Data Analytics: Titanic RPart, HClust, R Forest

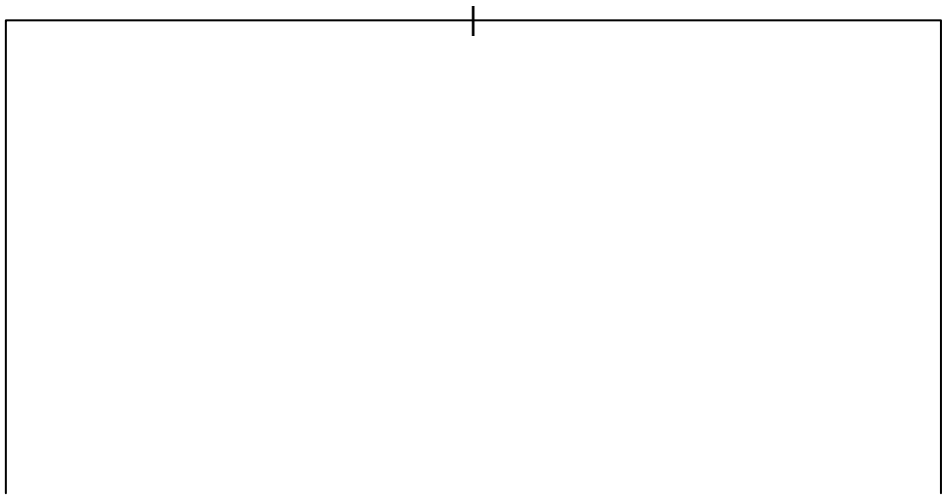
Brendan Donnelly

October 30, 2020

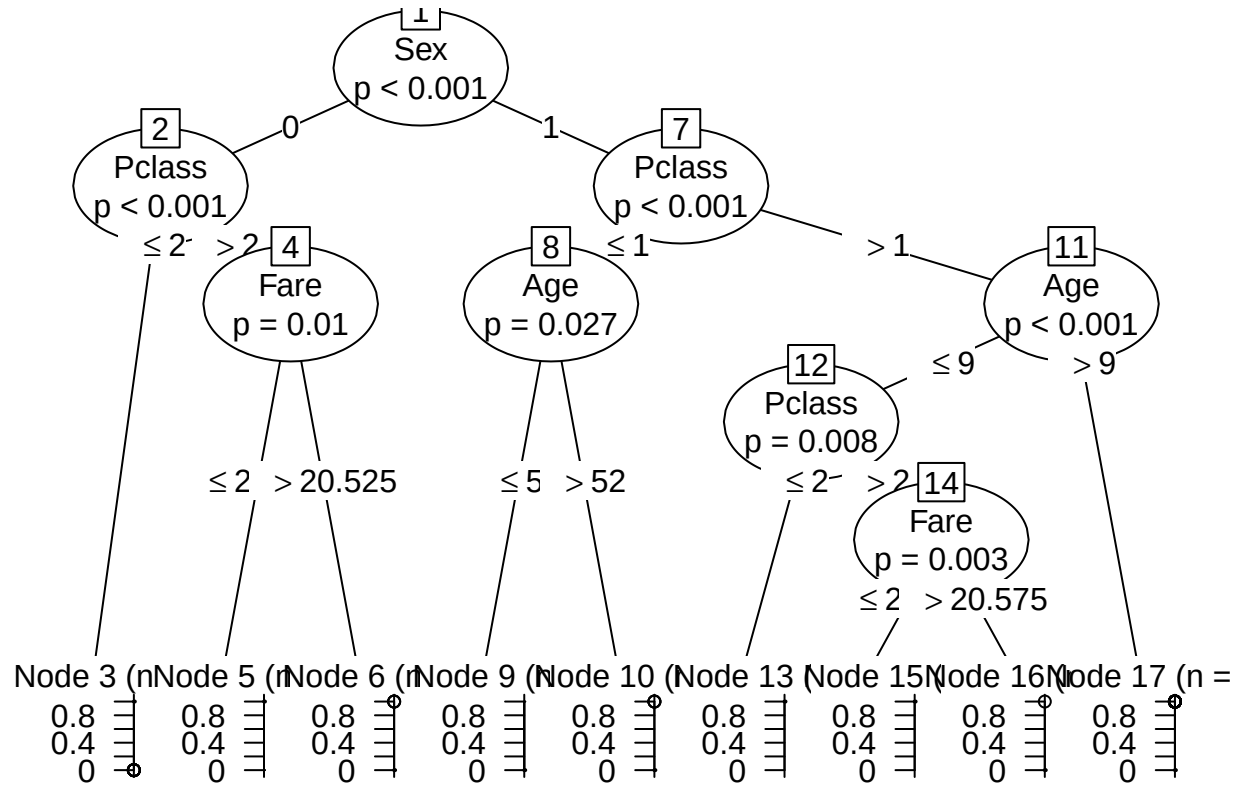
```
#to remove lists in env
#remove(list = ls())
library(ISLR)
library(tidyverse)
library(rpart)
library(tree)
library(party)
library(randomForest)
library(caret)
library(e1071)
library(titanic)


#Used data from Titanic library, came w/ titanic train and test set, no need for seed creation
data("titanic_train")
data("titanic_test")

#remove NA
titanic_train <- na.omit(titanic_train)
titanic_test <- na.omit(titanic_test)

#convert male, female to 1,0
titanic_train <- titanic_train %>%
  mutate(Sex = factor(Sex,
                      levels = c("male","female"),
                      labels = c(1,0)))
titanic_test <- titanic_test %>%
  mutate(Sex = factor(Sex,
                      levels = c("male","female"),
                      labels = c(1,0)))
```

RPart, Ctree

```
titanic_rpart <- rpart(Survived ~ ., data = titanic_train)
plot(titanic_rpart)
```



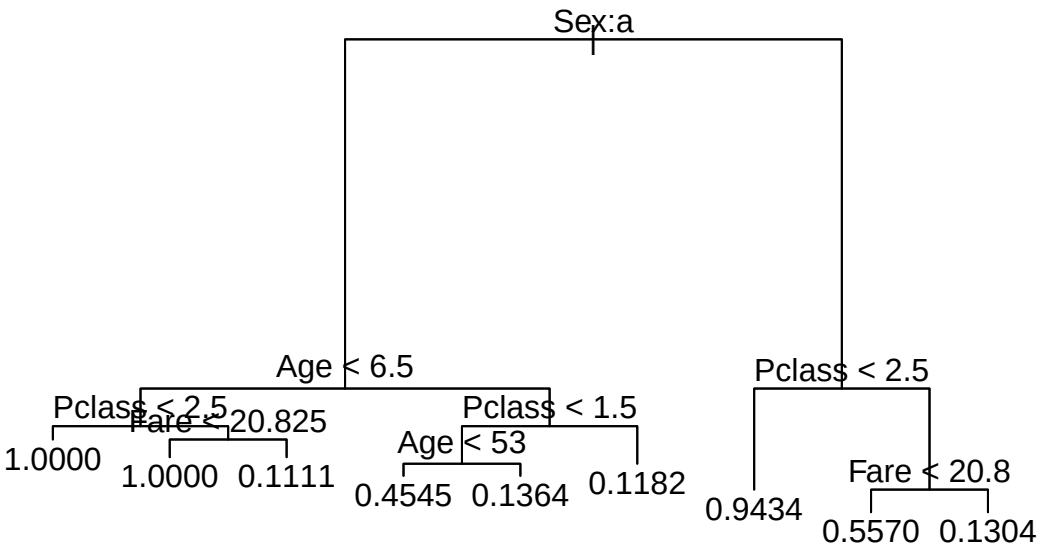
```r
require(party)
treeTitanic<- ctree(Survived~ Pclass + Age + Fare + Sex, data = titanic_train)
plot(treeTitanic)
```




```
tr<- tree(Survived~  Pclass + Age + Fare + Sex, data = titanic_train)
tr$frame
```

```
##       var   n         dev      yval splits.cutleft splits.cutright
## 1     Sex 714 172.2128852 0.4061625             :a              :b
## 2     Age 453  73.9072848 0.2052980           <6.5            >6.5
## 4  Pclass  24   5.3333333 0.6666667           <2.5            >2.5
## 8  <leaf>  10   0.0000000 1.0000000
## 9    Fare  14   3.4285714 0.4285714        <20.825         >20.825
## 18 <leaf>   5   0.0000000 1.0000000
## 19 <leaf>   9   0.8888889 0.1111111
## 5  Pclass 429  63.1794872 0.1794872           <1.5            >1.5
## 10    Age  99  23.4141414 0.3838384            <53             >53
## 20 <leaf>  77  19.0909091 0.4545455
## 21 <leaf>  22   2.5909091 0.1363636
## 11 <leaf> 330  34.3909091 0.1181818
## 3  Pclass 261  48.3065134 0.7547893           <2.5            >2.5
## 6  <leaf> 159   8.4905660 0.9433962
## 7    Fare 102  25.3431373 0.4607843          <20.8           >20.8
## 14 <leaf>  79  19.4936709 0.5569620
## 15 <leaf>  23   2.6086957 0.1304348
```

```
plot(tr)
text(tr)
```




```
cforest(Survived~  Pclass + Age + Fare + Sex, data = titanic_train, controls=cforest_control(mtry=2, mi
```

```
##
##   Random Forest using Conditional Inference Trees
##
## Number of trees:  500
##
## Response:  Survived
## Inputs:  Pclass, Age, Fare, Sex
## Number of observations:  714
```

```
#get tid of non numeric/int/dbls
train_y <- titanic_train[][2]
train_X <- titanic_train[c(-2,-4,-9, -11, -12)]

#convert y to number
train_y['Survived'] <- as.numeric(train_y['Survived'])


fitTitanic <- randomForest(Survived~ Pclass + Age + Fare + Sex, data = titanic_train, importance = TRUE
print(fitTitanic)   # view results
```

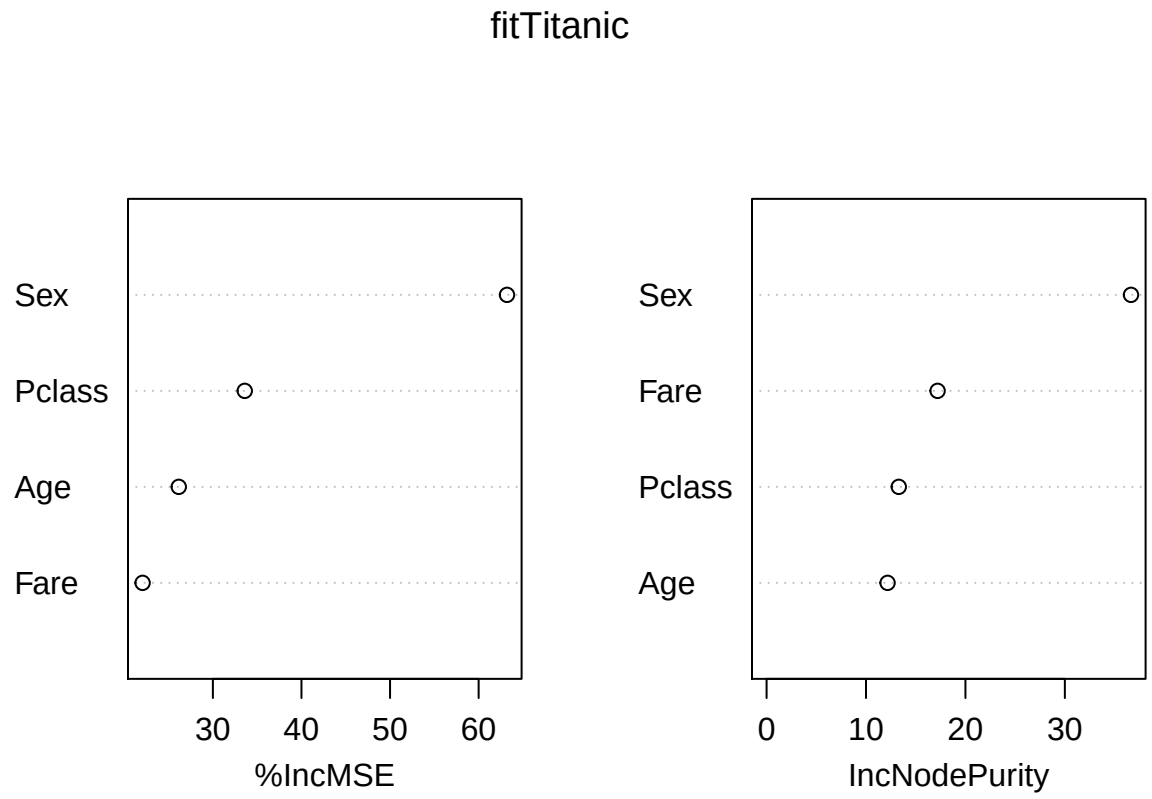
```
##
## Call:
##  randomForest(formula = Survived ~ Pclass + Age + Fare + Sex,      data = titanic_train, importance =
##                Type of random forest: regression
##                      Number of trees: 500
```

```
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.1358449
##                     % Var explained: 43.68
```

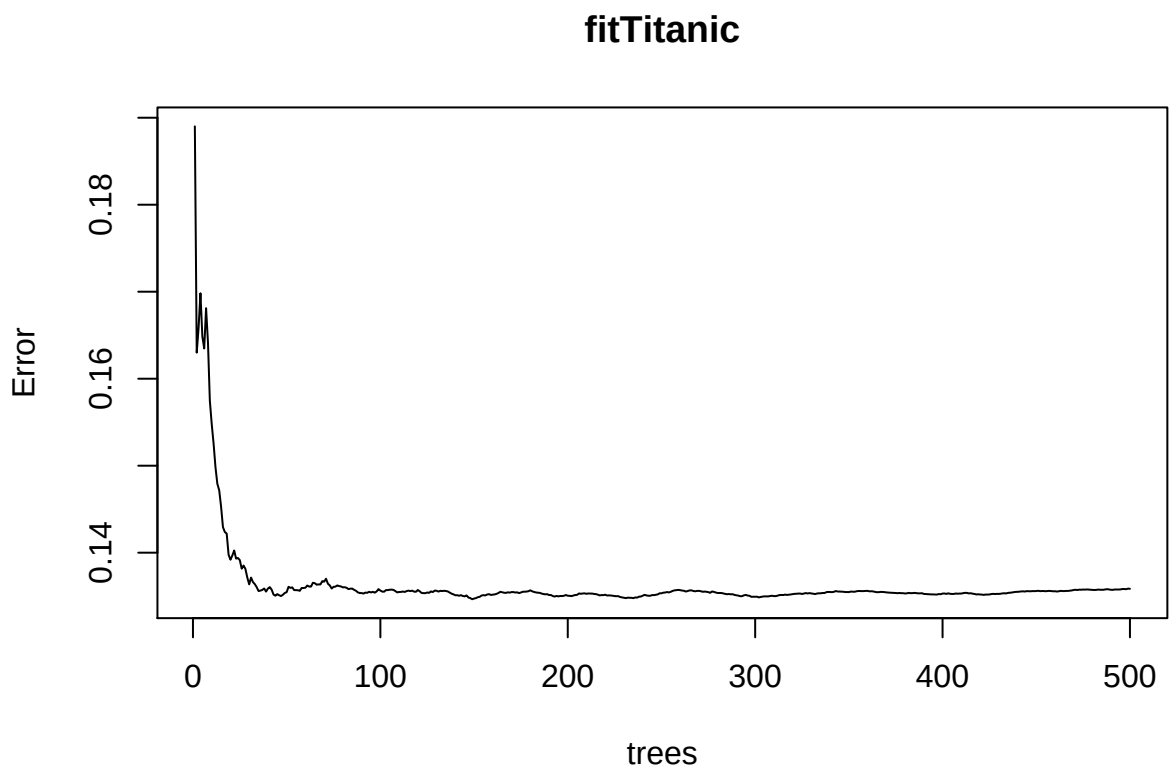
```
importance(fitTitanic) # importance of each predictor
```

```
##         %IncMSE IncNodePurity
## Pclass 33.59822      13.29555
## Age    26.15031      12.16976
## Fare   22.06555      17.20028
## Sex    63.20500      36.65367
```

```
varImpPlot(fitTitanic)
```




```
plot(fitTitanic)
```



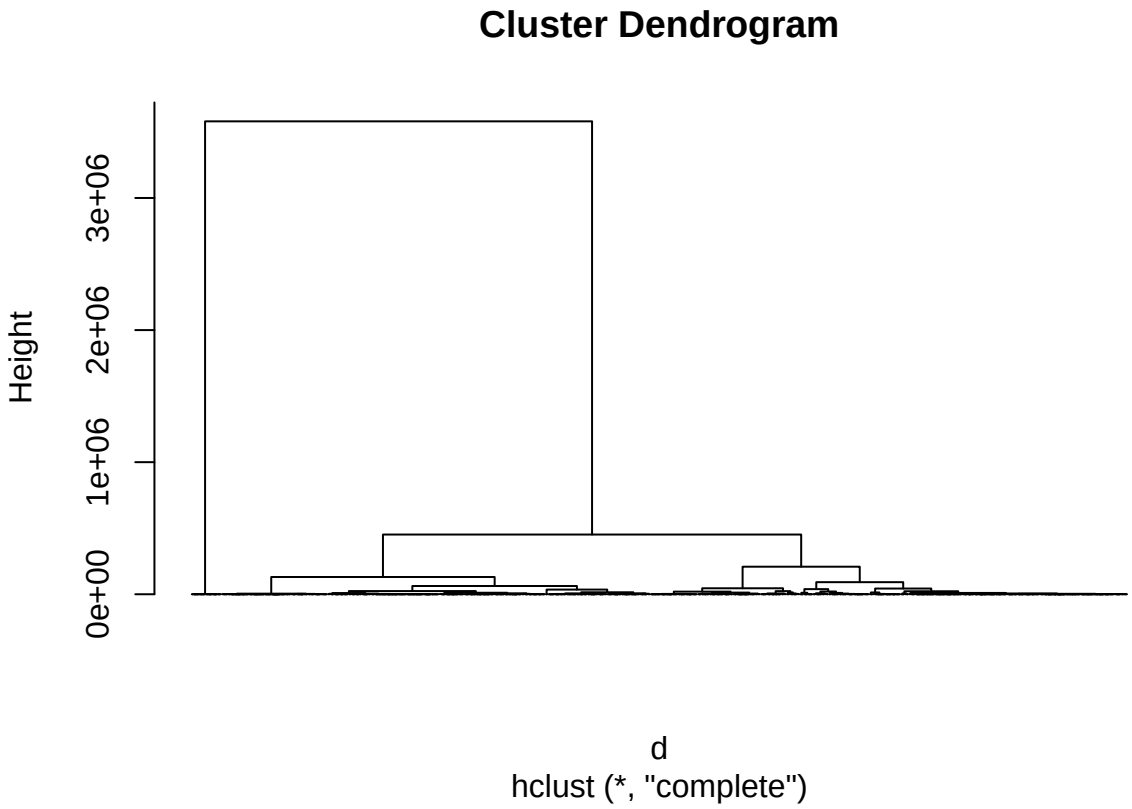

```
getTree(fitTitanic,1, labelVar=TRUE)
```

```
##    left daughter right daughter split var split point status    prediction
## 1              2              3      Fare    10.33540     -3  4.159664e-01
## 2              4              5      Fare     7.13335     -3  1.401869e-01
## 3              6              7       Age     7.50000     -3  5.340000e-01
## 4              8              9      Fare     2.50000     -3  6.666667e-02
## 5              0              0      <NA>     0.00000     -1  1.597633e-01
## 6             10             11       Age     1.50000     -3  8.125000e-01
## 7             12             13       Sex     1.00000     -3  5.044248e-01
## 8             14             15       Age    30.50000     -3  1.875000e-01
## 9             16             17      Fare     0.00000     -3 -2.220446e-16
## 10            18             19    Pclass     2.50000     -3  1.000000e+00
## 11            20             21       Sex     1.00000     -3  7.000000e-01
## 12             0              0      <NA>     0.00000     -1  2.212766e-01
## 13            22             23       Age    12.50000     -3  8.110599e-01
## 14             0              0      <NA>     0.00000     -1  7.500000e-01
## 15             0              0      <NA>     0.00000     -1 -5.551115e-17
## 16             0              0      <NA>     0.00000     -1  1.000000e+00
## 17            24             25       Age    48.00000     -3 -2.220446e-16
## 18            26             27    Pclass     1.50000     -3  1.000000e+00
## 19             0              0      <NA>     0.00000     -1  1.000000e+00
## 20            28             29      Fare    26.95000     -3  6.666667e-01
## 21            30             31       Age     3.50000     -3  7.222222e-01
## 22            32             33       Age     8.50000     -3  8.333333e-02
## 23            34             35    Pclass     2.50000     -3  8.536585e-01
```

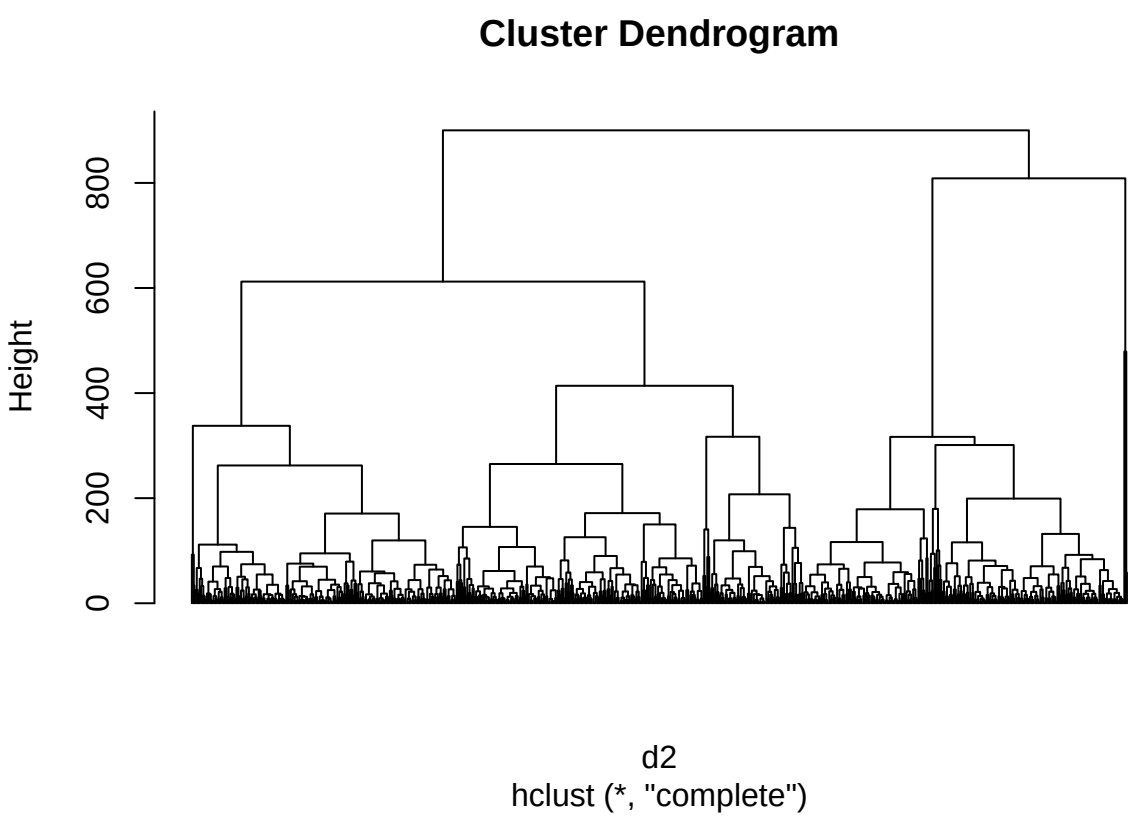
```
## 24            36             37    Pclass     0.00000     -3 -2.220446e-16
## 25             0              0      <NA>     0.00000     -1  0.000000e+00
## 26             0              0      <NA>     0.00000     -1  1.000000e+00
## 27            38             39      Fare    33.00210     -3  1.000000e+00
## 28             0              0      <NA>     0.00000     -1  1.000000e+00
## 29            40             41       Age     2.50000     -3  4.285714e-01
## 30            42             43       Age     2.50000     -3  2.857143e-01
## 31            44             45      Fare    20.64165     -3  1.000000e+00
## 32             0              0      <NA>     0.00000     -1  5.000000e-01
## 33             0              0      <NA>     0.00000     -1  0.000000e+00
## 34            46             47    Pclass     1.50000     -3  9.401198e-01
## 35            48             49      Fare    15.97500     -3  4.736842e-01
## 36             0              0      <NA>     0.00000     -1  0.000000e+00
## 37            50             51    Pclass     0.00000     -3 -2.220446e-16
## 38             0              0      <NA>     0.00000     -1  1.000000e+00
## 39             0              0      <NA>     0.00000     -1  1.000000e+00
## 40             0              0      <NA>     0.00000     -1  0.000000e+00
## 41             0              0      <NA>     0.00000     -1  6.000000e-01
## 42            52             53      Fare    26.95000     -3  3.333333e-01
## 43             0              0      <NA>     0.00000     -1  0.000000e+00
## 44             0              0      <NA>     0.00000     -1  1.000000e+00
## 45            54             55       Age     0.00000     -3  1.000000e+00
## 46             0              0      <NA>     0.00000     -1  9.882353e-01
## 47            56             57      Fare    20.25000     -3  8.902439e-01
## 48             0              0      <NA>     0.00000     -1  7.222222e-01
## 49             0              0      <NA>     0.00000     -1  2.500000e-01
## 50             0              0      <NA>     0.00000     -1  0.000000e+00
## 51             0              0      <NA>     0.00000     -1 -1.665335e-16
## 52             0              0      <NA>     0.00000     -1  1.000000e+00
## 53             0              0      <NA>     0.00000     -1  0.000000e+00
## 54             0              0      <NA>     0.00000     -1  0.000000e+00
## 55             0              0      <NA>     0.00000     -1  1.000000e+00
## 56            58             59      Fare    12.67500     -3  9.736842e-01
## 57             0              0      <NA>     0.00000     -1  8.181818e-01
## 58            60             61       Age    29.50000     -3  1.000000e+00
## 59            62             63      Fare    13.25000     -3  9.615385e-01
## 60             0              0      <NA>     0.00000     -1  1.000000e+00
## 61            64             65      Fare     0.00000     -3  1.000000e+00
## 62            66             67       Age    28.00000     -3  9.500000e-01
## 63            68             69       Age     0.00000     -3  1.000000e+00
## 64             0              0      <NA>     0.00000     -1  0.000000e+00
## 65             0              0      <NA>     0.00000     -1  1.000000e+00
## 66             0              0      <NA>     0.00000     -1  6.666667e-01
## 67             0              0      <NA>     0.00000     -1  1.000000e+00
## 68             0              0      <NA>     0.00000     -1  0.000000e+00
## 69             0              0      <NA>     0.00000     -1  1.000000e+00
```

```
# look at rfcv
#prediction <-predict(fitTitanic, train_X)
#confusionMatrix(prediction, train_y$Survived)
```

```
d<- dist(titanic_train, method = "euclidean")
hcl <- hclust(d, method = "complete")
plot(hcl, cex = 0.001, hang = -1)
```




```
#only factoring non-categorical , non-chars
d2<- dist(train_X, method = "euclidean")
hcl <- hclust(d2, method = "complete")
plot(hcl, cex = 0.001, hang = -1)
```


Cluster Dendrogram
Height
0
200
400
600
800
d2
hclust (*, "complete")