STA 141 Assignment 4 Report

Shuxin Li

912525987

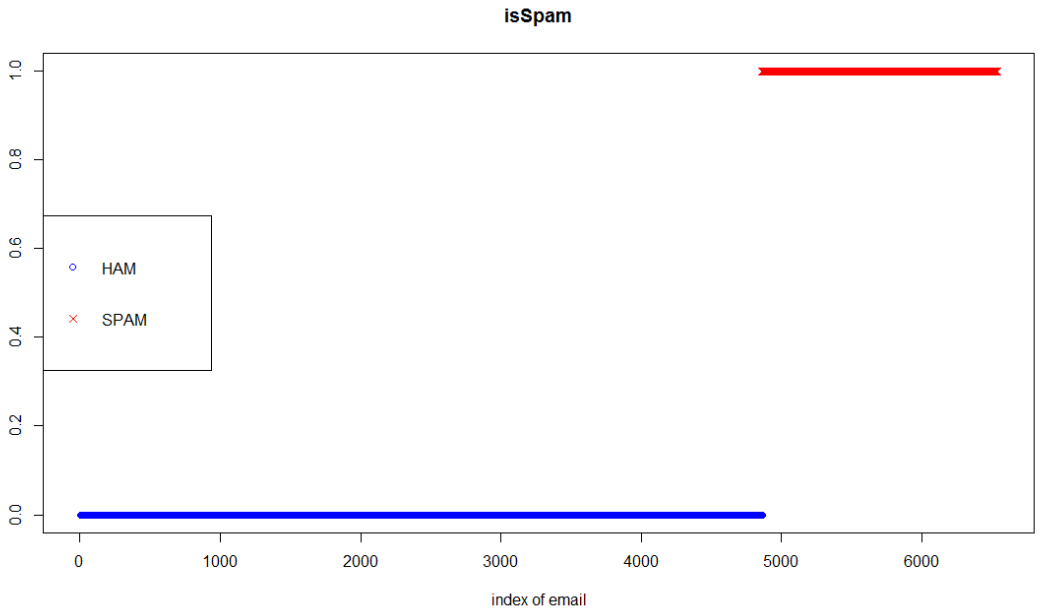
- **Methods explanation**

First, we mainly use three methods to examine if the variable has ability to distinguish SPAM from the emails. One of the method is to draw the scatterplot to show the effectiveness of the variable directly. The second is to draw a boxplot to show the distribution of variables' values depending on the SPAM emails and HAM emails. The last method is to compute the frequency of FALSE and TRUE value in HAM and SPAM emails, then compute the ratio of TRUE value in HAM and the ratio of TRUE value in SPAM. We can see that if the two ratios have obvious difference, it indicates that the variable has a good ability to distinguish SPAM from the emails. Otherwise, it indicates that the variable may not be used to distinguish SPAM from the emails. The first and third method is adapted for the logical variable, in addition, the first and second method is adapted for another type of value.

- **Variables effectiveness test**

1. The variable "isSpam": whether mail is Spam (TRUE) or Ham (FALSE). You can compute this from the name of the messages in the top-level list of messages.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

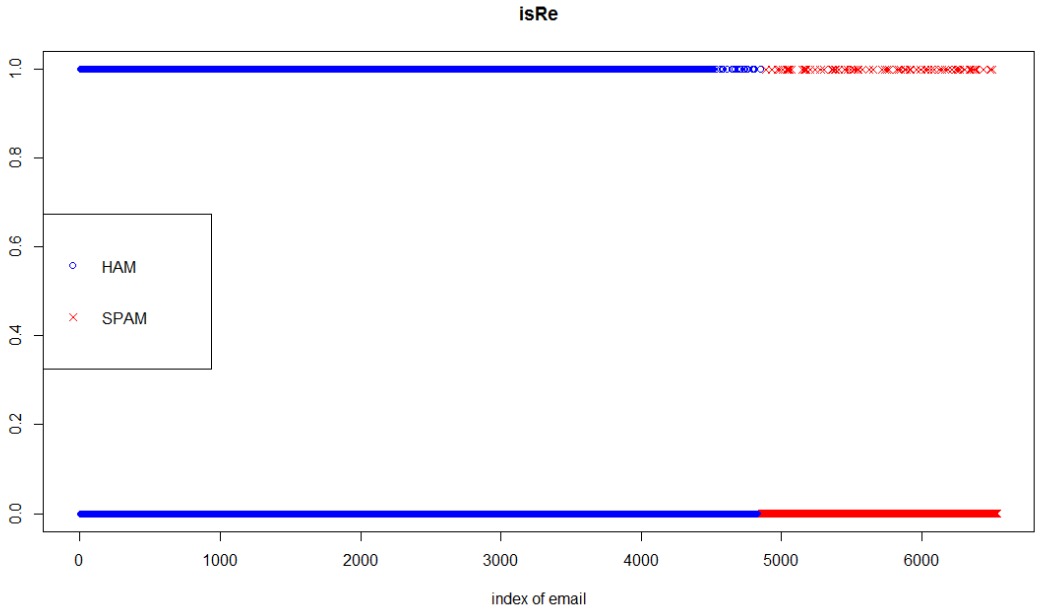
The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
                   0                             1
```

From the scatterplot, it indicates that all the SPAM is TRUE and all the HAM is not TRUE.
From the table of ratio, the ratio of TRUE in HAM is zero, however, the ratio of TRUE in SPAM is 1. Thus, there is a largest difference between the two ratios whch indicates the variable is the best.

2. The variable “isRe”: if the string Re: appears as the first word in the subject of the message.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

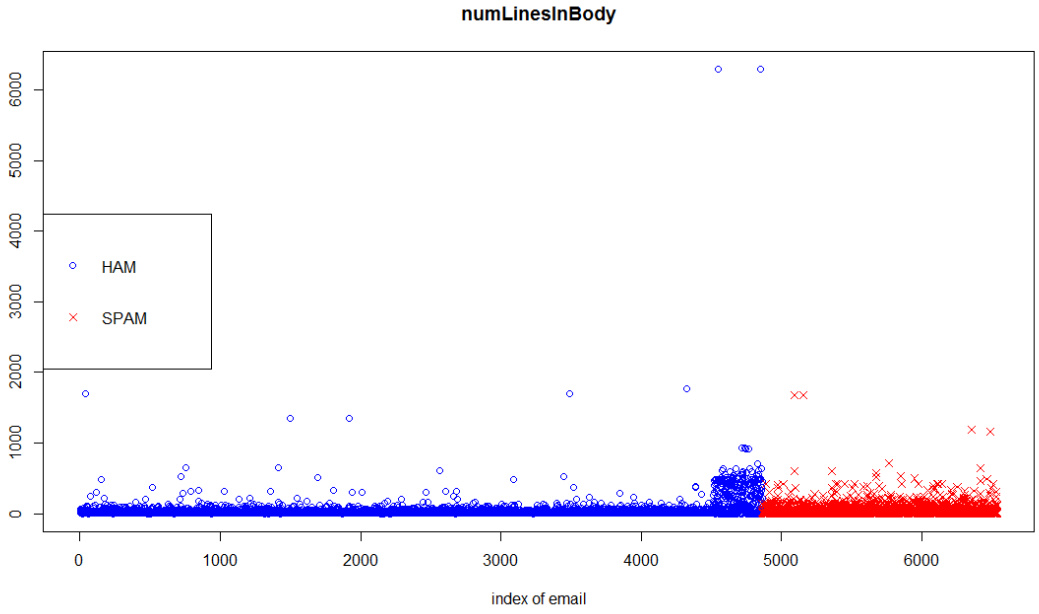
The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

```
ratio of TRUE in HAM      The ratio of TRUE in SPAM
           0.43323747                     0.04588796
```

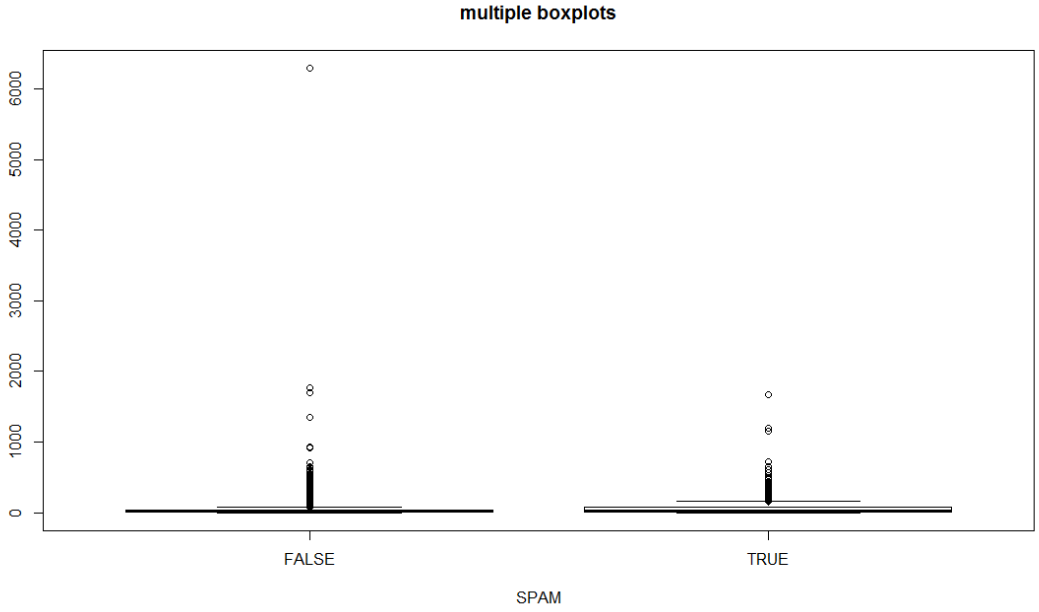
From the scatterplot, we just can see that there seems less SPAM to have Re as the first word.
From the table of ratio, the ratio of TRUE in HAM is 0.433 and the ratio of TRUE in SPAM is 0.046. It shows an obvious difference which means if a message has “Re” as the first word in the subject of the message, it is more possible to be a HAM not SPAM.

3. The variable "numLinesInBody": a count of the number of lines in the body of the email message.

Method1: scatterplot



Method2: boxplot



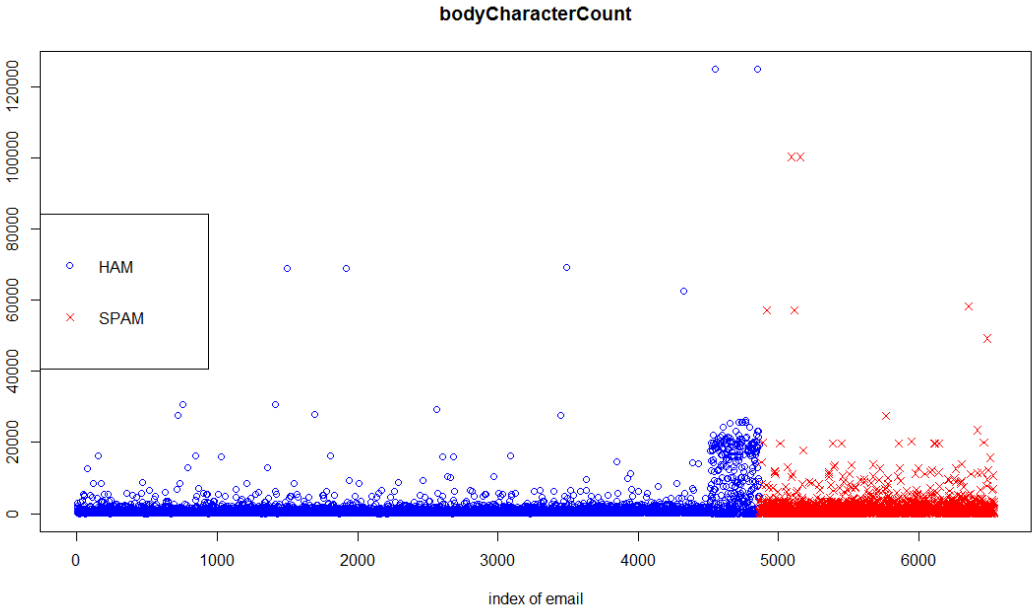
From the scatterplot, it indicates that the variable is not good to distinguish the SPAM emails from HAM emails.

From the boxplot, it indicates that there is no obvious difference between SPAM and HAM eamils

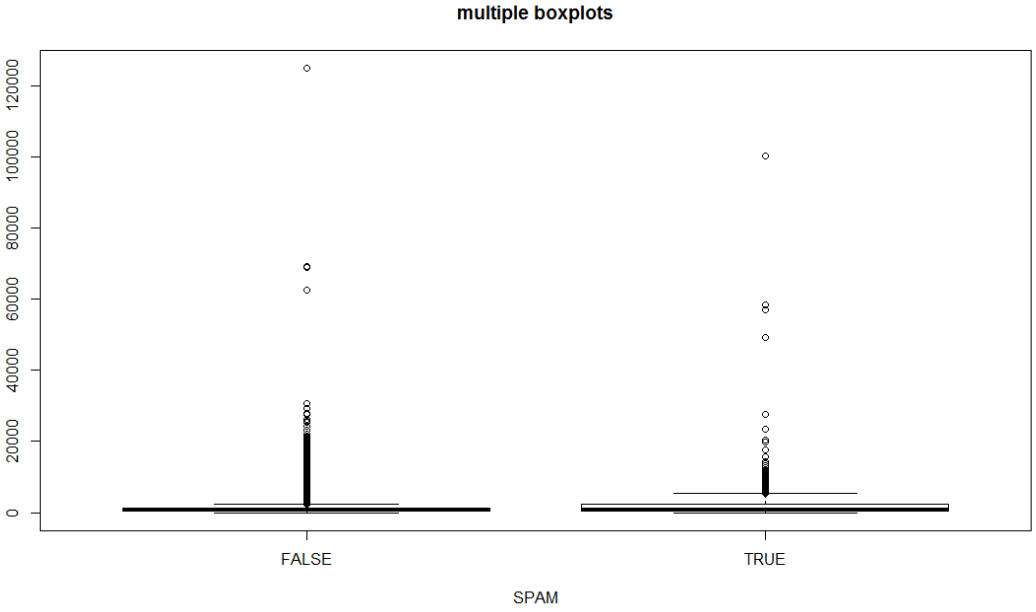
Thus, it is not a good variable to examine the SPAM email.

4. The variable "bodyCharacterCount": the number of characters in the body of the email message.

Method1: scatterplot



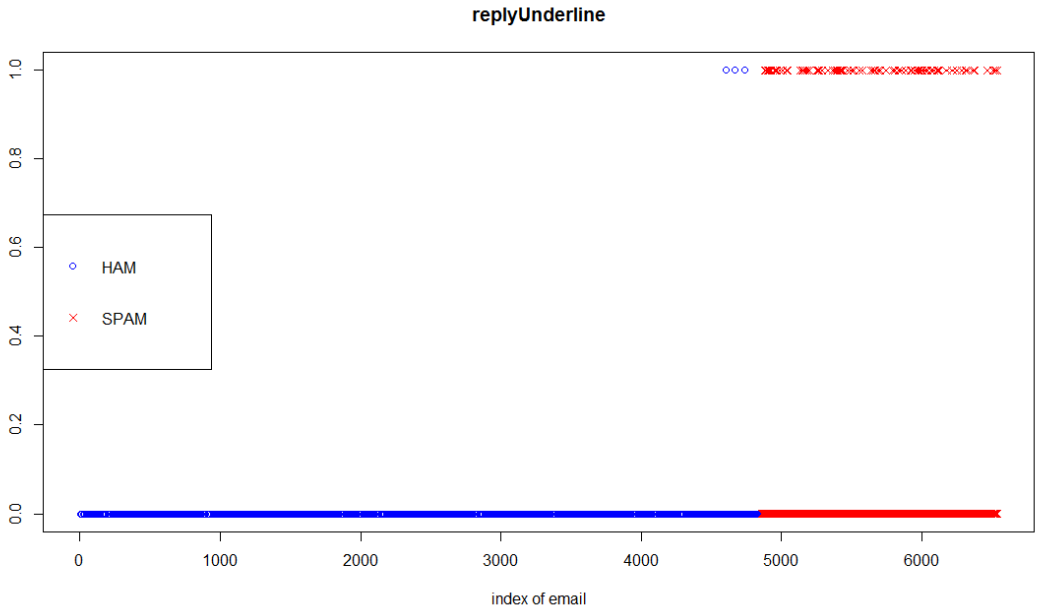
Method2: boxplot



From the scatterplot, it indicates that the variable is not good to distinguish the SPAM emails from HAM emails.
From the boxplot, it indicates that the distributions of the number of characters in the body of the email message depending on the SPAM and HAM seems not much different.
Thus, it is not a good variable to examine the SPAM email.

5. The variable "replyUnderline": whether the Reply-To field in the header has an underline and numbers/letters.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

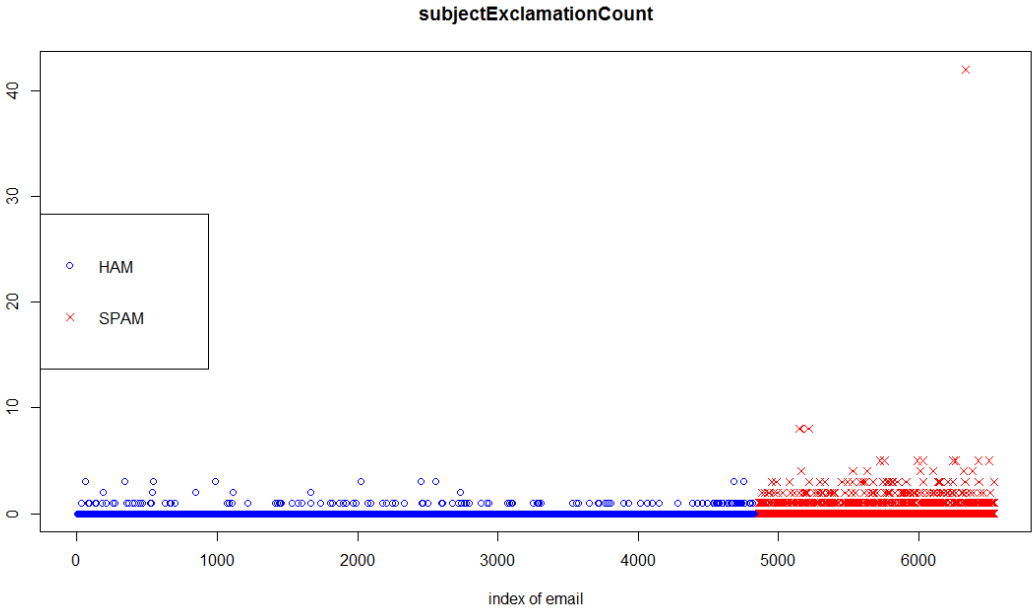
```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
         0.002220577                   0.114838710
```

From the scatterplot, it indicates that HAM emails almost do not have an underline and numbers/letters.

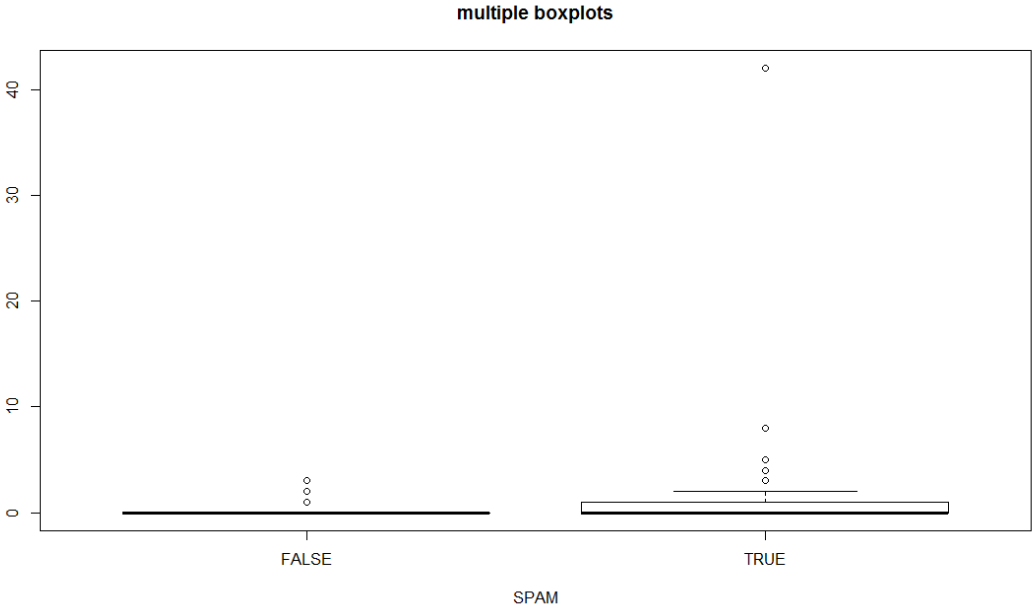
From the table of ratio, the ratio of TRUE in HAM is 0.00222, however, the ratio of TRUE in SPAM is 0.1148, which indicates that if the variable is true, we tend to believe that it is a SPAM email. Thus, the variable is partly good for examining the SPAM email.

6. The variable "subjectExclamationCount": a count of the number of exclamation marks (!) in the subject of the message.

Method1: scatterplot



Method2: boxplot

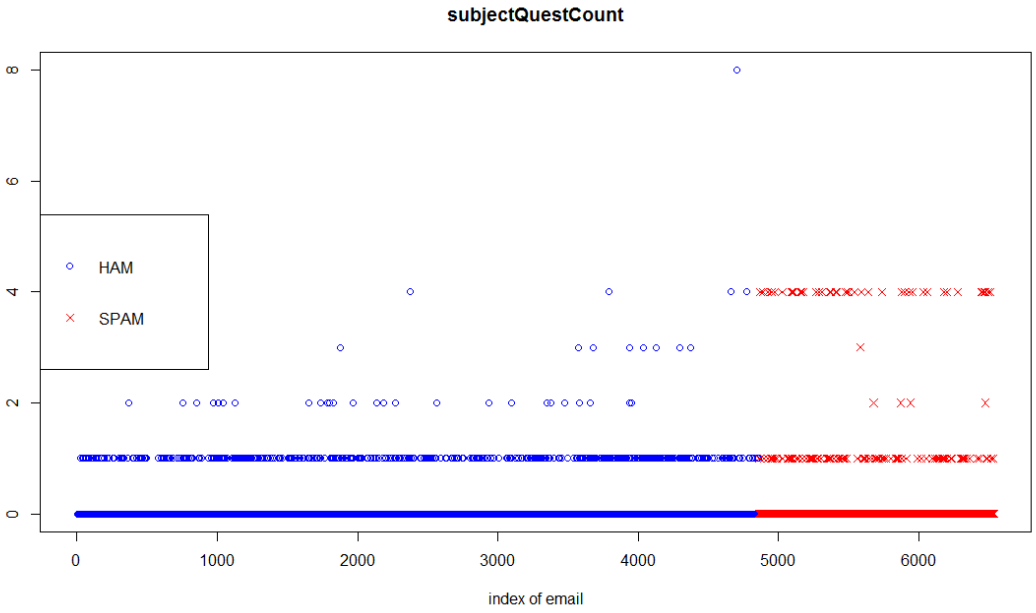


From the scatterplot, it indicates that SPAM emails seem to have more number of exclamation marks than HAM emails.

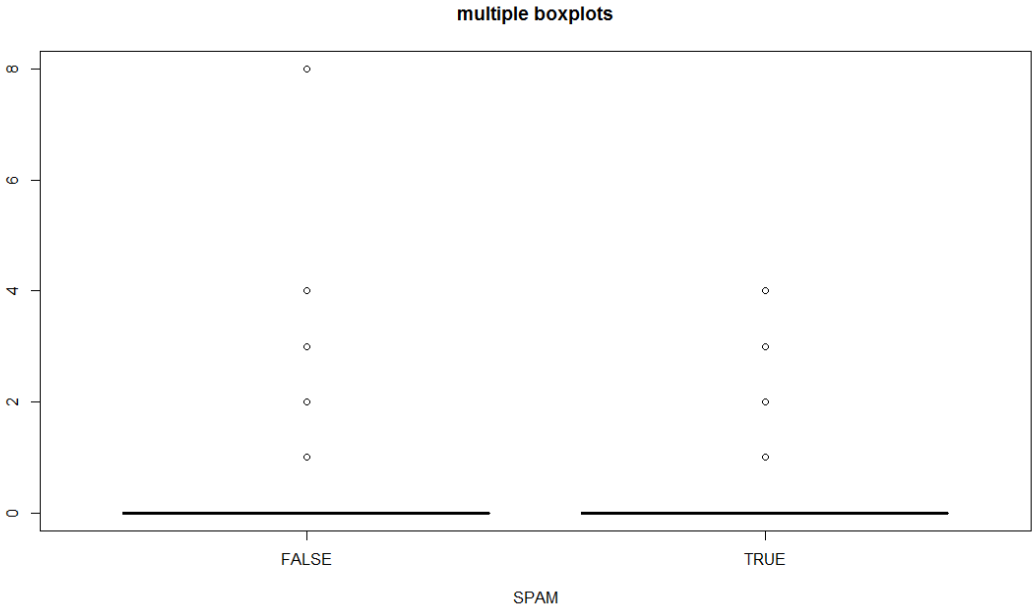
From the boxplot, it indicates that the median of two distributions are close, however, the Q3 of the distribution based on SPAM is larger than the distribution based on HAM, which means if the number of exclamation marks great than 0, it is more possible to believe the email is SPAM.

7. The variable "subjectQuestCount": the number of question marks in the subject.

Method1: scatterplot



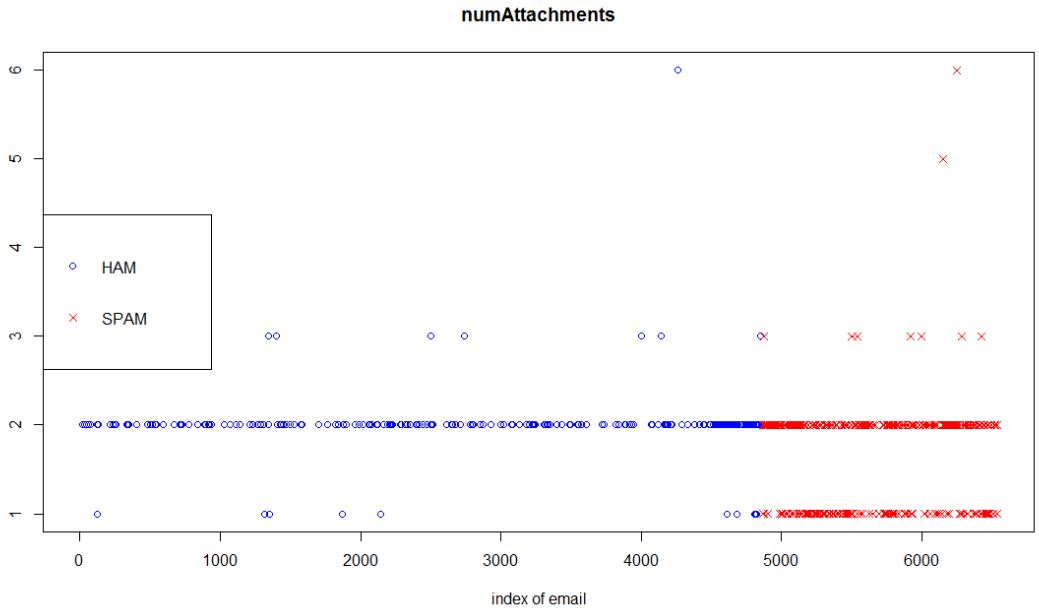
Method2: boxplot



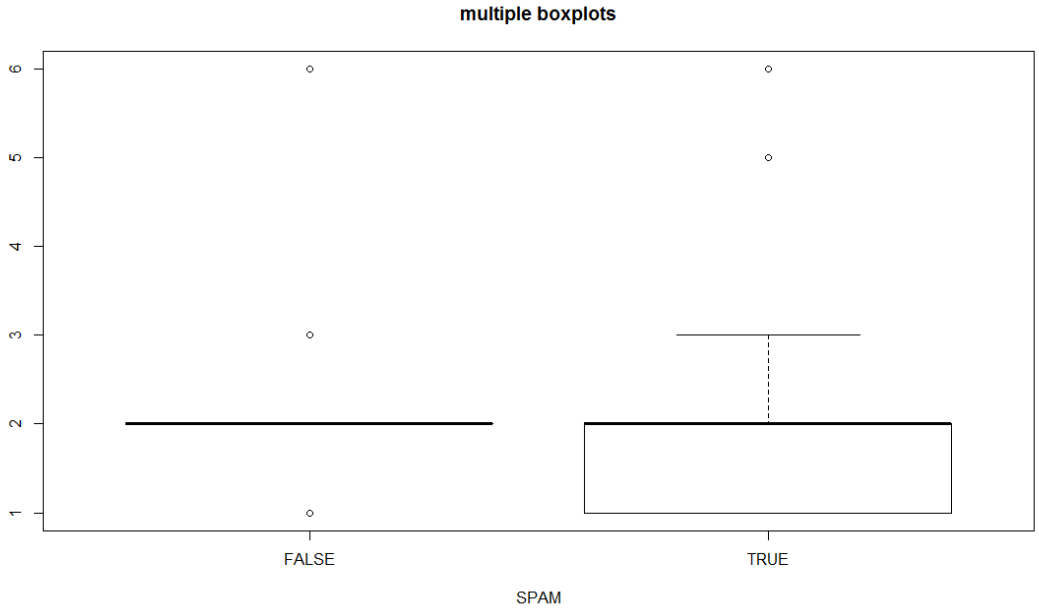
From the scatterplot, it indicates that the number of question marks in the subject in SPAM and HAM emails is not obviously different.
From the boxplot, it indicates that the number of question marks in the subject in SPAM and HAM emails has nearly distribution.
Thus, it is not a good variable to detect SPAM emails.

8. The variable "numAttachments": the number of attachments in the message.

Method1: scatterplot



Method2: boxplot

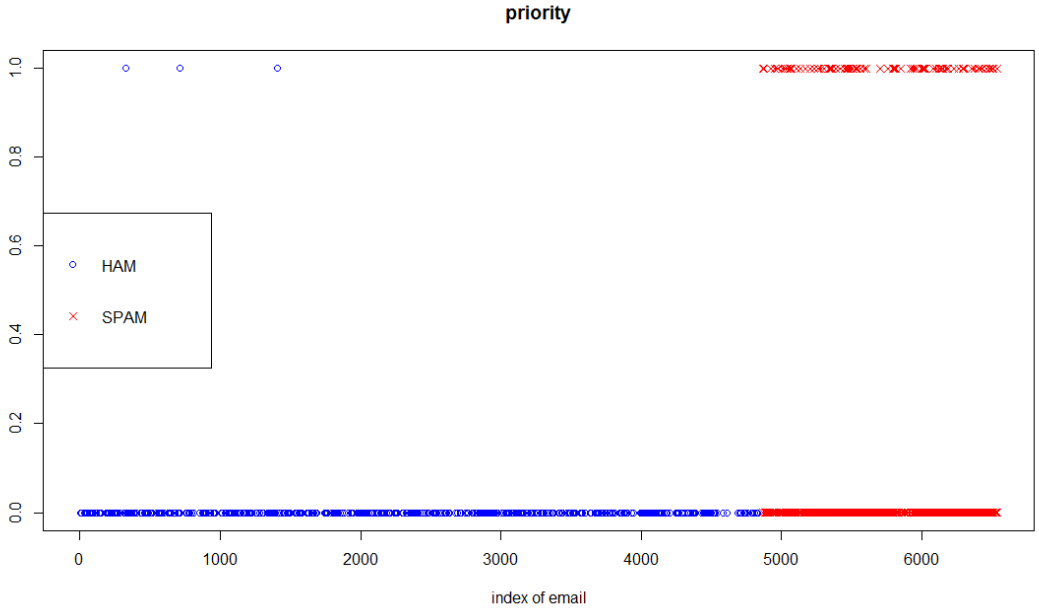


From the scatterplot, it indicates that if the email has one attachment, it is more possible to be SPAM email.

From the boxplot, it indicates that the distribution of number of attachments in the message based on SPAM is different from the distribution based on HAM, especially when the email has one attachment.

9. The variable "priority": whether the message's header had an X-Priority or X-Msmail-Priority that was set to high.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

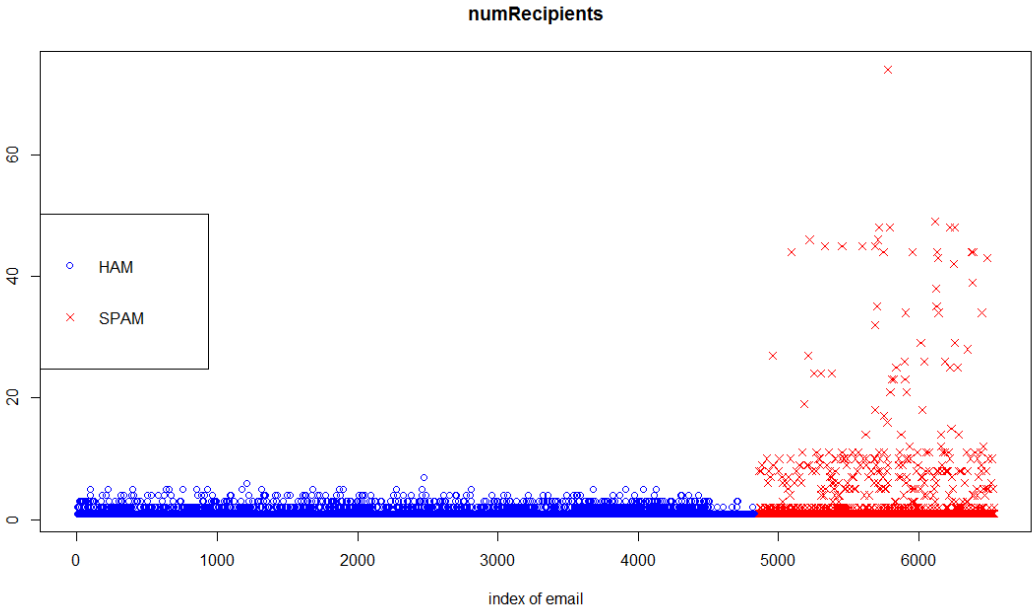
```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
          0.005366726                   0.182163188
```

From the scatterplot, it indicates that there are much more SPAM emails than HAM emails whose message's header had an X-Priority or X-Msmail-Priority that was set to high.

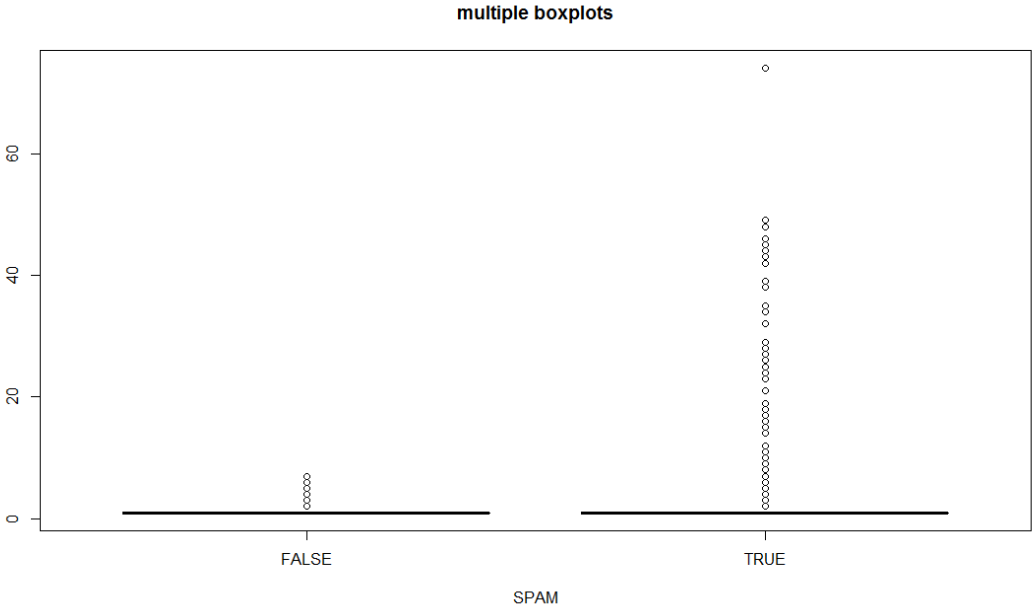
From the table of ratio, the ratio of TRUE in HAM is 0.0054, however, the ratio of TRUE in SPAM is 0.18216. Thus, if the variable is TRUE, we tend to believe that the email is SPAM.

10. The variable "numRecipients": the number of recipients in the To, Cc fields

Method: scatterplot



Method2: boxplot

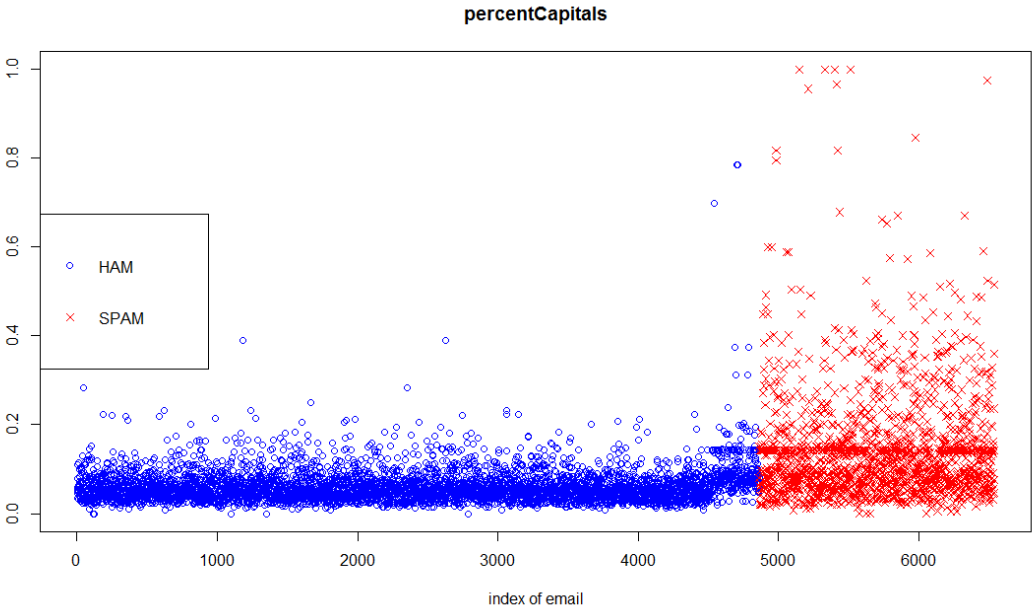


From the scatterplot, it indicates that SPAM emails seem to have more number of recipients in the To, Cc fields.

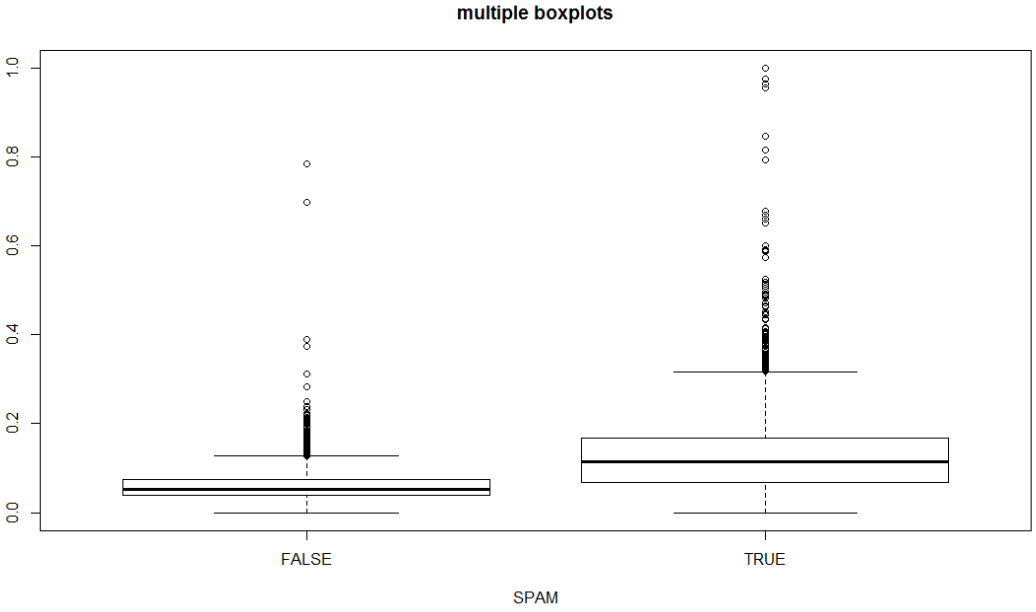
From the boxplot, it indicates that if the number of recipients in the To, Cc fields is more than 10, we tend to believe that it is a SPAM email.

11. The variable "percentCapitals": the percentage of the characters in the body of the email that are upper case (excluding blanks, numbers, and punctuation)

Method1: scatterplot



Method2: boxplot

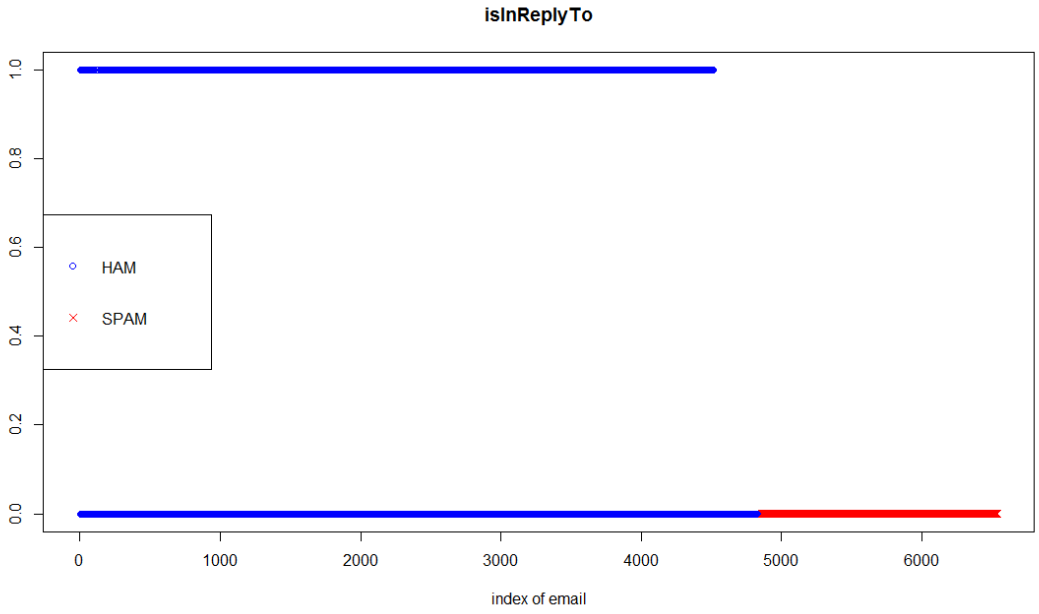


From the scatterplot, it indicates that the SPAM email seems to have larger percentage of the characters in the body of the email that are upper case.

From the boxplot, it indicates that the median of the percentage of the characters in the body of the SPAM email is larger than the median of the percentage of the characters in the body of the HAM email, which demonstrate that if the percentage is more than 0.1, we tend to believe that it is a SPAM email.

12. The variable "isInReplyTo": whether the header of the message has an In-Reply-To field.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

ratio of TRUE in HAM	The ratio of TRUE in SPAM
0.4019326	0.0000000

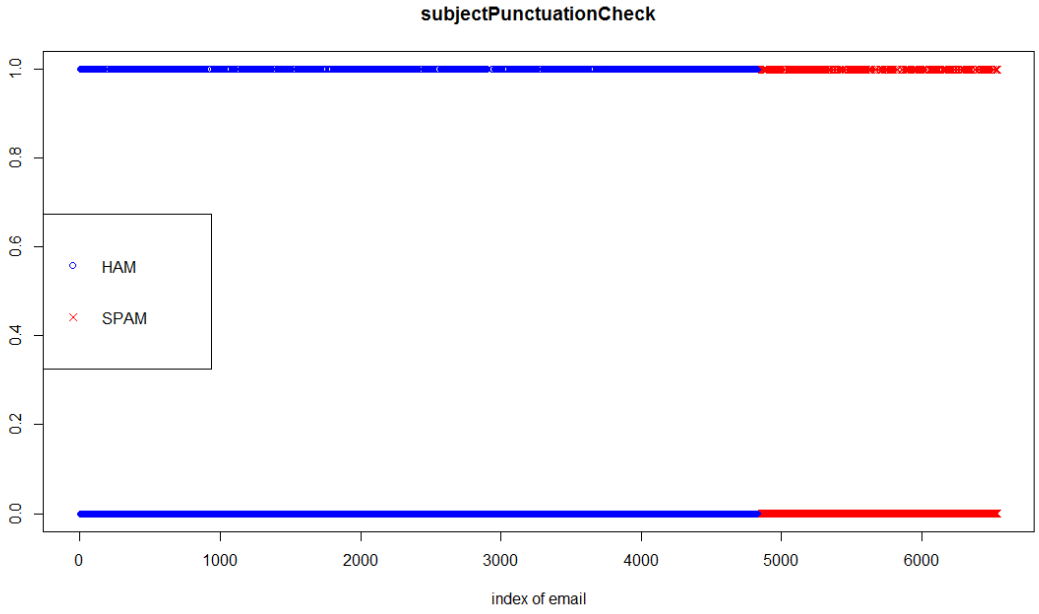
From the scatterplot, it indicates that SPAM do not have the In-Reply-To field

From the table of ratio, the ratio of TRUE in HAM is 0.4, however, the ratio of TRUE in SPAM is 0.

It indicates that if the variable is true, it is a HAM email, but it can not detect the SPAM email.

13. The variable "subjectPunctuationCheck": whether the subject has punctuation or digits surrounded by characters, e.g. V?agra and pay1ng, but not New!

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

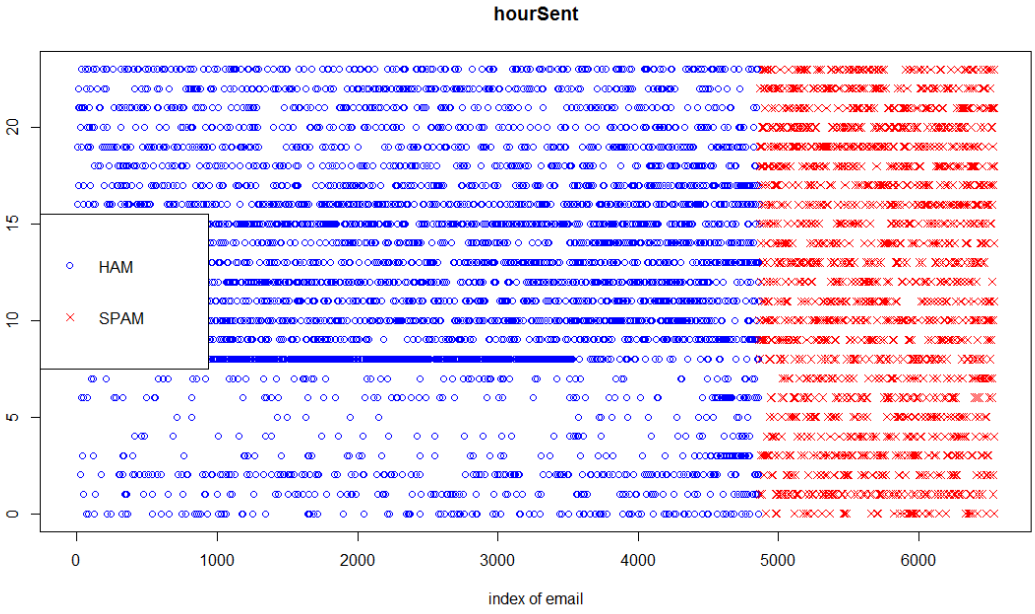
```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
           0.2578189                     0.1909308
```

From the scatterplot, it indicates that the variable can not be used in distinguishing the SPAM from the HAM.

From the table of ratio, the ratio of TRUE in HAM is 0.2578, however, the ratio of TRUE in SPAM is 0.1909, which indicates that the variable is not good at detecting SPAM emails.

14. The variable "hourSent": the hour in the day the mail was sent (0 -- 23)

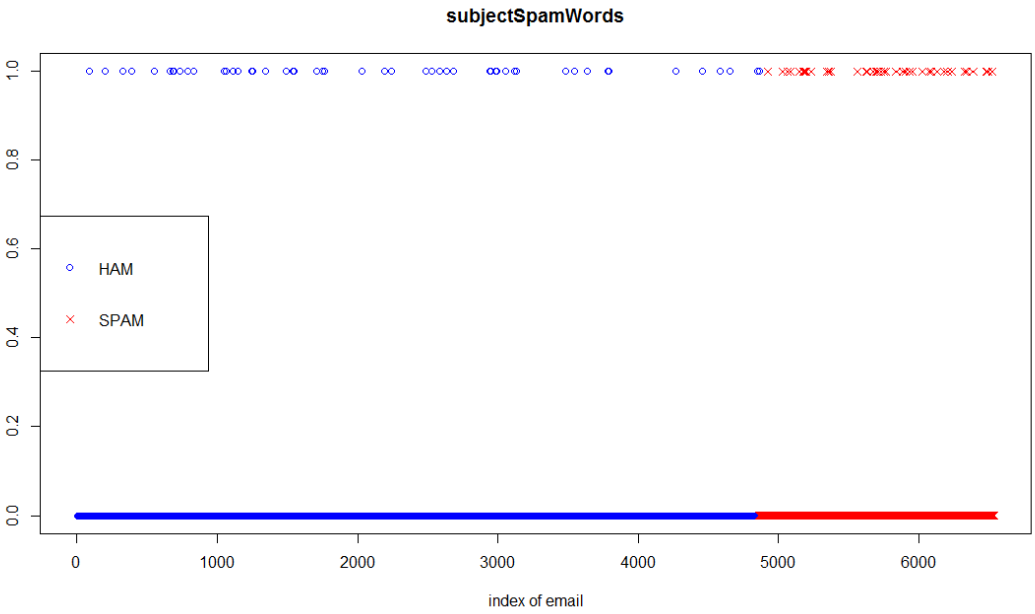
Method1: scatterplot



From the scatterplot, it indicates that there is no obviously relationship between the the type of email and the hour in the day the mail was sent. Thus, the variable can not be used to detect the SPAM emails.

15. The variable "subjectSpamWords": whether the subject contains one of the following phrases: viagra, pounds, free, weight, guarantee, millions, dollars, credit, risk, prescription, generic, drug, money back, credit card.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

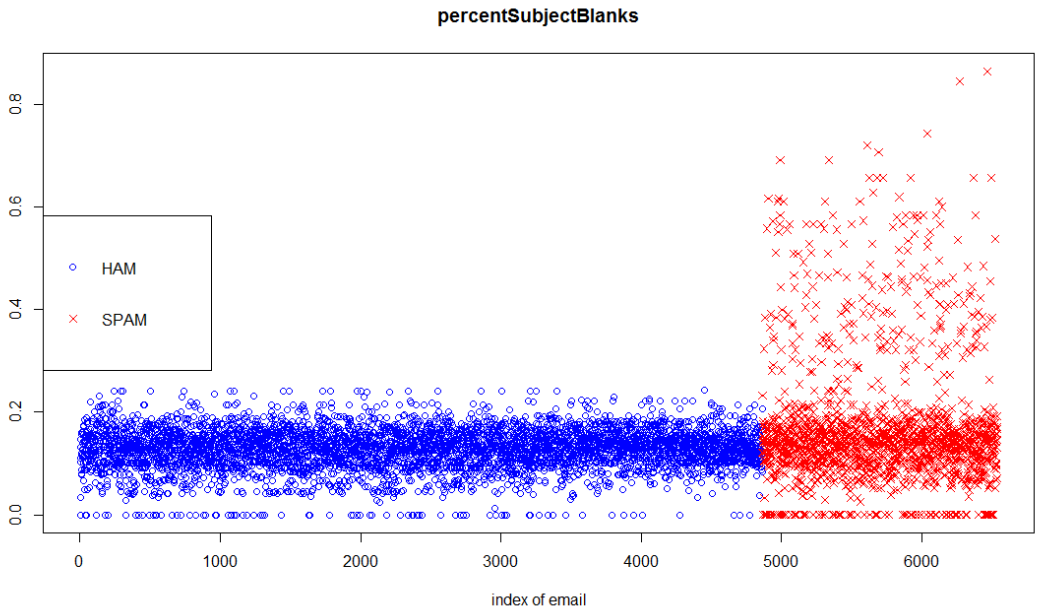
The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

```
ratio of TRUE in HAM      The ratio of TRUE in SPAM
           0.01028807                      0.02863962
```

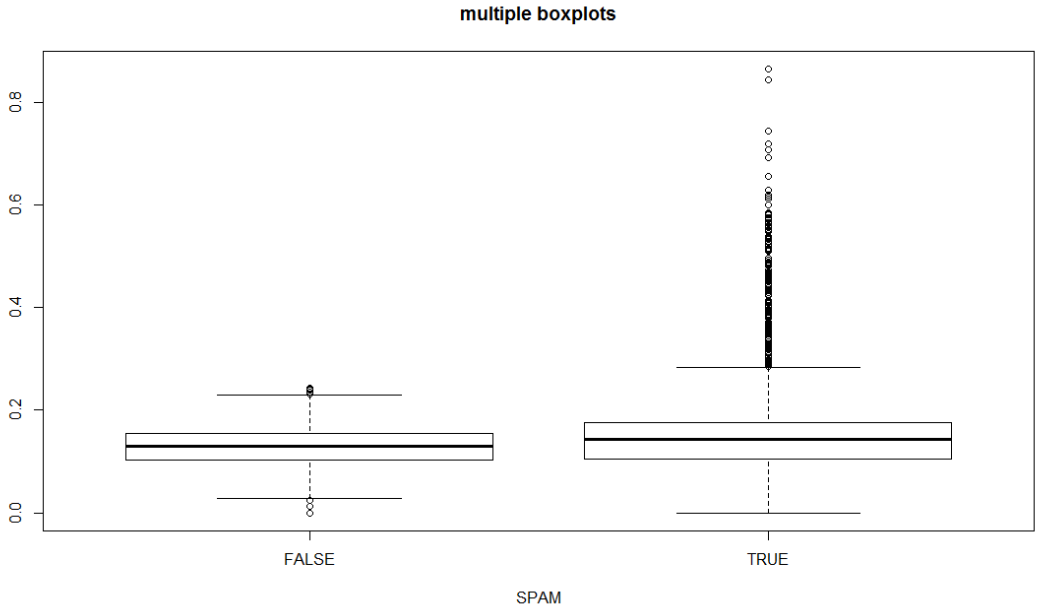
From the scatterplot, it indicates that both SPAM email and HAM email have a few spam words. From the table of ratio, the ratio of TRUE in HAM is 0.0102, however, the ratio of TRUE in SPAM is 0.028, which indicates that the variable may not be used to detect the SPAM emai for the two ratios are close.

16. The variable "percentSubjectBlanks": the percentage of blanks in the subject.

Method1: scatterplot



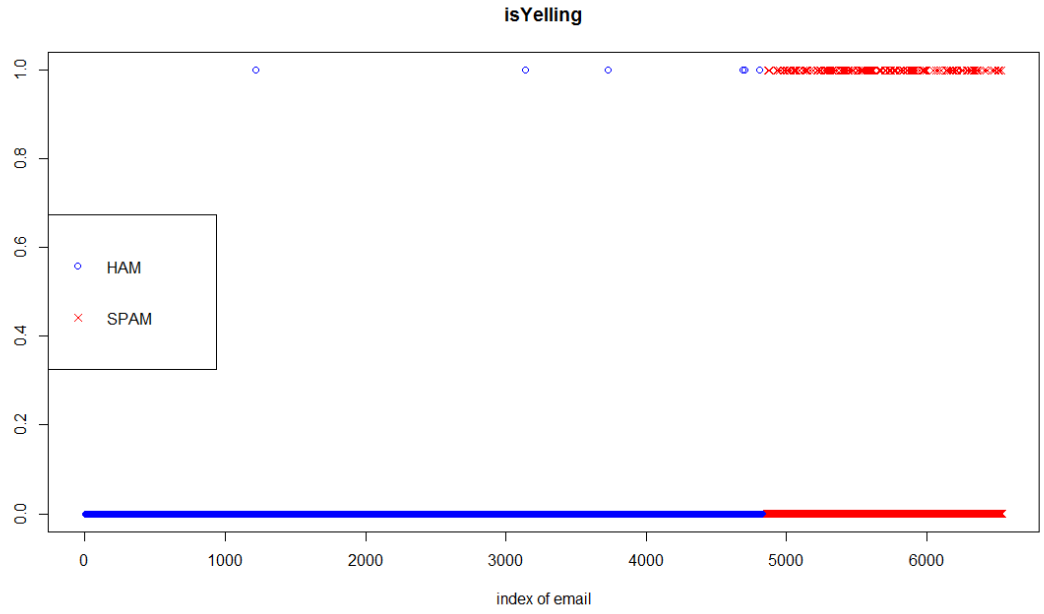
Method2: boxplot



From the scatterplot and the boxplot, it indicates that if the percentage of blanks in the subject is larger than 0.2, it is more likely to be the SPAM email.

17. The variable "isYelling": whether the Subject of the mail is in capital letters.

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

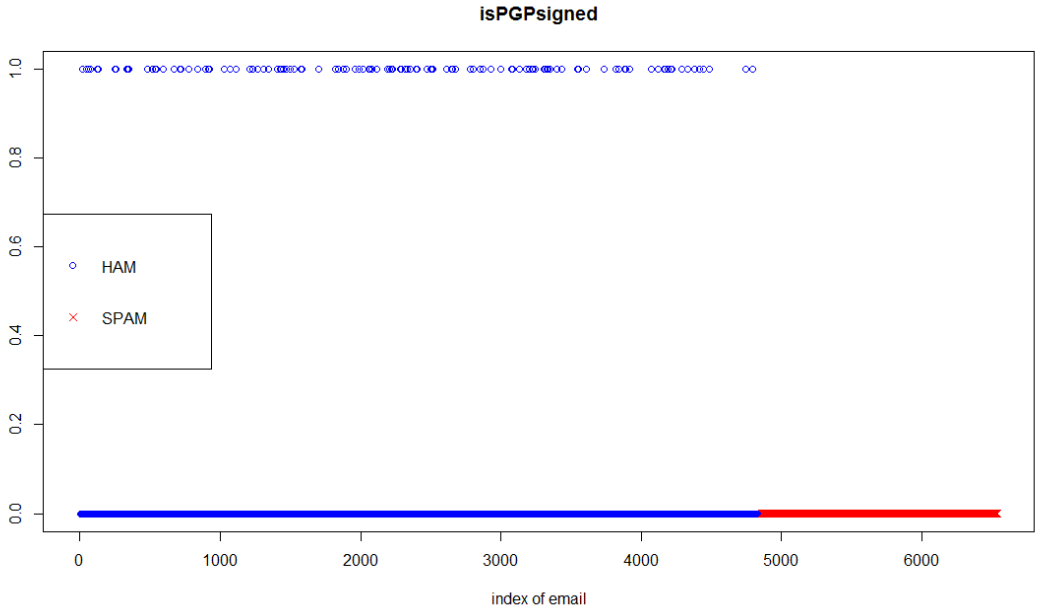
ratio of TRUE in HAM	The ratio of TRUE in SPAM
0.001440329	0.092482100

From the scatterplot, it indicates that if the subject of the email is in capital letters, it is more likely to be SPAM emails.

From the table of ratio, the ratio of TRUE in HAM is 0.0014, however, the ratio of TRUE in SPAM is 0.092. It indicates that when the variable is TRUE, we tend to believe that the email is SPAM

18. The variable "isPGPsigned": indicates whether the mail was digitally signed (e.g. using PGP or GPG)

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

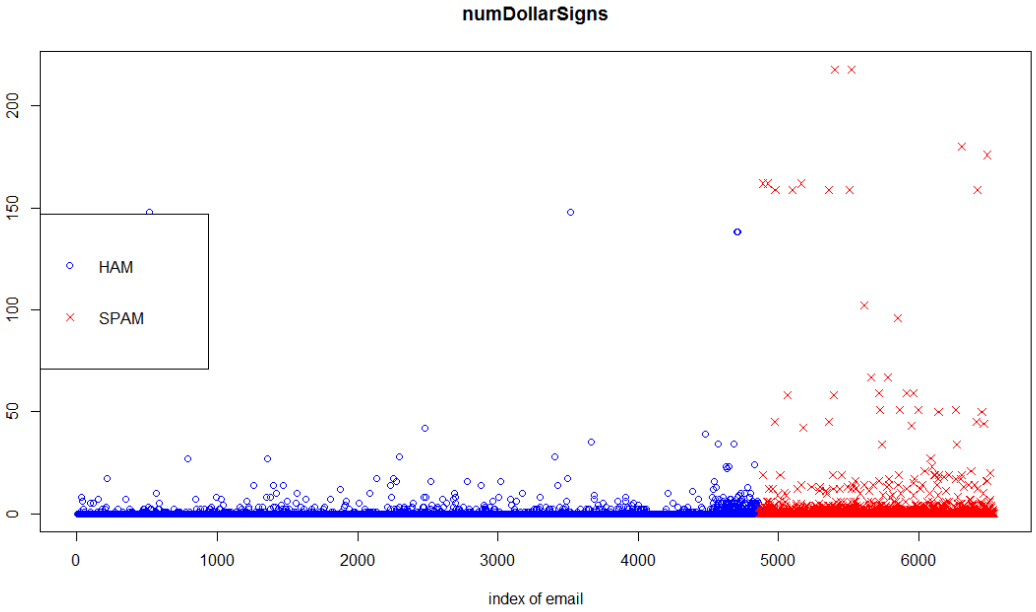
```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
          0.03120499                   0.00000000
```

From the scatterplot, it indicates that no SPAM emails were digitally signed.

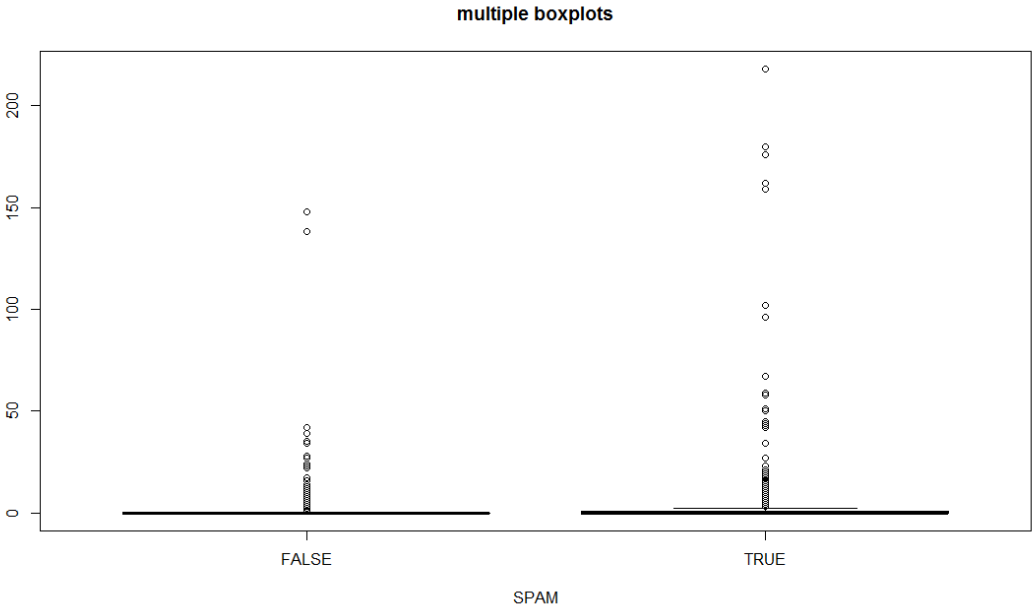
From the table of ratio, the ratio of TRUE in HAM is 0.031, however, the ratio of TRUE in SPAM is 0. It indicates that the variable can not be used to detect the SPAM email.

19. The variable "numDollarSigns": the number of dollar signs in the body of the message.

Method1: scatterplot



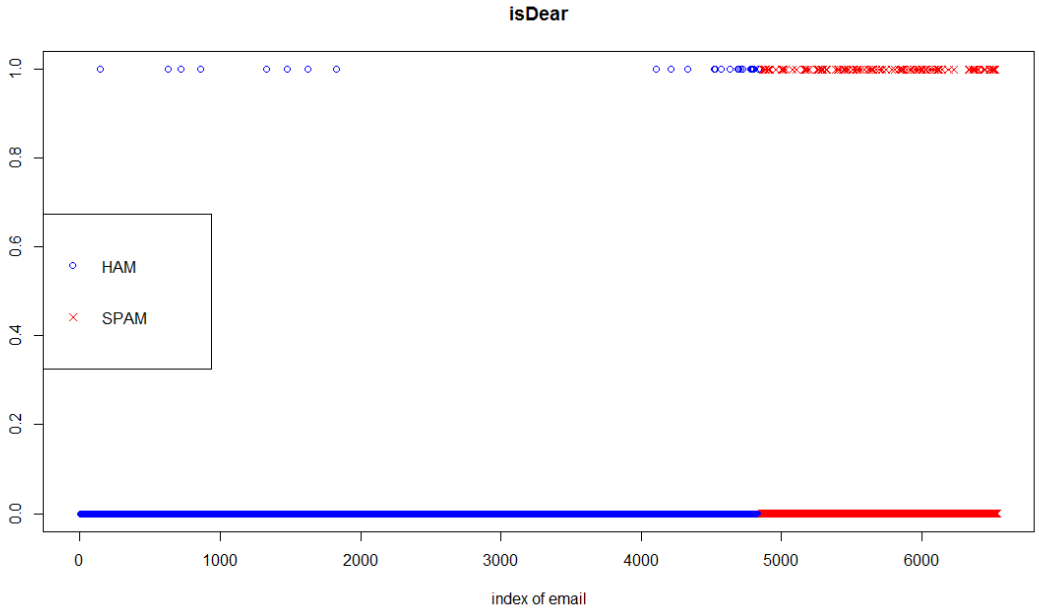
Method2: boxplot



From the scatterplot and boxplot, it indicates that if the number of dollar signs is larger than 50, it is more likely to be a SPAM email.

20. The variable "isDear": whether the message body contains a form of the introduction Dear **...**

Method1: scatterplot



Method2: ratio of TRUE value in HAM and SPAM

The ratio of TRUE value in HAM and the ratio of TRUE value in SPAM is

```
ratio of TRUE in HAM     The ratio of TRUE in SPAM
         0.005345395                   0.067978533
```

From the scatterplot, it indicates that when the message body contains a form of the introduction Dear, it is more likely to be a SPAM email.

From the table of ratio, the ratio of TRUE in HAM is 0.00534, however, the ratio of TRUE in SPAM is 0.068, which indicates that when the variable is true, we tend to believe it is a SPAM email.

- Summary

Among the 20 variables except the first variable.

The variables which can detect the SPAM emails well **on some special conditions** are below.

Variable "replyUnderline"

Variable "priority"

Variable "numRecipients"

Variable "percentCapitals"

Variable "percentSubjectBlanks"

Variable "isYelling"

Variable "isDear"

The variables which can detect the SPAM emails **on some special conditions** but not accurate are

Variable "subjectExclamationCount"

Variable "numAttachments"

The others variables can not detect the SPAM emails.

Code

```
>load("E:\\U course\\TrainingMessages.rda")
> trm=trainMessages

# 1
>isSpam<-function(i){
   # whether mail is Spam (TRUE) or Ham (FALSE)
   grepl("[S|s]pam",names(trm)[i])
}



# 2
>isRe<-function(i){
   # if the string Re: appears as the first word in the subject of the message
   if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

   j=grep("^Subject$", names(trm[[i]]$header))
   num=unlist(gregexpr("^[R][e]",trm[[i]]$header[j]))
   num[1]>0
}



# 3
>numLinesInBody<-function(i){
   # a count of the number of lines in the body of the email message
      length(trm[[i]]$body[which(trm[[i]]$body!="")])
}



# 4
>bodyCharacterCount<-function(i){
   # the number of characters in the body of the email message
     totalnum=unlist(gregexpr("[A-Z|a-z]",trm[[i]]$body))
     total=length(totalnum[totalnum>0])
     return(total)
}



# 5
>replyUnderline<-function(i){
   # whether the Reply-To field in the header has an underline and numbers/letters
   message=trm[[i]]$header[grepl("^[R|r]eply-[T|t]o$", names(trm[[i]]$header))]
```

```
   if(sum(grep("^[R|r]eply-[T|t]o$", names(trm[[i]]$header)))==0) return(0/0)

   else
      grepl("_[A-Z|a-z|0-9]", message)
}


# 6
>subjectExclamationCount<- function(i){
   # a count of the number of exclamation marks (!) in the subject of the message
   if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

   else
      j=grep("^Subject$", names(trm[[i]]$header))
      num=unlist(gregexpr("[!]",trm[[i]]$header[j]))
      num=length(num[num>0])
      return(num)
}


# 7
>subjectQuestCount<-function(i){
   # the number of question marks in the subject
   if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

   else
      j=grep("Subject", names(trm[[i]]$header))
      num=unlist(gregexpr("[?]",trm[[i]]$header[j]))
      num=length(num[num>0])
      return(num)
}


# 8
>numAttachments<-function(i){
   # the number of attachments in the message.
   if(sum(grep("attachments", names(trm[[i]])))==0) return(0/0)

   else
   length(names(trm[[i]]$attachments))
}


# 9
```

```
>priority<-function(i){
   # whether the message's header had an X-Priority or X-Msmail-Priority that was set to high
   if(sum(grep("X-Priority|X-Msmail-Priority", names(trm[[i]]$header)))==0) return(0/0)

   else
      k=grep("X-Msmail-Priority", names(trm[[i]]$header))
      j=grep("X-Priority", names(trm[[i]]$header))
      if(sum(j)!=0 & sum(k)!=0)
      {
         return(grepl("1|2",trm[[i]]$header[j])| grepl("High",trm[[i]]$header[k]))
      }

      else
         l=grep("X-Priority|X-Msmail-Priority", names(trm[[i]]$header))
         return(grepl("1|2|High",trm[[i]]$header[l]))
}


# 10
>numRecipients<-function(i){
   # the number of recipients in the To, Cc fields
   if(sum(grep("To|Cc", names(trm[[i]]$header)))==0) return(0/0)

   else
      k=grep("To", names(trm[[i]]$header))
      j=grep("Cc", names(trm[[i]]$header))

      if(sum(j)!=0 & sum(k)!=0)
      {
            return(length(gregexpr("@",trm[[i]]$header[j])[[1]]>0)+length(gregexpr("@",trm[[i]]$h
            eader[k])[[1]]>0))
      }

      else
         l=grep("To|Cc", names(trm[[i]]$header))
         return(length(gregexpr("@",trm[[i]]$header[l])[[1]]>0))
}


# 11
>percentCapitals<-function(i){
   # the percentage of the characters in the body of the email that are upper case
   totalnum=unlist(gregexpr("[A-Z|a-z]",trm[[i]]$body))
   total=length(totalnum[totalnum>0])
```

```
   uppernum=unlist(gregexpr("[A-Z]",trm[[i]]$body))
   upper=length(uppernum[uppernum>0])
   return(upper/total)
}


# 12
>isInReplyTo<-function(i){
   # whether the header of the message has an In-Reply-To field.
   sum(grep("^In-Reply-To$", names(trm[[i]]$header)))!=0
}


# 13
>subjectPunctuationCheck<-function(i){
   # whether the subject has punctuation or digits surrounded by characters
   if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

   else
      j=grep("Subject", names(trm[[i]]$header))
      num=gregexpr("[a-z|A-Z][0-9|[:punct:]][a-z|A-Z]",trm[[i]]$header[j])[[1]]
      length(num[num>0])>0
}


# 14
>hourSent<-function(i){
   # the hour in the day the mail was sent (0 -- 23)
   if(sum(grep("^Date$", names(trm[[i]]$header)))==0) return(0/0)

   else
      j=grep("^Date$", names(trm[[i]]$header))
      num=gregexpr(":", trm[[i]]$header[j])[[1]][1]
      substr(trm[[i]]$header[j],num-2,num-1)
}


# 15
>subjectSpamWords<-function(i){
# whether the subject contains one of the following phrases: viagra, pounds, free, weight, guarantee, millions, dollars, credit, risk, prescription, generic, drug, money back, credit card.
   if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

   else
```

```
    j=grep("Subject", names(trm[[i]]$header))
    sum(grep("viagra|pounds|free|weight|guarantee|millions|dollars|credit|risk|prescription
            |generic|drug|money back|credit card",trm[[i]]$header[j]))>0
}
# 16
>percentSubjectBlanks<-function(i){
  # the percentage of blanks in the subject
  if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

  else
    j=grep("Subject", names(trm[[i]]$header))
    totalnum=unlist(gregexpr(".", trm[[i]]$header[j]))
    total=length(totalnum[totalnum>0])
    blanknum=unlist(gregexpr(" ", trm[[i]]$header[j]))
    blank=length(blanknum[blanknum>0])
    return(blank/total)
}


# 17
>isYelling<-function(i){
  # whether the Subject of the mail is in capital letters
  if(sum(grep("^Subject$", names(trm[[i]]$header)))==0) return(0/0)

  else
    j=grep("^Subject$", names(trm[[i]]$header))
    sum(grep("[a-z]",trm[[i]]$header[j]))==0
}


# 18
>isPGPsigned<-function(i){
  # indicates whether the mail was digitally signed (e.g. using PGP or GPG)
  if(sum(grep("^[C|c]ontent-[T|t]ype$", names(trm[[i]]$header)))==0) return(0/0)

  else
    j=grep("^[C|c]ontent-[T|t]ype$", names(trm[[i]]$header))
    num=unlist(gregexpr("[S|s]igned", trm[[i]]$header[j]))
    length(num[num>0])>0
}


# 19
>numDollarSigns<-function(i){
```

```
   # the number of dollar signs in the body of the message
   num=unlist(gregexpr("[$]", trm[[i]]$body))
   length(num[num>0])
}
# 20
>isDear<-function(i){
   # whether the message body contains a form of the introduction Dear ...
   num=unlist(gregexpr("^Dear", trm[[i]]$body))
   length(num[num>0])>0
}
```