Emmanuel Cortez

Dr. Garcia

ADTA 5410.16W – Frisco

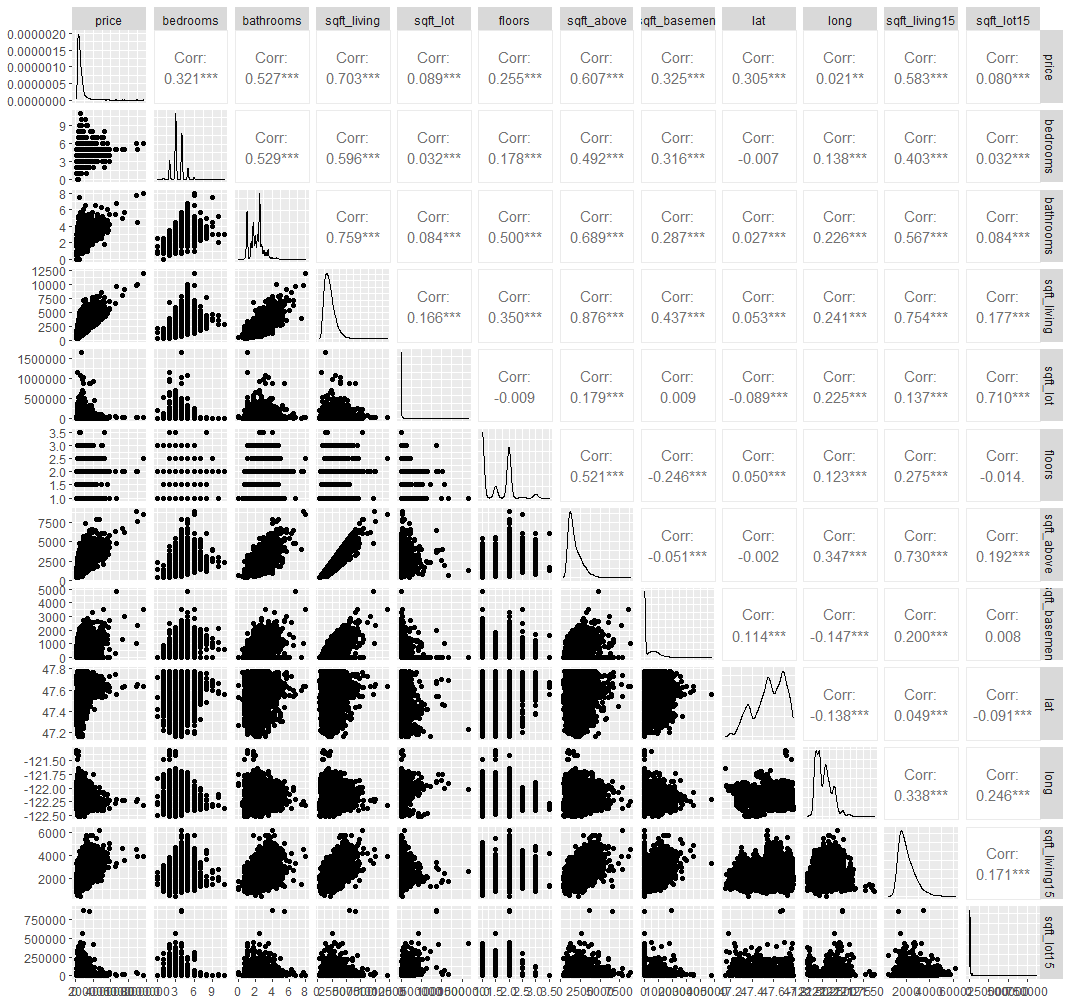
14 February 2021

Multiple Linear Regression for Price of King County Homes

**Tasks**

1. Fitting multiple linear regression model in training data set (King)
2. Base Model
3. EDA

While exploring the training set, King, all the data was imported as numeric and integer values. For proper analysis, I applied data type transformations for variables like view and condition to be ordinal and zipcode and waterfront to be factors. We also see that several continuous variables have a positively skewed distribution, however, this is to be expected when looking at home data as most data points have lower values when considering number of bathrooms and even home price. Through the alias function, I identified two variables that caused an error that read “there are aliased coefficients in the model” while trying to obtain variance inflation factors (VIF). The two culprits were sqft\_living and sqft\_above. The alias function confirms that both variables are linearly dependent which is indicative of multicollinearity, but we could see a preview of that in our correlation matrix. We will lean on our stepwise methods to reduce the variables that will give us the best model.



1. Developing Base Model

The base model contains all variables except ID, zipcode, sqft\_above, sqft\_living, sqft\_basement, yr\_built, yr\_renovated, lat and long. This left the base model to 11 total predictor variables. I used forward stepwise regression and all subsets regression to reduce the base model following data preparation and exploratory data analysis. While there were no issues with creating the functions in RStudio, it appeared that for both the forward stepwise and all subsets models indicated the highest number of terms possible as the best subset per Residual Sum of Squares, Adjusted R-square, Mallow’s Cp Statistic, and Bayesian Information Criterion. It recommended a model with 12 variables. This may be a result of violation of assumptions in some capacity, but I proceeded forward in creating my outputs with the following model:

twelve.lm <- lm(price ~ grade + lat + view + yr\_built + waterfront + bathrooms + condition + bedrooms + long + renovated + floors +

sqft\_lot, data=King)

my\_pred <- predict(twelve.lm, newdata=KingTest)

1. I initiated a best subsets regression to fit the optimal set variables to the model having a max of variables as supported by the forward stepwise regression.
2. I used two cross-validation methods to compare prediction error within the training set. First method used was split-sample cross-validation (SSCV). Second method used was 10 k-fold cross-validation (kFCV).
3. Predicted selling price of homes in test data set (KingTest)