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.