CS 457 - Homework Assignment 3: Exploratory Data Analysis

Due Date: Friday, February 05 at 11:59 pm

Purpose:
Demonstrate exploration of data via creation of statistical tables and visualizations using Python and tell interesting stories and insights around your analysis.

Points: 100

Part 1 (25 points)

Use `loan_small.csv` dataset

Below are some suggestions that might help you cleaning the data and make it more suitable for EDA

- Exclude "months" text in attribute term. For example, "36 months" can be replaced by 36
- emp_length can also be cleaned. Remove + and < symbols and try to make it a numerical attribute with your logic.
- loan_status can be converted into binary attribute such as "good" or "bad" based on given values. For example, "Fully paid" is considered as "good" loan status
- attributes such as mths_since_last_delinq and mths_since_last_record need imputation (fill in missing values)

Part 2 (75 points)

You do not need to use all the columns/attributes for Part 2. Use your imaginations to come up with interesting Univariate and Bivariate analysis.

- Generate appropriate summary (count, mean, median or mode) tables using group keyword in pandas.
    - Include at least two tables analysis or results
- Generate appropriate visualizations for Univariate analysis
    - At least one bar chart
    - At least one histogram
- Generate appropriate visualizations for Bivariate analysis
    - At least one scatter plot (continuous vs continuous)
    - At least one visualization for (discrete vs continuous)
    - One correlation plot
- Generate one Multivariate visualization (optional)

Include appropriate titles and labels for all the visualization and tables. **Interpret all the results. No points will be given without explanation.**

Deliverable: Submit a ipynb file containing your code, outputs and explanations. Include homework title, your name and your email on top of your ipynb code file.