EDA_cvieira

2025-06-26

Introduction

For this project, the “email” data set will be analyzed. This data set represents incoming emails for the first three months of 2012 for David Diez’s gmail account. This project aims to observe any correlation of marketing give-away language (‘winner’), to positively human-identified spam emails (‘spam’). If any possible relationships are observed, this indicator may be used as a tool to assist in identifying spam emails.

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE)
library(ggplot2); 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
```

```
email <- read.table("../data/email.txt", header=TRUE, sep = "\t") %>%
  mutate(spam = factor(spam, levels = c(0,1), labels = c("Not Spam", "Spam")))
```

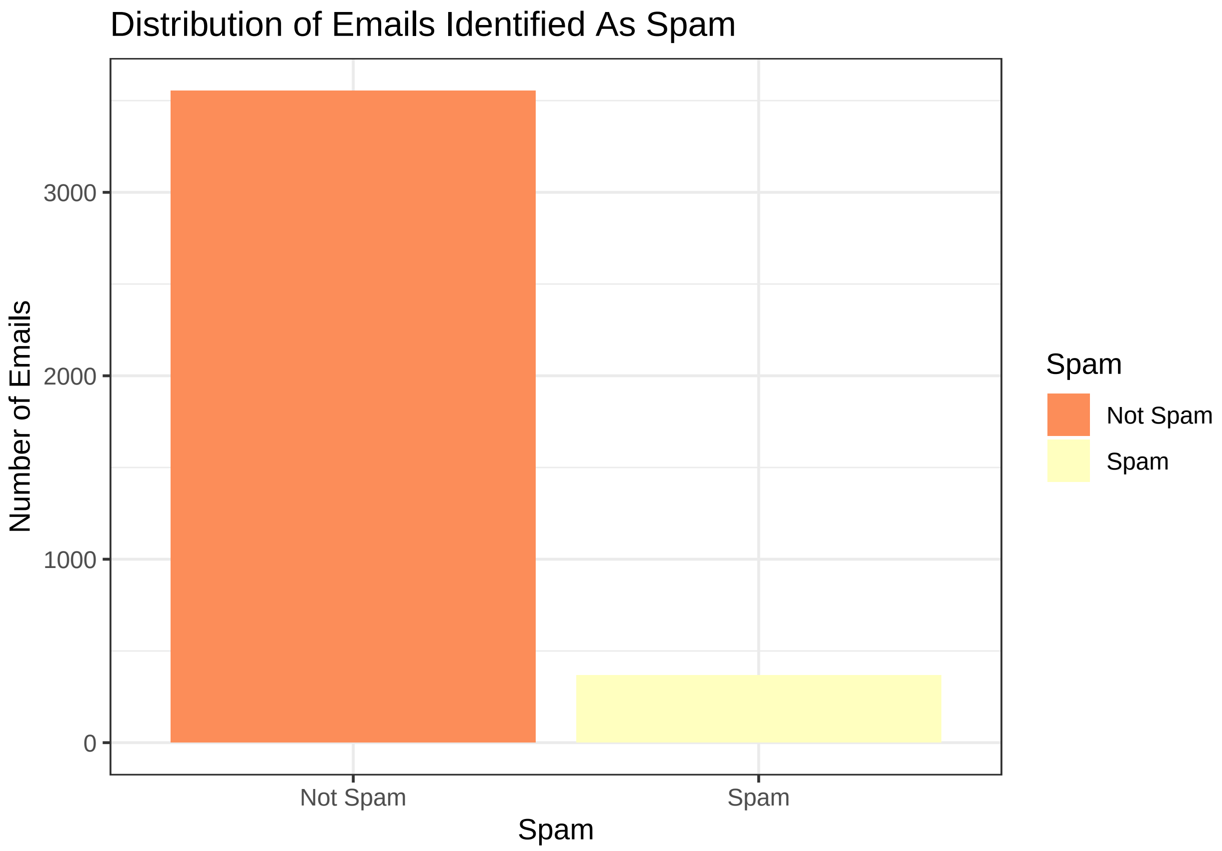
Univariate Exploration

The first variable to explored is ‘spam.’ This column of the data-set shows which of the emails are positively identified as spam. The figure below shows a little over 9% of total emails have been marked as spam - with the rest not being spam.

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

```
##
## Not Spam     Spam
##     3554      367
```

```r
ggplot(email, aes(x=spam, fill=spam)) +
  geom_bar() +
  ggtitle("Distribution of Emails Identified As Spam") +
  xlab("Spam") +
  ylab("Number of Emails") +
  scale_fill_brewer(palette = "Spectral", name = "Spam") +
  theme_bw()
```

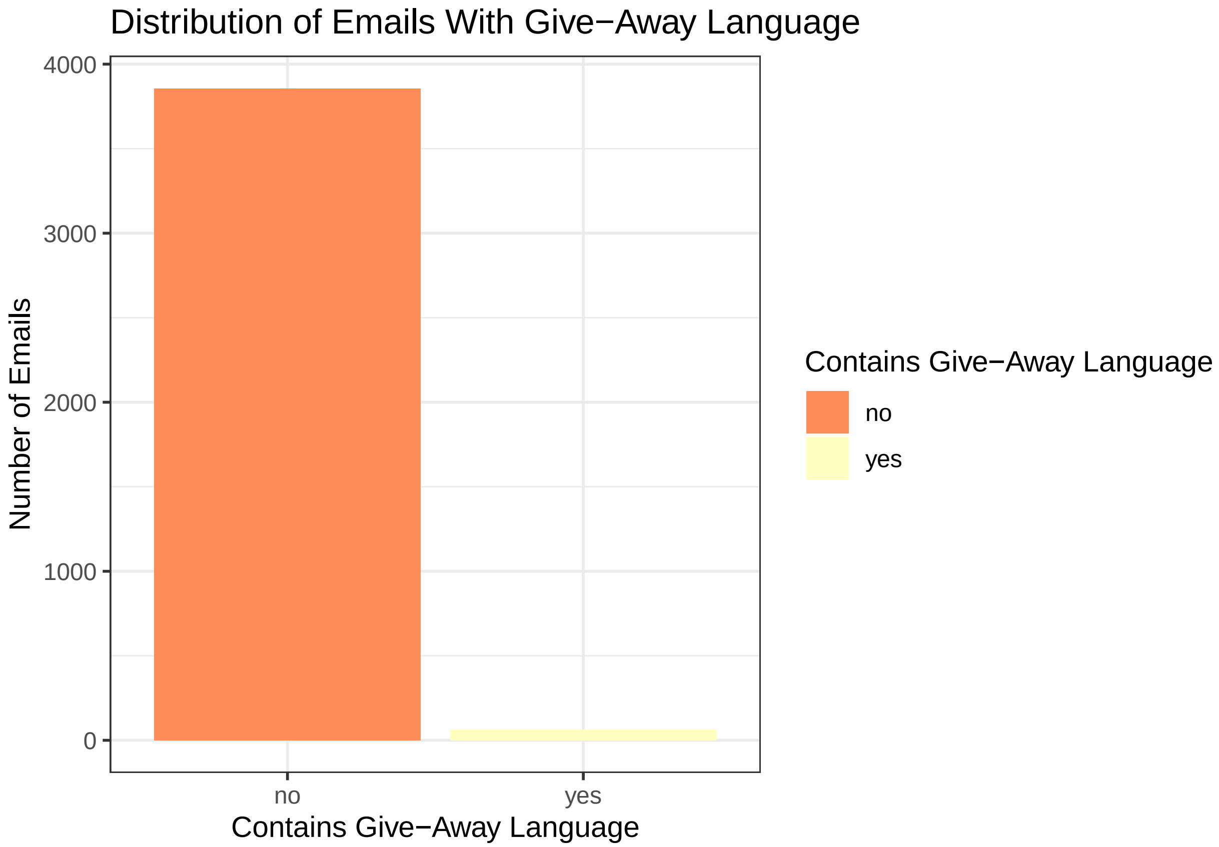


The second variable to be explored is presence of give-away language, or emails containing phrases like “winner.”

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

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

```r
ggplot(email, aes(x=winner, fill=winner)) +
  geom_bar() +
  ggtitle("Distribution of Emails With Give-Away Language") +
  xlab("Contains Give-Away Language") +
  ylab("Number of Emails") +
  scale_fill_brewer(palette = "Spectral", name = "Contains Give-Away Language") +
  theme_bw()
```



The bar chart above shows the vast majority of the emails do not contain give-away language. The distribution on its own is not very interesting. Observing the distribution of give-away language within spam emails may yield more interesting results. In the next section, how the two variables correspond with one another will be explored.

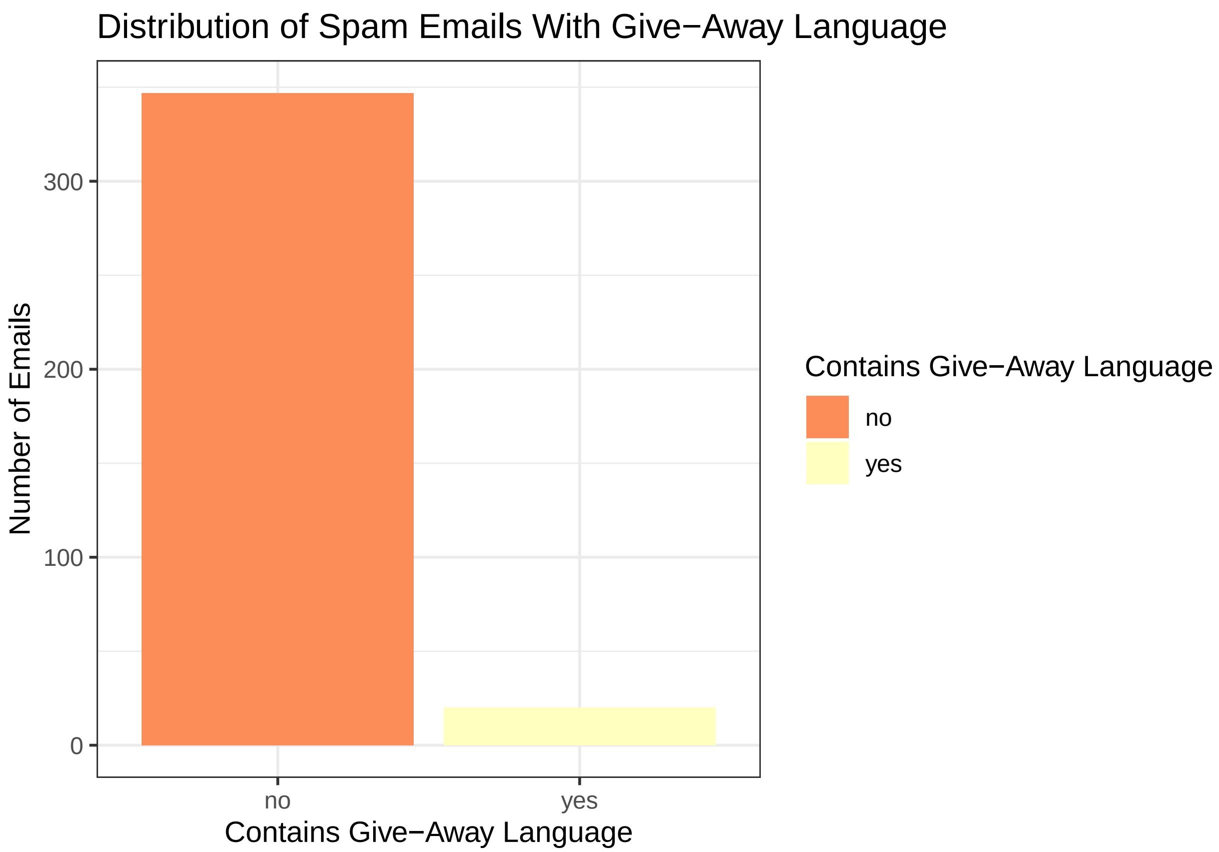
Bivariate Exploration

```
winner_email <- filter(email, winner > 0, spam == "Spam")

table(winner_email$winner)
```

```
##
##  no yes
## 347  20
```

```
ggplot(winner_email, aes(x=winner, fill=winner)) +
  geom_bar() +
  ggtitle("Distribution of Spam Emails With Give-Away Language") +
  xlab("Contains Give-Away Language") +
  ylab("Number of Emails") +
  scale_fill_brewer(palette = "Spectral", name = "Contains Give-Away Language") +
  theme_bw()
```



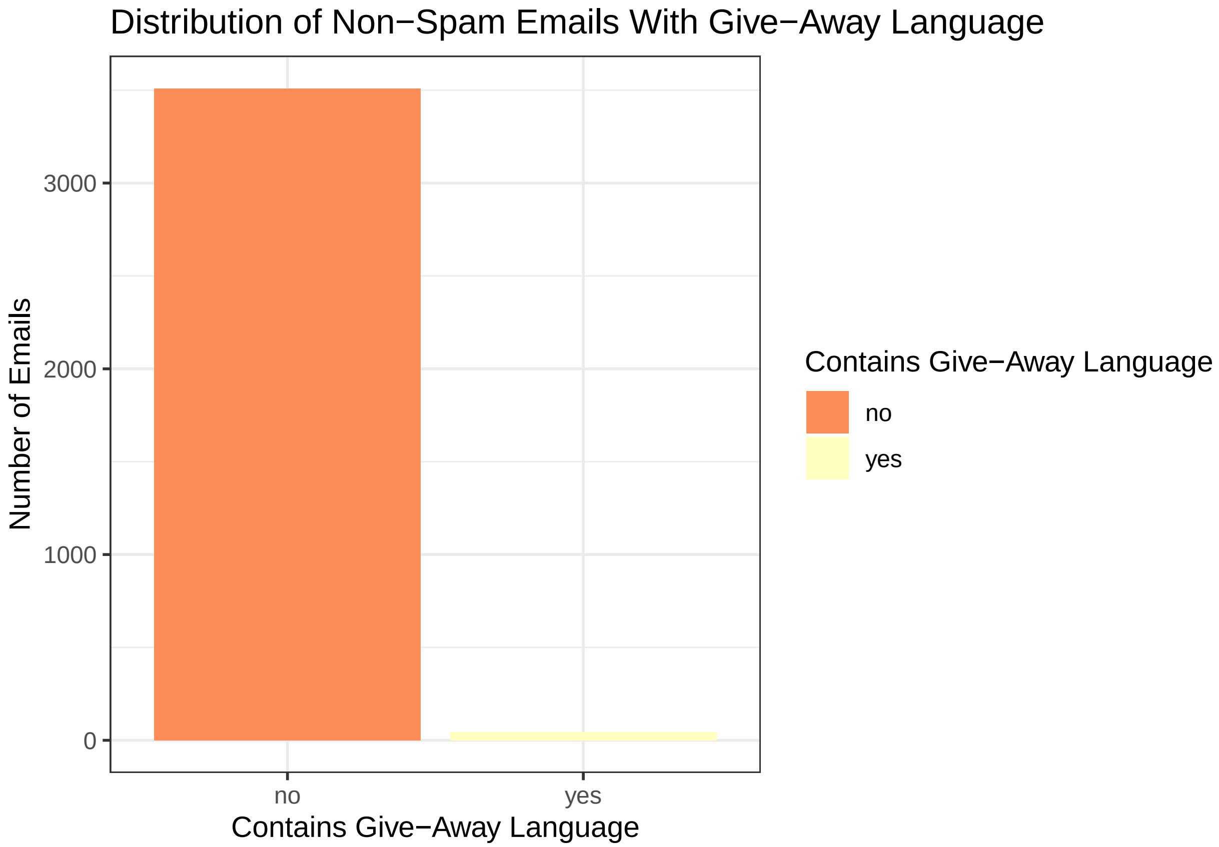
The distribution shows even after narrowing the scope to only spam emails, the majority of the emails do not contain give-away language. Around 6% of spam emails contain give-away language. As opposed to the 1% of non-spam emails that also contain give-away language as seen below.

```
nonwinner_email <- filter(email, winner > 0, spam == "Not Spam")

table(nonwinner_email$winner)
```

```
##
##   no  yes
## 3510   44
```

```
ggplot(nonwinner_email, aes(x=winner, fill=winner)) +
  geom_bar() +
  ggtitle("Distribution of Non-Spam Emails With Give-Away Language") +
  xlab("Contains Give-Away Language") +
  ylab("Number of Emails") +
  scale_fill_brewer(palette = "Spectral", name = "Contains Give-Away Language") +
  theme_bw()
```



Conclusion

The original hypothesis this exploration set out to prove was: emails with give-away language included can be an indication the email is spam. As can be seen from the figures above, there is some indication that give-away language is an indicator. The percentage of emails with give-away language in non-spam emails is around 1% while the number of spam emails with give-away language is 5%. Due to the percentage increase being very small, as well as give-away language being uncommon in emails overall, it is not reasonable to use this one indication to identify spam. Using this data set as an example, if the give-away language indicator were used to identify spam, it would result in 44 false positives compared to the 20 real catches. This means around 70% of the total catches would be wrong.