# Lending Club Case Study

By Abhinav Swarup & Srikaran R

Git : https://github.com/abhinavswarupgit/Lending-Case-Study-Upgrad-

# Problem Statement

- Business Understanding
    - Lending Club is the  largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures
    - There are two types of risks associated.
        - Not approving the loan to credible customers results in business loss
        - Approving loan to defaulters also contributes to business loss

- Business Objectives
    - Provide recommendations to identify customers who are likely to default
    - Use the approved loan data to create patterns and support the recommendations

- Solution Strategy
    - Use Exploratory Data Analysis
    - Identify Patterns
    - Provide recommendations on top risk factors for loan approval process

# Solution Approach

## Data Preparation

- Source and load data obtained
- Clean data for Null/0 values
- Remove/Impute columns containing blank/irrelevant data

## Data Selection

- Create additional categorical variables
- Create a correlation matrix to identify positive and negative relations
- Identify a list of features based on the correlation and domain knowledge , for analysis

## Univariate Analysis

- Create box plot to analyze the distribution of selected variables
- Create distplot to find the frequencies of the selected features
- Create point plot for categorical variables
- Identify loan default percentages with the selected features

## Bivariate Analysis

- Use pairplots to identify correlation between the chosen features
- Use line plots between features to identify correlation

# Assumptions & Business Constraints

- **Loan Status** is the Target Variable for the current analysis, and we will be dropping the records containing values as 'Current' as we cannot infer whether the applicant will default or pay the entire loan amount back
- Features containing more than 50% of null values are dropped
- Features having same values for all records and unique values for all records are dropped
- Highly correlated features are dropped
- Features Employee length and Number of public record bankruptcies are imputed to their mode values
- Features reflecting customer behavior variables that are not available at the time of application of loan are not considered for analysis
- Features like interest rates and employee length are standardized for better analysis
- Critical numerical features are categorized for segmented analysis
- For the feature annual income, couple of outlier values are dropped as they did not have any impact on the analysis

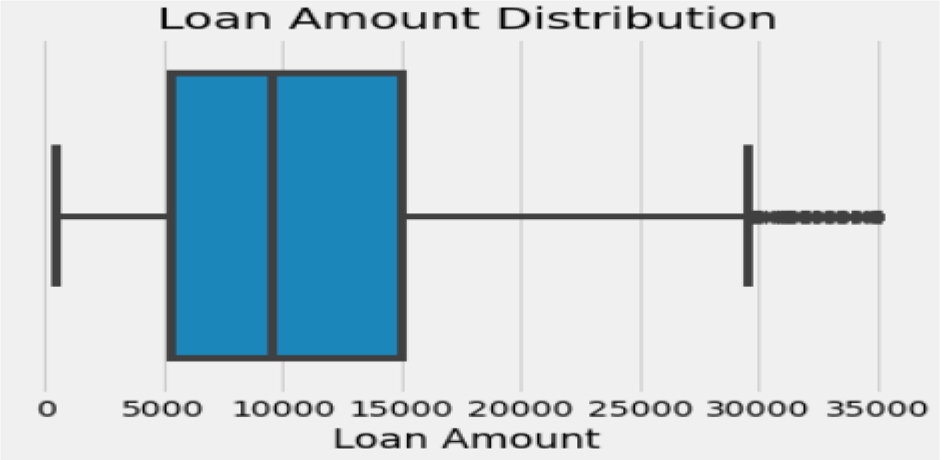
# Feature Selection

- Based on the correlation matrix and business knowledge, The following features were selected for analysis

  - Loan Amount
  - Interest Rate
  - Term
  - Annual Income
  - Loan Grade
  - Debt to Income Ratio
  - Loan Purpose
  - Revolving Utilization of Credit
  - Employee Length
  - Public Rec Bankruptcies

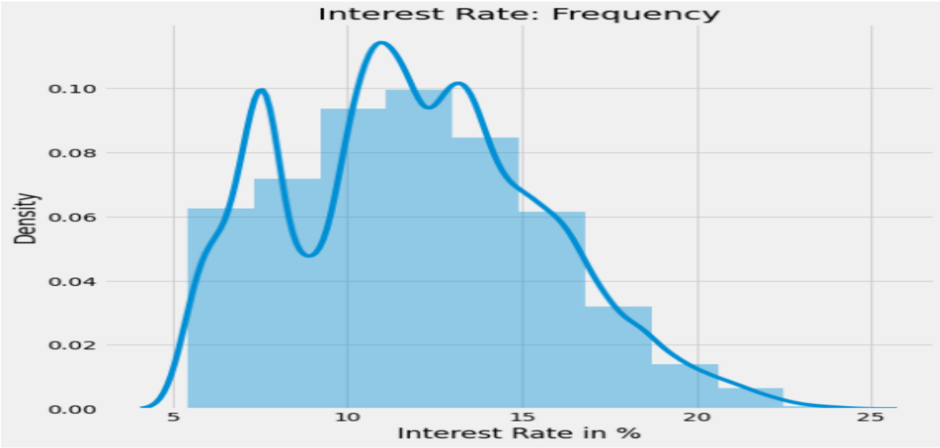
# Univariate Analysis

- Loan Amount



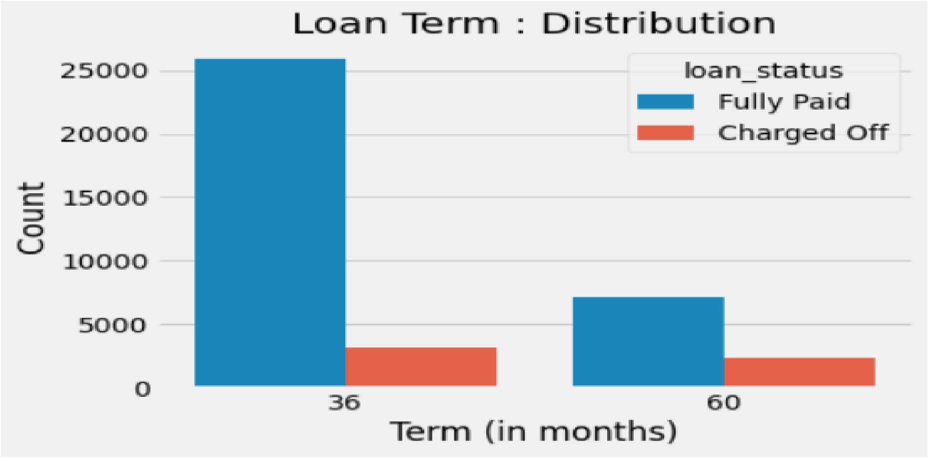

Most loans are given between $5000 and $15000

- Interest Rate



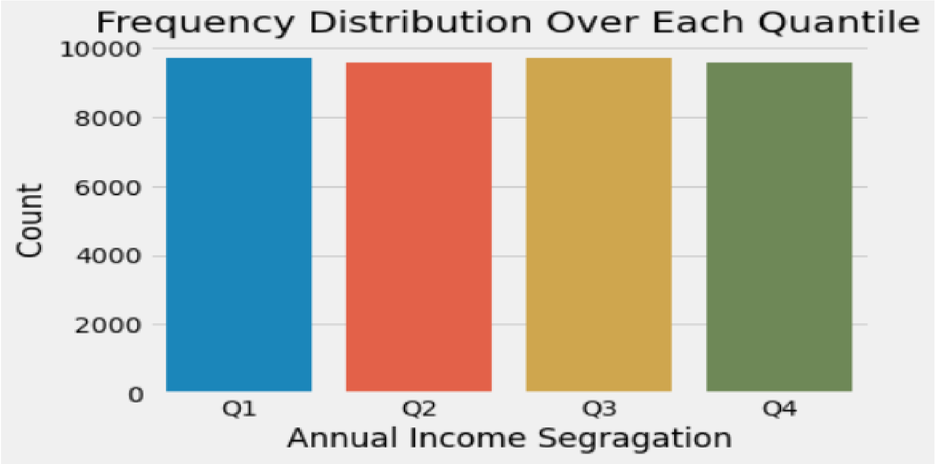

Most Loans are given on the interest rate of 9% to 14%

- Term




Most loans are taken for 36 months

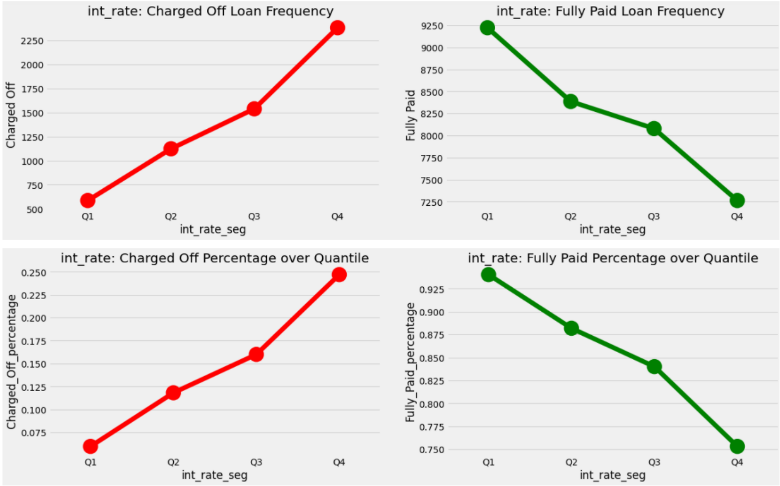
- Annual Income



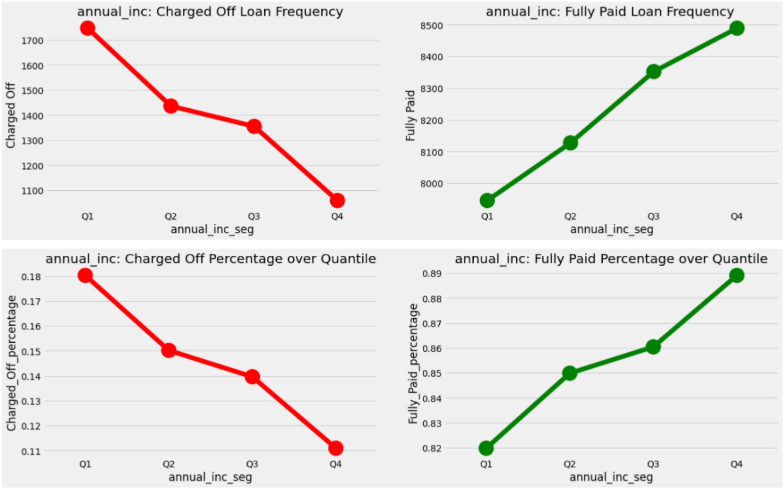

The income are equally distributed across the category

# Segmented Univariate Analysis

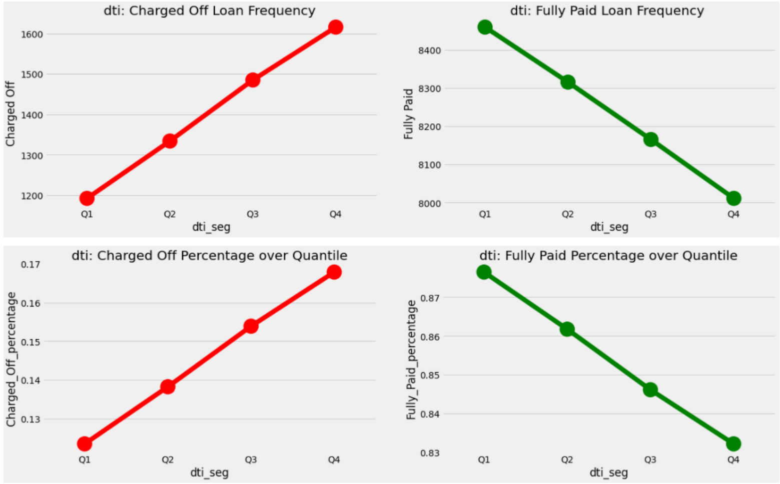
- Loan status is plotted against different categories of selected Features
- The below features with high values show a higher correlation to charged off loans except for Annual Income which has a negative correlations



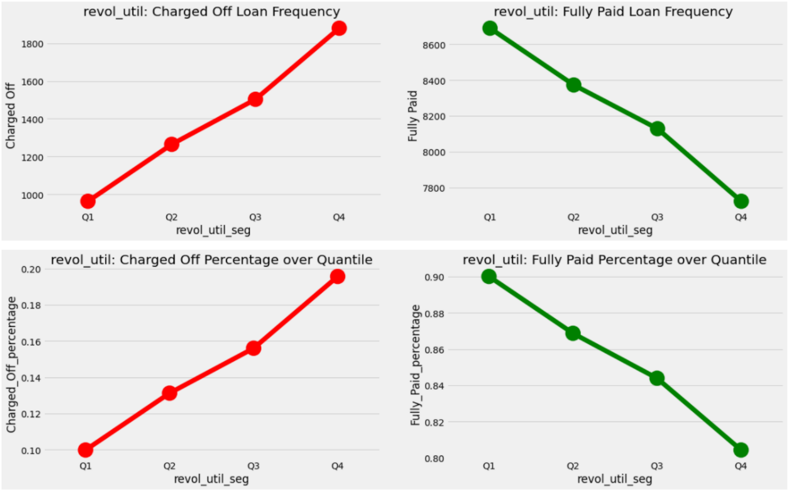

Interest Rate




Annual Income



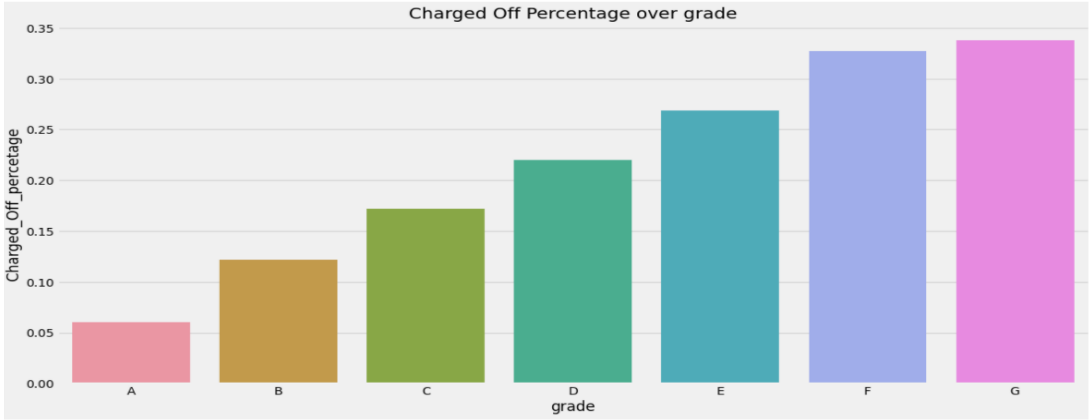

DTI



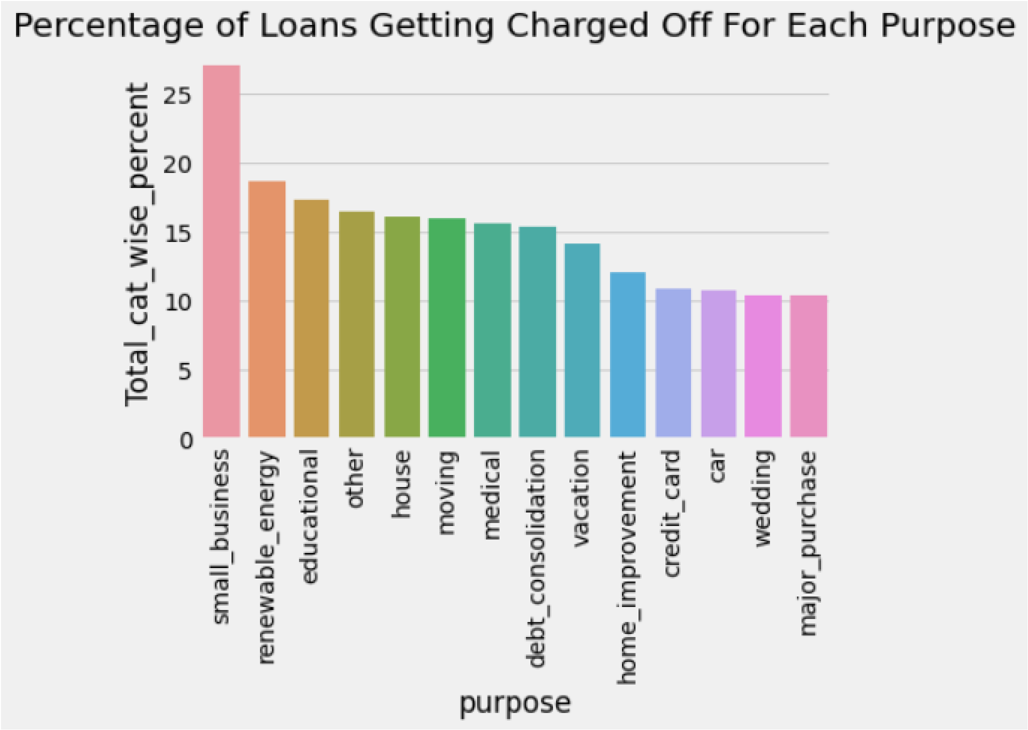

Revolving Utilization of Credit

# Segmented Univariate Analysis

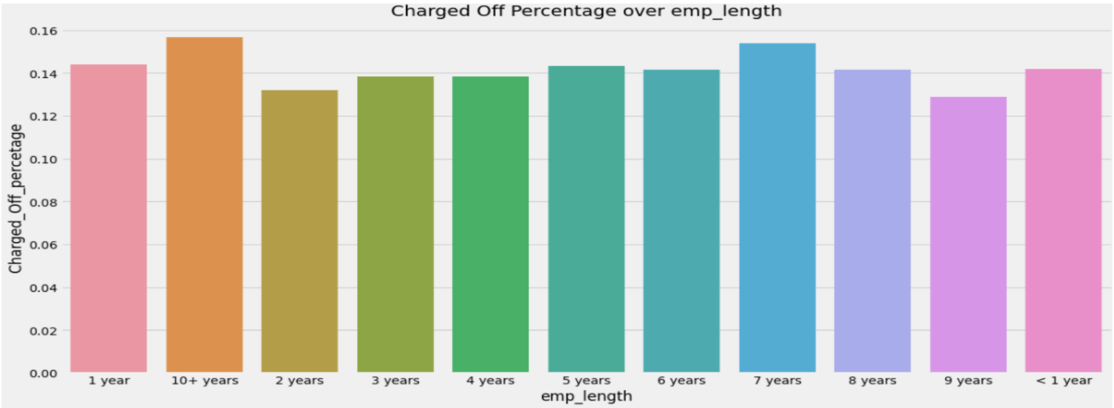
- Risk factor increases when we move from Loan Grade A to G
- There is no direct impact of number of years of job experience on loan charged off
- Small business is more riskier compared to other purpose of loans




Loan Grade



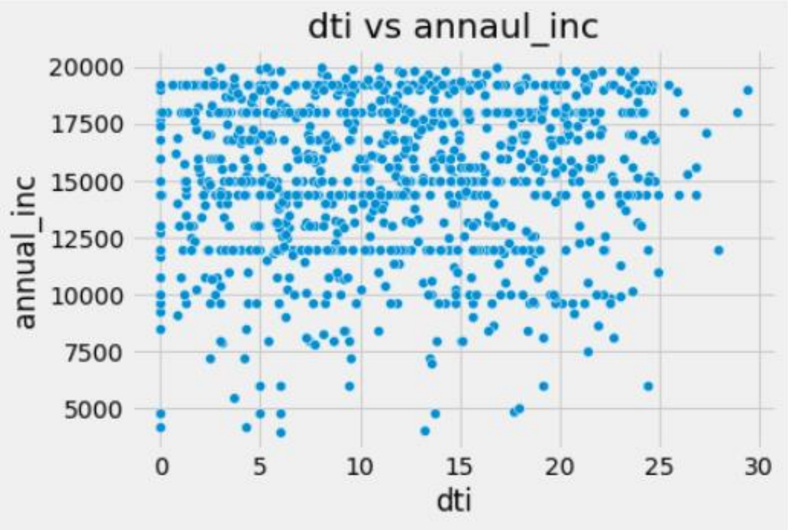

Purpose

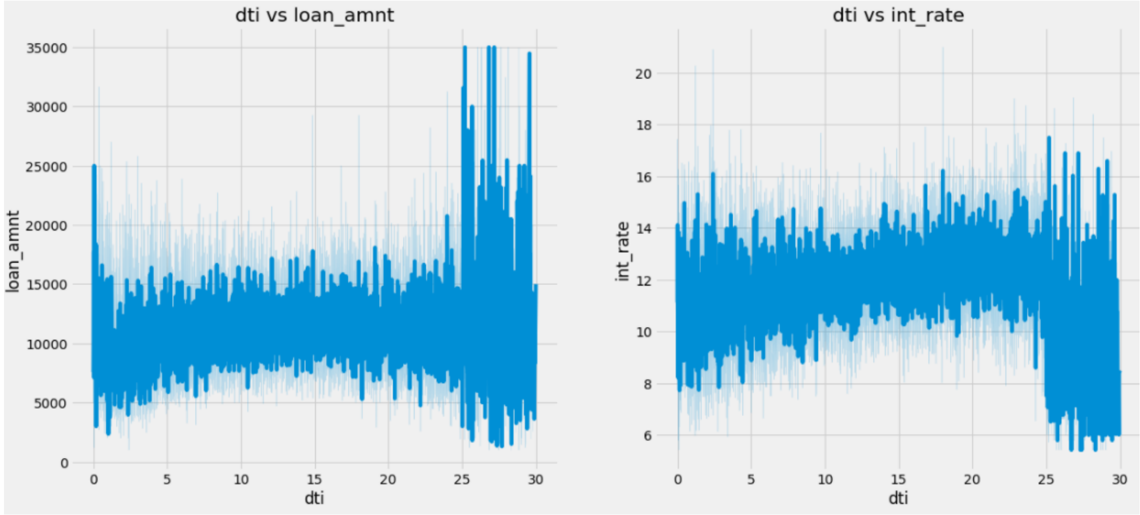


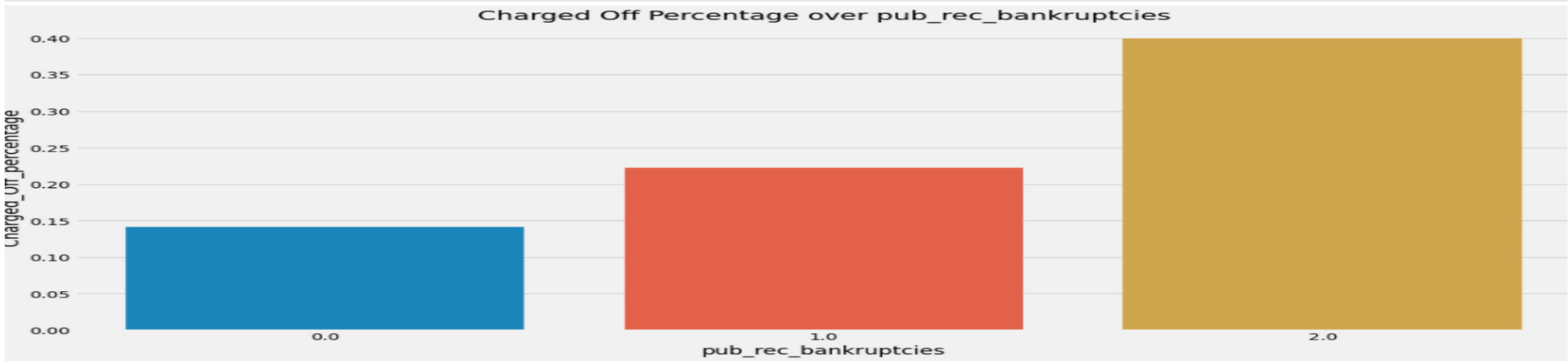

Employee Length

# Bivariate Analysis

- Scatter plot and Line plots are prepared against the selected features to identify correlation
- DTI vs Annual Income has less significance where else loan amount is spread with DTI and Interest rate are low too
- Applicants having public bankruptcies are defaulters

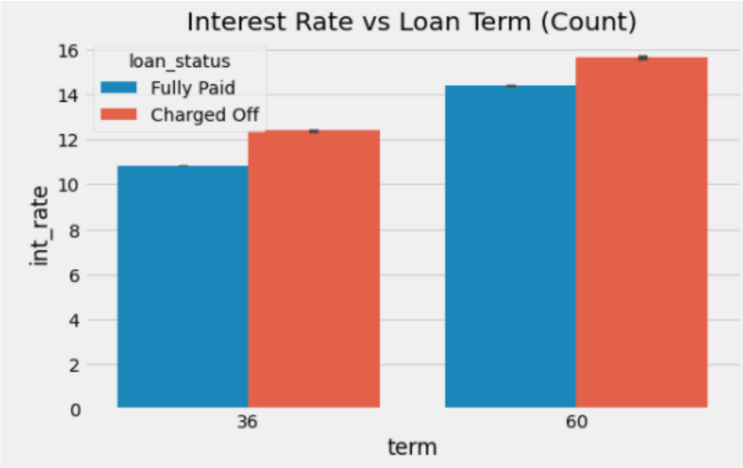


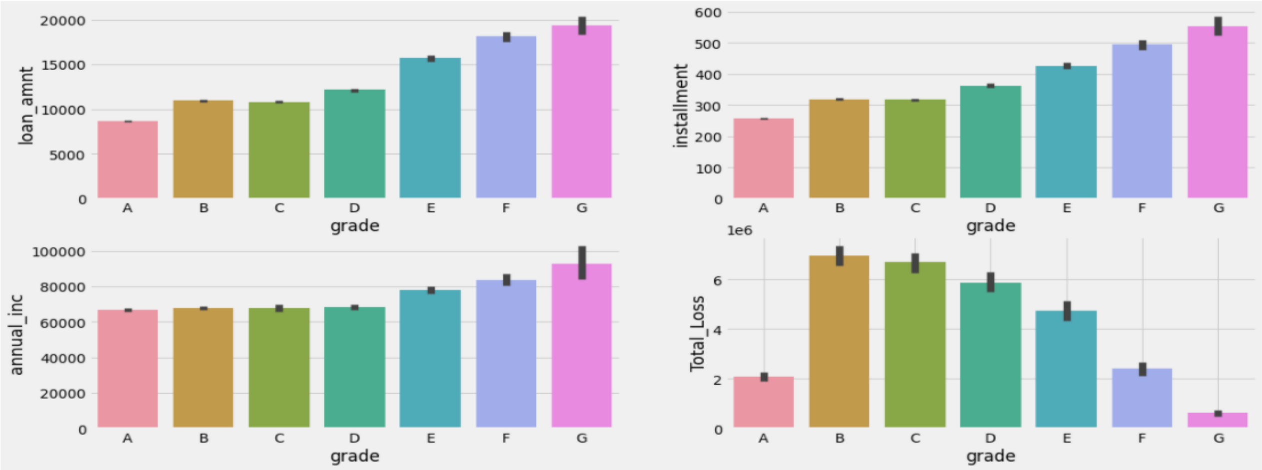



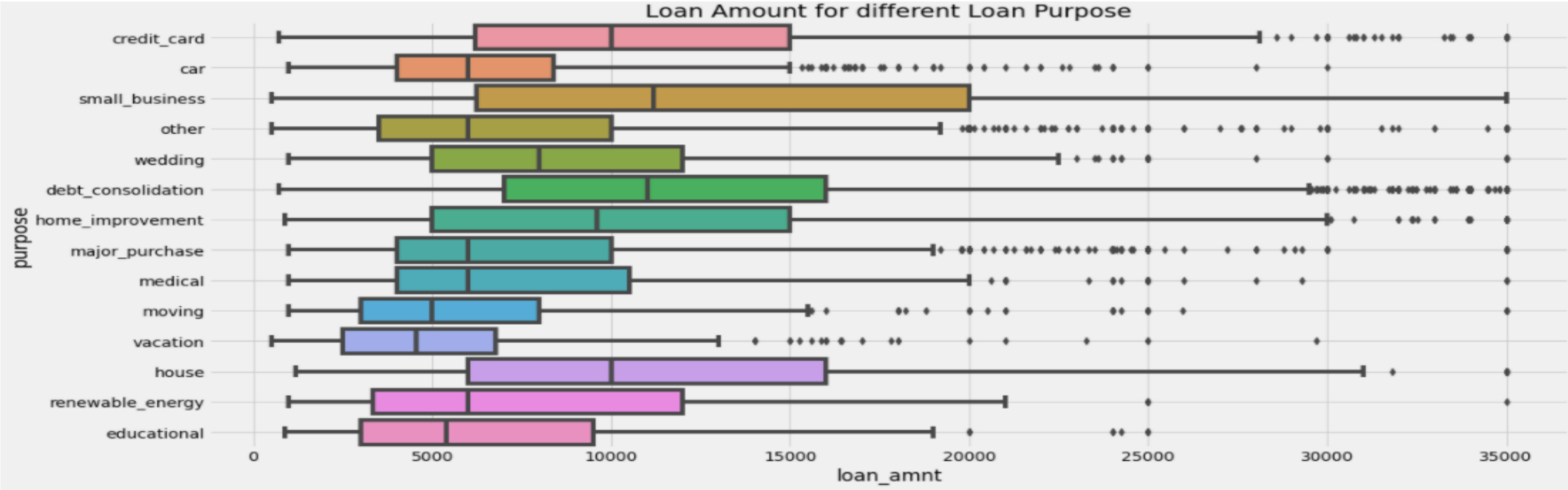



# Bivariate Analysis

- Grade G are given for high income applicants and for high loan amounts and higher installments
- 60 months loans are given at higher interest rate
- Higher amount of loan are taken for small business and debt consolidations



# Recommendations

- Based on the Exploratory Data Analysis performed, the following are the recommendations which are the driving factors in deciding the loan risk levels

  - Higher **Debt to Income** ratios corresponds to more defaulting
  - Small Business tends to be more risker **Purpose** compared to others
  - **Loan Grades** F & G are more riskier when combined with higher interest rates
  - Applicants with **Public Record of Bankruptcies** are sure to default
  - Applicants with higher **Annual Income** are less riskier compared to lower income group

  All the above features have to be looked together in determining the risky applicant and decision should not be taken based on one single feature mentioned