The Influence of Text Characteristics for Email Classification

Group 22

```r
library(ggplot2)
library(dplyr)
library(moderndive)
library(gapminder)
library(skimr)
library(tidyverse)
library(gt)
library(patchwork)
library(gridExtra)
library(broom)
library(knitr)
library(GGally)
```

```r
email<-read.csv("C:/Users/70652/Desktop/STATS5085 Data Analysis Skills/Project 2/DAS-Group-22
```

```r
email$yesno<-as.factor(email$yesno)
```

1 Exploratory Data Analysis

1.1 Correlation

```r
ggpairs(email[,1:6]) +
  theme(plot.background = element_rect(
    fill = "transparent",
    colour = NA,
    size = 1))
```

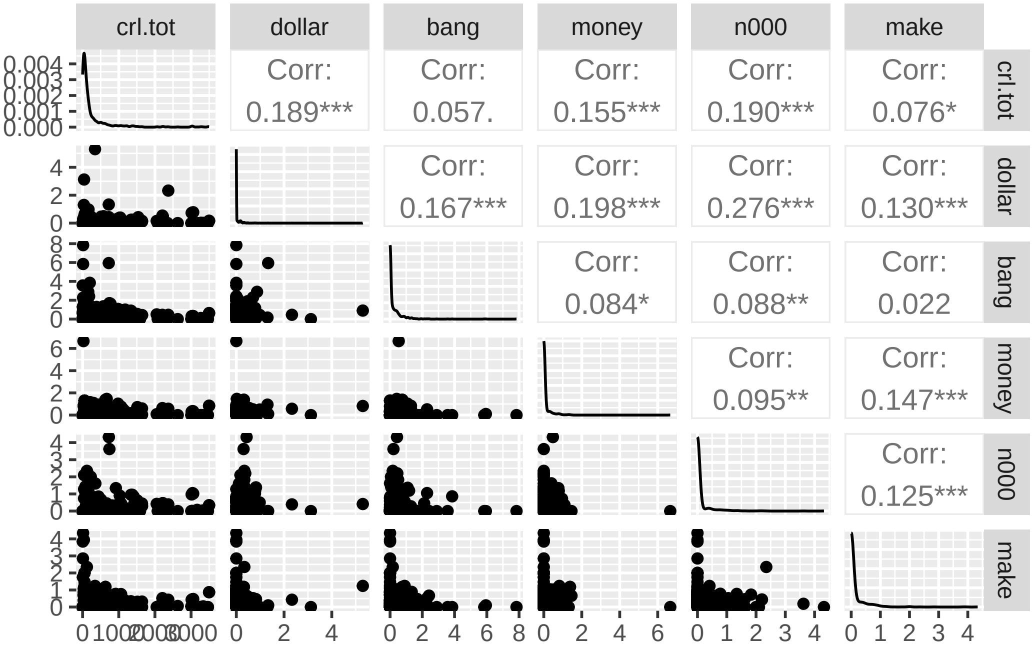

Figure 1: Correlations between each variables.

1.2 Data Visualization

```
ggplot(email, aes(x = yesno, y = crl.tot)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = "Uninterrupted sequences of capitals",
       title = "Spam indictor with total length of uninterrupted sequences of capitals")
```

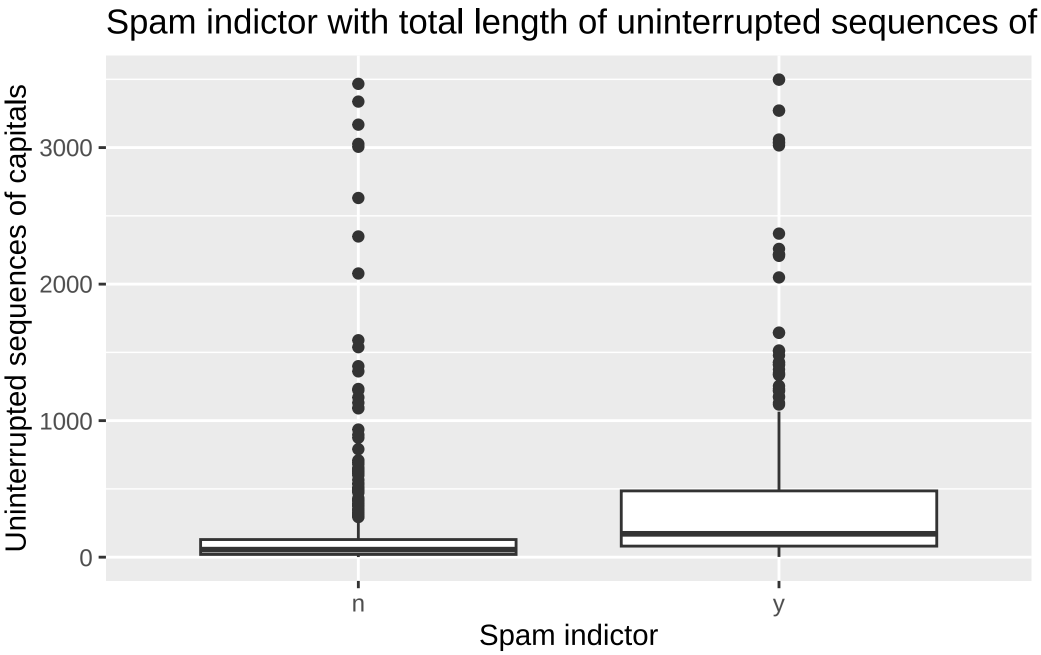

Figure 2: Boxplot of total length of uninterrupted sequences of capitals.

```r
ggplot(email, aes(x = yesno, y = dollar)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = "Occurrences of the dollar sign",
       title = "Spam indictor with occurrences of the dollar sign")
```

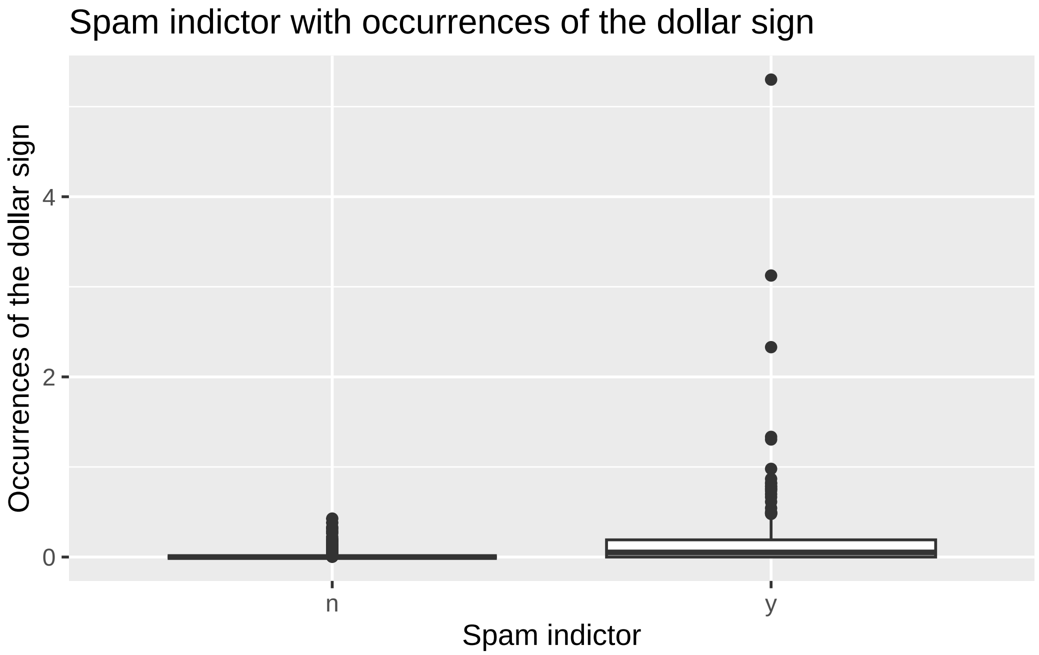

Figure 3: Boxplot of occurrences of the dollar sign.

```r
ggplot(email, aes(x = yesno, y = bang)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = 'Occurrences of "!"',
       title = 'Spam indictor with occurrences of "!"')
```

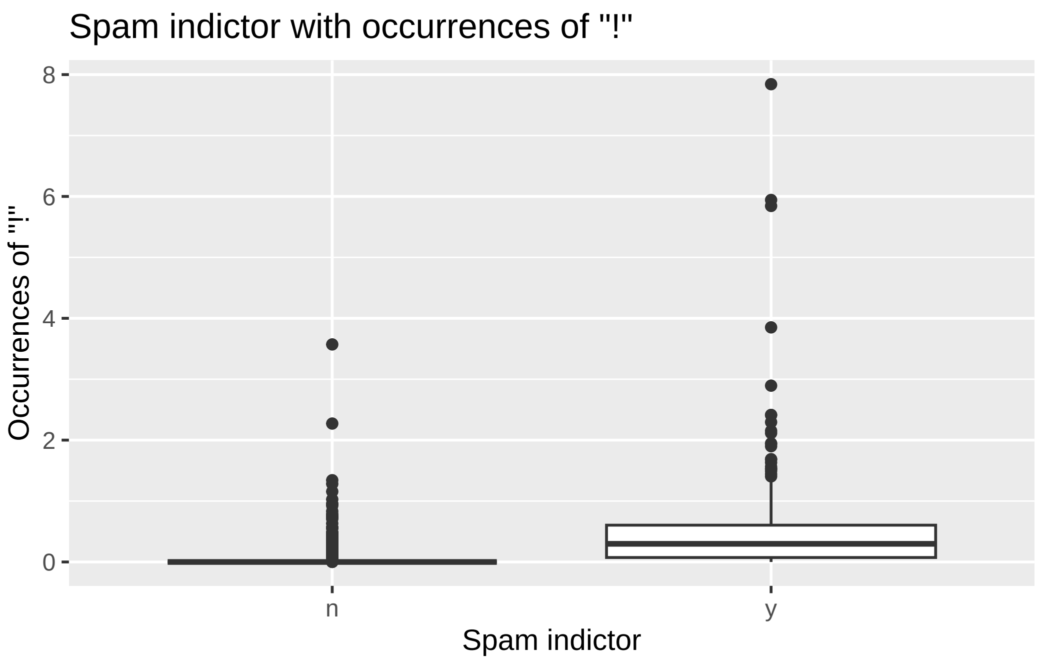

Figure 4: Boxplot of occurrences of '!'.

```
ggplot(email, aes(x = yesno, y = money)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = 'Occurrences of "money"',
       title = 'Spam indictor with occurrences of "money"')
```

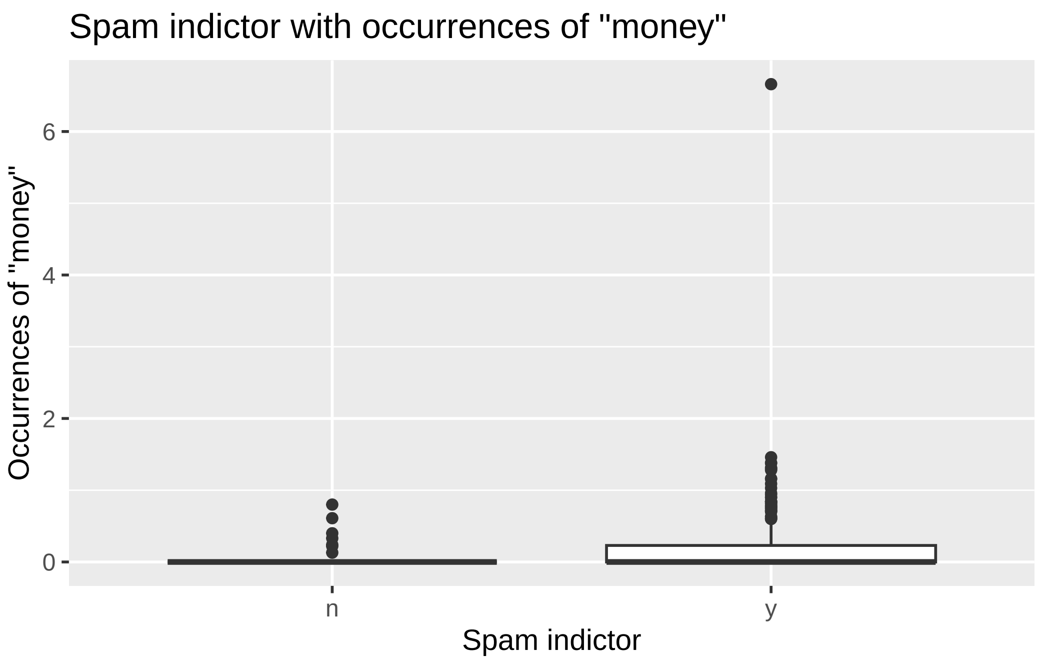

Figure 5: Boxplot of occurrences of “money”.

```
ggplot(email, aes(x = yesno, y = n000)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = 'Occurrences of "000"',
       title = 'Spam indictor with occurrences of "000"')
```

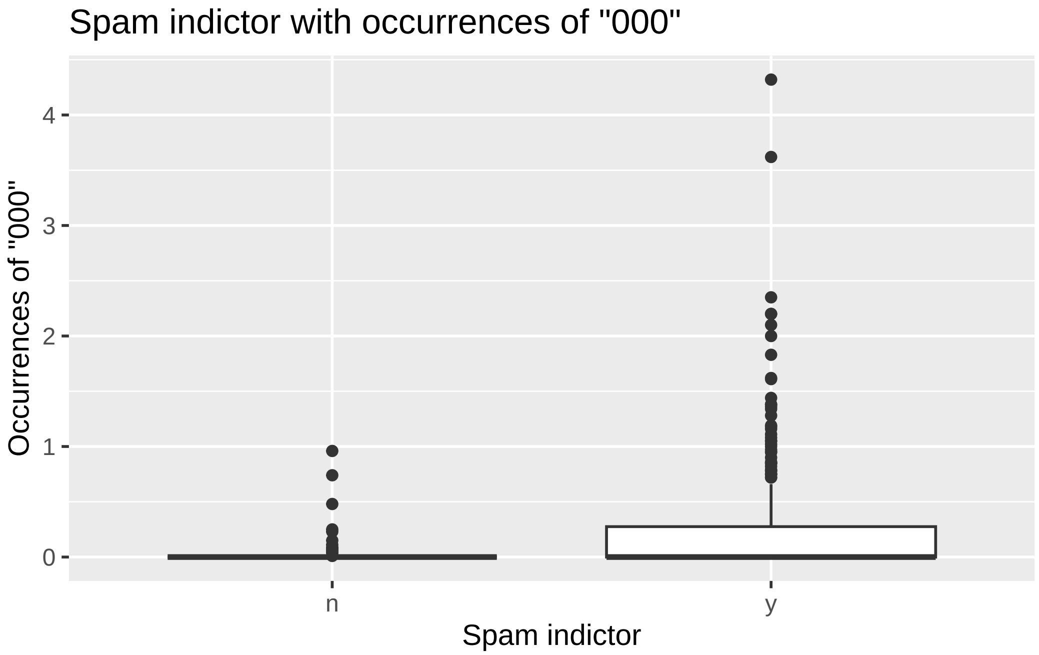

Figure 6: Boxplot of occurrences of ‘000’.

```
ggplot(email, aes(x = yesno, y = make)) +
  geom_boxplot() +
  labs(x = "Spam indictor", y = 'Occurrences of "make"',
       title = 'Spam indictor with occurrences of "make"')
```

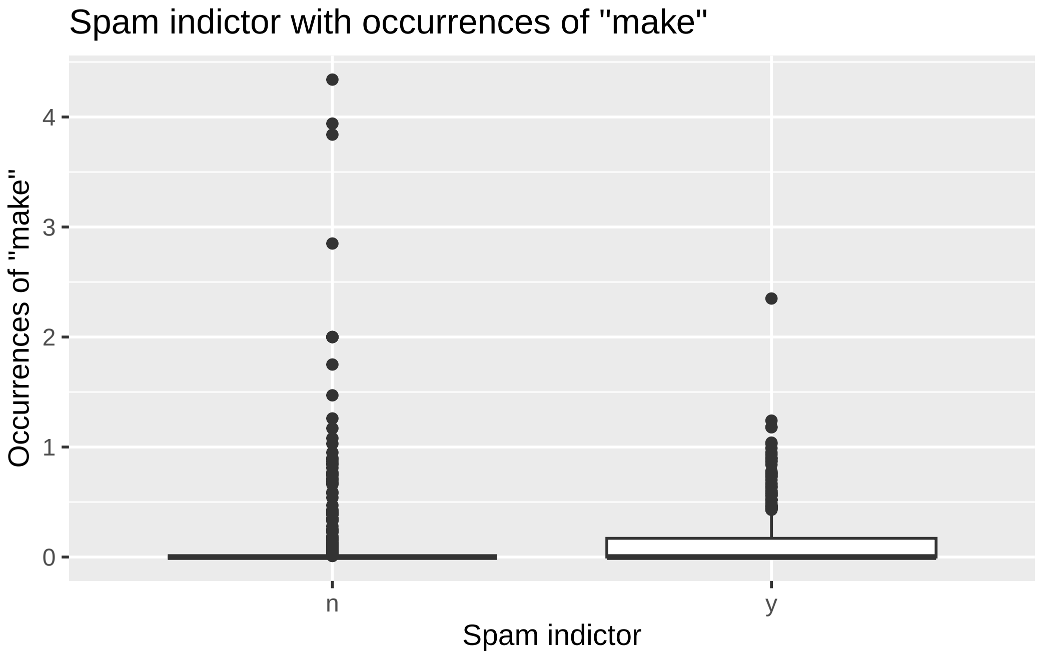

Figure 7: Boxplot of occurrences of 'make'.

2 Formal Data Analysis

```
model1 <- glm(yesno ~ crl.tot+dollar+bang+money+n000+make, data = email,
             family = binomial(link = "logit"))
```

```
summary(model1)
```

```

Call:
glm(formula = yesno ~ crl.tot + dollar + bang + money + n000 +
    make, family = binomial(link = "logit"), data = email)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.8101190  0.1254636 -14.427  < 2e-16 ***
crl.tot      0.0005502  0.0001886   2.917 0.003533 **
dollar       8.1346140  1.5396484   5.283 1.27e-07 ***
bang         2.9172085  0.3363971   8.672  < 2e-16 ***
```

```
money        5.9724851  1.2455257   4.795 1.63e-06 ***
n000         3.4827736  1.0261134   3.394 0.000688 ***
make        -0.4553154  0.4065463  -1.120 0.262731    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1234.17  on 919  degrees of freedom
Residual deviance:  752.44  on 913  degrees of freedom
AIC: 766.44

Number of Fisher Scoring iterations: 7
```

```
mod1coefs <- round(coef(model1), 2)
```

```
model2 <- glm(yesno ~ crl.tot+dollar+bang+money+n000, data = email,
             family = binomial(link = "logit"))
```

```
summary(model2)
```

```


Call:
glm(formula = yesno ~ crl.tot + dollar + bang + money + n000, 
    family = binomial(link = "logit"), data = email)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.8455026  0.1229712 -15.008  < 2e-16 ***
crl.tot      0.0005579  0.0001887   2.956 0.003115 ** 
dollar       8.1812828  1.5395890   5.314 1.07e-07 ***
bang         2.9348590  0.3371484   8.705  < 2e-16 ***
money        5.8334954  1.2421522   4.696 2.65e-06 ***
n000         3.4273127  1.0224981   3.352 0.000803 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1234.17  on 919  degrees of freedom
Residual deviance:  754.04  on 914  degrees of freedom
```

```
AIC: 766.04

Number of Fisher Scoring iterations: 7
```

```r
mod2coefs <- round(coef(model2), 3)
```

$$
\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \alpha + \beta_{crl.tot}\cdot \text{crl.tot} + \beta_{dollar}\cdot \text{dollar} + \beta_{bang}\cdot \text{bang} + \\
&\quad \beta_{money}\cdot \text{money} + \beta_{n000}\cdot \text{n000} \\
&= -1.846 + 0.001\cdot \text{crl.tot} + 8.181\cdot \text{dollar} + 2.935\cdot \text{bang} + 5.833\cdot \text{money} + 3.427\cdot \text{n000}
\end{aligned}
$$

```r
confint(model2) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-2.0926399	-1.610104
crl.tot	0.0001855	0.000936
dollar	5.3097543	11.355237
bang	2.3039250	3.626788
money	3.6506969	8.565015
n000	1.6521512	5.709387