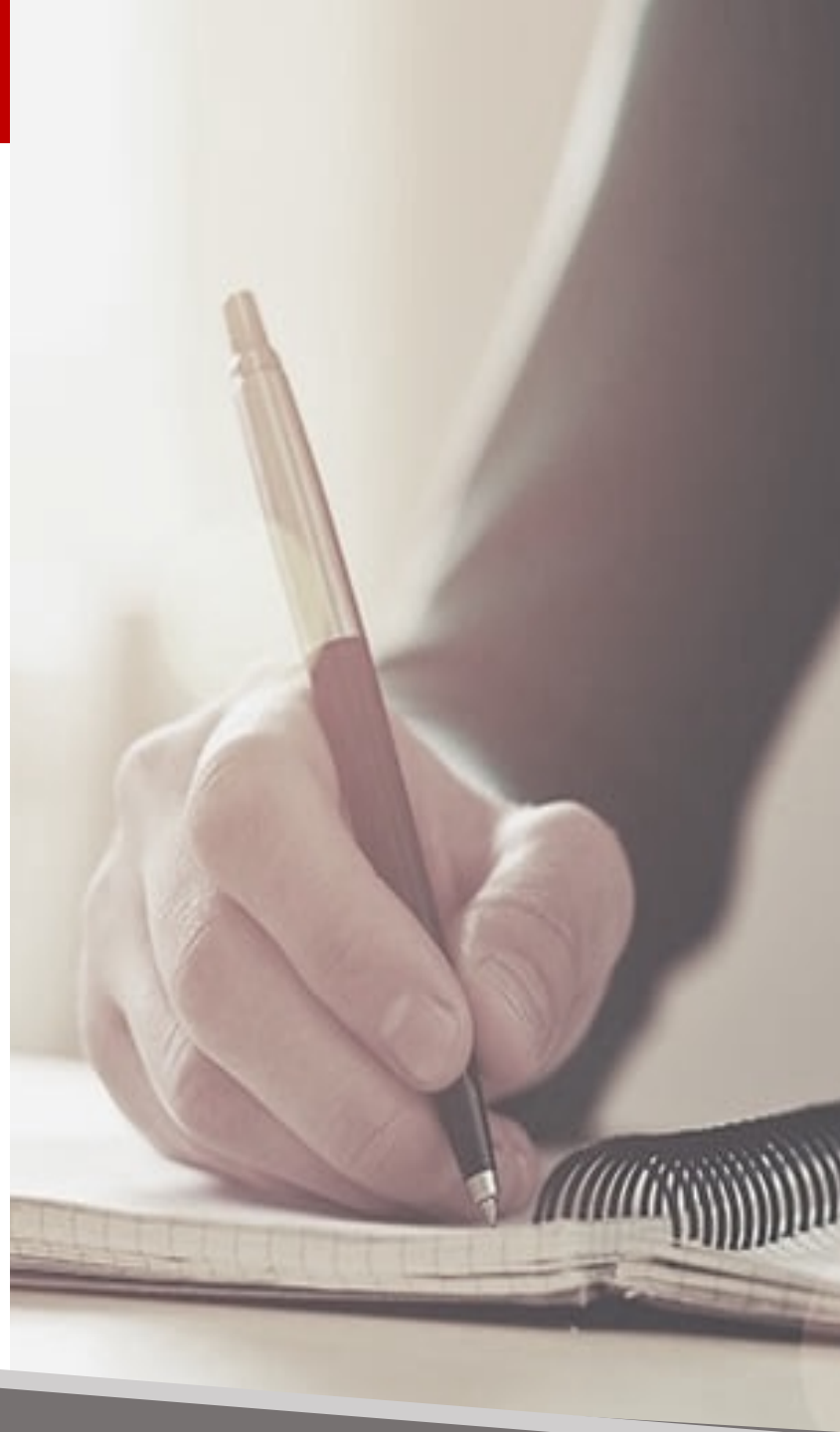
LENDING CLUB CASE STUDY

Submitted By:
Keya Bhattacharjee

TABLE OF CONTENTS

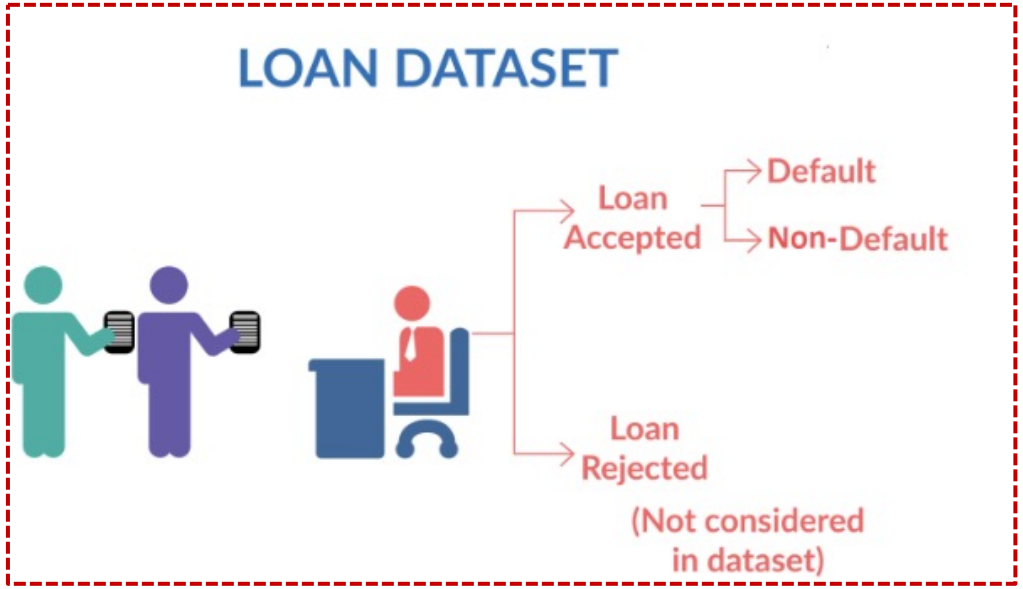


PROBLEM STATEMENT

Lending Club

A consumer finance company specialises in lending various types of loans to urban customers.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.



Two **types of risks** are associated with the bank's decision:

Risks

Business loss

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

Financial loss

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 (credit loss) for the company

If the company approves the loan, there are 3 possible scenarios

Loan Accepted

Fully Paid

Applicant has fully paid the loan.

Current

Applicant is in the process of paying the instalments

Charged off

Applicant has not paid the instalments in due time. These are the '**Defaulters**'

OBJECTIVE

Objective :

Identification of risky applicants using EDA

Lending loans to the 'risky' applicants is the largest source of financial loss (called credit loss). In other words, borrowers who **default** cause the largest amount of loss to the lenders.

Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

APPROACH

Data Understanding

Correct interpretation of the of the data and the meanings of the variables. Understanding the type of data in each column

Data Cleaning & Manipulation

Data quality issues are addressed , data is converted to a suitable and convenient format to work and manipulation of strings and dates is done correctly wherever required

Data Analysis

Analysis the data with the most appropriate method . Univariate , segmented univariate & Bivariate analysis is done

Graphs & Plots

Appropriate plots are created to present the results of the analysis and the relevant insights

Observations

Observations from the data analysis are listed and actionable recommendation s are provided

UNDERSTANDING THE DATA

- Uploading the relevant libraries
- Loading and reading the data
- The description of each column is given in the data Dictionary file
- Checking if the dataframe is uploaded correctly
- Checking the number of rows and columns – There are 39717 rows and 111 columns
- Fetching more information about the dataframe using dataframe.info()
- Checking the types of data in the columns ; there are three types of data – int, float and object
- The description of the variable are in the ‘Data Dictionary’ file

CLEANING OF THE DATA

- Once the dataframe is loaded correctly, we need to clean the data
- Checking the percentage of the missing values in each column , there are 54 columns with all null values
- Dropping the columns which have all null values. 54 columns were deleted, the total number of remaining columns now is 57
- Dropping the columns - 'member_id','desc', 'url', 'pymnt_plan', 'initial_list_status', 'policy_code', 'application_type', 'acc_now_delinq', 'delinq_amnt' ; they are not useful for the analysis
- Dropping the columns - 'chargeoff_within_12_mths', 'tax_liens', 'collections_12_mths_ex_med' as they have only 0 or NA values and are not helpful
- Replacing the NA Values with 'Unknown' in the columns 'mths_since_last_delinq' and 'mths_since_last_record'
- Remaining number of columns now is 45

ANALYSIS OF THE COLUMNS

Column : ‘emp_title’

Description : The job title supplied by the Borrower when applying for the loan.
Analysis Done : Count of each value is calculated
Observation : Most of the values are unique and does not reveal any useful information
Action Taken : Drop the column

Column : ‘emp_length’

Description : 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.
Analysis Done : Count of each value is calculated
Segmented Univariate Analysis done and a stacked bar graph is plotted for Employment Length Vs Loan Status
Observation : The fraction of the three categories of loan status across all the employment length groups are similar and does not reveal any conclusion
Action Taken : We will retain the column

Column : ‘home_ownership’

Description : The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
Analysis Done : Count of each value is calculated, replacing the ‘NONE’ values with ‘OTHER’
Segmented Univariate Analysis done and a stacked bar graph is plotted for Home Ownership Vs Loan Status
Observation : The fraction of the three categories of loan status across all the four types of home ownership are similar and does not reveal any conclusion
Action Taken : We will retain the column

ANALYSIS OF THE COLUMNS

Column : 'verification_status'

Description : Indicates if income was verified by LC, not verified, or if the income source was verified
Analysis Done : Count of each value is calculated,
Segmented Univariate Analysis done and a stacked bar graph is plotted for Verification Status Vs Loan Status
Observation : The fraction of the three categories of loan status across all verification status types are similar and does not reveal any conclusion
Action Taken : Retain the column

Column : 'purpose'

Description : A category provided by the borrower for the loan request.
Analysis Done : Segmented Univariate Analysis done and graphs are plotted for Purpose Vs Loan Status
Observation : The highest percentage of defaulters are those who have borrowed the money for small business.
Most of the loans were taken for the purpose of debt consolidation & paying credit card bill.
Number of charged off count is also high for these loans purposes
Action Taken : Retain the column

Column : 'title'

Description : This is the loan title provided by the borrower
Analysis Done : Count of each value is calculated
Observation : This column does not reveal any useful information
Action Taken : Drop the column

ANALYSIS OF THE COLUMNS

Column : ‘mths_since_last_delinq' and 'mths_since_last_record'

Description : Indicates the number of months since the borrower's last delinquency and the number of months since the last public record.
Analysis Done : The NA values are replaced with ‘Unknown’ . Count of each value is calculated,
Observation : Most of the values are Unknown, so we cannot derive any reliable conclusion from these columns.
Action Taken : Drop the columns

Column : ‘last_pymnt_d’

Description : Indicates the last month payment was received
Analysis Done : The year is extracted from the date, and a new column ‘Year of last payment’ is created, the NA fields are replaced by ‘Not recorded’
Stacked bar graph of Loan Status Vs Last Payment Year is plotted
Observation : From the graph we see that all the unpaid loans prior to 2016 have been charged off
There is almost a decreasing trend (except 2009) in the percentage of defaulters across the years
Action Taken : Retain the column

Column : ‘next_pymnt_d’

Description : Indicates the next scheduled payment date
Analysis Done : The missing values are replaced with ‘Unknown’ . Count of each value is calculated,
Observation : Most of the values are Unknown, so we cannot derive any reliable conclusion from these column.
Action Taken : Retain the column for now

ANALYSIS OF THE COLUMNS

Column : 'last_credit_pull_d'

Description : Indicates the most recent month LC pulled credit for this loan
Analysis Done : The year is extracted from the date, and a new column 'Year of last credit pull' is created, the NA fields are replaced by 'Not recorded'
Segmented Univariate Analysis of Loan Status Vs Year of Last Credit pull is done and graphs are plotted.
Observation : From the first graph we see that the percentage of the charged off count is highest for the credit pull last done in 2009.
The last credit pull for maximum number of loan applications was done in 2016.
The count of the charged off loan applicants is also showing a somewhat increasing trend.
Action Taken : Retain the column

Column : 'pub_rec_bankruptcies'

Description : Indicates the number of public record bankruptcies
Analysis Done : Replacing the NA values with 'Unknown' . Calculating the count of each value
Stacked bar graph of Loan Status Vs Bankruptcies is plotted.
Observation : Higher the value of recorded bankruptcy, higher is the percentage of the defaulters.
The number of recorded bankruptcy is only a very small percentage of the total loan applicants.
Action Taken : Retain the column

Column : 'int_rate'

Analysis Done : Stacked bar graph of Loan Status Vs Interest Rate is plotted.
Observation : Higher the interest rate, higher is the percentage of the defaulters.
Action Taken : Retain the column

ANALYSIS OF THE COLUMNS

Column : 'grade' and 'Sub_grade'

Description : These columns represent the LC assigned grade and sub grades.
Analysis Done : Segmented Univariate Analysis of Loan Status Vs Loan Grade is done and graphs are plotted.
Observation : The count of loan application accepted is highest for Grade B, followed by Grade A and Grade C
The percentage of Charged off count is lowest for Grade A, followed by B, C, D, E , F and G in order
The absolute count of charged off applicants is high for B, c and D as is visible from the second graph
We can ignore the sub Grade column as the grade column is sufficient for the analysis
Action Taken : Retain the grade column, ignore the sub-grade column

Column : 'addr_state' and 'zip_code'

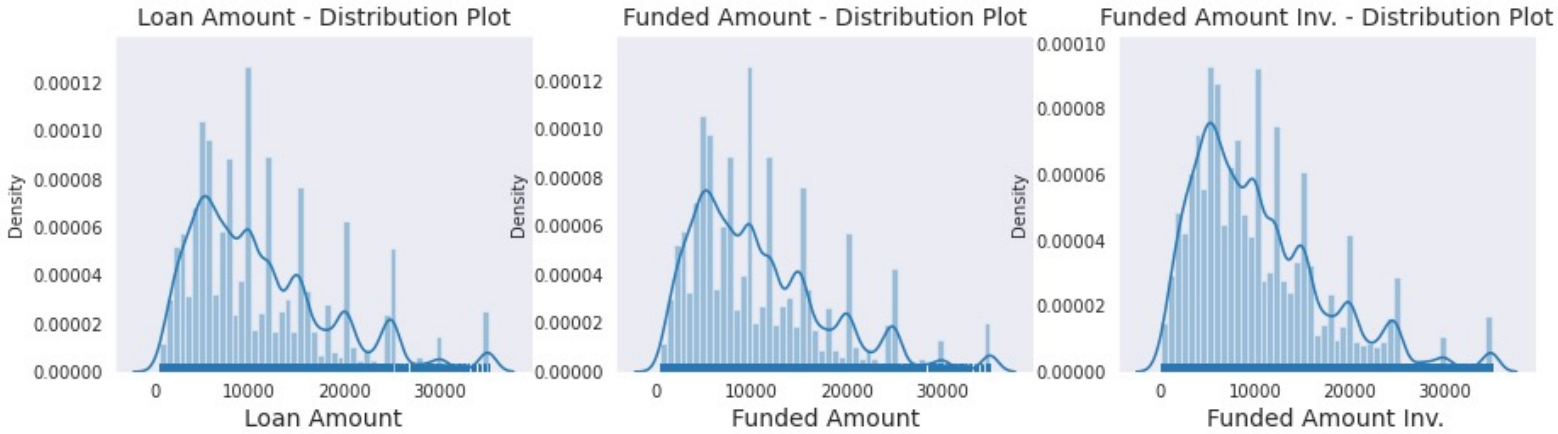
Description : These columns represent the address of the borrower
Analysis Done : Segmented Univariate Analysis of Loan Status Vs Address State is done and graphs are plotted.
Observation : The percentage of charged off count for NE is significantly high as w.r.t to it loan accepted count.
The loan accepted count is highest for CA(California) followed by NY(New York).
The charged off count is also highest for CA, followed by FL and NY. This is because of the high count of loan accepted counts
We can ignore the sub Grade column as the grade column is sufficient for the analysis
Action Taken : Retain the columns

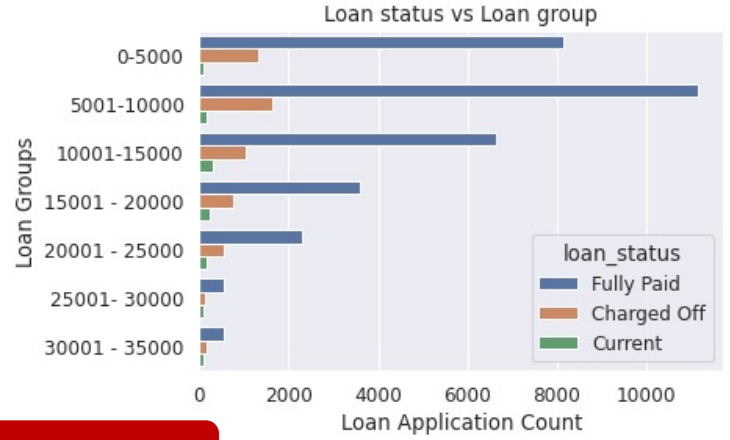
After we have dropped the unnecessary columns and filled up the missing values in the other columns, let us know check the status of the remaining columns.

ANALYSIS OF THE COLUMNS

Column : 'loan_amnt','funded_amnt' and 'funded_amnt_inv'

Description : There are three columns which indicates the loan amount applied for and funded
Analysis Done : Distribution graphs of each of these variable is plotted against the loan status
Observation : Distribution of amounts for all three looks very much similar.
It means that loan amount applied by borrower was more or less granted the same by the Lending Club
Action Taken : We will work with only loan amount column 'loan_amnt' for rest of our analysis.





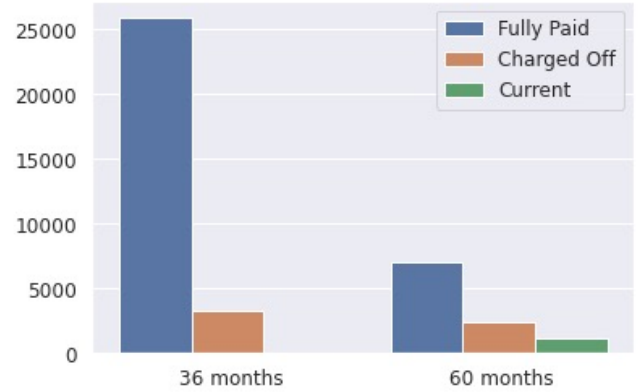
Column : 'loan_amnt'

Analysis Done : Box plot is plotted for the column 'Loan_amnt'
The loan amount is grouped into multiples of 5000 and a separate column is created
Bar graph is plotted for Loan amount Group Vs Loan Status for each loan amount group
Observation : We see that the maximum of the loan amount is between 5001 - 10000, followed by 5000 and 10000 -15000,
The box plot shows that 75% of the loan amount is below 15000
The count of charged off is highest for 5001 - 10000 followed by 5000 and 10000-15000.

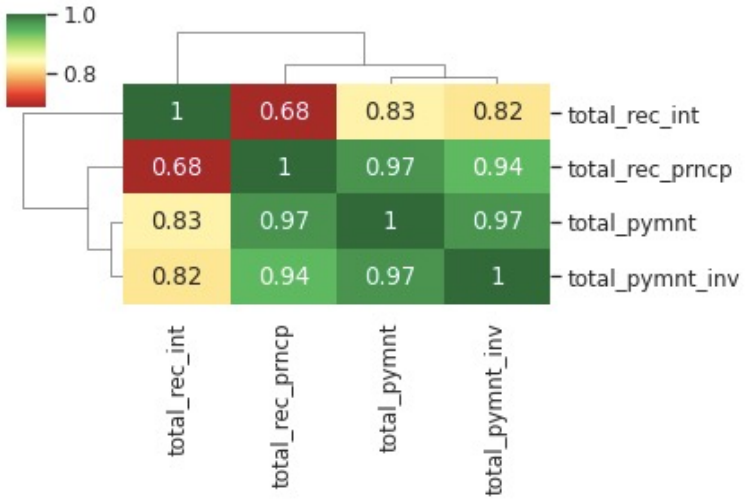
ANALYSIS OF THE COLUMNS

Column : 'term'

Description : This column shows the number of payments on the loan. Values are either 36 or 60 months
Analysis Done : Calculating the value of each term
Univariate Analysis - Loan Paying Term Vs Loan Status, bar graph plotted
Observation : The plot shows that those who had taken loan to repay in 60 months had more % of number of applicants getting charged off as compared to applicants who had taken loan for 36 months.



Column : 'total_pymnt' , 'total_pymnt_inv','total_rec_prncp' and 'total_rec_int'



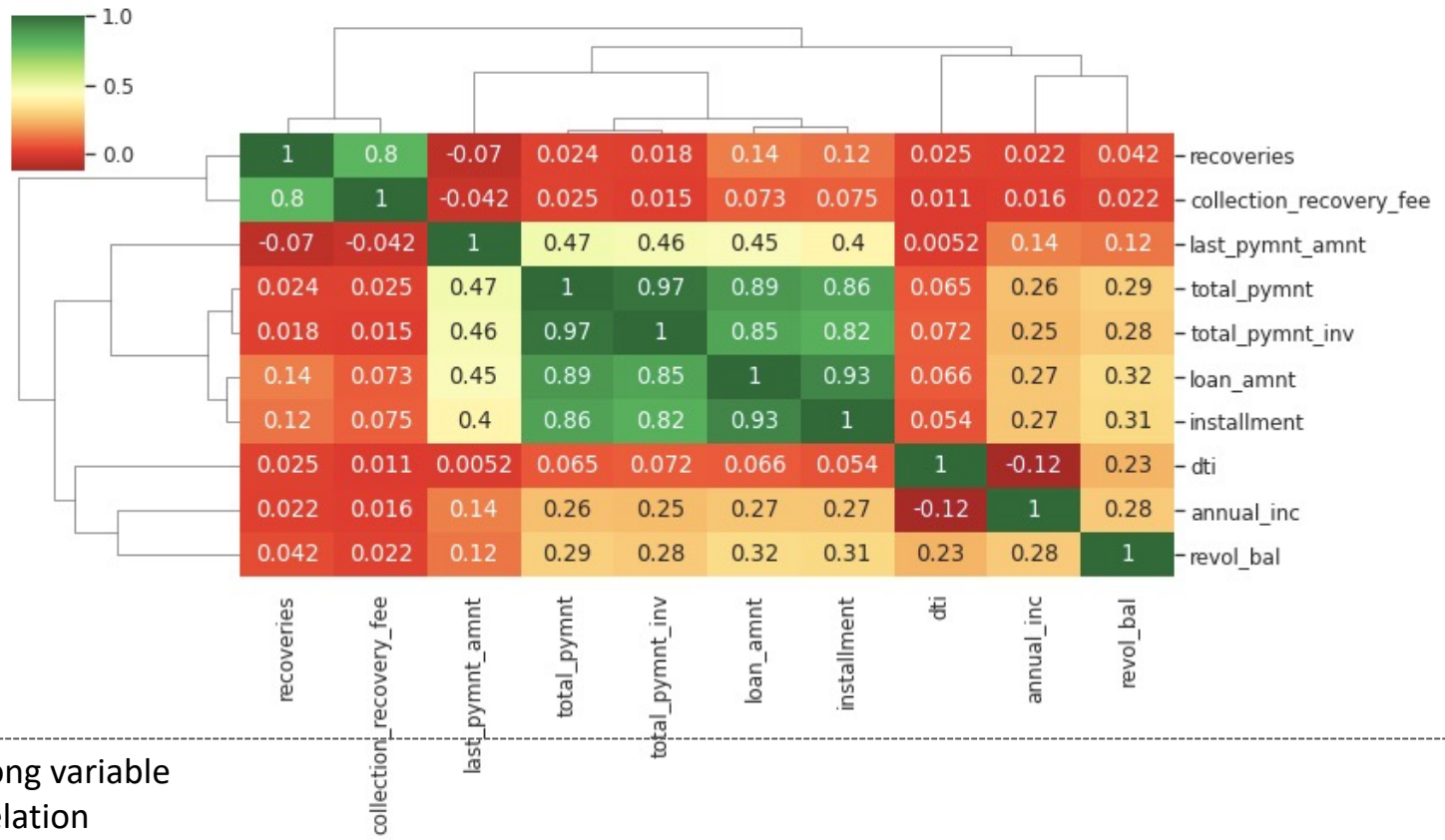
Description : These columns indicates the total payment received as part of the total funded amount, the invested amount. The principal and the interest received
Analysis Done : Calculating the correlation among the four columns
Observation : There is a strong correlation between 'total_pymnt','total_pymnt_inv', 'total_rec_prncp'. There is a positive correlation between all the four columns, although the correlation of 'total_rec_int' with the other three columns is slightly lesser.
Action: We can , therefore, use the column 'total_payment' for further analysis to represent these four columns

ANALYSIS OF THE COLUMNS

Columns

1.'loan_amnt’
2.'installment’
3.'annual_inc’
4.'dti’
5.'total_pymnt’
6.'total_pymnt_inv’
7.'last_pymnt_amnt’
8.'recoveries’
9.'collection_recovery_fee’
10.'revol_bal'

These are numeric variables



Analysis Done : Calculating the correlation matrix among variable

Observation : Most of the variables have positive correlation

Annual income with DTI(Debt-to-income ratio) is negatively correlated. That means when annual income is low DTI is high & vice versa.

There is a strong positive correlation between 'total_pymnt’, 'total_pymnt_inv’, ‘loan_amnt’ and ‘installment’. This means, higher the loan amount, higher is the payment towards it

Variables 'recoveries’ and 'collection_recovery_fee’ also shows a strong positive correlation, which means more spending on trying to do the recovery is resulting in more recovery

Variable 'recoveries’ and 'collection_recovery_fee’ has negative correlation with ‘last_pymnt_amnt’, which indicates if the last payment is high, then less amount is being spent on recovering the loan

Action: We will use the column 'total_pymnt’ for further analysis

CONCLUSION

Variable	Summary
addr_state	The charged off count is also highest for CA, followed by FL and NY. This is beacuse of the high count of loan accepted counts
earliest_cr_line	The count of charged off loans are also highest for these year 2000, followed by 1999 and 1998
grade	The percentage of Charged off count is lowest for Grade A, followed by B, C, D, E , F and G in respective order
int_rate	Higher the interest rate leads to higher charged off percentage
issue_d	The number of charged off is also increasing across the years consequently
last_credit_pull_d	Charged off count is highest for the credit pull last done in 2009.
last_pymnt_d	There is almost a decreasing trend (except 2009) in the percentage of defaulters across the years
loan_amnt	Higher the loan amount, higher is the percentage of charged off candidates
open_acc	The count of charged off loans are also highest for candidates with open credit 7,6,8,and 9
pub_rec_bankruptcies	Higher the value of recorded bankruptcy, higher is the percentage of the defaulters.
purpose	Highest percentage of defaulters are those who have borrowed the money for small business.
term	The percentage of charged off candidates are higher for 60 month term than 36 month

Keya Bhattacharjee

+91 9836304633

kbhjee@gmail.com