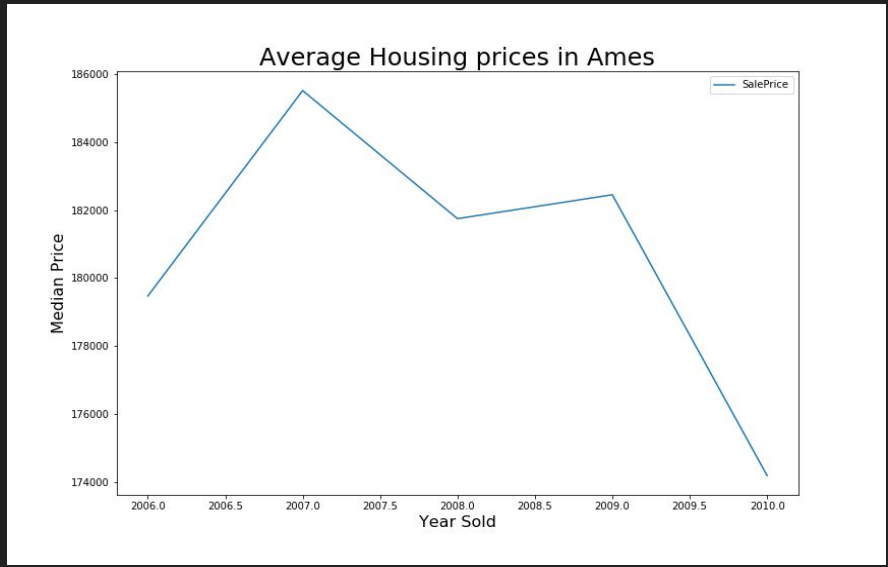
Predicting House Values in Ames Iowa

Project Overview

1. Background Information
2. Problem Statement
3. Data analysis
4. Optimal feature selection
5. Creating a model
6. Interpreting and tuning
7. Key takeaways
8. Recommendations

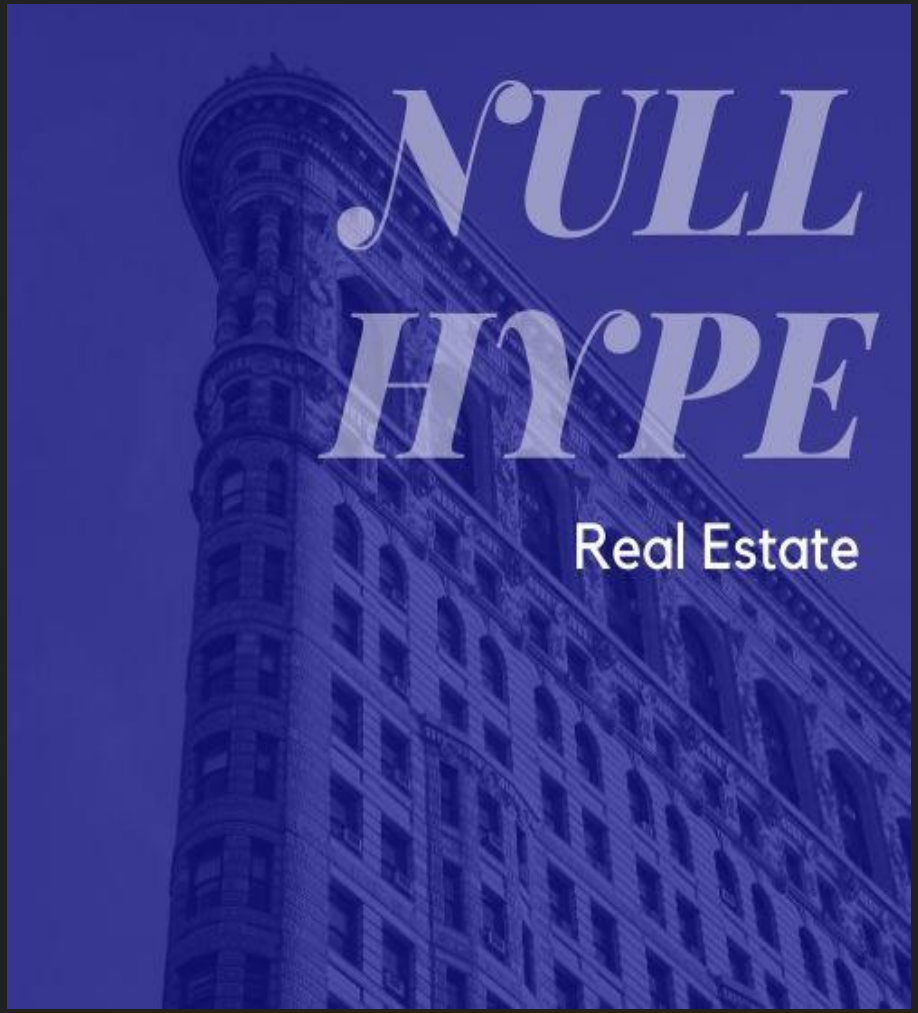
Ames housing at a glance

- Data - residential properties sold in Ames from 2006 to 2010
- We have almost 80 variables that pertain to each house sold
- The data is both qualitative and quantitative in nature
- Per the 2010 census ~ pop 59,000
- Home to Iowa State (Cyclones)
- Subprime mortgage crisis



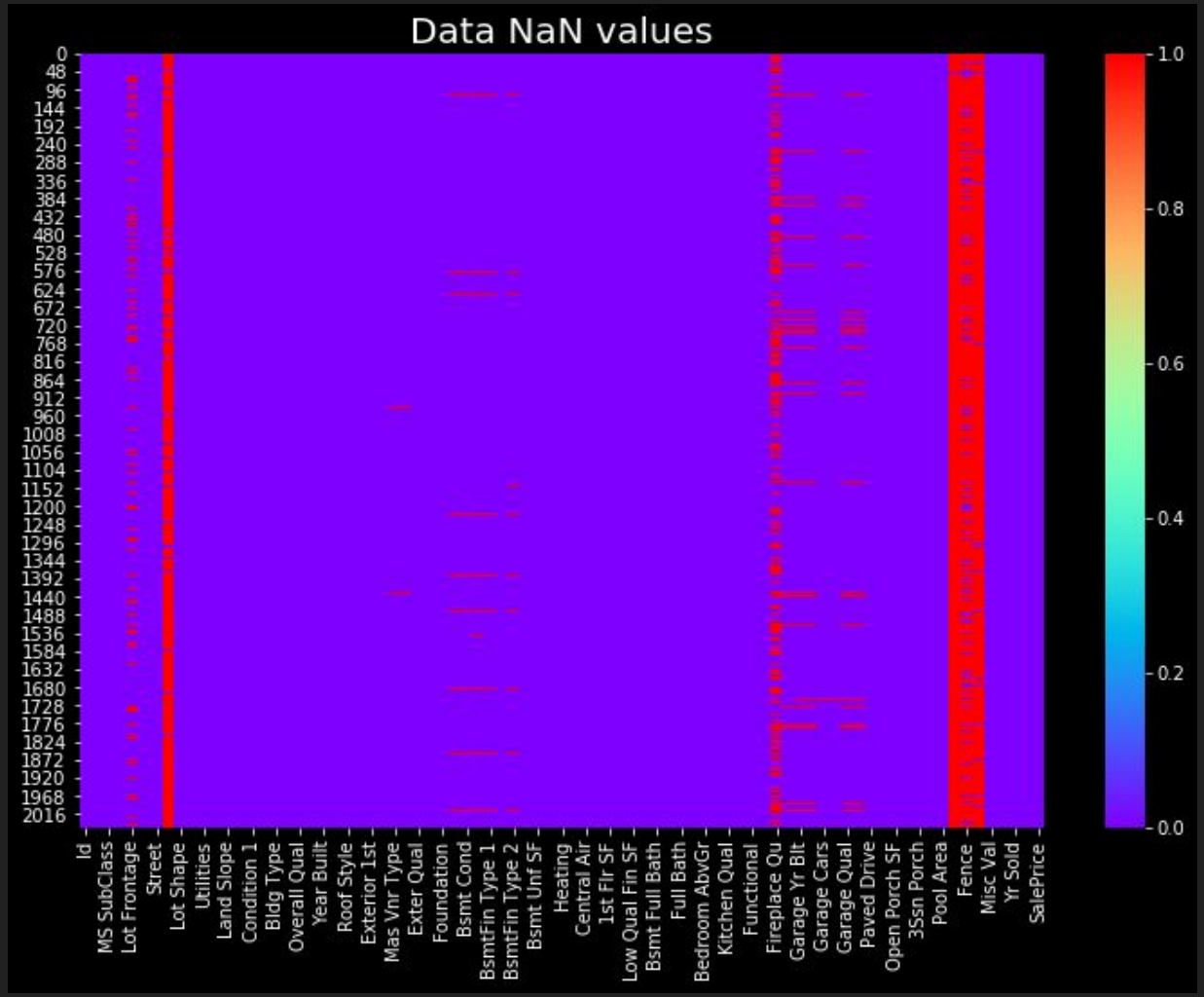
Problem Statement

Since the mortgage crisis, the once prominent Null Hype Real Estate group have begun to cede market share to local competitors. It appears their predictive pricing models are in need of a severe upgrading and they have tasked us to do statistical analysis on the Ames data and use it to *create a better forecasting model*. They will use this new model to accurately capture the Ames housing market trends and better reflect their client's true property value.

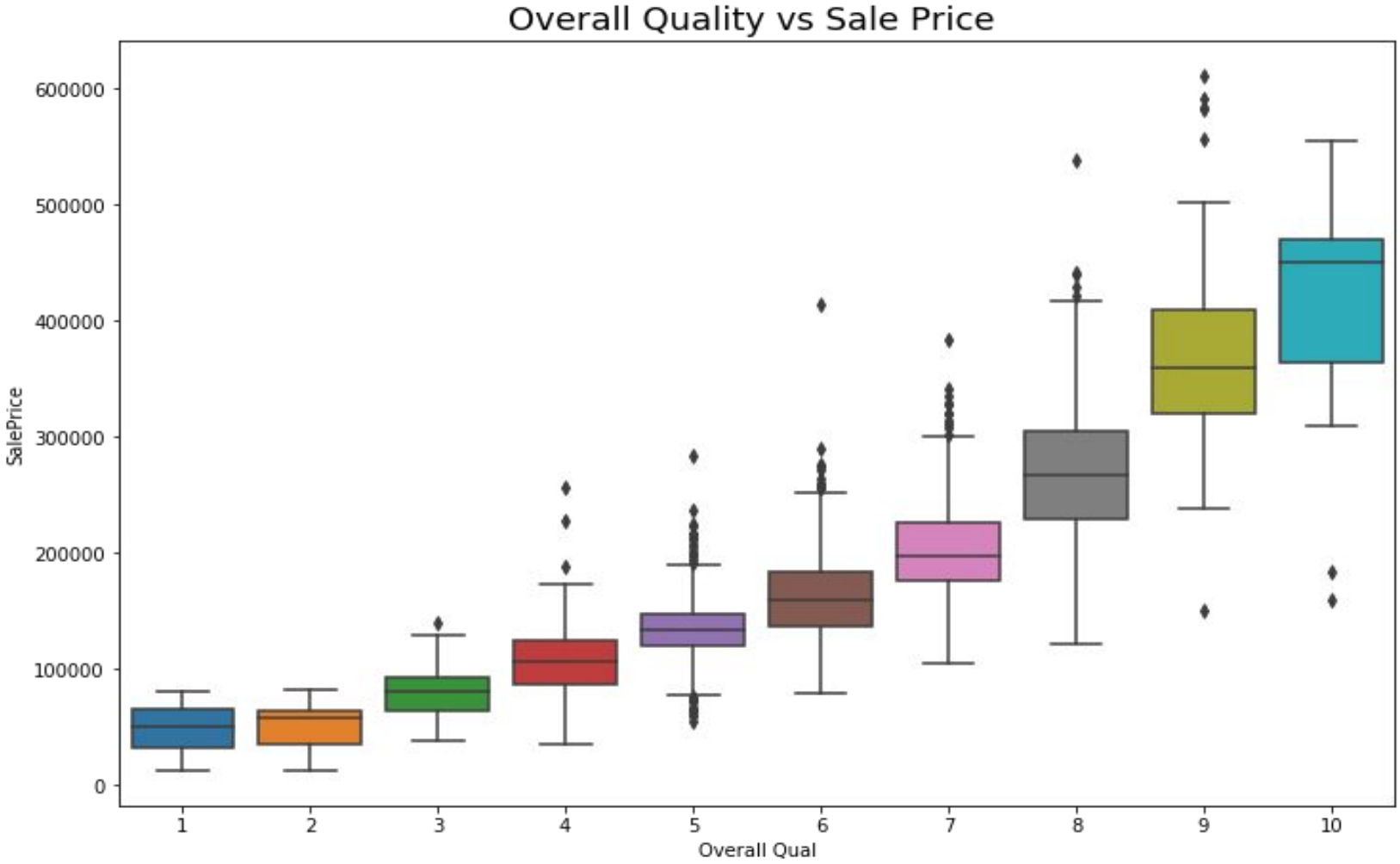


Data Exploration

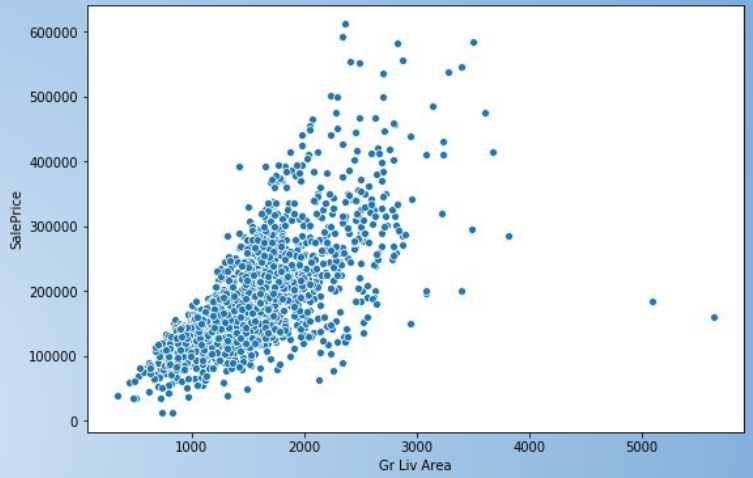
- Missing and ‘na’ values
- Impute correct values
- Usefulness
- Checking for correlation
- Multicollinearity
- Outliers
- Scale variables

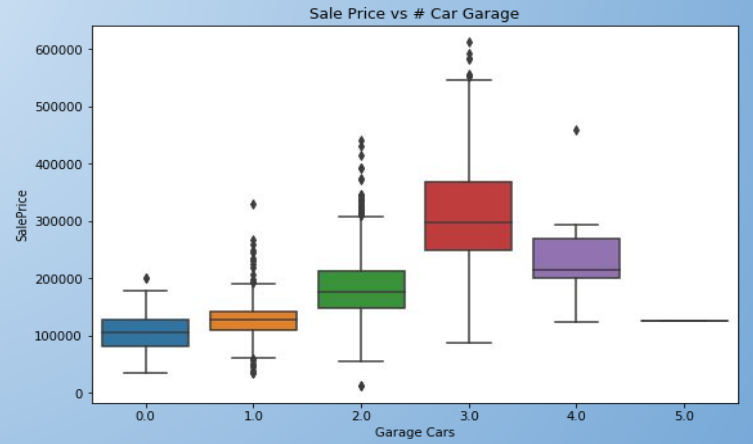


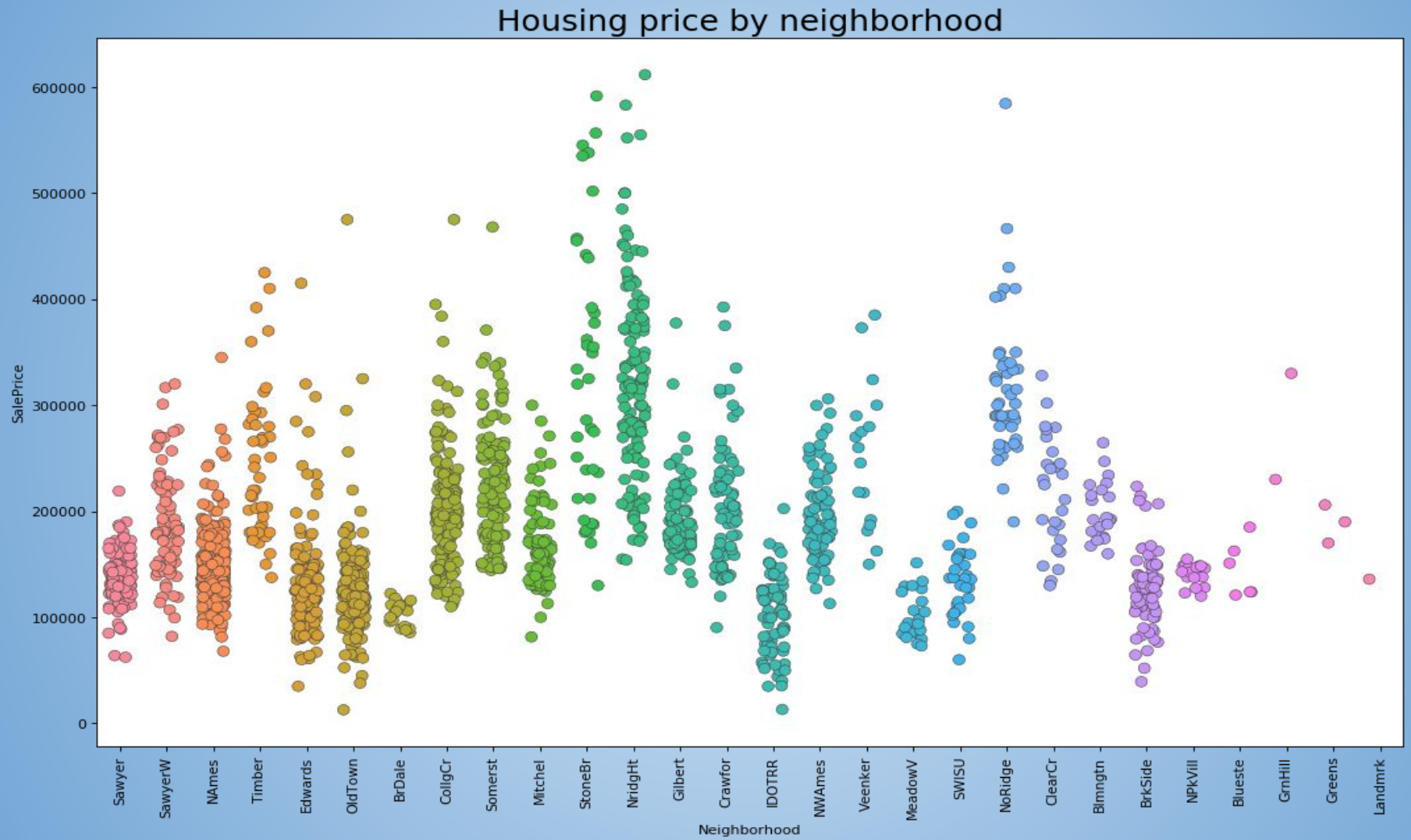

Sale Price Correlation
Id
Lot Frontage
Lot Area
Overall Qual
Year Built
Year Remod/Add
Exter Qual
Bsmt Qual
Bsmt Cond
Total Bsmt SF
Heating QC
Central Air
Gr Liv Area
Kitchen Qual
TotRms AbvGrd
Fireplaces
Fireplace Qu
Garage Finish
Garage Cars
Garage Area
Garage Qual
Paved Drive
Mo Sold
Yr Sold
SalePrice
total bath
total sf
age
-0.051
0.33
0.3
0.8
0.57
0.55
0.71
0.61
0.23
0.63
0.46
0.28
0.7
0.69
0.5
0.47
0.54
0.56
0.65
0.65
0.29
0.29
0.033
-0.015
1
0.63
0.81
-0.57
SalePrice
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

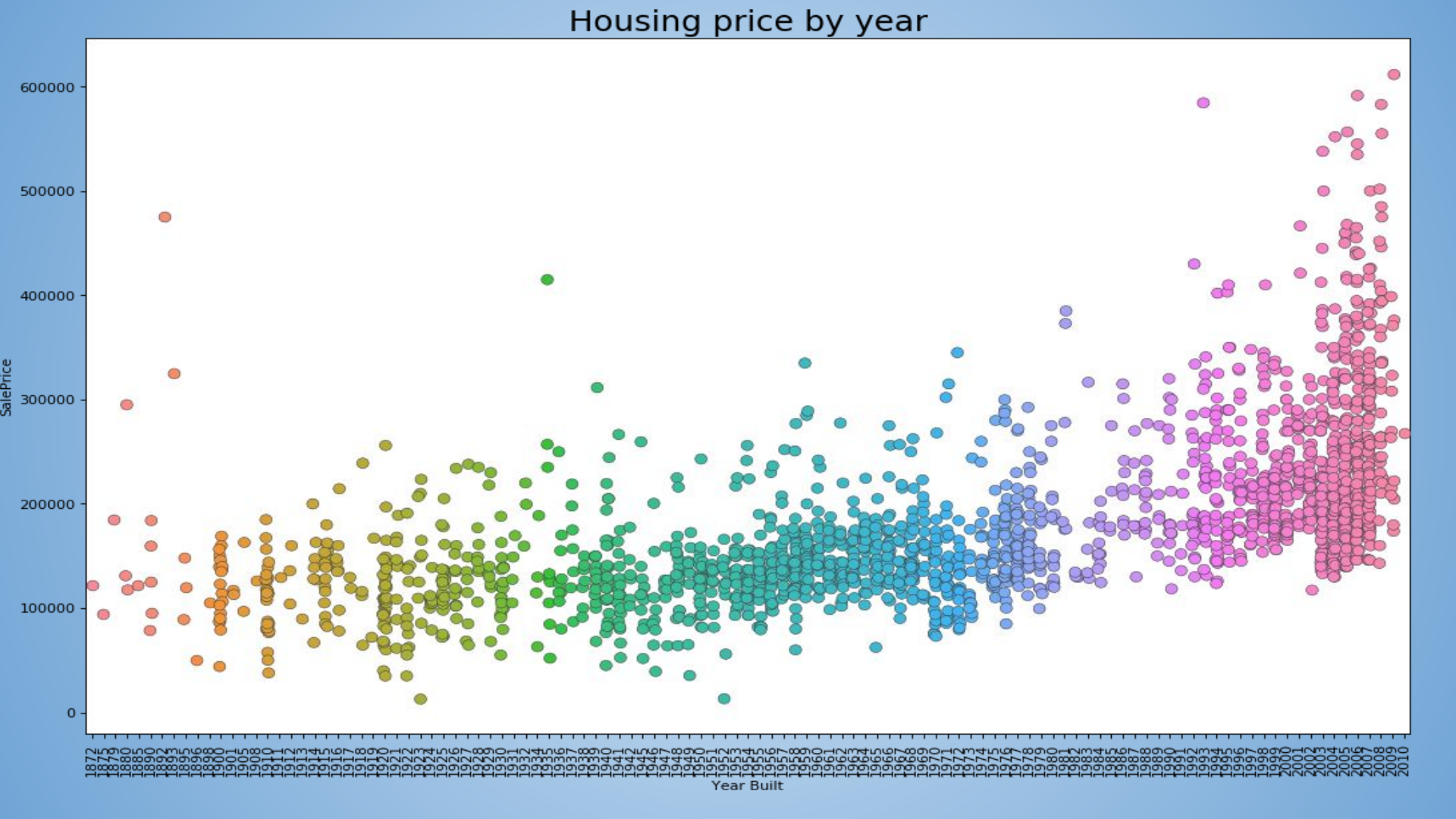

Overall Quality vs Sale Price
SalePrice
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
Overall Qual

Findings and Feature Selection






Housing price by neighborhood
SalePrice
0
100000
200000
300000
400000
500000
600000
Sawyer
SawyerW
NAmes
Timber
Edwards
OldTown
BrDale
CollgCr
Somerst
Mitchel
StoneBr
NridgHt
Gilbert
Crawfor
IDOTRR
NWAmes
Veenker
MeadowV
SWISU
NoRidge
ClearCr
Blmngtn
BrkSide
NPkVill
Blueste
GrnHill
Greens
Landmrk
Neighborhood


Housing price by year
SalePrice
600000
500000
400000
300000
200000
100000
0
Year Built

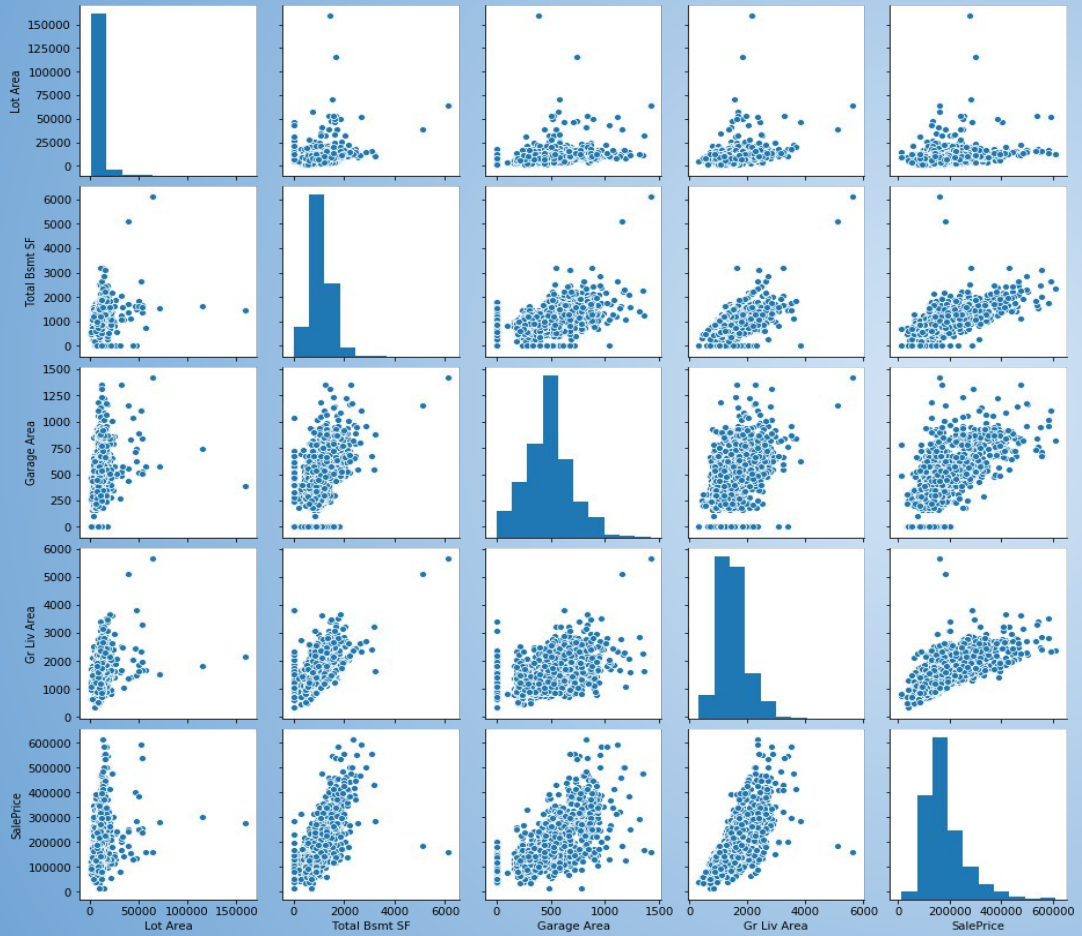
Model Preparation

- Ensure data is formatted and clean
- Remove redundant info
- Area and quality are main focus
- Numerically scaled important nominal features
- Include interaction terms and dummied variables

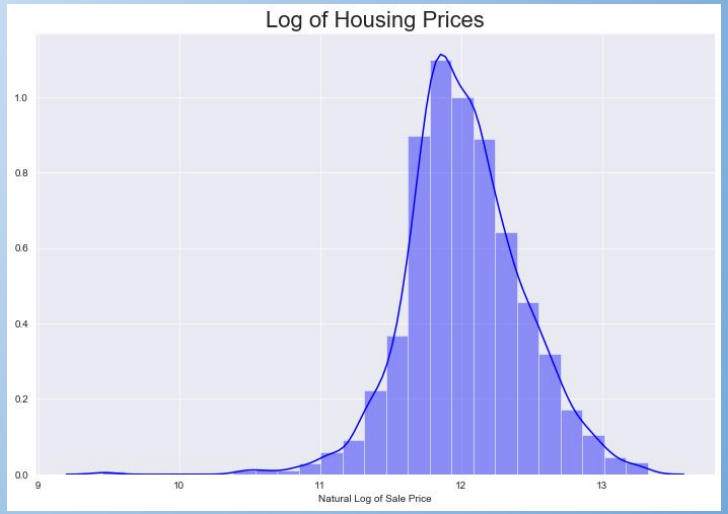
Linear Regression Model Analysis

- Started basic with forward feature selection
- Train, fit and check cross val scores
- Add or subtract terms depending on results
- Total area and overall quality had biggest impact on our model’s performance
- Initially the model was slightly underfit as the test scores were usually higher
- Scores improved in general with more variables as expected...but

Model Tuning



- **Model results drastically improved once the natural log of the sale price was used**
- **The model was also helped by add ordinal and converted nominal categories**
- **High correlation amongst variables**



Interpretations

- Area and quality have the most influence on accuracy of model
- Higher end features like fireplaces and pools raise property values
- The age of the house negatively impacts its worth, whereas newer and remodeled houses fetch higher prices
- The mean housing price can vary tremendously from one neighborhood to the next and this should be considered when determining property value
- The tuned regression model performs well with the data it has seen, with about 85% of the variance explained by our chosen predictors
- This model would do well in towns throughout the Midwest with similar demographics to Ames, but probably wouldn’t scale well to a larger city

Final recommendations:

- Remodeling goes a long way to increasing the property value
- Exterior quality extremely important
- Focus on kitchen!