Kaggle Competition - Ames, Iowa Housing Prices

Colin Mondi

Agenda

01	Problem Statement
02	Profiling Efforts
03	Analysis Findings
04	Data Cleansing / Pre-Processing
05	Model Development / Testing
06	Conclusion

Overview

Problem Statement: Iowa State University, located in Ames Iowa, is interested in expanding their campus. Due to most of the surrounding land not being up for sale, the University must determine the predicted sales price in order to provide the best offers for property owners.

Approach:

1. Retrieved and profiled *Ames Housing Data*
2. Identified key patterns/trends in the data
3. Cleansed and regenerated a model-ready version of the dataset
4. Developed, tested, and altered multiple different types of regression models to predict house sales prices
5. Identified the appropriate model with the most accurate price predictions

Profiling Efforts

High-Level Summary

Profiling Findings

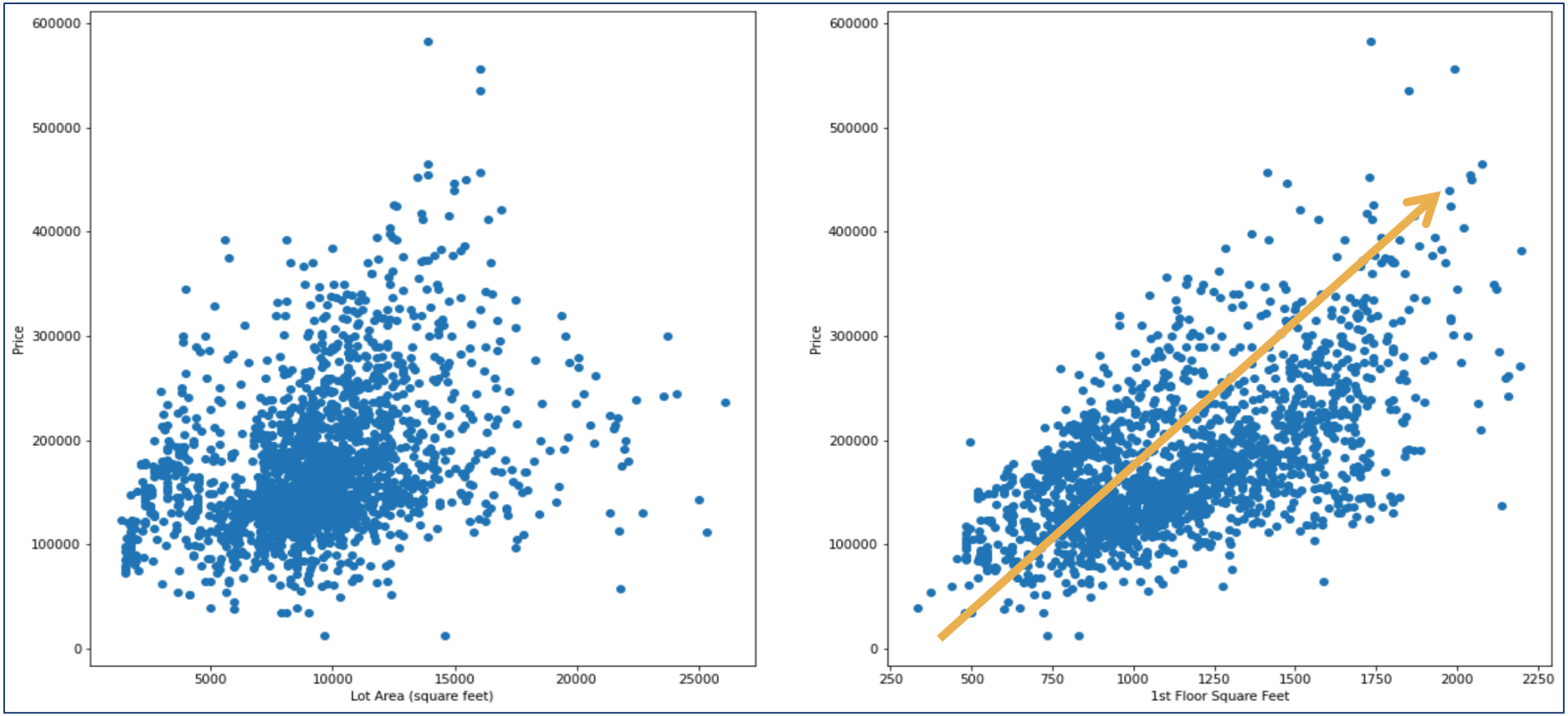
- *Data-specific statistics:*
  - ~80 columns, ~2,000 rows
  - Majority of data present; minimal or no instances of NULLs across columns
  - Mix of integer and object data types
- The average house price is ~$180,000 dollars
- 100+ years of sales prices available for homes
- Neighborhoods including *Brookside* and *Old Town* have the lowest average housing prices
- Neighborhoods including *Northridge Heights* and *Northridge* have the highest average housing prices

Key Columns

- The following columns were further investigated based off of the correlation exercises:
  - ***garage_finish - garage_area***
  - ***year_built***
  - ***gr_liv_area***
  - ***1st_flr_sf***
  - ***overall_qual***

Analysis - Square Footage

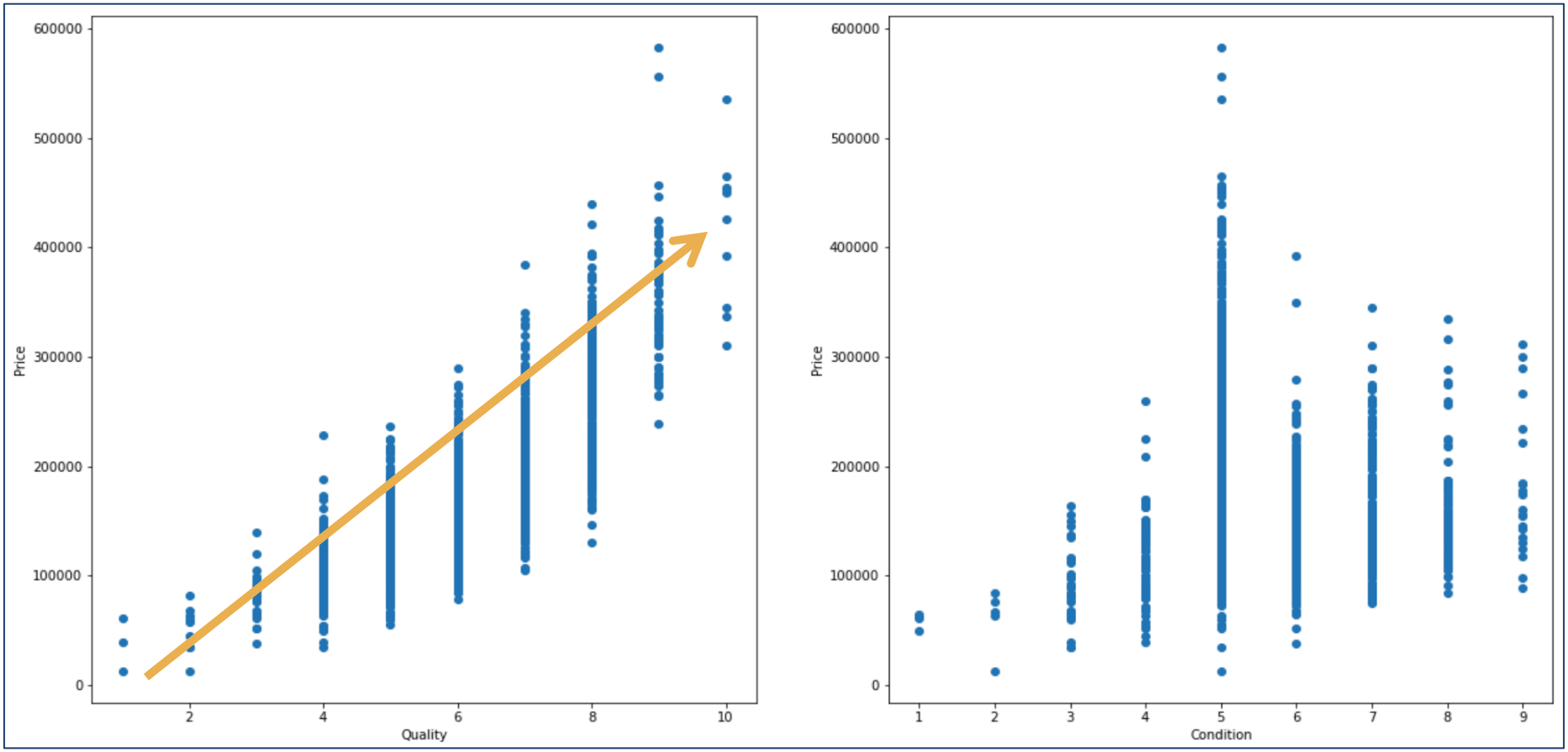
The first-floor square footage of a property building has a bigger impact on sales price in comparison to the square footage of the entire lot.



The university is planning on building new facilities on these lots. Just looking at these two variables alone, the ideal property would 1) have a small house and 2) a large lot

Analysis - Quality & Condition

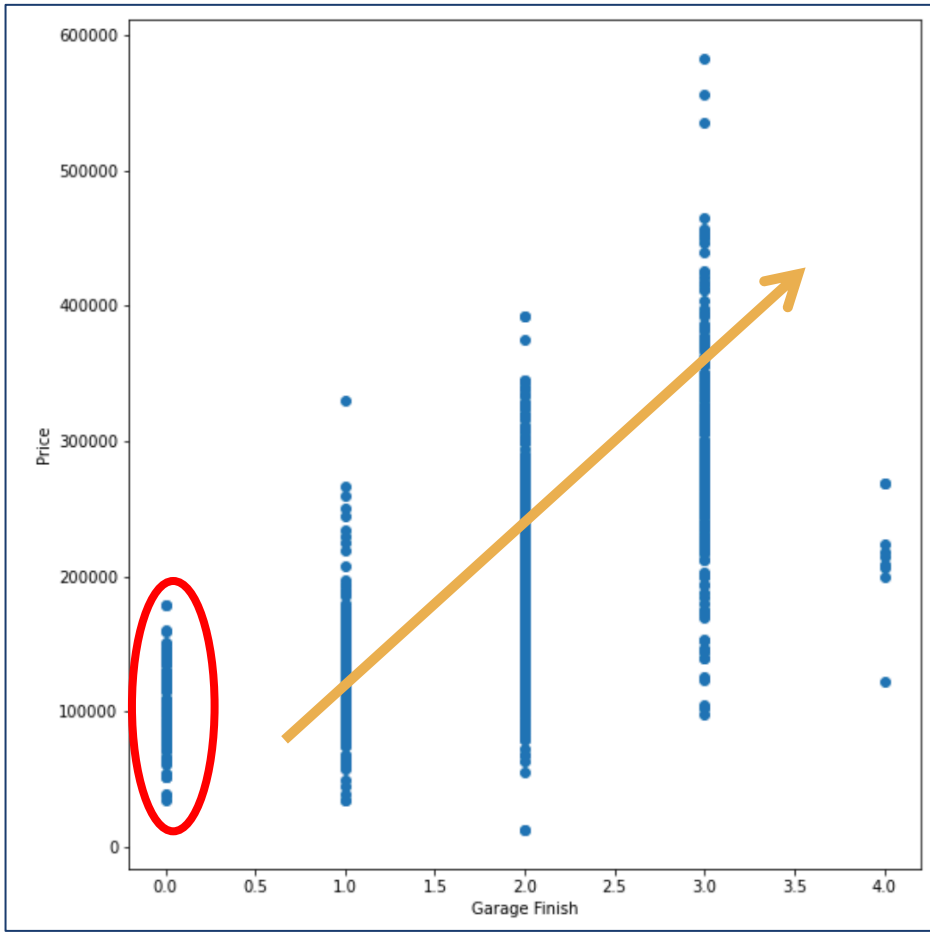
The overall quality is heavily correlated with sales price while the overall condition shows almost no relationship with sales price.

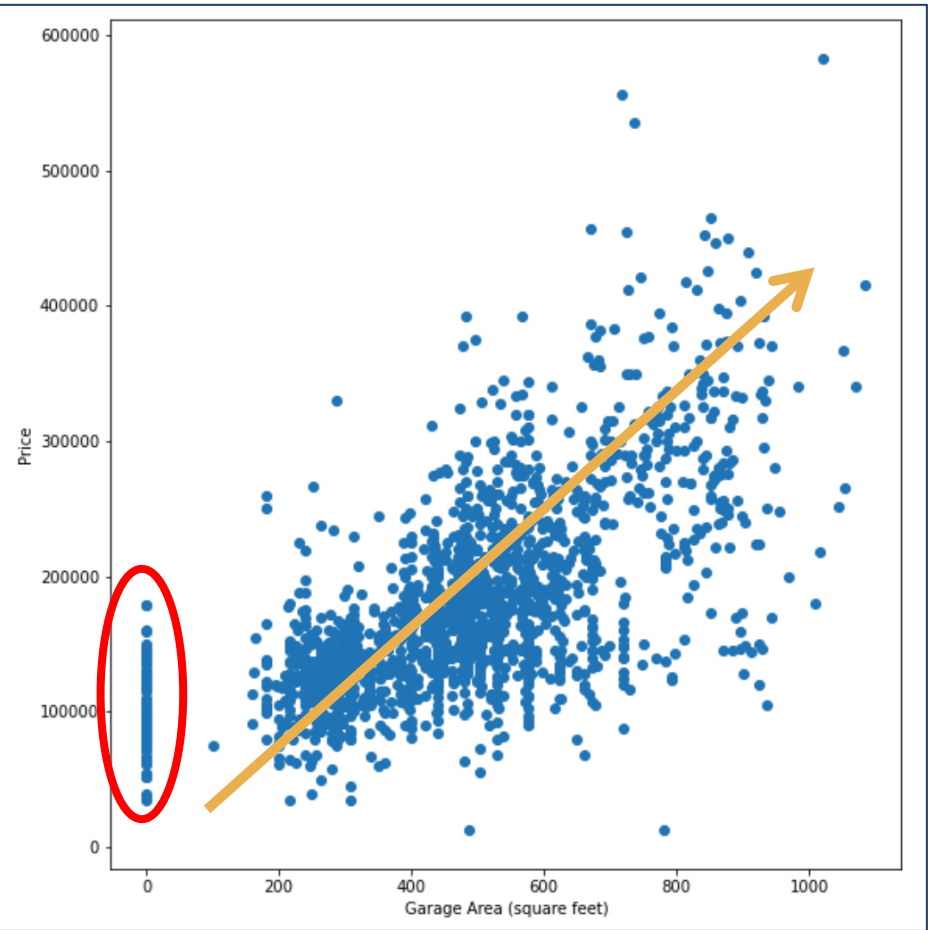


Unexpected difference in correlation with sales price for condition versus quality; assumed these two would show similar correlations with price

Analysis - Garage

Both the garage area and garage finish show a linear relationship with sales price, but there are some potential outliers at play here for garage area.

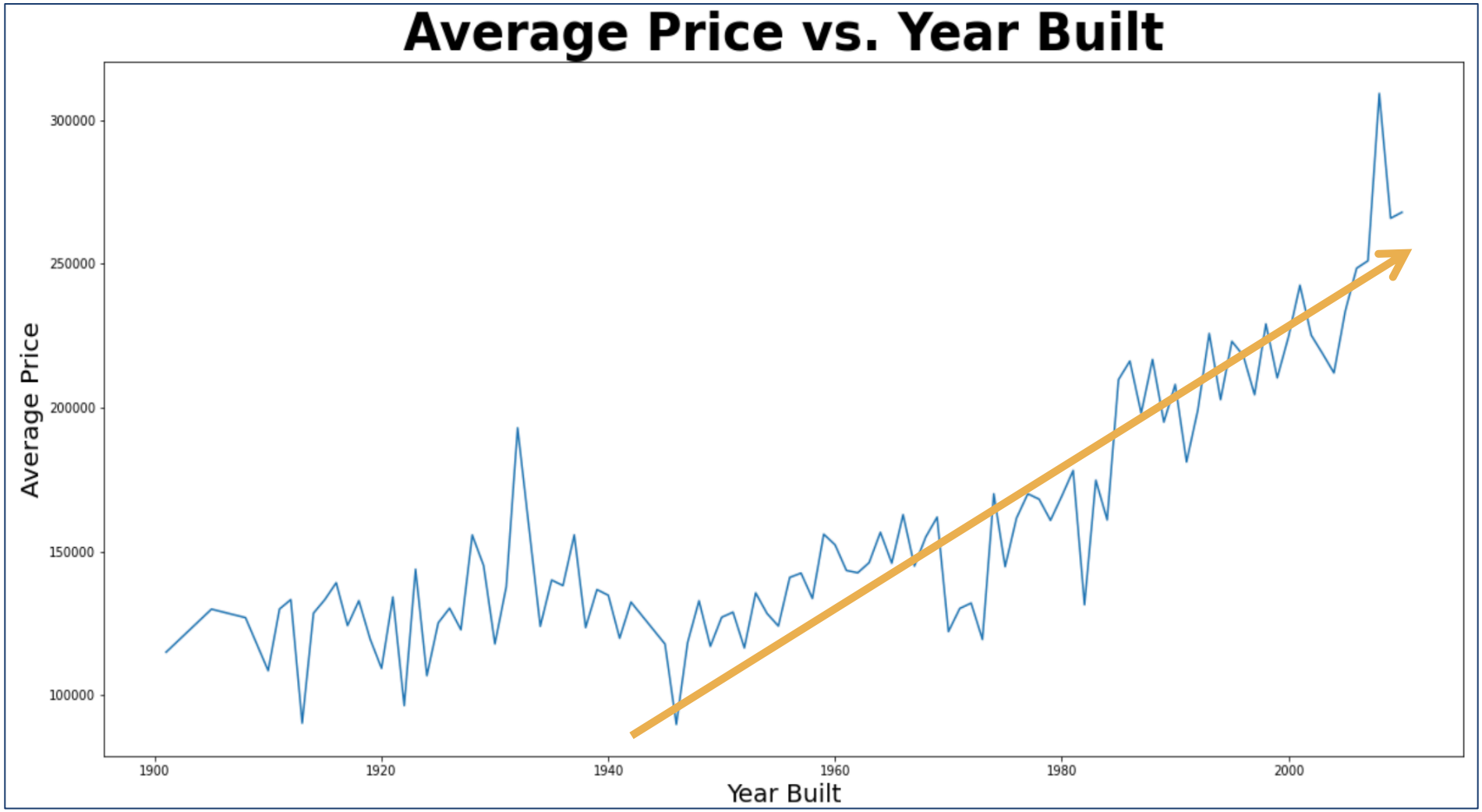




Unexpected high count of houses with no garage (circled in red above)

Analysis - Change in House Prices YoY

There's an expected increase in the change in average housing price YoY in recent years, but prior to ~1940 this wasn't necessarily the case.



Data Cleansing / Pre-Processing

	Actions
Data Cleansing	▪ Removed columns with 1,000+ NULL values; ***5 total*** ▪ Replaced remaining NULL values with either ***1) mean of column values*** or ***2) 0*** ▪ Removed outliers for key model variables; ***7 total*** ▪ Added two new categorial columns for groupings; ***lot size & year built*** ▪ One-Hot Encoded ordinal-type columns; ***18 total***
Pre-Processing	1. Multiple iterations of feature lists established for model testing ▪ Started with all variables positively correlated with sales price; modified from there 2. Standardized data 3. Created dummy columns ▪ 20 object-type variables

*After model testing was complete, the finalized feature list included **199** total variables*

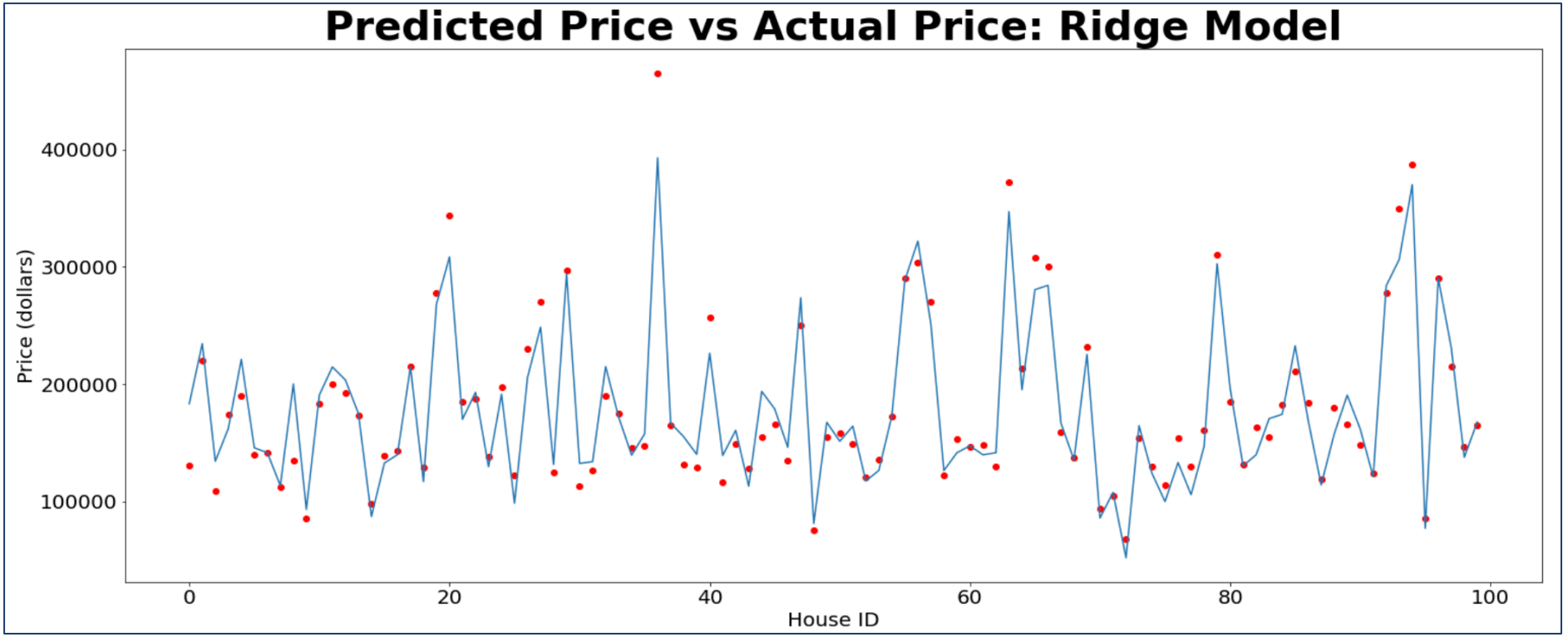
*NOTE*** - dummy columns reflect majority of these*

Model Development / Testing

Process

Model Testing

- Applied the same data cleansing & pre-processing actions to the test data that was originally made to the train data
- Tested 4 different model types; **Linear Regression, Ridge, Lasso, Elastic Net**
- Utilized GridSearch to view how each individual variable was performing; adjusted feature list (in 'Pre-Processing') using these results



Model Results

Ridge	▪ **R^2 score:** 0.91 ▪ **RMSE:** 22,501 ▪ **Alpha:** 5.09
Elastic Net	▪ **R^2 score:** 0.88 ▪ **RMSE:** 23,215 ▪ **Alpha:** 1.0
Lasso	▪ **R^2 score:** 0.91 ▪ **RMSE:** 22,487 ▪ **Alpha:** 41.32
Linear Regression	▪ **R^2 score:** 0.88 ▪ **RMSE:** 22,527 ▪ **Alpha:** n/a

Summary Findings

- The **lasso** & **ridge** models performed best; comparing the R^2 & RMSE values, they're almost identical
  - Alpha values show the only difference; *5.09* for ridge, *41.32* for lasso
- The **elastic net** & **linear regression** models didn't perform as well as lasso / ridge

Conclusions

Recommendation:
*After cleansing, pre-processing, and prepping all 4 models, the **ridge model** is the best predictor of housing prices for Ames, Iowa.*

Although the lasso and ridge model had very similar R^2 & RSME values, the ridge model will generally perform better when the prediction is a function of multiple variables.

Source