**Statistical Inference:**

After we finished the data wrangling and prepare our data ready for modeling, we had to do some statistical analysis especially with this linear regression model we were using, statistics certainly pay very important role.

Firstly, we had to discuss the concept of BLUE (Best Linear Unbiased Estimators) which was very important for the assumption we were going to make for this model to work. These are the important assumptions for this model:

* The residual term must be normally distributed which mean it has mean of zero and standard deviation of 1
* The residual term must be homoscedasticity and independent
* There should be no perfect collinearity
* The sample data set must be random

In the modeling phase, we had to split the data into training and test set so that we can avoid over-fitting. We used Skit-learn library to split the data using 70/30 splits and results showed: Training and testing set sizes (21131, 32) (9056, 32). With this training set, we had to train the data with linear regression by fitting the target variable with independent variables. Next, we used this fitting value to compare with the test data set to generate the accuracy score or a matric of MSE. The accuracy scores were as following:

**Testing Score: 0.411476963149**

**Training MSE: 0.276040455527**

**Testing MSE: 0.269870736571**

The score of 0.41 was not necessarily good but also not so bad either and that is why we believe using regularization might be able to improve the results of this score. The MSE of training and test were very similar which indicated there should be no over-fitting presented in the model. In addition to this result, we also had not forget the statistical result summarized below. The table described each coefficient with its parameter values and standard error as well as p-value. There were two main coefficients that are not statistically significant: the bedroom and G\_E. They were not significant because their p-value were higher than 5% of significance level we assumed for this model. The R2 known as the coefficient of determination for the model is 0.42 which was not bad either and it reflected how much information that the independent variables helped describe the data. The adjusted R2 was 0.419 which was slightly lower than R2.

In order to improve our result, we would use regularization like L1 or Lasso to intervene. Lasso would perform both variable selection and regularization. After we ran Lasso, we got our result as following:

Testing Score: 0.34325587227

Training MSE: 0.306110557796

Testing MSE: 0.301153923281

These results were a little worse than the normal linear regression. The score was only 0.34 whereas previous score was 0.41, so the model was worse off than before. The MSE for both train and test were still similar to each other 0.306 and 0.301, respectively. However, even the score of Lasso did not improve, this model was still better in the sense that it reflected what coefficient were important and what not. The result of these terms were presented below: