Preprocess:

1. Through histogram, the data is actually skewed, in order to make the target normal-distributed, the transformation of taking log form should be made.

```
combi$Sale_Price = log(combi$Sale_Price + 1)
```

2. Deal with missing values. I found that only Garage_Yr_Blt has 159 missing values, based on common sense, garage is usually built with the house. Besides there are 2227 of 2930 whose Garage_Yr_Blt == Year_Built. Therefore I use Year_Built to fill in the missing values.

3. Feature selection. Considering all data are from the same state, their Longitudes and Latitudes are similar, so “Longitude” and “Latitude” should be dropped. Aside from that, I found a few dominant features which a large portion of observations only take the specific value. Those dominant columns should be dropped too.

4. Handle categorical data. There are 38 categorical data which is over 50 % of features. I use “caret” to perfrom one-hot-encoding to make all predictors numeric.

Model 1:

RandomForest is considered first but the RSME is high.
Therefore, I use a simple lasso model by using cv.glmnet to find to lambda.
Accuracy: 0.124

Model 2:

Xgmboost
Accuracy: 0.128

Running time of code: system 244.60
Computer system: Aspire V5-473PG, @1.80GHz 2.4GHz 8.00GB
Acknowledgment：
Lasso + GBM + XGBOOST - Top 20% (0.12039 on Leaderboard) using R （Aniruddha Chakraborty）