BUAN6356_Homework 4_Group 7

15/11/2019

```
#install packages
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load( ISLR,tidyverse,ggplot2, leaps, data.table, rpart, rpart.plot,gbm, MASS, caret, randomFo
theme_set(theme_classic())

search()
```

```
##  [1] ".GlobalEnv"           "package:glmnet"       "package:foreach"
##  [4] "package:Matrix"       "package:corrplot"     "package:randomForest"
##  [7] "package:caret"        "package:lattice"      "package:MASS"
## [10] "package:gbm"          "package:rpart.plot"   "package:rpart"
## [13] "package:data.table"   "package:leaps"        "package:forcats"
## [16] "package:stringr"      "package:dplyr"        "package:purrr"
## [19] "package:readr"        "package:tidyr"        "package:tibble"
## [22] "package:ggplot2"      "package:tidyverse"    "package:ISLR"
## [25] "package:pacman"       "package:stats"        "package:graphics"
## [28] "package:grDevices"    "package:utils"        "package:datasets"
## [31] "package:methods"      "Autoloads"            "package:base"
```

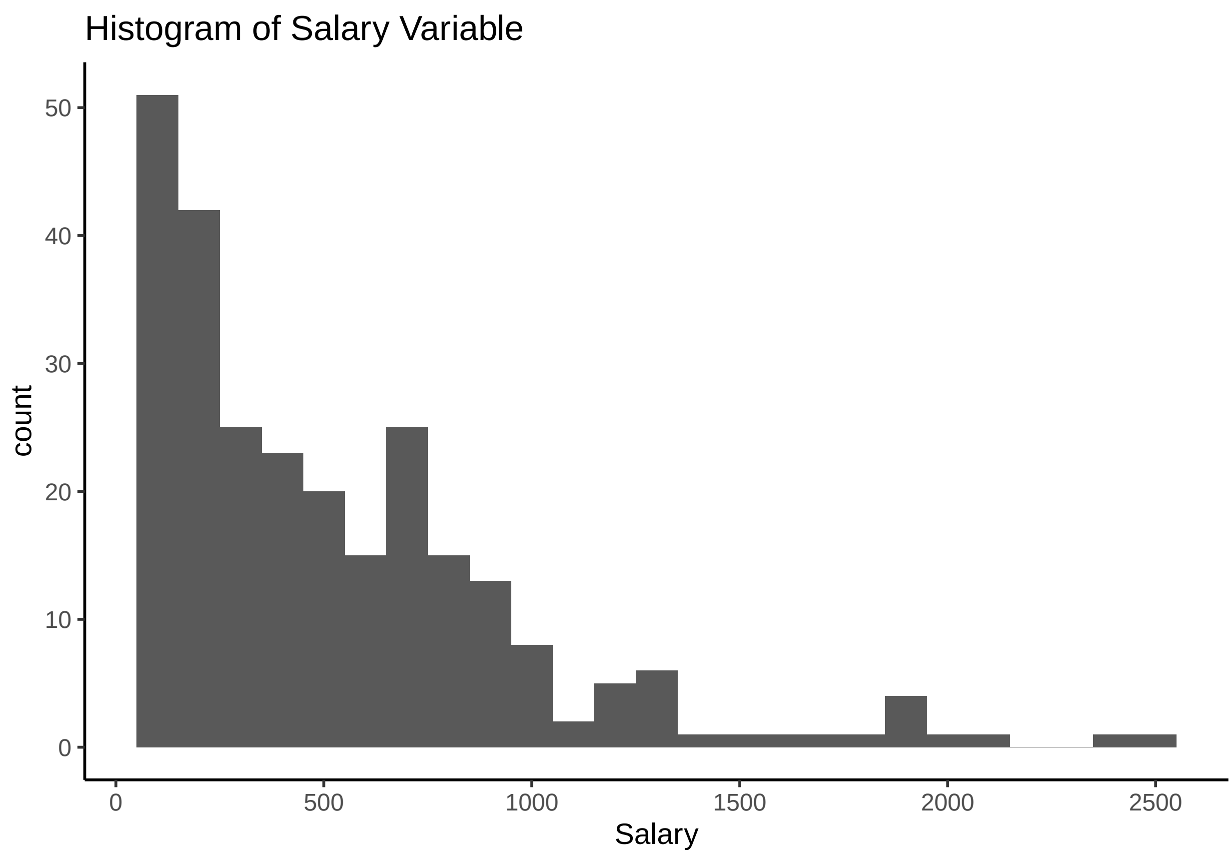
Answer 1

```
data(Hitters)
Hitters.df <- data.frame(Hitters)

HittersModified.df <- Hitters.df[!(is.na(Hitters$Salary) | Hitters$Salary==""), ]
```

There are 322 observations in original dataset Hitters. After removing Nulls from Salary column, we are left with 263 observations. 59 observations were removed where Salary was null.

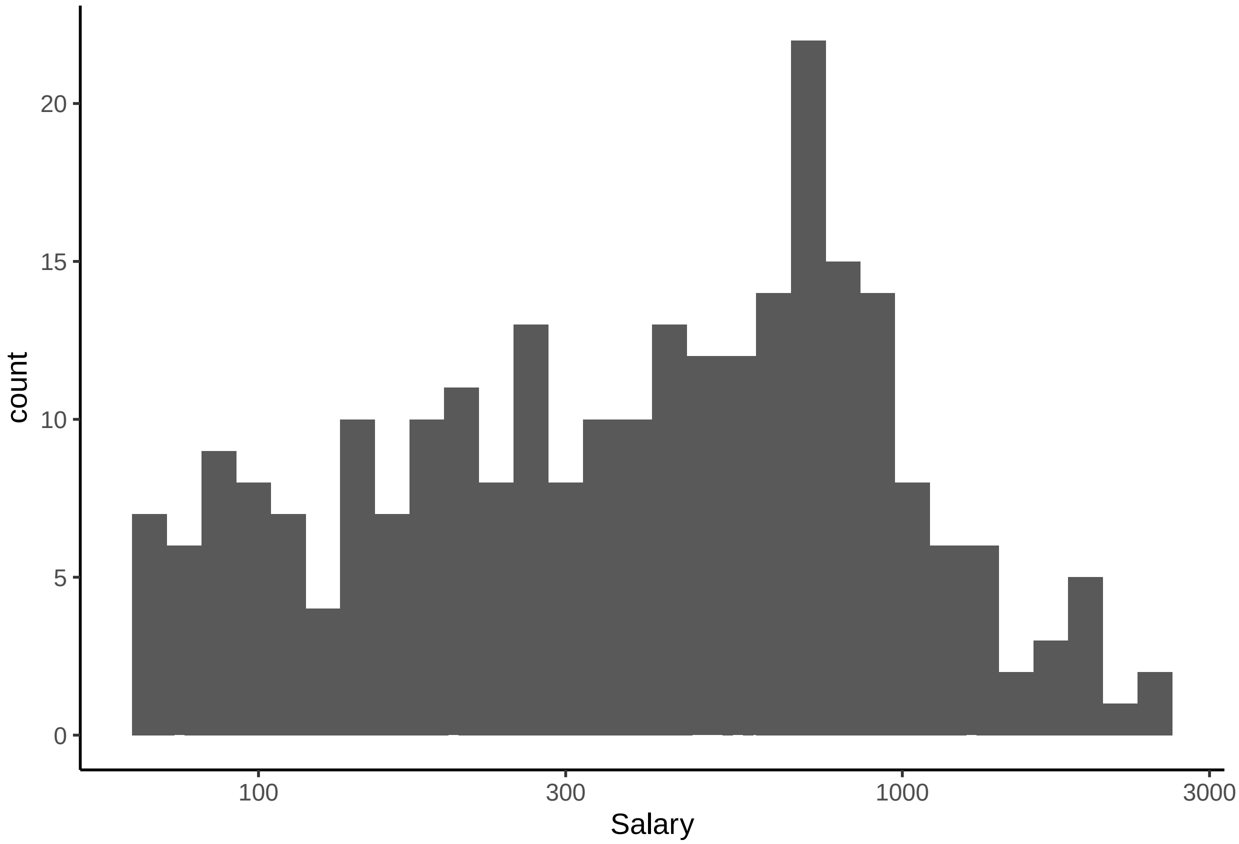
Answer 2

```
ggplot(HittersModified.df) +
  geom_histogram(aes(x = Salary), binwidth = 100) +
  ggtitle("Histogram of Salary Variable")
```




```
ggplot(HittersModified.df, aes(x = Salary)) + geom_histogram() + scale_x_log10() + stat_bin(bins = 100)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



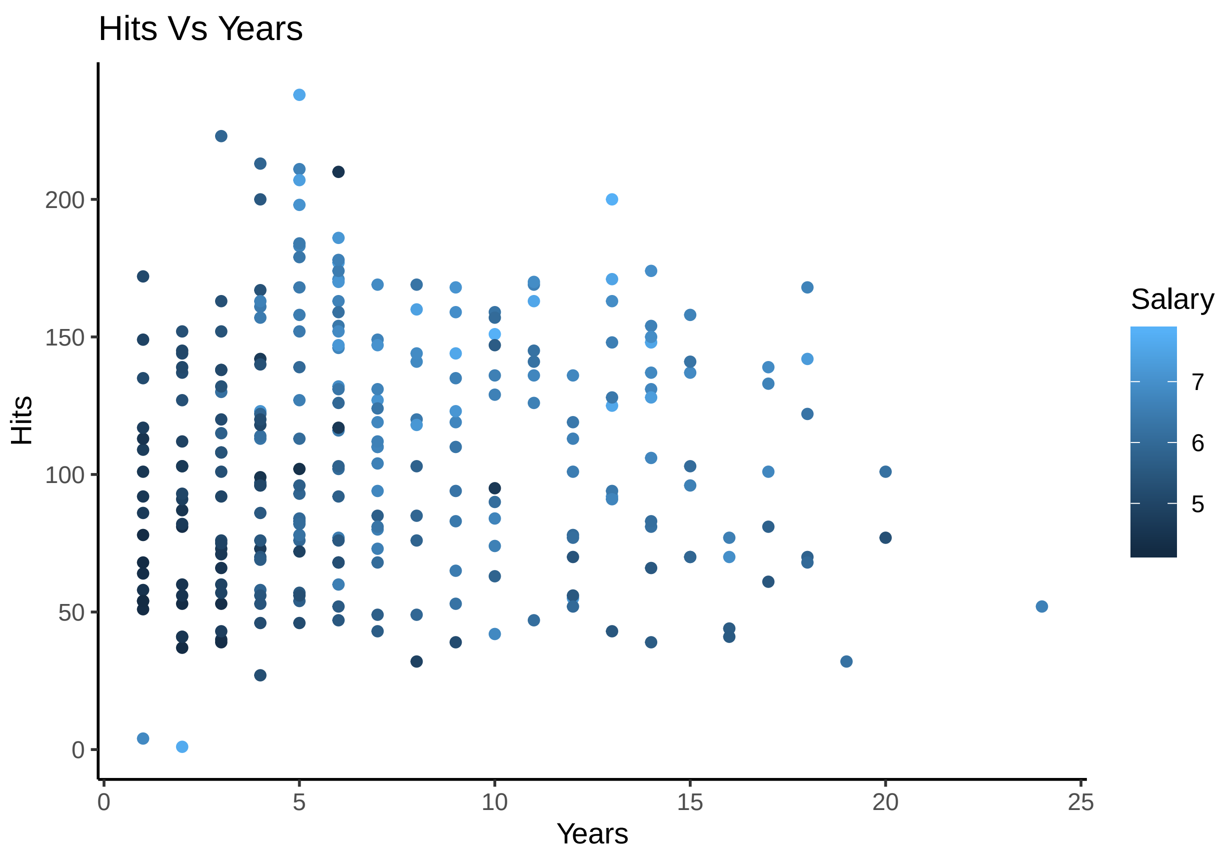

```
set.seed(42)
#skewness(HittersModified.df$Salary)
#skewness(log1p(HittersModified.df$Salary))
HittersModified.df$Salary <- log(HittersModified.df$Salary)
```

We first plot histogram for salary variable in order to check the skewness. We find that it is right skewed as expected, which means only few players receive high salaries than other players.

After performing log transformation, we see that the skewness is corrected and the distribution is almost normal.

Answer 3

```
ggplot(HittersModified.df) +
  geom_point(aes(x = Years, y = Hits, color = Salary)) +
 # scale_colour_manual(values=c("red", "blue","green"))
  ggtitle("Hits Vs Years")
```



As seen from the scatterplot above, as the Years and Hits increase, the log(Salary) also increaes. This is indicated by the color coding. Lighter the color shade, more is the value of log(Salary). It can be interpreted that more number of hits, more salary is offered to the players. Also, a player with more experience gets a higher pay.

Answer 4

We will perform a linear regression model of log Salary on all the numerical predictors.

```
#linear regression
require(leaps)
set.seed(42)
HittersModified.lm <- lm(Salary ~ ., data = HittersModified.df)

#regsubsets
search <- regsubsets(Salary ~ ., data = HittersModified.df)
summary_regsubsets <- summary(search)
summary_regsubsets$bic
```

```
## [1] -117.0304 -156.4291 -159.2777 -159.2182 -159.0885 -157.9207 -157.1229
## [8] -156.1954
```

```r
which.min(summary_regsubsets$bic)
```

```
## [1] 3
```

```r
#show models
summary_regsubsets$which
```

```
##   (Intercept) AtBat  Hits HmRun  Runs   RBI Walks Years CAtBat CHits
## 1        TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
## 2        TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   TRUE FALSE
## 3        TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE
## 4        TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE   TRUE FALSE
## 5        TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE  TRUE
## 6        TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE  TRUE
## 7        TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE
## 8        TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  FALSE FALSE
##   CHmRun CRuns  CRBI CWalks LeagueN DivisionW PutOuts Assists Errors
## 1  FALSE  TRUE 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   FALSE      TRUE   FALSE   FALSE  FALSE
## 6  FALSE FALSE FALSE  FALSE   FALSE      TRUE   FALSE   FALSE  FALSE
## 7  FALSE  TRUE FALSE   TRUE   FALSE     FALSE    TRUE   FALSE  FALSE
## 8  FALSE  TRUE FALSE   TRUE   FALSE      TRUE    TRUE   FALSE  FALSE
##   NewLeagueN
## 1      FALSE
## 2      FALSE
## 3      FALSE
## 4      FALSE
## 5      FALSE
## 6      FALSE
## 7      FALSE
## 8      FALSE
```

```r
# show models
#sum$which
```

When running the subset selection algorithm using regsubsets and using BIC on log(salary), we find that the 3rd model is the best model since it has the lowest BIC. The predictors included in this model are Hits, Walks and Years.

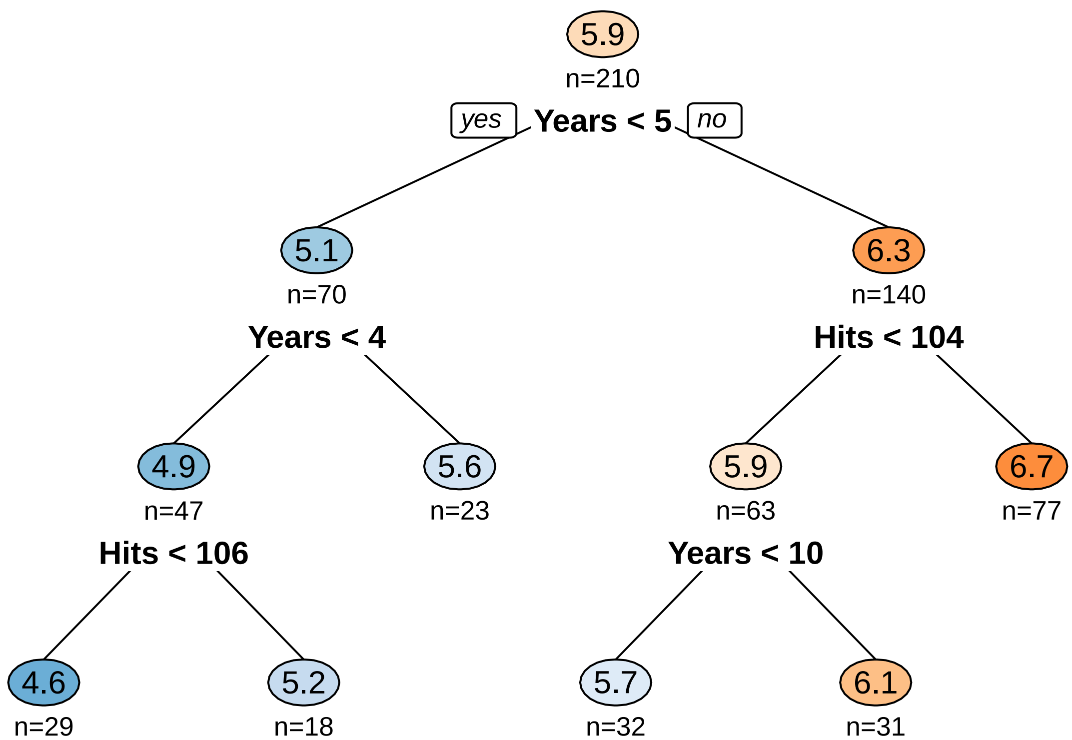
Answer 5

```r
library("data.table")
HittersModified.dt <- setDT(HittersModified.df)
```

```r
# **Split the data into training (80%) and validation/test set (20%)**
set.seed(42)
training.index <- sample(1:nrow(HittersModified.df), 0.8*(nrow(HittersModified.df)))
Hitters.train <- HittersModified.df[training.index, ]
Hitters.valid <- HittersModified.df[-training.index, ]
```

Answer 6

```r
# Generate regression tree
set.seed(42)
hitters.train.regtree <- rpart(Salary ~ Years + Hits, data = Hitters.train, method = "anova")


prp(hitters.train.regtree, type = 2,extra=1, under = TRUE, split.font = 2,
    varlen = -10, box.palette = "BuOr")
```




```r
rpart.rules(hitters.train.regtree, cover = TRUE ) # find rules
```

```
##  Salary                                            cover
##     4.6 when Years <  4       & Hits <  106          14%
##     5.2 when Years <  4       & Hits >= 106           9%
```

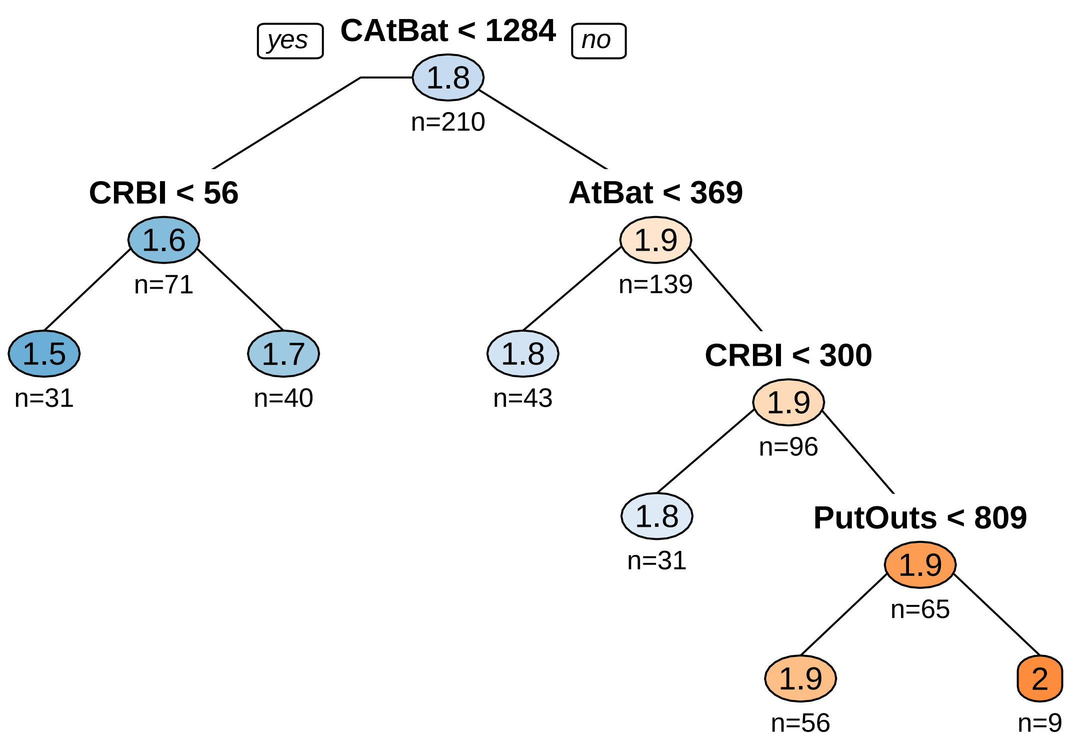
```
##      5.6 when Years is 4 to  5                       11%
##      5.7 when Years is 5 to 10 & Hits <  104         15%
##      6.1 when Years >=       10 & Hits <  104         15%
##      6.7 when Years >=        5 & Hits >= 104         37%
```

The players who have played atleast for 5 years and having hits greater than or equal to 104 are getting the highest salaries.

The rule is when Years >= 5 & Hits >= 104. 37% of the players receive highest salaries.

Answer 7

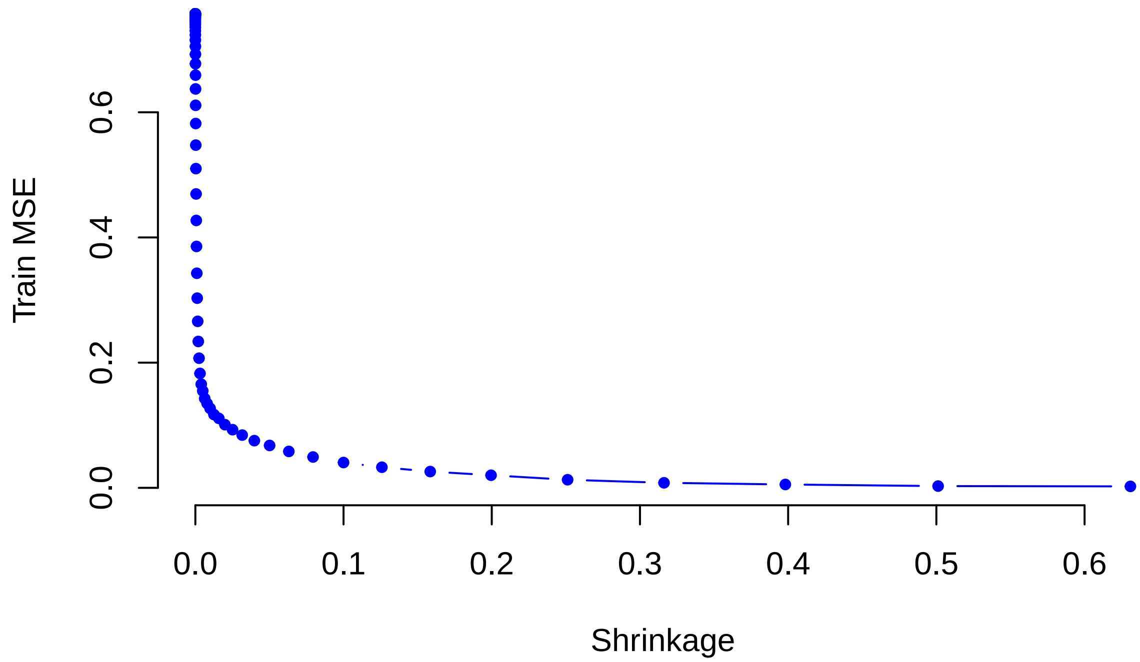
```r
set.seed(42)

# regression tree using all predictors
hitters.train.regtree.allpred <- rpart(log(Salary) ~ ., data = Hitters.train)

prp(hitters.train.regtree.allpred, type = 1,extra=1, under = TRUE, split.font = 2,
    varlen = -10, box.palette = "BuOr")
```



```r
rpart.rules(hitters.train.regtree.allpred, cover = TRUE) # find rules
```

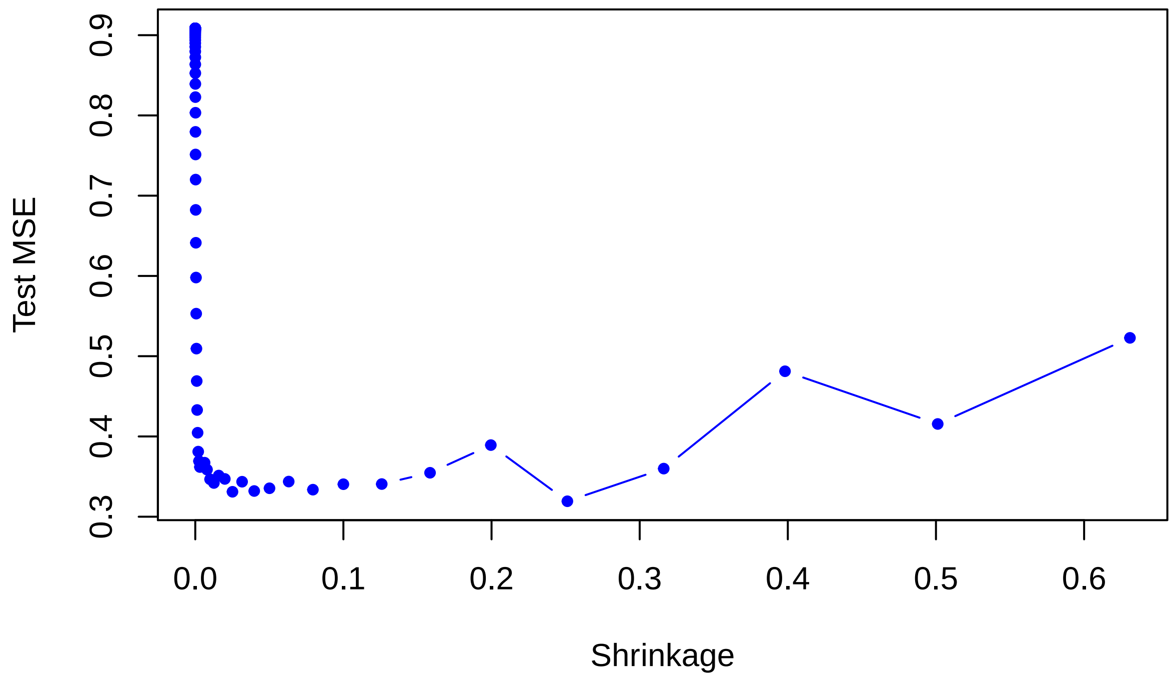
```
##  log(Salary)                                                                    cover
##          1.5 when CAtBat <  1284 & CRBI <   56                                   15%
##          1.7 when CAtBat <  1284 & CRBI >=  56                                   19%
##          1.8 when CAtBat >= 1284              & AtBat <  369                     20%
##          1.8 when CAtBat >= 1284 & CRBI <  300 & AtBat >= 369                    15%
##          1.9 when CAtBat >= 1284 & CRBI >= 300 & AtBat >= 369 & PutOuts <  809   27%
##          2.0 when CAtBat >= 1284 & CRBI >= 300 & AtBat >= 369 & PutOuts >= 809    4%
```

```r
pows <-  seq(-10, -0.2, by=0.1)
lambdas <-  10 ^ pows
length.lambdas <-  length(lambdas)
train.errors <-  rep(NA, length.lambdas)
test.errors <-  rep(NA, length.lambdas)

for (i in 1:length.lambdas) {
  boost.hitters <-  gbm(Salary ~ . , data=Hitters.train,
                        distribution="gaussian",
                        n.trees=1000,
                        shrinkage=lambdas[i])
  train.pred <-  predict(boost.hitters, Hitters.train, n.trees=1000)
  test.pred <-  predict(boost.hitters, Hitters.valid, n.trees=1000)
  train.errors[i] <-  mean((Hitters.train$Salary - train.pred)^2)
  test.errors[i] <-  mean((Hitters.valid$Salary - test.pred)^2)
}

plot(lambdas, train.errors, type="b",
     xlab="Shrinkage", ylab="Train MSE",
     col="Blue", pch=20, bty = "n")
```



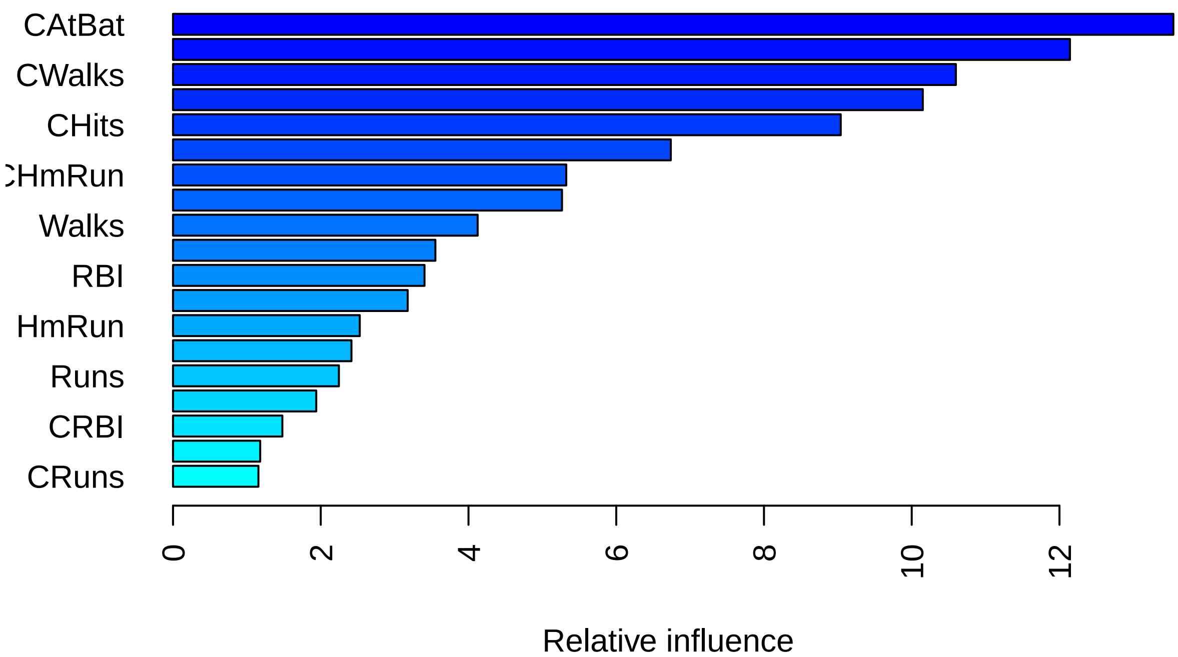
Answer 8

```
#For range of shrinkage values - test dataset
plot(lambdas, test.errors, type="b",
     xlab="Shrinkage", ylab="Test MSE",
     col="blue", pch=20)
```



Answer 9

```r
set.seed(42)
vboost.valid <- gbm(log(Salary)~., data=Hitters.valid, distribution = "gaussian", n.trees=1000)
summary(vboost.valid , las = 2)
```




```
##                 var   rel.inf
## CAtBat       CAtBat 13.541844
## Assists     Assists 12.142009
## CWalks       CWalks 10.597692
## Errors       Errors 10.149989
## CHits         CHits  9.039227
## AtBat         AtBat  6.738765
## CHmRun       CHmRun  5.323274
## PutOuts     PutOuts  5.265818
## Walks         Walks  4.123809
## Division   Division  3.551728
## RBI             RBI  3.404372
## Hits           Hits  3.176598
## HmRun         HmRun  2.528063
## League       League  2.415812
## Runs           Runs  2.245564
## NewLeague NewLeague  1.938999
## CRBI           CRBI  1.480379
## Years         Years  1.180096
## CRuns         CRuns  1.155961
```

CAtBat:13.541844, Assists:12.142009, CWalks:10.597692, Errors:10.149989 and CHits 9.039227 are the top 5 most important variables in the same order.

Answer 10

```
library(randomForest)
set.seed(42)
rf.hitters <-  randomForest(Salary ~ . , data=Hitters.train,
                            ntree=1000, mtry=19)
rf.pred <-  predict(rf.hitters, Hitters.valid)
mean((Hitters.valid$Salary - rf.pred)^2)
```

```
## [1] 0.2442542
```

The test set MSE value after applying bagging to the training dataset is 0.2442542.