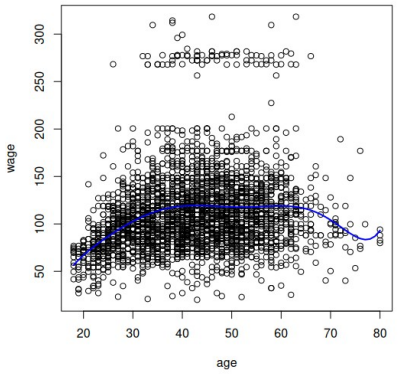
Question 6

a) First I split the Wage data into a testing and training set. I performed 10 different polynomial regressions, and found the MSE error for each of them. The model with the lowest MSE according to my calculations was model 7 with up to an X^7 term. I then ran an ANOVA test, which told me that the highest exponent model I should use is X^3, but for the plot I used X^7. After plotting, I got:

```
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    2998 5022216
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 **
4    2995 4771604  1      6070   3.8143  0.050909 .
5    2994 4770322  1      1283   0.8059  0.369398
6    2993 4766389  1      3932   2.4709  0.116074
7    2992 4763834  1      2555   1.6057  0.205199
8    2991 4763707  1       127   0.0796  0.777865
9    2990 4756703  1      7004   4.4014  0.035994 *
10   2989 4756701  1         3   0.0017  0.967529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

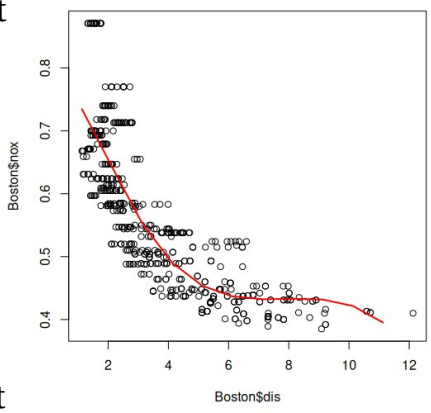
```
> print(which.min(errors))
[1] 7
```

b) This part hopefully worked, but not quite as well. I was able to do the cuts and run models, but my plotting would not work. Based on my models, the least CVM error came from the model with 10 cuts (I checked cutting from 2 to 10 times). This makes sense since the data was so populated in the 50-200 range for wage



Question 9

d) A spline model with DOF 4 means 1 knot, so I picked the knot right at the median of dis, meaning I would fit 2 different equations. My resulting fitted graph is:



e) I now chose to split into 4 knots, by the quartiles of dis. The more knots that I had, according to the summary of my fit were not significant, but I plotted them anyways.

f) I'm not sure how to do this without writing a bunch of for loops, and I've already been working on this homework for almost 2 hours. I'm not sure how to cross validate a number of knots, but It seems as if a lower number would be better, since 1 knot fit really well and 4 knots made further coefficients not significant.