LENDING CLUB CASE STUDY

SIDHANT DAS

AYUSH SRIVASTAVA

OVERVIEW

- One of the largest consumer loan marketplaces gives loans to customers post validating the eligibility of their loan amount and interest rate.
- This process involves assessing many parameters which includes their salary, home location past loan experience etc.
- The given analysis is the study of one of the datasets provided to us which contains details of the loan applicants and the corresponding consumer and loan attributes and whether they paid the loan fully/paying it currently or have defaulted in their payments.

TECHNOLOGY USED IN THE ANALYSIS

- Github repository to host the details of the analysis.
- Comma separated delimited file which is the source of the data.
- Python code, a programming language used for Data Analysis
- Python libraries used in the study
  - Pandas
  - Matplotlib
  - seaborn

DATA CLEANING

- Inefficient data like high NULL values or single value columns etc. which needed to be removed to improve the data analysis. Some of the criteria along with column on which cleaning was done have been shared.

Data Cleaning Step	Columns removed
list of columns with all values as nulls	'mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit'
Checking for columns with all single unique values as they will not contribute to our analysis	'pymnt_plan', 'initial_list_status', 'collections_12_mths_ex_med', 'policy_code', 'application_type', 'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens'
Dropping columns with high % of null records	'next_pymnt_d', 'mths_since_last_record', 'mths_since_last_delinq'
Dropping a few more num_cols which don't seem relevant in the analysis	'collection_recovery_fee', 'funded_amnt', 'funded_amnt_inv', 'recoveries', 'revol_bal', 'total_pymnt', 'total_rec_int', 'total_rec_late_fee', 'total_rec_prncp', 'last_pymnt_d', 'last_credit_pull_d', 'last_pymnt_amnt', 'total_pymnt_inv', 'earliest_cr_line', revol_util, last_pymnt_d, last_credit_pull_d
Removing patters for loan defaulters and customers with loans paid off, loan_status = 'Current'	loan_df = loan_df[loan_df['loan_status'] != 'Current']

DATA IMPUTATION

We found many missing values which cannot be removed, hence to get help in certain numeric calculations we have substituted the missing values as mentioned below.

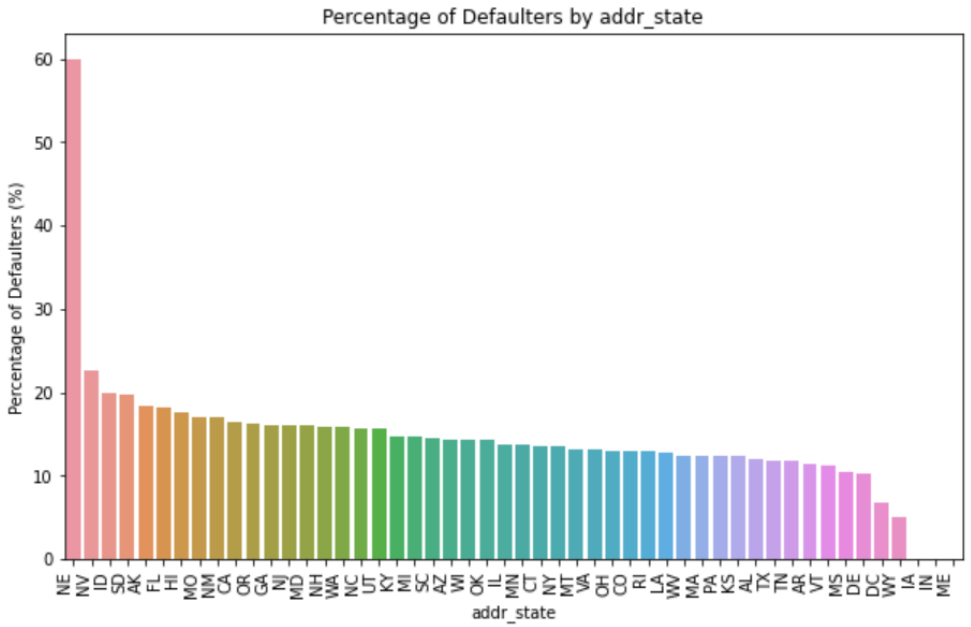
- Updating the nulls in pub_rec_bankruptcies column with -1 which would denote 'Not known'
- Imputing 'NA' for missing emp_length values
- Removing the % sign from the column and typecasting it to numeric datatype
- Removing the % sign from the column and typecasting it to numeric datatype
- Since the values 'Verified' and 'Source Verified' are more or less the same, we'll merge the values as 'Verified'

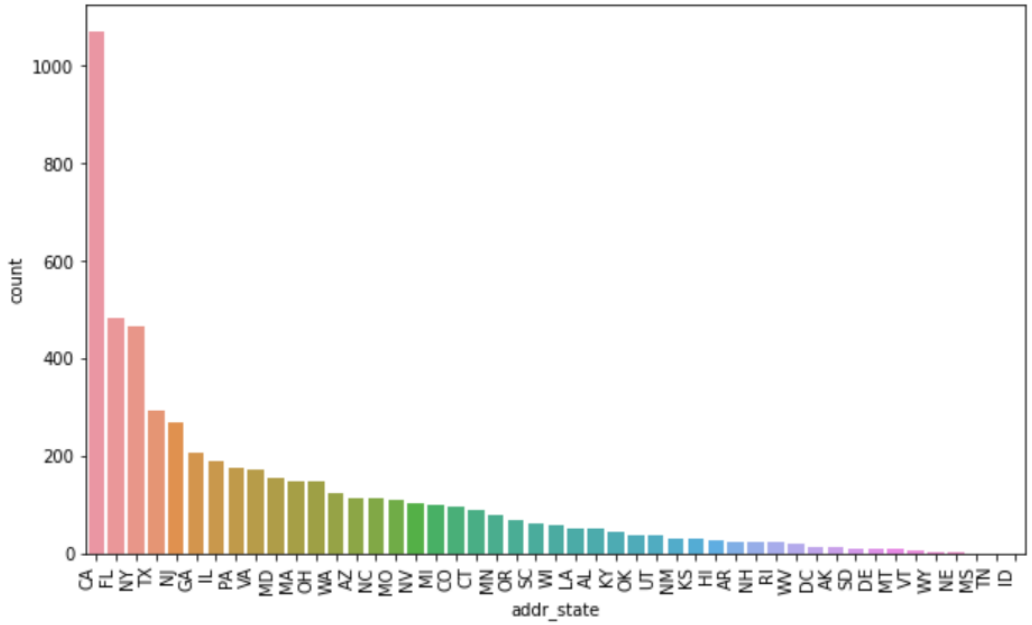
KEY VARIABLES IDENTIFIED IN THE ANALYSIS INSIGHTS

Key Variables	Variable Types
Employment length	Categorical
Home ownership	Categorical
Interest rate	Numerical
Loan purpose	Categorical
Loan amount	Numerical
Debt-to-income ratio	Numerical
Inquiry in last 6 months	Categorical
Loan Term	Categorical
Loan Grade and sub-grade	Categorical

BORROWER STATE LEVEL ANALYSIS

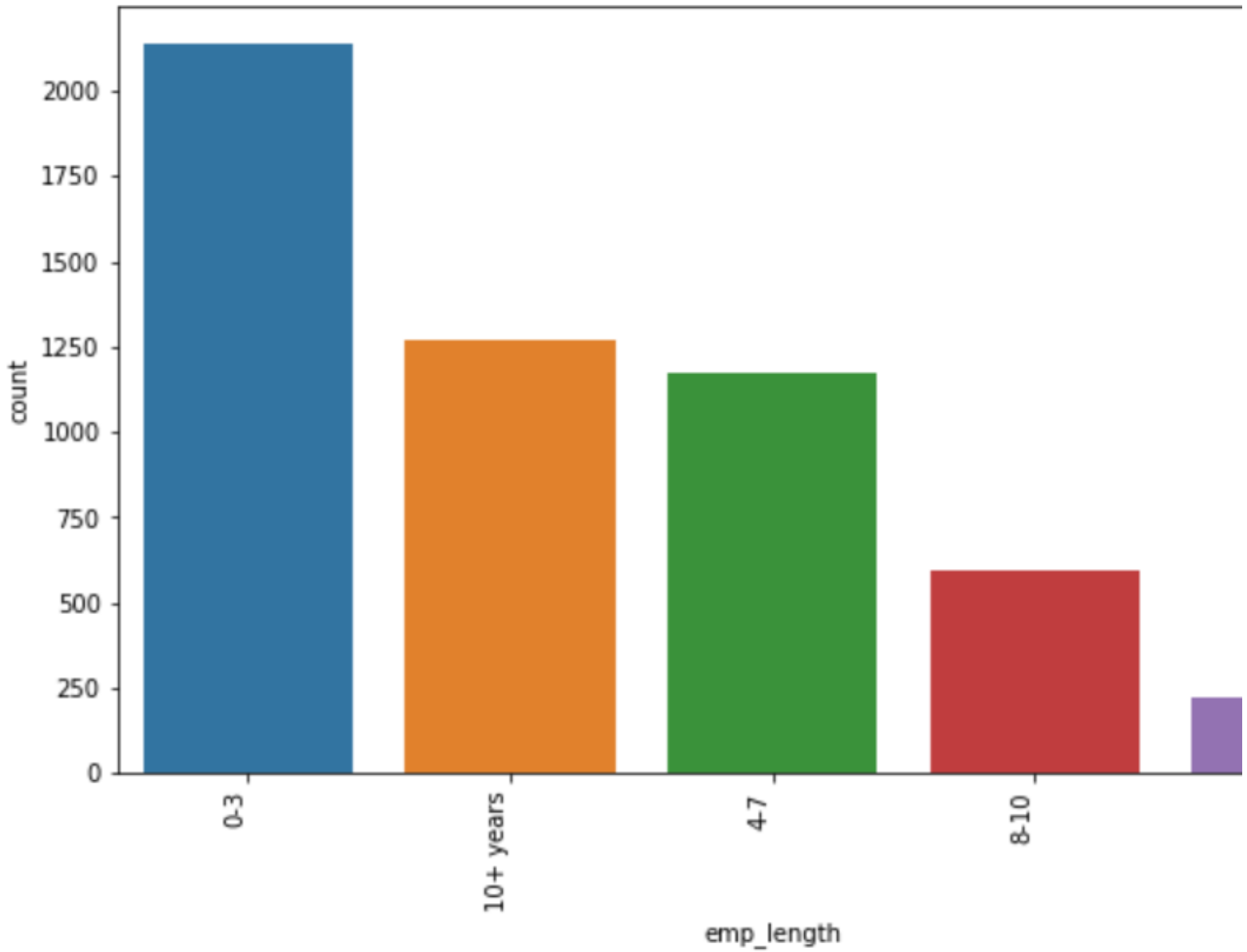
- CA has the highest no. of defaulters, followed by FL and NY. Percentage of loan defaults is highest (~60%) in NE though. Average percentage ranges between 10% - 20%. CA has 17% defaulting, while FL has 19% and NY has ~14% defaulting.





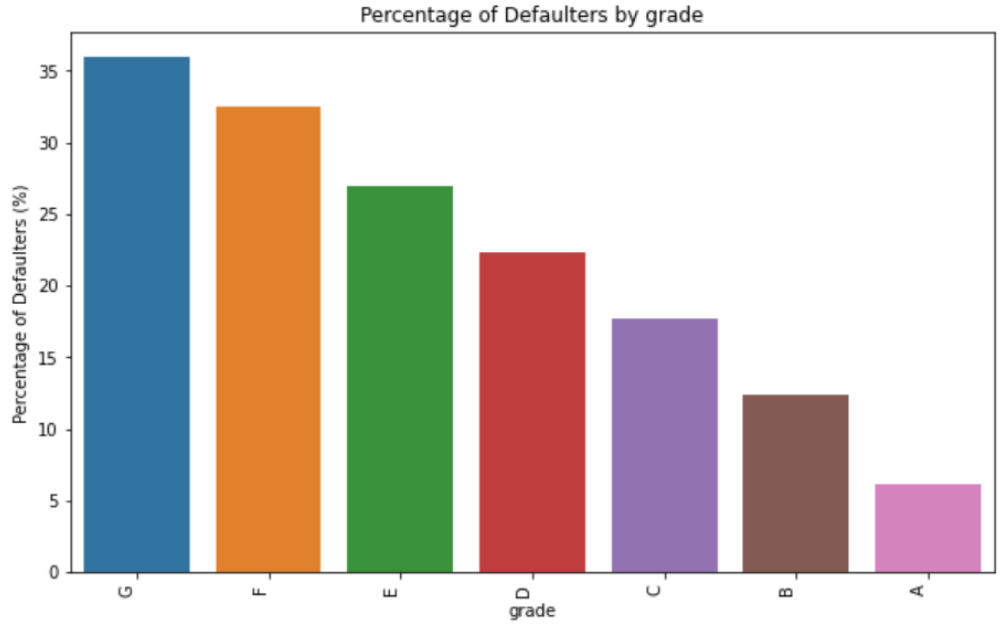
BORROWER WORK EXPERIENCE OBSERVATION

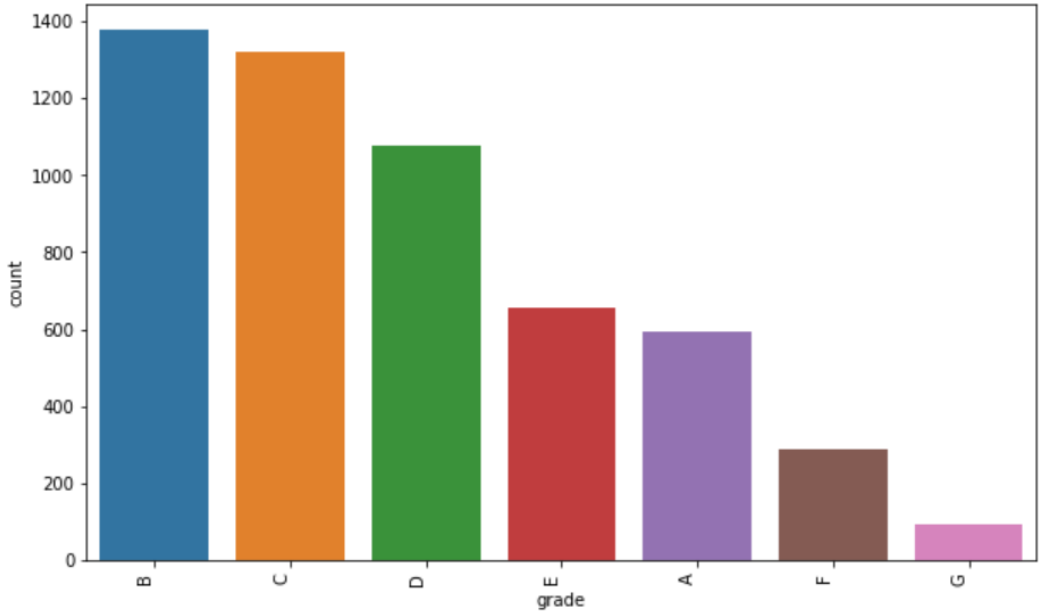
- Highest no. of defaulters are with experience 0-3 years



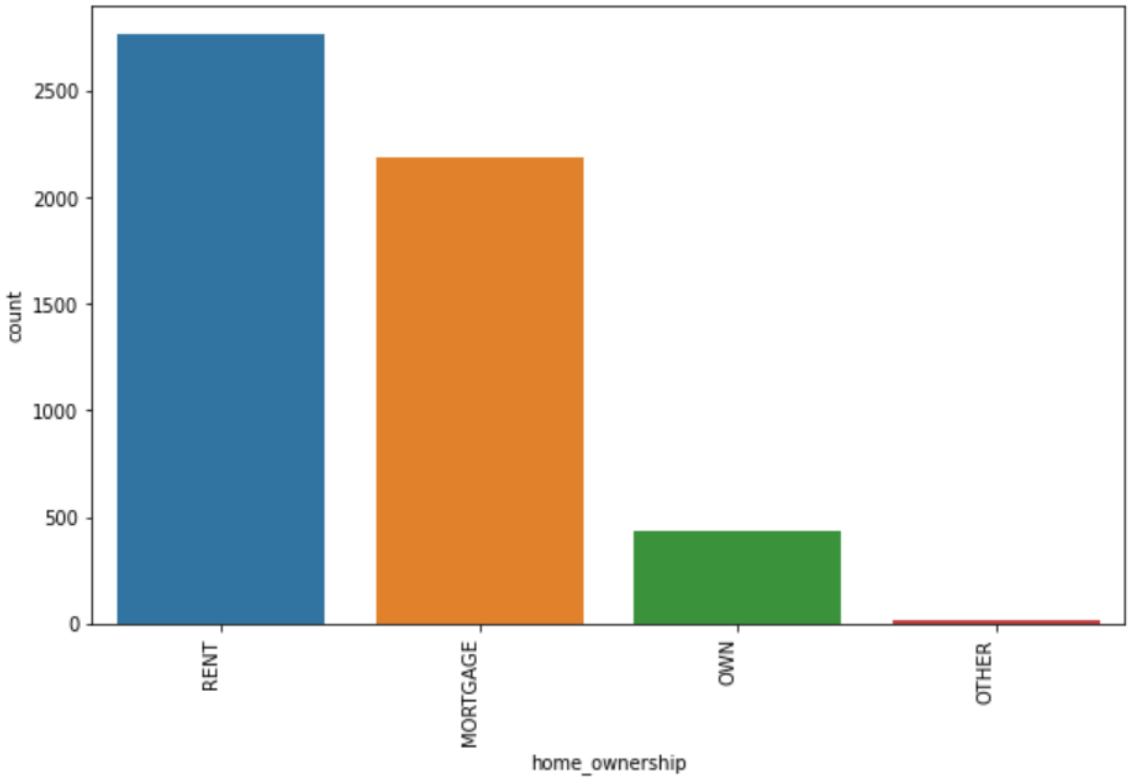
ANALYSIS ON THE LOAN GRADE ON THE DEFAULT PATTERN

1. Loan Grade B has the highest defaulters followed by C and D. However, percentage of defaulters seem to be increasing with decrease in loan grade quality.
2. For example, Loan grades A & B with highest credit quality have low percentage of defaults between 6% - 12%. At the same time Loan grade G with lowest credit quality has highest percentage of defaults (~35%).

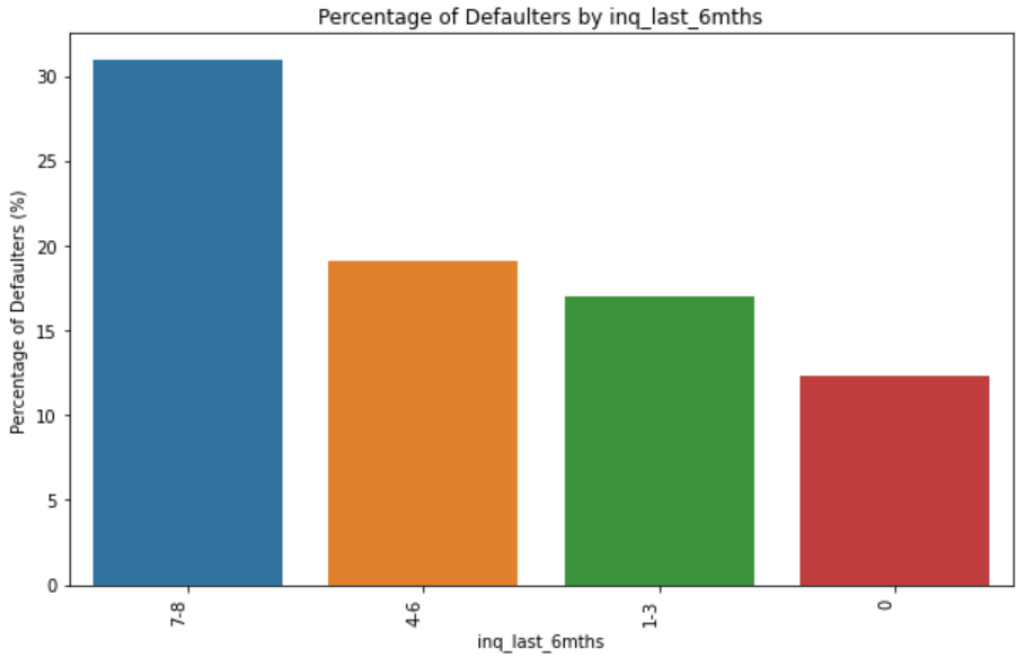




Majority of the defaulters don't own a house and live in a rented house.

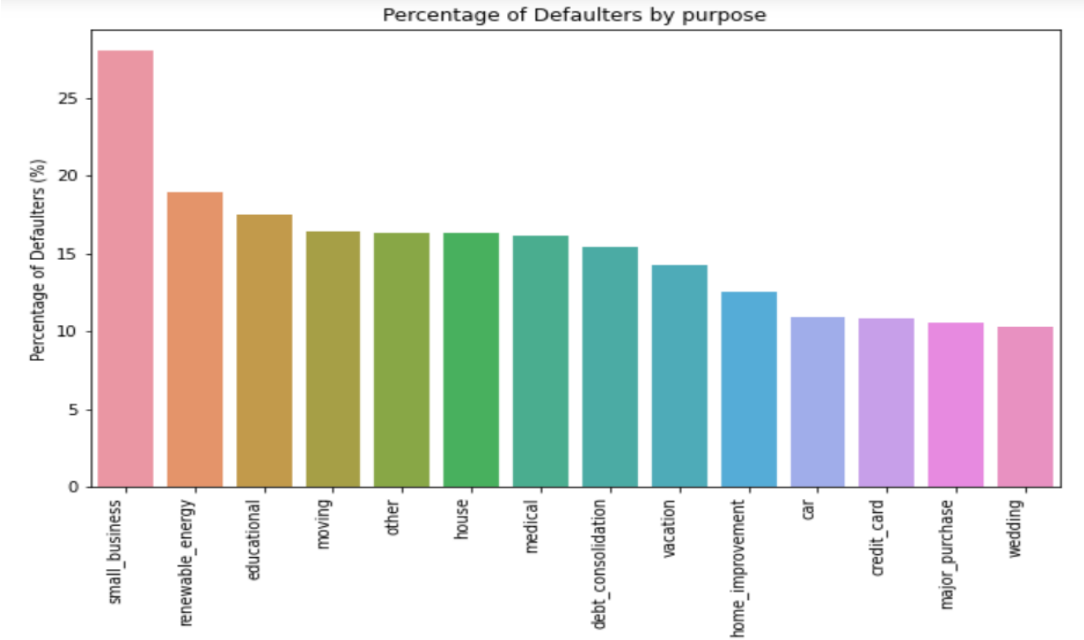


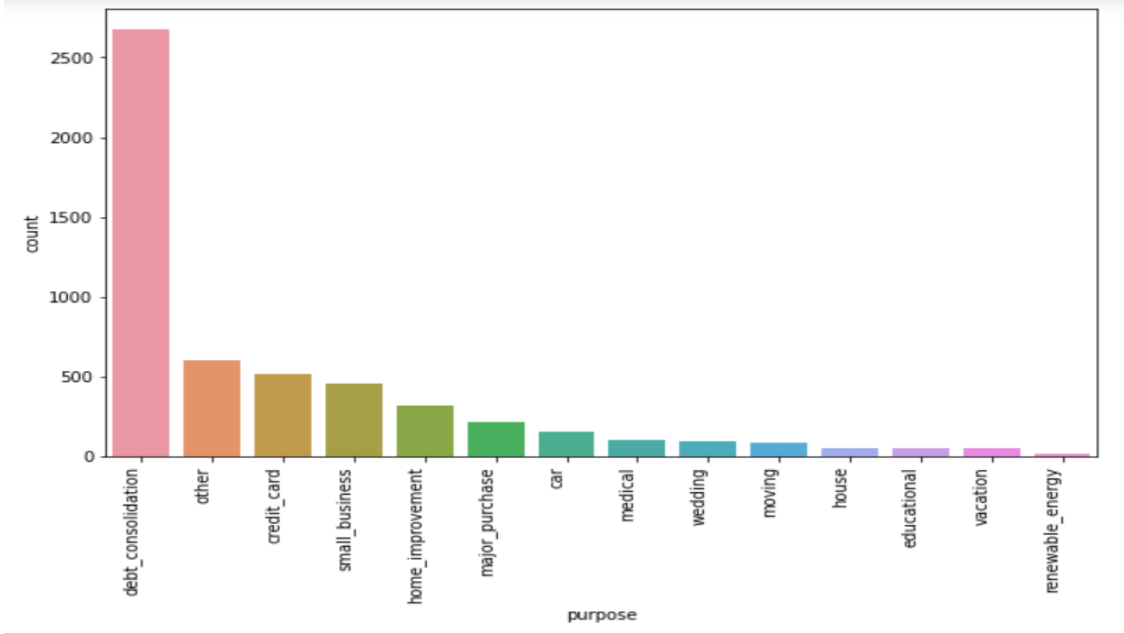
Majority of the defaulters have 1-3 inquiries in the last 6 months. However, Percentage of defaulters is highest (~30%) for people who have 8 counts of loan inquiries in the last 6 months.



LOAN PURPOSE ANALYSIS - I

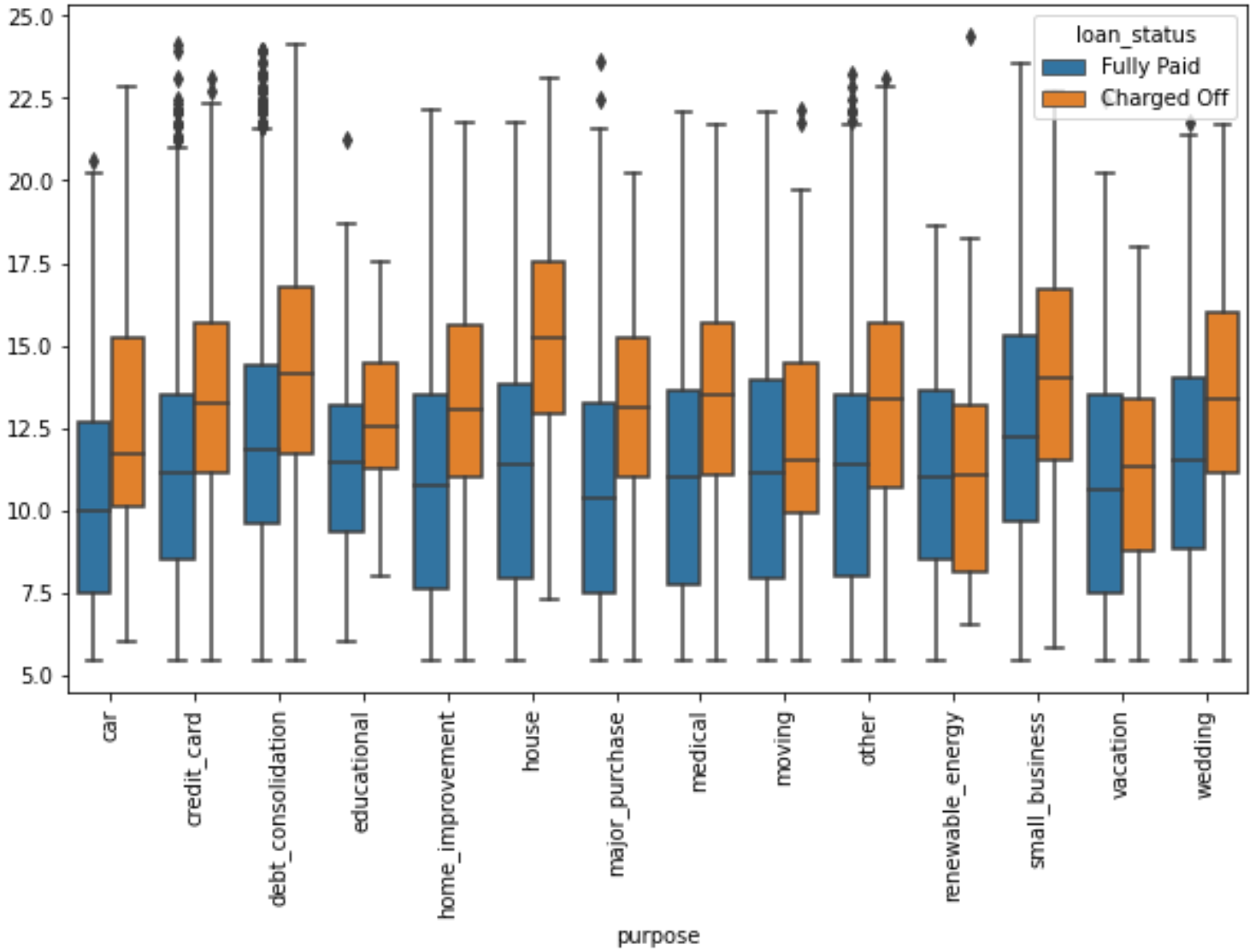
- Most defaulters purpose of taking the loan has been debt consolidation. Percentage of loan defaulters increases to ~28% from the average of ~15% for applicants who need the loan to finance their small business.





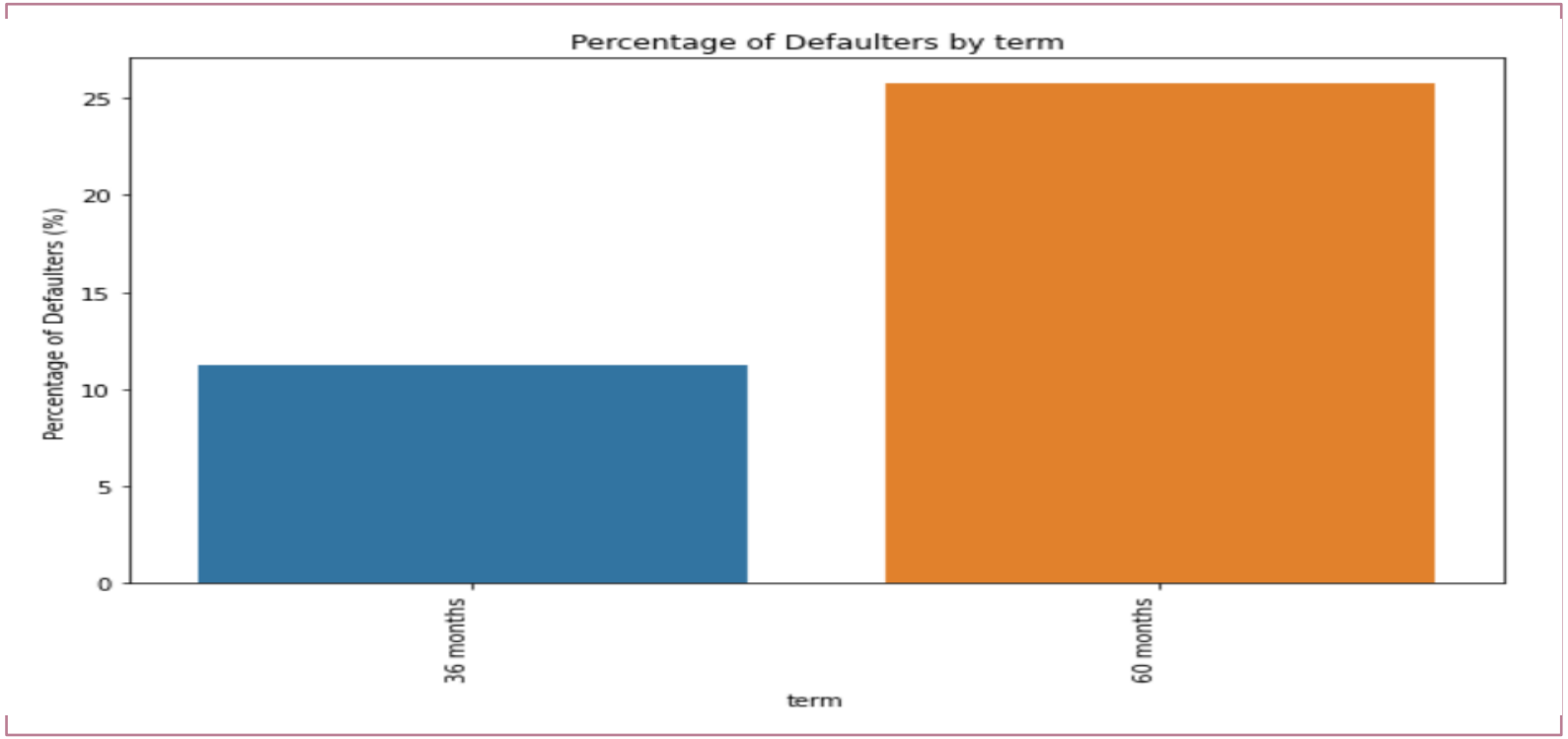
LOAN PURPOSE ANALYSIS - II

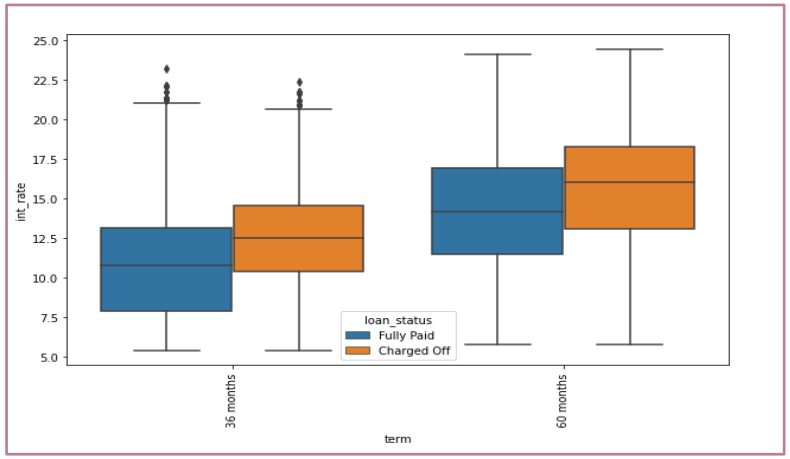
- The reason behind high defaulters for loan purpose as debt consolidation and small_business is the higher interest in these types of loans.

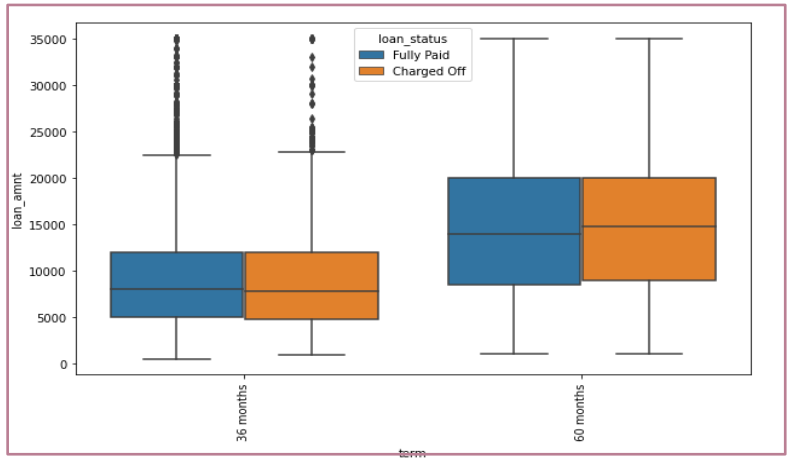


OBSERVATION ON LOAN TERM ON DEFAULTER BEHAVIOUR

- While most of the defaulters are ones with 36 months' loan term. But the percentage of defaulters doubles (~25%) for 60 months' term as compared to 36 months' term (~ 12%).
- This is because the loan amount and interest rates are much higher for loans with 60 months’ term.

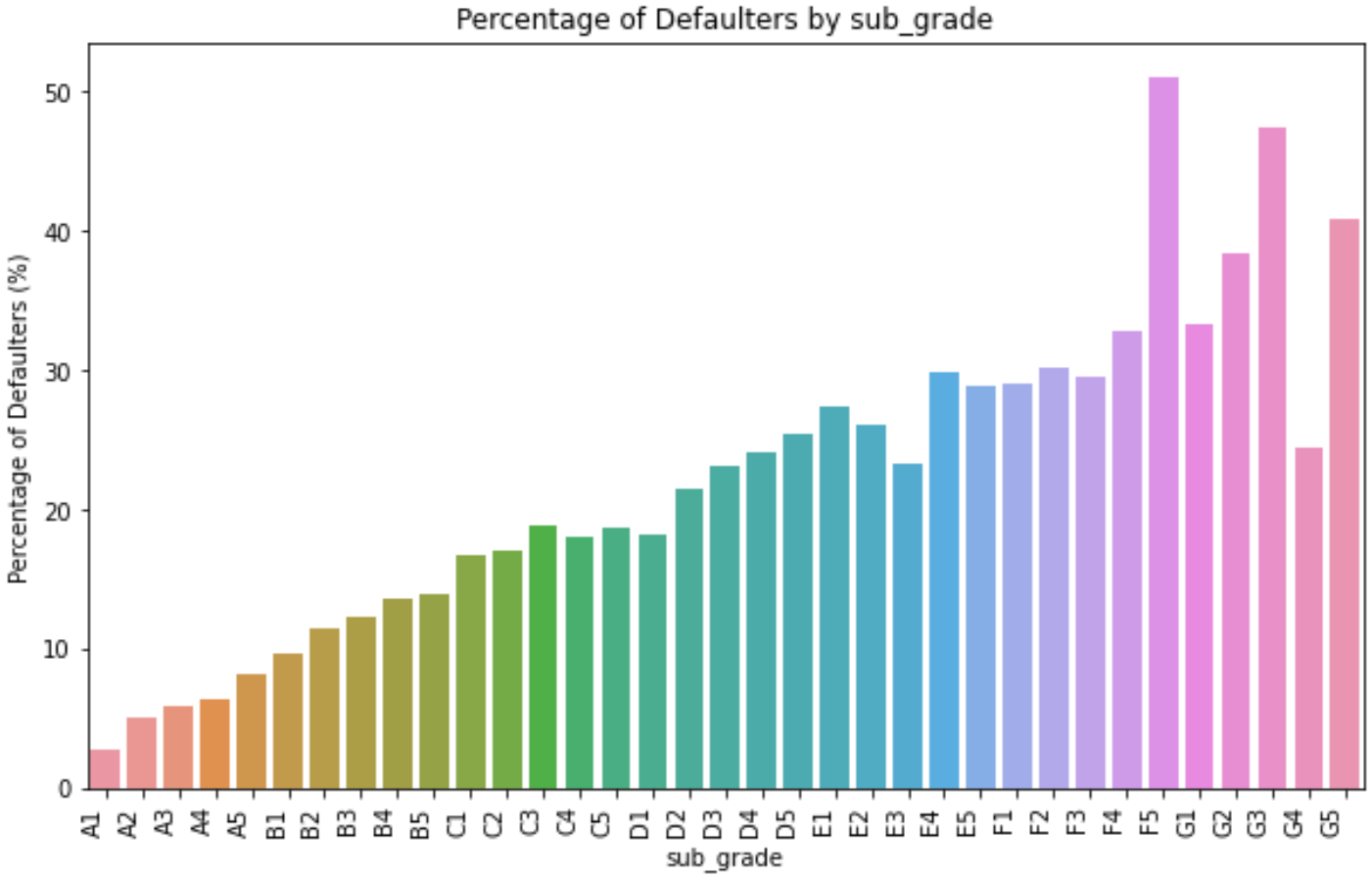






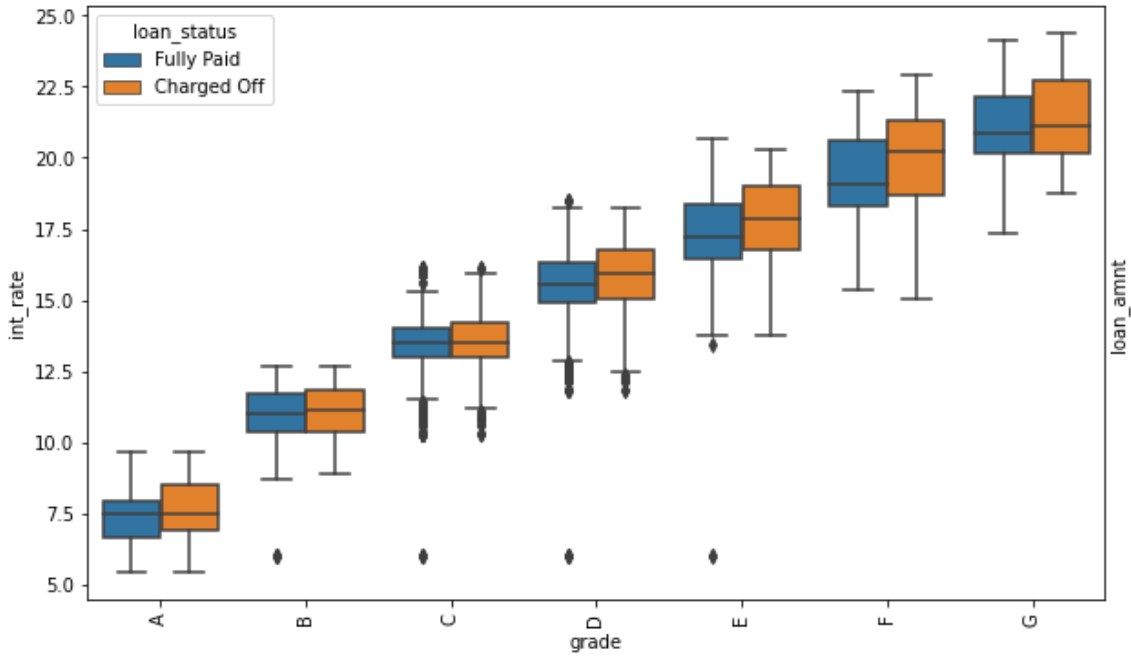
PERCENTAGE OF DEFAULTER BASIS OF LOAN SUB GRADE

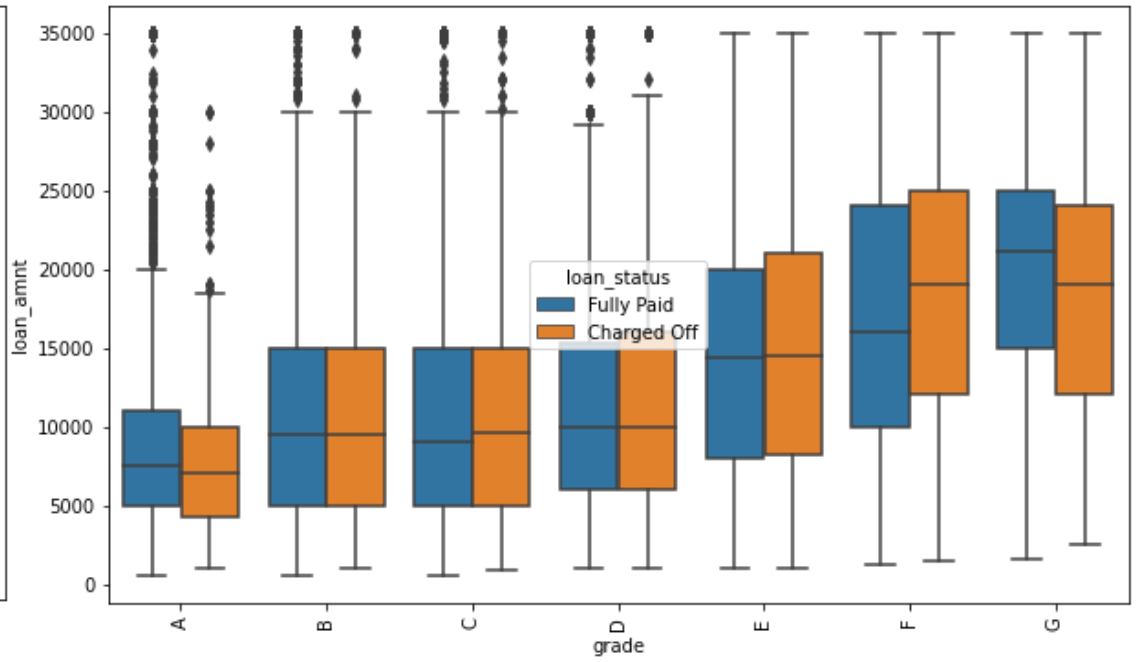
- Similar to grades, the probability of defaulting a loan increases with decreasing sub-grade (A1 to G5) with F5 having the highest percentage of loan defaulters (~50%).



EFFECT OF INTEREST RATE AND LOAN AMOUNT ON LOAN GRADE

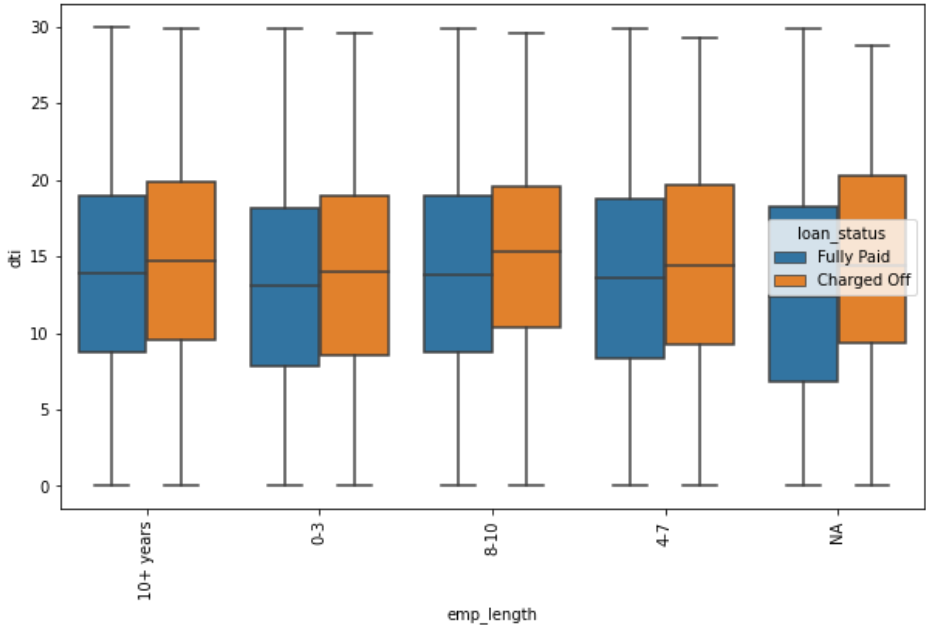
- The given charts show the correlation between Loan amount vs Loan grade and Loan interest rate vs Loan Grade.
- The Average Loan Amount assigned to loan grade G is higher than other grades and is charged at higher interest percentage.
- The trade-off between loan amount and interest rate for lower graded loans should be reconsidered, since they are high defaulting consumers.

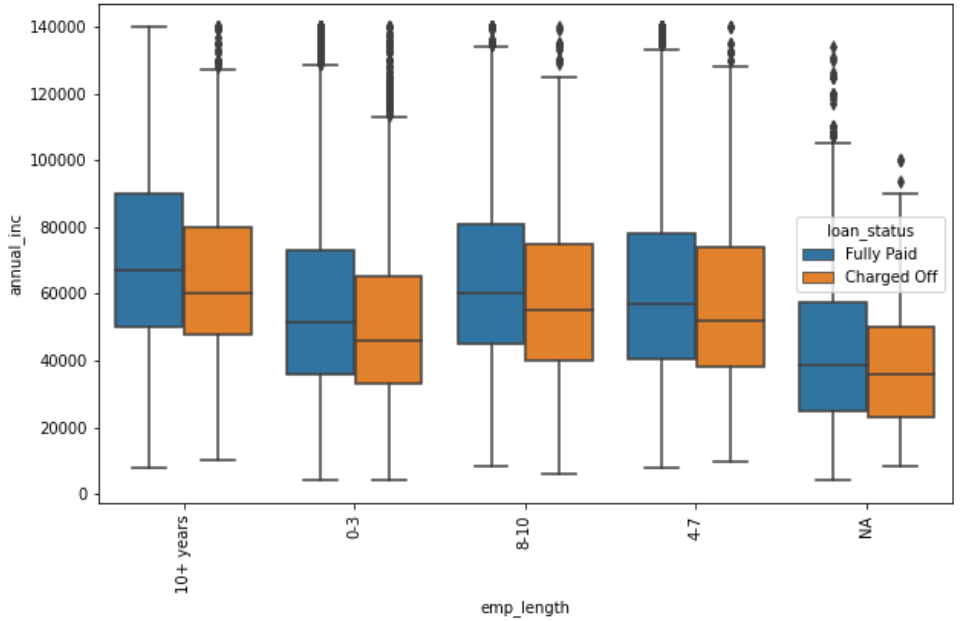




EFFECT OF SALARY AND DTI OF BORROWER ON LOAN STATUS

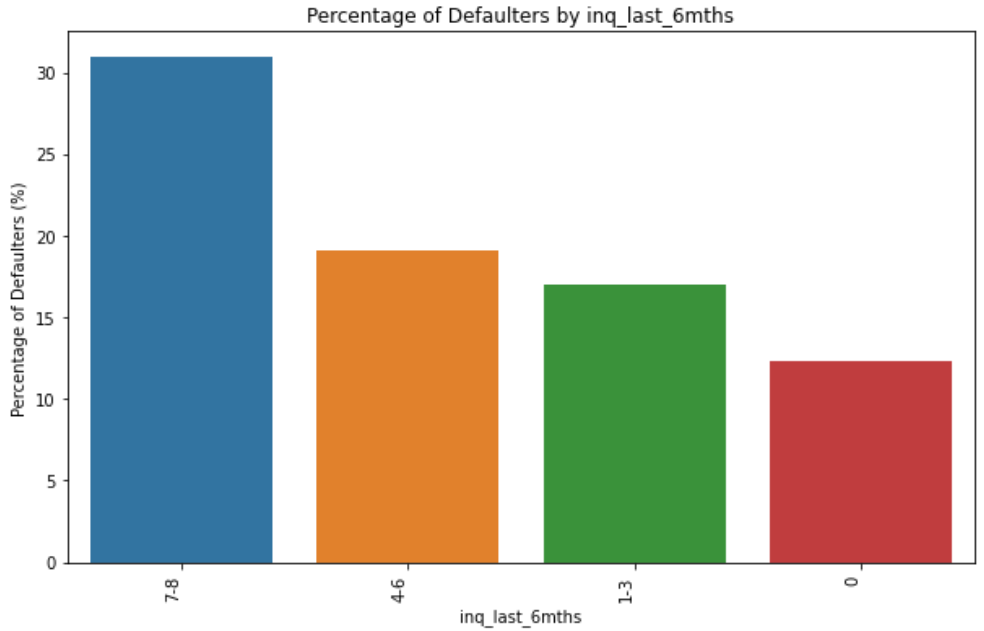
- The box plots shown below represent that annual income is directly proportional to the employee length. Similarly, higher annual income across the bands corresponds to lower chance of defaulting in the loans.
- Similarly, higher debt-to-income ratio infers higher chance of defaulting irrespective of the employment length for that person.

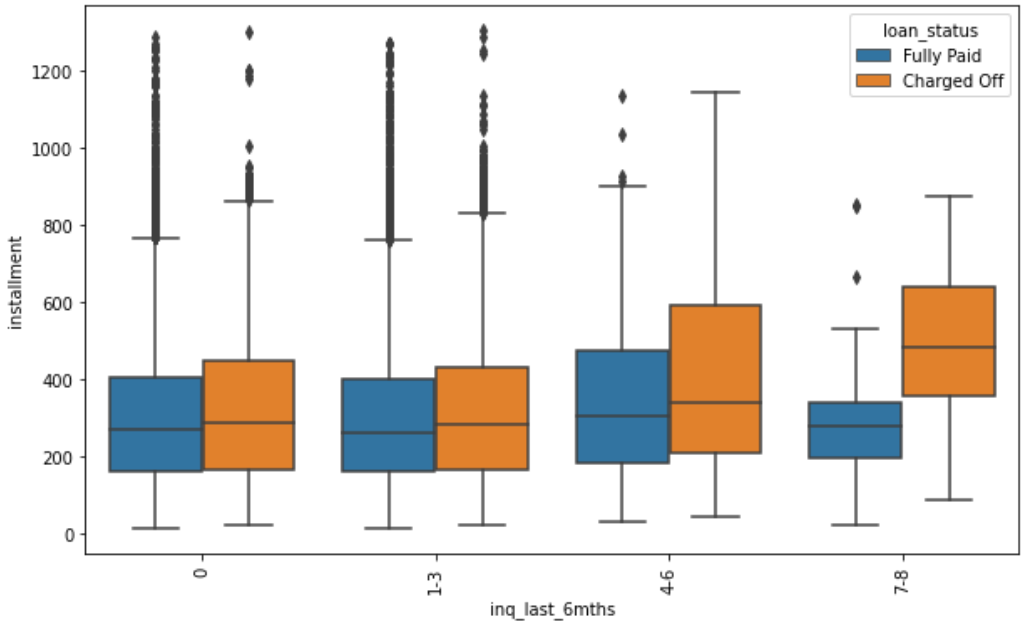




INQUIRIES IN LAST 6 MONTHS AND ITS EFFECTS

- Loan defaulters with 4+ inquiries in the last 6 months have much higher installments than applicants with lower number of inquiries. This needs to be re-evaluated as well, since higher no. of inquiries may represent a desperate individual for a loan and approving a loan with higher installments may result to defaulting.





CONCLUSIONS

- Key Drivers for loan defaults
  - States like CA and NE.
  - Loan applicants with 1-3 years and 10+ years of experience.
  - Loan grades with lower credit quality.
  - Loan applicants with rented accommodation.
  - Loan purpose – Especially small businesses and debt consolidation
  - Loan applicants with 60 month' term (higher loan amounts and interest rates).
  - Lower credit quality sub-grades, especially F5 (higher loan amounts and interest rates).
  - High installments for loan applicants with 4+ inquiries in last 6 months.
  - High DTI.