Lending Club Case Study
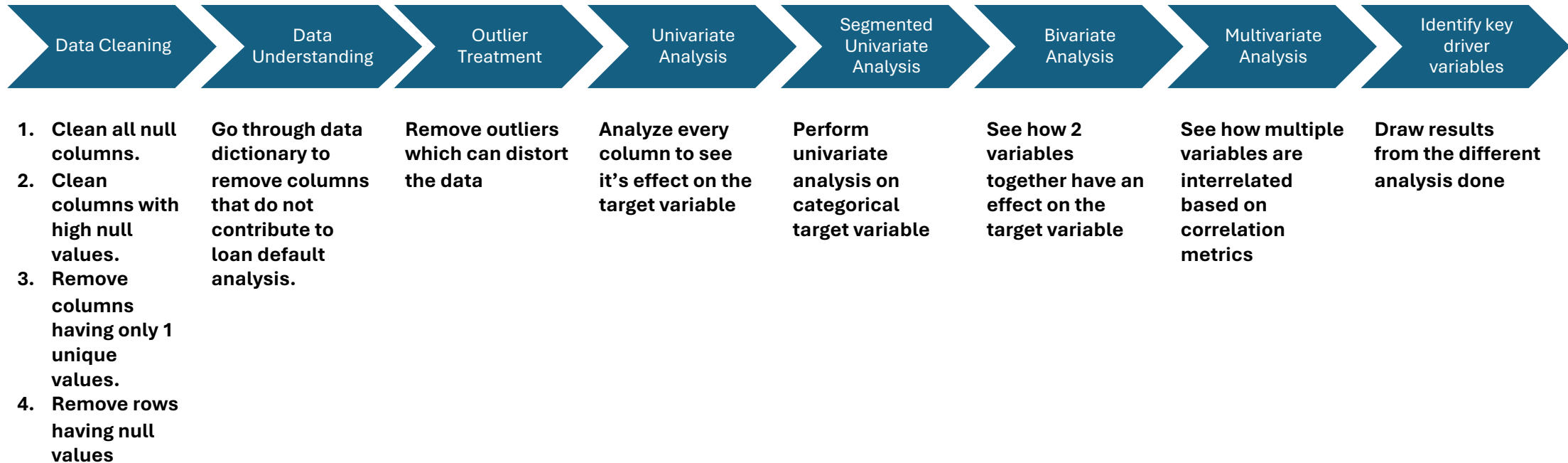
Group Members :

Jehanat Jamil

Amogh Vardhan

Case study process

| Data Cleaning | Data Understanding | Outlier Treatment | Univariate Analysis | Segmented Univariate Analysis | Bivariate Analysis | Multivariate Analysis | Identify key driver variables |

1. **Clean all null columns.**
2. **Clean columns with high null values.**
3. **Remove columns having only 1 unique values.**
4. **Remove rows having null values**

Go through data dictionary to remove columns that do not contribute to loan default analysis.

Remove outliers which can distort the data

Analyze every column to see it's effect on the target variable

Perform univariate analysis on categorical target variable

See how 2 variables together have an effect on the target variable

See how multiple variables are interrelated based on correlation metrics

**Draw results from the different analysis done**

Univariate Analysis

- Applicants with who stay in a rented accommodation have a higher chance of default
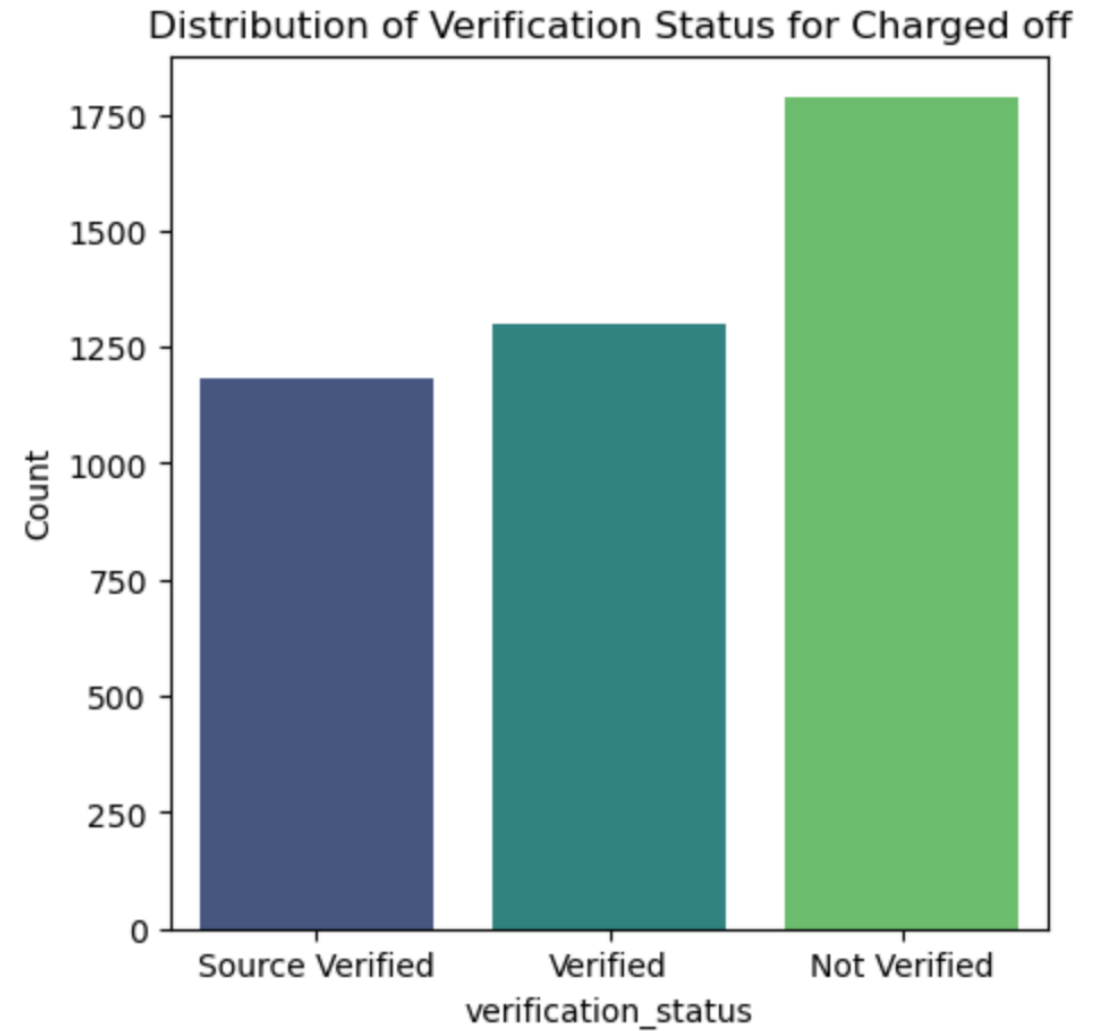

Distribution of Home Ownership where loan status is Charged off

- Applicants who have taken loan for debt consolidation are likely to default

Distribution of Purpose where loan status is Charged off

Distribution of Verification Status for Charged off
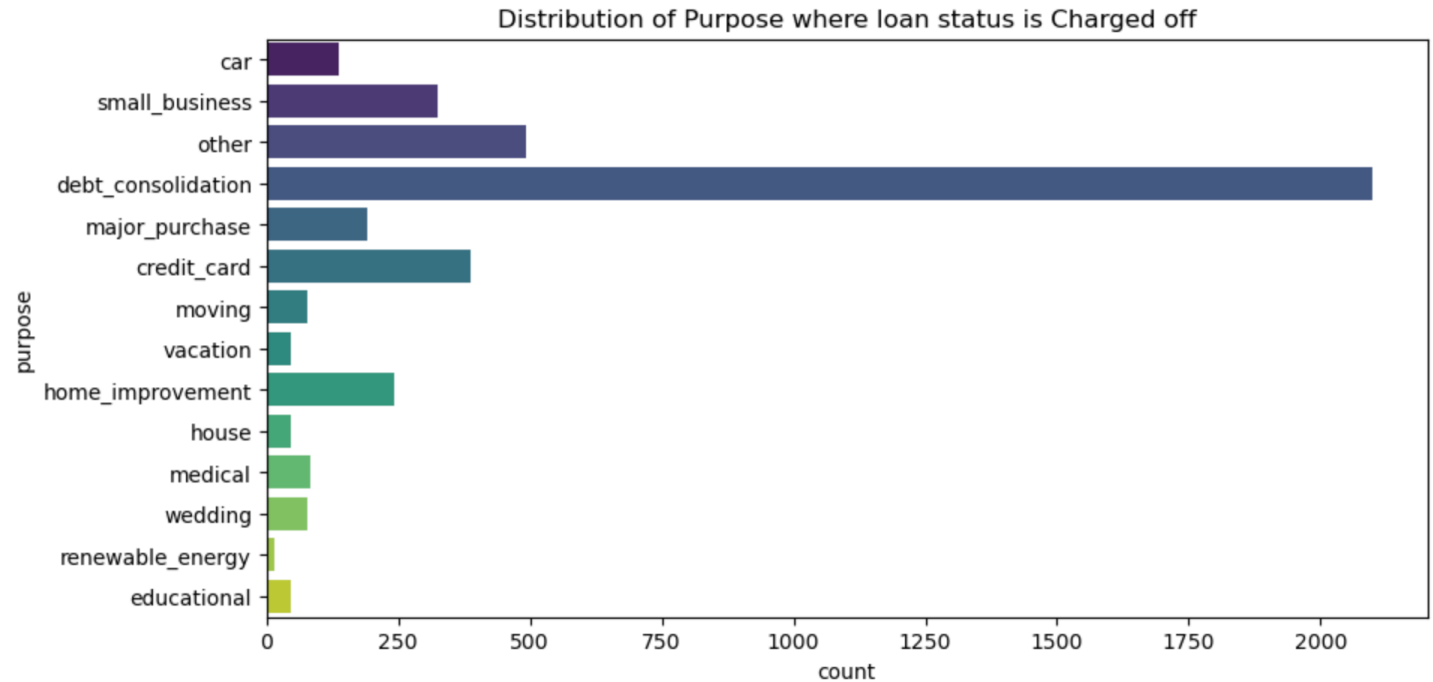
- Applicants who are not verified before giving a loan are likely to default

- Most of the defaults happened during the year of 2011. It is likely linked to the financial crisis - black Monday that caused a recession in the US.

Distribution where Loan Status is Charged Off
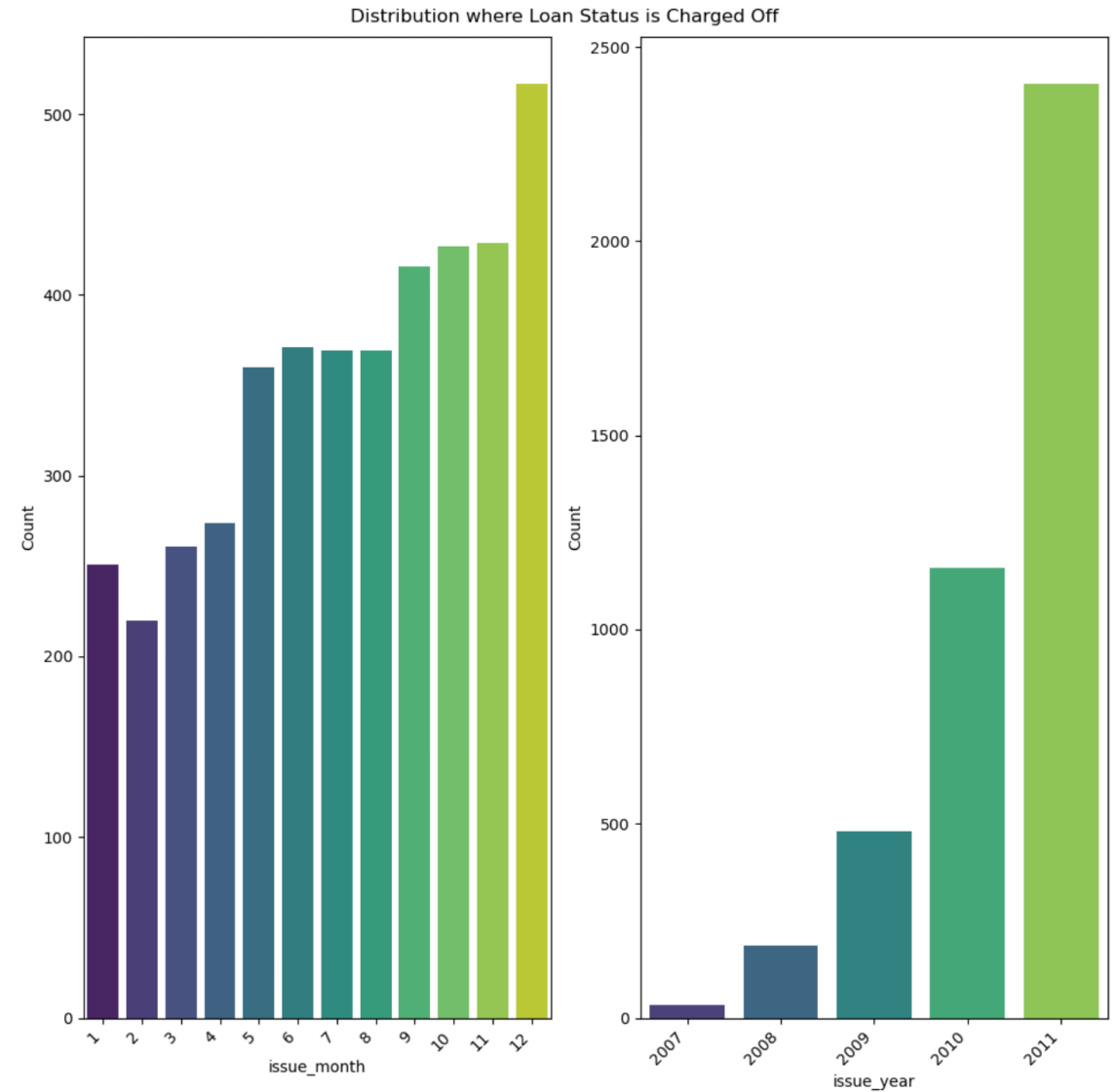
Derived Metrics

- - The number of open credit lines in the borrower's credit file is in the range between 10-19.

- - The amount of credit the borrower is using relative to all available revolving credit is 60-80.

- - The total number of credit lines currently in the borrower's credit file is 20-37.

- - The total amount of income range is 31,000-58,000.
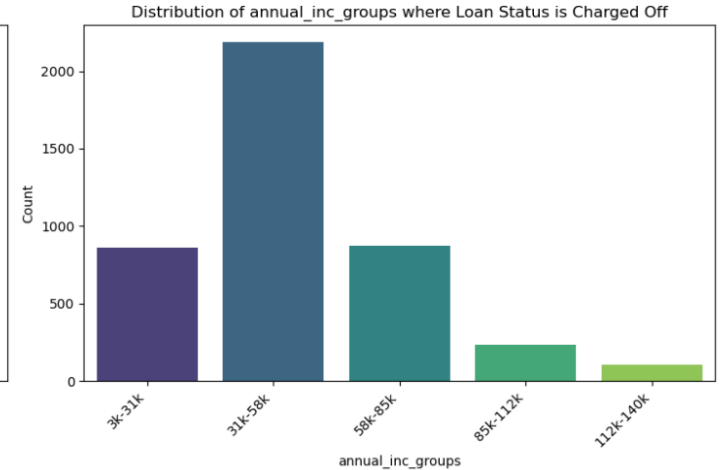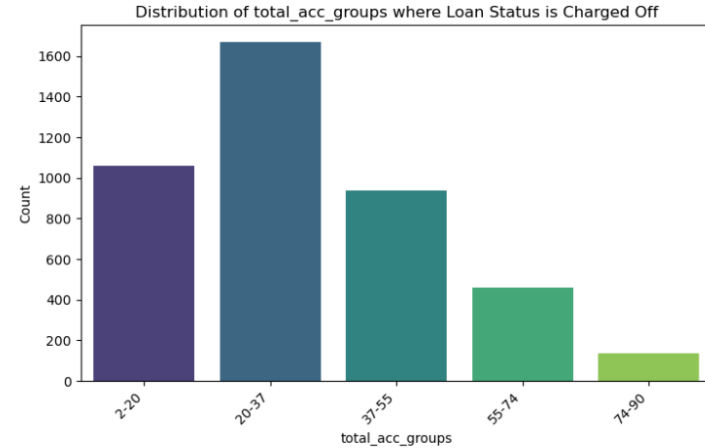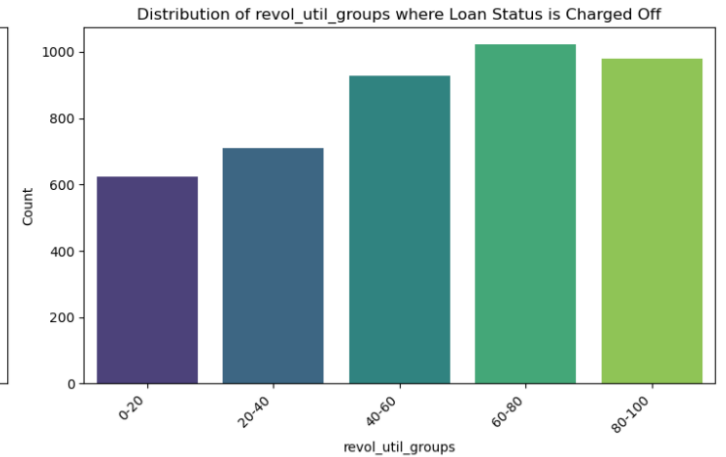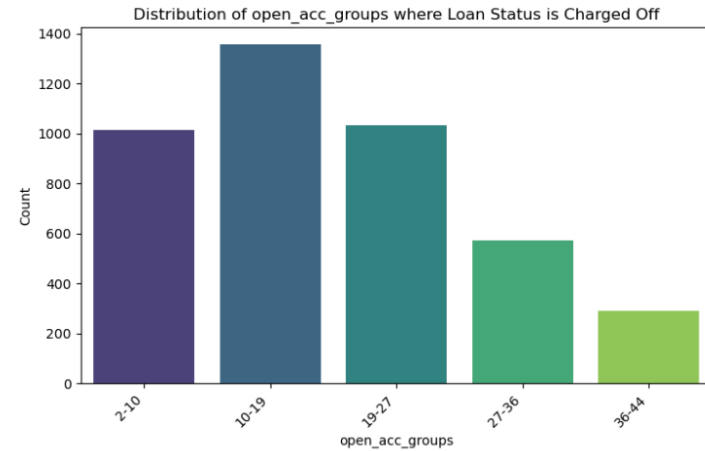
Bivariate Analysis

- loan taken for / by small businesses is the most riskiest since they have the highest default rate

Bar plot of Percentage of defaults by Loan purpose
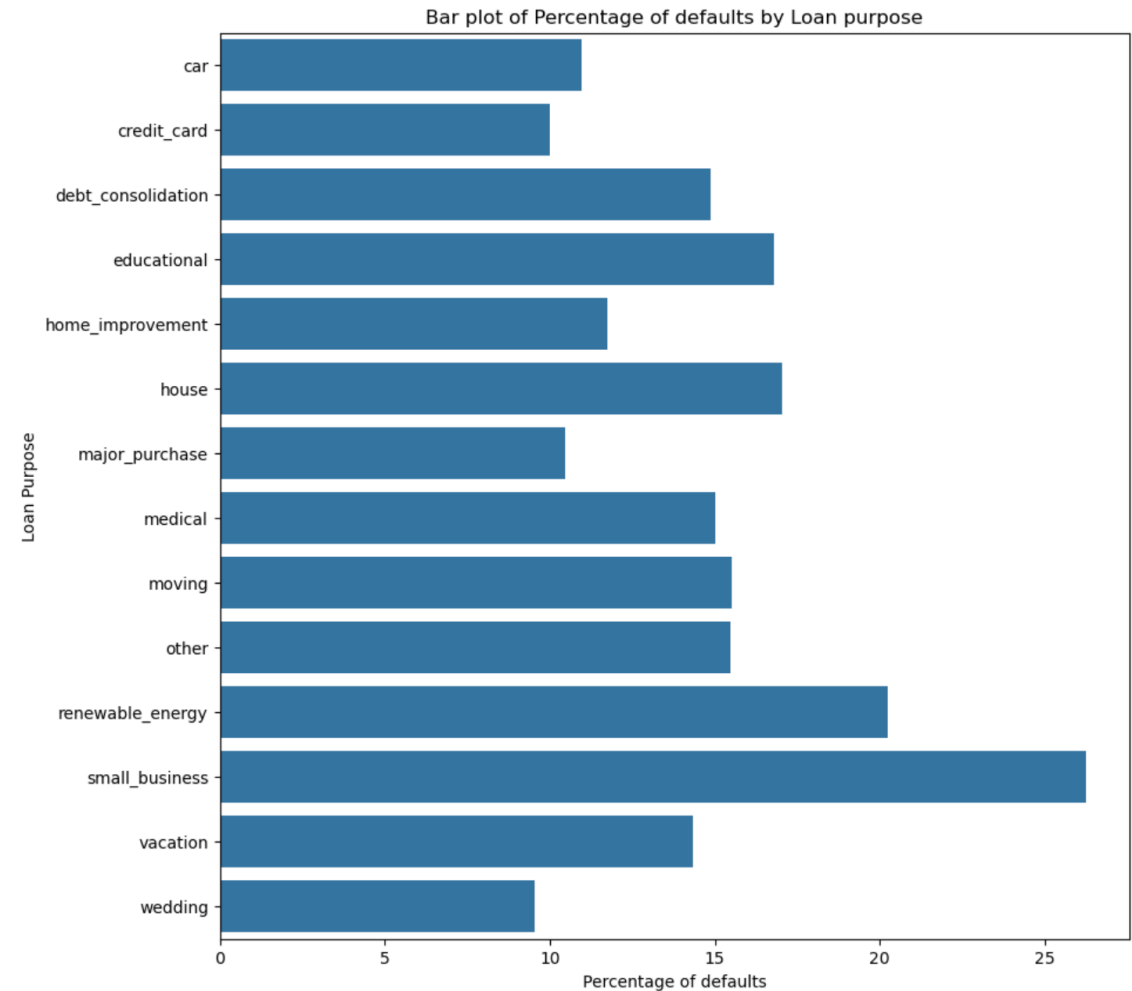
- A very high percentage of defaults are happening for applicants from state NE. More stringent review needs to be done for such applicants

Heatmap of Correlation Matrix (Numeric Variables)
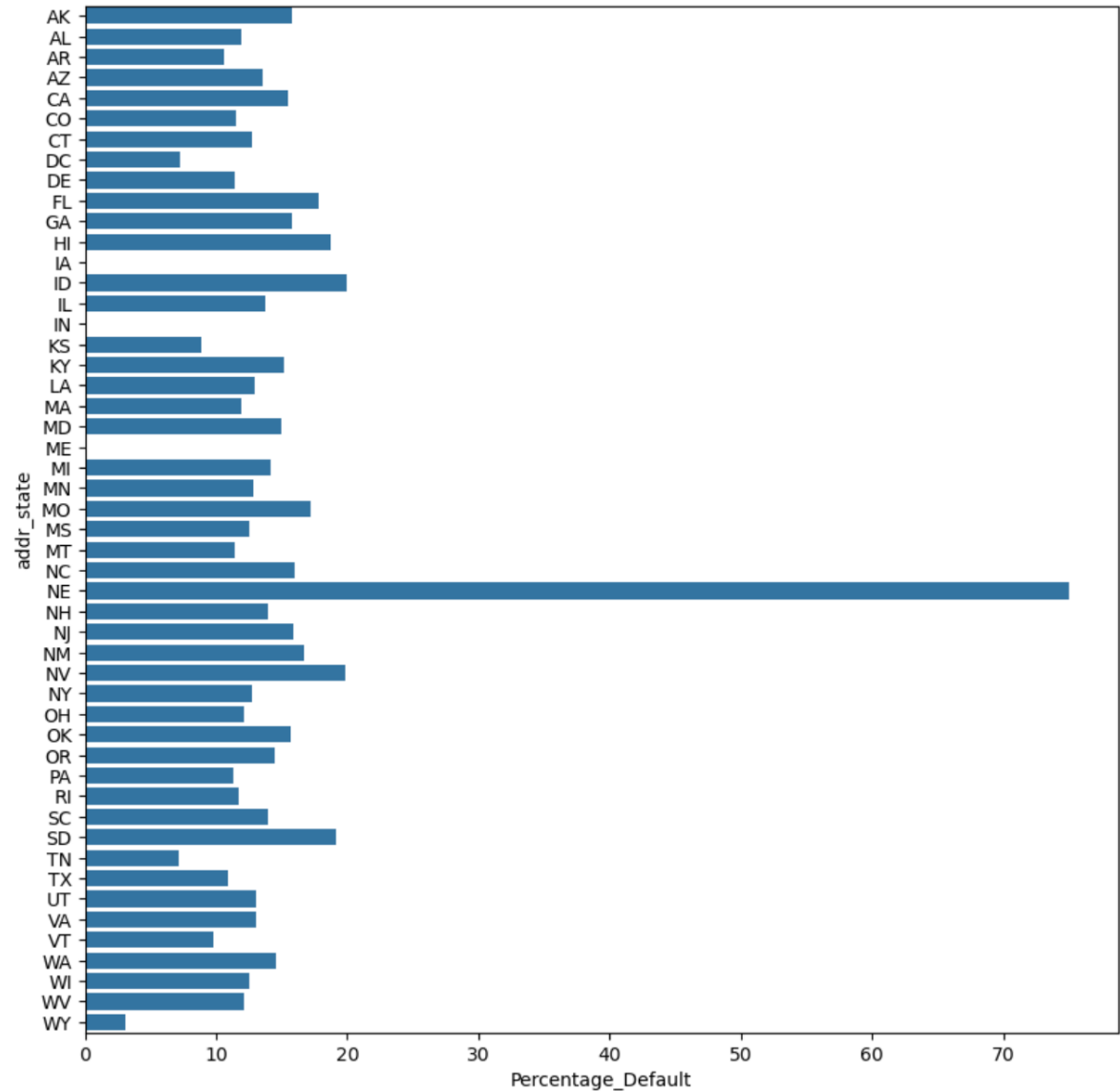
- From the correlation matrix, we see that Installment and loan_amnt are highly correlated & revol_bal, revol_util_per is moderately co-related
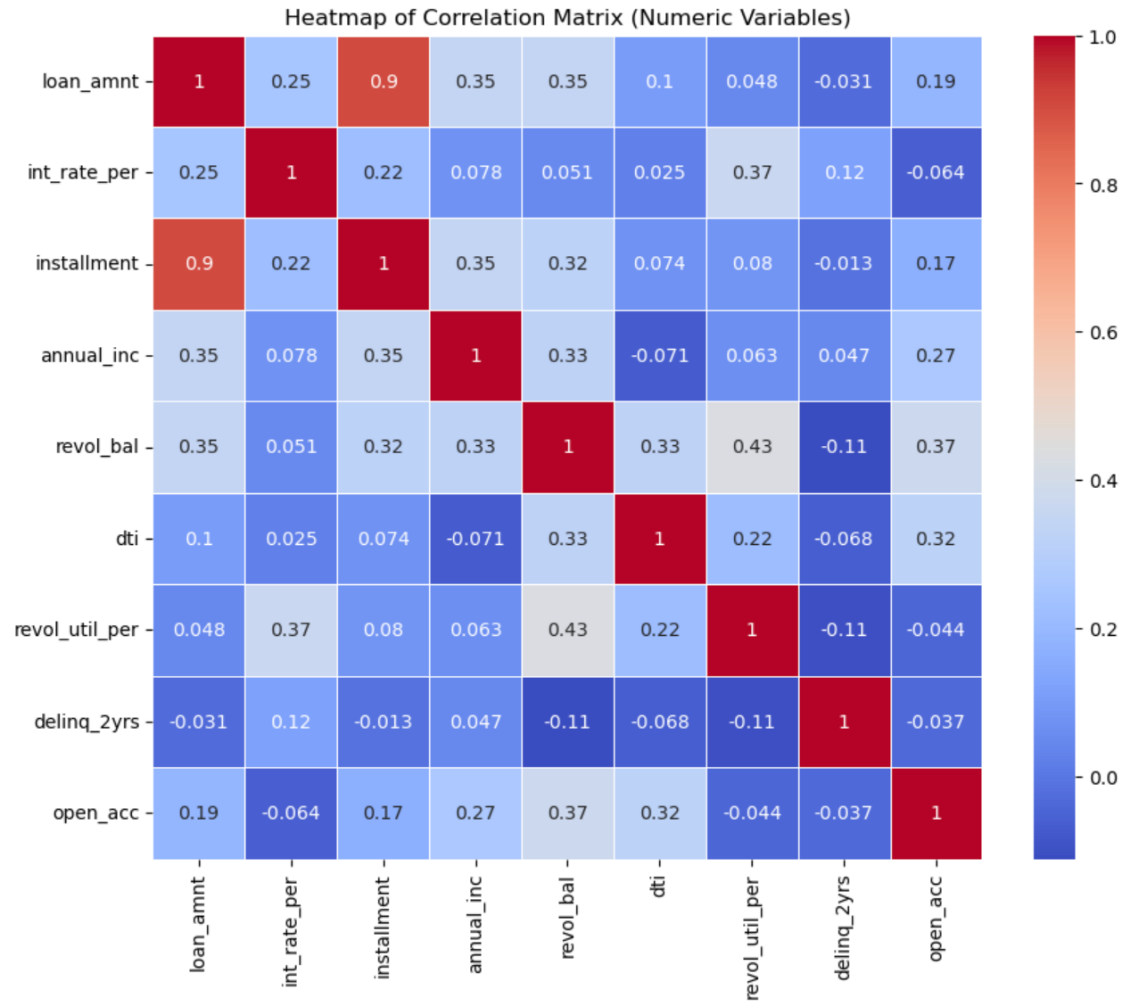
- The applicants with less interest has high probability of paying the loans and the applicants who obtain loans at higher interest rates most likely default

Annual Income Distribution by Loan Status
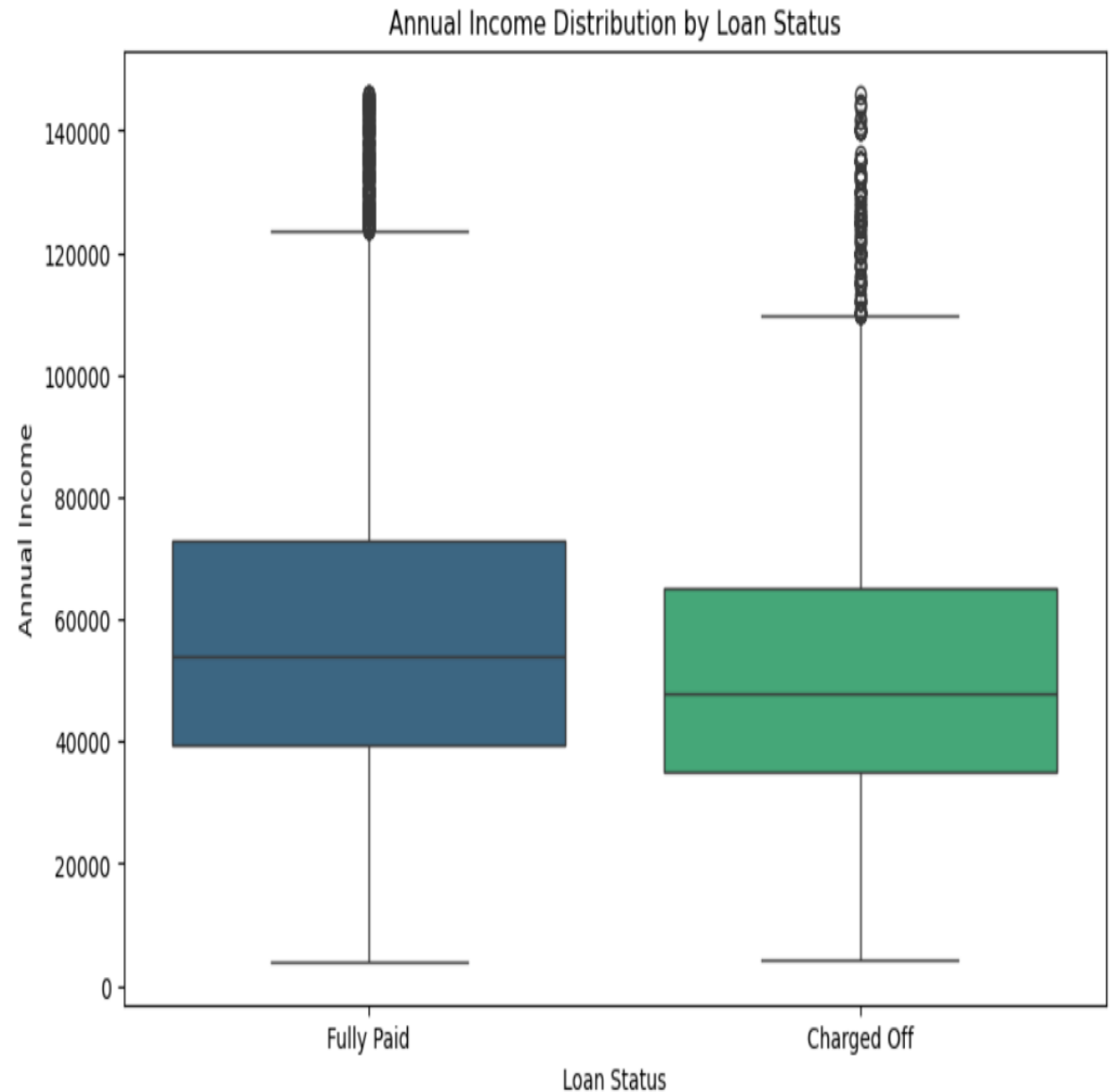
- The applicants with more annual income has high probability of paying the loans