LENDING CLUB – CASE STUDY

SUBMISSION

Name: Omkar Joshi , Gadicherla Sameer

Batch – March 31’ 2021

Problem statement

- You work for 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.

- Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

- 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 aim of this case study.

- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Methodology

Data Arrangement & Cleaning → **Univariate Analysis** → **Bivariate Analysis** → **Recommendation**

Data Arrangement & Cleaning

- Drop all the rows with missing values
- Drop all the columns with single values and high missing percentages
- Impute values to necessary columns

Univariate Analysis

- Perform segmented univariate analysis and check the defaulting nature.
- Analysing the spread of each numeric variable

Bivariate Analysis

- Considering the major categories, work with other variables to get better insight.
- Checking average defaulting rate in each of the main category given a sub-category

Recommendation

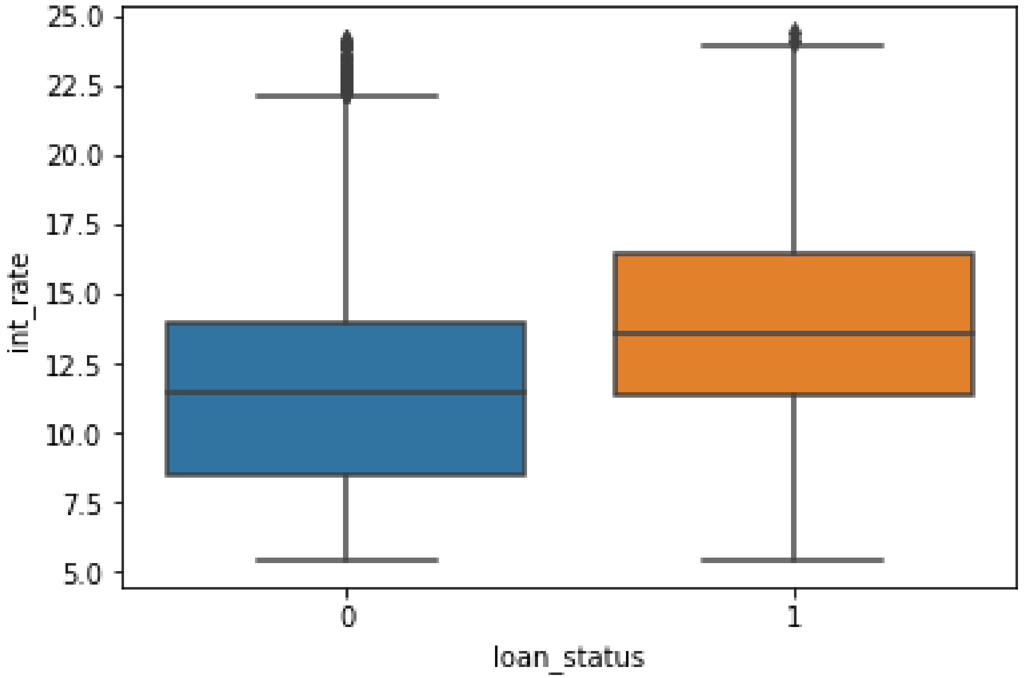
- Note down the spread in default percentage for each variable analysis and at the end, decide based on the variance of this percentage that who are the strong indicators.

Data Cleaning

- Total dataset contains 39717 rows and 111 columns.
- We found 54 such columns where all the values were null and removed those columns.
- Removed all the rows with missing last payment due
- Removed columns with single value throughout rows.
- dropped rows for "current" loan_status as they would not add much value to defaulting behaviour. Marked "charged off" as 1 and "Fully Paid" as 0 for simplifying our analysis.
- 'inq_last_6mths', 'initial_list_status', 'out_prncp_inv', 'total_pymnt_inv', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt', 'last_credit_pull_d', 'policy_code', 'application_type', 'member_id', 'funded_amnt_inv', 'pymnt_plan', 'url', 'title', 'zip_code', 'acc_now_delinq', 'delinq_amnt', 'funded_amnt', 'total_rec_prncp' are the droped columns.
- - Number of rows dropped: 1388(3.4%)
- - Number of columns dropped: 82(73.8%)

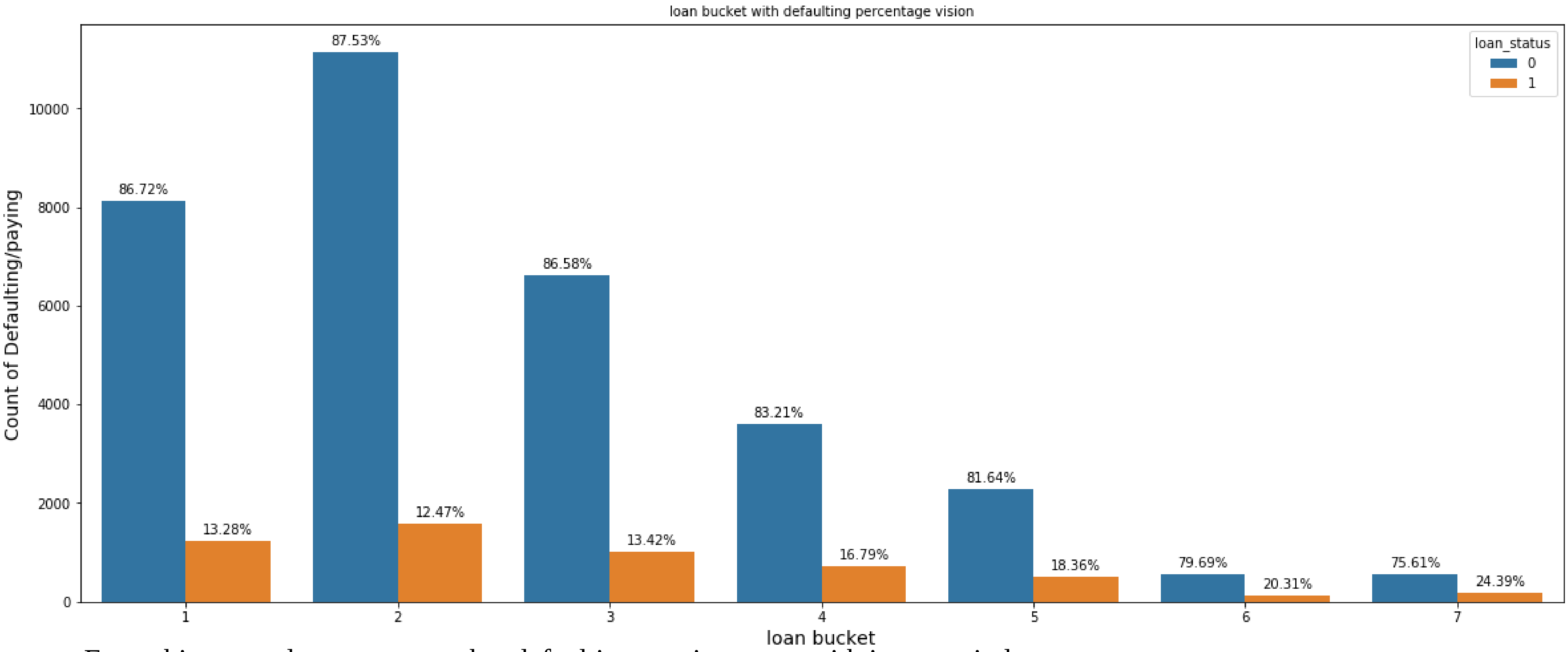
Univariate Analysis

- For univariate analysis we have used different types of plots which gave some meaningful insights.
- There are many plots drawn. We will show only unique plots in this ppt.



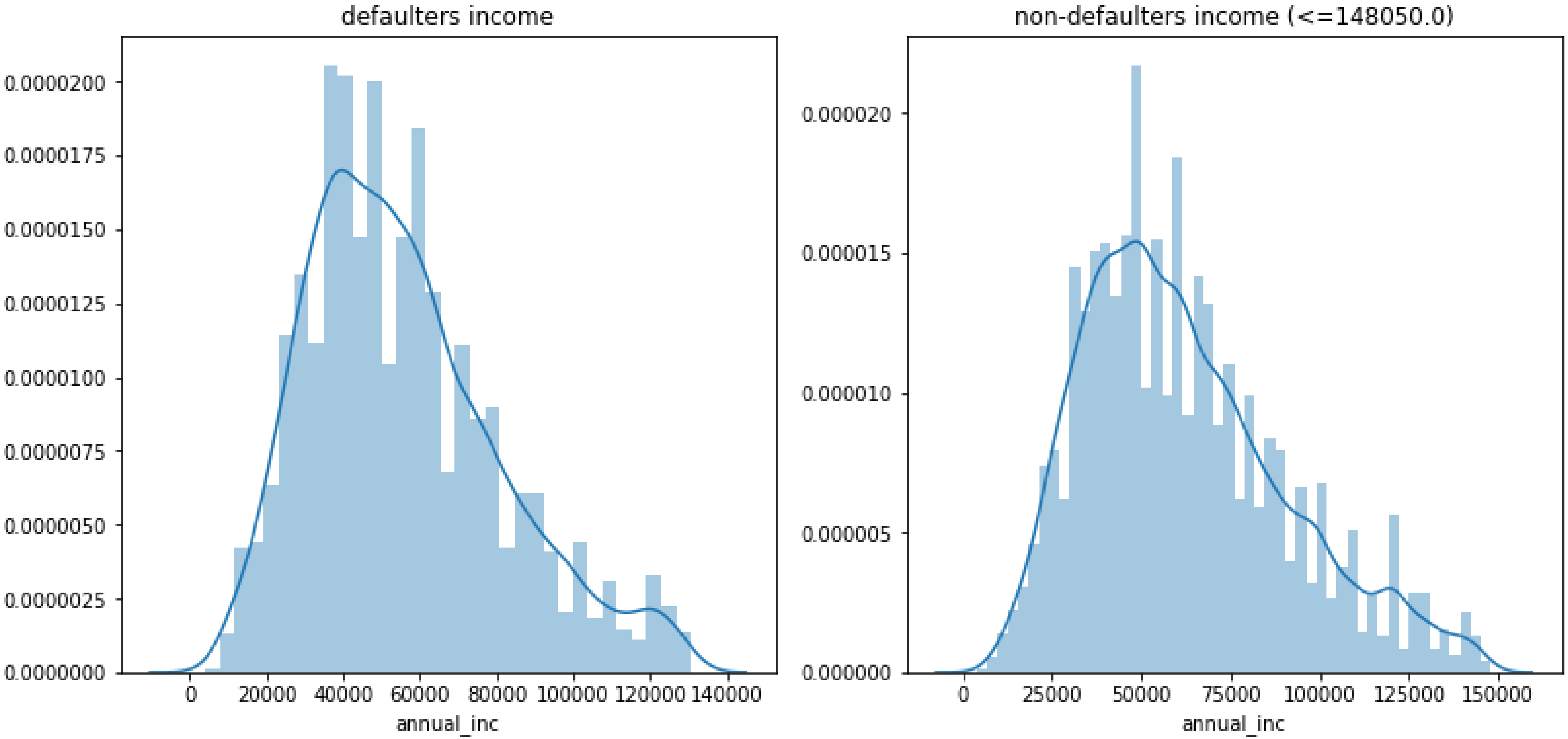
- From this boxplot we can see that defaulters have high interest rates compared to non-defaulters

Unviariate Analysis



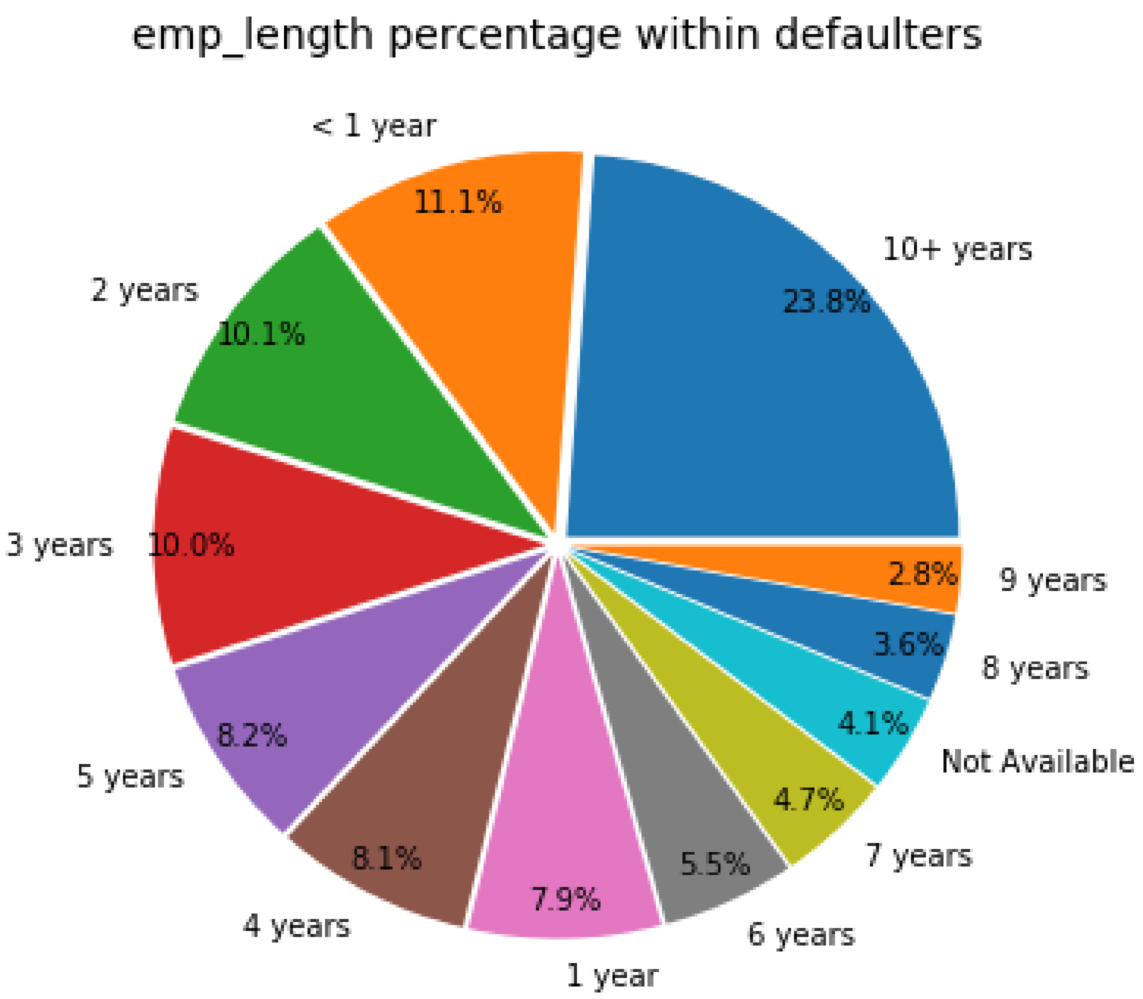
- From this countplot we can see that defaulting rate increases with increase in loan amount

Univariate Analysis



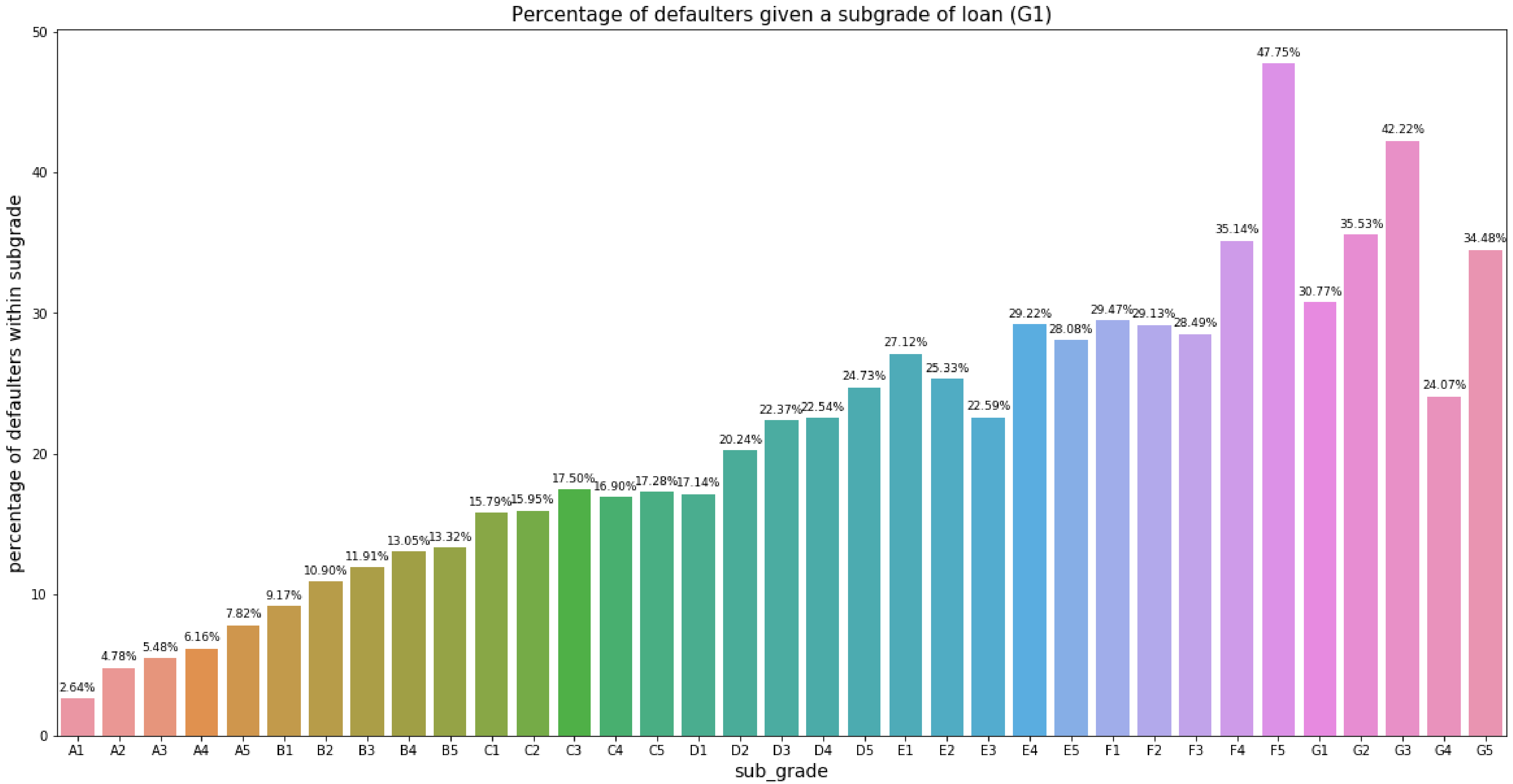
- From this histogram we can see there is normal distribution in income of defaulters and non-defaulters

Univariate Analysis



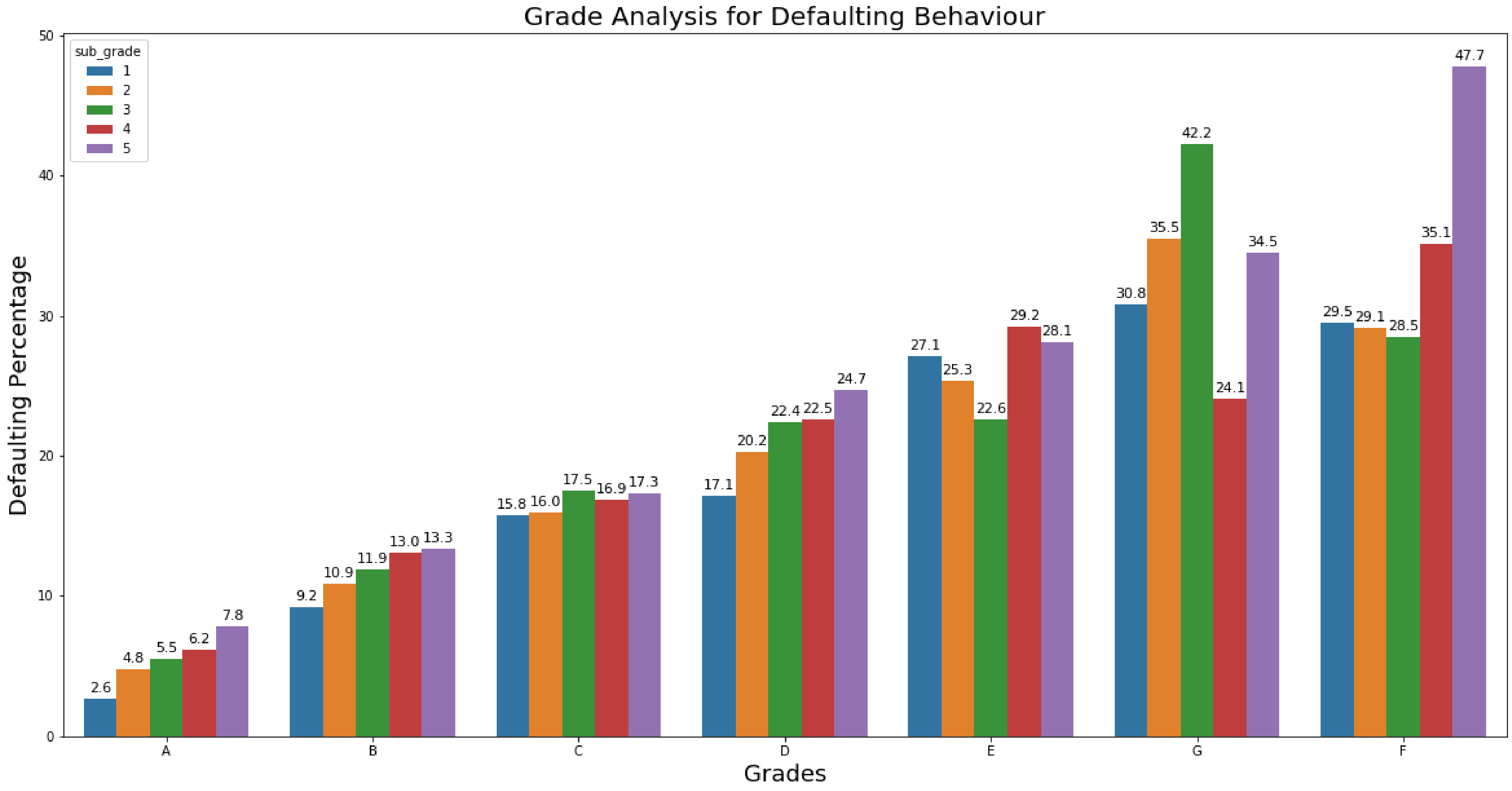
- From this pie chart we can see there are maximum defaulters over 10 years of experience

Univariate Analysis



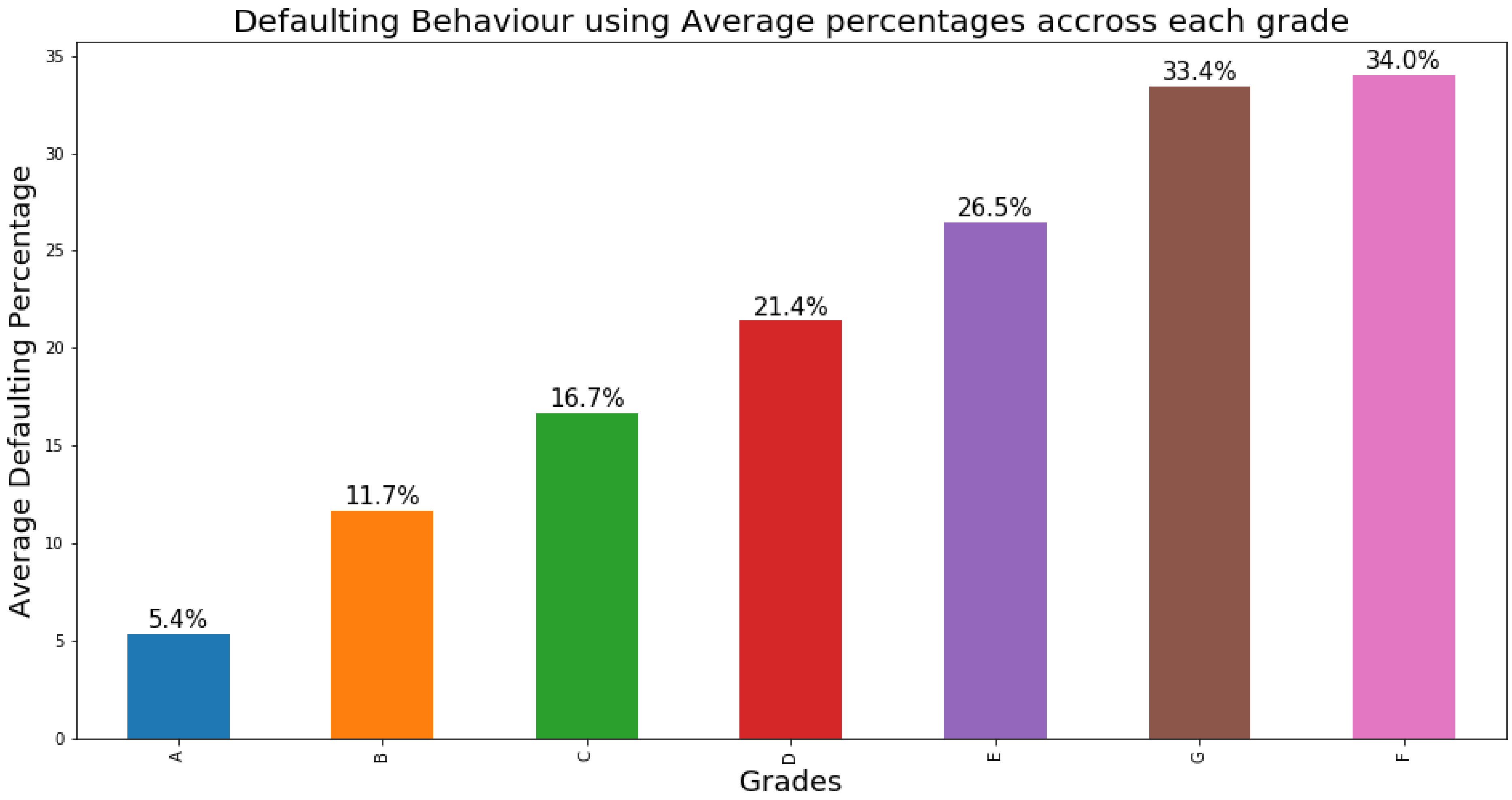
- From this bar chart we can see defaulting rate increases as grade increases. Within grade defaulting increases from 1-5 subgrade.

Bivariate Analysis



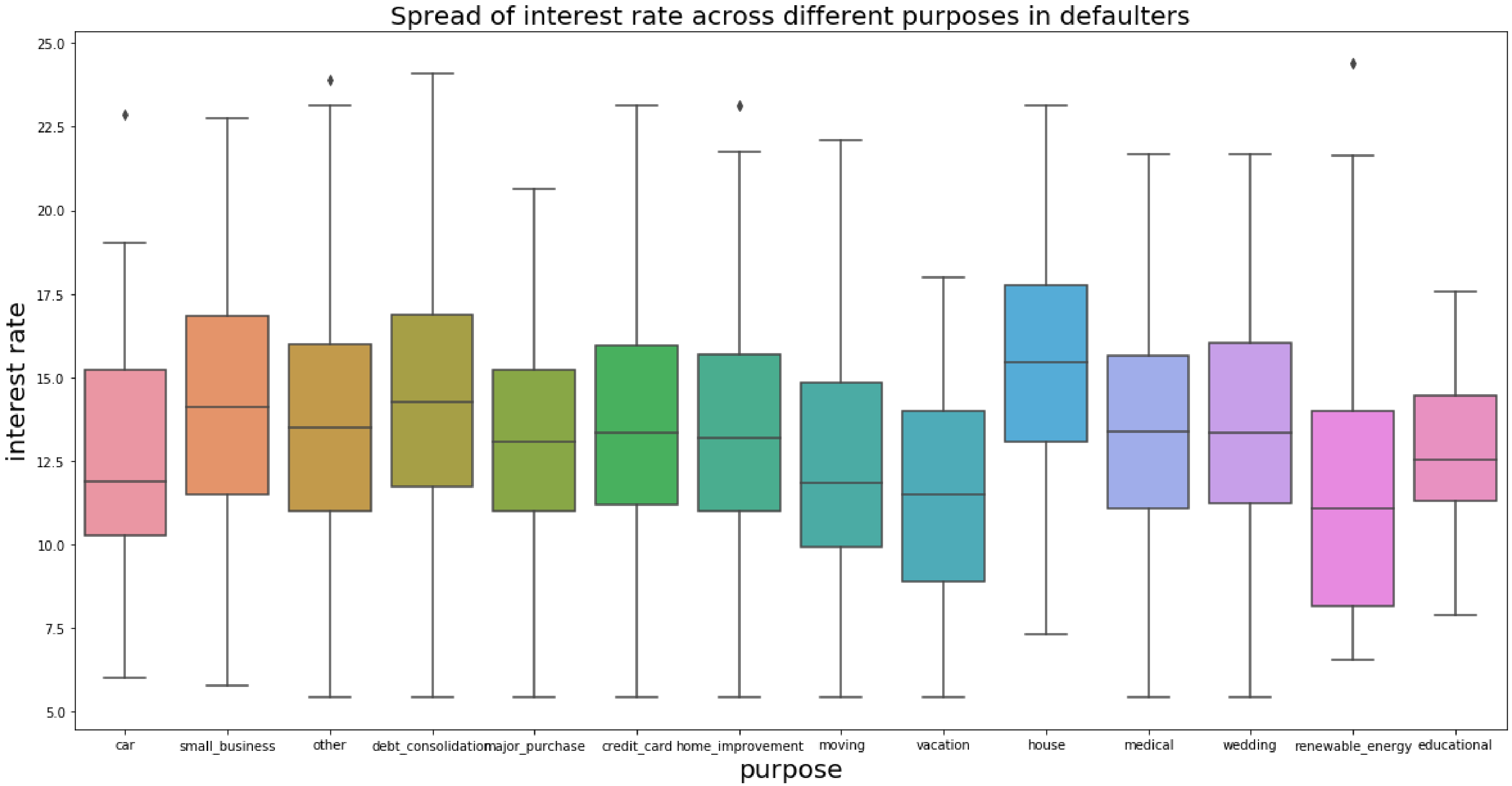
- From this bar chart we can see defaulting rate increases as grade increases. Within grade defaulting increases from 1-5 subgrade.

Bivariate Analysis



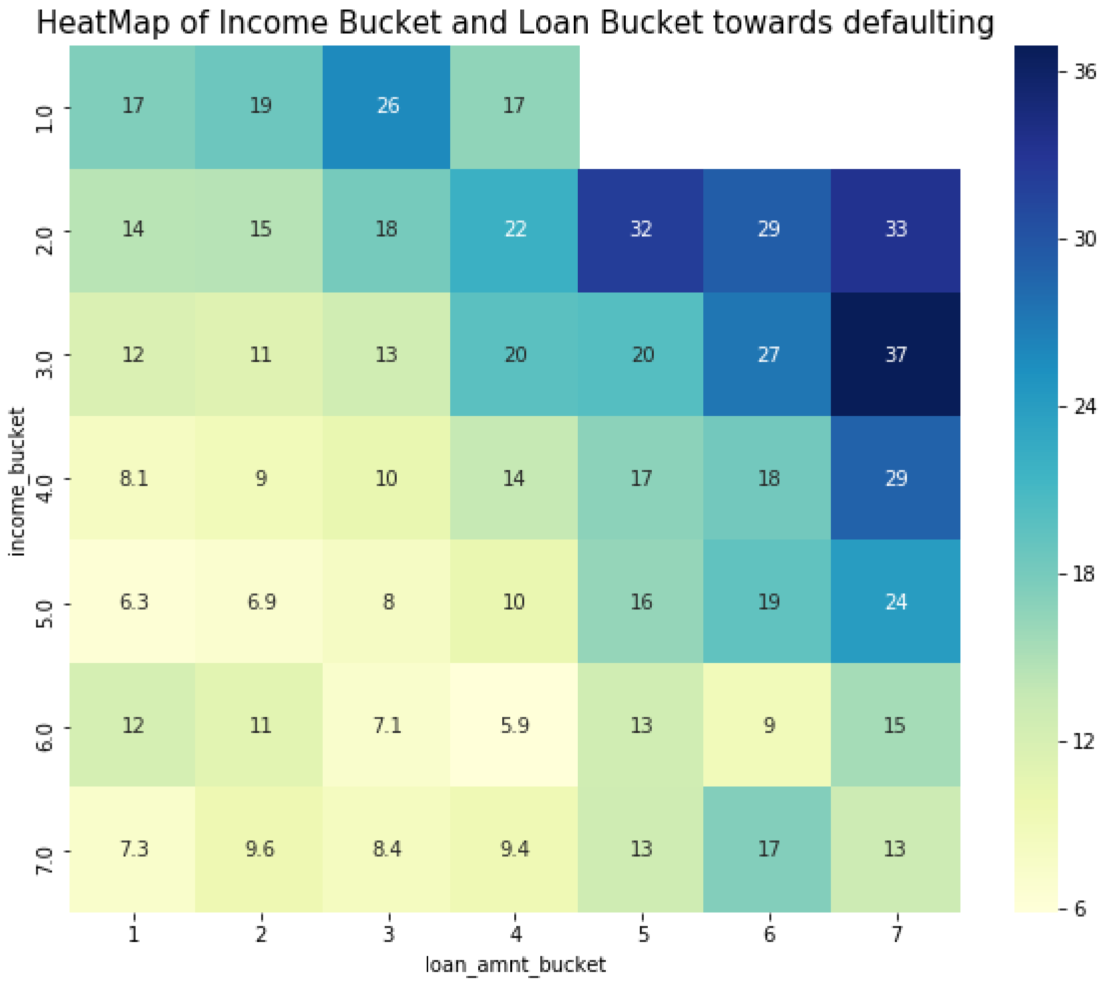
- From this bar chart we can see average defaulting rate increases a grade increases. Within grade defaulting increases from 1-5 subgrade.

Bivariate Analysis



- From this boxplot we can see defaulters who buy house loans have higher interest rates. Defaulters buying renewable_energy loans have very less interest rates.

Bivariate Analysis



- From this heatmap we can see defaulters whose incomelies between 50K to 75K and who purchase loan above 150K have highest defaulting rate.

Feature Importance

	default_variance
delinq_2yrs	50.00
sub_grade	45.11
grade_subgrades	28.62
grade	27.00
pub_rec_bankruptcies	26.21
pub_rec	22.22
interest_bucket	19.52
home_ownership	18.75
purpose_loan_buckets	17.25
purpose	16.43
purpose_verification	15.86
term	14.60
purpose_income	14.57
purpose_interest	14.06
income_loanamnt	12.88
loan_amnt_bucket	11.92
emplength_interest	10.54
emp_length	9.04
income_bucket	8.01
verification_status	4.64

- Based on these percentages, we get the top indicators of the analysis..
- These are the difference between max and min defaulting percentages in a particular category
- For bivariate, these are the average differences across the sub-categories.

Recommendations

Grade and Subgrade:

- These have shown a lot of variation in the defaulting percentages, thereby being a strong indicator.
- We can recommend to go with any subgrade within lower grades like A and B as we have seen the trend that as the grade increases,

the defaulting percentage also increases. - Within a grade, we can go for lower subgrade number for less defaulting percentage. A1 instead of A5 etc..

Interest Rate:

- When grouped into three categories, we have observed that the higher the interest rate, more the defaulters. - So its better to give loans in the normal or lower range. - recommended to keep the interest rate below 15% - Higher the interest , more the defaulting across all employment lengths.

Home Ownership:

- Though we see a variation of 18% in this, actually its because there are 0% defaulters in NONE category which has only 3 records. - If we ignore that we see a similar range of 14-18% defaulting in other categories like rental, mortgage, individual. - This is **NOT** a strong indicator

Purpose:

- We have received a lot of good insight with this particular column - When seen alone, - we saw that borrowers taking loans for **small businesses** end up defaulting around 25% of the times. - Where as giving loans in categories like **marriage, major_purchase, car and credit_card** have **very less risk.** - When looked along with some other categories: - *Verification-status:* **small businesses** tend to **default more**, so not recommended to give to them. - *Interest-Rate* : Lower the interest rate, less the defaulting percentage across all the categories. - *Loan-amnt* : We see that avg default rate in **major purchase** across all the loan buckets is **very less**. Small businesses tend to have more avg default rate. So avoiding small_business loans is better. - *Income-bucket* : **car loans have less default rates**. Small business end up defaulting more across all income buckets.

Recommendations

Term:
- It's recommended to **give loan for 3 years** instead of 5 years , as we see a default spike from 10% to 25% between these two.

Loan Amnt :
- We can observe more defaulting in **30000-35000** loan bracket(7) across all emp_length, recommended to give loans below 30000
- If the loan amount is below **15000**, then we see **less defaulting**.
- As the loan amount increases the defaulting rate is also increasing.

Employment length:
- defaulting percentage in 10 out of 12 categories is almost near to 13%
- When looked at only defaulter, we see that 10+ years had 23% of them.

Income Bucket:
- We see less defaulting in between **75000 to 150000** i.e., buckets **4-5-6, good to give loans in this range**.
- Higher the salary, lesser the default rate.
- Borrowers with income 5000-10000 are having the highest defaulting percentage of 23.4%, so better avoid loans in this range.\