Decision Trees

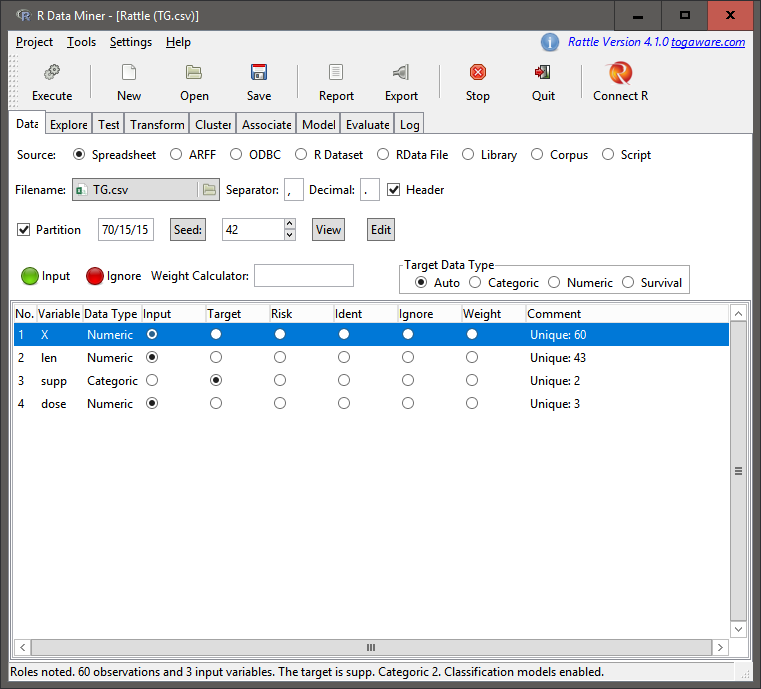
Dean D'souza

2016-08-08

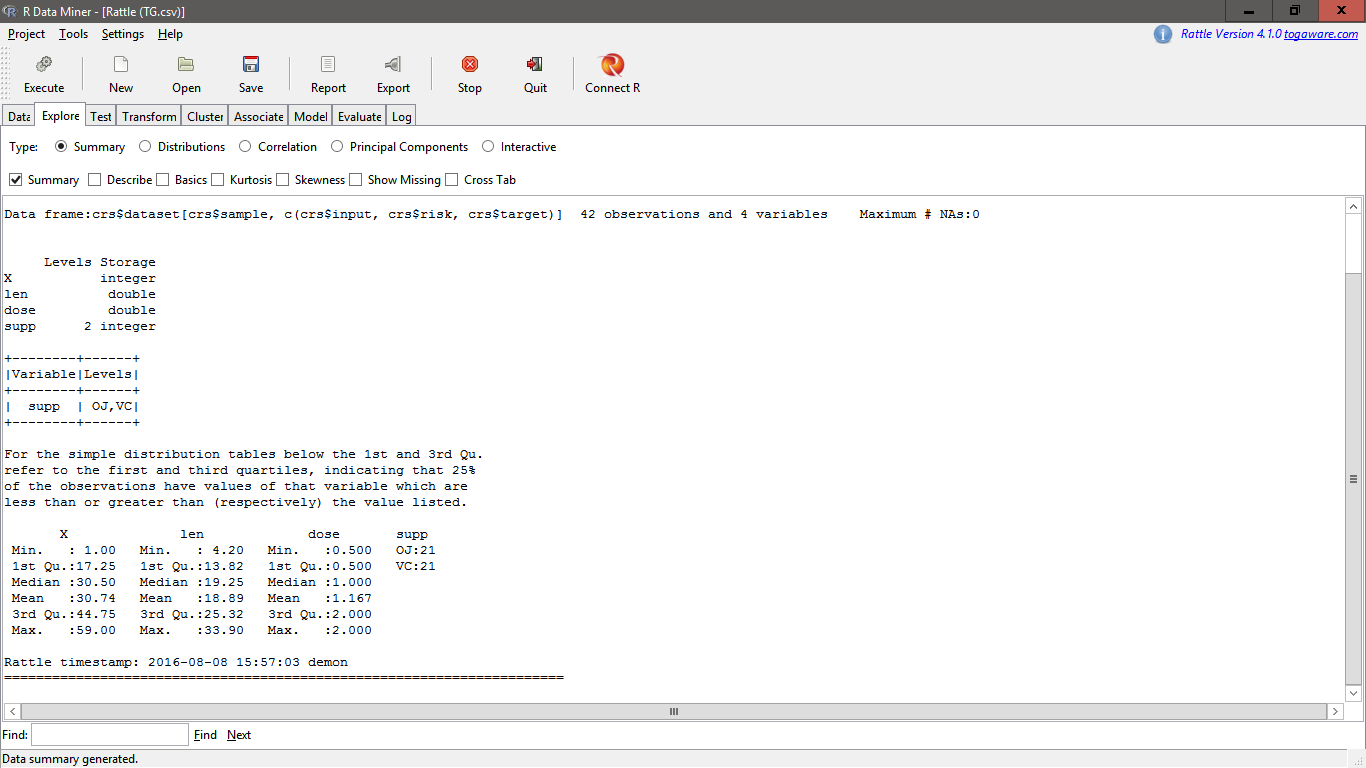
## Soltuions:

For the purpose of this assignment I chose the ToothGrowth dataset from base R. The description of the dataset is as follows:

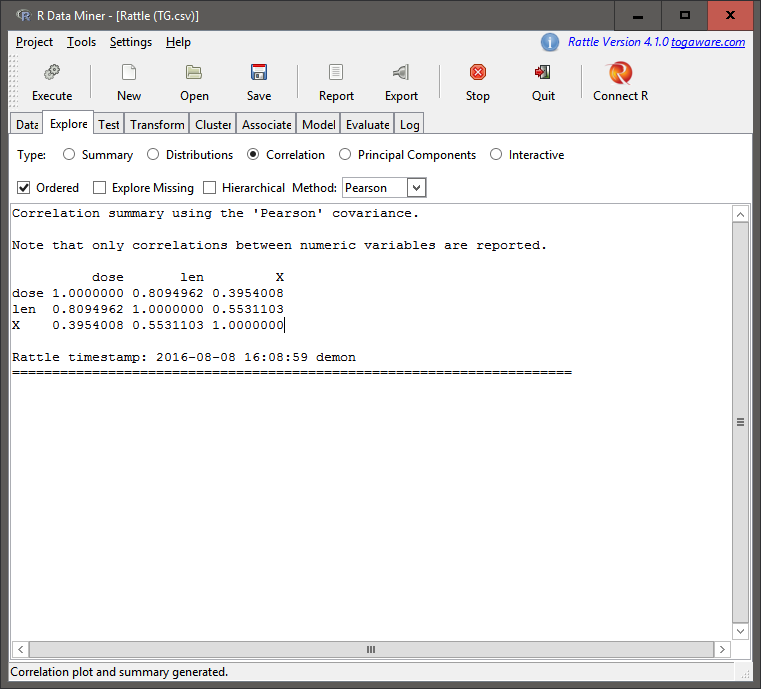
* **First, build a decision tree using Rattle just as we did here on a data set of your choosing.**
* Take several screen shots of your work and paste into a MS Word, LibreOffice, of pdf file, explain each step
* Tell me what you learned from the decision tree

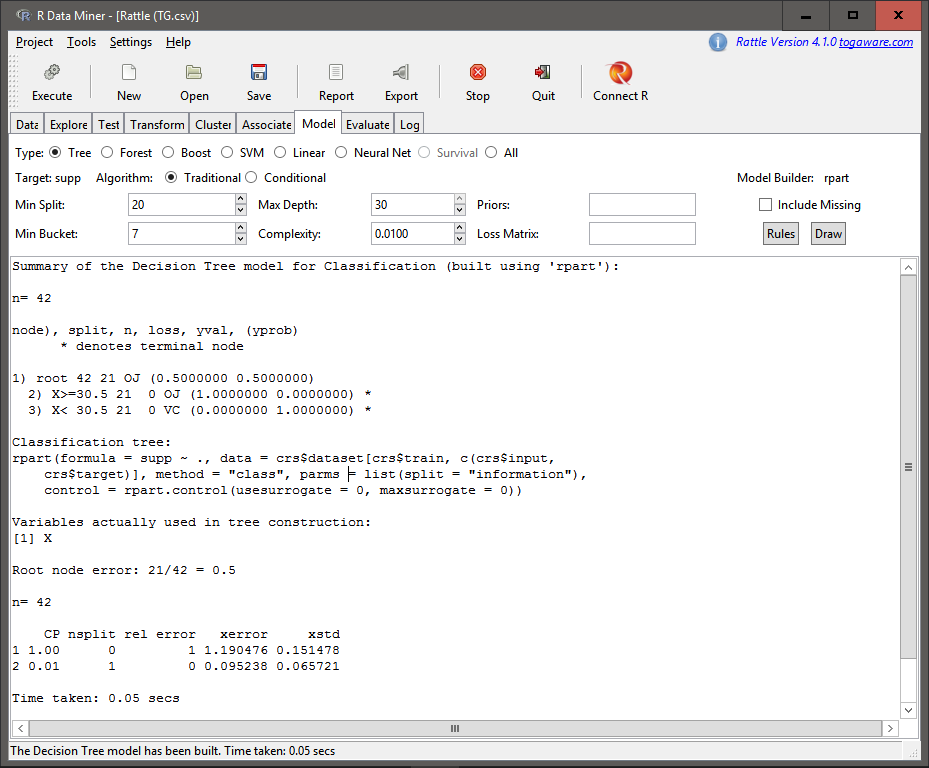
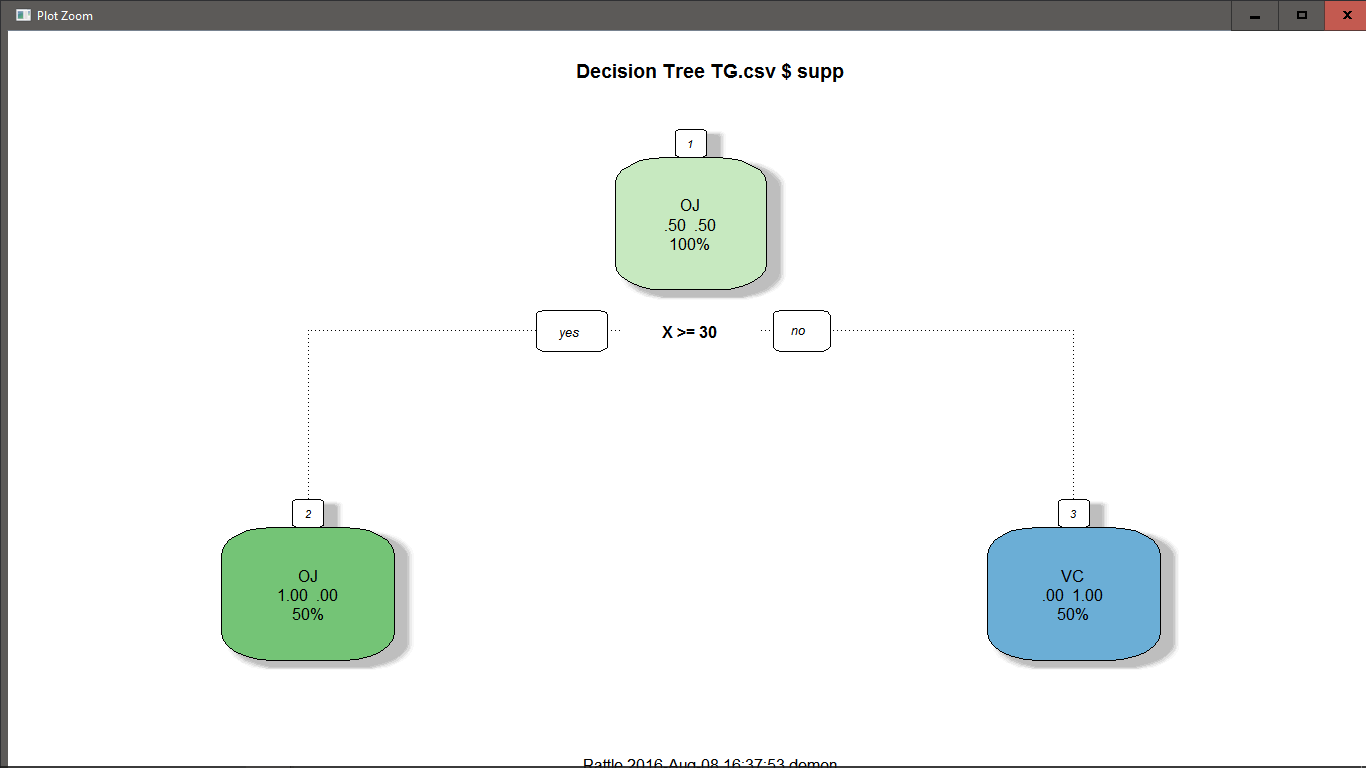


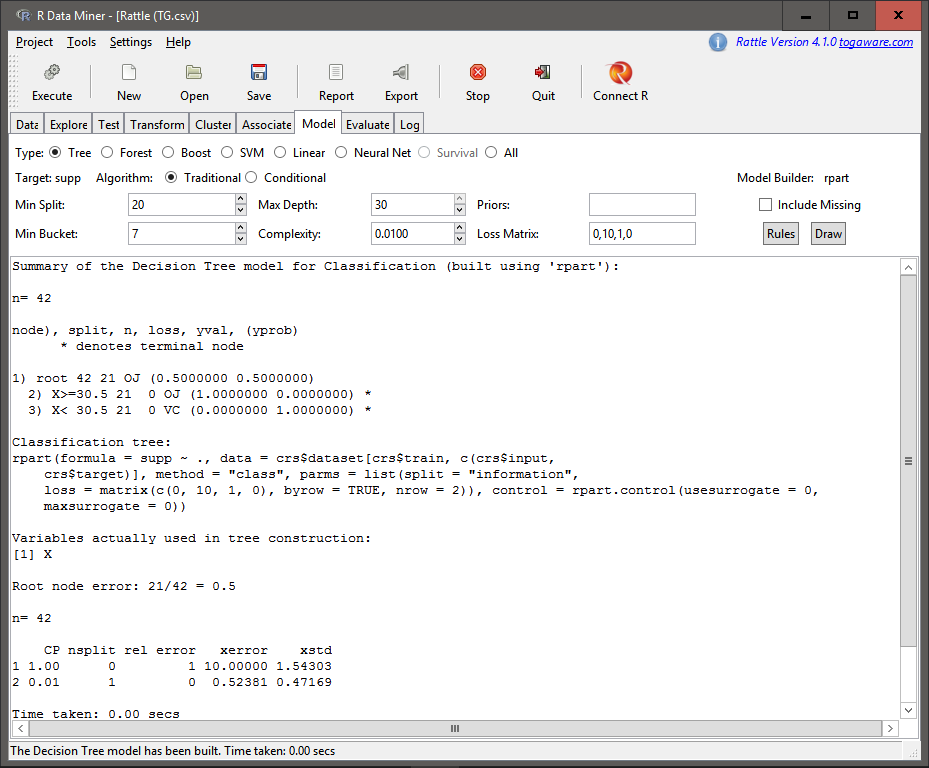
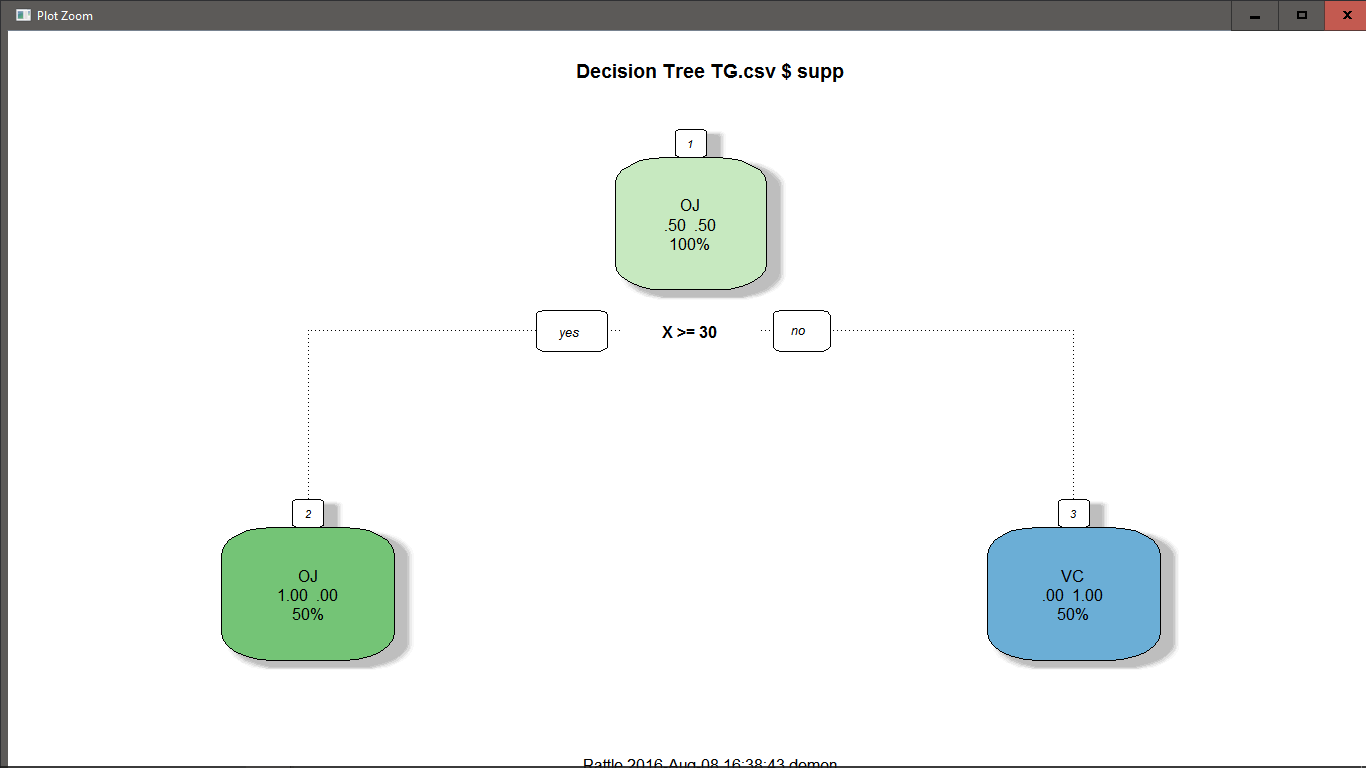
Screenshot1

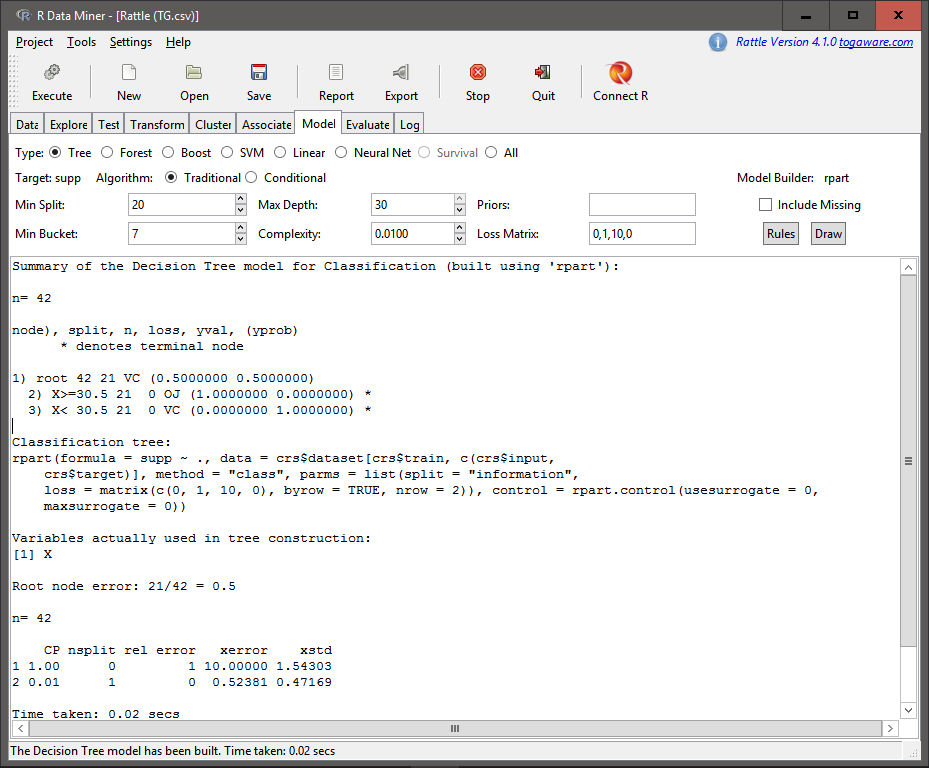
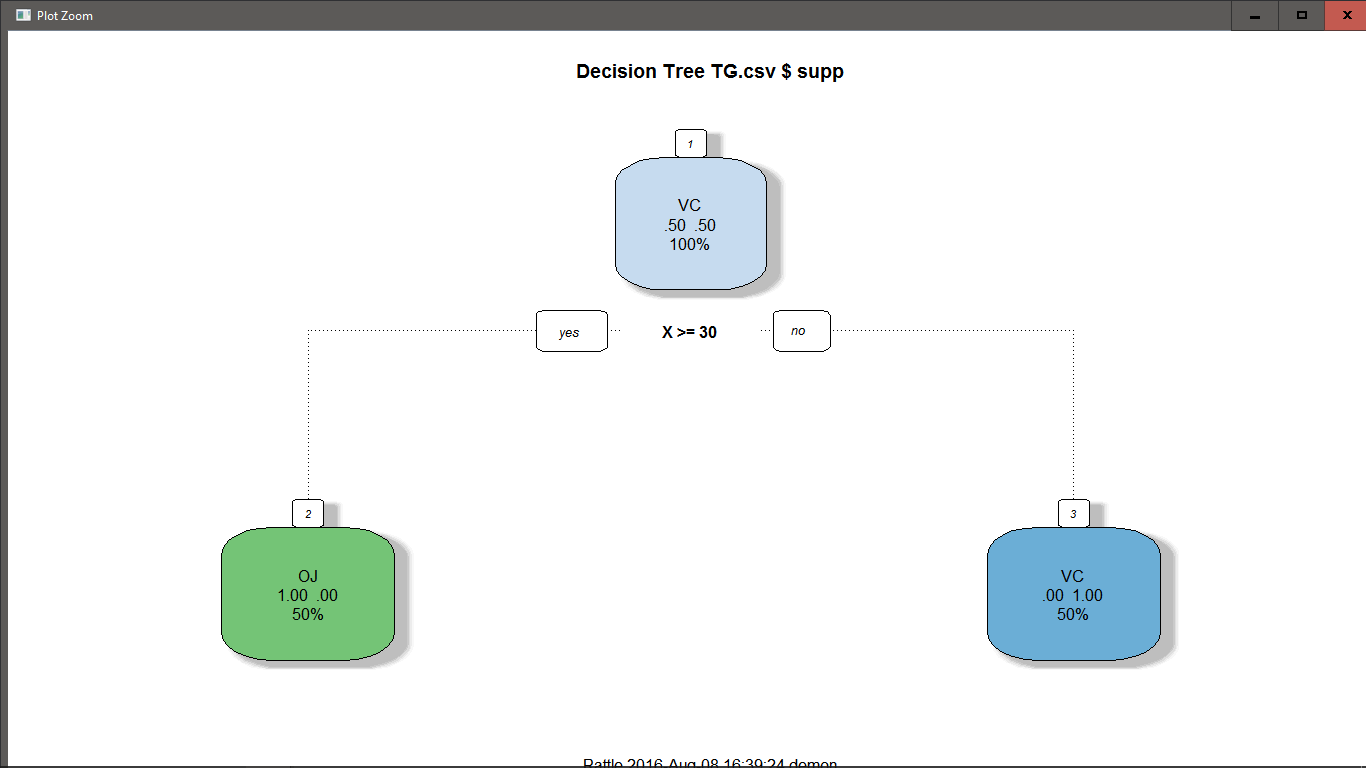


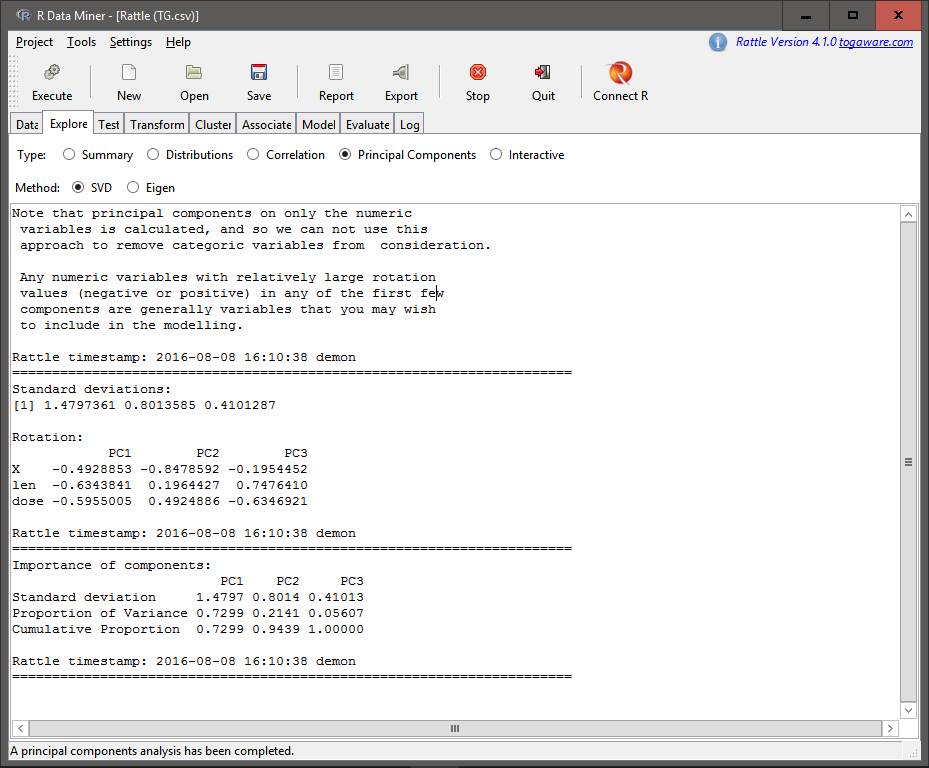
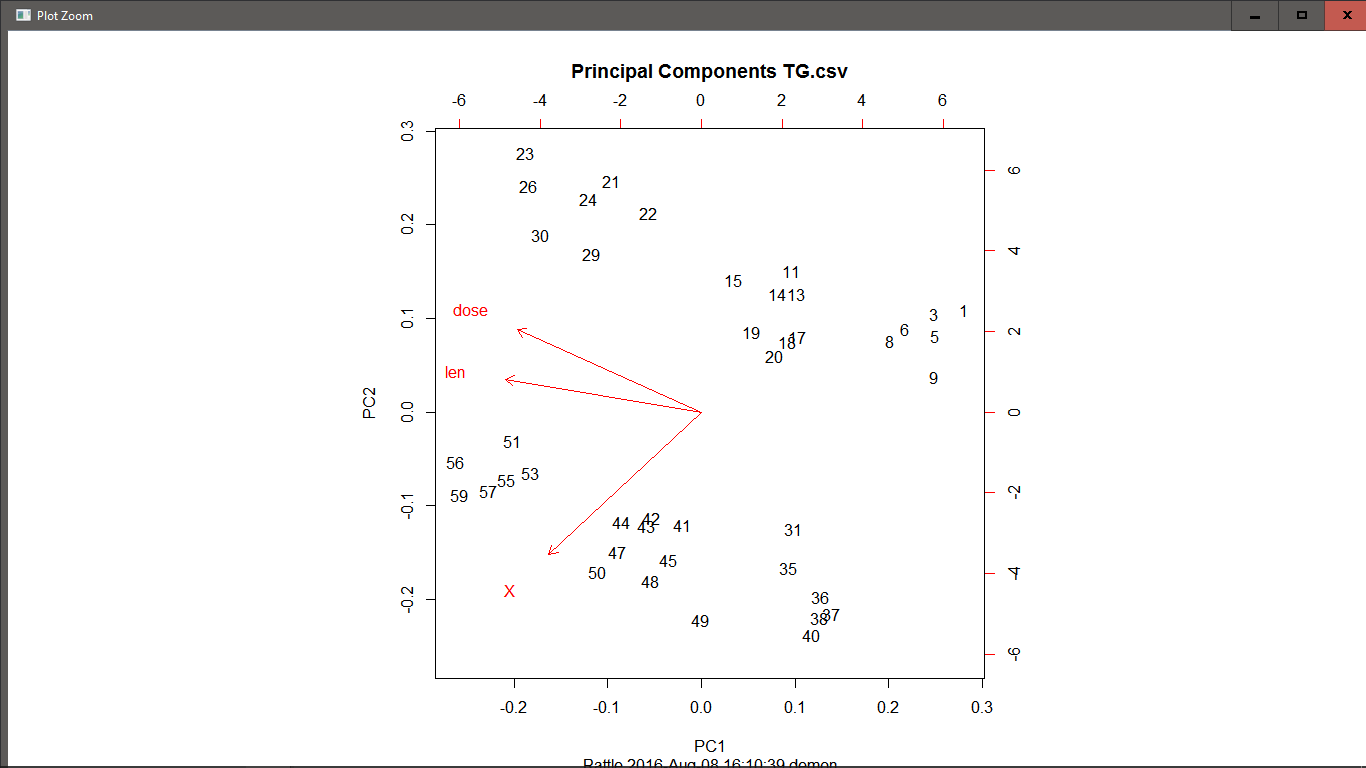
Screenshot2

* **Build another model of your choice using Rattle on the same data set**
* You can run Association Analysis, PCA, Regression, .whatever
* Tell me what you learned from this analysis and from the decision tree combined

# Setting Seed for reproducible research  
set.seed(695949)  
  
# Creating correlation matrix  
TG\_cor<-cor(data.matrix(TG))  
TG\_cor

## len supp dose  
## len 1.0000000 -0.2438927 0.8026913  
## supp -0.2438927 1.0000000 0.0000000  
## dose 0.8026913 0.0000000 1.0000000

#Checking which variables can be removed  
findLinearCombos(data.matrix(TG))

## $linearCombos  
## list()  
##   
## $remove  
## NULL

# Creating Training and Test Data partitions  
TG\_sv<-createDataPartition(TG$supp, p=.75,list=FALSE)  
TG\_train<-TG[TG\_sv,]  
TG\_test<-TG[-TG\_sv,]  
  
# Creating the linear model  
TG\_lm1<-lm(as.numeric(supp)~len+dose+len\*dose, data = TG)  
  
# Details of the linear model  
summary(TG\_lm1)

##   
## Call:  
## lm(formula = as.numeric(supp) ~ len + dose + len \* dose, data = TG)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.70669 -0.41031 0.03634 0.43684 0.75926   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.96595 0.35143 5.594 6.88e-07 \*\*\*  
## len -0.05163 0.02040 -2.530 0.0142 \*   
## dose 0.26610 0.44884 0.593 0.5557   
## len:dose 0.00757 0.01806 0.419 0.6767   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4716 on 56 degrees of freedom  
## Multiple R-squared: 0.1698, Adjusted R-squared: 0.1254   
## F-statistic: 3.819 on 3 and 56 DF, p-value: 0.01466

# Rebuilding the model  
TG\_lm2<-lm(as.numeric(supp)~len, data = TG)  
  
# Details of the new model  
summary(TG\_lm2)

##   
## Call:  
## lm(formula = as.numeric(supp) ~ len, data = TG)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.67063 -0.43510 -0.02031 0.46281 0.74254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.802456 0.170271 10.586 3.61e-15 \*\*\*  
## len -0.016077 0.008394 -1.915 0.0604 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4932 on 58 degrees of freedom  
## Multiple R-squared: 0.05948, Adjusted R-squared: 0.04327   
## F-statistic: 3.668 on 1 and 58 DF, p-value: 0.06039

# Performing anova  
anova(TG\_lm2,TG\_lm1)

## Analysis of Variance Table  
##   
## Model 1: as.numeric(supp) ~ len  
## Model 2: as.numeric(supp) ~ len + dose + len \* dose  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 58 14.108   
## 2 56 12.452 2 1.6554 3.7222 0.03036 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Predicting the with the test dataset  
TG\_p<-predict(TG\_lm2, TG\_test)  
  
# Observing the results  
summary(as.numeric(TG$supp))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 1.5 1.5 2.0 2.0

summary(as.numeric(TG\_train$supp))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 1.5 1.5 2.0 2.0

summary(as.numeric(TG\_test$supp))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 1.5 1.5 2.0 2.0

summary(TG\_p)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.330 1.435 1.491 1.492 1.537 1.651

* **Tell me why you prefer RStudio over Rattle or the opposite or some other tool of your choice over Rstudio. Just looking for opinions here, but I am curious HOW you evaluate the different tools.**