```
#Iris Clustering code

library(readxl)#to read excel
library(plyr)#split and put data
library(DescTools) #For Mode
library(dplyr) #For Pipe Operator
library(Hmisc) #For rcorr Function & Missing Values Treatment
library(QuantPsyc) #For lm.beta function
library(ggpubr) #For advanced QQ Plots
library(caret) #For Data manipulation
library(imputeMissings) #For imputing missing Values
library(purrr) #For Missing Values
library(naivebayes) #For naive Bayes
library(rpart.plot) #For DT Graph
library(psych) #descriptive stats
library(factoextra)#for pca operation
library(cluster)#for clustering
library(dplyr)#data manipulation
library(reshape2) #data reshape purposes
library(plotrix) #3D imaging
library(scatterplot3d)#3D scatterplot
library(corrplot) #plotting correlation
install.packages("DescTools")
install.packages("dplyr")
install.packages("Hmisc")
install.packages("QuantPsyc")
install.packages("caret")
install.packages("imputeMissings")
install.packages("purrr")
install.packages("naivebayes")
install.packages("rpart.plot")
install.packages("psych")
install.packages("factoextra")
install.packages("cluster")
install.packages("dplyr")
install.packages("reshape2")
install.packages("plotrix")
install.packages("scatterplot3d")
install.packages("corrplot")


#getting the data set
iris<- read.csv("C:\\Users\\sujoydutta\\Desktop\\Data analysis\\Datasets\\Clustering\\Iris.csv")


#seeing the data set
View(iris)
str(iris)


#checking null values
map(iris, ~sum(is.na(.)))

#summary statistics
describe(iris)


#dropping unnecessary columns
irisk<- iris[-c(1,6)]
irisk
```

```
#viewing outliers
boxplot(irisk)$out


# Eliminating outliers using Quartile method
iqr <- IQR(irisk$SepalWidthCm)
Q <- quantile(irisk$SepalWidthCm, probs=c(.25, .75), na.rm = FALSE)
eliminated <- subset(irisk, irisk$SepalWidthCm > (Q[1] - 1.5*iqr) & irisk$SepalWidthCm <
(Q[2]+1.5*iqr))
iqr <- IQR(eliminated$SepalWidthCm)
Q <- quantile(eliminated$SepalWidthCm, probs=c(.25, .75), na.rm = FALSE)
irisk <- subset(eliminated, eliminated$SepalWidthCm > (Q[1] - 1.5*iqr) & eliminated$SepalWidthCm
< (Q[2]+1.5*iqr))
remove(eliminated)

#scaling the data set
irissd <- scale(irisk)
head(irissd)


# Determining optimal number of clusters using Elbow Method
set.seed(123)

# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(irissd, k, nstart = 10 )$tot.withinss
}


# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15
```

```
# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)
print(wss_values)

#viewing the optimal number of clusters
fviz_nbclust(irissd, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")

# using the correct number of clusters
k4 <- kmeans(irissd, centers = 4, nstart = 25)
str(k4)

#plotting the cluster
fviz_cluster(k4, data = irissd)

#summary of k4
k4
```