Lending Club Case Study

Made by : Ramesh Krishnachari
Yash Gupta

Problem Statement

- The dataset given contains the information about past loan applicants and whether they ‘defaulted’ or not.
- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- 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.

Approach

Step 1 : Data Cleaning

1. Check the columns that contains more than 50% of missing values
   a. Delete these columns as they don't have sufficient data to perform data analysis
2. Dropping columns that will not effect analysis
   a. Columns where all rows contain same value(Single Valued). Eg : pymnt_plan
   b. Columns that denote ID of applicants. Eg : member_id
3. Dropping columns that are unnecessary based on business understanding
   a. Columns that denote customer behaviour as they are not available at the time of applying loan and thus they cannot be used as predictors for credit approval. Eg : earliest_cr_line
4. Dropping columns based on assumptions and will not effect analysis
   a. Columns that are of object data type and assumed not necessary. Eg : url
5. Dropping Rows where loan_status is current
   a. Since applicant is in process of paying loan these rows will not help in finding factors decide whether to lend money to borrower

Result 1 : *We have finally 38577 rows and 17 columns to be analysed*

Approach

Step 2 : Data Manipulation

1. Convert percentages stored as string to floating point number in column
   - int_rate
2. Extracting number of months as int from column
   - term
3. Check columns with less percentage of missing values
   - If their data type is object impute missing value with most frequently appearing value
   - If their data type is float or int impute with mean

Result 2 : *We can see there are no columns with missing values*

Step 3 : Outlier Detection

1. Plotting boxplots
2. Removing outliers using .quantile
3. Using .shape function on pandas dataframe to check rows having outliers were dropped

Result 3 :*We have 34640 rows and 17 columns to be analysed*

Approach

Step 4 : Derived Metrics

1. Type Driven
   1. Extracting month as int from issue_d column and storing in issue_month to find how months affect the loan status irrespective of year
2. Business Driven
   1. Making new column cia which stores total amount to be paid by borrower after the given term using compound interest formula
   2. amount_payable = principal amount(1 + interest rate) ^ time in years
   3. We want to analyse how amount to be paid at end of term effects loan status

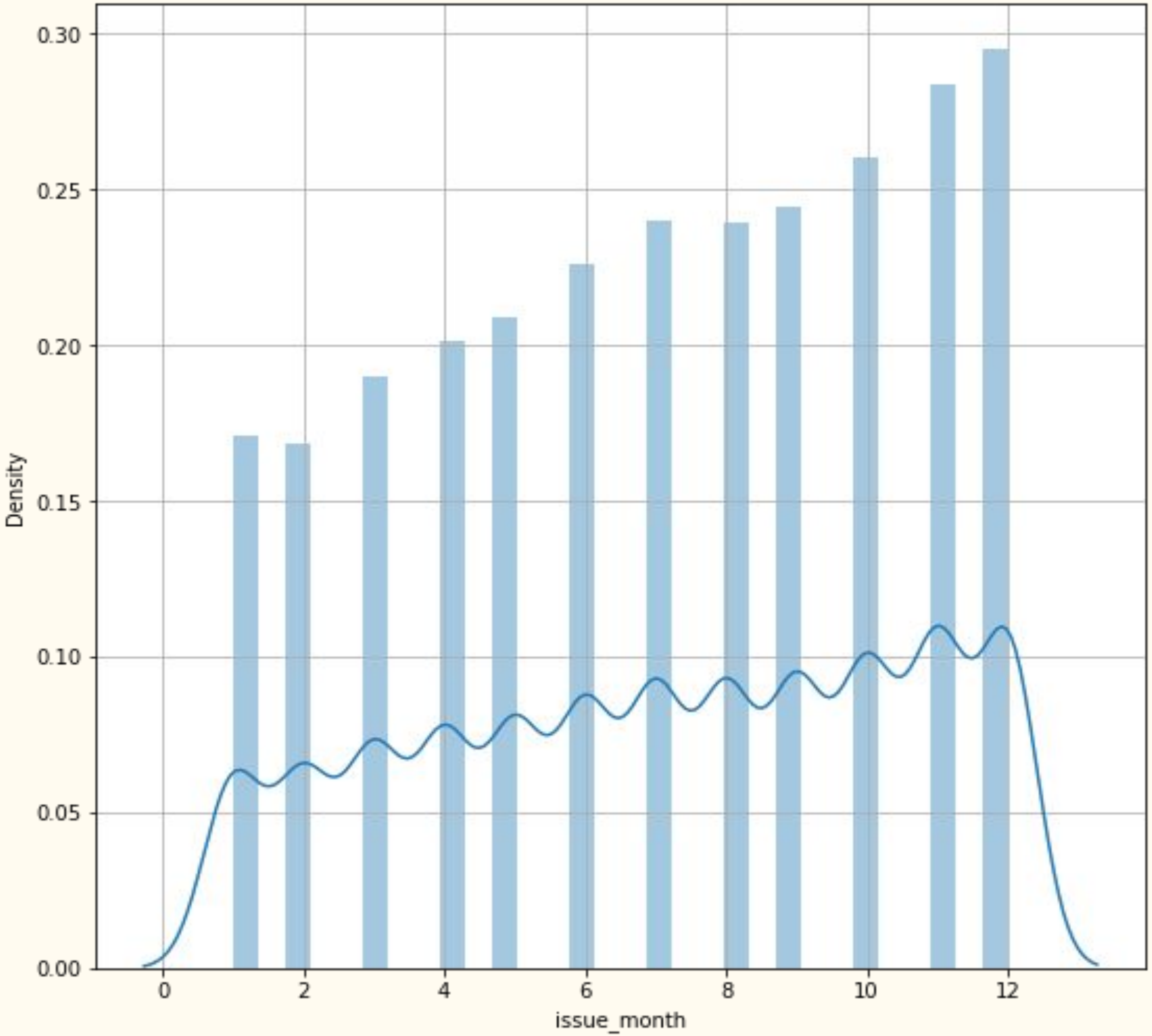
Result 3 :*We have 34640 rows and 19 columns to be analysed*

Step 5 : Data Analysis

1. Univariate Analysis
2. Segmented Univariate Analysis
3. Bi-Variate Analysis

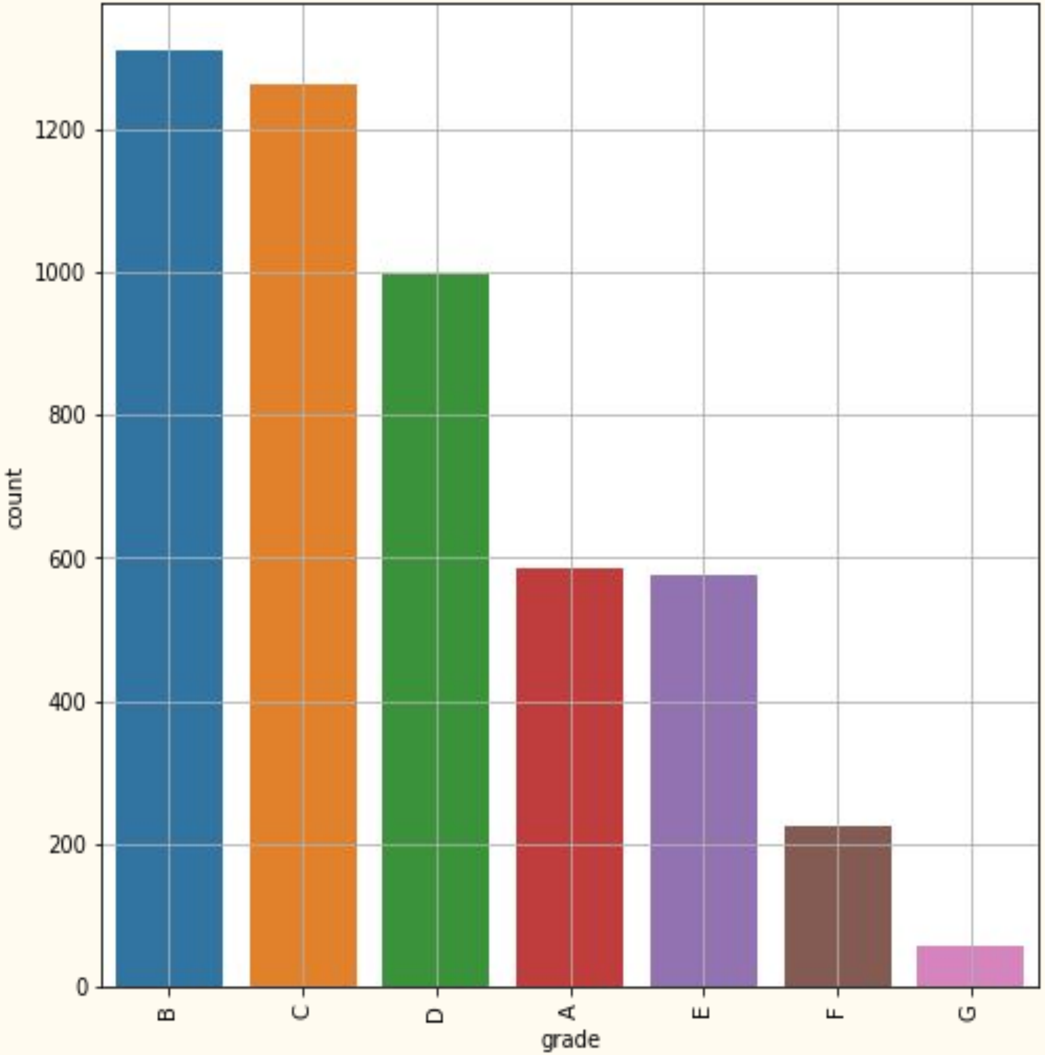
Univariate Analysis

1. Using data driven metrics we created new column issue_month which denotes the month which the loan was funded
2. Using seaborn library plotted distplot and passed above column as data
3. **Insight** : We can see most loan applications were funded in month of december.



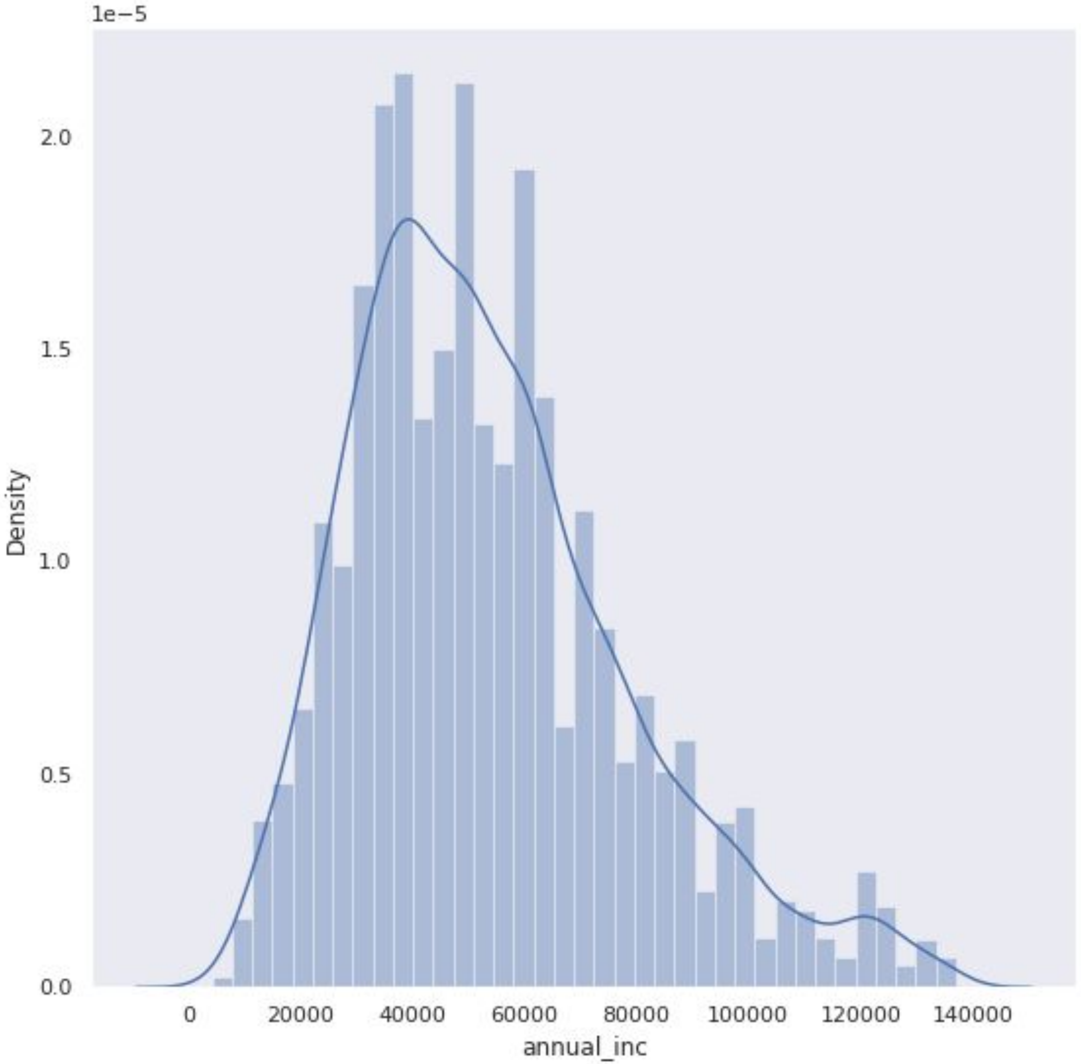
Segmented Univariate Analysis

1. We segmented dataset into two subsets.
2. First subset by name lcs_chargedOff contains rows for which loan_status is charged off.
3. From lcs_chargedOff we picked column grade and passed this column data to countplot present in seaborn library.
4. **Deciding Feature 1 : grade**
   a. Loan applications of grade C and grade D are very likely to default



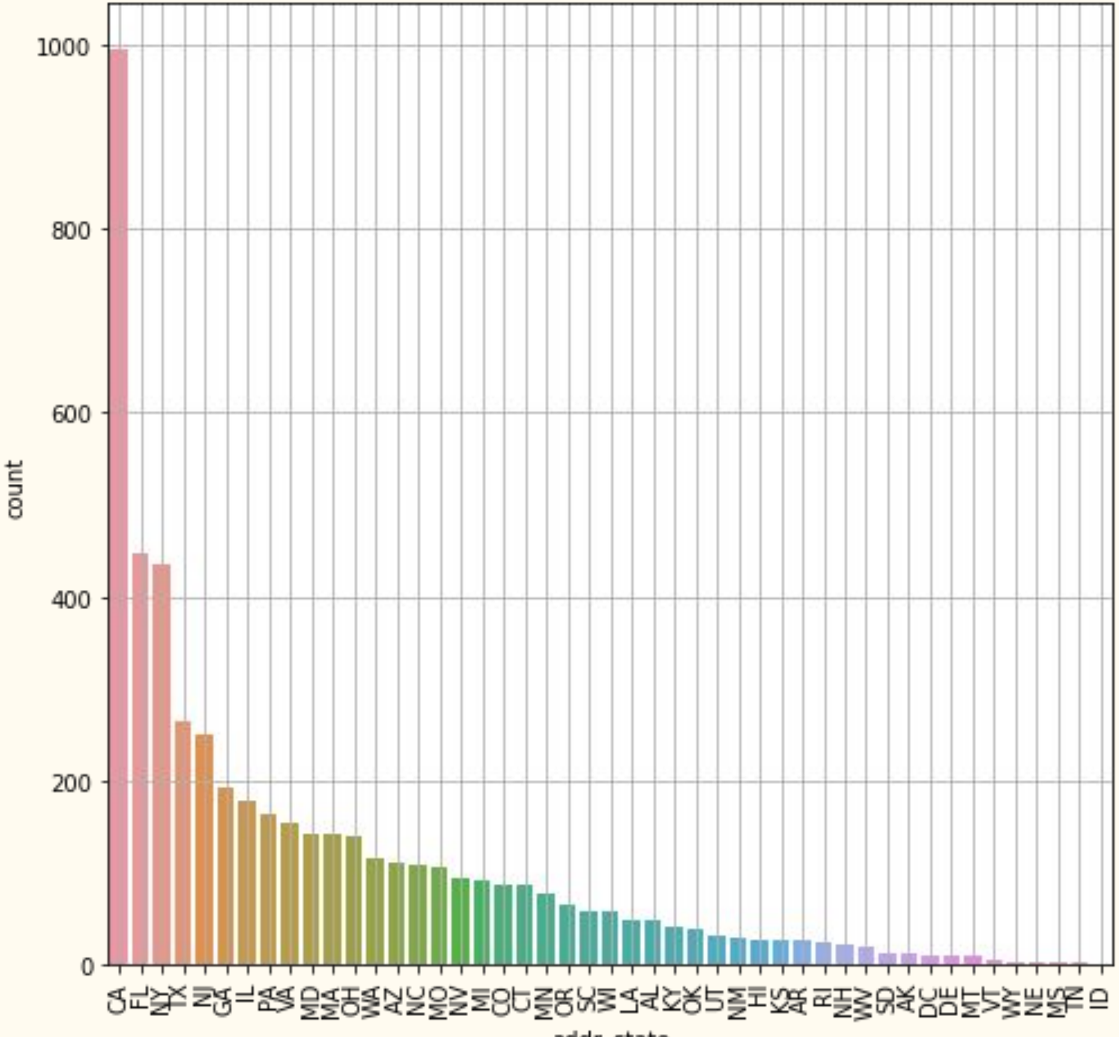
Segmented Univariate Analysis

1. We segmented dataset into two subsets.
2. First subset by name lcs_chargedOff contains rows for which loan_status is charged off.
3. From lcs_chargedOff we picked column annual_inc and passed it into distplot present in seaborn library.
4. **Deciding Feature 2 : annual_inc**
   a. Applicants with annual income 40,000 USD are very likely to default



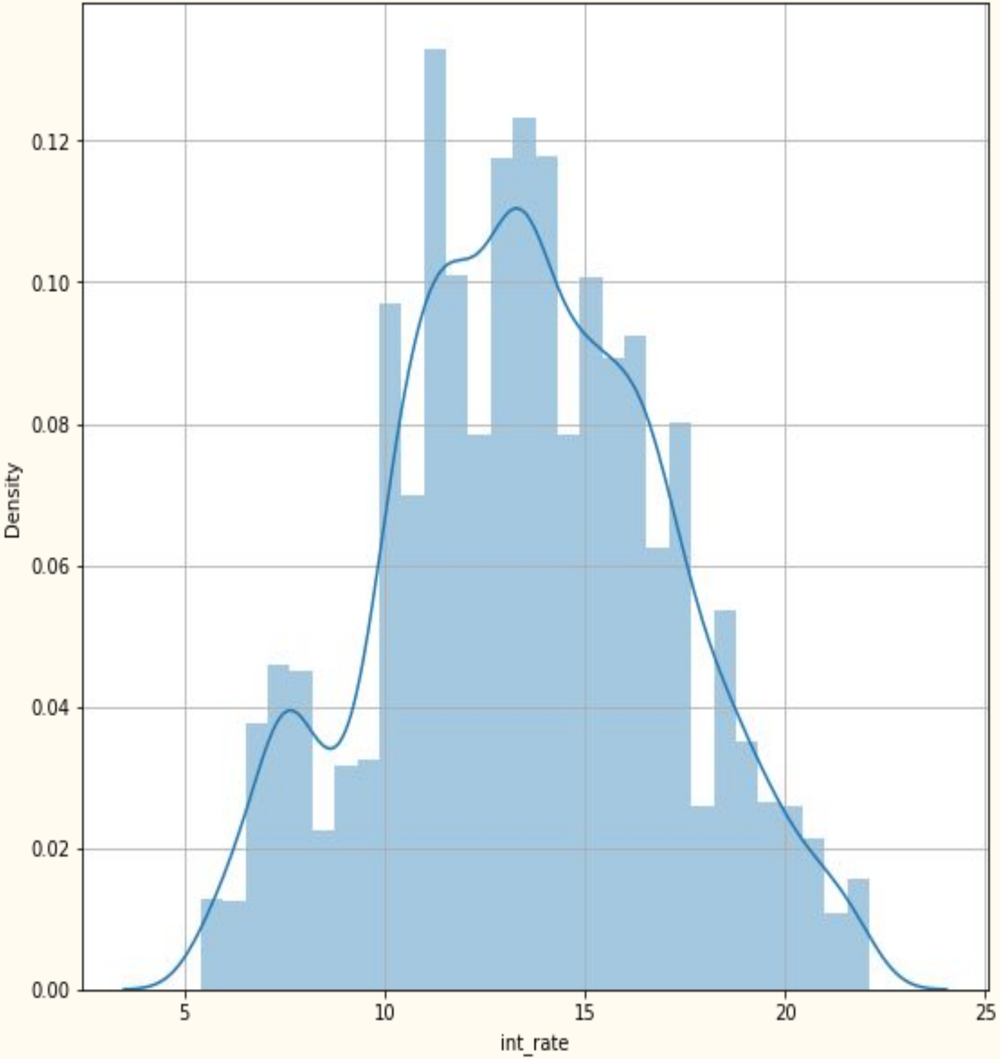
Segmented Univariate Analysis

1. We segmented dataset into two subsets.
2. First subset by name lcs_chargedOff contains rows for which loan_status is charged off.
3. From lcs_chargedOff we picked column addr_state and passed it into countplot present in seaborn library.
4. **Deciding Feature 3 : addr_state**
    a. Applicants from state of California and Florida will most likely default.



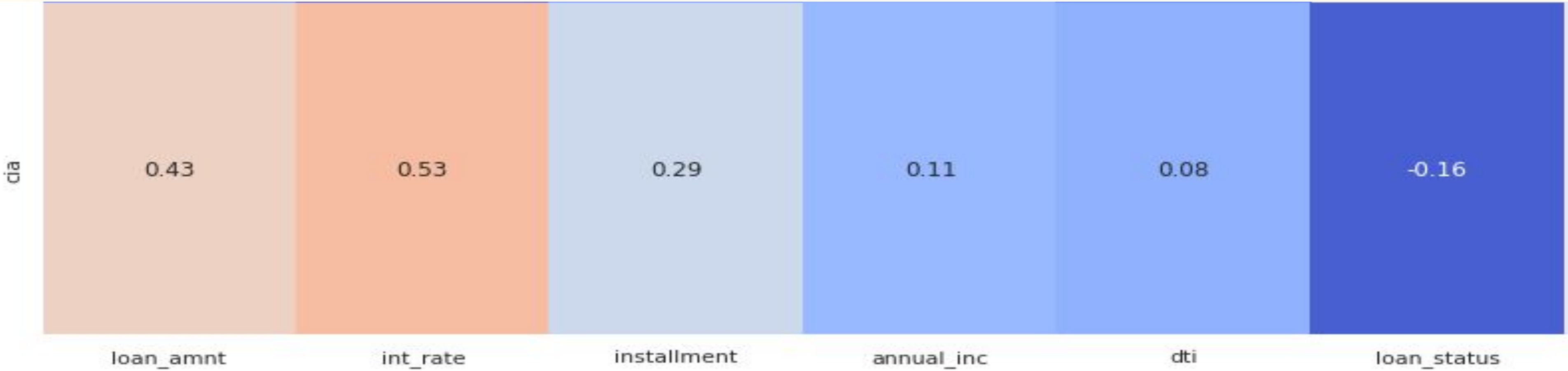
Segmented Univariate Analysis

1. We segmented dataset into two subsets.
2. First subset by name lcs_chargedOff contains rows for which loan_status is charged off.
3. From lcs_chargedOff we picked column int_rate and passed it into distplot present in seaborn library.
4. **Deciding Feature 4 : int_rate**
    a. Loan Applications with interest rate of approxly 12.5% will most likely default.



Bi-Variate Analysis

1. Using business driven metrics we created new column cia amount to be paid at the end of tenure.
2. Using seaborn library plotted heatmap which denotes correlation between loan characteristics variables
3. **Deciding Feature 5 : cia**
   a. We can see as the amount to be paid at end of term is negatively correlated with loan status.It means if amount to be paid is higher ,applicants will most likely default.



Conclusion/Recommendations

Conclusion 1 : Applications of low grade level should be avoided or should be given low priority.

Conclusion 2 : Applicants with annual income of 40,000 USD or less should be given loan at interest rate lower than 10%.

Conclusion 3 : People from state of California or Florida should be given loan only after thorough verification.