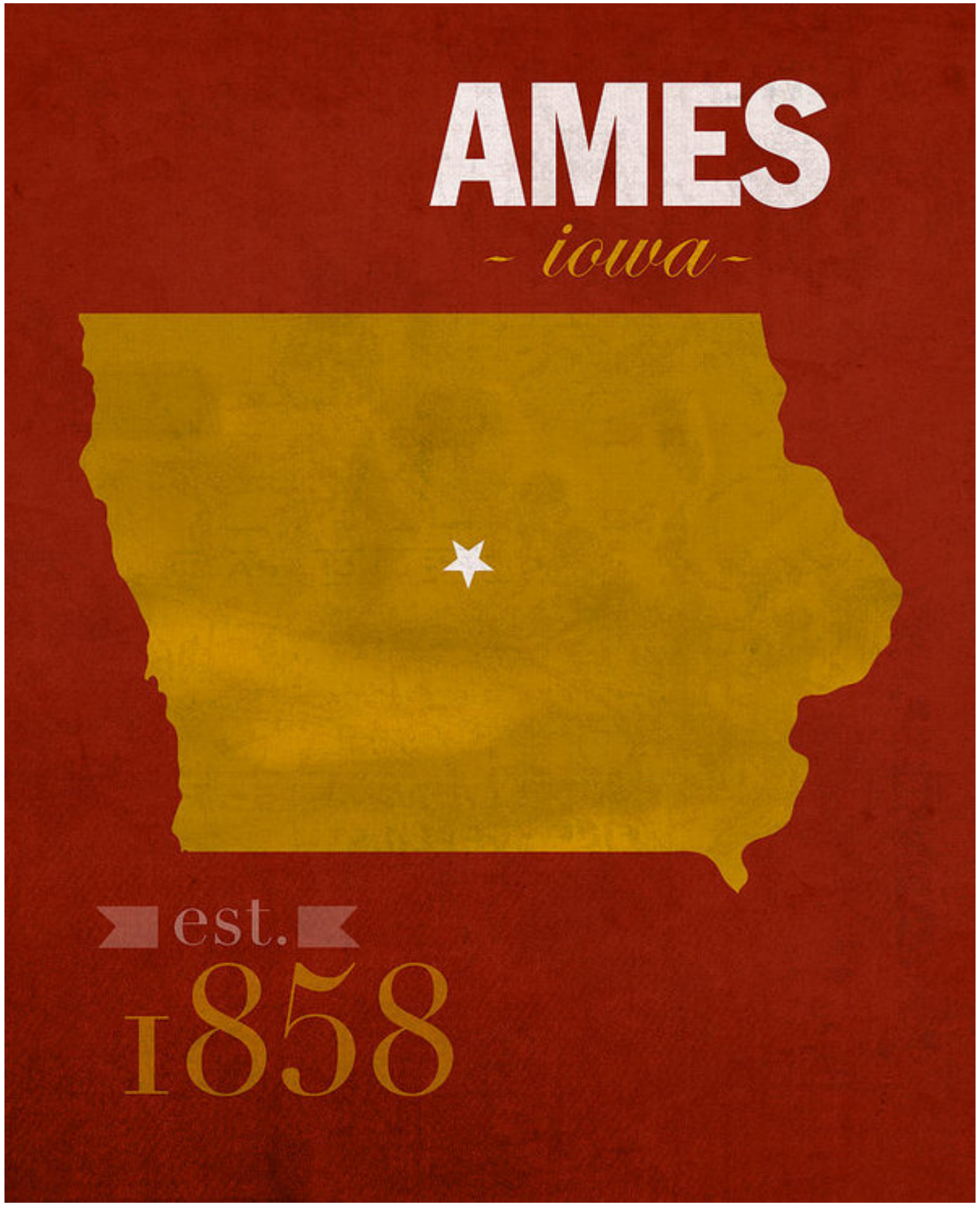
My Presentation of Project 2

Dylan Lunde

What’s our Data?

Ames, Iowa Housing Data!

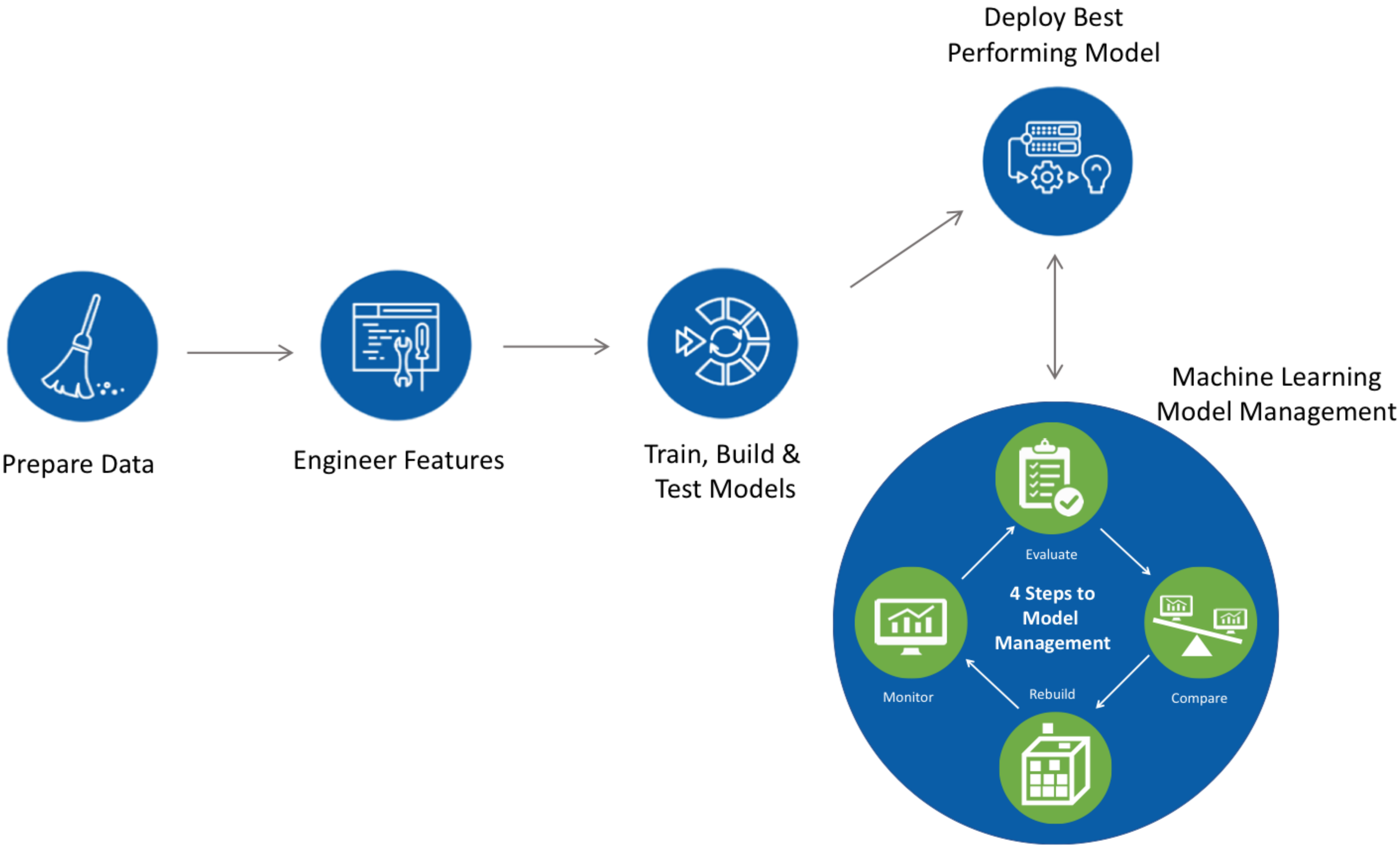


What’s our problem?

We want to predict the sale price of a home!

How do we do it?!

We make and train a model using data!


Deploy Best Performing Model
Prepare Data
Engineer Features
Train, Build & Test Models
Machine Learning Model Management
Evaluate
4 Steps to Model Management
Monitor
Rebuild
Compare

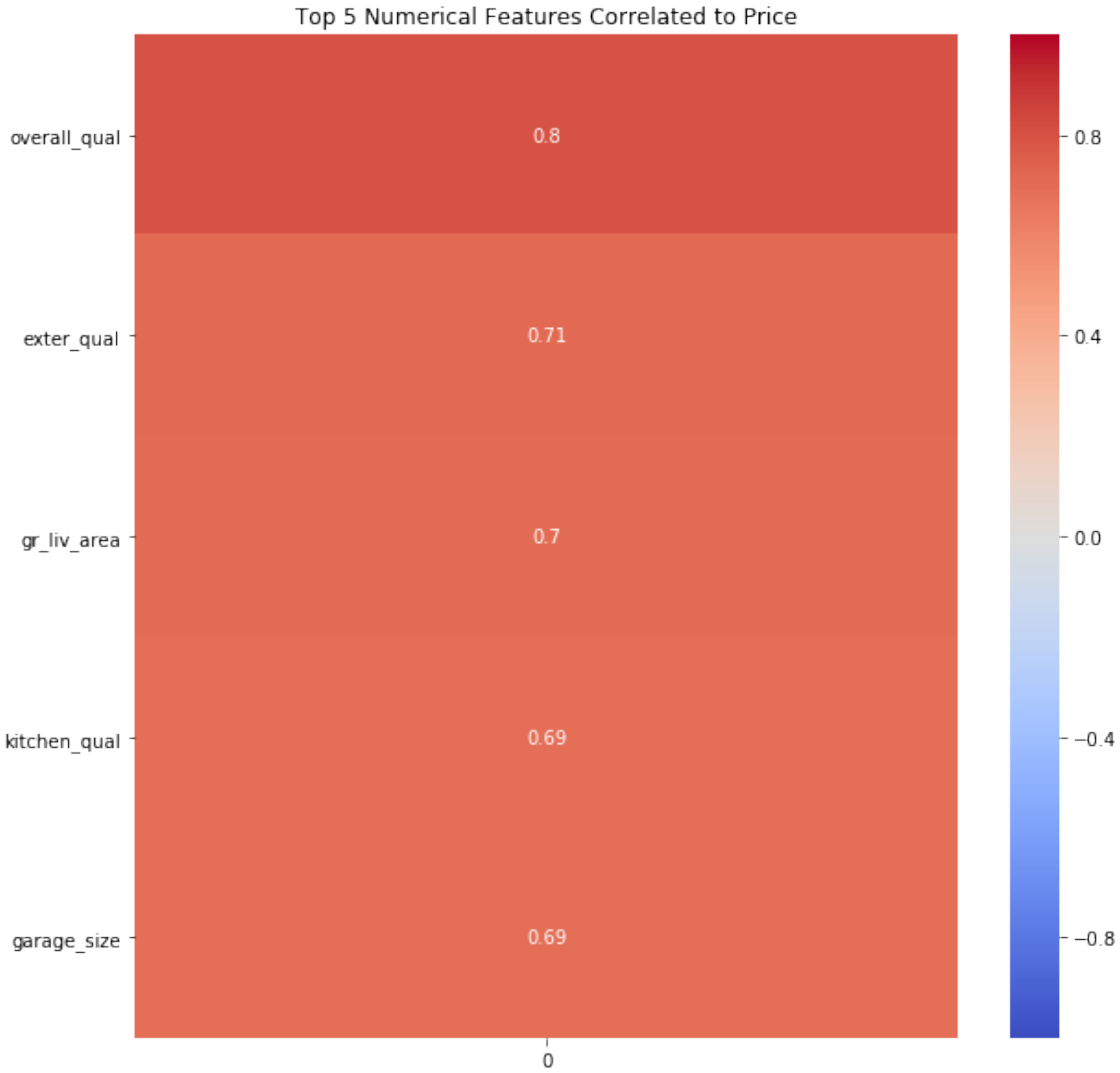
Initial Exploration

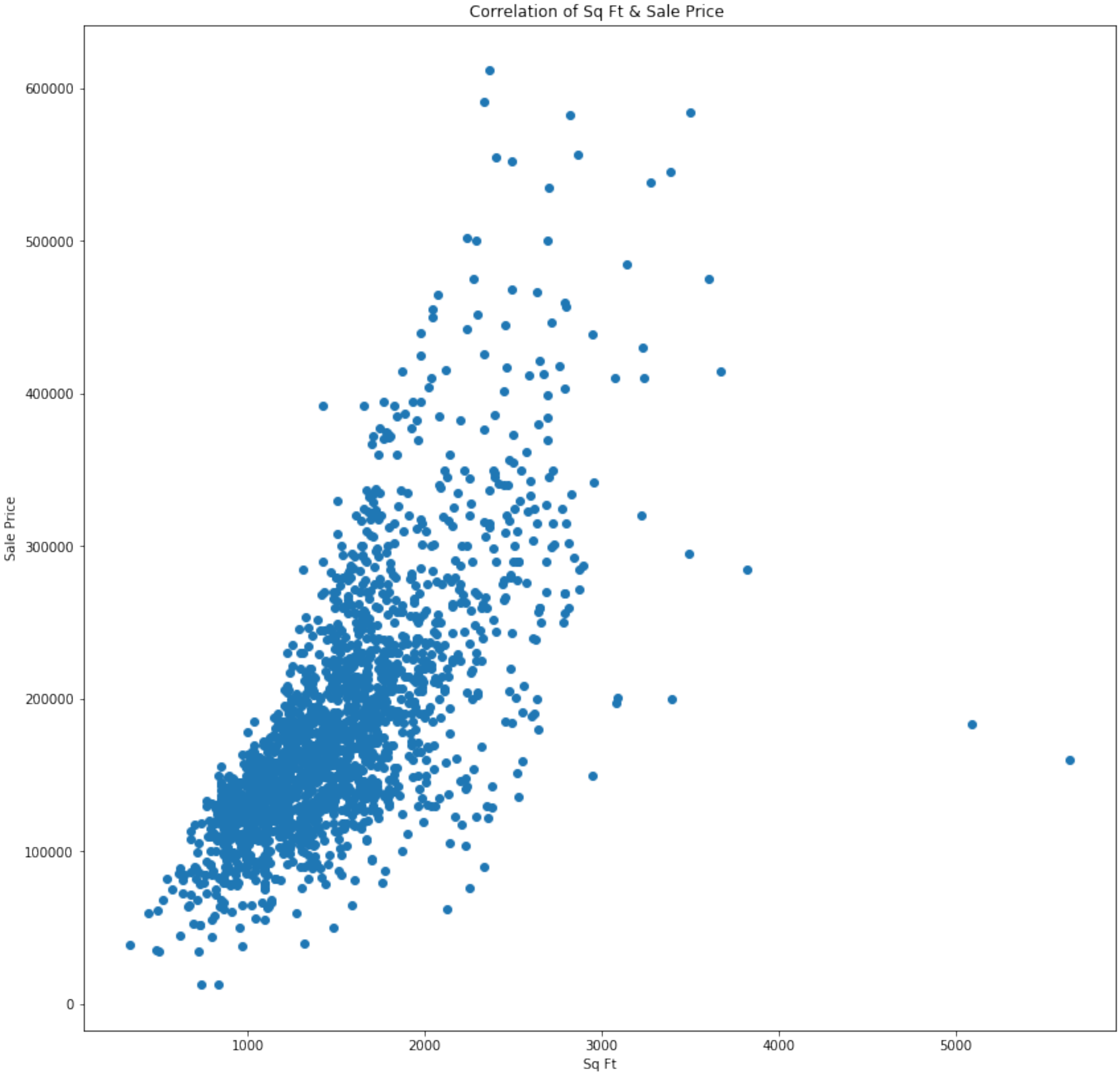
- Over 2,000 houses
- 81 Initial Variables including:
    - Overall Quality
    - Total Sq Footage (not including Garage and Basement)
    - Amount of bedrooms, full and half bathrooms
    - Lot Size
- Data was messy but manageable and insightful

- Began by finding the features of a home that correlated highly with sale price such as Total Sq Footage & Amount of Bedrooms
- Then got rid of unnecessary features and those with extremely low correlations between .5 and -.5 such as:
    - The PID
    - The year and month it was sold
    - The miscellaneous value of a miscellaneous feature
- Note that this is for the overall data. Miscellaneous features could definitely impact sale price of a home but there were not enough to include in my overall model
- Through exploration I found ordinal data that could be used numerically…

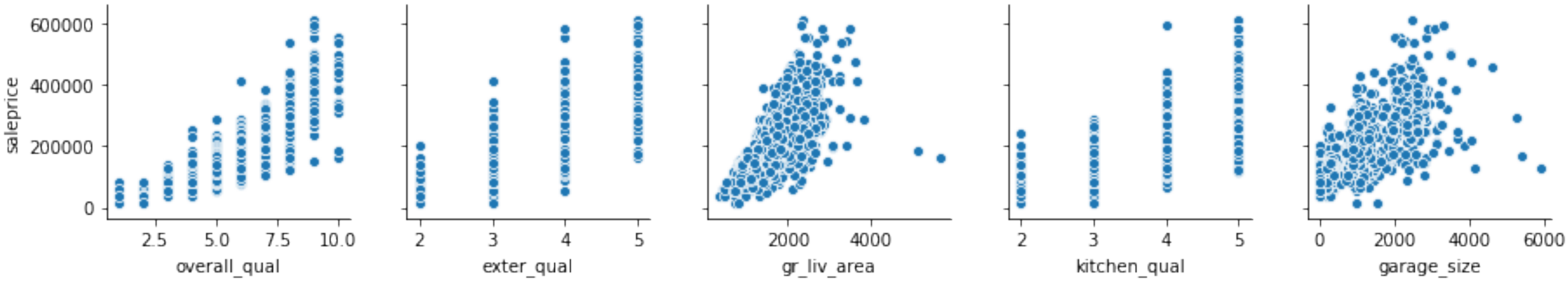
Numerical Correlation

- **Some Ordinal Features Converted to Numerical:**
    - Kitchen Quality
    - Exterior Quality
- **Features Exhibiting Collinearity:**
        - Garage Area & Garage Cars
- **Feature Engineered:**
    - Year Built
- **Surprise Non Collinearity:**
    - Beds, half & full baths did not have high correlation

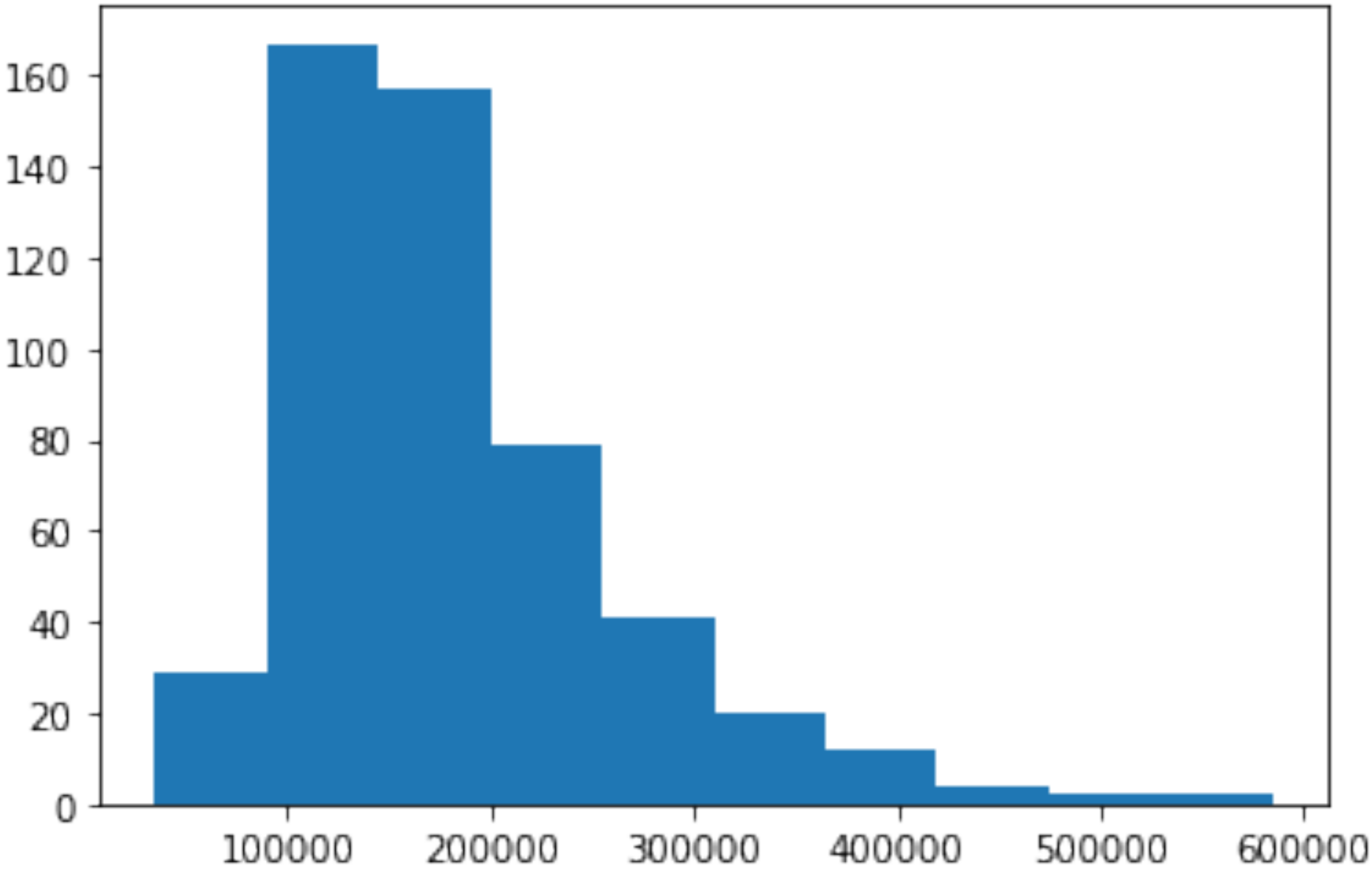

Top 5 Numerical Features Correlated to Price
overall_qual
exter_qual
gr_liv_area
kitchen_qual
garage_size
0.8
0.71
0.7
0.69
0.69
0
0.8
0.4
0.0
−0.4
−0.8


Correlation of Sq Ft & Sale Price
Sale Price
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Sq Ft

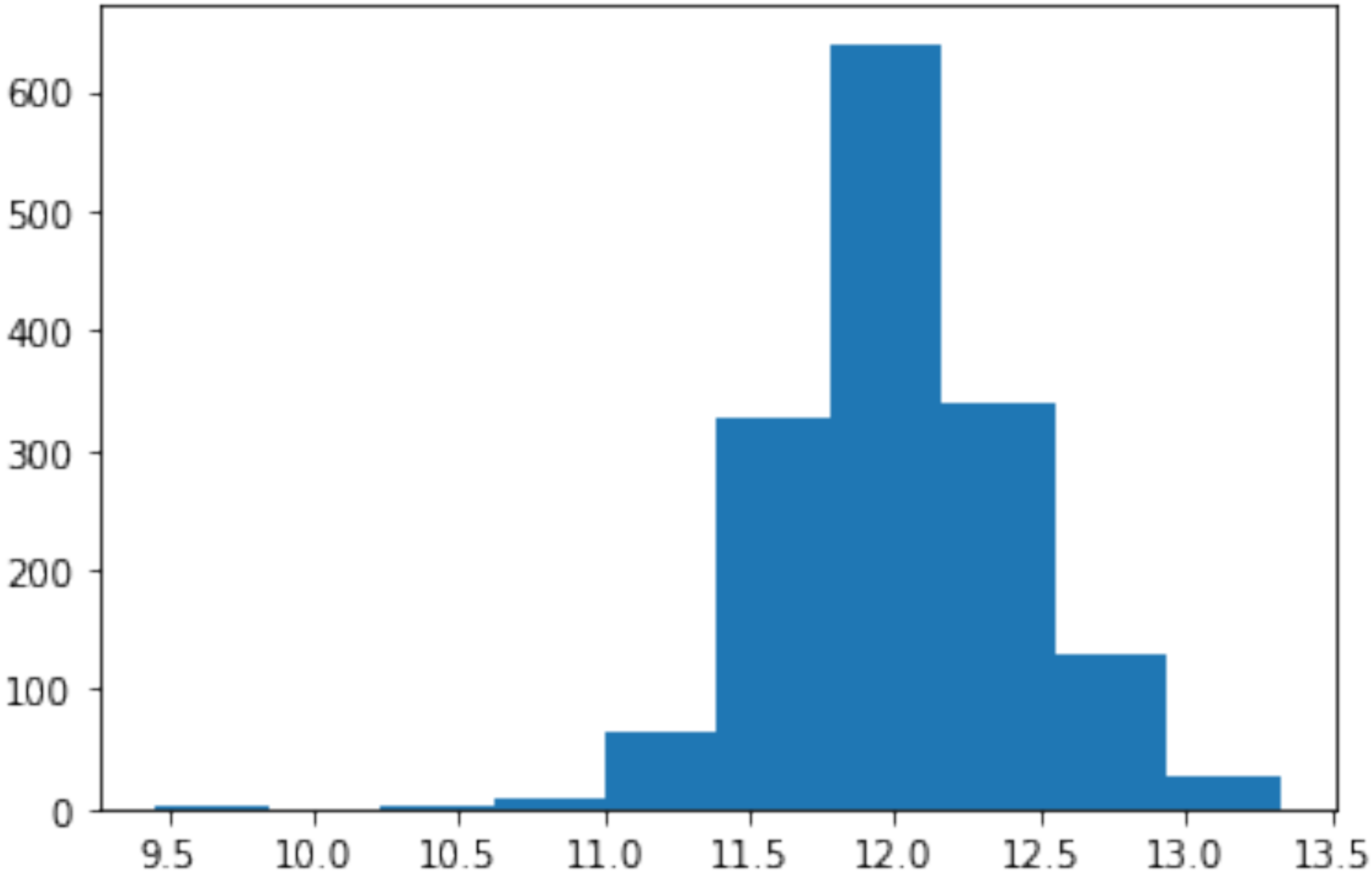
Pair plot showing Top 5 Features by Correlation



Distribution of my Sale Price Test Data



Distribution of that same Sale Price Data Transformed



- Linear Regression R2 scores were usually between 80 and 90 when using my Numerical Features. I tested multiple models where I included all my numerical features or only 5 or 10 etc.

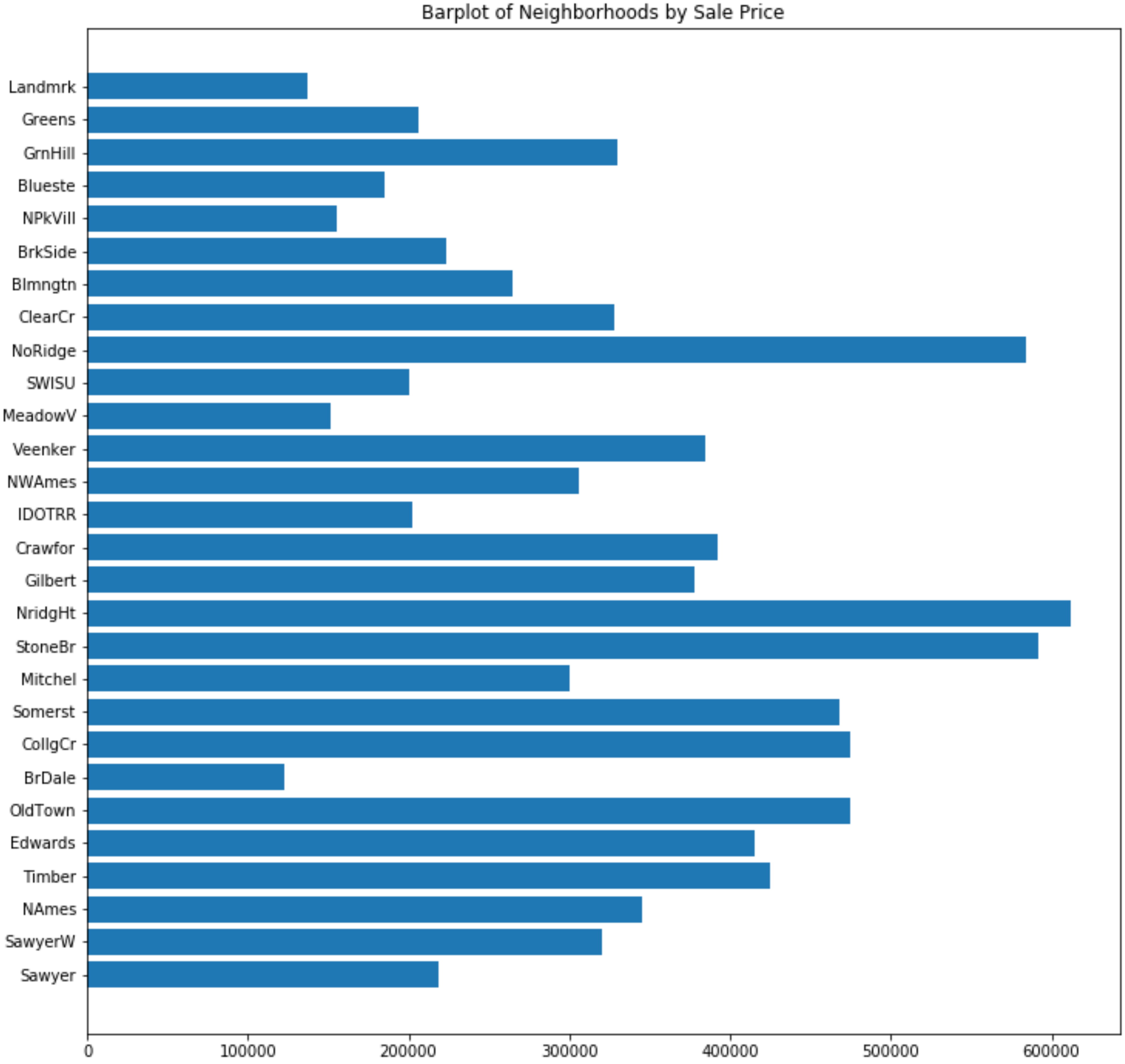
- Using Lasso & Ridge Regression Techniques did not noticeably improve my scores.

- For example, an R2 score of 85 means that 85% of the variance in my data can be explained using the features I included in my model

- To improve my model I could have created and explored dummy variables for the categorical features that were not ordinal, especially for neighborhoods and possibly zip codes.

- I could have used Polynomial Features on the features I did use

- Though there is room for improvement on the model I created, it is generalized and not specific to this data set.


Barplot of Neighborhoods by Sale Price
Landmrk
Greens
GrnHill
Blueste
NPkVill
BrkSide
Blmngtn
ClearCr
NoRidge
SWISU
MeadowV
Veenker
NWAmes
IDOTRR
Crawfor
Gilbert
NridgHt
StoneBr
Mitchel
Somerst
CollgCr
BrDale
OldTown
Edwards
Timber
NAmes
SawyerW
Sawyer
0
100000
200000
300000
400000
500000
600000

What can we gain from all this?

- I suggest that the homes with the highest sale prices will be determined by the following features:
    - Total Sq Footage
    - Overall Quality
    - Exterior Quality
    - Kitchen Quality
    - Garage Size
    - Basement Size/Area