

# **BEST FEATURES**

TO FETCH HIGHER  
HOUSE SALE  
PRICES

---

Group 2: Adi, Priscilla, Yong Lim, Zhi Hong





## TABLE OF CONTENTS

**01**

### **Background and Problem Statement**

Who we are and what we are trying to solve

**02**

### **Dataset**

The data we are using and our preparation

**03**

### **Exploratory Data Analysis**

Finding trends in our data

**04**

### **Features Engineering**

Preparing our data on features for modelling

**05**

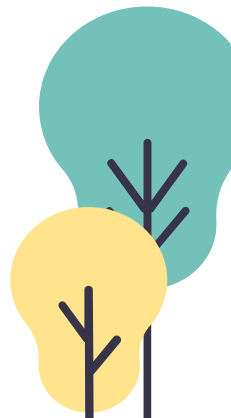
### **Modeling Sale Price**

Using regression model to predict sale prices

**06**

### **Findings & Recommendations**

Summarising our findings and future works



## Background

### OUR COMPANY

MyHome is an online property listing platform that connects home sellers with potential buyers. The website also provides data-driven recommendations on price trends based on details of the listings to help users optimise their bid/sell prices.

### PROBLEM STATEMENT

As analysts with MyHome, the aim of this study is to:

- 1) Create a regression model to predict home sale prices based on the listing details
- 2) Recommend the top 4 features that can fetch higher sale prices for residential properties

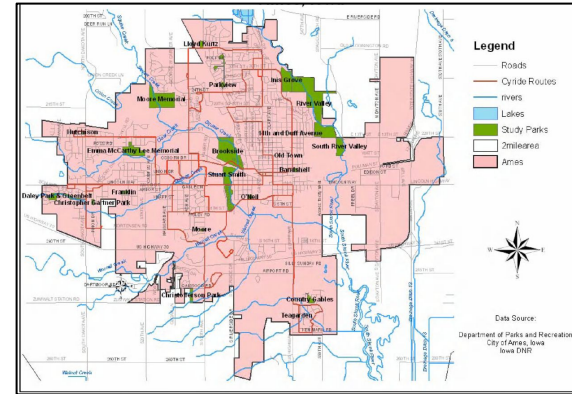


**MYHOME**  
REAL ESTATE AT A CLICK

## Data Used for Analysis

### Ames housing dataset (2006 – 2010)

- 80 variables, 2051 sales entries
- Variables such as sale prices, house size, quality, location, etc
- Continuous, ordinal and nominal variables



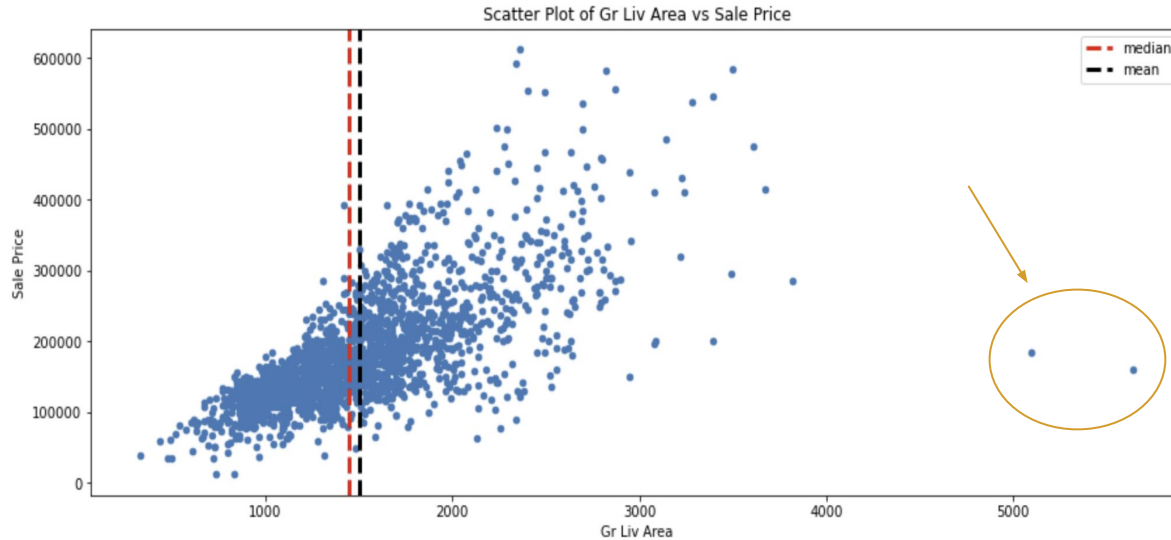
**Map of neighbourhood  
locations in Ames**



## Data Cleaning

- 1. Rename column headers to more intuitive names**
- 2. Change variable data type to facilitate calculations later**
- 3. Map ordinal variables to numeric scale**
- 4. Removed outlier data rows based on sale price**
- 5. Check and impute missing values**
- 6. Drop columns**
  - Too many missing data points
  - Too many of the same values within the column (80% threshold)
  - No significant sale price change observed across categories
  - Identifiers (does not affect trend)
  - Replaced by engineered variables (to be discussed later)

# Outliers based on Living Area vs Sale Price

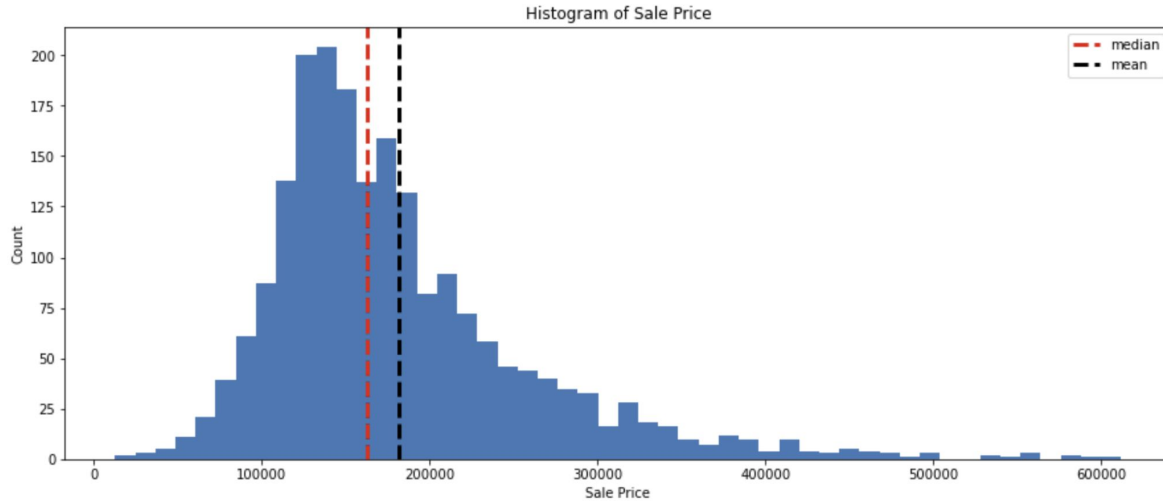


2 outliers are found based on the data:

- Living Area > 4,000 sqft
- Sale Price < \$300,000

Inappropriate interpretation of the Sale Price based on the Living Area sqft

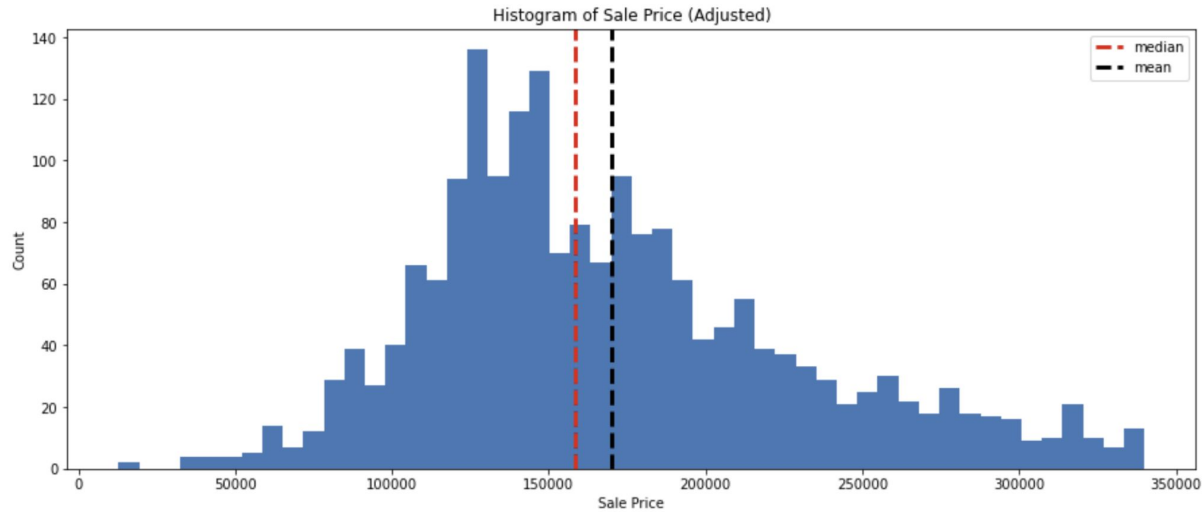
# Outliers based on Sale Price



- To remove high variability outside the upper and lower quartiles on the Sale Price
- Skewness = 1.556



# Adjusted Sale Price Histogram



- 95 rows are removed
- Skewness after adjustment = 0.66





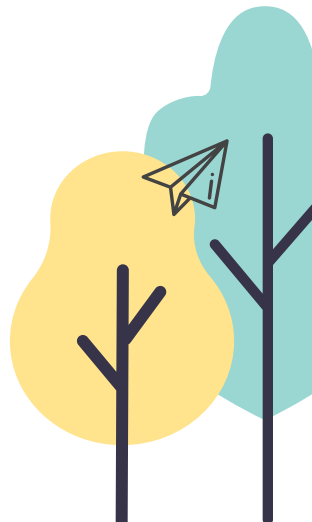
# Nominal Features to be Excluded

From the histogram of the nominal features, the features below will not be significant since more than **80%** is the same type thus will not be considered:

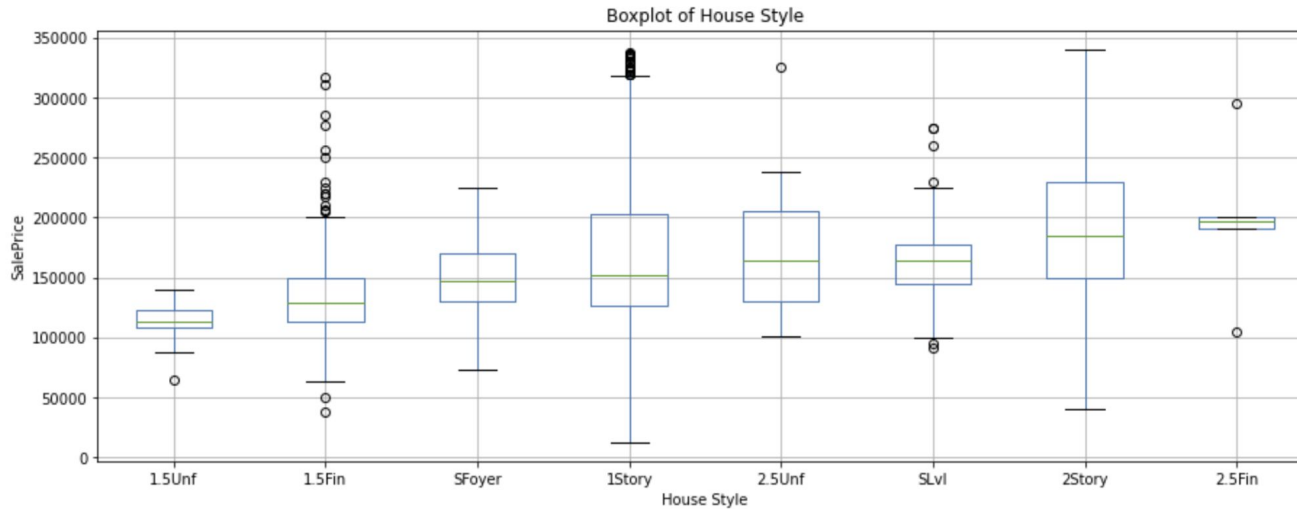
- Street
- Alley
- Land Contour
- Condition 1
- Condition 2
- Bldg Type
- Roof Style
- Roof Matl
- Heating
- Central Air
- Misc Feature
- Sale Type

The features also will not be included in the model:

- MS SubClass, overlapped with Bldg Type & Year Built and difficult to interpret
- MS Zoning, overlapped with Neighborhood
- PID, for indexing only
- ID, for indexing only



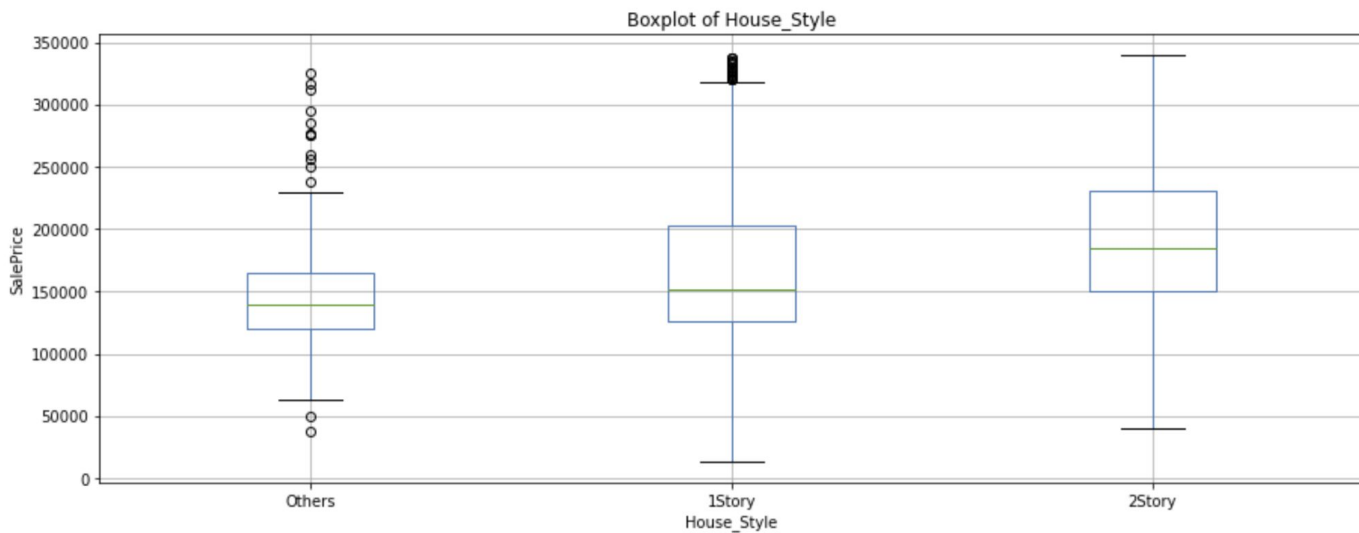
# Boxplot of House Style with Sale Price



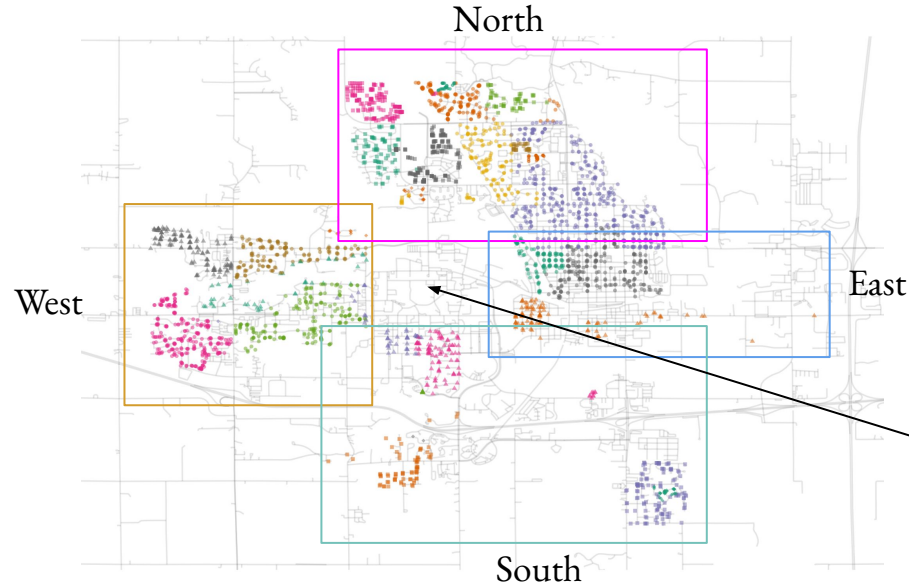
House Style	Proportion
1 Story	0.515
2 Story	0.286
1.5 Fin	0.111
SLvl	0.047
SFoyer	0.026
2.5 Unf	0.007
1.5 Unf	0.006
2.5 Fin	0.003



# Boxplot of House\_Style with Sale Price



# Neighborhood Map of Ames



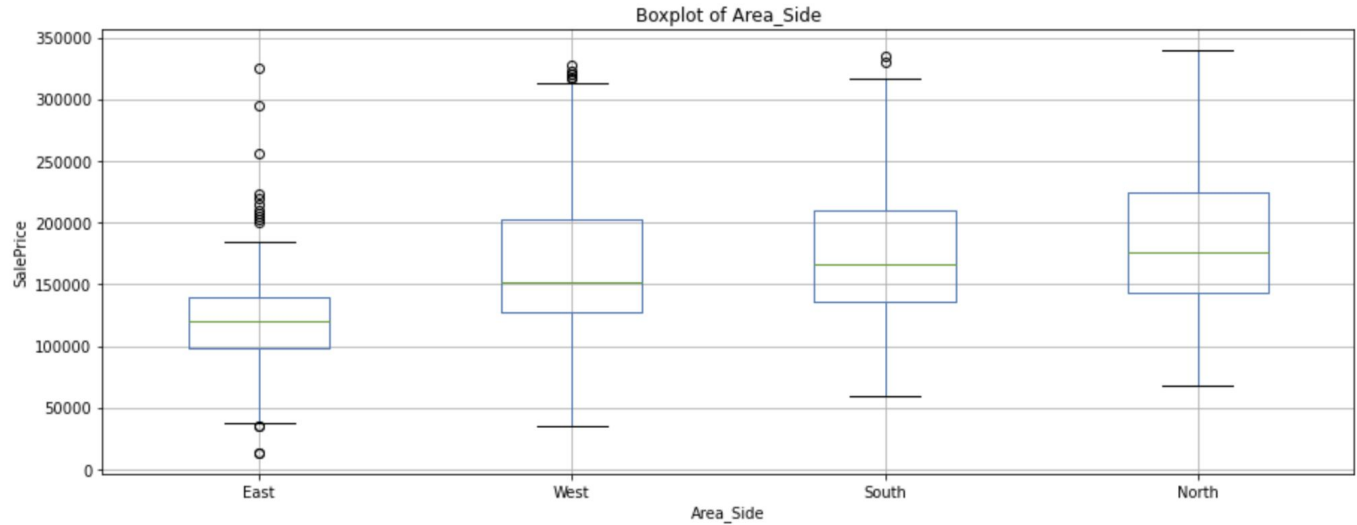
Group based on the location as of the IOWA State University

Iowa State University

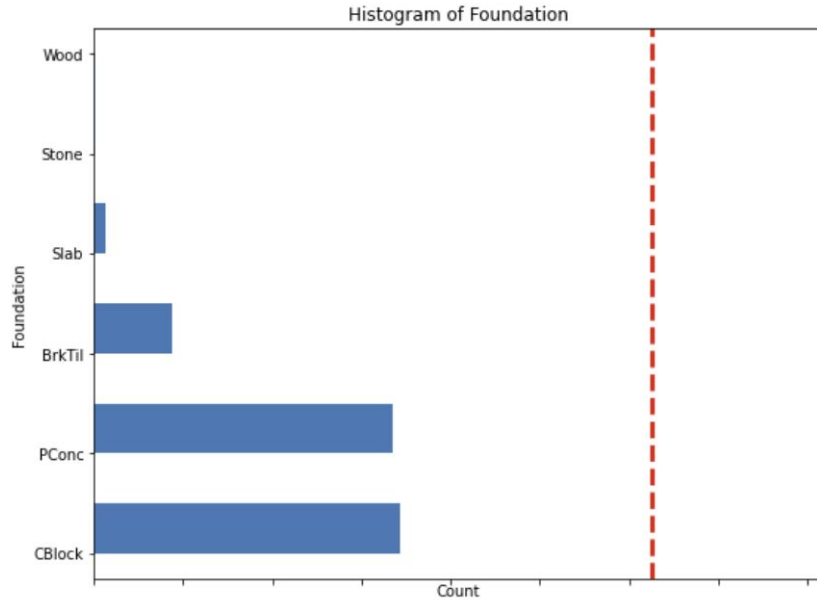
- |                      |                  |   |                       |               |
|----------------------|------------------|---|-----------------------|---------------|
| • North_Ames         | • Gilbert        | • Crawford                                | • Clear_Creek         | • Blueste     |
| • College_Creek      | • Sawyer         | • Iowa_DOT_and_Rail_Road                  | • Meadow_Village      | • Greens      |
| • Old_Town           | • Northwest_Ames | • Timberland                              | • Briardale           | • Green_Hills |
| • Edwards            | • Sawyer_West    | • Northridge                              | • Bloomington_Heights | • Landmark    |
| • Somerset           | • Mitchell       | • Stone_Brook                             | • Veenker             |               |
| • Northridge_Heights | • Brookside      | • South_and_West_of_Iowa_State_University | • Northpark_Villa     |               |

# Correlation of Area with Sale Price

After grouping, the feature to be used for modelling to check if location affects the sale price



# Histogram of Foundation



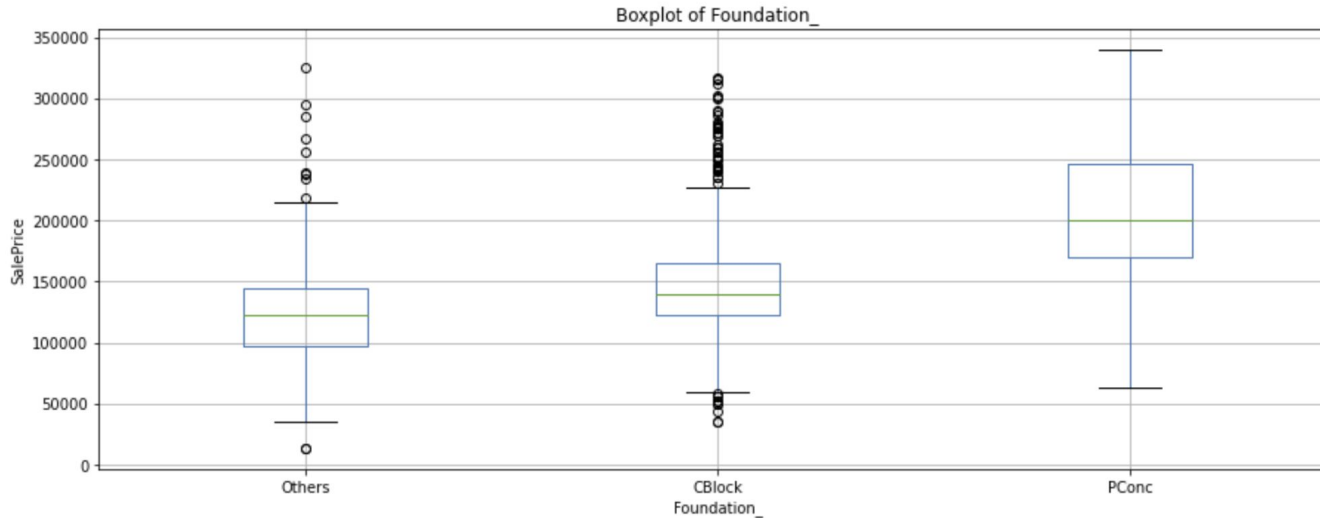
Grouping based on the value counts into 3 categories:

- Cinder Block
- Poured Concrete
- Others

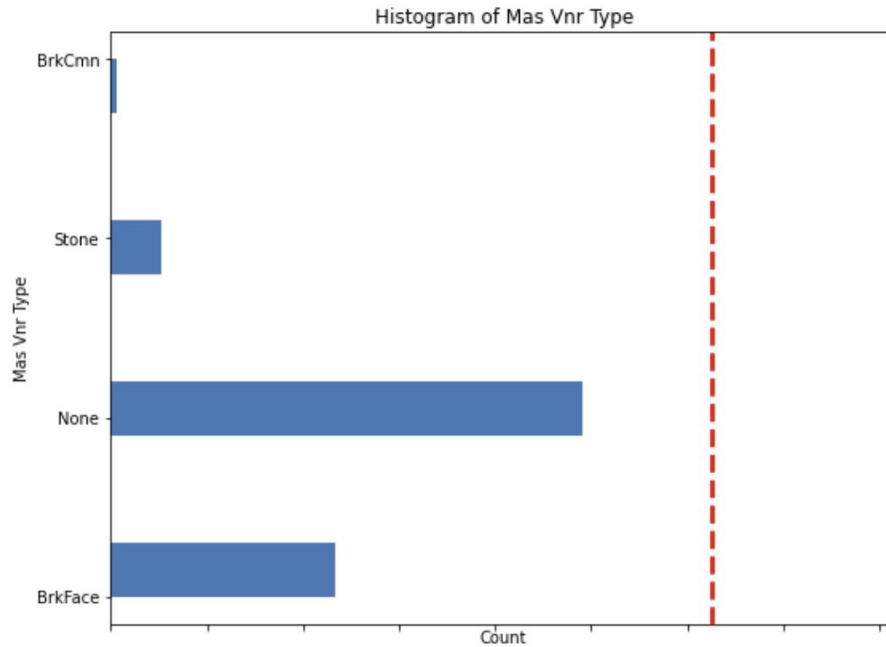


# Correlation of Newly Categorised Foundation with Sale Price

There is slight difference in the interquartile range between each category



# Histogram of Masonry Veneer Type



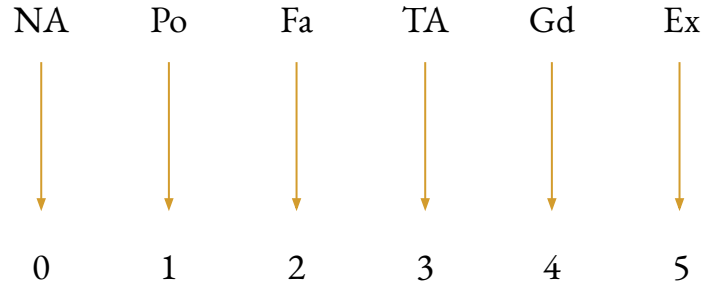
- Grouping based on availability



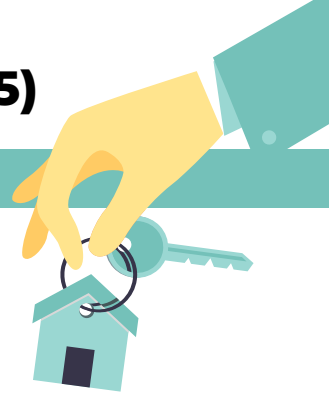


# Mapping Ordinal Features

- Ordinal features values are to be replaced with numeric scale with 0 as the lowest)
- Combined with Continuous Feature to find the correlation with SalePrice and features



# Correlation of Numerical Features with Sale Price ( $p > 0.5$ )



Features	Correlation (p)
Overall Qual	0.789617
Gr Liv Area	0.677887
Exter Qual	0.669849
Kitchen Qual	0.647822
Garage Cars	0.626066
Garage Area	0.615088
Year Built	0.600998
Total Bsmt SF	0.589474
Bsmt Qual	0.588699
Year Remod/Add	0.573572
1st Flr SF	0.566783
Garage Finish	0.566342
Full Bath	0.556303
Fireplace Qu	0.514253

- Quality for every area has high correlation with the Sale Price
- Area also has high correlation
- High collinearity features:
  - Fireplaces and Fireplace Qu ( $p = 0.86$ )
  - Overall Qual and Exter Qual ( $p = 0.72$ )
  - Total Bsmt SF and 1st Flr SF ( $p = 0.79$ )
  - TotRmsAbvGrd and Gr Liv Area ( $p = 0.8$ )
  - Garage Cars and Garage Area ( $p = 0.89$ )

## Feature Engineering (1 of 2)

- Total Bath:

$$\text{Full Bath} + 0.5(\text{Half Bath}) + \text{Bsmt Full Bath} + 0.5(\text{Bsmt Half Bath})$$

- Bedroom to Bathroom Ratio:

$$\text{Total Bedroom} / \text{Total Bath}$$

- Age of the property as of the year sold:

$$\text{Year Sold} - \text{Year Built}$$

- Age of the property as of last renovation:

$$\text{Year Sold} - \text{Year Add/Remod}$$

- Basement QC:

$$\text{Total Basement Area} \times \text{Basement Qual}$$



## Feature Engineering (2 of 2)

- Fireplace QC:

$$\text{No of Fireplaces} \times \text{Fireplace Quality}$$

- Overall + External Quality:

$$\text{Overall Quality} \times \text{External Quality}$$

- Area Room Ratio:

$$\text{Gr Liv Area} / \text{TotRmsAbvGrd}$$

- Garage Ratio:

$$\text{Garage Area} / \text{Garage Cars}$$

- Ext\_Facilities\_Area:

$$\text{Wood Deck Area} + \text{Oper Porch Area} + \text{Enclosed Porch Area} + \text{Three Season Porch Area} + \text{Screen Porch Area}$$



# Correlation of New Features with Sale Price



Features	Correlation (p)
Total_Bath	0.64
BR_Bath_Ratio	-0.43
Age	0.6
Age_Renov	-0.57
Bsmt_QC	0.61
Fireplace_QC	0.5
Ovrll_Exter_Qual	0.79
Area_Rm_Ratio	0.56
Garage_Ratio	-0.031
Ext_Facilities_Area	0.37

- Features to be retained with p value > 0.5:
  - Total\_Bath, remove Full Bath
  - Age, remove Year Built
  - Age\_Renov, remove Year Add/Remod
  - Bsmt QC, remove Total Bsmt SF
  - Area\_Rm\_Ratio
- Garage Cars feature to be removed as well  
(Garage Area p-value higher than Garage Cars)

# Features to be Included in the Model

**TOTAL FEATURES : 26**

## **NOMINAL FEATURES (DUMMIFIED):**

- House Style
- Foundation
- Neighborhood Area
- Mas Veneer Availability

## **NEW FEATURES :**

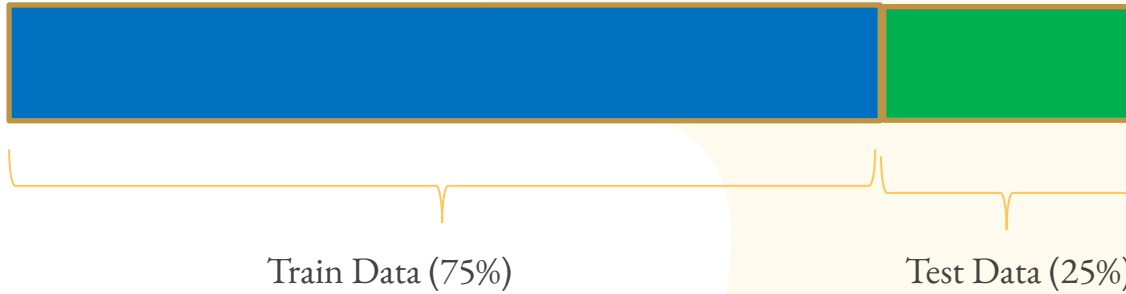
- Total\_Bath
- Age
- Age\_Renov
- Bsmt\_QC
- Area\_Rm\_Ratio

## **NUMERICAL FEATURES :**

- Overall Qual
- Exter Qual
- Bsmt Qual
- 1st Flr SF
- Gr Liv Area
- Kitchen Qual
- Fireplace Qu
- Garage Finish
- Garage Area



# Train Test Split



# Regression Model

## Linear Regression



Score: 0.85285

Cross-Validation

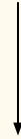
## Lasso Regression



Score: 0.85272

Cross-Validation

## Ridge Regression



Score: 0.85295

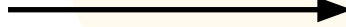




# $R^2$ Accuracy

## Ridge Regression

$R^2$  Score  
0.826154



Root Mean Square Error (RMSE)  
22392

---

## Linear Regression

$R^2$  Score  
0.8263



Root Mean Square Error (RMSE)  
22393

## Lasso Regression

$R^2$  Score  
0.8257



Root Mean Square Error (RMSE)  
22393



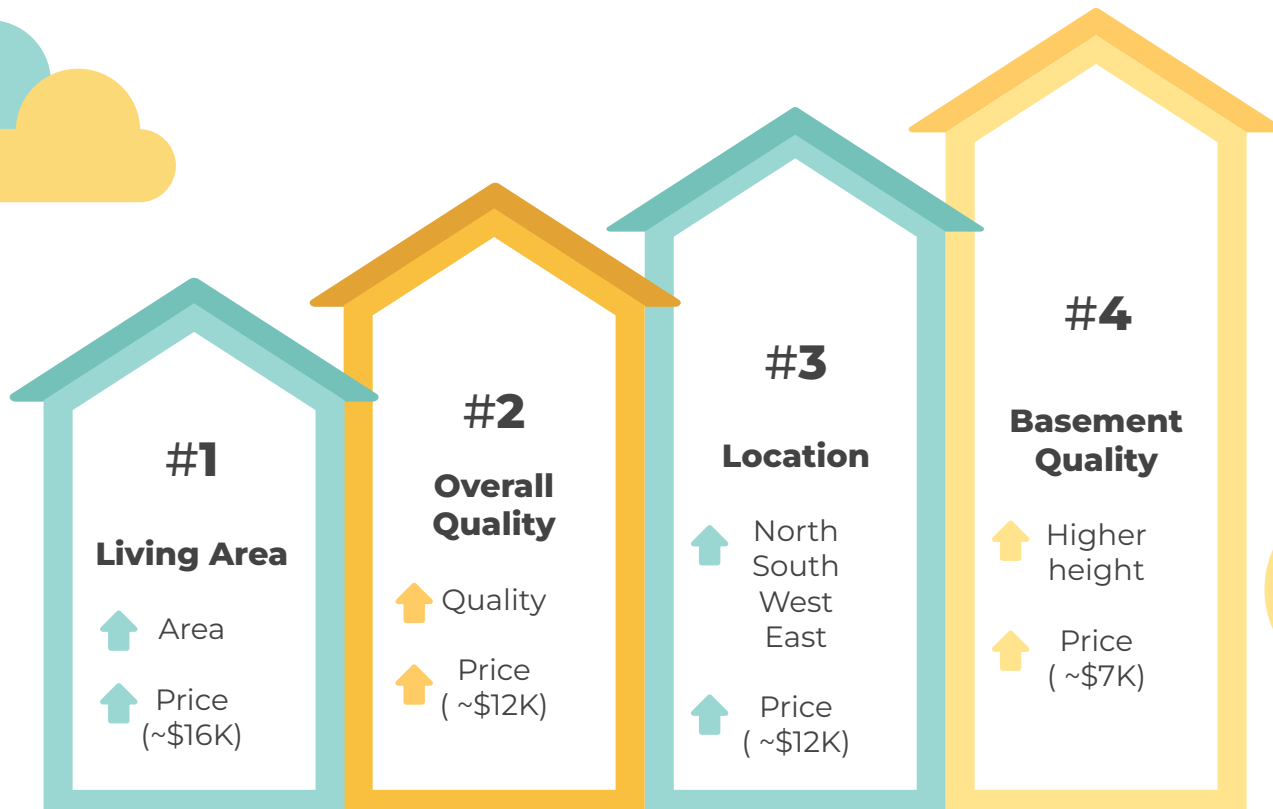


## Conclusions

1. Various area and quality of the house showed a strong correlation with the housing prices
2. Neighborhood location also showed a strong correlation with the housing prices.
3. Ridge Regression shows the best performance in predicting the sale price of the house



## Recommendations





## Recommendations (For buyers)

1. Invest in renovation
2. Invest in:
  - Bigger house
  - House located in North
  - House with higher basement height
3. For buyers with budget constraint, keep a lookout on cheaper deal (e.g. East location) so that they can allocate their budget accordingly based on their preference





## Recommendations (For sellers)

1. Highlight the quality of the house
2. Highlight the basement height
3. Ask for higher price if has a big house or located in North





## Future Actions

1. Update data set with current year data 2011
2. Important features to make it compulsory during data collection to reduce missing values
3. Research on new features in view home buyer/seller's perspective may change (e.g. with younger generation buying house)



## OUR TEAM

**Adi Hartanto Kusuma**

**Priscilla Ng**

**Tan Yong Lim**

**Kwek Zhi Hong**



# THANKS

Does anyone have any questions?

myhome.com



CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), and infographics & images by [Freepik](#).

**Please keep this slide for attribution.**







## Data Cleaning

- **Rename columns to more intuitive names**
- **Change variable data type to facilitate calculations later**
- **Map ordinal variables to numeric scale**
- **Removed outlier data rows based on sale price**
  
- **Check and impute missing values**
  - With suitable value depending on data type and inference from other columns
  
- **Drop columns**
  - Too many missing data points
  - Too many of the same values within the column (80% threshold)
  - No significant sale price change observed across categories
  - Identifiers (does not affect trend)
  - Replaced by engineered variables (to be discussed later)



## Dropped columns

- **Low correlation with sales price (corr < 0.5)**

'Heating QC', 'Fireplaces', 'TotRms AbvGrd', 'BsmtFin Type 1', 'BsmtFin SF 1', 'Mas Vnr Area', 'Paved Drive', 'Bsmt Exposure', 'Open Porch SF', 'Garage Qual', 'Garage Cond', 'Wood Deck SF', 'Electrical', 'Lot Frontage', 'Half Bath', '2nd Flr SF', 'Bsmt Cond', 'Bsmt Full Bath', 'Lot Area', 'Bsmt Unf SF', 'Bedroom AbvGr', 'Functional', 'Screen Porch', 'Exter Cond', '3Ssn Porch', 'Pool QC', 'Pool Area', 'Utilities', 'Mo Sold', 'BsmtFin Type 2', 'BsmtFin SF 2', 'Misc Val', 'Yr Sold', 'Bsmt Half Bath', 'Land Slope', 'Overall Cond', 'Low Qual Fin SF', 'Kitchen AbvGr', 'Enclosed Porch', 'Fence', 'Lot Shape'

- **Identified as non-important features**

'Year Built', 'Year Remod/Add', 'Full Bath', 'Total Bsmt SF', 'Garage Area', 'Garage Year Built'

- **More than 80% of same values:**

'Street', 'Alley', 'Land Contour', 'Condition 1', 'Condition 2', 'Bldg Type', 'Roof Style', 'Roof Matl', 'Heating', 'Central Air', 'Misc Feature', 'Sale Type'



## Dropped columns

### **Identifiers (no impact on trend):**

- ID and PID

### **Overlap with other variables:**

- MS Zoning, overlapped with Neighborhood
- MS SubClass, overlapped with Bldg Type & Year Built and difficult to interpret

### **No significant change in sale price between categories in the variable**

- Garage Type
- Lot config
- Exterior Material 1 and 2



## Other data cleaning steps

### **Treating null values**

- Lot Frontage: 330 rows filled with mean lot frontage value grouped by Lot Shape
- Masonry Veneer Type and Masonry Veneer Area: 22 rows changed to null (no feature in property)
- Basement Quality: 1 row filled with NA (no feature in property)
- Basement Exposure: 4 rows filled with NA (if feature exists) and null if no feature exists
- BsmtFin Type 2: Replace the null value with NA and Unf (based on the most value counts) if the SF is not zero
- Bsmt Full Bath and Bsmt Half Bath – input '0'

### **Change data type**

- Change the value of MS SubClass from int to string