Prediction of Housing Sale Price using the Ames Housing Data Set

DSI-23 Project 2
Ray Tan, Joey Kang, Samuel Ng, Lee Shi Min

Mispricing Of Housing Units: A $39B Problem

An information asymmetry & moral hazard problem: Real estate agents exploit info advantage and convince clients to sell their houses too cheaply

- Levitt & Syverson (2008) found that homes owned by real estate agents sold for **3.7% more than other houses**!

  The U.S. housing market in 2010 at a glance...

$278,000	**$10,000**	**4.18 mil**	**$39 bil**
Average Sale Price of Houses Sold	Est. Cost of Imperfect Info	Houses Sold in the U.S. in 2010	Loss in Housing Value assuming 95% of Homes Sold were not Property Agent Owned

Sources:

1. https://ideas.repec.org/a/tpr/restat/v90y2008i4p599-611.html
2. https://fred.stlouisfed.org/series/ASPUS
3. https://www.statista.com/statistics/226144/us-existing-home-sales/

Problem Statement

We want to help **uninformed home sellers** understand what constitute as fair housing prices by developing a regression model to **predict the sale prices of houses**. Specifically, we use linear models, i.e. ordinary least squares (OLS), Ridge and Lasso regressions.

A successful housing price prediction model should be able to predict housing prices with error term or **root mean squared error that is ideally lower than $10,000** (i.e. the cost of imperfection information in the housing market).

Data Sets Obtained From Kaggle

Housing sales data in Ames, Iowa from 2006 to 2010

- Contains a range of categorical, ordinal and continuous variables to capture unit-specific housing features (e.g. lot area, overall quality, neighbourhood)
- Data is randomly split into the train and the test sets:

Train Set	Test Set
● 2051 observations ● 80 features ● 1 target variable: sale price	● 878 observations ● 80 features (same as train)

Methodology

01	**Features Selection**	• Avoid overfitting • Easily understandable model • Use pair plots to identify variables with linear r/s to sale price
02	**Data Cleaning**	• Impute missing values • Remove outliers
03	**Features Engineering**	• Engineer custom features • Add interaction features • Assign numerical rank values for ordinal variables • Dummify categorical variables
04	**Model Preparation**	• Train-test split • Scaling
05	**Model Selection & Deployment**	• OLS, Ridge and Lasso Regressions • Cross Validation • Model Evaluation using R2 scores

Selection Of Features For Model Using Pair Plots

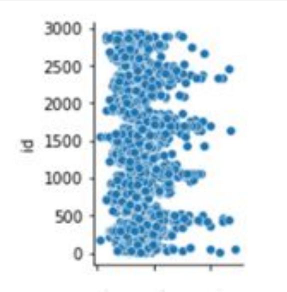
Pair plots are useful to provide visualisations to help us assess:

1. Whether variable has a **linear relationship** with target sale prices

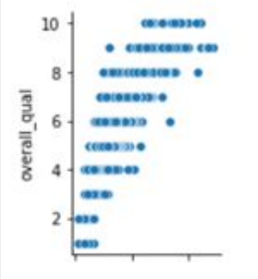
2. The **amount of variations** within each variable

3. Possible **collinearity and relationship** between similar variables

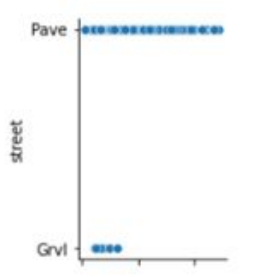
Sample Pair Plots of Independent Variables with Target (i.e. Housing Sale Prices)



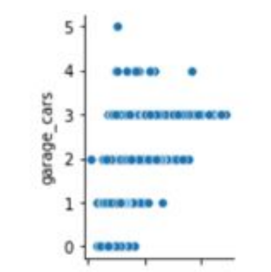
Non-linear

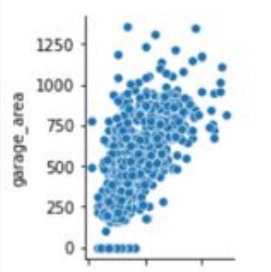


Linear



Feature with low variation





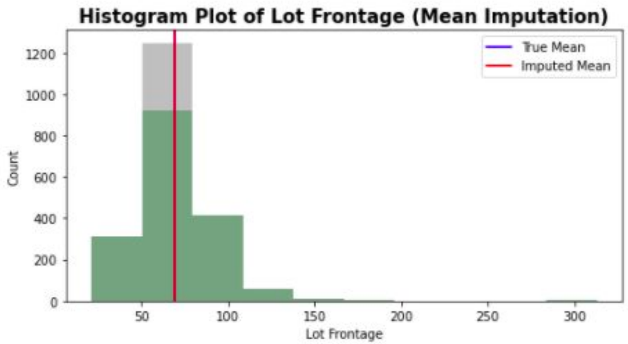
Related variables

Data Cleaning - Missing Values

	Missing Values
LotFrontage	330
BsmtQual	55
BsmtCond	55
TotalBsmtSF	1
BsmtFullBath	2
BsmtHalfBath	2
FireplaceQu	1000
GarageArea	1
GarageQual	114
GarageCond	114

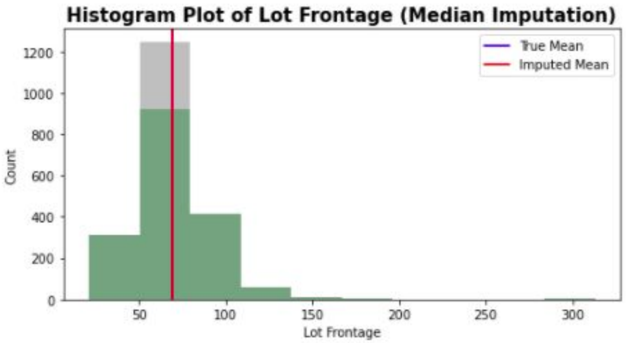
Method 1:
Imputation with **Mean**

```
Mean Lot Frontage Before Mean Imputation: 69.05520046484602
Mean Lot Frontage After Mean Imputation: 69.05520046484634
```



Method 2:
Imputation with **Median**

```
Mean Lot Frontage Before Median Imputation: 69.05520046484602
Mean Lot Frontage After Median Imputation: 68.88542174549
```

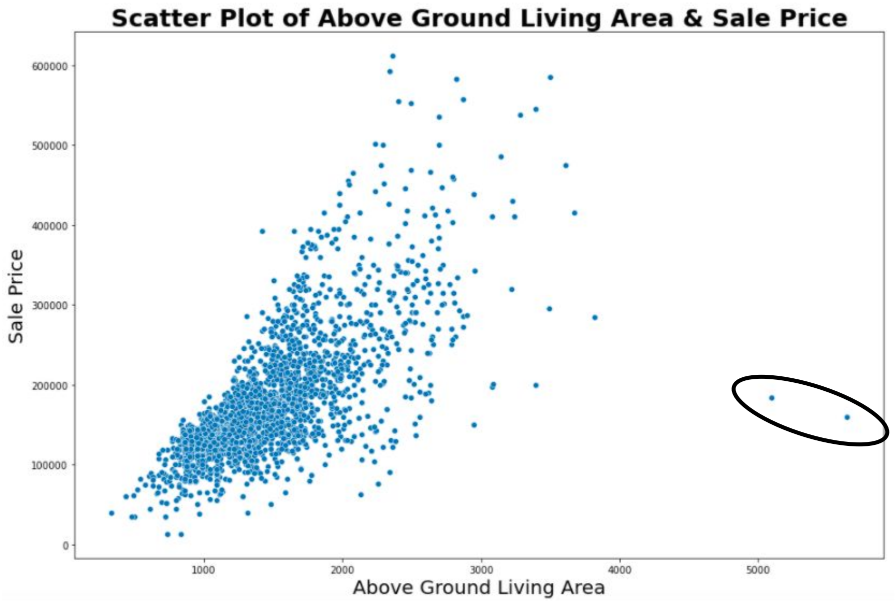


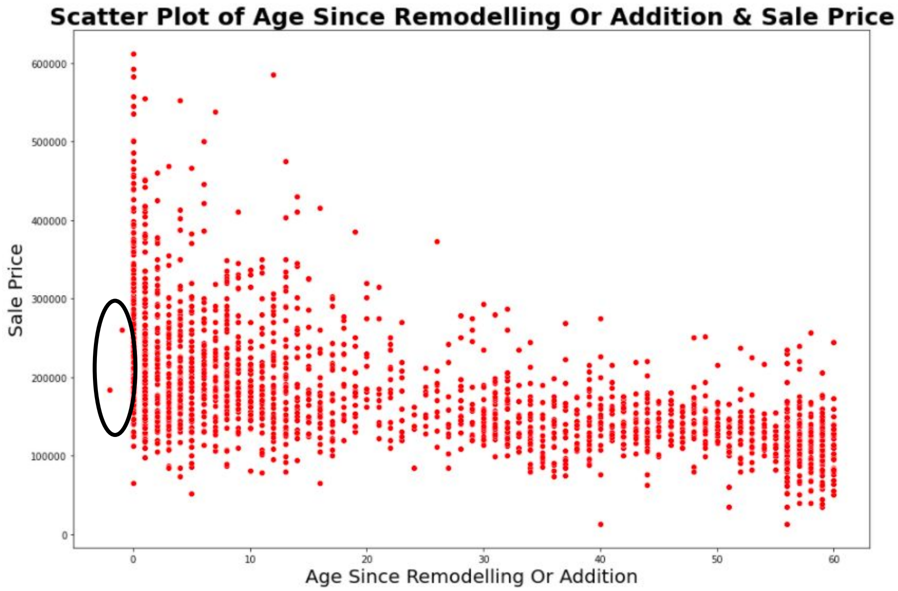
Data Cleaning - Missing Values

	Missing Values
LotFrontage	330
BsmtQual	55
BsmtCond	55
TotalBsmtSF	1
BsmtFullBath	2
BsmtHalfBath	2
FireplaceQu	1000
GarageArea	1
GarageQual	114
GarageCond	114

- Missing values were actually ‘NA’ values
- ‘NA’ values arised due to the absence of that feature
- They were replaced with 0 since these were numerical features

Data Cleaning - Outliers





Feature Engineering

Custom Features	
Age Since Remodelling or Addition	**= Year Sold - Year Remodelled or Added**
Total Rooms	**= Bathrooms + Bedrooms + Kitchen**
Porch Area	**= Open Porch + Enclosed Porch + 3 Season Porch + Screen Porch**
Neighbourhood Score	**= (to be explained later)**

Interaction Features
Overall Quality Condition
Exterior Quality Condition
Basement Quality Condition
Garage Quality Condition

= Quality x Condition

Feature Engineering

Ordinal Features
Exterior Quality Exterior Condition Basement Quality Basement Condition Garage Quality Garage Condition

Mapped strings with numerals

Ex	5
Gd	4
TA	3
Fa	2
Po	1

Nominal Features
Lot Shape Lot Configuration Land Contour Land Slope Year Sold

One-hot encoded (Dropped first)

Feature Engineering - Neighbourhood Score

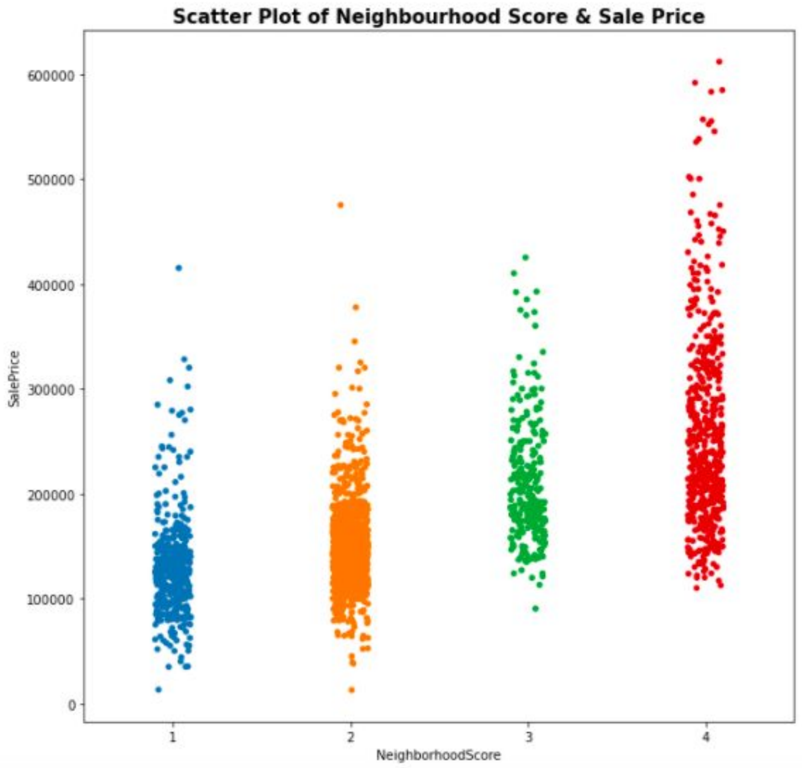
<u>Context</u>: Price of a house is determined not only by property-specific traits but also by **neighbourhood-specific traits** (i.e. the environment within which the house resides in)

<u>Idea</u>: The more favourable the neighbourhood, the higher the price with which the house can be sold for

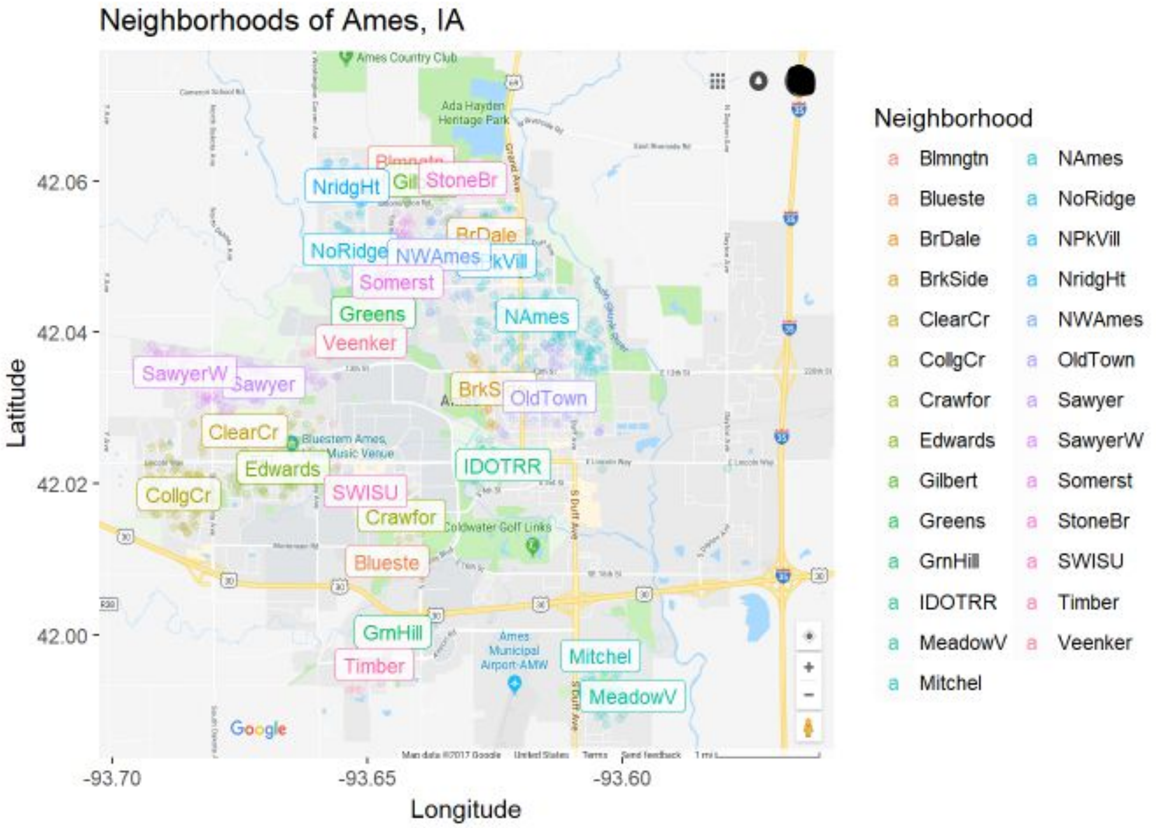
<u>Data</u>: Lack of features that could be considered neighbourhood-specific traits → **Had to improvise**

<u>Theory</u>: A more desirable neighbourhood would have:

1. Positive off-site features (e.g. park, greenbelt)
2. Typical, non-damaged, non-deducted houses
3. Higher overall quality and condition houses
4. Higher exterior quality and condition houses



Neighbourhoods In Ames, Iowa



Checking For Multi-Collinearity

What is Multi-Collinearity?

- It is when a variable can be linearly predicted from the one or more variables with a substantial degree of accuracy

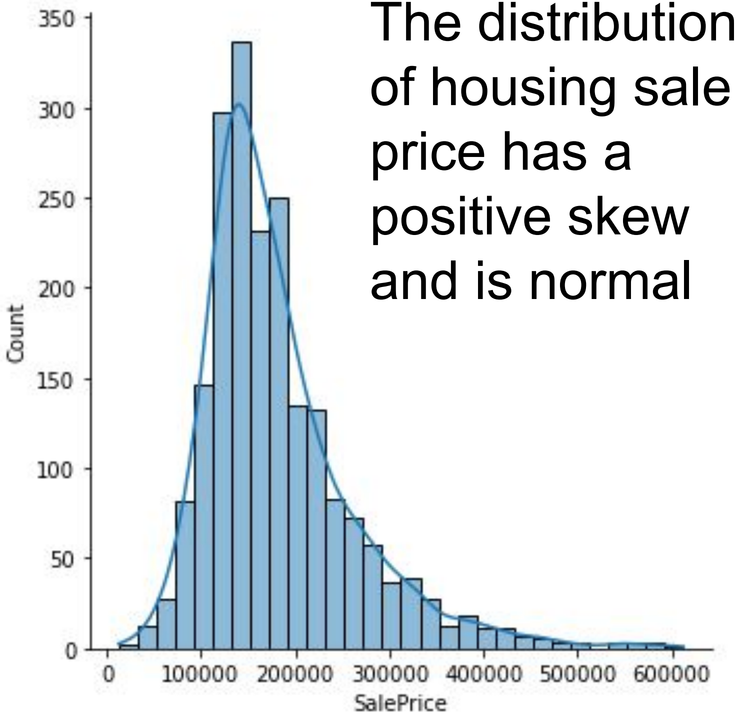
Issues of Multi-Collinearity

- Undermines the statistical significance of an independent variable.

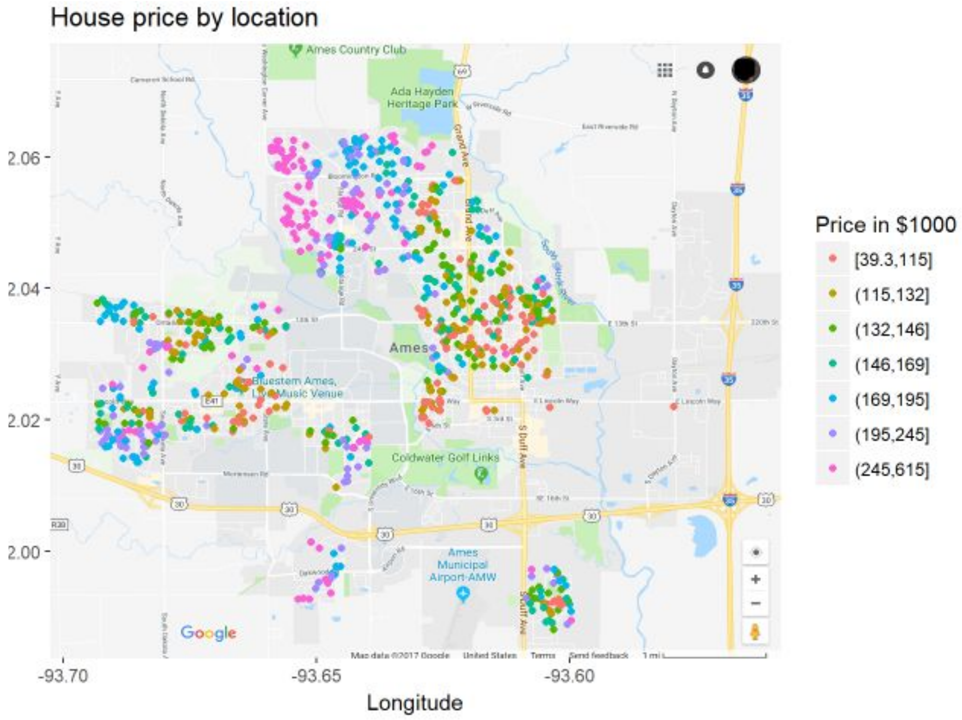
Threshold Selected = 0.8

Kept	**Dropped** (on the basis they are less correlated with sale price)
Age Since Remodeling or Addition	Year Remodelled or Added
Exterior Quality	Exterior Quality Condition
Fireplace Quality	Fireplaces
Garage Quality	Garage Condition
Pool Quality	Pool Area

Distribution Of Sale Price



The distribution of housing sale price has a positive skew and is normal

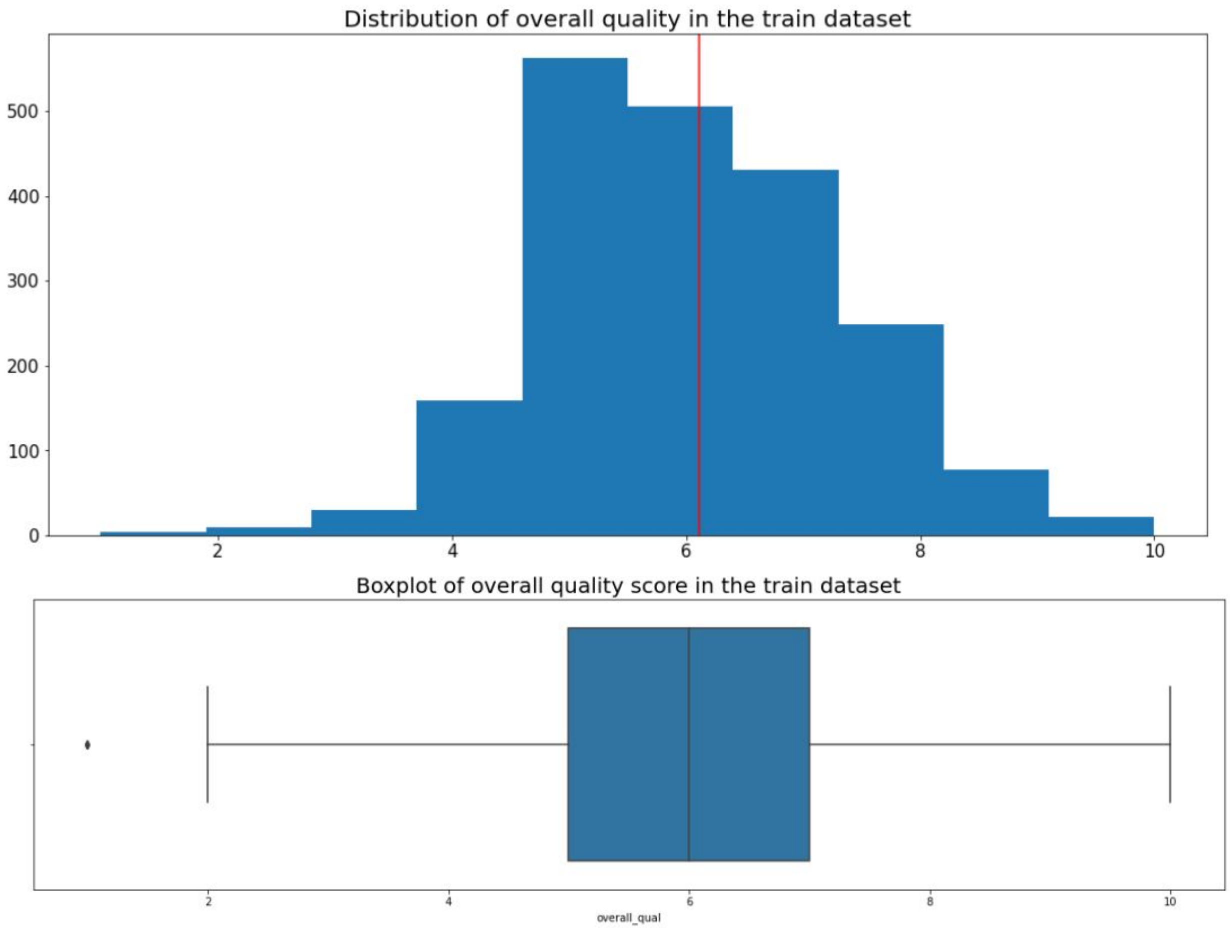


Distribution Of Overall Quality

(variable most correlated with Sale Price)

Similarly, normally distributed with a positive skew (but less so)

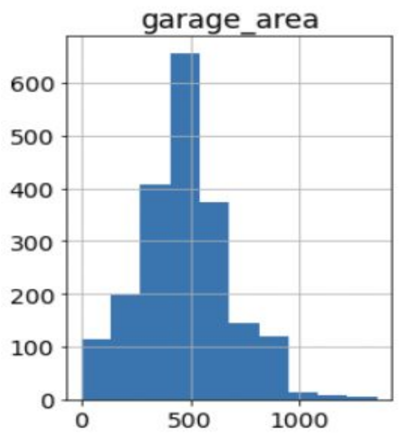
The mean overall quality is 6.11 out of 10 (see red line below).

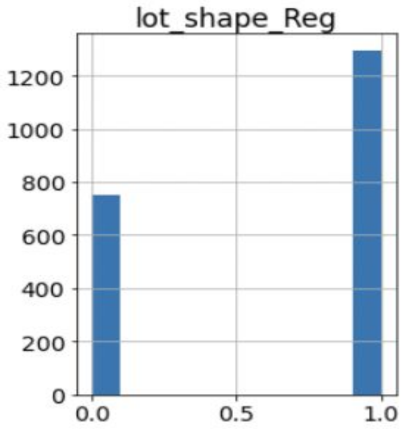


Model Preparation

- Dependent variable: **Sale Price**
- Independent variables: **49 variables** such as garage_area, overall_qual, lot_shape etc

- Train-test-split of **70-30**
  - Aligned with industry norms
- **Scaling** so that model is not impacted because of variables with large magnitude





Model Evaluation

	Linear Regression	LassoCV	RidgeCV
Cross Validation	0.8794	0.8798	0.8795
Train R^2	0.8886	0.8878	0.8884
Test R^2	0.8612	0.8520	0.8618
Difference in R^2	0.0274	0.0258	0.0266

- Selected the LassoCV as it has the
  - highest cross validation score; and
  - smallest difference in R^2 between the train and test set

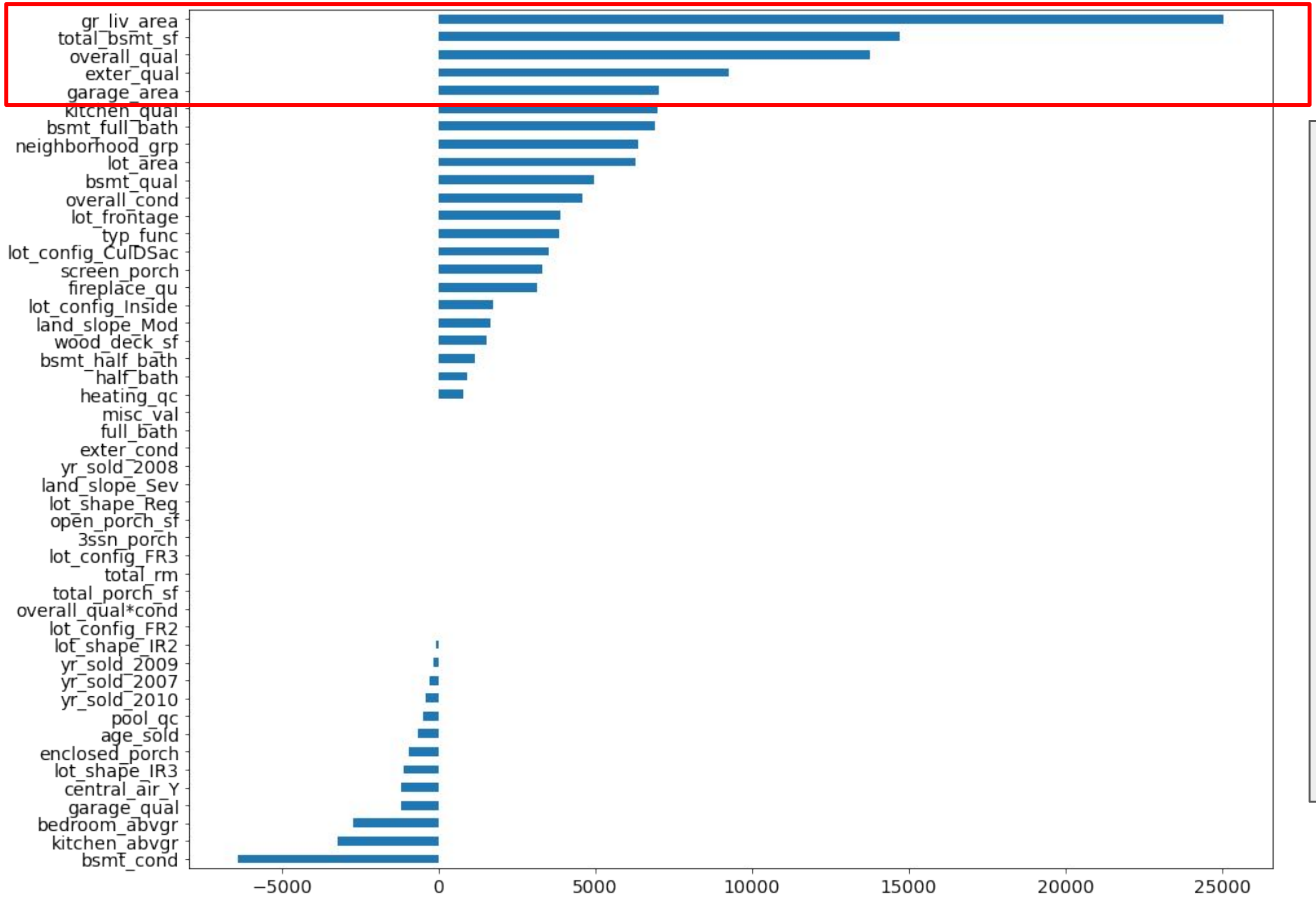
**The higher the R^2 the better*

Model Selected: Lasso Model

Lasso RMSE	<	Baseline RMSE
40,427		77,737

- Shrinks coefficient estimates towards zero (i.e. eliminating irrelevant variables) to make the model less complex
- Reduces variance so that model can generalise to new data better

Primary Findings



Size and quality of the house are important

- Above Ground Living Area
- Total Basement Area
- Overall Quality
- Exterior Quality
- Garage Area

Recommendations

- Sellers to pay attention to housing quality and size when setting sale price
- In addition, some of the key factors to note are as follows:
  - **Above ground living area** and **total basement area** will influence the sale price more so than the lot area and garage area
  - The **quality** of the overall, exterior, garage, etc has a larger effect on sale price than the condition
  - Houses situated in **neighborhoods** with higher scores can sell at a slight premium

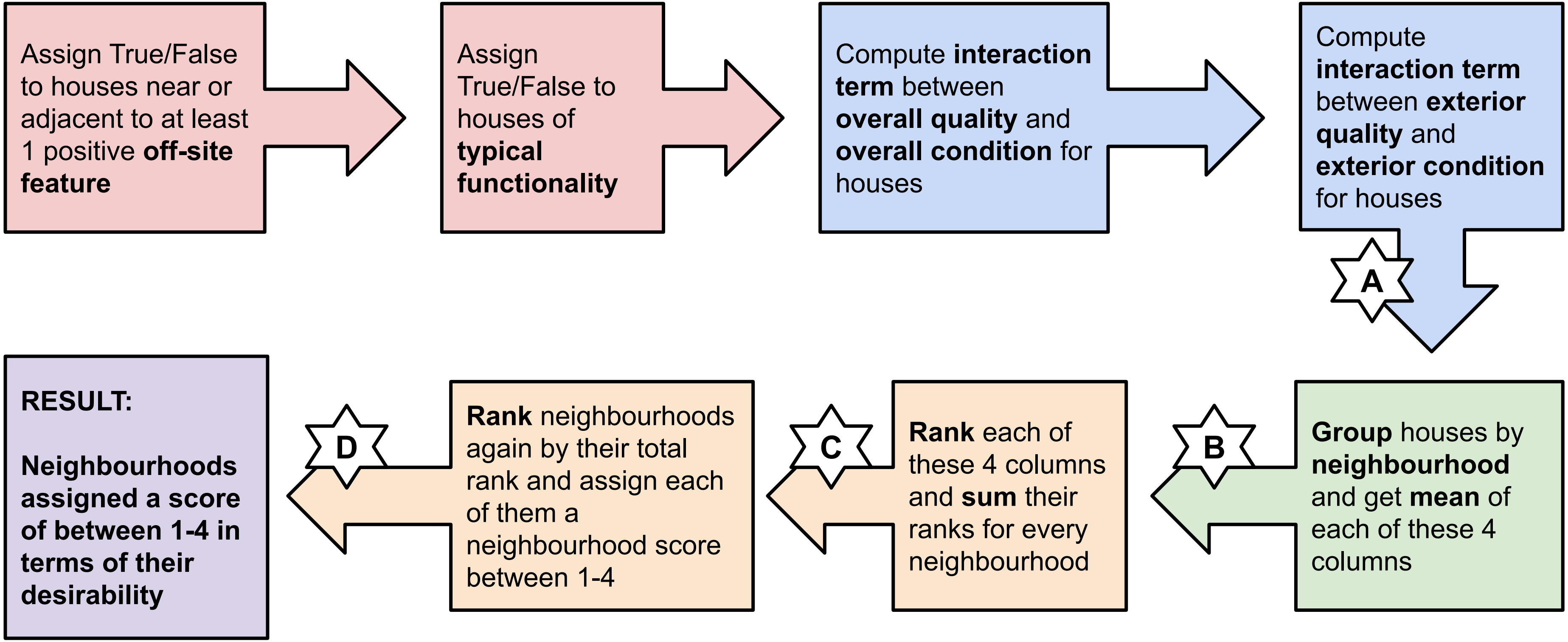
Conclusion

- Model RMSE is ~$40,000 > $10,000 (cost of imperfection information in the housing market) → **model can be improved**
- Consider including all variables and use lasso regularization to identify the variables

Next Steps: Consider gathering housing data of other states so that the model can be expanded to other parts of the US instead of limiting to Iowa

Annex

Feature Engineering - Neighbourhood Score



Feature Engineering - Neighbourhood Score

A

	Neighborhood	PosOffSiteFeature	TypFunctional	OverallQualCond	ExterQualCond
0	Sawyer	0	1	48	12
1	SawyerW	0	1	35	12
2	NAmes	0	1	35	12
3	Timber	0	1	25	9
4	SawyerW	0	1	48	9

B

	Neighborhood	PosOffSiteFeature	TypFunctional	OverallQualCond	ExterQualCond
0	Blmngtn	0.000000	1.000000	35.909091	12.000000
1	Blueste	0.000000	1.000000	38.666667	11.000000
2	BrDale	0.000000	0.947368	31.105263	9.000000
3	BrkSide	0.013158	0.894737	33.078947	9.407895
4	ClearCr	0.000000	0.740741	33.407407	10.259259

C

	Neighborhood	PosOffSiteFeatureRank	TypFunctionalRank	OverallQualCondRank	ExterQualCondRank	TotalRank	FinalRank
0	Blmngtn	9.0	24.5	18.0	22.0	73.5	21.5
1	Blueste	9.0	24.5	22.0	17.0	72.5	19.0
2	BrDale	9.0	13.0	7.0	3.5	32.5	5.0
3	BrkSide	19.0	7.0	11.0	7.0	44.0	10.0
4	ClearCr	9.0	1.0	13.0	14.0	37.0	6.0

D

	Neighborhood	FinalRank	DesirabilityScore
0	IDOTRR	1.0	1.0
1	Edwards	2.0	1.0
2	MeadowV	3.0	1.0
3	SWISU	4.0	1.0
4	BrDale	5.0	1.0