Exploratory Data Analysis

Lending Club Case Study

Facilitator: Uvaraj Thulasiram

Team member: Vaishali Makwana

Agenda

1. Introduction
2. Data Understanding
3. Data Preparation
4. Exploratory Data Analysis
5. Recommendations

1. Introduction

1.1 Objective

To analyze the loan dataset of a leading lending company, identify risks and issues in applicant's borrowing patterns, and prepare a comprehensive case study. The case study will highlight key risk factors and include observations and recommendations to mitigate potential credit losses.

- Develop a robust decision-making framework for loan applications based on identified risk factors.
- Ensure applications from applicants with the potential to repay the loan are not rejected unnecessarily.
- Prevent the approval of applications from applicants who are likely to default.

2. Data Understanding

The initial data exploration revealed several key observations to streamline the dataset for effective analysis, as outlined below:

- The dataset contains 39,717 rows and 111 columns.
- Most columns are of type object, with some date-like columns (e.g., issue_d) and categorical columns (e.g., grade, sub_grade, term, loan_status, verification_status).
- Columns such as id, member_id, url, and desc are irrelevant for risk analysis and can be removed.
- Several columns either have unique values for all rows or are entirely empty; these columns do not contribute to analysis and should be removed.
- 54 columns contain null values for all rows and can be safely deleted.
- Columns with only one unique value (e.g., collections_12_mths_ex_med, chargeoff_within_12_mths, tax_liens, pymnt_plan, policy_code, etc.) are non-informative and can also be removed.
- The columns loan_amnt, funded_amnt, and funded_amnt_inv provide overlapping information. Only loan_amnt (borrower's requested amount) is necessary.
- addr_state and zip_code show inconsistent distributions; this requires further investigation. Columns like next_pymnt_d lack sufficient data for meaningful analysis and can be excluded.

3. Data Preparation – Data Cleaning

Further data refinement and analysis revealed the following observations to ensure a clean and focused dataset for accurate insights:

- Columns with more than 50% null values (mths_since_last_record, mths_since_last_delinq, etc.) were removed.
- Relevant columns like emp_length, pub_rec_bankruptcies, and revol_util with some null values were retained, but records with nulls in critical fields like revol_util were dropped.
- A derived column may be created from the title field for categorical grouping (e.g., "Consolidation," "Home Improvement") or the column can be dropped if deemed unnecessary.
- Numeric columns like `loan_amnt` and `annual_inc` contain extreme outliers that significantly impact visualization and analysis.
- An interquartile range (IQR) method was used to identify outliers, revealing that 60% of checked columns had less than 8% outliers, which can be sliced for better results.
- Columns such as last_pymnt_d and emp_title were removed due to limited relevance, while others were retained for further analysis.
- Removing all outliers would result in excessive data loss, so only columns with minimal outliers will be adjusted.

3. Data Preparation – Data Engineering

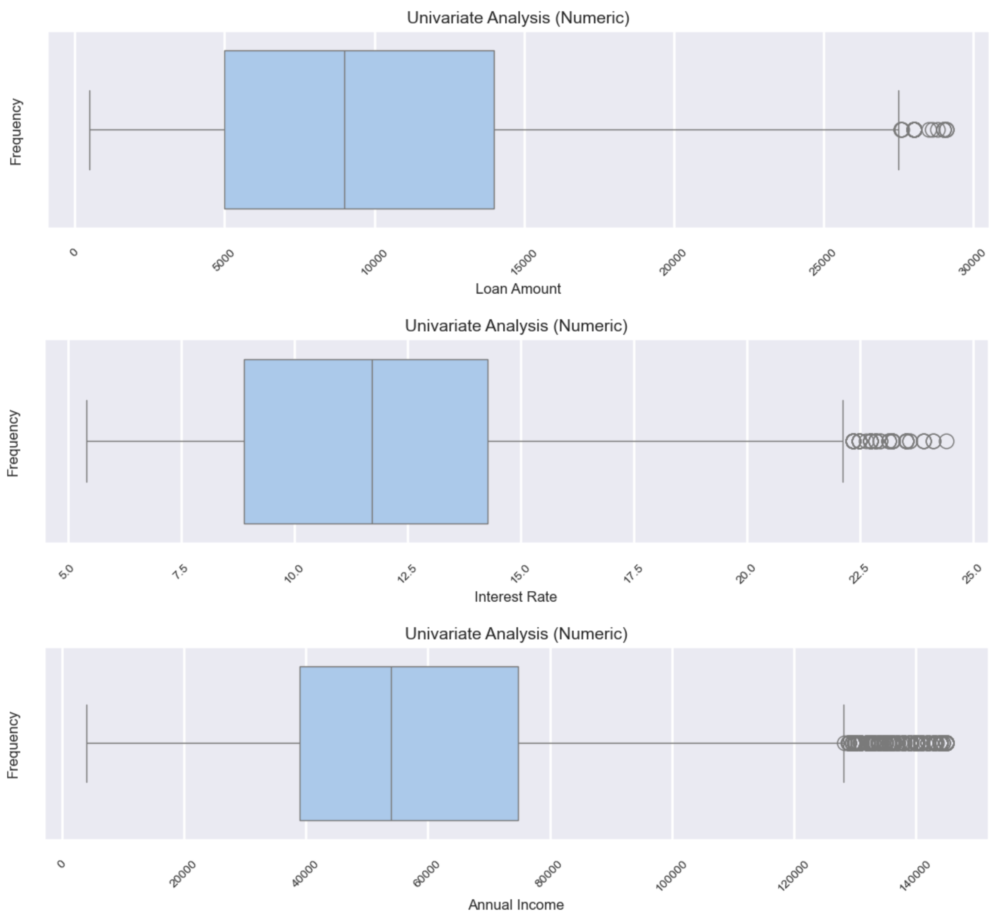
The following data engineering steps were implemented to enrich the dataset with derived features and optimize its structure for analysis:

- Derived columns d_earliest_cr_line_month and d_earliest_cr_line_year were created from the earliest_cr_line date column.
- Additional date-based columns (month, quarter, and year) will be extracted from issue_d and last_credit_pull_d columns.
- Created income-based categories (High Income, Middle Class, Low Income) derived from the annual_inc column.
- Derived credit health categories (Excellent, Good, Average, Poor, Critical) based on revol_util, where higher values indicate lower credit health.
- Categorized credit risk levels (Too many, Many, Moderate, Few, Very few) based on the number of open credit lines (open_acc).
- Grouped and converted relevant columns into categorical data types for better analysis and efficiency.

4. EDA – Univariate Analysis (Numeric)

As the initial step in exploratory data analysis, univariate analysis was performed on numeric columns to derive key insights.

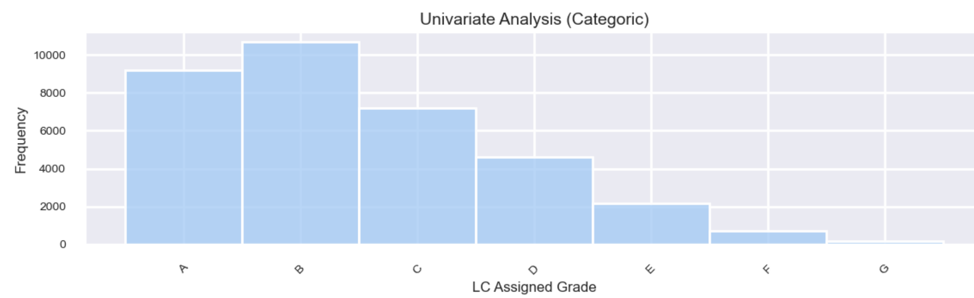
- Most borrowers take loans ranging between $5,000 and $14,000, with a few borrowing more than $27,000.
- The average interest rate is around 12%, but it can go as high as 25%, requiring further investigation into the circumstances leading to such high rates.
- Borrowers' annual incomes typically range from $40,000 (25th percentile) to $75,000 (75th percentile).

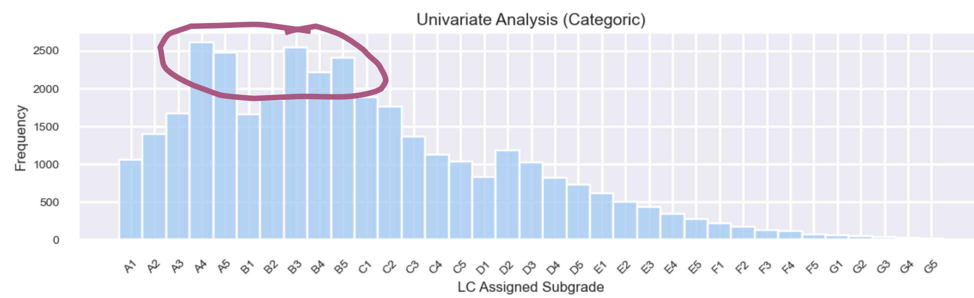


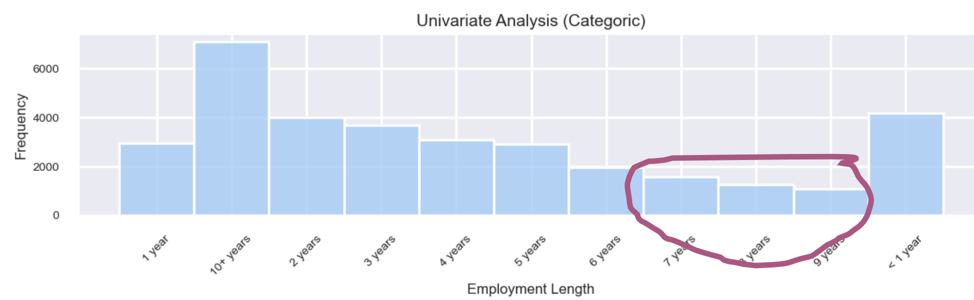
4. EDA – Univariate Analysis (Categoric)

After analyzing the numeric columns, univariate analysis was performed on categoric columns to derive key insights.

- Most loans were offered to Grade B and Grade A customers.
- Within the grades, most of the loans were offered to sub grades, A4, A5, B3, B5 and B4.
- The customers who borrowed the least are the ones with employment length ranging from 7 to 9 years.



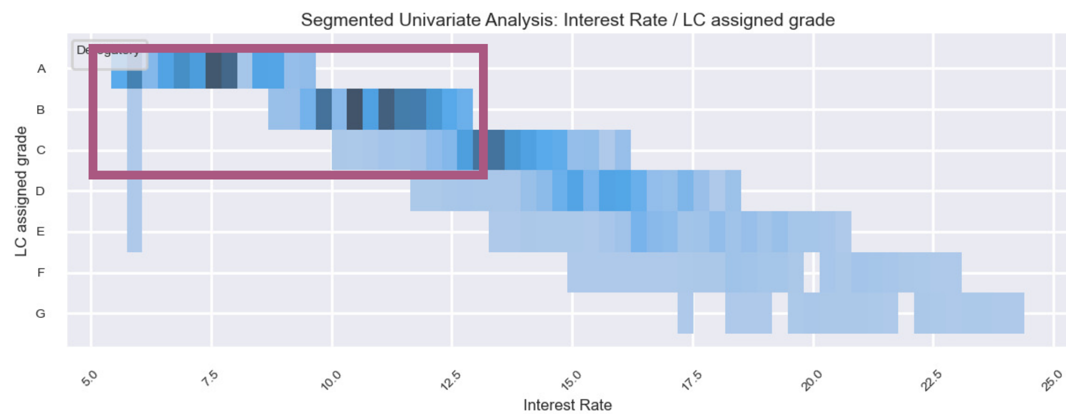


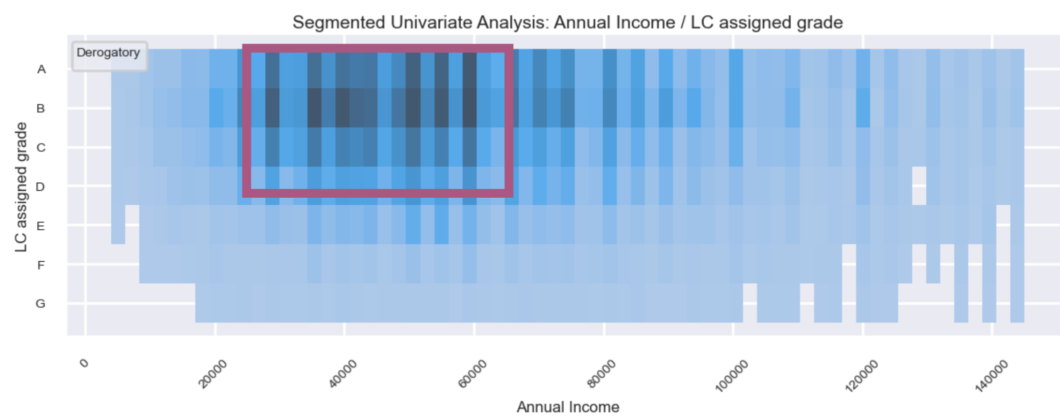


4. EDA – Univariate Analysis (Segmented)

Doing segmented univariate analysis on the numeric columns segmented by categories, we get the following insights,

- As expected, the better interest rates were given to the Grade A and B customers.
- It also reveals that, the grades are an ordered collection. The lower the grade the higher the risk of approving a loan application in future.
- The customers who borrowed the most belong to the better grades.
- They're neither in the high- or low-income category, but middle class.





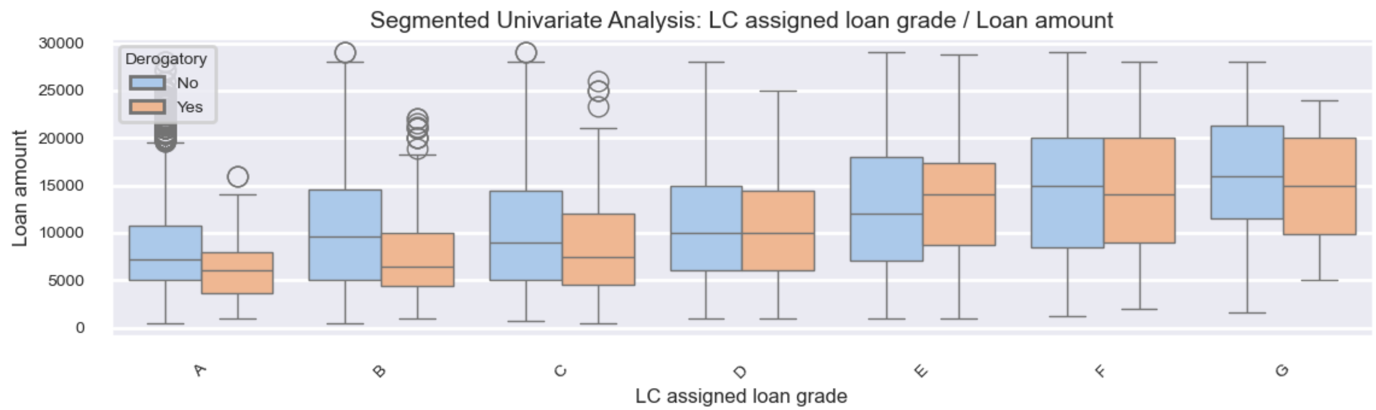
4. EDA – Univariate Analysis (Segmented)

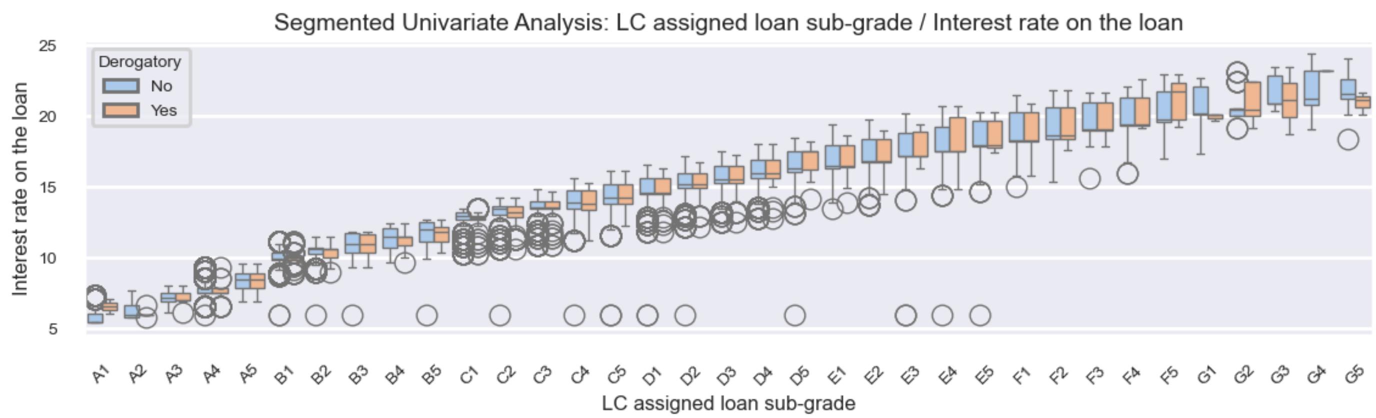
Doing further segmented univariate analysis reveal the following,

- The borrowers with lower grades borrowed more than the one with higher grades

- The interest rate for those borrowers with lower sub grades are much higher which proves again that the grades are inversely proportional to risk.

- The customers with lower grades show more disrespectful attitude. This insight is derived from the number of public derogatory records.

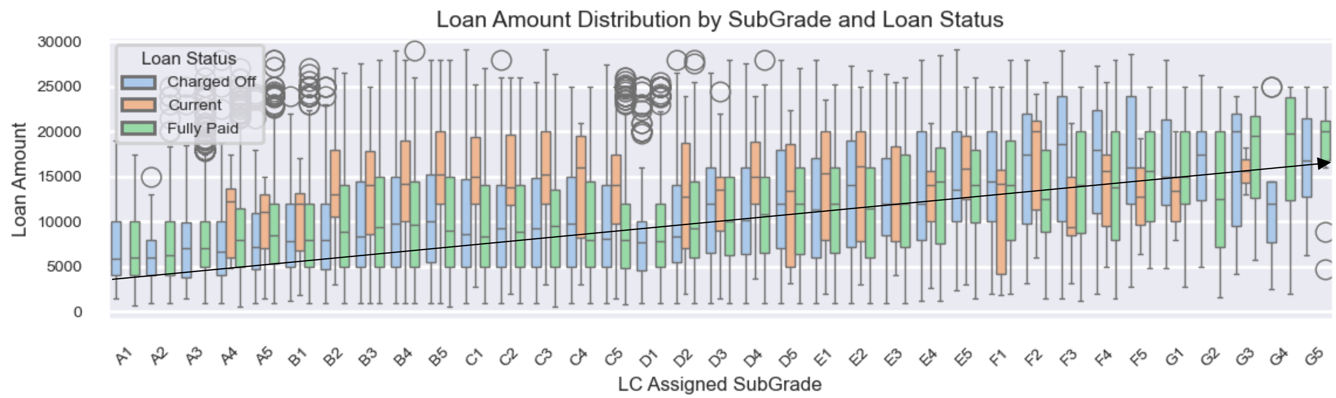


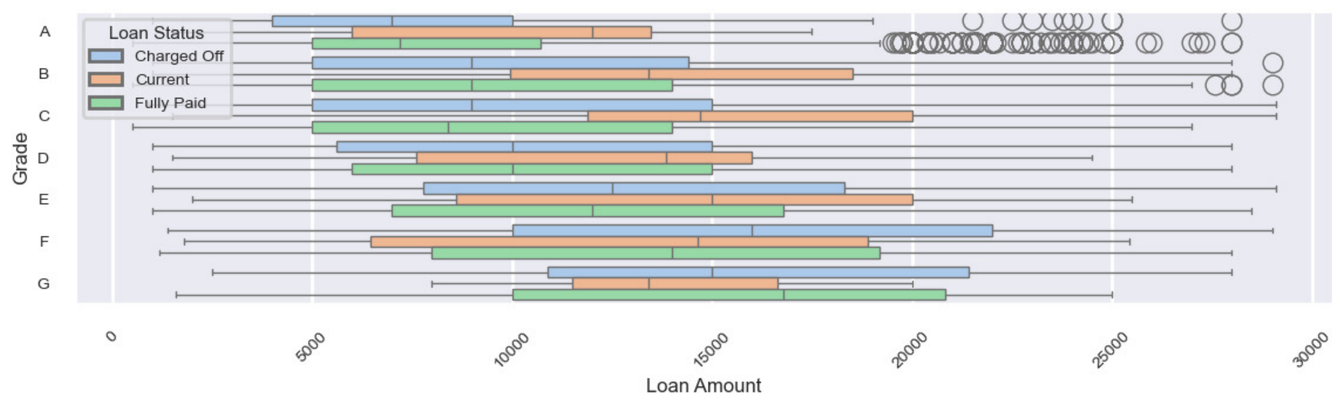


4. EDA – Bivariate Analysis

Bivariate analysis helps validate assumptions and clarify insights derived from univariate analysis.

- The first chart on the right shows that borrowers with lower subgrades tend to take larger loans at higher interest rates and default more frequently, leading to a higher number of charged-off loans.
- The second chart reinforces this observation by illustrating the distribution of loan amounts across different grades.





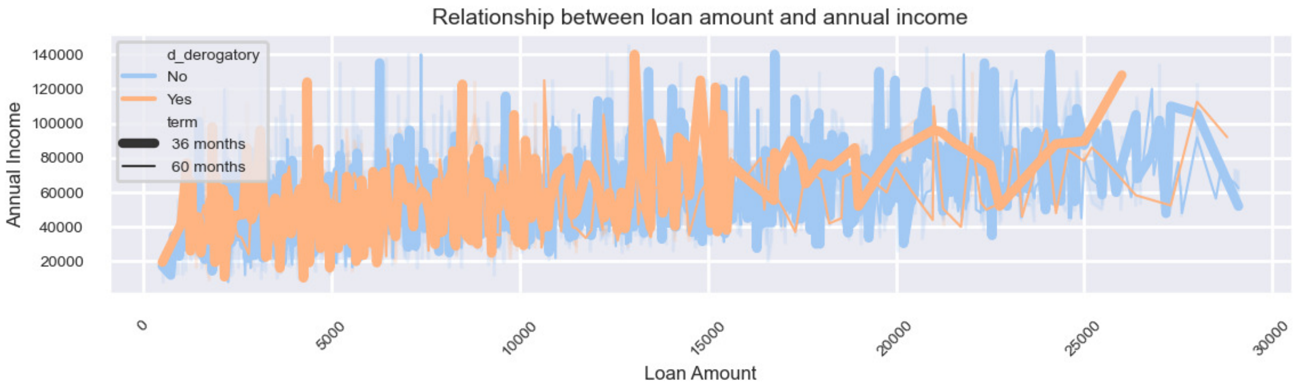
4. EDA – Bivariate Analysis

Heatmap based on correlation matrix between numeric variables

Key insights from the heatmap,

- A strong correlation (93%) between loan_amnt and installment highlights that higher loans directly result in larger installment amounts.
- 47% correlation between revol_util and int_rate suggests borrowers with higher credit utilization often face higher interest rates.
- A moderate positive correlation (36%) between annual_inc and loan_amnt indicates that higher-income borrowers tend to take larger loans.
- Public derogatory records and bankruptcies are strongly associated. An 85% correlation indicates high risk.
- Surprisingly debt-to-income ratio has minimal impact on loan_amnt and annual_inc

4. EDA – Multivariate Analysis



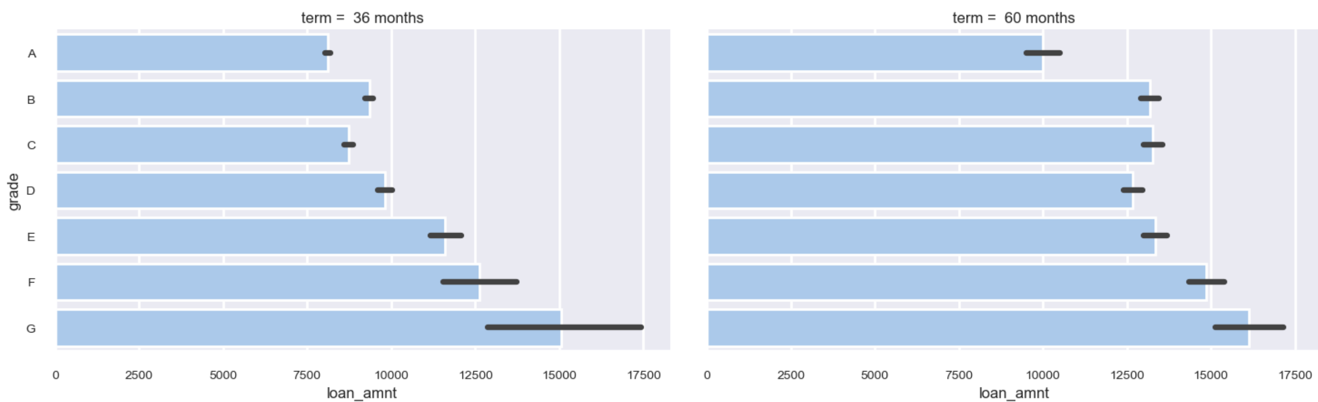
Line plot legends:

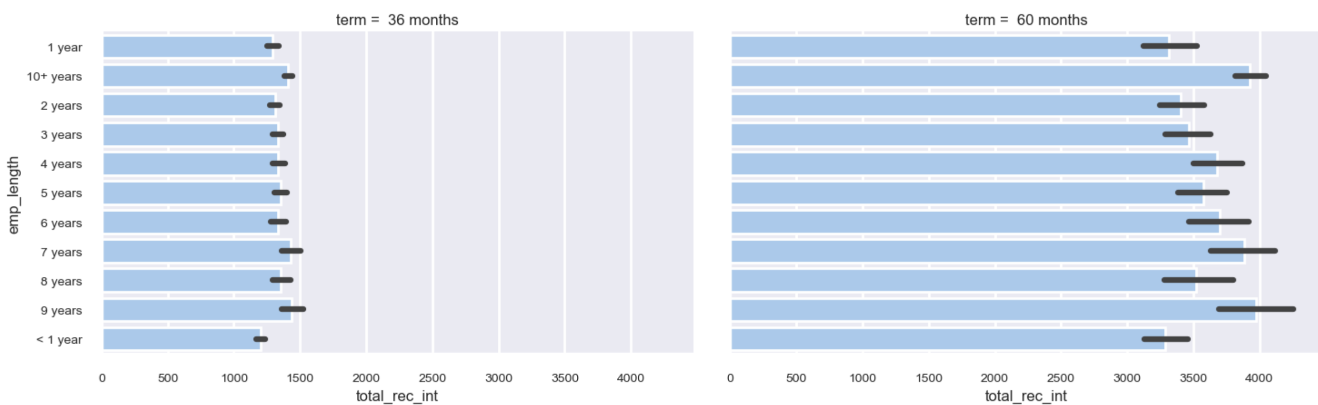
1. *Public Derogatory records are categorized to [‘Yes’, ‘No’] and indicated by [‘Blue’, ‘Orange’] line colors.*
2. *Loan Term is indicated by the thickness of the lines.*

Key insights from the multivariate analysis using Seaborn’s lineplot,

- There are fewer derogatory behavior from customers who has borrowed higher loan amount.
- The density of derogatory behavior is very high when the loan amount is less than $15,000/-
- The loan amount is gradually increasing as the annual income, which reconfirms higher correlation between the two variables.

4. EDA – Multivariate Analysis





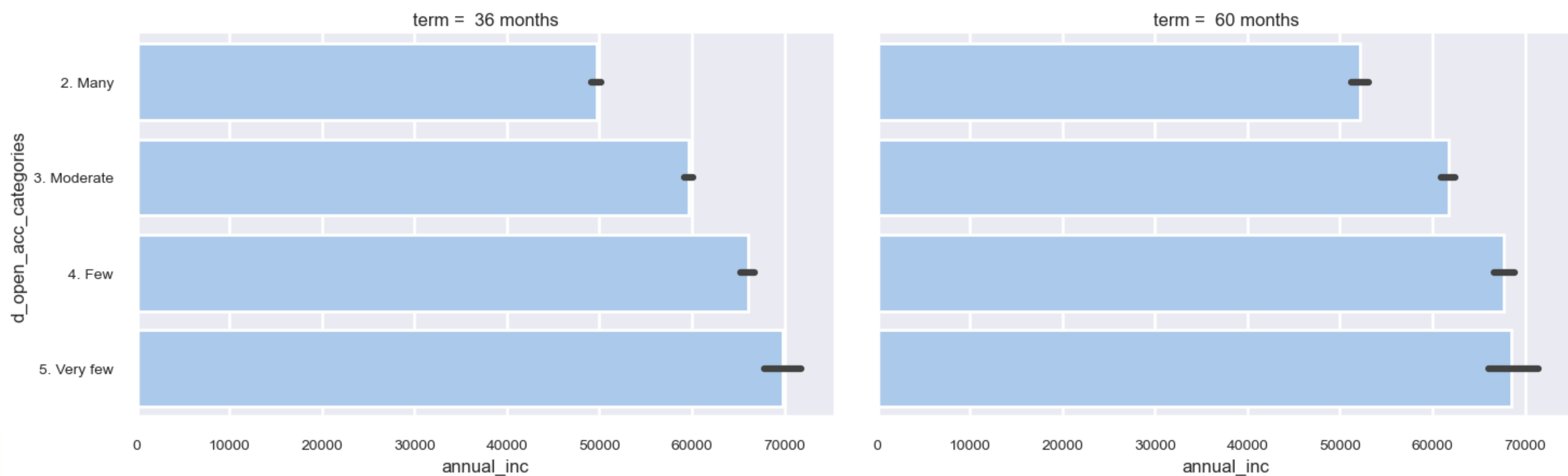
Insights from multivariate analysis using Seaborn's catplot

- Long term loan amounts are relatively higher compared to short term loan amounts across all grades
- Long term loans fetch more interest than short term loans across all employee lengths.
- Hence offering long term loans will be more profitable and as well have less burden on the customer as the installment amount is inversely proportional to the loan term

4. EDA – Multivariate Analysis

Comparing the annual income against the category of open credit lines based on the loan term gives the following insights,

- Customers with low income tend to have more open credit lines, which directly translates to high risk.



5. Recommendations

Recommendation Category	Action	Positive Focus on Higher-Grade Customers
Positive	Focus on Higher-Grade Customers	• Prioritize loan offers to customers with credit grades A, B, and C. • Explore targeted marketing campaigns to attract high-grade customers within the 6-9 years of employment segment. • Consider offering competitive interest rates for high-grade customers to remain competitive while potentially increasing loan volume. • Analyze the profitability of this segment while considering the lower interest rates.
Positive	Incentivize Good Behavior	• Implement loyalty programs or reward systems for customers with consistently positive payment histories. • Educate customers on the importance of maintaining good credit scores and minimizing derogatory records. • Offer financial counseling services to assist customers in improving their creditworthiness. • Monitor the impact of these incentives on loan approval rates and customer retention.
Positive	Promote Longer Loan Terms	• Educate customers on the benefits of longer loan terms, such as lower monthly installments and potential long-term cost savings. • Offer flexible loan term options to cater to individual customer needs and financial situations. • Clearly communicate the impact of loan term on total interest payments to ensure transparency. • Monitor customer satisfaction and repayment rates across different loan terms.
Positive	Incorporate Debt-to-Income Ratio	• Implement a robust debt-to-income (DTI) ratio assessment as a key factor in loan approval decisions. • Establish clear DTI thresholds for different loan products and risk categories. • Develop a scoring system that incorporates DTI along with other relevant credit risk factors. • Regularly review and adjust DTI thresholds based on market trends and internal risk assessments.
Negative	Avoid Over-reliance on Low-Risk Segments	• Avoid over-concentrating loan portfolios in the low-risk (high-grade) segment, as this may limit overall profitability. • Diversify lending strategies to include moderate-risk segments while maintaining appropriate risk controls. • Continuously monitor the risk-return profile of the loan portfolio and adjust lending strategies accordingly.

Thank you

Facilitator: Uvaraj Thulasiram

Team member: Vaishali Makwana