When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- 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

Contents

- ❑Requirement
- ❑Data understanding
- ❑Data cleaning
- ❑Data analysis
- ❑Recommendations

Requirement

The aim is to identify the 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.

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:
    a. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    b. **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'.
    c. **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 understanding

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

Libraries used:

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Datetime

➢The Loan dataset contains **39,717** records comprising of **111** features

Data cleaning

Total number of columns having missing values = **68** (61.26%)

- **54** columns/features have *NULL* values
- **14** columns/features have missing values (*NOT NULL*)
  - ✓ There are 10 features which has <30% missing values
  - ✓ There is only one feature (*desc*) which has >30% & <60% missing values
  - ✓ There is only one feature (*mths_since_last_delinq*) which has >60% & <90% missing values
  - ✓ *Only 2* features (*mths_since_last_record & next_pymnt_d*) who have more than >90% & <100% missing values

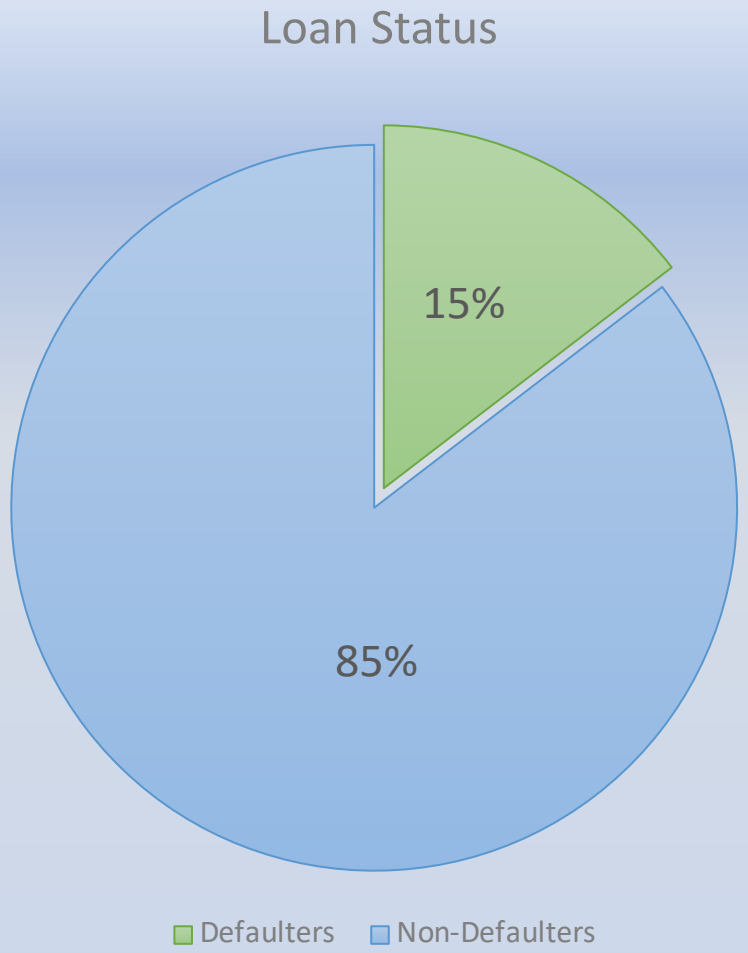
There are no rows with high percentage of missing values.

Data analysis

Few important features which will help us in identifying whether the loan applicant will be a *Defaulter* or *Non-Defaulter*:

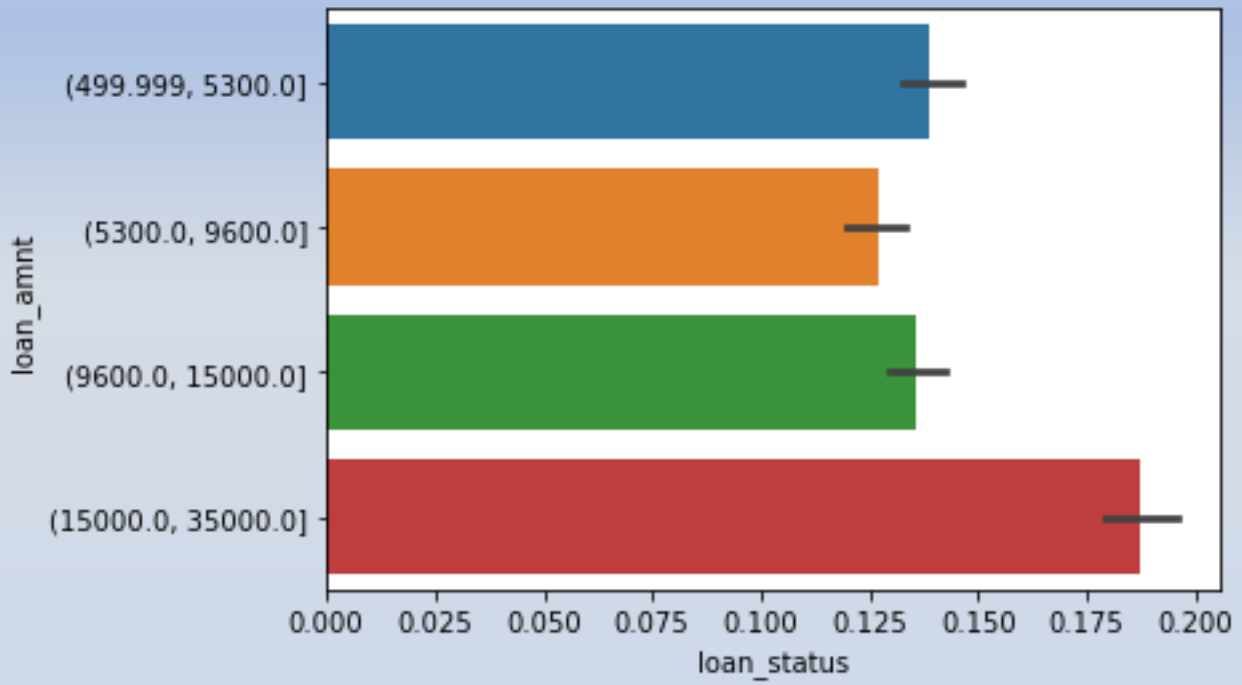
- **loan_amnt**
- **Term**
- **int_rate**
- **Grade**
- **sub_grade**
- **emp_length**
- **home_ownership**
- **annual_inc**
- **verification_status**
- **issue_d**
- **purpose**
- **dti**

The target feature/column is **loan_status**

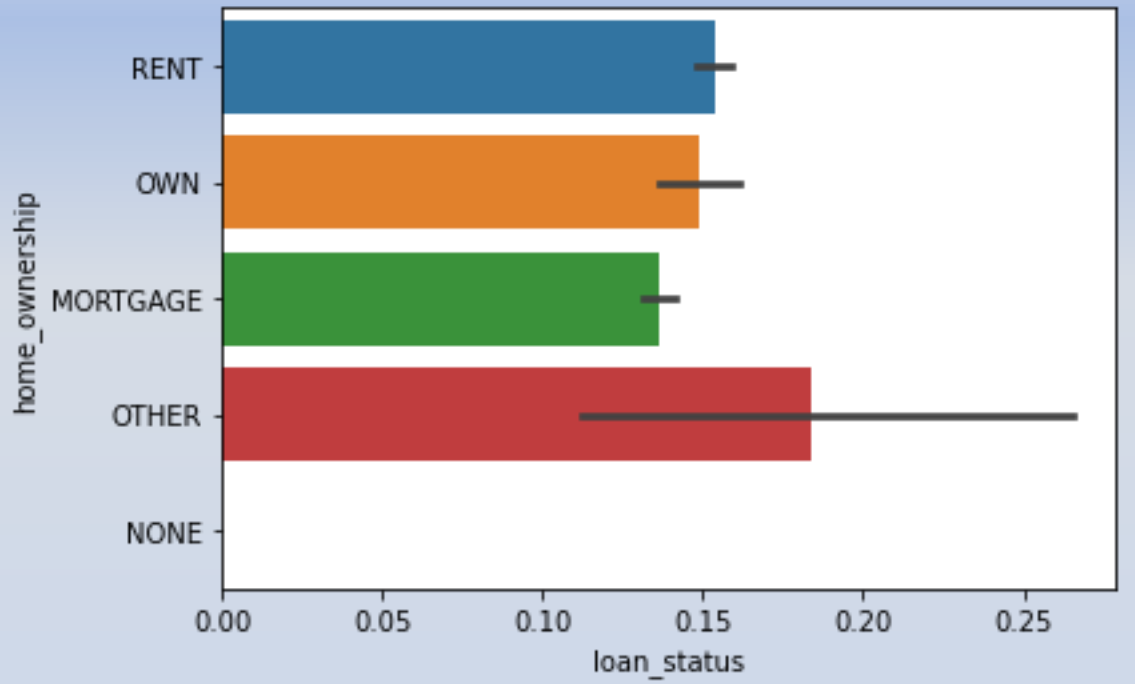


Defaulters = 14.6 %

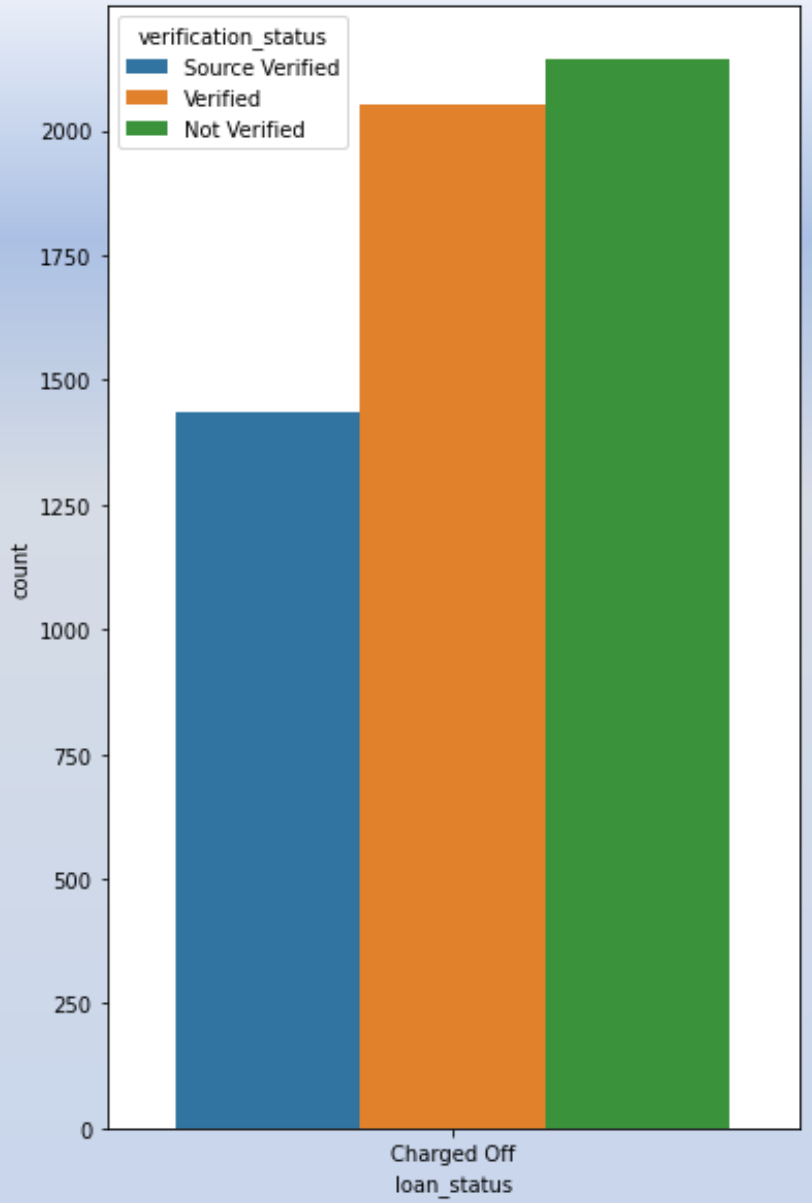
Non-Defaulters = 85.4 %

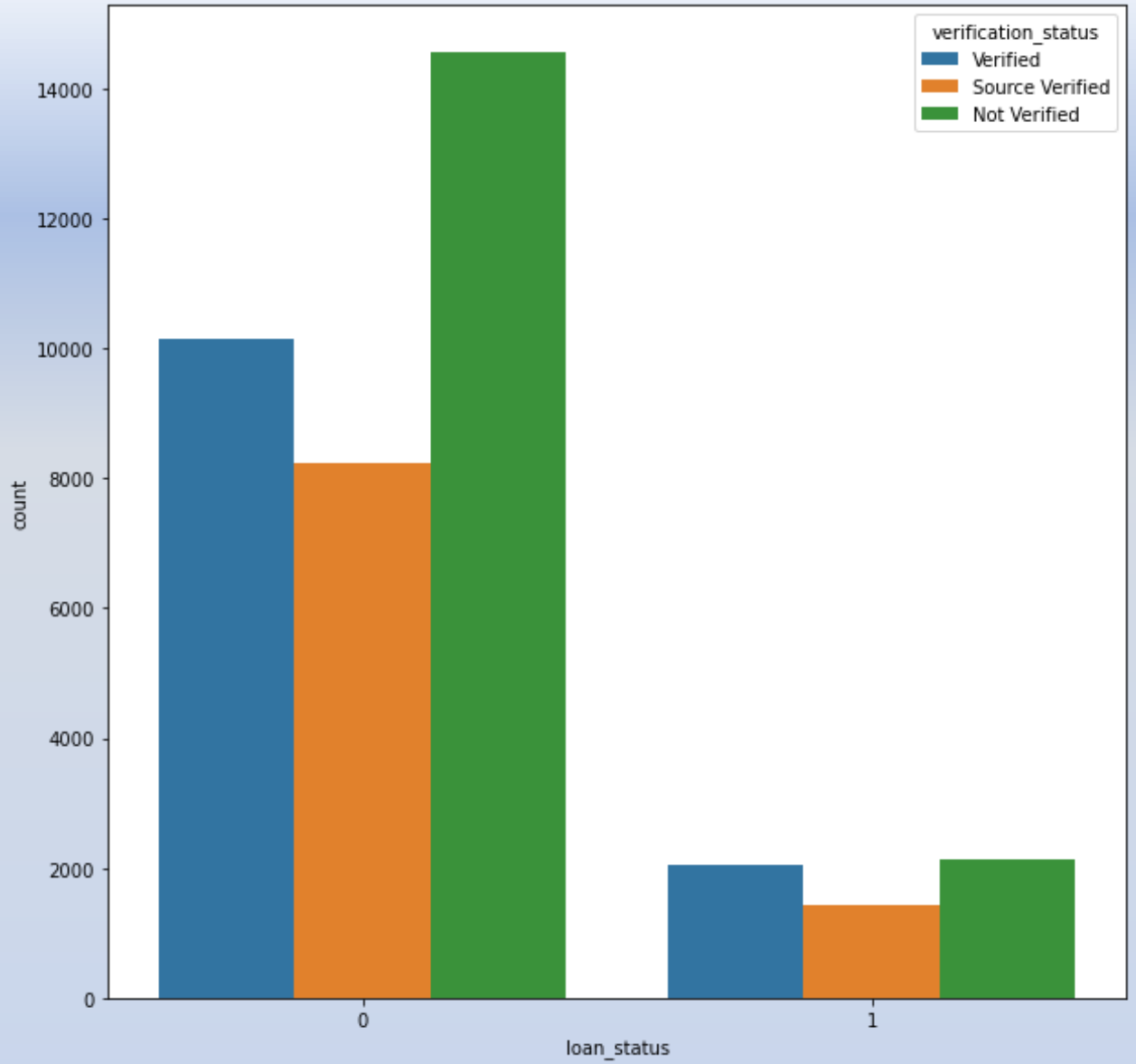


Default rate is high when loan amount is in the range of **(15000, 35000)**

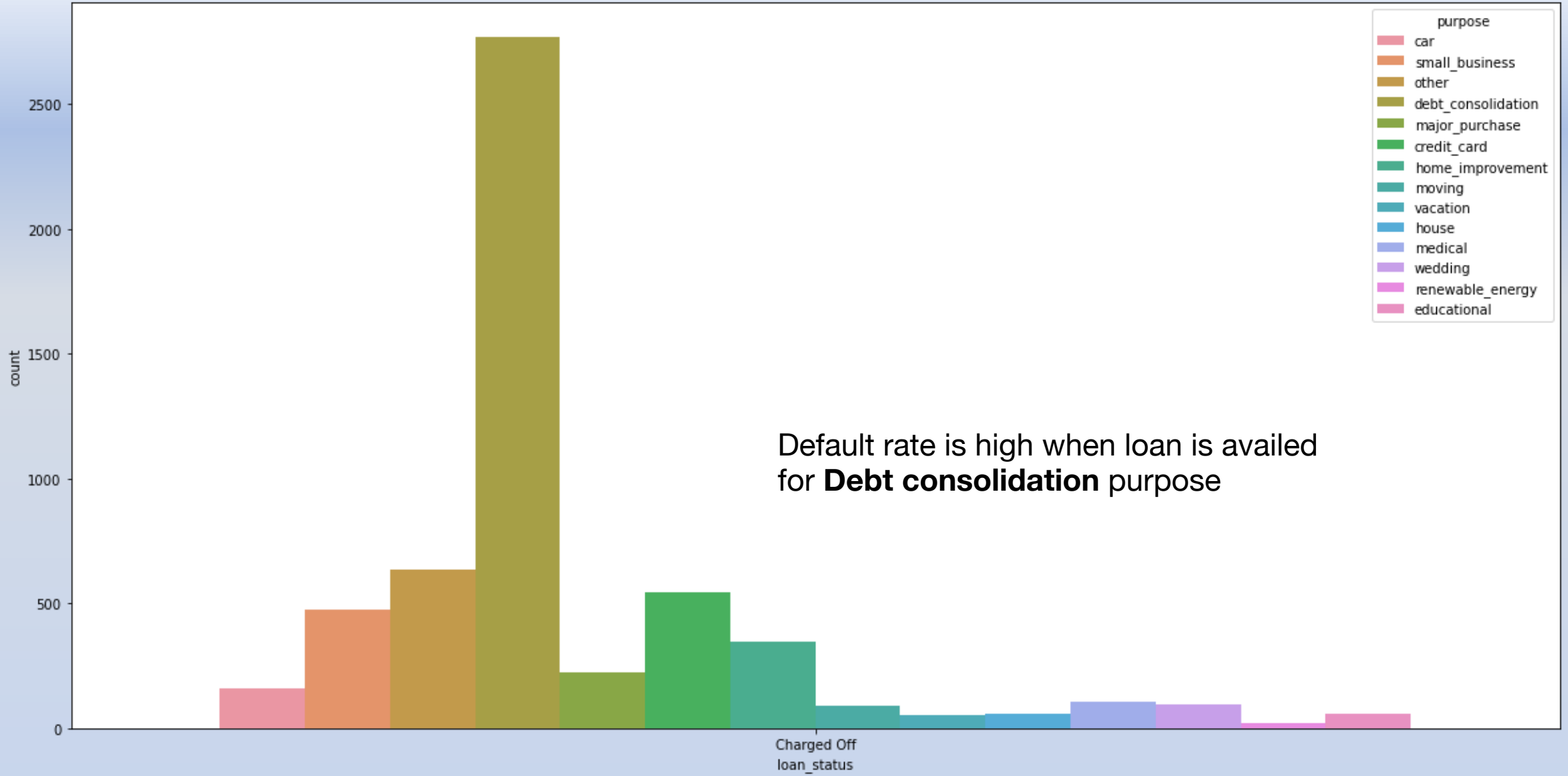


Default rate is high in home ownership of type **Others**

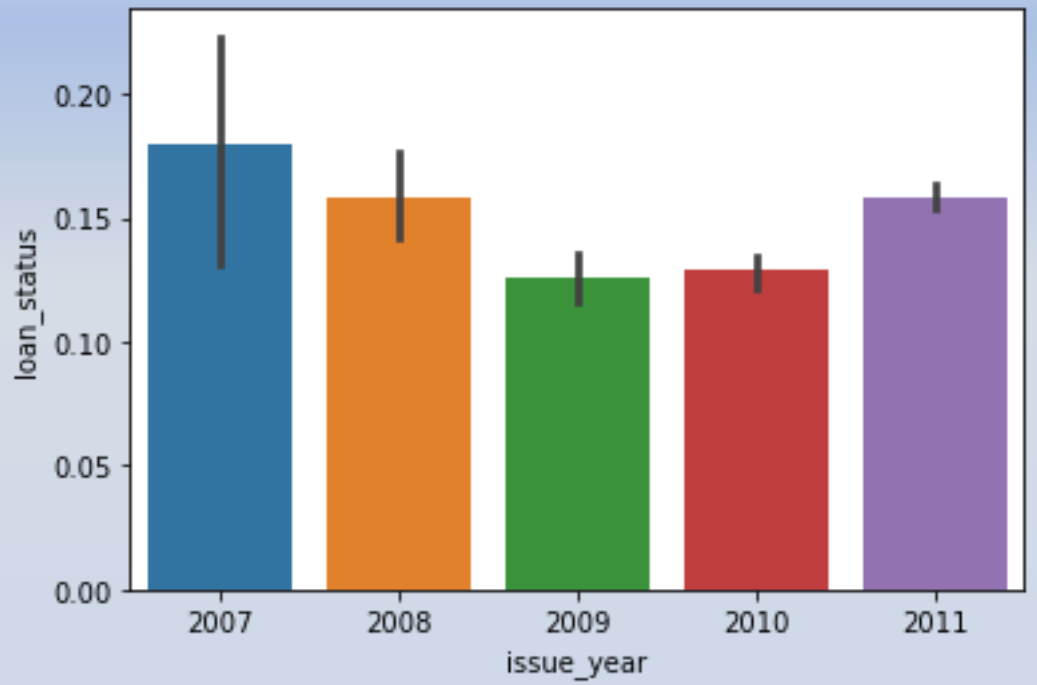




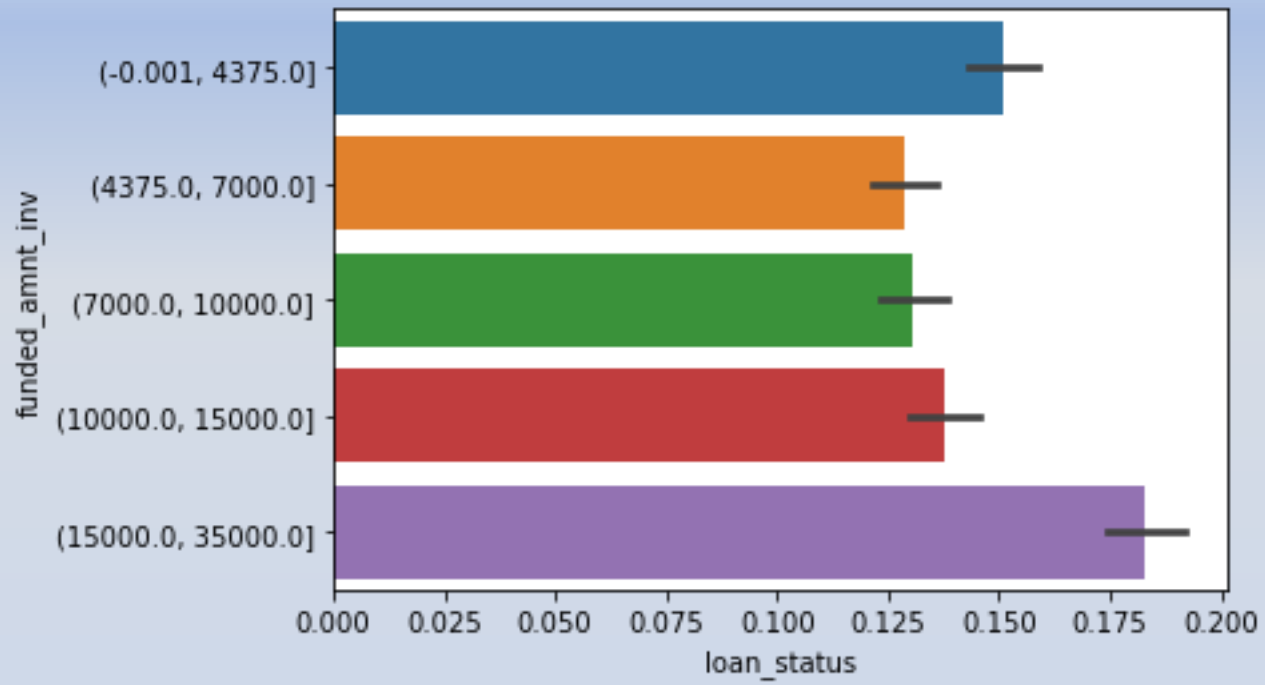
Not Verified loans have high chances of being Default



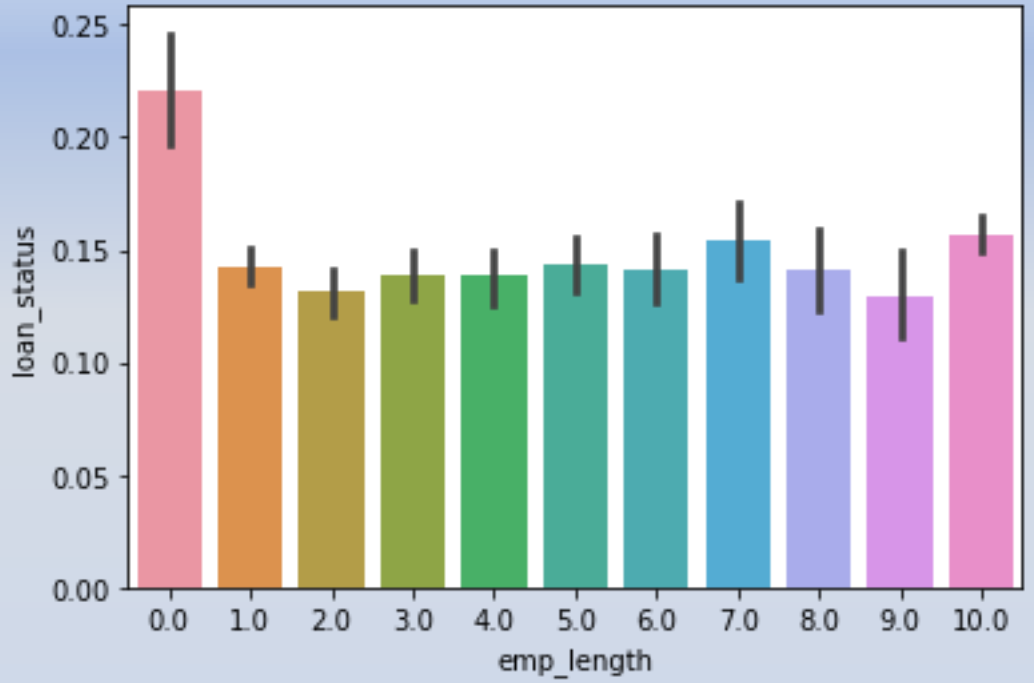
Default rate is high when loan is availed for **Debt consolidation** purpose



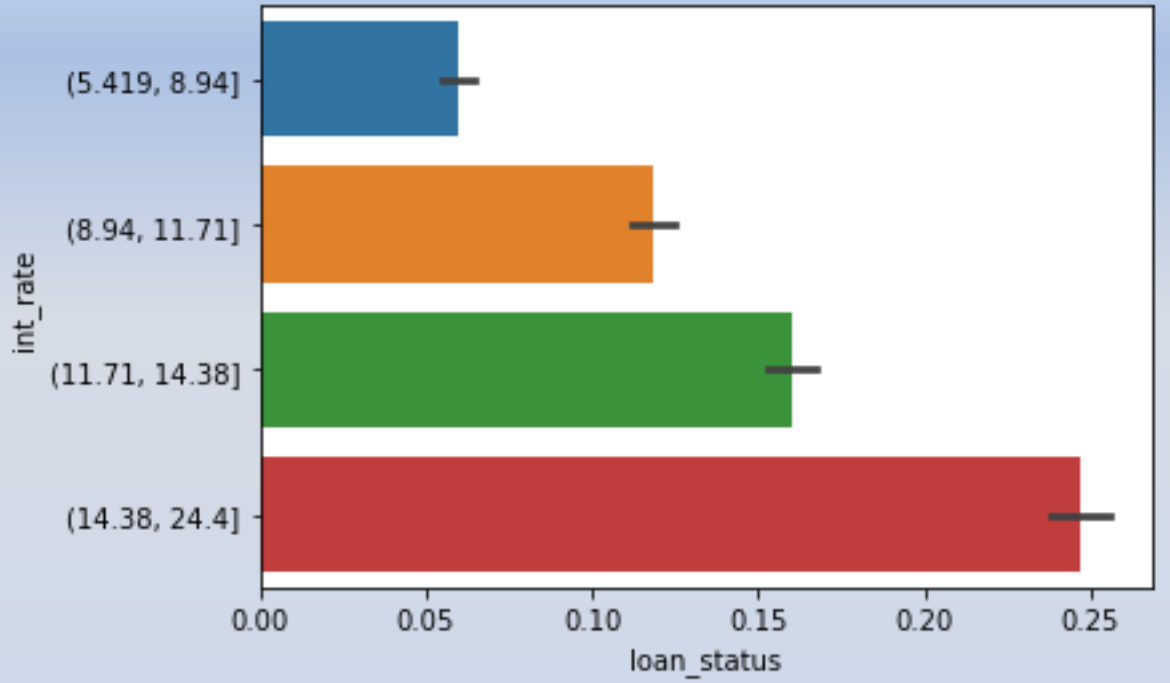
Loans availed in the year **2007** have maximum default rate



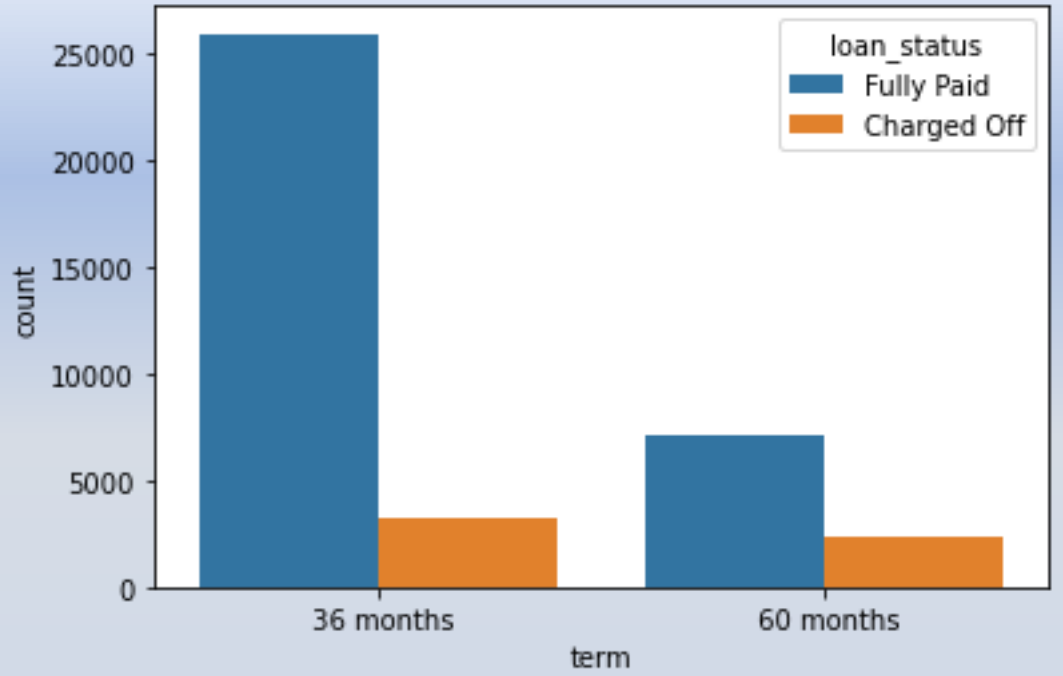
Default **percentage** is high when investor invests the amount in the range of **(15000, 35000)**



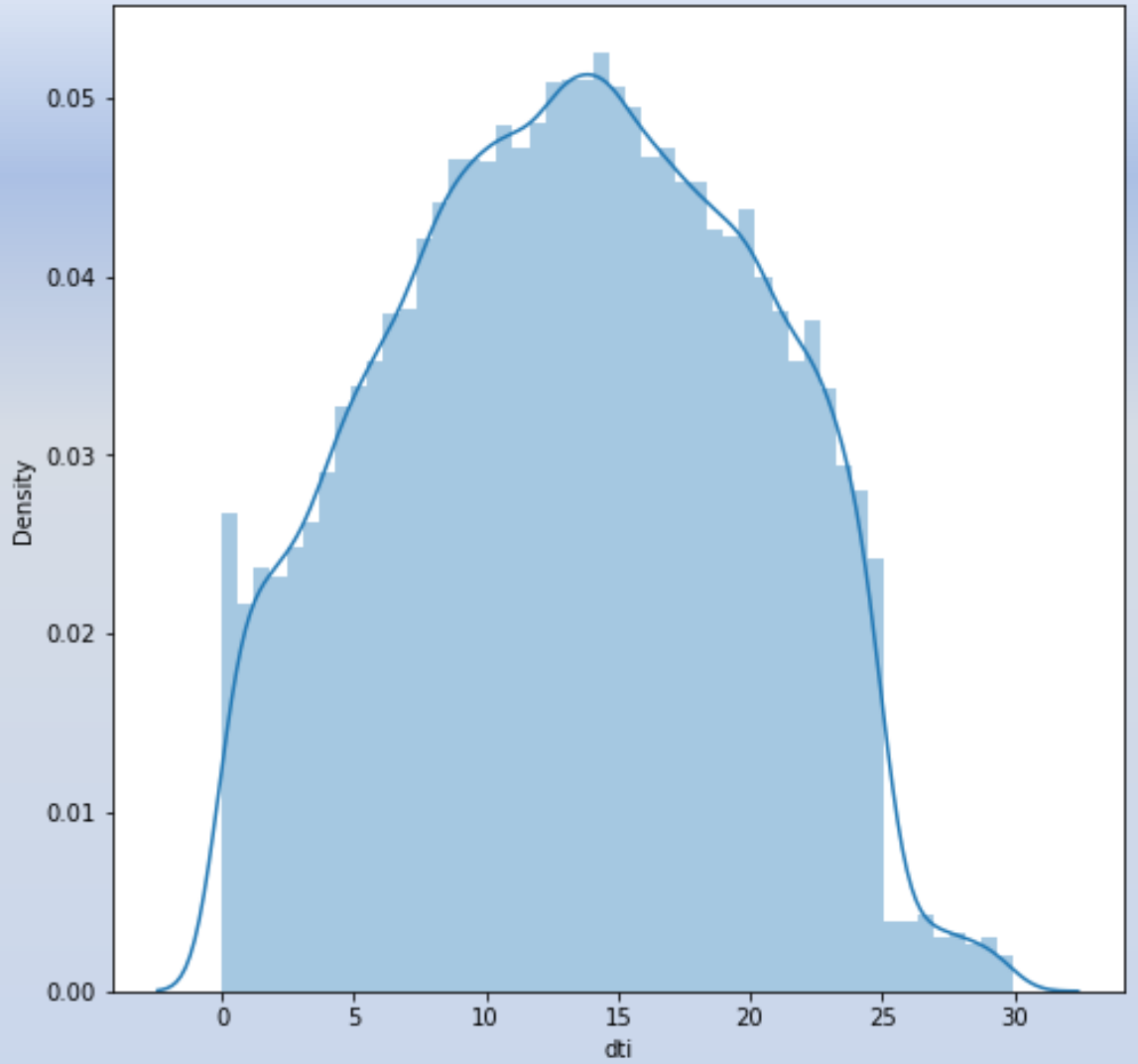
Loan applicants whose employee length is *n/a* (**0**) have high chances of getting defaulted. Next category is employees having more than **10+ years**



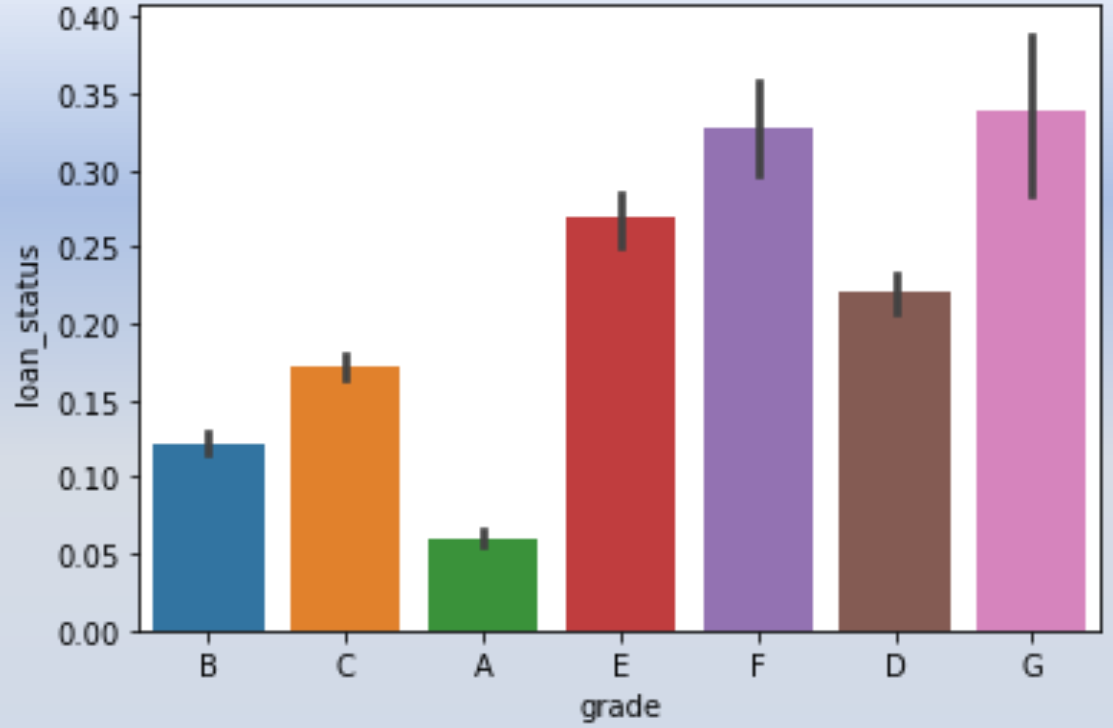
Default **percentage** is high when loans have a interest rate of more than **14%**



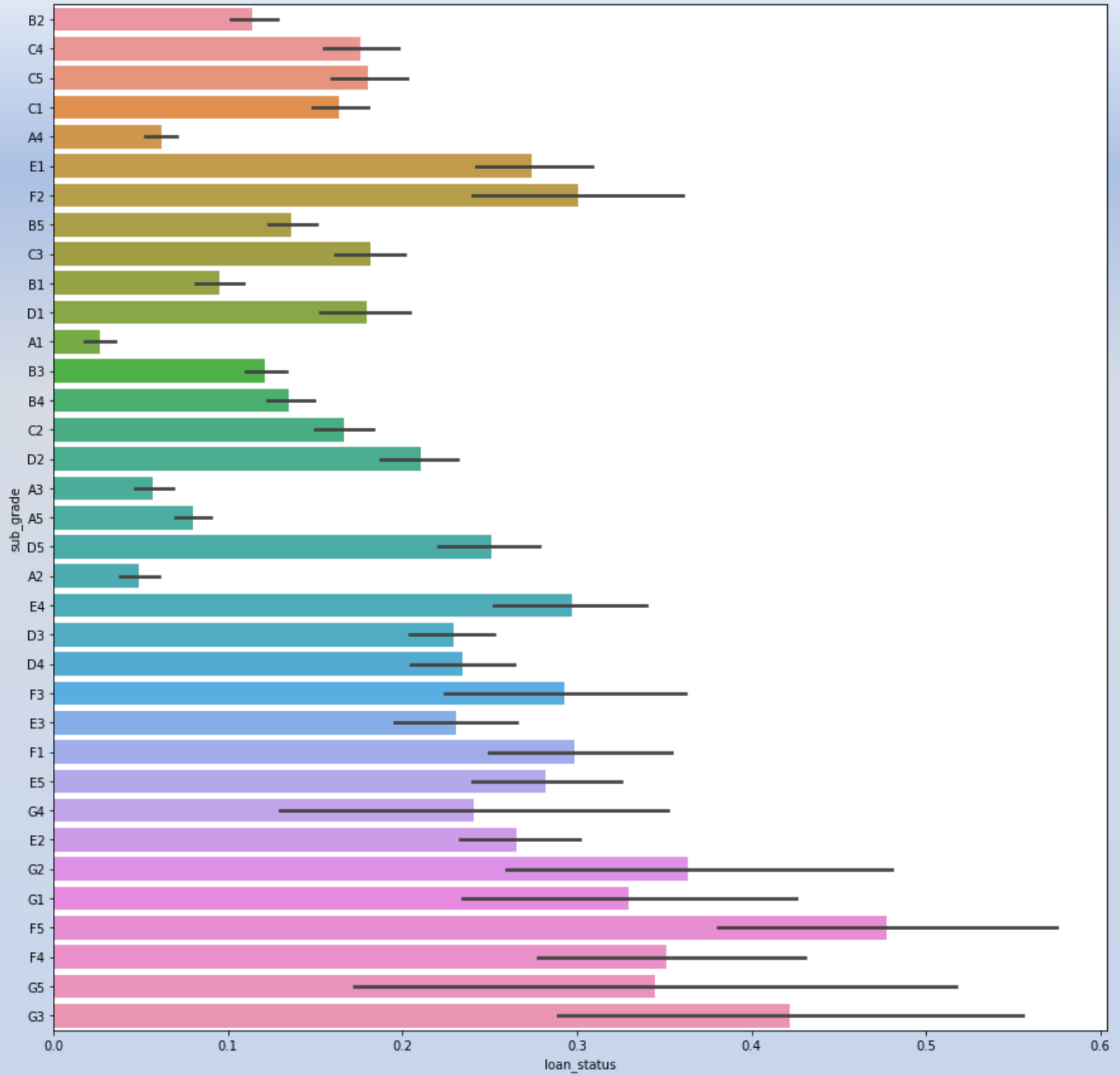
Default rate is high when loan term is **36 months**

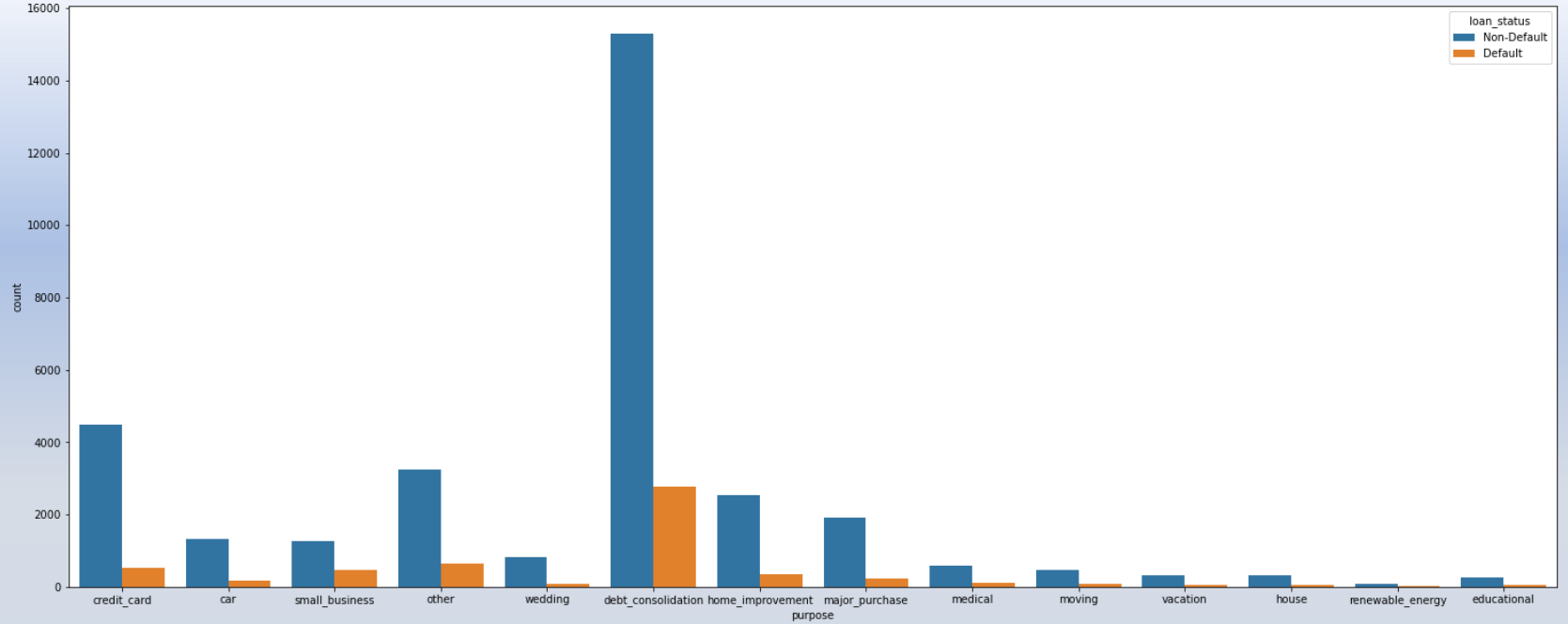


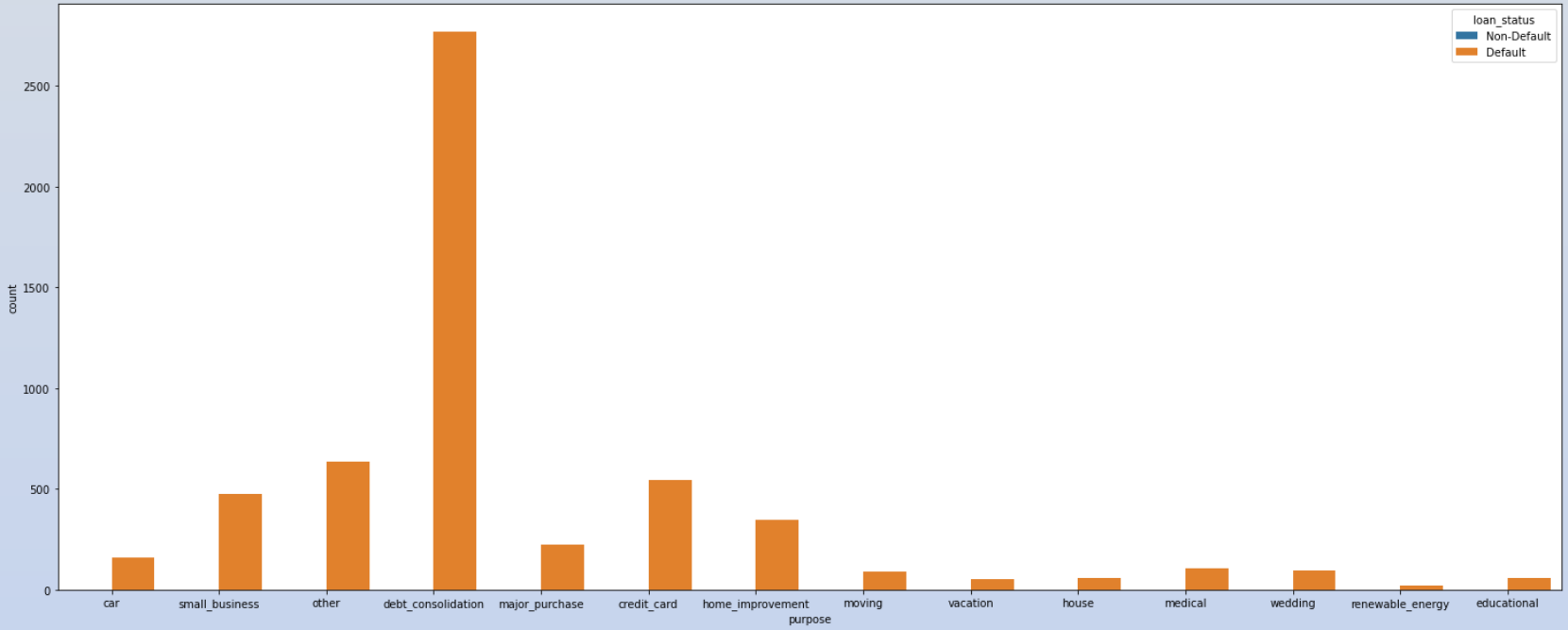
Default rate is high when *dti* is in the range of **(13, 25)**



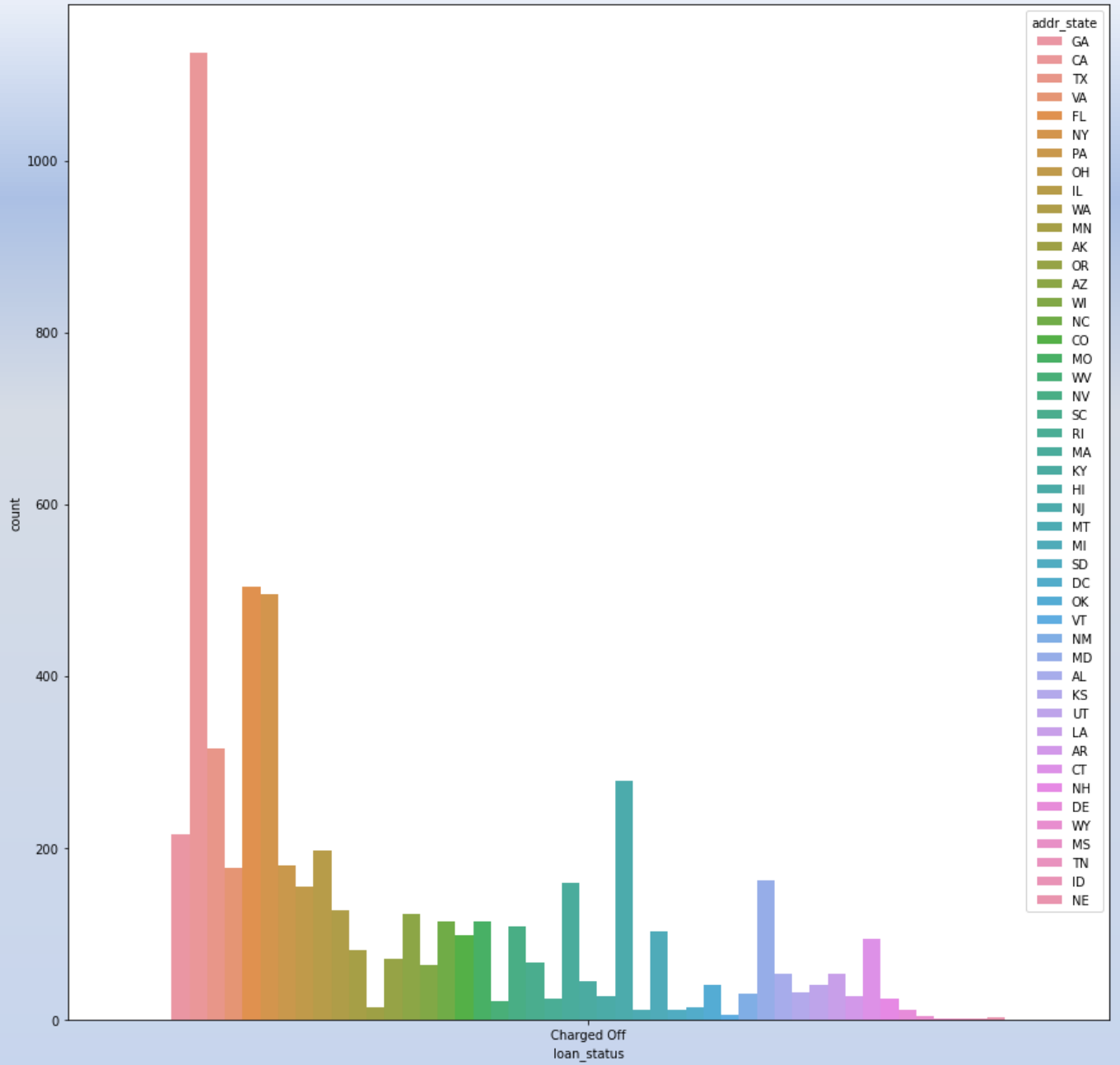
Default **percentage** is high in the grades belonging to **F** and **G** having sub-grades **F5** and **G3** respectively



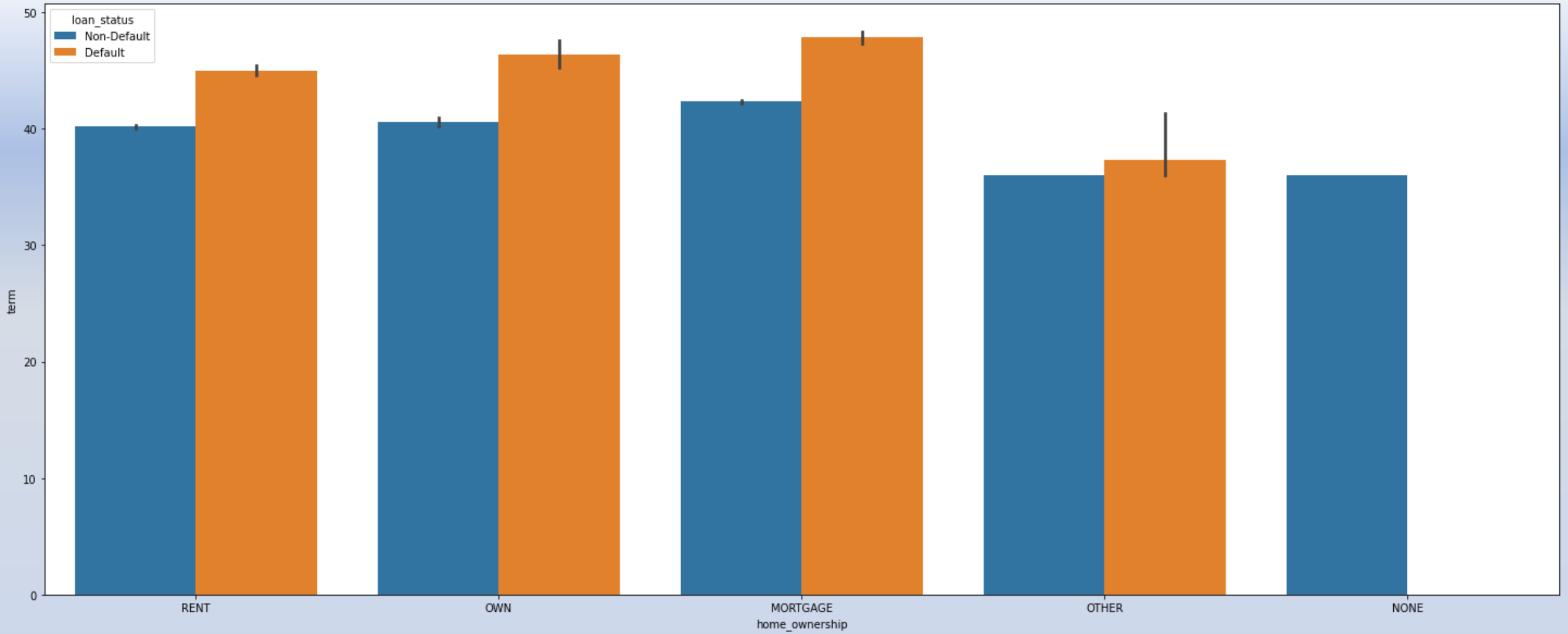




Default rate is high
for **debt_consolidation** category



Default rate is very high in **CA** (California) state



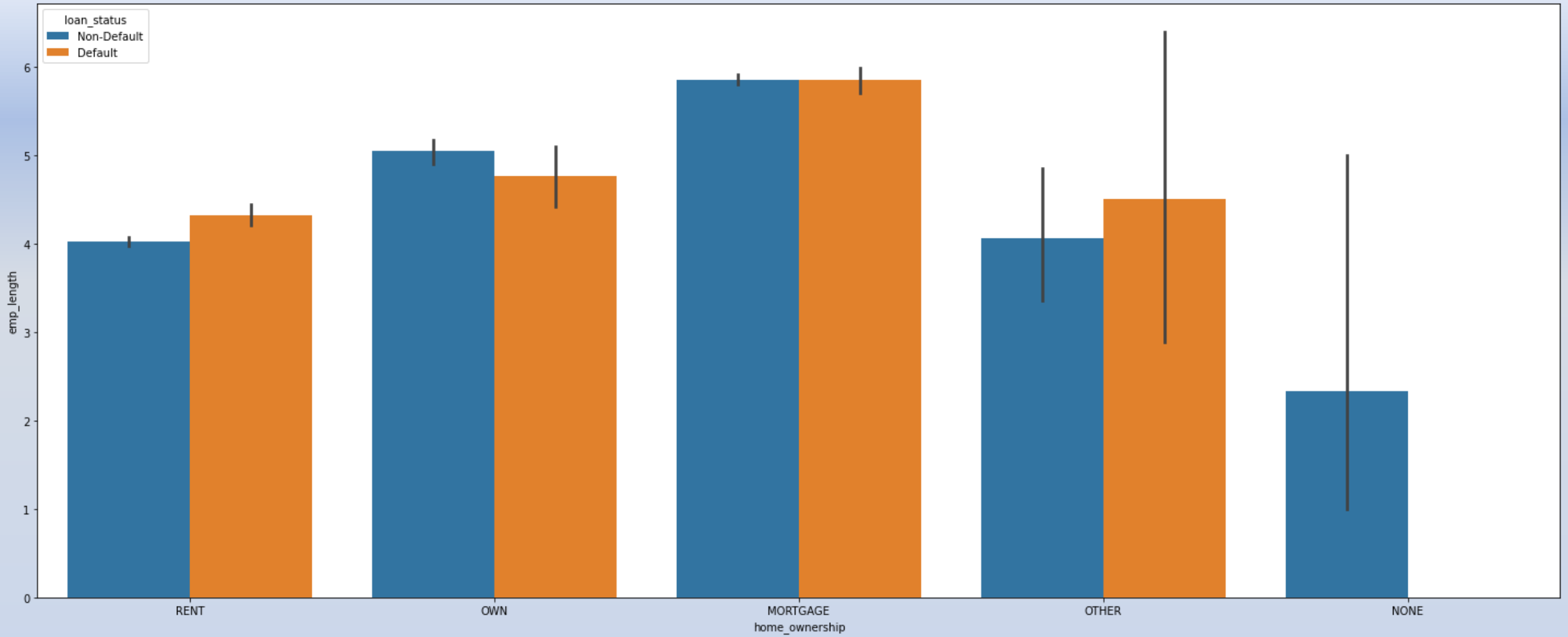
Default **percentage** is high if term is **more than 36 months** across following categories of home_ownership:

1.OTHER --> 100%

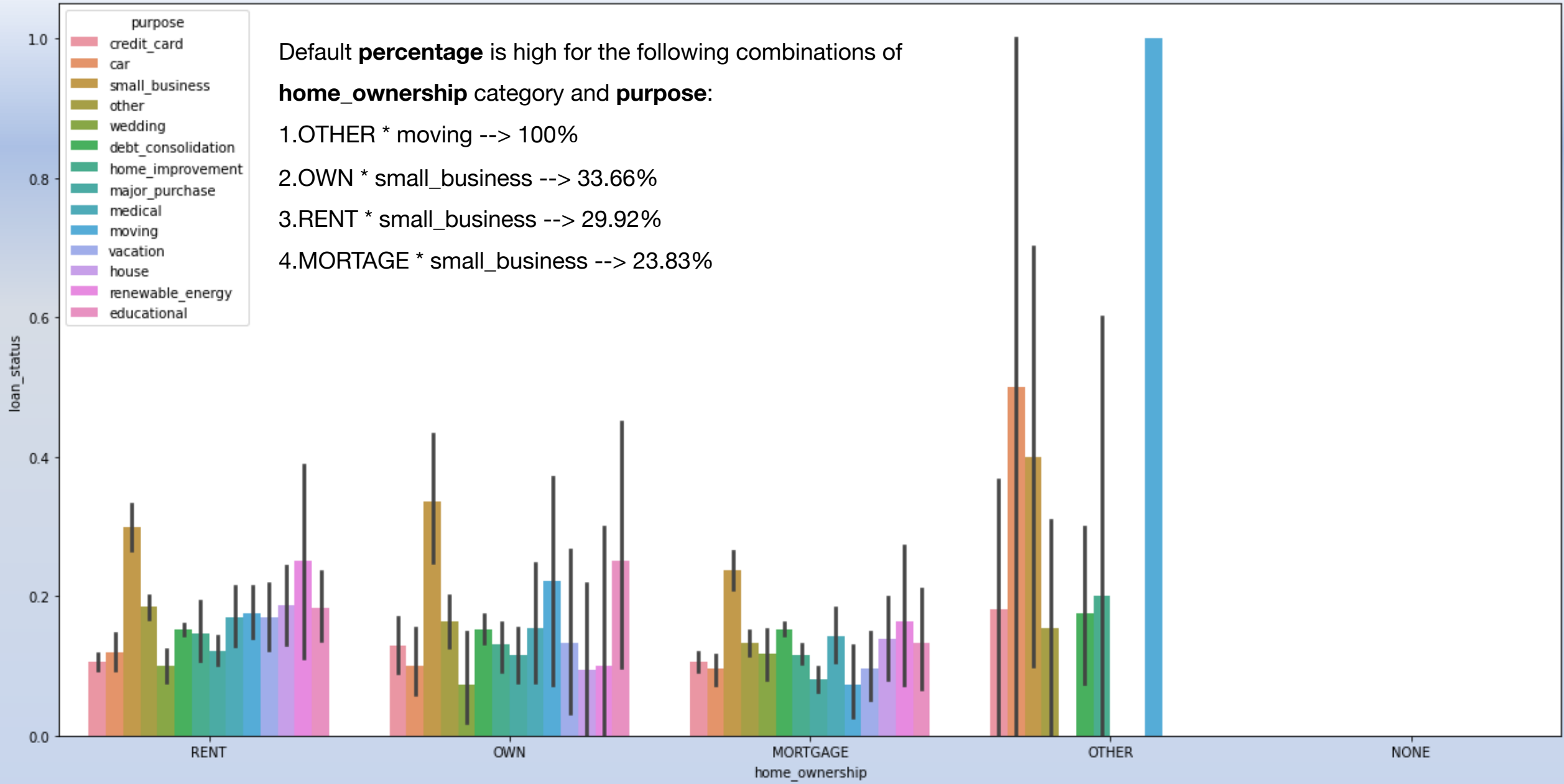
2.OWN --> 28.34%

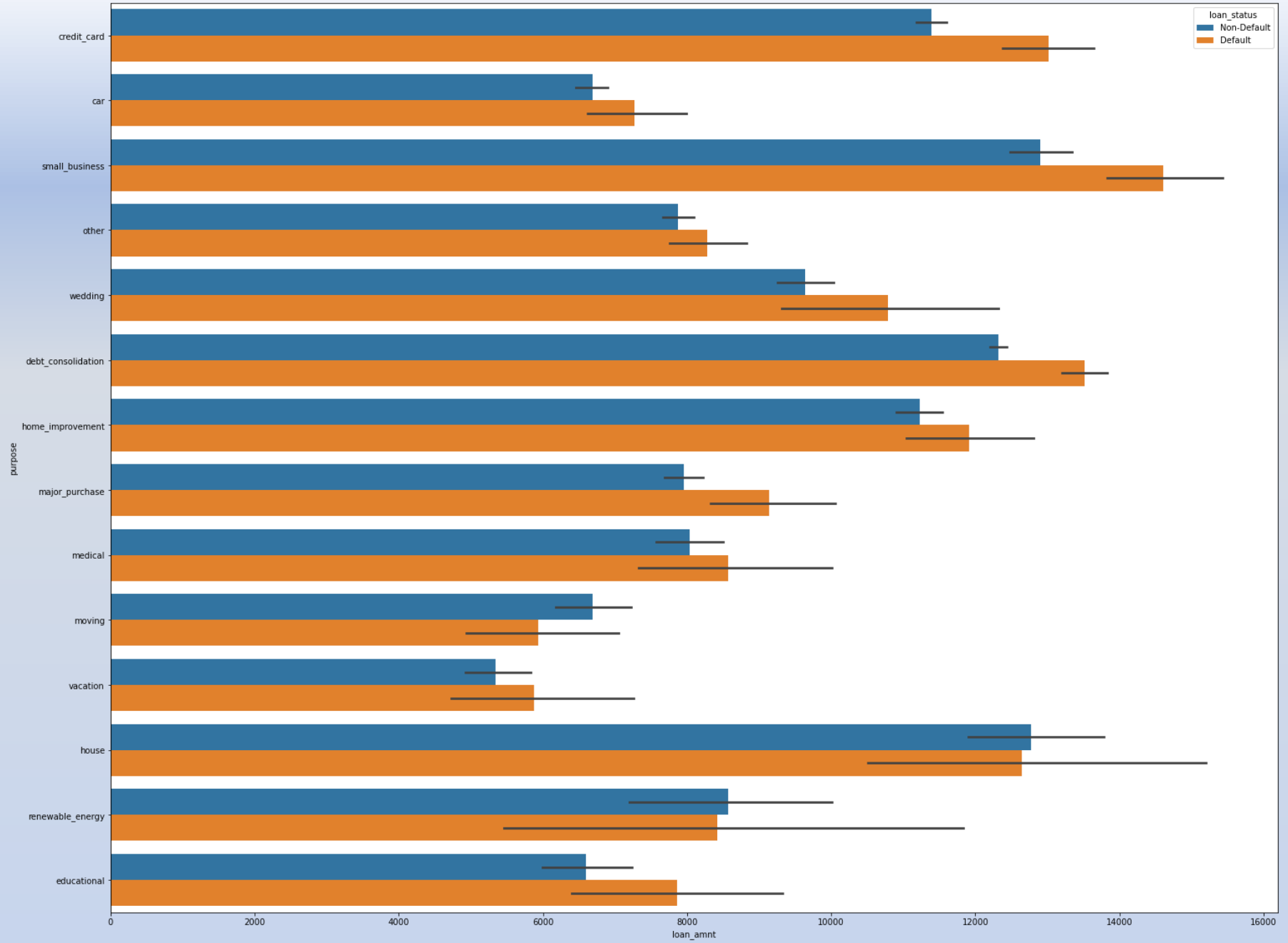
3.RENT --> 28.04%

4.MORTGAGE --> 22.84%

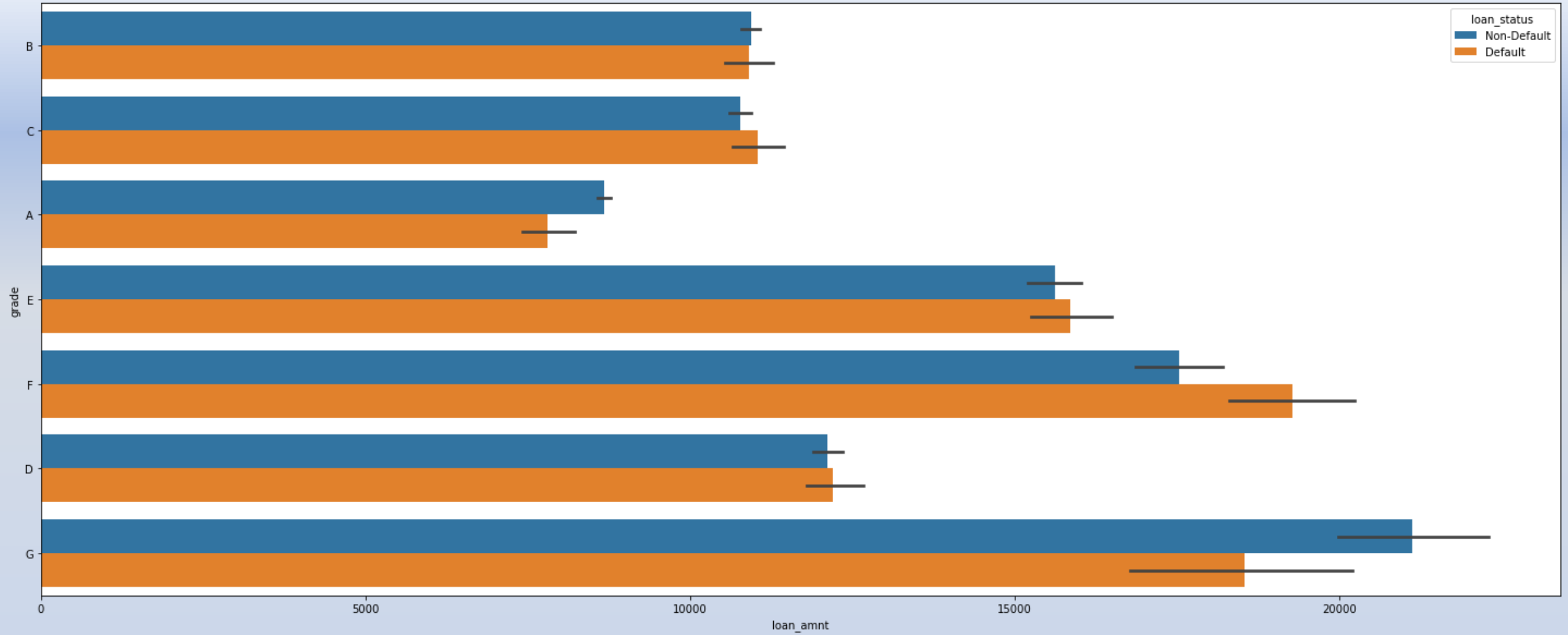


Maximum defaulters are found in **OTHER** category of home_ownership whose emp_length is **more than 4 years**

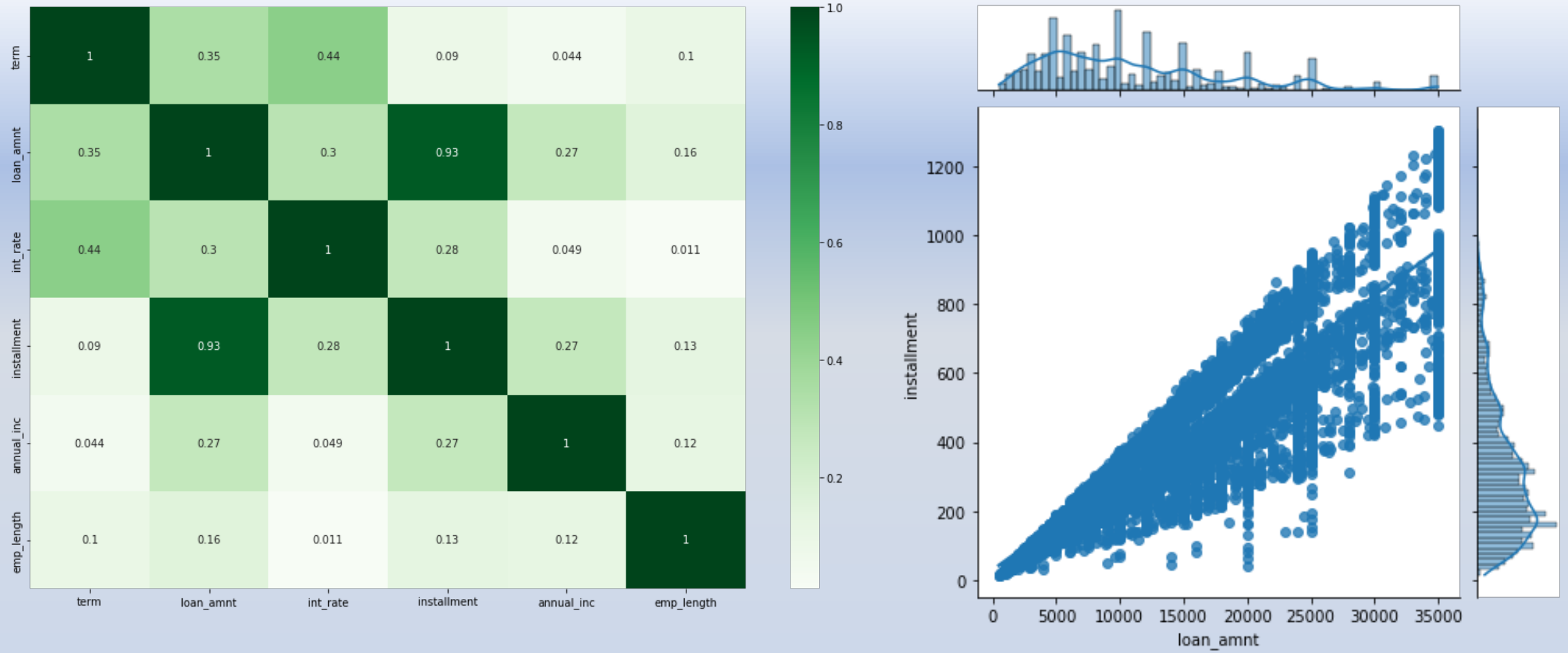

purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
Default **percentage** is high for the following combinations of
home_ownership category and **purpose**:
1.OTHER * moving --> 100%
2.OWN * small_business --> 33.66%
3.RENT * small_business --> 29.92%
4.MORTAGE * small_business --> 23.83%
loan_status
1.0
0.8
0.6
0.4
0.2
0.0
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership



Default **percentage** is higher when loan_amnt issued is higher (**>10000**) for small_business purpose



Default **percentage** is 45.1% for grade F when loan_amnt is greater than **15000**



From the heatmap and regression plot, it is very clear that there is a high correlation between loan_amnt and installment (**0.93**) i.e., as the loan amount increases, the instalments also increases.

Recommendations

Features which are highly influencing the Default v/s Non-Default loans are as follows:

- Loan amount
- Term
- Grade
- Home Ownership
- Purpose
- Verification status

Points to consider before approving the loan:

1. Default **percentage** is high when loan amount is in the range of (15000, 35000)
2. Not Verified loans have high chances of being Default
3. Default rate is high when loan is availed for Debt consolidation purpose
4. Default **percentage** is high when investor invests the amount in the range of (15000, 35000)
5. Default **percentage** is high when loans have a interest rate of more than 14% and loan term is 36 months
6. Default **percentage** is high in the grades belonging to F and G
7. Maximum defaulters are found in OTHER category of home_ownership whose emp_length is more than 4 years
8. Default **percentage** is higher when loan_amnt issued is higher (>10000) for small_business purpose