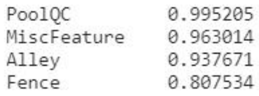
# House Price Prediction Report

Tao Shan

### Describe the data

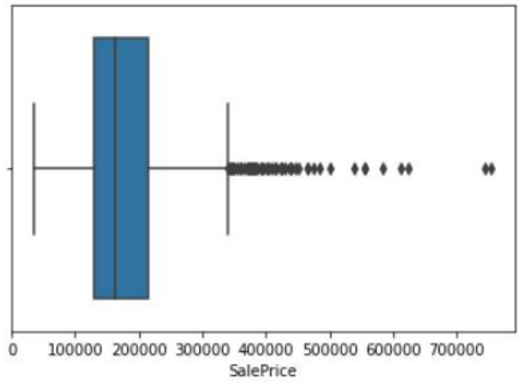
The size of the dataset is (1460, 81). The dataset describes house price, including building information, property, year of building, and others (80 attributes in total). The data has different types, including numeric and objects. Some of the attributes have lots of missing values.



```
PoolQC         0.995205
MiscFeature    0.963014
Alley          0.937671
Fence          0.807534
```

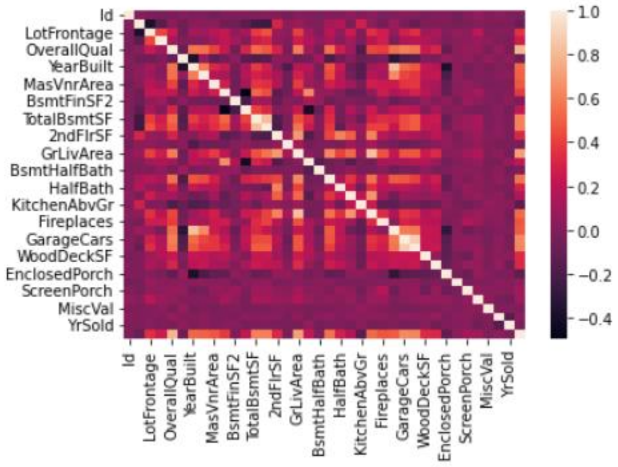
(Top missing values with missing value > 80%)

We found that PoolQC, MiscFeature, Alley, and Fence have many missing values.



(boxplot for the target attribute)

SalePrice is our target attribute. We can see the data has outliers when the price is roughly higher than 350K. Most of the price ranges in 140K – 210K (between 25th and 75th quantile)



(heatmap for all numeric attributes)

The heat map shows the linear relationship between pairs of numeric attributes. By python script, I found no pairs have a very high correlation, so there is no problem with multicollinearity.
To test the skewness, I found all columns with skewness lower than -3 or greater than 3. These columns will treat in preprocessing steps.

**Objective**

This model aims to find the relationship between the target sale prices and all other non-target attributes. We want to find new houses' prices before selling them. I'm using three linear models to find the linear relationship in this project. To compare these three models, I'm using the RMSE score and 10-fold's cross-validation score to estimate the model performance.

**Steps**

1. Data Cleaning

I drop the columns with more than 80% missing values for missing values. Also, I drop columns with too many unique values (more than the number of variables / 2). For other missing values, for categorical data, I'm using the mode to fill the missing values. For numerical data, I'm using the mean to fill the missing values.

1. Feature Engineering

Then, I use different methods for categorical data and numerical data. For categorical data, I'm using the hash encoder for categorical columns with more than four unique values and one-hot encoding for categorical columns with less than or equal to 4 unique values. For numerical columns, I'm using the power transformer for skewed data to reduce the skewness. Then use the robust scaler to minimize the problem of outliers.

**Models**

The models I have chosen are linear regression, lasso regression and rigid regression.
The 10-folder cross-validation with the RMSE score result shows below:
Linear regression: 35139
Lasso regression: 35130
Ridge regression: 34862

From the result score metrics, Ridge regression is the best model. Ridge regression performs L2 regularization to detect multicollinearity. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are significant—the results in predicted values being far away from the actual values.

**Findings**

The model results in a linear relationship (Y = b0 + b1x1 + b2x2 + …) between price and other attributes. It is easy to see how the model suggests the relationship. Before we sell the house, we can predict the future price and helps the house seller to determine its price.

### Revisit

Since linear methods' accuracy is not high enough, and the dataset is too small to analyze, I'll use new approaches such as logistic regression, random forest, xgboost, and deep learning to do a better model. Also, I'll use more data found on government websites, Kaggle, Google, and other open data sources.

### Appendix

Data: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Code: https://www.kaggle.com/taos2000/house-price-prediction-linear-model