


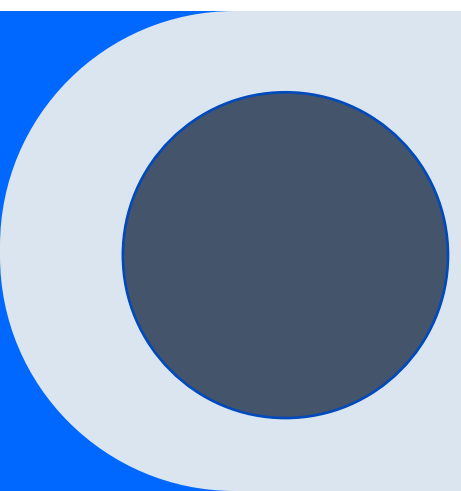


Lending Club Case Study

Group Facilitator – Parameshwari.R

Group Member – Pamela Roy

Batch – MLC51 EPGP ML&AI



Agenda

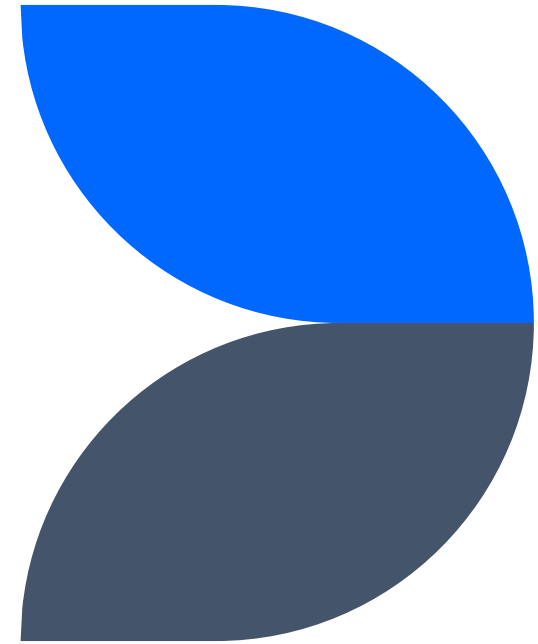
- Problem Statement
- Data Understanding
- Data Cleaning & Imputing
- Data Analysis
- Recommendations
- Meet the team

Problem Statement

We work for a **consumer finance company** which specializes 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. The data provided as a part of this case study contains information about past loan applicants and whether they 'defaulted' or not. 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 Understanding

Data files provided – loan.csv , Data_Dictionary.csv

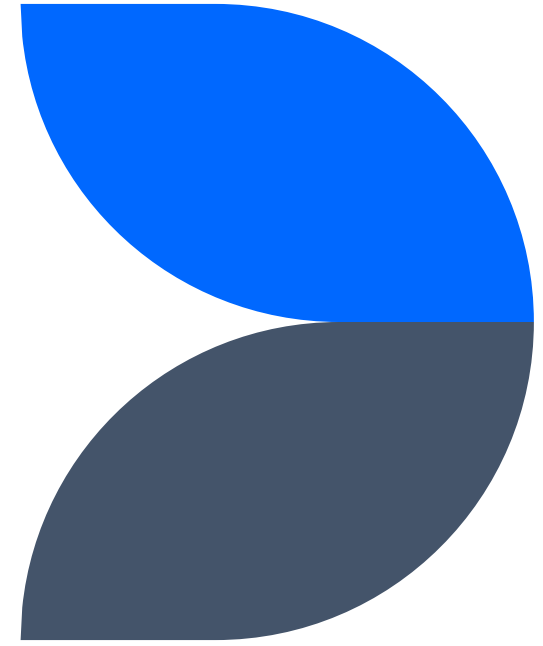


Data Understanding

On analysing loan.csv file below are the observations :

- Number of rows : 39717
- Number of columns : 111
- Data type of columns : float64(74), int64(13), object(24)
- Number of Categorical variables : 24
- Number of Numeric variables : 87
- Number of empty rows : 0
- Number of empty columns = 54 (Needs to be dropped)
- Number of Duplicate rows = 0
- Unique identifiers : id , member_id (Not needed for generic analysis , can be dropped)
- Customer behavior variables (not available during loan applications – can be dropped) : delinq_2yrs , earliest_cr_line , inq_last_6mths , open_acc , pub_rec , revol_bal , revol_util , total_acc , out_prncp , out_prncp_inv , total_pymnt , total_pymnt_inv , total_rec_prncp , total_rec_int , total_rec_late_fee_recoveries , collection_recovery_fee , last_pymnt_d , last_pymnt_amnt , last_credit_pull_d , application_type

DATA Cleaning & Imputing



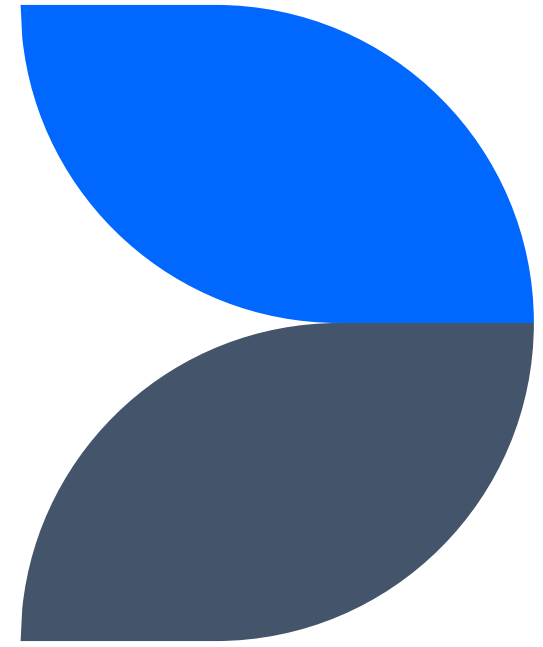
Data Cleaning (File : loan.csv)

- **Identified Columns with NULL Values for 80 to 100% of the entries**
 - These columns were dropped
 - Left with 57 Columns
- **Customer Behavior variables which were not available at the time of loan application were removed**
 - 21 columns were dropped (delinq_2yrs , earliest_cr_line , inq_last_6mths , open_acc , pub_rec , revol_bal , revol_util , total_acc , out_prncp , out_prncp_inv , total_pymnt , total_pymnt_inv , total_rec_prncp , total_rec_int , total_rec_late_fee , recoveries , collection_recovery_fee , last_pymnt_d , last_pymnt_amnt , last_credit_pull_d , application_type
 - Left with 36 columns
- **Other columns which do not add much value to the analysis were removed**
 - 24 columns were dropped (id , member_ID , pymnt_plan , funded_amnt_inv, installment .. Etc)
 - Left with 13 Columns to analyze
 - Columns available after deleting the columns : loan_amnt,term,int_rate, grade, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, purpose, dti, pub_rec_bankruptcies
- **Checked for row entries with missing values**
 - 1075 entries for emp_length were missing
 - 697 pub_rec_bankruptcies were missing

Data Imputing (File : loan.csv)

- **Missing Values in emp_length – 1075 entries**
 - updated with MODE of the column i.e : Most frequently occurring value .
In this case the missing fields were updated with 10+Years
- **Missing Values in pub_rec_bankruptcies– 697 entries**
 - updated with MODE of the column i.e : Most frequently occurring value
- **Entries having ‘NONE’ value in home_ownership field**
 - Replaced by ‘OTHER’
- **Removed % from int_rate column and convert the data type to numeric for analysis**
- **Checked loan_status fiels**
 - 82.98 % have status ‘Fully Paid’ (Applicant has fully paid off)
 - 14.15% have status ‘**Charges off**’ (Applicant has been charged off)
We are interested in this category .
 - 2.87% have status ‘Current’ (Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates will not be labelled as 'defaulted'.
Hence , these rows were dropped .

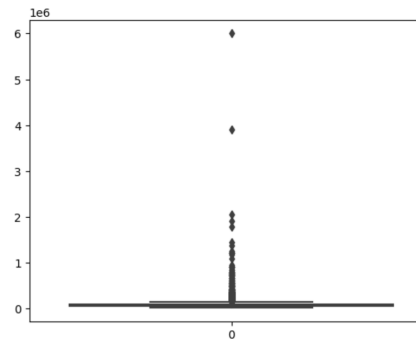
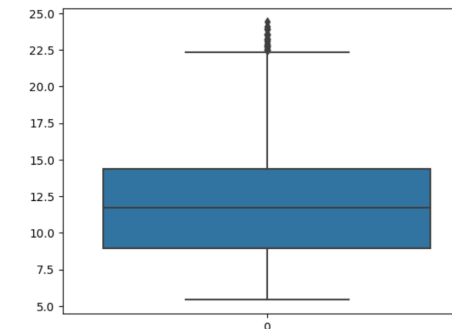
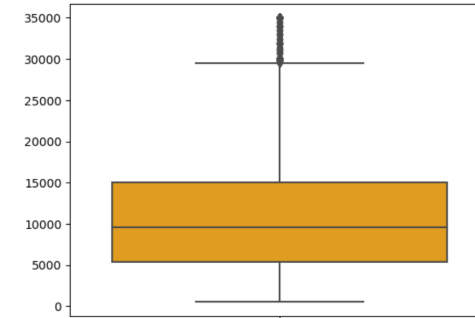
Univariate Analysis



Removal of Outliers (File : loan.csv)

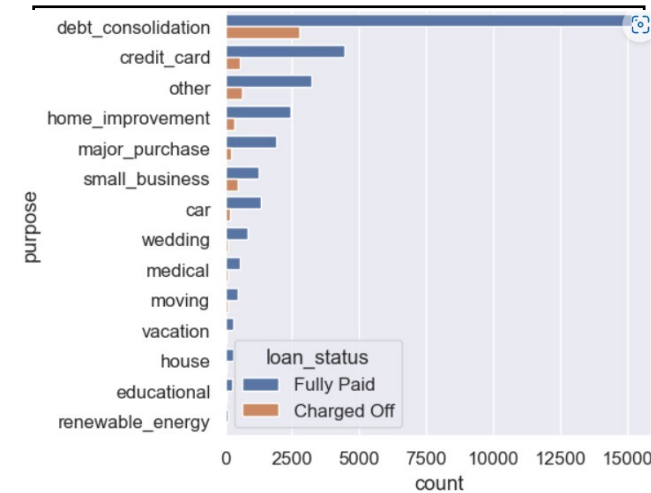
- **Outliers detected using box plots for `loan_amount`**
 - The distribution of data is continuous hence , outlier data not deleted
- **Outliers detected using box plots for `int_rate`**
 - The distribution of data is continuous hence , outlier data not deleted
- **Outliers detected using box plot for `annual_inc` field**
 - Rows within 99 to 100 percentile was deleted (Below pic shows the percentile values indicating huge difference in values)

0.00	4000.0
0.90	115000.0
0.95	140004.0
0.96	150000.0
0.97	165000.0
0.98	187000.0
0.99	234000.0
1.00	6000000.0

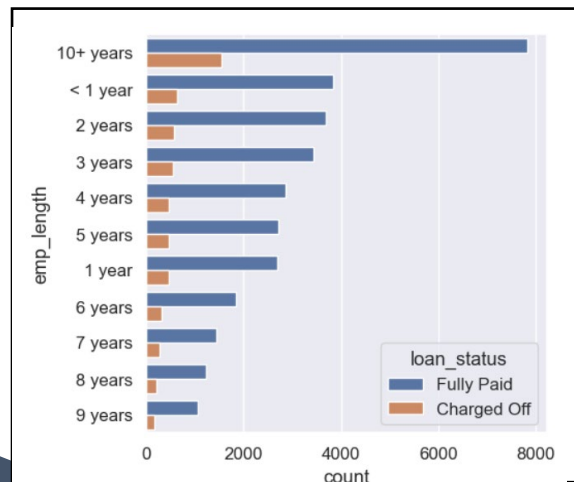


Univariate Analysis(File : loan.csv)

- Analysing the **Purpose of LOAN** for applicants with loan status = Charge off
 - Applicants who has status 'charge_off' has mostly applied for loan for **DEBT consolidations** .



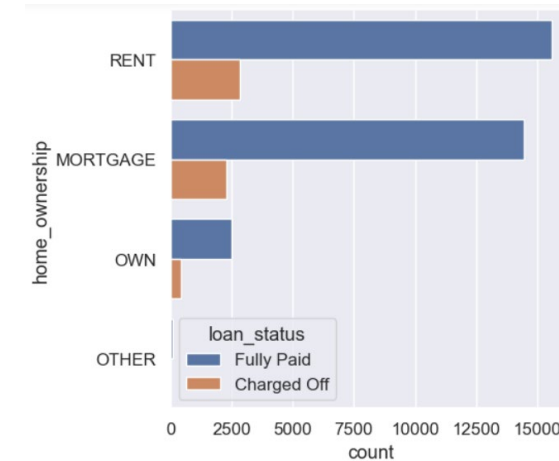
- Analysing the **Employment length** for applicants with loan status = Charge off



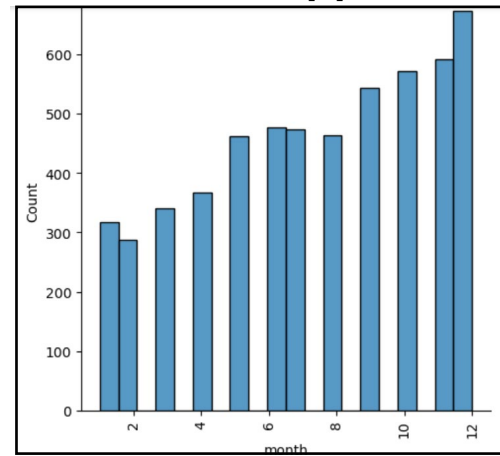
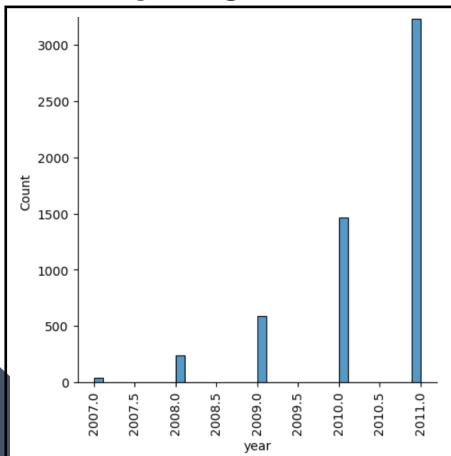
- The graph below indicates that the applicants who has **employment more than 10 years** are in the Status = Charge off category .

Univariate Analysis(File : loan.csv)

- Analysing the **Home Ownership** field for applicants with loan status = Charge off
 - Applicants who are staying on RENT or have MORTGAGED their home has status 'charge_off' .



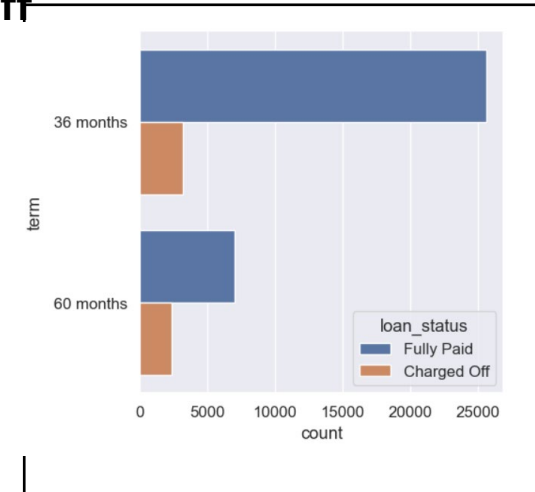
- Analysing the **Loan issue date** for applicants with loan status = Charge off



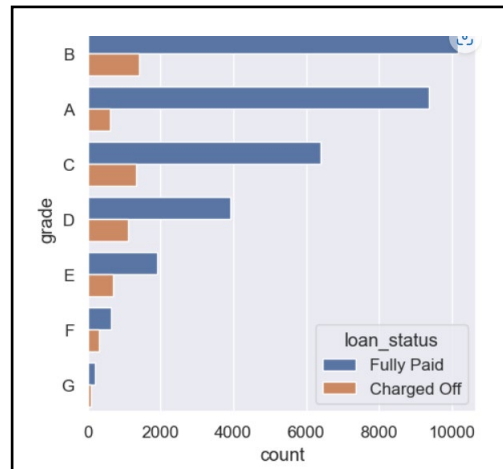
-Analysing the loan issue date data we find that the Applicants who has status 'charge_off' has mostly applied for loan in **2011 in the month of December** .

Univariate Analysis(File : loan.csv)

- Analysing the **Payment Terms** for applicants with loan status = Charge off
 - Analysing the Payment terms we don't find much difference in data distribution for the applicants who has status as 'charge_off.



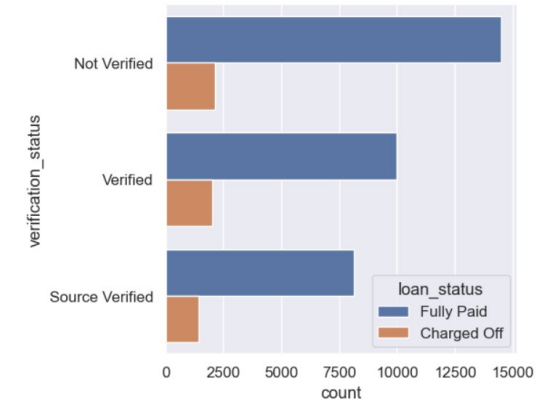
- Analysing the **GRADE** field for applicants with loan status = Charge off



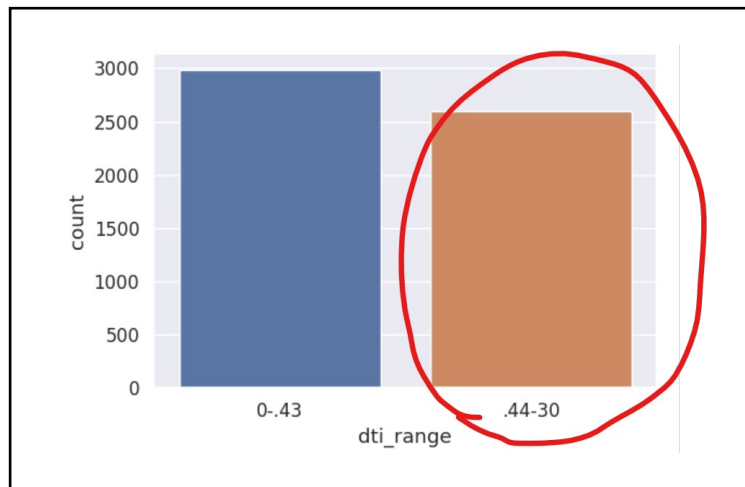
- The graph on the left indicates that the applicants in **GRADE – B , C , D** are in the group who has Status = Charge off category .

Univariate Analysis(File : loan.csv)

- Analysing the **Verification Status** for applicants with loan status = Charge off
 - There is a big set of Applicants who are in NOT VERIFIED category having status 'charge_off' .



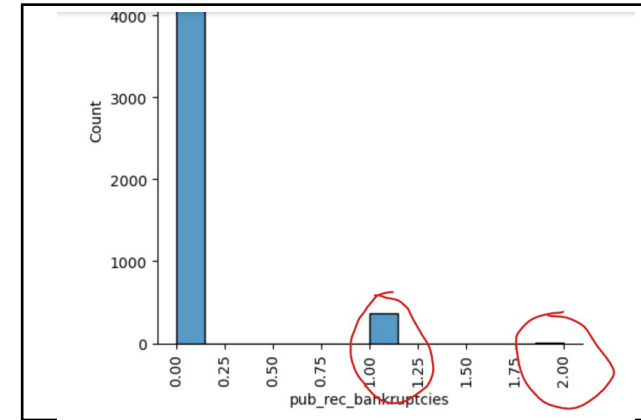
- Analysing the **DTI (Debt to Income Ratio)** for applicants with loan status = Charge off



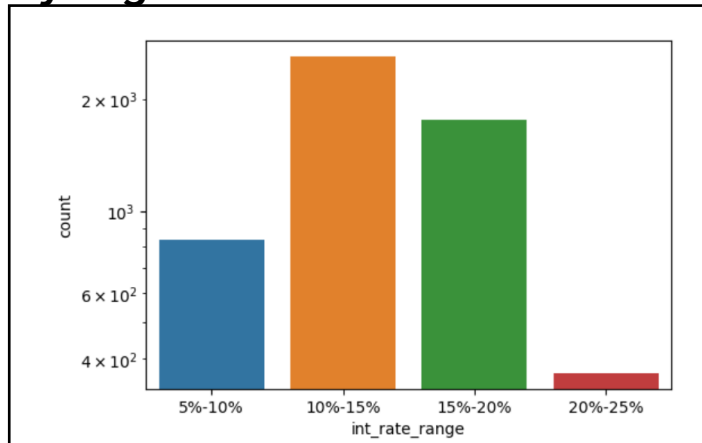
- On doing research we got to know that banks normally reject loans for applicants who has a DTI ration **greater than .43** . We have binned the DTI values to 0 to .43 and remaining in another category for all applicants whose loan status is charged_off . Highlighted in RED in the graph in left are the applicants who we are interested in .

Univariate Analysis(File : loan.csv)

- Analysing the **Public Record Bankruptcies** for applicants with loan status = Charge off
 - Analysing the pub_rec_bankruptcies for the applicants who has status as 'charge_off', we are interested in the small set of **applicants who has Bankruptcies recorded**.



- Analysing the **Interest Rate** field for applicants with loan status = Charge off

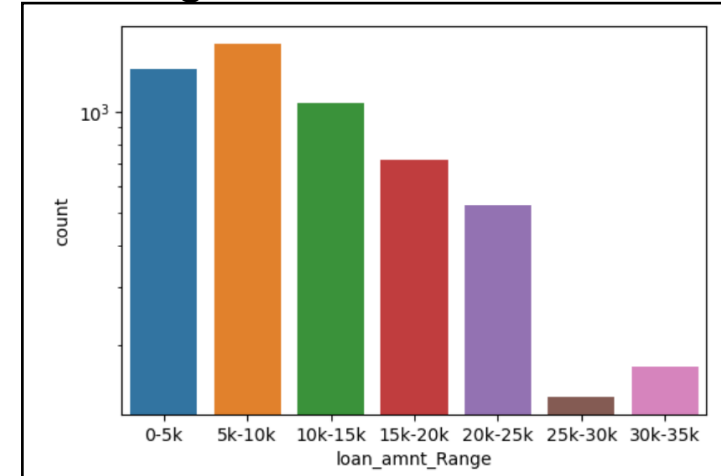


- After analysing the Interest rate field we find that most of the applicants with status = charge off have taken loan at **10 to 15% interest** or **15 to 10% interest**.

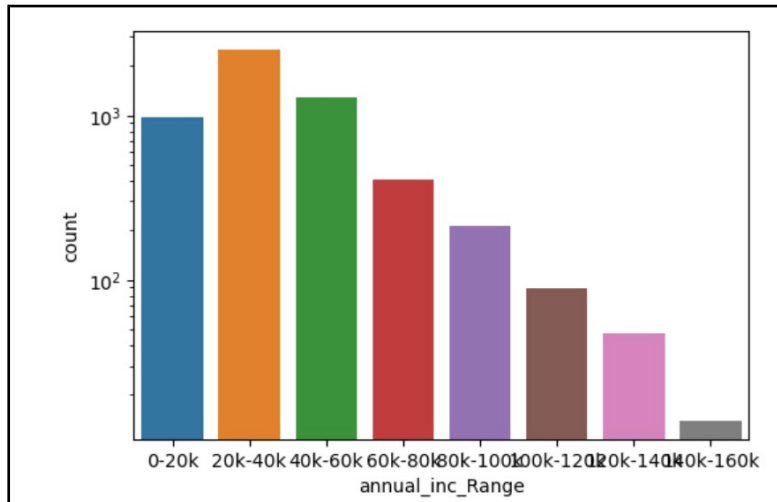
Univariate Analysis(File : loan.csv)

- Analysing the **Loan Amount** for applicants with loan status = Charge off

- Analysing the LOAN Amount field we find applicants who has status as 'charge_off' has mostly applied loans in **lower ranges** within 15 thousand .

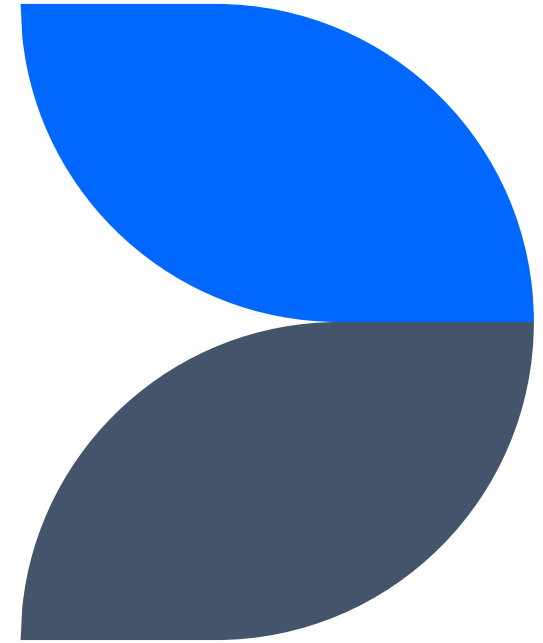


- Analysing the **Annual Income** field for applicants with loan status = Charge off



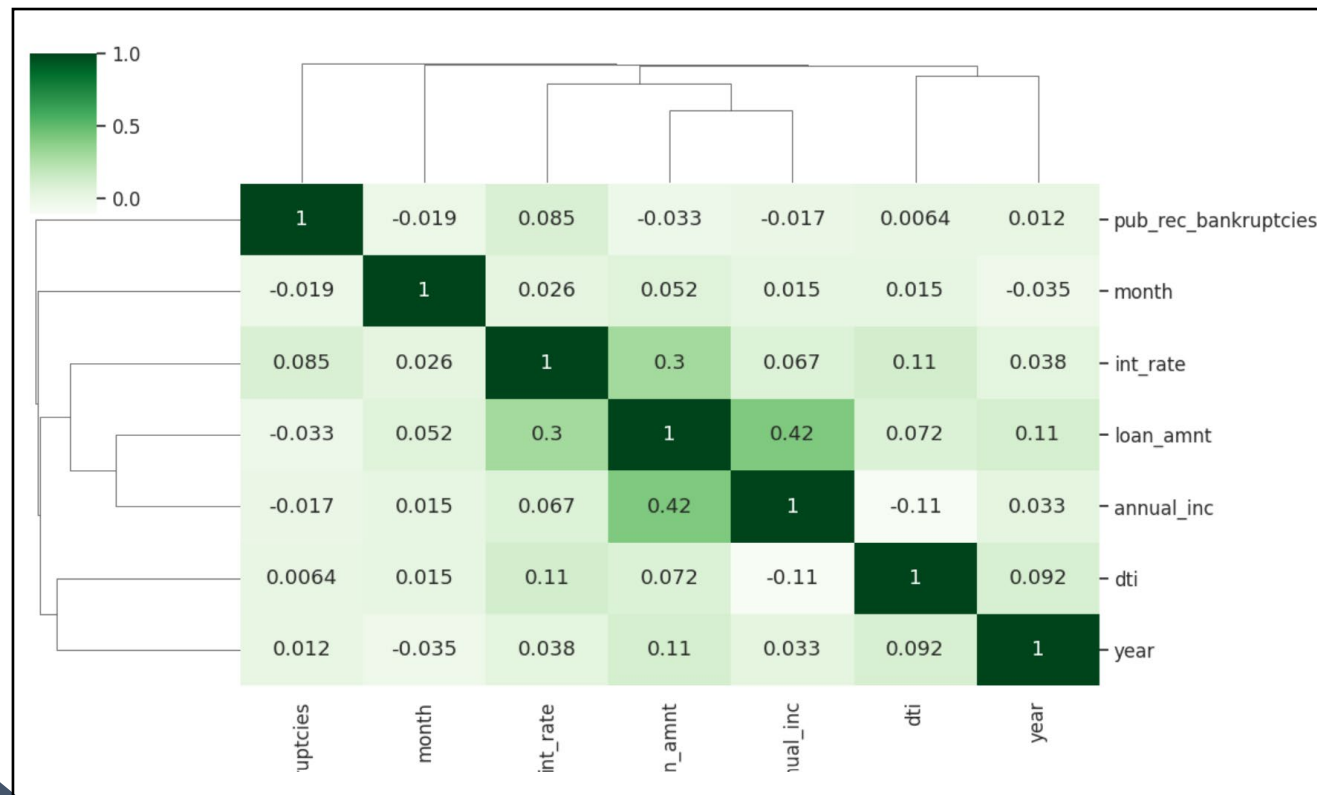
- Graph in the left indicates that the applicants who got loan status as charge off are mostly in **lower income group** .

Bivariate Analysis



Bivariate Analysis(File : loan.csv)

- **Finding co relations within all the numeric variables**
 - Variables chosen for this analysis : 'loan_amnt', 'int_rate', 'annual_inc', 'purpose', 'dti', 'pub_rec_bankruptcies', 'year', 'month', 'dti_range', 'int_rate_range', 'loan_amnt_Range', 'annual_inc_Range'

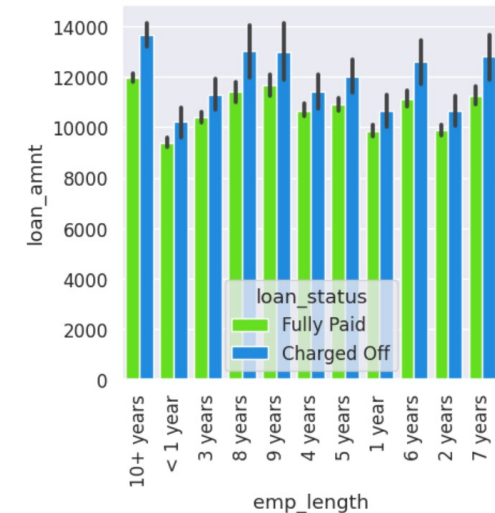


Observations :

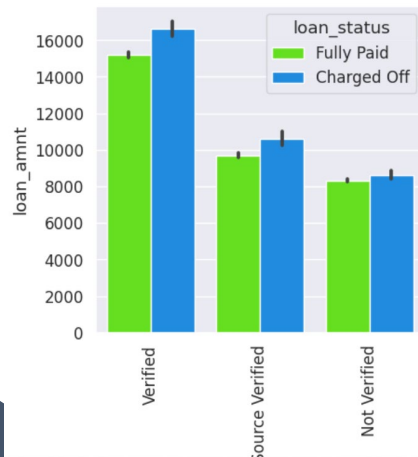
1. We see +ve correlation between Loan amount & annual income .
2. We observe +ve correlation between loan_amount and interest rate
3. +ve correlation exists between pub_rec_bankruptcies & interest rate
4. Some amount of correlation exists between DTI and interest rate

Bivariate Analysis(File : loan.csv)

- **Analysis between Employment length & loan amount for applicants with loan status as Charged Off**
 - The graph in the right shows applicants who has more than 10 years of employment has applied for higher amount of loan .



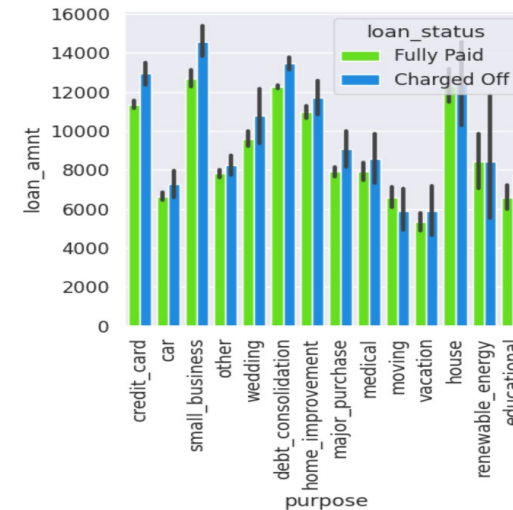
- **Analysis between loan amount & verification status for applicants with loan status as Charged Off**



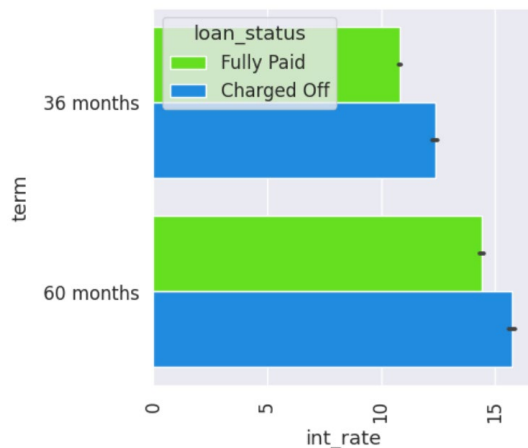
- The graph in the left indicates that applicant who had were verified got higher loans . But there is a good amount of applicants who were not verified but still got loan . The verified customers who were charged off needs to be analyzed further

Bivariate Analysis(File : loan.csv)

- **Analysis between Purpose & loan amount for applicants with loan status as Charged Off**
 - The graph in the right shows more or less even distribution for small business , credit cards , debt consolidation .



- **Analysis between Interest Rate & Terms for applicants with loan status as Charged Off**



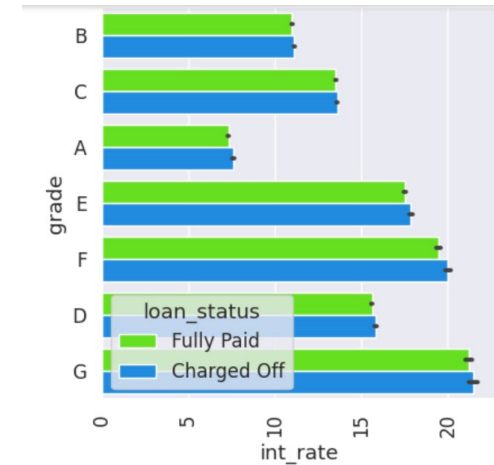
- The graph in the left indicates that the Interest rate is higher if the terms of loan is higher .
This is very obvious - nothing conclusive coming out

Bivariate Analysis(File : loan.csv)

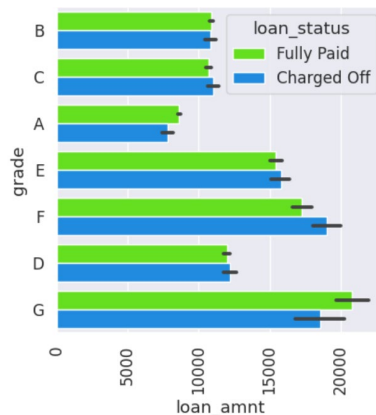
- **Analysis between GRADE & Interest Rate for applicants with loan status as Charged Off**

- The graph in the right indicates for applicants of lower grade the rate of interest charged was low .

Grade A was charged the lowest interest rate while Grade G were charged the highest .



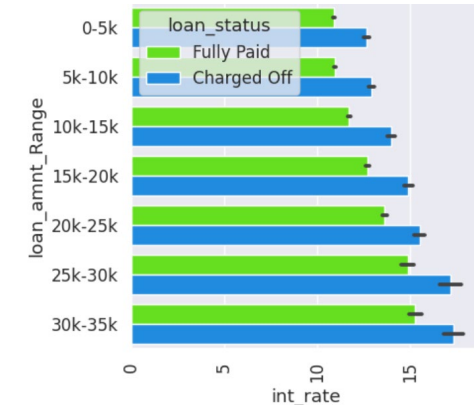
- **Analysis between Loan Amount & GRADE for applicants with loan status as Charged Off**



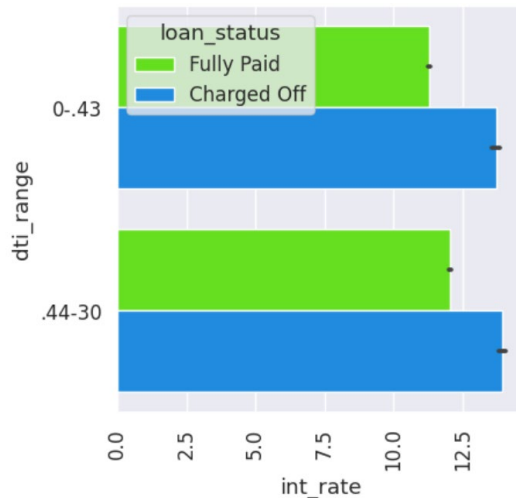
- In the graph to left we observe the Higher Grade with higher loan amount was charged off .
Need to watch out for these .

Bivariate Analysis(File : loan.csv)

- **Analysis between Interest Rate and loan amount for applicants with loan status as Charged Off**
 - The graph in the right indicates for rates were consistently high for all loan amount values .



- **Analysis between DTI Range & Interest Rate for applicants with loan status as Charged Off**

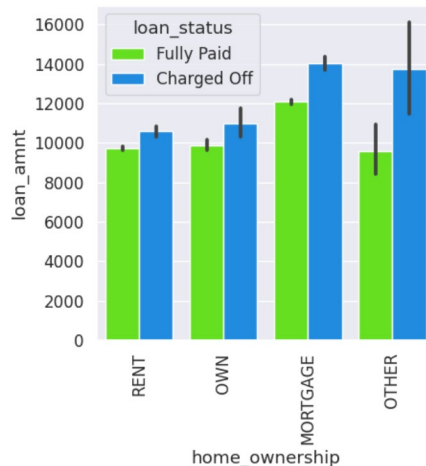


- In the graph to left we observe interest rates were high for both categories of DTI Range .

Bivariate Analysis(File : loan.csv)

- Analysis between Home Ownership and annual income for applicants with loan status as Charged Off
 - The graph in the right indicates that the Charge off was higher for applicants who are have Home ownership status as MORTGAGE or OTHERS

- Analysis between Home Ownership & loan amount for applicants with loan status as Charged Off



- In the graph to left we observe applicants who are have Home ownership status as MORTGAGE or OTHERS has higher loan values have been charged off

Bivariate Analysis(File : loan.csv)

- Analysis between Loan amount and month for applicants with loan status as Charged Off

-In the right graph we can observe that the charge off percentage is high for the loan sanctioned in the month of December



- Analysis between Loan amount & year for applicants with loan status as Charged Off

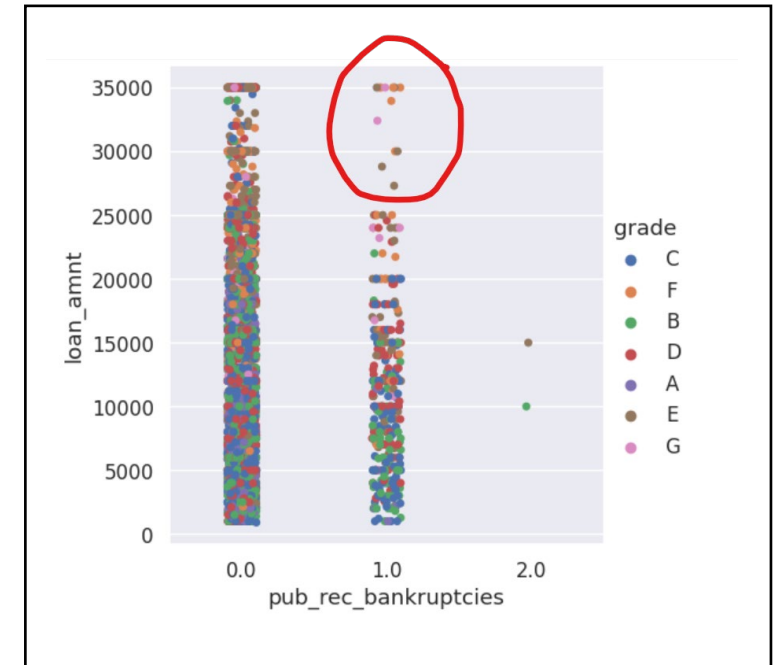


- In the graph to left we observe applicants who have taken loan in 2011 are having highest charge off percentage

Multivariate Analysis (File : loan.csv)

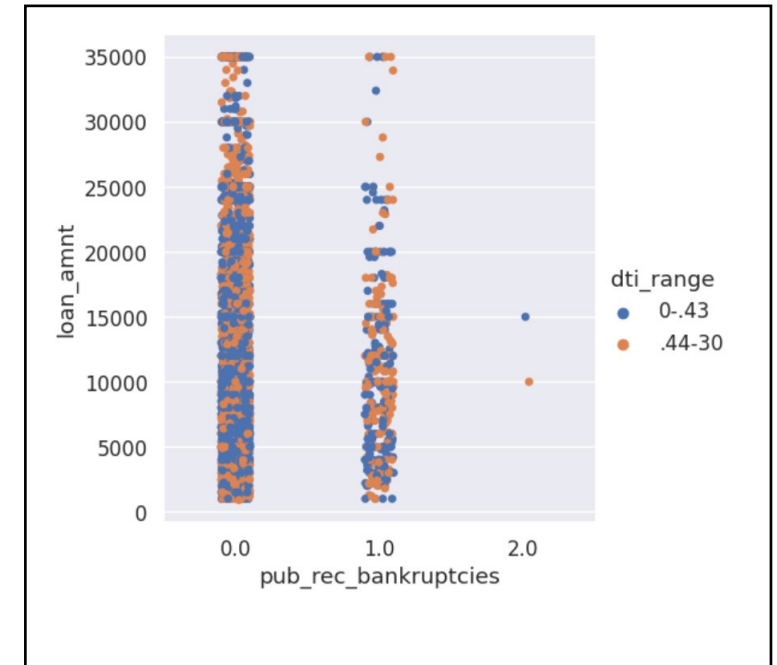
- Analysis between loan amount , GRADE and Public bankruptcy records for applicants with loan status as Charged Off
 - The graph indicates applicants having loan_status as Charged off , who have also declared bankruptcy and are in higher grade has applied for higher loan amount .

Possible candidates as defaulters

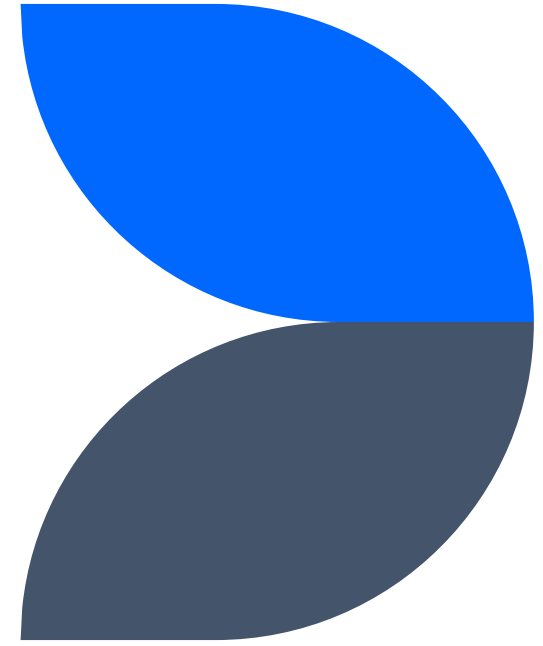


Multivariate Analysis (File : loan.csv)

- Analysis between loan amount , DTI Range and Public bankruptcy records for applicants with loan status as Charged Off
 - Observation from the graph : Applicants with DTI range $> .43$ who have applied for higher bank loan & have declared bankruptcy can be possible defaulters , Also high charged off applicants with higher loan value and are in a range of DTI $> .43$ can be defaulter



Recomendations



Recommendations (File : loan.csv)

Financial company need to consider the following driving factors to decide whether the applicants are risky(Defaulters / loan status is charged off) or not before approving the loan:

- 1.Applicant with **grade** E,F,G are less likely to repay the loan.
- 2.**Verification** of applicant should be done before approving the loan.
- 3.**Higher loan amount** and **high interest** rate for **small business group** will put the financial company at risk.
- 4.Applicant who are in **Mortgage** and **Rent** home are not likely to replay the loan.
- 5.Applicant with **less loan term** are likely to become defaulter.
- 6.Applicants with **DTI range** > .43 who have applied for higher bank loan & have declared **bankruptcy** can be possible defaulters , Also high charged off applicants with higher loan value and are in a range of DRI > .43 can be defaulter .
- 7.Applicant with **higher employment length** are seen to apply for bigger loan amounts.
- 8.Applicants are less likely to repay the loan if the loan is sanctioned in the **month of December**.

Meet our team



Parameshwari.R

Group Facilitator



Pamela Roy

Group Member



Thank you

Parameshwari.R

paramshetty100@gmail.com

Pamela Roy

bhattacharya.pam@gmail.com