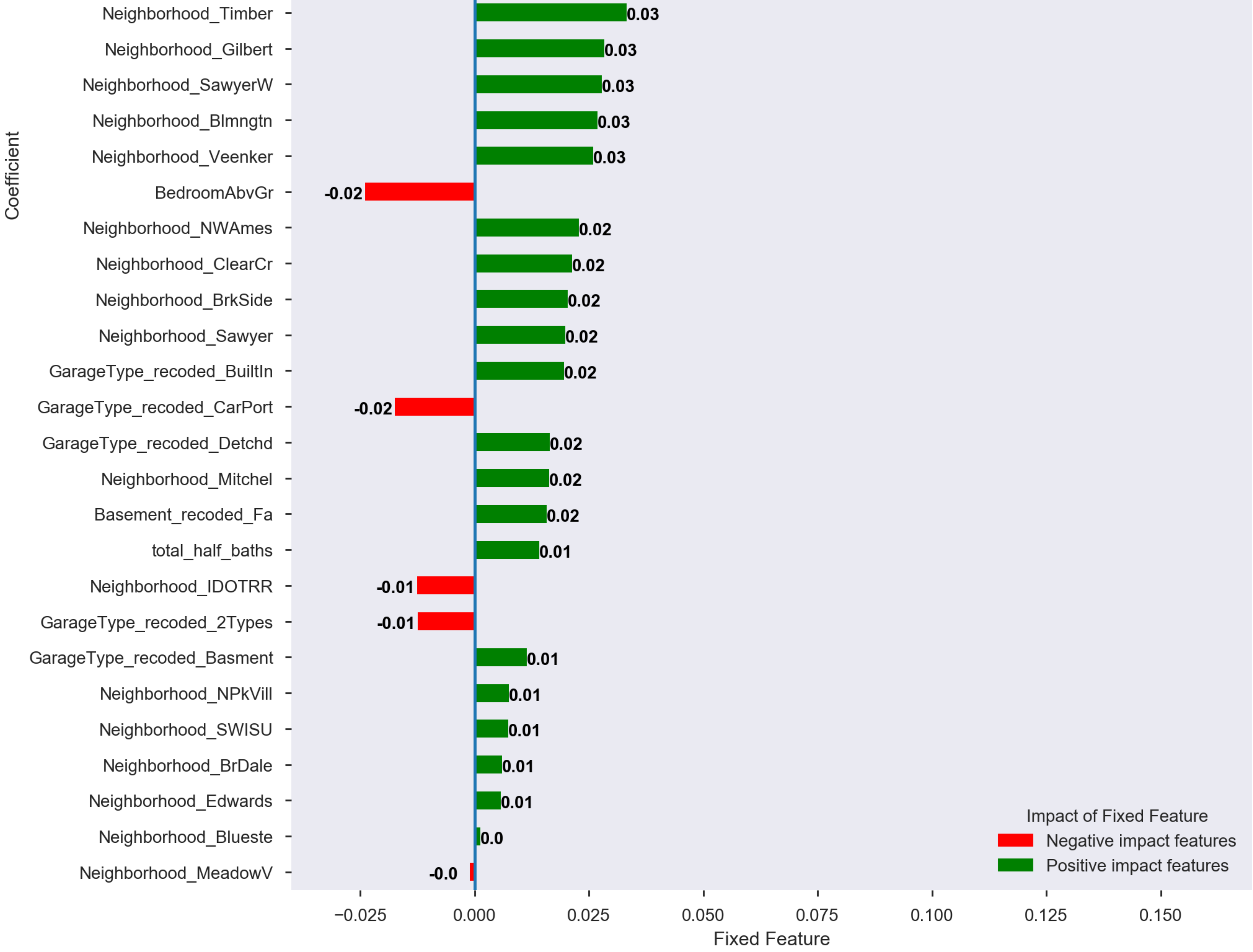
Selected Model

	Linear Regression	**Lasso Regression**	**Ridge Regression**
Cross-Validated Model (CV=10) Mean R2 (Std)	83.35% *(2.12%)*	82.97% *(2.04%)*	79.89% *(7.33%)*
Pre-2010 Model R2	85.05%	85.08%	84.48%

Training model on pre-2010 data with 2010 data used as holdout for testing

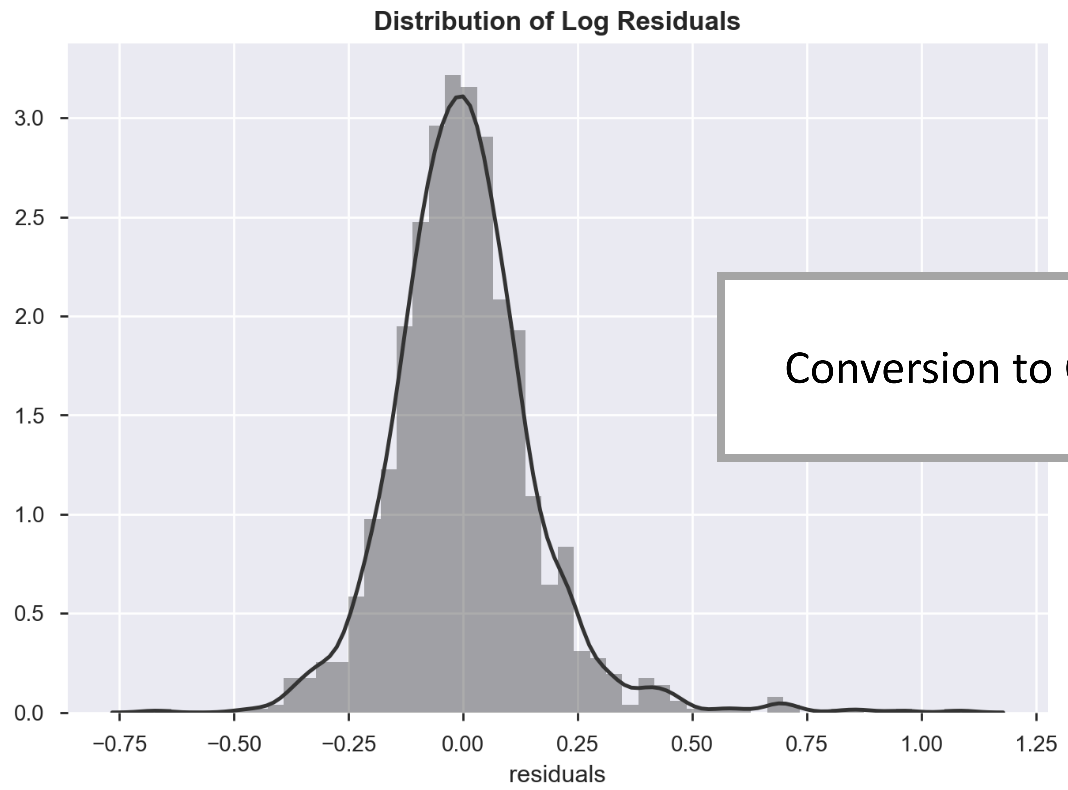
General Living Area size is the strongest predictor of sale price (log), with the height of the basement also (surprisingly) a key predictor. College Creek, Northridge Heights, Somerset, and Northridge are key neighborhoods contributing to sale price prediction



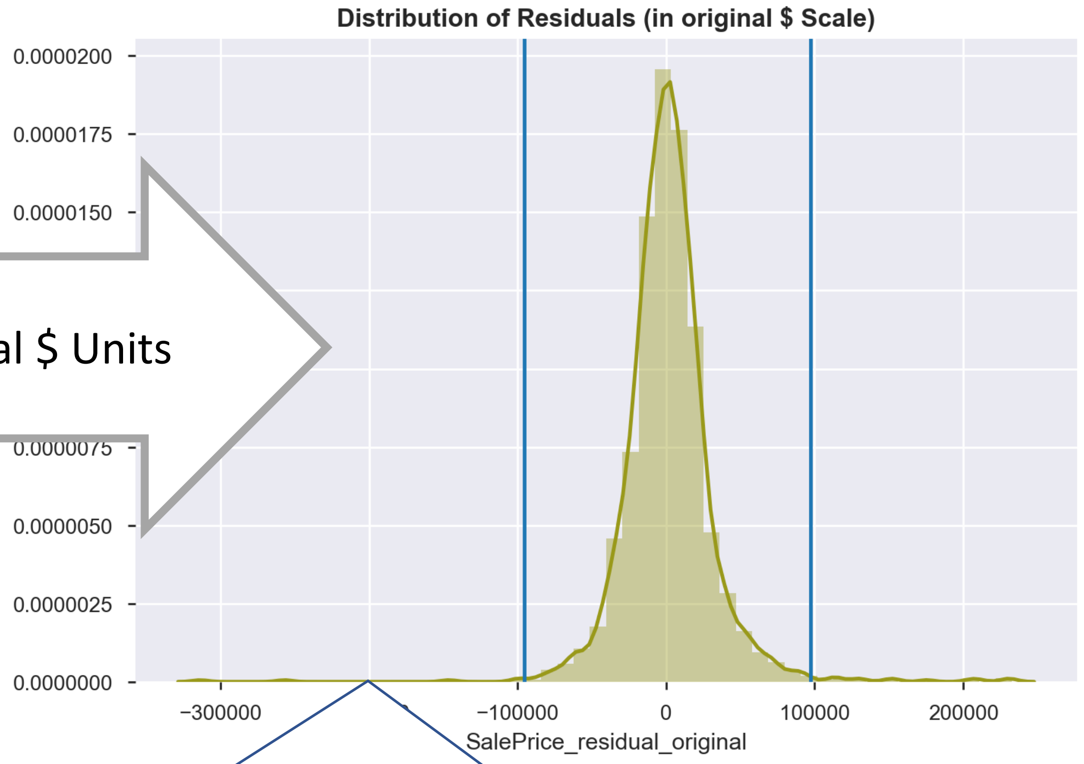

Coefficient
Neighborhood_Timber
0.03
Neighborhood_Gilbert
0.03
Neighborhood_SawyerW
0.03
Neighborhood_Blmngtn
0.03
Neighborhood_Veenker
0.03
BedroomAbvGr
-0.02
Neighborhood_NWAmes
0.02
Neighborhood_ClearCr
0.02
Neighborhood_BrkSide
0.02
Neighborhood_Sawyer
0.02
GarageType_recoded_BuiltIn
0.02
GarageType_recoded_CarPort
-0.02
GarageType_recoded_Detchd
0.02
Neighborhood_Mitchel
0.02
Basement_recoded_Fa
0.02
total_half_baths
0.01
Neighborhood_IDOTRR
-0.01
GarageType_recoded_2Types
-0.01
GarageType_recoded_Basment
0.01
Neighborhood_NPkVill
0.01
Neighborhood_SWISU
0.01
Neighborhood_BrDale
0.01
Neighborhood_Edwards
0.01
Neighborhood_Blueste
0.0
Neighborhood_MeadowV
-0.0
Impact of Fixed Feature
Negative impact features
Positive impact features
−0.025
0.000
0.025
0.050
0.075
0.100
0.125
0.150
Fixed Feature

Renovate- able Features

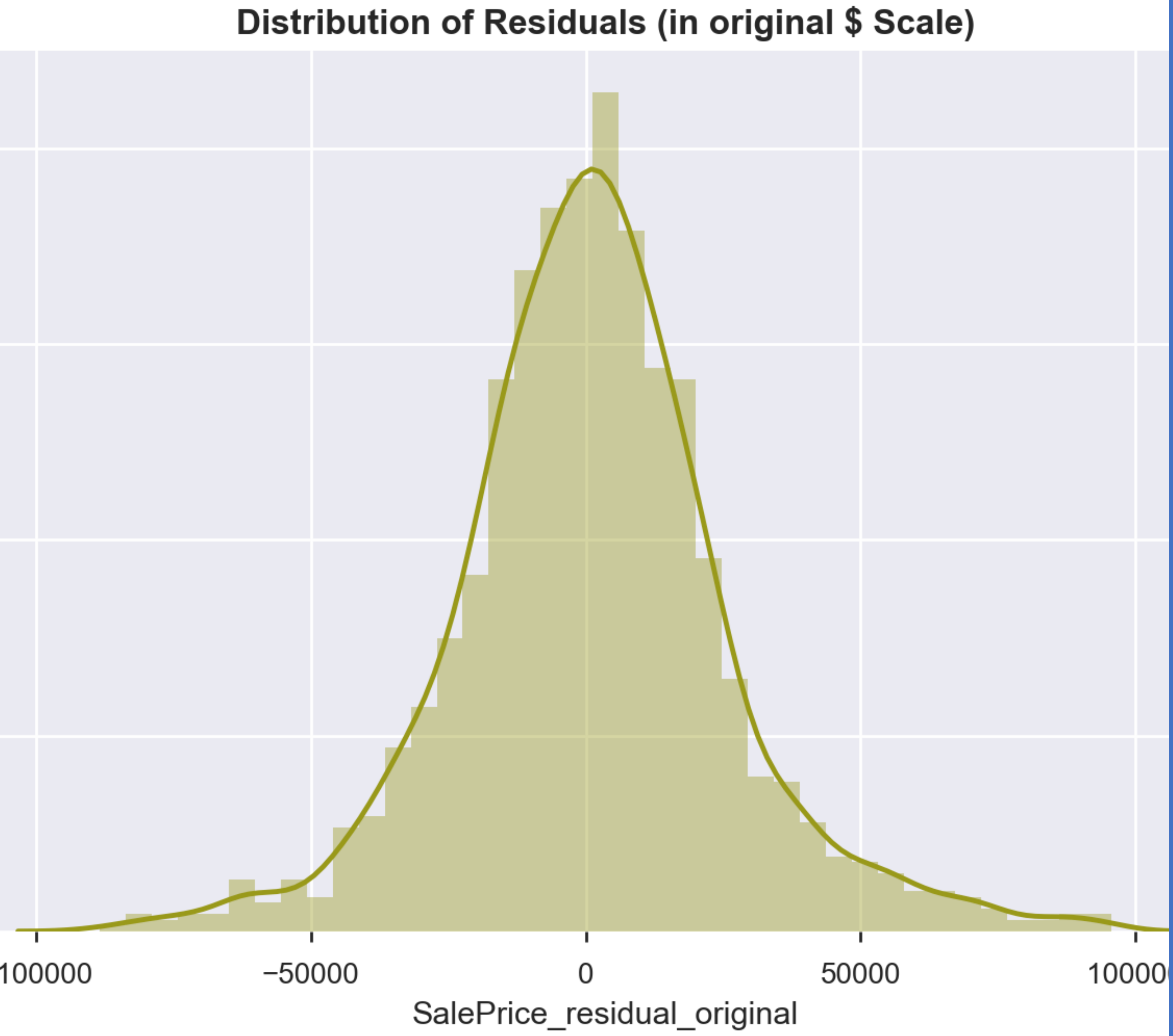
Given objective to identify expected ROI for particular renovate-able features, the residuals from the fixed model were used. Target transformed to original scale unit to aid interpretation as normality not impacted.



Conversion to Original $ Units

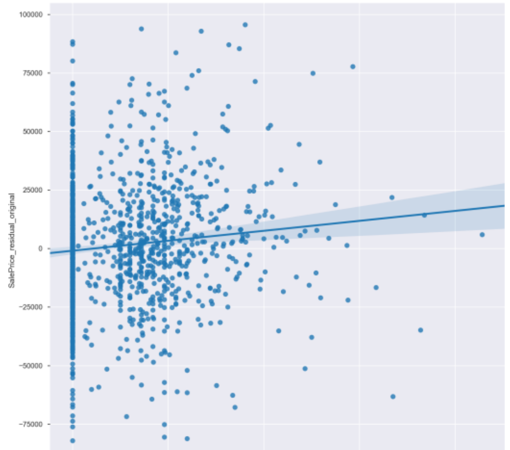


As there were some large residuals, a decision was made to remove **extreme** outliers (IQR Range * 3 denoted by lower and upper blues lines.)

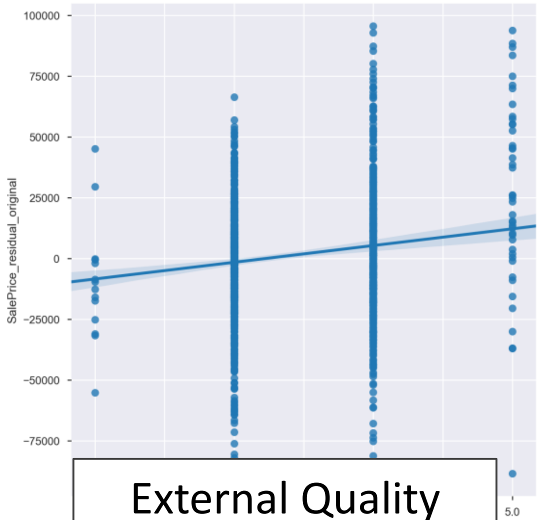


The removal of extreme residuals resulted a loss of 35 properties for modelling but resulted in a normally distributed target, which should help improve modelling accuracy and generalisation

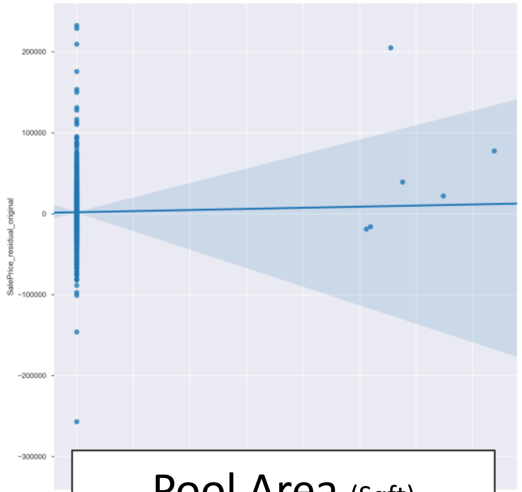
Exploration of relationships to target suggested most renovate-able features had some positive contribution to predicting sale price residuals



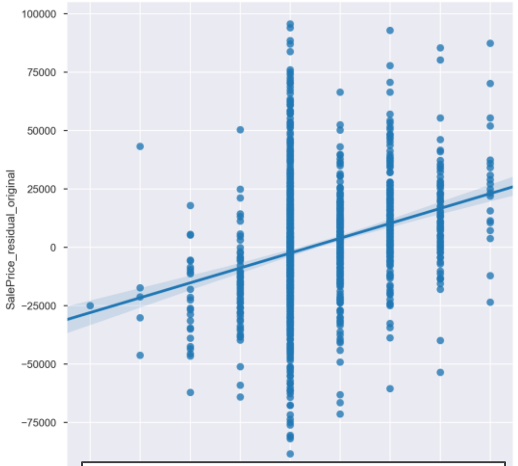
Wood Deck Size (Sqft)



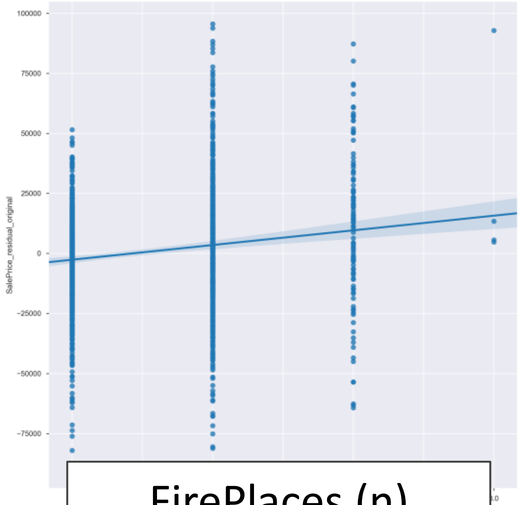
External Quality



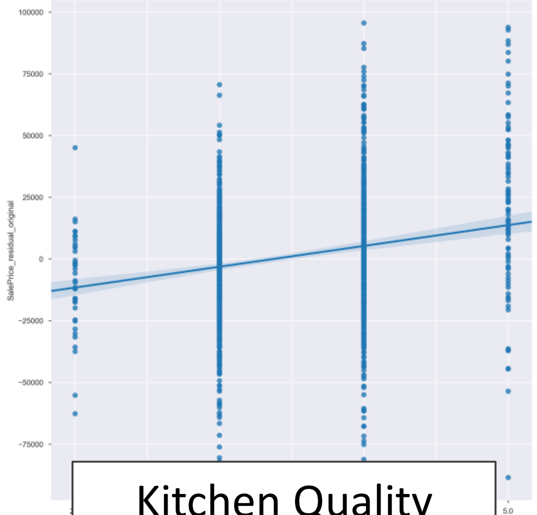
Pool Area (Sqft)



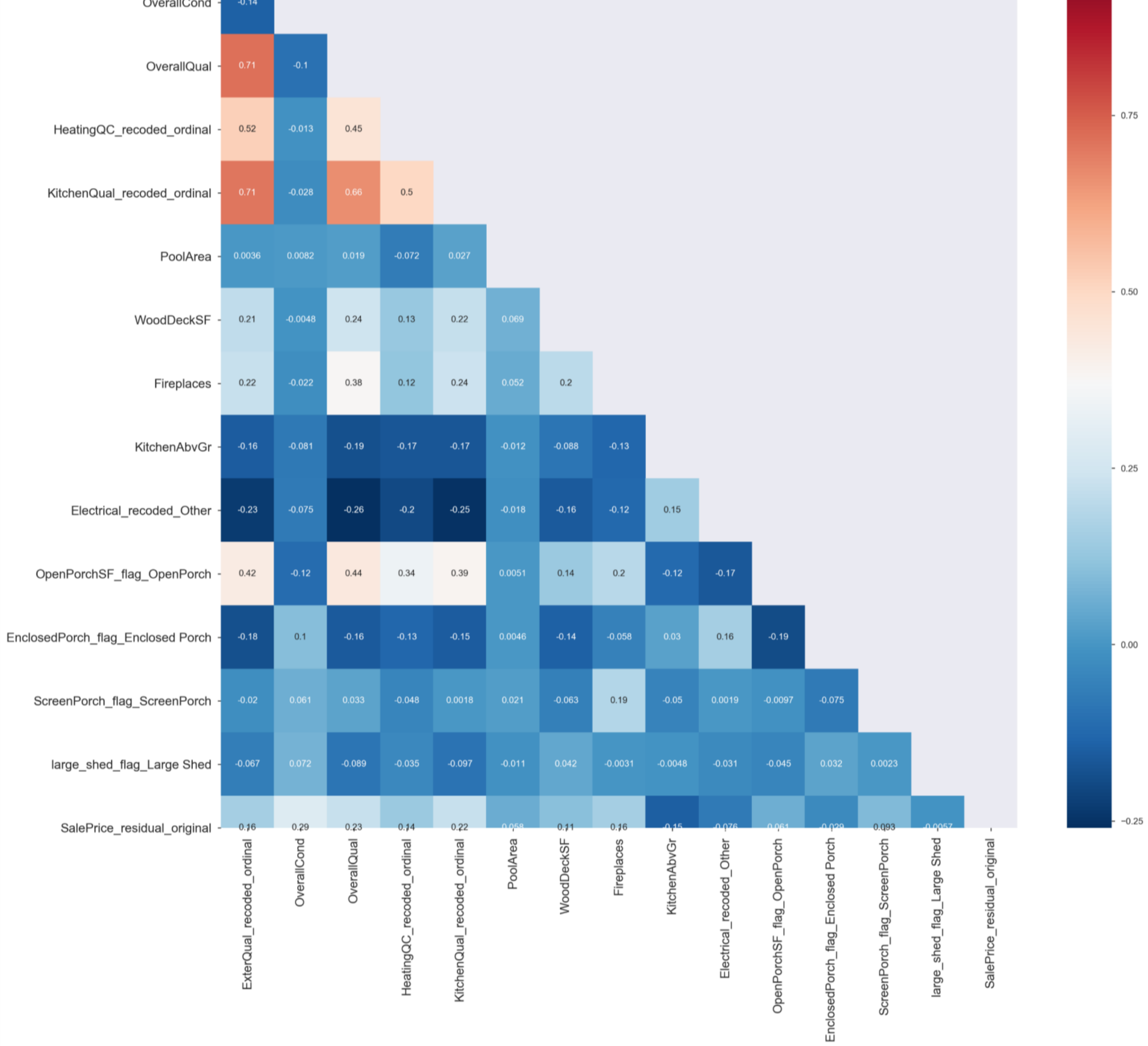
Overall Condition



FirePlaces (n)



Kitchen Quality

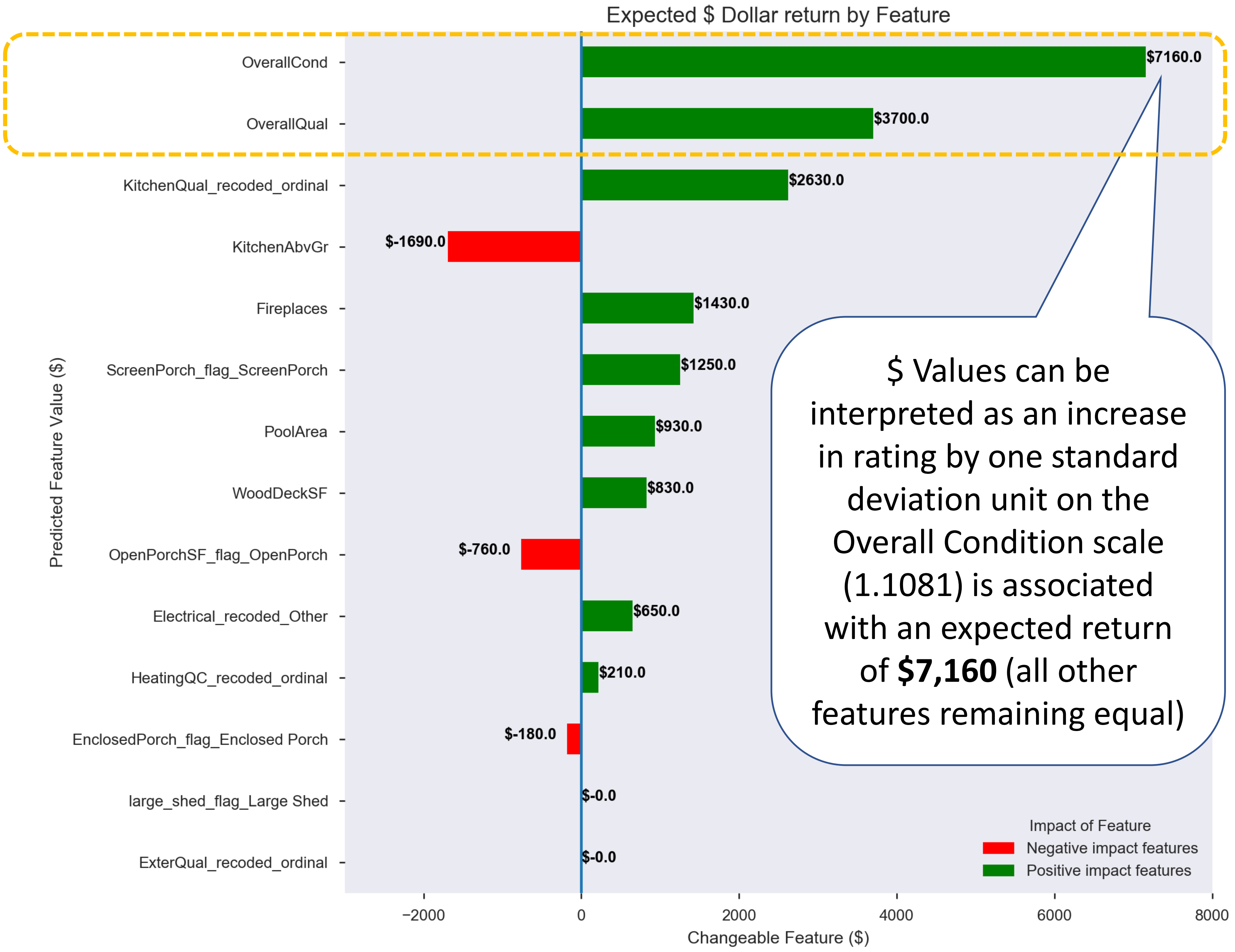


Investigation of correlations matrix among features suggests that multi-collinearity **unlikely** to be an issue

A linear and lasso regression model yielded similar results, with lasso being slightly better. However, performance was ultimately low, with less than 15% of variance in sale price residuals explained by renovate-able features

	Linear Regression	**Lasso Regression**
Cross-Validated Model (CV=10) Mean R2 (Std)	**14.47%** *(7.51%)* Predicted K-Fold cross-validation R2 scores: [-0.04, 0.13, 0.12, 0.13, 0.15, 0.19, 0.13, 0.18, 0.25, 0.21]	**14.59%** *(7.32%)* Predicted K-Fold cross-validation R2 scores [-0.04, 0.14, 0.12, 0.13, 0.15, 0.20, 0.13, 0.17, 0.25, 0.21]

As can be seen, overall condition and overall quality showed the highest expected ROI



How to use this information

There are opportunities to use the two models to identify ‘under-valued’ properties and estimate ROI from renovating

- **Fixed Features Model**
  - Use the fixed feature model help identify potentially 'undervalued' properties.
    - Where expected ROI is more than price to purchase/make changes then ‘undervalued’ property
  - Fixed features are more of a driver of sale price then renovate-able features

- **Renovate-able Features Model**
  - Calculate the expected ROI if improvement made in respective property feature
  - **Caution on this model:** Recommend only make decision where substantial expected value increase is at least 3x that of estimated renovation cost

When considering fixed features, focus on high impact fixed features but also try to capture as many possible sale price drivers as possible

- **Size of General living area**

- **Height/Quality of Basement**:
  - Focus on properties with basements more than 90cm (rated as good or excellent)

- **Target College Creek, Northridge Heights, Somerset and Northridge**
  - Identifying undervalued properties in these neighborhoods may provide greater opportunities to improve value

When renovating…

- **Focus on improving overall condition and quality**
- **Only basic improvements to existing kitchen**

Other observations

- Seek to sell in latter half of year (preferably September)
- Focus on newer properties as this has a known impact on property price

Thank you