KNN Problem 1.1

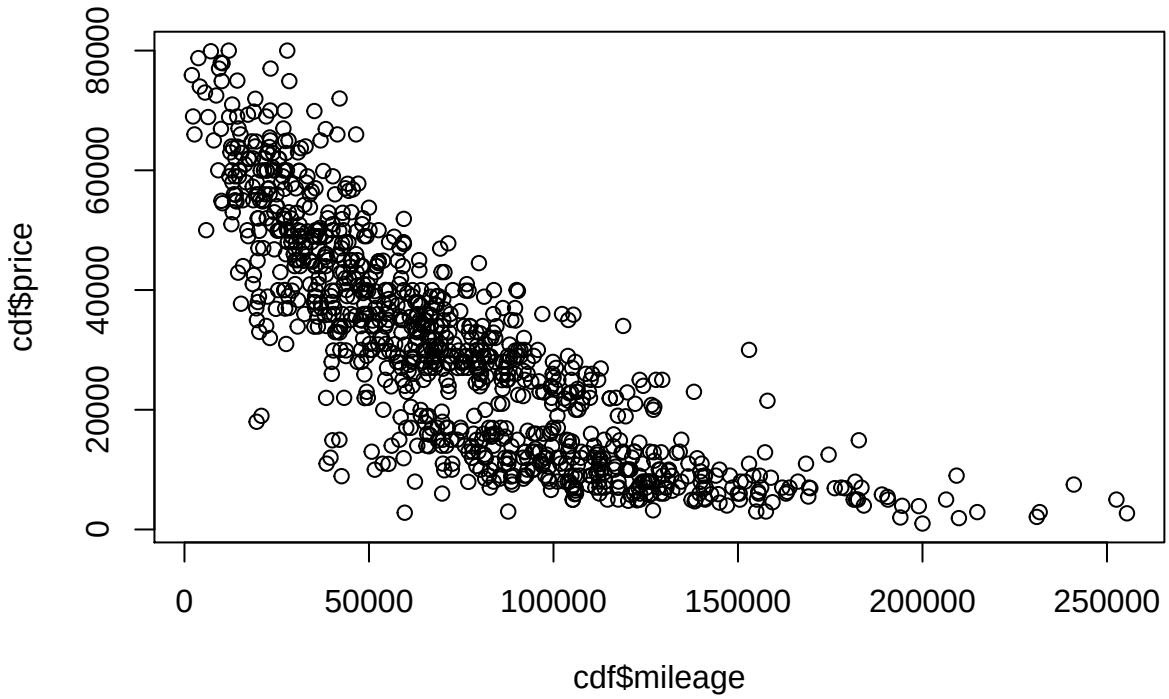
Rob

January 17, 2017

Read in the Data and have a look

Use **read.csv** to read in the data from the webpage.

```
cdat = read.csv("susedcars.csv")
cdf = cdat[,c(1,4)] #c1 is y=price c4 is mileage
plot(cdf$mileage,cdf$price)
```




```
print(dim(cdf))
```

```
## [1] 1000    2
```

There certainly is a relationship and it makes sense.
The more mileage on the car, the less it sells for.
Also, it is a non-linear relationship.

Try a linear fit and KNN for Various values of k

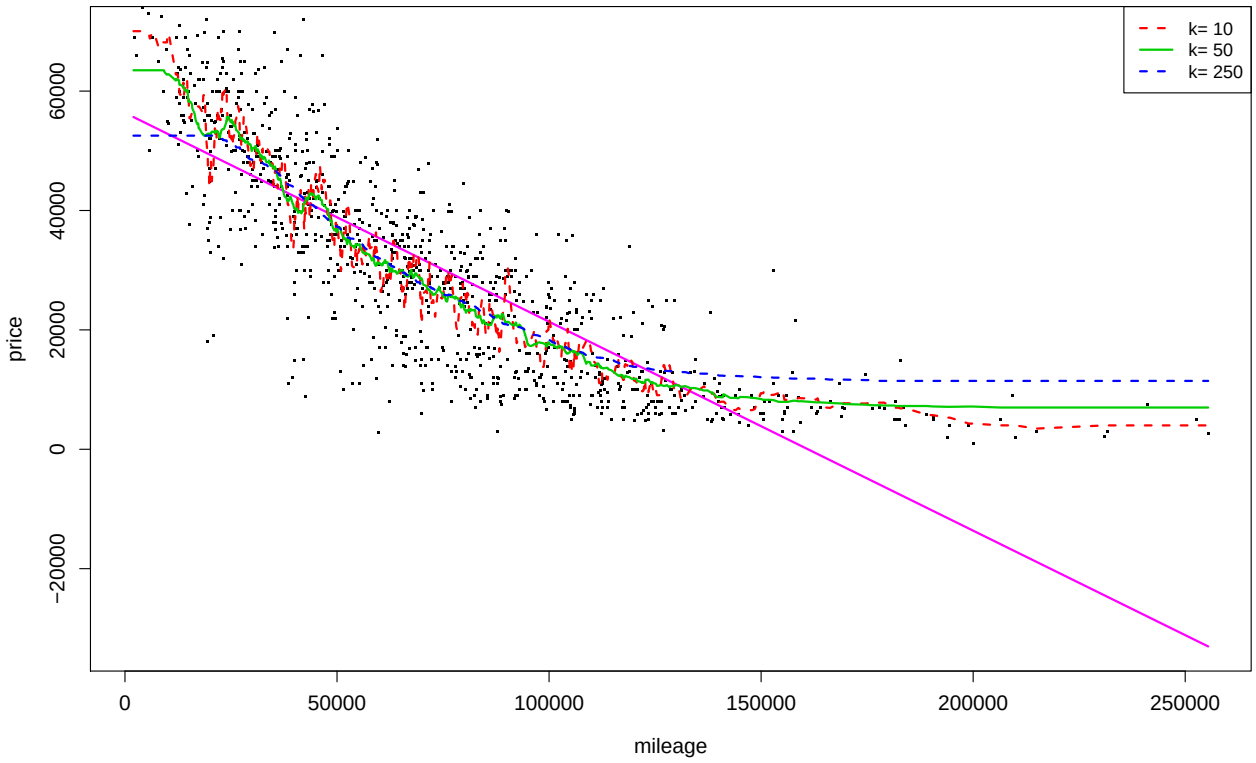
```
lmf = lm(price~mileage,cdf)
print(summary(lmf))
```

```
##
## Call:
## lm(formula = price ~ mileage, data = cdf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32670  -7063    239   6293  37024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.636e+04  6.706e+02   84.04   <2e-16 ***
## mileage     -3.500e-01  7.870e-03  -44.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10670 on 998 degrees of freedom
## Multiple R-squared:  0.6646, Adjusted R-squared:  0.6643
## F-statistic:  1978 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
test = data.frame(mileage=sort(cdf$mileage))

lmfit = predict(lmf,test)

dokv = c(10,50,250)
nk = length(dokv)
n=nrow(cdf)
knnf = matrix(0.0,n,nk)
library(kknn)
for(i in 1:nk) {
   kf = kknn(price~mileage,cdf,test,k=dokv[i],kernel = "rectangular")
   knnf[,i]=kf$fitted.values
}
plot(range(cdf$mileage),range(cbind(lmfit,knnf)),type="n",xlab="mileage",ylab="price",
              cex.axis=1.2,cex.lab=1.2)
points(cdf$mileage,cdf$price,pch=".",cex=2.2)
lines(test$mileage,lmfit,col="magenta",lwd=2)
ltyv = c(2,1,2)
for(i in 1:nk) {
   lines(test$mileage,knnf[,i],lty=ltyv[i],col=i+1,lwd=2)
}
legend("topright",legend=paste0("k= ",dokv),lty=ltyv,co=1+1:nk,lwd=2)
```



The linear fit is bad for small mileages and *really* bad for big mileages.

Predict at 100,000

Now let get prediction for the linear model and knn with $k = 50$ given $x = mileage = 100000$.

```
pdatf = data.frame(mileage=100000)
lmp = predict(lmf,pdatf)
knp = kknn(price~mileage,cdf,pdatf,k=50,kernel = "rectangular")$fitted.values[1]
cat("linear, knn:",lmp,knp,"\n")
```

```
## linear, knn: 21362.33 17761.14
```