Project 2: Ames Housing Data

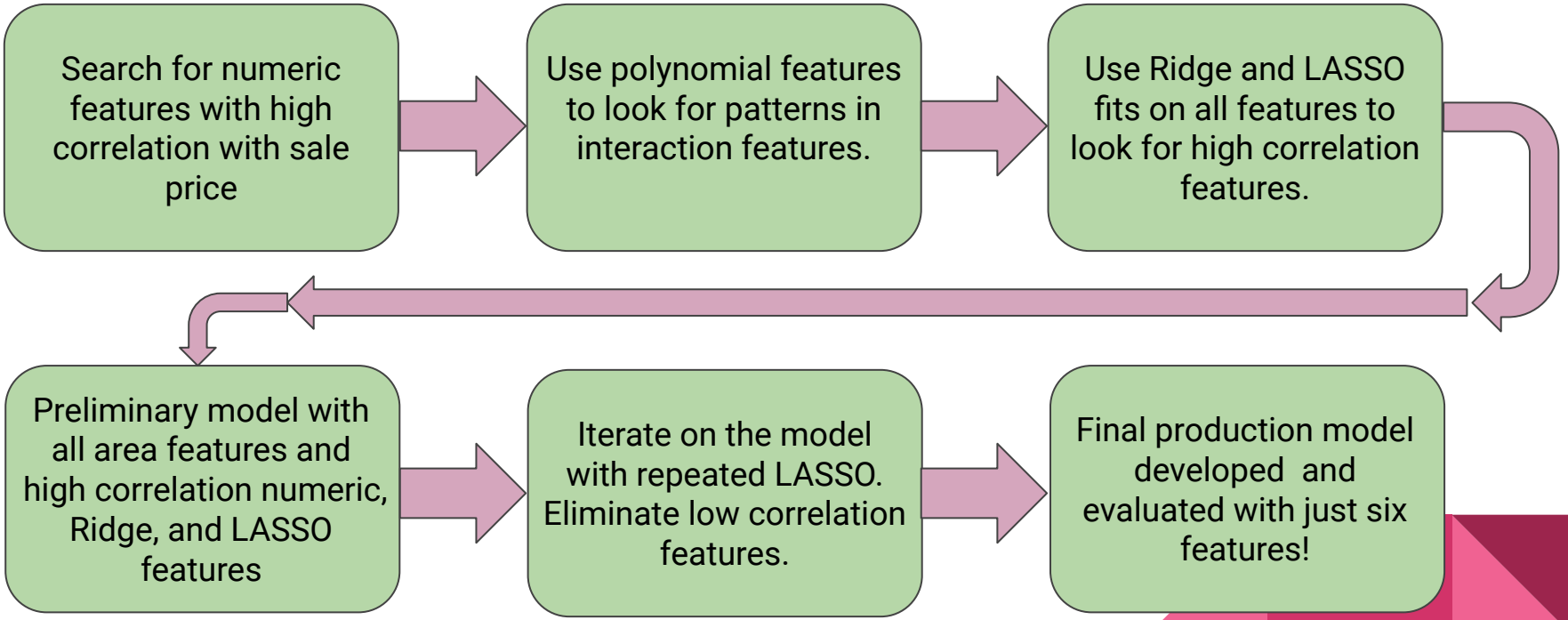
By Joey Notaro

Problem Statement

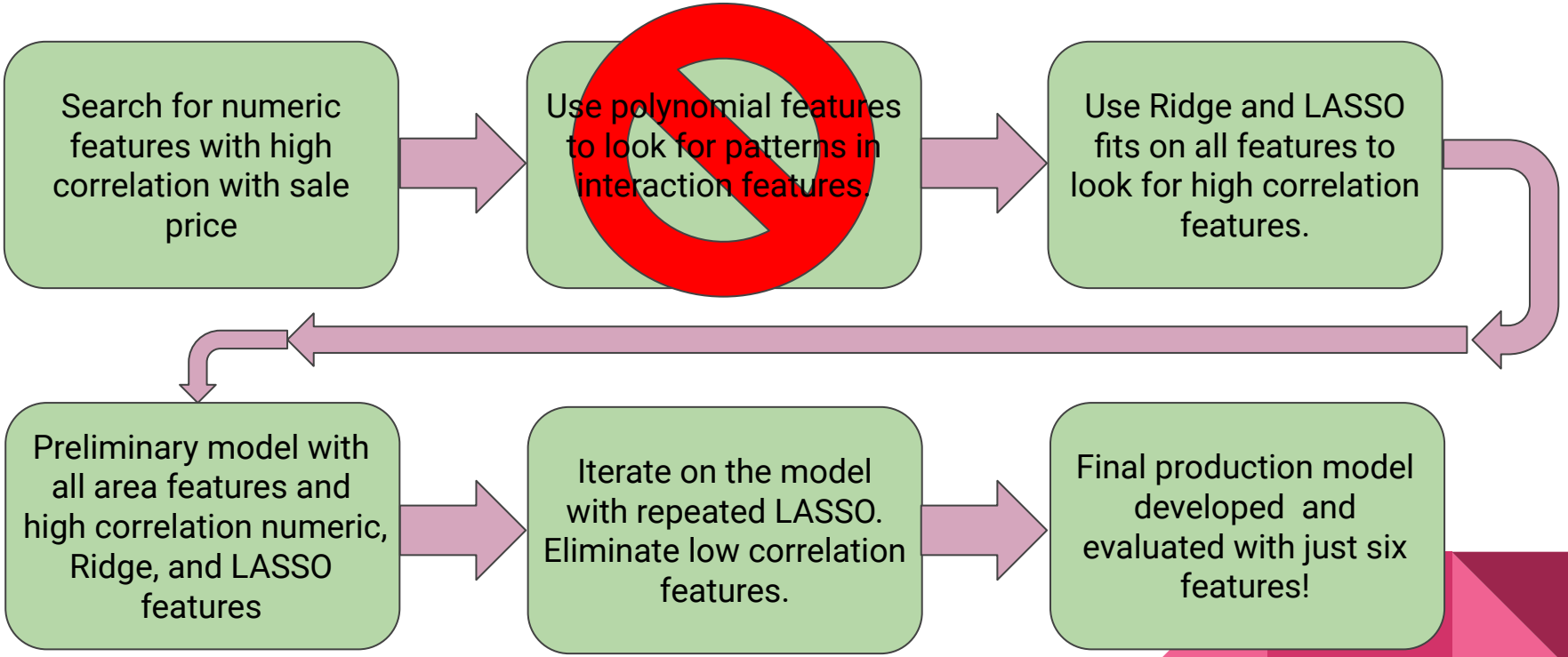
- An architectural firm in Ames, IA is interested how to best design home for better sales prices.

- *For different features of new home construction, which **features' square footage** has the greatest predictive power in determining the home's sale price? Based on the model developed, how should these features be **minimized or maximized** to raise the overall sales price?*

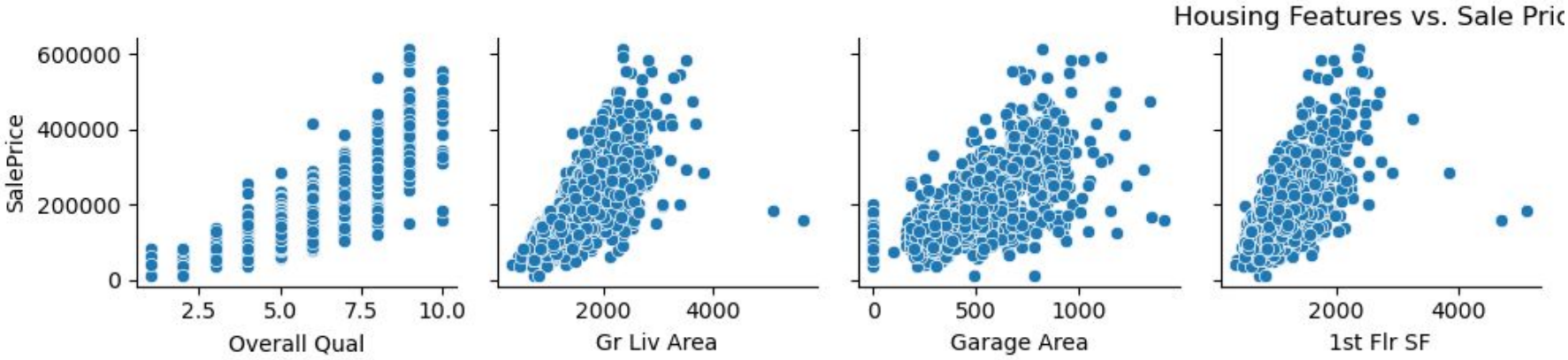
Methodology



Methodology



Promising Numeric Features



Correlation Coefficient			
~0.80	~0.70	~0.65	~0.62

Preprocessing and Feature Selection

- Encoded all categorical features as numeric (285 features in total!)
- Tried polynomial features to scan for interaction patterns. (Not helpful!)
- Returning to 285 features, fitted Ridge and LASSO regularization after standard scaling all features.

Ridge and LASSO Fits

Ridge	
Optimal Alpha	~599
Training Score	~0.90
Test Score	~0.89

LASSO	
Optimal Alpha	~673
Training Score	~0.91
Test Score	~0.91

Preliminary Model

- Identify all area or square footage features from data dictionary (most included in highly correlated numeric features.)
- Take top five high magnitude correlation features from both Ridge and LASSO fits.
- Narrow from 285 to 27 features.

Model Tuning

- Well fit but too many area and extraneous features.
- Standard Scale and LASSO fit five times, eliminating features with low magnitude correlation coefficients.

Preliminary Model	
Cross-Validation Score	~0.80
Training Score	~0.85
Test Score	~0.87

Production Model

- From 27 to 6 features:
  - Overall Quality
  - Living Area
  - Year Built
  - Basement Finished Square Footage
  - Garage Area
  - Lot Area

Production Model	
Cross-Validation Score	~0.75
Training Score	~0.78
Test Score	~0.84

Feature Loadings

	Coefficient (St. Devs.)	Standard Deviation	Coefficient (Units)
Overall Qual	35143.95846840979	1.4262705424542637	24640.46
Year Built	10528.397421552243	30.177889068708122	348.88
Gr Liv Area	22999.960314463213	500.4478294652996	45.96
Garage Area	8824.518776844956	216.1351017782383	40.83
BsmtFin SF 1	10116.465289422347	461.19504113765447	21.94
Lot Area	6770.160990820356	6742.488909496557	1.0

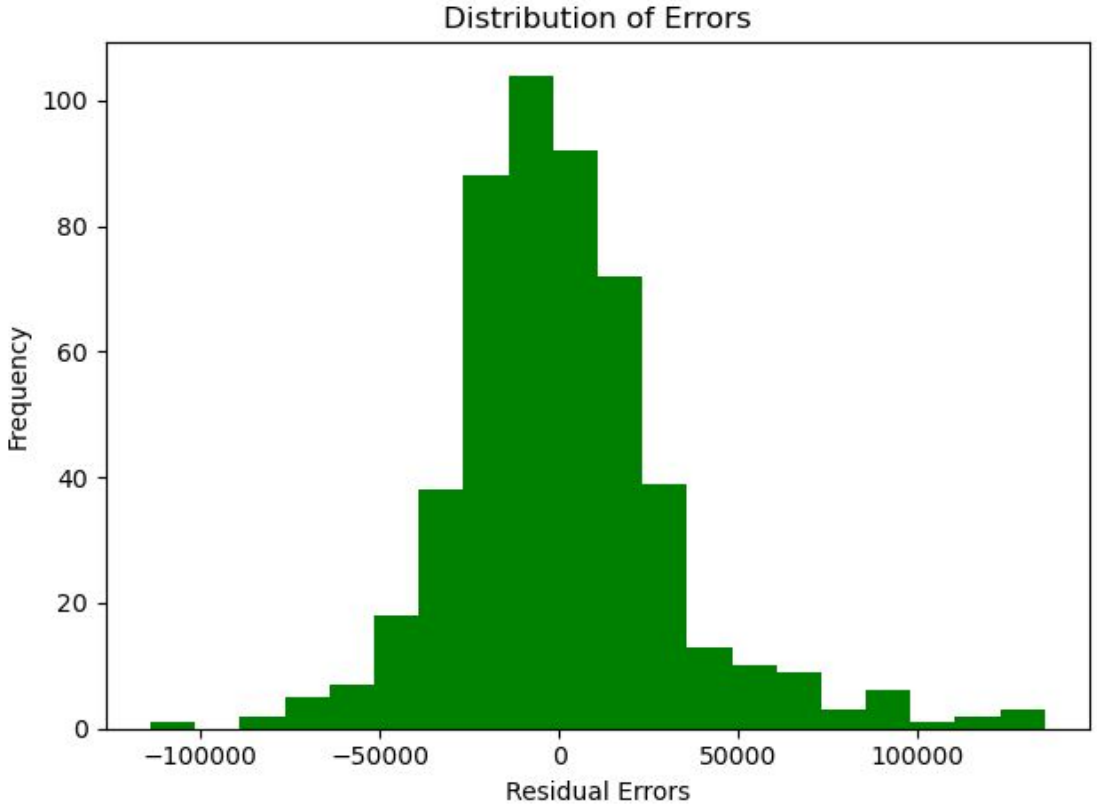
Feature Loadings

- Holding all else constant, this is the dollar increase in sale price for every 1 square foot increase in feature area.

	Coefficient (Units)
Overall Qual	24640.46
Year Built	348.88
Gr Liv Area	45.96
Garage Area	40.83
BsmtFin SF 1	21.94
Lot Area	1.0

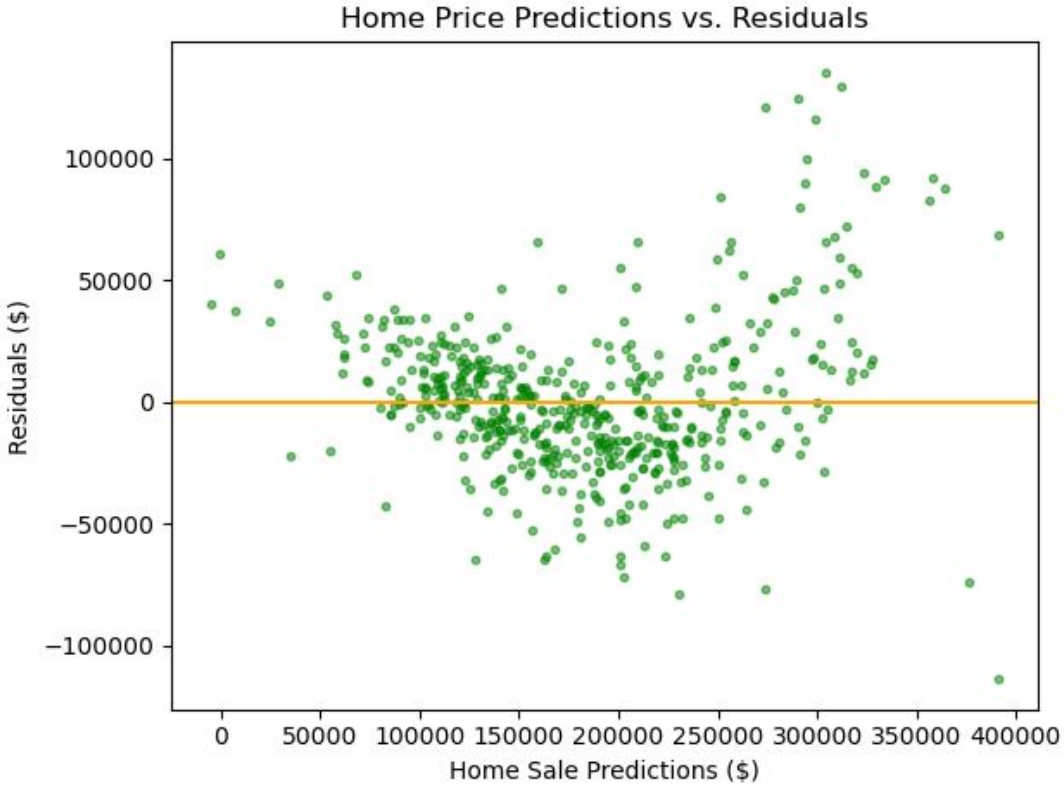
Evaluation of Assumptions

- Errors are normally distributed and centered around zero.



Evaluation of Assumptions

- Some heteroscedasticity, so model may violate some assumptions of randomness of errors.
- Could take better account of outliers.



Finding #1

- Holding all else constant, a one square foot increase in the above ground living area is correlated with a $45.96 increase in the sale price.
- Holding all else constant, a one square foot increase in the garage area is correlated with a $40.83 increase in the sale price.
- Holding all else constant, a one square foot increase in the finished basement is correlated with a $21.94 increase in the sale price.
- Holding all else constant, a one square foot increase in the lot area is correlated with a $1.00 increase in the sale price.

Business Recommendation #1

- Design homes with larger living areas and garages.
- If a home is designed with a basement, make sure that the design is a finished basement.
- Lot area increase value but only marginally. Focus on the home design.

Finding #2

- Quality materials matter. Holding all else constant, an increase of one unit of quality on a scale of 1-10 correlated with a sale price increase of nearly $25,000.

- The year built matters as well, but we can hold it constant as this is all new home construction.

Business Recommendation #2

- Partner with construction company and contractors who will use high-quality materials within budget.

- Kitchen quality was one of the last quality features to get eliminated during model iteration, so focus on quality kitchen materials when budget is a constraint.

Finding #3

- Outdoor features matter less.

- All porches were eliminated in model iteration early on except screen porches.

- Pool area actually had a negative correlation with sale price.

Business Recommendation #3

- Do not focus on designing outdoor features.
- If a porch is desired, design a screen porch.
- Pools may not be worth it.

Finding #4

- Design inside a space is more important than the size of the space on its own.

- Total basement square footage, and first and second floor square footage were outperformed by the individual area features in the production model.

Business Recommendation #4

- A second floor may not add much value to a home.
- A larger living area is more valuable than a larger first floor.
- A finished basement is more valuable than just having a larger basement.