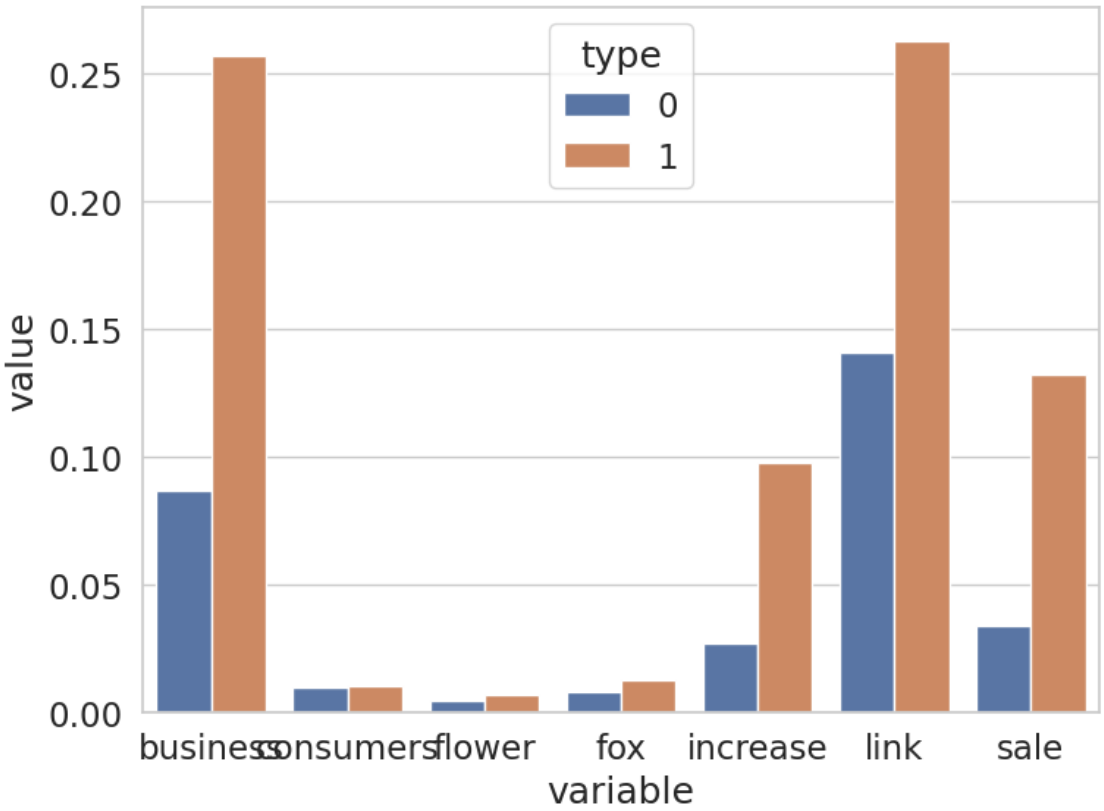
0.1 Question 1

Discuss one attribute or characteristic you notice that is different between the two emails that might relate to the identification of a spam email.

One characteristic that we may be able to use to determine if an email is spam or not is generality. Spam emails will never be personalized as they are meant to be sent out in mass quantities to varies people, therefore we can look out for emails that do not use anything that may pertain to the receiver specifically such as a name or address.

Create your bar chart with the following cell:

```
In [17]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))
         sns.barplot(data = temp, x = 'variable' , y = 'value', hue = 'type')
         plt.tight_layout()
         plt.show()
```



0.2 Question 6c

Explain your results in Q6a and Q6b. How did you know what to assign to `zero_predictor_fp`, `zero_predictor_fn`, `zero_predictor_acc`, and `zero_predictor_recall`?

the zero predictor is a function that classifies all mail as ham, that explains why we got such a high number in zero_predictor_fn and why we calculated 0 for zero_predictor_fp. For 6b. zero_predictor_acc is the proportion of emails that we were able to correctly classify, in our case it was 74%. In regards to the zero_predictor_recall since we know that recall is calculated using TP/TP+ FN we know that recall has to be 0 since in TP = 0 because of how we throw everything into inbox regardless of whether or not its spam.

0.3 Question 6e

Is the number of false positives produced by the logistic regression classifier greater than the number of false negatives produced?

There are much more false negatives than there are false positives.

0.4 Question 6f

How does the accuracy of the logistic regression classifier (computed in Question 5) compare to the accuracy of the zero predictor (computed in Question 6b)?

My zero predictor model got an accuracy of about 75%, the logistic regression model in comparison has a slightly worse accuracy at about 74%

0.5 Question 6g

Given the word features provided in Question 4, discuss why the logistic regression classifier may be performing poorly.

Hint: Think about how prevalent these words are in the email set.

The word features that were given were fox,business,link,sale,increase,consumers,flower. Words such as business are very common in both spam and ham emails thus causing the logistic regression to potential perform worse.

0.6 Question 6h

Would you prefer to use the logistic regression classifier or the zero predictor classifier for a spam filter? Why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

Despite it performing slightly worse I would perfer to have the logistic regression model as the zero predictor would mark everything as ham. This in turn nullifies the entire point of the classifier and a 74% accuracy is still acceptable for use.