

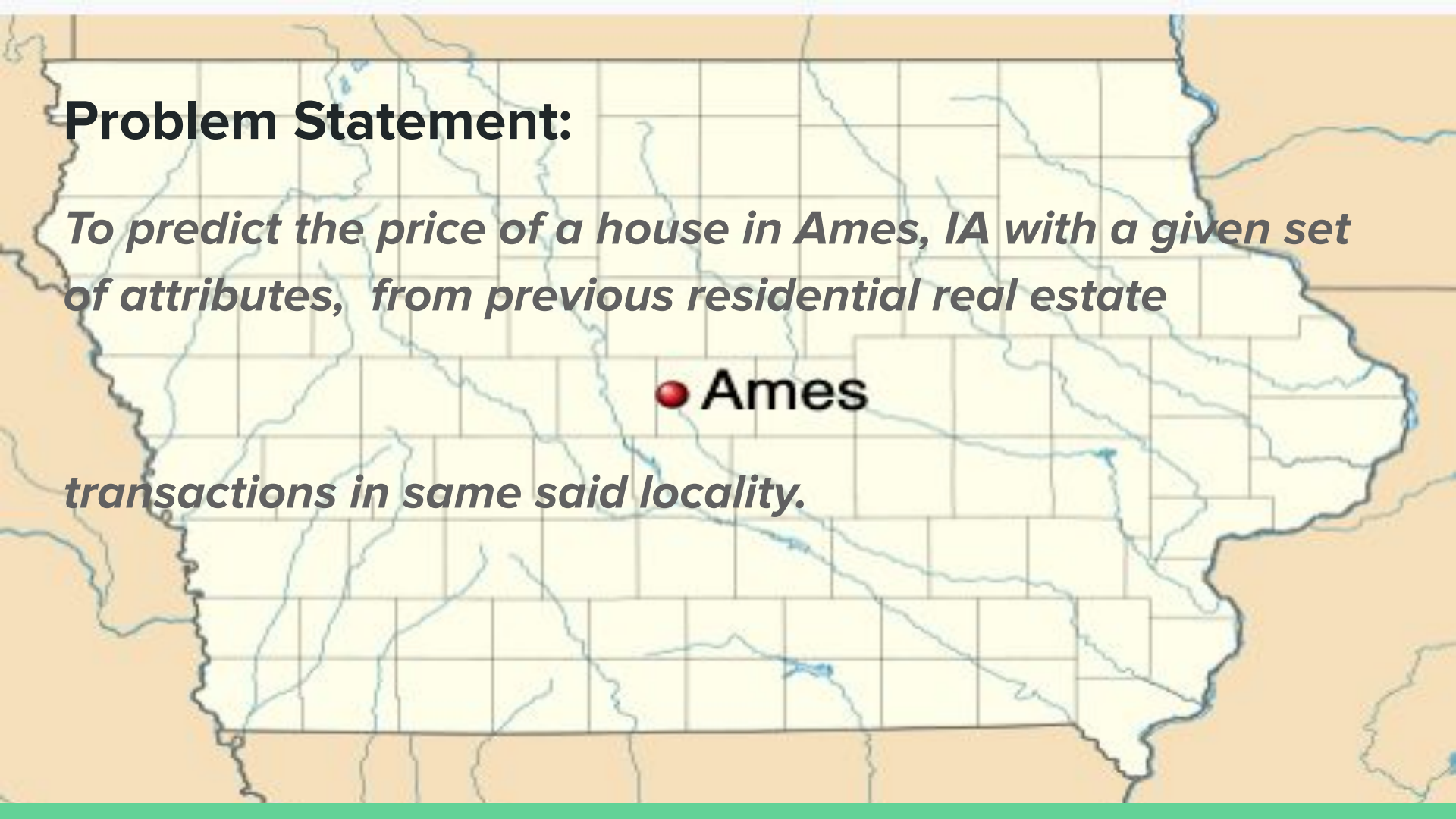
# Campaign to Project Prices

---

Immersion into the Ames, IA Housing Market

## **Problem Statement:**

*To predict the price of a house in Ames, IA with a given set of attributes, from previous residential real estate*

A map of the state of Iowa, showing its county boundaries and major river networks. A red dot is placed in the central part of the state, representing the location of Ames. The word "Ames" is written in black text next to the red dot.

**Ames**

*transactions in same said locality.*

# Given Data

Top Level Description

- Where
  - What of Style, Size, Condition
  - Granualized Whats
    - How Many
    - Condition
  - Transactional
    - When
    - Sale Type
    - Price
-

# Data Cleaning

- Main learning pick-up takeaways
  - Right away dropped anything over 1000 missing (looking back, wouldn't do that now)
    - Sherlock Holmes -- The Dog That Didn't Bark
  - With those in the low hundreds....
    - At first, filled in with the mean and went on my way
    - Backtracked with the superior method
      - Bit coded a new column of item missing or not
      - Filled in with the minimal value based on column context
    - A number missing/not missing columns showed up in models as useful (aka not dropped)

[illegible][illegible]

# Iteration 1: The Numericals

## From the Heatmap....

- Pure numericals like lot sizes, square footages
- Quasi-numericals that are categorical but have rank cardinality like house condition
- Missing/not missing bits

## It was cool seeing....

- Missing/not missing bits showed up as correlated and useful

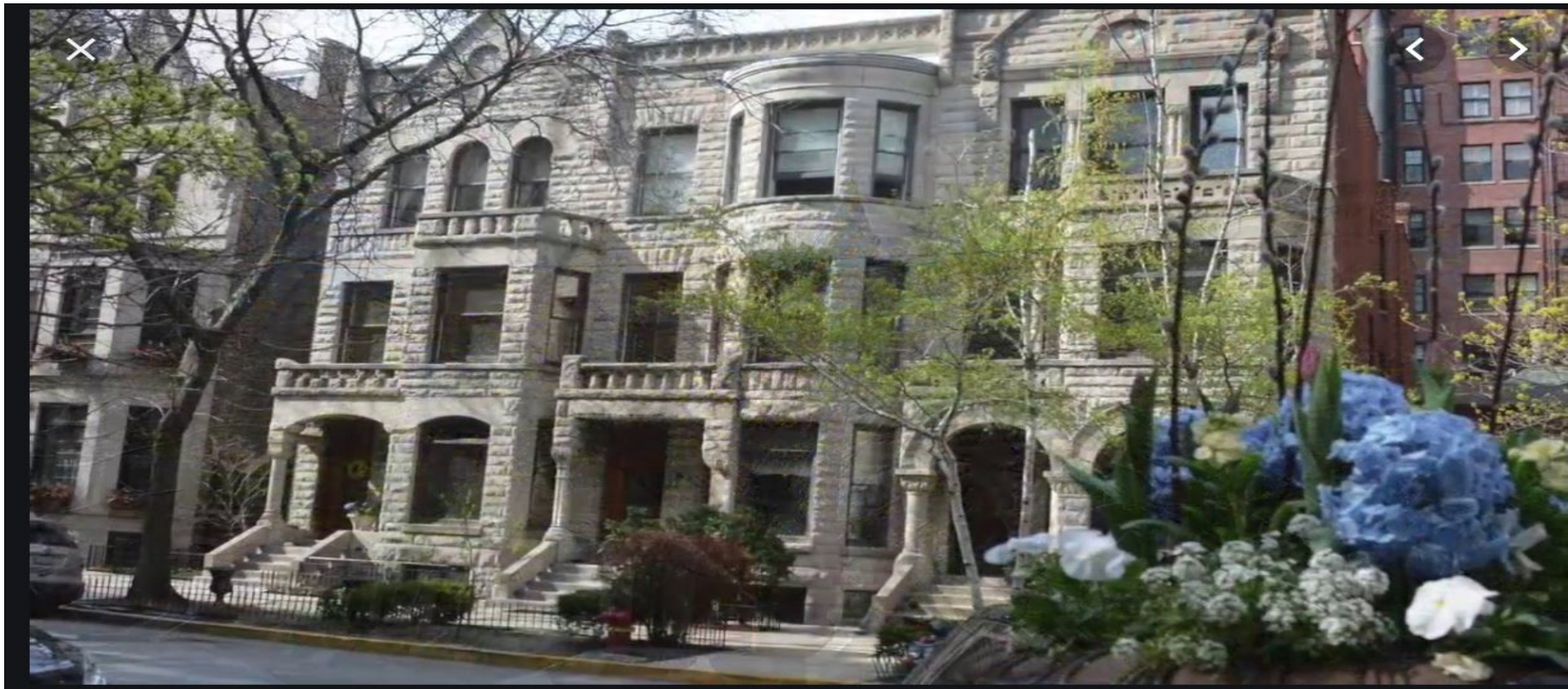
**OLS p values pointed out features to cull**

## Iteration 2: Dummied Up Selected Categoricals

- CONSIDERED **NEIGHBORHOODS** THE ***MOST VALUABLE CATEGORICAL***
- Deemed Building Type (single fam vs. townhouse for example) valuable
- Deemed Building Style (1 story vs 2 story vs split level for example) valuable
- Sale Type -- Financing Employed is info on buyer's financial strength; easy credit goosing prices perhaps; also NEW HOMES demarcated in here
- Transformed Sale Months to Selling Seasons (theory being spring selling season stronger than other seasons)

Takeaways: performance improved vis a vis baseline Iteration 1; OLS culled again





Worst house in the Gold Coast....





Best house in West Englewood

Moving to Elmhurst  
just to join York HS  
X-Country Team



# East Aurora fought and received a Naperville zip

When people talk about the east side of Aurora, they're usually implicitly excluding the parts which are far enough east to be in DuPage County. The DuPage County sections offer modern, suburban-style housing. They have more in common with Naperville, the well-off suburb just east of Aurora, than they do with the urban parts of Aurora

Read more:

<http://www.city-data.com/forum/chicago-suburbs/1745768-east-aurora-vs-west-aurora.html#ixzz61MoNOvDu>

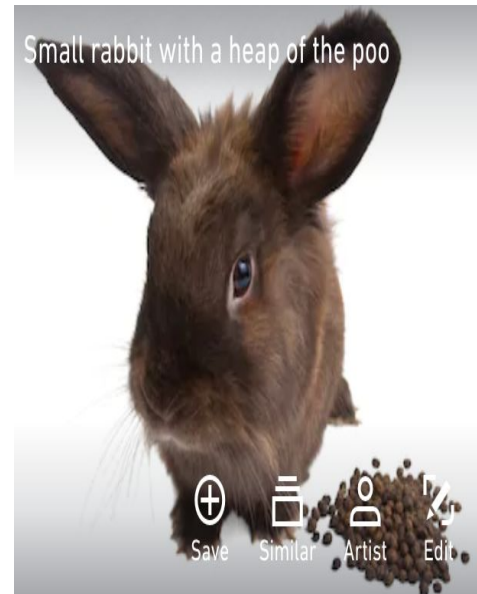
# Ignore a Munger-ism at your own risk



and



Is still



## Iteration 3: Feature Engineering

The hot dog/mustard example will uncover unsuspected relationships in other domains, but the Primacy of Location in real estate is such an lalapalooza 800 pound gorilla that it wound up being a dead end here. Or at least my attempts came to nought.

That made me even more eager to heavily pick on the location aka neighborhood feature. I wound up squaring it. That may be overfit in other domains, but it is justifiable here given the domain specific dynamic and realities.



## Iteration 3: Feature Engineering cont'd

Neighborhoods was also poly featured with:

- Overall Quality (of the house)
- Total Rooms Above Grnd

Also finally, the data was collected from the tail end of the housing boom into the housing bust, so Selling Season was poly featured with Sale Year.

## Iteration 3: Feature Engineering cont'd

Results: The neighborhood squareds and other terms made the correlations cut. A lot if not most survived the typical OLS culling hack.

A surprise was that NONE of the Season-Sale Year features made correlation, much less get to the OLS buzzsaw. Right through the biggest economic setback since the Great Depression. For what it's worth Summer Years was the best of the bunch. Some outside research performed points to smaller center of the country markets didn't much participate on the way up but didn't fall much on the way down. Ames wasn't Las Vegas or Miami.

Much improved performance vs Iteration 2 and baseline Iteration 1. OLS culls again

# Ridge and Ridge CV

Small shrinkage R square vs RL Iteration 3

# Lasso

Small shrinkage R square vs RL Iteration 3

# Final Words

---

Conclusion: LR 3, OLS 3, Ridge and Lasso had very close  $R^2$

RL and OLS did slightly better than Ridge or Lasso

RL had tiny bit better performance than OLS

Picked LR over OLS for kaggle for that bit of performance since there's really only bragging rights involved

If I had real money or skin in the game, I'd go with OLS --- best value of performance vs streamlining