Course Project

Housing Price Prediction

[1]. Wang Tianqi 1401213465
[2]. Matta Uma Maheswara Reddy 1601213442

- Data Description
- Feature Engineering
- Regression Models
- Conclusion

Data Description

- **Variable to be predicted:** SalePrice
- **Dataset Size**
- Train: 1460   Test: 1458
- **Features**
- Numerical: 36   Categorical: 43
- **Metrics**
- Root-Mean-Squared-Error (RMSE)
- **Goal**
- To predict the SalePrice and minimize the RMSE

Data Description

- Numerical

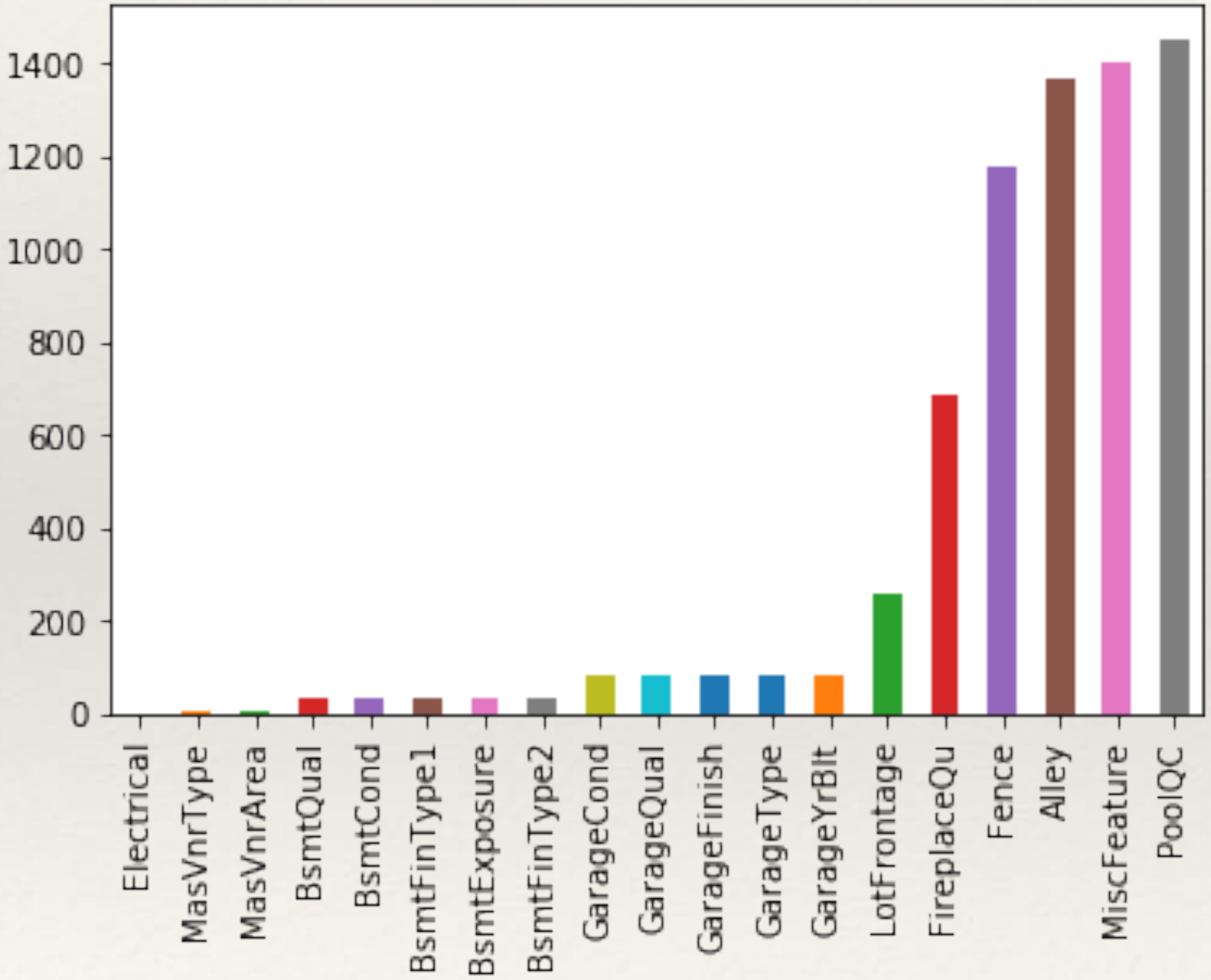
Variable	Description
GrLivArea	Above grade (ground) living area square feet
YearBuilt	Original construction date
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
LotArea	Lot size in square feet
1stFlrSF	First Floor square feet
FullBath	Full bathrooms above grade
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Fireplaces	Number of fireplaces

- Categorical

Variable	Description
OverallQual	Rates the overall material and finish of the house
Overallcond	Rates the overall condition of the house
MSZoning	Identifies the general zoning classification of the sale
Utilities	Type of utilities available
HouseStyle	Style of dwelling
Exterior1st	Exterior covering on house
Foundation	Type of foundation
Heating	Type of heating

Feature Engineering

- Missing Value



- Five features have over 50% missing values, so we delete those features

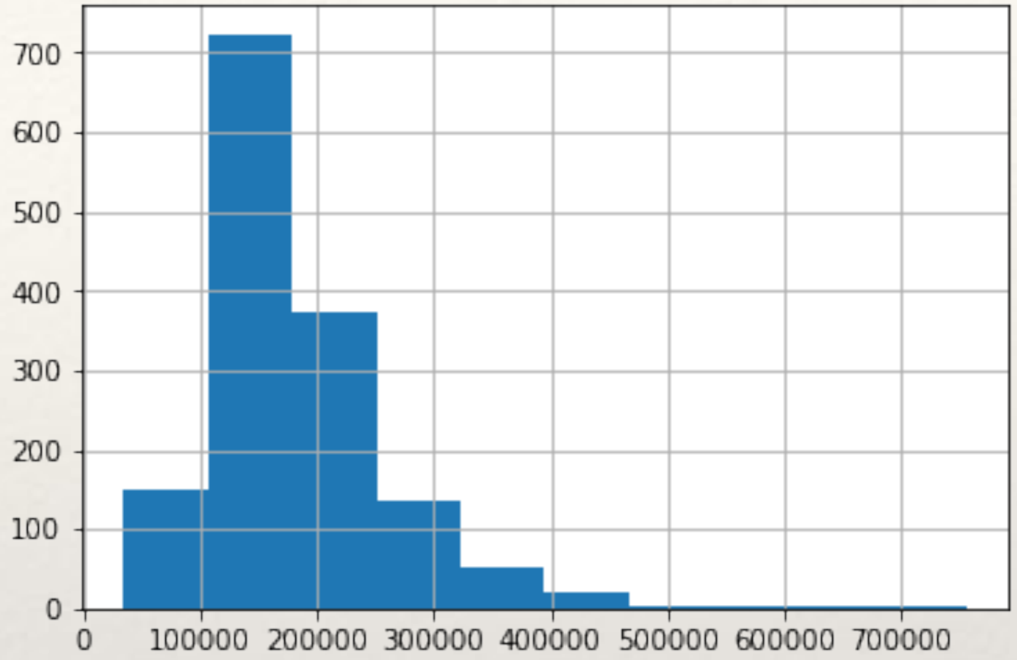
- Replace the numeric missing value (Nan) with mean of their column

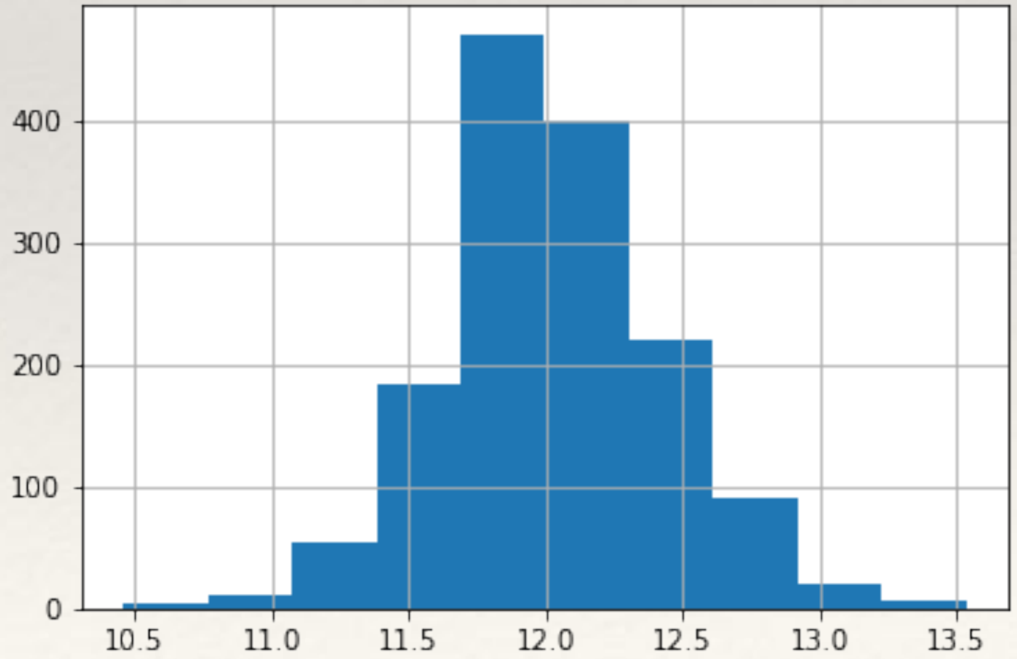
Feature Engineering

- **Categorical Features**

  Create dummy variables

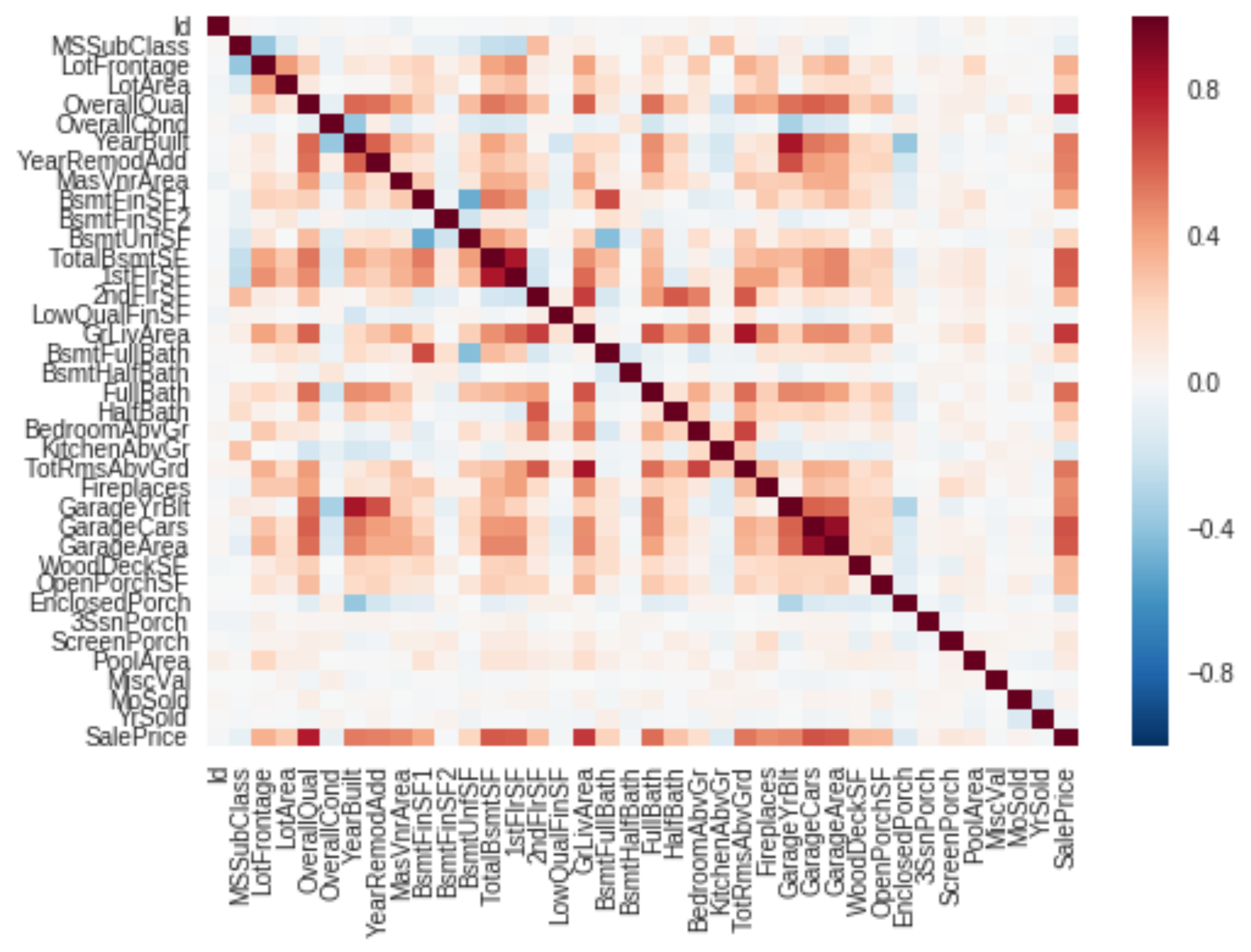
- **Numerical Features**
- Unskew some those highly skewed features (Skewness>1)
- SalePrice → Log Transformation





Feature Engineering

- Correlation Analysis



- **Feature Reduction**
  - Manually pick five features
  - PCA
  - Regularization

Regression Models

- **Linear Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Gradient Boosting Regression**
- **Linear Regression with Regularization**
- Lasso; Ridge;

Regression Models

Model	All Feature	PCA	Selected Feature
Linear Regression	0.1652	0.1495	0.1834
Gradient Boosting	0.1253	0.1743	0.1857
Decision Tree	0.2051	0.2858	0.2499
Random Forest	0.1529	0.2083	0.1962
Ridge	0.1273	0.1495	0.1834
Lasso	0.1231	0.1492	0.1834

Conclusion

- The best fit model is Lasso with all features, the model eventually picked 110 variables and eliminated the other 178 variables

- **Potential Improvement**
- Use GridSearchCV to better fit our parameters
- Go deeper about feature engineering
- Further process about outliers