**Data Mining**

**Assignment 2**

**November 6th, 2017**

**Surabhi Asati**

**Va7892**

**Task 1: Import the dataset**

Rcode:

mydata <-read.csv("/Users/surbhiasati/Desktop/Zombie Dataset.csv",header = TRUE)

str(mydata)

(mydata$Class <- as.factor(mydata$Class))

library(ggplot2)

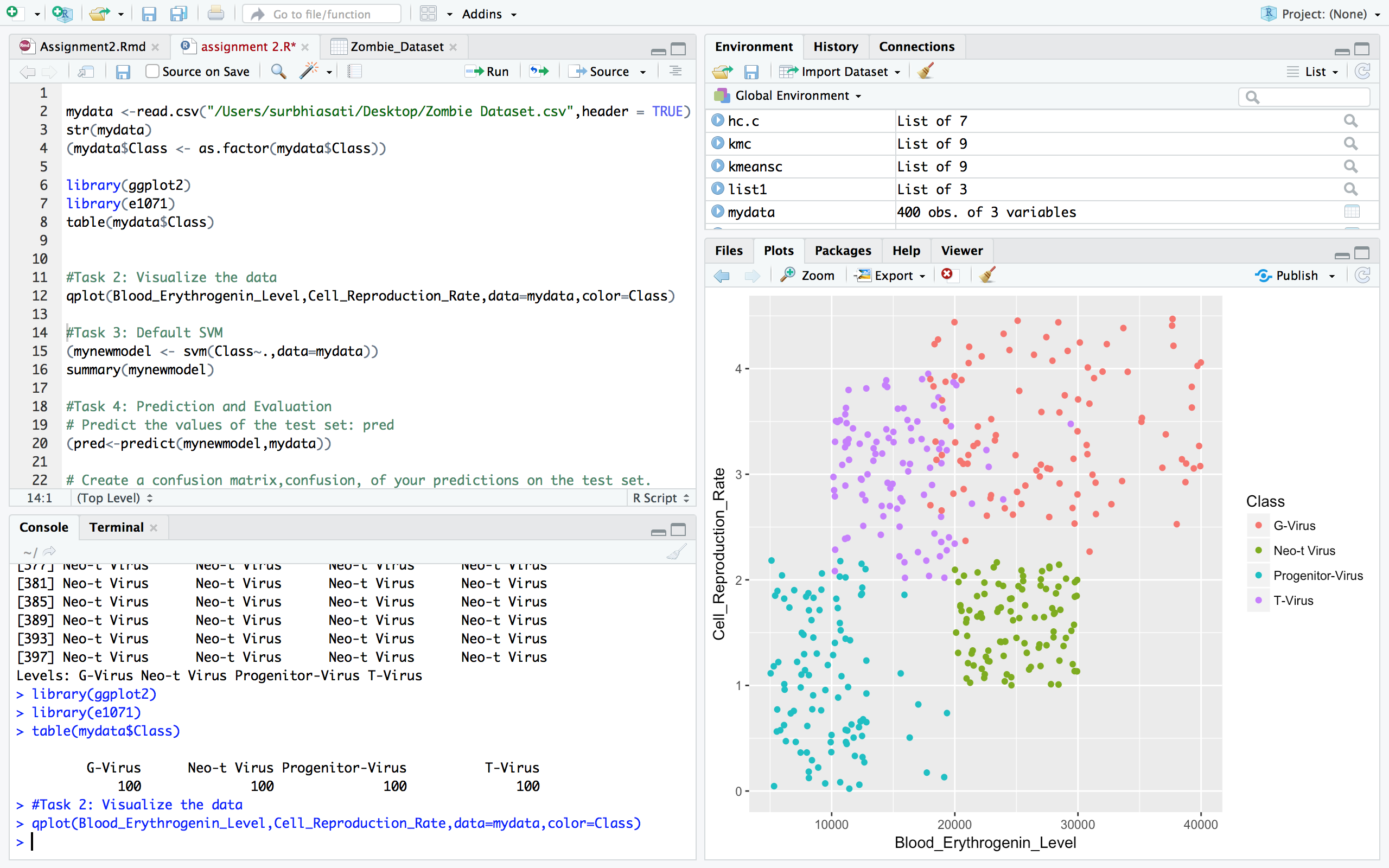
library(e1071)

table(mydata$Class)

**Task 2: Visualize the data**

From the plot, which two viruses seem to be mixed together in some section?

**T-Virus and G-Virus** seems to be mixed together in some section.



**Task 3: Default SVM**

How many support vectors are used to classify the T-Virus?

A-: T-Virus has **23** number of vectors

R Result:

Call:

svm(formula = Class ~ ., data = mydata)

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

gamma: 0.5

Number of Support Vectors: 125

( 45 41 16 23 )

Number of Classes: 4

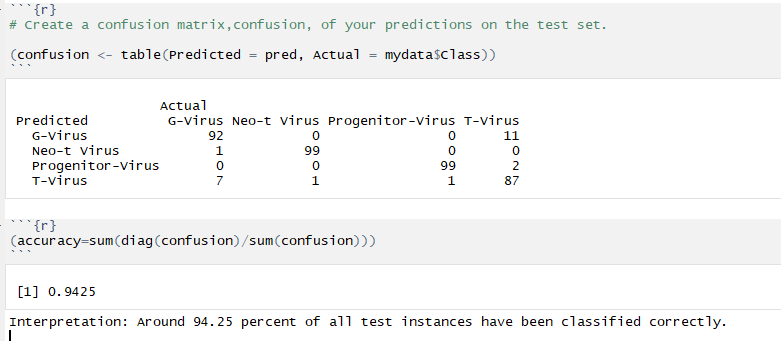
Levels:

G-Virus Neo-t Virus Progenitor-Virus T-Virus

**Task 4: Prediction and Evaluation**

# Predict the values of the test set in pred

(pred<-predict(mynewmodel,mydata))

****

Calculate the accuracy of the default classifier:

R code: accuracy=sum(diag(confusion)/sum(confusion))

(92+99+99+87)/400 = 94.25 which interpret that around 94.25 percent of all test instances have been classified correctly.

**Task 5: Tune the Model:**

tunedmodel<- tune(svm,Class~.,data=mydata,

ranges=list(cost=seq(0.1,10,1),gamma=2^(0.25:0.5)))

summary(tunedmodel)

**How many supporting vectors are there in the best model?**

Number of Support Vectors: 122

**Calculate the accuracy of the best model. How does it perform compared to the default model?**

**Task 6: Prepare Data for Clustering**

# Remove the class column

Ans: dataset=mydata[c(1,2)] OR

mydata$Class=NULL

#Second, standardize the data set. And show your code.?

Ans: scaledata=scale(mydata)

**Task 7: Hierarchical Clustering**

Q: Calculate the pairwise distance matrix. Do Not view the distance matrix.

Ans: R Code: pairwisedistance=dist(scaledata)

# Hierarchical Clustering Using Dentrogram (

hc.c=hclust(pairwisedistance,method="complete")

