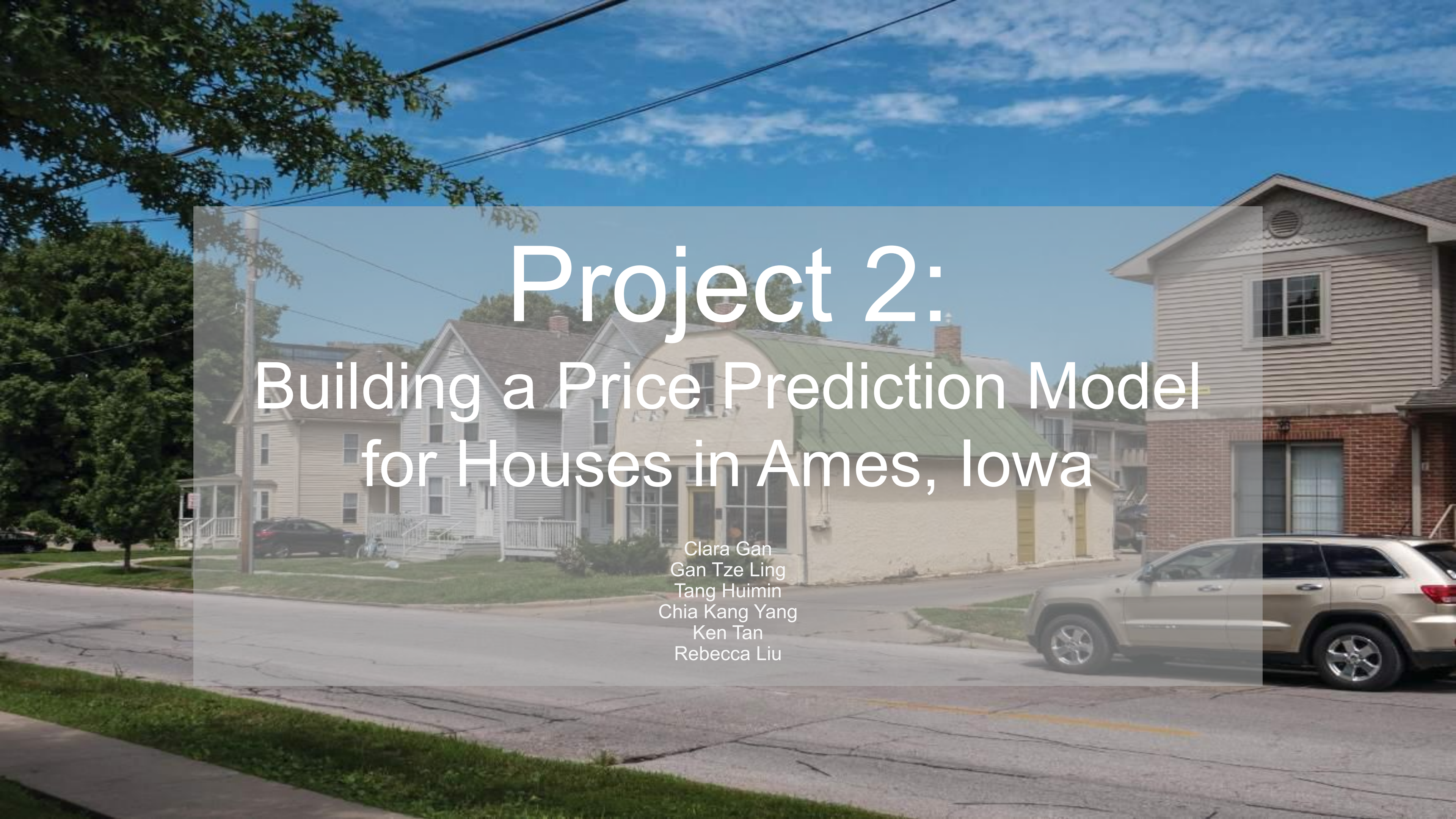
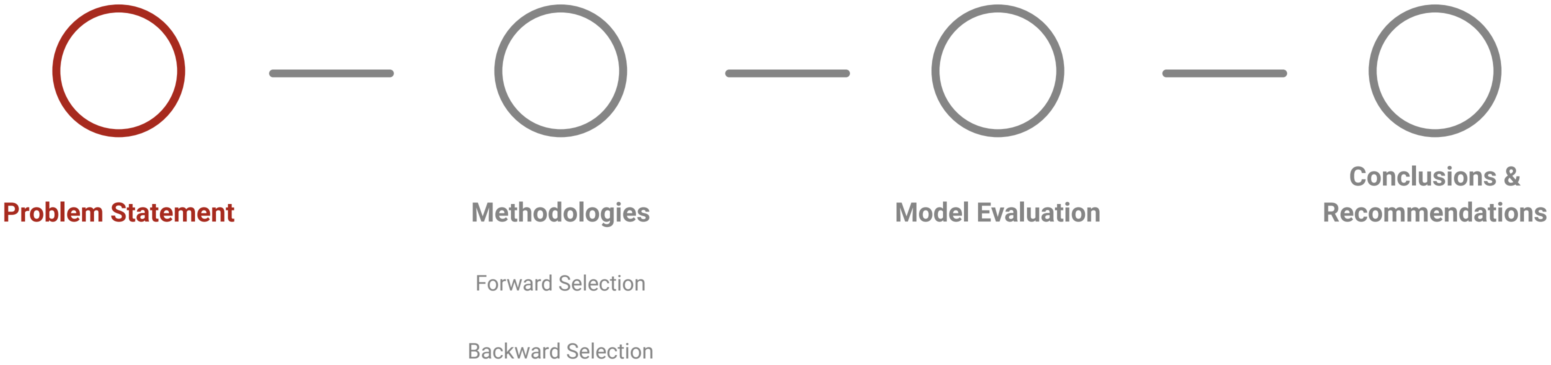


Project 2:

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

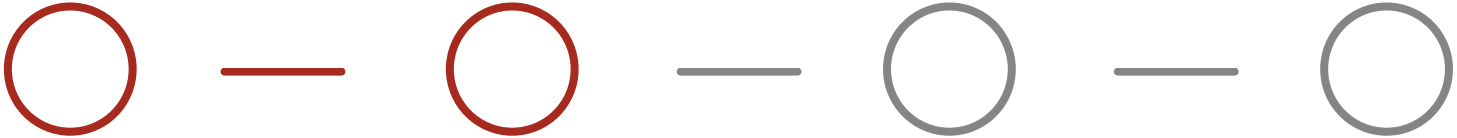
Model Evaluation

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

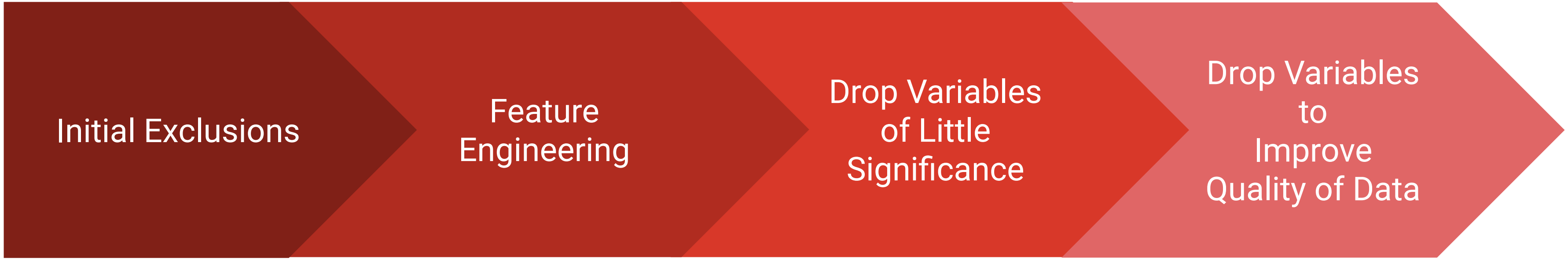
Forward Selection

- 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



Initial Exclusions

Feature Engineering

Drop Variables of Little Significance

Drop Variables to Improve Quality of Data

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

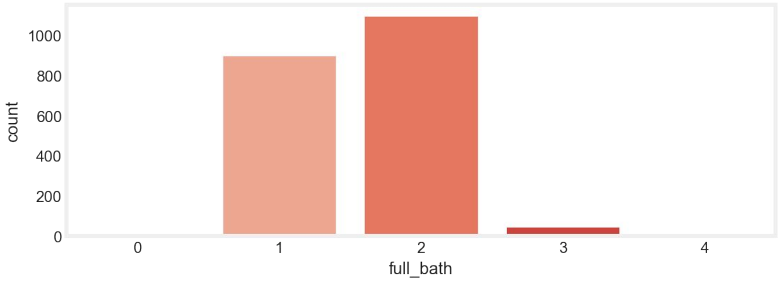
VS

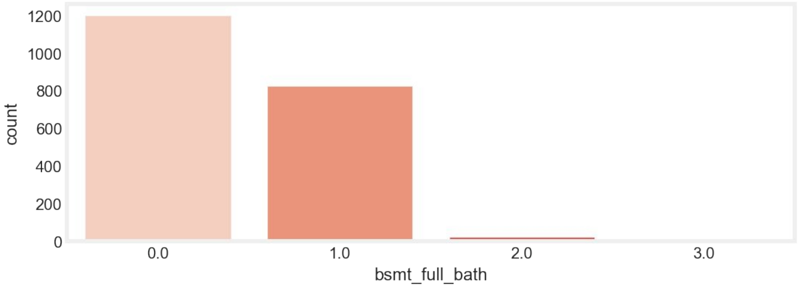
Garage Area
Size of garage in square feet

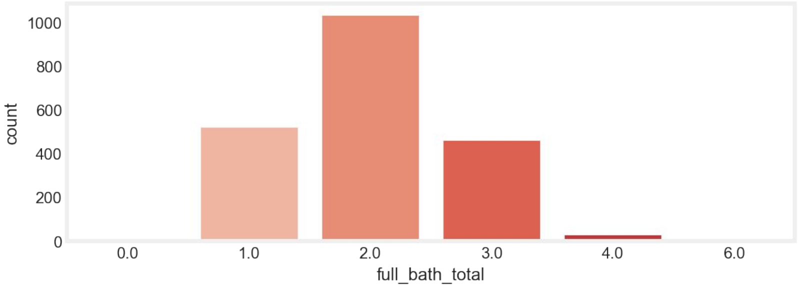
Forward Selection: Judgment Call

Feature Engineering

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





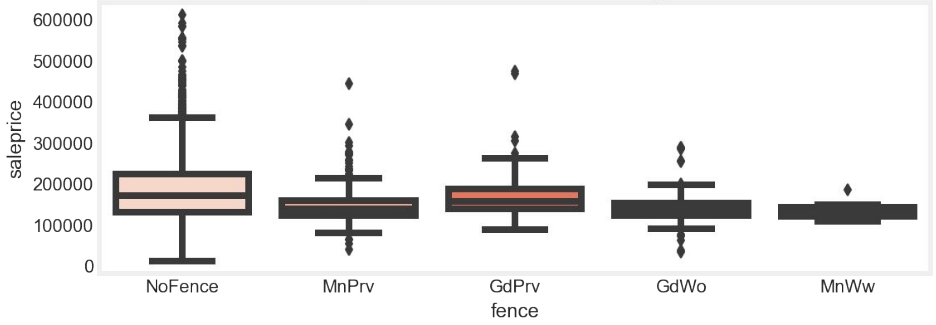


Forward Selection: Judgment Call

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

Properties with Fences vs. Saleprice



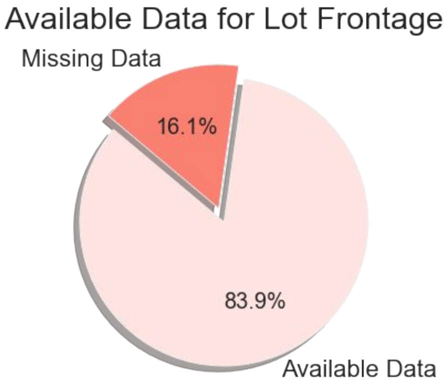
From the above boxplot, we can see that regardless of the quality of the fences, the mean prices for the properties appear to be relatively similar.

Forward Selection: Judgment Call

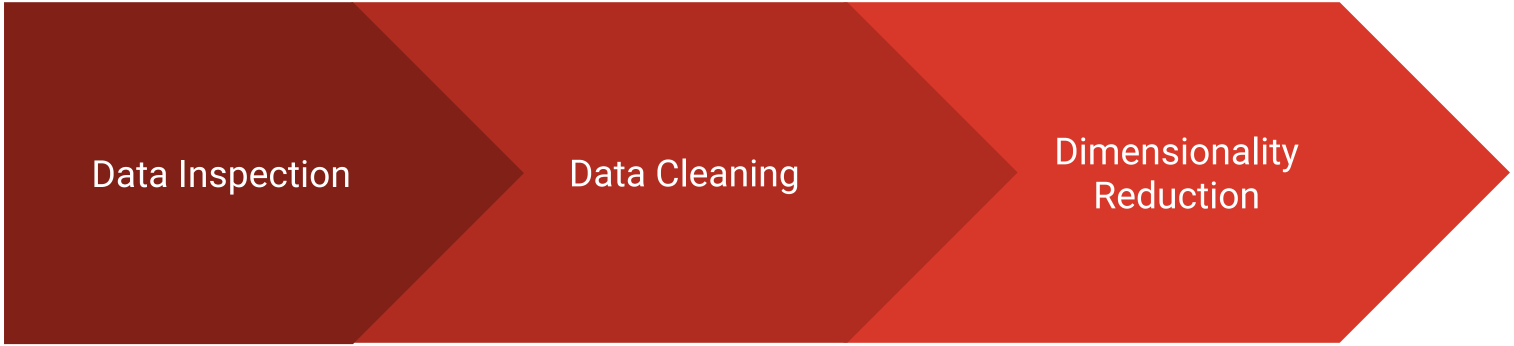
Drop Variables to Improve Quality of Data

- 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



Forward Selection: Systematic Approach



Data Inspection

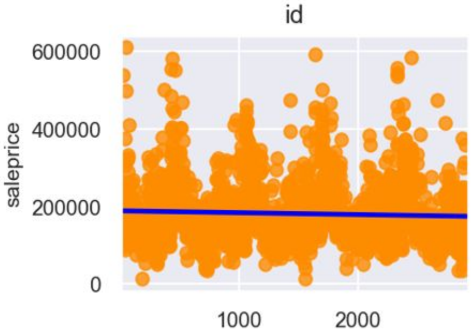
Data Cleaning

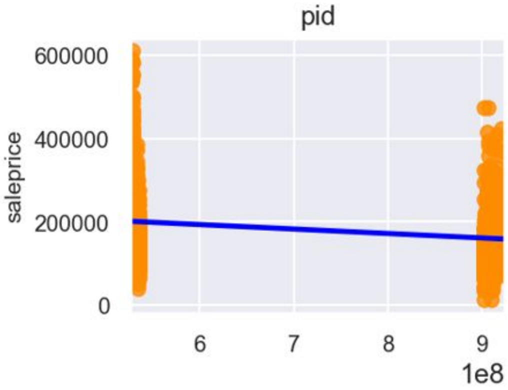
Dimensionality Reduction

Forward Selection: Systematic Approach

Data Inspection

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

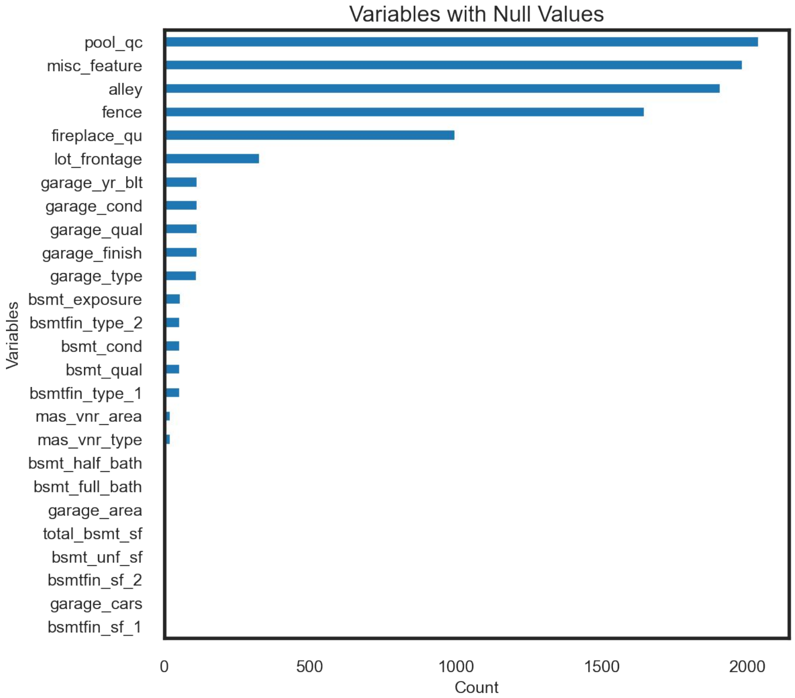




Forward Selection: Systematic Approach

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.

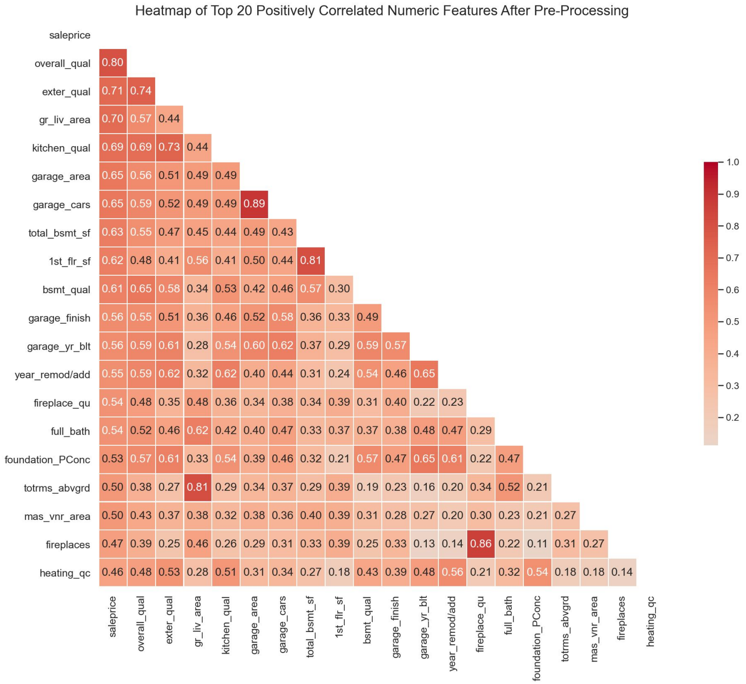
- Check for anomalous data and correct/ remove them, if necessary.
    - For example, ‘future-dated’ ‘Year’ values.

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.



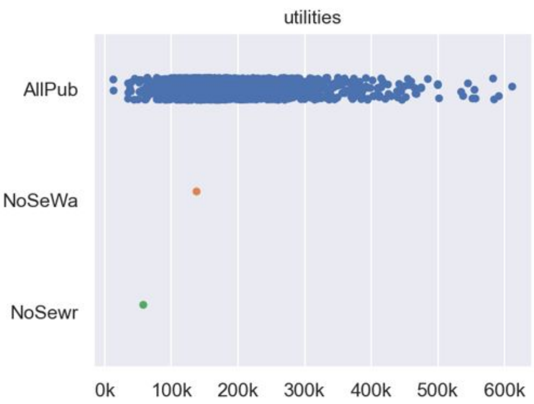
Heatmap of Top 20 Positively Correlated Numeric Features After Pre-Processing

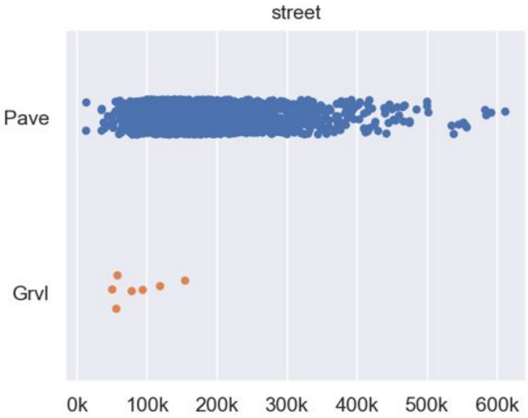
	saleprice	overall_qual	exter_qual	gr_liv_area	kitchen_qual	garage_area	garage_cars	total_bsmt_sf	1st_flr_sf	bsmt_qual	garage_finish	garage_yr_blt	year_remod/add	fireplace_qu	full_bath	foundation_PConc	totrms_abvgrd	mas_vnr_area	fireplaces	heating_qc
saleprice																				
overall_qual	0.80																			
exter_qual	0.71	0.74																		
gr_liv_area	0.70	0.57	0.44																	
kitchen_qual	0.69	0.69	0.73	0.44																
garage_area	0.65	0.56	0.51	0.49	0.49															
garage_cars	0.65	0.59	0.52	0.49	0.49	0.89														
total_bsmt_sf	0.63	0.55	0.47	0.45	0.44	0.49	0.43													
1st_flr_sf	0.62	0.48	0.41	0.56	0.41	0.50	0.44	0.81												
bsmt_qual	0.61	0.65	0.58	0.34	0.53	0.42	0.46	0.57	0.30											
garage_finish	0.56	0.55	0.51	0.36	0.46	0.52	0.58	0.36	0.33	0.49										
garage_yr_blt	0.56	0.59	0.61	0.28	0.54	0.60	0.62	0.37	0.29	0.59	0.57									
year_remod/add	0.55	0.59	0.62	0.32	0.62	0.40	0.44	0.31	0.24	0.54	0.46	0.65								
fireplace_qu	0.54	0.48	0.35	0.48	0.36	0.34	0.38	0.34	0.39	0.31	0.40	0.22	0.23							
full_bath	0.54	0.52	0.46	0.62	0.42	0.40	0.47	0.33	0.37	0.37	0.38	0.48	0.47	0.29						
foundation_PConc	0.53	0.57	0.61	0.33	0.54	0.39	0.46	0.32	0.21	0.57	0.47	0.65	0.61	0.22	0.47					
totrms_abvgrd	0.50	0.38	0.27	0.81	0.29	0.34	0.37	0.29	0.39	0.19	0.23	0.16	0.20	0.34	0.52	0.21				
mas_vnr_area	0.50	0.43	0.37	0.38	0.32	0.38	0.36	0.40	0.39	0.31	0.28	0.27	0.20	0.30	0.23	0.21	0.27			
fireplaces	0.47	0.39	0.25	0.46	0.26	0.29	0.31	0.33	0.39	0.25	0.33	0.13	0.14	0.86	0.22	0.11	0.31	0.27		
heating_qc	0.46	0.48	0.53	0.28	0.51	0.31	0.34	0.27	0.18	0.43	0.39	0.48	0.56	0.21	0.32	0.54	0.18	0.18	0.14	

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.

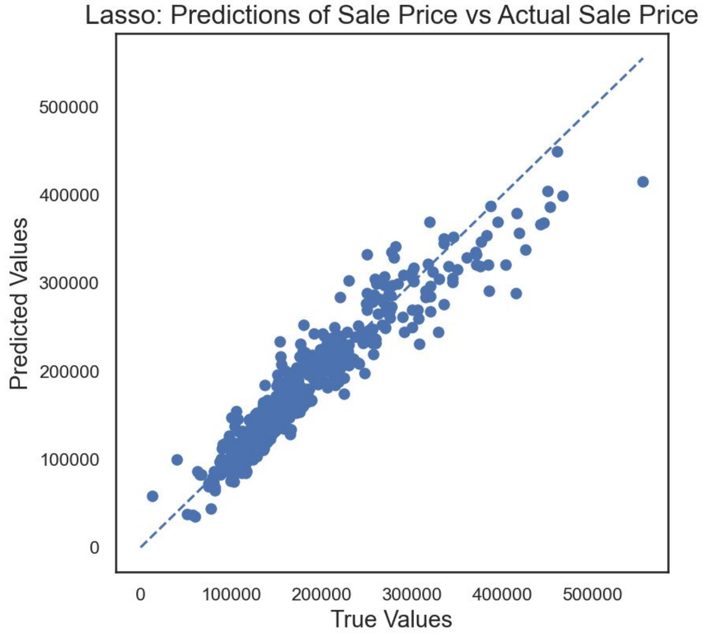


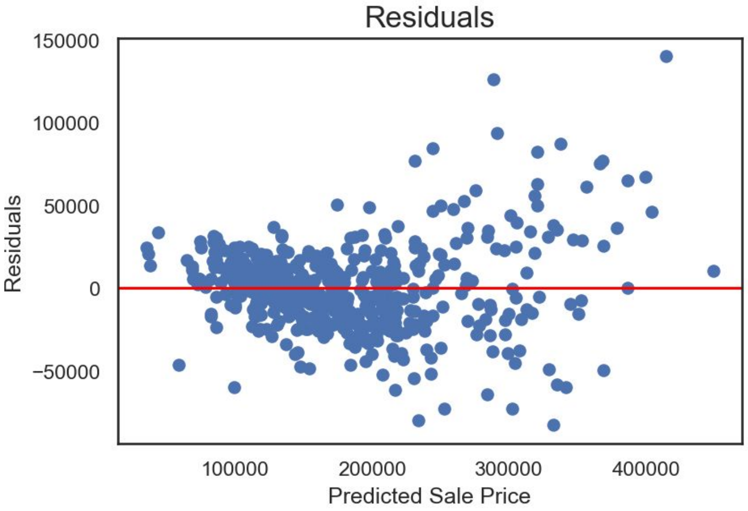


Forward Selection: Regression Model

Best performing model → Lasso Regression

Training R2	88.3%
Test R2	89.2%





Backward Selection

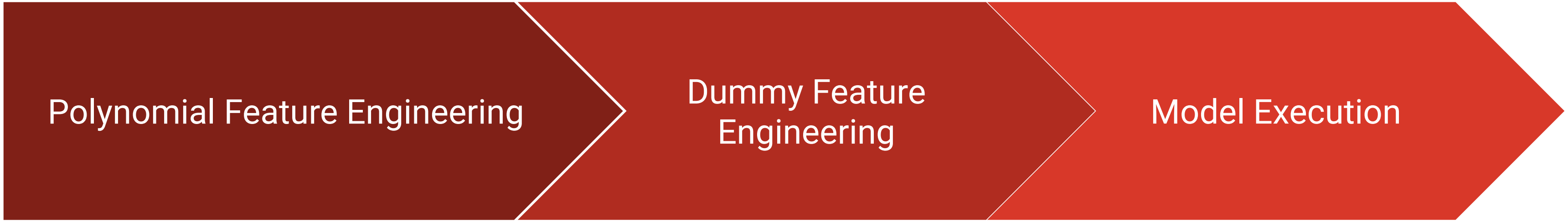
Underlying principle - More Data is Better. There are hidden interactions and nuances behind different data points that we want the machine and algo to pick up.

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 multicollinerarity 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

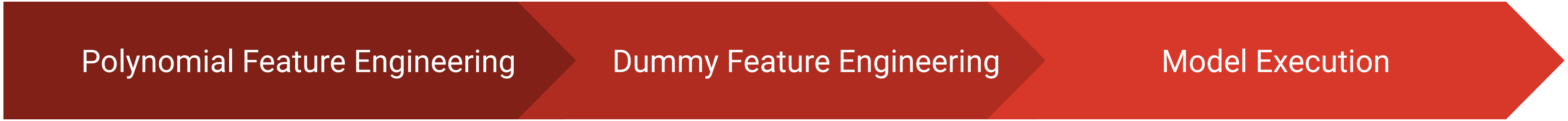
Dummy Feature Engineering

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

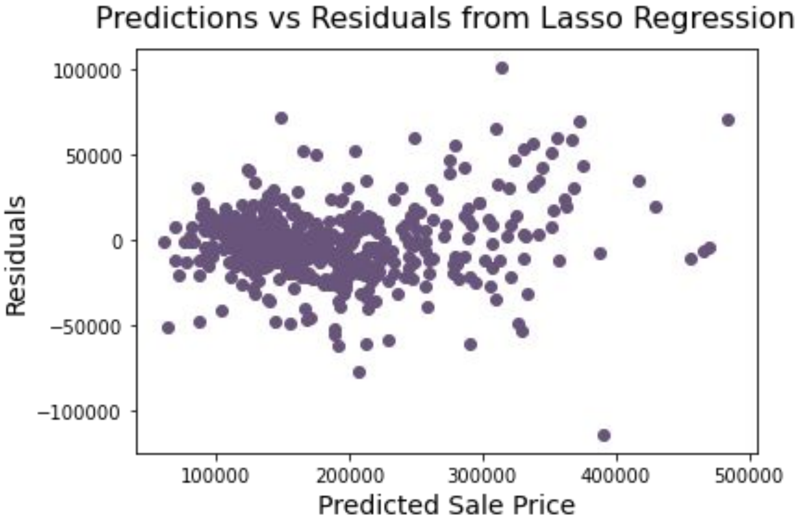
1. Reassemble the training data and experiment with different parameters across the major model classes.
2. Deploy optimisation techniques to identify the optimal parameters where possible.
3. Decide on the best performing and interpretable model.

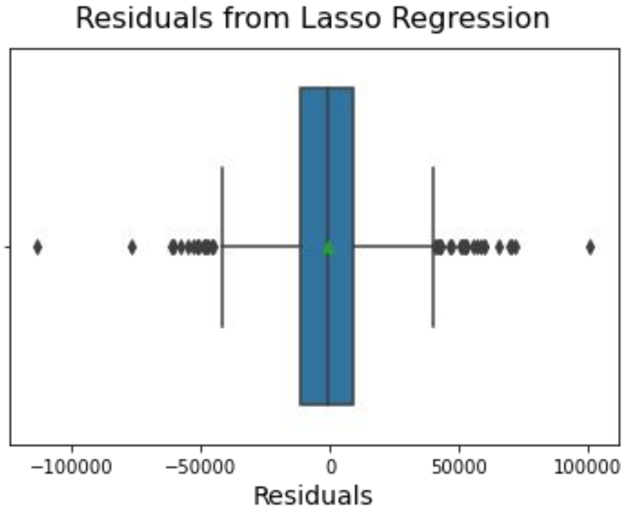
Backward Selection

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%

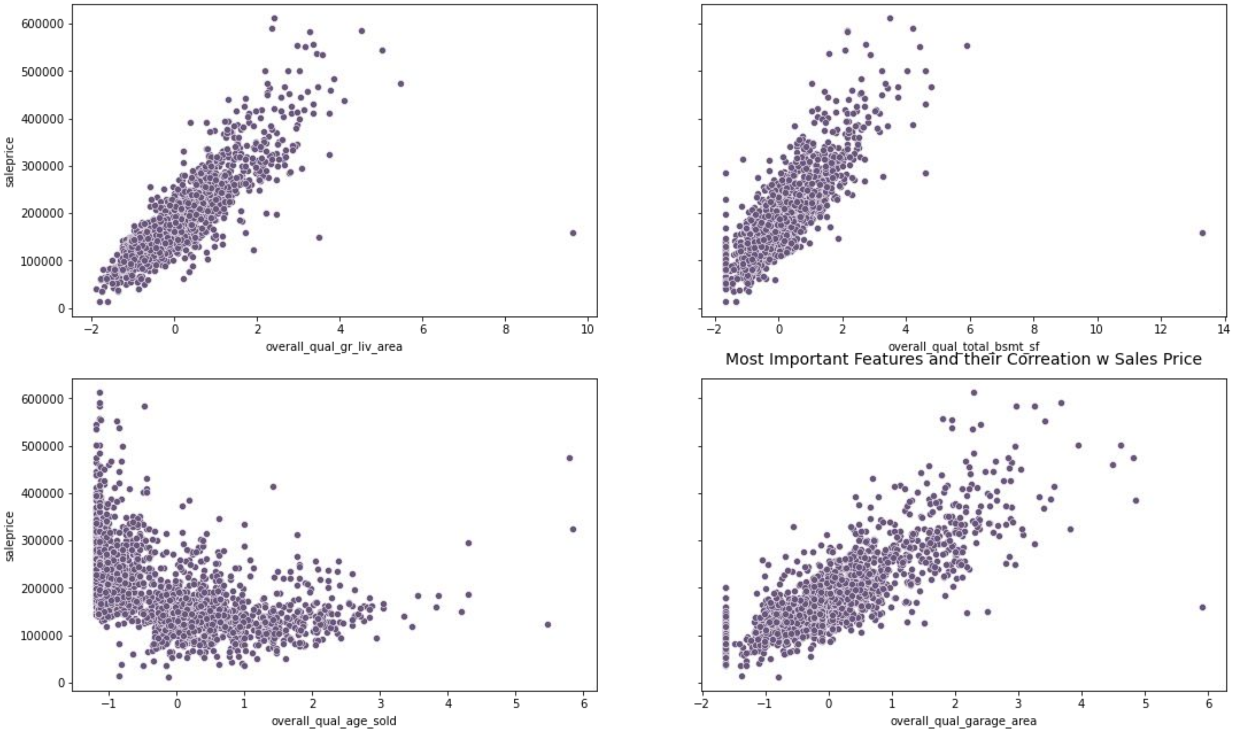




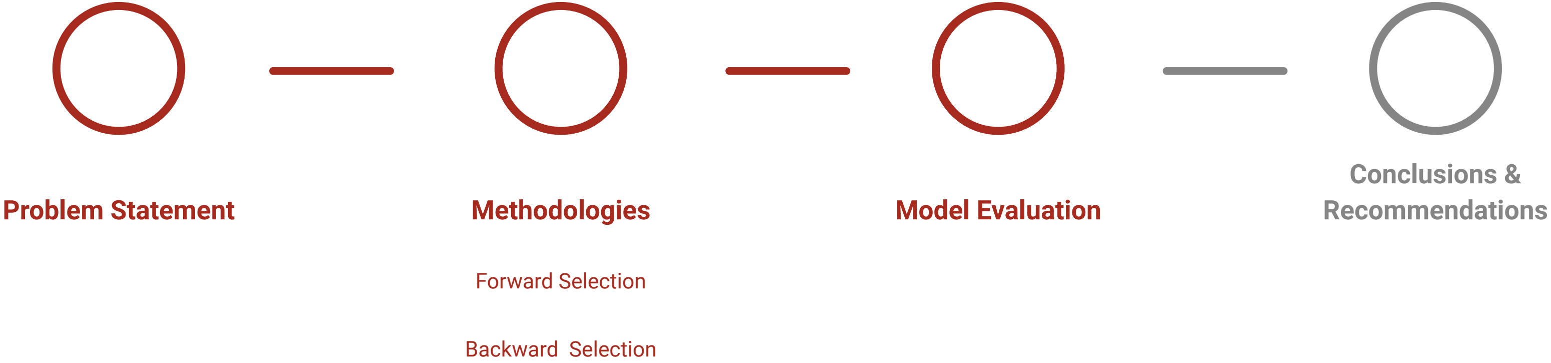
Backward Selection - Selected Features by Lasso

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.



Most Important Features and their Correation w Sales Price



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
Tr	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%
Te	92.8%	88.4%	89.2%	88.7%	87.5%	88.1%

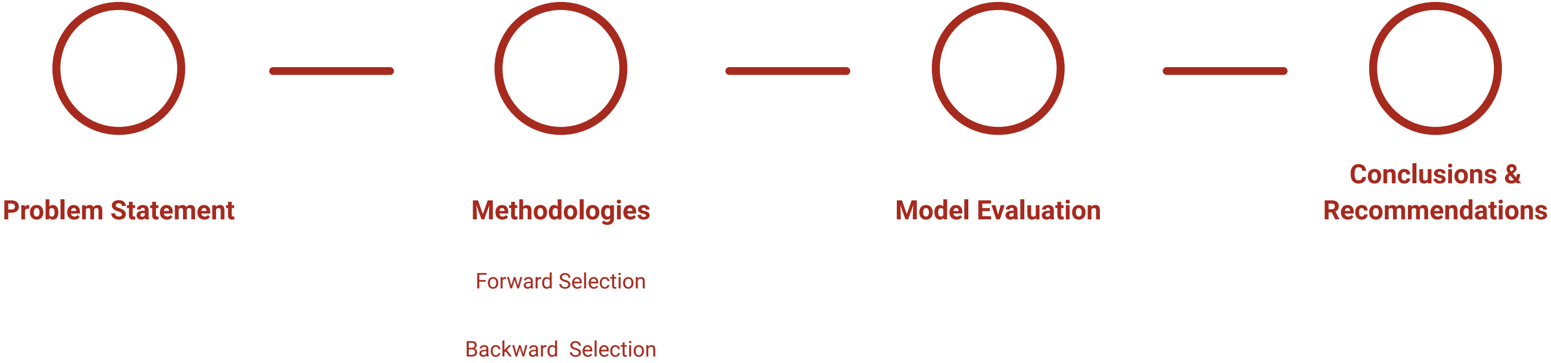
Features

Train R2	91.9%	87.7%	88.3%	85.9%	88.7%	86.9%

Model Evaluation

- Backward selection method gave the highest R^2 with a large number of coefficients.

- The forward approaches were neater, more time and computationally efficient, and easier to understand. However, this resulted in lower R^2 compared to Backward selection.



Problem Statement

Methodologies

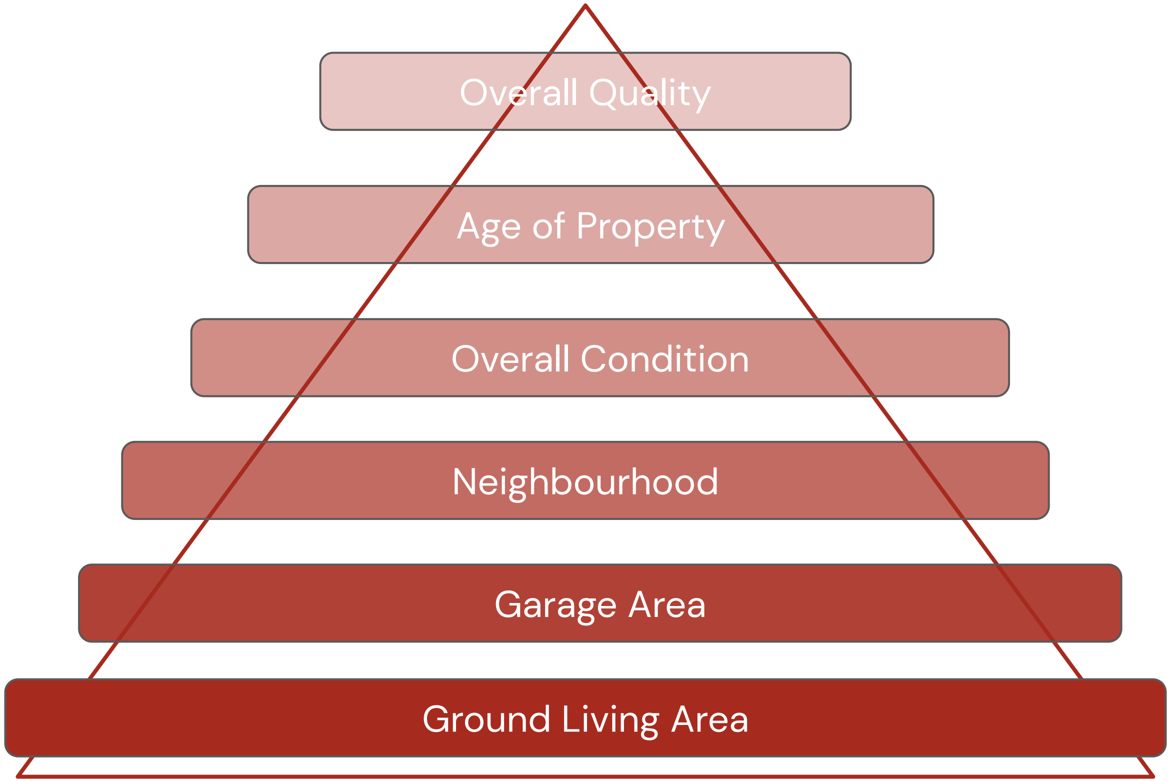
Forward Selection

Backward Selection

Model Evaluation

Conclusions &
Recommendations

Conclusion



Overall Quality

Age of Property

Overall Condition

Neighbourhood

Garage Area

Ground Living Area

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. Broden target market by positioning it as an up and coming city with education district.

T1
H4
A1
N1
K5
S1