MATH130_FINAL

Ella Andrew

2022-09-26

INTRODUCTION

```
email <- read.table("/Users/ellaandrew/Desktop/MATH130/Data/email.txt", header=TRUE, sep="\t")
```

In this document, I will be exploring the 'email' data set from our course website, which is representative of all of the emails that David Diez received through his gmail account in January through March of 2012. I am specifically doing some digging on whether the identity of an email as a "spam" email has any impact on the number of times that a dollar sign ($) or the word "dollar" appears in the email.

I will be pursuing this question by looking at two variables:

1. 'spam': a variable that indicates whether or not an email was a spam email
2. 'dollar': a variable that indicates the number of times that a dollar sign or the word "dollar" appears in an email

UNIVARIATE EXPLORATION

The first variable that I chose to explore in the 'email' data set was 'spam'. This variable simpply takes all of the emails within the data set and determines whether they were or were not spam emails. However, it sorts the emails into '0' and '1' labels, rather than 'no' and 'yes' labels.

```
email$spam_fac <- factor(email$spam, labels=c("no", "yes"))
table(email$spam, email$spam_fac, useNA="always")
```

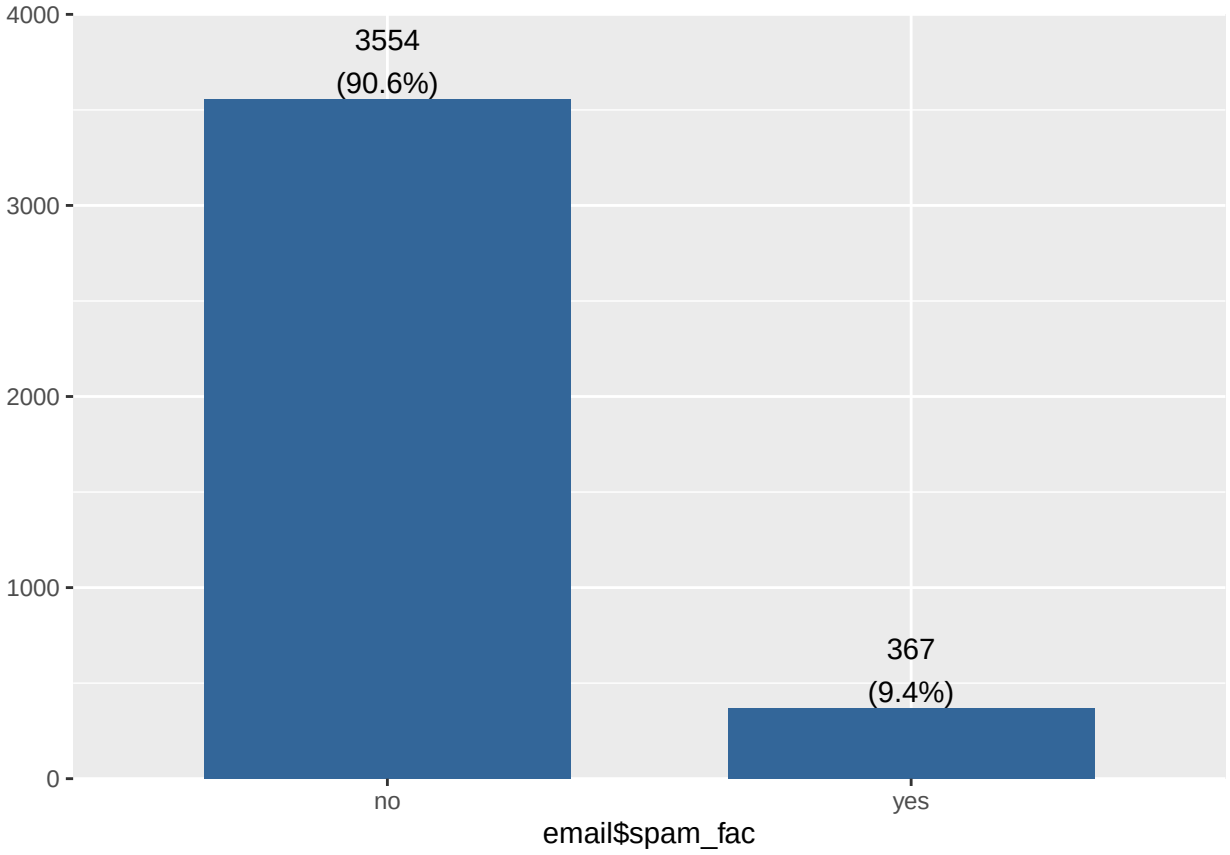
```
##
##          no  yes <NA>
##   0    3554    0    0
##   1       0  367    0
##   <NA>    0    0    0
```

In order to present 'spam' as categorical data, I changed the previous label '0' to 'no', and I changed the previous label '1' to 'yes'. Now the data is presented as categorical rather than quantitative.

```
library(sjPlot)
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
plot_frq(email$spam_fac)
```

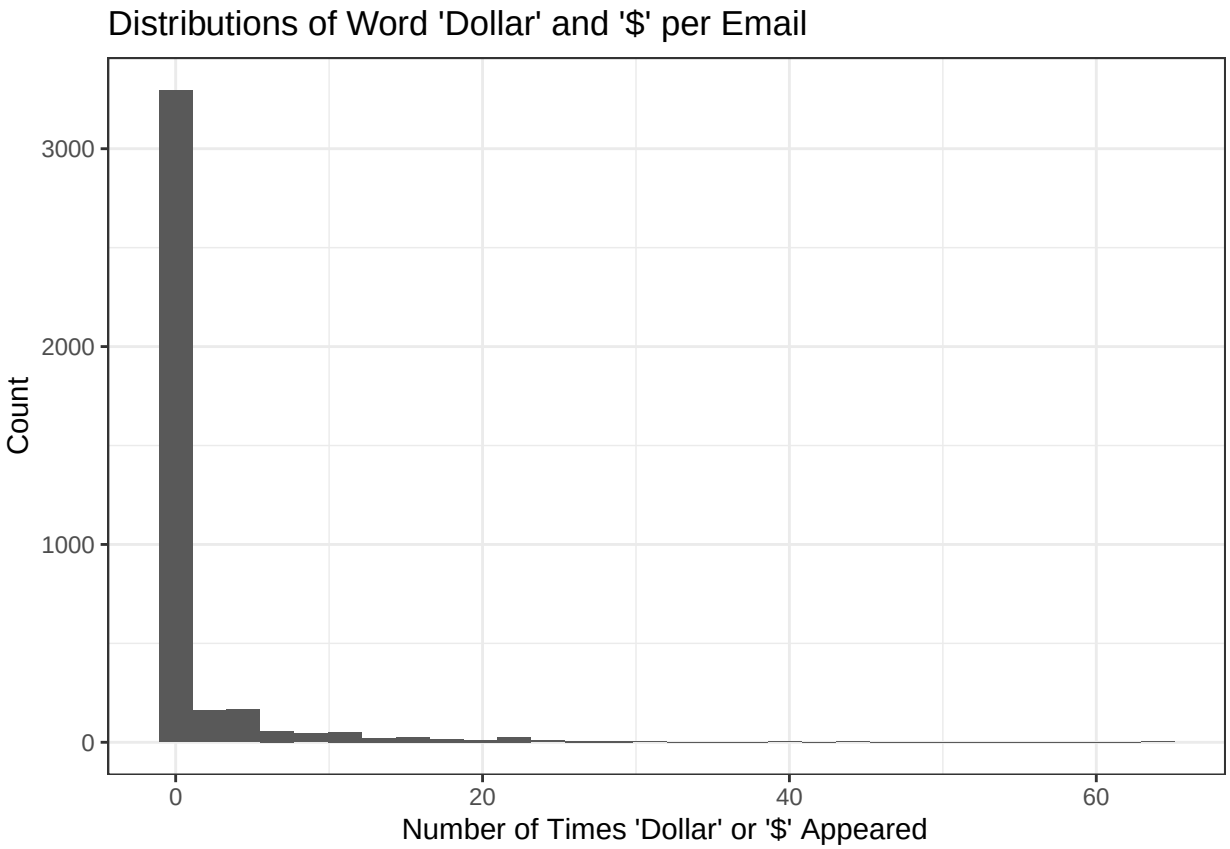


This bar chart shows that 90.6% of the emails that David Diez received through his gmail account in the first three months of 2012 were NOT spam emails, whereas 9.4% of the emails WERE spam emails.

Next, I wanted to explore the variable 'dollar', which indicates the number of times that a dollar sign or the word "dollar" appeared in every email. Now, the vast majority of the emails in this data set had zero occurrences of a dollar sign or the word "dollar," which made the bar chart of the variable a little difficult to read. So I also created a summary of the data to tell us a little more about the variable.

```
library(ggplot2)
ggplot(email, aes(x=dollar)) + geom_histogram() + theme_bw() + ggtitle("Distributions of Word 'Dollar' a
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



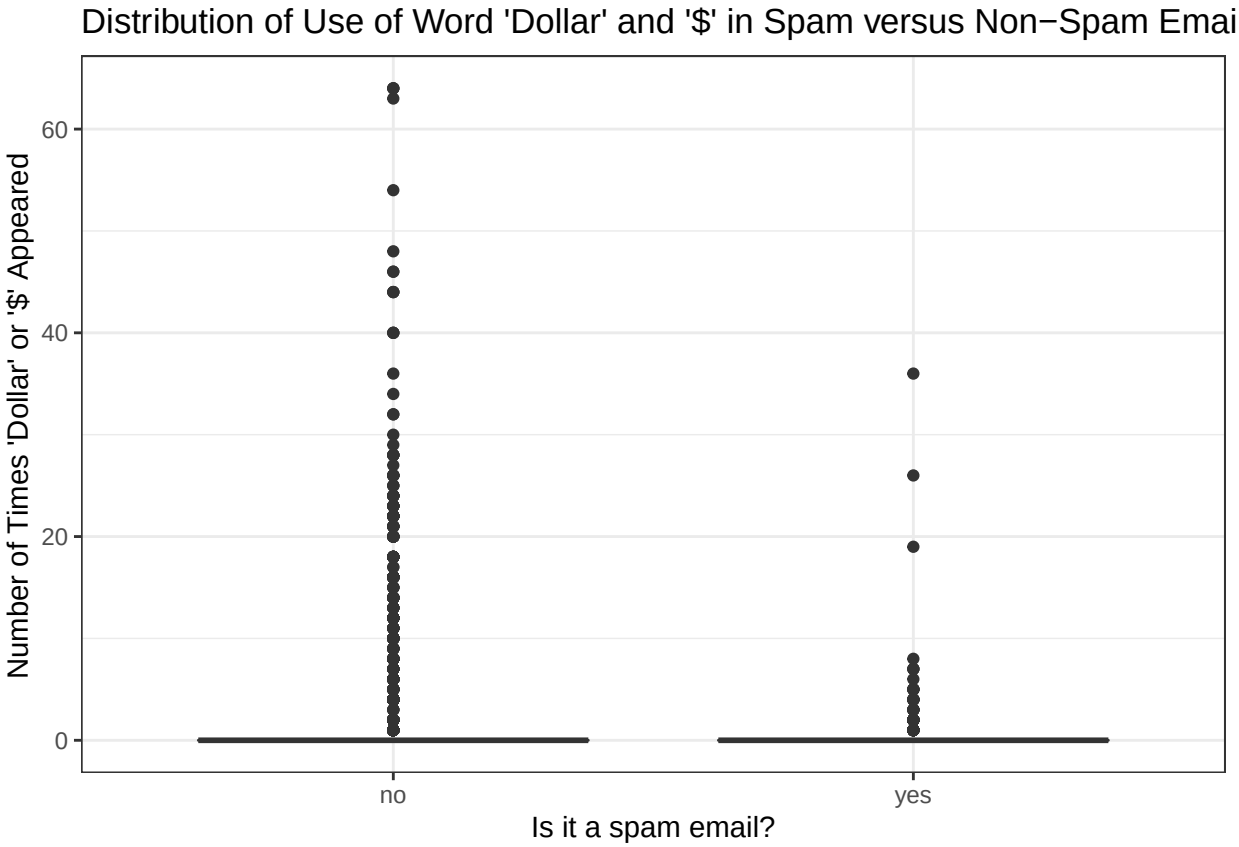

```
summary(email$dollar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.467   0.000  64.000
```

This tells us that, for the most part, emails did not contain a dollar sign or the word "dollar." However, it also shows us that the mean of the variable is likely skewed to the right of the graph due to large outliers (such as the maximum, 64).

BIVARIATE EXPLORATION

```
ggplot(email, aes(y=dollar, x=spam_fac)) + geom_boxplot() +theme_bw() + ggtitle("Distribution of Use of
```



This chart shows that, for the most part, BOTH spam and non-spam emails did NOT contain a dollar sign or the word "dollar." However, there were many emails in which a dollar sign or the word "dollar" appeared in both categories. Apparently, the non-spam emails generally had higher numbers for their 'dollar' observations than the spam emails.

CONCLUSION

The maximum number of times that a dollar sign or the word "dollar" appeared in this data set occurred in a NON-SPAM email, rather than a spam email. Not only this, but the non-spam email category actually has several observations that had more occurrences of dollar signs and the word "dollar" than the spam emails did. I found this surprising, because I had assumed that the spam emails would contain the maximum for the data set overall, and that it would appear to be higher on the y-axis in this box-plot in general due to having more occurrences of dollar signs and the word "dollar." Spam emails tend to try to hook people in through opportunities to make money.

Upon further thought, however, this distribution of data makes sense to me. David Diez, according to our course website, is the author of a book called "An Open Intro Statistics Textbook." The data distribution makes much more sense to me now that I know that the person who was sending and receiving emails containing a dollar sign or the word "dollar" was a professor (or at least a professional) in the field of statistics. It makes me curious to see these two variables compared to one another in the email account of someone who does not work with numbers for a living.