
Determining Property Value in Ames, Iowa

— Aerika Song —

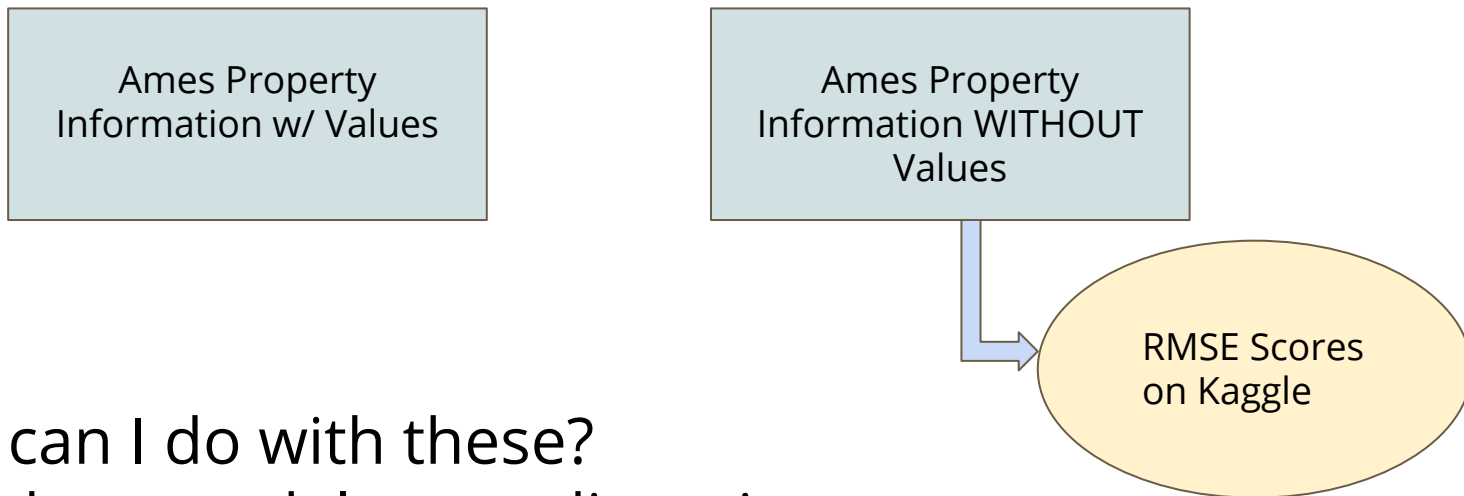
A little background...

- **NOT** a real estate agent
 - Data Science Student at General Assembly
 - I have a property in Ames, Iowa I want to value and sell!
-
- Problem: How am I going to do that?
 - Predictive model

Overview

- Data cleaning
- Picking variables
- Model creation
- Predictions
- Using my model

What we have:

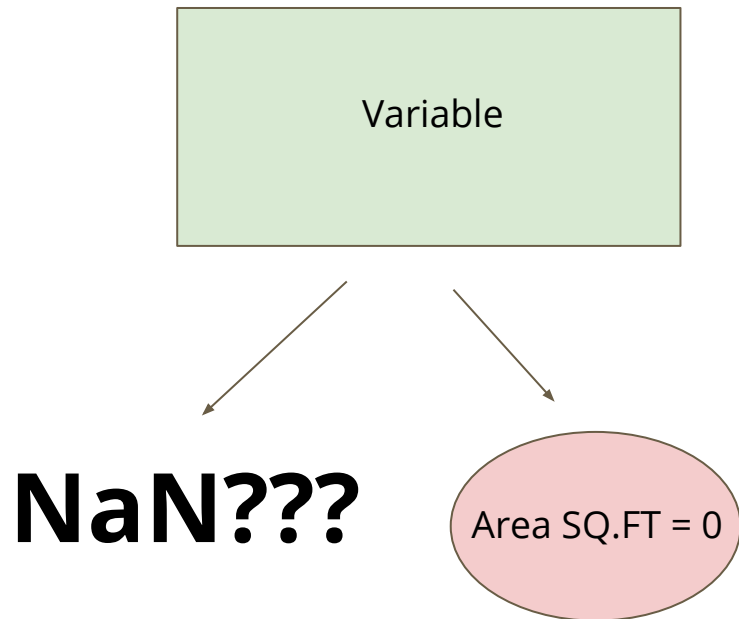


What can I do with these?

1. Make a model to predict prices.
2. Submit predicted prices for data without price values → RMSE Scores

1. Data Cleaning

- Assume NaN Cells = cells with blank space
- If Area SQ.FT = 0....
- Probably left blank because there wasn't a descriptive category because there is none!
 - Change to "NA"
- Applied similar mindset to all variables that had NaN values.
 - I.e. Pool, Misc Features, Garage Type, etc.



1. Data Cleaning (continued)

- Python's describe
 - Minimum priced property
 - Maximum priced property
 - Examined
- NaN Single Values → Median
 - 2nd Basement Types → Median = Unfinished
- Any strange outliers? Dropped.
 - Ie. Mansion sized property for only \$200,000
- NaN critical sale price predictor? Dropped.
 - Ie. Basement w/o ANY information
- Post-Model Outliers
 - Identified and removed if beneficial



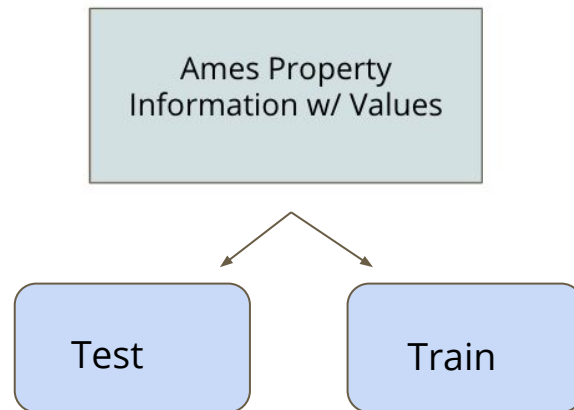
2. Picking Variables

- What are obvious predictors of a property's value?
 - Mindset: I'm buying a house, what are the things that I first notice?
 - Correlation Values
- Dummy-valued categorical data
- Single Linear Regression Model
 - Criteria: Variable Affects +/- \$30,000
 - Added to list of variables



3. Created My Model

- Split data WITH sale prices into:
 - 1) Training Data 2) Testing Data
 - Predicted Prices vs. Actual Sale Price
- Target: Sale Price
- Optimization: LOTS OF BACK & FORTH
- Final %'s of explained variability of prices:



Training Data

94% Variability
Explained

Testing Data

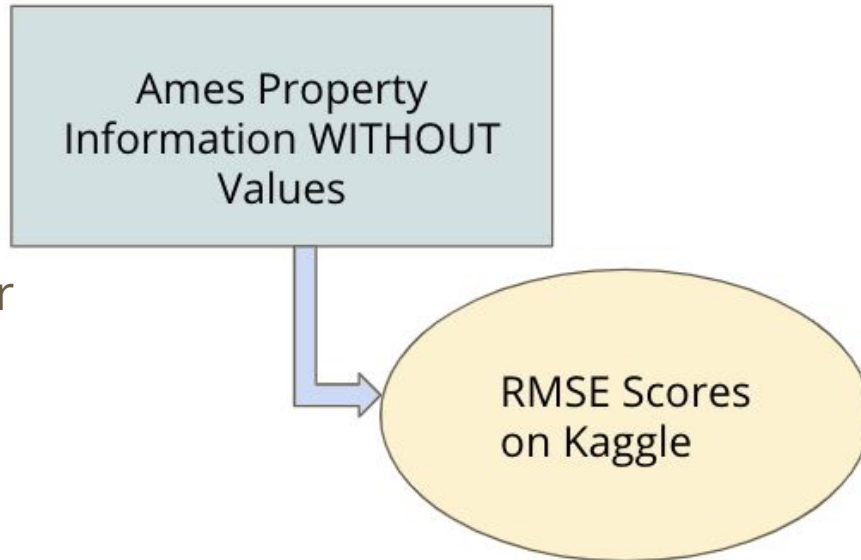
92% Variability
Explained

Splitting Training Data (5 Parts)

92% Variability
Explained (Mean)

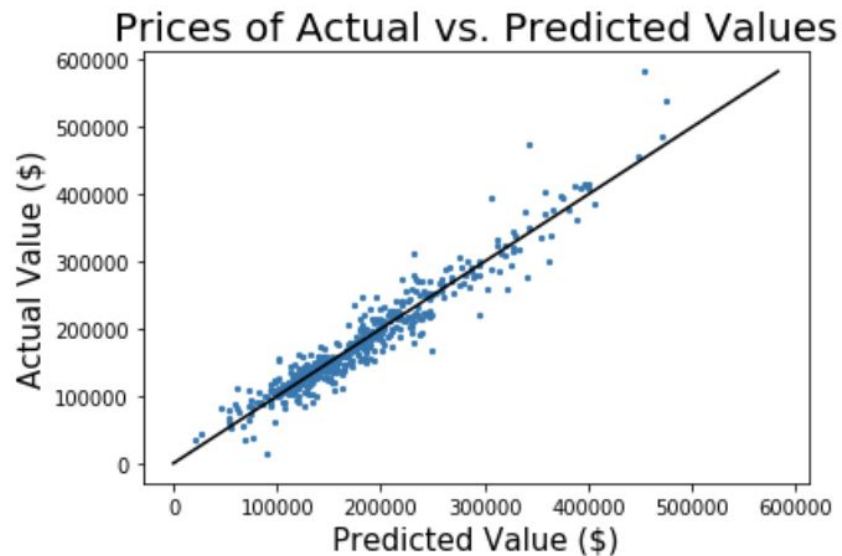
4. Predictions

- Using model
 - Made predictions for →
- Submission on Kaggle
- Examine RMSE Scores
- Lower RMSE Scores = better model



Caution to Note

- Big chunk of data in \$100,000 - \$300,000 range
- Not many higher value properties to model



Sources

<https://stackoverflow.com/questions/49065837/customize-the-axis-label-in-seaborn-jointplot>

<https://stackoverflow.com/questions/34001751/python-how-to-increase-reduce-the-fontsize-of-x-and-y-tick-labels/34004236>

<https://stackoverflow.com/questions/29813694/how-to-add-a-title-to-seaborn-facet-plot>

<https://stackoverflow.com/questions/28638158/seaborn-facetgrid-how-to-leave-proper-space-on-top-for-suptitle>

<https://blog.goodaudience.com/jupyter-notebooks-the-real-way-to-use-them-5b4417ea77ba>

<https://stackoverflow.com/questions/28914078/filter-out-rows-based-on-list-of-strings-in-pandas>

<https://stackoverflow.com/questions/14463277/how-to-disable-python-warnings>