Predicting Property Sale Price and Identifying Features to Enhance Price

Prepared by: Lindy Tan

Problem Statement

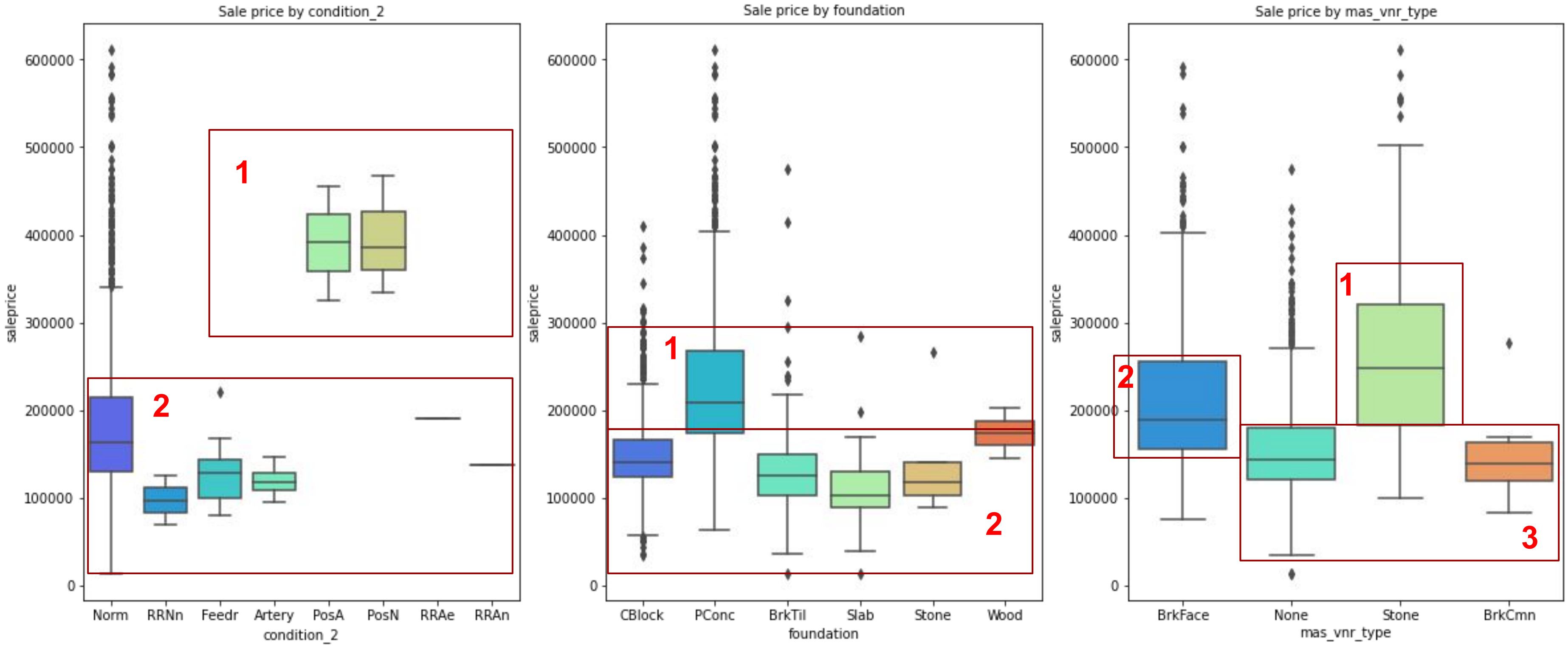
- As a property consultancy firm, we help property owners maximize the value and selling price of their properties.

- Through this project, we hope to:
  - Help property owners identify features which are most important to predict sales price
  - Provide our customers with a tool which can (i) provide quick estimates of their property potential selling prices, and (ii) help them identify which aspects of their properties they can improve on to enhance their selling prices

- Target audience: Customers (i.e. property owners)

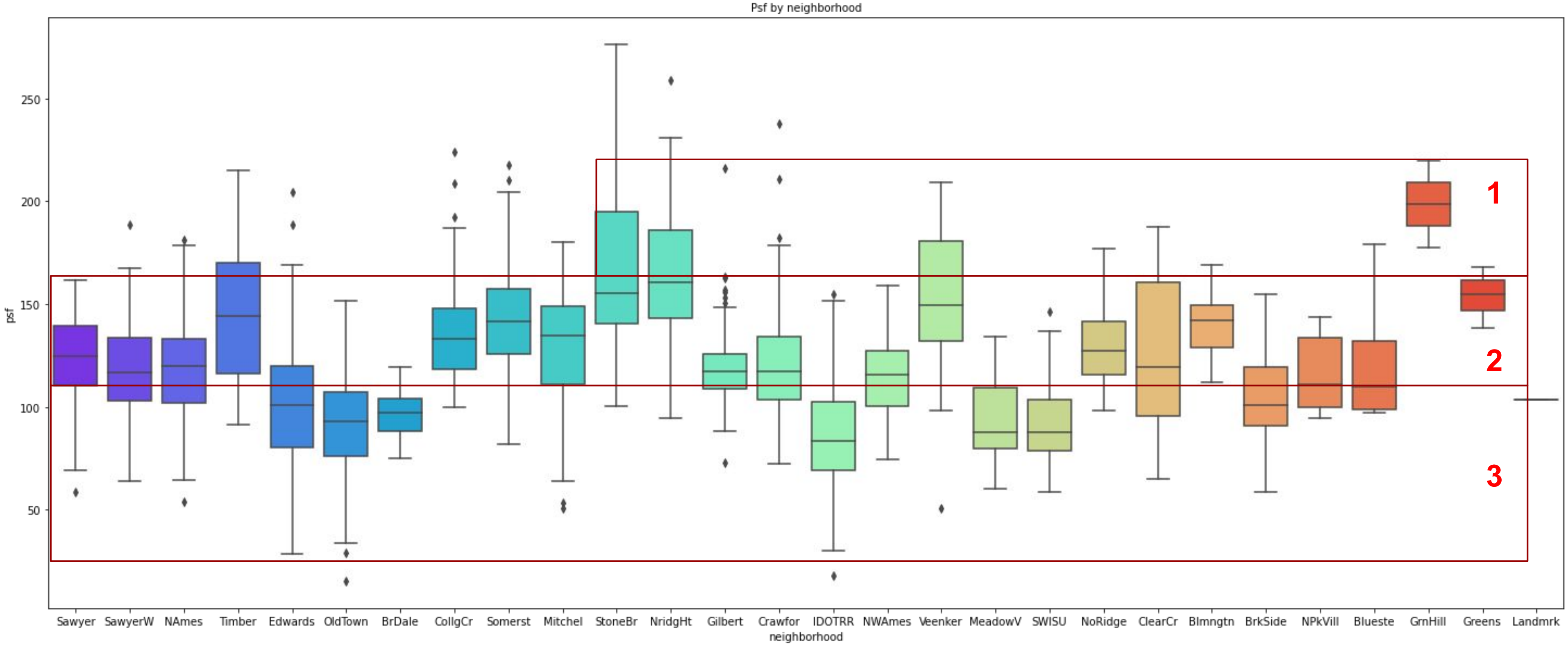
Insights from EDA: Regrouping proximity to positive off-site feature, type of foundation and masonry veneer type

Regrouping categorical features based on similarity in their sale price distributions across groups



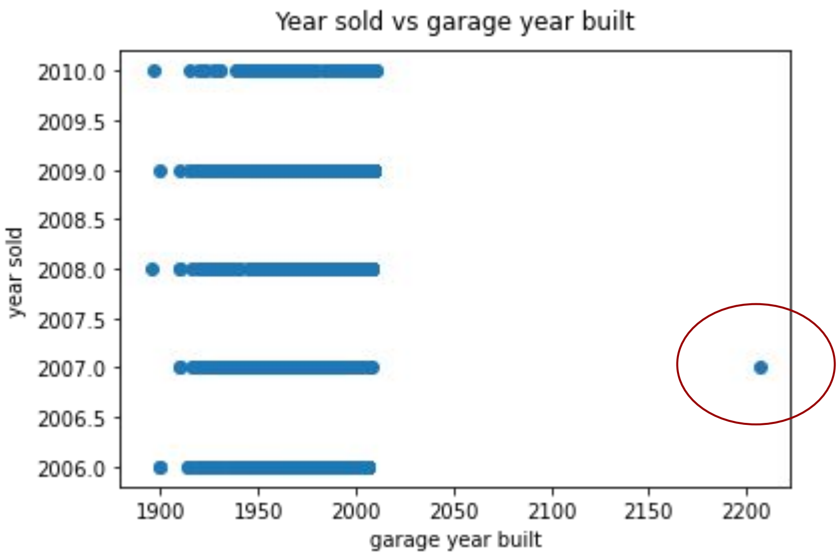
Insights from EDA: Regrouping neighbourhood

Regrouping neighbourhood based on similarity in their per square foot (psf) distribution



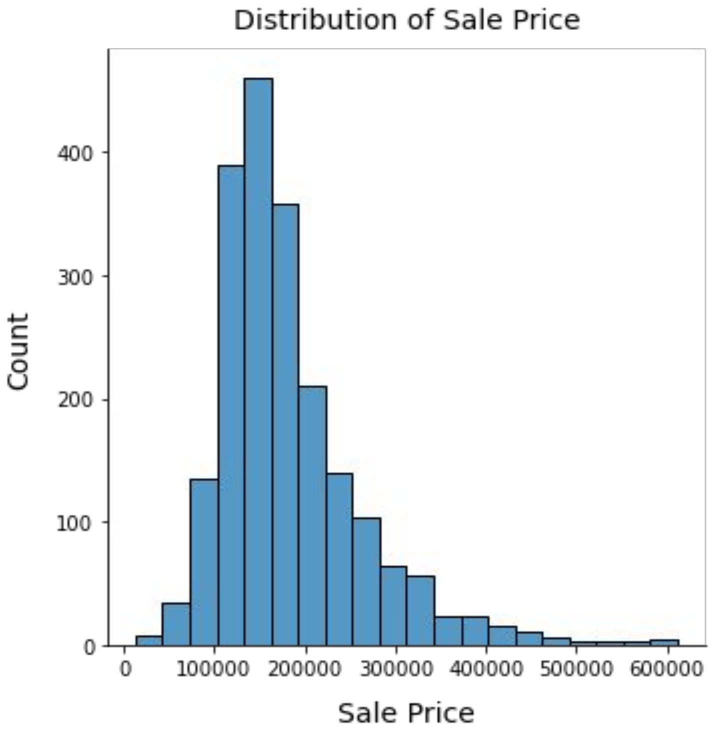
Insights from EDA: Identifying and removing outliers

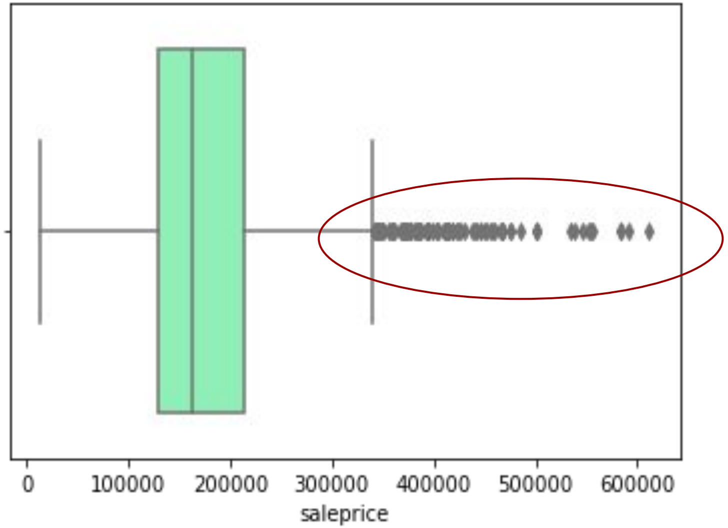
- 2 outliers where gross living area are more than 4,000 square feet
- Since they are rare, we remove them as they can skew results



- 1 outlier where garage year built is 2207 but house sold in 2007
- Set value of garage year built to 2007, same as the year sold/remodelled

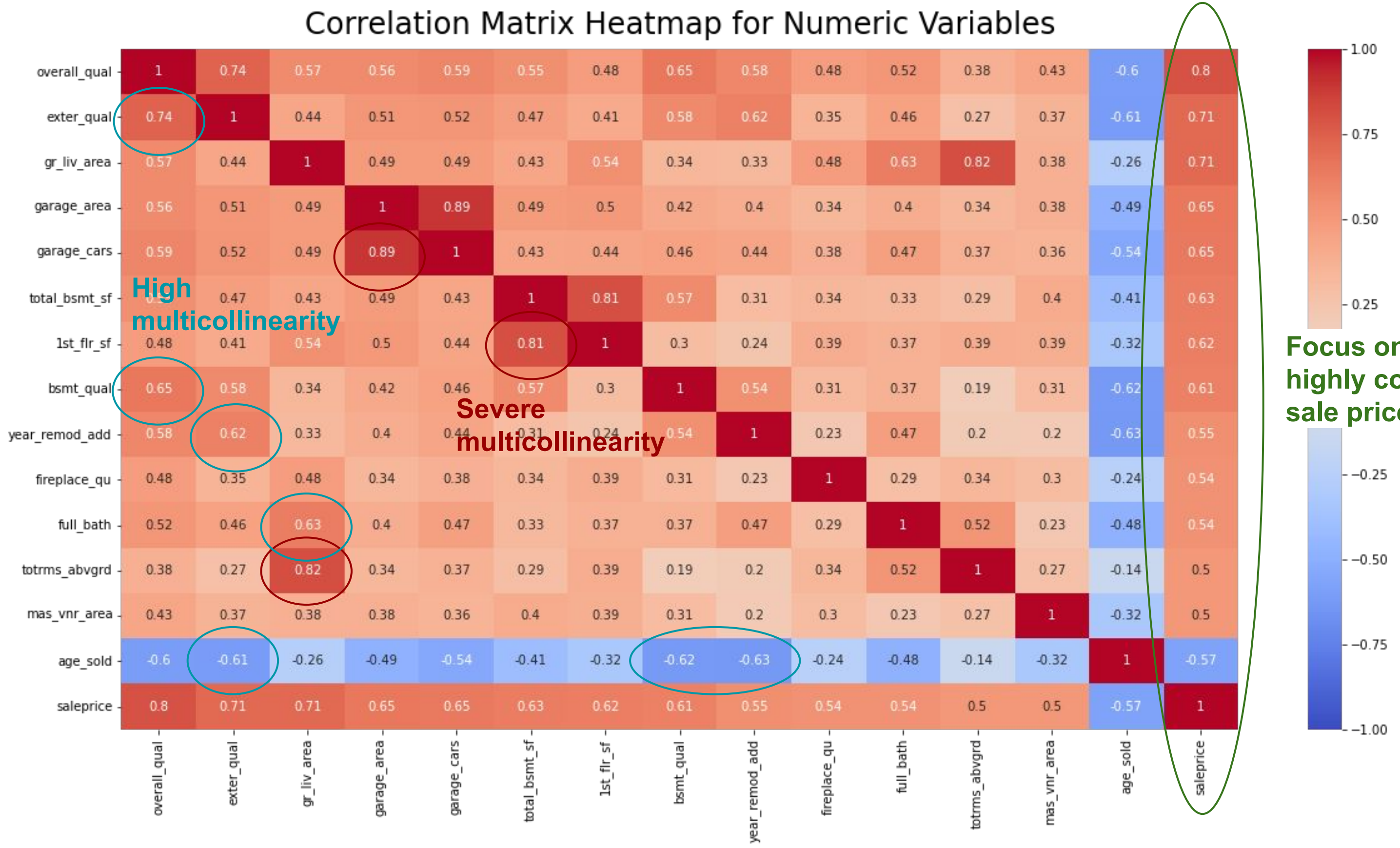
Insights from EDA: Sale price is right-skewed with outliers above $350,000.





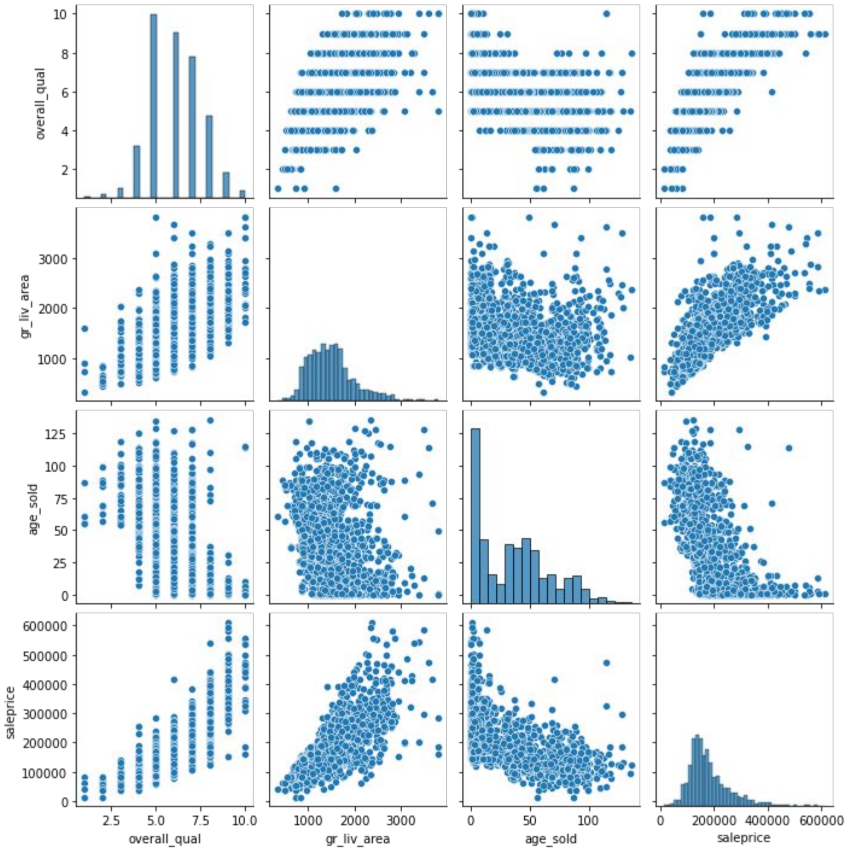
- Quite a number of outliers
- Do not deviate too much from the boxplot maximum
- Keep them in the dataset and see if we can identify features/characteristics which attribute to the high sale prices

Numeric Features: Some features correlated with sale price have severe multicollinearity with each other and similar distributions so they are excluded

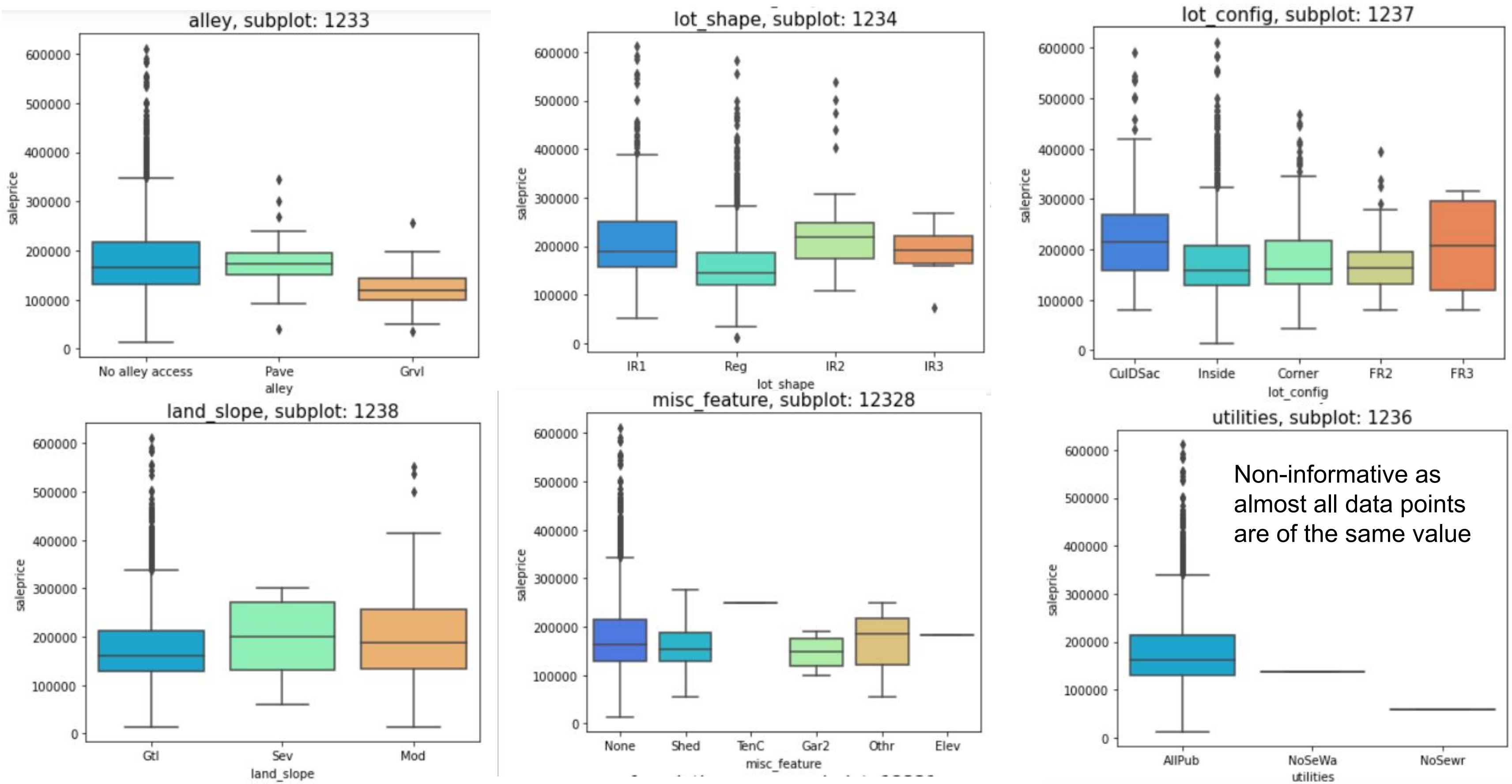


Focus on features highly correlated with sale price

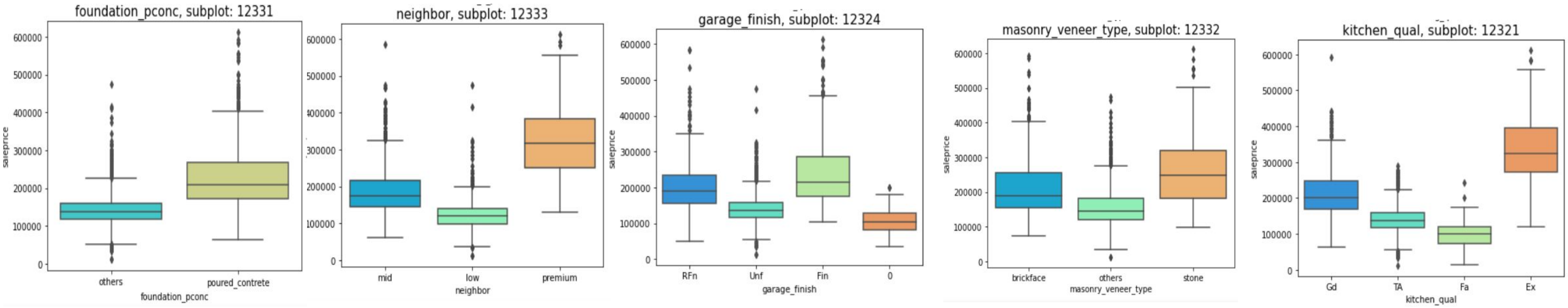
Numeric Features: Some features have non-linear relationship with sale price (e.g. overall quality) which polynomial features may help to address

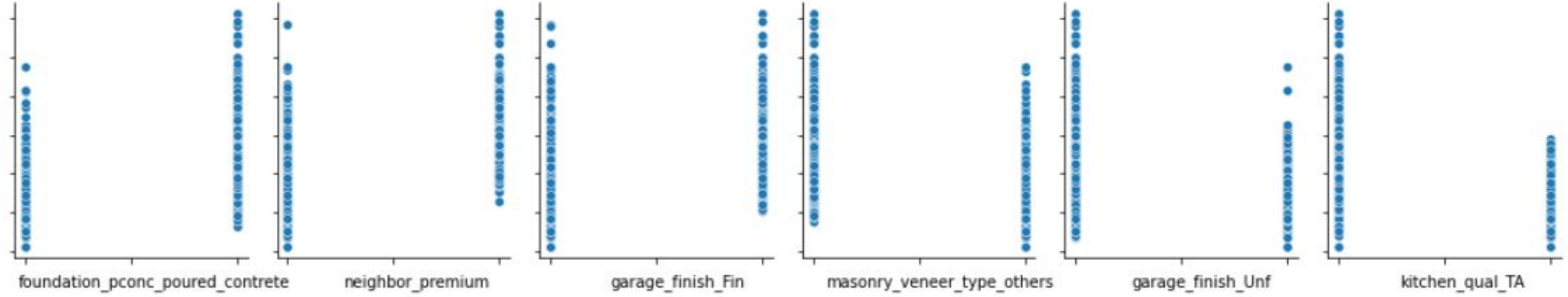


Category Features: Features with similar distribution of sale price across groups and unlikely to affect sale price can be excluded. Same for non-informative features.



Category Features: Features with varying distribution of sale price across groups and correlated with sale price should be included





Model Exploration: Performed Linear, Ridge, Lasso and Elastic Net on an extensive list of features, essential features and polynomial transformation of essential features

Model	Feature Selection	Linear	Ridge	Lasso	Elastic Net
1	• 10 numeric features with strong correlation with sale price (|correlation| > 0.5) without severe collinearity with each other (|correlation| < 0.8) ○ Overall qual, gross living area, garage area, 1st floor sqft, basement qual, year remodeled, fireplace qual, number of full bath, masonry veneer area, and age • All categorical variables (transforming to 156 dummies) except those with similar across-group sale price distribution	Hyperparameter: None CV: 2.7x10^17 RMSE: 3.4x10^13	Hyperparameter: α = 105 CV: 32,887 RMSE: 26,494 **Perform better than linear regression, as there are many features. Regularisation shrinks coefficients closer to 0, simplifying model.**	Hyperparameter: α = 585 CV: 31,790 RMSE: 26,361	Hyperparameter: α = 0.4 l1 ratio: 0.5 CV: 32,711 RMSE: 27,289
2	• 7 numeric features with strong correlation with sale price (|correlation| > 0.5) without severe collinearity with each other (|correlation| < 0.6) ○ Overall qual, gross living area, garage area, 1st floor sqft, fireplace qual, masonry veneer area, and age • 5 categorical features (transforming to 12 dummies) with high correlation with sale price (|correlation| > 0.4) ○ Neighborhood, foundation poured concrete, masonry veneer type, garage finish, kitchen qual	Hyperparameter: None CV: 33,454 RMSE: 28,311 **Performs better than Model 1 due to fewer yet important features**	Hyperparameter α = 13 CV: 33,443 RMSE: 28,308 **Perform comparable to linear regression, as fewer yet important features are used (i.e. lesser regularisation)**	Hyperparameter: α = 5 CV: 33,454 RMSE: 28,309	Hyperparameter: α = 0.2 l1 ratio: 0.95 CV: 33,442 RMSE: 28,308
3	• Transforming the 7 numeric features in Model 2 into 35 polynomial features • Same dummy features as Model 2	Hyperparameter: None CV: 30,039 RMSE: 25,696 **Performs better than Model 2 due to accounting of non-linear relationship**	Hyperparameter: α = 8 CV: 29,655 RMSE: 25,256 **Some regularisation is performed but lesser than in Model 1 (as evident by smaller α)**	Hyperparameter: α = 98 CV: 29,677 RMSE: 25,328	Hyperparameter: α = 0.2 l1 ratio: 0.95 CV: 29,651 RMSE: 25,258

Final Model: Linear Regression on 7 numeric features and 5 categorical features as it performs reasonably well and is easily interpretable for property owners

- Model performs reasonably well in providing home owners an estimate of their potential selling prices
  - Prediction error ~$28,000 (~15% deviation from mean sale price of ~$180,000)
  - This is not much higher than the prediction error ~$25,000 (~14% deviation from mean sale price) based on Lasso Regression on polynomial features

- Only 12 features required
  - **Quick and easy**: Minimal features for home owners to to input into the tool
  - **Accurate**: Obtain reasonably accurate estimates of their potential selling prices

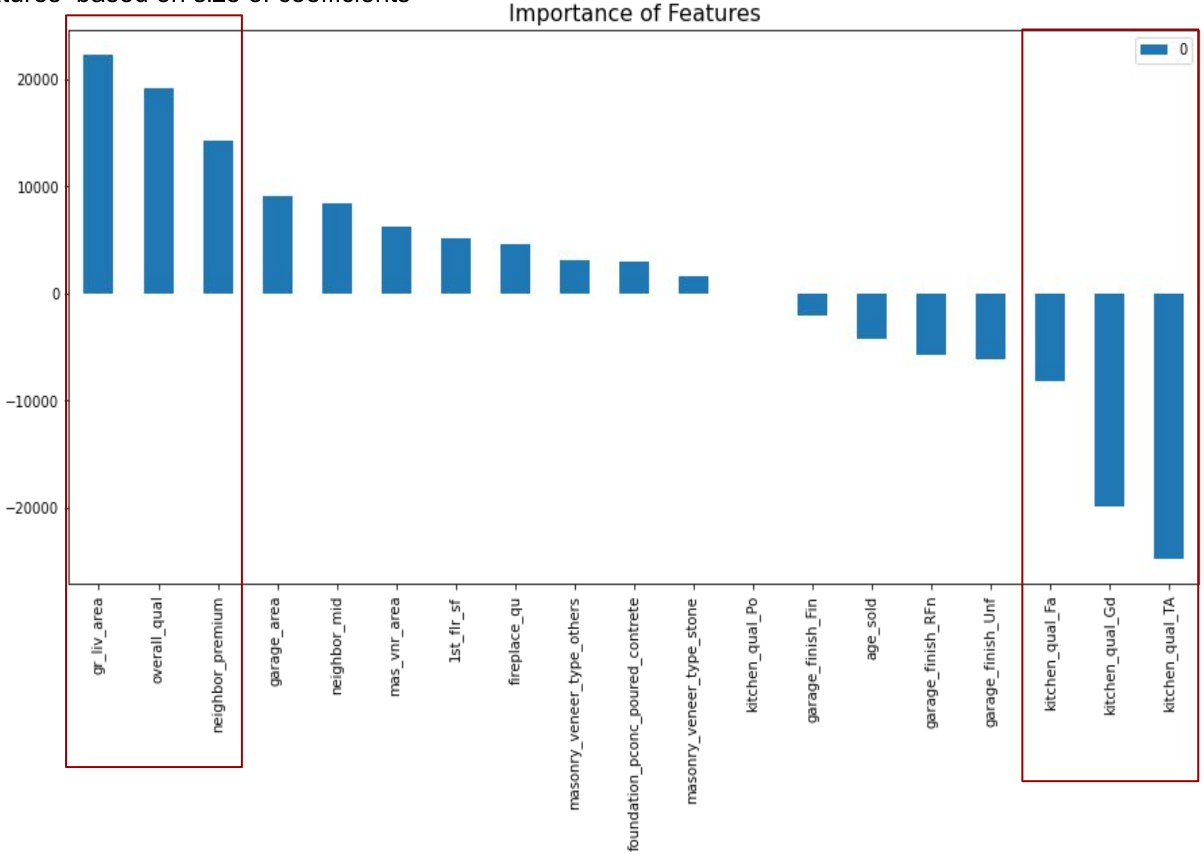
Numeric	**Categorical**
Overall quality	Neighborhood
Gross living area	Foundation poured concrete
Garage area	Masonry veneer type
1st floor sqft	Garage finish
Fireplace quality	kitchen quality
Masonry veneer area	
Age of property	

Recommendation: Property owner can improve kitchen quality, fireplace quality, using the right materials for masonry veneer and foundation to enhance selling price

Rank the importance of features based on size of coefficients

Features which affect sale price most (in decreasing importance):

- Living area (sq ft)
- Rating of overall material and finish of house
- Kitchen quality being excellent
- Neighbourhood in premium location (i.e. NridgHt, GrnHill and StoneBr)

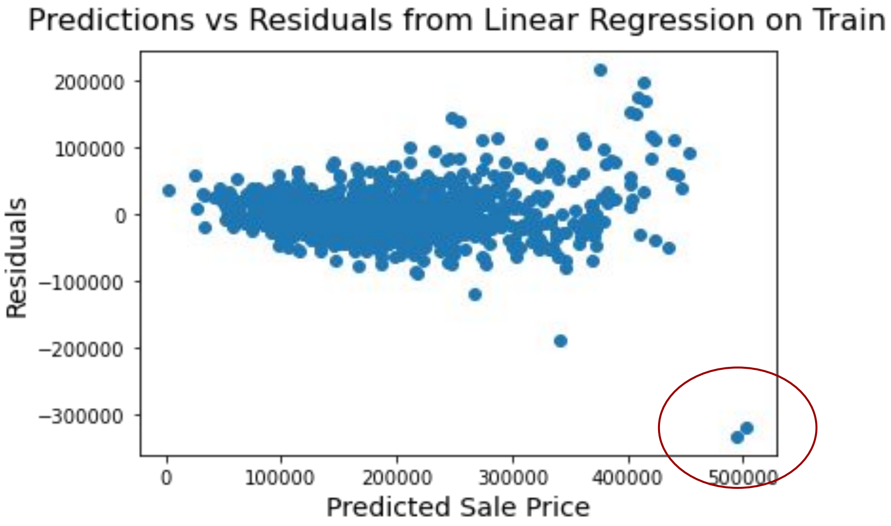


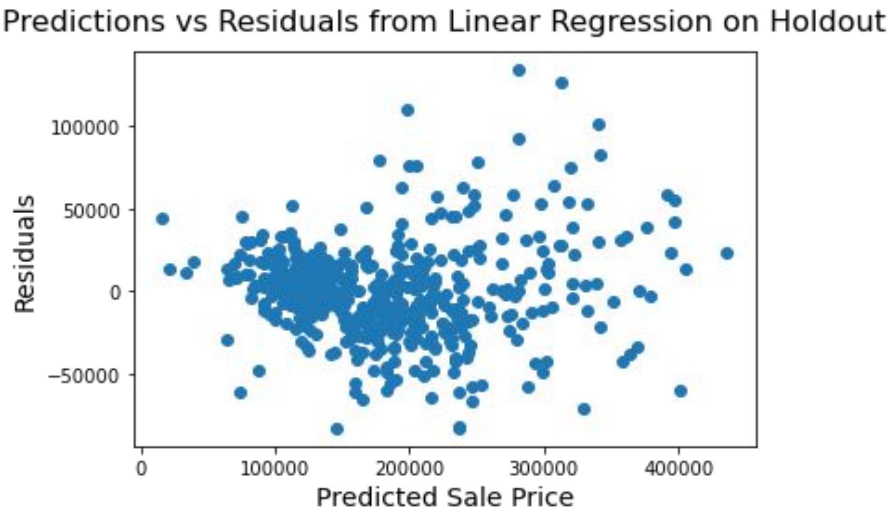
Features which home owners can improve to enhance selling price:

- Fireplace quality - improve rating (*for every unit increase in rating, average selling price increases by $5,000*)
- Kitchen quality - improve to excellent
- Masonry veneer type - use brick common/cinder block, followed by stone (instead of brick face)
- Foundation - use poured contrete

Error Analysis: Residuals are randomly scattered around 0, but there are 2 data points with high residuals

- Residuals plot for both train and holdout data centre randomly around 0.
- 2 data points in train set are not predicted well by the final model (i.e. deviation of more than $300,000)
- Further deep-dive into these shows that these properties are under-valuated
    - Predicted price ~$500,000 but sold at less than $200,000





Trade-off: Model found to perform the best may not be selected as final model to address the problem statement

- Lasso, Ridge and Elastic Net on polynomial transformation of the 7 numeric features and 5 categorical performs the best

- Linear Regression without polynomial transformation of numeric features is selected as final model due to interpretability with some trade off in terms of increased prediction error
  - Size of the coefficient is more interpretable
  - Can use to identify features which are most important to predict sales price and features which can be improved to enhance property selling prices

Learning Points

- Regularisation, through Ridge, Lasso and Elastic Net, can be used to shrink coefficients closer to 0, dropping out insignificant features

Pros of Regularisation	**Cons of Regularisation**
• Performs better than Linear regression especially when there are many features • Prevent overfitting of model, and a more simplified model, which can better generalise and predict unseen data • As number of features increases, the hyperparameter (α) increases, having more penalisation on the coefficients	• Hard to interpret the coefficients of the model

Areas for Future Work

- Explore more models with varying features to determine the linear regression model which can best predict property sale price