

```

print("Yasser Ashraf Gaber"                                22010409")
#use readxl//install.packages("readxl")
library(readxl)
#import the excel file to a data frame.
fruit_prices<-read_excel("D:\\1.A.Semester3\\Intro_to_data_science\\project\\Fruit Prices
2020 .xlsx")
#print the data frame,using print method with n paramater to read all rows in the file.
print(fruit_prices,n=63)

#Data Cleaning
#1-duplicated rows are in your data
if(sum(duplicated(fruit_prices))==0){
  paste("No Duplicated Values ",sum(duplicated(fruit_prices)))
}else{
  paste("Number of duplicated rows :",sum(duplicated(fruit_prices)))
}

#2-which rows have duplicated values
duplicated(fruit_prices)

#3-Select only all distinct rows//install.packages("dplyr")
library(dplyr)
print(distinct(fruit_prices),n=63)
#4-sum of NA values
if(sum(is.na(fruit_prices))==0){
  paste("There aren't NA Values",sum(is.na(fruit_prices)))
}else{
  paste("Number of NA values :",sum(is.na(fruit_prices)))
}

#unsupervised technique -> k-means
#retailprice and CupEquivalentPrice columns
fruit_prices_k<-fruit_prices[,c(3,8)]
print(fruit_prices_k,n=63)
Kmean_clustering_fruits<-kmeans(fruit_prices_k,centers = 2)
Kmean_clustering_fruits

#Supervised learning technique ->Decision Tree
#columns from 1 to 10
fruit_prices_tree<-fruit_prices[1:10,]
fruit_prices_tree
#use rpart
library(rpart)
tree<-rpart(Form ~ Fruit + Yield + RetailPriceUnit,
             data =fruit_prices_tree , minsplit=2)
tree
library(rpart.plot)
rpart.plot(tree1)

```