SSC 442 Lab 3

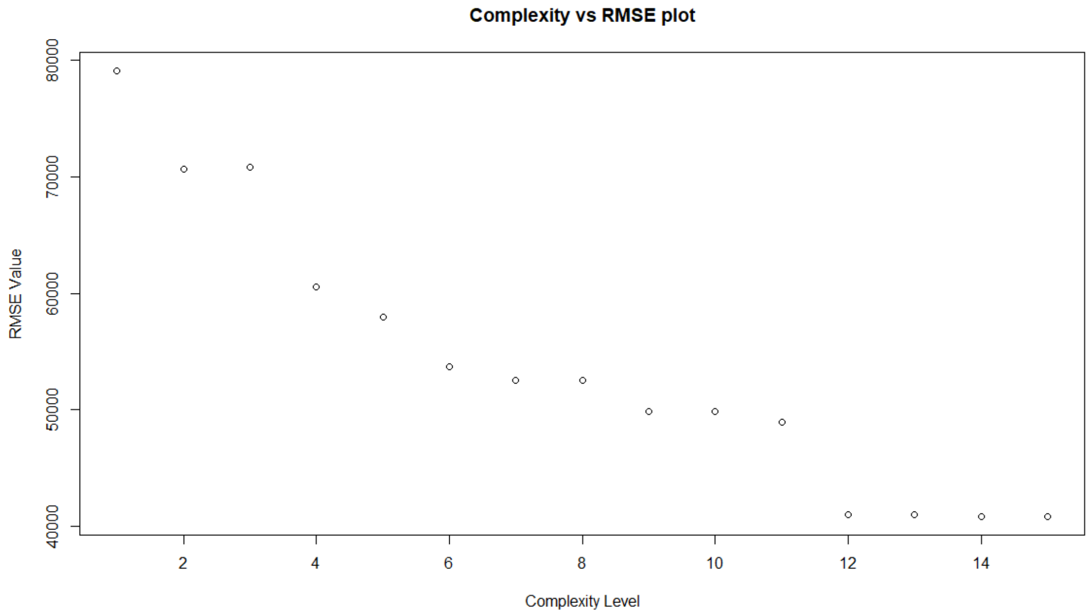
Team Awesome (19) - Sayem Lincoln, Joshua Schwimmer, John Townshend.

2/12/2020

Ex 1 Part 3

A chart plotting the model complexity as the x-axis variable and RMSE as the y-axis variable

```
plot(x=compPlot[,1],y=compPlot[,2], xlab = "Complexity Level", ylab = "RMSE
Value", main="Complexity vs RMSE plot")
```



Describe any patterns you see.

Ans - A scatter plot that forms a decreasing linear pattern.

Do you think you should use the full-size model? Why or why not?

Ans - A full strecthed model presents a plot that is very strecthed out, as the complexity increases the RMSE values decrease gradually, I say gradually because a pattern can be seen, as the RMSE to complexity values stay constant for when 2 or 3 new vaaribales are added to the model but the RMSE decrease when more then 3 variables are added to the model. When a full model is used the RMSE vs complexity plot stayed the same compared to the three previous models' plots so the full model presents the final ourtcome of how all the previous models have progressed.

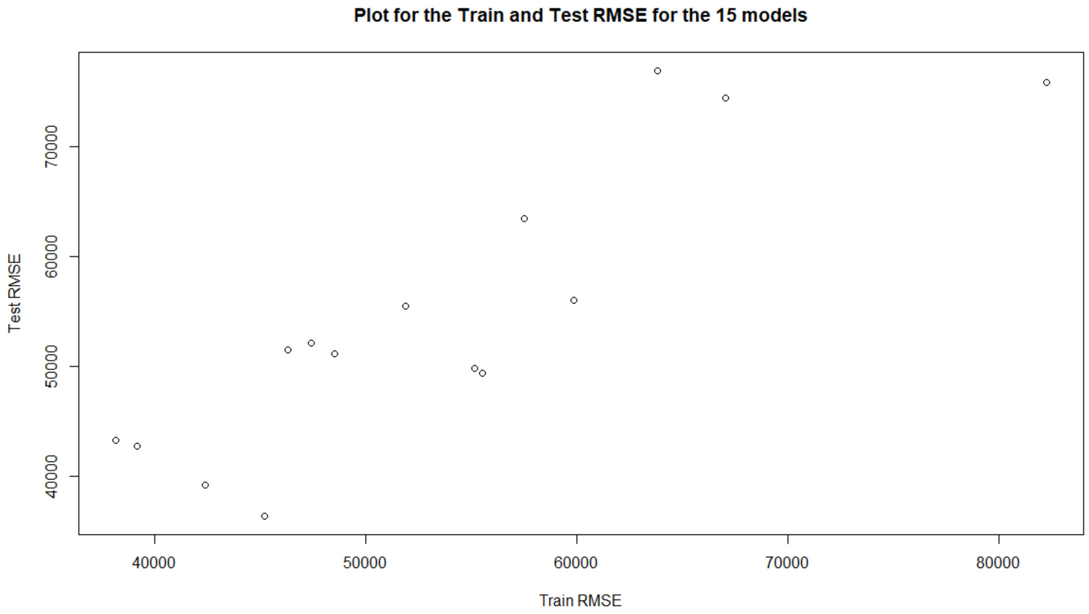
What criterion are you using to make this statement?

Making it on the basis of how the outcome came to be and how the RMSE value went down instead of going up, as the error between the models decreased and a proper value for RMSE got produced.

Ex 2 Part 1

Plot for the Train and Test RMSE for the 15 models

```r
plot(x=rmseVal[,1],y=rmseVal[,2], xlab = "Train RMSE", ylab = "Test RMSE",
main ="Plot for the Train and Test RMSE for the 15 models")
```



Ex2 Part 3-

Task -

In a PDF write-up, describe the resulting model. Discuss how you arrived at this model, what interactions you're using (if any) and how confident you are that your group's prediction will perform well, relative to other groups.

Answer –

Our resulting model values are Train RMSE: 44072.28 and Test RMSE: 37741.19.

We received these outcomes from our model by running selected variables through our model. We selected three variables from the Ames dataset making them our predictor variables as we believe these variables can present us with the best correlation causation scenario when regressed against our response variable SalePrice.

The model we are using is model 12: lm(data2$SalePrice ~ MSSubClass+LotFrontage+LotArea+YearBuilt+YearRemodAdd+MasVnrArea+BsmtFinSF1+ BsmtFinSF2+BsmtUnfSF+TotalBsmtSF+X1stFlrSF+X2ndFlrSF, data = data2)

We interacted the predictor variables against the response variable through a linear model regression function, then to get a predicted RMSE value we ran our final resulting model from Ex 1 Part 2 (the one with 15 variables and complexities) through our Test and Train function, we did this in order for the outcome to get tested and trained through our model until we received an outcome that has the lowest RMSE possible.

Additionally, it can be intuitively implied that these variables play a part on a property's Sale Price, so regressing SalePrice on these variables can present us with a smaller RMSE value for both Train and Test components.

We are confident with our outputs and we believe that our model has regressed properly on the variables that we chose, and an optimal RMSE values for both Train and Test component was outputted.