CREDIT RISK ANALYSIS CASE STUDY

SUBMISSION

Group Name: The MBBS Group

1. Murali Banala
2. Barkha Garg
3. Bhumesh Arkal
4. Suryateja Rallapalli

Context & Objective

The analysis activities in this Case Study focus on a consumer finance company (Anonymous) 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. An incorrect assessment of the applicant profile may lead to company losing the business due to the following:

- 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

With the aim of reducing the probability of the company incurring losses, there could be either of the below 2 decisions the company can make when a fresh loan application is received:

- Loan accepted: If the company approves the loan, there are 3 possible 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 instalments 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.

So the focus of this case study is to understand (using the provided loan data (loan.csv)), how the consumer and the loan attributes influence the tendency of Default and highlight the key indicators that could lead to a customer Defaulting the loan. We have used Jupyter notebook to build the Python code using various EDA techniques.

Data Cleaning & Formatting

Data Loading & Data Cleaning

Step 1: Loading the loan data file

a) The loan data (loan.csv) was loaded in Data Frame **loan_data**.

b) The file was loaded using the encoding standard ISO-8859-1 to avoid any encoding issues that might arise

c) The total count of records in loan_data data frame was 39717 and 111 columns

Step 2: Cleaning the Data

a) There were total 111 columns in the dataset

b) Around 54 columns had all the values as null. So these 54 columns were dropped entirely

c) Dropped the columns with still high null percentage of values: mths_since_last_delinq (~65%), mths_since_last_record (~93%), next_pymnt_d (~97%), desc (~33%)

d) On exploring a little about the Credit Risk Analysis Domain, found that few columns viz., member_id, url, zipcode, initial_list_status, collection_recovery_fee, policy_code, emp_title, title, last_credit_pull_d, earliest_cr_line, total_pymnt, total_rec_int, total_rec_late_fee, total_rec_prncp, last_credit_pull_d, last_pymnt_amnt, last_pymnt_d, collections_12_mths_ex_med, collection_recovery_fee, earliest_cr_line, initial_list_status will not be contributing to our analysis. Hence these columns were also dropped from the dataset.

e) The column pymnt_plan had only one value for all the records, so this was dropped too as this will not contribute in our analysis

f) column emp_length had close to 3% nulls and this would be required in our analysis, the rows with this value as null were dropped so that our data is not inflated with any nulls or imputed value.

g) So we were left with 35 columns and 38642 (approx. 97%) rows after performing these cleaning activities

Data Cleaning & Formatting (Cntd..)

Removing duplicates from the dataset

Step 1: id column represents each loan request uniquely, so checked the duplicity of this value

Step 2: No duplicated were found. Hence no further records were dropped

Converting to proper Datatypes and Format

Step 1: Columns like term, int_rate, revol_util had values like 36 months, 3%. Such data was standardized and the string months & % were dropped to store these values in numeric format

Step 2: column issue_d was originally of type object, so this was appropriately formatted and converted to Date datatype

Data Cleaning & Formatting (Cntd..)

- For ease of analyzing, buckets were created for values of attributes line Annual income, dti, funded amount.
- These buckets will ease the analysis when plotting distributions
- Below is the range if values in each bucket for each attribute

Range	Bucket
annual_inc <= 30000	0
30000 < annual_inc <= 35000	1
35000 < annual_inc <= 40000	2
40000 < annual_inc <= 45000	3
45000 < annual_inc <= 55000	4
55000 < annual_inc <= 65000	5
65000 < annual_inc <= 75000	6
75000 < annual_inc <= 90000	7
90000 < annual_inc <= 120000	8
120000 > annual_inc	9

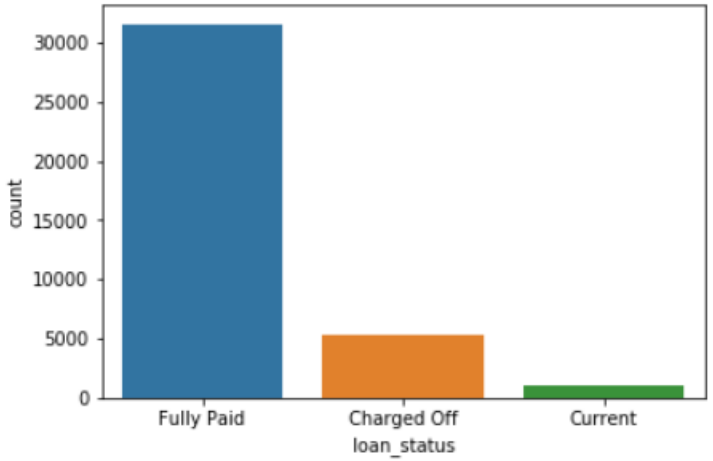
Range	Bucket
dti <= 4	0
4 < dti <= 6	1
6 < dti <= 8	2
8 < dti <= 10	3
10 < dti <= 12	4
12 < dti <= 16	5
16 < dti <= 18	6
18 < dti <= 20	7
20 < dti <= 22	8
dti > 22	0

Range	Bucket
funded amount<=2000	0
2000 < funded amount <= 4000	1
4000 < funded amount <= 6000	2
6000 < funded amount <= 8000	3
8000 < funded amount <= 10000	4
10000 < funded amount <= 12000	5
12000 < funded amount <= 14000	6
14000 < funded amount <= 16000	7
16000 < funded amount <= 22000	8
funded amount > 22000	9

Spread of Data 1

In this, we analysed important attributes in terms of their occurrences in the data and the spread of their values to draw some insights out of those:

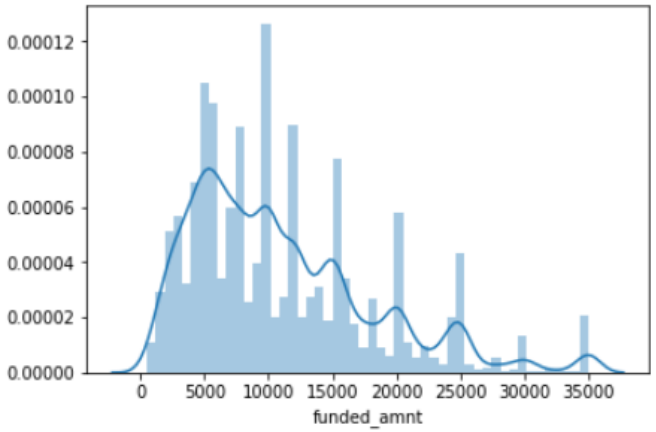
Step 1: Number of loan applications by various loan statuses:

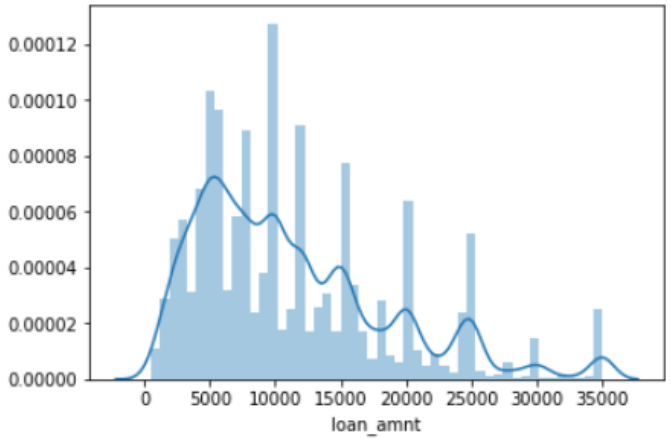


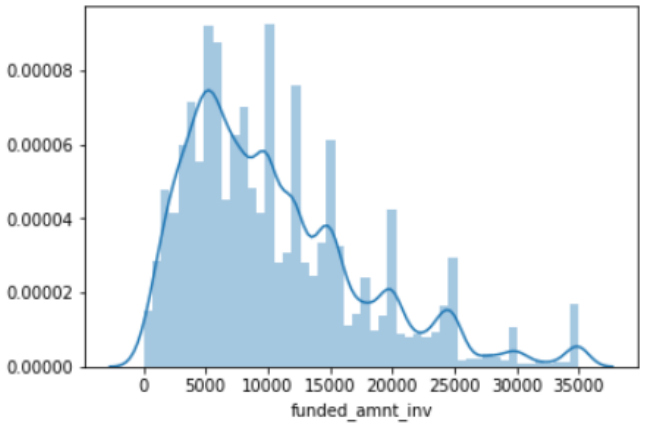
Maximum of the loans are settled Fully while a plenty of them are Defaulted. We will find out the factors of Default in further analysis

Spread of Data 2

- This is to see a trend of amount for loan amount, funded amount and funded amount that has been invested



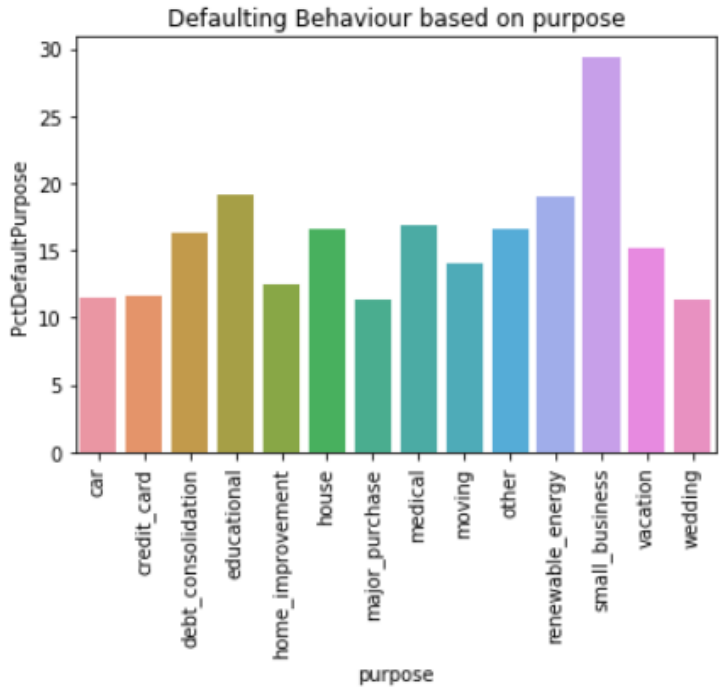




- Observation: The pattern of funds in each of the 3 categories is the same. So that implies there is no deficiency of funds. What loan amount is funded gets proportionally invested and so there is adequate balance between funding & credit

Univariate Data Analysis (Cntd..)

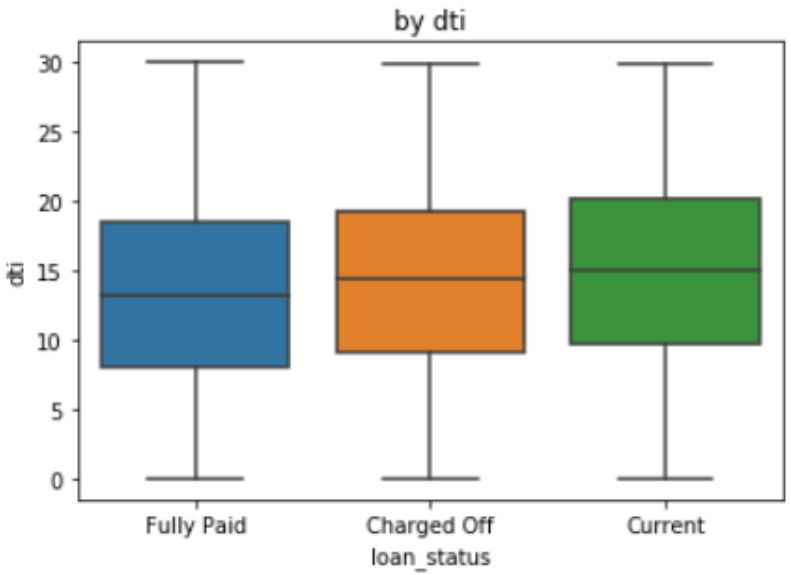
Step 2: Defaulting behavior based on the % of defaults for various loan purposes

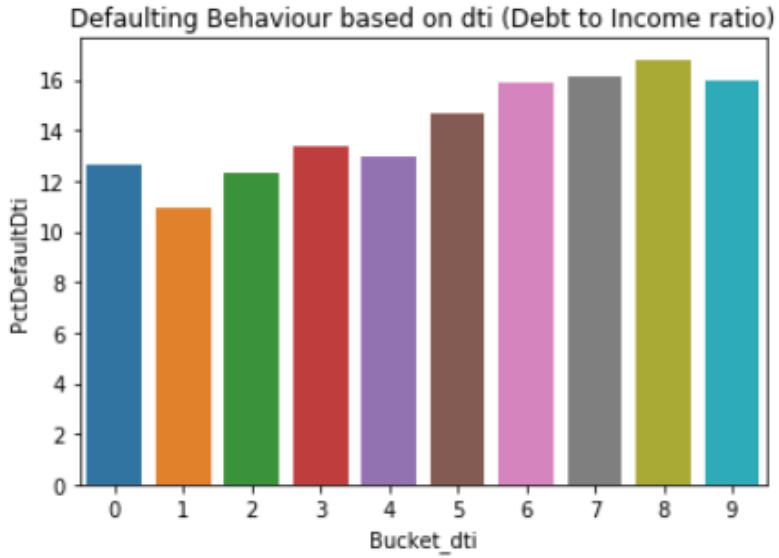


There are few purposes for which percentage of default is slightly higher than average - for example - small_business, educational, renewable_energy

Univariate Data Analysis (Cntd..)

- Step 3: Influence of dti (Debt-to-Income Ratio)

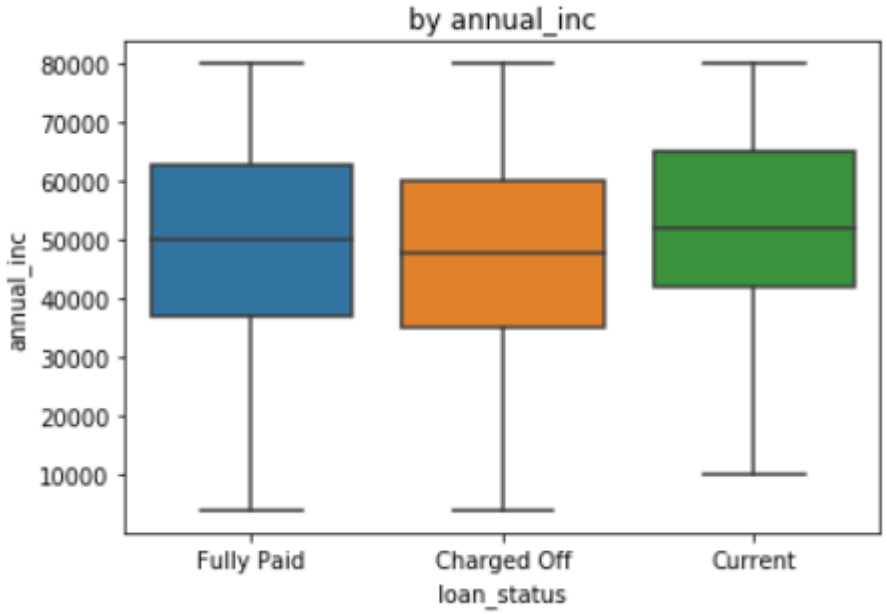


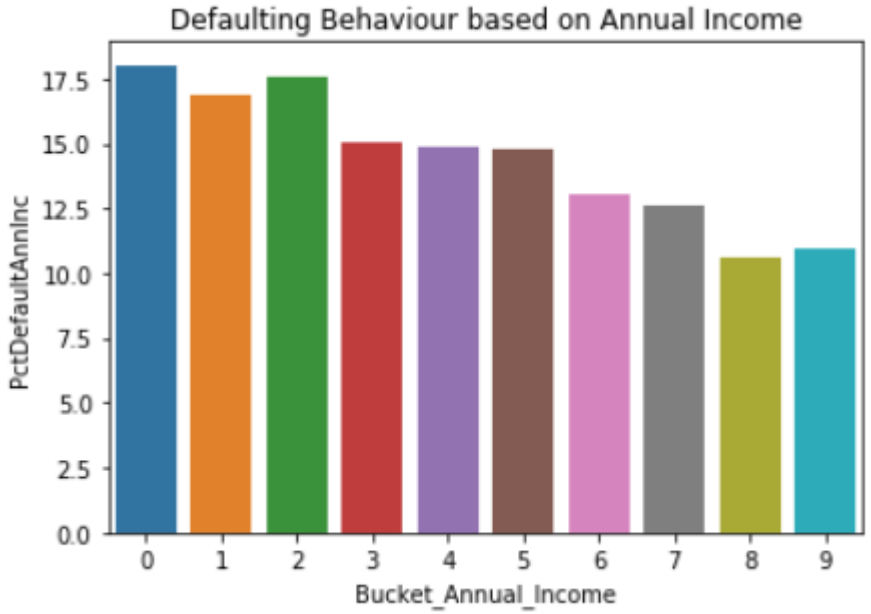


- The median dti for Charged off loans is a bit higher than the Fully Paid loans. This means that borrowers with higher dti (more debts) are slightly more leaned towards Defaulting as they already have lot of other debts to pay in a month
- Also shown in the second figure, as dti increases, percentage of default increases too. So dti is certainly a strong indicator of default

Univariate Data Analysis (Cntd..)

- Step 4: Spread of annual income of the applicants for various loans and the Defaults

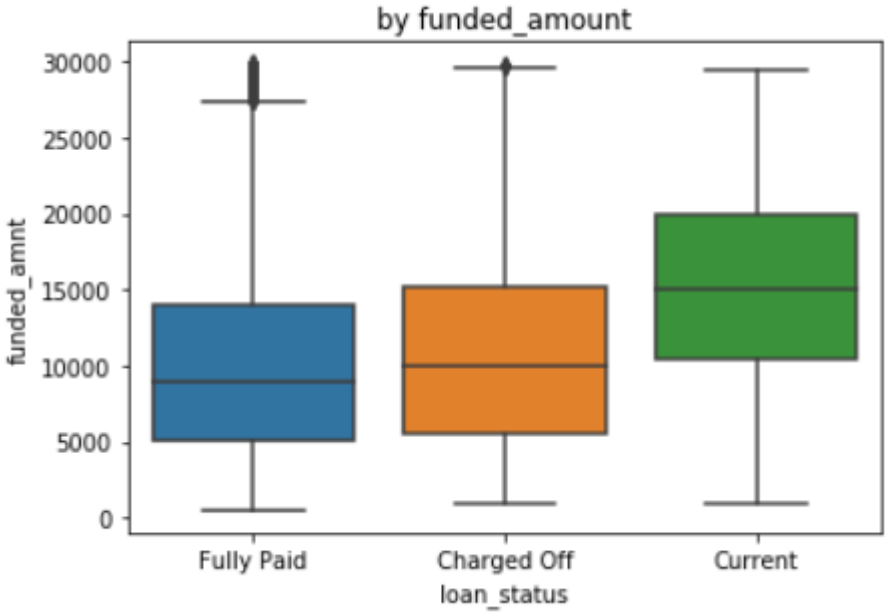


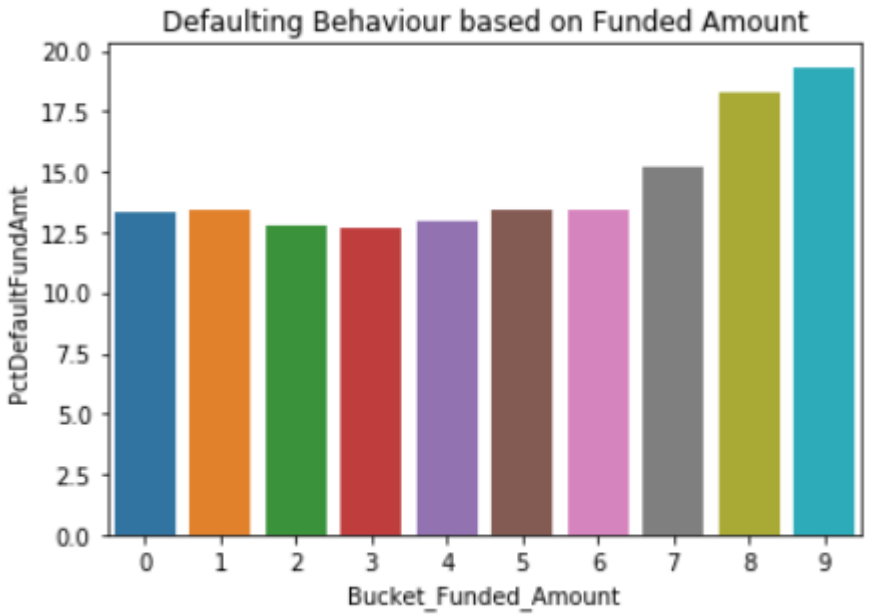


- The median annual inc for Fully Paid is higher than the Charged off loans
- Also, as the second plot shows, a lower annual income caused higher % of Defaults
- So annual income is also a potential driver of a loan being Defaulted

Univariate Data Analysis (Cntd..)

- Step 5: Spread of funded amount for various loans and the Defaults

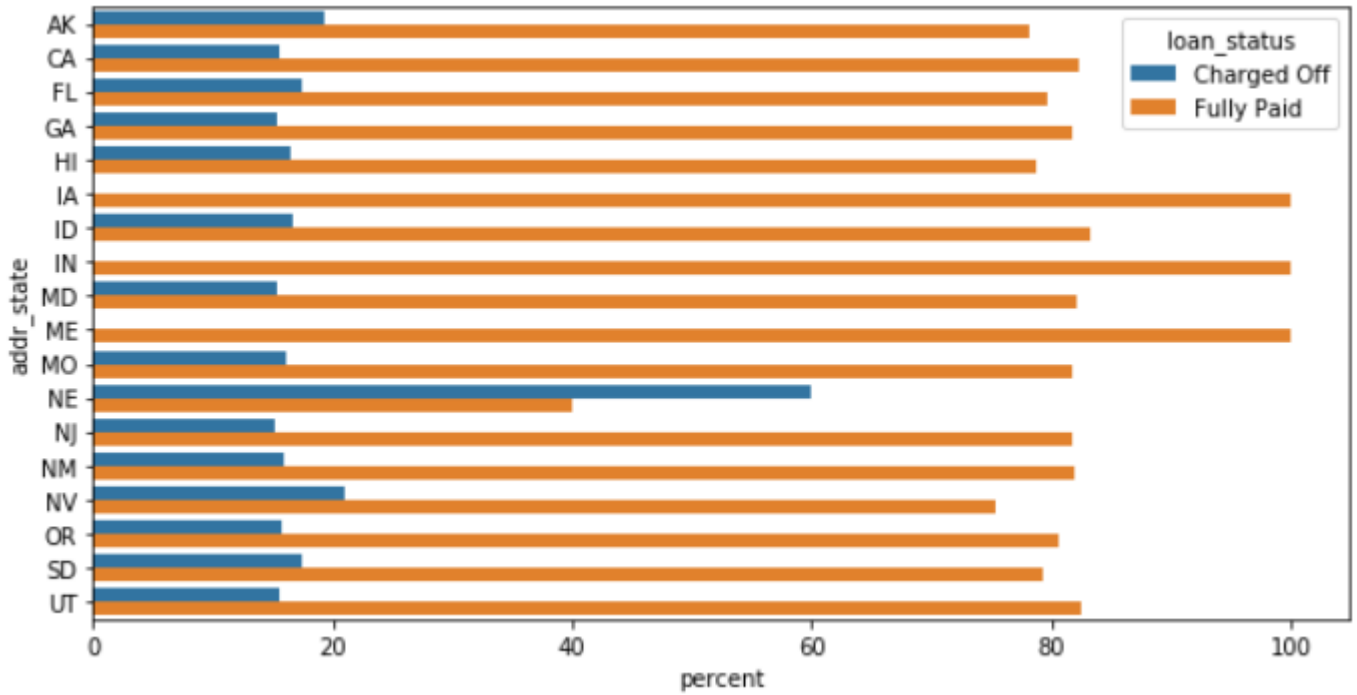


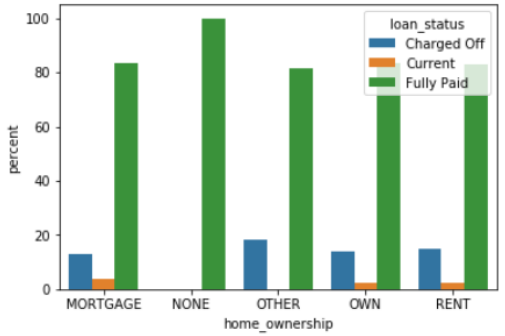


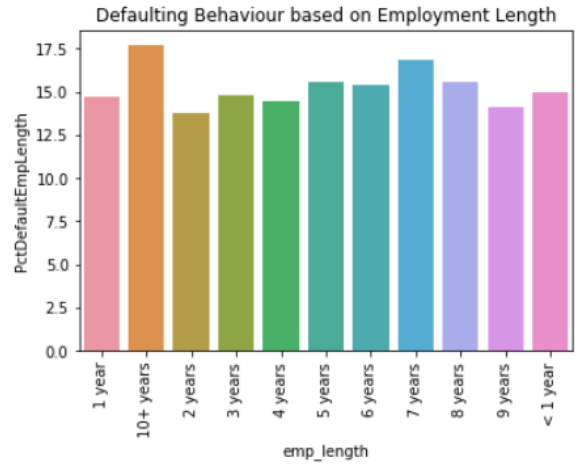
- Also, the median amount funded for Fully Paid loans is lower than the Charged off loans
- Also, as the second plot shows, as the amount funded increases the % of Defaults also goes up
- So funded amount also contributes to a loan being Defaulted

Univariate Data Analysis (Cntd..)

- Step 6: Loan status by employment length, home ownership & employment length
-



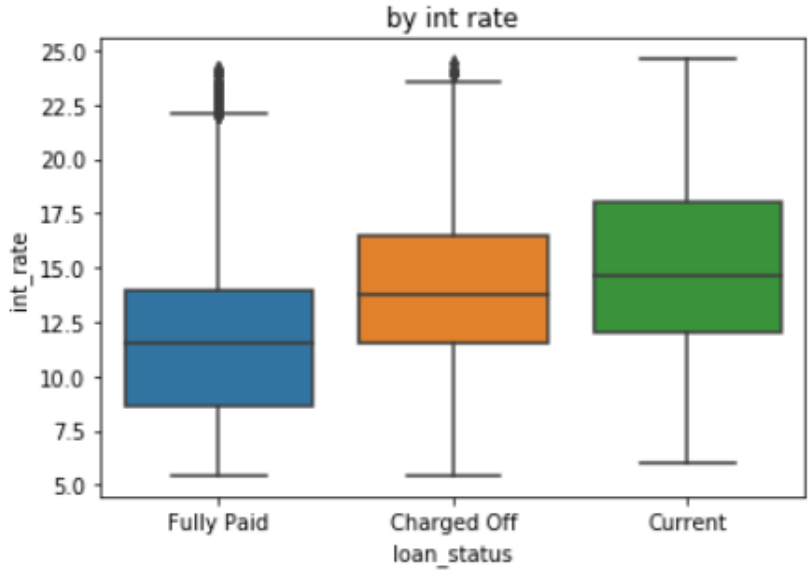




- As seen above, emp_legth and home ownership status do not appear to contribute to loan defaulting as there is not much deviation
- There are quite a few states (out of a total of 135 states) where the loan getting Defaulted are over 15%. So may be when sanctioning new loans, LC might review the loan application from that state more strictly and approve loan only if other influencing factors are well within the range

Univariate Data Analysis (Cntd..)

- Step 7: Influence of Interest Rate on loan defaults

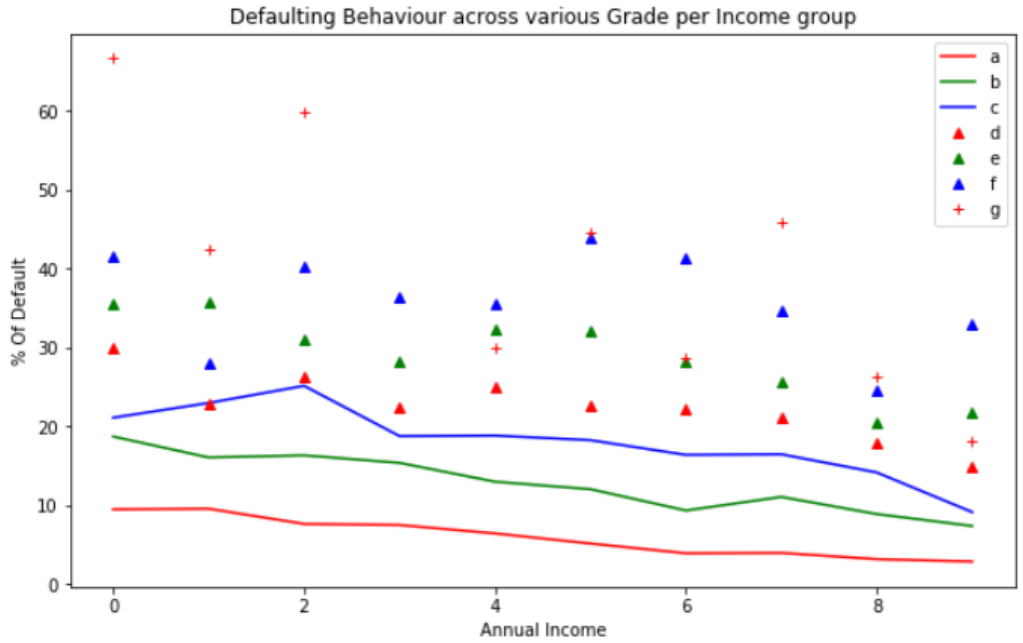


grade	
A	7.343809
B	11.052452
C	13.604269
D	15.776128
E	17.751342
F	19.806970
G	21.422283

- The median interest rate for Charged off loans is a lot higher than the Fully Paid loans. This means that borrowers getting loans at a higher interest rate are likely to default as applicants have to pay higher amount (A higher interest rate corresponds to a higher grade of loan like F, G, H etc.)

Bivariate Analysis

a) Grade and Income Group¶



Observations

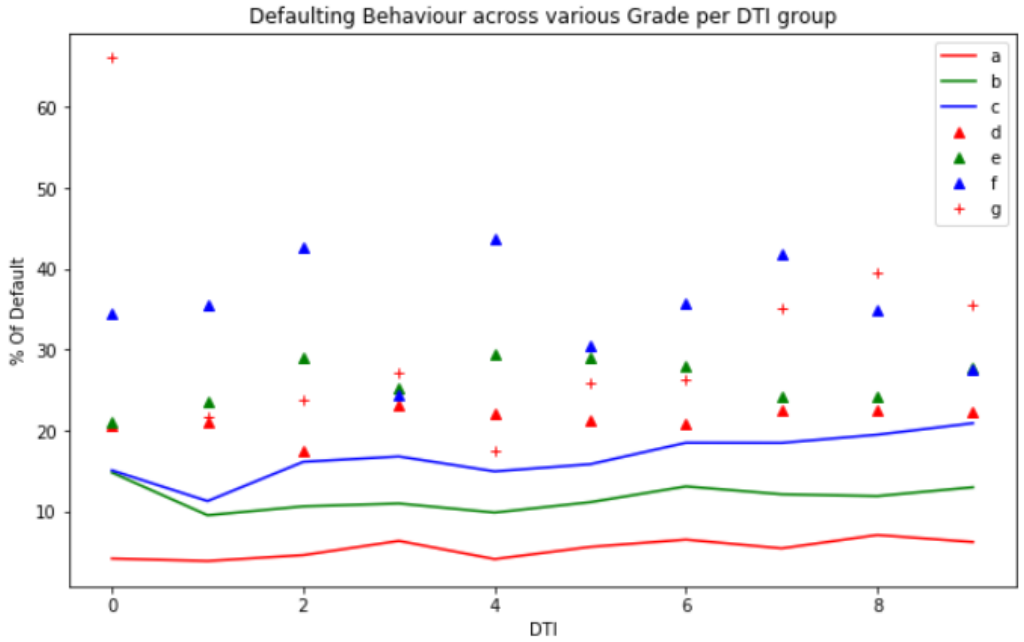
→ We can see from the above plot that for Grade E, F, G, likelihood of default increases in lower income groups.

→ As the loan grade goes up, chances of defaults go high as higher loan grade implies higher interest rate

→ From this we can suggest business to not take high risk with multiple drivers (Annual Income, DTI, Funded Amount, Purpose, Grade) .

Bivariate Analysis

- b) Grade and DTI Group



Observations

→ We can see from the above plot that for Grades E, F, G; likelihood of default increases in higher dti groups.

→ From this we can suggest business to not take high risk with multiple drivers (Annual Income, DTI, Funded Amount, Purpose, Grade) .

Conclusions

With the analysis performed on various parameters, following are the key observations:

- The attributes like Home Ownership, Employment Length, Address state do not strongly increase the probability of a loan being defaulted. We did see a slightly higher % of defaults in a few states. The business can review applications from those states with more scrutiny but there is no strong evidence that state results in Default
- The attributes like DTI (Debt to Income ratio), annual income, funded amount, loan, grade, purpose exhibit trend towards increasing the tendency of a loan being defaulted
- A higher grade implies a higher interest rate and so customer might not be able to repay. Likewise of a customer has a high dti (more debts to pay in a month), he may Default the loan
- With Bivariate analysis of Grade vs Annual Income, & Grade vs DTI, we saw low income and high grade plus high dti and high grade show a spike in default percentage.
- So, if a low income applicant with high DTI is funded a huge amount with a comparatively high interest rate, the business might go in loss as that might potentially result in a Default.

With all these observations, we can suggest business to not take high risk with multiple drivers (Annual Income, DTI, Funded Amount, Purpose, Grade) .