Loan Risk Assessment and Default Prediction: An Exploratory Data Analysis (EDA)

Presented by:
Bontha Tejaswini and Imliyangla Longkumer

INTRODUCTION

- The project aims to explore the factors that influence loan risk and default in a consumer finance company. It involves conducting an Exploratory Data Analysis (EDA) on past loan applicants' data to understand the patterns and attributes associated with loan defaults. Two types of risks are considered in this analysis:

  Risk of Not Approving a Repayable Loan: If the applicant is likely to repay the loan, not approving it results in a business loss.

  Risk of Default: If the applicant is likely to default, approving the loan may lead to a financial loss for the company.
- Objectives:

  - Estimating the driving factors which help in predicting whether the loan should be given or rejected to the customer.

  - Analyzing consumer attributes and loan attributes.
- Apart from applying the techniques ,we are also developing basic understanding of risk analytics in banking and financial services and understanding how data is used to minimize the risk of losing money while lending money to customers.

DATA LOADING AND INSPECTION

- We used libraries such as pandas,Numpy,seaborn,matplotlib.
- The dataset initially contained 111 columns and 39717 rows.
- We removed all the null columns along with columns that are irrelevant to this analysis(referring to the dictionary was crucial here).
- We also made modifications to specific column by converting them from object data type to float or integer for future analysis.

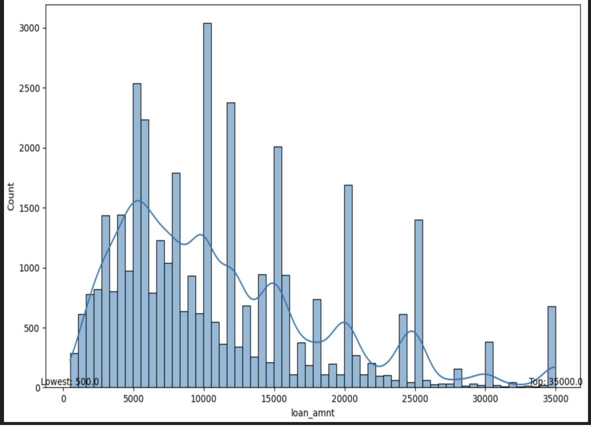
DATA UNDERSTANDING

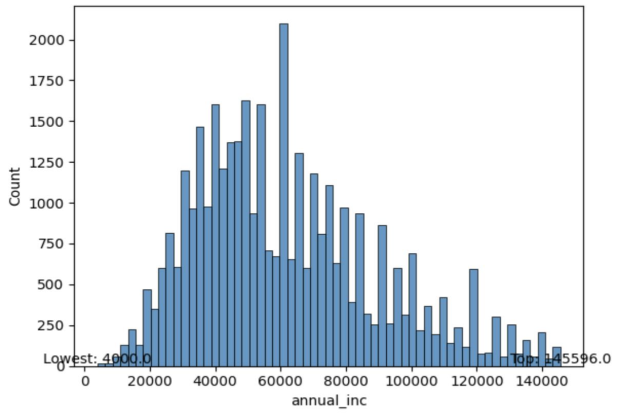
- As stated in the problem statement, the dataset contained information about the past loan applicants and who have either defaulted or not. Our objective is to identify the variables/factors that will likely lead to a loan applicant defaulting. After looking at the dataset, loan_status is the variable which informs us who have defaulted, who have not and the rest(current) which is irrelevant in this case.
- Based on this understanding we analysed how the rest of the relevant variables in the dataset behave, relate and/or fair against our target variable, loan_status.
- Some relevant variables were assumed to be the loam_amnt, annual_inc, emp_length, dti, etc.

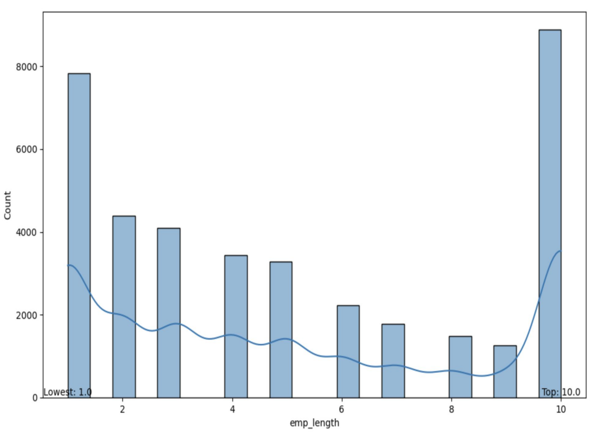
DATA MODIFICATION

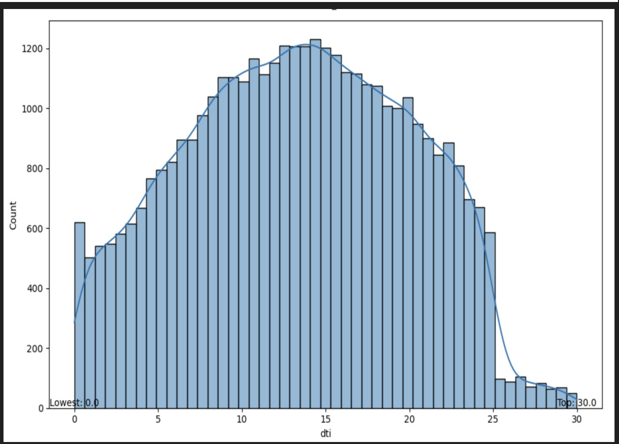
- Removing duplicate rows.
- Removing columns with 100-90 % missing values in columns
- Removing unnecessary variables
- Removing columns with single values.
- Converting obvious numerical column from object data type to integers or float.
- Removing outliers

DATA ANALYSIS - UNIVARIATE ANALYSIS









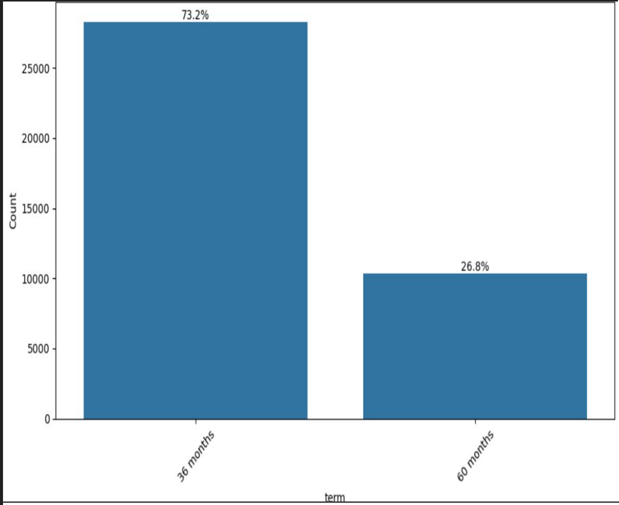
Observations:

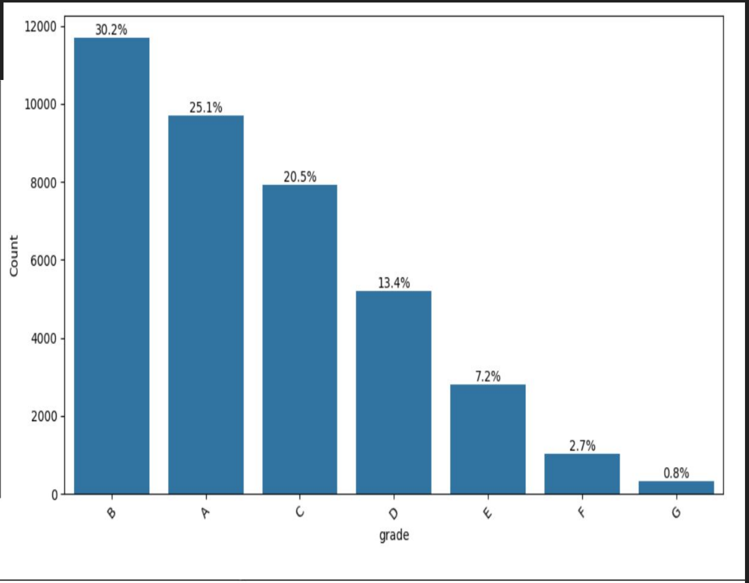
The histogram plots show the distribution of the numerical variables:

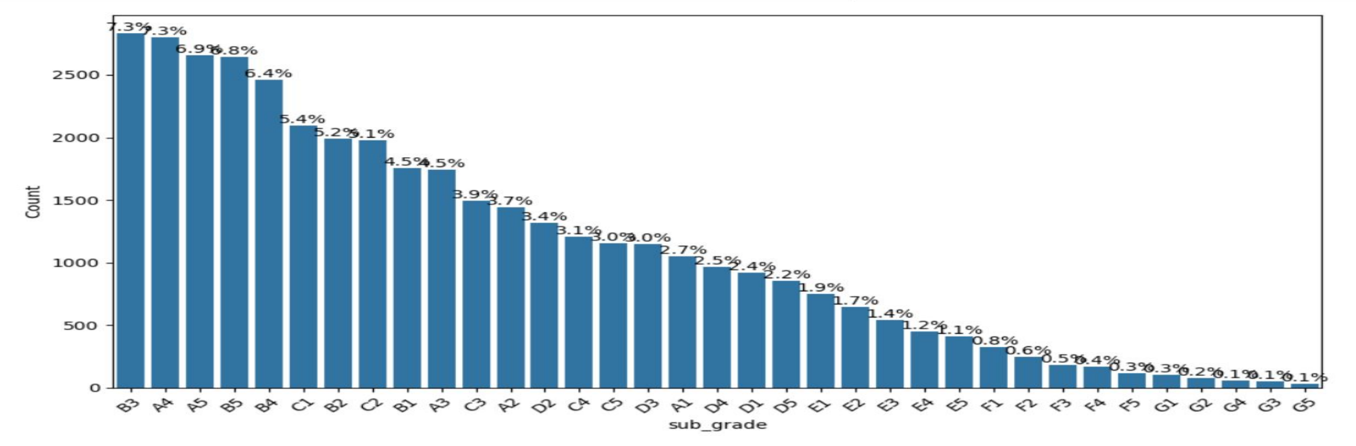
- Loan amount- largest $35000 an the smallest $500. The highest count being $10000
- int_rate- largest 24.6% an the smallest 5.6%. he highest count being 7-7.5 %(not shown here. Refer notebook)
- installment- largest 1300 an the smallest 15.7. he highest count being 100-200
- emp_length- longest 10 years an the shortest 1 year. he highest count being 10 years
- - annual_inc- largest $145596 an the smallest 4000 dollars.
- dti- lowest 0 and top 30. 15 being the highest count.

DATA ANALYSIS-CATEGORICAL VARIABLES

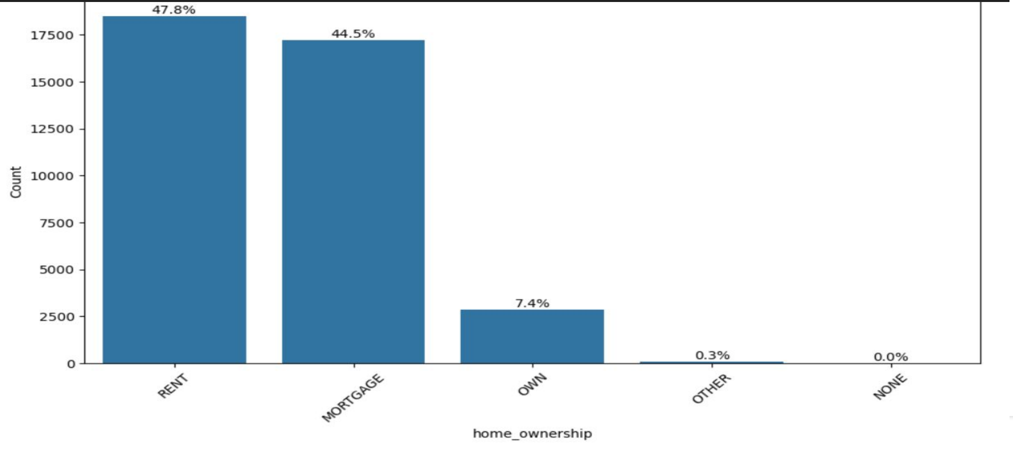
- Term- 73.2% of the loans were 36 months and 26.6% were 60 months.
- Grade- Grade b has the highest count whereas grade G has the lowest
- Sub-grade- sub-grade B3 has the highest count while grade G5 the lowest.

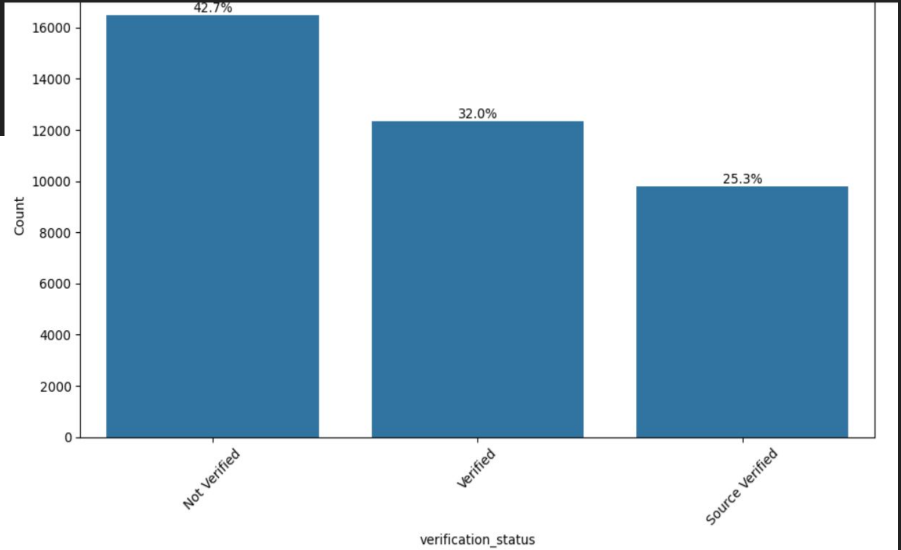


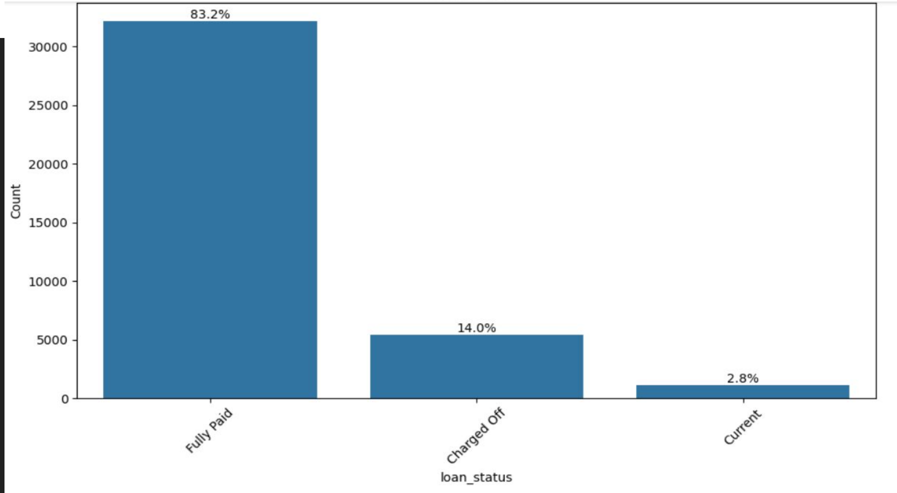




DATA ANALYSIS-CATEGORICAL VARIABLES







Home ownership- Applicants who rent has the highest count, followed by mortgage.
Verification status- Not verified has the highest count.
Loan_status- 83.2% are non-defaulters and 14% are defaulters.

DEFAULT RATE ANALYSIS

- After calculating the default rate of the loan_status with confirmation from the bivariate analysis, the default rate is estimated to be 14%
- A default rate of 14% means that approximately 14% of the loans in the dataset have defaulted. In this context, "default" typically refers to loans that have not been repaid or are in a state of delinquency.
- To classify loans as "good" or "bad" based on this default rate, it's common to use the rate as a threshold. In this case, loans with a default rate exceeding 14% (i.e., higher than 14%) are often categorized as "bad" loans, while loans with a default rate below 14% (i.e., lower than 14%) are considered "good" loans.

So, in this scenario:

- Loans with a default rate exceeding 14% would be classified as "bad" loans.
- Loans with a default rate below 14% would be classified as "good" loans.

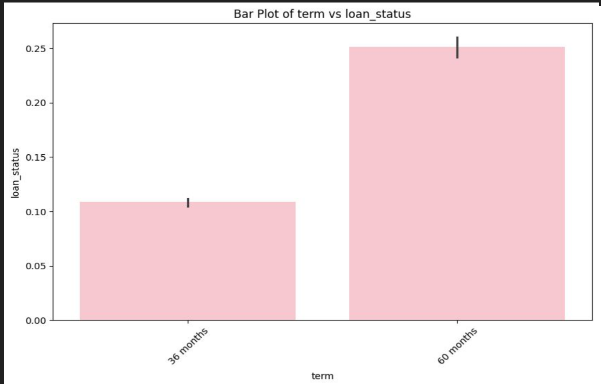
We determined the default rate across both categorical variables and numerical variable (by bucketing them) in the following slides

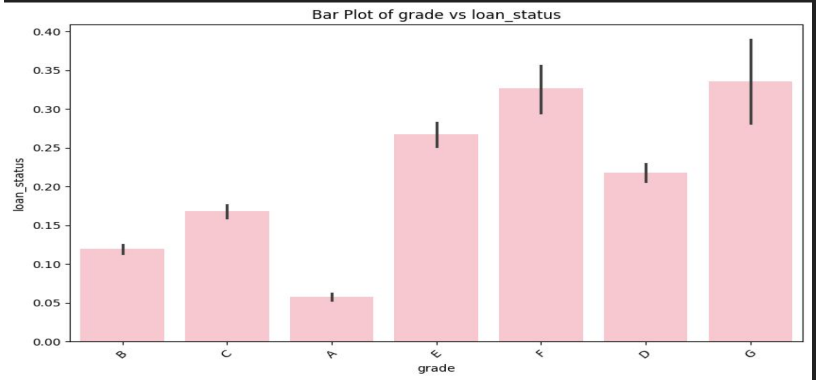
CATEGORICAL VARIABLE ANALYSIS

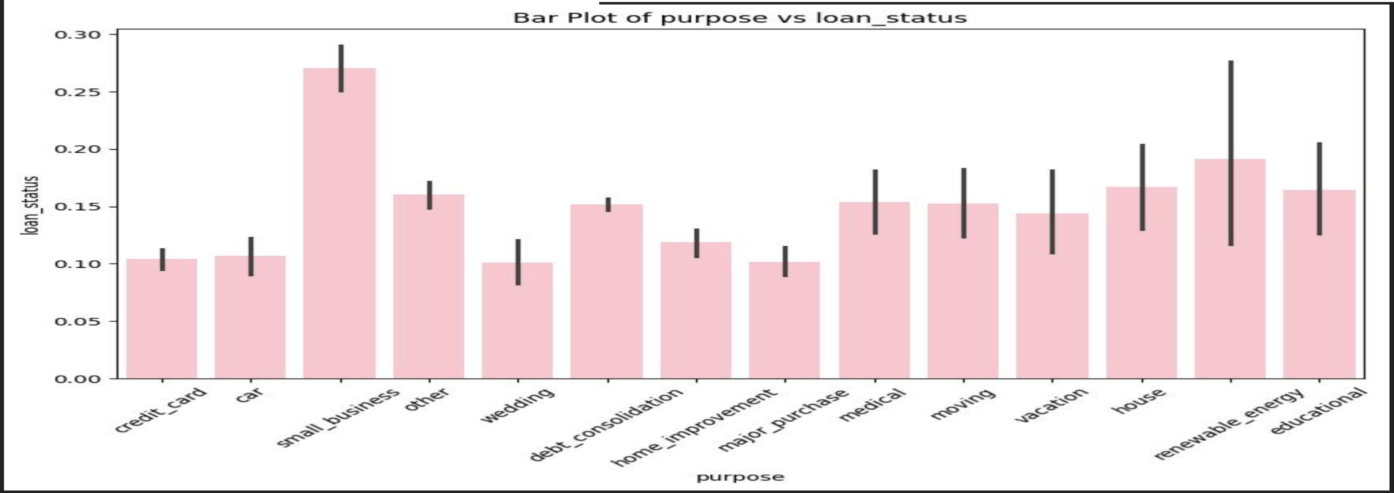
Term - Default rates are higher in the 60 months term than in the 30 months term

Grade- Grades G, F, E and D show higher default rates., G grade being the highest

Purpose- Small business have the highest default rate followed by renewable energy, educational, house and debt consolidation.


Bar Plot of term vs loan_status


Bar Plot of grade vs loan_status

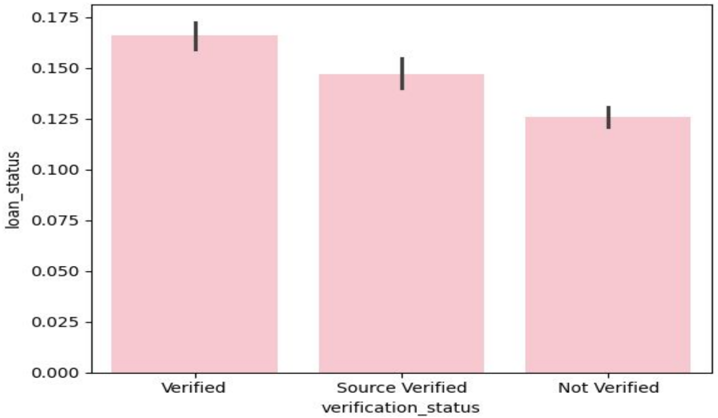

Bar Plot of purpose vs loan_status

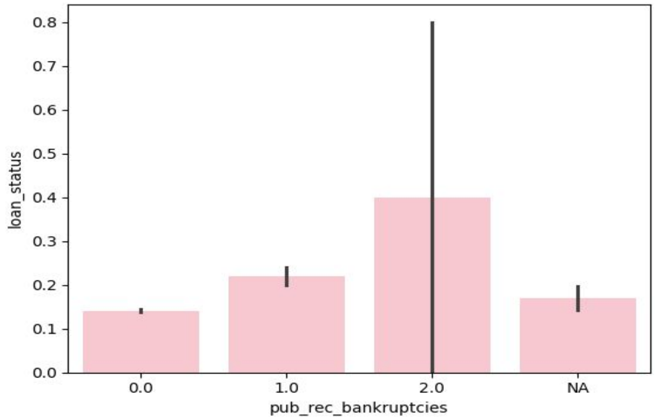
CATEGORICAL VARIABLE ANALYSIS

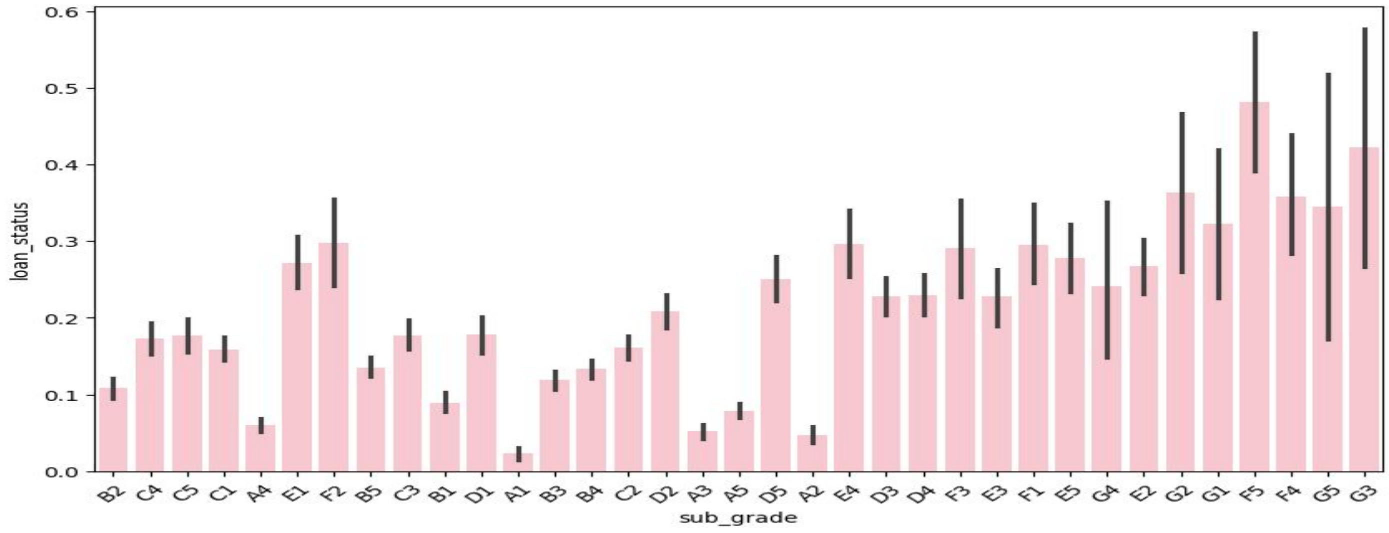
Verification Status- Income verified via the LC has higher default rate

Public record of bankruptcies(pub_rec_banckruptcies)- Cases where at least 2 bankruptcies have been recorded show higher default rate

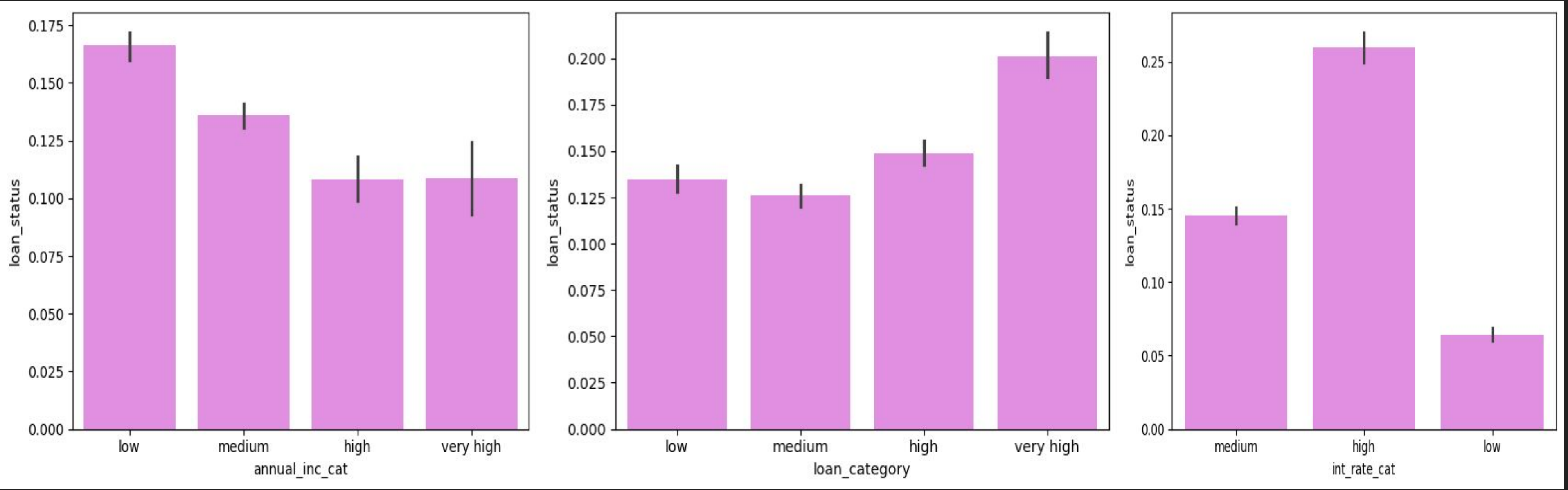
Subgrade- Like the grade variable, lower grade such as F5, G3, E2 show higher default rates.





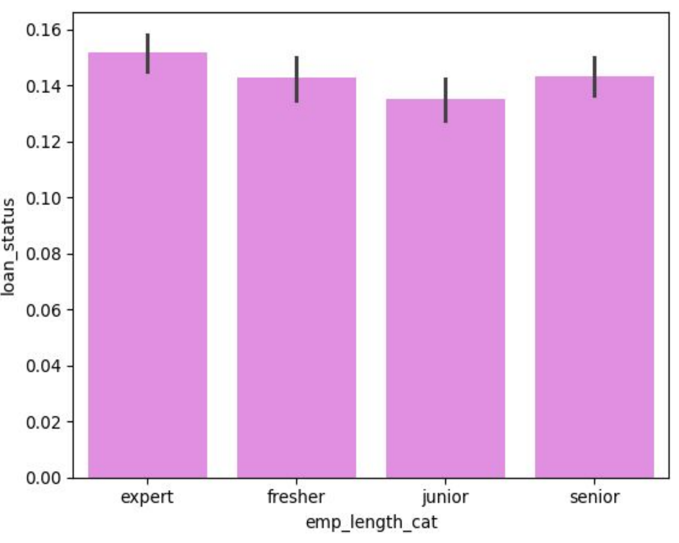


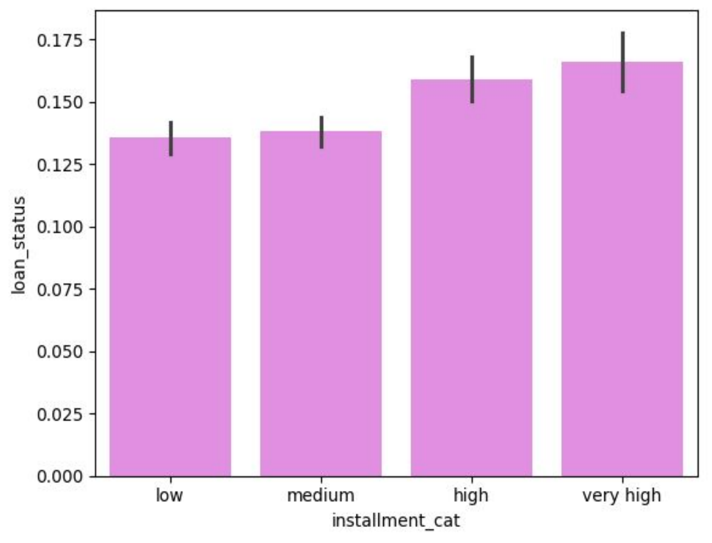
Categorical variables derived from numerical variables Analysis:

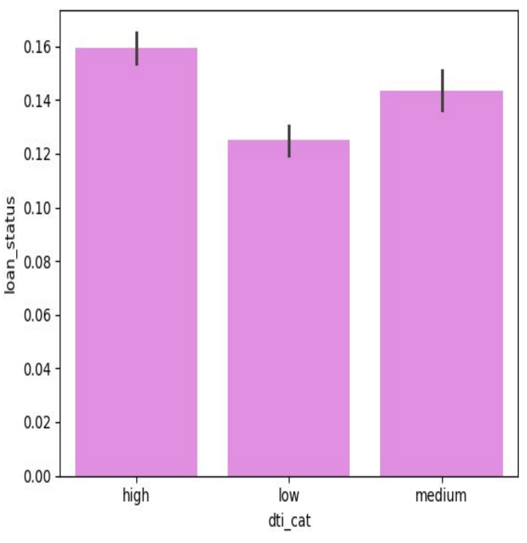


- Annual income- Lower income customers default more than high to very high income customers.
- Loan amount (loan_category)- Customers who take very high amount of loans default the most.
- Interest rate(int_rate_cat)- Highest default rate are seen for loan with high interest rate, followed by medium. Low interest rates have very low default rate.

Categorical variables derived from numerical variables Analysis:

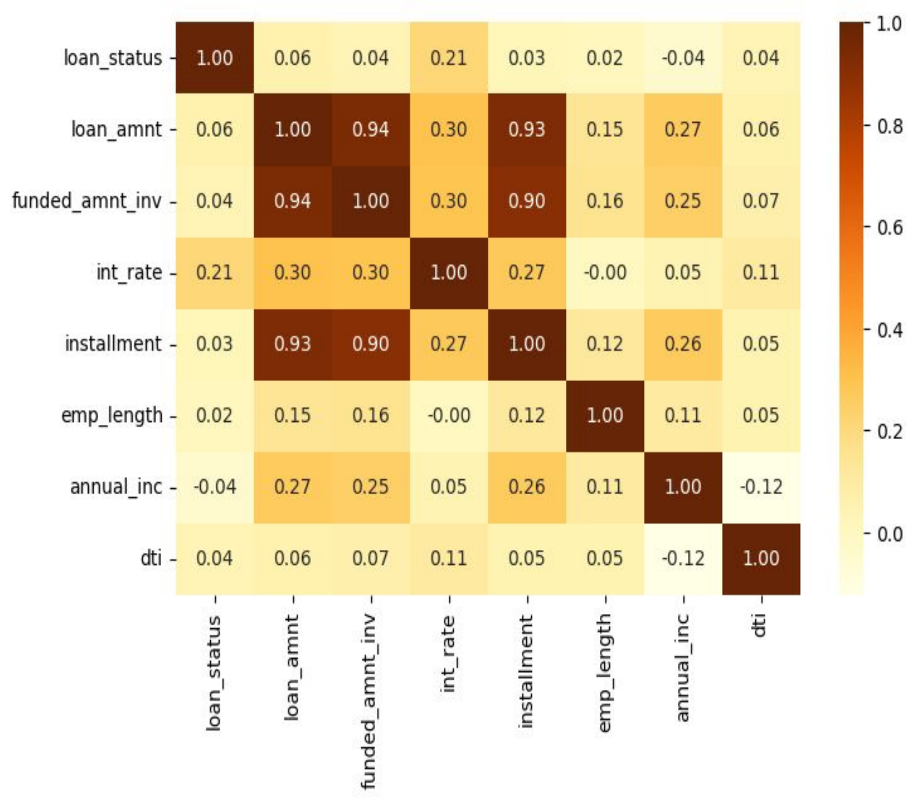






- Employment length(emp_length_cat)- Customers with more than 7 years of employment length slighted higher rate of defaulting than those below.
- Installment- Very high (>600) and high installments have higher default rate than medium to low.
- Debt to income ratio(dti_cat)- High dti has the highest default rate, followed by medium and low dti.

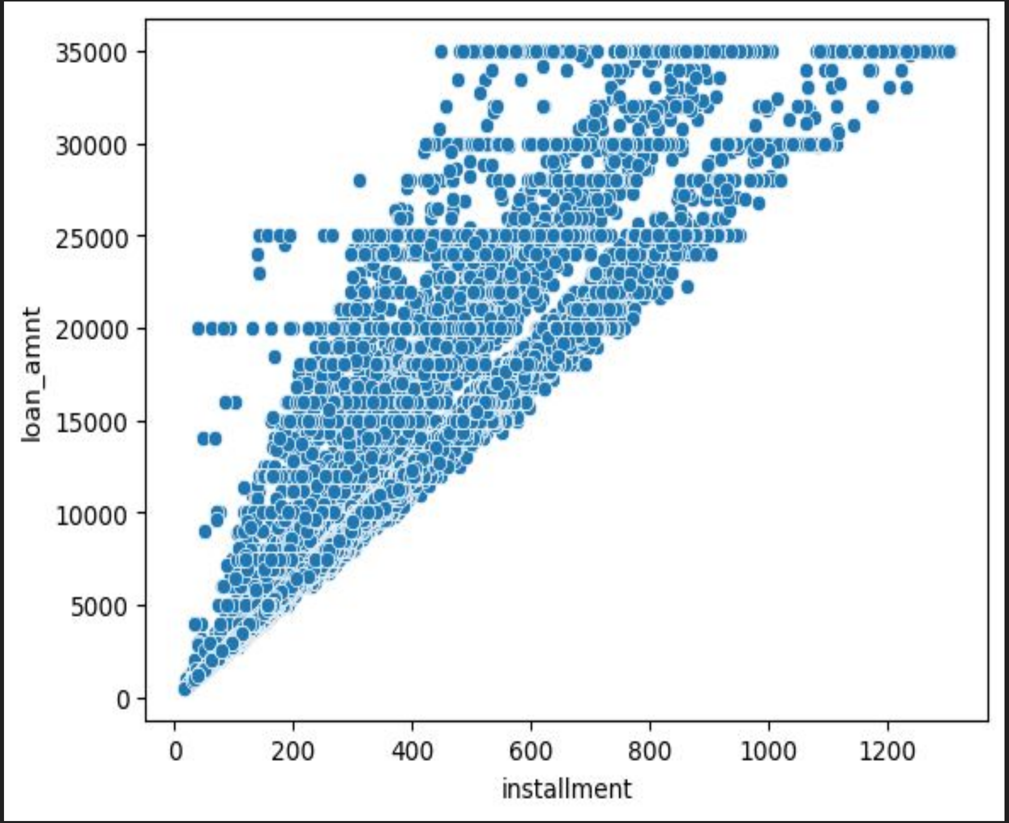
NUMERICAL VARIABLE ANALYSIS-CORRELATION



	loan_status	loan_amnt	funded_amnt_inv	int_rate	installment	emp_length	annual_inc	dti
loan_status	1.00	0.06	0.04	0.21	0.03	0.02	-0.04	0.04
loan_amnt	0.06	1.00	0.94	0.30	0.93	0.15	0.27	0.06
funded_amnt_inv	0.04	0.94	1.00	0.30	0.90	0.16	0.25	0.07
int_rate	0.21	0.30	0.30	1.00	0.27	-0.00	0.05	0.11
installment	0.03	0.93	0.90	0.27	1.00	0.12	0.26	0.05
emp_length	0.02	0.15	0.16	-0.00	0.12	1.00	0.11	0.05
annual_inc	-0.04	0.27	0.25	0.05	0.26	0.11	1.00	-0.12
dti	0.04	0.06	0.07	0.11	0.05	0.05	-0.12	1.00

- **Loan_amnt** and **funded_amnt_inv** have a correlation coefficient of 0.94 which is pretty high and expected as almost all the values in both the columns are the same.
- A correlation heatmap is a graphical tool which displays the correlation between multiple variables as a color matrix.
- Dark colour indicates that they are highly correlated and less shades or a lighter shade depicts less correlation.
- If the value is 1,it is said to be a positive correlation between the two variables,this indicates that one variable increases the other also increases.
- If the value is -1,it is said to be a negative correlation between the two variables.
- If the value is 0 there is no correlation.

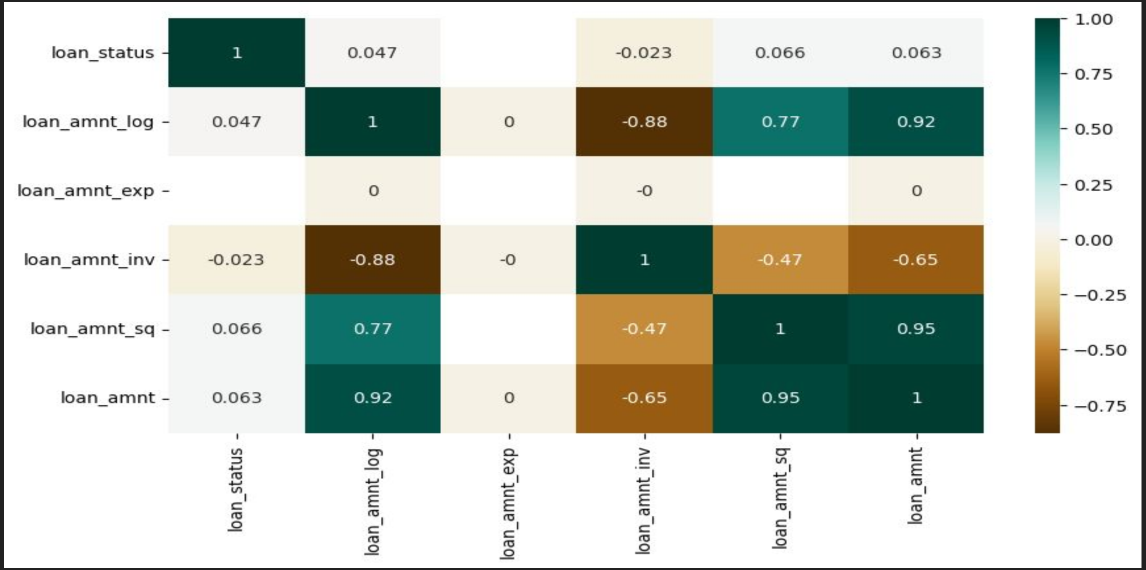
NUMERICAL VARIABLE ANALYSIS-SCATTER PLOT



- Based on the high correlation observed between loan_amont and installment in the heatmap, we wanted to investigate the linear relationship of these variables.
- The scatter plot on the left confirms the linear relationship between loan_amont and installment

DATA TRANSFORMATION

- Transforming the numerical variables to create pseudo variables such as log of the variable,square,inverse,cube and tanh.
- The need for exploring non-linear relationships arises from the desire to capture complex and non-linear patterns in the data that traditional linear models might overlook.



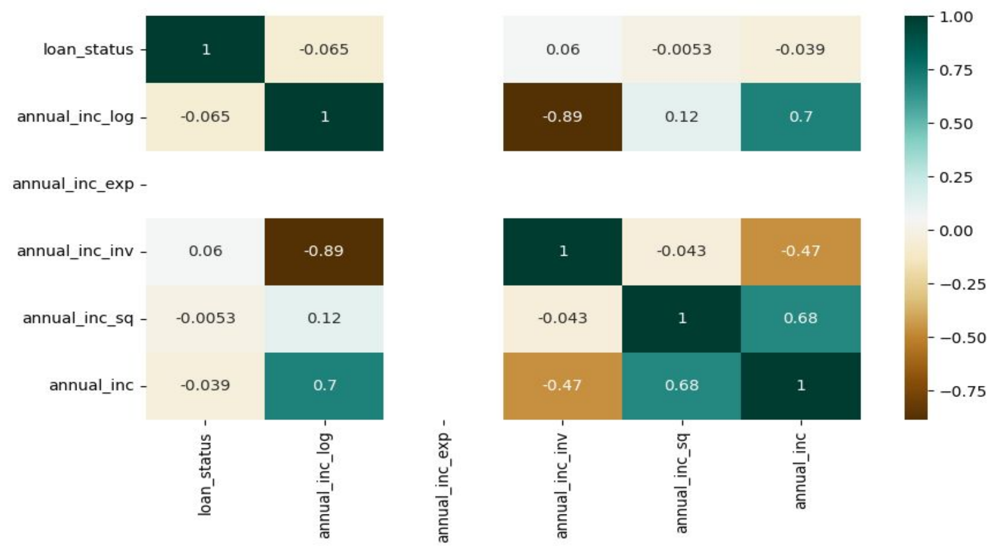
	loan_status	loan_amnt_log	loan_amnt_exp	loan_amnt_inv	loan_amnt_sq	loan_amnt
loan_status	1	0.047		-0.023	0.066	0.063
loan_amnt_log	0.047	1	0	-0.88	0.77	0.92
loan_amnt_exp		0		-0		0
loan_amnt_inv	-0.023	-0.88	-0	1	-0.47	-0.65
loan_amnt_sq	0.066	0.77		-0.47	1	0.95
loan_amnt	0.063	0.92	0	-0.65	0.95	1

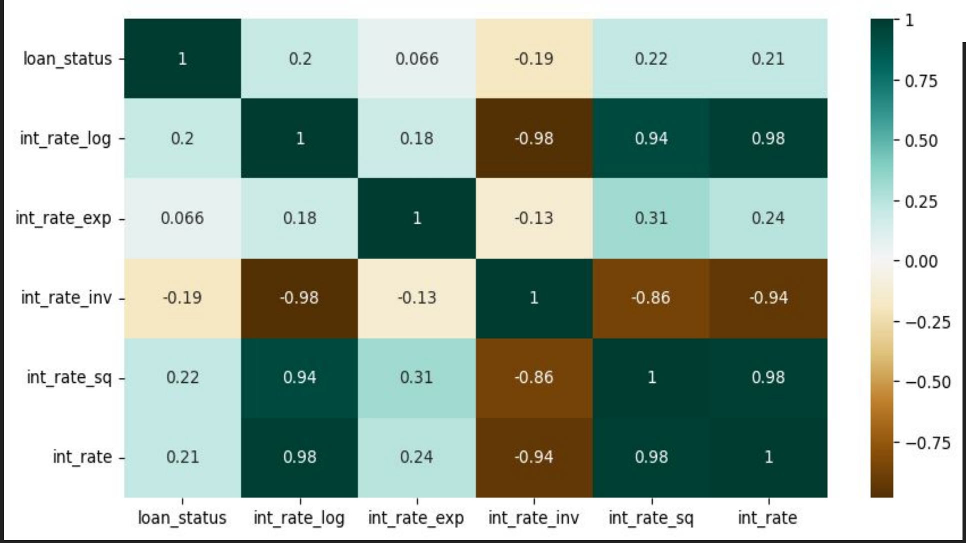
- The variable loan_amnt is transformed in the heat map above.
- Doesn’t show significant correlation with loan_status
- The inverse of loan_amnt shows a somewhat weak negative correlation with loan_status

DATA TRANSFORMATION

The heatmap on the right top:

- Annual_inc is the transformed variable here
- No significant correlation found with loan_status.

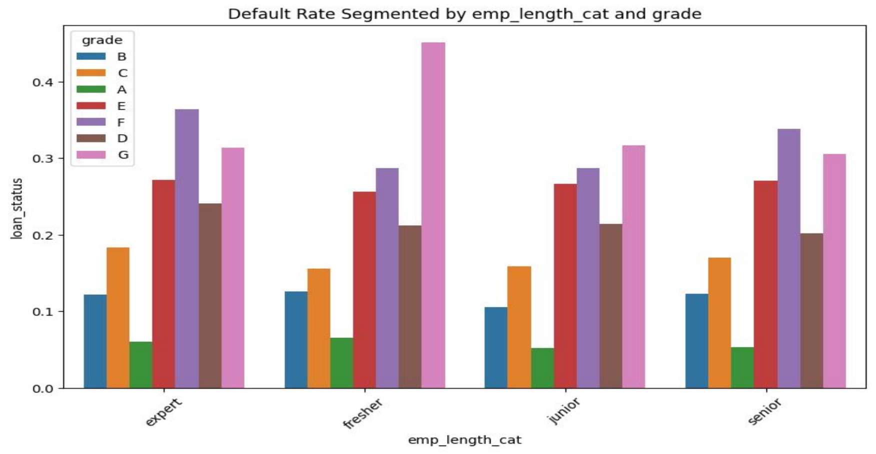


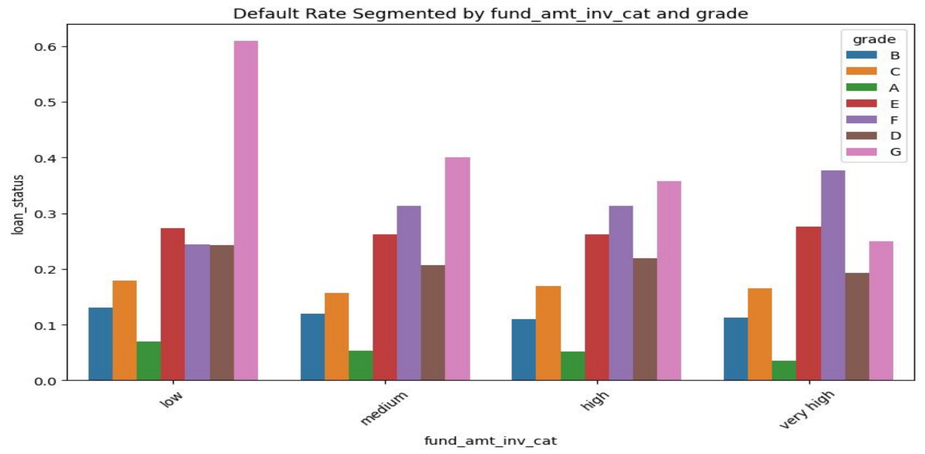


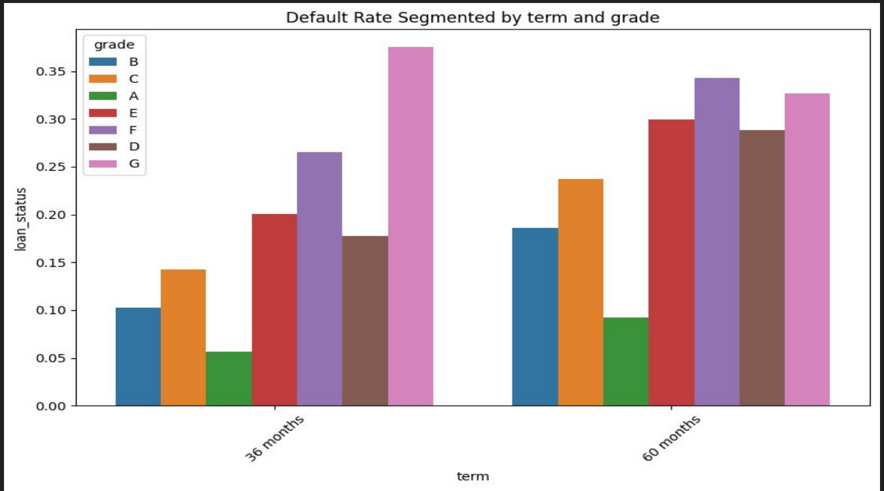
The heatmap on the left bottom:

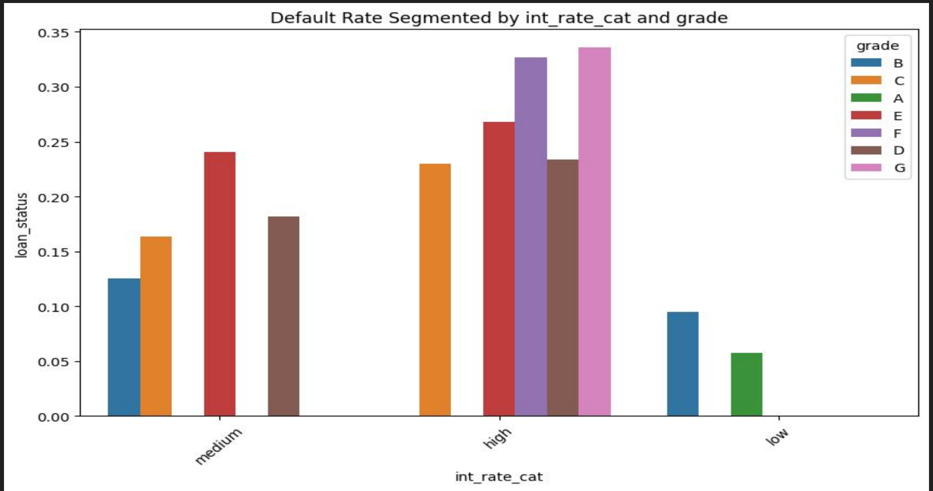
- int_rate(interest rate) is the transformed variable here.
- Int_rate_sq, int_rate and int_rate_log shows a weak to medium correlation with loan_status.

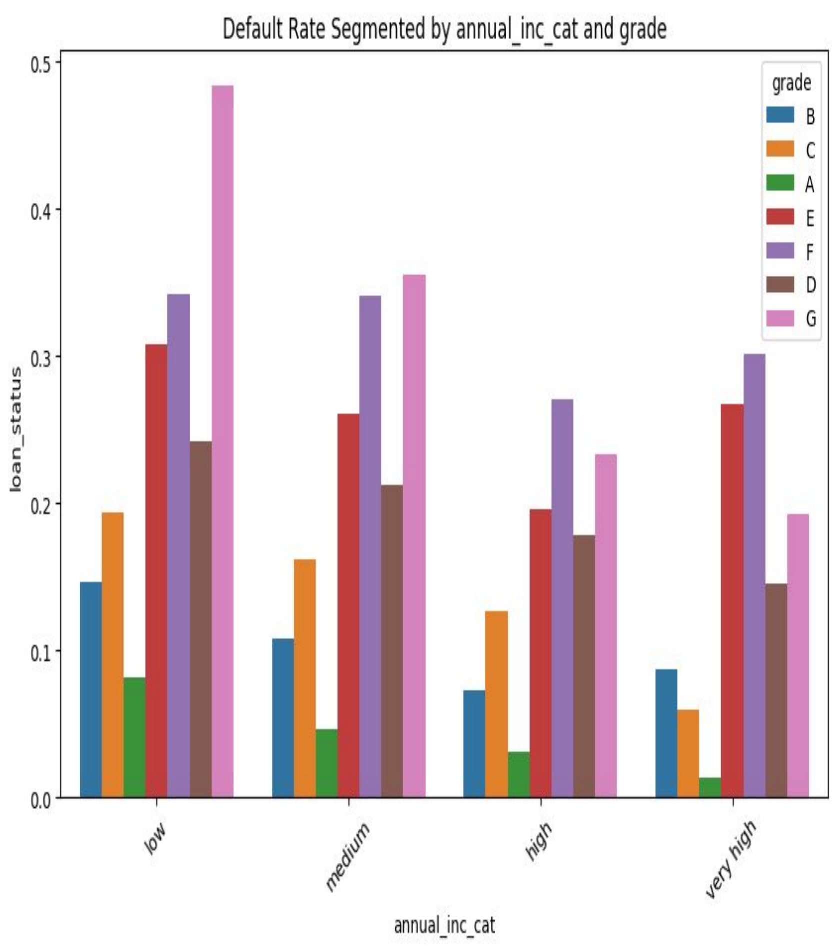
SEGMENTED BIVARIATE ANALYSIS


Default Rate Segmented by emp_length_cat and grade


Default Rate Segmented by fund_amt_inv_cat and grade


Default Rate Segmented by term and grade


Default Rate Segmented by int_rate_cat and grade



We compared the default rates across various variables, and some of the important predictors are purpose of the loan, interest rate, annual income, grade etc.

We have segmented the loan application across grade of loan,since that is a variable affecting many other variables-the type of applicants,interest rate,income and finally the default rate.

There are 6 types of loans based on grades:A,B,C,D,E,F,G.

- In case of the graph between loan_status and emp_length_cat, freshers with grade G have highest default rate.

In case of term and loan_status graph there are two types of term period one is 36 months and 60 months of duration

- 36 months grade G has highest loan_status >0.35 and when compared to 60 month grade F has loan_status varying between 0.30 -0.35.

There is a 6% increase in default rate as you go from high to low annual income.

CONCLUSION

The analysis of past loan applicants' data revealed crucial insights into factors influencing loan risk and default. The default rate, estimated at 14%, has been a pivotal reference point for categorizing loans as "good" or "bad." This analysis has led to several key findings:

- Univariate analysis provided a deeper understanding of the dataset, highlighting the distribution of numerical and categorical variables.
- Categorical variable analysis showed varying default rates based on factors such as loan term, grade, purpose, and verification status and variables like grade and sub-grades affected default rates the most.
- Numerical variables like annual income, loan amount, and interest rate were examined for their correlation with loan status and higher loan amount and interest rate showed higher default rates.
- Data transformation unveiled the potential influence of non-linear relationships and highlighted some weak to medium correlations.
- Higher interest rates showed higher default rates across all board.

Recommendations:

Based on our analysis, the following recommendations can be made:

- Risk Assessment: Implement more sophisticated risk assessment models that takes into account factors like interest rates, term, loan amount and purpose.
- Monitoring Loans: Continuously monitor loans with higher default risk, such as those associated with small businesses and high-interest rates, to reduce financial losses.
- Customer Segmentation: Segment customers based on risk profiles to offer tailored loan products and interest rates, minimizing default risks.