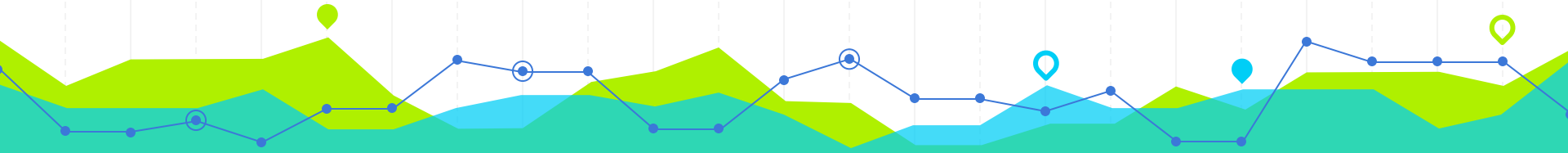
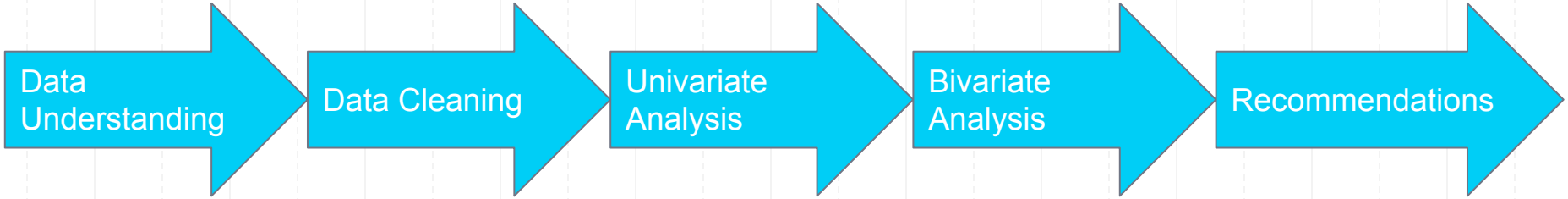


Lending Club Case Study

Group Members:
Gaurav Malhotra
Manik Garg

Approach


Data Understanding
Data Cleaning
Univariate Analysis
Bivariate Analysis
Recommendations

Data Understanding

About the Company

Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Backdrop

The company wants to understand the driving factors behind loan default, i.e. the driver variables which are strong indicators of default so that it can utilise this knowledge for its portfolio and risk assessment.

Statement

The group is expected to analyze the dataset containing data regarding loan applicants who have applied for loans in the past. With the help of EDA, the group will understand which variables affect the default tendency.

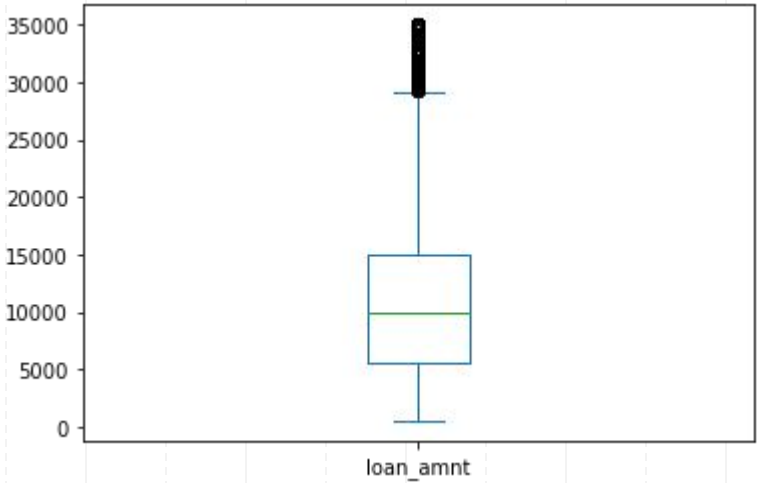
Data Cleaning

- ➔ Drop columns will null values.
- ➔ Handling missing values
- ➔ Convert values to correct data types like string to integers and float.

Univariate Analysis

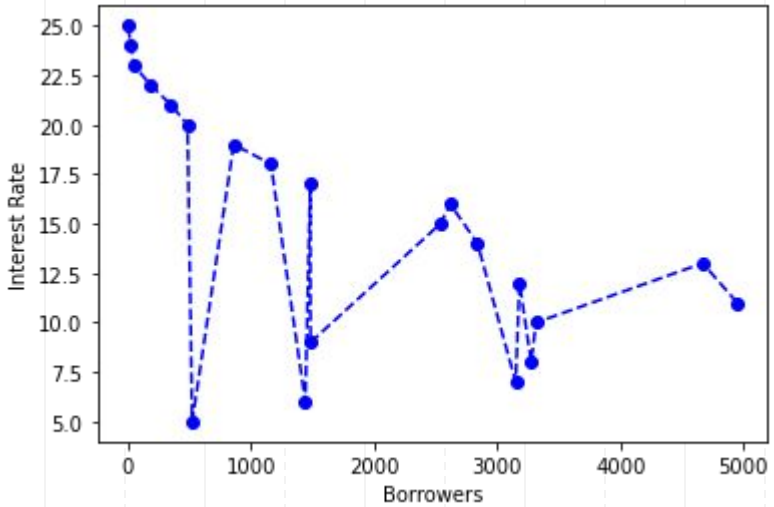
- ➔ Find the frequency of values occurring in the dataset for selected columns.
- ➔ Find the median(wherever possible) to get an idea of the data distribution for the variable.
- ➔ We have chosen median over average as it gives more optimal information about the data rather than average.
- ➔ Suggest variables for bivariate analysis.
- ➔ Plot the graphs of the selected variables for analysis.

Analysis of loan_amnt



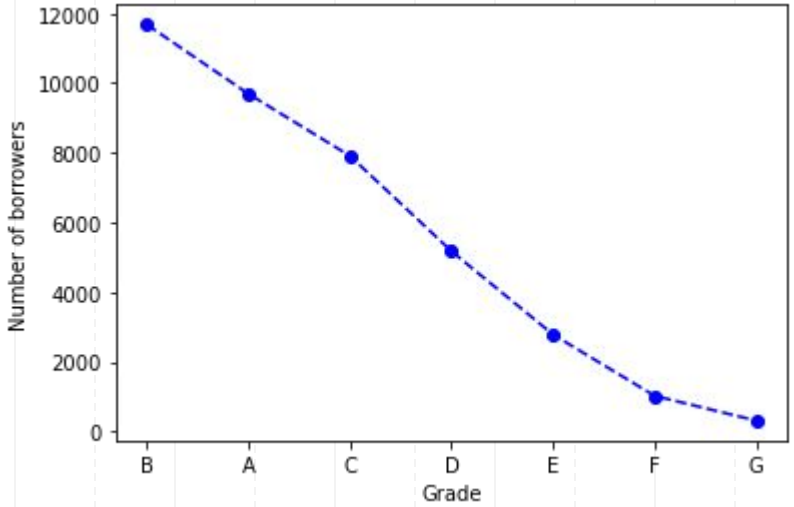
- ➔ Total unique Loan applied for by the borrower are 885.
- ➔ almost all our observation are in range of 30k to 35k

Analysis of int_rate



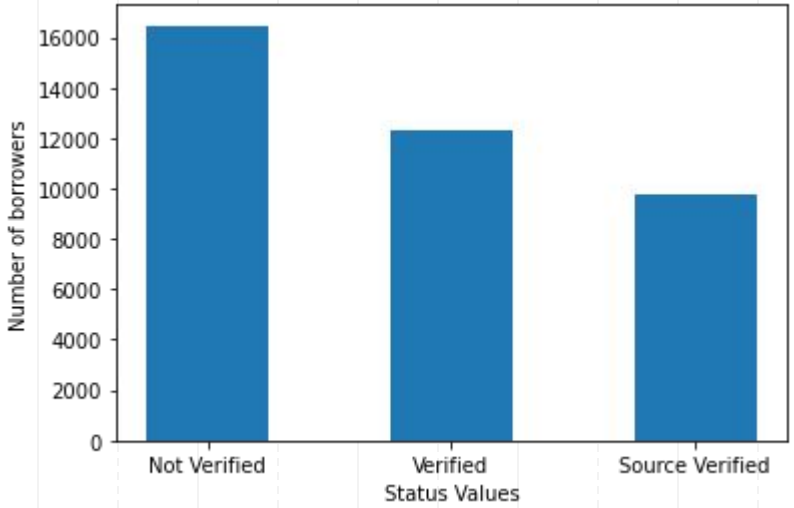
- ➔ There is a outlier with very high interest rate i.e. 25%
- ➔ Most of the interest rate lies 12% median
- ➔ This is a suitable candidate for bivariate analysis.

Analysis of grade



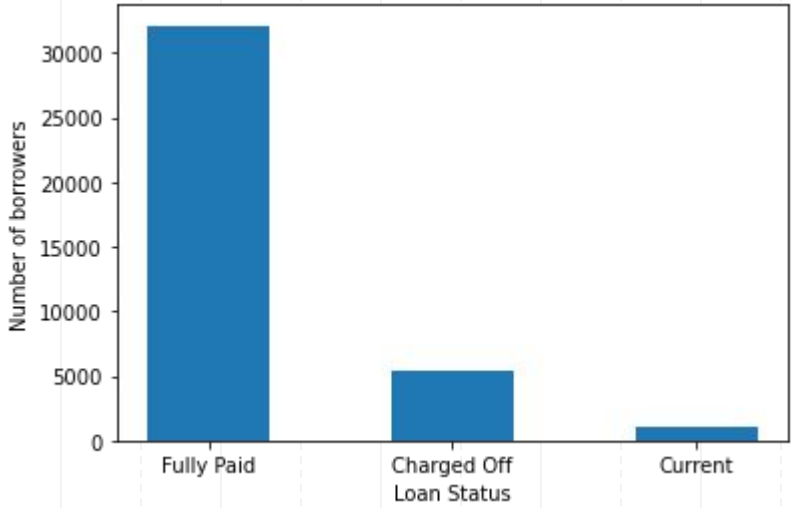
- ➔ The graphs seems very linear, so we can say that there are very less or no outliers.
- ➔ Based on this distribution we can say most of the borrowers fall into B grade

Analysis of verification_status



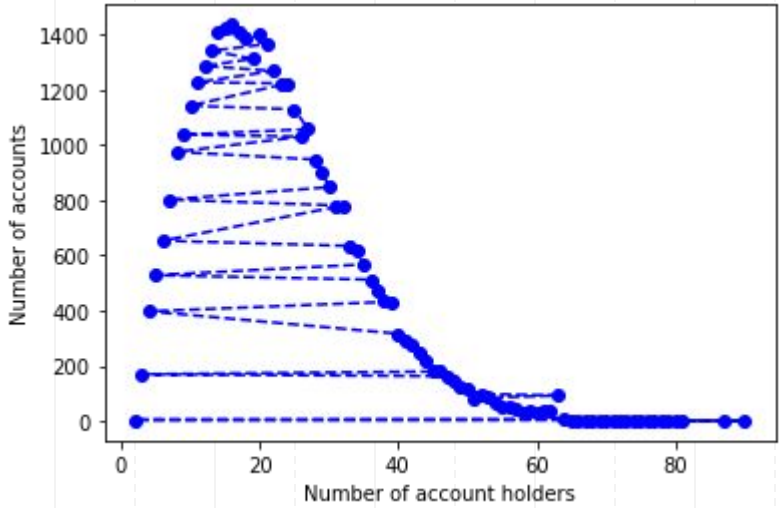
- Approximately 42% of the borrowers are not verified.
- We can infer that about 16000 who applied for loan are not verified and may be defaulters.
- This is a suitable candidate for bivariate analysis.

Analysis of loan_status



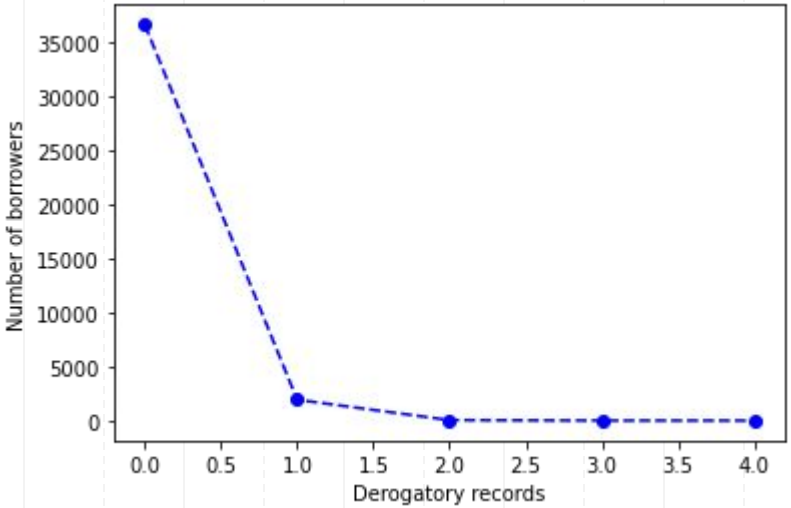
- ➔ % of Fully paid borrowers amounts to 83% approx.
- ➔ % of charged off borrowers is 14% approx.
- ➔ This is a suitable candidate for bivariate analysis.

Analysis of total_acc



- We can safely say that 60+ accounts can be treated a outliers.
- Also, from the graph we can see that median lies around 20.
- This is a suitable candidate for bivariate analysis.

Analysis of pub_rec

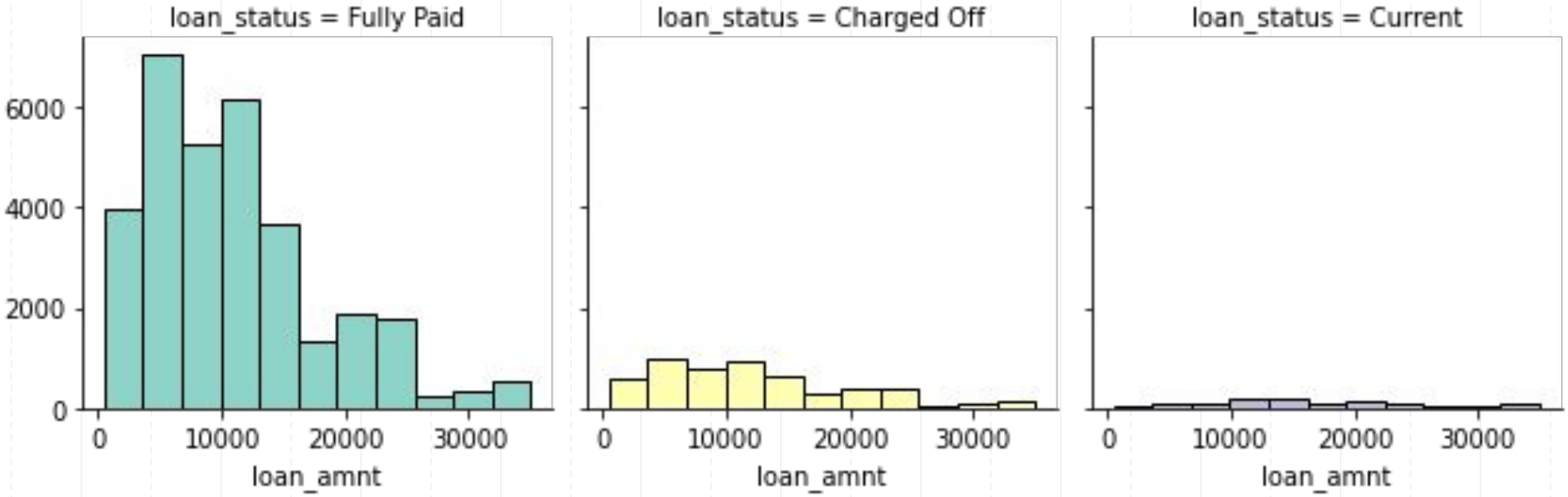


- ➔ Median of derogatory public records is 0.
- ➔ Based on this graph distribution, around 95% of the people have no derogatory public records, hence is not considerable.
- ➔ of use for bivariate analysis and we can safely discard this.

Bivariate analysis

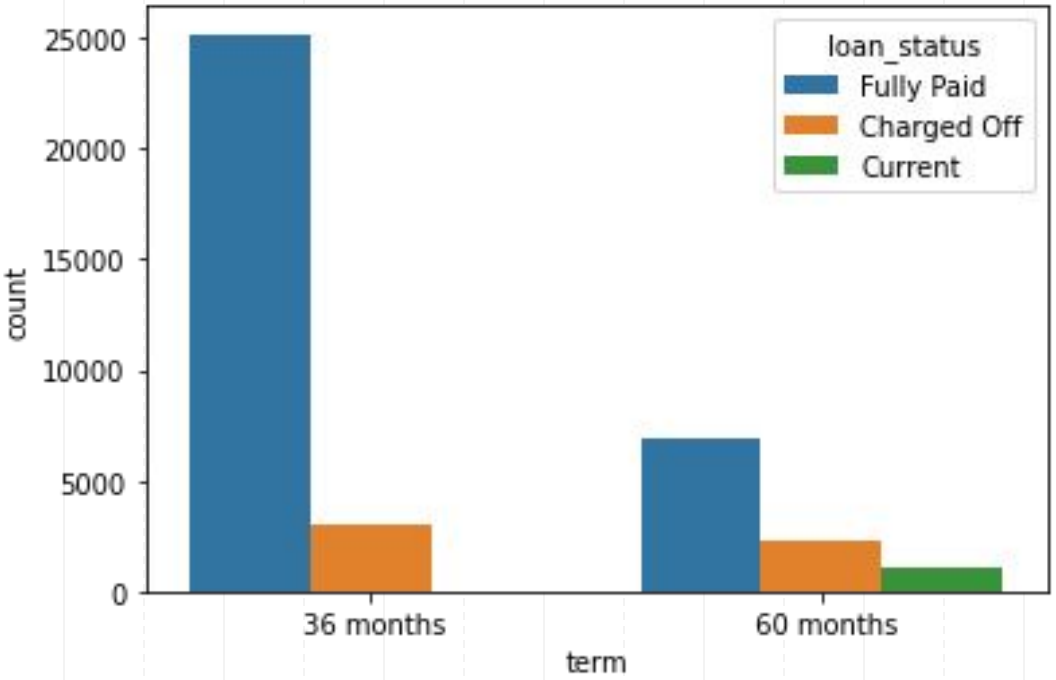
- ➔ Check how two variables affect each other
- ➔ Analysis their distribution
- ➔ Correlation Analysis

Bivariate analysis - Loan Amount vs Loan Status



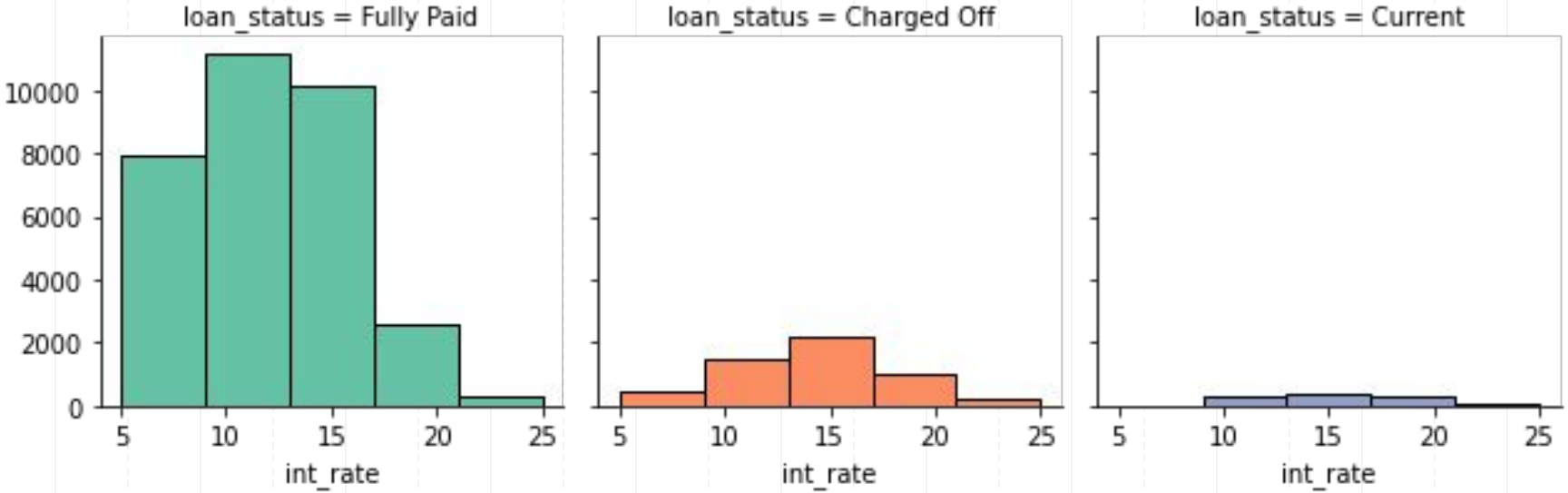
Most requested loan amount was in between 5k to 15k and has been paid fully but the charged off status is in the same range, so it becomes high risk and high reward.

Bivariate analysis - Term vs Loan Status



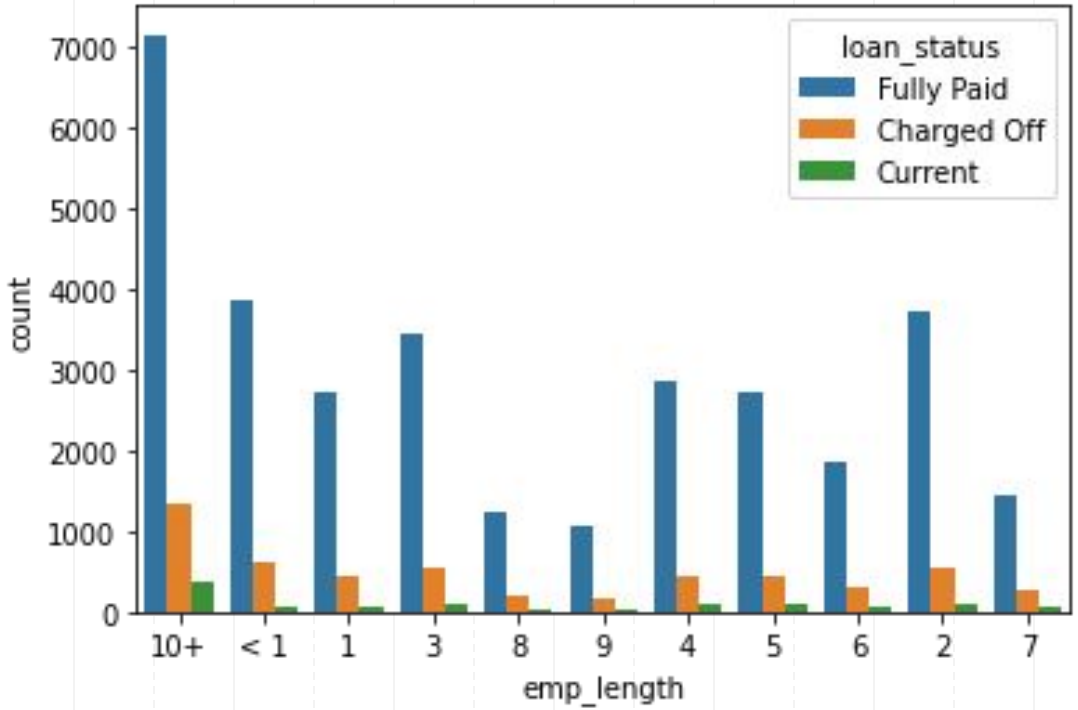
There are only two term rates present and clearly we can noticed highly requested term is 36 months and has been fully paid also

Bivariate analysis - int_rate vs Loan Status



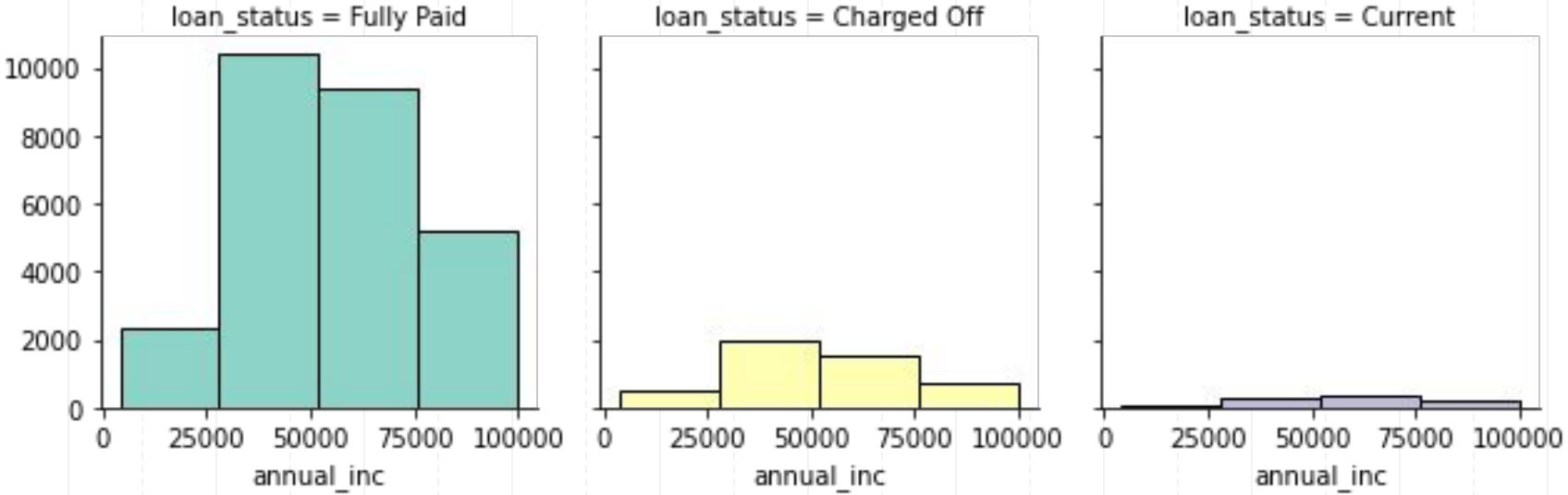
Most borrower who are unable to pay loan have interest rate of approx 15% and we noticed that lower interest rate allows borrower to fully pay there loan

Bivariate analysis - Emp_length vs Loan Status



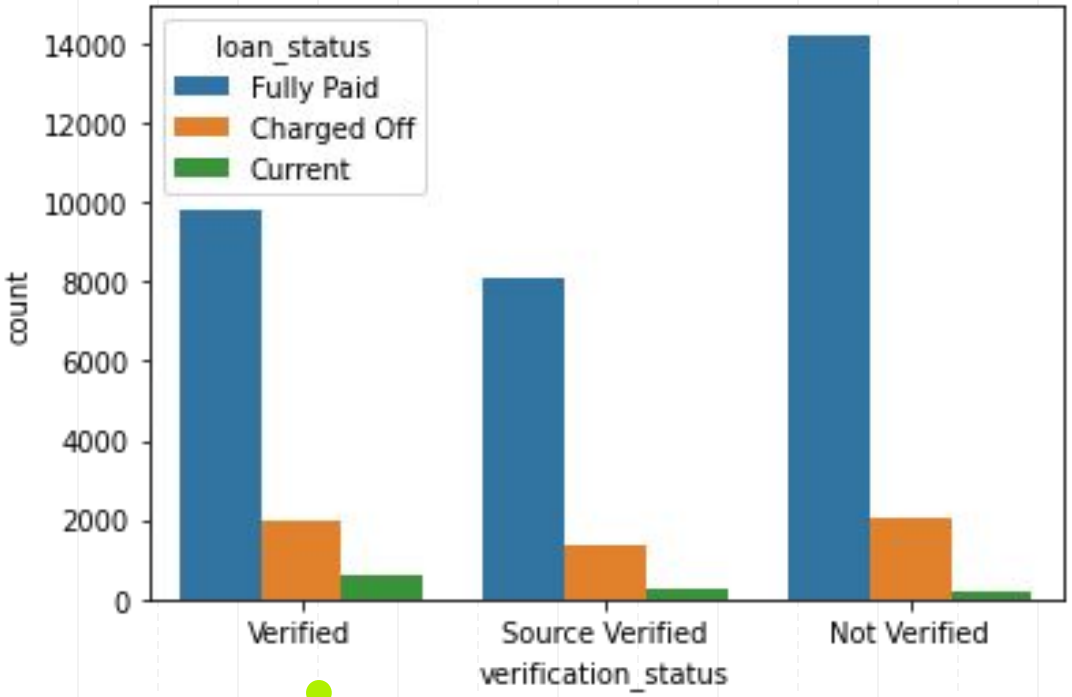
- Top insight is emp with 10+ exp are more reliable and pay there loan in full.
- Our hypothesis seems valid less emp length is more risky. While 10+ have same amount of charged off status - might indicate more variables are in play here to affect the thesis

Bivariate analysis - Annual_inc vs Loan Status



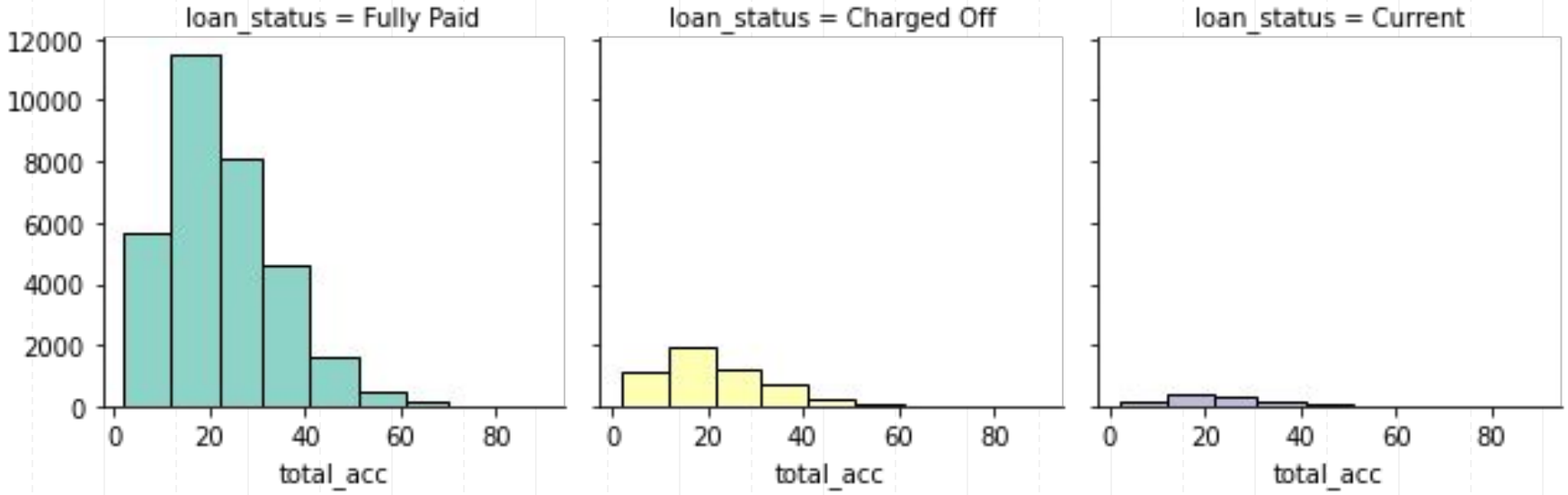
For this all three graphs seems in normal distribution hence does not contribute much to our hypothesis

Bivariate analysis - Verification_status vs Loan Status



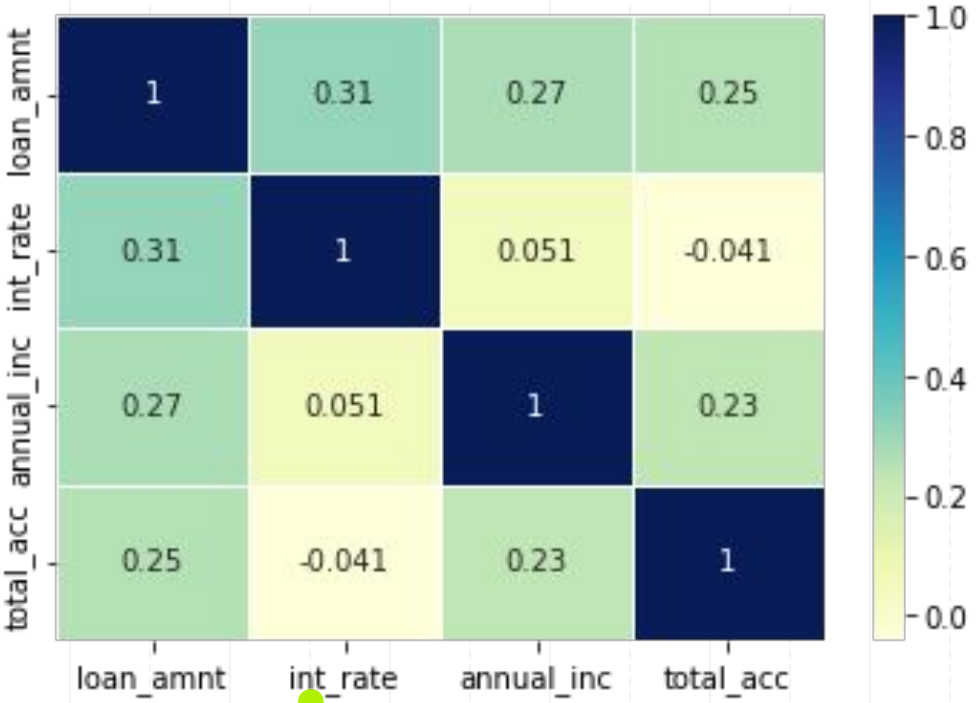
From the following visualization we can say emp verification status won't impact the load fully paid status, hypothesis is wrong here.

Bivariate analysis - Total_acc vs Loan Status



We may notice that emp with total_acc approx 10 to 20 have high repayment status.

Correlation



From this heat map we can say int_rate is correlated to loan_amnt

Recommendations

- ➔ Stop approving loans where amount/income is higher than 30%
- ➔ If amount asked is between 5k to 15k then its high risk and high reward for the LC.
- ➔ Term loan with 36 month are more reliable than with the other.
- ➔ We noticed that lower interest rate allows borrower to fully pay there loan. So if loan is given to risky customers we can give them with low interest. Again it's a risky move.
- ➔ Stop approving loans to people with prior bad record. Or at least stop approving high-value loans
- ➔ Employee with long exp are more reliable.
- ➔ We can do some cost cutting by not doing employee verification status because it doesn't impact the loan fully paid status.

Libraries Used

- ➔ Numpy
- ➔ Pandas
- ➔ MatplotLib
- ➔ Seaborn
- ➔ Pandas Profiling

THANK YOU