#install pacakages

#dplyr-transform and summarize tabular data with rows and columns.

install.packages("dplyr")

install.packages("readr")

install.packages("ggplot2")

library("dplyr")

library("ggplot2")

library(readr) # To read csv file

#Reading All Dataset(EMP)

train\_data<-read.csv("D:/ETL Hive/Demo/EMP.csv",header=TRUE)

head(train\_data)

#str- display the content of the list

str(train\_data)

#dim-for dimension

dim(train\_data)

nrow(train\_data)

ncol(train\_data)

#Exploration Data Analysis

#Boxplot

boxplot(train\_data$SALARY,col="blue")

boxplot(train\_data$SALARY,ylim=c(0,20000),col = "orange")

View(train\_data)

#histogram

hist(xlab="SALARY",ylab="AGE",train\_data$SALARY, col = "green",ylim = c(0,40),xlim = c(0,20000))

#to show the actual data points in histogram we use rug()

rug(train\_data$DEPARTMENT\_ID)

#barplot

table(train\_data$Employee\_Details)

count=table(train\_data$Employee\_Details)

percentage=table(train\_data$Employee\_Details)/99

percentage

barplot(percentage,col="blue",ylim=c(0.0,0.7),xlab = "Employee\_Details",ylab="percentage")

#create pie chart

pie(percentage)

#create the scatter plot

with(train\_data,plot(EMPLOYEE\_ID,SALARY))

#add annotation to plot

title("Salary and Department\_Id")

#ploting Tree

install.packages("partykit")

library(partykit)

tree <- ctree(EMPLOYEE\_ID~SALARY+MANAGER\_ID+DEPARTMENT\_ID, train\_data)

plot(tree, type='simple')

#Desion Tree confusion matrix

pred\_tree<- predict(tree,train\_data)

print("Confusion Matrix For decision tree");

table(predicated = pred\_tree,actual=train\_data$Employee\_Details)

tab1<-table(predicted = p1,Actual=train\_data$Employee\_Details)

tab2<-table(predicted = pred\_tree,Actual=train\_data$Employee\_Details)

print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))#Decision Tree Accuracy

#suffle the data from dataset

shuffle <- sample(1:nrow(train\_data))

head(shuffle)

t1<- train\_data[shuffle, ]

head(t1)

#dplyr-transform and summarize tabular data with rows and columns

#ISLR-introduction to statstical learing with application in R

#It provide the collection of data-sets used in the book 'An Introduction to Statistical Learning with Applications in R'.

library(ISLR)

library(naivebayes)

library(e1071)

A = data.frame(train\_data)

str(A)

sf = sample(1:3,nrow(A),replace = TRUE,prob = c(.30,.60,.10))

#barplot for predicate the data

barplot(table(sample(1:3,nrow(A),replace = TRUE,prob = c(.30,.60,.10))))

trd = A[sf == 1,]

tsd = A[sf == 2,]

model\_nb = naive\_bayes(Employee\_Details ~ ., data = trd)

model\_e1 = naiveBayes(Employee\_Details ~ ., data = trd)

pred\_nb = predict(model\_nb, tsd)

pred\_nb

table(pred\_nb,tsd$Employee\_Details)

pred\_e1 = predict(model\_e1, tsd)

pred\_e1

#classiication of single column from dataset

levels(train\_data$JOB\_ID)

table(train\_data$JOB\_ID)

plot(table(train\_data$JOB\_ID))

#analysis of salary in dataset

dim(train\_data)

sal <- train\_data[train\_data$SALARY > 0, ] # keep only rows with po

sal

sal1 <- train\_data[is.na(train\_data$SALARY) == FALSE,] # keep only those rows where base pay value is NOT missing (NA)

sal1

summary(train\_data$SALARY)