# Capstone One – Phase I: Dataset Selection and Project Plan

For this capstone project, I selected a dataset titled 'Financial Risk Assessment', which contains 15,000 records and 20 well-labeled columns. The dataset includes a healthy mix of numerical variables such as income, credit score, loan amount, and debt-to-income ratio, alongside categorical variables like gender, education level, employment status, and risk rating. It is formatted cleanly in a CSV structure, with minimal missing data, making it suitable for analysis. The dataset meets all the specified criteria: more than 3,000 rows, over 7 columns, mixed data types, and a high usability score based on its clarity and structure.

The business implications of this dataset are significant. By analyzing the relationships between borrower characteristics and financial indicators, we can identify which traits are most predictive of lending decisions. For example, understanding how income, employment history, and credit score influence loan amounts can help financial institutions design fairer, more efficient underwriting policies. Strategically, the analysis can guide data-driven risk segmentation, enhance targeting for financial products, and inform customer-level financial planning tools or educational outreach.

For the core of my capstone, I plan to use multiple linear regression to model the loan amount a borrower receives based on features such as income, credit score, debt-to-income ratio, assets, and employment status. This project will allow me to explore the strength and interpretability of linear models in a real-world finance context. The goal is to determine which variables most strongly influence loan size, quantify their impact, and provide business insights for loan officers and credit policy teams. Residual analysis and regression diagnostics will be used to evaluate the model's assumptions and performance.

In future phases, I will begin with a thorough exploratory data analysis (EDA), identifying key trends, outliers, and data quality issues. This will be followed by feature selection and encoding of categorical variables to prepare the data for regression modeling. After building and validating the linear regression model, I will interpret coefficients and visualize key relationships using statistical and business dashboards. If time permits, I may explore logistic regression or decision trees as alternative modeling techniques for related classification tasks (e.g., predicting risk rating). The final deliverable will include a report, presentation, and potentially an interactive visualization dashboard to communicate findings effectively.