Gremener Case Study

Shobana Lakshminarasimhan
Abhishek Chouhan
Aman Patni
Abhishek Anand

Problem Statement

- The goal of this case study is to identify the driver variables behind loan default.
- To identify the driving factors that can possibly identify the risky loan applicants

Data Preparation

- The input data is in .zip format
- The data is unzipped and the .csv file is being used for the analysis
- The data has 111 variables.
- First step is to identify the variables that do not contain any info.

Assumptions

- Pre Closure of loan is very much possible
- Exact parameters of grading is not known. But can see trends such as Public Derogatory Records, Public Record bankruptcies contributing towards it.
- Grading is done before a customer is onboarded / incepted.
- Number of Open Credit Lines are Credit Lines of the customer may or may not be available with the same institution
- Attributes : Interest Rate & Grade have causal relationship
- There is no known basis of Investors and LC distribution for Funded Amount
- Term Revolving Credit balance - means channelizing available credit line for loan repayment
- inq_last_6mths : these enquiries were made while a customer was already onboarded and could be associated with the institution for any duration until the loan in settled.
- acc_open_past_24mths : The accounts opened could be in same or different institution

Data cleaning – Removing variables

- Variables that contain only one value and/or NA can be removed from our analysis. 63 such variables in the input data are removed from our analysis.
  - LoanStatNew, acc_now_delinq, acc_open_past_24mths, all_util, annual_inc_joint, application_type, avg_cur_bal, bc_open_to_buy, bc_util, chargeoff_within_12_mths, collections_12_mths_ex_med, delinq_amnt, dti_joint, il_util, initial_list_status, inq_fi, inq_last_12m, max_bal_bc, 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_last_major_derog, mths_since_rcnt_il, 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, open_acc_6m, open_il_12m, open_il_24m, open_il_6m, open_rv_12m, open_rv_24m, pct_tl_nvr_dlq, percent_bc_gt_75, policy_code, pymnt_plan, tax_liens, tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_bal_ex_mort, total_bal_il, total_bc_limit, total_cu_tl, total_il_high_credit_limit, total_rev_hi_lim , url, verified_status_joint

Data cleaning – Removing Categorical variables

LoanStatNew	Description	Is needed for analysis	Comments
addr_state	The state provided by the borrower in the loan application	No	There are more loans with addr_state such as CA, NY, TX, FL, NJ and so on. Accordingly, the charged off loans are proportionately high in such states. Cannot make anything for the anallysis.
desc	Loan description provided by the borrower	No	Date when comment is added by the borrower and the comment on the purpose of taking the loan. There is a separate column for purpose. Also purpose is a crisp form of the description and can be used for the analysis
emp_title	The job title supplied by the Borrower when applying for the loan.*	No	Unable to make out anything from this column. Some kind of text processing is needed to categorize the employers based on the title for the analysis. Ignoring in the current analysis
id	A unique LC assigned ID for the loan listing.	No	Not needed for the analysis
member_id	A unique LC assigned Id for the borrower member.	No	One to one with the loan id for the given data and not needed for the analysis
url	URL for the LC page with listing data.	No	Redundant information - Has Loan Id which is same as id
zip_code	The first 3 numbers of the zip code provided by the b+B4orrower in the loan application.	No	Same as addr_state. Cannot make out anything from this zip_code data

Cleaning up the variables

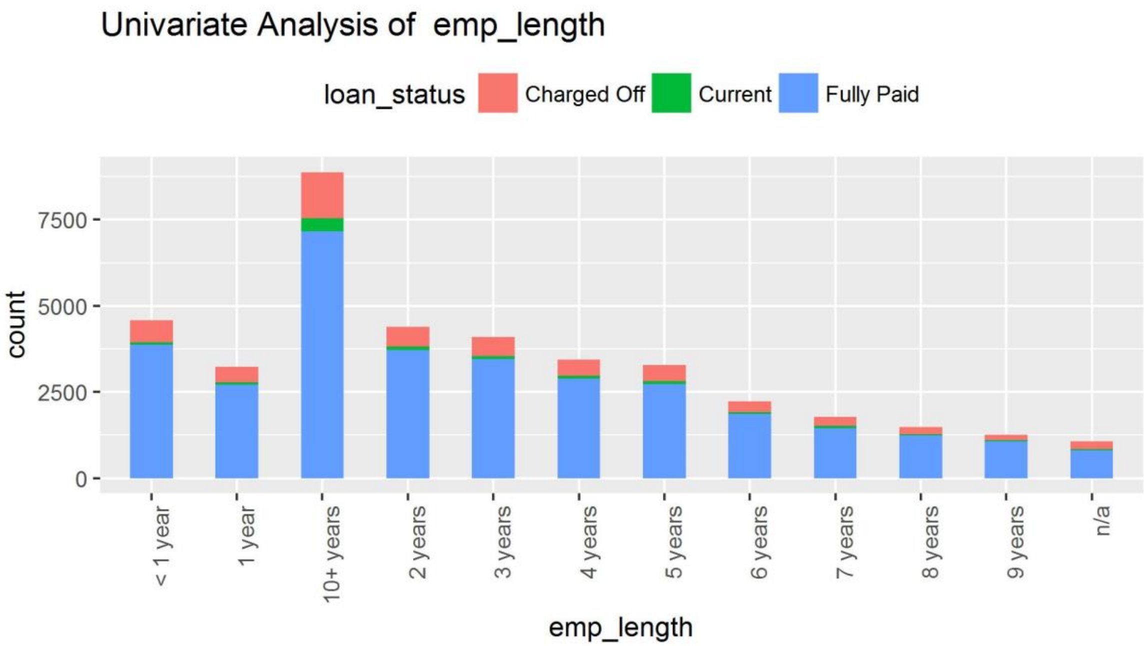
- Checked for duplicate rows and no duplicate rows are found
- Date variables – issue_d, last_pymnt_d, next_pymnt_d, earliest_cr_line and last_credit_pull_d are in month-day format
- To convert the date variables to Date format, a dummy date “01” is added and used use_PosixCT to convert the date variables to Date format.
- “term” column has “months” suffix, which is removed to convert it to integer data type
- “int_rate” and “revo_util” has %, which is removed to convert them to numeric values.

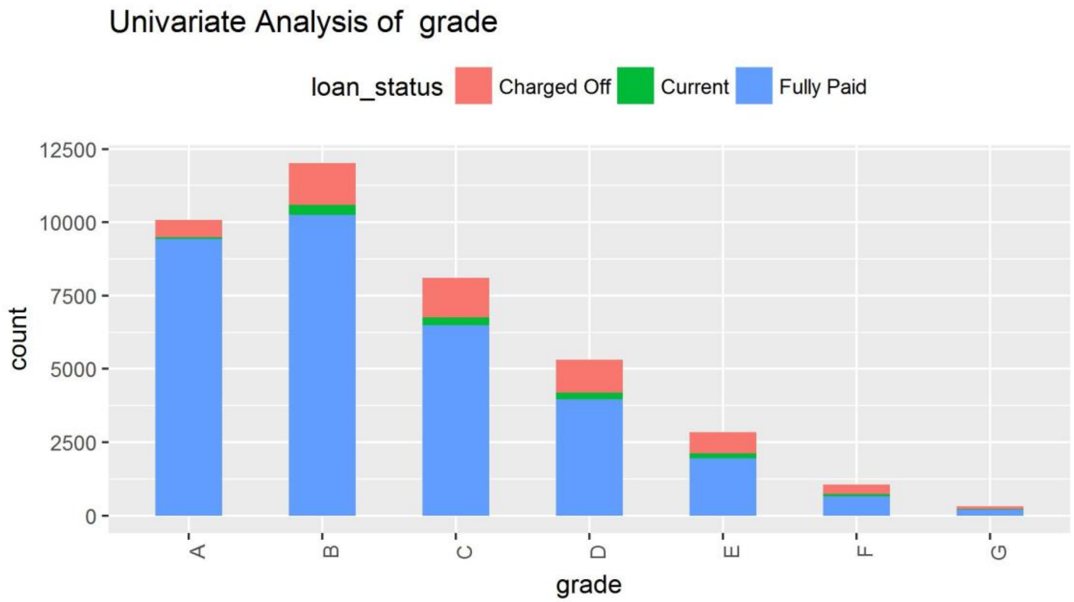
Note	Outliers are handled in the box plots.

Derived variables

- The difference between the issue_d and the last_pymnt_d is computed in terms of number of months and stored as act_pymnt_term.

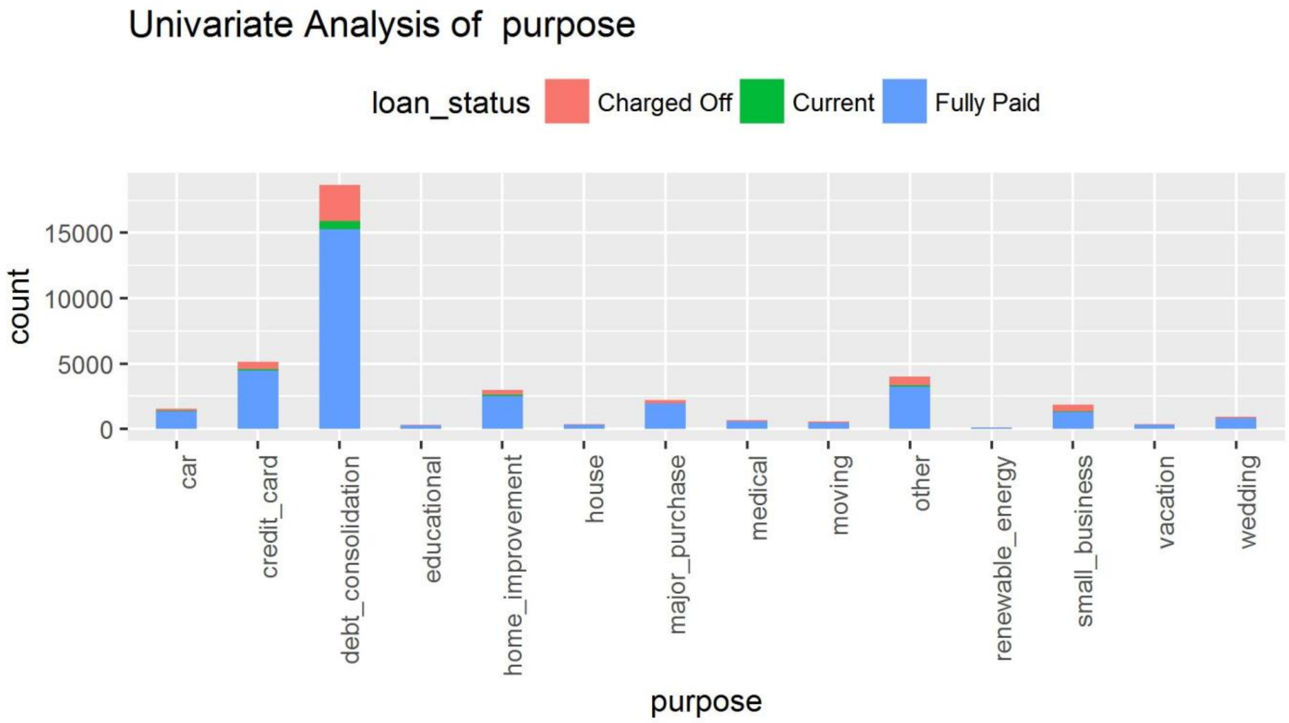
Univariate Analysis of Categorical Variables

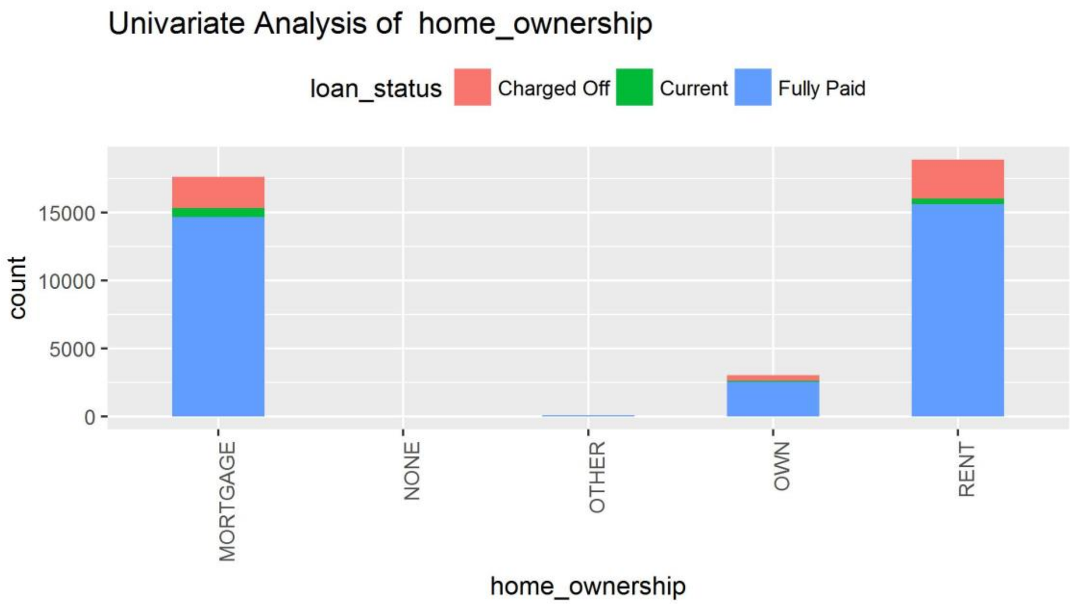




Inference:
More number of loans are given to applicants with 10+ Yrs of experience and more is the number of defaulters in this bracket
Number of defaulters are in increasing proportionally from Grade A to G

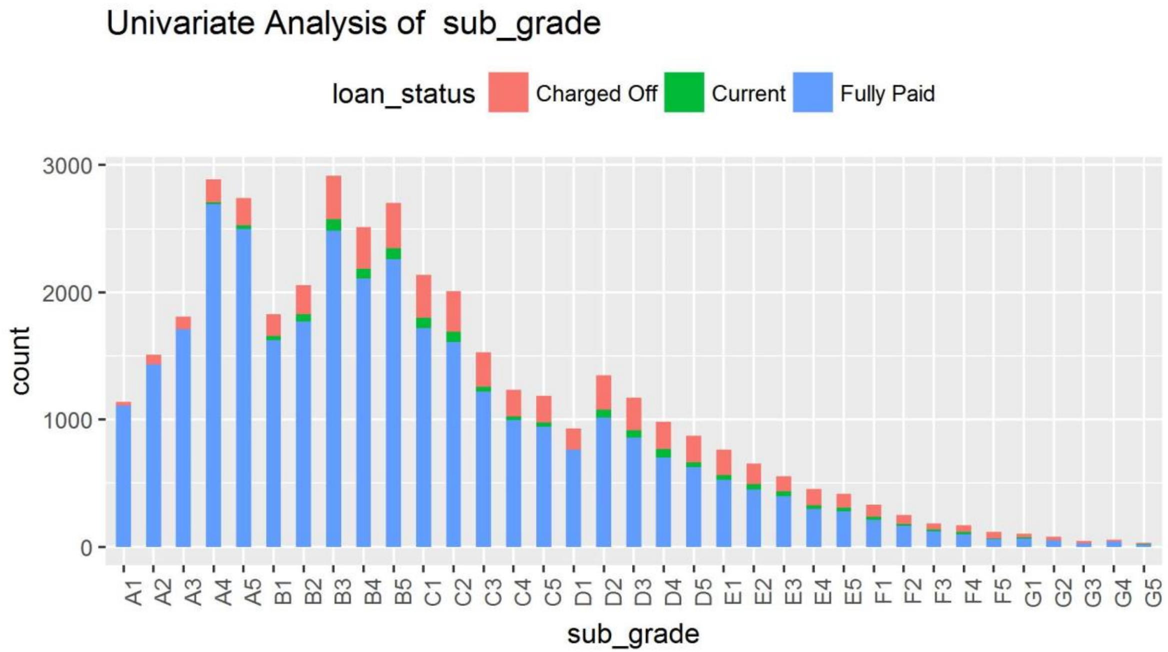
Univariate Analysis of Categorical Variables



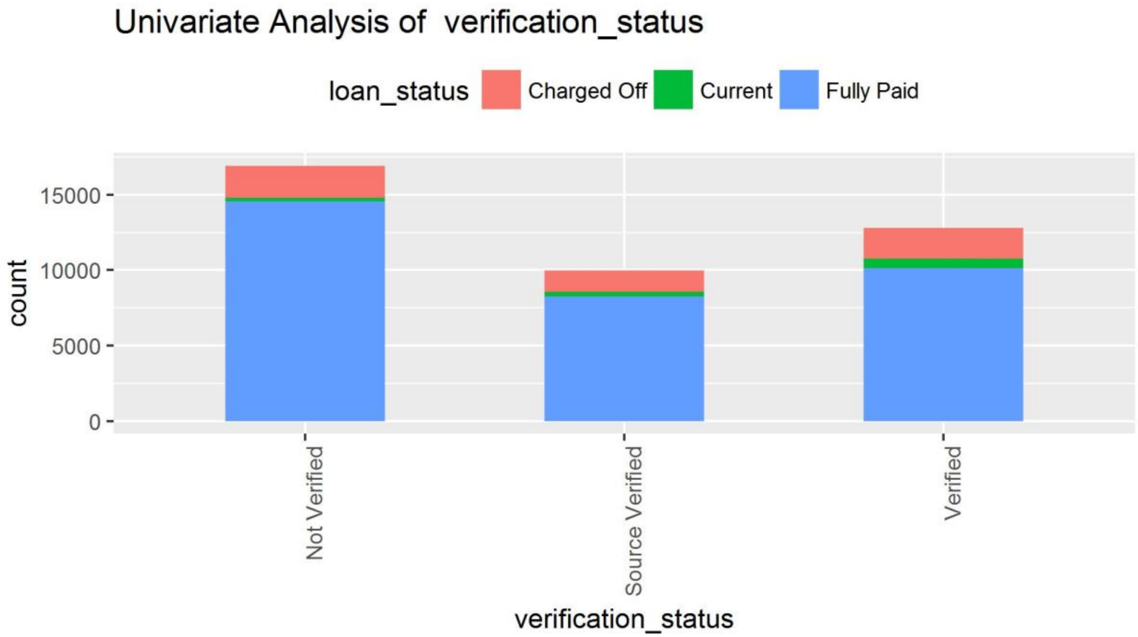


Inference:
Maximum number of loans taken in bracket Debt Consolidation and accordingly number of defaulters is alsi high
Defaulters are proportionally high for applicants who have accommodation on Rent

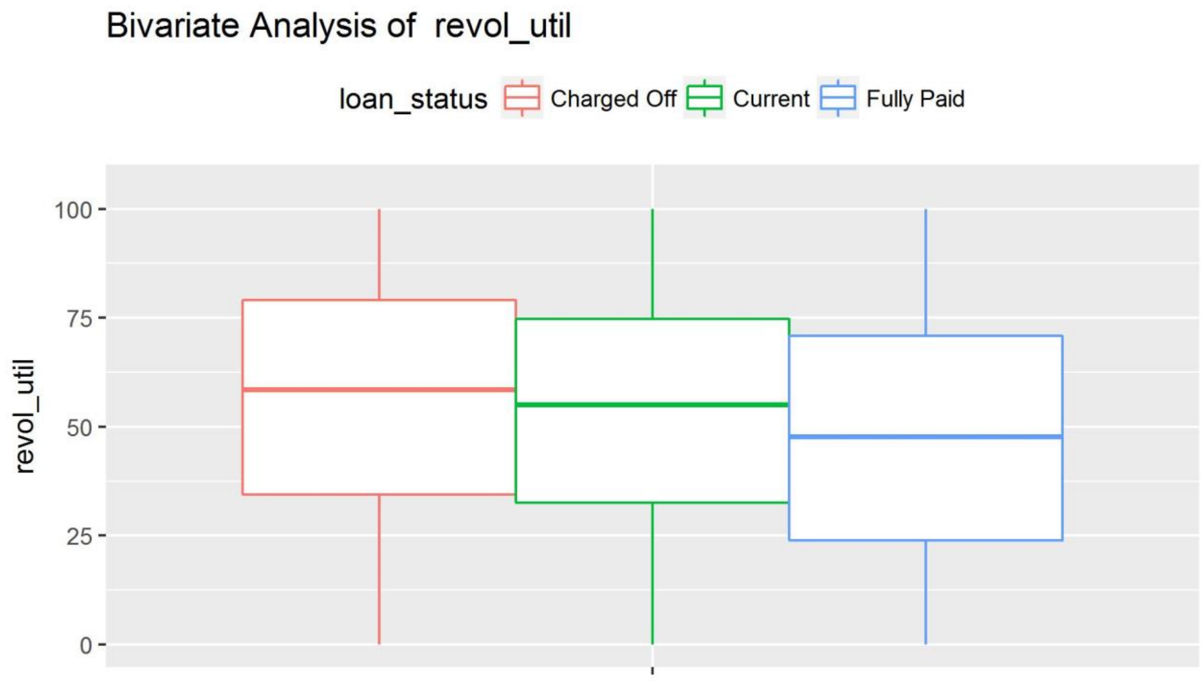
Univariate Analysis of Categorical Variables

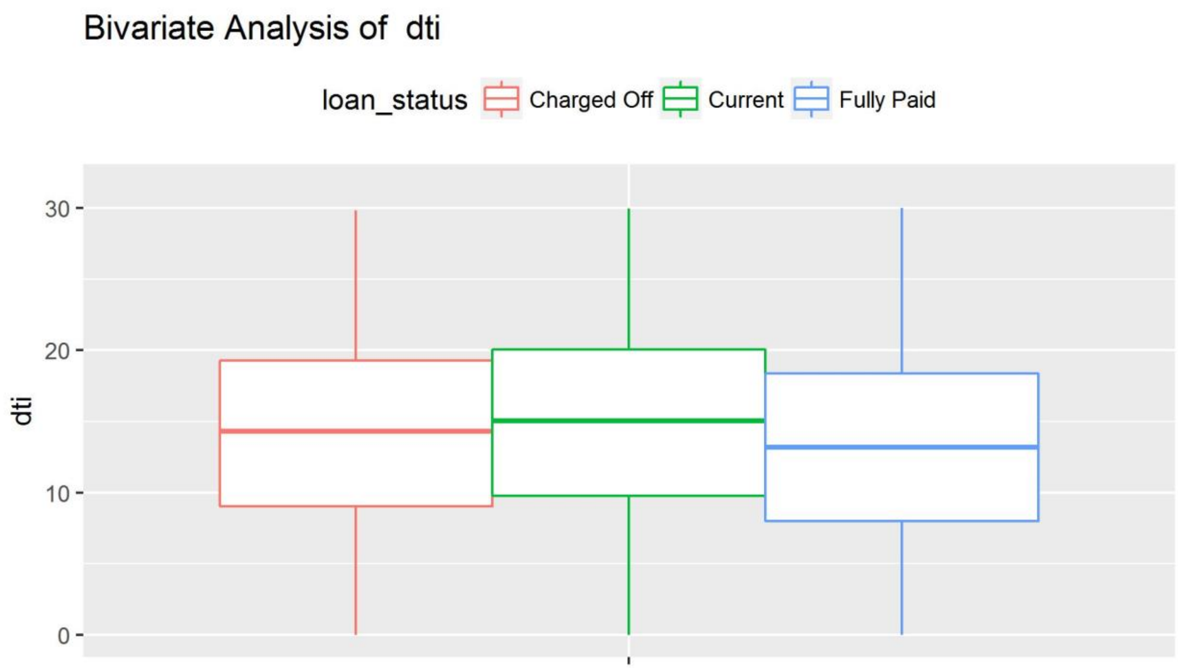


Inference:
Count of loans given decreases from Grade A to G
Equal proportion defaulter to fully paid was found with respect to Verification Status



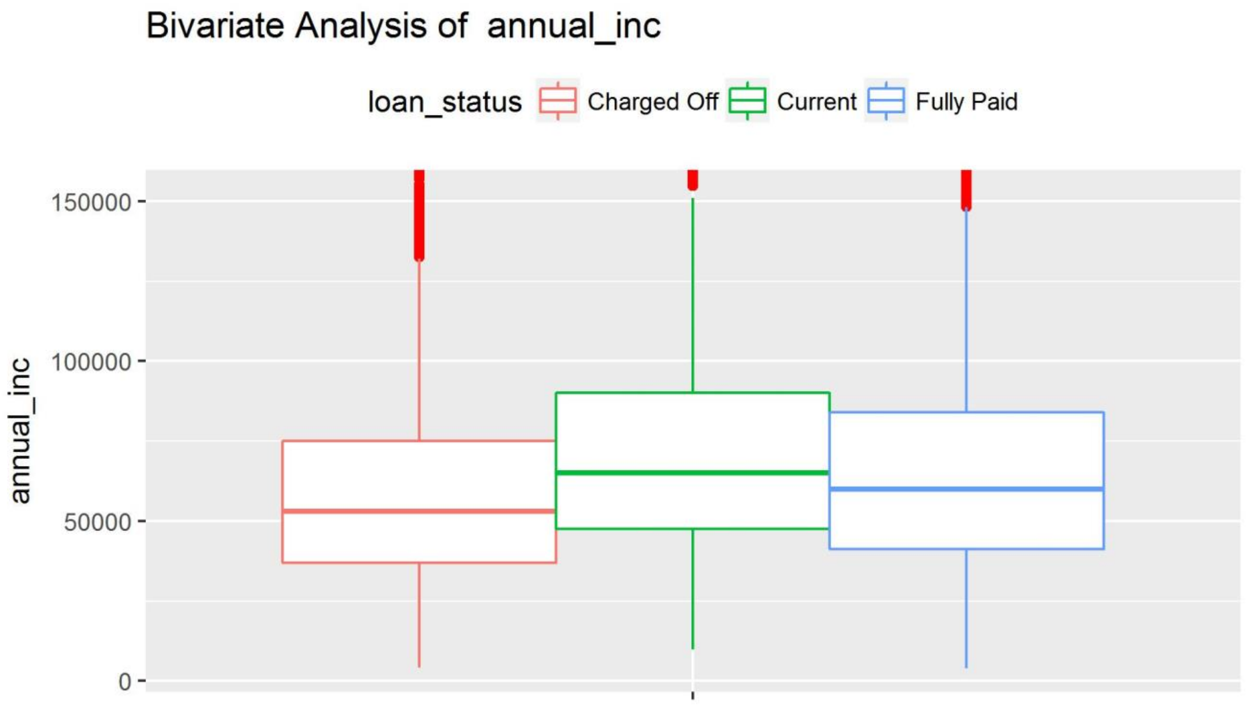
Bivariate analysis of variables with loan_status

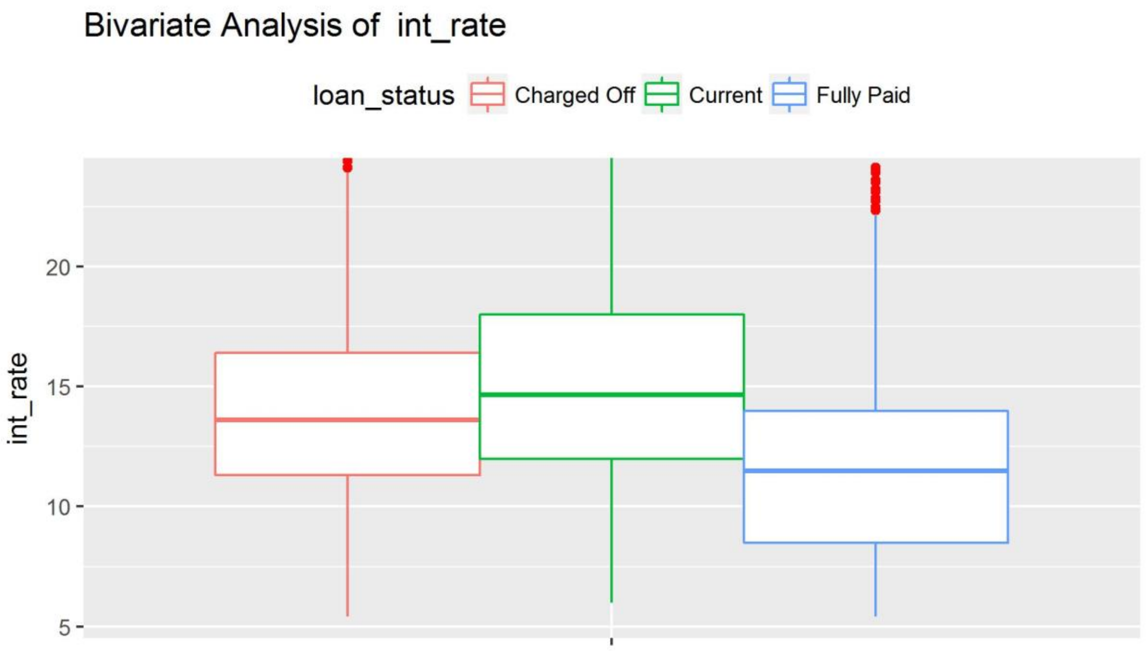




Inference:
Median for Revolving Utilization Rate is higher for Charged Off loans
Median for DTI is slightly higher for Charged off loans

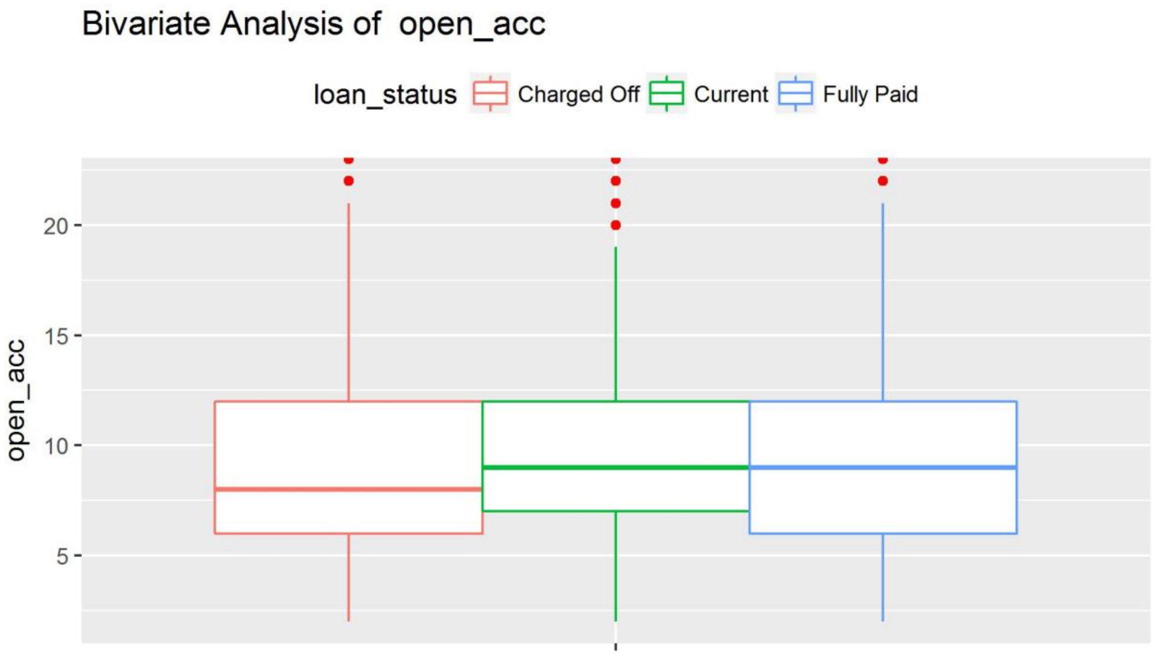
Bivariate analysis of variables with loan_status





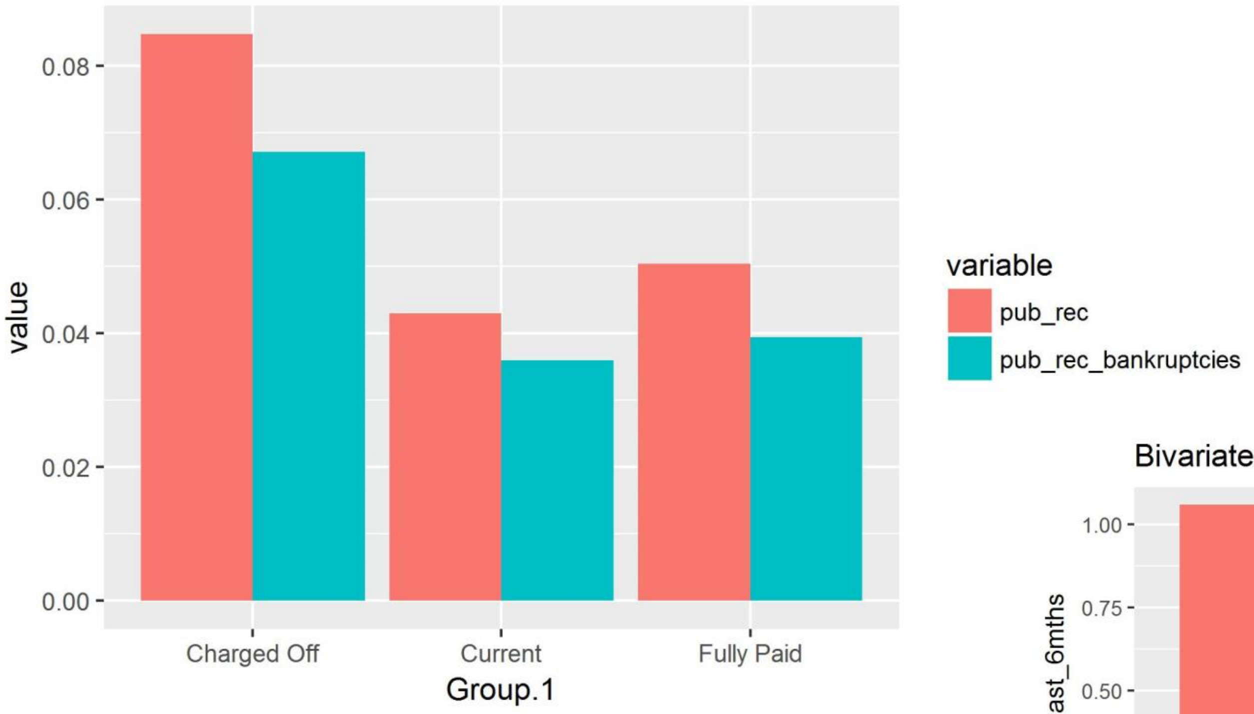
Inference:
Median for Average Annual Income is lower for Charged Off loans
Median for Interest Rate is higher for Charged off loans

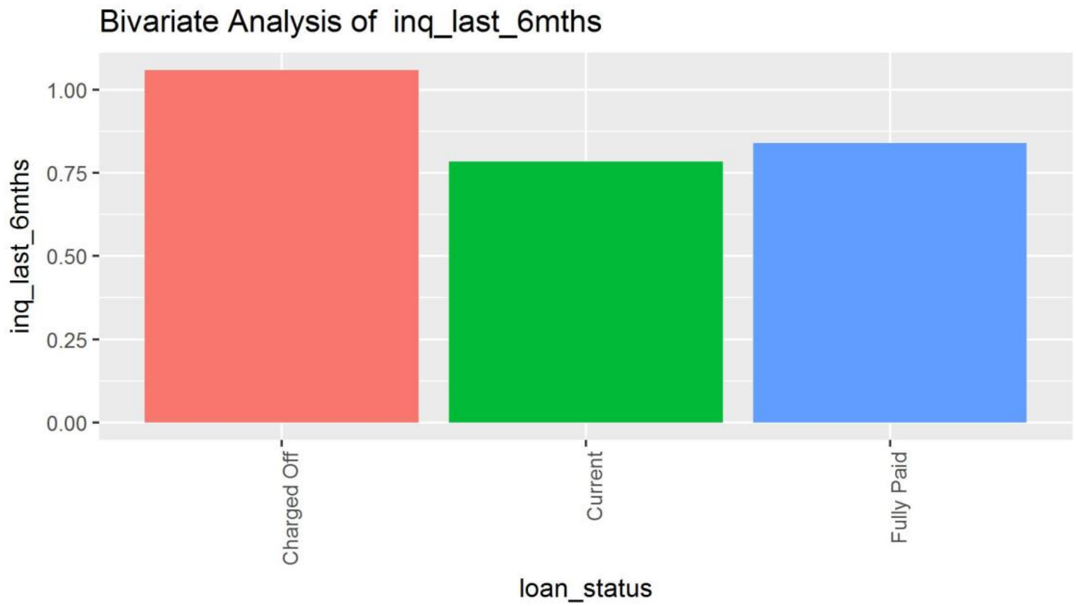
Bivariate analysis of variables with loan_status



Inference:
Median for Average Number of Open Account is less for Charged Off Loans

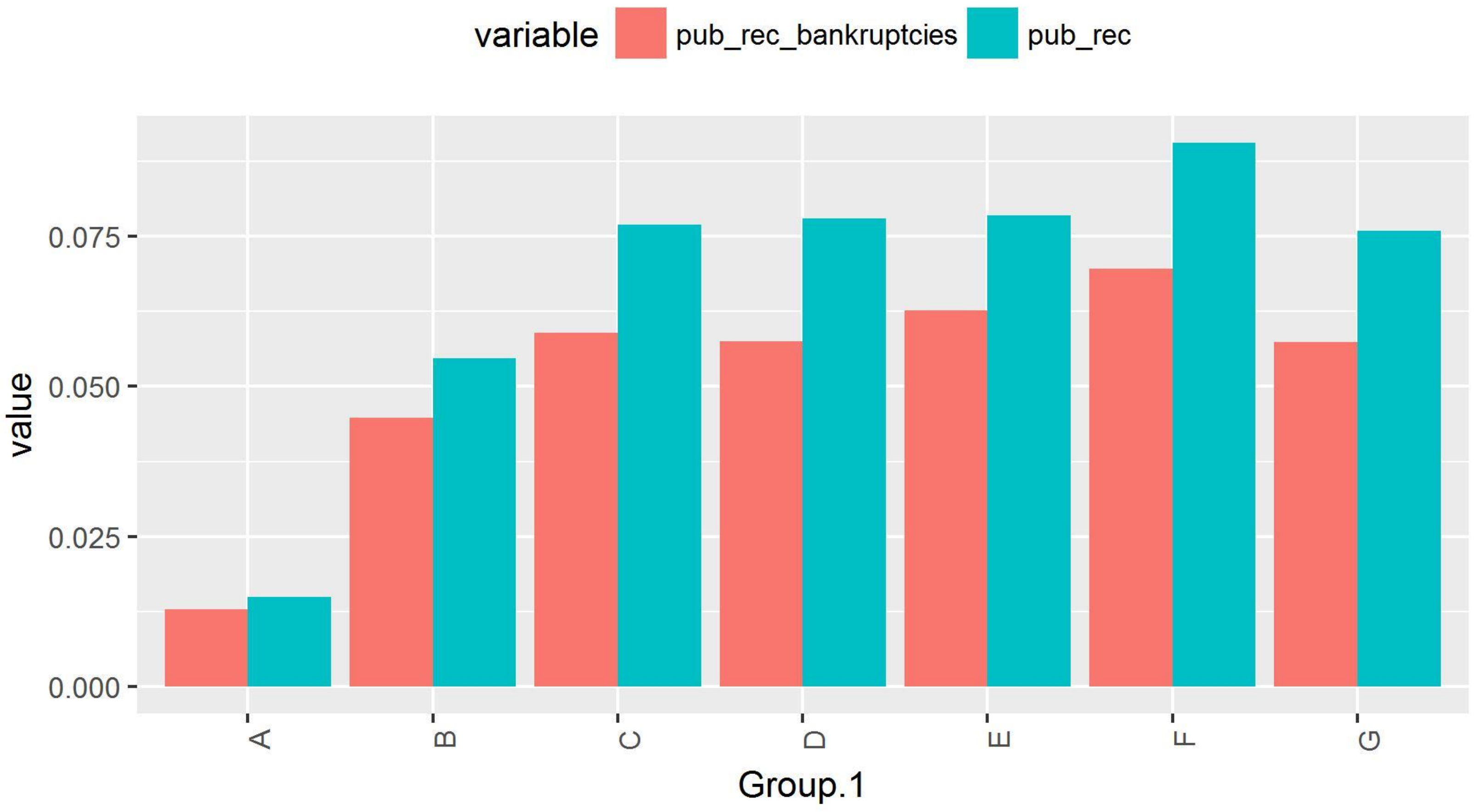
Multivariate analysis of variables with loan_status





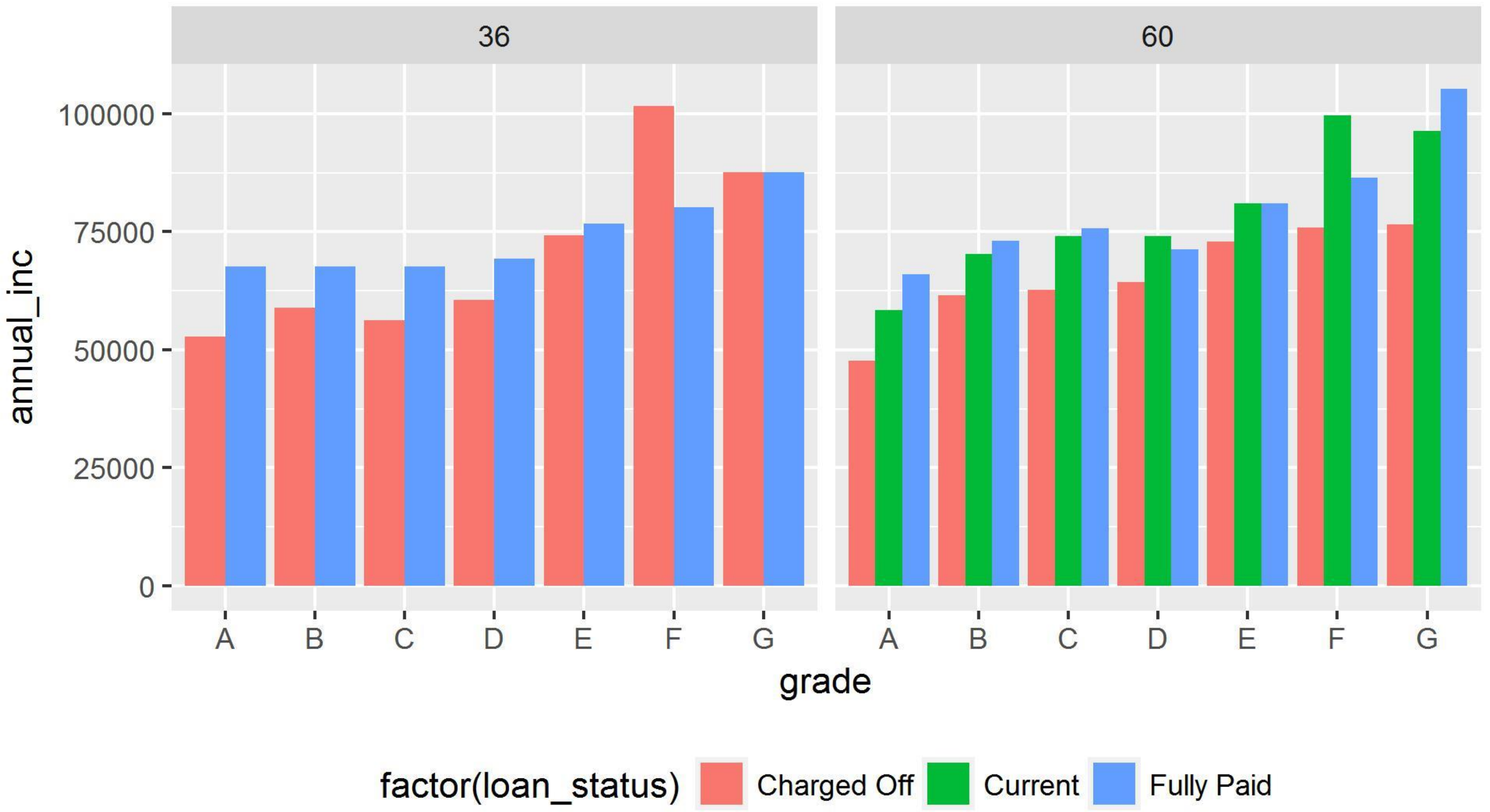
Inference:
Public Records & Number of Public Bankruptcies are Charged Off Loans
Inquiries made in last 6 months are higher for Charged Off Loans

Multivariate analysis of public records and public record bankruptcies for different grades



Inference:
Number of Public Derogatory records and Public Bankruptcies are in descending order of Grade from G to A

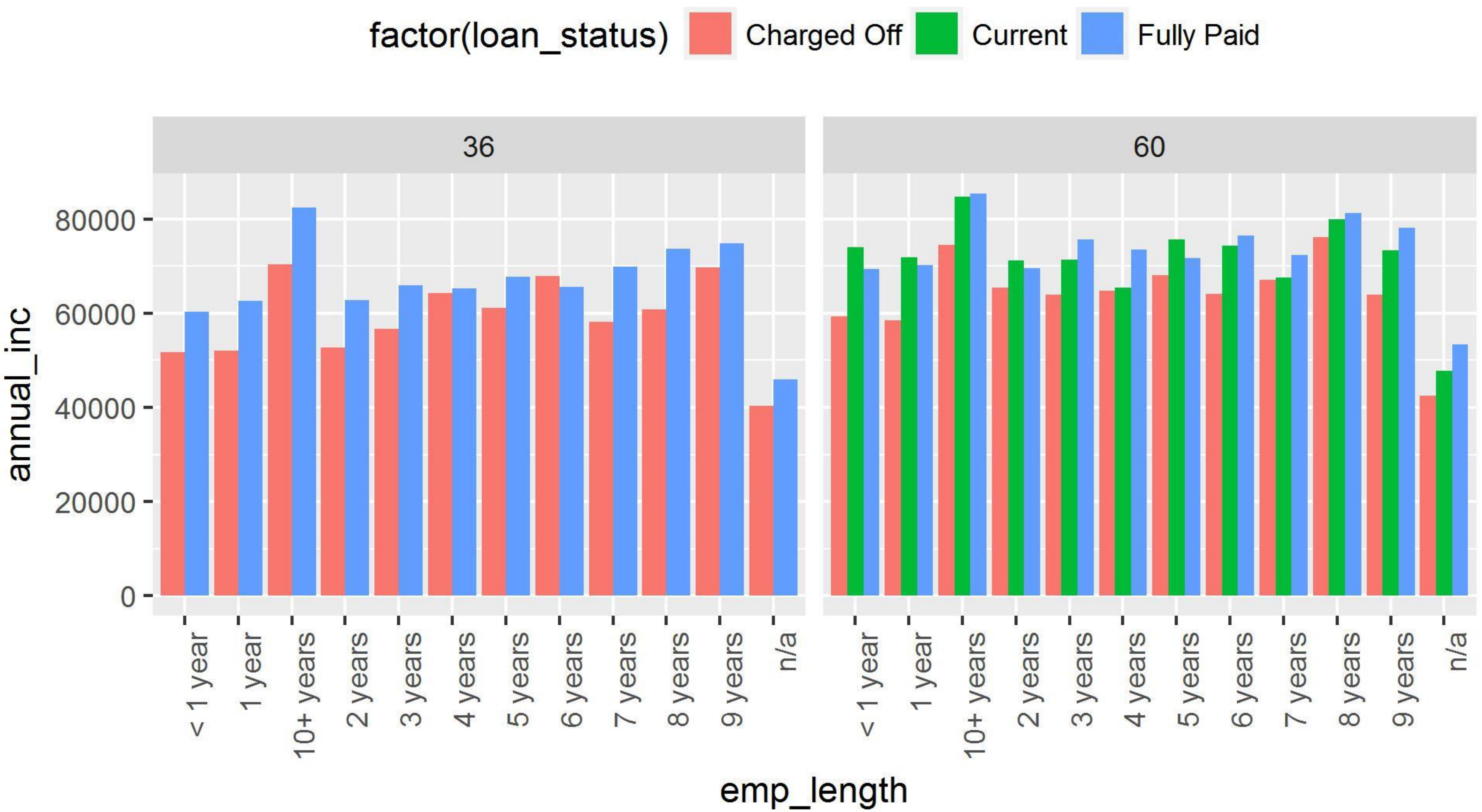
Multivariate analysis of annual income with loan_status for grades and different terms



Inference:

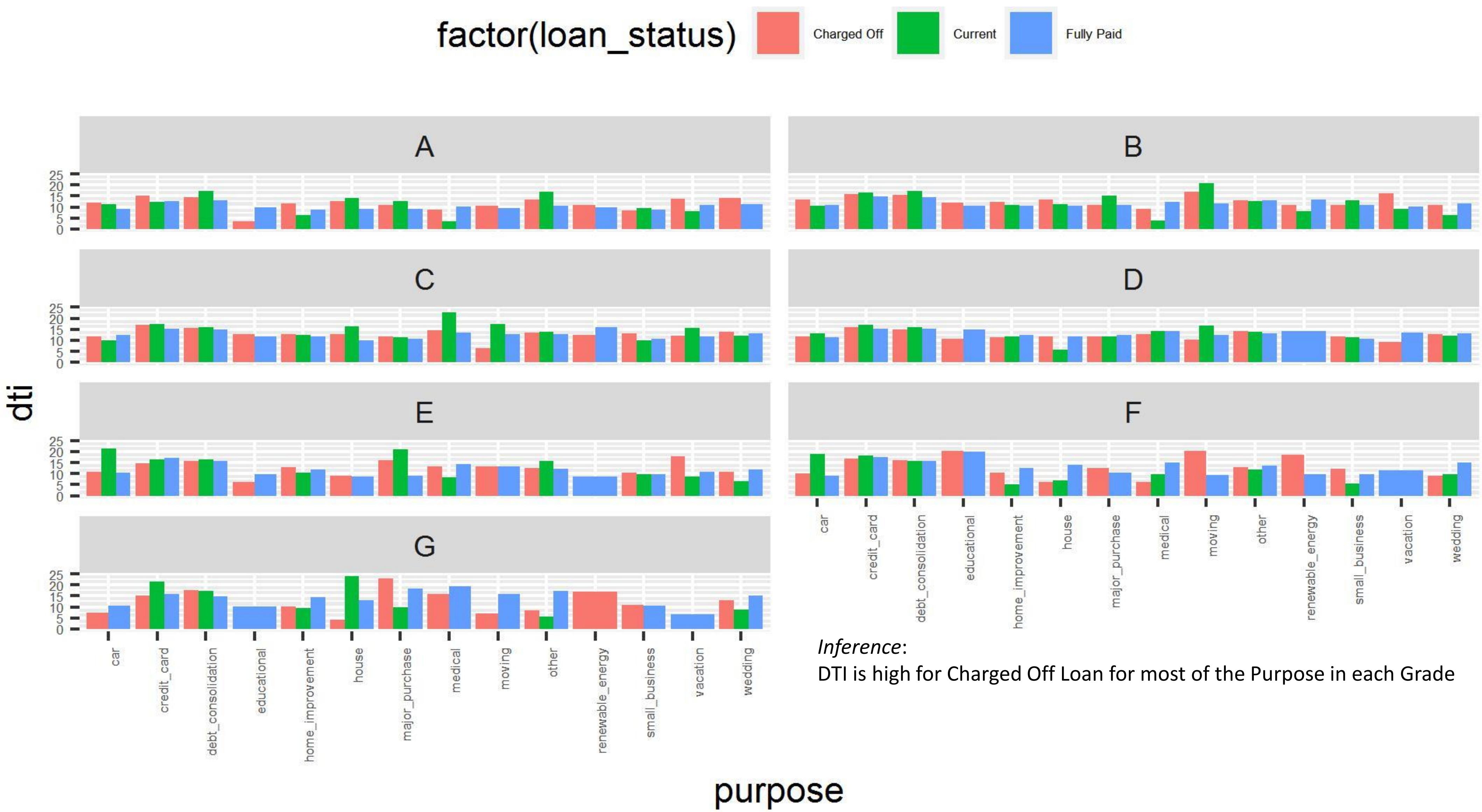
For each tenure – in each Grade - annual income is less for Charged Off loans

Multivariate analysis of annual income with loan_status for employment length and terms



Inference:
For each tenure – in each category of experience - annual income is less for Charged Off loans

Analysis of dti wrt loan_status, purpose and grade



Inference:
DTI is high for Charged Off Loan for most of the Purpose in each Grade

Conclusion

- Key Driving Factors: Number of public derogatory records and number of public rec bankruptcies are clearly an indicator of the Charged off loans.
- They are in the descending order from G to A>
- Next is the annual income. As seen from various plots, the annual income is lower for charged off loans and could be an indicator to defaulting
- Next is the DTI. The Debt to income ratio is higher for charged off loans compared to the fully paid loans and also can be indicator defaulting
- Next is the Purpose of loan. Maximum number of loans are provided for debt consolidation and maximum number of charged off loans are also there for the Debt consolidation. So purpose of the loans should also be considered for risky applicants