Ames Housing Data and Kaggle Challenge

Project 2

Esther Khor

The problem or challenge

To create the best linear regression model that is both high performing and generalisable in order to best predict sale prices of houses in Ames.

Our aim:

For property owners:

- Features to focus on in order to most effectively increase the value of their house sale price.

For potential property buyers:

- More efficiently determine what kind of houses to look for based on their budget or to help them set a reasonable budget based on the features they are looking for in their potential future house.

Summary of work done:

The dataset used is the Ames housing dataset from Kaggle and a detailed data dictionary can be found here.

Notebook 1: Preliminary EDA of 'train' dataset .

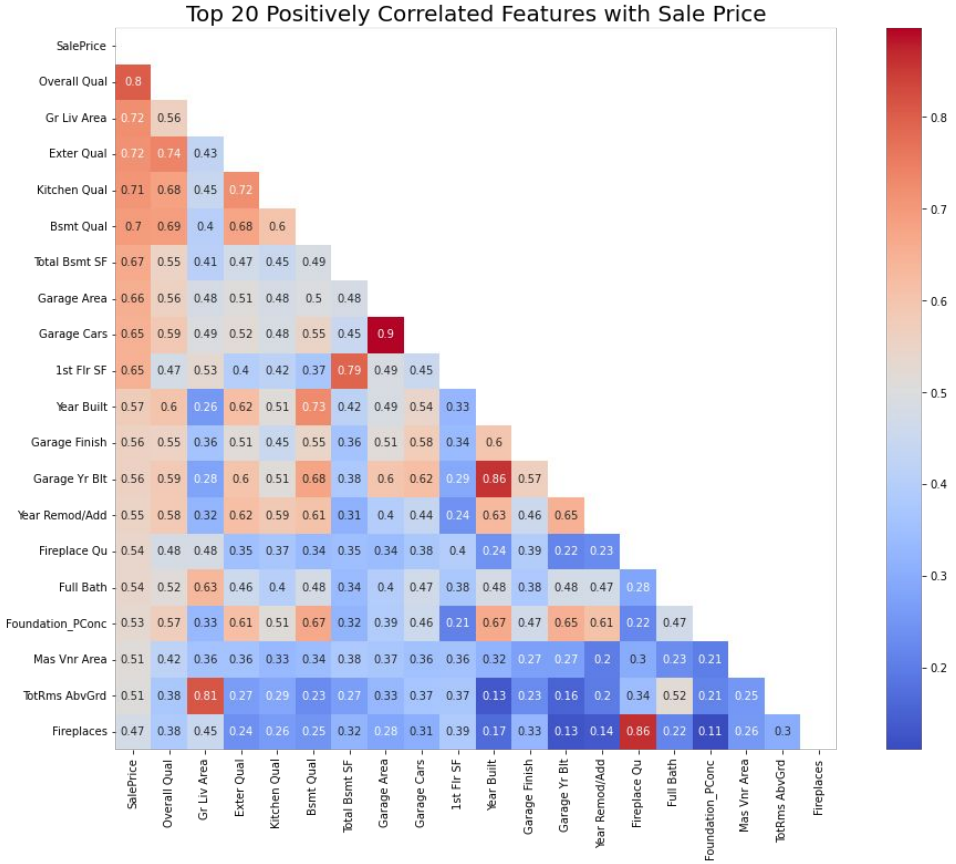
Notebook 2: Cleaning and Modifications of 'train' and 'test' datasets.

Notebook 3: Preprocessing and Feature Engineering.

Notebook 4: Modeling, Selection and Insights

1. Preliminary EDA of 'train' dataset

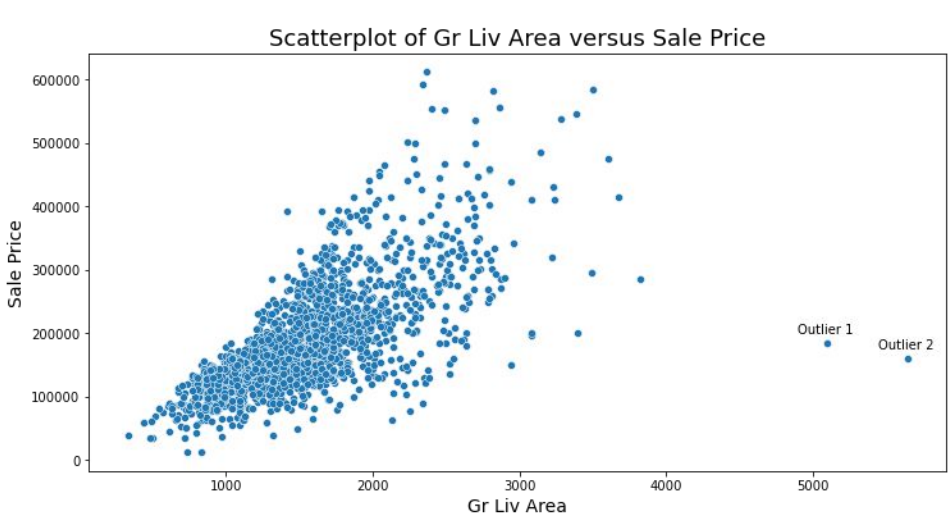
- To determine the extensiveness of data cleaning/ modifications
- Through the EDA done, it was determined that in order to create a feasible linear regression model, some work had to be done to reduce multicollinearity and skewness.

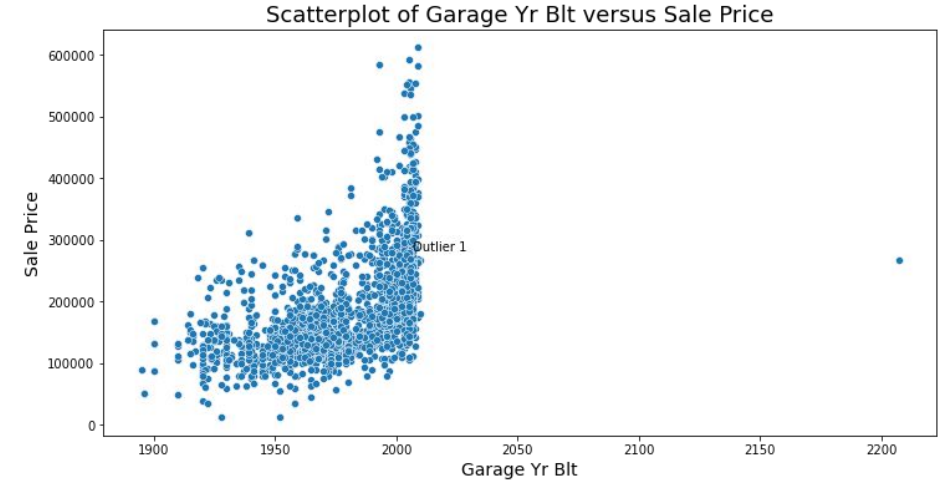


1. Preliminary EDA of 'train' dataset

2. Data Cleaning & Modifications

- Treat null/zero values in the datasets
- Identify and then treat or remove the outliers.
- Encode both ordinal and and nominal variables (228 variables, up from the initial 81)
- Done to both training and testing datasets





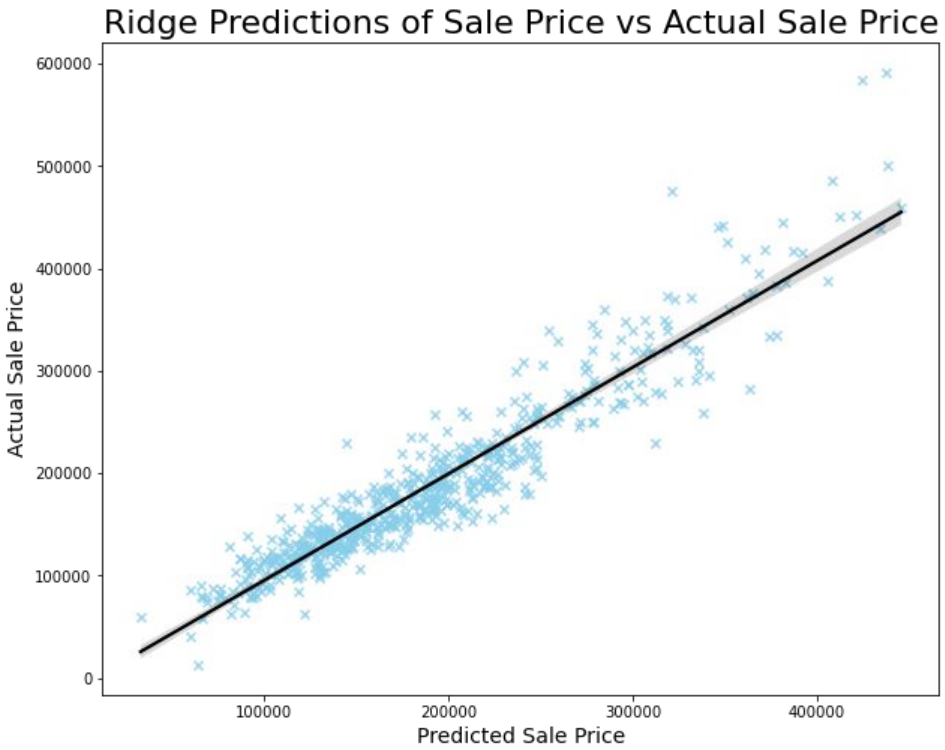
2. Data Cleaning & Modifications

3. Preprocessing and Feature Engineering.

- Reduce some of the redundant features (dimensionality reduction techniques):
  - correlation and
  - variance analysis
  - recursive feature elimination.

4. Modeling, Selection and Insights

- Model generation
  - Standard Linear Regression
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
- Evaluation
- Selection of the best performing model



4. Modeling, Selection and Insights

Model Limitations

- Generalises well to the city of Ames but may not generalise to other cities (differing external factors)
- Further preprocessing of the target variable (ie. add a log function to SalePrice to ensure a more normal distribution)
- Tradeoffs between interpretability and accuracy

Recommendations