# LENDING CLUB CASE STUDY

MANIKESH BOYINI

C-67, JULY 2024 BATCH

# OVERVIEW

- Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers.
- When the company receives a loan application, the company has to make a decision 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

# OBJECTIVE

- As a Data scientist for Lending Club, My goal is to analyze the data obtained from previous clients to help the company to make informed data driven decisions. This will help company operate efficiently and improve profitability.
- The goal of this case study is to identify these risky loan applicants.
- If one is able to ***identify these risky loan applicants***, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the one of the objective of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

# UNDERSTANDING DATA

- **Each row :** Represents a person's loan application data with the bank: Lending Club, with Each row is all the details.

| Description | Explanation | Description | Explanation |
|---|---|---|---|
| Number of rows | 39717 | Number of columns with all NAN values | 55 |
| Number of columns | 111 | Columns with one value | 6 |
| Time frame of data | June 2007 to Dec 2011 | Key columns | "loan_amnt", "funded_amnt", "funded_amnt_inv", "pub_rec_bankruptcies", "annual_inc", "term", "int_rate", "emp_length", "issue_year", |
| Data types of columns | Float type : 74 columns<br>Object type: 24 columns<br>Int type: 13 columns | | |

# CONTINUATION

| Categorical Columns | Date Columns | Numerical Columns |
| --- | --- | --- |
| Term, Grade, sub_grade<br>emp_title, emp_length<br>home_ownership,<br>verification_status<br>loan_status<br>pymnt_plan - N<br>url<br>purpose<br>title<br>addr_state<br>initial_list_status - f<br>application_type - INDIVIDUAL | issue_d<br>earliest_cr_line<br>last_pymnt_d<br>next_pymnt_d<br>last_credit_pull_d | loan_amnt, funded_amnt,<br>funded_amnt_inv, int_rate<br>Installment, annual_inc, dti<br>delinq_2yrs, inq_last_6mths<br>mths_since_last_delinq<br>mths_since_last_record<br>open_acc , pub_rec, revol_bal,<br>revol_util, total_acc<br>out_prncp, out_prncp_inv<br>total_pymnt, total_pymnt_inv ,<br>total_rec_prncp , total_rec_int,<br>total_rec_late_fee, Recoveries,<br>collection_recovery_fee ,<br>last_pymnt_amnt |

# DATA CLEANING

- All the columns with NA values should be dropped as this will not help in analysis.
- There are 55 columns with all NA values. Columns that should be dropped are (next_pymnt_d, 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)
- All the columns with zero values should be dropped.
- The columns where more than 65% of data is empty (mths_since_last_delinq, mths_since_last_record) should be dropped.
- Drop columns (emp_title, desc, title) as they are discriptive and text (nouns) and dont contribute to analysis.
- Drop customer behaviour columns which show data post the approval of loan. They contribute to the behaviour of the customer. Behaviour of the customer is recorded post approval of loan and not available at the time of loan approval. Thus these variables will not be considered in analysis and thus dropped`(delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type)`

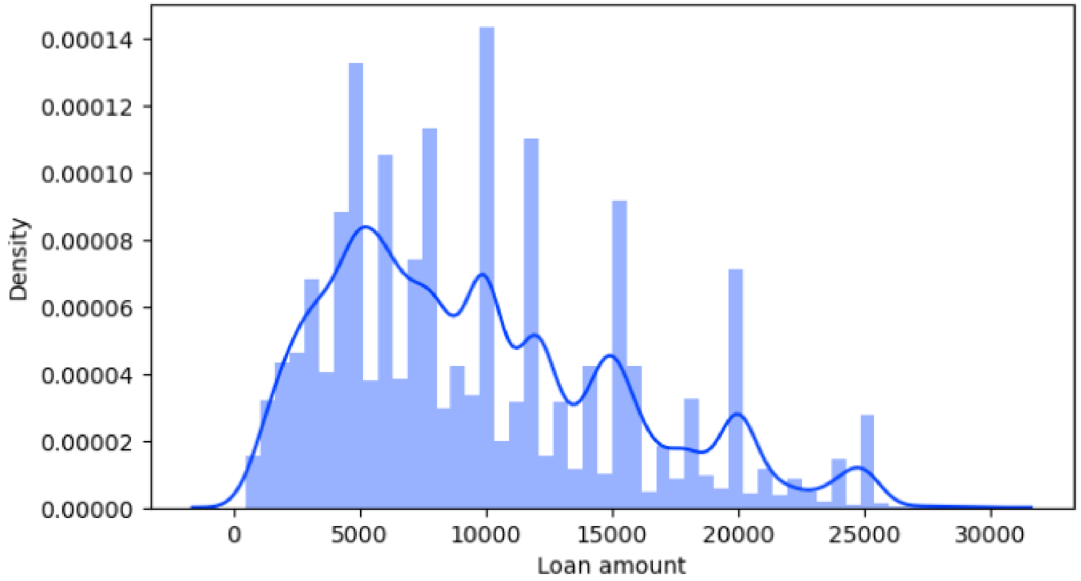
# CONVERT FORMATS AND STANDARDISATION

- Convert Column Format
  - . loan_amnt, funded_amnt, funded_amnt_inv columns are object type and will be converted to respective data type
  - int_rate, installment, dti columns are object type and will be converted to float
  - Take out "month" text from `term` column and convert to integer

- Standardize Values
  - .All currency values in the columns should be rounded off to 2 decimal places.

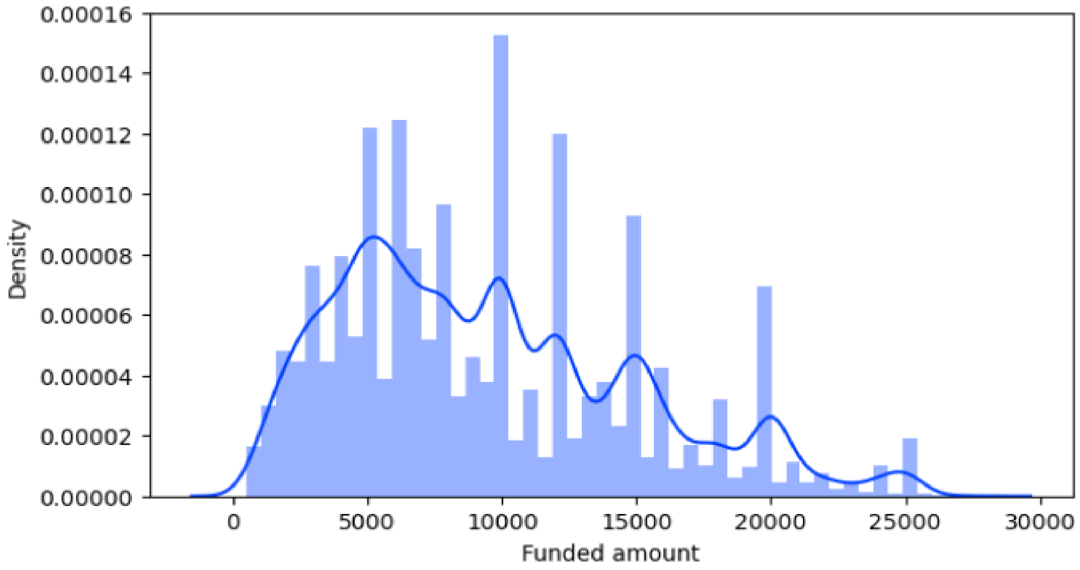
# UNIVARIATE ANALYSIS

- Loan Amount




**Most of the loan amount is in the range of 5000 to 13000 range**

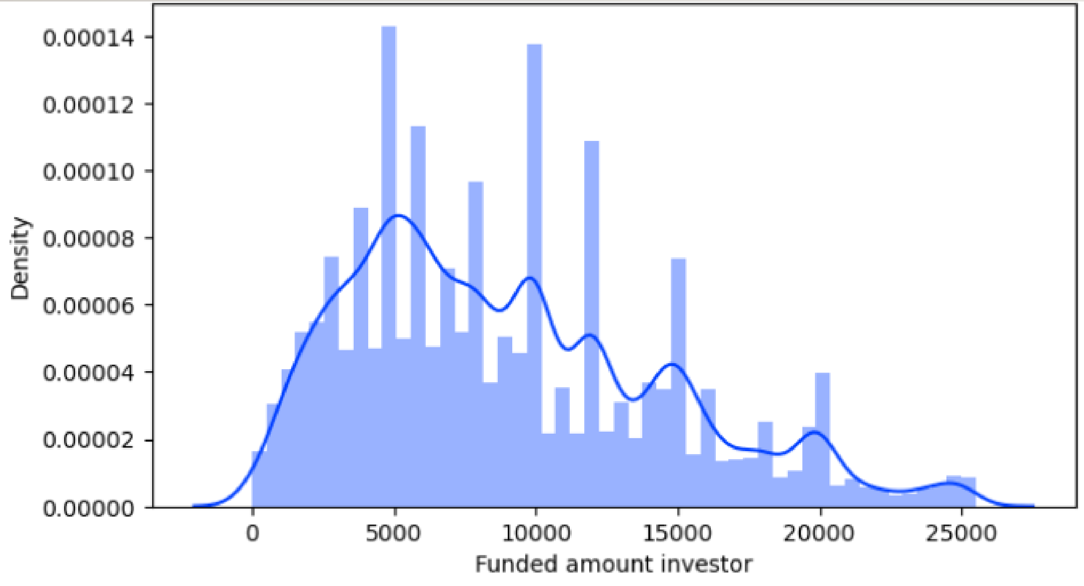
- Funded Amount




**Majority of the funded_amnt is in the range of 5000 to 13000**

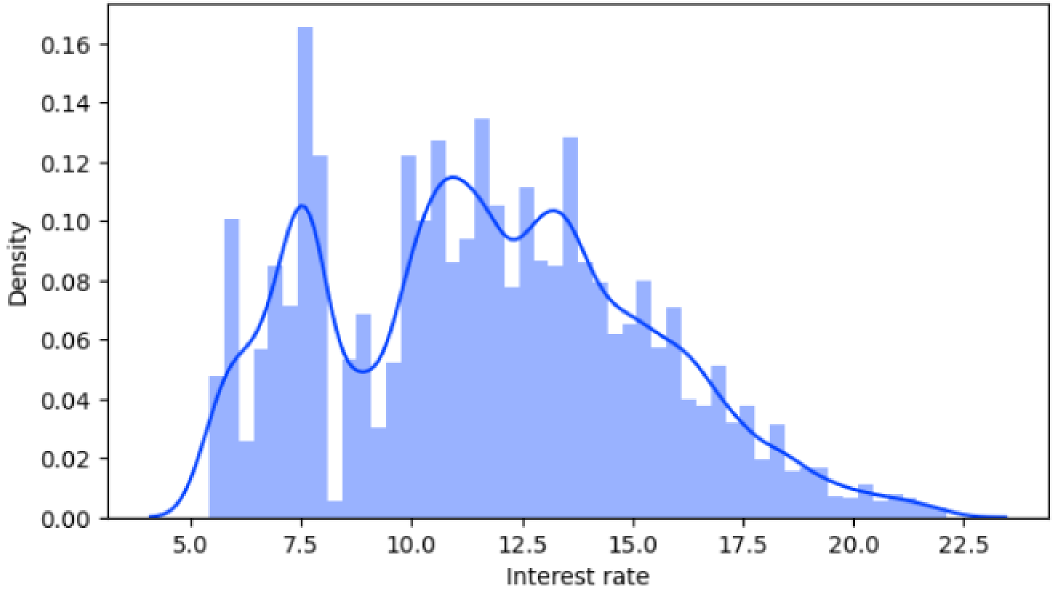
# UNIVARIATE ANALYSIS

- Fund Amount Inv




**Majority of the funded_amnt_inv is in the range of 5000 to 12000**

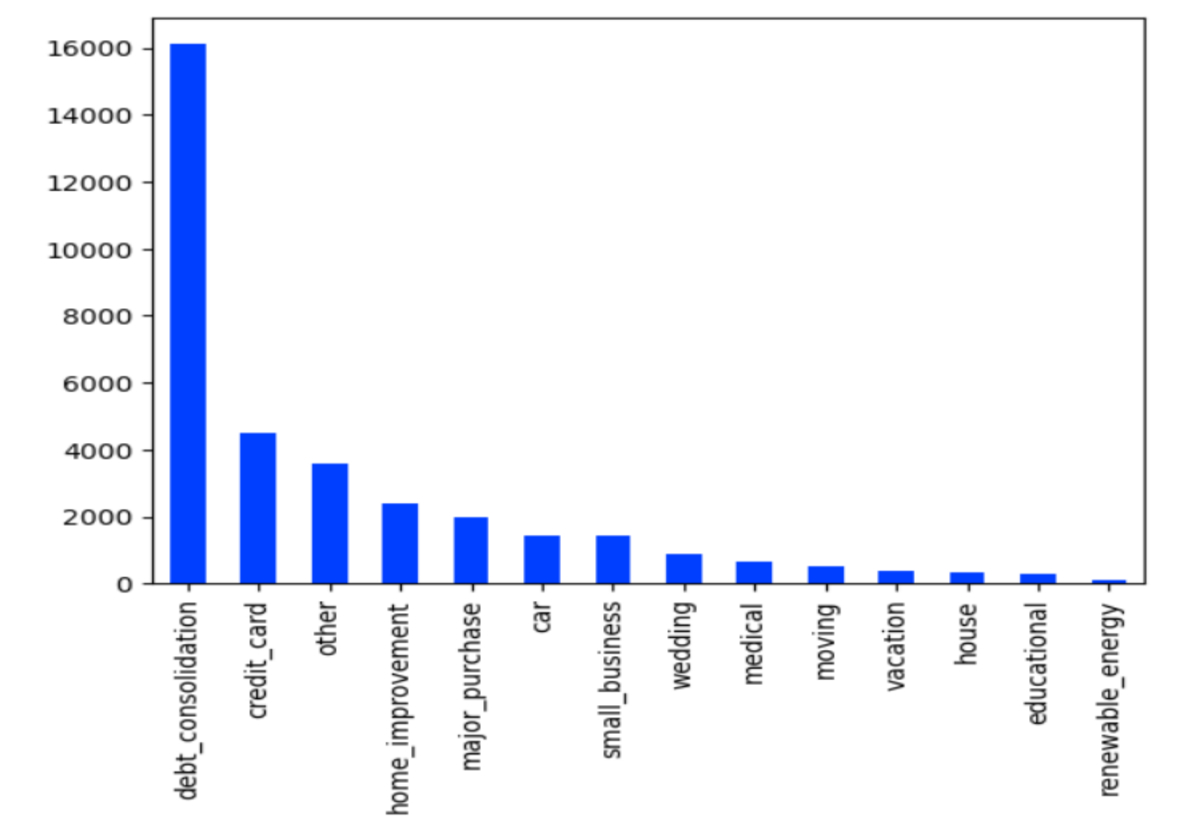
- Interest Rates



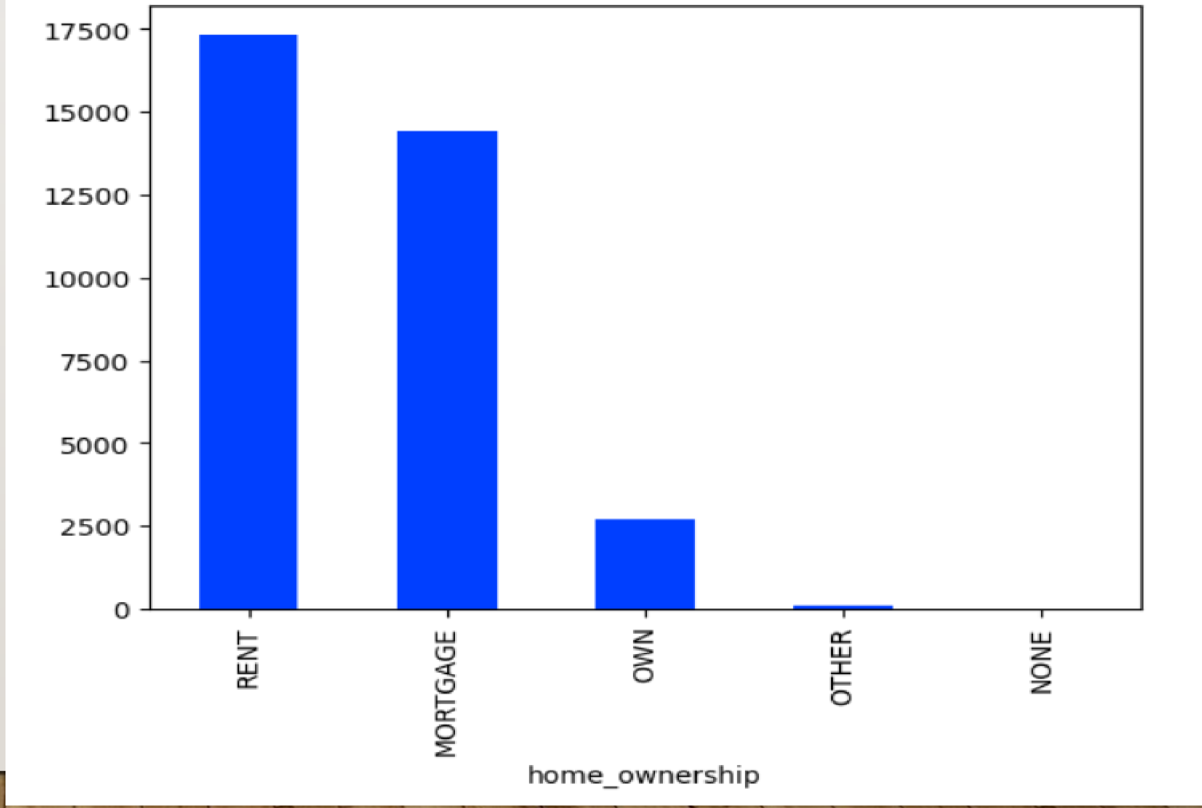

**Majority of the interest rate is in the range of 9% to 14%**

# UNIVARIATE ANALYSIS

- Loan purpose

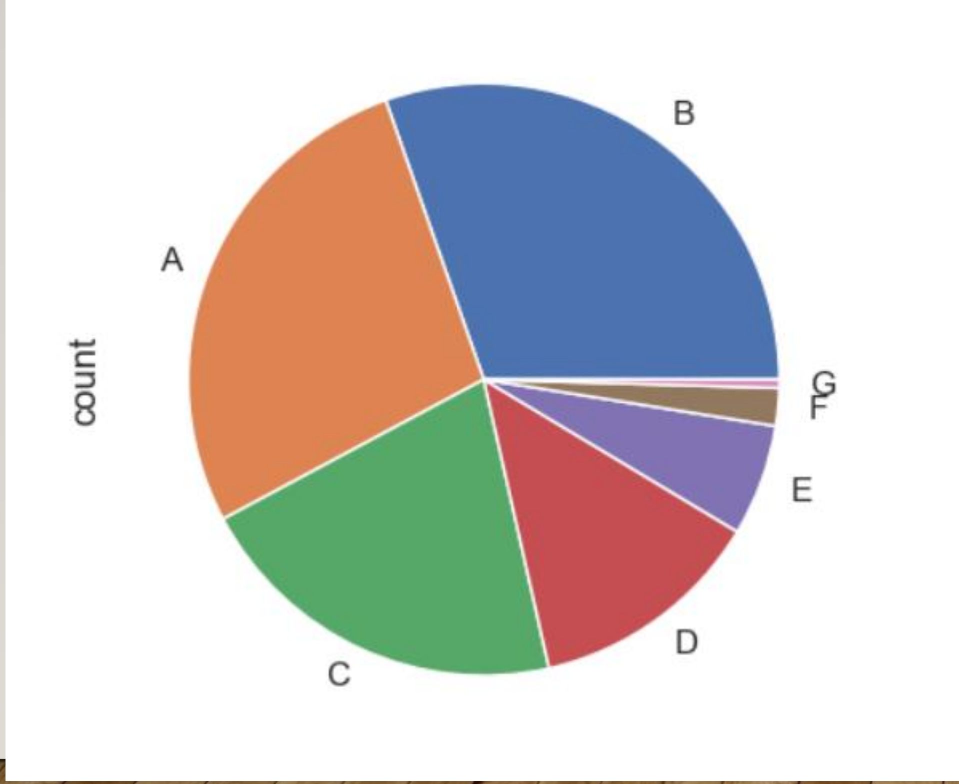



- Home ownership

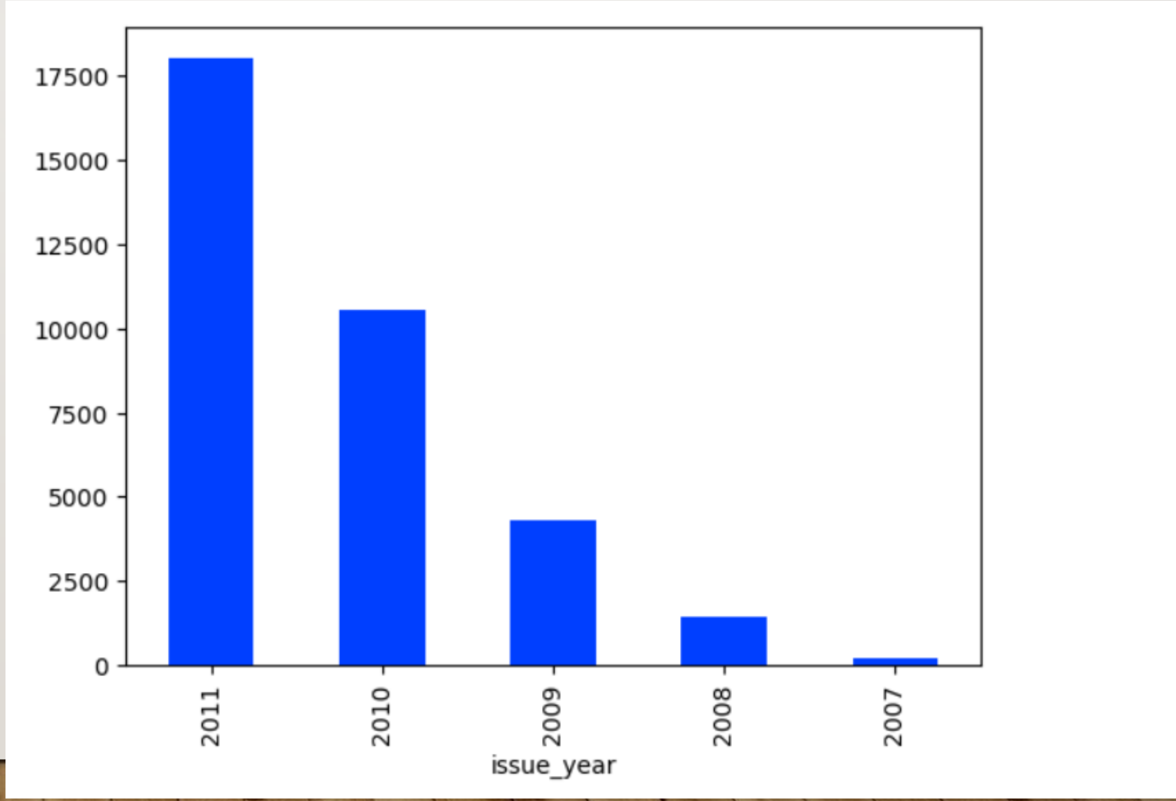


# UNIVARIATE ANALYSIS

- Grade




- Issue Year

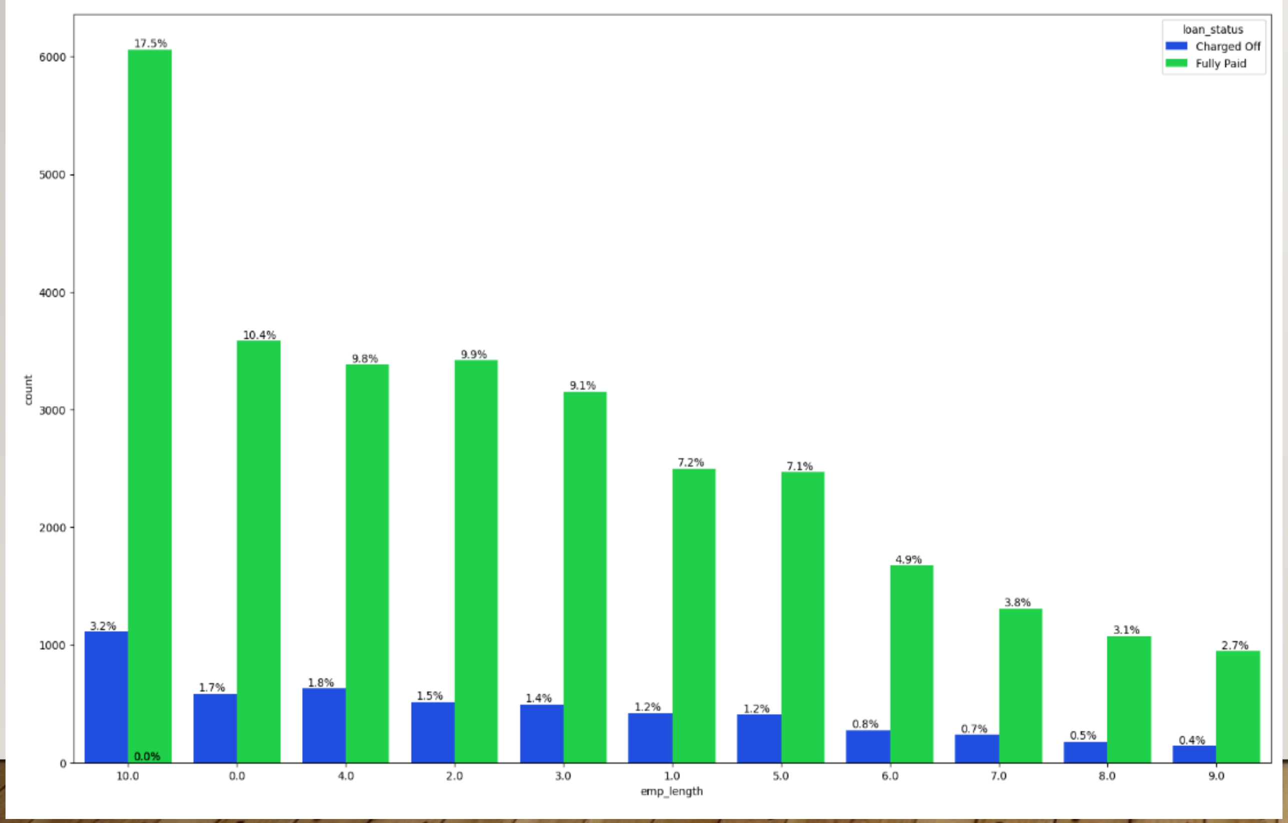


# UNIVARIATE ANALYSIS SUMMARY

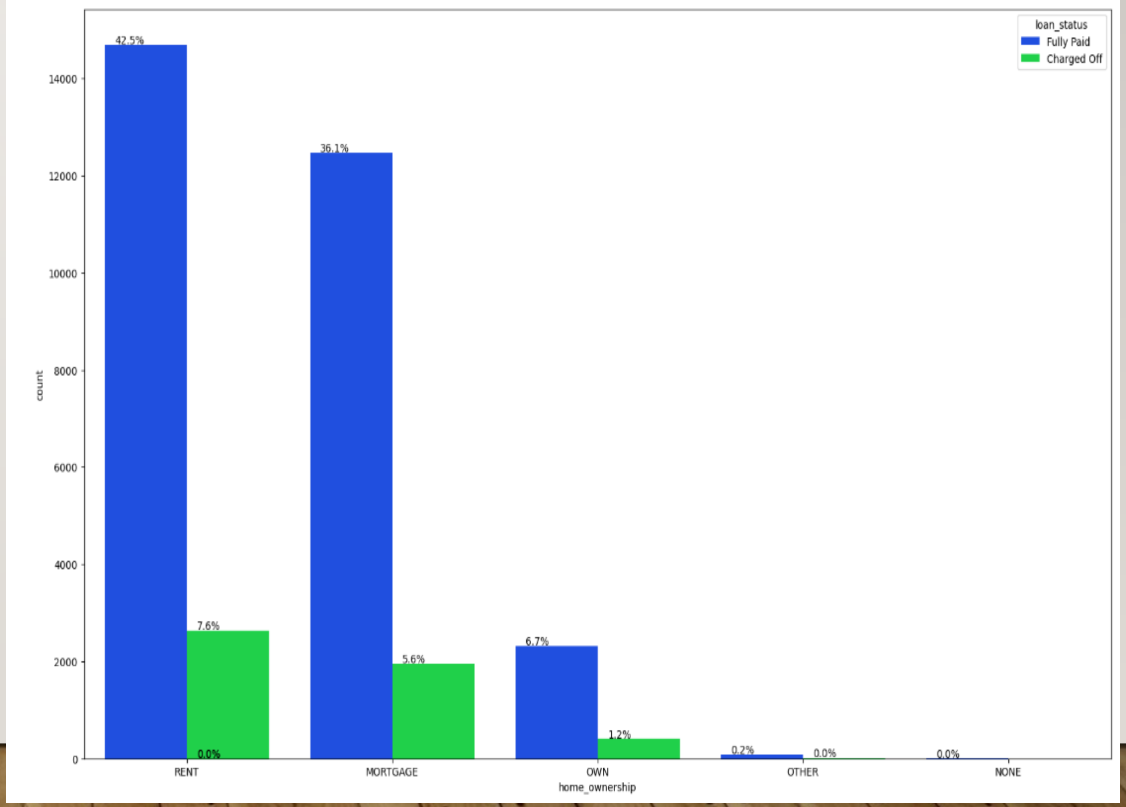
- Most customers earn 60000.0 per year
- Most customers live in a rented accommodation and morgage comes close second
- Most customers stay with one employer for more than 10 years
- Debt to earnings ratio is 0 for most of the customers
- Maximum number of customers ae from CA state.
- Most of the loan applicants have an annual income of 0-40 k
- Most of loans are taken in Q4 in a year
- Taking loans are increasing year after year
- least number of loans are taken in Q3

# BIVARIATE ANALYIS

- Employee Length vs loan status

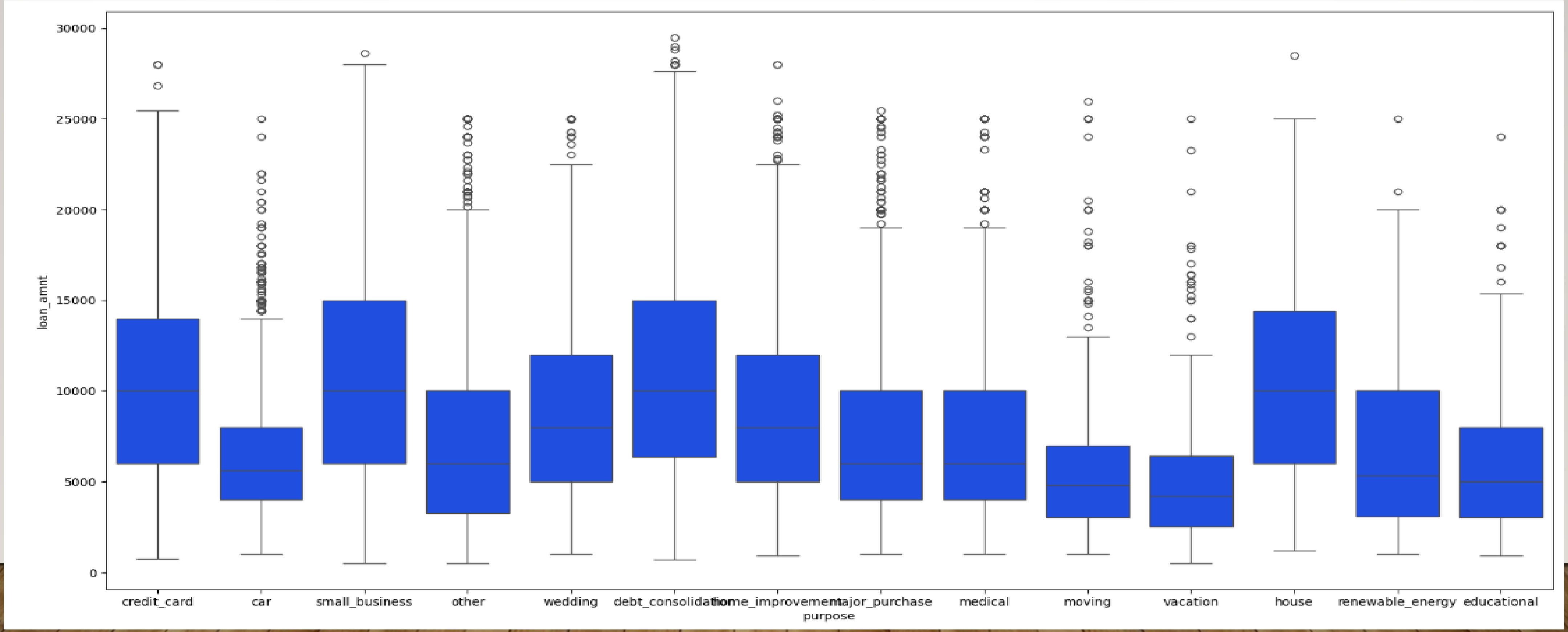



- Home ownership vs loan status



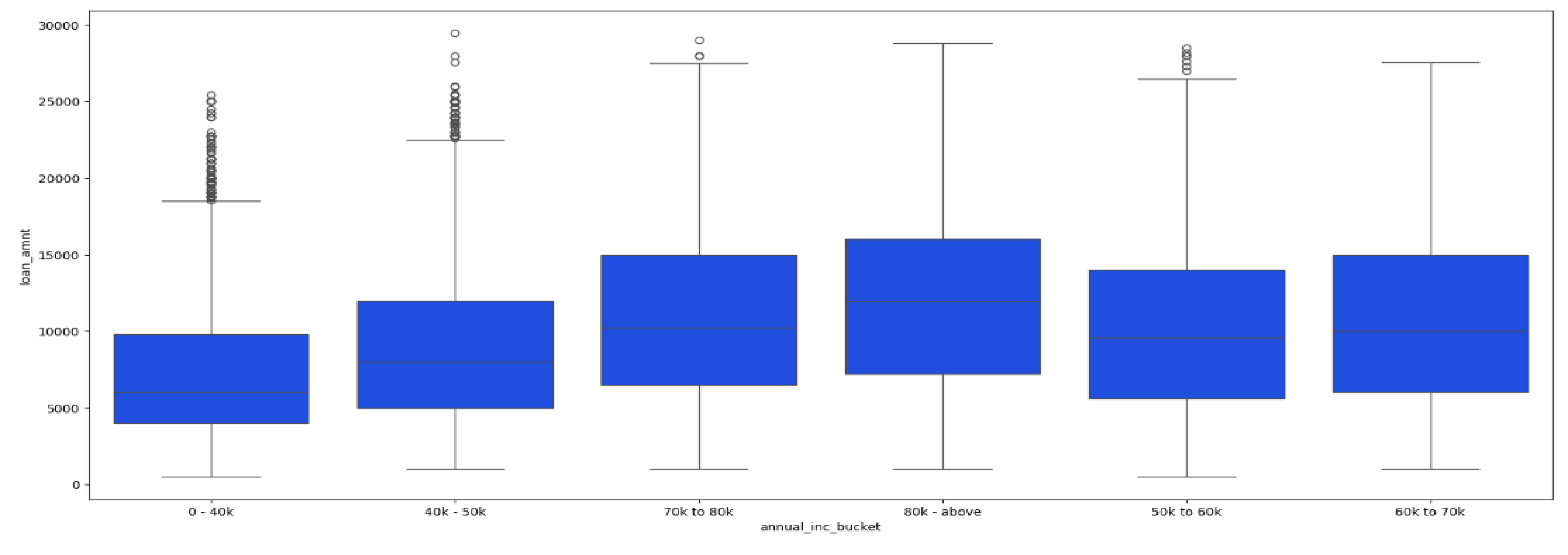
# BIVARIATE ANALYSIS

- Loan amount vs Purpose



# BIVARIATE ANALYSIS

- Loan amount vs annual income bucket



# BIVARIATE ANALYSIS SUMMARY

- Annual income between 0-40k has the highest fully_paid loan status across bucket
- Annual income between 0-40k also has the highest charged off status across buckets
- Annual income between 70-80k has the lowest fuly paid status and lowest charged off status.
- Bulk of the vacation loans tend to be lowest of all categories
- Small business and debt consolidation tend to be highest loan amount of all categories
- Overall highest Charge Off numbers are in the category of RENT and MORTGAGE
- The home_ownership status of mortgage and are at the highest risk of Charge Offs
- Mortgage status also has the highest range of loan amounts increasing the risk
- Highest Charge Offs are in the employee length of 10 Years and above
- High probablity of Charge Off's whose income range is less than 1 years

# CORELATION MATRIX (CLUSTER MAP)

Negative Correlation

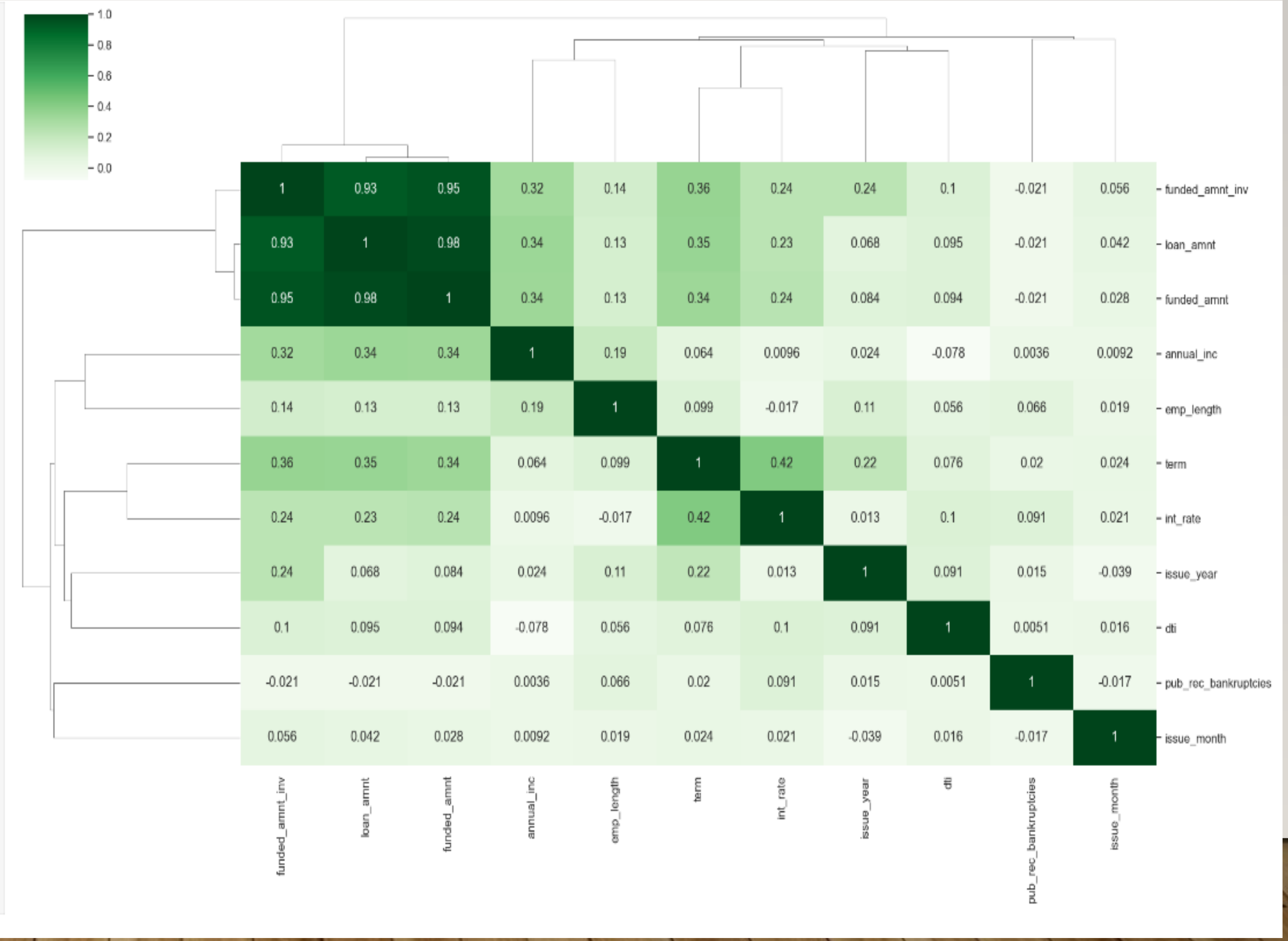
1. Loan amount, funded amount investor and funded amount has negative correlation with pub_rec_bankrupticies
2. Annual income has a negative correlation with dti

Strong Correlation

1. Term has a strong correlation with loan amount
2. Term has a strong correlation with interest rate
3. Annual income has a strong correlation with loan amount
4. Ffunded amount, funded amount inv and loan amount has strong corelation among each other

Weak Correlation

1. pub_rec_bankruptcies has weak correlation with most of the fields except employee length and interest rate
2. 2.Term shows weak corelation with dti



# CONCLUSIONS

- Lending club should check borrowers grade in order to reduce risk.
- Grade A applicants are more likely to pay full loan.
- Lending club should reduce the loans for 60 months tenure as they are prone to loan default.
- In contrast, 36 month term shows less loan default average
- Home ownership with Other are likely to default the loan and Mortgage may have more chance to fully pay the loan.
- Lower the dti ratio lower the change of loan default also, it is likely that low dti ratio borrower Likely to pay full loan.


Thank
You