# 624 HW 9

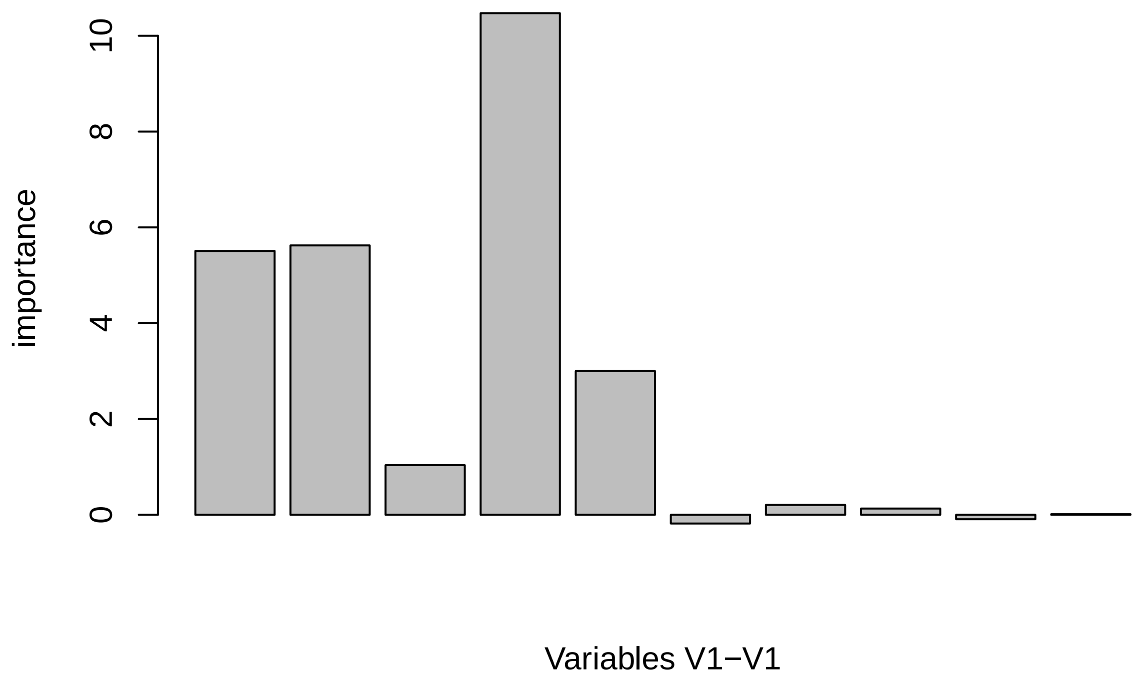
## Kuhn and Johnson Chapter 8

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## randomForest 4.6-14

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

## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin

##
## Attaching package: 'pre'

## The following object is masked from 'package:randomForest':
##
##     importance

## Loading required package: Rcpp

## Loading required package: rlang
```

### Problem 8.1

```
##          V1        V2        V3        V4        V5        V6        V7
## 1 0.1965959 0.8897369 0.4034173 0.9335958 0.7343655 0.3080857 0.7300751
## 2 0.7164260 0.1839942 0.8771072 0.5623151 0.1027748 0.2279233 0.6644855
## 3 0.3620857 0.7163158 0.4601120 0.2225171 0.8524531 0.3392558 0.9341949
## 4 0.3910775 0.2375733 0.5848327 0.2497158 0.7292472 0.3881883 0.5560306
## 5 0.8133072 0.3541920 0.6959593 0.5953801 0.7285362 0.9964300 0.6814503
## 6 0.4279599 0.1889772 0.3961759 0.3743723 0.5802106 0.2530935 0.3974592
##           V8        V9       V10         y
## 1 0.24125356 0.9239784 0.4639425 18.210088
## 2 0.98578585 0.4530842 0.4231528 14.829633
## 3 0.21314727 0.7413460 0.7615259 13.886334
## 4 0.02176553 0.6763241 0.2649187  9.044441
## 5 0.94450272 0.9189753 0.1579674 19.844821
## 6 0.09120397 0.6593991 0.2463080  8.882691
```

**a**

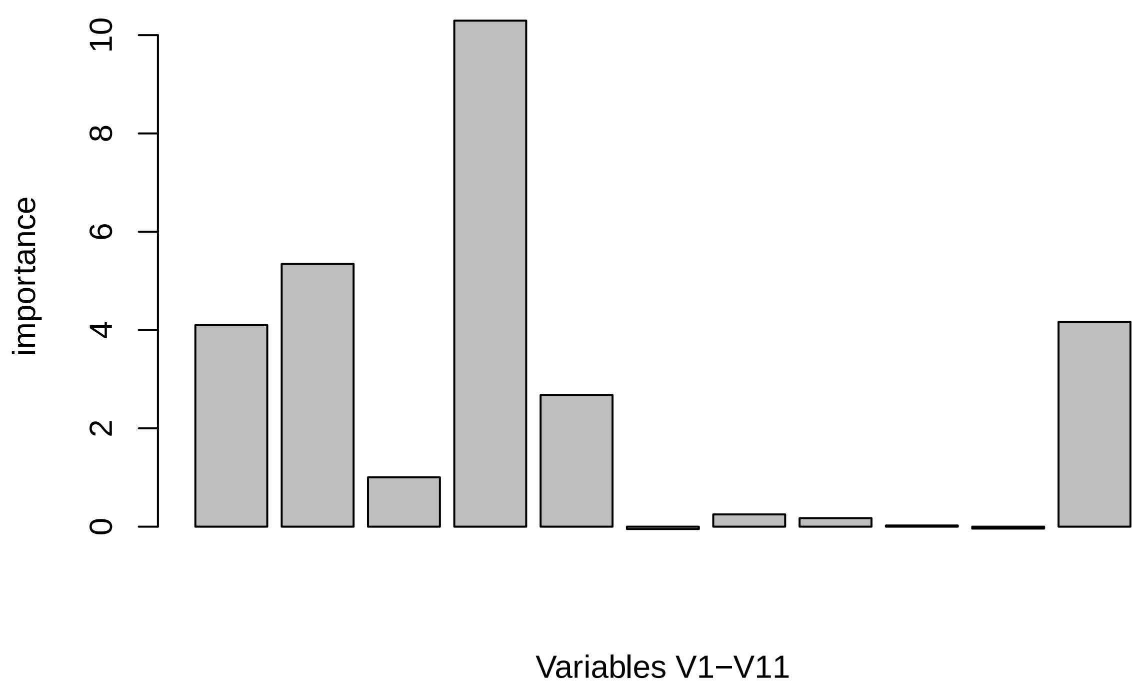



```
##       [,1]
##  [1,]  0.7
##  [2,]  1.9
##  [3,]  3.1
##  [4,]  4.3
##  [5,]  5.5
```

```
##  [6,]  6.7
##  [7,]  7.9
##  [8,]  9.1
##  [9,] 10.3
## [10,] 11.5
```

As we can see in the above chart, only variables 1-5 significantly effected the model.

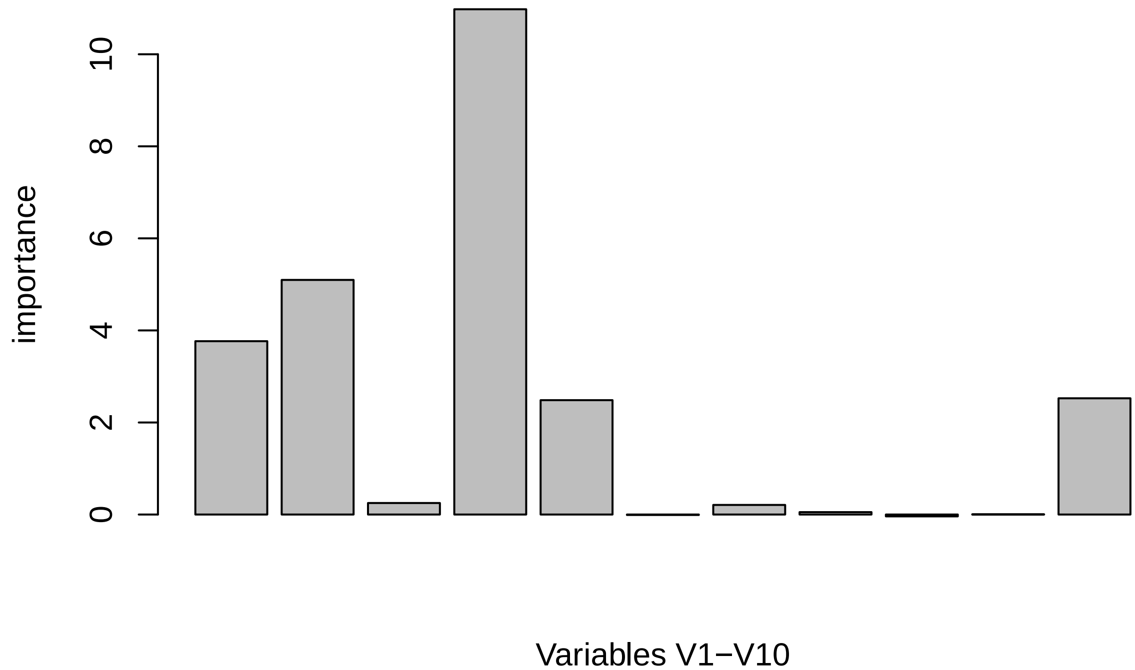
**b**




```
##       [,1]
##  [1,]  0.7
##  [2,]  1.9
##  [3,]  3.1
##  [4,]  4.3
##  [5,]  5.5
##  [6,]  6.7
##  [7,]  7.9
##  [8,]  9.1
##  [9,] 10.3
## [10,] 11.5
## [11,] 12.7
```

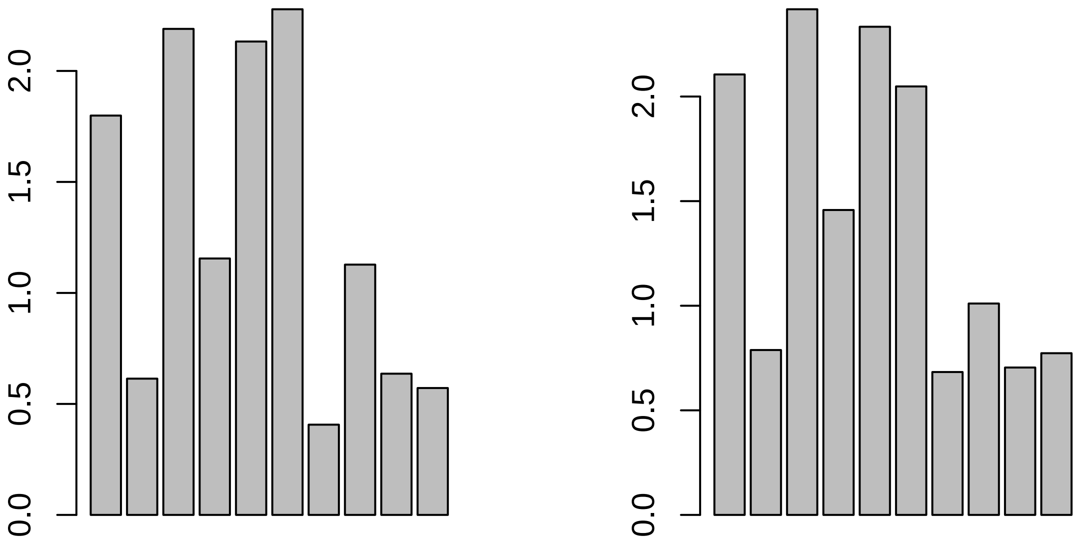
Next, I added an 11th variable that was highly correlated with V1. As we can see from the graph, it is a significant indicator. Additionally, it reduced the importace of V1 because V11 explains some of its variance.

**c**

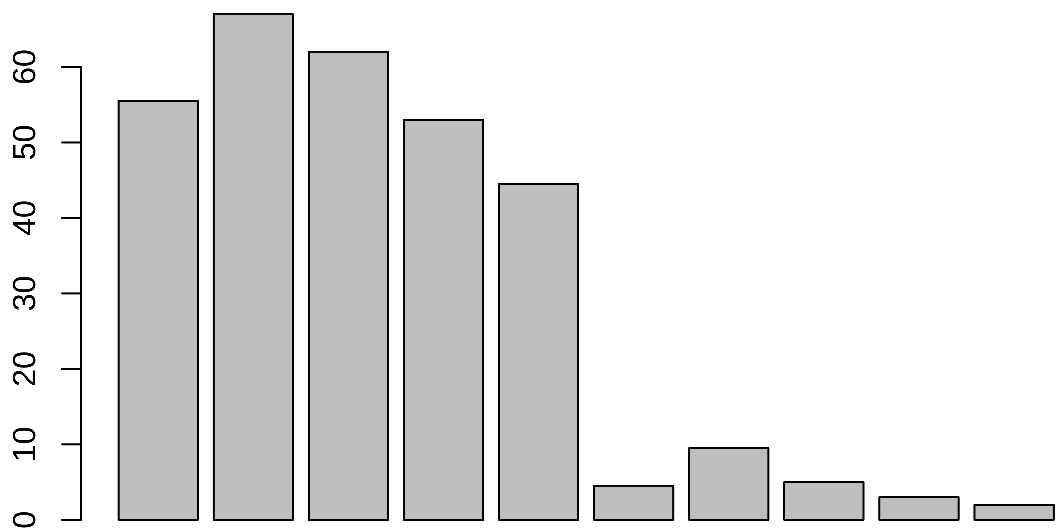


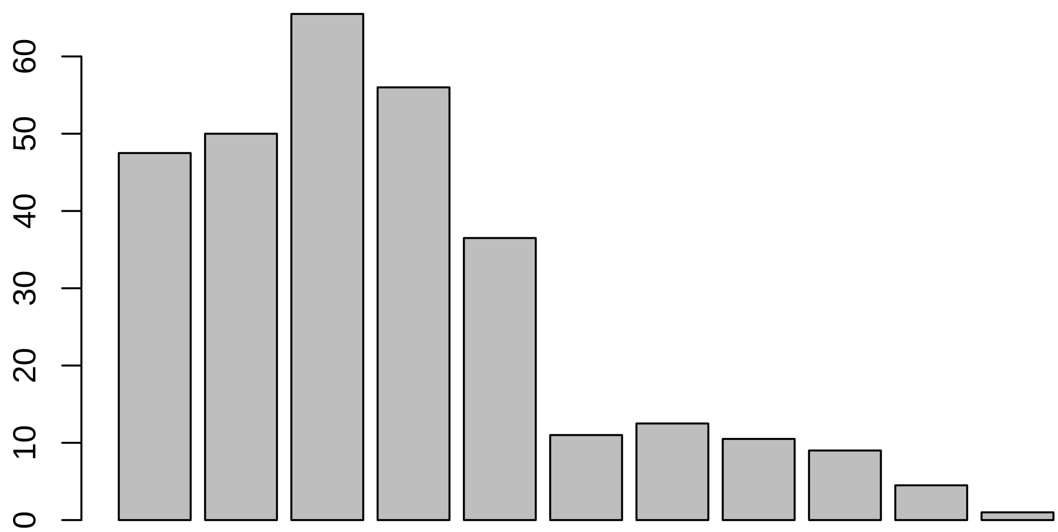

Using conditional trees, we find that fewer predictors are necessary. Namely, variable 3 becomes basically irrelevant.

**d**




Bagging seems to reduce the importance of all the previously important variables and likewise boosting the influence of the previously discarded variables (because explained variance must all equal 1).


0
10
20
30
40
50
60




Cubist tree building, however, gives the results we'd expect from above with the additional benefit of eliminating the importance of the dependent variable (V11).

## Problem 8.2

```
##          Overall
## High   5.1044426
## Low    0.1013638
## Middle 0.7867465
```

We can see that as granularity increases, importance tends to decrease. That is, as the range of our values increases, so does its importance.

## Problem 8.3

### a

By reducing the bagging fraction, variables with less explanatory power tend to be modelled separately from the more important variables. Conversely, there are fewer opportunities for models to be constructed from the traditionally important vectors. As learning rate increases, the marginal effect of a new tree on the model is increased– leading to a higher correlation factor. When a learning factor of .1 is used, each additional tree has less effect on the model than a learning factor of .2. This means more trees are needed, but reduces the likelihood of overfitting. That means the relationship bweetn learning rate and the number of trees is inverse.

### b

The model using a learning rate of .9 would have a tendency to overfit as each additional tree has a large marginal effect.

### c

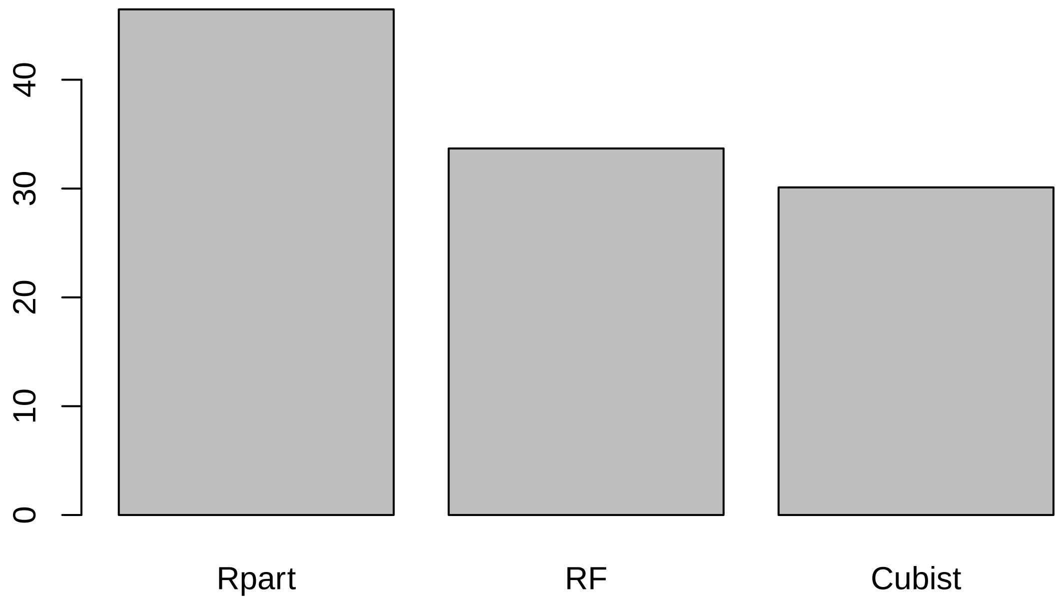
Interaction depth refers to the tree depth and the number of leaf nodes. As tree depth increases, the number of leaf nodes tends to increase leading to more data vectors coming into play. In both cases, we'd prefer a more uniform distribution of importance.

## Problem 8.7

### a

```
##
## Attaching package: 'imputeTS'

## The following object is masked from 'package:zoo':
##
##     na.locf

##
## Call:
## summary.resamples(object = resampling)
##
## Models: SingleTree, RandomForest, Cubist
## Number of resamples: 10
##
## MAE
##                   Min.   1st Qu.    Median      Mean   3rd Qu.     Max.
## SingleTree   0.8349514 0.9570279 1.1142687 1.1233198 1.2185423 1.508726
## RandomForest 0.5873856 0.8306705 0.8875993 0.8703800 0.9222134 1.208150
## Cubist       0.4685895 0.6423180 0.6659518 0.7688176 0.9578148 1.053937
##              NA's
## SingleTree      0
## RandomForest    0
## Cubist          0
##
## RMSE
##                   Min.   1st Qu.    Median     Mean  3rd Qu.     Max. NA's
## SingleTree   1.0154013 1.1992917 1.3977413 1.408004 1.484750 2.003905    0
## RandomForest 0.7767842 1.0243063 1.0933312 1.106106 1.209931 1.593887    0
## Cubist       0.6580057 0.7795592 0.8716059 0.972782 1.233067 1.294626    0
##
## Rsquared
##                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## SingleTree   0.1411297 0.3028180 0.4234355 0.4348887 0.5120323 0.7737295
## RandomForest 0.4366632 0.5791650 0.6848531 0.6716643 0.7870188 0.9048642
## Cubist       0.4150874 0.6130897 0.7349109 0.7137045 0.8201955 0.9359714
##              NA's
## SingleTree      0
```

```
## RandomForest    0
## Cubist          0
```




As we can see, the random forest model performs the best when using RMSE as the indicator, but does not beat the cubist model by much. ### b

```
## rpart2 variable importance
##
##   only 20 most important variables shown (out of 57)
##
##                          Overall
## ManufacturingProcess17  100.00
## ManufacturingProcess09   89.24
## ManufacturingProcess11   77.08
## BiologicalMaterial12     58.89
## ManufacturingProcess32   49.93
## BiologicalMaterial06     46.26
## BiologicalMaterial02     45.43
## ManufacturingProcess18   43.02
## ManufacturingProcess02   40.35
## BiologicalMaterial05     40.21
## ManufacturingProcess21   36.75
## BiologicalMaterial04     35.67
## ManufacturingProcess31   29.90
## ManufacturingProcess06   29.62
## BiologicalMaterial03     28.87
## ManufacturingProcess25   24.66
```

```
## ManufacturingProcess13   24.51
## BiologicalMaterial10     21.63
## BiologicalMaterial01     20.68
## BiologicalMaterial11     17.67
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 57)
##
##                        Overall
## ManufacturingProcess32  100.00
## BiologicalMaterial12     65.58
## BiologicalMaterial03     59.69
## ManufacturingProcess09   58.64
## BiologicalMaterial06     55.15
## ManufacturingProcess13   46.94
## BiologicalMaterial02     46.66
## ManufacturingProcess17   46.58
## ManufacturingProcess31   45.96
## BiologicalMaterial11     43.85
## ManufacturingProcess36   43.65
## BiologicalMaterial04     42.10
## BiologicalMaterial08     40.86
## BiologicalMaterial01     40.21
## ManufacturingProcess06   39.78
## BiologicalMaterial05     39.64
## BiologicalMaterial09     37.94
## ManufacturingProcess01   36.37
## ManufacturingProcess33   35.96
## ManufacturingProcess11   35.10
```

```
## cubist variable importance
##
##   only 20 most important variables shown (out of 57)
##
##                        Overall
## ManufacturingProcess32  100.00
## ManufacturingProcess09   49.19
## ManufacturingProcess17   46.77
## BiologicalMaterial03     39.52
## ManufacturingProcess29   30.65
## BiologicalMaterial02     30.65
## BiologicalMaterial06     27.42
## ManufacturingProcess22   25.81
## ManufacturingProcess04   23.39
## ManufacturingProcess34   18.55
## ManufacturingProcess27   15.32
## BiologicalMaterial08     14.52
## ManufacturingProcess26   14.52
## ManufacturingProcess01   14.52
## BiologicalMaterial01     13.71
## ManufacturingProcess24   12.90
## BiologicalMaterial10     12.10
## ManufacturingProcess45   11.29
```

```
## BiologicalMaterial12     11.29
## BiologicalMaterial04     10.48
```

By looking at the above summaries, we can see that Rpart and Cubist models have similar slopes of the importance curve where the random forest model has a much more shallow importance curve. Additionally, Manufacturing processes dominated the models over biological ones.

### c

Below, the optimal tree is described–using a single split of manufacting process 32 around the point .006. Likewise, we can confirm this by looking at the means of the yields of the respective subsets.

```
## n= 140
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 140 454.300300 40.18379
##    2) ManufacturingProcess32< 159.5 83 171.874100 39.26651
##      4) BiologicalMaterial12< 19.975 45  74.609320 38.57711
##        8) BiologicalMaterial05>=19.705 8   5.948750 36.99750 *
##        9) BiologicalMaterial05< 19.705 37  44.383230 38.91865
##         18) BiologicalMaterial05< 19.07 30  19.951230 38.63700
##           36) BiologicalMaterial09>=12.985 8   1.241487 37.94875 *
##           37) BiologicalMaterial09< 12.985 22  13.542240 38.88727
##             74) ManufacturingProcess09< 44.995 8   0.640800 38.23000 *
##             75) ManufacturingProcess09>=44.995 14   7.470486 39.26286 *
##         19) BiologicalMaterial05>=19.07 7  11.853170 40.12571 *
##      5) BiologicalMaterial12>=19.975 38  50.551180 40.08289
##       10) ManufacturingProcess17>=33.85 24   9.976196 39.41542 *
##       11) ManufacturingProcess17< 33.85 14  11.552090 41.22714 *
##    3) ManufacturingProcess32>=159.5 57 110.898300 41.51947
##      6) ManufacturingProcess06< 208.1 33  63.766020 40.96515
##       12) ManufacturingProcess04< 933.5 23  30.529530 40.53826
##         24) ManufacturingProcess02>=18.5 16   8.787344 40.02687 *
##         25) ManufacturingProcess02< 18.5 7   7.993943 41.70714 *
##       13) ManufacturingProcess04>=933.5 10  19.404810 41.94700 *
##      7) ManufacturingProcess06>=208.1 24  23.049730 42.28167
##       14) ManufacturingProcess17>=33.45 15   9.368773 41.79867 *
##       15) ManufacturingProcess17< 33.45 9   4.349400 43.08667 *
```

```
## [1] "Mean of 'less than' yield"
```

```
## [1] NaN
```

```
## [1] "Mean of 'more than' yield"
```

```
## [1] 40.18379
```