EDA Project

Ethan La Voie

9/23/2024

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readxl)
```

Introduction

```
email <- read.table("C:\\Users\\ethan\\OneDrive\\Desktop\\Math 130\\email.txt", header = TRUE, sep="\t"

str(email)
```

```
## 'data.frame':    3921 obs. of  21 variables:
##  $ spam        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ to_multiple : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ from        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cc          : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ sent_email  : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ time        : chr  "2011-12-31 22:16:41" "2011-12-31 23:03:59" "2012-01-01 08:00:32" "2012-01-01
##  $ image       : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ attach      : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ dollar      : int  0 0 4 0 0 0 0 0 0 0 ...
##  $ winner      : chr  "no" "no" "no" "no" ...
##  $ inherit     : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ viagra      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ password    : int  0 0 0 0 2 2 0 0 0 0 ...
##  $ num_char    : num  11.37 10.5 7.77 13.26 1.23 ...
##  $ line_breaks : int  202 202 192 255 29 25 193 237 69 68 ...
##  $ format      : int  1 1 1 1 0 0 1 1 0 1 ...
##  $ re_subj     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ exclaim_subj: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ urgent_subj : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ exclaim_mess: int  0 1 6 48 1 1 1 18 1 0 ...
##  $ number      : chr  "big" "small" "small" "small" ...
```

My data set is emails that were sent to a guy in 2012. the variables that I will be analyzing are whether or not the email was spam and if it said they were a winner or not. I'm hoping to figure out whether or not having winner in the email makes it a scam or not Spam

self explanatory, this is whether or not the email is confirmed spam or not. taking a good look at this will help me filter out which ones I need to look at. Those being spam being high priority as finding a pattern between those and winner will prove my theory.

Winner

If I Were to simply able to look at if the email said I was a winner and know its a scam it would make life easier, but you also want a lot of emails to say it so that it can be easy to find the spam.

Univariate Exploration

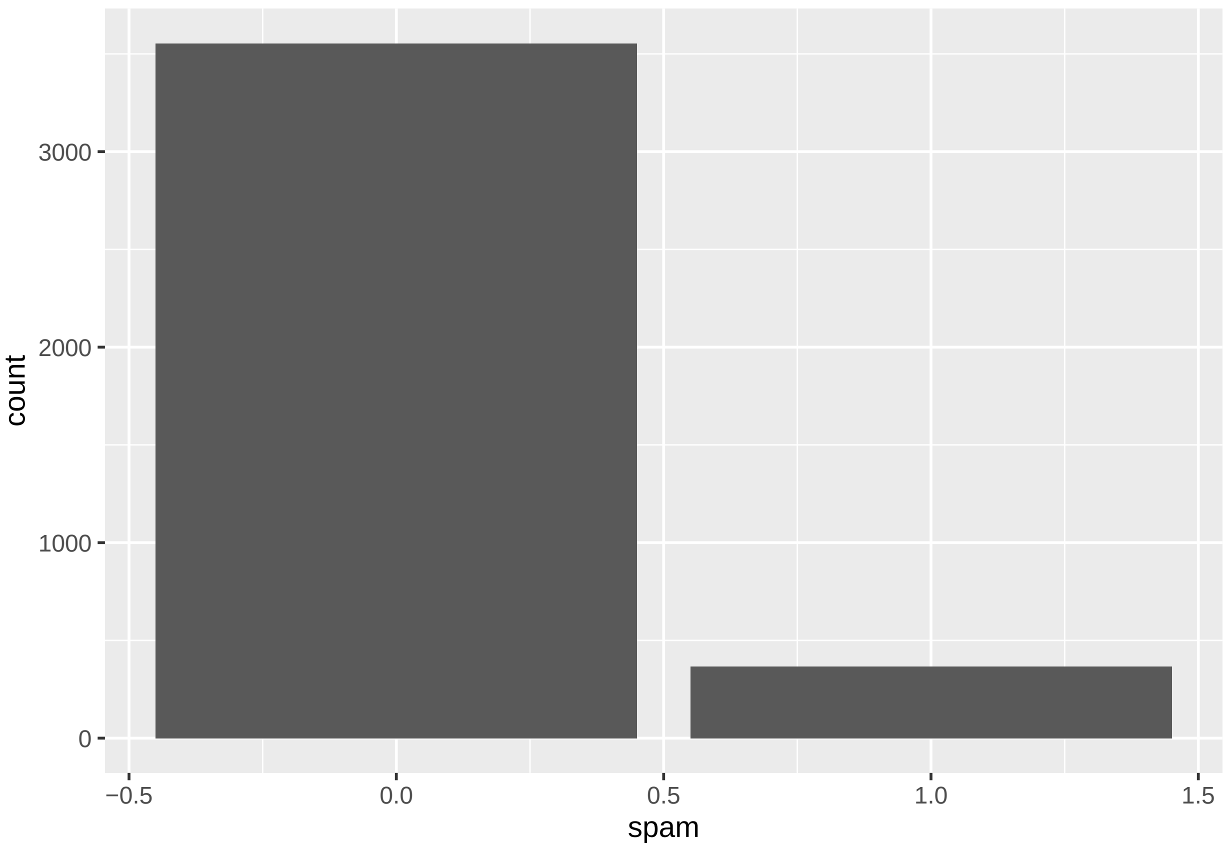
Lets take a look at these indivudially to see what we have got to work with.

```r
table (email$spam)
```

```
##
##    0    1
## 3554  367
```

This is just to get a good look at how many of these emails are spam or not

```r
ggplot(email, aes(x=spam)) + geom_bar()
```



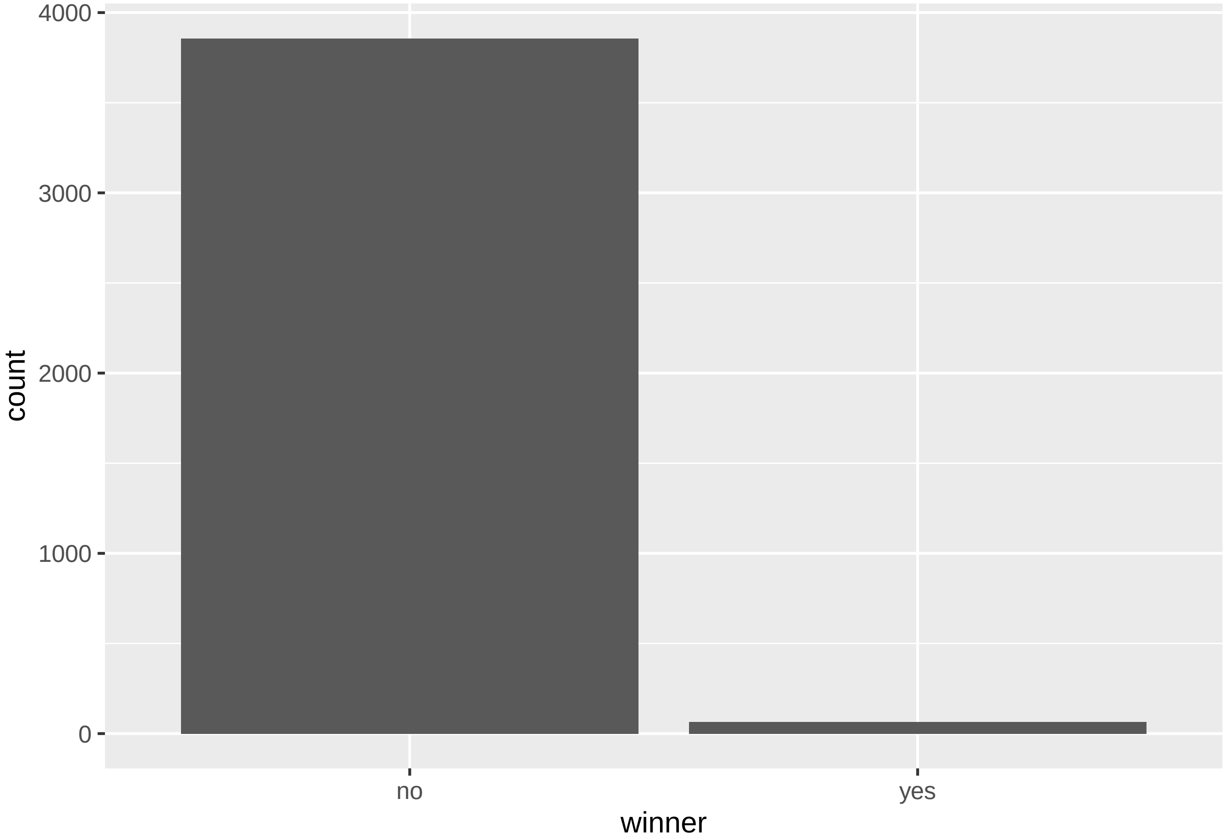
this shows that a lot of his emails were not spam which means we get to look at a smaller number.

```
table(email$winner)
```

```
##
##   no  yes
## 3857   64
```

64 out of 3921 emails said winner of some sort. not a very high number. Not great for the analysis but not terrible.

```
ggplot(email, aes(x=winner)) + geom_bar()
```



this bar graph shows in clear view that a very small number of emails are scams so this tells me that seeming that if it has a winner or not wont get all of the scams if any.

Bivariate Exploration

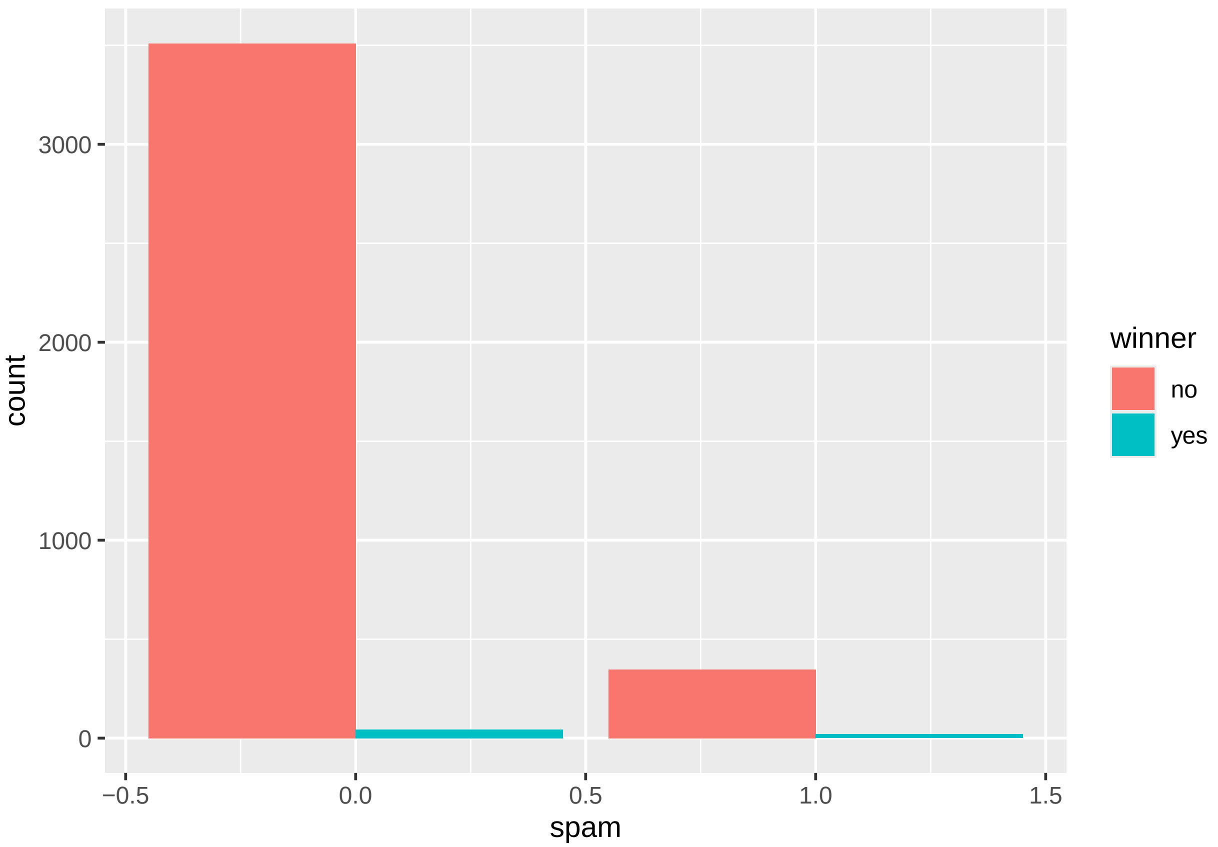
lets compare the two together to get a analysis going.

```
table(email$spam, email$winner)
```

```
##
##       no  yes
##   0 3510   44
##   1  347   20
```

1 means spam and yes means winner. Out of the 367 emails that were spam only 20 of them contained winner with the other 44 not being spam. Meaning its about 50/50 on whether or not a winner email is real or not.

```
ggplot(email, aes(x=spam, fill=winner)) + geom_bar(position = "dodge")
```



this graph is a good visual example of out comparison. The amount of spam and not spam containing winner is not high at all.

Conclusion

Surprisingly there was a low amount of spam emails that said winner and even lower were the ones that were actually spam. Unfortunately this proves my theory wrong, just looking at whether or not your email says that you're a winner is not enough proof to say that its a spam. what you can do with this information is give your grandma a break when she gets caught by that winner spam email as statistically you are more likely to run into a winner email that's not spam than one that is.