Lending Club Case Study

By -Neel Narayanan

The problem

Company

This company 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.

Context

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

Problem statement

To identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Challenges deep-dive

Challenge 1

Data quality Issues

Data quality issues are overlooked or are not identified correctly such as outliers, missing values and other data quality issues.

Challenge 2

Data cleaning & manipulation

Missing value imputation, outlier treatment and other kinds of data redundancies, etc.

Challenge 3

Data Analysis

The analysis successfully identify at least the 5 important driver variables (i.e. variables which are strong indicators of default).

Solution

EDA

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Implementation

Data Understanding

Number of rows x columns — Rows : 39717 X Columns: 111

Number of all null columns — 54 columns have been deleted

Number of all null rows — 0 rows deleted

Data Types — `float64(20), int64(13), object(24)`

Data Understanding

Important Features & Descriptions

annual_inc
The self-reported annual income provided by the borrower during registration.
desc
Loan description provided by the borrower
dti
A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
emp_length
Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title
The job title supplied by the Borrower when applying for the loan.*
funded_amnt
The total amount committed to that loan at that point in time.
funded_amnt_inv
The total amount committed by investors for that loan at that point in time.
grade
LC assigned loan grade

home_ownership
The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
installment
The monthly payment owed by the borrower if the loan originates.
int_rate
Interest Rate on the loan
issue_d
The month which the loan was funded
loan_amnt
The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status
Current status of the loan
sub_grade
LC assigned loan subgrade
term
The number of payments on the loan. Values are in months and can be either 36 or 60.
verification_status
Indicates if income was verified by LC, not verified, or if the income source was verified
zip_code
The first 3 numbers of the zip code provided by the borrower in the loan application.

Data Cleaning & Manipulations

Data Sanitization

#Cleaning simply all Null columns and rows which are not useful

There is a drop from 111 columns to 57 columns

#remove columns that have only 1 feature(redundant columns)

There is a drop from 57 columns to 48 columns

Dropping Columns based on missing data being more than 90%

There is a drop from 48 columns to 45 columns

```
#   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        39717 non-null  int64
 1   member_id                 39717 non-null  int64
 2   loan amnt                 39717 non-null  int64
 3   funded amnt               39717 non-null  int64
 4   funded amnt inv           39717 non-null  float64
 5   term                      39717 non-null  object
 6   int_rate                  39717 non-null  object
 7   installment               39717 non-null  float64
 8   grade                     39717 non-null  object
 9   sub grade                 39717 non-null  object
 10  emp_title                 37258 non-null  object
```

```
11  emp_length              38642 non-null  object
 12  home_ownership          39717 non-null  object
 13  annual_inc              39717 non-null  float64
 14  verification_status     39717 non-null  object
 15  issue_d                 39717 non-null  object
 16  loan_status             39717 non-null  object
 17  url                     39717 non-null  object
 18  desc                    26777 non-null  object
 19  purpose                 39717 non-null  object
 20  title                   39706 non-null  object
 21  zip_code                39717 non-null  object
 22  addr_state              39717 non-null  object
 23  dti                     39717 non-null  float64
 24  delinq_2yrs             39717 non-null  int64
 25  earliest_cr_line        39717 non-null  object
 26  inq_last_6mths          39717 non-null  int64
 27  open_acc                39717 non-null  int64
 28  pub_rec                 39717 non-null  int64
 29  revol_bal               39717 non-null  int64
 30  revol_util              39667 non-null  object
 31  total_acc               39717 non-null  int64
 32  out_prncp               39717 non-null  float64
 33  out_prncp_inv           39717 non-null  float64
 34  total_pymnt             39717 non-null  float64
 35  total_pymnt_inv         39717 non-null  float64
 36  total_rec_prncp         39717 non-null  float64
 37  total_rec_int           39717 non-null  float64
 38  total_rec_late_fee      39717 non-null  float64
 39  recoveries              39717 non-null  float64
 40  collection_recovery_fee  39717 non-null  float64
```

```
41  last_pymnt_d              39646 non-null  object
 42  last_pymnt_amnt           39717 non-null  float64
 43  last_credit_pull_d        39715 non-null  object
 44  pub_rec_bankruptcies      39020 non-null  float64
dtypes: float64(15), int64(10), object(20)
```

Manipulations

Dictionaries used to map the columns that are below

Term	Map value
36 months	1
60 months	2

Grades	Map Values
A	7
B	6
C	5
D	4
E	3
F	2
G	1

Home Ownership	Map Values
NONE	0
OTHER	1
ANY	2
RENT	3
MORTGAGE	4
OWN	5

Verification : -
Source Verified = 2 , Verified = 1, Not Verified = 0

Sub Grades : -
From A1 = 34 to G5 = 0

Issue Date(Loan issue date) : -
From Jan = 1 to Dec = 12

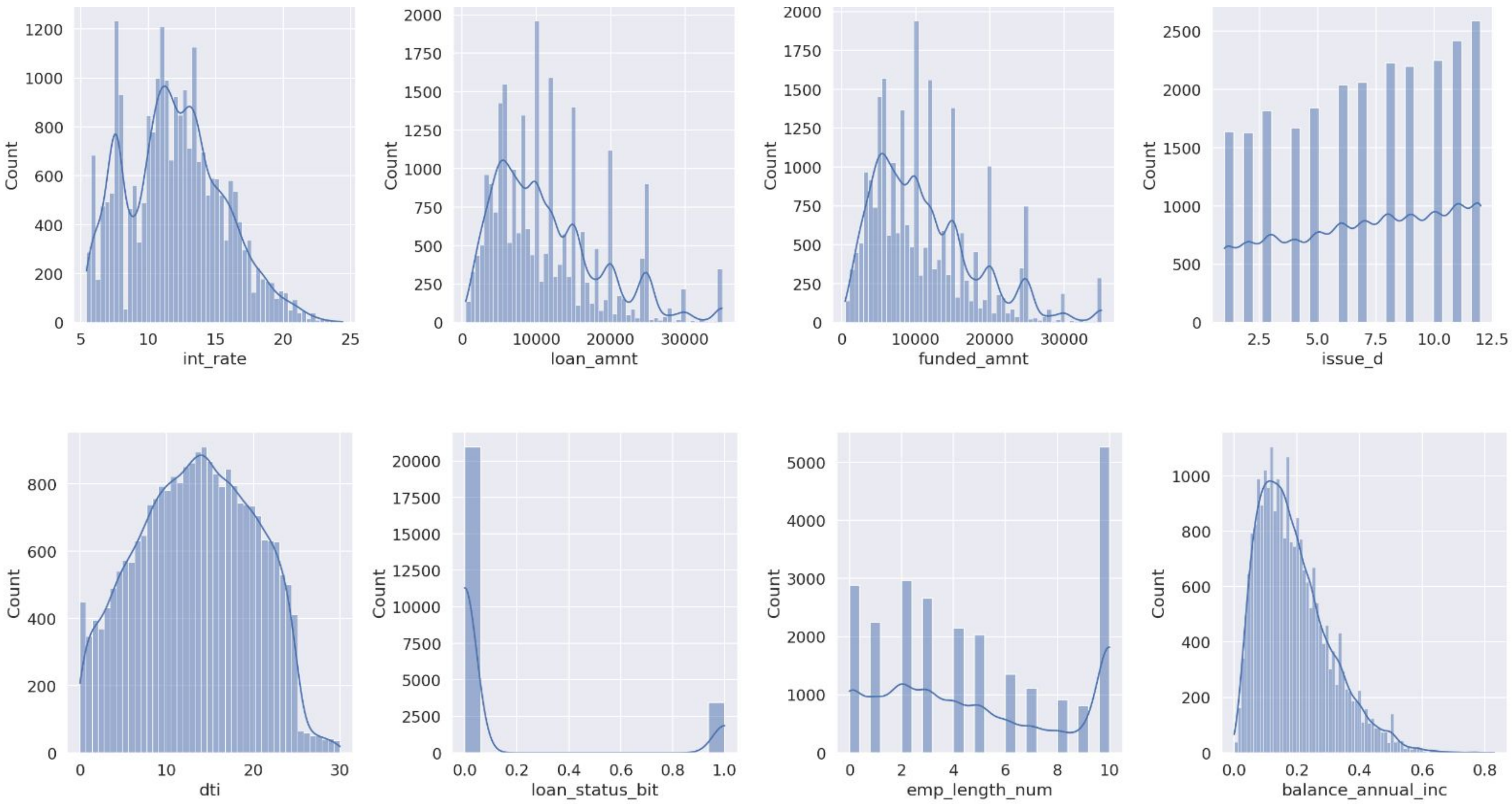
Loan Status : -
Fully Paid = 0 , Charged Off = 1 , Current = 2

Data visualization and feature engineering

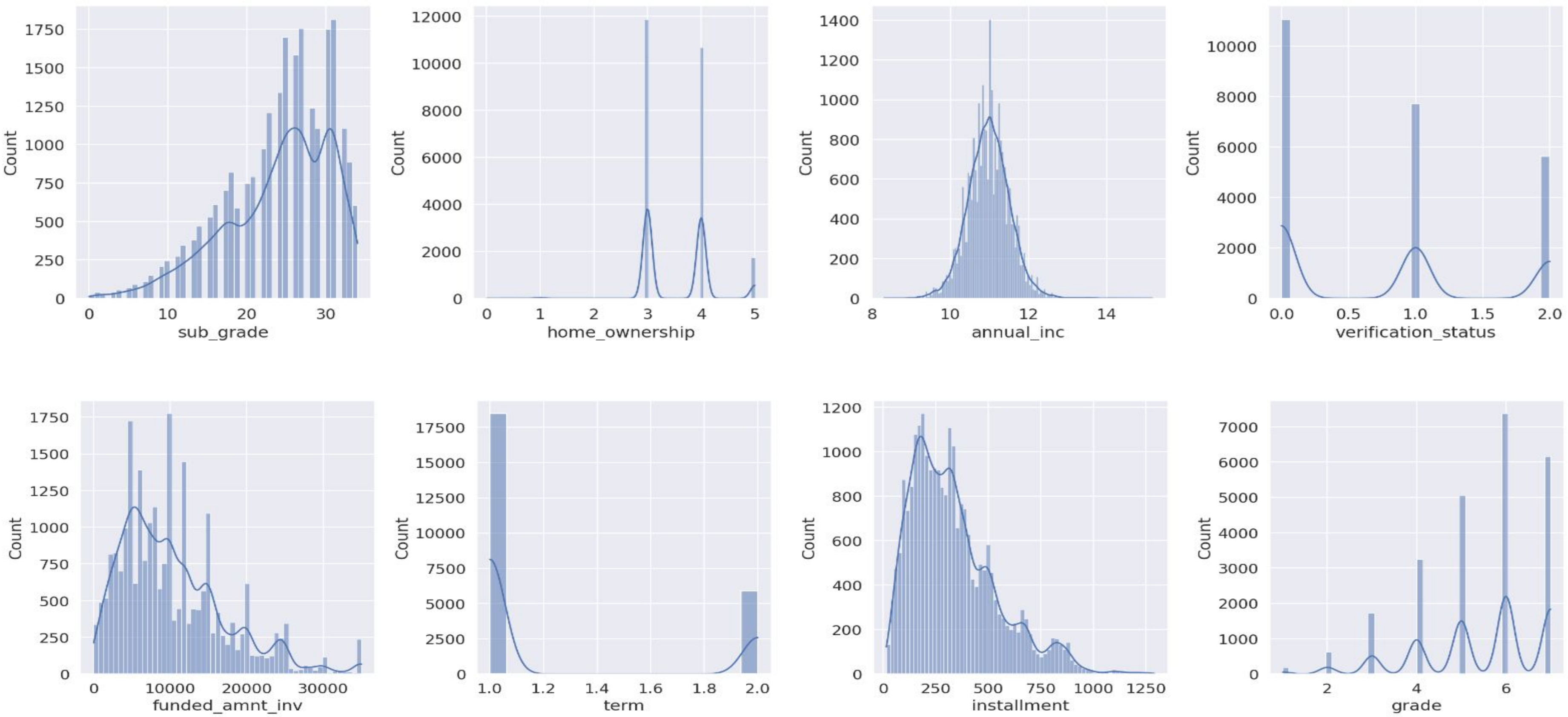
Univariate Analysis and distribution of features

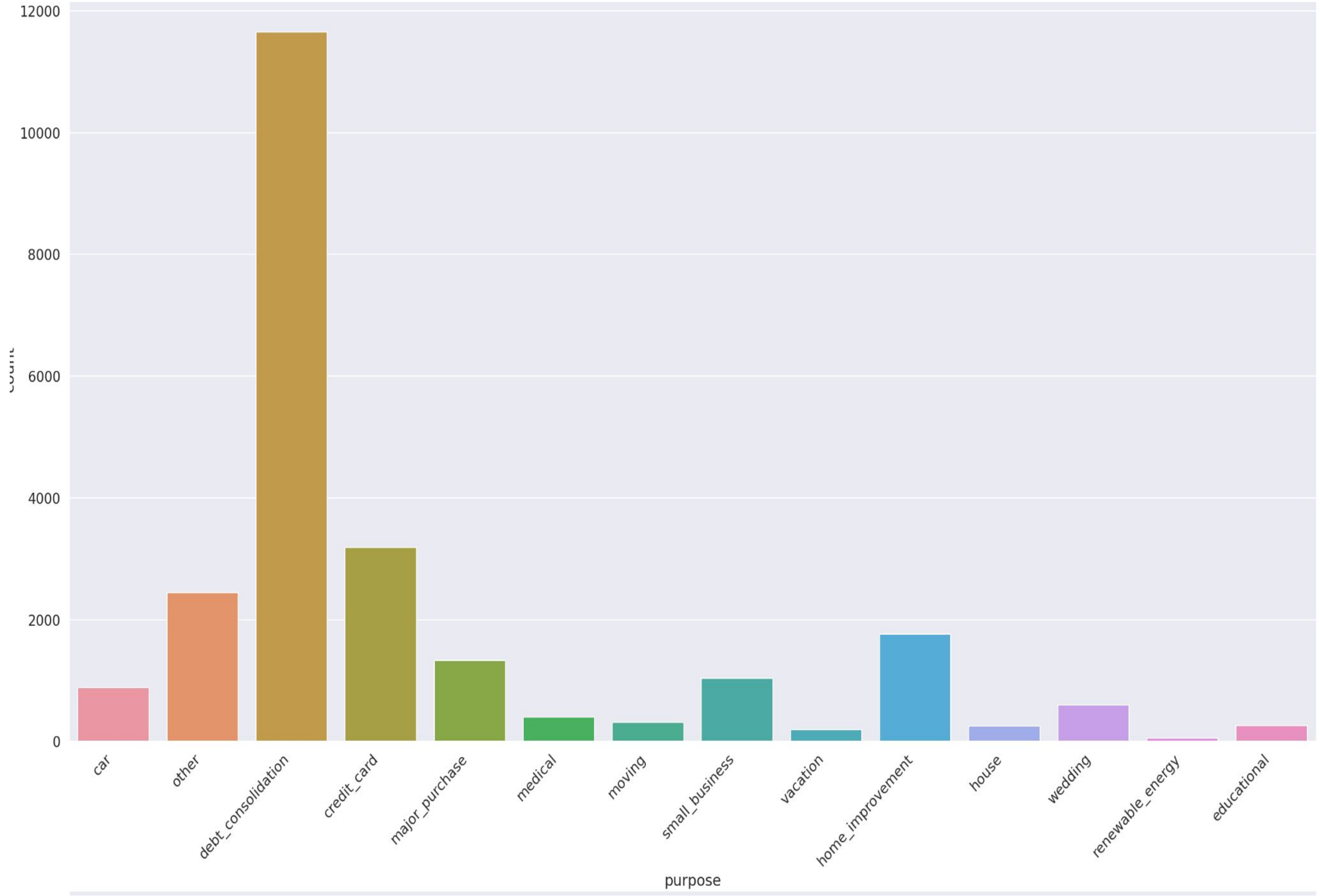
Insights of the below plots we have following conclusions

- The trend of Loan ,Funded and invested amount features are identical
- Found incremental trend as the grade and subgrade increases
- · There are lot of customers with income not verified
- · Identified that the average customers are having 16%-18% DTI ratio where 36% above are termed as high risk of defaulting
- There are huge sum distribution of cases with fully paid loan settlements
- There are huge sum distribution of cases who are have long employment
- There are huge sum distribution of cases with rented and mortgage houses



Count
int_rate
loan_amnt
funded_amnt
issue_d
dti
loan_status_bit
emp_length_num
balance_annual_inc

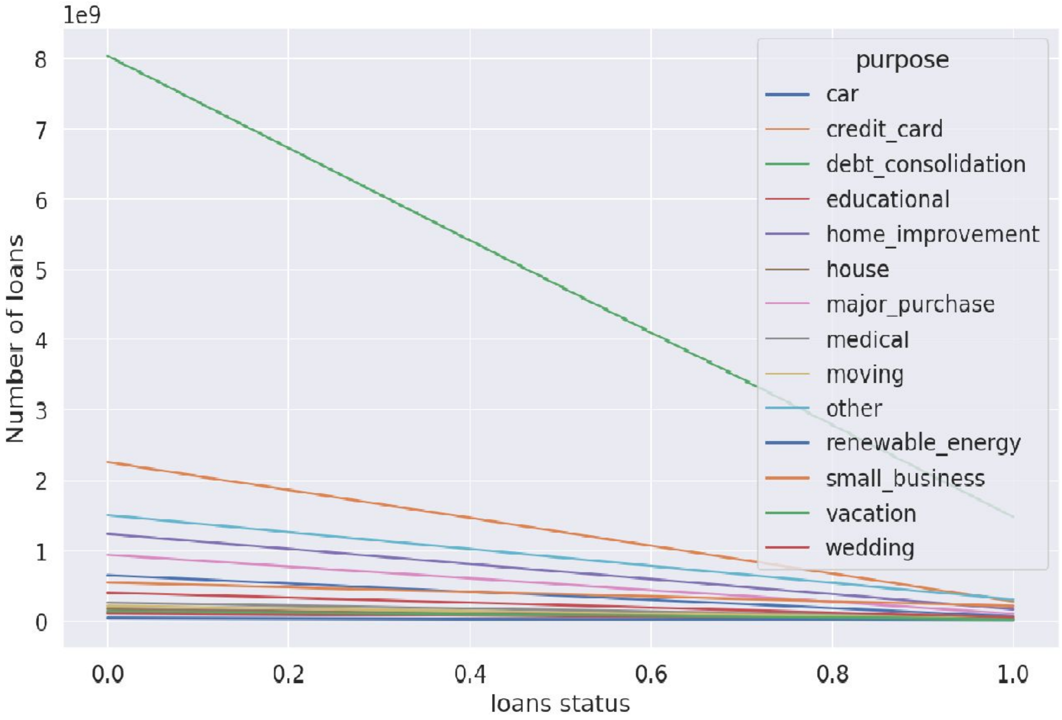

Count
sub_grade
0 10 20 30
0 250 500 750 1000 1250 1500 1750
Count
home_ownership
0 1 2 3 4 5
0 2000 4000 6000 8000 10000 12000
Count
annual_inc
8 10 12 14
0 200 400 600 800 1000 1200 1400
Count
verification_status
0.0 0.5 1.0 1.5 2.0
0 2000 4000 6000 8000 10000
Count
funded_amnt_inv
0 10000 20000 30000
0 250 500 750 1000 1250 1500 1750
Count
term
1.0 1.2 1.4 1.6 1.8 2.0
0 2500 5000 7500 10000 12500 15000 17500
Count
installment
0 250 500 750 1000 1250
0 200 400 600 800 1000 1200
Count
grade
2 4 6
0 1000 2000 3000 4000 5000 6000 7000



Debt Consolidation and credit related loans are higher in the distribution compared to other features with which we can how many are defaulted in the next analysis

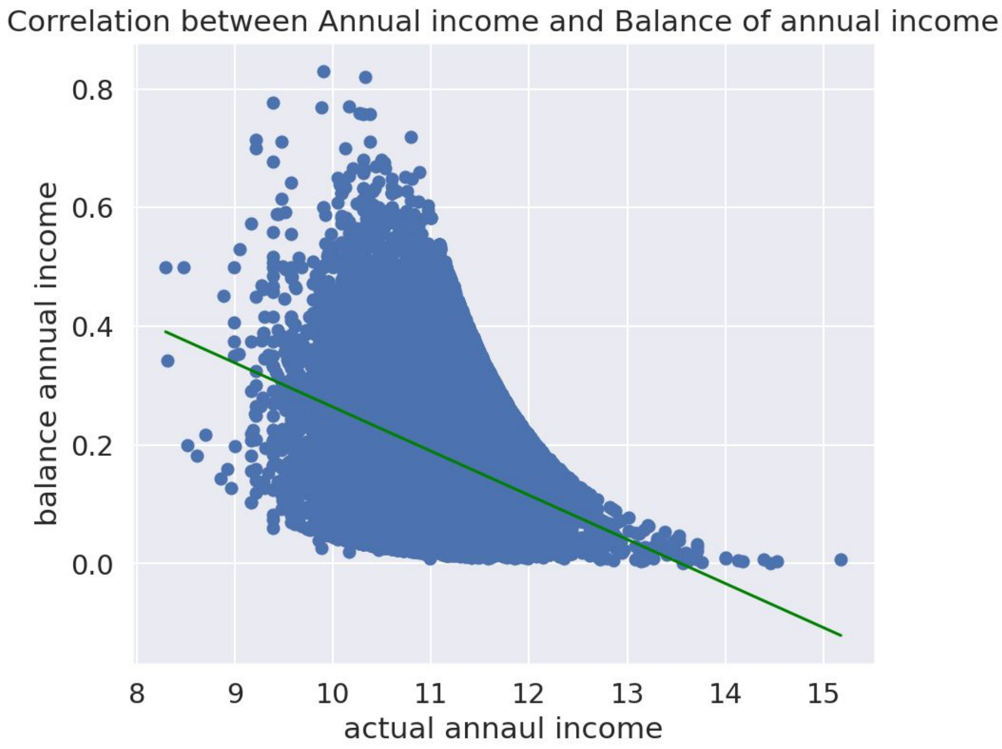
Bivariate Analysis & Multivariate Analysis

By analysing that how purposes are distributed over the overall counts of loans and its status
We got an insight that most number of loans are debt_consolidation and credit card which have been defaulted.



Loan Status : -
Fully Paid = 0 , Charged Off = 1 ,
Current = 2

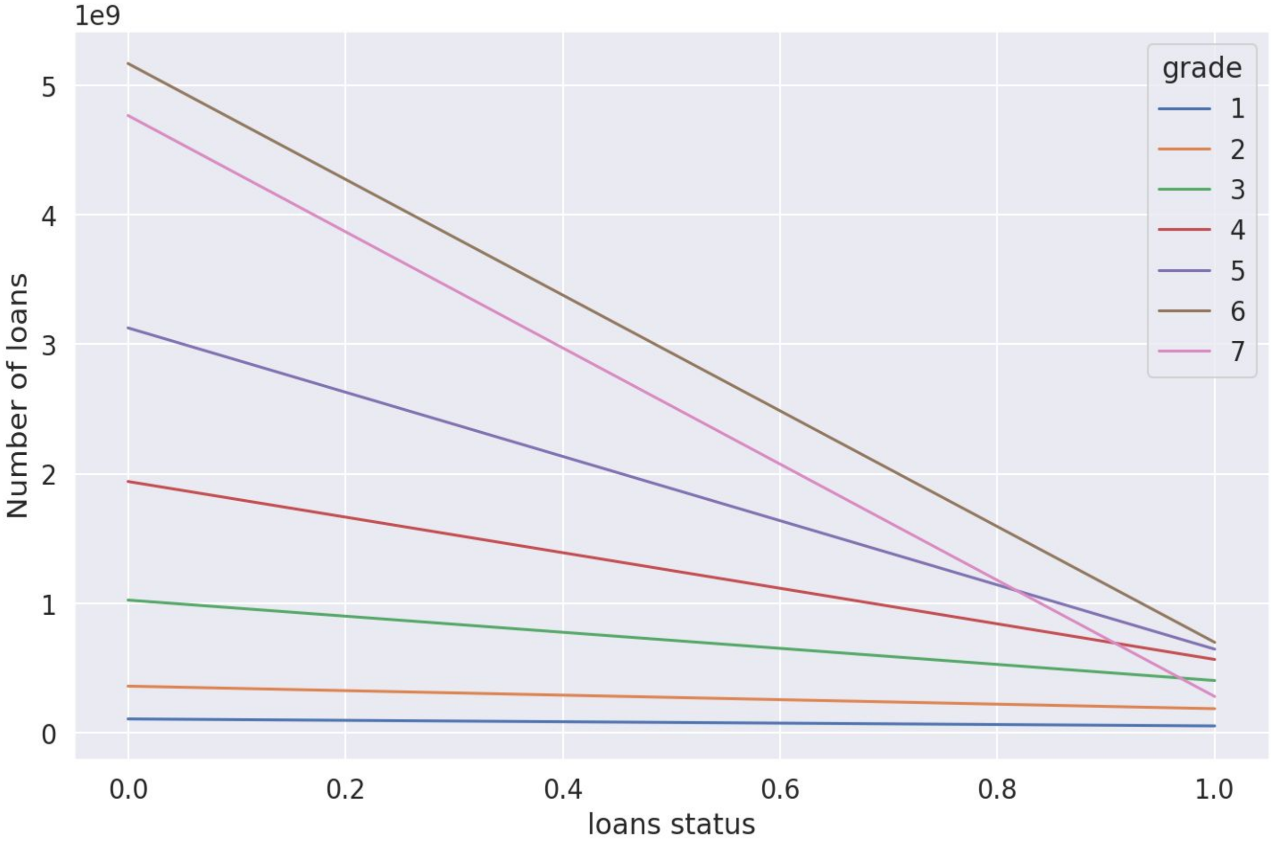
By analysing the correlation and distribution of annual income and balance annual income which is nothing but the balance of the income after the installments .



- The actual annual income increases the balance annual income decreases.
- We can conclude that when the balance decreases then the probability of defaulting also increases.
- Also if the Balance annual income increases the loan defaulting decreases.

#Now we can see the number of loans distributed over the grade and Status of the loan.

- The insight is as the grades a,b,c,d (please refer the mapping table below) have more number of loans and also have more number of loans defaulted compared to other grades this is the same in the case if sub grades

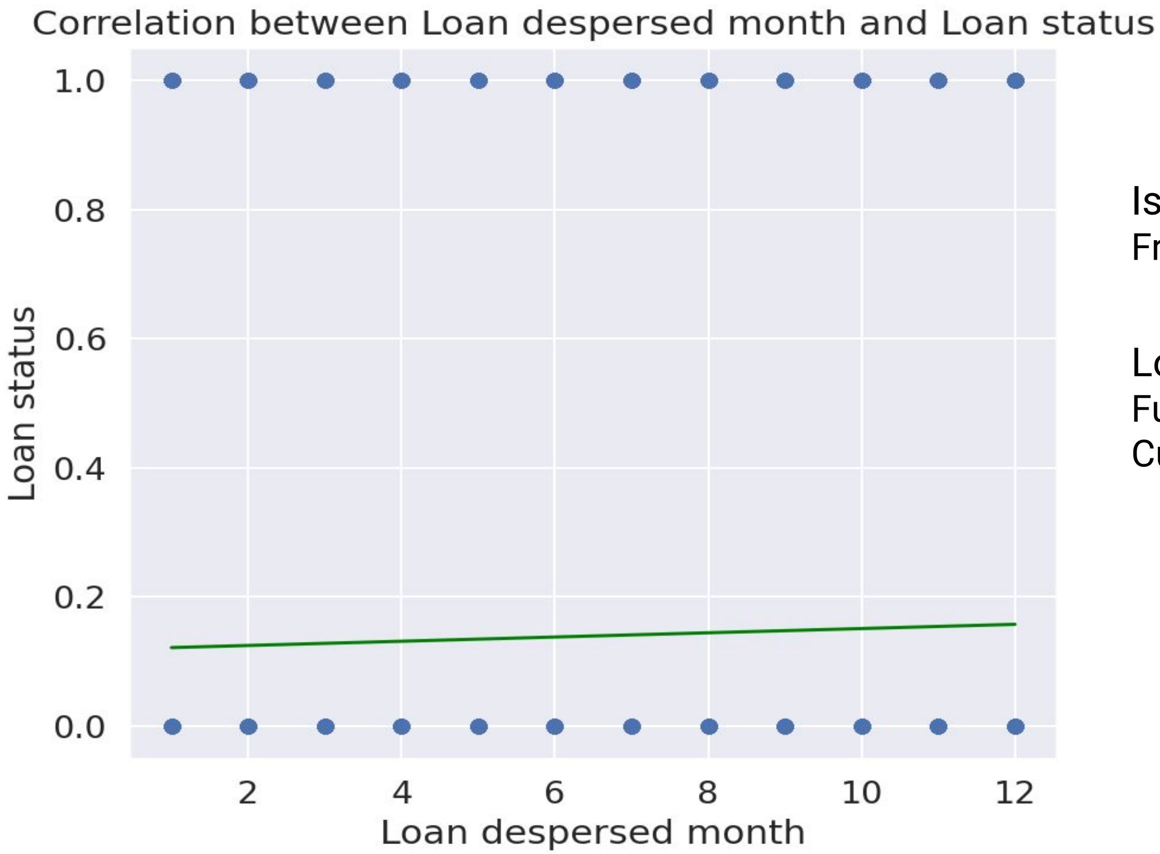


Grades	Map Values
A	7
B	6
C	5
D	4
E	3
F	2
G	1

Loan Status : -
Fully Paid = 0 , Charged Off = 1 , Current = 2

By analysing the correlation and distribution of Loan disbursed month and Loan Status.

- We can conclude that the most likely the loans sanctioned between 8 th to 12 th month got defaulted.



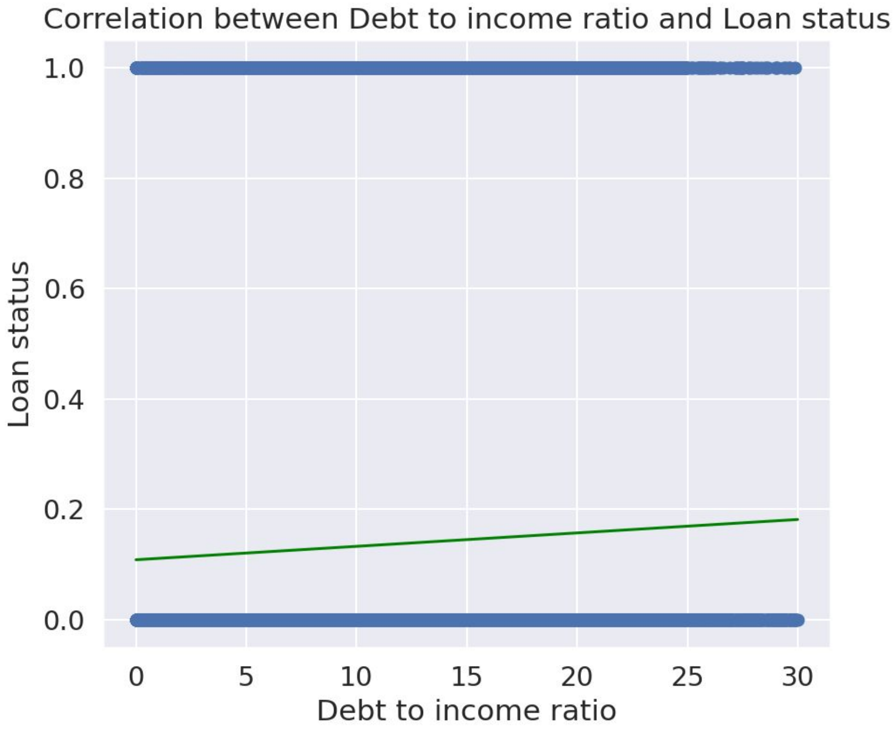
Mappings

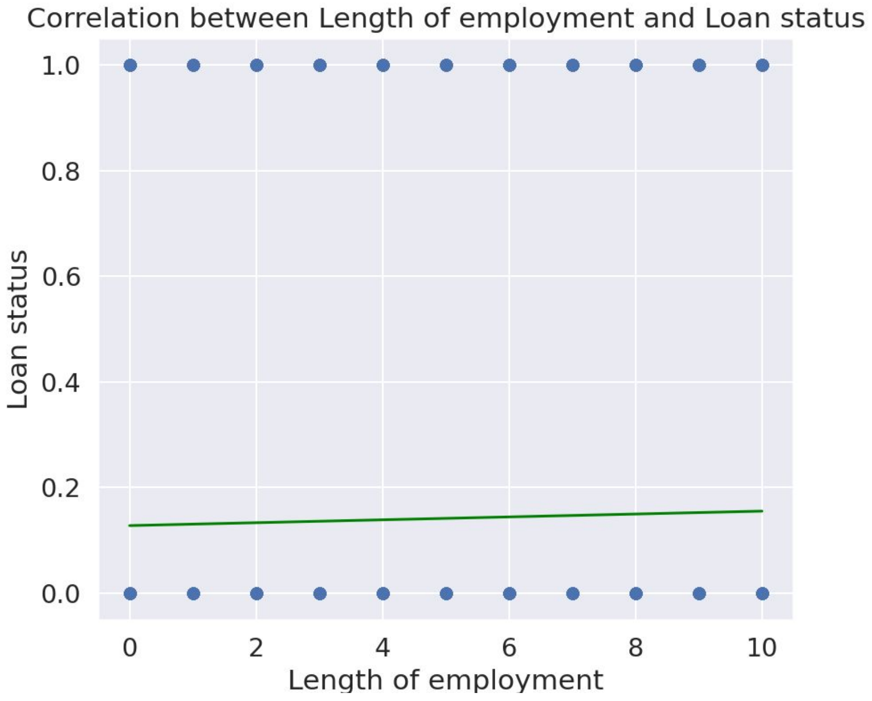
Issue Date(Loan disbursed date) : -
From Jan = 1 to Dec = 12

Loan Status : -
Fully Paid = 0 , Charged Off = 1 ,
Current = 2

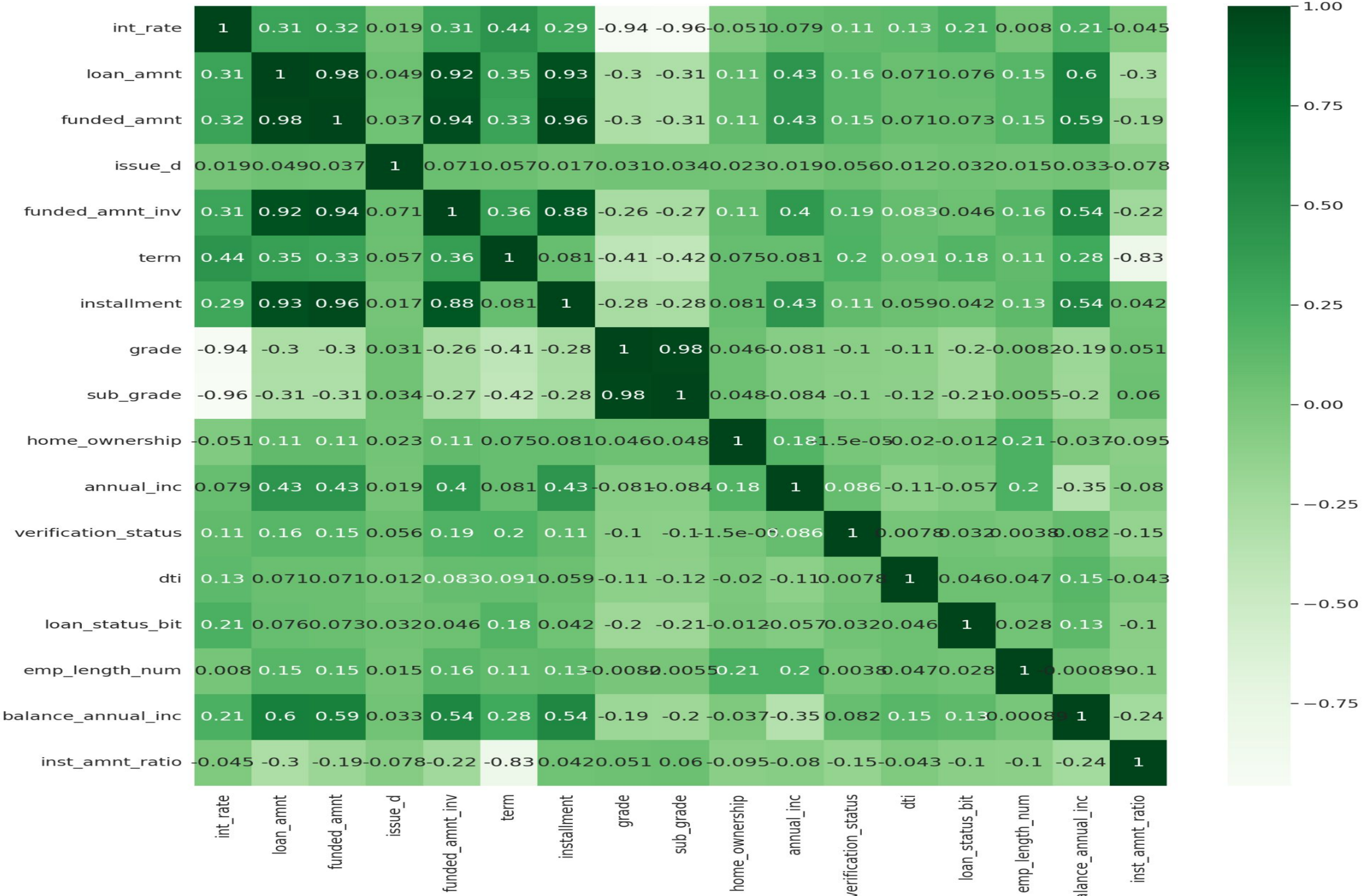
By analysing the correlation and distribution of Debt to income ratio and Loan status and Length of employment

- We can see that the debt to income ratio has a positive correlation with loan status that is if the the Debt to income ratio increases then there is is a high chance of defaulting the loan .
- We can point that as the length of employment increases the chance of loan defaulting increases.





Correlation Matrix



The Recommendations

Based on the correlation values the following variables have a positive correlation with the "loan_status_bit" feature:

- loan_amnt: 0.076499
- funded_amnt: 0.073149
- funded_amnt_inv: 0.046281
- term: 0.176951
- installment: 0.042059
- issue_d: 0.032325
- verification_status: 0.031629
- dti: 0.046339
- balance_annual_inc: 0.127828

The positive correlation coefficient suggests that as the value of these variables increases, the likelihood of loan default also increases.

On the other hand, the following variables have a negative correlation with the "loan_status_bit" feature:

- grade: -0.203700
- sub_grade: -0.206201
- annual_inc: -0.057067
- home_ownership: -0.011517
- emp_length_num: 0.027814

The negative correlation coefficient suggests that as the value of these variables decreases, the likelihood of loan default increases.

However, it is important to note that correlation does not necessarily imply causation, and other factors could be at play in determining the loan status.

Thank You