Predicting Sale Prices of Houses in Ames, Iowa

By Stephanie, Edward and Dave

Problem statement

This project aims to inform potential home sellers and property investors in Ames of the features that will increase the value of their homes. Given a set of characteristics of their property, they will be able to get an estimation of their property value from the prediction model.

This model will inform homeowners and investors of:

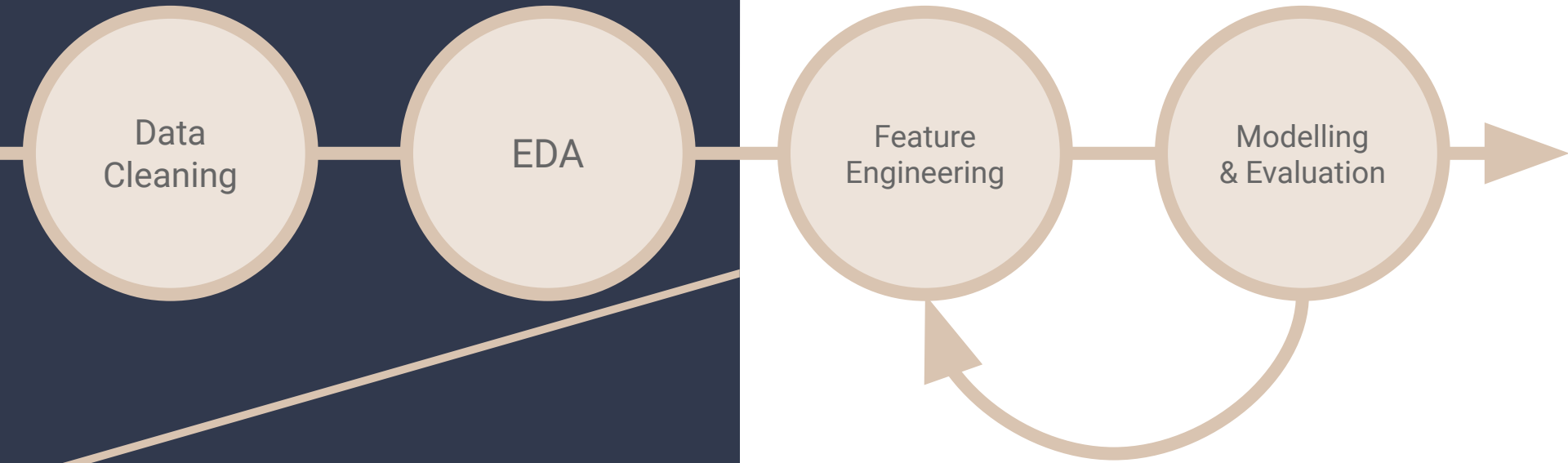
1. Top ten features that will add the most value to a property
2. Top ten features that will decrease the value of a property
3. Neighborhoods to invest in
4. How to increase the value of your property

Background about Ames, Iowa & the Ames Housing Dataset

Fast facts!

- Ames was ranked ninth on CNNMoney's "Best Places to Live" in 2010.
- In 2019, 33,391 out of the 66,258 people living in Ames were made up of students.
- Ames has a humid continential climate, with hot summers up to 39degC and cold winters down to -33degC.

Workflow


Data Cleaning
EDA
Feature Engineering
Modelling & Evaluation

Data cleaning
Outliers

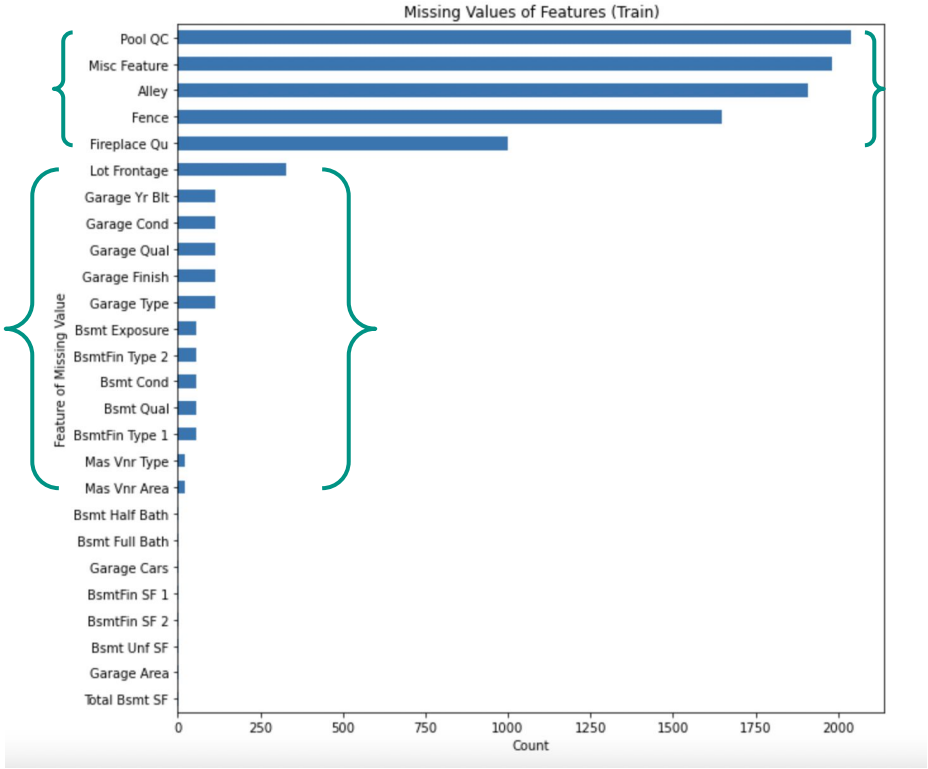
Two outlier points have been identified in the training set to be removed.

Data cleaning
Missing Values

Columns with more than 80% missing data values were dropped.

Since ‘Id’ and ‘PID’ are identification features without correlation to Sale Price, they will also be dropped.

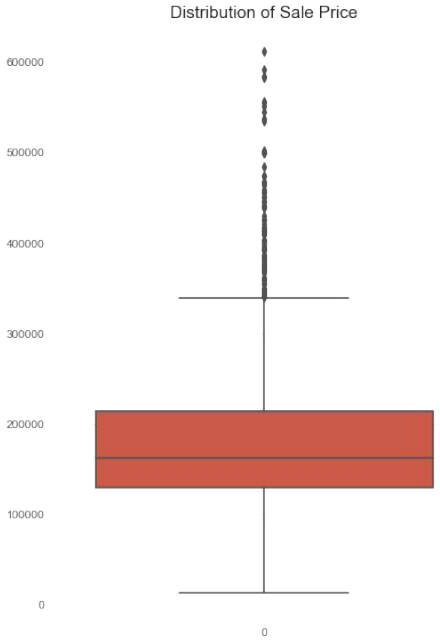
Other missing data values are imputed with values by logical inference e.g. missing values for ‘Lot Frontage’ is imputed with mean values as it is unlikely that a property does not have roads leading to it.



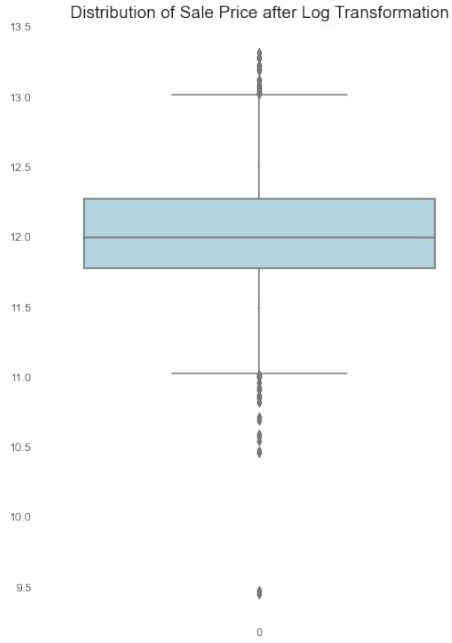
EDA
Sale Price

Our target variable ‘Sale Price’ has a highly skewed distribution. It resembles closer to a normal distribution after log transformation, more ideal for linear regression modelling.

Before



After

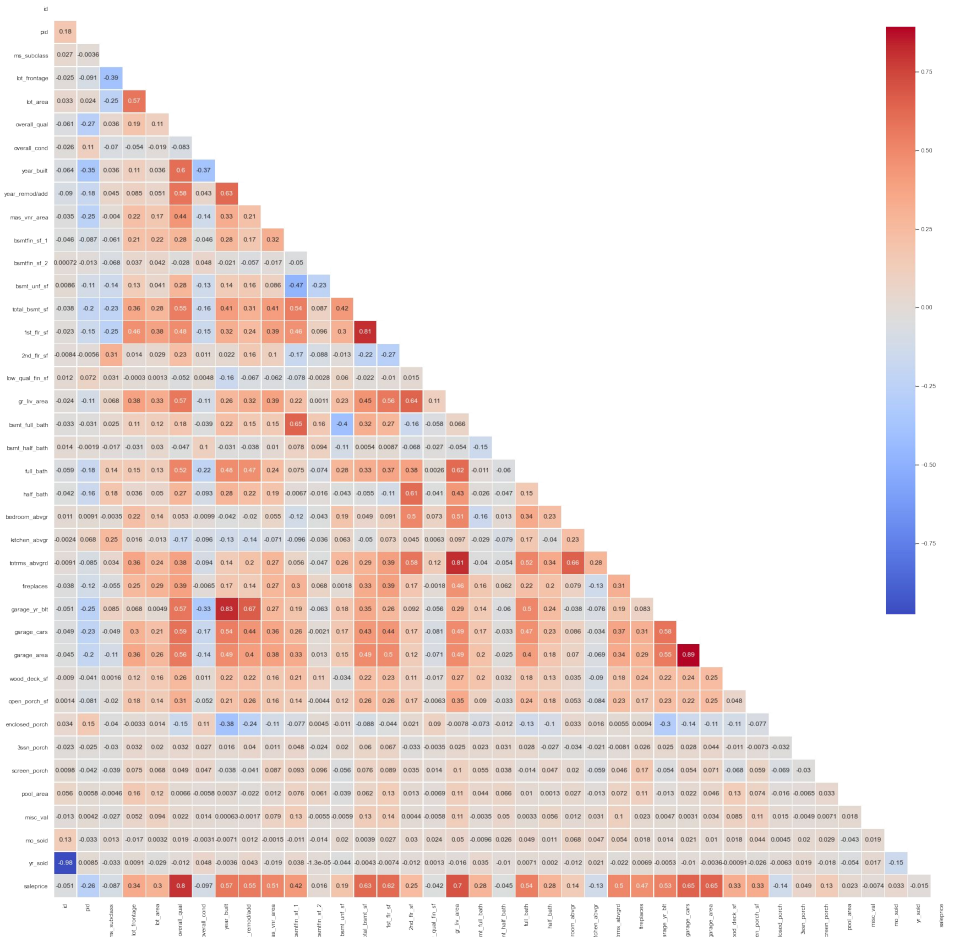


EDA
Multicollinearity

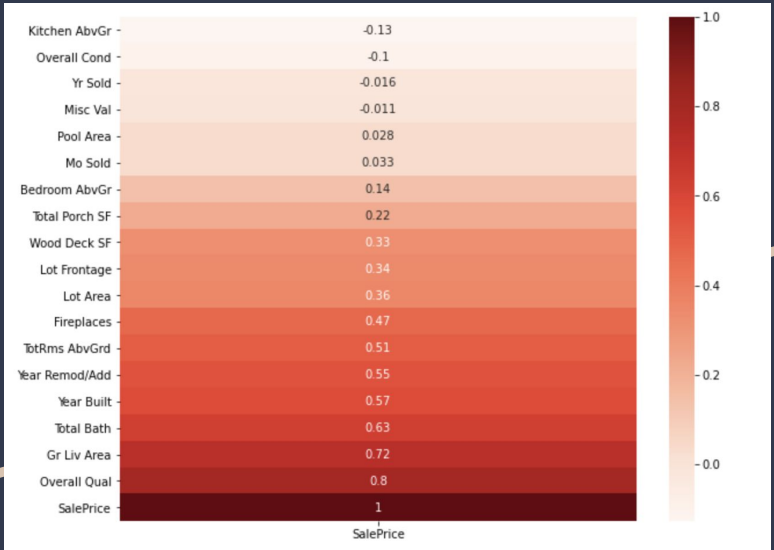
Some predictor variables were observed to have high collinearity values and were dropped from the training and test sets.

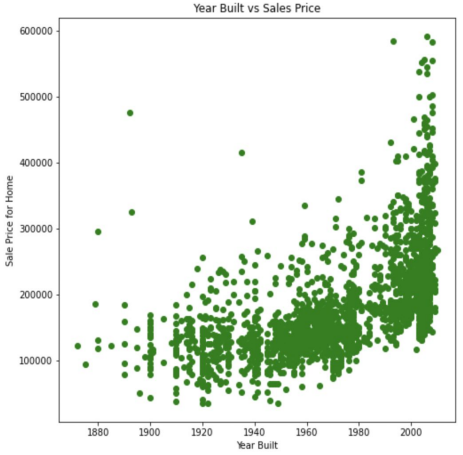
For example:
Area of basement finishes type 1 +
Area of basement finishes type 2 +
Area of basement unfinished =
Total area of basement

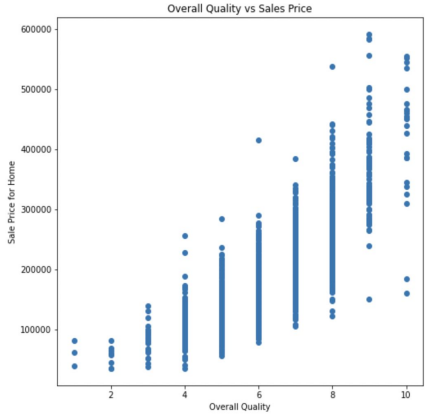
In such instances, the predictor variables with least correlation to the Sale Price are dropped.



EDA
Correlation to Sale Price







Feature Engineering

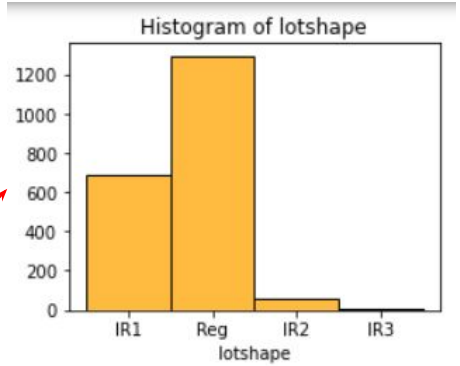
Hypothesis Testing

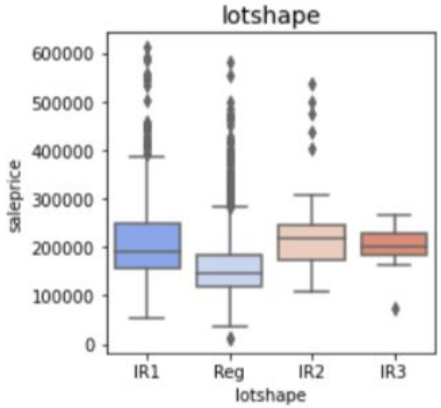
$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

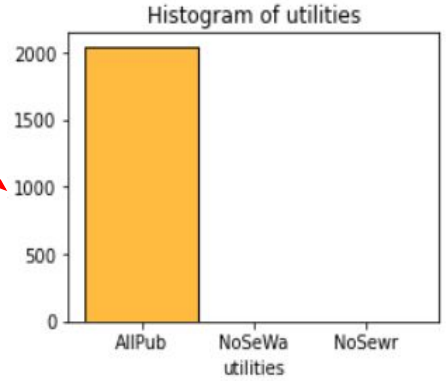
P-value = 2e-45 < 0.05

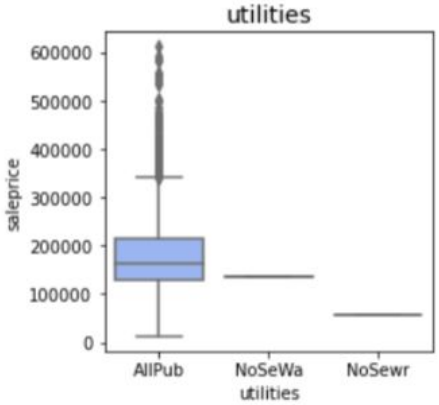
P-value = 0.135 > 0.05

Mapping Ordinal/Nominal Variables









Feature Engineering

Create Columns

One Hot Encoding Categorical Variables

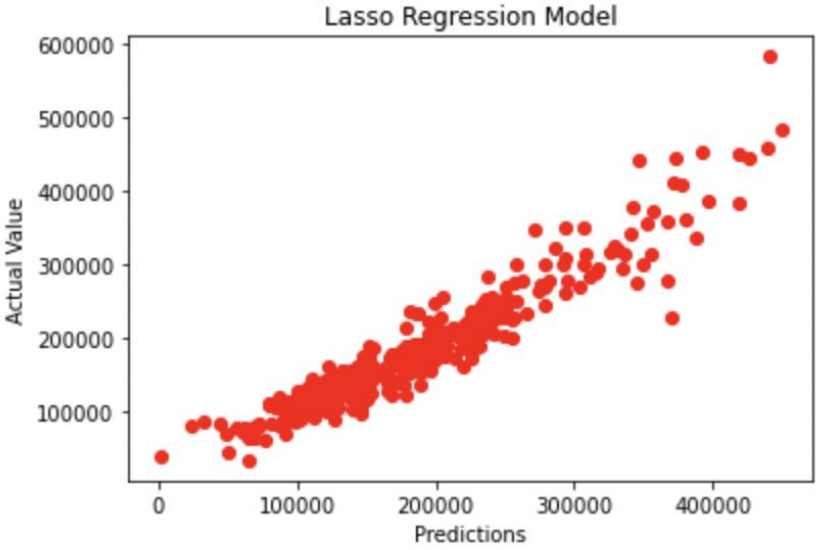
Grouping

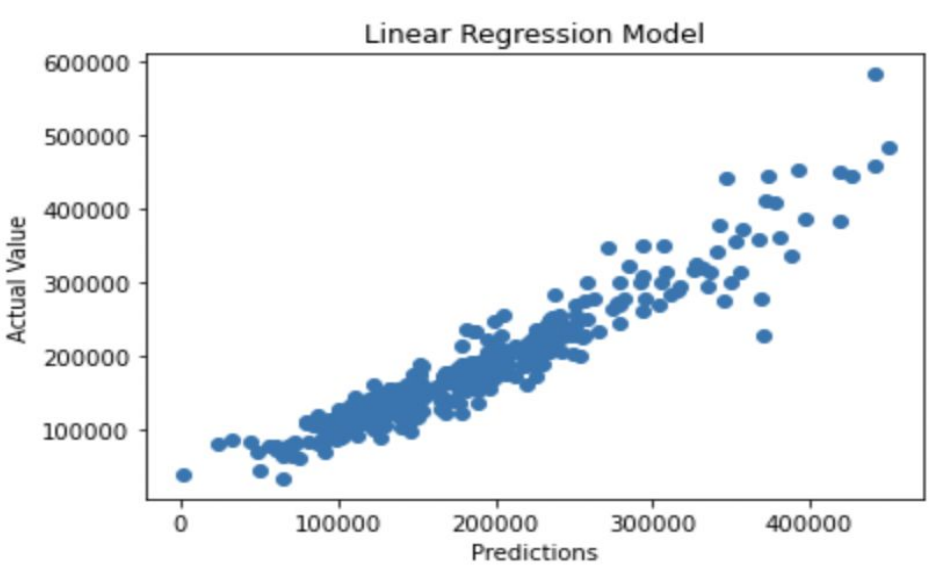
- Age = Year Sold - Year Built
- Two_floors?
- Bathroom = Fullbath + 0.5 Halfbath
- Porch?

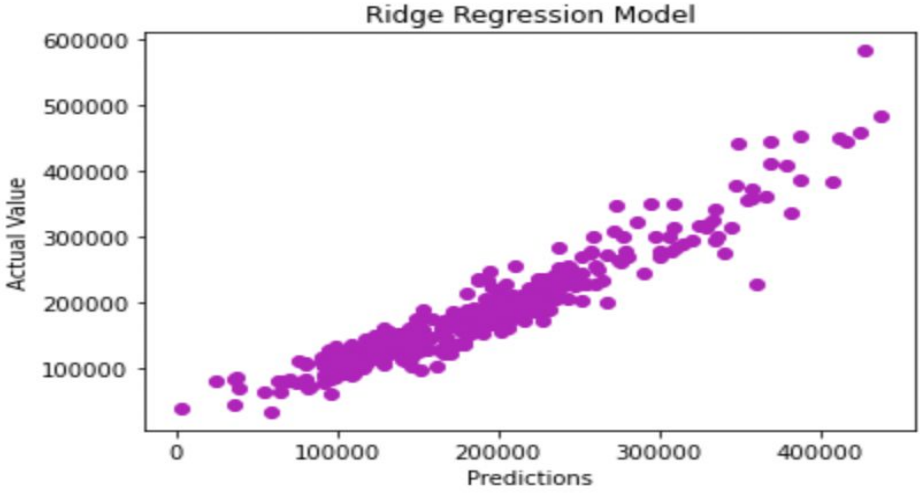
- For categorical data with no ordinal relationship

```
exterior_type = {'VinylSd': 'Vinyl',
                 'MetalSd': 'Metal',
                 'HdBoard': 'Wood',
                 'Wd Sdng': 'Wood',
                 'Plywood': 'Wood',
                 'CemntBd': 'Cement',
                 'BrkFace': 'Brick',
                 'WdShing': 'Wood',
                 'AsbShng': 'Others',
                 'Stucco': 'Others',
                 'BrkComm': 'Brick',
                 'Stone': 'Others',
                 'CBlock': 'Others',
                 'AsphShn': 'Others',
                 'ImStucc': 'Others'}
```

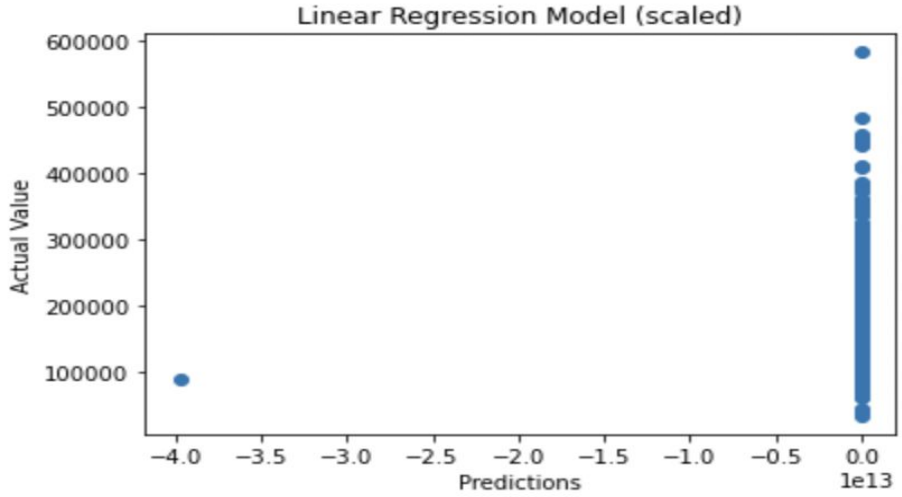
Modelling & Evaluation (1)
Unscaled Data

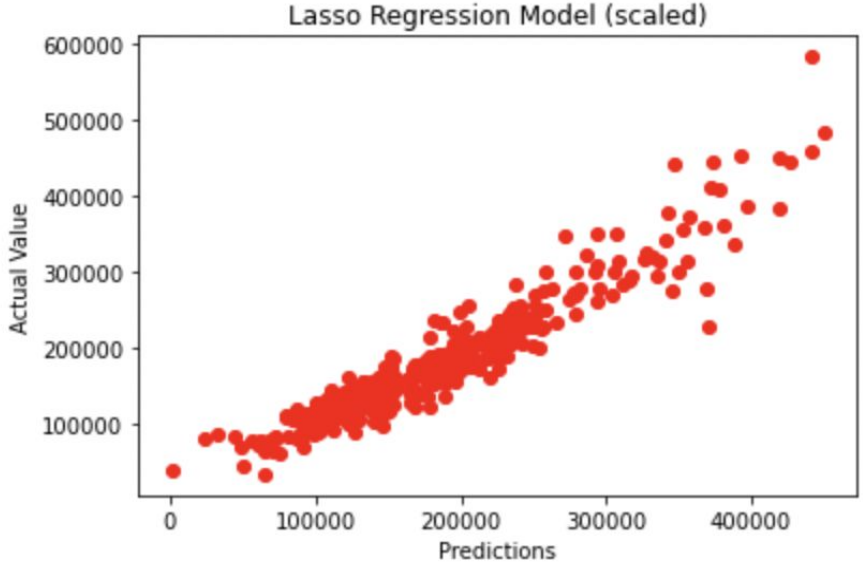


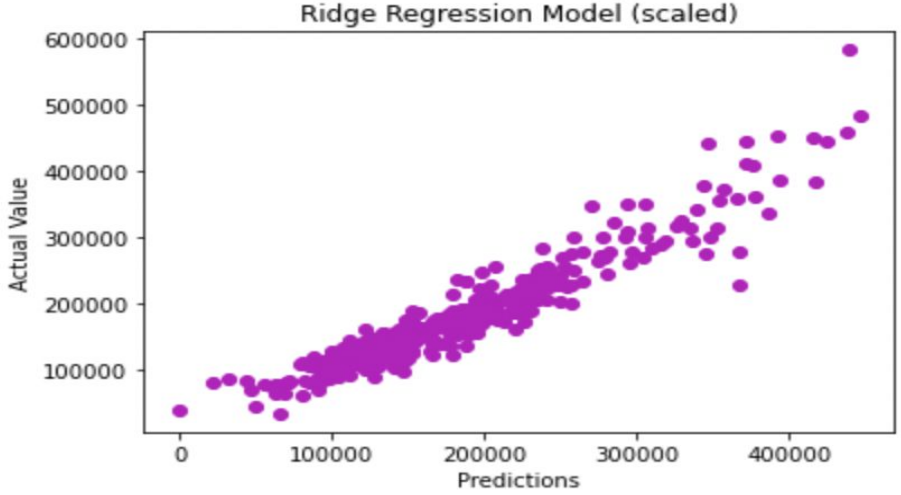




Modelling & Evaluation (2) Scaled Data

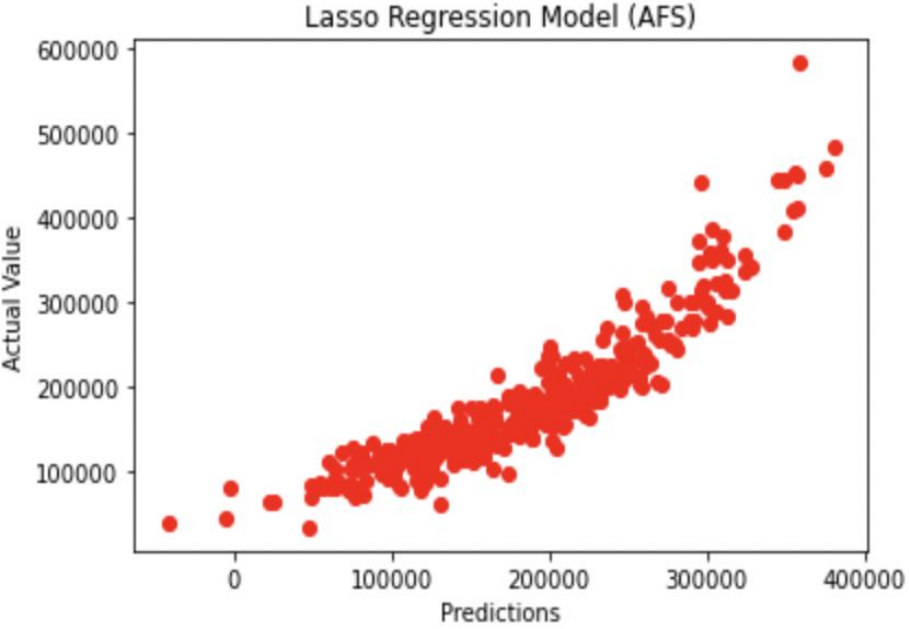


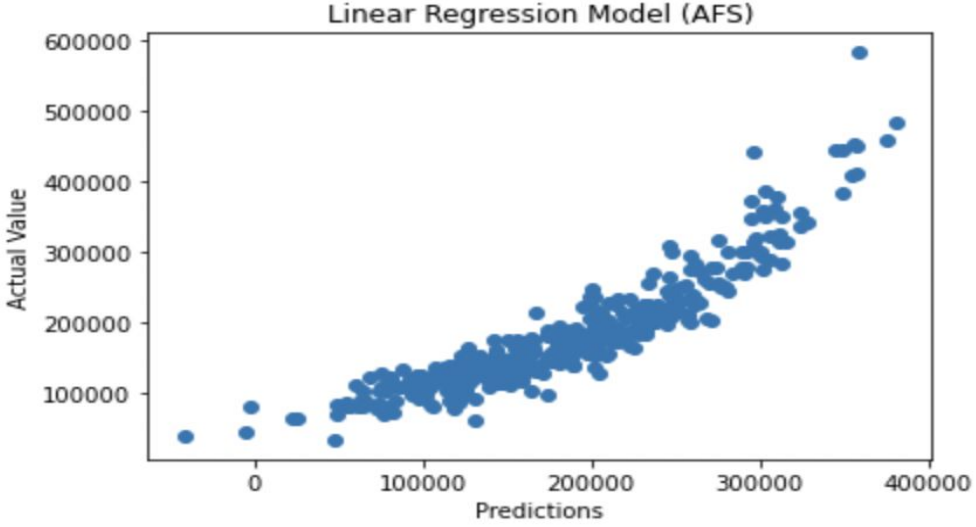


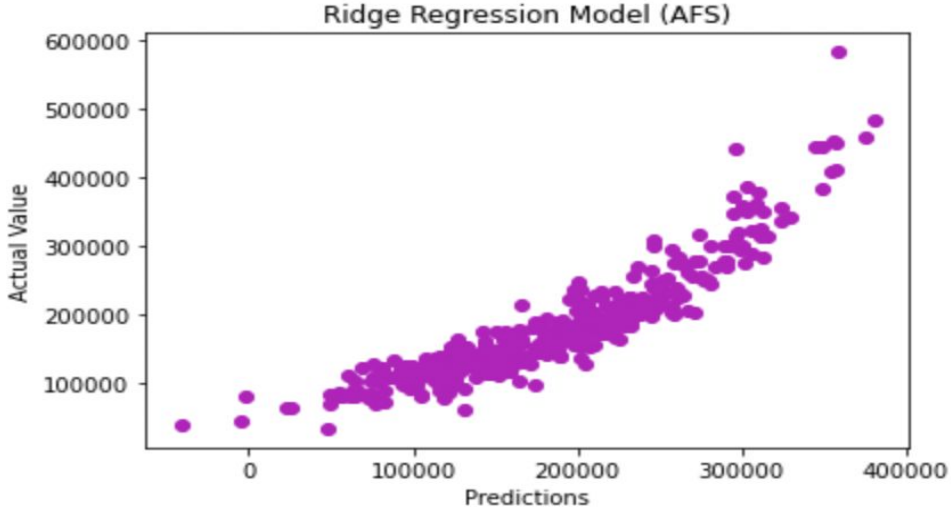


Modelling & Evaluation

Automated Feature Selection







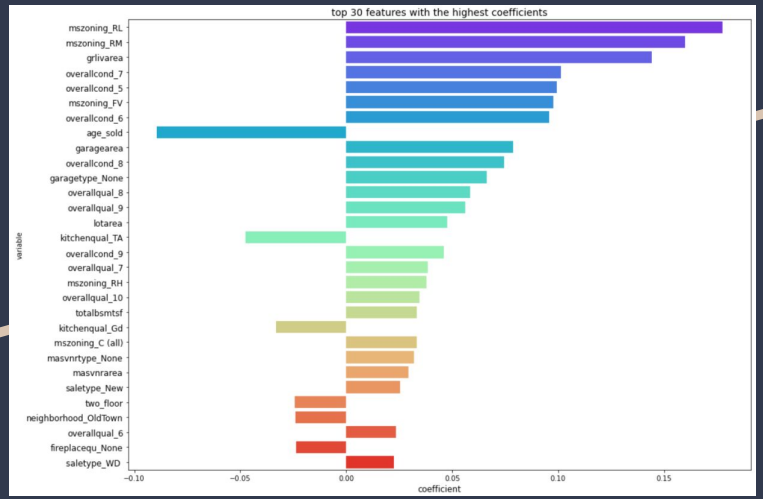
Comparison

Linear Regression
Lasso
Ridge

Model	RMSE	R2
baseline	0.42802367386108625	-0.0062235095763414 35
linear regression	24427117914.989464	-3.2772029489334305e+21
ridge	0.14525212317065 24	0.884121235050213
lasso	0.14488633816746765	0.8847041297129481
elastic net	0.14454781461850938	0.8852422725207754

Summary

Top features that increases and decreases the value of the property



Features that increases Property Sale Price

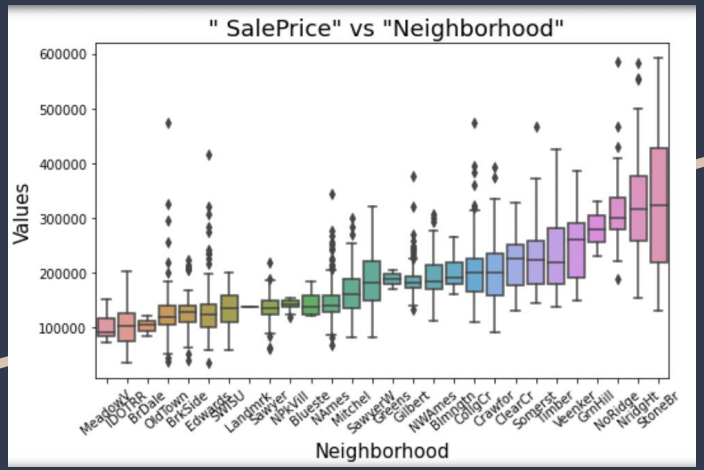
1. MS Zoning (Low-medium density)
2. Above Ground Living Area
3. Overall Condition (6-8)
4. Garage Area
5. Lot Area

Features that decreases Property Sale Price

1. Age of the property
2. Kitchen Quality (TA)
3. Neighbourhood (OldTown)
4. Fireplace (None)
5. 2-Story Property

Summary

Neighborhoods worthy of Investing



Highly Valued Neighborhoods

1. Stone Brooke
2. North Ridge Heights
3. Northridge
4. Green Hills
5. Veenker

Low Valued Neighborhoods

1. Meadow Village
2. Iowa DOT and Rail Road
3. Briardale
4. Old Town
5. Brookside

Conclusion & Further Recommendations

1. The 3 factors that increases the price of a home are quality, size and location.
2. Missing data values will always have an impact to our prediction model no matter how logically we try to drop or impute values.
3. Reducing number of features to a reasonable amount for data collection and modelling.
4. Categorical variables can be clustered and mapped to reduce the number of features used for modelling.
5. Fine tuning hyperparameters of the model to achieve better fitting.