AMOD 5240H – Term Project

Nov. 13, 2018

Due: Monday, Dec. 10 at 10 pm

The objective of this assignment is to allow you to explore a data set using the techniques we learned in the course.

A. Data

Choose a data set that you want to analyse. There are two possibilities:

1. Pick an already existing data set in an area of your interest. Here are some suggestions on websites where you can find data.

https://simplystatistics.org/2018/01/22/the-dslabs-package-provides-datasets-for-teaching-data-science/

https://www.kaggle.com/datasets

https://data.gov.uk/

https://github.com/caesar0301/awesome-public-datasets

https://www.toronto.ca/city-government/data-research-maps/open-data/

https://www.data.gov/finance/

https://open.canada.ca/data/en/dataset?q=financial+performance+data

https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/finance/

2. Create your own data set by collecting data in an area of your interest (heights of mountains to test for fit to Benford's law), or taking a sample (a sample of cars in the Trent parking lot to test whether colours are distributed uniformly), or completing an experiment. If you choose to go in this direction, make it as random as possible. However, it is OK if it is not totally random and you take this opportunity to think about the difficulties of obtaining random samples. If sampling, keep in mind that systematic sampling is very close to random sampling if no intrinsic order is found in the population. You wouldn't need to draw random numbers for the cars to visit in the parking lot, it would be good enough for this practical exercise to pick a parking lot at random, find a starting point at random and choose every kth (k chosen at random) car to measure.
   If you choose to sample or perform an experiment, I suggest that you limit your attention to non-human samples. Surveys and experiments involving humans are prone to many non-sampling errors that are harder to control.

B. Content

Your document must include the following sections:

1. **Description of the data set and provenance (10 points)**
   Describe the data set that you picked, its provenance and its structure. If you collected your own data, describe how you did it.

2. **Appropriateness for statistical analysis (10 points)**

Have the data been collected so that they are fit for statistical inference? Discuss. (If you conducted an experiment or took a sample, describe the flaws of your sampling method here. If you are using an existing data set, do you think it is a random sample of a population? If so, which one? What generalizations would you be justified in making?)

3. **Two Questions (10 points)**

   After observing the structure of your data, but before plotting it or summarizing it, choose two questions that you think you would like to answer with these data, and what technique(s) that you learned this term (or during your term project) would help you answer those questions. The questions do not have to be in statistical language – that comes later.
   Now pick **one** of the two questions as the one you are going to explore in the rest of the document.

4. **Data summary using R (15 points)**
   Summarize your data with numerical measurements and plots. Subset the data set to explore only the portions that relate to your question.

5. **Data analysis using R (35 points)**
   Analyse your data set using appropriate statistics and hypotheses. Assume that all assumptions for a statistical analysis are valid (even if in 2 you said they were not).

6. **Conclusion (10 points)**
   What can you conclude from your analysis? Were you able to answer your original question? If not, why not?

7. **References**
   Make sure that every source that you use is properly referenced.

C. Format (10 points)

A document that is as clean as possible of spelling mistakes, that is nicely organized, and that looks professional should be uploaded to blackboard as a pdf (a knit version of an Rmarkdown file) or a word document. Your document **should not include** the list of values from the data set, especially if that list is long. That is why you are being asked to summarize it. You do not have to include code for data summaries, but I would like to see the code for the test itself, so that I can check that it is done correctly.