Analysis of Ames House Prices

Problem Statement

To identify indicators/ features from the Ames Housing dataset which predict the sale price of housing in Ames Iowa. This information will help buyers and investors to find the best houses based on their preferences and needs. It will also help sellers to identify ways to increase their house's selling price. Linear regression models with and without regularization such as Lasso, Ridge and ElasticNet will be used to model the data and R squared and Root Mean Square Error will be used to evaluate and find the best model for production and prediction.

The predicted sale price based on the data from Kaggle will be submitted to the DSI Kaggle competition to evaluate the effectiveness of the model on unseen data.

Grouped Features

1. Age related features = ‘Year Built’'Mo Sold','Yr Sold','Year Remod/Add','Garage Yr Blt'

2. Condtions related features = Overall Qual', 'Overall Cond','Exter Qual', 'Exter Cond','Functional'

3. Location related features =' MS Zoning','Neighborhood','Condition 1','Condition 2'

4. Size related features = 'MS SubClass','Lot Frontage','Lot Area','Bldg Type','Bsmt Qual', 'BsmtFin SF 1', 'BsmtFin SF 2', 'Bsmt Unf SF','Total Bsmt SF','1st Flr SF', '2nd Flr SF', 'Low Qual Fin SF','Gr Liv Area','Garage Area','Wood Deck SF', 'Open Porch SF', 'Enclosed Porch', '3Ssn Porch','Screen Porch', 'Pool Area'

Grouped Features

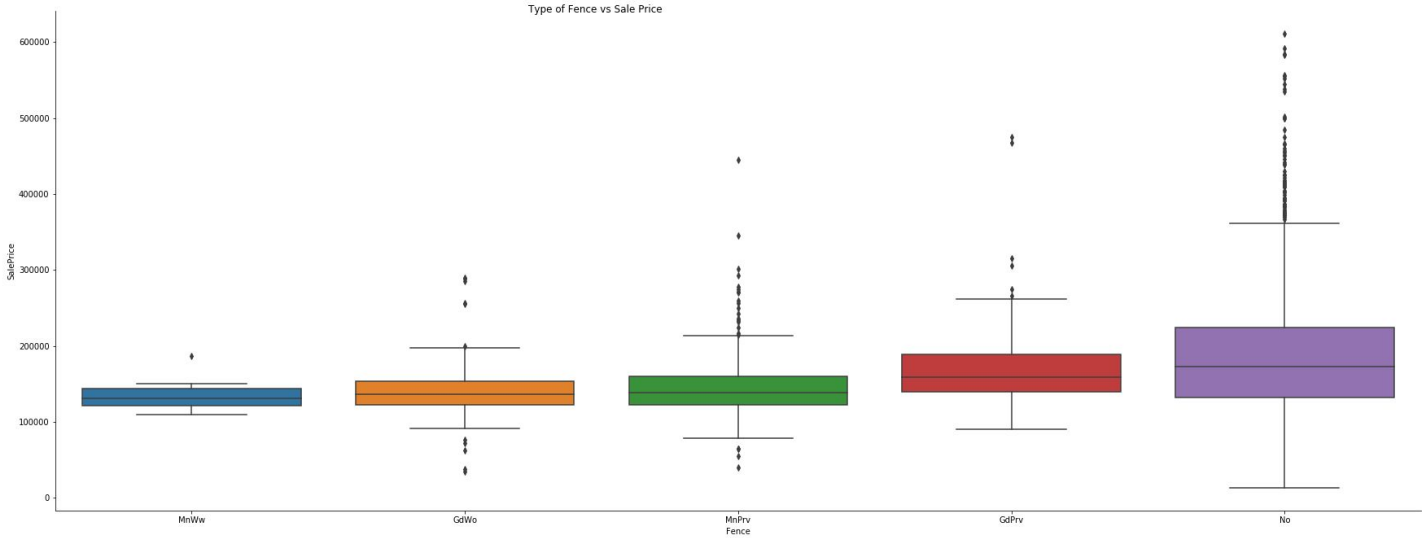
5. Interior related features = 'Utilities','Electrical','Bsmt Cond', 'Bsmt Exposure','BsmtFin Type 1','BsmtFin Type 2','Heating', 'Heating QC', 'Central Air','Bsmt Full Bath', 'Bsmt Half Bath', 'Full Bath','Half Bath', 'Bedroom AbvGr', 'Kitchen AbvGr', 'Fireplaces', 'Fireplace Qu','Garage Type', 'Garage Finish', 'Garage Cars', 'Garage Qual', 'Garage Cond'

6. Exterior related features= 'Lot Shape','Lot Config','House Style','Roof Style', 'Roof Matl','Exterior 1st', 'Exterior 2nd','Mas Vnr Type', 'Mas Vnr Area','Foundation','Paved Drive', 'Pool QC', 'Fence', 'Misc Feature','Misc Val'

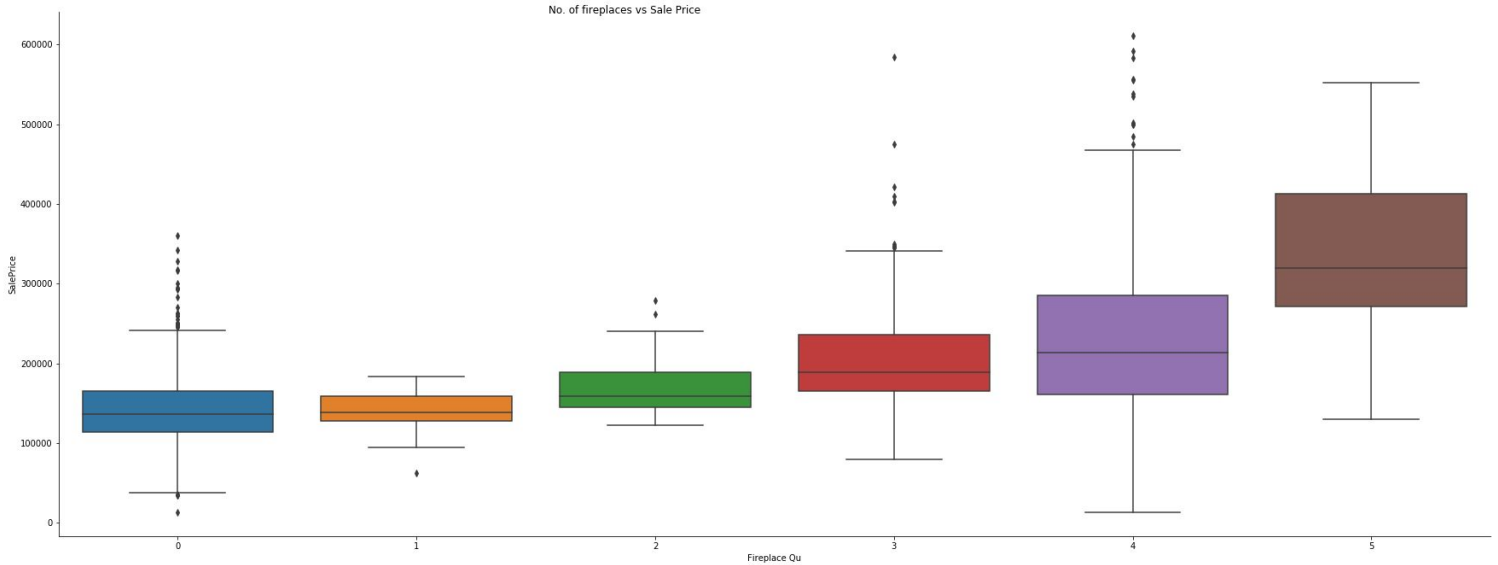
7. Area related features = 'Street','Alley','Land Contour','Land Slope'

8. Misc related features = 'Sale Type'

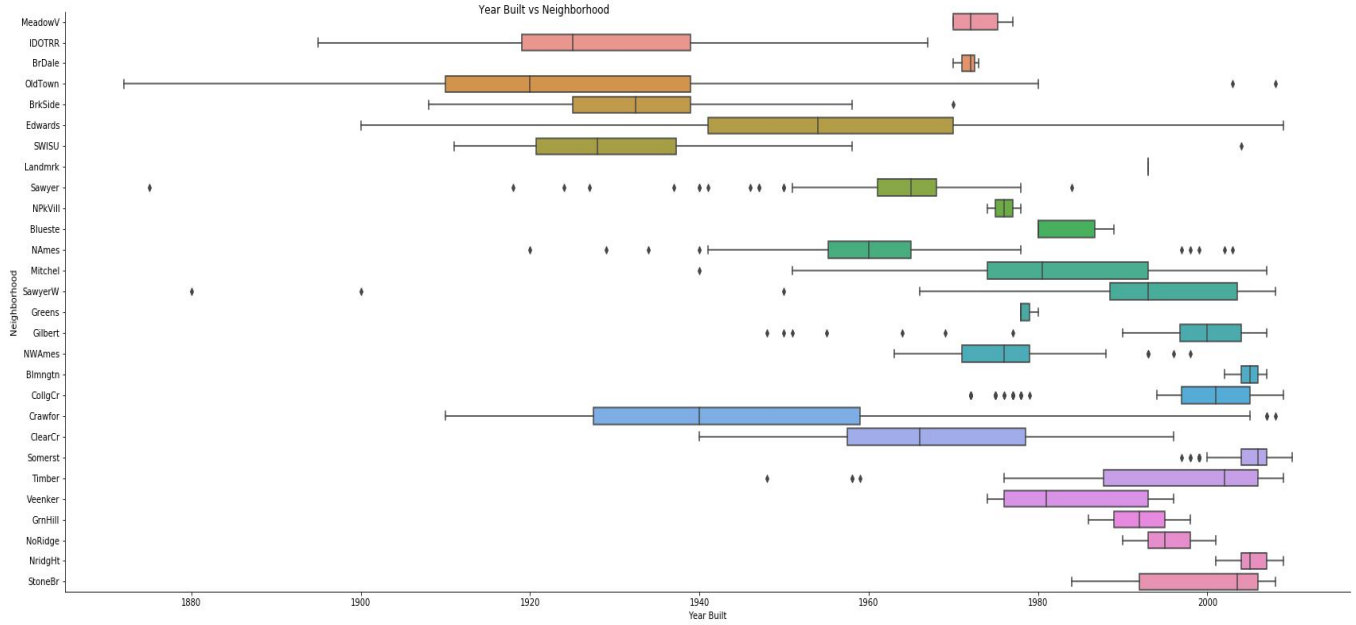
EDA and Cleaning



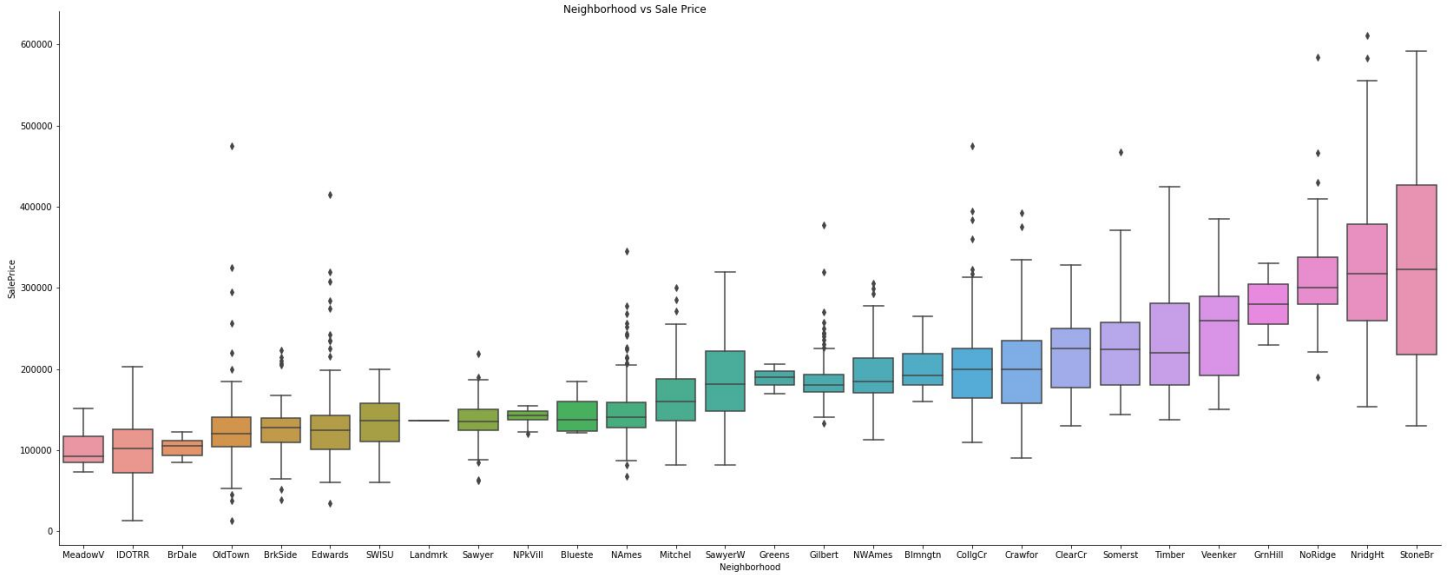
EDA and Cleaning



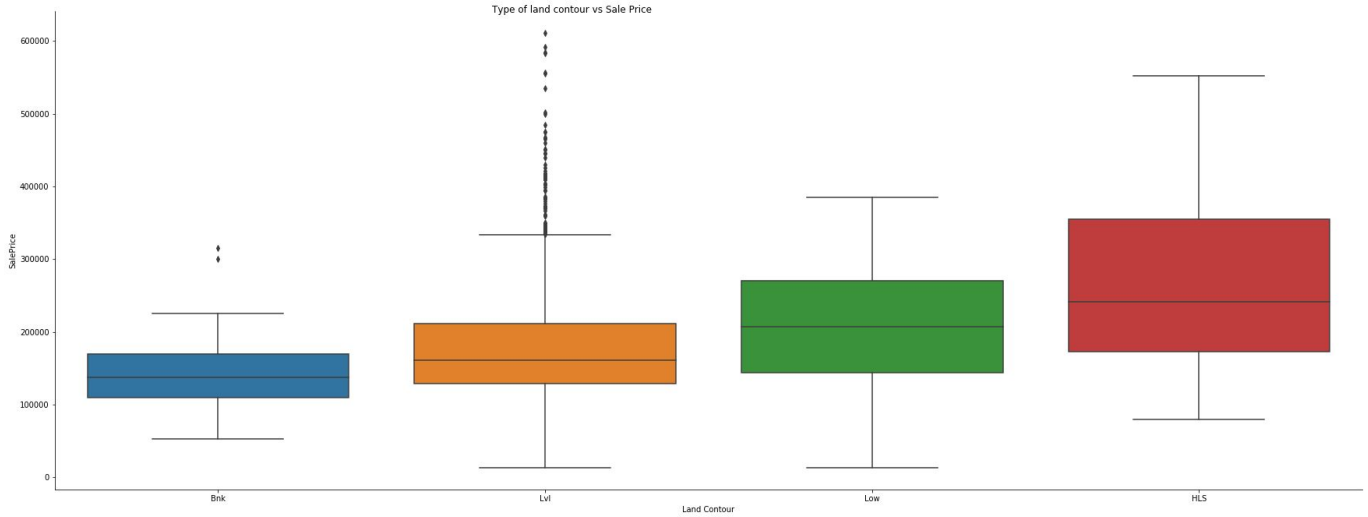
EDA and Cleaning



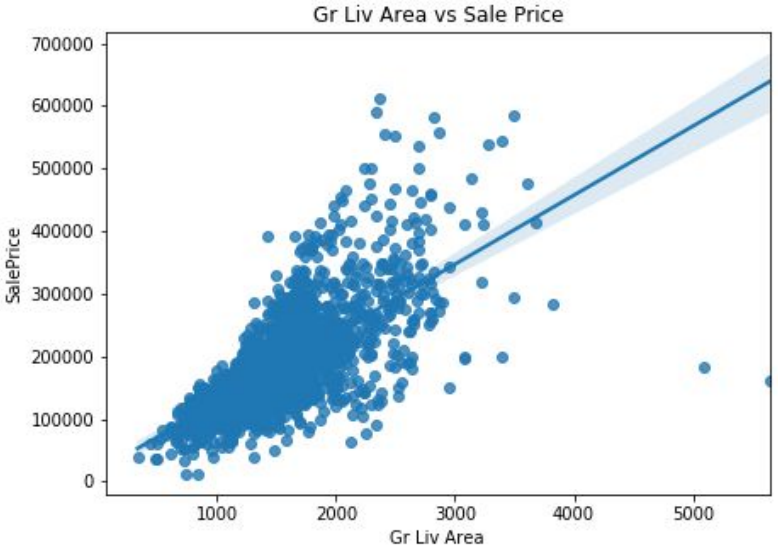
EDA and Cleaning

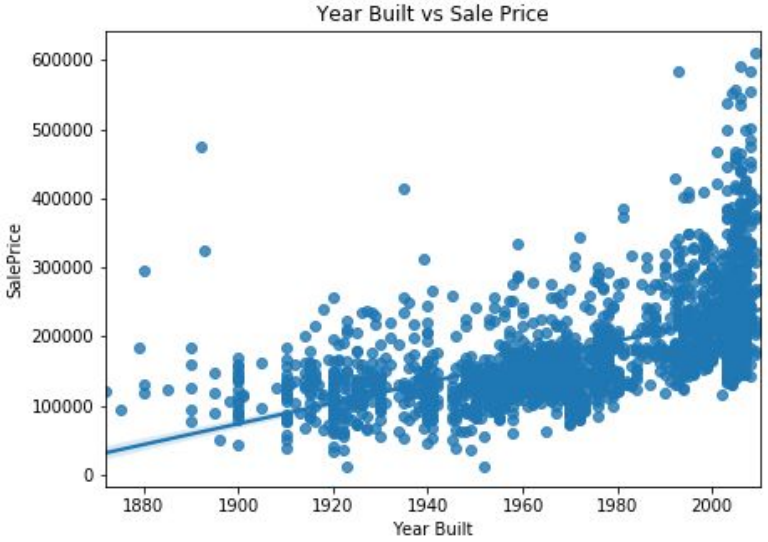


EDA and Cleaning



EDA and Cleaning





EDA and Cleaning

1. Boxplots and Scatterplots show a positive linear relation between the grouped features and sale price except for area features like land contours.

2. Boxplot of Neighborhood vs sale price vs year built shows that the newer neighborhoods have higher sale prices.

3. Age related features may have significant impact on sale price

4. Encoded null values with appropriate values

5. Encoded categorical values with appropriate values

Feature Selection

1. Dropped features that are less than 25% correlated with sale price
2. Used Recursive Feature elimination with LassoCV to eliminate features to 25 features

Preprocessing and Feature Engineering

1. Created features for “house age” and “ reno newness” as age of house and newness of renovation should have significant impact on sale price

2. Dropped any rows where ground floor area is more than 4000 sq ft as outliers are found above this threshold.

3. Dummy encoded categorical variables

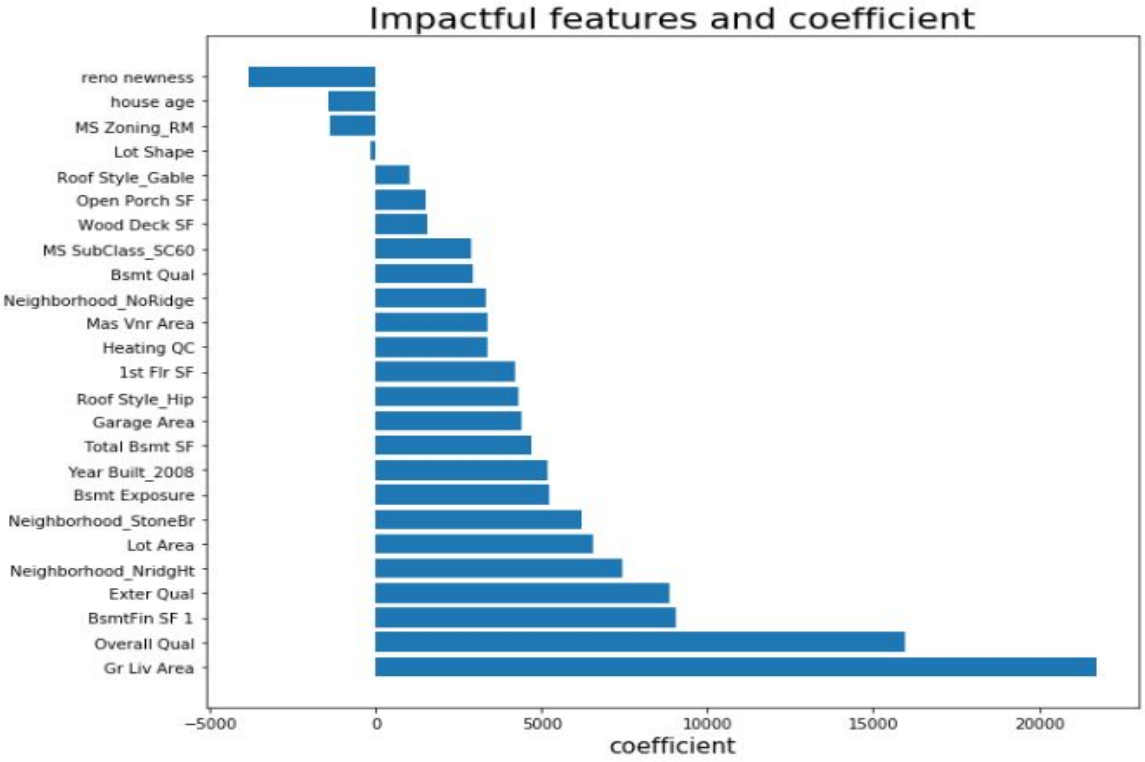
Model Benchmark

1. Linear Regression as baseline model

Model Tuning

1. Used Pipeline, Gridsearch CV to tune 3 models:
    a. LassoCV
    b. RidgeCV
    c. ElasticNet

Production Model and Insights



Production Model and Insights

1. Selected ElasticNet as production model as its R2 was the highest and RSME the lowest
2. Size and overall quality features have the largest impact on sale price
3. The newer neighborhoods such as Stone Brook identified a positive impact on sale price.
4. The newer the house and renovations the higher the impact on sale price
5. Having good quality exterior features such as Masonry veneer have a positive impact on sale price
6. Living in a medium density residential area has a negative impact on sale price, most likely more people would live in an area where houses are cheap
7. Having a a good quality fireplace has a significant postive impact on sale price, this could be due to Ames, Iowa, having average annual temperatures of 10 degrees Celsius.
8. Having a basement also has a significant postive impact on sale price which is also related to the size of the house
9. Houses built in 2008 have a positive impact on sale price, this could be due to housing prices reaching their peak prior to the financial crisis. Analysis during the EDA has also shown that sale prices after 2008 were lower.

Research

1) Ames, Iowa is a small university town.
2) Population less than 100,000 pax
3) Iowa State university is top employer
4) Average Annual Temperature is 10 degrees celsius
5) Research shows price increase in housing near universities

Conclusion and Recommendation

Analysis shows: distance from the university impacts sale price.

We recommend that for our model to be more accurate for other cities, information needed:

1) Distance from facilities such as universities
2) Weather
3) Economic activity
4) Demographics