Lending Club Case Study

Version 1.0

Author: Shangar Devi Natarajan
Nikhil George John

Problem Statement

▪ Able to identify risky loan applicants, such loans can be reduced thereby cutting down the amount of credit loss.

▪ Risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

▪ 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.

Data Handling and Cleaning

No missing values or incorrect data types before we proceed to the analysis stage

Column wise:

- All null value column has been removed.

- Dropped the columns which has customer behavior and will not be available for new customers. There is no point of analyzing those column where data will not be available for new customers

- Dropped the columns, if all values on column is same.

- Dropped column which has NA and 0 which will not be useful for analysis

- Removed desc column as NLP is not part of this project.

- Removed url column

Correcting Incorrect Data Types:

- Removed months string from terms

- Removed % from int rates

- Removed all special char from emp_length

- Replace verification_status "source verified" value with verified

Imputing Missing Values:

- Replacing emp length NA or empty with 0 value

- Replace home_ownership None value with Others

Data Handling and Cleaning

Row wise:

- Dropped current loan status as this is will not add value. Interested on charged off and fully paid loans.
- Checked id and member Id are unique. There is no duplicate rows because id and member_id rows as same as no of rows.
- Checked the no of rows which has missing NA or null. There are no empty rows in dataset.

Sanity Checks:

- All numeric data should be positive. Confirmed all are positive integers

Derived Column:

- Extracted month and year from issue_d.

Outliers Removal:

- All numeric data has been distributed well except annual_inc. So removing outlier on annual_inc. Data kept till 98 percentile

Analysis Approach – Univariate Analysis

- Distribution of loan_amnt , funded_amnt and funded_amnt_inv looks similar.

- Derived metrics – Planned to DTI. However data already has DTI.

Continuous data observation:

- Most of the Loan amounts are between 5000 15000
- Most of the funded amount are between 5000 15000
- Most of the interest rates between 8 to 14%
- Most of the installments are between 200 400
- Most of the DTI are between 12 18
- Most of the Annual income is in the range of 30,000 and 60,000

Analysis Approach – Univariate Analysis

Categorical data observation:

- 76% of loans are 36 months term and 24% loans are 60 months term
- Loan applicants contribution are Grade A (30%), Grade B(26%), Grade C (20%)
- 22% of employees are 10+ years and 20% employees are 1 year.
- 49% of loan applicants has rented home and 43% is mortgage
- 56% of source has been verified and 44% is not verified
- 85% of loan is fully paid and 15% of loan is charged off.
- 47% of purpose is dept consolidation and 13% is for credit card
- Given data set, 53% of loan provided on year 2011 and 30% of loan provided on 2010
- No of loan applied is slightly high on Oct, Nov and Dec

Analysis Approach – Segmented univariate analysis

Segmented variate analysis Observation:

- 60 months term charged off is 25% compare with 36 months term which is 11%. Prefer term is 36 months.

- Lower grade charged off rate is way high than higher grade. charged off for Grade E (27%), F (33%) and G(35%) , how ever Grade A (6%), B(12%), C(17%), D (22%)

- Sub grade: sub grade is aligning with Grade behaviour. charged off percentage is in ascending order from A1 to F5. Slight diffrence behaviour on G1 to G5 grade. However number of records contributed to G subgrade is way lower than other grades. we can ignore subgrade G*.

- Not much variant in charge_off percentage based on Emp_length

- Not much variant in charge_off percentage based on home_ownership

- verification status verified is just 3% higher than non_verified no major variance

- High charged off percentage purpose are small_business (28%). renewable_energy (19%)   educational (17%)

- addr_state   NE state has high charged off rate 60%. However only 5 records contributed. So Ignoring.

- pub_rec_bankruptcies 2 , charged off is 40% which is high . pub_rec_bankruptcies 1 is 22% pub_rec_bankruptcies 0 is 14%

- No of loans applied on Sep, Oct, nov, dec months are high compare to other months.

-

Analysis Approach – Bivariate Analysis:

Bivariate Analysis Observation

Prerequisite:

Converted numical variable in to categorical dti, funded_amt, Interested_rate and income. Analyzed all possible combination of 'emp_length','grade','home_ownership','pub_rec_bankruptcies','term','purpose','verification_status','month','dti_slab','funded_amnt_slab','int_rate_slab','income_slab' where chargedoff rate is more than 30%. this has been achived via loop.

Observations:

- funded_amnt_slab (30000.0, 35000.0], emp_len = 6 years and charged off is 35%
- [20.0, 24.0] interest slab has highest charged off rate
- lower grade has high charged off percentage
- Grade D and E charged off percentage is high for all months along with Grade F & G.
- pub_rec_bankrupcies = 1 and home_ownership is OTHER and charged off percentage is 67%
- int_rate slab (20.0, 24.0), home_ownership = OWN, RENT, charged off is very high 50%, 40%
- funded_amnt_slab (5000.0, 10000.0], pub_rec_bankrupcies = 2.0, charged off is 50%
- funded_amnt_slab 30000.0, 35000.0 pub_rec_bankrupcies = 2.0, charged off is 67%

Analysis Approach – Bivariate Analysis:

Bivariate Analysis Observation

- dti_slab 10.0, 15.0 and purpose = renewable_energy, charged off percentage is 35.
- dti_slab 25.0, 30.0 and purpose = renewable_energy, charged off percentage is 50.
- Purpose = small_business and funded_amnt_slab is [10000.0, 15000.0], [15000.0, 20000.0] [20000.0, 25000.0] charged off is 30+ %
- purpose = vacation, funded_amnt_slab = [15000.0, 20000.0], charged off = 42%
- purpose = medical,   funded_amnt_slab =(15000.0, 20000.0], charged_off = 50%, funded_amnt_slab = (30000.0, 35000.0] charged_off = 67%
- purpose = small_business funded_amnt_slab = (30000.0, 35000.0] , charged_off = 46%,
- int_rate_slab   (15.0, 20.0] and (20.0, 24.0], purpose = small_business , educational   , house , medical, moving , charged off percentage is more than 35%
- funded_amnt_slab (30000.0, 35000.0] , november and december, charged off is more than 35%
- int_rate_slab is 20.0, 24.0, from 7th month to 12th month, charged off is more than 35%
- funded_amnt_slab (30000.0, 35000.0], dti_slab   (25.0, 30.0], (15.0, 20.0] (20.0, 25.0] , charged off is more than 30%
- int_rate_slab (20.0, 24.0], all dti slab, charged off rate is more than 40%

Analysis Approach – Bivariate Analysis:

## Bivariate Analysis Observation

- Small business has more than 30% of charged off for most of the income slab.

- Lower income slab, Medical purpose has 30% of charged off.

- 5 to 10% interest rates, charged off percentage is low across all income slabs.

- When income increases charged off decreases for 5 to 10% interest rates.

- dti_slab 25 to 30, income slab 113800 to 132100, charged off is 40+%

- int_rate 20 to 24 slab has more than 35% of charged off regardless of employee experience

- term 60 months, home_ownership=other, charged off rate is very high

- higher slab of funded amount , bankrupcy = 1.0, charged off percentage is 45

- 10k to 15kof funded amount , bankrupcy = 2.0, charged off percentage is 90+

- int_rate_slab 20 to 24 int_rate slab, renewable energy purpose, charged off rate is 95+%

- High interest slab charged off rate is high. lending club are on track to identify person is likely to default and provided higher interest rates.

- purpose medical, funded_amt = 30k to 35k, charged off rate is >65%

- purpose medical, funded_amt = 25k to 30k, charged off rate is >45%

- purpose wedding , charged off rate is more than 30% for 30k to 35k fund amount.

- purpose vacation, charged off rate is more than 30% for 15k to 20k fund amount

- renewable energy purpose, dti_slab 25 to 30, charged off is 50%across all dti slab, small business charged off is more than 25%
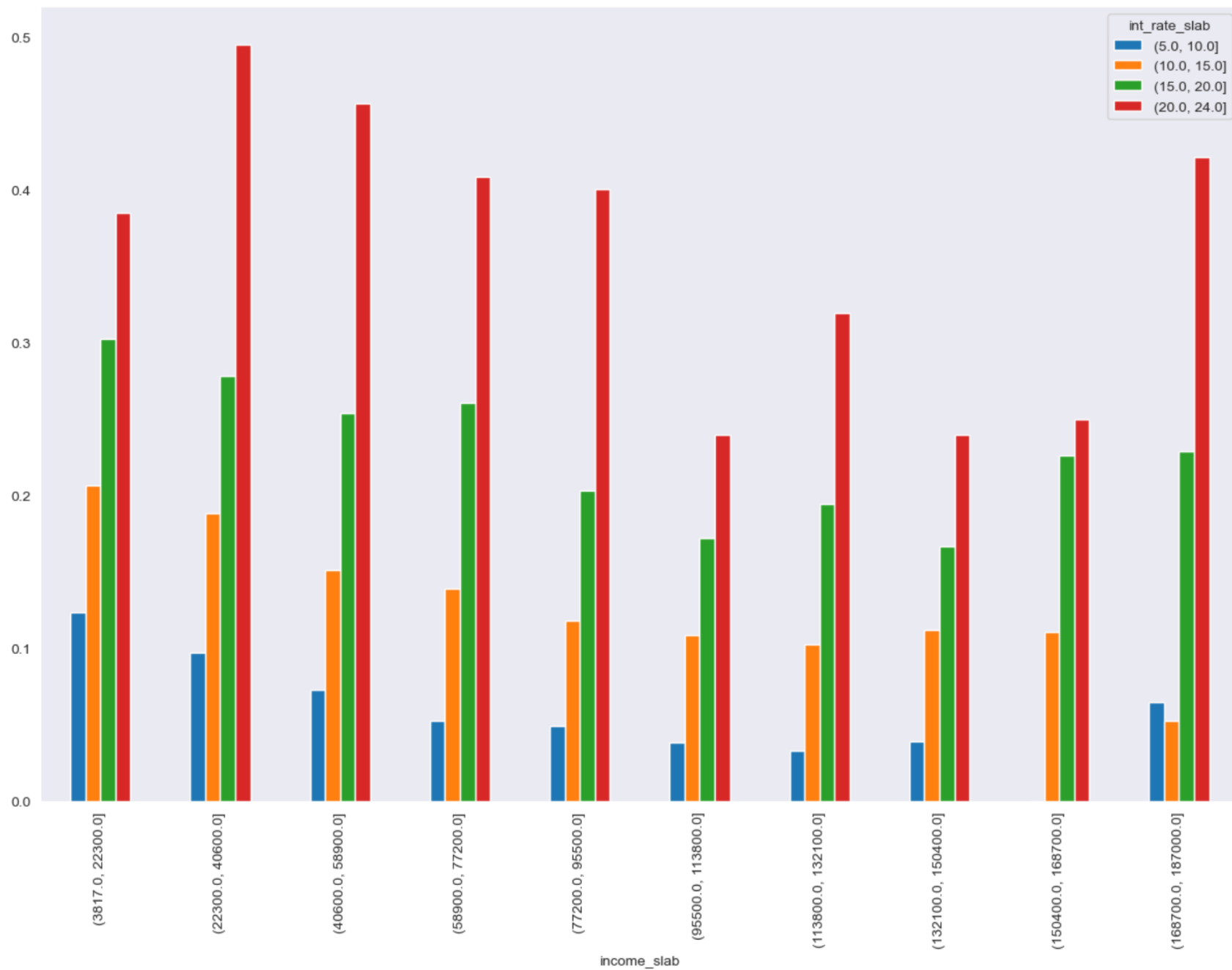
Business Insights:

- Most of the Loan amounts are between 5000 15000
- Most of the interest rates between 8 to 14%
- Most of the Annual income is in the range of 30,000 and 60,000
- 76% of loans are 36 months term and 24% loans are 60 months term
- 85% of loan is fully paid and 15% of loan is charged off.
- 60 months term charged off is 25% compare with 36 months term which is 11%. Prefer term is 36 months.
- Lower grade charged off rate is way high than higher grade.
- High charged off percentage purpose are small_business (28%). renewable_energy (19%)   educational (17%)
- No of loans applied on Sep, Oct, Nov, Dec month is high compare to other months.
- Grade D and E charged off percentage is high for all months along with Grade F & G.
- int_rate slab (20.0, 24.0), home_ownership = OWN, RENT, charged off is very high 50%, 40%
- Small business has more than 30% of charged off for most of the income slab.
- Lower income slab, Medical purpose has 30% of charged off.
- 5 to 10% interest rates, charged off percentage is low across all income slabs.
- When income increases charged off decreases for 5 to 10% interest rates.
- 10k to 15kof funded amount , bankrupcy = 2.0, charged off percentage is 90+
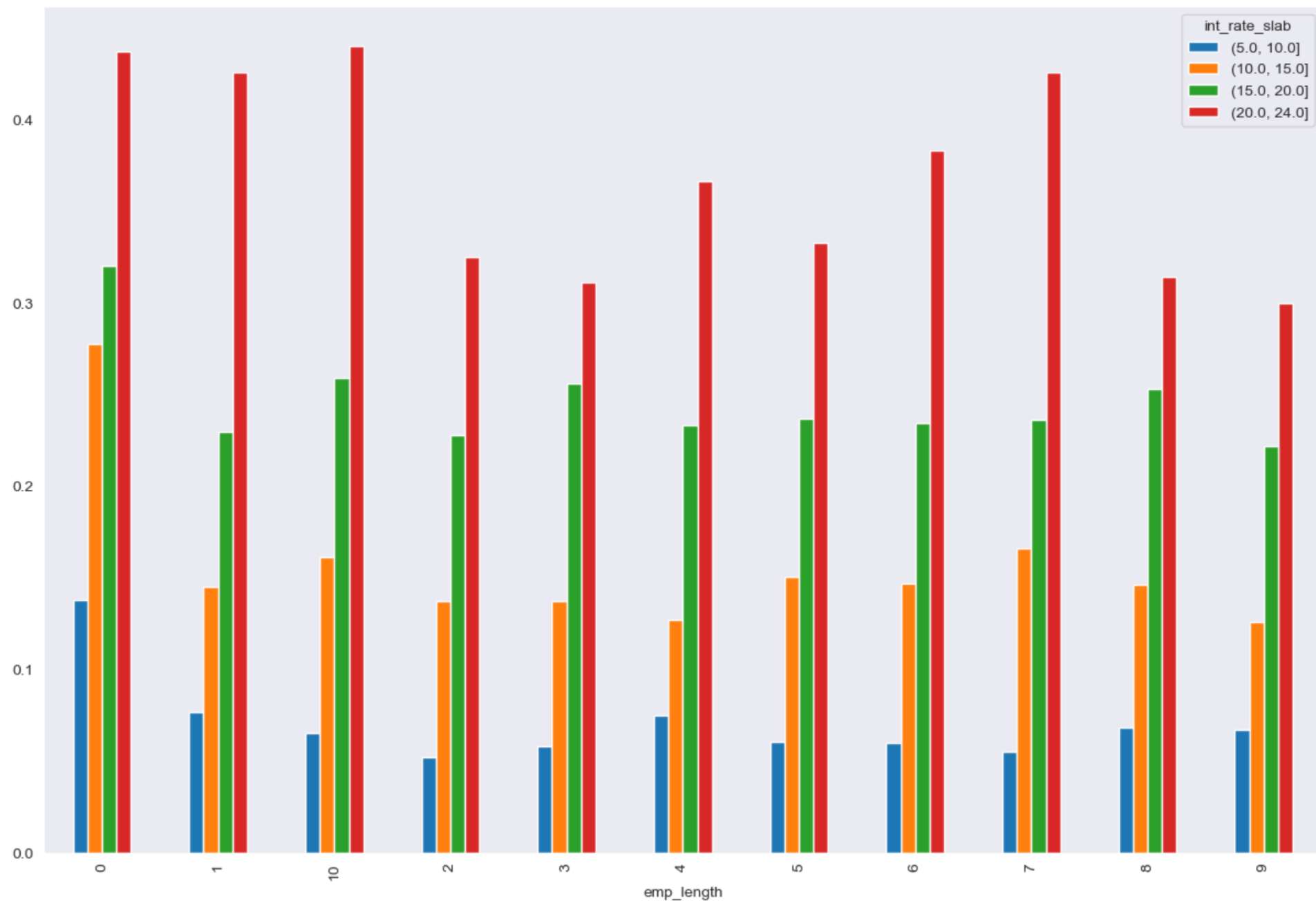
Business Insights:

- High interest slab charged off rate is high. lending club are on track to identify person is likely to default and provided higher interest rates.

- purpose medical, funded_amt = 30k to 35k, charged off rate is >65%

- purpose medical, funded_amt = 25k to 30k, charged off rate is >45%

- purpose wedding , charged off rate is more than 30% for 30k to 35k fund amount.

- purpose vacation, charged off rate is more than 30% for 15k to 20k fund amount
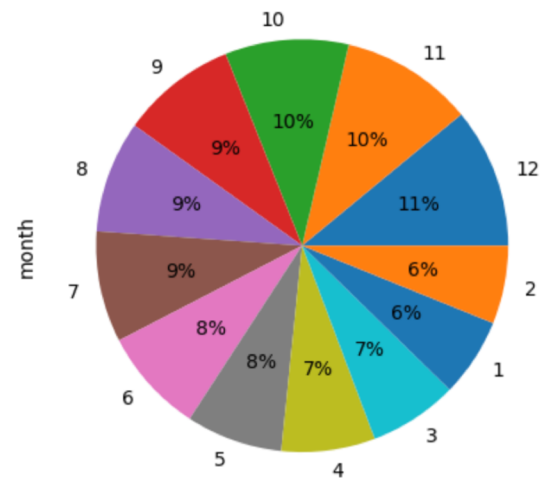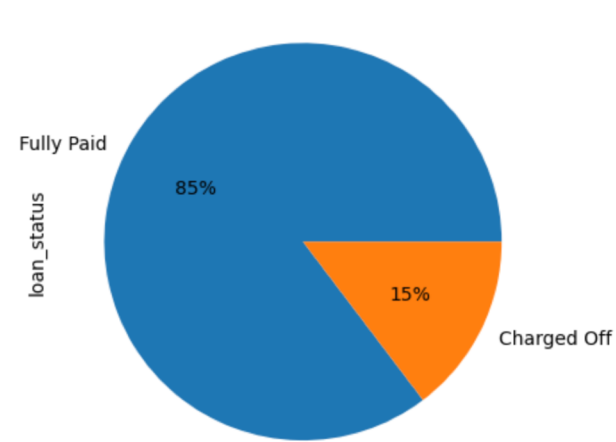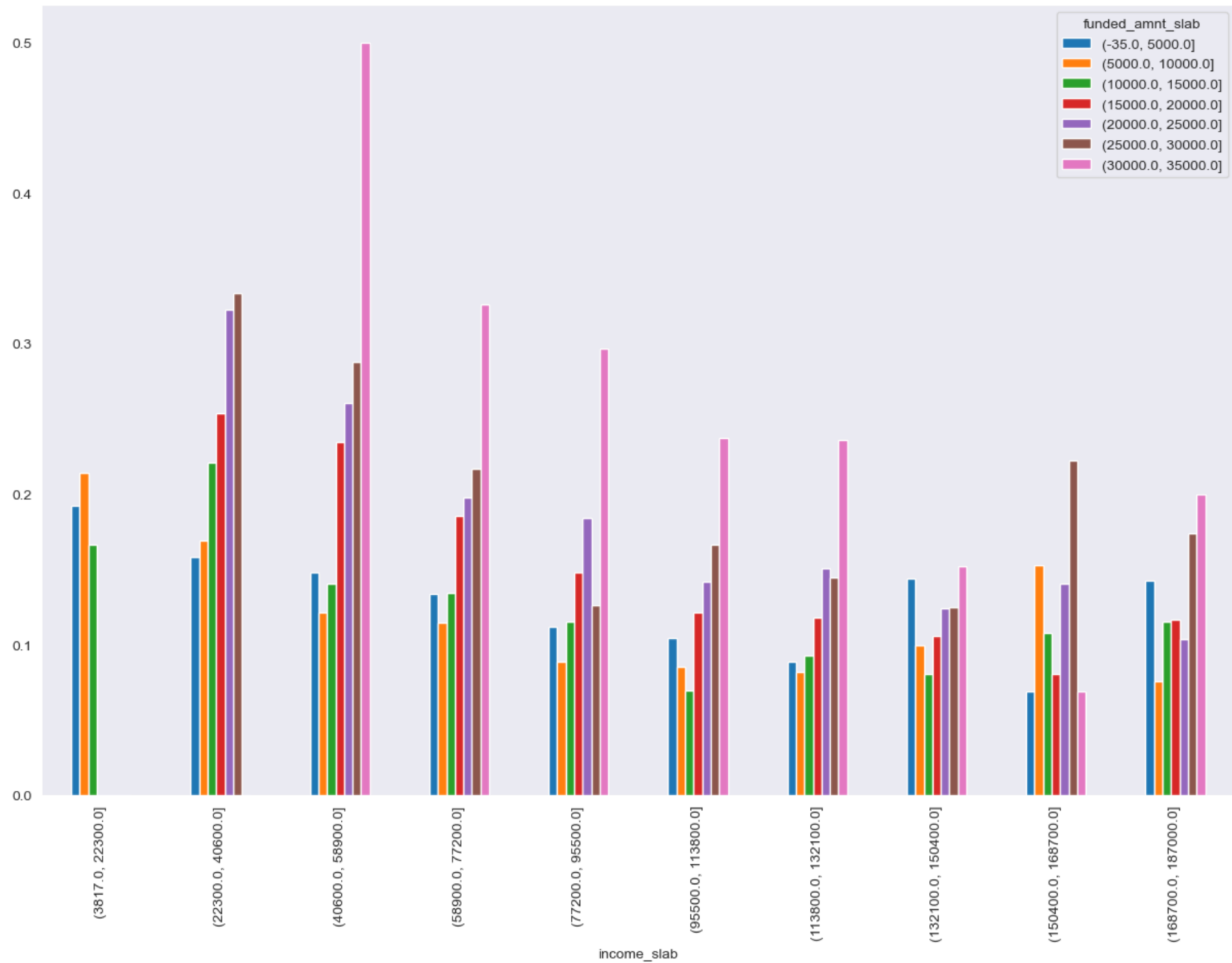
loan_status grade	Charged Off	Fully Paid	Tot_entries
A	6.06	93.94	9905
B	12.26	87.74	11452
C	17.43	82.57	7676
D	22.21	77.79	4976
E	27.03	72.97	2579
F	32.94	67.06	935
G	34.15	65.85	284

Thank You!