How to classify an eamil as spam

Group 25: Hang Cheng, Haowei Yan, Weijan Li, Wenli Lyu, Zehao Wang

1 Introductions

Spam has caused some distress in people's daily life, so identifying spam correctly becomes more and more important nowadays. This study aims to finding Which text characteristics influence whether an email will be classified as spam or not by analyzing the data shared with the UCI Machine Learning Repository.

2 Data Reading

```
# Load the necessary package
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
```

```
# Read CSV data
d25 <- read.csv("dataset25.csv")
```

```r
# select different data
d25.spam <- d25 %>%
  select(yesno,crl.tot,dollar,bang,money,n000,make)
d25.spam$yesno <- as.factor(d25.spam$yesno)
d25.spam$crl.tot <- d25.spam$crl.tot/10
```

According to the data, six main characteristics may exert an influence on classifying an email as spam. We divide "crl.tot" by 10 because the number is much larger than other data.

3 Analysis of Six Main Characteristics

We firstly analyze these characteristics separately.

3.1 Total length of uninterrupted sequences of capitals

```r
ggplot(data=d25.spam,aes(x=yesno,y=crl.tot,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

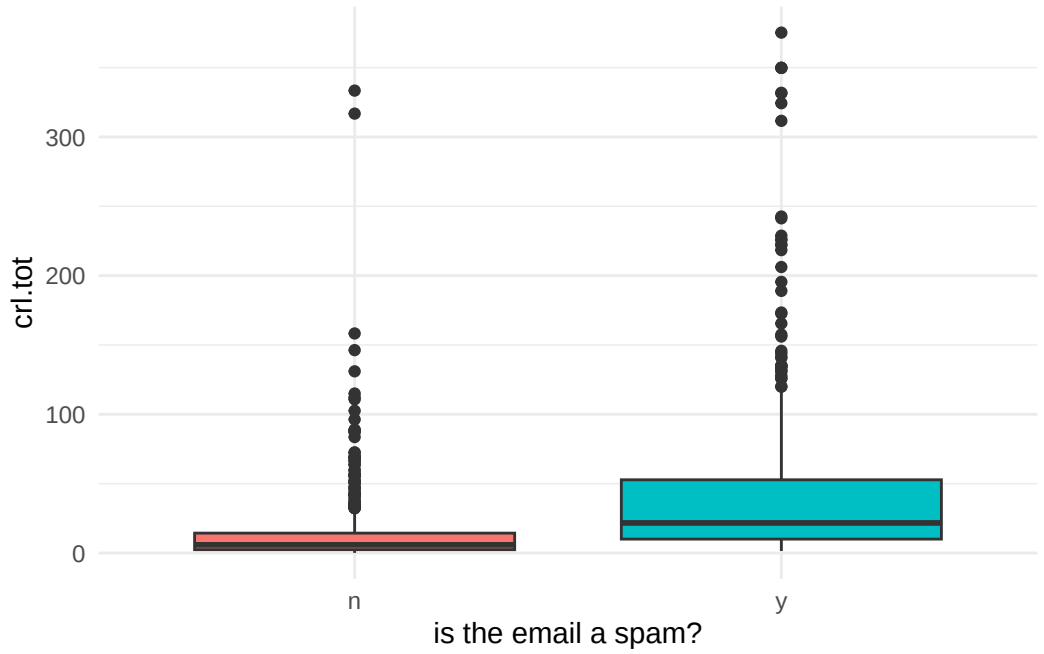

Figure 1: Total length of uninterrupted sequences of capitals in an email

The boxplot shows that, on average, there are more uninterrupted sequences of capitals in a spam than in a normal email.

3.2 Occurrences of the dollar sign

```
ggplot(data=d25.spam,aes(x=yesno,y=dollar,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

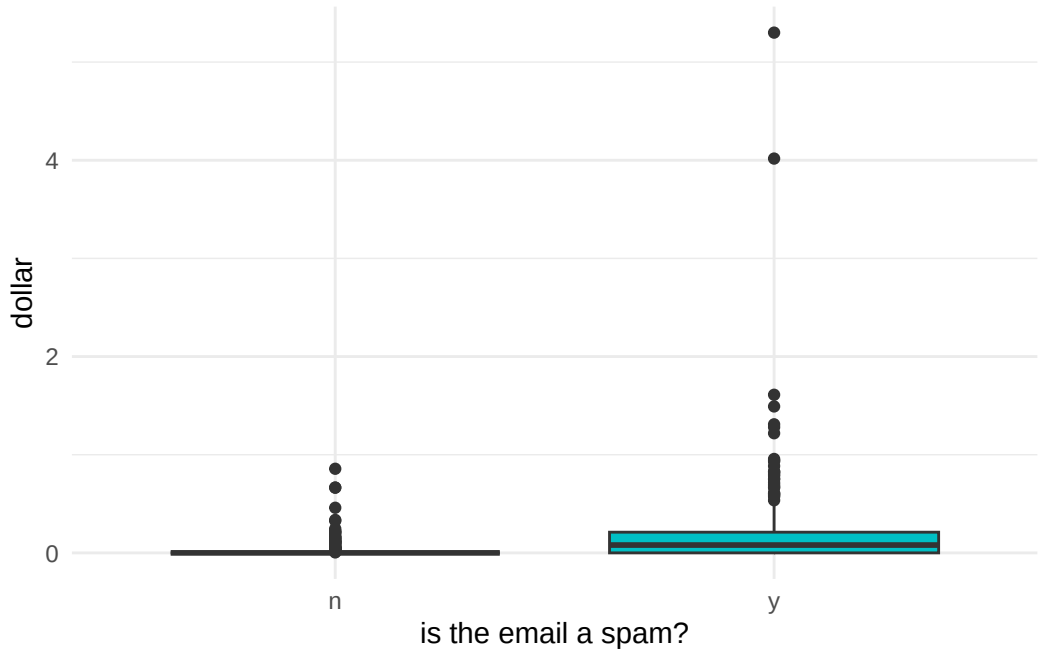

Figure 2: Occurrences of the dollar sign in an email

This graph shows that, on average, dollar sign appears more frequently in a spam.

3.3 Occurrences of '!'

```
ggplot(data=d25.spam,aes(x=yesno,y=bang,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

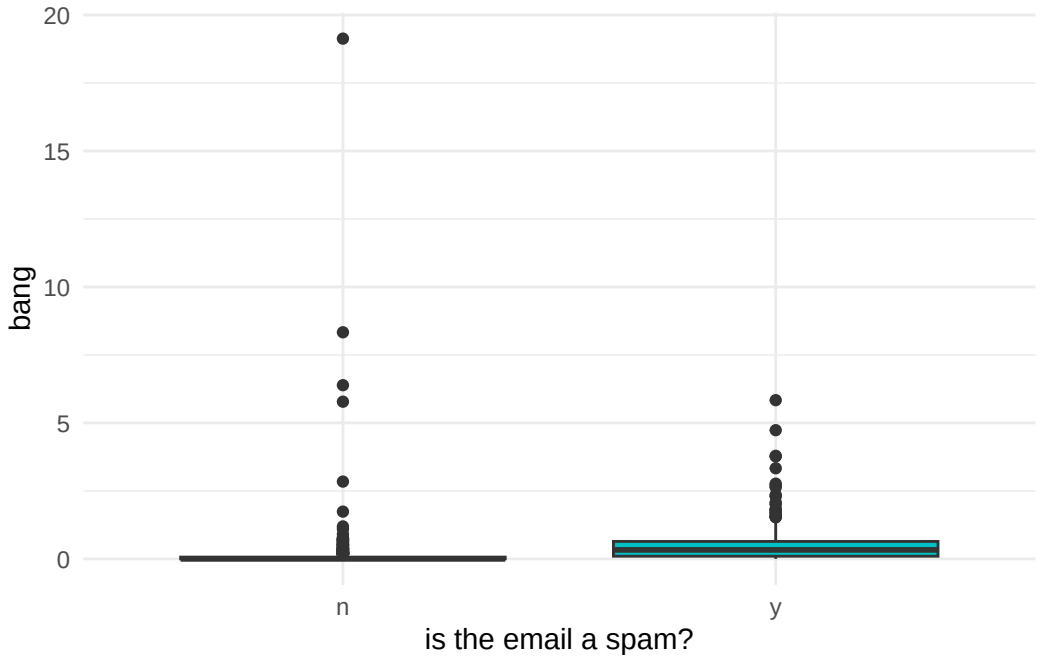

Figure 3: Occurrences of '!' in an email

This boxplot shows that, on average, exclamation mark tend to occur more in a spam.

3.4 Occurrences of 'money'

```
ggplot(data=d25.spam,aes(x=yesno,y=money,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

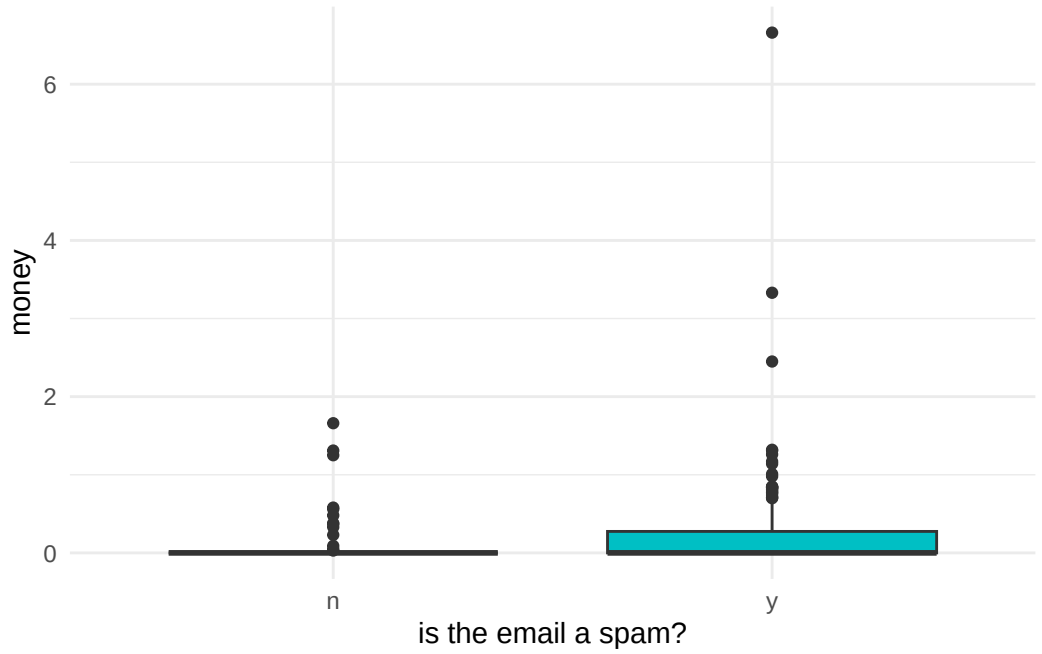

Figure 4: Occurrences of 'money' in an email

The graph shows that, on average, 'money' appears more frequently in a spam.

3.5 Occurrences of the string '000'

```
ggplot(data=d25.spam,aes(x=yesno,y=n000,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

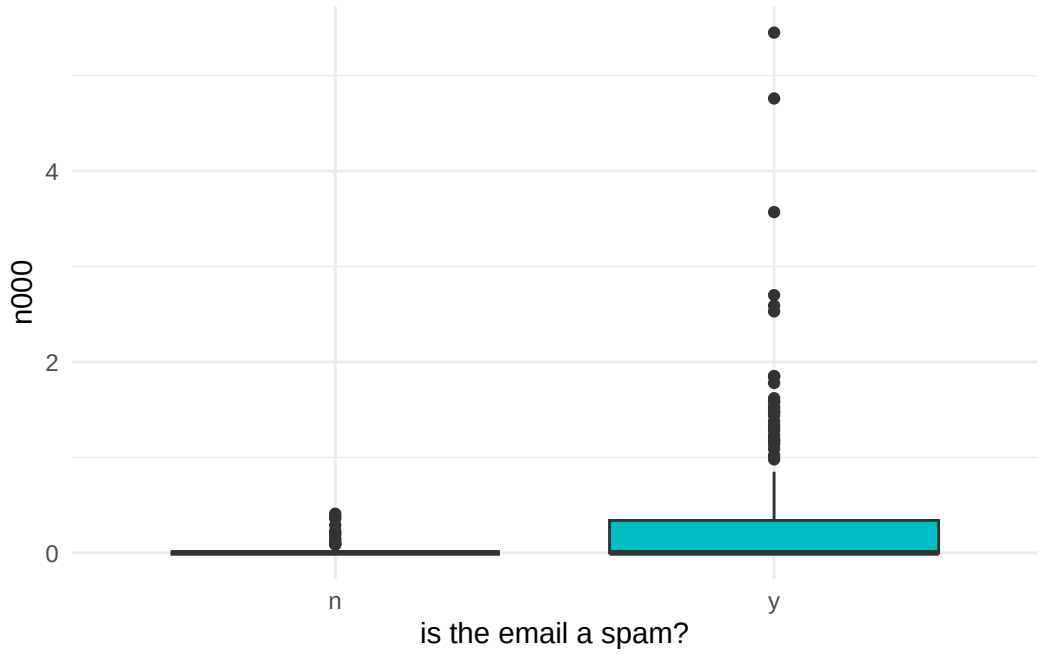

Figure 5: Occurrences of the string '000' in an email

The boxplot shows that, on average, the string '000' is more likely to occur in a spam.

3.6 Occurrences of 'make'

```
ggplot(data=d25.spam,aes(x=yesno,y=make,fill=yesno))+
  geom_boxplot()+
  labs(x="is the email a spam?")+
  theme_minimal()+
  theme(legend.position="none")
```

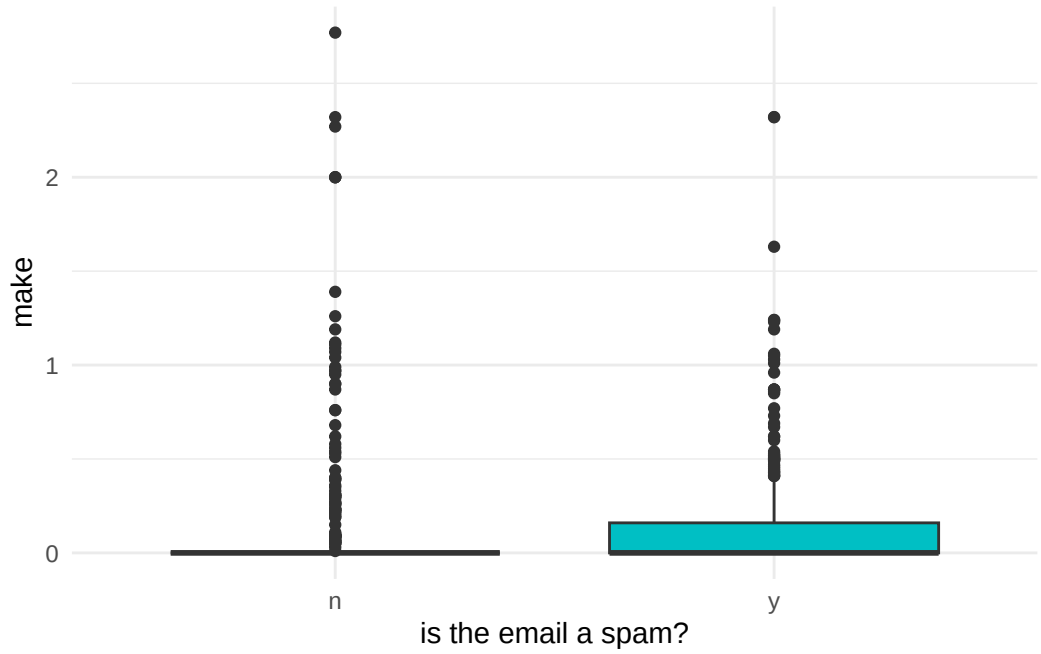

Figure 6: Occurrences of 'make' in an email

This graph shows that, on average, the occurences of 'make' in s spam is slightly more than in a noraml email.

4 Regression Results of the Data by using Generalized Linear Models

```
model.spam <- glm(yesno ~ crl.tot + dollar + bang + money + n000 + make, data = d25.spam, family = binomial(link = "logit"))

model.spam %>%
  summary()
```

```
Call:
glm(formula = yesno ~ crl.tot + dollar + bang + money + n000 +
    make, family = binomial(link = "logit"), data = d25.spam)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.711204   0.122790 -13.936  < 2e-16 ***
crl.tot      0.013173   0.002912   4.523 6.09e-06 ***
dollar       6.960462   1.211991   5.743 9.30e-09 ***
bang         0.730175   0.182056   4.011 6.05e-05 ***
money        3.518528   0.694623   5.065 4.08e-07 ***
n000         5.505005   1.201289   4.583 4.59e-06 ***
make         0.013481   0.308039   0.044    0.965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1241.77  on 920  degrees of freedom
Residual deviance:  809.54  on 914  degrees of freedom
AIC: 823.54

Number of Fisher Scoring iterations: 7
```

The coefficients of six characteristics are all positive, suggesting that spam tends to have more of these text characteristics. All the coefficients of the characteristics, except 'make', are significant because of the low p-values. Then we can obtain the effect of these characteristics by looking at the exponential values.

5 Residuals Analysis

```r
d25.spam$residuals <- residuals(model.spam, type = "deviance")
```

```r
ggplot(d25.spam, aes(x = residuals)) +
  geom_histogram(binwidth = 0.5, fill = "blue", alpha = 0.6) +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```

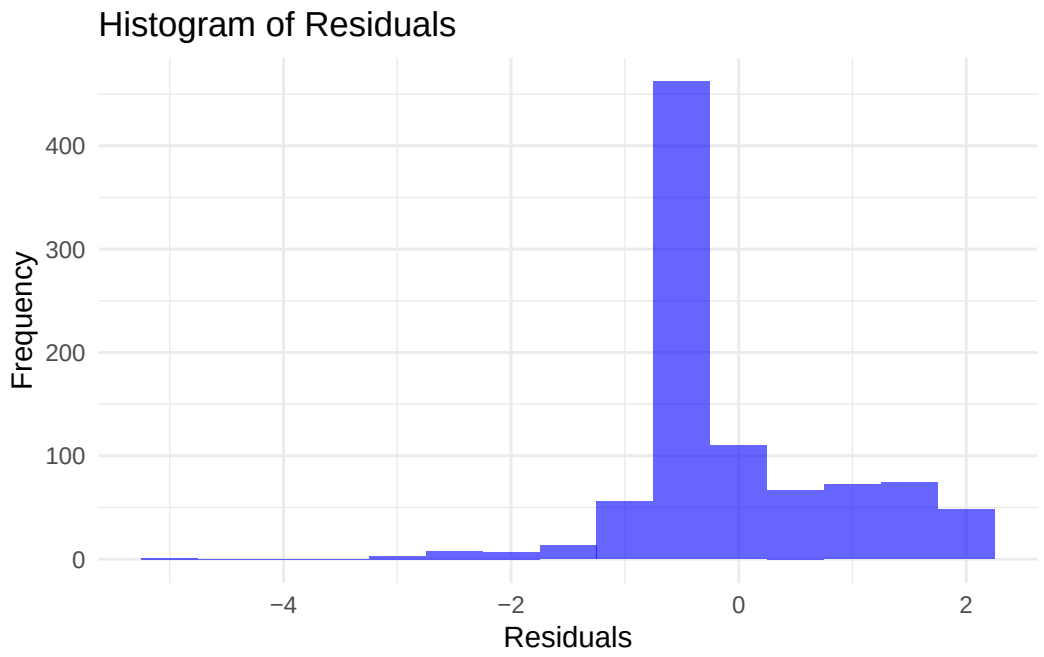

Figure 7: Histogram of Residuals

This residual analysis plot shows the residual is skewed distribution, it may not completely follow the normal distribution. This means that the model has systematic errors in some intervals.

```r
ggplot(d25.spam, aes(x = fitted(model.spam), y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
```

```
  labs(title = "Fitted value vs. Residuals", x = "Fitted values", y = "Residuals") +
  theme_minimal()
```

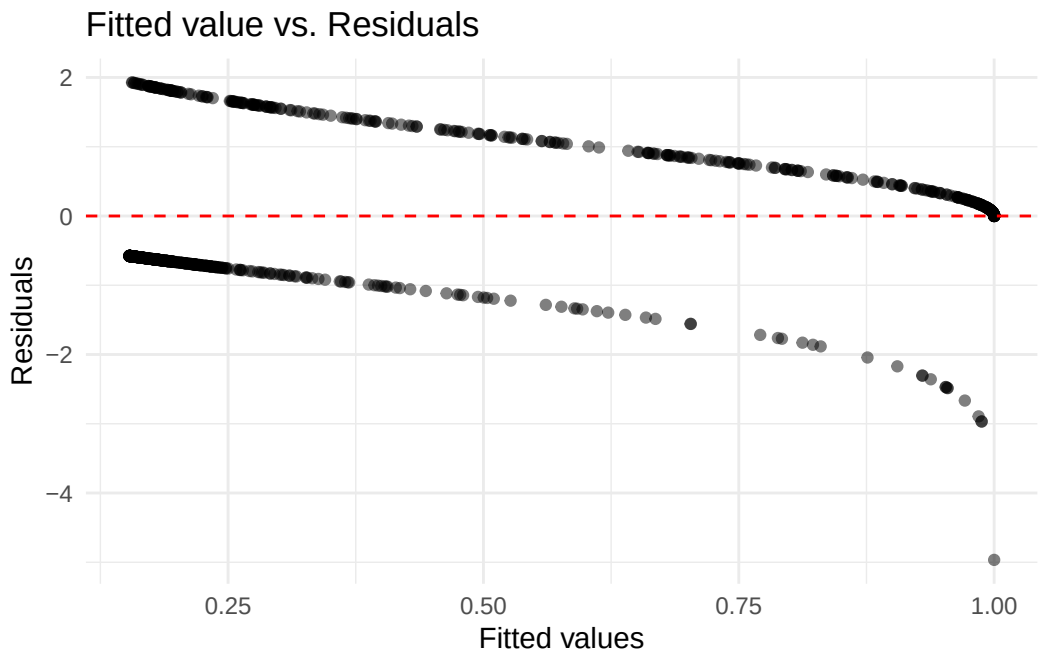

Figure 8: Residuals vs Fitted Values

This plot indicates that the residual are not randomly distributed, but show a systematic trend. Meanwhile, this plot means that the model has larger prediction errors for certain ranges, which may affect the reliability of the confidence interval.

```
ggplot(data=d25.spam,aes(sample=residuals))+
  stat_qq()+
  stat_qq_line(color="red",lwd=1)+
  labs(title="Q-Q Plot of Residuals", x="Theoretical Quantiles", y=" ")+
  theme_minimal()
```

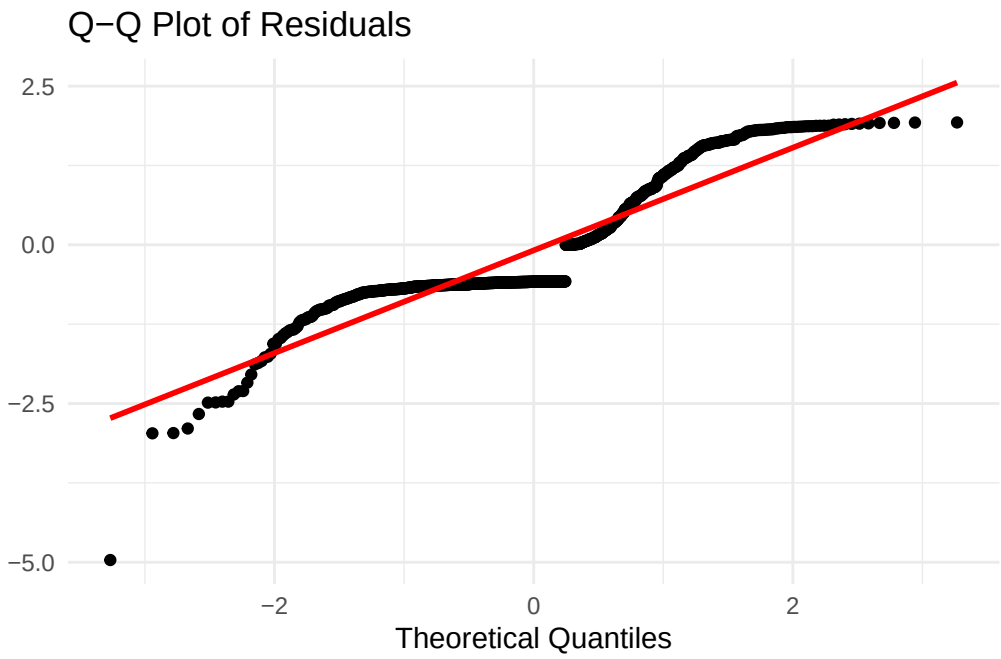

Figure 9: Q-Q Plot of Residualss

From the Q-Q Plot, residuals deviate from the reference line at most of the points, indicating deviations from normality.

6 Data Summary

```
plot_model(model.spam, show.values = TRUE, title = "Odds", show.p = FALSE, value.offset = 0.25)+
  theme_minimal()
```

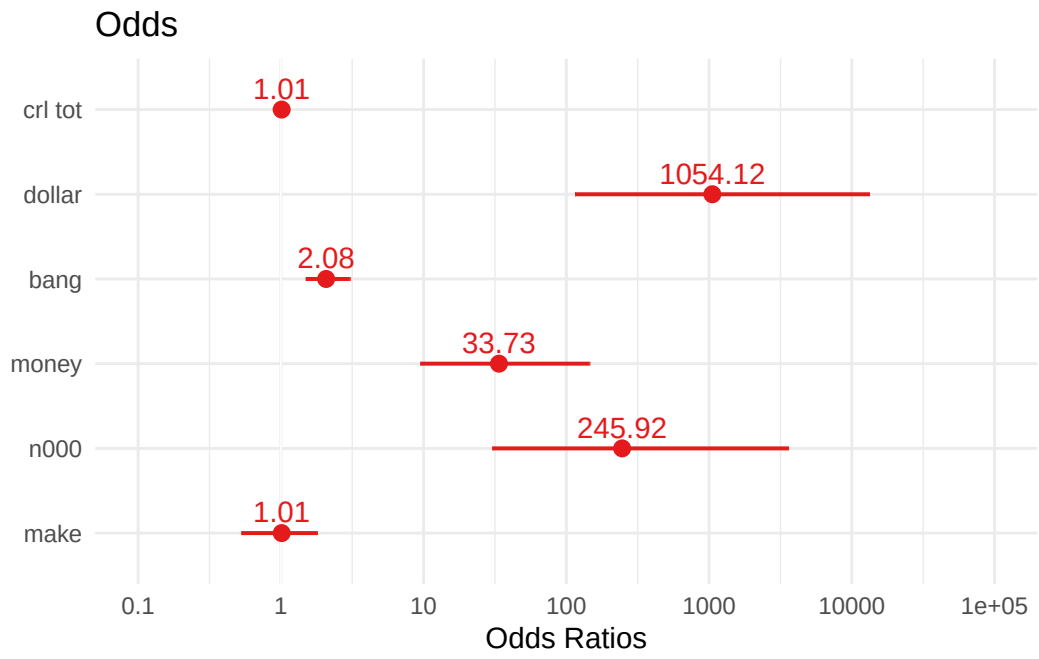

Figure 10: Odds of classifying emails as spam

According to the graph, when we look at a ten-characters length difference of uninterrupted sequences of capitals between two emails, the one having more uninterrupted sequences of capitals is 1.01 times more likely to be classified as spam than the one with less uninterrupted sequences of capitals. Also, with one unit increase in the occurrences of dollar sign, exclamation mark, character 'money', string '000' and character 'make', the higher one's odds of be classified as spam are 1054.12 times, 2.08 times, 33.73 times, 245.92 times and 1.01 times than those of the lower one respectively.

7 Conclusions

This analysis indicates that the occurrences of dollar sign in emails make them easiest to be classified as spam. The presence of the string '000' is also a characteristics that makes an email be identified as spam. Compared with them, the occurrences of 'money', bang

and ‘make’ seem to be not important in classifying emails as spam. The study also demonstrates that, with the increase of the length of uninterrupted sequences of capitals, emails are more likly to be identified as spam.