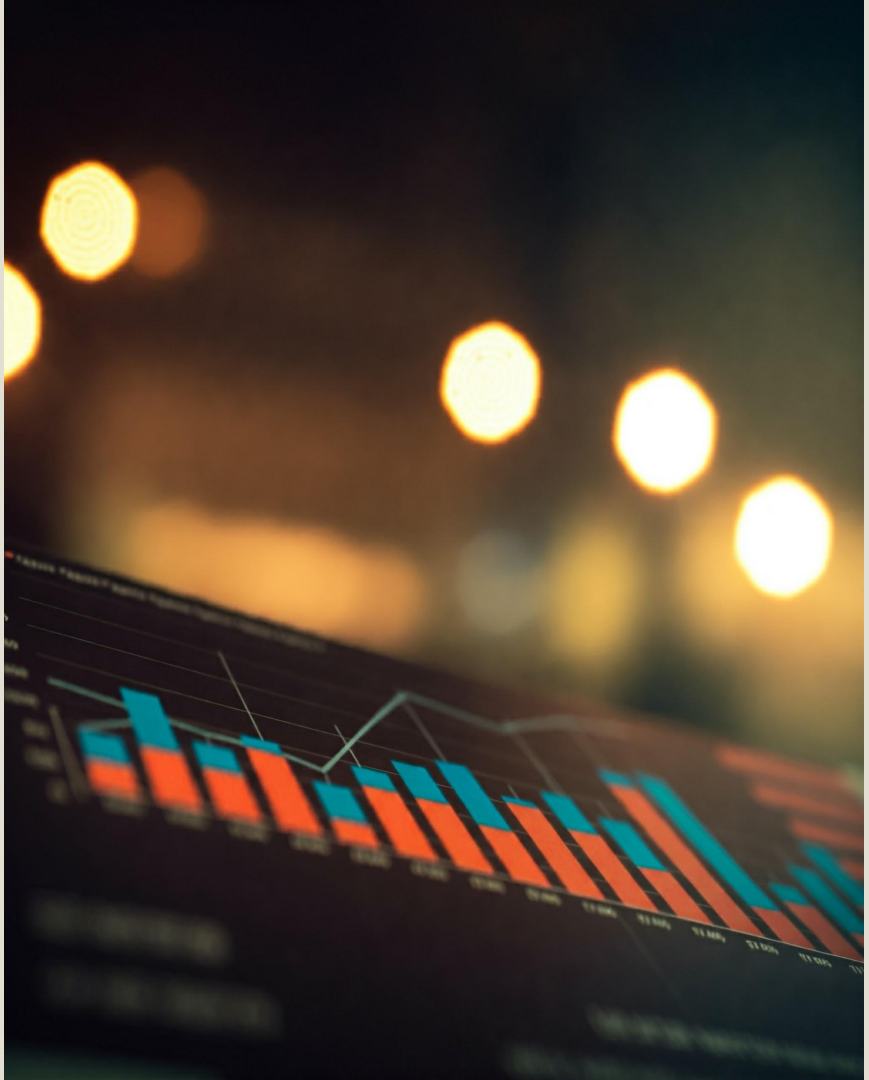
16 Nov 2024

Lending Club Case Study

Built as part of Upgrad EDA Activity



Agenda

- ➔ Problem Statement
- ➔ EDA
  - ◆ Data Cleaning
  - ◆ UniVariate & Segmented Univariate Analysis
  - ◆ BiVariate Analysis
  - ◆ Derived Metrics

Problem Statement

For Financial Institutions in lending business it very critical to identify right customer to lend, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss).

The Idea is to use EDA to identify the right attributes which can help predict ‘risky’ applicants from the Loan Data Set Shared from largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures between 2007 and 2011

Data Cleaning

Steps:

1. Fix rows and columns
   a. Identify columns with lot of Null
      i. Removed Columns with Zero Values Total Column With 0 values: 54
2. Some of below columns don't add any value to predict loan default:
   a. desc - this description of the loan doesnt help
   b. next_pymnt_d','mths_since_last_delinq','mths_since_last_record' - These columns don't add any value
   c. Pymnt_plan, initial_list_status, collections_12_mths_ex_med, policy_code,acc_now_delinq, application_type, pub_rec_bankruptcies, tax_liens,delinq_amnt also have been removed
3. Some columns like "id", "member_id", "url", "title", "emp_title" are redundant
4. Other columns deleted last_credit_pull_d,out_prncp_inv,total_pymnt_inv, funded_amnt, delinq_2yrs, revol_bal, out_prncp, total_pymnt, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, chargeoff_within_12_mths – These columns are not related to loan default
5. Refer to attached data dictionary for the final list of columns 21 columns considered
6. Drop rows with more than 5 null values
7.

Data Cleaning

Final Shape and Columns:

Shape:(39717, 21)

Columns:

'funded_amnt_inv', 'Term','int_rate', 'installment',
'Grade','sub_grade', 'emp_length',
'home_ownership', 'annual_inc','verification_status',
'issue_d', 'loan_status', 'purpose',
'zip_code','addr_state', 'dti',
'earliest_cr_line', 'inq_last_6mths', 'open_acc',
'revol_util', 'total_acc'

Data Cleaning

Treating Missing Values:

1. **Replace Employment Length & Revol_Util with Mode**

Standardizing Values:

1. **Round funded_amnt_inv**
2. **Remove decimal from installment**
3. **Remove months from Term column**
4. **Standardize Emp Length to have numbers from - 0 to 10**
5. **Convert interest rate from Text to float and remove ‘%’**

UniVariate, Segmented Univariate & Bi-Variate Analysis

Analysis of Numerical Values:

1. Treat inq_last_6months with IQR for lower and Upper bound
2. Treat Annual Inc for outliers with log function for analysis

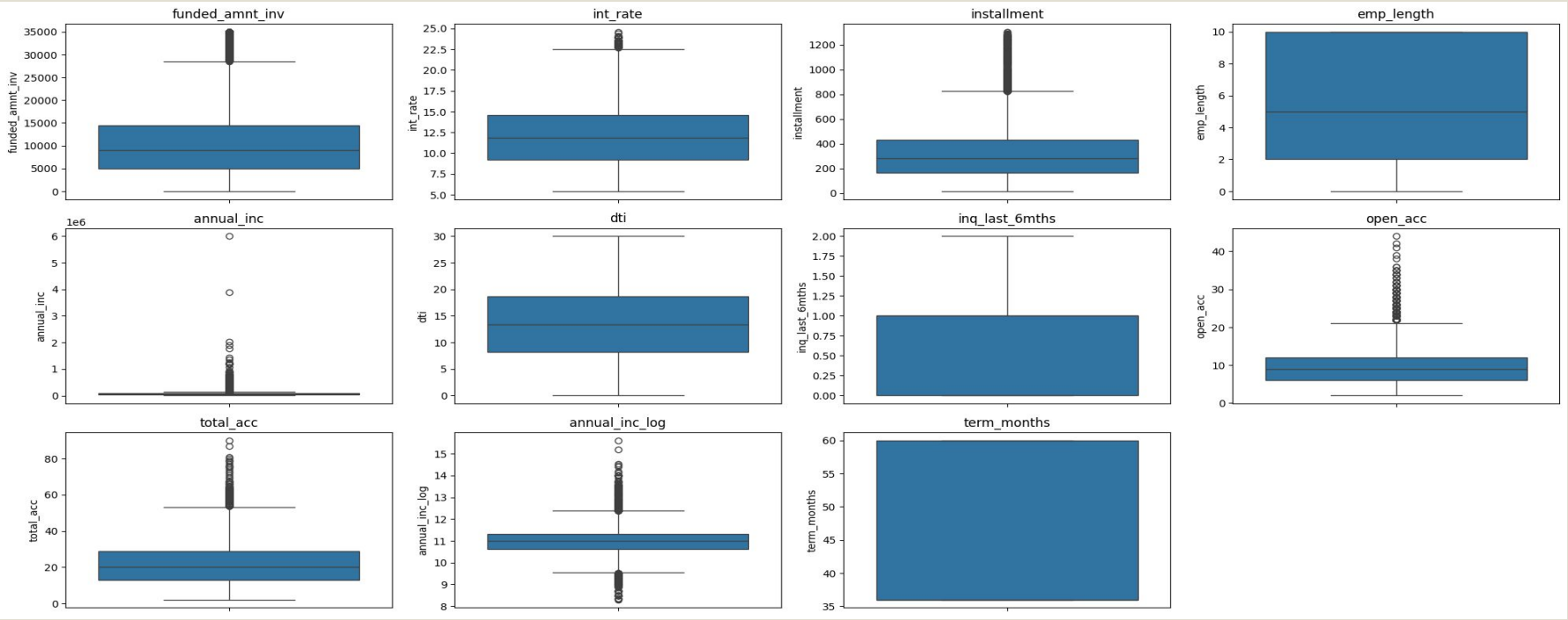
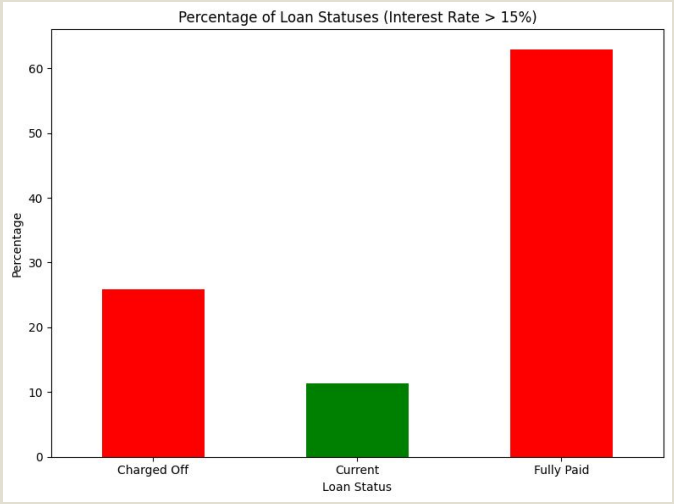


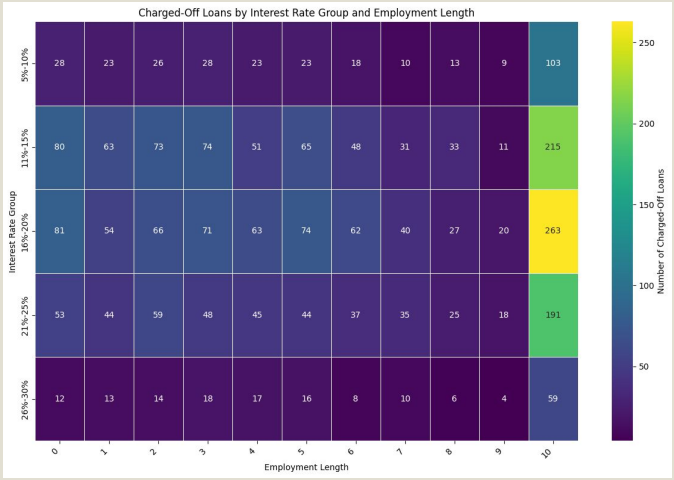
Image Post Treatment

UniVariate, Segmented Univariate & Bi-Variate Analysis

Observations for numerical values:(Segmented =Charged Off, Higher risk of default)

1. Funded Amount Analysis indicated that if the funding is above higher than the probability of default is higher , however the margin difference is not conclusive ,we may have to do hypothesis on this
2. Interest rate : If the interest is higher than 15% the likelihood of defaulting is higher. Hence lower rate of interest is one of ways to manage risk profiles. To analyse this further let's plot a. chart which loan status , interest rate and funding amount
3. Int Rate Groups 16%-20% are at high risk of default , validate payment capacity before giving loans at high ROI
4. Employment term more than 10 years and have high rate of interest i.e 16%-20% have high chance of risk
5. Loans taken under 31k have high risk of loan defaulting

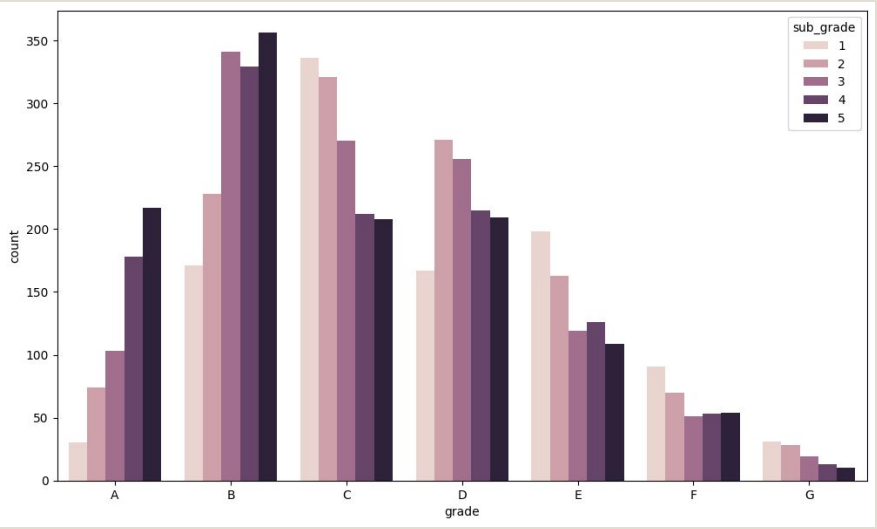


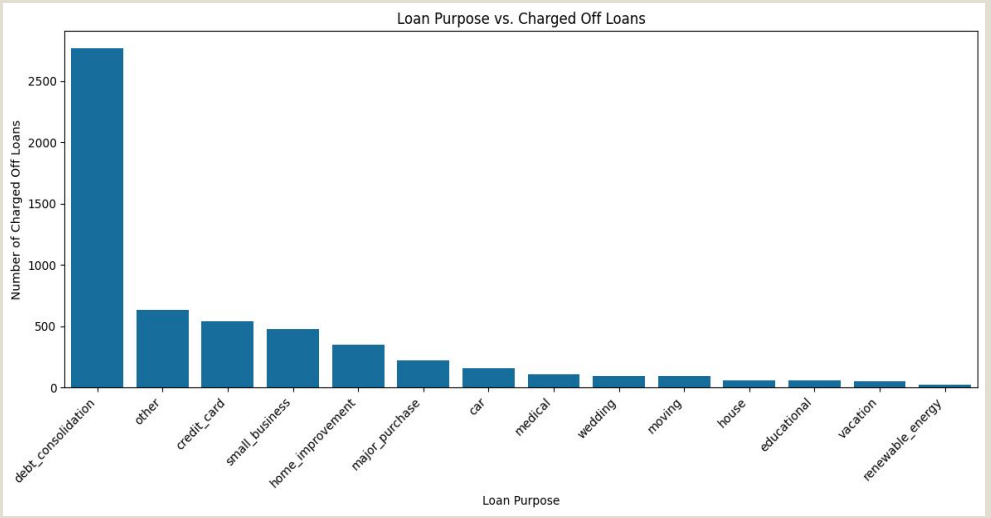


UniVariate, Segmented Univariate & Bi-Variate Analysis

Observations for Categorical Values:(Segmented =Charged Off, Higher risk of default)

1. LC Grade B and D have highest risk, within B and C , Sub Grades - B3 and C2 have the highest number of defaults
2. Loan applicants who stay in Rented houses have high probability of default
3. Loan taken for purpose of debt consolidation have very high probability of defaulting
4. While loan verification is important the impact is very limited on risk
5. CA State applicants have the highest amt of loan default risk

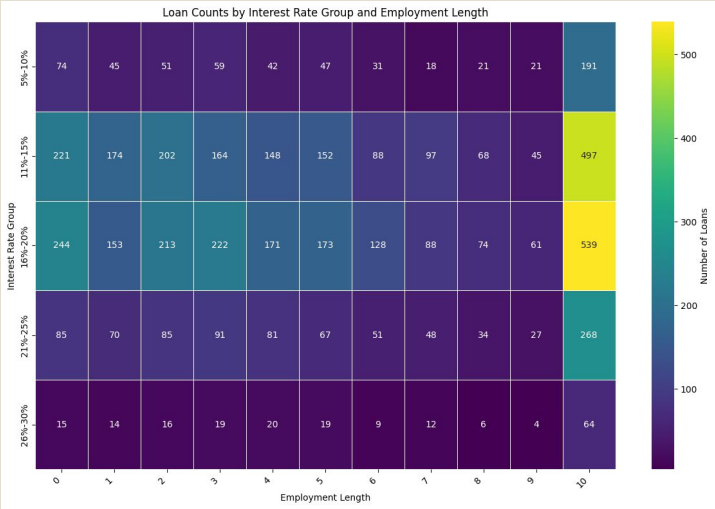


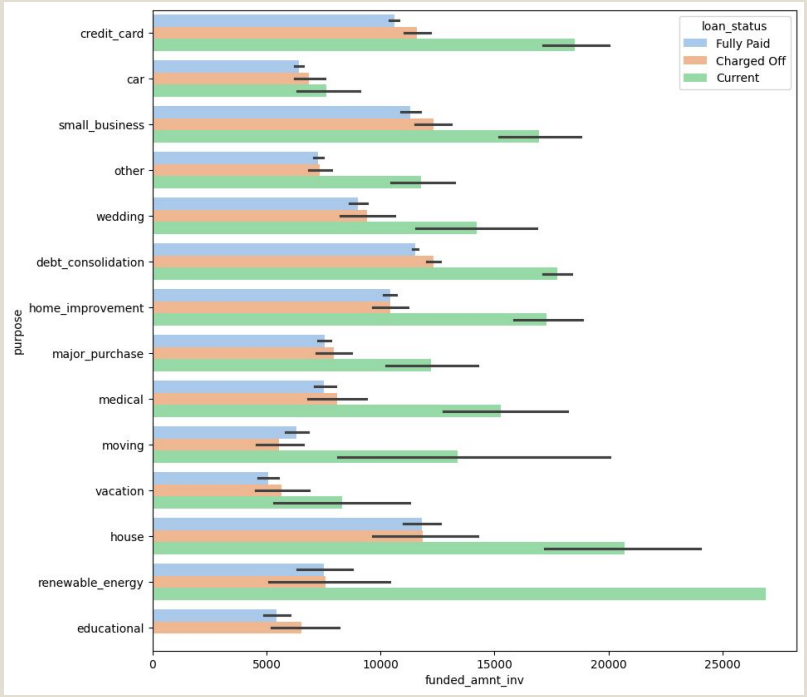


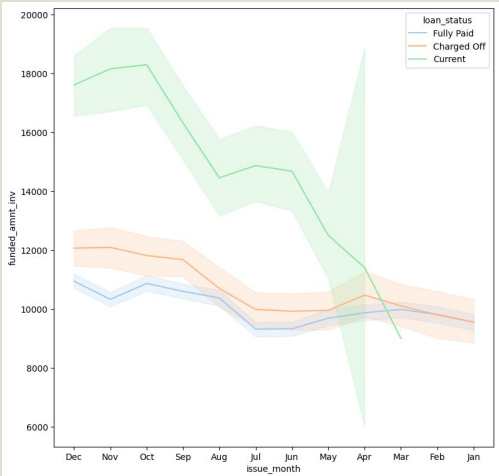
UniVariate, Segmented Univariate & Bi-Variate Analysis

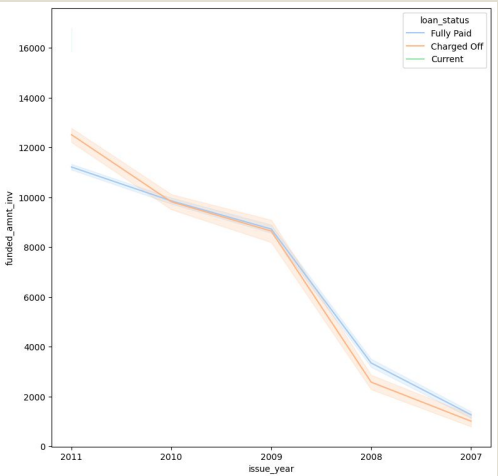
Observations for BiVariate Analysis:

1. Loan given in Dec Month has highest risk esp in the year 2011 Dec month applicants has highest risk of default
2. Loans given with funding amount <10k has highest risk, we should do through validation before given low value risk
3. Home Improvement, Credit Card and debt consolidation loan given below a certain threshold have clear indication of default
4. Based on home ownership , customer who stay in mortgage house and have income range from 70-80k have high risk of default
5. Higher the funding amount, higher the interest rate, however highest risk is when int rate is btw 15-16%
6. When loan is taken for small business , debt consolidation , house have a risk of loan defaulting
7. Loan when given to applicants with 10+ years employment exp and rate of interest 16-20% the risk of defaulting is higher.


Loan Counts by Interest Rate Group and Employment Length
Interest Rate Group
5%-10%
11%-15%
16%-20%
21%-25%
26%-30%
74 45 51 59 42 47 31 18 21 21 191
221 174 202 164 148 152 88 97 68 45 497
244 153 213 222 171 173 128 88 74 61 539
85 70 85 91 81 67 51 48 34 27 268
15 14 16 19 20 19 9 12 6 4 64
0 1 2 3 4 5 6 7 8 9 10
Employment Length
Number of Loans
500
400
300
200
100

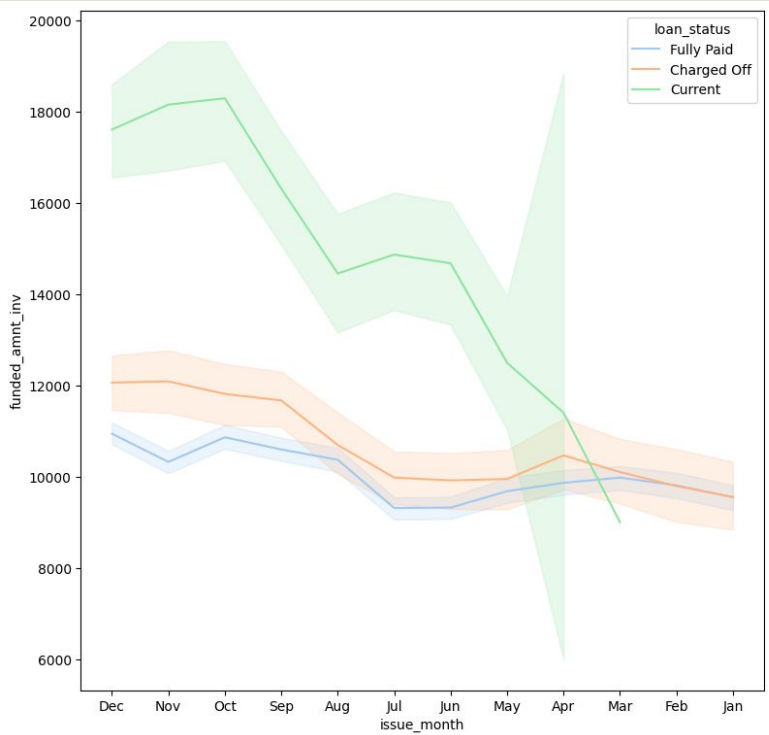

loan_status
Fully Paid
Charged Off
Current
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
purpose
0
5000
10000
15000
20000
25000
funded_amnt_inv

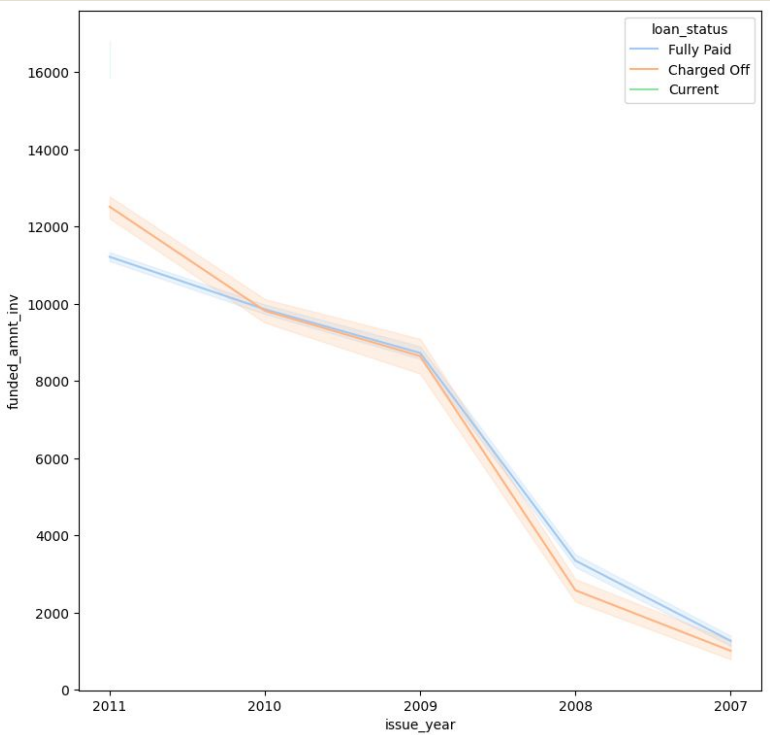

loan_status
Fully Paid
Charged Off
Current
funded_amnt_inv
20000
18000
16000
14000
12000
10000
8000
6000
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
issue_month


loan_status
Fully Paid
Charged Off
Current
funded_amnt_inv
16000
14000
12000
10000
8000
6000
4000
2000
0
2011
2010
2009
2008
2007
issue_year

Derived Metrics

When we separate Month and Year from issue_d we can analyse how it impact loan risk





THANK YOU