# Regression Analysis

On the King's County Housing Dataset

# Data Source

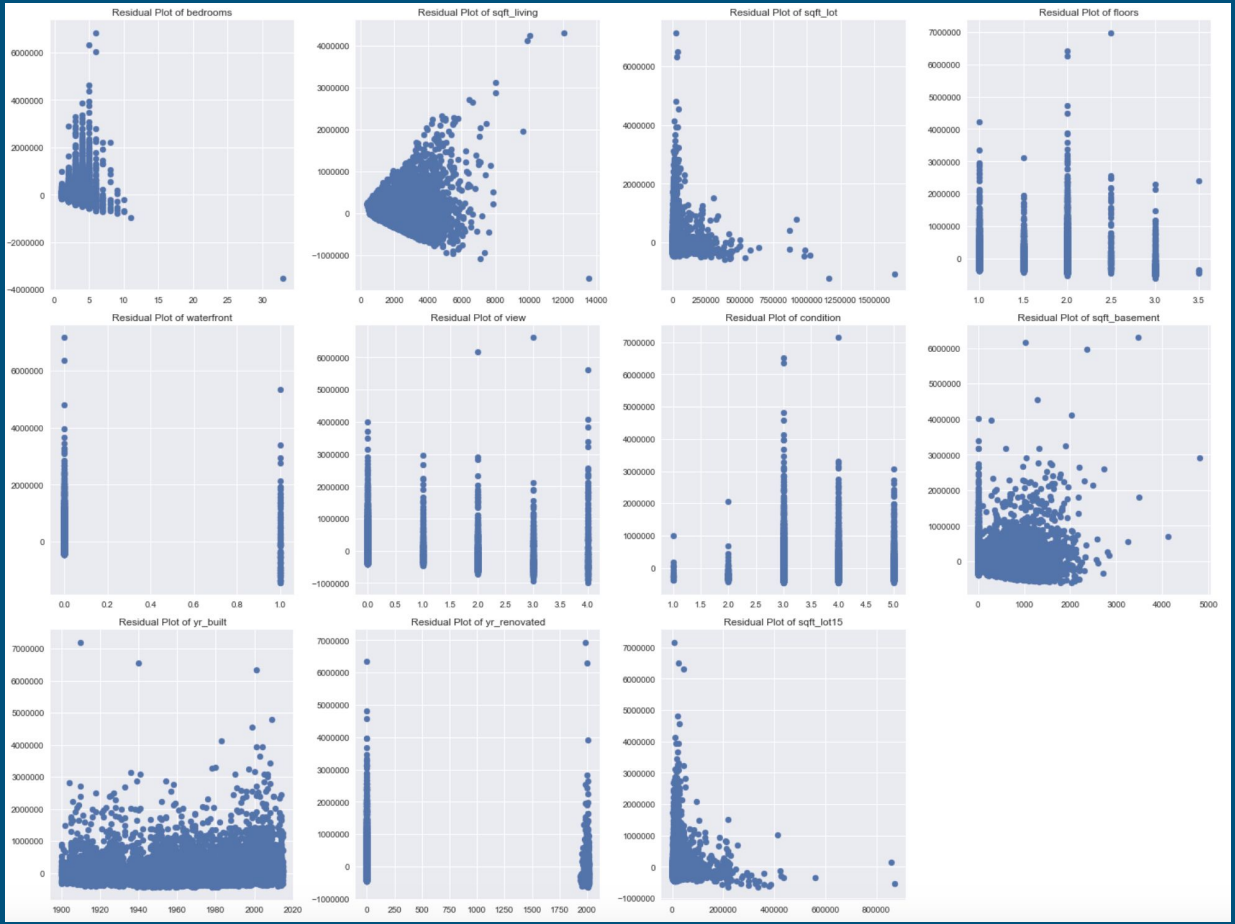
Kings County Data Set

Assumptions of homoscedasticity are violated

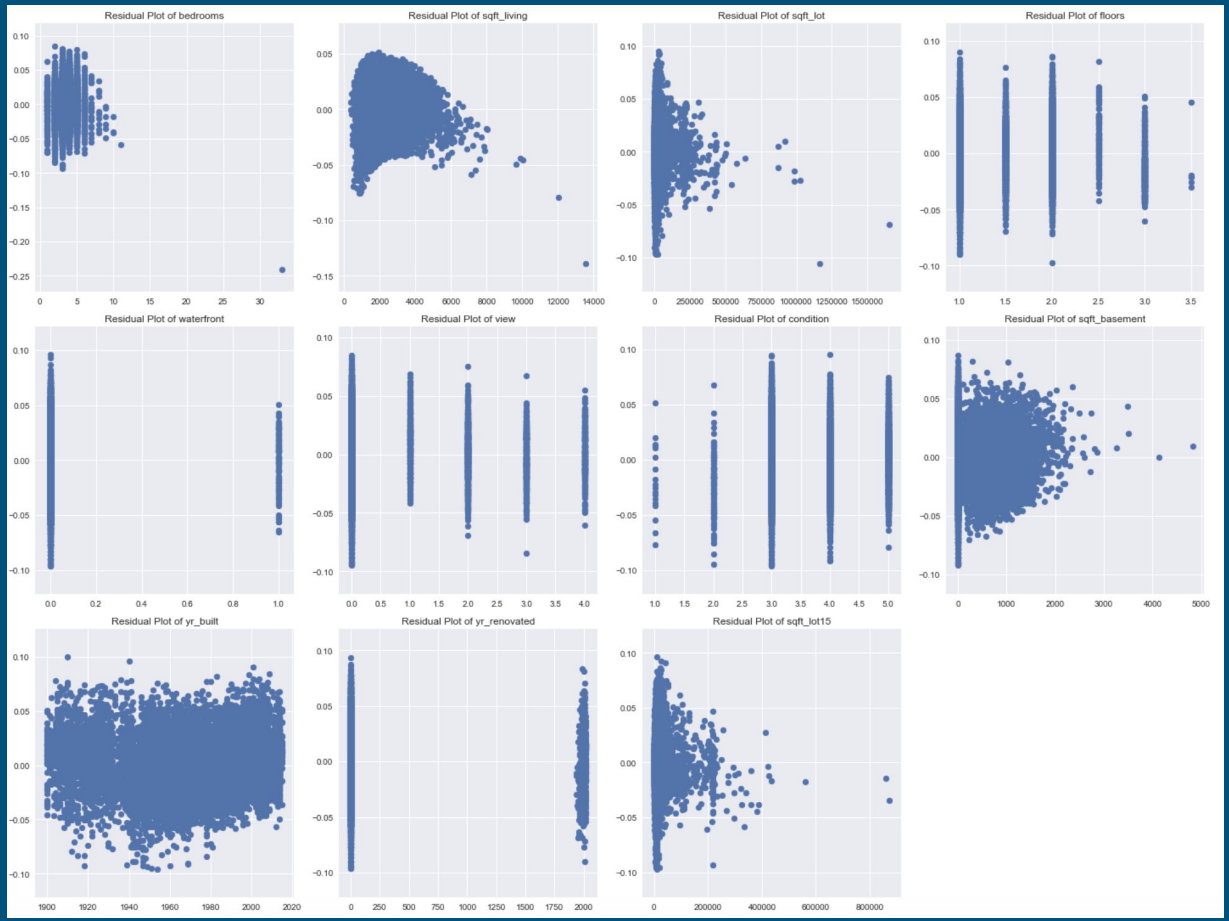
Assumptions of normality are also violated

Thus, we do a Box-Cox Transformation of the target data (price) to see whether we can fix model assumption violations

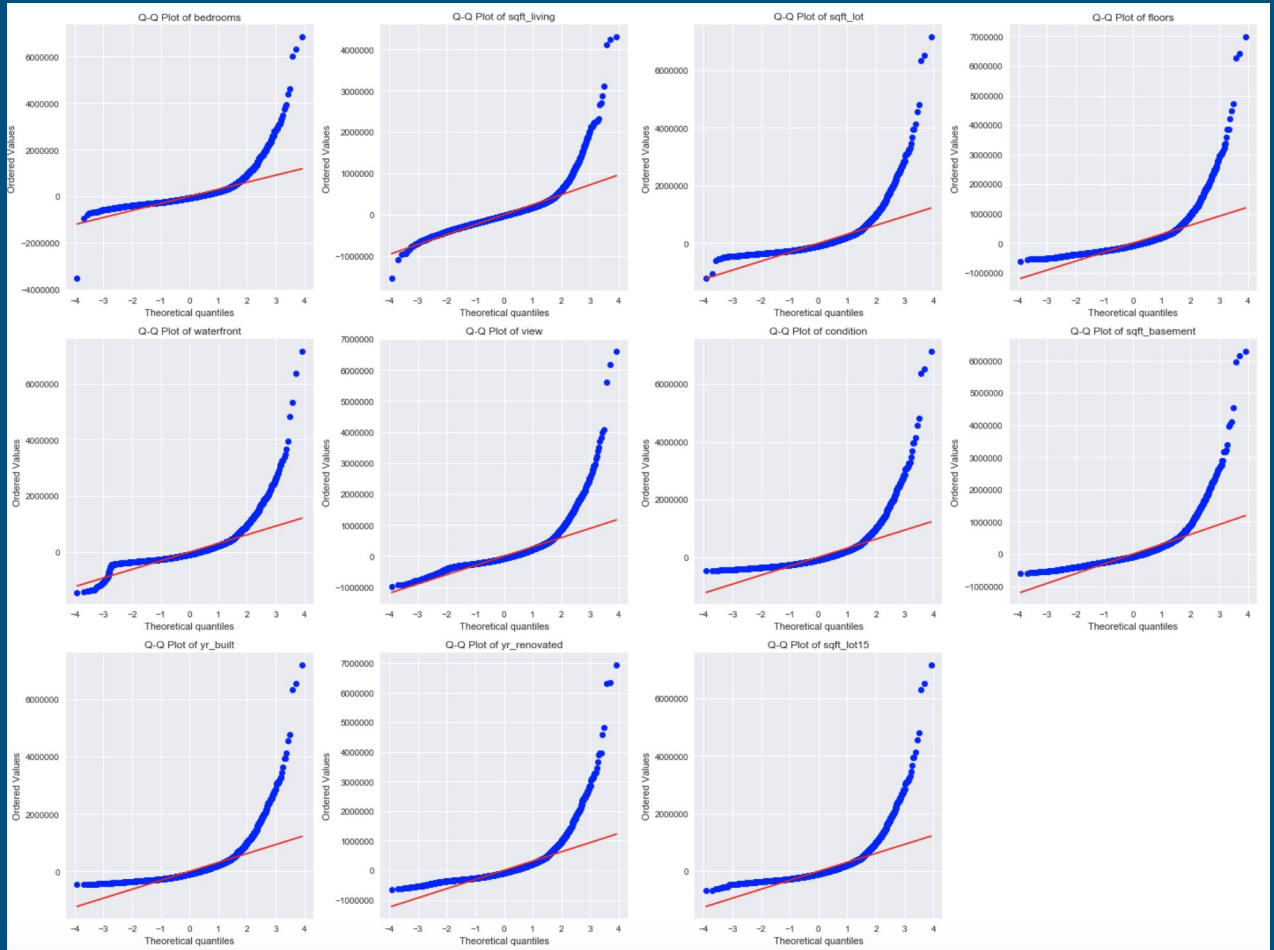
# Homoscedasticity: Before Box-Cox Transformation



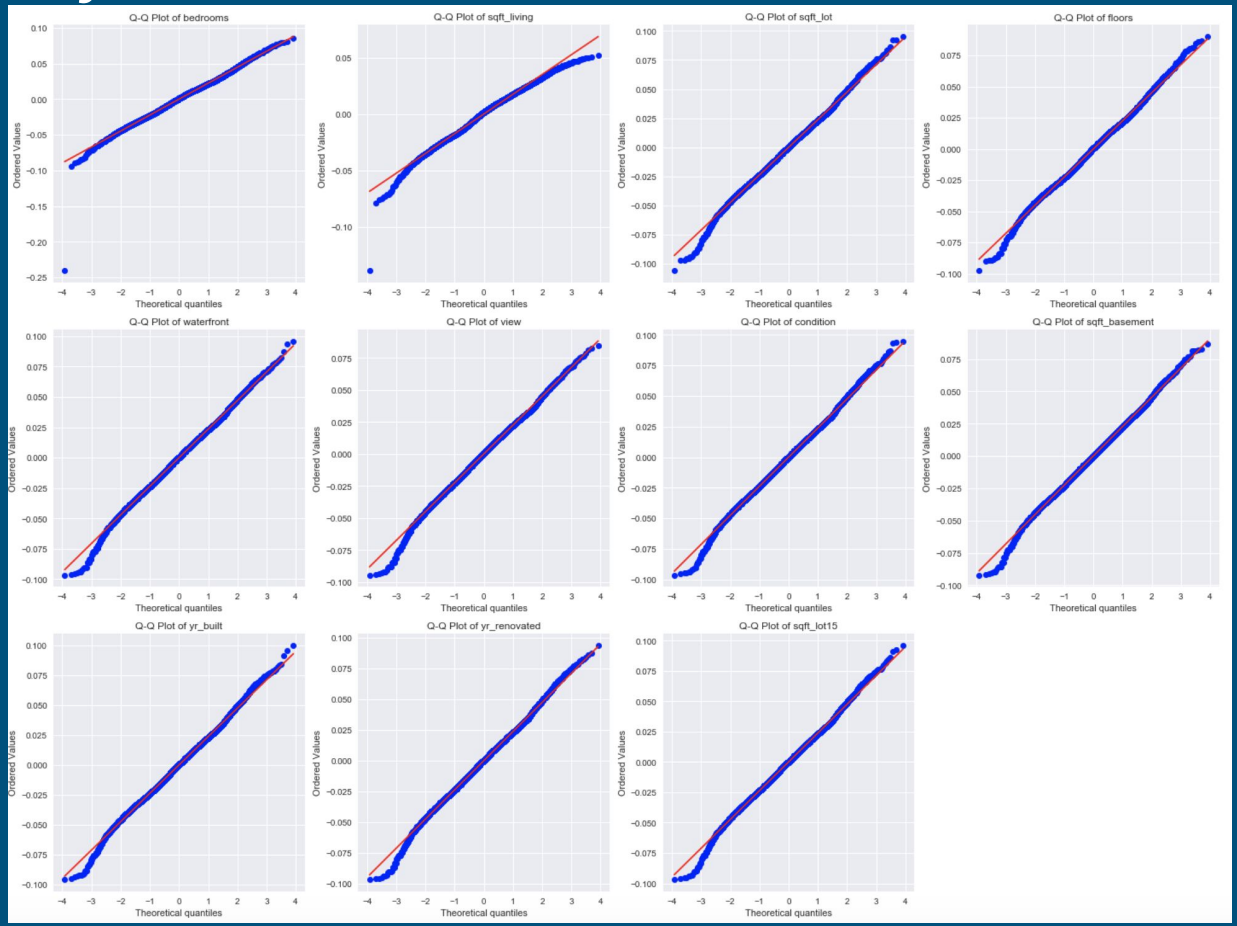
# Homoscedasticity: After Box-Cox Transformation



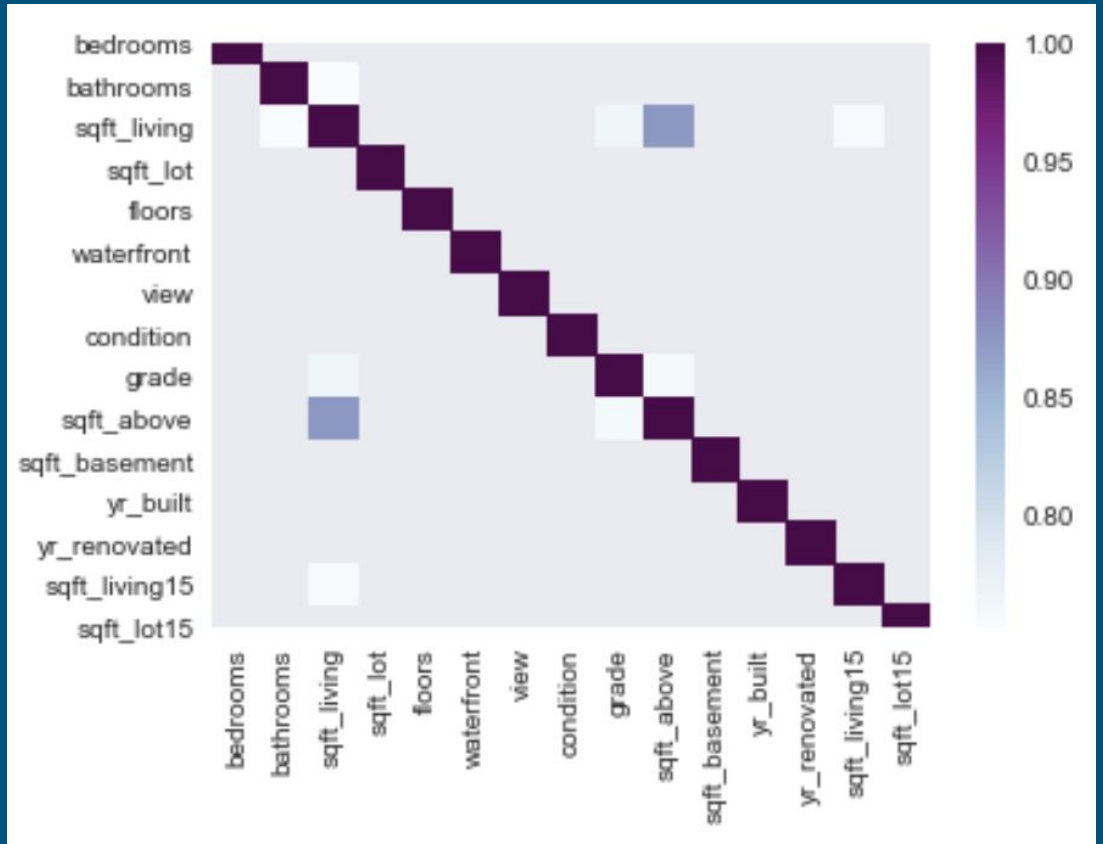
# Normality: Before Box-Cox



# Normality: After Box-Cox



# Eliminating Multicollinearity




As we can see above, 'bathroom', 'grade', 'sqft_above', and 'sqft_living15' are highly correlated, so for the stability of the model, we will exclude them from our predictor variables in our regression analysis.

# Turning Zipcode into Dummy Variables

Since there is a big difference of housing prices between poor and rich neighborhoods, it should be included in our multiple regression analysis, at the cost of greatly increasing the complexity of the model.

# Multiple Regression

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | price_boxcox | **R-squared:** | 0.833 |
| **Model:** | OLS | **Adj. R-squared:** | 0.832 |
| **Method:** | Least Squares | **F-statistic:** | 765.1 |
| **Date:** | Sun, 20 Oct 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 14:59:41 | **Log-Likelihood:** | 41288. |
| **No. Observations:** | 12343 | **AIC:** | -8.241e+04 |
| **Df Residuals:** | 12262 | **BIC:** | -8.181e+04 |
| **Df Model:** | 80 | | |
| **Covariance Type:** | nonrobust | | |

***price_boxcox*~**bedrooms+sqft_living+sqft_lot+floors+waterfront+view+condition+sqft_basement+yr_built+yr_renovated+sqft_lot15+zipcode_cat_98002+zipcode_cat_98003+zipcode_cat_98004+zipcode_cat_98005+zipcode_cat_98006+zipcode_cat_98007+zipcode_cat_98008+zipcode_cat_98010+zipcode_cat_98011+zipcode_cat_98014+zipcode_cat_98019+zipcode_cat_98022+zipcode_cat_98023+zipcode_cat_98024+zipcode_cat_98027+zipcode_cat_98028+zipcode_cat_98029+zipcode_cat_98030+zipcode_cat_98031+zipcode_cat_98032+zipcode_cat_98033+zipcode_cat_98034+zipcode_cat_98038+zipcode_cat_98039+zipcode_cat_98040+zipcode_cat_98042+zipcode_cat_98045+zipcode_cat_98052+zipcode_cat_98053+zipcode_cat_98055+zipcode_cat_98056+zipcode_cat_98058+zipcode_cat_98059+zipcode_cat_98065+zipcode_cat_98070+zipcode_cat_98072+zipcode_cat_98074+zipcode_cat_98075+zipcode_cat_98077+zipcode_cat_98092+zipcode_cat_98102+zipcode_cat_98103+zipcode_cat_98105+zipcode_cat_98106+zipcode_cat_98107+zipcode_cat_98108+zipcode_cat_98109+zipcode_cat_98112+zipcode_cat_98115+zipcode_cat_98116+zipcode_cat_98117+zipcode_cat_98118+zipcode_cat_98119+zipcode_cat_98122+zipcode_cat_98125+zipcode_cat_98126+zipcode_cat_98133+zipcode_cat_98136+zipcode_cat_98144+zipcode_cat_98146+zipcode_cat_98148+zipcode_cat_98155+zipcode_cat_98166+zipcode_cat_98168+zipcode_cat_98177+zipcode_cat_98178+zipcode_cat_98188+zipcode_cat_98198+zipcode_cat_98199

# Checking for Overfitting

We split the data set by to 80/20 rule. We did multiple regression on the training data set, then we use this predictor to compute the mean square error in the testing data set, and compare it with the MSE from the training dataset.

The MSE of the testing data set is 14.071, and the MSE of the training data set is 14.072, thus there is no overfitting, and regularization is not needed.