LENDING CLUB CASE STUDY ANALYSIS

Amit Kumar Das
Mukundhan S

PG Program in Machine Learning & Artificial Intelligence (C71)

upGrad

iiit-b
ज्ञानमुत्तमम्

CONTENTS

INTRODUCTION

Case Brief

- A consumer finance company which specialises in lending various types of loans to urban customers.
- When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile
- Two types of risks are associated with the bank's decision
  - ✓ Not approving the loan results in a loss of business to the company
  - ✓ Approving the loan to an applicant not likely to repay the loan, may lead to a financial loss for the company

Objective

- To identify patterns which indicate if a person is likely to default
- Leveraging the patterns to take decisions that may include denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc

APPROACH

Understand, Clean & Transform Data	UNIVARIATE Analysis	BIVARIATE Analysis	MULTIVARIATE Analysis	Recommendations
▪ Awareness about key columns/features ▪ Identify & treat null values, duplicates ▪ Identify & drop columns not relevant or useful ▪ Missing Values, Outliers Treatment ▪ Standardize, Manipulate, Derive to aid analysis ▪ Binning/Bucketing	▪ Categorical Unordered Analysis ▪ Categorical Ordered Analysis ▪ Categorical Quantitative Analysis ▪ Bar Plots	▪ Histogram, Box Plot, Scatter Plot	▪ Correlation Heatmap	▪ Key inferences from the data analysis ▪ Summary of key recommendations for the company (Lending Club)

DATA UNDERSTANDING

After the basic clean up (nulls, duplicates), a comprehensive assessment done to identify other columns that need to be removed. Snapshot of the assessment below

Sr #	Column Names	Remarks
1	*id, member_id*	Not required
2	*acc_now_delinq*	Empty
3	*pymnt_plan, initial_list_status*	*fixed value as n & f respectively for all*
4	*url, desc, emp_title*	Not useful for analysis
5	*title*	Too many distinct values, not useful
6	*zip_code*	Full info not available
7	*mths_since_last_delinq*	Only partial info available
8	*mths_since_last_record*	Approx 10% only has values
9	*revol_bal, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv*	Not useful for defaulter analysis
10	*total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee*	Not useful for defaulter analysis
11	*last_pymnt_d, last_credit_pull_d, last_pymnt_amnt, next_pymnt_d, chargeoff_within_12_mths*	Not useful for defaulter analysis
12	*collections_12_mths_ex_med, policy code, acc_now_delinq, delinq_amnt, tax_liens , application_type*	Only 1 value in the columns
13	*addr_state*	State info will not help in identifying likely defaulters

DATA CLEANING AND MANIPULATION

- ❑ Load required python libraries
- ❑ Load the loan.csv and try to visualize the dataset using inbuilt functions
- ❑ Identify columns which are completely null and then drop them
- ❑ Identify Columns which have 50% null and drop them
- ❑ Identify columns which do no have any effect on the analysis and drop them
- ❑ Identify columns which needs adjustments in datatype. Like object to int or float
- ❑ Identify rows which do not contribute anything to the analysis
- ❑ Check for rows whose values need changes and can be imputed with another values.

UNIVARIATE DATA ANALYSIS

❖ **Univariate analysis** is a statistical method used to analyze and summarize data sets consisting of **one variable**. It deals with the analysis of a single variable, rather than multiple variables, to understand its distribution, central tendency and dispersion.

❖ It was carried out for both **Categorical** and **Quantitative** Variables

Categorical

Unordered

- Home Ownership (home_ownership)
- Verification Status (verification_status)
- Purpose of the Loan (purpose)
- Address State of Loan application (addr_state)

Ordered

- Grade of Loan (grade)
- Sub-grade of Loan (sub_grade)
- Term of the loan (36 / 60 months) (term)
- Employment length of applicant (emp_length)
- Issue year (issue_year)
- Issue month (issue_month)

Quantitative

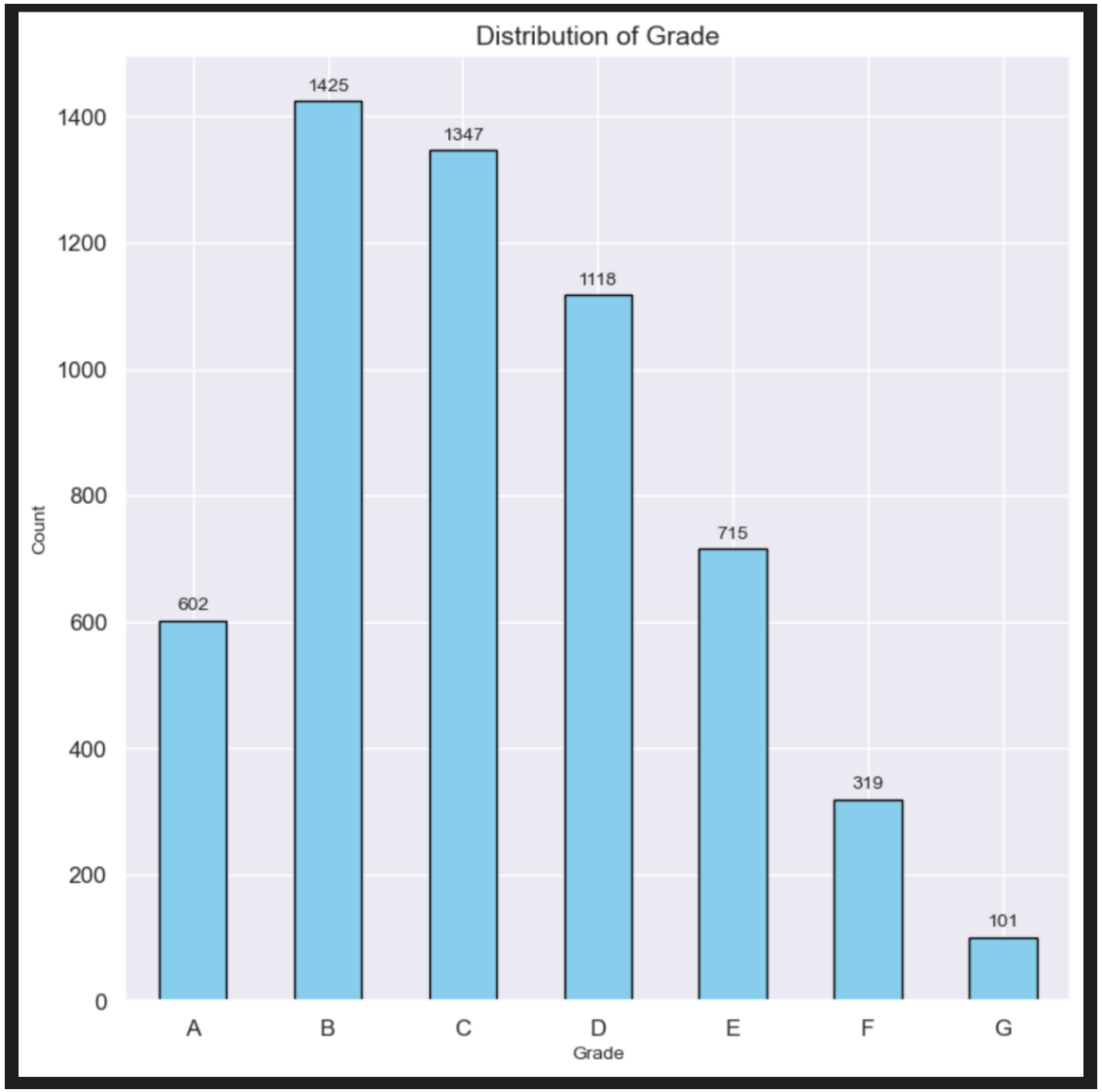
- Loan Amount (loan_amnt)
- Funded Amount (funded_amnt)
- Interest Rate (int_rate)
- Monthly Installments (installment)
- Annual Income of applicant (annual_inc)
- Debt-Income Ratio (dti)
- Delinquencies in 2yrs (delinq_2yrs)
- Number of open credit line (open_acc)
- Inquiries in last 6 months (inq_last_6mths)
- Revolving Balance (revol_bal)
- Total Payment (total_pymnt)
- Total Payment funded by investors (total_pymnt_inv)

UNIVARIATE DATA ANALYSIS

Loans against Grade

Key Inference

- 68% of defaulters are from the loan categories B, C and D
- A quarter of the defaulters are from B category.



upGrad

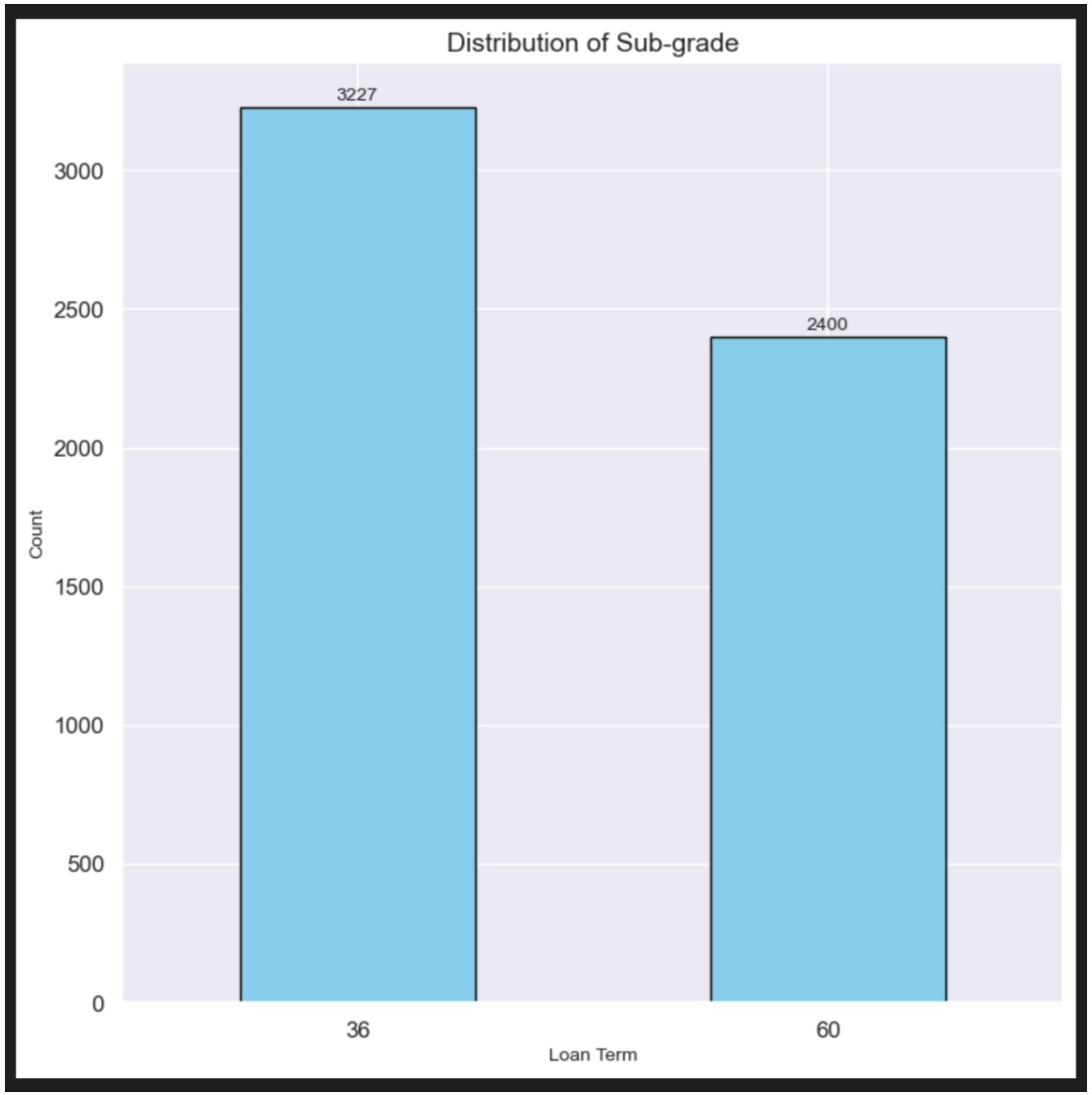
iimt-b
ज्ञानमुत्तमम्

UNIVARIATE DATA ANALYSIS

Defaulters against Loan Term

Key Inference

- There are 3227 defaulters who have been granted with term 36 months.
- This corresponds to about 57% of total defaulters.

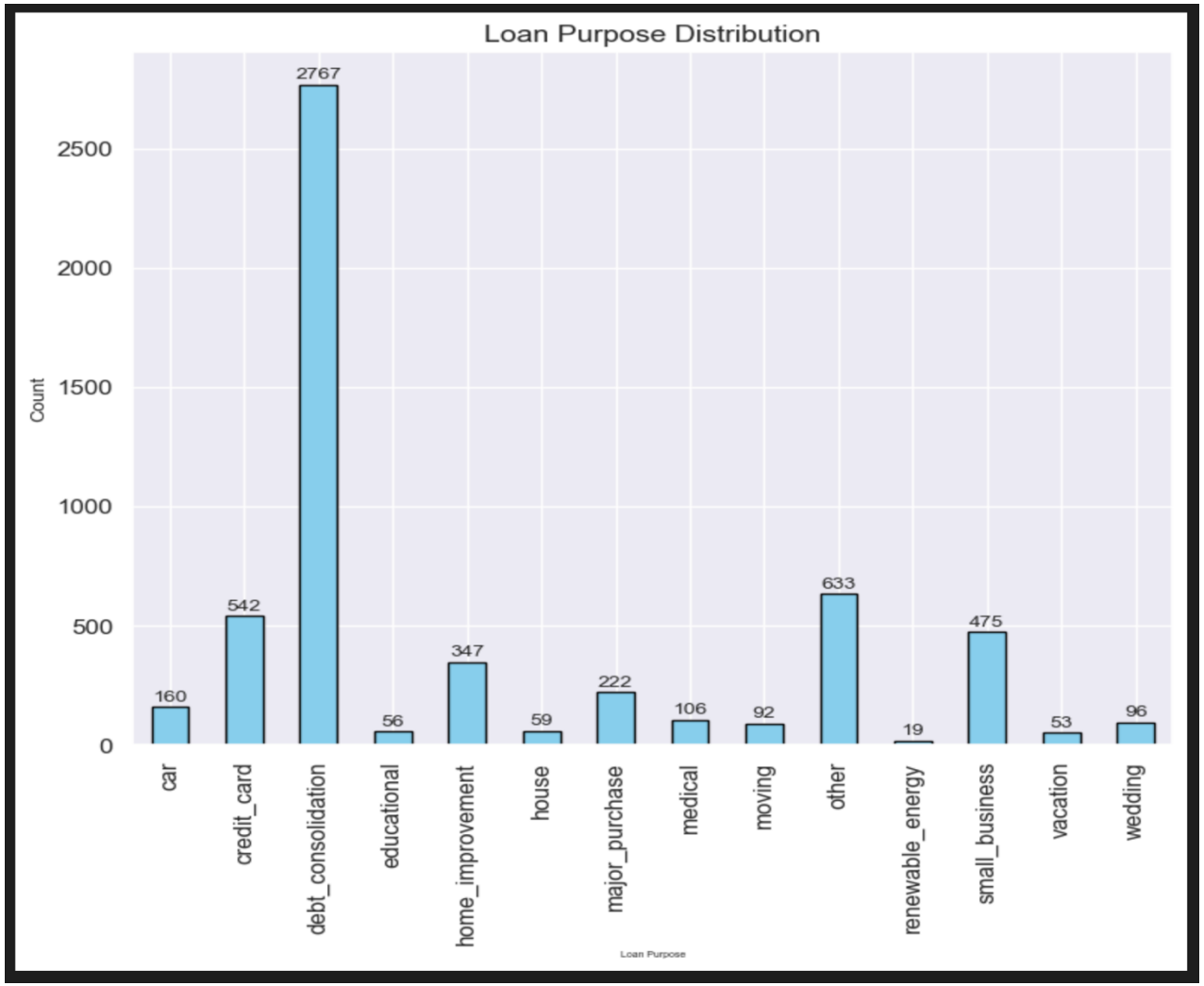


UNIVARIATE DATA ANALYSIS

Loan Defaulters against Loan Purpose

Key Inference

- Debt_Consolidation – 49%
- Other – 11%
- Credit Cards – 10%
- Small Business – 8%



upGrad

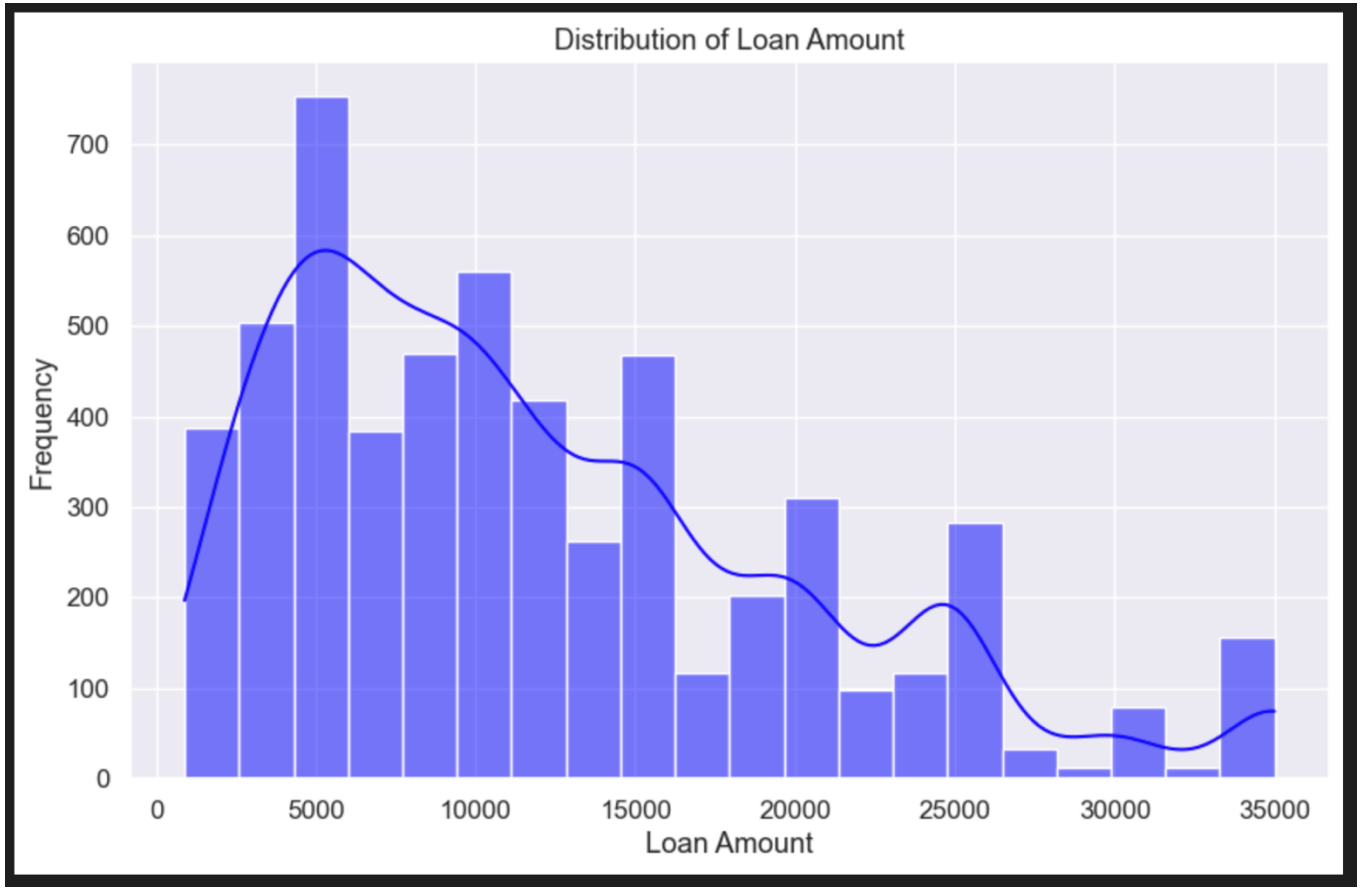
iit-b ज्ञानमुत्तमम्

UNIVARIATE DATA ANALYSIS

Loan Defaulters against Loan Amount

Key Inference

- Majority of the defaulters are granted $5000 as loan amount.
- From data analysis we see that 50th percentile is at $10000.



UNIVARIATE DATA ANALYSIS - KEY INFERENCES & OBSERVATIONS (1/2)

Categorical - Unordered

1. The high proportion of renters among charged-off loans suggests that renting might be associated with a higher risk of loan default. This could be due to various factors such as higher financial strain due to rent payments.
2. Homeowners, especially those with mortgages, might be perceived as less risky borrowers.
3. A larger proportion of verified loans are charged-off could also indicate that verification, while helpful in risk assessment, doesn't always prevent defaults.
4. The high proportion of charged-off loans for "debt_consolidation" might suggest that borrowers struggling with existing debt are more likely to default on new loans taken out to consolidate those debts.
5. Loans for "small_business" or "home_improvement" might carry higher risks due to the inherent uncertainties and potential financial challenges.
6. Applicants from CA, FL, NJ, NY and TX are expected to default frequently

Categorical - Ordered

1. The Loan grades B,C,D have higher tendency to default. This is likely due to the fact that lower grades are assigned to loans with higher perceived risk.
2. The sub-grades B3,B4,B5,C1,C2 have higher defaulting rate. Further, subgrades A1, A2, A3 have least default rate
3. The loans of 36months default more, indicating applicants with short duration loans tend to turn into defaulters.
4. Employment length of 10 yrs tend to default more which might be due to the reason that the number of loans given are more to these types of applicants
5. Employment length indicates that loans to applicants who are new to jobs are second highest in default.
6. The distribution for "Issue Year" indicates that loan defaulters are more in 2011, but from the trend it seems like this might be due to the more loans given year or year due to increased business.
7. The "Issue Month" distribution indicates that loans granted in the last quarter of the year tend to default more, which might be due to the financial stress at the end of year

UNIVARIATE DATA ANALYSIS - KEY INFERENCES & OBSERVATIONS (2/2)

Quantitative Variables

1. 50% of the defaulters(2594) are having loan amounts more than $10000. The box plot indicates right skewed plotting which indicates that the company should exercise caution when giving higher loan amount
2. We see that most of the defaulters are applicants who have been granted with higher interest rates. The 25th percentile is at 11.31% and 75th percentile is at 16.4%. To reduce the risk of default, the lending company should consider offering loans at lower interest rates when possible
3. Among loan applicants who charged off, it's observed that the majority of them had monthly instalment amounts falling within the range of $168-$457 USD. The firm would closely monitor other factors while granting loan to the applicants who fall in this bucket of instalments
4. Despite higher incomes in some cases, the loans were still charged off. This suggests income alone may not be a strong predictor of creditworthiness or repayment capability
5. Further, company would exercise caution when granting loans to the candidates who have annual income less than $50000
6. Company should restrict giving loans to individuals who have DTI more than 15%. This indicates stress in their financial status.
7. The variable delinquencies in 2yrs gives an interesting insight that almost all the defaulters have delinquencies as zero. Indicating that this may not help in deciding the approval of loan process.
8. Loans with a higher number of open credit lines might be associated with a higher likelihood of being charged off. So the firm should exercise caution while granting loans to individuals who have higher open credit lines.
9. A significant portion of charged-off loans had very few inquiries in the last 6 months, so it might suggest that first time applicants have tendency to default more.

BIVARIATE DATA ANALYSIS

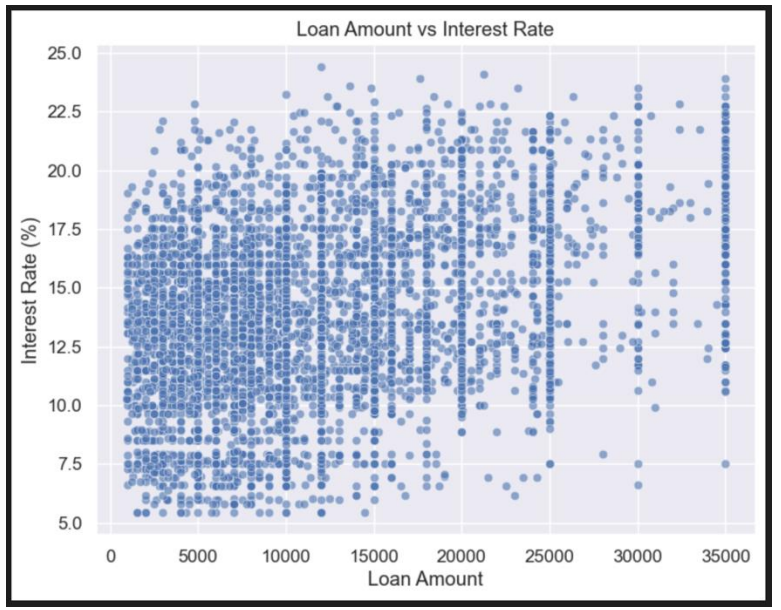
- **Bivariate analysis** is a statistical method that involves the simultaneous analysis of two variables (factors) and aims to determine the empirical relationship between them.
- The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables
- Analysis is done on the following variables
  - ✓ Loan Amount vs Interest Rate
  - ✓ Loan Amount vs Annual Income
  - ✓ Grade/Sub-Grade vs Interest Rate
  - ✓ Debt-to-Income Ratio (DTI) vs Loan Amount
  - ✓ Loan Purpose vs Loan Amount
  - ✓ Employment Length vs Loan Amount
  - ✓ Line plot to identify trends in loan_amnt over issue_year and issue_month.
  - ✓ Line plot to identify trends in int_rate over issue_year and issue_month.

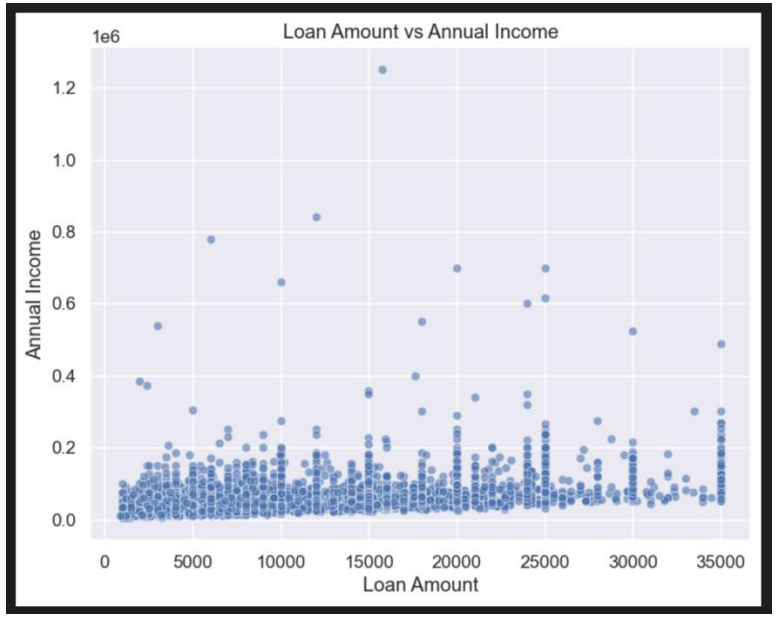
BIVARIATE DATA ANALYSIS

Loan Amount, Interest Rate and Annual Income

Key Inference

- There is a weak positive correlation between loan amount and interest rate for charged-off loans. This suggests that larger loans are slightly more likely to be charged off, possibly due to higher risk or difficulty in repayment.
- There is a weak negative correlation between loan amount and annual income for charged-off loans. This suggests that borrowers with lower annual incomes are slightly more likely to have their loans charged off.



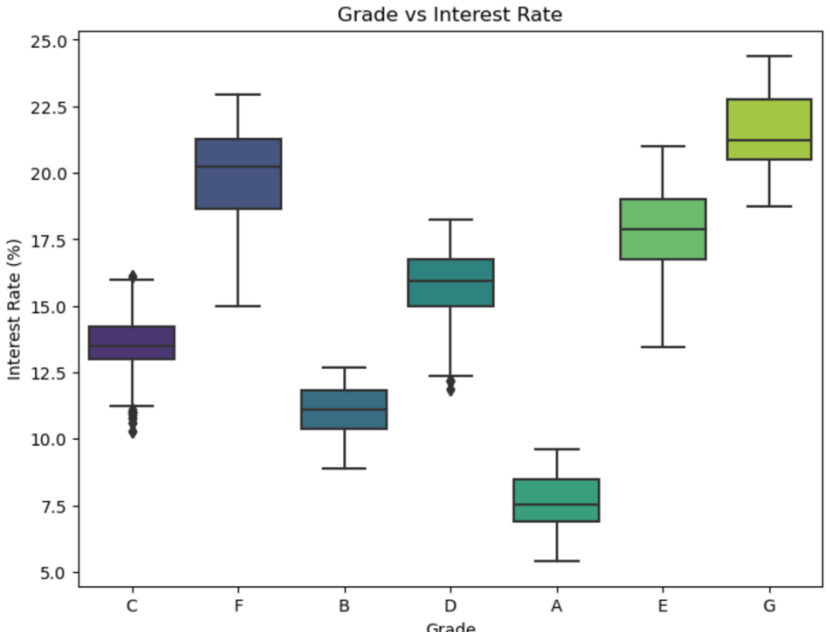


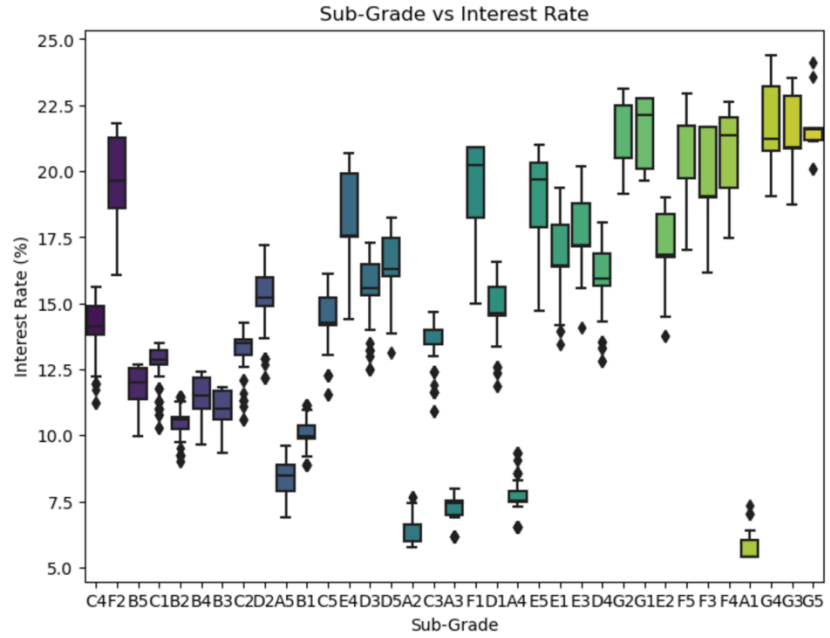
BIVARIATE DATA ANALYSIS

Interest Rate, Grade and Sub-Grade

Key Inference

- A strong positive correlation between loan grade and interest rate is evident. As the loan grade decreases (from A to G), the interest rate generally increases.
- The sub-grade plot provides a more granular view. Within each grade, the interest rate tends to increase as the sub-grade letter moves further from A (e.g., A1 to A5).



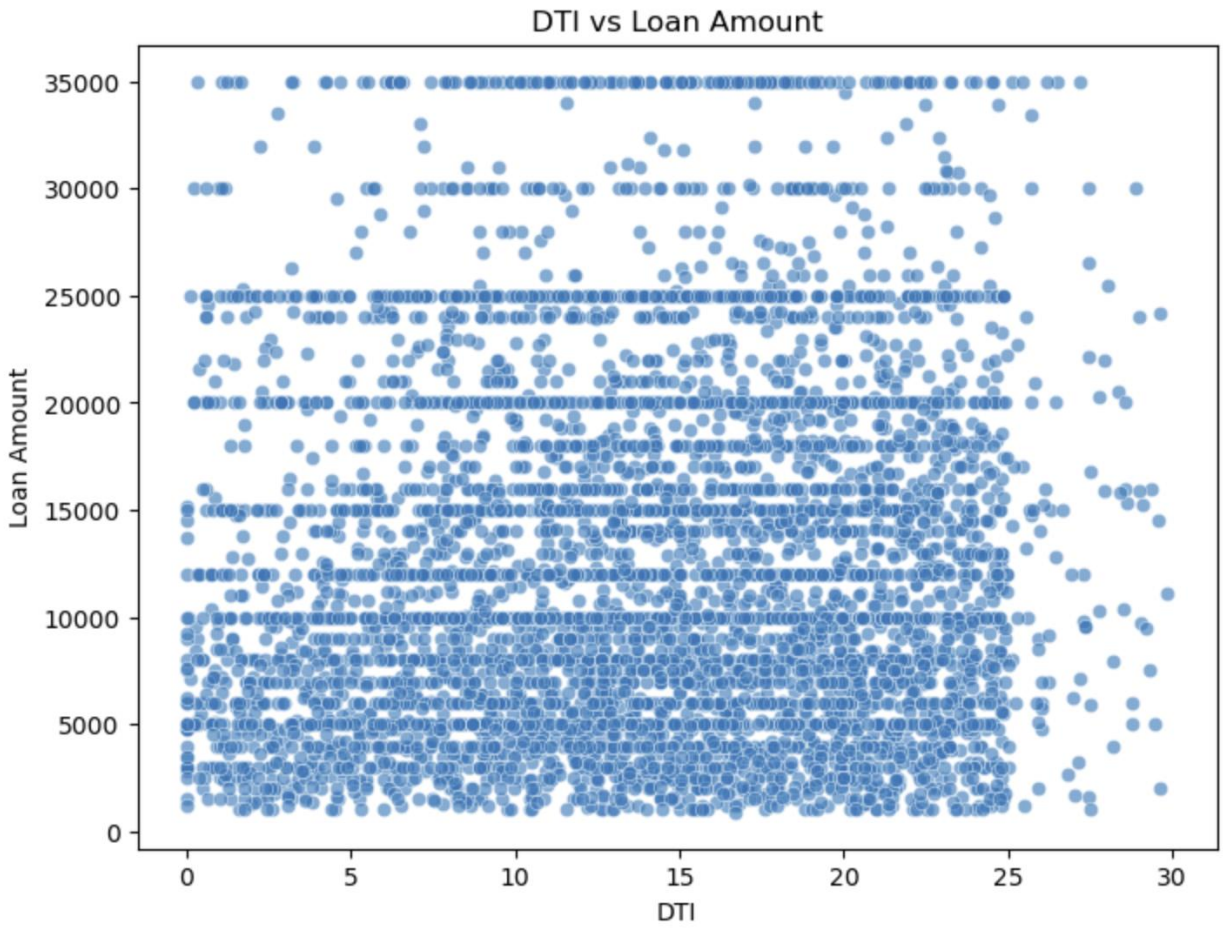


BIVARIATE DATA ANALYSIS

Loan Default Ratio against term and purpose

Key Inference

- There seems to be a slight positive correlation between DTI (Debt-to-Income Ratio) and Loan Amount. This suggests that borrowers with higher DTI might tend to borrow larger amounts.

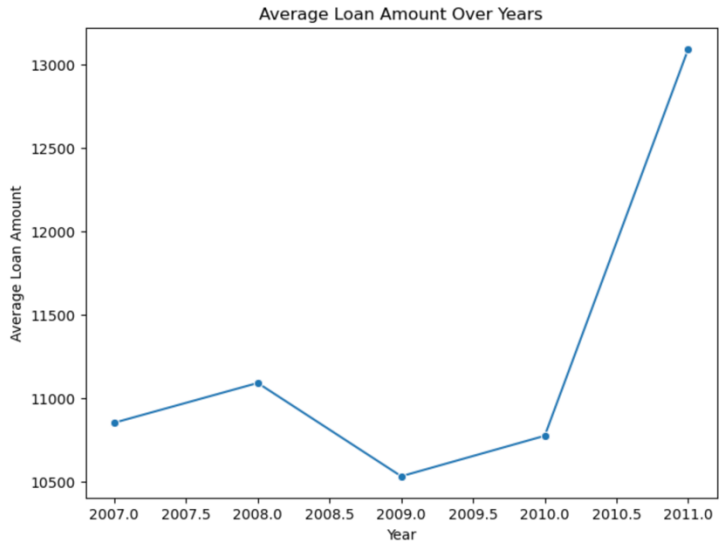


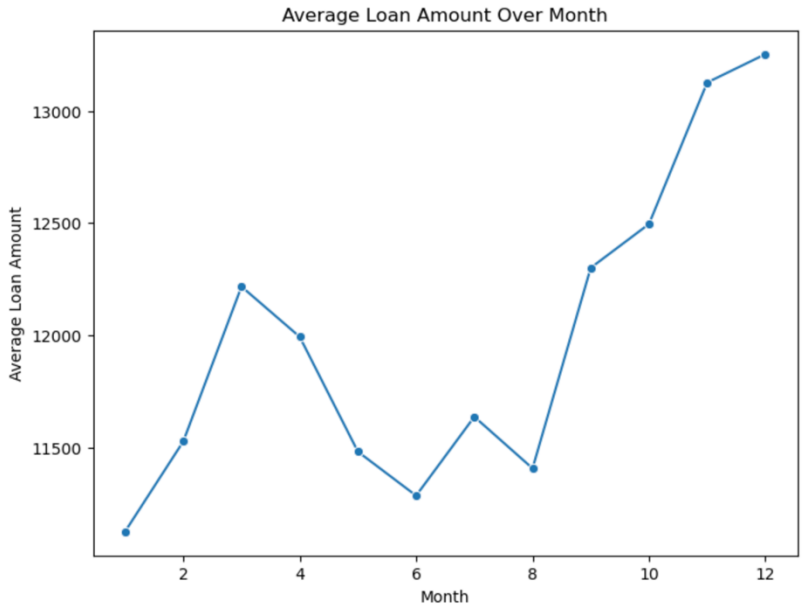
BIVARIATE DATA ANALYSIS

Loan Status against interest rate and purpose

Key Inference

- The plot shows a general upward trend in the average loan amount over the years from 2007 to 2011.
- The plot shows some fluctuations in the average loan amount across months. There is a noticeable increase in the average loan amount towards the end of the year.



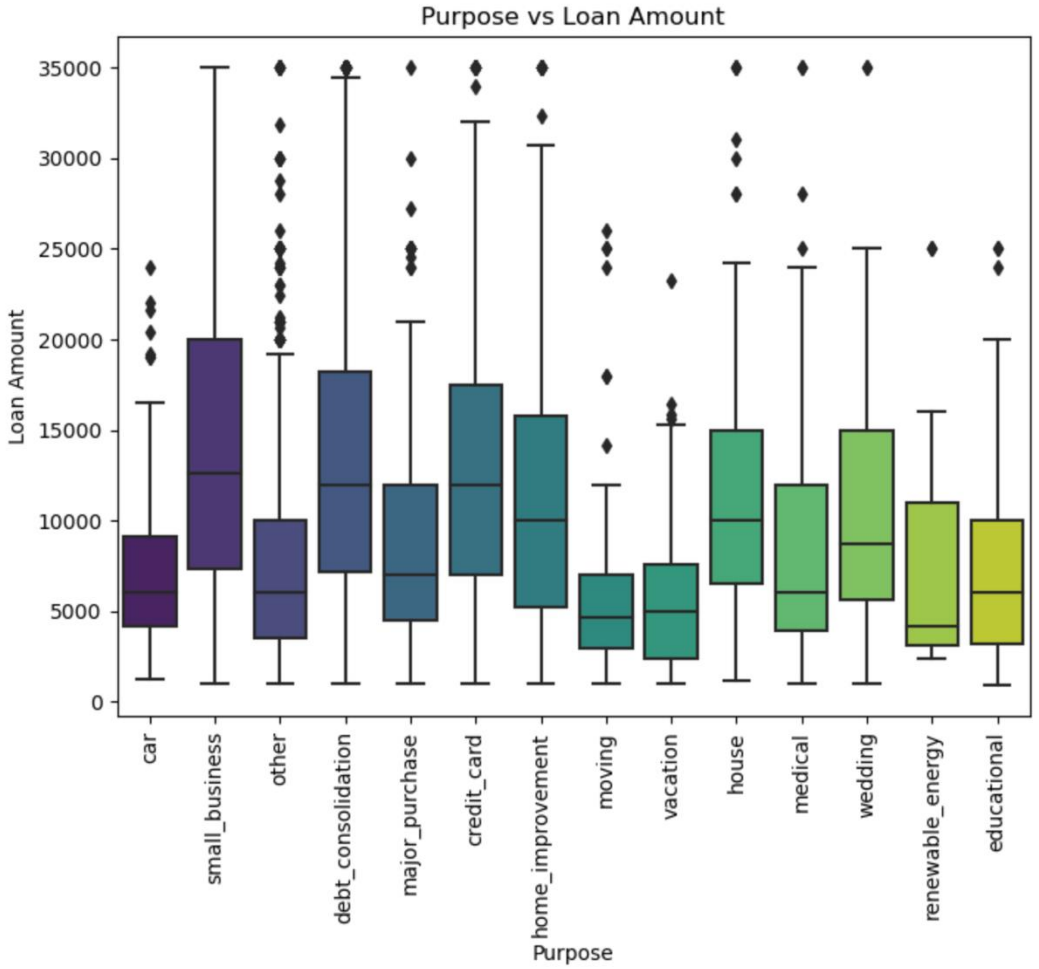


BIVARIATE DATA ANALYSIS

Grade against interest rate for each loan status

Key Inference

- The box plot shows that certain loan purposes are associated with higher loan amounts. For instance, loans for "small_business," "home_improvement," and "house" tend to have higher loan amounts compared to purposes like "credit_card" or "moving."



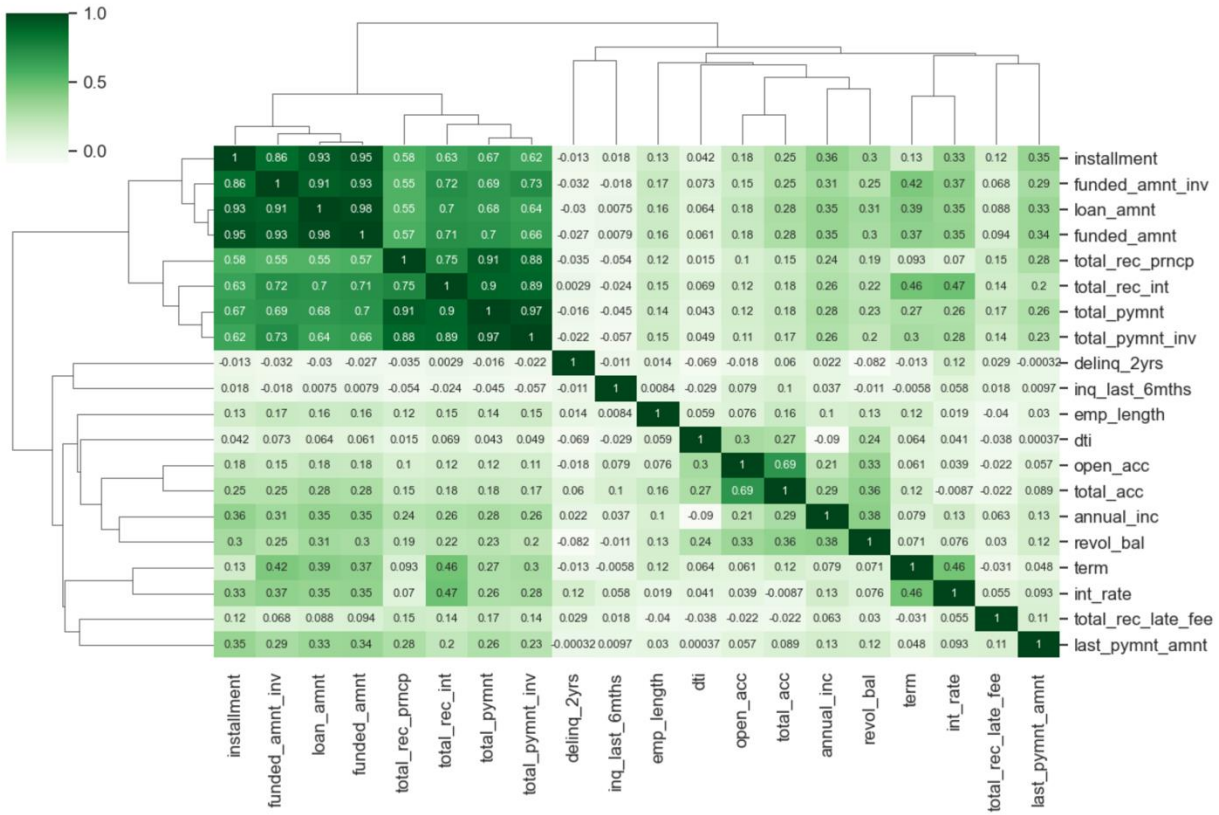
BIVARIATE DATA ANALYSIS - KEY INFERENCES & OBSERVATIONS

1. Larger loans are slightly more likely to be charged off, possibly due to higher risk or difficulty in repayment.
2. Borrowers with higher interest rates may be more likely to default, suggesting that lenders may be pricing in higher risk for these loans.
3. Borrowers with lower annual incomes are slightly more likely to have their loans charged off, possibly due to lower financial stability or higher debt burdens.
4. The strong correlation between grade and interest rate, and the increasing spread and outliers for lower grades, suggest that charged-off loans are likely concentrated in the lower grades (D-G) and their corresponding sub-grades.
5. The sub-grade level provides a more nuanced view of risk. Within each grade, the higher sub-grades (e.g., A5, B5) likely have a higher proportion of charged-off loans compared to the lower sub-grades (e.g., A1, B1), reflecting the lender's assessment of increasing risk within each grade.
6. Loan purposes with higher average loan amounts might have a higher risk of default if borrowers struggle to manage the larger debt burden.
7. The plot might reveal that certain loan purposes, such as "small_business" or "home_improvement," have a higher proportion of charged-off loans due to their inherent risks or the financial challenges associated with those projects.
8. The increase in average loan amount towards the end of the year is more pronounced for charged-off loans, it could suggest that borrowers might be taking on larger loans towards the end of the year for holiday spending or other seasonal expenses. This could potentially increase their financial burden and the risk of default.
9. Higher interest rates are associated with increased risk of default. Lenders may be charging higher rates to compensate for the perceived higher risk.

MULTIVARIATE ANALYSIS

To check if there is correlation or relationship among numeric variables

Key Inference



- funded_amnt, funded_amnt_inv, and loan_amnt show very high positive correlations
- total_pymnt, total_pymnt_inv, and total_rec_prncp (total principal received) have strong positive correlations
- dti (debt-to-income ratio) shows moderate positive correlations with open_acc (number of open credit lines) and total_acc (total number of credit lines)
- int_rate has a strong positive correlation with term

upGrad

iiit-b ज्ञानमुत्तमम्

MULTIVARIATE ANALYSIS - KEY INFERENCES & OBSERVATIONS

1. Strong Correlation within Loan Amounts - funded_amnt, funded_amnt_inv, and loan_amnt show very high positive correlations. This indicates that these variables are highly related, as expected.
2. Payment and Principal Correlations - total_pymnt, total_pymnt_inv, and total_rec_prncp (total principal received) have strong positive correlations. This suggests that borrowers who repay more in total also repay a larger portion of the original principal. total_rec_int (total interest received) has moderate to high correlations with total_pymnt and total_pymnt_inv. This is expected as interest is a component of the total payment.
3. Delinquency and Inquiry Relationships: - delinq_2yrs (number of delinquencies in the last 2 years) has a weak negative correlation with inq_last_6mths (number of inquiries in the last 6 months). This might suggest that borrowers with recent credit inquiries are less likely to have a history of delinquencies. However, the correlation is weak, so further investigation is needed.
4. Debt-to-Income Ratio (DTI) and Credit Utilization: - dti (debt-to-income ratio) shows moderate positive correlations with open_acc (number of open credit lines) and total_acc (total number of credit lines). This suggests that borrowers with a higher number of open credit lines tend to have a higher DTI..revol_bal (revolving balance) has a moderate positive correlation with dti. This indicates that borrowers with higher revolving balances are more likely to have a higher DTI.
5. Interest Rate and Loan Term - int_rate has a strong positive correlation with term. This is expected as longer loan terms typically have higher interest rates.

CONCLUSION - RECOMMENDATIONS (TOP INFUENCERS)

- Loan Interest Rates - Most of the defaulters are between 11% to 16%, which indicates financial stress induced due to the higher interest rates. The company should strive to reduce the interest rates wherever possible to reduce probability of defaulting

- DTI - High DTI means applicant have more debt which translates to increased financial burden. Company carefully study the creditworthiness of such individuals who have high DTI. Individuals with 15% and above should be avoided.

- Annual Income - From provided data we see that most of the defaulters are in bracket of $40k-$60k. Company should evaluate these individuals who fall in this bracket of annual income.

- Loan Grades and Sub-Grades - Loans granted to B,C,D amount to majority of defaulters. So, the firm should exercise caution while granting loans to applicants falling in this category. Further, the company pay attention to applicants who fall in the subgrades B3, B4, B5, C1, C2 as they have tendency to default frequently.

- Loan Purpose - Loan Purpose of Debt-Consolidation is highest in the defaulters. Loan purposes like "educational," "renewable_energy," and "small_business" have a very low number of observations among charged-off loans. Firm should refrain from granting loans to individuals having debt_consolidation as purpose.

- Loan Term - Considering 36 months are defaulting more, so Company should device more evaluation while granting short term loan of 3 years.

CONCLUSION - RECOMMENDATIONS (ANCILLARY)

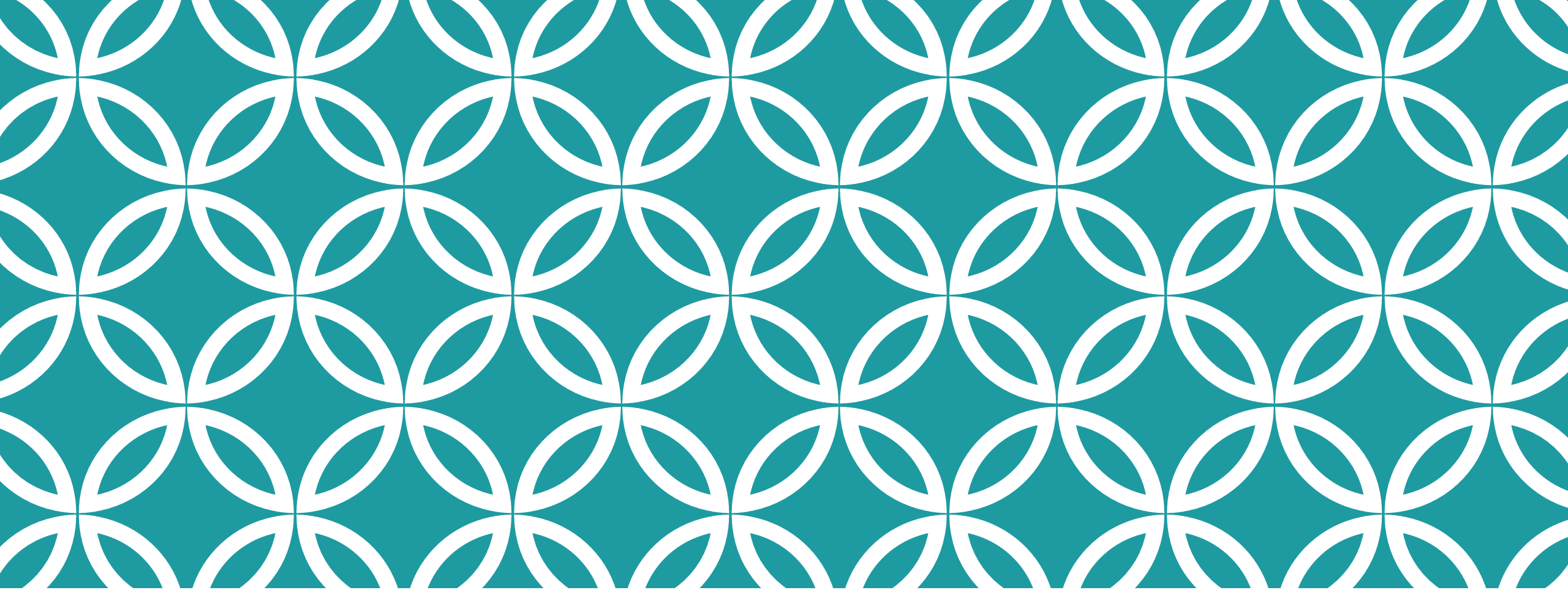
- Verification Status - It is interesting to note that the number of defaulters are more in case of "Verified" individuals. This indicates that company should strengthen the overall verification process.

- Employment Length - From the data it is evident that applicants who have 10+ years of experience tend to default more. This indicates that experience or employment legth may not give a very indicator or creditworthiness of the individual. Company should more closely evaluate the credit scores of such individuals.

- House Ownership - From the data it is clear that individuals in "Rented" house tend to default, indicating that such individuals find it difficult in managing Rent and Installments simultaneously.

- Geographic Indicators: Loan applicants from states like California (CA), Florida (FL), and New York (NY) are more likely to default. The company should check on regional risks, economic activities/trends and modify their loan strategies.

- Year and Month of Issue Date - It is evident to note that the defaulters are more from the 2011, along with the number of loan applicants which indicates that there is a good economic boost, however to minimise on the risks, company should make their approval process robust. Also it is interesting to note that the loans which are granted in last quarter of the year tend to default more, so the Firm should employ caution when granting loans during the holiday season.



THANK YOU

PG Program in Machine Learning & Artificial Intelligence (C71)

upGrad

iit-b
ज्ञानमुत्तमम्