LendingClub Case Study

Team :
Badari Prasad H
Saurabh Bhargava

July 2023 MLC53

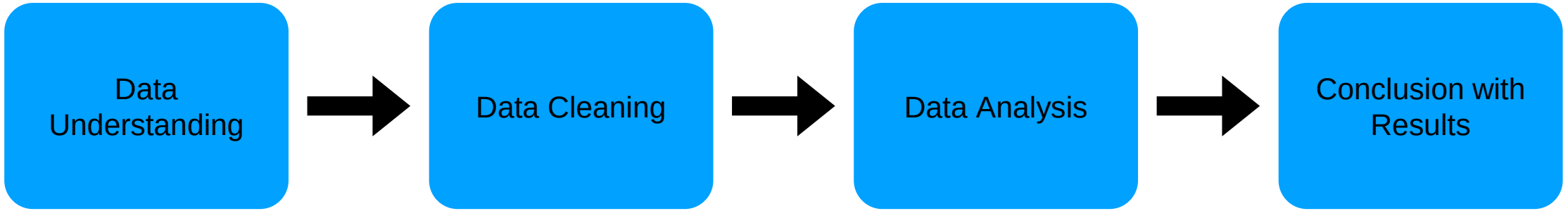
Objective

- **Company**: Lending Club 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.

- **Context:** The company aims to identify the key variables that influence loan default, i.e. the variables that strongly predict default. This knowledge can help the company with its portfolio and risk management.

- **Problem Statement:** Use exploratory data analysis (EDA) to examine the dataset of past applicants' information. Find out what factors make a borrower more likely to default on a loan. Use these findings to inform your decisions on whether to approve or reject a loan application, and how much interest and loan amount to offer.

Analysis Methodology



- Studying columns using dictionary.
- Identify the important columns in dataset that may contribute in analysis.

- Cleaning Missing values
- Removing Duplicate columns
- Removing rows with all zero/one value and null values.
- Removing unwanted columns which does not contribute to the analysis, ex : misc, emp title, desc etc…
- Convert the values to proper datatype such as float, int etc.

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Correlation Table
- Correlation Chart

- Insights and Observation

Data Understanding

- The dataset has 39717 rows and 111 columns.
- There are many columns which consists of null value and NAN values.
- The dataset has contains different variables like categorical and numerical data.
- Some of the columns only consists of NAN values.
- There are some columns which has string values also.
- The columns in the dataset has duplicate values.
- Our dataset needs a good amount of pre-processing.

Data Understanding, Useful categories…

Loan related	Customer related	Customer behaviour related
loan_amnt	home_ownership	delinq_2yrs
funded_amnt	term	pub_rec_bankruptcies
term	verification_status	revol_bal
int_rate	purpose	recoveries
installment	grade	purpose
annual_inc	pub_rec_bankruptcies	earliest credit line
dti	issue_d_month	loan_status
open_acc	Annual salary	
pub_rec	Grade	
total_acc		
total_pymnt		
total_pymnt_inv		
total_rec_prncp		
total_rec_int		

Data Cleaning and Manipulation/Impute

- Data cleaning comprises of dropping all duplicate values, dropping columns which does not have any data at all.
- Dropping columns which does not impact analysis like id, url, desc etc..
- Impute the data with mean/median/mode for numerical columns where data were missing, based on column and domain knowledge few columns were filled with value of zero. Example : mths_since_last_delinq, last_pymnt_d, revol_util etc…
- Normalizing the data for example, date can contain string like “<2008”, such data to be normalized using appropriate methodology, instead of dropping such rows. And another instance like employee length to be normalized and should contain only integer values instead of strings.
- Outcome of this step is the data which is ready to be taken to further analysis.

Data Analysis

- Data Analysis comprises of different steps in itself :
  - Univariate Analysis.
  - Bivariate Analysis.
  - Multivariate Analysis.
  - Creation of Correlation Table / Heatmap.

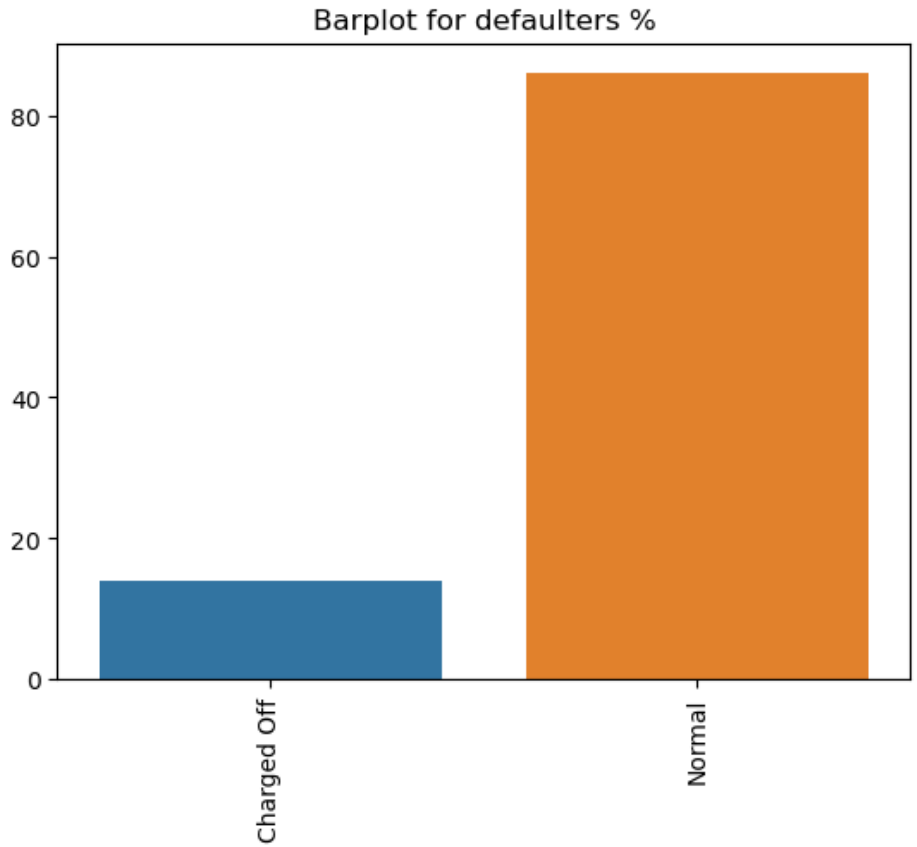
Data Analysis : Continued…

- Looking deeper into all the columns we see loan issue_d which is issue date when the loan was issued to the borrower.
- And as part of problem statement, it can be observed that the date range should be between 2007 to 2011, so check the value counts split the issue date into month issued and year loan was issued.
- And similar approach of creation of year and months are done for all other relevant columns which represents date.
- And using “issue_year” and “issue_month” details, data was further cleaned, and any data beyond the problem statement were dropped.

Univariate Analysis

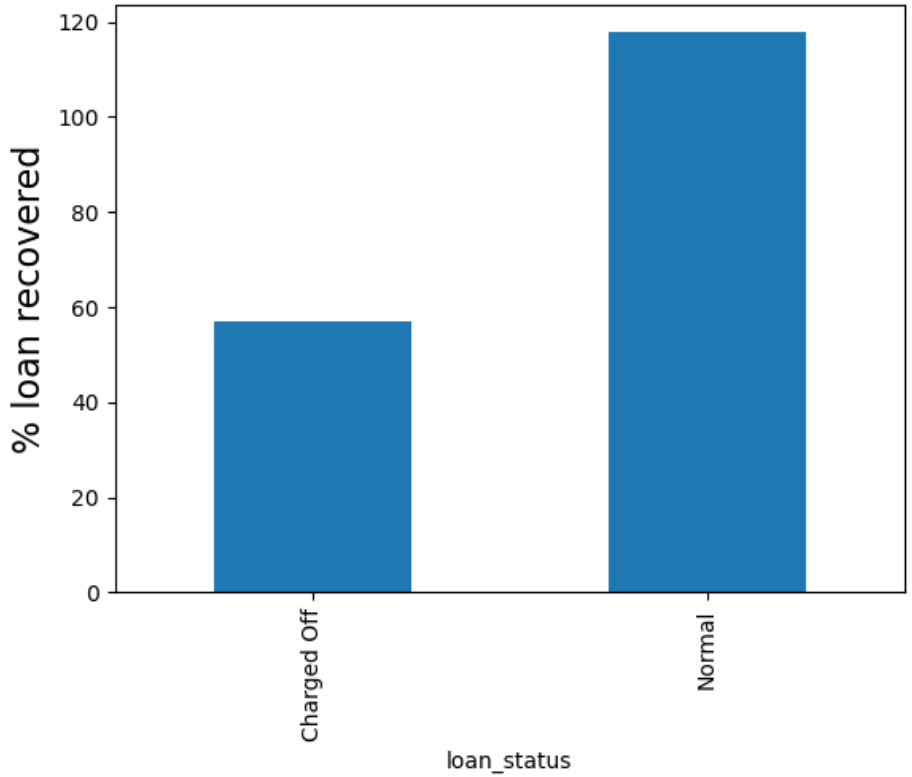
For any lenders its absolutely critical to understand :

- How many borrowers are defaulters in % terms : from analysis its found that **14%** of the borrowers defaulted on loan payment.



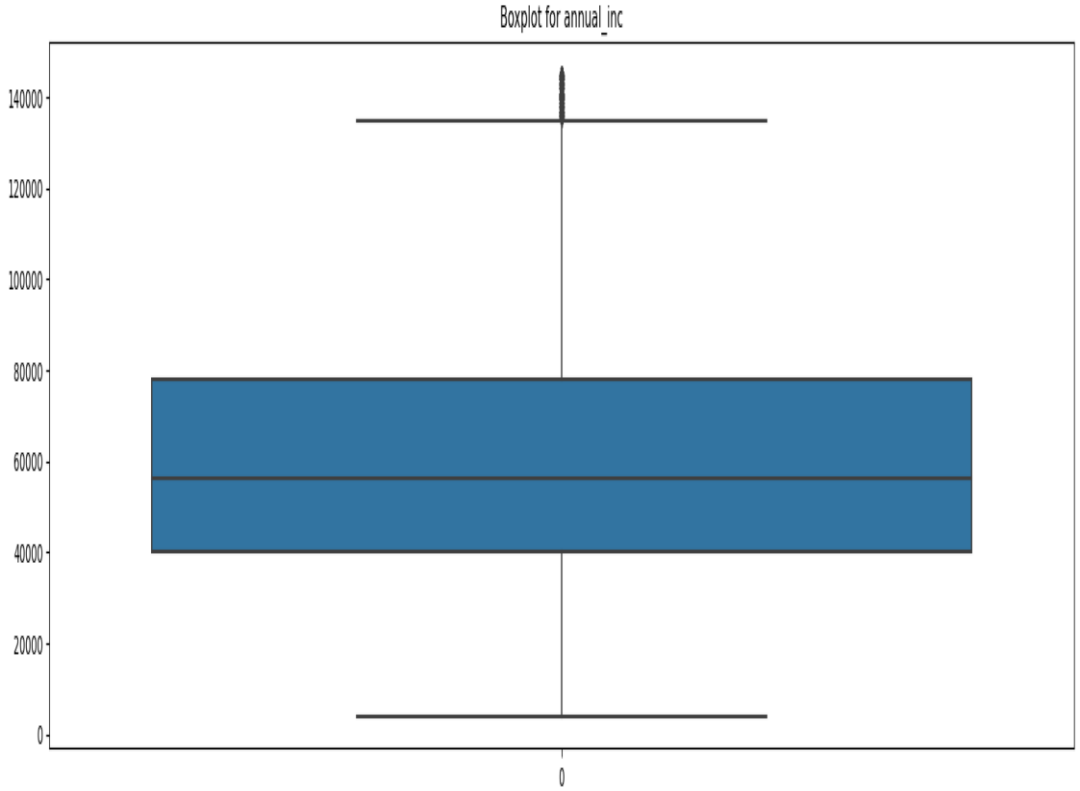
Univariate Analysis : Continued…

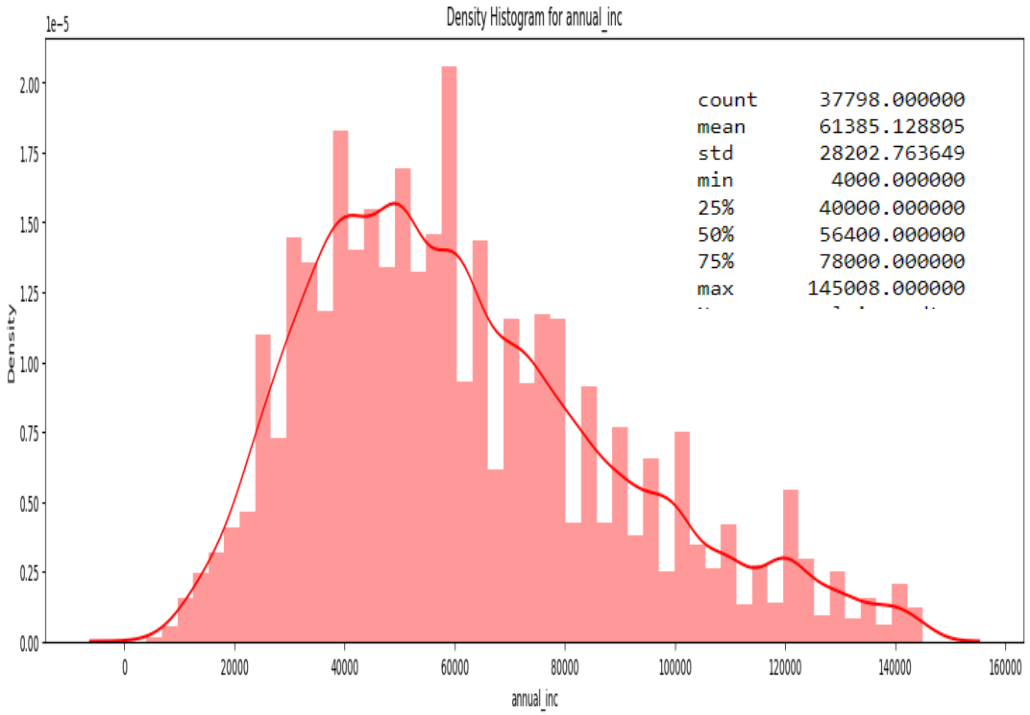
- Percentage of loan that was recovered from borrowers: **only 57%** of loans were recovered and out of fully recovered loans **17% is the** profit from below plot.
- And inference from these 2 charts : Business can afford an additional **16.49%** of offenders and any variable which increases % of offenders would lead to loss of business.



Univariate Analysis Continued : Annual Income column analysis

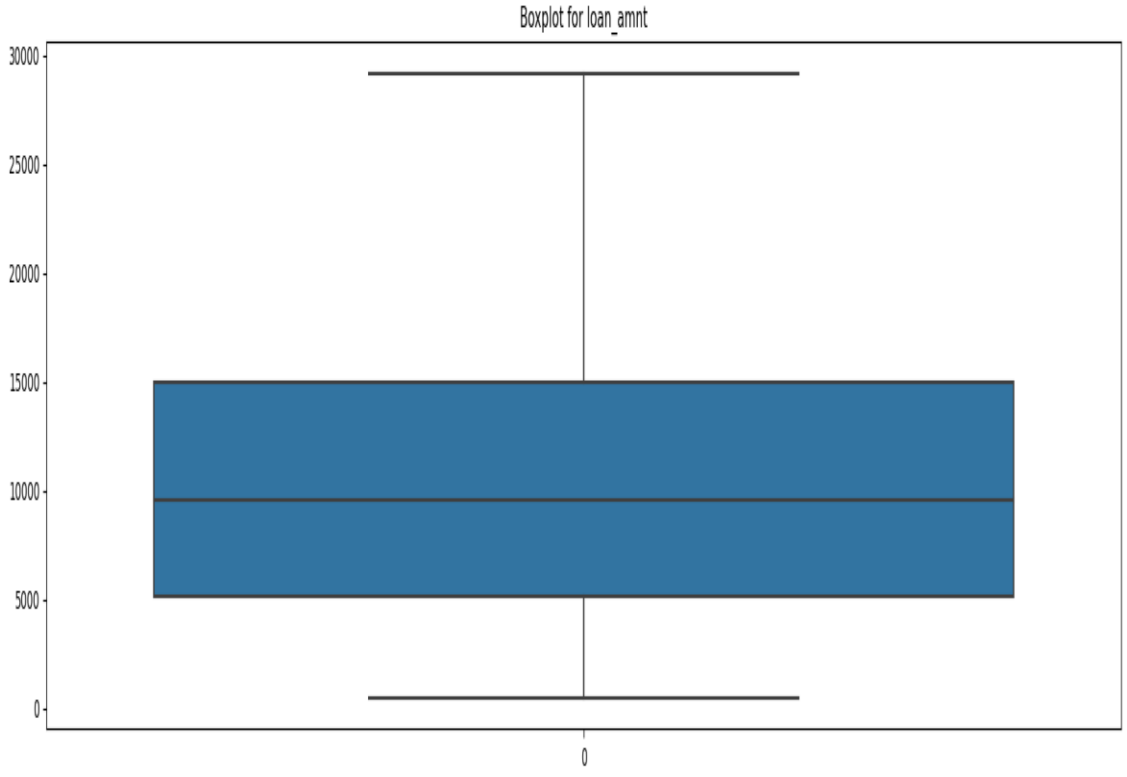
Using annual_inc column : annual income as disclosed by the borrower we can see that income ranges from $4000 to $145008. So borrowers above 50% quantile are most likely to pay the loan and others are most likely to default.





Univariate Analysis Continued : Loan Amount column analysis

Using loan_amnt column : Loan amount requested the borrowers; we can see that range is from min of $500 to max of $30000.



Inference from Univariate analysis

- **14%** of the borrowers defaulted on loan.
- **Only 57%** of loans were recovered.
- Out of fully recovered loans **17% is the** profit.
- The loan business is vulnerable to losses from defaults and low recoveries, and it needs to monitor its default rate closely and take preventive measures to keep it below **16.49%.**

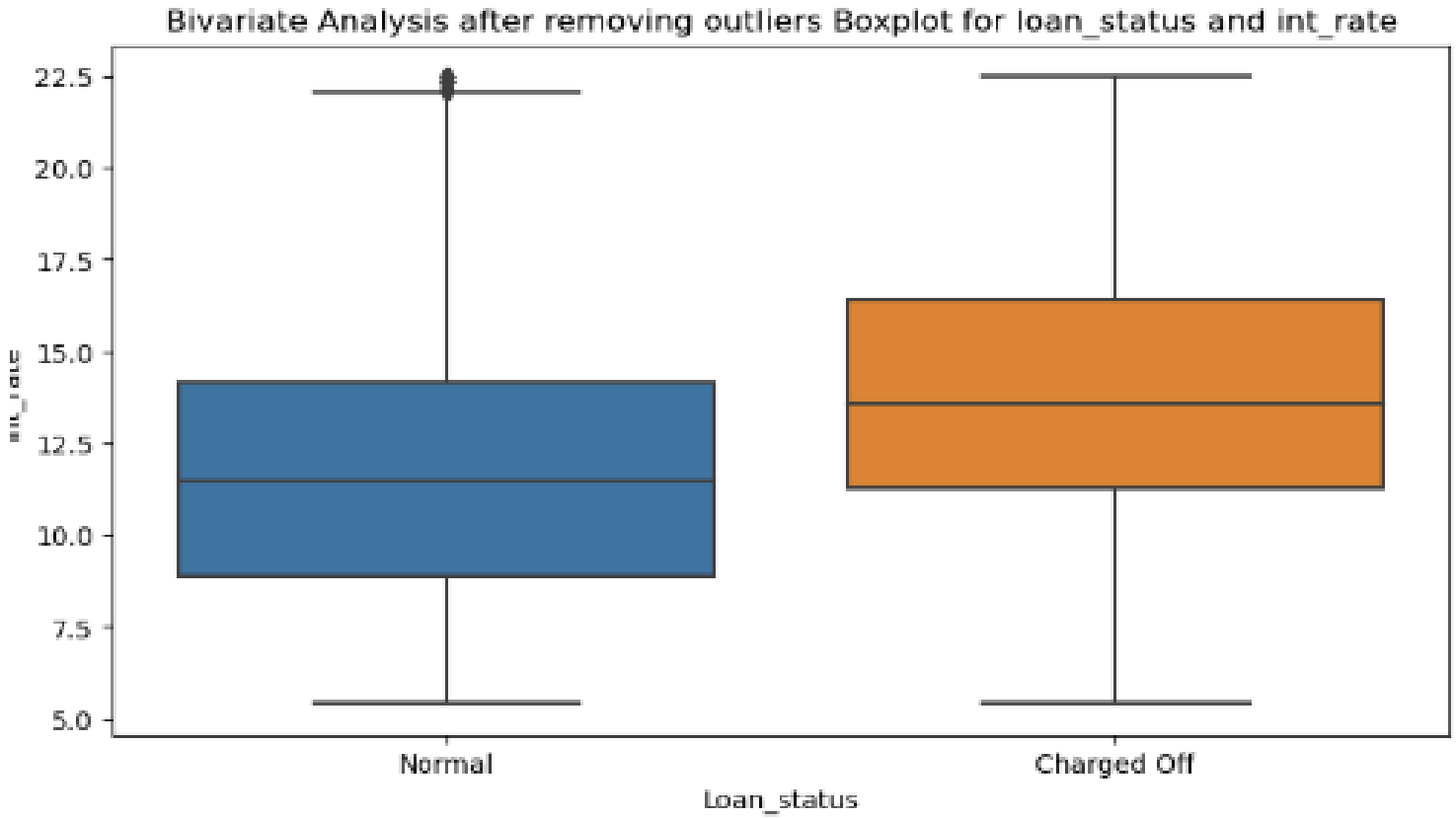
Bivariate Analysis

While doing bivariate analysis, following details emerged as important factors:

- Charged Off Borrower's have :
  - ✓ int_rate between range 11.26 and 16.4
  - ✓ open_acc between range 6.0 and 11.0
  - ✓ total_pymnt between range 2271.96 and 9158.71
  - ✓ total_pymnt_inv between range 1914.79 and 8485.97

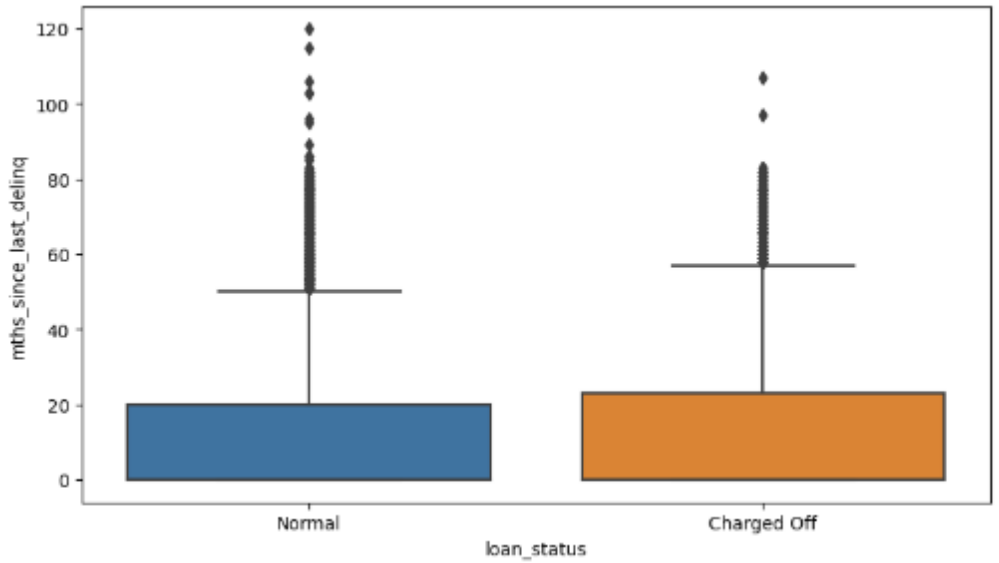
Bivariate Analysis. Continued.

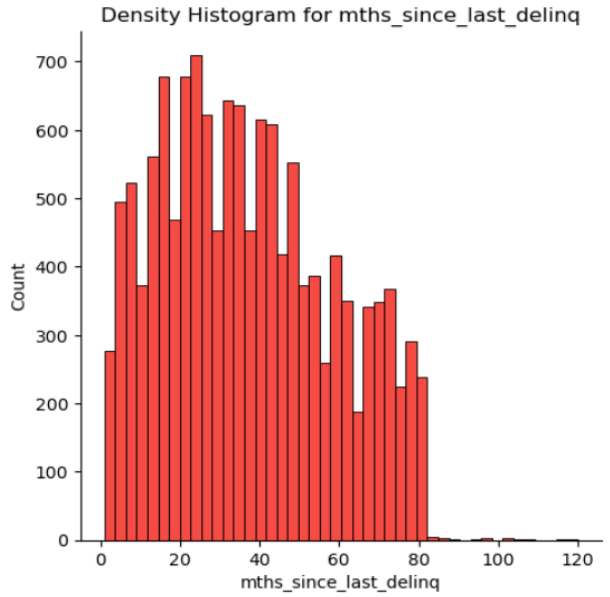
Plot for int_rate vs loan status : interest rate range between 11.26% and 16.4% quantiles. So compared to normal customer [which has q3 : 14.59%] the customers are likely to default according to afore mentioned quantiles.



Bivariate Analysis. Continued.

Plot for mths_since_last_delinq vs loan status

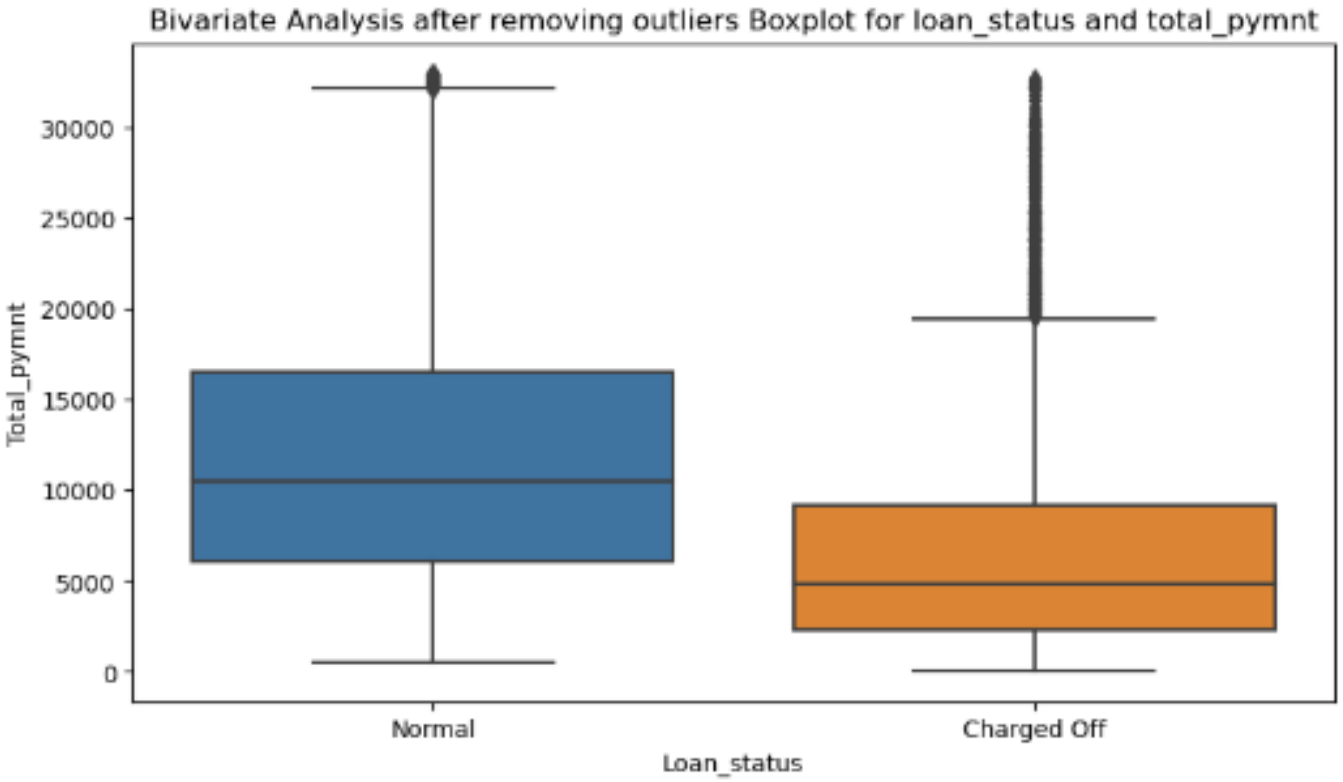




From above chat mths_since_last_delinq is highest between 20 to 40.

Bivariate Analysis. Continued.

Plot for total_pymnt vs loan status



Quantile details for Charged Off Borrower's has total_pymnt ranging between Q1 = 2271.96 and Q3 = 9158.71

Quantile details for all Borrower's with total_pymnt ranging between Q1 = 5594.00 and Q3 = 16553.74

Inference from Bivariate Analysis

Borrowers who default on their loans tend to have higher interest rates, moderate number of open accounts, and lower total payments and investments, and they often have a delinquency in the past two to three years.

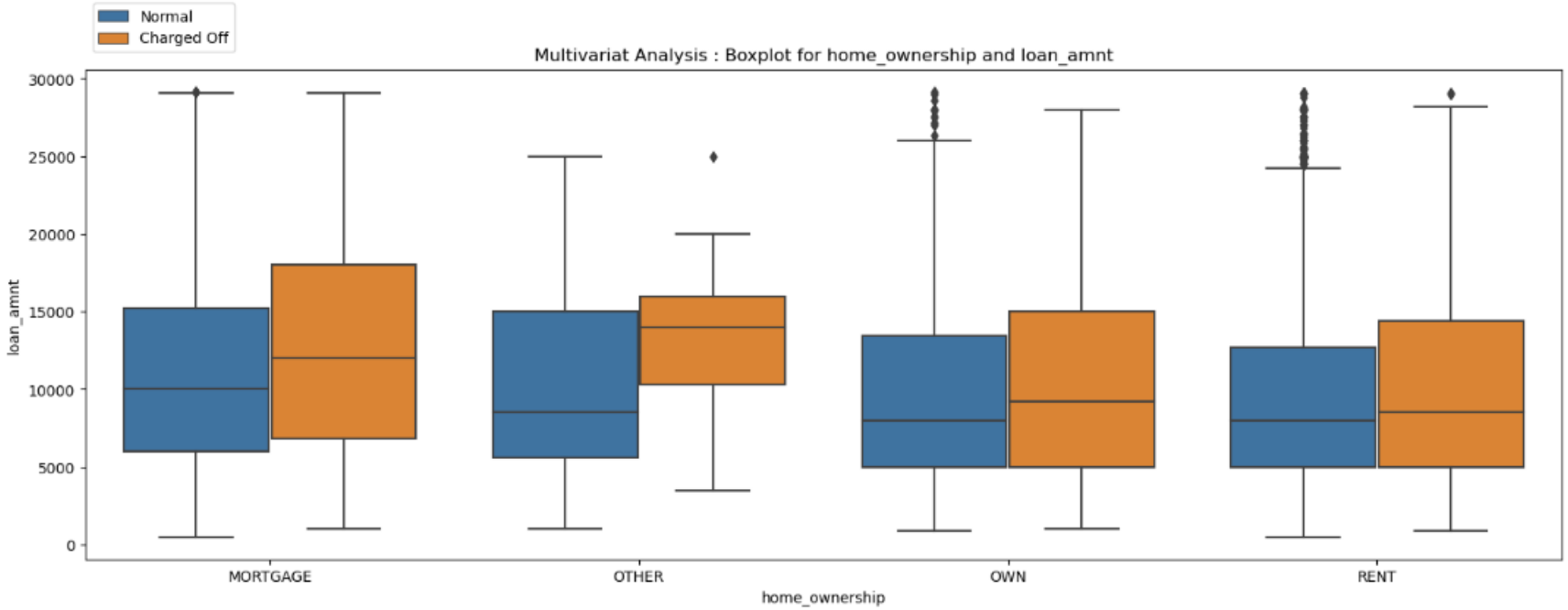
Multivariate Analysis

Multivariate analysis, was performed on combination of loan amount , home ownership , term , verification status (other variables) against following variables :

- purpose
- grade
- annual income
- int_rate
- mths_since_last_delinq

Multivariate Analysis : Home ownership vs other variables

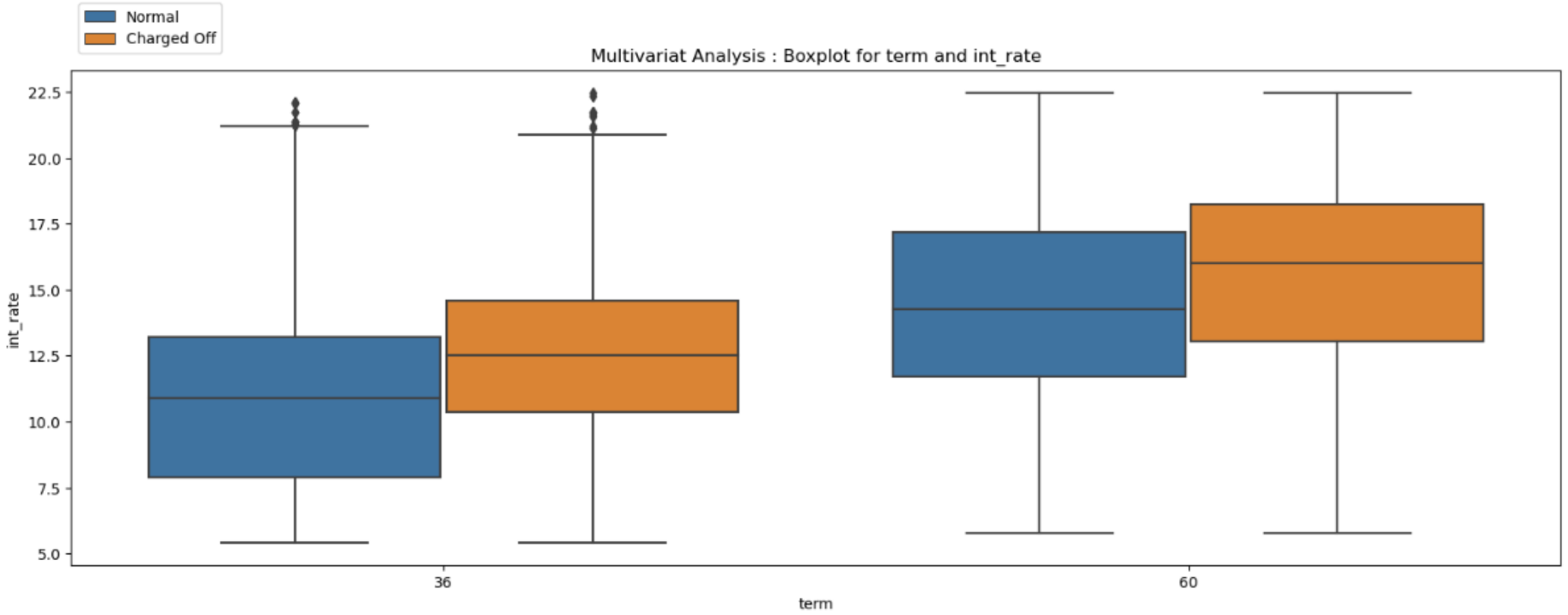
- A MORTGAGED borrower with annual_inc between 47508.0 and 84000.0 is likely to default.
- A MORTGAGED borrower with loan_amnt range 6800.0 and 18000.0 is likely to default.
- High interest rates between 11.03% and 16.29% affect all borrowers with different home ownership statuses equally, leading them to default.



- Borrowers with home_ownership status as OTHER and mths_since_last_delinq ranges between 0.0 and 22.0 are more likely to default.
- The percentage of default for mortgage and others are similar, which is higher as compared to percentage of defaults by borrowers who own the house or rent the houses.

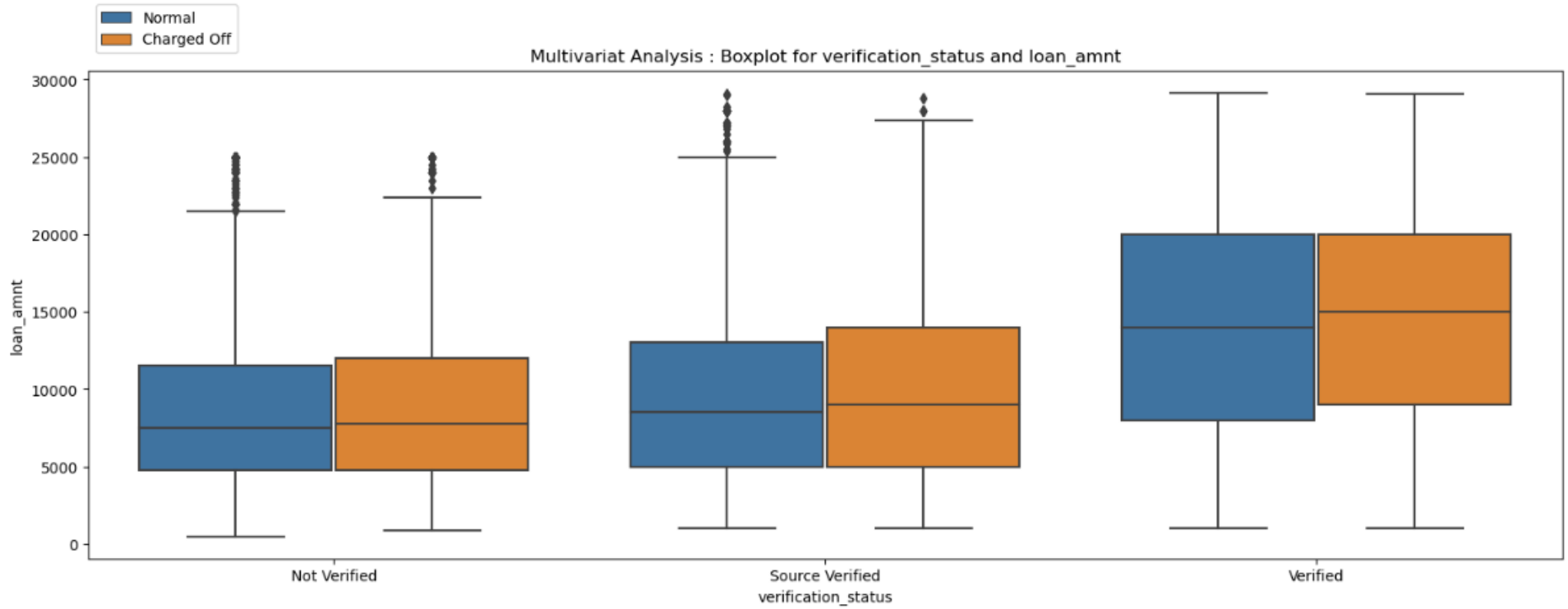
Multivariate Analysis : Term vs other variables

- While considering loan amount, likely hood of default for a borrower with term as 60 months is higher compared to borrower with term as 36 month.
- Borrowers who have a 36-month term tend to default at lower interest rates (10.37% to 14.59%) than borrowers who have a 60-month term, who start defaulting at interest rates between 13.06% and 18.25%.
- Borrowers with term as 60 months are more likely to pay total payment compared to default.



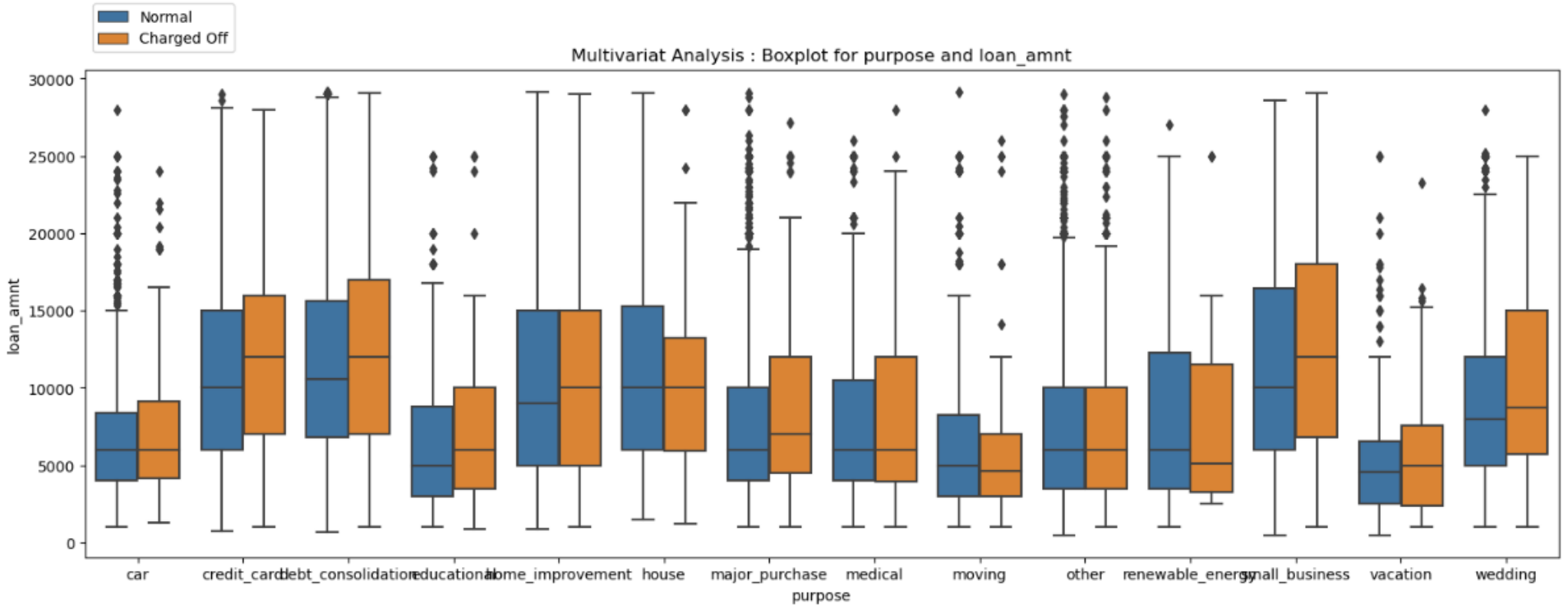
Multivariate Analysis Verification Status vs other variables

- Verification status does not seem to have a strong influence on the likelihood of defaulting, as borrowers who default have similar incomes regardless of their verification status.
- Verification status may be related to the amount of loan requested and granted, as borrowers who are verified tend to have higher loan amounts than those who are not verified, or source verified.
- Verification status may also be related to the amount of installment and total payment, as borrowers who are verified tend to have higher installments and total payments than those who are not verified, or source verified.
- Borrowers who are verified may have higher financial needs or expectations, which may lead them to borrow more and pay more, but also increase their risk of defaulting if they cannot meet their obligations.



Multivariate Analysis : purpose vs loan amount/others

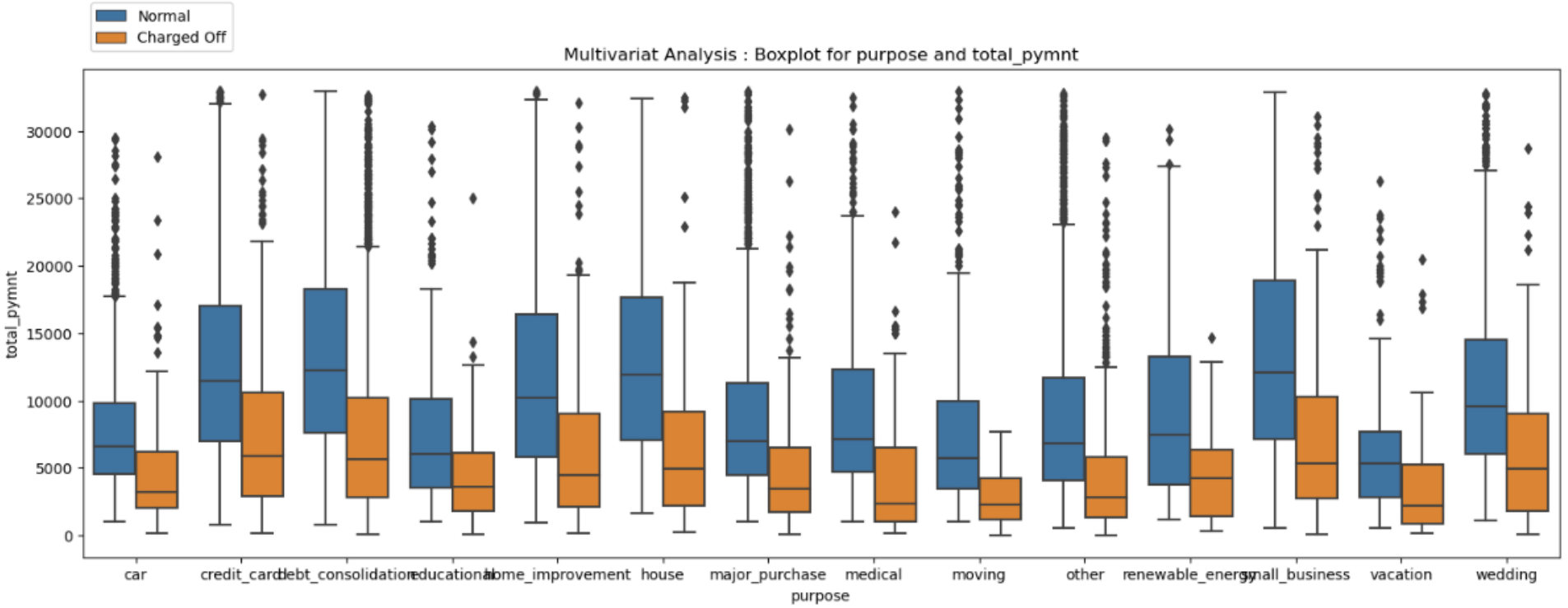
Purpose vs Loan amount/installment/int_rate/dti: Borrowers who default on their loans have different loan amount ranges depending on their purpose of borrowing, and those who borrow for debt consolidation, credit card, or small business have the highest loan amounts, while those who borrow for vacation, moving, or renewable energy have the lowest loan amounts.



Please refer to backup slides for multivariate analysis for purpose vs mths_since_last_delinq.

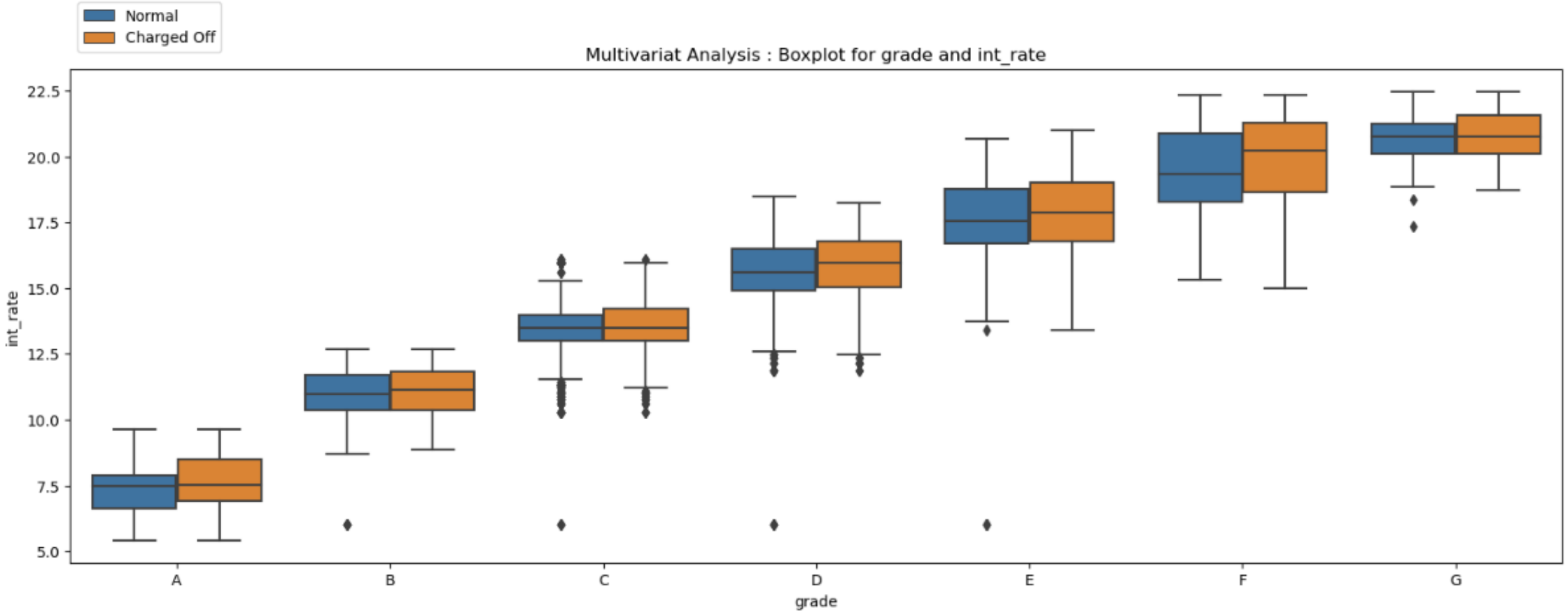
Multivariate Analysis : **purpose vs total_pymnt**

Purpose vs total_pymnt : Borrowers who default on their loans have different total payment ranges depending on their purpose of borrowing, and those who borrow for debt consolidation, credit card, or small business have the highest total payments, while those who borrow for vacation, moving, or medical have the lowest total payments.



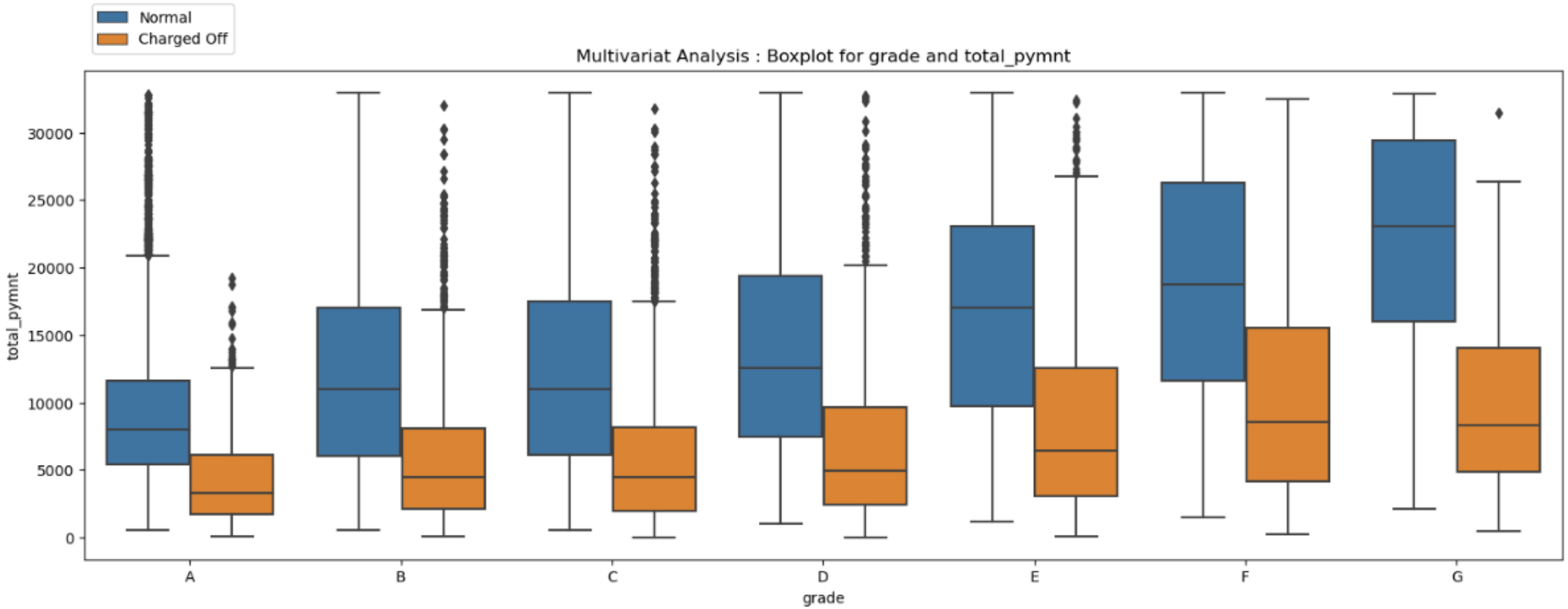
Multivariate Analysis : grade vs int_rate

Grade vs int_rate: Borrowers who default on their loans have different interest rate ranges depending on their grade, and those who have lower grades (E, F, G) have higher interest rates than those who have higher grades (A, B, C, D).



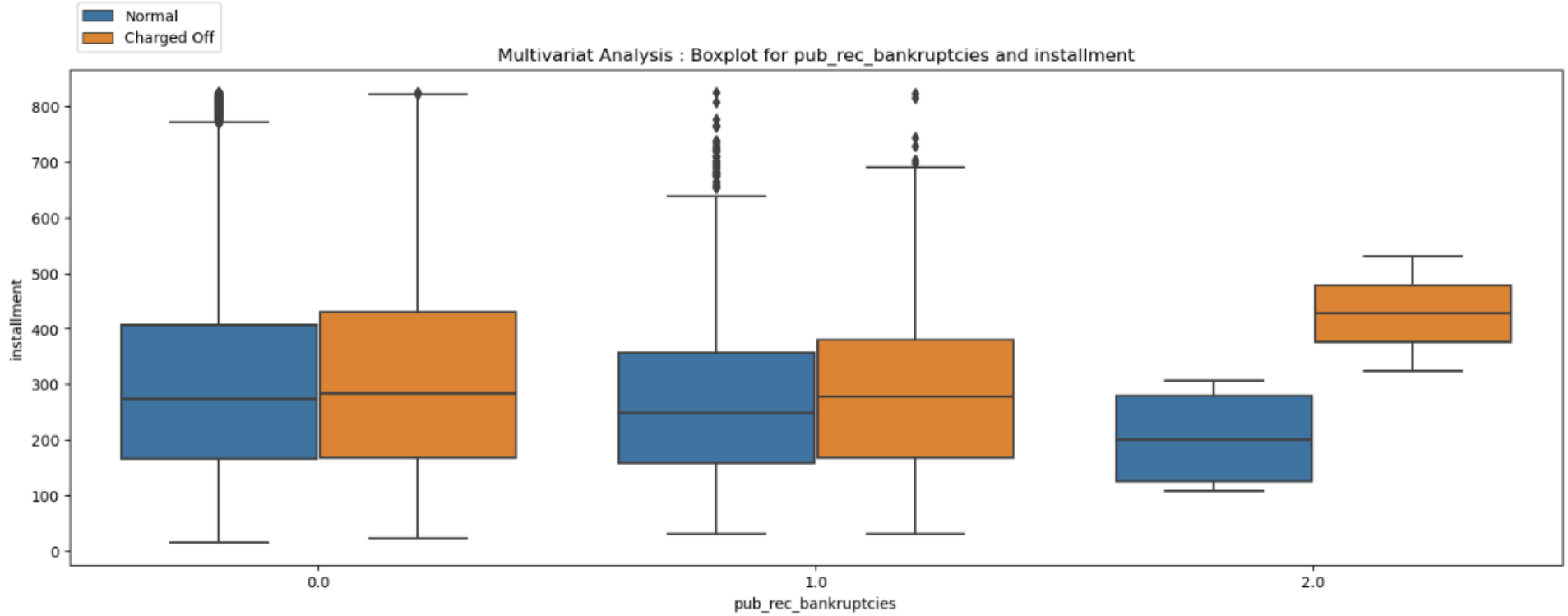
Multivariate Analysis : grade vs total_pymnt

Grade vs total_pymnt : Borrowers who default on their loans have different total payment ranges depending on their grade, and those who have lower grades (E, F, G) have higher total payments than those who have higher grades (A, B, C, D).



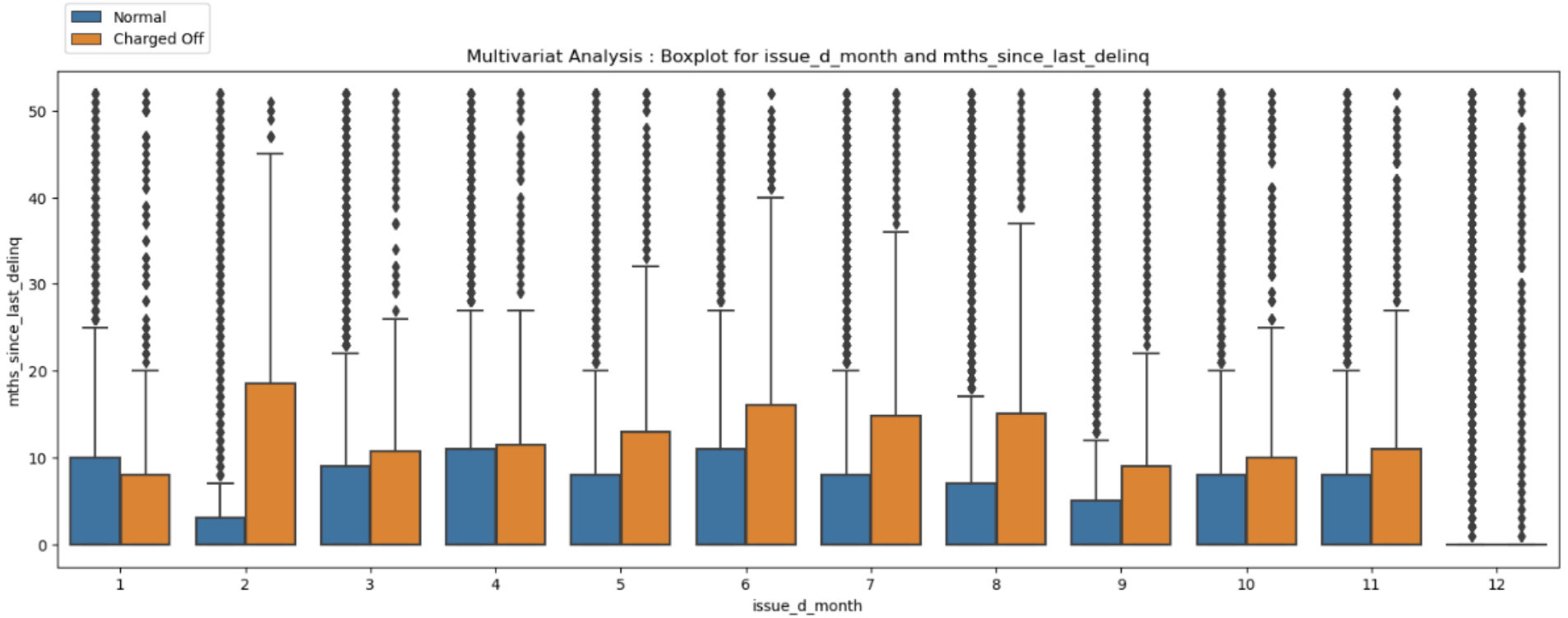
Multivariate Analysis : pub_rec_bankruptcies vs installments

pub_rec_bankruptcies vs installments: Borrowers who default on their loans have different installment ranges depending on their public record bankruptcies, and those who have more bankruptcies (2.0) have higher installments than those who have fewer bankruptcies (0.0 or 1.0).



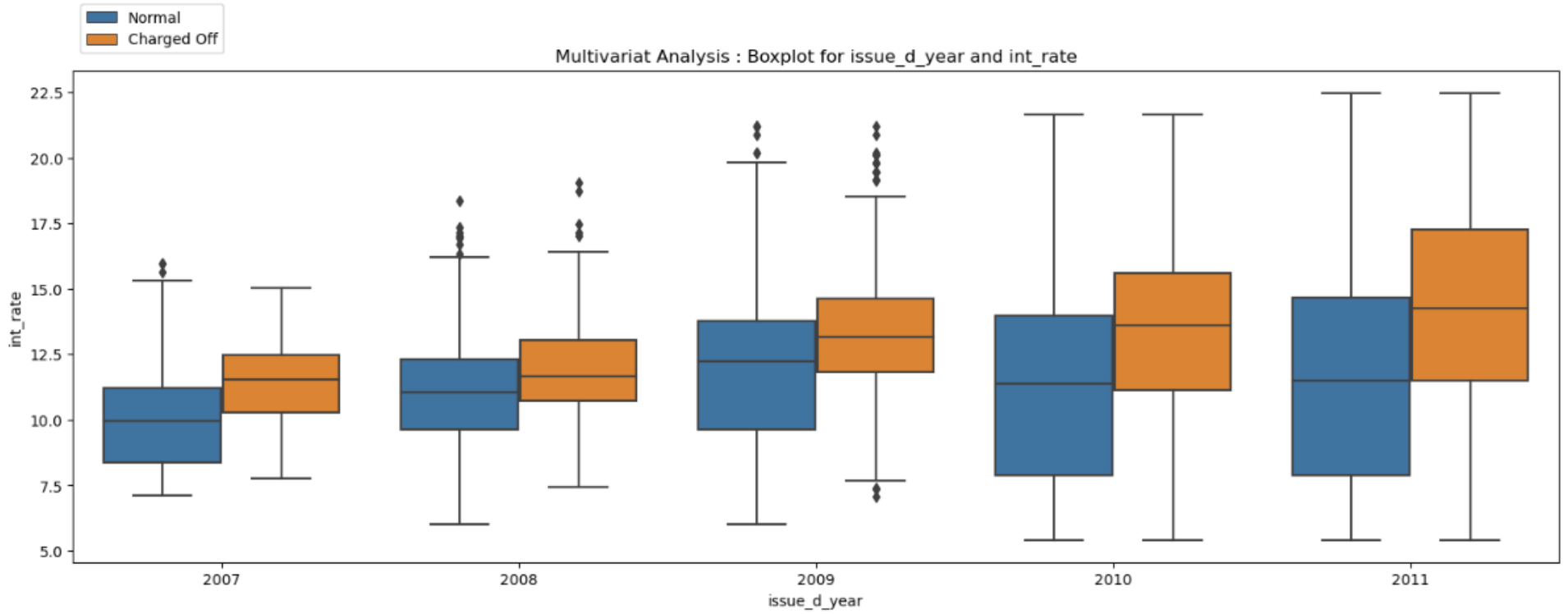
Multivariate Analysis : mths_since_last_delinq vs issue_d_month

mths_since_last_delinq vs issue_d_month: Borrowers who default on their loans have different months since last delinquency ranges depending on their issue month, and those who have issued loans in February (2) or June (6) have the longest delinquency history, while those who have issued loans in December (12) or January (1) have no delinquency history.



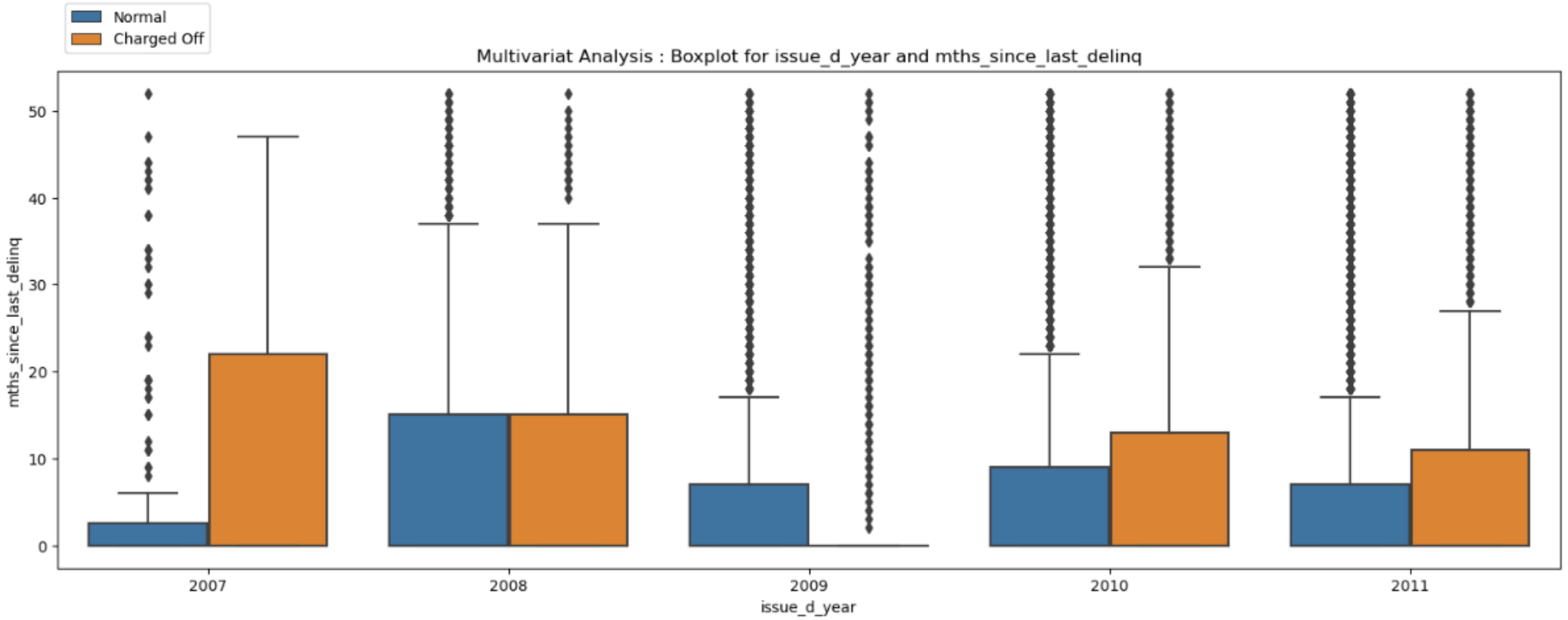
Multivariate Analysis : int_rate vs issue_d_year

Int_rate vs issue_d_year: Borrowers who default on their loans have different interest rate ranges depending on their issue year, and those who have issued loans in 2011 have the highest interest rates, while those who have issued loans in 2007 have the lowest interest rates.



Multivariate Analysis : mths_since_last_delinq vs issue_d_year

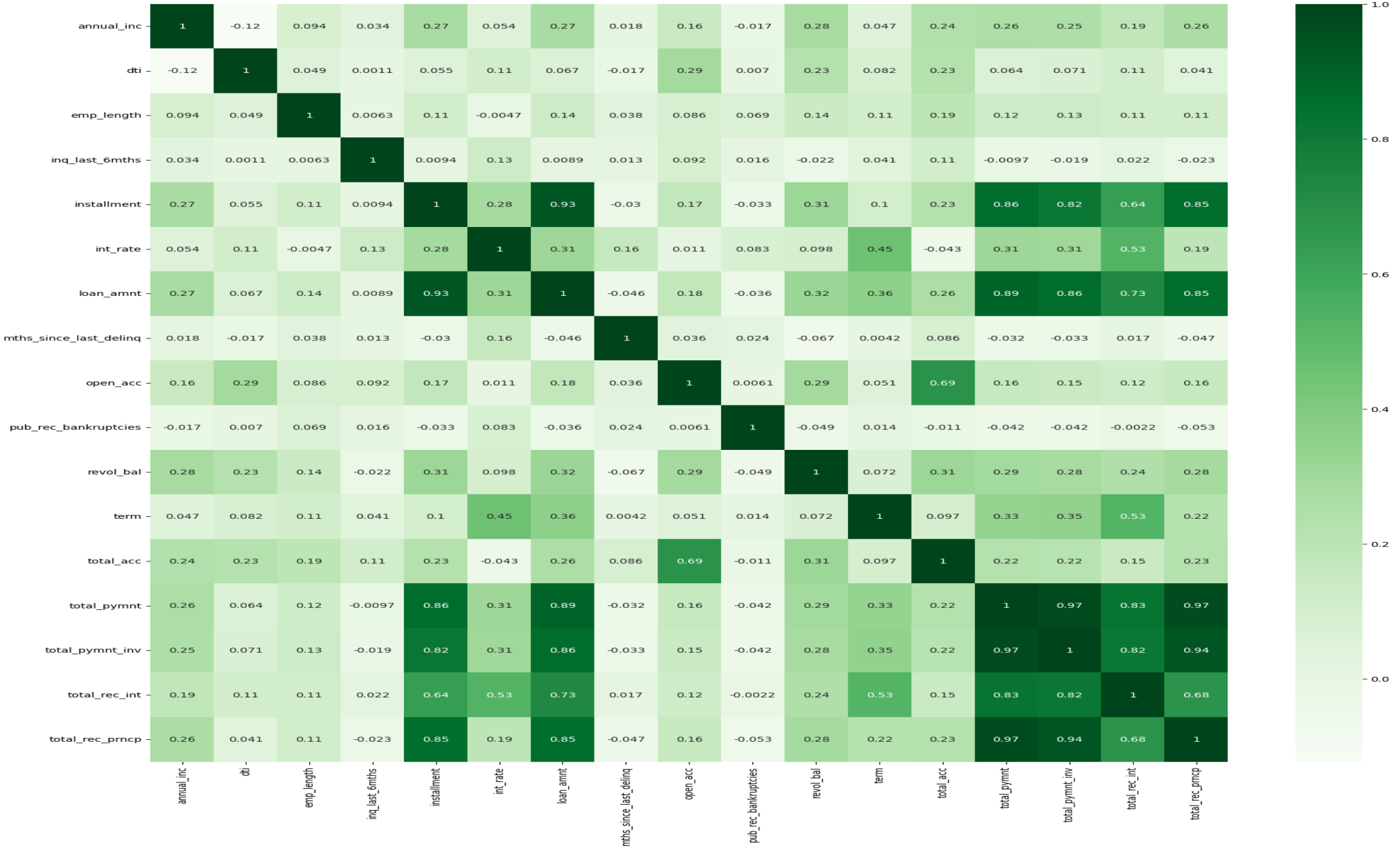
mths_since_last_delinq vs issue_d_year: The column mths_since_last_delinq in loan data shows the number of months since the borrower's last delinquency. In the given lines, it shows that borrowers who were charged off in 2011 had a range of 0.0 to 11.0 months since their last delinquency, borrowers charged off in 2010 had a range of 0.0 to 13.0 months since their last delinquency, borrowers charged off in 2009 had no delinquencies recorded for them and borrowers charged off in 2008 had a range of 0.0 to 15.0 months since their last delinquency.



Inferences from Multivariate Analysis (important pointers captured)

- Loan amount vs term: Borrowers with a 60-month term are more likely to default than those with a 36-month term.
- Interest rate vs term: Borrowers with a 36-month term default at lower interest rates than those with a 60-month term.
- Annual income vs verification status: Borrowers who default have similar incomes regardless of verification status, but verified borrowers have slightly higher incomes.
- Loan amount vs verification status: Borrowers who default have different loan amounts depending on verification status, and verified borrowers have the highest loan amounts.
- Annual income vs purpose: Borrowers who default have different incomes depending on purpose, and those who borrow for home improvement or house purchase have the highest incomes.
- Loan amount vs purpose: Borrowers who default have different loan amounts depending on purpose, and those who borrow for debt consolidation, credit card, or small business have the highest loan amounts.
- Grade vs loan amount: Borrowers who default have different loan amounts depending on grade, and those who have lower grades have higher loan amounts.
- Grade vs int_rate: Borrowers who default have different interest rates depending on grade, and those who have lower grades have higher interest rates.
- Grade vs total_pymnt: Borrowers who default have different total payments depending on grade, and those who have lower grades have higher total payments.
- Int_rate vs issue_d_year: Borrowers who default have different interest rates depending on issue year, and those who have issued loans in 2011 have the highest interest rates.

Correlation heat map



Inference from Correlation heatmap

The variables installment, loan_amnt, total_pymnt, total_pymnt_inv and total_rec_prncp have a high positive correlation with each other, which means that they are all related to the amount of money involved in the loan.

On the other hand, the variables annual_inc and dti have a high negative correlation, which means that as the annual income of the borrower increases, the debt-to-income ratio decreases.

Summary of Analysis

- **14%** of the borrowers defaulted on loan.
- **Only 57%** of loans were recovered.
- Out of fully recovered loans **17% is the** profit.
- The loan business is vulnerable to losses from defaults and low recoveries, and it needs to monitor its default rate closely and take preventive measures to keep it below **16.49%.**
- Borrowers who default on their loans tend to have higher interest rates, moderate number of open accounts, and lower total payments and investments, and they often have a delinquency in the past two to three years.
- Loan amount vs term: Borrowers with a 60-month term are more likely to default than those with a 36-month term.
- Interest rate vs term: Borrowers with a 36-month term default at lower interest rates than those with a 60-month term.
- Annual income vs verification status: Borrowers who default have similar incomes regardless of verification status, but verified borrowers have slightly higher incomes.
- Loan amount vs verification status: Borrowers who default have different loan amounts depending on verification status, and verified borrowers have the highest loan amounts.
- Annual income vs purpose: Borrowers who default have different incomes depending on purpose, and those who borrow for home improvement or house purchase have the highest incomes.

Summary of Analysis, continued..

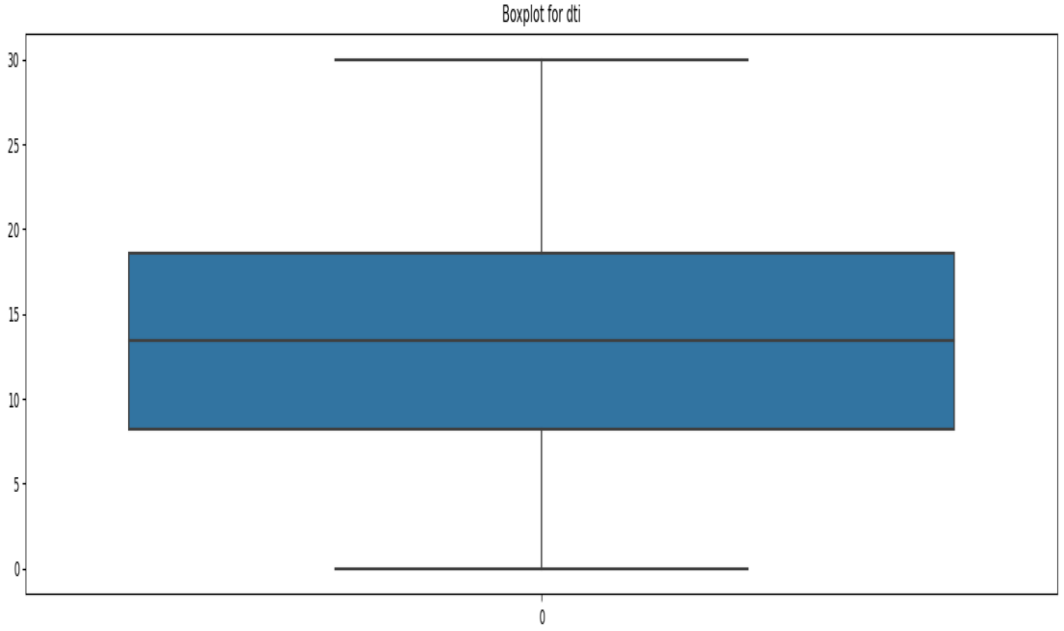
- Loan amount vs purpose: Borrowers who default have different loan amounts depending on purpose, and those who borrow for debt consolidation, credit card, or small business have the highest loan amounts.
- Grade vs loan amount: Borrowers who default have different loan amounts depending on grade, and those who have lower grades have higher loan amounts.
- Grade vs int_rate: Borrowers who default have different interest rates depending on grade, and those who have lower grades have higher interest rates.
- Grade vs total_pymnt: Borrowers who default have different total payments depending on grade, and those who have lower grades have higher total payments.
- Int_rate vs issue_d_year: Borrowers who default have different interest rates depending on issue year, and those who have issued loans in 2011 have the highest interest rates.

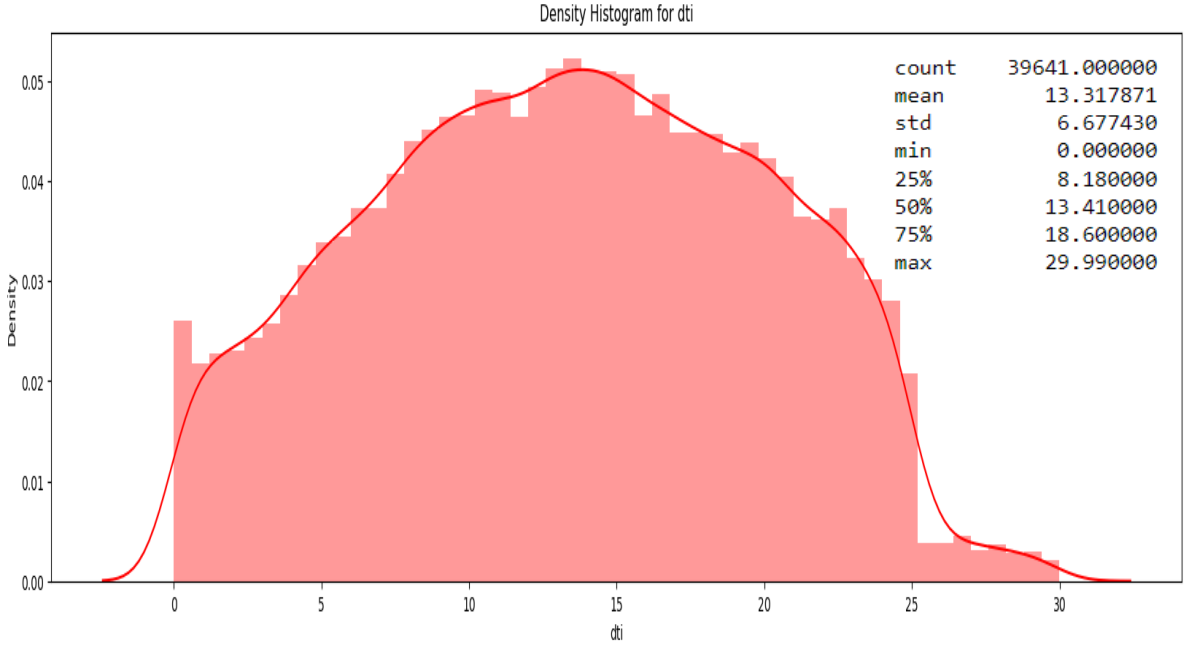
Thank you

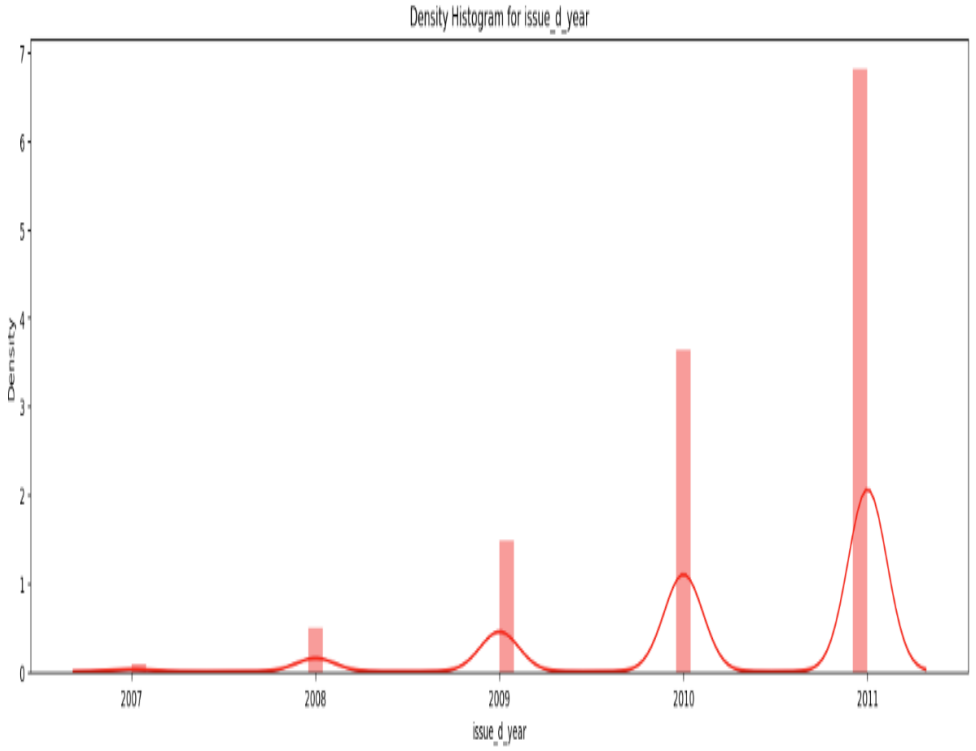
BackUp Slides For reference

Univariate Analysis : DTI ratio column analysis

Using dti ratio column : from analysis borrower's debt to Income ratio is above quantile 50% are more likely to default. Or borrowers with dti ranging from 13.41 to 18.6 are most likely to default.



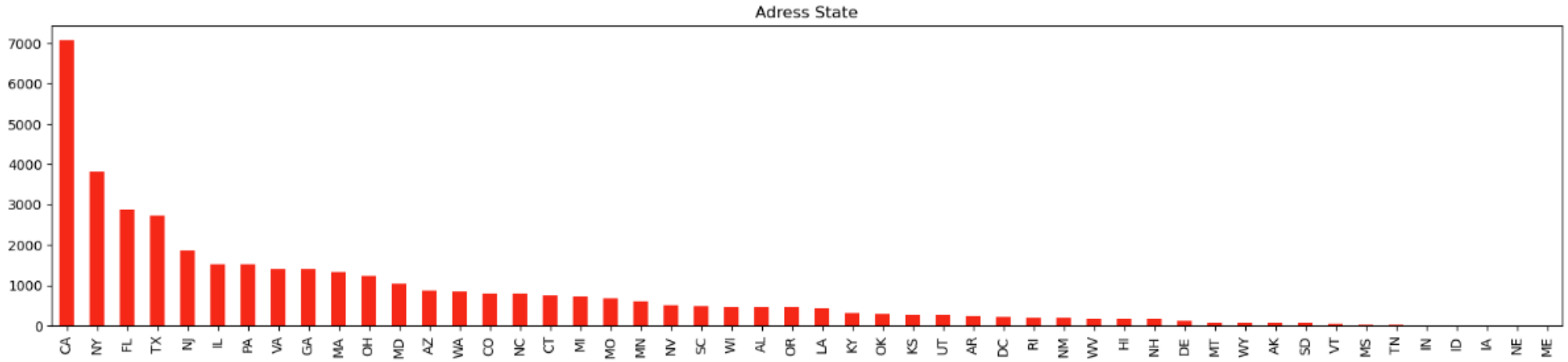




Univariate Analysis : Loan issued year column analysis

From chart the loan issue is increasing year by year.

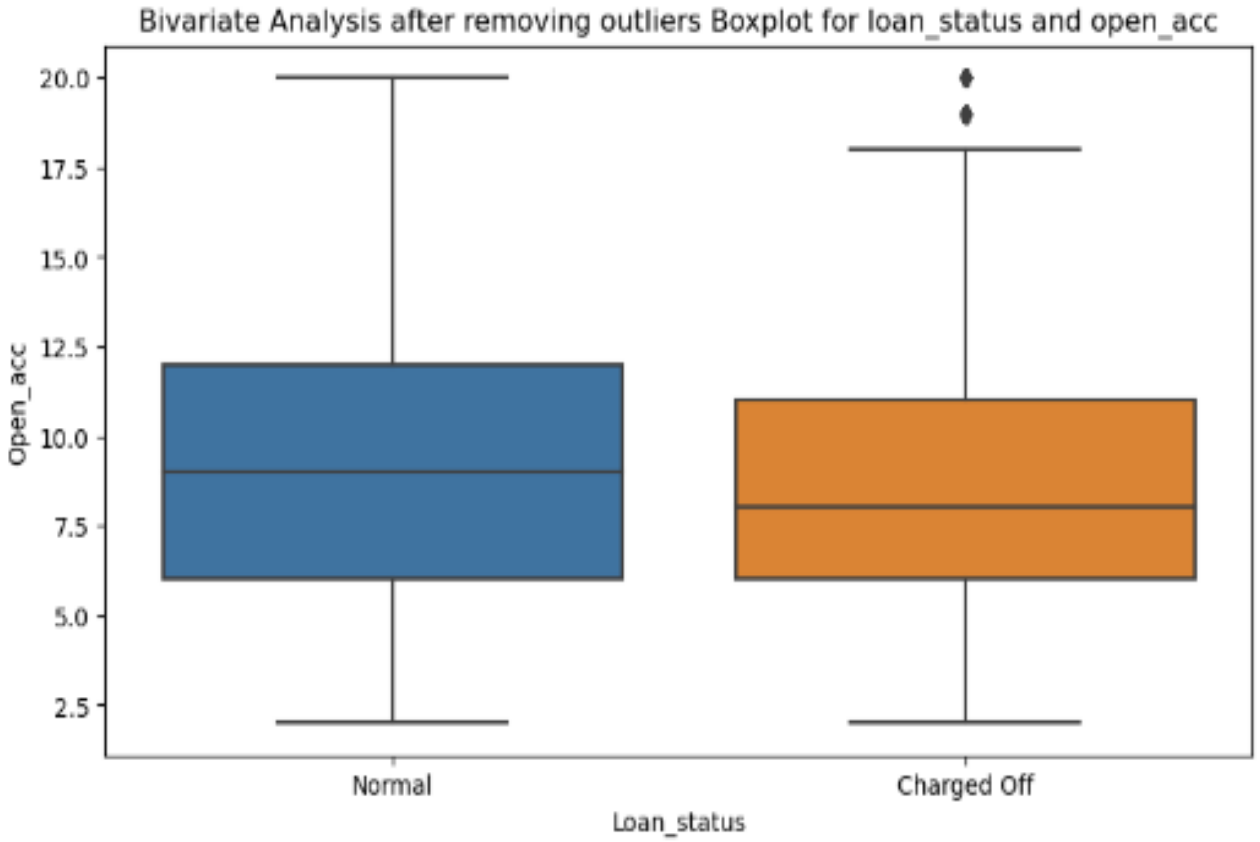
Univariate Analysis : States column analysis



State CA : has higher number of borrowers

Bivariate Analysis. Continued.

Plot for open_acc vs loan status

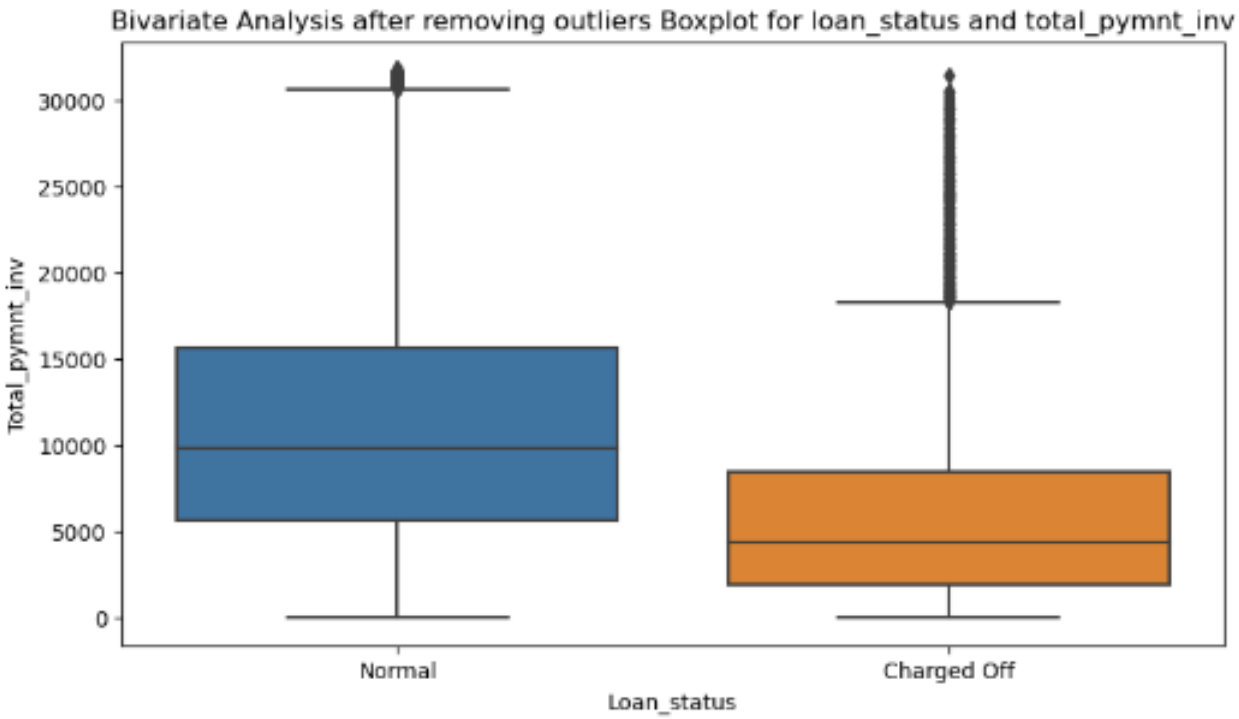


Quantile details for Charged Off Borrower's has open_acc ranging between Q1 = 6.0 and Q3 = 11.0

Quantile details for all Borrower's with open_acc ranging between Q1 = 6.0 and Q3 = 12.0

Bivariate Analysis. Continued.

Plot for total_pymnt_inv vs loan status

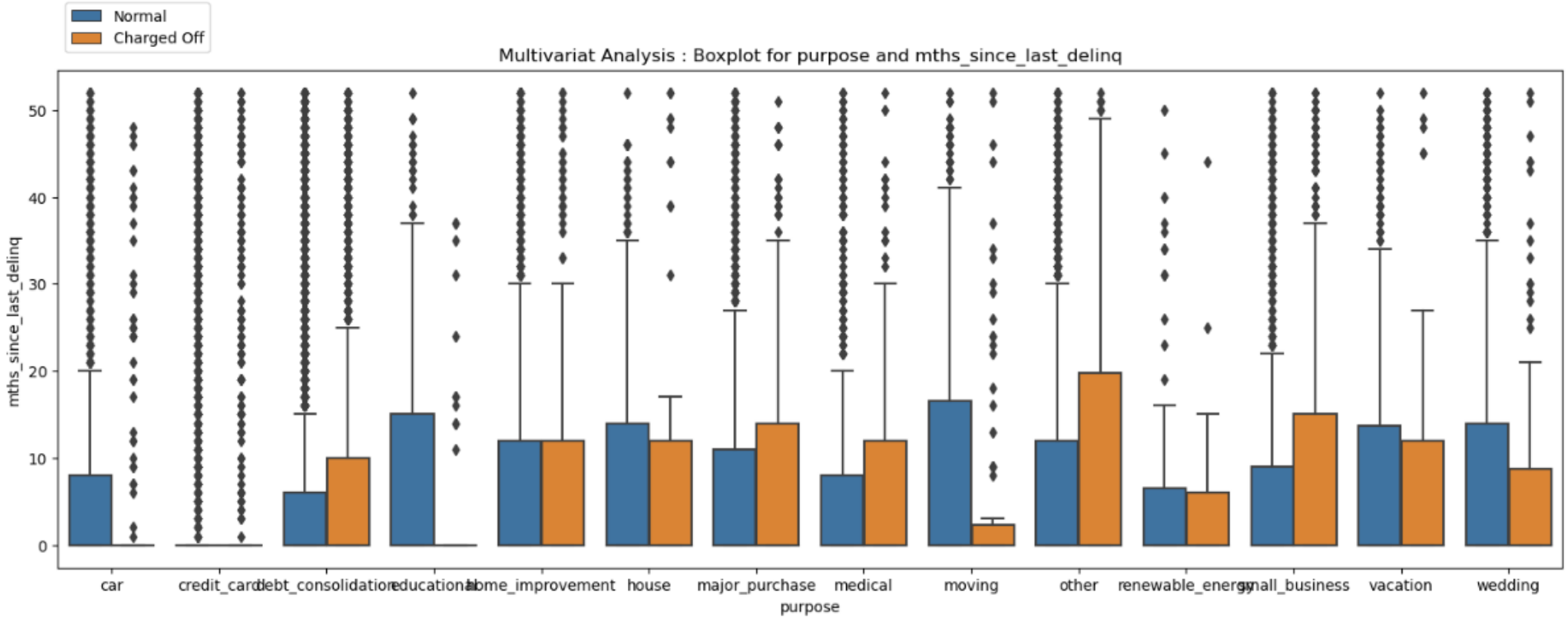


Quantile details for Charged Off Borrower's has total_pymnt_inv ranging between Q1 = 1914.79 and Q3 = 8485.97

Quantile details for all Borrower's with total_pymnt_inv ranging between Q1 = 5138.68 and Q3 = 15815.91

Multivariate Analysis : **mths_since_last_delinq vs purpose**

mths_since_last_delinq vs purpose: Borrowers who default on their loans have different months since last delinquency ranges depending on their purpose of borrowing, and those who borrow for other, major purchase, or small business have the longest delinquency history, while those who borrow for credit card, car, or educational have no delinquency history.



Multivariate Analysis : grade vs mths_since_last_delinq

Grade vs mths_since_last_delinq : Borrowers who default on their loans have different months since last delinquency ranges depending on their grade, and those who have higher grades (A, B) have no delinquency history, while those who have lower grades (C, D, E, F, G) have some delinquency history.