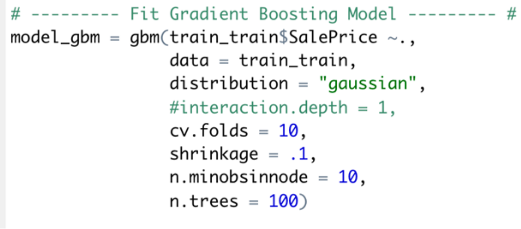
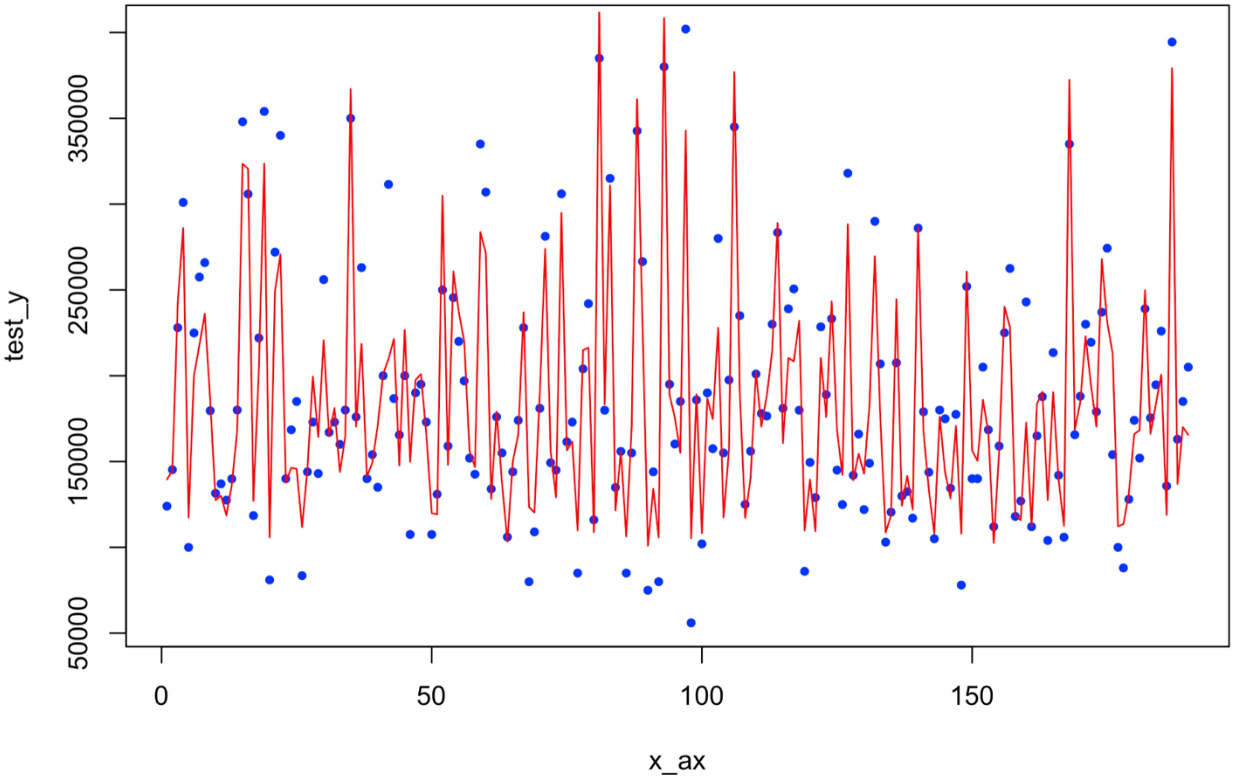
Machine Learning project process:

1. Drop variables with too many missing values
2. Impute missing values for train/test data set with Mode/mean
3. Scale the values for numerical data
4. Simplify year data into 5 years a block category
5. Some categorical features transformed into ordinal
6. Drop unique values > 90% of features
7. Split the training dataset to train and test (80% vs 20%)
8. Use gradient boosting to train the model and get R^2 = 0.886



```
# --------- Fit Gradient Boosting Model --------- #
model_gbm = gbm(train_train$SalePrice ~.,
               data = train_train,
               distribution = "gaussian",
               #interaction.depth = 1,
               cv.folds = 10,
               shrinkage = .1,
               n.minobsinnode = 10,
               n.trees = 100)
```

```
> cat('The R-square of the test data is ', round(rsq,3), '\n')
The R-square of the test data is  0.886
```



9. Re-train the model with all the training data
10. Predict the test data given with the model