Lending Club Case Study

Rasteh Nili

Introduction

- Lending club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.
- Lending loans to ‘risky’ applicants is the largest source of financial loss of the company .
  - Not approving the loan for applicants that are likely to pay will results in a loss of business for the company
  - In the other hand if the applicant is not likely to repay the loan, approving the loan will lead to a financial loss for the company
- If one is able to identify risky loan applicants, then credit loss of the company will be reduced

Business Objective

- The company wants to understand the **driving factors** behind loan default
  - Variables that are strong indicators of default.
- The company can utilise this knowledge for its portfolio and risk assessment
- The objective of this analysis is to use EDA to:
  1. Understand driving factors behind loan default
  2. Identify risky applicants “defaulted” using risk analytics on collected data from past applicants.

Methodology

Understanding Data

- Understand Variable types
- Using metadata to understand the terminology and meaning behind each variable
- Identify numerical variables
- Identify Categorical Variables

Data Cleaning

- Removing:
  - Variables with >90% NA
  - Low variant Variables with one unique value
  - Variables with similar information
  - String column with almost all values unique
  - Variables that are not considered risk factor
  - Duplicate rows
  - Rows with undesirable outcome
- Imputation for Missing values
- Clean spaces and unnecessary characters
- Convert date columns
- Derived Variables
  - Aggregate or expand columns

Univariate Analysis

- Distribution analysis for:
  1. Continuous variables using:
     - Box plot
     - Histogram
  2. Categorical Variable:
     - Bar plot
     - Pie chart

Segmented Univariate Analysis

1. Categorical Variables
   - Bar plot
2. Continuous Variables using:
   - Box plot
   - Violin plot
   - Statistical tests of significance

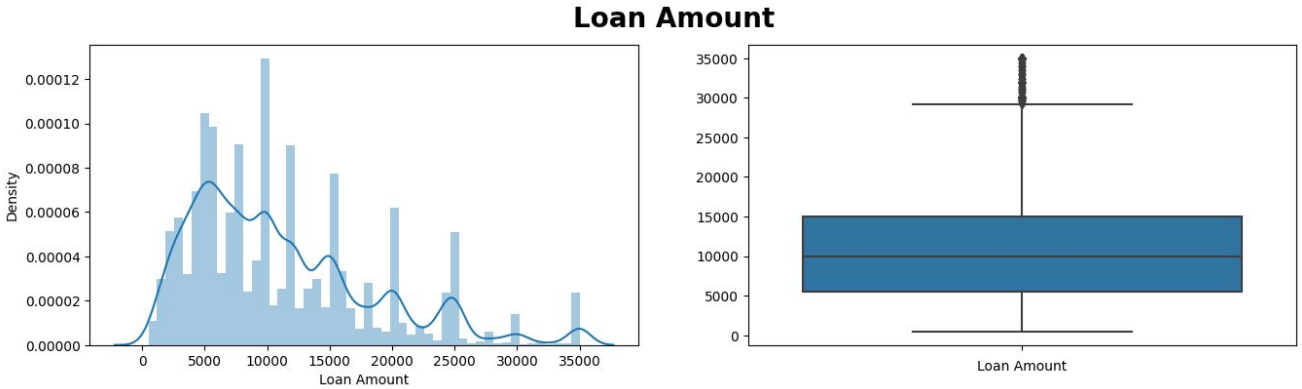
Bivariate Analysis

1. Categorical variables vs. Categorical Variables
   - Bar Plot
2. Categorical Variables vs. Continuous variables:
   - Box plot
   - Bar plot
2. Categorical Variable:
   - Bar plot
   - Pie chart
3. Continuous variable vs. Continuous Variable:
   - Correlation Heatmap
   - Pair Plot

Recommendation

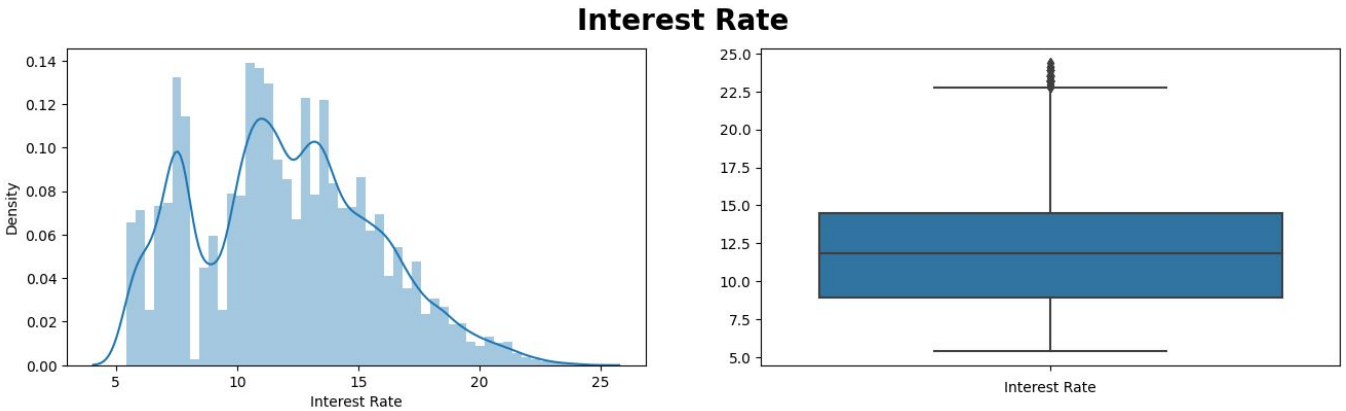
1. Identify at least the 5 important driver variables (i.e. variables which are strong indicators of default).
2. Drive Metrics:
   a. Type driven Metrics
   b. Business driven Metrics
   c. Data driven Metrics

Univariate Continuous variable

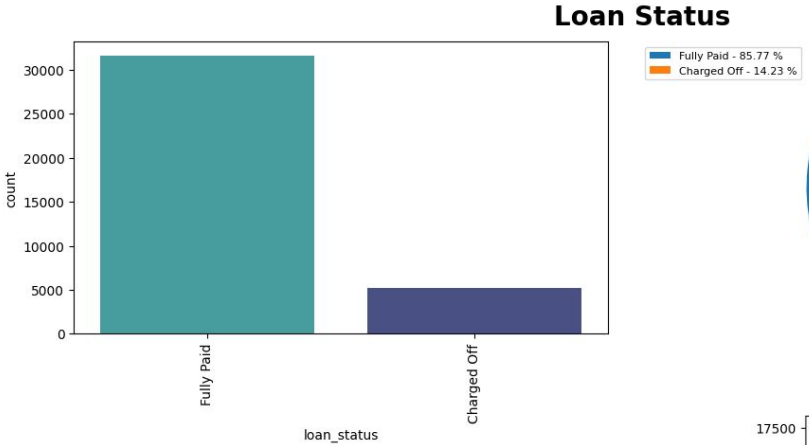


- Most of the loans are around $,5000
- There is a spike in $10,000 Loan indicating that this is a popular loan

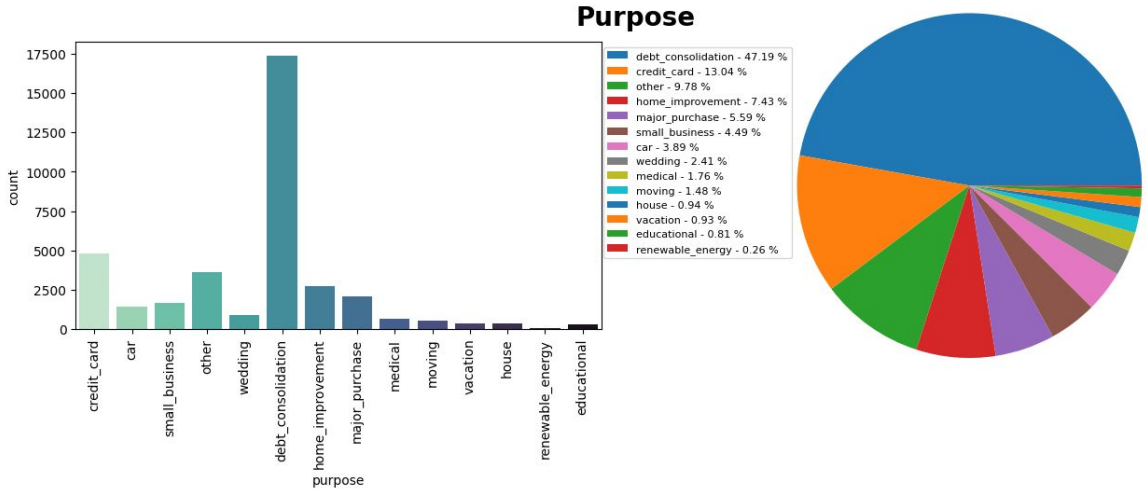
- Most Interest rates are between 10% to 15%



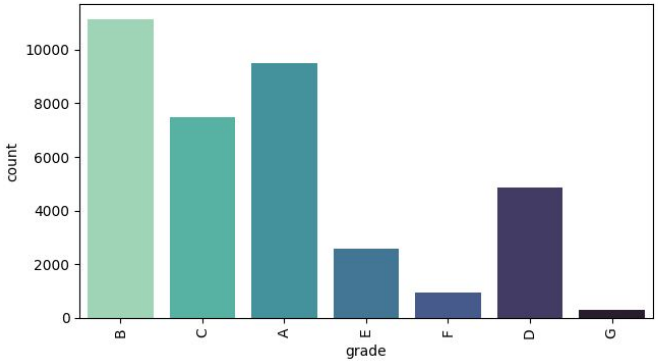
Univariate Categorical

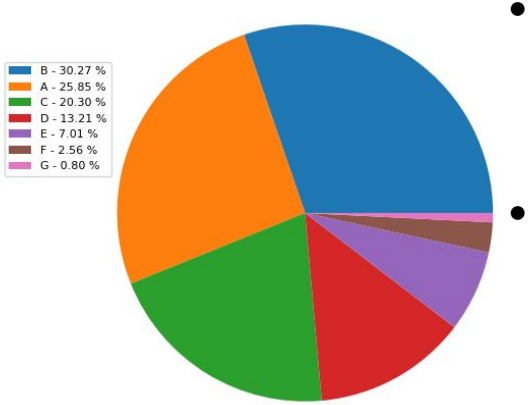


- Number of fully paid loans are 6 times of defaulted ones. So most of the applicants repay their depths.
- about 50% of these loans are for debt loans are for debt consolidation.

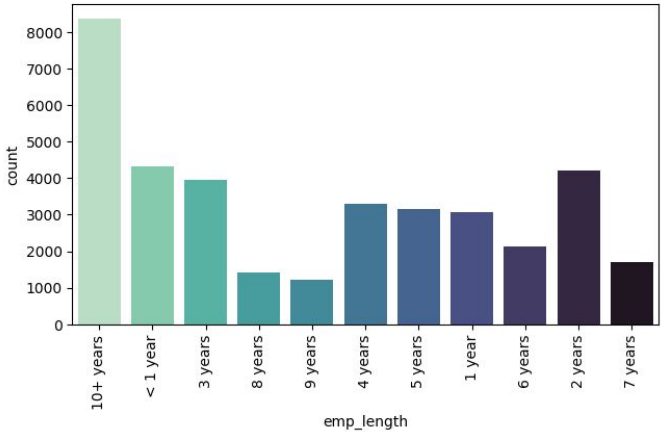


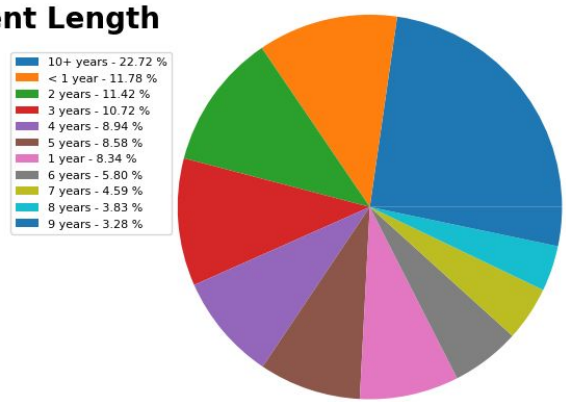
Univariate Categorical



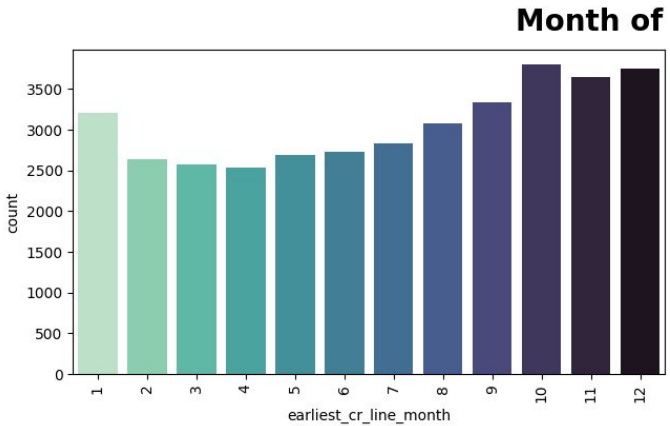


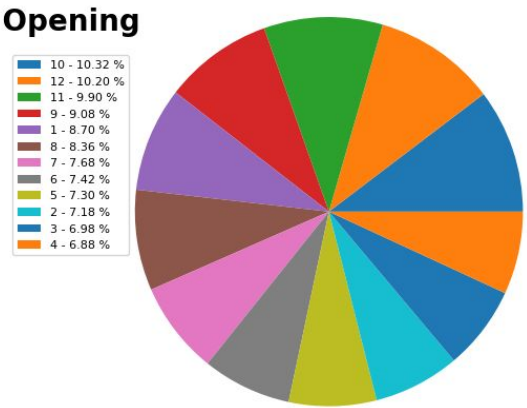
- About 80% of the loans are in first three grades of (A,B, C) and bellow 1% of applicants are in grade G of the loans
- About 50% applicants were employed for either more than 10 years or less than 2 year.



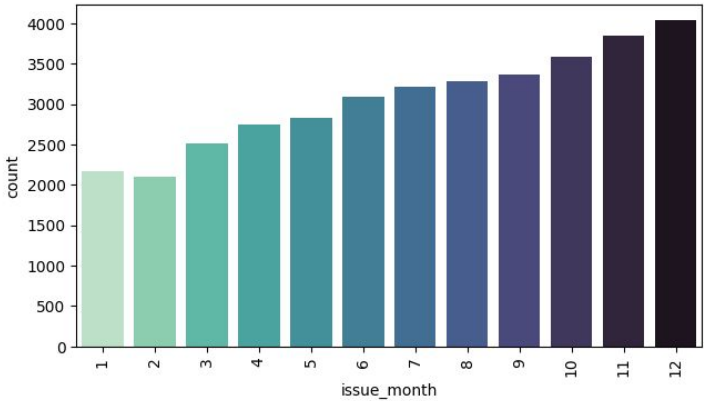


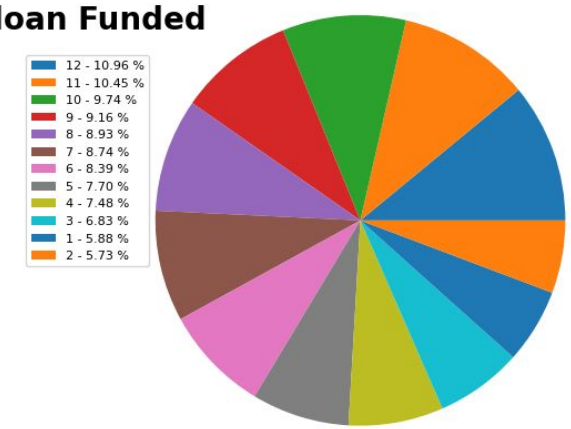
Univariate Categorical



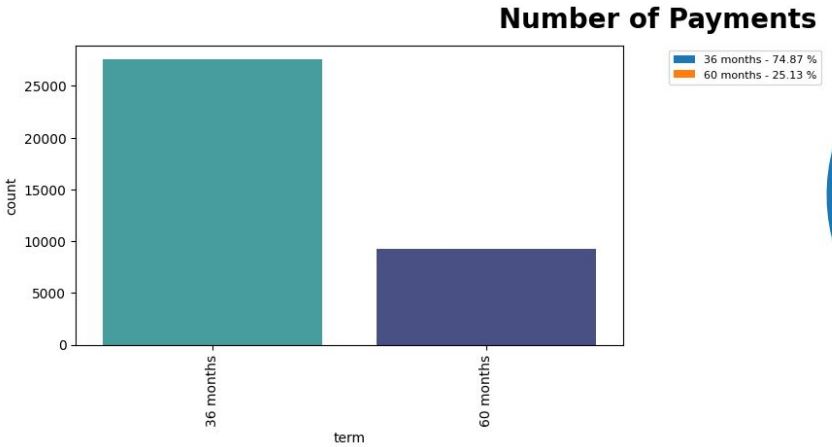


- More credit lines are opened in last three months of the year and in January. About 40% of credit lines are opened in those 4 months.
- Most loans are being founded in last three months of the year

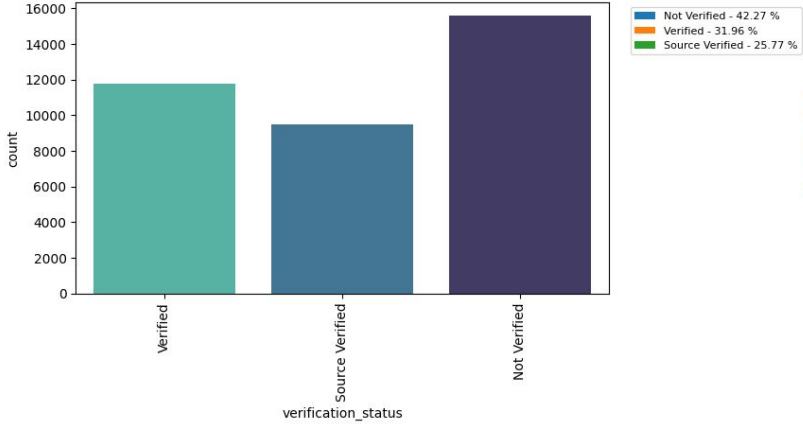




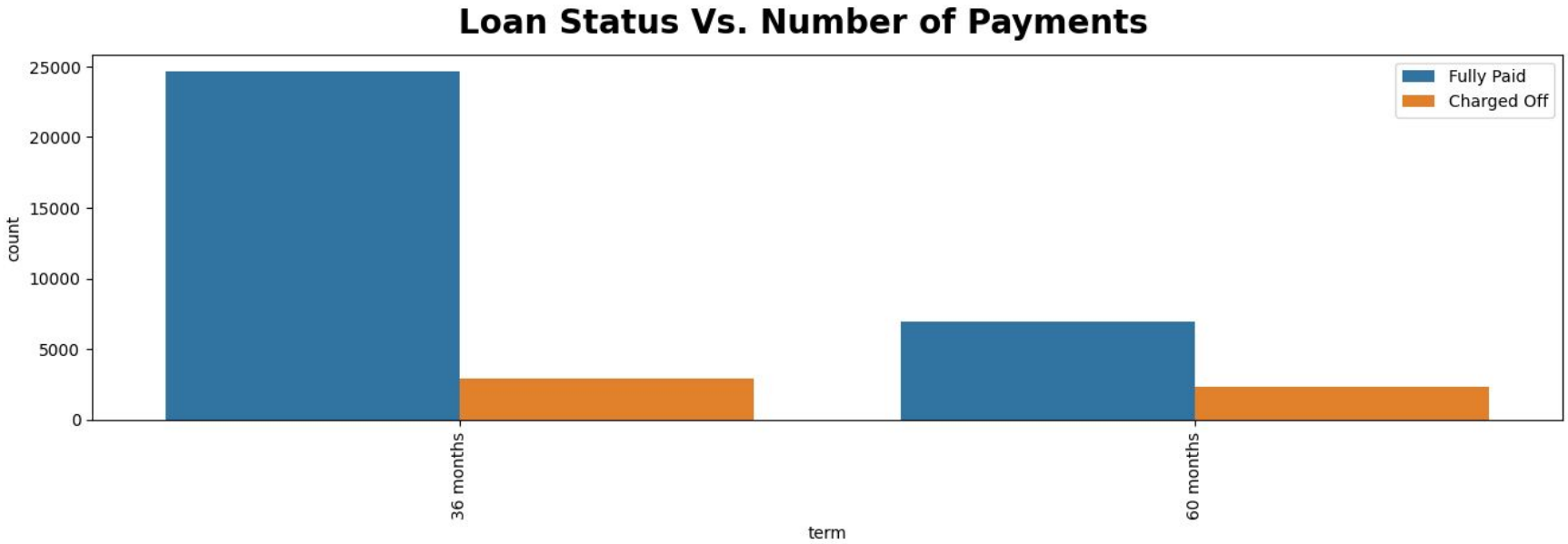
Univariate Categorical



- 75% of applicants have 36 months payment plan and only 25% have 60 months plan.
- The income of 47% of applicants who got the loan was not verified by LC.

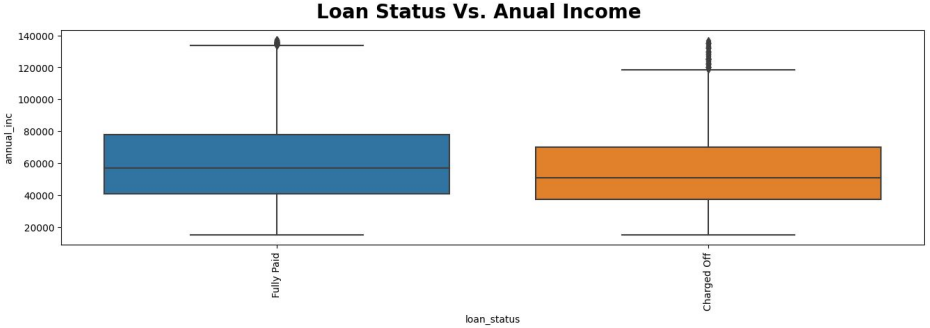


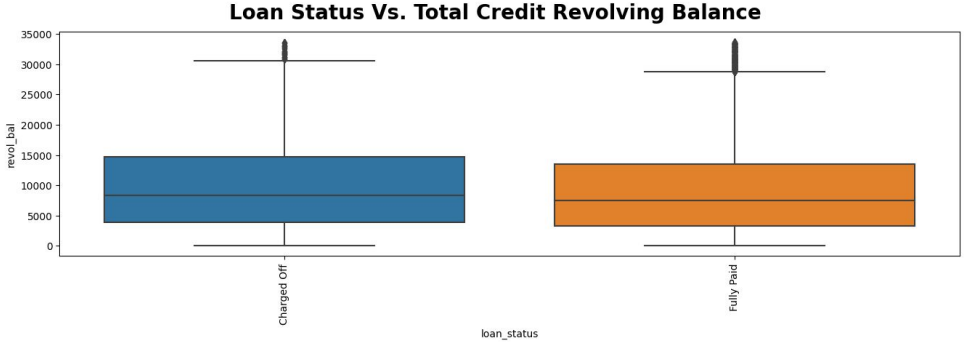
Segmented Univariate two Categorical Variable



While the number of charged off applicants in both 36 and 60 months term are very close, The number of applicants that paid their debts are 5 times greater in 36 months plan. Indicating that this 36 month plan is less risky. So 60 months plan is a risk factor, and applicants who apply for that plan are more likely to not pay their debt.

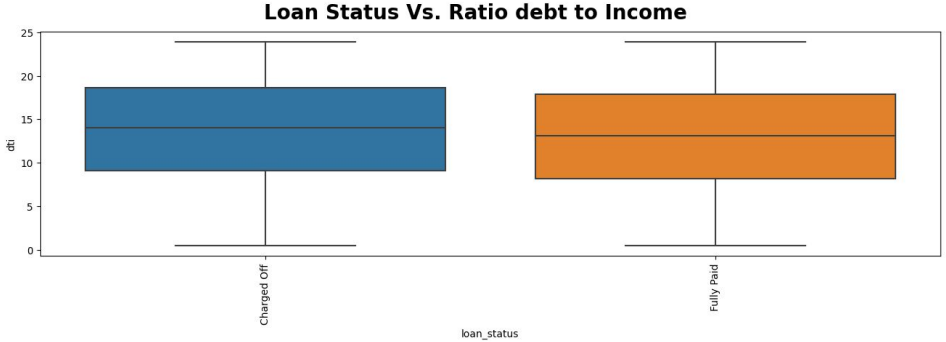
Segmented Univariate two Continuous Variable



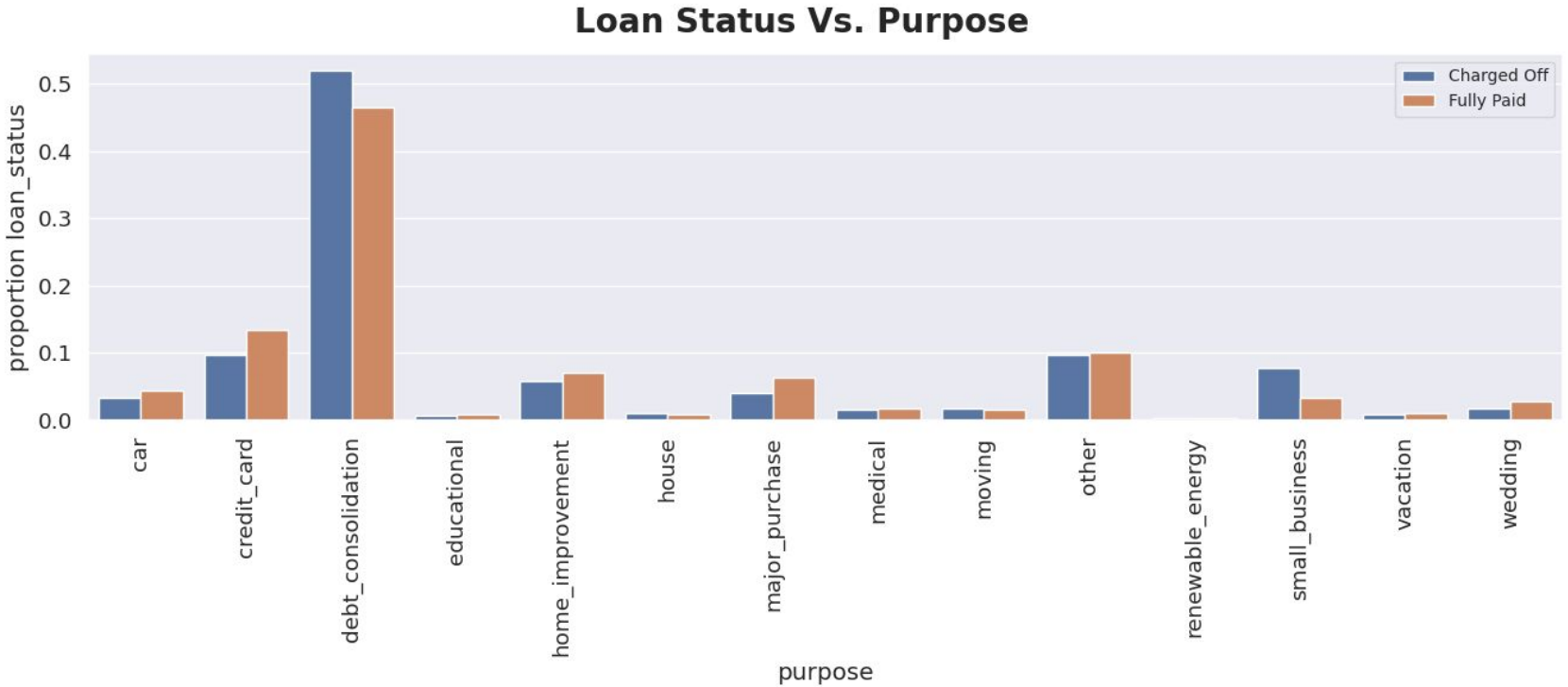


The Smirnova test have p-value of $1.099e^{-11}$ is showing a significant differences in Ratio debt to income for defaulted and fully paid applicants. So applicants with higher ratio of debt to income are more likely to defaulted.

These applicants also have significantly lower annual income and higher total credit revolving balance. So all these three factors or combination of them can be considered as risk factor.

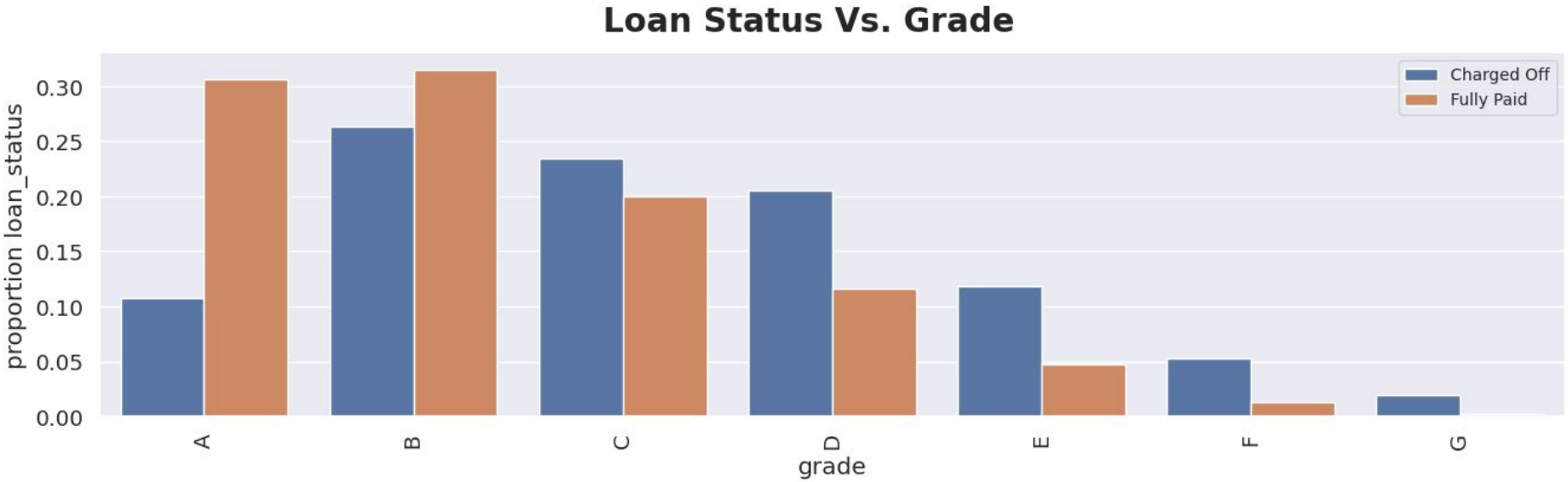


Bivariate Continuous Vs. Categorical Variables



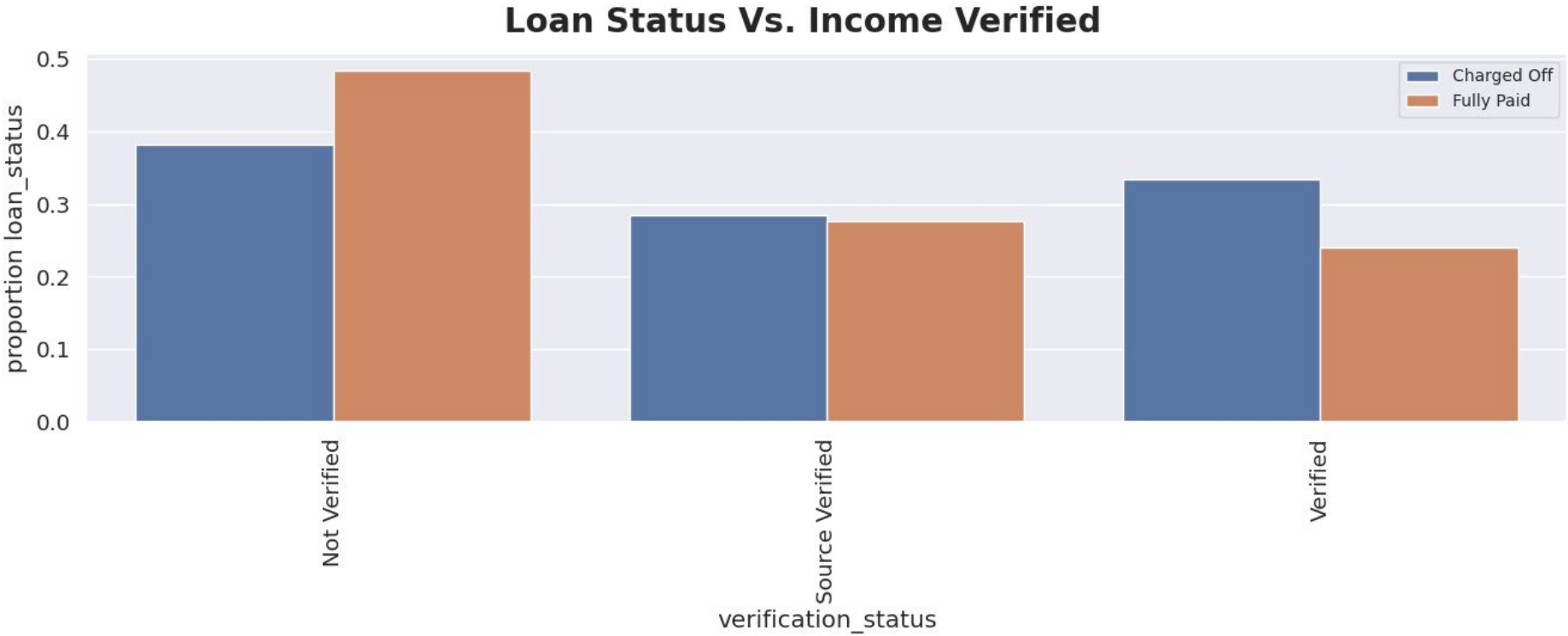
Applicants that take the loan for debt consolidation or small businesses are more likely to be defaulted. So loans for debt_consolidation and small businesses can be consider risky loans.

Bivariate Continuous Vs. Categorical Variables



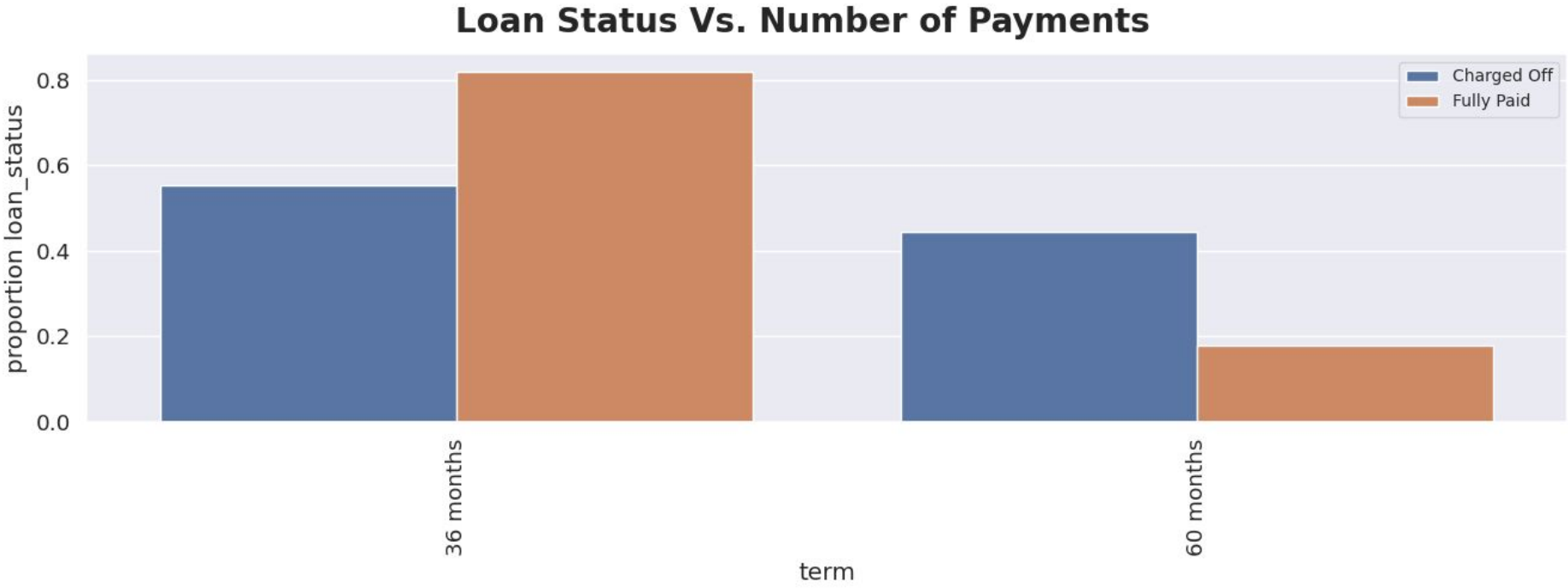
Since most applicants in grade D, E, F and G are defaulted we can consider these grades as high risk grades.

Bivariate Continuous Vs. Categorical Variables



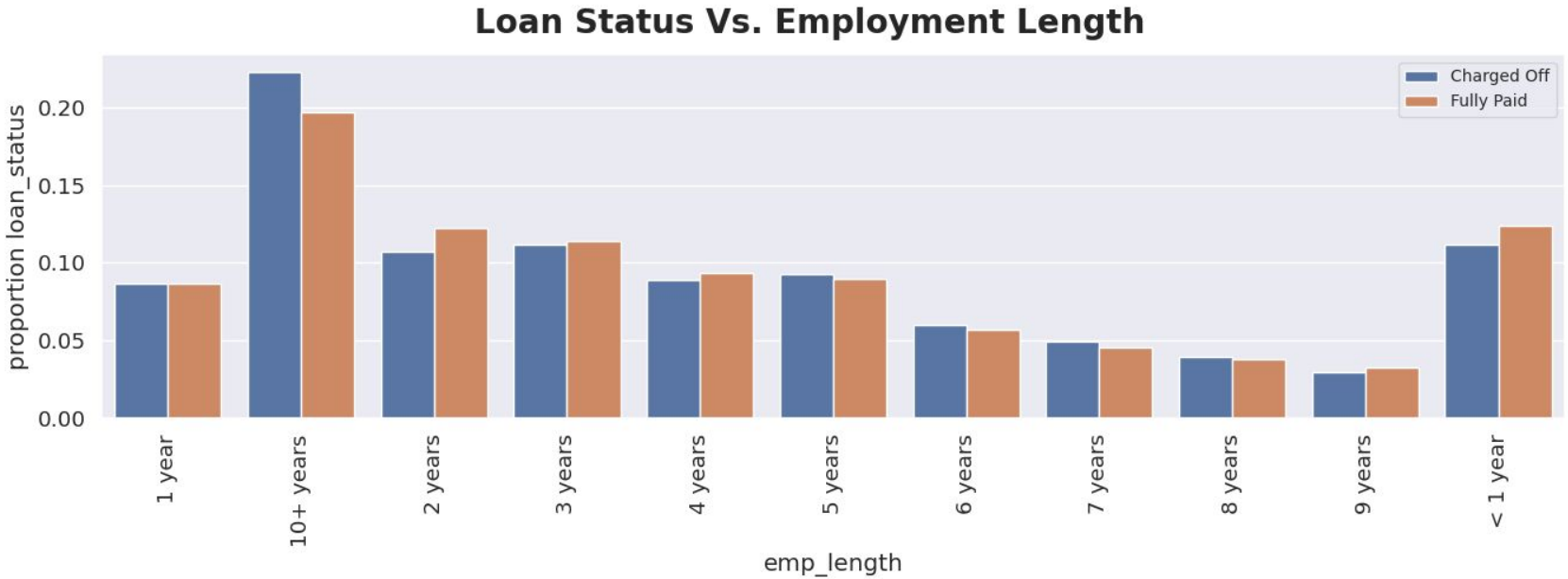
Applicants that their income been verified have more chance of being defaulted than those who didn't.

Bivariate Continuous Vs. Categorical Variables



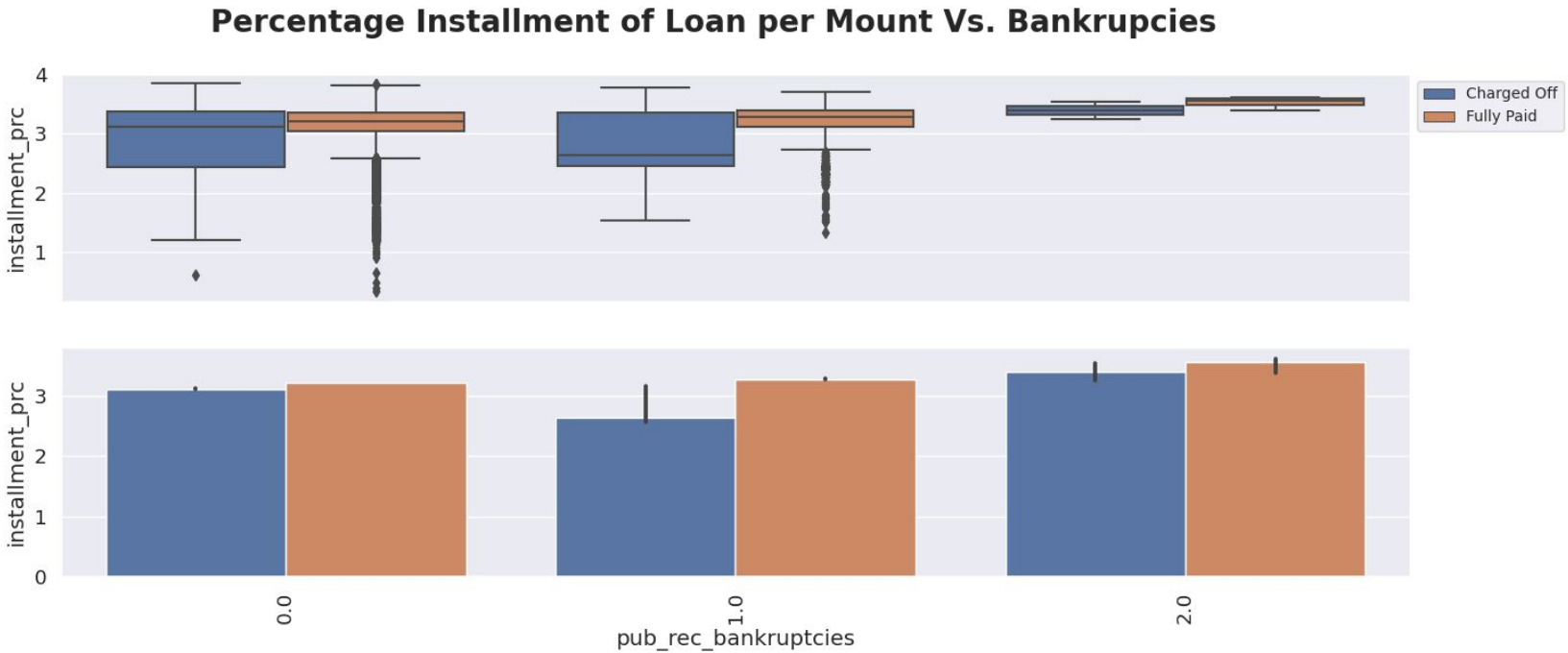
Applicant for 60 months plan are more likely to default.

Bivariate Continuous Vs. Categorical Variables



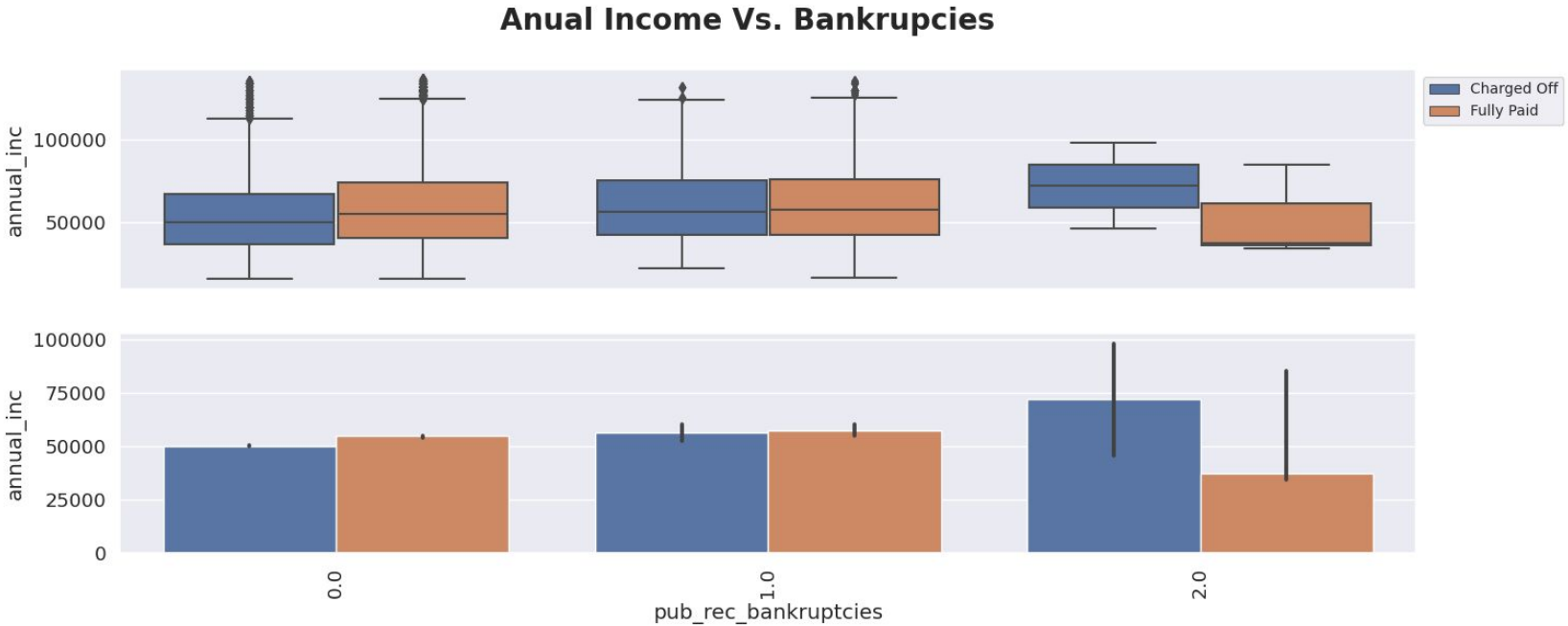
Applicants with 10+ years of employment have higher charged off proportion than fully paying. So those applicant are at higher risk of being defaulted.

Bivariate categorical vs continuous



The box plot for percentage installment tends to be lower for defaulted applicants with 2 recorded bankruptcies.

Bivariate categorical vs continuous



Applicants with 2 bankruptcy report and higher annual income are more likely to be defrauded. Based on the Box plot for Annual income vs. Bankruptcies the 25 percentile annual income for people with 2 bankruptcies is higher than 75 percentile annual income of applicants that fully paid off their loan.

Recommendations

1. Reduce number of loans for risky purposes like small businesses
2. Reduce loans in high risk graded D, E, F and G
3. Reduce number of long term (60 months) loans
4. Reduce giving loan to applicants with high credit revolving balance, low income and high debt income ratio