Lending Club Case Study

SUBMISSION

Name: Siddhant Sambit, Tina Ghosh

Msc in ML & AI batch

Batch Start Date – 30 December 2019

Email – siddhantsambit@gmail.com, ghosh.tina90@gmail.com

Problem solving methodology

- Data Loading
- Data Understanding
- Data Cleaning
- Univariate Analysis and Summary
- Bivariate Analysis of Categorical Data
- Bivariate Analysis of Continuous Data
- Bivariate Analysis Summary
- Additional Analysis(Derived Metrics)
- Conclusion

Project Brief :
For Lending Club Develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

- **Business and Data Understanding :**

  Lending club has to make a decision for loan approval based on the applicant's profile and information about past loan applicants and whether they 'defaulted' or not.

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 results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

In the Problem Statement we are only considering the risk of after loan approved . There can be three scenario :

- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

- Based on the above loan status need to decide whether a person 'defaulted' or not.**• Business objectives :**
  Need to understand the **driving factors / driver variables)** which are strong indicators of default So that Lending club can utilise this analysis for risk assessment.`

Data Understanding and Cleaning

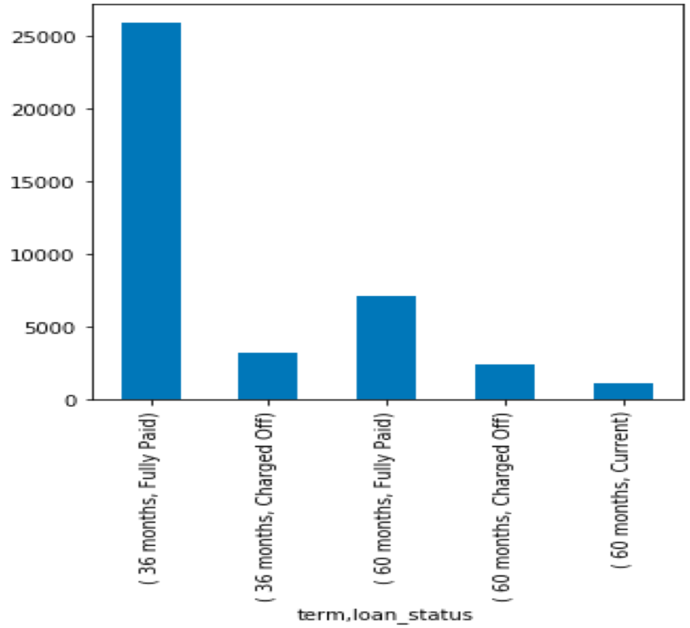
The input loan data provided contains 39717 records with 111 metrics associated.

Cleaning –

- Dropping records which are having loan amount equals to zero
- Dropping all columns which contains all null / missing values
- Dropping columns with more than 30% and not relevant for our analysis
- Removing % symbol and replacing null value with median for ‘revol_util’ column
- Converting all date field from Object to dateTime
- Replacing missing values with ' Job Title Not Provided ' where ‘emp_title’ is not provided
- Replacing null values with zero for column ‘pub_rec_bankruptcies’
- Converting loan_amnt from object to integer type.

Univariate Analysis

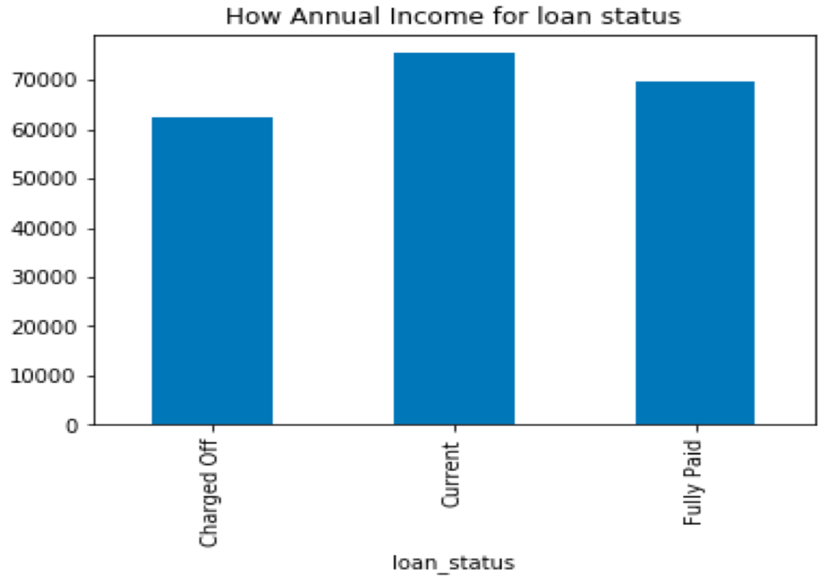
Analysis on loan term of consumers for different type of loan status



Conclusion: Consumers taking loan term 36 months are most defaulted one

Univariate Analysis

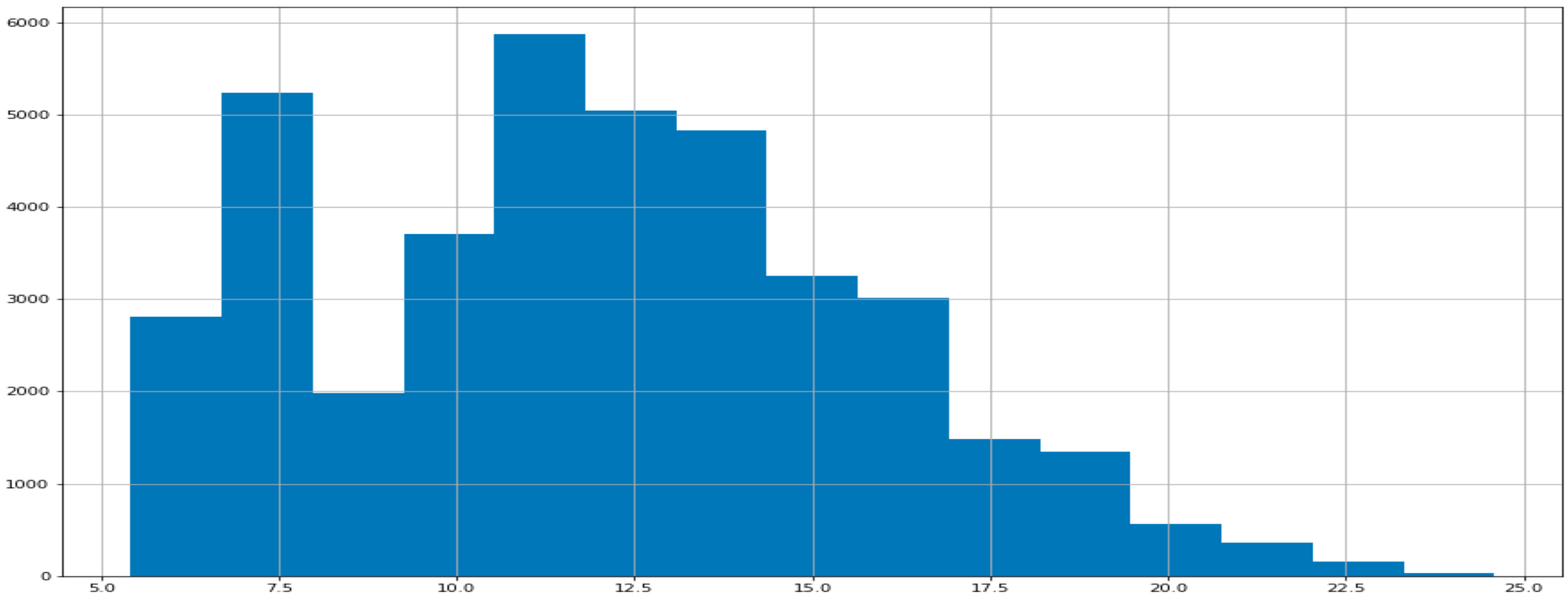
Analysis on annual income of consumers for different type of loan status



Conclusion: Consumers having less 60000 as annual income are most defaulted one

Univariate Analysis

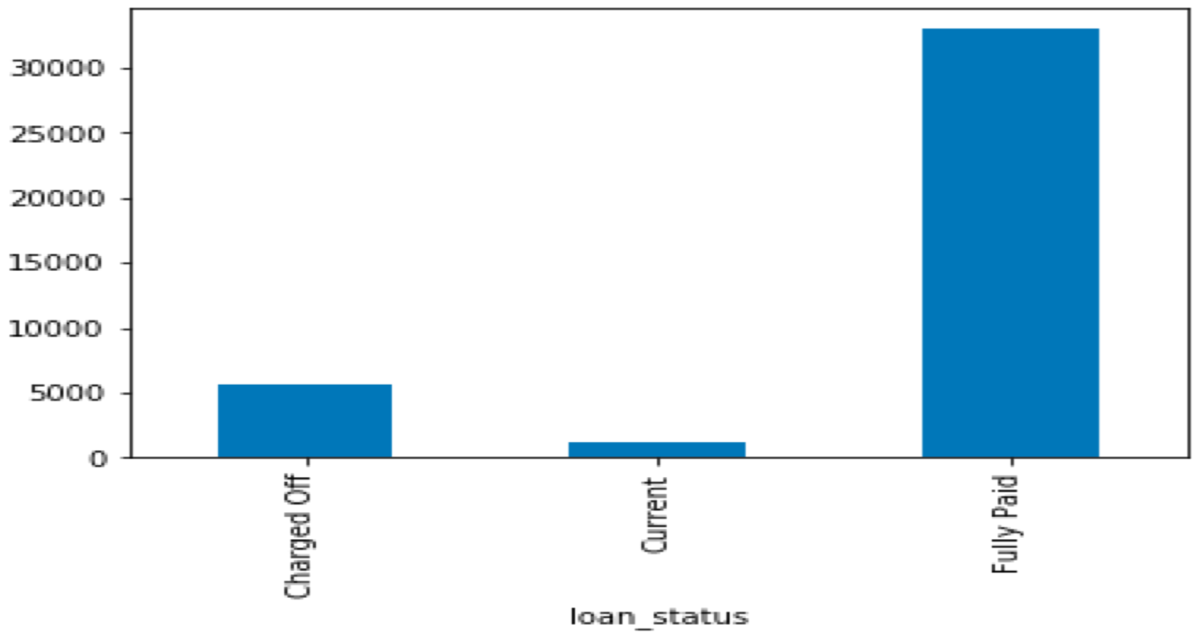
Analysis on interest rate variation



Conclusion: People are getting loan approved with interest rates varying within 10.1 % to 14.5%

Univariate Analysis

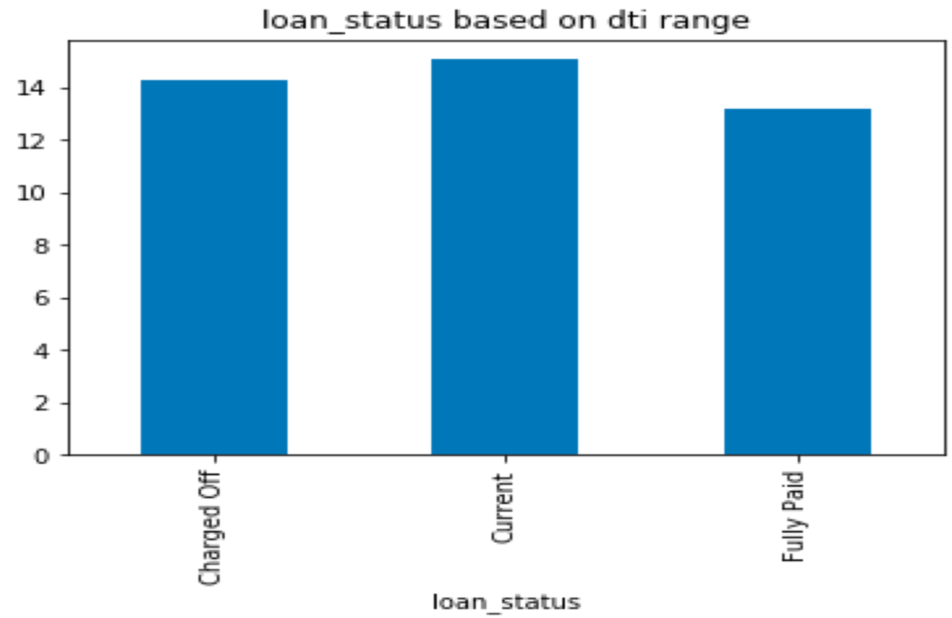
Analysis on enquiries for last 6 months



Conclusion: People who have fully paid off loans did maximum enquiry in last 6 months

Univariate Analysis

Analysis on dti ratio for different type of loan status



Conclusion: Using dti ratio we cannot conclude anything evident as all are near to same range

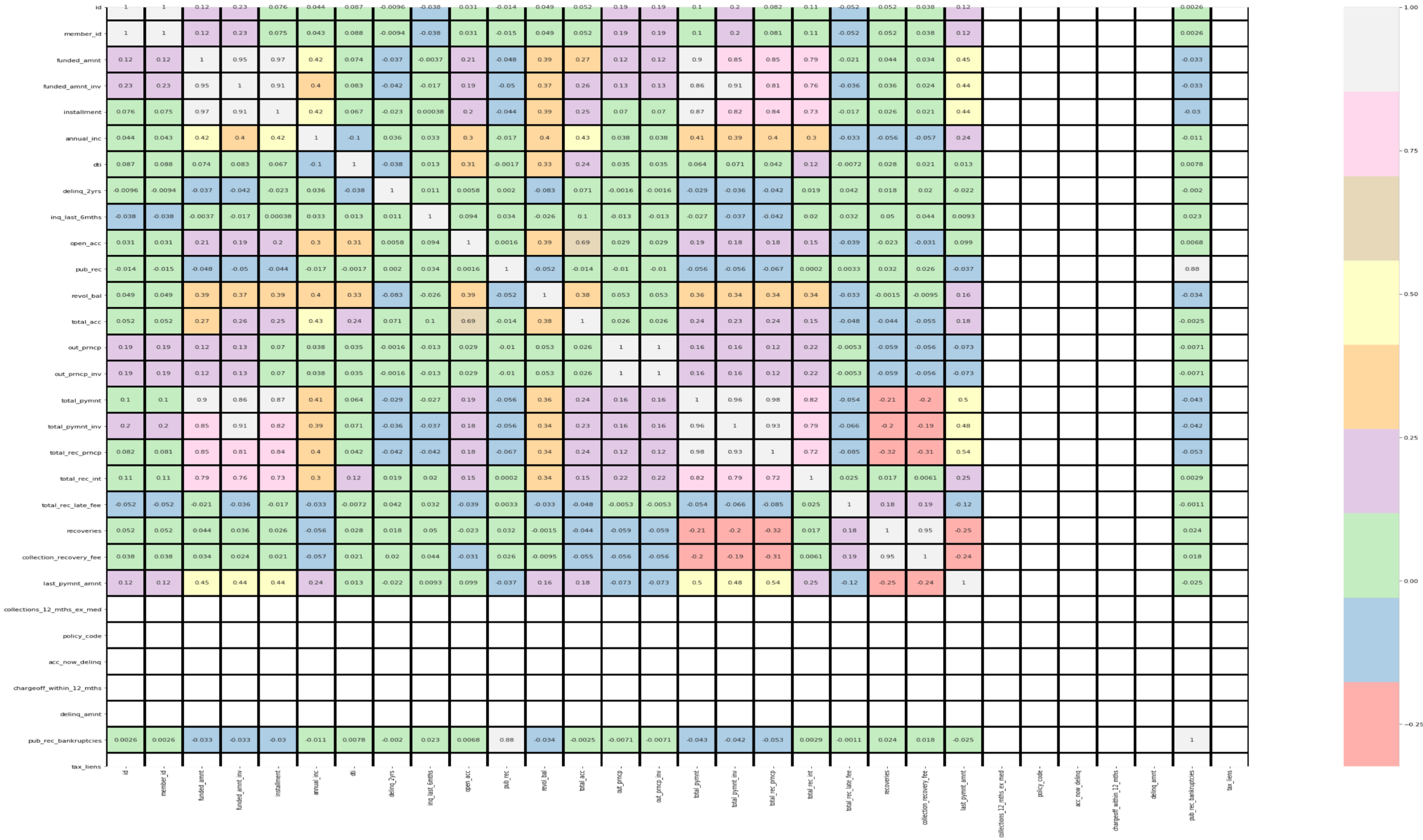
Univariate Analysis - Summary

Variable Name	Type Of Variable	Conclusion	
Loan_amnt	Quantative Variable	mean	683131
		std	210694
		min	54734
		75%	837755
		max	1077501
Funded_amnt_inv	Quantative Variable	mean	10397
		std	7128
		min	0
		50%	8975
		max	35000
Int_rate	Quantative Variable	mean	12.02
		std	3.72
		min	5.42
		25%	9.2
		max	24.5

Univariate Analysis - Summary

Variable Name	Type Of Variable	Conclusion	
home_ownership	Categorical Variable	Funded_amnt	If home_ownership is RENT they get more funded_amount
verification_status	Categorical Variable	Funded_amnt	If verification_status Not Verified then they gets more funded_amount which makes more changed-off
purpose	Categorical Variable	Funded_amnt	If purpose is debt consolidation(1) and credit card (2) people are getting more funding amount.
emp_length	Categorical Variable	Funded_amnt	If emp_length is more than 10 years people are getting more Funded_amount.

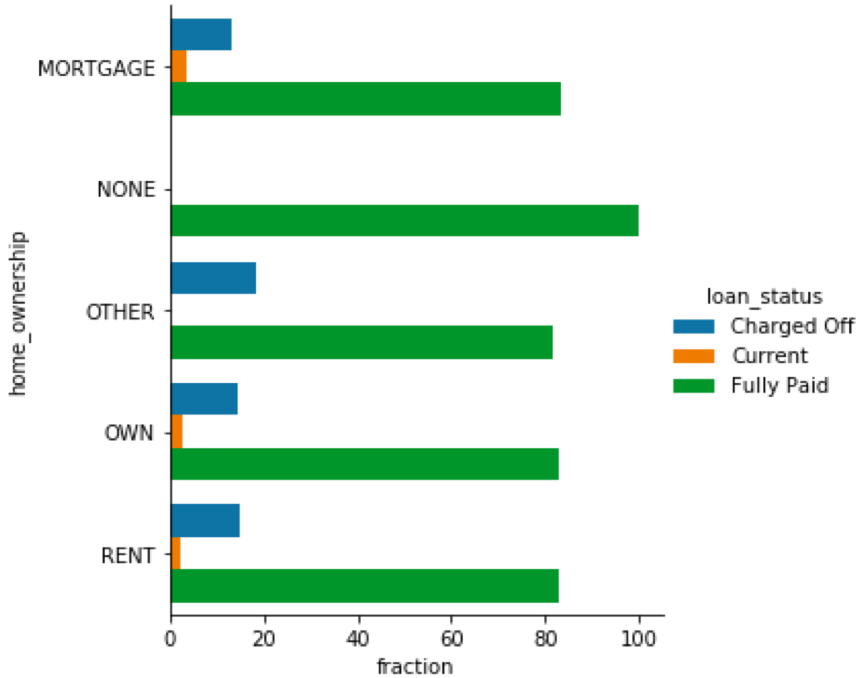
Bivariate – Correlation Matrix (Continuous Data)



There is no significant positive or negative correlation between continuous variables but funded amount and loan amount can be found strongly correlated

Bivariate Analysis Of Categorical Data

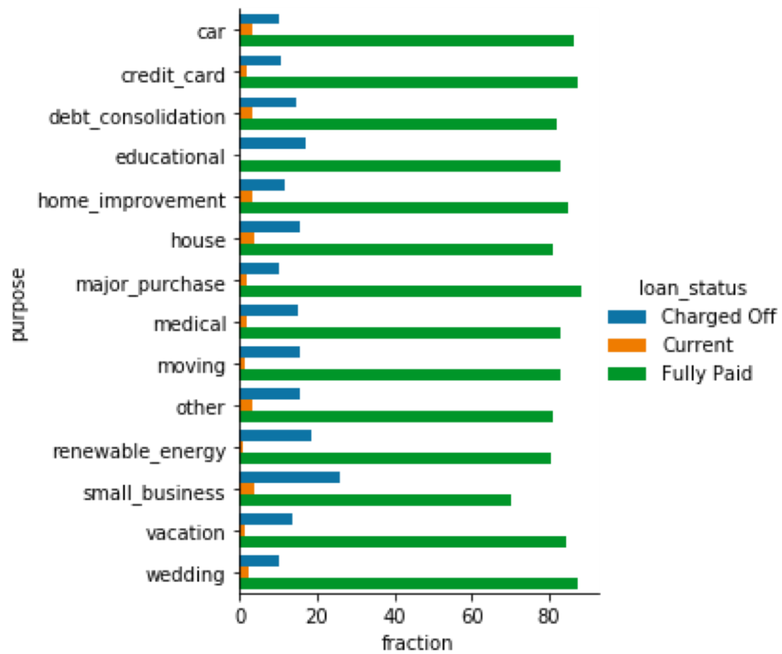
Analysis on home ownership type of consumers for different type of loan status



Conclusion: Consumers with home ownership as others are most defaulted one

Bivariate Analysis Of Categorical Data

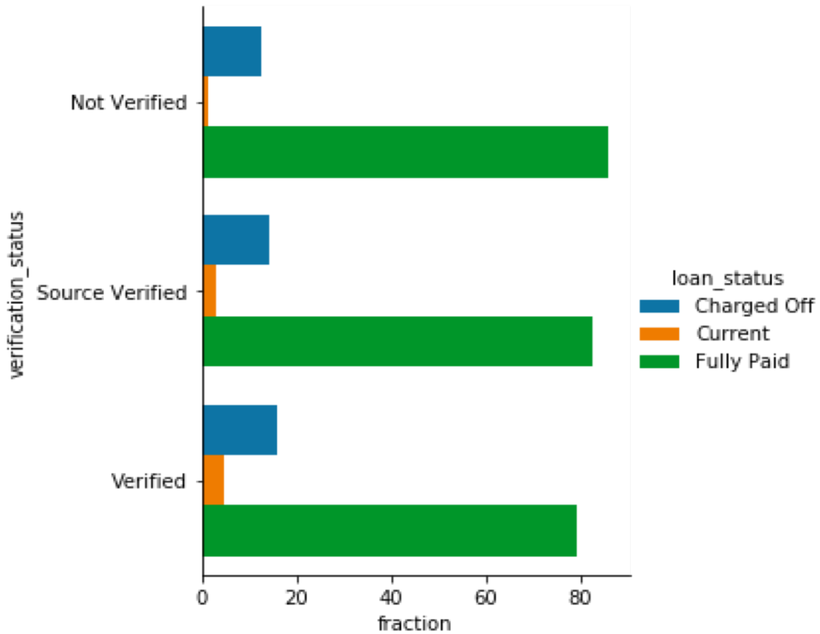
Analysis on purpose type of consumers for different type of loan status



Conclusion: Consumers with purpose as small business are most defaulted one

Bivariate Analysis Of Categorical Data

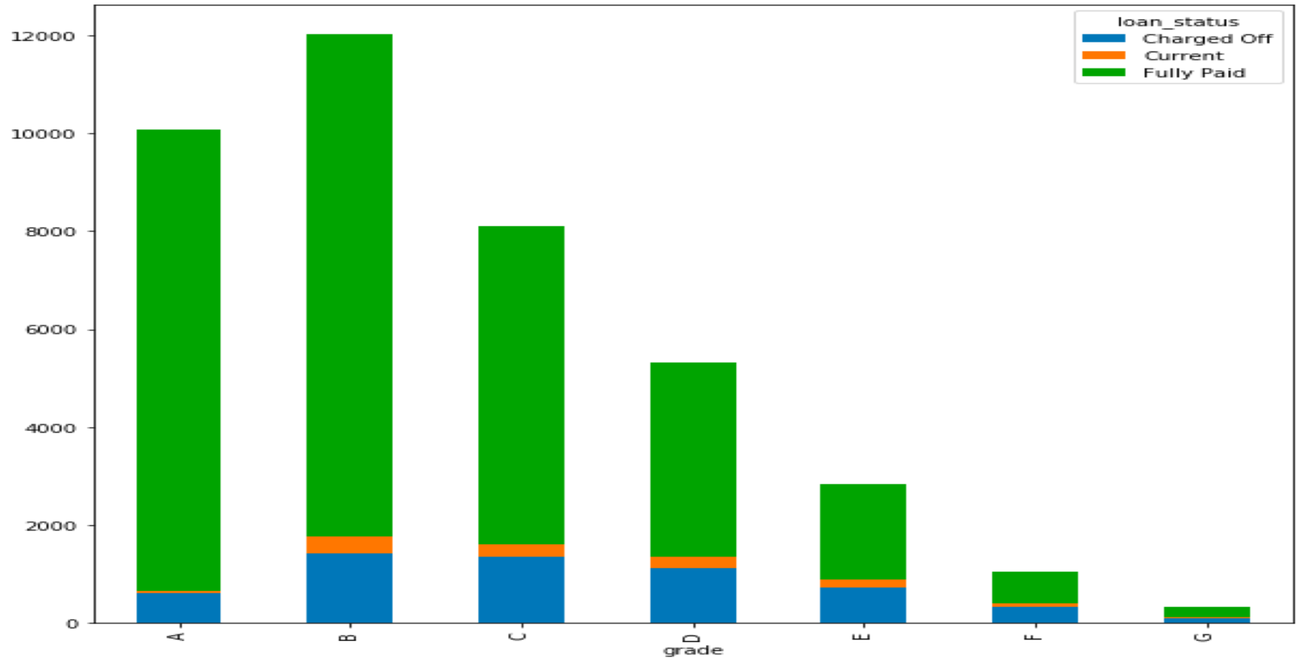
Analysis on verification status of consumers for different type of loan status



Conclusion : Consumers with verification status as verified are most defaulted one

Bivariate Analysis Of Categorical Data

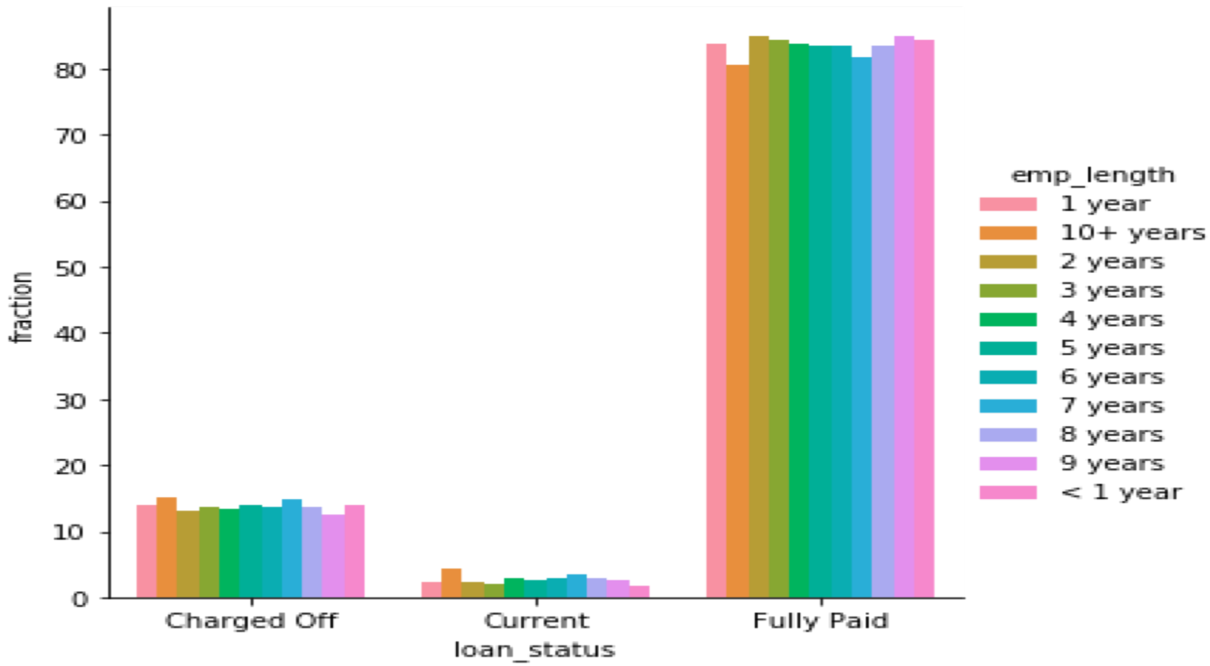
Analysis on grade for different type of loan status



Conclusion : Loans grade as B & C are most defaulted one

Bivariate Analysis Of Categorical Data

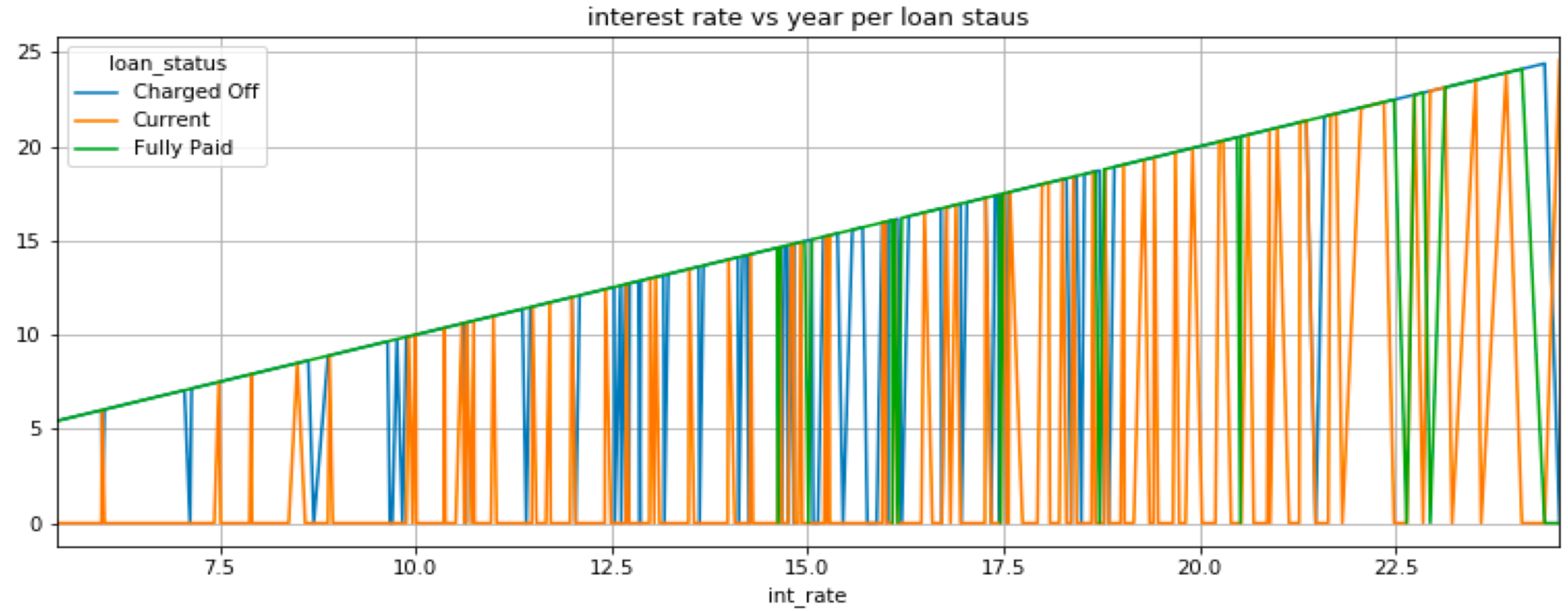
Analysis on employment length for different type of loan status



Conclusion : Employment length with less than 1 year, 1 year and 10+ years are most defaulted one

Bivariate Analysis Of Categorical Data

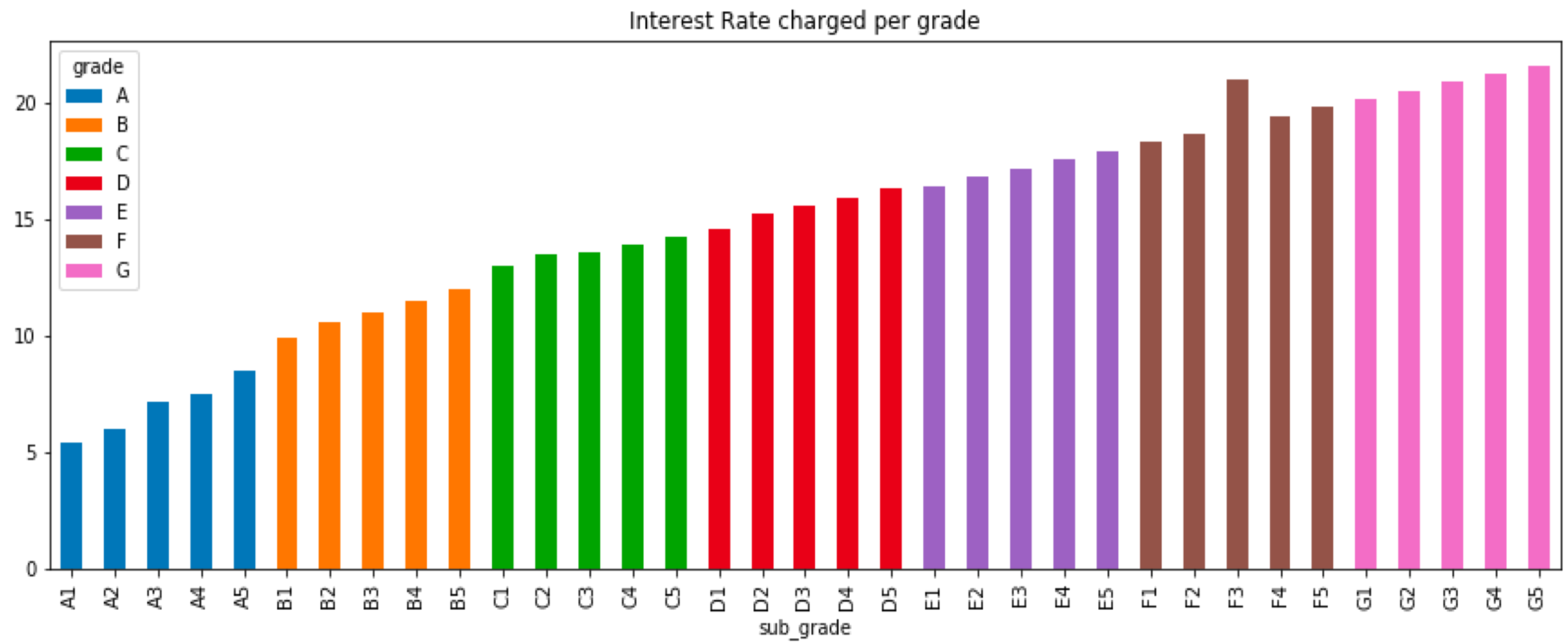
Analysis on interest rate for different type of loan status



Conclusion : Higher interest rates tend to have most defaulted one

Additional Analysis Based on Derived Metrics

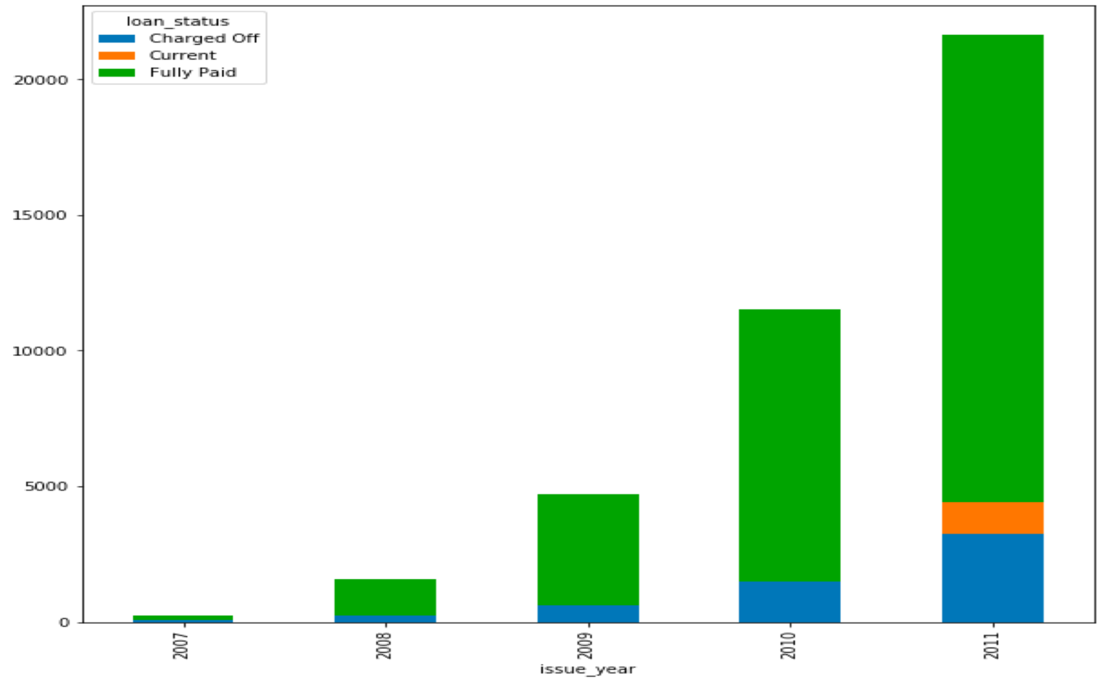
Analysis on interest rate per grade



Conclusion : Grades with higher interest may lead to charged off

Additional Analysis Based on Derived Metrics

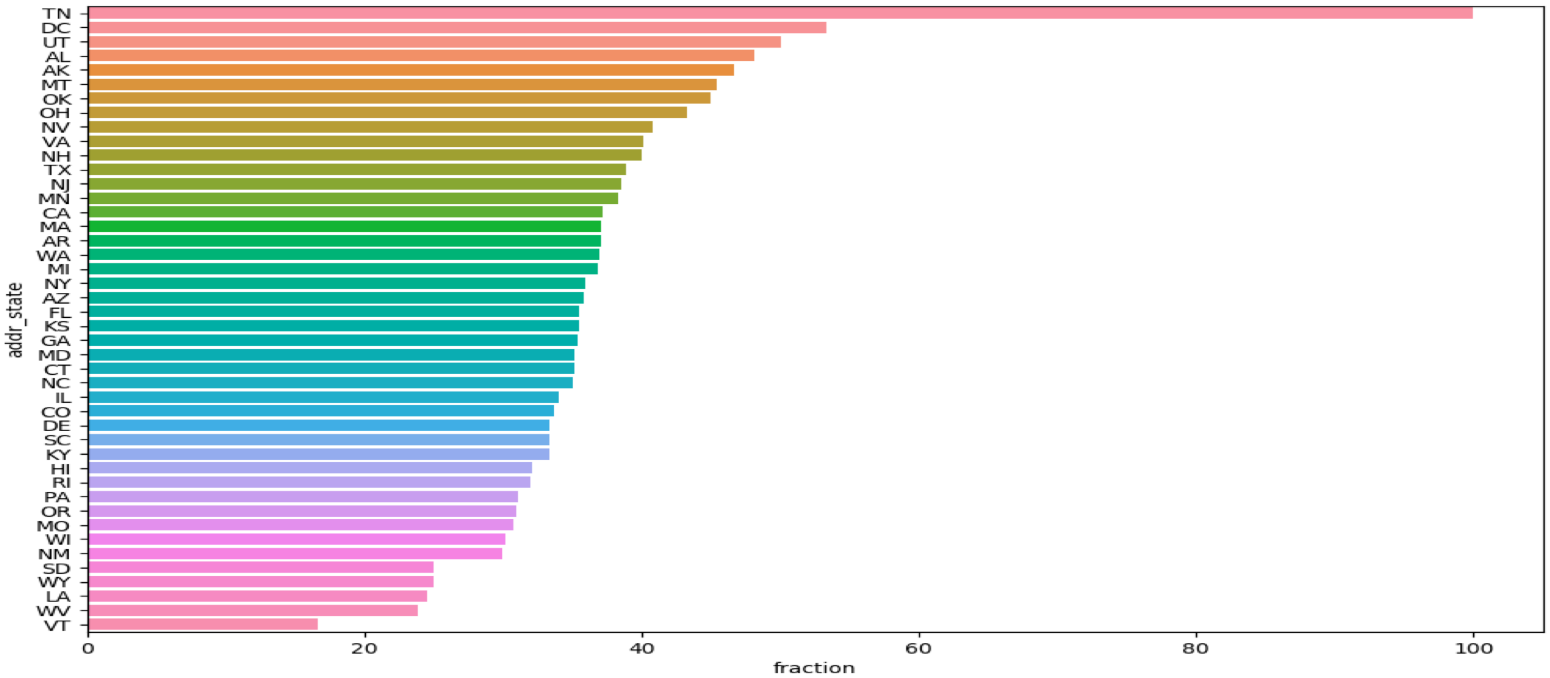
Analysis on issue year for different type of loan status



Conclusion : Loans issued in the year 2011 are most defaulted one

Additional Analysis Based on Derived Metrics

Analysis on address state for loan status verified as there were maximum defaulters



Conclusion : State list in decreasing order for maximum defaulter with verified address

Conclusion

Incase consumer applies for a loan detail analysis have been provided in each slide for driving a decision to avoid loss of business and finance to company.