The Price is (Always) Right

WHAT PRICE IS THE FAIR PRICE?

MARK YUNG, ANG JUN SIONG,
JONATHAN SOON, MERIKY LO ALEXANDER

MMJJ Realty: Iowa Chapter

Meriky

Realty Chief A♠

Never Been to Ames nor Iowa

Mark

Team Psychic

Needs More Mana

Jonathan

Intern K♥

Lowly Paid

Jun Siong

Swimmer 2♦

Perpetually Immersed (in Work or Water)

MMJJ Realty

BACKGROUND

You - High Profile Property Investors and Home Buyers

MMJJ have been hired to **estimate the price** for each of his **800+ properties**.

MMJJ REALTY

PROBLEM STATEMENT

What is the Fair Price?

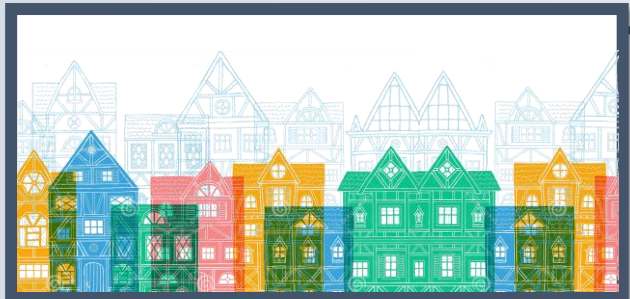
What are the Most Important Predictors of Home Prices? How?

Can we Predict Home Prices Accurately? How?

MMJJ REALTY

PROCEDURE & METHODOLOGY

Data Source

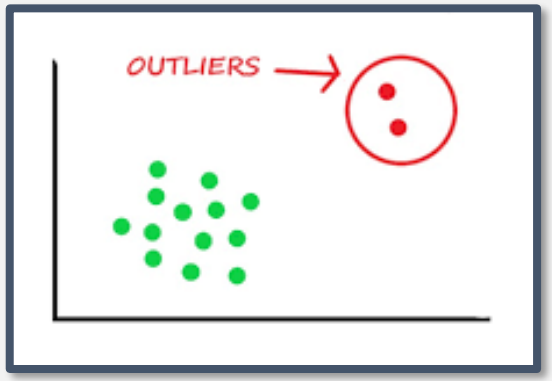


Ames Housing Data Set

- ~2000 Data Points for Home Sales Between 2006 to 2010
- 79 Numerical & Categorical Variables
- Go-to Introductory Regression Dataset

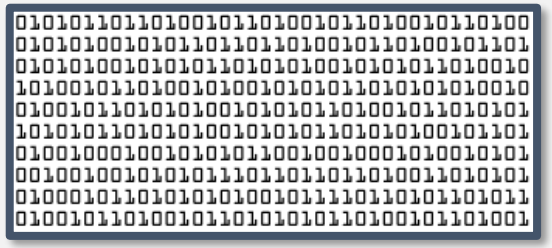
TRUMAN STATE UNIVERSITY

Basic Data Preparation



Remove Outliers

Remove Useless Data



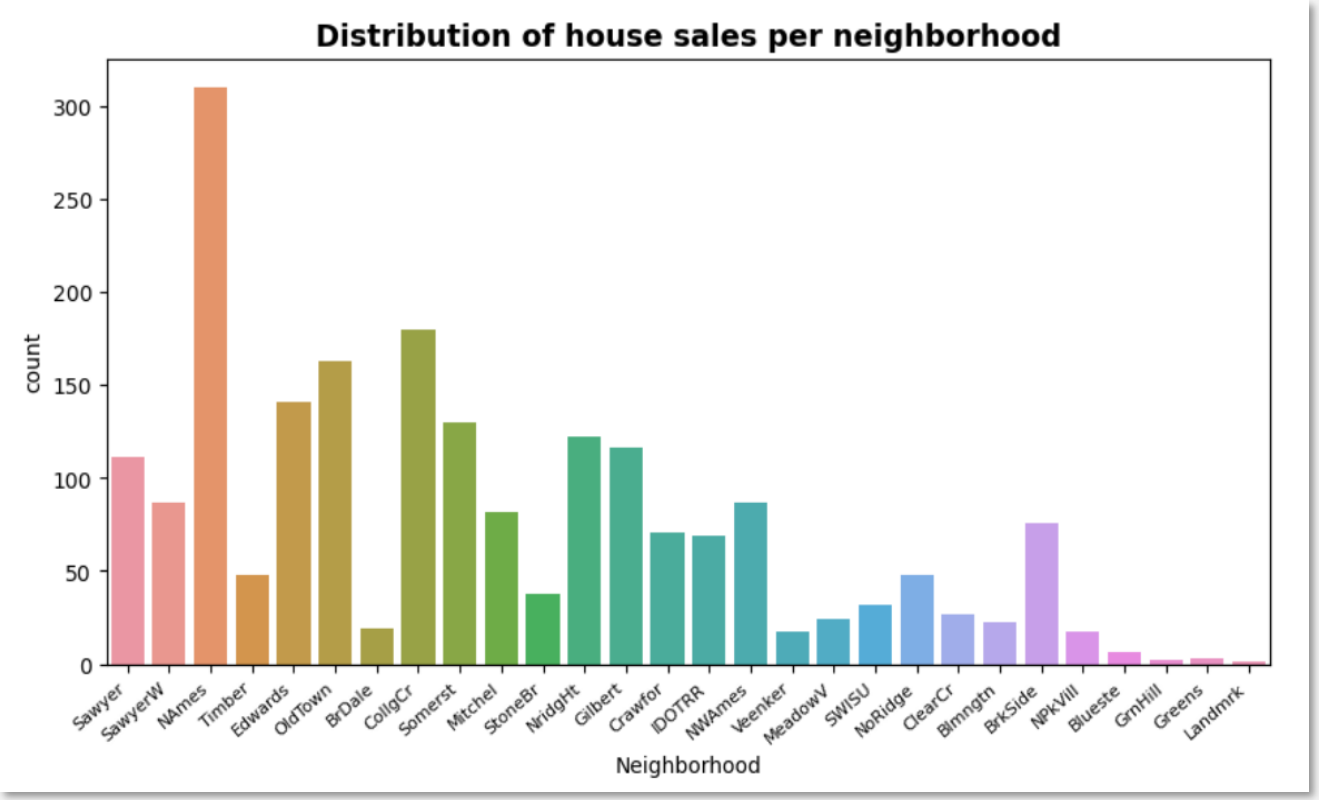
Data "Digitization"

Trim Sparse Variables

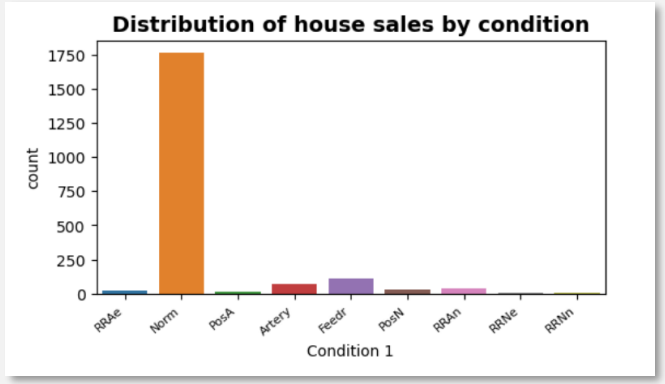
PROCEDURE & METHODOLOGY

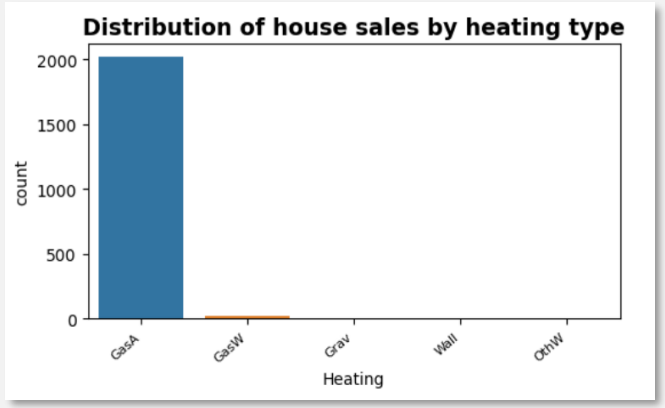
Exploratory Data Analysis

Understanding the Dataset



ID Low Variance Predictors





PROCEDURE & METHODOLOGY

Feature Engineering

Basement Size

	bsmtfin sf 1	bsmtfin sf 2	bsmt unf sf	total bsmt sf
0	533.0	0.0	192.0	725.0
1	637.0	0.0	276.0	913.0
2	731.0	0.0	326.0	1057.0
3	0.0	0.0	384.0	384.0
4	0.0	0.0	676.0	676.0

Garage Features

	garage score	garage finish	garage cond	garage qual	garage cars
0	144	3	4	4	3
1	144	3	4	4	3
2	64	2	4	4	2
3	192	4	4	4	3
4	96	2	4	4	3

Year Remod Feature

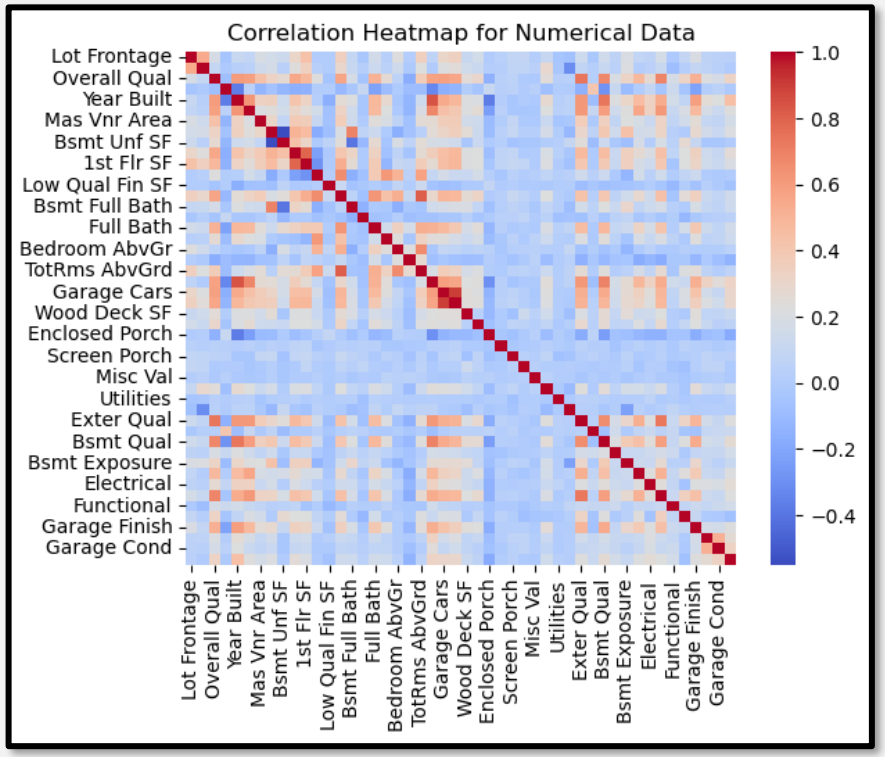
	age sold	year built	year remod/add	yr sold
0	5	1976	2005	2010
1	12	1996	1997	2009
2	3	1953	2007	2010
3	3	2006	2007	2010
4	17	1900	1993	2010

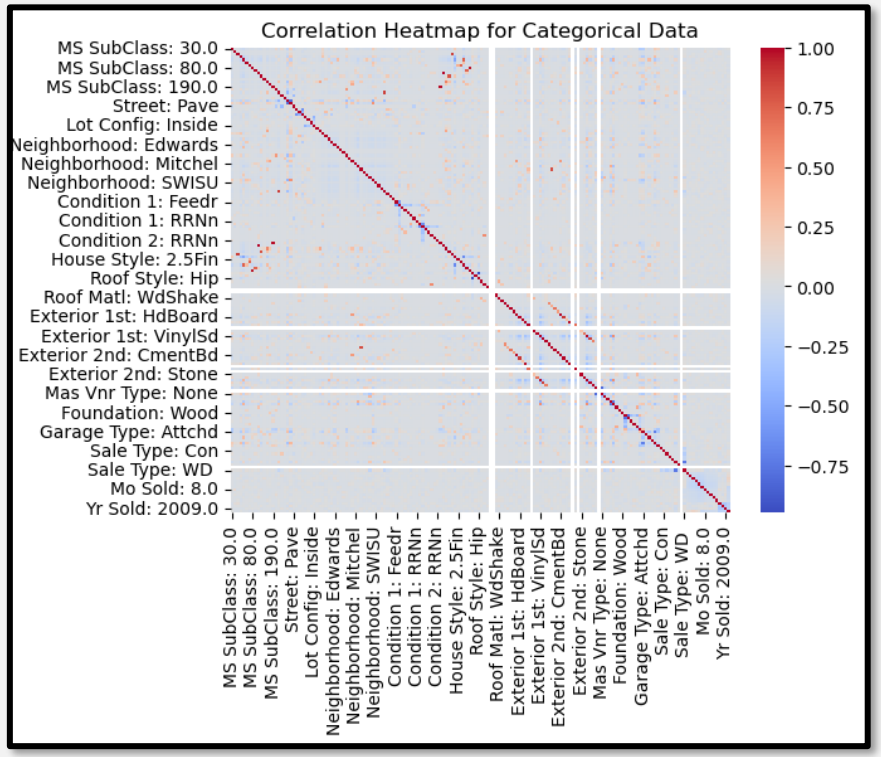
Floor SF Features

	1st flr sf	2nd flr sf	low qual fin sf	gr liv area
2043	1728	0	0	1728
2044	861	0	0	861
2045	1172	741	0	1913
2046	1200	0	0	1200
2047	1028	776	0	1804

PROCEDURE & METHODOLOGY

Variables Selection Stage 1: (Anti-) Multi-Collinearity





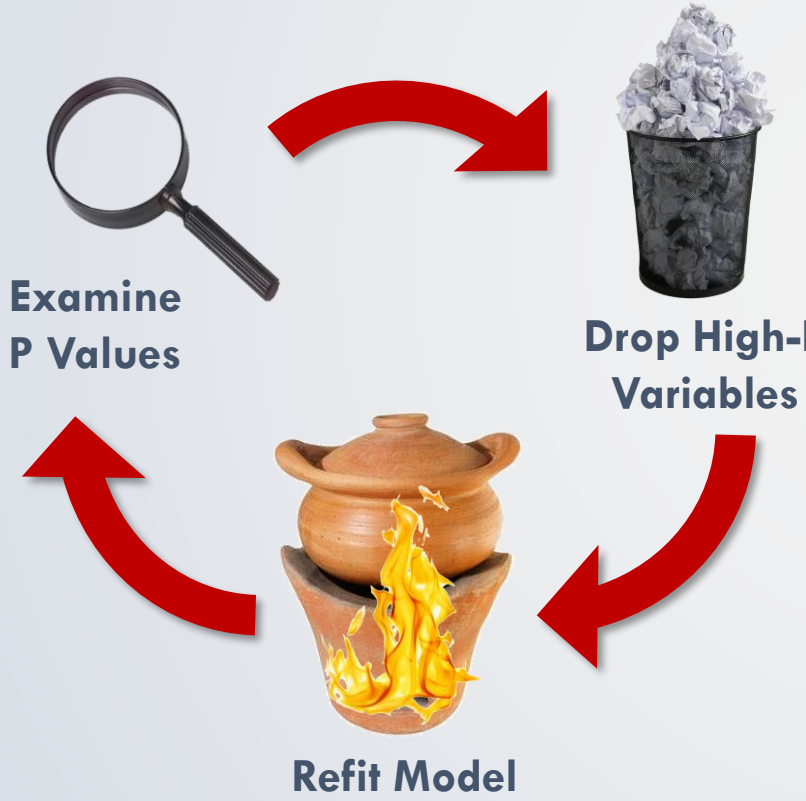
Removed

8

Predictor Variables

PROCEDURE & METHODOLOGY

Variables Selection Stage 2: Iterative Hypothesis Testing



Pick *Only* Significant Predictors

IN SUMMARY

79 Predictor Variables

From Initial Exploratory Data Analysis

70 Predictor Variables

(Anti-)Multi-Collinearity

49 Predictor Variables

Hypothesis Testing: 1st Fitting

41 Predictor Variables

Hypothesis Testing: 2nd Fitting

22 Predictor Variables

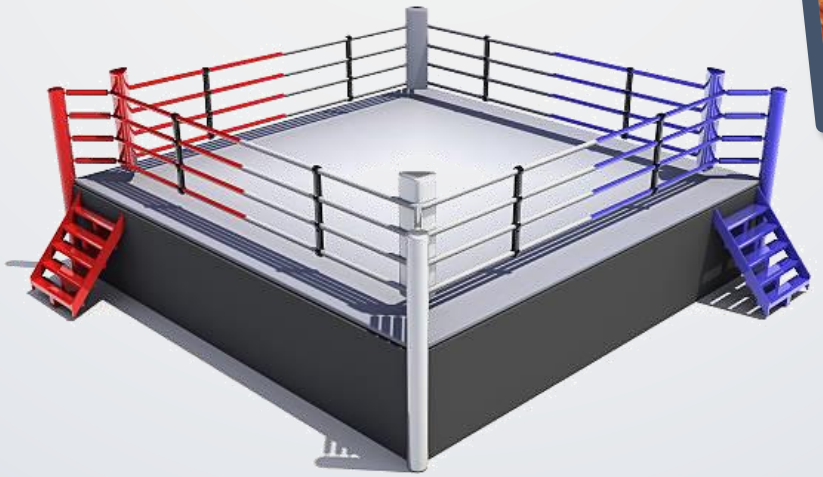
PRIMARY FINDINGS

WHAT PRICE IS THE FAIRPRICE?

MERIKY, JONATHAN, MARK, JUN SIONG

2000 Observations

22 Predictors



MMJJ REALTY

PRIMARY FINDINGS

MODELS SUMMARY

M1: Vanilla Model
(Ridge Regression)
Validation R^2: **0.91**
Kaggle RMSE: **$24,XXX**

10% **Observations Dropped for Having Empty Cells**

Id	Col 1	...	Col XX	Col XX	Col XX	Col XX	
1	Data	...	Data	Data	Data	Data	
2	MISSING		Data	Data	Data	Data	DROPPED
3	Data		MISSING	MISSING	Data	Data	DROPPED
4	Data		Data	Data	Data	Data	
...	...	...	...	...	...	...	
...	...	...	...	...	...	...	
YY	Data	...	MISSING	Data	Data	Data	
YY	Data	...	Data	Data	Data	Data	
YY	Data		Data	Data	Data	Data	

MMJJ REALTY

PRIMARY FINDINGS

MODELS SUMMARY

M1: Vanilla Model

(Ridge Regression)
Validation R^2: **0.91**
Kaggle RMSE: **$24,XXX**

M2: Drop Columns

(Ridge Regression)
Validation R^2: **0.87**
Kaggle RMSE: **$29,XXX**

4 **Variables Dropped for Having Empty Cells**

Id	Col 1	...	Col XX	Col XX	Col XX	Col XX
1	Data	...	Data	Data	Data	Data
2	Data		Data	Data	Data	Data
3	Data		MISSING	MISSING	Data	Data
4	Data		Data	Data	Data	Data
...	...	...	...	...	...	...
...	...	...	...	...	...	...
YY	Data	...	MISSING	Data	Data	Data
YY	Data	...	Data	Data	Data	Data
YY	Data		Data	Data	Data	Data

MMJJ REALTY

PRIMARY FINDINGS

MODELS SUMMARY

M1: Vanilla Model
(Ridge Regression)
Validation R^2: **0.91**
Kaggle RMSE: **$24,XXX**

M2: Drop Columns
(Ridge Regression)
Validation R^2: **0.87**
Kaggle RMSE: **$29,XXX**

M3: Impute Median
(Ridge Regression)
Validation R^2: **0.93**
Kaggle RMSE: **$21,XXX**

Median **Broad Imputation Method for Missing Data**

Id	Col 1	...	Col XX	Col XX	Col XX	Col XX
1	Data	...	Data	Data	Data	Data
2	Data		Data	Data	Data	Data
3	MEDIAN		MEDIAN	MEDIAN	Data	Data
4	Data		Data	Data	Data	Data
...	...	...	...	...	...	...
...	...	...	...	...	...	...
YY	Data	...	MEDIAN	Data	Data	Data
YY	Data	...	Data	Data	Data	Data
YY	Data		Data	Data	Data	Data

MMJJ REALTY

PRIMARY FINDINGS

MODELS SUMMARY

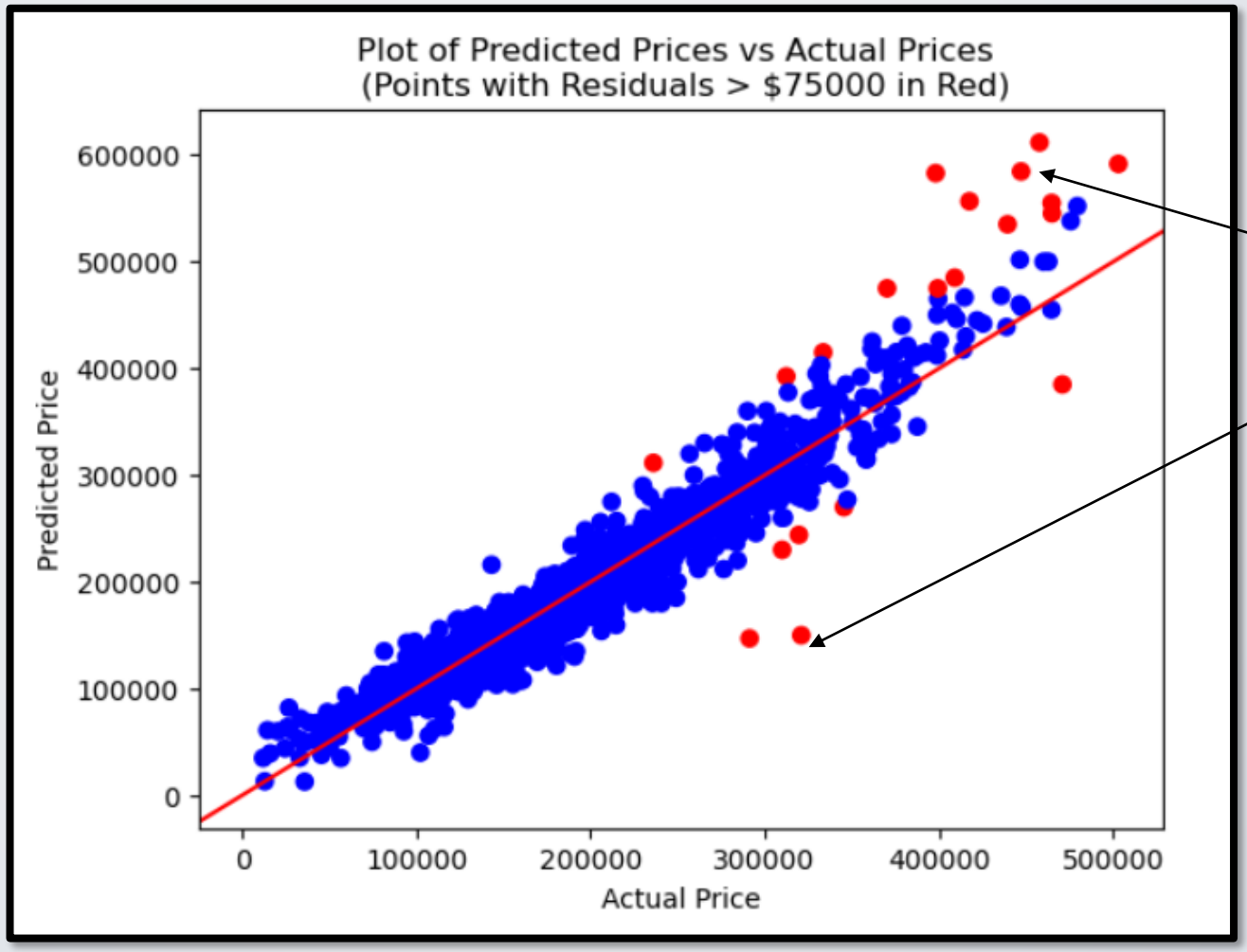
M1: Vanilla Model
(Ridge Regression)
Validation R^2: **0.91**
Kaggle RMSE: **$24,XXX**

M2: Drop Columns
(Ridge Regression)
Validation R^2: **0.87**
Kaggle RMSE: **$29,XXX**

M3: Impute Median
(Ridge Regression)
Validation R^2: **0.93**
Kaggle RMSE: **$21,XXX**

M4: Drop Outliers
(Ridge/ Lasso Regression)
Validation R^2: **0.94**
Kaggle RMSE: **$20,030**

Predicted Prices vs Actual Prices from M3



Model Refitted After Removing Outlying Observations (In Red)

MMJJ REALTY

CONCLUSION

Problem Statement Recap

What is the Fair Price?

What are the Most Important Predictors of Home Prices?

Can we Predict Home Prices Accurately?

Final Model Evaluation

0.94 R^2 + $20,030 Test RMSE on Kaggle

Most Important Variables in Predicting Home Prices

Neighborhood, Gross Living Area, Condition, Sales Type

Data Preparation Makes the Difference

Same Same (Dataset), But Vastly Different (Results)

MMJJ REALTY

RECOMMENDATIONS

Employ More Complex Models (GBR, Neural Nets)

- Unoptimized Gradient Boosting Regression: $18,XXX Test RMSE on Kaggle
- **BUT**, Involves Tradeoff in Interpretability

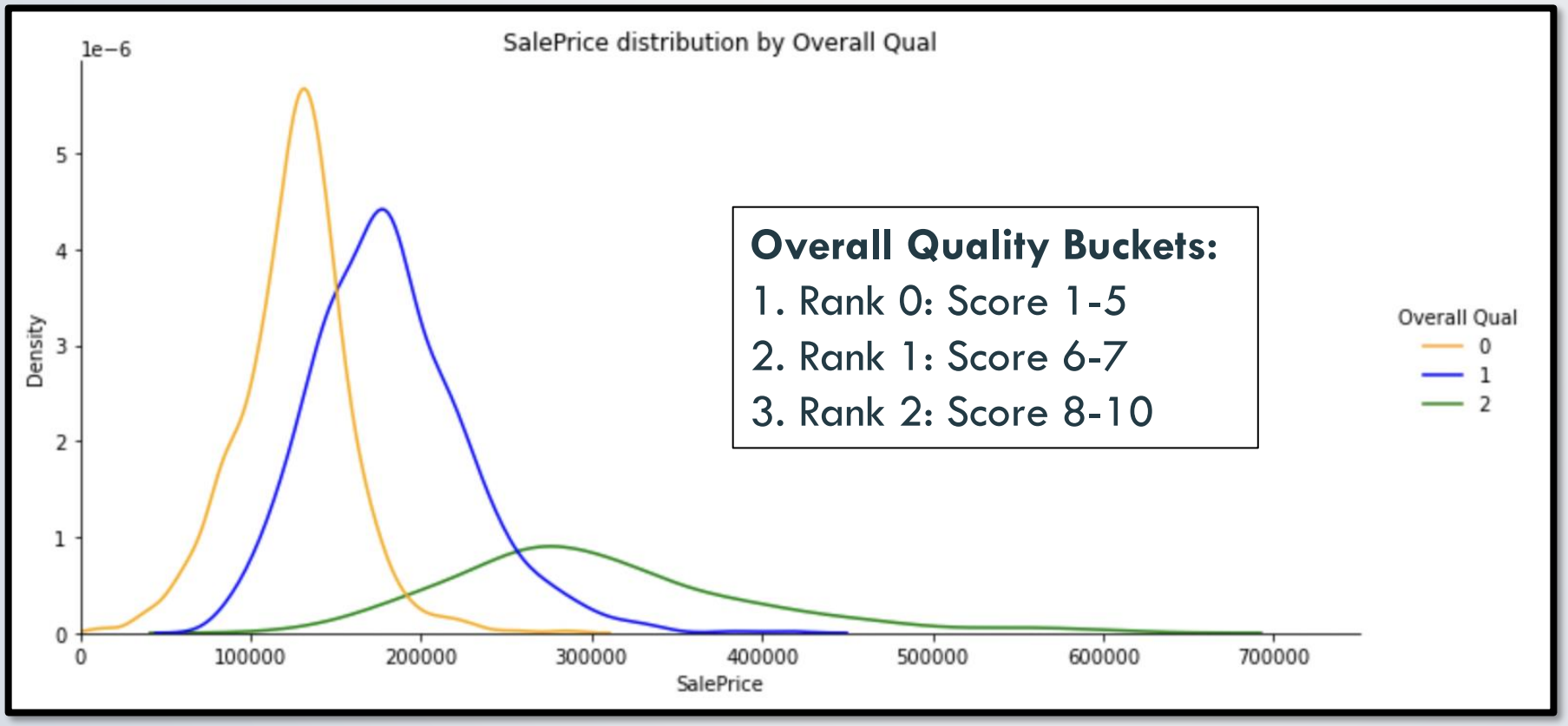
Bonus: Implement **Market Segmentation…**

MMJJ REALTY

MARKET SEGMENTATION

Market Segments by certain characteristics are affected by **different factors**. Oftentimes, these macro-effects translate into our models as fluctuations in feature *dependency*.

Translates into different factors and features **affecting model to varying degrees.**



MMJJ REALTY

MARKET SEGMENTATION

Tailored Models:
Different Market Segments, Different Behaviors

Segments have **different correlation behaviors** ➡ **BEST** Performance, with Enough Data

	OQ0	OQ0_corr	OQ1	OQ1_corr	OQ2	OQ2_corr
1	1st Flr SF	0.516779	Gr Liv Area	0.658482	Kitchen Qual	0.517902
2	Gr Liv Area	0.487283	Garage Area	0.541589	Gr Liv Area	0.513535
3	Total Bsmt SF	0.465055	Garage Cars	0.526667	Garage Cars	0.512368
4	Central Air_Y	0.439907	Full Bath	0.502532	TotRms AbvGrd	0.501728
5	BsmtFin SF 1	0.437947	1st Flr SF	0.499822	Exter Qual	0.501463
6	Paved Drive	0.410663	Total Bsmt SF	0.492332	Mas Vnr Area	0.496856
7	Year Built	0.408489	TotRms AbvGrd	0.440296	Bsmt Qual	0.495505
8	BsmtFin Type 1	0.402716	Kitchen Qual	0.431487	Garage Area	0.490732
9	Fireplaces	0.390328	Exter Qual	0.420268	Fireplace Qu	0.432934
10	Garage Area	0.385854	Year Remod/Add	0.407365	1st Flr SF	0.410804

RMSE values for each of the subsets:

1. Test RMSE(Overall Qual = 0): 14,550
2. Test RMSE(Overall Qual = 1): 18,713
3. Test RMSE(Overall Qual = 2): 42,278

MMJJ REALTY


OUR PRICE IS ALWAYS THE FAIR PRICE
NO ABSD, FOREIGNERS ELIGIBLE,
DM US ON SLACK NOW!!!

RECOMMENDATIONS

QUESTIONS?

MMJJ REALTY

RECOMMENDATIONS

THANKS!

MMJJ REALTY