Gramener EDA Case Study

Submitted by -

Aman Rai

Utkarsh Kant

Mukesh Tibrewala

Deva Sharma

Business Understanding

You work for a consumer finance company which specialises in lending various types of loans to urban customers. 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

Business Objectives

Business Objective:

Gramener is looking for the attributes in a applicant profile which can help them in deciding whether to approve or decline the loan application.

Goal of Analysis:

To find out the relation between the different attributes and their impact on loan default. And suggest which attributes contributes a significant difference in Loan Default.

Gramener may choose to utilize this knowledge for its portfolio and risk assessment of new loan applicants

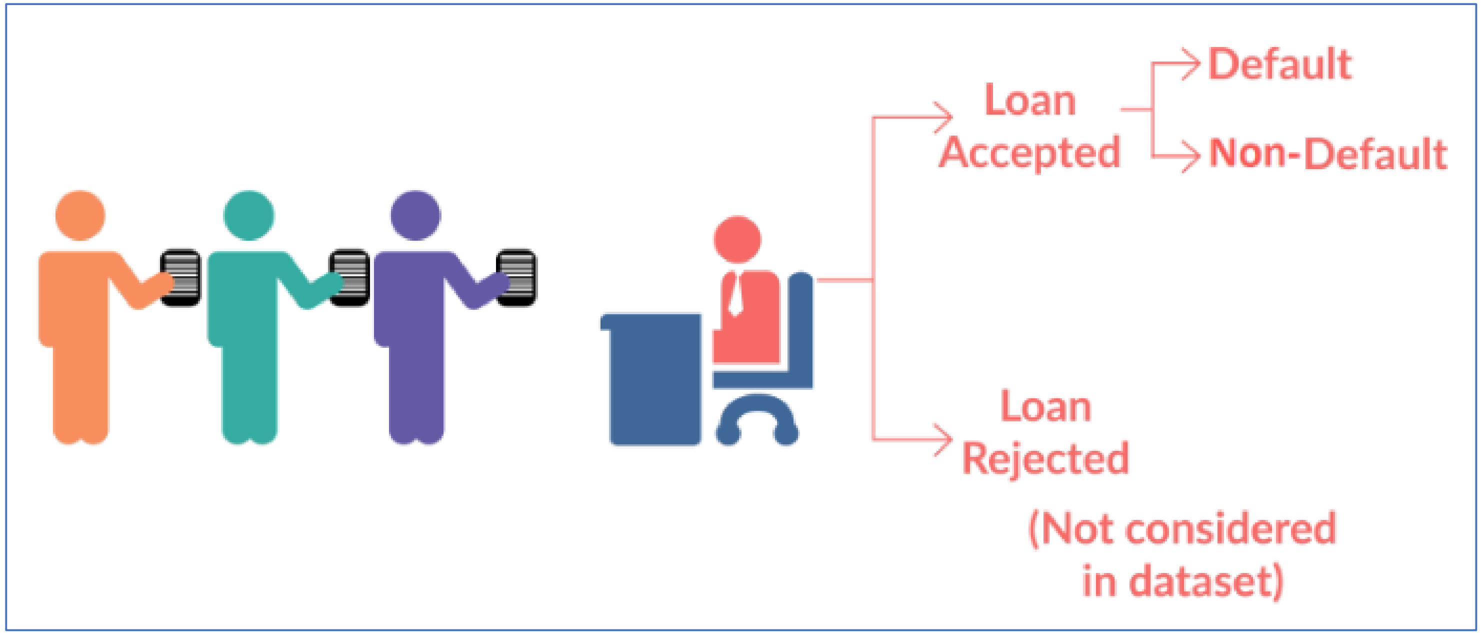
Loan Decision Criterion

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

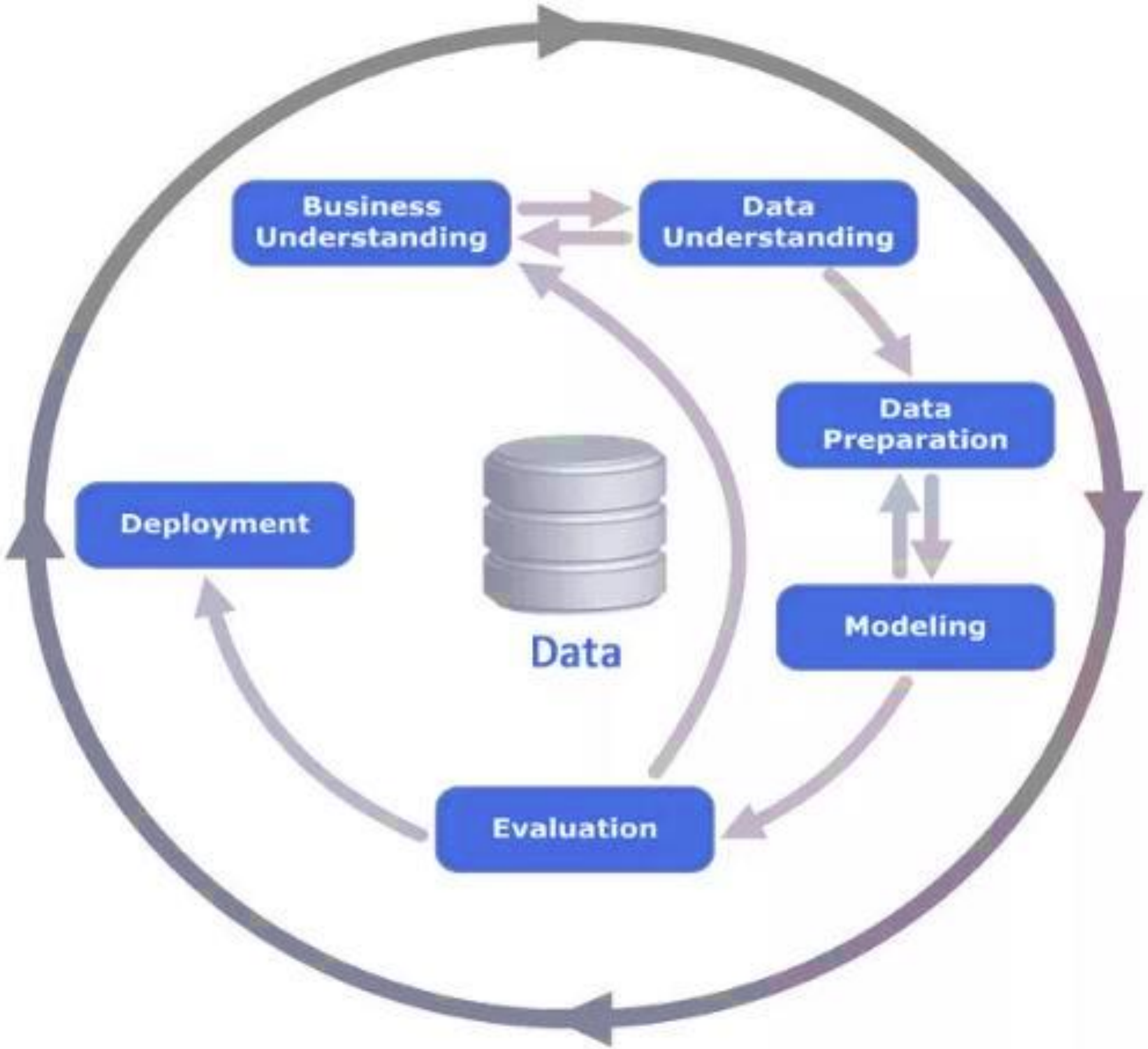
1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

   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) **Chargedoff:** 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)



CRISPDM Methodology



Understanding the Data

The company has come across some important attributes in order to understand the behavior of their approved loan customers w.r.t. the loan defaulters.

Thus, the lending company has decided to work only on these variables to mitigate the future risk. The driver variables considered for this case study are:

annual_inc	-	Annual Income of applicant
loan_amnt	-	The listed amount of the loan applied for by the borrower
funded_amnt	-	The total amount committed to that loan at that point in time
int_rate	-	Interest Rate on the loan
grade	-	LC assigned loan grade
dti	-	Debt to income ratio
emp_length	-	Employment length in years
purpose	-	A category provided by the borrower for the loan request.
home_ownership	-	The home ownership status provided by the borrower during registration
loan_status	-	Current status of the loan

Data Cleaning

Data Cleaning Steps:

1. Verify that the dates are imported correctly.
2. Remove all columns that don't change. **Justification**: There is no variance, it cannot help us to determine the reason for default. We can save memory and analysis, plotting and data frame transformations are faster
3. Identify all columns that don't provide any value. **E.g.** the 'url' column: It provides us a link to access MORE data, but we doesn't have any username and password to access the data.
4. Remove all columns that need Text Processing. We are not going to perform any NLP/ NLU and therefore such columns are not very useful
5. Identify all redundant columns and remove them. **E.g.** The purpose of loan is a drop down which is already a categorical variable. We don't need the title column as it becomes redundant.
6. Identify all columns that don't have any other value other than NA and 0. Remove such columns.
7. Since we are going to identify defaulter status based on Employer Name, we better scan through the 'emp_title' column and consolidate all employer names. **E.g.** ARMY and US ARMY are same. Walmart, WALMART, Walmart and WALMART are same.
8. Few columns such as Rate of Interest have been read as characters due to the presence of the "%" symbol. Convert all such instances to numeric
9. Since we are dealing with the aggregate, we may not need the primary keys in this analysis such as 'id' and 'member_id'. We can remove these as well.

Analysis

Uni-varate Analysis

1. For Categorical Variables
2. For Numeric Variables
3. For Segmented Variables

Bi-variate Analysis

1. By keeping 'loan_status' fixed in one of the columns
2. Scatter plots

Understanding the Data

The company has come across some important attributes in order to understand the behavior of their approved loan customers w.r.t. the loan defaulters.

Thus, the lending company has decided to work only on these variables to mitigate the future risk. The driver variables considered for this case study are:

annual_inc	-	Annual Income of applicant
loan_amnt	-	The listed amount of the loan applied for by the borrower
funded_amnt	-	The total amount committed to that loan at that point in time
int_rate	-	Interest Rate on the loan
grade	-	LC assigned loan grade
dti	-	Debt to income ratio
emp_length	-	Employment length in years
purpose	-	A category provided by the borrower for the loan request.
home_ownership	-	The home ownership status provided by the borrower during registration
loan_status	-	Current status of the loan

Conclusion – Top Observations

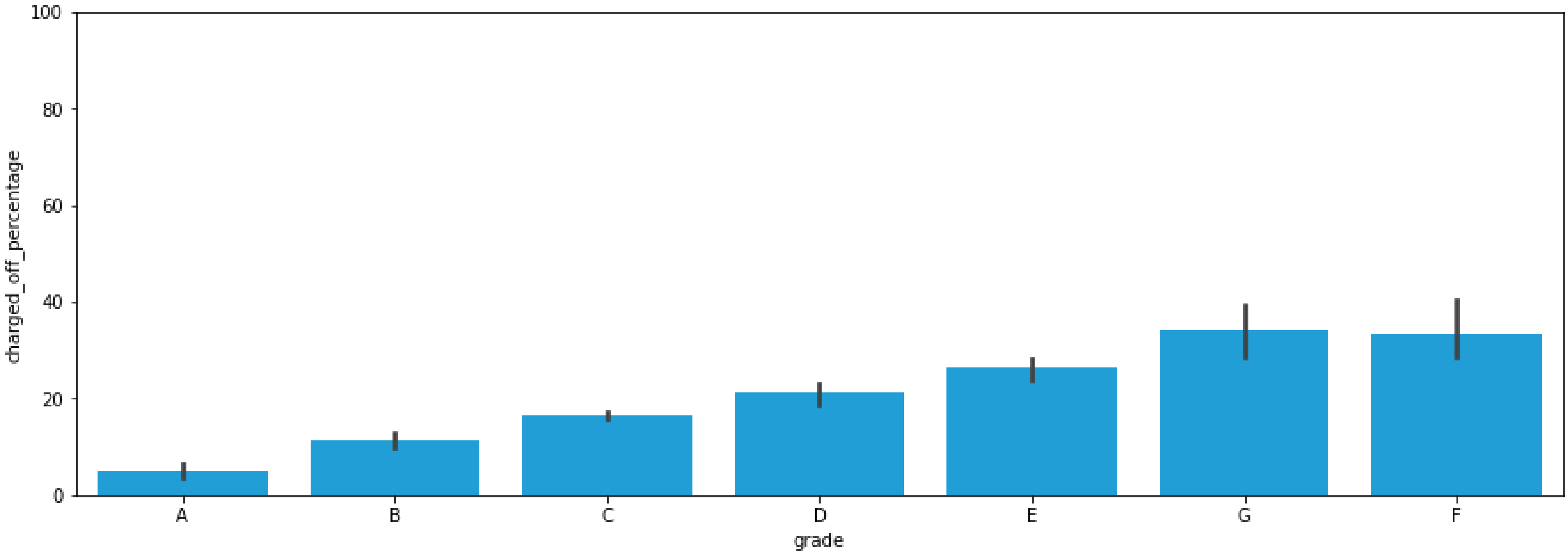
Below are the Final Inferences from our analysis

1. **grade & sub_grade** (LC assigned loan grade and loan subgrade.)
   - HIGHER (alphabatically higher i.e. B>A) grade & sub_grade relates to more likely to default on loans.
2. **home_ownership** (The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.)
   - While home_ownership does not have significant impact, but OTHER status relates to more likely to default on loans.
3. **total_rec_prncp** (Principal received to date)
   - LESSER total_rec_prncp relates to more likely to default on loans.
4. **total_pymnt_inv** (Payments received to date for portion of total amount funded by investors)
   - LESSER total_pymnt_inv relates to more likely to default on loans.
5. **total_pymnt** (Payments received to date for total amount funded)
   - LESSER total_pymnt relates to more likely to default on loans.
6. **last_pymnt_amnt** (Last month payment was received.)
   - LESSER last_pymnt_amnt relates to more likely to default on loans.
7. **int_rate** (Interest Rate on the loan)
   - HIGHER int_rate relates to more likely to default on loans.

Conclusion – Top Observations

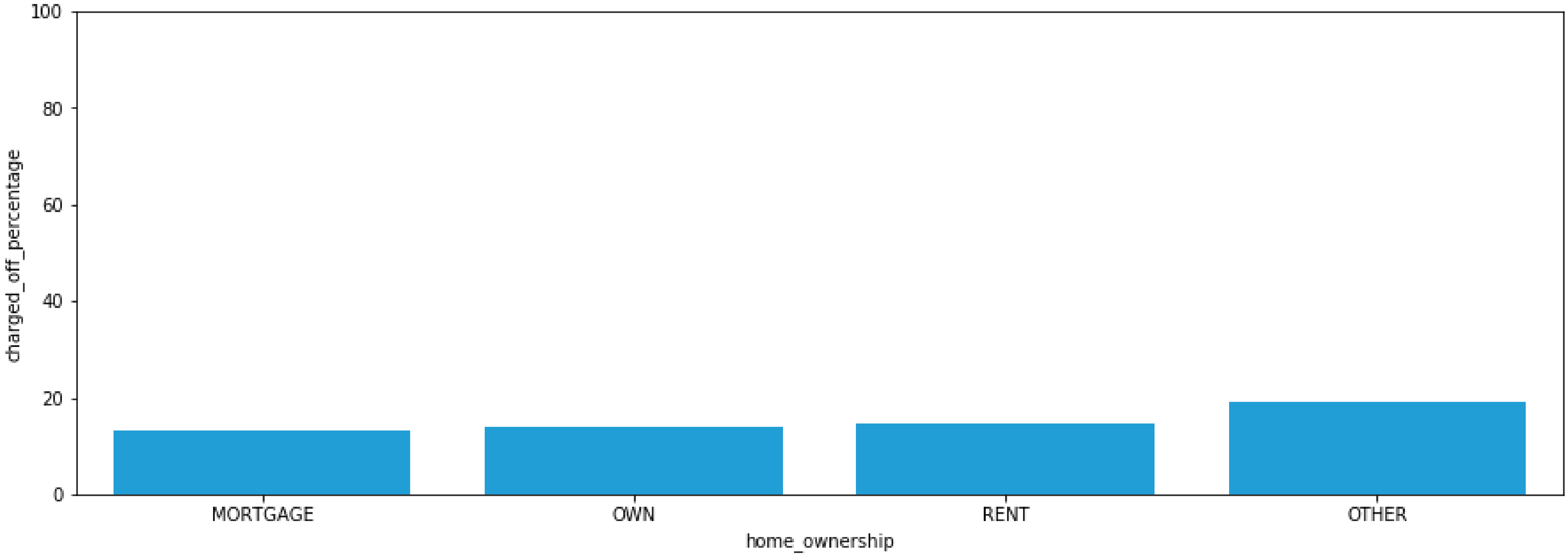
8. **revol_util** (Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.)
   - HIGHER revol_util relates to more likely to default on loans.
9. **inq_last_6mths** (The number of inquiries in past 6 months (excluding auto and mortgage inquiries).)
   - HIGHER inq_last_6mths relates to more likely to default on loans.
10. **loan_amnt** (The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.)
    - HIGHER loan_amnt relates to more likely to default on loans.
11. **dti** (A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.)
    - HIGHER dti relates to more likely to default on loans.
12. **annual_inc** (The self-reported annual income provided by the borrower during registration.)
    - LESSER annual income relates to more likely to default on loans.
13. **total_acc** (The total number of credit lines currently in the borrower's credit file)
    - LESSER total_acc relates to more likely to default on loans.

grade & sub_grade



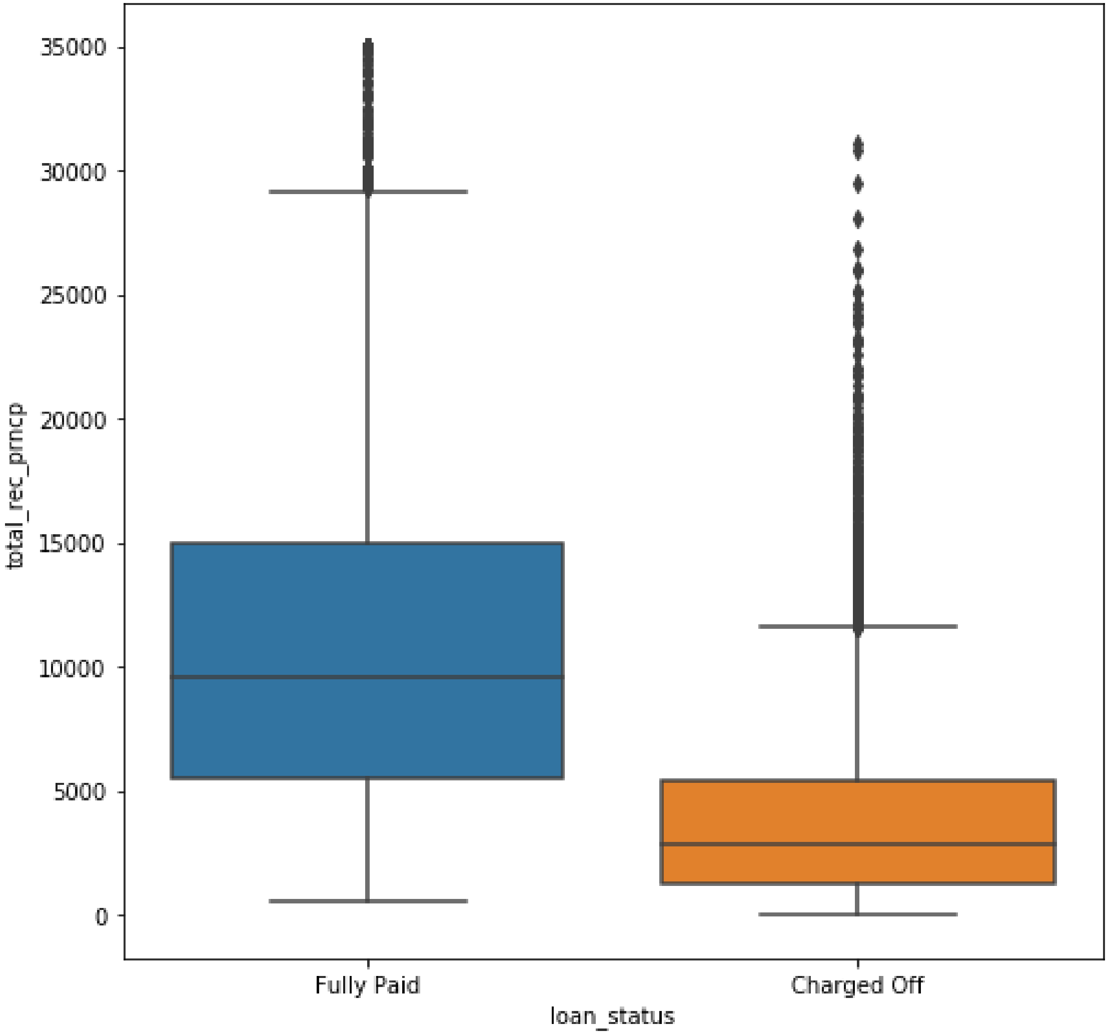
HIGHER (alphabatically higher i.e. B>A) **grade** & **sub_grade** relates to MORE likely to default on loans.

home_ownership



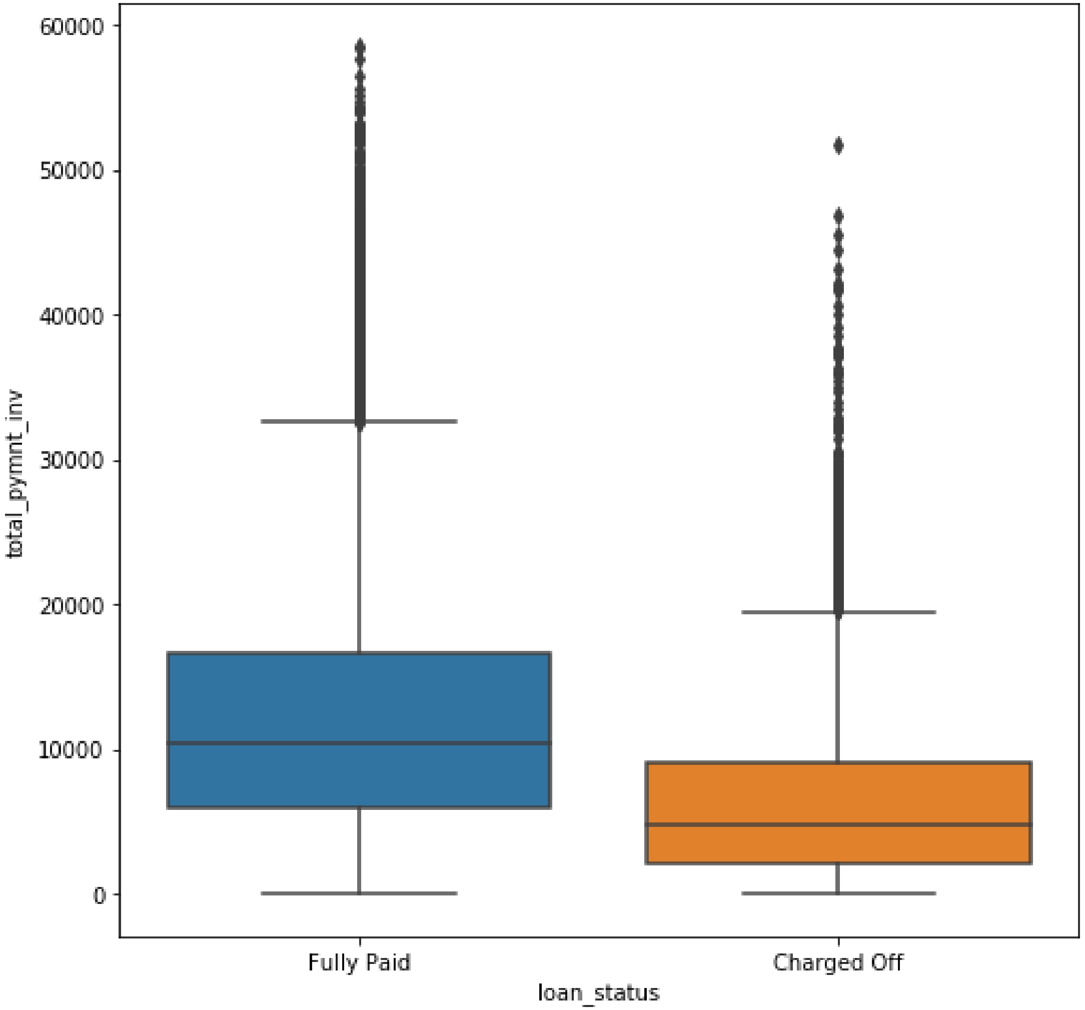
While **home_ownership** does not have significant impact, but OTHER status relates to more likely to default on loans.

total_rec_prncp



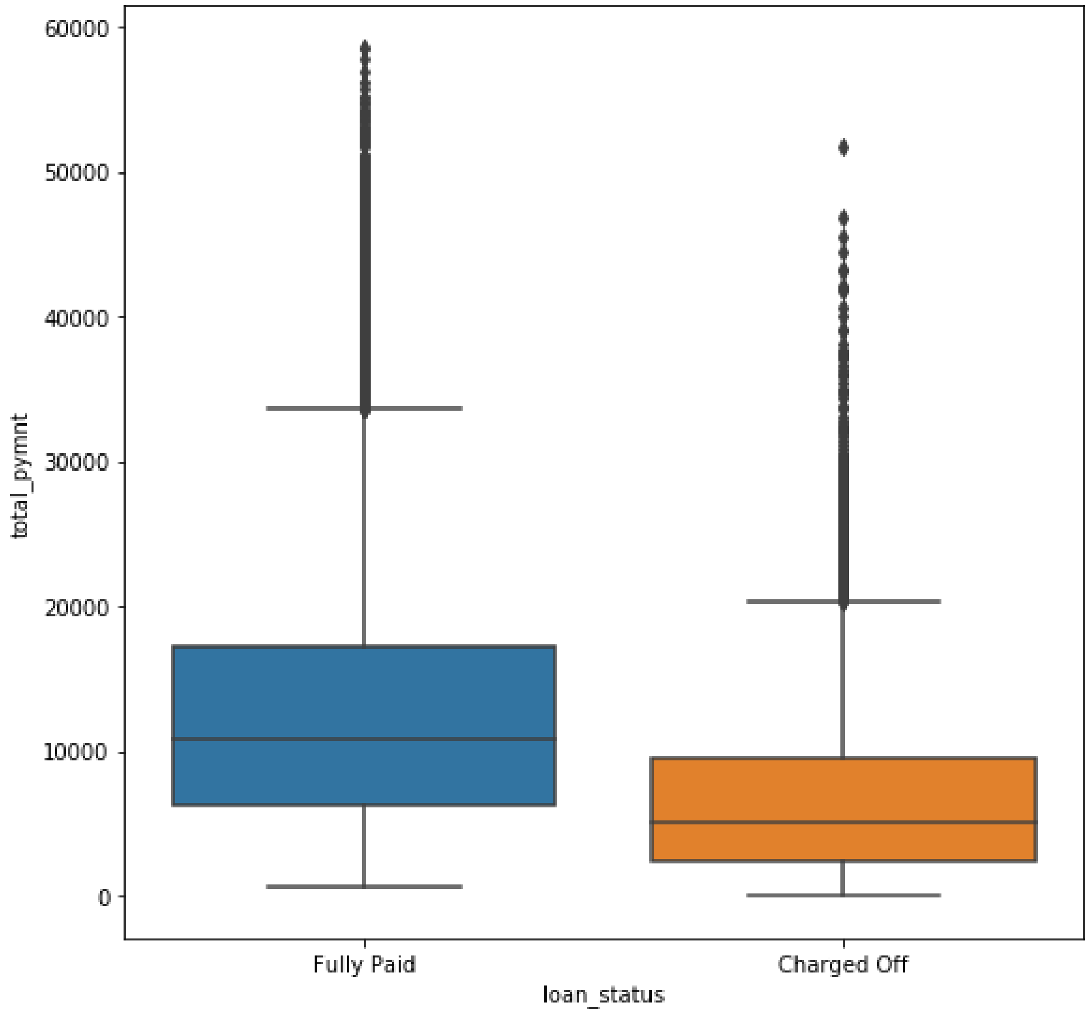
LESSER **total_rec_prncp** relates to MORE likely to default on loans.

total_pymnt_inv



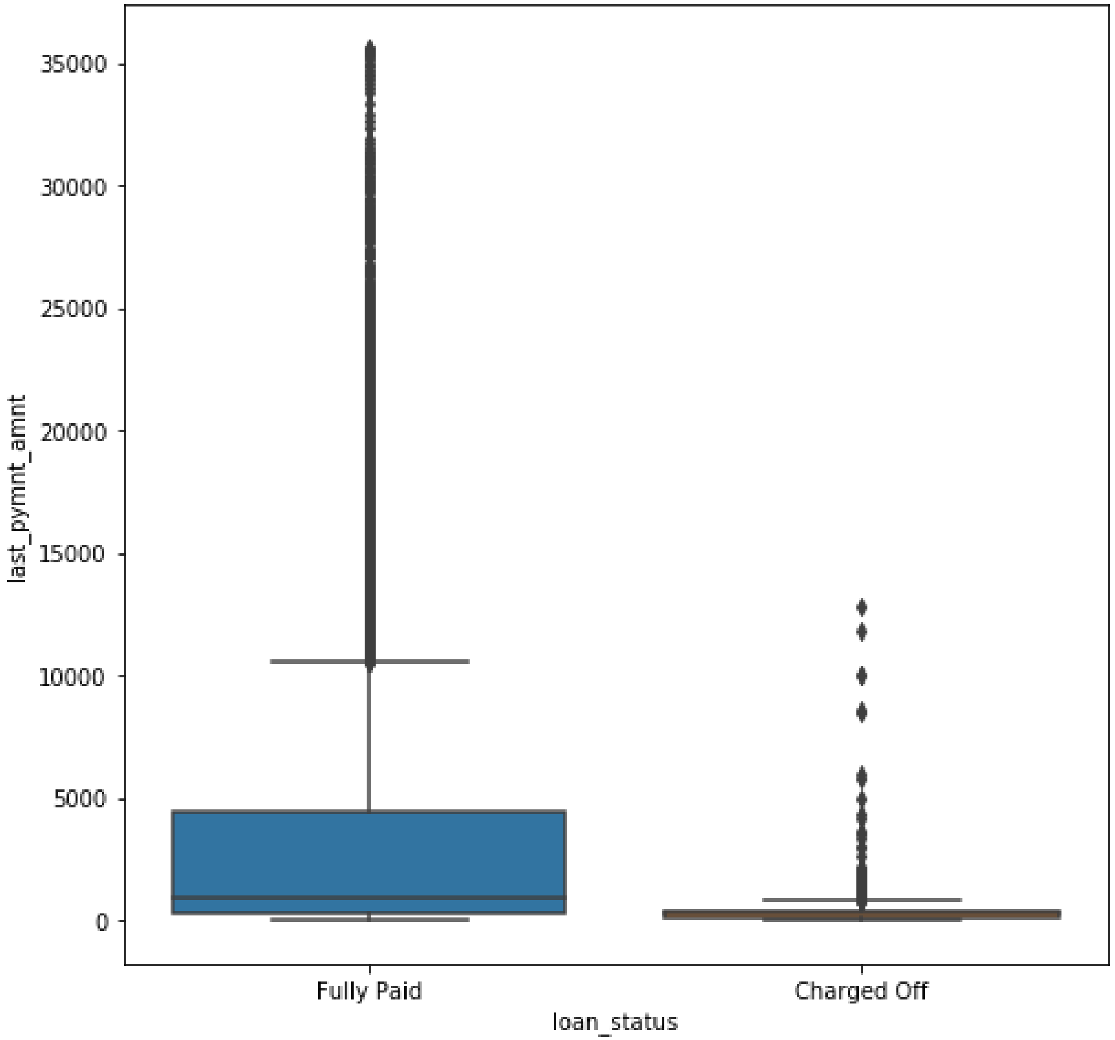
LESSER **total_pymnt_inv** relates to MORE likely to default on loans.

total_pymnt



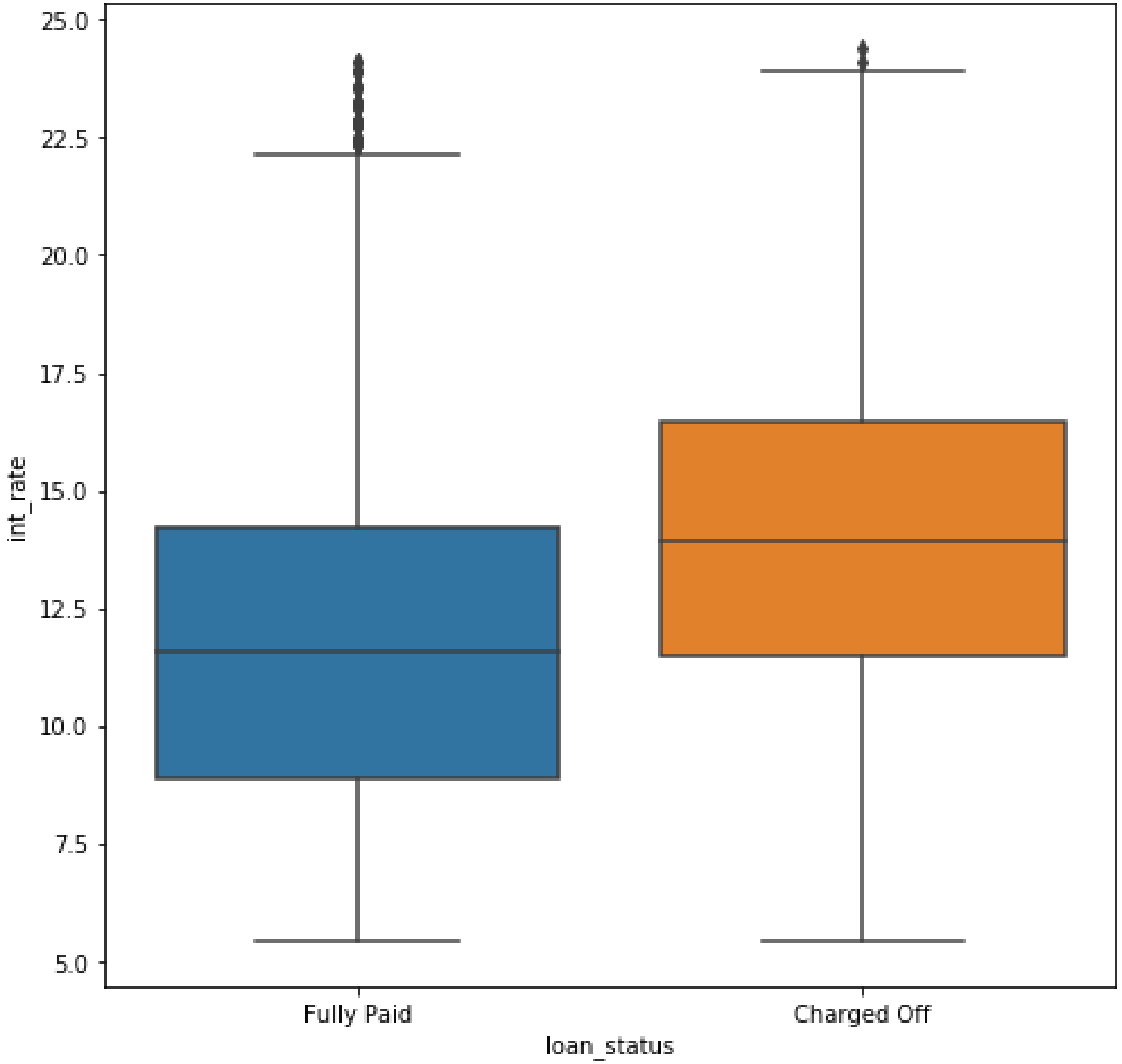
LESSER **total_pymnt** relates to MORE likely to default on loans.

last_pymnt_amnt



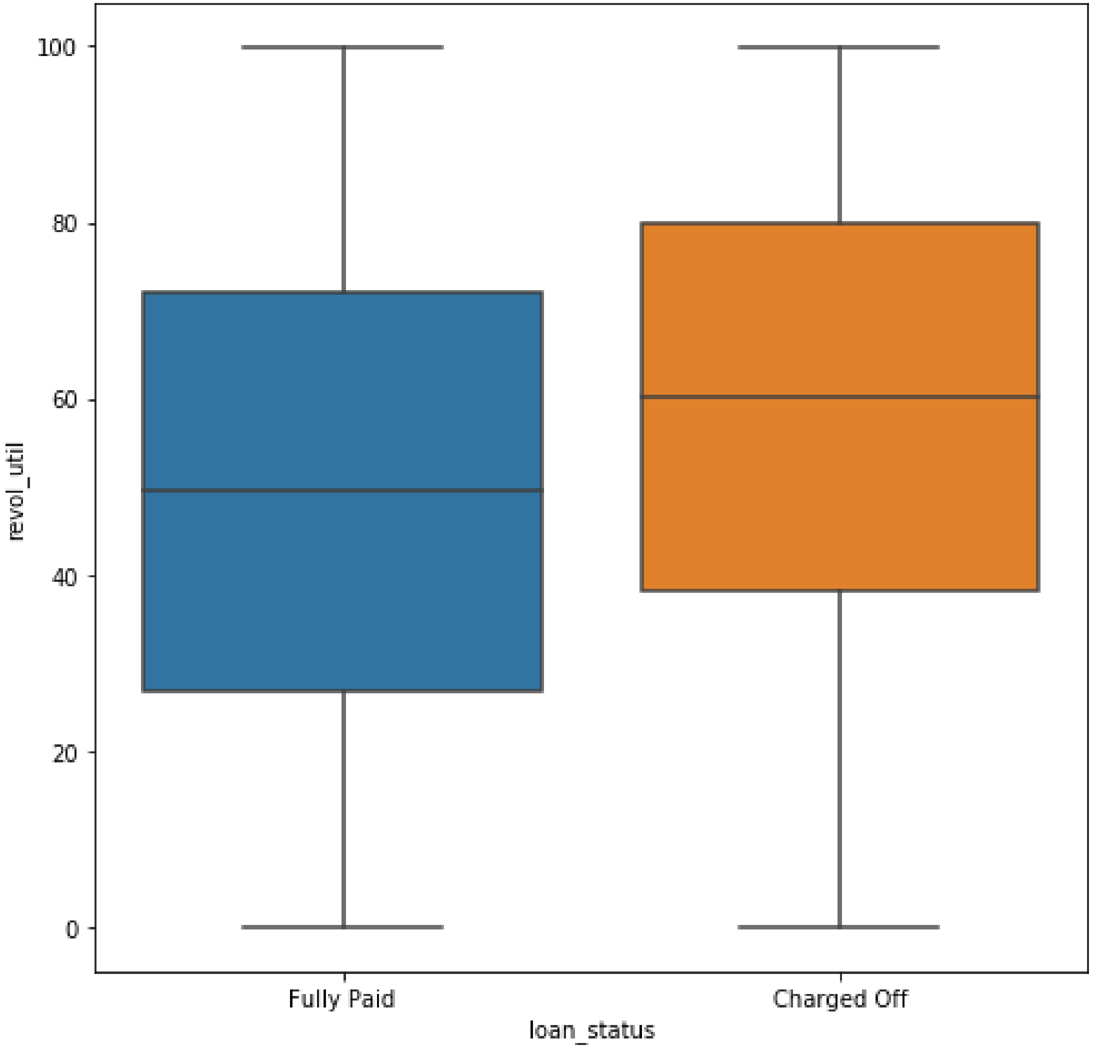
LESSER **last_pymnt_amnt** relates to MORE likely to default on loans.

int_rate



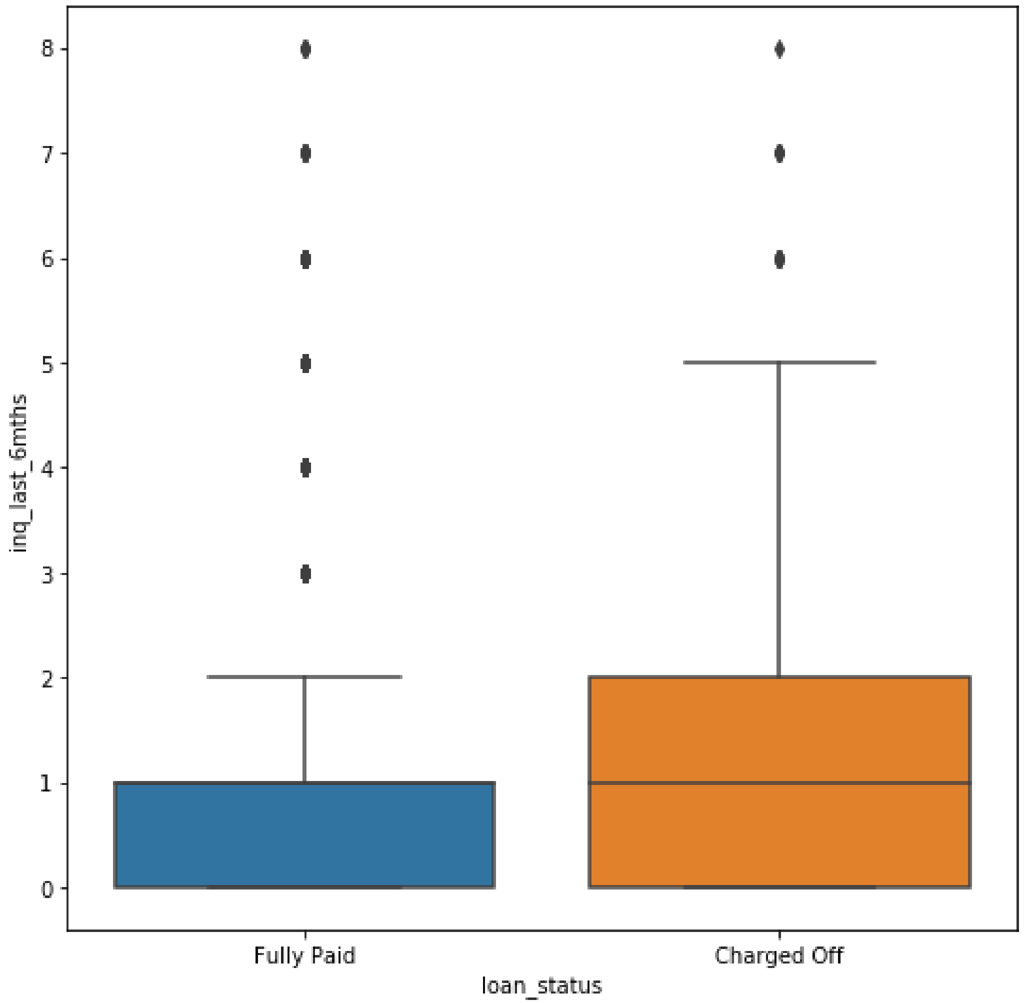
HIGHER **int_rate** relates to MORE likely to default on loans.

revol_util



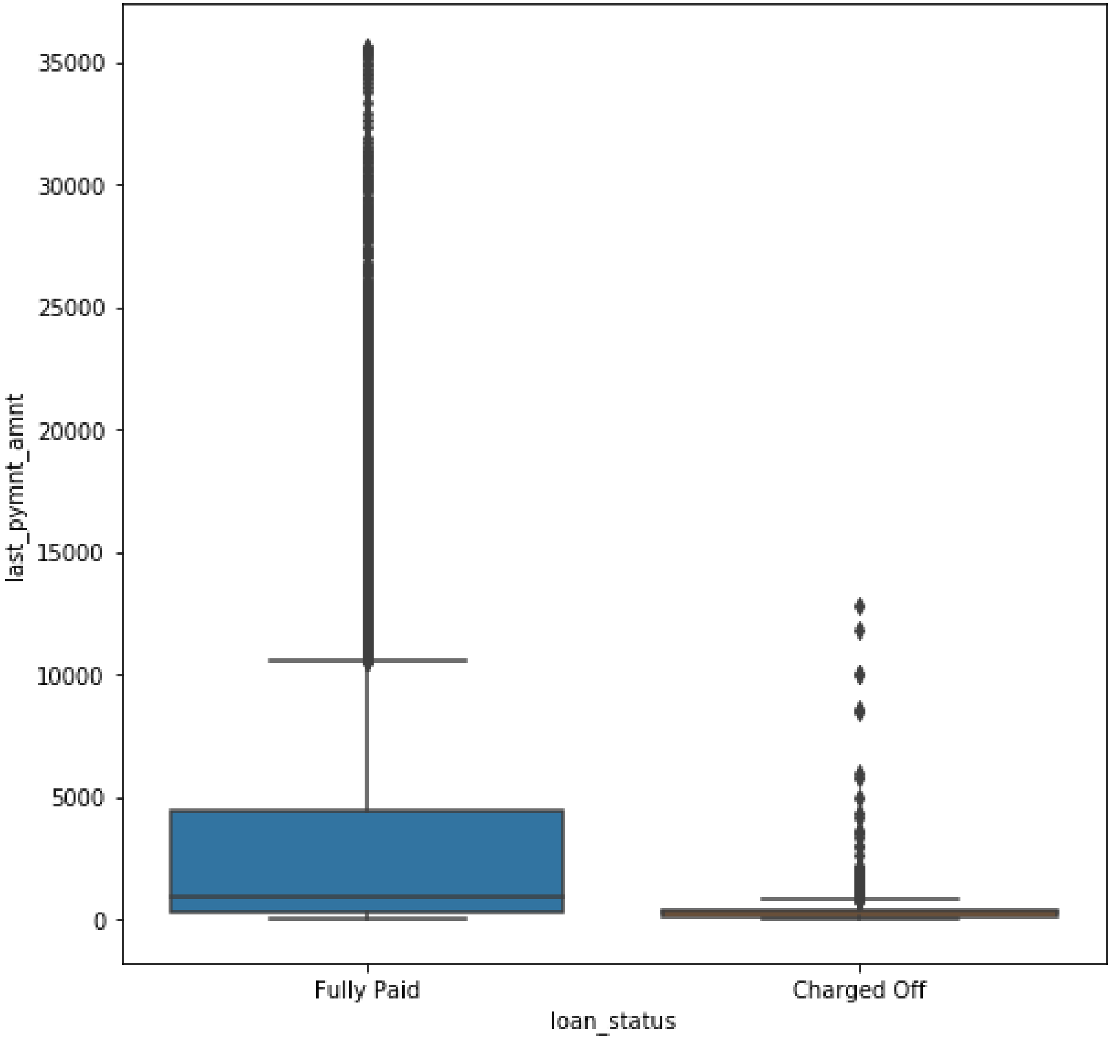
HIGHER **revol_util** relates to MORE likely to default on loans.

inq_last_6mths



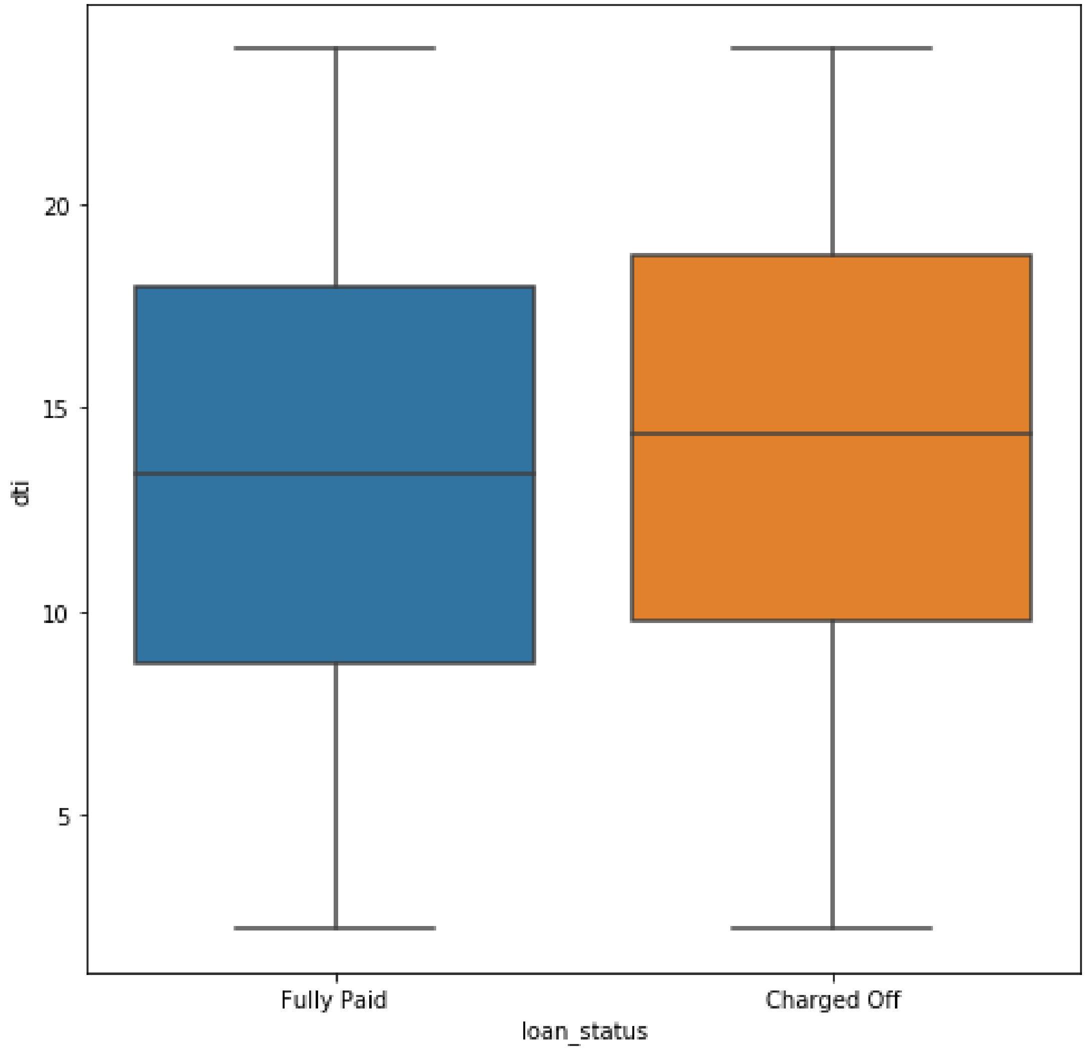
HIGHER **inq_last_6mths** relates to MORE likely to default on loans.

loan_amnt



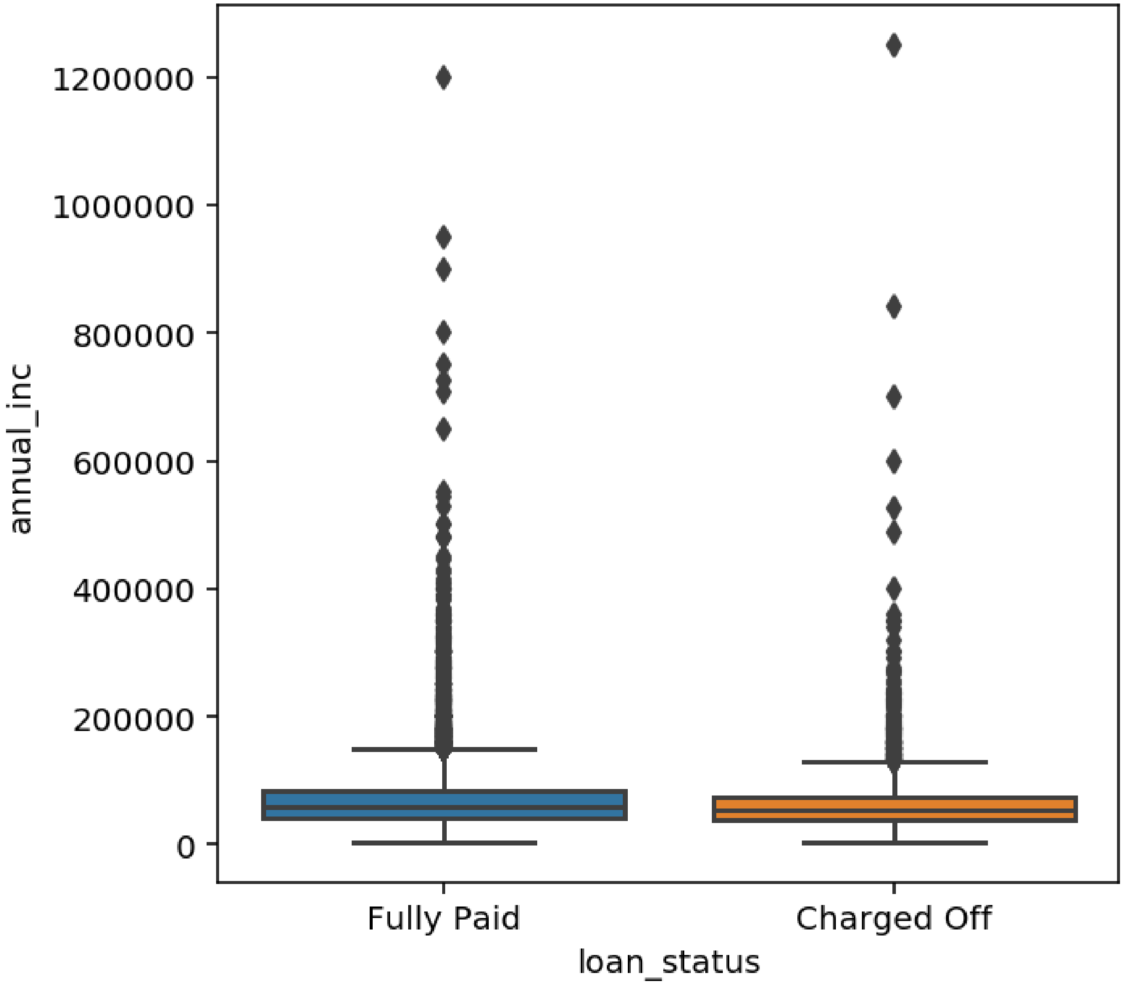
HIGHER **loan_amnt** relates to MORE likely to default on loans.

dti



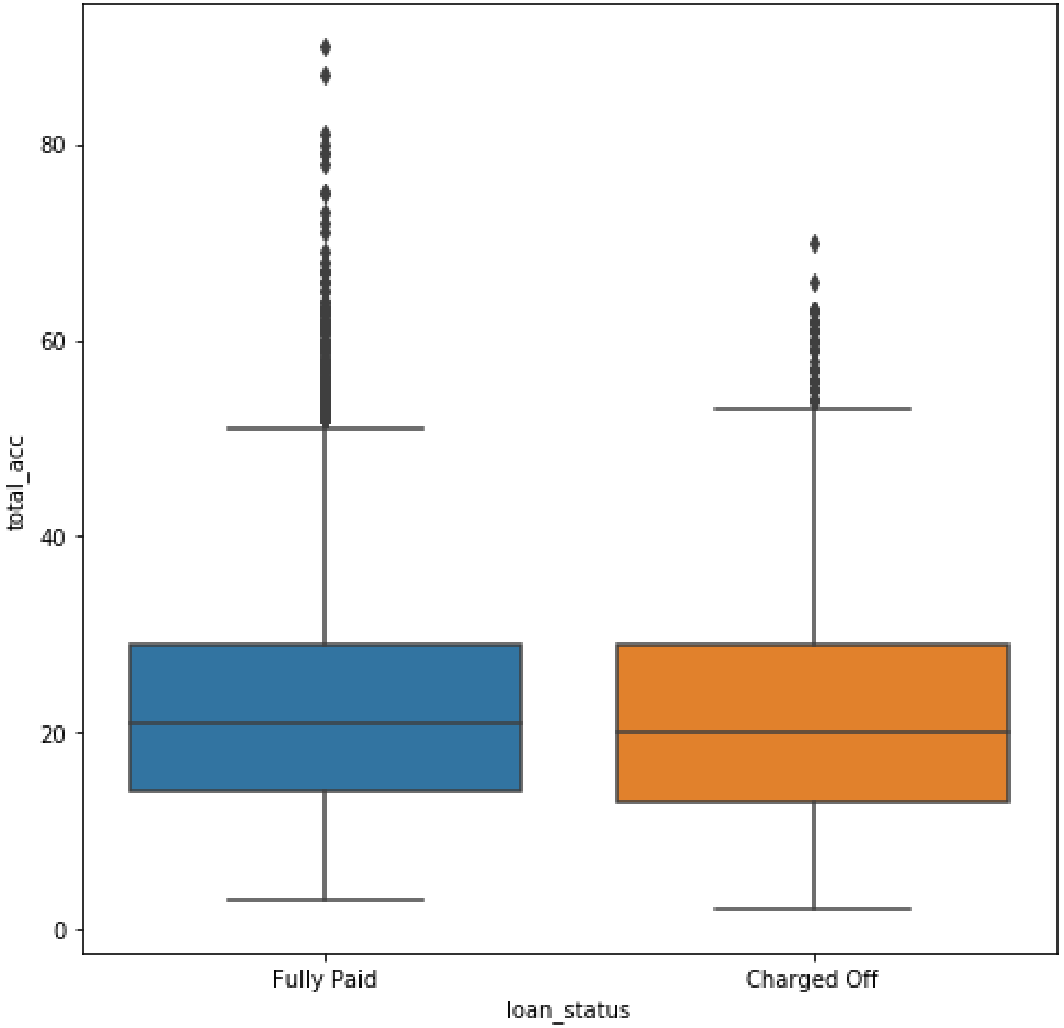
HIGHER **dti** relates to MORE likely to default on loans.

annual_inc



LESSER **annual income** relates to MORE likely to default on loans.

total_acc



LESSER **total_acc** relates to MORE likely to default on loans.

Thank you!