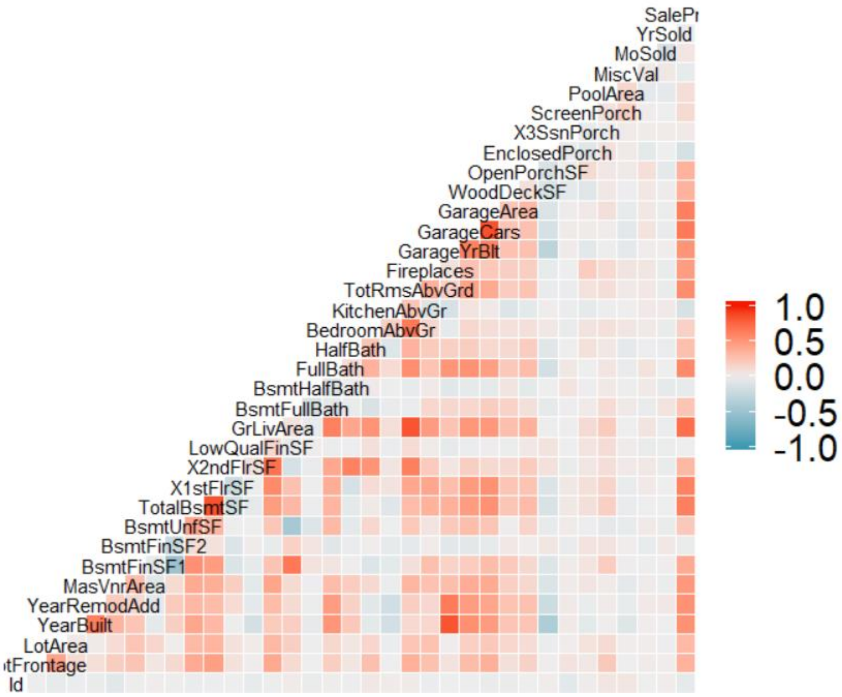
Group 11

2/12/2020

Lab 3 Exercise 2

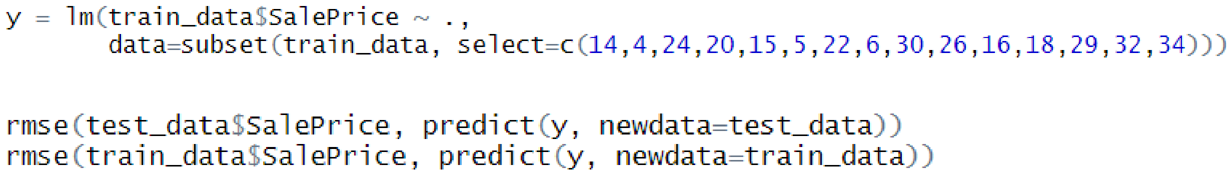
In this exercise, our goal is to build a model to predict Sale Price using any method and number of regressors, with the lowest possible RMSE. We tried a variety of different methods before we made the best model. At first we used the forward selection method that ended up with 15 regressors we thought would be good predictors for sale price. However, the result from that was not ideal.

To better decide which regressors we should use for prediction, we created a heatmap showing how each variable is correlated with each other. We tried using all of the regressors that have high correlations with sale price, yet we didn't think the resulting RMSE was low enough.



Then we wrote a function that first sets up a vector to hold our regression variables that the function decides works best and a vector of all our parameters possible. Then it loops through each parameter and calculates the RMSE with the added parameter and any of the parameters that have been added to our regression vector. Then once we find the parameter that creates the lowest RMSE we remove it from the parameter vector and into the regression vector. Once we loop over the entire parameter vector without lowering our RMSE we end the function. Using this function, we were able to choose which regressors to use. and the variables we ended up using are: Yearbuilt, MasVnrArea, YearRemodAdd, Fullbath, KitchenAbvGr, BsmtFinSF1, Fireplaces, Miscval, WoodDeckSF, BsmtHalfBath, Halfbath, X3SsnPorch, Yrsold. All but five regressors we ended up with are the ones that are highly correlated with sale price on the heatmap.

The model we ended up with is:

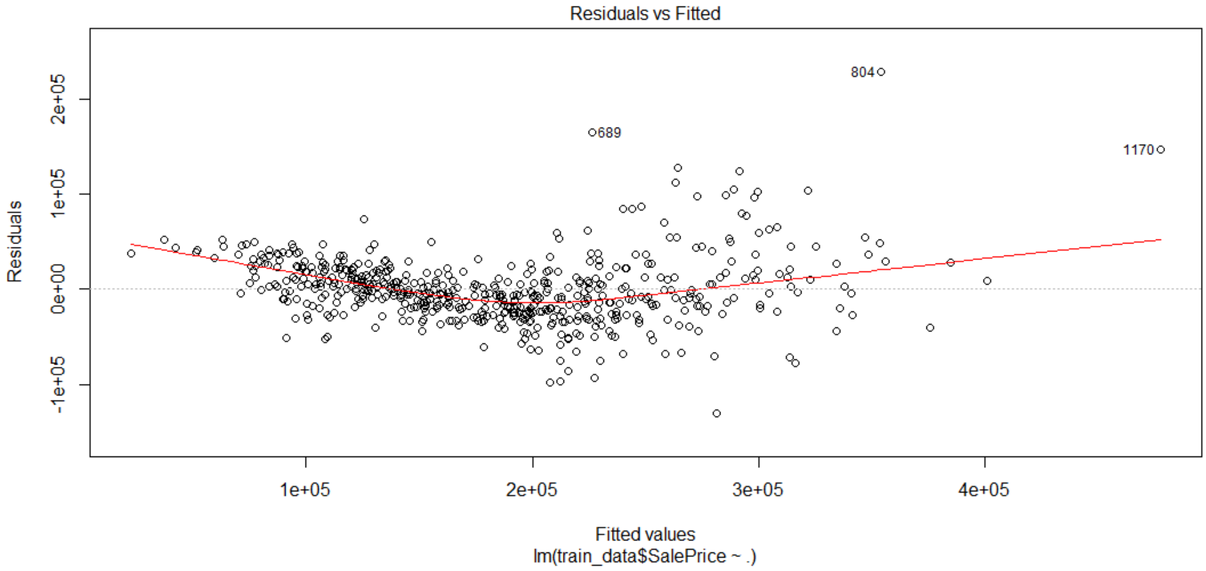


```
y = lm(train_data$SalePrice ~ .,
       data=subset(train_data, select=c(14,4,24,20,15,5,22,6,30,26,16,18,29,32,34)))

rmse(test_data$SalePrice, predict(y, newdata=test_data))
rmse(train_data$SalePrice, predict(y, newdata=train_data))
```

The lowest RMSE on the testing data is **47472.73**

Here’s what our model looks like.



This graph proves that the model we developed is solid and it would be a practical tool to predict sale price, given our chosen regressors. We are confident that we did well among the student groups because our model is strong, not only in a theoretic sense but also in a computational sense. Also, because we put in a lot of effort in this lab and the function we wrote worked very well and we are very proud of it.