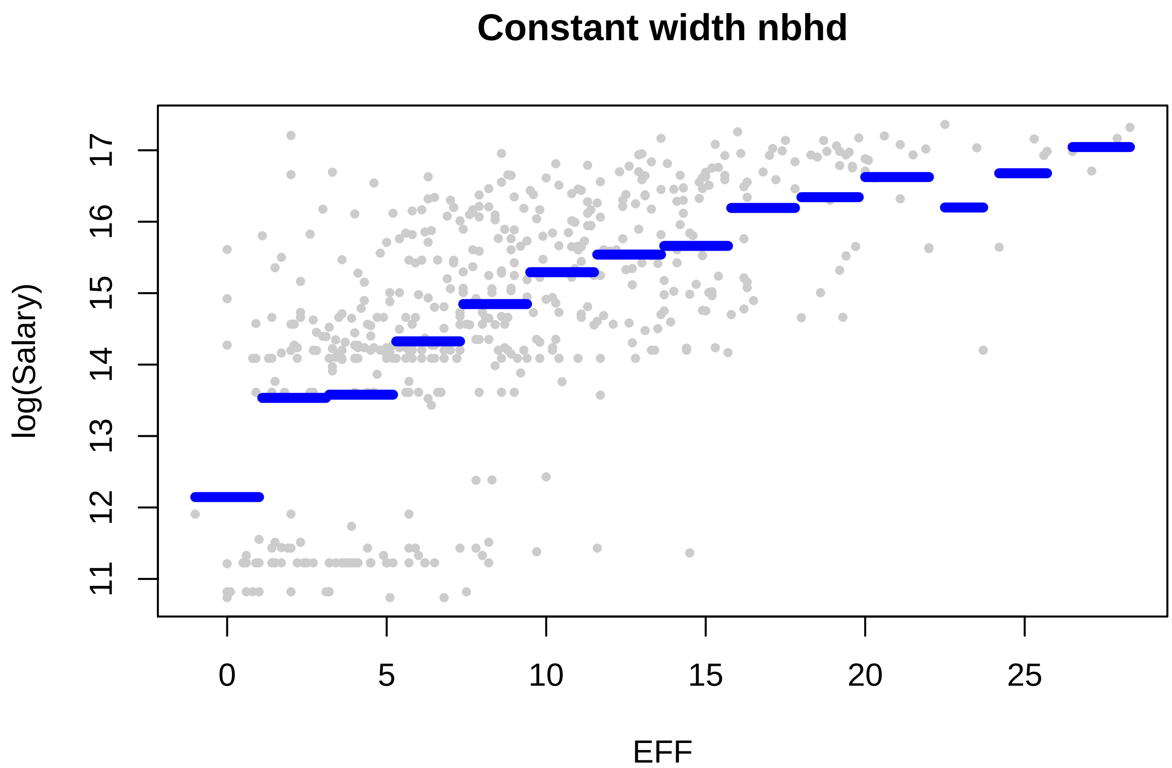
444 project code

Zizhou Wang

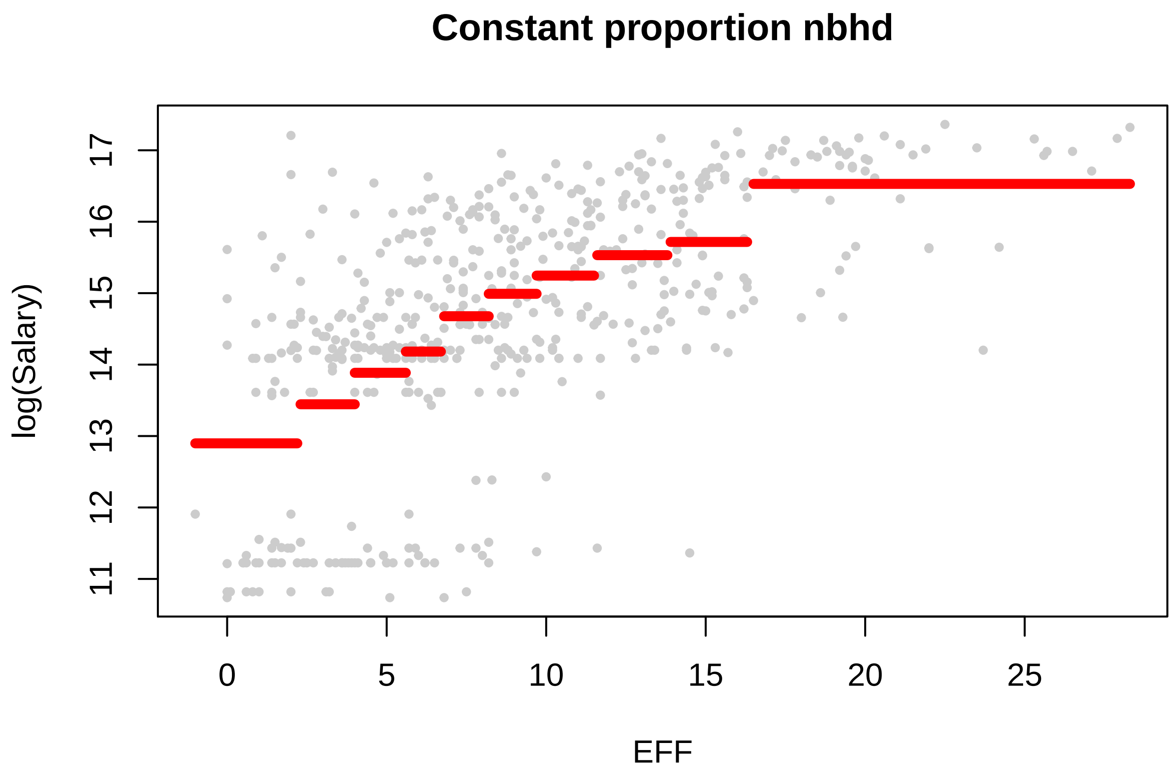
April 6, 2018

```
##   ID           Name   Salary Team AGE GP  W  L  MIN  PTS  FGM  FGA FGPER
## 1  1  Stephen Curry 34682550  GSW  30 51 41 10 32.0 26.4  8.4 16.9  49.5
## 2  2   LeBron James 33285709  CLE  33 76 46 30 37.1 27.6 10.6 19.4  54.7
## 3  3   Paul Millsap 31269231  DEN  33 32 17 15 29.3 14.8  5.4 11.2  48.2
## 4  4 Gordon Hayward 29727900  BOS  28  1  0  1  5.3  2.0  1.0  2.0  50.0
## 5  5  Blake Griffin 29512900  DET  29 58 28 30 33.8 21.3  7.5 17.0  43.8
## 6  6     Kyle Lowry 28703704  TOR  32 71 53 18 32.3 16.6  5.2 12.1  43.3
##   TPM TPA TPPER FTM FTA FTPER OREB DREB REB AST TOV STL BLK  PF   FP DD2
## 1 4.2 9.8  42.3 5.5 5.9  92.1  0.7  4.4 5.1 6.1 3.0 1.6 0.2 2.2 43.8   5
## 2 1.8 4.9  36.1 4.6 6.3  73.0  1.2  7.4 8.6 9.1 4.2 1.5 0.9 1.7 54.5  47
## 3 1.1 2.9  36.6 2.9 4.2  70.7  1.4  4.8 6.3 2.8 1.9 1.2 1.1 2.6 31.3   1
## 4 0.0 1.0   0.0 0.0 0.0   0.0  0.0  1.0 1.0 0.0 0.0 0.0 0.0 1.0  3.2   0
## 5 1.9 5.5  34.8 4.4 5.6  78.6  1.3  6.1 7.3 5.7 2.8 0.7 0.3 2.4 38.8  16
## 6 3.1 7.6  40.9 3.0 3.5  85.9  0.9  4.7 5.6 6.8 2.3 1.1 0.2 2.5 35.3  22
##   TD3 PLUSMINUS Position Country Draft.Round Draft.Number SGap        WR
## 1   0       9.5        G     USA           1            7    1 0.8039216
## 2  16       0.6        F     USA           1            1    1 0.6052632
## 3   0       2.3        F     USA           2           47    1 0.5312500
## 4   0       3.0        F     USA           1            9    1 0.0000000
## 5   3       1.1        F     USA           1            1    1 0.4827586
## 6   3       5.0        G     USA           1           24    1 0.7464789
##    PRA
## 1 22.5
## 2 28.3
## 3 16.0
## 4  2.0
## 5 20.6
## 6 19.8
```

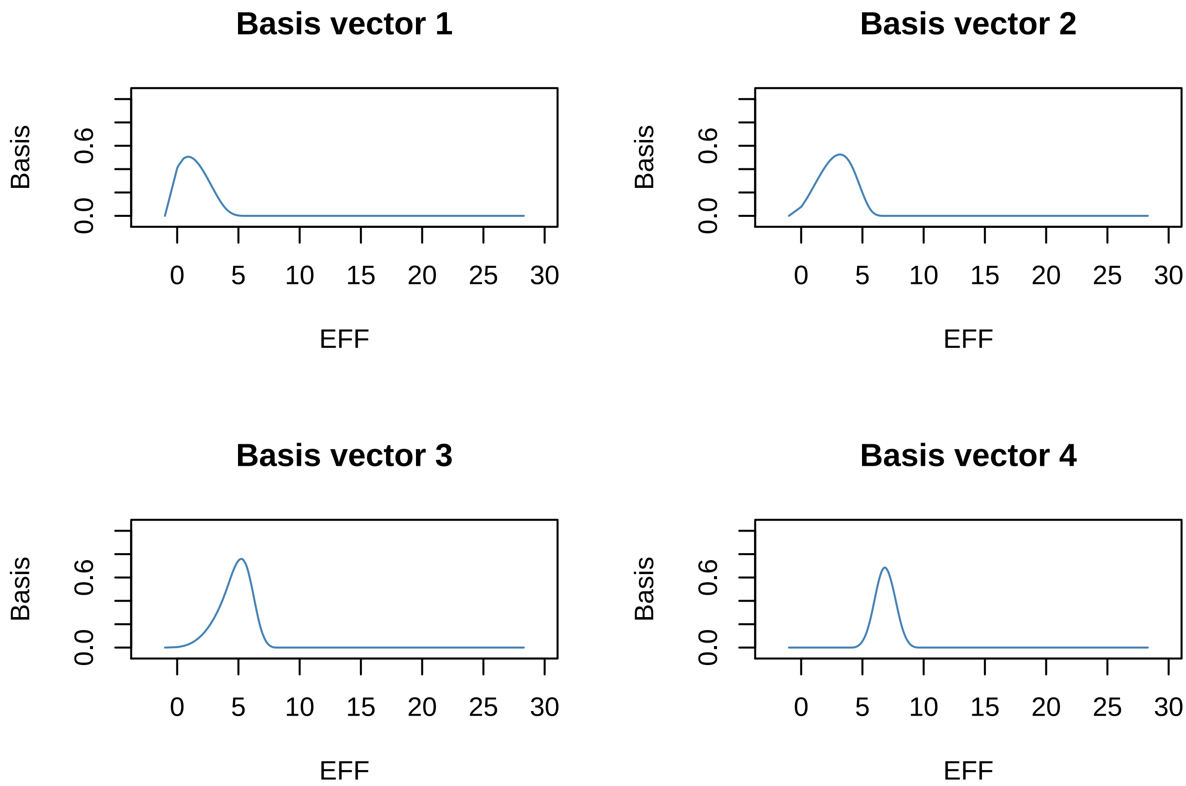
One of the most straight forward way to evaluate the performance of an NBA player is to look at his "Points per Game", "Assists per Game", and "Rebounds per Game", which are the 3 most popular statistics in NBA. We initially tried to find a relationship between the PRA(Points + Rebounds + Assists per game) and the Salary of an NBA player. However, we realized that it will almost always introduce a bias, because it does not tell us the full image of the player's ability. For example, Points are usually easier to get compared to Assists and Rebounds. When a player scores, they will either get two points or three points, potentially earning an extra Free Throw, which counts as one more point. When a player gets an Assist or a Rebound, the count only goes up by 1. Having 10 Rebounds or 10 Assists after a game is considered a good performance, but having 10 Points for a game is usually average. The PRA also introduces a heavier weight on the player's offensive ability than his defensive ability on the court, since Points, Assists, and Offensive Rebounds all happen at the front court. Therefore, we found a better way to determine the efficiency of an NBA player, which is to look at his EFF, calculated by EFF = PTS + REB + AST + STL + BLK - FGM - FTM - TOV, where all variates are averaged per game. The EFF takes Steal (STL), Block (BLK), Field Goal Missed (FGM), Free Throw Missed (FTM), and Turn Over (TOV) into account, which adds the defensive ability (STL and BLK) and inefficiency (FGM, FTM, TO) into the equation.

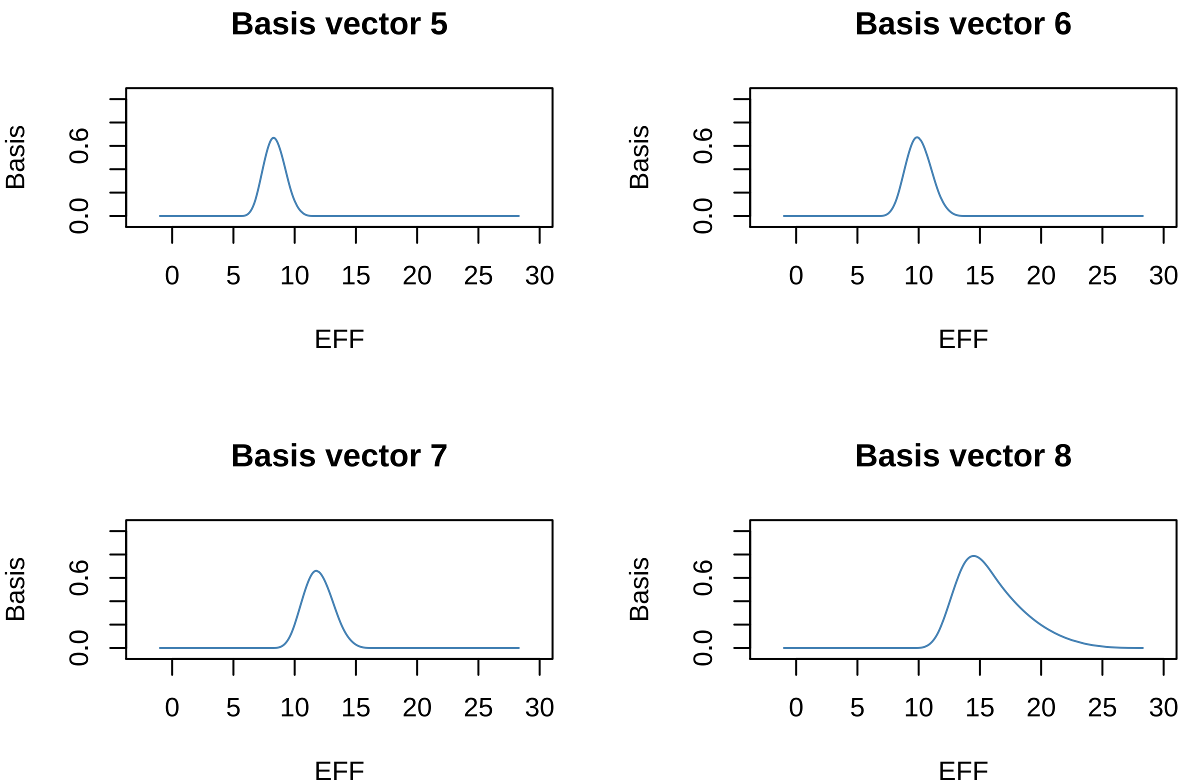


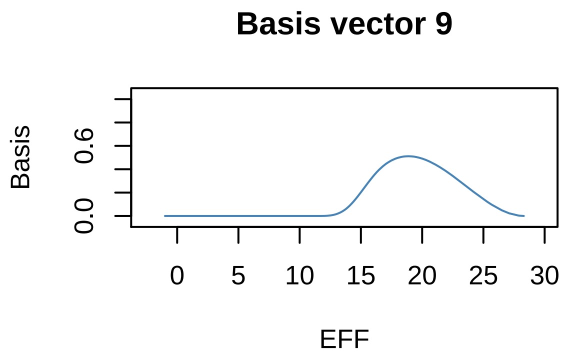
Constant width nbhd
log(Salary)
11
12
13
14
15
16
17
0
5
10
15
20
25
EFF

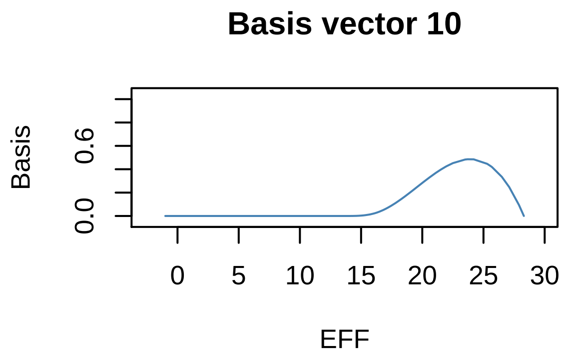


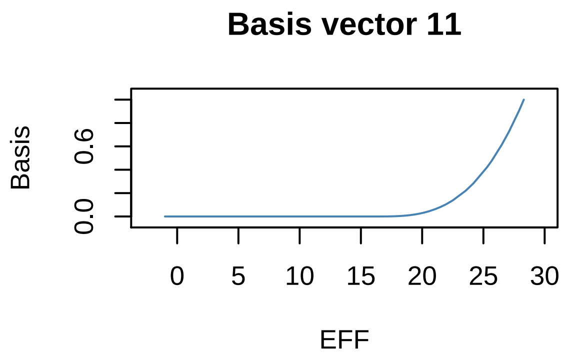
We first want to see what our data look like when EFF is plotted aginst log(Salary), even though our data look to be bimodel, we can still observe an increasing trend, according to the piece wise fitting using neighbourhood.


Basis vector 1
Basis vector 2
Basis vector 3
Basis vector 4
Basis
0.0
0.6
0
5
10
15
20
25
30
EFF


Basis vector 5
Basis vector 6
Basis vector 7
Basis vector 8
Basis
0.0
0.6
0
5
10
15
20
25
30
EFF

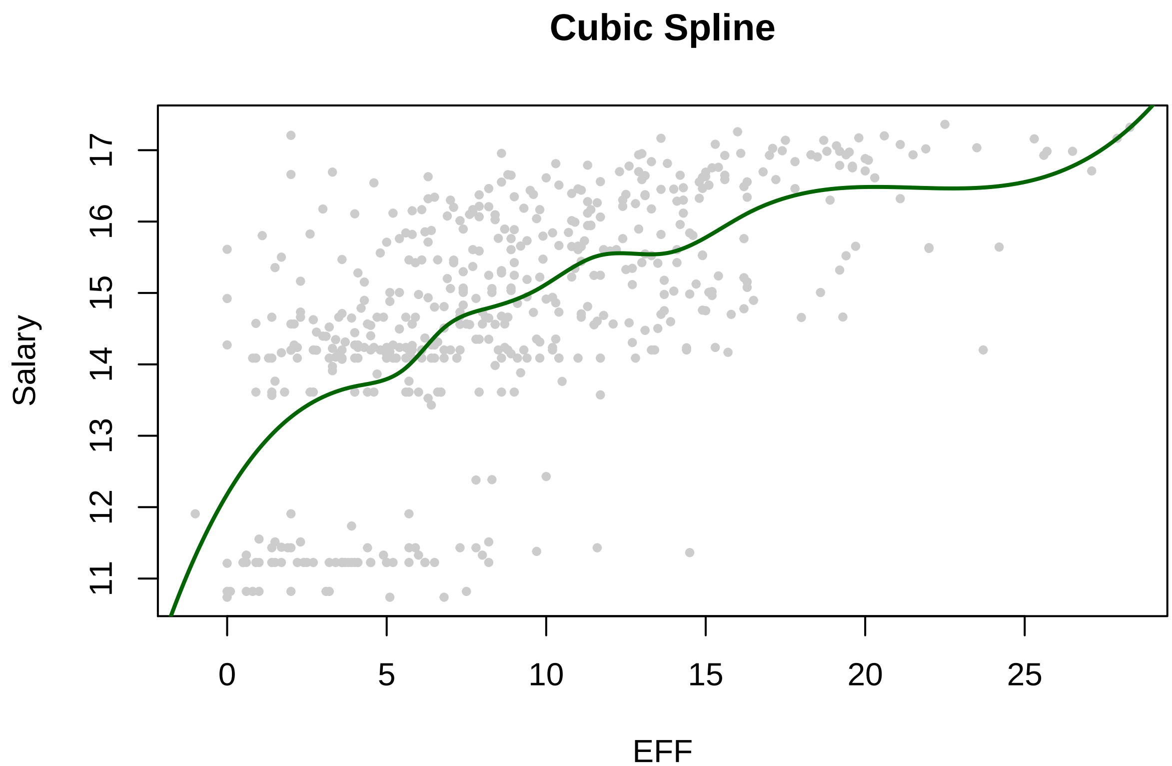






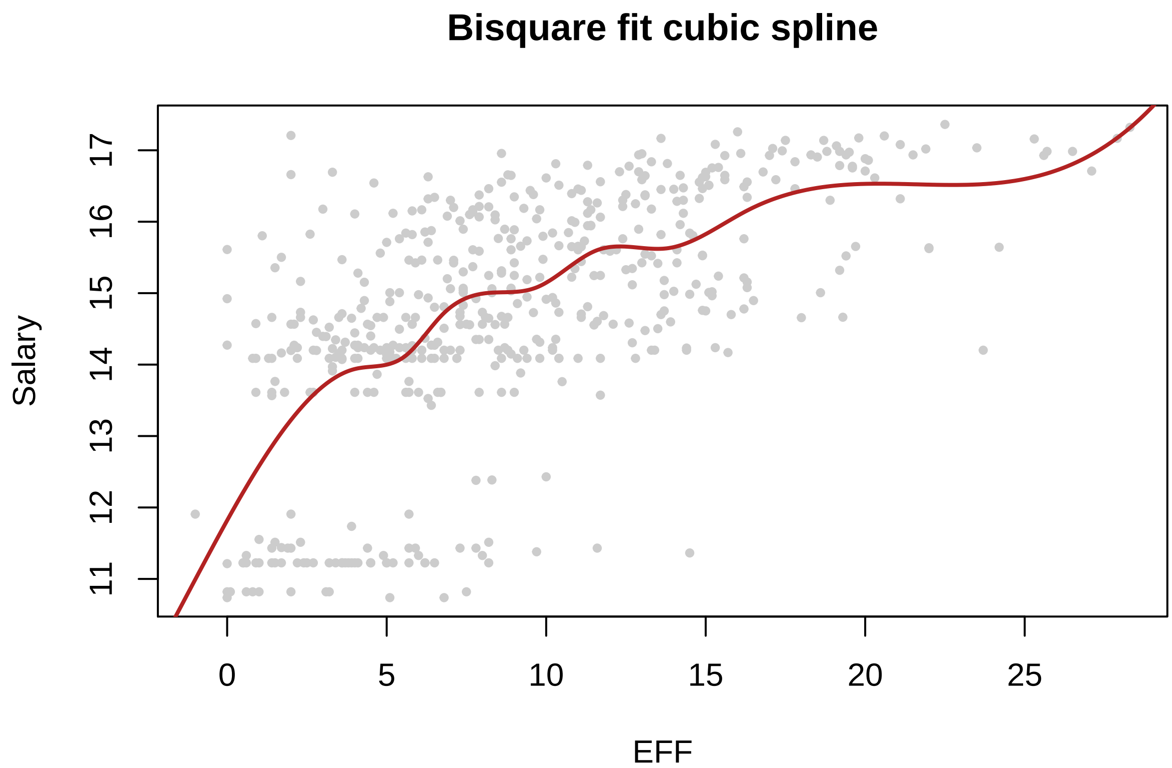
We then try to fit a cubic spline to our data. First we need to get its basis functions for our fitted model, which can be illustrated by plotting them as a function of EFF. The basis functions are clearly not polynomials. The estimated smooth will be a linear combination of these basis functions.

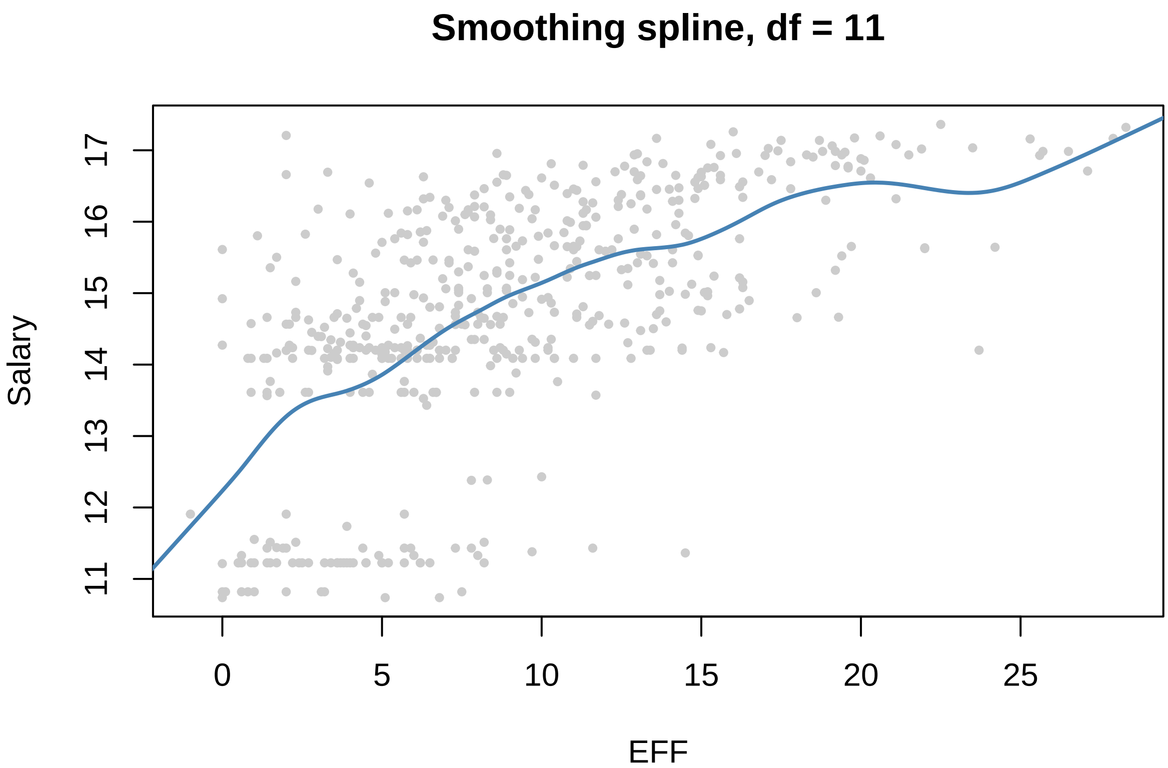
```
## Warning in bs(x, degree = 3L, knots = structure(c(4, 5.6, 6.76, 8.2, 9.7, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases
```



We then fitted the cubic spline to the data.

```
## Warning in bs(x, degree = 3L, knots = structure(c(4, 5.6, 6.76, 8.2, 9.7, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases
```


Bisquare fit cubic spline
Salary
11
12
13
14
15
16
17
0
5
10
15
20
25
EFF


Smoothing spline, df = 11
Salary
11
12
13
14
15
16
17
EFF
0
5
10
15
20
25

Random Forest

We would like to utilize random forest to determine the importance of explanatory variates.

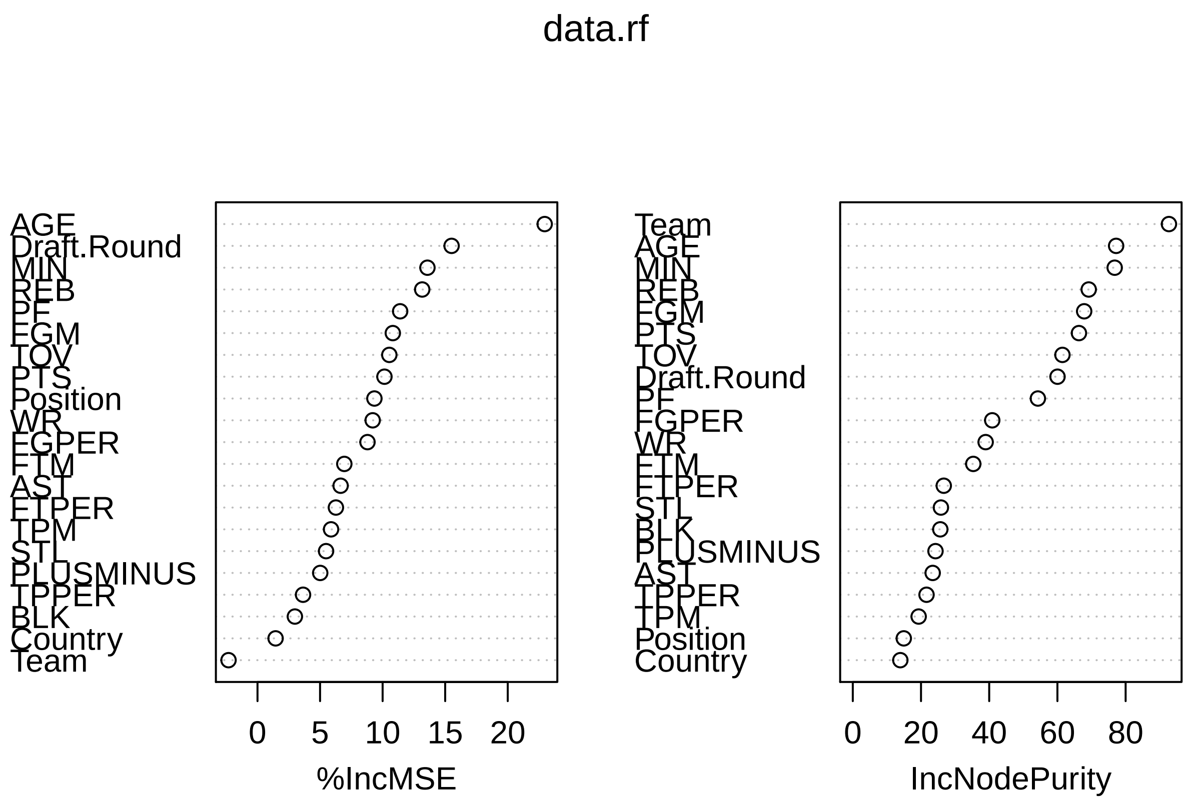
```
## Warning: package 'randomForest' was built under R version 3.4.4
```

```
## randomForest 4.6-14
```

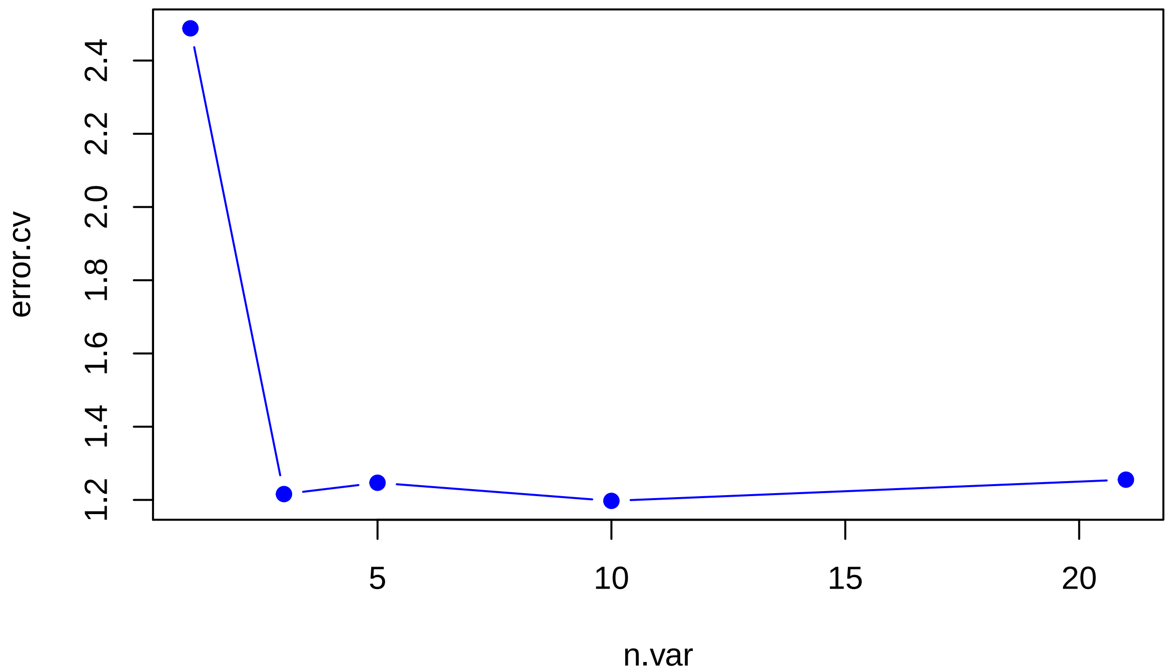
```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##             IncNodePurity
## PTS              66.31385
## REB              69.18051
## AST              23.42569
## TOV              61.46742
## STL              25.83868
## BLK              25.65737
## Team             92.70357
## WR               38.95579
## AGE              77.21489
## FGM              67.84526
## FGPER            40.89163
## TPM              19.31425
## TPPER            21.62336
## FTM              35.32419
## FTPER            26.67255
## PF               54.27138
## PLUSMINUS        24.24912
## Position         14.95129
## Country          13.93867
## MIN              76.79506
## Draft.Round      60.02632
```

```
##               %IncMSE
## PTS         10.145684
## REB         13.163556
## AST          6.641795
## TOV         10.534684
## STL          5.480177
## BLK          2.987393
## Team        -2.315631
## WR           9.205577
## AGE         22.952298
## FGM         10.815210
## FGPER        8.790611
## TPM          5.878884
## TPPER        3.640674
## FTM          6.938165
## FTPER        6.267423
## PF          11.401504
## PLUSMINUS    5.012516
## Position     9.342765
## Country      1.446441
## MIN         13.582576
## Draft.Round 15.510622
```

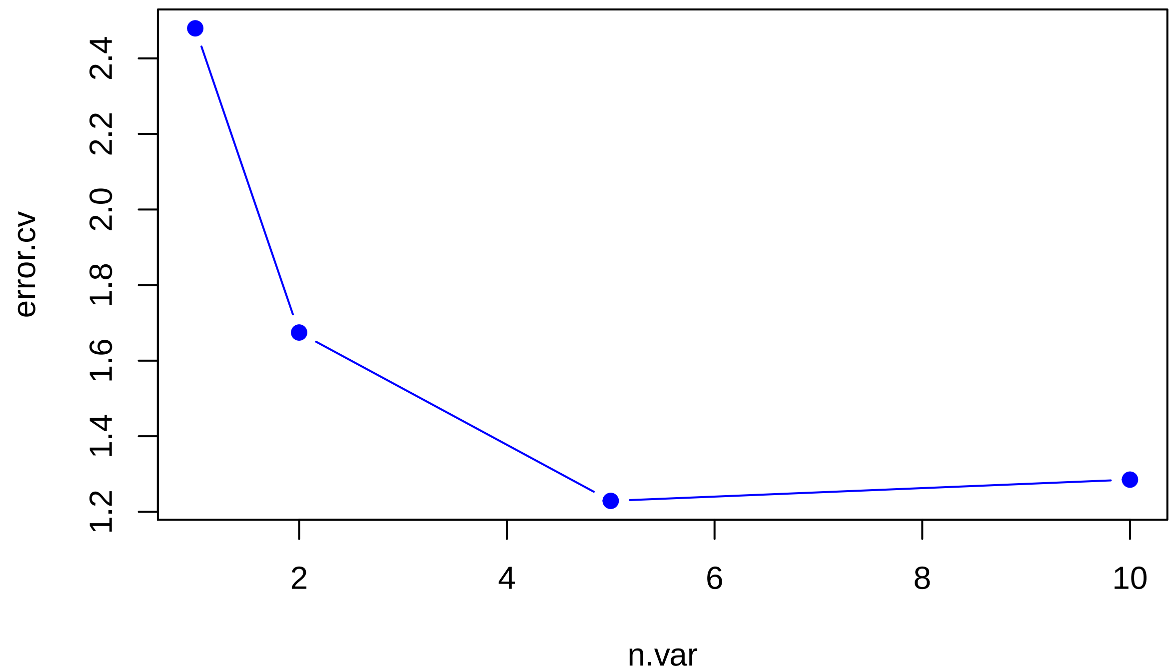


data.rf
AGE
Draft.Round
MIN
REB
PF
FGM
TOV
PTS
Position
WR
FGPER
FTM
AST
FTPER
TPM
STL
PLUSMINUS
TPPER
BLK
Country
Team
0
5
10
15
20
%IncMSE
Team
AGE
MIN
REB
FGM
PTS
TOV
Draft.Round
PF
FGPER
WR
FTM
FTPER
STL
BLK
PLUSMINUS
AST
TPPER
TPM
Position
Country
0
20
40
60
80
IncNodePurity

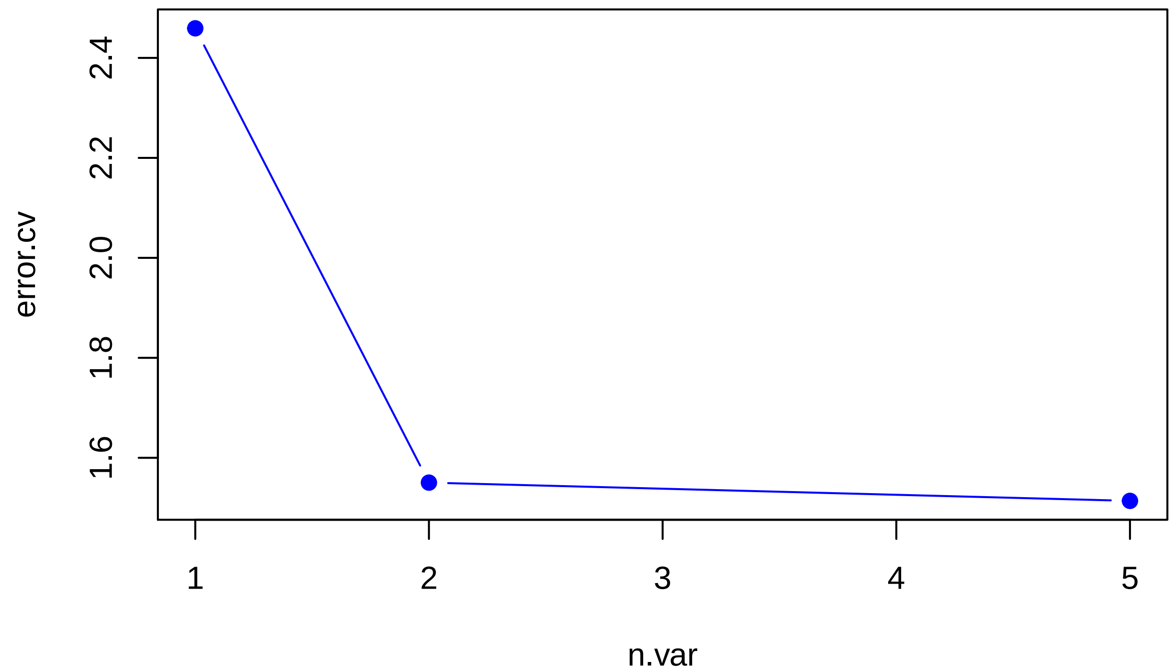


We used PTS, REB, AST, TOV, STL, BLK, Team, WR, AGE, FGM, FGPER, TPM, TPPER, FTM, FTPER, PF, PLUSMINUS, Position, Country, MIN, and Draft.Round against log(Salary) for the random forest. We did not choose to include Draft.Number because it is a categorical variate with 60 different potential values, but random forest does not accept categorical predictors with more than 53 categories.

The result suggests that the error of cross validation is the lowest for 10 explanatory variates, at about 1.19. We then choose the top 10 most important variates based on RSS, Team, MIN, FGM, PTS, AGE, REB, Draft.Round, PF, TOV, and FTM, and run the process again.



The result from the second run suggests that the error of cross validation is the lowest when there are 5 explanatory variates, at around 1.19. We then choose the top 5 most important variates againbased on RSS, which are Team, MIN, FGM, PTS, and AGE, and run the process again.

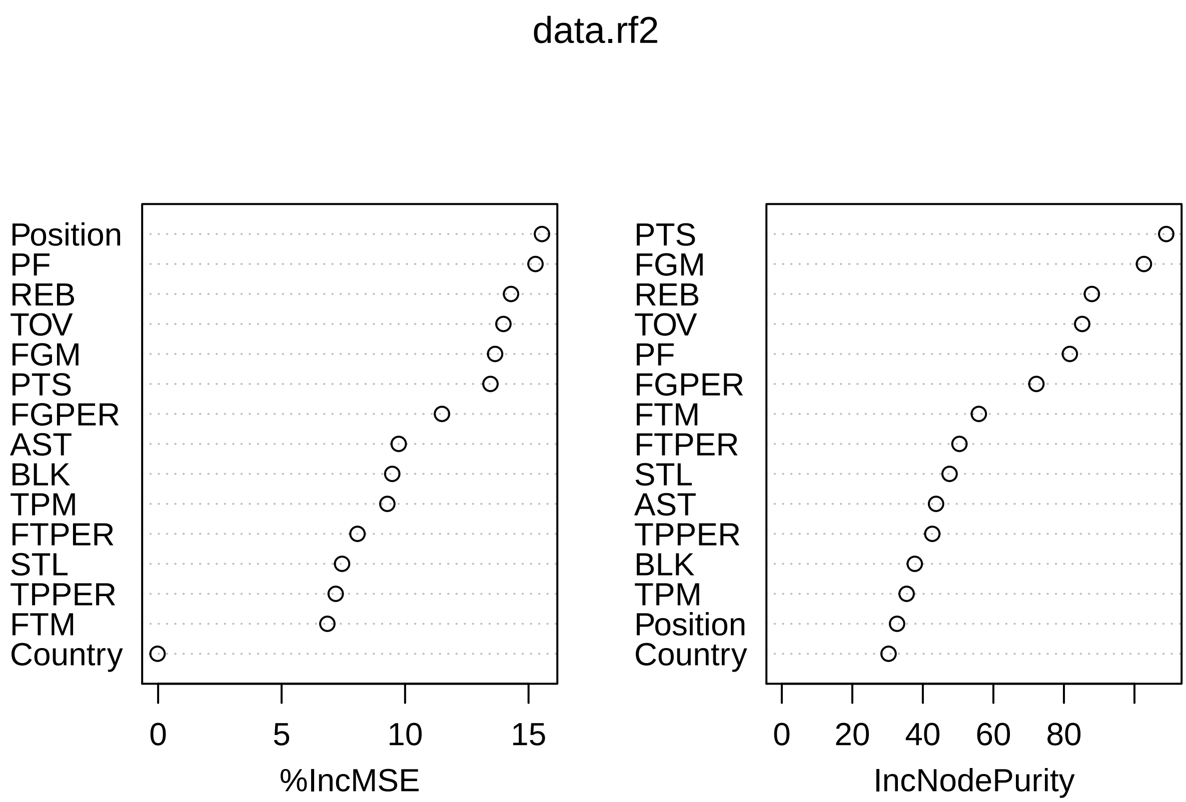


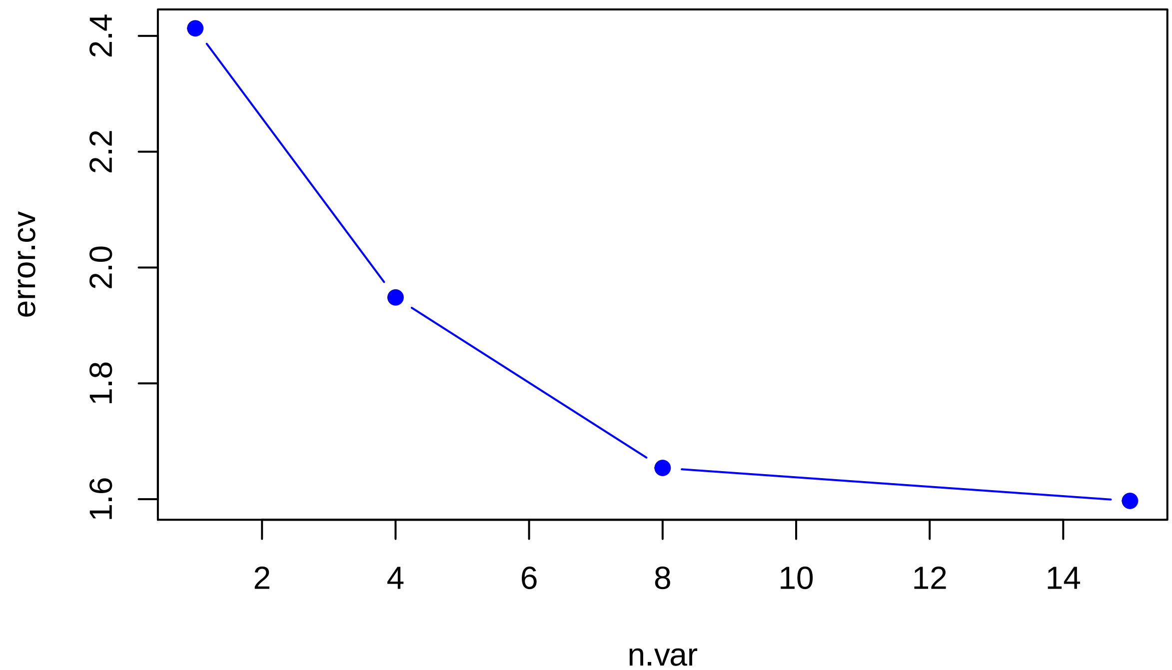
The result from the third run suggests that the error of cross validation is the lowest when there are 5 explanatory variates, at around 1.59. We can say that the Team, MIN, FGM, PTS, and AGE are important variates based on cross validation. Also, AGE seems to be the most important for predictive purposes.

To make a more sensible modelling exercise would consider how Salary might depend on just those explanatory variates that were under the control of the NBA player. Therefore, we removed Team, MIN, WR, AGE, Draft.Round, and PLUSMINUS. Age is obviously an uncontrollable variate. We think players rarely have control for Team, MIN, and Draft.Round since it does not depend on players' previous NBA performance, instead, these would depend on the decisions from coach and the organization Also, WR and PLUSMINUS have a lot to do with the teammates of the NBA player we are trying to analyze, so we decided to take these out of consideration as well. This move left us with 15 explanatory variates.

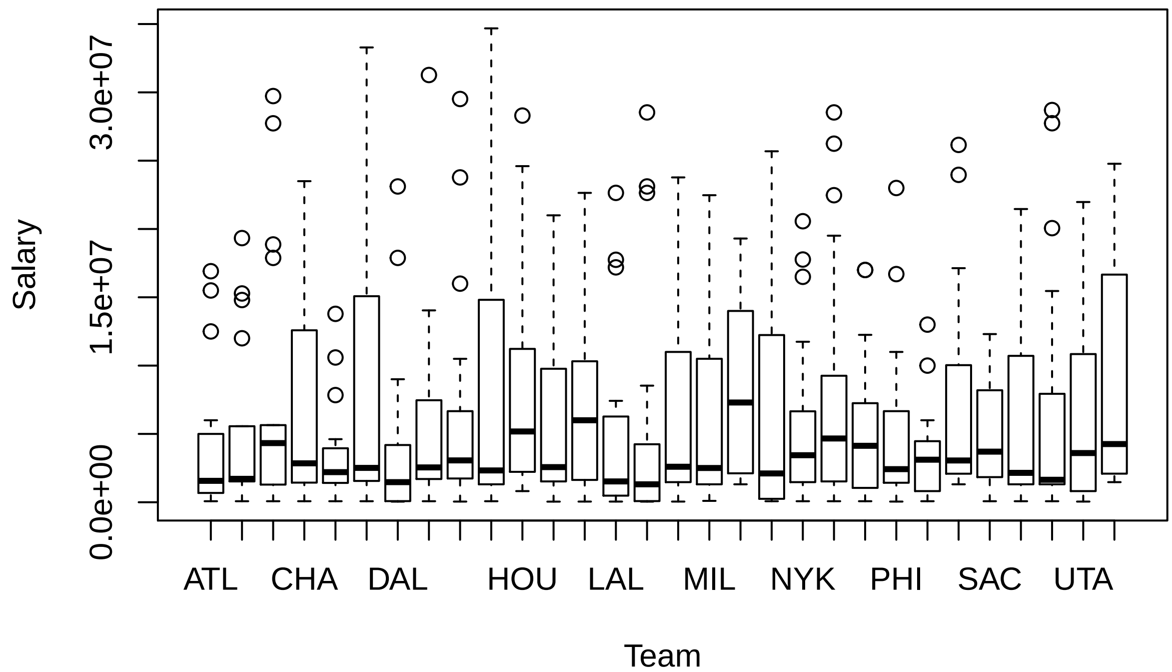
```
##          IncNodePurity
## PTS          109.01022
## REB           87.92080
## AST           43.74299
## TOV           85.18036
## STL           47.57607
## BLK           37.70856
## FGM          102.68892
## FGPER         72.19427
## TPM           35.40049
## TPPER         42.67947
## FTM           55.86730
## FTPER         50.38130
## PF            81.68347
## Position      32.68509
## Country       30.27948
```

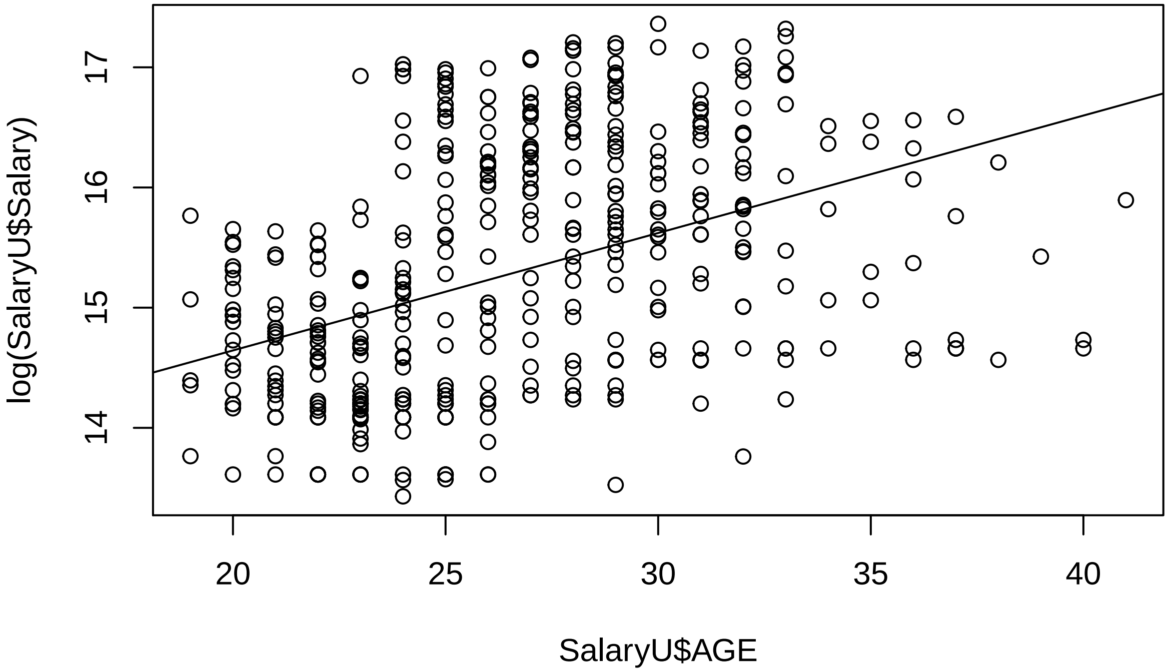
```
##             %IncMSE
## PTS      13.45676759
## REB      14.29023733
## AST       9.74211629
## TOV      13.98210382
## STL       7.44861084
## BLK       9.48017036
## FGM      13.64514900
## FGPER    11.49439423
## TPM       9.28224115
## TPPER     7.19087229
## FTM       6.85360588
## FTPER     8.07314266
## PF       15.27999289
## Position 15.54249201
## Country  -0.02396414
```

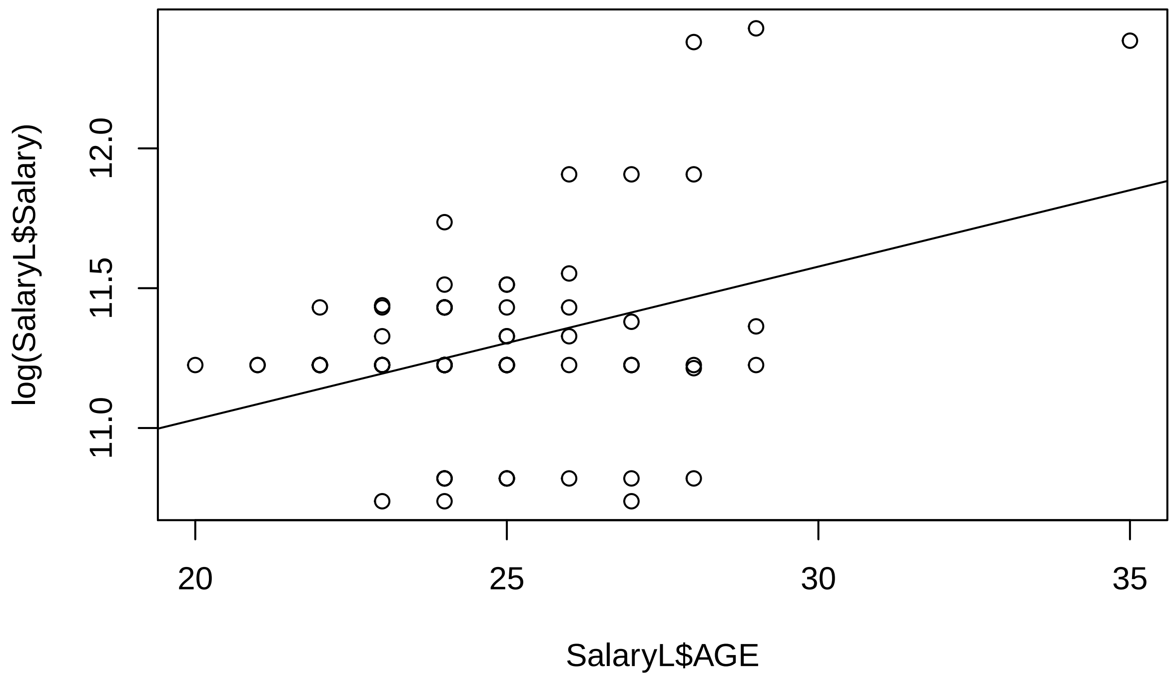




The cross validation suggests that all 15 explanatory variates are important. The result also suggests that FGM, TOV, REB, and PTS are the most important.


Salary
0.0e+00
1.5e+07
3.0e+07
ATL
CHA
DAL
HOU
LAL
MIL
NYK
PHI
SAC
UTA
Team


log(SalaryU$Salary)
14
15
16
17
20
25
30
35
40
SalaryU$AGE


log(SalaryL$Salary)
12.0
11.5
11.0
20
25
30
35
SalaryL$AGE