EDA Case Study Assignment

Gramener Case study Explanation

About Case Study:-

1.User can take lone through the such as interest rate, monthly instalment, tenure etc.

2.Some popular products are credit card loans, debt consolidation loans, house loans, car loans etc.

Business Objective:-

- To identify key variables that strongly indicate the likelihood of default, and to potentially utilize these insights in the loan approval or rejection decision-making process.

Data Understanding:-

Types of variables :-

- Customer (applicant) demographic
- Loan related information & characteristics
- Customer behavior (if the loan is granted)

Customer's Demographics	Loan related information & characteristics	Customer behavior (if the loan is granted)
Employment Length	Loan Amount	Earliest Credit Line
Employment title	Funded Amount	Revolving Balance
Annual income	Funded Amount Investment	Recoveries
Zip Code	Interest Rate	Application type
Description	Lone Status	Lone Purpose

Decision Matrix

Loan Accepted - Three Scenarios

- 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 installments 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).

Key Columns

Customer Demographics

- Annual Income (annual_inc) - Annual income of the customer. Generally higher the income, more chances of loan pass
- Home Ownership (home_ownership) - Whether the customer owns a home or stays rented. Owning a home adds a collateral which increases the chances of loan pass.
- Employment Length (emp_length) - Employment tenure of a customer (this is overall tenure). Higher the tenure, more financial stability, thus higher chances of loan pass
- Debt to Income (dti) - The percentage of the salary which goes towards paying loan. Lower DTI, higher the chances of a loan pass.
- State (addr_state) - Location of the customer. Can be used to create a generic demographic analysis. There could be higher delinquency or defaulters demographically.

Excluded Columns

- Customer Behaviour Columns - Columns which describes customer behavior will not contribute to the analysis. These attributes are not significant for consideration towards the loan approval/rejection process.
- Granular Data - Columns which describe details which are very specific to company, may not be required for the analysis. For example, “grade” column may be relevant for creating business outcomes but “sub grade” is be very granular and will not be used in the analysis.
- **54 columns** contain **NA** values only, and these columns will be removed namely: acc_open_past_24mths, all_util, annual_inc_joint, avg_cur_bal, bc_open_to_buy, bc_util, dti_joint, il_util, 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, 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, verification_status_joint
- Other irrelevant columns such as id, member_id, emp_title, desc, title, url which does not contribute to the analysis will be dropped as well.

Challenges deep-dive

Challenge 1

Find relevant attributes

To find columns which are really significant for analysis and contribute to discover patterns.

Challenge 2

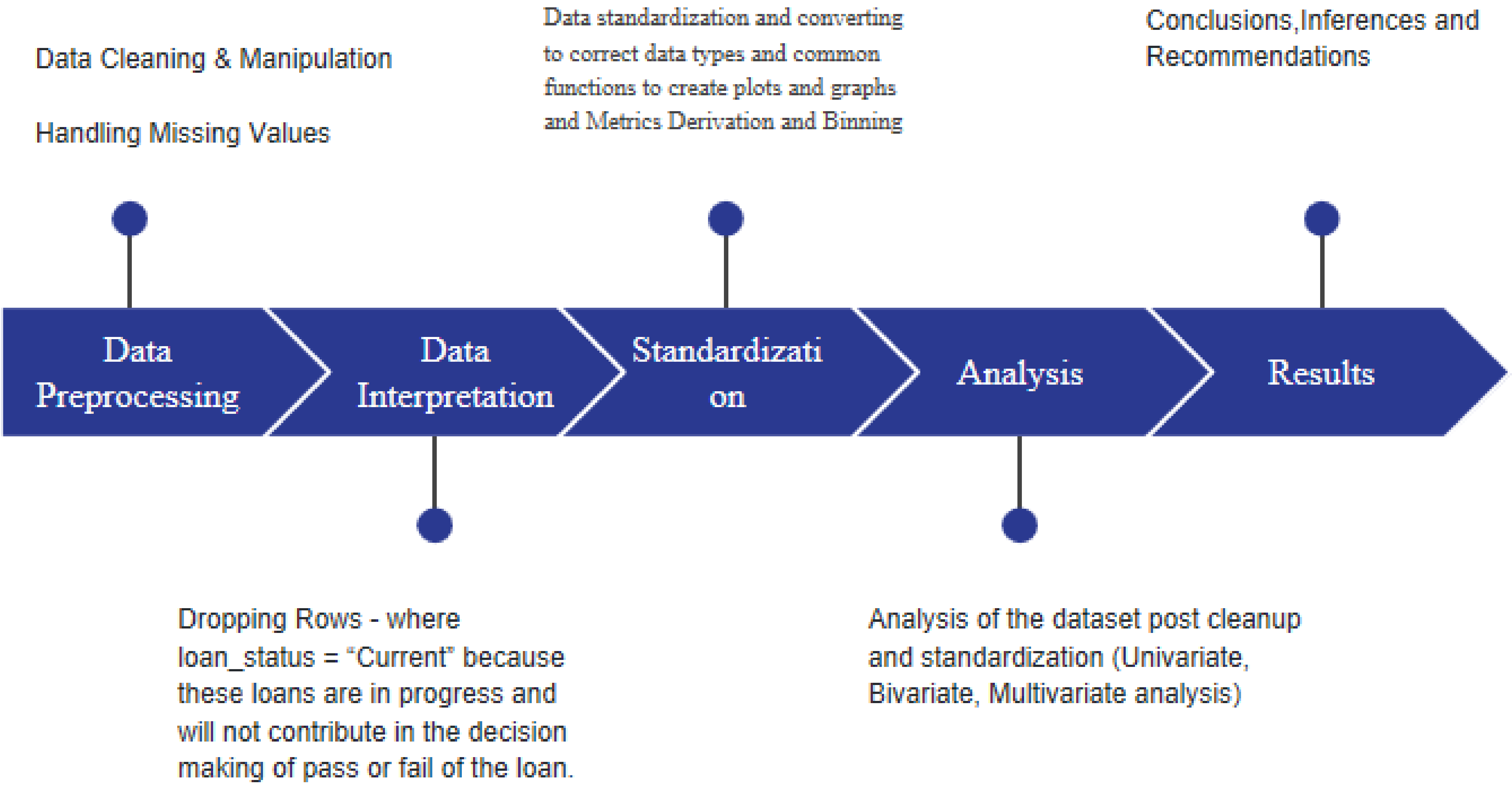
Data Correction

Some of the columns has missing values and data is not in the correct format so data cleansing and standardization is needed. Neatly segregation of categorical and numerical columns.

Challenge 3

Discover Relations in attributes

Once data will be in correct format then on remaining columns we need to do analysis and find patterns which affects `loan_status` variable with respect to other variables by using EDA (Univariate, Bivariate, Multivariate) Analysis.



Data Preprocessing

Data Cleaning & Manipulation

Handling Missing Values

Data Interpretation

Dropping Rows - where loan_status = "Current" because these loans are in progress and will not contribute in the decision making of pass or fail of the loan.

Standardizati on

Data standardization and converting to correct data types and common functions to create plots and graphs and Metrics Derivation and Binning

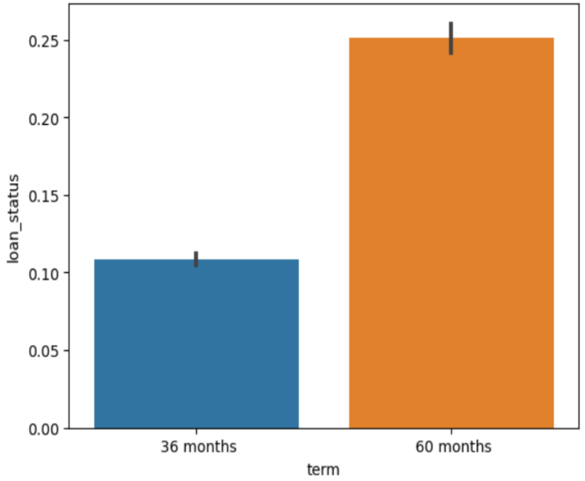
Analysis

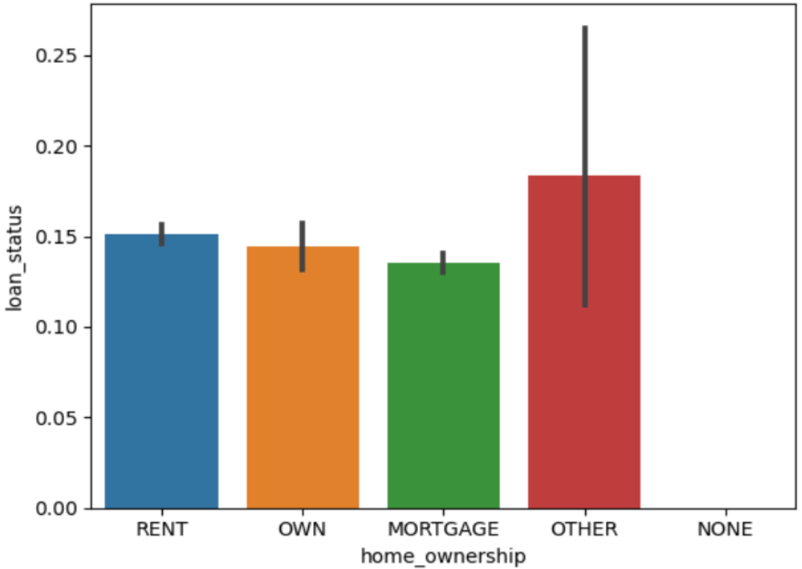
Analysis of the dataset post cleanup and standardization (Univariate, Bivariate, Multivariate analysis)

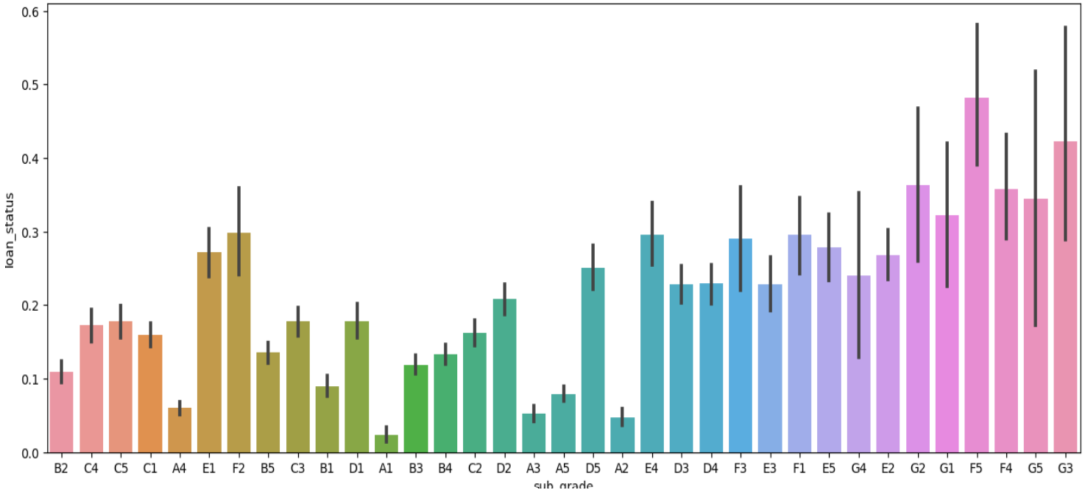
Results

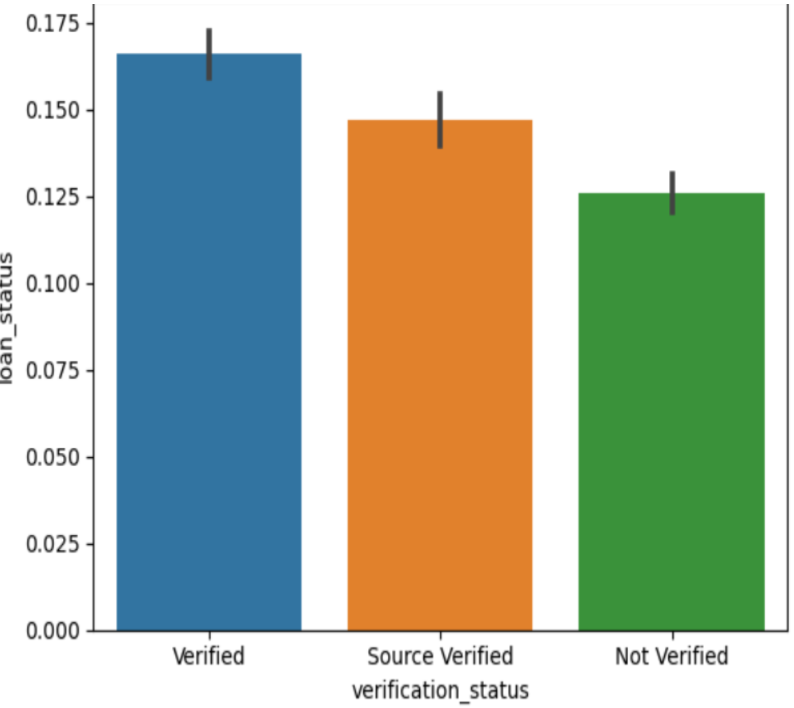
Conclusions,Inferences and Recommendations

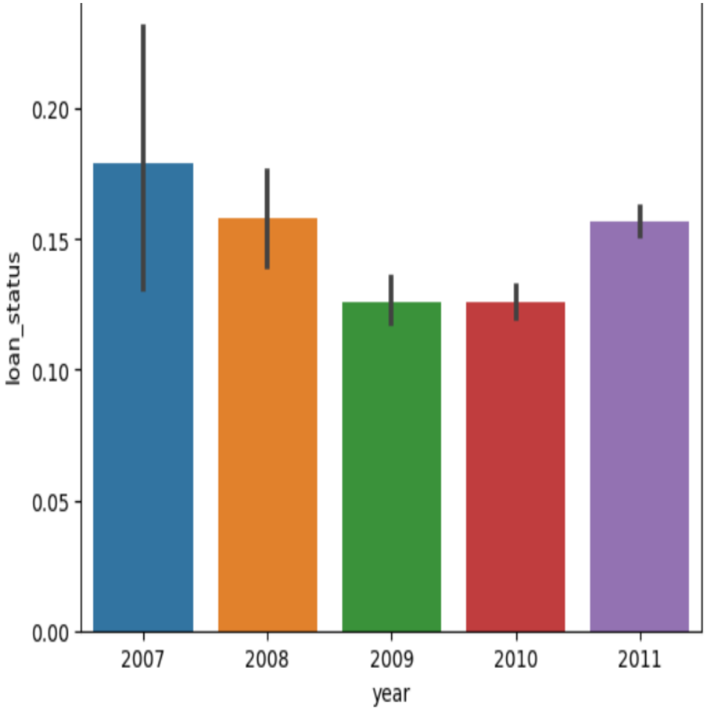
Univariate Analysis

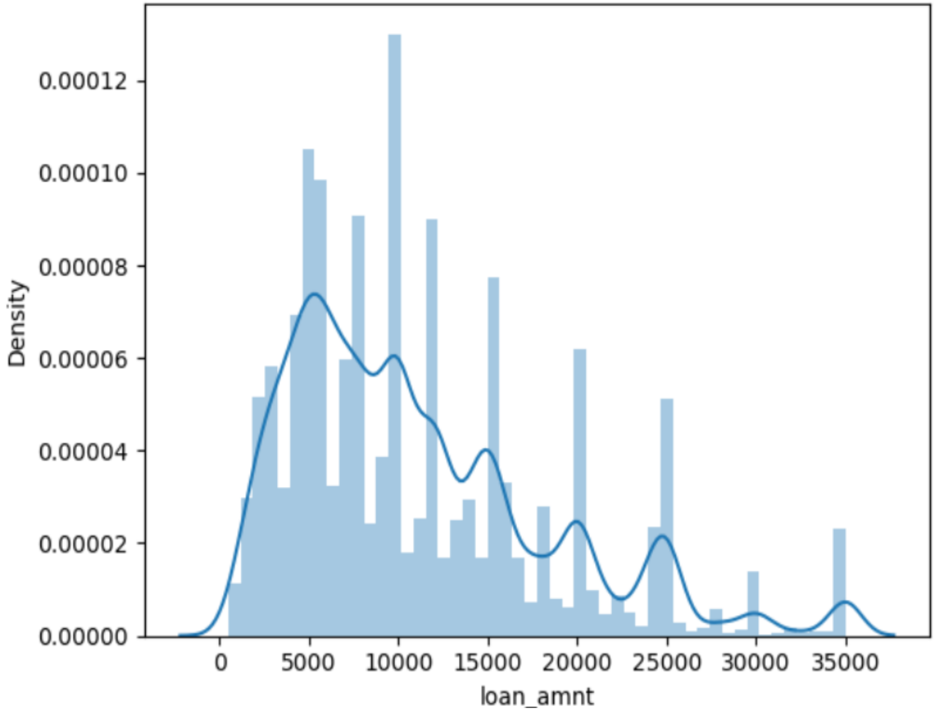








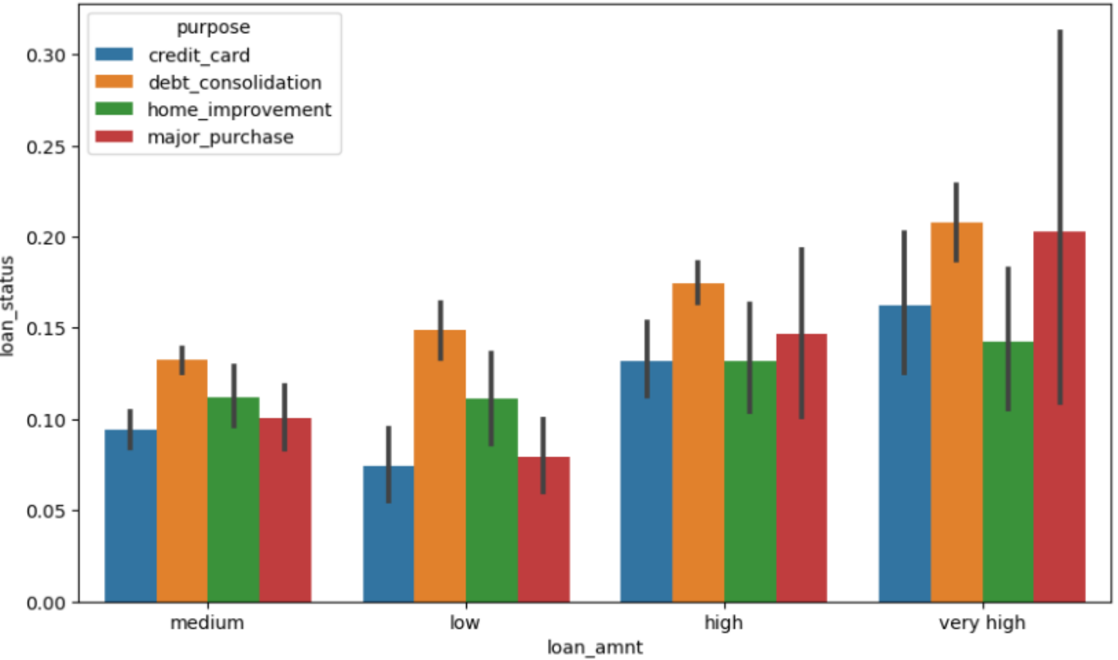


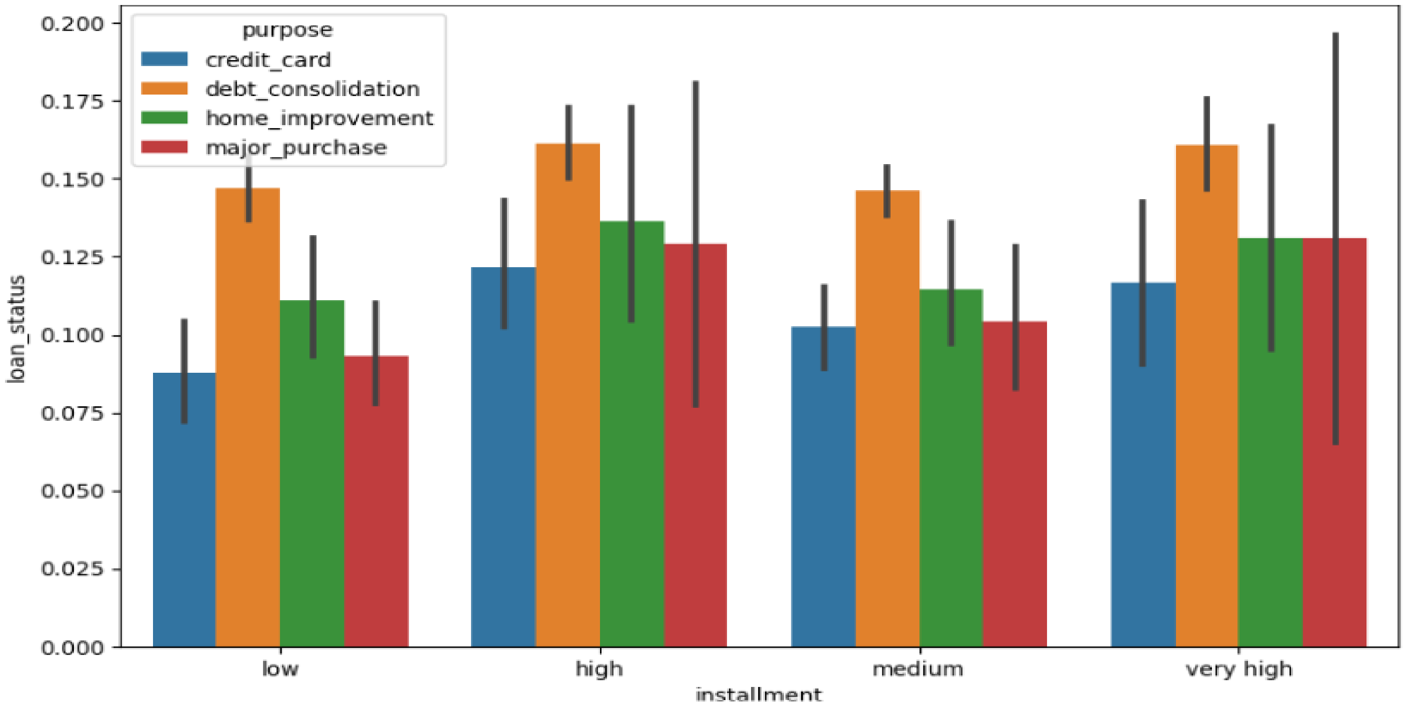


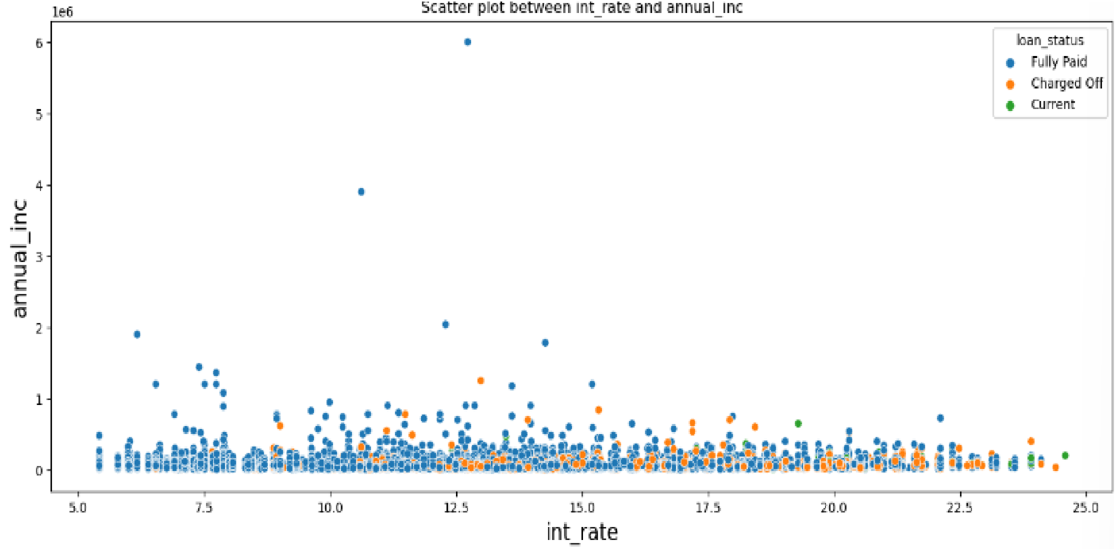
Data Cleaning & Pre-processing

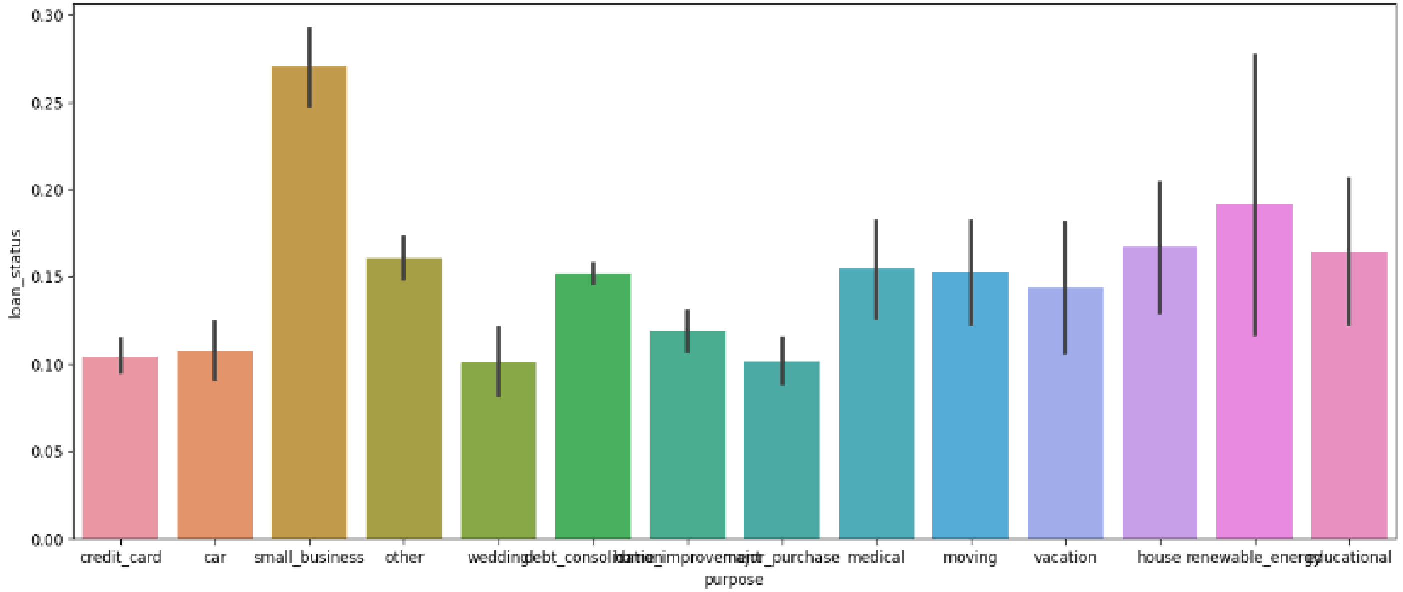
- While loading the dataset, some of the variables had mixed data types so they have to be converted accordingly for further analysis.
- There are many columns with null values. So they had to be dropped as they won't play a role in the analysis of the dataset.
- Some columns has only a single unique value so it does not make any sense to include it as part of our data analysis. 9 columns had such unique values such as ['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'] and they were removed.
- Dropped records where loan_status="Current" as the loan in progress cannot provide us insights as to whether the borrower is likely to default or not.
- Common functions were created for repeating common operations like plotting bar graphs, box plots, histograms, count plots, binning etc.
- Converted columns like debt to income (dti), funded amount (funded_amnt), funded amount investor (funded_amnt_inv), interest rate (int_rate) and loan amount (loan_amnt) to float to match the data. Also converted loan date (issue_d) to DateTime (format: yyyy-mm-dd).

Bivariate Analysis









Conclusion

- The **grade** of loan goes from A to G, the **default rate increases**. This is expected because the grade is decided by Lending Club based on the riskiness of the loan.
- 2.) **Purpose: small business loans** default the most, then **renewable energy** and **education**.
- 3.) **Term: 60 months** loans default more than **36 months loans**.
- 4.) **Verification status**: surprisingly, **verified loans** default more than **not verified**.
- 5.) The easiest way to analyse how default rates vary across **continuous variables** is to bin the variables into **discrete categories**.
- 6.) **Higher the loan amount, higher the default rate.**
- 7.) **High interest rates default more**, as expected.
- 8.) **High dti translates into higher default rates**, as expected.
- 9.) **The higher the installment amount, the higher the default rate.**
- 10.) **Lower the annual income, higher the default rate**.
- 11.) Most loans are debt consolidation (to repay other debts), then credit card, major purchase etc. **Debt consolidation loans have the highest default rates**

Thank you

Jagriti kumari