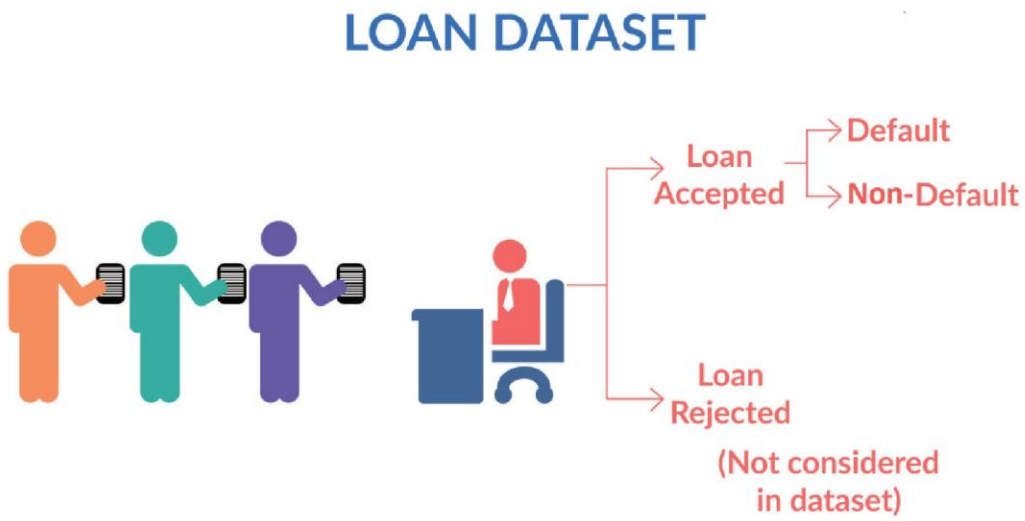
Lending Club Case Study



Problem Statement:

The Lending Club is a peer-to-peer lending platform that connects borrowers and investors. The company has provided loan data for a period of time to perform exploratory data analysis (EDA) to gain insights into the lending behaviour of borrowers and investors. The objective of this analysis is to identify patterns and trends in the data that can help understand the factors that affect loan performance and default rates. The analysis will involve examining the loan data to gain insights into the loan characteristics such as loan amount, interest rate, term, purpose, borrower's credit history, and employment status. The analysis will also explore the relationship between loan performance and other variables such as the borrower's income, debt-to-income ratio, and credit score. The goal of the analysis is to provide insights

that can help investors make better decisions and reduce the risk of loan defaults.

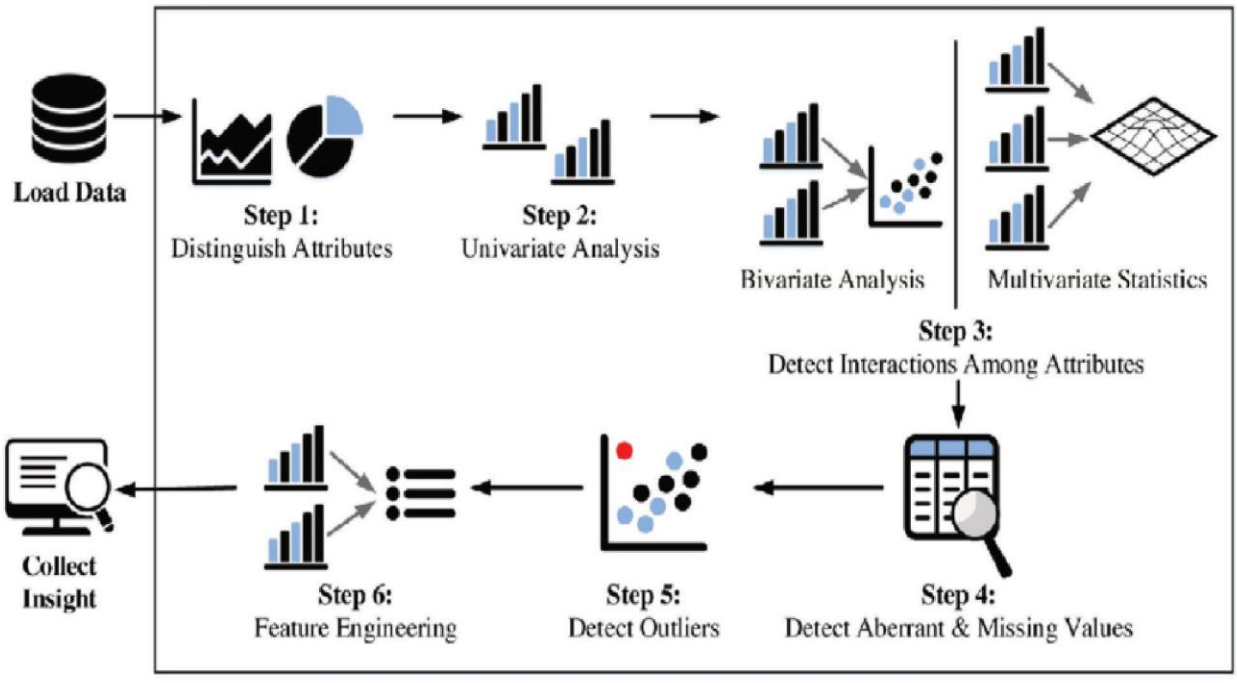
Data Understanding:

- loan_amnt: The total amount of the loan that was issued to the borrower.
- term: The length of time over which the loan is scheduled to be repaid (either 36 months or 60 months).
- int_rate: The interest rate assigned to the loan.
- installment: The monthly payment due on the loan.
- grade: The assigned loan grade (A through G) based on the borrower's credit score.
- sub_grade: The assigned subgrade (1 through 5) based on the borrower's credit score.
- emp_title: The job title provided by the borrower during application.
- emp_length: The length of time the borrower has been employed.
- home_ownership: The type of home ownership (rent, own, mortgage, etc.)
- desc: The discription for which the borrower is taking the loan
- title: The title for which the the loan is borrowed
- annual_inc: The borrower's self-reported annual income.
- verification_status: Indicates if income was verified by LC, not verified, or if the income source was verified.
- issue_d: The month the loan was issued.
- loan_status: Current status of the loan (Fully Paid, Charged Off, Late, etc.).
- pymnt_plan: Indicates if a payment plan has been put in place for the loan.
- purpose: The purpose of the loan (debt consolidation, credit card refinancing, etc.).
- zip_code: The first three digits of the borrower's zip code.

- addr_state: The state where the borrower resides.
- dti: Debt-to-Income ratio, calculated by dividing monthly debt payments by monthly income.
- delinq_2yrs: The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
- earliest_cr_line: the month in which the borrower first opened their oldest credit account that is listed on their credit report.
- inq_last_6mths: The number of inquiries made to the borrower's credit report in the last 6 months.
- open_acc: The number of open credit lines in the borrower's credit file.
- pub_rec: The number of derogatory public records in the borrower's credit file.
- revol_bal: The total revolving balance is the amount of credit that the borrower has used and has not yet paid back, and is an important factor in assessing the borrower's creditworthiness
- revol_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to the total credit available.
- total_acc: The total number of credit lines in the borrower's credit file.
- initial_list_status: The initial listing status of the loan.
- out_prncp: Remaining outstanding principal for total amount funded.it shows the remaining balance on the loan at the time when the data was collected.
- out_prncp_inv: Remaining outstanding principal for portion of total amount funded by investors. It represents the outstanding principal amount on the loan that is yet to be paid off by the borrower for the part of the loan funded by investors. It is similar to out_prncp, but takes into account the amount of the loan funded by investors rather than the total loan amount.
- total_pymnt: refers to the total payment received from the borrower on the corresponding loan, including the principal amount and any interest or fees charged.
- total_pymnt_inv: Payments received to date for portion of total amount funded by investors.
- total_rec_prncp: Principal received to date. refers to the total amount of principal that has been received up to now. This includes the part of the monthly payment that goes towards paying off the principal amount of the loan.

- total_rec_int: Interest received to date.
- total_rec_late_fee: Late fees received to date.
- recoveries: post charge off gross recovery. refers to the amount of money a lender is able to collect from a borrower after the borrower has defaulted on their loan.
- collection_recovery_fee: post charge off collection fee.
- last_pymnt_d: Last month payment was received.
- last_pymnt_amnt: Last total payment amount received.It includes both the principal and interest portion of the payment.
- next_pymnt_d: Next scheduled payment date.
- last_credit_pull_d: The month in which the borrower's credit report was last pulled for the purpose of granting credit.
- pub_rec_bankruptcies: Number of public record bankruptcies.

Steps Involved in Exploratory Data Analysis:



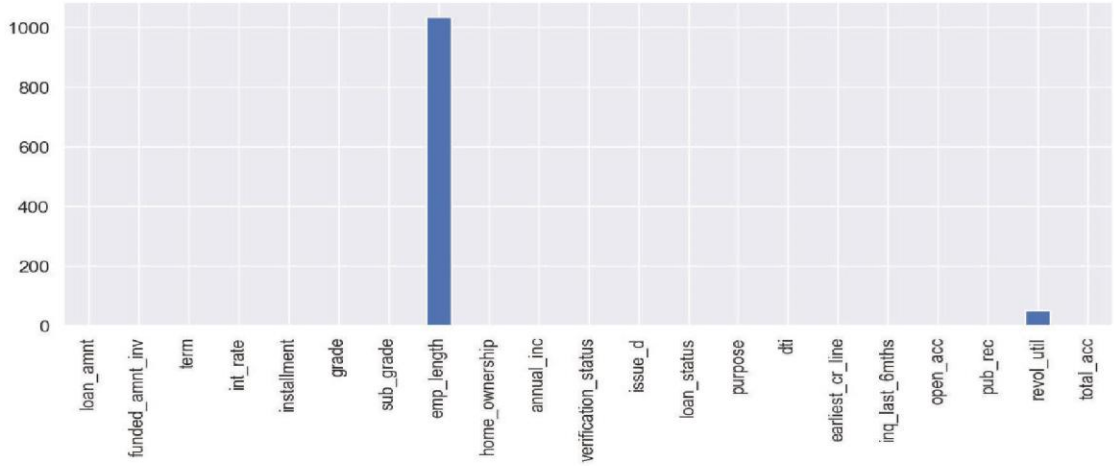
Step 1: Understanding and Loading the Data followed by Observations

- There are Total 39717 number of records and 111 attributes in the dataset
- The Target column which is loan status has 3 distinguish values
    - Fully Paid
    - Current
    - Charged Off
- In order to determine the pattern of defaulters we don't need the data where borro wer is
  currently paying the loan so we can remove the rows from the dataset where the loa n status is
  current
- After checking for the unique values in the data it has been observed that there are many
  Features in the dataset that have all the unique values as well as there are many feat ures where all the values are similar along with the features where all the values are Nan in such cases, we can remove all of the above features
- There are few features such as URL , desc that is description which also do not add m eaning or do not help in determining the pattern to be the defaulter we can remove such features from the
  Data
- Look for the description of the data in order to understand it in a better way
- Look for Categorical and Numerical Columns in the data

	loan_amnt	funded_amnt_inv	installment	annual_inc	dti	inq_last_6mths	open_acc	pub_rec	total_acc
count	38577.000000	38577.000000	38577.000000	3.857700e+04	38577.000000	38577.000000	38577.000000	38577.000000	38577.000000
mean	11047.025430	10222.481123	322.466318	6.877797e+04	13.272727	0.871737	9.275423	0.055422	22.052648
std	7348.441646	7022.720644	208.639215	6.421868e+04	6.673044	1.071546	4.401588	0.237804	11.425861
min	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	2.000000	0.000000	2.000000
25%	5300.000000	5000.000000	165.740000	4.000000e+04	8.130000	0.000000	6.000000	0.000000	13.000000
50%	9600.000000	8733.440000	277.860000	5.886800e+04	13.370000	1.000000	9.000000	0.000000	20.000000
75%	15000.000000	14000.000000	425.550000	8.200000e+04	18.560000	1.000000	12.000000	0.000000	29.000000
max	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	8.000000	44.000000	4.000000	90.000000

Step 2: Cleaning the Data Missing value imputation

Check for Missing value

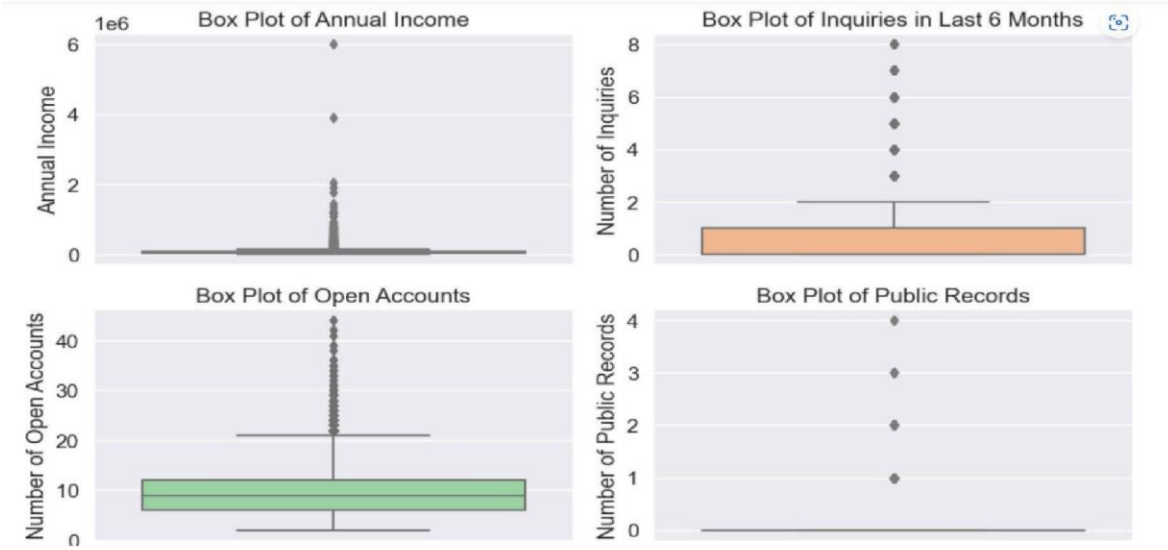


- emp_length has 2.68% null values in it that is 1033 values are null
- revol_util has 0.13% null values in it that is 50 values are null

Step -3 Standardization of the Data

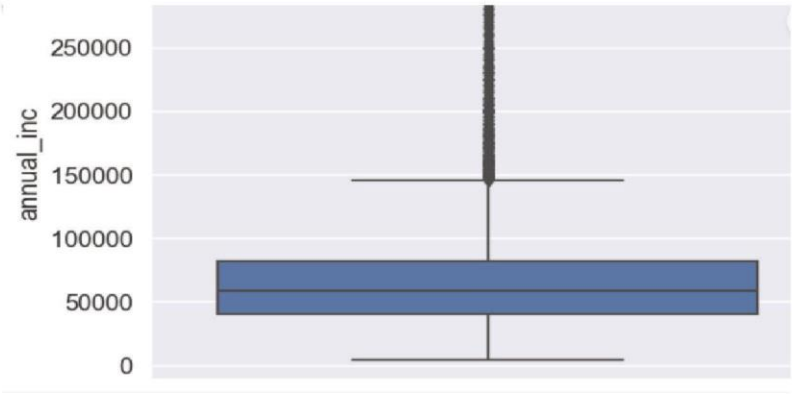
- Analyse the columns and wherever possible and important we can convert the Values to Numerical by removing the symbols like % or $ whatever is present or simply by Replacing the values in the numbers

Step -4 Outliers Detection & Treating them



- It was important to treat the Annual income column and replace the outliers with the threshold values instead of directly removing them from the data. The rest of the 3 columns have the quite continuous distribution so we are not handling them

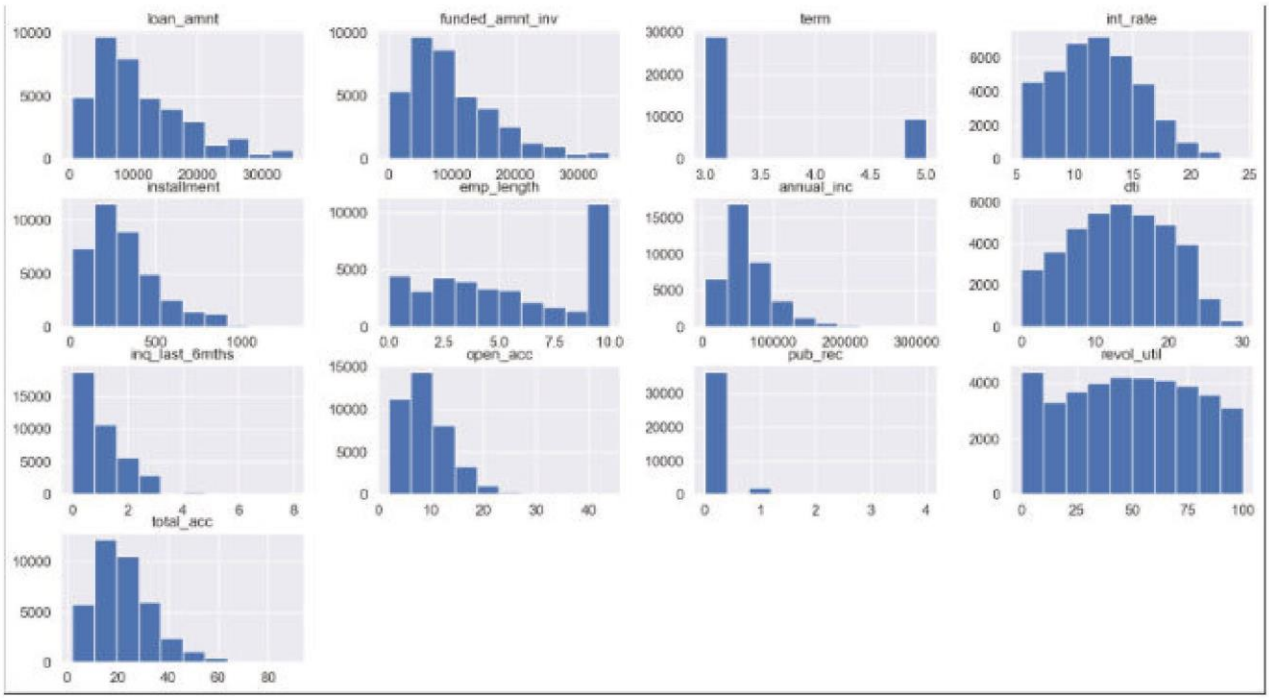
- After treating the outliers in the Annual Income column



Step – 5 Duplicate Data Inspection

- There are no Duplicate values in the data

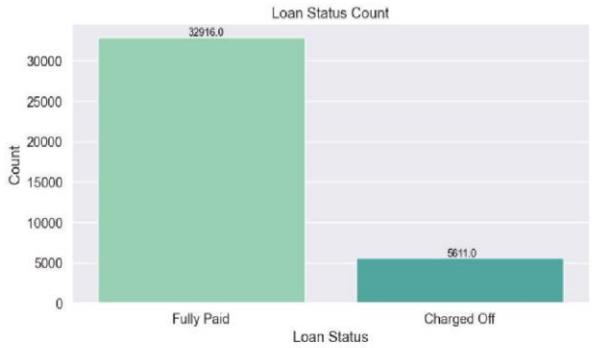
Step -6 Data Distribution

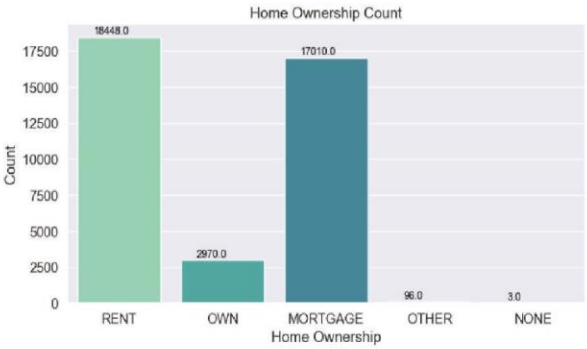


- Looking at the above Distribution plots the features have been observed to be skewed that is they are not normally distributed
    - loan amount
    - funded amount invested
    - instalments
    - inquiry in last 6 months
    - open accounts
    - annual income
    - total accounts
- all of the above features are right skewed wherein the tail is extended to the right side making it positively skewed data
- The values are not equally distributed and that's why it is a skewed data
- There are features that are normally distributed which means that the values are equally distributed in the data such as
    - revol_util
    - dti

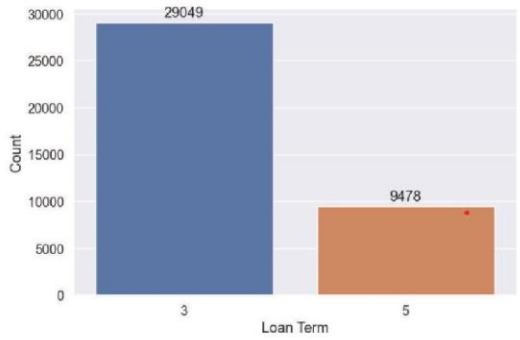
- emp_length

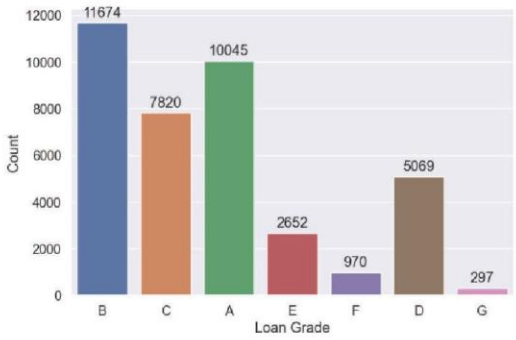
step – 7 Univariate Analysis

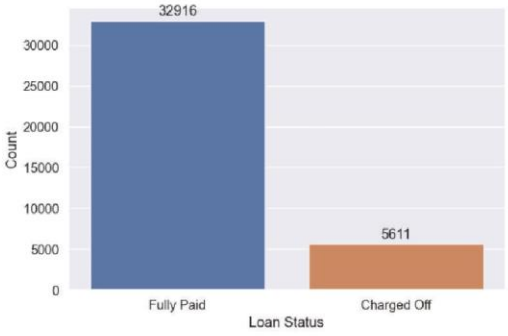




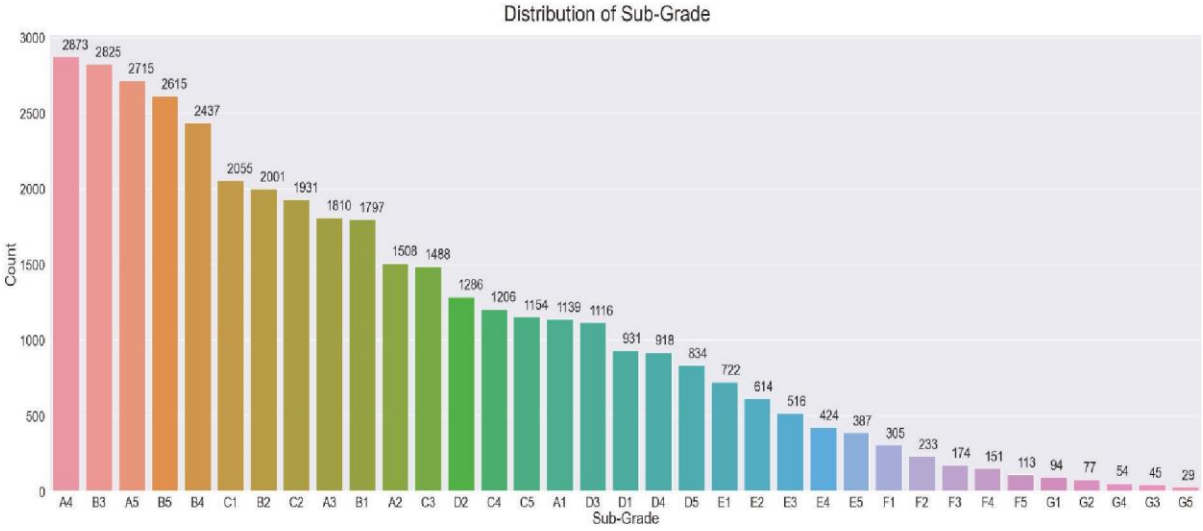
- There are 5611 people who are in the defaulters list
- There are more people being on rent who applies for the loan followed by the people who kept the property on the mortgage



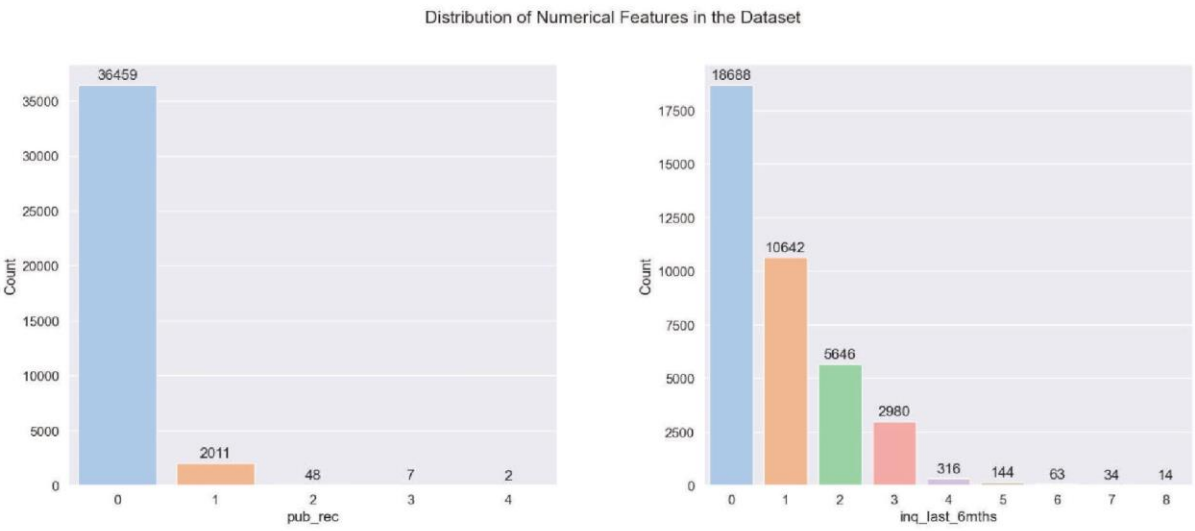




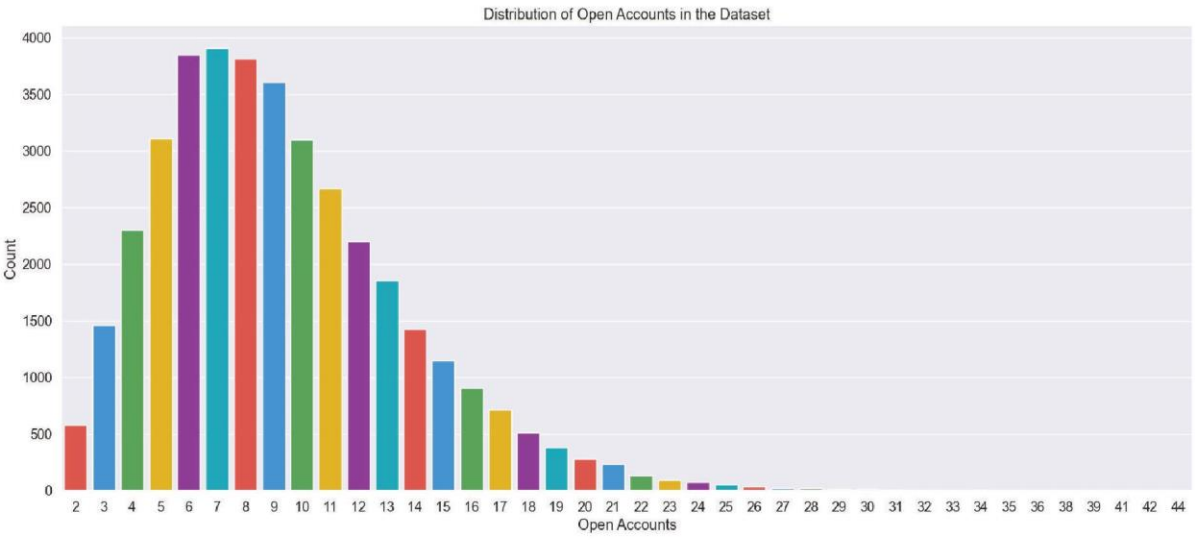
- The loan is given for 3 or 5 years of time period to the borrowers. There are people more people who are supposed to pay back the loan within 3 years of time period
- There are in all 7 major categories that are A, B, C, D, E,F, G. The people who have been classified in the B grade are more in number to apply for the loan
- There are 3 types of people for whom the loan has been approved that is the people whose documents or status is verified by the bank who's lending the money to people followed by source verified people whose background verification is done by the third party and last the people whose status is not at all verified. There are a greater number of people whose status is not verified and still given the loans which may take a toll on the bank who is lending the money to such people whose status is not verified so it is recommended to verify the status before approving the loan for a particular person

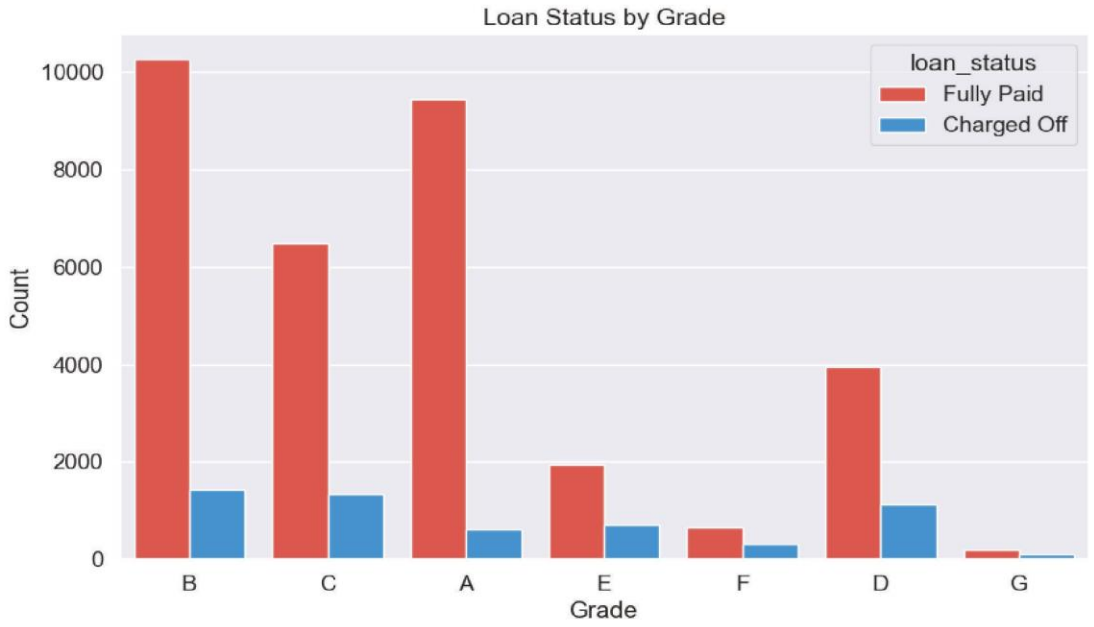


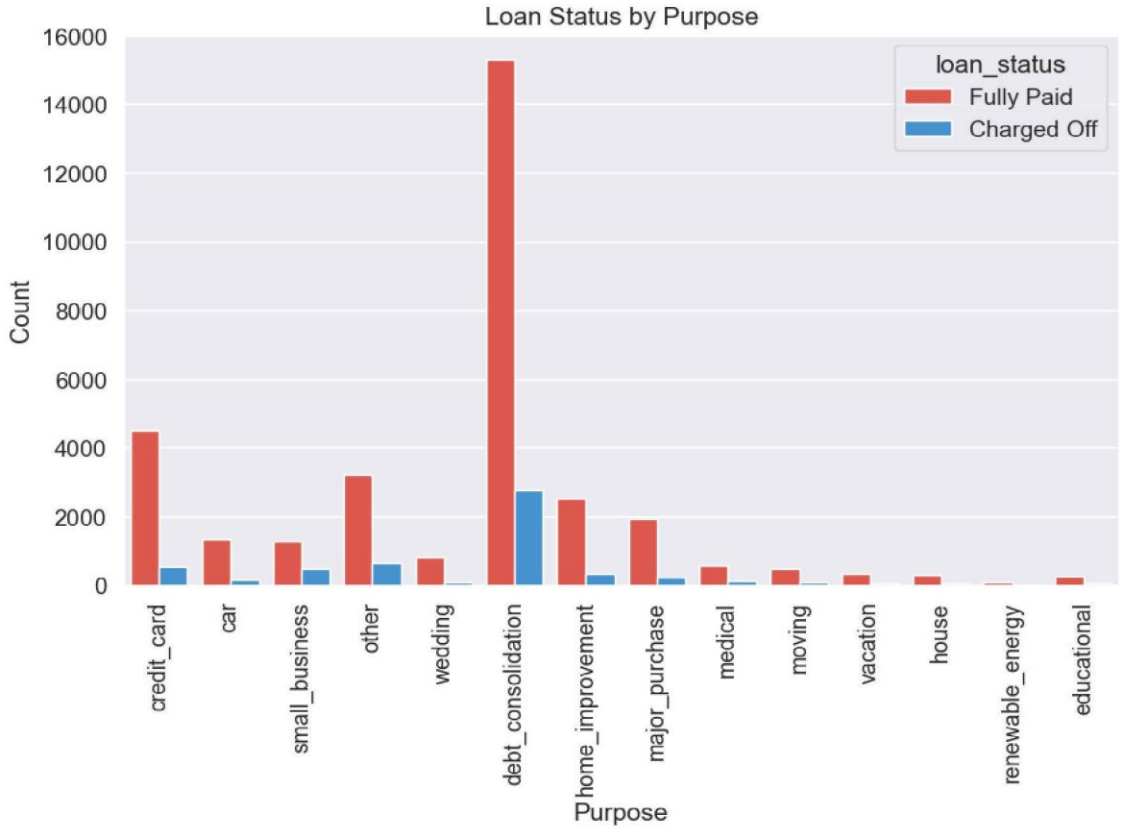
- Grade is further divided into 5 parts as sub grade category for example in the grade A there are 5 sub grades for that such as A1, A2, A3, A4, A5 and similar for the rest of the grades
- The people from sub grade A4 are the major part of borrowing the loan



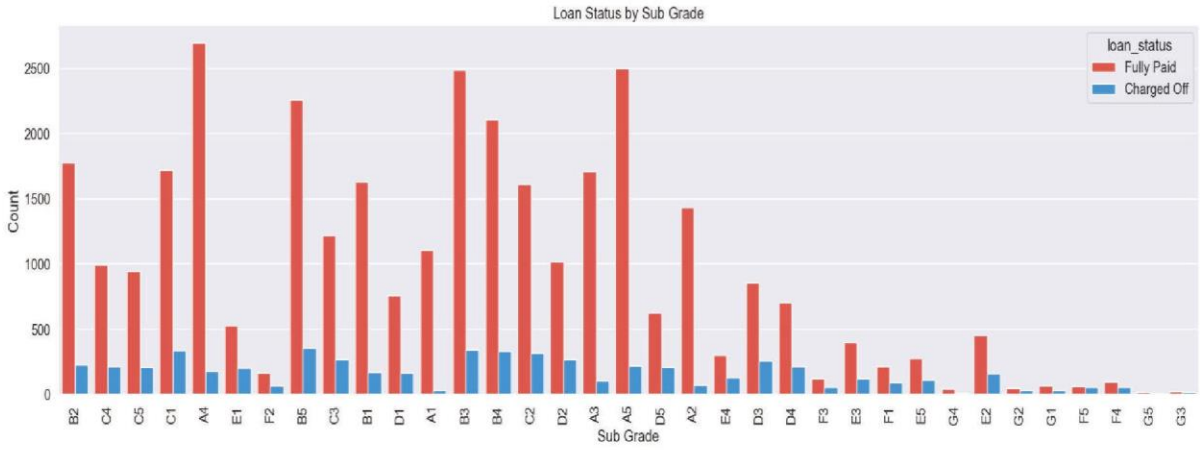
- The people who applied for the loan have either 0 public records or they have 1 or more public records which we get to know while doing the background verification check of the person
- There are a greater number of people who belongs to no public record area which is good for the bank as we can still rely on the people having no public records than that of the people who tends to have some or the other public records
- we have the feature called inquiry in last 6 months whose information is also filled while doing the background check in order to know the history of the loan borrower wherein either there are 0 or more inquiries made to that person's account in order to know the credit history of the person



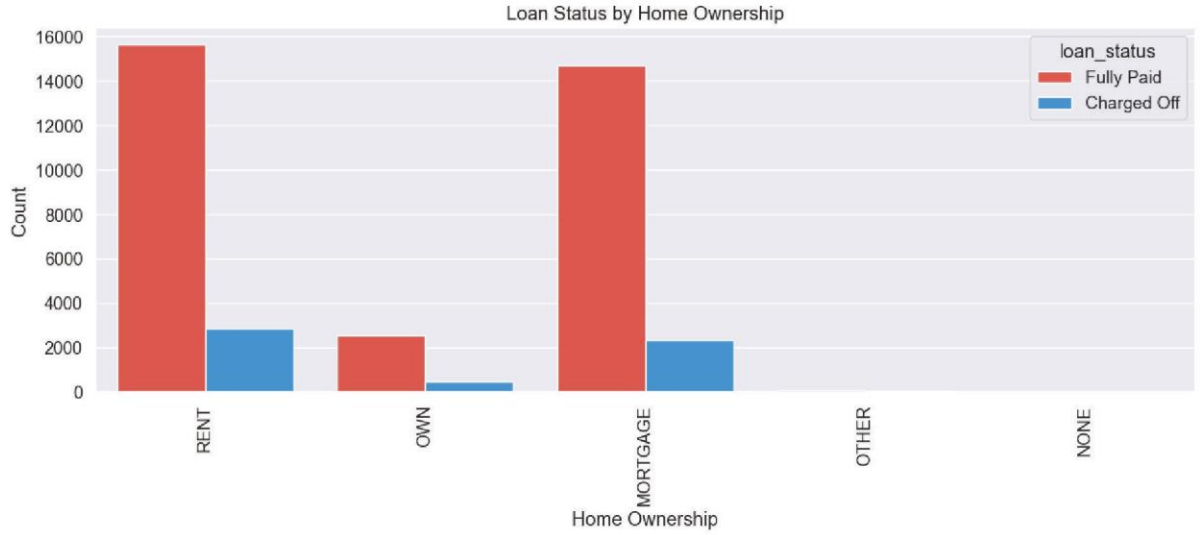




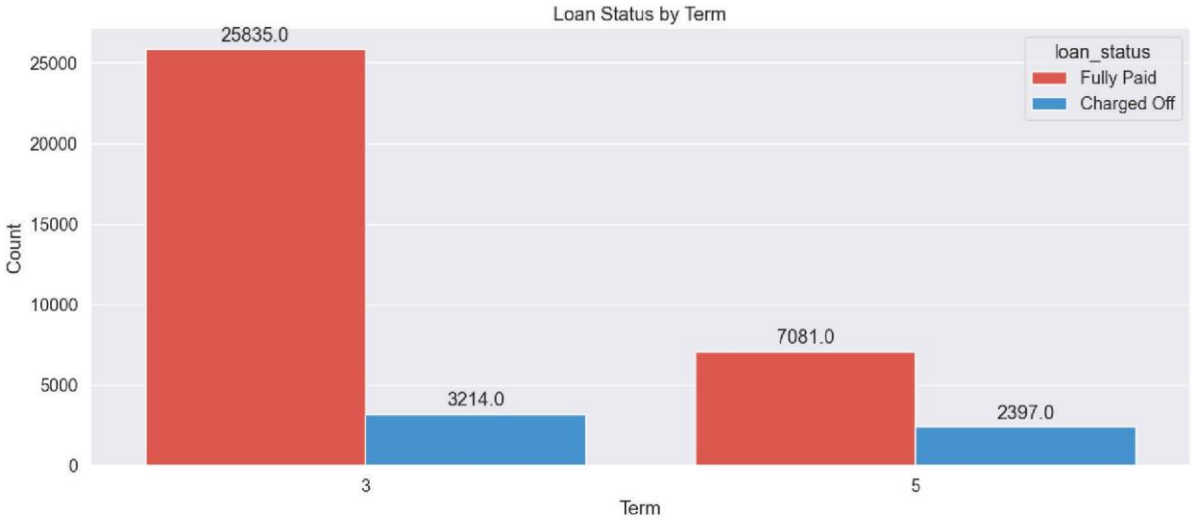
- The people who are most likely to be defaulter belongs to the grade B followed by grade D as per the above observations
- The people who take the loan for debt consolidation purpose that is to clear other loans are more likely to be in the list of defaulters



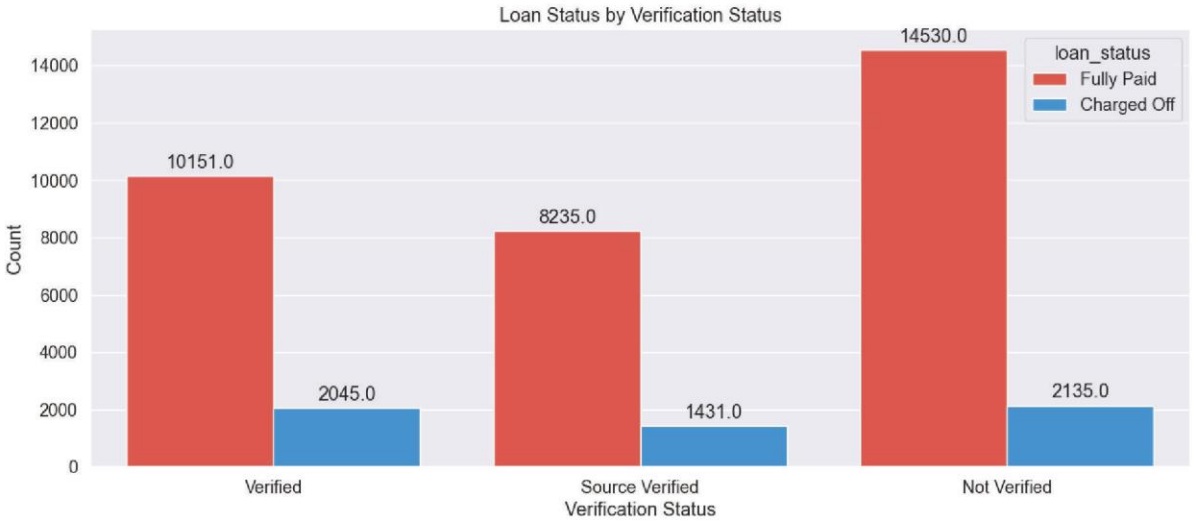
- The People who are more likely to be the defaulters belongs to sub grade B5 followed by sub grade B4 and so on



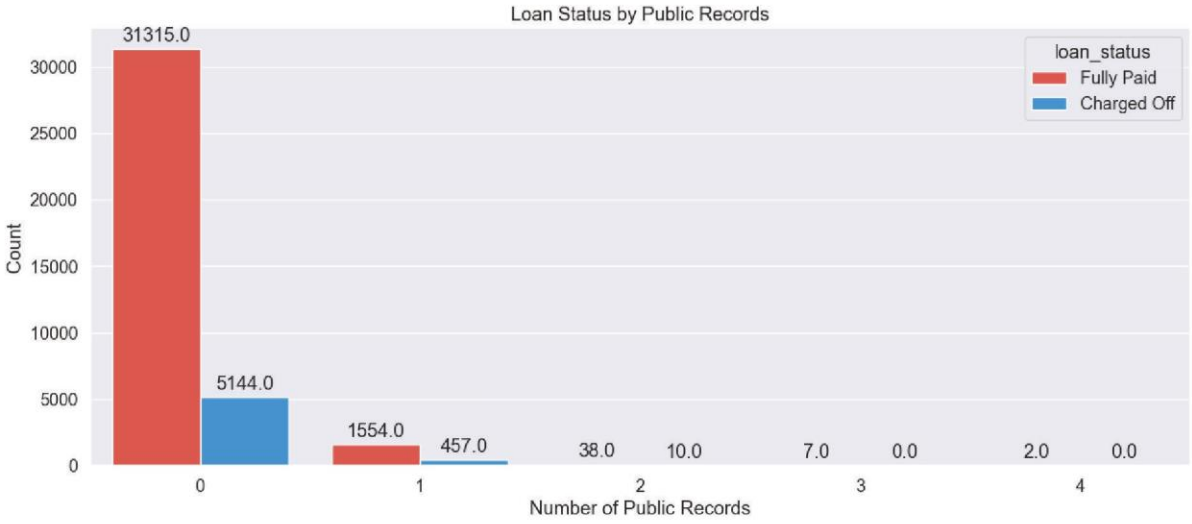
- There are people who are more likely to default the loans who lives in Rental Apartments



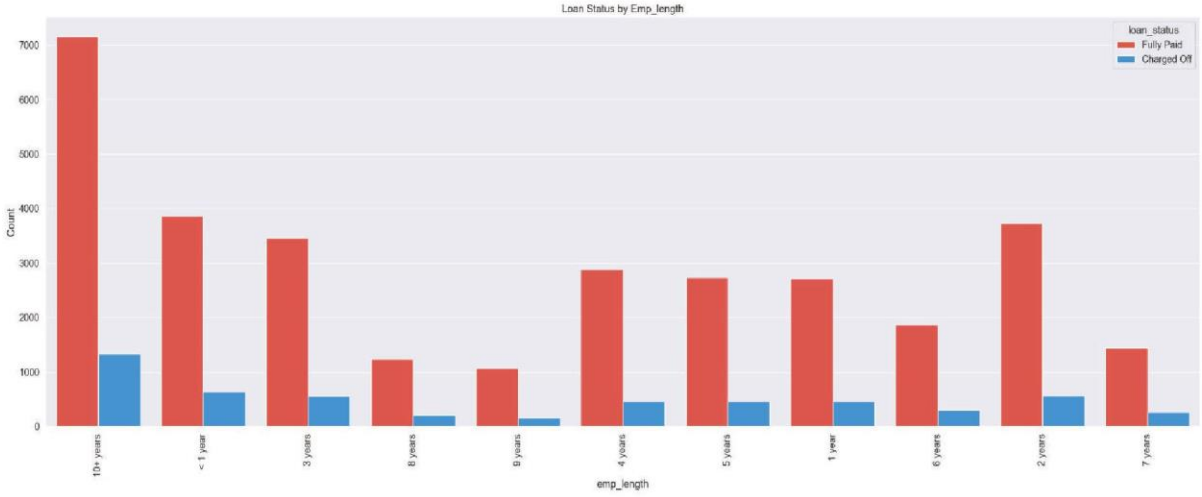
- The people whose time period is less that is 3 years to repay the loan are more likely to be the defaulters in comparison with the people whose time period is more to repay the loan



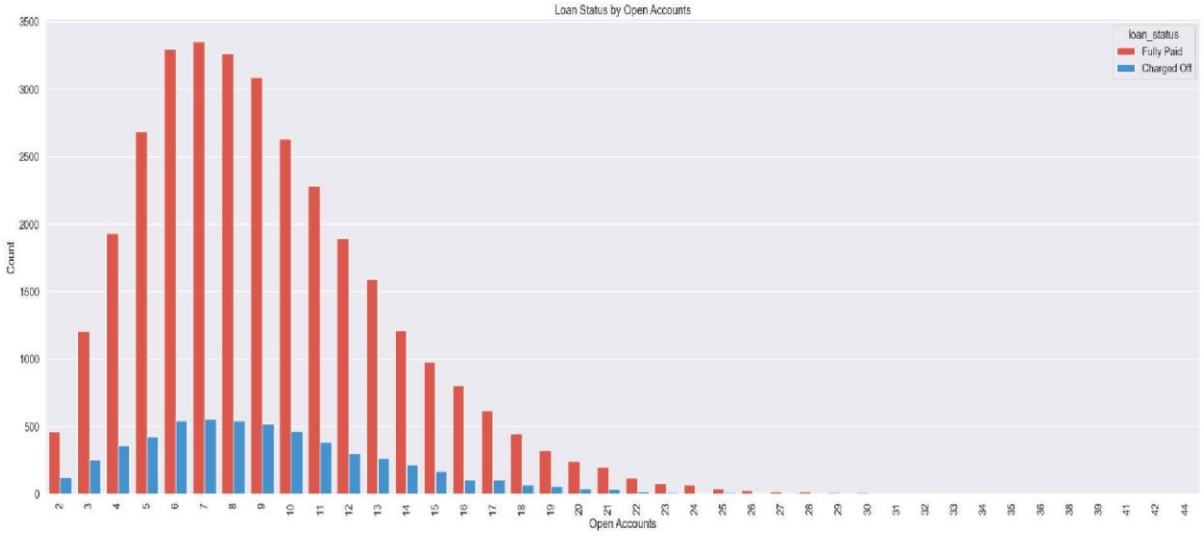
- The people whose verification is not done and for whom the loan is approved without any background check are most likely to be defaulters so it is recommended not to approve the loan without any background check of the loan borrowers

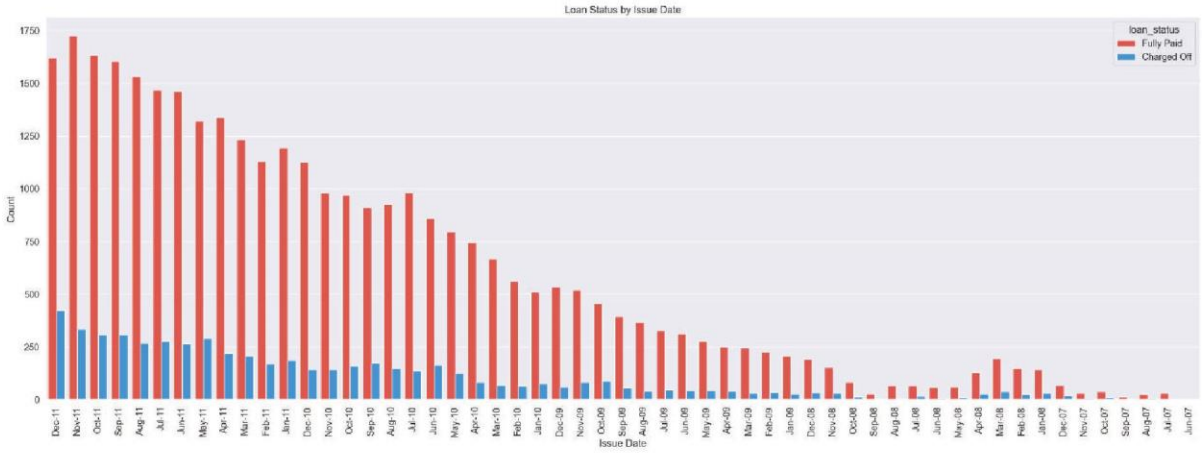


- There are people whose public records are zero are still tend to be in the defaulters list so we cannot blindly trust the borrower by just checking the public records it is recommended to go through few more documents to know the behaviour of the borrower



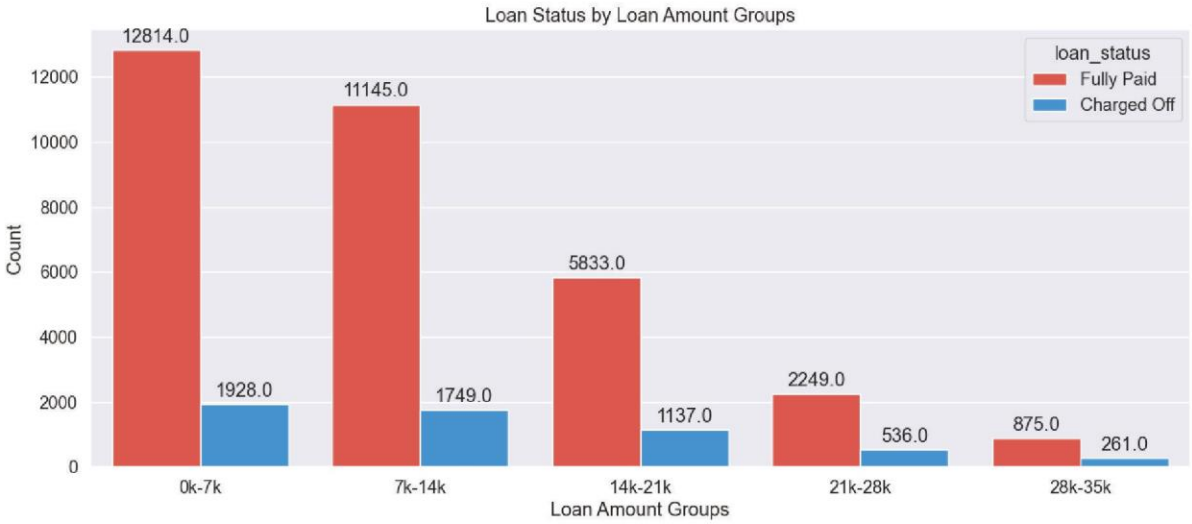
- The people who have more employment time period are the majority in the defaulters list in comparison with the others



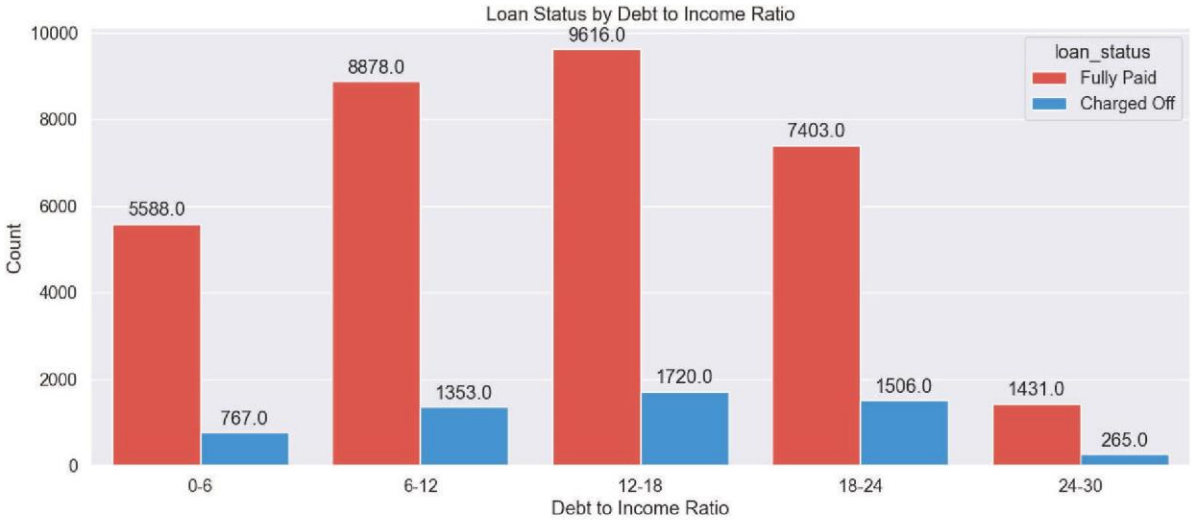


- In the month of December 2011, the people who have taken the loans are likely to be the defaulters followed by May 2011 and so on
- This may happen that there is any financial crisis in that country in those time periods so it is recommended to take in account these kinds of situations also while approving for the loans

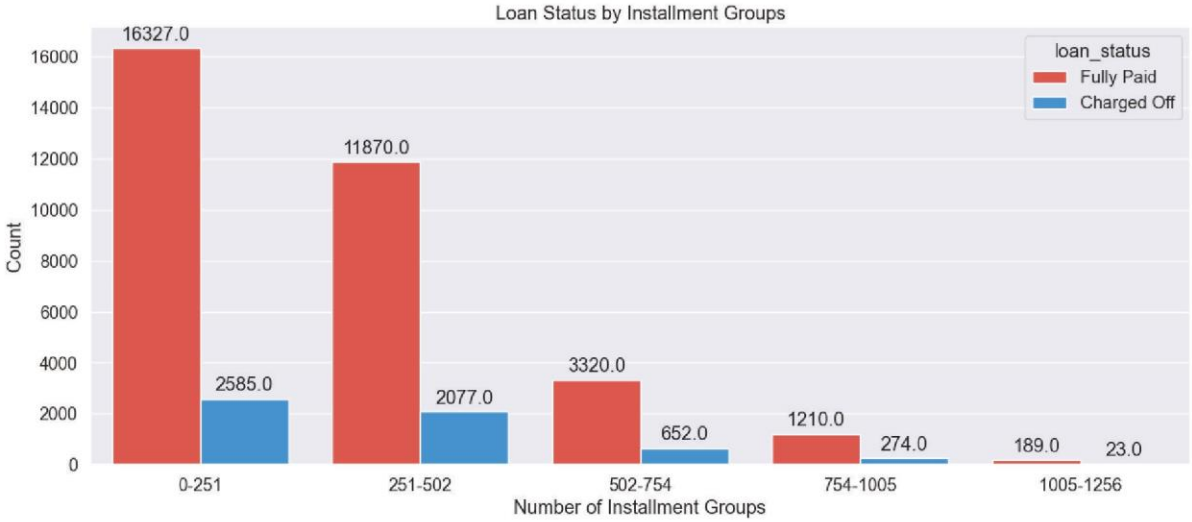
- We are using the binning approach in order to gather some more information regarding the features that effects on defaulters
- In this approach we are creating the 5 bins for selected features having equal records in each of the bins
- let's create the bins for below features
    - funded_amnt_inv
    - int_rate
    - open_acc
    - revol_util
    - total_acc
    - annual_inc
    - loan_amnt
    - dti
    - installment



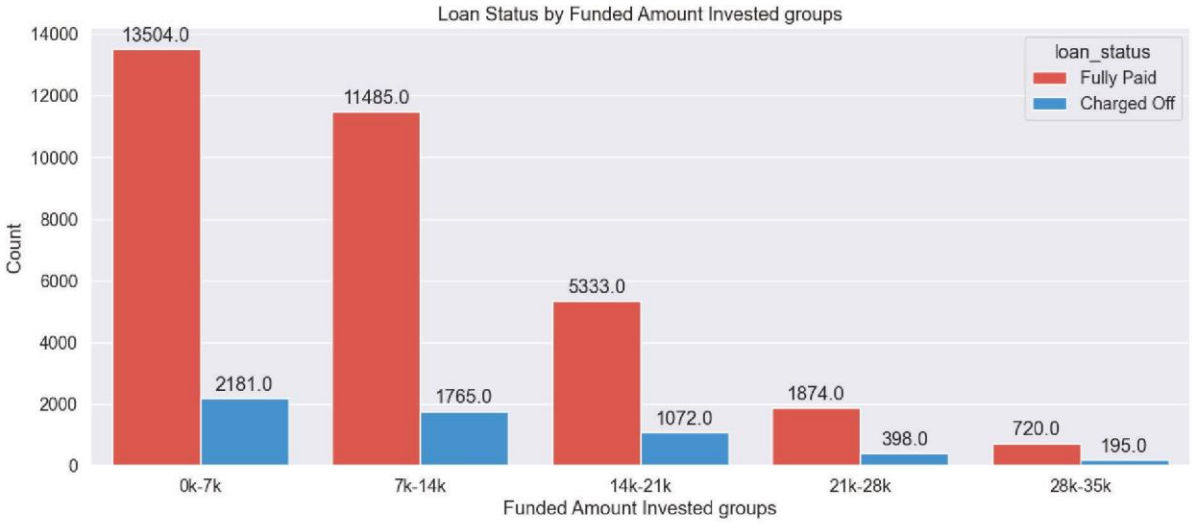
- People who take the loan amount in the range of 0 - 7000 are more likely to be the defaulters the reason could be that they can't afford to take much loan repay to the bank



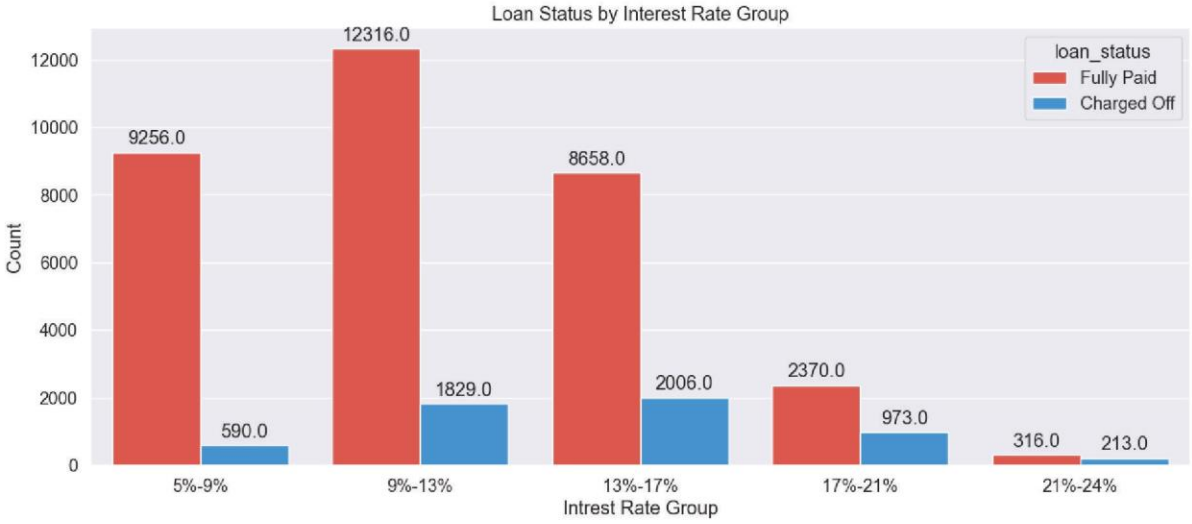
- The people whose debt-to-income ratio is in between 12-18% are most likely to be in the defaulters list



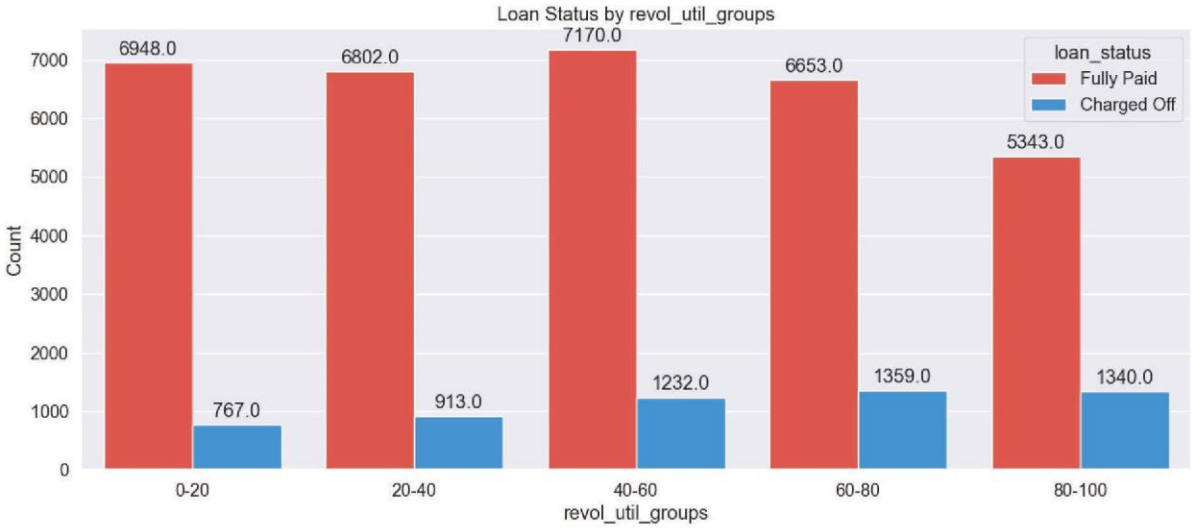
- The people who have installments of 0 - 251 are likely to be the defaulters although being small amount they are unable to repay the loan



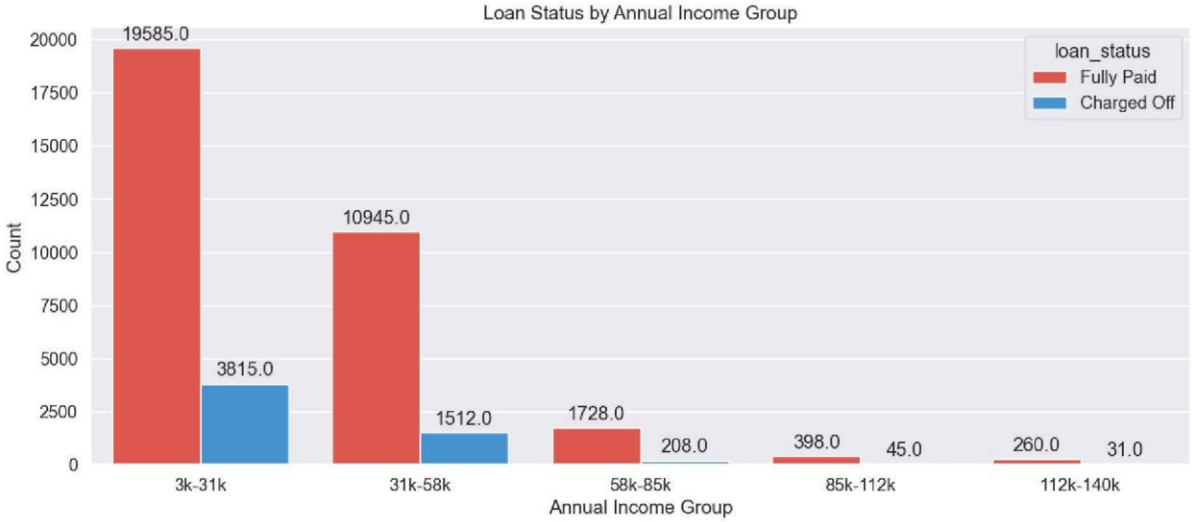
- The people for whom the invested amount is in the range of 0 - 7000 are most likely to be the defaulters



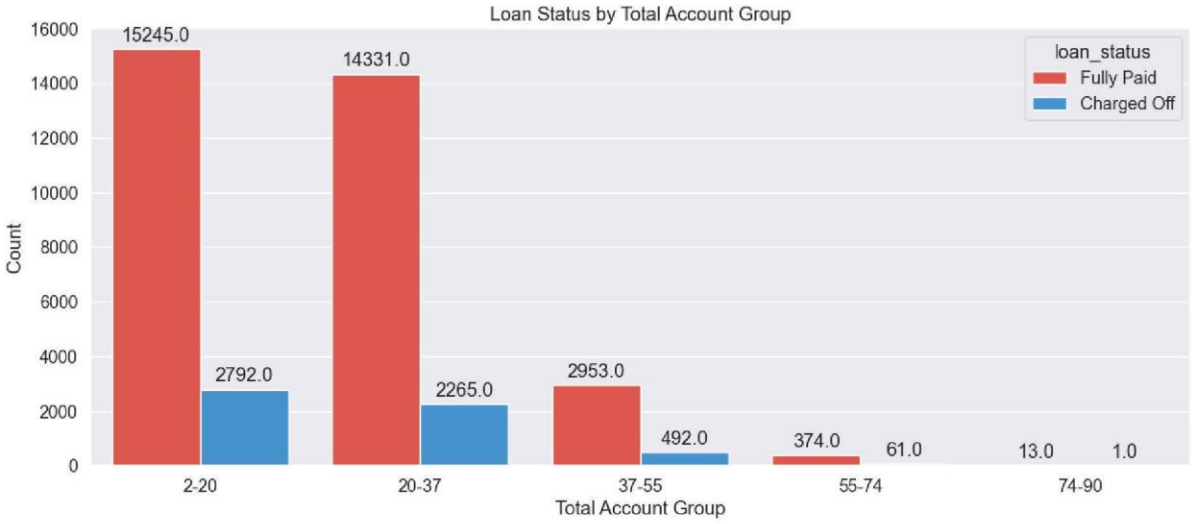
- The people whose interest rate is in the range of 13-17% are tend to default the laon followed by the people having the interest range between 9-13%



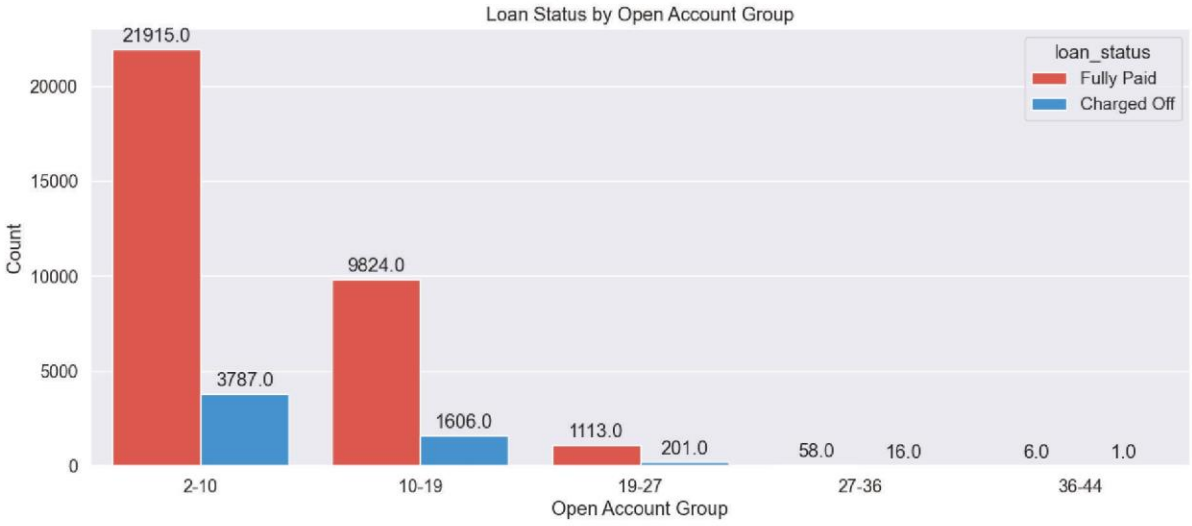
- The people who are in the group of 60-80 revolving util are more likely to default the loans



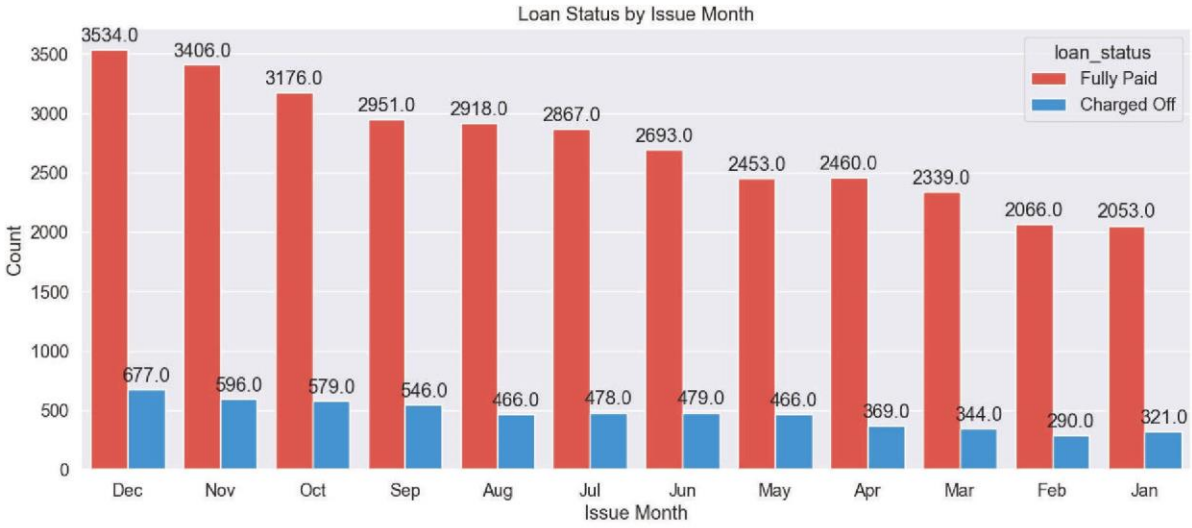
- The people who have less income that is in the range of 3 - 31k are more likely to be the defaulters



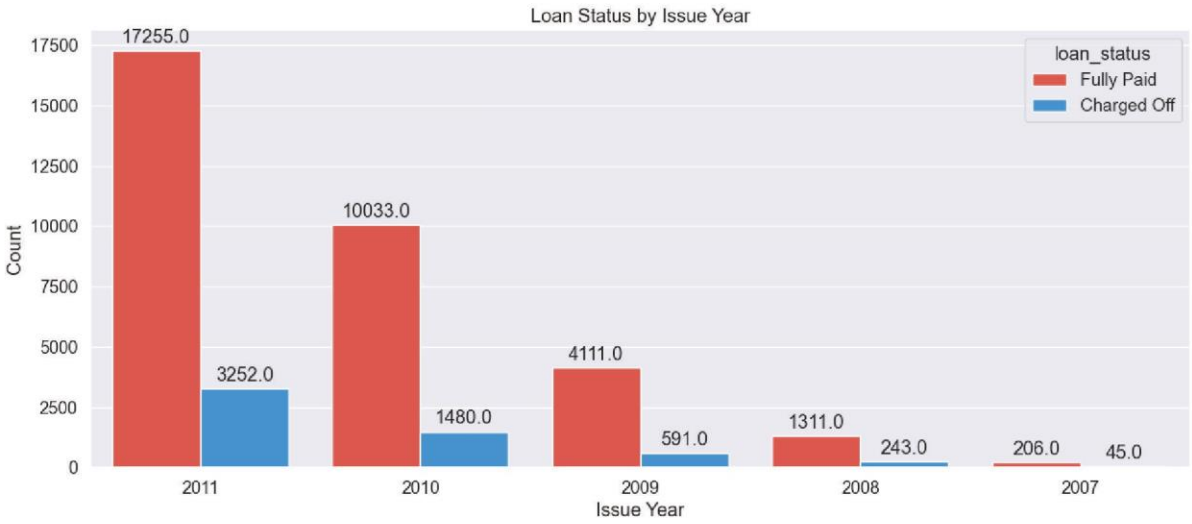
- The people have Total Accounts in the range of 2-20 are more likely to be the defaulters



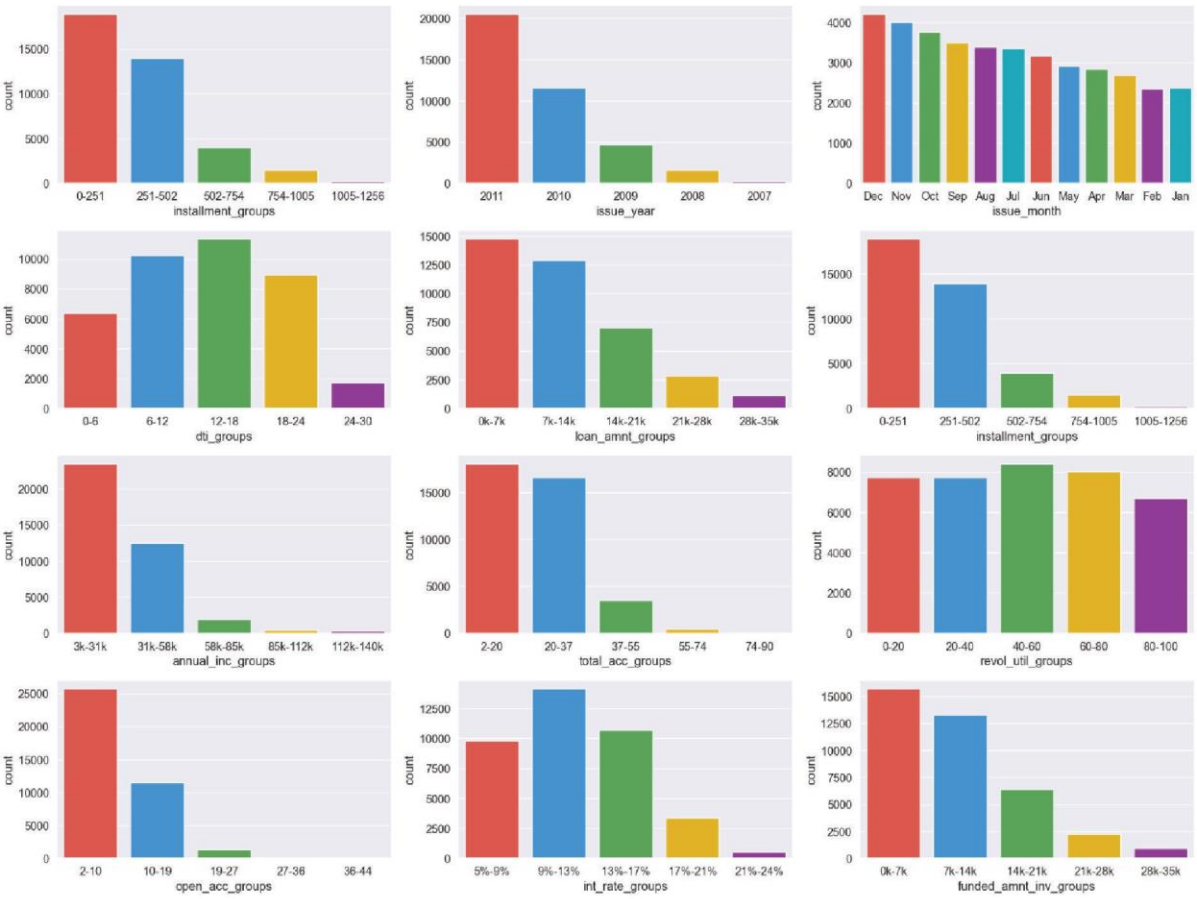
- The people who have the open accounts in the range of 2-10 are more likely to be the defaulters as the reason could be that they are unable to manage their finances

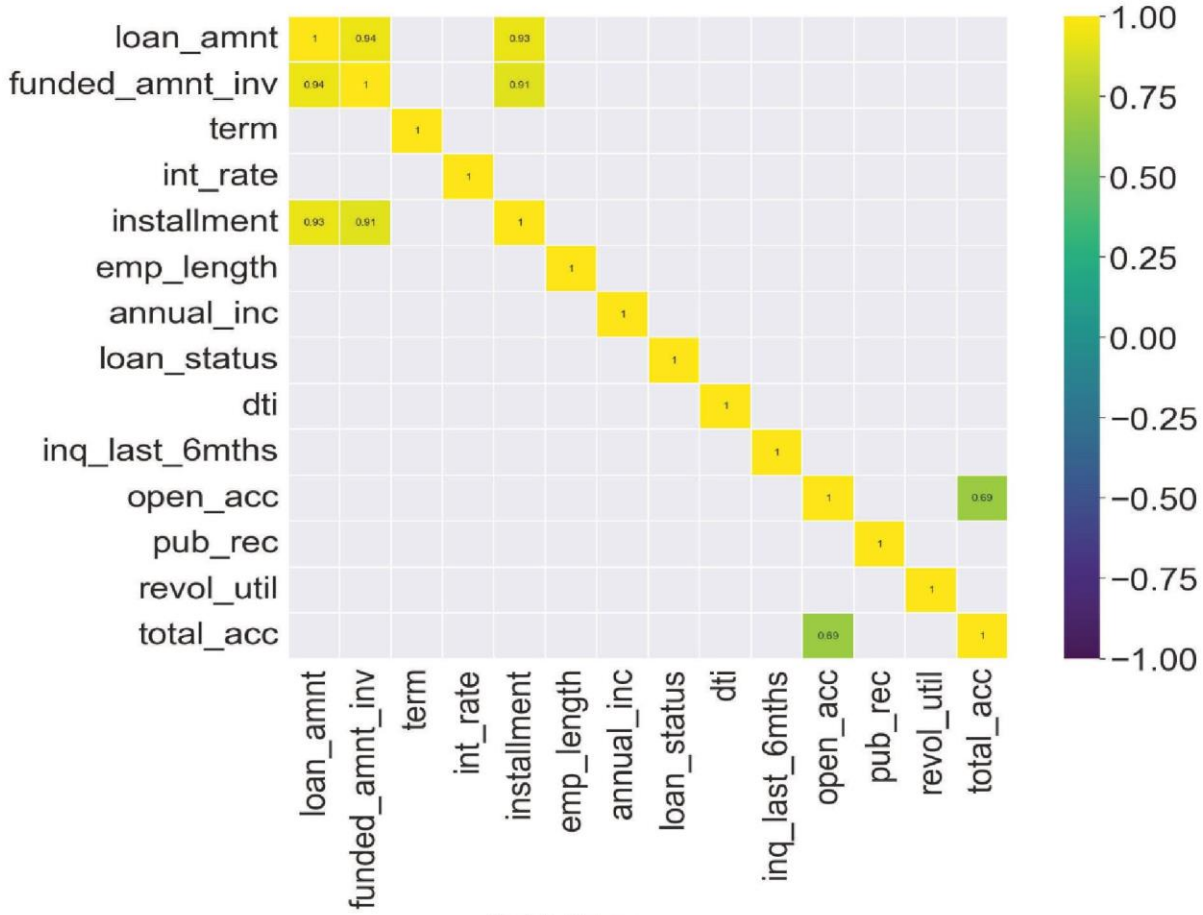


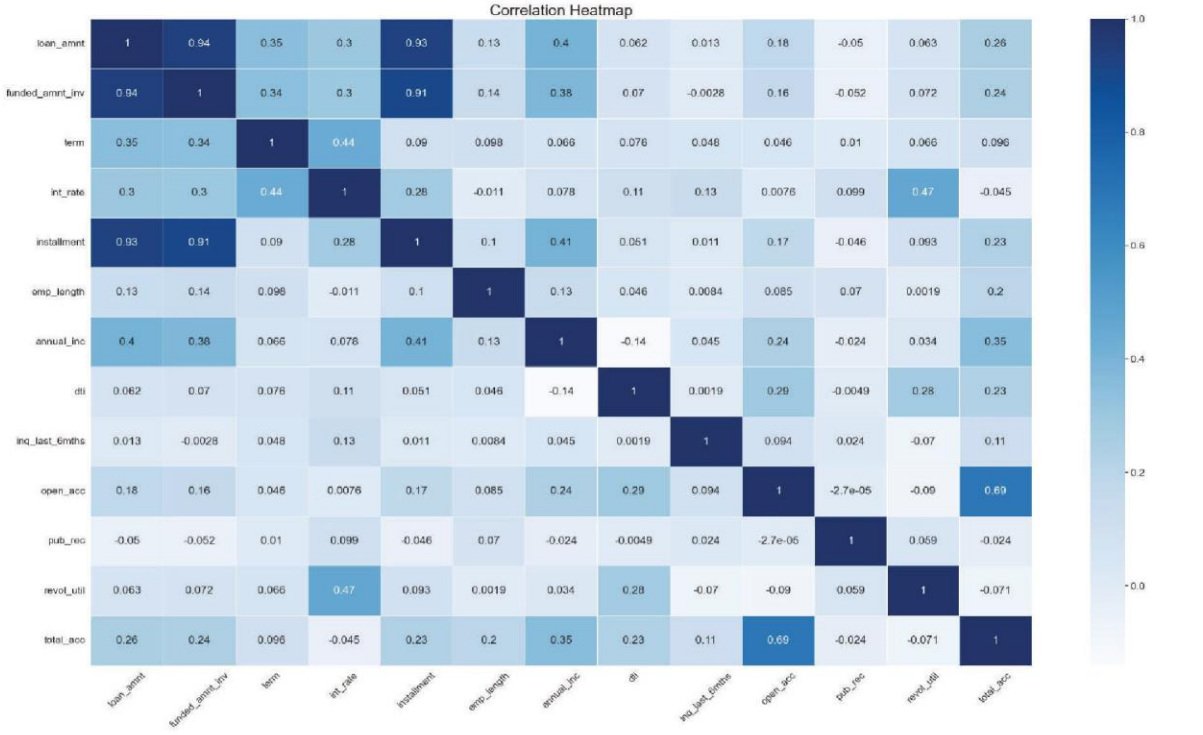
- The people who have taken the loan in the month of the December are more likely to be in the list of defaulters



- The people who have taken the loan in the year of 2011 are more likely to be the defaulters
- The reason for the above case could be that the country could have the financial crisis


count
0-251
251-502
502-754
754-1005
1005-1256
installment_groups
2011
2010
2009
2008
2007
issue_year
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
issue_month
0-6
6-12
12-18
18-24
24-30
dti_groups
0k-7k
7k-14k
14k-21k
21k-28k
28k-35k
loan_amnt_groups
3k-31k
31k-58k
58k-85k
85k-112k
112k-140k
annual_inc_groups
2-20
20-37
37-55
55-74
74-90
total_acc_groups
0-20
20-40
40-60
60-80
80-100
revol_util_groups
2-10
10-19
19-27
27-36
36-44
open_acc_groups
5%-9%
9%-13%
13%-17%
17%-21%
21%-24%
int_rate_groups
funded_amnt_inv_groups


loan_amnt
funded_amnt_inv
term
int_rate
installment
emp_length
annual_inc
loan_status
dti
inq_last_6mths
open_acc
pub_rec
revol_util
total_acc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Heatmap
loan_amnt
funded_amnt_inv
term
int_rate
installment
emp_length
annual_inc
dti
inq_last_6mths
open_acc
pub_rec
revol_util
total_acc

- Data Analysis using Profiling report.html

Recommendations :

- The people who have taken the loan in the month of the December are more likely to be in the list of defaulters
- The people who have taken the loan in the year of 2011 are more likely to be the defaulters
- In the month of the December there are many festivals which happens in the US so people are more likely to get the loans and are being charged off
- The reason for the above case could be that the country could have the financial crisis
- The people who have the open accounts in the range of 2-10 are more likely to be the defaulters as the reason could be that they are unable to manage their finances Since they need to maintain minimum balance in each of the accounts which leads not able to repay the loan
- The people whose verification is not done and for whom the loan is approved without any background check are most likely to be defaulters so it is recommended not to approve the loan without any background check of the loan borrowers
- Try to avoid giving the loans to the people who belongs to grade B and D or do a good background check for them
- The people who take the loan for debt consolidation purpose that is to clear other loans are more likely to be in the list of defaulters. It is recommended to avoid giving the loans for debt consolidation purpose.