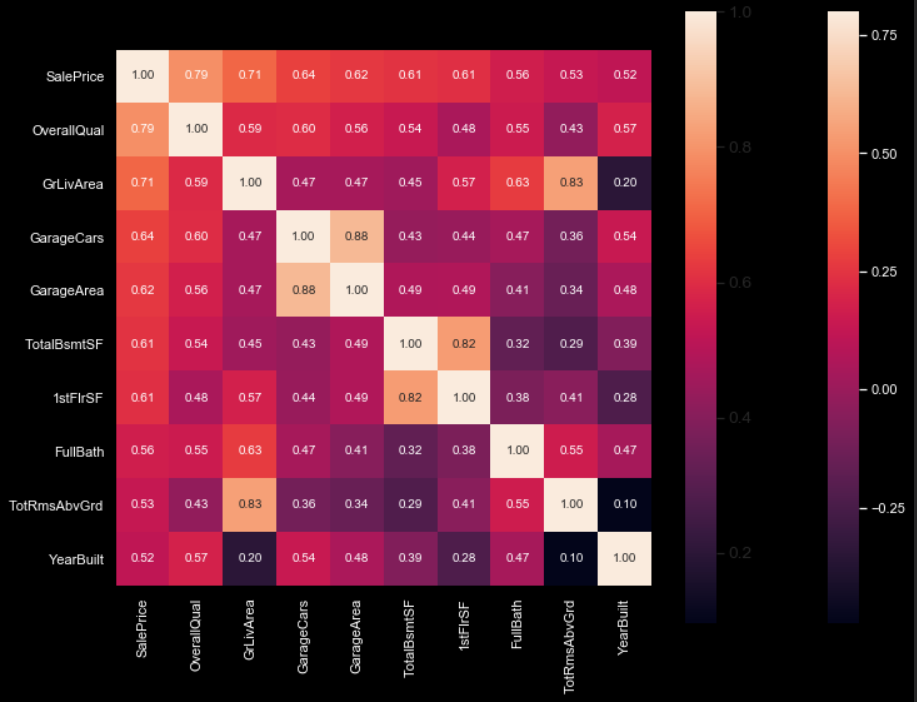
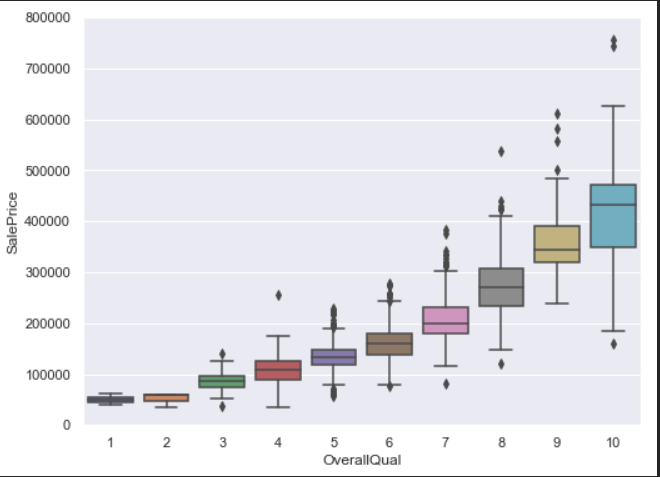
**Problem**

This Kaggle competition’s goal is to have the data scientist predict final home price with the included explanatory variables.

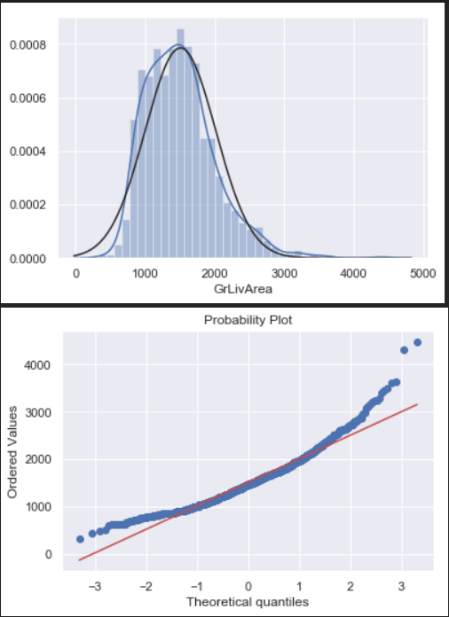
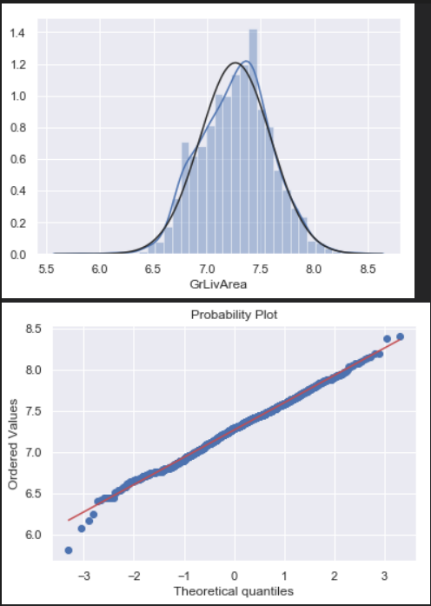
**Data**

The housing data contains 1459 rows and 79 descriptive variables as well as the sale price. Much of the data is linearly related to the sale price. Below are the 10 highest corelated variables with SalePrice. GarageCars and GarageArea showed collinearity so only one could be used in a linear model.



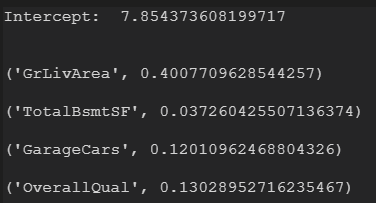


SalePrice and several of the independent variables showed non normality and needed a log transformation to be useful in the model. As an example of the efficacy of the log transformation, the before and after histogram and probability plots for GrLivArea are shown below.



**Analysis**

I chose to use a linear regression model due to the linear nature of the data. Chosen independent variables and their coefficients are below.

****

The model scored 80.6% on the train data.