Lending Club Case Study

EXPLORATORY DATA ANALYSIS

- KARTHIK GADEPALLI

- RADHIKA MADHANA

Table of Contents

Problem Statement

Problem:

You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
2. 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 loan data given contains information about past loan applicants and whether they 'defaulted' or not. 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.

Objective:

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Constraints:

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
    1. Fully paid: Applicant has fully paid the loan (the principal and the interest rate).
    2. Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    3. Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan.
2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset).

Data Sourcing

- Loan.csv file contains the complete loan data for all loans issued through the time period 2007 to 2011. It has 39717 Rows and 111 Columns.
- Data_Dictionary which describes the meaning of all attributes given in Loan.csv file.

Data Cleaning

- Identified 57 columns with more than 50% missing values that doesn’t participate in analysis are dropped.
- ‘desc’ and ‘title’ columns which has text/description that doesn’t participate in analysis are dropped.
- There are 1140 rows with loan_status='current' which are deleted as current loan status records are not used for analysis.
- Created a subset of loan data by removing 25 Behavioral Columns and columns that are uniqueness in nature which has only one Unique value.
- After Data cleaning process by dropping the columns not required for analysis we are left with 26 columns.

Data Manipulation

Fix Missing Values:

- Handled the missing values for categorical variables ‘pub_rec_bankruptcies’, ‘emp_length’ column by filling with mode value.
- Handled the missing values in numerical variable ‘revol_util’ column by filling with mean value.

Standardize values :

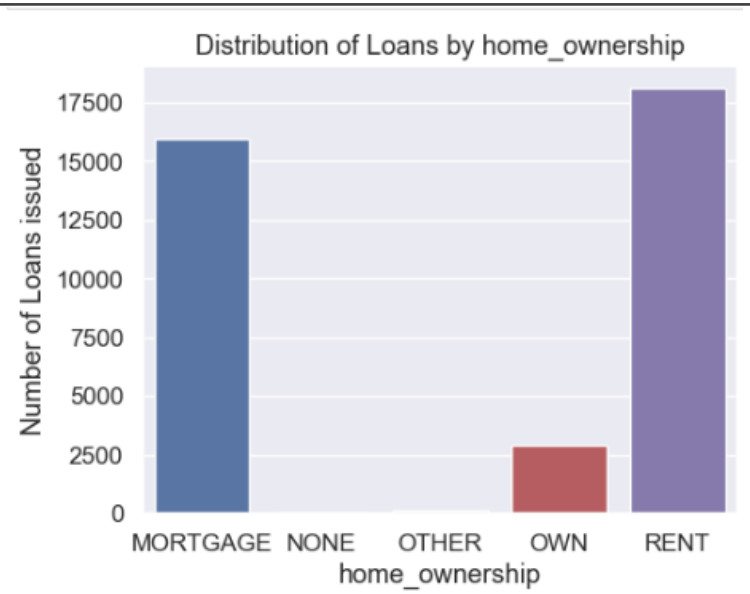
Below four columns can be categorized as continuous variables, hence converting them into int/float case to case:

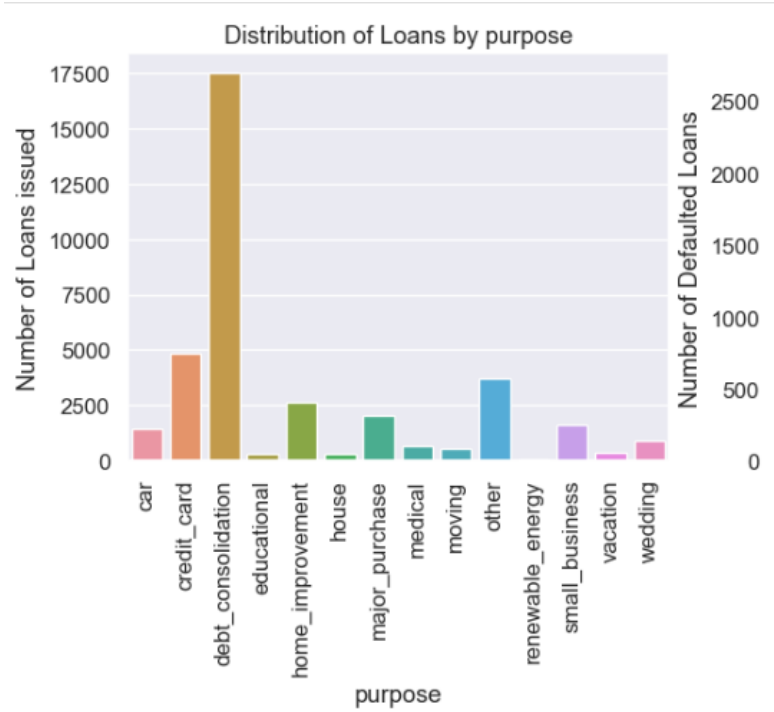
- revol_util: Standardise this value by trimming off % and convert to numeric value.
- int_rate: Standardise this value by trimming off % and convert to numeric value.
- term: Standardise this value by trimming off 'months' and convert to numeric value.
- emp_length: Standardise this value by trimming off 'years' and convert to numeric value. <1 year is treated as 0 and 10+ years is treated as 10.

Data Manipulation

- Outliers are removed for numerical columns by using quantile mechanism.
- Imputed the missing values by using mean or mode.
- All the columns contain non-null values now.
- Created derived columns 'issue_month' and 'issue_year' from issue_d columns which will be used for analysis.
- Created bins for columns 'int_rate’ 'open_acc’, 'revol_util’, 'total_acc’, 'annual_inc’, 'installment’, 'dti' and created new derived columns for them which is used for better data analysis.

Univariate Analysis - Home Ownership, Purpose

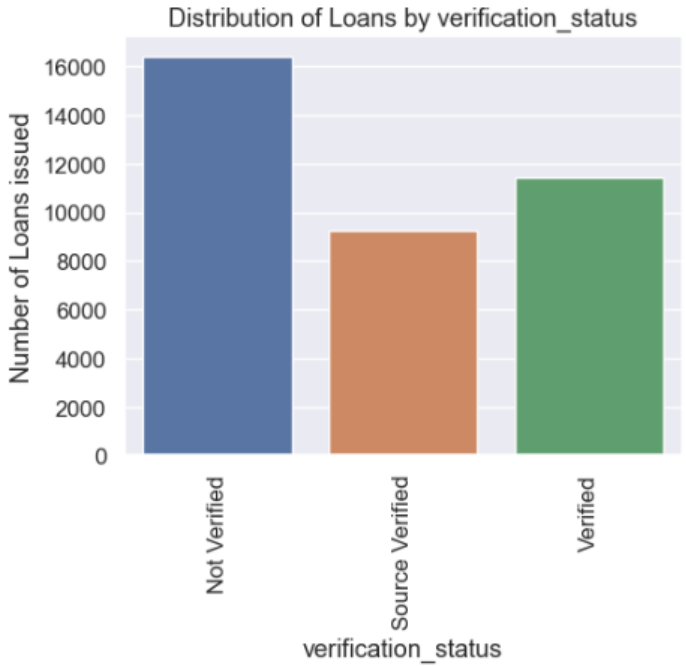


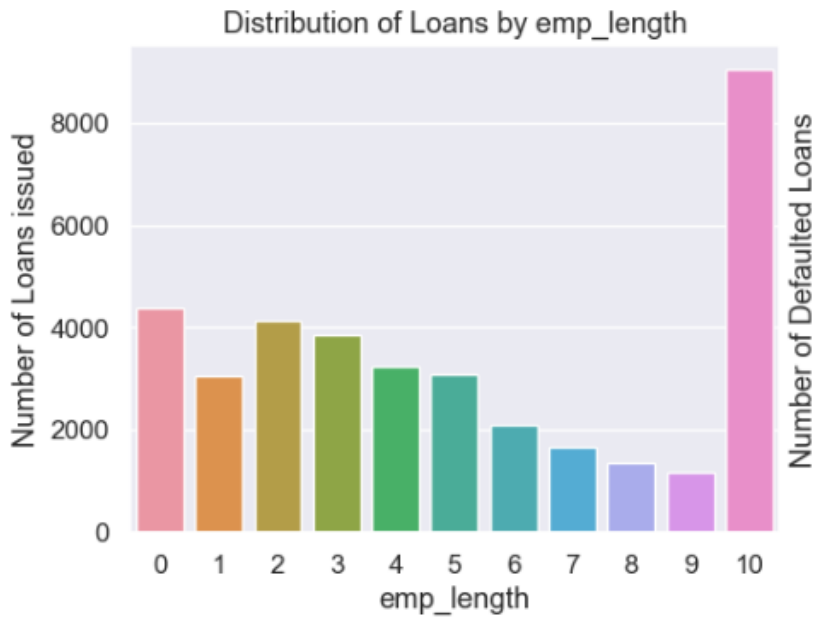


Conclusions drawn:

1. Majority of the loans were taken by applicants living on Rent followed by Mortgage.
2. Majority of loans were taken by applicants for debt_consolidation followed by credit card.

Univariate Analysis - Verification Status, Employee Length

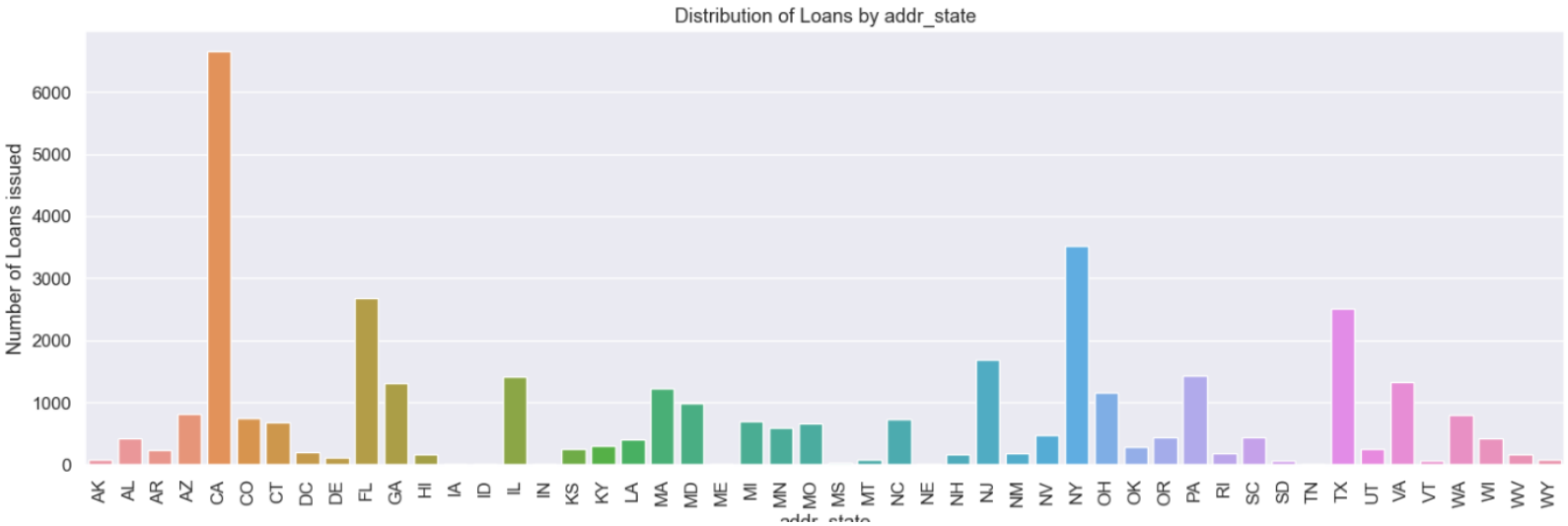


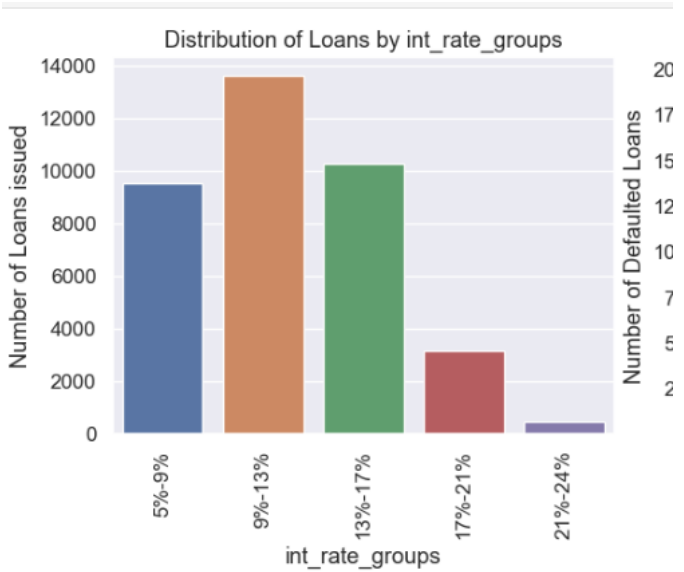


Conclusions drawn:

1. Majority of the loans taken by applicants are not verified which is alarming sign.
2. Majority of loans were taken by applicants having 10+ years of experience.

Univariate Analysis - State, Interest Rate



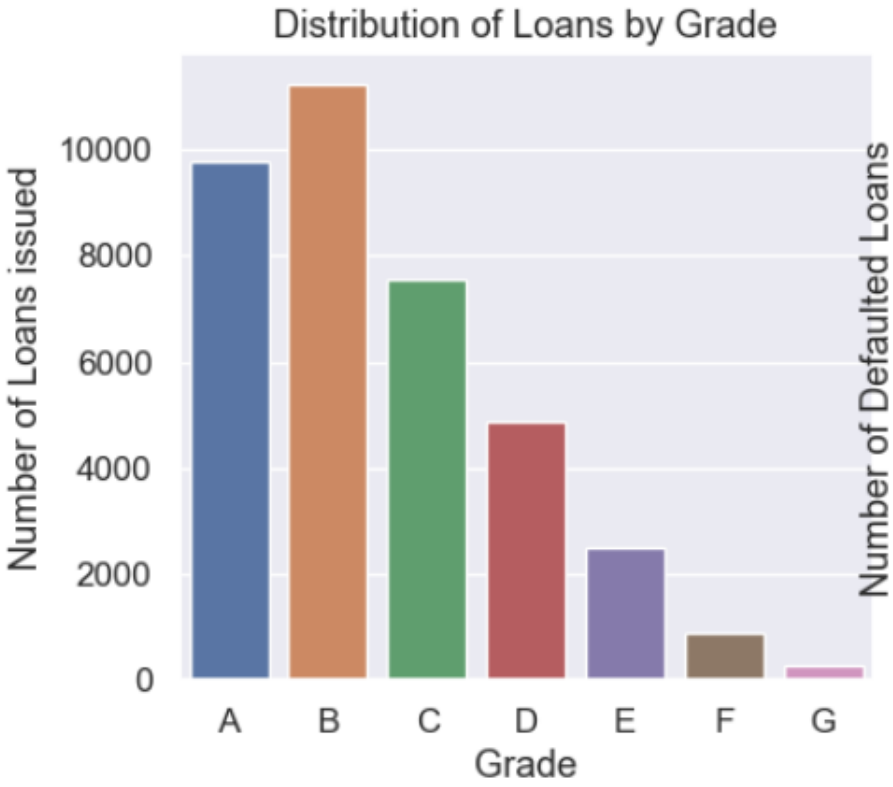


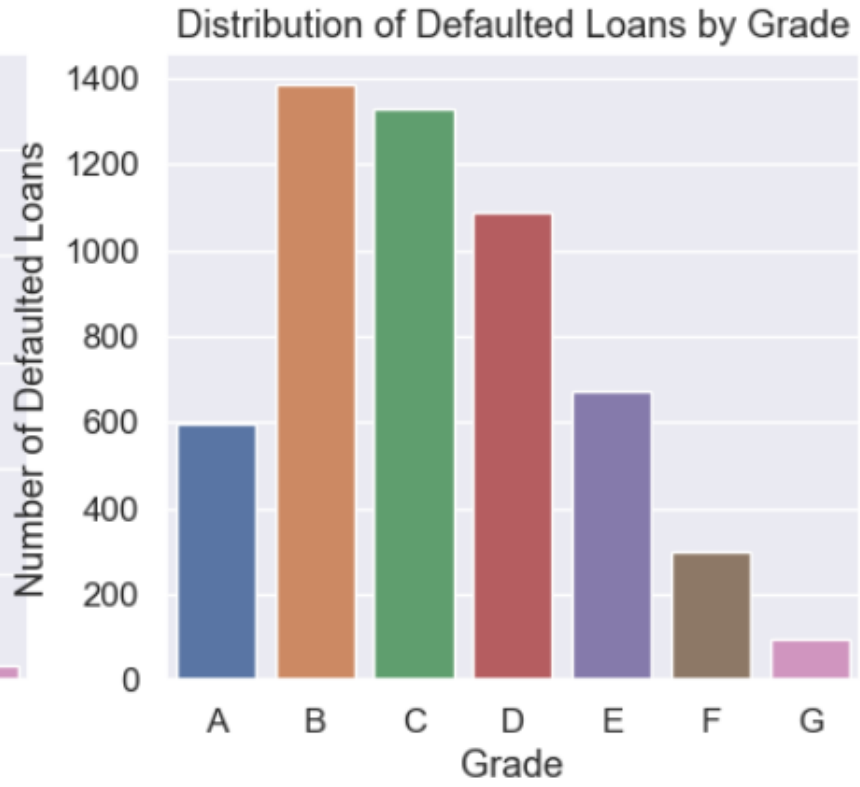
Conclusions drawn:

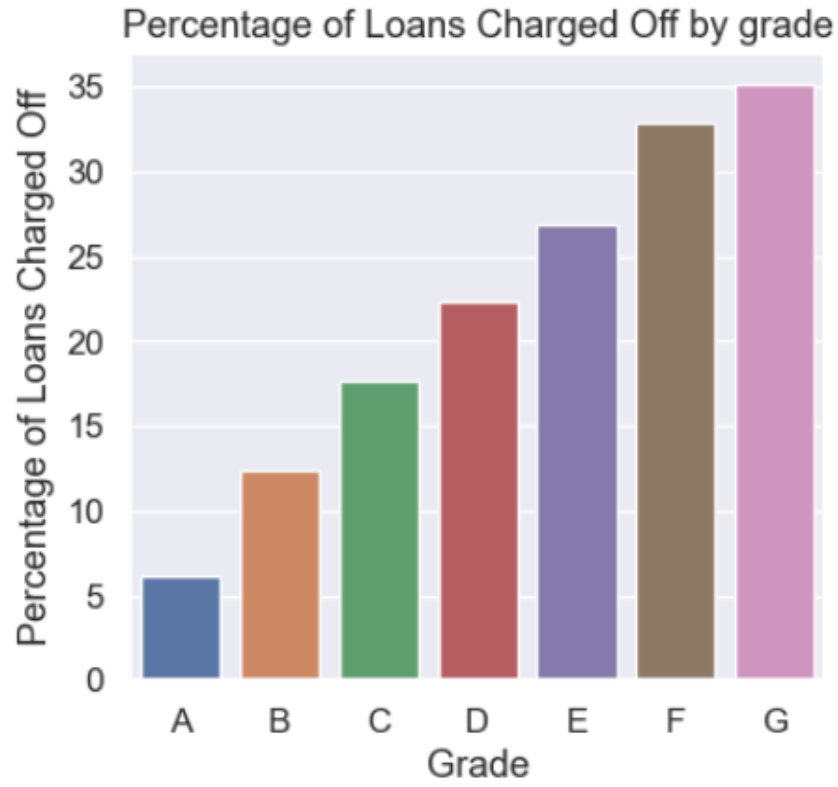
1. Majority of the loan's applicants are from CA state.
2. Majority of applicants rate of interest for loan is in the range of 9% to 13%.

Segmented Univariate Analysis
and
Bivariate Analysis for Categorical variables

Analysis of Loans by Grade



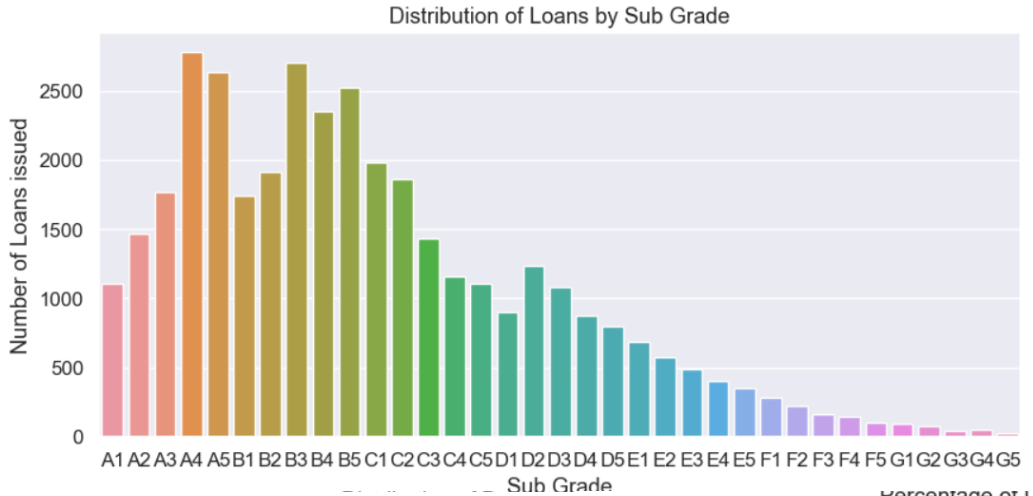


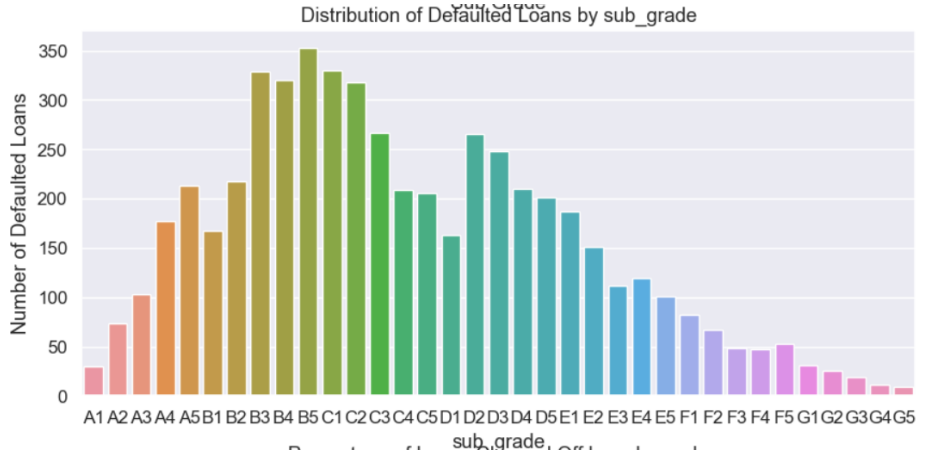


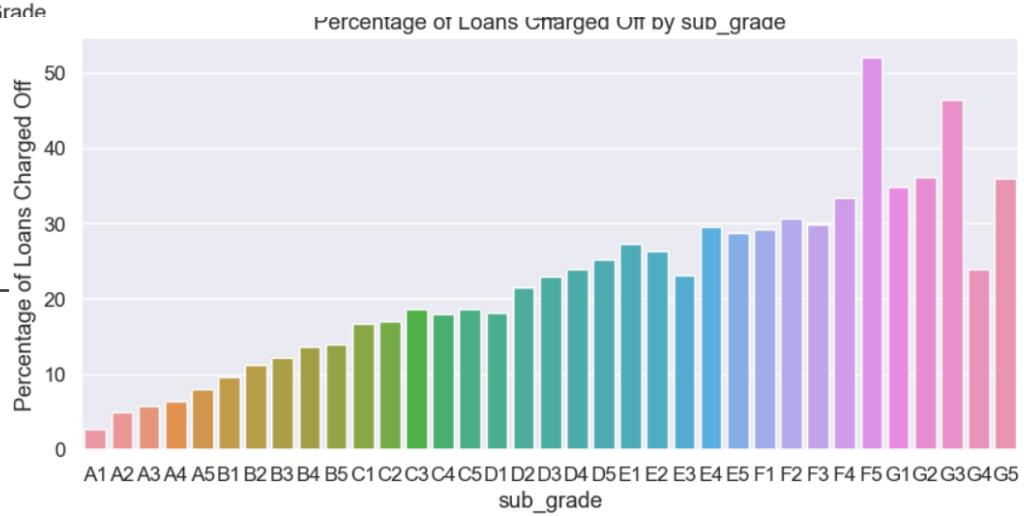
Conclusions drawn

1. More number of loans disbursed were of Grade B type, followed by Grade A type of loans.
2. More number of loans were charged-off by Grade B and C type.
3. Considering the percentage of loans charged-off Grade wide, clearly, we can see a trend that Grade G loans were riskier followed by F, E, D, C, B, A.

Analysis of Loans by SubGrade



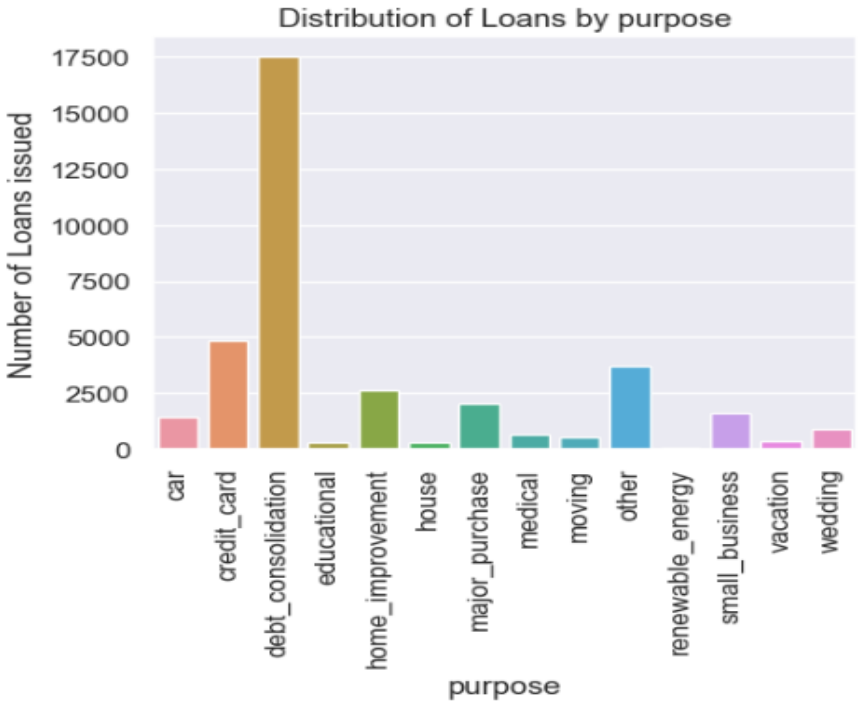


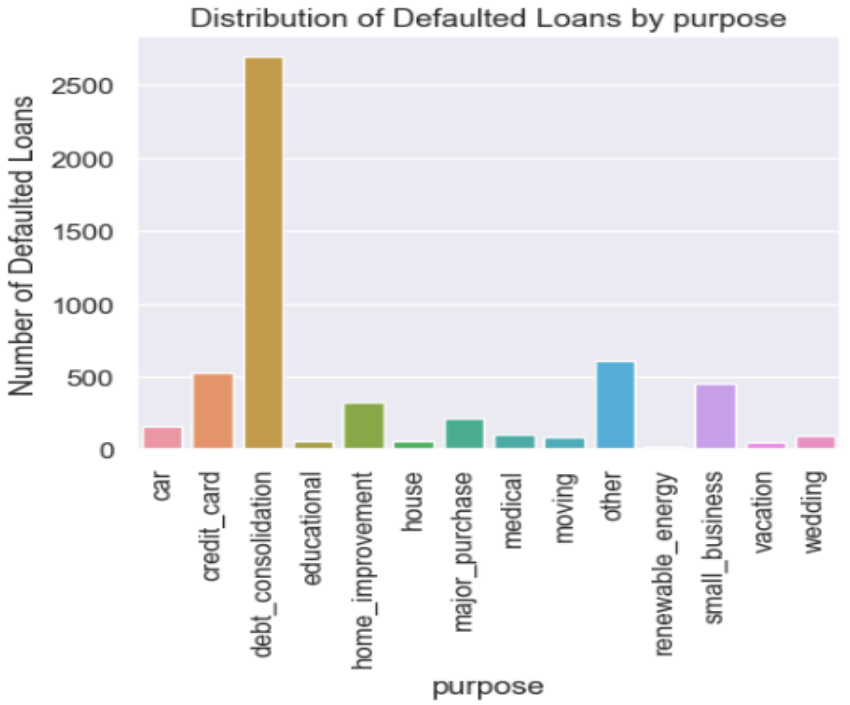


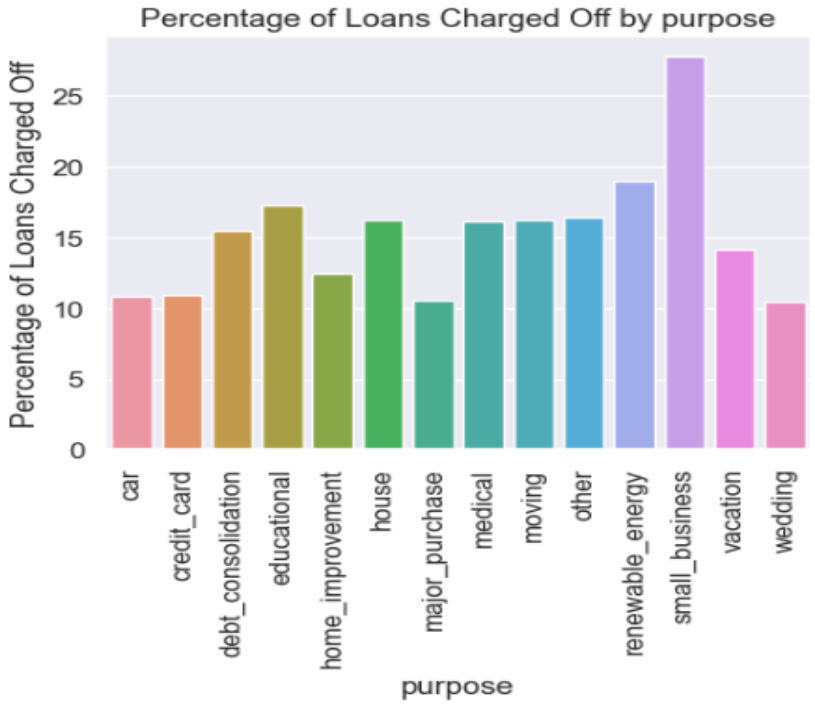
Conclusions drawn

1. More number of loans disbursed were of Sug-grades A4, B3, A5, B5, followed by others.
2. More number of loans defaulted were from Sub-grades B5, followed by B3, C1, B4, C2.
3. More than 50% of F5 sub-grade type loans were defaulted, followed by G3 sub-grade type.

Analysis of Loans by Purpose



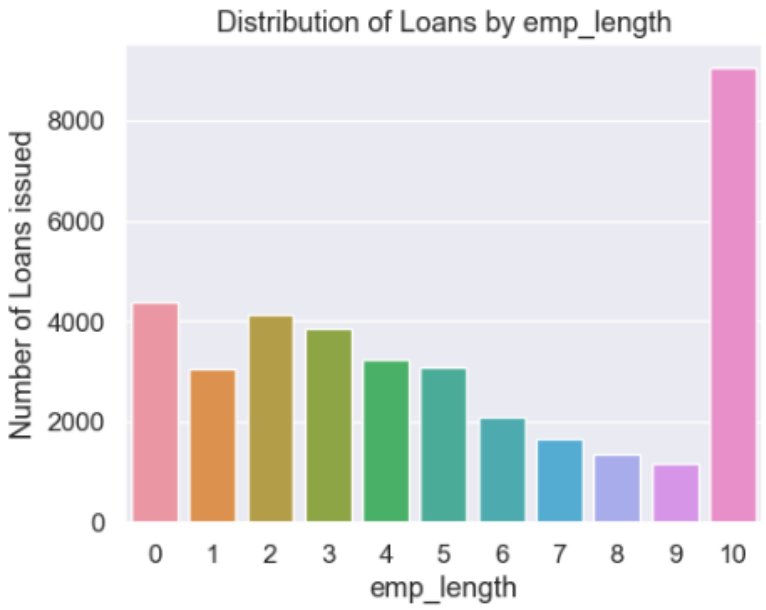


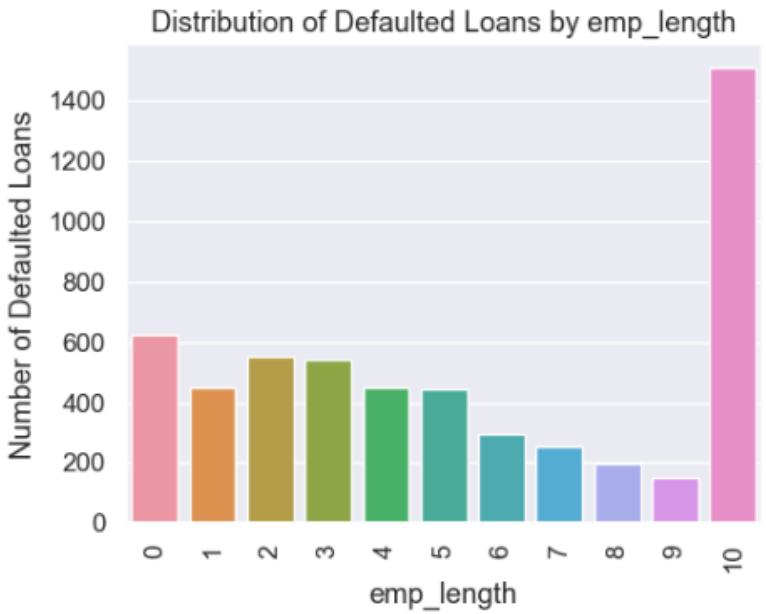


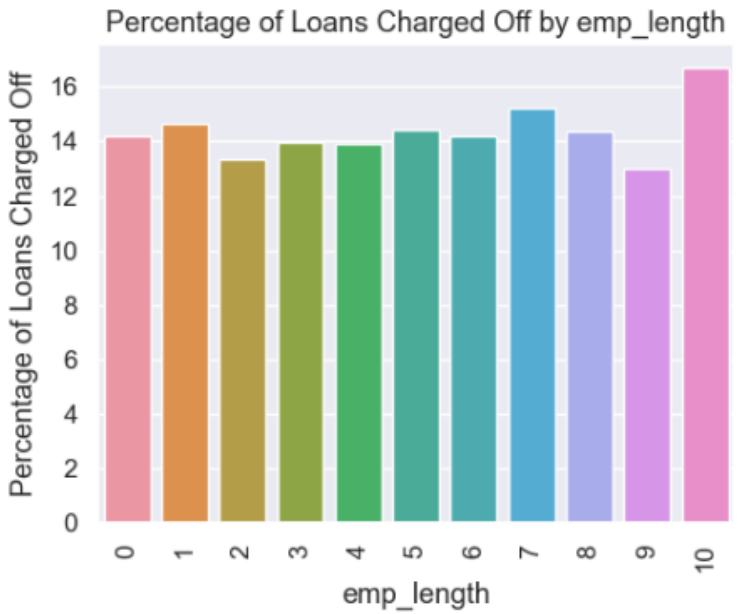
Conclusions drawn

1. Predominantly, loans were taken by those who want to use them for debt-consolidation. This gives a picture that they are trying to rotate the loan amount from one to other, which is not bad sign for the company
2. Loans were also taken by those who wish to pay the credit card bills.
3. Number of defaulters were highest in debt_reconsolidation category as the number of loans taken were highest in this category.
4. Percentage of defaulters were highest in the category of 'small_business' purpose.

Analysis of Loans by Employee length



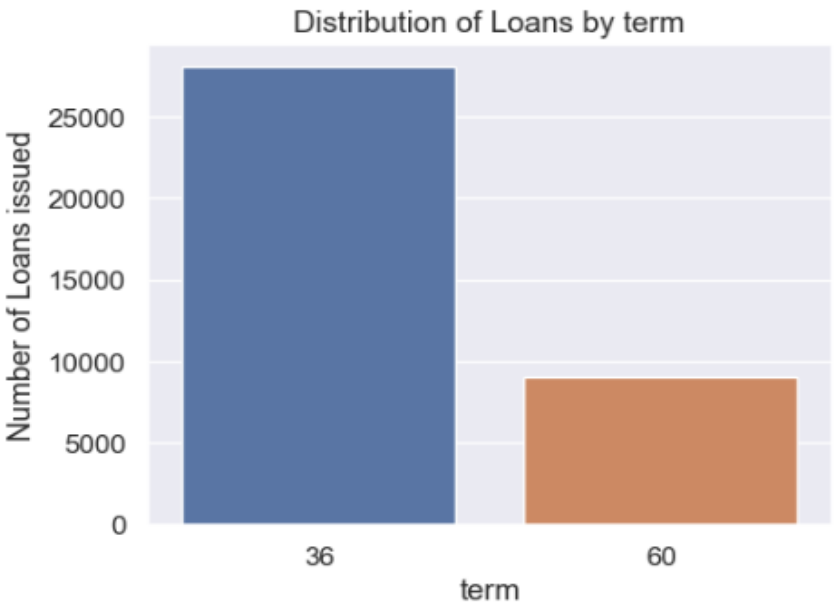


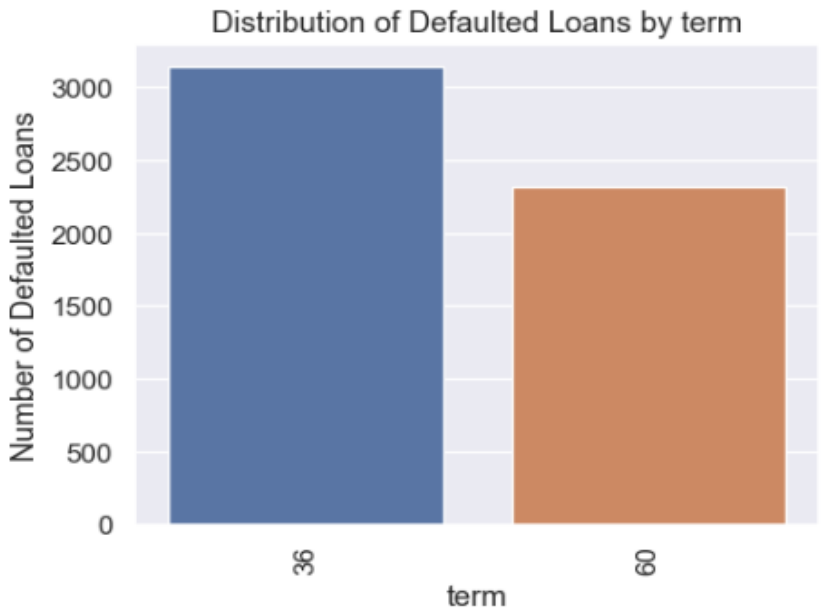


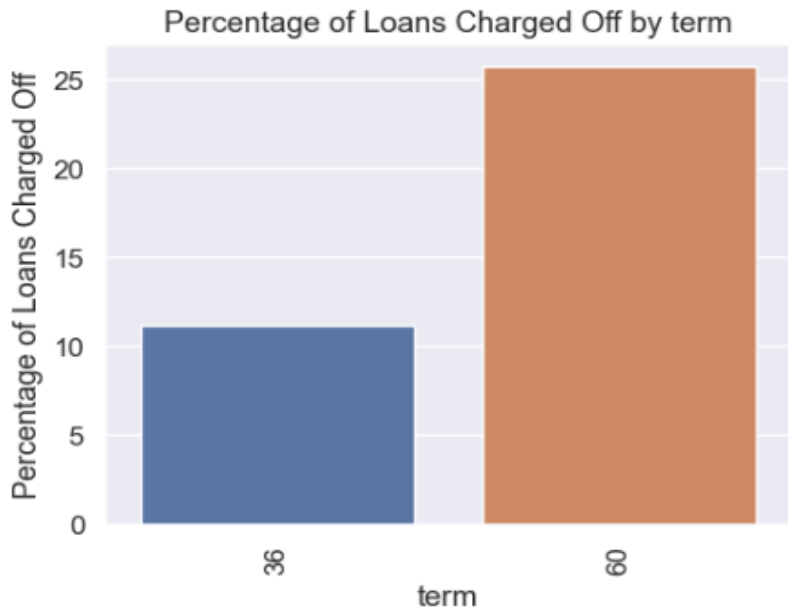
Conclusions drawn

1. Large number of loans were disbursed to those whose employment is 10+ years.
2. Another interesting point is that those with <1 year employment were more prone to taking loans.
3. Defaulters' percentage is almost same across all emp_length categories. 10+ years category is slightly on the higher side.

Analysis of Loans by term



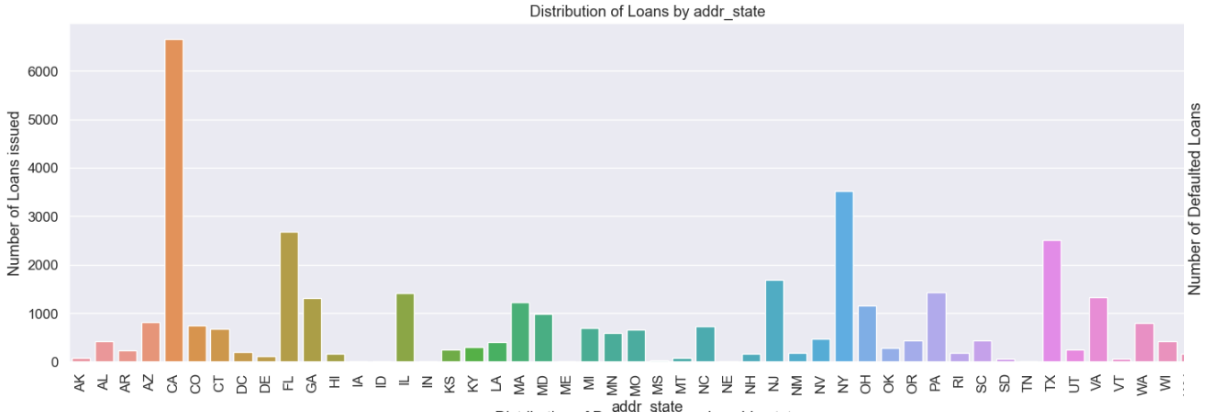


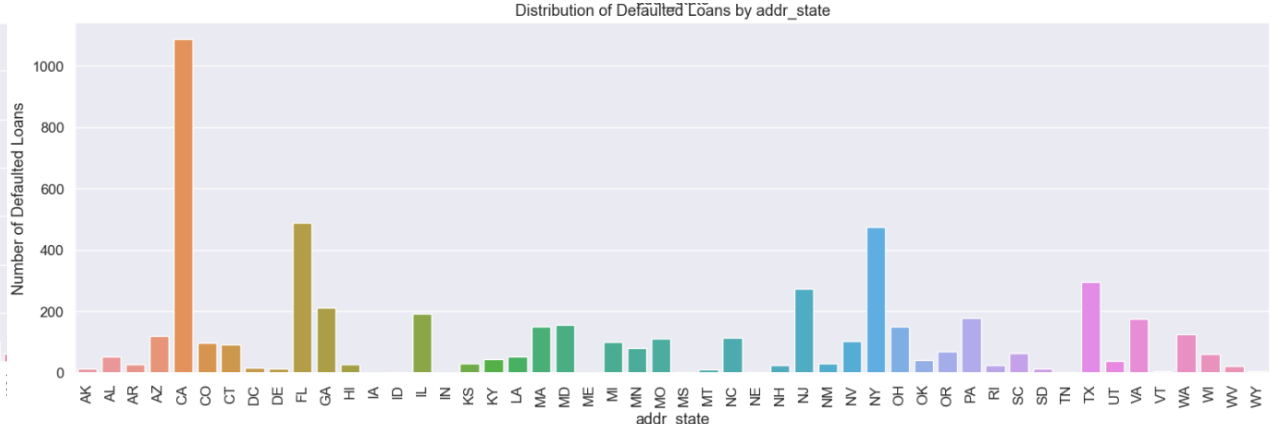


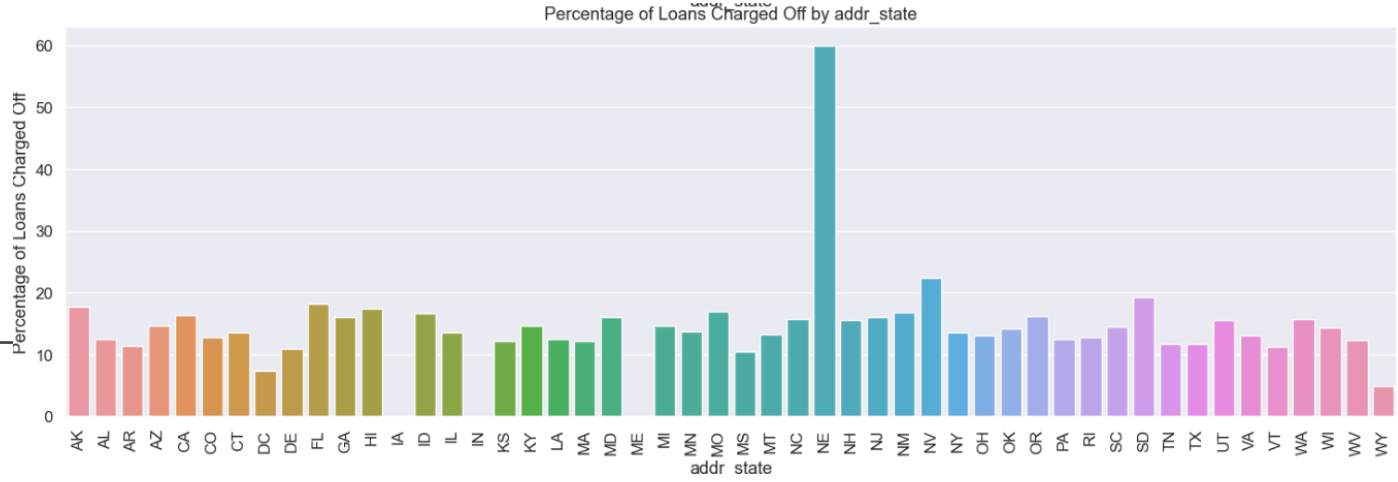
Conclusions drawn

1. Majority of the loans were taken and defaulted by those who opted for shorter loan term, i.e., 36 months.
2. In contrast percentage of defaulters were high in those who opted for 60 months term.

Analysis of Loans by State





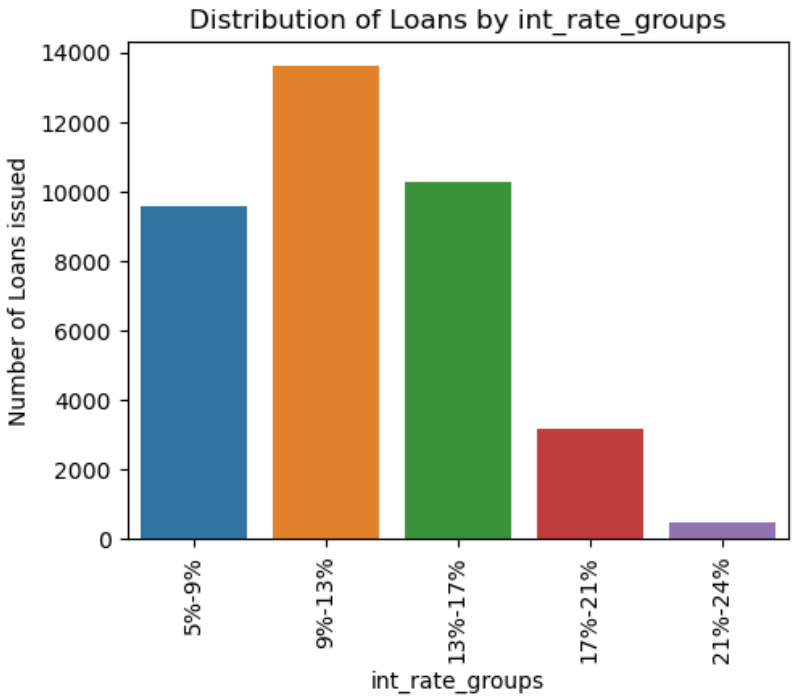


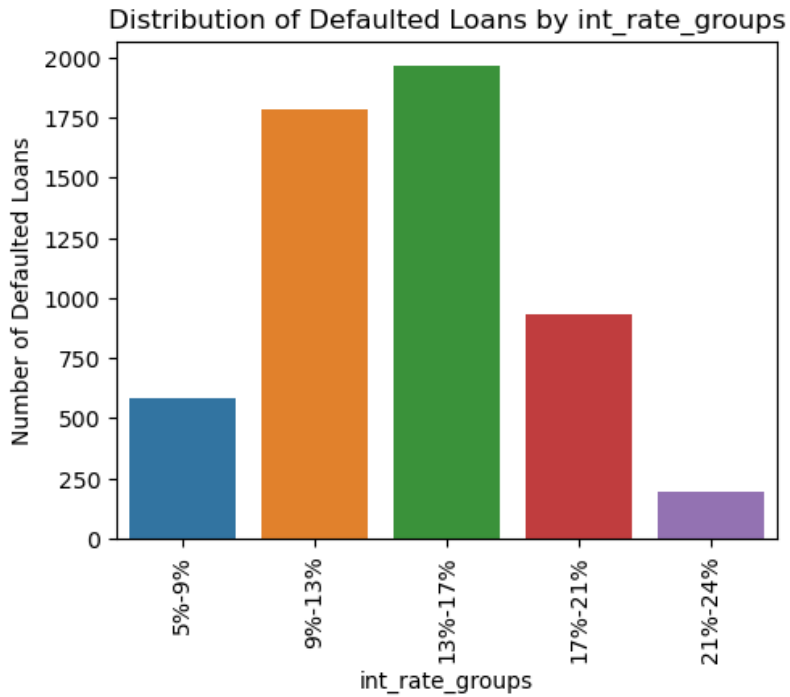
Conclusions drawn

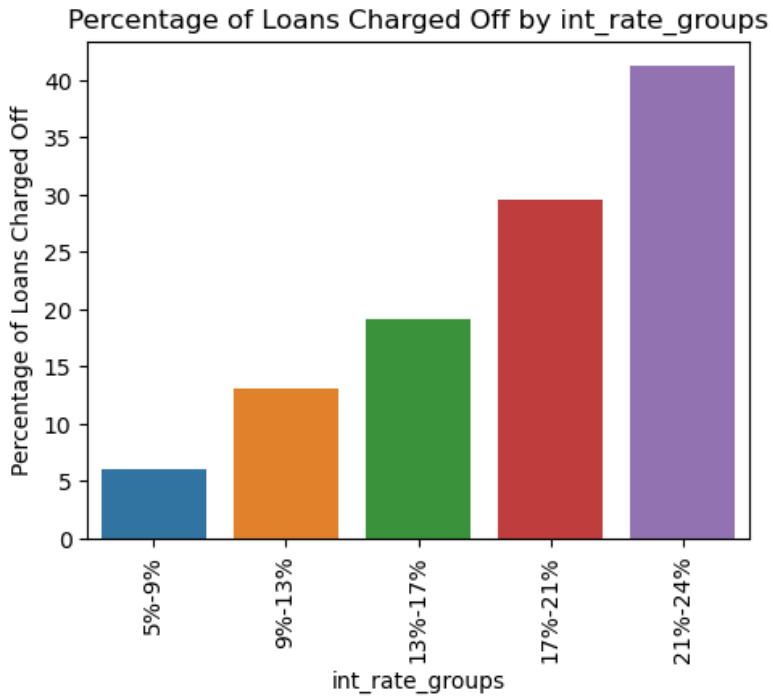
1. Majority of the loans were taken from CA state followed by NY state and TX state owing to the strong business in these states.
2. NE state defaulter percentage is quite high ~60%.

Bivariate Analysis of continuous variables using binning

Analysis of Interest rate groups



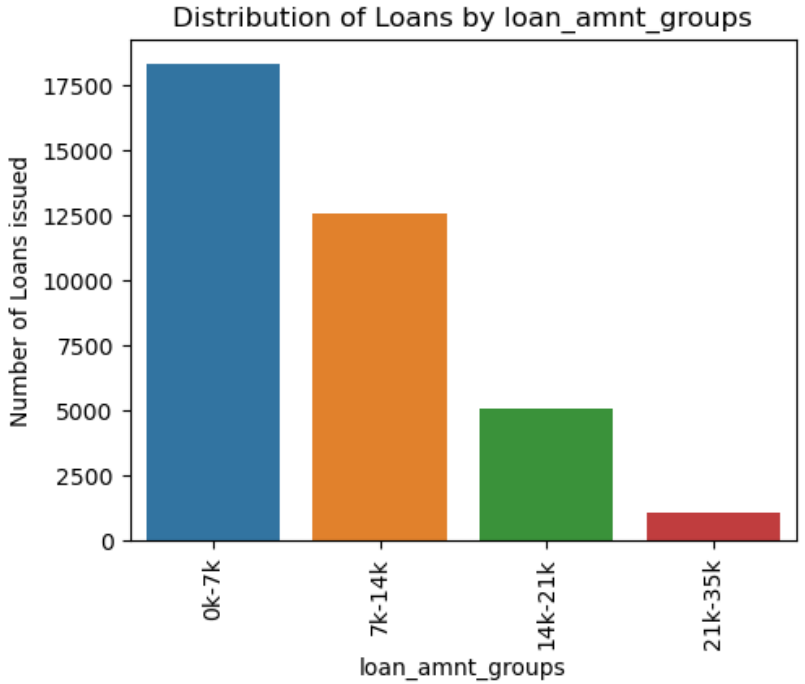


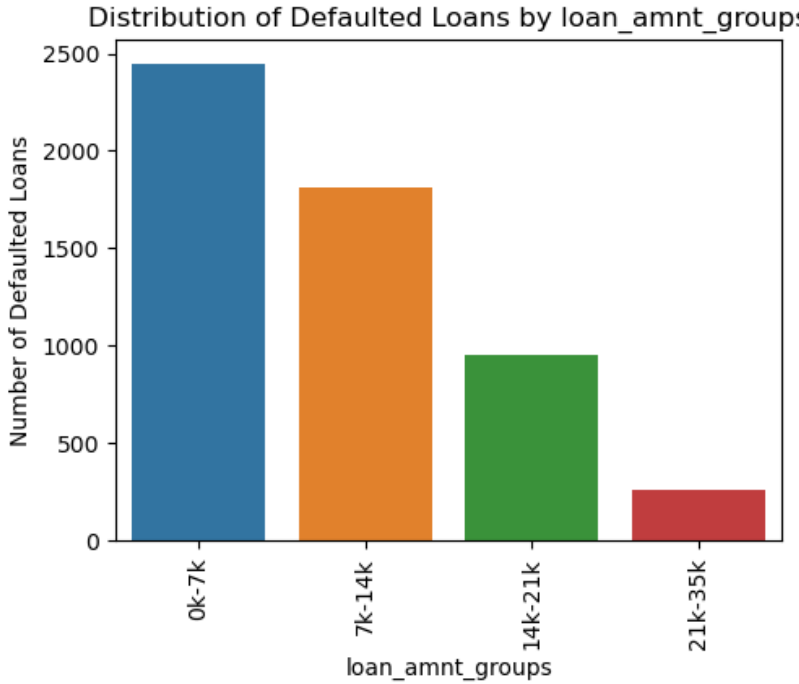


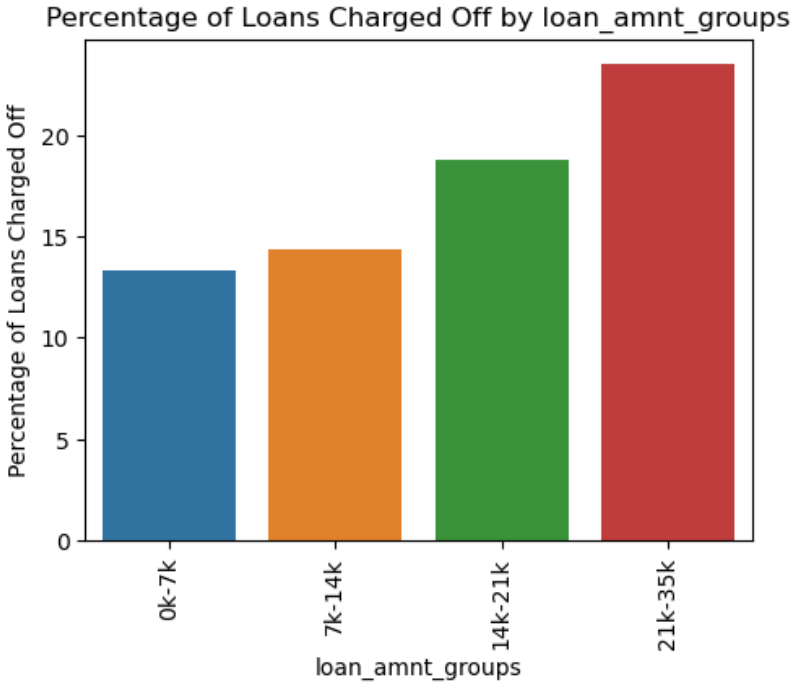
Conclusions drawn

1. More loans were disbursed under 9%-13% interest rate bracket.
2. More loans were defaulted under 13%-17% interest rate bracket.
3. We can clearly see the trend that charged-off loan percentage is increasing as the rate of interest is increasing.

Analysis of Loan amount groups



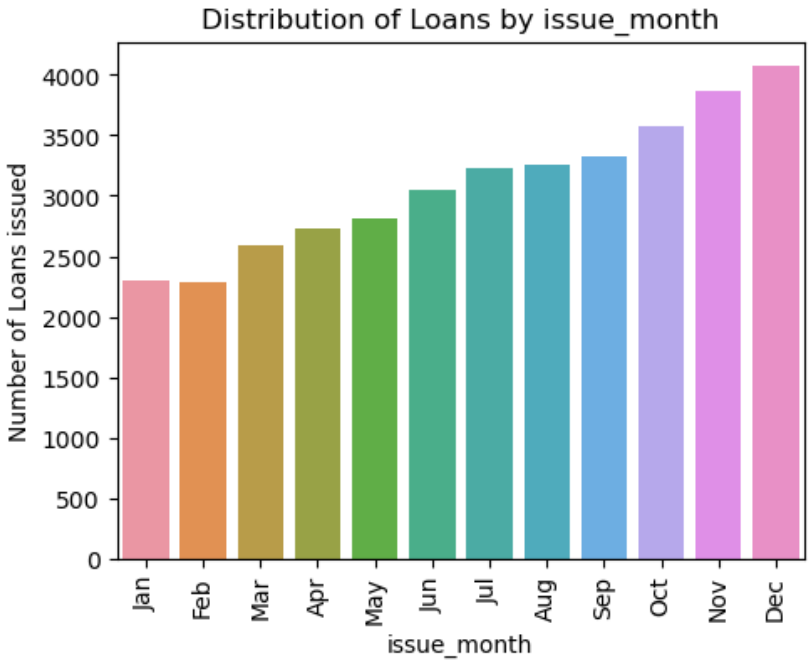


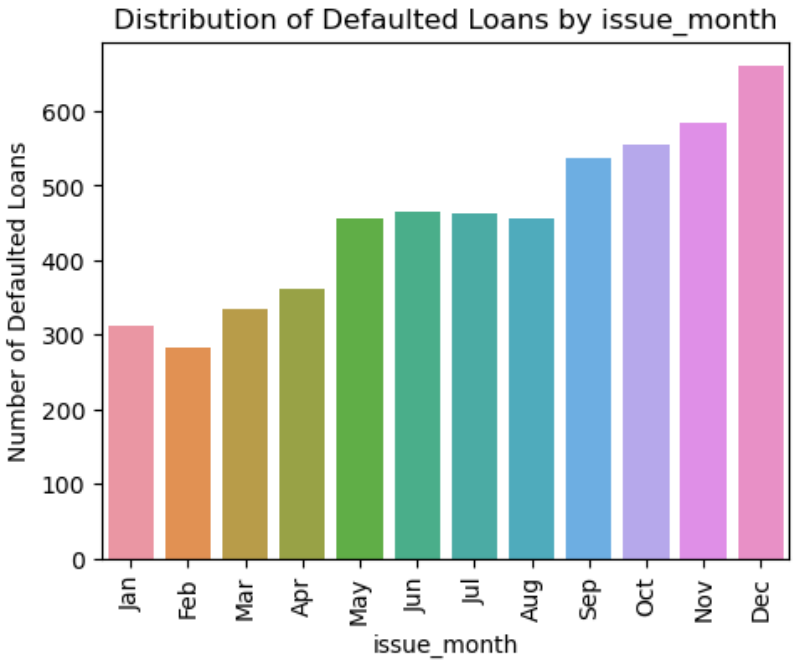


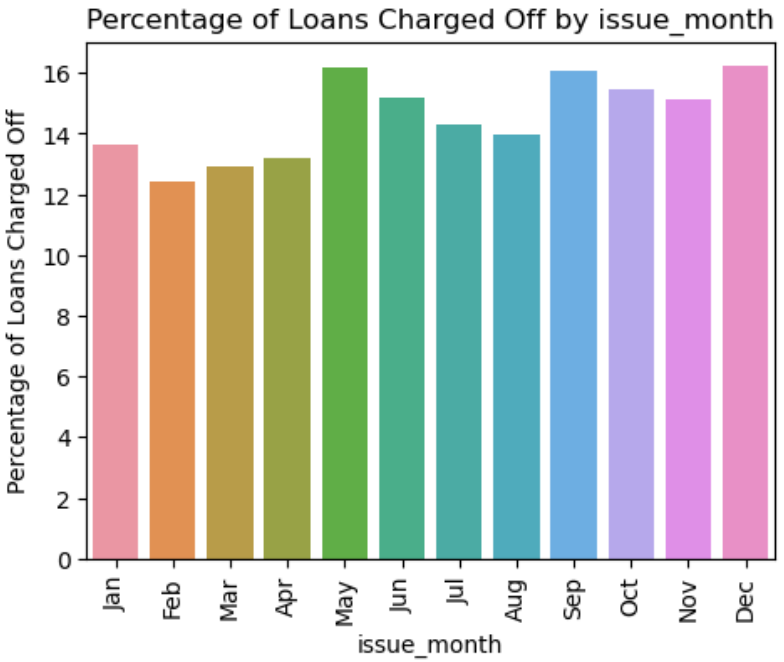
Conclusions drawn

1. More loans were disbursed to with loan amounts under 7k.
2. Trend is that % of defaulting increases as the loan amount increases, >20% of loans with loan amount >21k were **defaulted**.

Analysis of Loan issue month



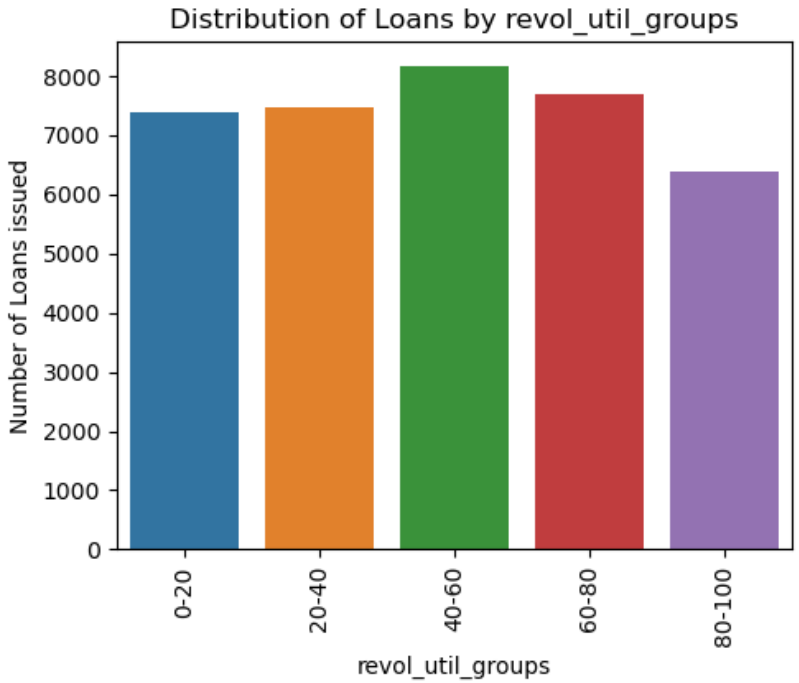


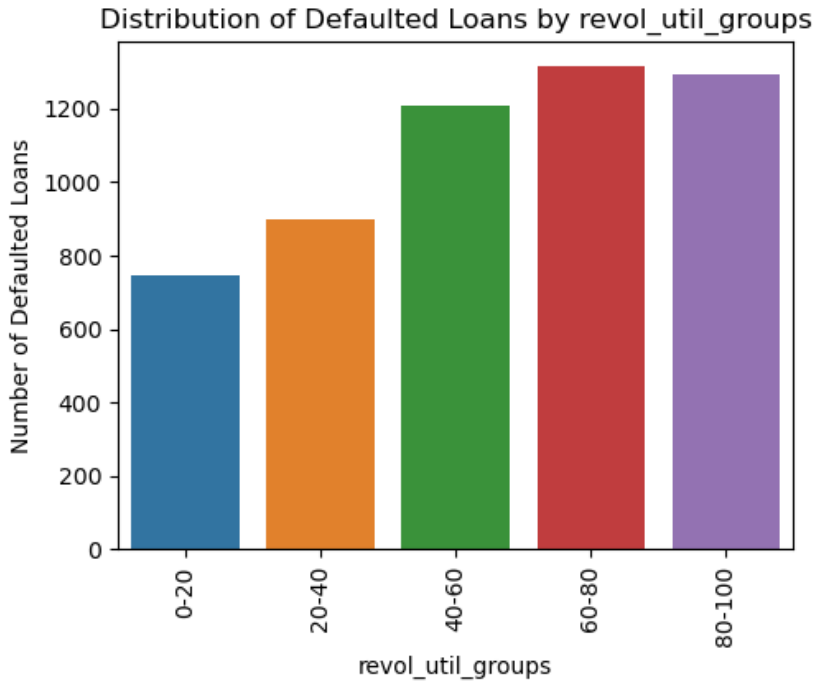


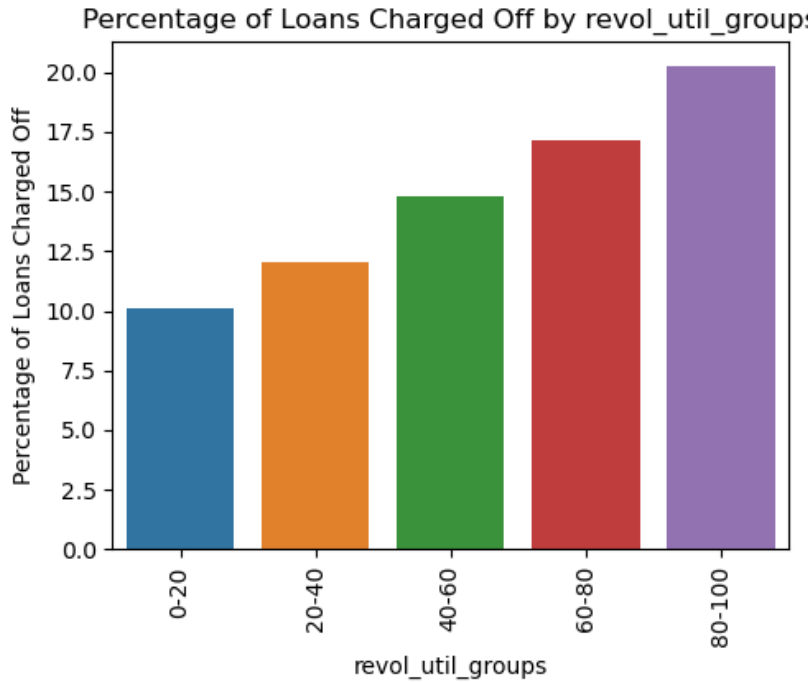
Conclusions drawn

1. Number of loans taken and defaulted kept on increasing from Jan to Dec and peaked in Dec.
2. Percentage of loans defaulted were highest when issued in May, Sep and Dec.

Analysis of Revolving utilization groups





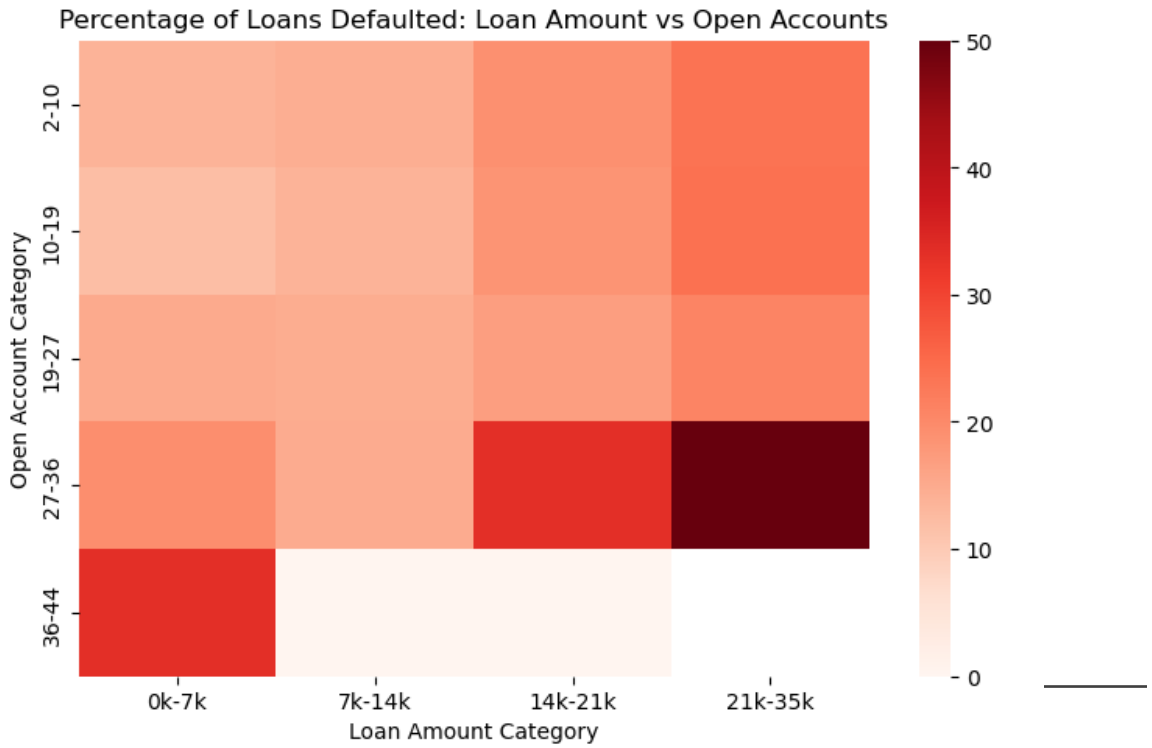


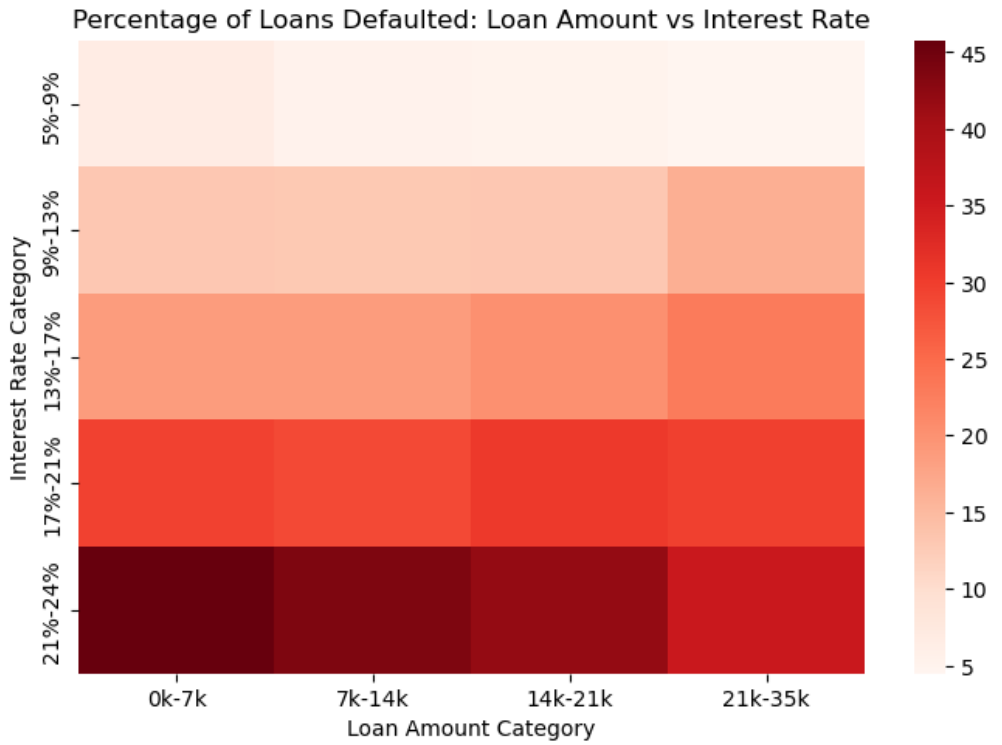
Conclusions drawn

1. Loan borrowers with high credit line utilization (>40%) are defaulting loans more.
2. We can clearly see the trend that charged-off loan percentage is increasing as the credit revolving line utilization percent is increasing.

Bivariate Analysis using Heatmaps

Loan amount category vs other quantitative variables

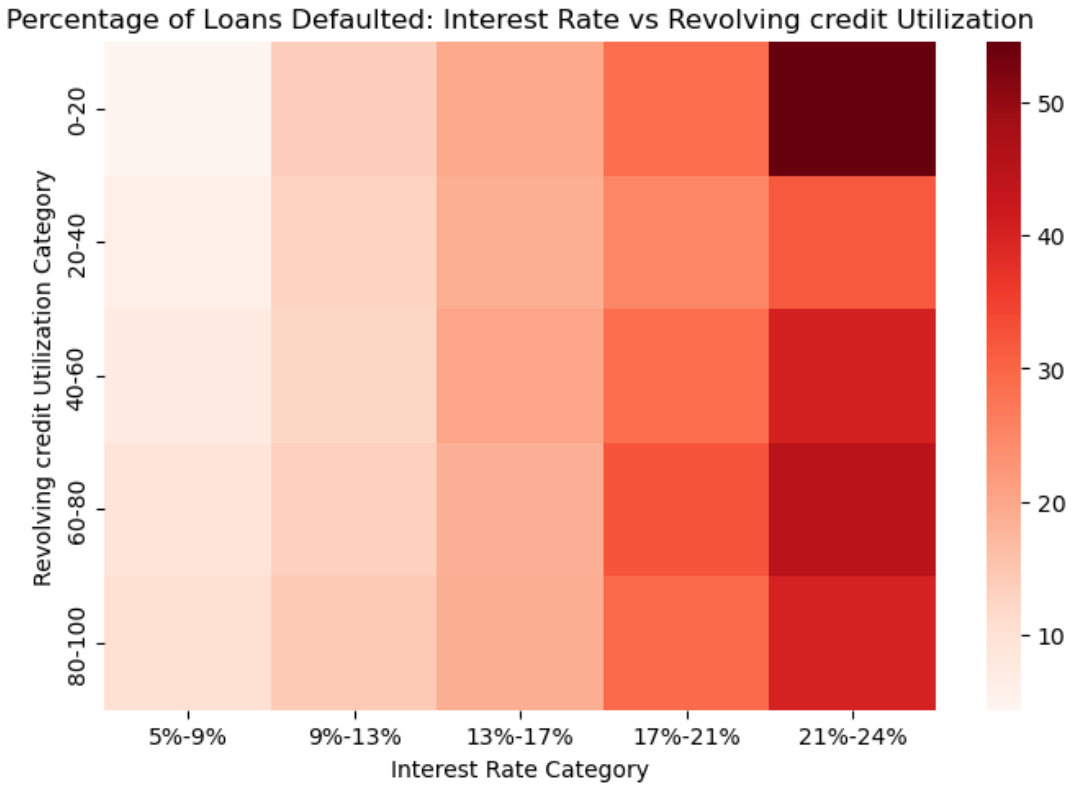


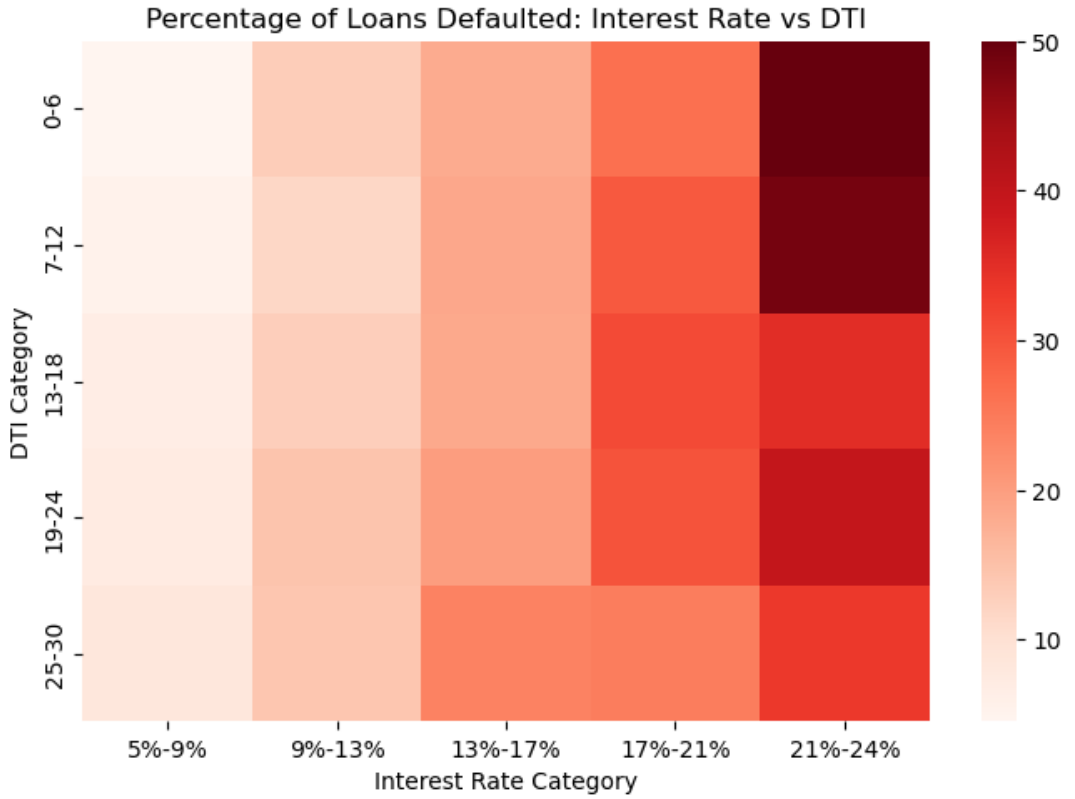


Conclusions drawn

1. More than 50% of borrowers with Loan amount between 21k & 35k and Active credit lines between 27 & 36 **defaulted**.
2. More than 45% of borrowers with Loan amount less than 7k and Interest rates between 21% & 24% **defaulted**.

Interest rate category vs other quantitative variables

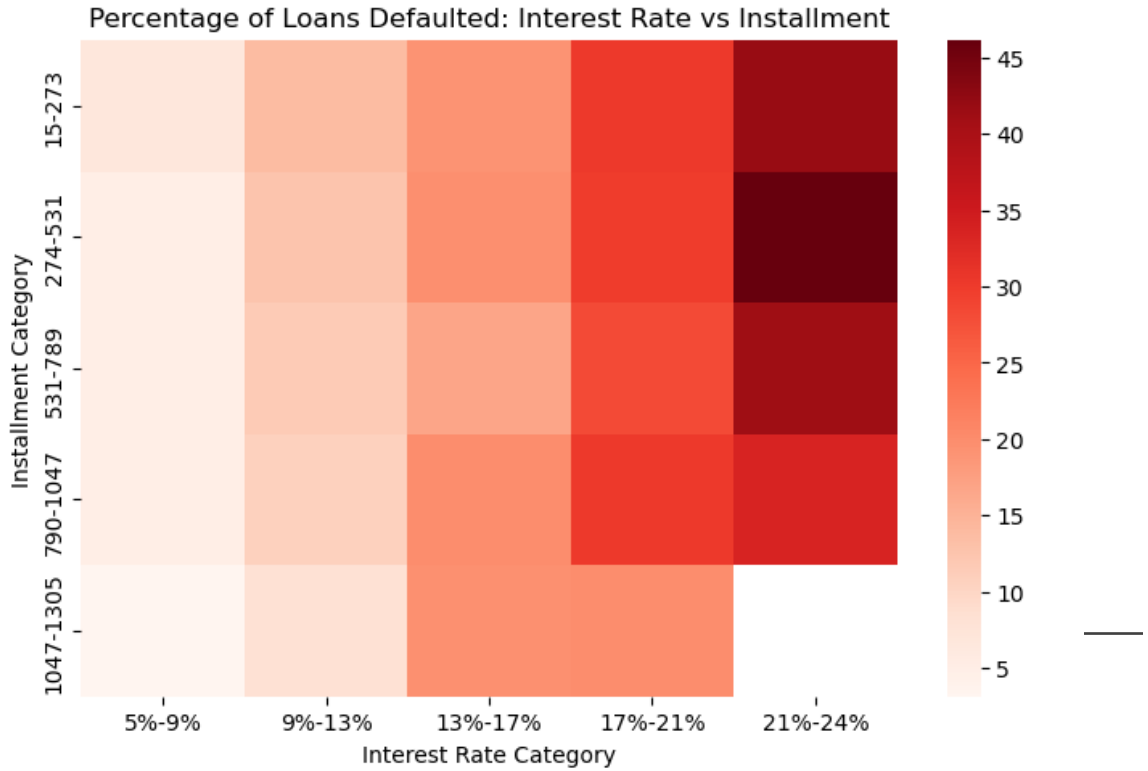


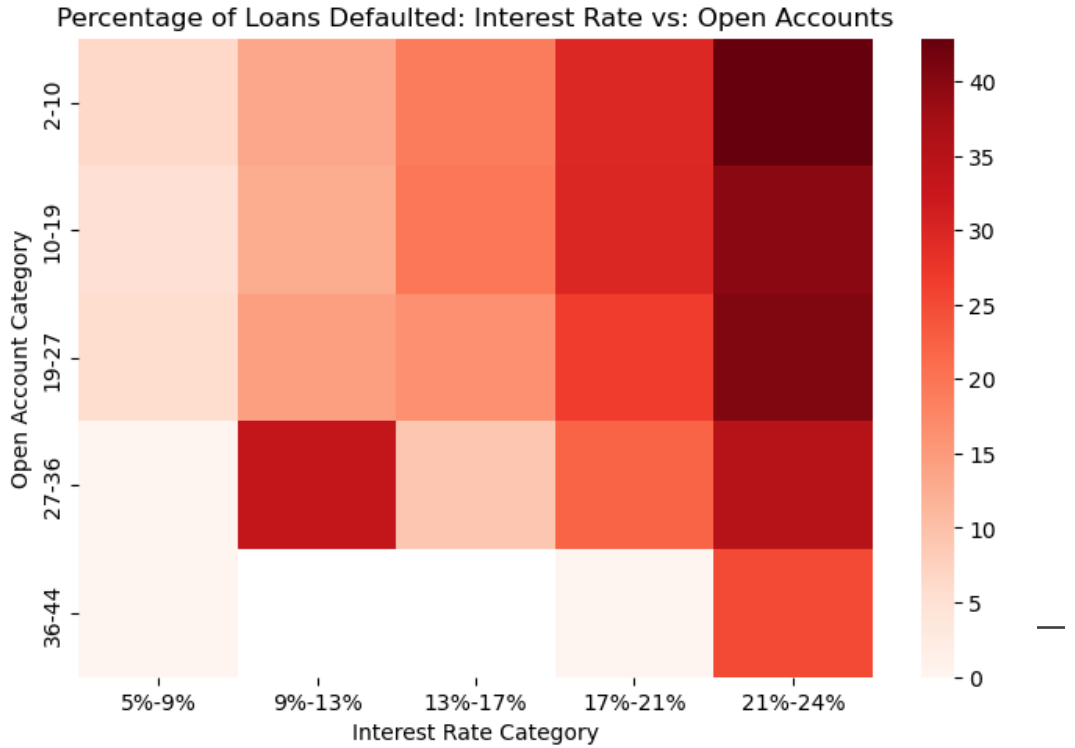


Conclusions drawn

1. More than 54% of borrowers with Interest rates between 21% & 24% and Revolving credit utilization between 0% & 20% **defaulted**.
2. More than 50% of borrowers with Interest rates between 21% & 24% and Debt to income ratio between 0% & 6% **defaulted**.

Interest rate category vs other quantitative variables

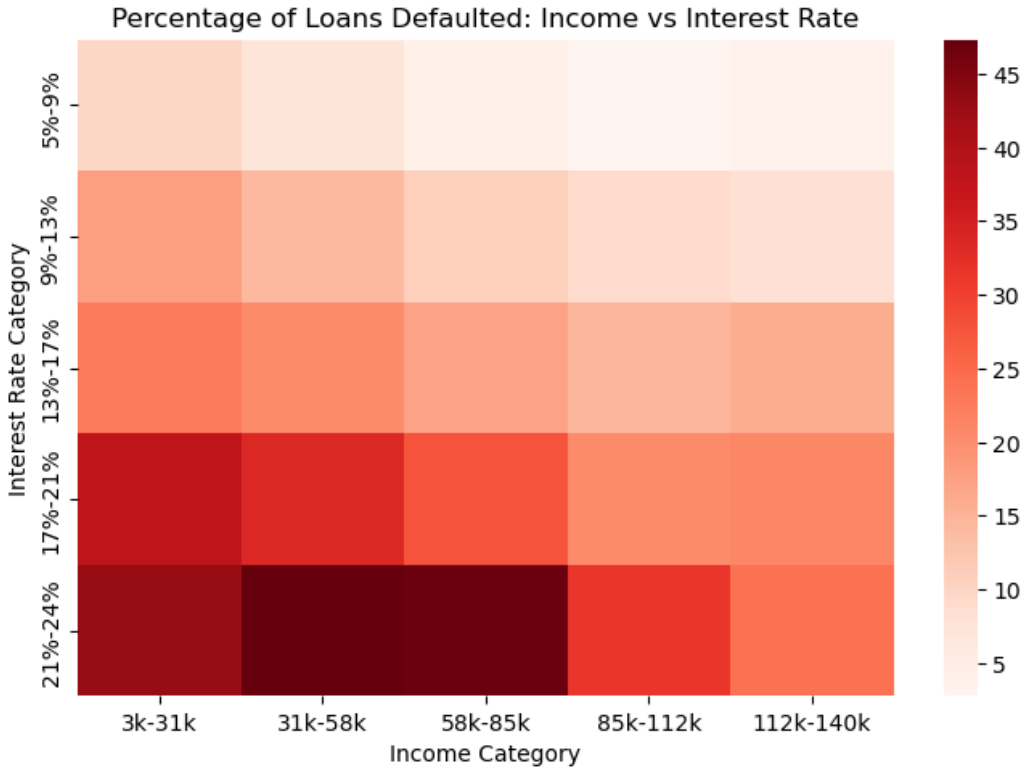


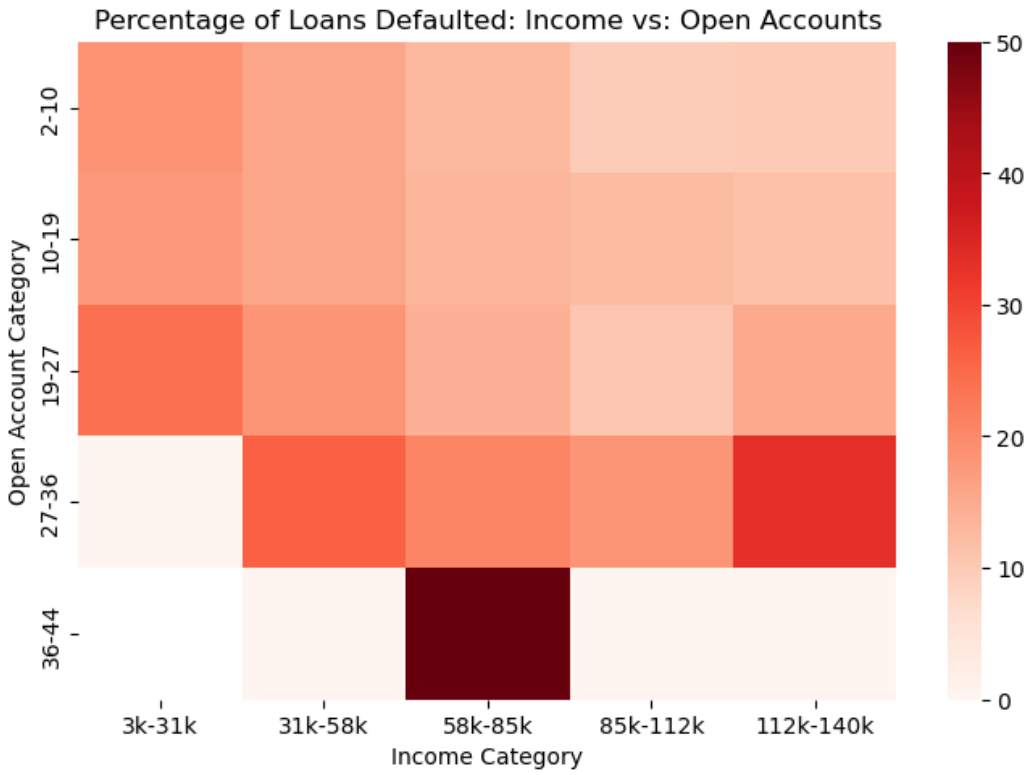


Conclusions drawn

1. More than 45% of borrowers with Interest rates between 21% & 24% and Installment amount between 274 & 531 **defaulted**.
2. More than 40% of borrowers with Interest rates between 21% & 24% and Active credit lines between 2 & 10 **defaulted**.

Annual income category vs other quantitative variables





Conclusions drawn

1. More than 50% of borrowers with Annual income between 58k and 85k and Active credit lines between 36 & 44 **defaulted**.
2. More than 47% of borrowers with Annual income between 31k and 85k and Interest rates between 21% & 24% **defaulted**.

Conclusion

1. Grade G loans were riskier followed by F, E, D, C, B, A.
2. More than 50% of F5 sub-grade type loans were defaulted, followed by G3 sub-grade type.
3. Percentage of defaulters were highest in the category of 'small_business' purpose.
4. Percentage of defaulters were high in those who opted for 60 months loan term than 36 months loan term.
5. Even though majority of the loans taken and defaulted are highest in CA state, NE state defaulter percentage is highest with ~60%.
6. Charged-off loan percentage is increasing as the ‘rate of interest’ and ‘revolving credit utilization’ is increasing.
7. More than 50% of loans with below criteria are charged-off:
   a. Loan amount between 21k & 35k and Active credit lines between 27 & 36.
   b. Interest rates between 21% & 24% and Revolving credit utilization between 0% & 20%.
   c. Interest rates between 21% & 24% and Debt to income ratio between 0% & 6%.
   d. Annual income between 58k and 85k and Active credit lines between 36 & 44.