My Process:

* I first set the working directory, and loaded in the first data set which I saved as “houseprice.csv”
* I looked at the data types. What stuck out to me was how Zipcode ( a categorical variable) was being coded as an integer
* Some columns I felt could be better represented. I felt that the Year of the home sale itself was not super relevant, as all the homes were being sold in either 2014 or 2015. So I thought that the age of the house could be more valuable. I created a variable called Age\_of\_Home that took the Year the house was being sold, and subtracted the year the house was built. I then took the variable called “Age\_of\_Home” and categorized it as an older, or newer house. Any house with an age over 25 years I categorized as “old” for my dummy variable
* The data provided us with the sqft of basements. I know when you get your house valued, that they do not usually count the square footage of the basement. So I thought it could be more helpful to create a dummy variable that says whether the house has a basement or if it doesn’t.
* I wanted to see whether the geographic location within King County affected the home price. So I set 47.590627, -122.253413 (the tip of Mercer Island) as my point of reference for computing geographic location. I tried to consider that a lot of King County remains unincorporated, so I made the point of reference central to where most residences are at (not the exact CenterPoint of King County). Using the lat and long I created two new dummy variables. One called “East Side” using the long, and one called “North”. I left out Westside and South to avoid multi-collinearity
* The variable grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design. I broke it into dummy variables: avg\_construction\_design, 7-10, poor\_construction\_design 5-6, and unconstructed\_construction\_design 1-3. I left out creating a variable for excellent construction design (which would be 11-13) to avoid multi-collinearity.
* I decided to break view into dummy variables as well. Bad view had ratings 0,1. Acceptable view 2,3. I left out excellent view(4) to avoid multi-collinearity
* I created a variable called sqft\_yard. Since the lot incorporates the size the home, I wanted to know how big each home’s yard is. So I did sqft\_lot – sqft\_above to get the square footage of the yard
* I created a variable called “renovated”. This helped solve the problem that unrenovated houses were given a 0 value, and renovated houses were documented in the year they were renovated 2015, for example. So if a house was renovated, it will have a 1 value. If not, 0.

Before removing for outliers, I checked the data and it had a total of 12,093 rows, and a total of 37 variables

* When removing outliers I used a statistical method that would take each column in the data and rcreate a boxplot with the outliers. I saved the outliers to a variable called outliers. I then saved my data frame as x. I had it remove all outliers in mydata frame x. Then I resaved the data frame again as mydata, which excluded all outliers. I removed outliers for bedrooms (left with 11,794 rows), bathrooms(left with 11,550 rows), sqft\_living (left with 11,346), sqft\_lot(10,157), sqft\_above (left with 9,912 rows),sq\_basement (left with 9,750 rows), sqft\_living15( 9,627 rows), and sqft\_lot15 (9,390 rows).
* There didn’t seem to be outliers for floors, year built.
* We didn’t remove outliers for grade, condition, (since they are on such a small scale). We also did not remove outliers for the dummy variables we just created since they are coded only as 0,1.
* We decided to only remove outliers for things specific to an average house in King County. For example, a home must have a bathroom and homes without bathrooms were considered outliers, and therefore removed from the data set.

Part 2 Training and Validation Split

* Set the seed using 666
* Created the split 60% training (5,634 rows), 40% validation (3,756 rows)
* We built the model on the training set.
* We tried the three methods for reducing the predictors, and decided to use BACKWARD elimination for the basis of our model. That is because it had the highest adjusted r-squared value
* When testing the different elimination methods on various variables throughout the process, backward elimination always provided the highest accuracy. Forward, and both stepwise regression were only slightly less accurate, and yielded about the same adjusted r-squared each time