Project Title

Exploratory Analysis of Borrower Characteristics and Loan Conditions (EDA & SQL)

Project Plan

Team

Harmandeep Singh

Isis Hassan

1. Project Scope

This project focuses on exploratory data analysis (EDA) and SQL-based descriptive insights to identify patterns and relationships in loan data.

The analysis is observational and descriptive in nature. It does not include predictive modeling, classification, causal inference, or risk scoring.

2. Project Objective

The objective of this study is to explore relationships between **borrower characteristics** (credit score, annual income, and employment status) and **loan attributes** (interest rate, loan amount, and loan purpose) using exploratory data analysis and SQL-based queries.

3. Research Questions

RQ1: How does the credit score have influence on deciding Loan attributes?

a. *How does credit score relate to interest rate?*
b. *How does credit score relate to loan amount?*
c. *How does loan purpose vary across different credit score ranges?*

RQ2: Annual Income and Loan Amount Size

- *How does a borrower's annual income relate to the loan amount issued?*

RQ3: Employment Status and Loan Characteristics

- *How do loan amounts differ by employment status?*
- *How do interest rates vary across employment statuses?*
- *How does loan purpose distribution differ by employment status?*

4. Assumptions

- The dataset represents realistic, historical-like loan and borrower information.
- Interest rate and loan amount reflect the lender's assessment of borrower risk.
- The data is internally consistent and suitable for exploratory analysis.
- Missing or invalid values, if present, are handled appropriately.

5. Project Phases

Phase 1: Data Understanding & Cleaning
Deadline: 27.01.2026

- ☑ ~~Load the dataset~~
- ☑ ~~Initial inspection using head(), shape, info(), and describe()~~
- ☑ ~~Data Cleaning: our dataset is synthetic, but we will check anyway~~
    - ☑ ~~Check missing values, duplicates, and unique values~~
    - ☑ ~~Remove duplicate rows~~
    - ☑ ~~Fix incorrect data types~~
    - ☑ ~~Handle missing and invalid values~~
- ☑ ~~Drop irrelevant columns (id → index)~~
- ☑ ~~Save the cleaned dataset as a new CSV file~~

Phase 2: EDA- Univariate Analysis

To understand individual variable distributions, identify outliers, and detect category imbalances.

Deadline: 28.01.2026

- Analyze numerical variables using summary statistics, histograms, and boxplots.
- Analyze categorical variables using frequency counts and bar plots.
- Identify distributions, outliers, dominant categories, and imbalances.

- Univariate Analysis
    - ☑ ~~annual_income (borrower characteristic)~~
    - ☑ ~~credit_score (borrower characteristic)~~
    - ☑ ~~employment_status (borrower characteristic)~~
    - ☑ ~~loan_amount (loan characteristic)~~
    - ☑ ~~interest_rate (loan characteristic)~~
    - ☑ ~~loan_purpose (loan characteristic)~~

Phase 3: EDA- Bivariate Analysis

To examine relationships between borrower characteristics and loan attributes.

Deadline: 28.01.2026

- ☑ ~~Numerical vs Numerical analysis using scatter plots and correlation~~
- ☑ ~~Numerical vs Categorical analysis using boxplots and grouped means~~
- ☑ ~~Categorical vs Categorical analysis using cross-tabulations (pd.crosstab())~~

Phase 4: EDA to SQL Questions

Deadline: 29.01.2026

- Review key findings from univariate and bivariate EDA.

• Translate observed patterns into clear SQL questions • Ensure all SQL questions are relevant to the project objective and based on the cleaned dataset.
Phase 5: Presentation (PPT)
Deadline: 29.01.2026
Deliverable: 10-slide presentation
• Slides outline: 1) Project title & objective 2) Dataset overview 3) Research questions 4) Data cleaning summary 5) Univariate analysis highlights 6) Bivariate analysis highlights 7) Key EDA insights 8) SQL questions & examples 9) Key findings 10) Conclusion & next steps

6. Final Deliverables

- ☑ ~~Cleaned dataset (CSV file)~~
- ☑ ~~Final, well-documented analysis notebook~~
- ☑ ~~List of SQL questions with corresponding queries~~
- ☑ ~~Presentation summarizing key findings~~

Analysis Rule

Every plot or table must be followed by one to two concise sentences explaining:

- The observed pattern
- Its relevance to the research question

Resources:

- Kaggle Dataset