**Question #2**

Part 1

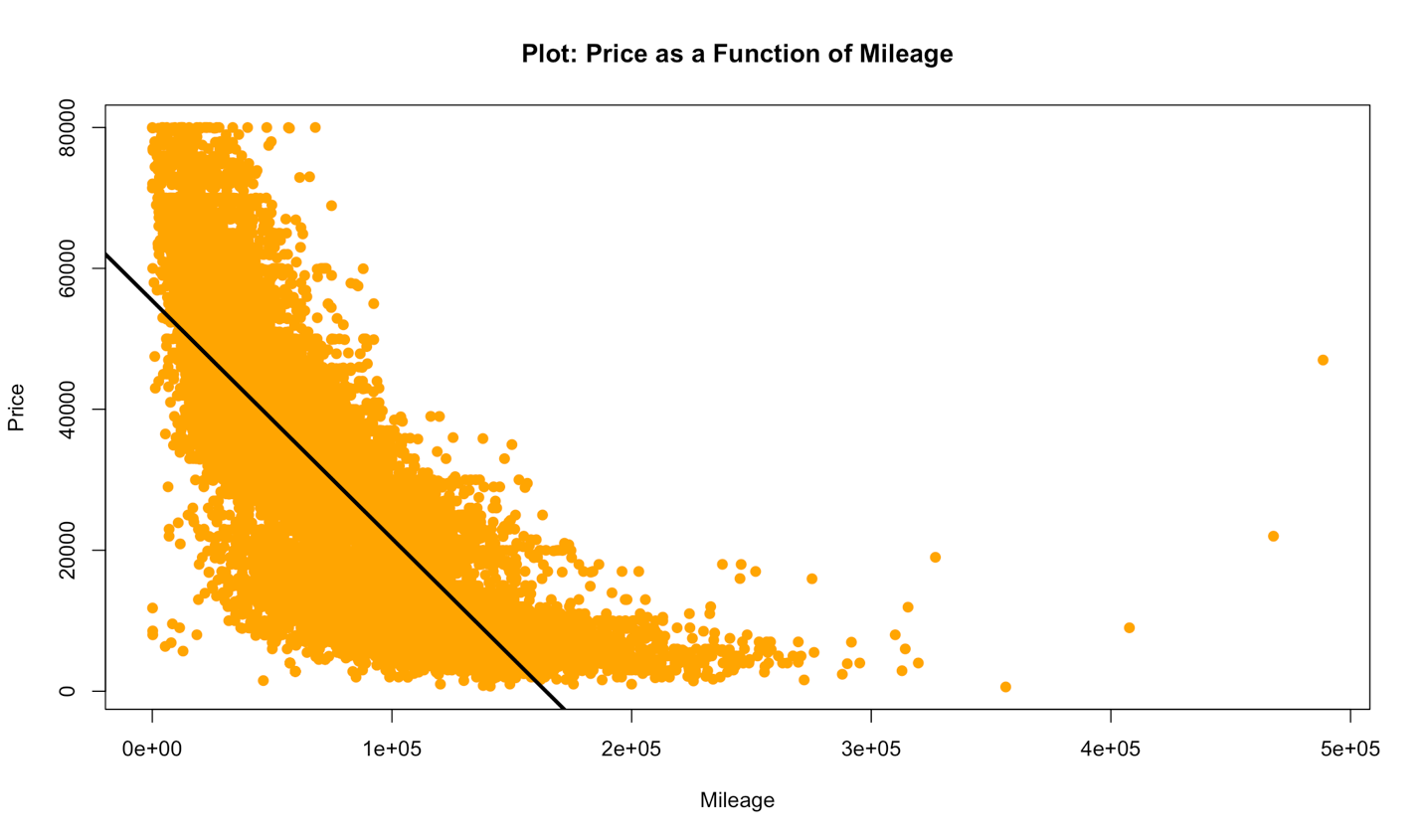
The cars data can be used to understand the relationships between variables such as prices and mileage. Prices of the cars as a function of mileage, year of their make, how many owners each car has had and the sound system can be predicted using this data. How the prices vary by region can also be explored using this data.

Part 2

Splitting the data into two samples:

* Training set has 15047 observations
* Testing set has 5016 observations

Part 3



Scatter plot of Price vs. Mileage with best linear regression fit in black.

Part 4

Inserting the code here b/c we couldn’t figure out how to fit the polynomial function to the line.

*library(boot)*

*n.folds = 2*

*poly<-2*

*n = ntrain*

*cvmean = rep(0,length(poly))*

*for (j in n.folds){*

*testIndex <- j*

*trainingTemp <- cars\_data[-testIndex,]*

*testTemp <- cars\_data[testIndex,]*

*yVar <- trainingTemp$price*

*xVar <- trainingTemp$mileage*

*for (i in poly){*

*model <- lm(price ~ poly(mileage,i), data=trainingTemp)*

*model\_prediction <- predict(model,testTemp)*

*residualFold <- (testTemp$price-model\_prediction)^2*

*msError <- sum(residualFold)*

*cvmean[i] <- cvmean[i] + msError*

*}*

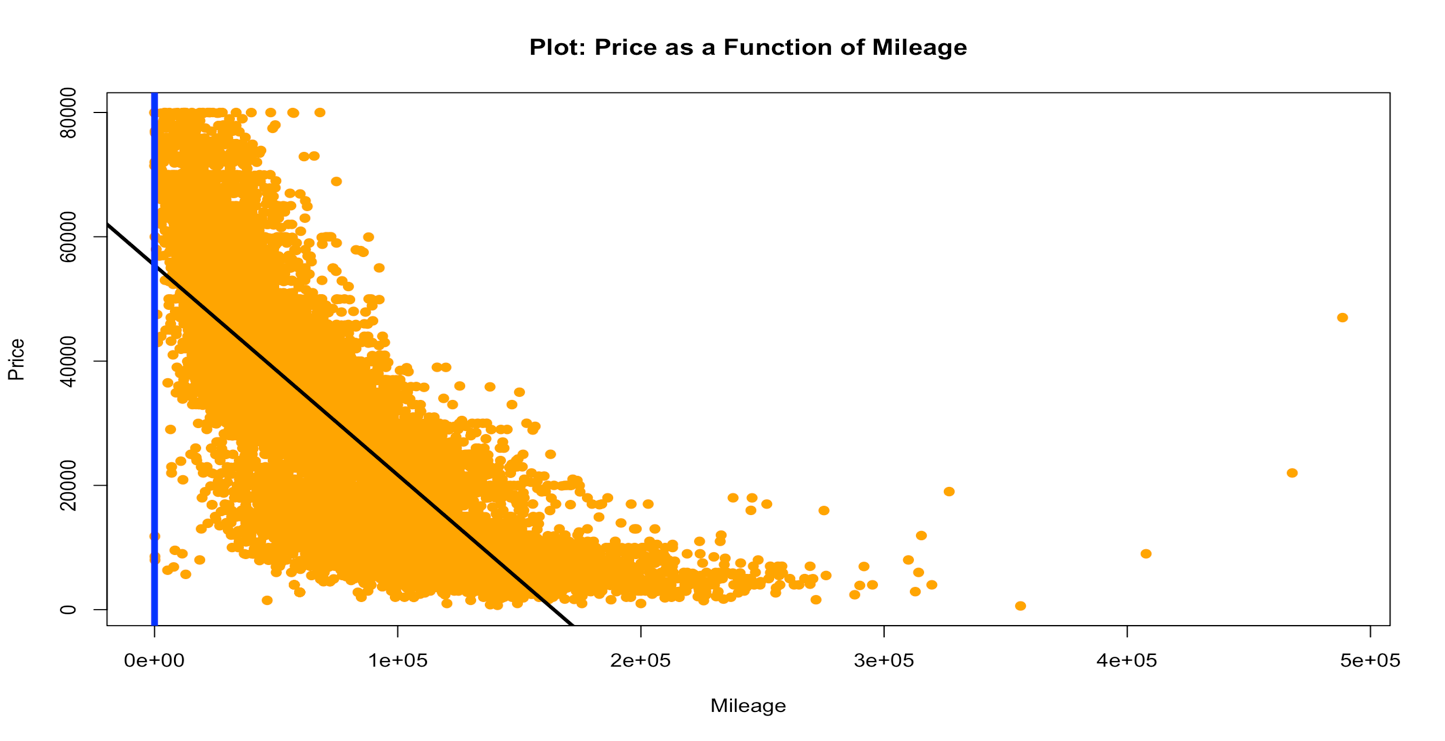
*}*

*plot(cars\_data$mileage,cars\_data$price, main="Plot: Price as a Function of Mileage", ylab = "Price", xlab = "Mileage", col="orange",pch=19)*

*abline(lm(cars\_data$price~cars\_data$mileage),col="black",lwd=3)*

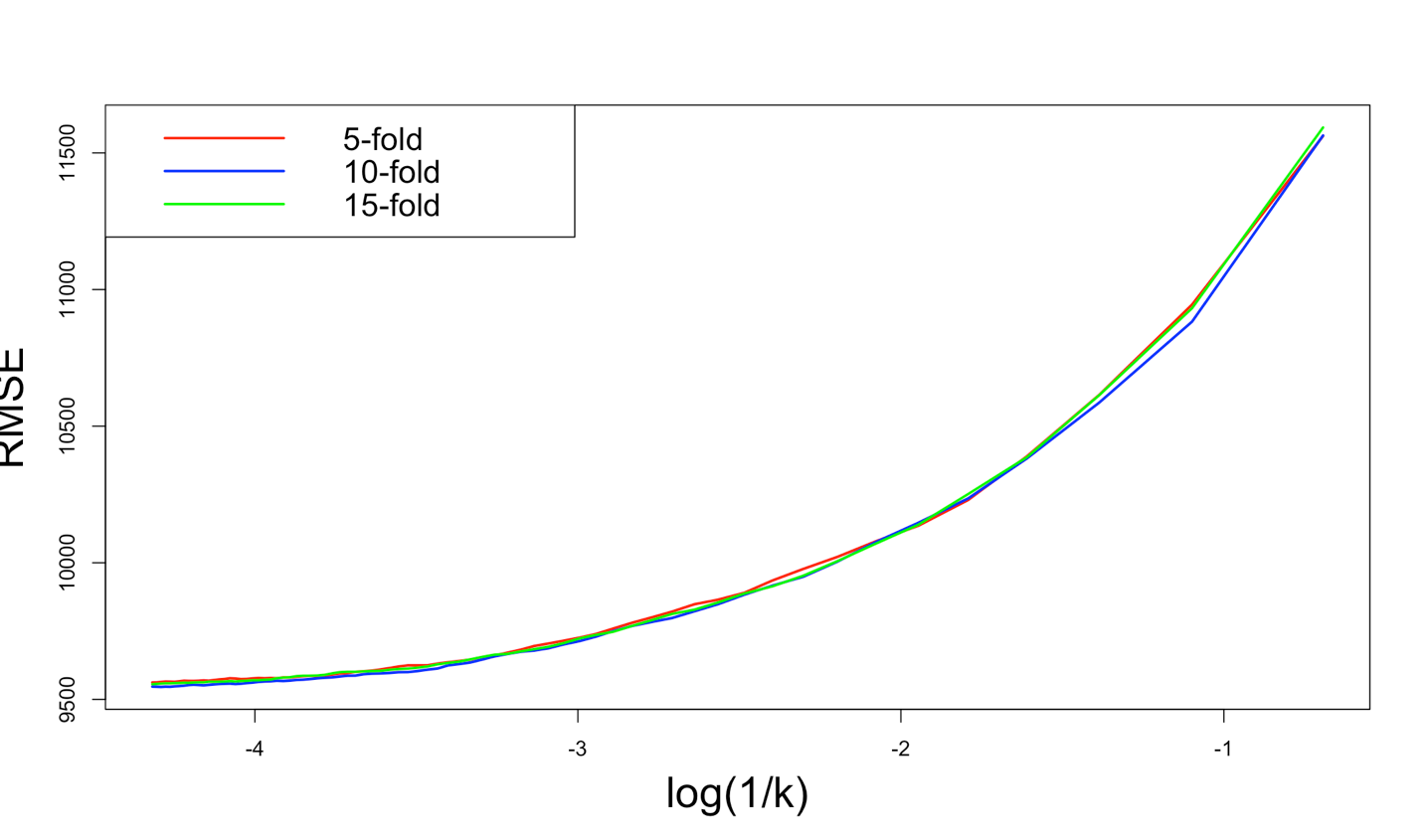
*abline(lm(cars\_data$price~poly(cars\_data$mileage,1)), col="blue", lwd=5)*

*summary(lm(cars\_data$price~poly(cars\_data$mileage,11)))*

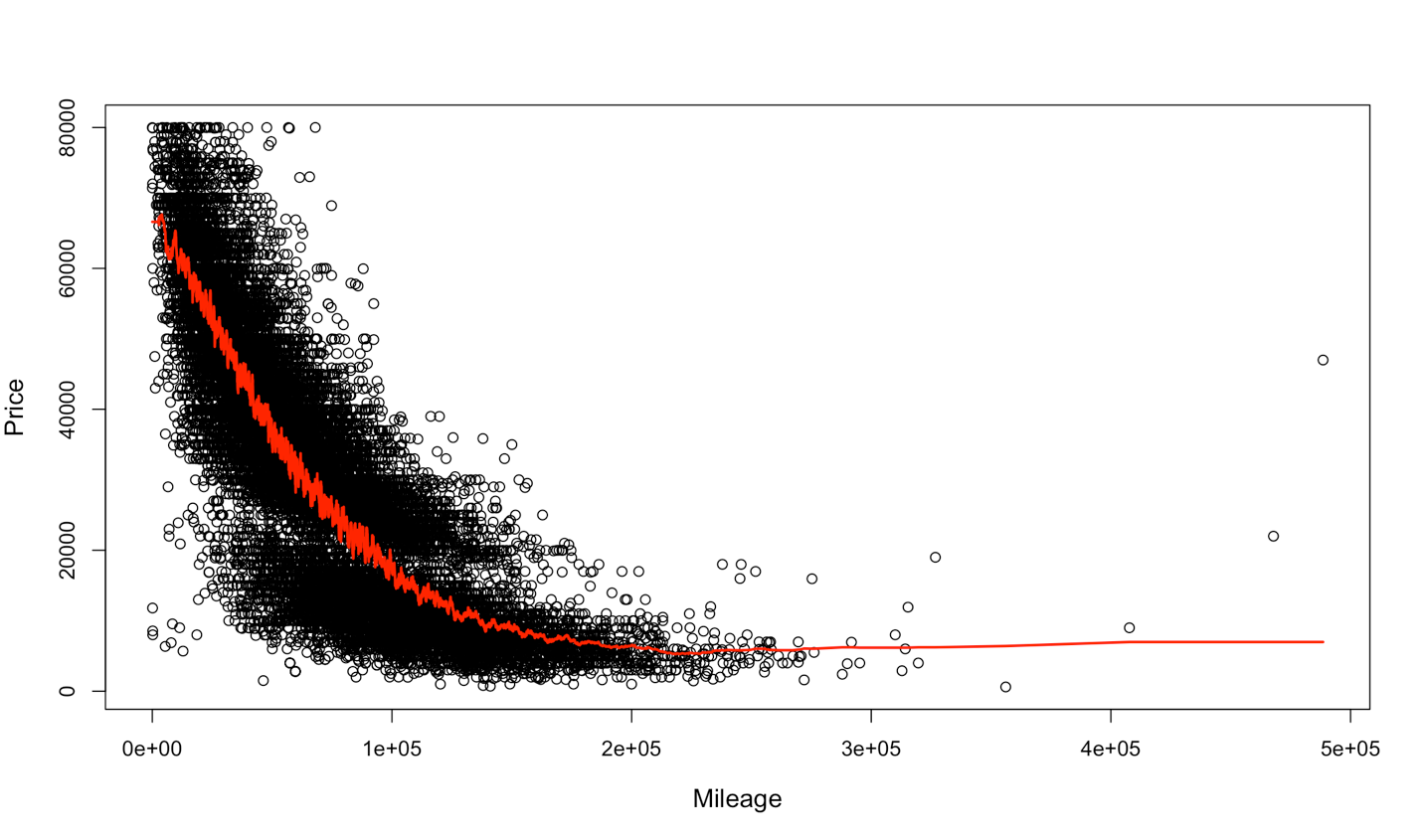


The blue line isn’t correct but the polynomial degree with the least mean squared error is 11

Part 5 & 6 Using k-nn

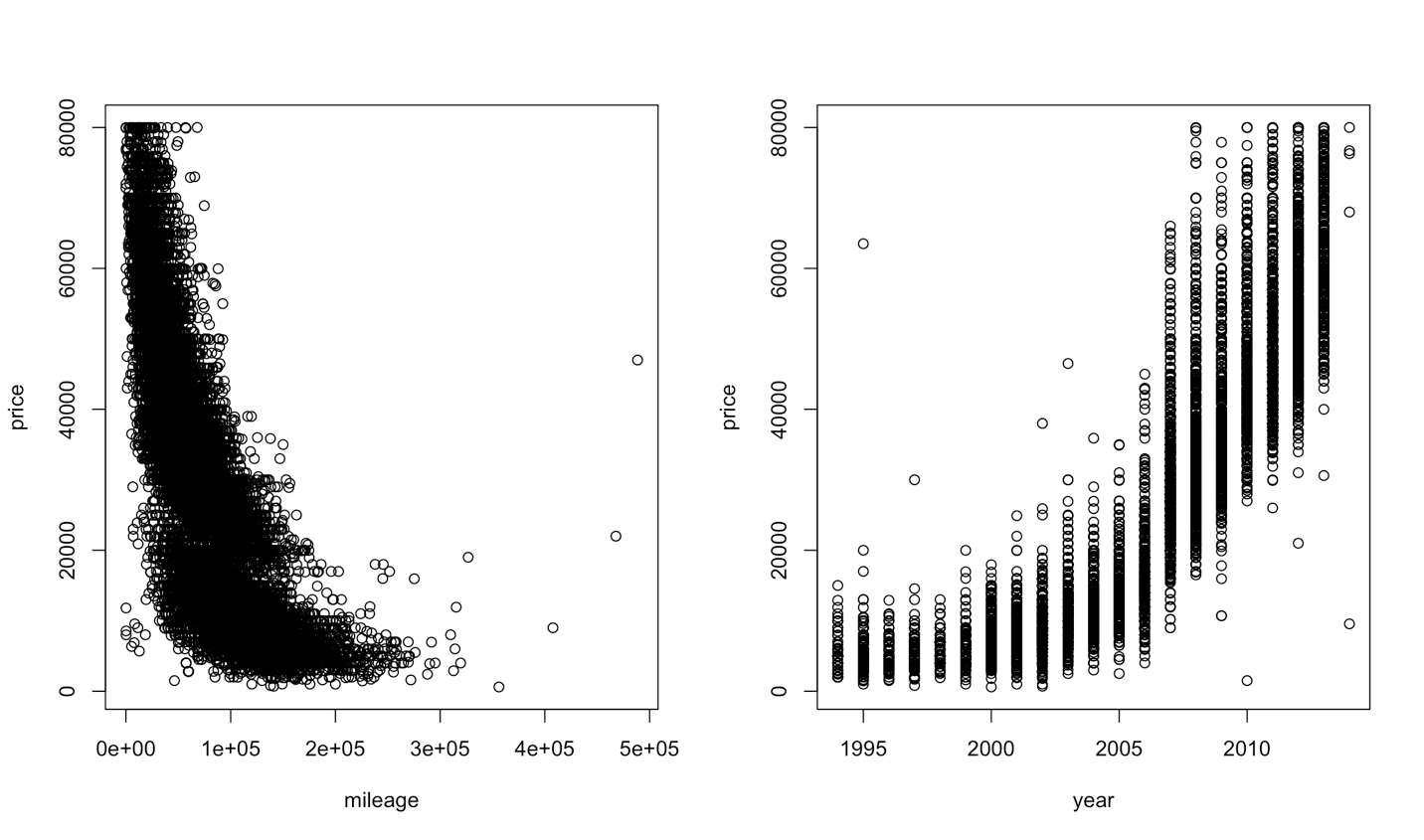


We ran the cross validations across three different folds and looked for the RMSE against log(1/k)

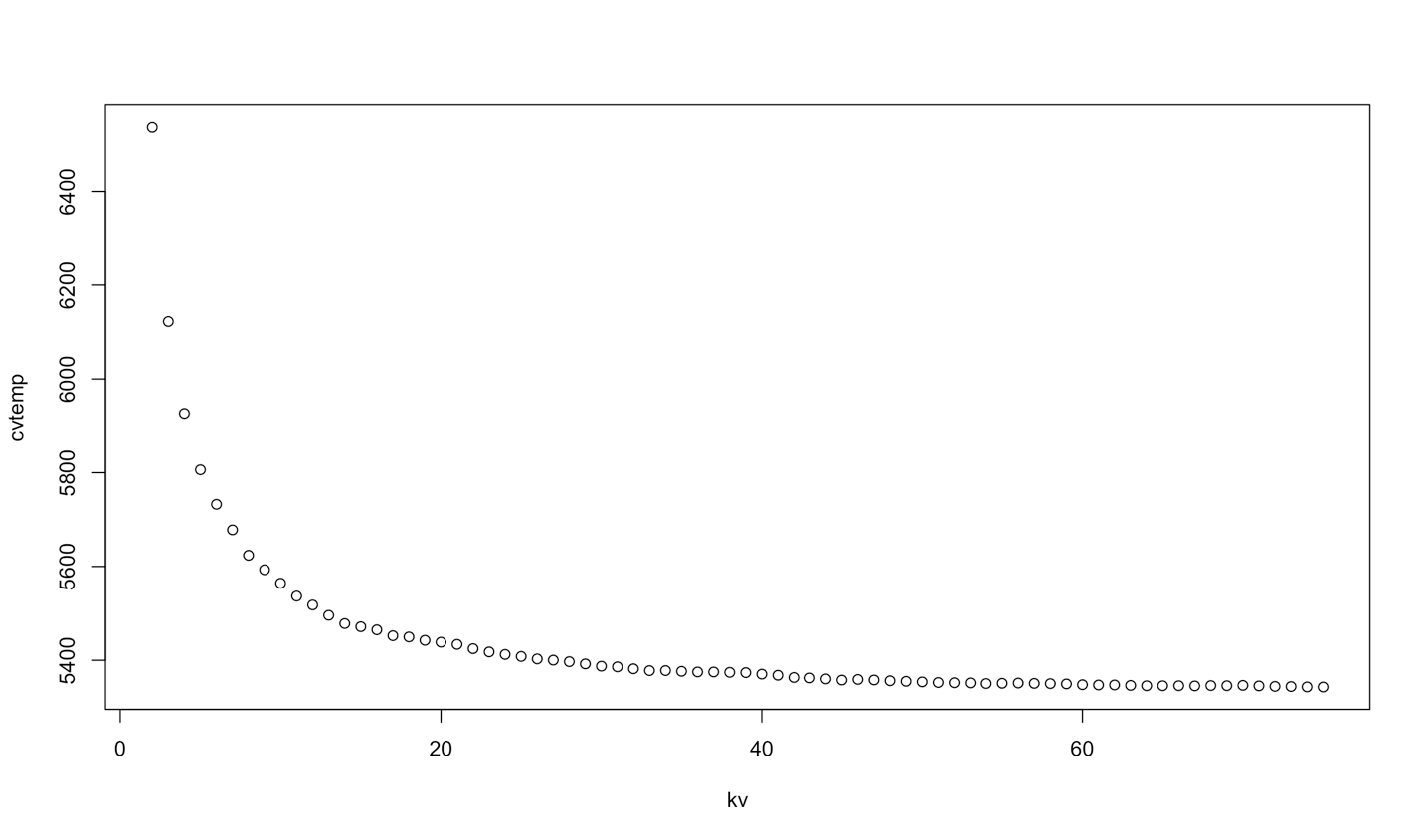


We fit the line using k = 75

*Part 6 using k-nn*



Two scatter plots plotting different x-variables to predict price



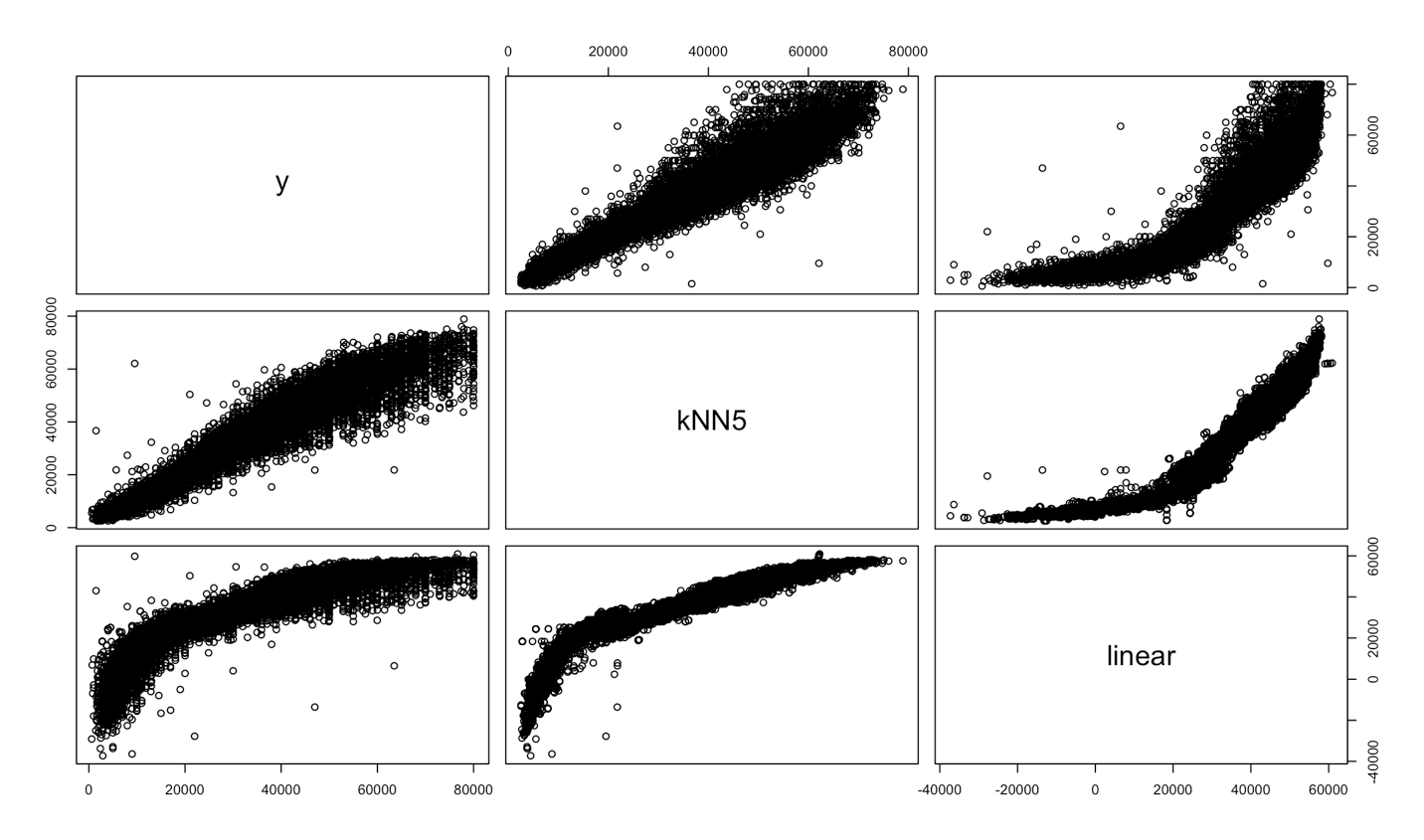
We ran CV temp across 10 folds through 75 k factors

y kNN5 linear

y 1.0000000 0.9658295 0.9115898

kNN5 0.9658295 1.0000000 0.9438781

linear 0.9115898 0.9438781 1.0000000



We refit the data using k=5

knn predicted value: 34358.6

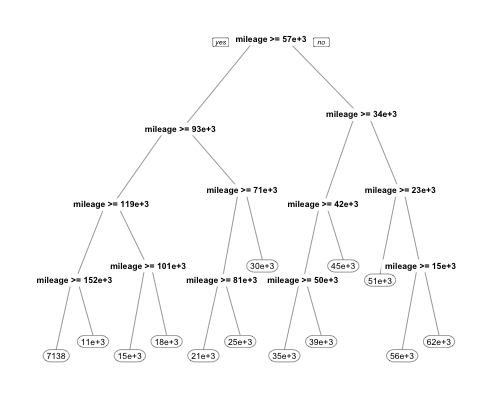
We predicted that the price of the car is $34,358.6 with a mileage of 50,000 and was made in the year 2008

The linear model predicts that the price is $36,328.48

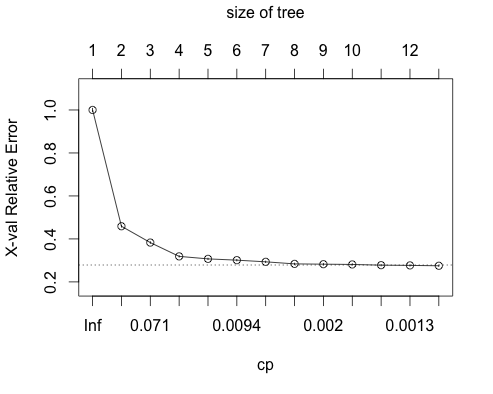
After we adjust the scaling, the linear model predicts a price of $43,874.53 if we use mileage and year as explanatory variables

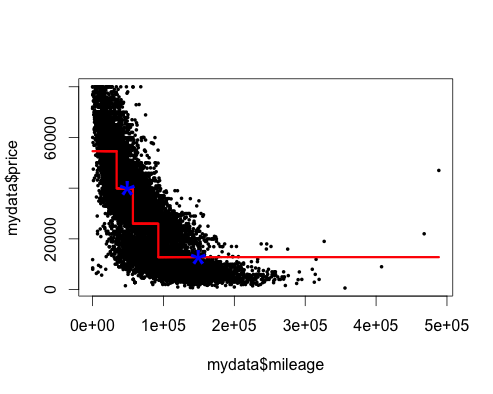
Parts 5 and 6 using Regression Trees

Below is our initial tree based on just Price and Mileage

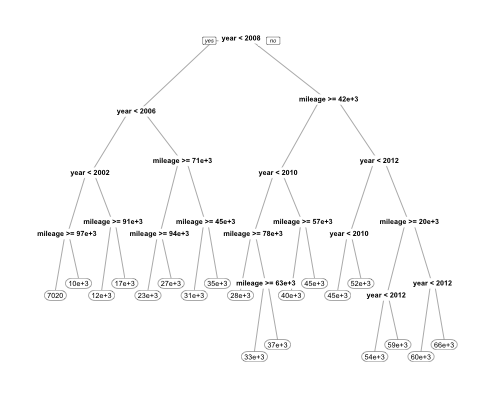


Our initial pre-pruned tree. The below cross validation chart shows us that a cp of .001 is optimal providing us a tree size of 13. We would use the k-nn method due to the MSE.

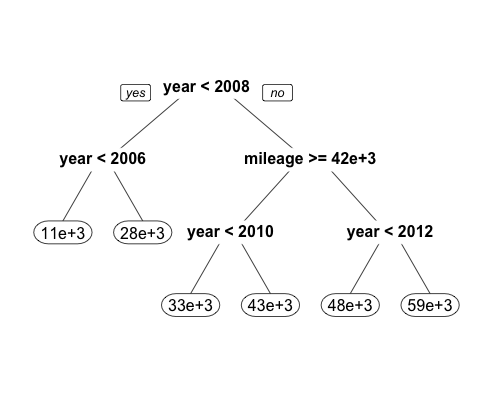




This shows our training fitted values in a step function. We can use this to predict values given either a set price or mileage value. The blue values represent our predicted price given the car has 50,000 miles or 150,000 miles for respective predicted prices of roughly $40,000 and $17,000 according to the chart.



The above tree represents a pre-pruned tree based on both mileage and year. The tree below is the pruned version.



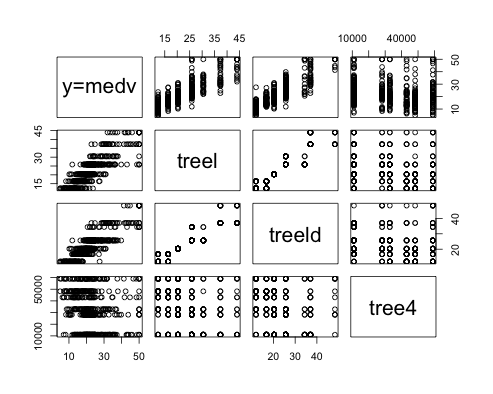
y=price treel treeld tree4

y=price 1.0000000 0.8338431 0.8560183 -0.1519301

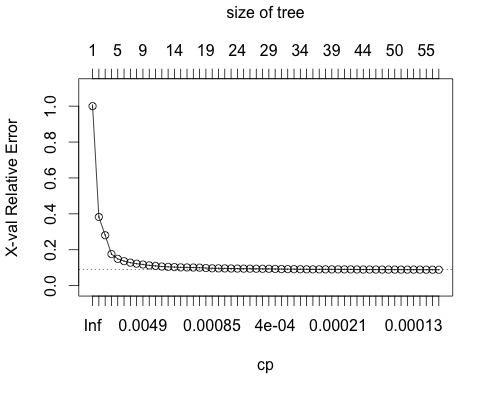
treel 0.8338431 1.0000000 0.9364702 -0.2340959

treeld 0.8560183 0.9364702 1.0000000 -0.2239759

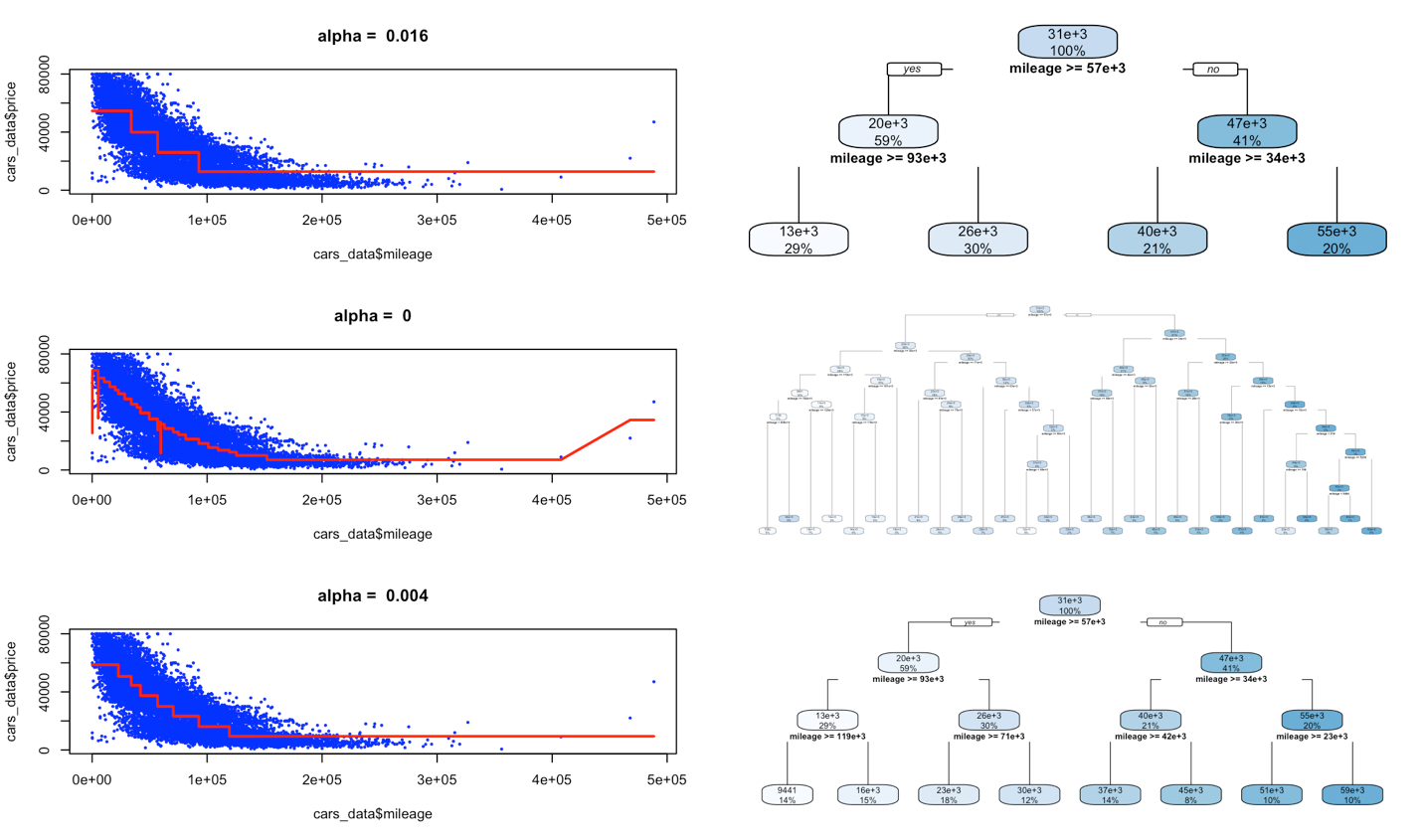
tree4 -0.1519301 -0.2340959 -0.2239759 1.0000000



The above correlation matrix and subsequent charts describe the fits of our three trees, the first is our initial fit tree, the second is our tree fit with just price and mileage and the third includes price, mileage and year.



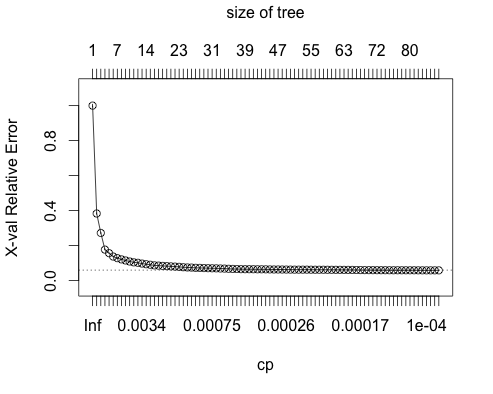
Plotting cross validation we determine that our best CP (the lowest value below the dotted line) is 1e-04 and gives us a size of tree of 58.



The top chart represents a tree that has high bias, while the bottom charts show a tree with low bias. The middle tree balances the bias and variance to provide a better tree. Here our best tree is has more nodes than the other two options. We can see that our tree now with a size of 58 is superior to our original tree with a size of 13 from a predictability standpoint with the addition of year as a predictor.

Question 7

Cross Validation Chart



Again our optimal cp is the lowest point below the dotted line and is 0.0001006511 giving a tree size of 86. This tells us that we are best off using year and mileage as predictors of price as the rest of the factors have minimal predictive power. We have 29 splits (nsplit+1) and minimum xerror of .2733849, which is the minimum xerror.

CP nsplit rel error xerror xstd

25 0.0001678106 28 0.2657576 0.2733849 0.003679587