Business problem

A decision tree & random forest model can be built with target variable 'Sale'

Inference from the data

Data talks about sales depends on income, advertising , population, price and other various categories in the data with 400 observations

Data set size

Data give is found to be a continuous data ,categorical data for which classification model can be performed getting deeper into the data analysis and its behavior

Exploratory data analysis

By using summary of the data set we can find the mean, median, mode for different columns in the data set with this we can find how the data is some of the variance and standard deviation of data set

> #second moment

> var(CompanyData$Income)

[1] 783.2182

> var(CompanyData$Income)

[1] 783.2182

> var(CompanyData$Population)

[1] 21719.81

> var(CompanyData$Price)

[1] 560.5844

> var(CompanyData$Age)

[1] 262.4496

>

> sd(CompanyData$Income)

[1] 27.98604

> sd(CompanyData$Population)

[1] 147.3764

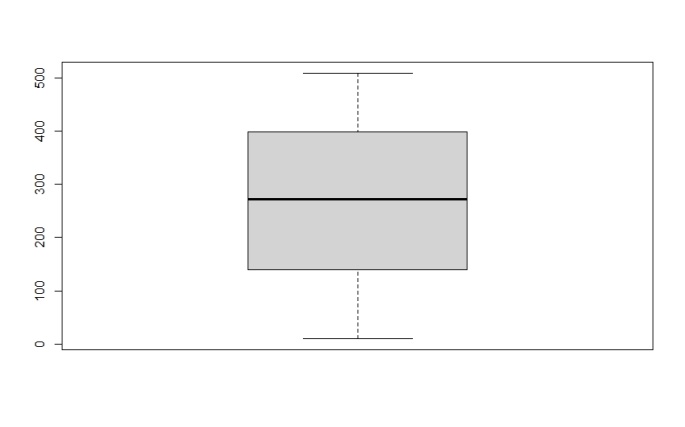
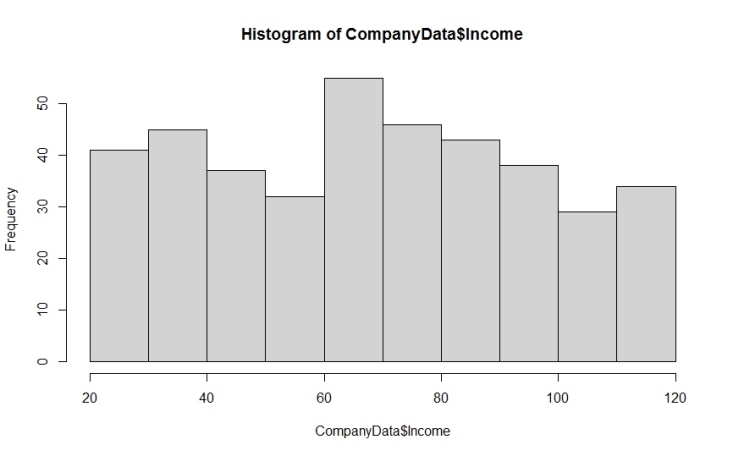
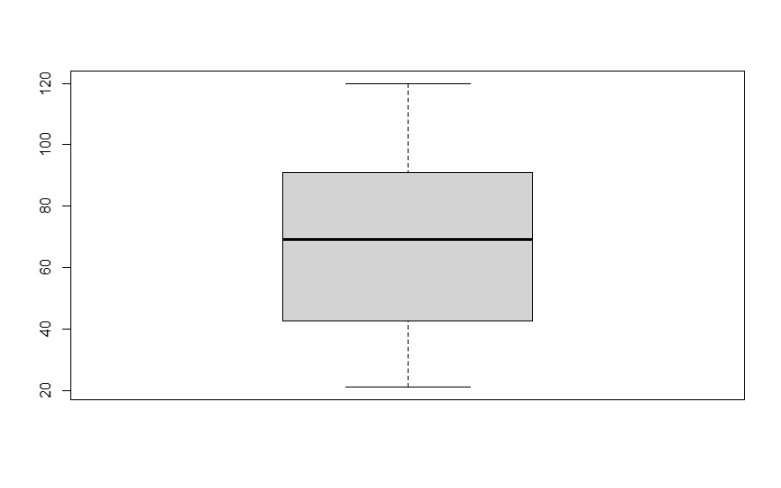
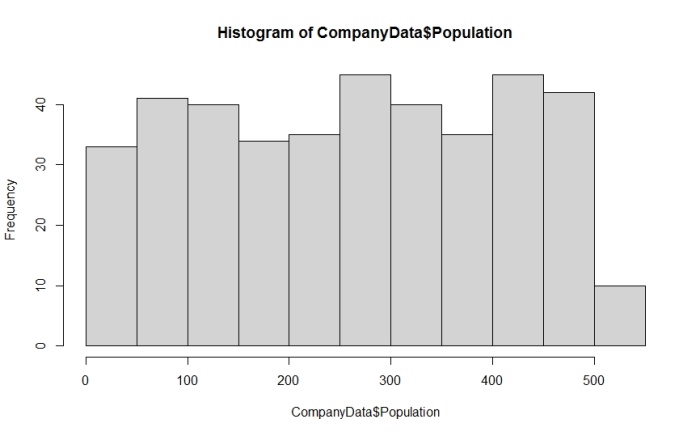
> sd(CompanyData$Price)

[1] 23.67666

> sd(CompanyData$Age)

[1] 16.2003

Graphical representation



In the above histogram and boxplot of population and income is shown in the population data the values are lies in the range of 30 to 45 and there is no outliers in the data it is clearly shown by boxplot where in the income the values lies in the range of 40 to 50 with equal distribution and there is no outliers skewness and kurtosis for population and income

> skewness(CompanyData$Income)

[1] 0.04925888

> kurtosis(CompanyData$Income)

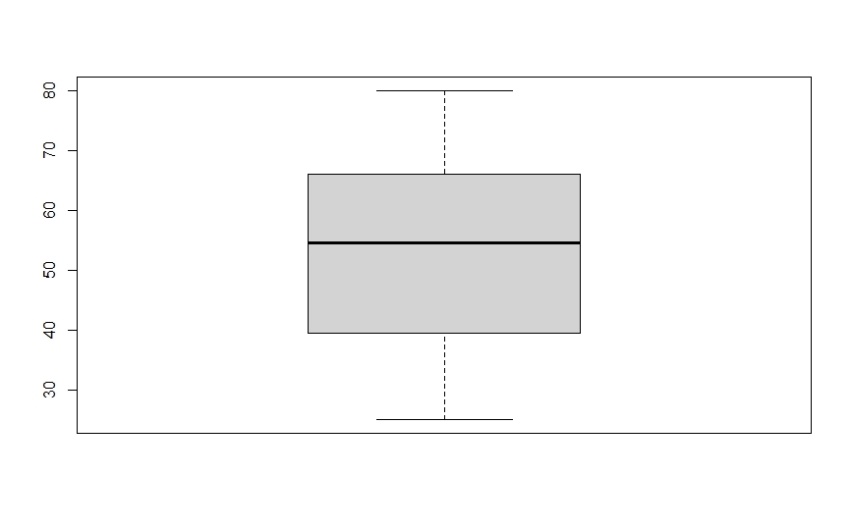
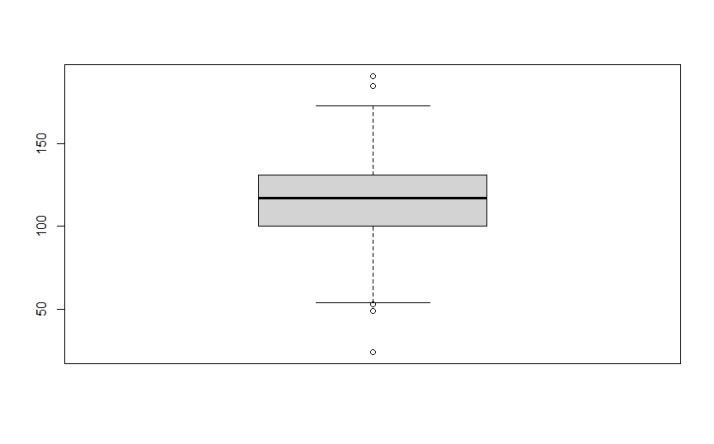
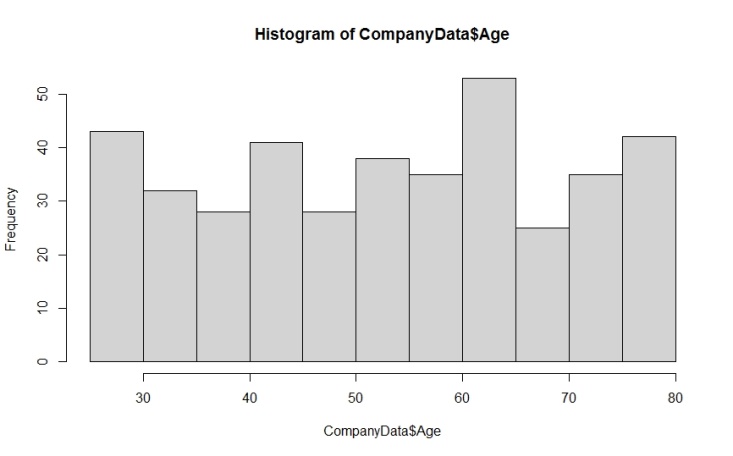
[1] 1.913267

> skewness(CompanyData$Population)

[1] -0.05103434

> kurtosis(CompanyData$Population)

[1] 1.797696



In the above representation price data is normally distributed where the age close to equal distribution but the age data has outliers in the data set where in price there is no outliers skewness of price and age id left skeweed with the kurtosis and skewness value of

> skewness(CompanyData$Price)

[1] -0.1248159

> kurtosis(CompanyData$Price)

[1] 3.431294

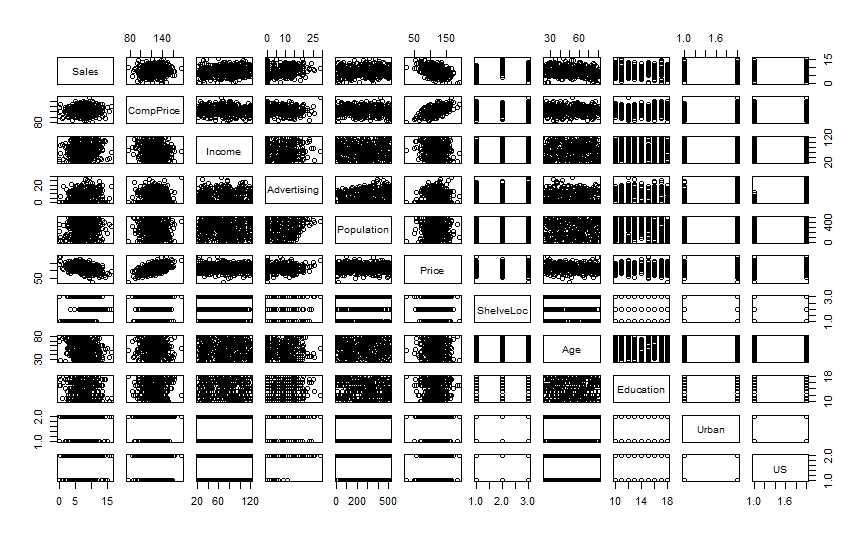
> skewness(CompanyData$Age)

[1] -0.076892

> kurtosis(CompanyData$Age)

[1] 1.864776

Bivariate analysis



Bivariate analysis helps to relate the variables with other variables and we can find the types of correlation

Data modeling

In this at first the continuous data is converted into categorical data of yes and no and then if we want to change the yes and no in to 0 and 1 we can use as.factor function to convert after converting all the categorical values by leveling the data we can perform the decision tree and random forest modeling in the classification model we want to split the data first for train and test data adte that we want to do leveling proportion of yes and no for train and test data id

> prop.table(table(companydatatrain$highsale))

no yes

0.728 0.272

> prop.table(table(companydatatest$highsale))

no yes

0.6933333 0.3066667

Using the library function c5.0 we can perform decision tree and accuracy for train and test is 96% and 80%

Cross table for test data

Total Observations in Table: 150

| predicted default

actual default | no | yes | Row Total |

---------------|-----------|-----------|-----------|

no | 92 | 12 | 104 |

| 3.246 | 8.926 | |

| 0.885 | 0.115 | 0.693 |

| 0.836 | 0.300 | |

| 0.613 | 0.080 | |

---------------|-----------|-----------|-----------|

yes | 18 | 28 | 46 |

| 7.338 | 20.180 | |

| 0.391 | 0.609 | 0.307 |

| 0.164 | 0.700 | |

| 0.120 | 0.187 | |

---------------|-----------|-----------|-----------|

Column Total | 110 | 40 | 150 |

| 0.733 | 0.267 | |

---------------|-----------|-----------|-----------|

For train data

Total Observations in Table: 250

| predicted default

actual default | no | yes | Row Total |

---------------|-----------|-----------|-----------|

no | 178 | 4 | 182 |

| 15.628 | 41.827 | |

| 0.978 | 0.022 | 0.728 |

| 0.978 | 0.059 | |

| 0.712 | 0.016 | |

---------------|-----------|-----------|-----------|

yes | 4 | 64 | 68 |

| 41.827 | 111.949 | |

| 0.059 | 0.941 | 0.272 |

| 0.022 | 0.941 | |

| 0.016 | 0.256 | |

---------------|-----------|-----------|-----------|

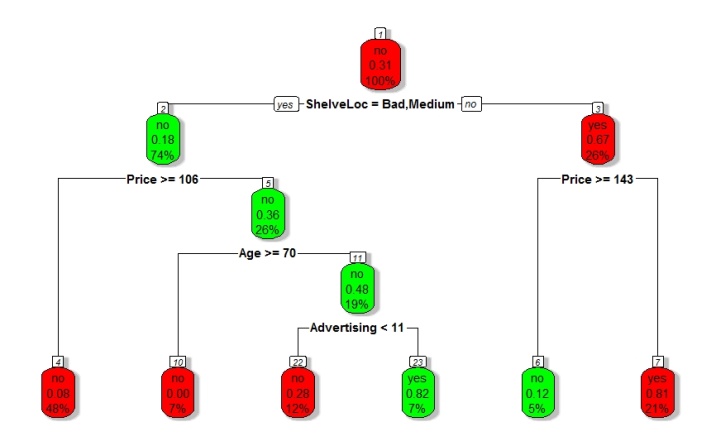
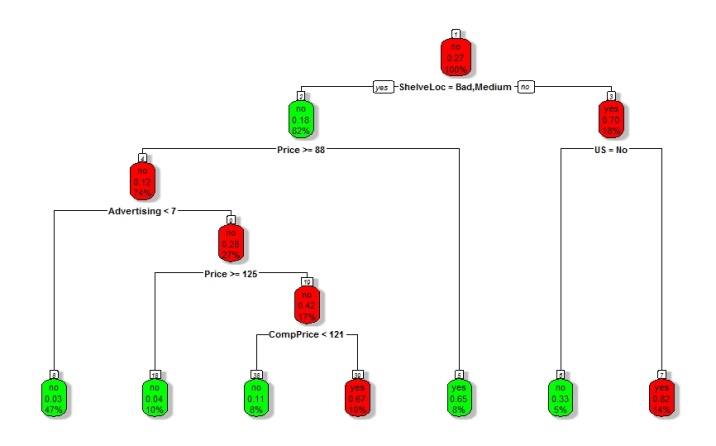
Column Total | 182 | 68 | 250 |

| 0.728 | 0.272 | |

---------------|-----------|-----------|-----------|

Decision tree flow chart for train data

In this shlveloc is taken as root node and the branches of price , us, advertise with this terminal node are found by calculating this way accuracy is better than taking other columns as root node



Train data

Random forest

In this random forest it takes different no decision tree to perform the action in this middle line is the value analyzed by random forest above and below lines are the interval, for random forest the train and test data accuracy is 96% and 82% respectively

