NAME   : Hariprasad K K

REG NO : 19BCE7079

COURSE : Data_analytics

TOPIC   : Email classification- Spam or Ham

DATE   : 18-11-2021

```r
if(!require("kernlab")){
install.packages("kernlab")
}
```

```
> if(!require("kernlab")){
+   install.packages("kernlab")
+ }
Loading required package: kernlab
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
e proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/SATWIKA/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/kernlab_0.9-29.zip'
Content type 'application/zip' length 2849843 bytes (2.7 MB)
downloaded 2.7 MB

package 'kernlab' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\SATWIKA\AppData\Local\Temp\RtmpATMVOs\downloaded_packages
```

```r
library(kernlab)
data(spam)
str(spam[,1:5])
```

```
> library(kernlab)
> data(spam)
> str(spam[,1:5])
'data.frame':   4601 obs. of  5 variables:
 $ make   : num  0 0.21 0.06 0 0 0 0 0 0.15 0.06 ...
 $ address: num  0.64 0.28 0 0 0 0 0 0 0 0.12 ...
 $ all    : num  0.64 0.5 0.71 0 0 0 0 0 0.46 0.77 ...
 $ num3d  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ our    : num  0.32 0.14 1.23 0.63 0.63 1.85 1.92 1.88 0.61 0.19 ...

> table(spam$type)

nonspam    spam
   2788    1813
```

```r
install.packages("caTools")
library(caTools)
```

```
> install.packages("caTools")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
e proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/SATWIKA/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
also installing the dependency 'bitops'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/bitops_1.0-7.zip'
Content type 'application/zip' length 42425 bytes (41 KB)
downloaded 41 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/caTools_1.18.2.zip'
Content type 'application/zip' length 316450 bytes (309 KB)
downloaded 309 KB

package 'bitops' successfully unpacked and MD5 sums checked
package 'caTools' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\SATWIKA\AppData\Local\Temp\RtmpATMVOs\downloaded_packages
> library(caTools)
Warning message:
package 'caTools' was built under R version 4.0.5
```

set.seed(32)

split = sample.split(spam$type, SplitRatio = 0.7)

trainSpam = subset(spam, split == T)

testSpam = subset(spam, split ==F)

names(spam)

```
> set.seed(32)
> split = sample.split(spam$type, SplitRatio = 0.7)
> trainSpam = subset(spam, split == T)
> testSpam = subset(spam, split ==F)
> names(spam)
 [1] "make"            "address"           "all"               "num3d"            "our"          "over"
 [7] "remove"          "internet"          "order"             "mail"             "receive"      "will"
[13] "people"          "report"            "addresses"         "free"             "business"     "email"
[19] "you"             "credit"            "your"              "font"             "num000"       "money"
[25] "hp"              "hpl"               "george"            "num650"           "lab"          "labs"
[31] "telnet"          "num857"            "data"              "num415"           "num85"        "technology"
[37] "num1999"         "parts"             "pm"                "direct"           "cs"           "meeting"
[43] "original"        "project"           "re"                "edu"              "table"        "conference"
[49] "charSemicolon"   "charRoundbracket"  "charSquarebracket" "charExclamation"  "charDollar"   "charHash"
[55] "capitalAve"      "capitalLong"       "capitalTotal"      "type"
```
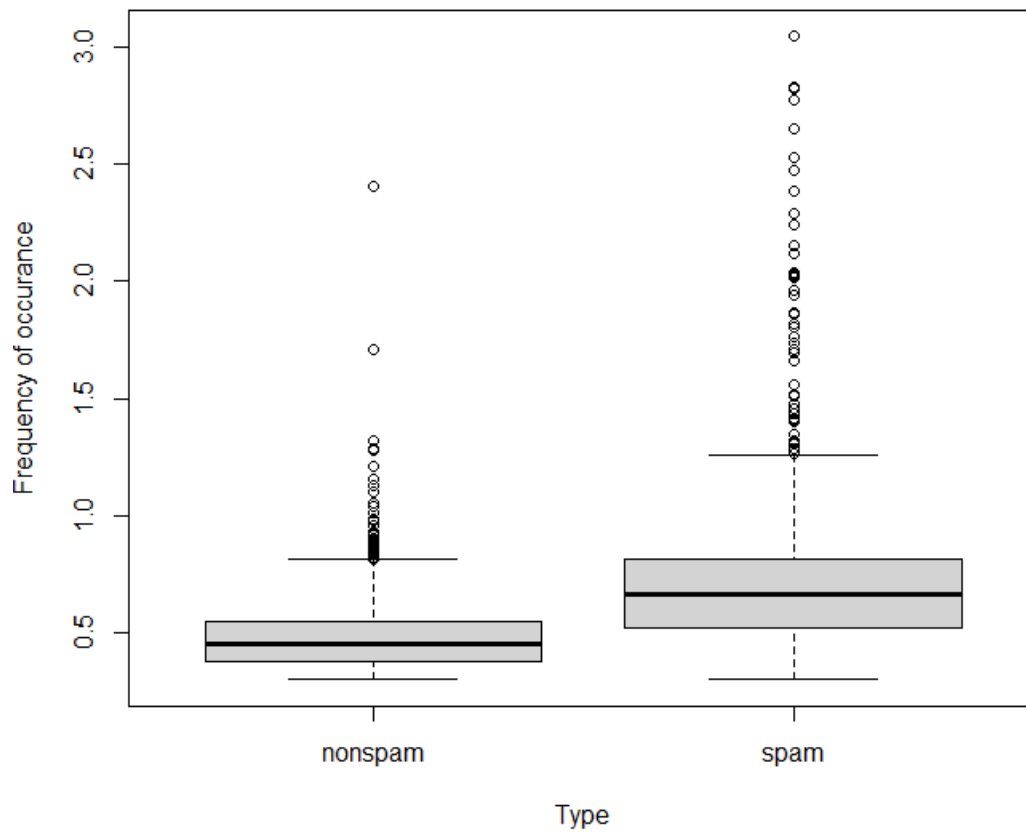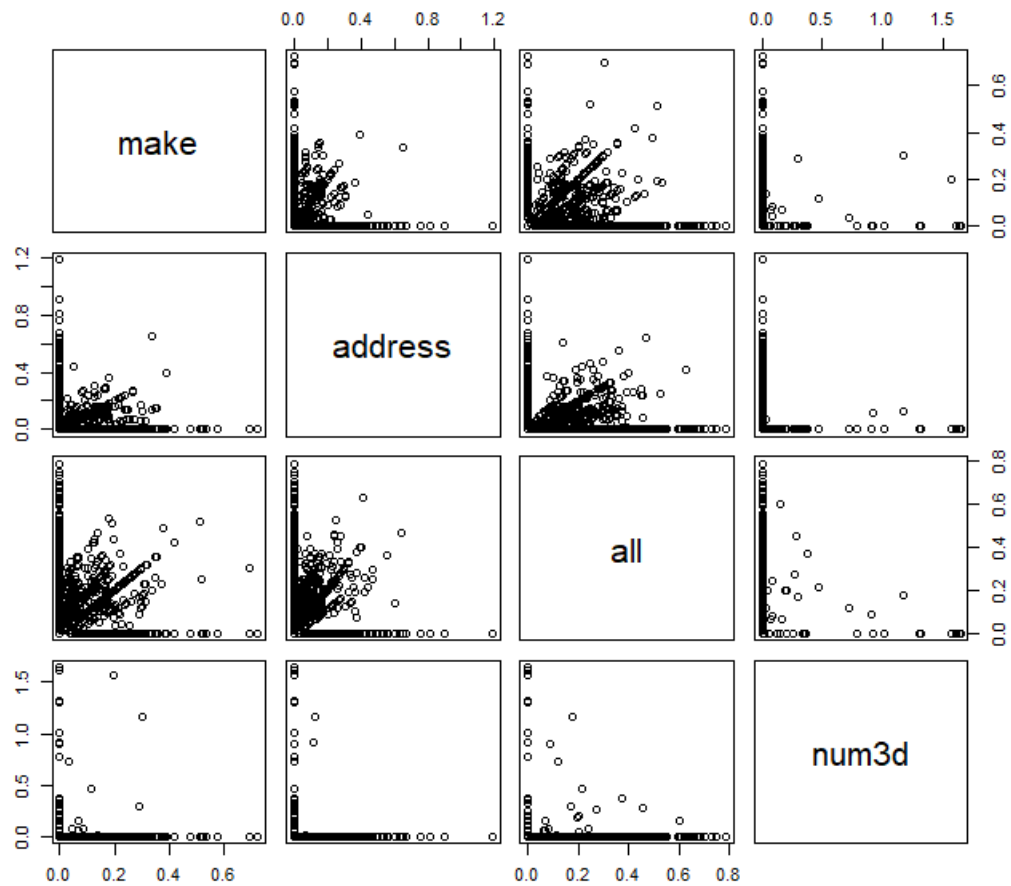
head(spam)[,1:6]

```
> head(spam)[,1:6]
  make address  all num3d  our over
1 0.00    0.64 0.64     0 0.32 0.00
2 0.21    0.28 0.50     0 0.14 0.28
3 0.06    0.00 0.71     0 1.23 0.19
4 0.00    0.00 0.00     0 0.63 0.00
5 0.00    0.00 0.00     0 0.63 0.00
6 0.00    0.00 0.00     0 1.85 0.00
```

```
plot(log10(trainSpam$capitalAve+1) ~trainSpam$type,
ylab = "Frequency of occurance",xlab = "Type")
```



```
plot(log10(trainSpam[,1:4]+1))
```

```
mdist = dist(t(log(trainSpam[,1:55]+1)))
hclustering = hclust(mdist)
plot(as.dendrogram(hclustering))
```

```
trainSpam$numType = as.numeric(trainSpam$type)-1
costFunc = function(x,y){
sum(x != (y > 0.5))
}
cvError = rep(NULL, 55)
cvError
```

```
> trainSpam$numType = as.numeric(trainSpam$type)-1
> costFunc = function(x,y){
+    sum(x != (y > 0.5))
+ }
> cvError = rep(NULL, 55)
> cvError
NULL
```

```
install.packages("boot")
library(boot)
```

```
> install.packages("boot")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
e proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/SATWIKA/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/boot_1.3-28.zip'
Content type 'application/zip' length 640731 bytes (625 KB)
downloaded 625 KB

package 'boot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\SATWIKA\AppData\Local\Temp\RtmpATMVOs\downloaded_packages
> library(boot)
Warning message:
package 'boot' was built under R version 4.0.5
```

```r
suppressWarnings(
for(i in 1:55){
lmFormula = reformulate(names(trainSpam)[i], response = "numType")
glmFit = glm(lmFormula, family = "binomial", data = trainSpam)
cvError[i] = cv.glm(trainSpam, glmFit, costFunc, 2)$delta[2]
}
)
predModel = suppressWarnings(
glm(
numType ~ charDollar+charExclamation+remove+money+free,
family = "binomial",
```

```
data = trainSpam
)
)
pred_y = as.character(
ifelse(
as.numeric(predict(predModel, testSpam))>0.5,
"spam",
"nonspam"
)
)
crossTab = table(pred_y,testSpam$type)
crossTab
accuracy = sum(crossTab[-c(2:3)])/sum(crossTab[1:4])
accuracy
```

```
> suppresswarnings(
+    for(i in 1:55){
+      lmFormula = reformulate(names(trainSpam)[i], response = "numType")
+      glmFit = glm(lmFormula, family = "binomial", data = trainSpam)
+      cvError[i] = cv.glm(trainSpam, glmFit, costFunc, 2)$delta[2]
+    }
+ )
>
> predModel = suppresswarnings(
+    glm(
+      numType ~ charDollar+charExclamation+remove+money+free,
+      family = "binomial",
+      data = trainSpam
+    )
+ )
> pred_y = as.character(
+    ifelse(
+      as.numeric(predict(predModel, testSpam))>0.5,
+      "spam",
+      "nonspam"
+    )
+ )

> crossTab = table(pred_y,testSpam$type)
> crossTab

pred_y     nonspam spam
  nonspam      810  190
  spam          26  354
> accuracy = sum(crossTab[-c(2:3)])/sum(crossTab[1:4])
> accuracY
> accuracy = sum(crossTab[-c(2:3)])/sum(crossTab[1:4])
> accuracy
[1] 0.8434783
```