Ames, Iowa

An exploration of house prices

GA DSI-20
Samuel Cheah

GENERAL ASSEMBLY

Photo from: Ames Chamber of Commerce

Housing prices are hard to understand

- Buyers want to know:
  - “Is this house being priced fairly?”
  - “If I only care about certain features, what is the price range I should expect?
- Sellers/developers want to know:
  - “How can I increase the value of my house?”
  - “If I want to develop a new project, where and how should I do it?”

Problem Statement

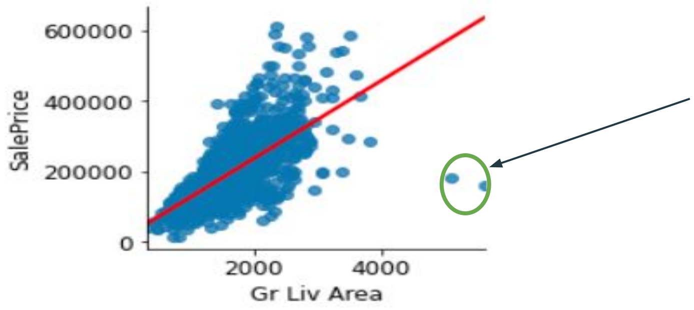
Can we use a regression model to predict house prices given a set of variables?

A Brief Walkthrough

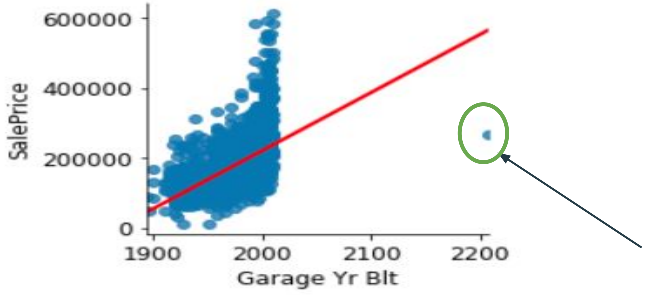
Steps taken to prepare data for modelling

Data Cleaning

Removing outliers
 - dropped houses >4000 sqft.



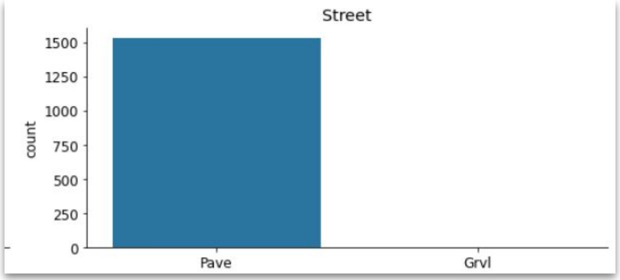
Fixing erroneous data
 - garages can't be built in the future



Data Cleaning

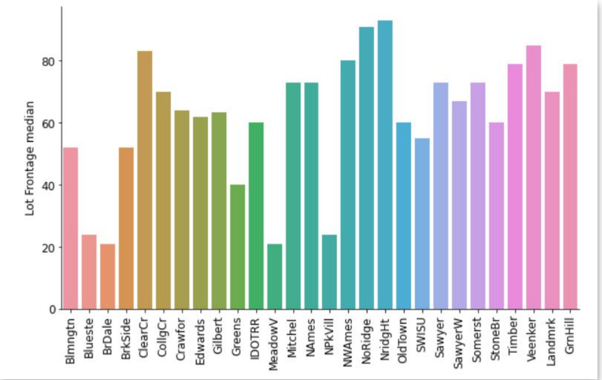
Dropping columns

- some columns had too many majority values



Imputing null values

- Using the median/mode to fill in null values

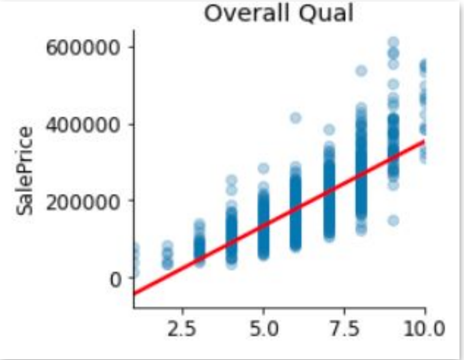


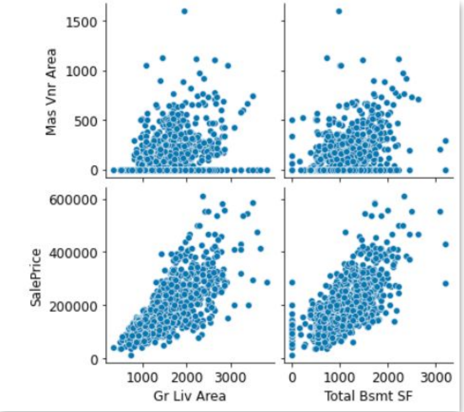
Exploratory Data Analysis

Scatterplots to identify correlations between variables and Sale Price.

- Variables related to quality/condition, and square footage show strong correlation to sale price.

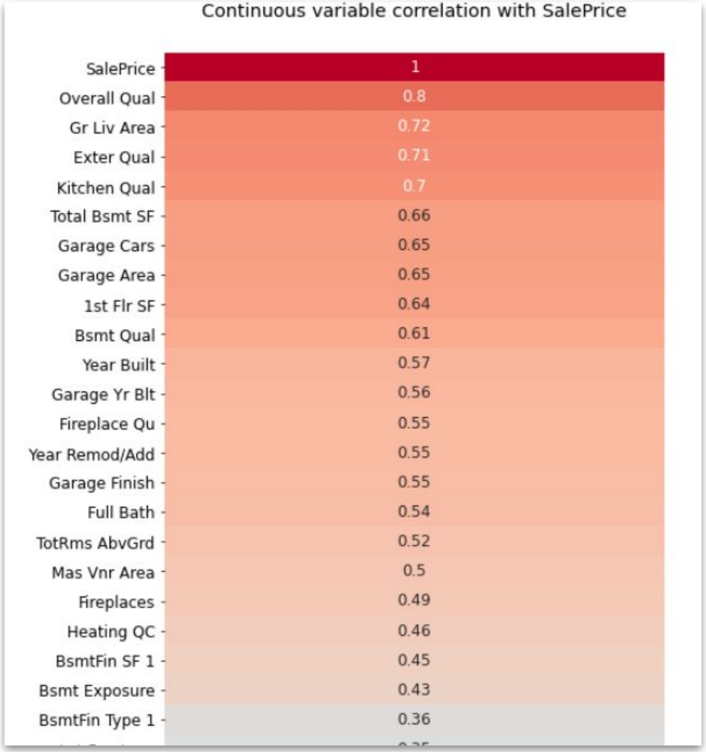
Pairplots to identify correlations between pairs of variables and Sale price.





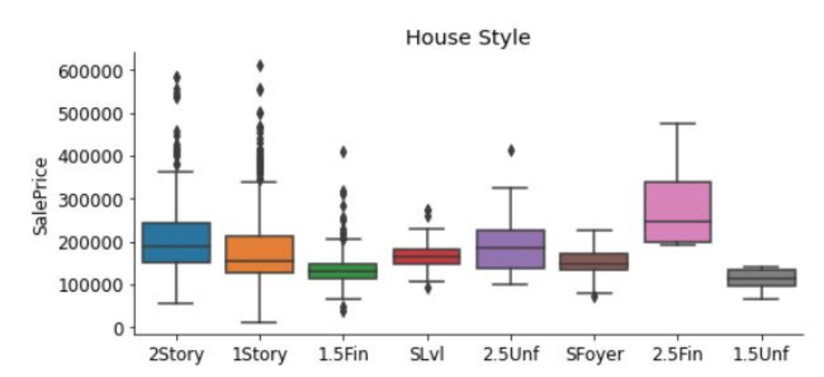
Exploratory Data Analysis

Heatmaps to rank variables by strength of correlation to Sale Price.



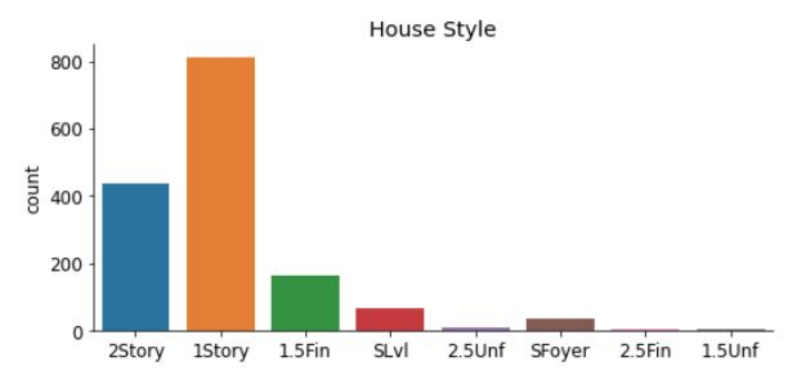
Exploratory Data Analysis

Boxplots to visualize distribution of Sale Prices across categorical data.



Count plots to visualize the distribution of values within a given feature.

- Features such as Heating were further dropped here due to low variation between categories.



Feature Engineering

Explored some interaction effects:

Year Built * Gr Liv Area

- Interaction term produces stronger correlation to sale price.

	Year Built	Gr Liv Area	Year Built * Gr Liv Area	SalePrice
Year Built	1.000000	0.237710	0.284415	0.570446
Gr Liv Area	0.237710	1.000000	0.998698	0.717142
Year Built * Gr Liv Area	0.284415	0.998698	1.000000	$\boxed{0.738471}$
SalePrice	0.570446	0.717142	0.738471	1.000000

Fireplaces * Fireplace Qu

- Interaction term did not appear to improve correlation to sale price.

	Fireplace * Qu	Fireplaces	Fireplace Qu	SalePrice
Fireplace * Qu	1.000000	0.960668	0.913630	0.532866
Fireplaces	0.960668	1.000000	0.864605	0.486506
Fireplace Qu	0.913630	0.864605	1.000000	0.550830
SalePrice	0.532866	0.486506	0.550830	1.000000

Pre-processing

Ordinal variables were encoded on an integer scale.

Excellent	5
Good	4
Average	3
Fair	2
Poor	1
None	0

Nominal variables were dummy encoded.

All data was also scaled appropriately.

	Land Contour
0	Bnk
1	Lvl
2	Lvl
3	Lvl
4	Lvl
...	...
1531	Bnk
1532	Lvl
1533	Lvl
1534	Lvl
1535	Lvl

	Land Contour_Bnk	Land Contour_HLS	Land Contour_Low	Land Contour_Lvl
0	1	0	0	0
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	0	0	0	1
...	...	...	...	...
1531	1	0	0	0
1532	0	0	0	1
1533	0	0	0	1
1534	0	0	0	1
1535	0	0	0	1

Model Evaluation

Round 1

As a first run, we used all available variables as predictors.

Model	Alpha	Train RMSE	Validation RMSE
Lasso with n_alphas=100	63.42	22069.57	33338.92
Ridge	12.75	21974.65	34909.88
MLR	na	22544.14	2.28e^17
Baseline(mean)	na	na	79511.73

This acts as our new baseline to benchmark the effects of our final selected predictors.

Overfitting is very severe here.

- Is evidenced by the difference between the train and validation RMSE.

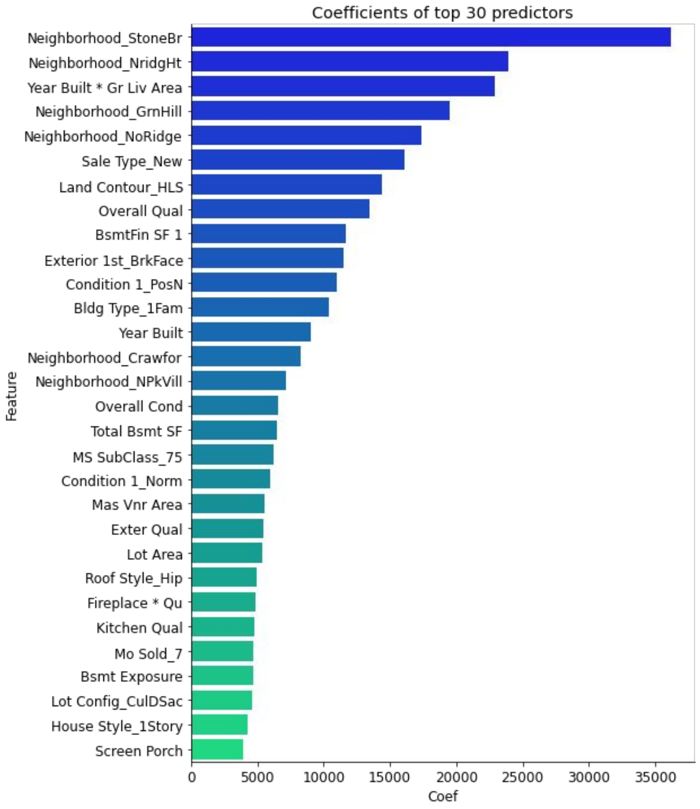
Model Evaluation

First run of Lasso model produces **115 non-zero coefficients.**

We select a few sets of predictors to test.

The best combination after testing:

- **Top 30 most positive coefficients**



Model Evaluation

Round 2

Testing all models using Top 30 variables with positive coefficients.

Model	Alpha	Train RMSE	Validation RMSE	Kaggle RMSE
Lasso	21.55	23735.29	24499.2	29382.85
Ridge	2.12	23819.92	24565.27	29357.91
MLR	na	23712.35	24625.63	29457.31
Baseline(mean)	na	na	79511.73	83689.59

Model Evaluation

Round 2

Observations:

Less overfitting

- Train RMSE ↑, validation RMSE ↓.
    - i.e. Increase Bias, reduce Variance.
- Difference between train and validation RMSE <1000.

Not perfectly generalizable yet

- Kaggle RMSE is still higher than validation RMSE.

Model	Alpha	Train RMSE	Validation RMSE	Kaggle RMSE
Lasso	21.55	23735.29	24499.2	29382.85
Ridge	2.12	23819.92	24565.27	29357.91
MLR	na	23712.35	24625.63	29457.31
Baseline(mean)	na	na	79511.73	83689.59

Discussion

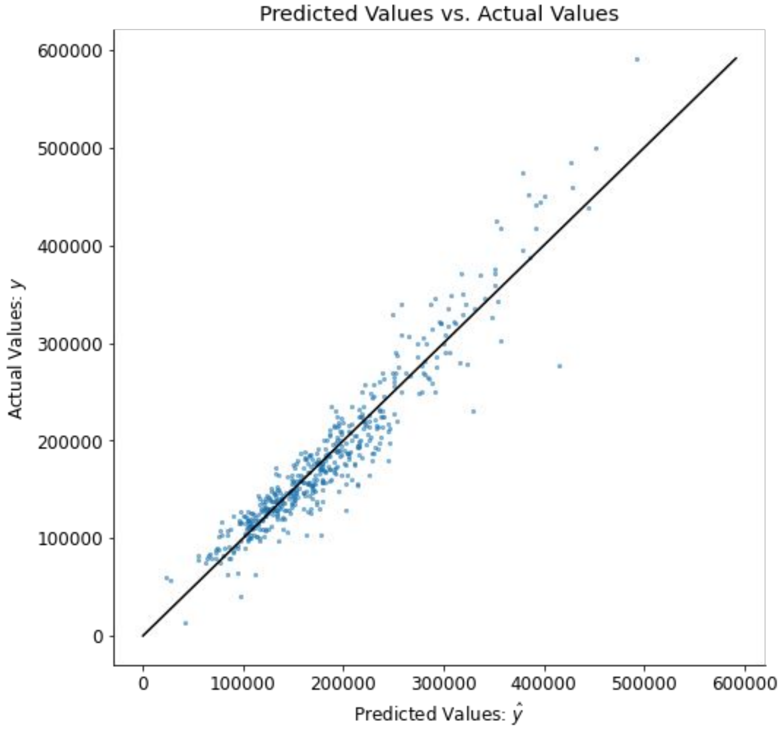
Final model: **Lasso** with **30** predictors, alpha=21.55.

- Best predictive power out of all models tested.
- **Unable to perform inference** on it, as Lasso is a biased estimator
    - Biases coefficients towards 0.

Discussion

Predictions vs Actual

- Decent performance for prices between 0 and $300,000
- Higher variance of residuals for expensive houses as there is less data to learn from.



Discussion

Neighborhood, **quality ratings**, and **square footage** related features are good predictors of price.

Expected price increase by neighborhood:

Stone Brook: **$36812**

Northridge Heights: **$23968**

Greenhill: **$19473**

Every unit increase in Overall quality increases house value by **$13471.**

Every unit increase in (Year Built * Gr Liv Area) increases house value by **$22,869.**

N.B. These refer to unit increases of the scaled variable.

Recommendations

For buyers:

- If location is not an issue, houses outside expensive neighborhoods will get you better quality and more space per dollar.

For sellers:

- Invest in upgrading the finish or workmanship of your house.
- Remodeling/renovating to add rooms can increase the value of your house.

For developers:

- Focus on developing new projects in high value neighborhoods like Stone Brook.

Recommendations

Improvements to our model:

Explore interaction effects:

- Year Built * Gr Liv Area was a significant interaction term.

Create a unique model for each neighborhood:

- Different features may have different significance depending on neighborhood.

Experiment with model complexity:

- Our model uses 30 predictors. Increasing the complexity by adding predictors may result in better performance. See Appendix.

Appendix

Effect of number of predictors on model performance:

30 predictors

```
Optimal alpha: 21.5540000000019
Lasso CV mean: -24456.775019057943
Lasso RMSE on train set: 23735.287596709277
Lasso RMSE on validation set: 24499.186358450745
```

45 predictors

```
Optimal alpha: 10.551817265617263
Lasso CV mean: -24118.39057722564
Lasso RMSE on train set: 23092.067818102507
Lasso RMSE on validation set: 23575.942824855232
```

60 predictors

```
Optimal alpha: 12.697823077378194
Lasso CV mean: -23831.76949917916
Lasso RMSE on train set: 22482.287246065916
Lasso RMSE on validation set: 22632.7231172807
```