# LENDING CLUB CASE STUDY

Submitted by:

Srishti Kedia
Kiran Bharat Pawar

# PROBLEM STATEMENT

- You work for 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:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

# DATA UNDERSTANDING

# OVERVIEW

- We have been provided with the Private Data of Lending Club. The complete loan data for all loans issued through the time period 2007 to 2011.

- We also have a data dictionary which describes the meaning of these variables.

- The dataset has 39717 rows and 111 columns.

# DATA FEATURES

- Each row in the data has a unique id for the loan listing and member_id for the borrower.

- Each row contains these data about the loan offered:
  - loan_amnt - Listed amount of the loan applied the borrower
  - funded_amnt - Amount committed for the loan
  - funded_amnt_inv - Amount committed by the investors for the loan
  - term - No. of payments of loan in months
  - int_rate - Interest rate on the loan
  - installment - Monthly amount that the borrower needs to pay
  - issue_d - The month and year in which loan was funded
  - dti - Ratio of monthly payment / monthly income
  - out_prncp - Remaining outstanding principal from total amount funded.
  - total_pymnt - Payment received to date for total amount funded.
  - total_rec_int - Interest received to date
  - total_rec_late_fee - Late fee received to date
  - recoveries - post charge off gross recovery
  - collection_recovery_fee - post charge off collection fee
  - last_payment_d - Last month payment was received
  - last_payment_amt - Last amount received
  - next_payment_d - Next scheduled payment date

# DATA FEATURES (CONT.)

- Various categorical information about the loan:
    - grade and sub_grade - Grade/ Quality score assigned to loan
    - loan_status - Current status of the loan.
    - payment_plan - Payment plan has been put in place or not.
    - desc - Loan description provided by the borrower.
    - purpose - Category of the loan request
    - title - Title of the loan request

- Each row contains various information about the borrower:
    - emp_title - Job title of the borrower
    - emp_length - Time the borrower has been employed
    - home_ownership - Home ownership status of the borrower
    - annual_inc - Annual income reported by the borrower
    - verification_status - Indication of whether income was verified or not
    - delinq_2y - No. of due incidents in borrower's credit file
    - mths_since_last_delinq - Months since last due incident
    - open_acc - No. of open credit lines in borrower's credit file
    - pub_rec - No. of derogatory public records
    - revol_bal - Total credit revolving balance
    - revol_util - The amount of credit the borrower is using relative to all available revolving credit.

# DATA FEATURES (CONT.)

- Geographic information about the borrower:
  - zip_code - Zip code provided by the owner
  - addr_state - State provided by the owner

- Available Loan statues:
  - Fully paid
  - Charged Off
  - Current

- Home ownership status has 4 values:
  - RENT
  - OWN
  - MORTGAGE
  - OTHER
- Various purpose of the loans are given which can help us in deciphering what causes the most defaulters.

- Verification status will tell us if the borrower's income was verified by the organisation or not.

# DATA CLEANING

# OVERVIEW

- There are various ways to clean the data:
    - Fix rows and columns
    - Fix missing values
    - Standardise values
    - Fix invalid values
    - Filter data

- We will use the above strategies to clean the data before starting the analysis

# FILTER DATA

- First, lets filter out "Charged Off" loans as we will only be dealing with such loans according to the problem statement
    - We have 5627 rows and 111 columns after the filtration.

- Remove all columns with greater than 65% missing percentage as it will skew the data.
    - We have 5627 rows and 55 columns after the filtration

- There are no duplicated rows in the dataset

- There are no empty rows in the dataset

- Check for all columns which have only 1 unique value in the dataset and remove them as they will not help in the analysis.
    - We have 5627 rows and 37 columns after the filtration

- Remove the columns which will not impact whether a borrower defaulted or not.
    - We have 5627 rows and 19 columns after the filtration.

- Columns after filtrations are: ‘loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp_length', 'home_ownership', “annual_inc', 'verification_status', 'issue_d', 'purpose', 'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'pub_rec_bankruptcies'

# FIX ROWS AND COLUMNS

- Convert the “term” values by replacing months for analysis.
  - Values after  fixing are
    - 60
    - 36
- Remove % and convert interest rate to float
  - Values after  fixing are like
    - 15.7 ..

- Round of values to 2 decimal points for float columns for easy interpretation

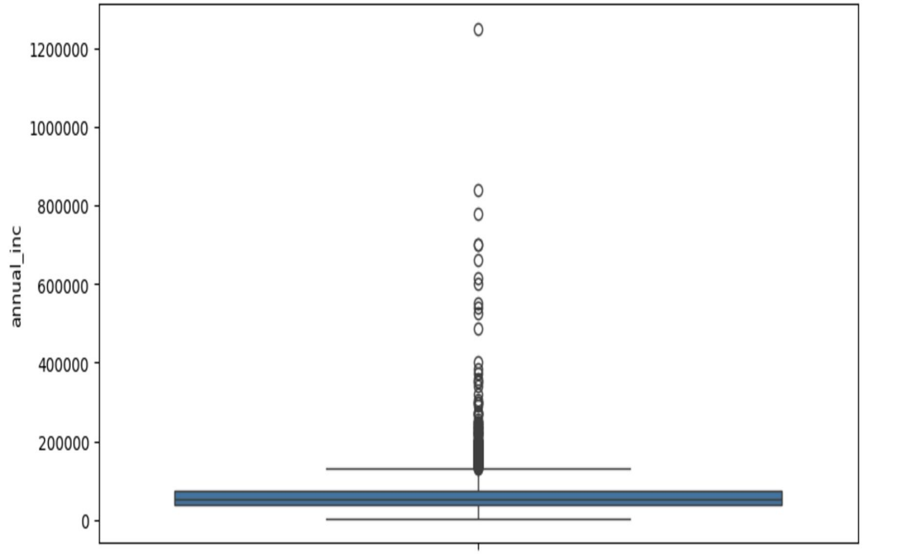
- Convert issue date to datetime object

- Convert employee length to int using a mapping where 10 means 10+ and 0 means < 1

# REMOVE OUTLIERS

- Checking if annual income has outliers..
  - As we can see the max and min have huge differences from IQR which can be supported by the boxplot.
  - Remove the outliers from the annual income. (Anything outside 1.5IQR)
  - We have 5367 rows and 19 columns

- The dti ratio most values are b/w 8 to 18 does not seem to have outlier.



# STANDARDISE VALUES

- Employee Length column has 225 missing values. Let's impute it.
  - Find out annual income of employees with null emp_length
  - Median is 36000
  - Mode of employee length is 10
  - Let's fill emp_length with 10 as these people tend to have high income and are mostly business owners.

- Verification status has 3 categories: Sources Verified, Verified and Not Verified.
  - We can combine Verified and Source Verified into one category.

- Impute pub_rec_bankruptcies with -1 to not consider them.

- Now the data is fully cleaned with no missing values.

# DERIVED VALUES

- Derive New columns to help in analysis.

- Separate out the issue year, quarter and month columns
  - Issue_year
  - Issue_month
  - Issue_quarter

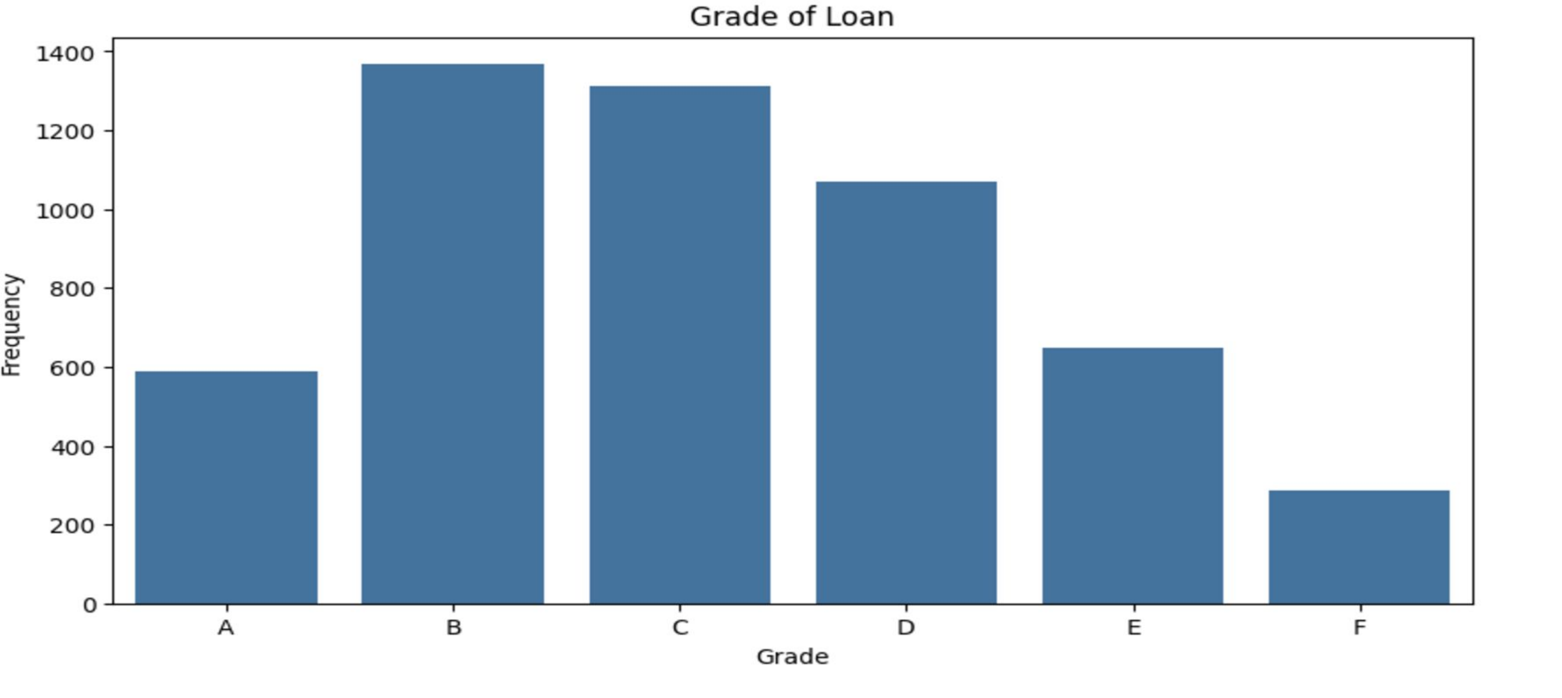
- Divide the float columns into multiple ranges for easy analysis and interpretation
  - Loan Amount : loan_amt_range( 0-5000, 5000-10000...)
  - Interest Rate: int_rate_range(0-2, 2-4...)
  - Annual income: annual_inc_range(0-20000, 20000-40000...)
  - Debt to income ratio: dti_range(0-2, 2-4, 4-6..)

# ANALYSIS

# UNIVARIATE ANALYSIS

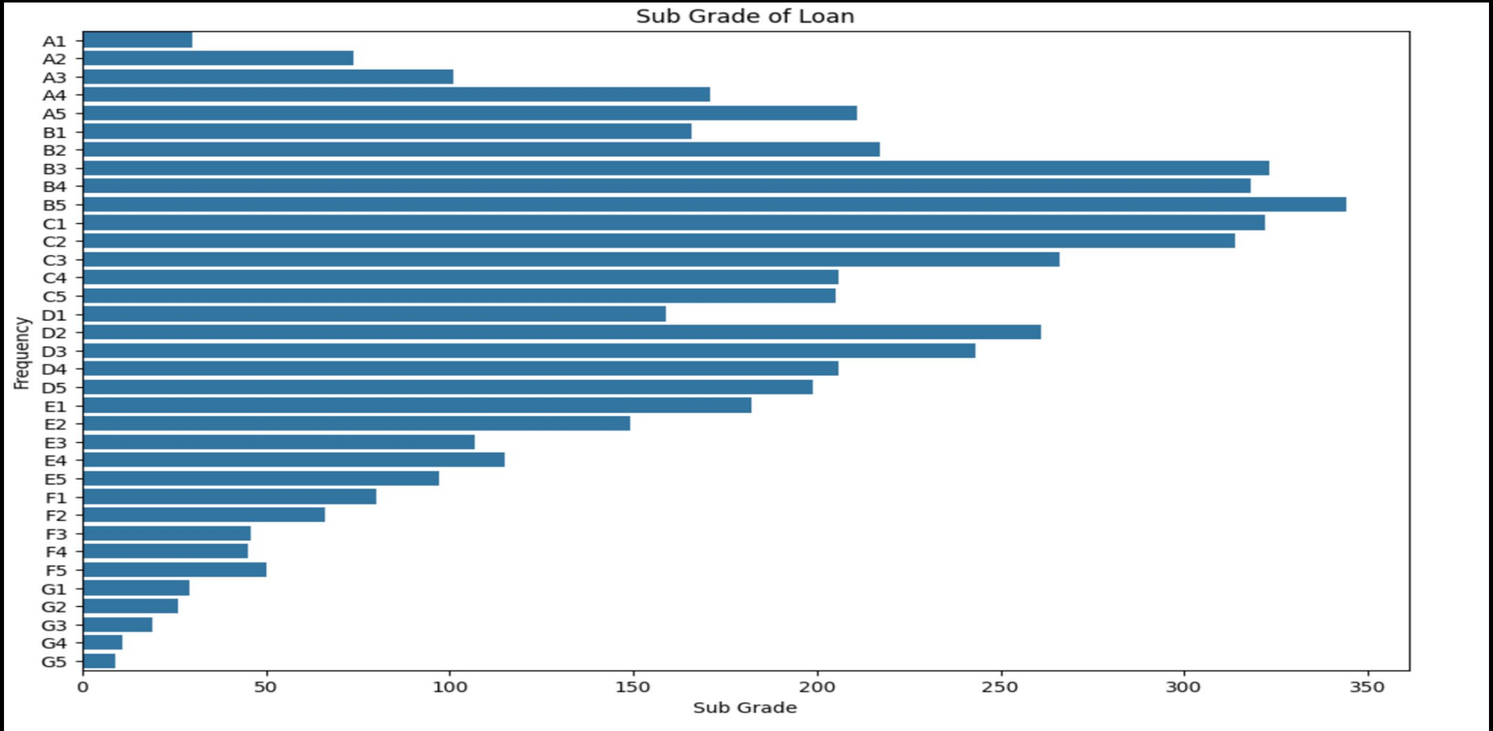
# ORDERED CATEGORICAL VARIABLES

# GRADE



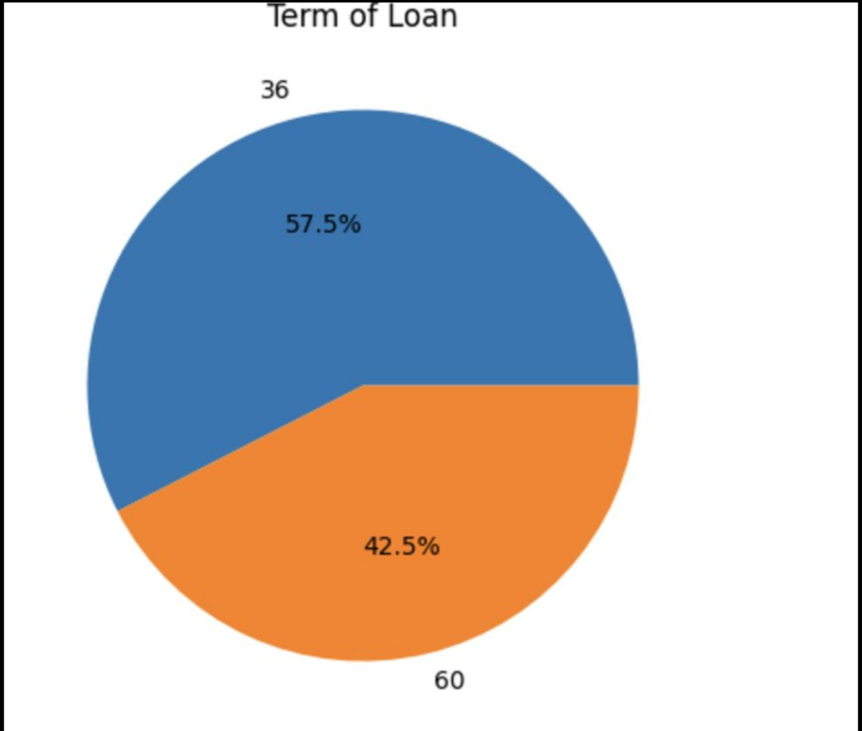
Grade B and C have highest number of defaulters indicating that such loan grades contribute to more defaulters.

# SUBGRADE



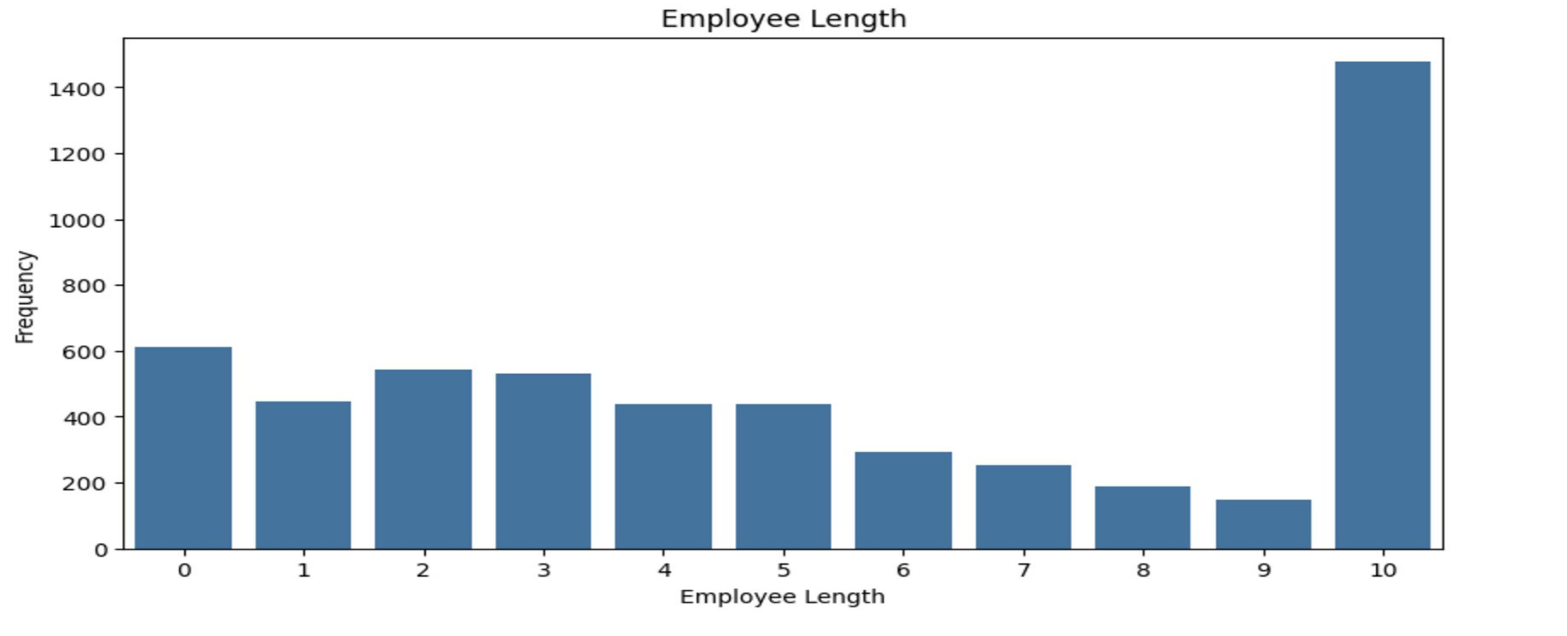
The same interpretation from the previous graph that sub grades of B have the highest defaulter with the highest defaulter being of "B5" sub-grade

# TERM OF LOAN



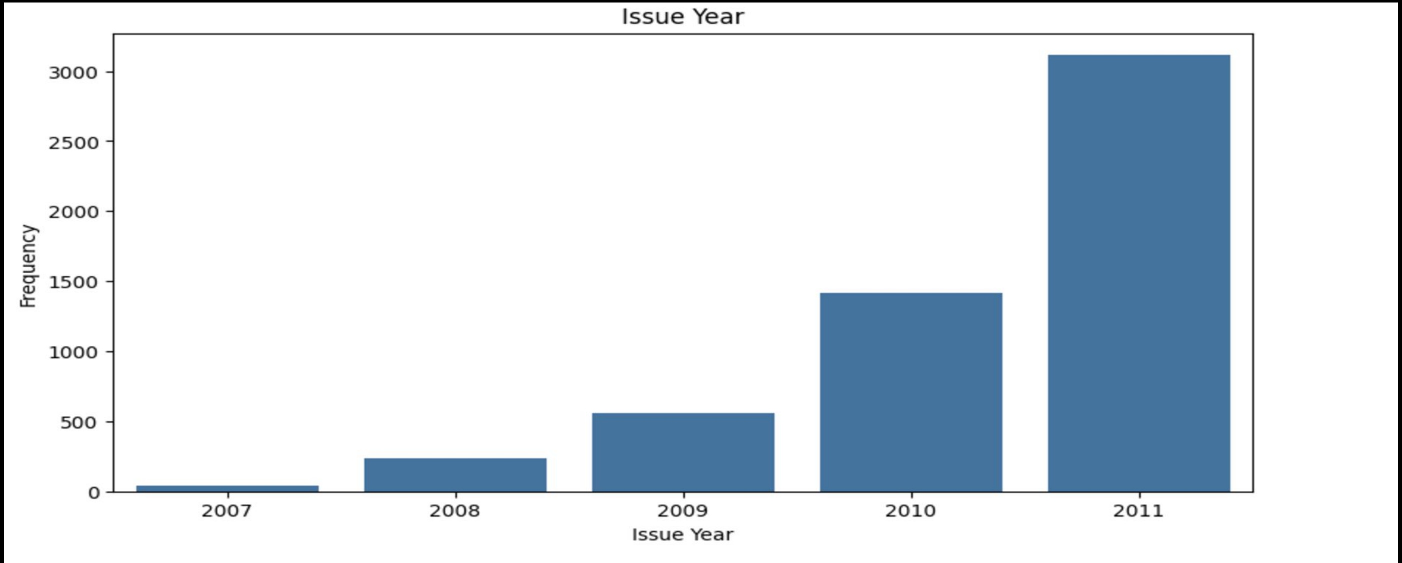
As we can see, short term loans have more defaulters, which indicates that people fail to pay more often when they take loan for a shorter term.

# EMPLOYMENT LENGTH



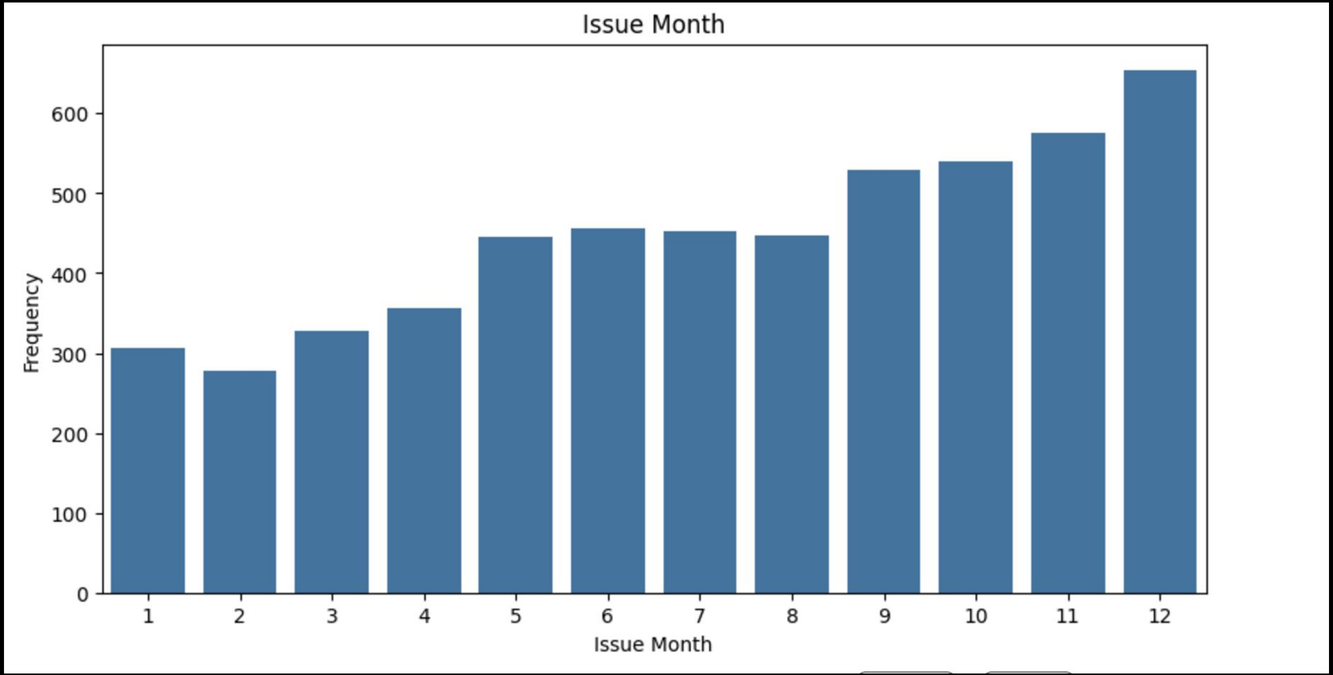
Employees with the highest work duration are defaulting the most. This is surprising and is an important note for the organisation to not consider this a huge bonus when approving a loan.

# ISSUE YEAR



Year 2011 has the most number of defaulters. This could indicate financial changes / recession during the year.

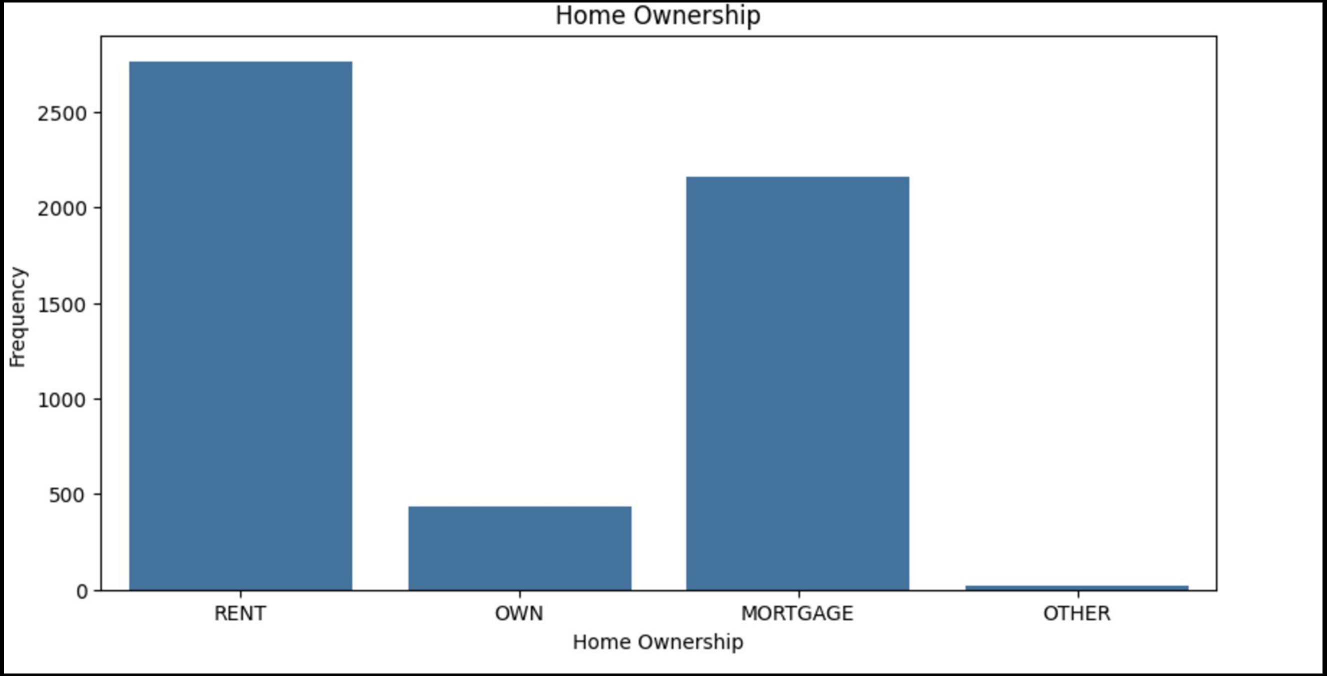
# ISSUE MONTH



Loans taken in the final quarter(Q4) especially dec has the highest number of defaulters indicating holiday Spending leading to financial strain.

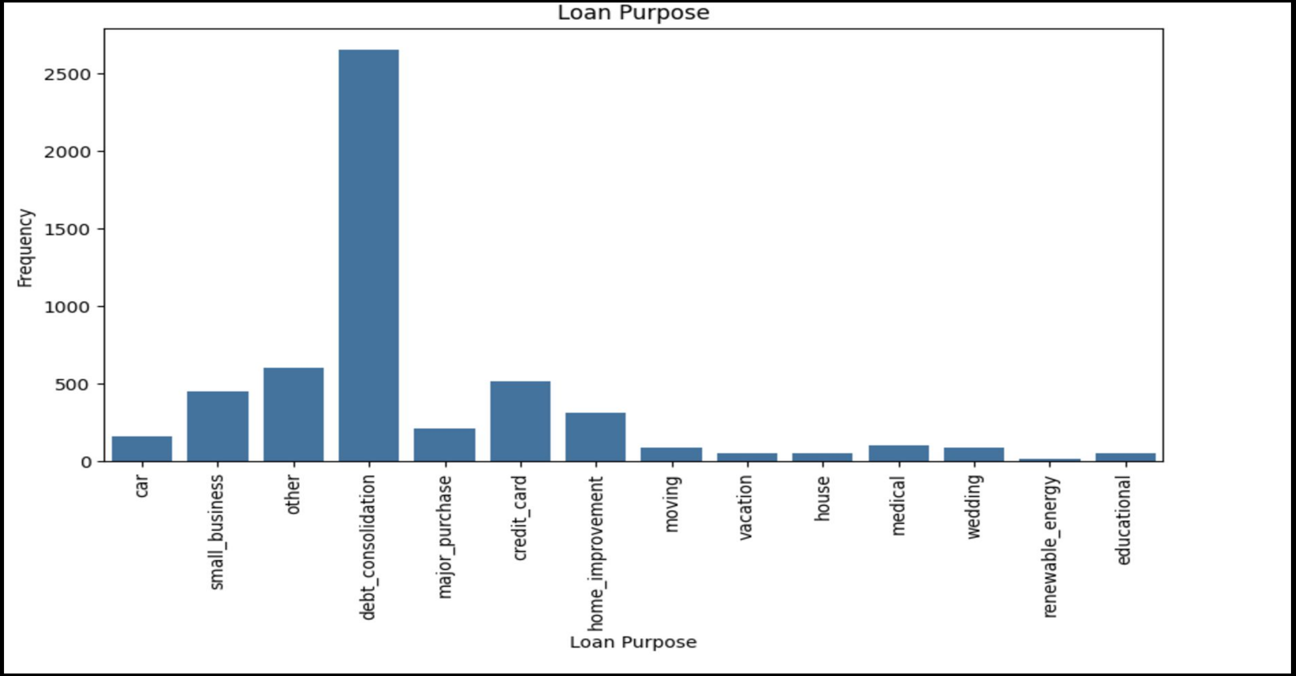
# UNORDERED CATEGORICAL VARIABLES

# HOME OWNERSHIP



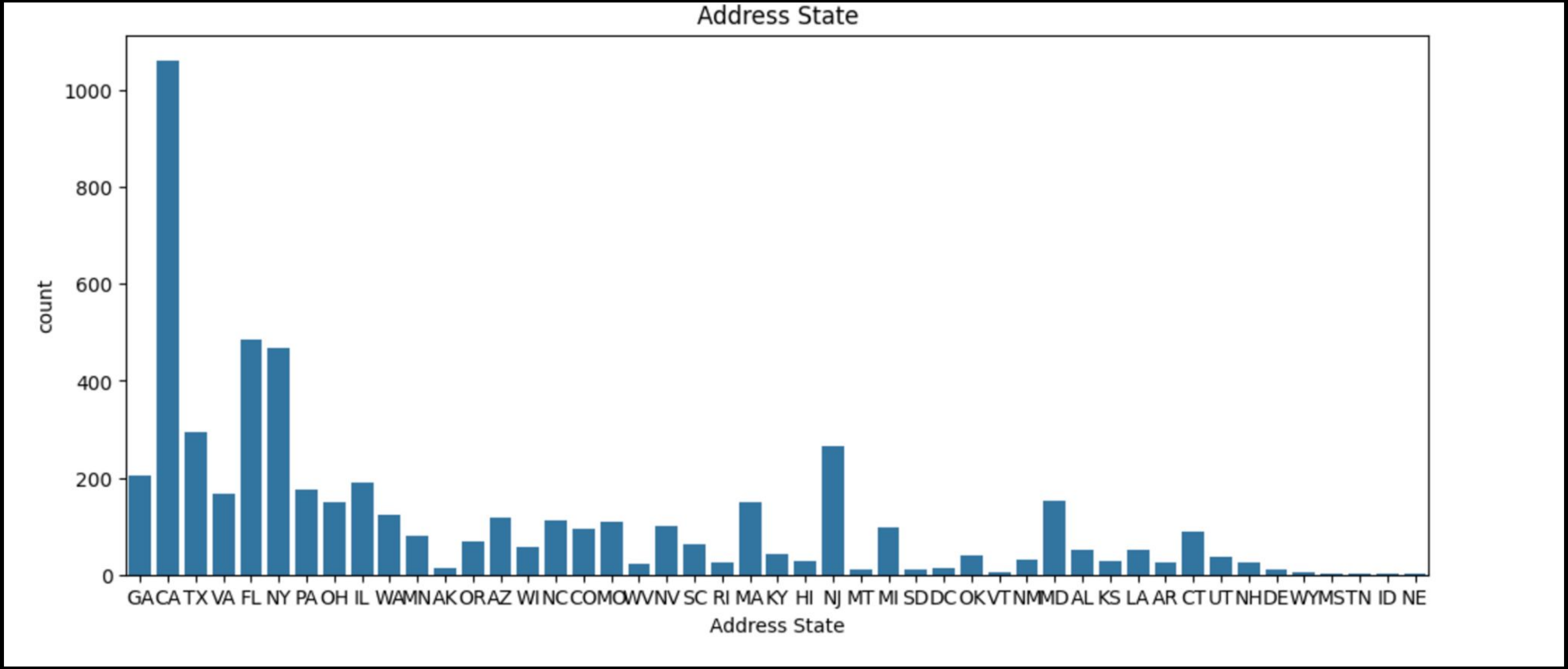
Borrowers living in rented / mortgaged homes form majority of defaulters. This is a strong factor to assess for the company before approving a loan.

# LOAN PURPOSE



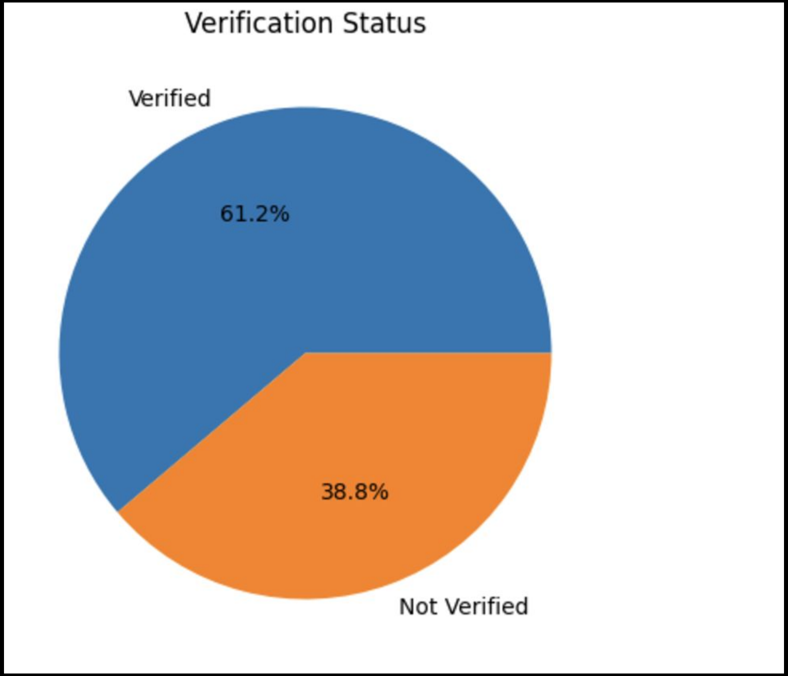
Applicants who have Debt consolidation purpose form the highest number of defaulters. Lending companies need to assess their portfolios carefully.

# ADDRESS STATE



California(CA) has the highest number of defaulters. Stricter assessments needs to be done for this state.

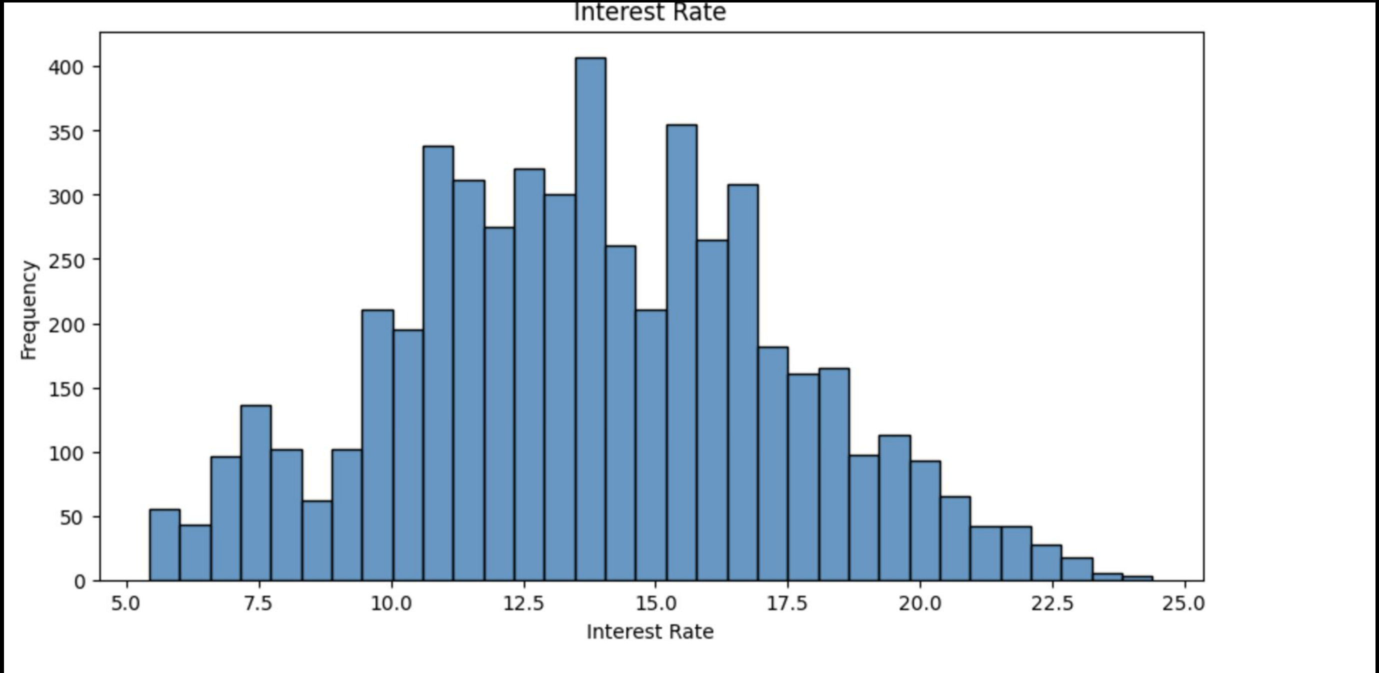
# VERIFICATION STATUS



The verified ones form higher % of defaulters. This means the verification is not proper and lending company should take strict actions against the verification department.

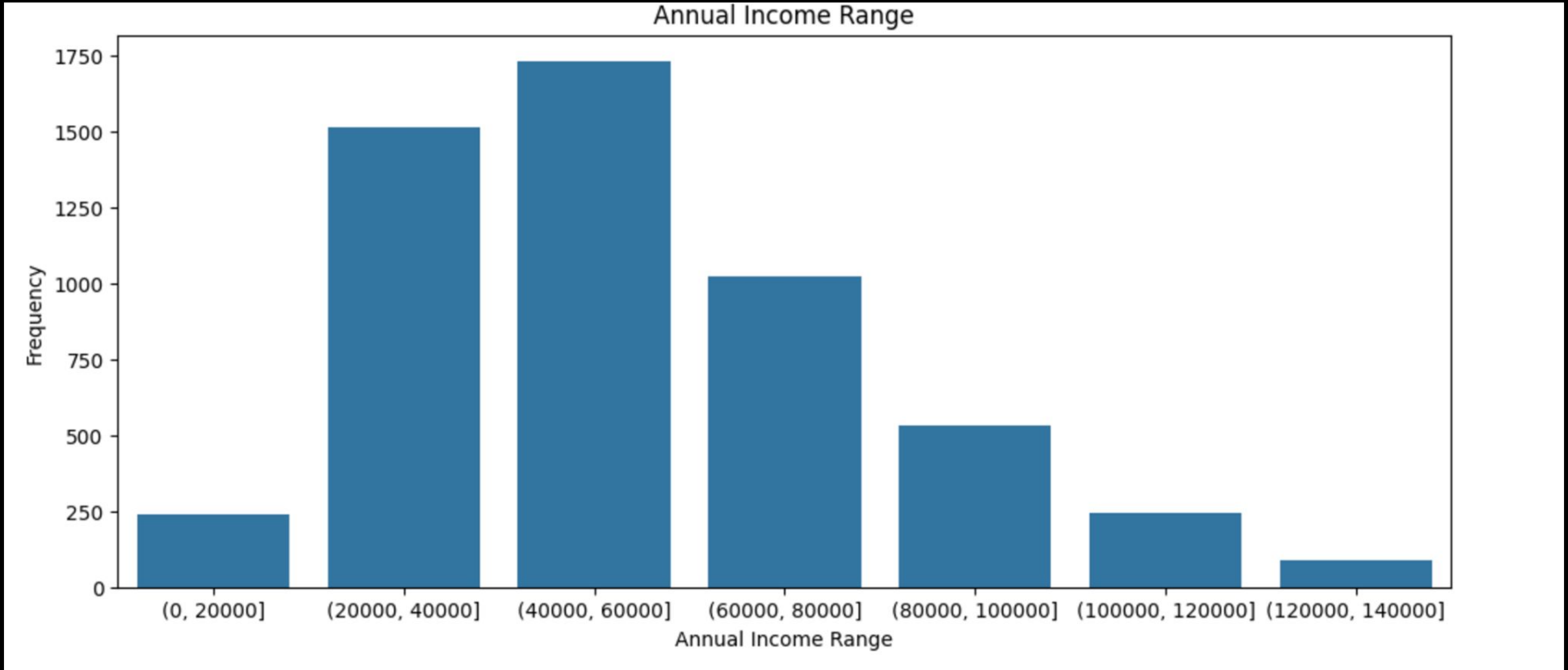
# QUANTITATIVE VARIABLES

# INTEREST RATE



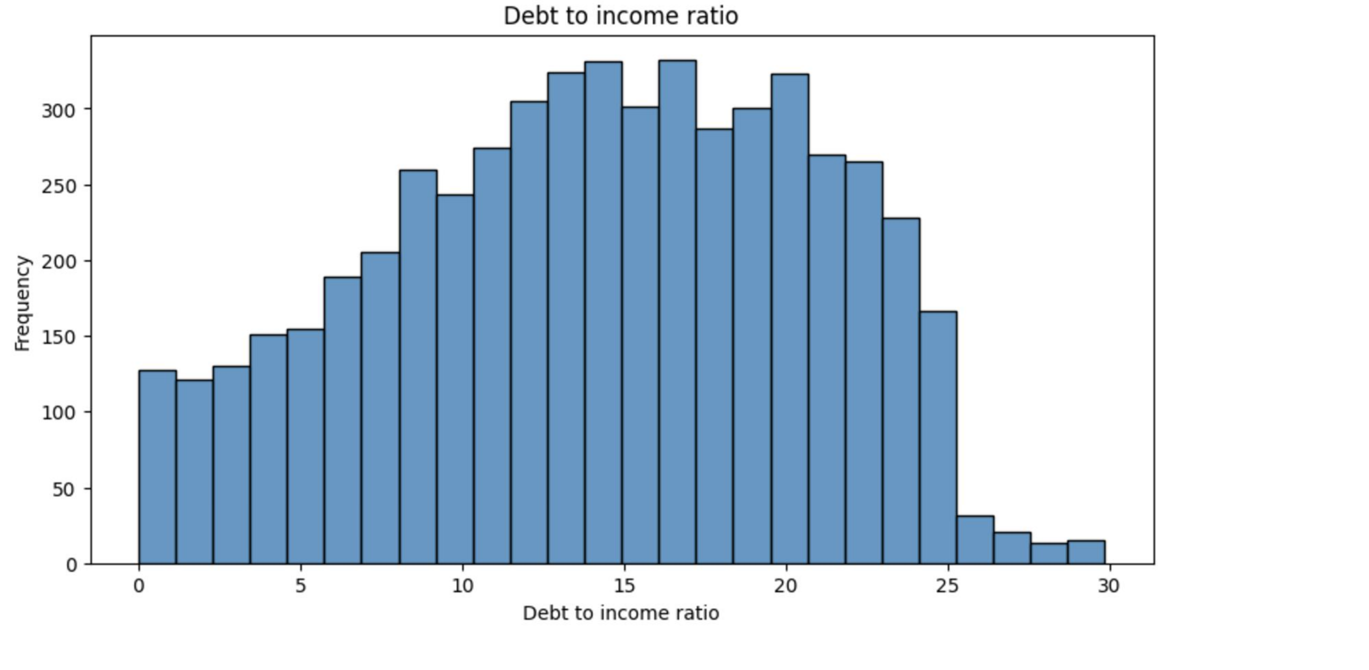
A considerable portion belonged to the interest rate bucket of 13%-17%. To reduce the risk of default, the lending company should consider offering loans at lower interest rates when possible.

# ANNUAL INCOME



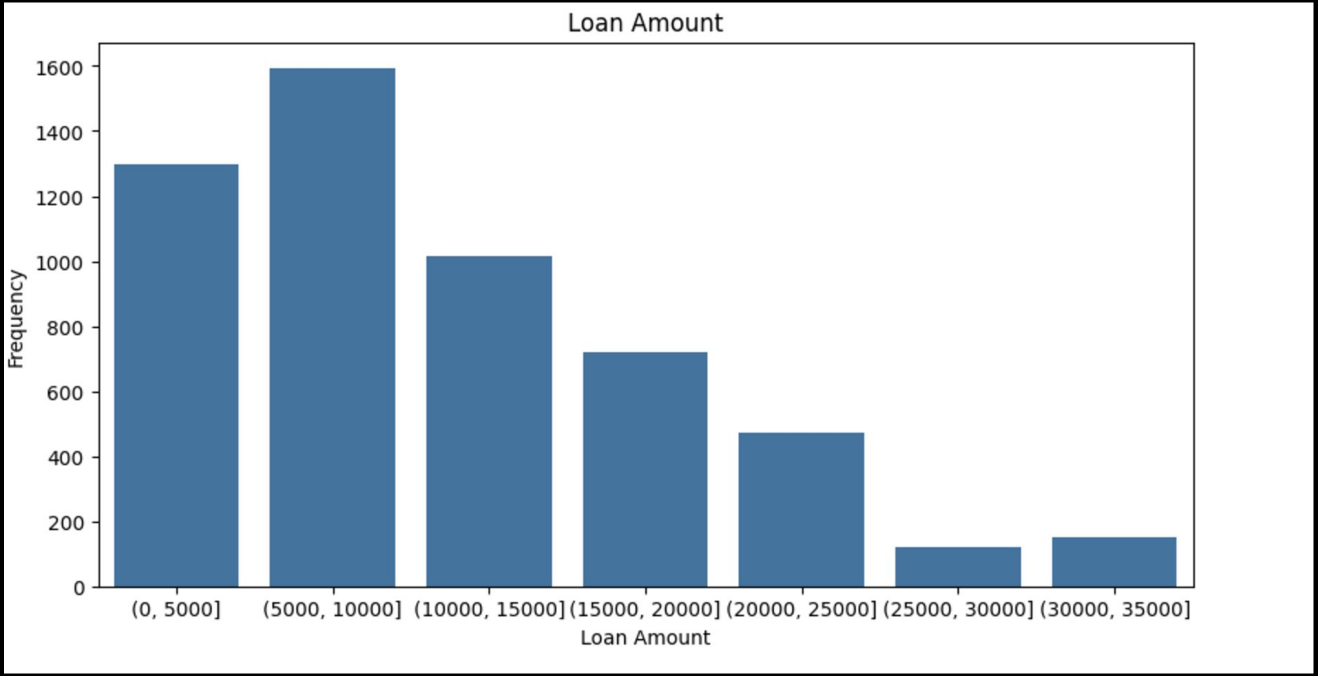
Borrowers with annual income b/w 20000 and 60000 form the most number of defaulters. So, the company should prioritise people with more income when providing loans to avoid risk.

# DEBT TO INCOME RATIO



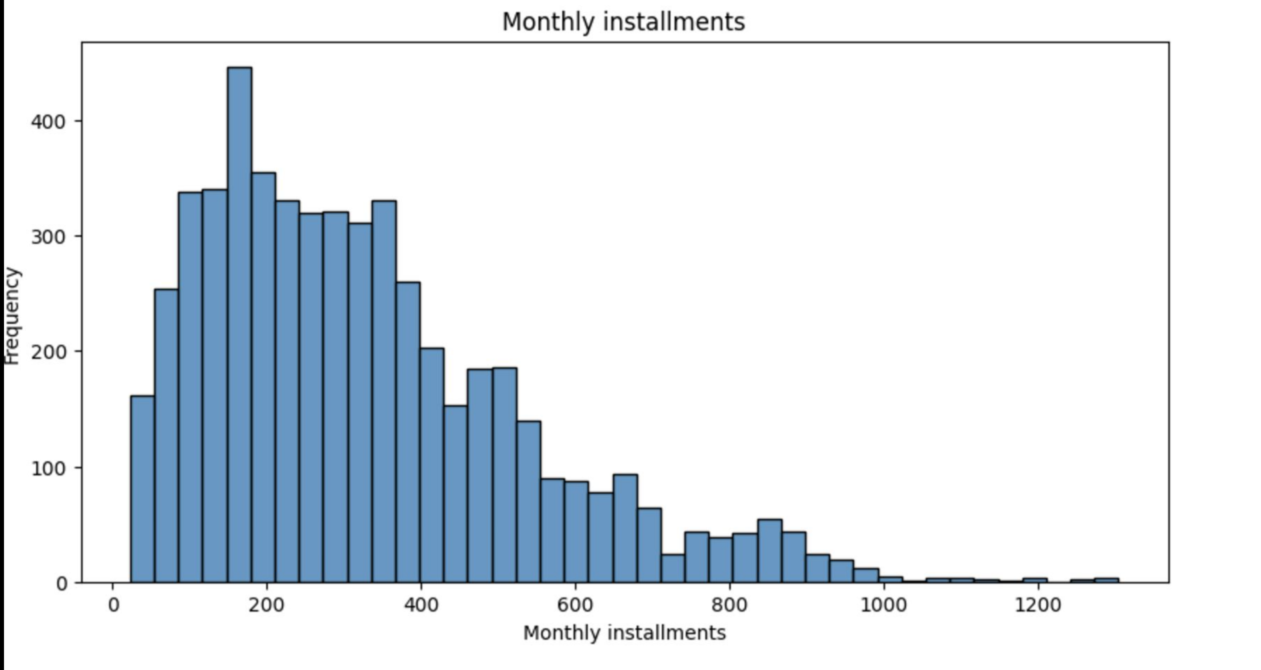
People with very high debt-to-income form the highest number of defaulters. There should be stricter cut off of dti ratio of the borrowers.

# LOAN AMOUNT



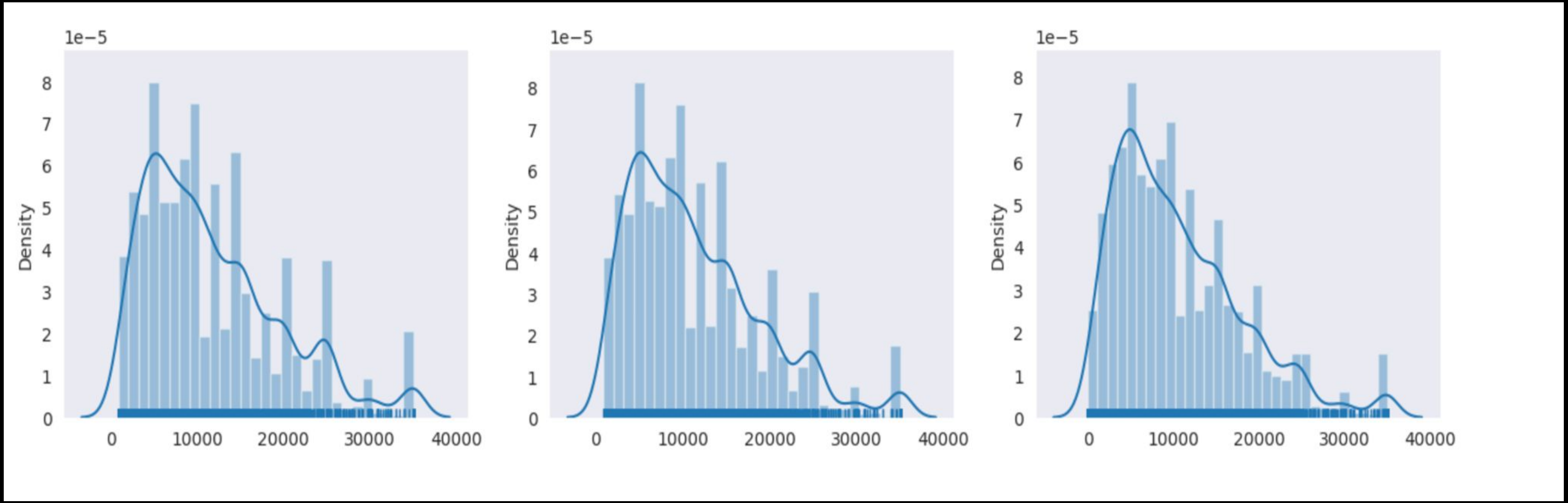
Surprisingly, lower loan amounts form highest % of defaulters. Generally, smaller loans are provided to people with lower credit score, which could be a reason for the loan not getting paid.

# MONTHLY INSTALLMENTS



Loans with a monthly installment of around 200 USD have the most defaulters. This could again indicate lower credit score borrowers.

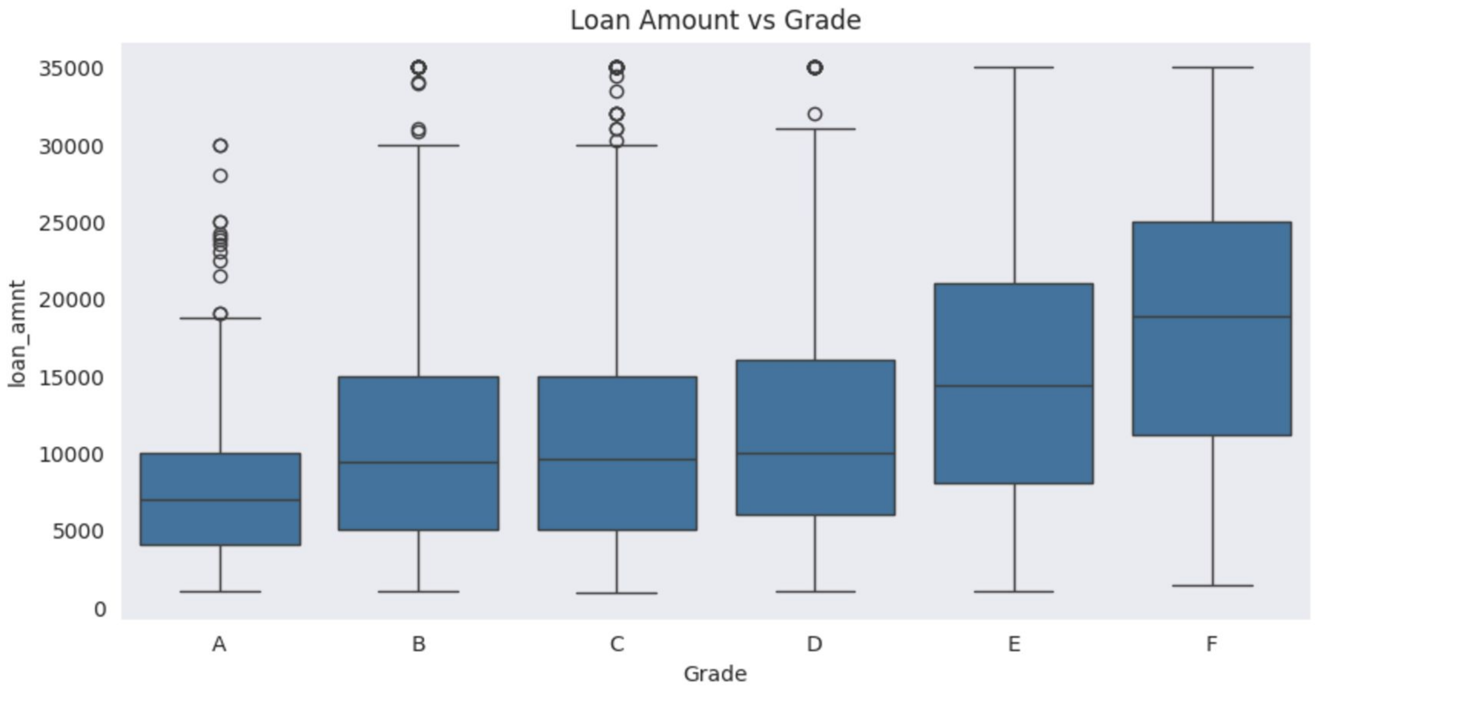
# LOAN AMOUNT & FUNDED AMOUNT



As we can see distribution plot for loan amount, funded amount and funded amount inv is exactly the same, so we will only use loan_amnt out of the three for further analysis.

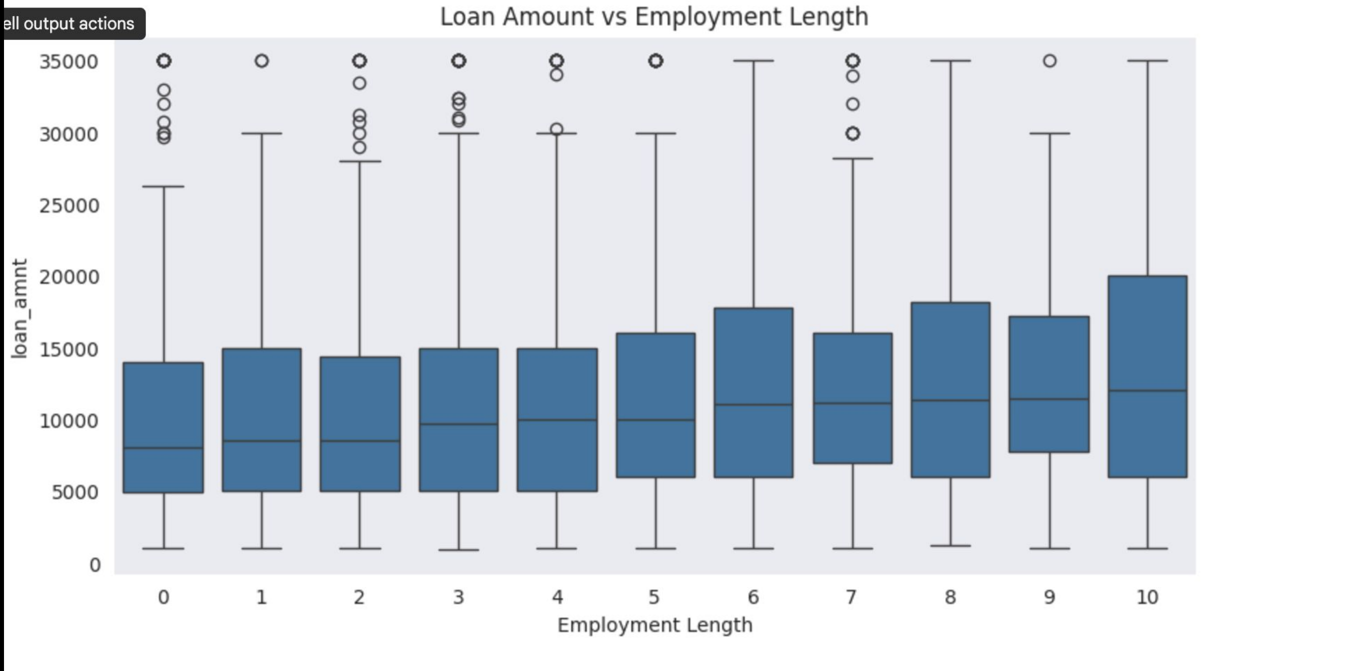
# BIVARIATE ANALYSIS

# LOAN AMOUNT VS GRADE



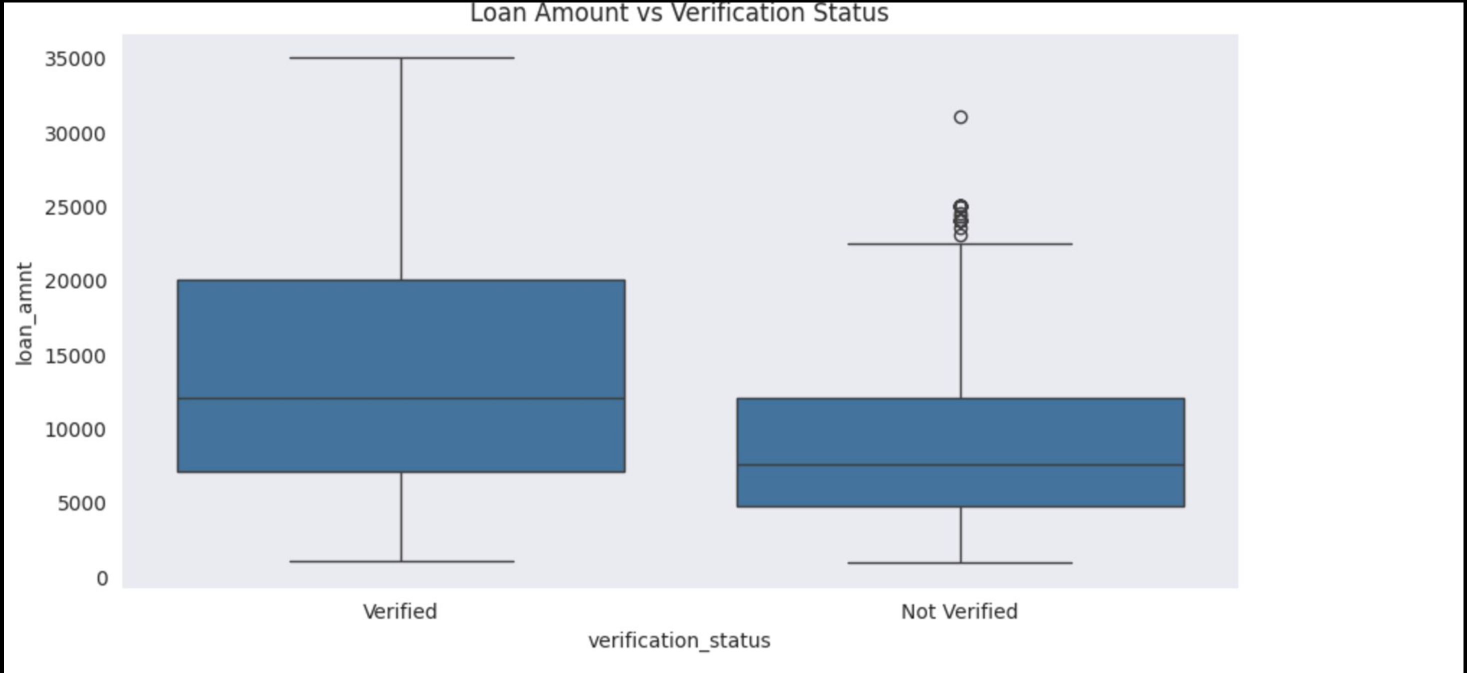
Higher loan amounts tend to have worse grade which can imply bad credit scores, high-risk browser profile. So the company should carefully assess lending high amount to borrowers with bad LC grade.

# LOAN AMOUNT VS EMPLOYEE LENGTH



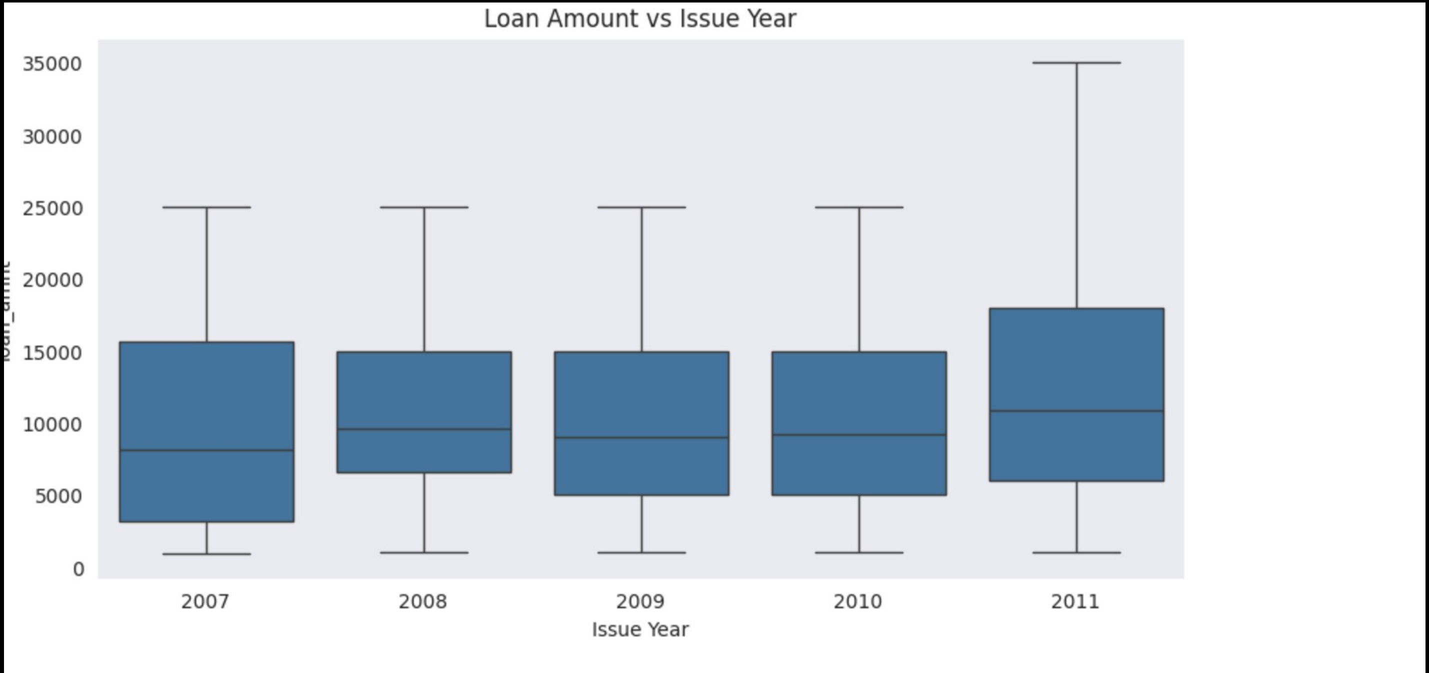
As we see the loan amount does not vary widely with borrower's employment length since the median loan amount of all employment length is b/w 7500 and 10000

# LOAN AMOUNT VS VERIFICATION STATUS



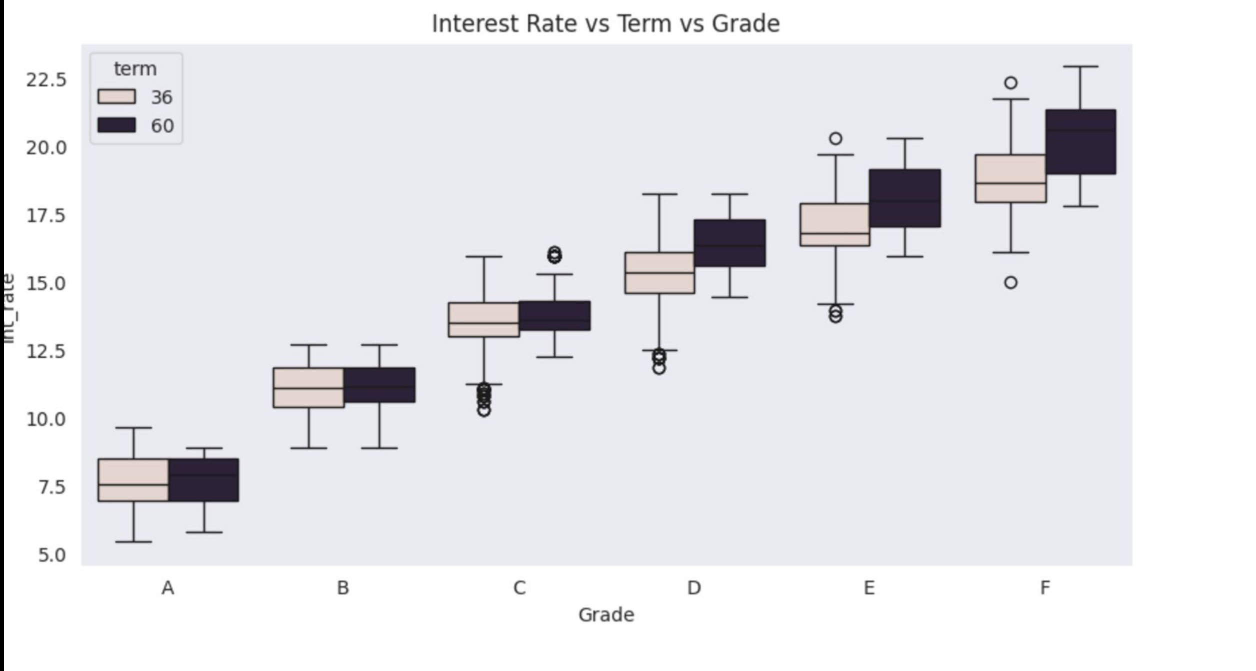
As we can see in the graph, higher loan amounts have more counts and are verified more. This could explain more verified defaulters. As higher loans tend to be riskier and verified easily by the company.

# LOAN AMOUNT VS ISSUE YEAR



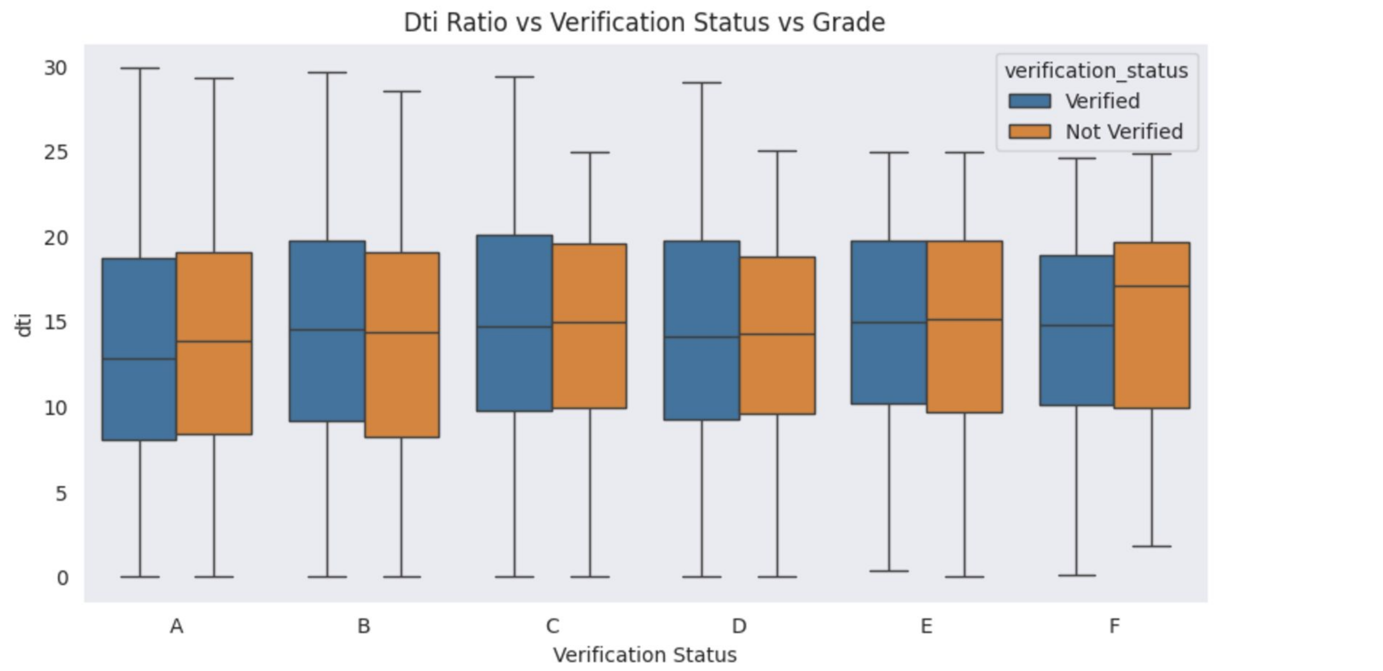
Borrowers in 2011 have applied for higher loan amounts as compared to other years. This could again prove previous issue_y analysis that some recession / financial challeneges would have occured in this year

# INTEREST RATE VS TERM VS GRADE



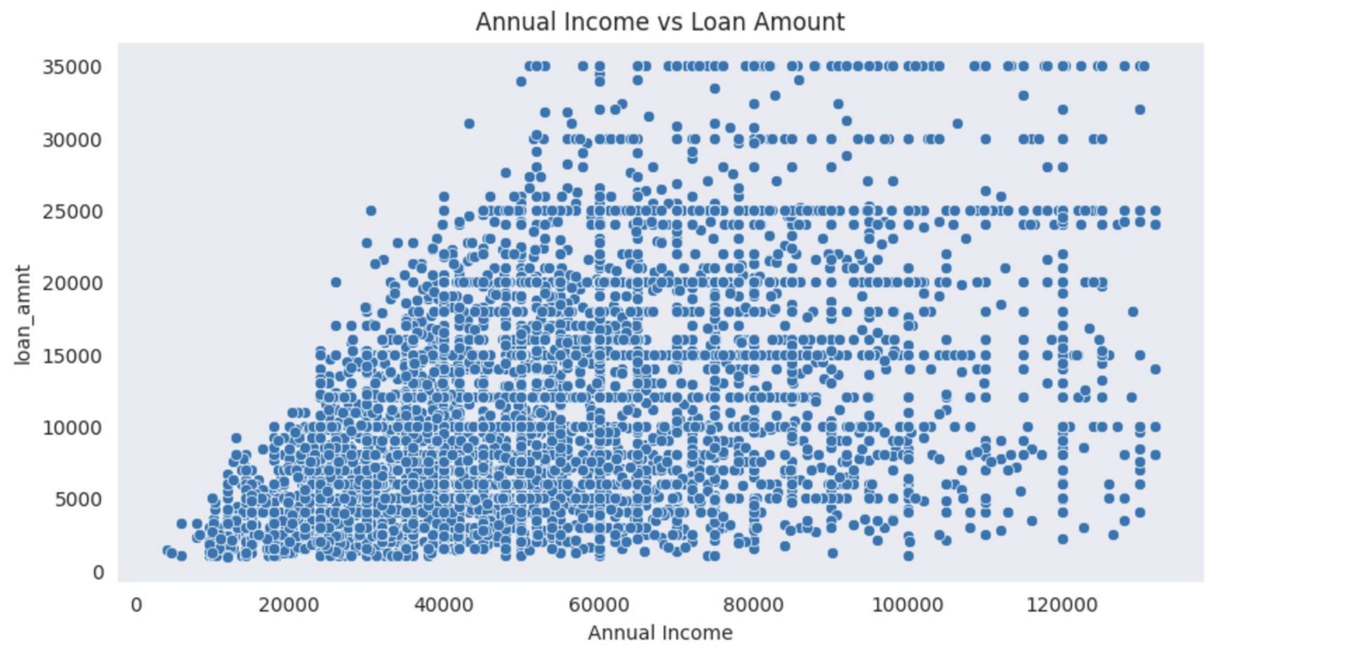
As we can see from the chart above, higher interest rates are for larger term ( 60 months) and have worse LC grade.

# DTI RATIO VS VERIFICATION STATUS VS GRADE



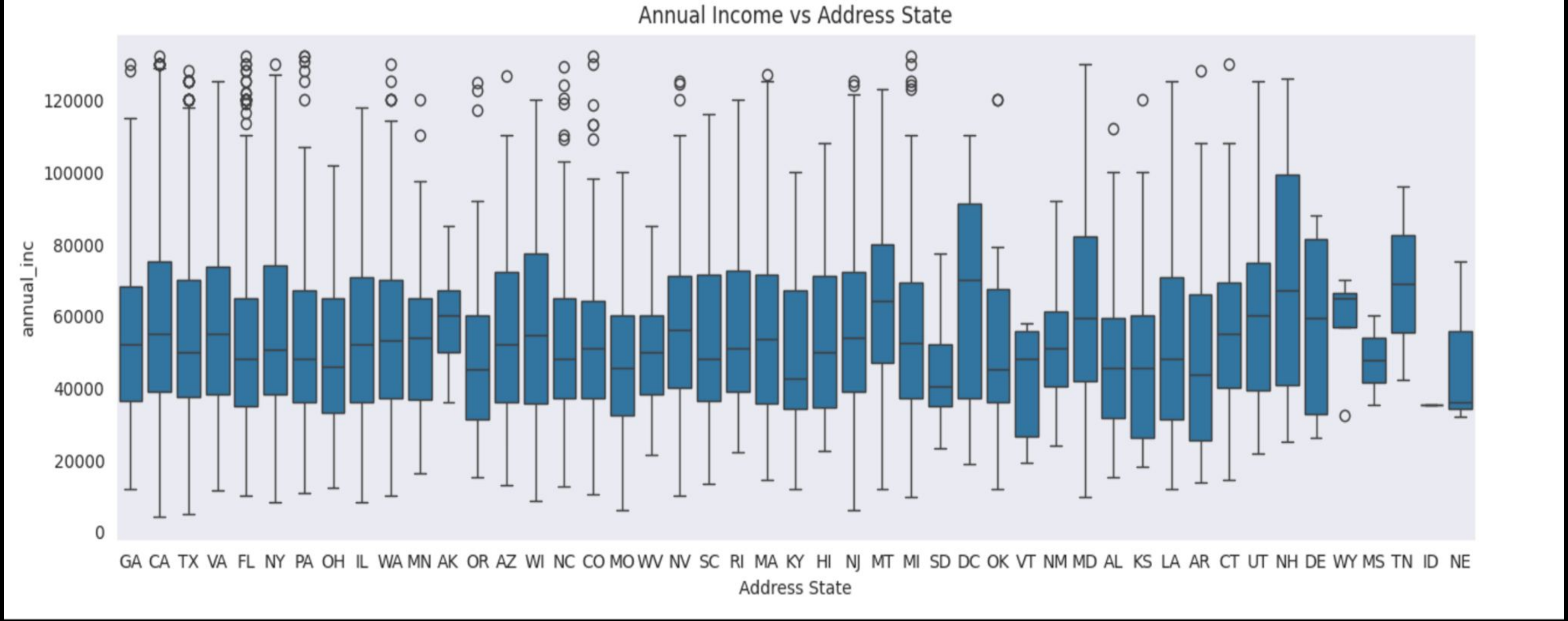
Debt to Income ratio does not vary widely with grade or verification status

# ANNUAL INCOME VS LOAN AMOUNT



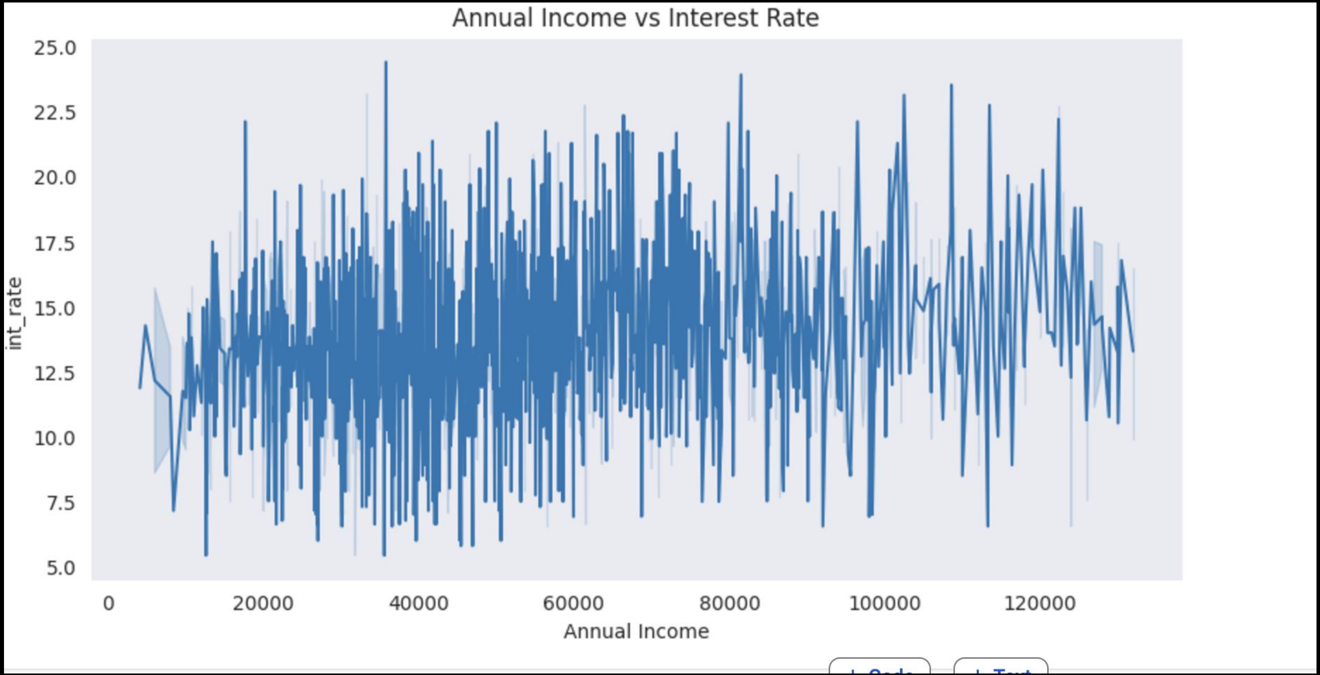
There are borrowers with annual income lower than 60000, taking loans of 30000 or more which is risky and must be assessed

# ANNUAL INCOME VS ADDRESS STATE



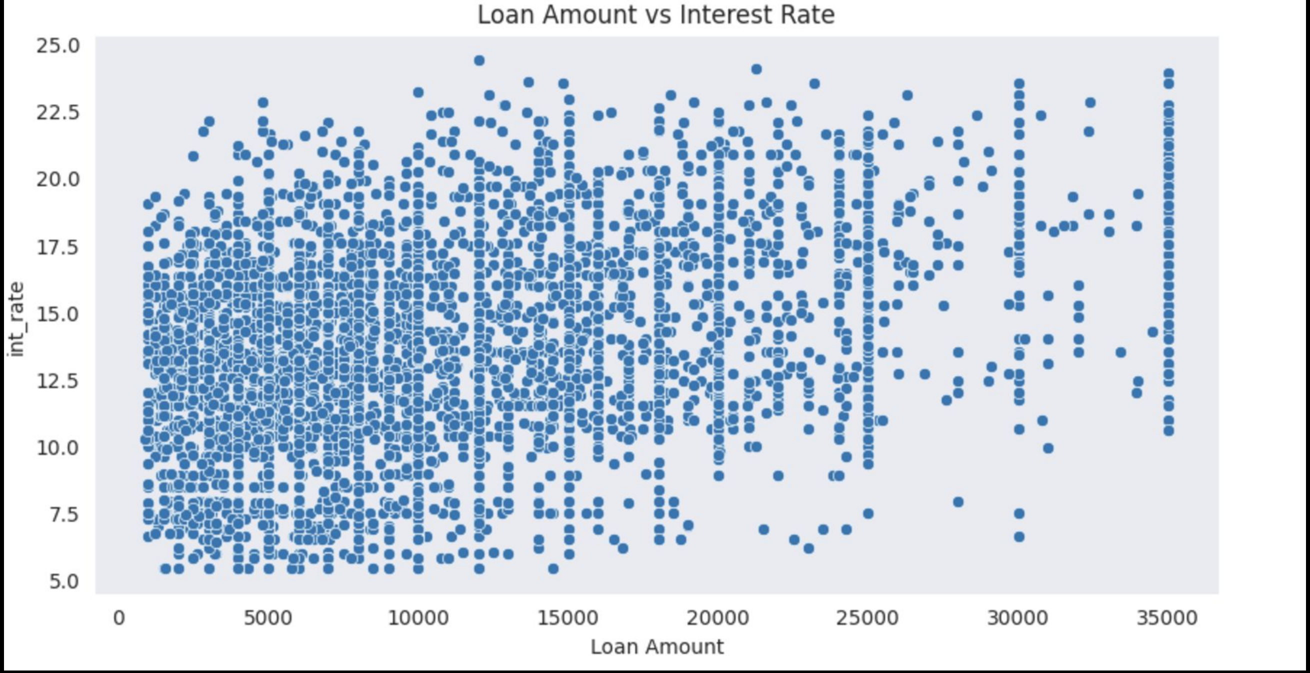
Washington DC and New Hemisphere have higher annual income indicating more employment in these states.

# ANNUAL INCOME VS INTEREST RATE



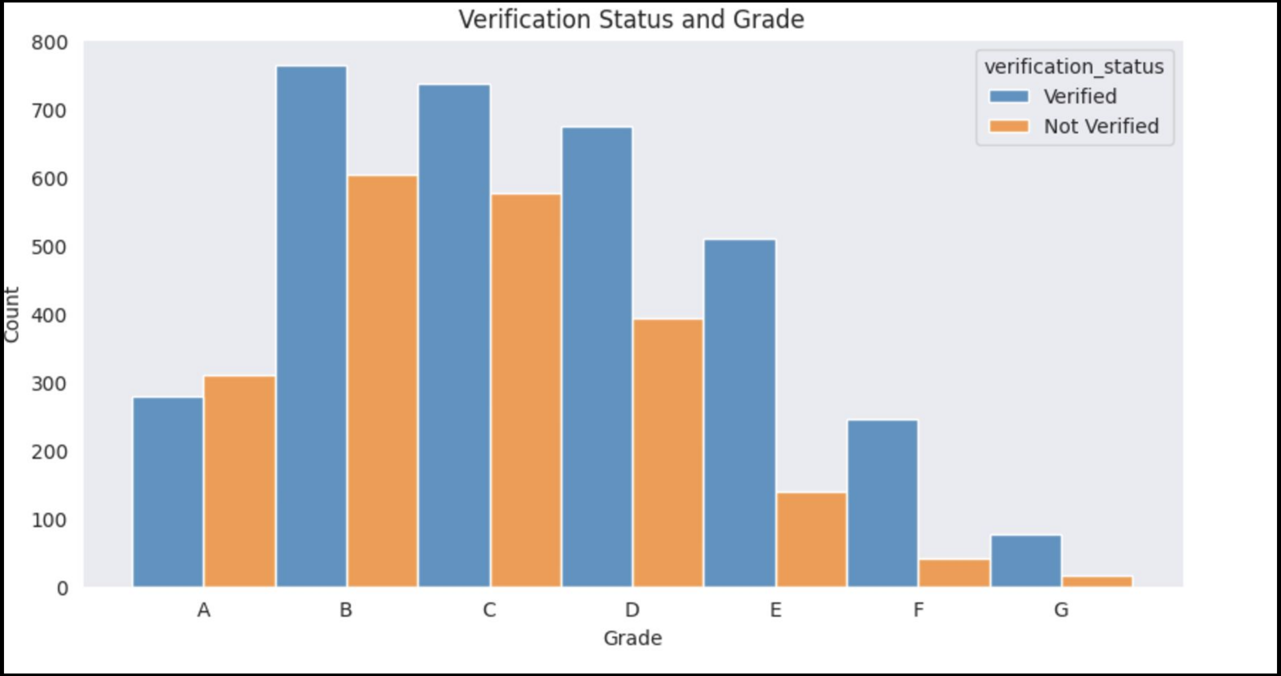
There is no clear pattern b/w annual income and interest rates. But very low income are given very low interest rates.

# LOAN AMOUNT VS INTEREST RATE



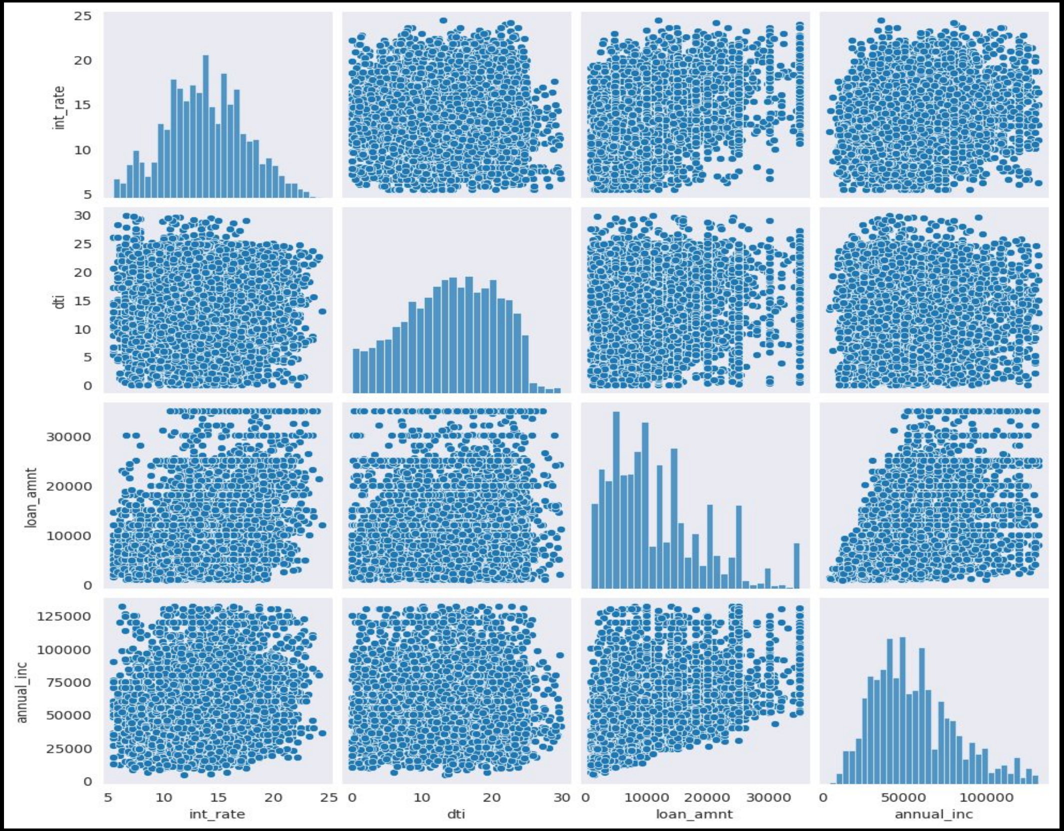
As the loan amount increases, interest rate increases, but the most number of defaulters have extremely low loan amount. This could indicate defaulters with very low credit scores.

# VERIFICATION STATUS VS GRADE



For good grade i.e A, there are more non verified defaulters than verified ones, but for remaining ones the number of verified defaulters is more than that of non verified.

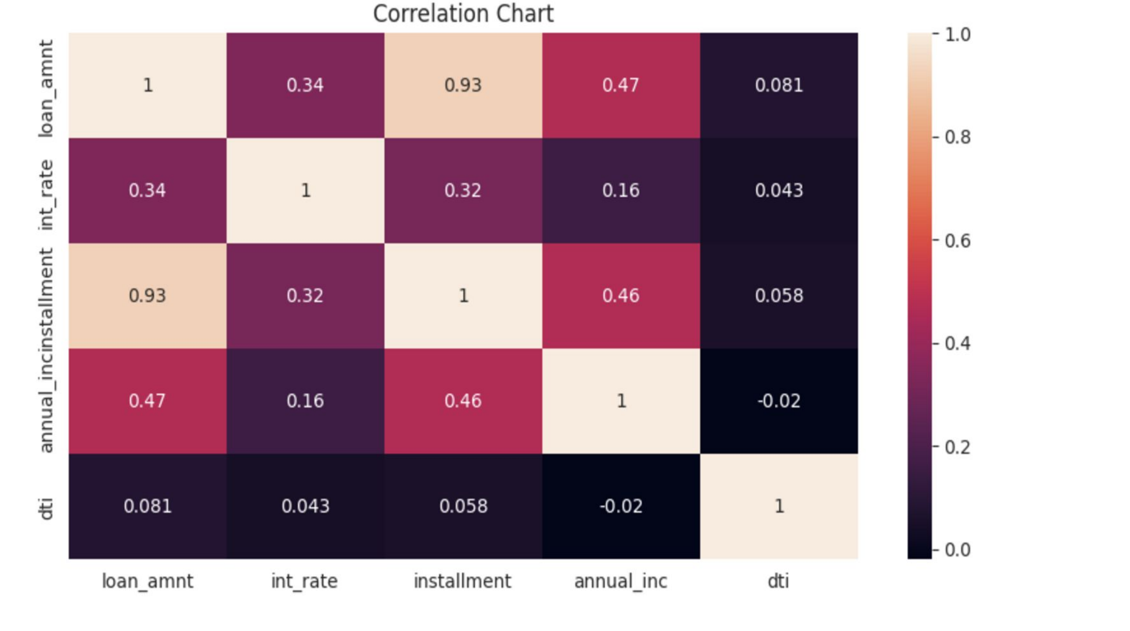
# PAIR PLOT OF ALL QUANTITATIVE VARIABLES



Relationship between various quantitative variables can be figured out from the above graph:

1) Interest rate and loan amount are directly related, as loan amount increases, interest rate increases
2) Loan amount and annual income are also directly related with few exceptions which must be taken into account for better risk management.
3) Annual income and dti ratio are inversely related but few exceptions/ outliers exist which must be taken into account for better risk management.

# CORRELATION B/W QUANTITATIVE VARIABLES



Correlation chart verifies the above points:

1) Annual income and dti ratio are inversely related, as annual income increase, dti ratio decreases
2) Installment rate and loan amount are highly positive correlated which means larger loan amounts have bigger terms.
3) Loan amount and annual income also have a positive correlation but its less than 0.5 which is a concern.
4) Installment and annual income are positive correlated too which means people with higher annual income opt for higher installments.

# KEY ANALYSIS AND RECOMMENDATIONS

- Grade B has very high defaulter count.
  - This could indicate borrowers with average credit scores
  - So borrowers with this grade should be assessed properly before approving their loan.

- Short term loans ( 36 months) have more defaulters.
  - Company should encourage borrowers to take long term loans and educate them about this.

- Borrowers who have been employed for more than 10 years contribute to most number of defaulters.
  - Higher employment length should not be given a lot of weightage when approving loans.

- Year 2011 has the maximum defaulters indicating recessions.
  - Any recession / financial challenges in a year should be predicted the company.
  - Stricter regulations should be implemented in such years for approving a loan.

- Quarter 4 and December month has the maximum defaulters.
  - This could indicate more financial burdens and holiday spending.
  - Company should assess any loans approved more carefully in Q4

# KEY ANALYSIS AND RECOMMENDATIONS (CONT.)

- Borrowers living in rented / mortgaged homes form majority of defaulters.
    - This could indicate more financial burdens for such borrowers
    - This is a strong factor to assess for the company before approving a loan.

- Borrowers  whose purpose is Debt consolidation form the highest number of defaulters.
    - Such profiles should be assessed thoroughly and avoided if any risk is found.

- California has the highest number of defaulters
    - Company should modify regulations based on state's defaulter count.
    - Stricter regulations need to be done for California

- Verified borrowers are defaulting more than non verified borrowers.
    - This could indicate faulty verification process
    - Verification team of the company should be questioned and the process needs to be reassessed.

- A considerable portion belonged to the interest rate bucket of 13%-17%.
    - To reduce the risk of default, the lending company should consider offering loans at lower interest rates when possible.

# KEY ANALYSIS AND RECOMMENDATIONS (CONT.)

- Borrowers with annual income b/w 20000 and 60000 form the most number of defaulters.
  - So, the company should prioritise people with more income when providing loans to avoid risk.

- People with very high debt-to-income form the highest number of defaulters.
  - There should be stricter cut off of dti ratio of the borrowers.

- Surprisingly, lower loan amounts form highest % of defaulters.
  - Generally, smaller loans are provided to people with lower credit score, which could be a reason for the loan not getting paid.
  - Borrower's credit score should be assessed carefully and financial education should be provided.

- Loans with a monthly installment of less than 200 USD have the most defaulters.
  - This could again indicate lower credit score borrowers and verify the above point.

- Higher loan amounts tend to have worse grade which can imply bad credit scores, high-risk browser profile.
  - So the company should carefully assess lending high amount to borrowers with bad LC grade.

- Higher loan amounts have more counts and are verified more.
  - This could explain more verified defaulters ass higher loans tend to be riskier and are verified easily by the company. Stricter regulations need to be implements for higher loan amounts

# KEY ANALYSIS AND RECOMMENDATIONS (CONT.)

- Higher interest rates are for larger term ( 60 months) and have worse LC grade.
    - Bad LC grade should be considered as an important factor when approving loans.

- There are borrowers with annual income lower than 60000, taking loans of 30000 or more which is risky and must be assessed
    - Stricter cut offs should be applied on loan amounts based on the annual income of borrowers.

- As the loan amount increases, interest rate increases, but the most number of defaulters have extremely low loan amount.
    - This could indicate defaulters with very low credit scores.
    - Credit score should be considered an important factor.

- For good grade i.e A, there are more non verified defaulters than verified ones, but for remaining ones the number of verified defaulters is more than that of non verified.
    - Verification should not be considered an important factor for bad LC grade borrowers.

- Loan amount and annual income also have a positive correlation but its less than 0.5 which is a concern
    - Stricter cut offs for loan amount needs to be implemented

# THANK YOU