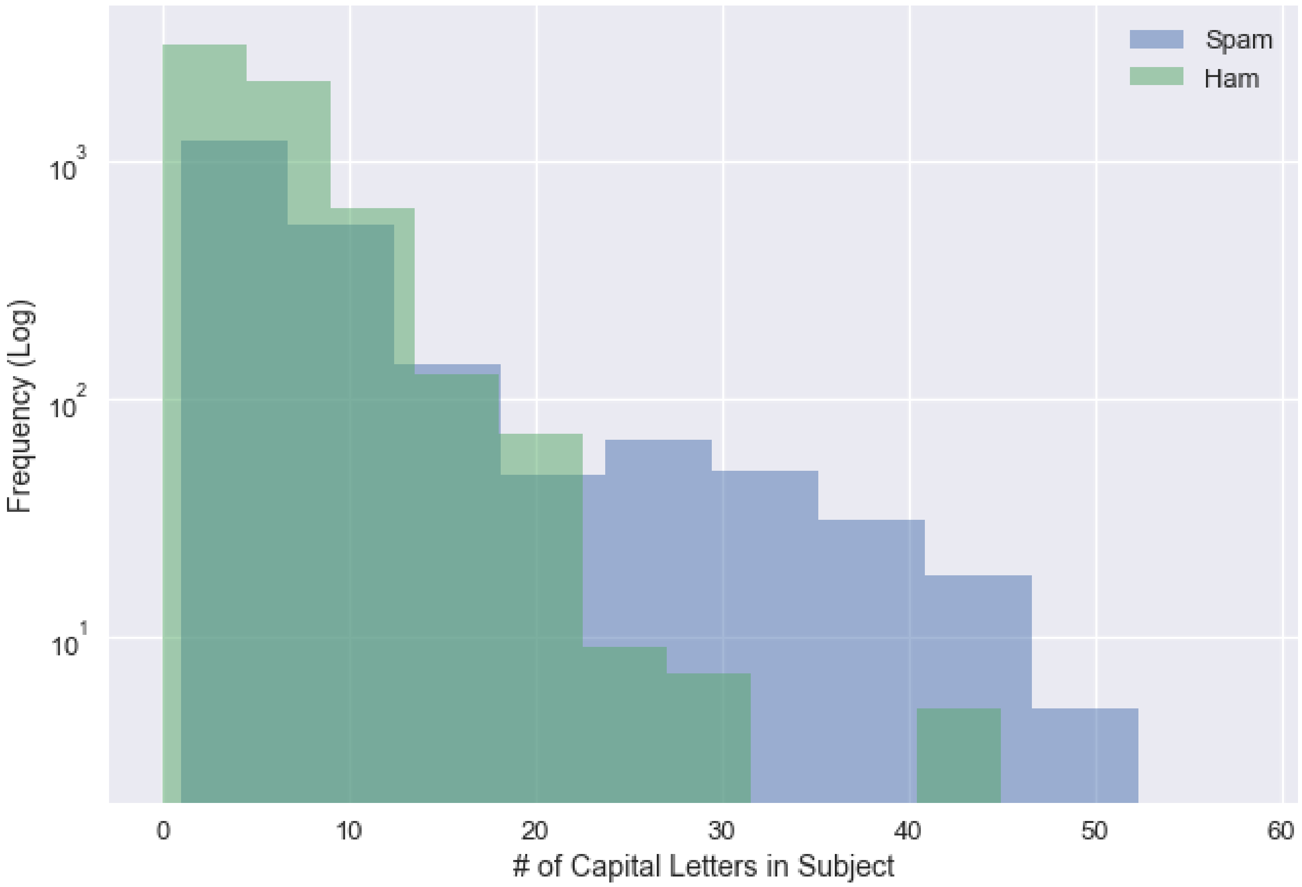
In [619]:

```
output:
```

I created a histogram that displays the number of capital letters that exist in the subject field of spam emails vs. ham emails, with the y-axis logged. We can see that the plot is skewed to the right, with a peak in the range 0-10 for both ham and spam emails. Despite seeing higher frequencies in the 20-50 range for spam emails, it disproved my original belief that capital letters in subject fields are indicative of spam emails - it's hard to specify a definite cut-off to determine whether the email is spam or ham.


Spam
Ham
Frequency (Log)
10^3
10^2
10^1
# of Capital Letters in Subject