M358K - Homework 1

posted on: September 10th, 2018

due: September 24th, 2018

Emails: which variables are useful to distinguish spam?

In this homework, you will explore the dataset `emails` to help answering the question: which variables are useful to distinguish spam vs regular emails?
The Emails dataset. `data/emails.csv` on Canvas.
Variable description. `data/emails-descrip.txt` on Canvas.
Number of questions in this homework: 3.
Maximal points possible: 6 from writeup, 2 from code, 4 from presentation. This gives a total of **10 points**.

Question 1

`spam` vs `exclaim_mess`: descriptive analysis with numerical variables

1. Make a dotchart, a boxplot, a histogram and a violin plot of `spam` vs `exclaim_mess`.

2. Which of the above plots are useful for describing the relationship between these two variables? What do those plots convey? Why are the other plots not as useful?

3. Summarize the relationship between `spam` and `exclaim_mess` in a couple of sentences.

Question 2

Sometimes it is useful to recode variables. `spam` vs `exclaim_mess` is an example.

1. Recode `exclaim_mess` into four values: $0, 1, 2, >= 3$. Call this new variable `exclaim_mess.recode`. What is the type of this new variable?

2. Produce a table and a mosaic plot of `spam` vs `exclaim_mess.recode`. What do they reveal?

3. Summarize the relationship between `spam` and `exclaim_mess.recode` in a couple of sentences.

4. Why is it reasonable to recode `exclaim_mess`?

5. How would your summary on the relation between `spam` and `exclaim_mess` change if you had recoded it into 5 values? 10 values? 3 values? Which regroup is most reasonable, and why?

Question 3.

1. Run a descriptive analysis for `spam` vs X for each of the 20 variable X in the dataset. For each analysis, include ONE plot and/or ONE table that is most informative, and write a short sentence summarizing the relationship between `spam` and X.

2. Based on your findings, give a list of variables that you think should be included for further analysis.