LENDING CLUB Case Study

Problem Statement**:**

Business Understanding

You work for a **consumer finance company** that specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must decide 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

The data given below contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns that 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.

In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

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

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

→ **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

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*)

Business Objectives

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.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (*called credit loss*). 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. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one can identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicant's using EDA is the aim of this case study.

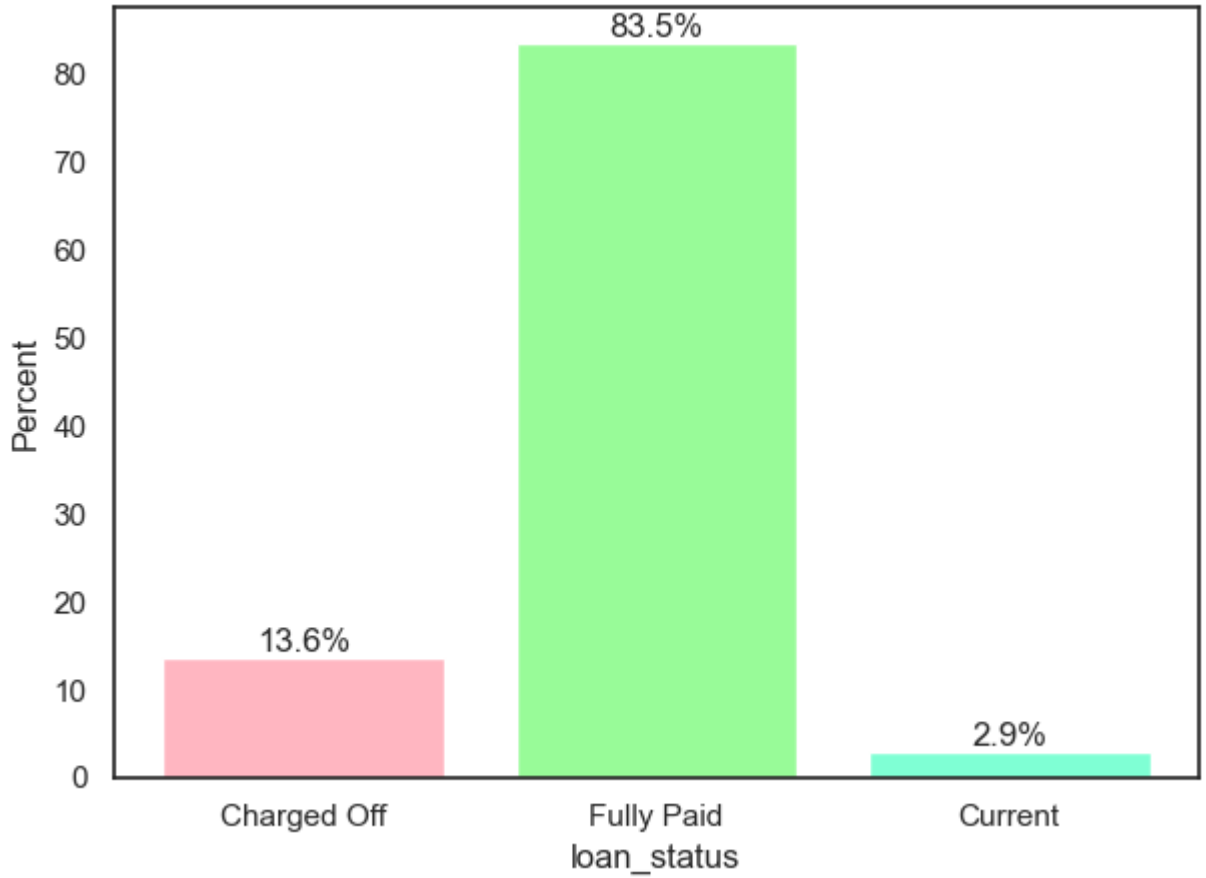
In other words, the company wants to understand the **driving factors** (***or driver variables***) behind loan default, i.e. the variables that are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics (*understanding the types of variables and their significance should be enough*).

Exploratory Data Analysis

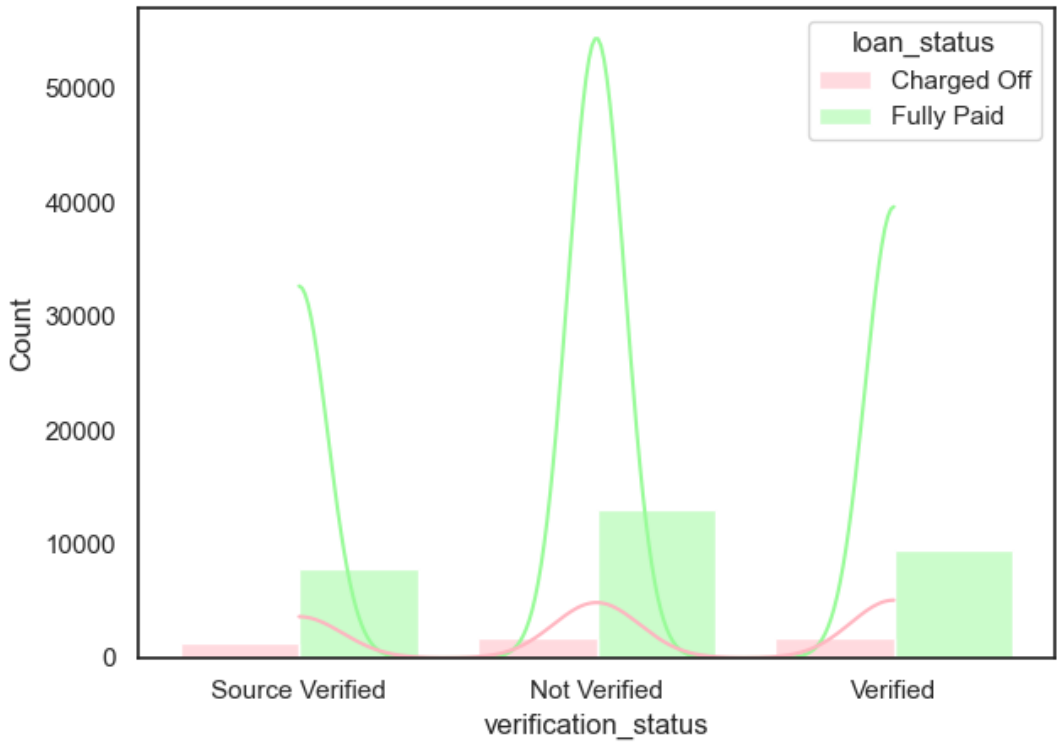
Univariate Analysis

Plot for **loan_status**



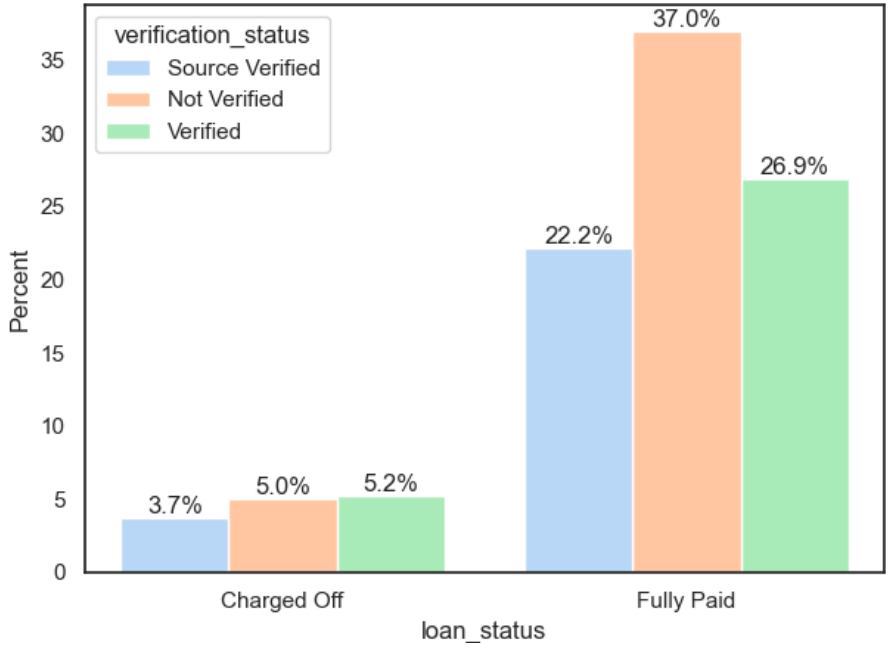
→ ***13.6%*** of loans are *'Charged-Off'*
→ ***83.5%*** of loans are *'Fully-Paid'*

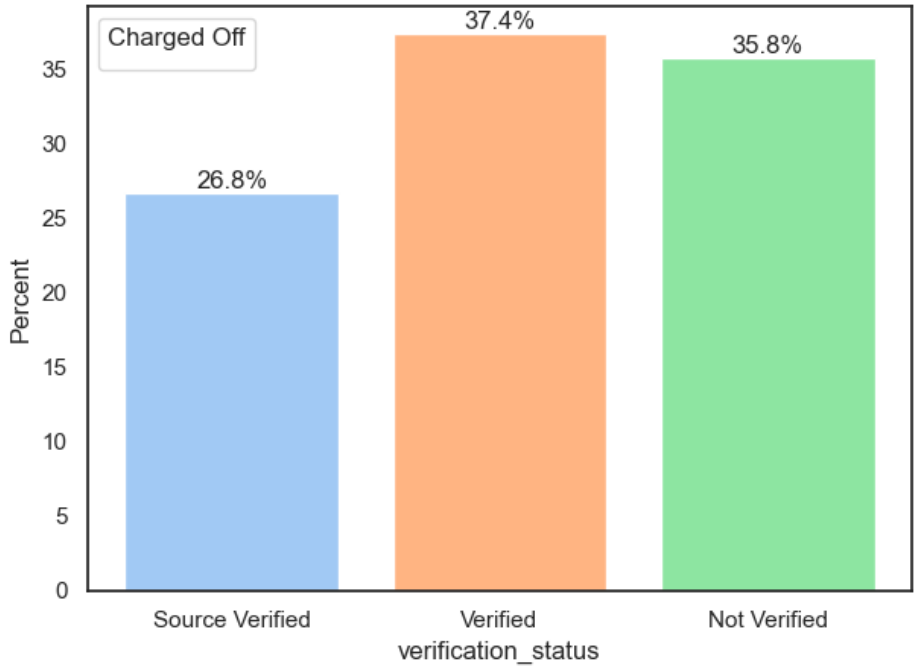
Plots for **verification_status**



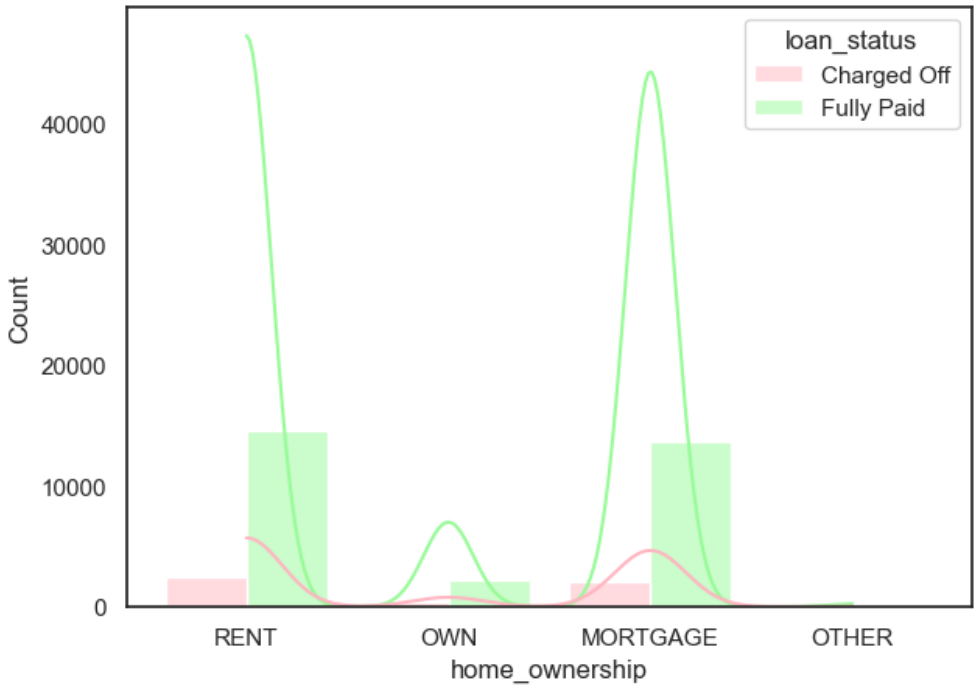
→ ***37.0%*** of **Not Verified** loans against 'loan_status' are *Not-defaults*

→ ***37.4%*** of **Verified** loans against 'Charged Off' are *Defaults*



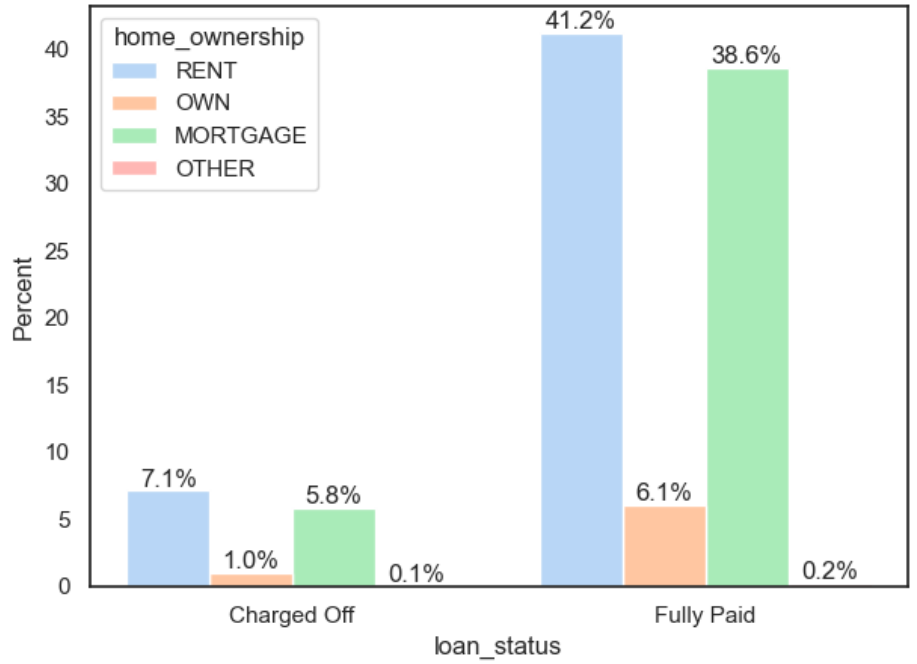


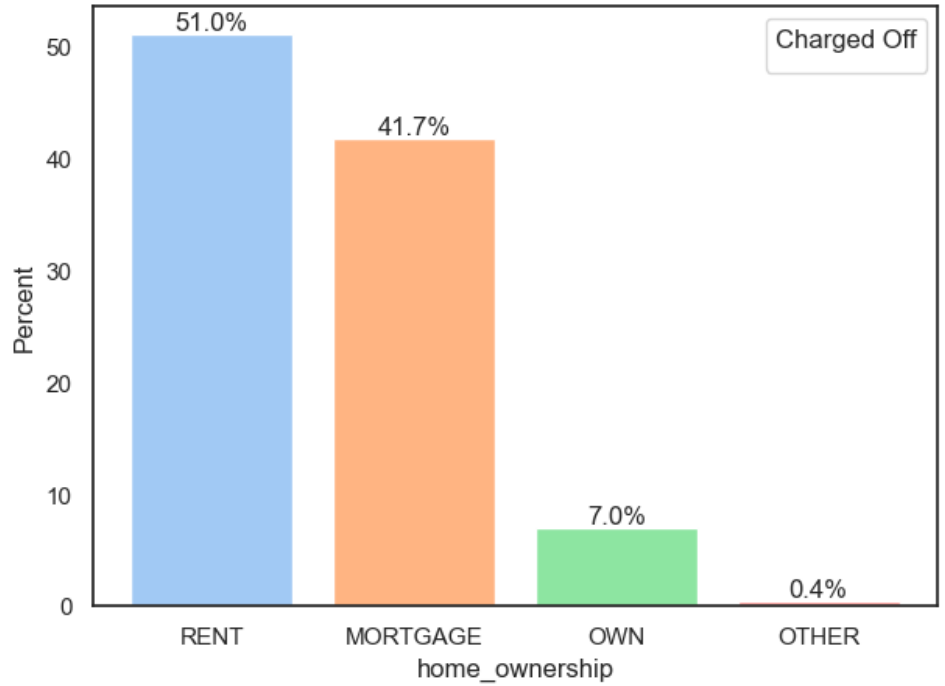
Plots for **home_ownership**



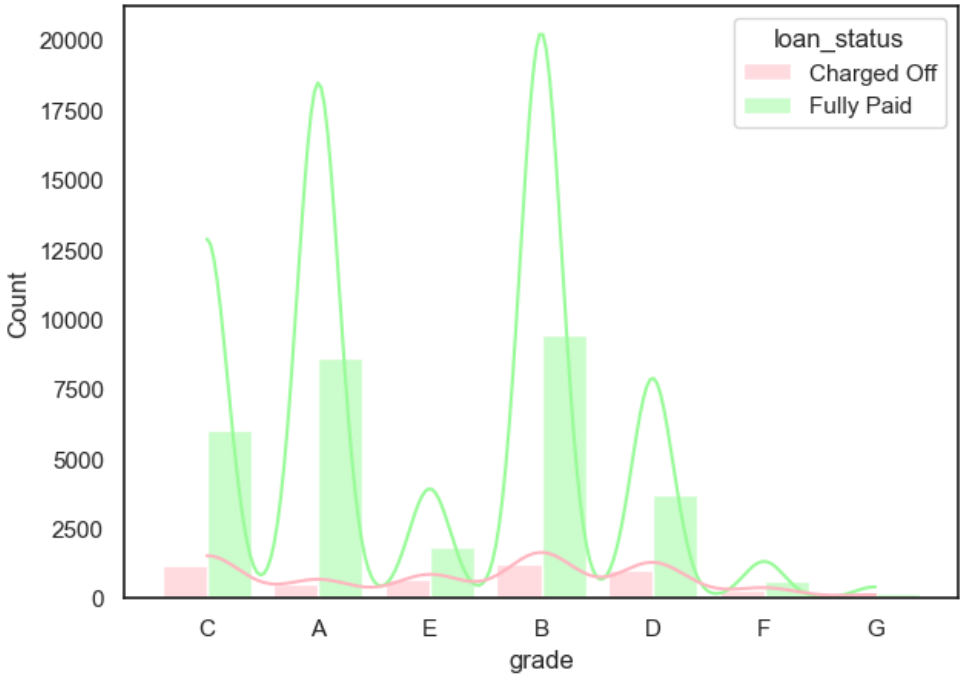
→ ***41.2%*** of **RENT** loans against 'loan_status' are *Not-defaults*
→ ***38.6%*** of **MORTGAGE** loans against 'loan_status' are *Not-defaults*

→ ***51.0%*** of **RENT** loans aginst 'Charged Off' are *Defaults*
→ ***41.7%*** of **MORTGAGE** loans aginst 'Charged Off' are *Defaults*



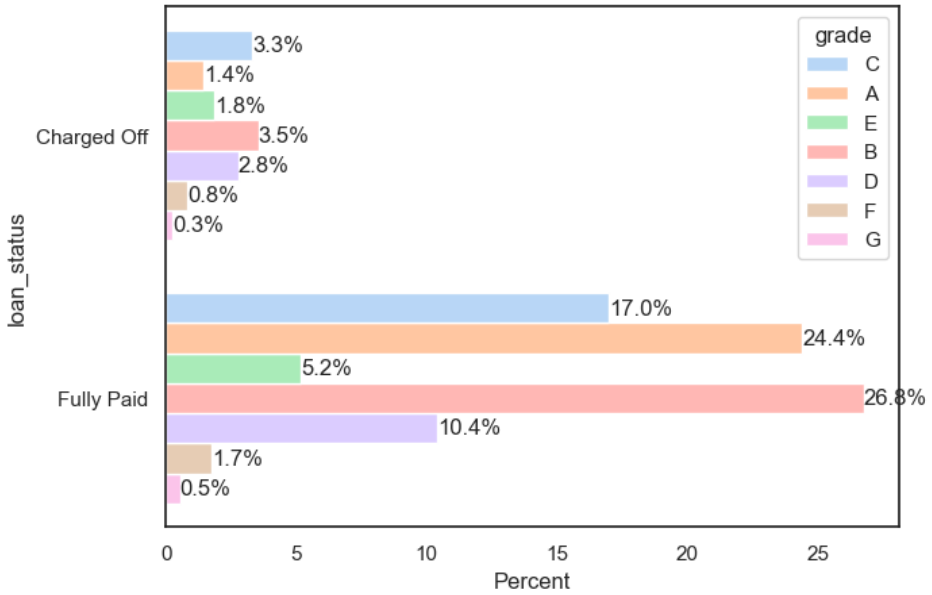


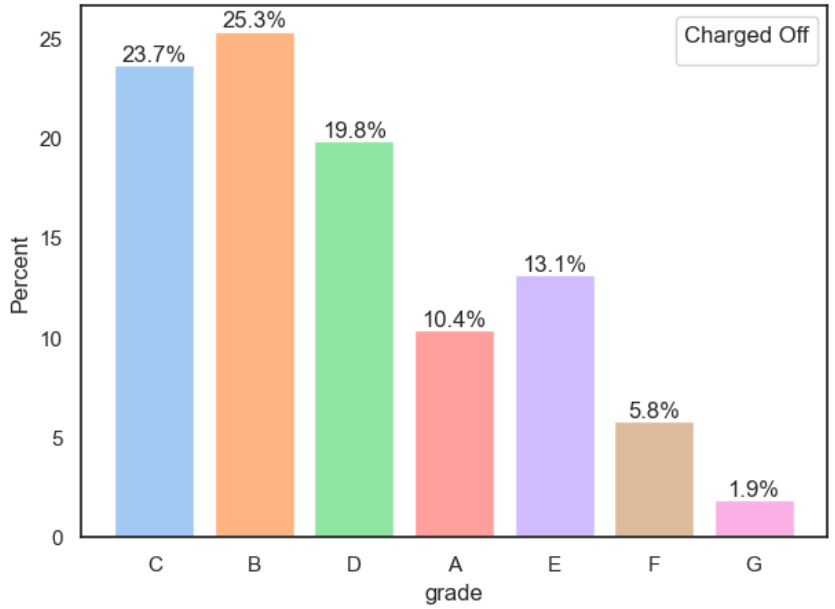
Plots for **grade**



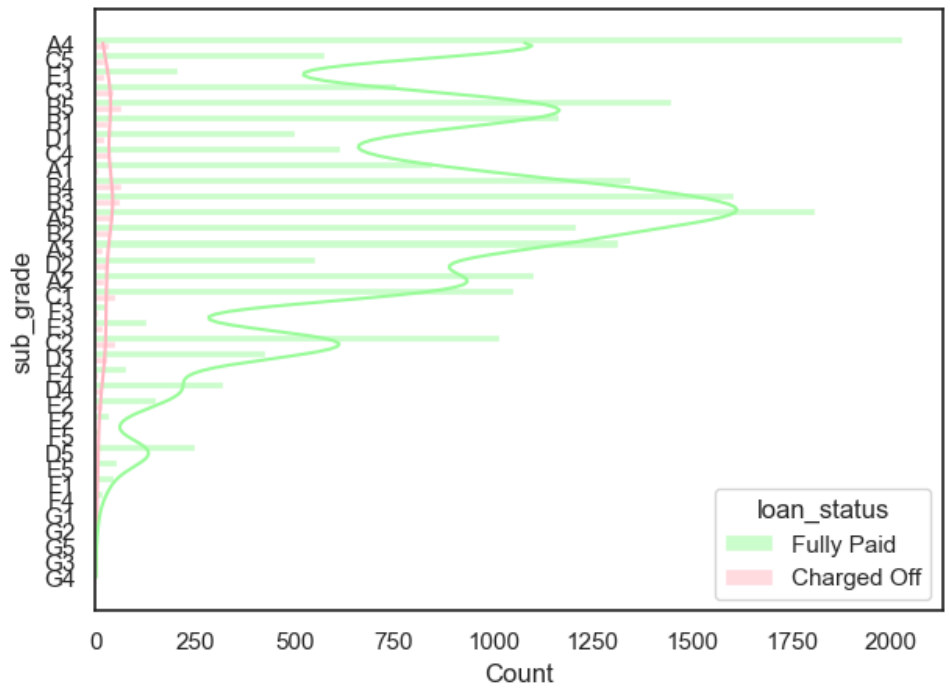
→ ***26.8%*** of 'grade' **B** loans against 'loan_status' are *Not-defaults*
→ ***24.4%*** of 'grade' **A** loans against 'loan_status' are *Not-defaults*

→ ***25.3%*** of 'grade' **B** loans against 'Charged Off' are *Defaults*
→ ***23.7%*** of 'grade' **C** loans against 'Charged Off' are *Defaults*
→ ***19.8%*** of 'grade' **D** loans against 'Charged Off' are *Defaults*



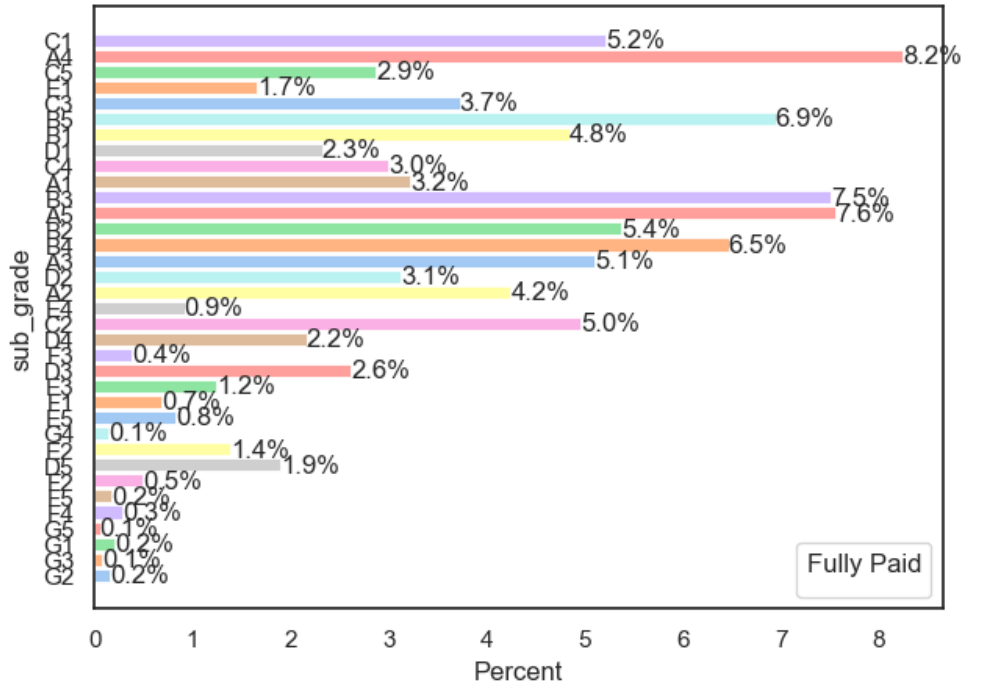


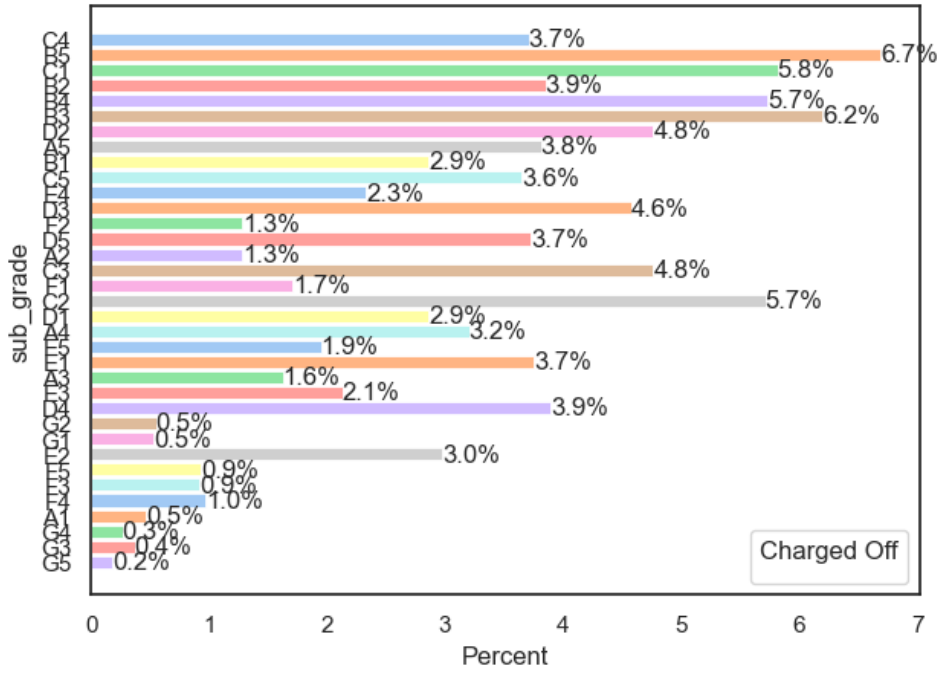
Plots for **sub_grade**



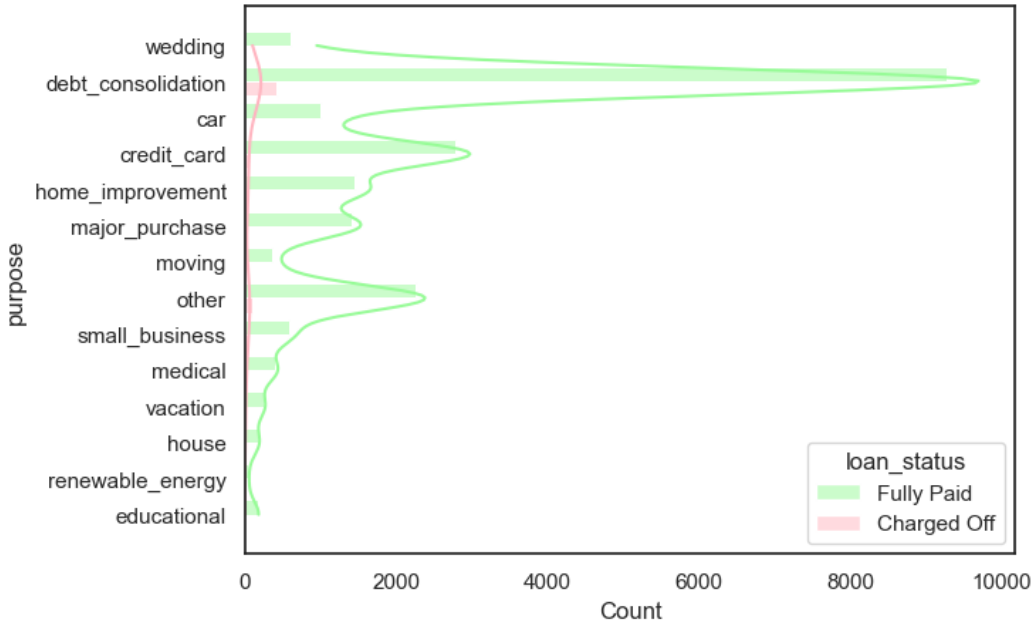
→ ***8.2%*** of 'sub_grade' **A4** loans against 'Fully Paid' are *Not-defaults*
→ ***7.6%*** of 'sub_grade' **A5** loans against 'Fully Paid' are *Not-defaults*
→ ***7.5%*** of 'sub_grade' **B3** loans against 'Fully Paid' are *Not-defaults*

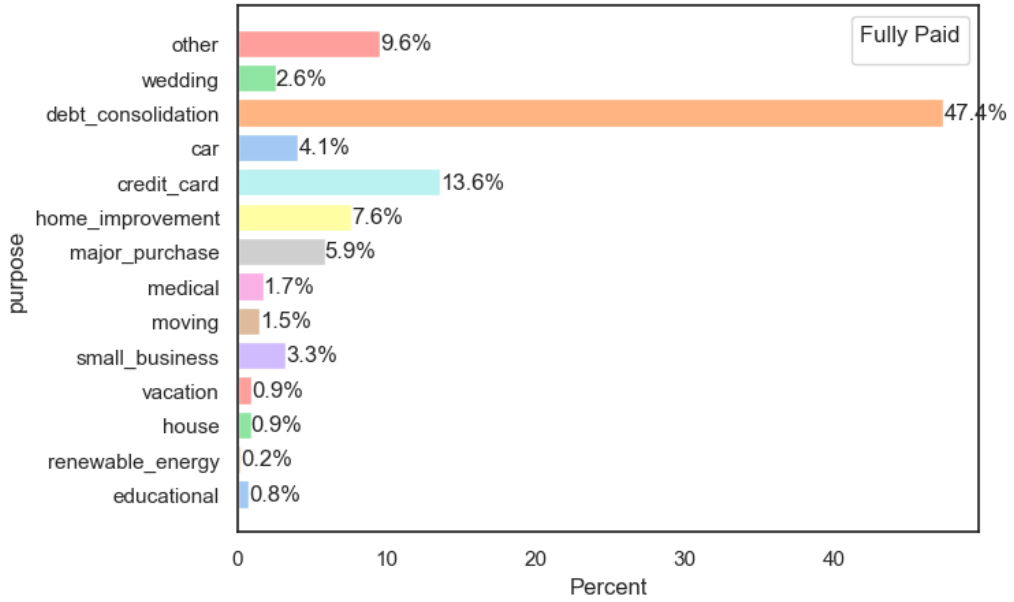
→ ***6.7%*** of 'sub_grade' **B5** loans against 'Charged Off' are *Defaults*
→ ***6.2%*** of 'sub_grade' **B3** loans against 'Charged Off' are *Defaults*
→ ***5.8%*** of 'sub_grade' **C1** loans against 'Charged Off' are *Defaults*
→ ***5.7%*** of 'sub_grade' **B4** & **C2** loans against 'Charged Off' are *Defaults*

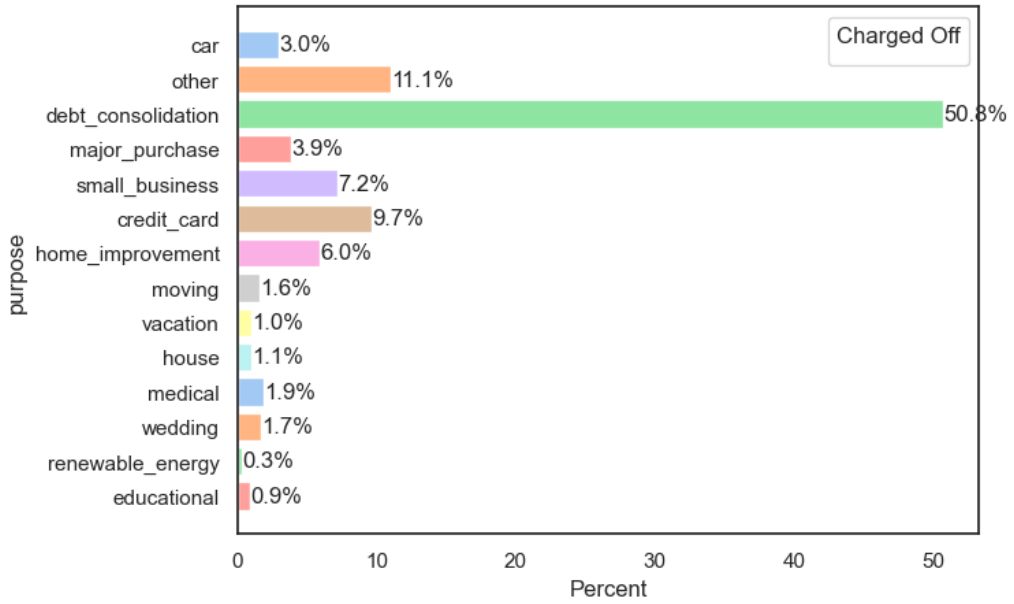




Plots for **purpose**



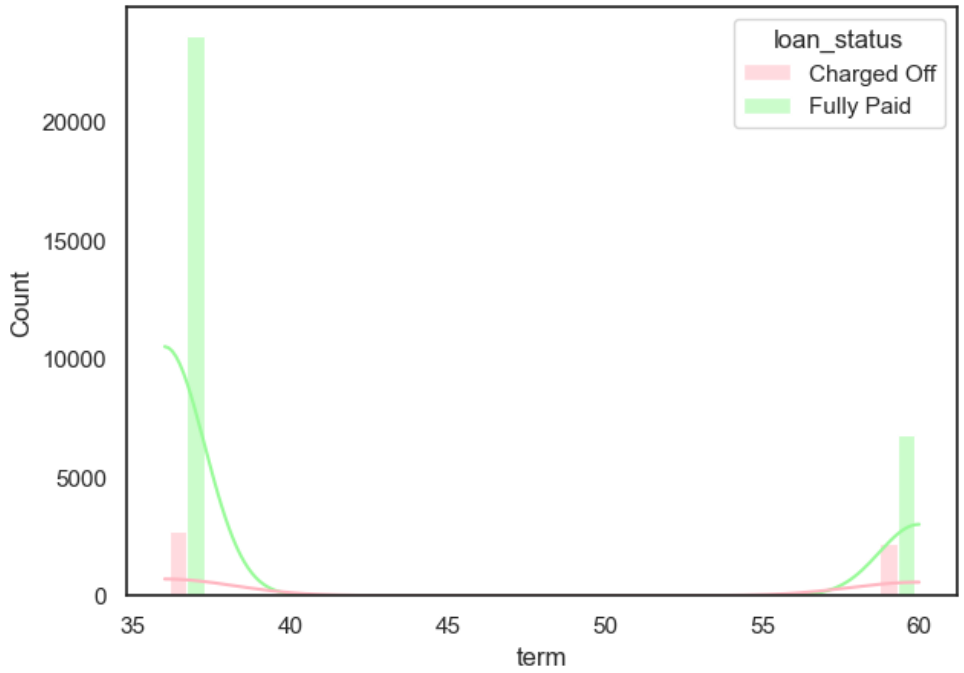




→ ***47.4%*** of **debt_consolidation** loans against 'Fully Paid' are *Not-defaults*

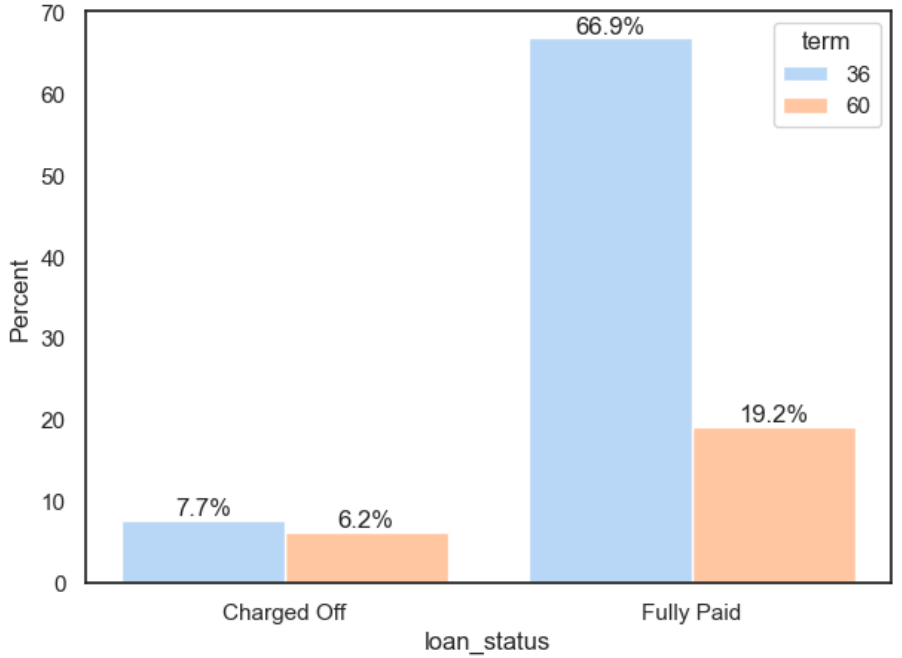
→ ***50.8%*** of **debt_consolidation** loans against 'Charged Off' are *Defaults*

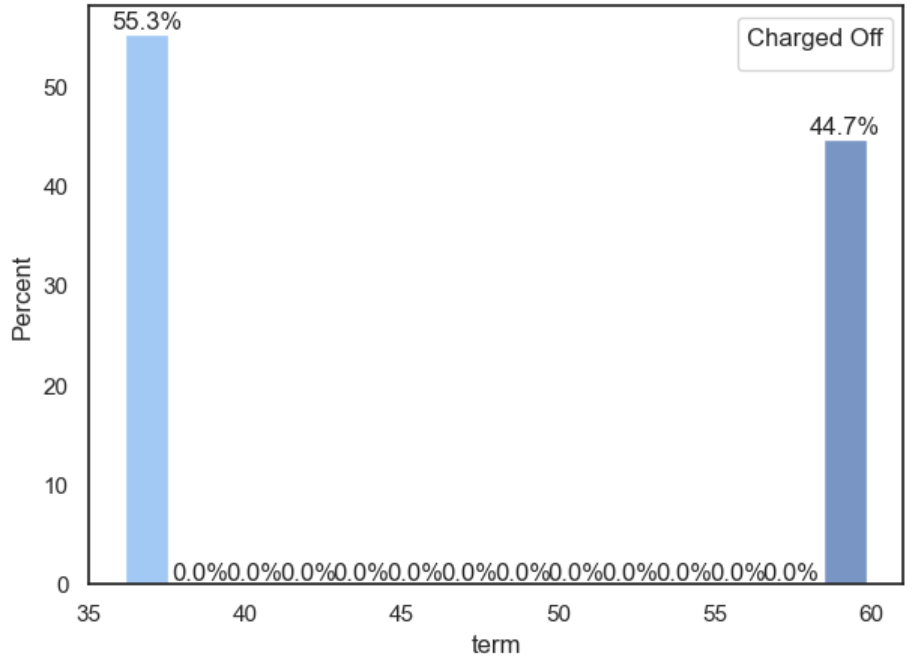
Plots for **term**



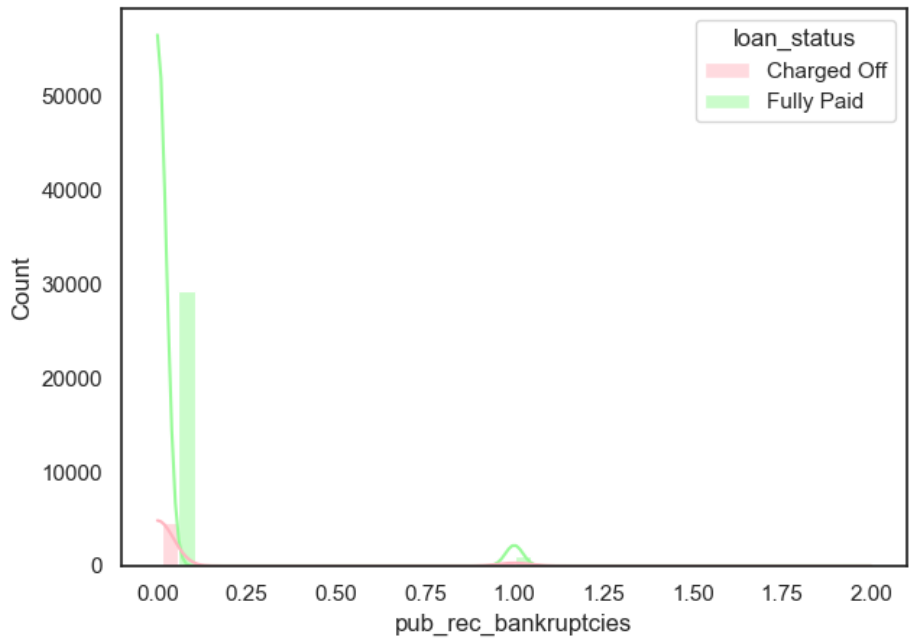
➔ ***66.9%*** of SHORT 'term' **36** *months* loans against 'loan_status' are *Not-defaults*
➔ ***19.2%*** of LONG 'term' **60** *months* loans against 'loan_status' are *Not-defaults*

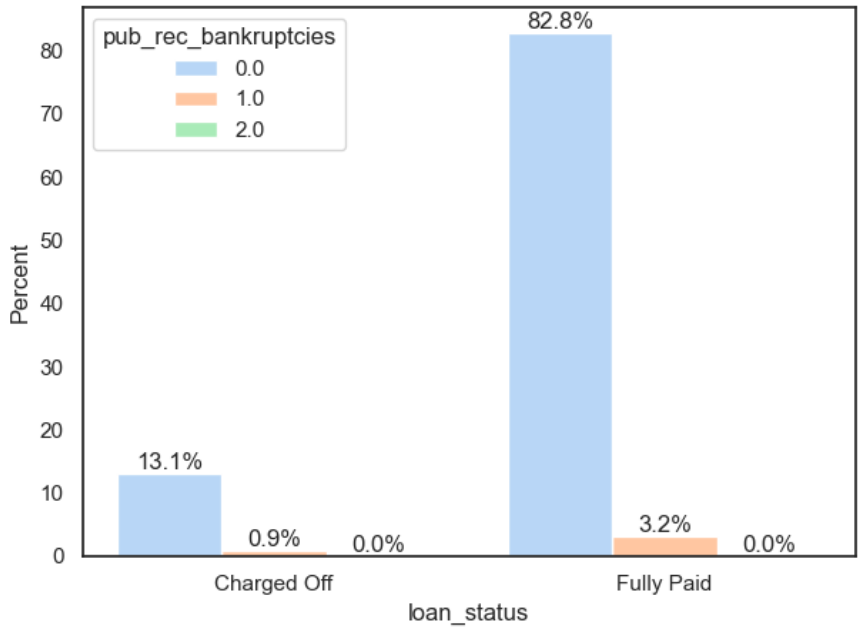
➔ ***55.3%*** of SHORT 'term' **36** *months* loans against 'Charged Off' are *Defaults*
➔ ***44.7%*** of LONG 'term' **60** *months* loans against 'Charged Off' are *Defaults*

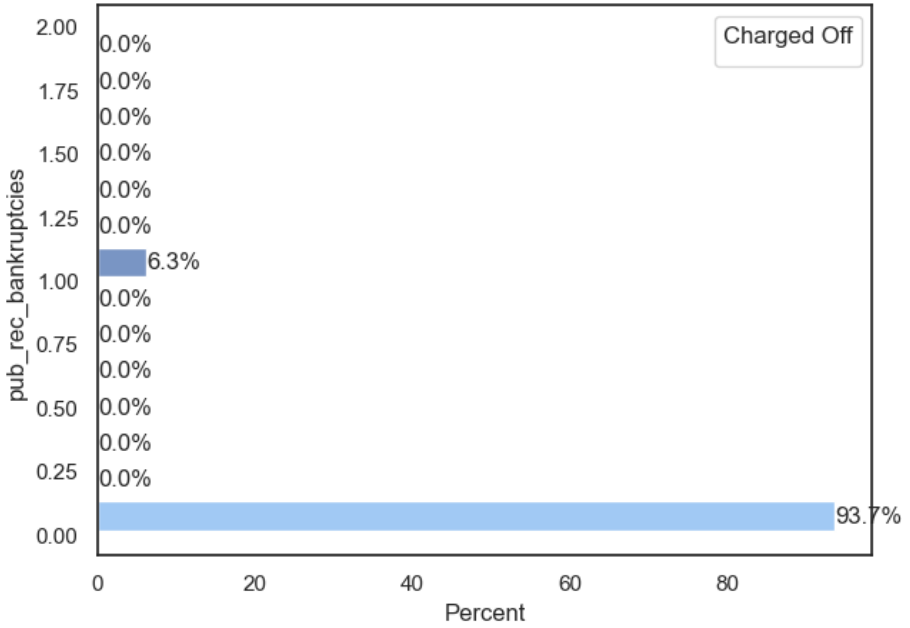




Plots for pub_rec_bankruptcies



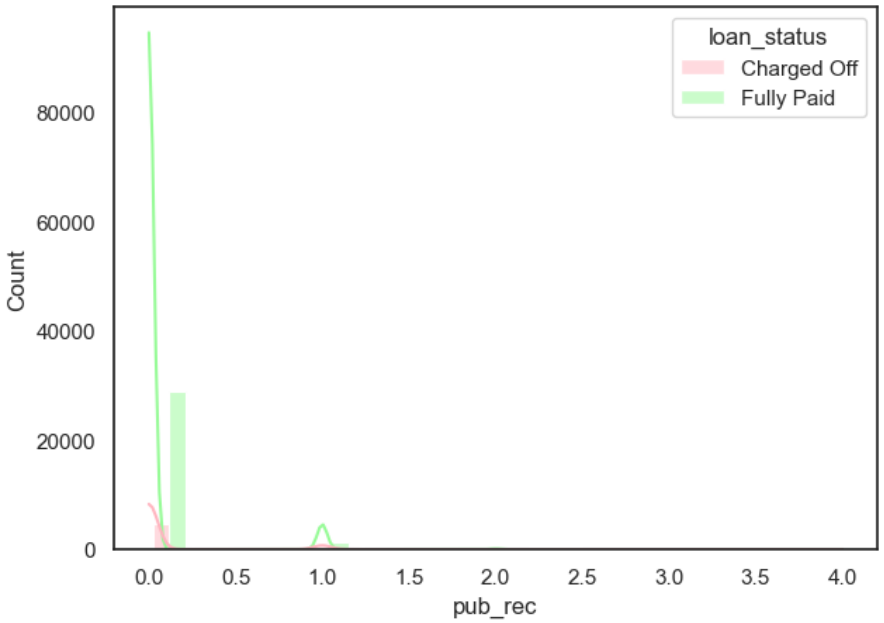


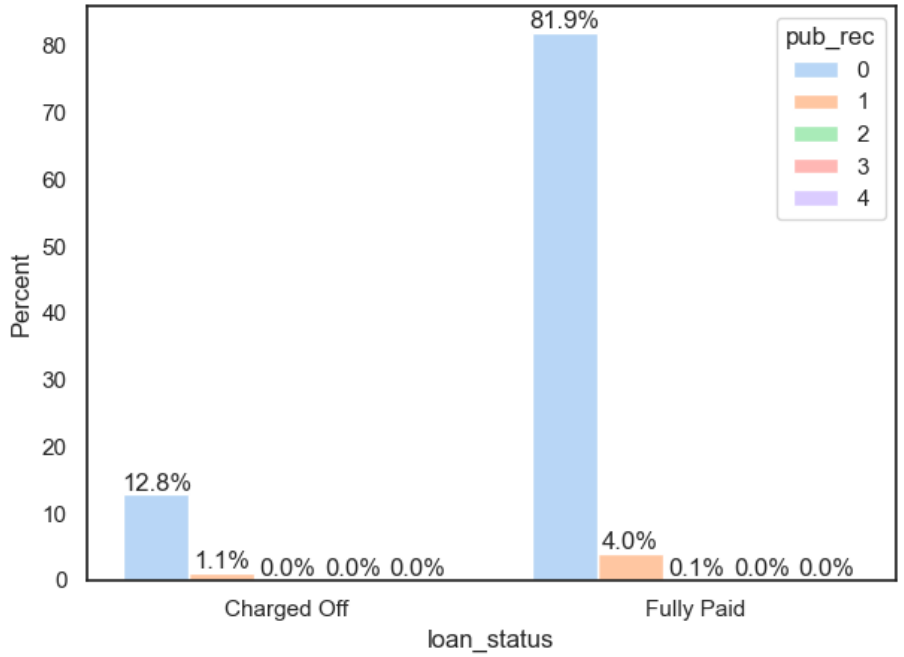


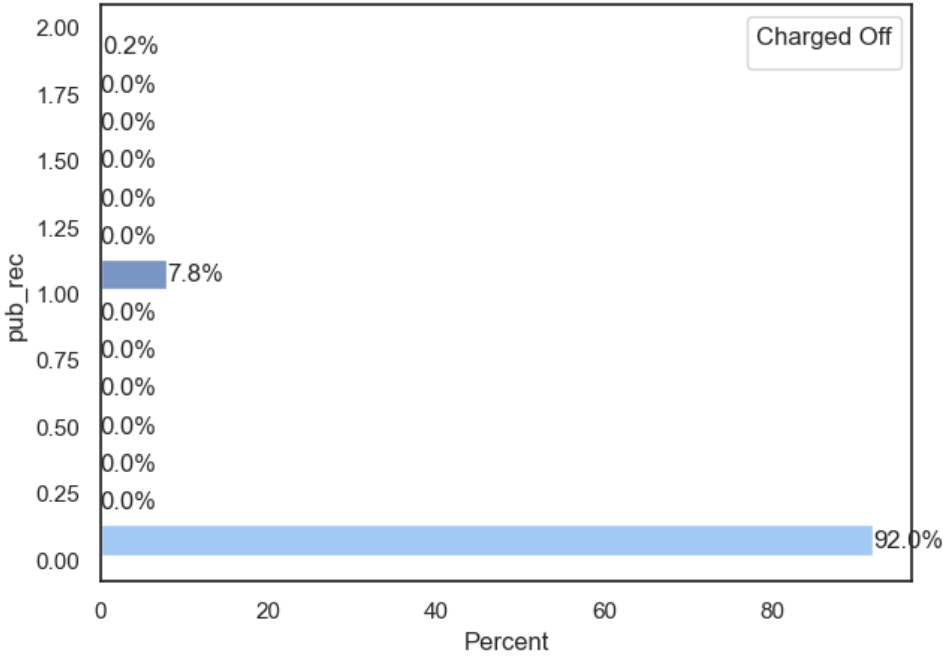
→ ***82.8%*** of 'pub_rec_bankruptcies' **0** loans against 'loan_status' are *Not-defaults*
→ ***13.1%*** of 'pub_rec_bankruptcies' **0** loans against 'loan_status' are *Defaults*
→ ***3.2%*** of 'pub_rec_bankruptcies' **1** loans against 'loan_status' are *Not-defaults*

→ ***93.7%*** of 'pub_rec_bankruptcies' **0** loans against 'Charged Off' are *Defaults*
→ ***6.3%*** of 'pub_rec_bankruptcies' **1** loans against 'Charged Off' are *Defaults*

Plots for pub_rec



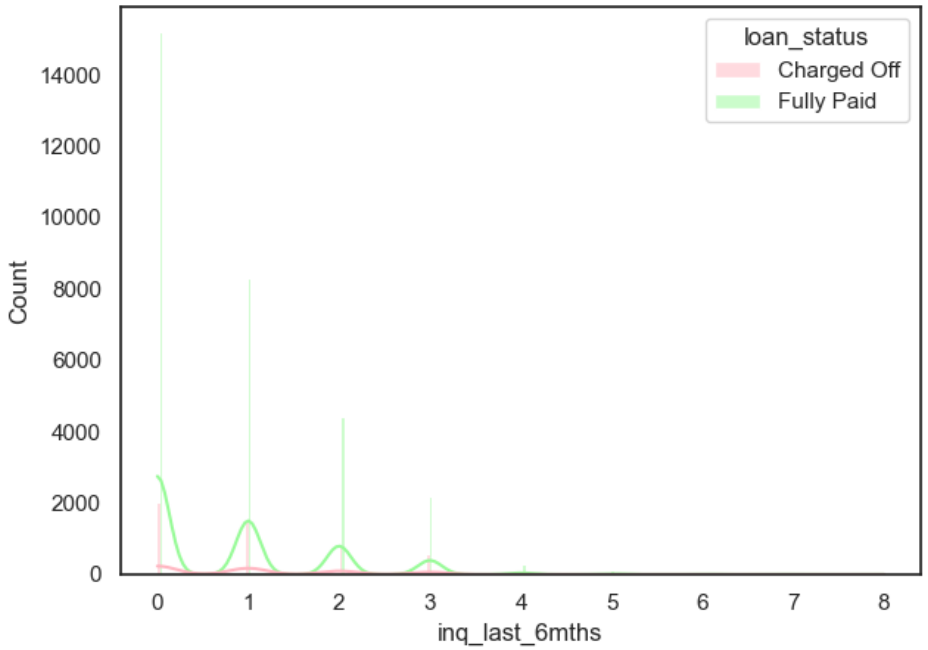


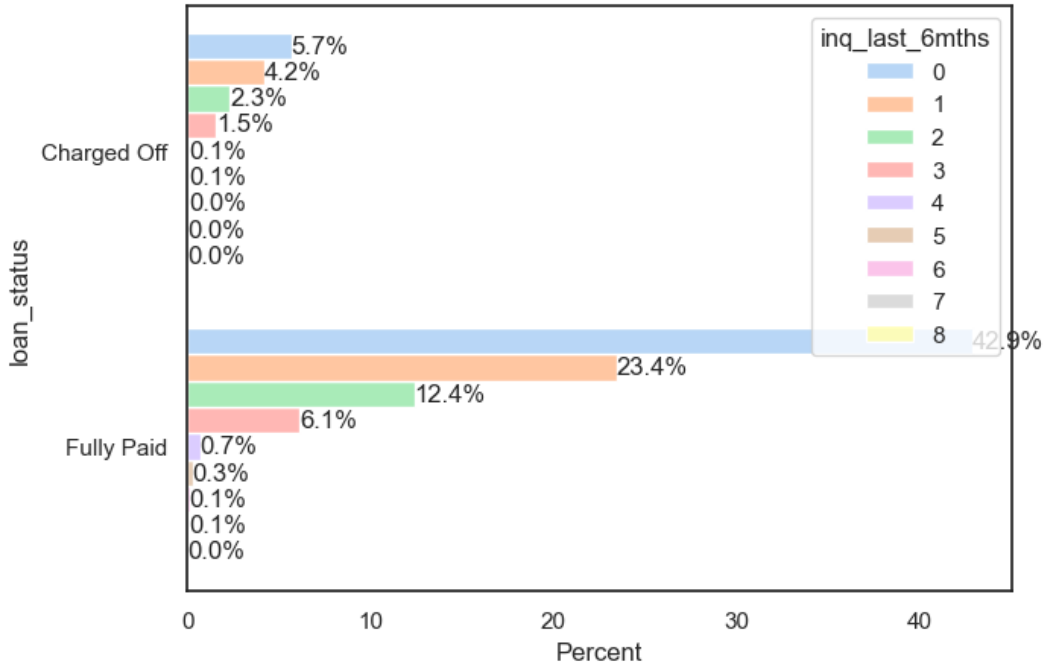


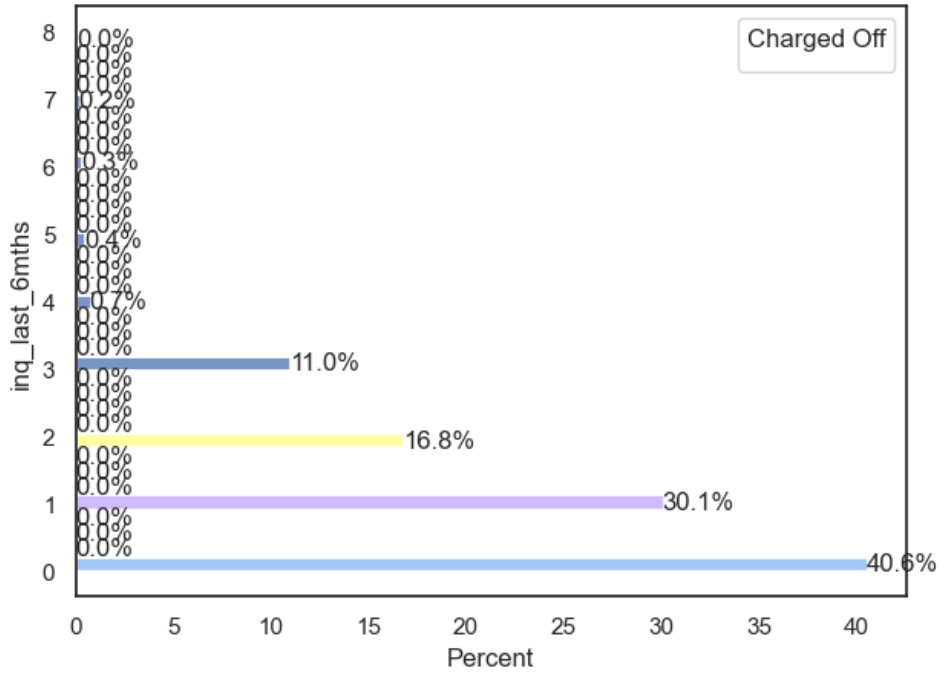
➔ ***81.9%*** of 'pub_rec' **0** loans against 'loan_status' are *Not-defaults*
➔ ***12.8%*** of 'pub_rec' **0** loans against 'loan_status' are *Defaults*
➔ ***4.0%*** of 'pub_rec' **1** loans against 'loan_status' are *Not-defaults*

➔ ***92.0%*** of 'pub_rec' **0** loans against 'Charged Off' are *Defaults*
➔ ***7.8%*** of 'pub_rec' **1** loans against 'Charged Off' are *Defaults*

Plots for inq_last_6mths



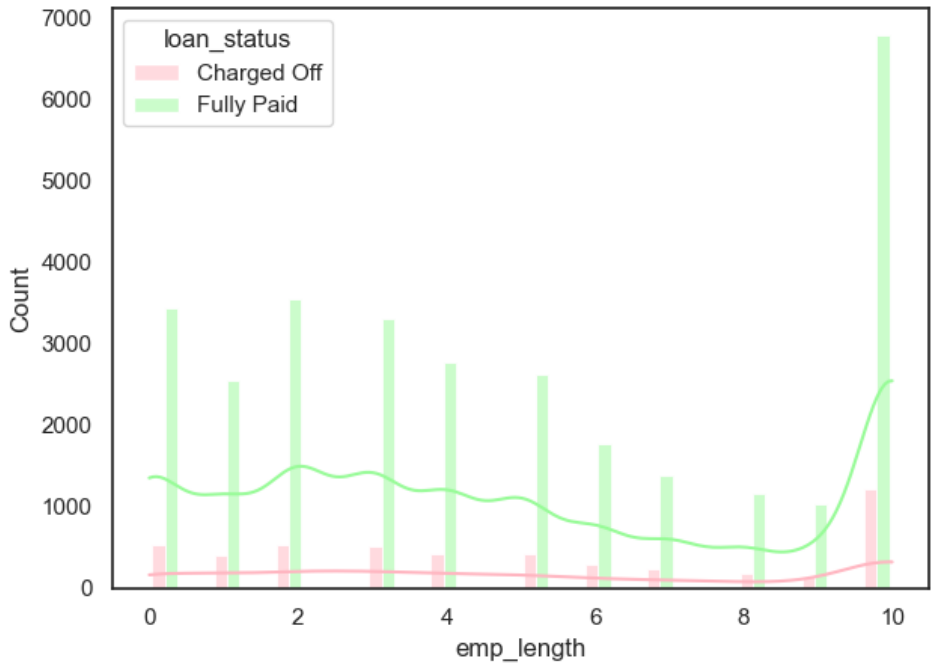


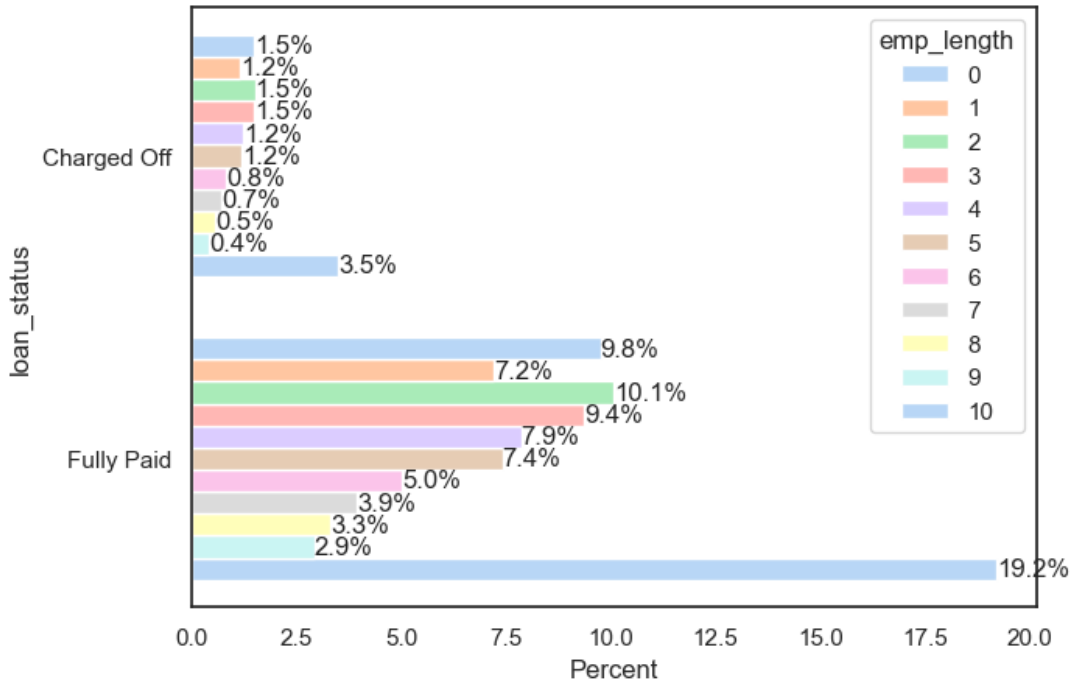


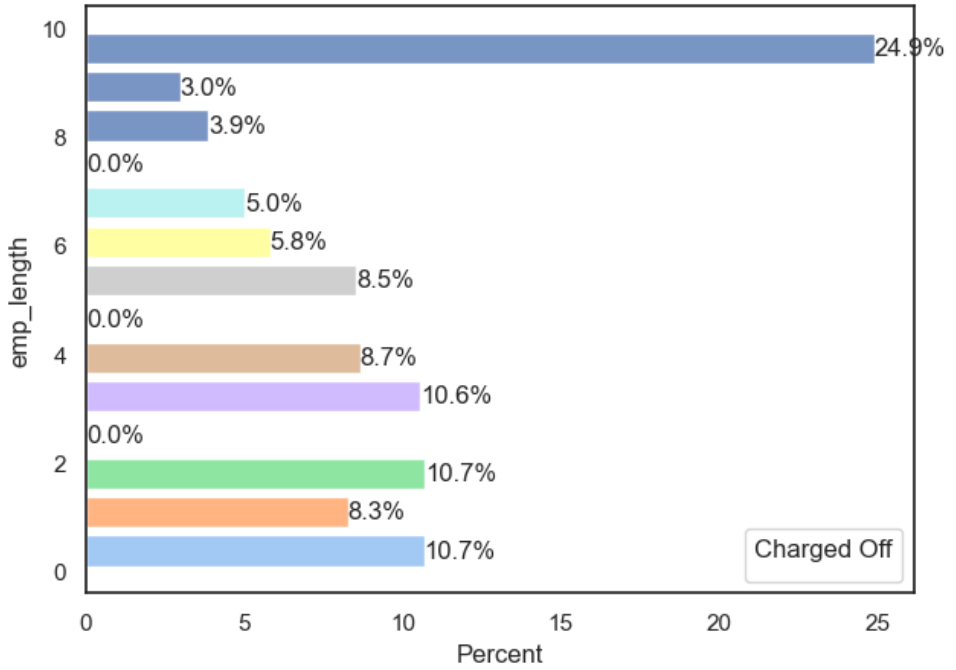
→ ***42.9%*** of 'inq_last_6mths' **0** loans against 'loan_status' are *Not-defaults*
→ ***23.4%*** of 'inq_last_6mths' **1** loans against 'loan_status' are *Not-defaults*
→ ***12.4%*** of 'inq_last_6mths' **2** loans against 'loan_status' are *Not-defaults*

→ ***40.6%*** of 'inq_last_6mths' **0** loans against 'Charged Off' are *Defaults*
→ ***30.1%*** of 'inq_last_6mths' **1** loans against 'Charged Off' are *Defaults*
→ ***16.8%*** of 'inq_last_6mths' **2** loans against 'Charged Off' are *Defaults*
→ ***11.0%*** of 'inq_last_6mths' **3** loans against 'Charged Off' are *Defaults*

Plots for emp_length



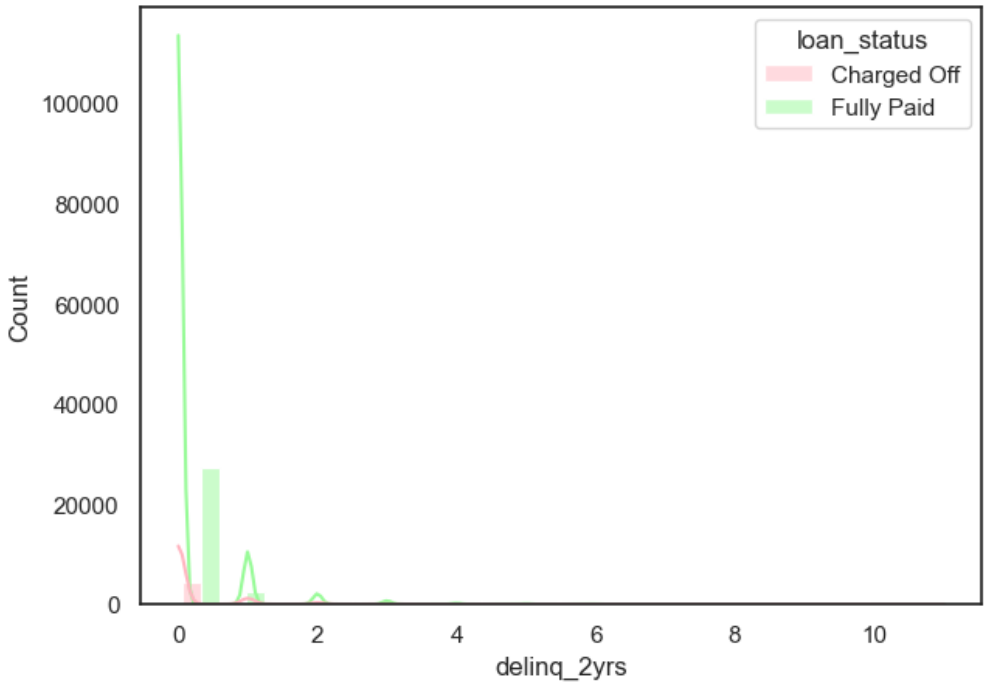




→ ***19.2%*** of 'emp_length' **10** years loans against 'loan_status' are *Not-defaults*

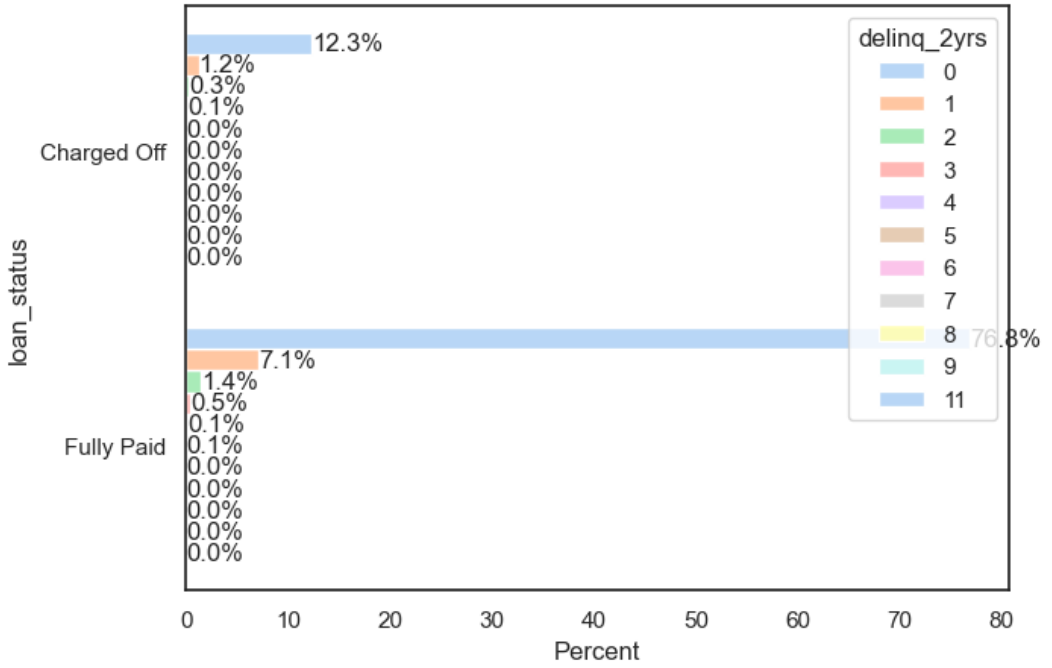
→ ***24.9%*** of 'emp_length' **10** loans years against 'Charged Off' are *Defaults*

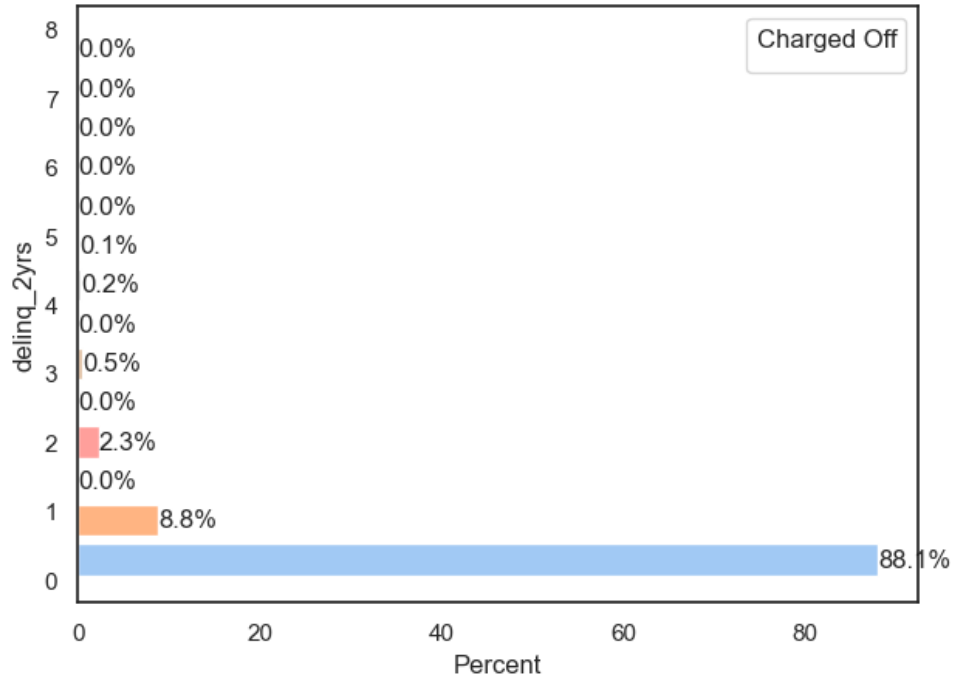
Plots for delinq_2yrs



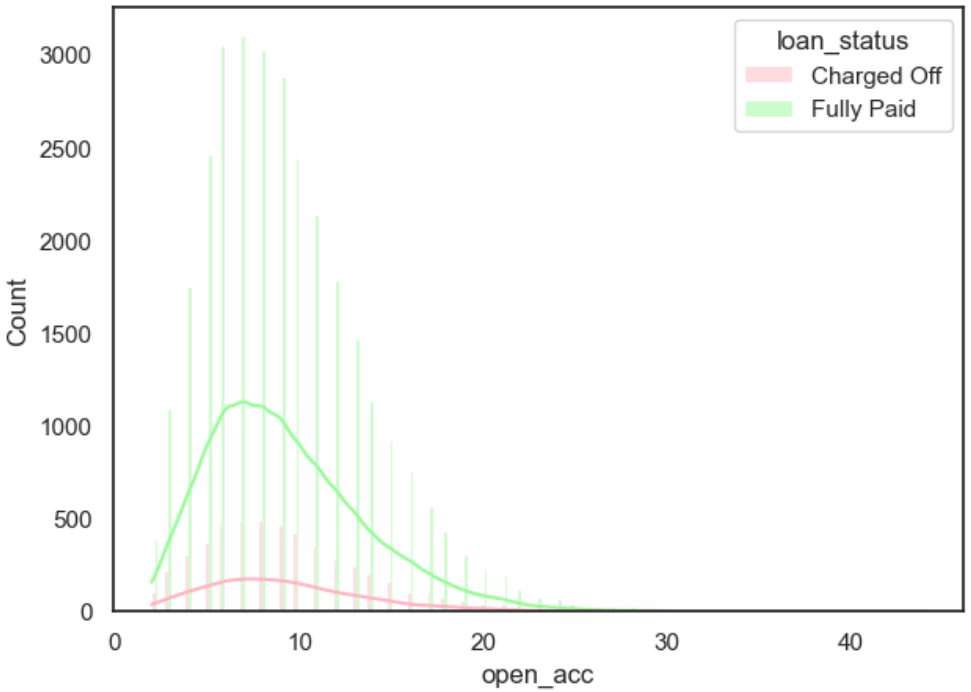
→ ***76.8%*** of 'delinq_2yrs' **0** loans against 'loan_status' are *Not-defaults*

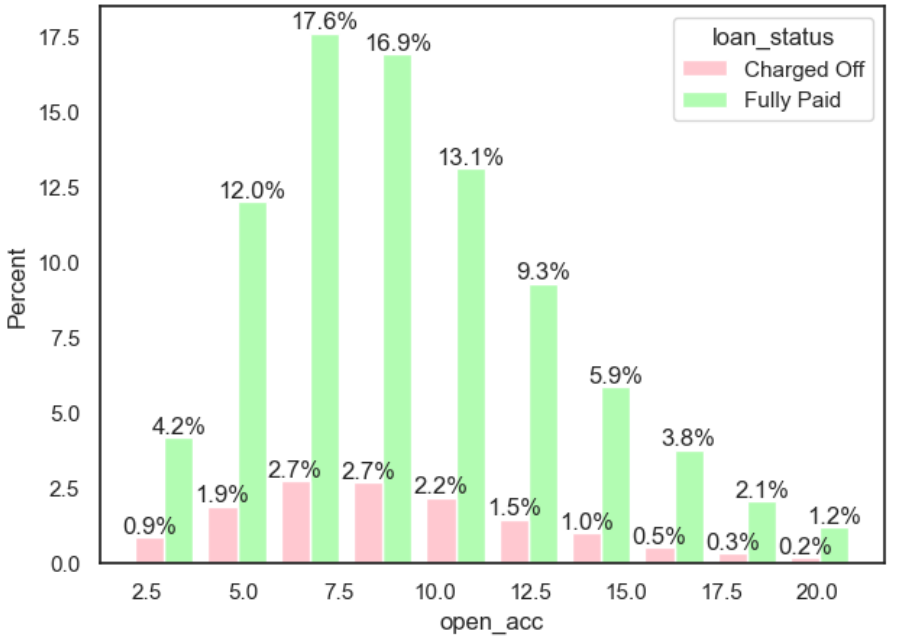
→ ***88.1%*** of 'delinq_2yrs' **0** loans against 'Charged Off' are *Defaults*

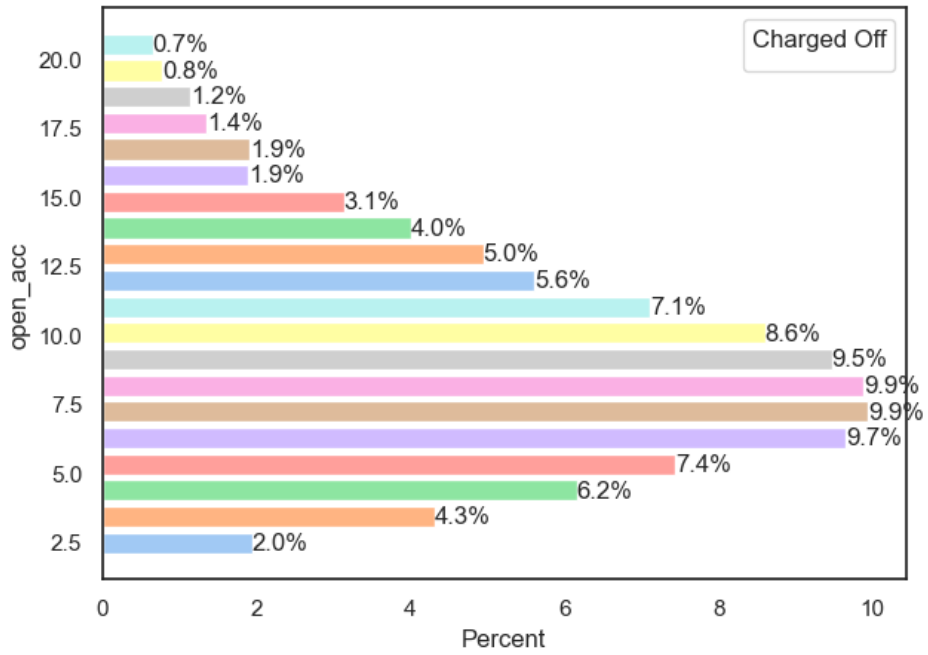




Plots for open_acc



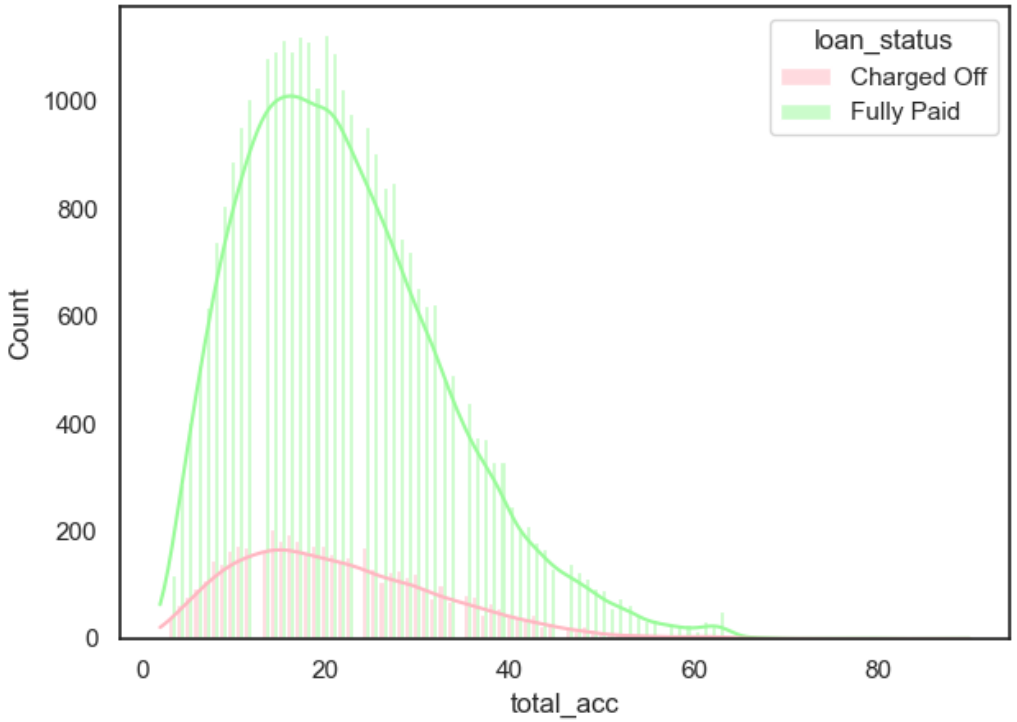


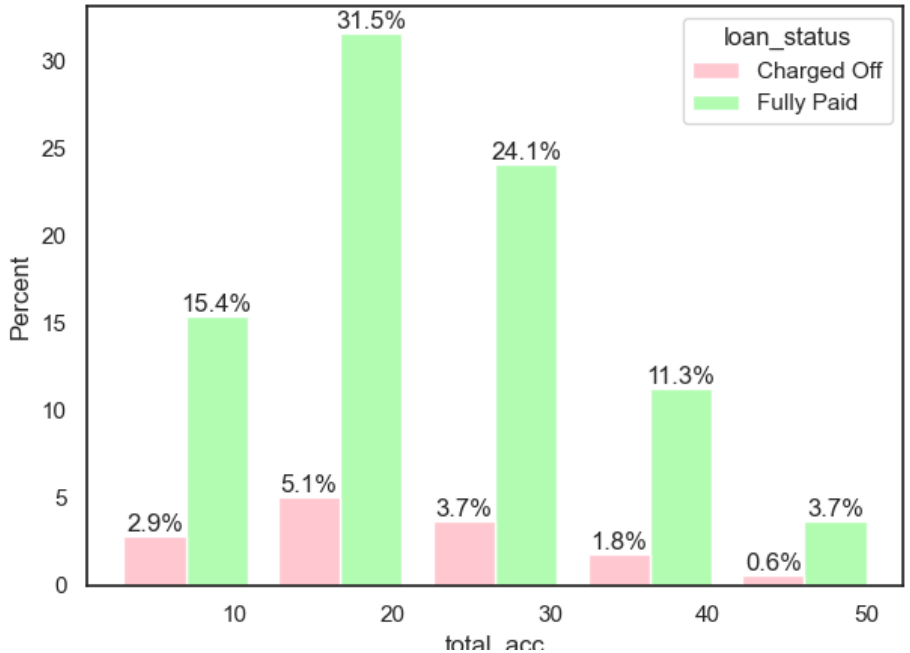


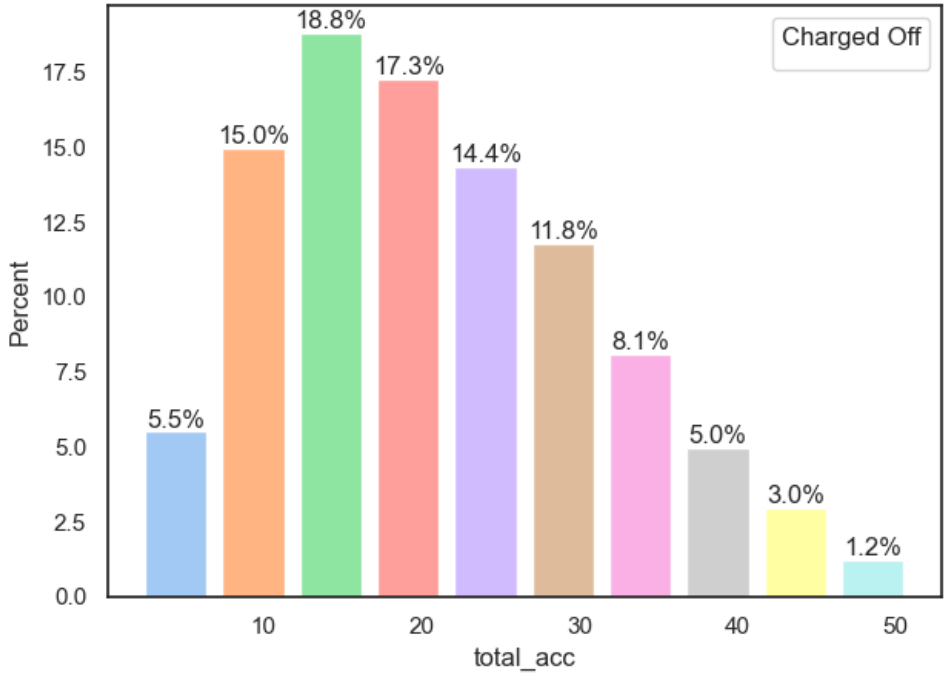
→ ***>10%*** of 'open_acc' are between **5-12.5** loans against 'loan_status' are *Not-defaults*

→ ***>7%*** of 'open_acc' are between **5-11** loans against 'Charged Off' are *Defaults*

Plots for total_acc



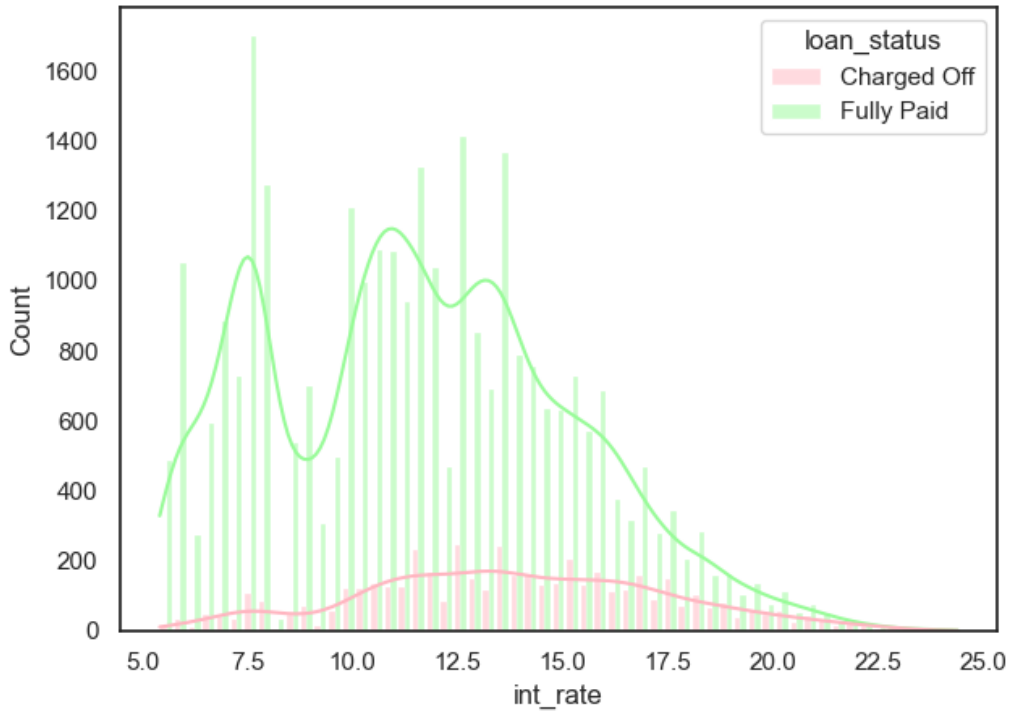


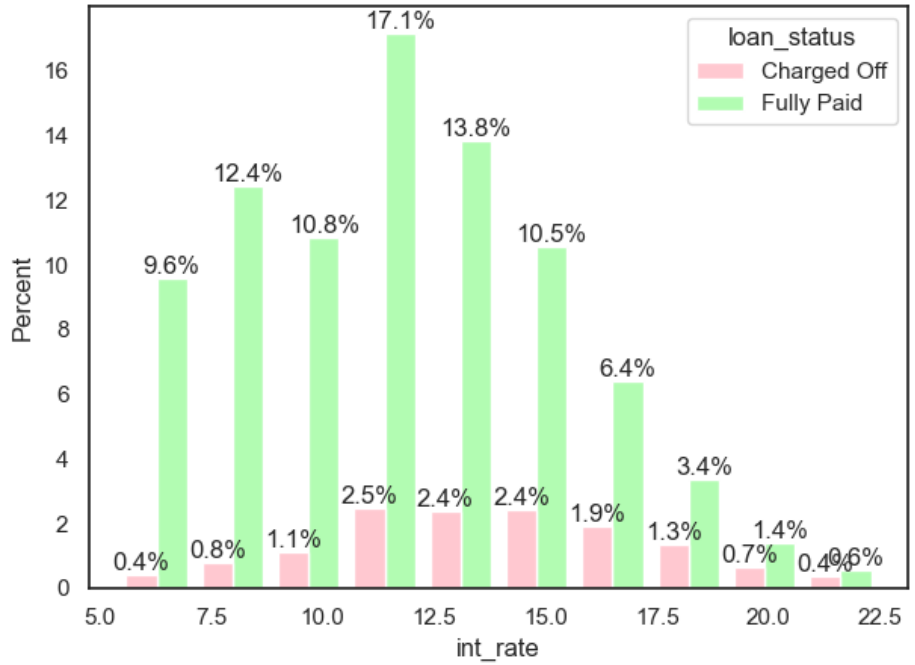


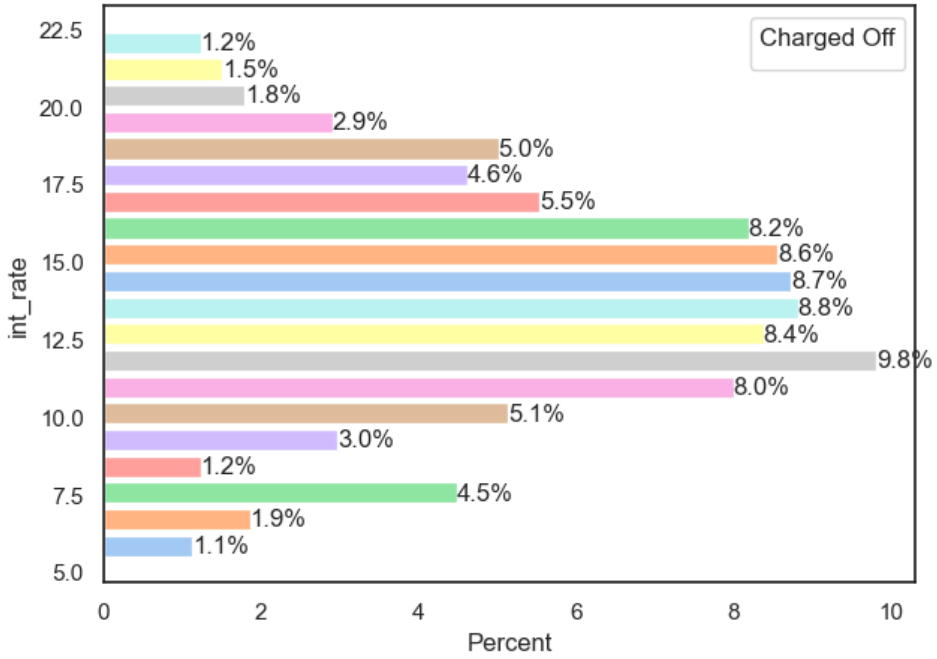
→ ***31.5%*** of 'total_acc' **20** loans against 'loan_status' are *Not-defaults*
→ ***24.1%*** of 'total_acc' **30** loans against 'loan_status' are *Not-defaults*

→ ***>12%*** of 'total_acc' are between **10-30** loans against 'Charged Off' are *Defaults*

Plots for int_rate



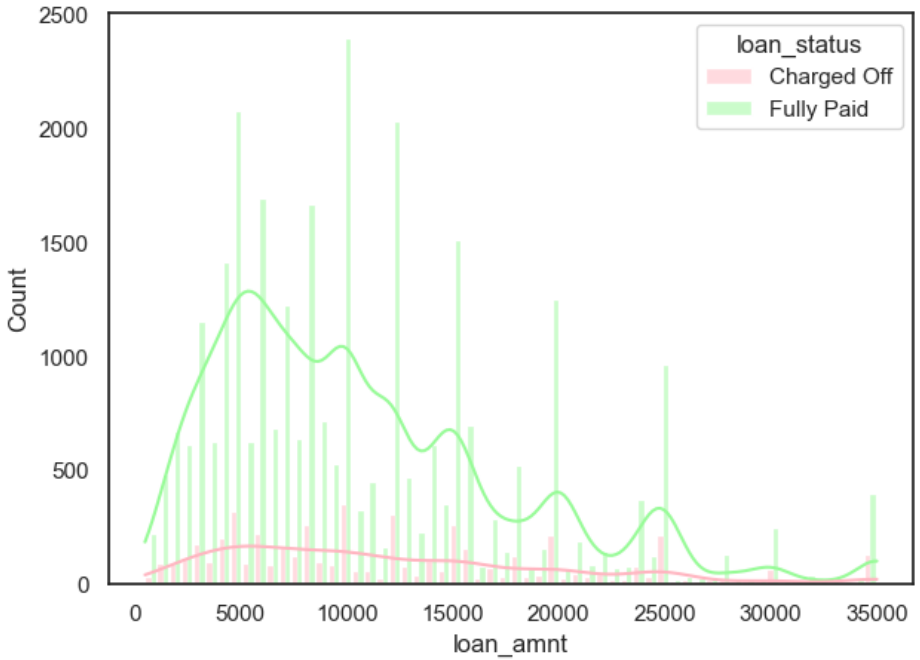


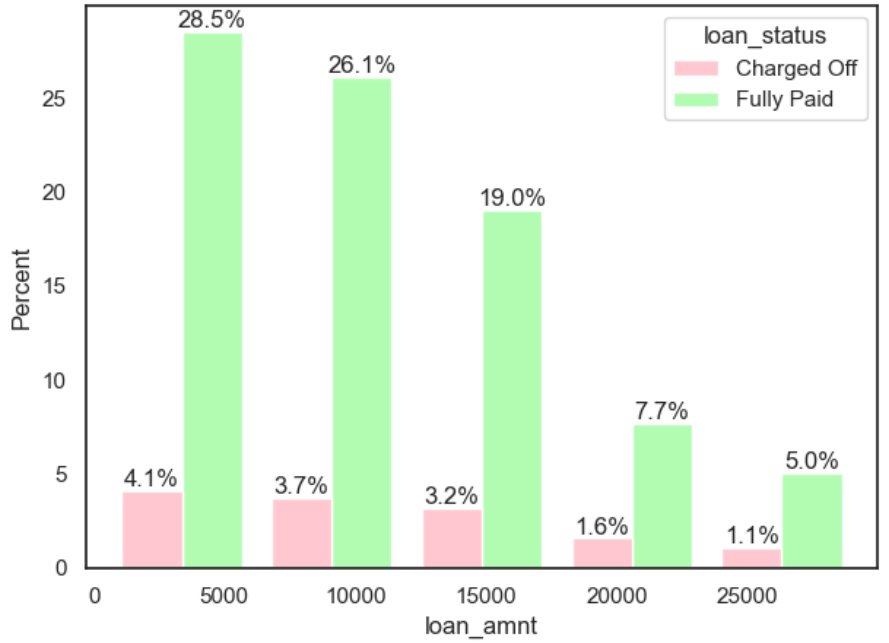


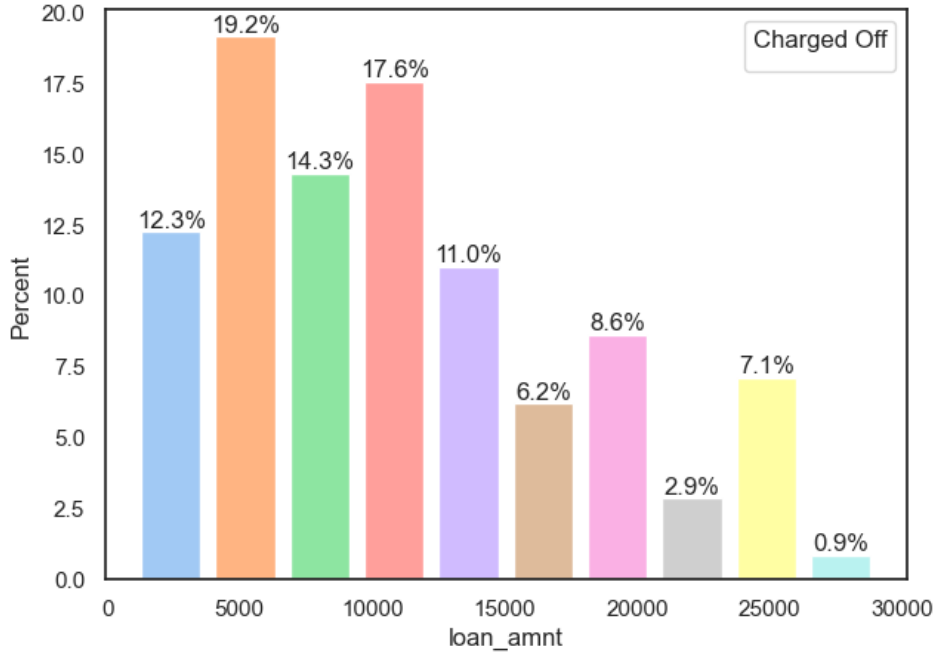
➔ ***17.1%*** of 'int_rate' **12** loans against 'loan_status' are *Not-defaults*
➔ ***13.8%*** of 'int_rate' **13** loans against 'loan_status' are *Not-defaults*
➔ ***12.4%*** of 'int_rate' **7** loans against 'loan_status' are *Not-defaults*

➔ ***>8%*** of 'int_rate' are between **11-16** loans against 'Charged Off' are *Defaults*

Plots for loan_amnt



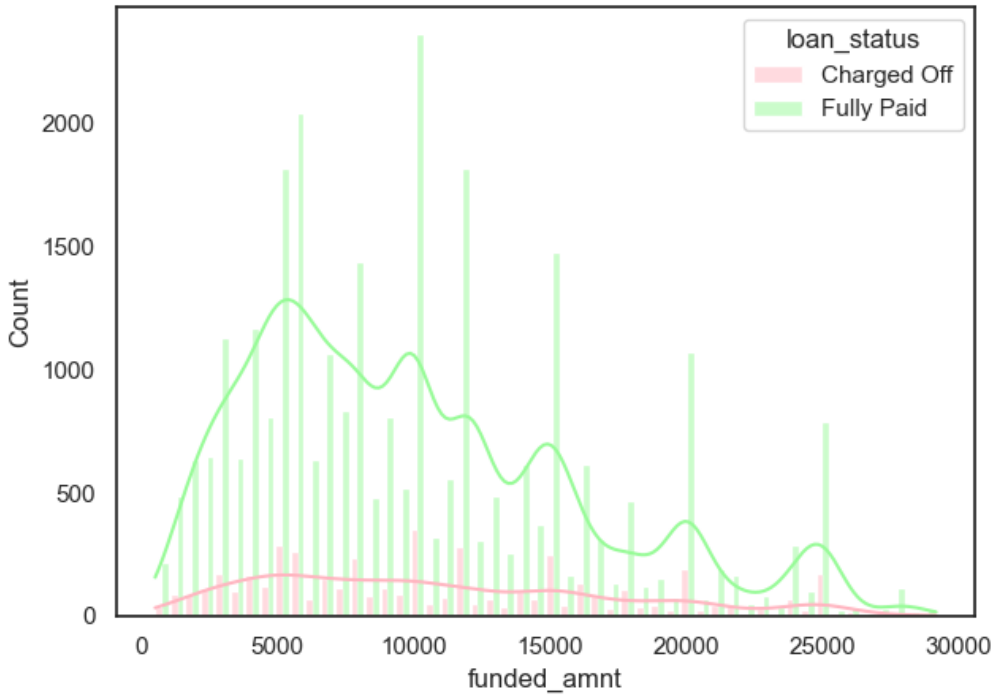


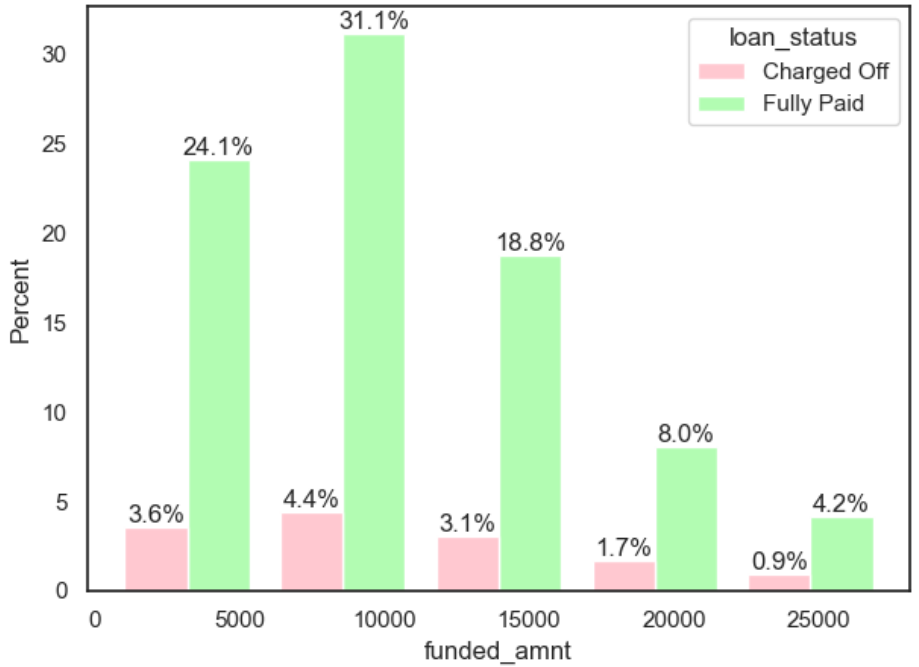


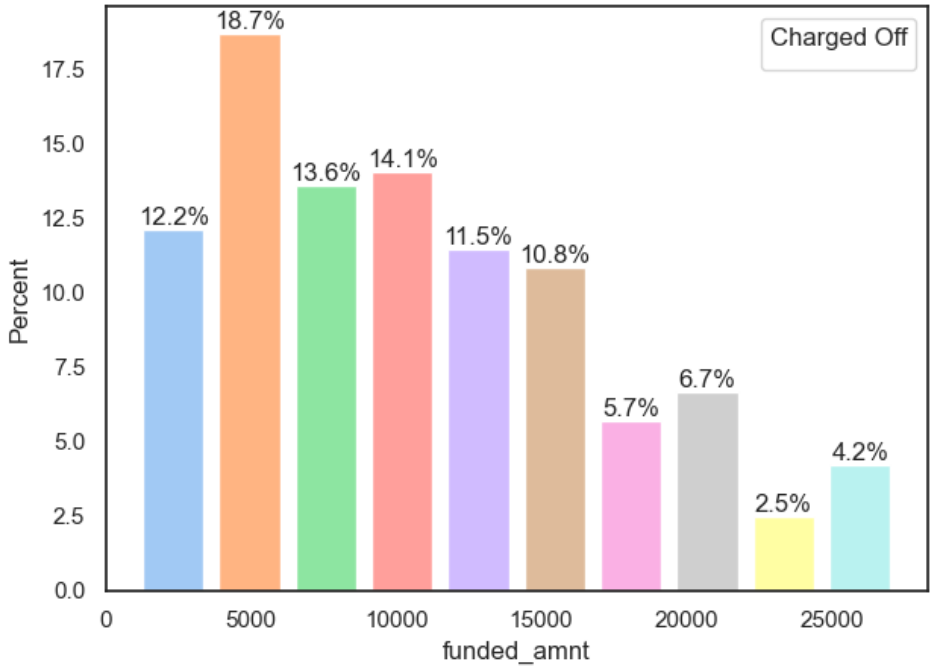
→ ***28.5%*** of 'loan_amnt' **5k** loans against 'loan_status' are *Not-defaults*
→ ***26.1%*** of 'loan_amnt' **10k** loans against 'loan_status' are *Not-defaults*
→ ***19.0%*** of 'loan_amnt' **15k** loans against 'loan_status' are *Not-defaults*

→ ***19.2%*** of 'loan_amnt' **5k** loans against 'Charged Off' are *Defaults*
→ ***17.6%*** of 'loan_amnt' **10k** loans against 'Charged Off' are *Defaults*

Plots for funded_amnt



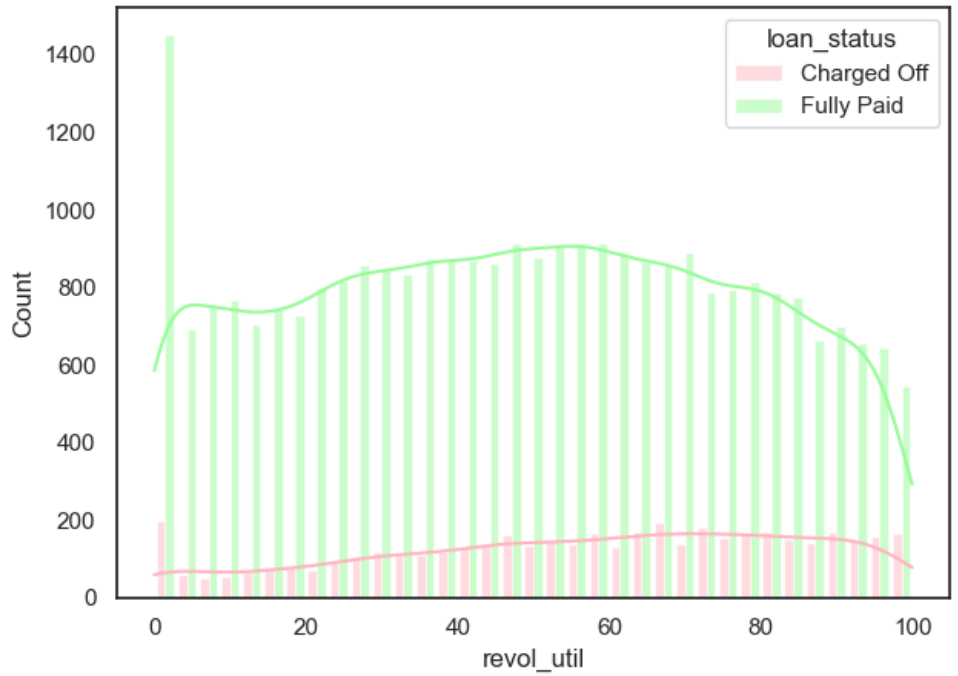


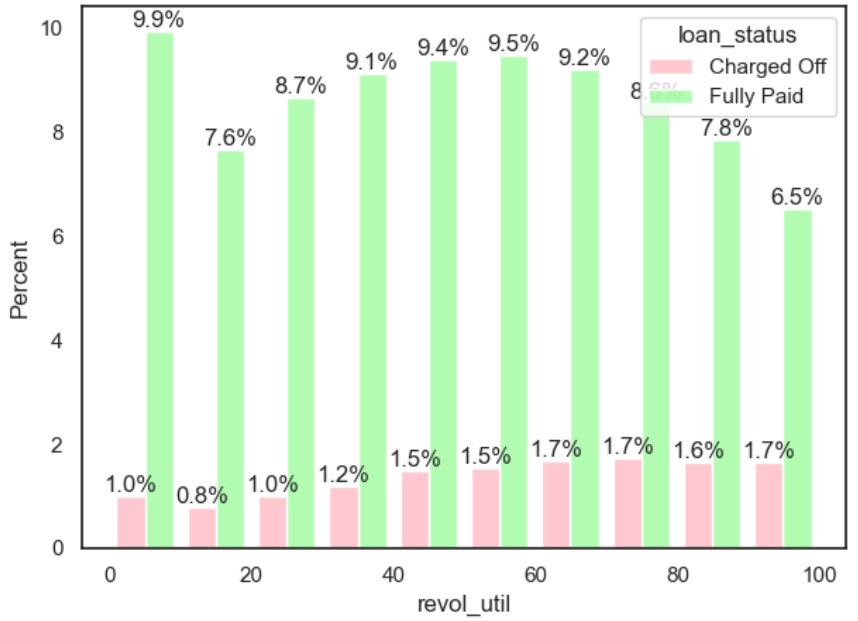


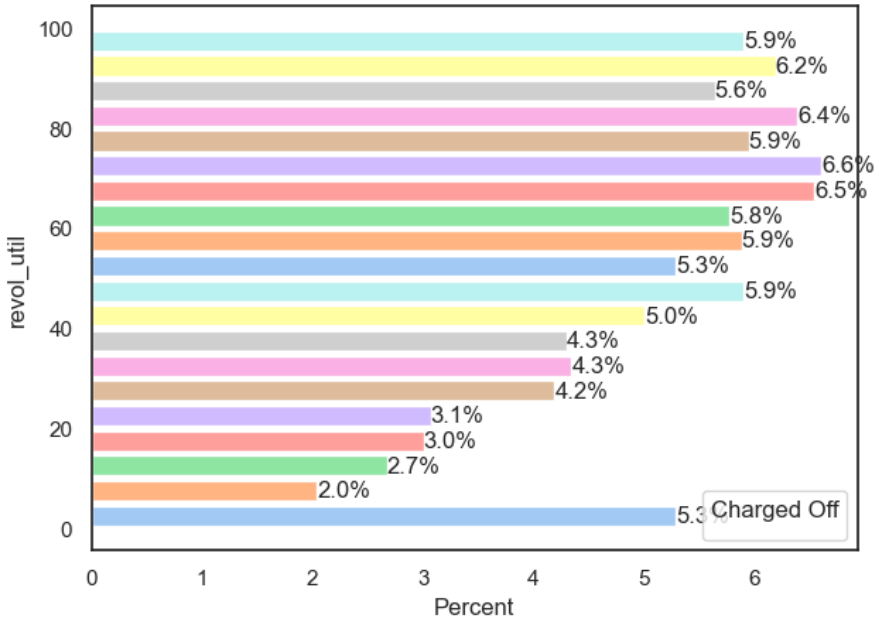
→ ***31.1%*** of 'funded_amnt' **5k** loans against 'loan_status' are *Not-defaults*
→ ***24.1%*** of 'funded_amnt' **10k** loans against 'loan_status' are *Not-defaults*
→ ***18.8%*** of 'funded_amnt' **15k** loans against 'loan_status' are *Not-defaults*

→ ***18.7%*** of 'funded_amnt' **5k** loans against 'Charged Off' are *Defaults*

Plots for revol_util



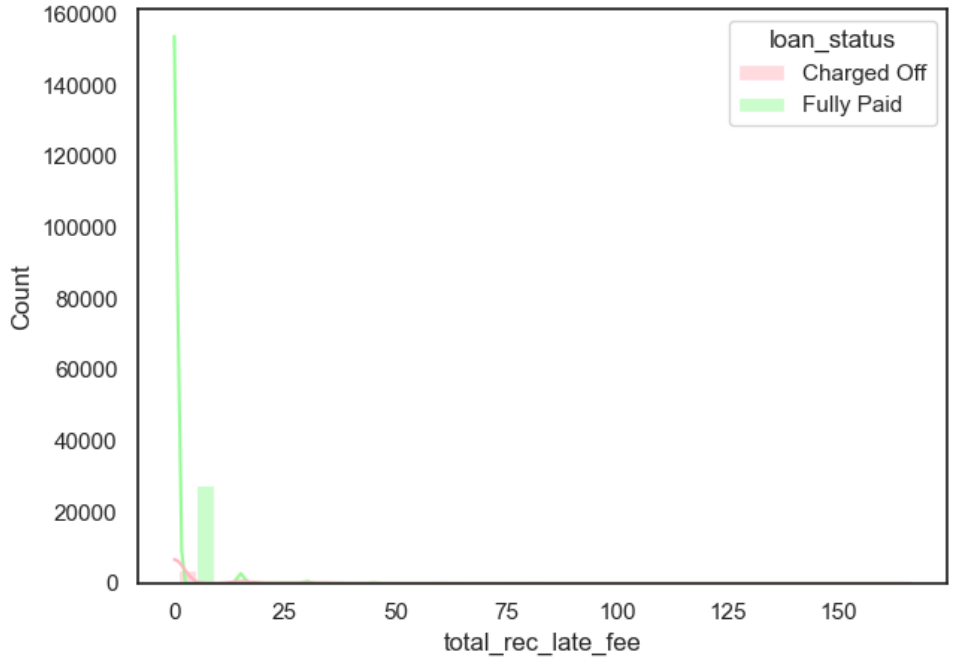


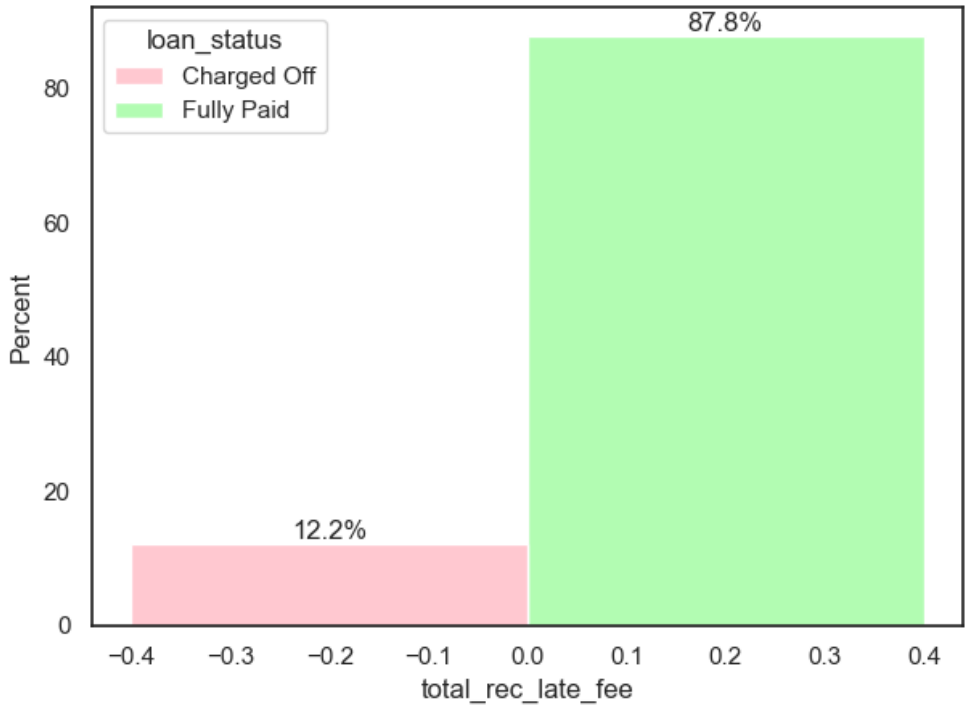


➔ ***9.9%*** of 'revol_util' is **5** loans against 'loan_status' are *Not-defaults*

➔ ***>5%*** of 'revol_util' is **>40** loans against 'Charged Off' are *Defaults*

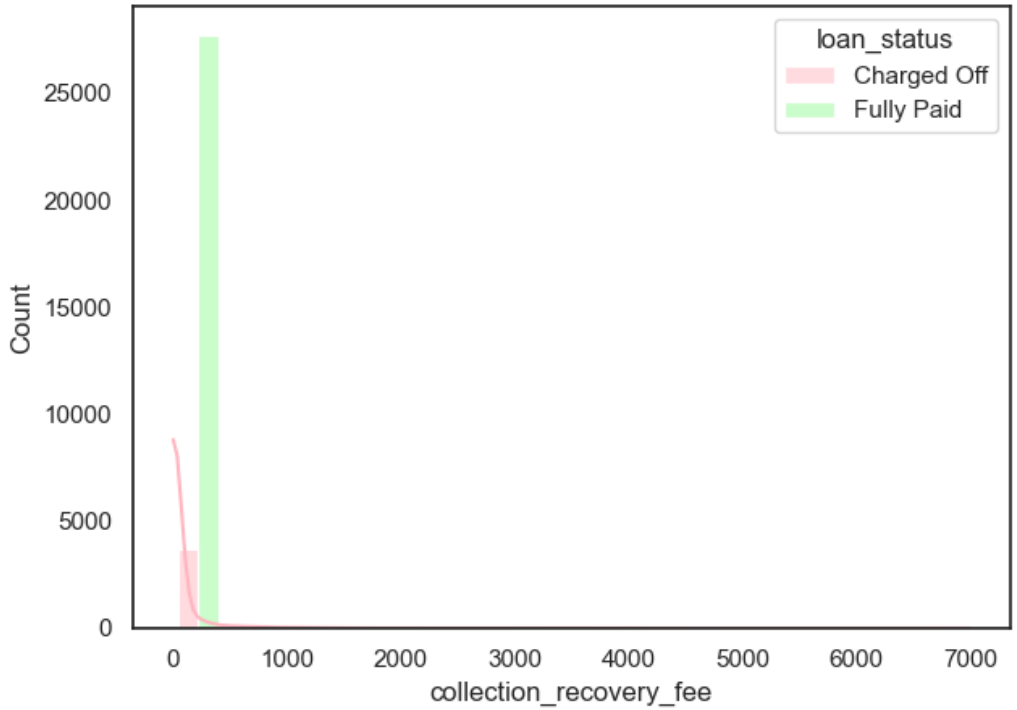
Plots for total_rec_late_fee

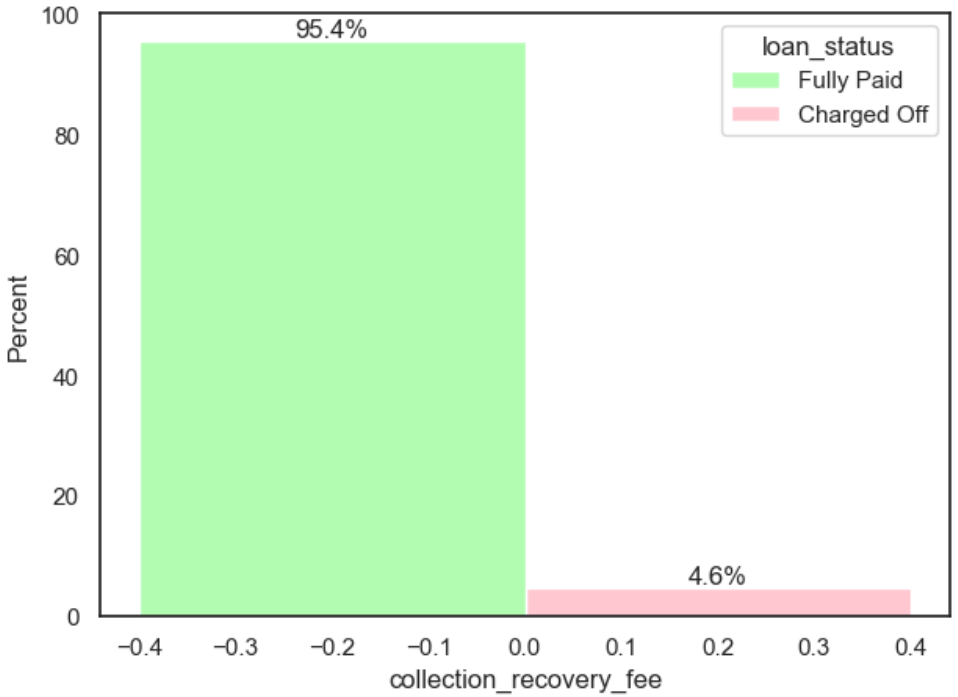




→ ***87.8%*** of 'total_rec_late_fee' are between **0.0** to **0.4** loans against 'loan_status' are *Not-defaults*
→ ***12.2%*** of 'total_rec_late_fee' are between **0.0** to **-0.4** loans against 'loan_status' are *Defaults*

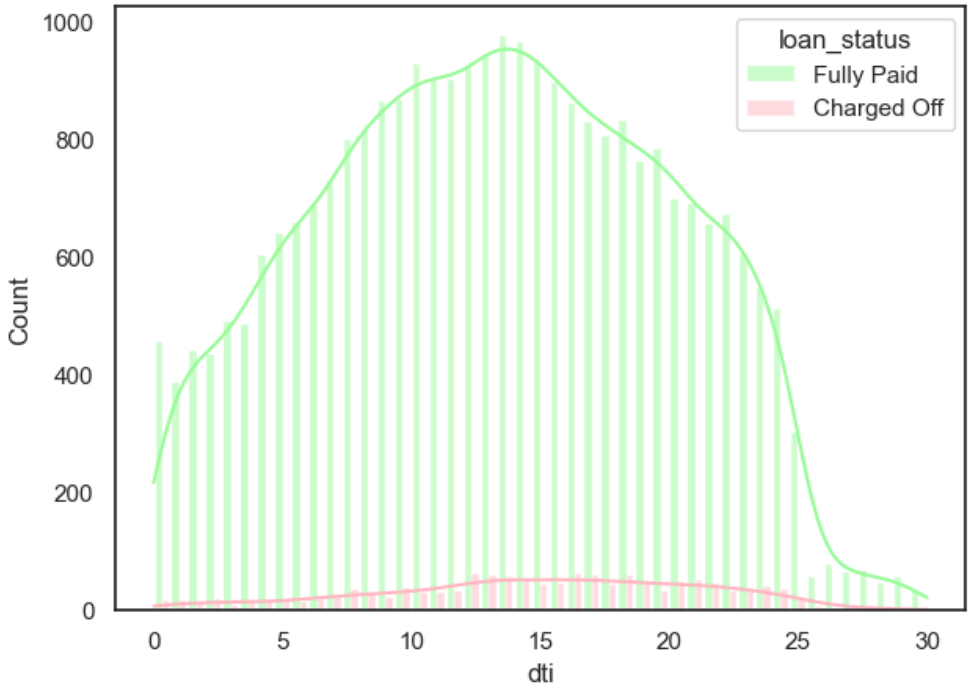
Plots for collection_recovery_fee

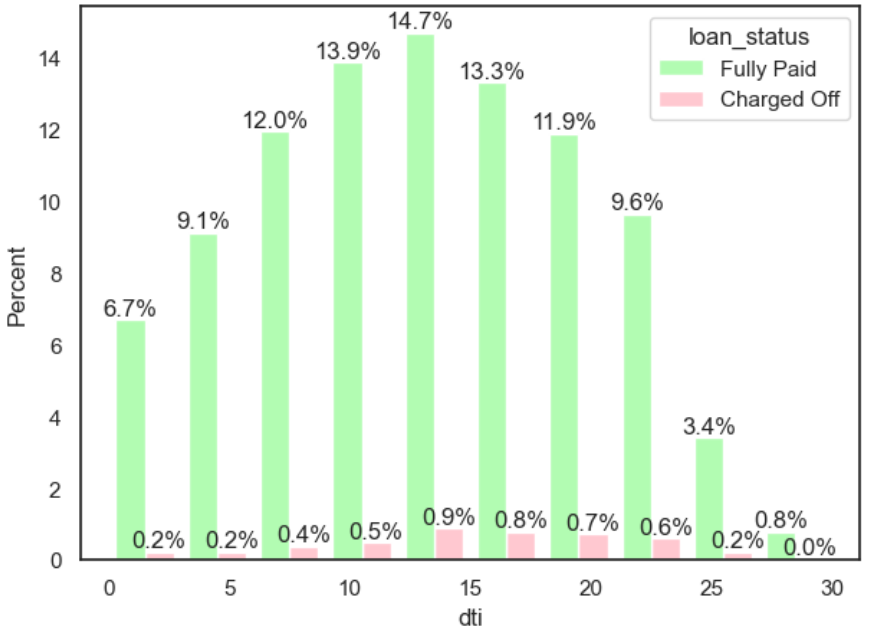


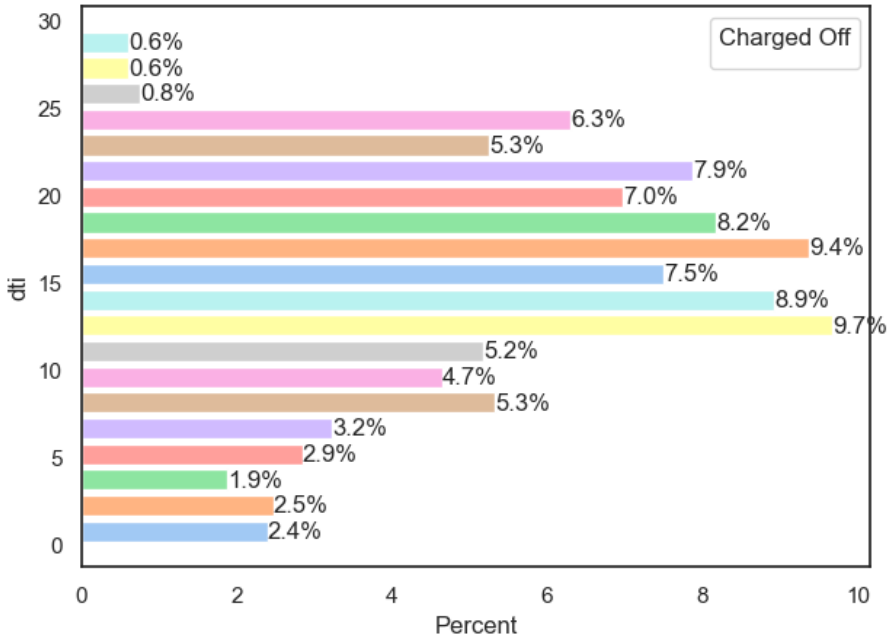


→ ***4.6%*** of 'collection_recovery_fee' are between **0.0** to **0.4** loans against 'loan_status' are *Defaults*
→ ***95.4%*** of 'collection_recovery_fee' are between **0.0** to **-0.4** loans against 'loan_status' are *Not-defaults*

Plots for dti



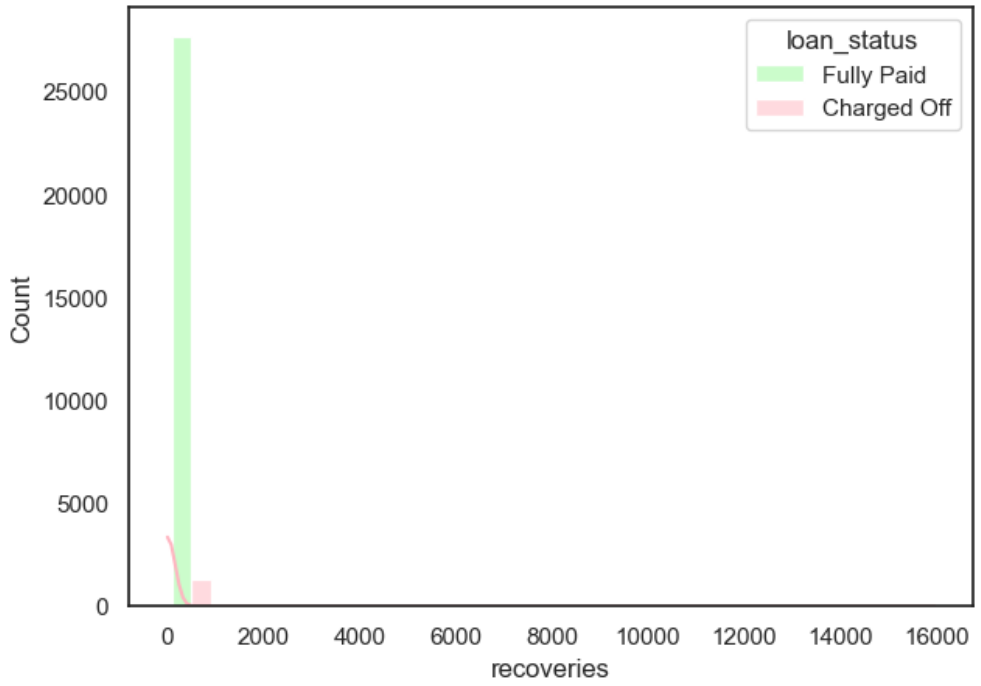


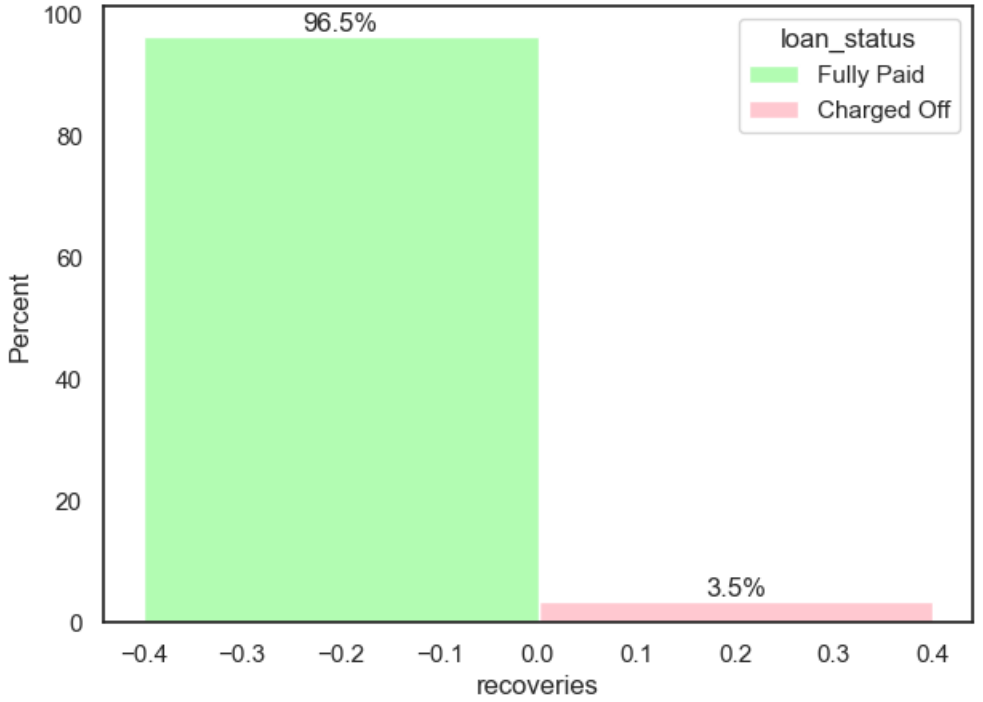


→ **14.7%** 'dti' have Simmetry arround **14** value loans against 'loan_status' are *Not-defaults*

→ ***>7%*** 'dti' are between **12-22** value loans against 'Charged Off' are *Defaults*

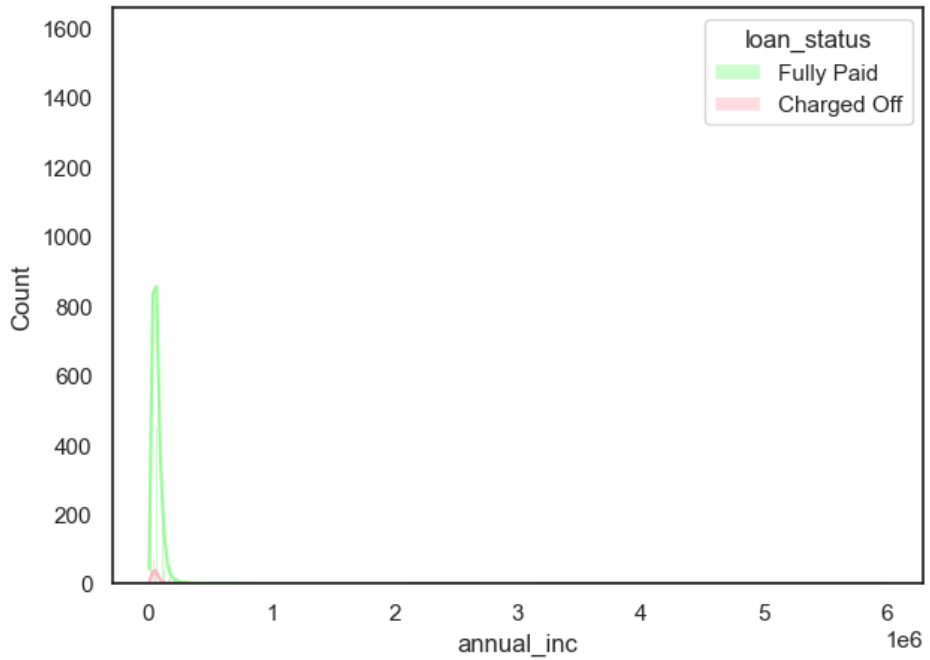
Plots for recoveries

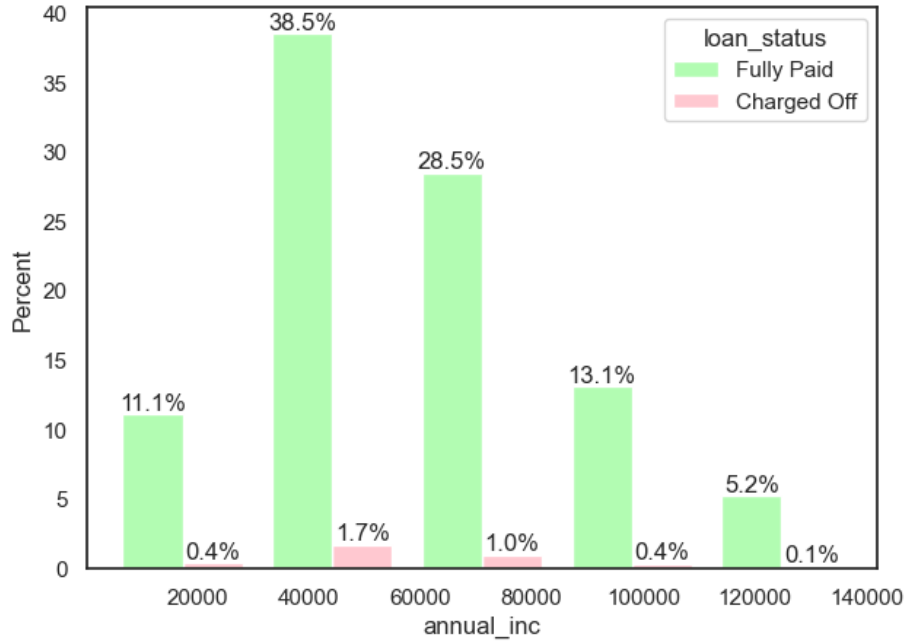


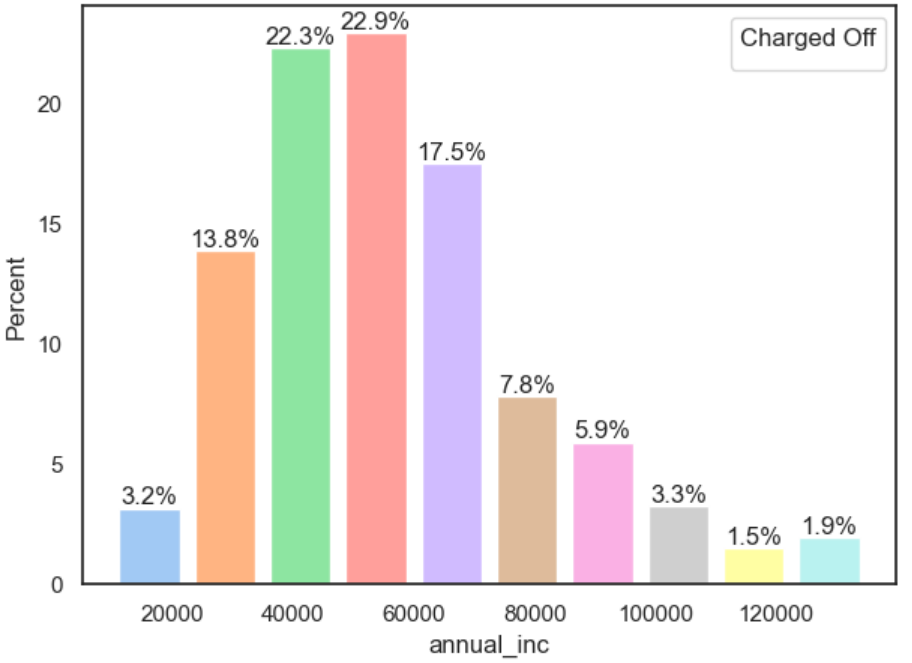


→ ***3.5%*** of 'recoveries' are between **0.0** to **0.4** loans against 'loan_status' are *Defaults*
→ ***96.5%*** of 'recoveries' are between **0.0** to **-0.4** loans against 'loan_status' are *Not-defaults*

Plots for annual_inc



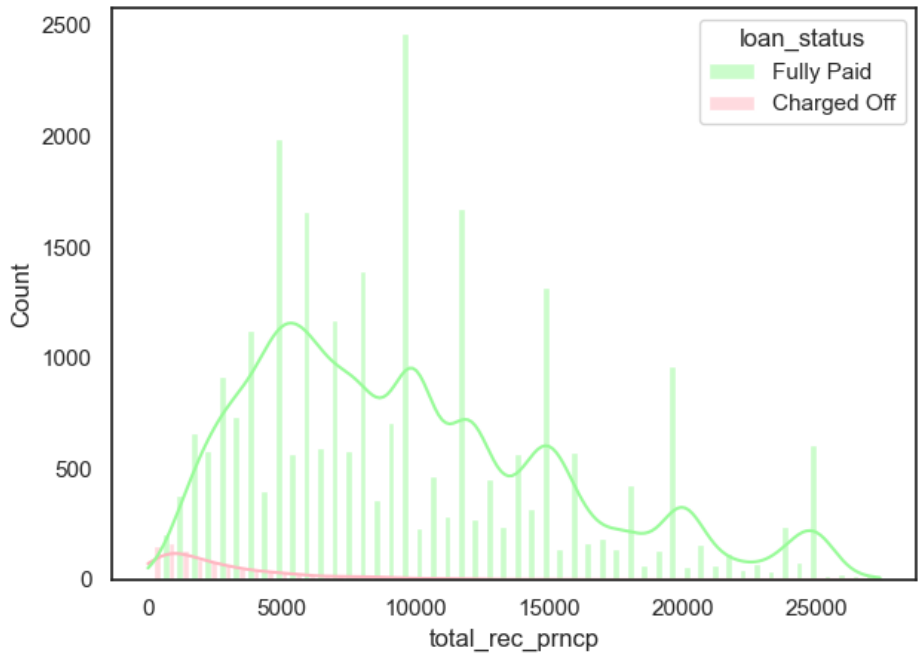


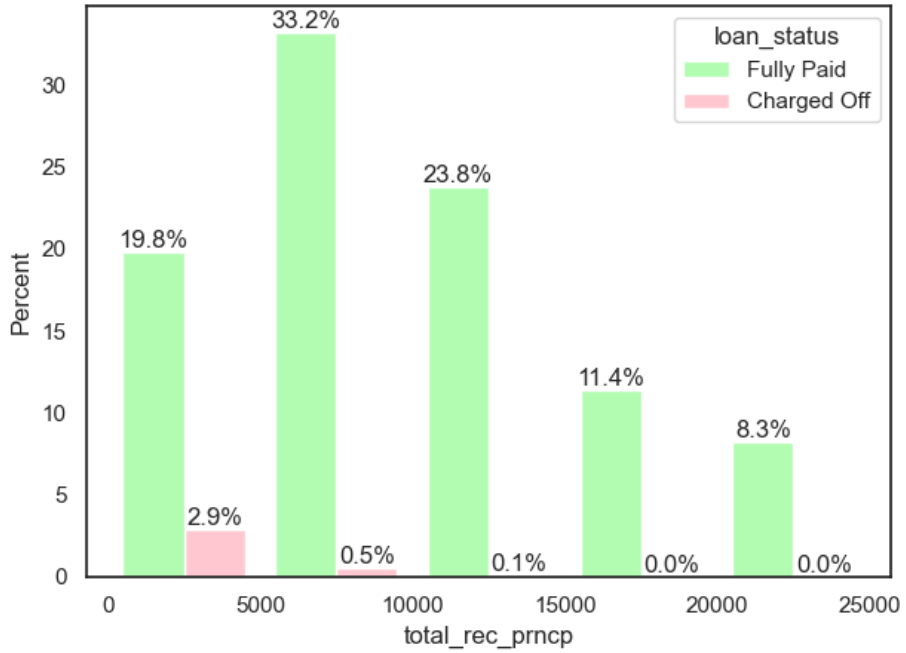


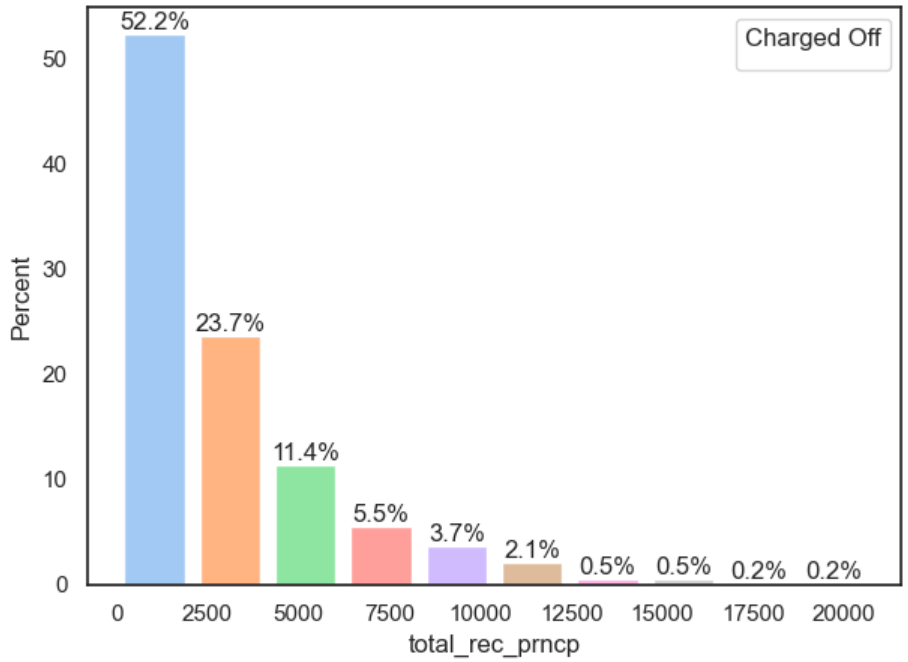
→ ***38.5%*** of 'annual_inc' **40k** loans against 'loan_status' are *Not-defaults*
→ ***28.5%*** of 'annual_inc' **70k** loans against 'loan_status' are *Not-defaults*

→ ***22.9%*** of 'annual_inc' **55k** loans against 'loan_status' are *Defaults*
→ ***22.3%*** of 'annual_inc' **40k** loans against 'loan_status' are *Defaults*
→ ***17.5%*** of 'annual_inc' **65k** loans against 'loan_status' are *Defaults*
→ ***13.8%*** of 'annual_inc' **30k** loans against 'loan_status' are *Defaults*

Plots for total_rec_prncp





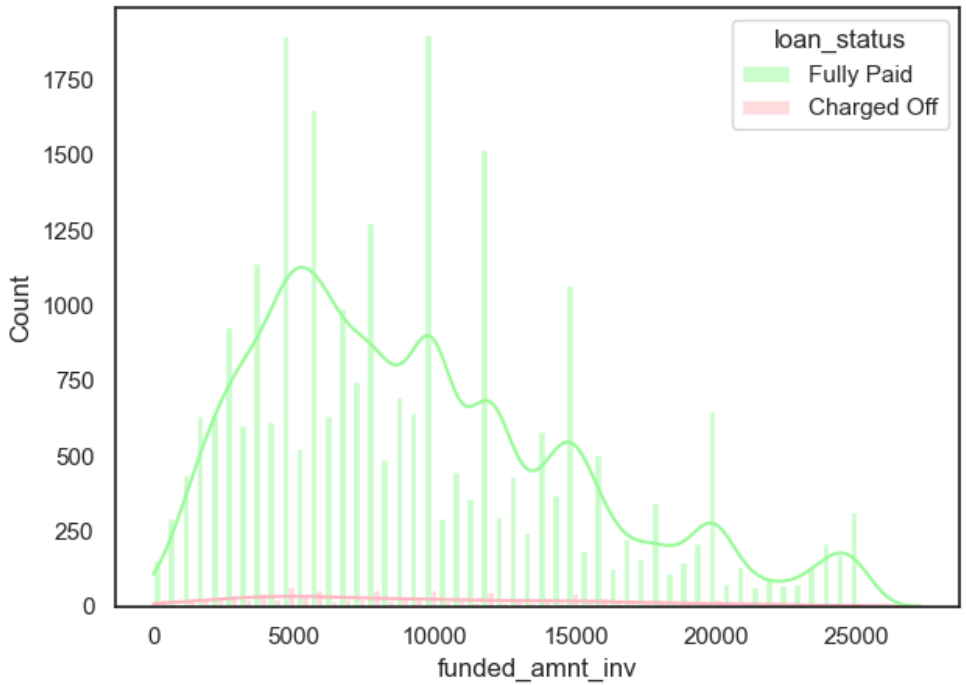


→ ***33.2%*** of 'total_rec_prncp' **10k** loans against 'loan_status' are *Not-defaults*
→ ***23.8%*** of 'total_rec_prncp' **5k** loans against 'loan_status' are *Not-defaults*
→ ***19.8%*** of 'total_rec_prncp' **15k** loans against 'loan_status' are *Not-defaults*

→ ***52.2%*** of 'total_rec_prncp' **1k** loans against 'loan_status' are *Defaults*
→ ***23.7%*** of 'total_rec_prncp' **3k** loans against 'loan_status' are *Defaults*
→ ***11.4%*** of 'total_rec_prncp' **5k** loans against 'loan_status' are *Defaults*

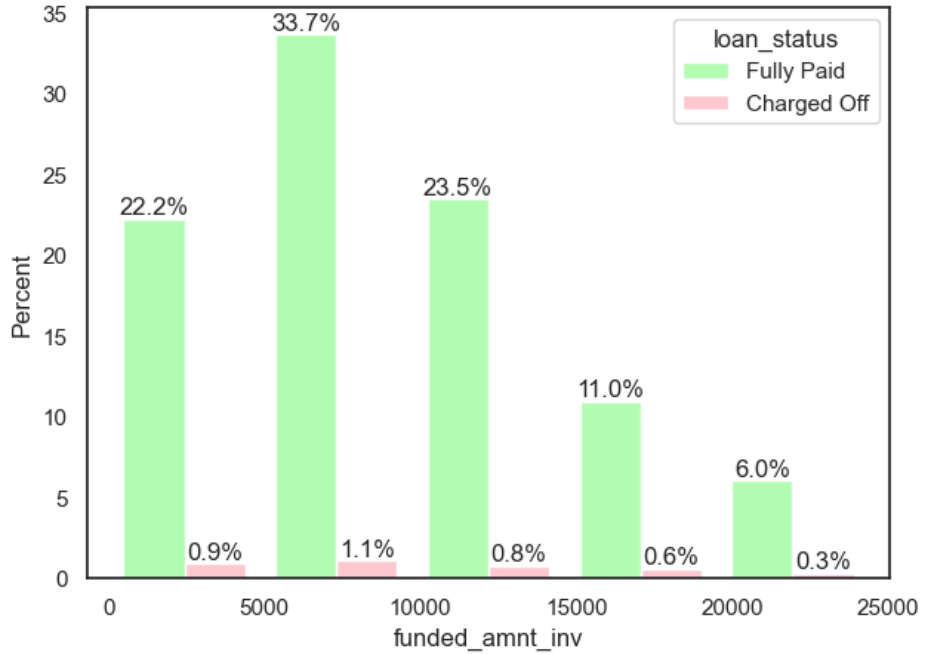
→ *Graph gradually decreased with an increase in 'total_rec_prncp'*

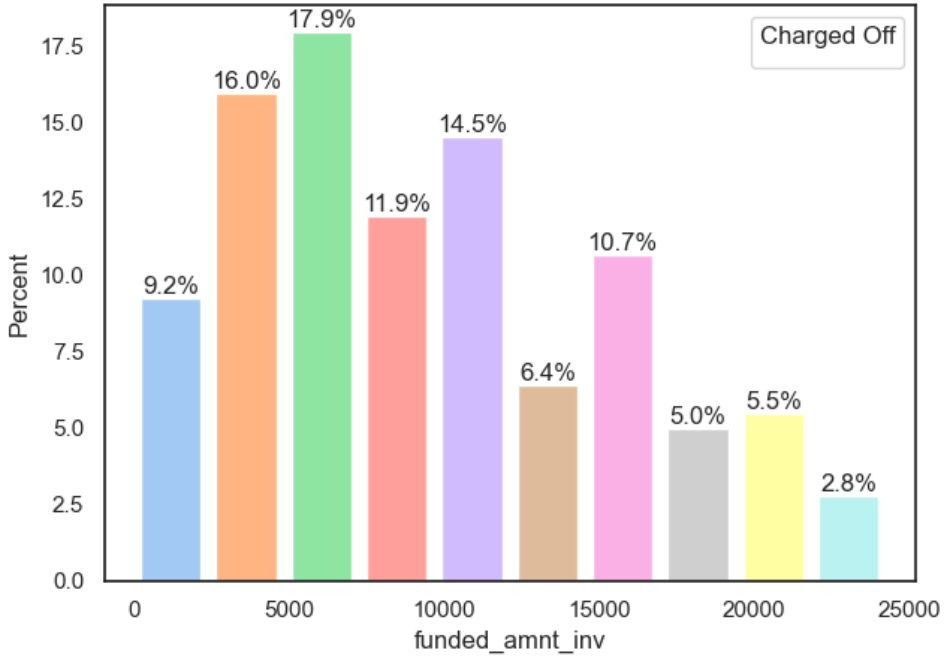
Plots for funded_amnt_inv



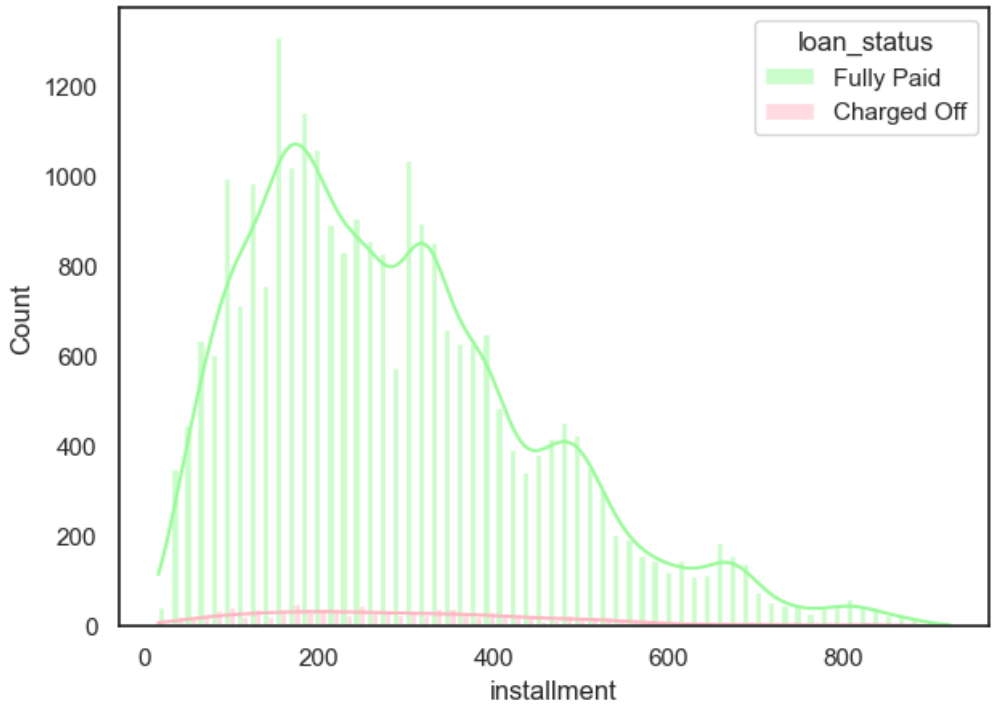
→ ***33.7%*** of 'funded_amnt_inv' **7k** loans against 'loan_status' are *Not-defaults*
→ ***23.5%*** of 'funded_amnt_inv' **11k** loans against 'loan_status' are *Not-defaults*
→ ***22.2%*** of 'funded_amnt_inv' **2k** loans against 'loan_status' are *Not-defaults*

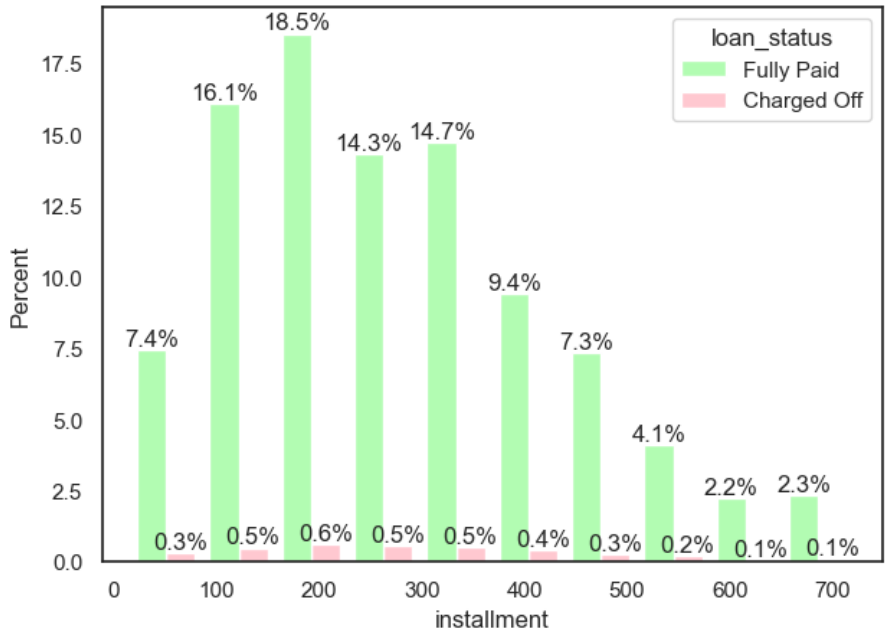
→ ***17.9%*** of 'funded_amnt_inv' **5k** loans against 'loan_status' are *Defaults*
→ ***16.0%*** of 'funded_amnt_inv' **4k** loans against 'loan_status' are *Defaults*
→ ***14.5%*** of 'funded_amnt_inv' **11k** loans against 'loan_status' are *Defaults*

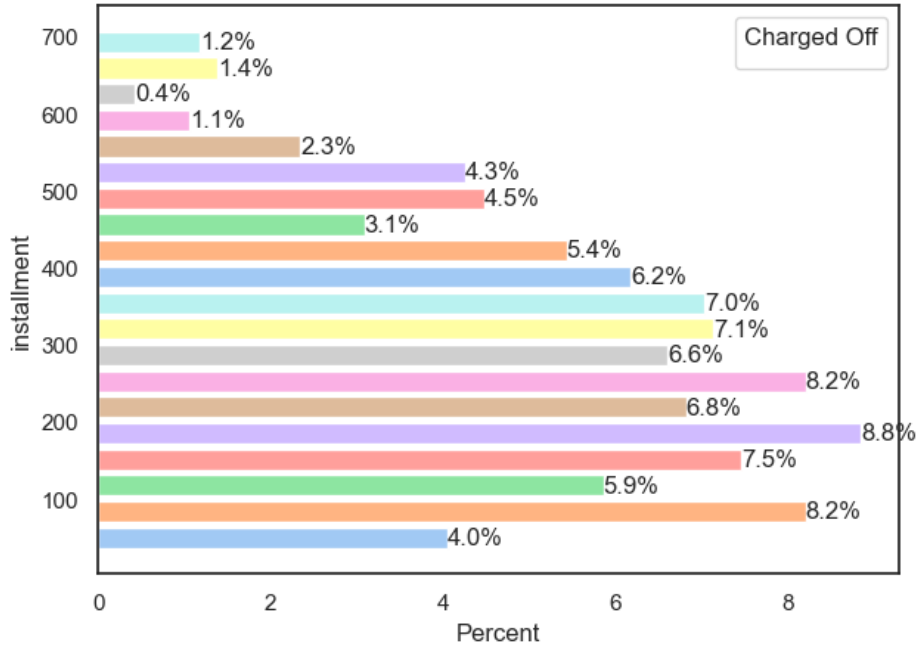




Plots for installment



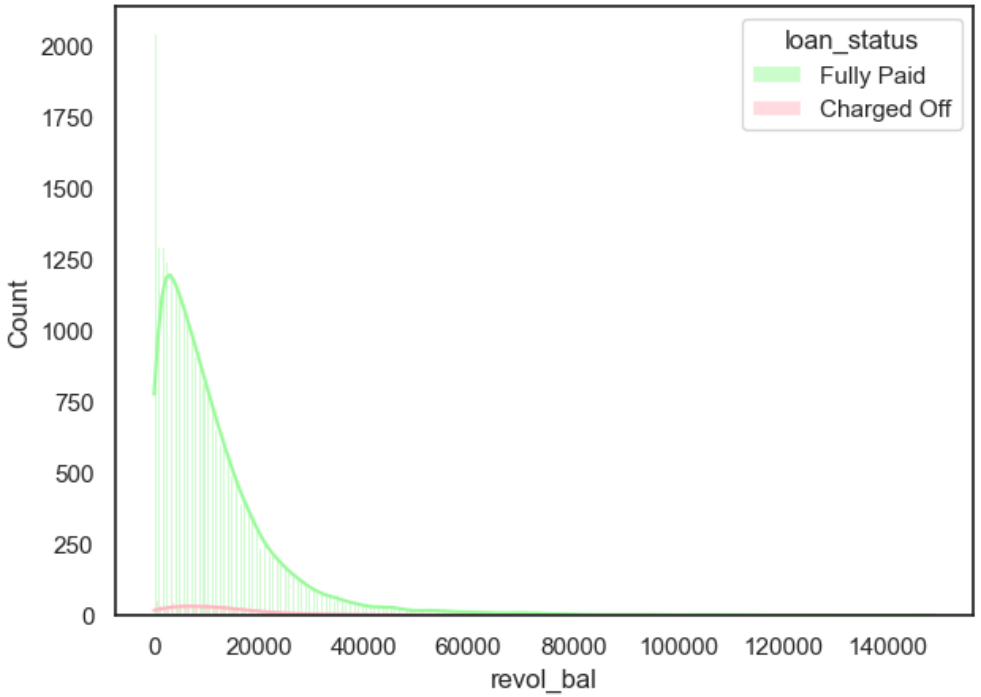




→ ***>14%*** of 'installment' are between **100-300** loans against 'loan_status' are *Not-defaults*

→ ***>5%*** of 'installment' are between **100-400** loans against 'Charged Off' are *Defaults*

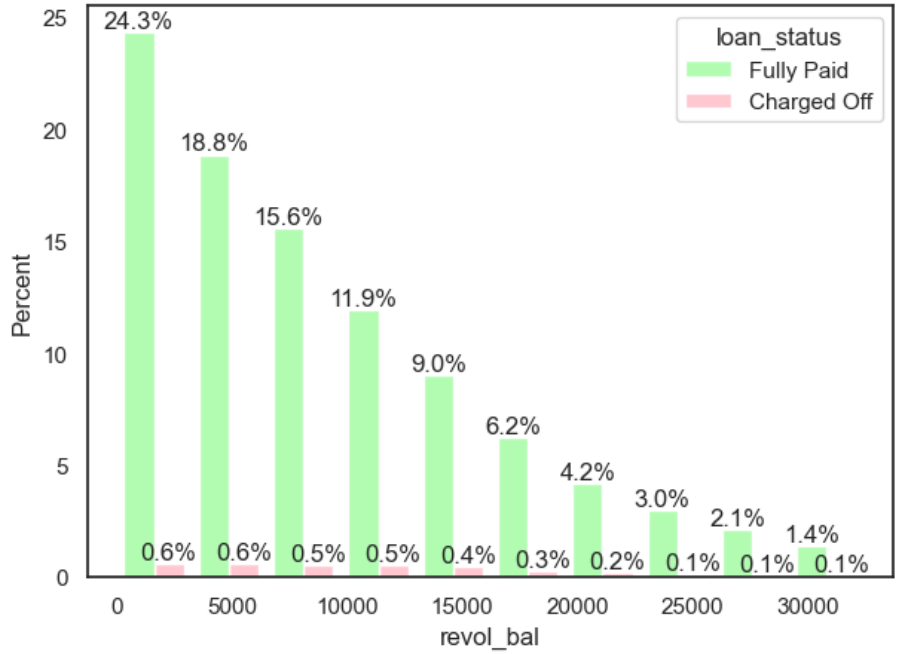
Plots for revol_bal

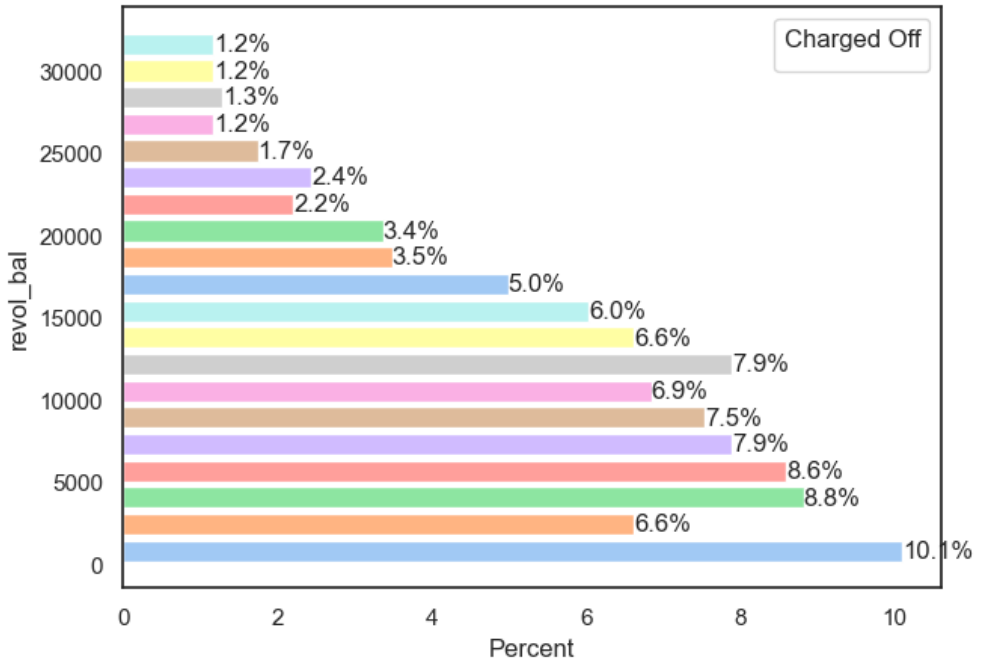


→ LOW 'revol_bal' loans against 'loan_status' are *Not-defaults*

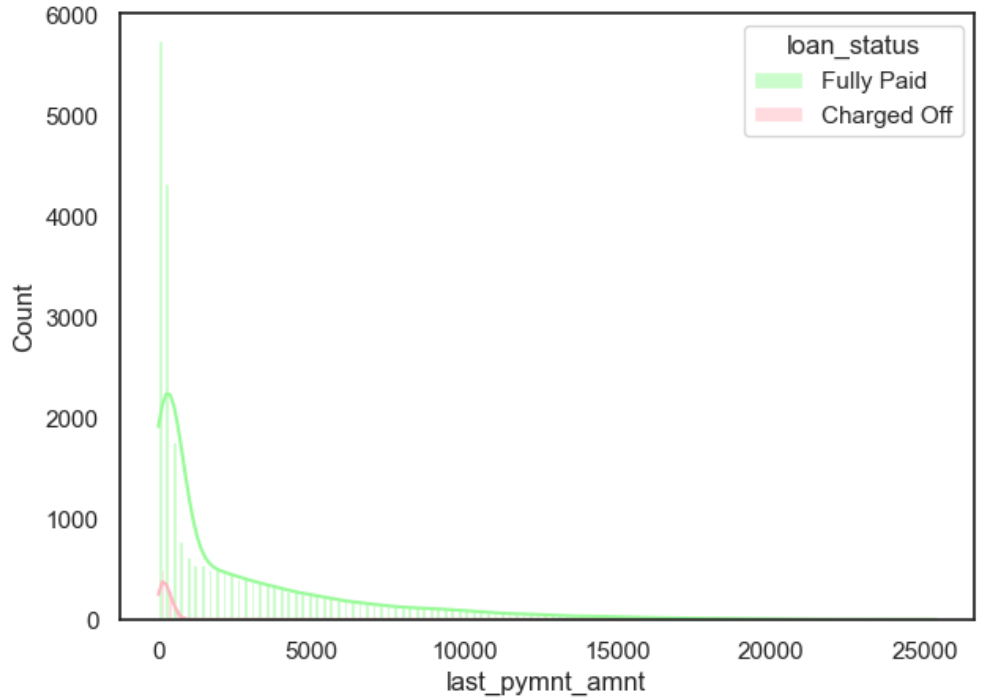
→ *Graph gradually decreased with an increase in 'revol_bal'*

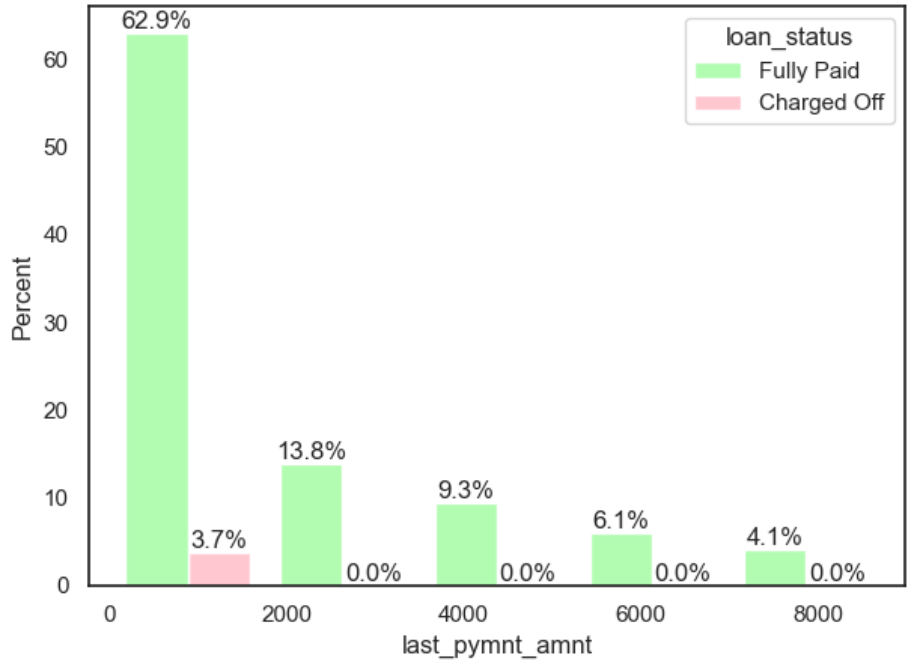
→ LOW 'revol_bal' loans against 'Charged Off' are *Defaults*

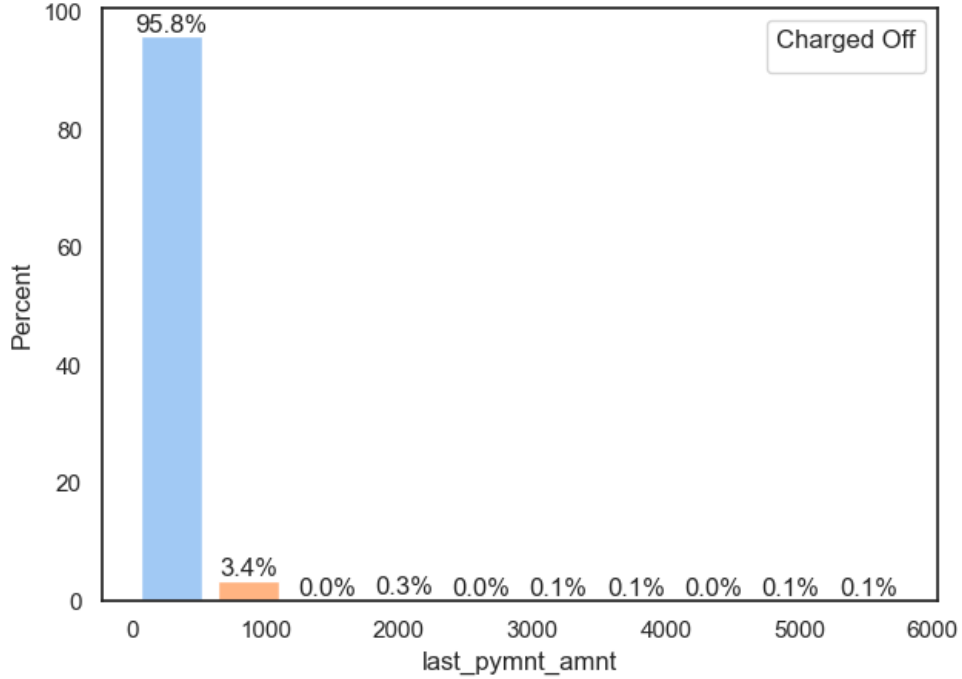




Plots for last_pymnt_amnt



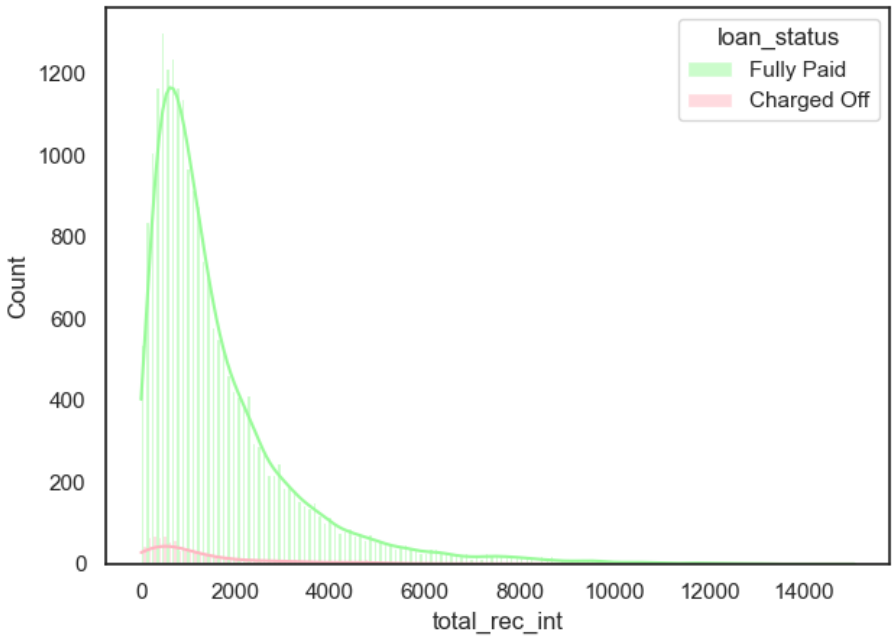




→ ***62.9%*** of 'last_pymnt_amnt' **1k** loans against 'loan_status' are *Not-defaults*

→ ***95.8%*** of 'last_pymnt_amnt' **1k** loans against 'Charged Off' are *Defaults*

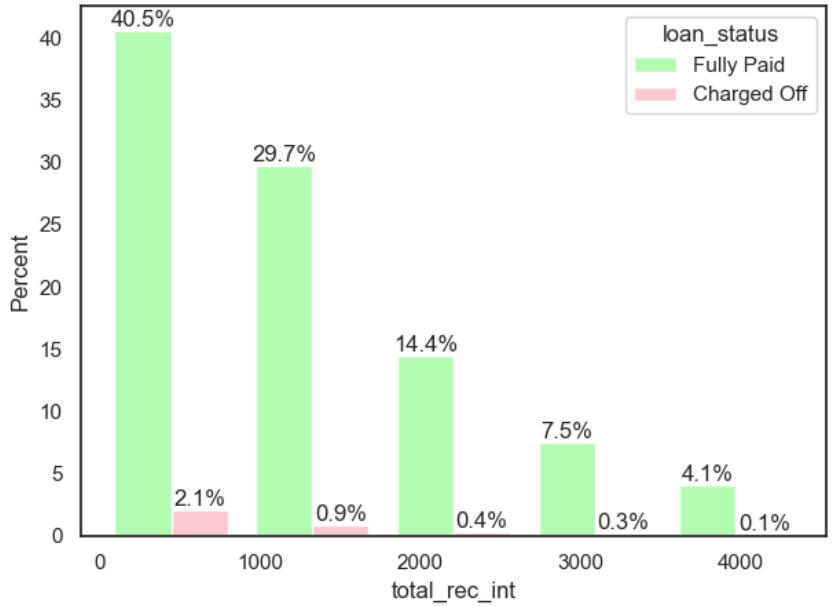
Plots for total_rec_int

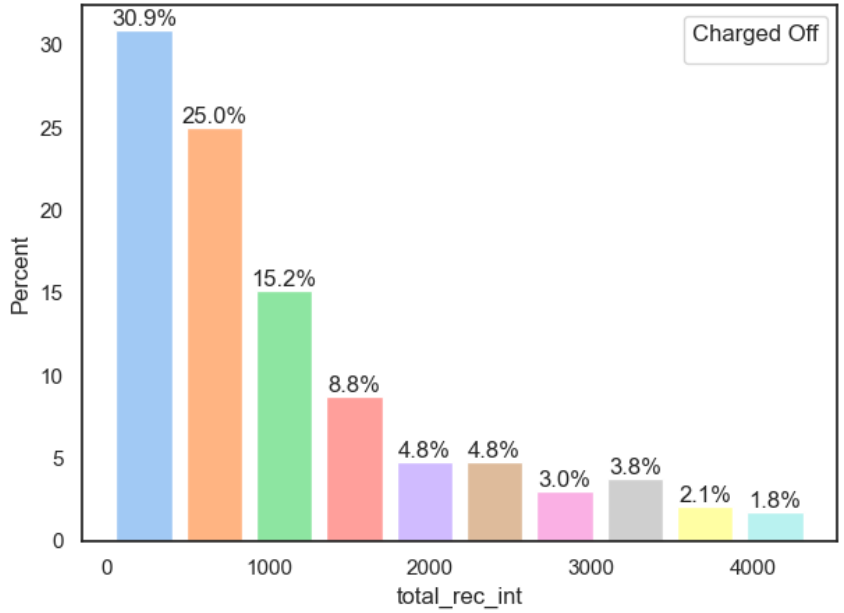


→ LOW 'total_rec_int' loans against 'loan_status' are *Not-defaults*

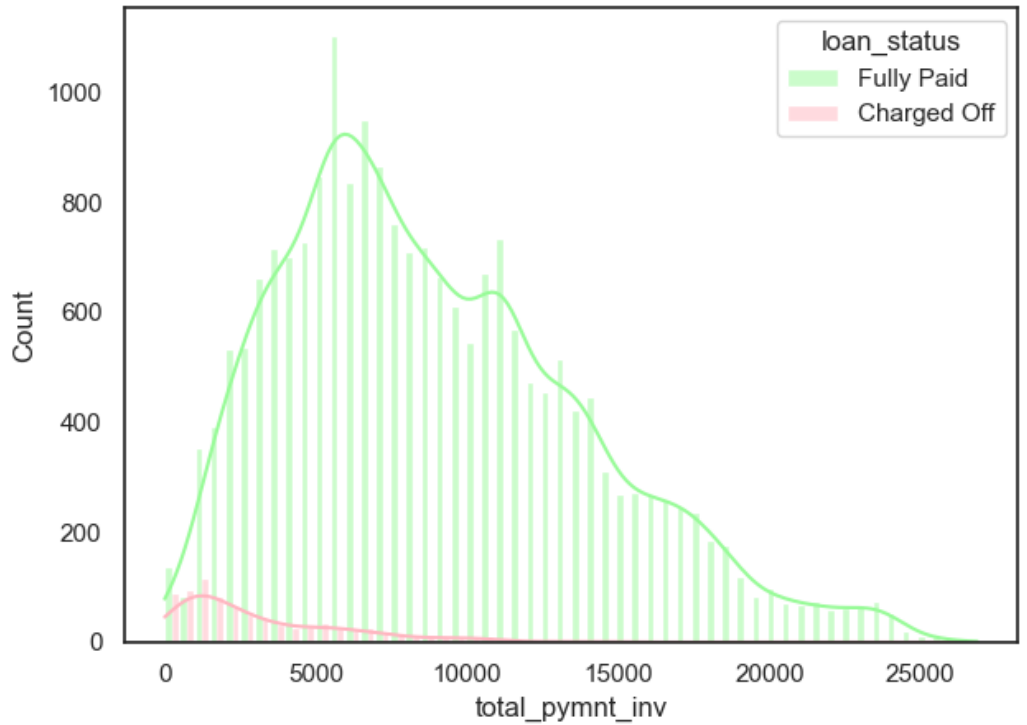
→ LOW 'total_rec_int' loans against 'Charged Off' are *Defaults*

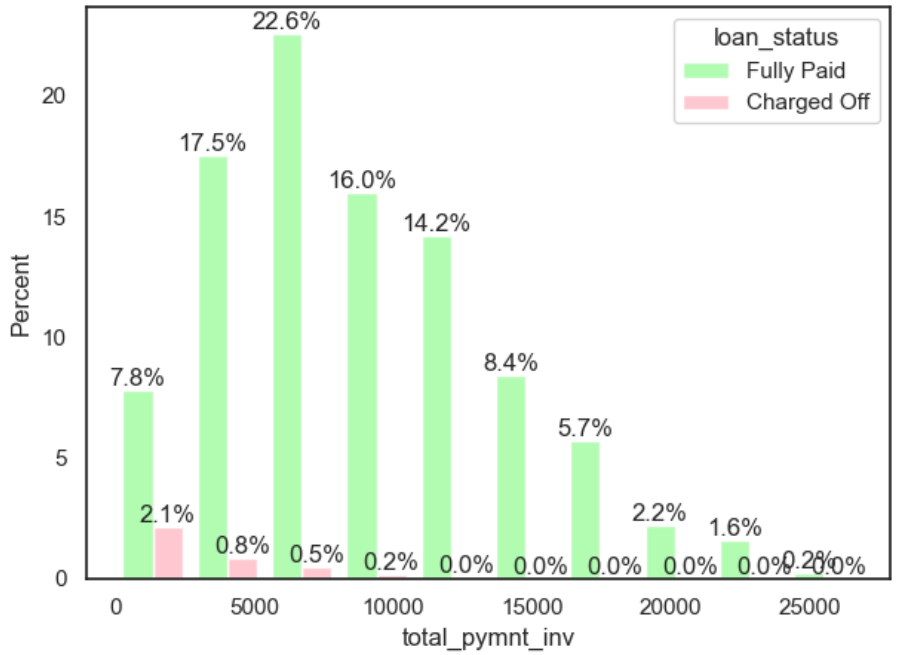
→ *Graphs gradually decreased with an increase in 'total_rec_int'*

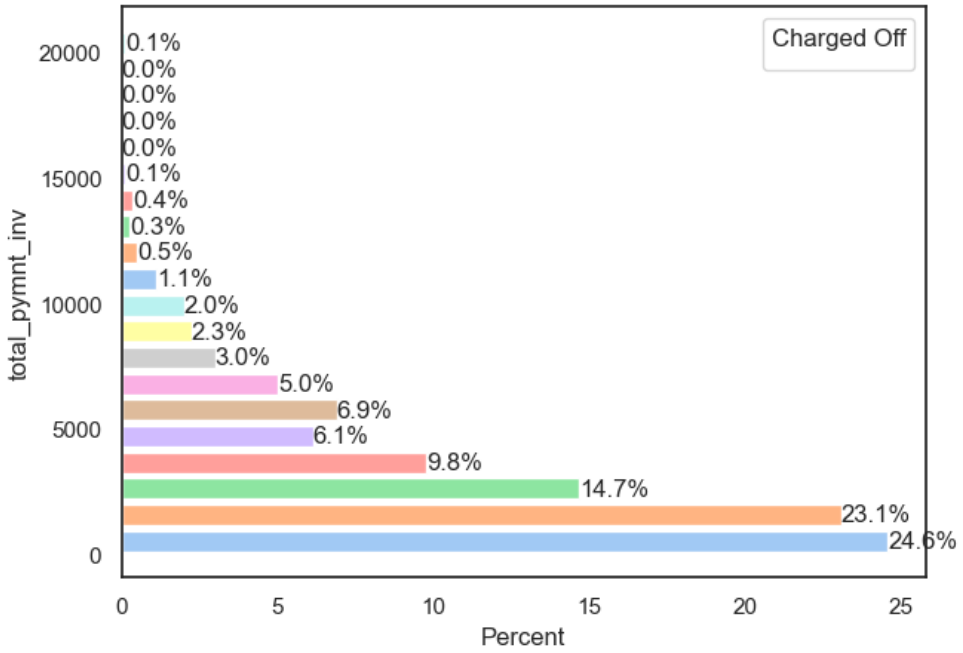




Plots for total_pymnt_inv





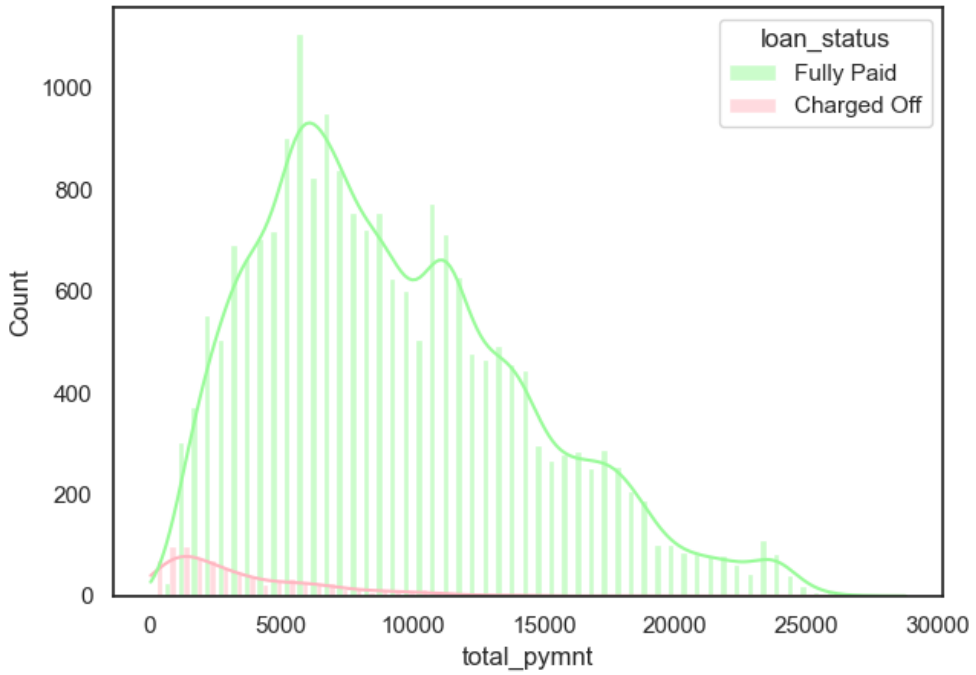


→ ***>16%*** of 'total_pymnt_inv' are between **5k-10k** loans against 'loan_status' are *Not-defaults*

→ LOW 'total_pymnt_inv' loans against 'Charged Off' are *Defaults*

→ *Graph gradually decreased with an increase in 'total_pymnt_int'*

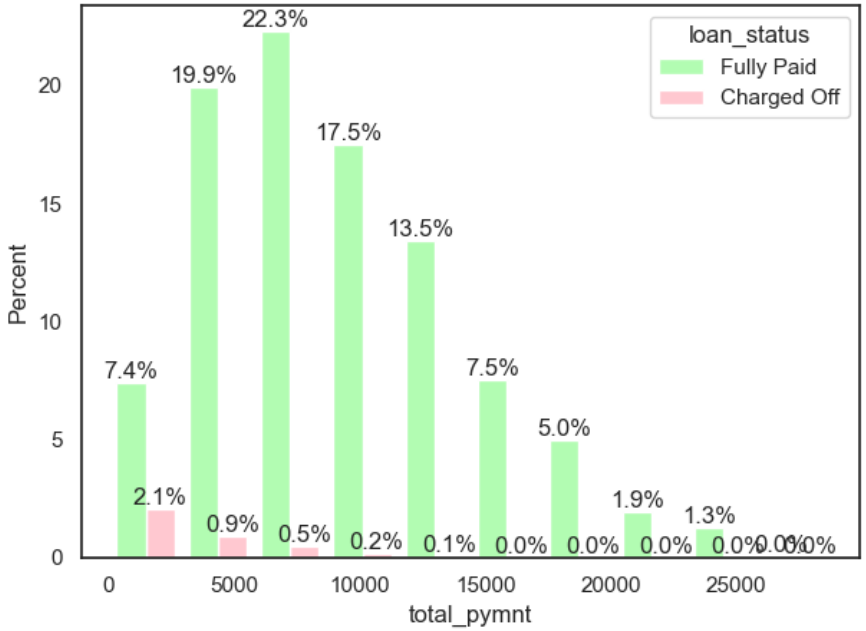
Plots for total_pymnt

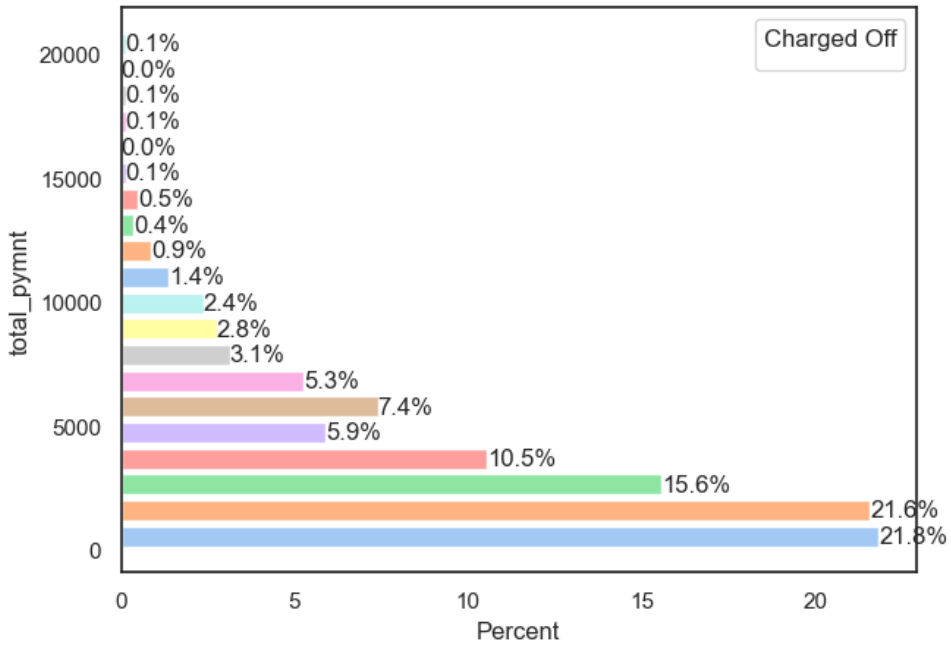


→ ***>17.5%*** of 'total_pymnt' are between **5k-10k** loans against 'loan_status' are *Not-defaults*

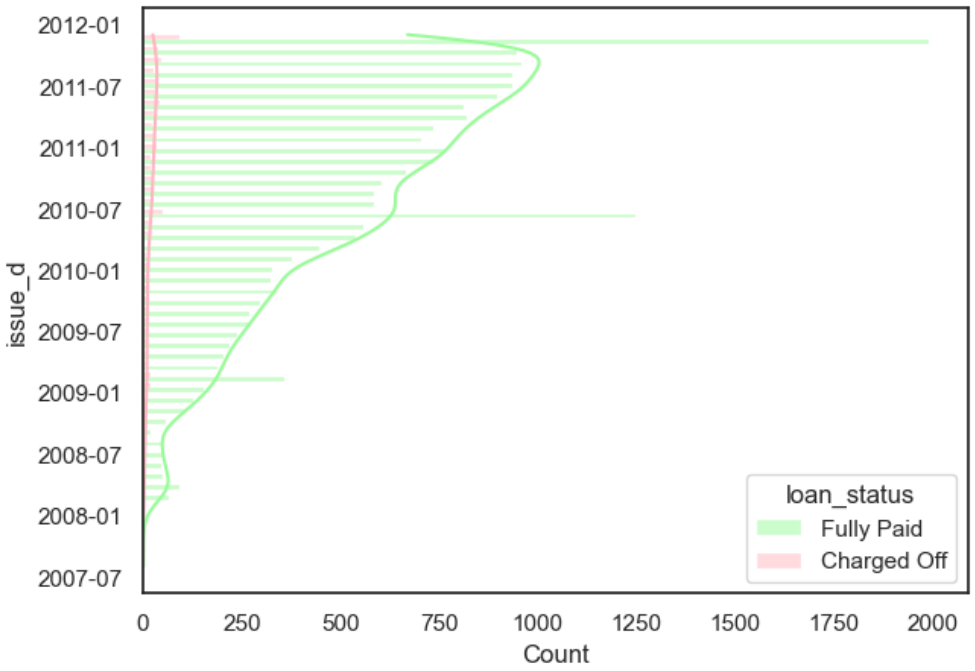
→ LOW 'total_pymnt' loans against 'Charged Off' are *Defaults*

→ *Graph gradually decreased with an increase in 'total_pymnt'*



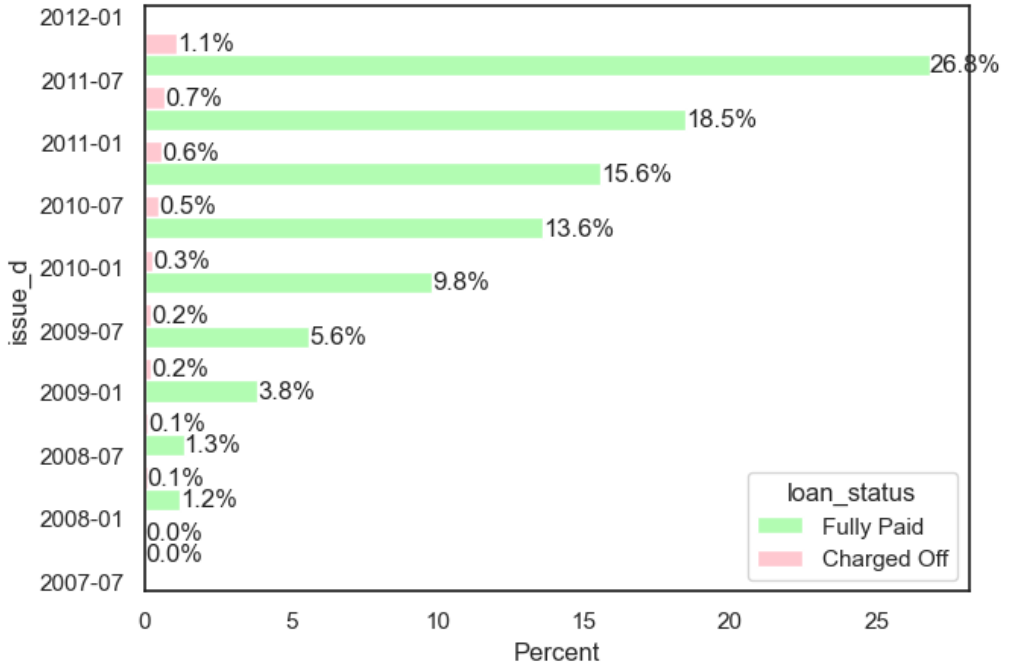


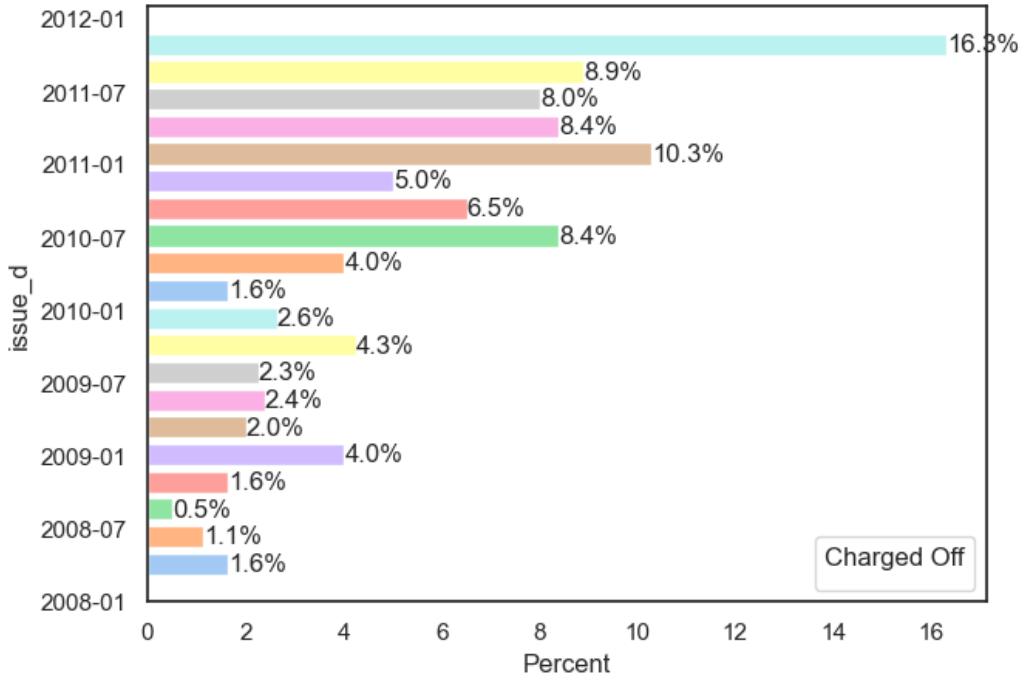
Plots for issue_d



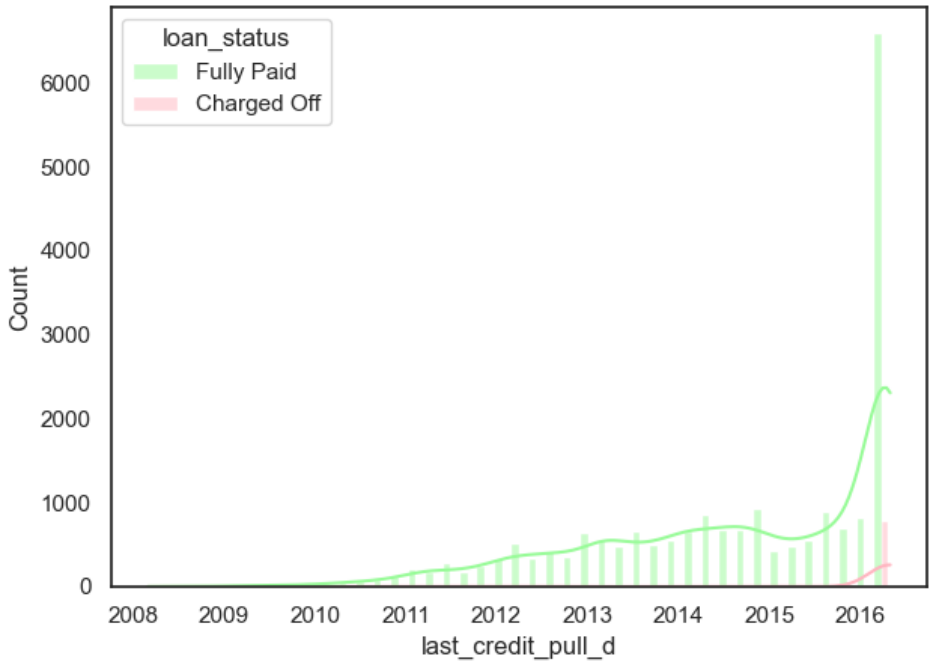
→ RECENT **2012** 'issue_d' loans against 'loan_status' are *Not-defaults*

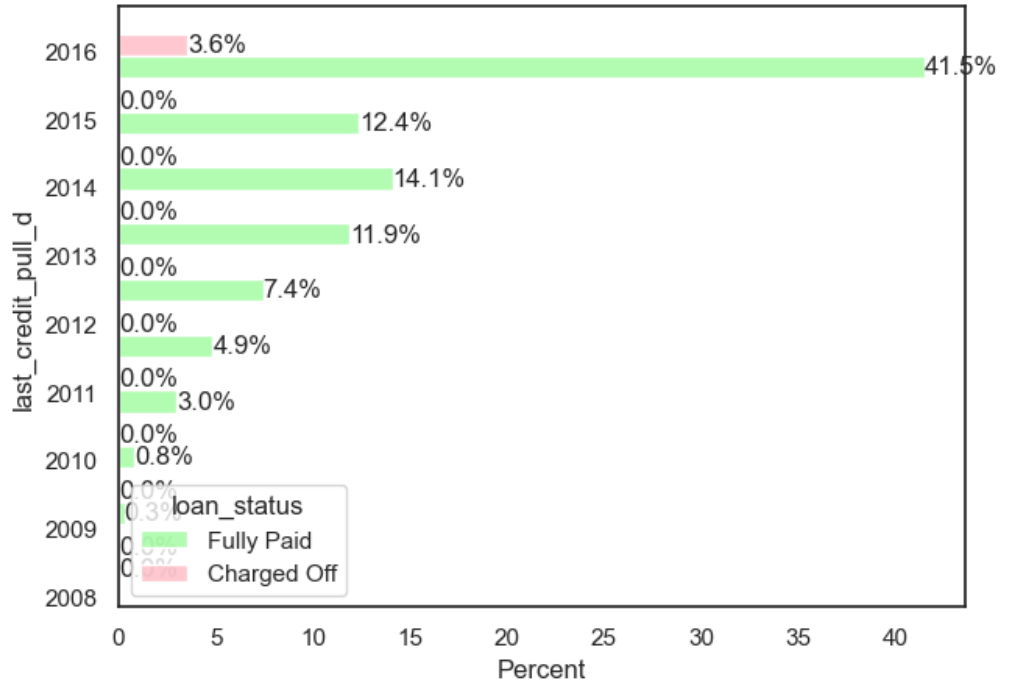
→ **2012** 'issue_d' loans against 'Charged Off' are *Defaults*

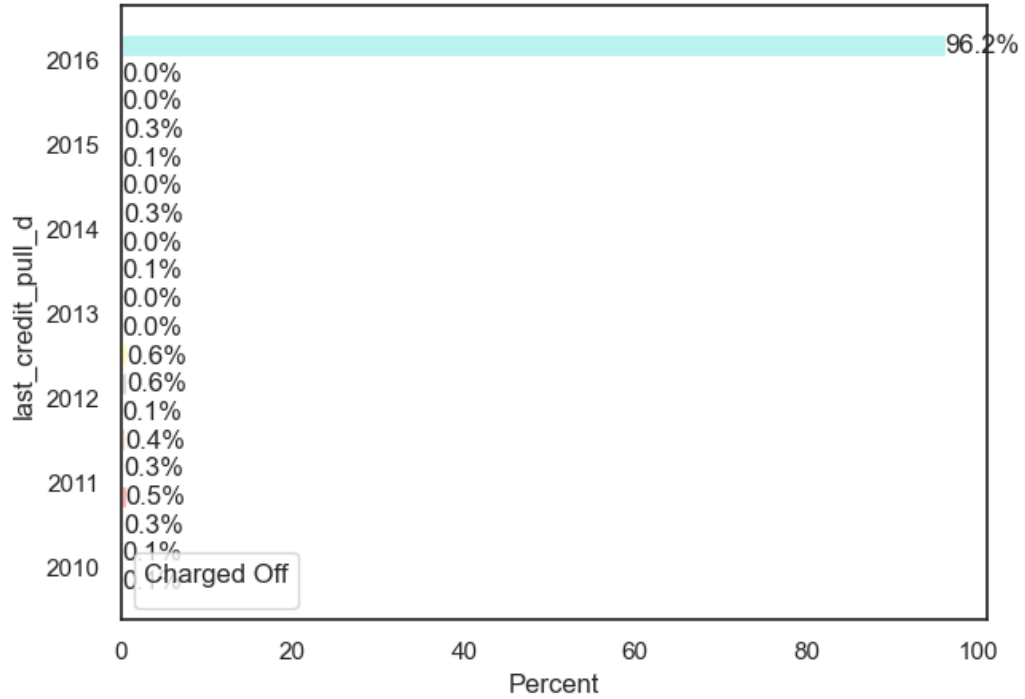




Plots for last_credit_pull_d



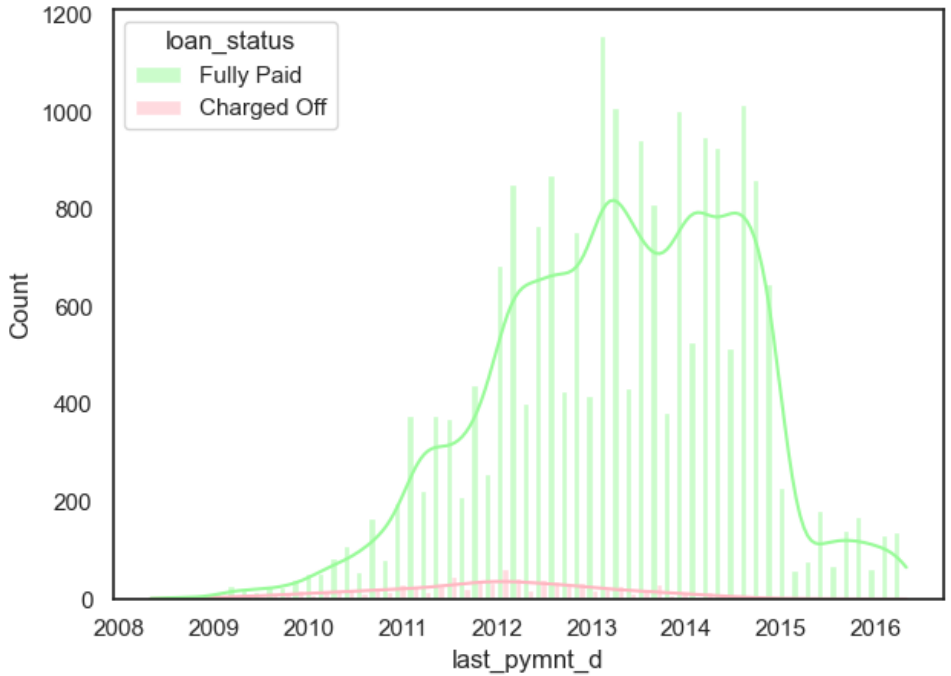




→ ***41.5%*** of 'last_credit_pull_d' year **2016** loans against 'loan_status' are *Not-defaults*

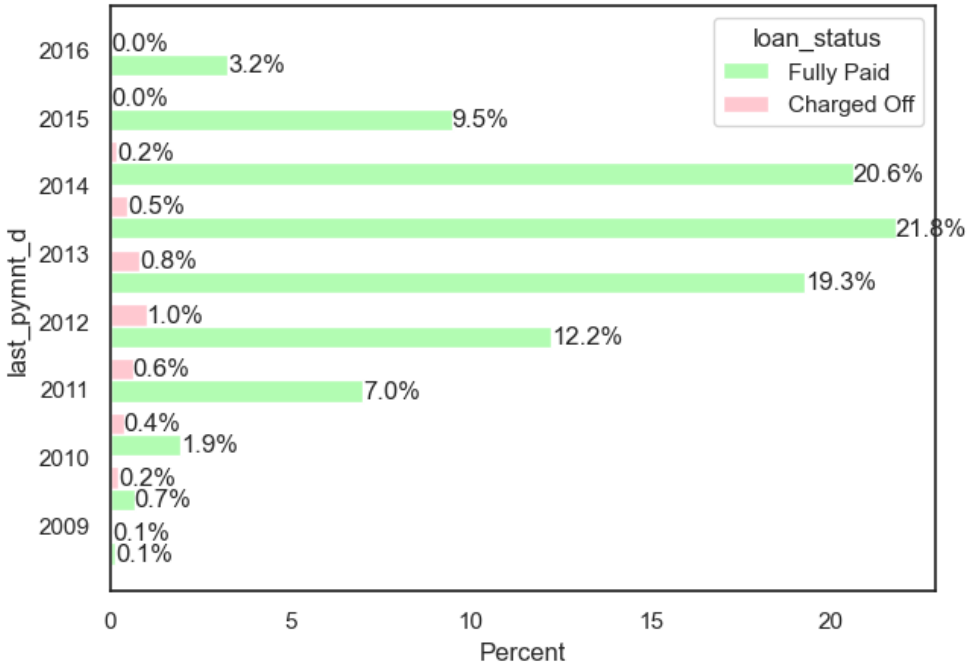
→ ***96.2%*** of 'last_credit_pull_d' year **2016** loans against 'Charged Off' are *Defaults*

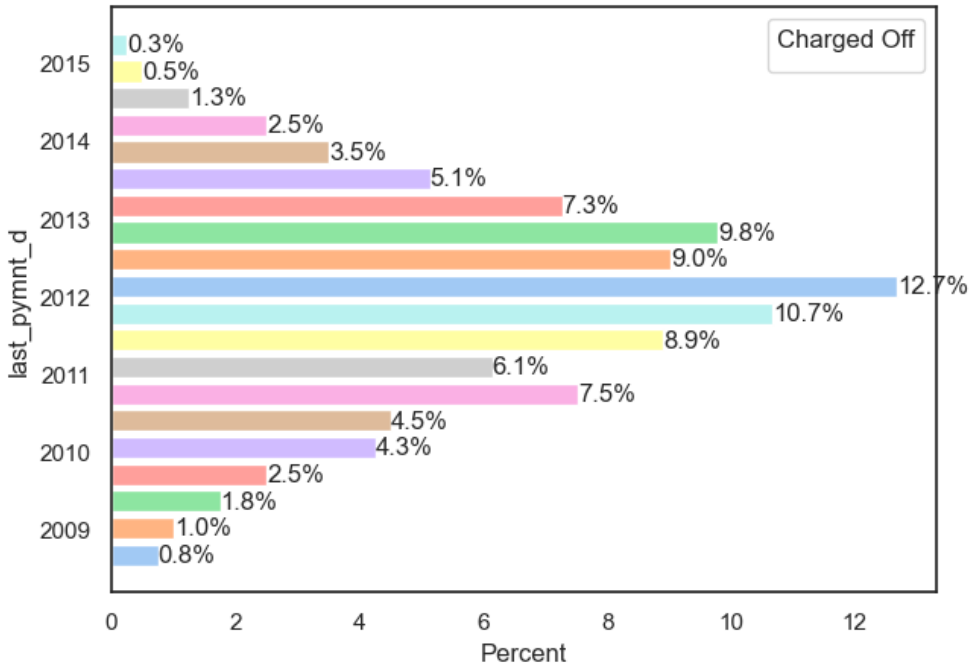
Plots for last_pymnt_d



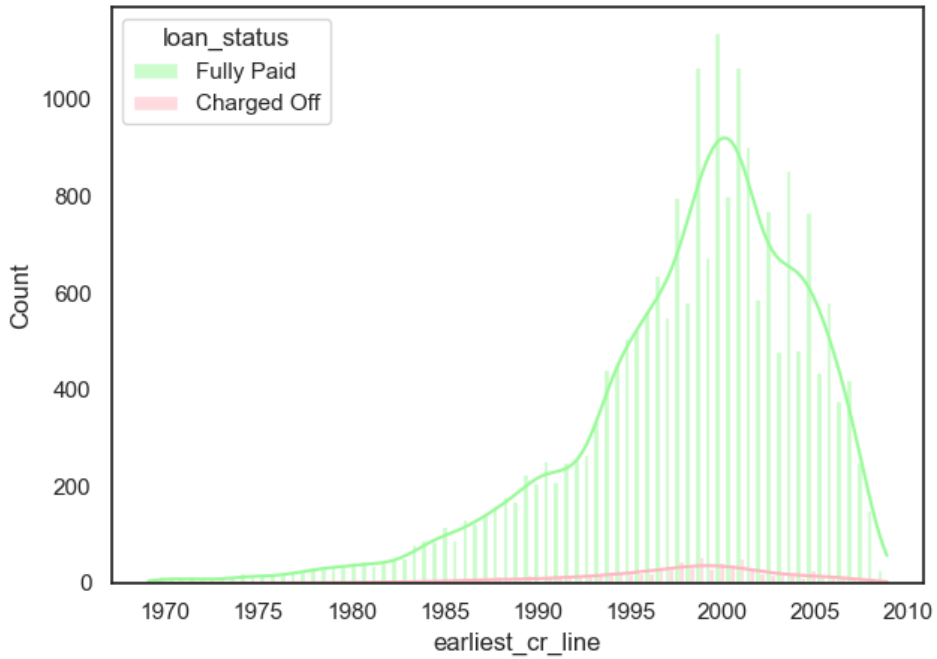
→ ***>20%*** of 'last_pymnt_d' are between year **2013-2014** loans against 'loan_status' are *Not-defaults* [after COVID]

→ ***12.7%*** of 'last_pymnt_d' year **2012** loans against 'Charged Off' are *Defaults* [COVID]



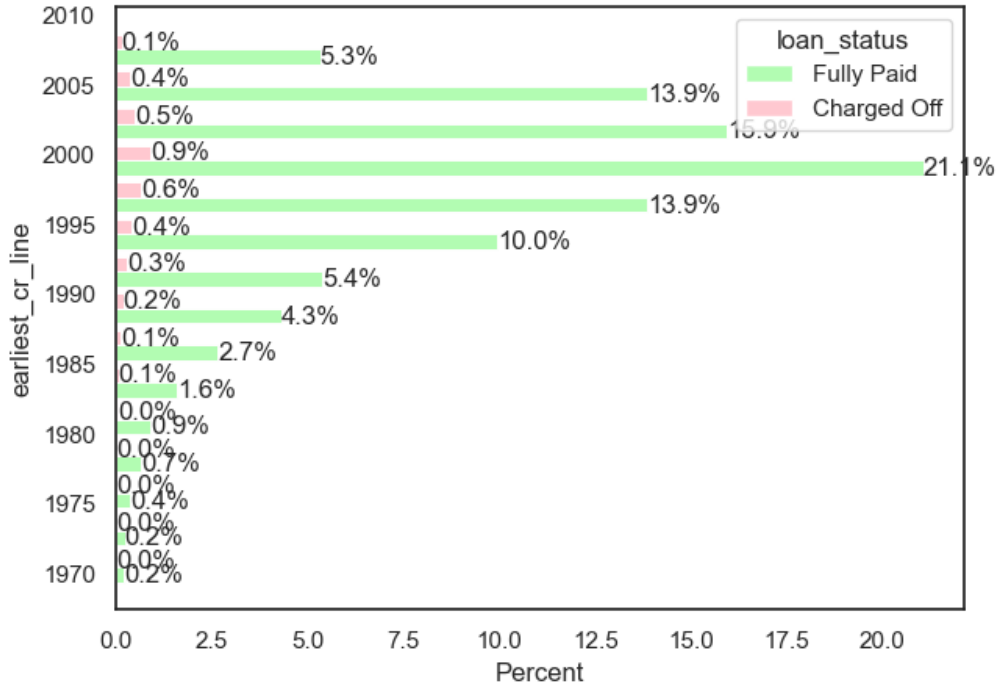


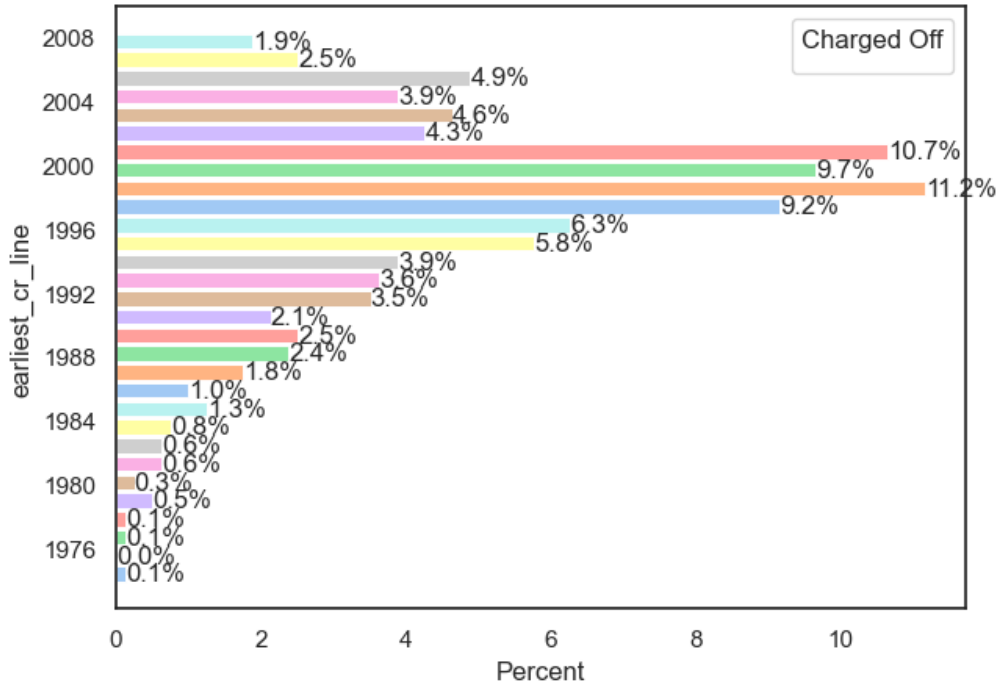
Plots for earliest_cr_line



→ ***>14%*** of 'earliest_cr_line' are between years **1995-2005** loans against 'loan_status' are *Not-defaults*

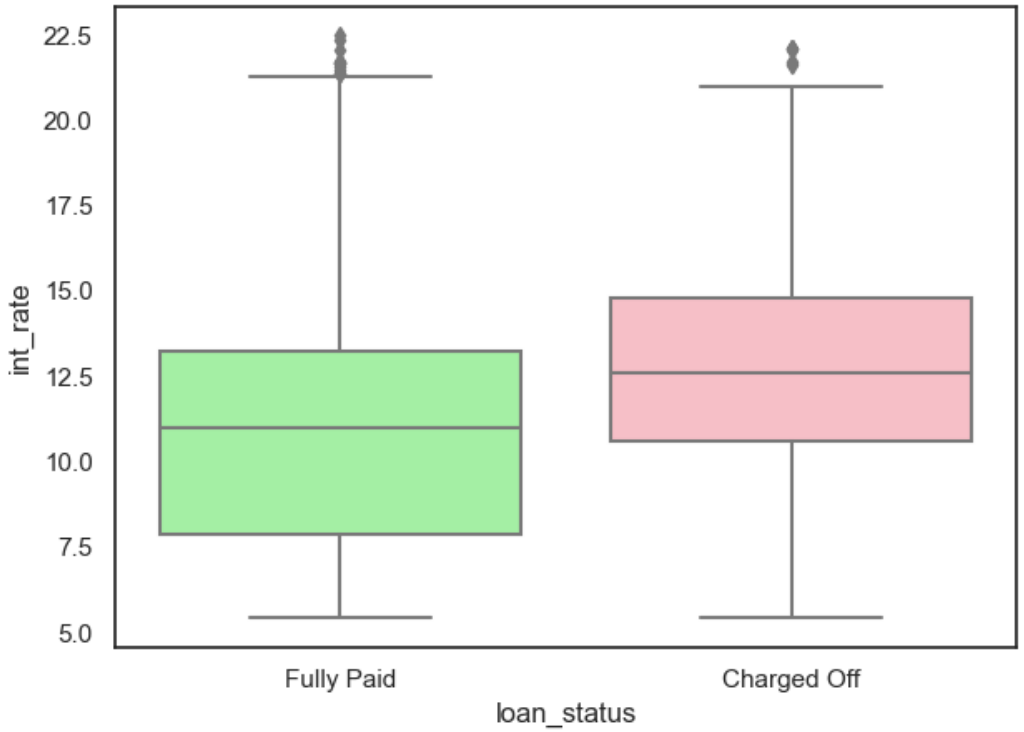
→ ***>9%*** of 'earliest_cr_line' are between years **1998-2001** loans against 'Charged Off' are *Defaults*





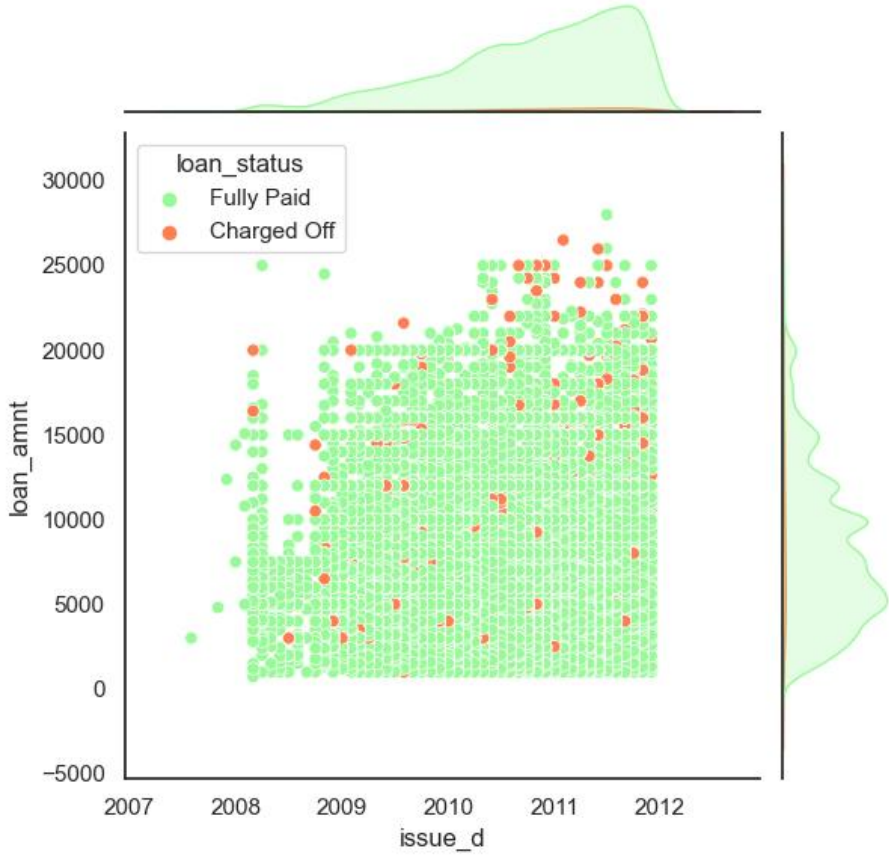
Bivariate Analysis

loan_status vs int_rate



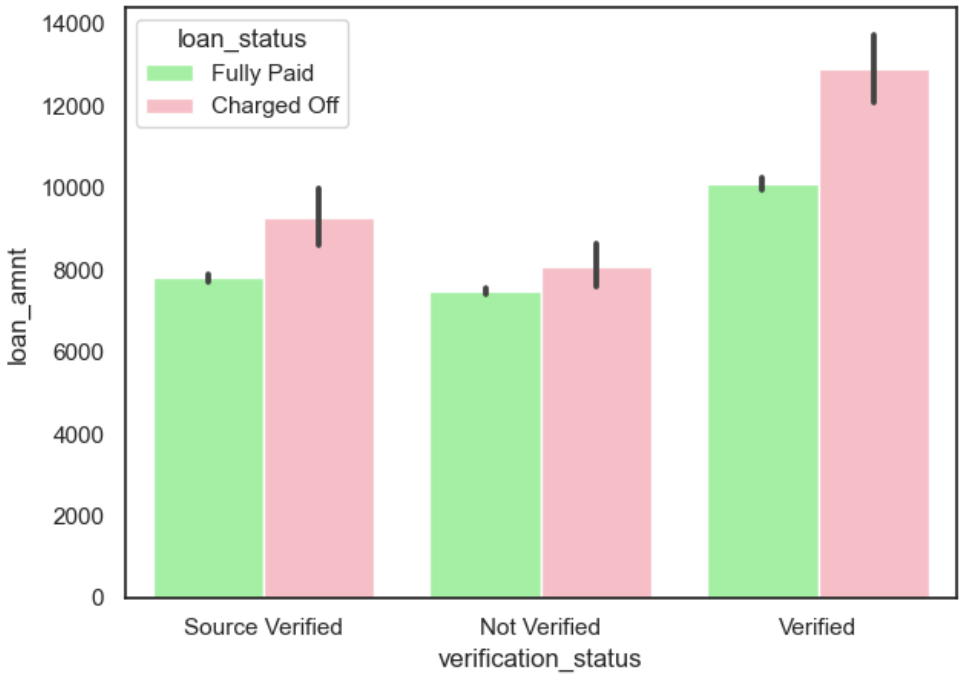
→ HIGH 'int_rate' have more *Defaults* but needs to impose HIGH 'int_rate' on *Defaults* [***CONTRADICT***]

issue_d vs loan_amnt vs loan_status



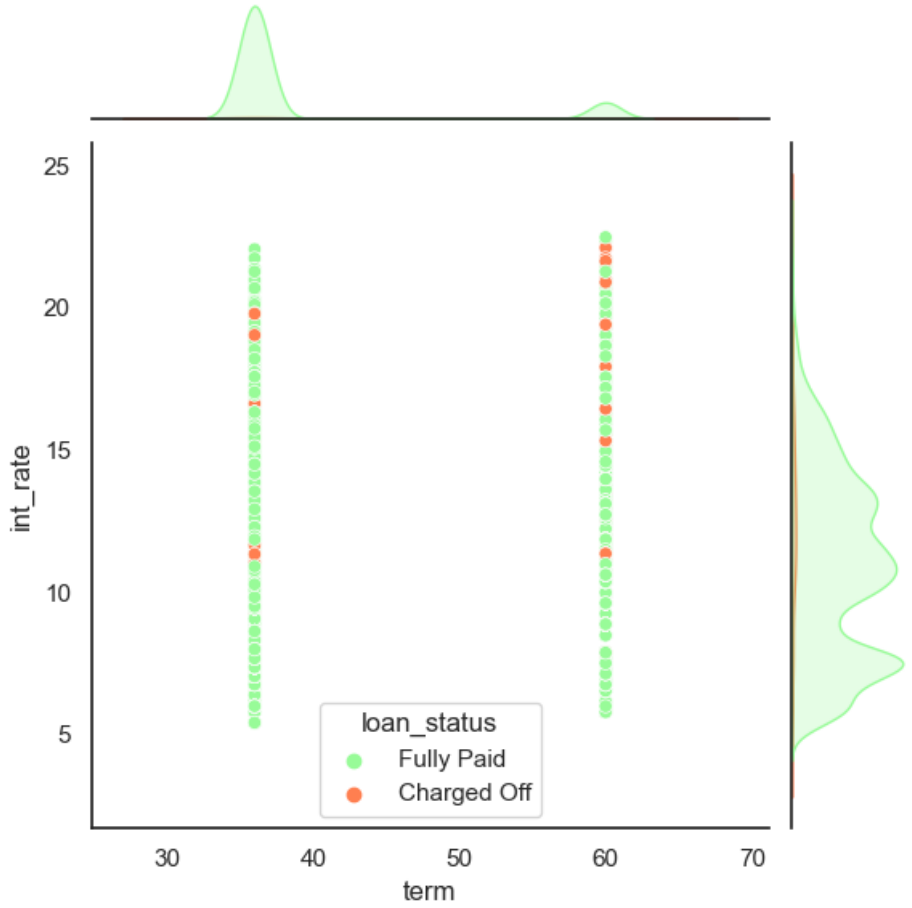
→ 'loan_amnt' increased with recent years, HIGH 'loan_amnt' after COVID have high *Defaults* [***RISK***]

verification_status vs loan_amnt vs loan_status



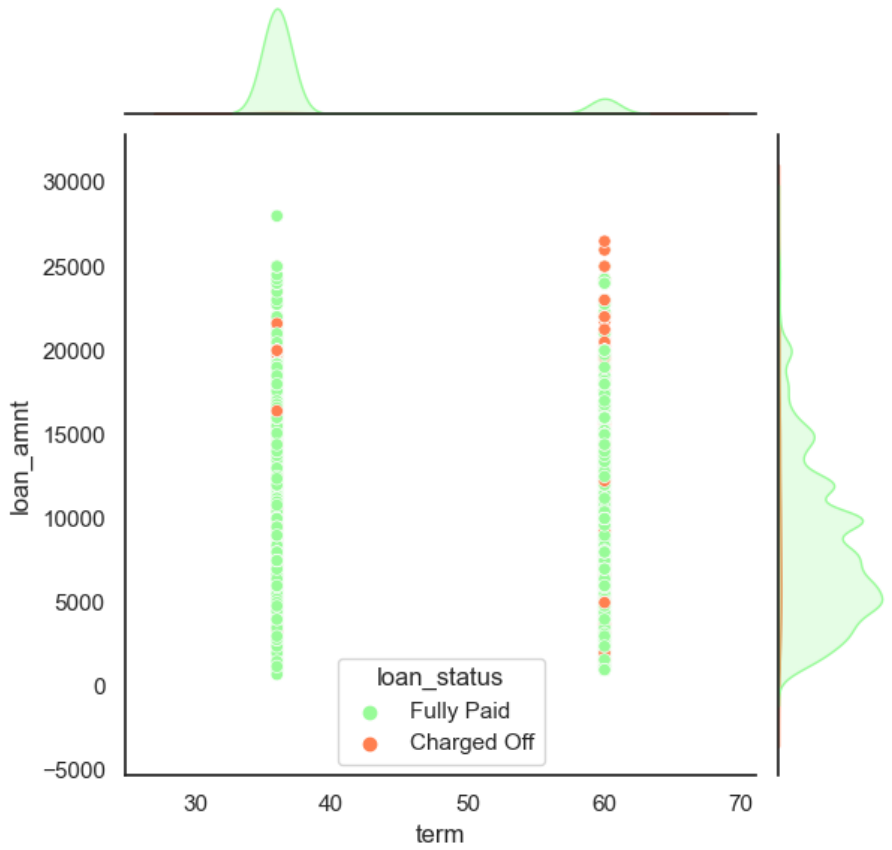
→ HIGHER 'loan_amnt' are **Verified** more often and have huge *Defaults* [***RISK***]

term vs int_rate vs loan_status



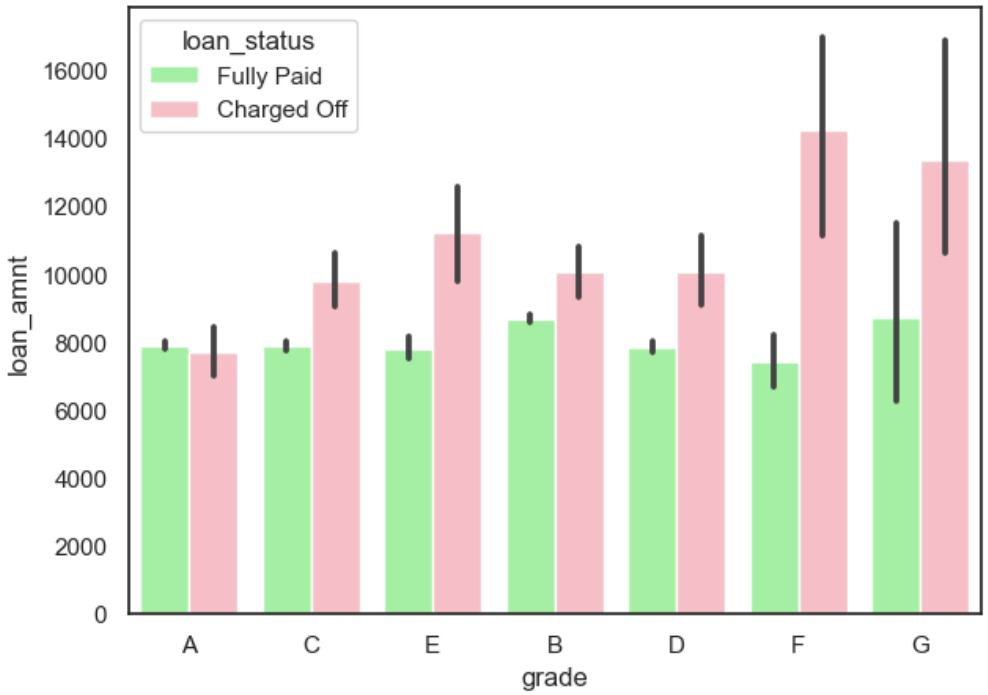
→ HIGH 'int_rate' for LONG 'term' has more *Defaults* [***RISK***]

term vs loan_amnt vs loan_status



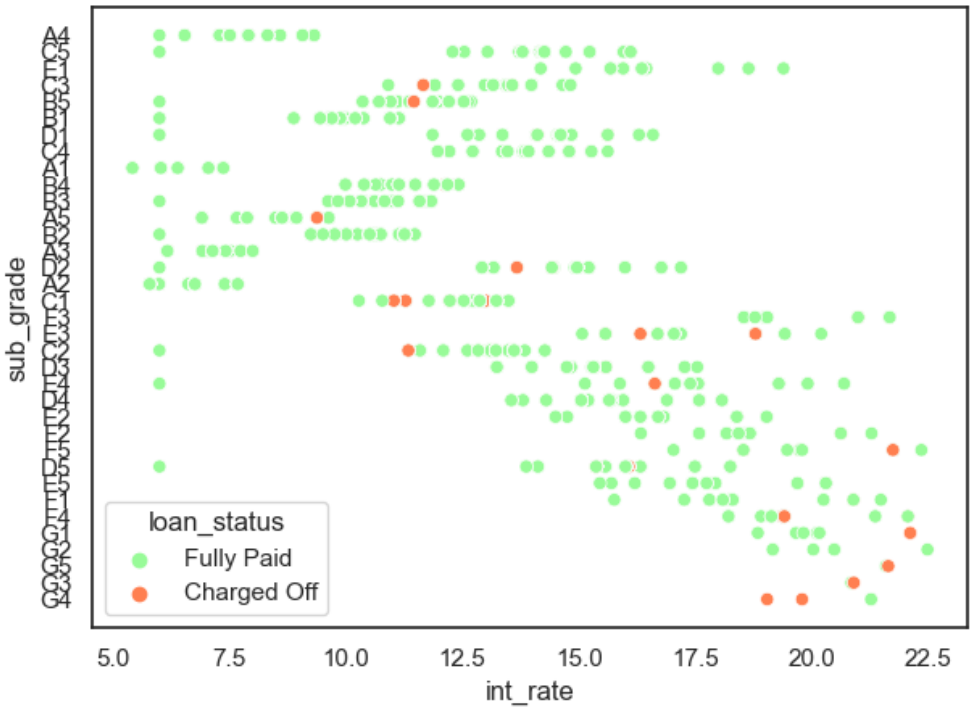
→ HIGH 'loan_amnt' for LONG 'term' have high *Defaults* [***RISK***]

grade vs loan_amnt vs loan_status



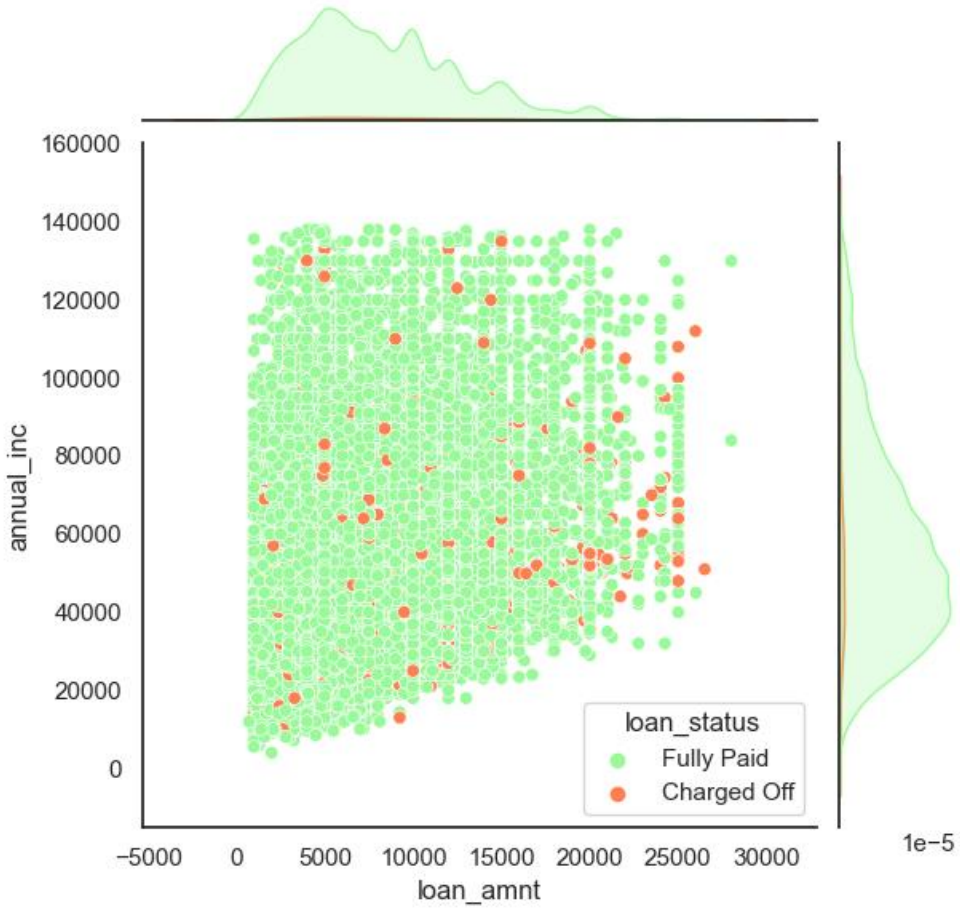
→ HIGHER 'loan_amnt' for LOWER 'grade' are high *Defaults* [***RISK***]

int_rate vs sub_grade vs loan_status

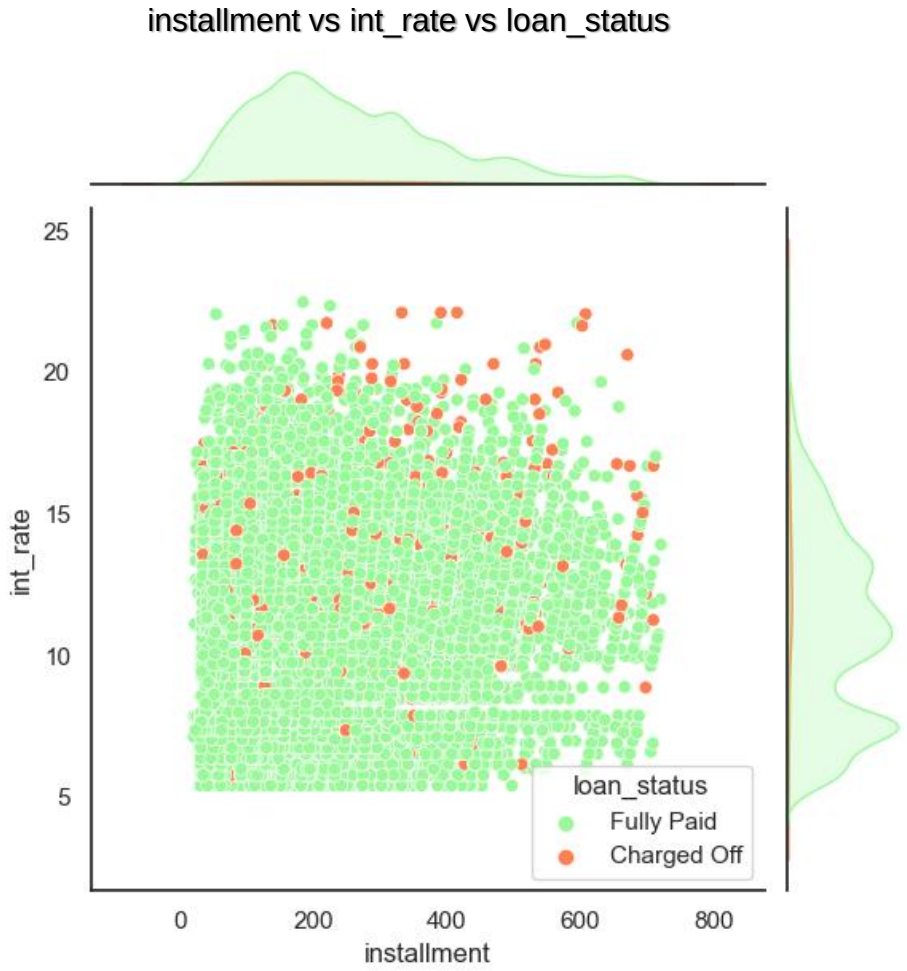


→ LOWER 'sub_grade' have HIGHER 'int_rate' [***SAFE***]

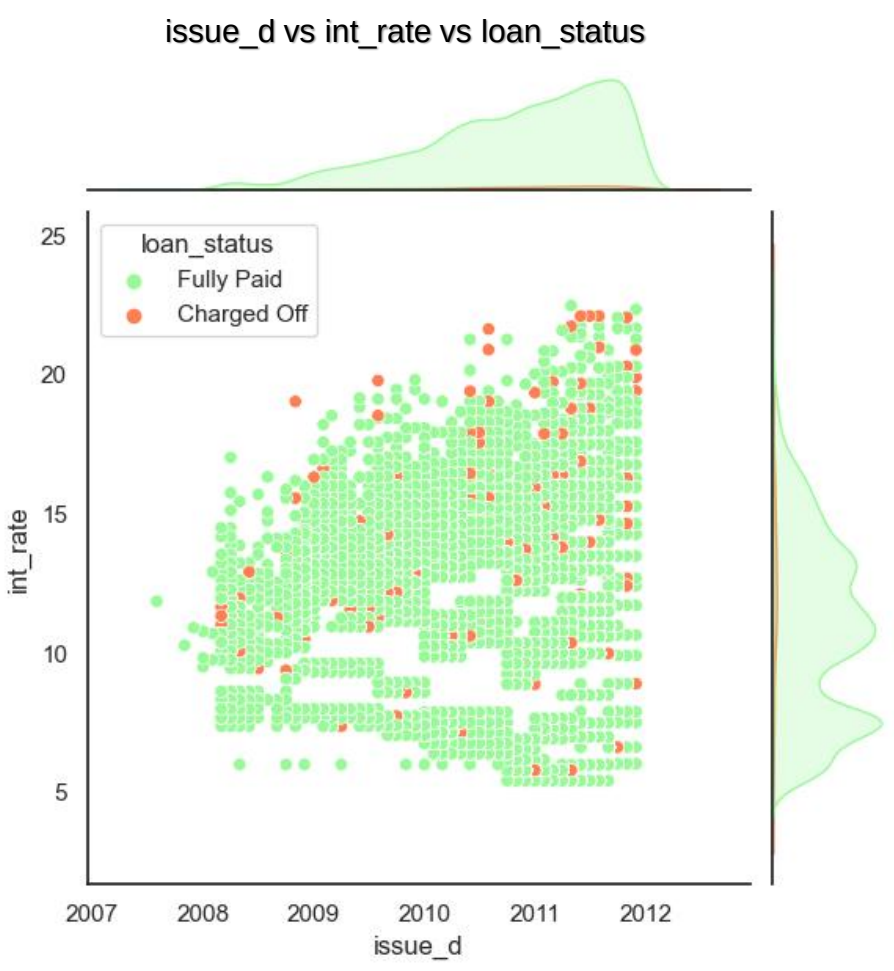
loan_amnt vs annual_inc vs loan_status



→ HIGHER 'loan_amnt' for LOWER 'annual_inc' have *Defaults* [***RISK***]

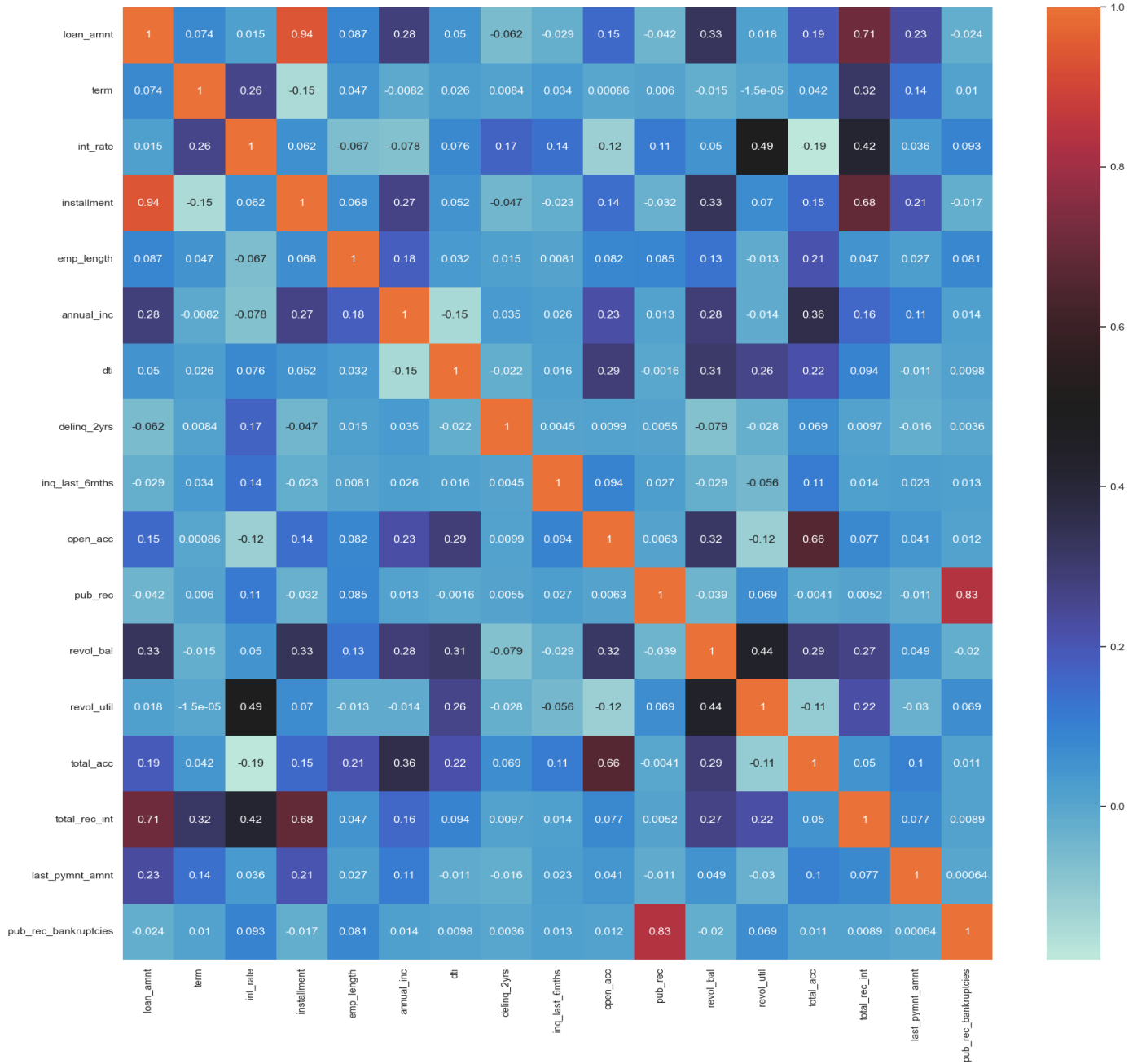


→ HIGH 'int_rate' for LONG 'instalment' have high *Defaults* but to Compensate it is necessary [***CONTRADICT***]



→ HIGH 'int_rate' after COVID have high *Defaults* [***RISK***]

HEAT-MAP



→ 'int_rate' is 49% correlated to 'revol_util' [***SAFE***]
→ 'loan_amnt' is 33% correlated with 'revol_bal' [***RISK***]