Exercise 1, Part 3

The chart shows a clearly downward sloping line; as complexity increases, RMSE always decreases, but by differing amounts. There is a large difference in RMSE when changing the model size from 1 to 3, but the difference gets smaller and smaller as you add more. I do not think that we should use the full-sized model because it would be overfitting the data and therefore would give us inaccurate results. In the lab, it says that we may have to fit to a specific dataset-not just the full-size model- in order to allow our model to predict more accurately and not overfit. This is why I believe we should not use the full-size model.

Exercise 2, Part 3

The final model we ended up with is SalePrice ~ GrLivArea*PoolArea + YearBuilt*GarageCars + TotalBsmtSF + BedroomAbvGr + BsmtFinSF1 + TotRmsAbvGrd*KitchenAbvGr + YearRemodAdd*Fireplaces + ScreenPorch*LotArea + WoodDeckSF , data = train_data). We interacted several variables together such as GrLivArea and PoolArea, YearBuilt and GarageCars, TotRmsAbvGRd and KitchenAbvGR, YearRemodAdd and FirePlaces, and ScreenPorch and LotArea. We interacted these variables because they seemed to be correlated with eachother and they helped bring down the RMSE. We also deleted the variable “MSSubClass” from our regression because it resulted in a lower RMSE when we did not include it. Our resulting RMSE for the model is 34,485.71 which is quite a bit lower than where we started, at 38,090.61. I am confident that my group did well on this project, but if I had to guess where we stood with other groups I would say about middle of the road (but I hope we did better than that).