Lending Club Case Study

Siddharth Ghule

Problem Statement

- Background:
  - We work for a consumer finance company which specialises in lending various types of loans to urban customers. This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface. The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders.
- Aim:
  - Identify risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

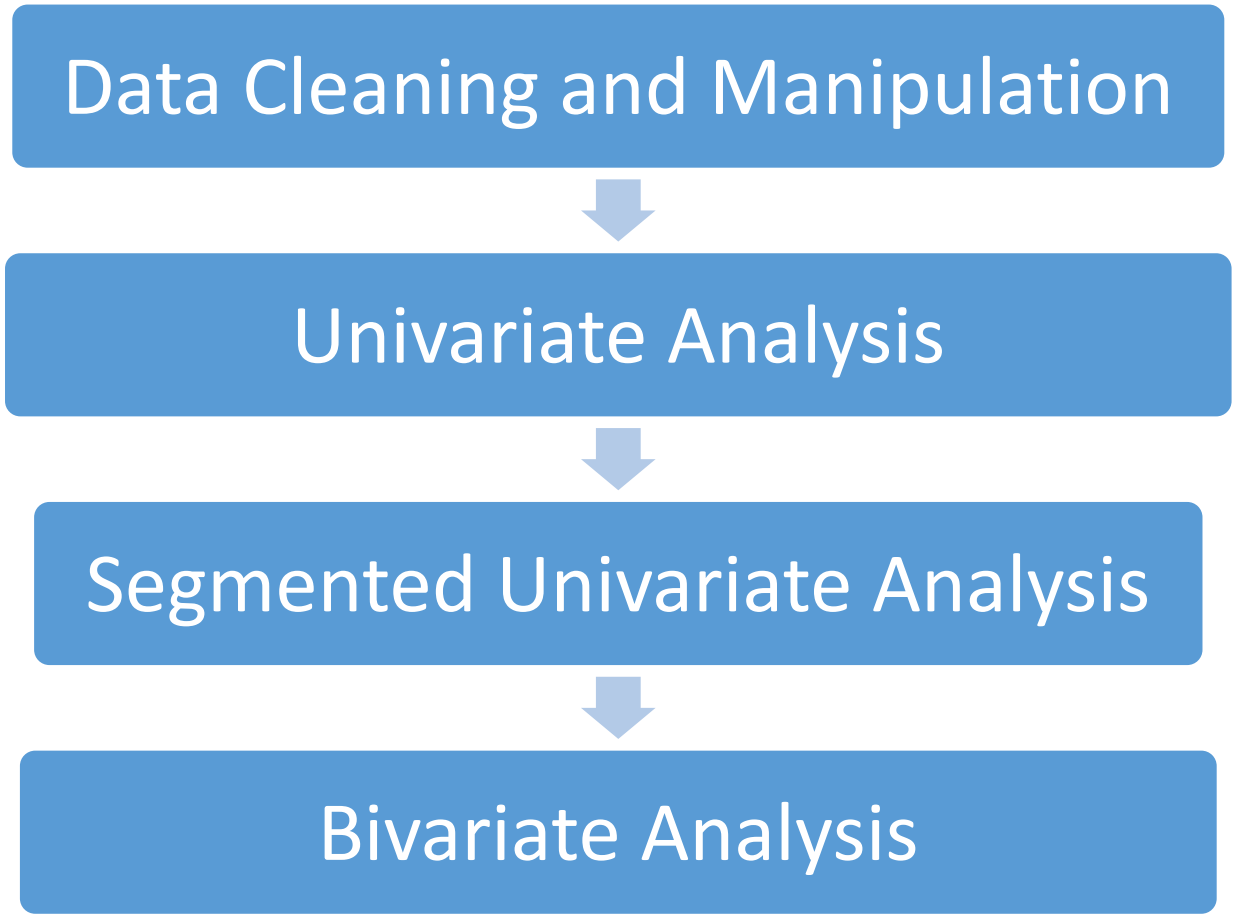
Data

- We are given loan data for all loans issued through the time period 2007 t0 2011
- Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not provided.
- There are 39,717 data points and 111 variables.
- loan_status is the target variable with following possible values:
  - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
  - **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'.
  - **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

Data Issues

- There are a few rows and columns with null/missing values
- Some columns have identical values for all the rows
- There variables with incorrrect datatypes (eg. last_pymnt_d its string but it should be datetime)
- Target variables have irrelevent value (eg. loan_status = Current)
- Many variables are highly correlated

Approach


Data Cleaning and Manipulation
Univariate Analysis
Segmented Univariate Analysis
Bivariate Analysis

Data Cleaning and Manipulation

- **Drop Unnecessary Columns:**
  Columns such as 'id','member_id','url','zip_code' do not contain useful infor mation. Hence, we drop these columns.
- **Remove Identical Columns:**
  As identical value do help in any way we drop these columns.
- **Filter Data:**
  In this anlaysis, we are only interested in customers who have either fully p aid their loan or defaulted.
  Hence, we drop customers who are currently paying installments.
- **Fix Incorrect Data Types:** We remove symbols such as ‘%’ from numerical variables and convert them to either float or int. We also convert date variables to datetime type variables.

Data Cleaning and Manipulation

- **Remove Highly Correlated Variables:** Even after removing unnecessary variables we are still left with large number of them (>30).
  The correlated variables contain similar information, there analysis is doesn't provide more information. Therefore, we remove highly correlated variables using correlation analysis.
- **Derived Metrics:** we create new metrics from issue_d and last_pymnt_d. We extract month and year from these variables.

Univariate Analysis

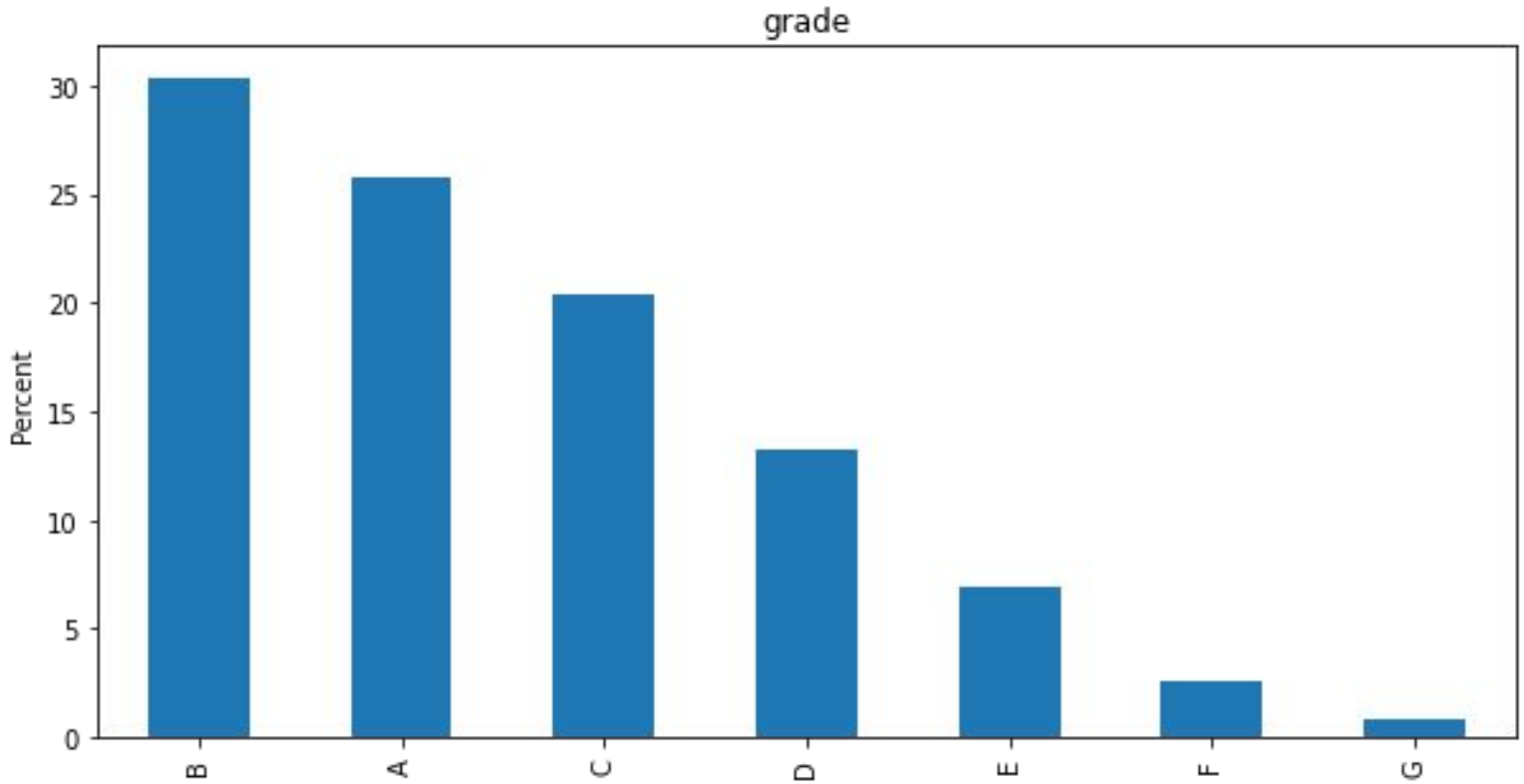
We perform univariate analysis on categorical and continuous variables. Here, we briefly describe the protocol of univariate analysis:

- **Protocol for categorical variables:** We plot the distribution of categorical variable as histogram. We have used percentage of data points instead of counts/frequency as percentages are easy to comprehend.
- **Protocol for categorical variables:** We compute different metrics (eg. mean, 25th quantile, 50th quantile, etc.). We also plot box plot and distribution using frequency histogram.

Next, we show univariate analysis only for the strong indicators identified in this analysis.

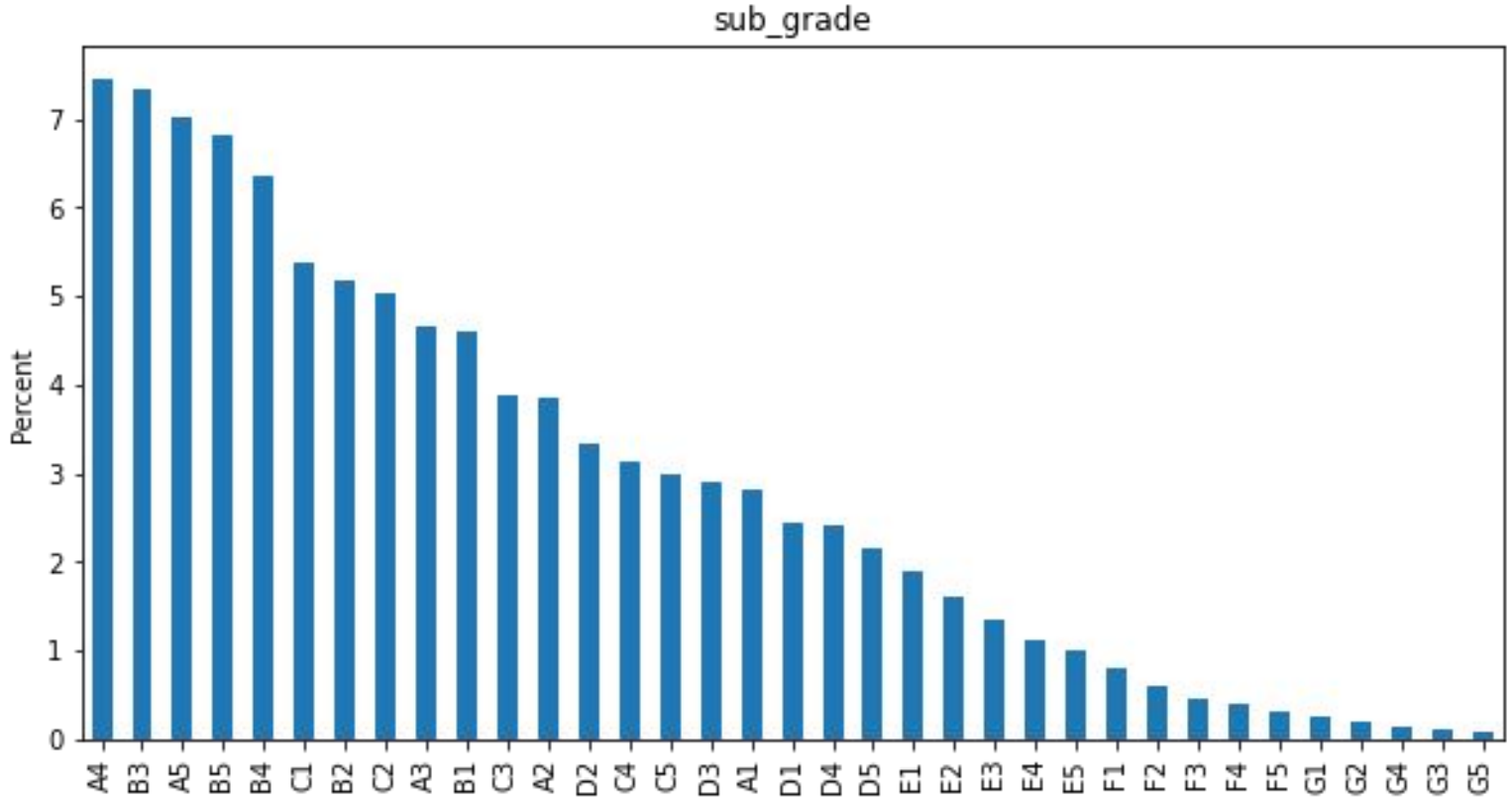
Univariate Analysis (categorical variable)

- **grade:** It can be seen that most borrowers have B grade. And Borrowers with G grade are lowest in number.



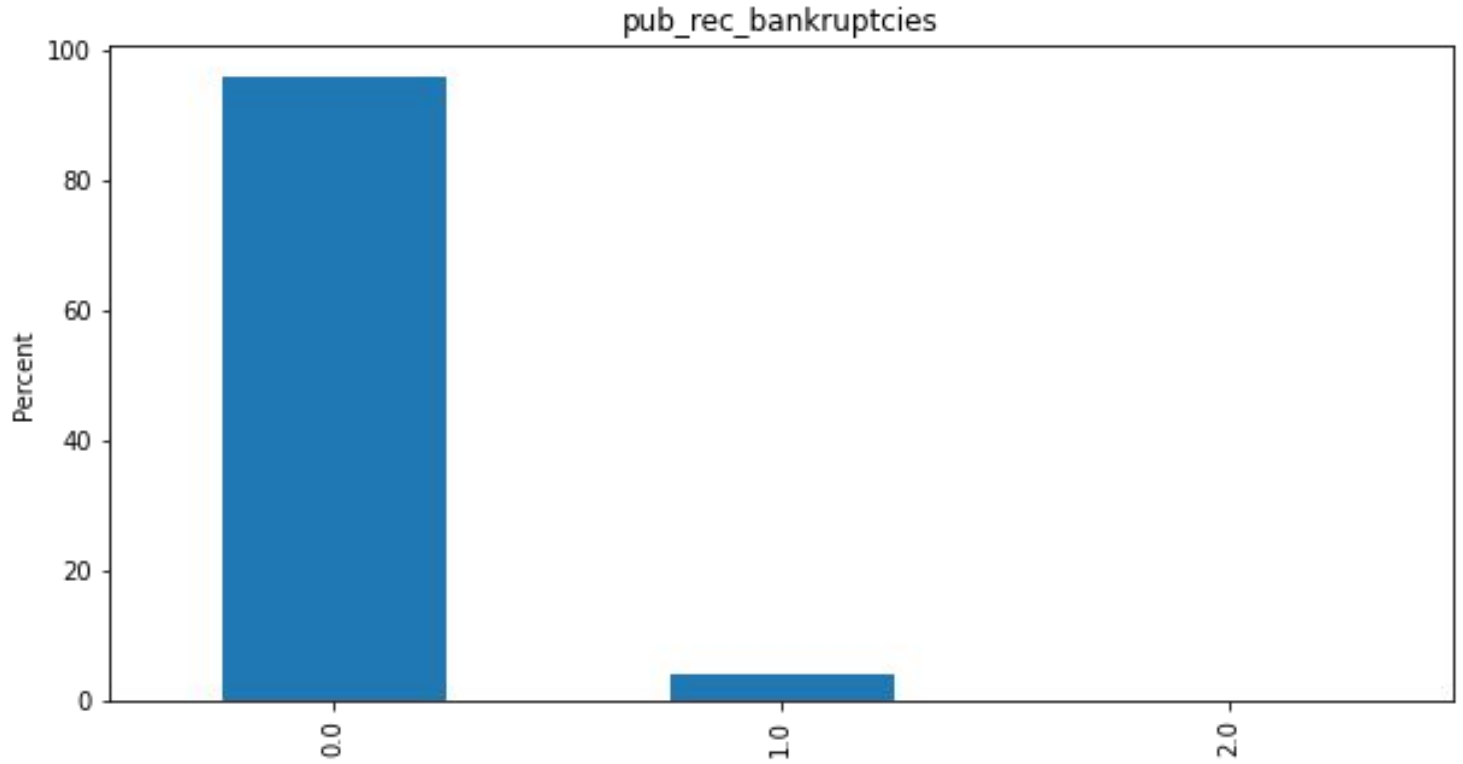
Univariate Analysis (categorical variable)

- **sub_grade:** It can be seen that most borrowers have A4 sub grade. And Borrowers with G5 sub grade are lowest in number.



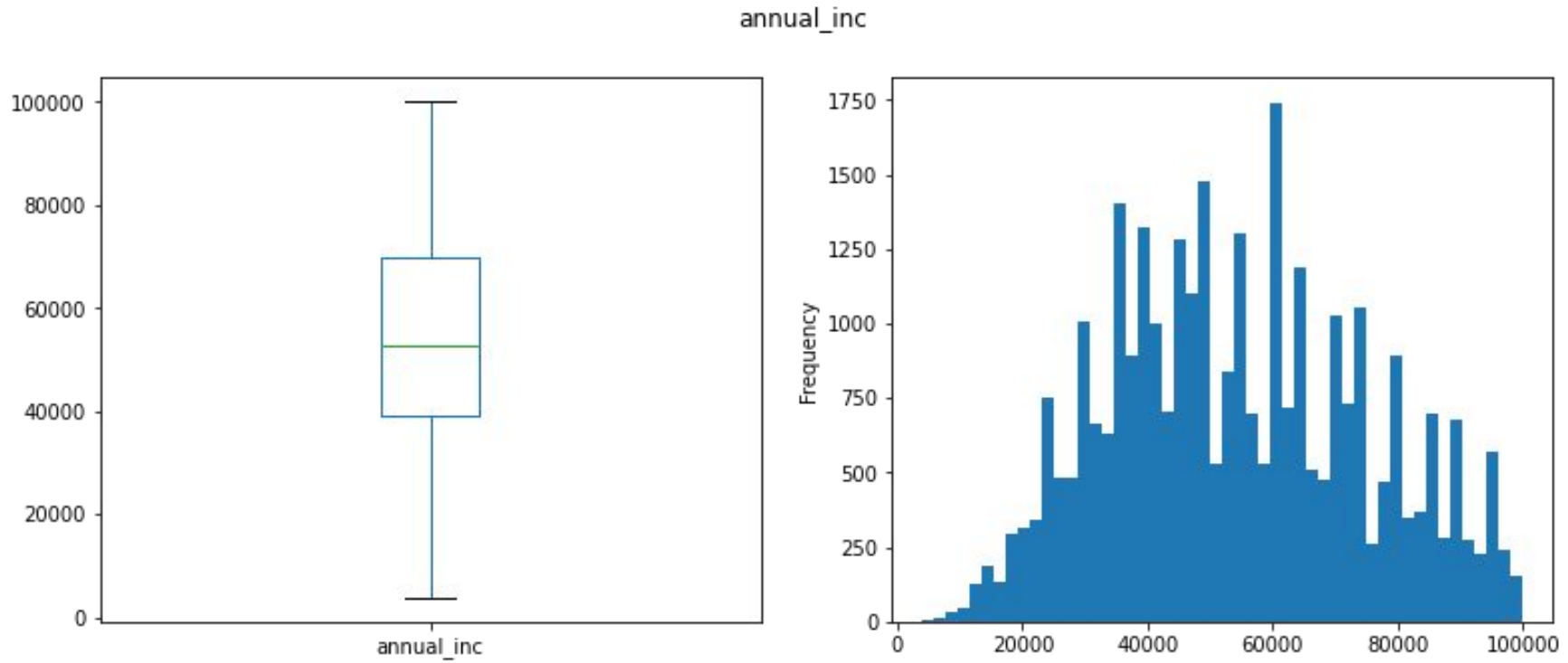
Univariate Analysis (categorical variable)

- **pub_rec_bankruptcies:** It can be seen that most of the borrowers (>90%) have no public records of bankruptcies against them.



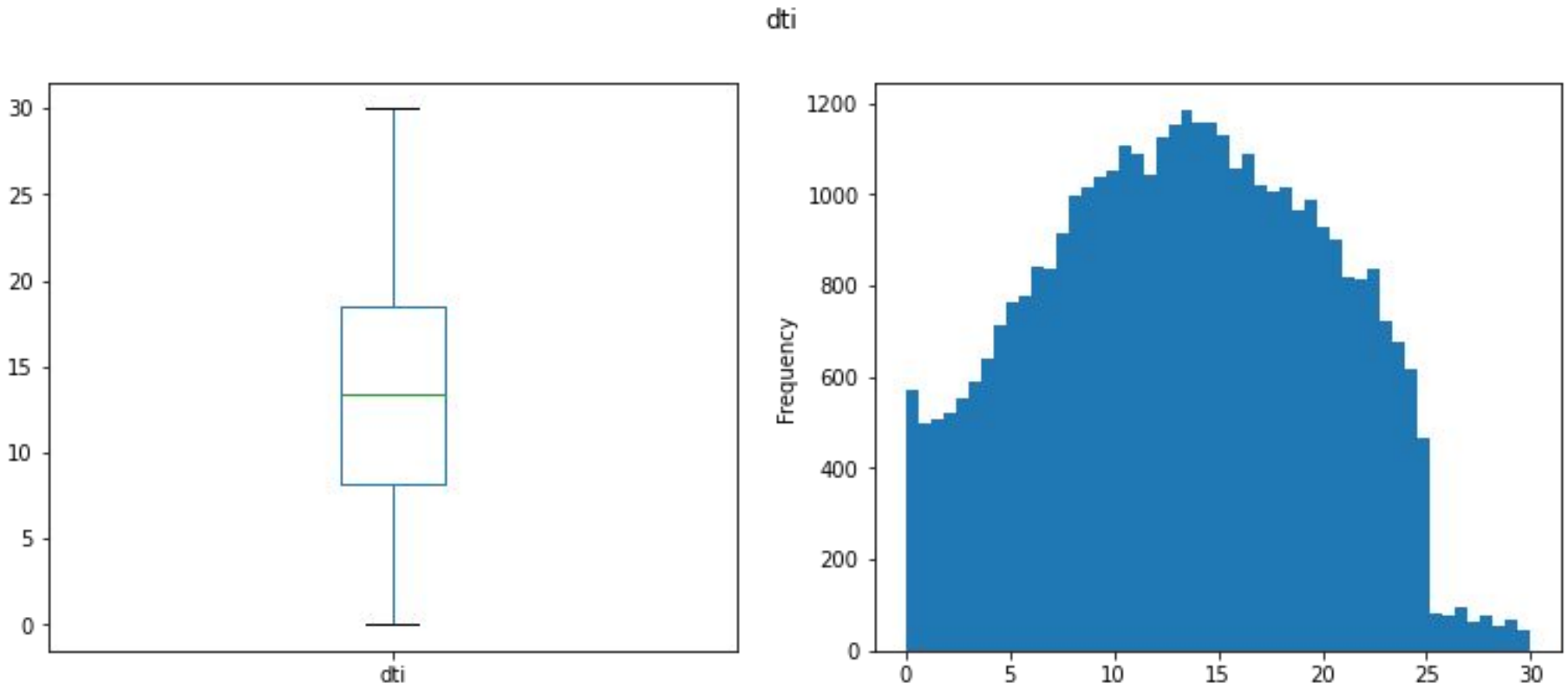
Univariate Analysis (continuous variable)

- **annual_inc:** The average annual income of borrowers is $54499 and median income of $52800.



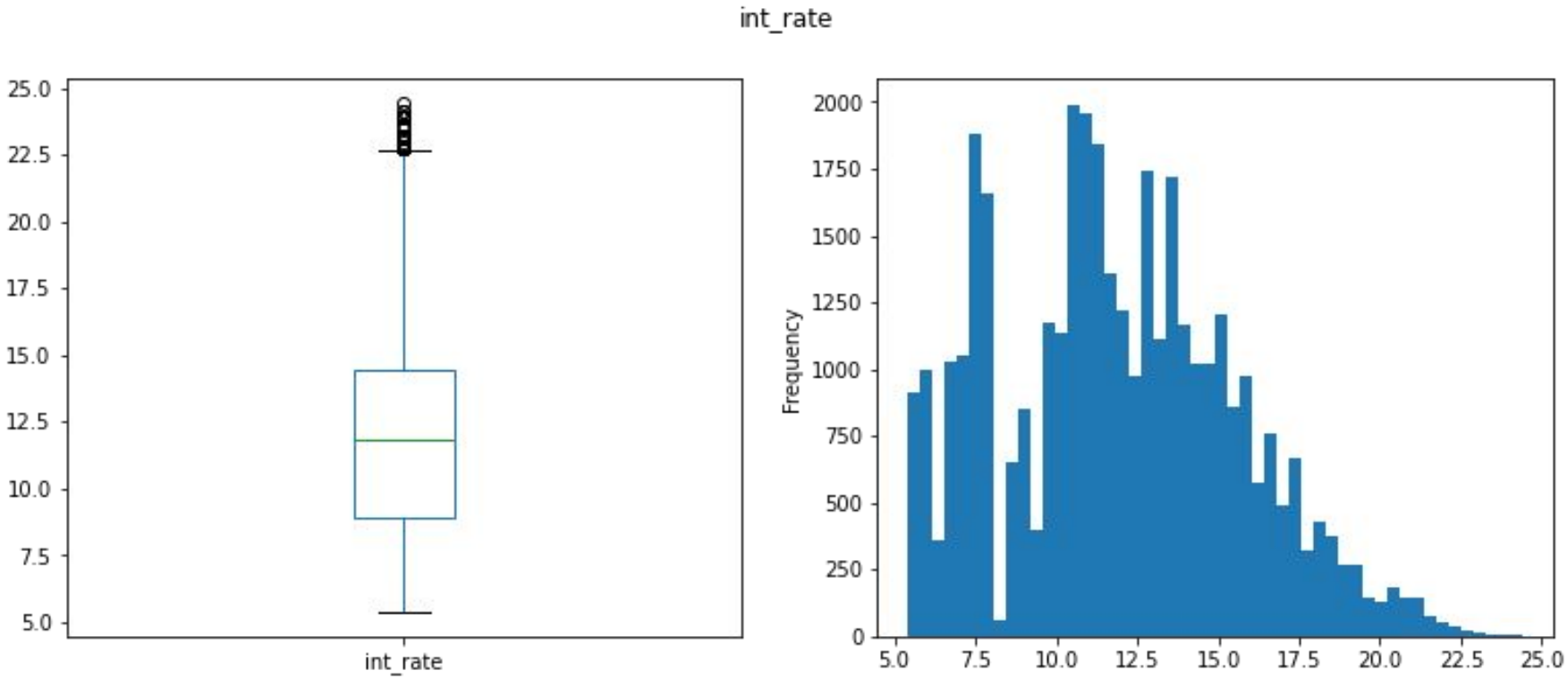
Univariate Analysis (continuous variable)

- **dti:** The average dti of borrowers is 13.29 and median dti is 13.39.



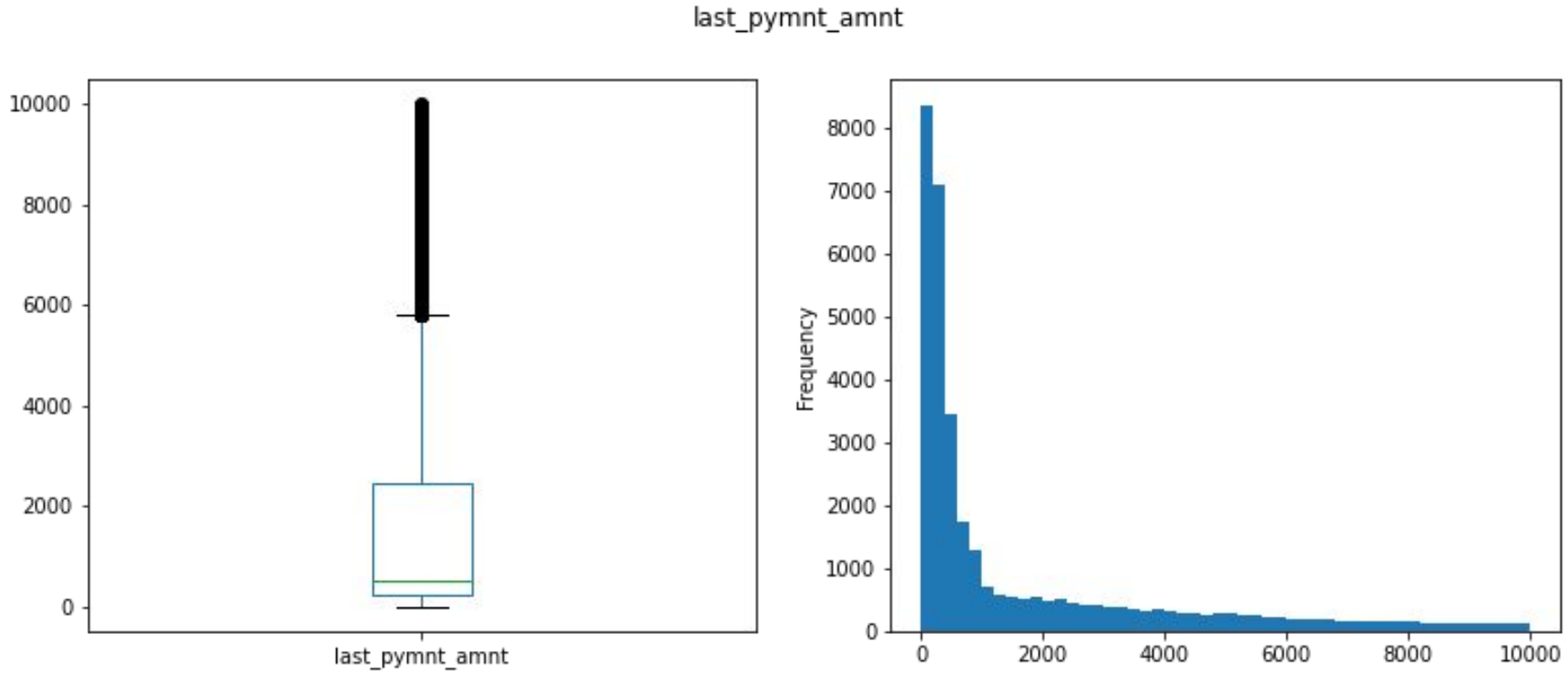
Univariate Analysis (continuous variable)

- **int_rate:** The average interest rate on the loans is 11.96% and median interest rate is 11.83.



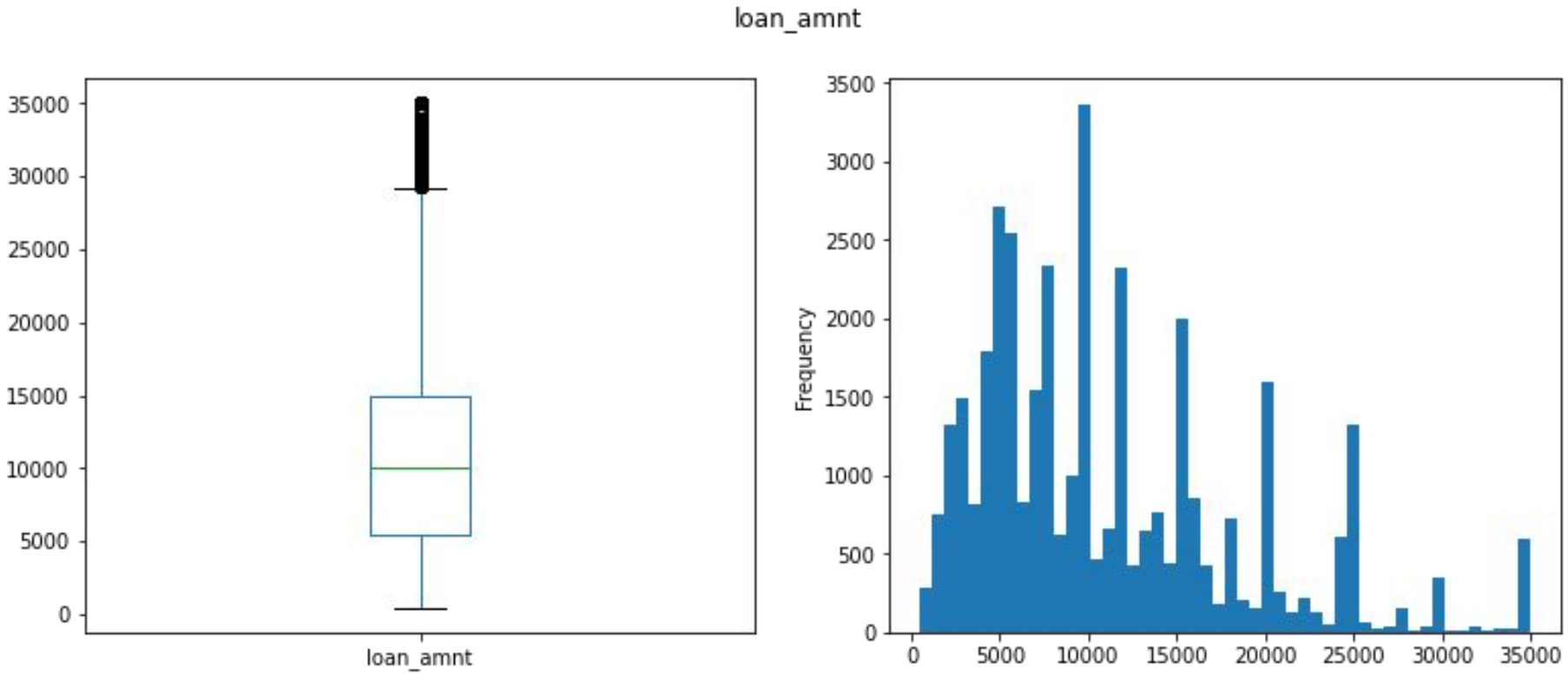
Univariate Analysis (continuous variable)

- **last_pymnt_amnt:** The average amount of last payment received from the borrowers is $1723 and median payment is $491



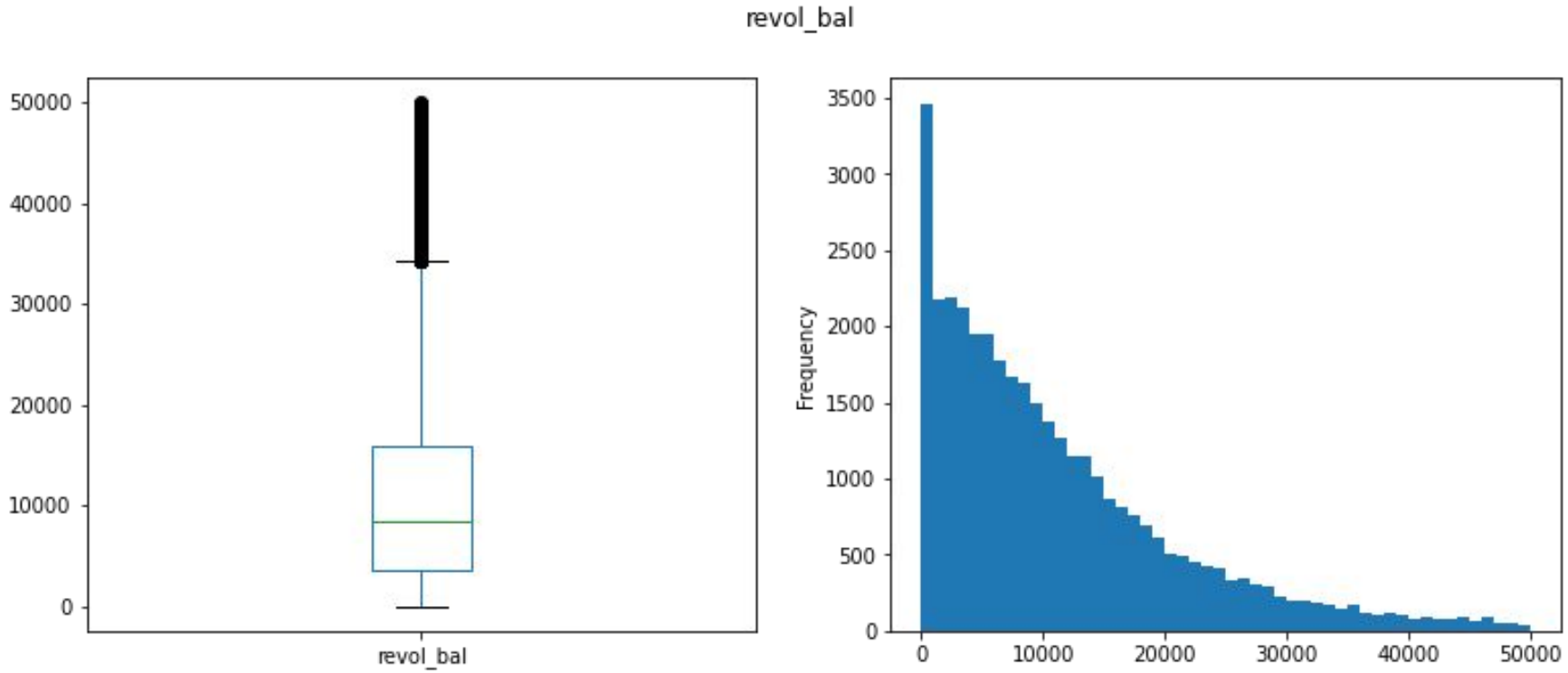
Univariate Analysis (continuous variable)

- **loan_amnt:** The average loan amount is $11131 and median loan amount is $10000.



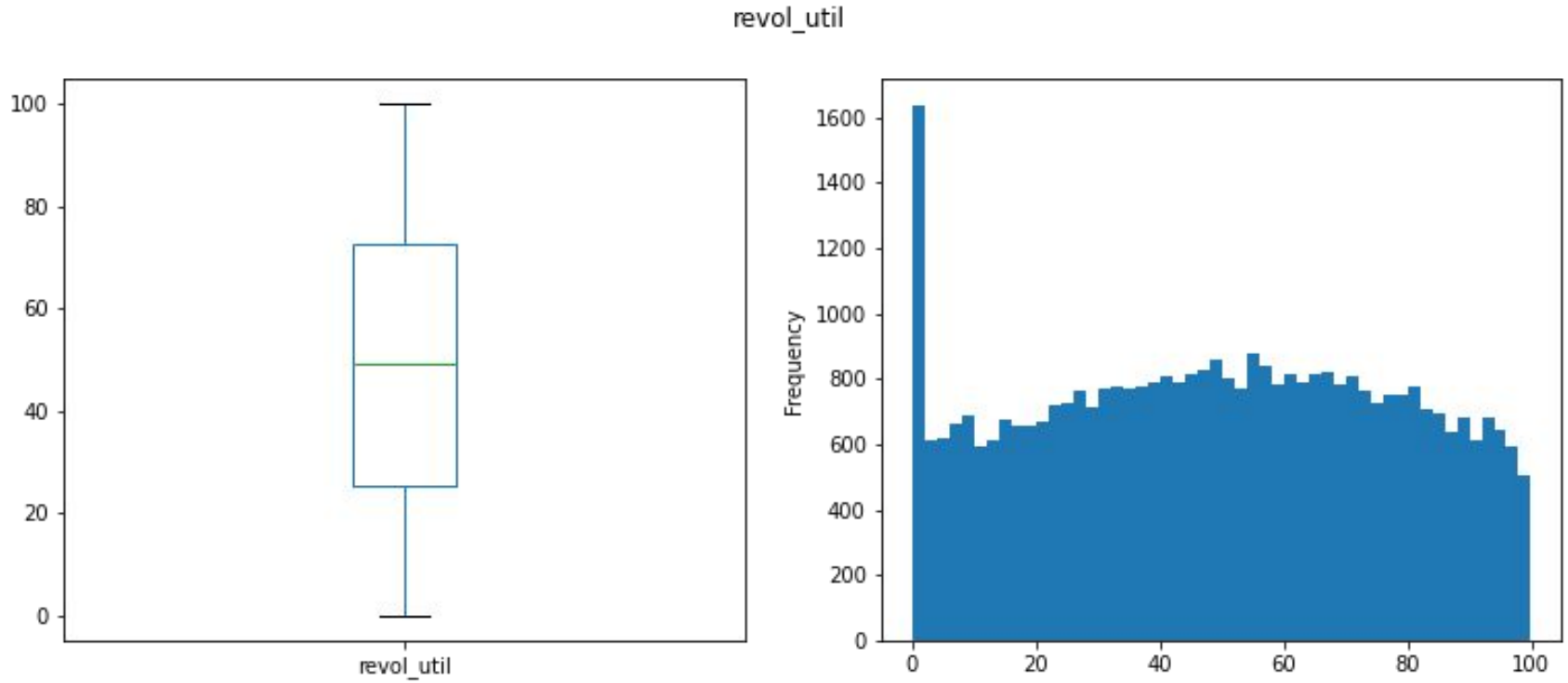
Univariate Analysis (continuous variable)

- **revol_bal:** The average revolving balance of borrowers is $11189 and the median revolving balance of borrowers is $8473.



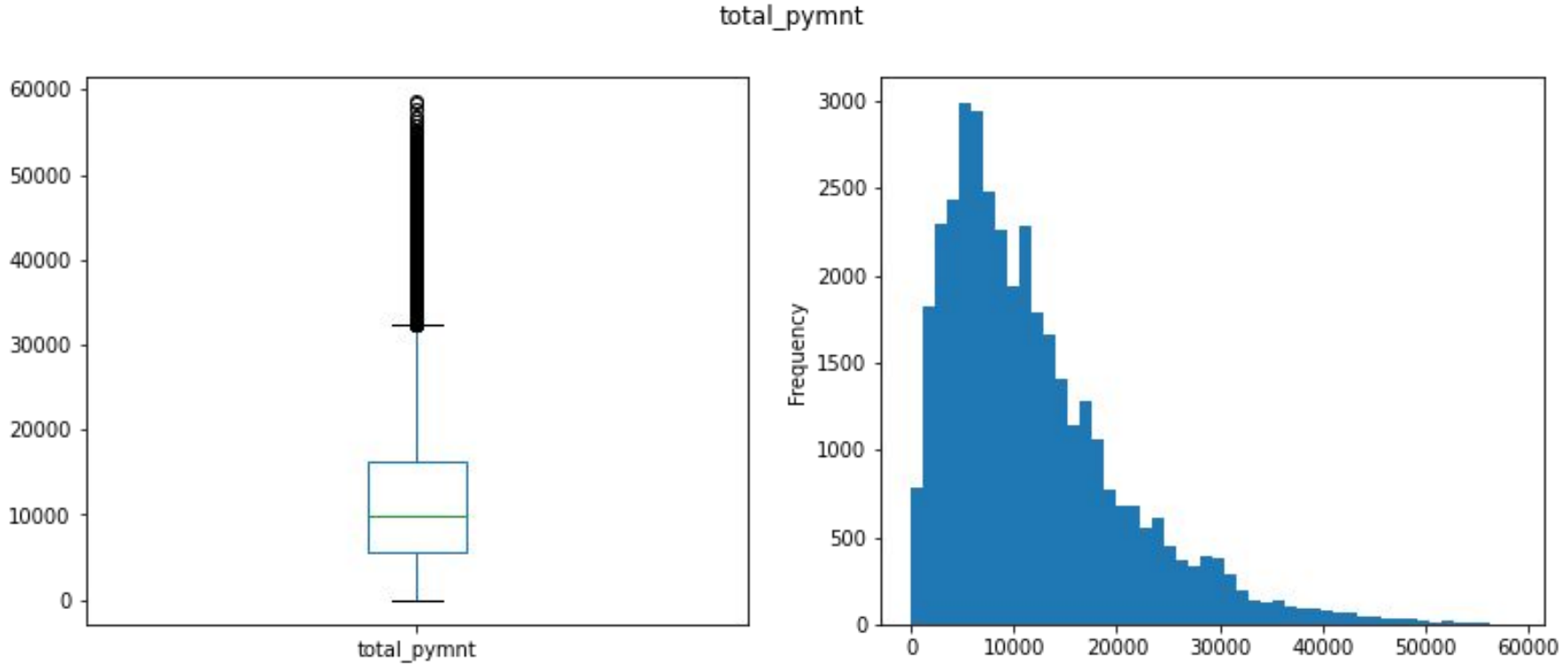
Univariate Analysis (continuous variable)

- **revol_util:** The average revolving line utilization rate of the borrowers is 48.86% and the median rate is 49.30%.



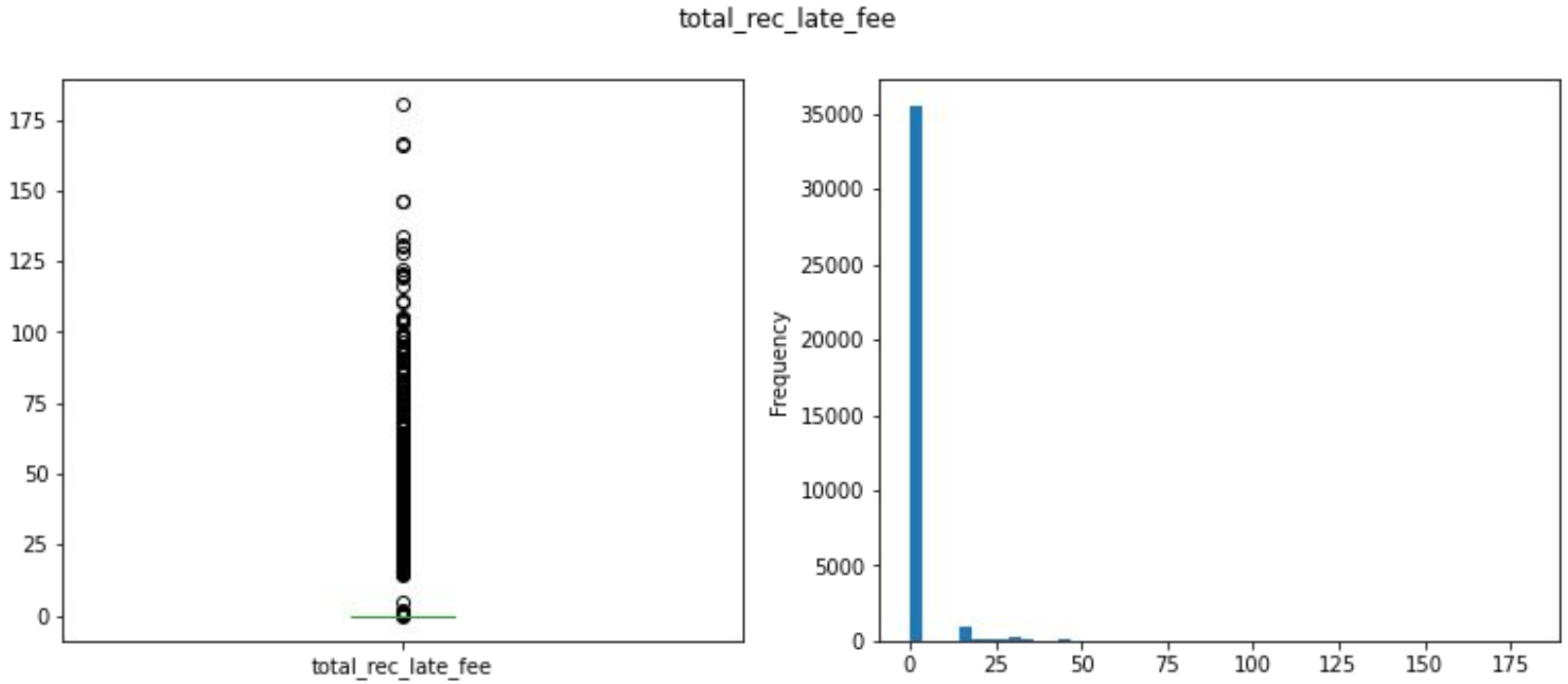
Univariate Analysis (continuous variable)

- **total_pymnt:** The average of total payment received from the borrowers is $11985 and the median payment received is $9811.



Univariate Analysis (continuous variable)

- **total_rec_late_fee:** The mean late fees received from the borrowers is $1.3 and the median of late fee is $0. Most of the borrowers have 0 late fees.



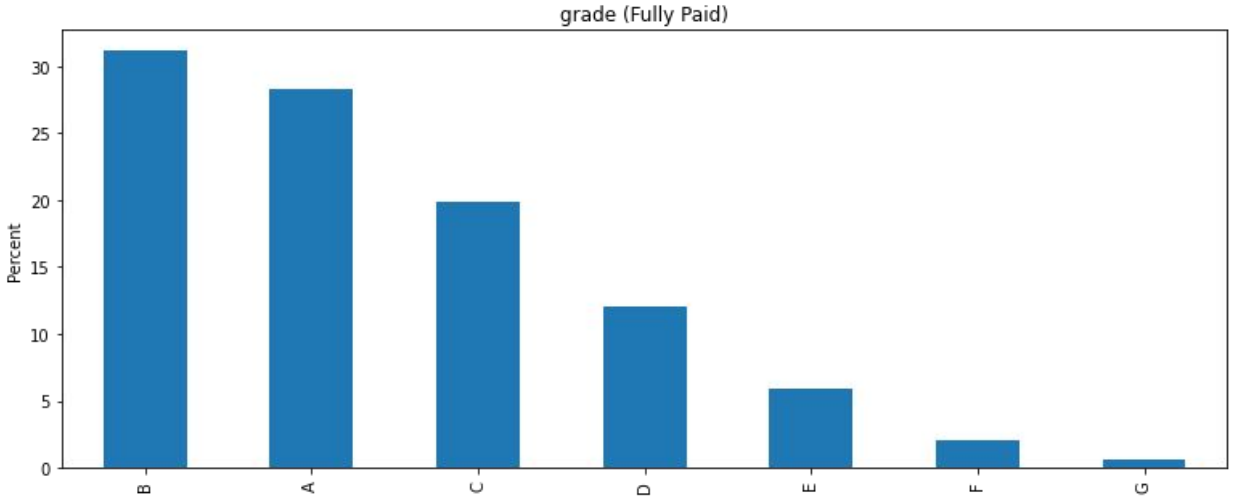
Segmented Univariate Analysis

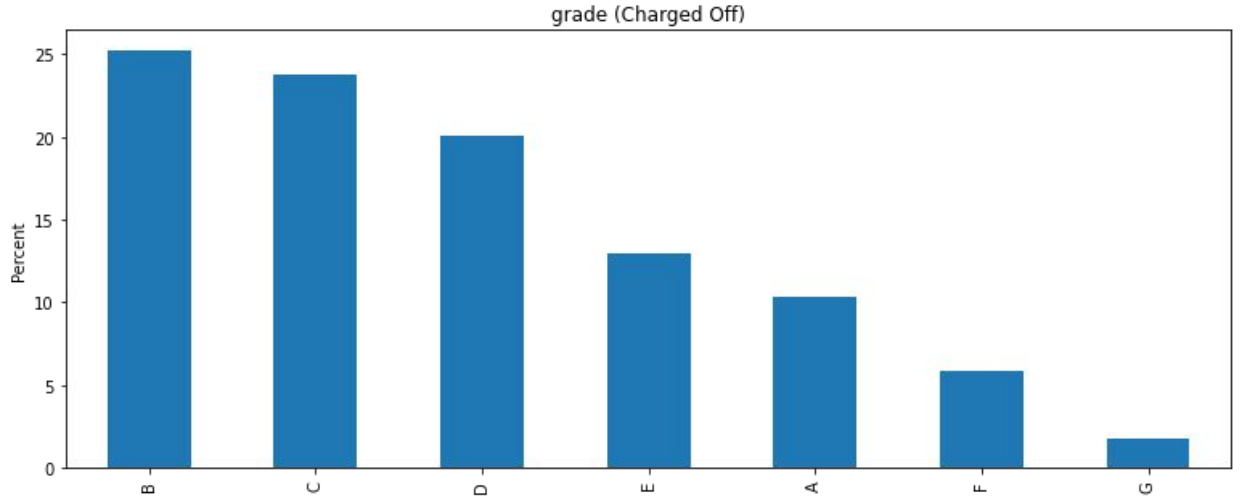
- For segmented univariate analysis, we grouped data into two segments based on loan_status (ie. Fully Paid and Charged Off) and performed univariate analysis on each of them. The groups were formed based on the variable of interest (ie. loan_status). We used the same protocol outlined above in the univariate analysis of categorical and continuous variables.

Next, we show segmented univariate analysis only for the strong indicators identified in this analysis.

Segmented Univariate Analysis (categorical variable)

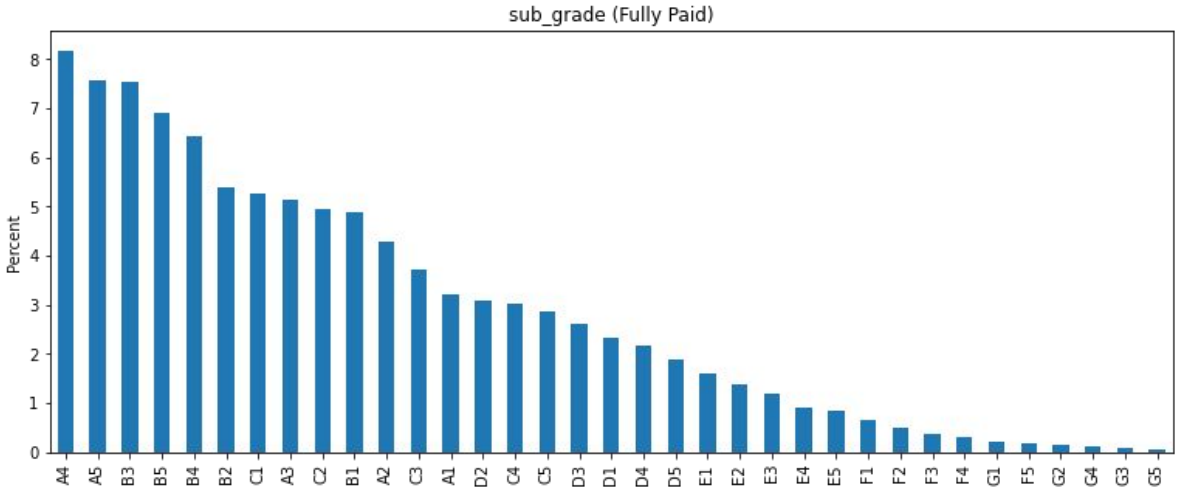
- **grade:** From the distributions below, we can seen that grade B have large number of Fully Paid customers and Defaulters. However, grade A has more percentage of Fully Paid than Defaulter. It looks like Grade A borrowers are most profitable. Whereas Grad B and C borrowers are risky.

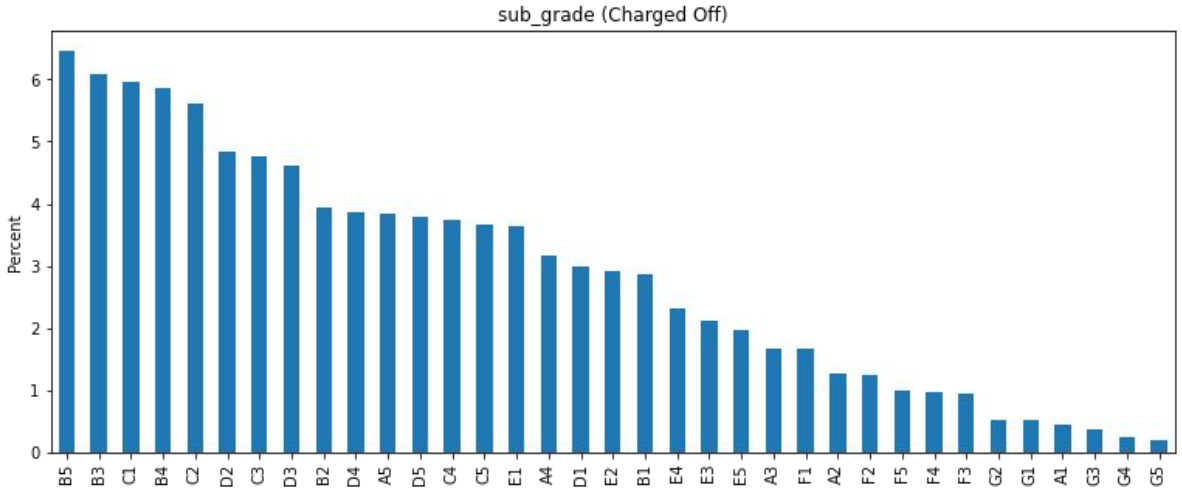




Segmented Univariate Analysis (categorical variable)

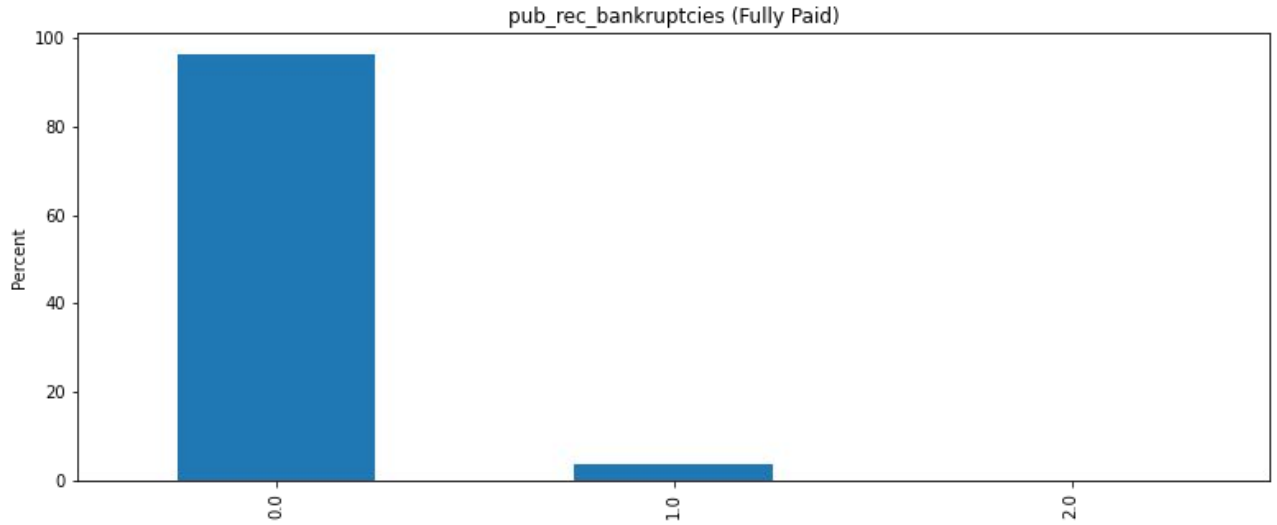
- **sub_grade:** Borrowers with B5 and B3 sub grades are most risky. Borrowers with sub grade of A4 and A5 are most profitable.

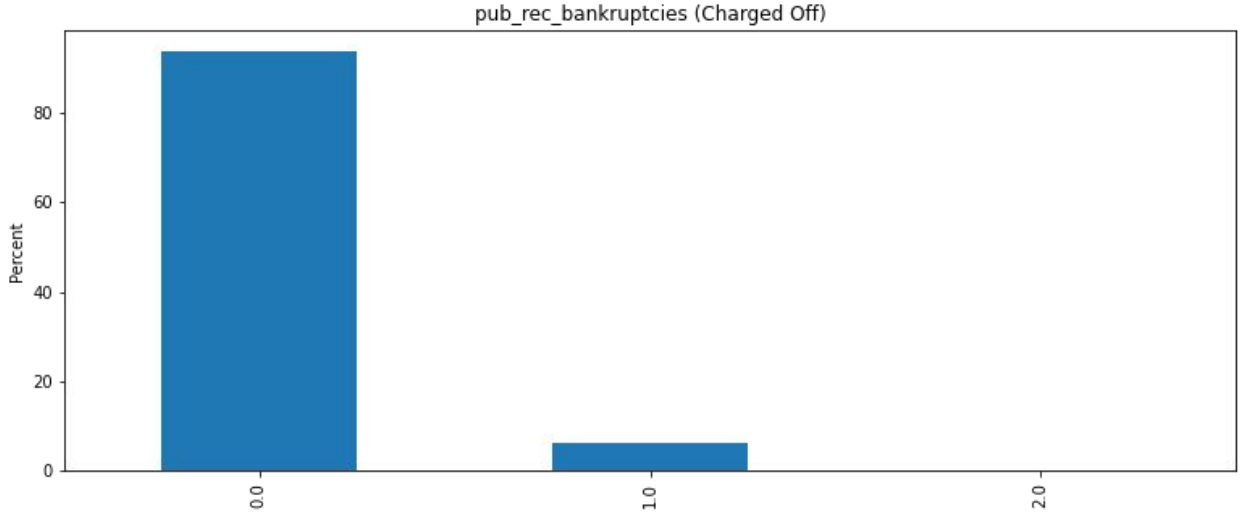




Segmented Univariate Analysis (categorical variable)

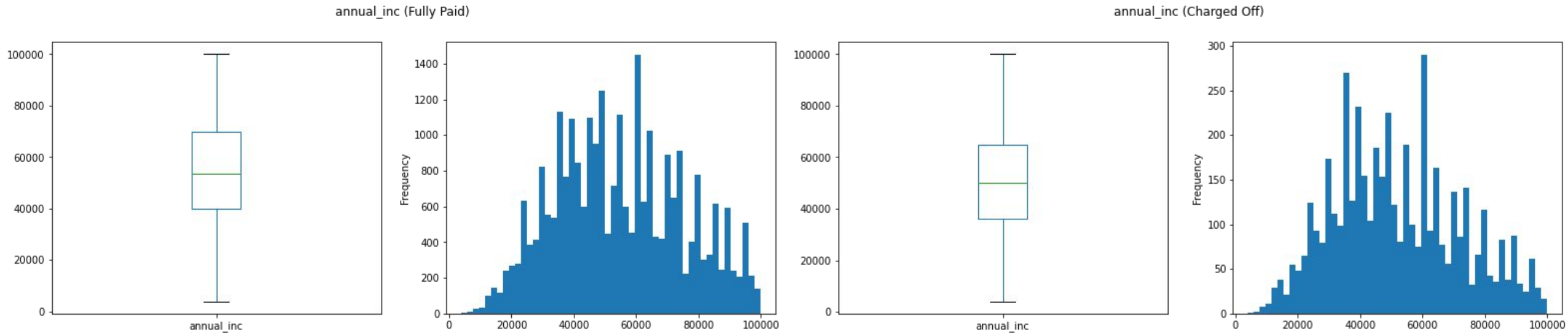
- **pub_rec_bankruptcies:** It is difficult to draw conclusion from these distributions as they show very similar patterns.





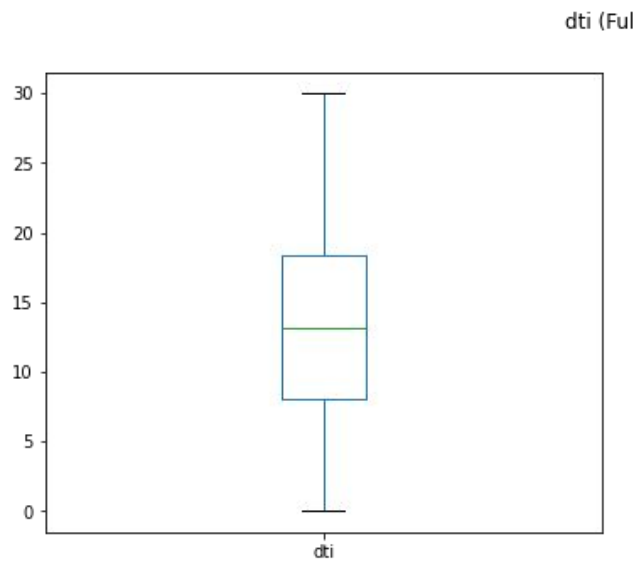
Segmented Univariate Analysis (continuous variable)

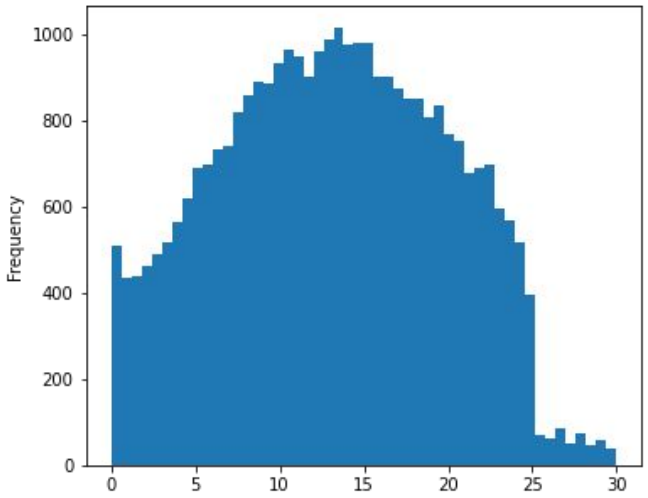
- **annual_inc**

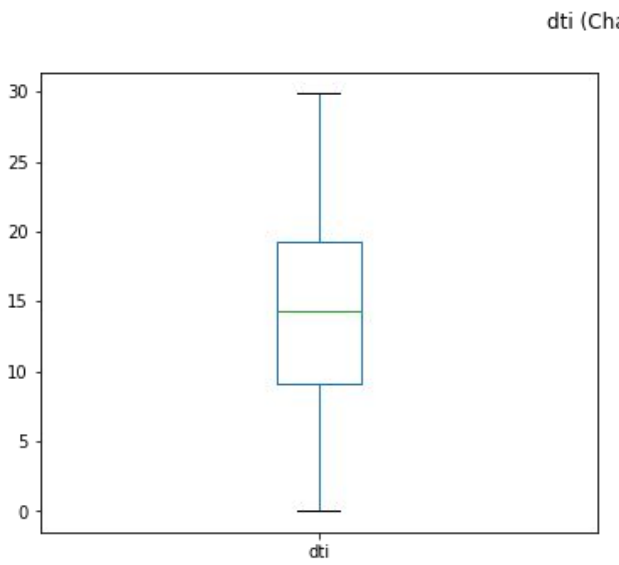


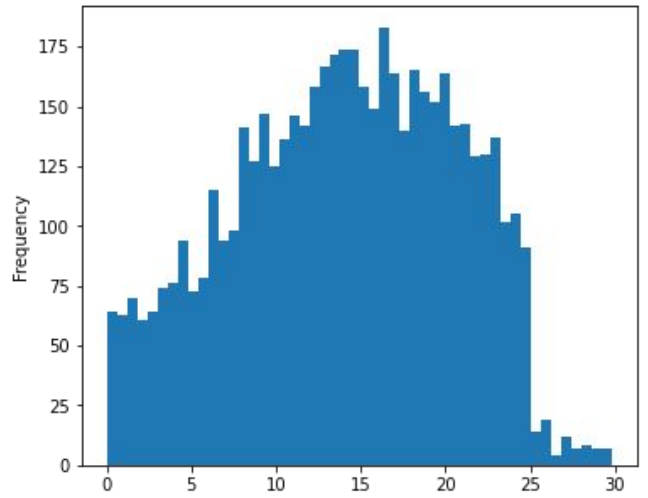
Segmented Univariate Analysis (continuous variable)

- **dti**



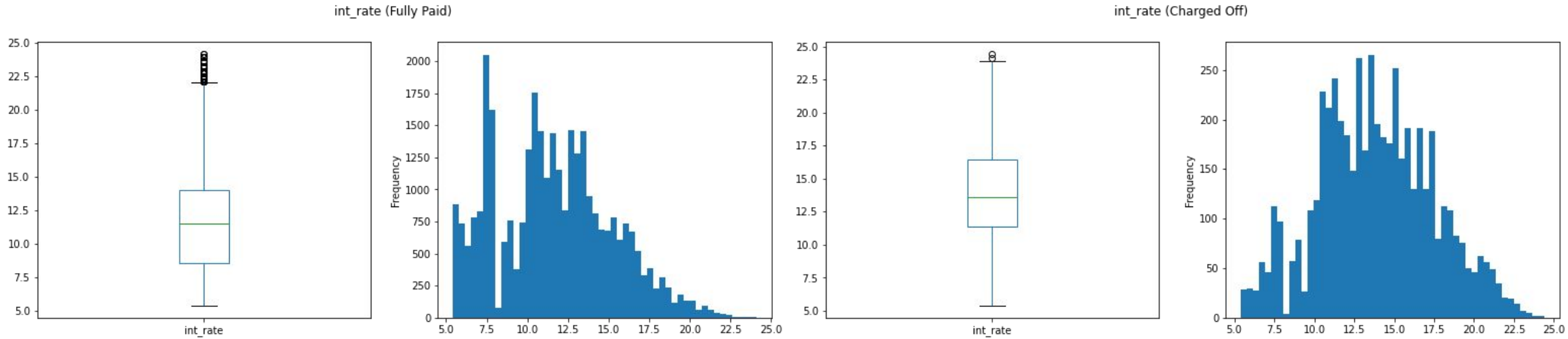






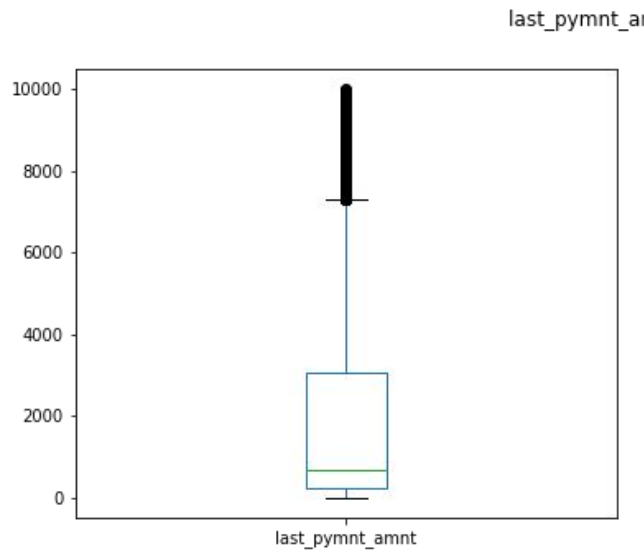
Segmented Univariate Analysis (continuous variable)

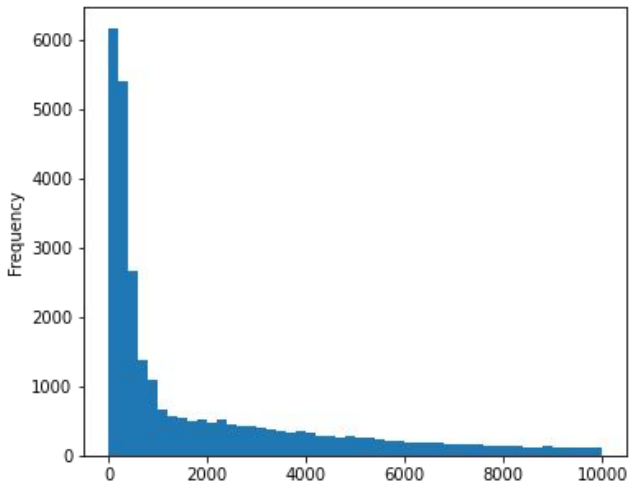
- **int_rate**

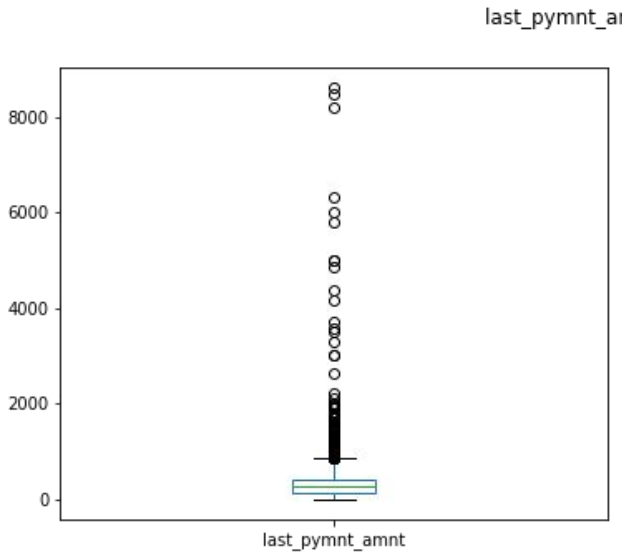


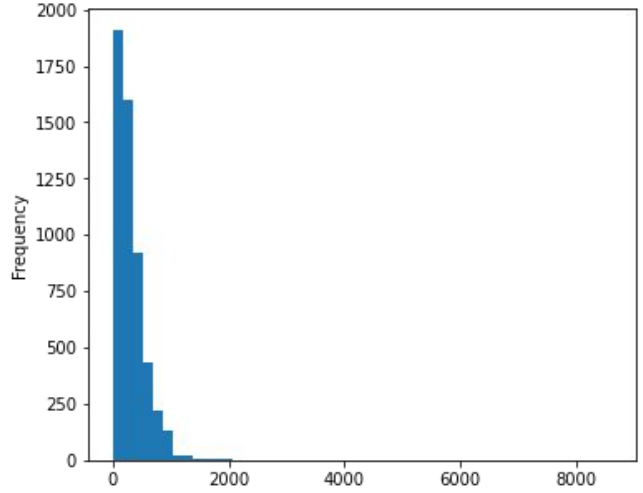
Segmented Univariate Analysis (continuous variable)

- **last_pymnt_amnt**



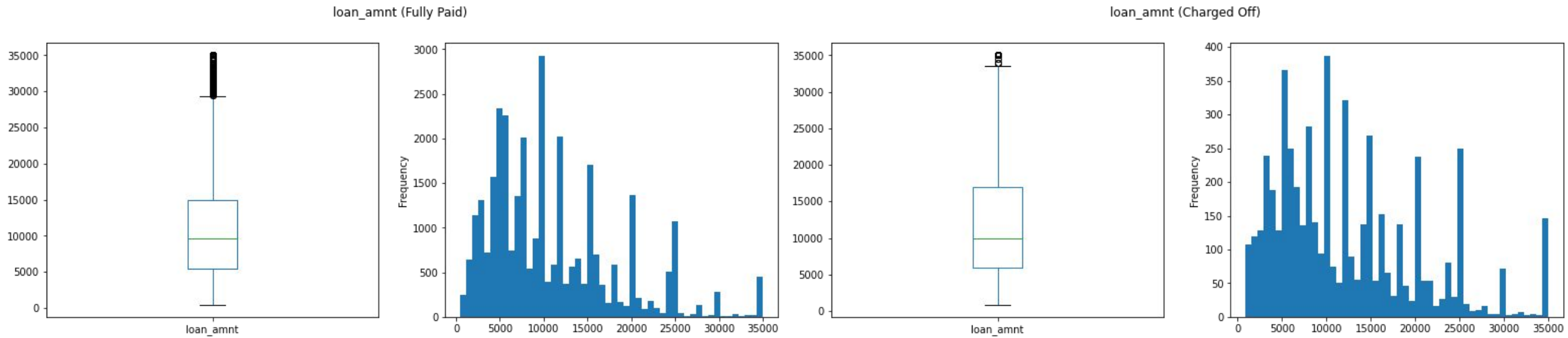






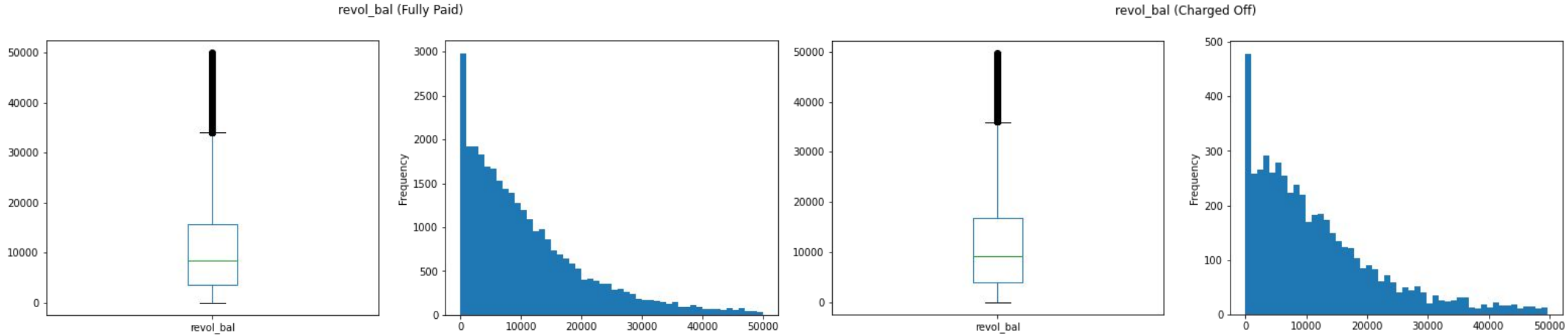
Segmented Univariate Analysis (continuous variable)

- **loan_amnt**



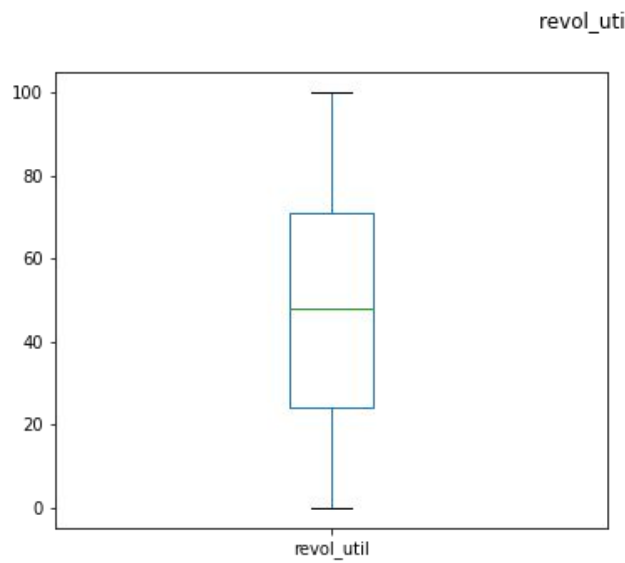
Segmented Univariate Analysis (continuous variable)

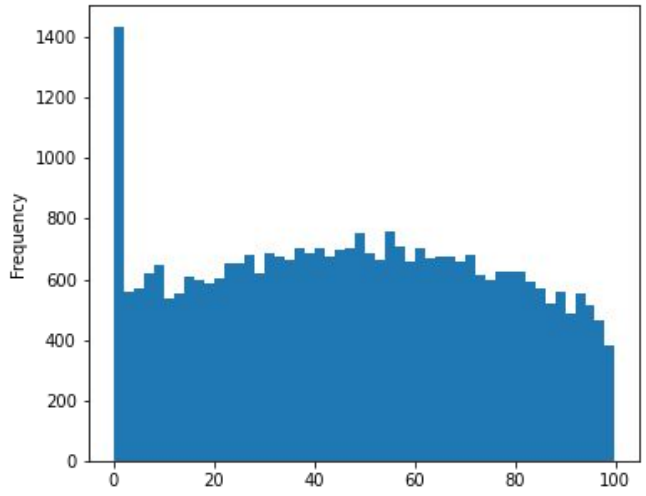
- **revol_bal**

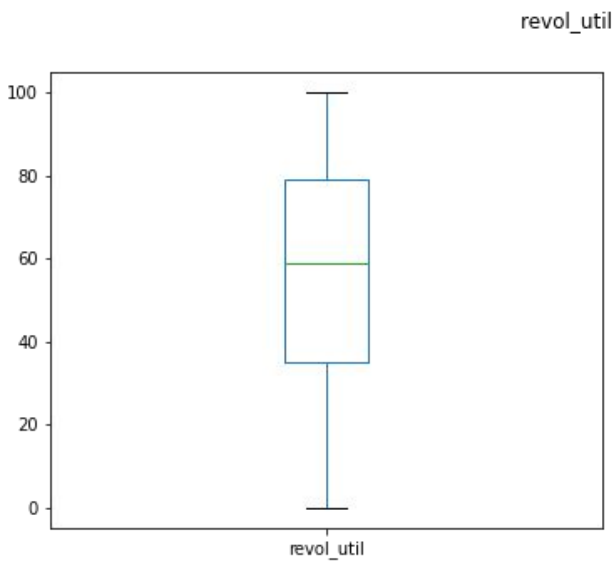


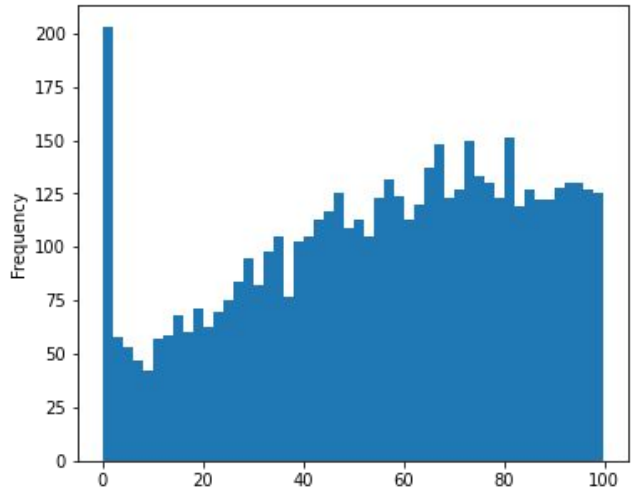
Segmented Univariate Analysis (continuous variable)

- **revol_util**



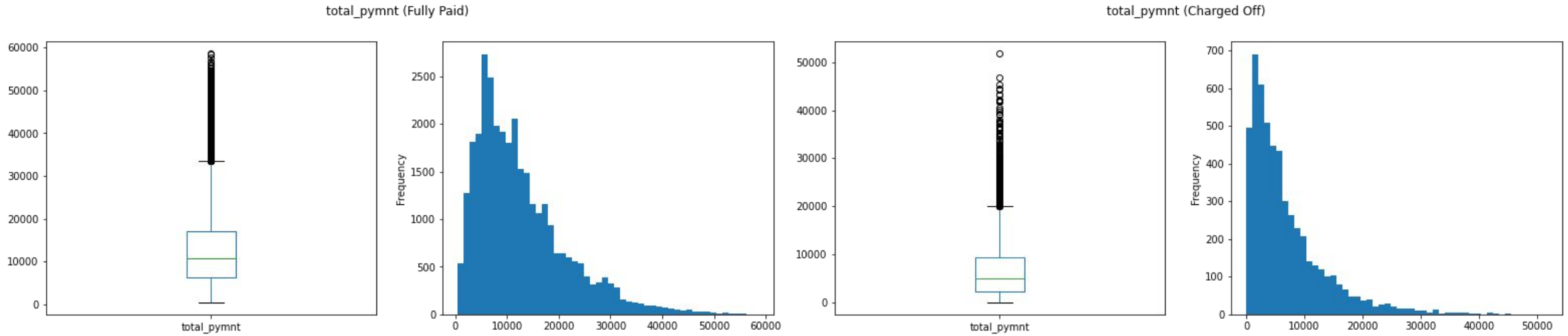






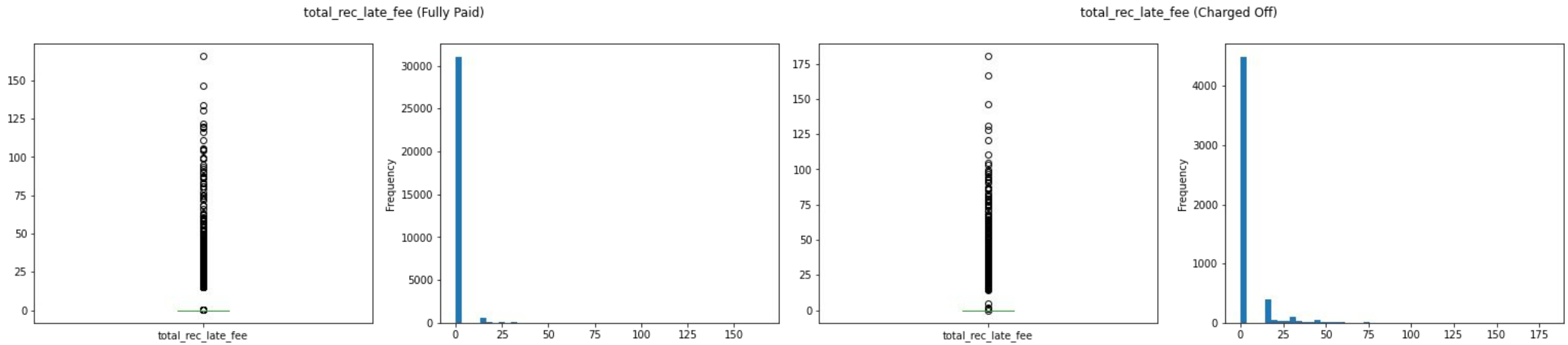
Segmented Univariate Analysis (continuous variable)

- **total_pymnt**



Segmented Univariate Analysis (continuous variable)

- **total_rec_late_fee**

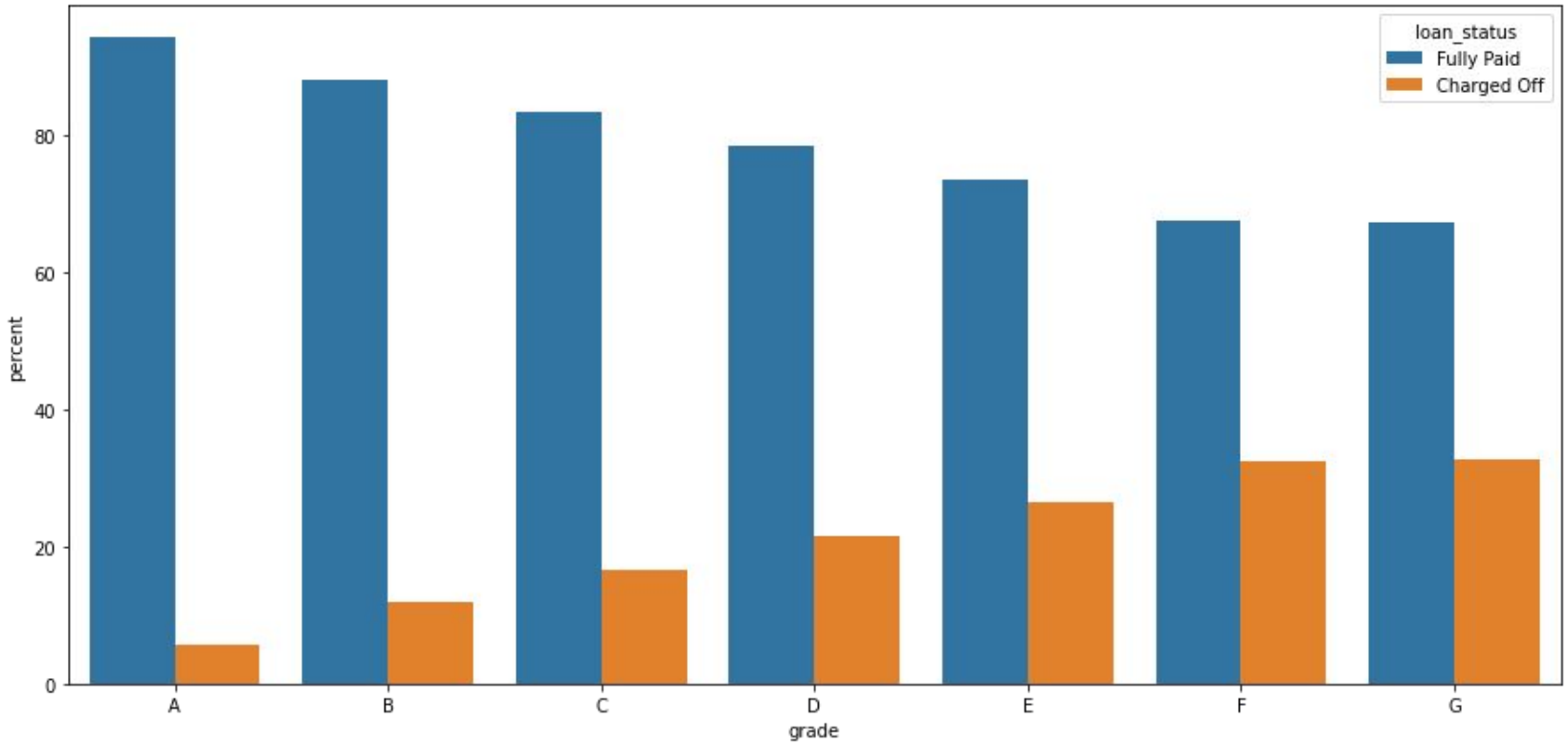


Bivariate Analysis

- Bivariate analysis was performed with respect to loan_status as this is the quantity of interest.
- We identified important driver variables for default mainly through bivariate analysis.
- Next, we show results of bivariate analysis and give reasons as to why we think the variable is a driver variable for default.

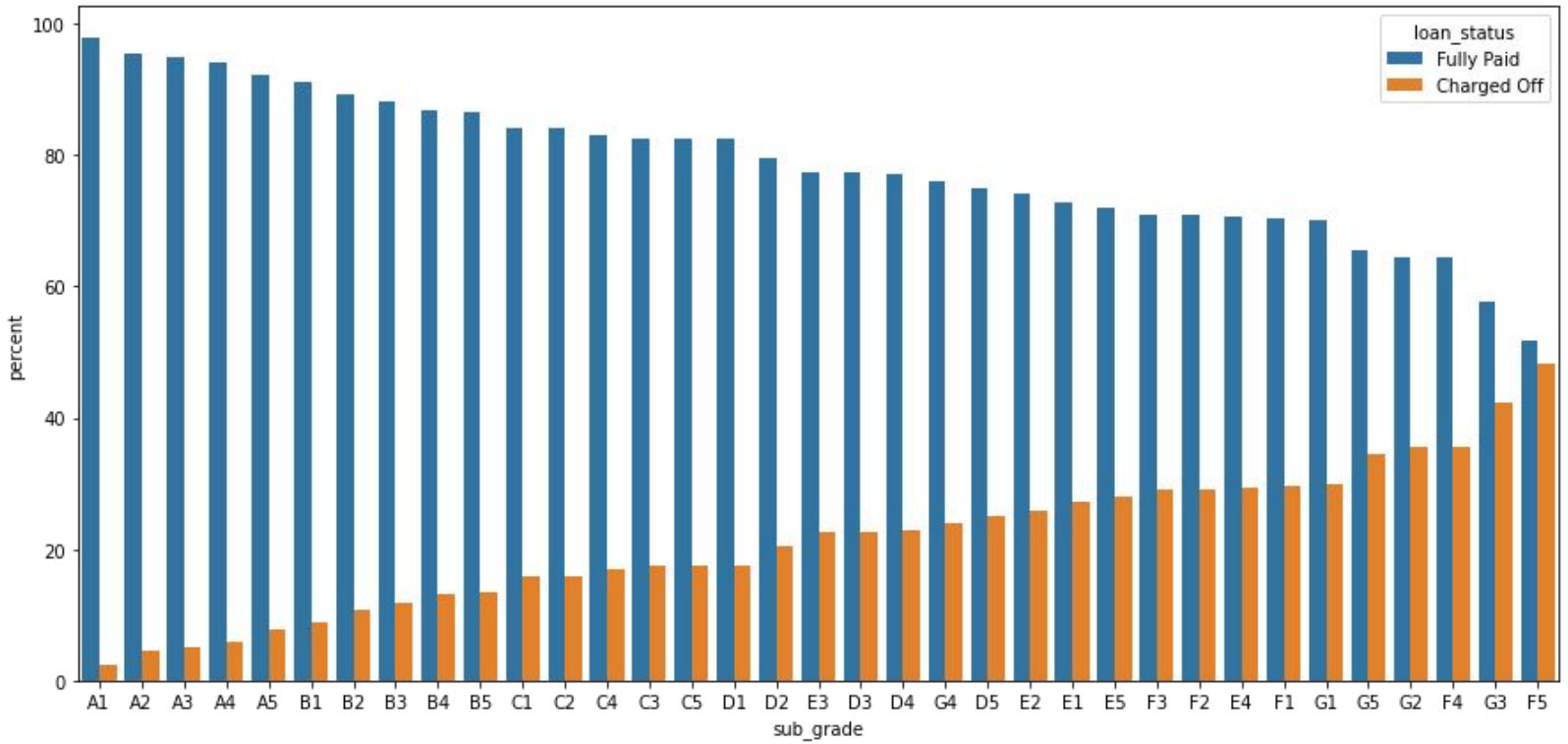
Bivariate Analysis (categorical variable)

- **grade:** Borrowers having grades E, F, G are risky are more likely to default.



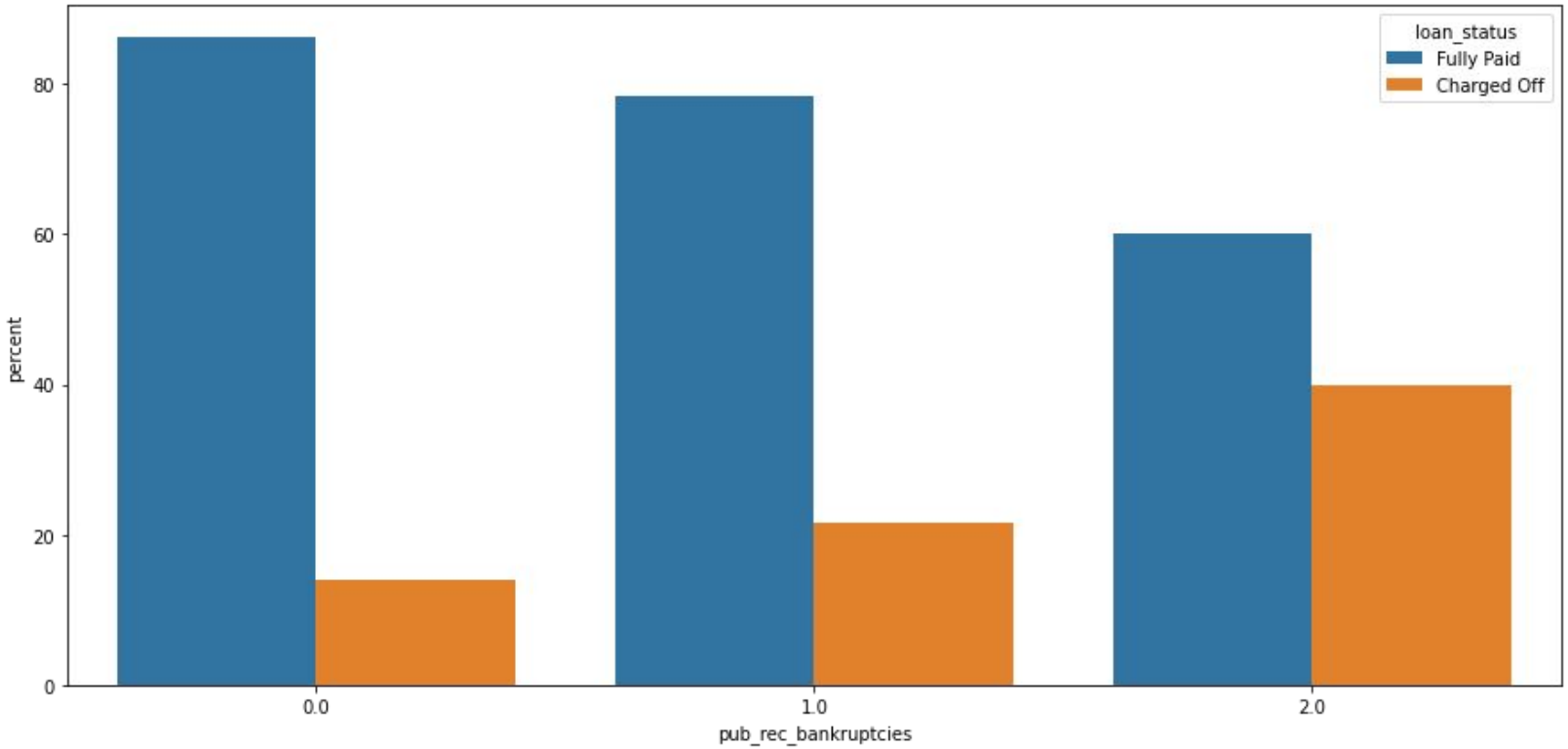
Bivariate Analysis (categorical variable)

- **sub_grade:** Borrowers with sub grade F5, G3 and F4 are more risky than others.



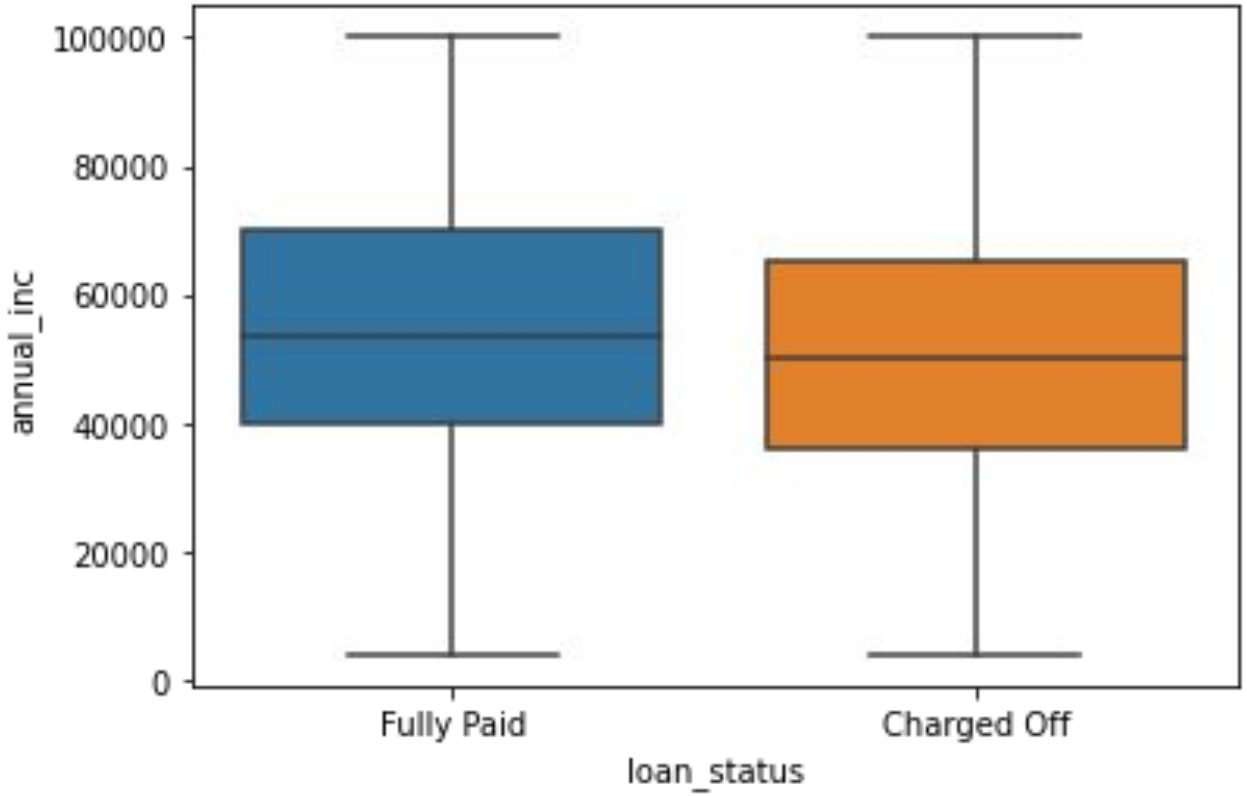
Bivariate Analysis (categorical variable)

- **pub_rec_bankruptcies:** Borrowers with high number of public bankruptcies are more likely to default.



Bivariate Analysis (continuous variable)

- **annual_inc:** Borrowers with low annual income are more likely to default.



Bivariate Analysis (continuous variable)

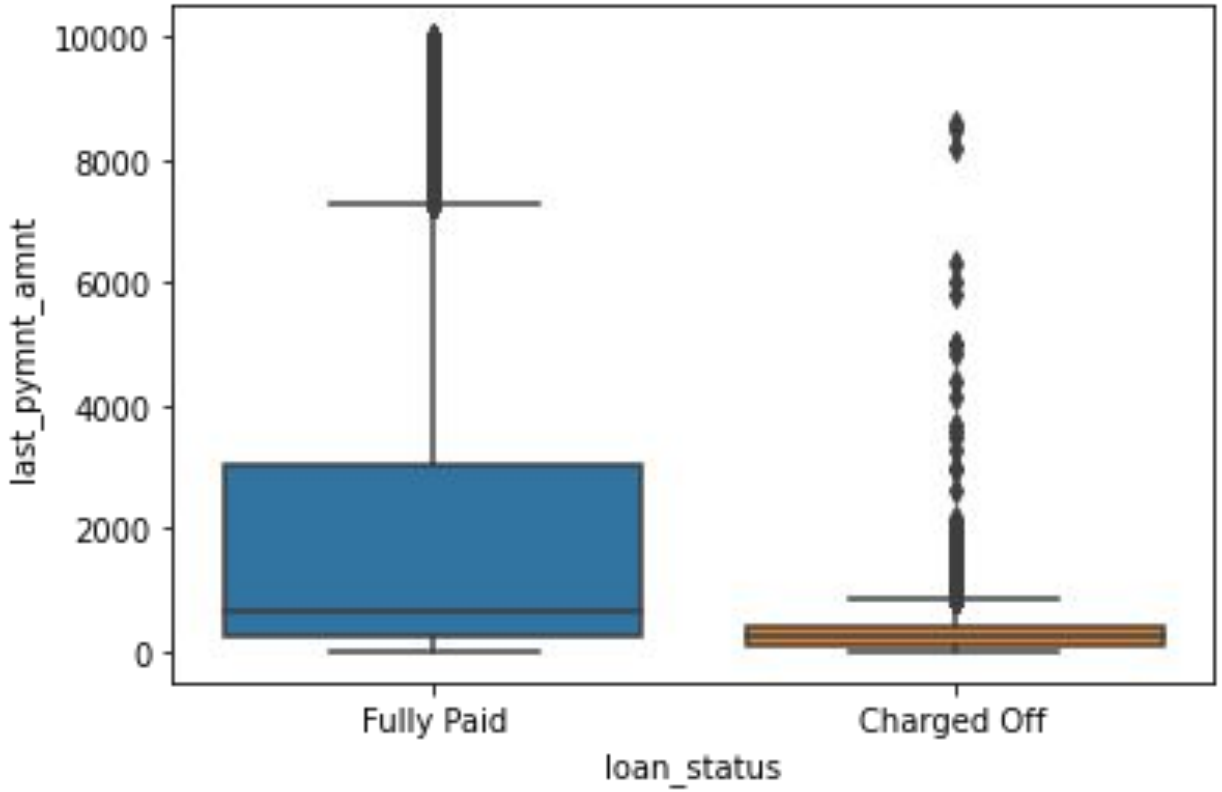
- **dti:** Borrowers with high dti are more likely to default.

Bivariate Analysis (continuous variable)

- **int_rate:** Borrowers with high interest rates are more likely to default.

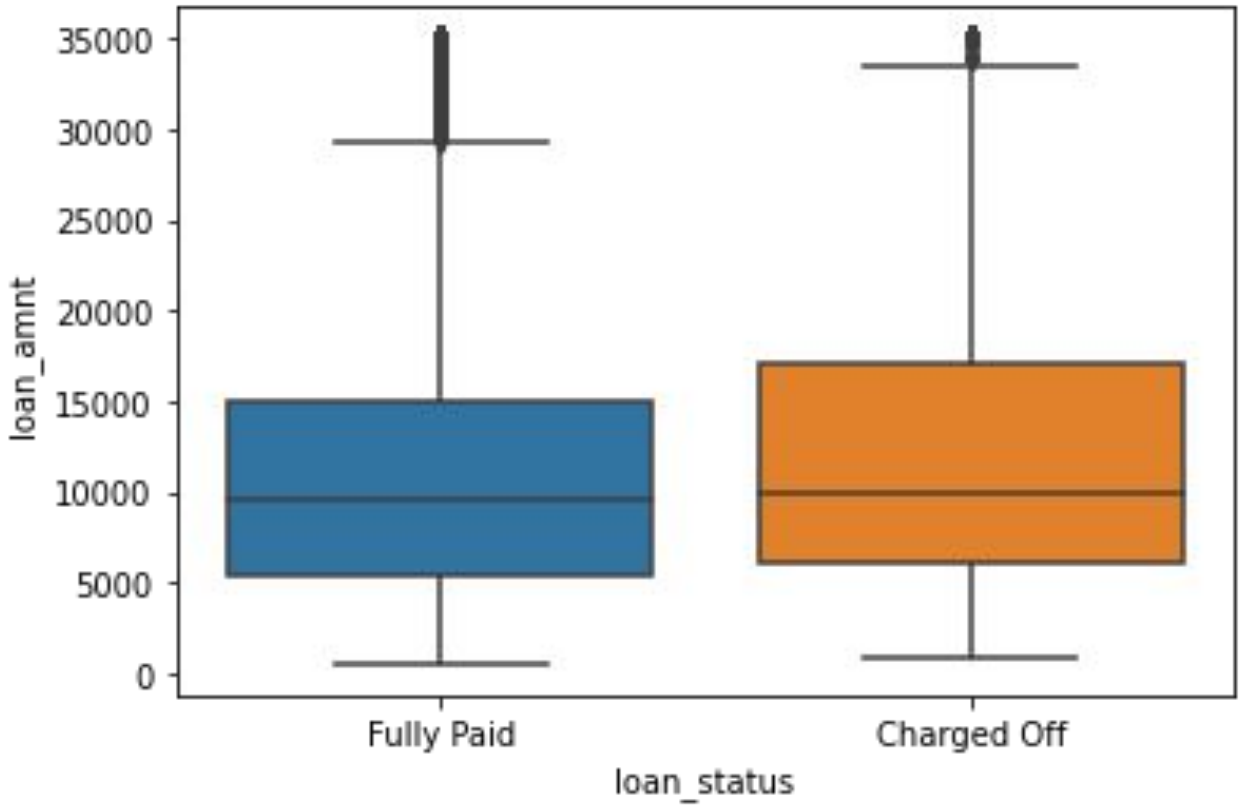
Bivariate Analysis (continuous variable)

- **last_pymnt_amnt:** Borrowers who paid low amount as a last payments are more likely to default.



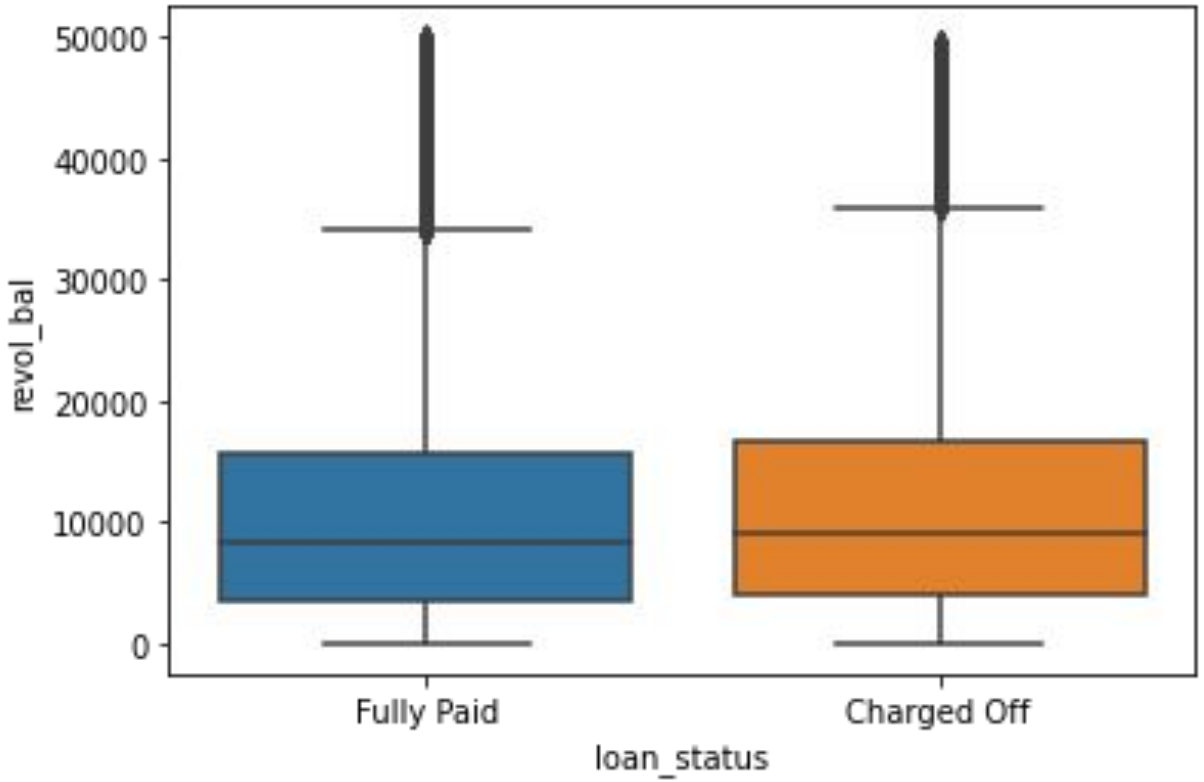
Bivariate Analysis (continuous variable)

- **loan_amnt:** Borrowers who have high loan amount are more likely to default.



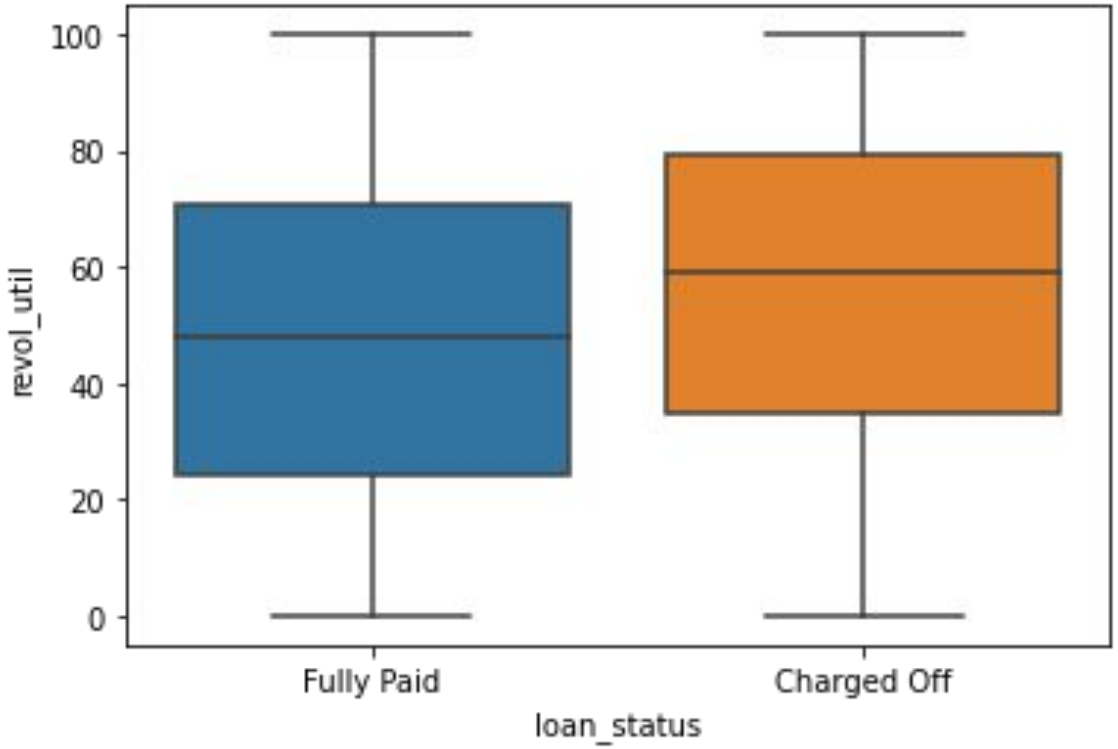
Bivariate Analysis (continuous variable)

- **revol_bal:** Borrowers with high revolving credit balance are more likely to default.



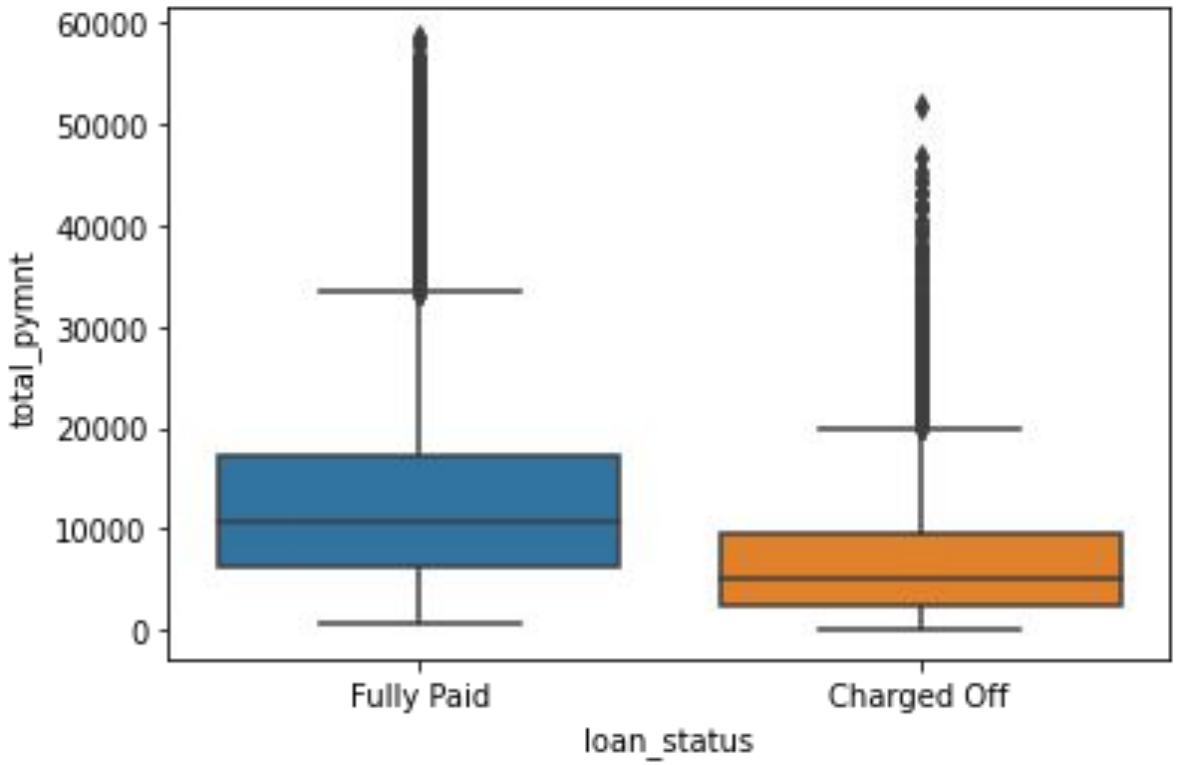
Bivariate Analysis (continuous variable)

- **revol_util:** Borrowers with high revolving utilization rate are more likely to default.



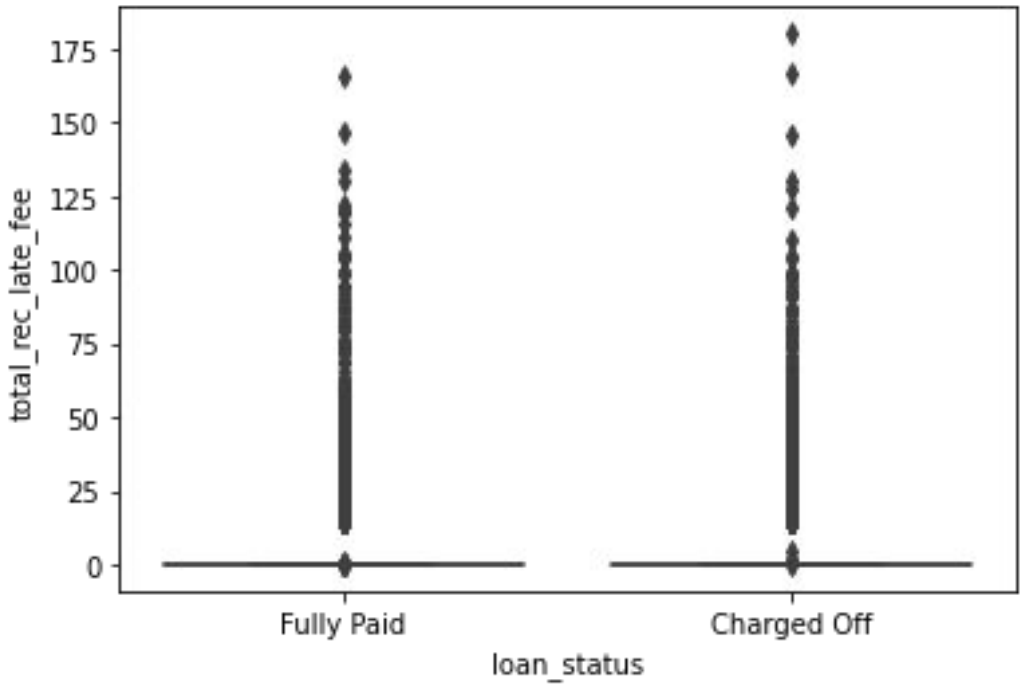
Bivariate Analysis (continuous variable)

- **total_pymnt:** Lower total payment received from the borrower is the indicator of possible defaulter.



Bivariate Analysis (continuous variable)

- **total_rec_late_fee:** Higher late fees indicate that borrower is delinquent. Hence, more likely to default. Due to large number of borrowers with 0 late fees box plot is not useful. So, we compare metrics as shown in the table below.



	Mean	q=25%	q=50%	q=75%
Charged Off	4.44	0	0	0
Fully Paid	0.87	0	0	0

Conclusions

- Univariate analysis was quite useful to get familiar with data and understand its properties. It also helped us to identify outliers in the variables.
- We didn’t find segmented univariate analysis that useful.
- Bivariate analysis was the most insightful for identifying driver variables that are indicators of default.
- We identified following 13 driver variables which are strong indicator of default: grade, sub_grade, pub_rec_bankruptcies, annual_inc, dti, int_rate, last_pymnt_amnt, loan_amnt, revol_bal, revol_util, total_pymnt, total_rec_late_fee

Guidelines for Risk Assessment

We should be careful while approving loans for the applicant having following properties:

- Applicant has G, E or F grade.
- Applicant has F5, G3 or F4 sub grade.
- Applicant has public record of bankruptcies.
- Applicant has below the average annual income (<$50,000).
- Applicant has high dti (>14).
- Applicant has taken loan with high interest rate (>13%).
- Applicant’s last payment/installment was partial.
- Applicant has take loan of high amount (>$12,000).

Guidelines for Risk Assessment

- Applicant has high revolving balance (>$11,899).
- Applicant has high revolving utilization rate (>55%).
- Applicant has delay the payment/installment.