

The background image shows a suburban street scene. On the left, there are green trees and a utility pole. In the center, a row of houses is visible, including a prominent yellow house with a green roof. To the right, a two-story house with white siding and a brick base is partially visible. A gold SUV is parked on the street in the foreground on the right. The sky is blue with some light clouds.

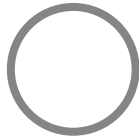
Project 2:

Building a Price Prediction Model for Houses in Ames, Iowa

Clara Gan
Gan Tze Ling
Tang Huimin
Chia Kang Yang
Ken Tan
Rebecca Liu



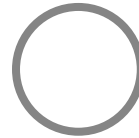
Problem Statement



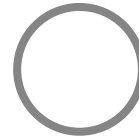
Methodologies

Forward Selection

Backward Selection



Model Evaluation



**Conclusions &
Recommendations**

Problem Statement

Home sellers often anchor their offer price to avoid underselling, while home buyers, due to information asymmetry, often pay different prices for houses with similar features.

As a property consultancy firm, we aim to build a model to identify the features that are most important to predict sales price of houses. This would allow us to provide our clients with a tool that gives estimates of their potential selling prices, and help them identify which aspects of their properties they can improve on to enhance their selling prices.



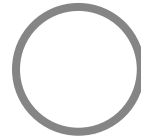
Problem Statement



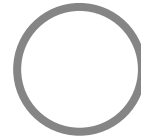
Methodologies

Forward Selection

Backward Selection



Model Evaluation



**Conclusions &
Recommendations**

Methodologies

Forward Selection

- Selecting variables before running the models
- Two approaches:
 - a. Based on judgement call
 - b. Based on systematic approach

Backward Selection

- Entire dataset was clean before running the model
- All variables were kept for modelling

Forward Selection: Judgment Call



Forward Selection: Judgment Call

Initial Exclusions

- Identified variables based on judgement call that did not affect sale price
 - E.g. ID, PID, roof material
- Simplifying variables by identifying overlapping variables
 - E.g. garage type vs garage cars

Garage Cars

Size of garage in
car capacity

VS

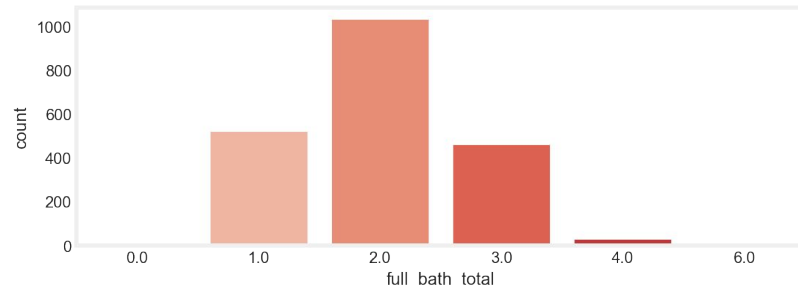
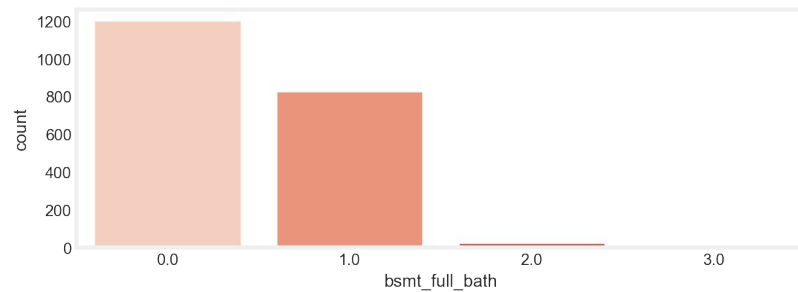
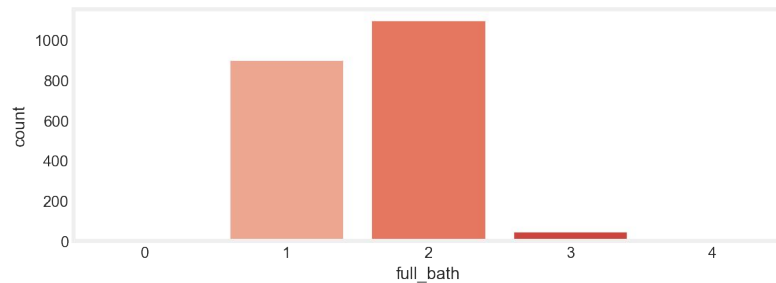
Garage Area

Size of garage in
square feet

Forward Selection: Judgment Call

Feature Engineering

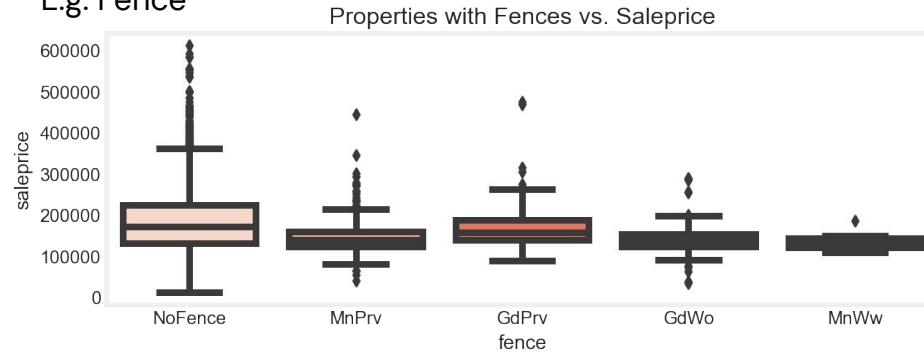
- Simplifying variables and combining variables
 - For example, summing total number of bathrooms.



Forward Selection: Judgment Call

Drop Variables of Little Significance

- If the quality of the variable does not significantly affect the average sale price of the property
- E.g. Fence



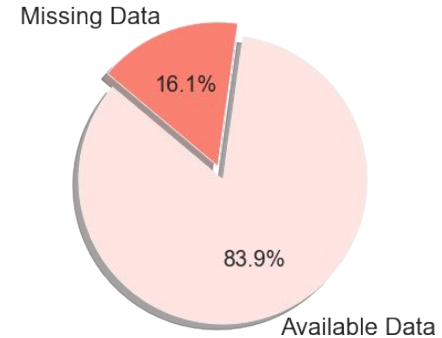
From the above boxplot, we can see that regardless of the quality of the fences, the mean prices for the properties appear to be relatively similar.

Forward Selection: Judgment Call

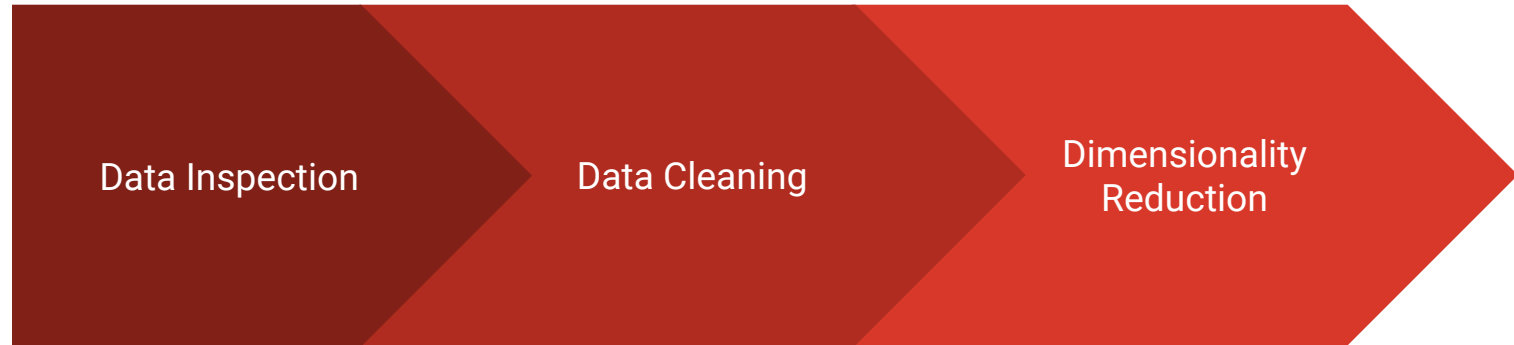
Drop Variables to Improve Quality of Data

- If the sum of null values in the variable was significant.
 - E.g. Lot Frontage; ~20% null values
- If there are categories with few data points
 - E.g. Agricultural zones or split foyer

Available Data for Lot Frontage



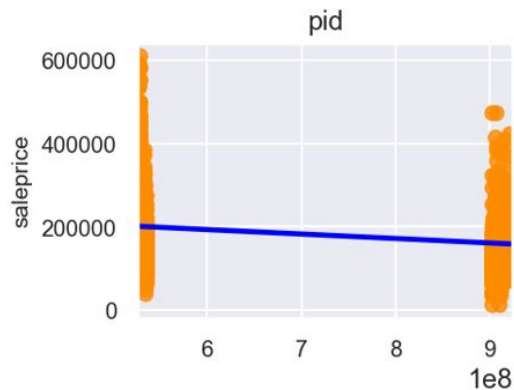
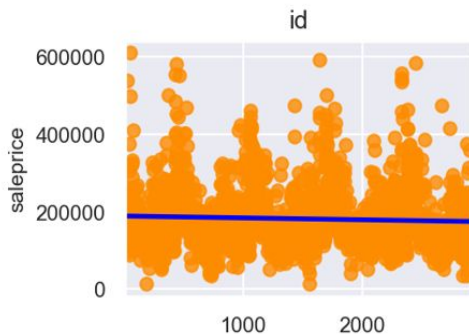
Forward Selection: Systematic Approach



Forward Selection: Systematic Approach

Data Inspection

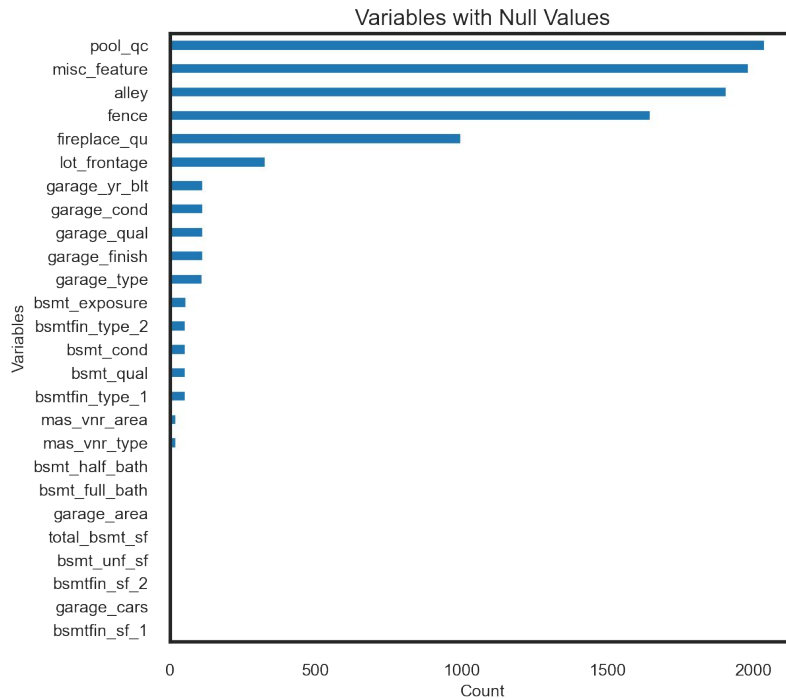
- Identify variables that are unlikely to affect sale price.
 - For example, variables such as 'ID' and 'PID'.



Forward Selection: Systematic Approach

Data Inspection

- Identify null values.
- For example, 'Pool Quality',
'Miscellaneous Features', etc.



Forward Selection: Systematic Approach

Data Cleaning

- Missing data fields are cleaned, imputed or dropped, according to each column's needs.
- Check for anomalous data and correct/ remove them, if necessary.
 - For example, 'future-dated' 'Year' values.

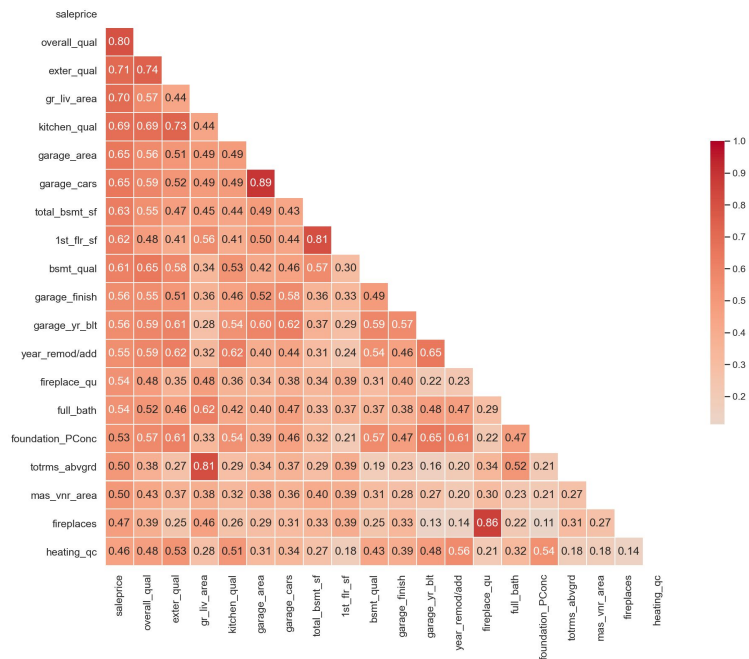
ID	Garage Year Built	Year Remodeled
2261	2207	2007

Forward Selection: Systematic Approach

Dimensionality Reduction

- Apply high correlation filter
 - Drop one variable of each pair of variables whose pairwise correlation exceeds a preset threshold.
 - Only retain one variable of each pair of variables that show a decent or high correlation with sale price.

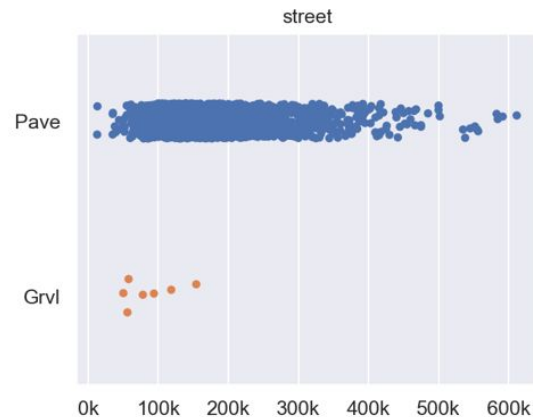
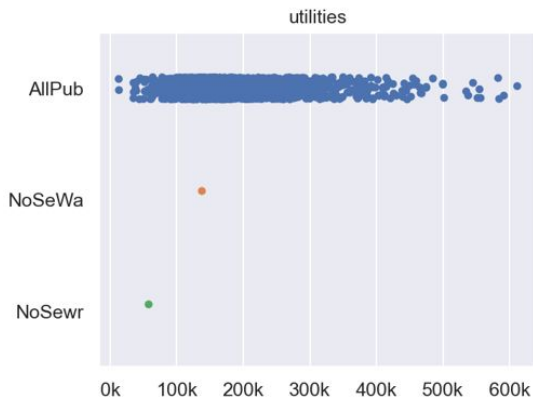
Heatmap of Top 20 Positively Correlated Numeric Features After Pre-Processing



Forward Selection: Systematic Approach

Dimensionality Reduction

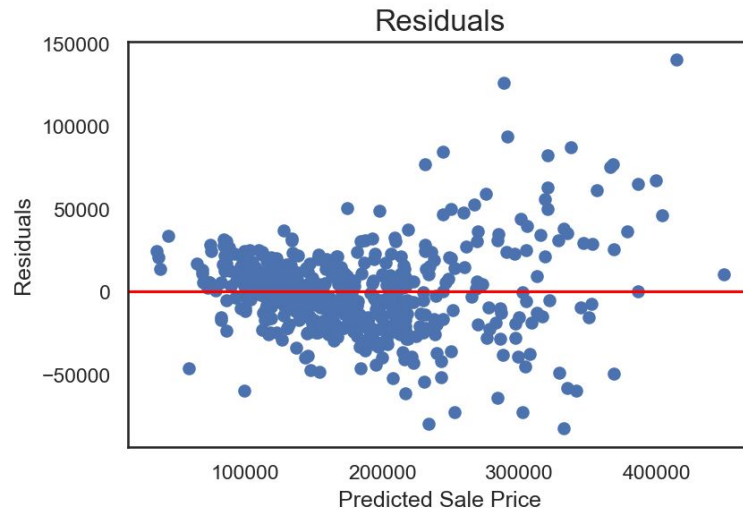
- Apply low variance filter
 - Variables with a variance lower than a preset threshold are removed.
 - This is because variables with a low variance will have little effect on the sale price.
 - For example, values for 'Utilities' column are all 'All Public Utilities' except for two values, and values for 'Street' column are all 'Paved' except for seven values.



Forward Selection: Regression Model

Best performing model → Lasso Regression

Training R²	88.3%
Test R²	89.2%



Backward Selection

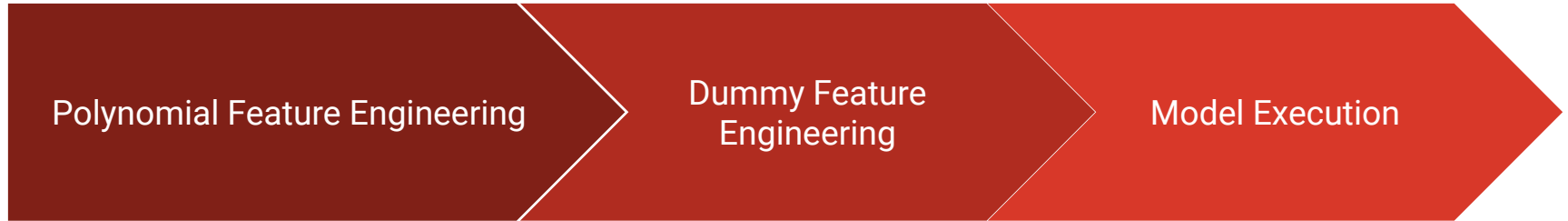
Underlying principle - More Data is Better. There are hidden interactions and nuances behind different data points that we want the machine and algo to pick up.

Approach - Focus on information gain. Employ feature engineering to amplify the information our models can extract from the base data-set. We then refine, adapt and iterate accordingly to achieve the best possible result.

Implementation - Clean up the data set to ensure data quality and logical consistency is present. Avoid throwing out data unless we are absolutely certain that they are redundant. Do feature engineering on a big scale.

Mild multicollinearity issues? Not a major issue - we rely on standardisation and regularisation to handle these for us. After all, polynomial feature engineering (a standard technique) creates multiple highly correlated features which are then used subsequently in regressions.

Backward Selection - Process Flow



Backward Selection - Polynomial Feature Engineering



Polynomial Feature Engineering

1. Split the training data into numerical and categorical dfs.
2. Create new numerical features as desired, i.e. `age_property_sold`.
3. Apply polynomial feature engineering with 2 degrees and interaction terms included.
4. Apply standardisation techniques to the output from 2.

Backward Selection - Dummy Feature Engineering



Polynomial Feature Engineering

Dummy Feature Engineering

1. Clean and ensure all categorical features have logical and well separate categories.
2. Create new categorical features as desired, i.e. seasons.
3. Apply dummy feature engineering to all categorical features.

Backward Selection - Model Execution



Polynomial Feature Engineering

Dummy Feature Engineering

Model Execution

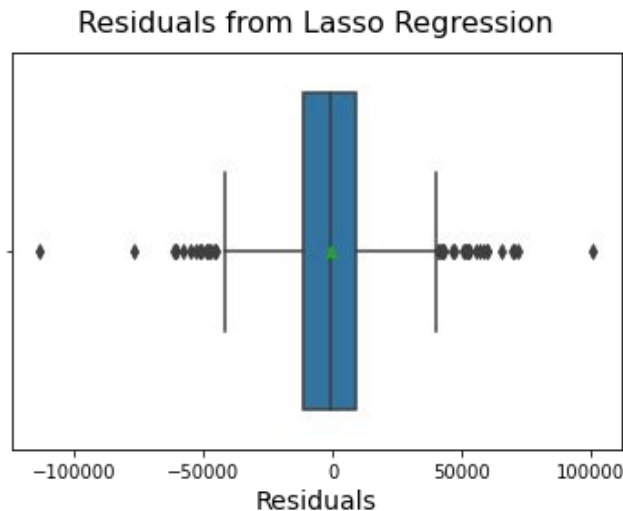
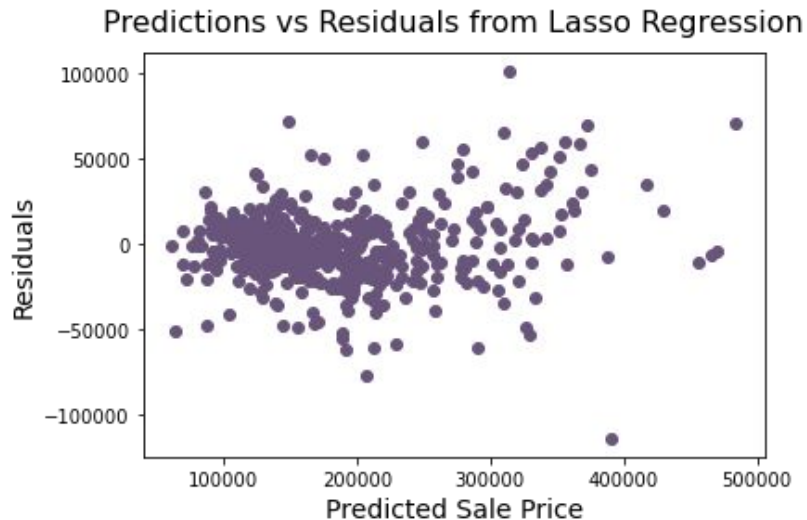
1. Reassemble the training data and experiment with different parameters across the major model classes.
2. Deploy optimisation techniques to identify the optimal parameters where possible.
3. Decide on the best performing and interpretable model.

Backward Selection

Best performing model → Lasso Regression

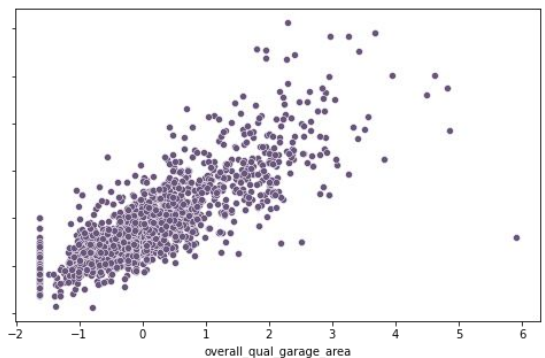
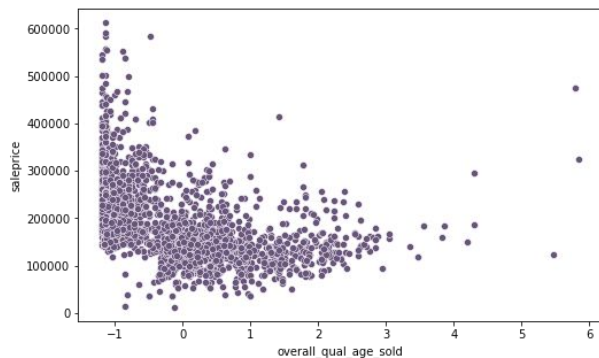
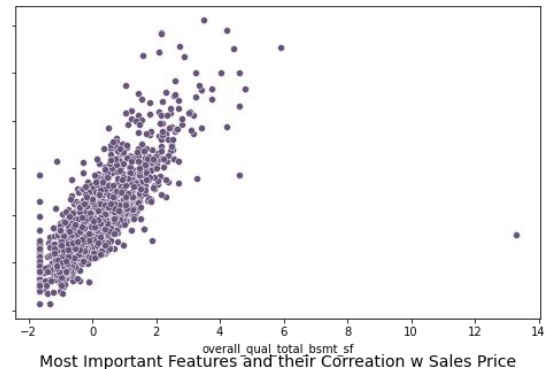
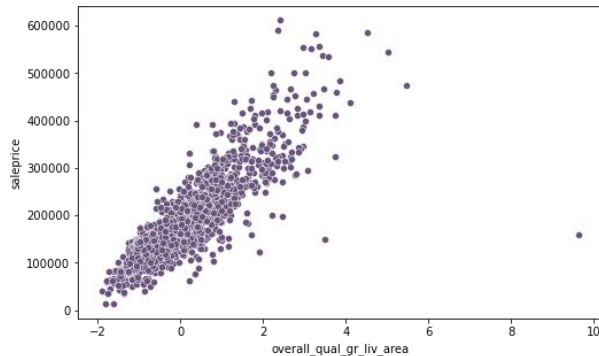
Why? Lasso works well with large data-sets with many features (like the training dataset used here) and performs feature selection automatically by shrinking the coeffs of less important features.

Training R2	92.0%
Test R2	92.8%



Backward Selection - Selected Features by Lasso

1. The majority of the features “selected” by Lasso as most important were interaction features.
2. Some of them exhibited a direct relationship with sales price, but others had less indirect relationships.





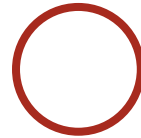
Problem Statement



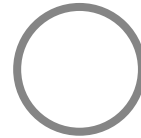
Methodologies

Forward Selection

Backward Selection



Model Evaluation



**Conclusions &
Recommendations**

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features

1029

38

76

19

35

43

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Backward selection gave the highest R2

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features 1029 38 76 19 35 43



Backward selection gave the highest R2

- Less time efficient
- More computationally intensive

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Backward selection gave the highest R2

- Less time efficient
- More computationally intensive
- Results were harder to interpret

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Forward selection gave lower R2 scores

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Forward selection gave lower R2 scores

- More time efficient
- Less computationally intensive

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Forward selection gave lower R2 scores

- More time efficient
- Less computationally intensive
- Results were easier to understand

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

In general, all models performed well

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Problem Statement

- **Predictive tool** to help clients estimate price of their properties
- **Gain insights** to inform recommendations on how clients can enhance selling price

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features 1029 38 76 19 35 43

Problem Statement



- Predictive tool** to help clients estimate price of their properties
- Gain insights** to inform recommendations on how clients can enhance selling price

	Kang Yang	Ken	Huimin	Rebecca	Clara	Tze Ling
1	Overall Quality + Ground Living Area	Ground Living Area	Ground Living Area	Overall Quality	Neighborhood	Neighborhood
2	Overall Quality + Total Basement Area	Overall Quality	Overall Quality	Age of Property	Ground Living Area	Land Contour
3	Masonry Veneer Area + Pool Area	Neighborhood	Neighbourhood	Overall Condition	Sale Type	Near positive feature
4	Overall Quality + Basemt Finished Area	Basement Exposure	External Quality	Size of Property	MS Zoning	Exterior Quality
5	Basemt Finished Area + Pool Area	Garage Area	Garage Cars	No. of Full Baths	Overall Quality	Exterior Covering
Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Test R2	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

Features	1029	38	76	19	35	43
----------	------	----	----	----	----	----

Problem Statement

- **Predictive tool** to help clients estimate price of their properties
- ➡ **Gain insights** to inform recommendations on how clients can enhance selling price



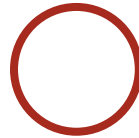
Problem Statement



Methodologies

Forward Selection

Backward Selection

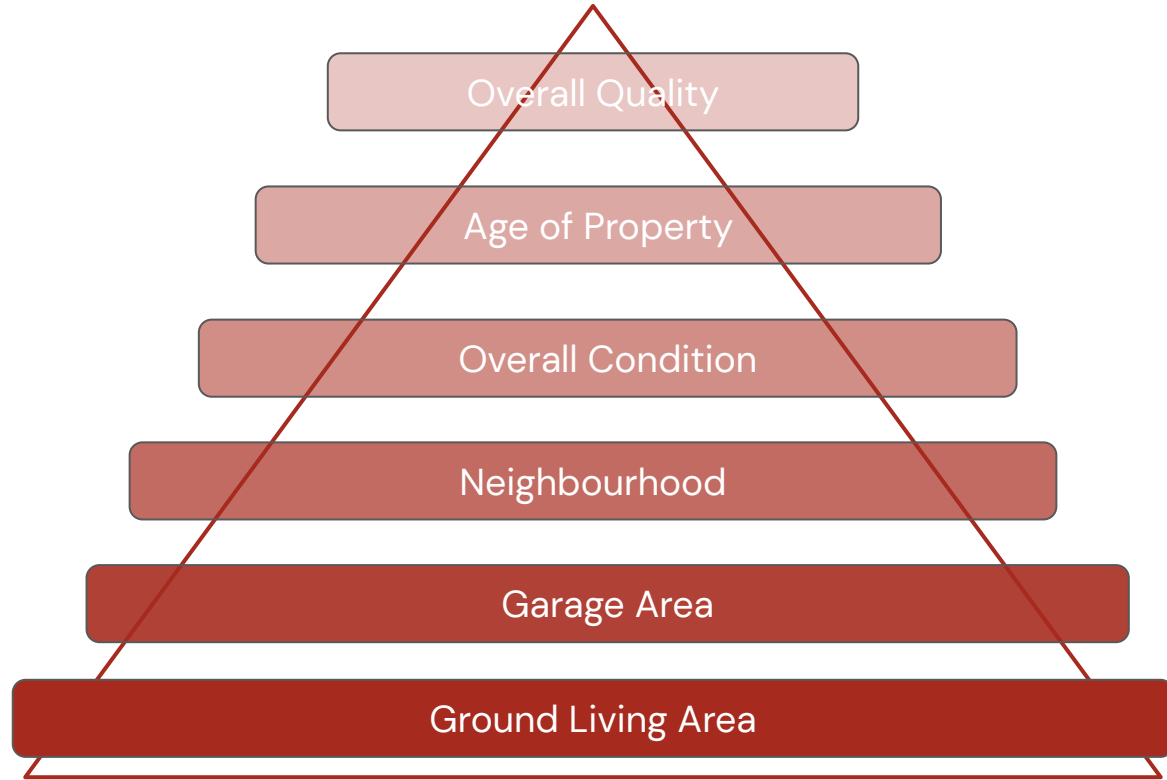


Model Evaluation



**Conclusions &
Recommendations**

Conclusion



Recommendations

Focus on Quality

Ensure materials and finish used for the house is of high quality

Being Young is Attractive

Older houses sell for much less than newer houses.

Location Matters

Use the neighborhood of the property as an anchoring point for the market sale price of the property.

- + Northridge Heights, Northridge, Stone Brook
- Edwards, Northwest Ames, Old Town

Recommendations

Garage Size

Build a new garage if not present

Expand the garage area

Indoor > Outdoor

Extend ground living area if there are large area of unused land, as it more valuable than lot area.

Less emphasis can be placed on urban landscape.

Put in Effort for Maintenance

Maintain the overall condition of the house to maximize resale value

Interesting Features



Fireplaces

- Build them if you haven't already



Bathroom Count & Type

- Total number of bathroom matters, especially full bathrooms; consider converting half bathrooms to full if possible



Lot Area

- Despite size being an important factor for several features, lot area does not seem to affect sale price much

Next Steps?

01

Proximity to Amenities

As location is a top feature, the data collection of the proximities of amenities (such as schools, supermarkets and restaurants) could be included to further narrow down the neighbourhood features that has a high correlation with the sale prices of properties.

02

Intangible Variables

Variables that increases the quality of one's life, which may rival the location of the property as the most important feature for the purchase of a property. For example, the safety index & crime rate of the neighborhood, transport accessibility, and noise & air quality.

Next Steps?

03

Access to continuous flow of transactional data

The sale prices are only accurate to a certain extent due to the lack of recent sales data. A real-time machine-learning model with access to a continuous flow of transactional data would be necessary to keep the model updated. Data collection could be improved to deal with the missing data fields upfront.

04

Change your Target Audience

For neighborhoods that appear to be less desirable to live in, such as the South & West of Iowa State University, perhaps changing the target audience of property buyers from new families to families with schooling children. Broaden target market by positioning it as an up and coming city with education district.



T₁

H₄

A₁

N₁

K₅

S₁