Semester Project

### Group 6: Ada Li, Xinyu Wang, Jingsi Wu,Yixuan Huang

1. **Introduction**

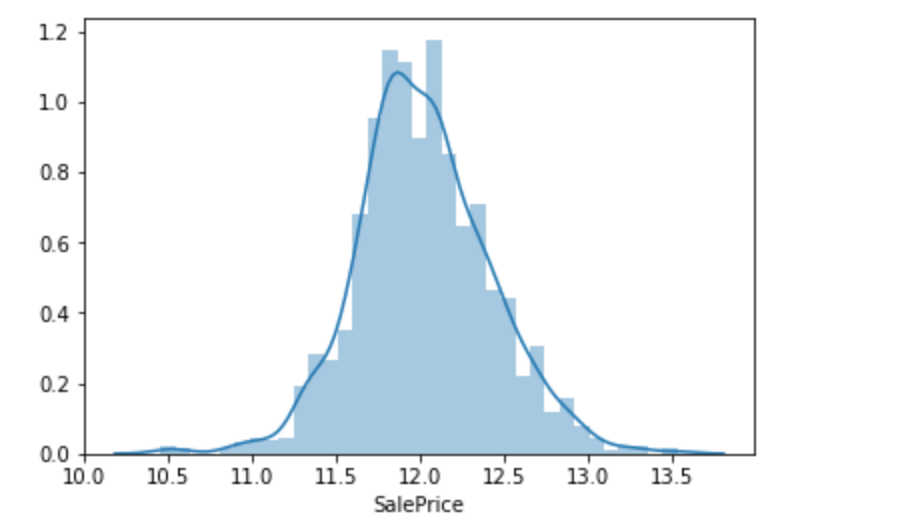
Kaggle competition: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

The goal of this competition is to predict sales prices of each home in Ames, Iowa. We choose this competition as an exercise in order to apply the data mining techniques like data feature engineering, Multiple linear regression, Random forest, XGBoost, ect.

1. **Initial Data Exploration**

The Dataset consists of 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. There are 1460 rows in the training set and 1459 rows in the test set.

We have found that for the target variable SalePrice, since its the original distribution is skewed, we decide to perform a log transform of it to fix this problem.

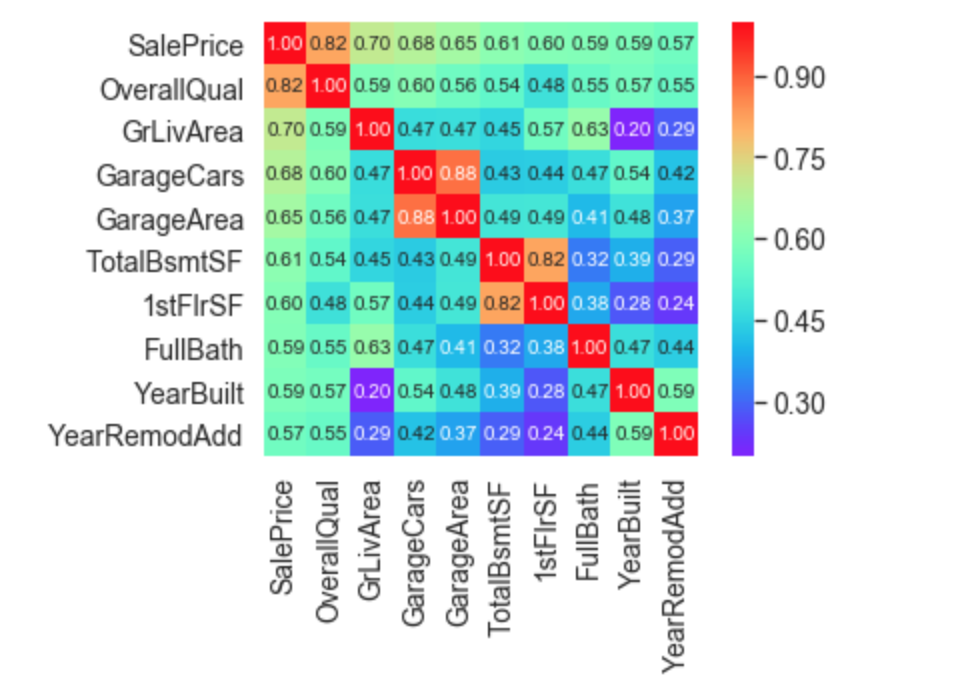


(our target after transformation)

We then try to check the relationship between the target variable and explanatory variables to have a look at the most useful predictors.

* For Numeric Variable

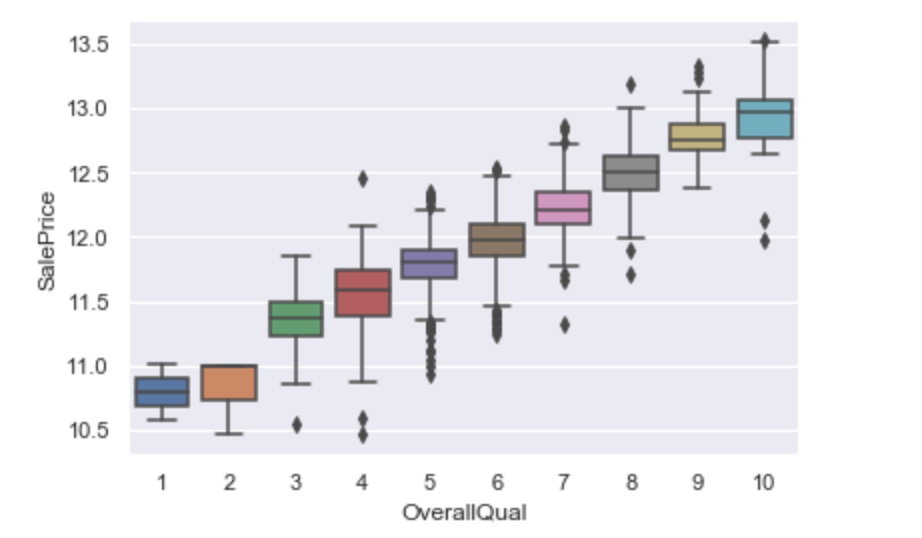
By checking the correlation between the target variable and numerical variable, we can find the explanatory variables which have the highest relationships(correlation) with the target variable. So we use these variables when we build our models. We also deal with multicolinearity problem by analysing relationships between those numeric variables.



(variables with highest linear relationship with our target)

* For Categorical Variables

By checking relationships between our target and some categorical variables, we can also find certain variables like OveralQual (Overall material and finish quality), CentralAir(Central air conditioning) have a strong relationship with the target variable.



(relationship between ‘OverallQual’ and ‘SalePrice)

1. **Data Processing and feature engineering**

Now we need to do the data processing and select the feature which will be used in predicting the sales price. From the heat map above we can see that there are lots of features in this dataset and some of them are highly correlated, so we choose to drop the highly correlated features to avoid multicollinearity. We also drop the columns with too many null values under certain criterions.

After that we encode the categorical variable with LeaveOneOut and then normalize the features.

So far we have got the new training set (with 64 features) and test set (with 64 features) which can be used to train and test the predictive model.

1. **Training & Evaluation**

Our next approach is training our dataset using different algorithms, including Multiple Linear regression, Penalized regression (Ridge Regression, Lasso Regression), Random Forest and XGBoost.

We check the accuracy of those algorithms by using k-fold cross validation. The performance of those algorithms are the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Multiple Linear regression** | **Ridge Regression** | **Lasso Regression** | **Random Forest** | **XGBoost** |
| Accuracy | 0.9986(+/-0.0048) | 0.99862 | 0.99866 | 0.9946(+/-0.0057) | 0.9973(+/-0.0049) |

Based on the performance, we choose to use the **Multiple Linear regression** to do the prediction.

1. **Kaggle group name and screenshots of the Kaggle leaderboard**

