



## PROJECT 2

# Ames Housing Analysis

Poa Miao Er  
Poh Yong Quan  
Tan Jun Pin



## Content

1. Introduction
2. Problem Statement
3. Dataset Introduction
4. Data Cleaning
5. EDA & Visualisation
6. Models
7. Findings
8. Conclusions
9. Recommendations

# Introduction

With the ever-rising cost of living, essential needs are getting more expensive.

Shelter or house always contributed to the biggest part of spending.

New home buyer at Ames, Iowa would like to own a home and invest at the same time

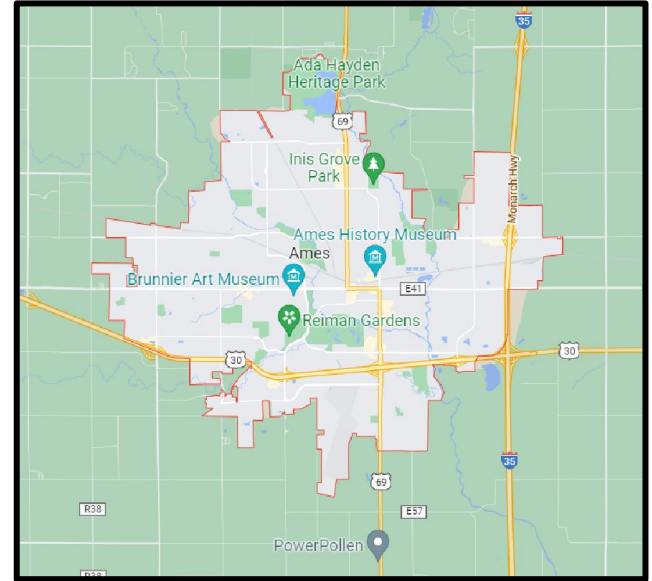
Houses are liability and asset at the same time. After purchasing as the price could increase in the future.



# Problem Statement

Ensuring new home buyers could yield highest return of investment from their new house purchase, the objective for our projects are:

1. To identify the best and worst permanent and non-permanent features affecting the sale price in Ames, Iowa.
2. To identify the non-permanent feature of house that could lead to high Sale Price after renovation.



# Datasets Introduction

- House Transaction at Ames, Iowa between the year 2006 to 2010
- Consists of 80 columns of the house features and the transaction price
- Features have both continuous variables and discrete variables
- List of complete data dictionary [\[here\]](#)

The screenshot shows the Kaggle competition page for 'DSI-US-11 Project 2 Regression Challenge'. At the top, it says 'Community Prediction Competition'. The challenge title is 'DSI-US-11 Project 2 Regression Challenge' with the subtitle 'Predict the price of homes at sale for the Ames Iowa Housing dataset'. It indicates '89 teams · 2 years ago'. A navigation bar includes 'Overview', 'Data' (which is highlighted), 'Code', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and a 'Late Submission' button. Below the navigation bar is a 'Data Description' section. It states 'There are three files:' and lists them: 'train.csv' (all training data), 'test.csv' (test data for predictions), and 'sample\_sub\_reg.csv' (example submission format). At the bottom, there is a link for 'Codebook / Data Dictionary'.

Community Prediction Competition

## DSI-US-11 Project 2 Regression Challenge

Predict the price of homes at sale for the Ames Iowa Housing dataset

89 teams · 2 years ago

Overview **Data** Code Discussion Leaderboard Rules Team My Submissions **Late Submission** ...

### Data Description

There are three files:

- **train.csv** -- this data contains all of the training data for your model.
  - The target variable ( `SalePrice` ) is removed from the test set!
- **test.csv** -- this data contains the test data for your model. You will feed this data into your regression model to make predictions.
- **sample\_sub\_reg.csv** -- An example of a correctly formatted submission for this challenge (with a random number provided as predictions for `SalePrice` . Please ensure that your submission to Kaggle matches this format.

[Codebook / Data Dictionary](#)

# Data Cleaning - Outliers

Error	Outliers	After Data Cleaning
Outliers in 'Lot_Area'	159,000 sqft & 115,149 sqft	Removed from dataset

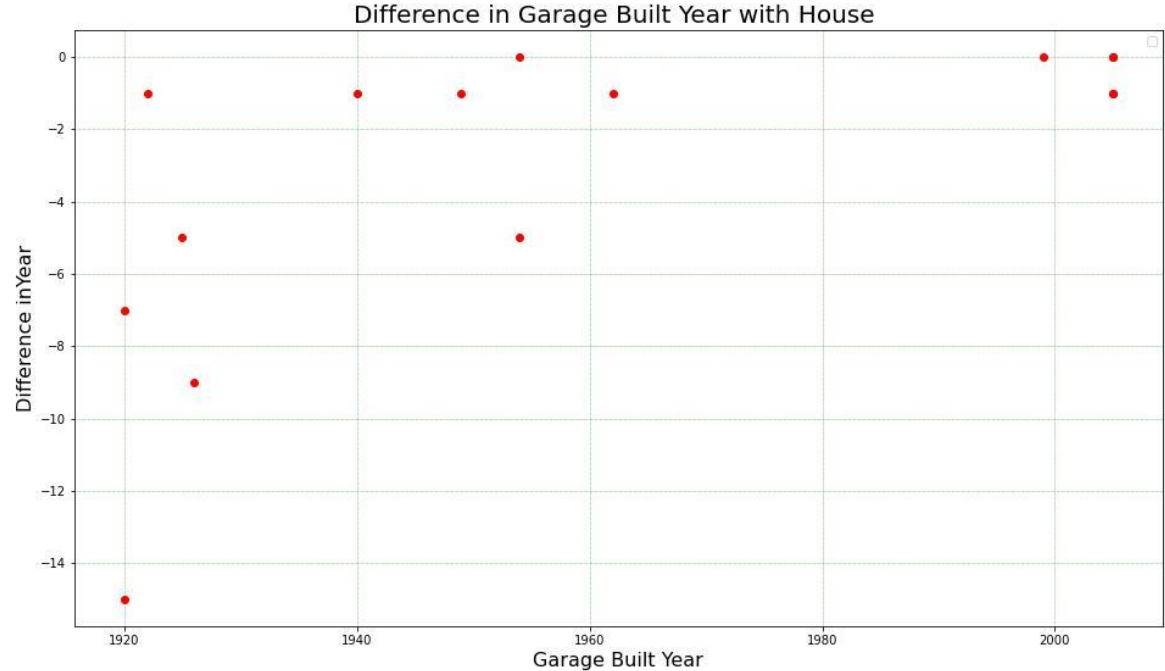


# Data Cleaning

Error	Before data Cleaning	After Data Cleaning
'Year_Remod/Add' is earlier than 'Year_Built'	Id = 851 Year_Built = 2002 Year_Remod/Add = 2001	Id = 851 Year_Built = 2002 Year_Remod/Add = 2002
'Year_Sold' is earlier than 'Year_Built', 'Year_Remod/Add' & 'Garage_Yr_Blt'	Id = 2261 Garage_Yr_Blt = 2207 Yr_Sold = 2007	Id = 2261 Garage_Yr_Blt = 2007 Yr_Sold = 2007
	Id = 2181 Year_Remod/Add = 2009 Yr_Sold = 2007	Id = 2181 Year_Remod/Add = 2009 Yr_Sold = 2009
	Id = 1703 Year_Remod/Add = 2008 Yr_Sold = 2007	Id = 1703 Year_Remod/Add = 2008 Yr_Sold = 2008
'Garage_Yr_Blt' is earlier than 'Year_Built'	0.5% (11 units) of the datasets has 'Garate_Yr_Blt' earlier than 'Year_Built'	<b>Remained</b> (To clarify in later part)

# Data Cleaning on 'Garage\_Yr\_Blt' is earlier than 'Year\_Built'

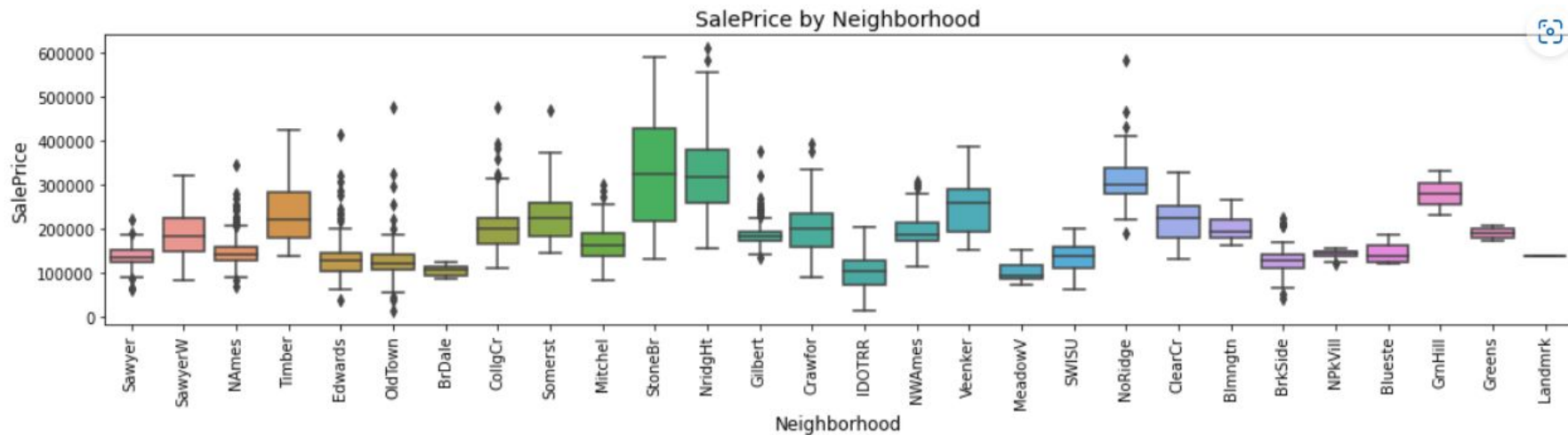
- 11 units has garage built before house
- 5 units built on 1920s
- 6 units built between 1930 - 2005
- More common in 1920s
- Homebuilders built garage first so they could live in it while financing and constructing the house \*
- No changes on the data



Source:

*\*THE GARAGE: ITS HISTORY AND PRESERVATION*

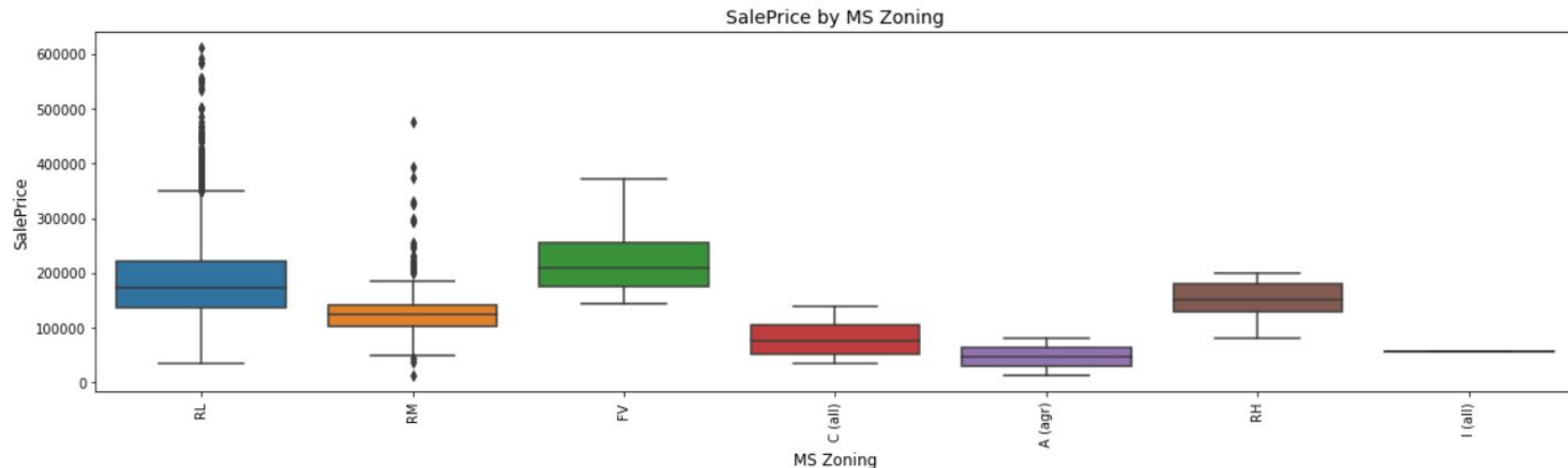
# EDA & Visualisation



- Stone Brook and Northridge Heights: nearly similar median price, not exceed each others' interquartile ranges
- Stone Brook and Mitchell: interquartile ranges do not intersect on the sale price scale

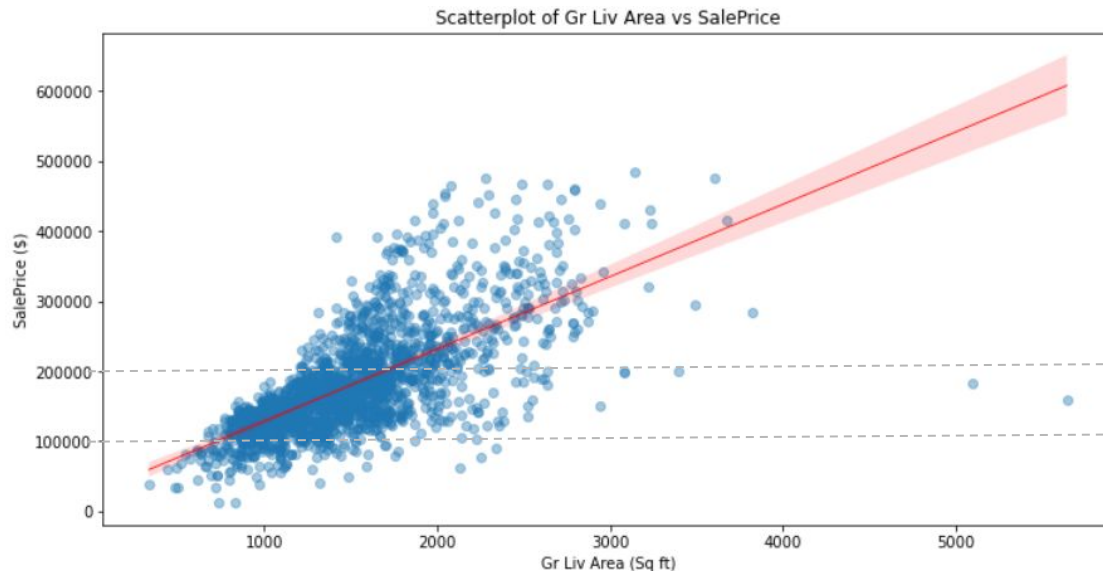


# EDA & Visualisation



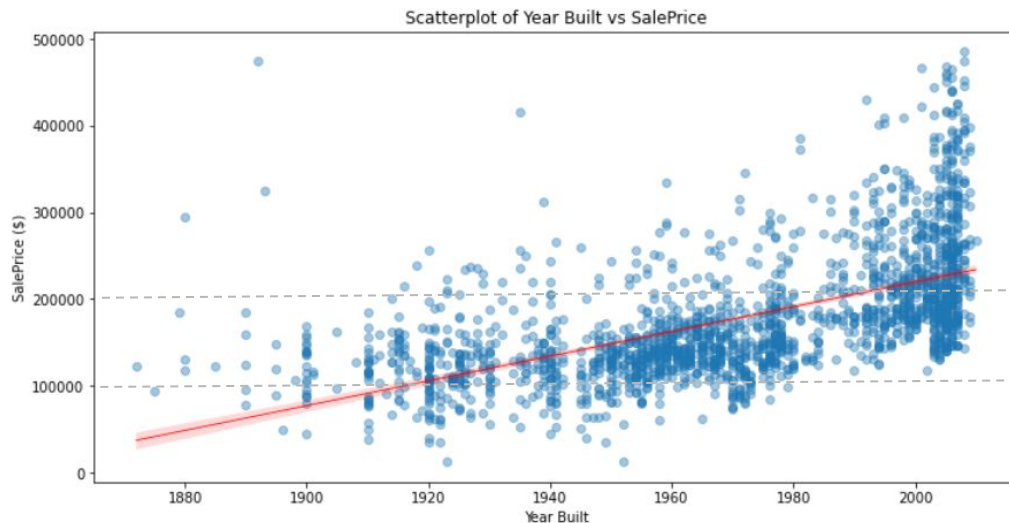
- The highest median: FV -- Floating Village Residential
- The lowest median: C(all) -- Commercial properties
- Outlier: RL -- Residential Low Density, probably due to poor zone and rich neighborhood

# EDA & Visualisation



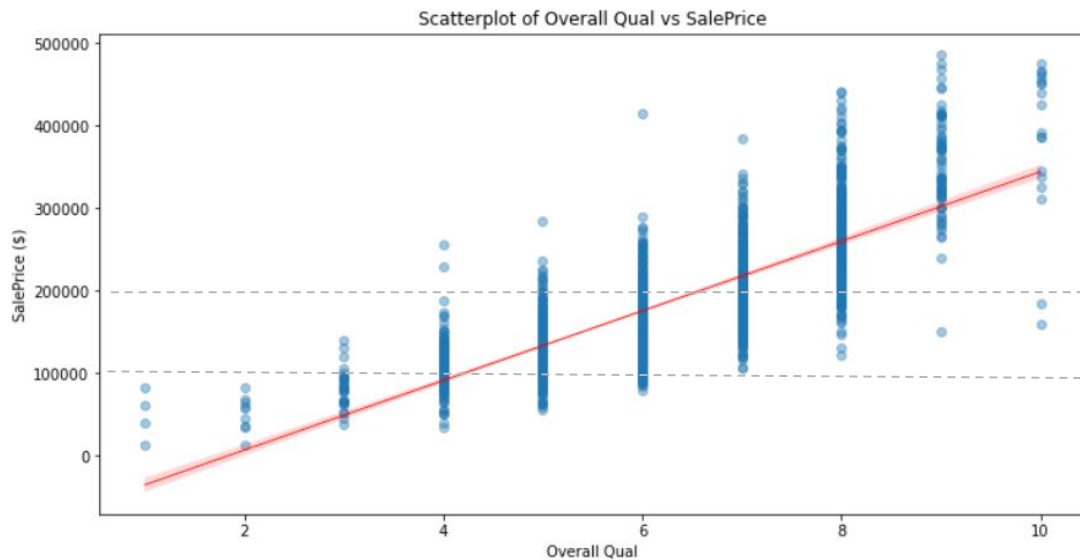
- Upward linear regression
- Sale price tend to increase when the ground living area increase
- clustered between about 1,000 to 2,000 square feet and at the price range of \$100,000 to \$200,000

# EDA & Visualisation



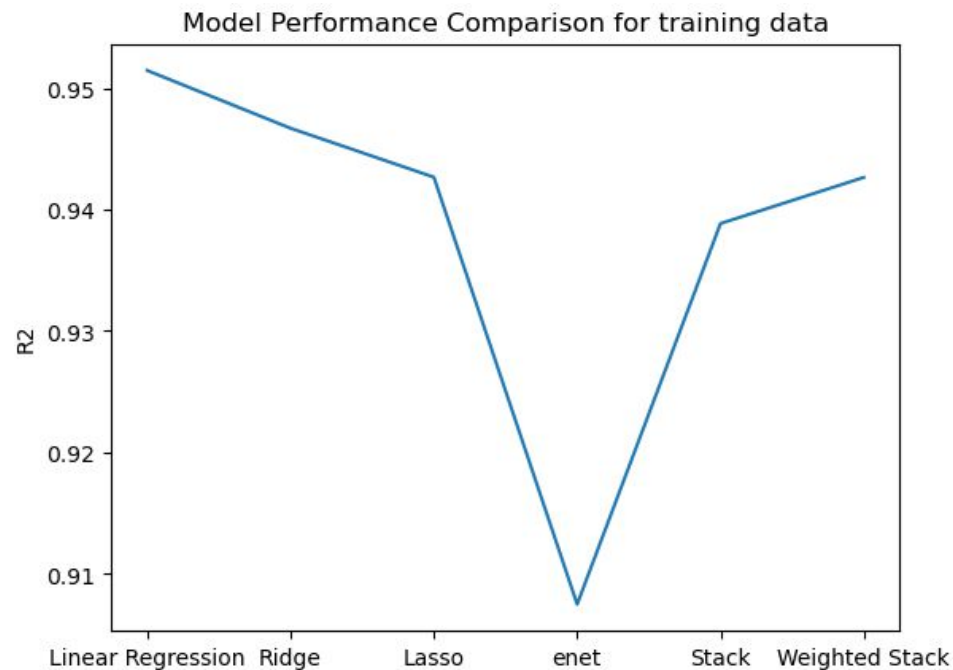
- Upward linear relationship
- More people tend to buy houses which were built after year 2000 and at the price range of \$100,000 to \$200,000

# EDA & Visualisation



- Upward linear relationship
- tend to buy houses at the rating of 5-7 and at the price range of \$100,000 to \$200,000.

# Models

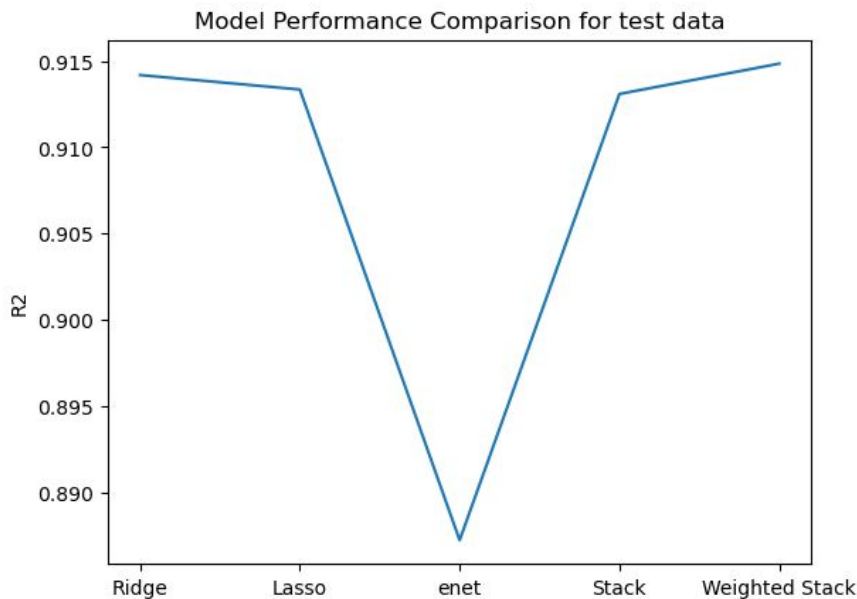


- Linear Regression, Ridge and Lasso were used for modelling
- Simple stacking of model to check if performance improves
- More tuning for ElasticNet

Simple Stack	$\text{mean}(\text{Ridge}, \text{Lasso}, \text{ElasticNet})$
Weighted Stack	$\text{Ridge} * 0.5 + \text{Lasso} * 0.3 + \text{ElasticNet} * 0.2$

# Models

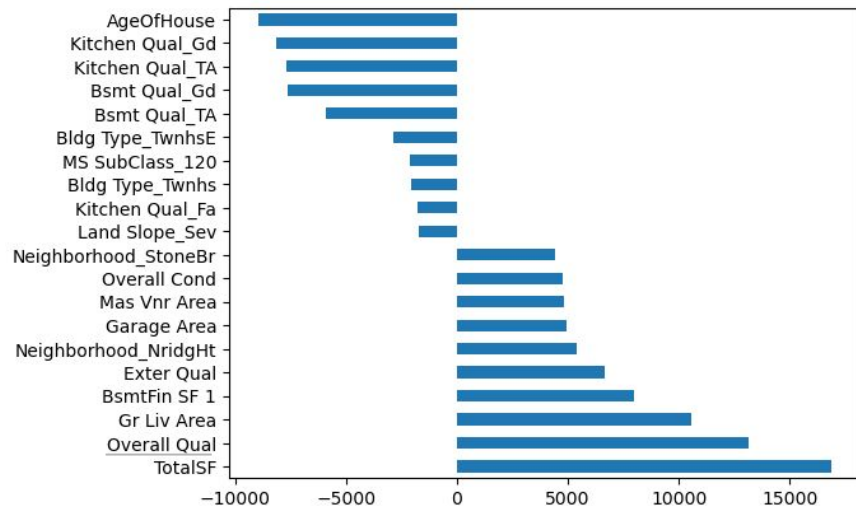
R2 Validation Score	
Linear Regression	$-6.31 \times 10^{23}$
Ridge	0.914
Lasso	0.913
Enet	0.887
Stack	0.913
Weighted Stack	0.915



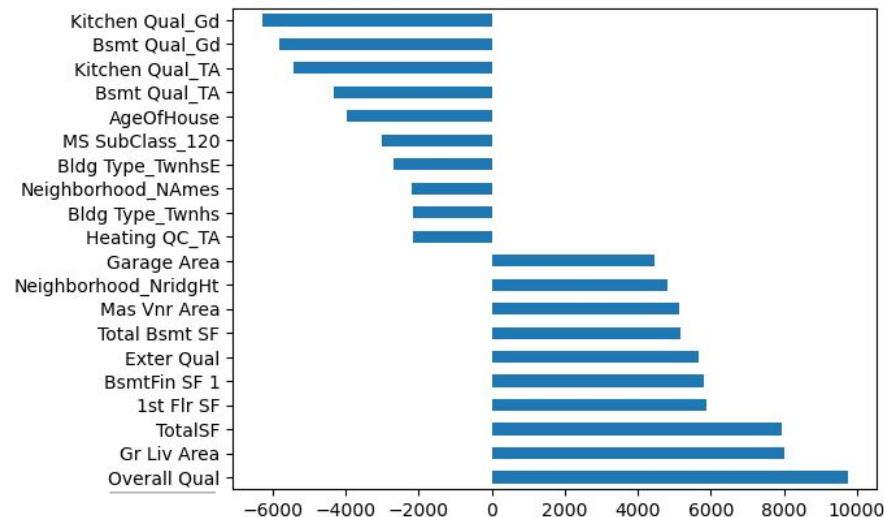
- Severe overfitting for the linear regression model seen in training vs test scores
- Regularised models do better out the training data set
- Slight improvement when the models are stacked

# Findings

## Lasso Coefficients



## Ridge Coefficients



- Overall Quality of the house appears on both of Lasso and Ridge coefficients and has been given a very large coefficient. North Ridge and Stone Brook, Ground Living Area
- Age of house has a large negative coefficient
- **52.27%** of features got zeroed out in Lasso; Fireplace & Total Rooms

# Conclusion

Problem Statements	Suggestions								
Identify the best and worst permanent and non-permanent features affecting the sale price in Ames, Iowa	<table><tr><th>Positive Coefficient</th><th>Negative Coefficient</th></tr><tr><td>Overall Quality</td><td>Age of house</td></tr><tr><td>Living Area</td><td>Average/Bad Kitchen</td></tr><tr><td>Neighborhood</td><td>Average/Bad Basement</td></tr></table>	Positive Coefficient	Negative Coefficient	Overall Quality	Age of house	Living Area	Average/Bad Kitchen	Neighborhood	Average/Bad Basement
Positive Coefficient	Negative Coefficient								
Overall Quality	Age of house								
Living Area	Average/Bad Kitchen								
Neighborhood	Average/Bad Basement								
Ensuring new home buyers could yield highest return of investment from their new house purchase	<ul style="list-style-type: none"><li>• Proposal to get house with lower overall quality, and remake &amp; remodel to improve overall quality</li><li>• Northridge and Stone Brook; Northern part of Ames</li><li>• Big Ground Area</li></ul>								



# Recommendations



Model could include economic data such as interest rate and employment



Retrain models in the future in case of data drift



Expand datasets to other states of US

README.md - must always be named like this

Start with EDA notebook and save that file as csv file

Start with another notebook and use that csv file

Image and data to be saved in separate folders for neater

Pycarat can be used but need to use linear regression

To extract betas from the sklearn models.

Code is always inappropriate for a non-technical audience

Highlight model name and package name

Shap package

Omitted variable that is additional variables