Problem Statement:

A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

Approach - A decision tree can be built with target variable Sale (we will first convert it in categorical variable) & all other variable will be independent in the analysis.

> Company\_Data <- read\_csv("C:/Users/Dell/Desktop/ExcelR Assignments/11. Decision Tree/Company\_Data.csv")

> head(Company\_Data)

# A tibble: 6 x 11

Sales CompPrice Income Advertising Population Price ShelveLoc Age Education Urban US

<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <chr> <chr>

1 9.5 138 73 11 276 120 Bad 42 17 Yes Yes

2 11.2 111 48 16 260 83 Good 65 10 Yes Yes

3 10.1 113 35 10 269 80 Medium 59 12 Yes Yes

4 7.4 117 100 4 466 97 Medium 55 14 Yes Yes

5 4.15 141 64 3 340 128 Bad 38 13 Yes No

6 10.8 124 113 13 501 72 Bad 78 16 No Yes

> str(Company\_Data)

Classes ‘spec\_tbl\_df’, ‘tbl\_df’, ‘tbl’ and 'data.frame': 400 obs. of 11 variables:

$ Sales : num 9.5 11.22 10.06 7.4 4.15 ...

$ CompPrice : num 138 111 113 117 141 124 115 136 132 132 ...

$ Income : num 73 48 35 100 64 113 105 81 110 113 ...

$ Advertising: num 11 16 10 4 3 13 0 15 0 0 ...

$ Population : num 276 260 269 466 340 501 45 425 108 131 ...

$ Price : num 120 83 80 97 128 72 108 120 124 124 ...

$ ShelveLoc : chr "Bad" "Good" "Medium" "Medium" ...

$ Age : num 42 65 59 55 38 78 71 67 76 76 ...

$ Education : num 17 10 12 14 13 16 15 10 10 17 ...

$ Urban : chr "Yes" "Yes" "Yes" "Yes" ...

$ US : chr "Yes" "Yes" "Yes" "Yes" ...

> summary(Company\_Data)

Sales CompPrice Income Advertising Population

Min. : 0.000 Min. : 77 Min. : 21.00 Min. : 0.000 Min. : 10.0

1st Qu.: 5.390 1st Qu.:115 1st Qu.: 42.75 1st Qu.: 0.000 1st Qu.:139.0

Median : 7.490 Median :125 Median : 69.00 Median : 5.000 Median :272.0

Mean : 7.496 Mean :125 Mean : 68.66 Mean : 6.635 Mean :264.8

3rd Qu.: 9.320 3rd Qu.:135 3rd Qu.: 91.00 3rd Qu.:12.000 3rd Qu.:398.5

Max. :16.270 Max. :175 Max. :120.00 Max. :29.000 Max. :509.0

Price ShelveLoc Age Education Urban

Min. : 24.0 Length:400 Min. :25.00 Min. :10.0 Length:400

1st Qu.:100.0 Class :character 1st Qu.:39.75 1st Qu.:12.0 Class :character

Median :117.0 Mode :character Median :54.50 Median :14.0 Mode :character

Mean :115.8 Mean :53.32 Mean :13.9

3rd Qu.:131.0 3rd Qu.:66.00 3rd Qu.:16.0

Max. :191.0 Max. :80.00 Max. :18.0

#As per Problem statement sales column need to be converted into Categorical form.

#As Mean of sales column is 7.49 i.e. aprox 8 converting it into 1 for Company\_Data$Sales>= 8 and into 0 for Company\_Data$Sales< 8.

> Categorical\_Company\_Data <- ifelse(Company\_Data$Sales>= 8, yes = 1,0)

> head(Categorical\_Company\_Data)

[1] 1 1 1 0 0 1

#Deleting 1st column from dataset to merge the categorical new column into it

> Company\_Data <- Company\_Data[-1]

> Company\_Data <- cbind(Categorical\_Company\_Data,Company\_Data)

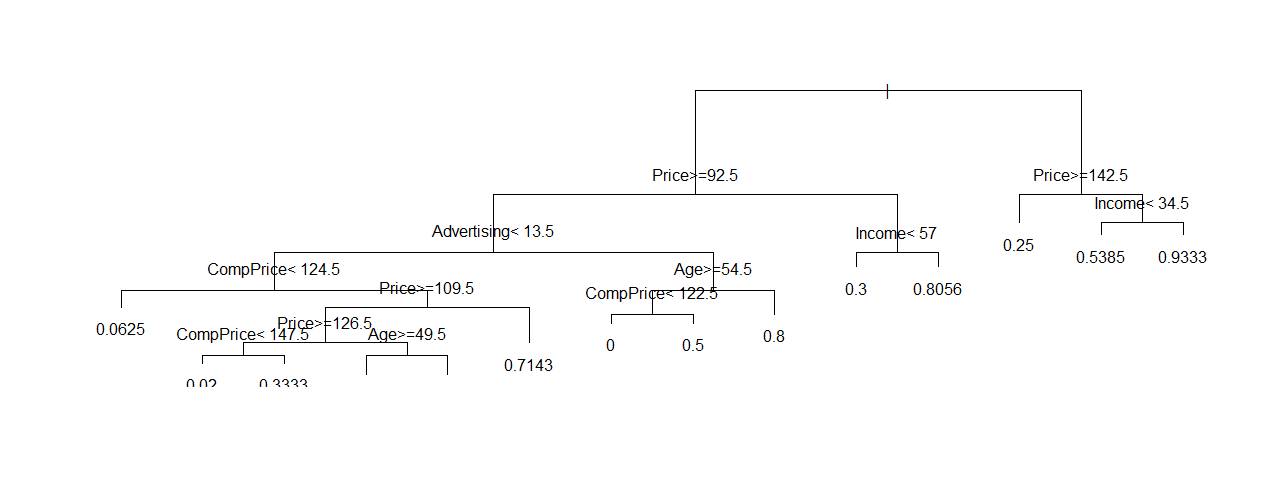
> cor(Company\_data)

#Building a regression tree using rpart Simple model

> model\_cart1 <- rpart(Categorical\_Company\_Data~.,data=Company\_Data,method="anova")

> plot(model\_cart1)

> text(model\_cart1)



> pred\_Categorical\_Company\_Data <- predict(model\_cart1,Company\_Data)

> rmse\_Categorical\_Company\_Data <- sqrt(mean((pred\_Categorical\_Company\_Data-Company\_Data$Categorical\_Company\_Data)^2))

> rmse\_Categorical\_Company\_Data

[1] 0.3327733

#Building a function to find adjusted R sqr values.

>Adjusted\_RSqred <- function(pred, obs, formula = "corr", na.rm = FALSE) {

n <- sum(complete.cases(pred))

switch(formula,

corr = cor(obs, pred, use = ifelse(na.rm, "complete.obs", "everything"))^2,

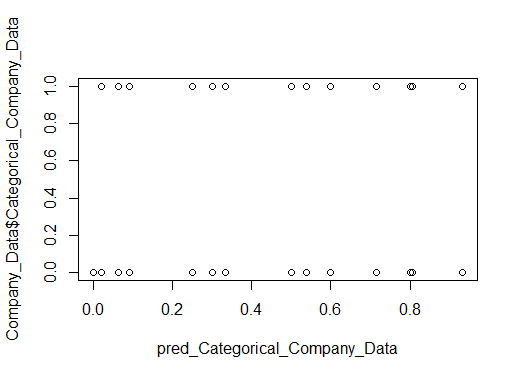
traditional = 1 - (sum((obs-pred)^2, na.rm = na.rm)/((n-1)\*var(obs, na.rm = na.rm))))

}

> Adjusted\_RSqred(pred\_Categorical\_Company\_Data,Company\_Data$Categorical\_Company\_Data)

[1] 0.5422156

> plot(pred\_Categorical\_Company\_Data,Company\_Data$Categorical\_Company\_Data)



> cor(pred\_Categorical\_Company\_Data,Company\_Data$Categorical\_Company\_Data)

[1] 0.7363529