GRAMENER CASE STUDY

SUBMISSION

Group Name:Datascience Riders

1. Member name:Amit Kumar Singh
2. Member name: Karthik Bhat
3. Member name :Shijesh Velayudhan
4. Member name : Vikas Lalchandani

- Classification Line of Business.

- **XYZ is 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.
- Two **types of risks** are associated with the bank's decision:
- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- 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** for the company
- XYZ 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.

Note : Assumption we make the name of company as XYZ throughout the documents.

Business Objective

- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.
- If one is able 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.
- 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.

Aim of Case Study

The aim of team 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.

- We have work for the company which 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 utilize this knowledge for its portfolio and risk assessment.
- We have the dataset .It contains the complete loan data for all loans issued through the time period 2007 to 2011, and using this data is finally the goal and objective to provide the list of Risky applicant.

Data Analysis

Understanding the Requirement:

- The requirement is pretty simple to identify the best Applicant and more profit for company by maximize the best loan application and reduce the risk by identifying the risky applicant.

Understanding the Data:

- We have used the provided list of attributes ,features of the applicant for the period from 2007 to 2011 to analyses and provide the accurate predict model.

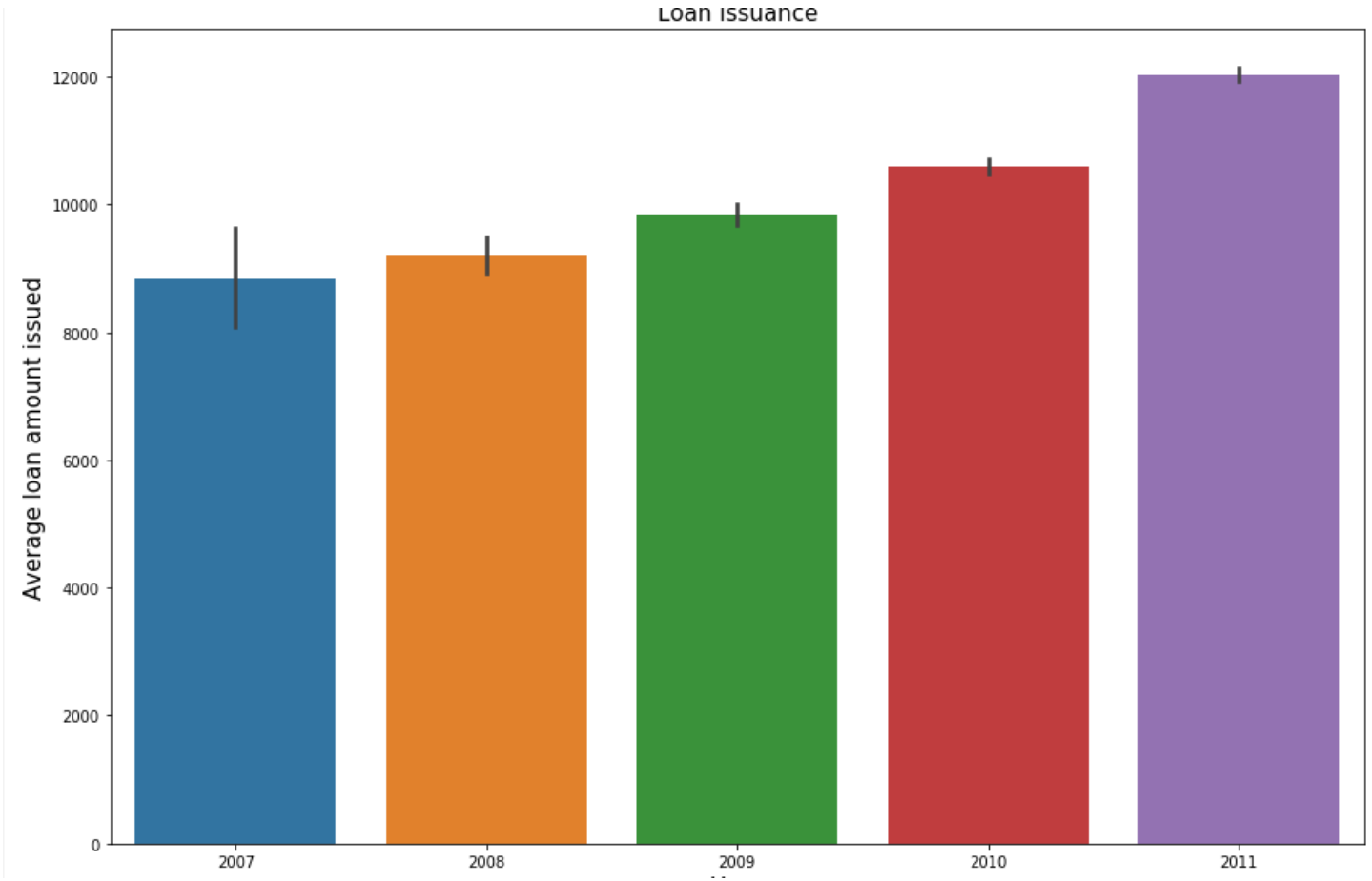
Risk

- Potential loss of revenue.

Tools used

- Python is used for data cleansing ,modelling ,analysis, sampling and many more.
- Tableau and Python is used for plotting the significant results.

Exploring Data and Analysis representation before performing any action to provide initial understanding



Approach and Methodology

- Logical understanding and derive new variables
- Cleaning Method of unnecessary and unwanted data or group of data including variate or higher variate variables value.
- Deriving the Clean Data Set,
- Data Analysis using the features available.
- Identifying the set of records for the business the records justifying the best policy for the company to identify the good applicant.

Logical understanding and derive new variables

All the logical and meaning of the data set is given in the excel sheet for which we used most as part our analysis is listed below. Some are listed below

bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
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.
id	A unique LC assigned ID for the loan listing.
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)

Cleaning Method

- There are many irreverent data in data set we have removed from the set
- There are many columns which are not required for Data analysis which we removed
- There are many columns name whose column name is not properly given we have renamed
- There are many column having irreverent values or null value which we have removed
- More than 50% of columns have been removed
- Cleaning method involves the removing the outliers when not required
- Data set have information and required to be part of analysis has been filtered for further analysis.

Data Analysis Approach

- Identifying the Univariate Variables.
- Identifying the Bivariate and combination of Variable for the Analysis.
- Exploring the data arrived after the Univariate and Bivariate Analysis.
- Using Charts,Group,Functions etc to derive important aspects and supports the analysis.
- Deriving the Best case study which provides the approach a solution to the bank as a analysis tool to get the reasons for the charged-off, most profit strategy,Region,Group etc

Data Representation – Experience Category of Loan Status

- **Number of Most Loan Taker : Junior :**

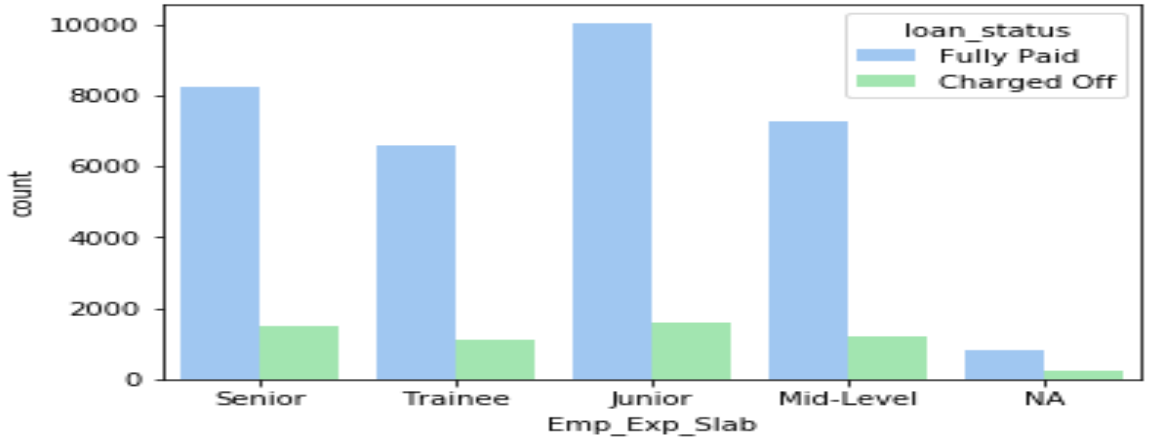
  Defaulter in Junior is 13 %

- **Number of Most Paid of : Senior**

  Defaulter in Senior is 18%

```
Emp_Exp_Slab   loan_status
Junior         Charged Off      1581
               Fully Paid      10048
Mid-Level      Charged Off      1227
               Fully Paid       7269
NA             Charged Off       227
               Fully Paid        803
Senior         Charged Off      1488
               Fully Paid       8225
Trainee        Charged Off      1088
               Fully Paid       6571
Name: id, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x1984acc0>
```



Data Representation – Interest Rate Category of Loan Status

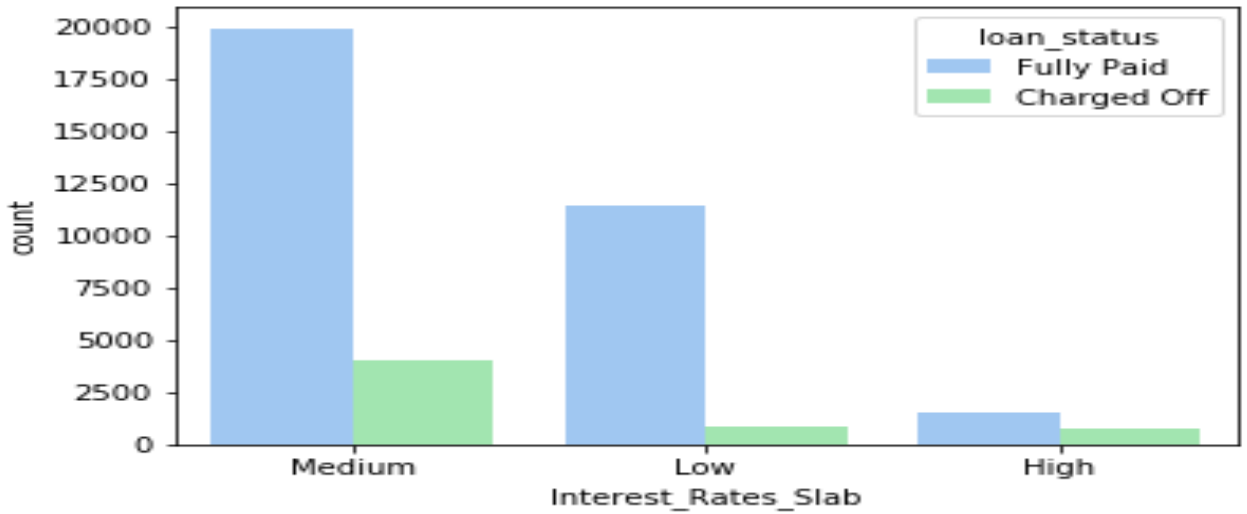
- % of charged-of In High I.R% -32%
- % of charged-of in Mid I.R% - 16%
- % Charged-of in Low I.R% - 06%

Remarks : High interest rate

Yields more charged-off

```
Interest_Rates_Slab  loan_status
High                 Charged Off       731
                     Fully Paid       1502
Low                  Charged Off       830
                     Fully Paid      11486
Medium               Charged Off      4050
                     Fully Paid      19928
Name: id, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x198af7b8>
```



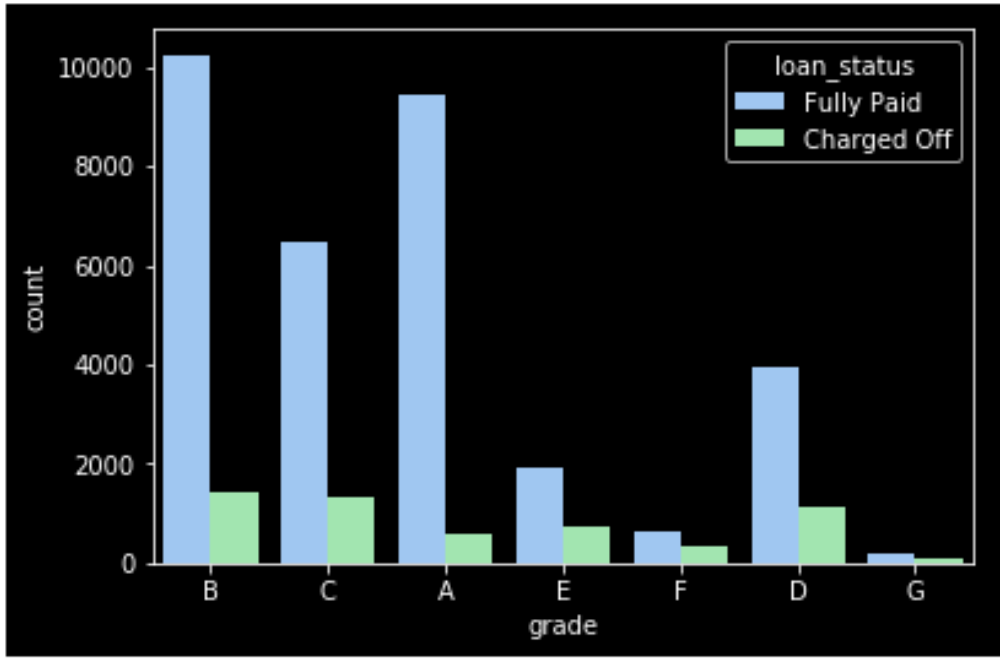
Data Representation – Grade Category of Loan Status

Charged off % grade wise:

B	25.396543
C	23.935127
D	19.871681
E	12.671538
A	10.728925
F	5.631795
G	1.764391

Fully Paid % grade wise:

B	31.136833
A	28.688176
C	19.677361
D	12.012395
E	5.896828
F	1.986876
G	0.601531



Remarks :

We can clearly see that percentage of Charged off are greater than Fully paid in below Grades:

C,D,E,F,G

While percentage of Fully paid is greater in:

A and B

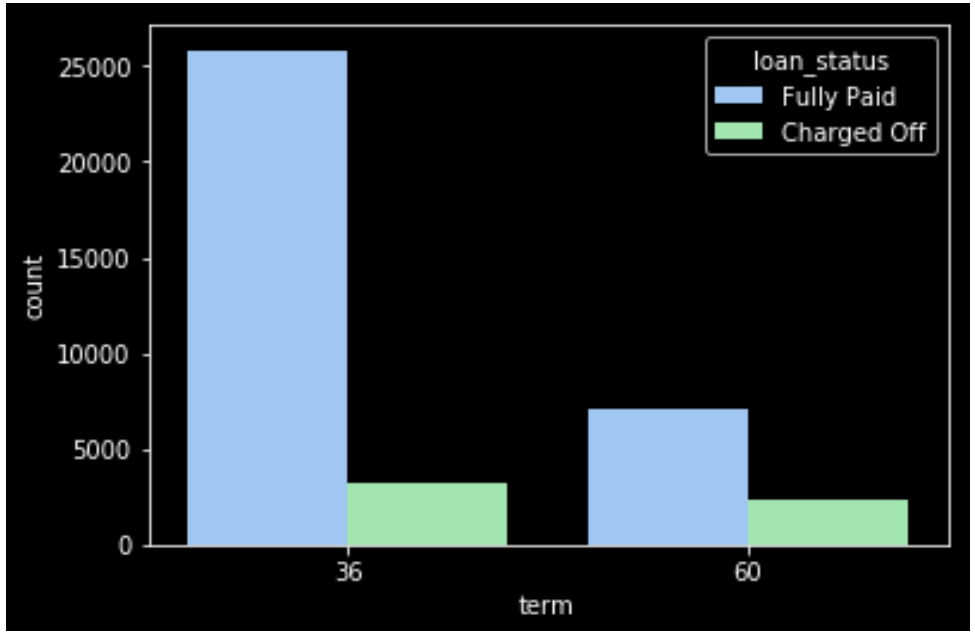
Data Representation – Term Category of Loan Status

Charged off % term wise:

36	57.280342
60	42.719658

Fully Paid % grade wise:

36	78.487666
60	21.512334



Remarks :

Higher % of charged off in 60 Month term and

Higher % of Fully paid in 36 Month Term

Data Representation – Verification Status Category of Loan Status:

Charged off % term wise:

Not Verified	38.050258
Verified	36.446266
Source Verified	25.503475

Fully Paid % grade wise:

Not Verified	44.142666
Verified	30.839106
Source Verified	25.018228

Remarks : Charged off % is higher even for verified customers.

Data Representation – Delinquency Age Category of Loan Status

Charged off % Age wise:

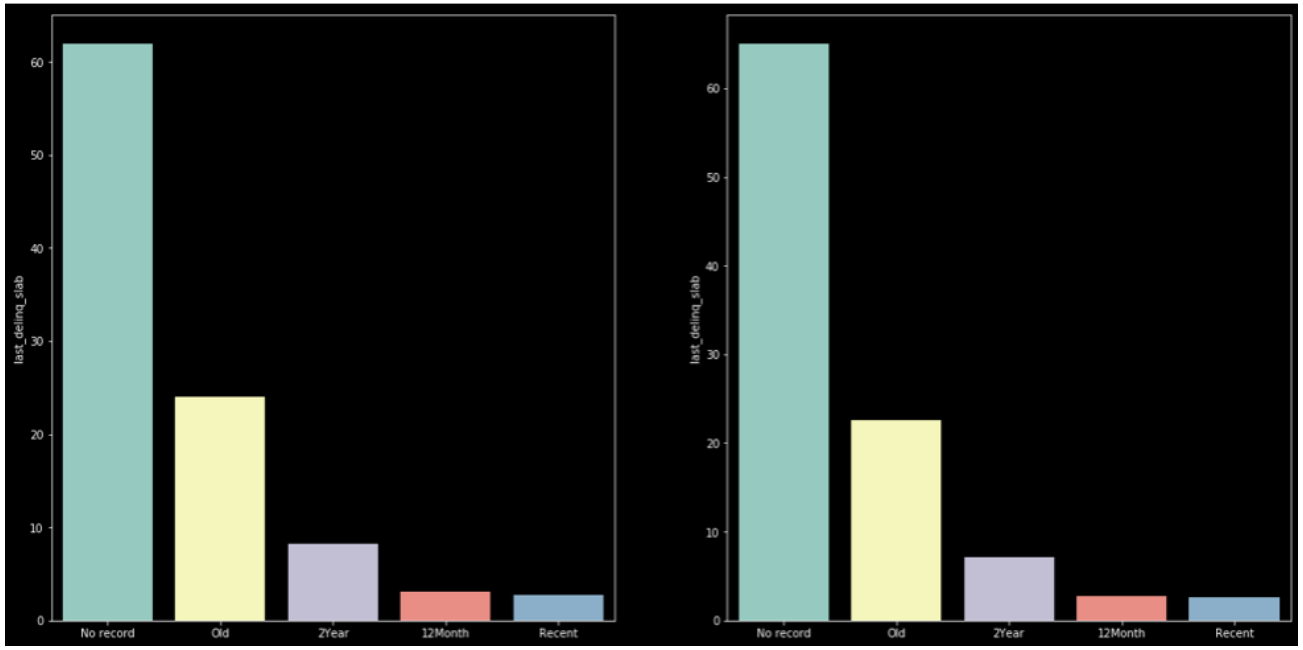
No record	61.914097
Old	23.970772
2Year	8.216004
12Month	3.118874
Recent	2.780253

Fully Paid % Age wise:

No record	64.995747
Old	22.627294
2Year	7.096853
12Month	2.694738
Recent	2.585369

Remarks :

Where there is no delinquency recorded,
% of fully paid customers is higher than charged off.



Data Representation – Revolving line utilization rate Category of Loan Status

Charged off % Rate wise:

High	29.691677
Very High	29.656033
Medium	23.632151
Low	17.020139

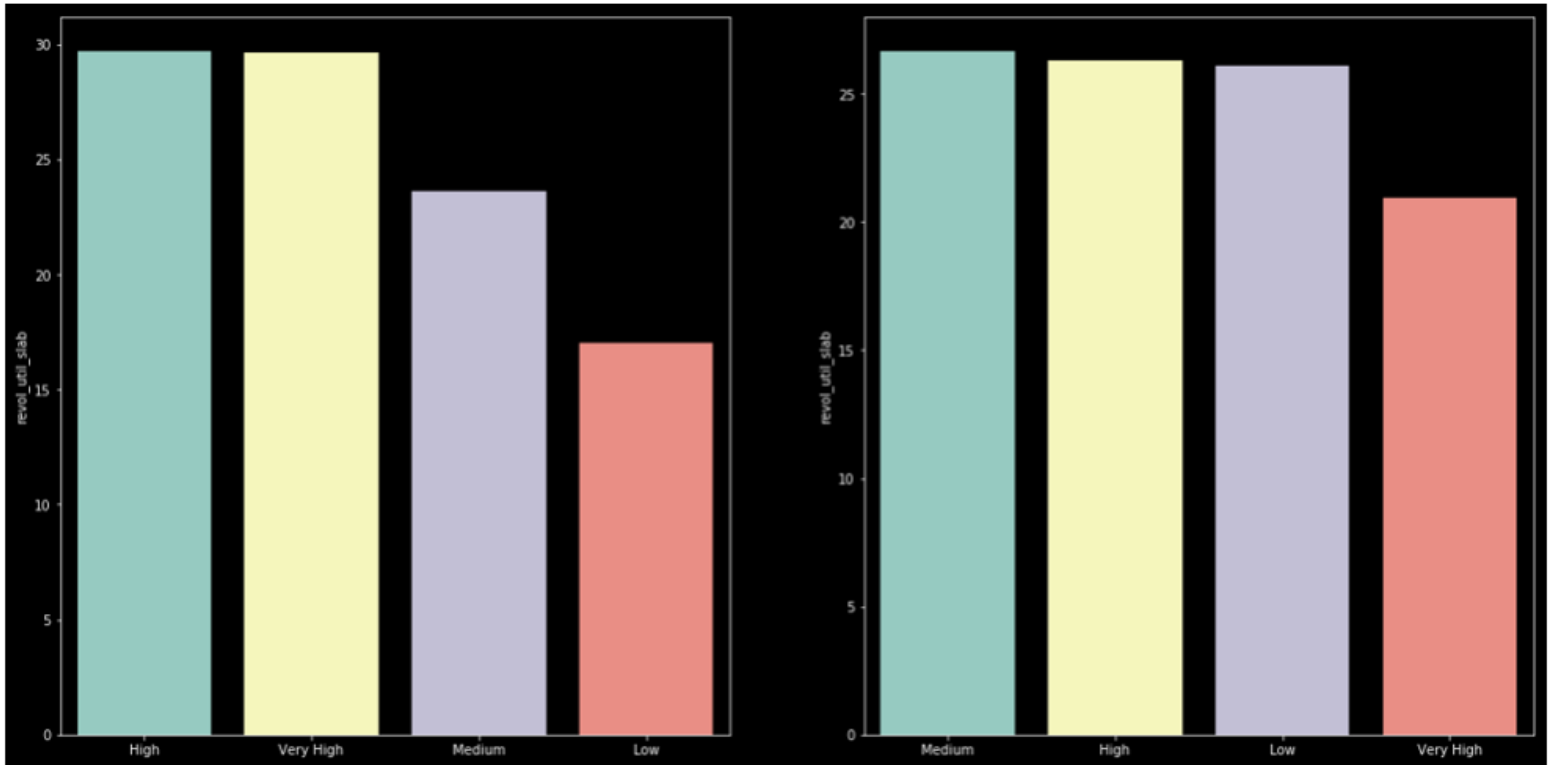
Fully Paid % Rate wise:

Medium	26.655730
High	26.291165
Low	26.096731
Very High	20.956374

Remarks :

Charged Off % is higher for:
Very High and High Utilisation rate.

Fully Paid % is higher for:
Low and Medium Utilisation rate.



Data Representation – number of inquiries in past 6 months Rate Category of Loan Status

Charged off % Rate wise:

Low	87.078952
Medium	11.905186
High	0.784174
Very High	0.231688

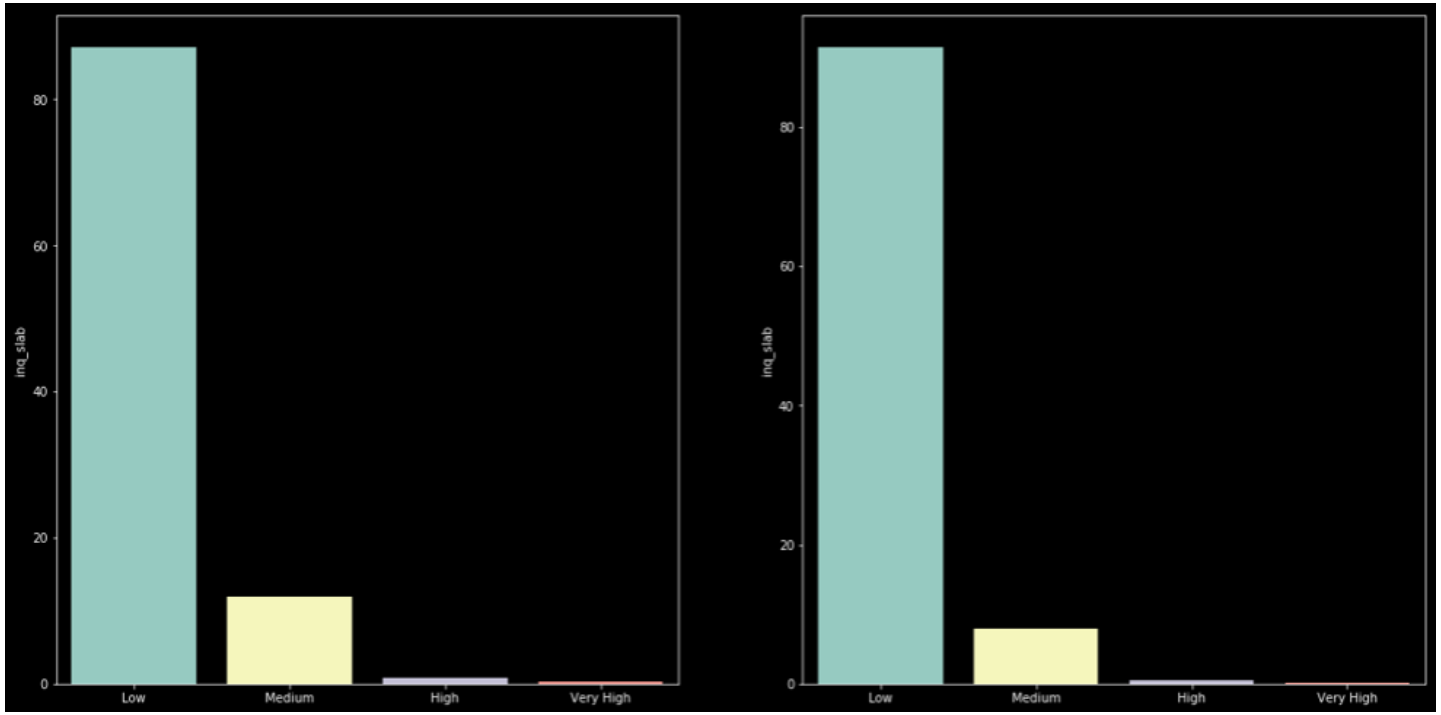
Fully Paid % Rate wise:

Low	91.414510
Medium	7.983959
High	0.495200
Very High	0.106331

Remarks :

Charged Off % is higher for:
Medium, High and Very High enquiries.

Fully Paid % is higher for:
Low enquiries.



Facts for few outliers

- Yeah, that's true and the reason why this is happening is that the outliers are not well defined in the int_rate case. What we mean to say is that the values are pretty close and large discrepancies aren't easily observed. So generally in such cases, we don't go for outlier analysis. A better idea would be to categorize them into specific buckets (like low, high, etc.)and continue your analysis. For the annual income case, yes you can go for an outlier analysis but again categorization is a much better idea.

Relevant Observations

- A credit line is a pool of money available for borrowing. Also known as a line of credit, these loans have a maximum limit, and borrowers have the option of borrowing any amount up to that limit
- In simplest terms, it is an arrangement between a financial institution, usually a bank, and a customer, that established the maximum amount of a loan that the customer can borrow.
- While doing univariate on loan term of charged off records found 60 months above as majority defaulted compared to 36 months.
- There is a high chance of Charge Off for High Enquiries, Revolving utilization and Interest rates.
- Bad Grades and 60 Month terms are also highlight Charge offs.
- If there are previous delinquencies Charge off can be expected.

Assumptions

Lending Club or any peer to peer company determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees. Grades are like ratings for the borrowers. Any borrower with Rating A is highly likely to get a loan as compared to a borrower with Grade D . Grades are based assigned to borrowers on various metrics like their Credit History ., monthly income etc