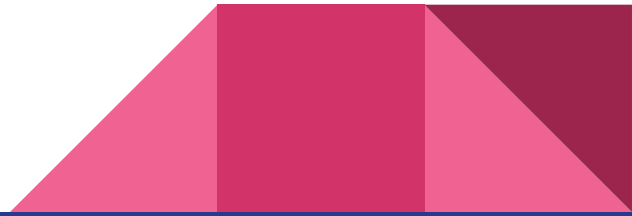
To become a Housetradamus

By:
Richard Ling
4/10/2020

Outline

- What are we doing?
- Data Overview
- Data Cleaning
- Data Processing
- Outcomes
- Comments/Questions



What are we doing?

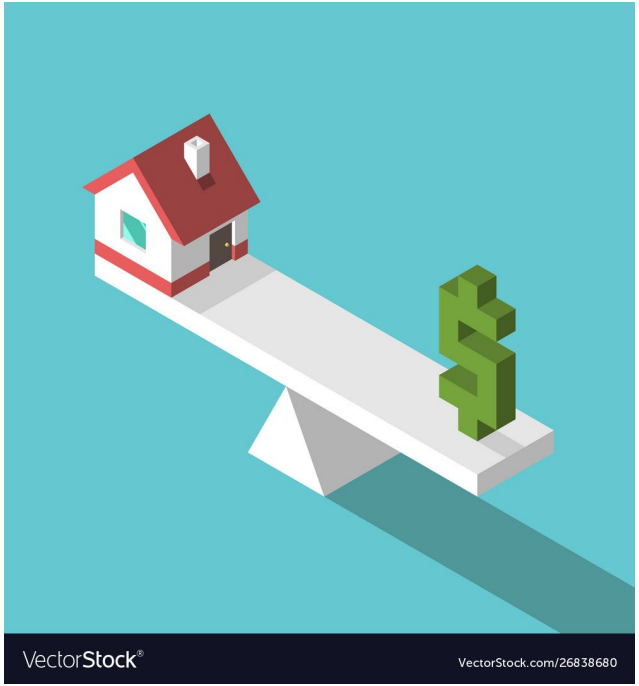
Data from 2006 to 2010

Goal:

Important attribute for houses

and

Predict house prices!



Data Overview

Total 81 features:

38 Quantitative (condition rating, lot area, bedroom counts, etc)

2 IDs (ID # for government)

40 Qualitative (house material, foundations, etc)

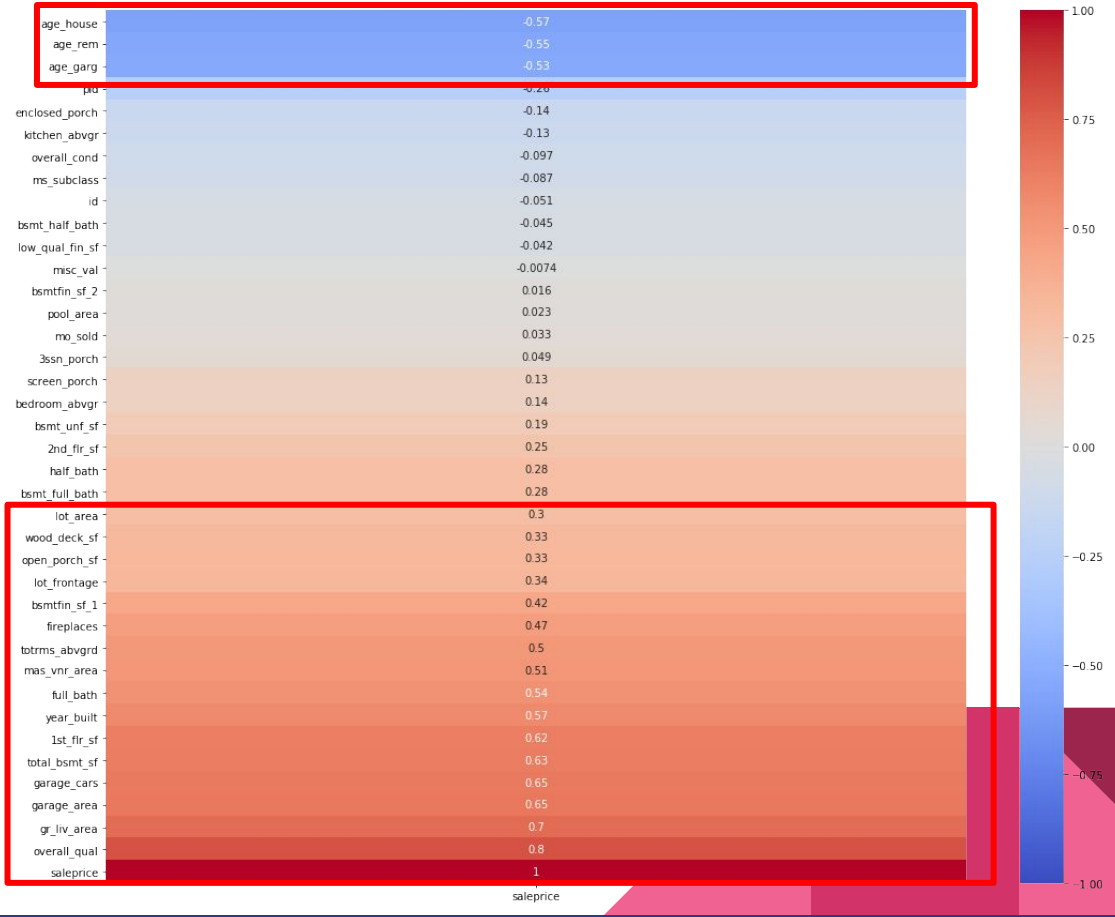
Question: These features all affect the sale price?

Answer:

Data Cleaning - Quantitative

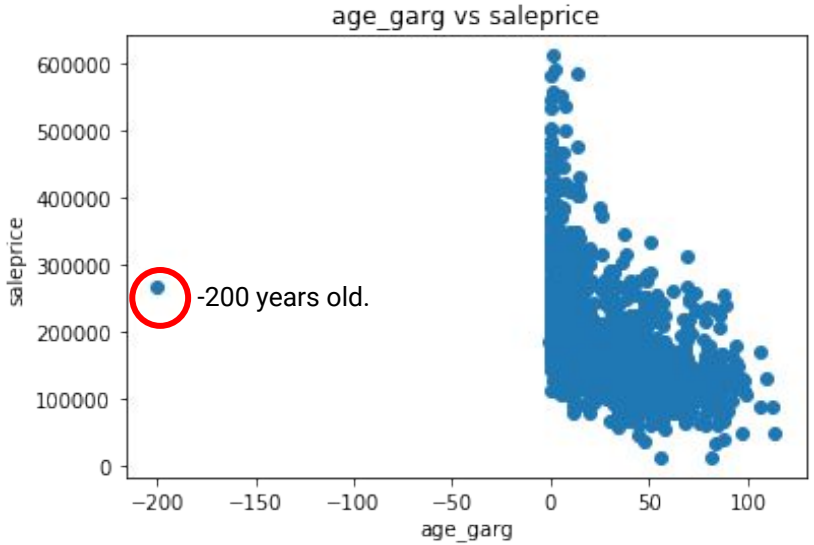
1) Ignore IDs
2) Turn years into age.
3) Heatmap for correlation > 0.30

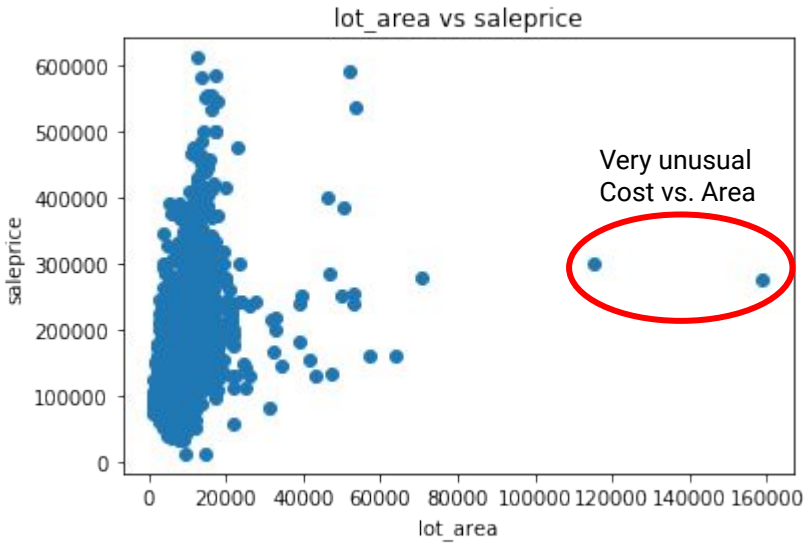
Result : 20 features



Data Cleaning - Quantitative

4) Scatter plot and filter outliers. For example:



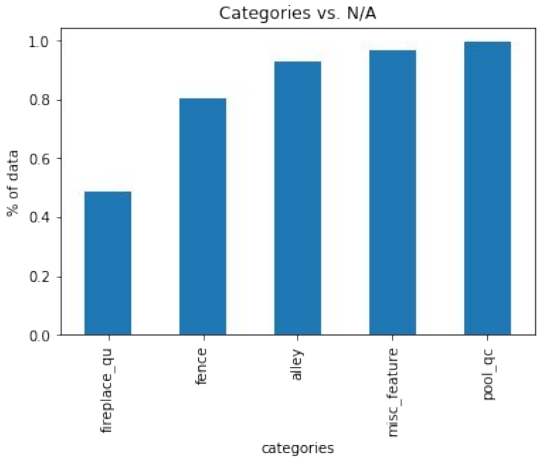


5) Check the NAs:

lot_frontage: 330.
Fill with the means of its neighborhood

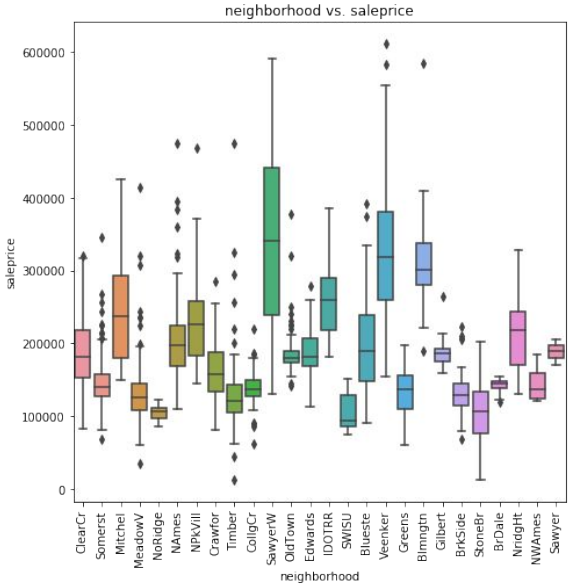
Data Cleaning - Qualitative

Drop categories with high % NAs.



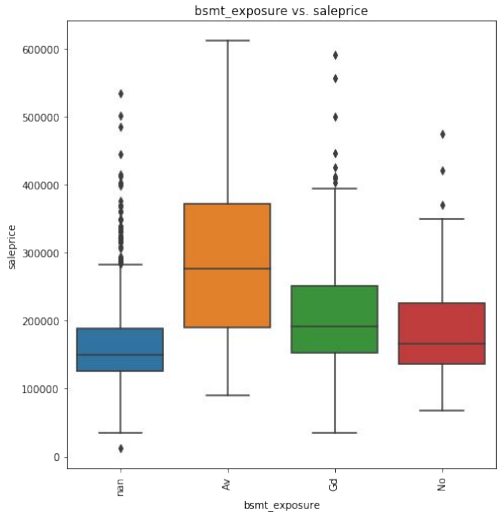
Basically entire column is the same.

Boxplot and choose high variations.



A lot of price change.

Ignore small price change.



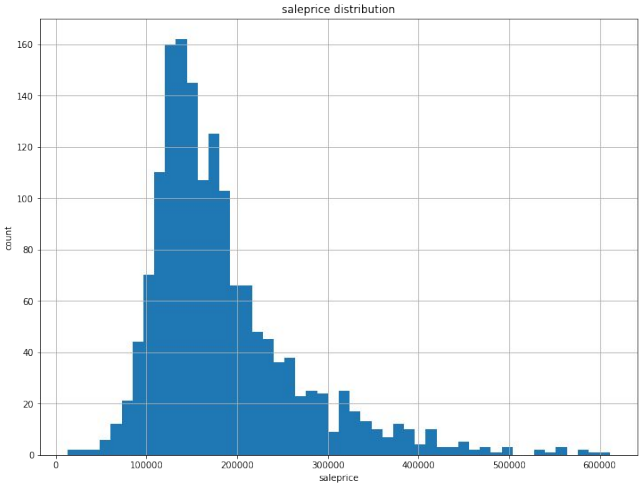
Result: 20 Features

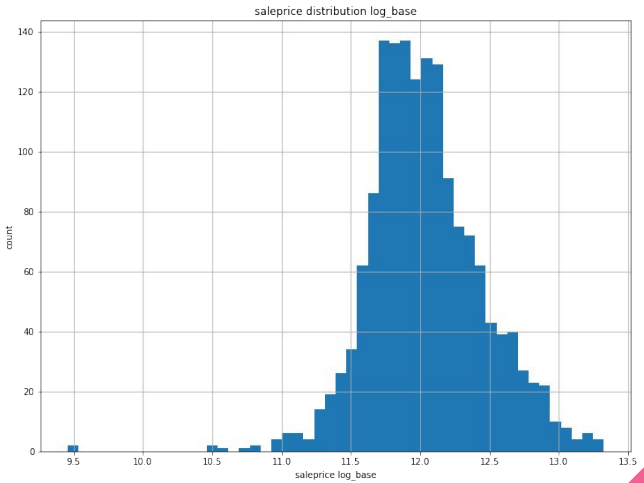
Data ready, set, go!

#1 Setup: X = 40 features
y = sale price

#2 One-hot encoding on qualitative.
Result: X total of 150 features.

#3 log transformation. For example:





Normal distributed except for 1 data point. Not a outlier.

Data ready, set, go!

75/25 train test split. Default.

Results:

cv_score (5 fold)	Train	Test	Kaggle Score
Linear Regression	0.914	0.856	23.83 K
Ridge	0.954	0.911	23.80 K
LassoCV	0.947	0.920	23.28 K

Scores about the same for train and test.

Ridge and Lasso adjusted slightly on some features.

Outcome

Model:

- Capable of prediction with given features.

Limitations:

- ONLY for ~2006 to ~2010
- SIMILAR geographical location
- SIMILAR economic condition

Improvements:

- More iterations.
- Separate categories.

We became a Housetradamus!

Thank you!

Comments / Questions?