# Lending Club Case Study: Assessment-1

Prepared by : Joshy PJ & Sheetal R

# Agenda

- Business Problem and their objectives
- Analysis Overview
- Key Observations

# Lending Club Case Study

## Brief about the Company

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.


## Business Problem

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

## Objectives

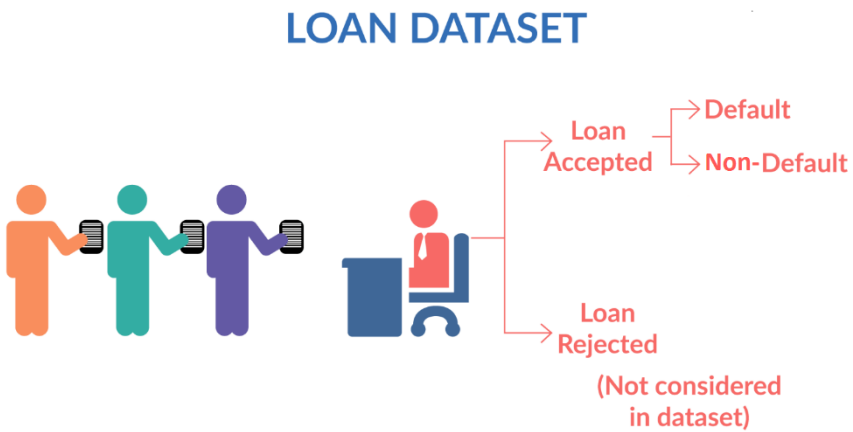
Company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

# Lending Club Case Study

upGrad

Data flow –

- How Lending Club process their client request
- How they define the applicant is default or not




**Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)

**Current:** Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

**Charged-off:** Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

# Lending Club Case Study

upGrad

## Data Provided

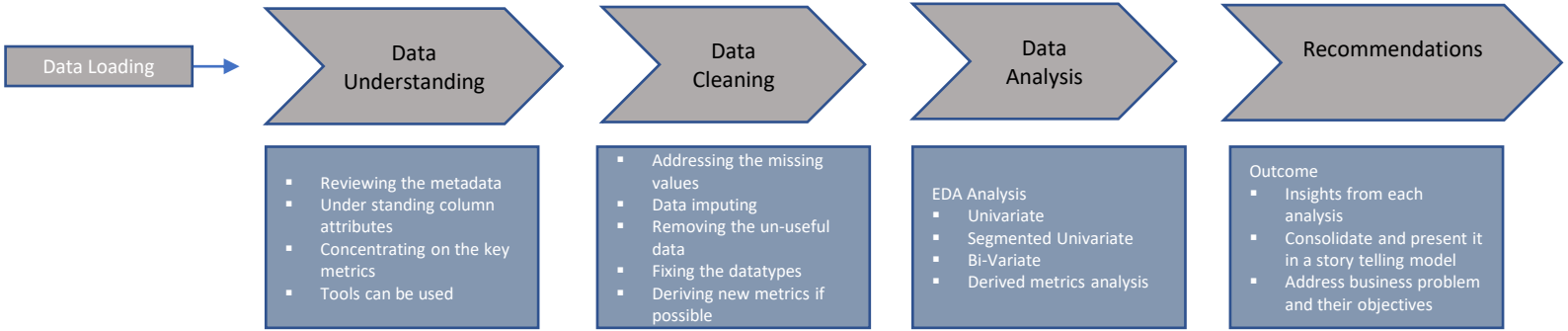
- Lending Club Provided the historical data which contains
  - Applicant demographic characteristics
  - Applicant behavior characteristics
  - Loan characteristics
  - Also provided the metadata around the attributes

## Assessment Methodology

- Based on the data , problem statement and their objectives Exploratory Data Analysis is the suitable approach to get insights from the available data
- Following session will be talking more about the phases in Exploratory data analysis methodology

## Exploratory Data Analysis

1. Data understanding
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis
4. Recommendations



# Lending Club case study - Exploratory Data Analysis

upGrad

## Data Understanding

- Provided data in the csv format
- It has 39717 rows and 11 columns
- Columns contains data around loan characteristics , applicant demographic and behavior characteristics
- Loan_status column talks about - is it ‘Fully Paid’ ,’Charged Off’ or ‘Current’
- Loan_status = ‘Current’ meaning , currently running loans – thus not considered for this case study
- Most of the columns are showing data types as objects- need to address this based on the data
- There are columns will ‘null values’ – this can be addressed based on the null value percentage

# Lending Club case study - Exploratory Data Analysis

upGrad

## Data Cleaning

- Looking for the null values and its percentage
  - There are 57columns with null value percentage more than 65
- Methodology for removing those null values
  - Having these columns will not make any sense ;thus, dropping these based on the null value percentage threshold of above 50.
  - 57 columns were dropped from the analysis
  - There are columns with customer specific behavior data , this data may not help for the new applicant; thus, removed those
- Imputing those null values based on the available and business problems
  - Now there are only 7 columns with missing values
    - Missingno library used for anlaysing the missing data
    - desc - talks about loan description – data is in long string format- will not make any sense
    - title, emp_length.emp_title - are important , thus removing the missing rows
    - pub_rec_bankruptcies,chargeoff_within_12_mths,collections_12_mths_ex_med,tax_liens – are with very fewer missing values – thus marking it as 'NAN'
- Formatting the column datatypes based on the data and its meaning
  - All the datatypes are showing as object
  - Based on its data types and metadata information , necessary datatypes applied
  - Derived Month and Year from the Issue_d column
- Filtering the data
  - Based on the loan_status – Current applicant can be removed from the assessment
  - Finding the outliers in each columns and remove those extreme outliers
    - Eg: annual_inc greater than 230000 can be removed based on its 99% percentile value and max gap.

## Missing value percentage overview

| Column Name | Null Value Percentage |
|---|---|
| verification_status_joint | 100.0 |
| annual_inc_joint | 100.0 |
| mo_sin_old_rev_tl_op | 100.0 |
| mo_sin_old_il_acct | 100.0 |
| bc_util | 100.0 |
| bc_open_to_buy | 100.0 |
| avg_cur_bal | 100.0 |
| acc_open_past_24mths | 100.0 |
| inq_last_12m | 100.0 |
| total_cu_tl | 100.0 |
| inq_fi | 100.0 |
| total_rev_hi_lim | 100.0 |
| all_util | 100.0 |
| max_bal_bc | 100.0 |
| open_rv_24m | 100.0 |
| open_rv_12m | 100.0 |
| il_util | 100.0 |
| total_bal_il | 100.0 |
| mths_since_rcnt_il | 100.0 |
| open_il_24m | 100.0 |
| open_il_12m | 100.0 |
| open_il_6m | 100.0 |
| open_acc_6m | 100.0 |
| tot_cur_bal | 100.0 |
| tot_coll_amt | 100.0 |

| Column Name | Null Value Percentage |
|---|---|
| mo_sin_rcnt_rev_tl_op | 100.0 |
| mo_sin_rcnt_tl | 100.0 |
| mort_acc | 100.0 |
| num_rev_tl_bal_gt_0 | 100.0 |
| total_bc_limit | 100.0 |
| total_bal_ex_mort | 100.0 |
| tot_hi_cred_lim | 100.0 |
| percent_bc_gt_75 | 100.0 |
| pct_tl_nvr_dlq | 100.0 |
| num_tl_op_past_12m | 100.0 |
| num_tl_90g_dpd_24m | 100.0 |
| num_tl_30dpd | 100.0 |
| num_tl_120dpd_2m | 100.0 |
| num_sats | 100.0 |
| num_rev_accts | 100.0 |
| mths_since_recent_bc | 100.0 |
| num_op_rev_tl | 100.0 |
| num_il_tl | 100.0 |
| num_bc_tl | 100.0 |
| num_bc_sats | 100.0 |
| num_actv_rev_tl | 100.0 |
| num_actv_bc_tl | 100.0 |
| num_accts_ever_120_pd | 100.0 |
| mths_since_recent_revol_delinq | 100.0 |
| mths_since_recent_inq | 100.0 |

| Column Name | Null Value Percentage |
|---|---|
| dti_joint | 100.0 |
| total_il_high_credit_limit | 100.0 |
| mths_since_last_major_derog | 100.0 |
| next_pymnt_d | 97.0 |
| mths_since_last_record | 93.0 |
| mths_since_last_delinq | 65.0 |

# Lending Club case study - Exploratory Data Analysis

upGrad

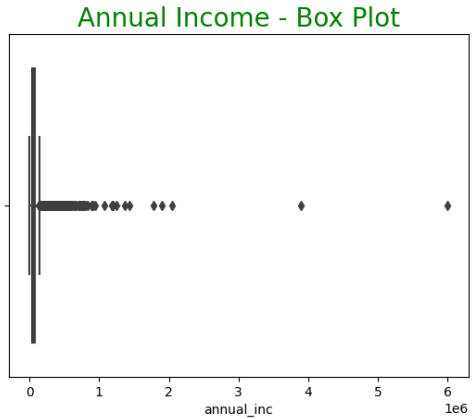
Final columns and its Significance

| Column Name | Significance |
|---|---|
| Id | Applicant ID |
| loan_amnt | Requested Amount |
| funded_amnt | Approved amount by Lending Club |
| funded_amnt_inv | Approved amount by Investor |
| Term | Tenure of the loan |
| int_rate | Interest Rate |
| Grade | Applicant Grade , based on the profile decided by Lending Club |
| sub_grade | Grade break-up or subcategories |
| emp_title | Applicant's title , based on their jobs |
| emp_length | Applicant's work experience |
| home_ownership | Applicant Home Ownership |
| annual_inc | Applicant Home |
| verification_status | Applicant background verification Status |
| loan_status | Charged Off or Fully Paid or Current |
| purpose | Loan taking purpose |
| addr_state | Applicant Location - State |
| dti | Debit to income ratio |
| pub_rec_bankruptcies | whether applicant is brakrupted or not - how many times |
| issue_d_month | Derived columns from Issue_d |
| issue_d_year | Derived columns from Issue_d |
| loan_outcome | 0 if Fully Paid , 1 if Charged Off - Derived column from loan_status |

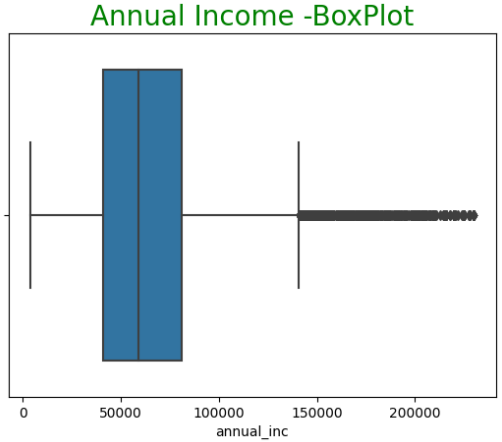
Final column set after cleaning and changing its attributes

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 37202 non-null | object |
| 1 | member_id | 37202 non-null | object |
| 2 | loan_amnt | 37202 non-null | int64 |
| 3 | funded_amnt | 37202 non-null | int64 |
| 4 | funded_amnt_inv | 37202 non-null | float64 |
| 5 | term | 37202 non-null | object |
| 6 | int_rate | 37202 non-null | float64 |
| 7 | installment | 37202 non-null | float64 |
| 8 | grade | 37202 non-null | object |
| 9 | sub_grade | 37202 non-null | object |
| 10 | emp_title | 37202 non-null | object |
| 11 | emp_length | 37202 non-null | object |
| 12 | home_ownership | 37202 non-null | object |
| 13 | annual_inc | 37202 non-null | float64 |
| 14 | verification_status | 37202 non-null | object |
| 15 | issue_d | 37202 non-null | object |
| 16 | loan_status | 37202 non-null | object |
| 17 | pymnt_plan | 37202 non-null | object |
| 18 | url | 37202 non-null | object |
| 19 | purpose | 37202 non-null | object |
| 20 | title | 37193 non-null | object |
| 21 | zip_code | 37202 non-null | object |
| 22 | addr_state | 37202 non-null | object |
| 23 | dti | 37202 non-null | float64 |
| 24 | initial_list_status | 37202 non-null | object |
| 25 | collections_12_mths_ex_med | 37155 non-null | float64 |
| 26 | policy_code | 37202 non-null | object |
| 27 | acc_now_delinq | 37202 non-null | object |
| 28 | chargeoff_within_12_mths | 37202 non-null | object |
| 29 | delinq_amnt | 37202 non-null | int64 |
| 30 | pub_rec_bankruptcies | 37202 non-null | float64 |
| 31 | tax_liens | 37170 non-null | float64 |
| 32 | issue_d_year | 37202 non-null | int64 |
| 33 | issue_d_month | 37202 non-null | int64 |

## Addressing the outliers








Percentiles value for Annual Incomes

80%, annual income = 90000.0
90%, annual income = 115000.0
95%, annual income = 140000.0
99%, annual income = 230000.0
100%, annual income = 6000000.0

There is a huge gap between , max and 99% of Annual Income – so annual income > 230000 dropped from the dataset

## Data Analysis

Exploratory Data Analysis (EDA) is a process of analyzing and summarizing a dataset by using visual and statistical methods to uncover patterns, relationships, and insights that inform further analysis and modeling.

There are many methods under this category and mainly 3 methods are used in this assessment

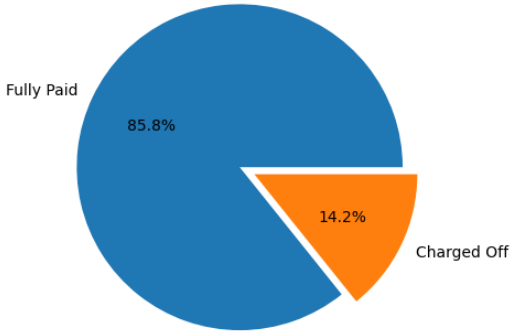
- ❑ Univariate Analysis
- ❑ Segmented Univariate Analysis
- ❑ Bi-variate Analysis

Based on the available columns and their data types , these are divided in to 3 catogories

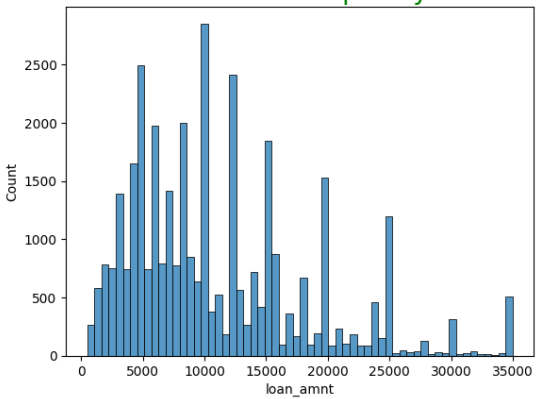
| Continuous variables | Categorical variables | Derived variables |
|---|---|---|
| loan_amnt | id | annual_inc_bins |
| int_rate | term | int_rate_bins |
| annual_income | grade | loan_amnt_bins |
| dti | sub_grade | dti_bins |
| pub_rec_bankruptcies | emp_title | loan_outcome |
| issue_d_month | emp_length | |
| issue_d_year | home_ownership | |
| | verification_status | |
| | loan_status | |
| | purpose | |
| | addr_state | |

## Univariate Analysis

### Loan Status Charged - Off vs Fully Paid




### Loan Amount frequency Plot




### Observation:

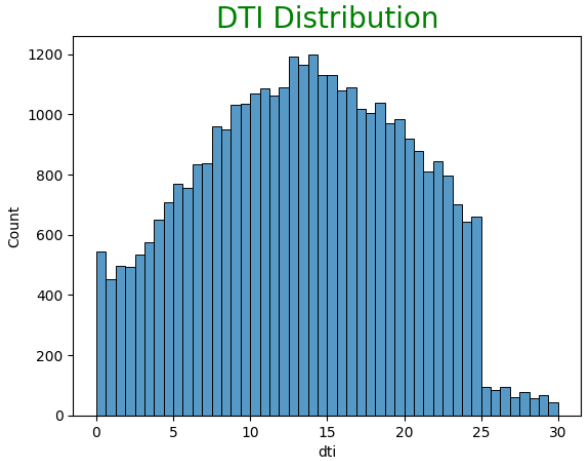
Based on the available data set , 85.8% of the loans are Fully Paid and 14.2% are Charged Off

- Fully Paid 30686
- Charged Off 5095

### Observation:

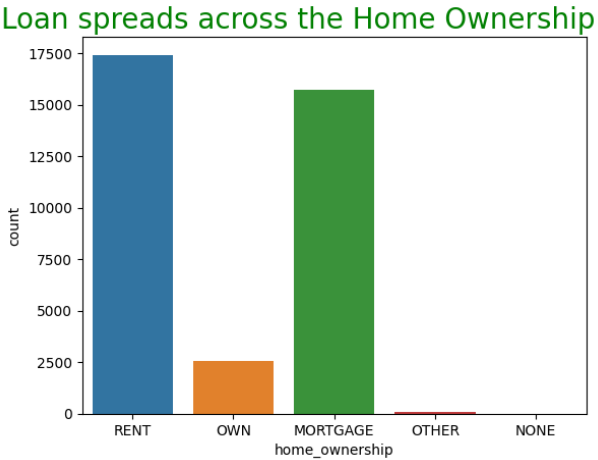
Most of the loans are issued for loan amount in the 5000 pool

## Univariate Analysis




Observation:

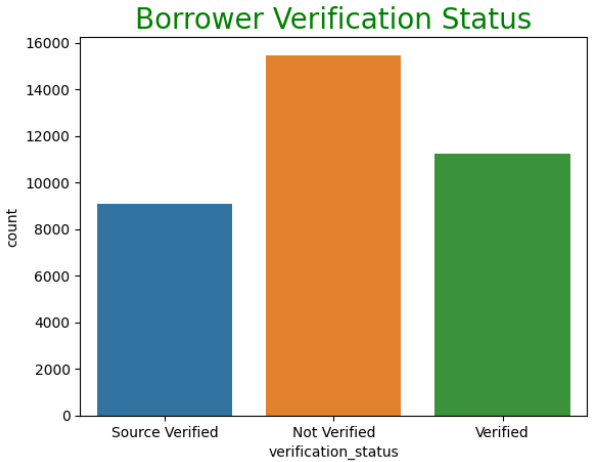
'dti' (Debt To Income) is peaking at 14 to 15 range




Observation:

Most of the loans are issued for clients who is staying at Rented House or at Mortgaged House

## Univariate Analysis




Observation –

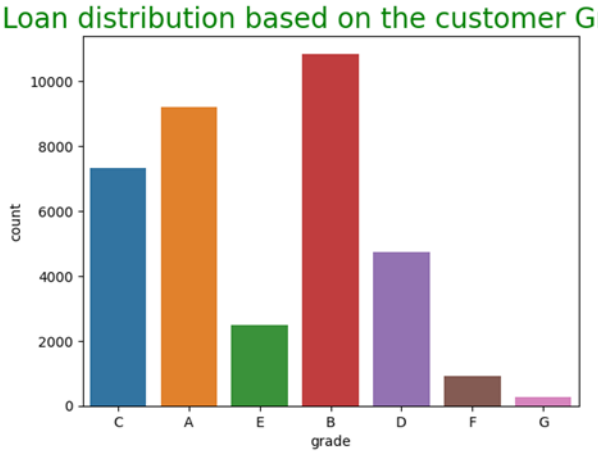
3 categories of Verification_status :

Source Verified - client's income verified ,

Verified - client verified by LC

Not Verified

There are scenarios where verification is not done , need to correlate it with loan status to check any relation here. This will be done under Bivariate analysis

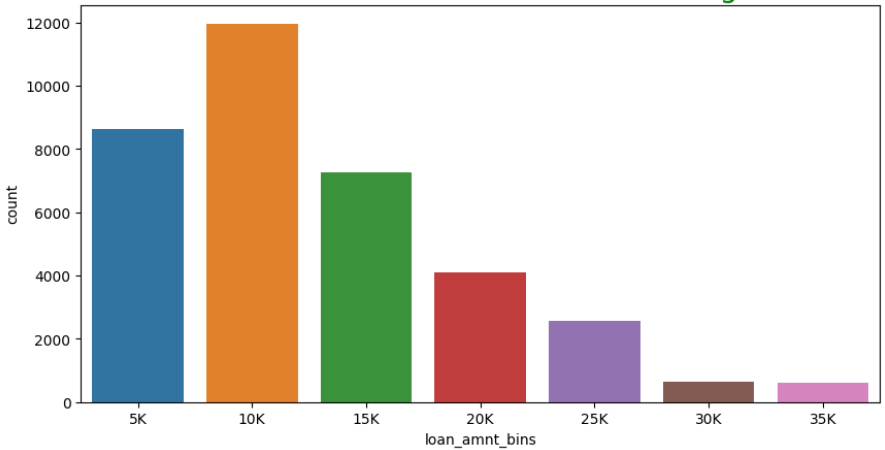



Observation:

- Loan grades are assigned based on both the borrower's credit profile and the nature of the contract.
- 'A' grade loans represent the lowest risk while 'G' grade loans are the riskiest.
- B and A grade loans are issued the most, while riskier grade loans like E,F,G are issued less

## Segmented -Univariate Analysis

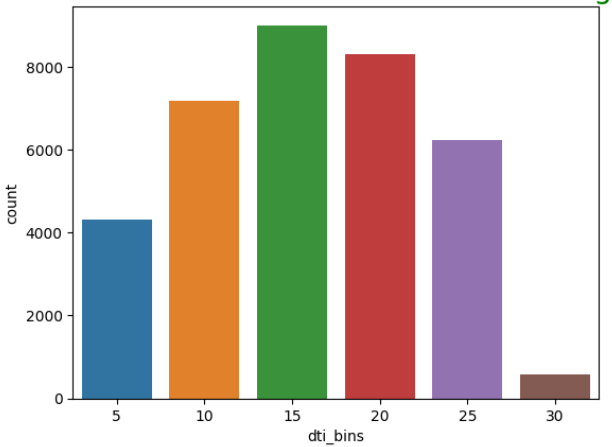
Loan Amount -Based on the Binning




Observation:

Based on the Binning , majority of the loan amount between 5k to 10K range

DTI Distribution based on the Binning

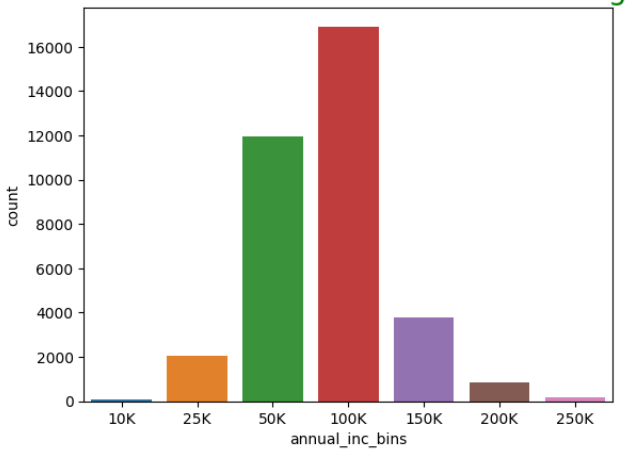



Observation:

'dti' (Debt To Income) is peaking from 10 max at 15 , then slowly down towards 20 range

## Segmented Univariate - Analysis

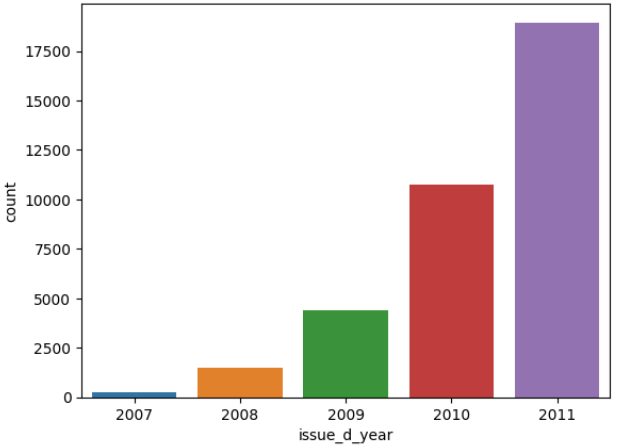
Annual Income -Based on the Binning




Observation:

- Most of the loans are issued for loan_amnt - 5000 -10000 range

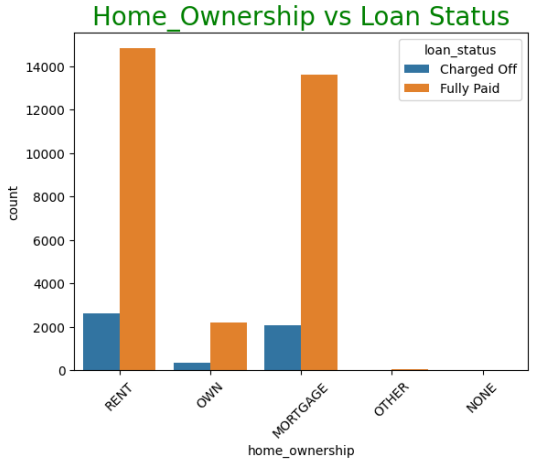
Loan Issued Year -Based on the Binning



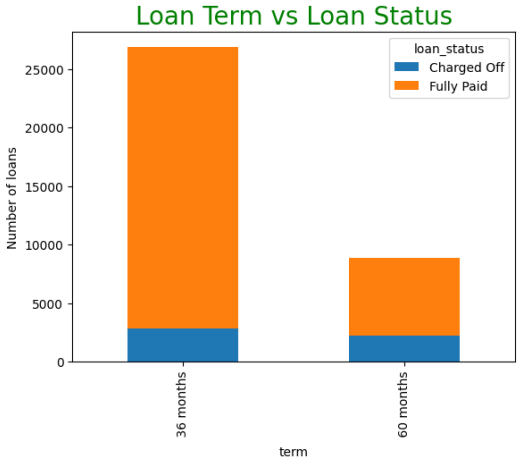

Observation:

- More Loans are issued in the year of 2011
- As year increased , number of loans issued also increased - linear relation

## Bivariate - Analysis

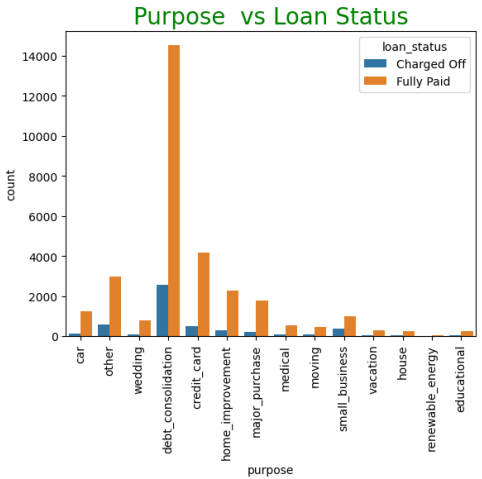


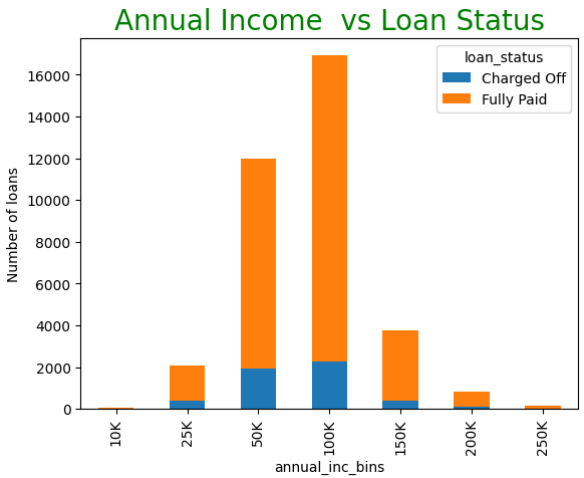

Observation:
Charged Off loans are relatively lower for clients those who own the house compared to other categories




Observation:
Close to 70% of the loans were for 36months and remaining are for 60months

## Bivariate - Analysis



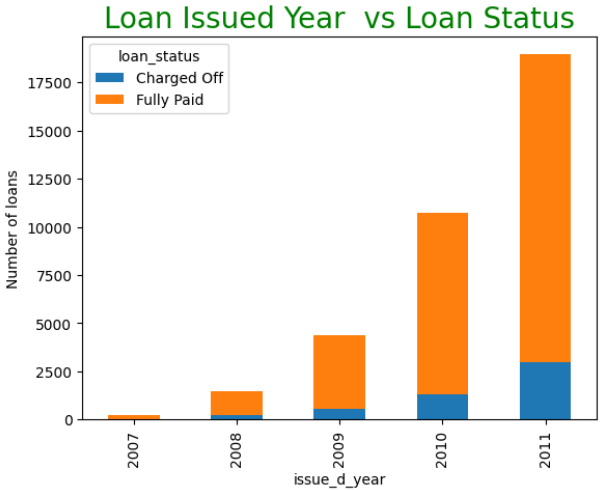

### Observation:

- Most of the fully paid clients have opted loan for 'debt_consolidation' purpose
- Charged off loans are also higher for 'debt_consolidation' purpose compared to other purposes of loan
- Most of the loans are issued for 'debt_consolidation' purpose
- Also there are clients taken loan for the credit card payment and some are not able to clear it

### Observation:

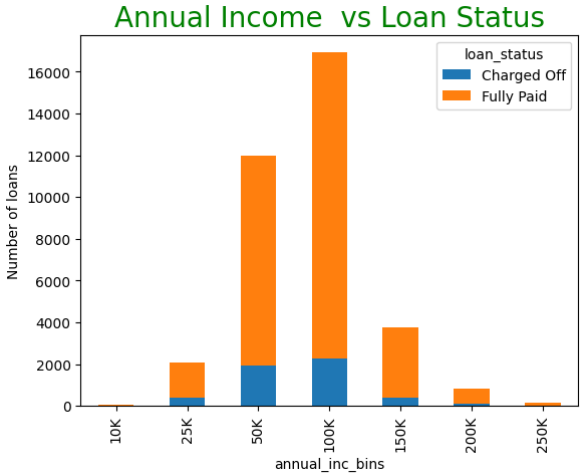
- People with annual_inc < 100000 are applied for more loans
- Out of that 20-25% of the people were defaulters
- People with salary > 10000 , taking less number of loans and majority fully paid the amount

## Bivariate - Analysis




### Observation:

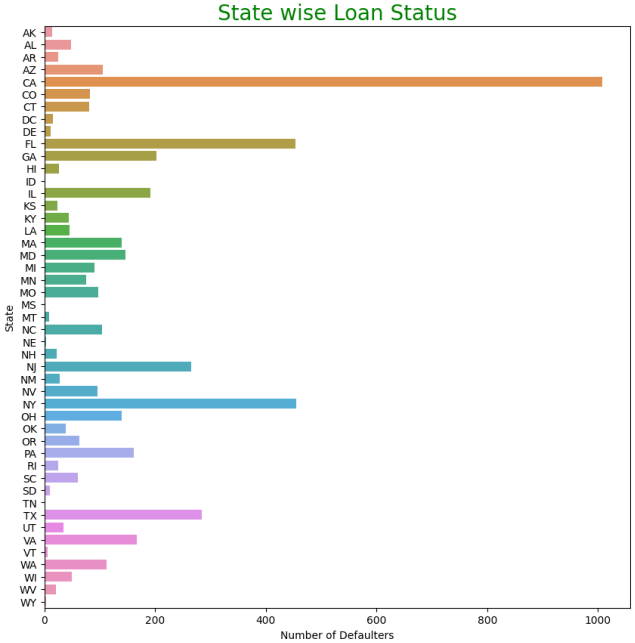
- The number Charged Off and Fully Paid loans have linear growth over the years 2007 - 2011
- most of the loans are issued in the year of 2011
- There was an econmic recession in US @ 2011 , may be this can be one of the reason for more defaulters in that time frame




### Observation:

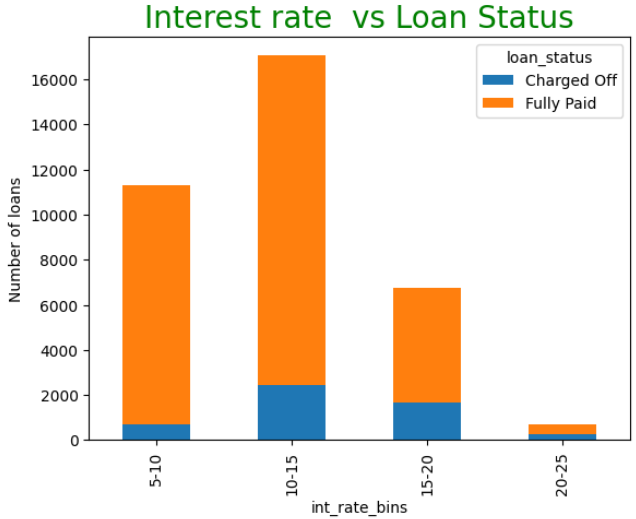
- People with annual_inc < 100000 are applied for more loans
- Out of that 20-25% of the people were defaulters
- People with salary > 10000 , taking less number of loans and majority fully paid the amount

## Bivariate - Analysis




### Observation:

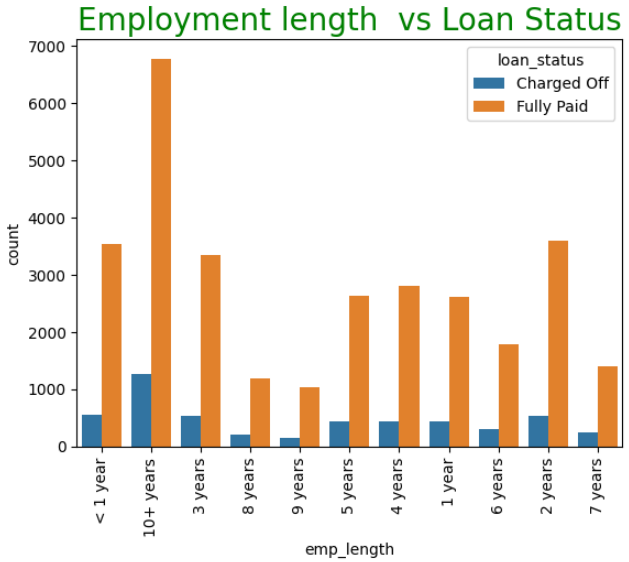
- More defaulters are from the state 'CA'
- second defaulters are from Florida and New York




### Observation:

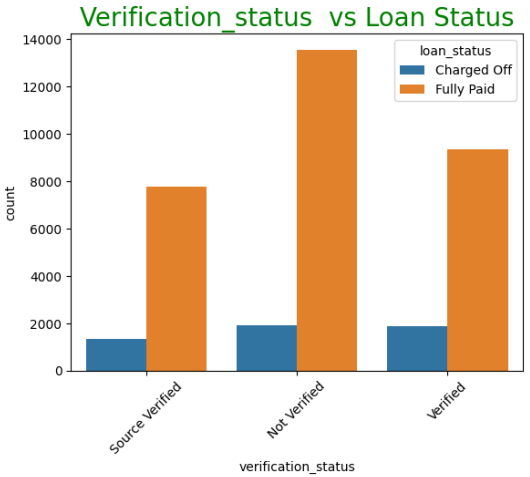
- If the interest rate is lower, the loans are likely to be fully paid
- majority of the people took loan at the interest rate of 10-15% , out of which close to 15% are defaulters
- with the increase of interest rate, more chances of defaulters

## Bivariate - Analysis




### Observation:

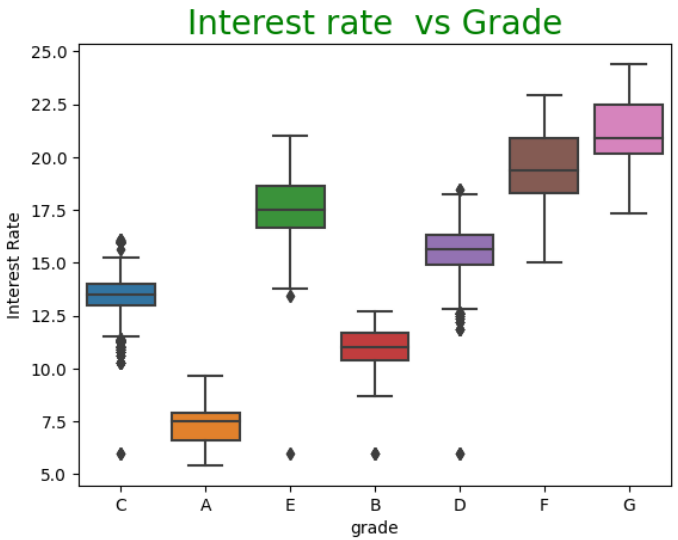
- Employees with 10+ years tend to charge off more than other employees
- less than 10 year employment length , each category showing same share in charged off

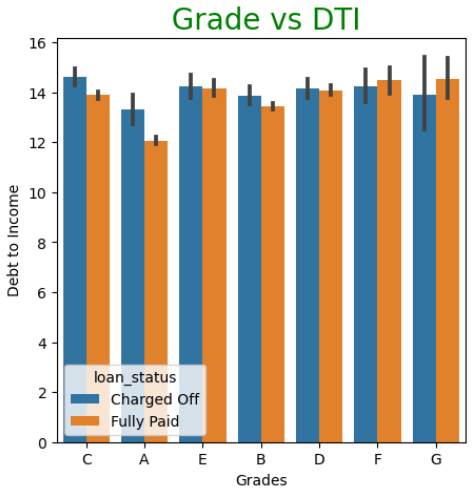



### Observation:

- Applicant with 'Source Verified' showing very likely to default compared to others
- So its an eye opener - Verifying the Source of the applicant can reduce the defaulter case.

## Bivariate - Analysis




### Observation:

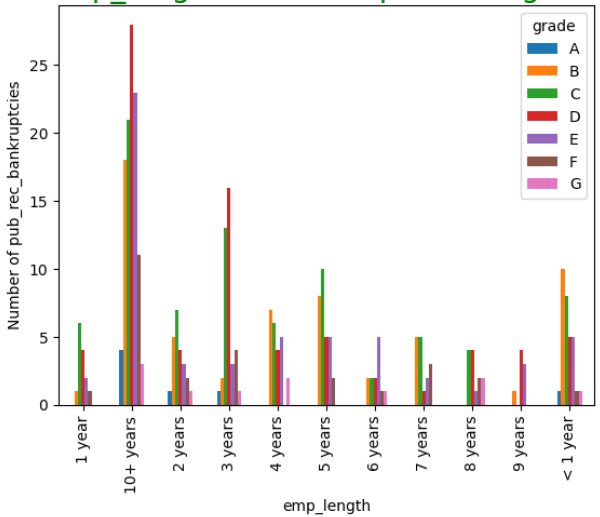
- As the grades A,B,C,D,E,F,G are categorized from low to high risk,the interest rate increases as the risk increases
- Good candidates are Applicant with 'A' category with less interests are tend to clear the loan

### Observation:

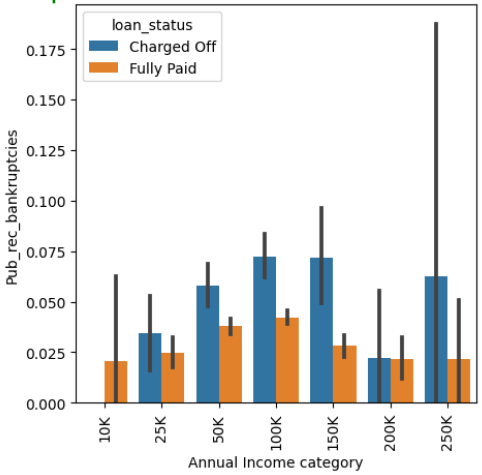
- Fully paid - grade A loans have the lowest debt_to_income values

## Bivariate - Analysis

### Emp_length vs bankruptcies vs grade




### bankruptcies vs annual income vs loan_status



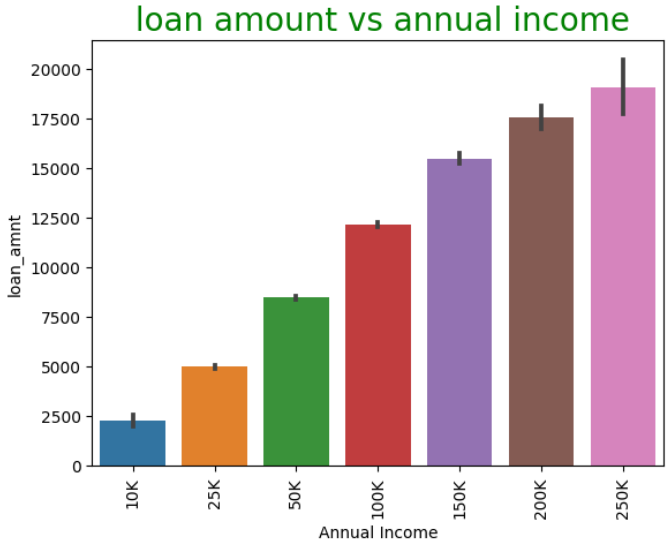

### Observation:

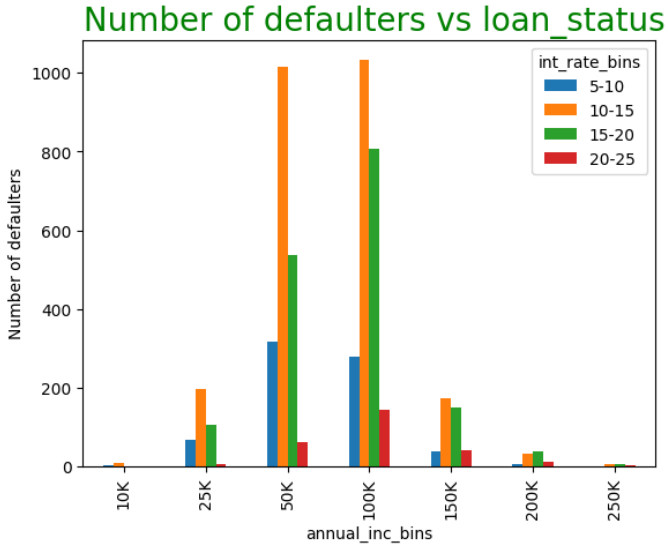
- Many bankruptcies tend to occur with Charged Off clients having 10+ years working experience and mostly they are of D-grade loans

### Observation:

- There are no Charged Off loans for annual income <10000
- Clients with pub_rec_bankruptcies tend to get charged off

## Bivariate - Analysis




### Observation:

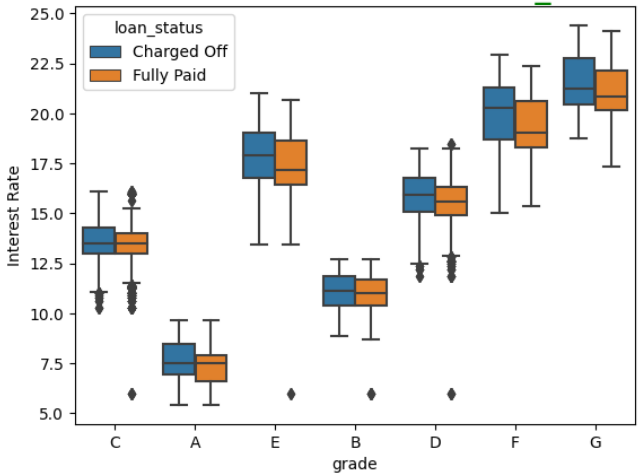
- As annual income increases, loan amount also increases (shows positive correlation)

### Observation:

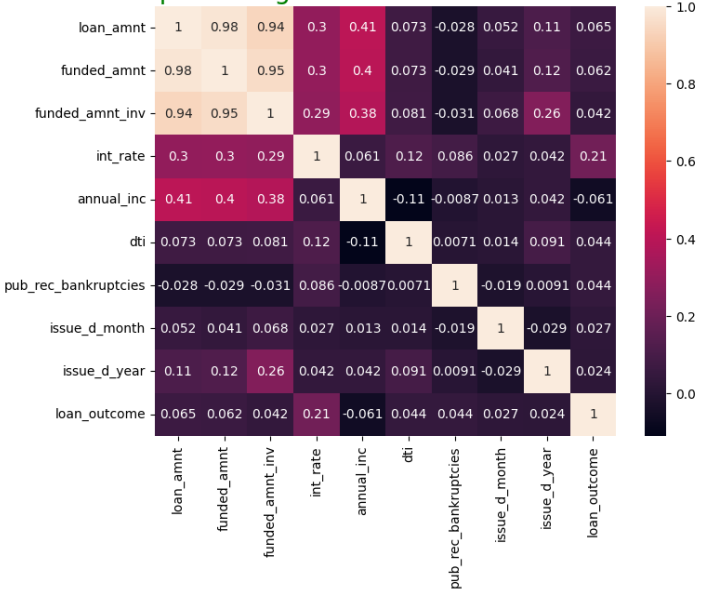
- Many defaulters tend to have[50000-100000] annual income and most probably the interest rates of these loans fall under [10%-15%] category

## Bivariate - Analysis

### Interest rate vs Grade vs Loan_Status




### HeatMap showing correlation with other attributes



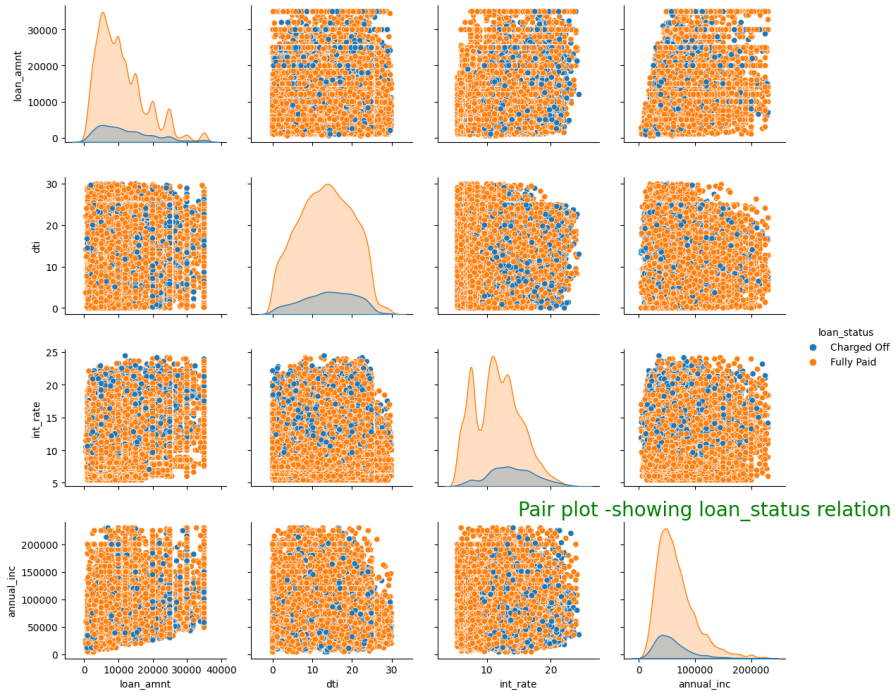

### Observation:

- As loan_amount increases, interest rate also increases (shows positive correlation)
- Above category 'C' , more likely to default. This can be due their profile plus and higher interest rate

### Observation:

- Loan_amnt and annual_inc are slightly correlated
- Int_rate and loan_amnt are slightly correlated

## Pair Plot – Correlation with multiple variables against Loan Status




Pair plot -showing loan_status relation

### Observation:

- Interest rate and the dti are the key factors for the defaulers
- when the loan amount and interest rates are high , people tend to default

## Observation from the Univariate analysis

- Based on the sample data , 85.8% of the loans are fully paid (non-defaulters) and 14.2% of the loans are charged off (defaulters)
- Most of the loans are issued for loan_amnt - 5000 to 10000 range (10k Bin)
- Majority of the loans are issued against the applicant profile with B and A and very less towards other catogory like C,D,E,F,G
- The frequency of 'dti' (Debt To Income) is high around 14 to 15 range, which means most of the issued loans have debt to income ratio around 15
- 70% loans are issued for '36 months' and 30% loans are issued for '60 months'
- Most of the loans are issued for clients who's home ownership type 'RENT' or 'MORTGAGE'
- Majority of the loan issued for the people whose background verification done correctly either at Source or with available data
- More people took loan for the 'debt_consolidation' and credit card payment

## Observation from the Segmented Univariate analysis

- Charged Off loans are less for the clients those who own the House compared to other categories of home ownership
- Most of the people took loan for the 'debt_consolidation' purpose and credit card payment
- The people who took loan for debt_consolidation shows more default, compared to other cases
- Mean of debt_to_income (DTI) increases linearly across the splitted debt_to_income buckets / bins
- Clients having annual_inc > 100000 are likely to pay the loan fully
- Employees with 10+ years tend to get charged off more than other employees
- The number of Charged Off and Fully Paid loans have linear growth over the years 2007 - 2011
- More defaulters are from the state 'CA'
- Most of the issued loans have 'dti' values populated around 14 to 15
- If the interest rate is lower, the loans are likely to be fully paid
- The number of issued-loans linearly increase across the years 2007 - 2011

## Observation from the Bivariate analysis

- Grades A,B,C,D,E,F,G are categorized from low to high risk, the interest rate increases as the risk increases
- Majority of the loans issued to the people with Grade B and A
- Fully paid - grade A loans have the lowest debt_to_income values
- Income and Loan amount showing the positive correlation same with the interest rate
- Majorify of the Charged Off clients having 10+ years working experience and mostly they are of D-grade loans
- Charged Off clients tend to have higher interest rates compared to Fully paid clients
- Annual income of fully paid clients is slightly higher than that of Charged Off clients
- There are no Charged Off loans for annual income <10000
- Clients with pub_rec_bankruptcies tend to get charged off
- High interest rate,higher Grade level(> B) tend to default more compared to others

## Key insights from the analysis

Key attributes which deciding whether a loan applicant tend to default or not :

- - Annual Income
- - Verification Status
- - Grades
- - Interest rates
- - DTI
- - Pub_rec_bankruptcies

Additional points to consider while approving the loan

- - Verification of documents / Source
- - Income in the range 50000-100000 are tend to default
- - with past pub_rec_bankruptcies
- - Grades above D grade( E,F,G ) are tended to default
- - Clients from 'CA' state showing defaulter nature

upGrad

# Thank You!