Predicting Housing Prices Ames, Iowa

Tanya Shapiro | Dec 23rd, 2021

What might influence prices?

AGE

SIZE & SPACE

QUALITY & CONDITION

LOCATION

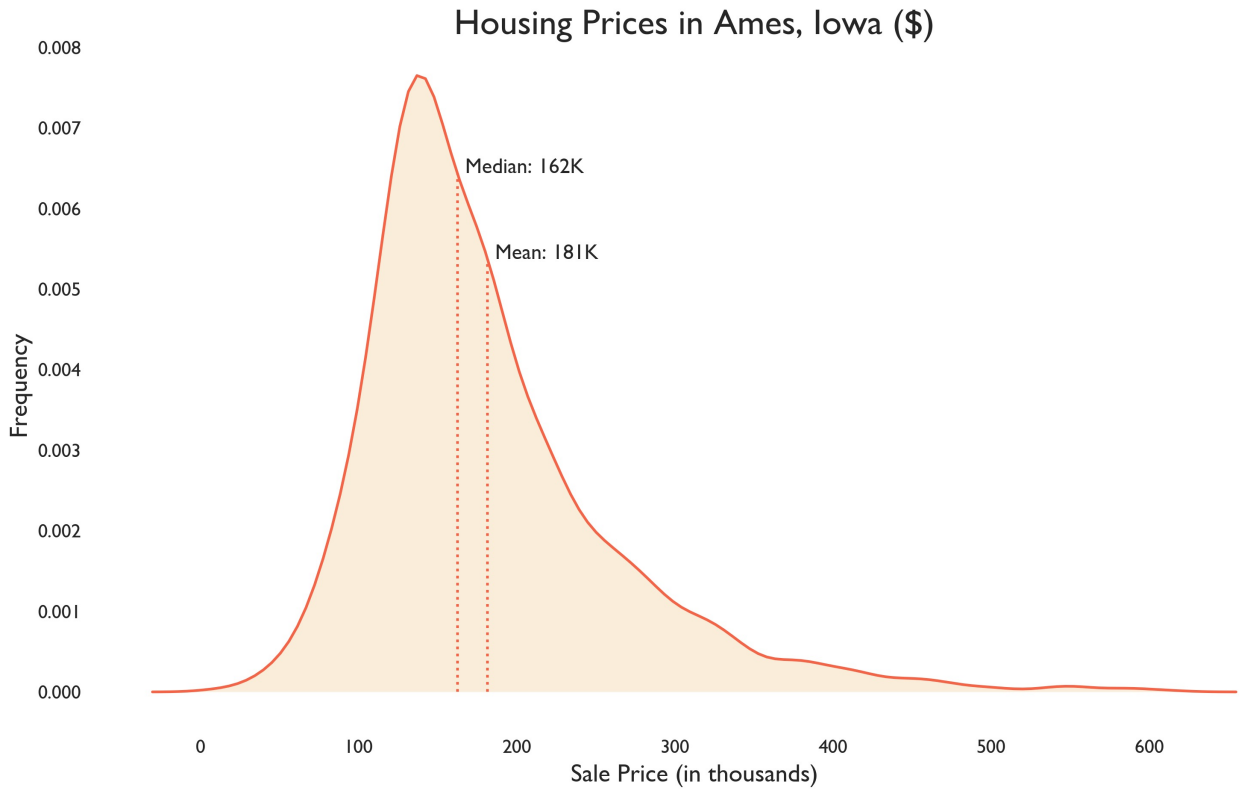
Do we have data to test?

FACTORS	EXAMPLES VARIABLES
Age	Year Built, Year Remodeled, Year Sold
Home Size & Space	Above Grade Ground Living Area, Finished Basement Area, Number of Rooms
Quality & Condition	Overall, Exterior, Basement, Garage, Heating
Location	Neighborhood, MS Zoning

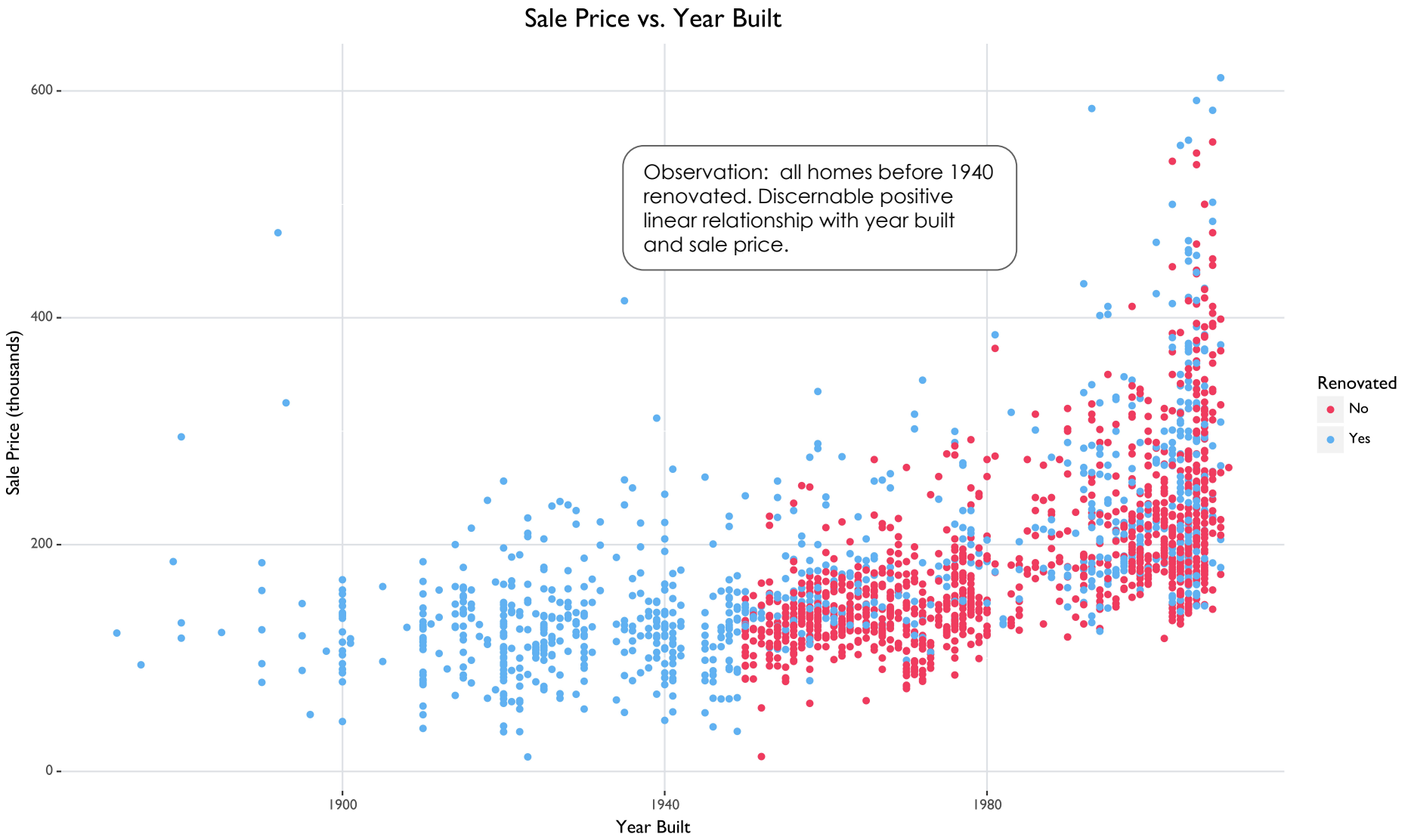
But first - what does the market look like?

Initial Observations:

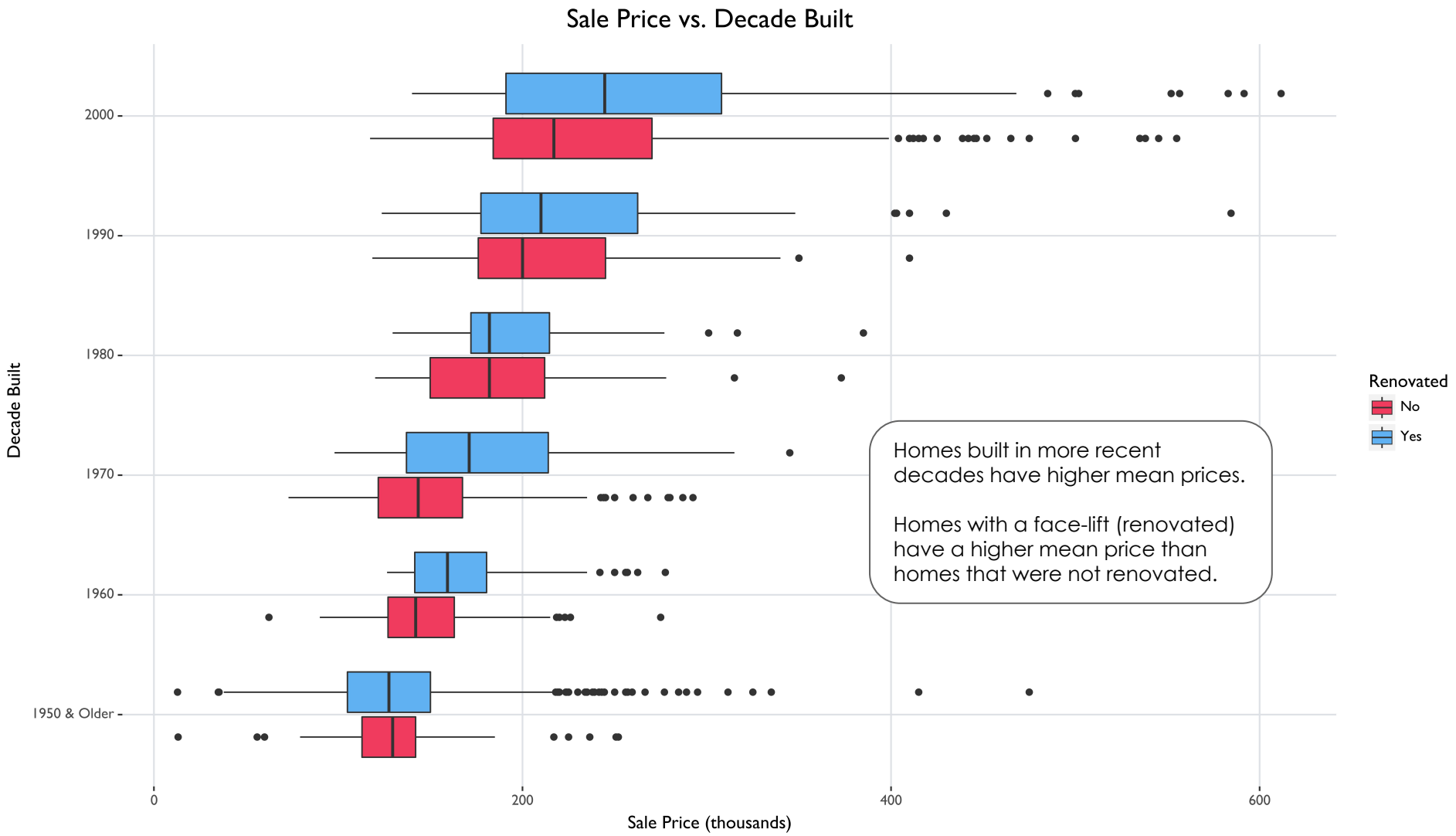
- **Average:**181K
- **Median:** 162K
- There are some big home outliers skewing the data (>400K)



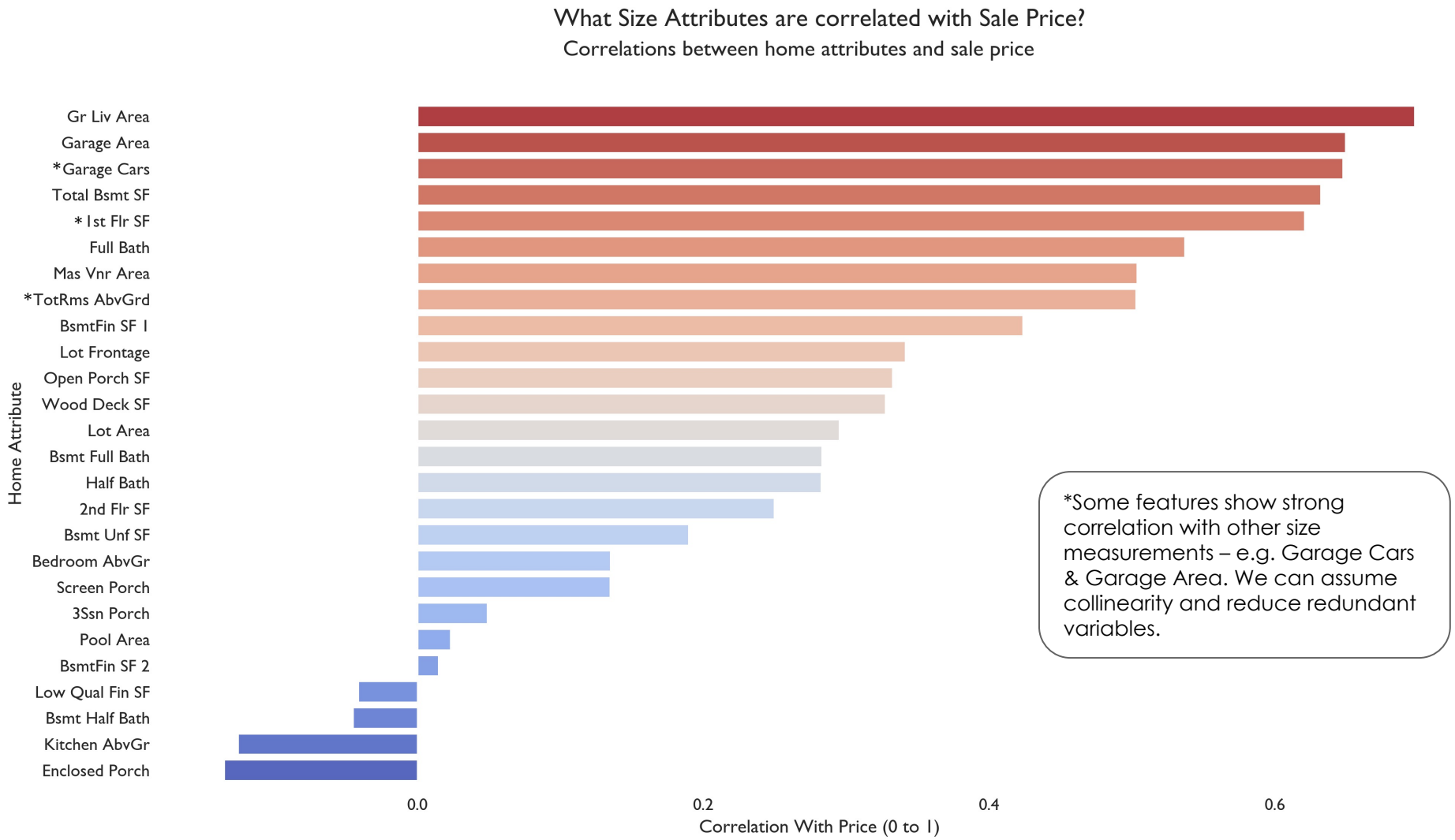
Age: The Newer The Better ($)



Age Cont'd (By Decade)

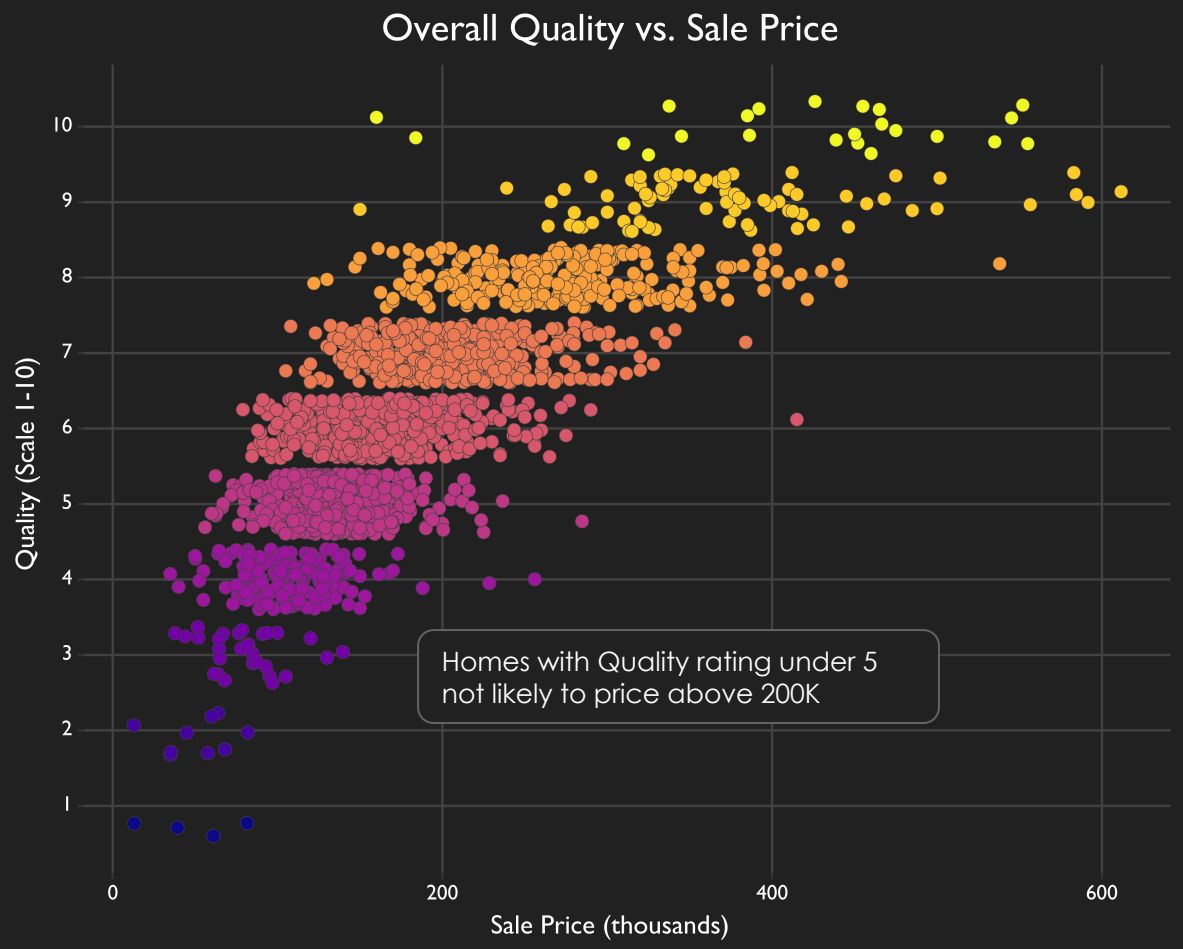


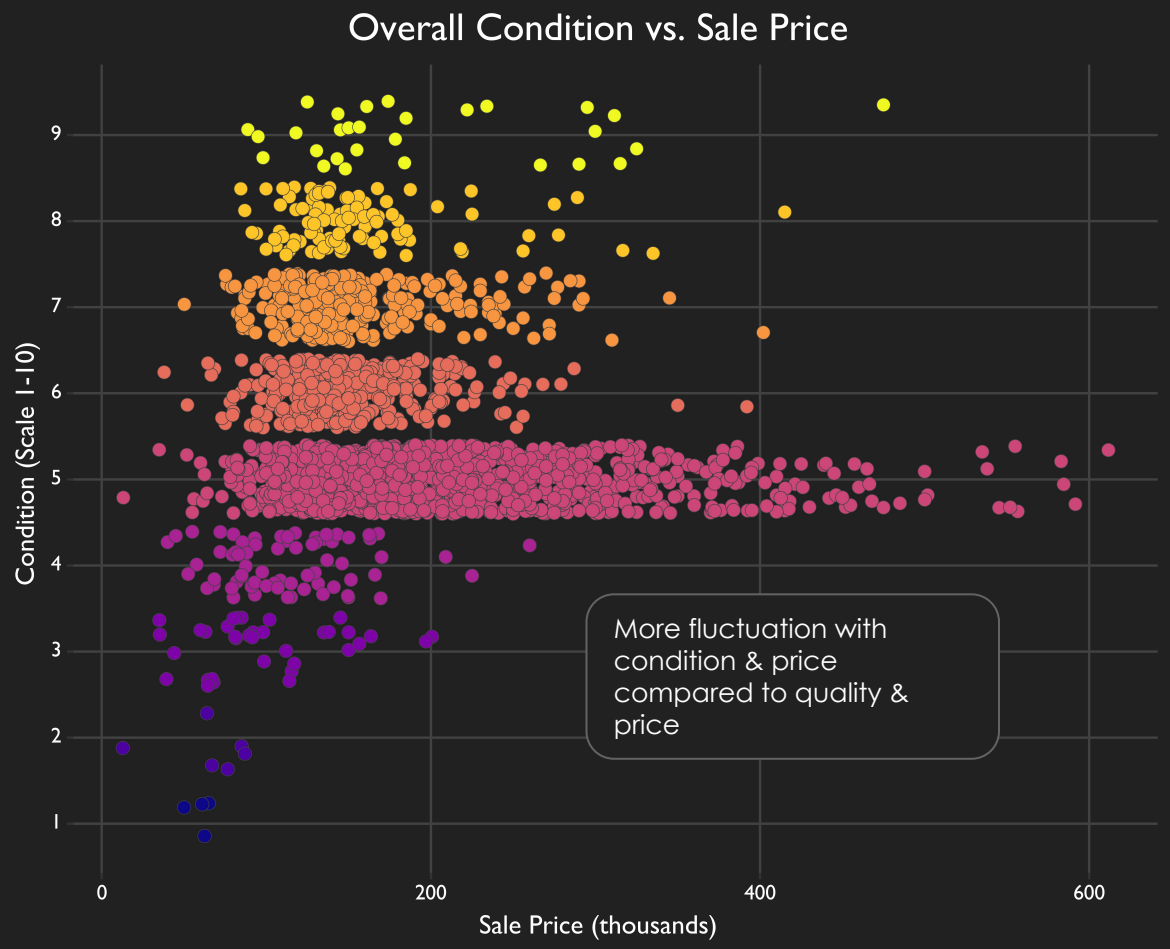
Size: Living, Garage, & Basement SF are important



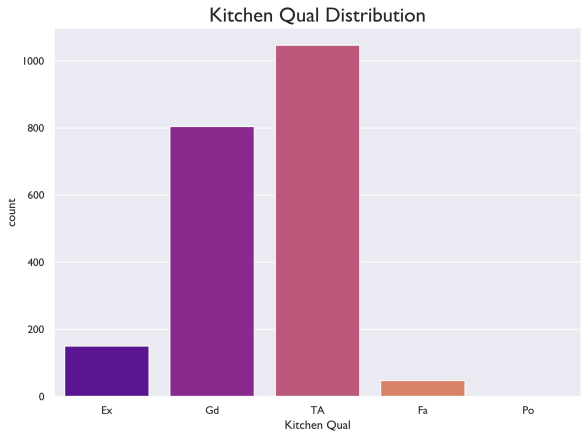
*Some features show strong correlation with other size measurements – e.g. Garage Cars & Garage Area. We can assume collinearity and reduce redundant variables.

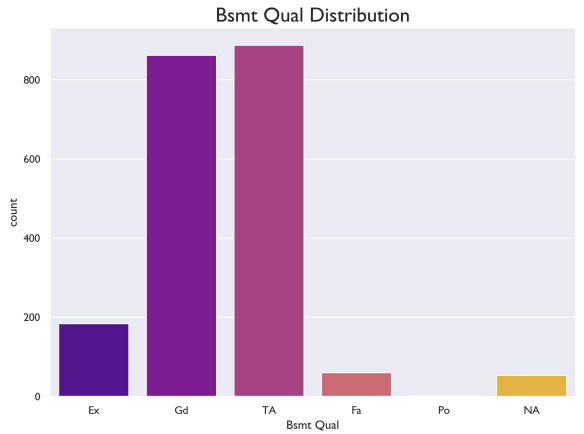
Quality: Positive Relationship with Price

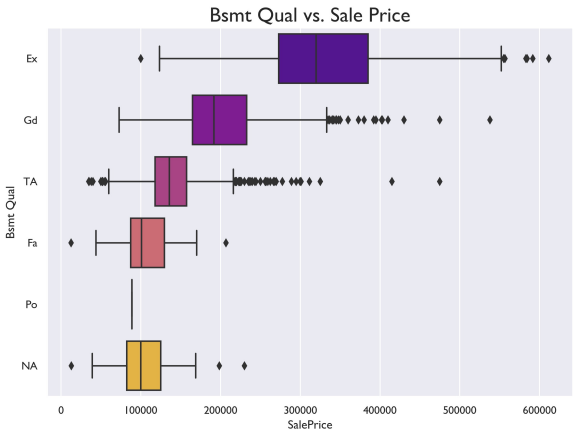


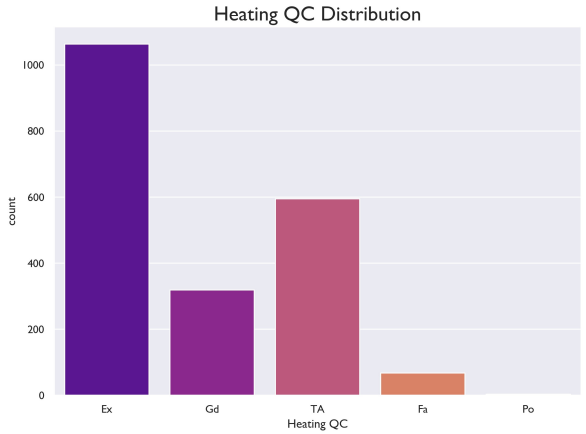


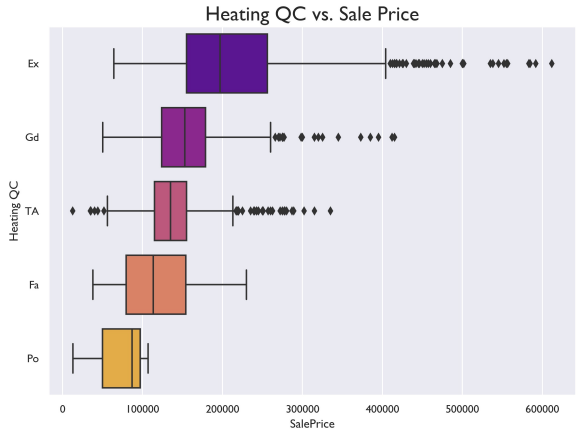
Pricing Discrepancies based on Other Quality Measures

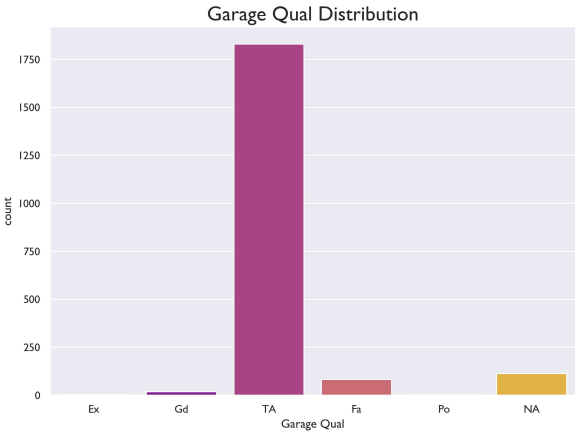


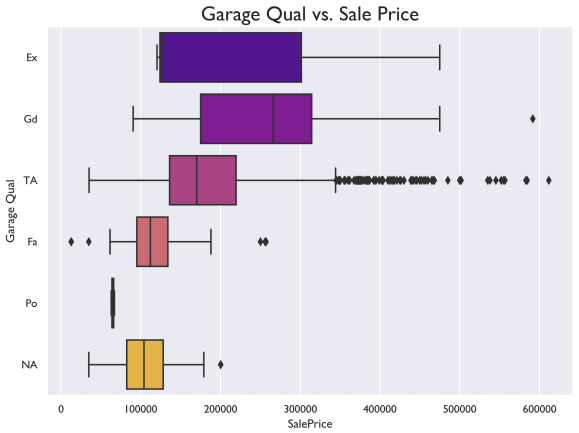






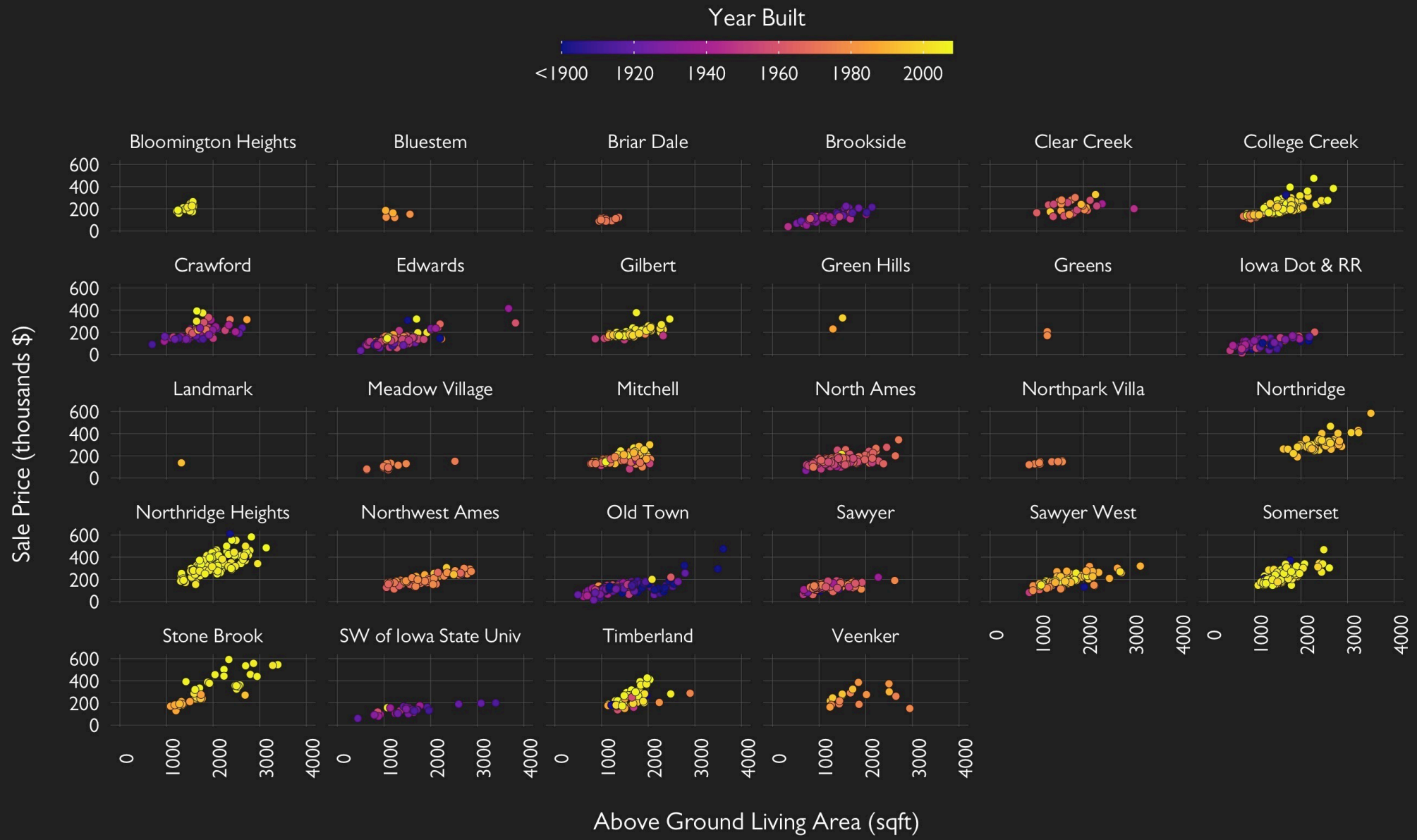






All quality features worth considering…except Garage

Neighborhood: Size & Price



Methodology

- **Outliers**. Some mega mansion homes > 4000 sqft
- **Missing Values.** Replaced appropriate "0" or "None" or mean based on other attributes.
- **Feature Selection.** Selected based on correlation to price, removed variables with high collinearity.
- **Categorical Values.** Used pandas "get dummies."
- **Models Applied**. Linear, Ridge, and Lasso.

Desc	Train	Test	RMSE
1 - LR Numeric Variables Only	0.826	0.842	31,568
2 - LR Numeric & Categorical Variables	0.897	0.901	24,959
3 - LR Overall Qual as Dummy Variable	0.903	0.906	24,361
4 - LR Overall Qual Dummy & Sqft Outliers	0.865	0.882	26,882
5 - Ridge	0.903	0.907	24,275
6 - LassoCV	0.903	0.908	24,148
7 - Ridge with Alpha 26 (GridSearch Tuning)	0.903	0.908	24,129

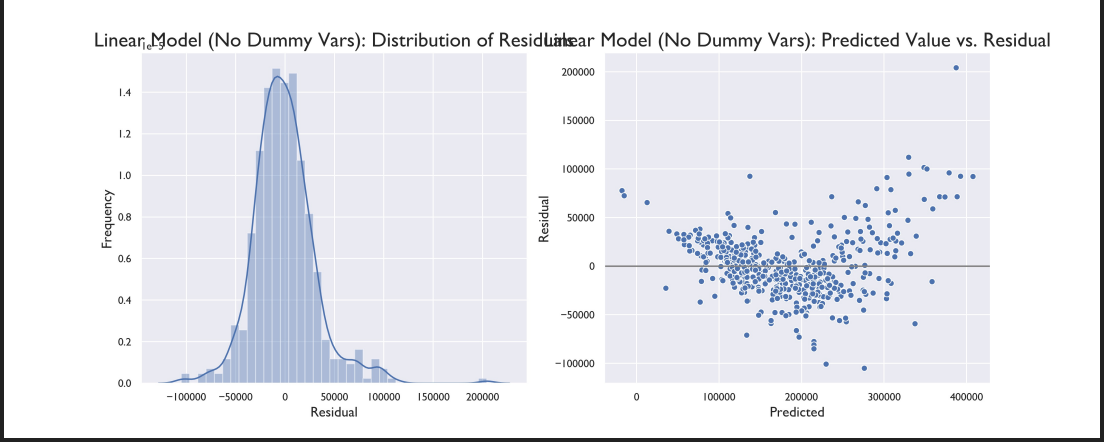
Model Results

Model Ingredients

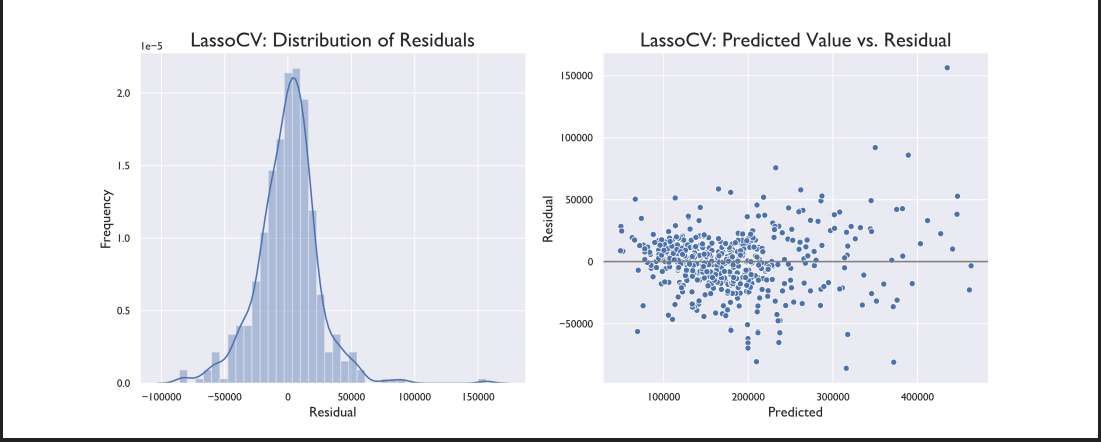
Numeric Variables: Year Built, Year Remod/Add, Above Ground Living Area, Garage Area, Total Basement Sqft, Overall Condition

Categorical Variables: MS Zoning, Neighborhood, Basement Quality, Overall Quality, Garage Quality, Heating Quality, Exterior Quality, Foundation, Masonry Veneer Type

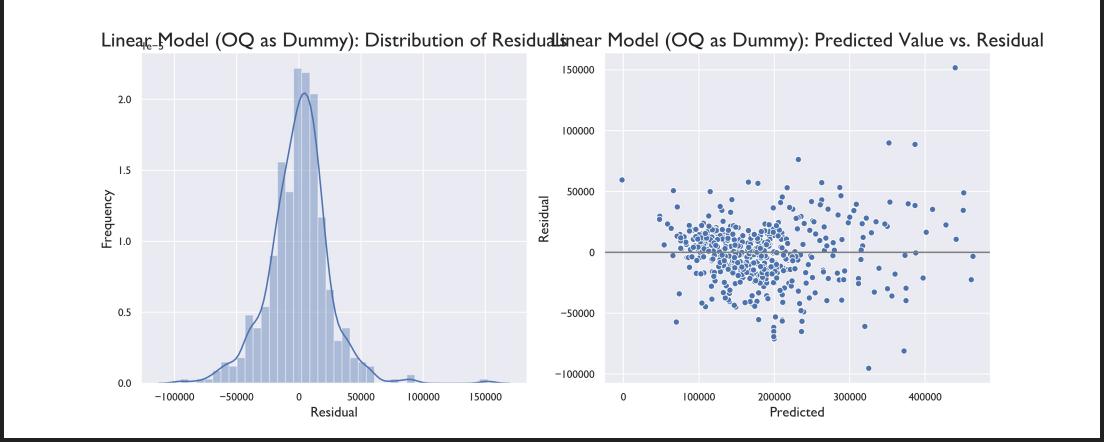
Linear (No Dummy Var)



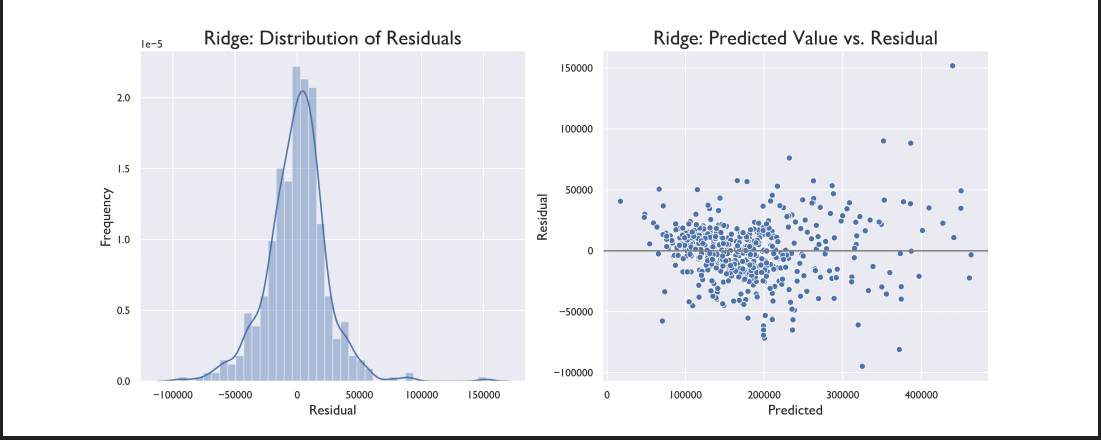
LassoCV



Linear (With Dummy – Overall Qual)



Ridge



Conclusions & Considerations

- **Model is weak at predicting expensive homes** - from residual analysis, there are some extreme outliers. Most likely due to right-skewed nature of sale price distribution.
- **Feature Engineering** - there may be unseen variables that have a relationship with price or variables that can be factored together to produce a better variable. For instance, distance to Iowa State University may have an impact on price, future research could create a mean distance based on Neighborhood.
- **Explore Ordinal Variables** - ordinal variables may not be evenly weighted, e.g. "Excellent" can be 2pts higher than "Good." What gets the best score? Dummified or numeric?
- **Consider Other Models** - regression model may be too simple for predicting house price. Future research should test other supervised learning models, e.g. Decision Tree.
- I**nferences** – homeowners should consider renovation and maintenance of their home with time – improving quality and condition could increase value (sale price)