Ames Housing Data Challenge

A Statistical Investigation on Factors that Influence the Home Value

Kai Zhao

Agenda

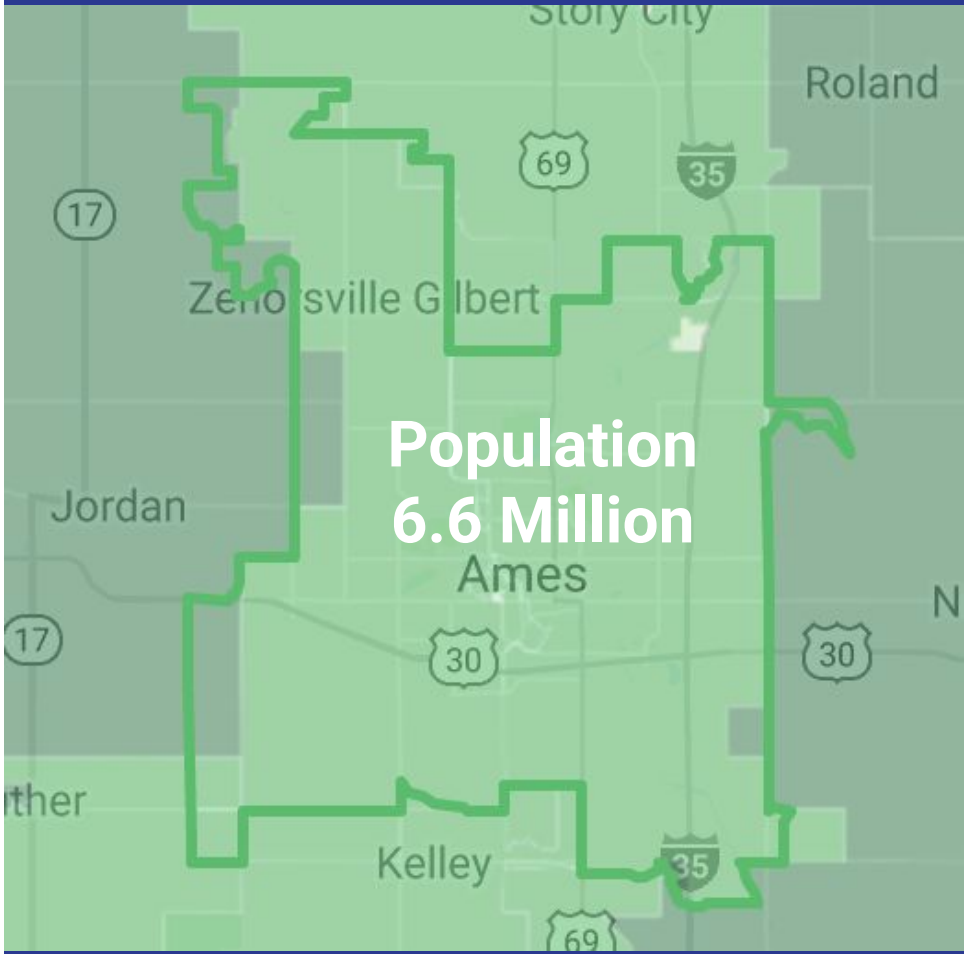
1. Purpose of This Study
2. Background & Data
3. Methodology
   a. EDA & Data Cleaning
   b. Feature Engineering
   c. Data Transformation
   d. Modeling
4. Results
5. Conclusion & Discussion

Picture Source: shutterstock

Agenda

1. Purpose of This Study
2. Background & Data
3. Methodology
   a. EDA & Data Cleaning
   b. Feature Engineering
   c. Data Transformation
   d. Modeling
4. Results
5. Conclusion & Discussion

Picture Source: shutterstock



The Sales Data

Features (columns): 78

Training Data (rows): 2,051

Testing Data (rows): 879

Time: 2006 - 2010

Picture Source: shutterstock

Features

18 Continuous	14 Discrete	23 Nominal	23 Ordinal
Measurable	*Countable*	*Categorical*	*Rankable*
Lot Area	Year/Month	Street/Alley	Quality
Gr Liv Sqft	Garage Car	Neighborhood	Condition
Garage Area	Full Bath	Bldg Type	Slope

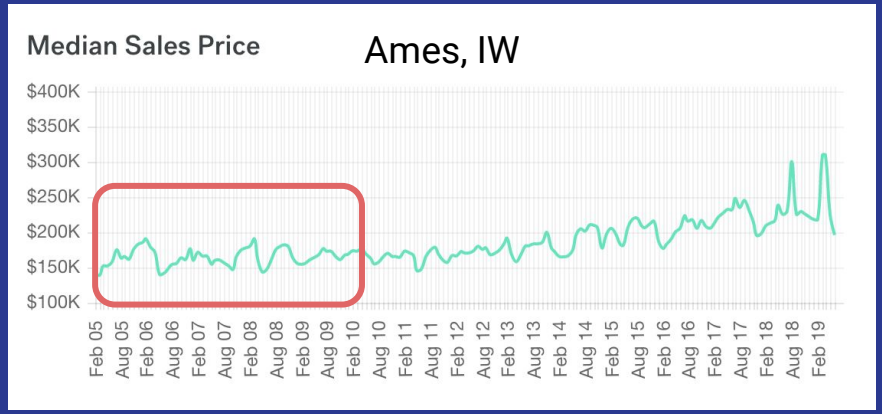
The Sales Data

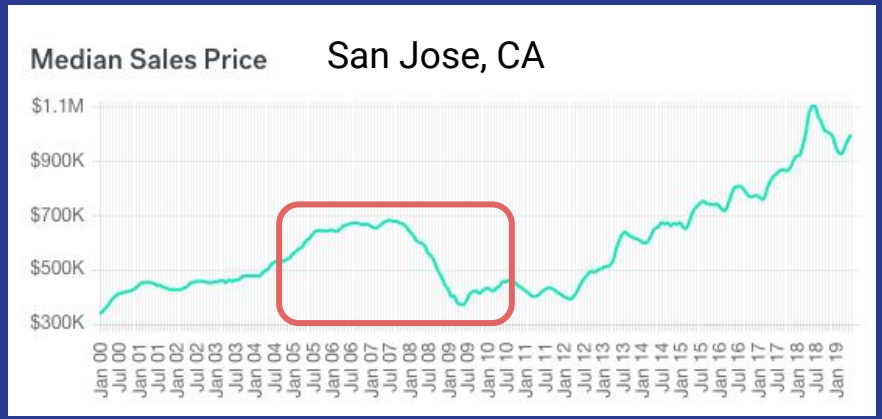
Features (columns): 78

Training Data (rows): 2,051

Testing Data (rows): 879

Time: 2006 - 2010

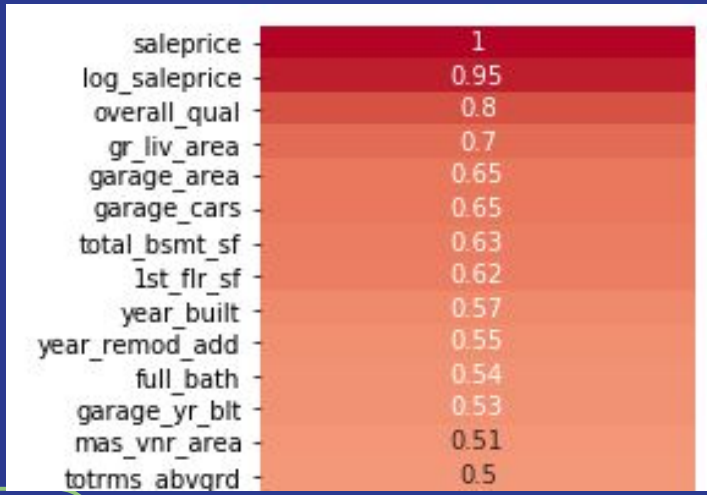


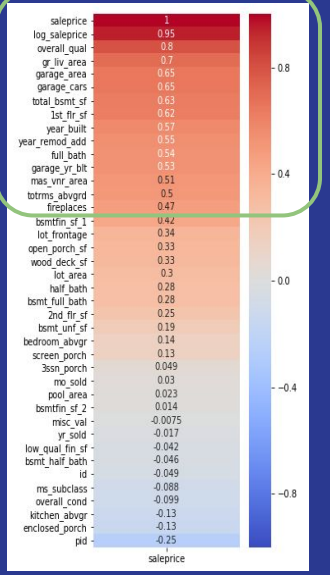


Picture Source: shutterstock

Agenda

1. Background & Data
2. Purpose of the Study
3. Methodology
   a. EDA & Data Cleaning
   b. Feature Engineering
   c. Data Transformation
   d. Model Selection
4. Results
5. Conclusion & Discussion





Data Cleaning

```
if type(feature) in ['Continuous', 'Discrete']:

    if "NaN" == "missing":

        feature.replace("NaN", 0)

    else: # No Value

        feature.ignore()
else: # Nominal and Ordinal

    if ("NaN" == "missing") and 'Nan'.count() > 10:

        feature.ignore()

    else: # No Value

        'NaN'.delete()
```

Index	Fture	Missing Value	%	Note	Action
4	lot_frontage	330	16.09	No Value	ignore feature
7	alley	1911	93.17	no alley access	n.a.
26	mas_vnr_type	22	1.07	missing Value	ignore feature
27	mas_vnr_area	22	1.07	missing Value	ignore feature
31	bsmt_qual	55	2.68	no basement	n.a.
32	bsmt_cond	55	2.68	no basement	n.a.
33	bsmt_exposure	58	2.83	55 no basement, 3 missing values, id: 1797, 67, 2780	delete: 1797, 67, 2780
34	bsmtfin_type_1	55	2.68	no basement	n.a.
35	bsmtfin_sf_1	1	0.05	id: 1342, no basement, should be 0	correct to 0
36	bsmtfin_type_2	56	2.73	55 no basement, 1 missing value (id. 445)	delete: 445
37	bsmtfin_sf_2	1	0.05	id: 1342, no basement, should be 0	correct to 0
38	bsmt_unf_sf	1	0.05	id: 1342, no basement, should be 0	correct to 0
39	total_bsmt_sf	1	0.05	id: 1342, no basement, should be 0	correct to 0
48	bsmt_full_bath	2	0.1	id: 1342 & 1498, no basement, should be 0	correct to 0
49	bsmt_half_bath	2	0.1	id: 1342 & 1498, no basement, should be 0	correct to 0
58	fireplace_qu	1000	48.76	no fireplaces	n.a.
59	garage_type	113	5.51	no garage	n.a.
60	garage_yr_blt	114	5.56	113 no garage, 1 missing value, id: 2237	delete: 2237
61	garage_finish	114	5.56	113 no garage, 1 missing value, id: 2237	delete: 2237
62	garage_cars	1	0.05	id: 2237, missing value	delete: 2237
63	garage_area	1	0.05	id: 2237, missing value	delete: 2237
64	garage_qual	114	5.56	113 no garage, 1 missing value, id: 2237	delete: 2237
65	garage_cond	114	5.56	113 no garage, 1 missing value, id: 2237	delete: 2237
73	pool_qc	2042	99.56	no pool	n.a.
74	fence	1651	80.5	no fense	n.a.
75	misc_feature	1986	96.83	no misc feature	n.a.

Feature Engineering

Features

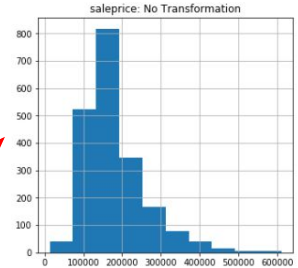
18 Continuous	14 Discrete	23 Nominal	23 Ordinal
Measurable	*Countable*	*Categorical*	*Rankable*
Lot Area	Year/Month	Street/Alley	Quality
Gr Liv Sqft	Garage Car	Neighborhood	Condition
Garage Area	Full Bath	Bldg Type	Slope
Stay as Is	Dummify	Dummify	???
Age of the House			Discrete Weighted? / Dummify

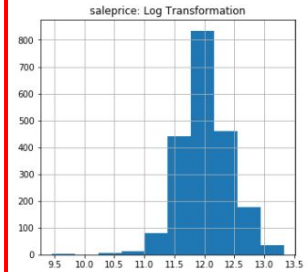
The Data Transformation

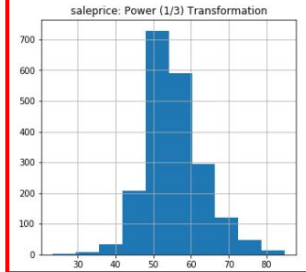
No Trans

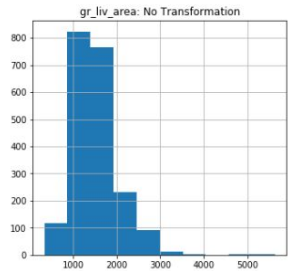
Log

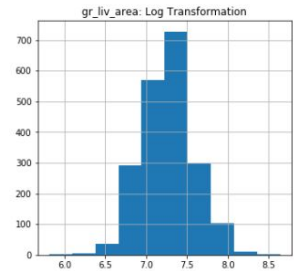
Cubic Root

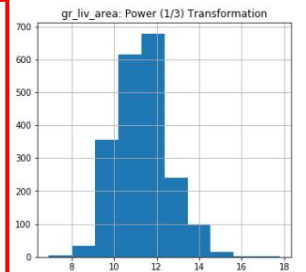


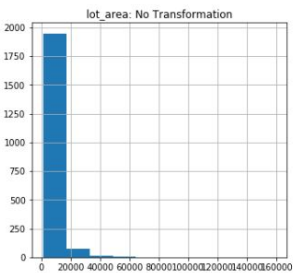


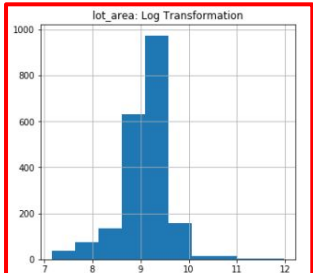


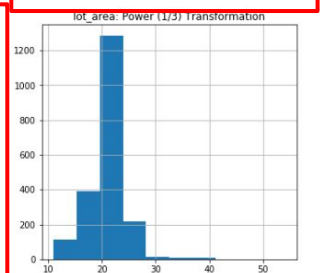




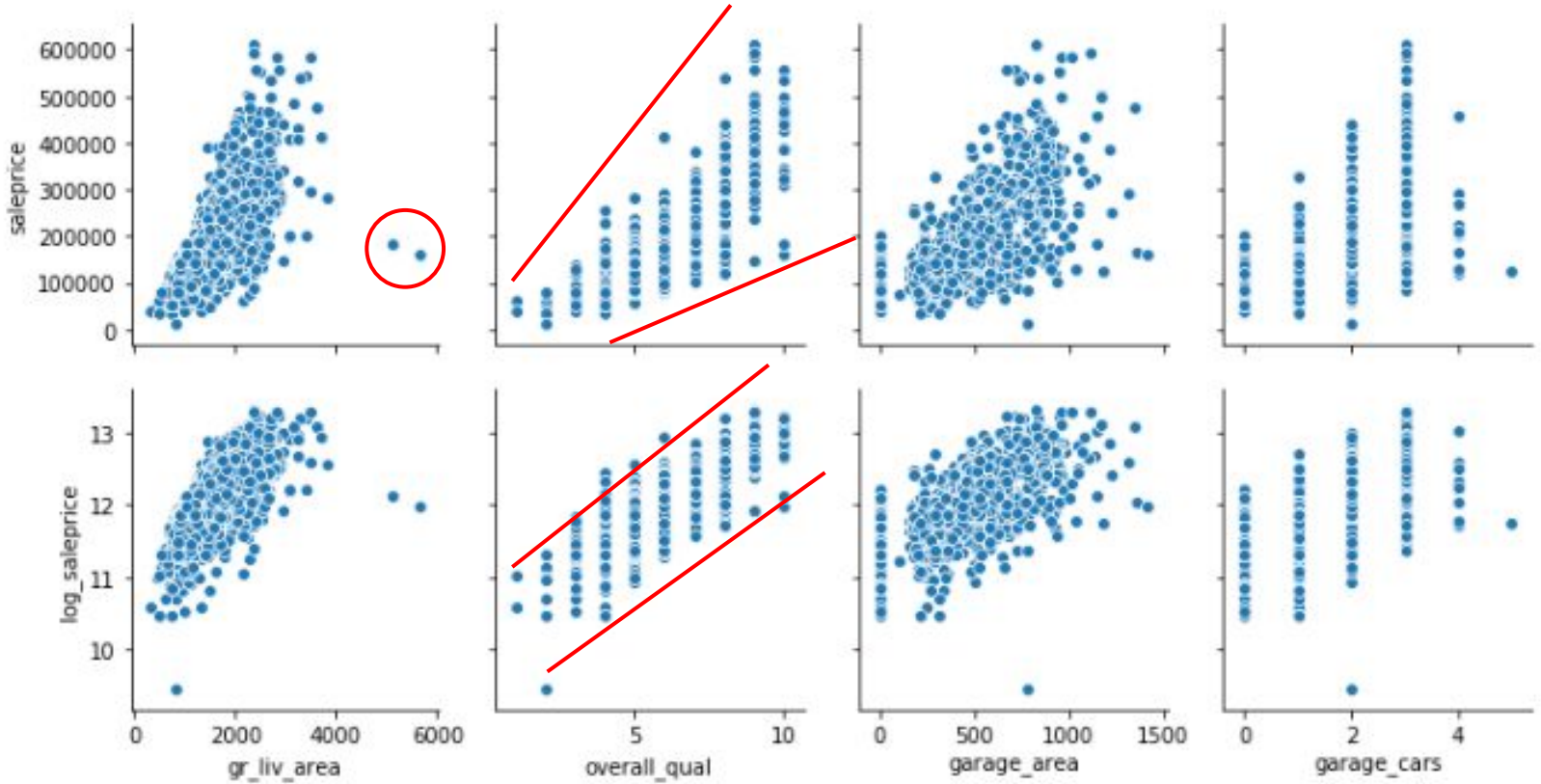








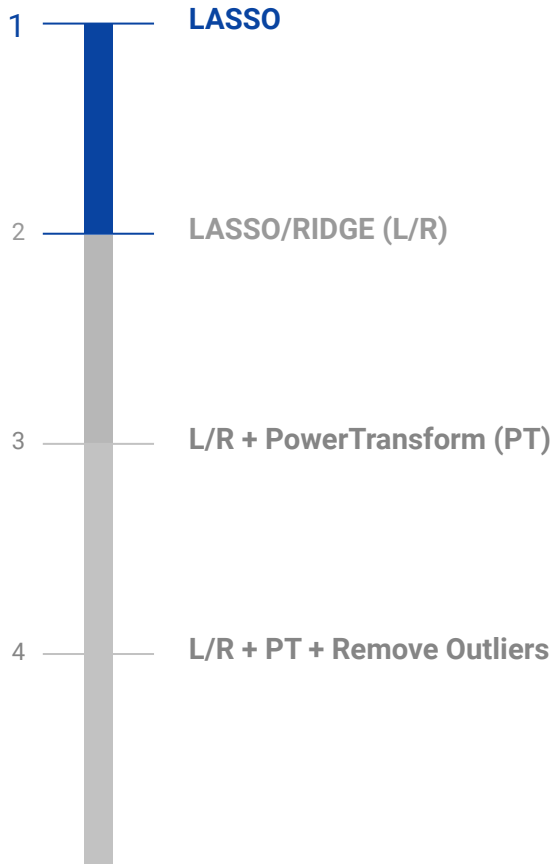
The Data Transformation



Modeling

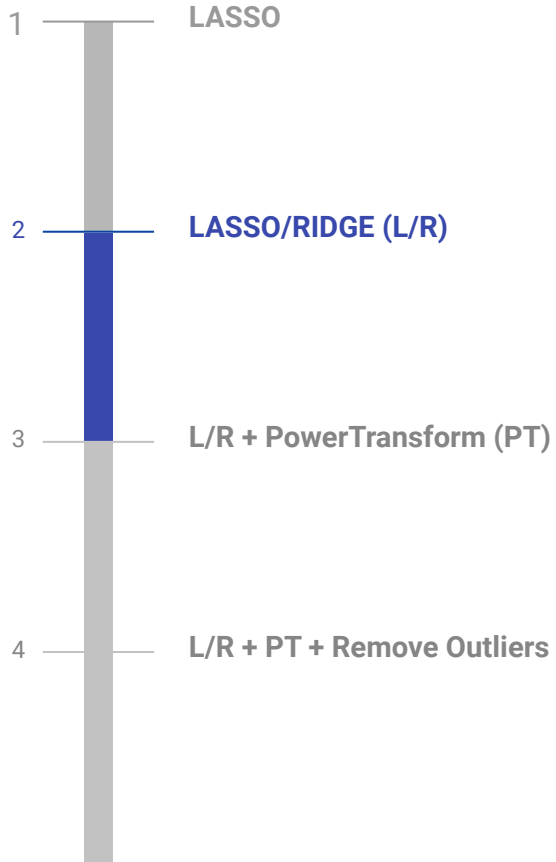
1. **LASSO**
2. LASSO/RIDGE (L/R)
3. L/R + PowerTransform (PT)
4. L/R + PT + Remove Outliers

Modeling



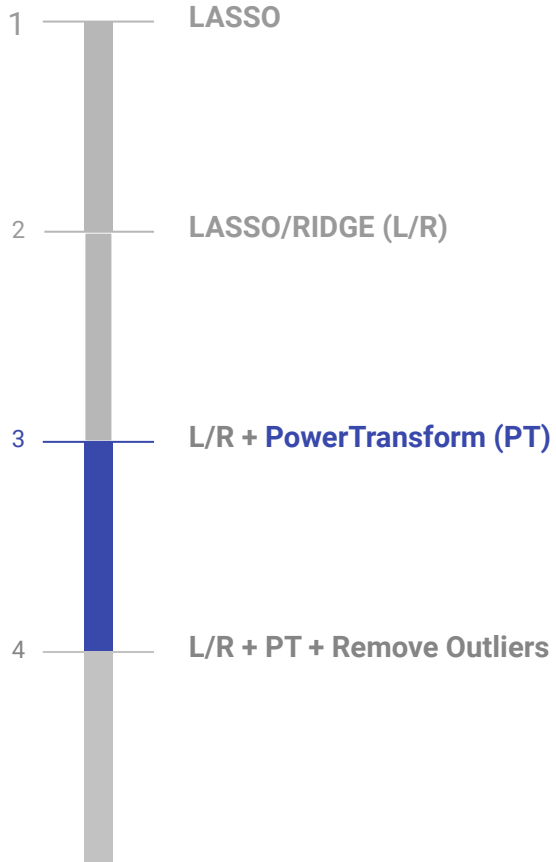
33,702

Modeling



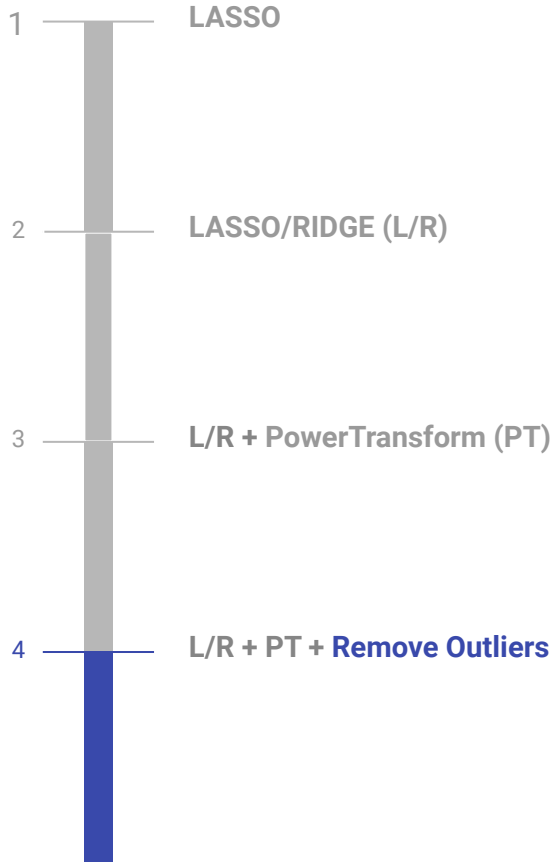
29,358

Modeling



23,180

Modeling



21,612

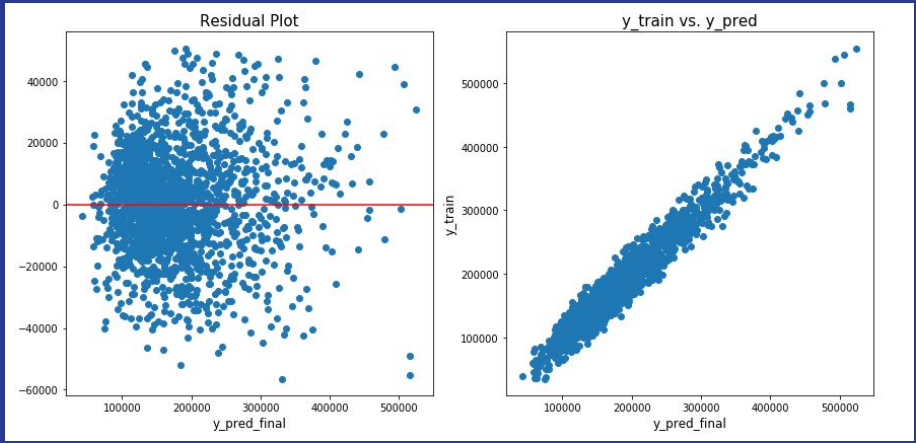
Agenda

1. Purpose of This Study
2. Background & Data
3. Methodology
   a. EDA & Data Cleaning
   b. Feature Engineering
   c. Data Transformation
   d. Modeling
4. Results
5. Conclusion & Discussion

Picture Source: shutterstock

Final Model: LASSO
R2 Score: 0.945
Training MSE: 17,987
Kaggle Score: 21,612

Features: 20
Outlier Deleted: 41



Conclusion & Recommendations

01	Gross Living Area	• **Owner/Seller**: maximize the sqft • **Buyer/Investor**: search one with addition potential
02	Total Basement Size Basement Finish	• **Size, Exposure, Finish** • **Owner/Seller**: enjoyment, long term value • **Buyer/Investor**: ROI (wortht ½ as above ground) Market: buyer vs. seller
03	Kitchen Quality	• **High Use & High Traffic (as bathroom)** • **Owner:** lifestyle, long term value • **Investor**: Watch out for local home price cap • Market: buyers vs. sellers
04	Lot Area	• **Adds potential to value increase** • **Due Diligence:** Permit, restriction, and easements
05	Overall Quality	• Keep good maintenance. • Seller's market: fix = additional money to pay (buyer) • Buyer's market: fix = sell cheaper

Conclusion & Recommendations - Basement

1. Walk-out Basement

```
(df_bsmt['open_porch_sf'] != 0).mean()
```

```
0.7638190954773869
```

Conclusion & Recommendations - Basement

2. Standard lot basement

Conclusion & Recommendations - Basement

3. Garden lot basement

Conclusion & Recommendations

01	Gross Living Area	• **Owner/Seller**: maximize the sqft • **Buyer/Investor**: search one with addition potential
02	Total Basement Size Basement Finish	• **Size, Exposure, Finish** • **Owner/Seller**: enjoyment, long term value • **Buyer/Investor**: ROI (wortht ½ as above ground) Market: buyer vs. seller
03	Kitchen Quality	• **High Use & High Traffic (as bathroom)** • **Owner:** lifestyle, long term value • **Investor**: Watch out for local home price cap • Market: buyers vs. sellers
04	Lot Area	• **Adds potential to value increase** • **Due Diligence:** Permit, restriction, and easements
05	Overall Quality	• Keep good maintenance. • Seller's market: fix = additional money to pay (buyer) • Buyer's market: fix = sell cheaper

Conclusion & Recommendations - Kitchen

Home Feature Keyword (All homes)	Effect (% homes sell for above expected values)	Most Common Metro
Steam oven	34%	Los Angeles, CA
Professional appliance	32%	Los Angeles, CA
Wine cellar	31%	Los Angeles, CA
Steam shower	31%	Chicago, IL
Pot filler	27%	Dallas, TX
Shed/Garage studio	26%	Los Angeles, CA
Heated floor	26%	New York, NY
Waterfall countertop	26%	Los Angeles, CA
Outdoor kitchen	25%	Dallas, TX
Prep sink	24%	Los Angeles, CA

Conclusion & Recommendations

01	Gross Living Area	• **Owner/Seller**: maximize the sqft • **Buyer/Investor**: search one with addition potential
02	Total Basement Size Basement Finish	• **Size, Exposure, Finish** • **Owner/Seller**: enjoyment, long term value • **Buyer/Investor**: ROI (wortht ½ as above ground) Market: buyer vs. seller
03	Kitchen Quality	• **High Use & High Traffic (as bathroom)** • **Owner:** lifestyle, long term value • **Investor**: Watch out for local home price cap • Market: buyers vs. sellers
04	Lot Area	• **Adds potential to value increase** • **Due Diligence:** Permit, restriction, and easements
05	Overall Quality	• Keep good maintenance. • Seller's market: fix = additional money to pay (buyer) • Buyer's market: fix = sell cheaper

Additional Study

1. Interaction between different features? (Feature Engineering)

2. How market change affect the valuation of features.

3. Study of human behavior: should some buyer info be included as a variable for better prediction?


THE DAILY SHOW
WITH TREVOR NOAH

Thank You!