Predictive modelling: Housing prices in Ames, Iowa

DSI-18 Project 2

Sahaj Chawla - November 2020

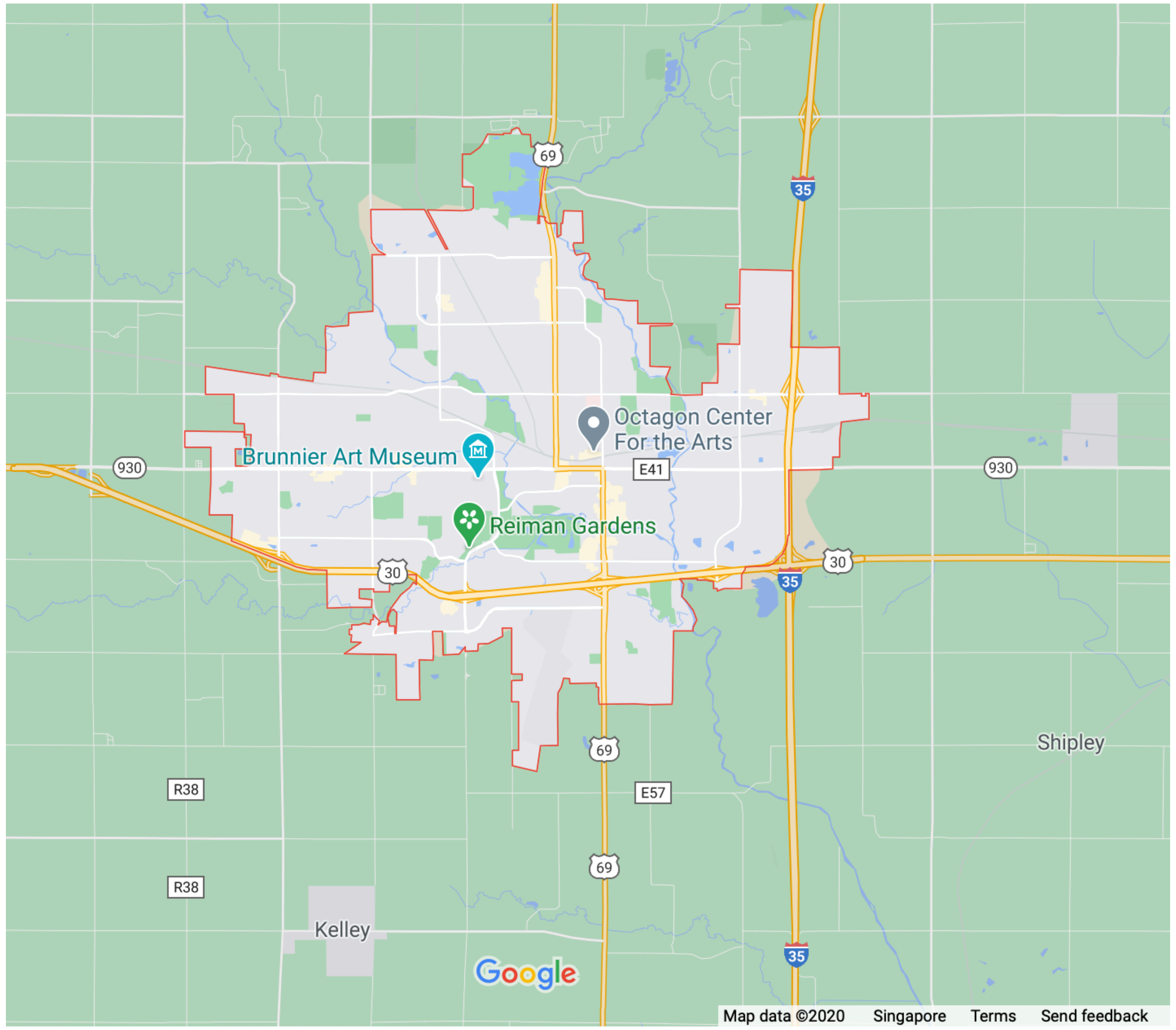
Background and Context

Why are we here?

- **Business Objective** - How can we maximise sales prices for prospective house sellers in Ames, Iowa?

- **Data Science Objective** - Which features are predictive of home sale prices, and how much value do they add?

    - *Bonus objective - which modelling approaches lead us to the most accurate predictions?*



Overview of the data

What are we working with?

- 2050 records of home sales in Ames, Iowa from 2006 to 2010.
- Data contains 80 ‘features’/variables including, but not limited to -
    - (Categorical) Type of housing/sale
    - (Continuous) Year of sale/remodelling/construction
    - (Continuous) Square footage of houses/bedrooms/garage/basement
    - (Ordinal) Rating quality of overall house/kitchen/basement/heating etc.
- **Our target variable for this analysis is to derive sales price.**

Analysis Workflow

What did we do?

1. Data cleaning: Removing outliers, standardising categorical variables.
2. Exploratory Data Analysis: Check correlations to guide initial hypothesis and feature selection.
3. Feature Engineering: Reducing the noise and amplifying the signal
4. Model Iteration and Selection: Preparing data with train/validate splits, then running ridge/lasso/elasticnet models for further feature selection. Models were compared on RMSE scores.
5. Model Evaluation: Understand what's working and what can be improved for the model, along with any caveats.

Exploratory Data Analysis

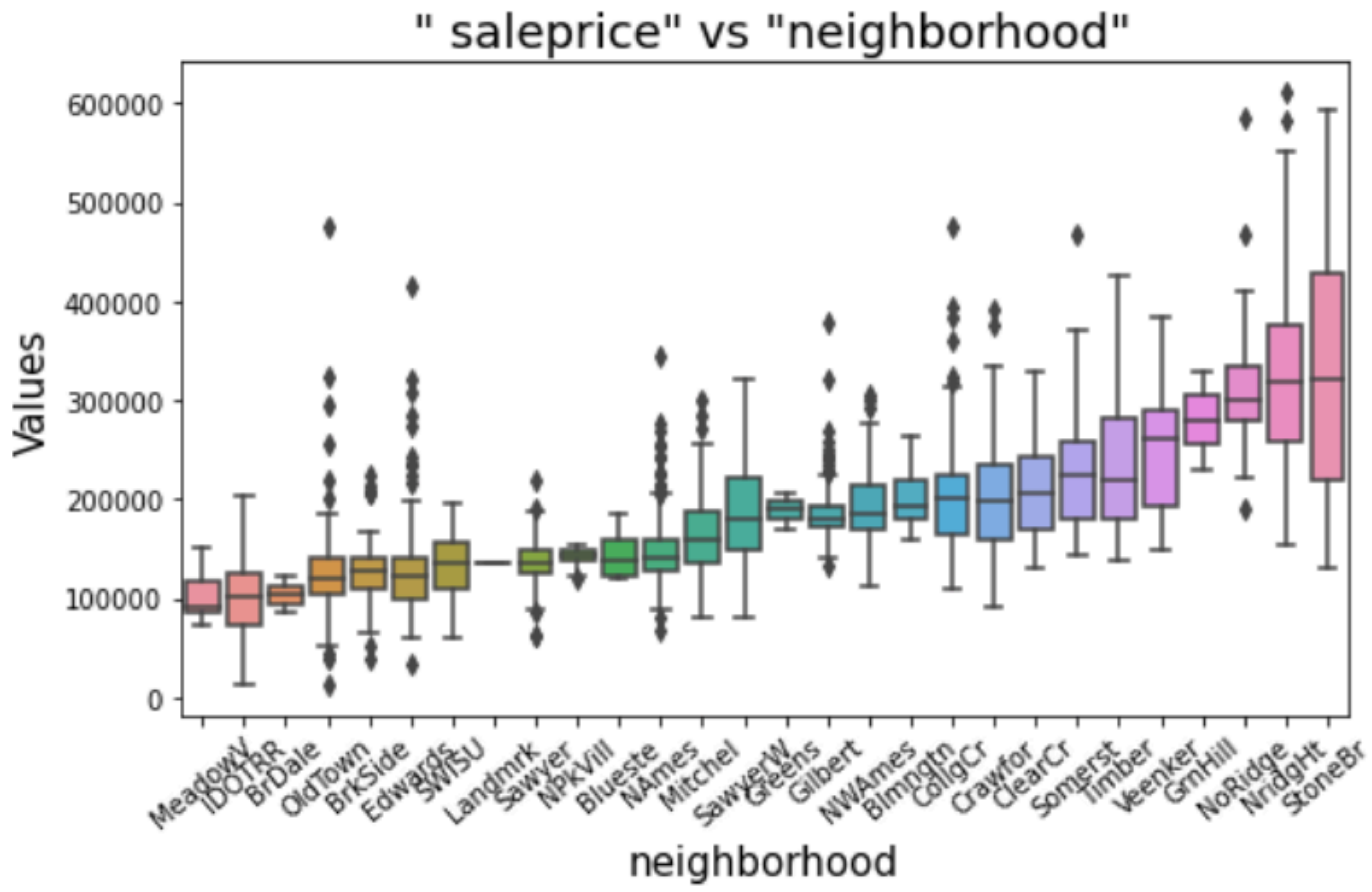
What did we learn?

Some surprises and some ‘no-brainers’

- Numerical data was looked at via correlation, and categorial via boxplots

- It is evident that certain neighbourhoods are more attractive than others.

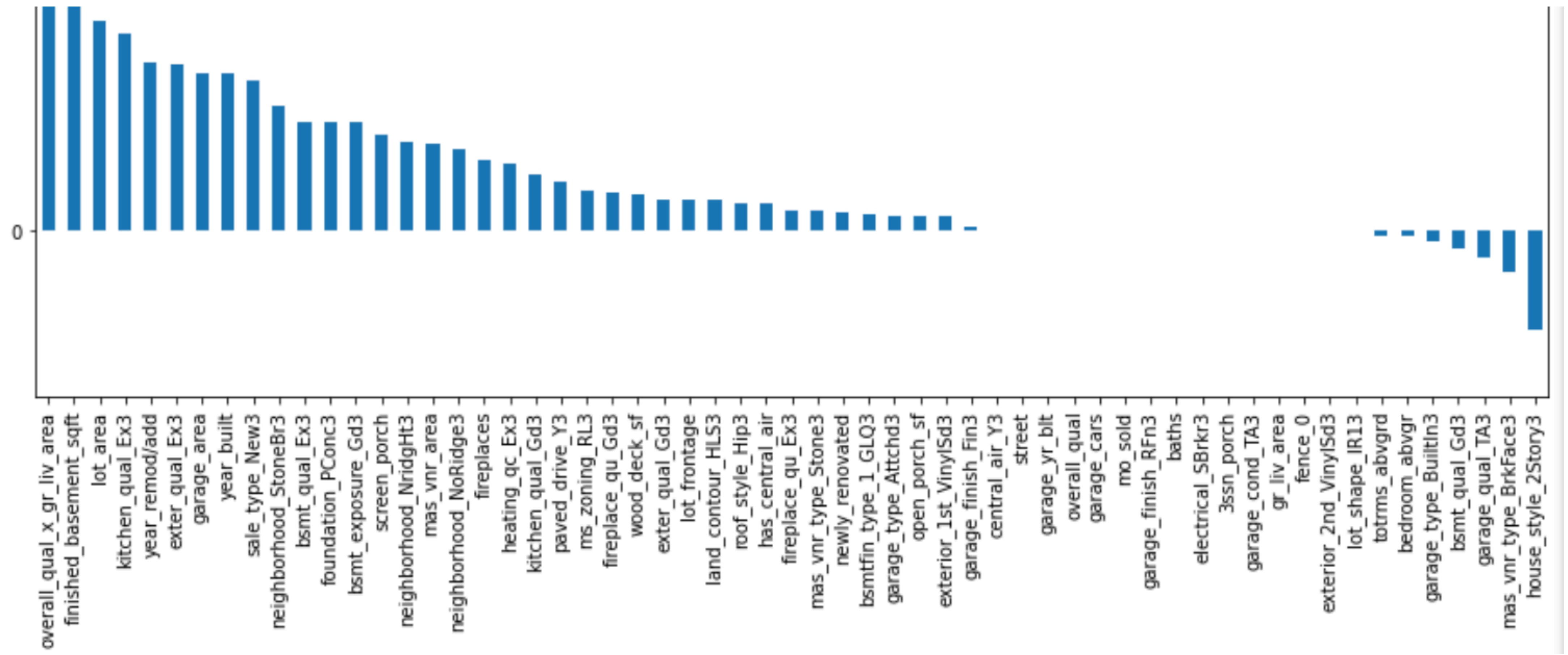
Feature	Correlation w/ sale price
Overall quality	0.8
Living area	0.72
Garage area	0.66
Garage cars	0.65
Baths	0.63
Year Built	0.57
Year remodelled	0.55
Masonry Veneer Area	0.51
Total Rooms above grade	0.51
Fireplaces	0.47



Feature Selection and Data Modelling

How did we select features and decide on the best model?

- After detecting some collinearity in variables, I opted to use an embedded method and let a regularisation model assist in feature selection. LassoCV helped reduce the feature list from 60 to 30.
- Certain categorical features were amplified to boost the signal and reflect better in regression models
- All the final models appeared to be reasonably accurate with low variance, as such ridge model was picked (since coefficients would not be zeroed out)

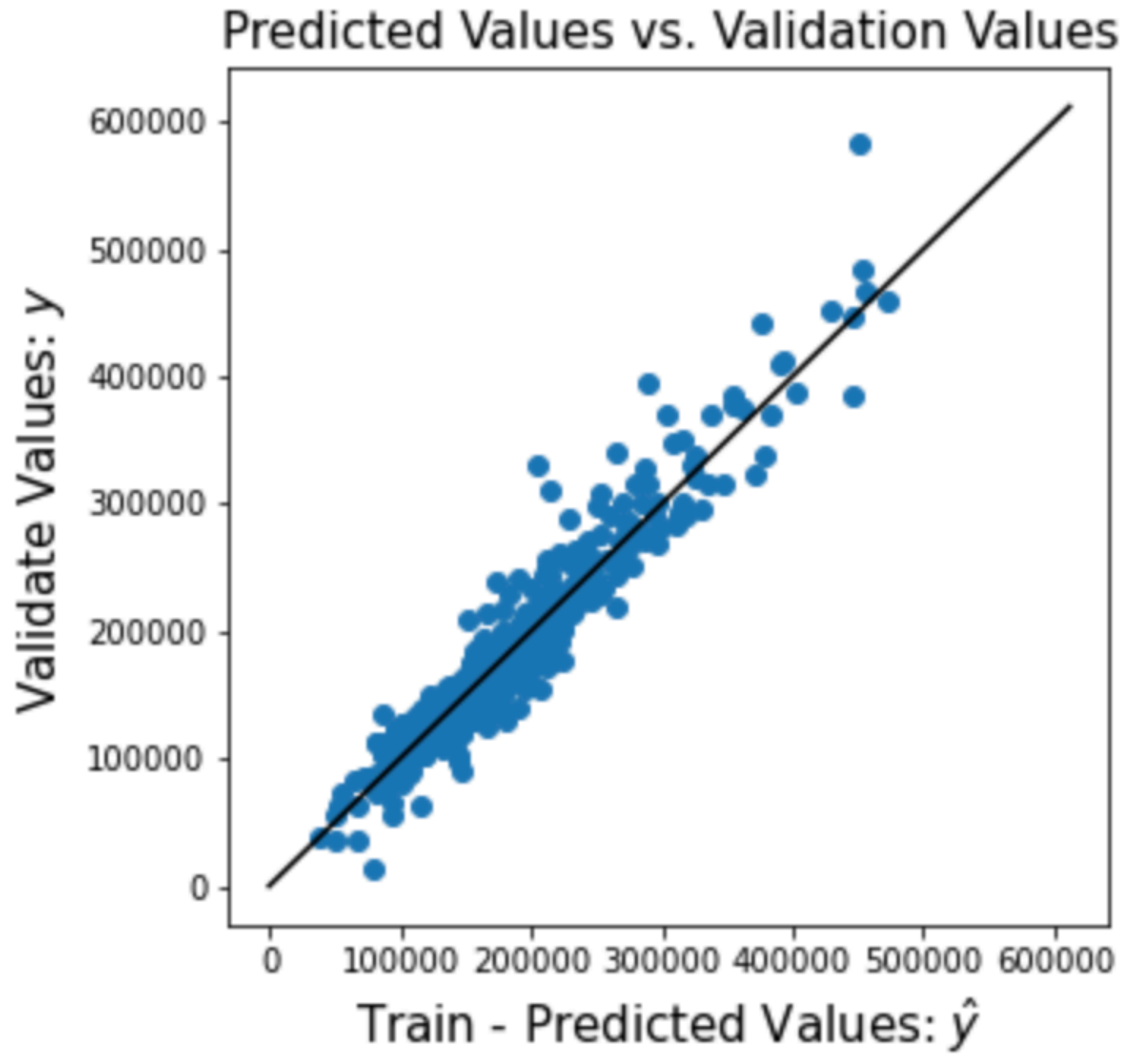


Model	Train RMSE	Validate RMSE
Linear Regression	22588	22565
Lasso	22589	22564
Ridge	22591	22582
ElasticNet	23888	24056

Evaluation of the chosen model

What is it good at and how can it be improved?

- After fitting for optimal alpha, this model is able to **account for approximately 91% of the variation in Sale Price of a property** (adjusted R2 score) and is able to **predict the Sales Price within $24,000** (RMSE)
- Further exploration of the residuals of this model revealed that it is not as accurate at predicting higher values. In the future, perhaps a non-linear model will be a better fit to lower the bias at the higher end of price.
- Caveats and areas of improvement:
    - Limited to Ames, and may not be generalisable to other cities.
    - 2006-2010 is during US subprime crisis, causing property price fluctuations
    - A more robust dataset with buyer demographic information could possibly help us segment buyers to provide more targeted recommendations.



Conclusions

So what have we learned about the impact of features on housing sale prices?

Feature	Impact on sale price in $
overall_qual_x_gr_liv_area	35292.462291
finished_basement_sqft	11308.618639
lot_area	6964.636429
kitchen_qual_Ex3	6808.022352
exter_qual_Ex3	6104.373549
sale_type_New3	5385.419446
garage_area	5225.568410
year_remod/add	5203.524185
bsmt_qual_Ex3	4211.705075
neighborhood_StoneBr3	4208.267459
year_built	4043.423839
bsmt_exposure_Gd3	3746.145623
neighborhood_NridgHt3	3409.039331
foundation_PConc3	3360.449824
screen_porch	3191.665864
neighborhood_NoRidge3	2748.362612
fireplaces	2540.974487
kitchen_qual_Gd3	2461.401288
heating_qc_Ex3	2250.289739
mas_vnr_area	2045.219786
exter_qual_Gd3	1829.767230
paved_drive_Y3	1784.546687
lot_frontage	1681.283121
ms_zoning_RL3	1377.763654
fireplace_qu_Gd3	1349.825052
wood_deck_sf	1146.320313
land_contour_HLS3	1116.848613
roof_style_Hip3	1105.219473
central_air_Y3	987.613536

To make it more actionable for home sellers, I will lump these features into groups

Group	Consists of	Combined Impact
Interaction	• Overall quality + living area	$35,000
Area	• Basement sq footage • Lot area • Garage area • Wood deck sq footage	$27,000
Quality rating	• Kitchen quality • Exterior quality • Basement Quality • Fireplace quality	$25,000
Location	• Residential low-density zone • Northridge • Northridge heights • Stonebrook • Land contour - hillside	$12,000
Home age	• Year of remodelling • Year built	$9,000
Additions	• Fireplaces • Roof style • Central Air	$5,000

Recommendations

How can this analysis help to inform seller decisions?

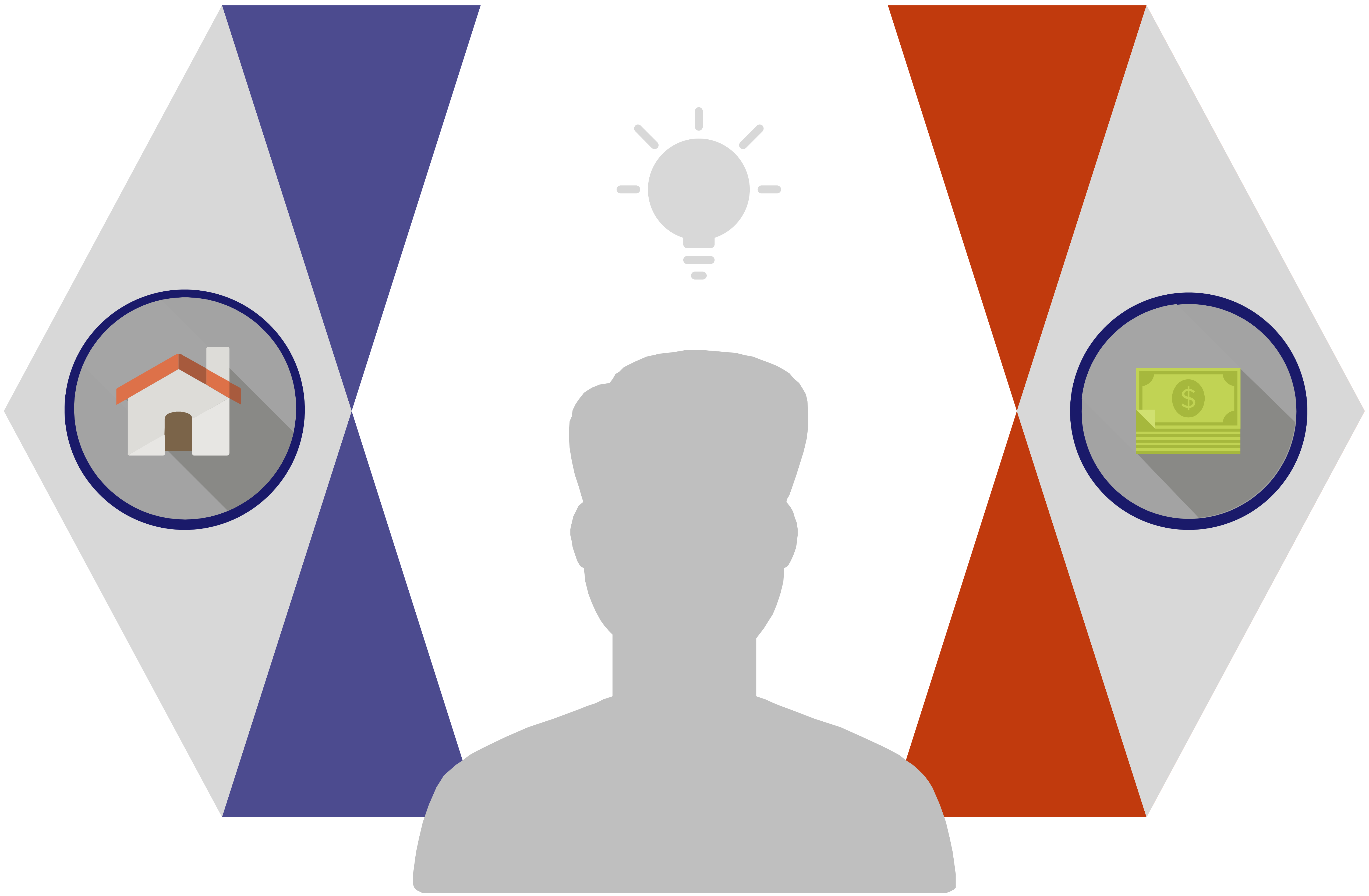
Given that it will be unlikely or extremely difficult to increase any continuous variables (such as lot frontage or square footage), I have decided to base recommendations on 2 groups of categorial variables as these can be changed by sellers.

Quality Ratings

Installation of new fixtures and fittings could lead to an increase in quality ratings, eg-

Having excellent kitchen quality will result in $6773 increase in sale price

Having excellent exterior quality will result in $6055 increase in sale price



Home Additions

Adding new features to your home can also drive up sale price, eg-

Having a paved drive will result in $1964 increase in sale price

Having a hip style roof will result in $1107 increase in sale price