Lending Club Case Study

Manikandan Krishnamurthy Vembu

Deepenti Jha

Problem Statement

- This case study aims at providing an understanding of how real life business problems are solved using Exploratory Data Analysis (EDA) techniques.
- The domain of the case study relates to Banking and Financial Services industry
- The data provided is that of a financial institution – Lending Club (LC) who is the largest online loan marketplace, providing personal loans, business loans, and financing of medical procedures
- When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan will result in loss of business
  - If the applicant is not likely to repay the loan, i.e. (s)he is likely to default, then approving will lead to loss

Objective

- The data provided for this case study contains information about the past loan applications and whether they defaulted or not.
- We will be attempting to find if there are any patterns which indicate whether the applicant would default on the loan or not using EDA techniques. This information would then be used by the company to make more informed decisions while approving loans.
- The patterns or the driving factors (or driver variables) that lead to loan default is the primary goal of this case study.

Dataset Analysis

- Two files are provided as part of the data set:
  - Loan data set – a csv file that contains records of past loan applications and their status along with various other attributes
  - Data Dictionary – that contains what each column in the file represent
- The Loan Data Set contains details about post approved loan applications only; therefore rejections are not included.
- We will exploring the data and perform EDA analysis to attempt at finding patterns of defaulters.

Approach

- The raw data contains more than 39000 lines and 100+ column attributes for each record
- Filter rows and keep only required columns for the analysis

Rows Analysis

- There are no summary rows such as total , subtotal etc
- No Header and Footer rows are provided
- For this analysis, only rows with the categorical variable **loan_status** whose values are **Fully Paid** and **Charged Off** will be considered as our aim is to find defaulting patterns.
  Therefore, the rows with **Current** (the loans that are being paid and active) is not relevant and are removed.

Approach

Column Analysis

- Drop Columns
  - The columns in the dataset have more than 70% values as empty (NULL) or N/A. These would not be considered for the analysis
- Columns (**id, member_id)** are dropped as these are only identification columns and not used in the analysis
- Columns (emp_title, desc, title, url) are not relevant to the analysis and are dropped
- Following columns that represent customer behaviour are not relevant either and are dropped
  - (delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type)

Approach

Column Analysis

- In all only the following columns are considered for this analysis:
  - loan_amnt,funded_amnt,funded_amnt_inv,term,int_rate,grade,emp_length,home_ownership,annual_inc, verification_status, issue_d, loan_status, purpose,addr_state,dti
- Converted the following column values for the analysis

Column Name	Type of Manipulation
Term	Remove extra space character
Int_rate	Remove % from the values and store as float64 datatype

Approach

Column Analysis

- Add new columns

Column Name	Details
Issue_year	Stores year component(e.g. 2011) from issue_d column. The column is coverted to int64 format
Emp_length_c	Emp_length converted to integer and stored in this column as follows: < 1 year: 0 2 years: 2 3 years: 3 4 years: 4 5 years: 5 6 years: 6 7 years: 7 8 years: 8 9 years: 9 10+ years: 10

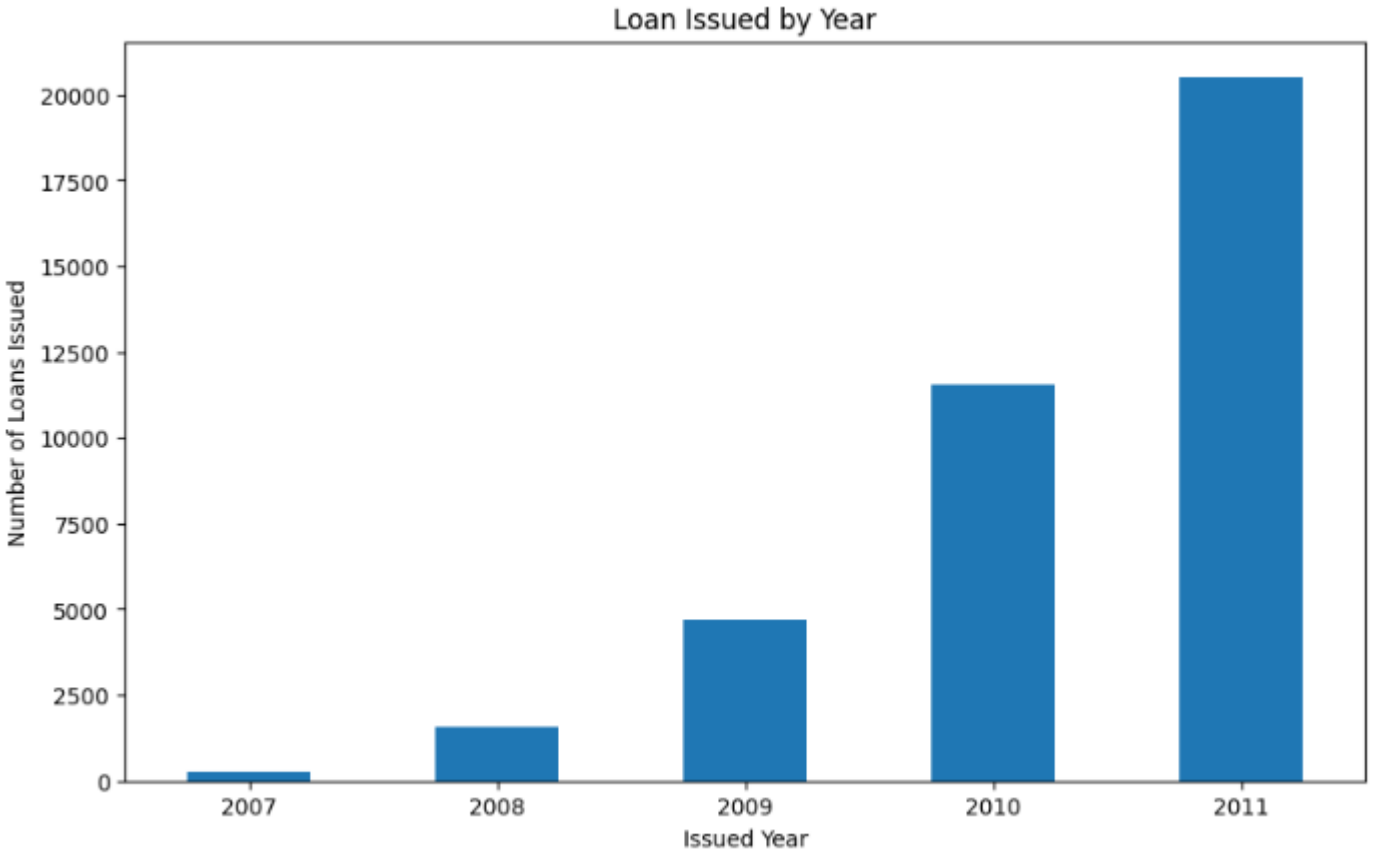
Case Study Analysis

With the data cleansed and the relevant columns only in the dataset now, the following Univariate and bi-variate/multi-variate analysis were done to identify patterns. Plots and observations of each of these are provided as well:

Univariate Analysis
Loans Issued by Year
Loans Issued by Term

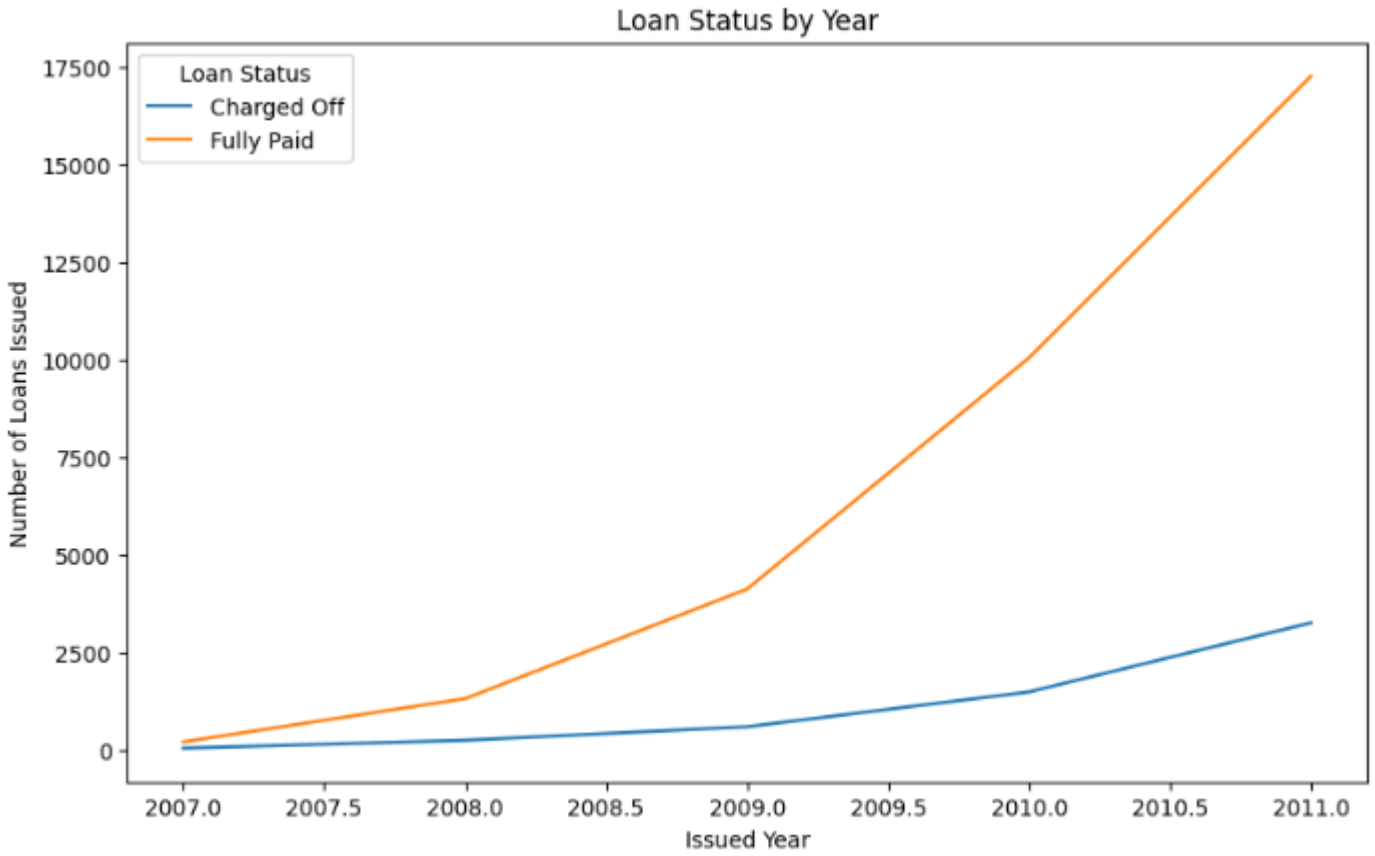
Bi-variate /Multi-variate Analysis
Loan Status by Year
Loan Status by Interest Rate
Loan Status by Grade
Loan Status by Home Ownership
Loan Status by Verification Status
Loan Status by reported purpose
Loan Status by State
Loan Status by DTI
Loan Status by Loan Amount Approved

Loans Issued by Year



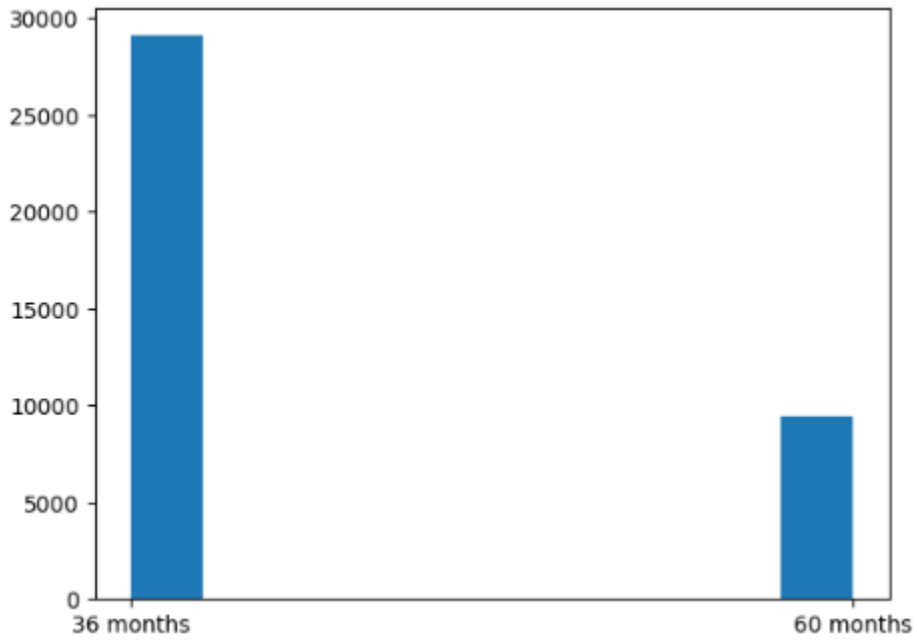
Observation: It seems the Lending Club company has been approving more number of loans over the years.

Loans Status by Year



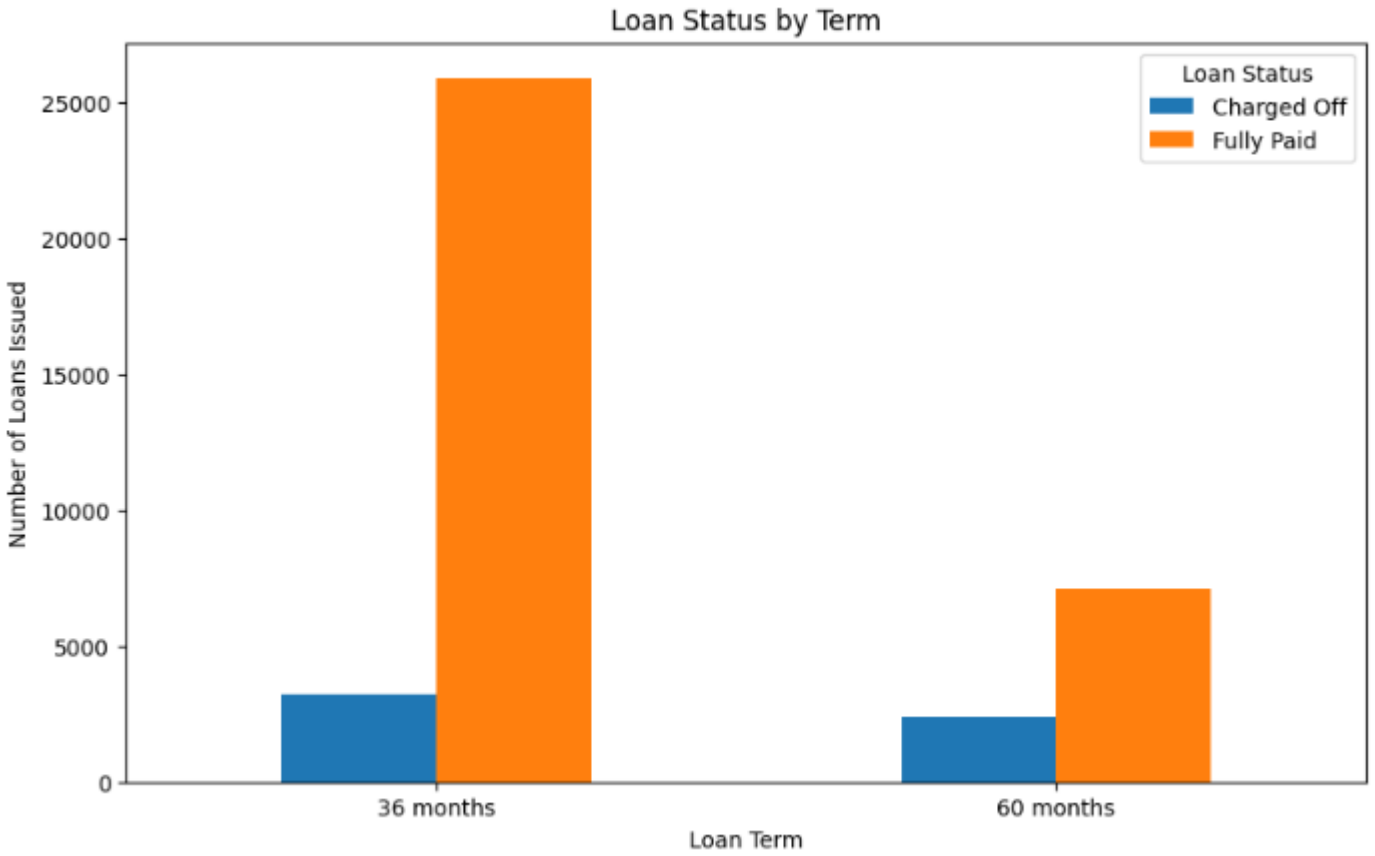
Observation: From the chart above, it is clear that the number of defaulters are also on the rise over the years. Though the Fully Paid loan numbers are substantial over the years, the company should definitely should take steps to bring the defaulter numbers down.

Loans Issued by Term



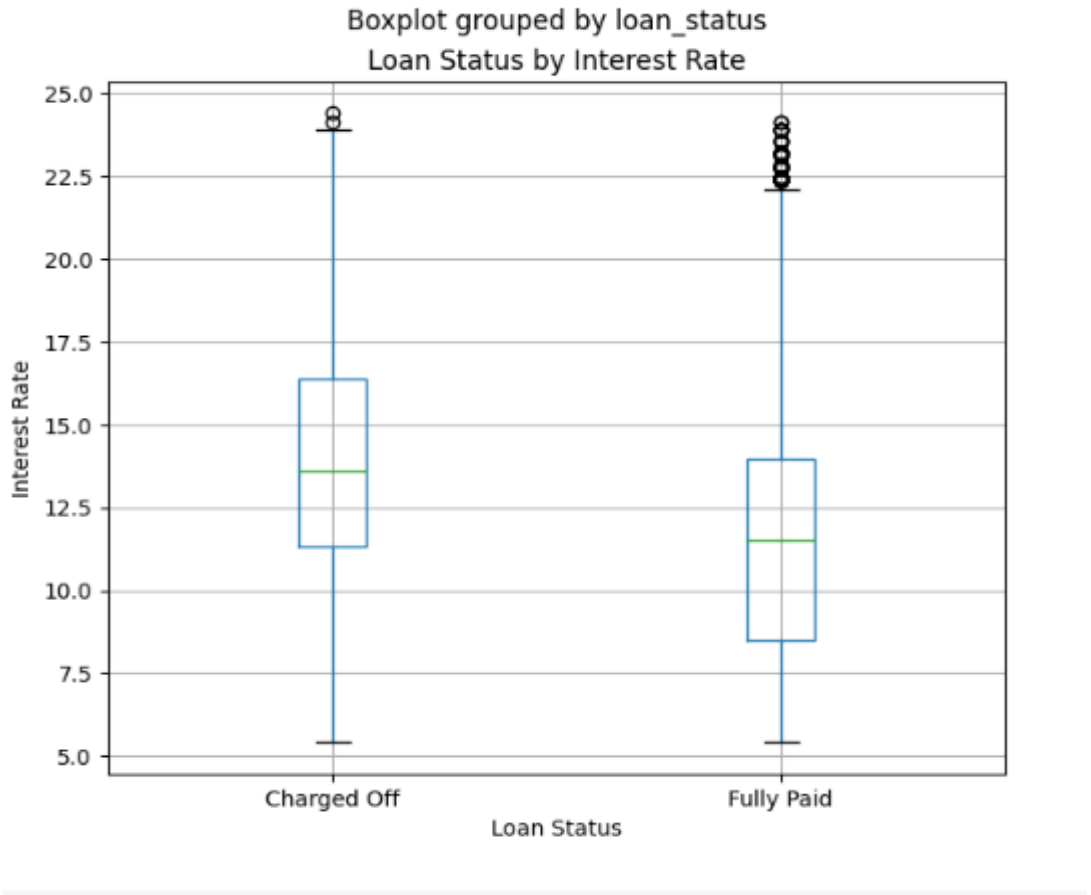
Observation: From the data set, there are only two term values – 36 months and 60 months tenure. The number of loans approved for 36 month tenure is very high, indicating that many applicants have preferred to have shorter time duration, most likely, due to loans being availed for personal needs.

Loans Status by Term



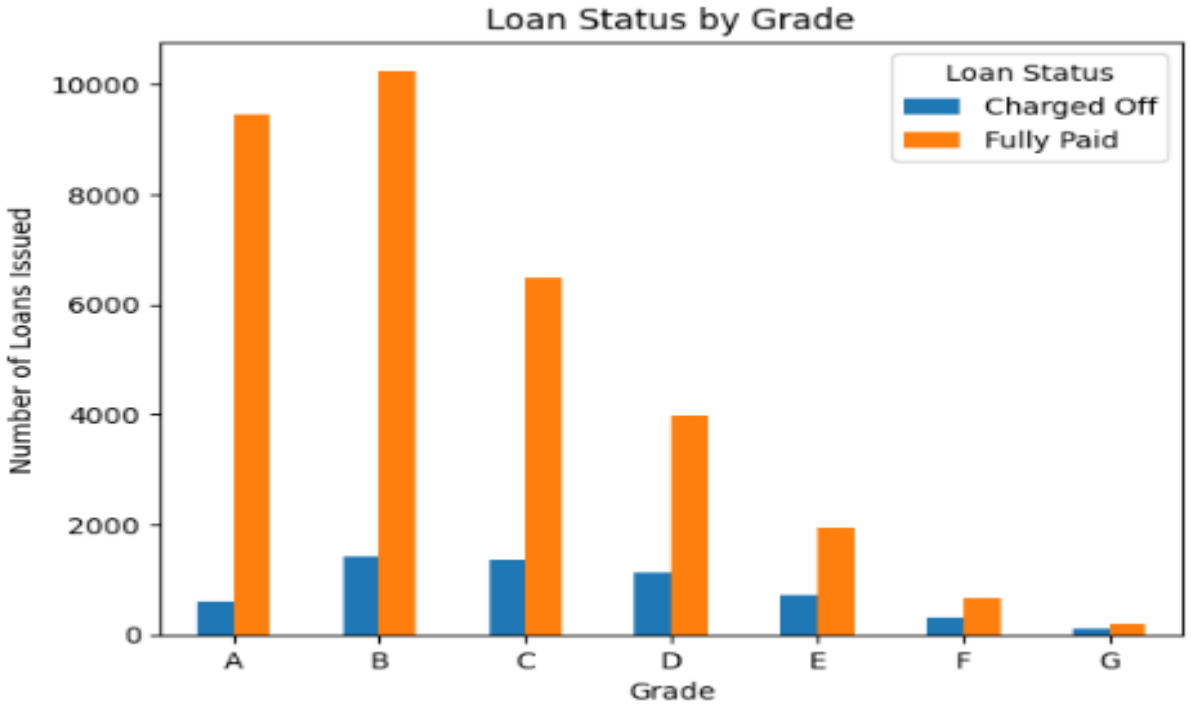
Observation: From the Loan status perspective, we can see that the defaulters on the 36 months are slightly higher than the ones that have taken for 60 months duration.

Loans Status by Interest Rate



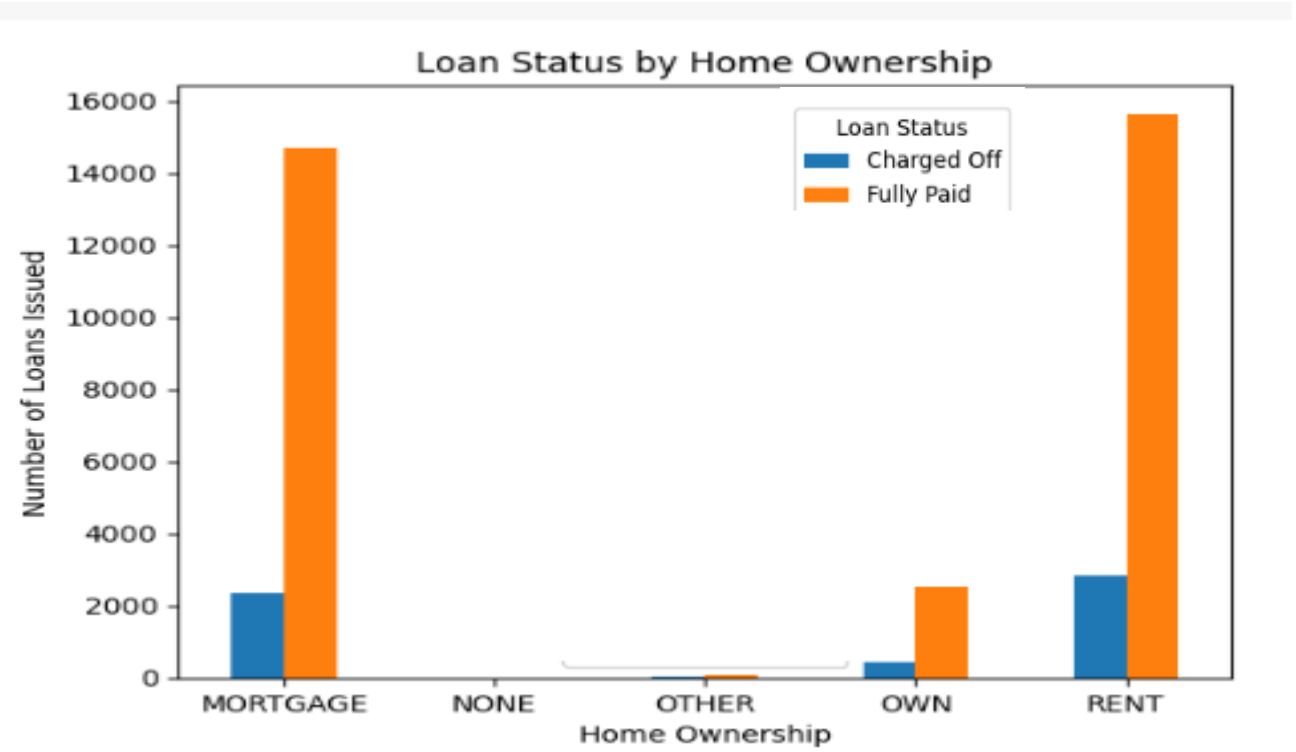
Observation: We did a box plot chart to find how loan status stacked up to interest rates. The defaulters had a median interest rate of 17% which is higher than 11% for the Fully Paid applications. It seems there has been some indicators that these people would default on the loan and hence had been charged a higher interest rate at the time of approval. Another observation is that there are outliers on the Fully Paid applications as well, probably first timers with higher interest rates but have eventually closed them.

Loans Status by Grade



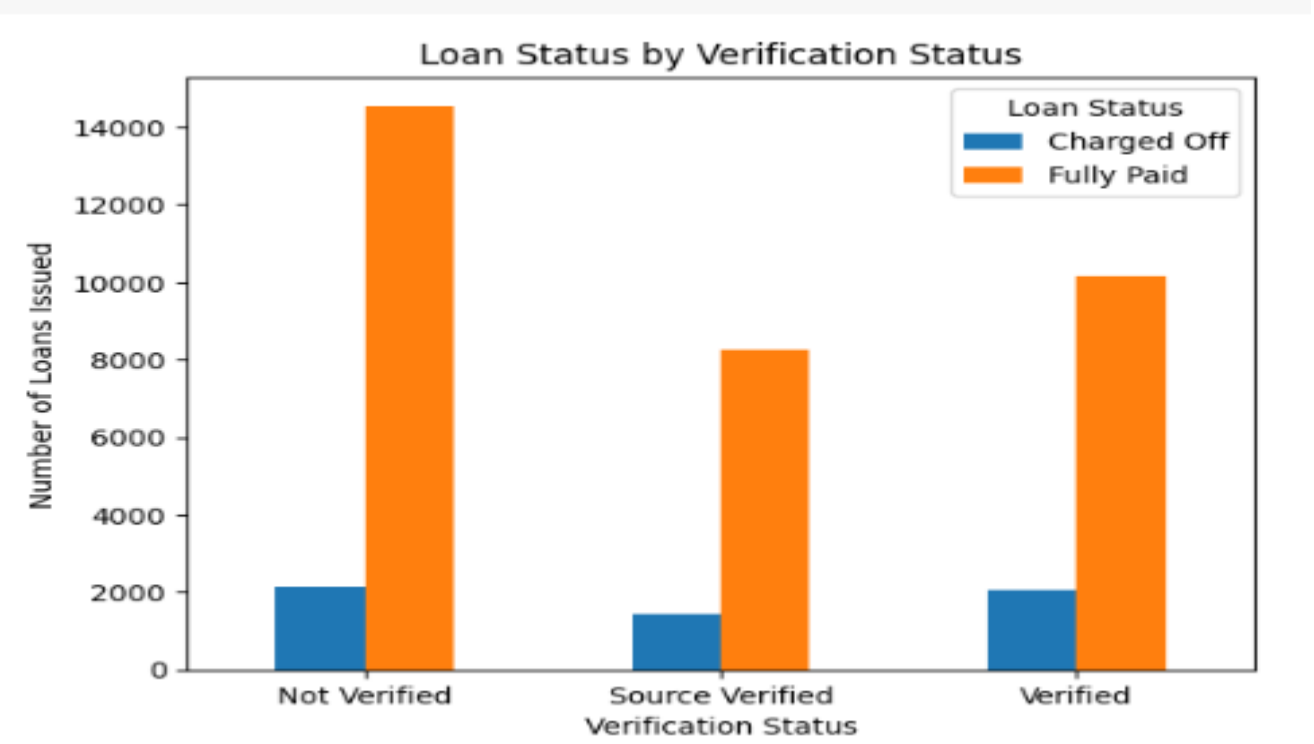
Observation: When we plotted the Grade attribute against the loan status, we see that people classified under “B” and “ C” grades have higher defaulters compared to others. The gauge used for classifying the grade could be reviewed by the loan approval team to reduce the risks.

Loans Status by Home Ownership



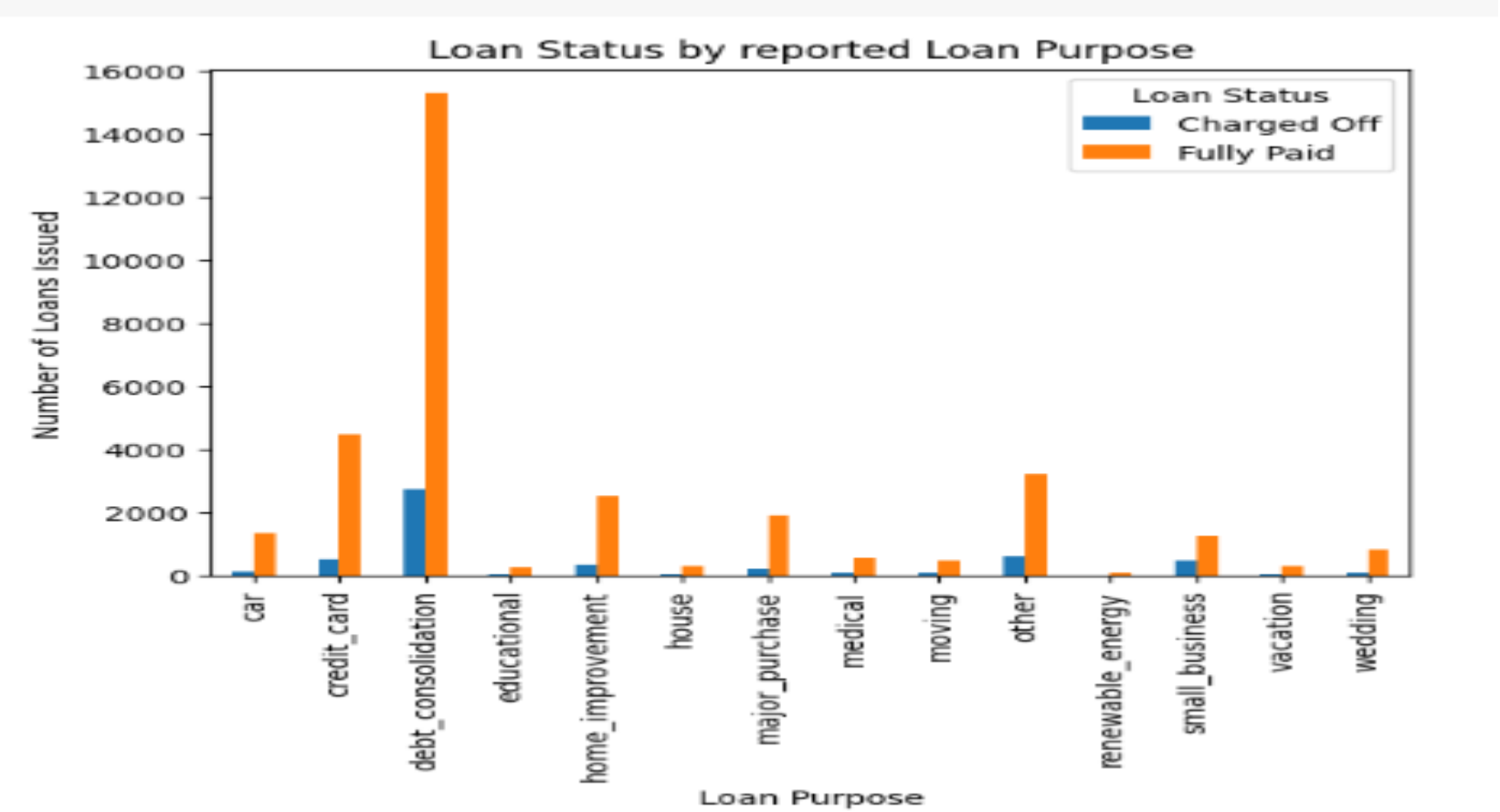
Observation: The graph above shows the relation between home ownership of borrowers against the loan status. People who are on Rent have defaulted more, closely followed by those who have their house mortgaged.

Loans Status by Verification Status



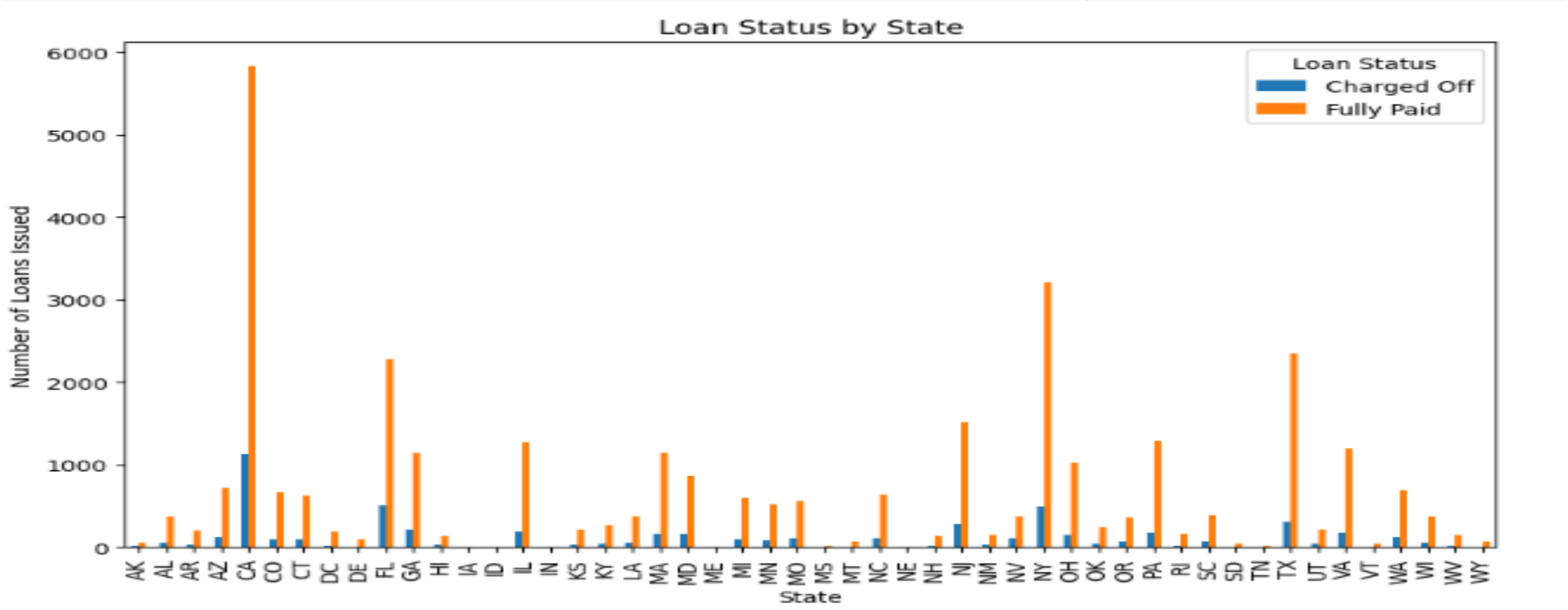
Observation: From the chart above, it is clear that the people who's verification status were Not Verified at the time of approving the loan have defaulted more than that of those that are Verified or Source Verified. The company can consider putting some stringent measures to probably approve loans only after Source Verified so as to reduce the amount of risks.

Loans Status by Purpose



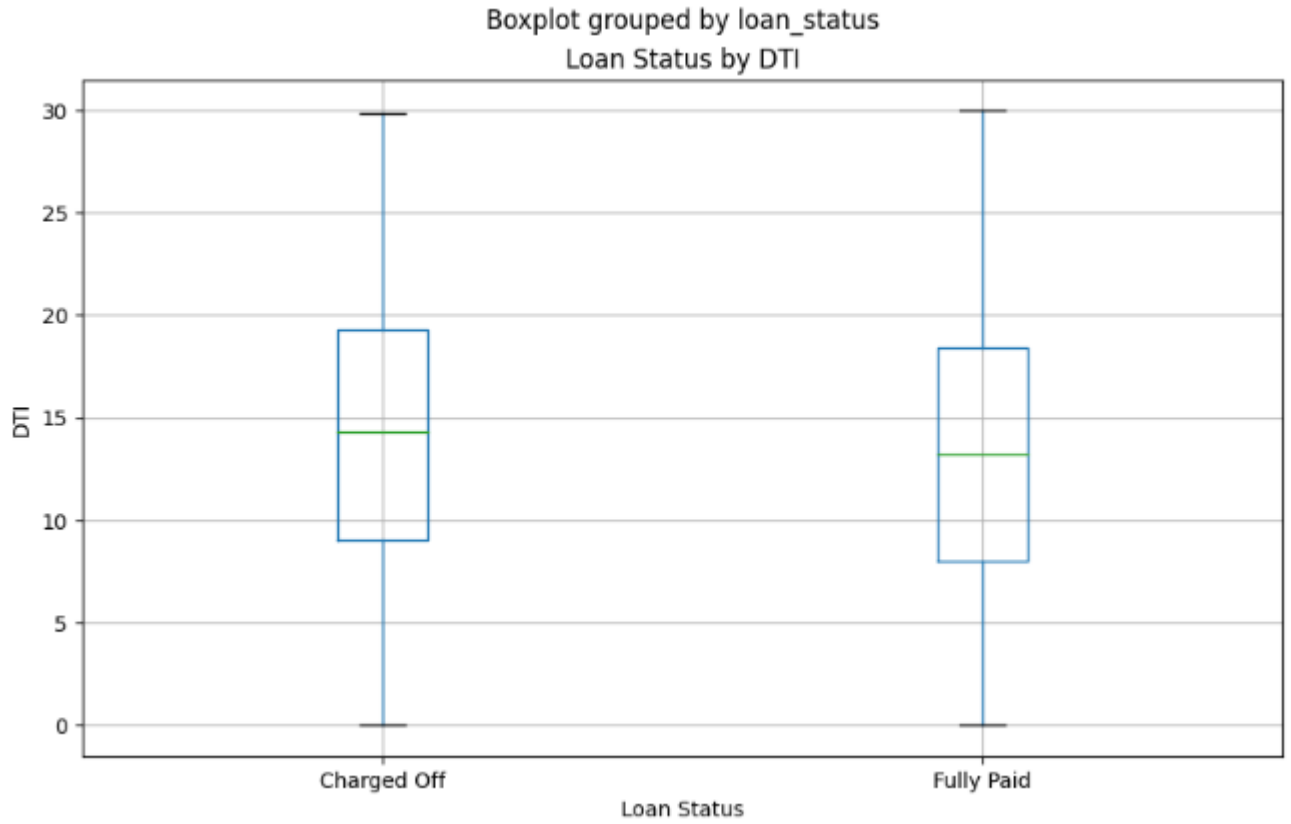
Observation: It is evident from the chart that people who indicated that they were availing loans due to debt consolidation have defaulted more. This is another parameter that the company should consider while approving loan applications.

Loans Status by Demographics



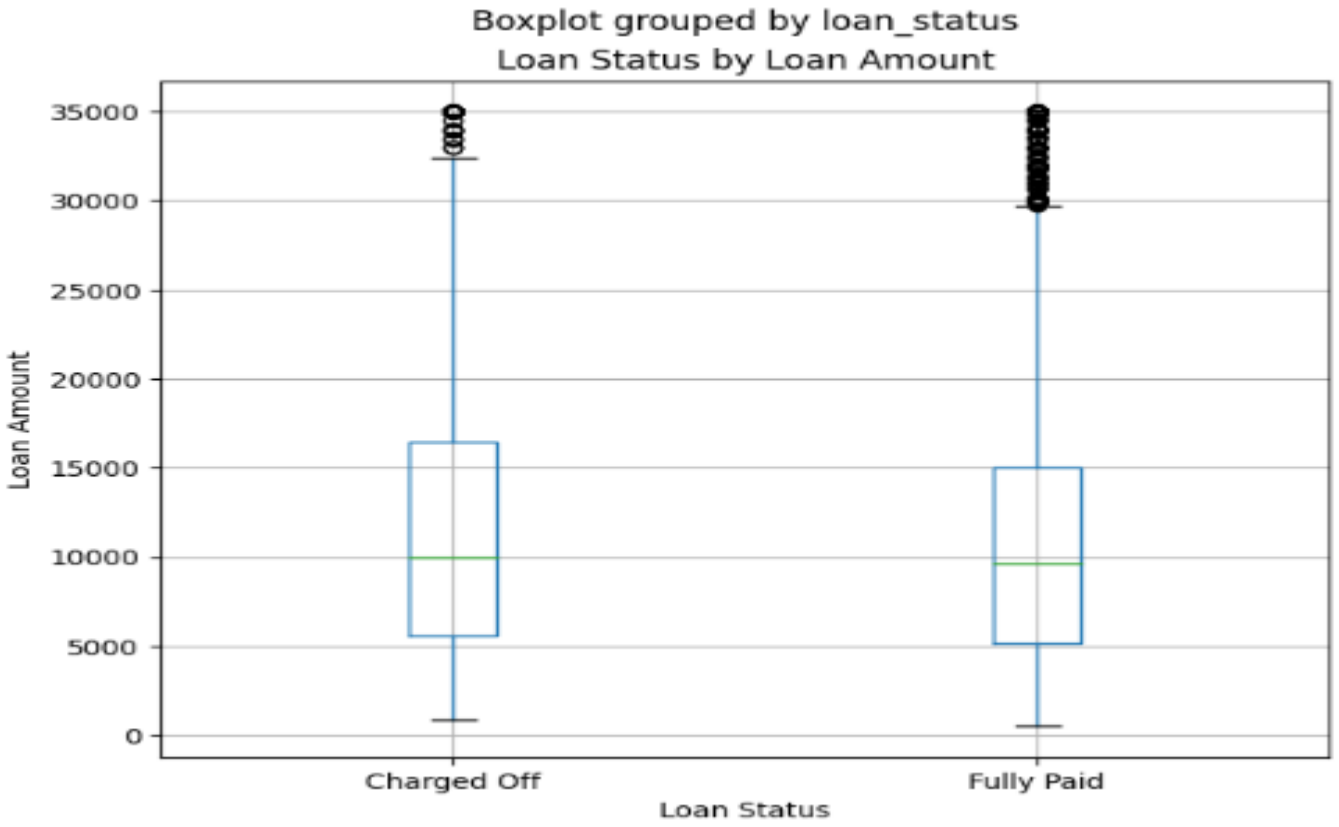
Observation: We inferred from the above chart that top 3 defaulters are residents of CA, NY and FL. Even though the Fully Paid ones follow the same pattern, Lending Club can consider applying more scrutiny for loan applications from these states.

Loans Status by Debt to Income (DTI) ratio



Observation: We did a box plot estimate of Debt to Income (DTI) ratio of the approved loans. This measure indicates how much debt the person owes against their income that they earn. It seems there is no significant difference between the defaulters and fully paid. We did not observe any outliers as well. One conclusion we could arrive at is that the people who defaulted may not have mentioned all the

Loans Status by Loan Amount



Observation: This chart tells us whether there is any relation between Loan Status and the Amount they have taken. Fully Paid applications have taken higher loan amount as there seems to be a lot of outliers beyond the IQR. Though there are some in this range for defaulters category as well, the median loan amount is almost the same.

Conclusion

- Based on the analysis we have done with the given data set, it seems there are patterns of defaults that the company can pay closer attention to while approving future loans.
- As the company grew over the years, the loan approvals have also seen the increase
- However, certain parameters where company could consider a more stringent approach is on: (a) while setting the interest rates, (b) when the applicant is applying for loan for their debt consolidation, (c) for the residents of CA, NY or FL and (d) how the Verification Status is arrived.
- These should help in reducing their loss and make more informed decisions while approving loans.