DSI Immersive 12

Ames Iowa Housing Prediction - Kaggle Challenge

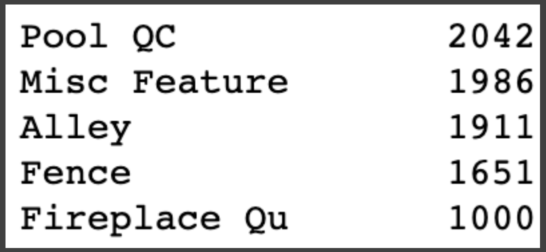
GABRIELLE JIAXIAN TAN

Problem Statement

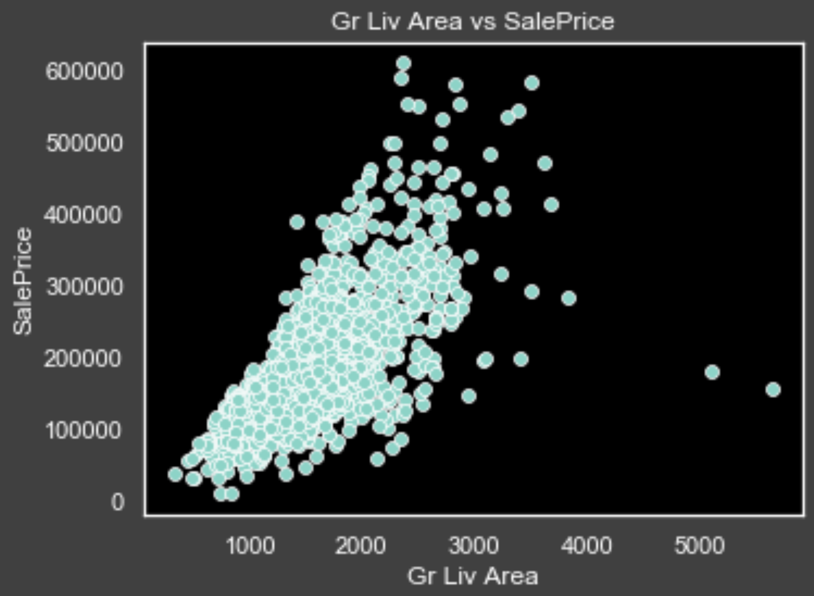
For each house ID in the Aimes Iowa Housing dataset, make predictions on the SalePrice using the test data based on certain characteristics with the lowest possible error in Neighbourhoods.

Process

- EDA and Data Cleaning
  - Set ID as index
  - Impute missing null values
  - Categorical Values vs Continuous Values
  - Drop columns ie PID, ID
  - Manage outliers



```
Pool QC            2042
Misc Feature       1986
Alley              1911
Fence              1651
Fireplace Qu       1000
```

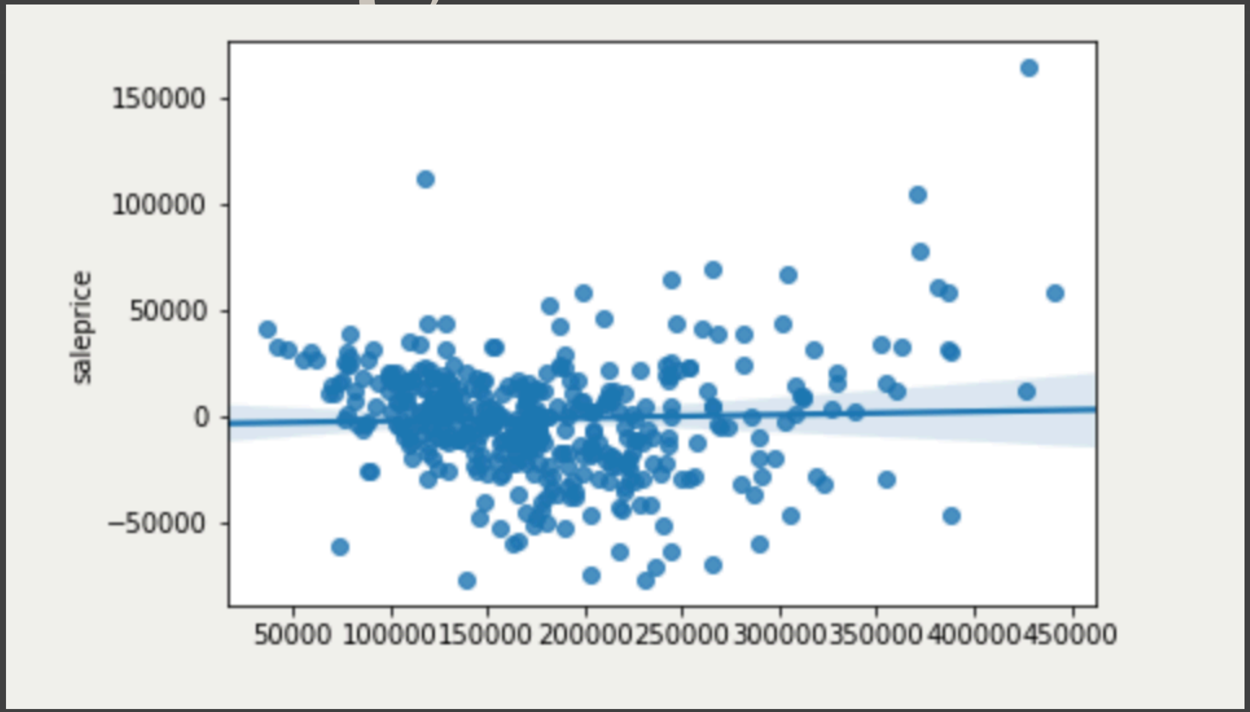


Diving into the Features

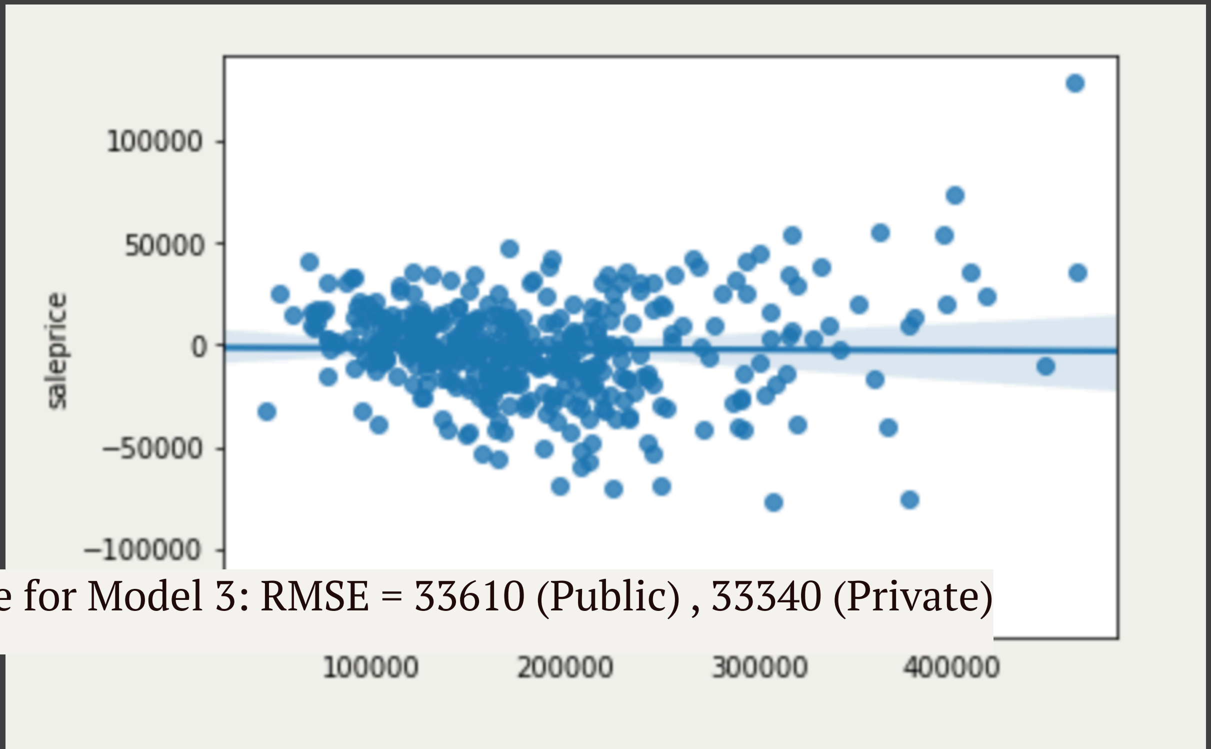
- Feature Processing via plotted distributions (scatter (continuous) and box plots (categorical))
    - Statistical Analysis
        - .mean()
    - Correlation
    - Removing collinear variables (aka highly correlated variables) ie bed baths to sq ft
    - One-hot encoded categorical variables : opt out
    - Drop features that will not be explored
    - Train-test-split
        - target variable - y - SalePrice
    - StandardScaler
- Feature Engineering
- Feature Selection

Modeling

Kaggle Score for Model 1: RMSE = 31730 (Public) , 33020 (Private)



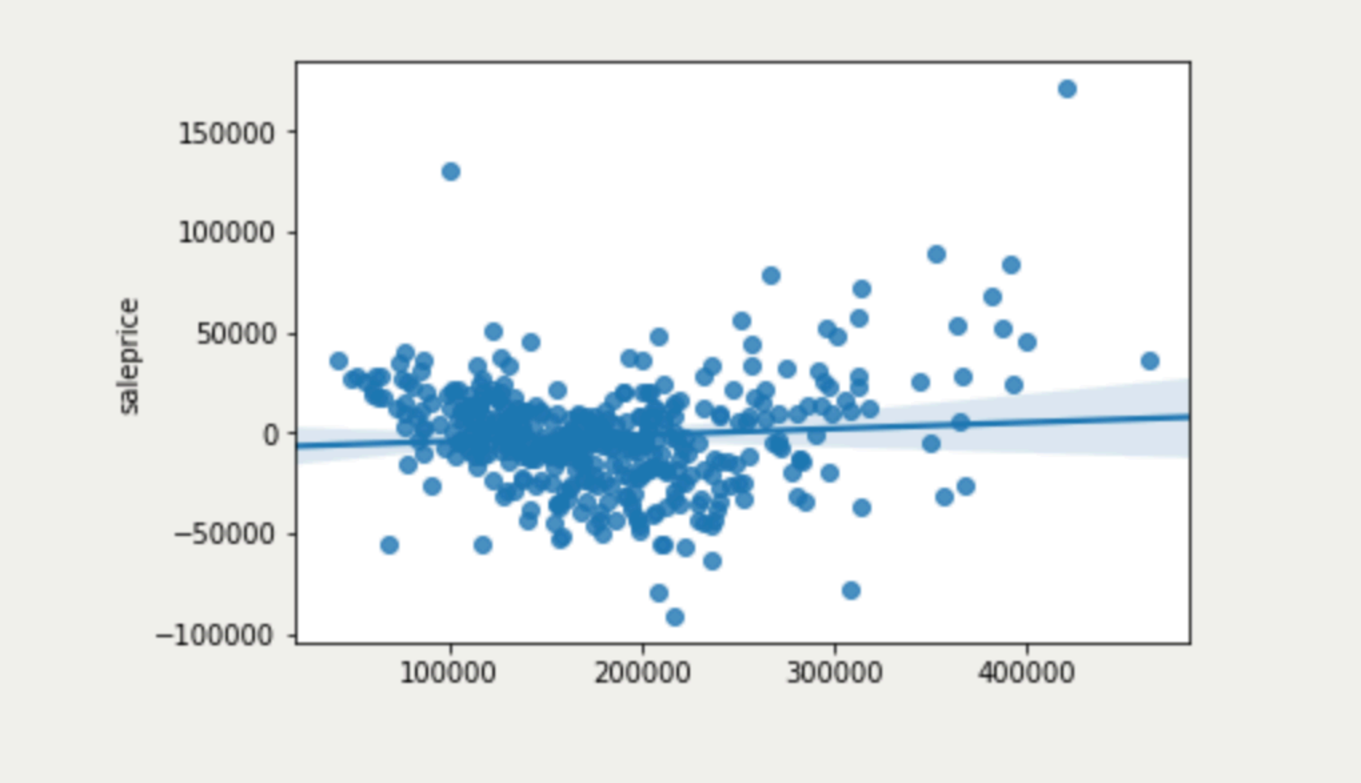
Kaggle Score for Model 2: RMSE = 28670 (Public) , 31710 (Private)

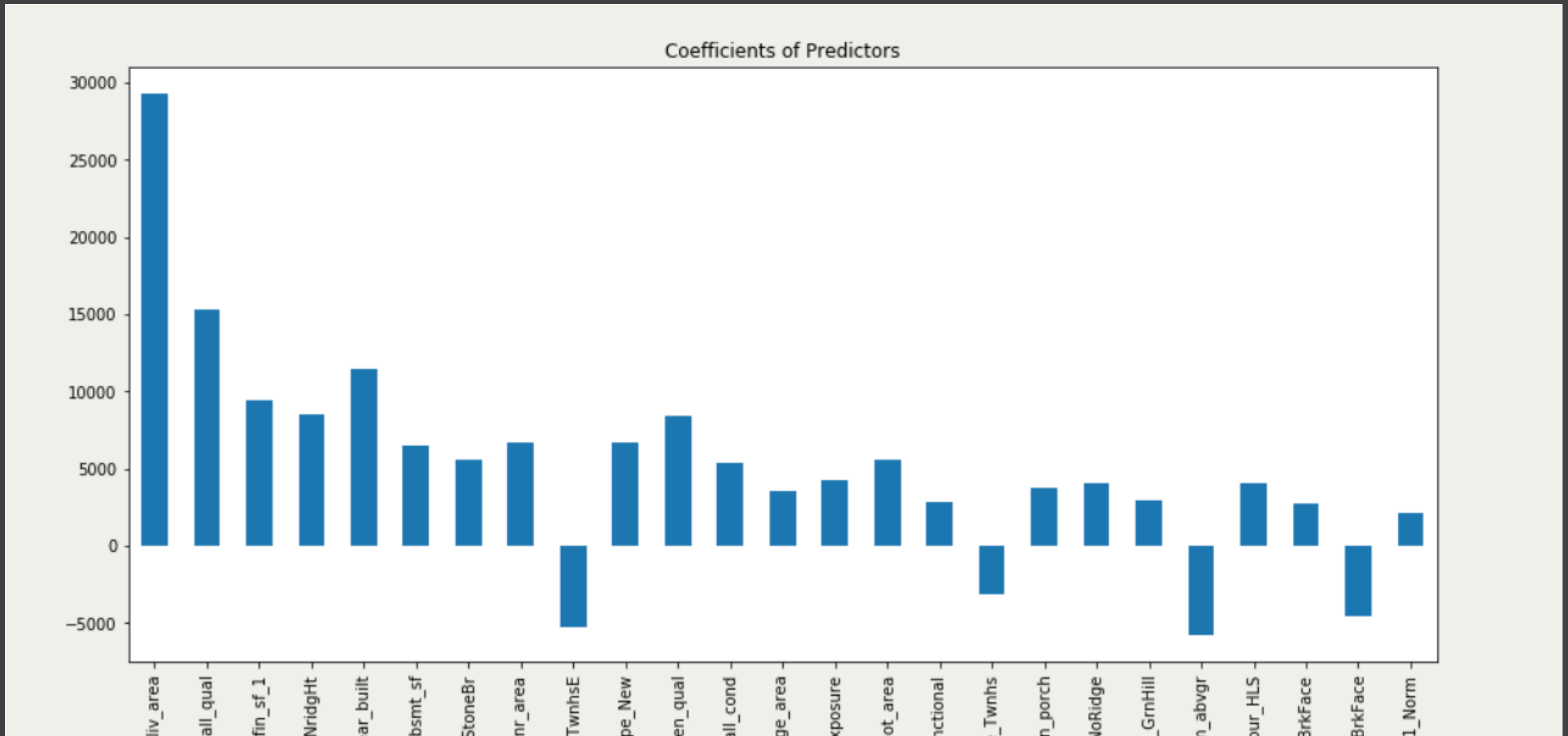


Kaggle Score for Model 3: RMSE = 33610 (Public) , 33340 (Private)

Modeling

Kaggle Score for Model 3: RMSE = 33610 (Public) , 33340 (Private)





Production Model

- Ridge Model
  - best performing Kaggle Model
  - Answers problem statement
- Interpretation
  - accuracy
  - any pattern to errors

Recommendations

- key features
  - to include to get the most value to a home: Overall Quality, Year Built
  - not to include: Bedrooms
- suggestions to improve the valuation of their homes
  - 1. explore median/mode of null values of that particular column
- neighbourhoods with maximum investment potential
  - Northridge Heights, Stone Brook, Green Hills

This Ridge model, although it can be generalise to other cities, some revisions to the dataset ie date, would make it a more comparable analysis though.