LENDING CLUB CASE STUDY (ML 38 BATCH)

ASHIS DAS

VENKATA SAI KUMARI

CASE STUDY OBJECTIVE

- In this case study, we are presented with loan paid and defaulted dataset. Using this dataset, our objective is to **identify key variables that impact loan default.**
- These key variables can thereafter be used as indicators by the company to reject loans to ‘risky’ applicants.

APPROACH

Data Cleaning

Data Cleaning refers to activities such as:

- Removing columns/rows with all/max NA records
- Columns Formatting

And more...

Data Preparation

In this step, we create certain transformations on dataset such as:

- Creating Derived Columns
- Standardizing data if needed

Univariate Analysis

The analysis starts here with evaluating each numerical and categorical variable through bar/boxplots so as to understand their frequency count and descriptive stats

Segmented Univariate Analysis

In this step, each numerical/categorical variable is compared with the target variable to see significance of the variable

Bivariate/Multivariate Analysis

Here, two or more variables are combined to understand their combined effect on the target variable

DATA CLEANING

Managing Null records in Column

- Columns which have more than 10,000 NAs were deleted (**58 such columns)**

Single data in columns

- There were **9 columns** which had single unique value. They were deleted

Deleting non relevant columns

- Some columns such as transactional data don't add value to our analysis. Hence, **deleted 22 such columns**

Loan Status "Current" Removed

- We are only interested in fully paid and charged off values in Loan Status. **1140 "Current" records deleted**

Outliers Treatment

- If max of a numerical column is greater than 50 times the median of that column, then we have deleted last 1%ile of that column. **395 rows deleted**

After Data Cleaning, we are left with:

- Total **38182 rows and 22 columns**

DATA PREPARATION

Standardize Columns

- **Columns such as int_rate, revol_util and term** needed further cleaning to remove '%' and 'term' mentioned in those columns

Derived Columns - Date

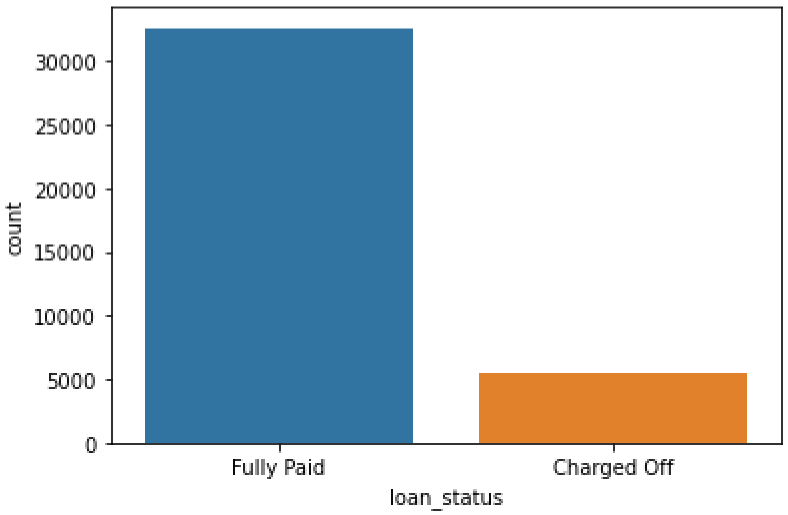
- The column **issue_d** was parsed to created two further columns of month and year

Derived Columns - Binning

- Certain numerical columns such as **'loan_amnt','int_rate', 'instalment', 'dti' and 'annual_inc' were binned**

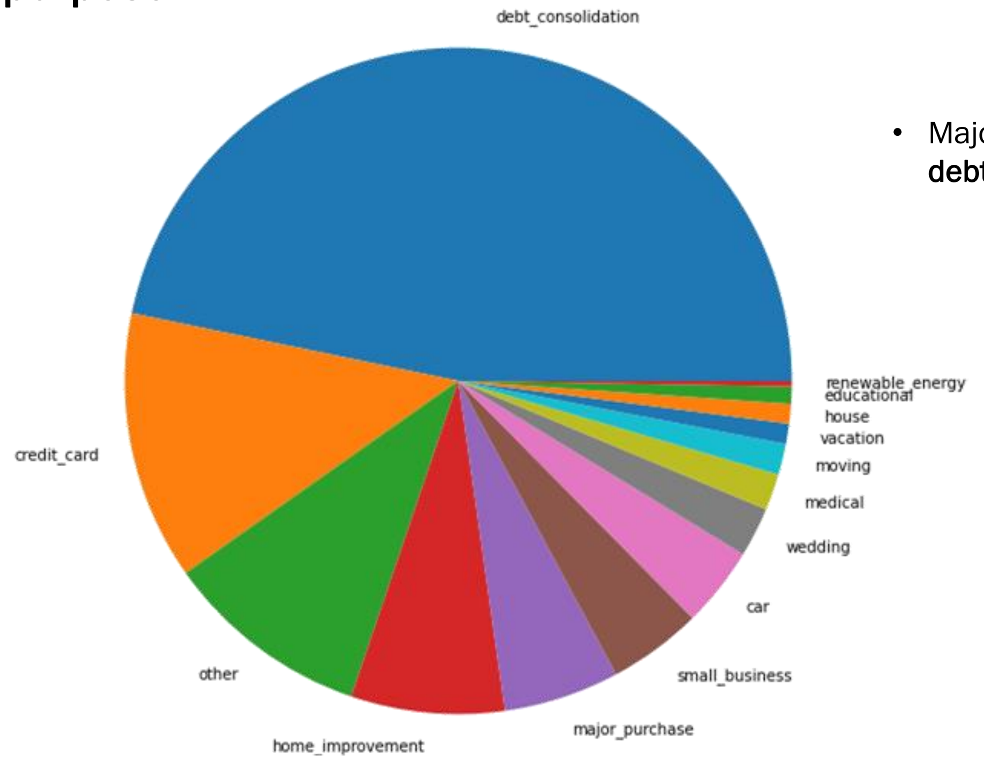
UNIVARIATE ANALYSIS - I

Target Variable: Loan Status



- Charged off records are almost **1:6** to Full Paid Record

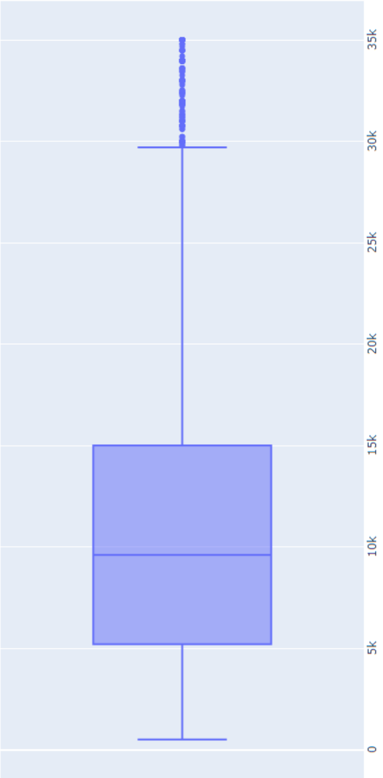
Pie Plots for analysing columns such as loan purpose



- Majority of loan applicants purpose is **debt consolidation**

UNIVARIATE ANALYSIS - II

Box Plot Analysis



- Descriptive stats such as mean, median, quartiles and SD for numerical variables such as loan amount, instalment, interest rate, dti etc. were understood **through box plot charts**

- Fig shows box plot for **loan amount**, where median value is 9600

Frequency analysis

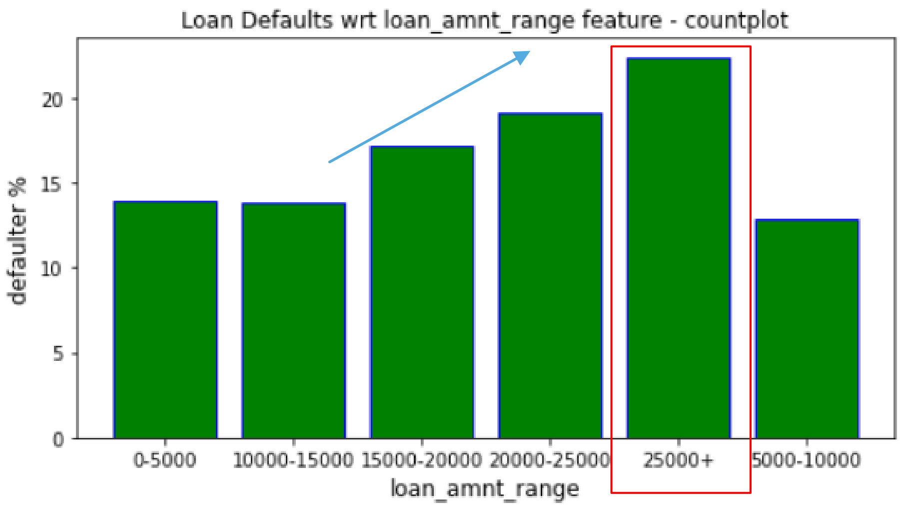
Variable	Max Frequency	Second Max Frequency
Home Ownership	Rent	Mortgage
Loan Amount	5000-10,000	0-5,000
Annual Income Range	25,000-75,000	50,000-75,000
Dti	10%-15%	15%-20%
Employee Length	10+	<1 year
Add_state	CA	NY
Grade of loan	B	A

- Frequency analysis provided a sense of data values present in each variable. **It also indicates any biases present in the sample**

SEGMENTED UNIVARIATE ANALYSIS - I

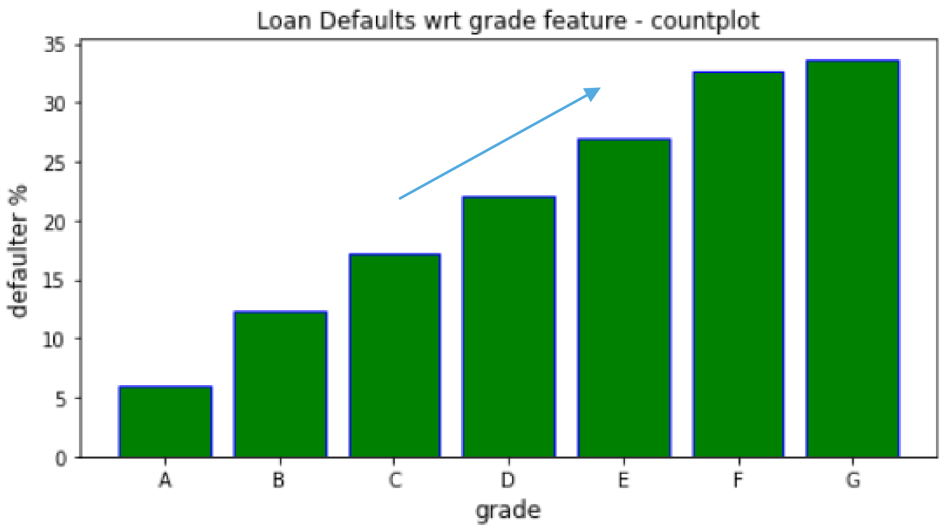
- This piece of analysis directly compares the target variable – loan status or default ratio (in our analysis) with one other variable. Key output of this analysis:
  - Identifying particular category of the independent variable, **where default ratio is the highest**
  - **Any trends** noticed between independent and Target Variable
- KEY Examples of segmented univariate analysis are shown below

Loan Amount vs Default Rate



With more loan amount range, **there is increase in default ratio.** Applicants which apply for more than 25,000 loan amount, their default ratio is more than **20%.**

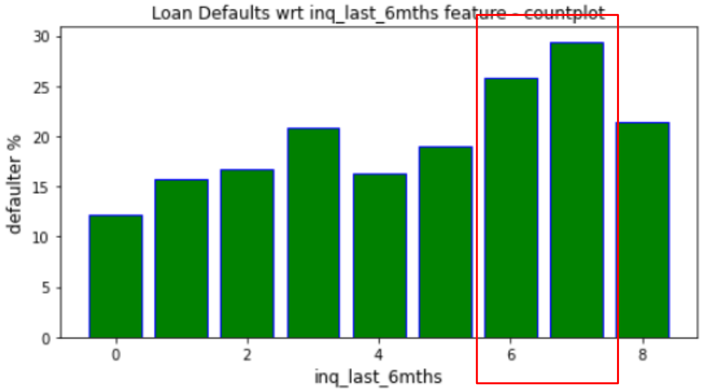
Grade vs Default Rate



With increase in loan Grade (A to G), default ratio increases from 5% to 35%

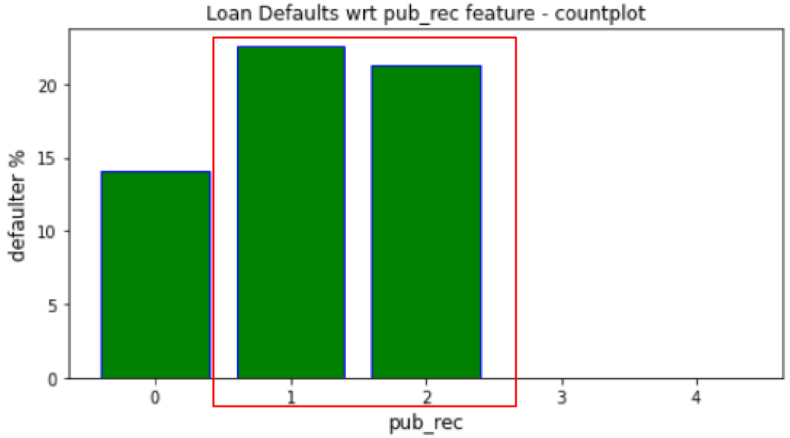
SEGMENTED UNIVARIATE ANALYSIS - II

Inquiries (Last 6 months) vs Default Rate



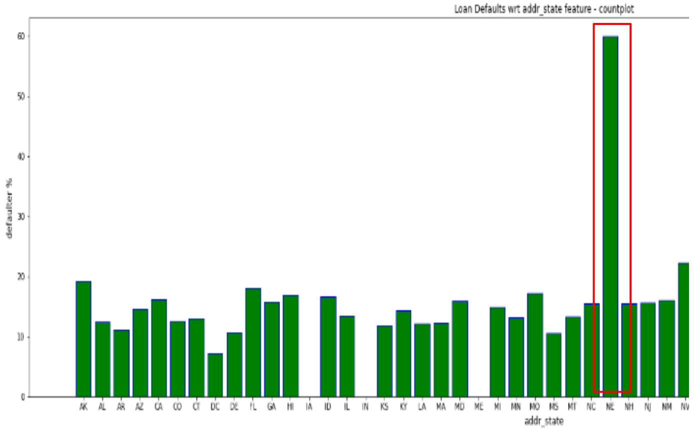
More than 6 inquiries in last 6 months could be an indicator for a risky profile.

Public Records vs Default Rate



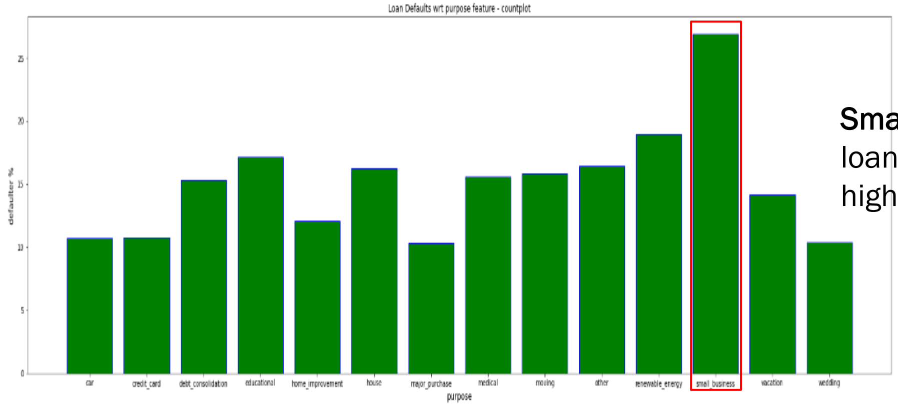
More than 1 public record can be seen as a good indicator to determine a risky applicant

Address State vs Default Rate



Applicants from NE Address State have 60% default rate

Purpose vs Default Rate

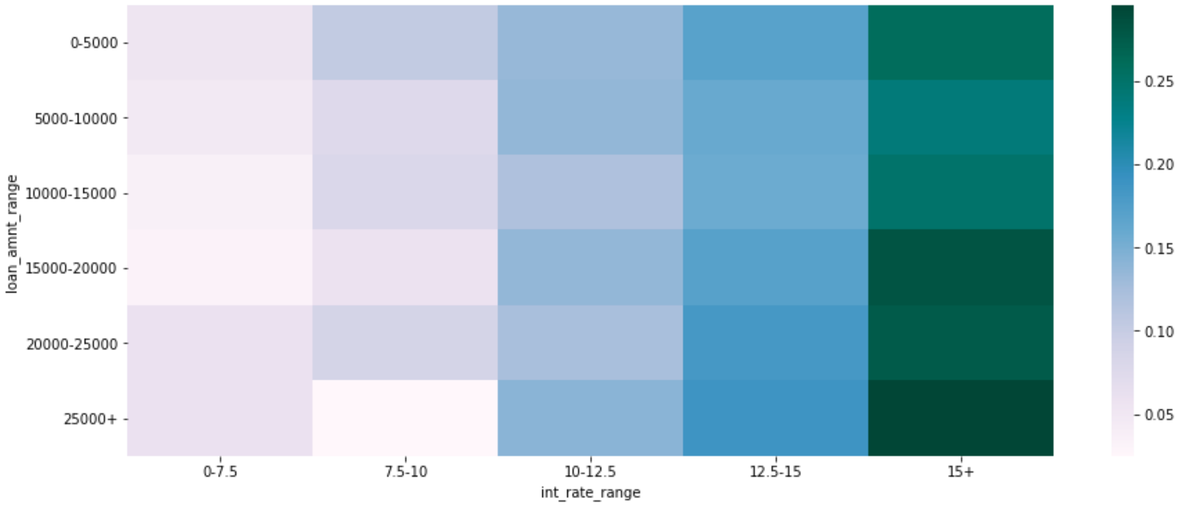


Small Business related loan purpose have the highest default rates

BIVARIATE ANALYSIS - I

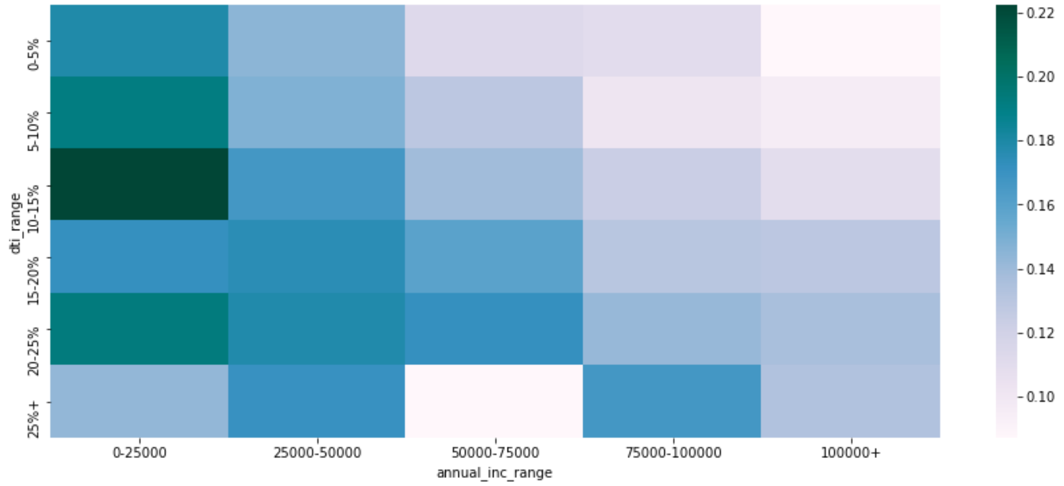
- This piece of analysis compares the target variable (loan status) vs combination of independent variables. Key charts have been shown in this slide and the next

Heat Map - Loan amount vs Interest Range ; Value = Default%



- Higher the interest rate with more loan amount, is an indicator of more risk for default.

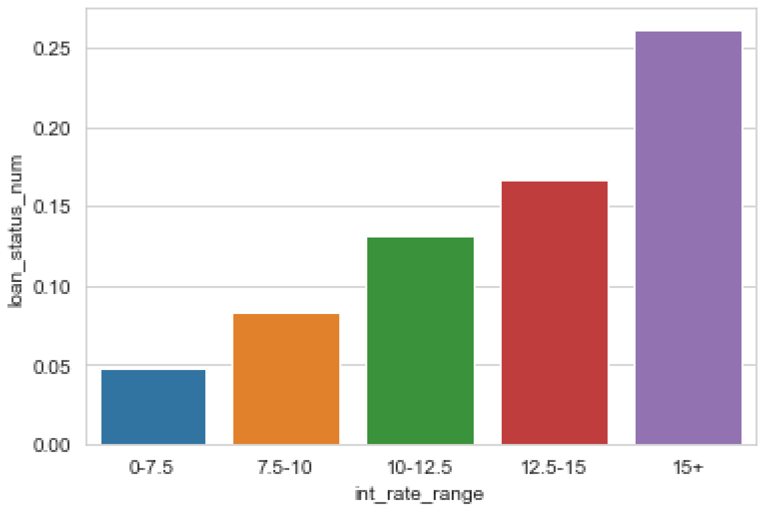
Heat Map - annual_inc_range vs dti_range ; Value = Default%

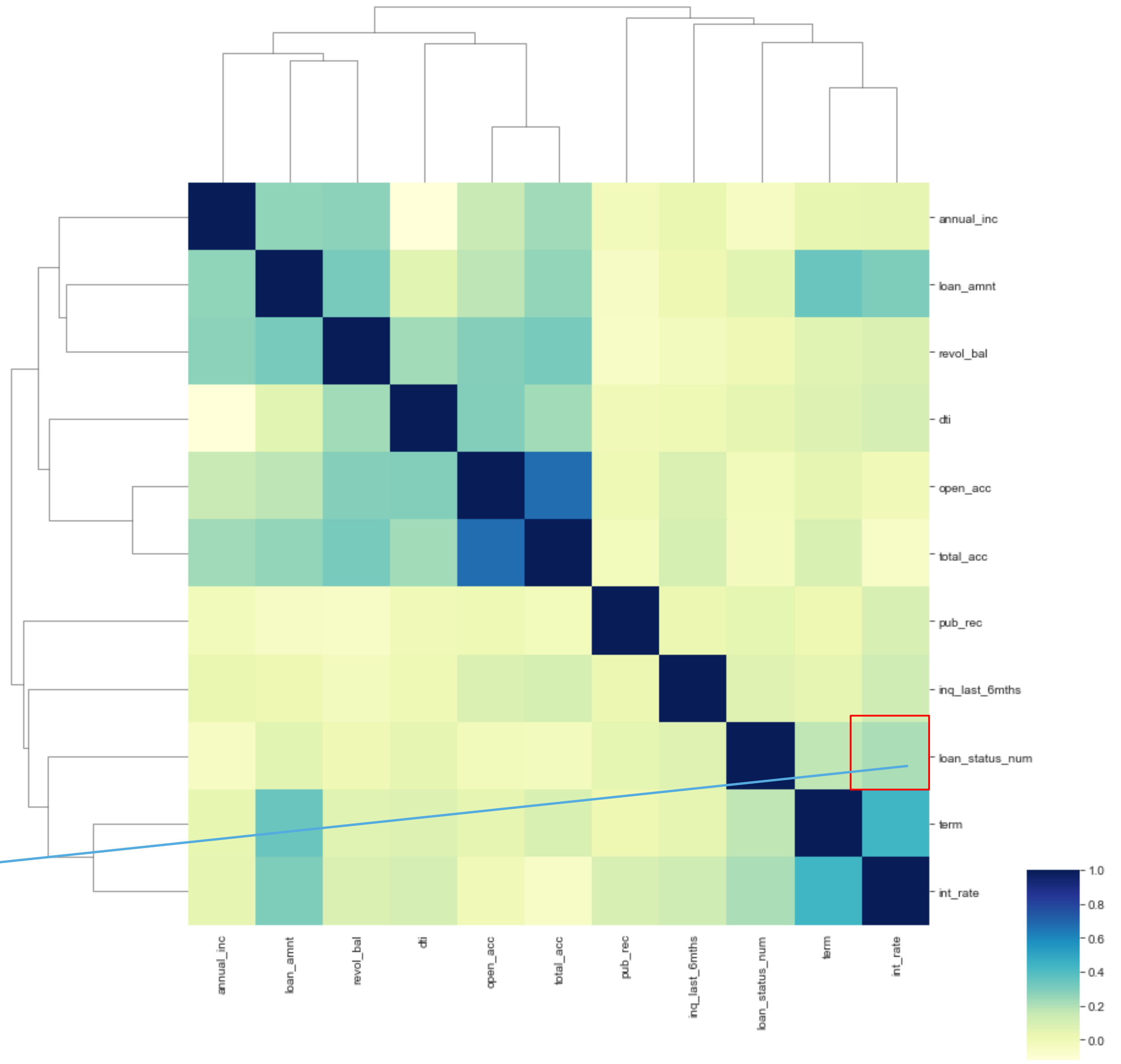


- A combination of dti range and annual_income, shows that for annual income below 25K, and dti 10% is an indicator of risky profile

BIVARIATE ANALYSIS - II

- **Clusters** have been created to group similar variables. For ex – loan amount & revol balance, and term & int_rate.
- **Open acc and total acc** are grouped under one cluster
- The correlation graph on the right for loan_Status shows moderate positive correlation with **int_rate, term and loan_amount**





CONCLUSION

After the data cleaning and exploratory data analysis, below are the key drivers for indicating if a loan applicant is likely to default:

1. **Loan Amount and Interest Rate Charged** – If the loan amount is greater than **25,000**, indicates a **20%-25%** probability of default based on historical data. The profile become more risky if interest rate charged exceeds 15%.

2. **Loan Term** – **60 months** loan term is more likely to not be paid compared with 36 months. It also has a correlation with point loan amount.

3. **Loan Purpose** – Small Business and Renewable energy related purposes tend to have more than **20%** default rate. This is followed up by loans taken for education or higher studies.

4. **Public Records** – Applicants with one or more public records need to be closely verified since they have higher defaults than candidates with no public records.

5. **State Address –** Data provides that applicants from NE state have a whopping **60%** default rate. So, a loan approver needs to be careful with applicants filing from that state.

6. **Loan Grade –** More the loan grade, higher the probability of defaulting. This is also observed in sub grade columns of Grade.

7. **DTI –** The debt to income ratio is also an important variable. Higher the ratio, more is the chance of defaulting.

THANK YOU

ASHIS DAS

VENKATA SAI KUMARI