Predicting House Prices in Ames, Iowa

Presentation Outline

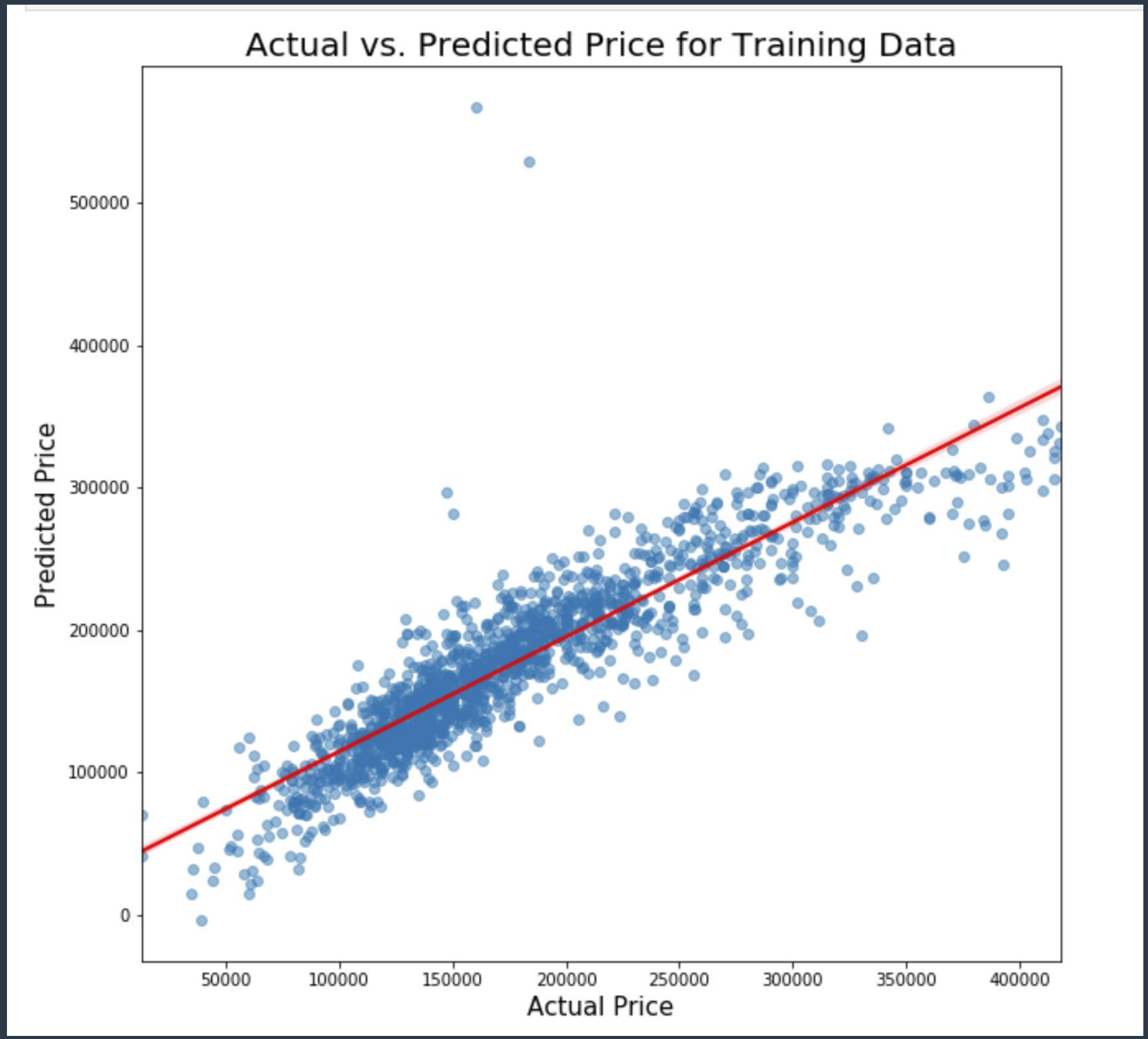
- Problem Statement
- Data exploration and cleaning process
- Developing a predictive model
- FAQs

Problem Statement

- How does house quality and square footage affect the sale price of a house?
- I will use housing data from Ames, Iowa to build a multiple linear regression model to predict housing prices in order to provide relevant information to realtors

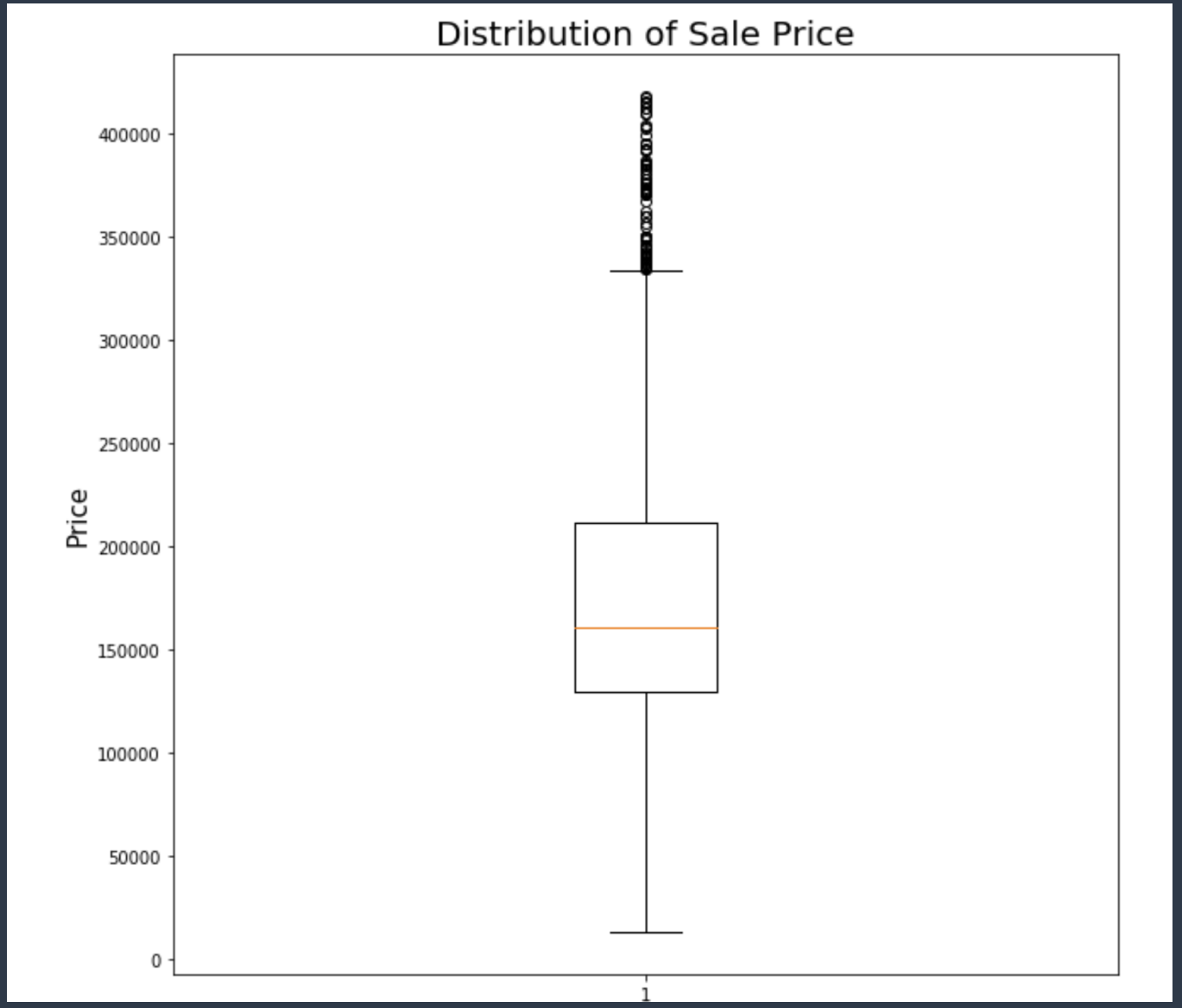
Data Summary

- Data from 2051 houses
  - Sale Price outliers!
  - Removed if more than 3 standard deviations away from the mean
  - Gr Liv Area outliers – 2 above 5000


Actual vs. Predicted Price for Training Data
Predicted Price
Actual Price
500000
400000
300000
200000
100000
0
50000
100000
150000
200000
250000
300000
350000
400000

Data Cleaning

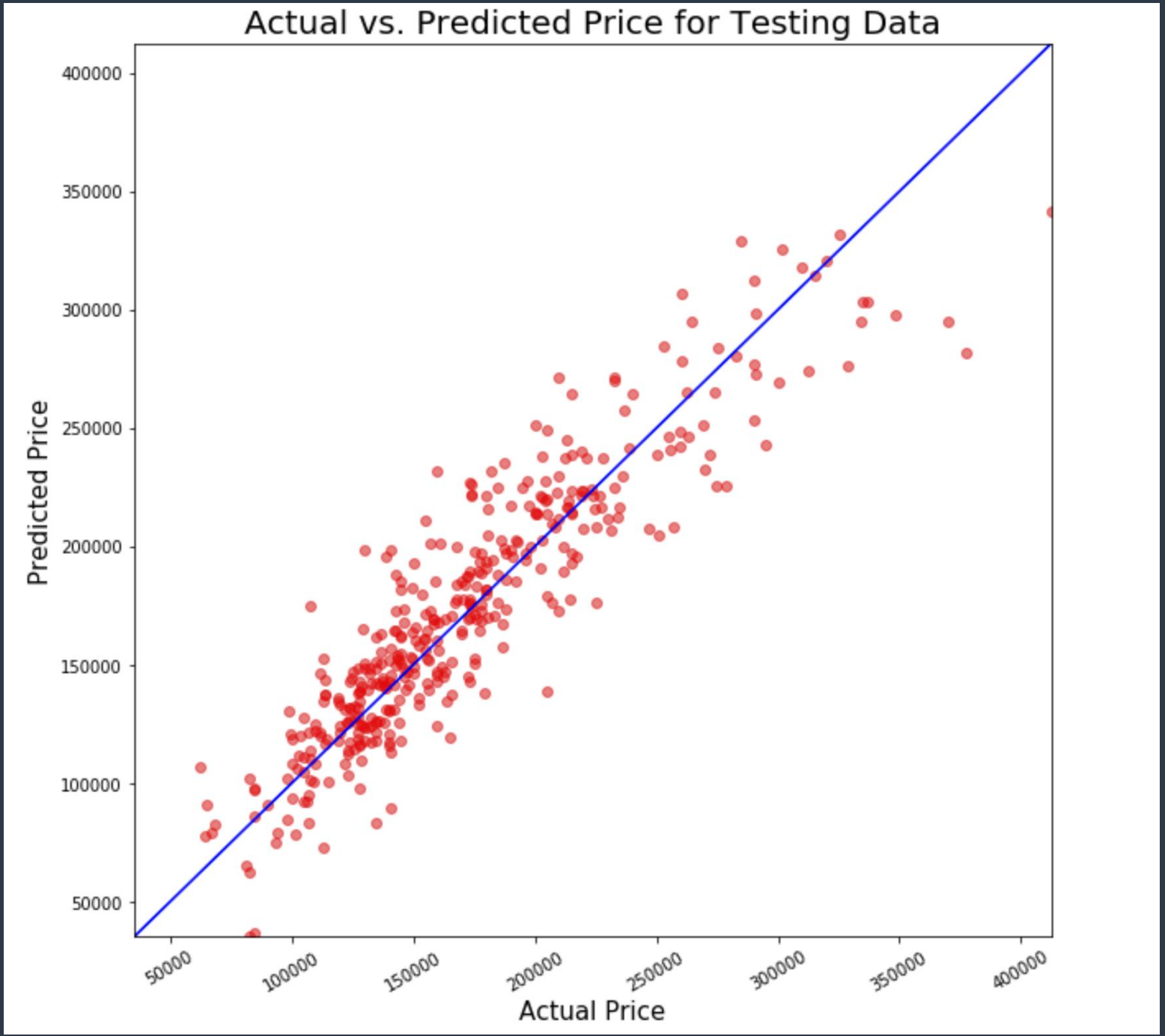
- Removed Outliers
  - 2014 houses
- Replaced missing values with the median
  - Garage Year Built



Predictive Model

- Factors selected based off correlation with the sale price
  - Overall Quality, Above Grade Living Area, Garage Area, Total Basement Square Feet, 1st Floor Square Feet, Year Built, Fireplace Quality
- R^2: how much variation does the model account for
  - Train: 86%
  - Test: 83%
- RMSE: average error in predicted price
  - Train: $26,649.66
  - Test: $23,911.80


Actual vs. Predicted Price for Training Data
Predicted Price
Actual Price
400000
350000
300000
250000
200000
150000
100000
50000
50000
100000
150000
200000
250000
300000
350000
400000

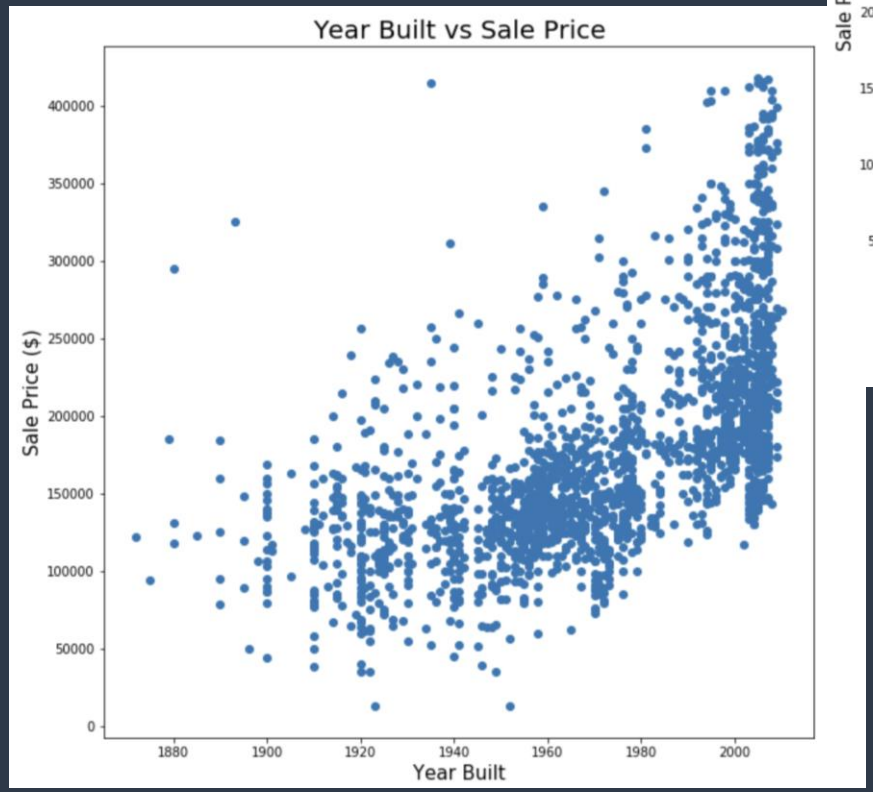

Actual vs. Predicted Price for Testing Data
Predicted Price
Actual Price
400000
350000
300000
250000
200000
150000
100000
50000
50000
100000
150000
200000
250000
300000
350000
400000

Frequently Asked Questions

- Which features appear to add the most value to a home?
- Which features hurt the value of a home the most?
- What are things that homeowners could improve in their homes to increase the value?
- What neighborhoods seem like they might be a good investment?
- Do you feel that this model will generalize to other cities? How could you revise your model to make it more universal OR what date would you need from another city to make a comparable model?

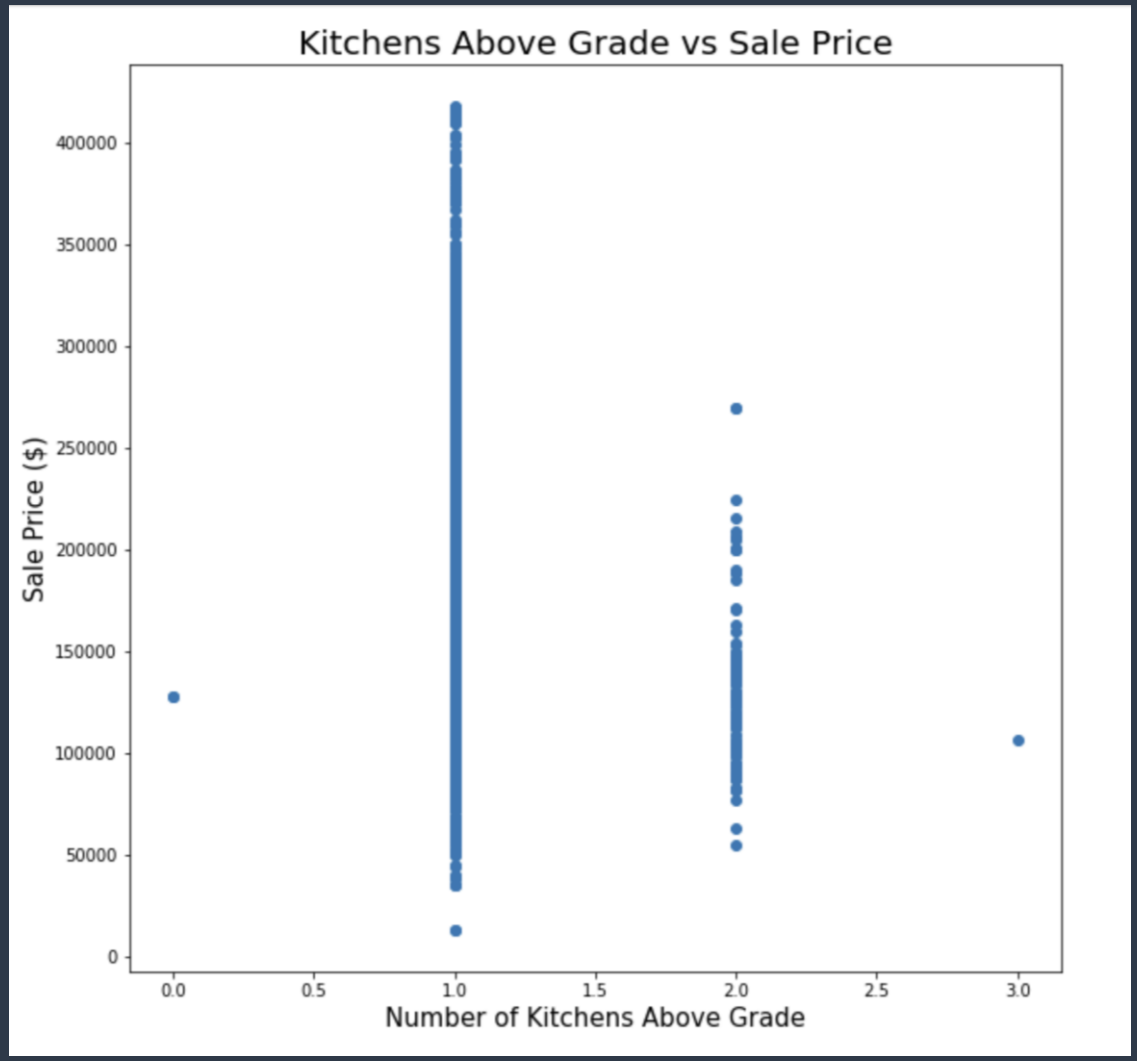
What model features add the most value?

- GR Living Area
  - As above grade living area goes up one square foot, sale price is predicted to increase $50.60 assuming all else stays the same
- Year Built
  - As the year built increases by year, the sale price is predicted to increase $459.94



What model features hurt value the most?

- Kitchen AbvGr
  - As the number of kitchens above grade, increase, the sale price is predicted to decrease $22,808.26
- Enclosed Poarch square footage also had a negative correlation to sale price, but was not used in the model

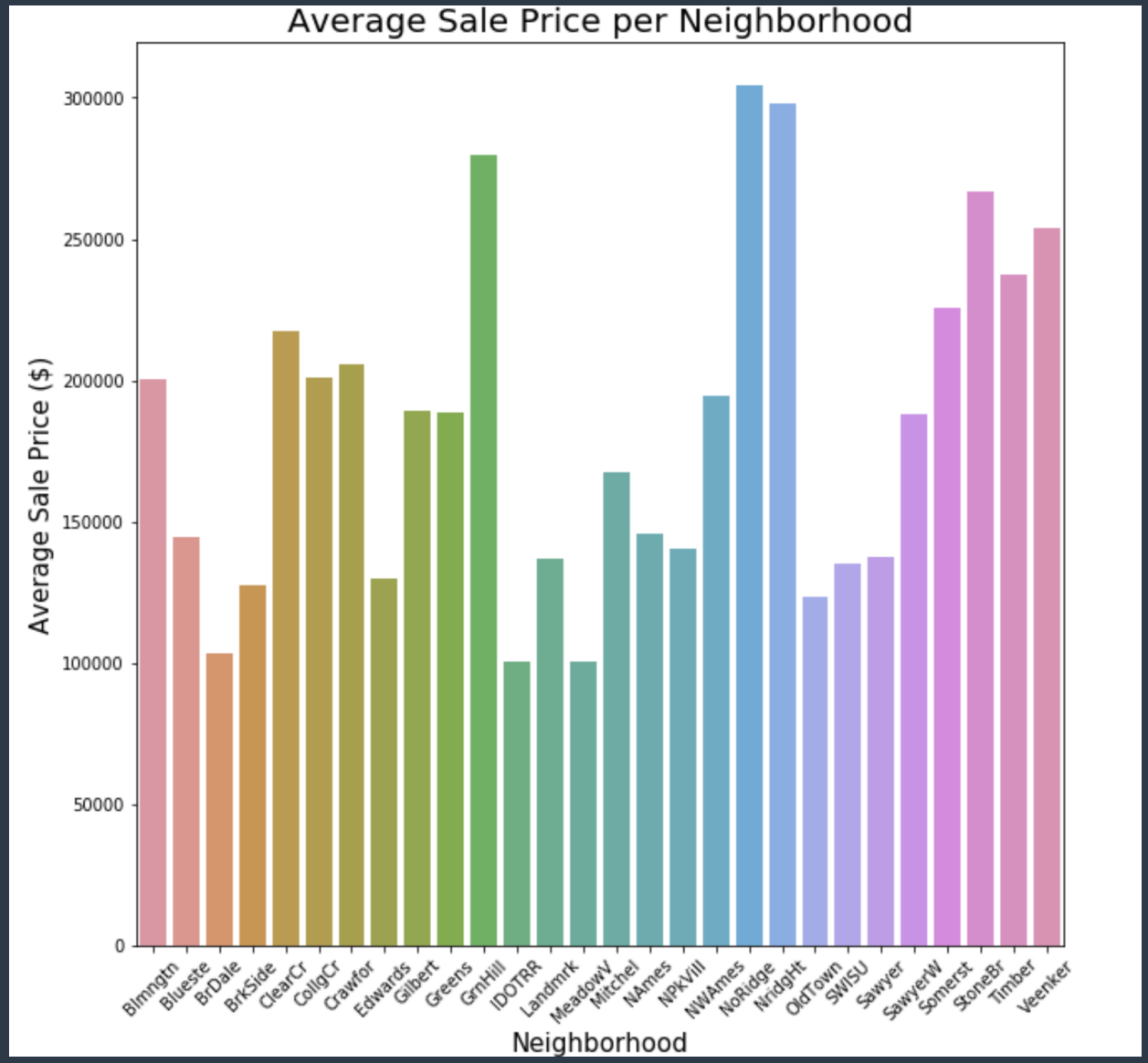


What should be improved before listing a home?

- Size
- Improvements – fix windows, repaint, update fixtures

Where are the best neighborhoods for investment?

- North Ridge: $304,469.29
- Northridge Heights: $298,050.07
- Green Hills: $280,000.00
- Stone Brook: $266,656.89



Can this model be used for other cities?

- Short answer: no
- We can try, but it might not be as accurate as this model
- If we gather more housing data from other cities: yes

References:

- Kaggle data dictionary
- https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
- Lessons!