EDA CASE STUDY: LENDING CLUB CASE

Prepared By:
Rijumani Deka
Saad Teli

PROBLEM STATEMENT

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.
- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

ANALYSIS STEPS FOLLOWED:

- Data understanding & preparation
- Data cleaning & manipulation
- Data analysis:
  - Univariate
  - Bivariate
- Presentation
- Conclusions

Data Understanding & Preparation

- Import and read data from .csv file creating data frame
- Inspect and analyze the shape of the data frame.
- Check if columns with entire null or missing values present in data frame.
- Find out unnecessary or not in use for pre-approval scenario columns

Data Cleaning & Manipulation

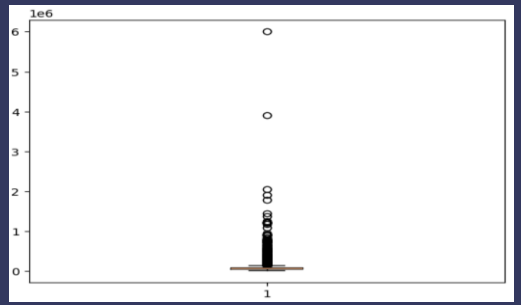
- Drop Columns with all Null values.
- Drop columns with more than 90% values as 0.
- Some columns after more analysis were found to be single valued, which doesn't really give any insight so have dropped those.
- We are dealing with a situation which occurs before a loan is approved, to make decision whether to approve a loan or not. Some of the fields/Columns seem to have no use during the process of loan approval, have removed those.
- We see that there are three types of loan status, i.e., 'Fully Paid', 'Charged Off', and 'Current'. We need to find out who are most likely to default loan repayment, and loans with Current status cannot give an insight. So, removed loans records with 'Current' status

Dealing with data quality issues

- The data frame initially had 111 columns.
- After removing columns with a majority of missing (NaN) values, we were left with 57 columns.
- Next, columns where more than 90% of the values were zero were removed reducing the number of columns to 46.
- We further eliminated columns with only a single unique value, leaving us with 43 columns.
- Finally, we reviewed the remaining columns for relevance and kept only those useful for analysis, resulting in a final dataset of 20 columns.
- Next, we checked for columns with missing value records. ‘emp_length’ & ‘revol_util’ were having missing values out of the 20 columns. As, ‘emp_length’ percentage of missing values is very low, we decided it was ok to input some default values. To deal with missing values of ‘revol_util’ we decided to drop missing value records.
- We decided to standardized ‘emp_length’, ‘revol_util’ & ‘int_rate’, after further inspection.

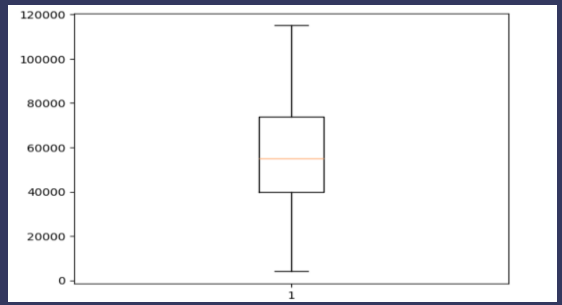
CHECKING IF ANNUAL INCOME HAS ANY OUTLIER

1.



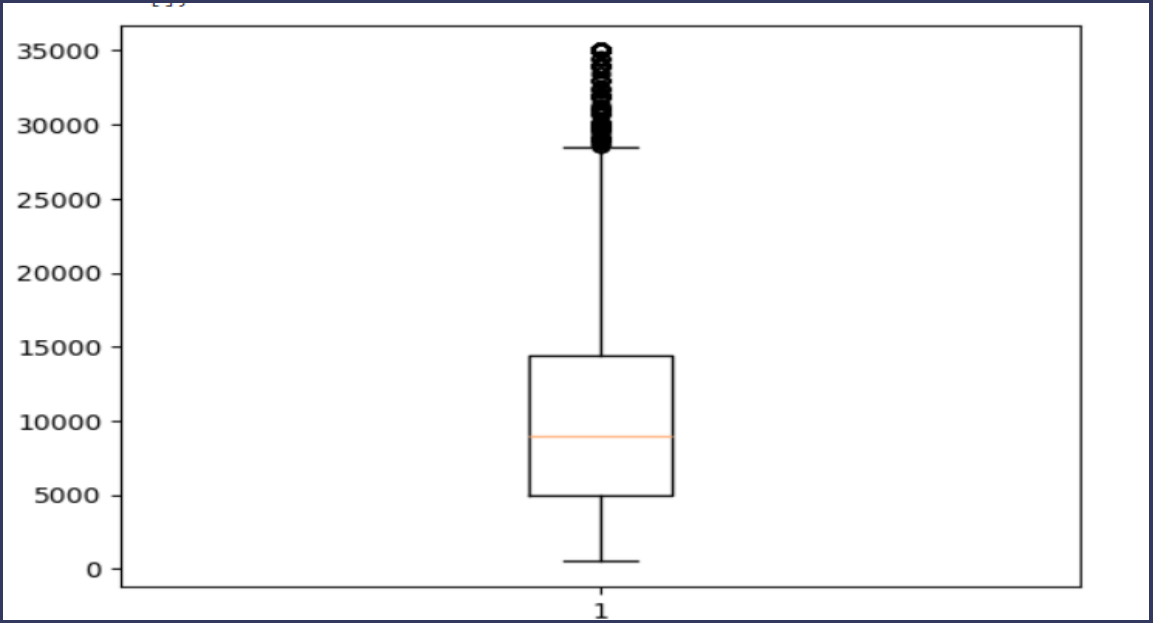
Outliers Exist, so we needed to deal with them.
Found quantile information for ‘annual_inc’.

2.



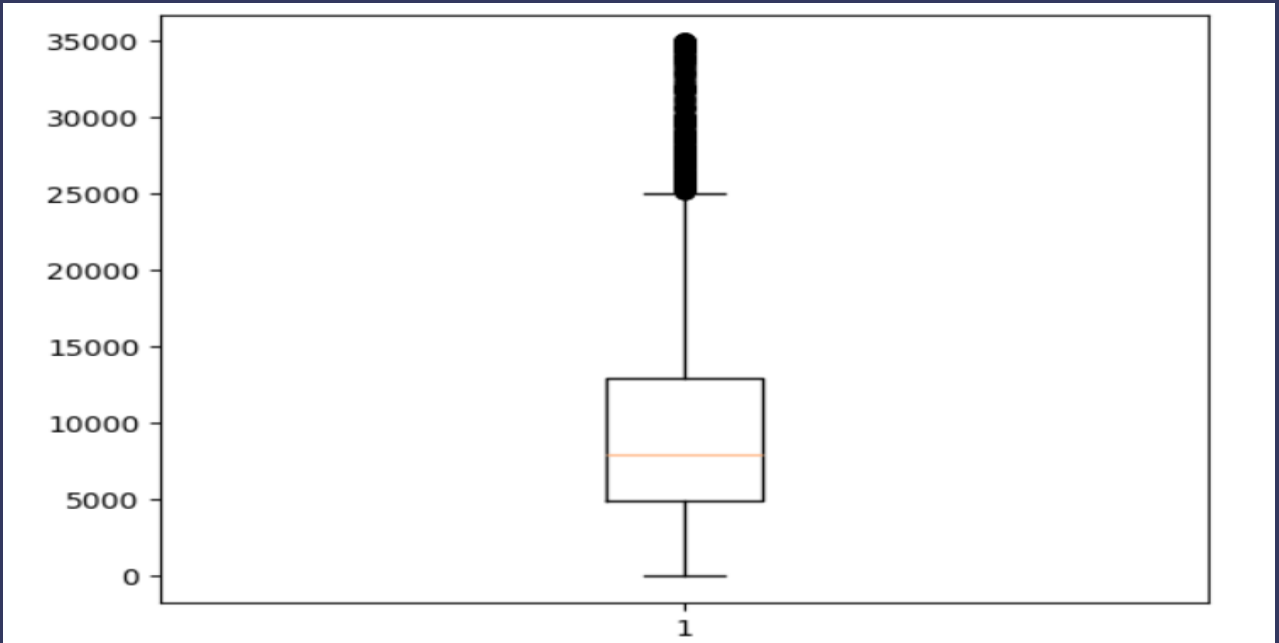
Dropped ‘annual_inc’ for quantile more than 90%.

CHECK LOAN_AMNT



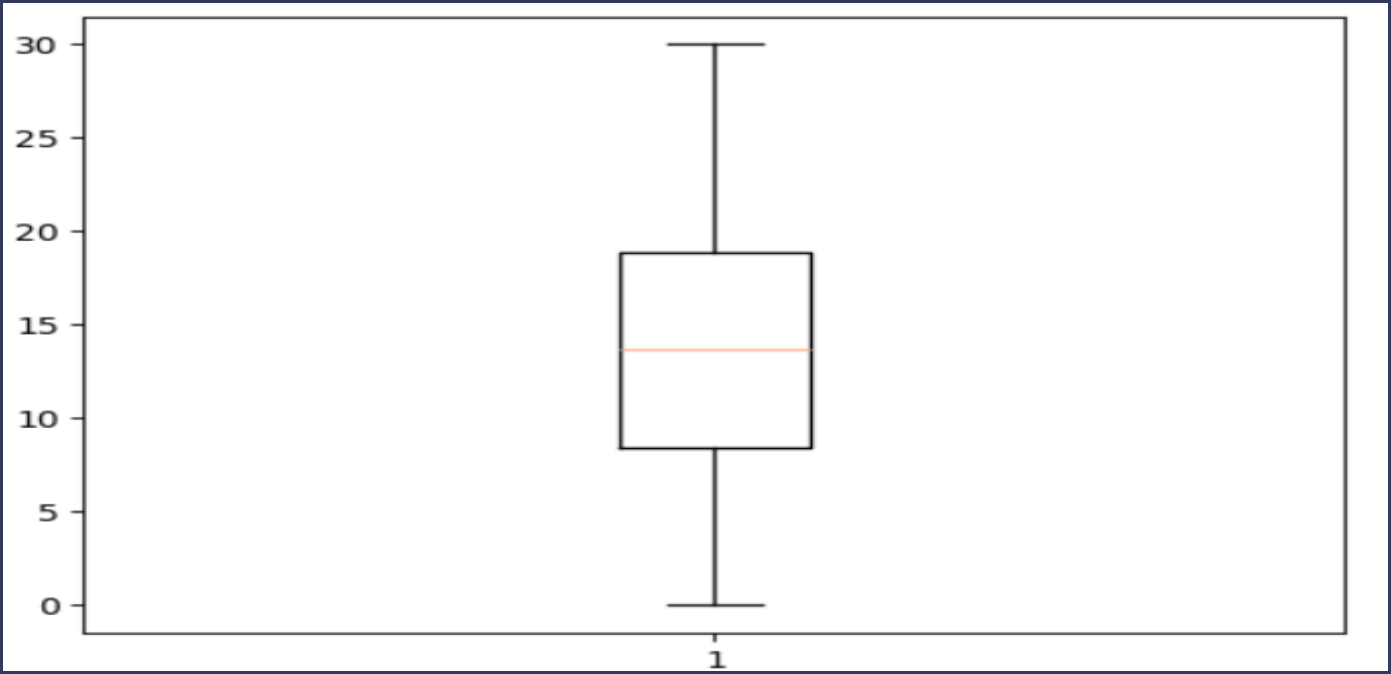
Doesn’t seem to require any data manipulation

CHECK FUNDED_AMNT_INV



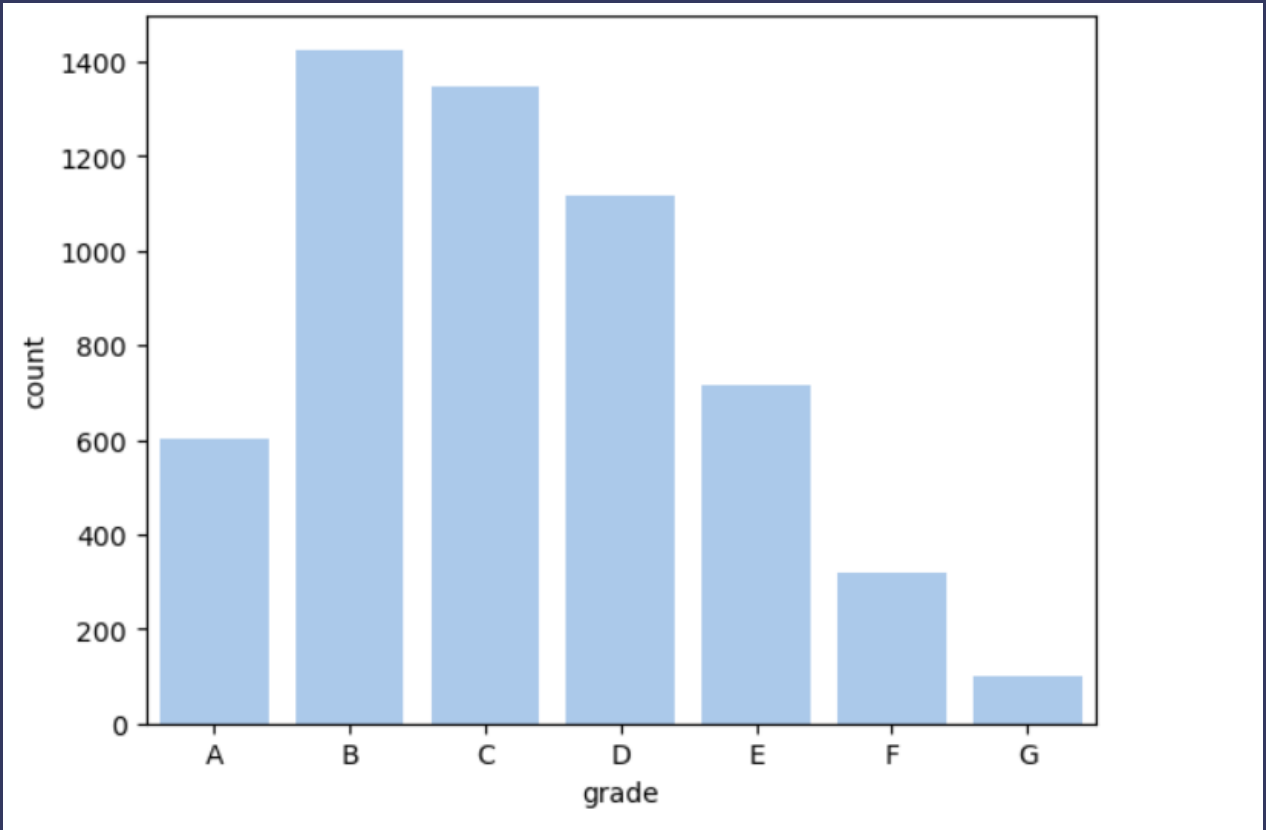
Doesn’t seem to require any data manipulation

CHECK DTI

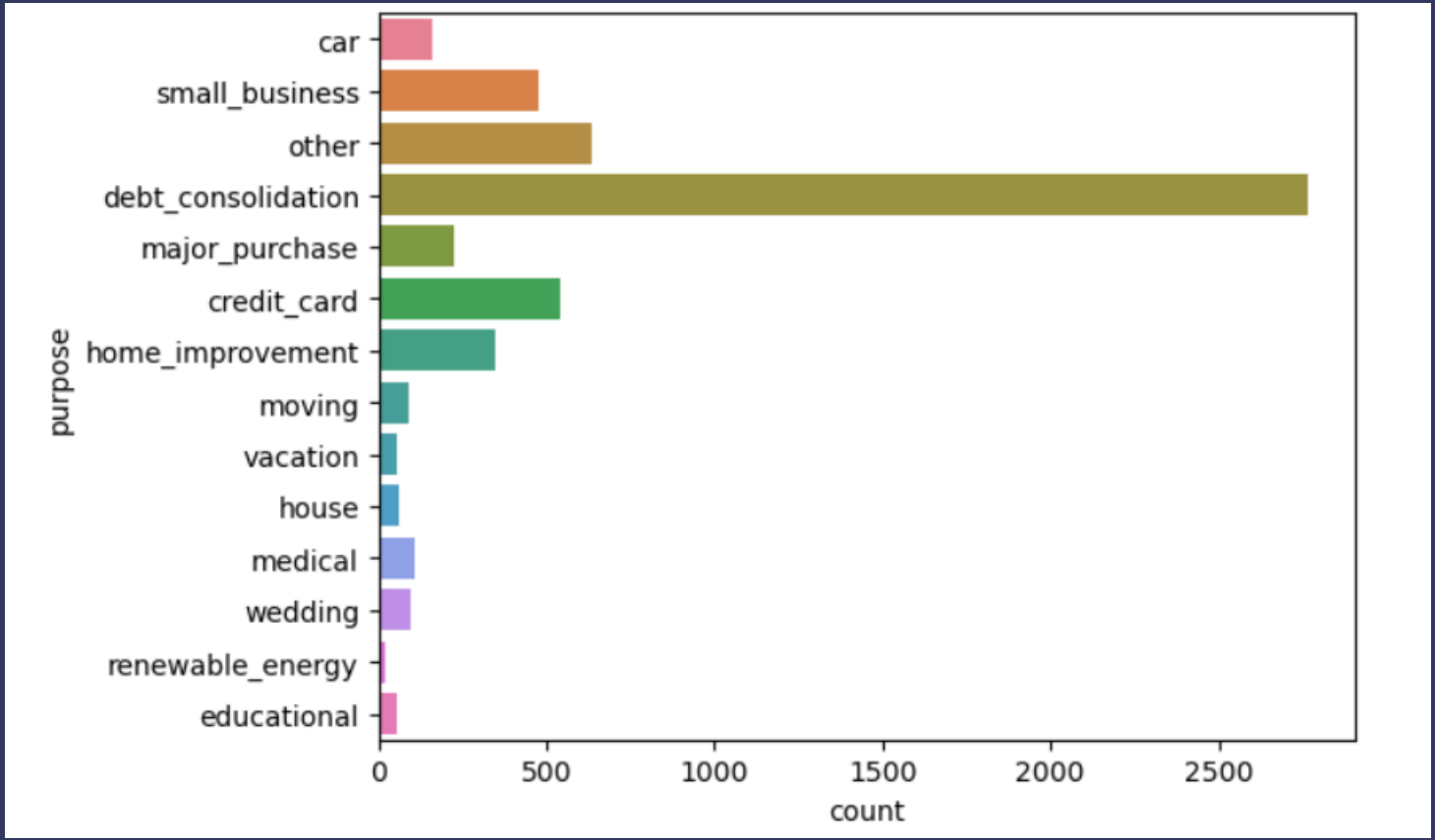


Doesn’t seem to require any data manipulation

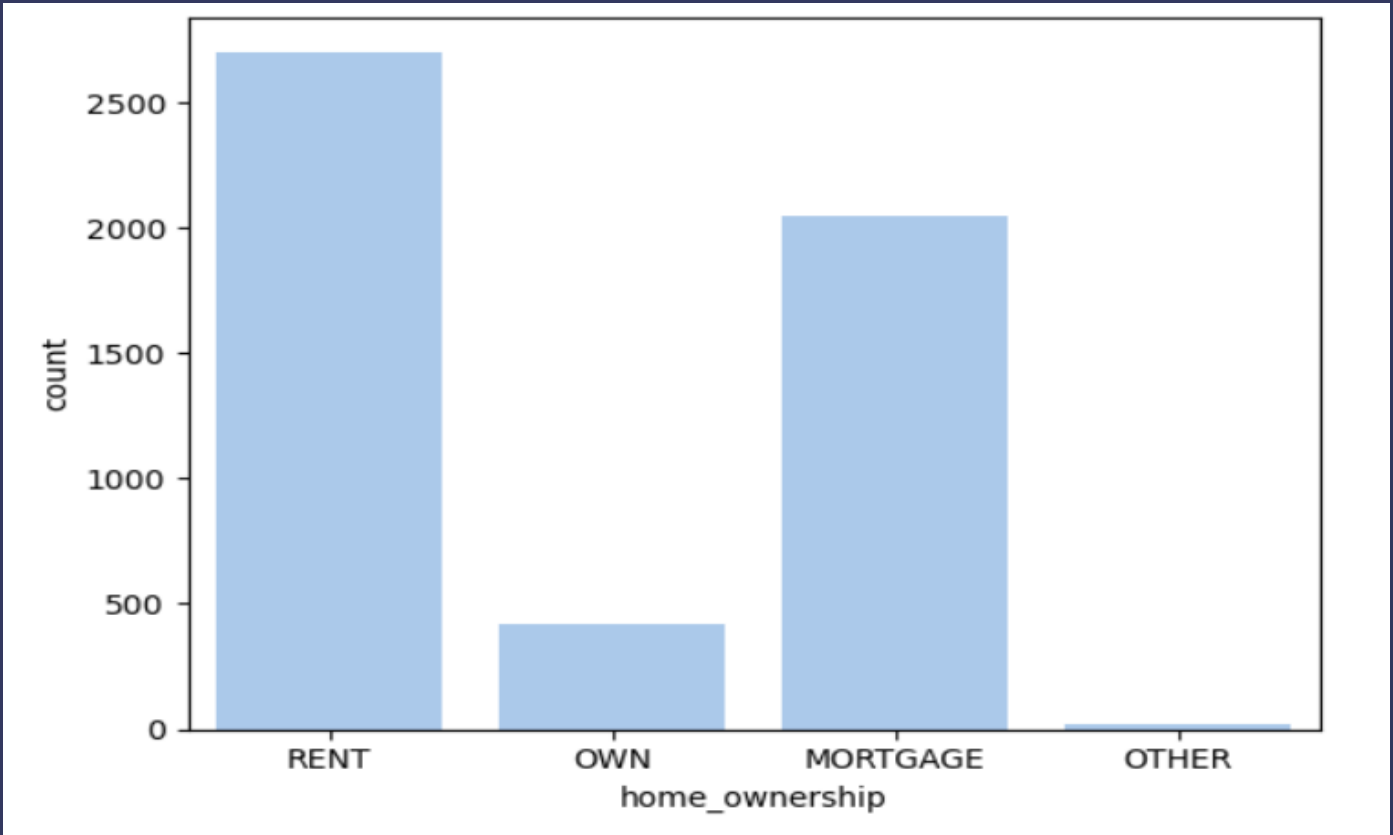
SUB-GRADING FOR LOAN STATUS CHARGED OFF



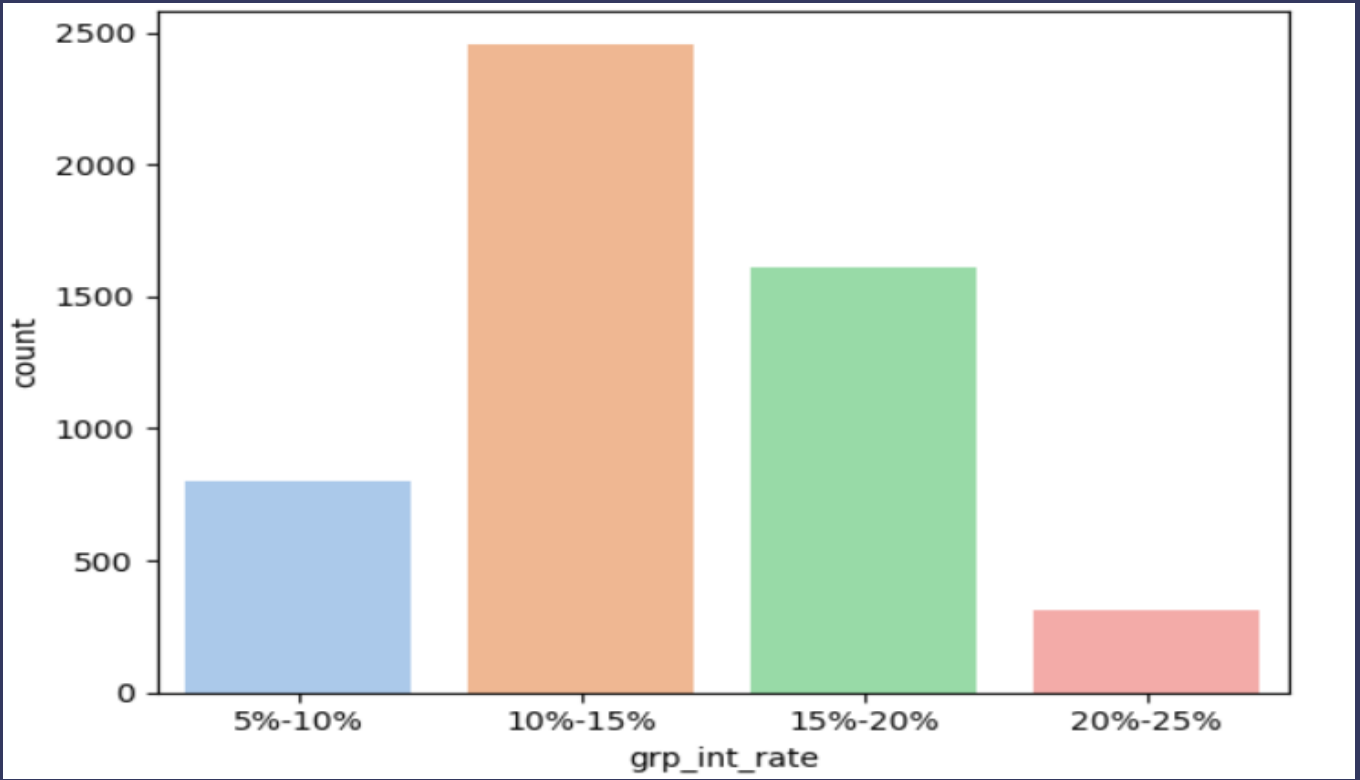
VISUALIZING PURPOSE FOR STATUS CHARGED OFF



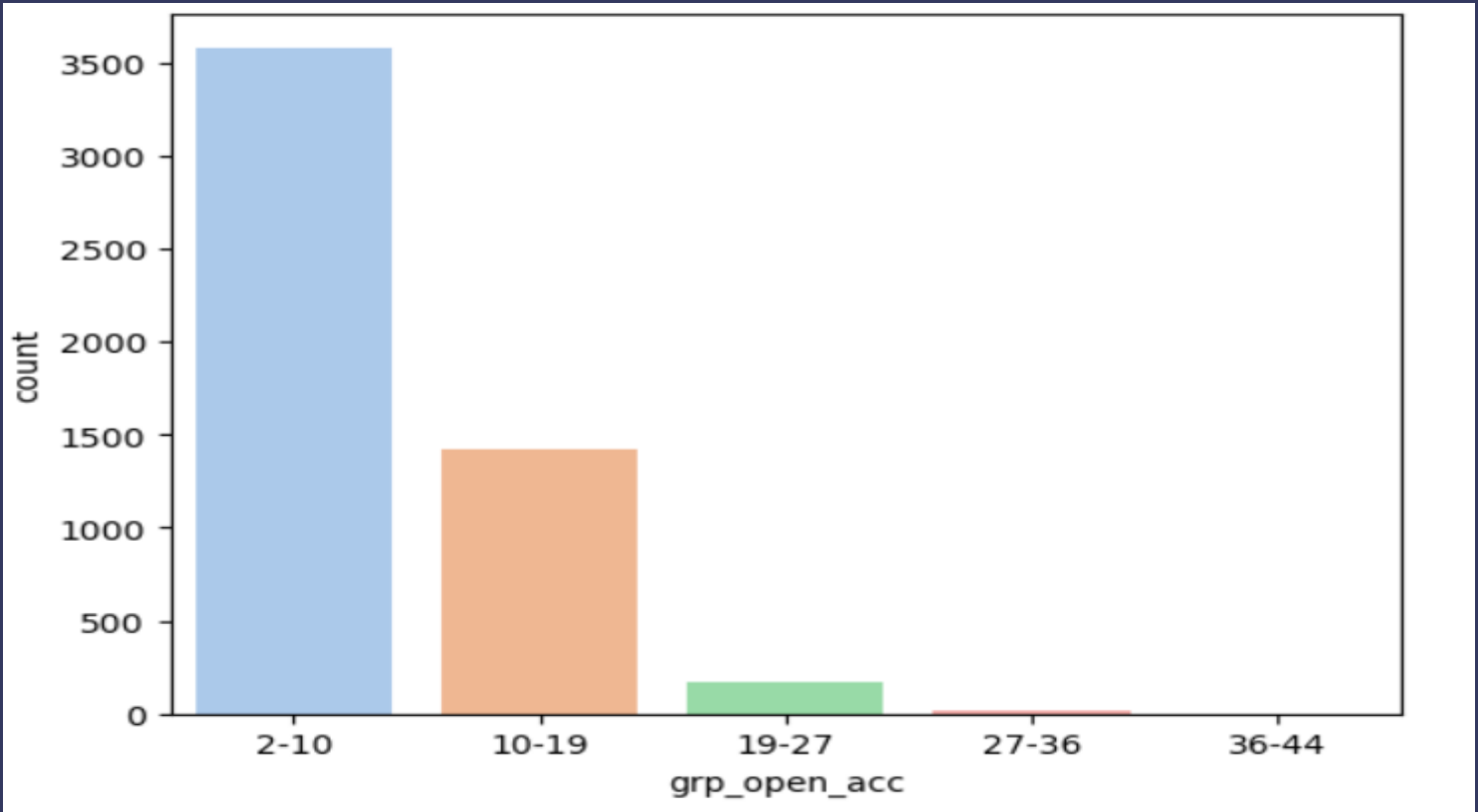
VISUALIZING HOME_OWNERSHIP FOR STATUS CHARGED OFF



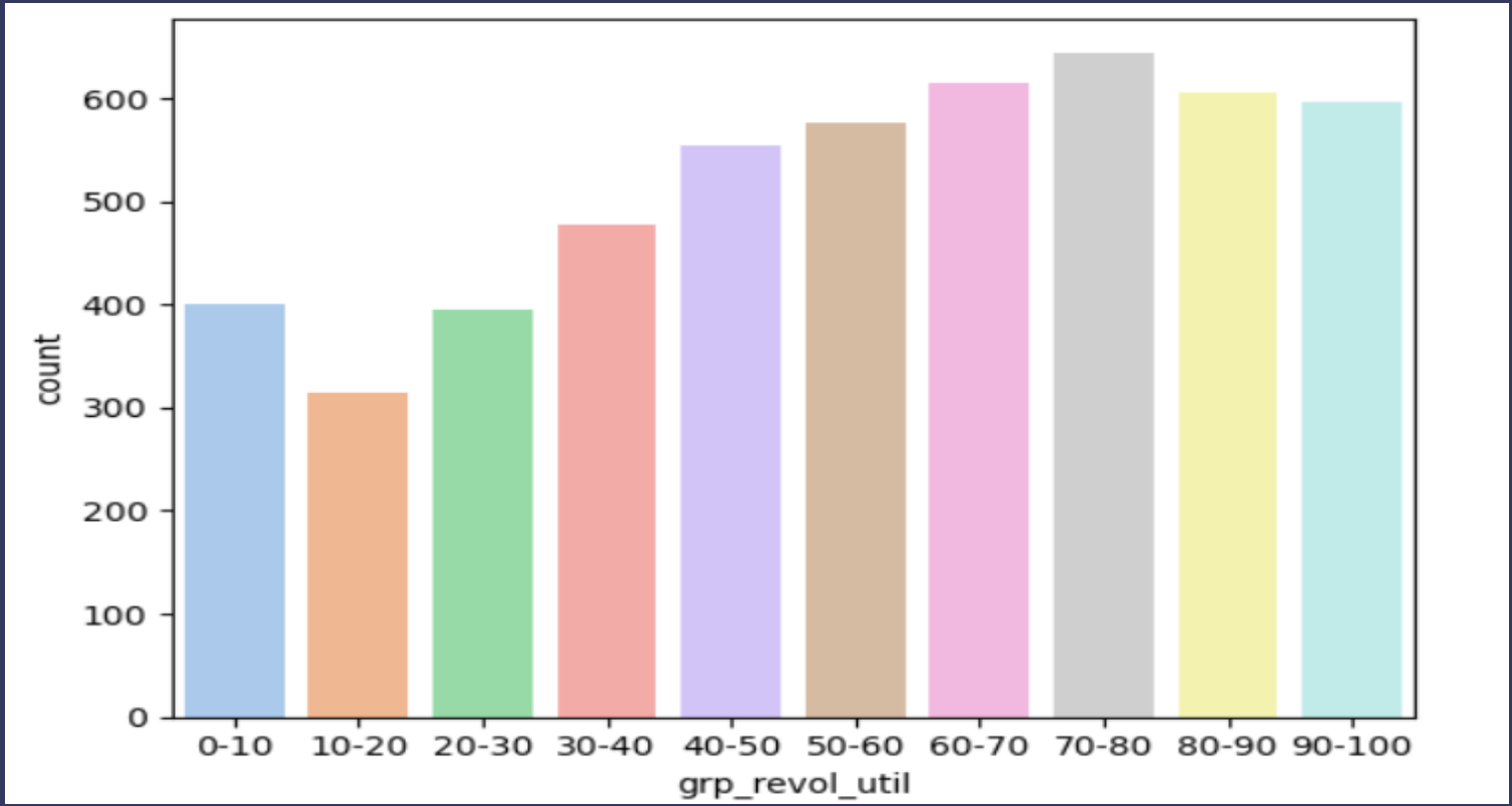
VISUALIZING INTEREST RATES FOR STATUS CHARGED OFF



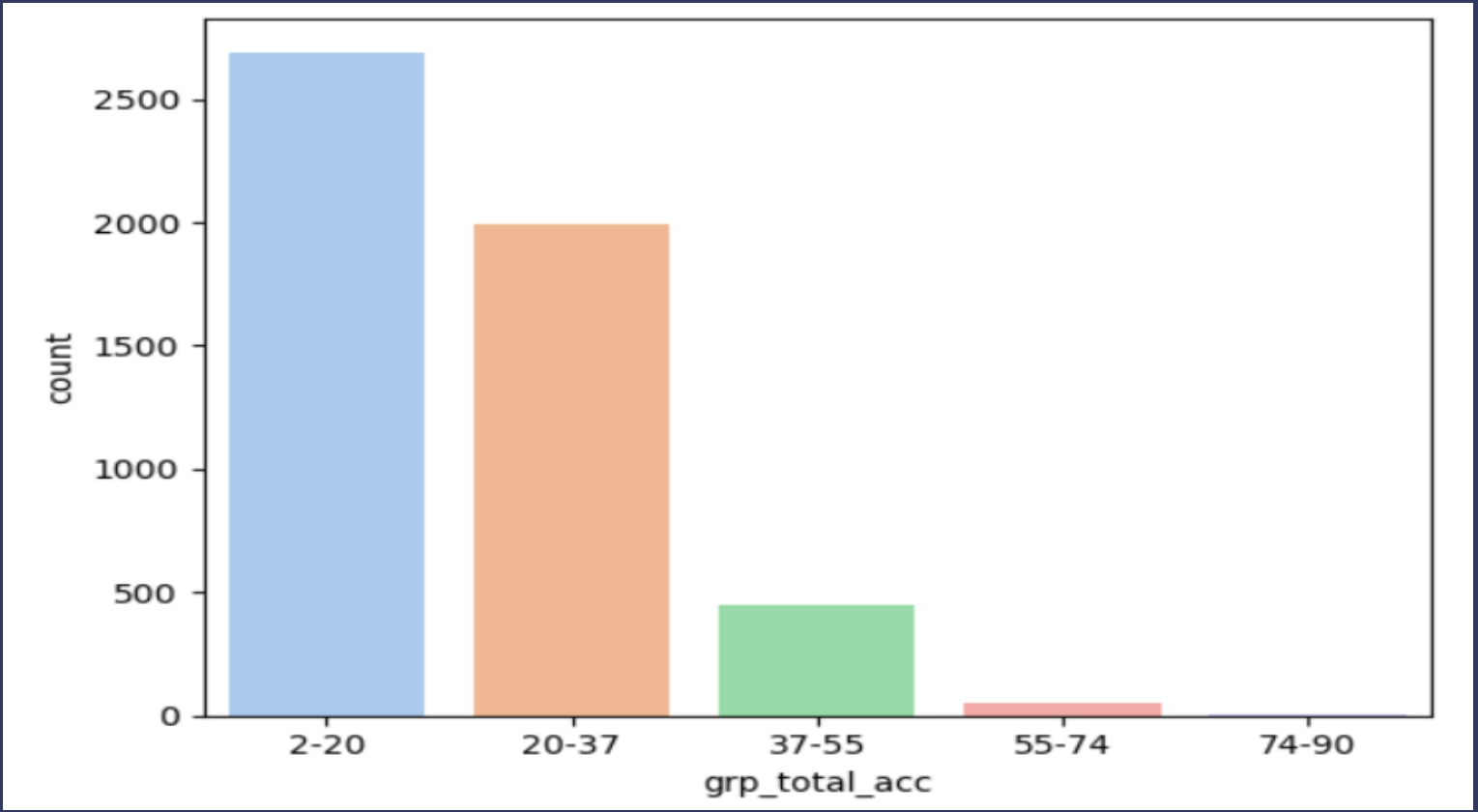
VISUALIZING OPEN ACCOUNTS FOR STATUS CHARGED OFF



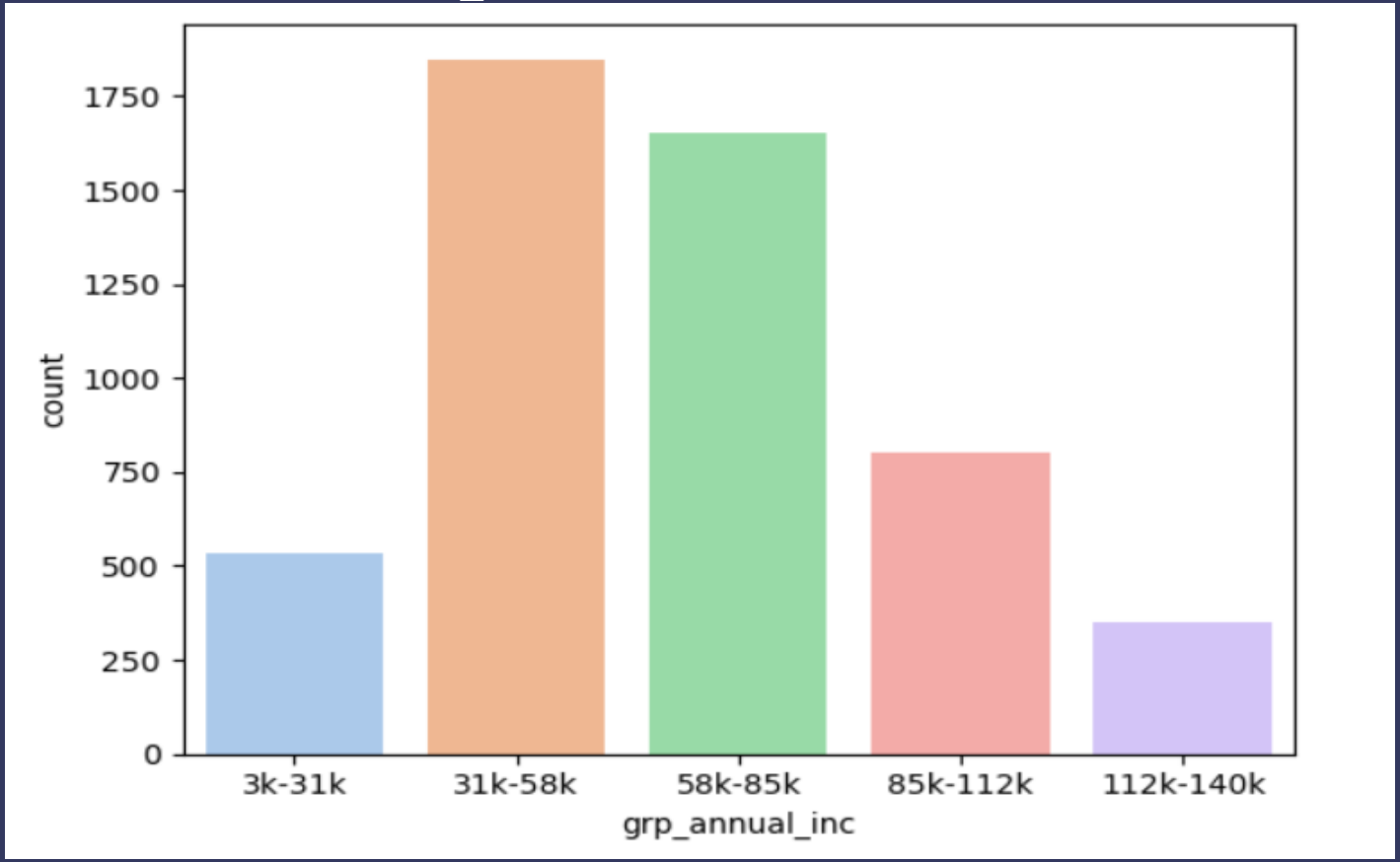
VISUALIZING REVOL_UTIL FOR STATUS CHARGED OFF



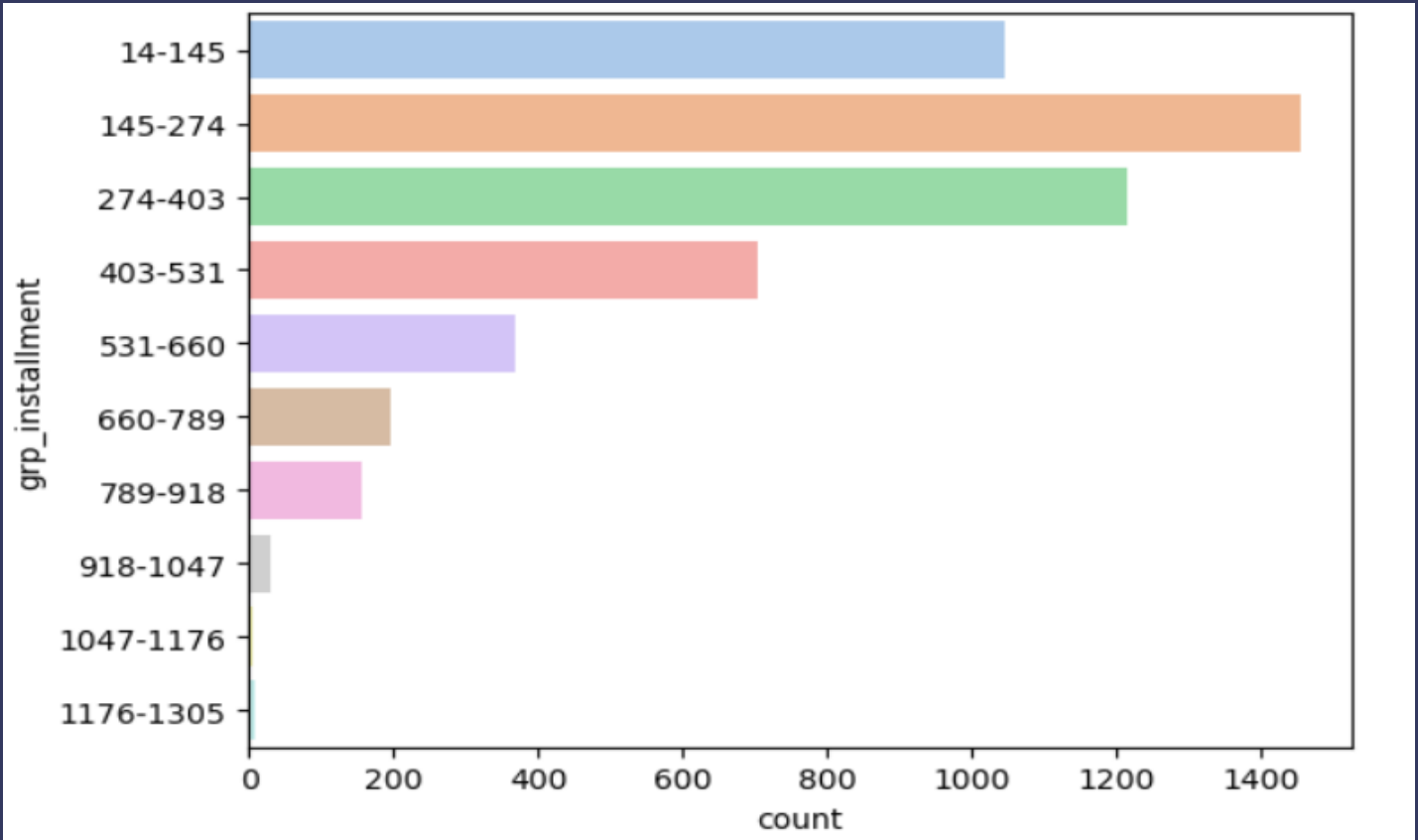
VISUALIZING TOTAL_ACC FOR STATUS CHARGED OFF



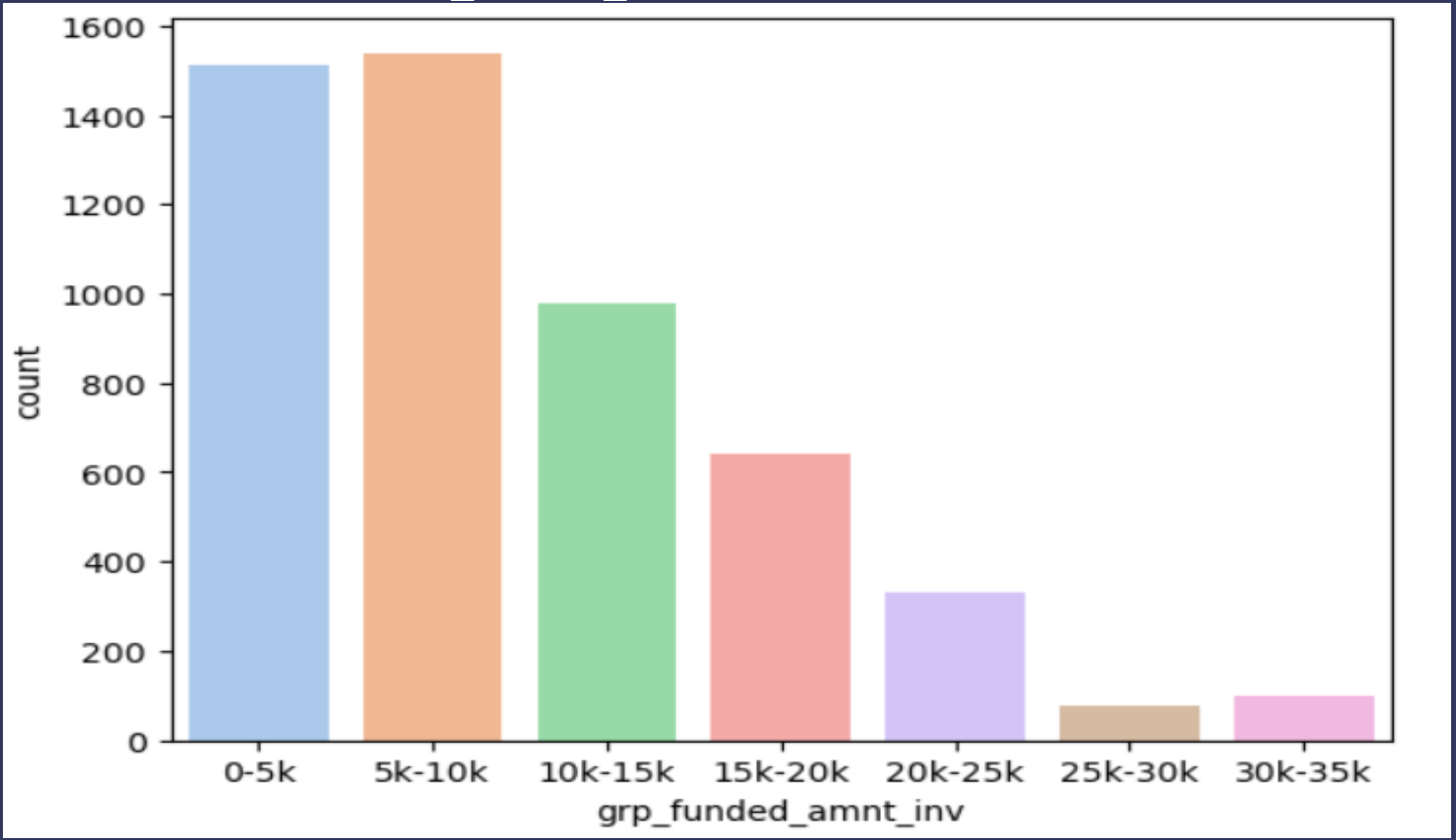
VISUALIZING ANNUAL_INC FOR STATUS CHARGED OFF



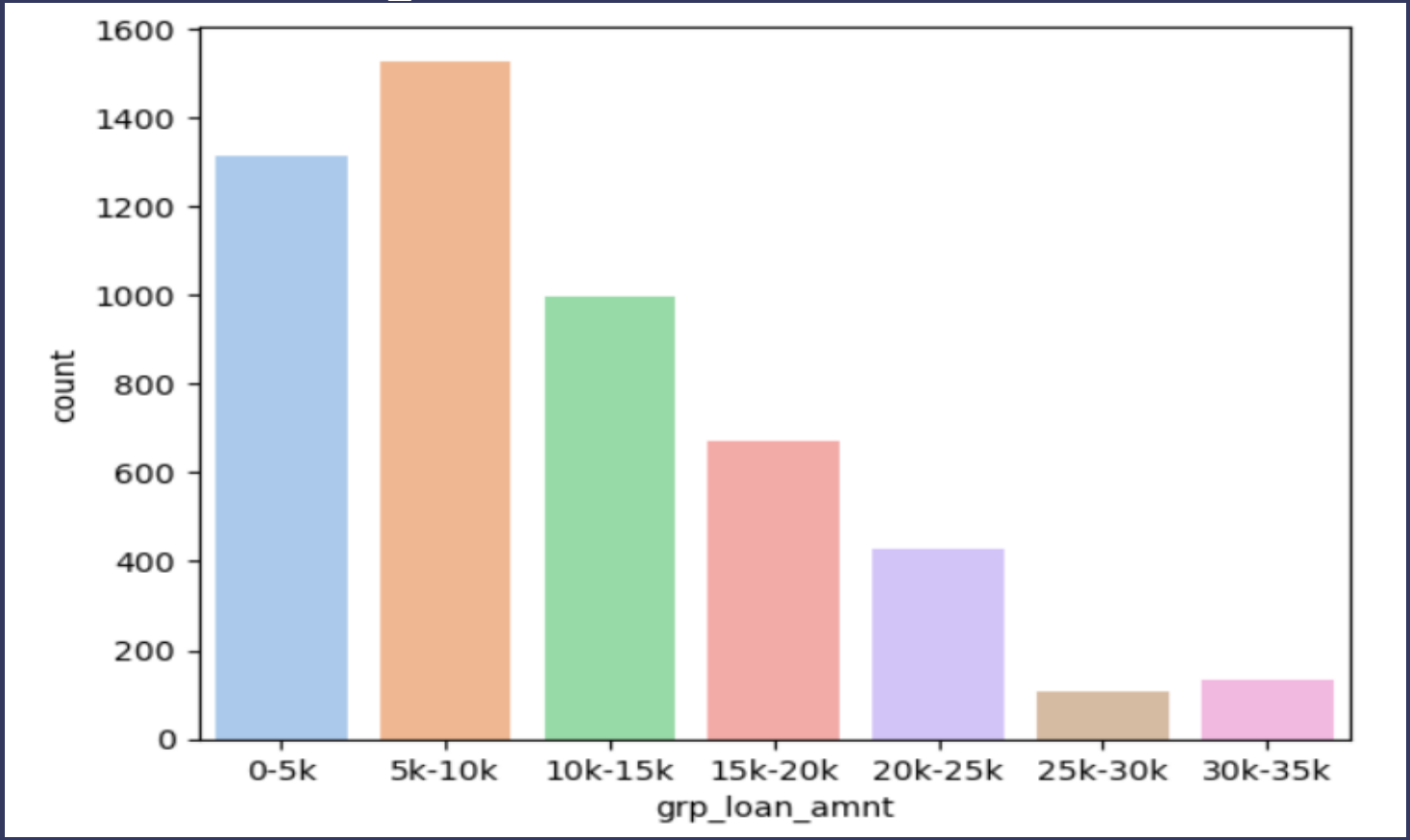
VISUALIZING INSTALLMENT FOR STATUS CHARGED OFF



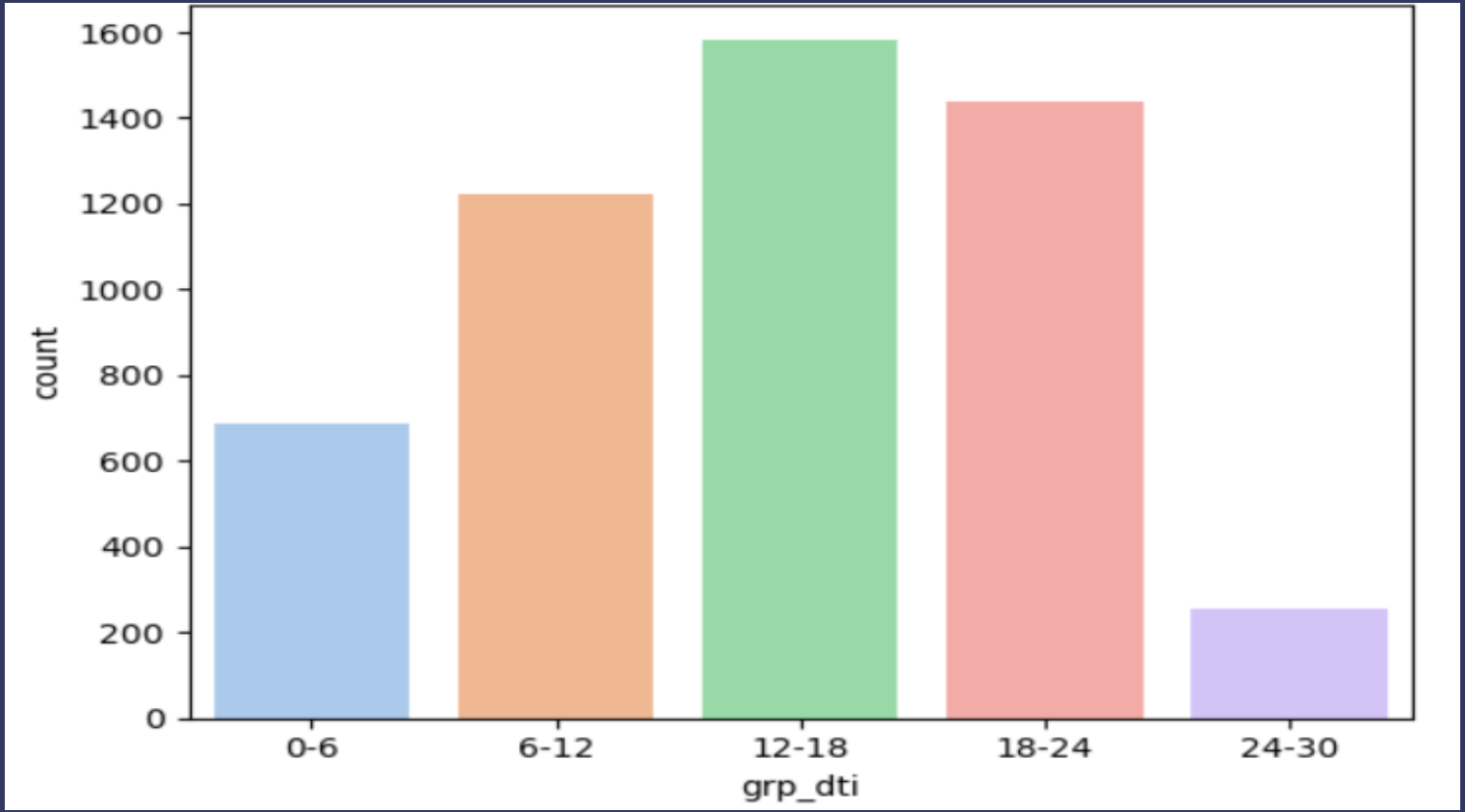
VISUALIZING FUNDED_AMNT_INV FOR STATUS CHARGED OFF



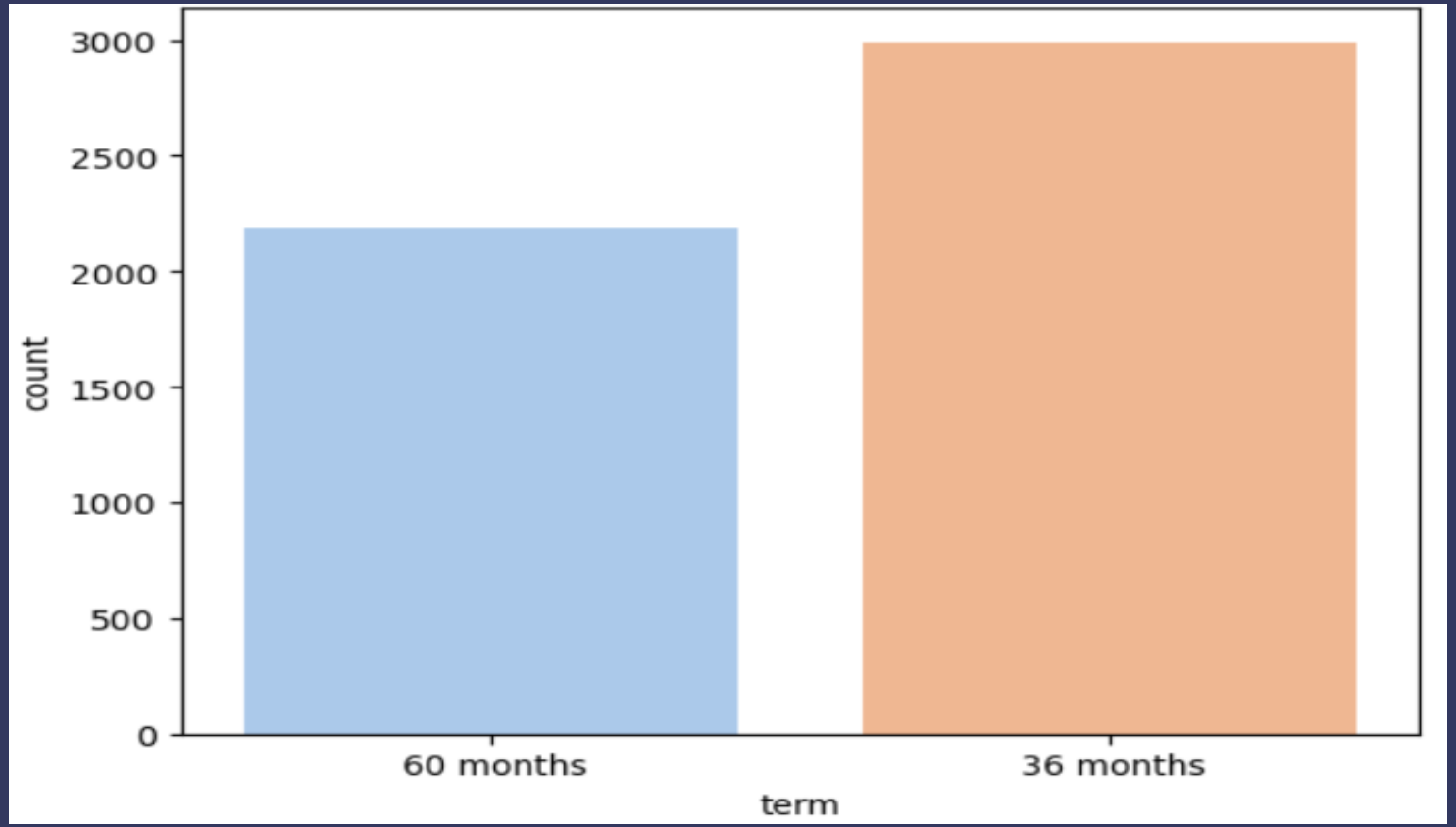
VISUALIZING LOAN_AMNT FOR STATUS CHARGED OFF



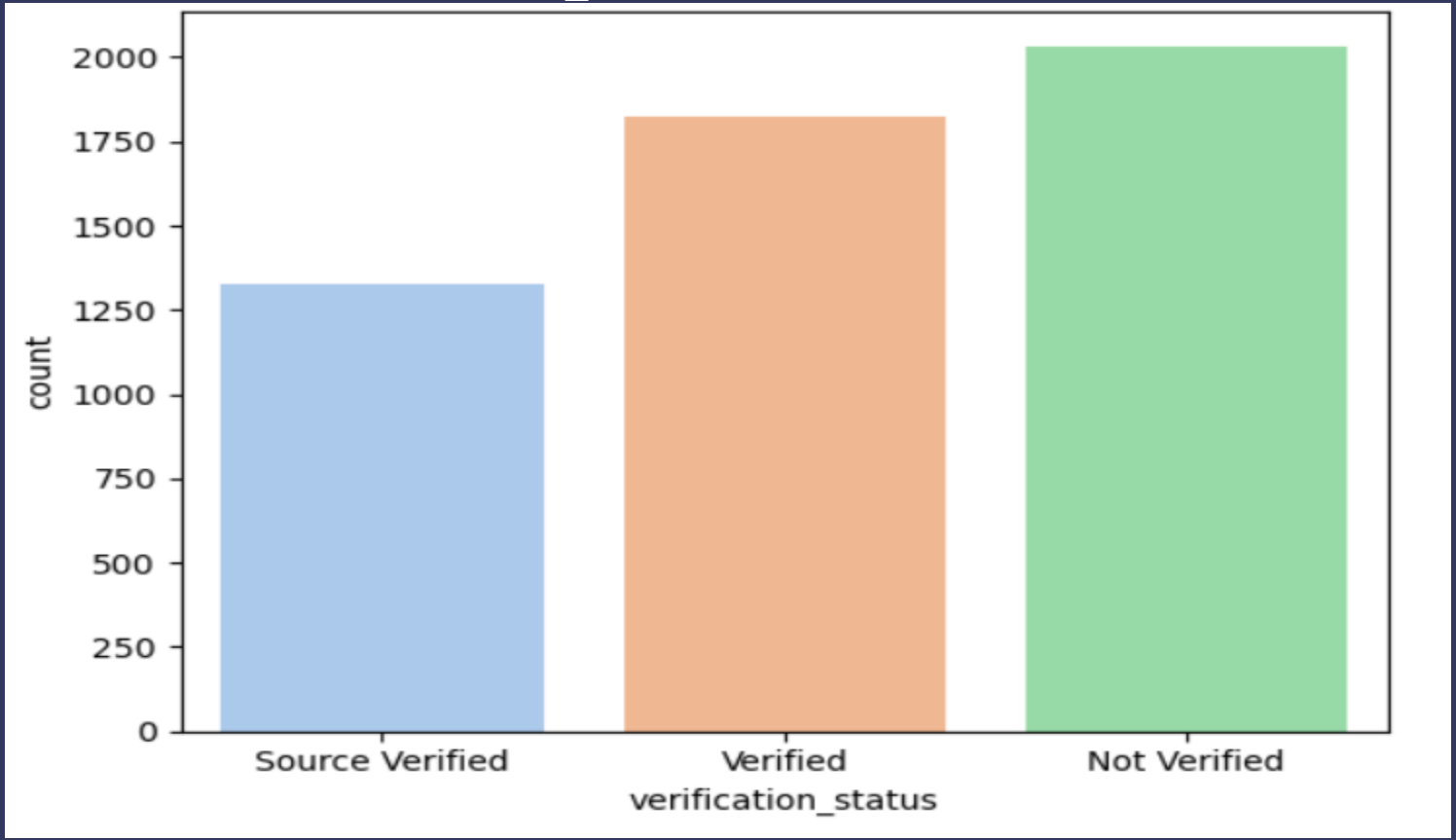
VISUALIZING DTI FOR STATUS CHARGED OFF



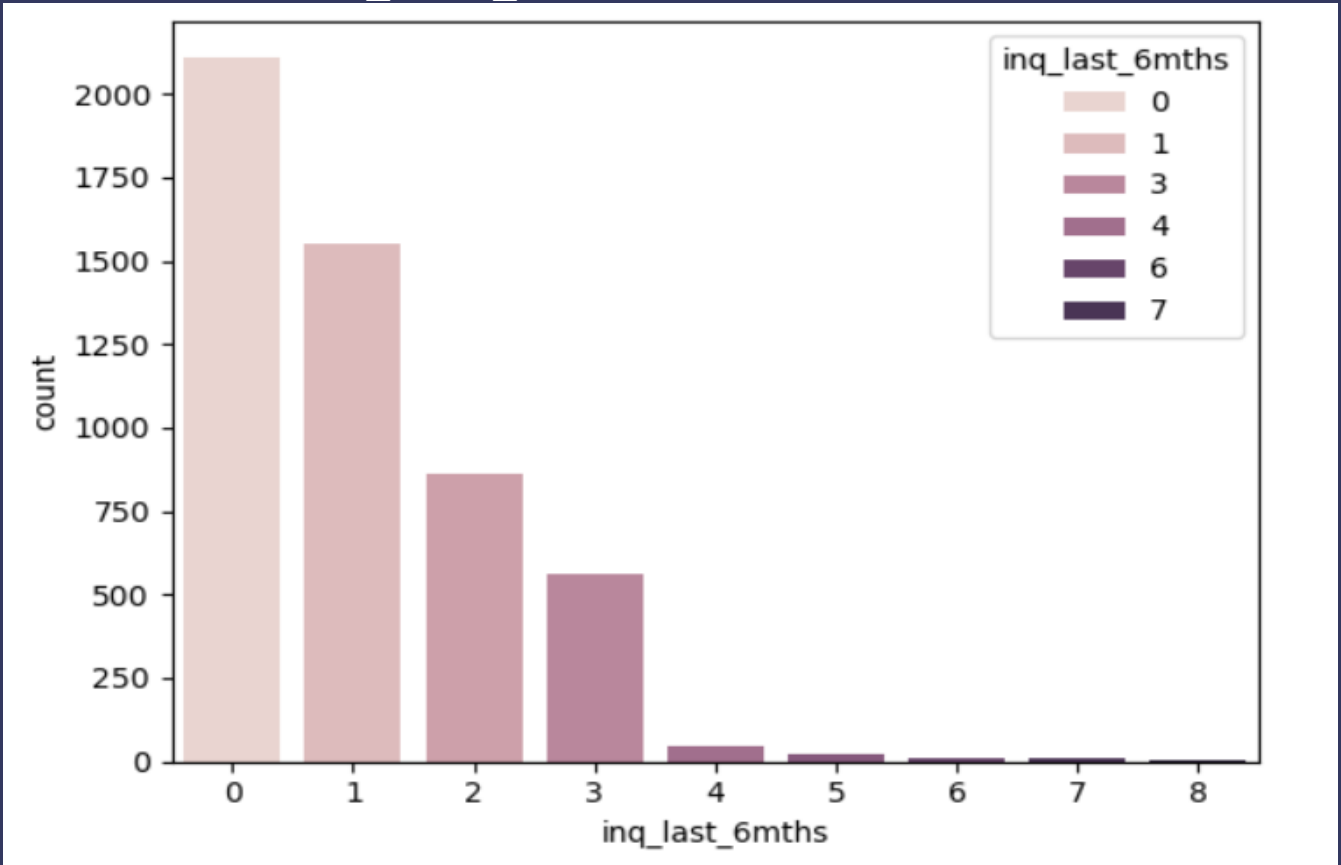
VISUALIZING TERM FOR STATUS CHARGED OFF



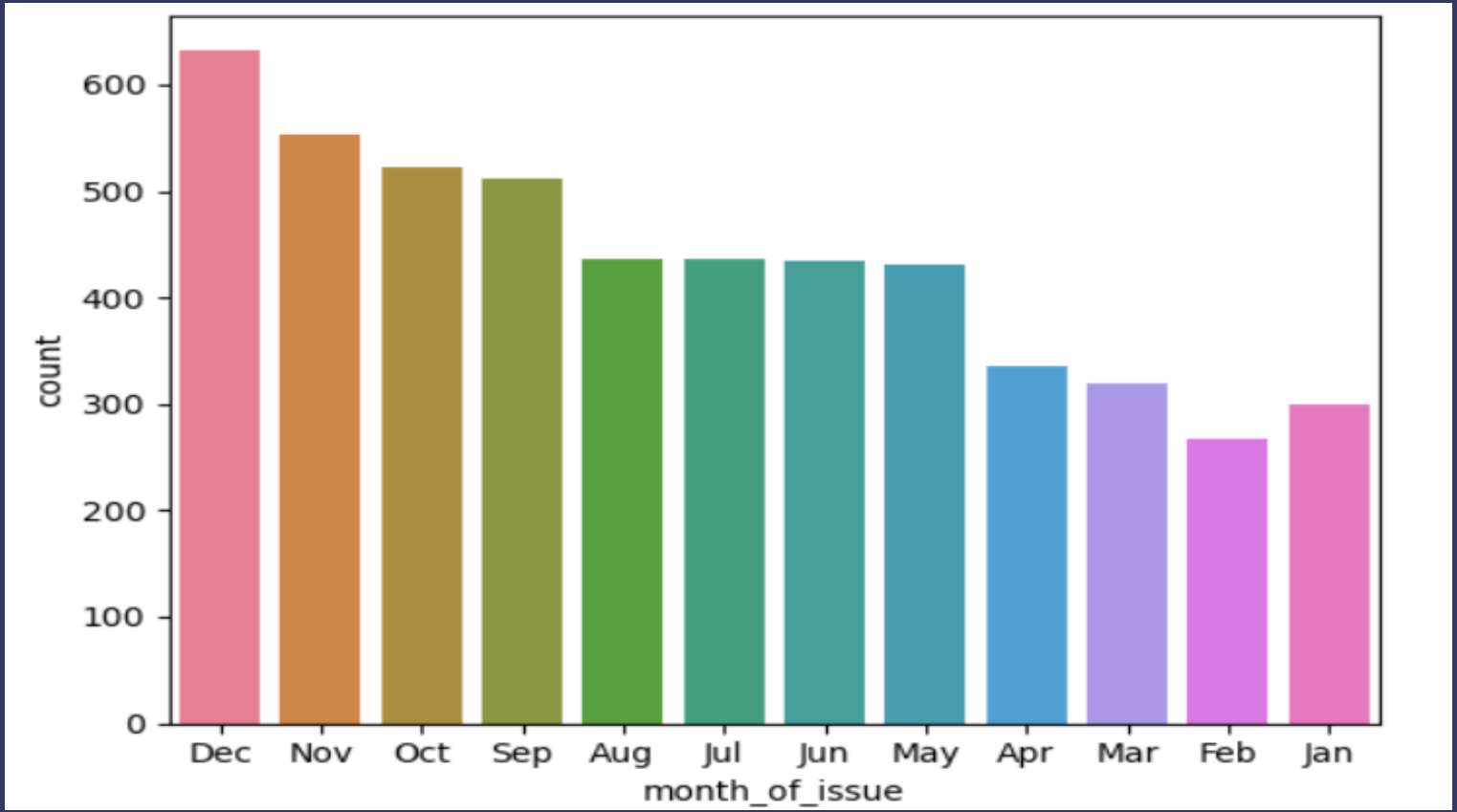
VISUALIZING VERIFICATION_STATUS FOR STATUS CHARGED OFF



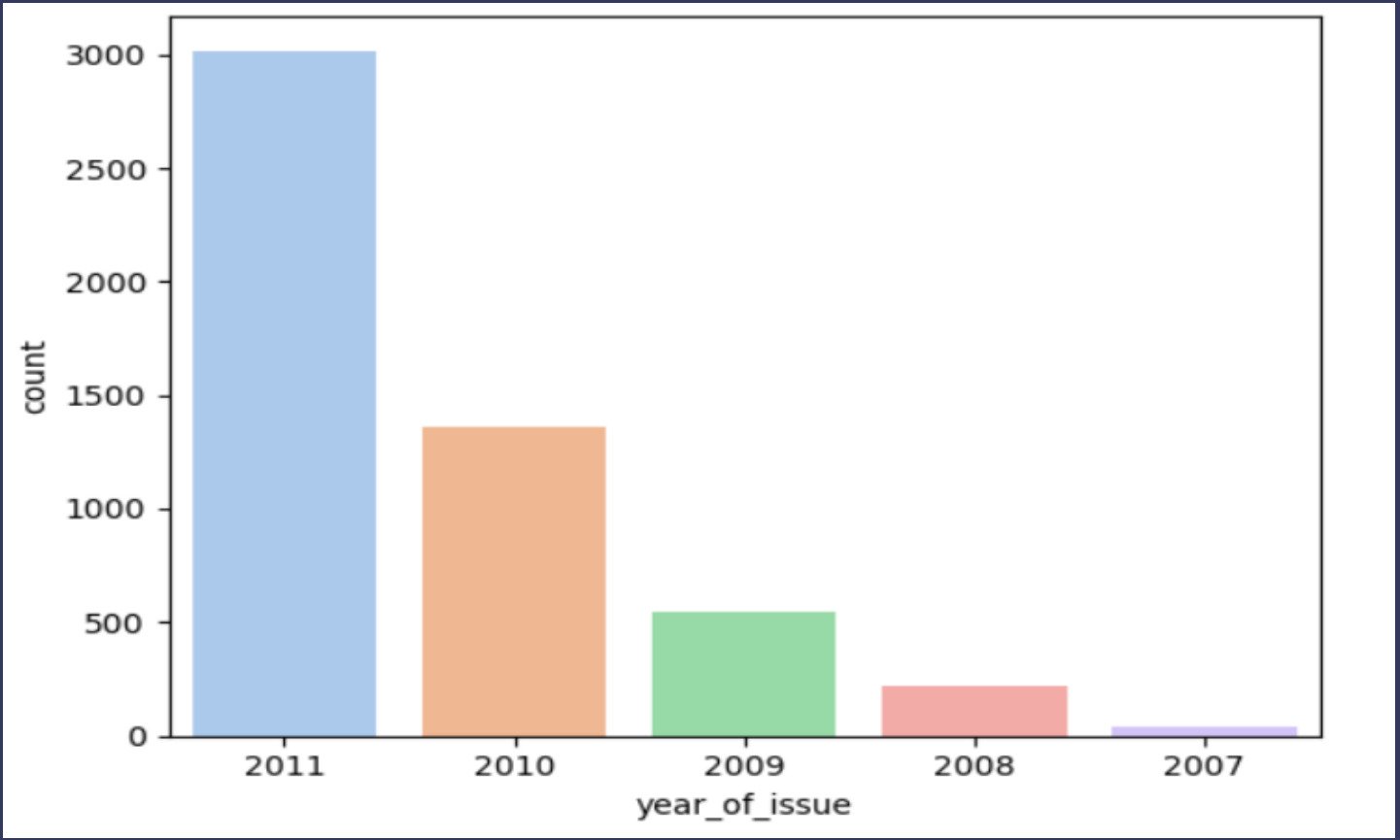
VISUALIZING INQ_LAST_6MTHS FOR STATUS CHARGED OFF



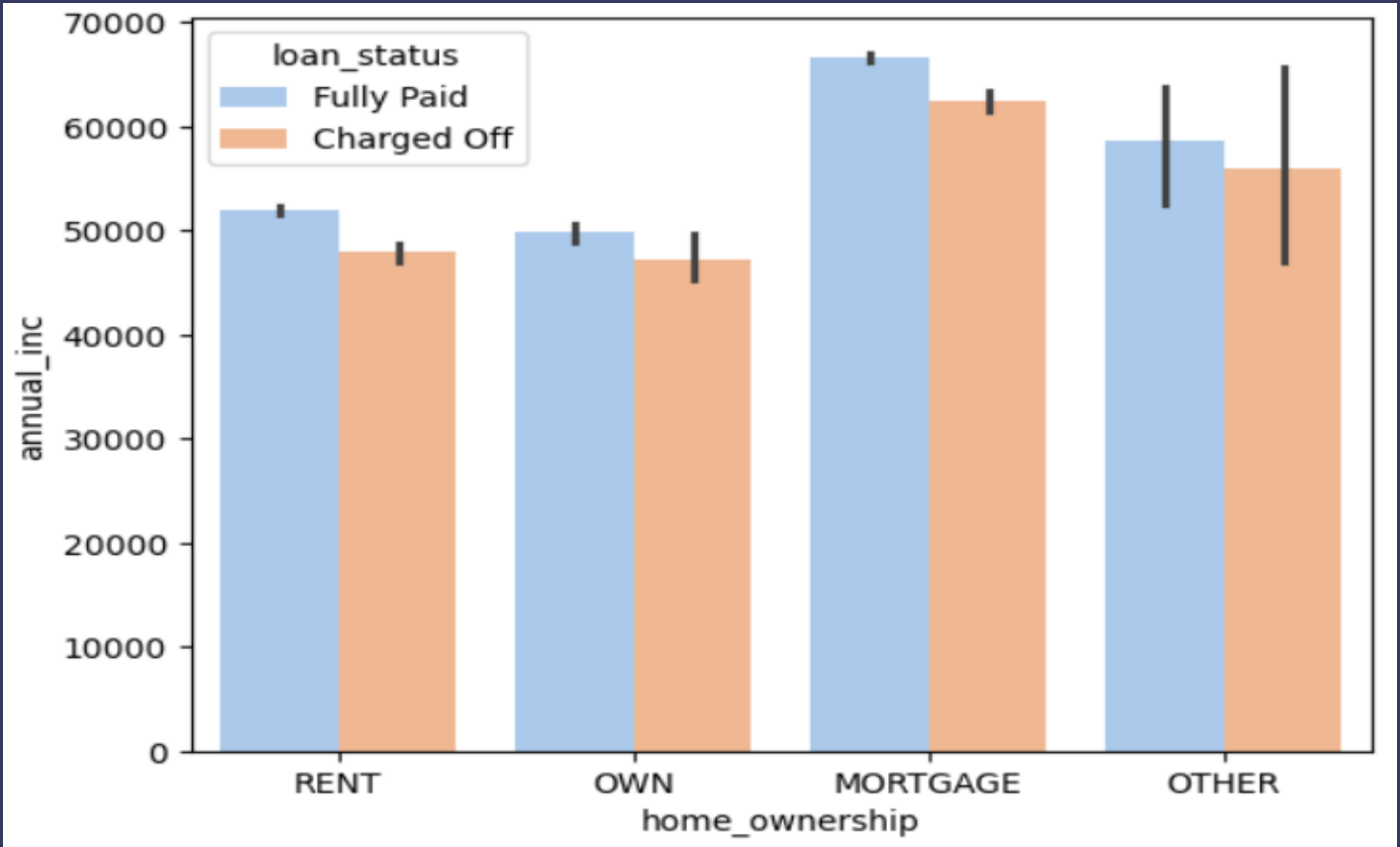
VISUALIZING MONTH OF ISSUE FOR STATUS CHARGED OFF



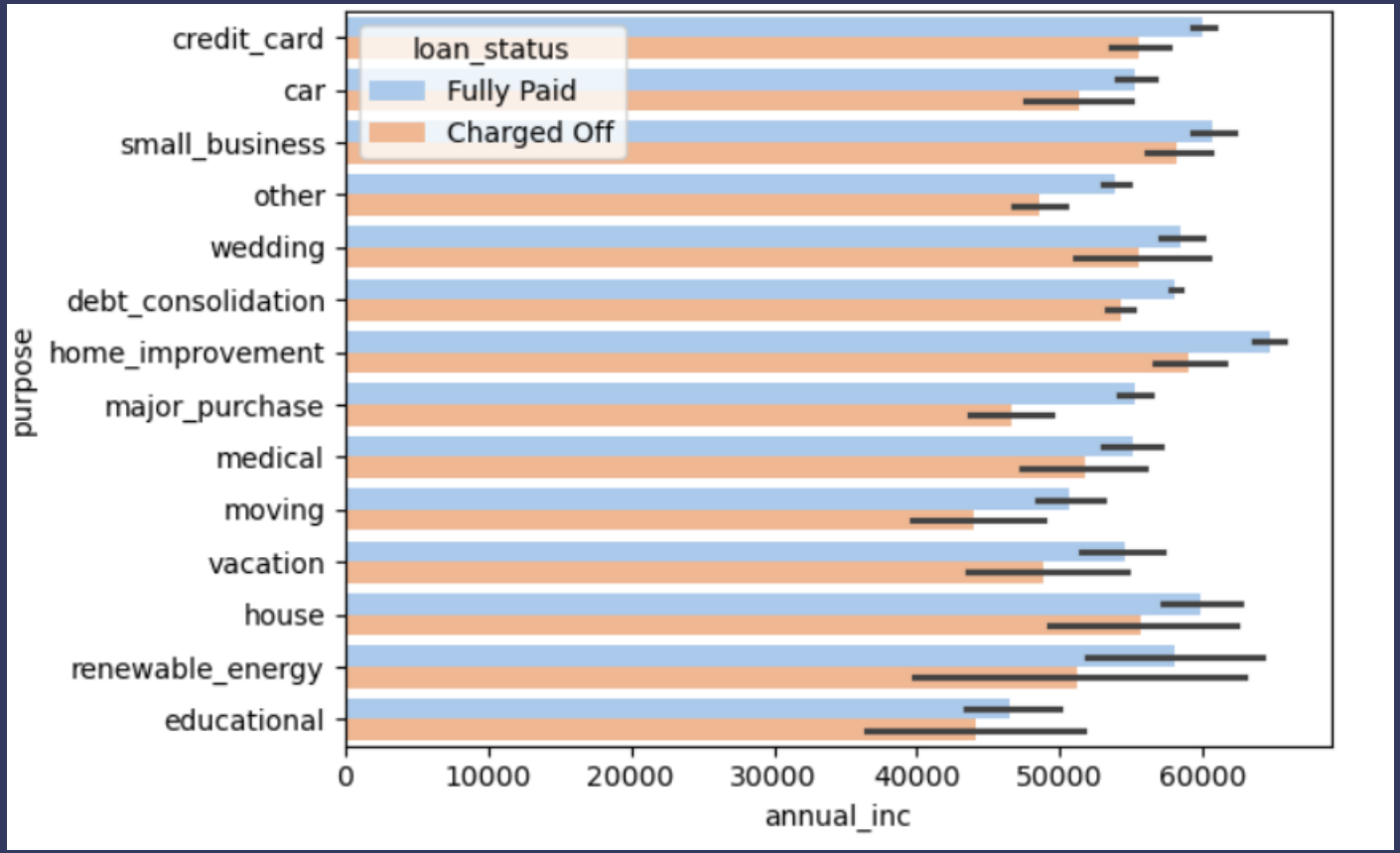
VISUALIZING YEAR OF ISSUE FOR STATUS CHARGED OFF



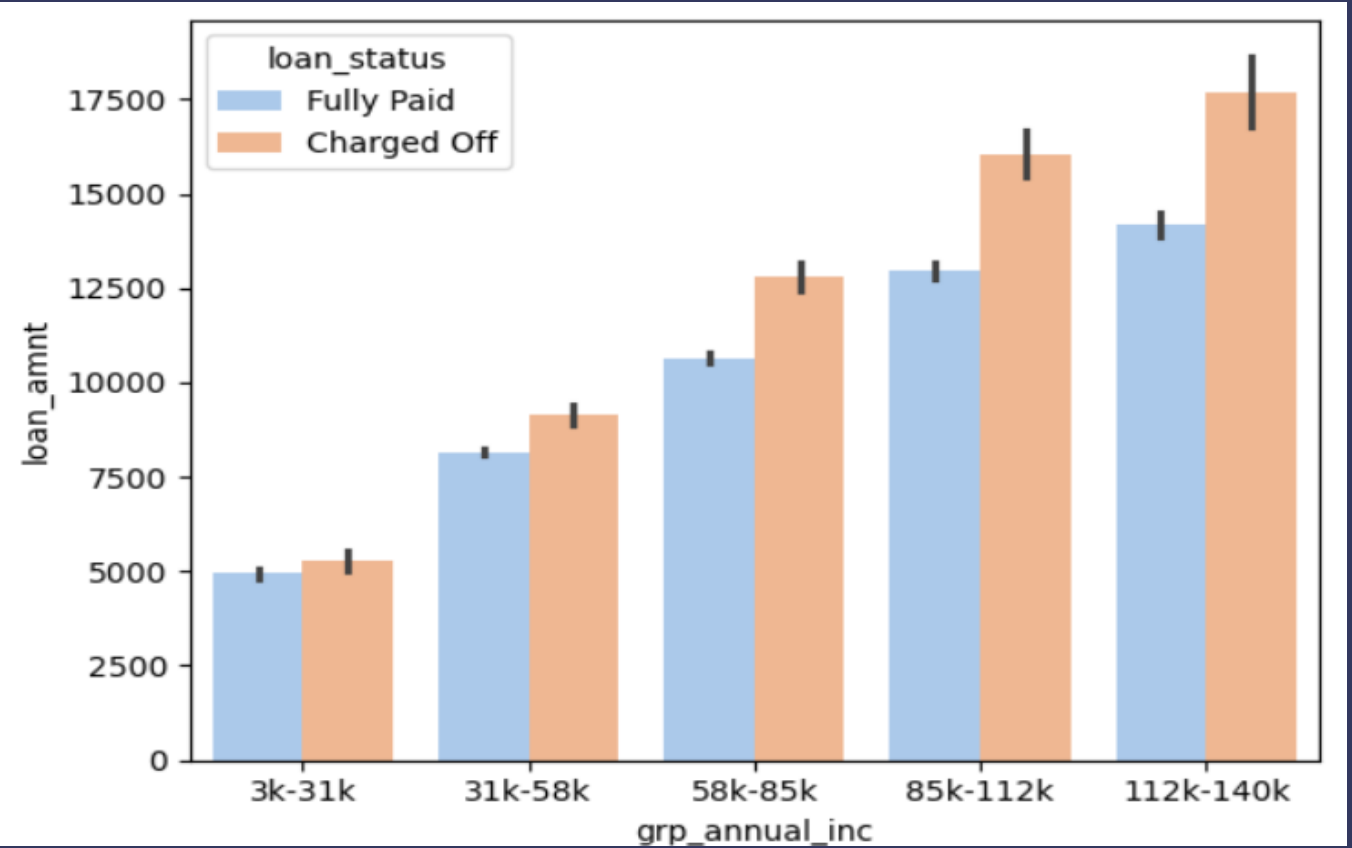
VISUALIZING ANNUAL INCOME VS HOME OWNERSHIP



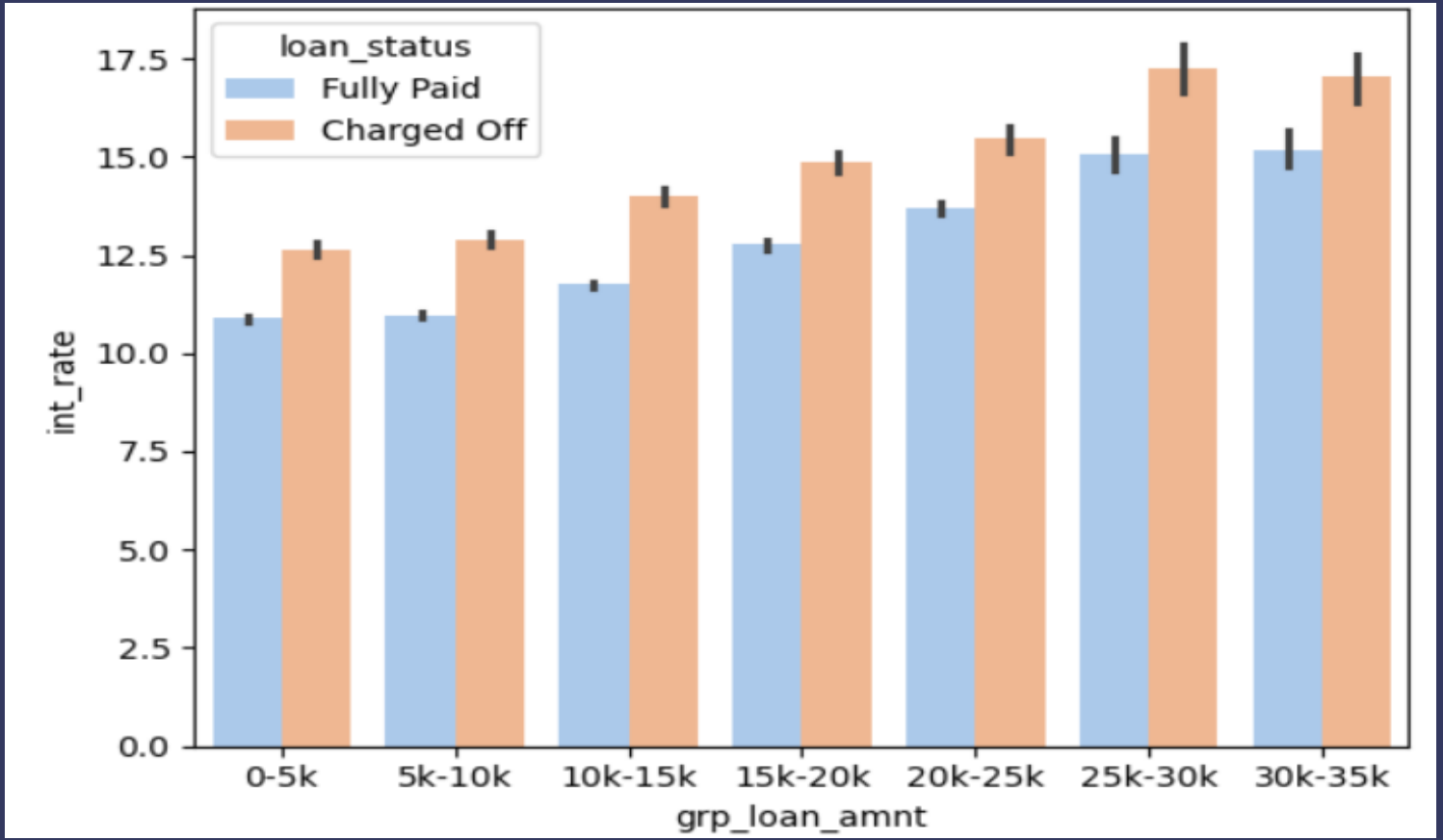
VISUALIZING ANNUAL INCOME VS LOAN PURPOSE



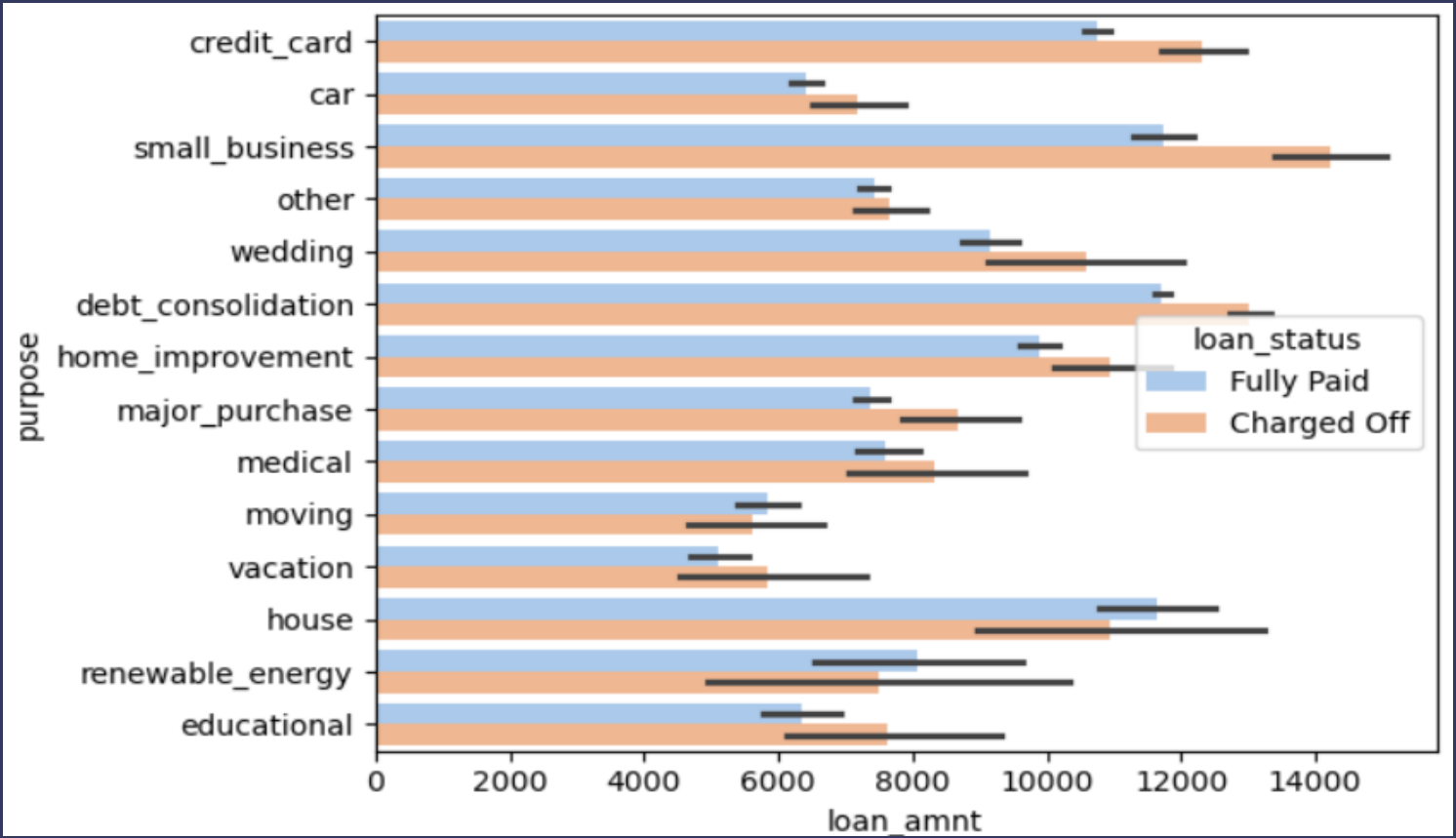
VISUALIZING ANNUAL INCOME VS LOAN AMOUNT



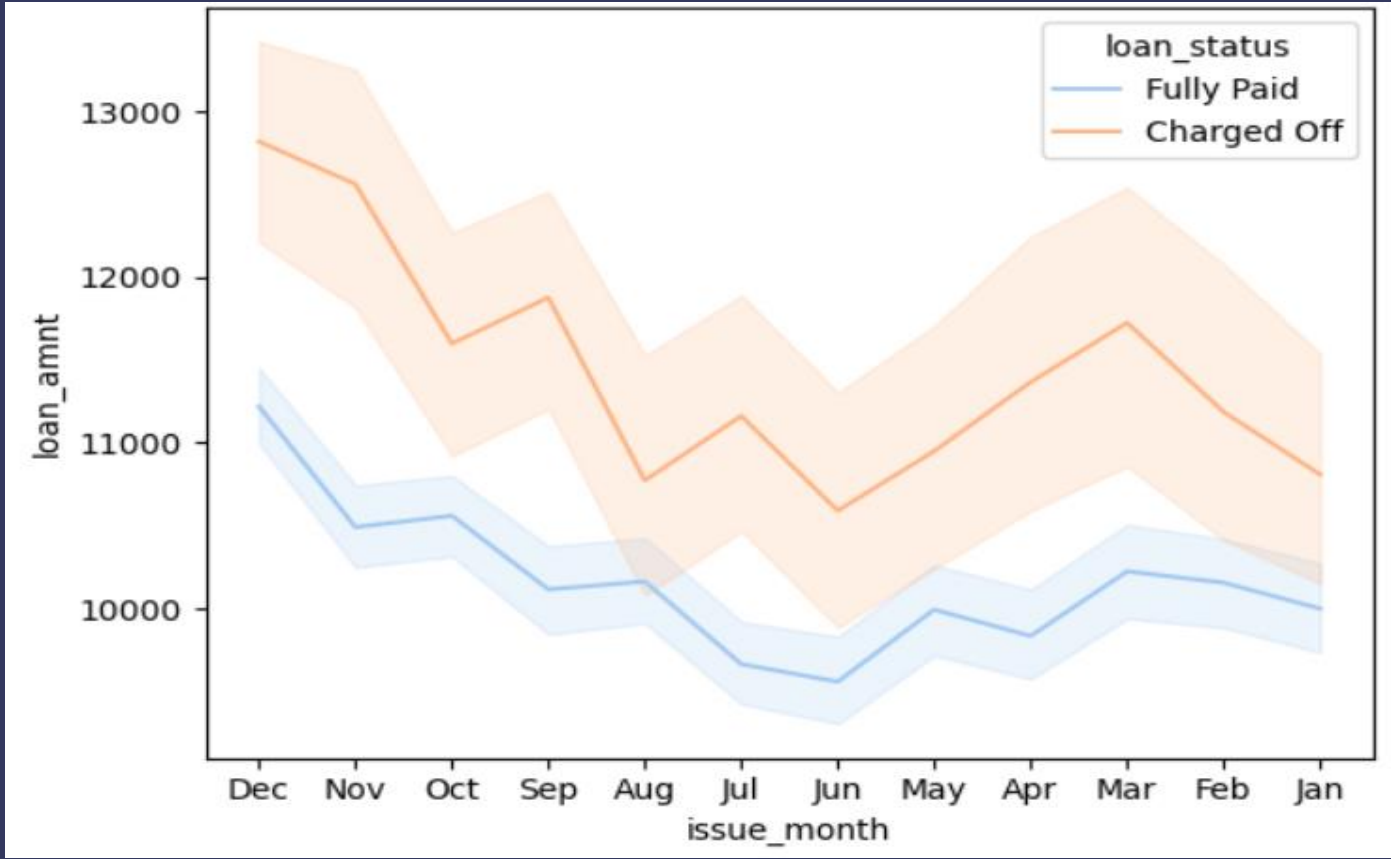
VISUALIZING LOAN AMOUNT VS INTEREST RATE



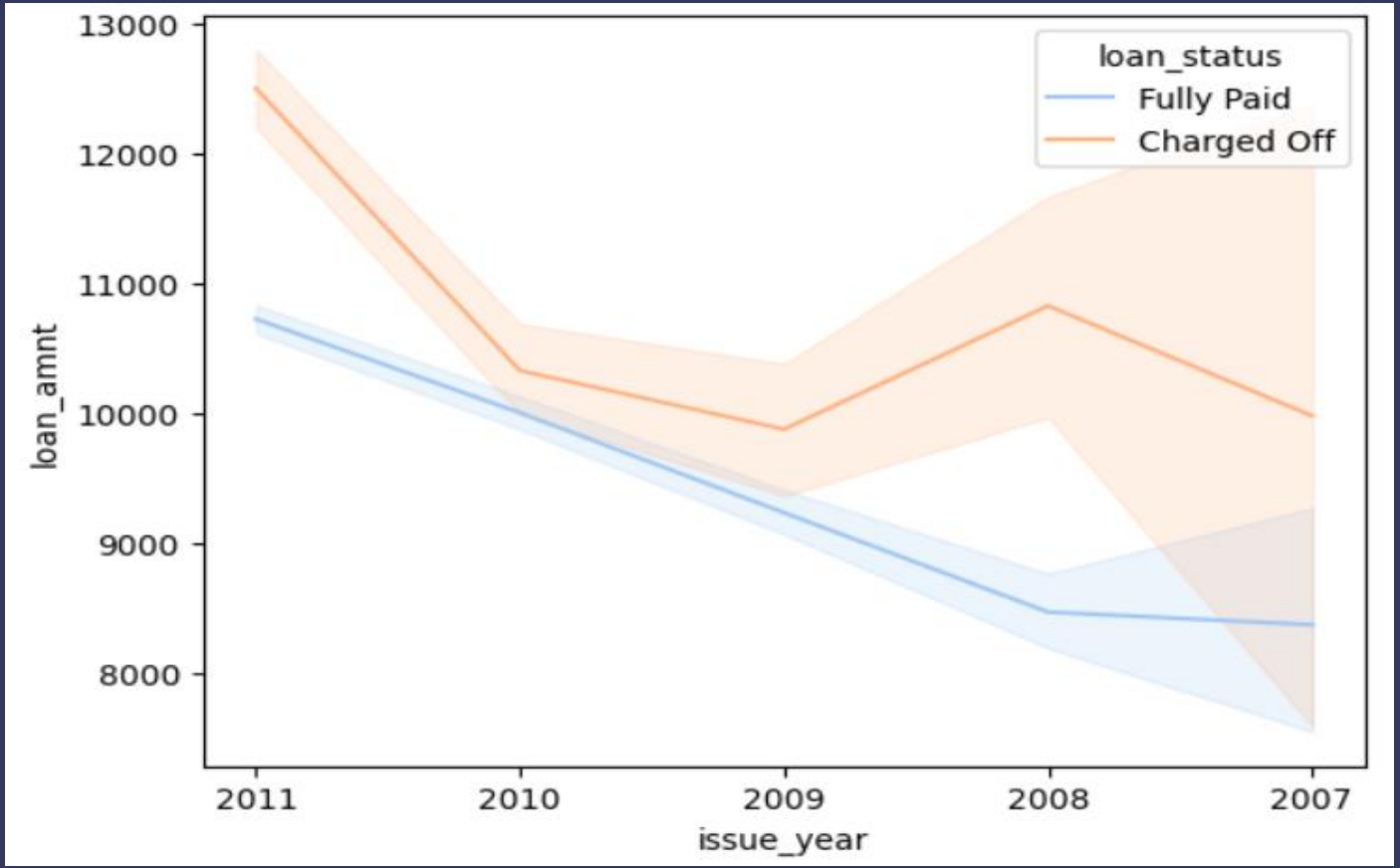
VISUALIZING LOAN VS LOAN PURPOSE



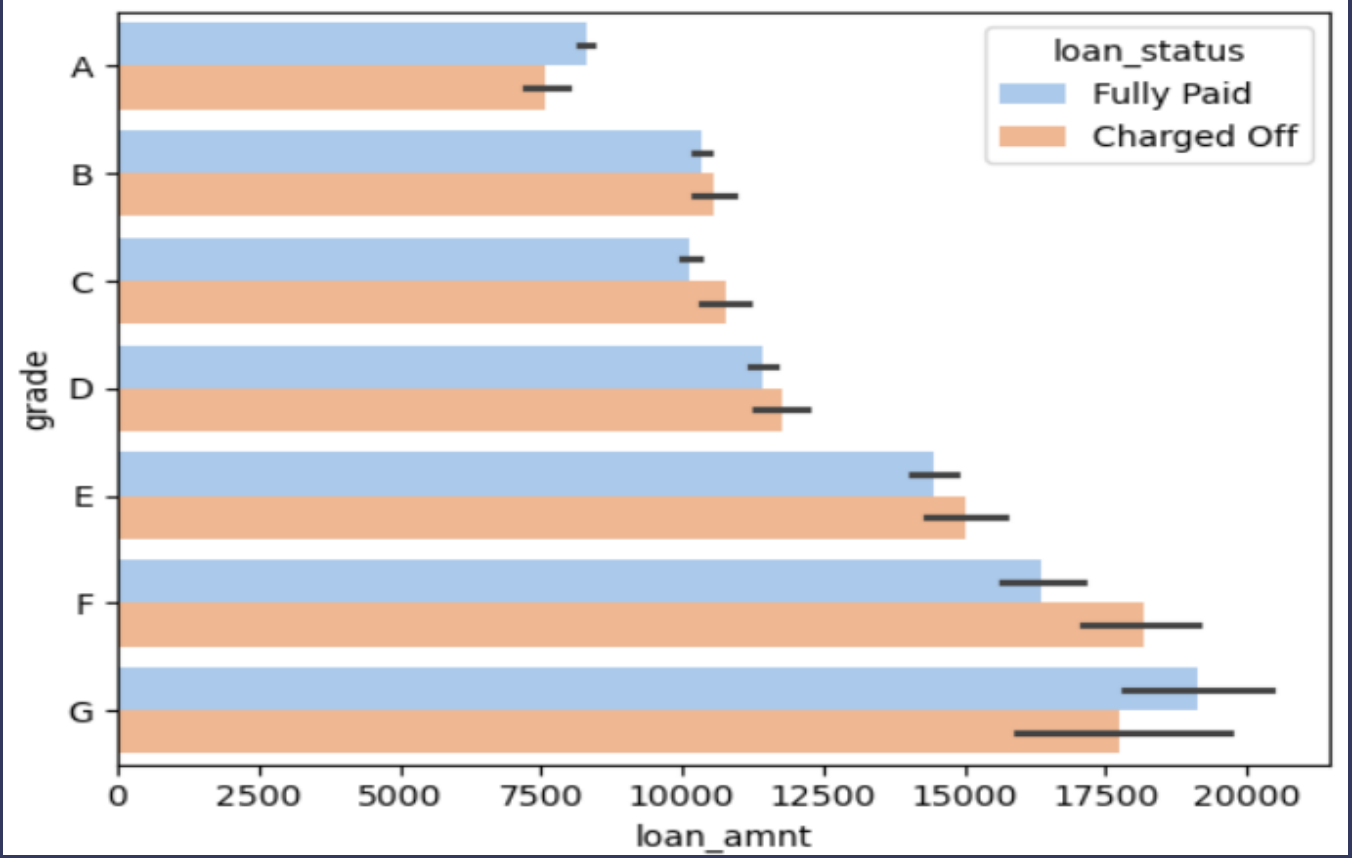
VISUALIZING LOAN AMOUNT VS MONTH ISSUED



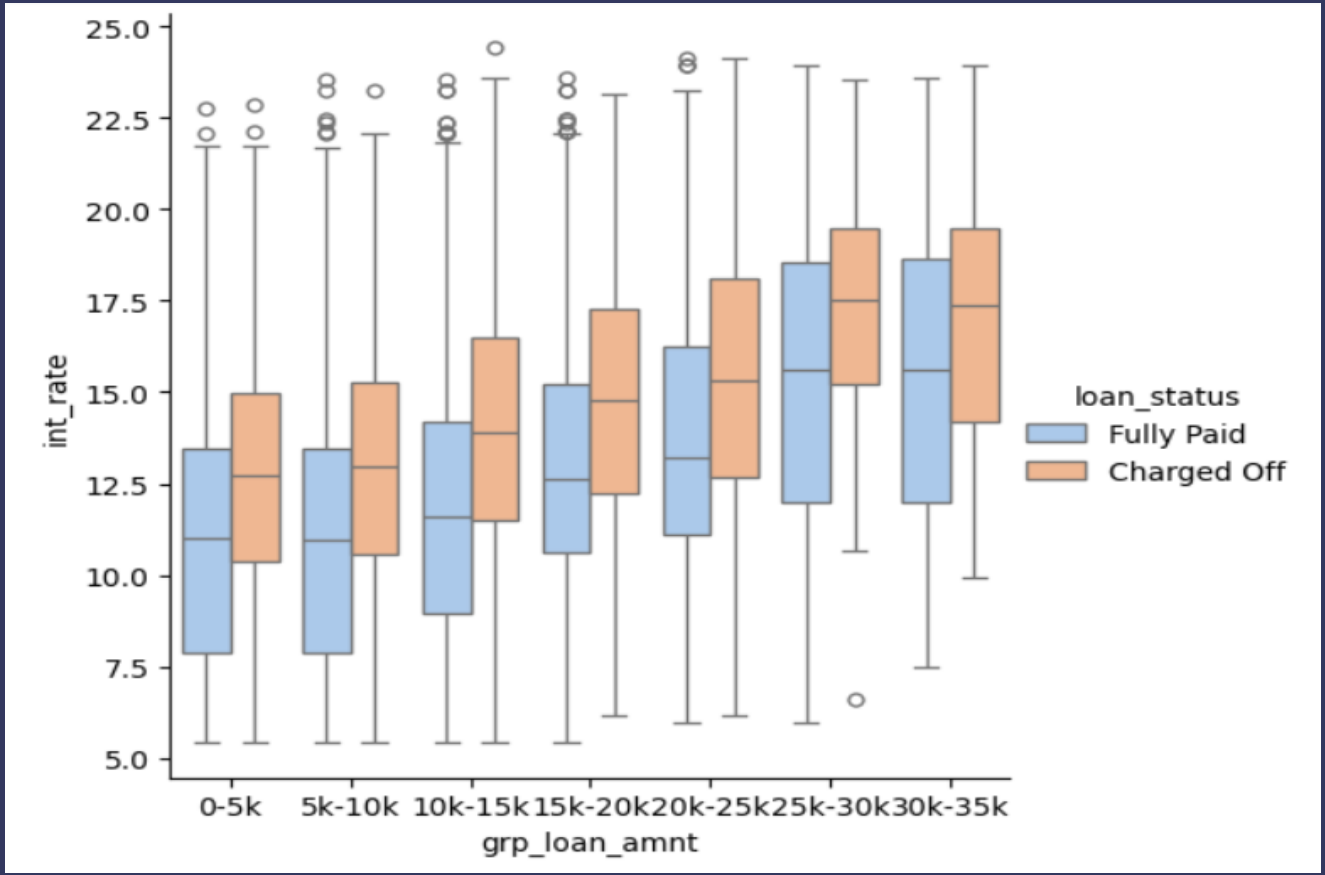
VISUALIZING LOAN AMOUNT VS YEAR ISSUED



VISUALIZING LOAN AMOUNT VS GRADE



VISUALIZING LOAN AMOUNT VS INTEREST RATES



CONCLUSIONS : BASED ON ALL THE ABOVE ANALYSIS, THERE IS MORE PROBABILITY OF DEFAULTING WHEN:

- the grade is B.
- taking loans to clear other debts
- having ownership of the house as Rent
- loans are taken for interest rates of 10-20%
- open credit lines in the borrower's credit file is 2-20
- the amount of credit the borrower has used relative to all available revolving credit its very much likely by 40-100%.
- annual income is between 31K - 58K
- installments are 145-274. Higher the installments, more likely to default.
- loan amounts applied for are really small, like, 5K-10K.
- DTI is between 12-18.
- loans applied for longer term of 36 months.
- verification status is 'Not Verified'.
- there are 0 inquiries in the last 6 months.

CONCLUSIONS : CONTINUING…

- loan is issued during the last months of the year, say, December.
- most defaulters are from the year of issue 2011.
- home ownership is Mortgage type, and income is between 60-70K. Even though, fully paid and charged off loan statuses are going hand in hand. We cannot really categorize on this data.
- loans are taken for home improvement, and have income in between 60-70K.
- annual income is 112-140K, with loan amount of 15-17.5K
- loan amount availed is 25-30K, with 16-17.5 interest rate.
- loan taken for small businesses with an amount 14-15k.
- loans are taken in the December month of the year in between 12-13K.
- loans taken in the year 2011, in between amount of 12-13k.
- loan taken between 17.5-20K by grade F.
- loans taken with interest rates of 17.5-20%, for amounts between 25-35K.

THANK YOU !