The given dataset is from Dream Club which connects borrowers with investors. We will use lending data from 2007-2010 and build a classifier model to predict whether or not the borrower has paid back their loan in full.

Dataset: [loan\_data.csv](https://bits-pilani.instructure.com/courses/1203/files/248044?wrap=1)[Download loan\_data.csv](https://bits-pilani.instructure.com/courses/1203/files/248044/download?download_frd=1)

Here are what the columns represent:

1. credit.policy: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
2. purpose: The purpose of the loan (takes values "credit\_card", "debt\_consolidation", "educational", "major\_purchase", "small\_business", and "all\_other").
3. int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
4. installment: The monthly installments owed by the borrower if the loan is funded.
5. log.annual.inc: The natural log of the self-reported annual income of the borrower.
6. dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
7. fico: The FICO credit score of the borrower.
8. days.with.cr.line: The number of days the borrower has had a credit line.
9. revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
10. revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
11. inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.
12. delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
13. pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).
14. not.fully.paid: whether or not the borrower paid back their loan in full

**QUESTIONS:**

Q-1: Load the dataset and print the metadata in the notebook. - 1M

Q-2: Create a countplot using sns showing the counts of loans by **purpose**,  with the color hue defined by **not.fully.paid**. Mention the insights from the same. - 1M

Q-3: Print a heatmap to check NULL values and Correlation values. - 1M

Q-4: Perform stratified splitting of train and test data to ensure similar class distribution. 1M

Q-5: Build a classifier model using the Decision Tree algorithm and plot the same. - (1+1)M

Q-5: Print confusion matrix and classification report before and after pruning the Decision tree. Write down your observations on the effects of pruning. - (1+1)M

Q-7: Find out the stratified cross-validation accuracy 1M.