Assignment 2

This assignment concerns a set of data from a crowd-sourced lending service, in loans-subset.csv. It has attributes of 5425 loans that were "Charged Off" (not paid in full) and an equal number that were paid in full. This information is in the final columns, Loans.loan_status. There are 10 additional variables:

Funded_amnt: the amount of money lent.
Loan_amnt: the amount of money requested.
Dti: debt to income ratio, excluding mortgage and proposed loan.
Emp_length: the number of years the borrower has been employed.
Installment: the monthly payment.
Annual_inc: borrower's annual income.
Revol_bal: the balance on all the borrower's revolving credit accounts.
Earlyear: the year in which the borrower first borrowed money.
Proputil: the proportion of the borrower's maximum revolving credit being utilized (this is actually given as a percentage, a number between 0 and 100).
Open_acc: the number of credit accounts the borrower has opened over the years.

Note: you should include all relevant code. Plots should be appropriately labeled.

1. (Justification 2 marks, screeplot 2 marks, score plots 3 marks, comments 2 marks)Perform principal components analysis on the numerical data. Justify your choice of covariance or correlation matrix. Include the screeplot; choose a number of dimensions and justify your choice. Make appropriate plots of the scores, using color or plotting symbol to indicate which points are "charged off" or paid in full. Comment on patterns or lack thereof.
2. (3 marks) Scale the chosen principal components (as in lab 3) and give the loadings of your chosen components. Use rounding/suppression to make this nice to look at.
3. (2 marks for loadings, 1 for comment on interpretability, 2 for names) Perform varimax rotation and give the new loadings. Are they more interpretable? Do your best at giving a name to each of your rotated components.
4. (2 marks for the table—leaving it as R output is OK, 2 for the comment) Find the linear discriminant function. Make a table of the cross-validation predicted classes vs the true classes. Comment on the level of accuracy.
5. (4 marks for the plots and summaries, 1 for each assumption commented on) Make suitable plots and summaries comparing the discriminant function scores for each group. Is the normal assumption reasonable? What about equal variance?
6. (2 for the loadings, 2 for the comments) Compute the correlations of the original variables with the discriminant function scores (loadings). Comment on which variables are most important.
7. (3 marks for the computations, 2 for comments) Compute the posterior probability of each group for the five new data points in loanprecition.csv. Do not refit the discriminant function, but use the fact that rather than being a 50/50 split between defaulters and payers, in fact only 14% of loans are charged off. Is there anyone that is clearly a high risk? Explain based on the probabilities you computed.

8. (2 marks) One variant of quadratic discriminant analysis allows you to use the t distribution rather than the normal distribution. This is more robust to unusual values. What are your thoughts about how helpful it would be here?
9. (5 marks for code and plots/tables, 2 for comments.) Try it (look at the qda help; you will use the method="t" argument), in such a way that it is suitable to compare with your output from question ~~3~~ 4. Make suitable plots to show how the posterior probabilities for each observation are similar/different (and whether they recapture the true classification). Comment on which method you prefer.

45 marks total, will be converted to a percentage.