General Overview

About Data Analysis Assignments

Data analysis assignments will consist of a project where you will obtain, clean, explore, analyze, and write-up a brief data analysis of an assigned data set. For each of the two data analysis assignments you will submit four parts (all are required).

1. A write-up of your data analysis explaining what question you were addressing, the data you used, how you applied statistical methods, and what the conclusions of your analysis were (less than 1,500 words)
2. One figure and a corresponding figure caption (less than 500 words)
3. A list of references for the statistical methods that you used. They should be referenced from the write-up in item 1.
4. An R script that reproduces the numbers you report in item 1 and the figure in item 2.

Data Analysis Assignments Grading

Your data analysis assignments will be graded by multiple other students in the class. Each student will assign a score using the rubric described below. Your score will be the median percent from the peer assessments, multiplied by 40.

The Data Analysis Rubric

You will assign a numerical score between 0 and 5 for each of the questions in the following rubric.

Item 1 - Write-up

- Does the analysis have an introduction, methods, analysis, and conclusions?
- Are figures labeled and referred to by number in the text?
- Is the analysis written in clear and understandable English?
- Are the names of variables reported in plain language, rather than in coded names?
- Does the analysis report the number of samples?
- Does the analysis report any missing data or other unusual features?
- Does the analysis include a discussion of potential confounders?
- Are the statistical models appropriately applied?
- Are estimates reported with appropriate units and measures of uncertainty?
- Are estimators/predictions appropriately interpreted?
- Does the analysis make concrete conclusions?
- Does the analysis specify potential problems with the conclusions?

Item 2 - Figure and caption

- Is the figure caption descriptive enough to stand alone?
- Does the figure focus on a key issue in the processing/modeling of the data?
- Are axes labeled and are the labels large enough to read?

Item 3 - References

- Does the analysis include references for the statistical methods used?

Item 4 - R script

- Can the analysis be reproduced with the code provided?

Data Analysis assignment 1

Data

For this analysis you will use the loans data available from here:

https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv
https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda

There is a code book for the variables in the data set available here:

https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf

Prompt

The data above consist of a sample of 2,500 peer-to-peer loans issued through the Lending Club (https://www.lendingclub.com/home.action). The interest rate of these loans is determined by the Lending Club on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores.

The purpose of your analysis is to identify and quantify associations between the interest rate of the loan and the other variables in the data set. In particular, you should consider whether any of these variables have an important association with interest rate after taking into account the applicant's FICO score. For example, if two people have the same FICO score, can the other variables explain a difference in interest rate between them?

What you should submit

Your data analysis submission will consist of the following components:

1. The main text of your document including a numbered list of references. This can be uploaded either as a pdf document or typed into the text box (not both!). The limit for the text and references is 2000 words. Your main text should be written in the form of an essay with an introduction, methods, results, and conclusions section.
2. One figure for your data analysis uploaded as a .png, .jpg, or .pdf file, along with a figure caption of up to 500 words.

Reproducibility

Due to security concerns with the exchange of R code, you will no longer be asked to submit code to reproduce your analyses. I still believe reproducibility is a key component of data analysis and I encourage you to create reproducible code for your data analysis.

Submission Deadline

You must submit your data analysis by February 18th, 2013 at 7:00AM UTC-5:00 (Baltimore time). No late days may be applied to the data analysis. Note that this is an extension of the original date posted on the class website.