M358K - Homework 1

posted on: September 12th, 2017

due: September 26th, 2017

Instructions and grading scheme

Submission instructions. Please submit three files on Canvas with the following names:

- **homework1-writeup**: this is a word file containing your answers. All figures, tables etc must be inside this file! For how to save R figures, see instructions on Canvas.
- **homework1-code-final.R**: this is a text file containing the R codes you used to compute the numbers, tables, and figures CONTAINED IN your report ONLY. This file IS graded, see below.
- **homework1-code-draft.R**: this is a text file containing the R codes you used to explore the data. This file should give an idea on how you explored the dataset, and how you arrived at code-final.R. It can tables/figures which are EXCLUDED from your report, comments, etc. For instance, the R history of your various sessions is enough.

Acceptable word files are: anything that can be opened by LibreOffice, OpenOffice, Microsoft Word, or equivalent software. Example file extensions are .doc, .docx, .odt.

Grading scheme. For the write up: on each question you can earn 0/1/2 points.

- 2 = correct answer

- 1 = partially correct answer, or correct answer but with muddy or missing justification
- 0 = incorrect answer, unreasonable answer

For code-final.R: you can earn 0/5 points.

- 5 = code-final.R runs on the given dataset, gives all the tables and figures included in your report.
- 0 = no code file OR code file does not run at all OR code file does not produce the reports' tables/figures OR code file raises a plagiarism red flag, etc

For code-draft.R: you can earn 0/5 points.

- 5 = The grader is sufficiently convinced that you explored this dataset on your own.
- 0 = no code file OR code file raises a plagiarism red flag OR code file is irrelvant, does not give an indication on how you arrived at code-final.R

Bonus points for presentation:
Write-up: +2 points for nice report(grammatically correct sentences, no rambling discussions, discussions exceed expectations, extra analysis of the dataset)
R code: +2 points for neat layout. (code adequately commented, clearly laid out, easy to understand).

Bonus questions
These are extra challenging questions, and are additional opportunities to score points. On each bonus question you can get 0/4 points.

Questions

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.
Maximal score on this set (excluding bonuses): 22
Maximal score on this set (including bonuses): 34

1. `spam` vs `format`: descriptive analysis with categorical variables
   (a) Produce a mosaic plot of `spam` vs `format`. What does this mosaic plot reveal?
   (b) Produce a table of `spam` vs `format`. What fraction of spam are formatted? What fraction of non-spam are formatted? What fraction of formatted emails are spam? What fraction of plain text emails are spam?
   (c) Summarize the relationship between `spam` and `format` in a couple of sentences.
2. Sometimes it is useful to recode variables. `spam` vs `exclaim_mess` is an example.
   (a) Make a dotchart, a boxplot, a histogram and a violin plot of `spam` vs `exclaim_mess`.
   (b) Which of the above plots are useful for describing the distribution of this variable? What do those plots convey? Why are the other plots not as useful? Summarize the relationship between `spam` and `exclaim_mess` in a couple of sentences.
   (c) Recode `exclaim_mess` into 4 values: $0, 1, 2, >= 3$. Call this new variable `exclaim_mess.recode`. What is the type of this new variable?
   (d) Produce a table and a mosaic plot of `spam` vs `exclaim_mess.recode`. What do they reveal?

(e) Summarize the relationship between `spam` and `exclaim_mess.recode` in a couple of sentences.

(f) Why is it reasonable to recode `exclaim_mess`?

(Bonus question) 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?

3. `spam` vs `num_char`: descriptive analysis with numerical variables

   (a) Make a dotchart, boxplot and violin plot of `spam` vs `num_char`.

   (b) Which of the above plots are useful for describing the distribution of this variable? What do those plots convey? Summarize the distribution of `num_char` in a couple of sentences.

(Bonus question) Run a descriptive analysis for `spam` vs X for each of the 20 variable X in the dataset. For each analysis, include one plot or one table that is most informative, write a short sentence summarizing the relationship between `spam` and X, and say "yes" or "no" to the question: Should this variable be included for further analysis?