Lending club case study

- By:
- Siddhartha Khatri
- Gangadhar Reddy

Problem Statement

Company and investors are not able to identify the applicants profile properly before loan approval and hence they are facing both business and financial loss.

Objective

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 (risky applicants) at a higher interest rate, etc.

Exploratory Data Analysis

Data Sourcing

- Campany already given the dataset contains the information about past loan applicants and whether they 'defaulted' or not.

- For this case study we need to focus only on provided dataset given by company.

Data Cleaning

01
Remove columns having only NAN Values

02
Remove columns having only zero values

03
Filter out variables which could not be used for defaulters

Data Cleaning – Variable Type

Variable related to applicants.

Variable related to loan characteristics.

Variable related to customer behavior. These variables generated after the loan is approved.

Selected variables

1. emp_length: Employment length could be directly relates to financial status of the borrower.

2. home_ownership: Home ownership could be directly relates to financial status of the borrower e.g. If borrower is already owning house or morgage or on rent.

3. annual_inc: Annual Income.

4. addr_state: Just kept it to analyse the state wise loan status.

5. dti: Debt to income ratio could play important role to identify defaulters.

6. loan_amnt: Loan Amount

7. funded_amnt: Funded Amount by Company

8. funded_amnt_inv: Funded Amount by Investors

9. term: Terms of loan

10. int_rate: Interest Rate
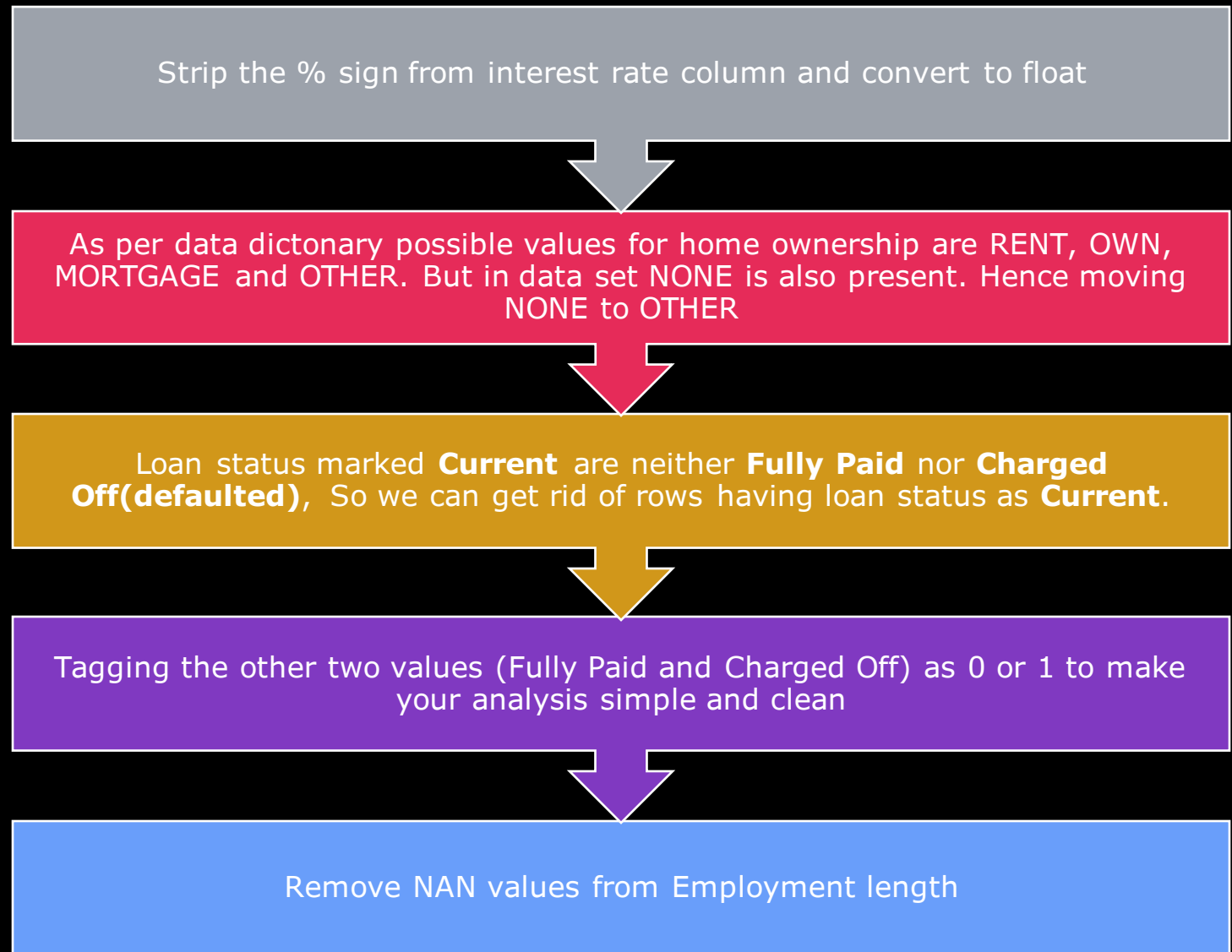
11. grade: Grades

12. sub_grade: Sub Grades

13. verification_status: Loan verification status

14. loan_status: Indicating defaulters. Possible values Charged Off (Defaulter), Fully Paid, Current

15. purpose: Purpose of loan

Data Cleaning

Strip the % sign from interest rate column and convert to float

As per data dictonary possible values for home ownership are RENT, OWN, MORTGAGE and OTHER. But in data set NONE is also present. Hence moving NONE to OTHER

Loan status marked **Current** are neither **Fully Paid** nor **Charged Off(defaulted)**, So we can get rid of rows having loan status as **Current**.

Tagging the other two values (Fully Paid and Charged Off) as 0 or 1 to make your analysis simple and clean

Remove NAN values from Employment length

Univariate Analysis

Unordered Categorical Variables

Home ownership: There are many people belong to mortgage category.

Verification Status: More than 16000 loans are not verified.

Loan Status: More than 5000 loans have Charged Off out of approximately 37000.

Purpose : People are taking loan mostly to do debt consolidation.

Address State: People from state with code CA, NY, FL, TX and NJ are the top 5 for having loans.

Ordered Categorical Variables

Employment Length: Most people taking loan having employment length more than 10 Years or less than 1 year.

Loan Terms: More than 25K loan are having terms of 36 months.

Grade: Most people avails loan at lower interest rates having grade A, B and C.

Sub Grade: Most people avails loan at lower interest rates having grade A1-A5, B1-B5 and C1-C5.

Quantitative Variables

Annual Income: Too many outliers present in the dataset. Filter out the 90th percentile from dataset. Mean annual income is around 58K.

DTI: Count of loan is very less when dti is greater than 25.

Loan Amount: Count of loan is maximum for 10K loan amount.

Funded Amount Investors: Count of loan is maximum for 5K and 10K amount invested by investors.

Interest Rate: People are intrested in loan for interest rate around 7 and 12.

Segmentation

Annual income of defaulters is relatively lesser than the annual income of fully paid borrowers. mean of annual income is relatively higher for fully paid.
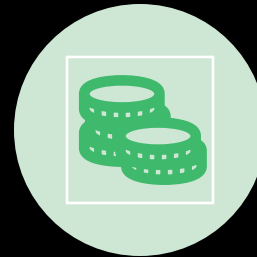
Debt to income ratio is relatively higher for defaulters. 75th and 50th quantile of dti is relatively higher for defaulters.

Loan amount are relatively higher for defaulters. 75th and 50th quantile of loan amount is relatively higher for defaulters.
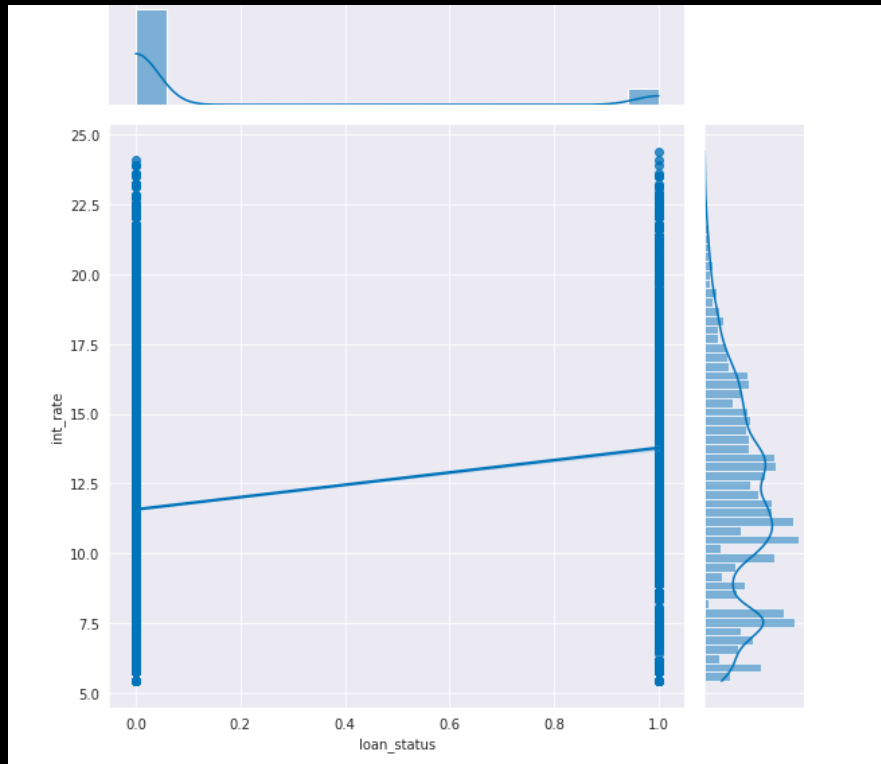
Amount invested by investors is relatively higher for defaulters. 75th and 50th quantile of funded amount by investors is relatively higher for defaulters.

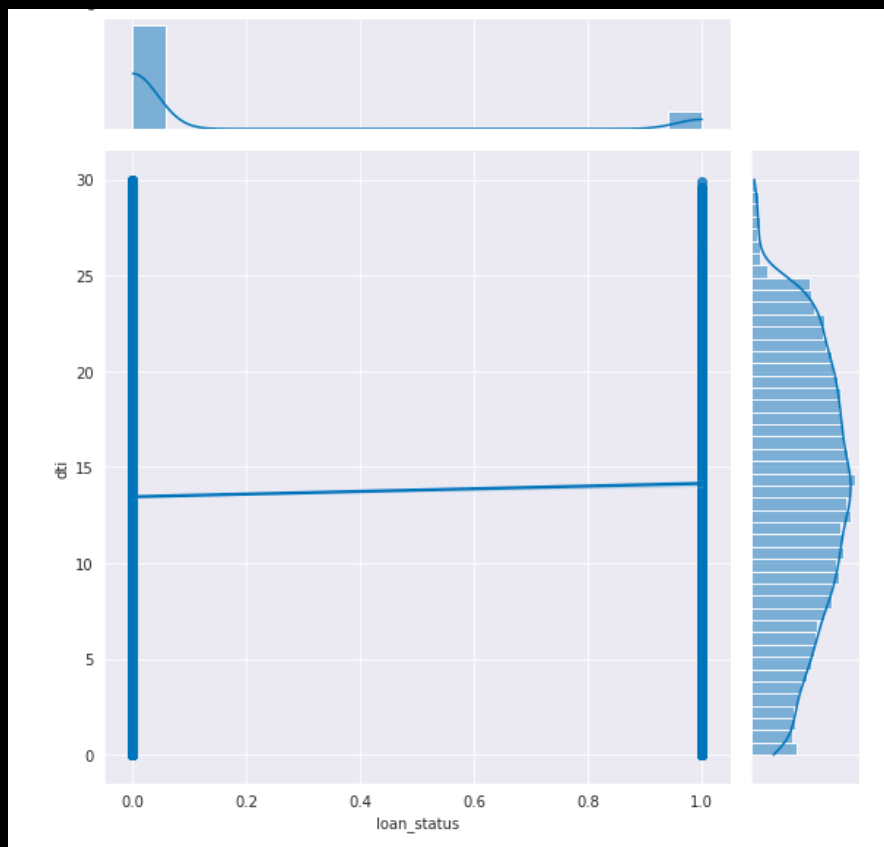# Bivariate Analysis

Interest Rate vs Loan Status

Interest rates for defaulters is slightly higher compare to fully paid borrowers.
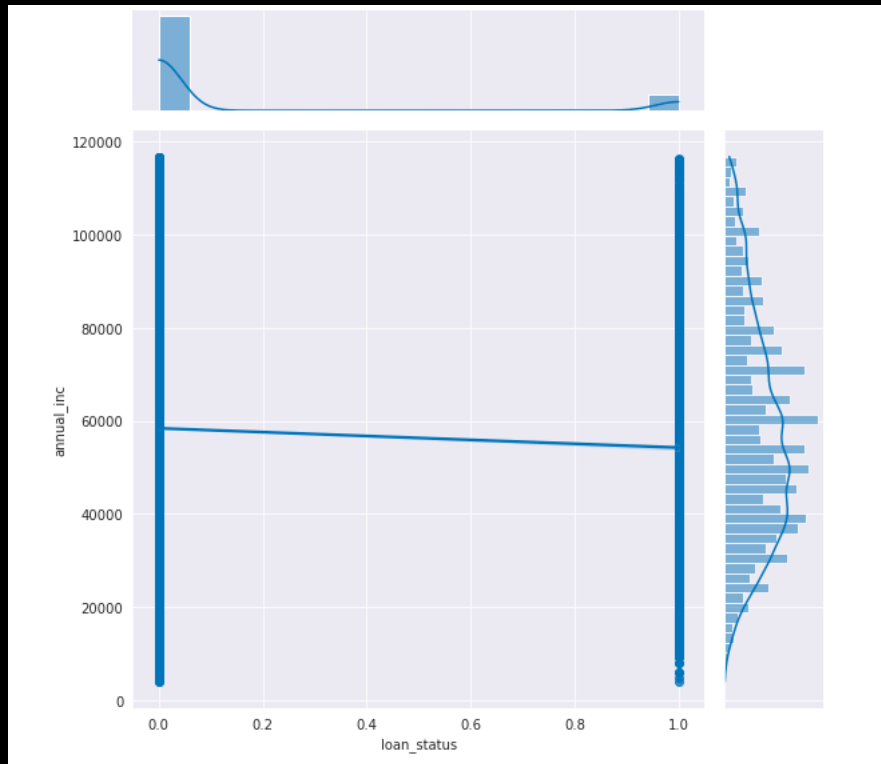
Annual Income vs Loan Status

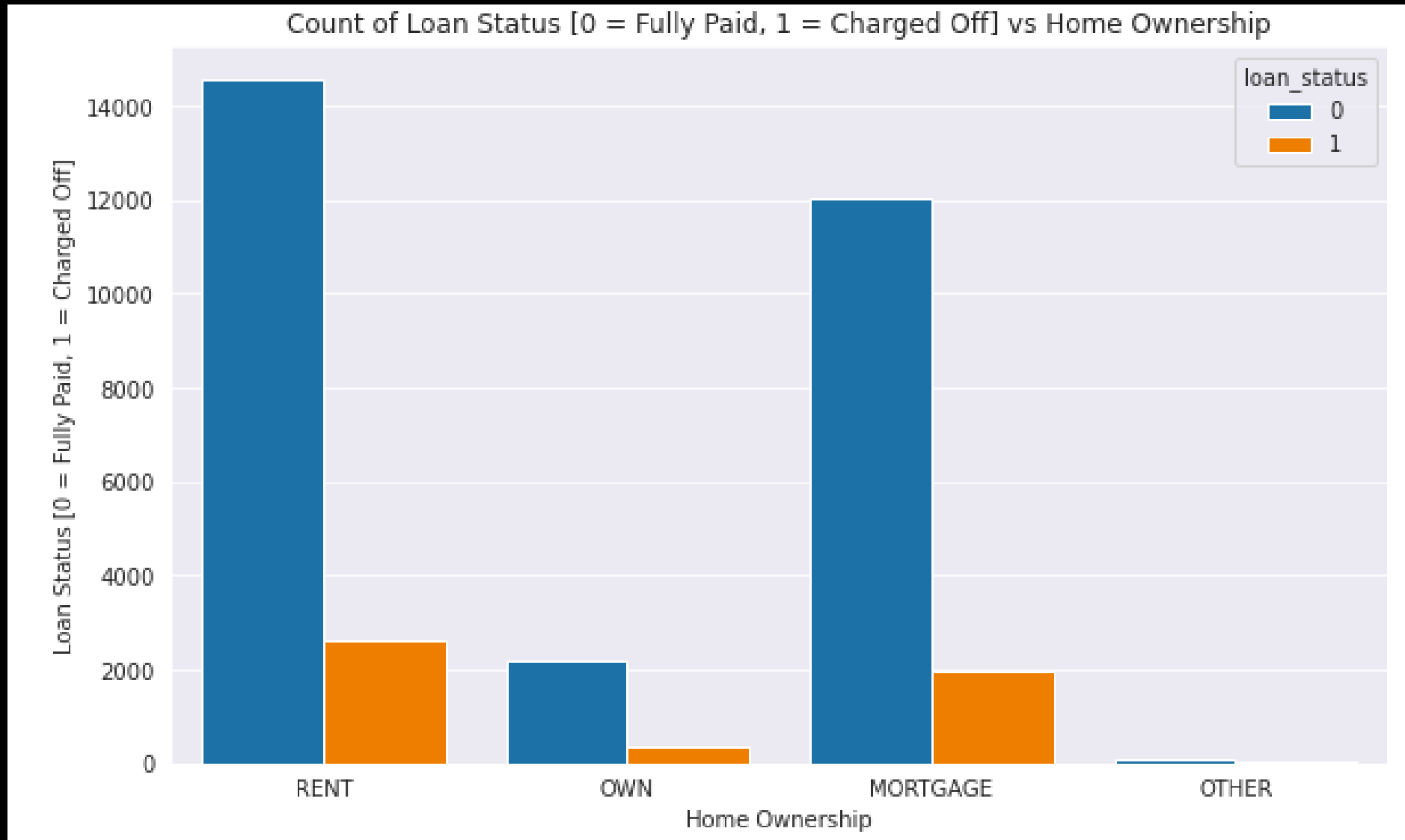Annual income of defaulters are slightly lesser compare to fully paid borrowers.
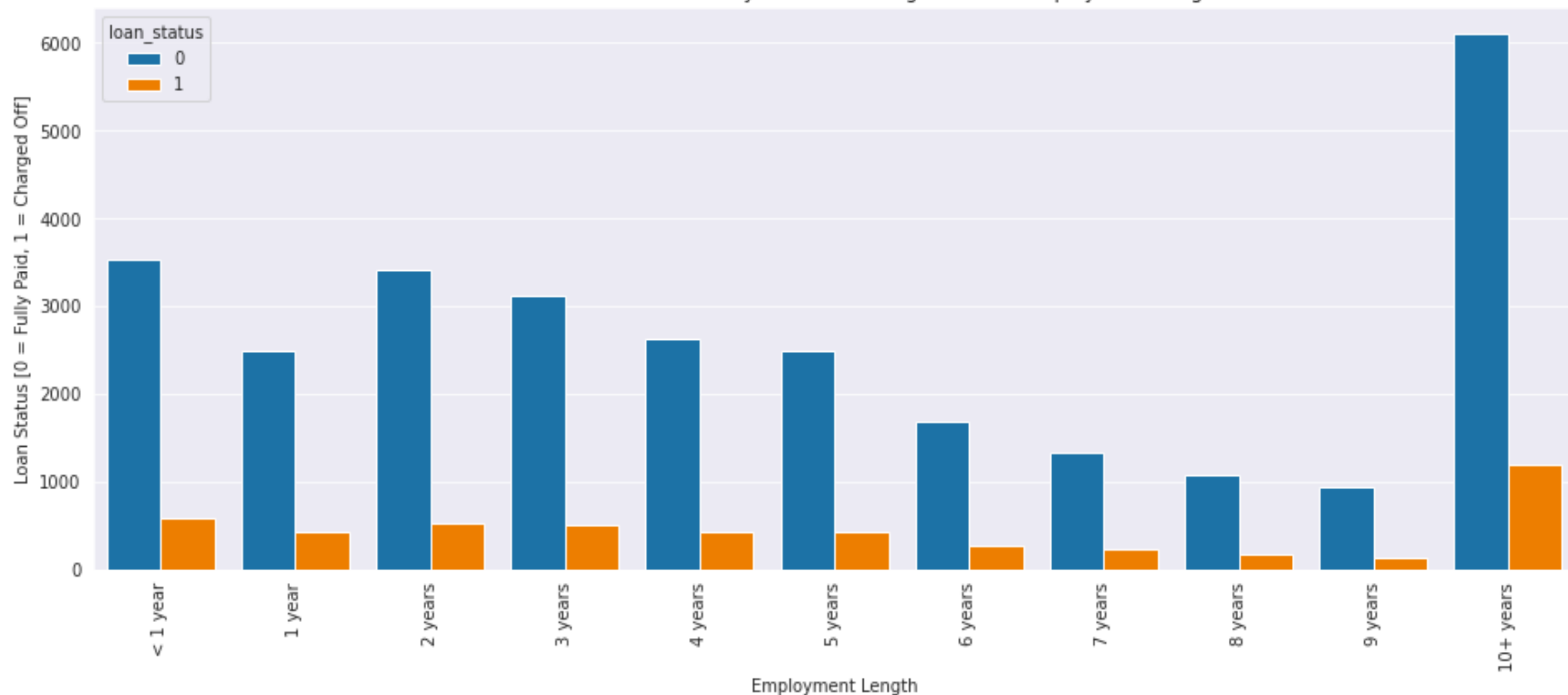
DTI vs Loan Status

DTI ratio is slightly more for defaulters.

Most defaulter belongs to RENT or MORTGAGE category of home ownership.

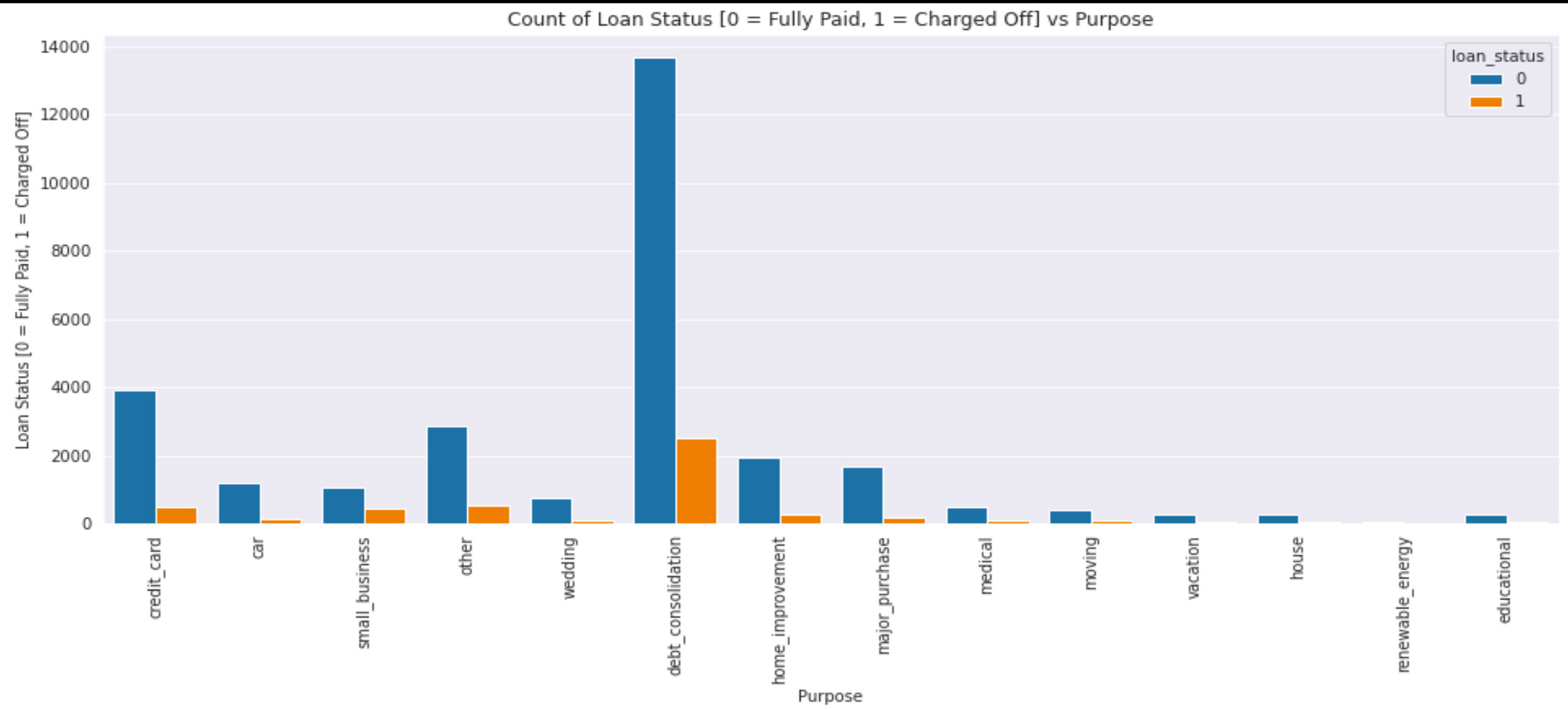Count of Loan Status [0 = Fully Paid, 1 = Charged Off] vs Home Ownership

More than 1000 defaulters are having employment length greater than 10+ years.

Count of Loan Status [0 = Fully Paid, 1 = Charged Off] vs Employment Length
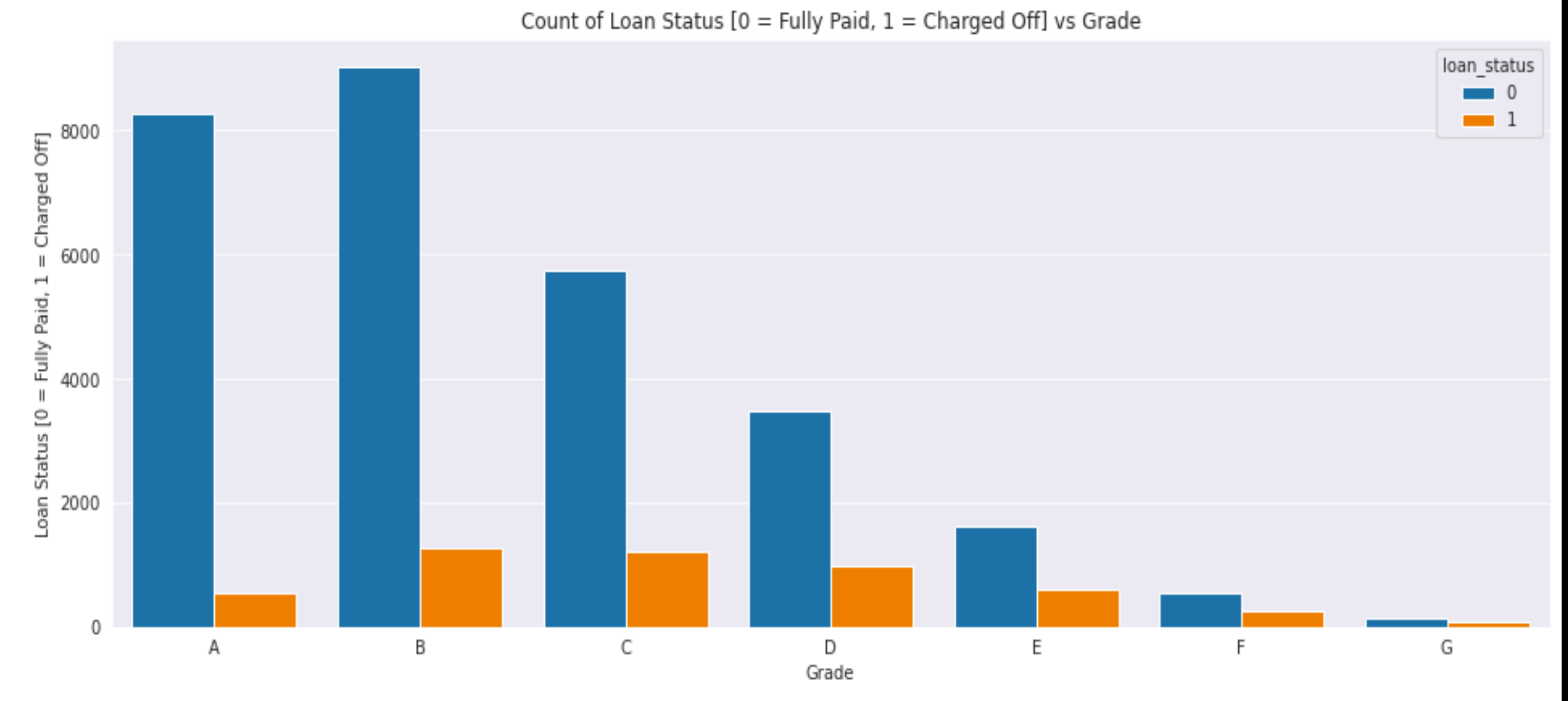
Number of defaulter are more than 3K for debt consolidation purpose.

Ratio of defaulters after grade B is quite heigher. Tendency of defaulters are more from group C to G.

Recommendations

- More than 4.5K defaulters are from MORTGAGE and RENT category of home ownership. It will be great to check the DTI and purpose for these categories before approval.

- More than 16K loans are not verified. It will be benifial to the comapny if they can reduce the number by strenthening the verification process.

- People are taking loan mostly for debt consolidation (one of the category in pupose), Approximate 3K loans are defaulters in this category. The total number is more than ~17K out of ~37K. This is a huge number. Company should think of their business approch. The number of loan for other purpose e.g. house, renewable energy and education is very low.

- More than ~1K borrowers become defaulters out of ~7K borrowers for 10+ years category of employment length. The ratio is 1 out of 7. It will be great to check the DTI and purpose for those application before approval.

- Approximately 13K loans belongs to category of less than year or more than 10 years. Please do proper verification before approval.

- There are more defaulters belongs to state CA, TX, FL, NY anf NJ as there are more loans from these state. Please do strenthen the process of verification in these states.

- DTI ratio should not be more than 13.5 to avoid defaulters.

- Average annual income should be greater than 58K (USD) to avoid defaulters.