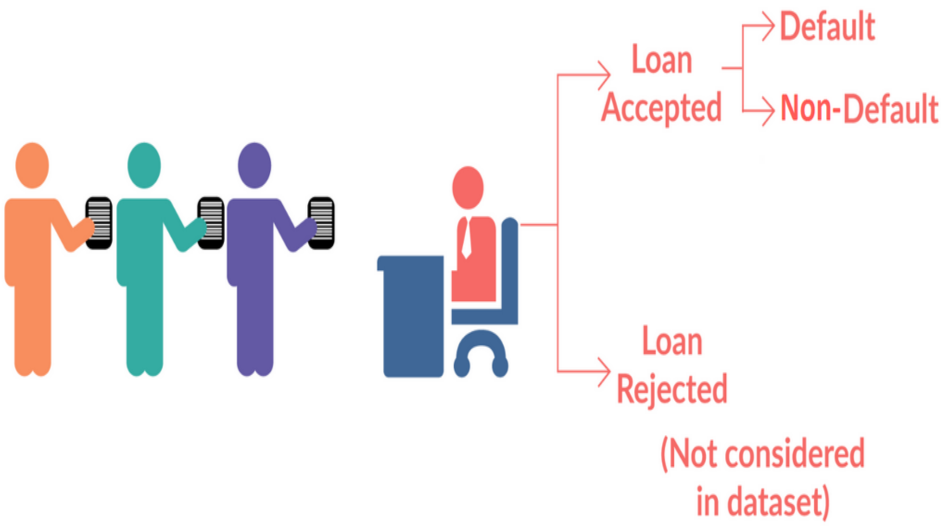
LOAN CASE STUDY

SUBMISSION

Team:
Abhinaba Chakraborty
Arunima Dasgupta
Koustav Bose
Lakshmi J

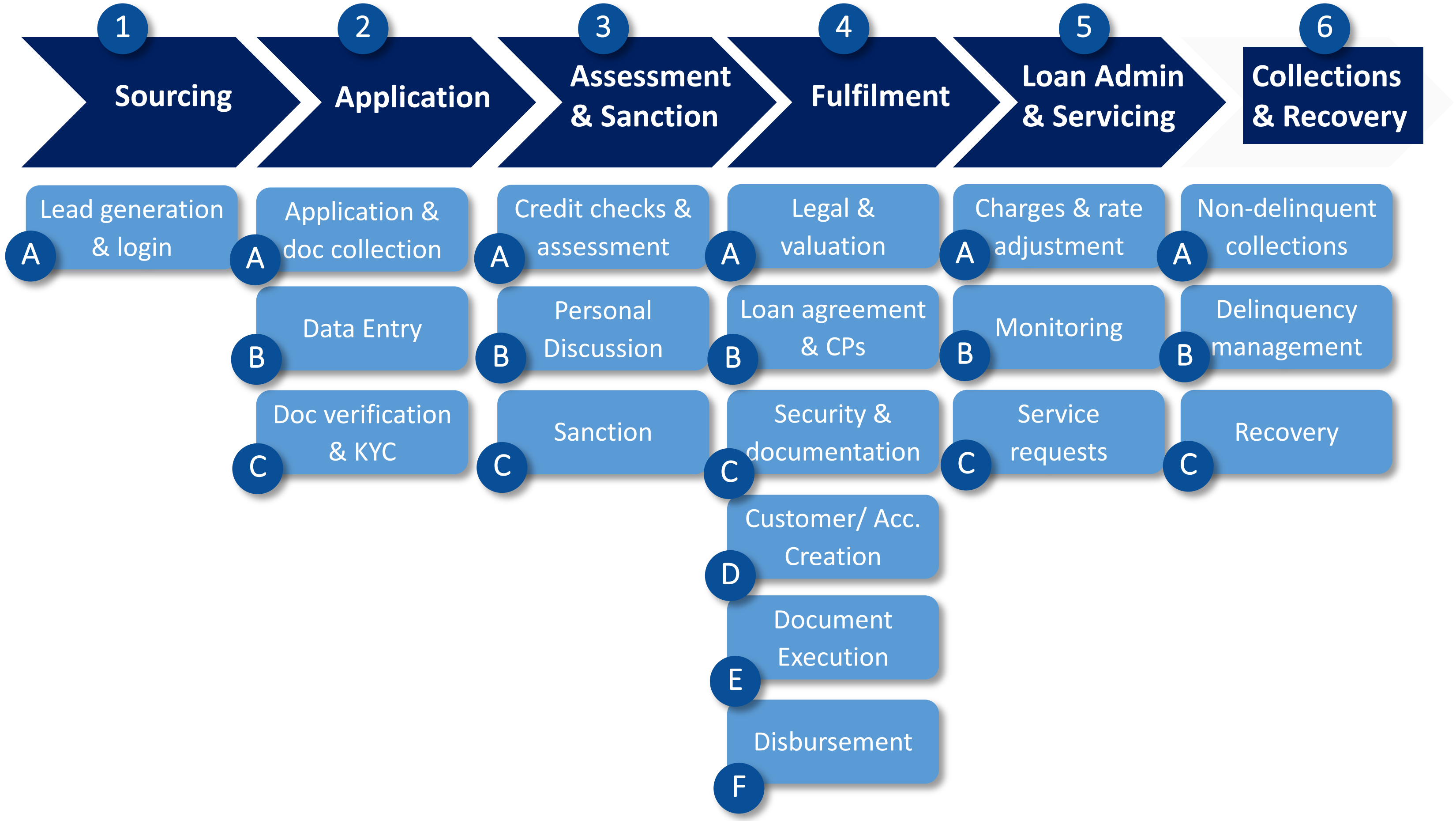
Business Objectives



- This consumer finance 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. '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.
- The company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

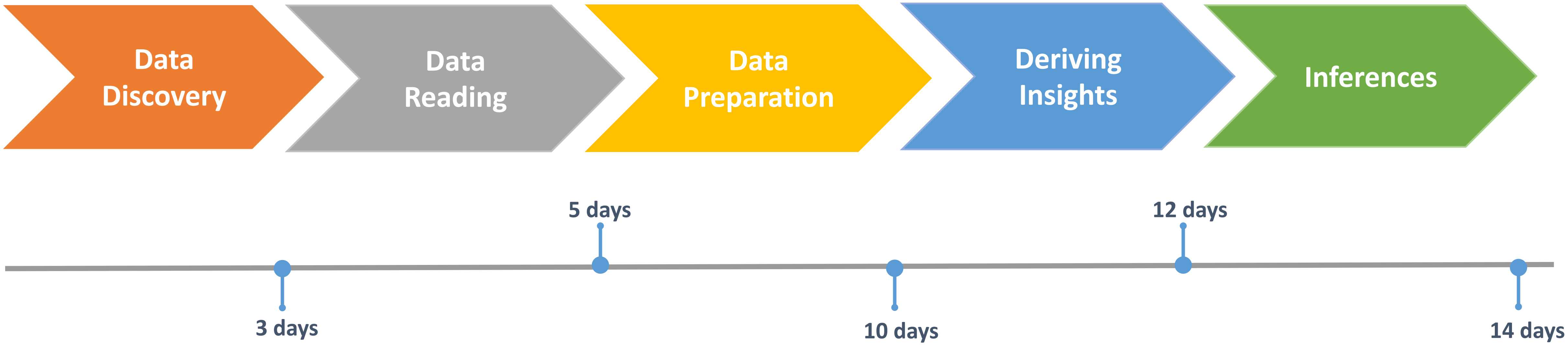
Process/ Operating Model

The following diagram shows a typical flow for an end to end loan management process



We will be primarily focusing on the stage **3A** of the given Process/ Operating model in this exercise

Problem solving methodology (Stage wise plan for 2 weeks)



- Understanding of business objective and current context
- Assess the resources in terms of technology and data
- Develop initial hypotheses
- Identify Data Sources
- Convert business context into an Analytics problem

- Reading/ importing the data
- Data Quality Checks
- Data Understanding

- Data Exploration and Variable Selection
- Removal of insignificant variable
- Removing & Imputing missing values
- Outlier removal
- Date format conversion
- Variable type casting
- Removal of invalid values

- Univariate analysis
- Segmented univariate analysis
- Bivariate analysis
- Feature Engineering or deriving new metrics out of the existing ones

- Analyze inferences from the model
- Create customized visualizations
- Communicate the results to stake holders

Data Cleaning

UpGrad

- **Removal of insignificant variables**

The data set consisted of seven variables which do not add any significant value for analysis. These variables are : **id, member_id, emp_title, url, desc, title, zip_code.** The justification for their removal is given in the jupyter notebook

- **Dropping the variables having single unique value/ single category**

Variables →

```
pymnt_plan
initial_list_status
collections_12_mths_ex_med
policy_code
application_type
acc_now_delinq
chargeoff_within_12_mths
delinq_amnt
tax_liens
```

- **Missing value treatment**

1. Variables having more than 50 % of missing values has been dropped from the data set, because imputing such a large number of missing values will lead to biasness in the data set.
2. Few variables with missing value has been imputed or dropped as necessary

Variable Name	Missing value %
emp_length	2.706650
revol_util	0.125891
last_pymnt_d	0.178765
last_credit_pull_d	0.005036
collections_12_mths_ex_med	0.140998
chargeoff_within_12_mths	0.140998
pub_rec_bankruptcies	1.754916
tax_liens	0.098195

For the date variable, missing values has been imputed with the previous date value

- **Outlier detection**

There are many numeric variable where outlier has been detected (very minimal percentage), example : Total income , revolving balance etc. Since removal of outlier is a subjective call , so based on the business need outlier treatment has been done.

- **Variables type casting (converting the variable type)**

➢ There are variables which are stored as a object while importing the data into python . These object type variable has been changed as category type for the ease of further manipulation .

Example : term, sub_grade, emp_length, home_ownership, verification_status, loan_status, purpose, addr_state

➢ There are few variables which are been converted into category type from the integer type , since they have less number of distinct values & for the specific business sense it is required .

Example : delinq_2yrs, inq_last_6mths etc.

➢ Variables which has % sign at end has been stored as object as default . These variables has been converted into float after stripping the % sign.

Example : revol_util , int_rate

➢ Variables which are having date has been formatted properly

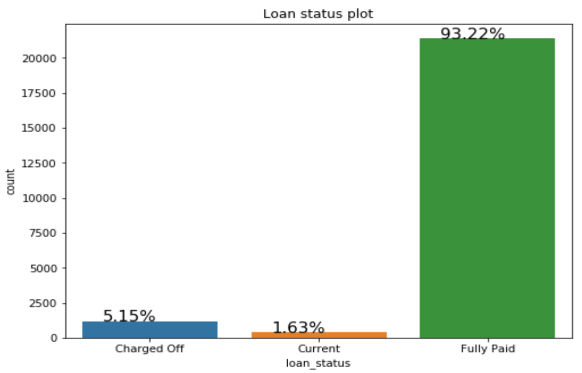
Example : issue_d,earliest_cr_line,last_pymnt_d,last_credit_pull_d

Insights

- **Loan status**

➢ From the adjacent graph, it clearly shows that around 5% of the loans issued by the company is being defaulted.

	rank	frequency
Fully Paid	1	21375
Charged Off	2	1181
Current	3	374



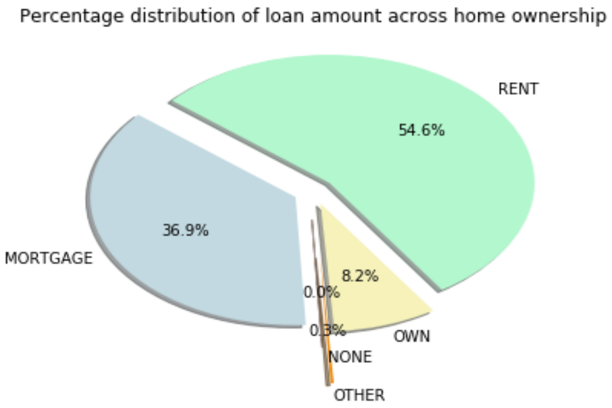
- **Term wise analysis**

➢ Long term loans are riskier as they have higher percentage of charge offs while most loans are usually short term and contribute to a large number of charge offs.

loan_status	Charged Off	Current	Fully Paid	Total	Charged Off %
term					
60 months	412	374	2873	3659	11.26
36 months	769	0	18502	19271	3.99

- **Home ownership wise analysis**

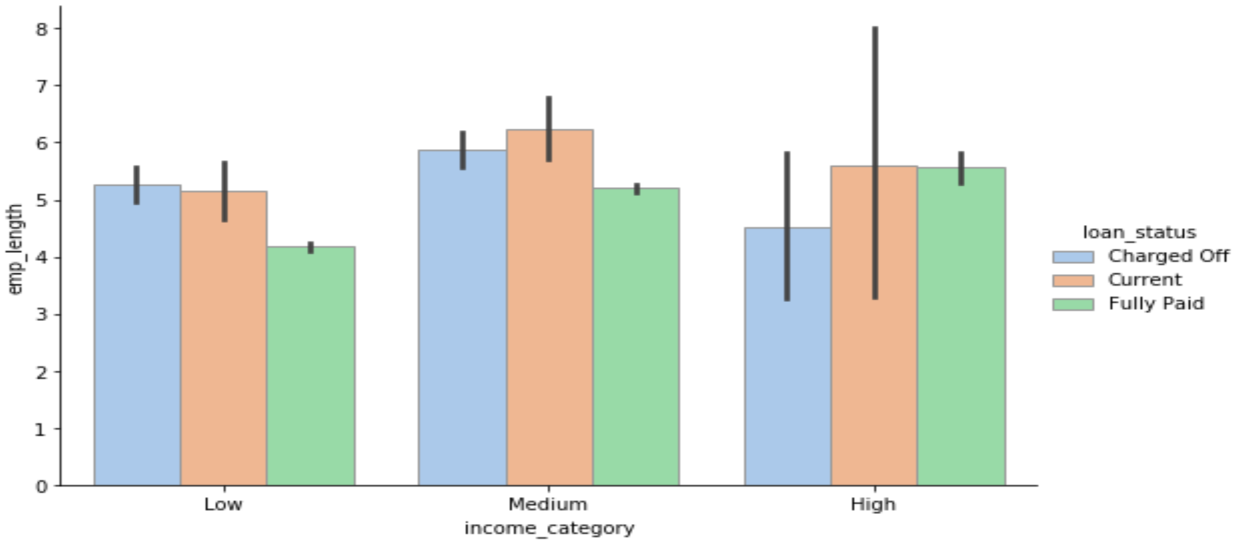
➢ From the chart we can see that most people who are charged off stay in rented property or in mortgage. So the financial company needs to be cautious while giving the loans to such people where there is no collateral for recovery.



Insights

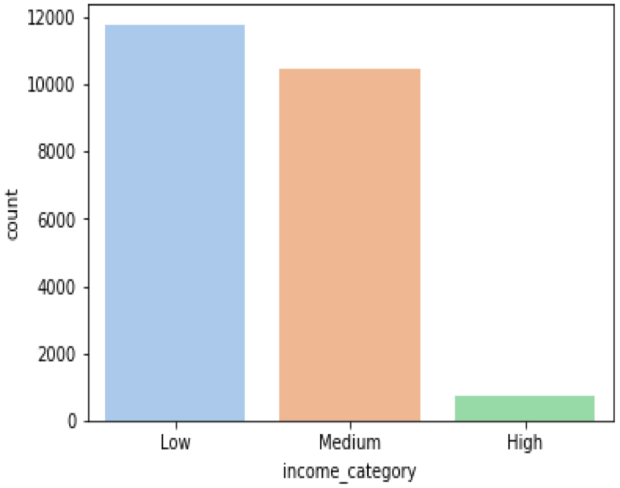
- **Employee length and Annual income analysis**

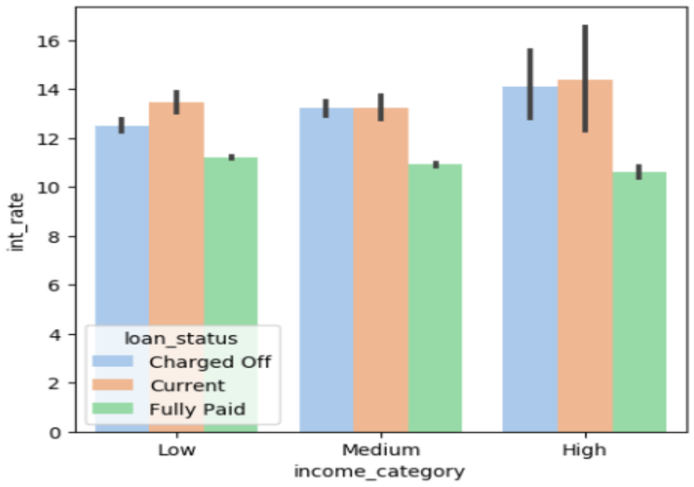
  - It is assumed that as the employment length increase, the income gets higher and so is capability for loan closure is also high.
  - Thus, borrowers having lesser experience, have less access to money and tend to default more often.
  - The adjacent graph shows a similar trend as expected.



- **Annual income and interest rate analysis**

  - From the adjacent table, it clearly shows that low income category people are the one of the main driving factor behind the loss of the company due to charge off loans .
  - In terms of interest rate, the lower income category are charged similar interest rates as medium and higher income groups which could be another reason for them to tend to default.

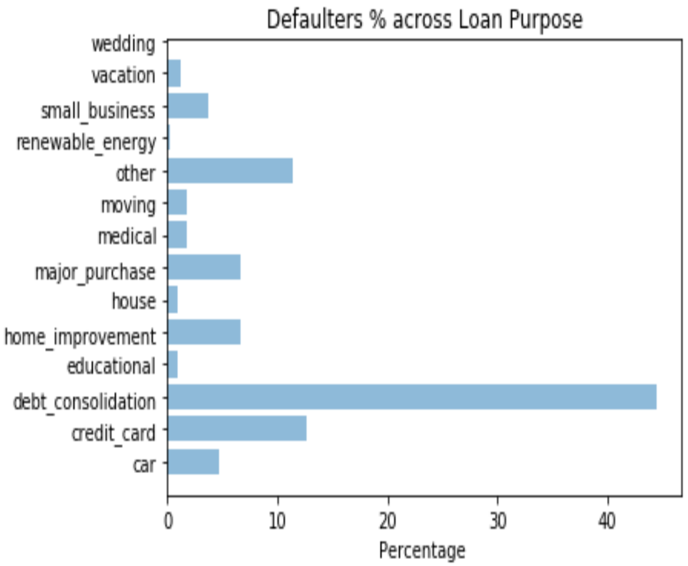


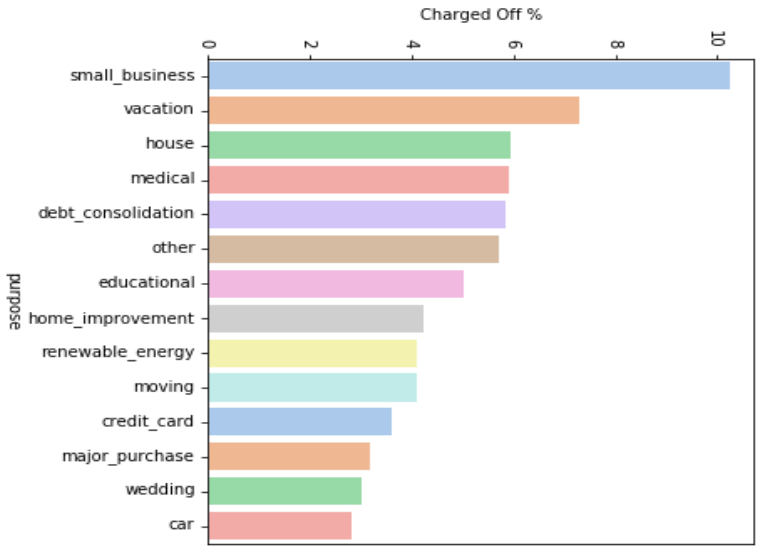


Insights

- **Loan purpose analysis**

➢ Even though we see a large number of borrowers requesting loans for debt consolidation, we see an interesting trend where people taking loans for purpose such as small business, vacation, house are riskier in terms of charge off tendency.





.

- **Verification analysis**

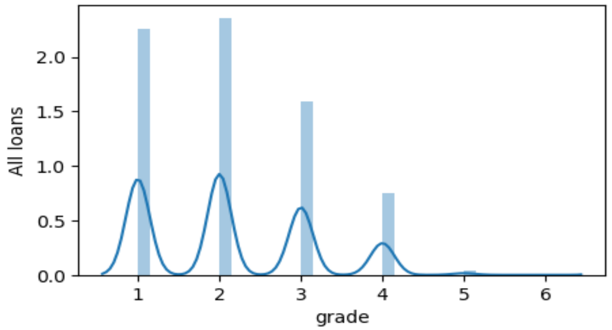
➢ The lending company processes loans after income and source of income verification for most of the loans. Yet there are a large number of charge off cases in spite of verification. This process needs to be looked into by the lending company to ensure proper identification of potential defaulters.

loan_status verification_status	Charged Off	Current	Fully Paid	Total	Charged Off %
Verified	357	99	4727	5183	6.89
Source Verified	284	136	5601	6021	4.72
Not Verified	540	139	11047	11726	4.61

- **Grade analysis**

.

➢ Grade B(2 in graph), followed by A(1 in graph) & C (3 in graph) lead to larger number of charge off whereas the higher loan grades (E(5), F(6), G(7)) are riskier as their chances of being charged off are higher.

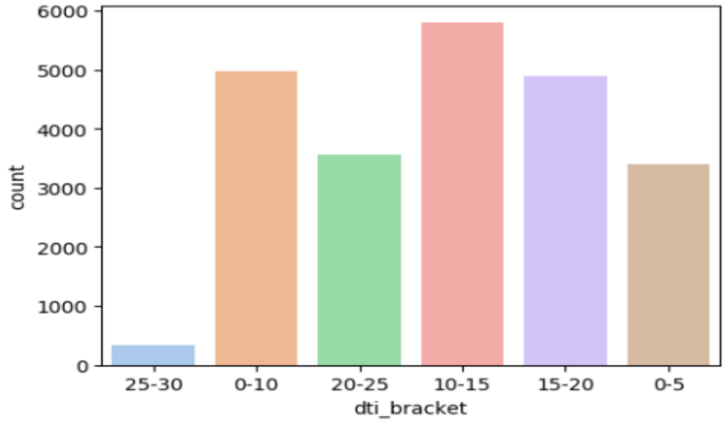


loan_status grade	Charged Off	Current	Fully Paid	Total	Charged Off %
7.0	10	1	31	42	23.81
6.0	40	12	163	215	18.60
5.0	108	25	761	894	12.08
4.0	181	57	2316	2554	7.09
3.0	303	93	4200	4596	6.59
2.0	369	158	6966	7493	4.92
1.0	170	28	6938	7136	2.38

Insights

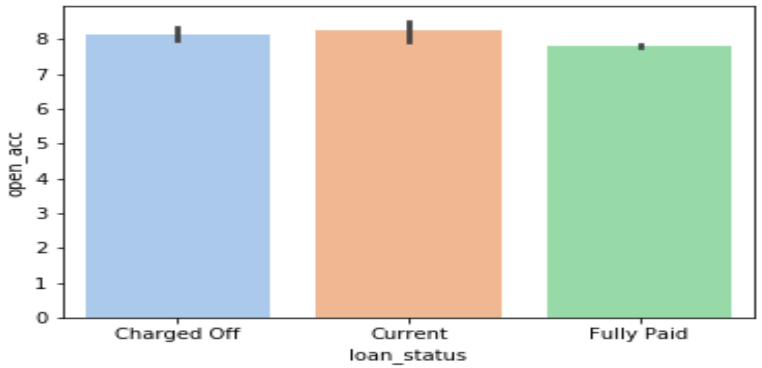
- **Debt to income ratio wise analysis**

➢ It is seen that the borrower's total monthly debt payments on the total debt obligations he owes (debt to income ratio) shows a larger tendency of default in the range of 10-15 among other ranges.



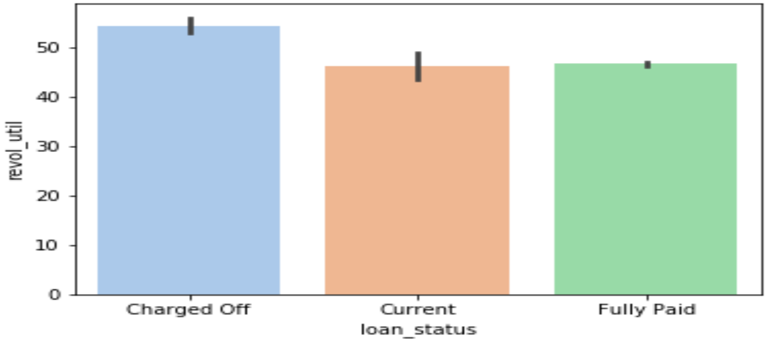
- **Open credit lines analysis**

➢ It is seen that the open credit lines are same across the types of loan status which indicates even the low income people also have similar credit lines to that of the medium and higher income people.



- **Revolving line utilization rate**

➢ It is seen that the revolving line utilization rate are greater than 50 among lower income people (less experienced) than higher income people (more experienced). Hence, causing tendency to default.



LOAN ATTRIBUTES	CONSUMER ATTRIBUTES
Loan request purpose (purpose)	Self reported annual income (annual_inc)
LC assigned loan grade(grade)	Employment length in years (emp_length)
Number of payments on the loan (term)	Total monthly debt payments on total debt obligations (dti)
Income/Source of income verified by LC (verification_status)	Revolving line utilization rate (revol_util)
Interest rate on the loan (int_rate)	Number of open credit lines (open_acc)
	Home ownership status (home_ownership)

Conclusion

- Most of the loads that have been charged off are the ones belonging to borrowers from low income brackets and having considerably less average employment lengths (~5 years).
- These borrowers are issued loans with high interest rates.
- Their debt to income ratio is high which means that they are already in debt and have comparable number of credit lines as their higher income borrower counterparts.
- On top of it, these borrowers have a higher revolving credit line utilization which means they are mostly living on credit.
- These borrowers have home ownership status of mortgaged or rent which further increases the risk of charge off.
- The verification process seems to have no effect on filtering out these kinds of borrowers and needs to be looked into by LC.
- Also, the loans which have a term of 60 seem to be a risker proposition and needs stringent verification.