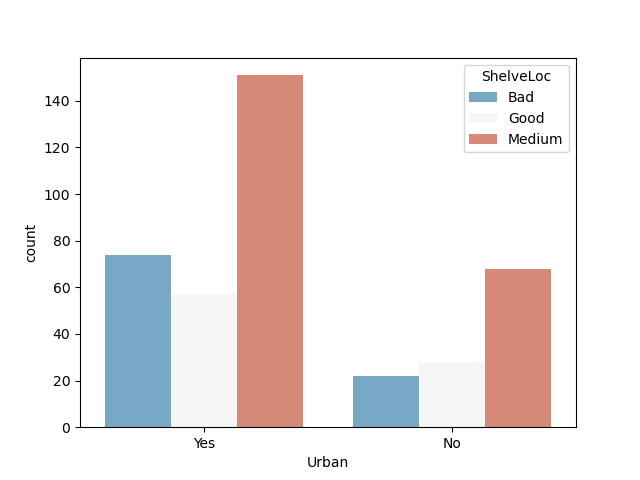
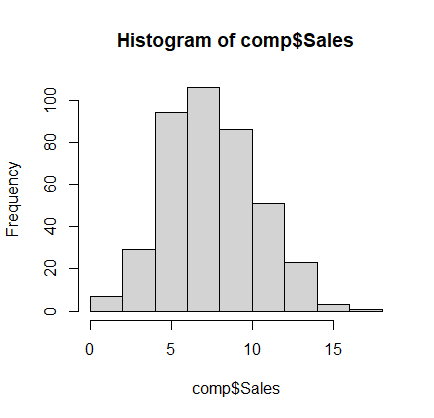
Decision tree and Random Forest

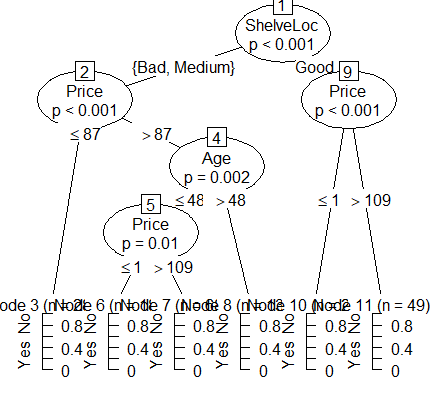
1. Company\_data:

* Checking Sales Distr {yes , No}
* # proportion of NO Yes are No Yes
* # 0.8025 0.1975 respectively
* ### Bad Good Medium of ShelveLoc
* ### 96 85 219
* # ShelevLoc quality Medium is of higher freq disrtr ver the entire dataset.

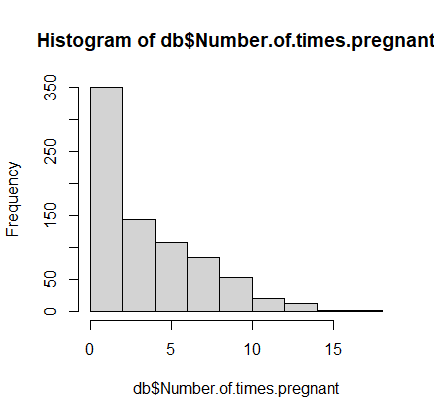




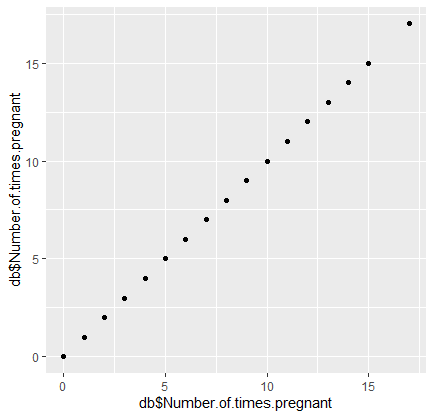
* Sales of company in Histogram
* Company Sales look the most top in between 5 -10 .By judging it has stable sales since the histogram looks normally distributed.
* Below is the Graph tree for the company data



* C5.p0 gave test accuracy of our model is 66 % accurate 0.665625.
* #Confusion matrix:
* # No Yes class.error
* #No 247 10 0.03891051
* #Yes 33 30 0.52380952.



2.) Diabetes Data set:

* The relation between number of times pregnant and its frequency is between 0-5 %.
* Age v Pregnancy for diabetes.
* 

The Numbers seems to correlate as with age and time. Pregnant.

#NO YES

#0.6510417 0.3489583 is the proportion of people having diabetes or not

test accuracy ::: 0.7402597 74% in C5.0

test accuracy ::: 0.7012987 .Accuracy reduced in Random Forest

3.) Fraud data set

# Good Risky

# 0.7933333 0.2066667

# only 20 % percent are observed to have risky Taxable\_income to be cross checked

Using Decision tree

[[99 42]

[28 11]]

Report:

precision recall f1-score support

0.0 0.78 0.70 0.74 141

1.0 0.21 0.28 0.24 39

accuracy 0.61 180

macro avg 0.49 0.49 0.49 180

weighted avg 0.66 0.61 0.63 180

Using Random Forest

[[16 21]

[ 5 78]]

Report:

precision recall f1-score support

0.0 0.80 0.92 0.85 144

1.0 0.14 0.06 0.08 36

accuracy 0.74 180

macro avg 0.47 0.49 0.47 180

weighted avg 0.66 0.74 0.70 180

# Accuracy improved from 61 to 74 by using Dtree to Random Forest