FinalProject_R_SuffolkDavid

David Suffolk

2/9/2020

Import Data

```
library(rpart)
library(rpart.plot)
```

Import Datasets

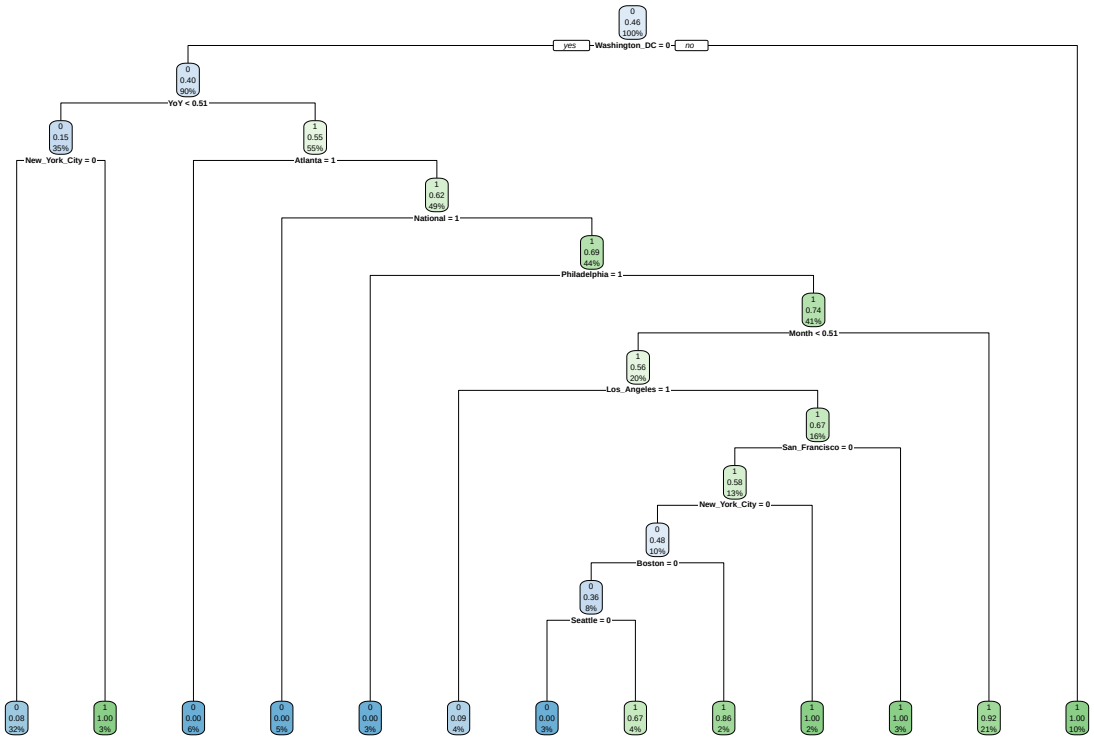
```
data_accounting = read.csv("df_Accountant.csv", header = TRUE)
data_RN = read.csv("df_RN.csv", header = TRUE)
data_DataAnalyst <- read.csv("df_Data_Analyst.csv", header=TRUE)
data_DataScientist <- read.csv("df_Data_Scientist.csv", header=TRUE)
data_FinancialAnalyst <- read.csv("df_Financial_Analyst.csv", header=TRUE)
data_GraphicDesigner <- read.csv("df_Graphic_Designer.csv", header=TRUE)
data_ITManager <- read.csv("df_IT_Manager.csv", header=TRUE)
data_Professor <- read.csv("df_Professor.csv", header=TRUE)
data_ProjectMgr <- read.csv("df_Project_Mgr.csv", header=TRUE)
data_Sales_Rep <- read.csv("df_Sales_Rep.csv", header=TRUE)
data_Software_Engineer <- read.csv("df_Software_Engineer.csv", header=TRUE)
data_Teacher <- read.csv("df_Teacher.csv", header=TRUE)
```

Function Building

```
decision_tree_job_title <- function(data){
  data_updated <- data.frame(data$Month,data$YoY,data$Atlanta,data$Boston,data$Chicago,data$Houston,data
  names(data_updated) <- c("Month","YoY","Atlanta","Boston","Chicago","Houston","Los_Angeles","National
  ran <- sample(1:nrow(data_updated), 0.9 * nrow(data_updated))
  nor <-function(x) { (x -min(x))/(max(x)-min(x)) }
  train_norm <- as.data.frame(lapply(data_updated, nor))
  data_train <- train_norm[ran,]
  data_test <- train_norm[-ran,]
  dtm <- rpart(Value_split~., data_train, method="class")
  rpart.plot(dtm, compress=TRUE, uniform=TRUE)
  p <- predict(dtm, data_test, type="class")
  confMat <- table(data_test$Value_split,p)
  accuracy <- sum(diag(confMat))/sum(confMat)
  return (accuracy*100)
}
```

Accountant

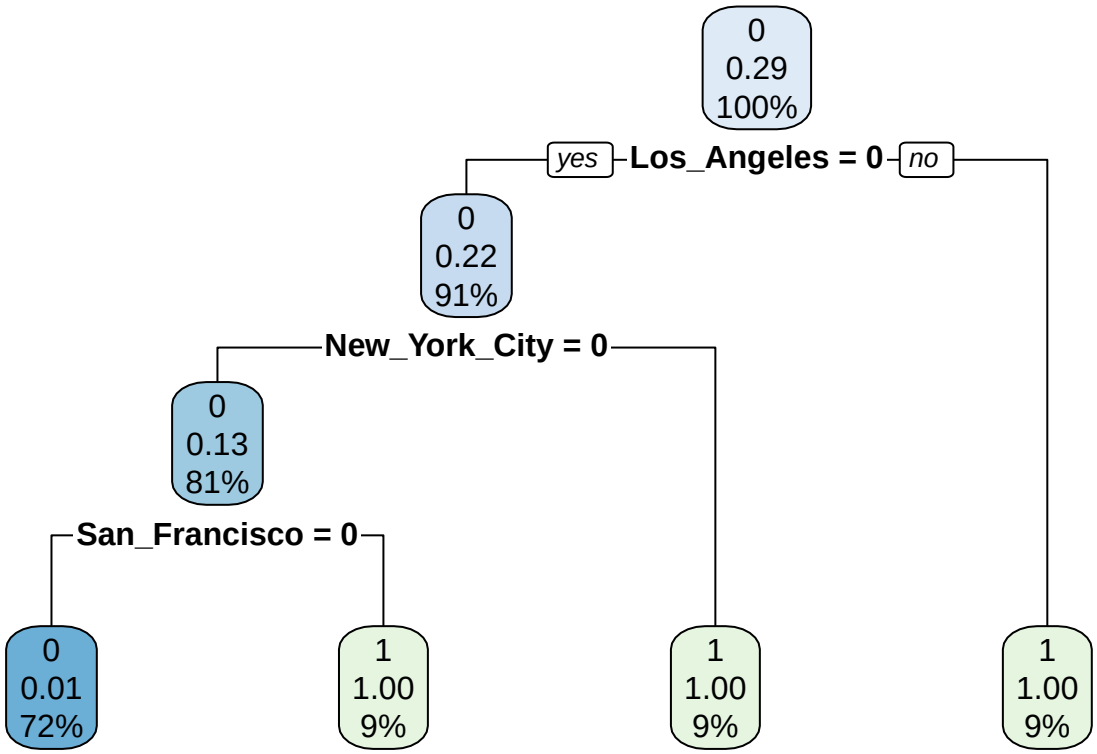
```
decision_tree_job_title(data_accounting)
```




```
## [1] 93.75
```

Registered Nurse (RN)

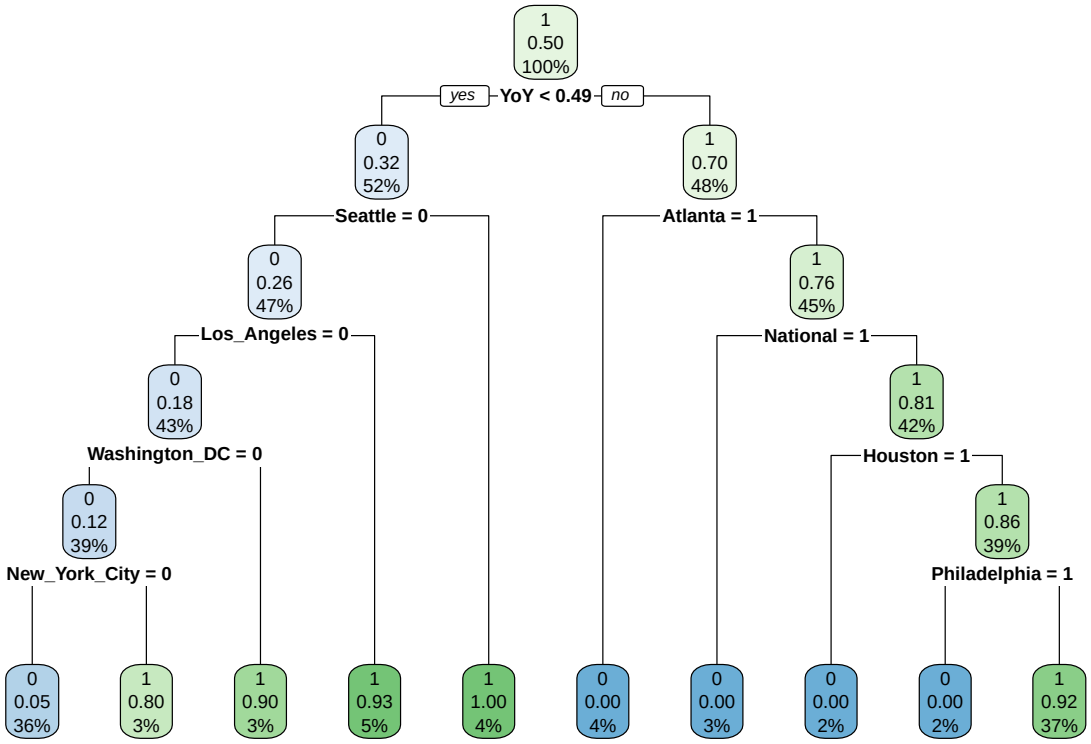
```
decision_tree_job_title(data_RN)
```




```
## [1] 100
```

Data Analyst

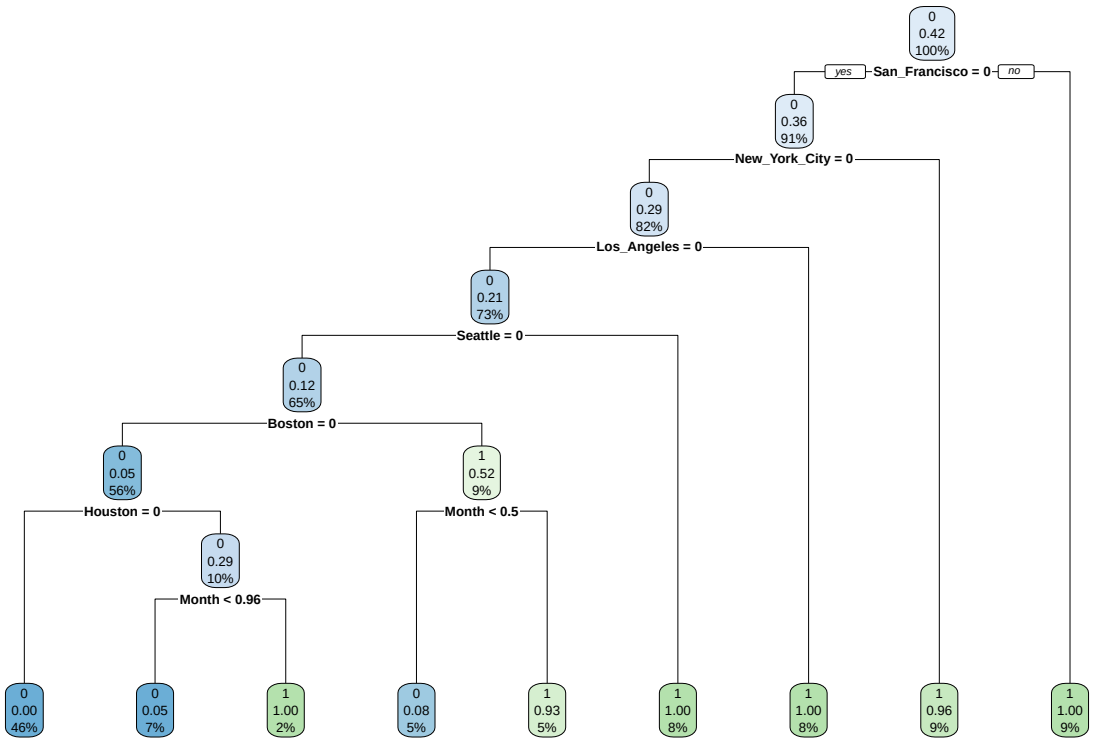
```
decision_tree_job_title(data_DataAnalyst)
```




```
## [1] 84.375
```

Data Scientist

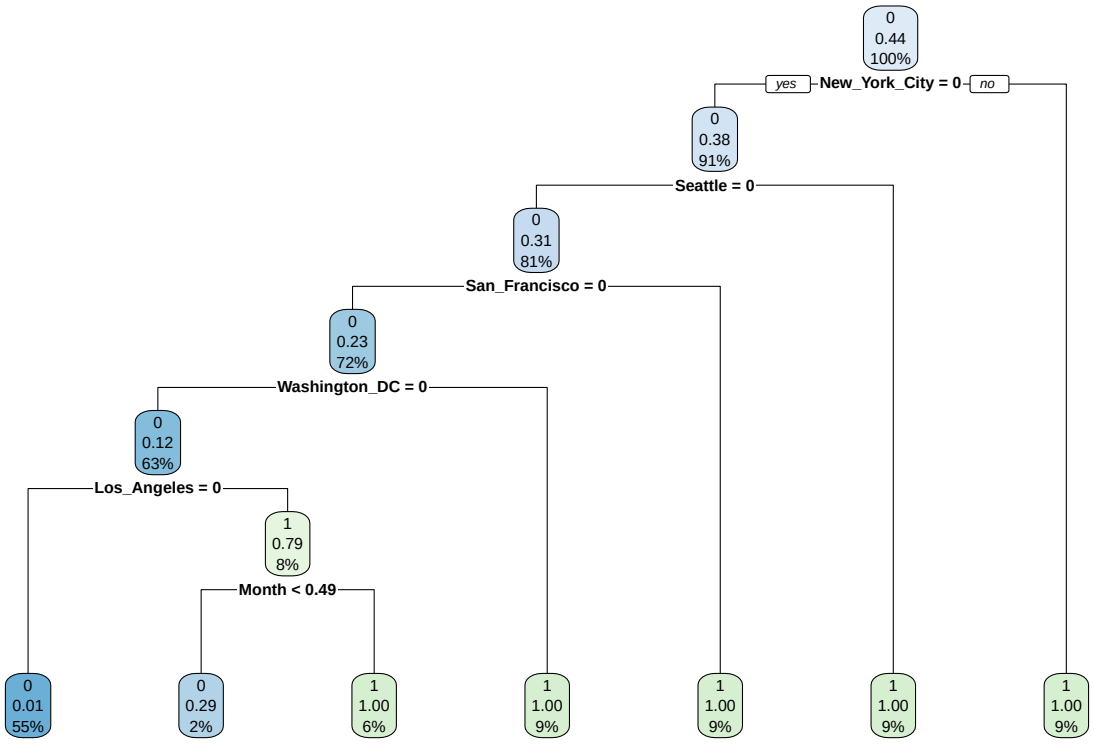
```
decision_tree_job_title(data_DataScientist)
```




```
## [1] 100
```

Financial Analyst

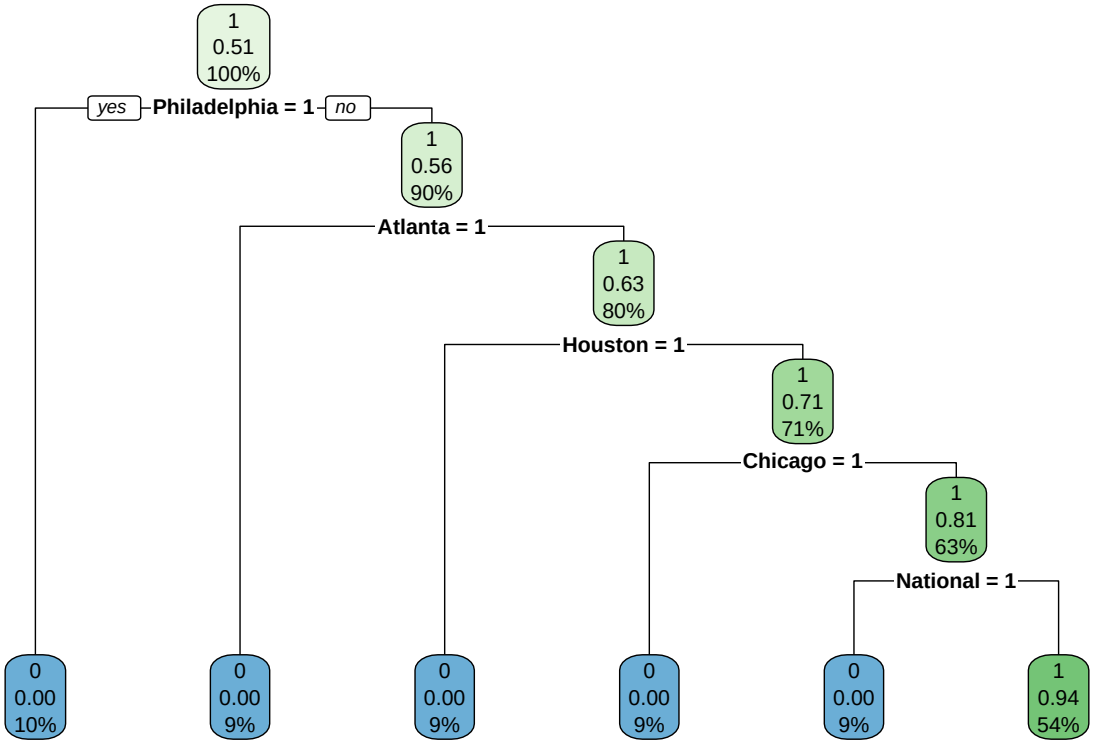
```
decision_tree_job_title(data_FinancialAnalyst)
```




```
## [1] 96.875
```

Graphic Designer

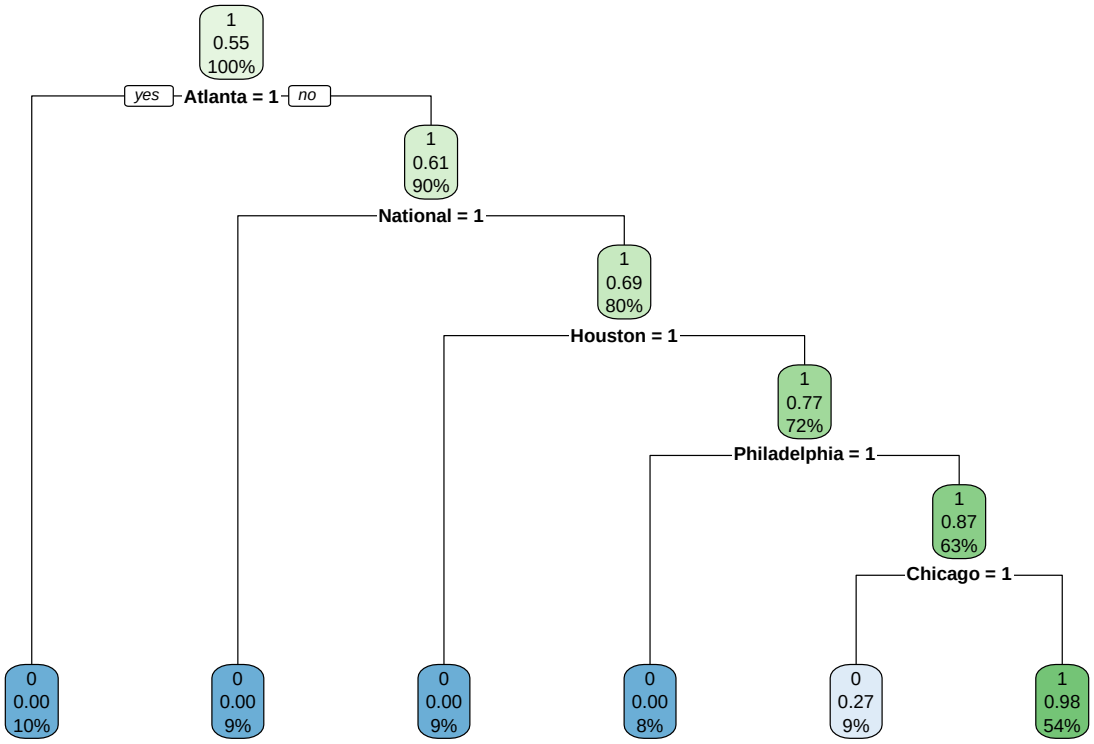
```
decision_tree_job_title(data_GraphicDesigner)
```




```
## [1] 93.75
```

IT Manager

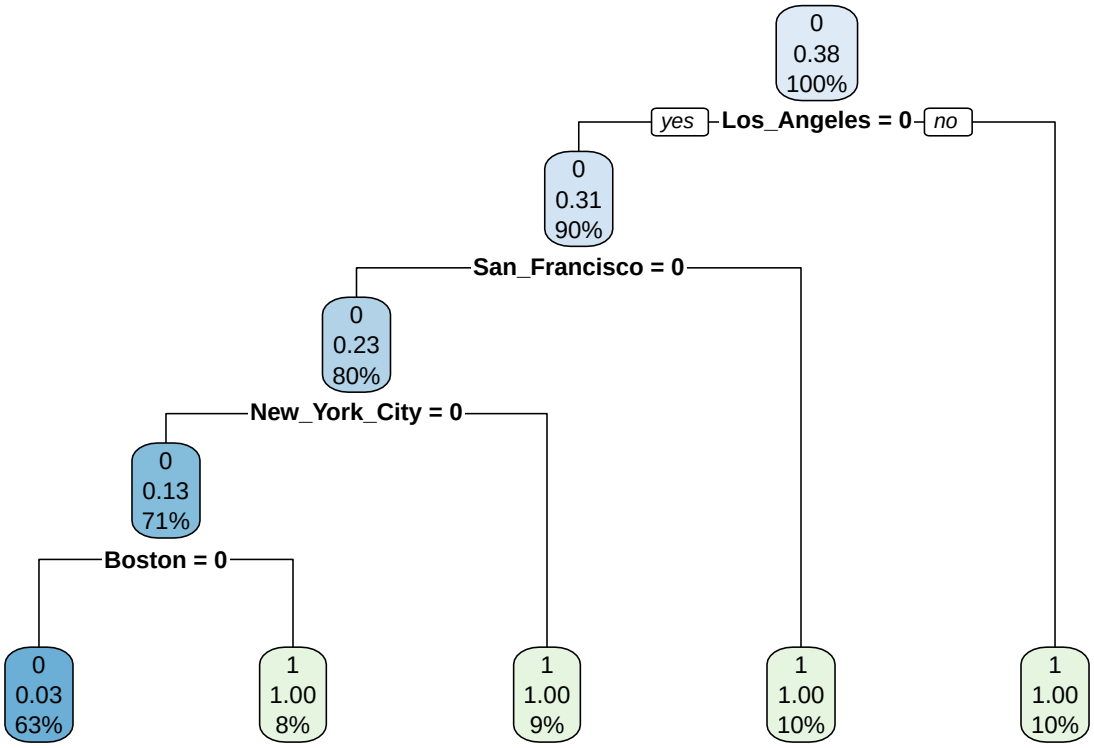
```
decision_tree_job_title(data_ITManager)
```




```
## [1] 100
```

Professor

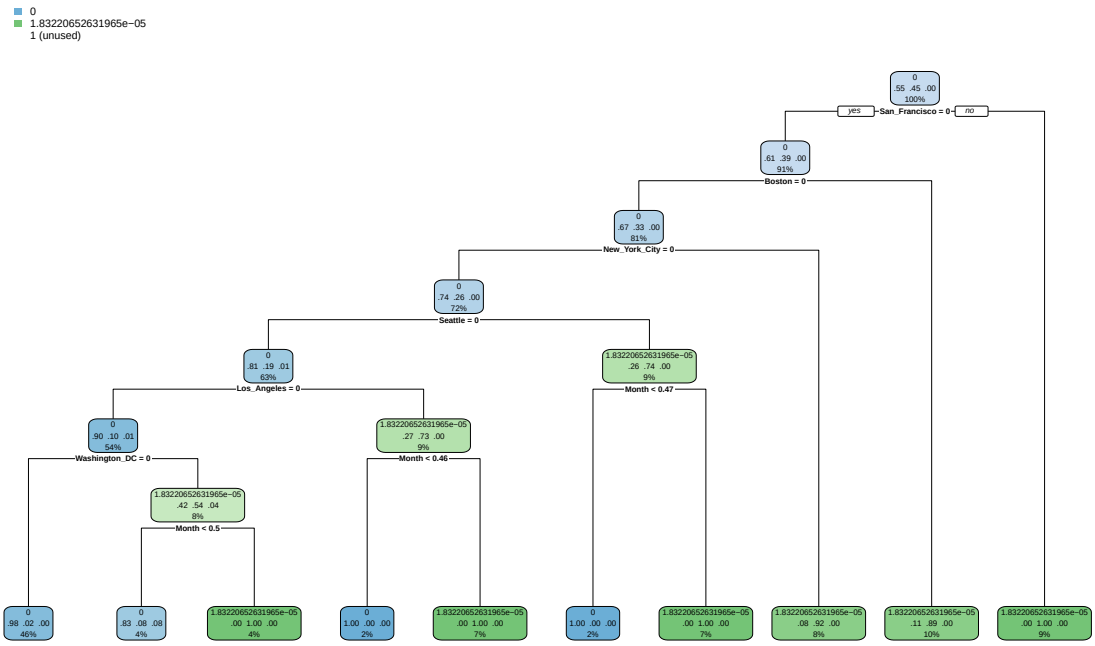
```
decision_tree_job_title(data_Professor)
```




```
## [1] 93.75
```

Sales Representative

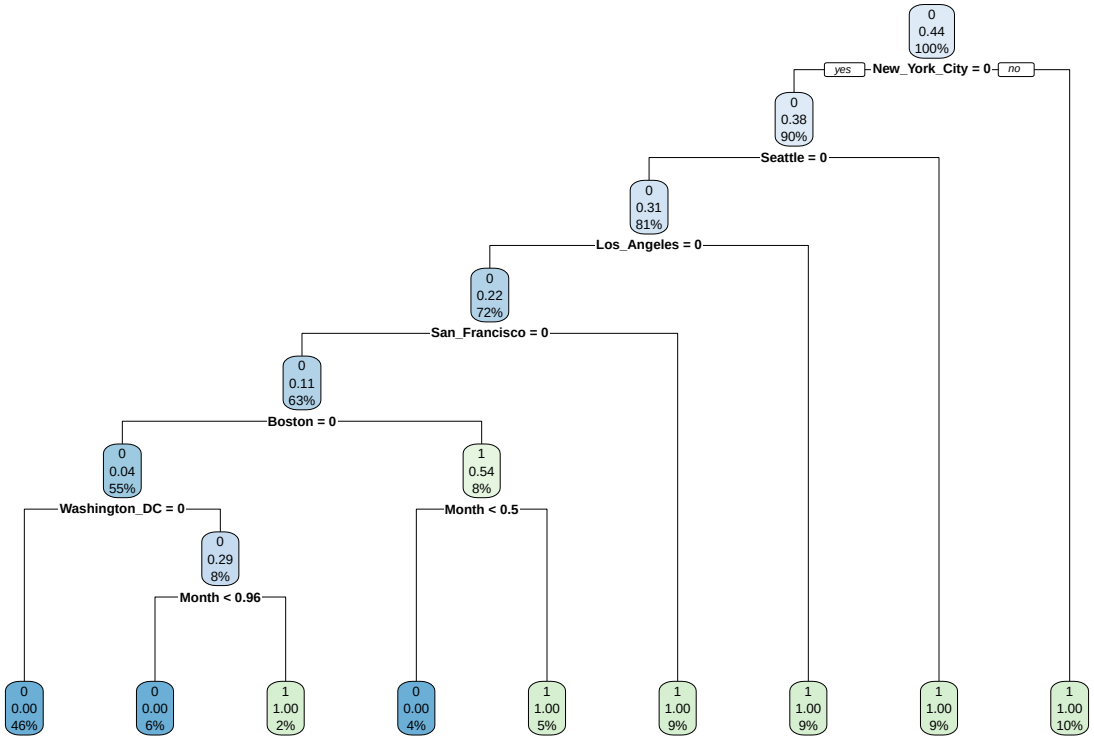
```
decision_tree_job_title(data_Sales_Rep)
```




```
## [1] 96.875
```

Software Engineer

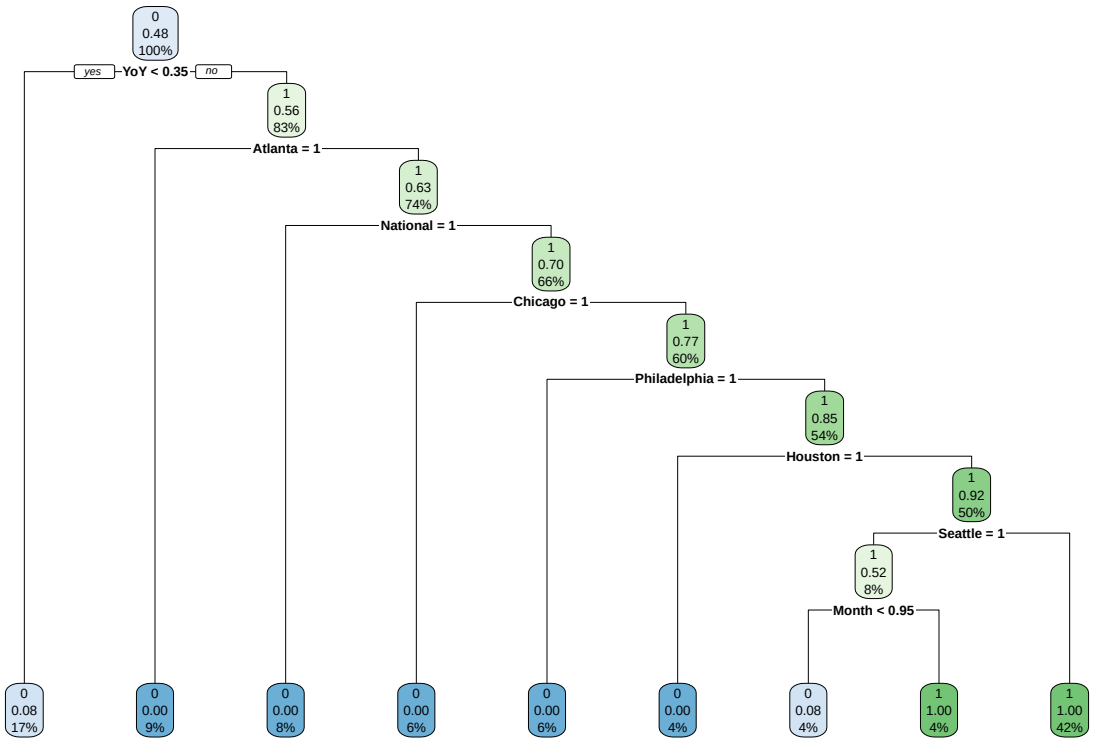
```
decision_tree_job_title(data_Software_Engineer)
```




```
## [1] 100
```

Teacher

```
decision_tree_job_title(data_Teacher)
```




```
## [1] 96.875
```

Data Scientist Bins

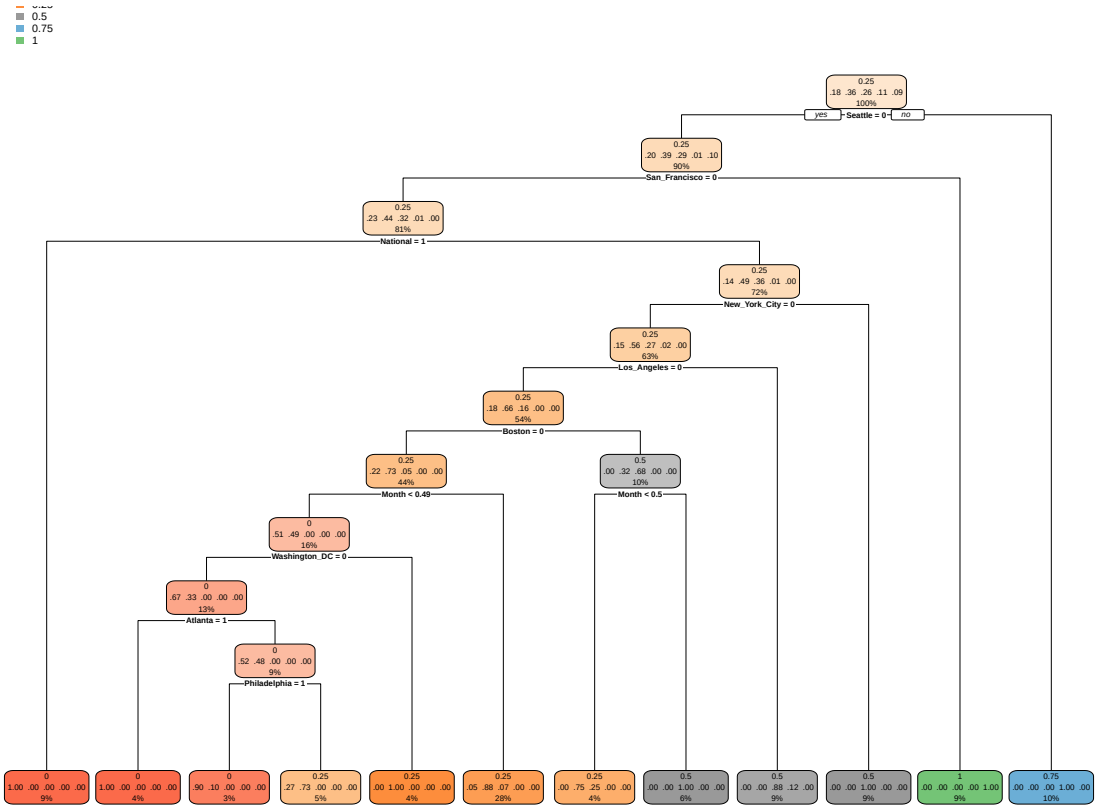
```
data_data_scientist_bins = read.csv("df_Data_Scientist_Bins.csv", header = TRUE)
head(data_data_scientist_bins)
```

```
##     X         Metro Dimension.Type  Month      Dimension         Measure
## 1  89      National      Job Title 201707 Data Scientist Median Base Pay
## 2 173       Atlanta      Job Title 201707 Data Scientist Median Base Pay
## 3 257 New York City      Job Title 201707 Data Scientist Median Base Pay
## 4 341   Los Angeles      Job Title 201707 Data Scientist Median Base Pay
## 5 425  Philadelphia      Job Title 201707 Data Scientist Median Base Pay
## 6 509       Houston      Job Title 201707 Data Scientist Median Base Pay
##    Value YoY Atlanta Boston Chicago Houston Los.Angeles National
## 1  95217 2.0       0      0       0       0           0        1
## 2  93069 1.6       1      0       0       0           0        0
## 3 112474 2.2       0      0       0       0           0        0
## 4 114061 1.3       0      0       0       0           1        0
## 5  94540 2.1       0      0       0       0           0        0
## 6 110859 0.6       0      0       0       1           0        0
##   New.York.City Philadelphia San.Francisco Seattle Washington.DC
## 1             0            0             0       0             0
```

```
## 2             0            0             0       0             0
## 3             1            0             0       0             0
## 4             0            0             0       0             0
## 5             0            1             0       0             0
## 6             0            0             0       0             0
##   Value_split Value_bins
## 1           0          0
## 2           0          0
## 3           1          2
## 4           1          2
## 5           0          0
## 6           0          2
```

```
decision_tree_job_title_bins <- function(data){
  data_updated <- data.frame(data$Month,data$YoY,data$Atlanta,data$Boston,data$Chicago,data$Houston,data
  names(data_updated) <- c("Month","YoY","Atlanta","Boston","Chicago","Houston","Los_Angeles","National
  ran <- sample(1:nrow(data_updated), 0.9 * nrow(data_updated))
  nor <-function(x) { (x -min(x))/(max(x)-min(x)) }
  train_norm <- as.data.frame(lapply(data_updated, nor))
  data_train <- train_norm[ran,]
  data_test <- train_norm[-ran,]
  dtm <- rpart(Value_bins~., data_train, method="class")
  rpart.plot(dtm, compress=TRUE, uniform=TRUE)
  p <- predict(dtm, data_test, type="class")
  confMat <- table(data_test$Value_bins,p)
  accuracy <- sum(diag(confMat))/sum(confMat)
  return (accuracy*100)
}
```

```
decision_tree_job_title_bins(data_data_scientist_bins)
```




```
## [1] 75
```

Large Data Import

```
all_data <- read.csv("df_all_final.csv", header = TRUE)
summary(all_data)
```

```
##        X                Metro        Dimension.Type     Month
##  Min.   : 31.0   Atlanta    : 335   Job Title:3685   Min.   :201707
##  1st Qu.:259.0   Boston     : 335                    1st Qu.:201802
##  Median :488.0   Chicago    : 335                    Median :201811
##  Mean   :490.2   Houston    : 335                    Mean   :201831
##  3rd Qu.:720.0   Los Angeles: 335                    3rd Qu.:201906
##  Max.   :957.0   National   : 335                    Max.   :201912
##                  (Other)    :1675
##              Dimension                Measure         Value
##  Accountant       : 319   Median Base Pay:3685   Min.   : 43577
##  Data Analyst     : 319                          1st Qu.: 60550
##  Data Scientist   : 319                          Median : 72828
##  Financial Analyst: 319                          Mean   : 78372
##  Graphic Designer : 319                          3rd Qu.: 95713
##  Professor        : 319                          Max.   :144533
##  (Other)          :1771
```

```
##       YoY           Value_bins    Registered.Nurse  Data.Scientist
##  Min.   :-7.000   Min.   : 5.00   Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 1.200   1st Qu.: 9.00   1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 2.000   Median :11.00   Median :0.00000   Median :0.00000
##  Mean   : 1.849   Mean   :12.17   Mean   :0.08657   Mean   :0.08657
##  3rd Qu.: 2.700   3rd Qu.:16.00   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   : 5.900   Max.   :25.00   Max.   :1.00000   Max.   :1.00000
##
##    Accountant        IT.Manager      Sales.Representative
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.04776   Mean   :0.08657
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##    Professor          Teacher        Software.Engineer Project.Manager
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.08657   Mean   :0.08657   Mean   :0.08657
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##  Financial.Analyst  Data.Analyst     Graphic.Designer     Atlanta
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.08657   Mean   :0.08657   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##      Boston            Chicago           Houston          Los.Angeles
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.09091   Mean   :0.09091   Mean   :0.09091   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##     National       New.York.City      Philadelphia     San.Francisco
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.09091   Mean   :0.09091   Mean   :0.09091   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##     Seattle        Washington.DC
##  Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000
##  Mean   :0.09091   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000
```

```
##  Max.   :1.00000   Max.   :1.00000
##
```

```
drops <- c("X", "Metro","Dimension.Type","Dimension","Measure","Value")
all_data_updated <- all_data[ , !(names(all_data) %in% drops)]
summary(all_data_updated)
```

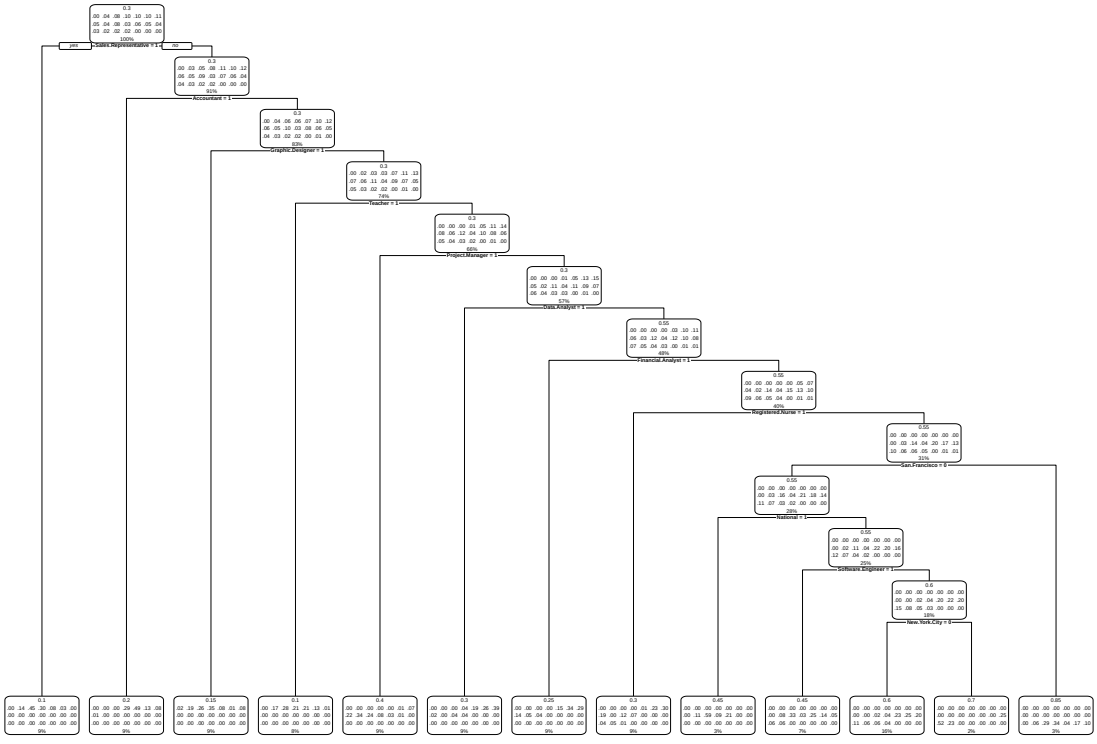
```
##      Month             YoY           Value_bins    Registered.Nurse
##  Min.   :201707   Min.   :-7.000   Min.   : 5.00   Min.   :0.00000
##  1st Qu.:201802   1st Qu.: 1.200   1st Qu.: 9.00   1st Qu.:0.00000
##  Median :201811   Median : 2.000   Median :11.00   Median :0.00000
##  Mean   :201831   Mean   : 1.849   Mean   :12.17   Mean   :0.08657
##  3rd Qu.:201906   3rd Qu.: 2.700   3rd Qu.:16.00   3rd Qu.:0.00000
##  Max.   :201912   Max.   : 5.900   Max.   :25.00   Max.   :1.00000
##  Data.Scientist      Accountant        IT.Manager
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.08657   Mean   :0.04776
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  Sales.Representative   Professor          Teacher
##  Min.   :0.00000      Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000      1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000      Median :0.00000   Median :0.00000
##  Mean   :0.08657      Mean   :0.08657   Mean   :0.08657
##  3rd Qu.:0.00000      3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000      Max.   :1.00000   Max.   :1.00000
##  Software.Engineer Project.Manager   Financial.Analyst  Data.Analyst
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.08657   Mean   :0.08657   Mean   :0.08657
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  Graphic.Designer     Atlanta            Boston           Chicago
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.08657   Mean   :0.09091   Mean   :0.09091   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##     Houston          Los.Angeles         National       New.York.City
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.09091   Mean   :0.09091   Mean   :0.09091   Mean   :0.09091
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##   Philadelphia     San.Francisco        Seattle        Washington.DC
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.09091   Mean   :0.09091   Mean   :0.09091   Mean   :0.09091
```

```
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
```

```
decision_tree_all <- function(data){
  ran <- sample(1:nrow(data), 0.9 * nrow(data))
  nor <-function(x) { (x -min(x))/(max(x)-min(x)) }
  train_norm <- as.data.frame(lapply(data, nor))
  data_train <- train_norm[ran,]
  data_test <- train_norm[-ran,]
  dtm <- rpart(Value_bins~., data_train, method="class")
  rpart.plot(dtm, compress=TRUE, uniform=TRUE)
  p <- predict(dtm, data_test, type="class")
  confMat <- table(data_test$Value_bins,p)
  accuracy <- sum(diag(confMat))/sum(confMat)
  return (accuracy*100)
}
```

```
decision_tree_all(all_data_updated)
```

```
## Warning: All boxes will be white (the box.palette argument will be ignored) because
## the number of classes in the response 21 is greater than length(box.palette) 6.
## To silence this warning use box.palette=0 or trace=-1.
```



```
## [1] 34.41734
```