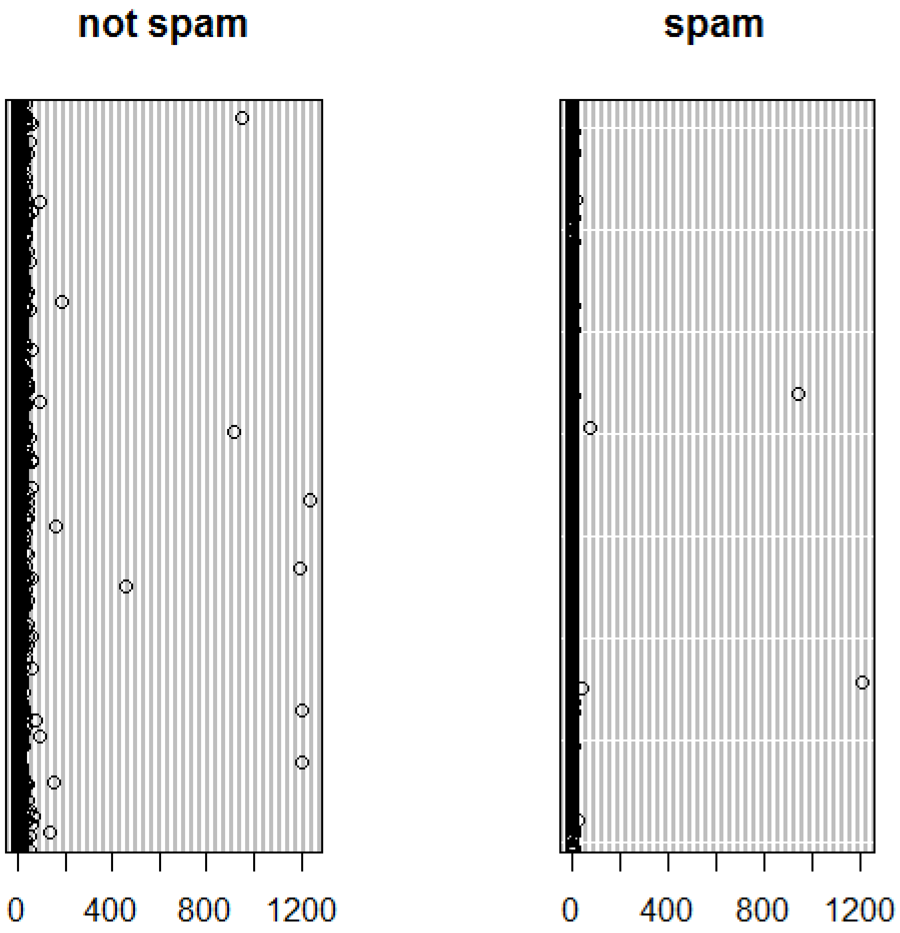
Home Work 1 Write up

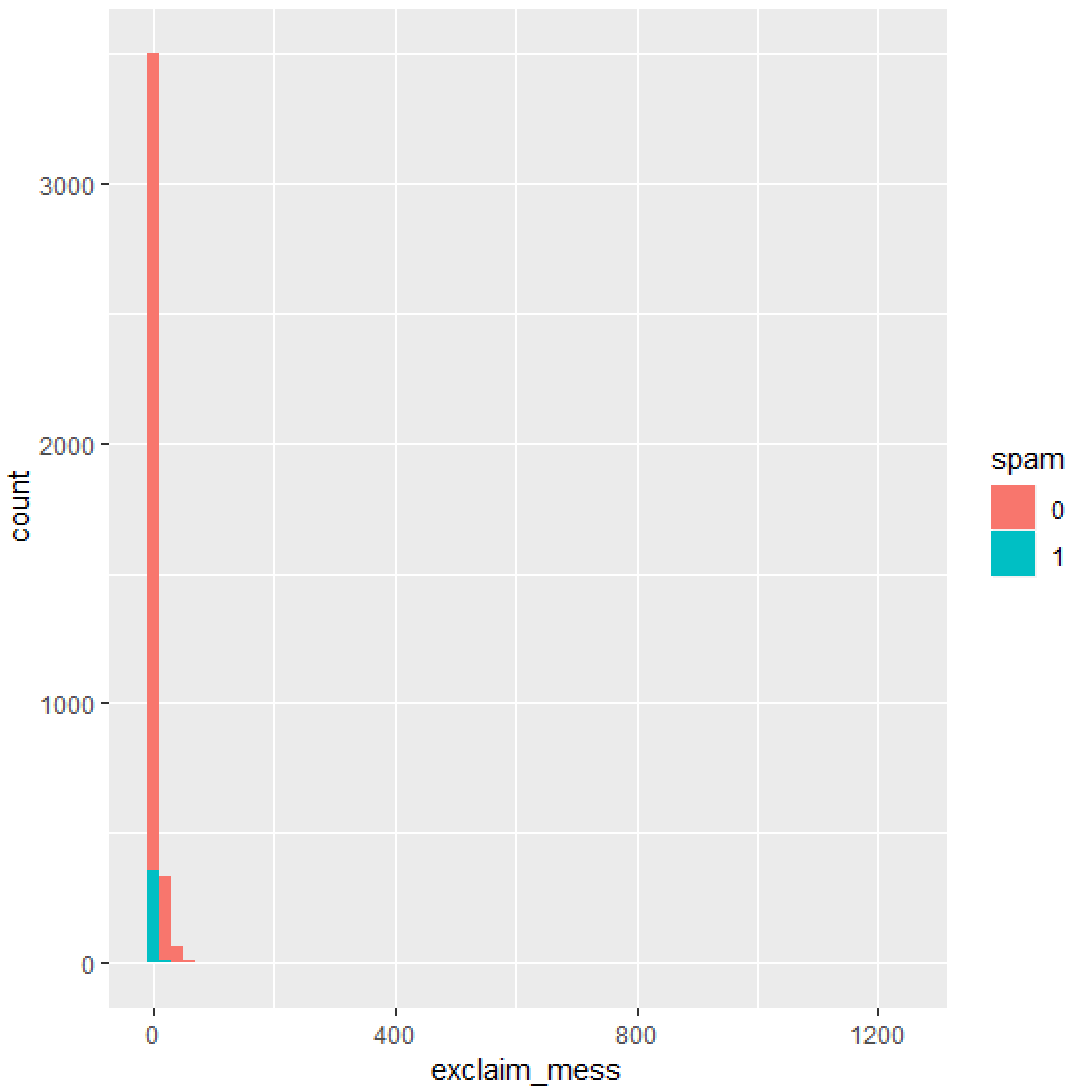
Question1

Part 1

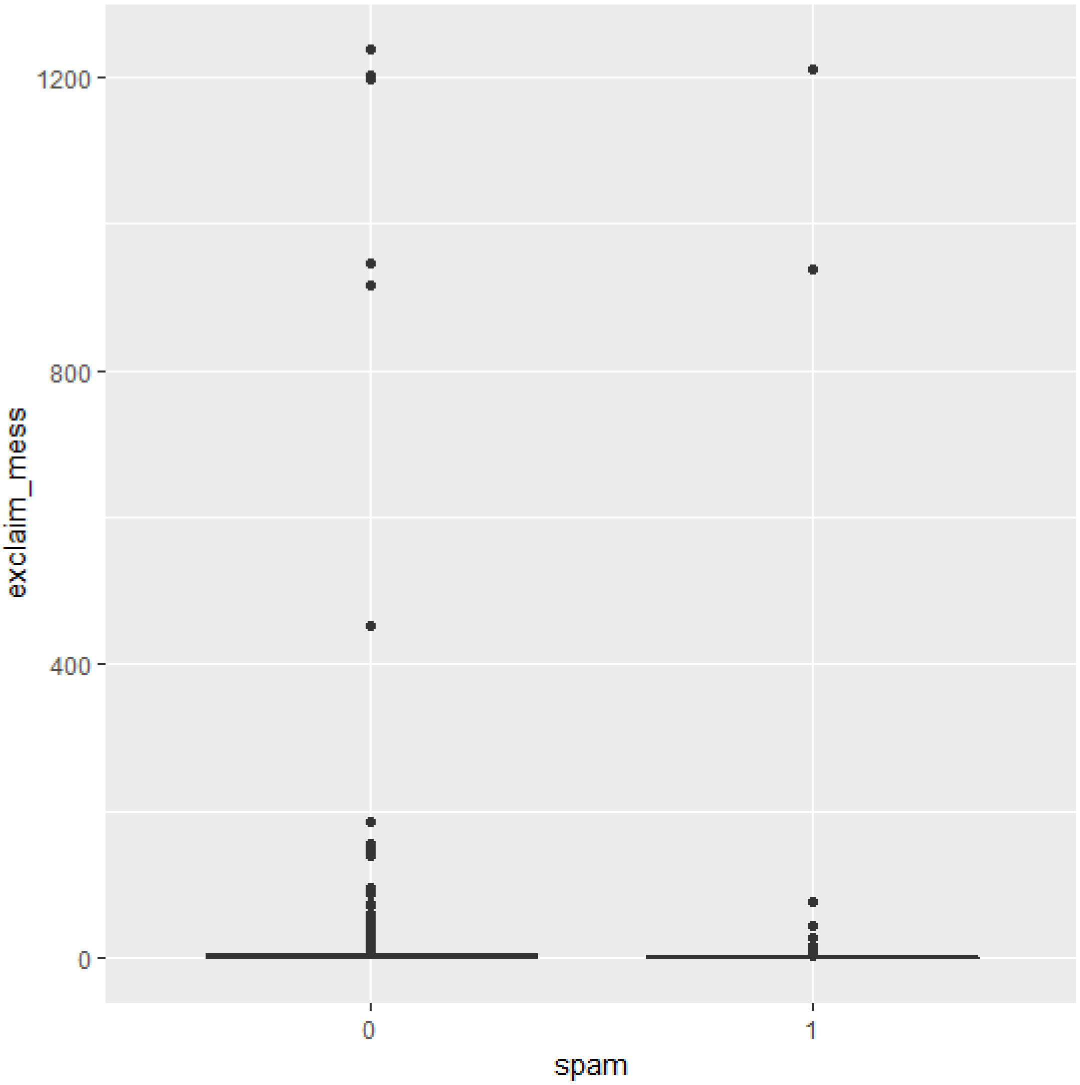
Dot chart of spam vs exclaim_mess



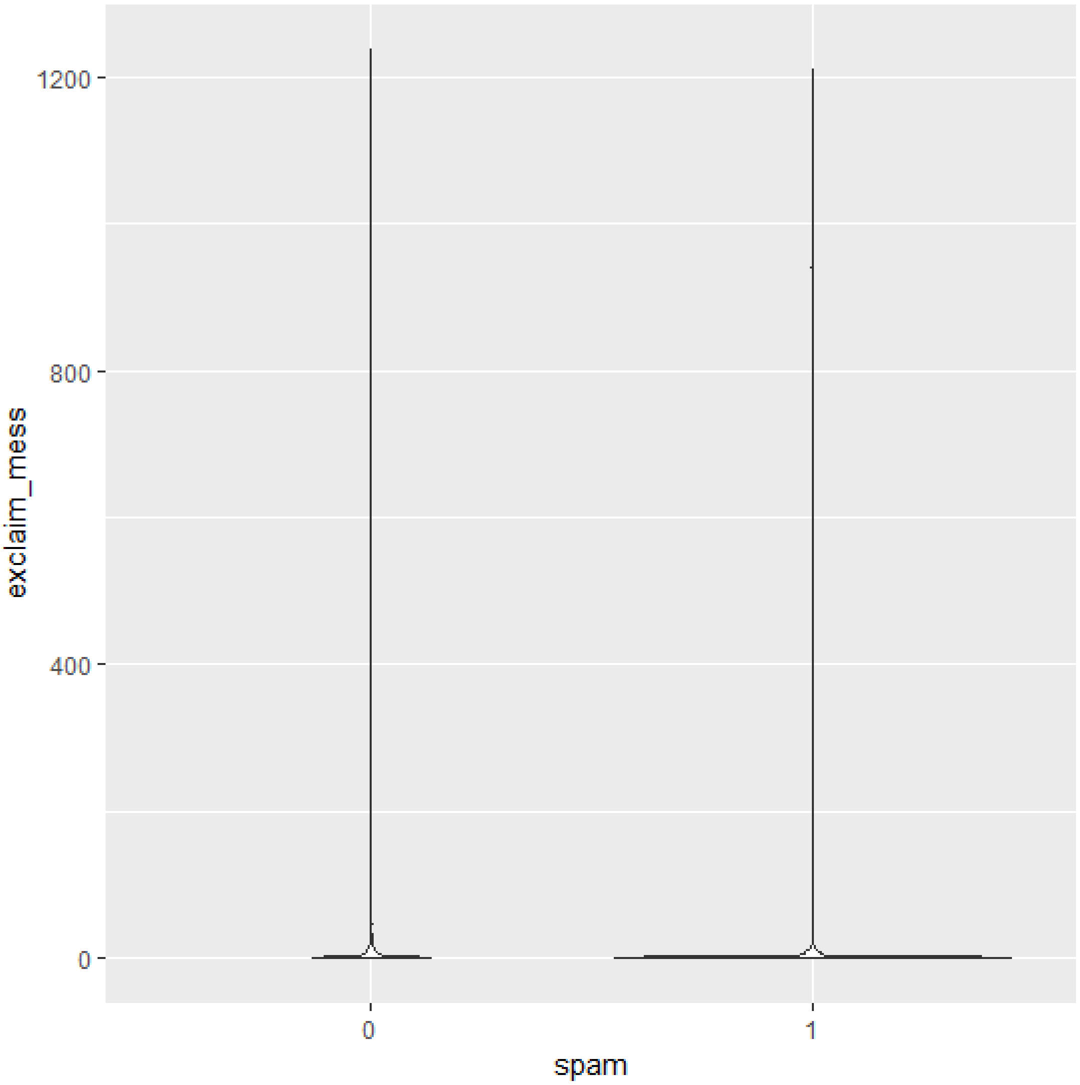
Histogram of spam vs exclaim_mess



Box plot of spam vs exclaim_mess



Violin plot of spam vs exclaim_mess



Part 2: Which of the above plots are useful for describing the relationship be-tween these two variables? What do those plots convey? Why are the other plots not as useful?

Violin plot would be the most useful for describing the relationship between two variables, because it shows how exclamation points are proportionally concentrated between spam and not spam email (concentration of datapoint around 0). As most emails have zero exclamation points it is hard to visualize that ratio between spam and not spam in dot chart and box plot. Histogram is not as useful since there are more not spam emails than spam emails, spam emails with exclamation points barely register on the above histogram. Histogram would be a better choice than dot chart and box plot, but violin plot is better to show the relationship between exclaim_mess and spam.

Part 3. Summarize the relationship between spam and exclaim mess in a couple of sentences.

Looking at the above graphs, it would be wise to not use the exclaim mess variable to determine spam because there are more emails with exclamation point which are not spam. Both spam and not spam emails have 0 exclamation points as median (as we can see from box plot). From the box plot and dot chart we can see there are only two spam emails with excessive exclamation points, but many not spam emails with excessive exclamation points. Hence, the above plots suggest emails with exclamation points have higher probability of being not spam.

Question 2

Sometimes it is useful to recode variables. spam vs exclaim mess is an example.

Part 1. Recode exclaim mess into four values: 0, 1, 2, >= 3. Call this new variable exclaim_mess.recode. What is the type of this new variable?

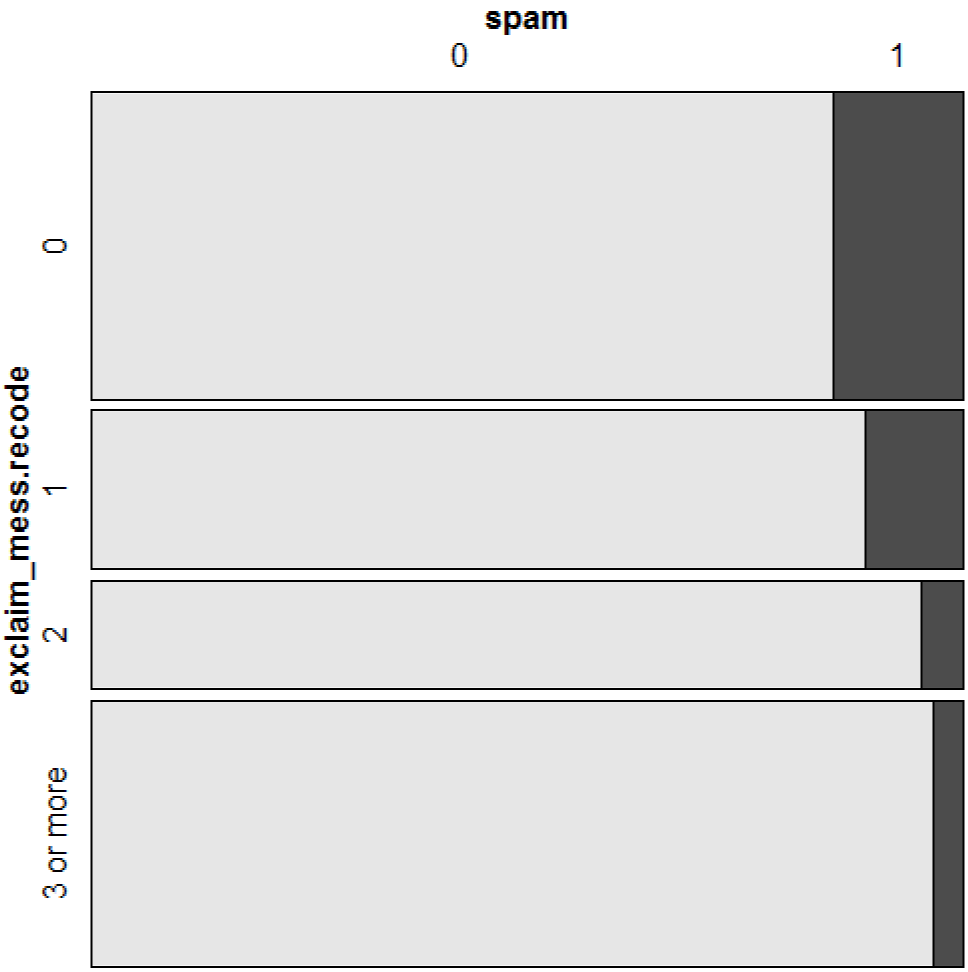
The type is, **categorical** for the new variable, exclaim_mess.recode

Part 2. Produce a table and a mosaic plot of spam vs exclaim mess.recode. What do they reveal?

The table:

	0	1	2	3 or more
0 (Not Spam)	1219	650	482	1203
1 (Spam)	216	83	25	43

Mosaic plot:



The table and mosaic plot confirm our assumption from previous plots that if there are 1 or more exclamation points in an email, it is more likely to be not spam. The table and mosaic plot show most spam with exclamation points have only 1 or 2 in an email.

Part 3. Summarize the relationship between spam and exclaim mess.recode in a couple of sentences.

Out of 367 spam emails 58.85% (216) had no exclamation point. Out of those who had exclamation points and were spam, 22.61% (83) had one, 6.81(25) had two, and 11.72% (43) had 3 or more. Compared to 3554 not spam emails, it had 34.30%, 18.29%, 13.56%, and 33.85% respectively in the same category. Based on the data, we can assume that spam emails are less likely to have more than one exclamation point than not spam.

Part 4. Why is it reasonable to recode exclaim mess?

Most emails (1435 out of 3921) have no exclamation points, such that their median was 0 and looked concentrated on our initial plots. We want to find out how many exclamation points were there in how many emails, so we can compare them to spam and not spam emails and find out a relationship. It is reasonable to recode exclaim mess, so we can analyze how many emails with X values exist, X being the number of exclamation mark in an email.

Part 5. How would your summary on the relation between spam and exclaim mess change if you had recoded it into 5 values? 10 values? 3 values? Which regroup is most reasonable, and why?

This will depend on the size of increments we pick, example if we can pick <2, <4, >=6 or we can pick 1, 2, >=3, each will give different data, but the conclusion would not change unless we pick extremely large increments with smaller values. The more values we pick more precise we can get in our conclusion, provided we pick reasonable increments. For example, picking 5 values of 0, 1, 2, 3, and >3 would give

	0	1	2	3	More than 100
0 (Not Spam)	1219	650	482	1192	11
1 (Spam)	216	83	25	41	2

We would be more precise with our conclusion because we have more data. Recoding with 10 values would be even more precise, however if we recode with 3 values we would be less precise.

Regrouping with 5 values would be most reasonable since we can come up with a precise conclusion without adding too many values. 3 values would be too little, and 10 values would be excessive in exchange for a little bit more precision.

Question 3.

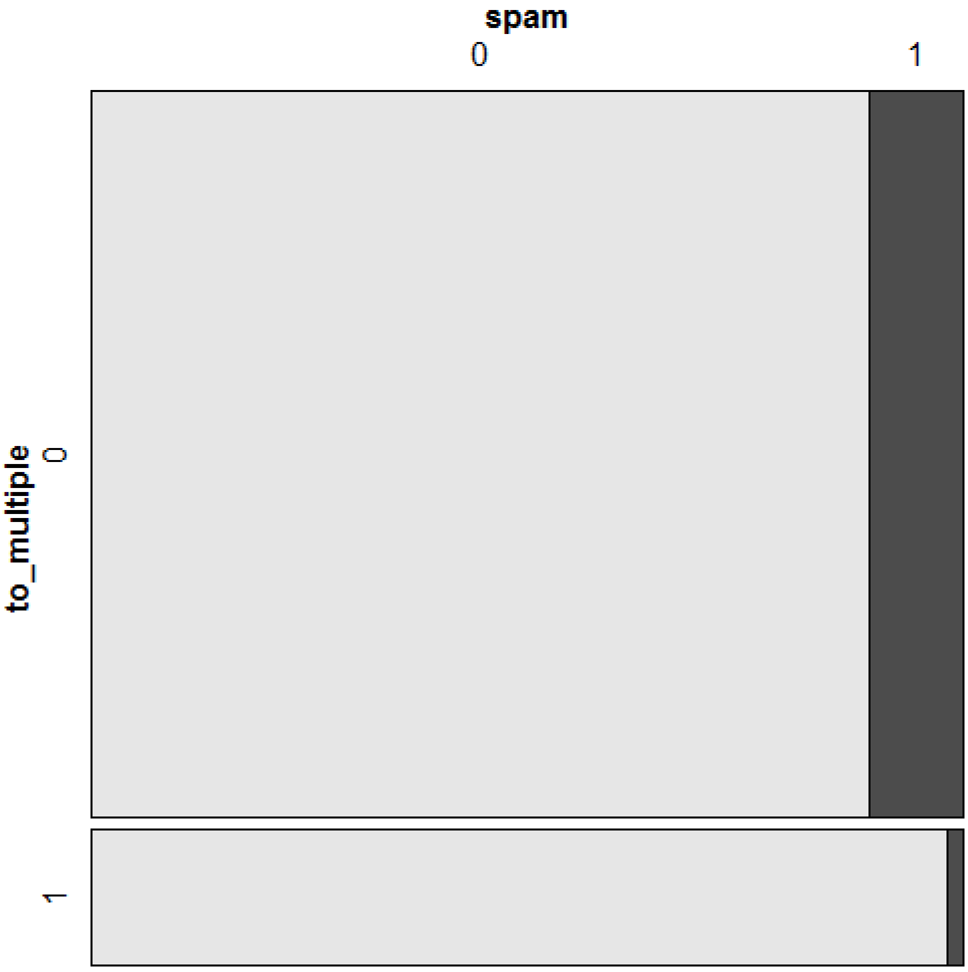
Part 1. Run a descriptive analysis for spam vs X for each of the 20 variable X in the dataset. For each analysis, include ONE plot and/or ONE table that is most informative, and write a short sentence summarizing the relationship between spam and X.

Spam vs to_multiple

Table:

	0 (Single recipient)	1 (multiple recipient)
0 (Not spam)	2946	608
1 (Spam)	355	12

Mosaic plot of spam vs to_multiple



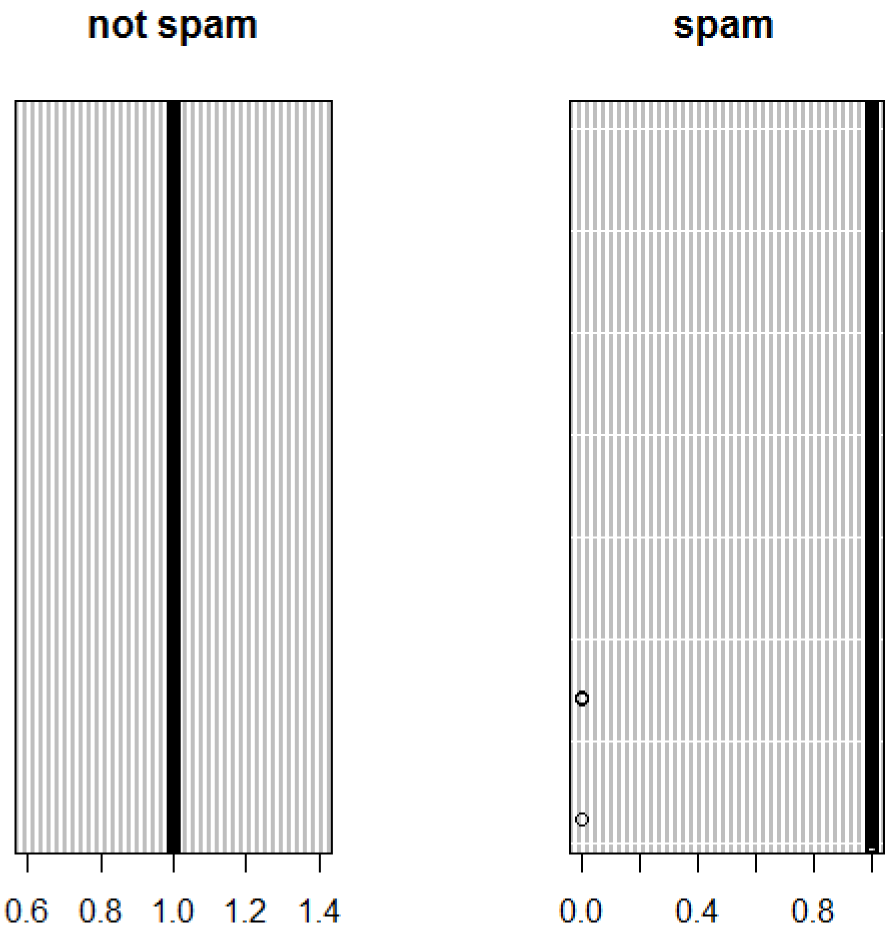
We can conclude from the table and graph that if emails were sent to multiple recipient they are more likely to be not spam. So, to_multiple is not a good indicator of spam because only 3.27% (12) out of 367 spam emails had multiple recipient compared to 17.11% (608) of not spam emails.

Spam vs from

Table:

	0 (name not listed)	1 (name listed)
0 (Not spam)	0	3554
1 (Spam)	3	364

Box plot



We can see all the not spam emails were listed from someone, whereas there were 3 spam emails that were not listed. So, if there is an email which is not listed under anyone's name it has a higher chance of being a spam.

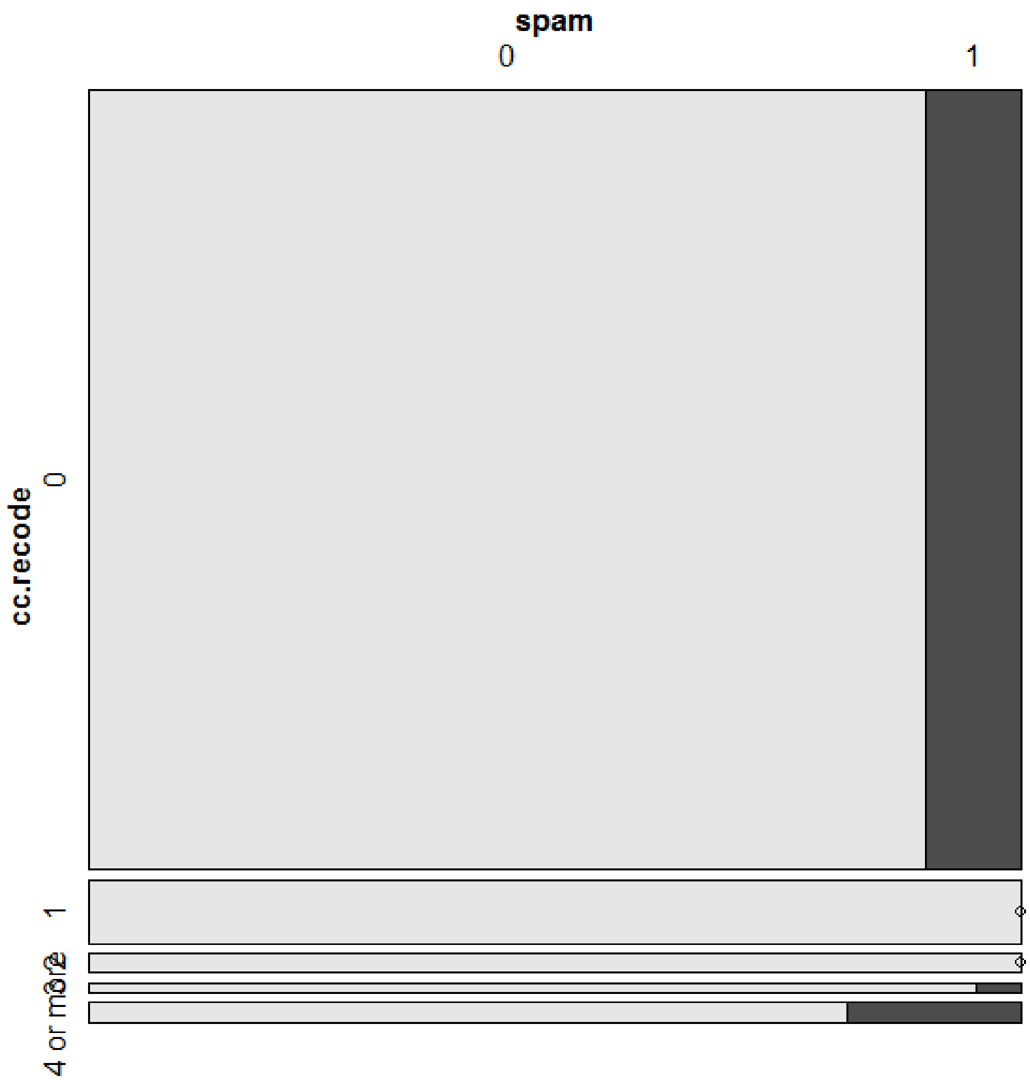
Spam vs cc

Table

Just like in the case of spam vs exclaim_mess in number two, we can analyze the data better if we recode this. If we recode into five values we get the following table

	0	1	2	3	4 or more
0 (Not Spam)	3087	278	80	39	70
1 (Spam)	349	0	0	2	16

Mosaic plot



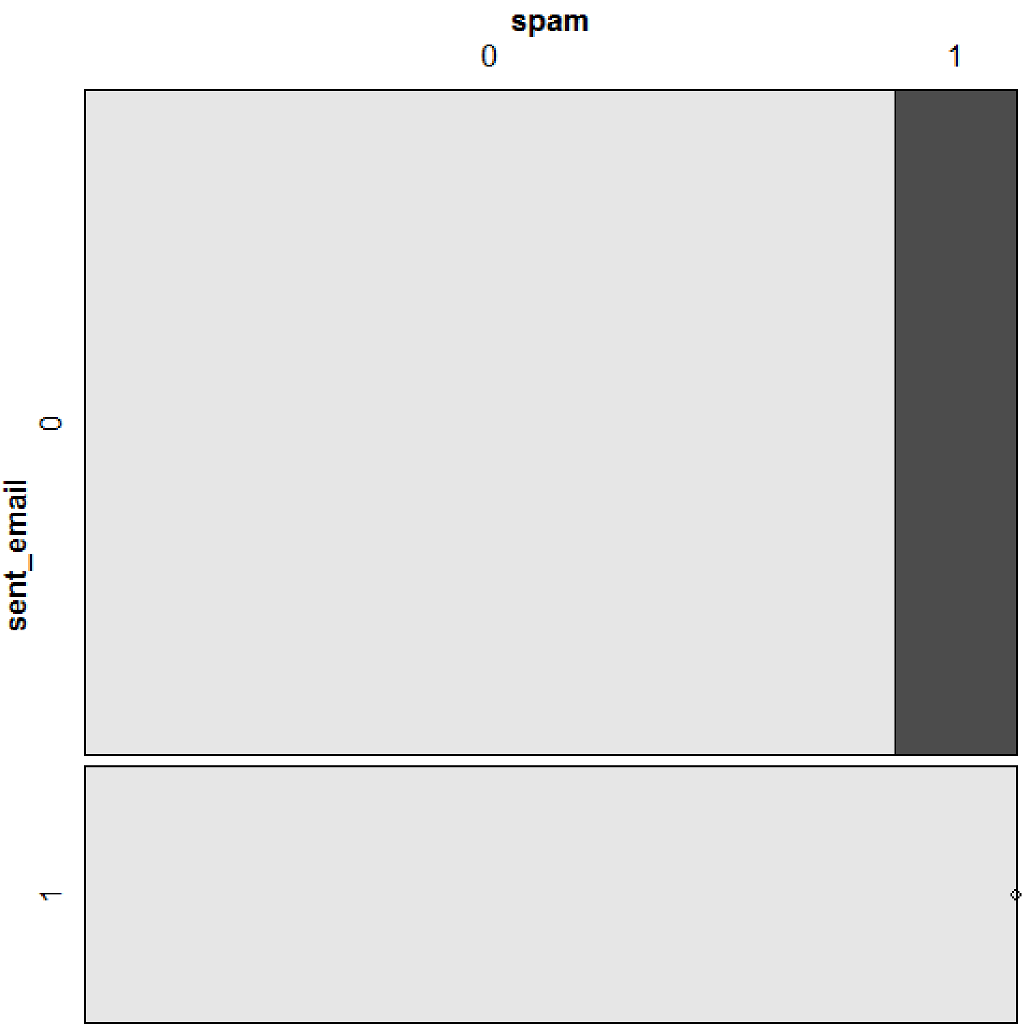
Most spam and not spam email did not have multiple recipents, but from the graph we can deduce that if there are 4 or more recipients in an email, it is slightly more likely to be spam.

Spam vs sent_email

Table

	0 (had not been contacted in the last 30 days)	1 (had been contacted in the last 30 days)
0 (Not spam)	2464	1090
1 (Spam)	367	0

Mosaic plot



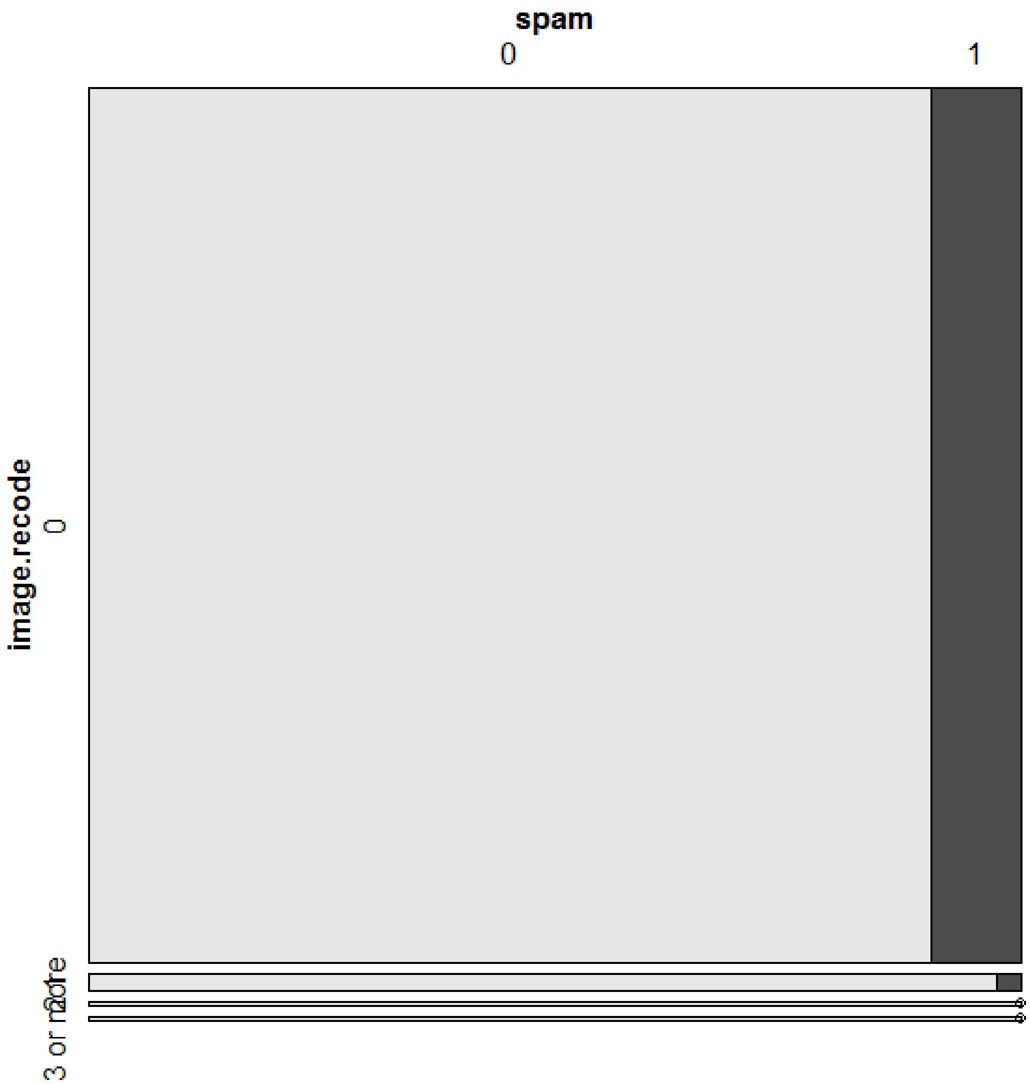
The table and the graph indicate that if the email is from some one who has been sent an email in the last 30 days, it is most likely not a spam because 0 cases exist.

Spam vs image

Table

	0	1	2	3 or more
0 (Not Spam)	3446	74	17	17
1 (Spam)	365	2	0	0

Mosaic Plot



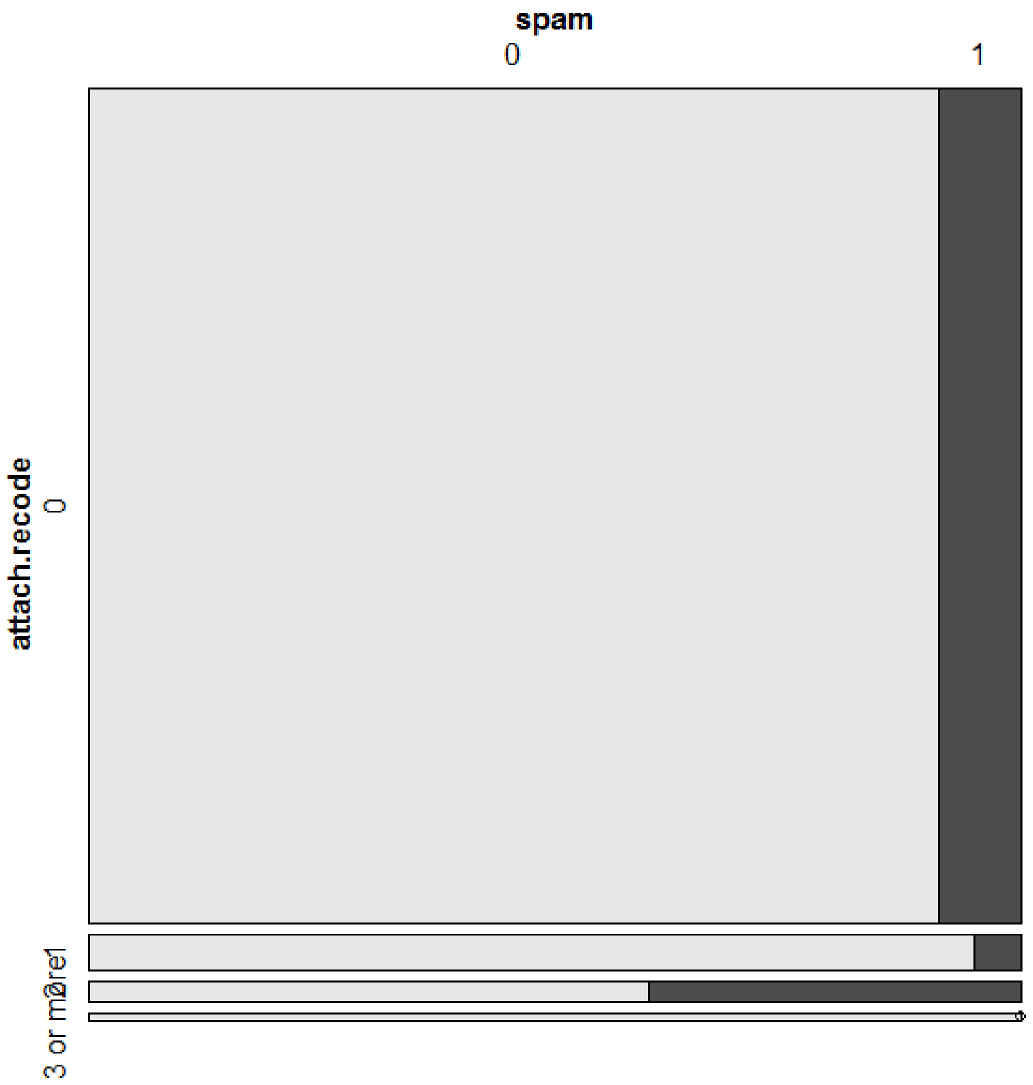
Most email had no images attached. However, if images are attached it is more likely that the email is not a spam

Spam vs attach

Table

	0	1	2	3 or more
0 (Not Spam)	3315	150	54	35
1 (Spam)	323	8	36	0

Mosaic plot



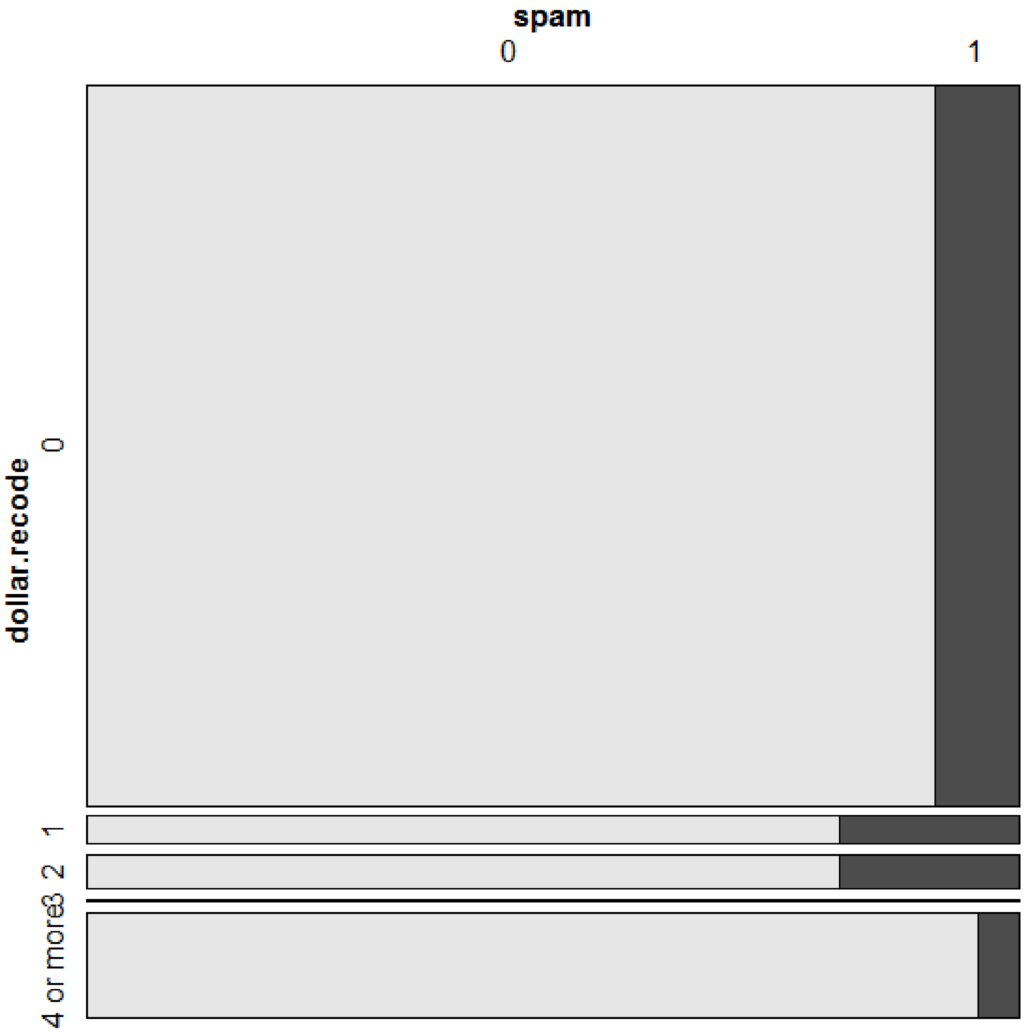
Most emails had no attachments. If there are 3 or more attachments, then it is likely that the email is not a spam. However, if there are exactly 2 attachments it is more likely to be a spam because there is a probability of 40% (36/90).

Spam vs dollar

Table:

	0	1	2	3	4 or more
0 (Not Spam)	2886	97	122	4	445
1 (Spam)	289	23	29	6	20

Mosaic plot:

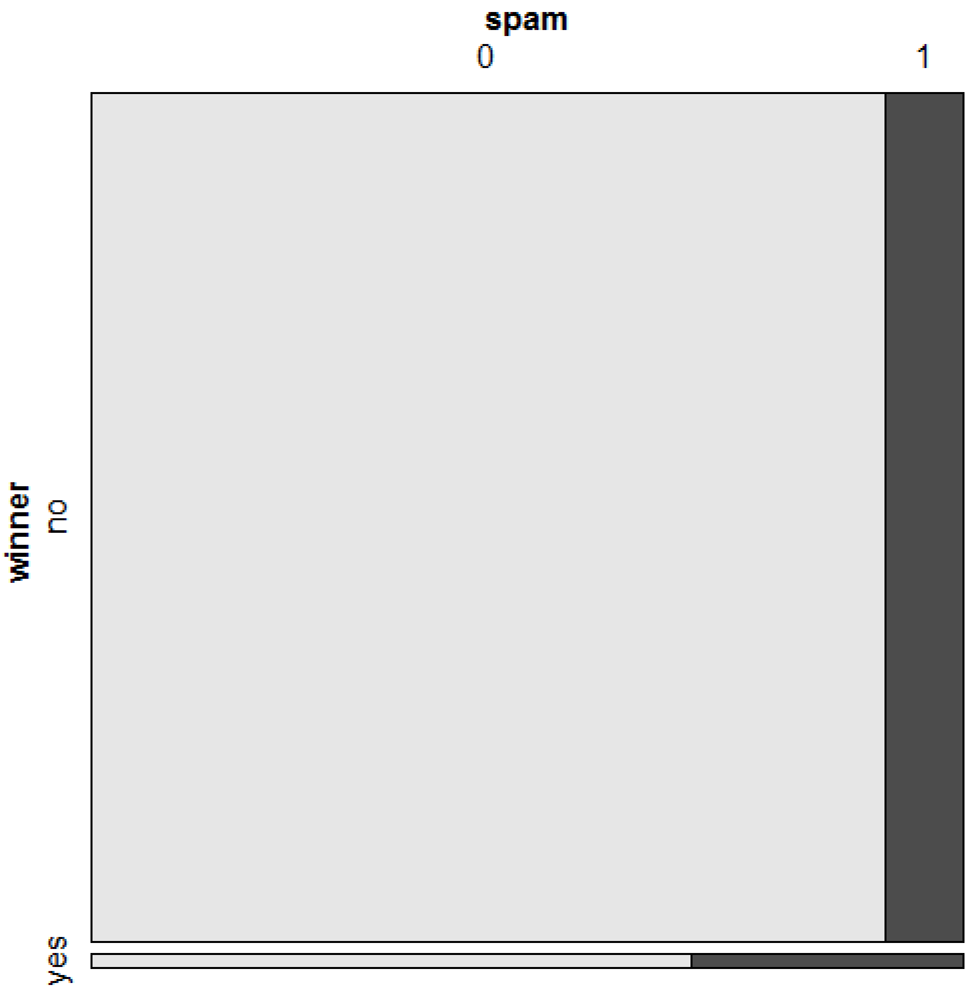


If we look at the dot chart for spam/not spam vs dollar the data would be scattered for both. However, from the mosaic plot we can deduce that if there is one- or two-dollar sign in the email it is more possible for it to be spam. However, if there are 4 or more-dollar signs in an email it is likely not spam.

Spam vs winner

	No(winner did not appear)	Yes(winner appeared)
0 (not spam)	3510	44
1 (spam)	347	20

Mosaic plot



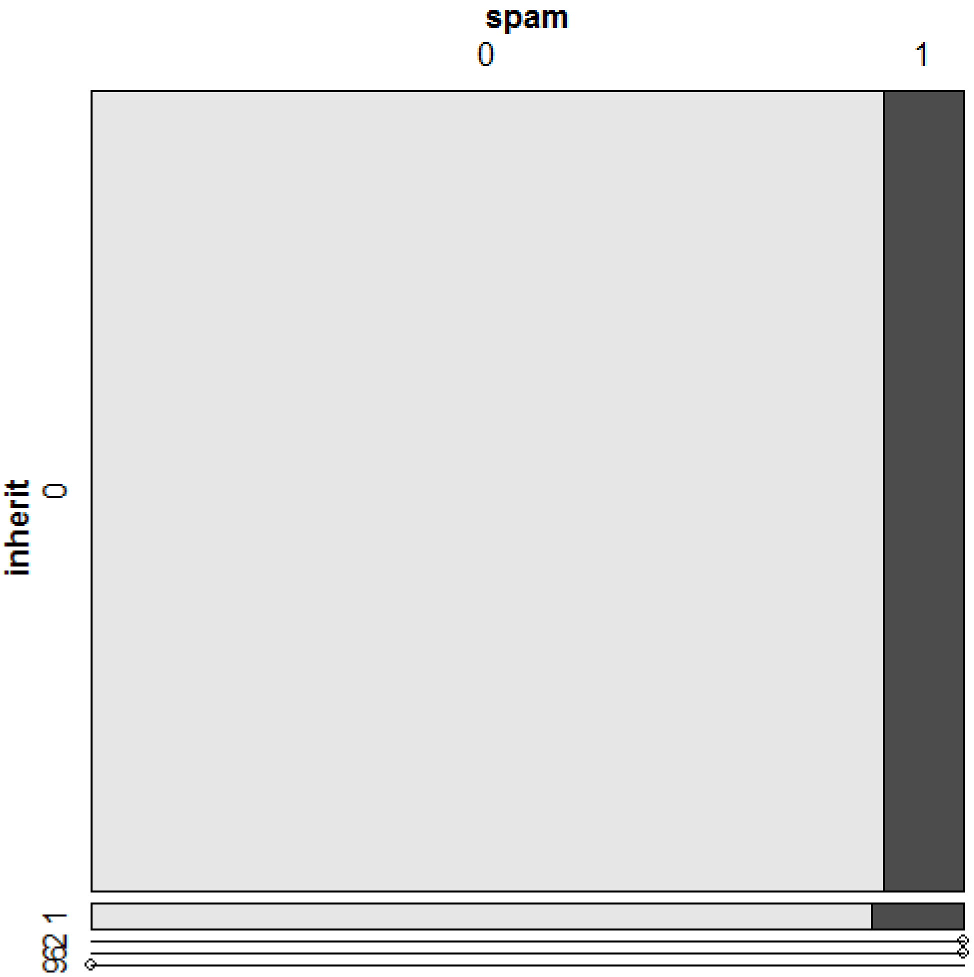
We can conclude from the mosaic plot that if the word winner appears in an email it is more likely to be spam.

Spam vs inherit

Table

	0	1	2	6	9
0 Spam	3440	109	3	2	0
1 Not spam	353	13	0	0	1

Mosaic plot



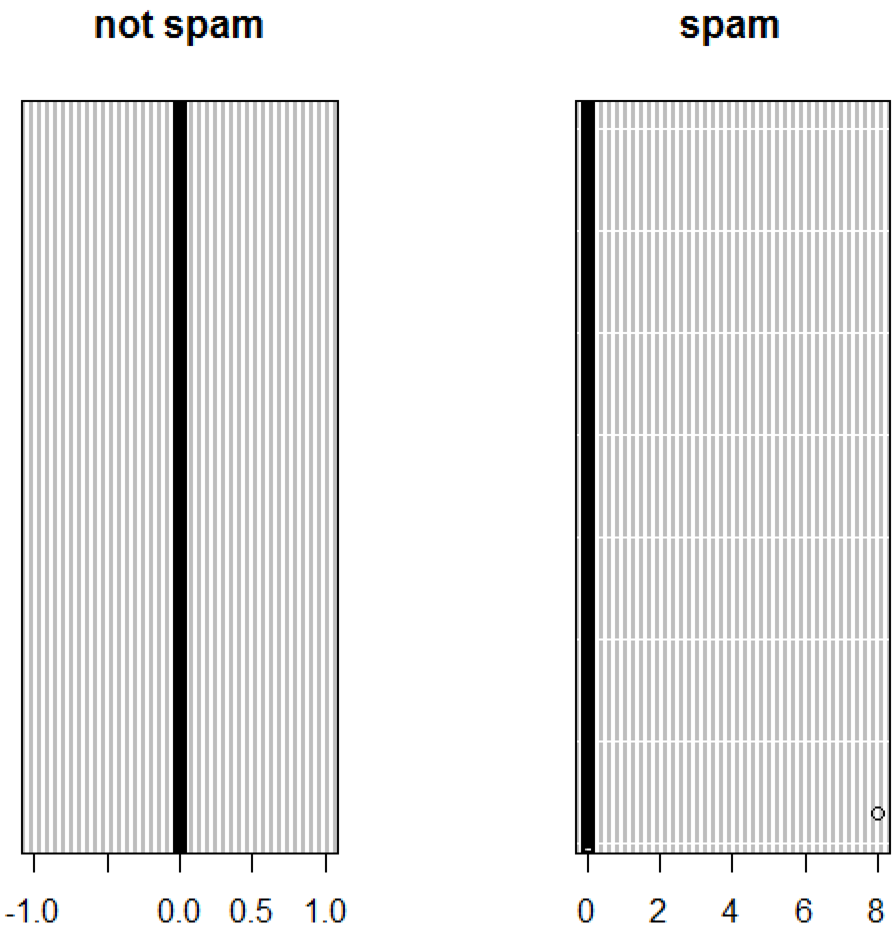
From the data we can only conclude that if the word inheritance appears on an email exactly once it has a slightly higher chance of being a spam.

Spam vs Viagra

Table

	0	8
0 (not spam)	3554	0
1 (spam)	366	1

Dot chart



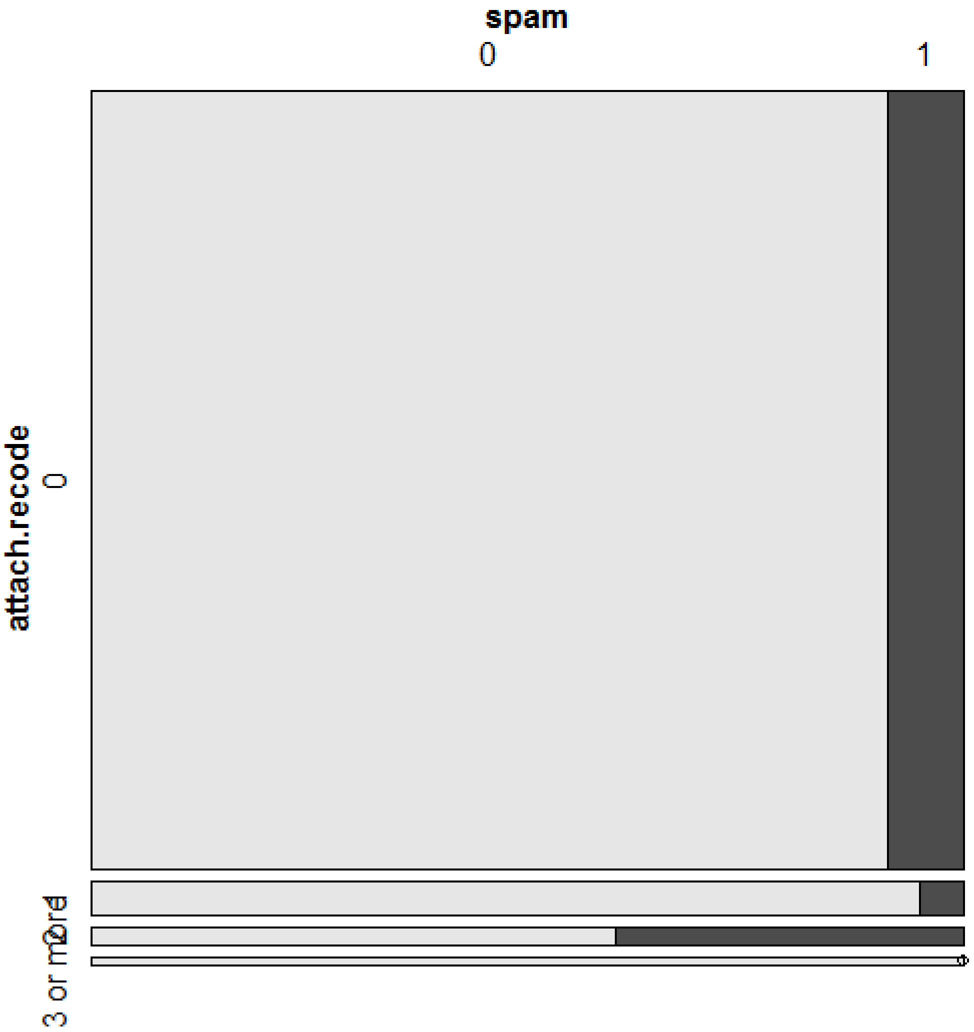
There is only one email which had the word Viagra which was repeated 8 times. It is a spam, but there is not enough data to come to any conclusions.

Spam vs password

Table

	0	1	2	3 or more
0	3446	20	37	51
1	363	2	2	0

Mosaic plot



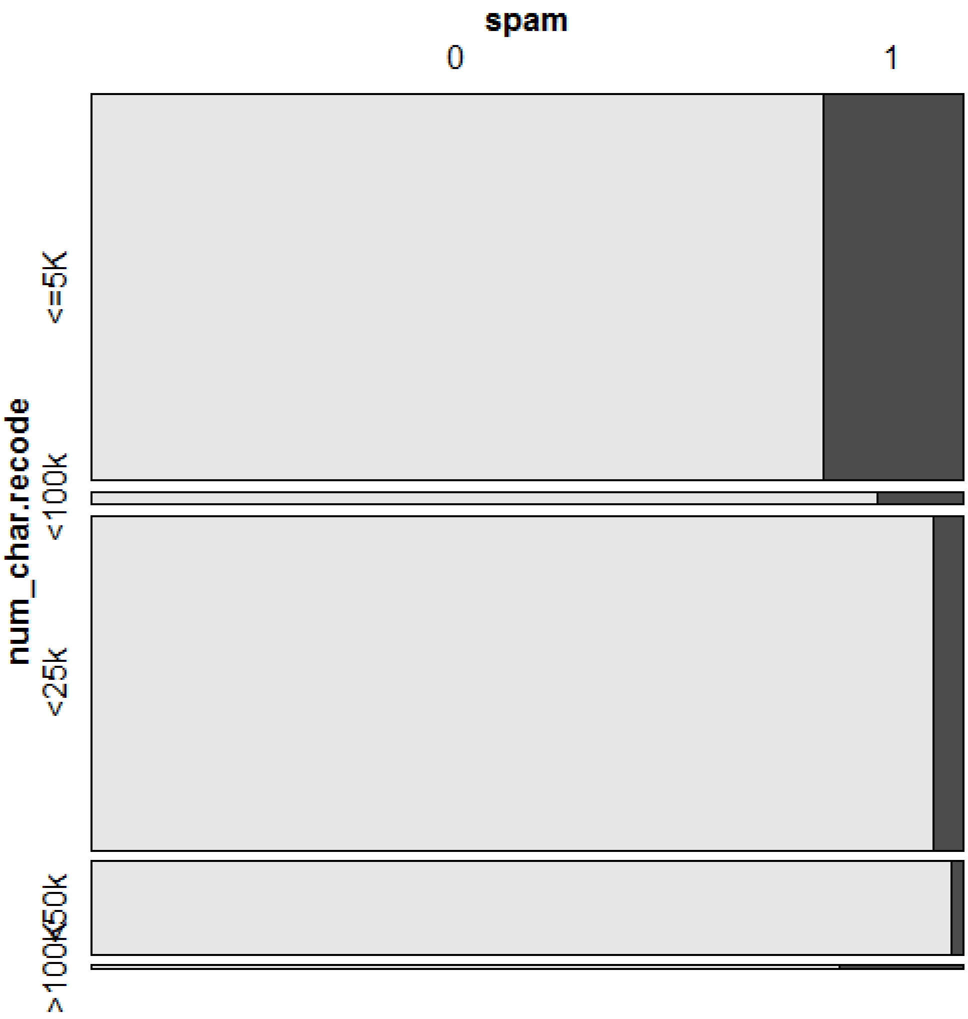
If the word password appears 1 and 3 or more times in an email it is most likely not spam. From the mosaic plot we would conclude that if it appears exactly 2 times it is most likely spam. However, if we look at the table there are only 2 emails where the word password appears 2 times, so there is not enough data to come to the second conclusion.

Spam vs num_char

Table

	Less than 5K	5k-25k	25k-50k	50k-100k	>100k
0	1534	1521	433	54	12
1	297	56	6	6	2

Mosaic plot



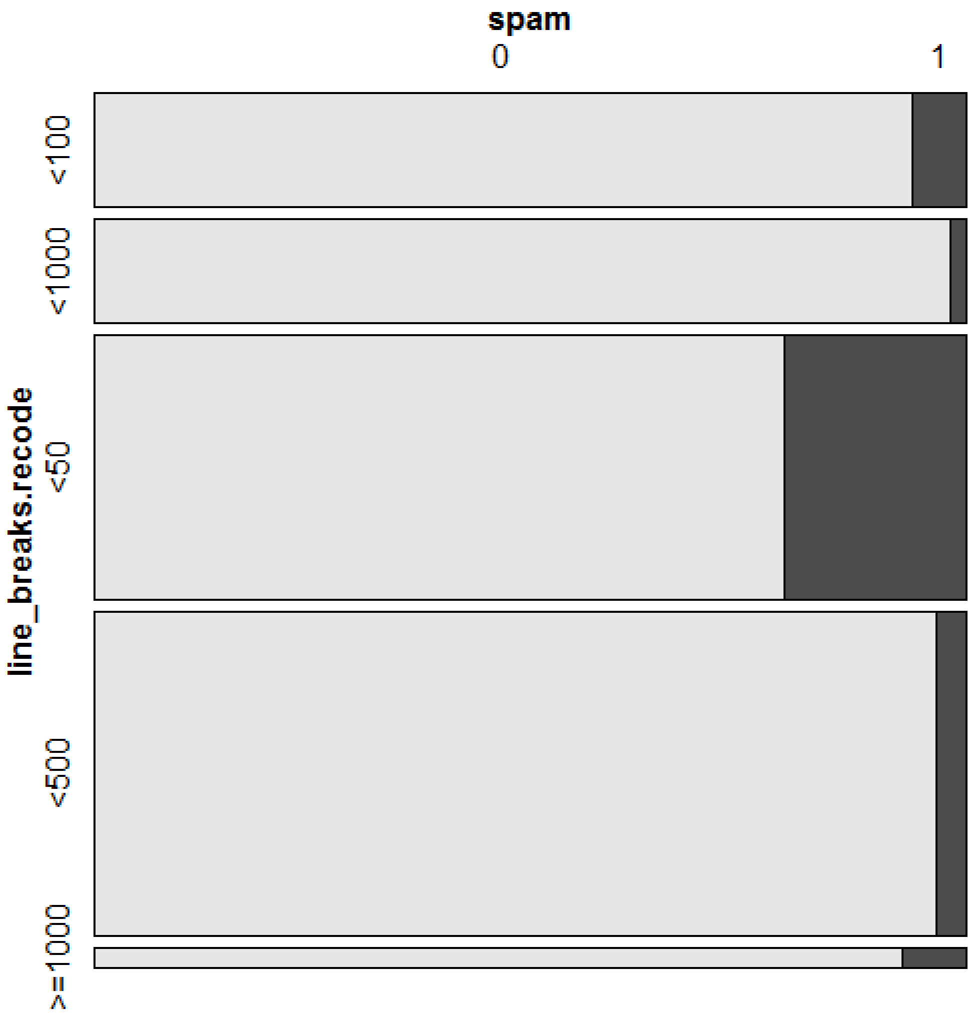
The data shows that if the number of characters in an email is higher than 5000 it is likely to be not spam. We can look more into tis variable on data less than 5000 on how it is distributed to come up with a more precise conclusion.

Spam vs line_breaks

Table:

	Less than 50	50-100	100-500	500-1000	>1000
0 (Not spam)	992	507	1483	485	87
1 (Spam)	261	34	56	9	7

Mosaic plot



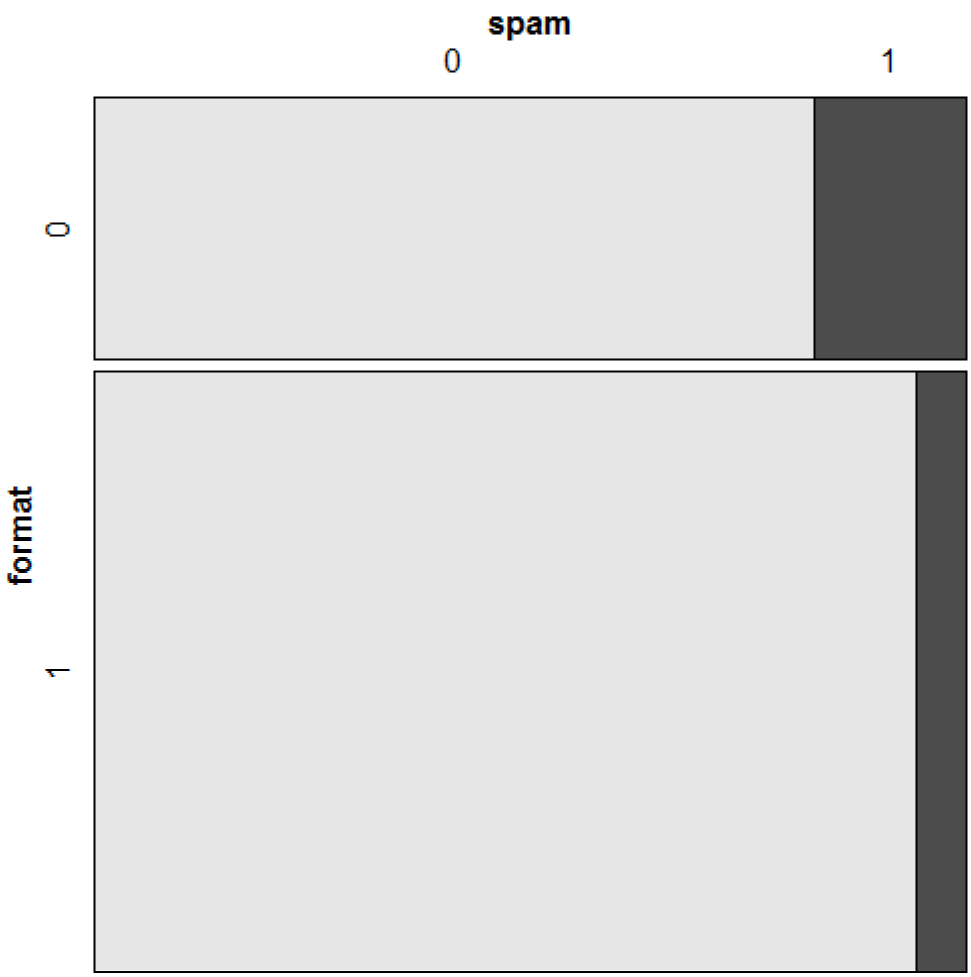
The greater the number of line breaks there are in an email, the more likely it is not a spam. Just like in the case of num_char we can look more into emails with lesser than 50-line breaks to come up with a precise number (lesser than 50) because more spam emails are concentrated within less than 50 line breaks category.

Spam vs format

Table:

	0(not written using html)	1(written using html)
0 (not spam)	986	2568
1 (spam)	209	158

Mosaic plot



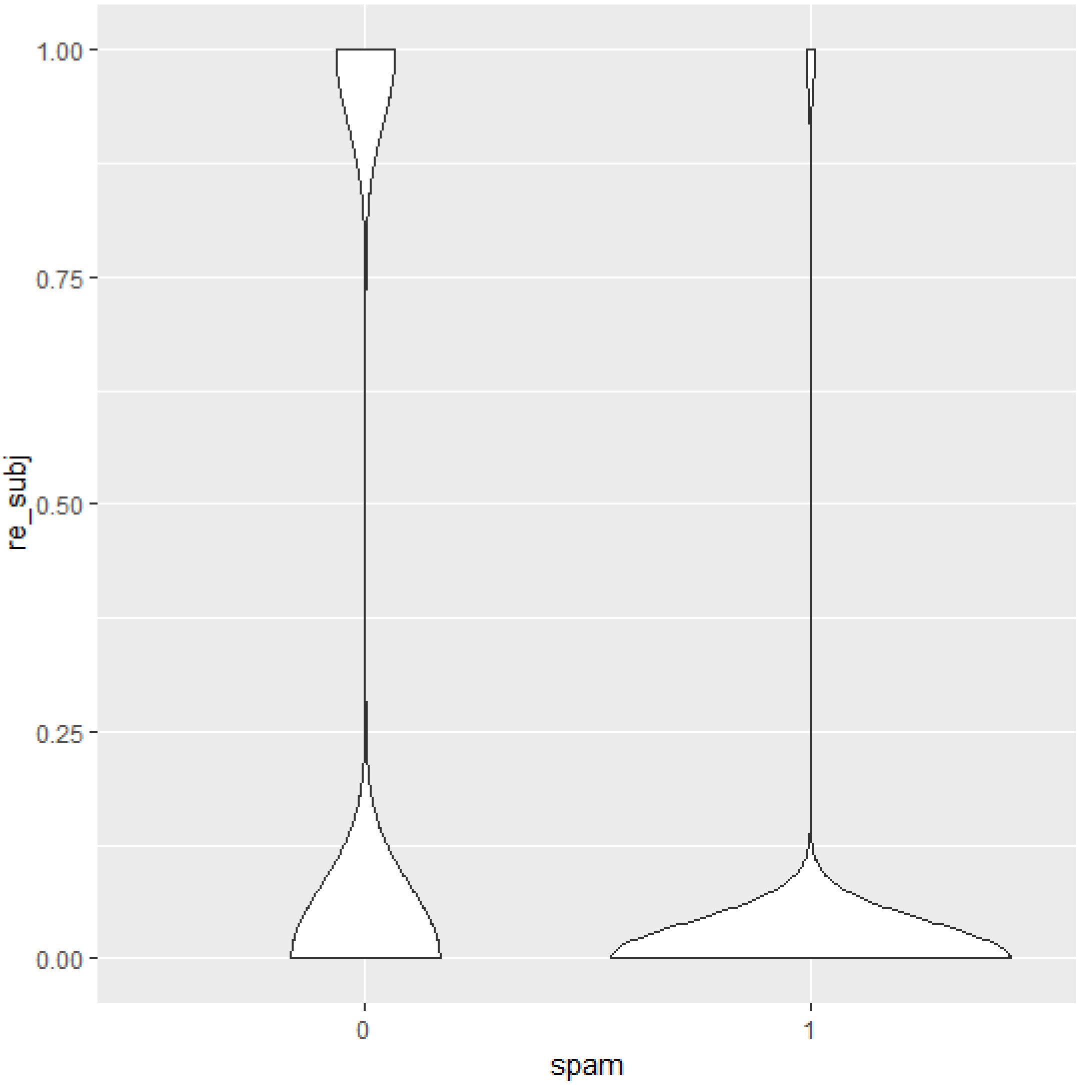
The data suggests that if an email was not written using HTML, it is more likely to be a spam.

spam vs re_subj

Table:

	0 (subj did not start with re)	1 (subj started with re)
0 not spam	2537	1017
1	359	8

Violin plot



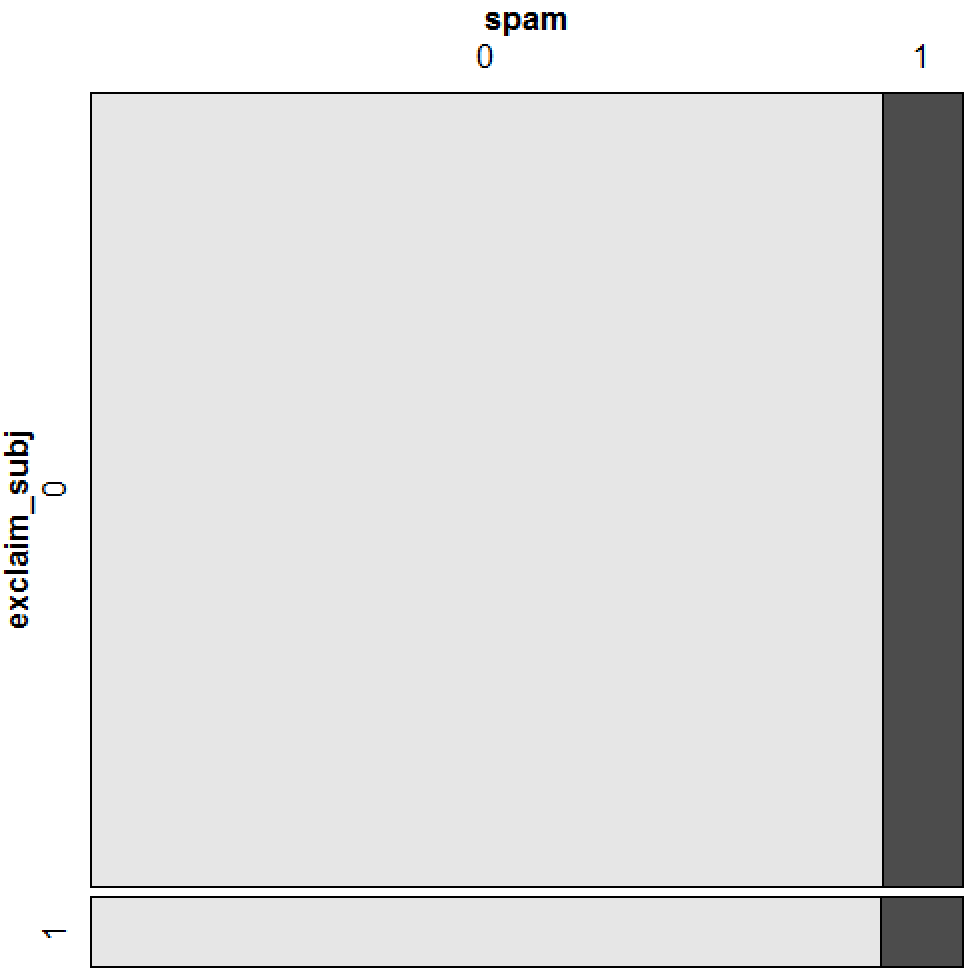
If the subject line started with re or its variation it is more likely not to be a spam. However, it is not 100% certain because there were 8 emails which started with re and was a spam.

Spam vs exclaim_subj

Table

	0 (subj did not have !)	1 (subj had !)
0 not spam	3269	285
1	337	30

Mosaic plot



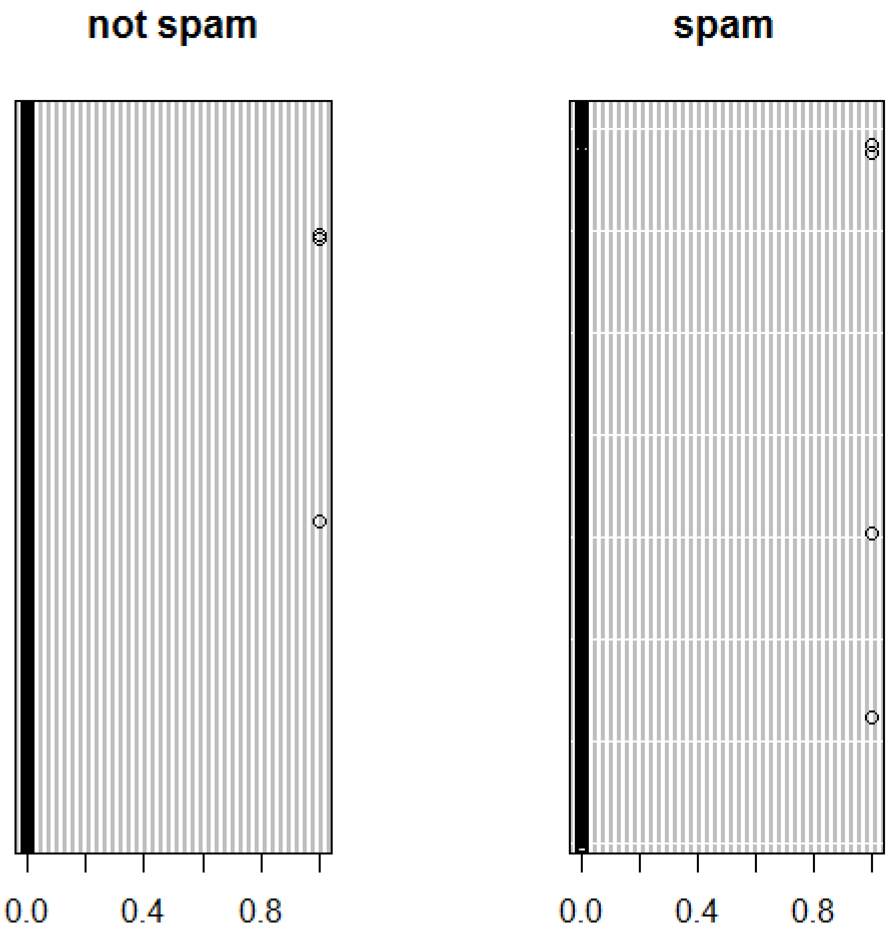
If the subject line had exclamation point, we can conclude from the graph that they are almost equally likely to be spam or not spam. However, the graph also shows that if the email had exclamation mark then, there a faint miniscule more chance of it being spam than not spam.

Spam vs urgent_subj

Table

	0 (email did not have urgent in subj)	1 (email had urgent in subj)
0 not spam	3551	3
1	363	4

Dot chart



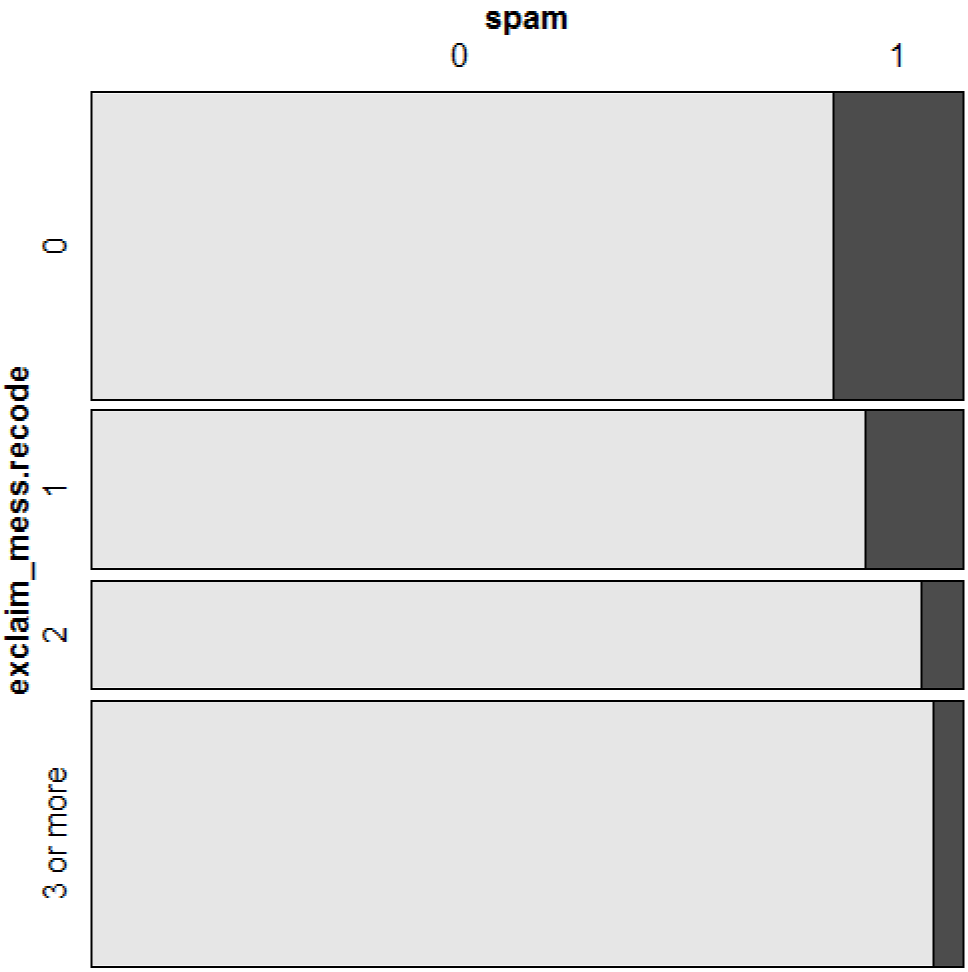
We can see from the dot chart that almost all emails did not have the word urgent in subject. Although, there were 4 spam emails compared to 3 not spam emails which had the word urgent in subject, proportionally not small emails have larger volume. So, we can conclude that if there is urgent in the subject line it is most likely not spam.

Spam vs exclaim_mess

The table:

	0	1	2	3 or more
0 (Not Spam)	1219	650	482	1203
1 (Spam)	216	83	25	43

Mosaic plot:



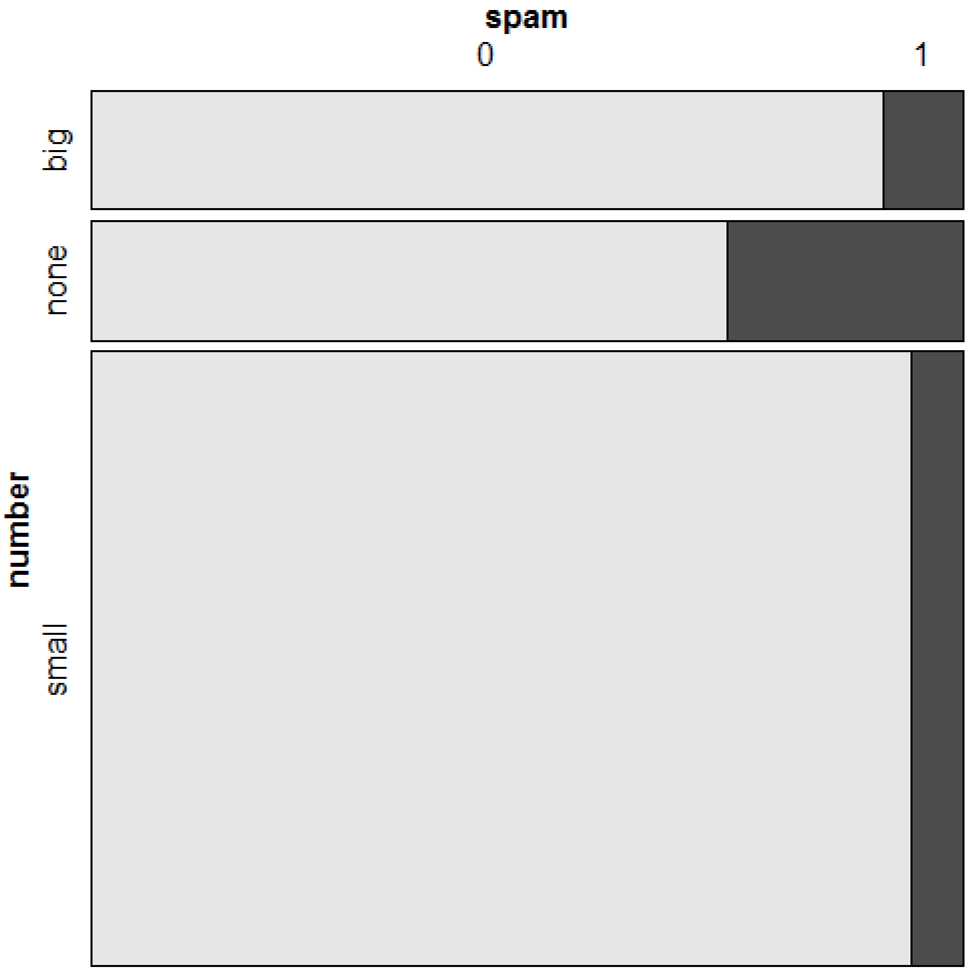
If there are 1 or more exclamation points in an email, it is more likely to be not spam. The table and mosaic plot show most spam with exclamation points have only 1 or 2 in an email.

Spam vs number

Table

	Big	None	Small
0	495	400	2659
1	50	149	168

Mosaic plot



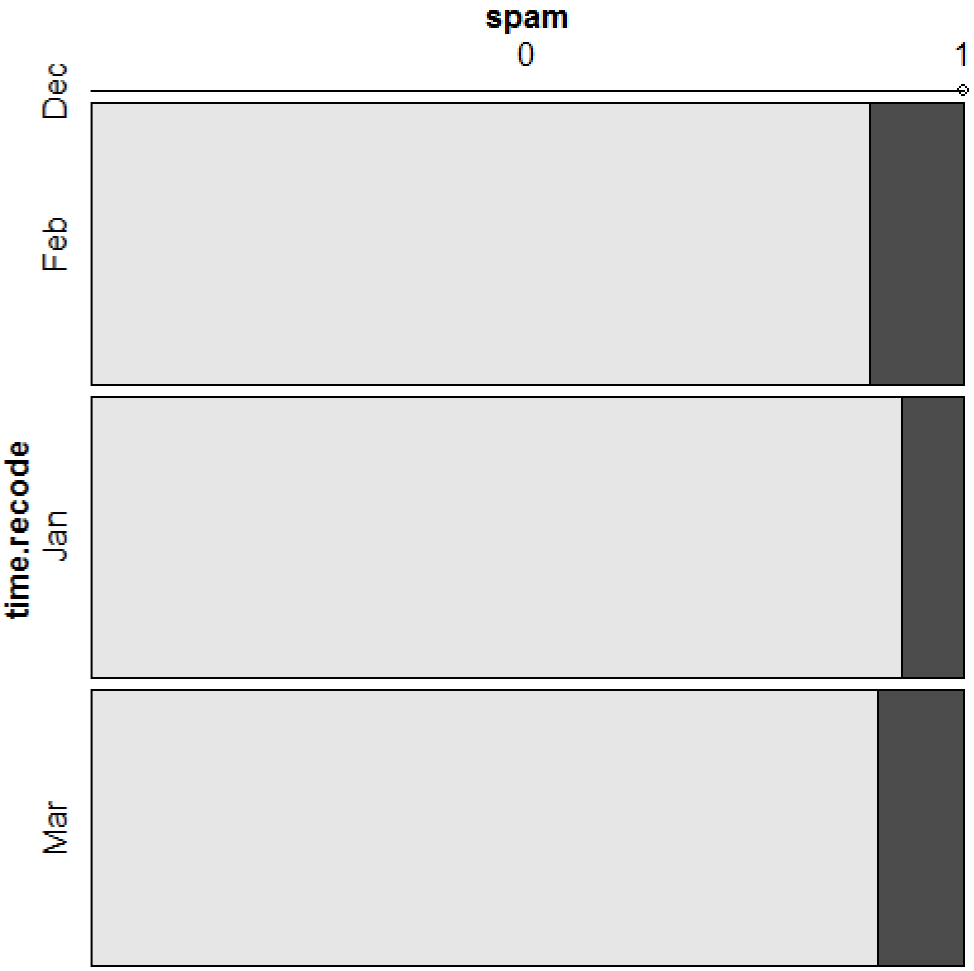
The plot suggests that if an email has a bigger number it is more likely to be a spam than if it were to have a smaller number (under a million).

Spam vs time

Table

	December (2012)	Janurary	February	March
0 Not spam	2	1217	1175	1160
1 spam	0	95	143	129

Mosaic plot



Based on the plot we can conclude that emails received in January are less likely to be spam than emails received during February and March. In December of 2012 only 2 emails were received since we are counting from the 31st , so it would be wise to ignore the two emails for this analysis.

Part 2. Based on your findings, give a list of variables that you think should be included for further analysis.

If I were to determine spam and not spam I would combine different parts of multiple variables to compare vs spam and not spam for my further analysis. I would use

from, cc, sent_email, image, attach, dollar, winner, password, num_char, line_breaks, format, urgent, and number

For example, I would like to see if emails with number bigger than a million has the word dollar in it.