UPSTART PRE-INTERVIEW QUESTION

In the CSV file that accompanies this document are a set of loans issued between October 2014 and September 2015. The performance of these loans has been observed through October 2015.

As of October 2015, each loan was marked (in the column *loan_status*) as "Current", "Fully Paid", or "Charged Off". "Current" means that the borrower is either on or ahead of schedule in terms of paying off his or her loan. "Fully Paid" means that the loan has been entirely paid off. "Charged Off" means that the borrower has stopped paying and defaulted on the loan.

In addition to loan status, other variables included in the dataset are a unique member id, loan amount, interest rate, the date the loan was issued, lower and upper endpoints of the range of the borrower's FICO score, annual income, and the date of the last payment that was made by the borrower.

Please answer the following questions. Along with your answers, feel free to send over any code you used to obtain them. (Showing your work can get you partial credit!)

(1) What is the relationship between FICO score and interest rate?
(2) What is the relationship between loan size and interest rate?
(3) Does it seem like the size of a loan and/or the borrower's FICO score are related to whether or not the loan ends up being charged off? How might this relate to your answer to the previous question?
(4) What percentage of loans have been charged off as of October 2015?
(5) Some of the loans which are marked "Current" in October 2015 will end up being charged off before they are paid in full. Using the fact that all of these loans are 3 years in length, estimate the probability that a randomly chosen loan would eventually become charged off, possibly after October 2015. Be sure to explain any assumptions you make.

P.S. We want you to spend a few hours on this **not** a few days! (But feel free to comment on any other interesting patterns or relationships in the data.)