# Load packages

library(ggplot2)

library(dplyr)

library(openintro)

# Compute summary statistics

email %>%

  group\_by(spam) %>%

  summarize(median(num\_char) , IQR(num\_char))

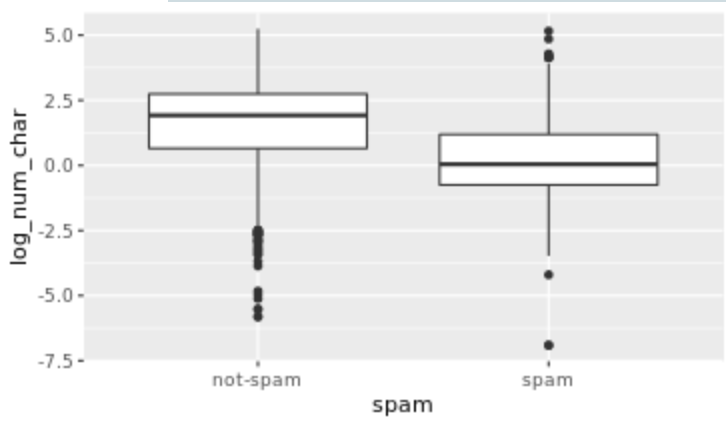
# Create plot

email %>%

  mutate(log\_num\_char = log(num\_char)) %>%

  ggplot(aes(x = spam, y = log\_num\_char)) +

  geom\_boxplot()



The median length of not-spam emails is greater than that of spam emails.

# Compute center and spread for exclaim\_mess by spam

email %>%

 group\_by(spam) %>%

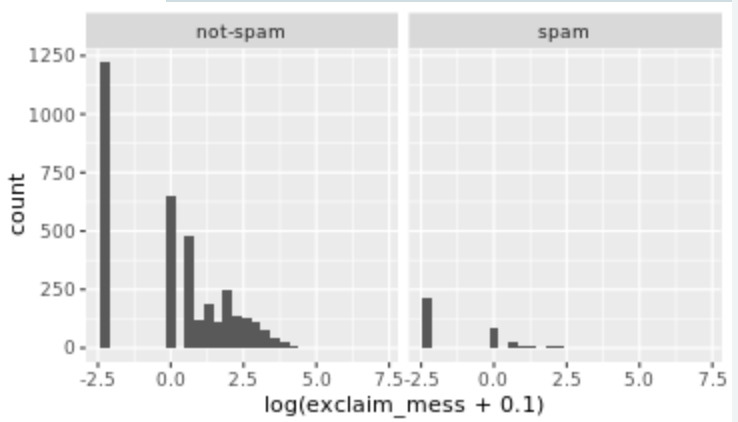
 summarize(median(exclaim\_mess) , IQR(exclaim\_mess))

# Create plot for spam and exclaim\_mess

ggplot(email, aes(x= log(exclaim\_mess+ 0.1))) +

geom\_histogram(position = "dodge") +

facet\_wrap(~spam)



There are more cases of spam in this dataset than not-spam.

# Create plot of proportion of spam by image

email %>%

  mutate(has\_image = image > 0) %>%

  ggplot(aes(x = has\_image, fill = spam)) +

  geom\_bar(position = "fill")



An email without an image is more likely to be not-spam than spam.

# Test if images count as attachments

sum(email$attach < email$image)

Output: 0 Hence, images are counted as attachments

# If you encounter an email with greater than 10 occurrences of the word "dollar", is it more likely to be spam or not-spam?

email %>%

  filter(dollar > 0) %>%

  group\_by(spam) %>%

  summarize(median(dollar))

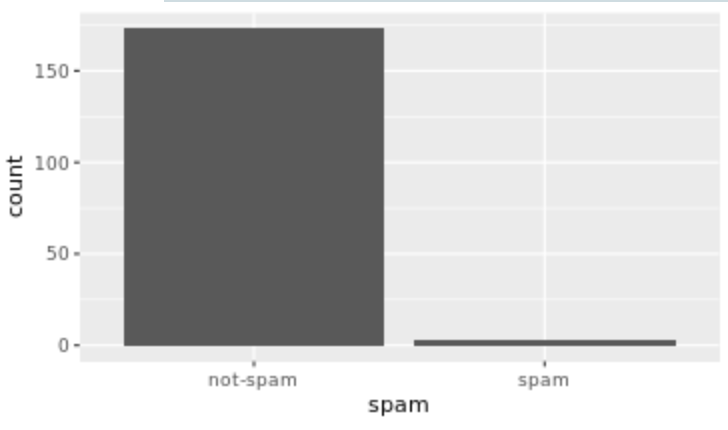
# Question 2

email %>%

  filter(dollar > 10) %>%

  ggplot(aes(x = spam)) +

  geom\_bar(position = "dodge")



# Reorder levels

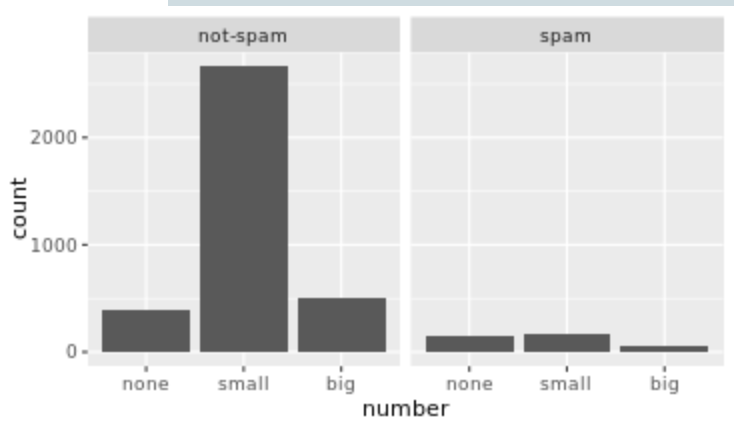
email$number <- factor(email$number, levels = c("none","small", "big"))

# Construct plot of number

ggplot(email, aes(x = number)) +

  geom\_bar() +

  facet\_wrap(~spam)



Given that an email contains no number, it is more likely to be spam.