LENDING CLUB CASE STUDY

Group Facilitator : Praveenkumar Periyaswamy
Team Member : Kiran Chakkilam

Objective

To analyze the loans dataset and find out the people who are likely to default

Problem Statement

We have been given a loans dataset with multiple columns. The applicant can either pay the loan or not pay the loan. We have to figure out if we can approve the loan to an applicant or not.

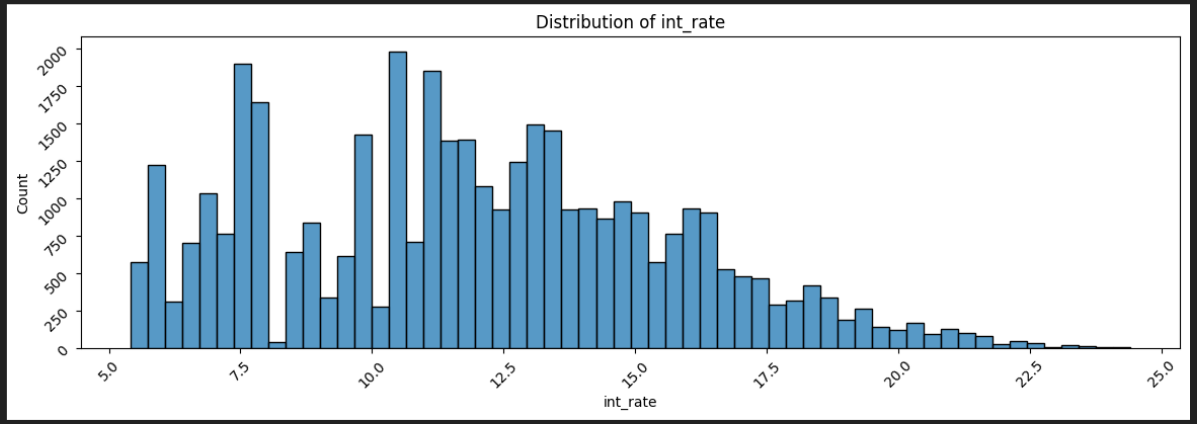
Business Objective

We have to minimize the amount of credit loss. If the applicant is likely to pay the loan, then not approving the loan is a loss of business to the company. Similarly, if the applicant can’t repay the loan, then approving the loan is again a loss to the company.

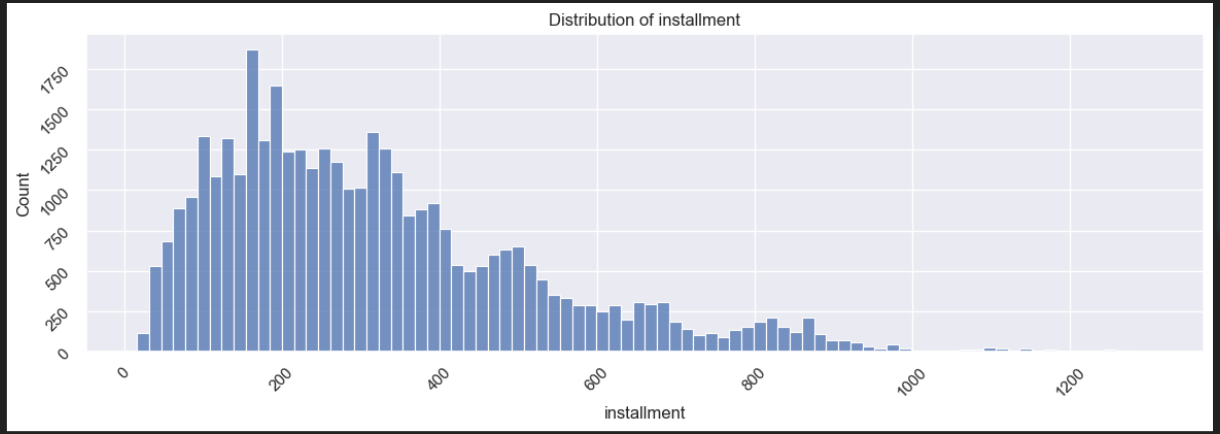
Data Cleaning

1. Drop the columns where the Null values are greater than equal to 50 %
2. Identify the Single Non null unique columns and removed them from dataset
3. Drop all irrelevant columns
4. Select only loan applicant with charge-off and Fully Paid loan status
5. Replace the Null values in emp_length column
6. Remove the data where pub_rec_bankruptcies column value is null
7. Standardization and data type conversion for int_rate ,term and issue_d columns
8. Derive the new columns from issue_d columns
9. Detect and treat outliers in annual_inc column

Univariate Analysis – Qunatitative variables



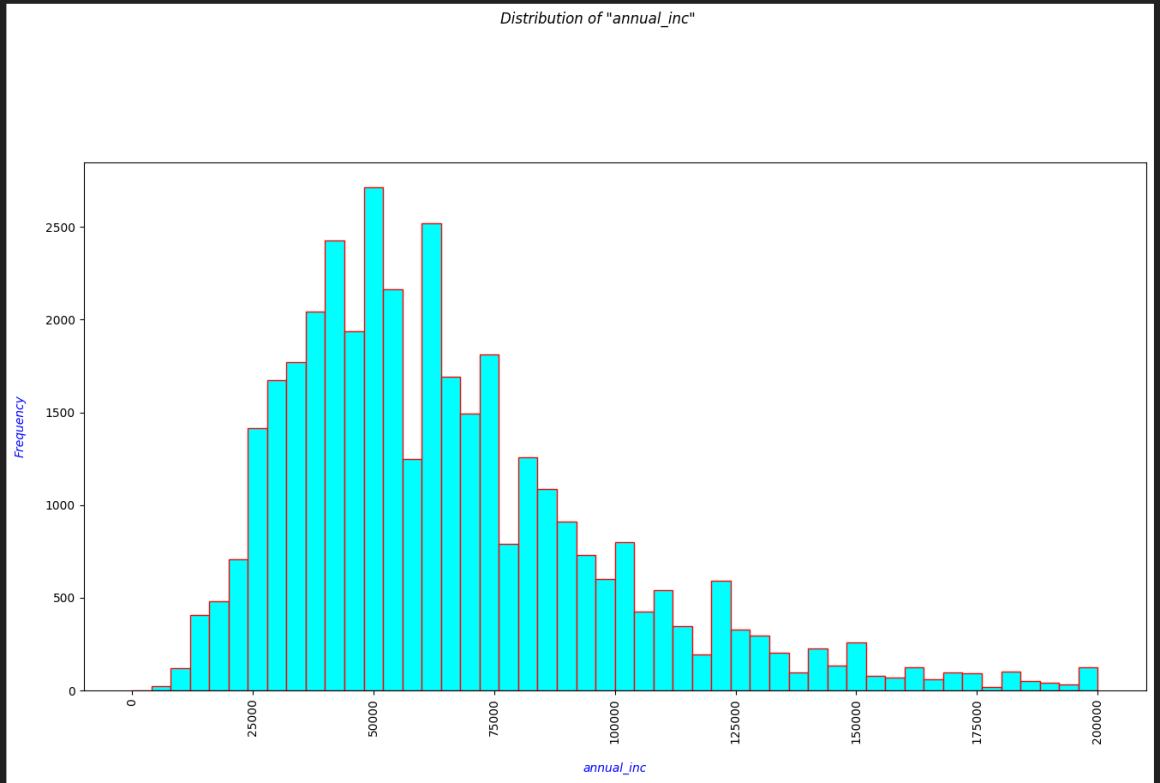
Usually, **int_rate** is offered in the range of 6-8% or 10-15%. May be the case that the interest rate is higher for the borrowers with higher debt-to-income ratios.



The Majority of borrowers are with in 400 months installment.

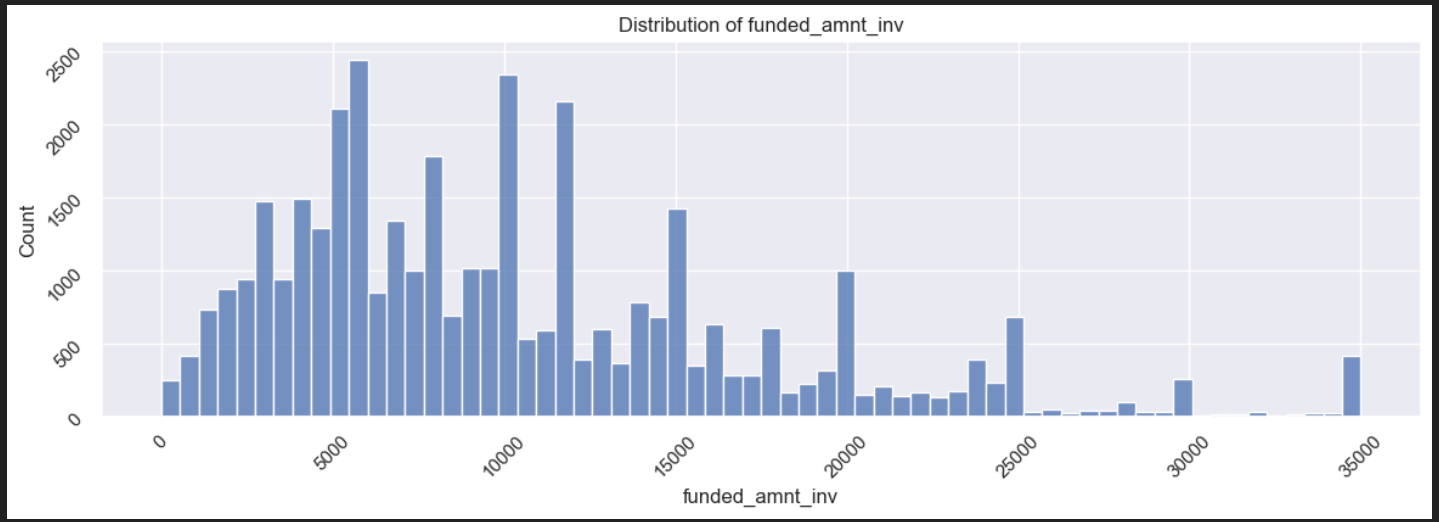
Univariate Analysis – Qunatitative variables

Most of the applicants are from 25000 to 75000 income range. This mean lender get most of the requests from less salaried individuals



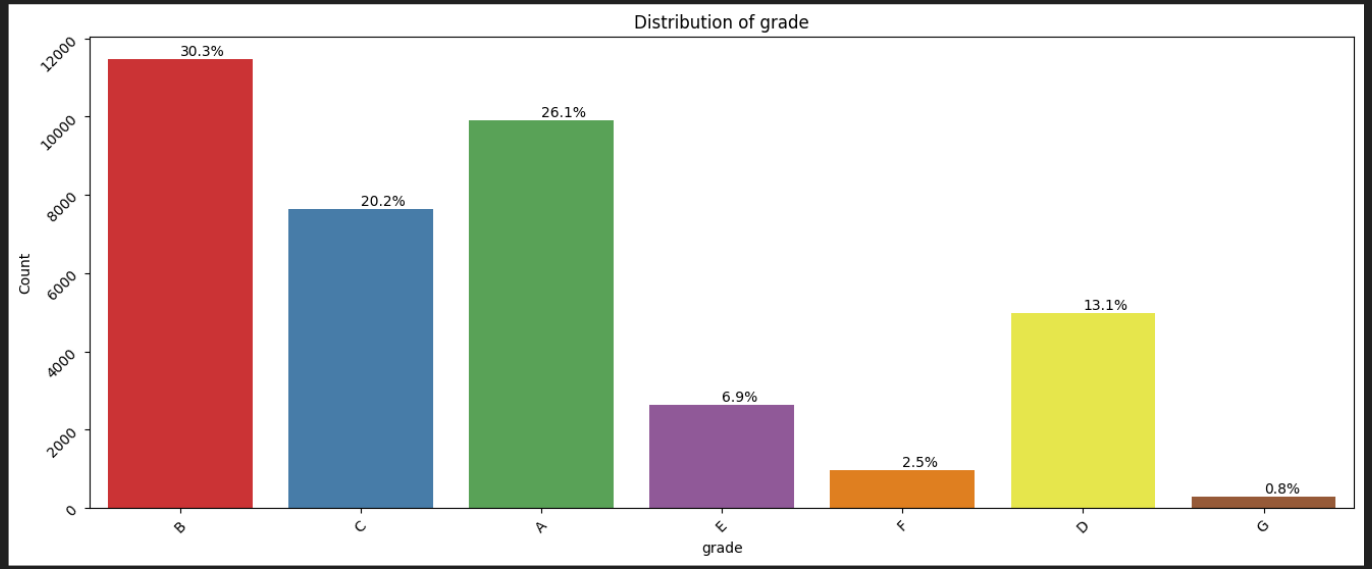
Univariate Analysis – Qunatitative variables

funded_**amnt_inv** - The majority of applicants are granted loan between 4,500 and 10,000.



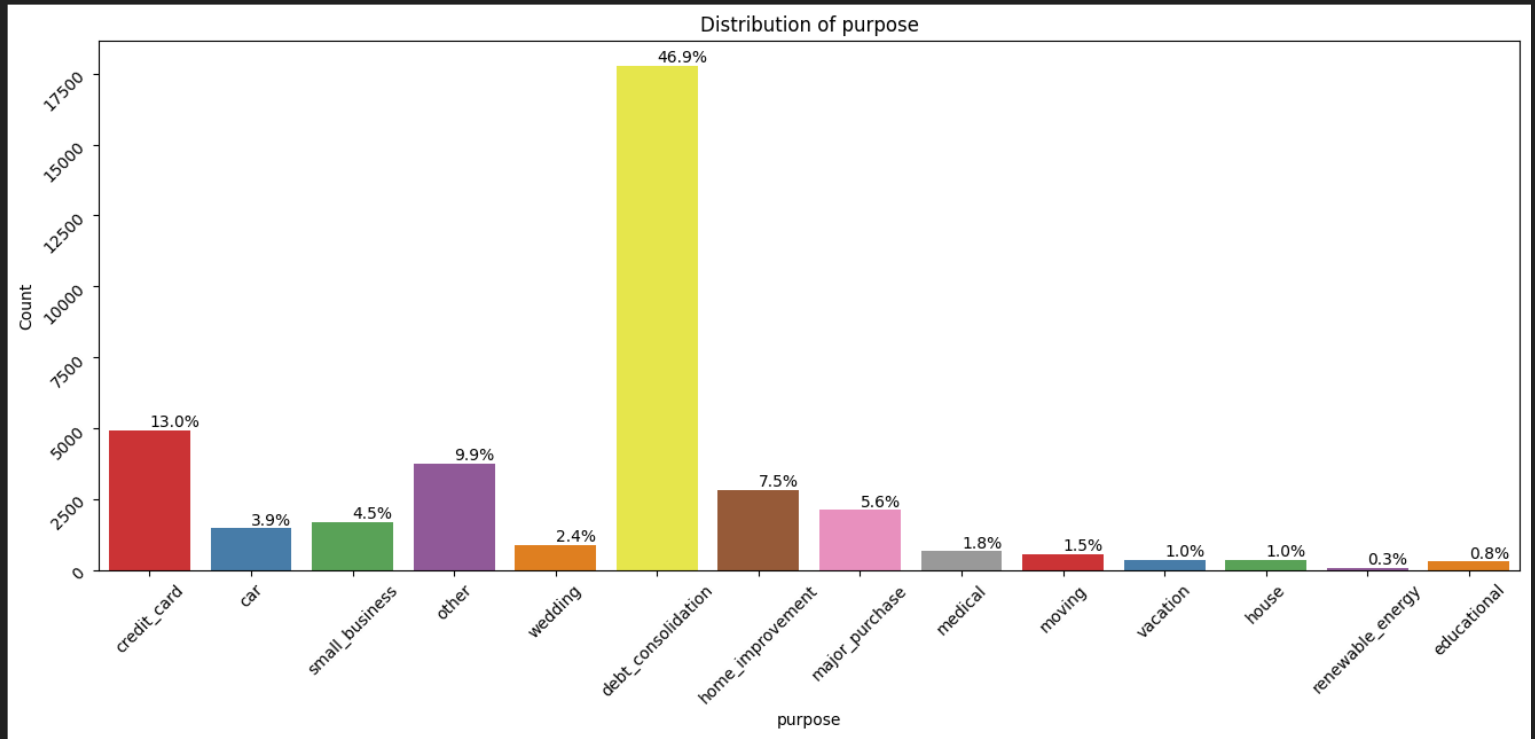
Univariate Analysis – Categorical variables

- Most of applicants are from **A** and **B** grade levels



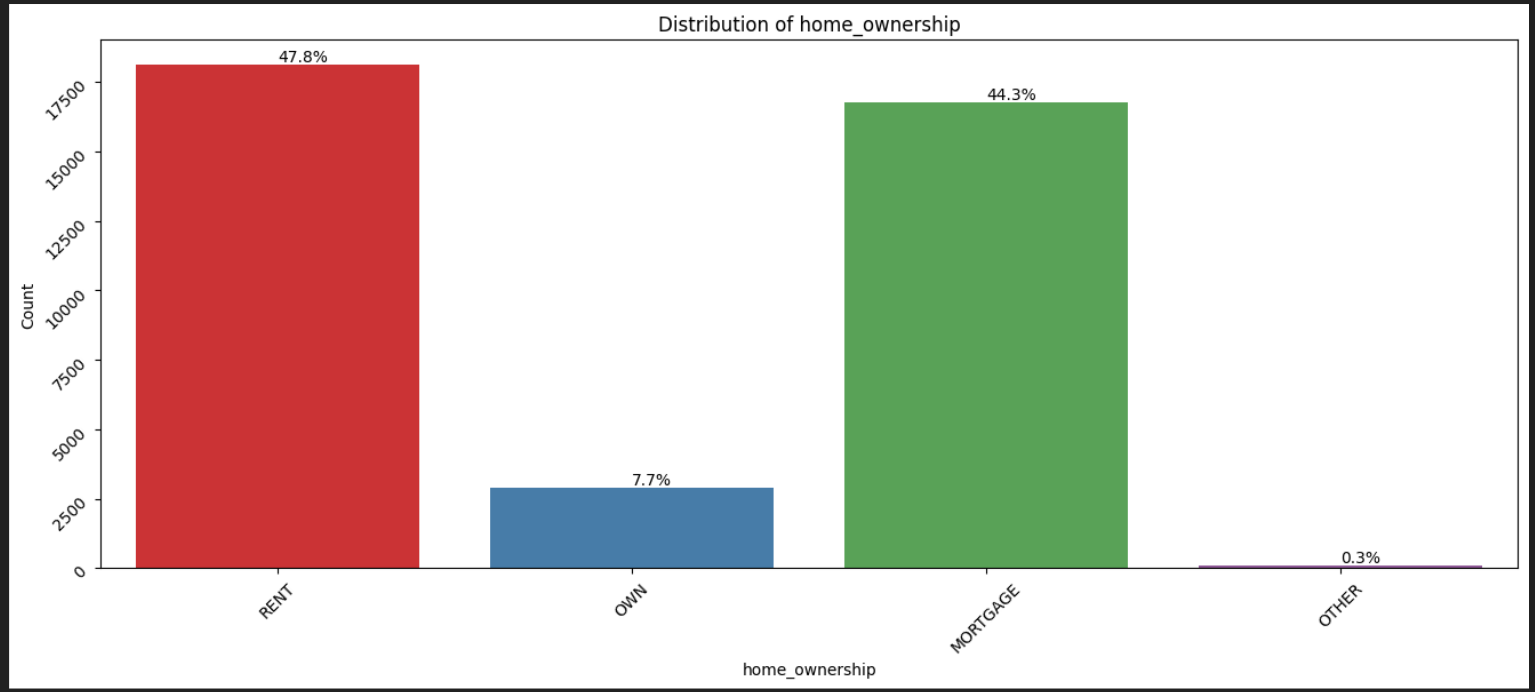
Univariate Analysis – Categorical variables

- Most of the people are taking loas for **Debt Consolidation purpose**



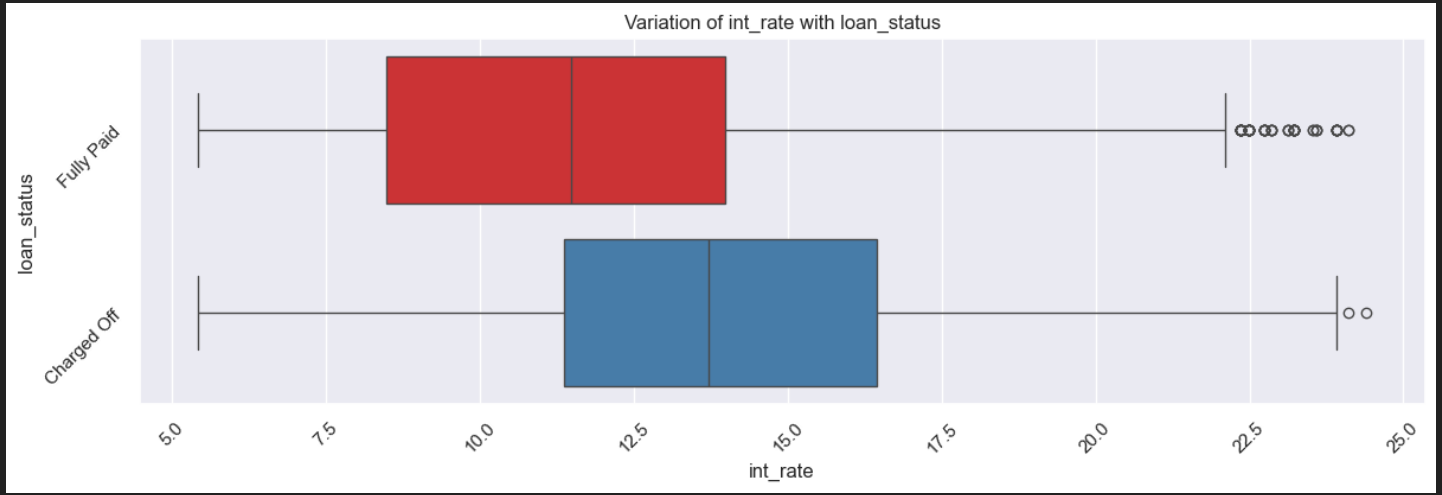
Univariate Analysis – Categorical variables

- Applicants on **RENT** and **Mortgage** are taking more loans



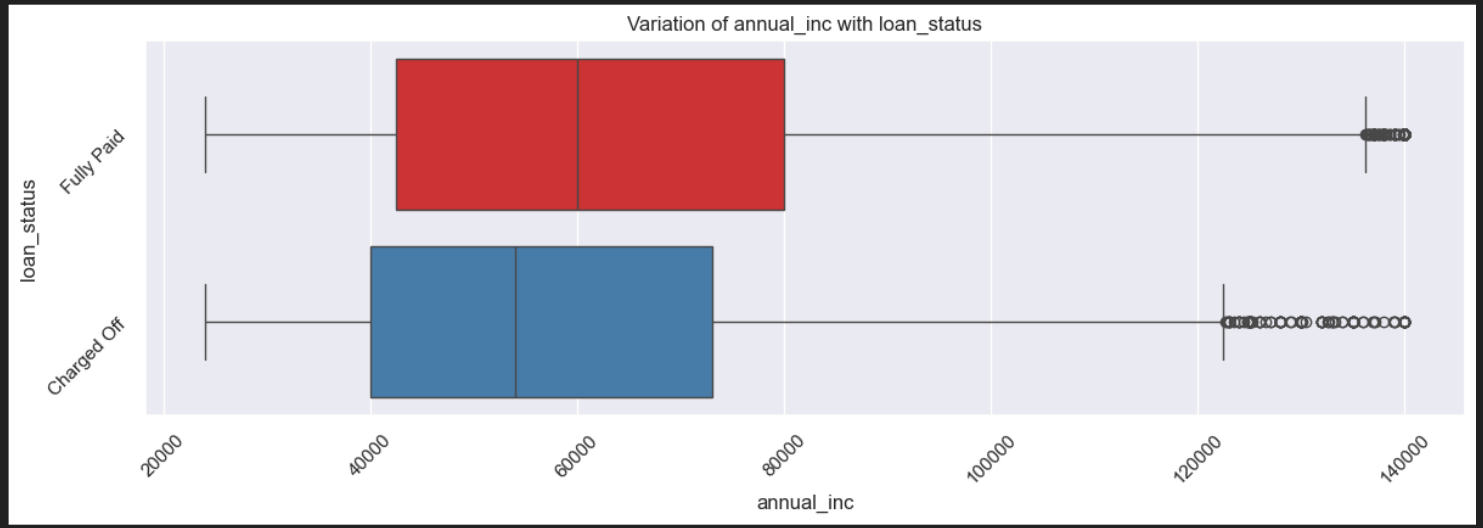
SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF NUMERICAL VARIABLES ON LOAN STATUSES

- The interest rates of fully paid loans and defaulted loans differ noticeably, with the defaulted loan interest rate being greater on average
- ***int_rate*** is likely to influence the loan status of the borrower. Thus, it is a driving factor behind loan default



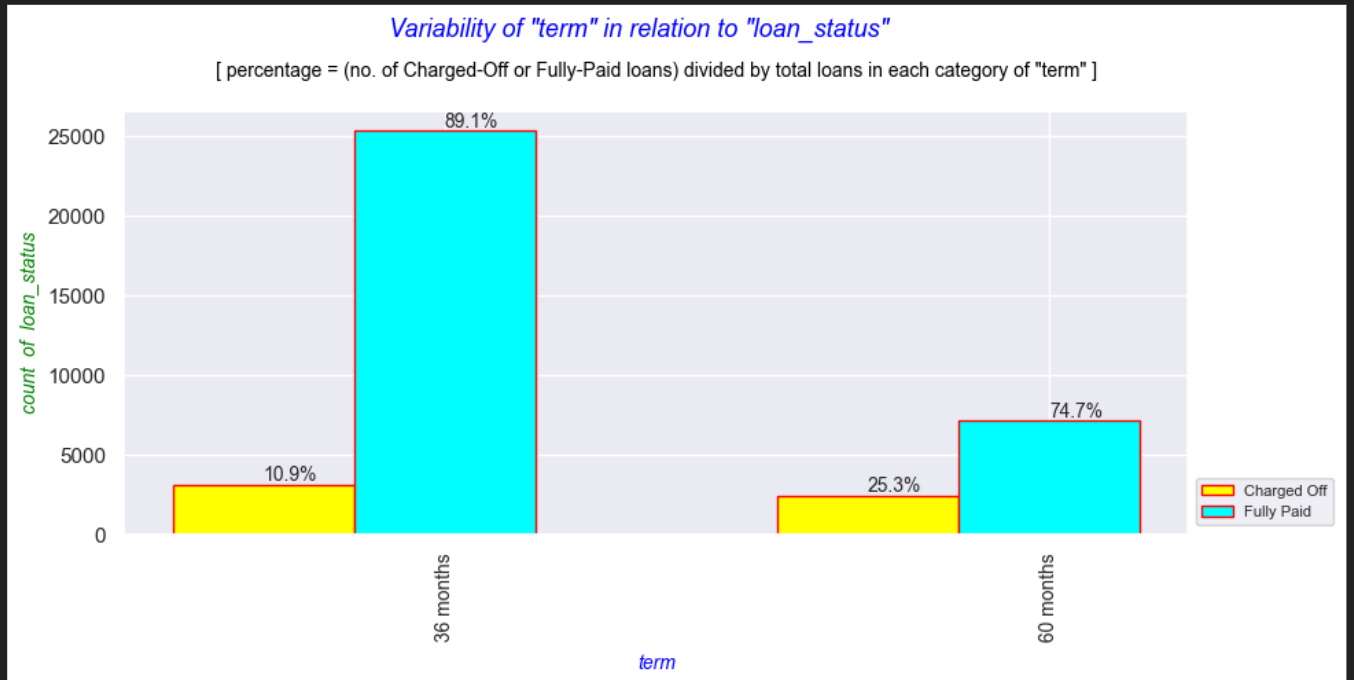
SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF NUMERICAL VARIABLES ON LOAN STATUSES

- The annual income of loan defaulters at 75th percentile is lower than those who have fully paid back their loans.
  But when we compare the 25th percentile incomes, they are different but not as significant as 75th percentile incomes.
  This shows that the borrowers with higher income are less likely to default



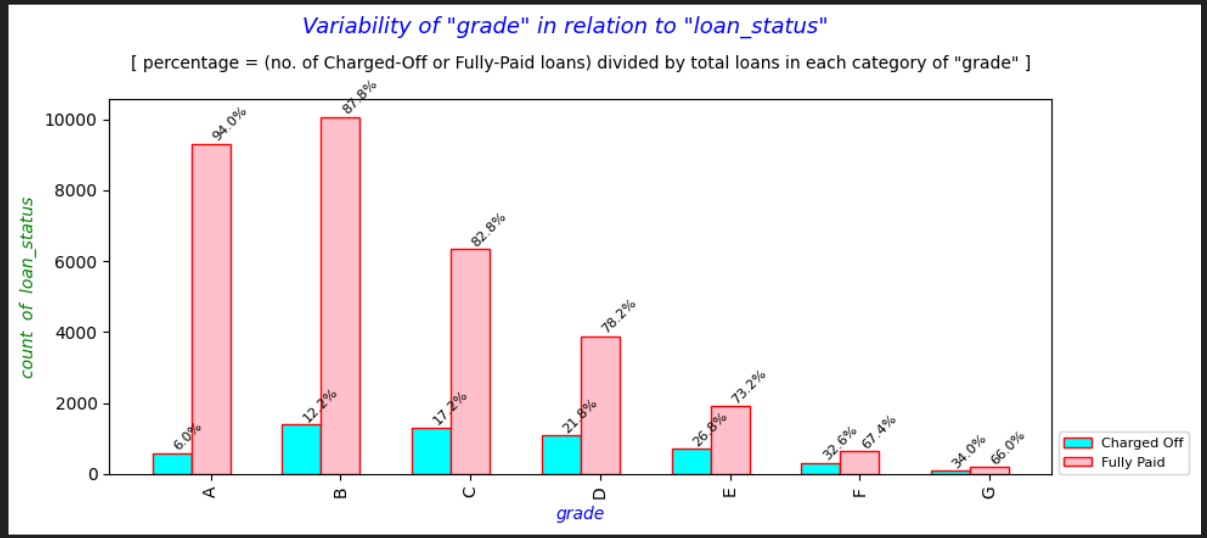
SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF CATEGORICAL VARIABLES ON LOAN STATUSES

- The percentage of defaulters in 60 months term more than that of 36 months. So clearly term is influenced variable for lending loans



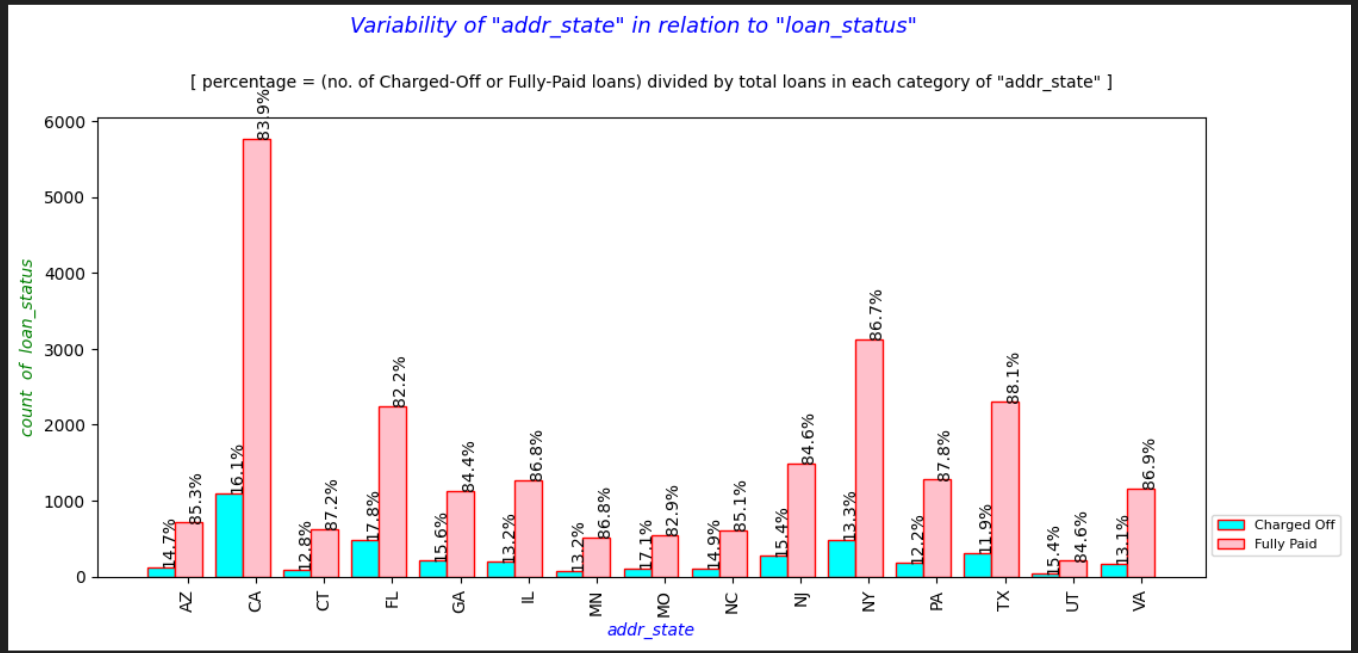
SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF CATEGORICAL VARIABLES ON LOAN STATUSES

- Observed that default rate percentage increases from A to G grade. It is evident that applicants with lower grade are more defaulting.

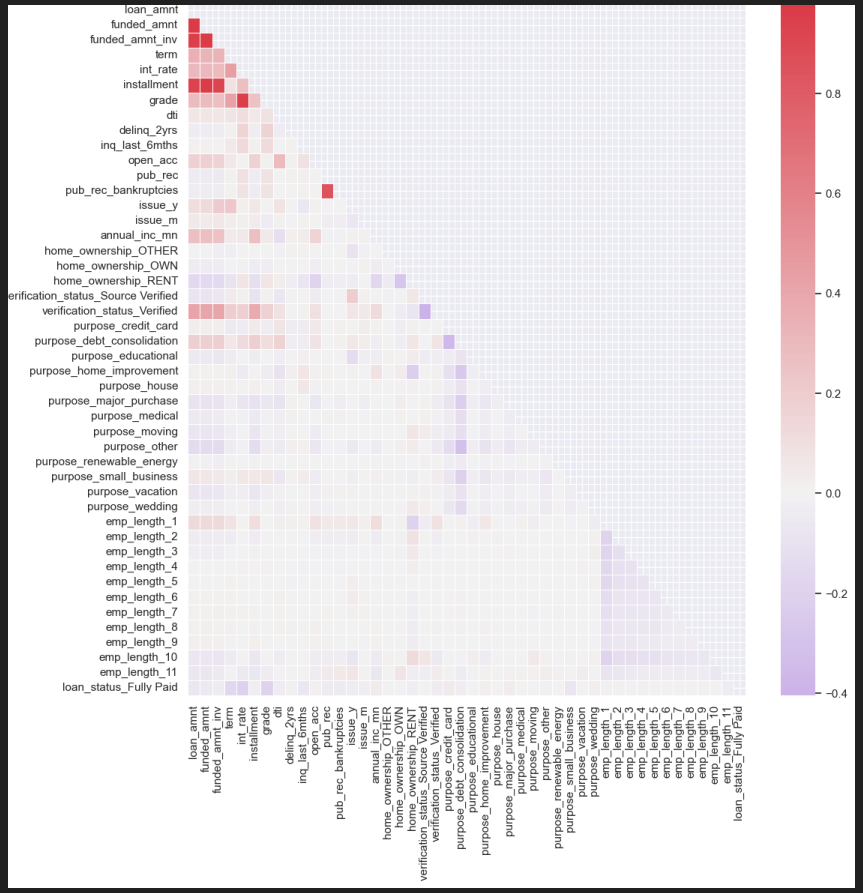


SEGMENTED UNIVARIATE ANALYSIS TO ASSESS THE INFLUENCE OF CATEGORICAL VARIABLES ON LOAN STATUSES

- Most of the borrowers are from the state California, followed by New York, Florida and Texas states.



Multivariate Analysis



Analysis from above matrix

- we can conclude that int_rate is highly correlated to grade of the borrowers. installment and funded_amnt_inv, loan_amnt, are highly correlated variables

Conclusion

- These are the key variable which influences default status - **int_rate, term, grade, subgrade, annual_inc , purpose** and **addr_state**
- These are the correlated variables in the dataset – **int_rate & grade** , **installment** & **funded_amnt_inv**