Homework4

Shraddha Hemant Kadam (sxk190069@utdallas.edu)

07/28/2020

```
pacman::p_load(data.table, MASS, ggplot2, dplyr, ISLR, RColorBrewer,
               rpart, rpart.plot, gbm, caret, tree,leaps, moments,randomForest, gains)
knitr::opts_chunk$set(echo = TRUE, fig.height=8, fig.width=12, fig.path = 'Figs/')
theme_set(theme_classic())
options(digits = 3)
```

```
data(Hitters)
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Andy Allanson      293   66     1   30  29    14     1    293    66      1
## -Alan Ashby         315   81     7   24  38    39    14   3449   835     69
## -Alvin Davis        479  130    18   66  72    76     3   1624   457     63
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575    225
## -Andres Galarraga   321   87    10   39  42    30     2    396   101     12
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133     19
##                   CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Andy Allanson       30   29     14      A        E     446      33     20
## -Alan Ashby         321  414    375      N        W     632      43     10
## -Alvin Davis        224  266    263      A        W     880      82     14
## -Andre Dawson       828  838    354      N        E     200      11      3
## -Andres Galarraga    48   46     33      N        E     805      40      4
## -Alfredo Griffin    501  336    194      A        W     282     421     25
##                   Salary NewLeague
## -Andy Allanson        NA         A
## -Alan Ashby        475.0         N
## -Alvin Davis       480.0         A
## -Andre Dawson      500.0         N
## -Andres Galarraga   91.5         N
## -Alfredo Griffin   750.0         A
```

```
tail(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Wayne Krenchicki   221   53     2   21  23    22     8   1063   283     15
## -Willie McGee       497  127     7   65  48    37     5   2703   806     32
## -Willie Randolph    492  136     5   76  50    94    12   5511  1511     39
## -Wayne Tolleson     475  126     3   61  43    52     6   1700   433      7
## -Willie Upshaw      573  144     9   85  60    78     8   3198   857     97
## -Willie Wilson      631  170     9   77  44    31    11   4908  1457     30
```

```
##                    CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Wayne Krenchicki   107  124    106      N        E     325      58      6
## -Willie McGee       379  311    138      N        E     325       9      3
## -Willie Randolph    897  451    875      A        E     313     381     20
## -Wayne Tolleson     217   93    146      A        W      37     113      7
## -Willie Upshaw      470  420    332      A        E    1314     131     12
## -Willie Wilson      775  357    249      A        W     408       4      3
##                    Salary NewLeague
## -Wayne Krenchicki      NA         N
## -Willie McGee         700         N
## -Willie Randolph      875         A
## -Wayne Tolleson       385         A
## -Willie Upshaw        960         A
## -Willie Wilson       1000         A
```

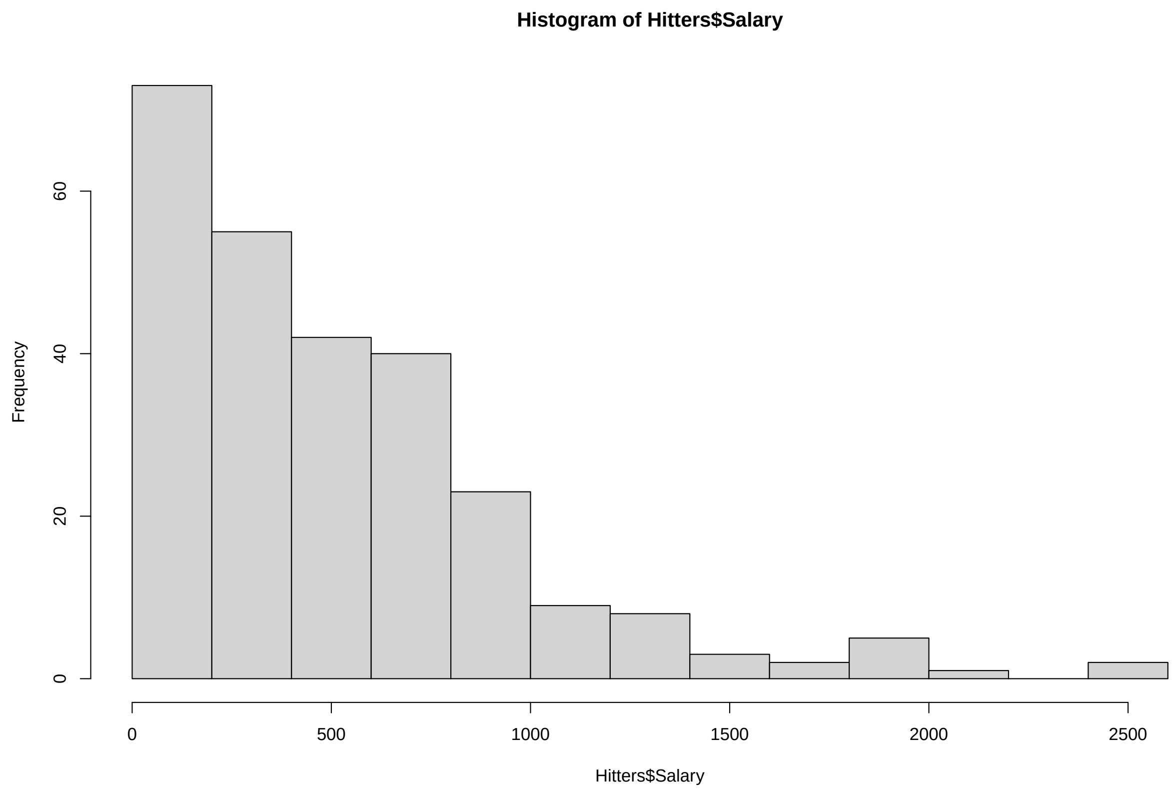
```r
colSums(is.na(Hitters))
```

```
##     AtBat      Hits     HmRun      Runs       RBI     Walks     Years    CAtBat
##         0         0         0         0         0         0         0         0
##     CHits    CHmRun     CRuns      CRBI    CWalks    League  Division   PutOuts
##         0         0         0         0         0         0         0         0
##   Assists    Errors    Salary NewLeague
##         0         0        59         0
```

```r
Hitters <- Hitters[!is.na(Hitters$Salary), ]
```

1 Ans: The dataset had 322 observations initially. When I checked for missing values in the dataset which had observations with unknown salary information, I found out that there were 59 observations having missing values. So those 59 observations were removed, leaving 263 observations in this dataset.

```r
hist(Hitters$Salary)
```

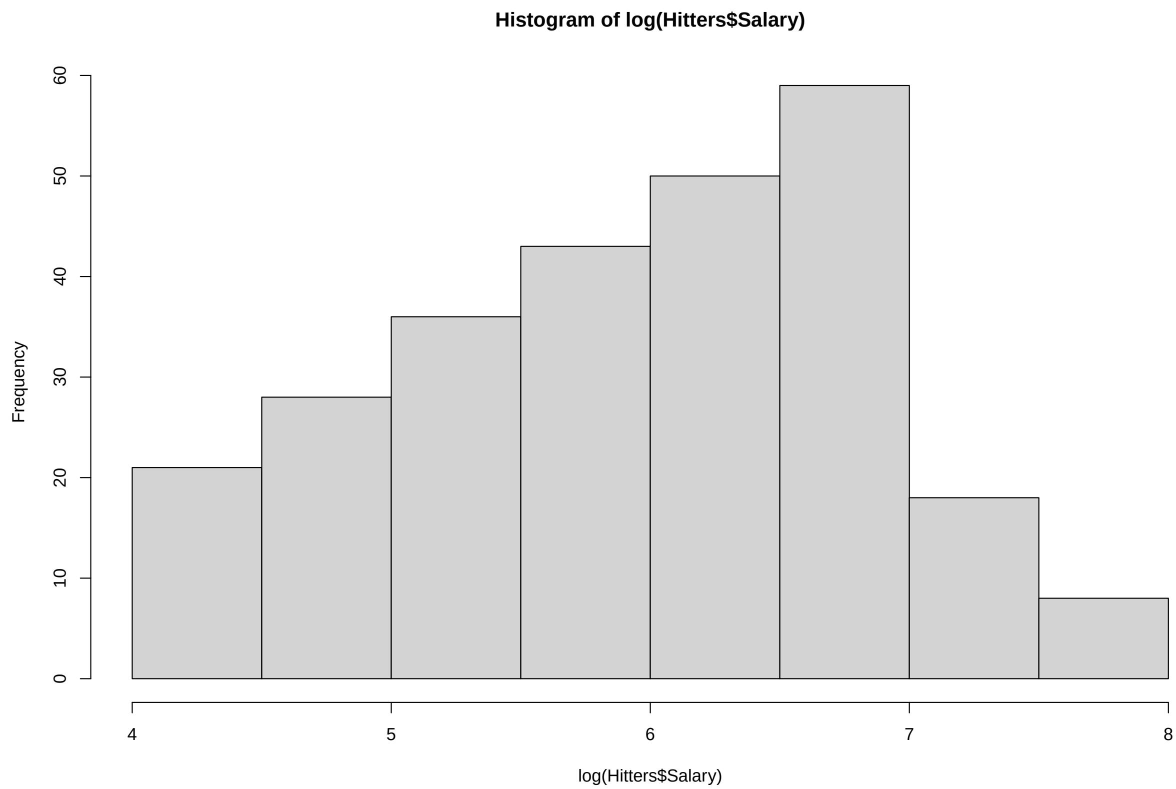



```r
skewness(Hitters$Salary)
```

```
## [1] 1.58
```

2 Ans: To transform the salaries using a natural log transformation, lets first visualize how our data looks like. By looking at the above histogram we can see that it is right skewed with a skewness of 1.58. Logarithmic transformation is transforming a highly skewed variable into a more normalized distribution. Hence, let us see what impact does log transformation of the salary have on skewness.

```r
hist(log(Hitters$Salary))
```



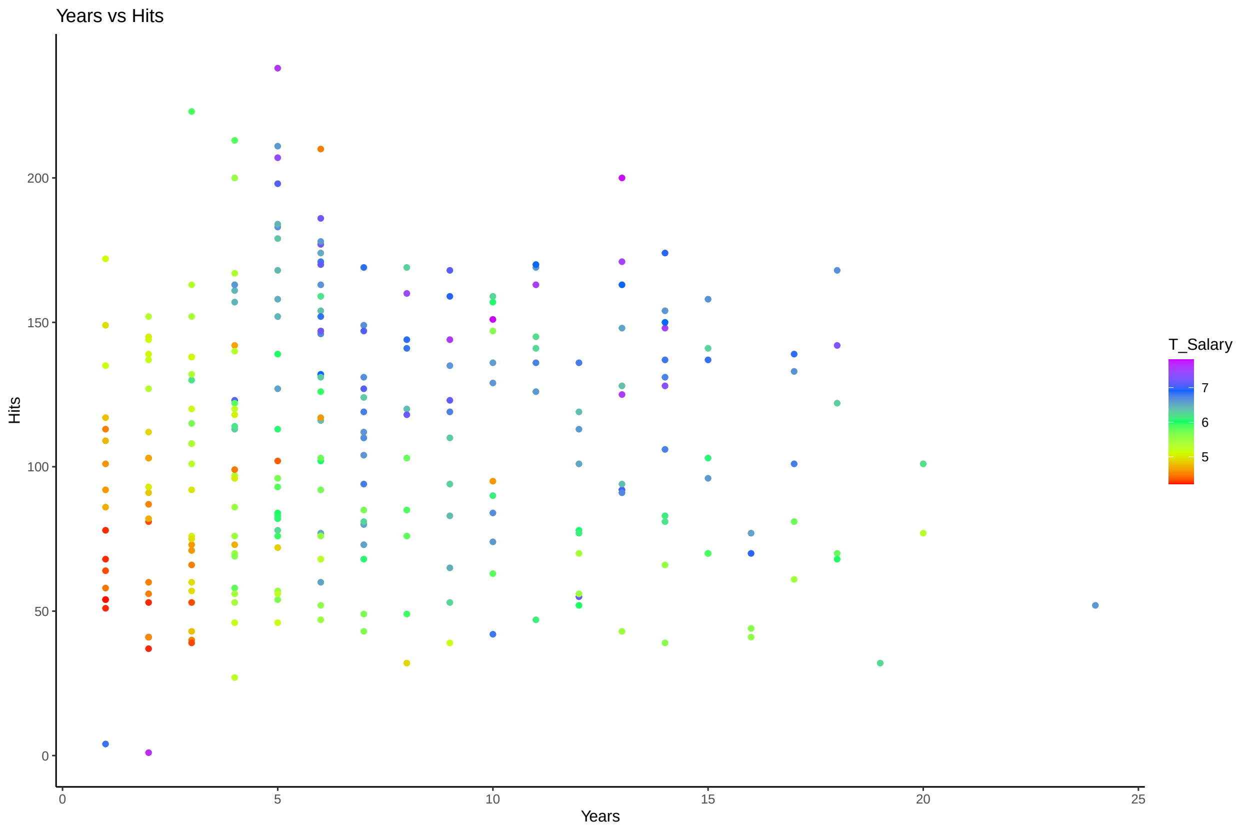

```
skewness(log(Hitters$Salary))
```

```
## [1] -0.181
```

From above we can see that the skewness has been significantly decreased to -0.181. So now we make changes in the dataset by transforming the salaries.

```
T_Salary<- Hitters[,19]
T_Salary<- log(T_Salary)
Hitters<-Hitters[,-19]
Hitters<-cbind(Hitters,T_Salary)
```

```
ggplot(Hitters, aes (x=Years, y=Hits))+ geom_point(aes(color = T_Salary)) +
  scale_color_gradientn(colours = rainbow(5)) +
  ggtitle("Years vs Hits")
```



3 Ans: From the above scatter plot we can see that by considering players having 0-5 years of experience, the salary is on the lower end of the spectrum regardless of having higher number of hits. And as number of years are increasing, the salary is also increasing. There are more number of players who have 7 or less years of experience. There is an outlier where players have higher salary as compared to the other players with the same number of experience and higher hits.

```r
options(digits=6)
search <- regsubsets(T_Salary ~ ., data = Hitters, nbest = 1,
                     nvmax = dim(Hitters),
                     method = "exhaustive")
sum <- summary(search)
sum$which
```

```
##    (Intercept) AtBat  Hits HmRun  Runs   RBI Walks Years CAtBat CHits CHmRun
## 1         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE
## 2         TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   TRUE FALSE  FALSE
## 3         TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 4         TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE   TRUE FALSE  FALSE
## 5         TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE  TRUE  FALSE
## 6         TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE  TRUE  FALSE
## 7         TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 8         TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 9         TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 10        TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 11        TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 12        TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 13        TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  FALSE FALSE  FALSE
## 14        TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE   TRUE FALSE  FALSE
```

```
## 15         TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE   TRUE  TRUE  FALSE
## 16         TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE   TRUE  TRUE  FALSE
## 17         TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  FALSE
## 18         TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  FALSE
## 19         TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE   TRUE
##    CRuns  CRBI CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1   TRUE FALSE  FALSE   FALSE     FALSE   FALSE   FALSE  FALSE      FALSE
## 2  FALSE FALSE  FALSE   FALSE     FALSE   FALSE   FALSE  FALSE      FALSE
## 3  FALSE FALSE  FALSE   FALSE     FALSE   FALSE   FALSE  FALSE      FALSE
## 4  FALSE FALSE  FALSE   FALSE     FALSE   FALSE   FALSE  FALSE      FALSE
## 5  FALSE FALSE  FALSE   FALSE      TRUE   FALSE   FALSE  FALSE      FALSE
## 6  FALSE FALSE  FALSE   FALSE      TRUE   FALSE   FALSE  FALSE      FALSE
## 7   TRUE FALSE   TRUE   FALSE     FALSE    TRUE   FALSE  FALSE      FALSE
## 8   TRUE FALSE   TRUE   FALSE      TRUE    TRUE   FALSE  FALSE      FALSE
## 9   TRUE FALSE   TRUE    TRUE      TRUE    TRUE   FALSE  FALSE      FALSE
## 10  TRUE FALSE   TRUE    TRUE      TRUE    TRUE   FALSE  FALSE       TRUE
## 11  TRUE FALSE   TRUE    TRUE      TRUE    TRUE   FALSE  FALSE       TRUE
## 12  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE      FALSE
## 13  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 14  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 15  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 16  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 17  TRUE FALSE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 18  TRUE  TRUE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
## 19  TRUE  TRUE   TRUE    TRUE      TRUE    TRUE    TRUE   TRUE       TRUE
```

```r
sum$bic
```

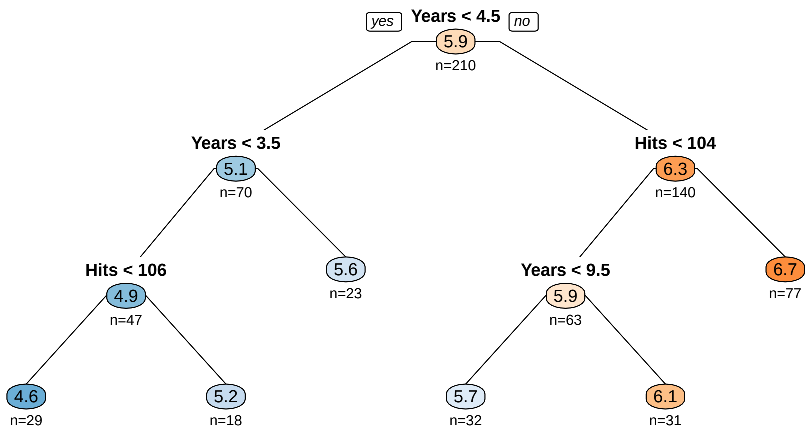
```
##  [1] -117.030 -156.429 -159.278 -159.218 -159.089 -157.921 -157.123 -156.195
##  [9] -152.765 -148.806 -144.596 -140.654 -136.548 -131.094 -125.711 -120.199
## [17] -114.713 -109.186 -103.614
```

4 Ans: We know that lower the BIC value better is the model, and in order to calculate the BIC value, I have used regsubsets() to perform best subset selection from the regression model in the above code with the exhaustive method. So if we consider above BIC values, "-159.278" is the lowest among all others which gives us the sub-model 3 in exhaustive search as the best model having the following predictors: 1.Hits 2.Walks 3.Years

5 Ans: Creating a training dataset consisting of 80% of the observations and test/validation dataset consisting of the remaining observations.

```r
set.seed(42)
train.index<-sample(1:nrow(Hitters), nrow(Hitters)*0.8)
train.df <- Hitters[train.index, ]
valid.df<- Hitters[-train.index, ]
```

```r
dpr.ct <- rpart(T_Salary ~ Years + Hits, data = train.df,
                  method = "anova")
prp(dpr.ct, type = 1, extra = 1, under = TRUE,roundint=FALSE,
    split.font = 2, varlen = -10,
    box.palette = "BuOr")
```




```r
for(i in 1:nrow(Hitters))
  {
if(((Hitters$Hits[i]>= 104) && Hitters$Years[i] >= 4.5))
  {
  print(row.names(Hitters)[i])
   }
}
```

```
## [1] "-Andre Dawson"
## [1] "-Alfredo Griffin"
## [1] "-Alan Trammell"
## [1] "-Buddy Bell"
## [1] "-Bob Brenly"
## [1] "-Bill Buckner"
## [1] "-Brett Butler"
## [1] "-Bo Diaz"
## [1] "-Bill Doran"
## [1] "-Brian Downing"
## [1] "-Brook Jacoby"
## [1] "-Bill Madlock"
## [1] "-Chili Davis"
## [1] "-Carney Lansford"
## [1] "-Cal Ripken"
## [1] "-Don Baylor"
## [1] "-Doug DeCinces"
## [1] "-Darrell Evans"
## [1] "-Dwight Evans"
```

```
## [1] "-Damaso Garcia"
## [1] "-Don Mattingly"
## [1] "-Dale Murphy"
## [1] "-Dave Parker"
## [1] "-Denny Walling"
## [1] "-Dave Winfield"
## [1] "-Eddie Milner"
## [1] "-Eddie Murray"
## [1] "-Frank White"
## [1] "-George Bell"
## [1] "-George Brett"
## [1] "-Gary Carter"
## [1] "-Gary Gaetti"
## [1] "-Gary Pettis"
## [1] "-Garry Templeton"
## [1] "-Gary Ward"
## [1] "-Glenn Wilson"
## [1] "-Harold Baines"
## [1] "-Hubie Brooks"
## [1] "-Jesse Barfield"
## [1] "-Jose Cruz"
## [1] "-Jody Davis"
## [1] "-Julio Franco"
## [1] "-Jim Gantner"
## [1] "-Jim Morrison"
## [1] "-Johnny Ray"
## [1] "-Jim Rice"
## [1] "-Kevin Bass"
## [1] "-Kirk Gibson"
## [1] "-Ken Griffey"
## [1] "-Keith Hernandez"
## [1] "-Kent Hrbek"
## [1] "-Keith Moreland"
## [1] "-Ken Oberkfell"
## [1] "-Leon Durham"
## [1] "-Lee Lacy"
## [1] "-Lloyd Moseby"
## [1] "-Larry Parrish"
## [1] "-Lou Whitaker"
## [1] "-Marty Barrett"
## [1] "-Mike Davis"
## [1] "-Mike Easler"
## [1] "-Mel Hall"
## [1] "-Mookie Wilson"
## [1] "-Ozzie Smith"
## [1] "-Paul Molitor"
## [1] "-Pat Tabler"
## [1] "-Ron Hassey"
## [1] "-Rickey Henderson"
## [1] "-Ray Knight"
## [1] "-Ron Oester"
## [1] "-Rafael Ramirez"
## [1] "-Ryne Sandberg"
## [1] "-Roy Smalley"
```

```
## [1] "-Robin Yount"
## [1] "-Steve Balboni"
## [1] "-Scott Fletcher"
## [1] "-Steve Garvey"
## [1] "-Steve Sax"
## [1] "-Tony Bernazard"
## [1] "-Tom Brunansky"
## [1] "-Tony Gwynn"
## [1] "-Tommy Herr"
## [1] "-Tony Pena"
## [1] "-Tony Phillips"
## [1] "-Tim Wallach"
## [1] "-Von Hayes"
## [1] "-Wally Backman"
## [1] "-Wade Boggs"
## [1] "-Willie McGee"
## [1] "-Willie Randolph"
## [1] "-Wayne Tolleson"
## [1] "-Willie Upshaw"
## [1] "-Willie Wilson"
```

6 Ans: From the above tree we can see that highest salary is transformed salary of 6.7, so the rule for players receiving the highest salary is Years should be 4.5 or more and Number of hits should be greater than or equals 104. The players likely to receive highest salaries according to this model are as stated above in the results.