Lending Club loan Cleaning and Analysis Structure

Cleaning

1) Standardising term column by making it numeric.
2) Making emp_length column numeric by removing extra symbols
3) Categorising home ownership into MORTGAGE, OWN, RENT, OTHER
4) Standardising isuue_d into MYD
5) Categorising verification_status into VERIFIED and NOT VERIFIED
6) Replacing blanks in emp_title into NA
7) Grouping annual_inc into inc_group
   a) Low: Less than $50,000
   b) Medium: $50,000 to $100,000
   c) High: More than $100,000
8) Removing Duplicates

Analysis

KPI –

- total loan Volume,
- Portfolio Risk Rate,
- Average Interest Rate,
- Average Customer Income

1. Risk Analysis: Default Rate by Loan Grade

- **Why we do it:** To check the quality of the loans. In your data, Grade is the strongest predictor of risk.
- **What it tells:** It shows that **Grade A** loans are very safe (~6% default rate), while **Grade G** loans are highly risky (~48% default rate). This validates the grading system.

2. Business Growth: Loan Volume over Time

- **Why we do it:** To track the business trajectory. Your data shows massive growth from 2012 to 2014, with a peak around 2014-2015.
- **What it tells:** Shows how many loans were issued each year. It highlights the "boom" years and potentially recent slowdowns (2016 seems lower in your dataset).

3. Purpose Analysis: Why are people borrowing?

- **Why we do it:** To understand customer needs.
- **What it tells:** "Debt Consolidation" is by far the biggest driver (over 230k loans), followed by "Credit Card" refinancing. These two categories make up the bulk of your business.

4. Income vs. Risk: Do Higher Earners Pay Back?

- **Why we do it:** To see if income is a good safety net.
- **What it tells:** Yes. Your data shows "Low" income earners (<$50k) have a default rate of ~24%, while "High" earners (>$100k) have a default rate of only ~15%.

5. The "Term Risk" Analysis (36 vs. 60 Months)

- **Why we do it:** To see if long-term loans are dangerous.
- **What it tells: This is a major finding.** In your data, 60-month loans have a default rate of **~32%**, while 36-month loans are only **~16%**. Long-term loans are nearly *twice* as risky.

6. The "Verification Paradox"

- **Why we do it:** To test if checking income helps.
- **What it tells: Counter-intuitive result.** In your data, "Verified" loans actually default *more* (~22%) than "Not Verified" loans (~15%).
    - *Business Context:* This often happens because banks only demand verification for borrowers who look "borderline" risky on paper, whereas rock-solid borrowers get automatic approval (Not Verified).