Creating the next Ames

Annette Paciorek, Kisaki Watanabe, Nillia Ekoue, Colin Ford

Agenda

...Objective

...Approach

...Data Exploration

...Feature Engineering

...Model Selection

...Results

...Conclusion

Here is why you should care

Market-sizing assumptions:

_764k new homes 2019 * 1/1000...

_764 * $160k median SalePrice...

_$122M total sale value * 5% fee =

$6.1M Opportunity

The Plan:

Use Ames data to project sale prices in similar communities.

Sell that data to developers and agents.

We approached this opportunity with three guidelines:

1. Domain Knowledge

    Don't expect to re-invent the drivers of home price.

2. Visualization

    Pictures often show what raw data can't.

3. Machine Learning

    Validate our thinking with *interpretable* models.

Domain Knowledge

How does a real estate agent price her inventory?

How does a buyer value a home?

Assumed Factors

Size: interior/exterior

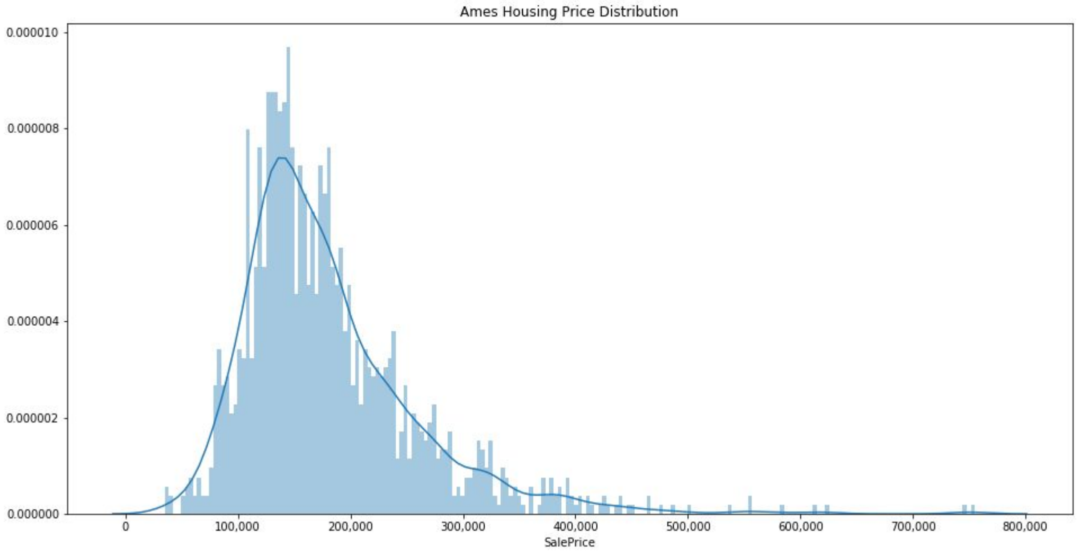
Location: proximity to school, work, crime

Quality: interior/exterior

Time: original build, remodel, sale date

Sale Price

When exploring the data, we first wanted to understand the distribution of our data:



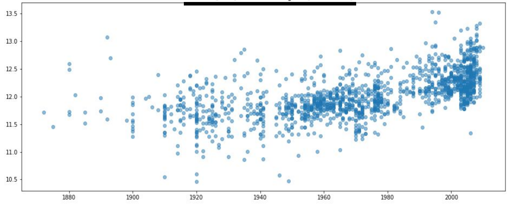
Takeaways

_Target customers build houses between **$130k-$214k**

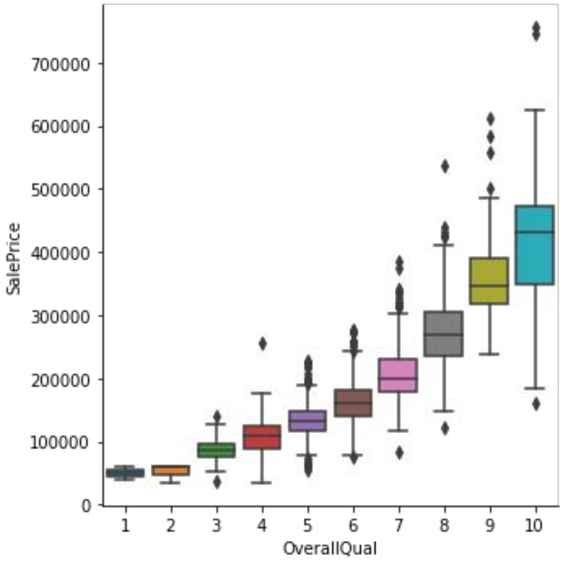
_**Few houses > $300k** led us to log transform Sale Price so that we could improve our prediction accuracy.

We then tested our **domain knowledge** by **visualizing** how variables changed with Sale Price:

Year Built

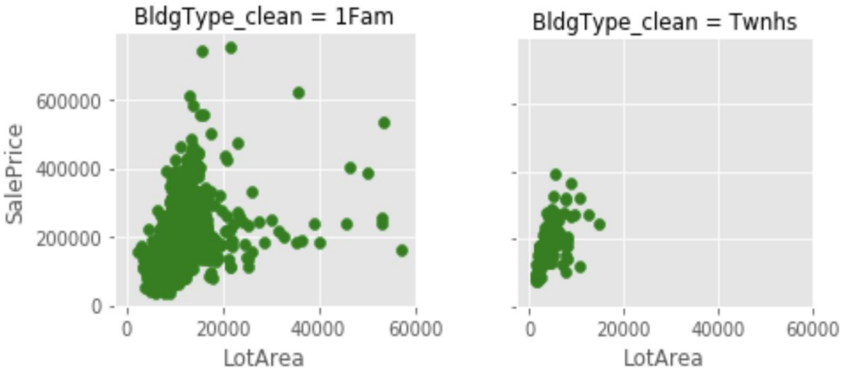


Quality



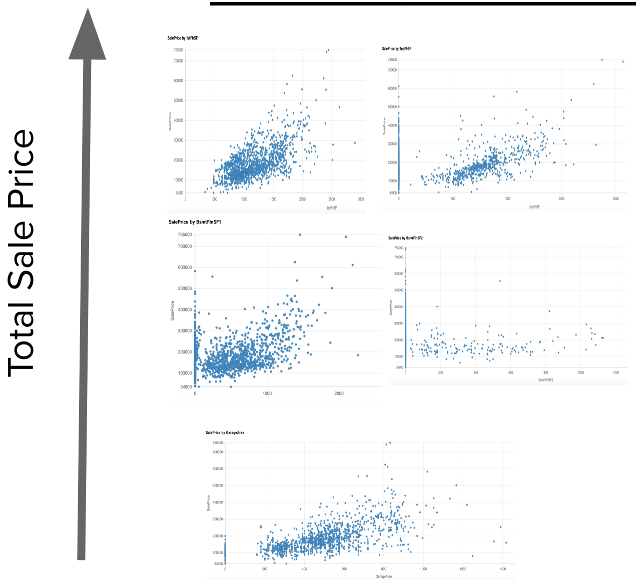
Lot Size

Yard matters for *Certain types* of dwellings

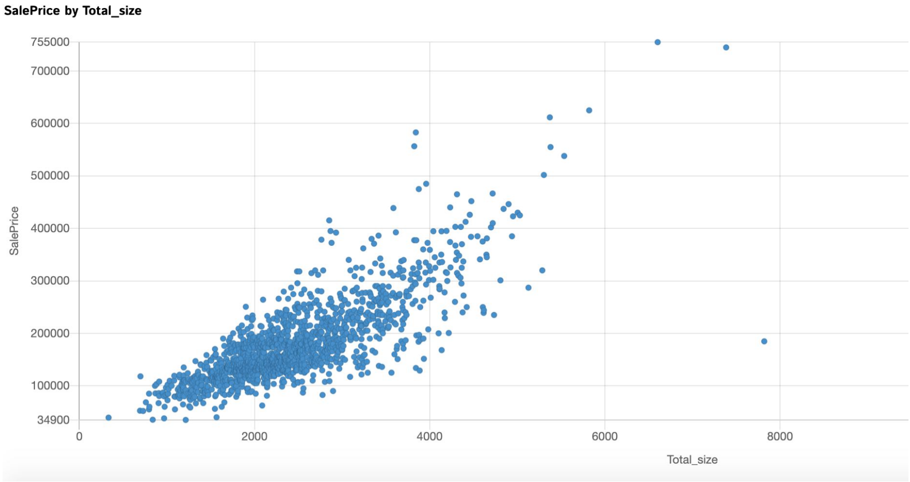


We aggregated several *interior* size features to reflect **domain knowledge*** about how size impacts price.

Individual Size Features



Total Size



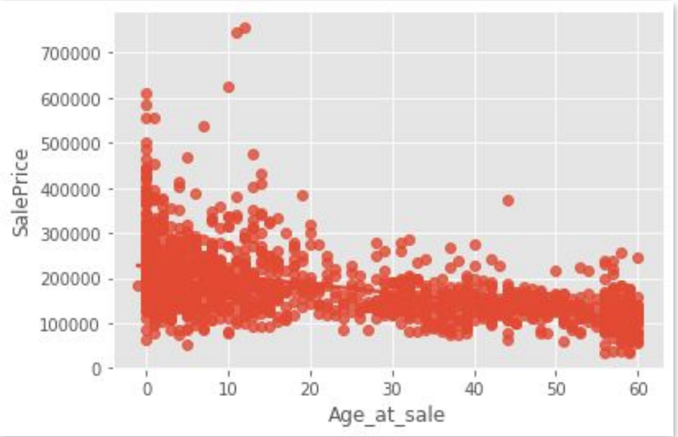
**Also avoids variance in model performance due to multicollinearity*

We also took creative liberty to create key price drivers we didn't immediately find in the data:

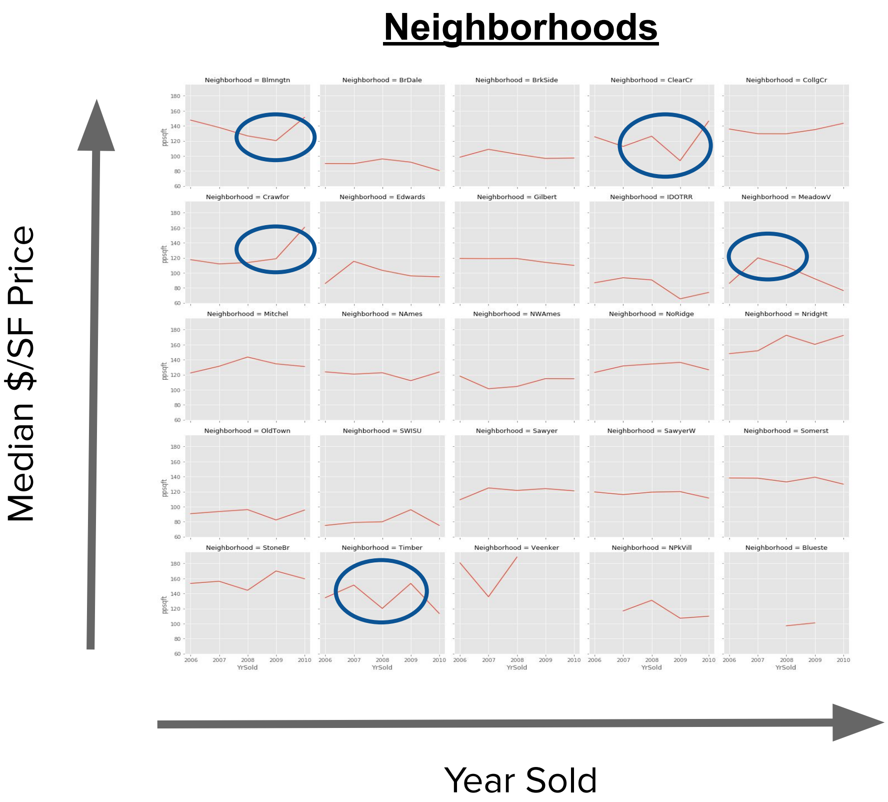
What we did

Engineer **Age @ Sale** column...
=
Lesser of **Remodel Date** & **Year Built...**
-
Sales Year

New column vs. Price



Some variables were significant only when grouped at a more granular level:



Takeaway

_Retained Sales Year in feature set

_Created binary variable for each neighborhood (multiple columns)

_Tested accuracy with and without Sales Year included

We ultimately created **2 feature sets**

Category	Feature-Heavy Set	Feature-Light Set
Location	Neighborhood	Neighborhood
Quality	Overall Quality	Overall Quality
Quality	Kitchen Quality	Kitchen Quality
Quality	Exterior Quality	Exterior Quality
Quality	Building Type	X
Quality/Size	Fireplaces	X
Size	Total Size	Total Size
Size	Lot Area	X
Size	Outdoor Porch Size	X
Time	Age at Sale	Age at Sale
Time	Year Sold	Year Sold

Each category of model has pros and cons.

	Simple Linear Regression	Mult. Linear + Automated Ft. Selection	Random Forests
Predictive Accuracy*	.23	.14	.15
Advantages	Fastest to implement, ease of understanding	Interpretability,predictive value,conservative feature selection	More resilient or stable against variability in the data.
Challenges/ Caveats	Sacrifice accuracy for simplicity; underfit	Higher bias - over-generalizes the relationship price and selected variables.	Expect high variance in predictive value.
Implementation Recommendation	Use to confirm coarse relationships, identify outliers and find avenues for further investigation	Use in production (external facing) to price houses for RE agents and developers.	Use internally to benchmark intuition, highlight price drivers among new neighborhoods.

Root mean squared error for each model* *(lower is better)***

Each type of model tested comes with advantages and disadvantages.

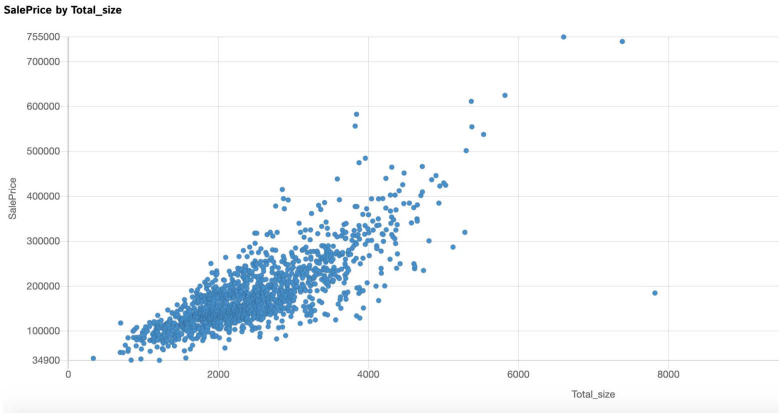
Model Complexity
Lowest to highest

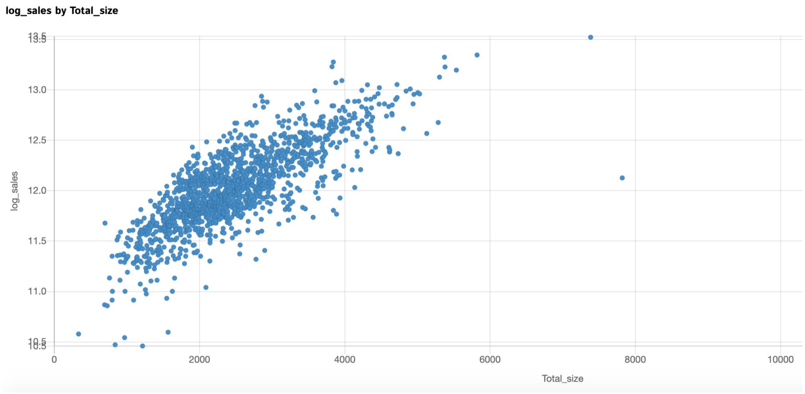
Model Description	Model Name	Feature Set to predict price	% of Variance Explained by the Model	Error of prediction vs. True Value
Basic Line graph	Simple Linear	Total Size only	0.644	0.238
Multiple variable linear	MLR	Heavy	0.87	0.14
Multiple variable linear	MLR	Light	0.87	0.14
Mult Variable Linear, Automates feature selection	AIC MLR	Heavy	0.86	0.164
Mult Variable Linear, Automates feature selection	AIC MLR	Light	0.89	0.14
Multiple variable linear + Penalty	Lasso	All	0.88	0.158
Multiple variable linear + Penalty	Lasso	Heavy	0.88	0.193
Multiple variable linear + Penalty	Lasso	Light	0.86	0.196
Decision Tree	Random Forest	All	0.973	0.164
Decision Tree	Random Forest	Heavy	0.978	0.152
Decision Tree	Random Forest	Light	0.977	0.162

Ask Us Questions

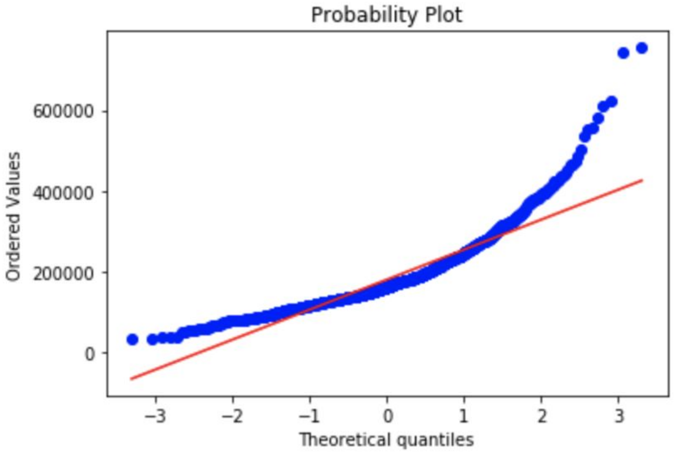
Appendix

When exploring the data, we first wanted to understand the range of our historical data along key variables:

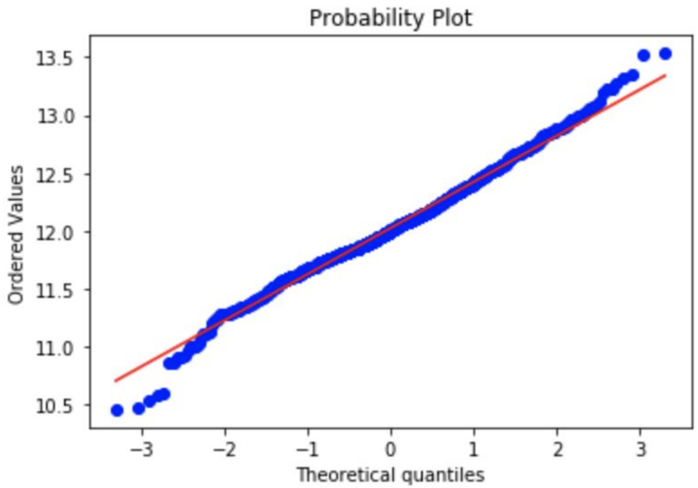




We log transformed our target variable in order to account for left skew of data:



Raw Sale Price Distribution



Log-transformed Sale Prices

Random Forest Details

	Original	Heavy	Light
No. of Features	18	11	7
Top 3 Features	Overall Quality, Ground Living Area, Garage Area	Total Size, Overall Quality, Age at Sale	Total Size, Overall Quality, Age at Sale
Parameters	max_features=2, min_samples_leaf=1, min_samples_split=2, n_estimators=100	max_features=2, min_samples_leaf=1, min_samples_split=2, n_estimators=100	max_features=2, min_samples_leaf=1, min_samples_split=2, n_estimators=100
RMSE (test)	0.164	0.152	0.162

Lasso Model Details

	Original	Heavy	Light
No. of Features	18	11	7
Top 3 Features	YearBuilt, Year RemodAdd, Ground Living Area	Total Size, Age at Sale,Outdoor Porch Size	Age at Sale,Total Size
Parameters	alpha tuned with gridsearchCV lasso2 = Lasso(warm_start = True, max_iter = 1e7) params = {'alpha':np.linspace(0.00162975083462060O, 0.001, 100)}	grid_search_lasso = GridSearchCV(estimator=lasso2, param_grid=params, cv=5)	alpha tuned with gridsearchCV
RMSE (test)	0.157	0.193	0.193

All

MLR & AIC Model Details

	MLR Heavy	MLR Light	Forward AIC Heavy	Forward AIC Light
No. of Features	11	7	10	6
Feature Importance	**+Overall Quality**	**+Overall Quality**	**-Year Sold**	**-Year Sold**
RMSE	.1433	.1403	.1705	.1403

Model Selection & Regularization

Model Selection:

- choosing the optimal model using AIC
- picking the model with the lowest RSS(or the highest R^2) via subset selection.

Regularization / shrinkage:

- Lasso: hyperparameter tuning using GridSearchCV
- Ridge:

Dimension Reduction Methods : Linear combination of predictors/ Random forests

- Principal Components Regression
- Partial Least Squares

Model Selection & Regularization

Dimension Reduction Methods : Linear combination of predictors/ Random forests

Highlight features importance

With the Lasso model, having 3 fireplaces and a kitchen in fair condition will drive down the price the price of a house while having 2 fireplaces will increase its value.

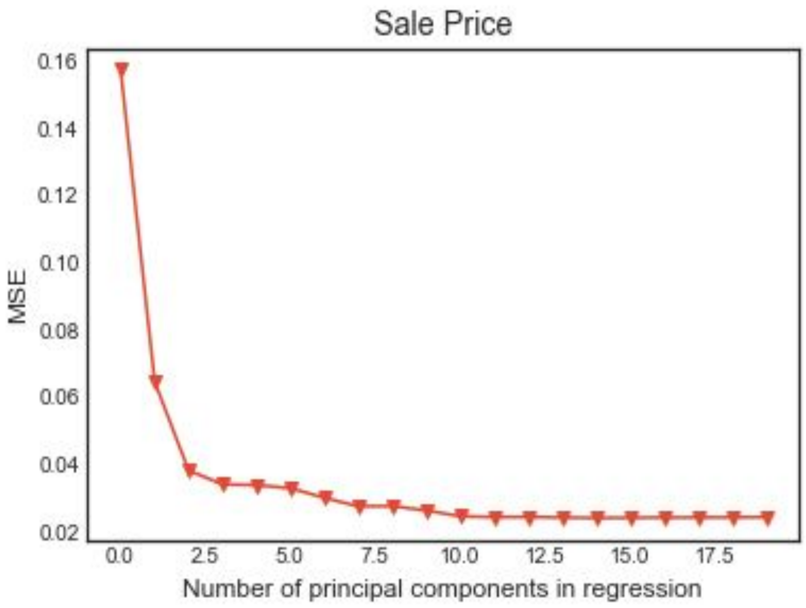
Prime Location: Nord Ridge Heights,

Motivation

“Let’s make millions

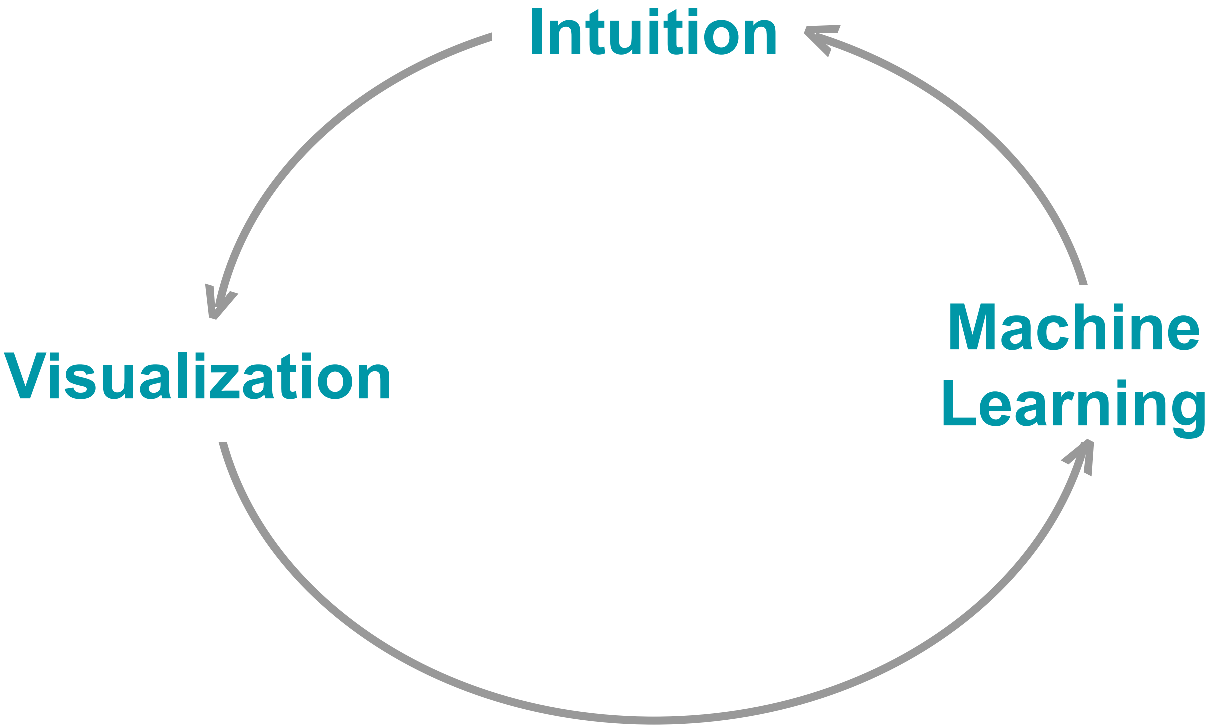
on the next Ames, Iowa.”(s)

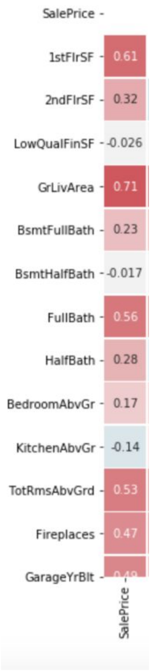
Features Selection

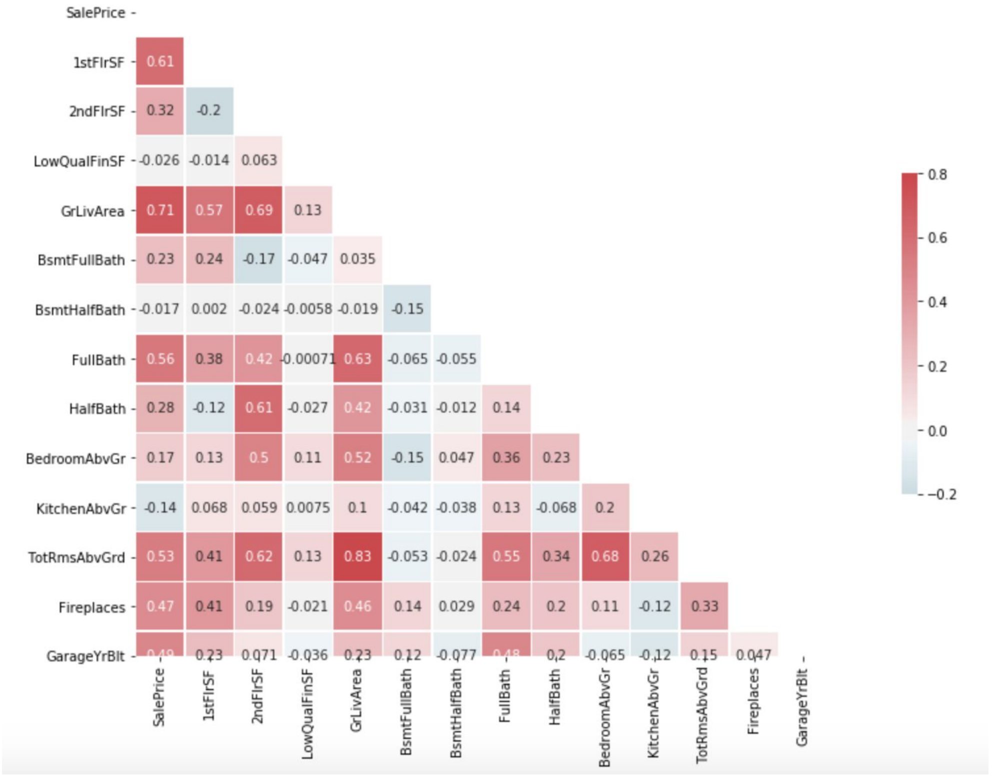


Maybe we can but Kisaki's Feature Importance graph here

These guidelines translated to the following iterative process:




SalePrice
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
0.61
0.32
-0.026
0.71
0.23
-0.017
0.56
0.28
0.17
-0.14
0.53
0.47


SalePrice
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
0.8
0.6
0.4
0.2
0.0
−0.2