LENDING CLUB CASE STUDY (ML-C47)

By Souvik Misra and Tanmaya R Sahu

TABLE OF CONTENTS

PROBLEM STATEMENT

Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

The purpose is 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 aim of this case study.

APPROACH TO SOLUTION

DATA CLEANING:

WE BEGIN WITH CLEANING THE DATASET. THIS INVOLVES REMOVING COLUMNS HAVING 30% OR MORE NULL VALUES, COLUMNS HAVING SINGULAR OR BUSINESS IRRELEVANT VALUES AND A FEW RESULTANT NULL ROWS. THEN WE FOLLOW WITH ALL THE ANALYSES PARALLELLY

UNIVARIATE ANALYSIS:

CHECKING THE DISTRIBUTION, COMPOSITION AND OTHER RELEVANT TRENDS FOR EACH REQUIRED COLUMN THAT CAN PROVIDE US WITH INSIGHTS

BIVARIATE ANALYSIS:

CHECKING HOW TWO VARIABLES FROM THE RELEVANT COLUMNS VARY WITH ONE ANOTHER, ESPECIALLY WHEN SEGMENTED ALONG LOAN STATUS

MULTIVARIATE ANALYSIS:

CHECKING HOW MORE THAN TWO VARIABLES VARY WITH ONE ANOTHER, ALSO SEGMENTED ALONG LOAN STATUS

By Souvik Misra and Tanmaya R Sahu

DATA CLEANING

NULL or Missing Value Columns
We start with identifying columns in the dataset that have more than 30% missing values. We could identify around 58 such columns, several of them almost entirely comprised of missing values

Columns having Single Values
Next, we proceed with dropping columns that only single value. Total nine Single valued columns removed.

Columns Created post Loan Approval and No Business Impact
We have Identified columns having unique id , created after Loan Approval , Column contains value same as other columns . Those columns are removed.

Removal records for missing Rows
We will remove records for missing rows as we can not filled values based on assumption.
Total percentage missing would be around 2%

Derived Columns & Refine Columns Values
Certain column datatypes will be converted and for some newer columns will be derived for better analysis

Final Data Set
Finally, we have our dataset ready for analyses

DATA CLEANING

```
Number of columns to be removed:  58
List of columns to be removed:
 ['verification_status_joint', 'annual_inc_joint', 'mo_sin_old_rev_tl_op', 'mo_sin_old_il_acct', 'bc_util', 'bc_open_to_buy', 'avg_cur_bal', 'acc_open_past_24mths', 'inq_last_12m', 'total_cu_tl', 'inq_fi', 'total_rev_hi_lim', 'all_util', 'max_bal_bc', 'open_rv_24m', 'open_rv_12m', 'il_util', 'total_bal_il', 'mths_since_rcnt_il', 'open_il_24m', 'open_il_12m', 'open_il_6m', 'open_acc_6m', 'tot_cur_bal', 'tot_coll_amt', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'num_rev_tl_bal_gt_0', 'total_bc_limit', 'total_bal_ex_mort', 'tot_hi_cred_lim', 'percent_bc_gt_75', 'pct_tl_nvr_dlq', 'num_tl_op_past_12m', 'num_tl_90g_dpd_24m', 'num_tl_30dpd', 'num_tl_120dpd_2m', 'num_sats', 'num_rev_accts', 'mths_since_recent_bc', 'num_op_rev_tl', 'num_il_tl', 'num_bc_tl', 'num_bc_sats', 'num_actv_rev_tl', 'num_actv_bc_tl', 'num_accts_ever_120_pd', 'mths_since_recent_revol_delinq', 'mths_since_recent_inq', 'mths_since_recent_bc_dlq', 'dti_joint', 'total_il_high_credit_limit', 'mths_since_last_major_derog', 'next_pymnt_d', 'mths_since_last_record', 'mths_since_last_delinq', 'desc']
```

There are some columns that do not represent a user's risk taking capability and are irrelevant to our analysis. A list of such columns are as follows:

- id
- member_id
- emp_title
- url
- title
- zip_code
- earliest_cr_line
- last_pymnt_d
- last_credit_pull_d

```
Number of Single Value columns:  6
List of columns to be removed:
 ['pymnt_plan', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq', 'delinq_amnt']
```

	percent_missing
emp_length	2.706650
pub_rec_bankruptcies	1.754916
chargeoff_within_12_mths	0.140998
collections_12_mths_ex_med	0.140998
revol_util	0.125891
tax_liens	0.098195
total_rec_prncp	0.000000
total_acc	0.000000
out_prncp	0.000000
total_pymnt	0.000000
recoveries	0.000000
total_rec_int	0.000000
total_rec_late_fee	0.000000

We could observe that the following columns are relevant only post the Charge-off i.e. default so they are not needed here. Therefore we will be removing them

1. collections_12_mths_ex_med
2. chargeoff_within_12_mths
3. tax_liens

*the column descriptions can be obtained from the data dictionary uploaded to the same repo

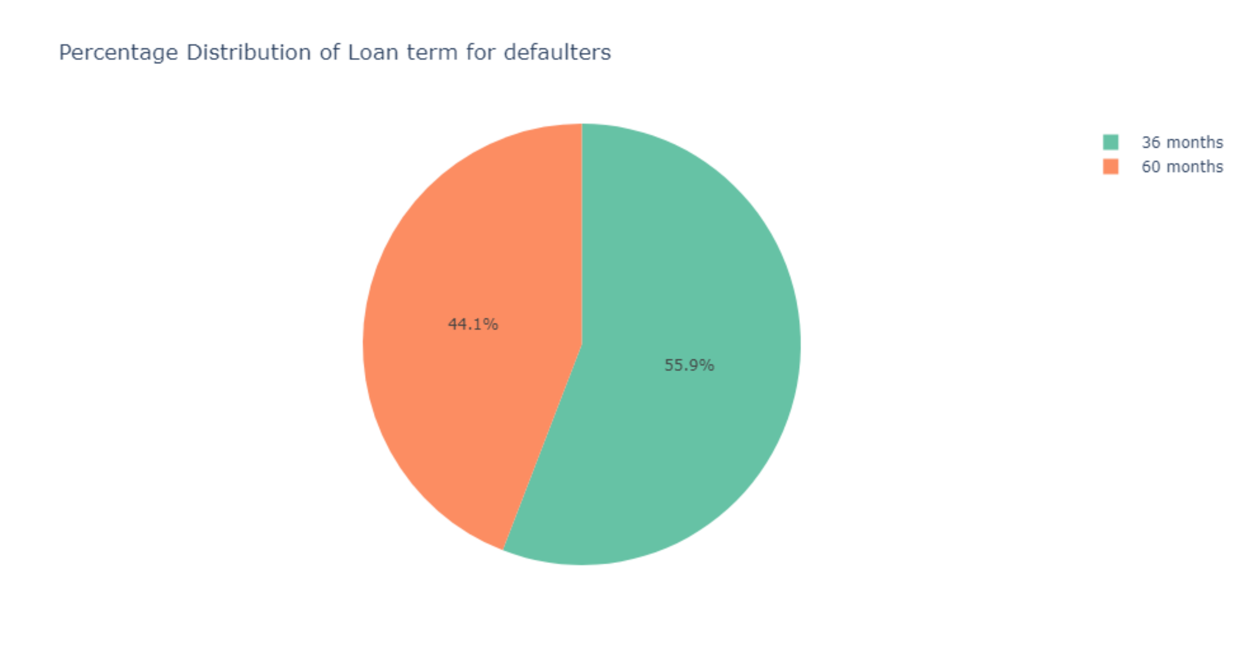
By Souvik Misra and Tanmaya R Sahu

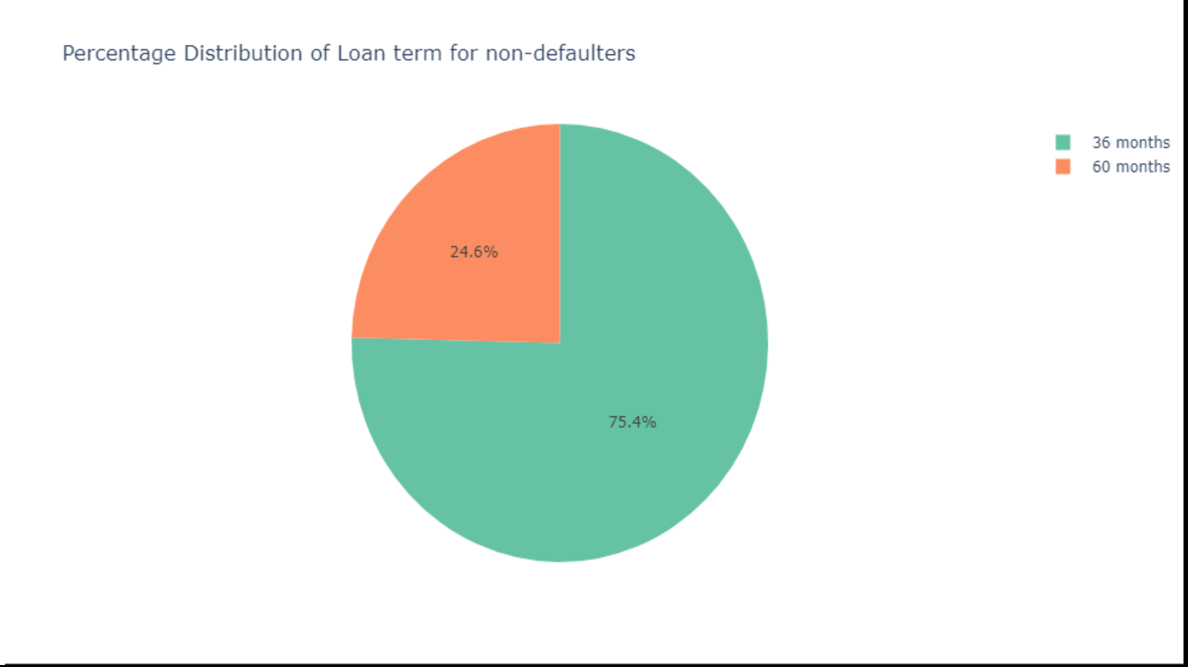
UNIVARIATE ANALYSIS

The first step of EDA analysis is Univariate Analysis and Univariate Segmented Analysis.
We have used some of columns for Univariate Analysis and Univariate Segmented Analysis between Defaulters and Non-Defaulters.
Out of that following columns are clearly indicates Defaulters based on comparison of values with Non-defaulters.

- **Loan Terms**
- **Annual Income**
- **Interest Rate**
- **Credit Revolution Utilization rates**
- **Home Ownership**
- **Loan Purpose**
- **DTI Rate**
- **Grade**
- **Issued Month**
- **State of Residency**
- **Revolving credit balance**

UNIVARIATE ANALYSIS





- **Loan Term** : Defaulters are 20% more likely to avail a 60 month loan term than non-defaulters

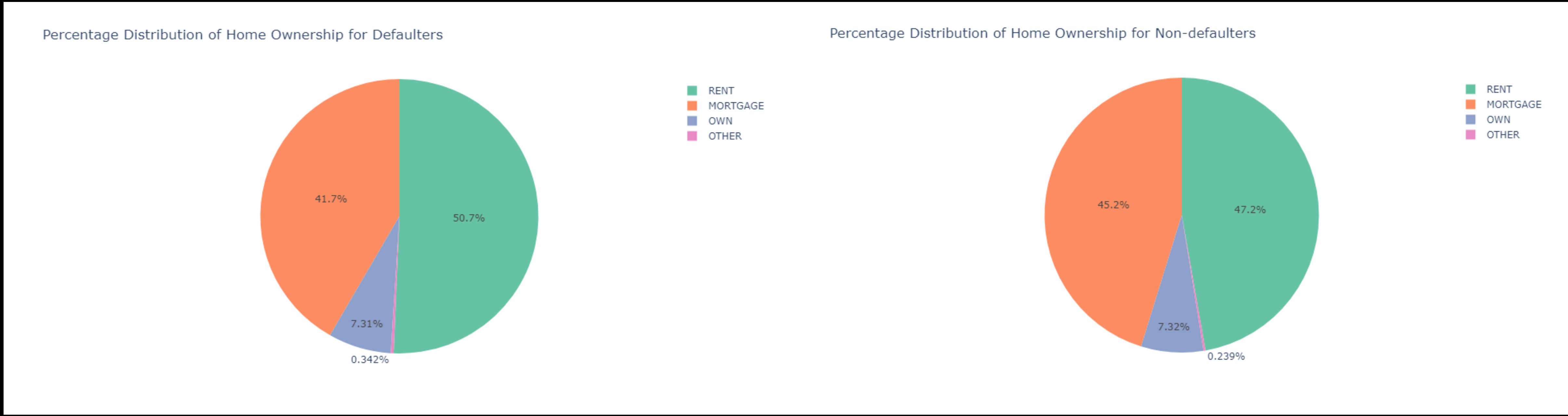
UNIVARIATE ANALYSIS (CONTD.)

- **Annual Income** : Defaulters annual income median level USD 4500 less than the non-defaulters. In other words, defaulters are low annual income group.

UNIVARIATE ANALYSIS (CONTD.)

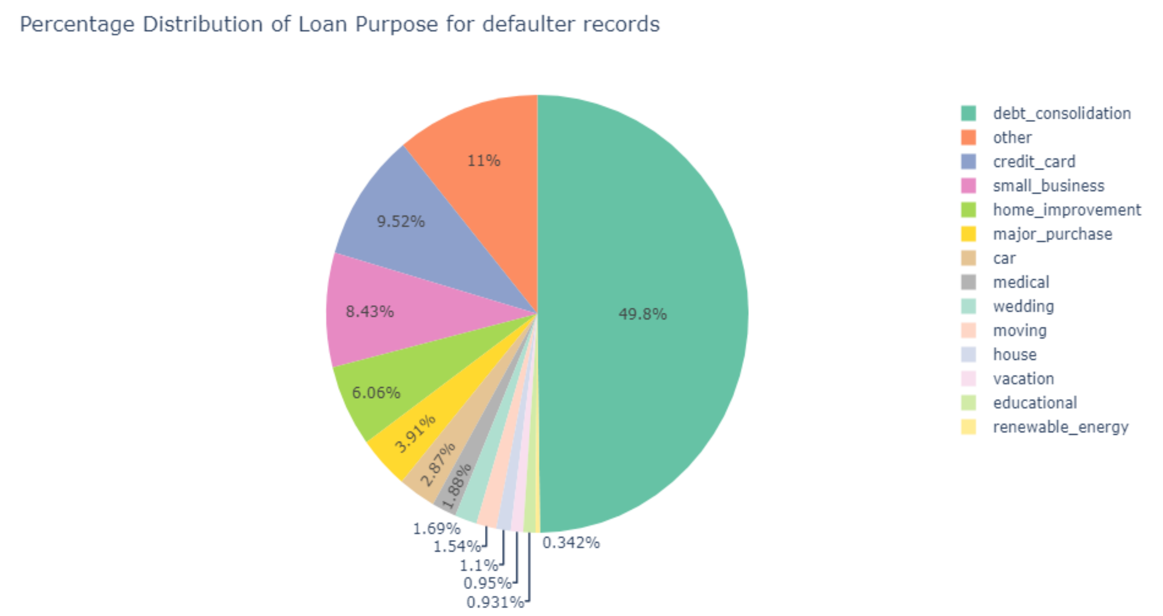
- **Credit Revolution Utilization rates:** For revolving Credit Limit Utilization rates, most defaulters are utilizing around 15% more than most of the non-defaulters. Therefore, higher rates especially above 60% can have a greater indication of default

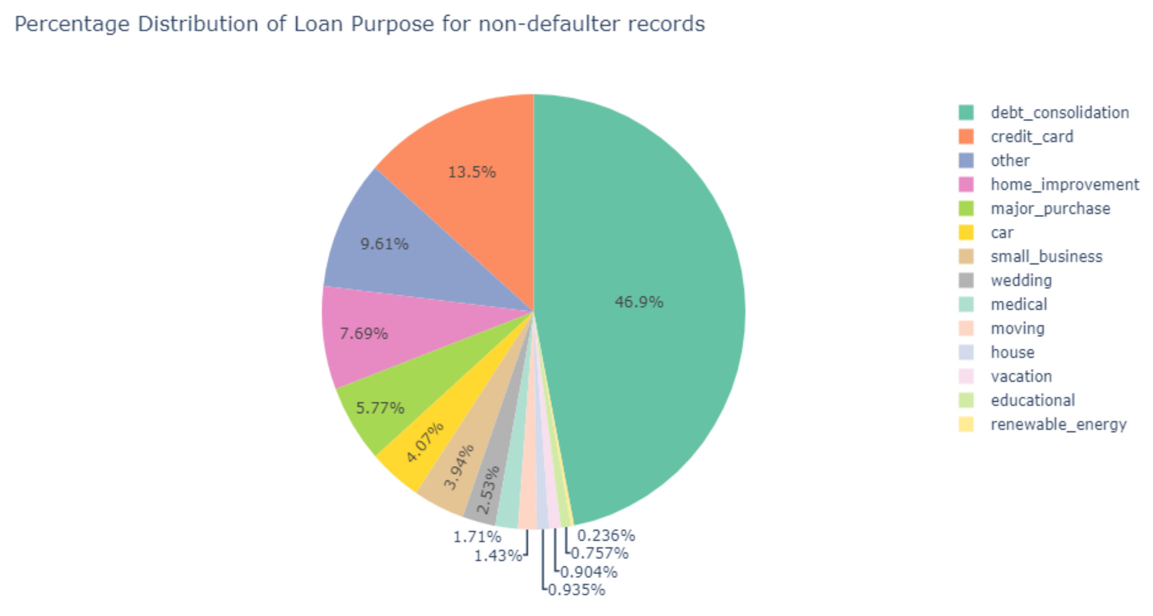
UNIVARIATE ANALYSIS (CONTD.)



- **Home Ownership :** 3.4% more defaulters tend to stay in Rented houses as compared to non-defaulters

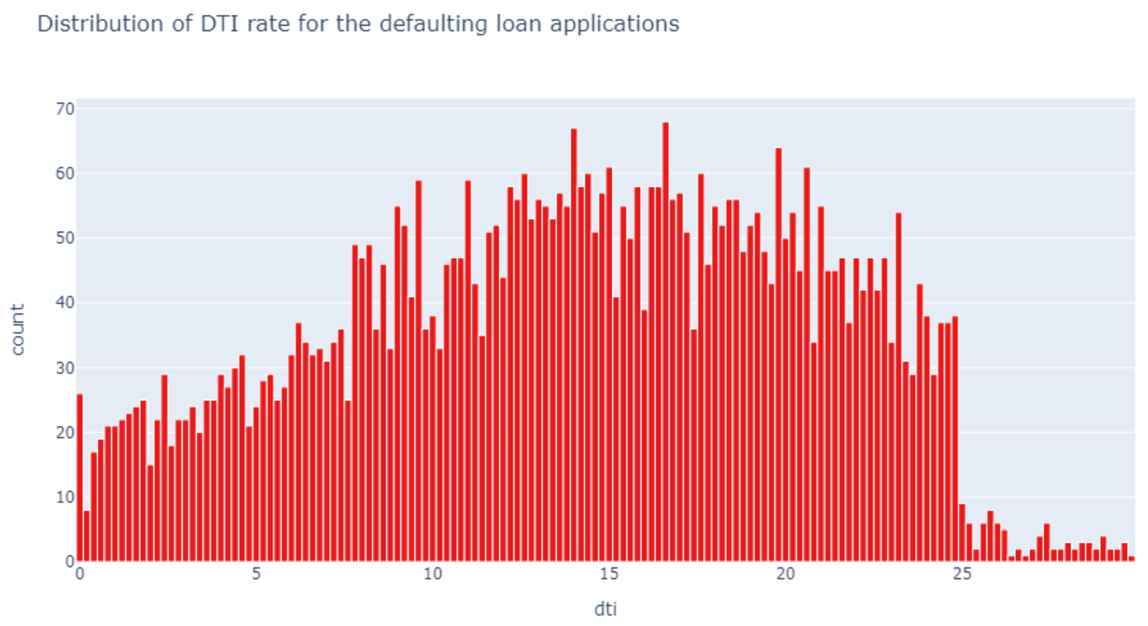
UNIVARIATE ANALYSIS (CONTD.)

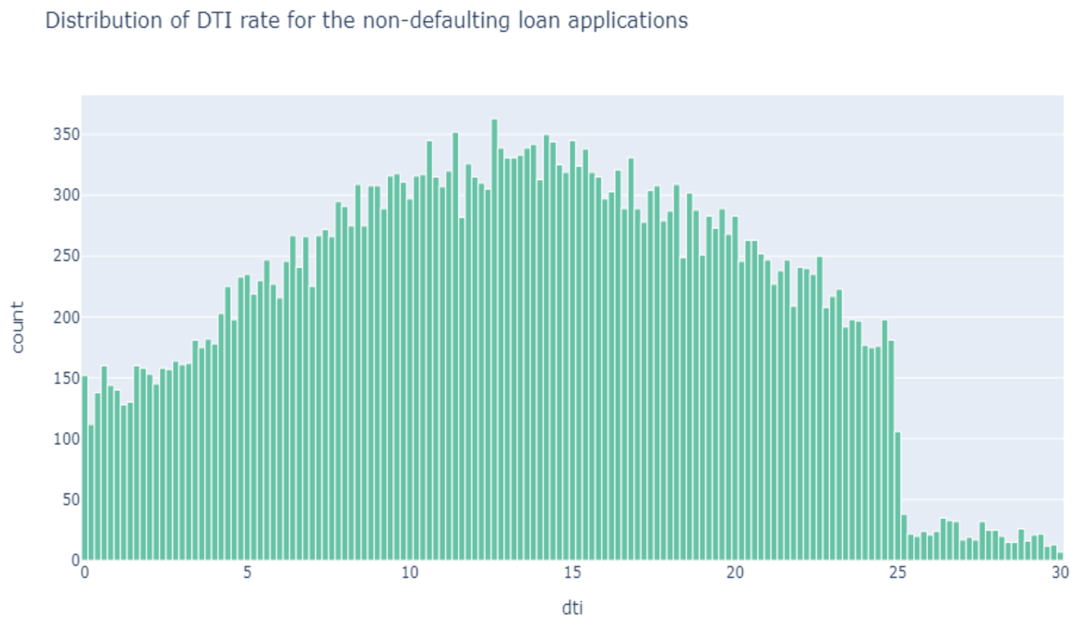




- **Loan Purpose:** Defaulters are around 3% more likely to avail loans for debt consolidation

UNIVARIATE ANALYSIS

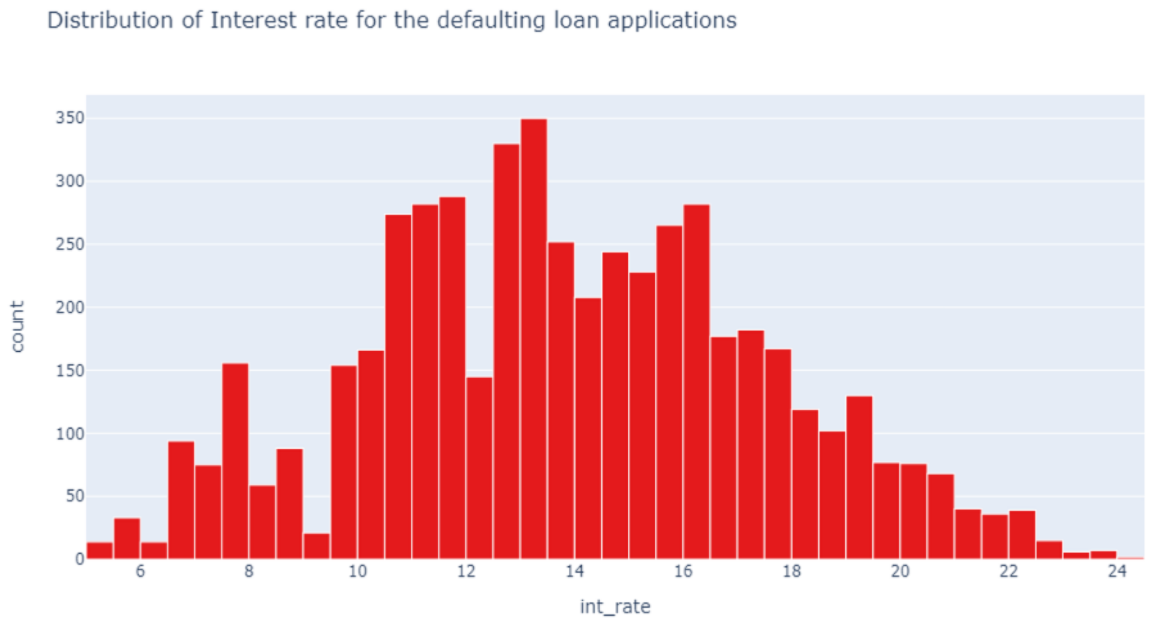


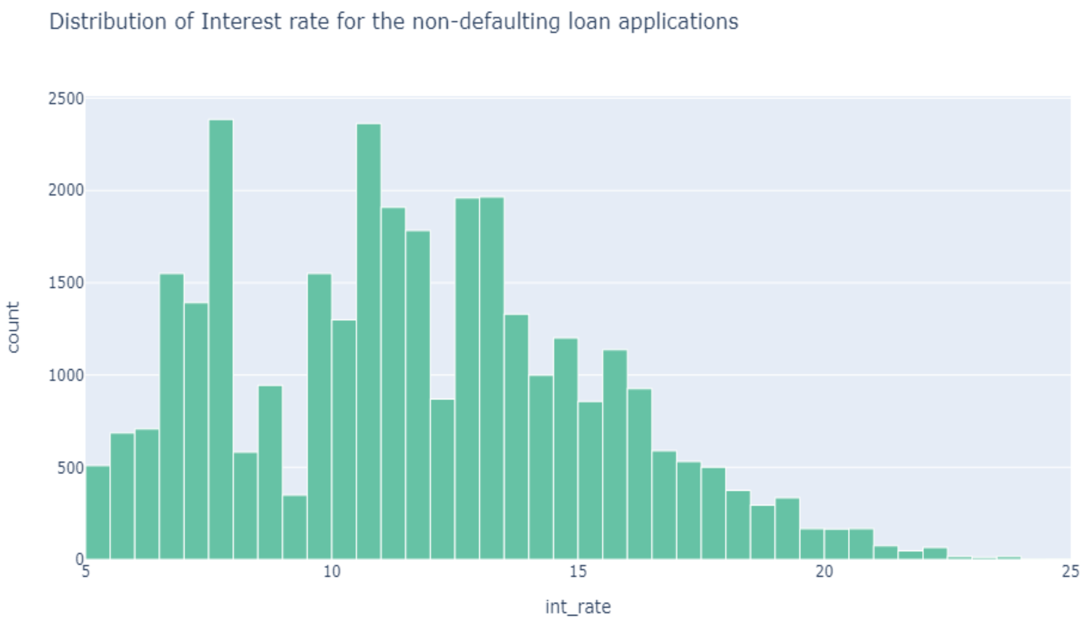


- **DTI Rate:** For DTI (Debt to Income) rates, most defaulters are likely to take on 4% more debt per unit income as compared to most of the non-defaulters. Therefore, higher DTI rates (>16%) can have a greater indication of default

By Souvik Misra and Tanmaya R Sahu

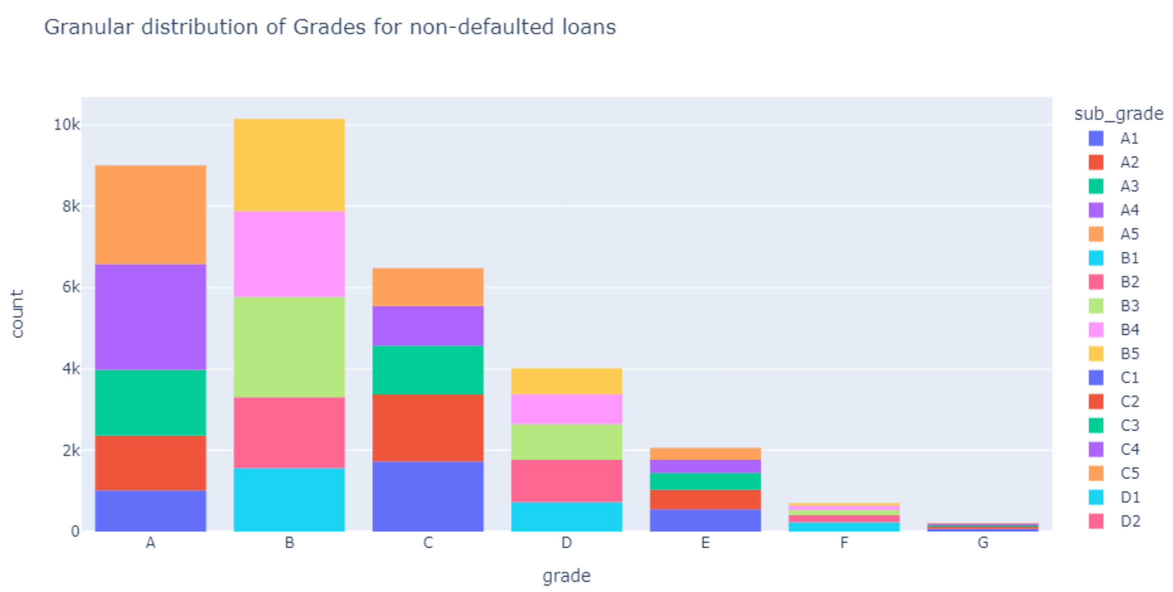
UNIVARIATE ANALYSIS





- **Interest Rate:** Interest rate is higher for Defaulters in comparison to Interest rate for Non-defaulters.

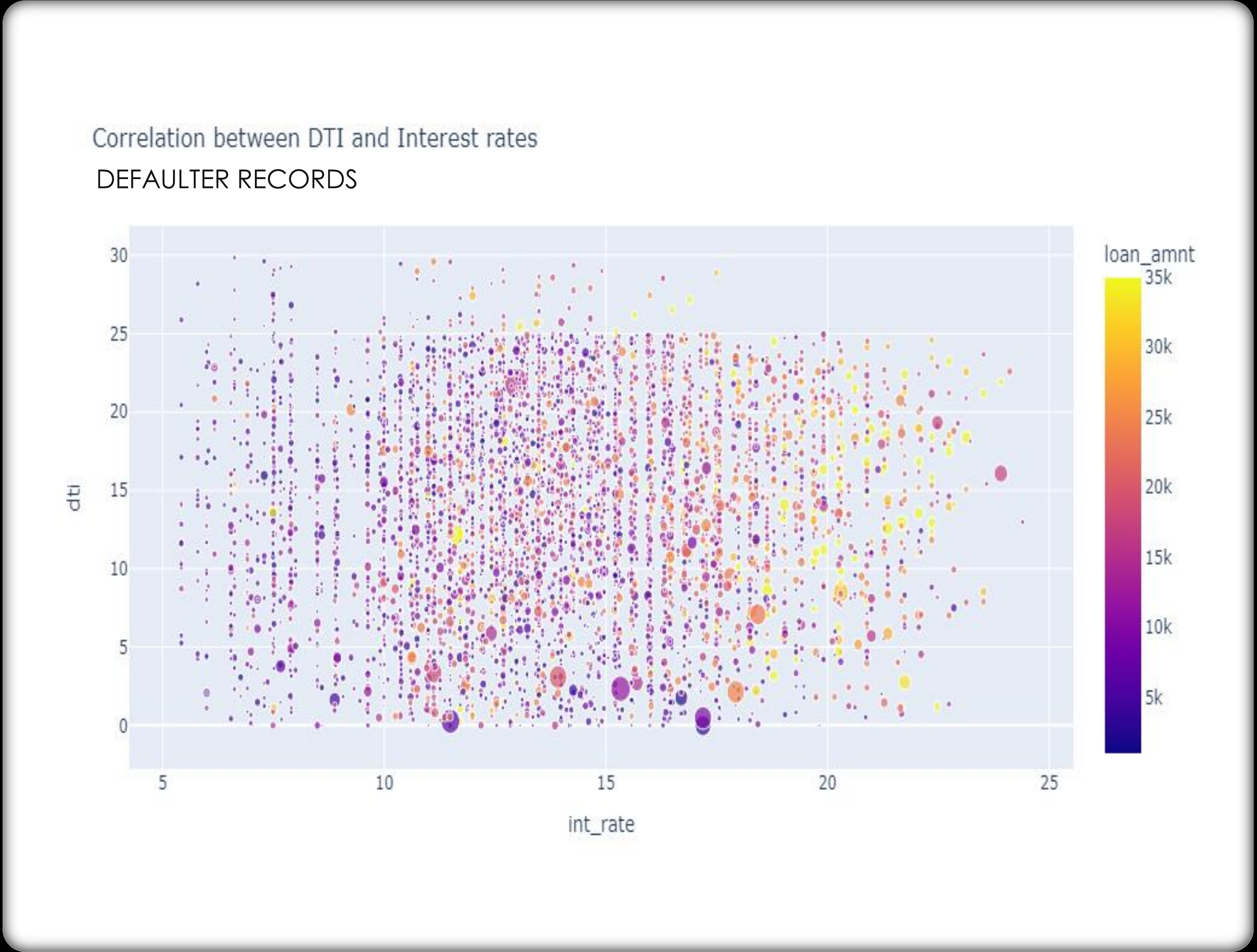
BIVARIATE & MULTIVARIATE ANALYSIS



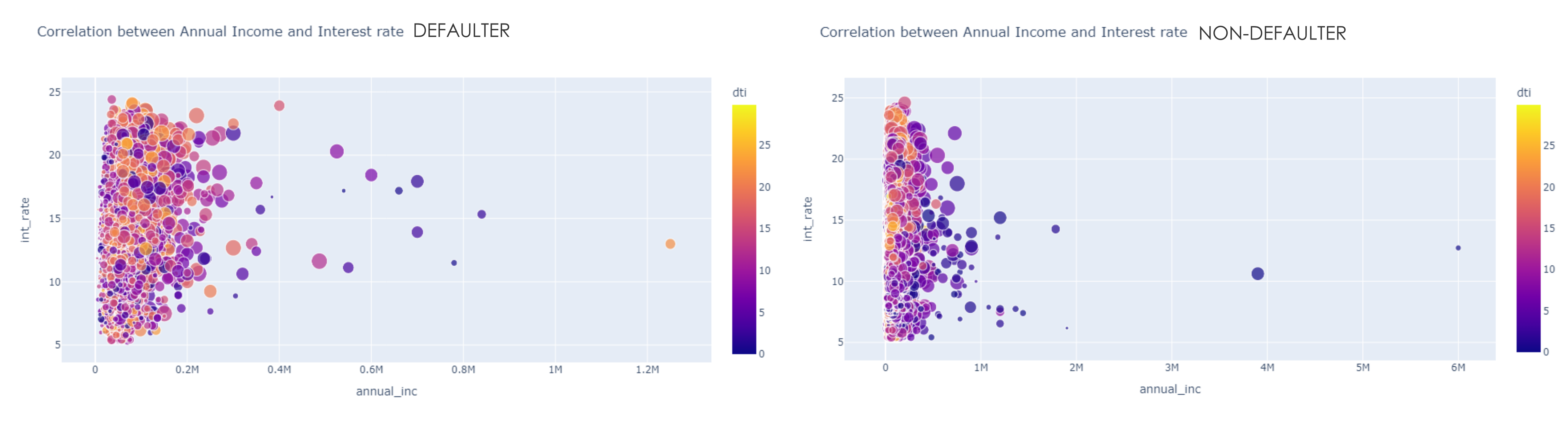
It is observed that although in both cases most of the applicants are from Grade B, the Grade A employees tend to have a greater proportion of non-defaulters whereas lower grades such as C, D, E etc tend to have greater proportion of defaulters. For subgrades however there is no clear and observable pattern

BIVARIATE & MULTIVARIATE ANALYSIS (CONTD.)

- It is also observed that in defaulters, applicants with higher annual income tend to have lower DTI and tend to stay within the interest rate of 11-18%
- On the other hand, in the same dataset defaulting applicants with lower annual income tend to go for higher loan amounts coupled with higher interest rates
- Size of the circle is proportional to the annual income



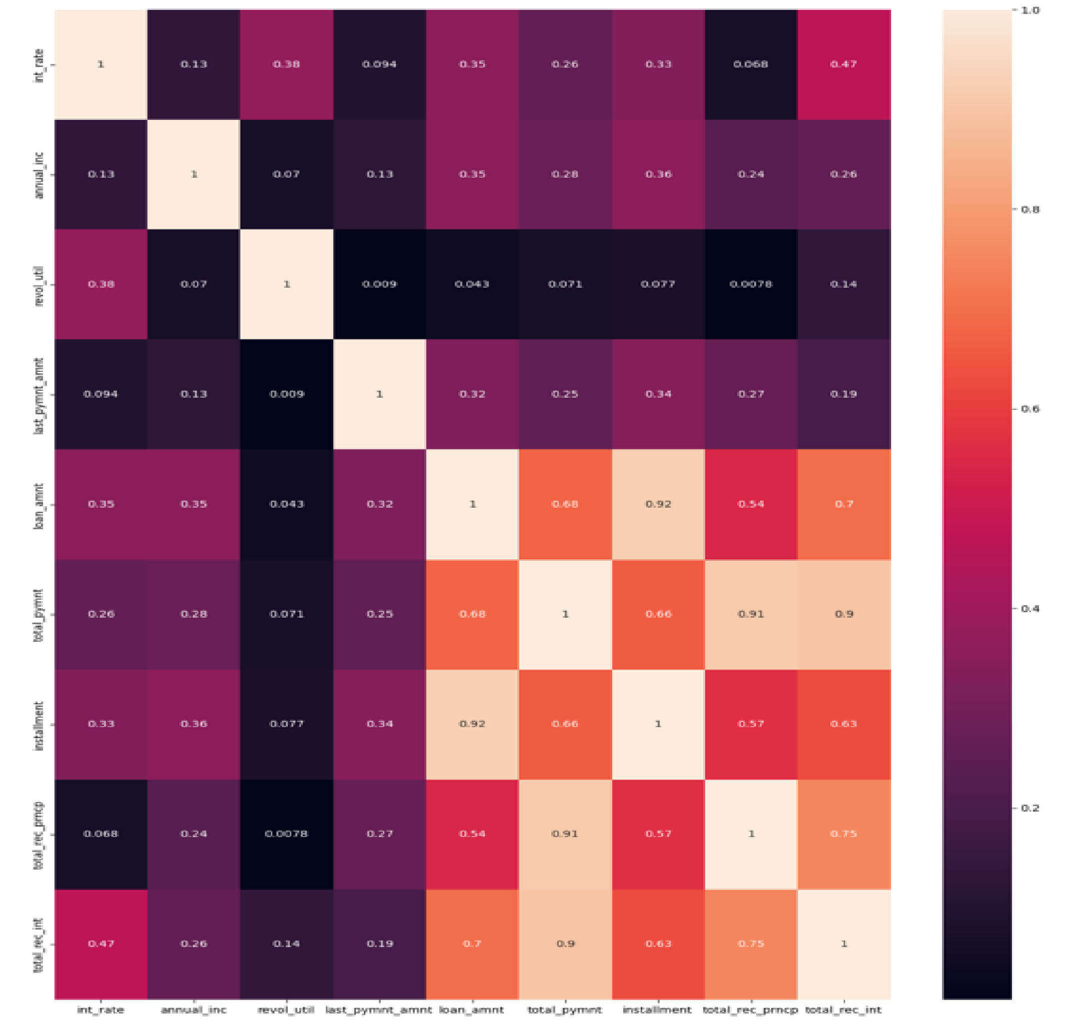
BIVARIATE & MULTIVARIATE ANALYSIS (CONTD.)



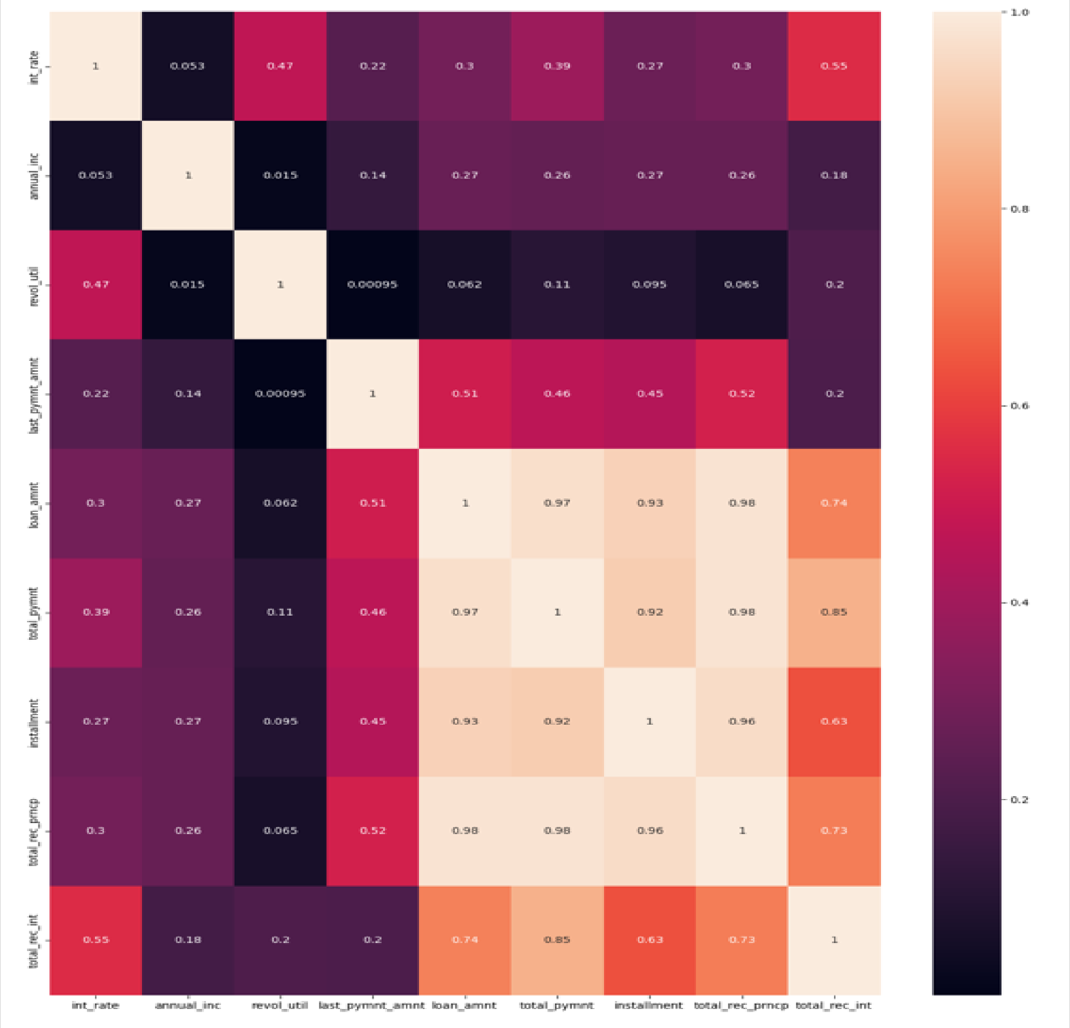
Non defaulters tend to earn higher annually than defaulters. While most defaulters earn within the USD 400K per annum, most non-defaulters are earning well up to USD 1 million per annum. Also higher interest rate implies higher DTI ratio for the applicants. Size of the circle is proportional to the annual income

BIVARIATE & MULTIVARIATE ANALYSIS (CONTD.)

DEFAULTER



NON-DEFAULTER



BIVARIATE & MULTIVARIATE ANALYSIS (CONTD.)

- As observed in the previous heatmap, for defaulters there is lower correlation between the listed amount of the loan applied for by the borrower , the monthly installment amount, the principal received to date and the payments received to date for total amount funded than non-defaulters where they are highly, almost perfectly correlated. This implies that potential defaulters are irregular in fulfilling their credit repayment obligations

is_defaulter	inq_grp	pub_rec	pub_rec_bankruptcies
NO	0-2	0.047995	0.037412
	2-4	0.064266	0.046070
	4-6	0.006410	0.000000
	6-8	0.000000	0.000000
YES	0-2	0.081882	0.063807
	2-4	0.098071	0.072347
	4-6	0.000000	0.000000
	6-8	0.000000	0.000000

is_defaulter	delinq_grp	pub_rec	pub_rec_bankruptcies
NO	0-2	0.049021	0.037901
	2-4	0.031250	0.026786
	4-6	0.111111	0.074074
	6-8	0.750000	0.250000
	8+	0.000000	0.000000
YES	0-2	0.082391	0.063805
	2-4	0.100000	0.050000
	4-6	0.250000	0.250000
	6-8	1.000000	1.000000

is_defaulter	total_acc_grp	pub_rec	pub_rec_bankruptcies
NO	2-10	0.043310	0.028011
	10-18	0.056643	0.043162
	18-26	0.057170	0.046275
	26-34	0.042905	0.033292
	34-42	0.036891	0.031519
	42-50	0.028961	0.025554
	50-58	0.024390	0.017738
	58-66	0.004926	0.004926
	66-74	0.000000	0.000000
	74-82	0.000000	0.000000
	82-90	0.000000	0.000000
YES	2-10	0.034910	0.020270
	10-18	0.095082	0.067541
	18-26	0.105431	0.087859
	26-34	0.086550	0.070175
	34-42	0.081197	0.066239
	42-50	0.074074	0.068783
	50-58	0.040000	0.040000
	58-66	0.027778	0.027778
	66-74	0.000000	0.000000

is_defaulter	open_acc_grp	pub_rec	pub_rec_bankruptcies
NO	2-8	0.049029	0.036709
	8-14	0.050181	0.039956
	14-20	0.044373	0.036438
	20-26	0.047170	0.032075
	26-32	0.093750	0.046875
	32-38	0.000000	0.000000
	38-44	0.000000	0.000000
YES	2-8	0.070331	0.051115
	8-14	0.086978	0.068305
	14-20	0.128352	0.109195
	20-26	0.105882	0.094118
	26-32	0.062500	0.062500
	32-38	0.000000	0.000000

BIVARIATE & MULTIVARIATE ANALYSIS (CONTD.)

BIVARIATE & MULTIVARIATE ANALYSIS

- From the previous results, we can come to the following conclusion that on an average the defaulters are more likely to have derogatory public records and bankruptcies, across delinquencies, inquiries and the number of credit lines present

SUMMARY

- The following factors should be considered while receiving a loan application from anybody to the Lending Club
  - Revolving Credit Line utilization rate
  - Annual income group
  - Employment Grade
  - Debt to Income ratio
  - Rate of interest as compared to annual income
  - Loan amount availed as compared to annual income
  - Number of derogatory public records
  - Number of public record bankruptcies
  - Loan term availed
  - Loan Purpose availed
  - Correlation between the listed amount of the loan applied for by the borrower , the monthly installment amount, the principal received to date and the payments received to date for total amount funded to observe the consistency in fulfilling their credit repayment obligations

By Souvik Misra and Tanmaya R Sahu