LENDING CLUB CASE STUDY

SUBMISSION

Group Name:

1. Harsh Prateek Singh
2. Ravi Venkatesh
3. Sharmistha Das
4. Viswa Vivek T

Lending Club Case Study Abstract

Business Model:

Lending Club provides online platform for investors (individuals, institutes etc) and borrowers (individuals , business etc.) for exchange the money as loan. Business model work towards to provide less interest rates to borrowers while keeping the investor interest without impacting.

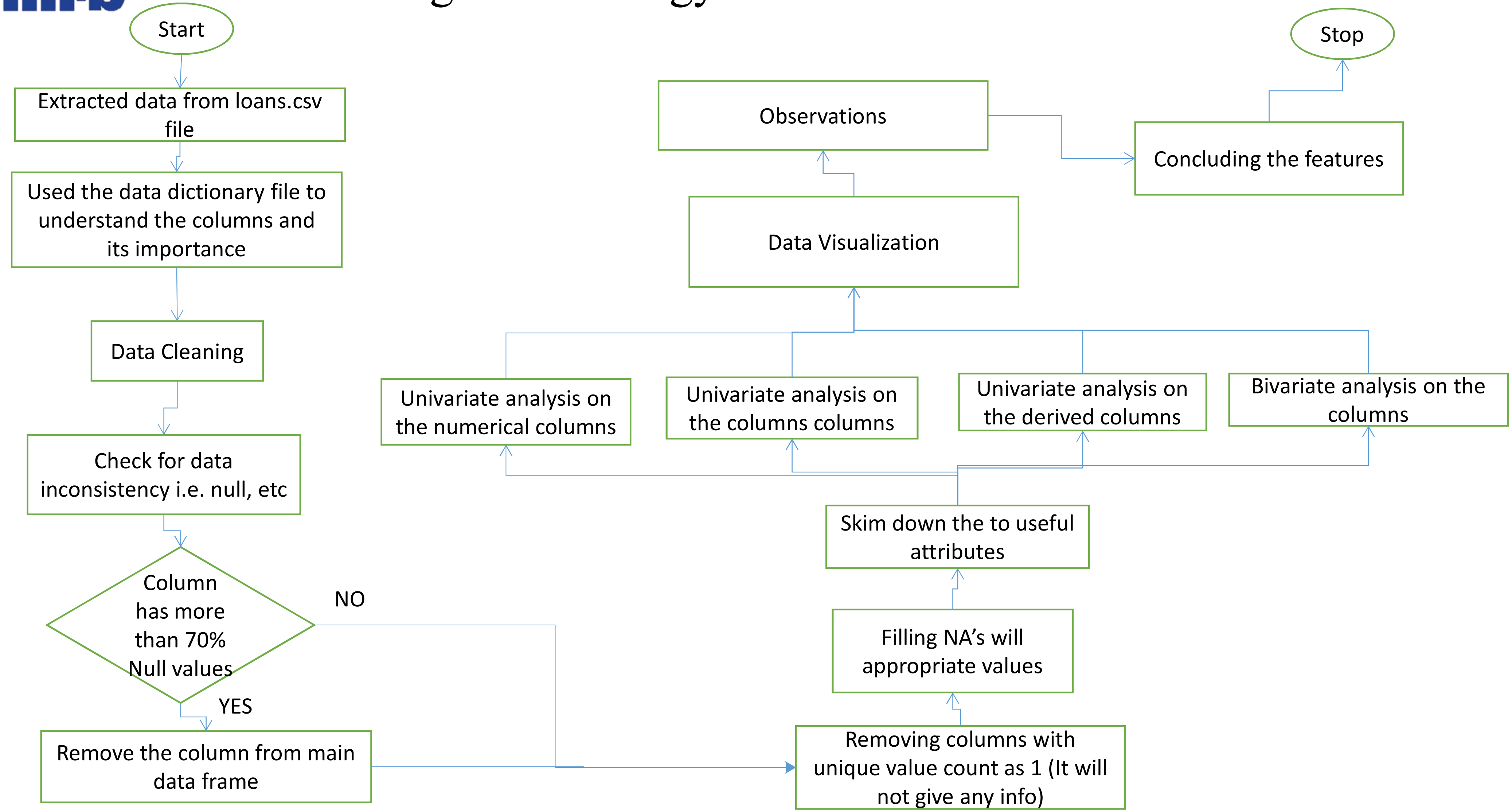
A loan is split into notes of $25 each, and investor can buy the notes in multiples of 25 based on investor potential. The amount is transferred to borrower. Borrower give monthly interest for investor and the loan will be closed based on loan term.

Business Case:

As loan lending always has risk of financial loss, in terms of loan default (Charged Off) , which is inevitable however financial losses can be reduced if such loan application can be identified.

Lending club 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 and to gain investors trust.

Problem solving methodology



Attributes Considered for Analysis

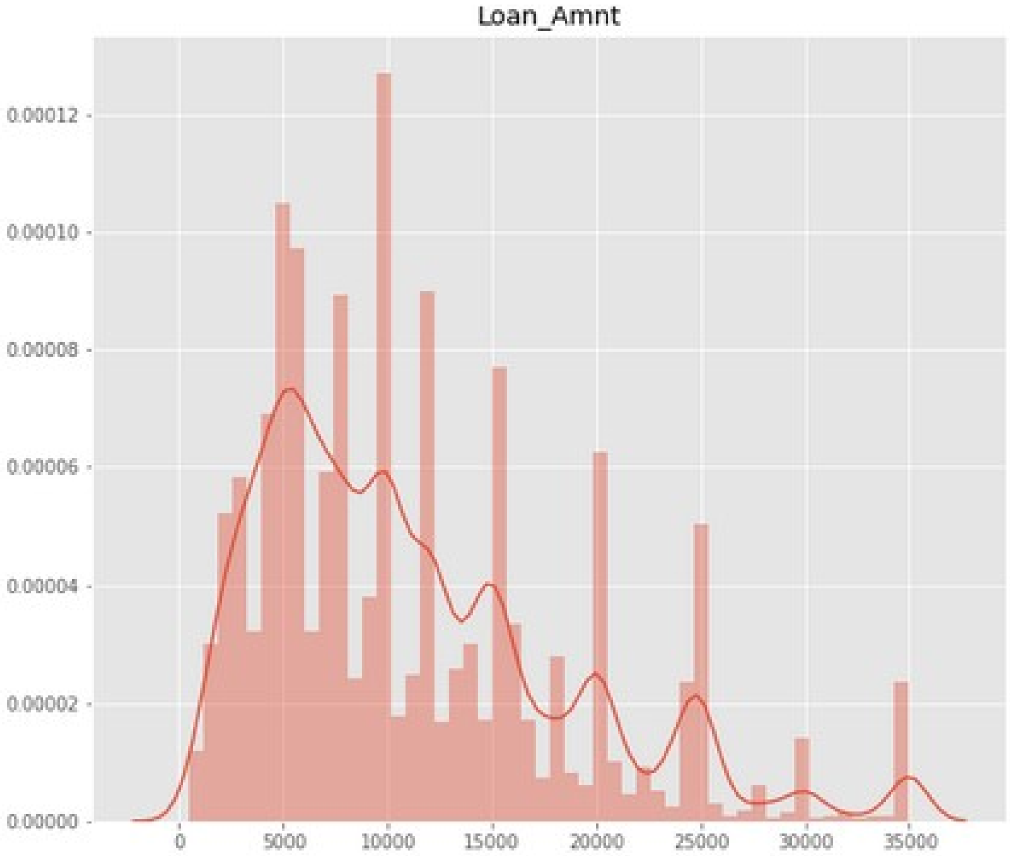
Attributes	Attributes
loan_amnt	delinq_2yrs
funded_amnt	earliest_cr_line
term	mths_since_last_delinq
int_rate	open_acc
sub_grade	pub_rec
emp_length	revol_bal
home_ownership	revol_util
annual_inc	total_acc
verification_status	out_prncp
issue_d	total_pymnt
loan_status	total_rec_prncp
Purpose	recoveries
addr_state	last_pymnt_d
dti	pub_rec_bankruptcies

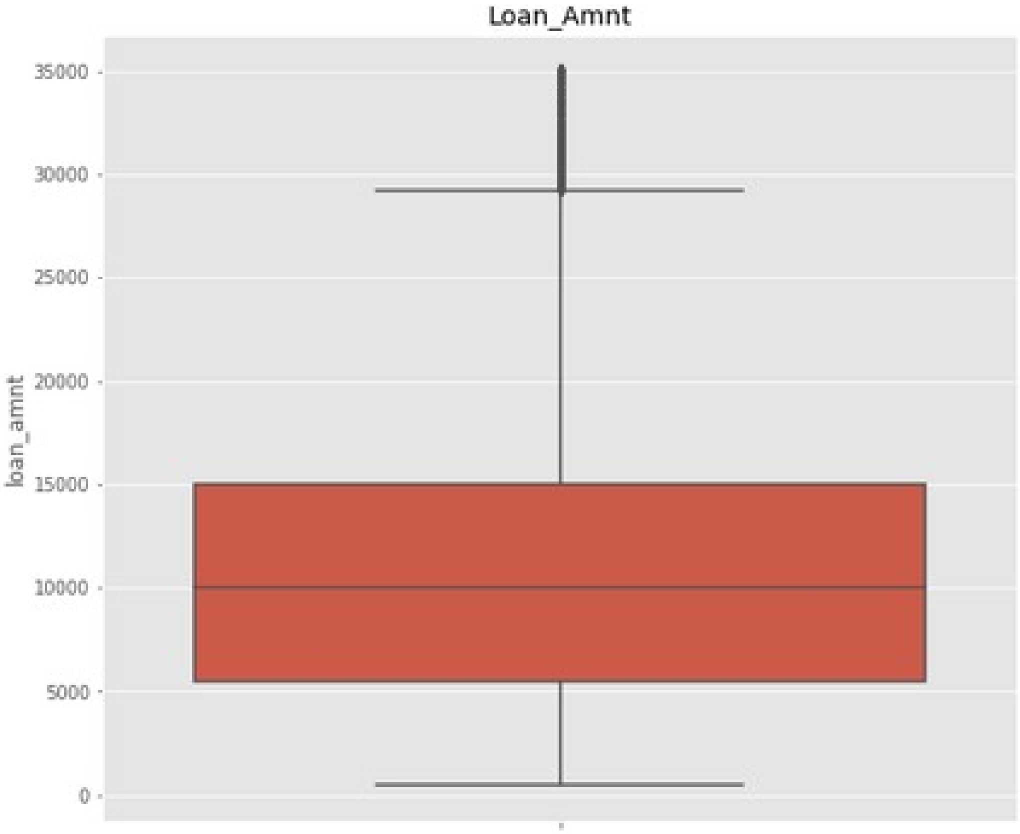
Analysis approach

Approach:

1. Load the CSV file in loan_df data frame
2. Use the Dictionary file to understand the attribute in the loan csv, and its importance
3. Remove all the columns with all the entries as NA.
4. Fix the data inconsistency such as data type, blanks etc.
5. Remove the columns with 1 unique value. As data dont have change hence we cant inferr anything.
6. Convert loan_status to caps to avoid an y grouping issue or sorting issue.
7. Create bins for Loam_amnt, Annual_inc, interest rate .
8. Create derived columns from Issue_d,.
9. Based of domain knowledge and team brain storming we concluded column required for analysis.
10. Filter out Numerical attributes for Numerical univariate analysis.
11. Filter out Numerical attributes for categorical univariate analysis.

Univariate Analysis -- Numerical

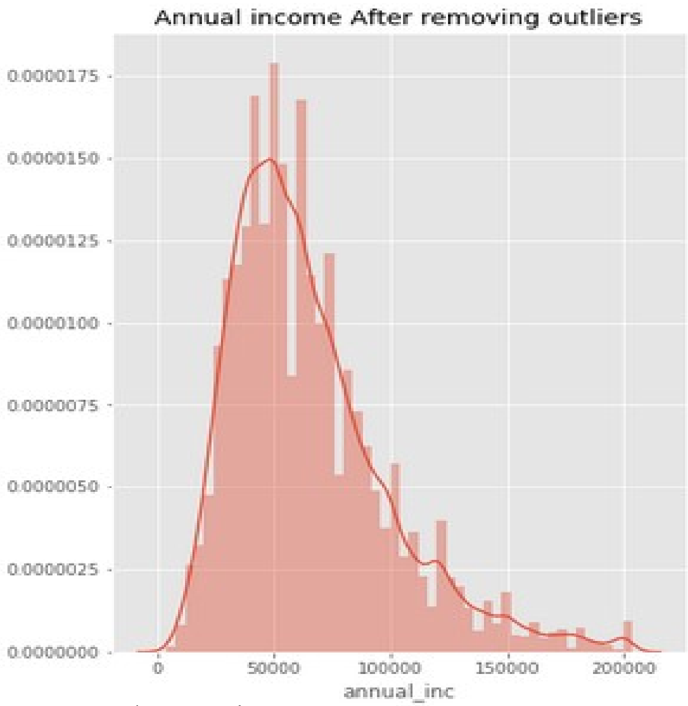


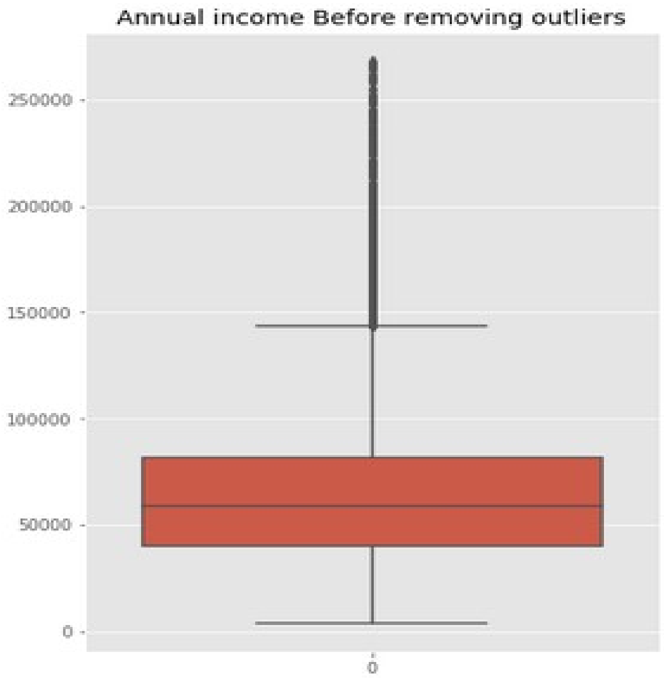


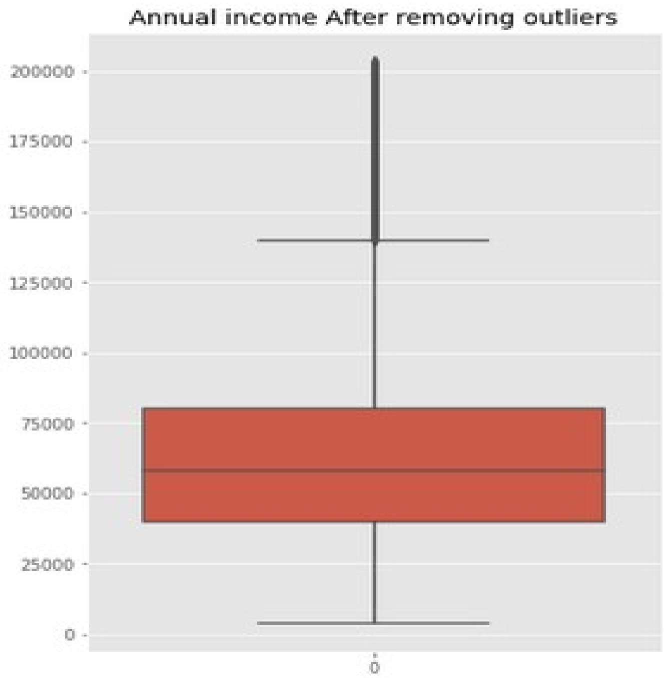
Observations

- 10000$ is the median of the population.
- The maximum number of loans applied between 6000-15000$

Univariate Analysis



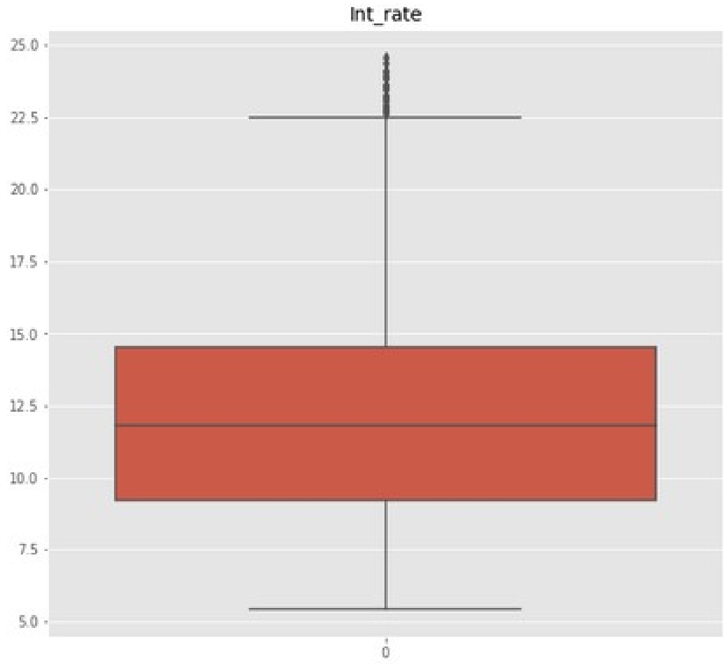


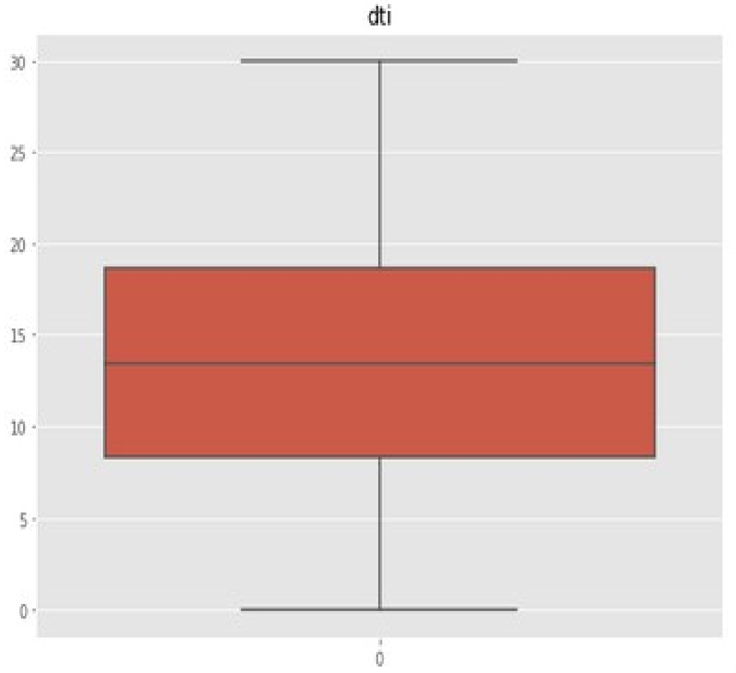


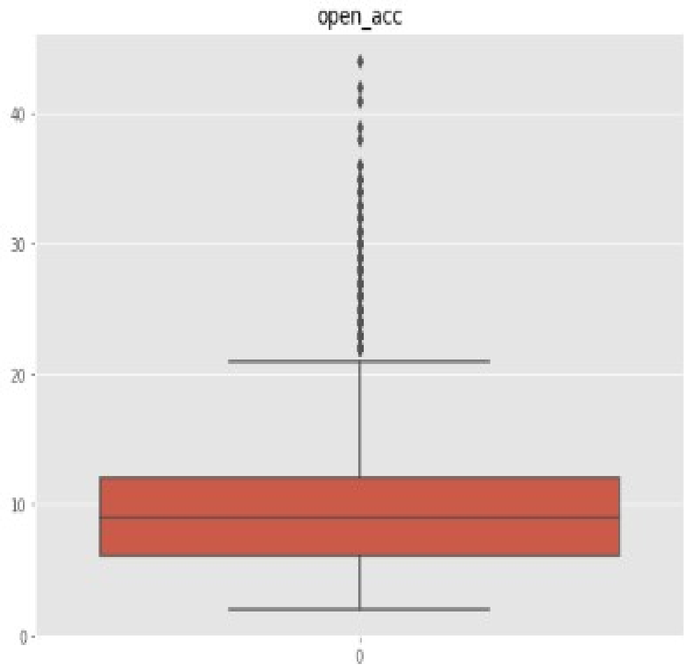
<u>Observations</u>

- Removing all outliers i.e more than 99.3 percentile as it hardly contributing to 0.83 % of all the charged off amount and affecting the population mean to large extent.
- 99.3 percentile coming to 269975 Dollar because data points effecting the data starting after 25000 Dollars
- The histograms is showing right skewed data after removed the outliers i.e.more than 99.3 percentile. It means it move the mean is increased and cant be considered for any analysis as it will give wrong information.
- Most of the annual income is between 40000$ to 90000 $ approx annually.

Univariate Analysis



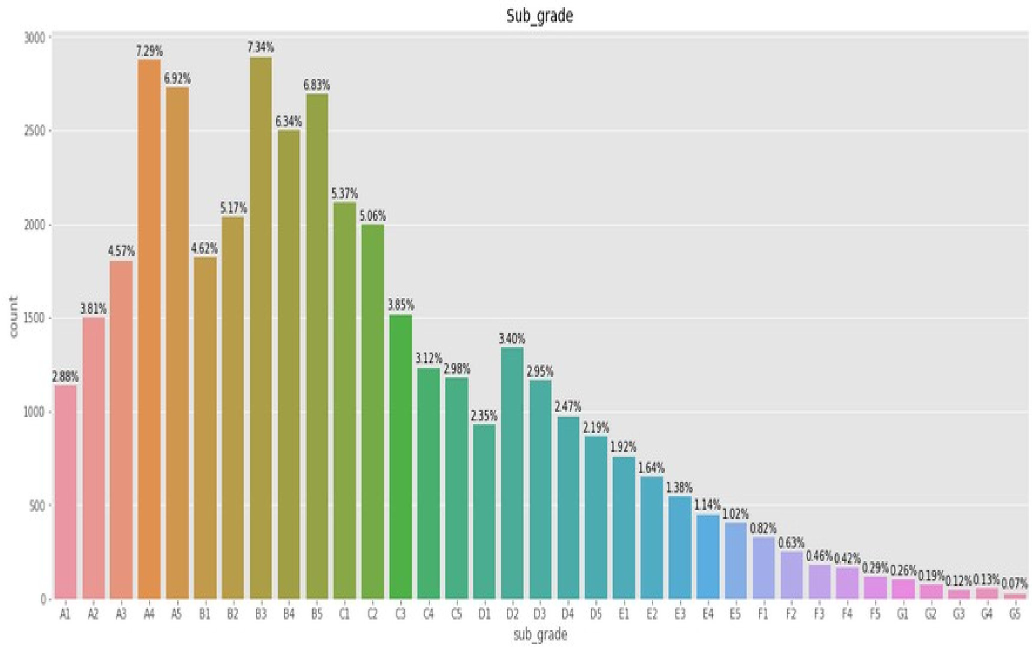


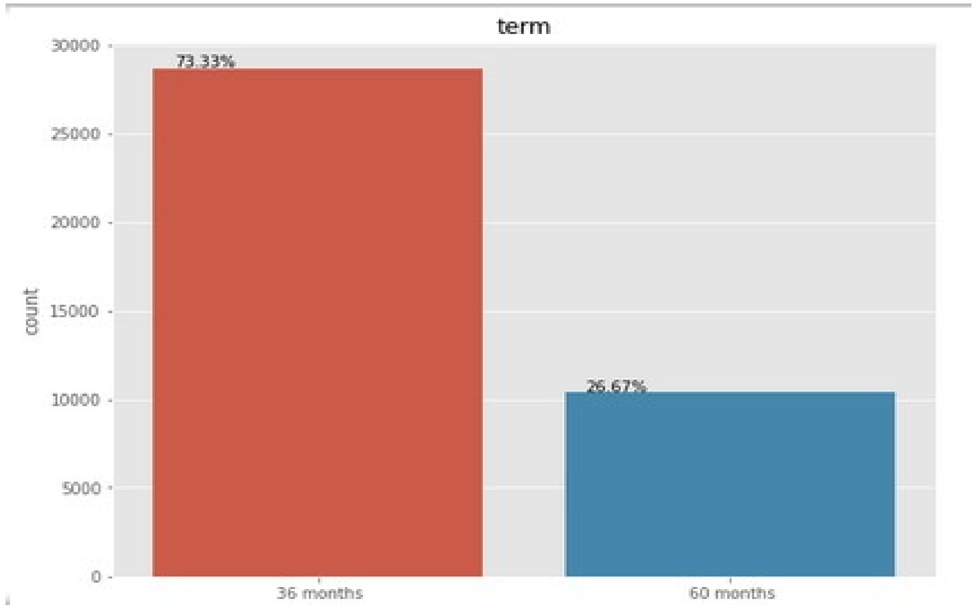


<u>Observations</u>

- Most of the borrowers are charged between 9% to 15% annually approx int_rate
- DTI(debt to income ratio): Most of the applicant DTI is between 8-19, the less the better.Less ratio is favorable for applicant.
- Open Credit Line (open_acc): Most of the applicant has 5-12 cerdit line.

Univariate Analysis -- Categorical



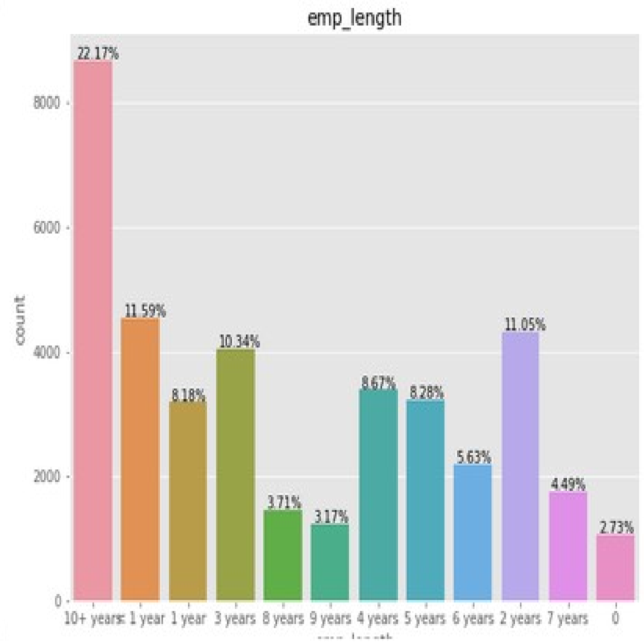


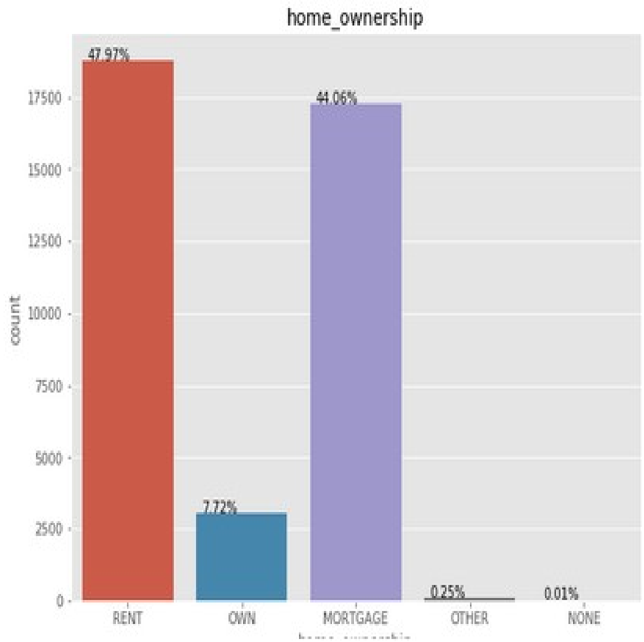
<u>Observations</u>

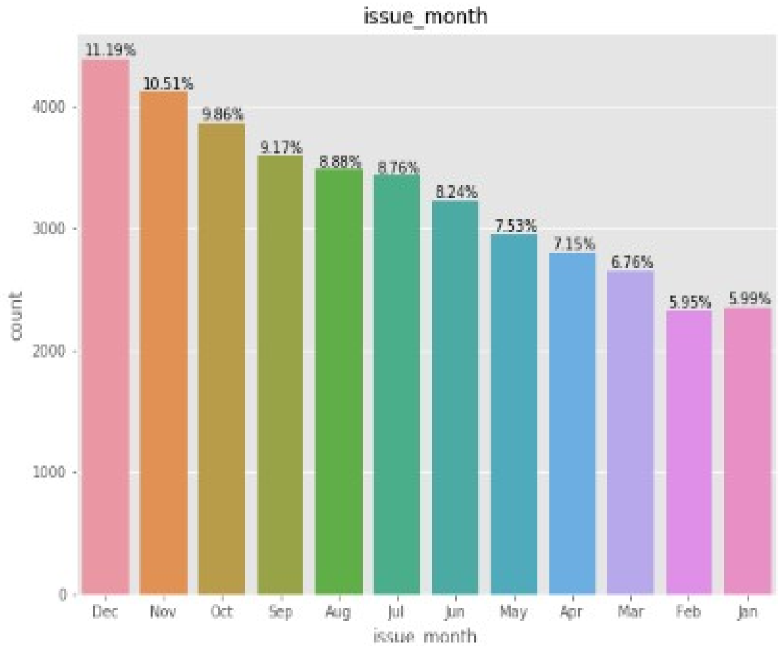
- 7.3% of borrower sub-graded in B3, followed by A5,A4,B5.
- Applicants prefer 36 months term over 60 months term

Univariate Analysis -- Categorical

UpGrad





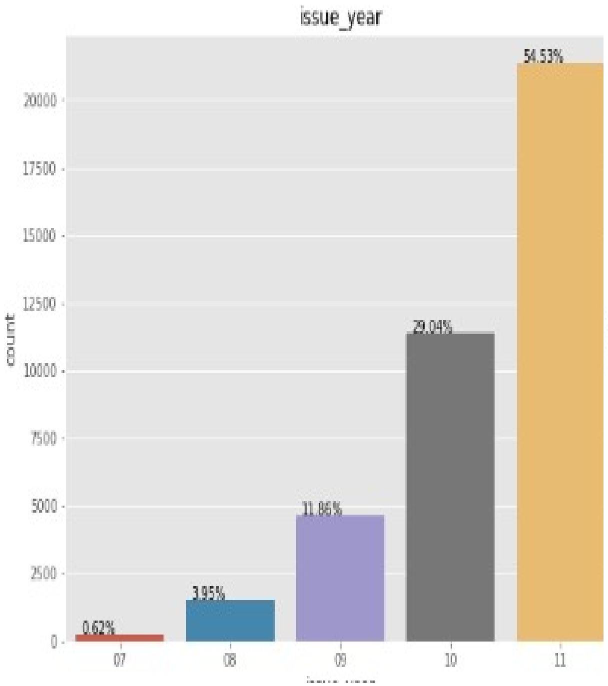


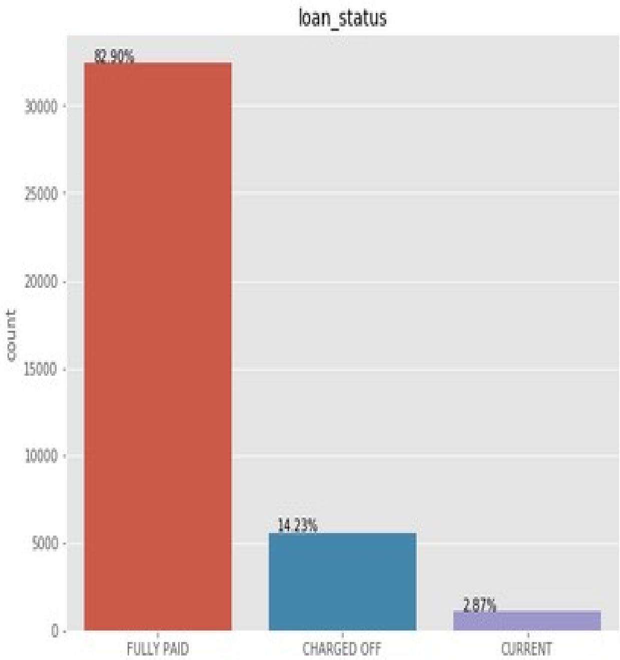
Observations

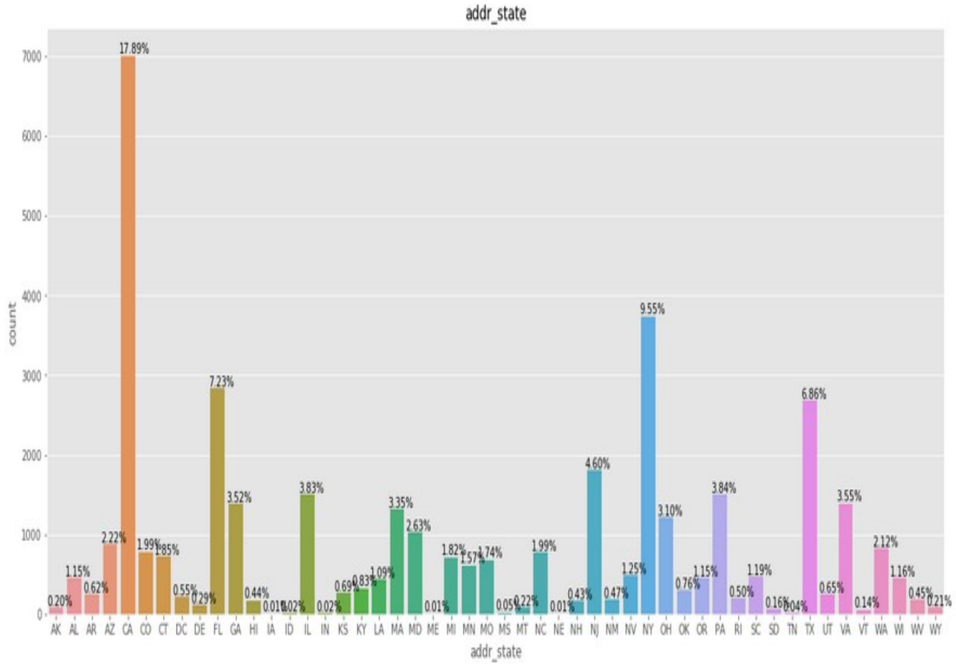
- Experience with more than 10 years tend to look for loans.
- People live in rented house also apply for loans followed by Mortgage homes
- Last quarter of the year is the time when more applicants reaches for loan.

Univariate Analysis -- Categorical

UpGrad







<u>Observations</u>

- Year on year more loans are issued which mean user buyer power is increasing.
- More people tend to pay the loan however 14% approx people not pay the loan
- California is potential state for more state

Bivariate Analysis

UpGrad

Mindset considered for Bivariate Analysis:

As we have to understand the attributes which are effecting the loan_status adversly. We have to deep dive into the analysis of each attribute's category wise along with analysis on population to understand if any category of an attribute contributing more toward the charged off .

Keeping the above mentioned thing in mind we have analysed the overall population of a attribute and category population.

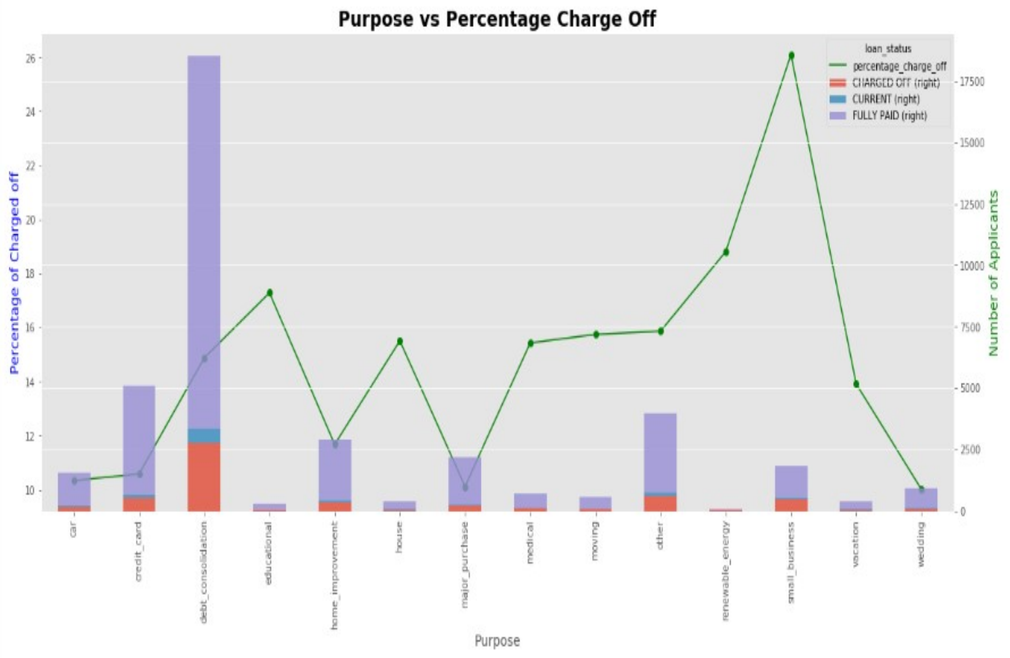
We have considered **"percentage of charged off "** as measure of category population charged off data. Below the formula used to calculate it.

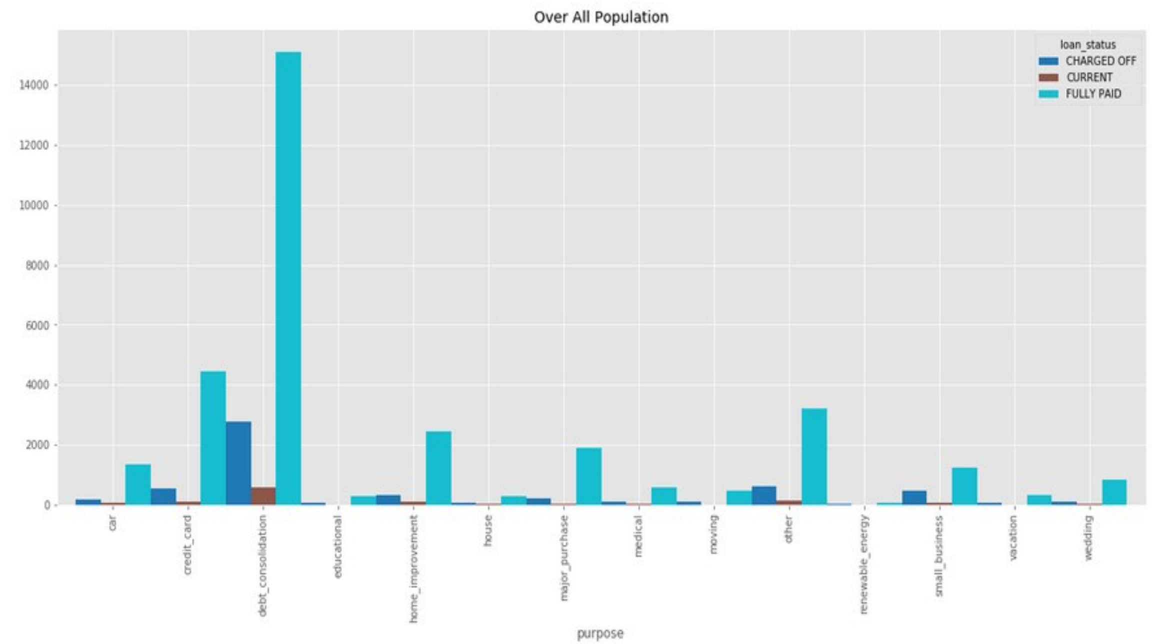
$$\text{Percentage of charged off} = \frac{\text{Charged off loan count for the category}}{\text{Total loan Count of the category}} * 100$$

Example:

loan_status total_acc	CHARGED OFF	CURRENT	FULLY PAID	All	percentage_charge_off
2	1	0	3	4	25.00
3	42	3	137	182	23.08
4	79	5	335	419	18.85
5	90	9	452	551	16.33
6	105	9	567	681	15.42
7	132	15	681	828	15.94
8	171	17	815	1003	17.05
9	166	23	890	1079	15.38
10	187	21	981	1189	15.73
11	198	37	1040	1275	15.53

Bivariate Analysis

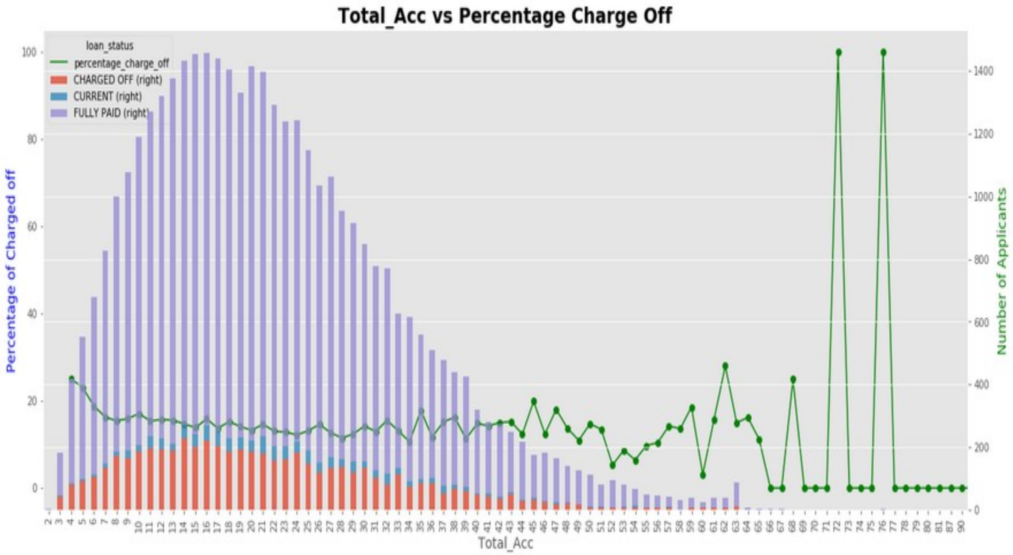


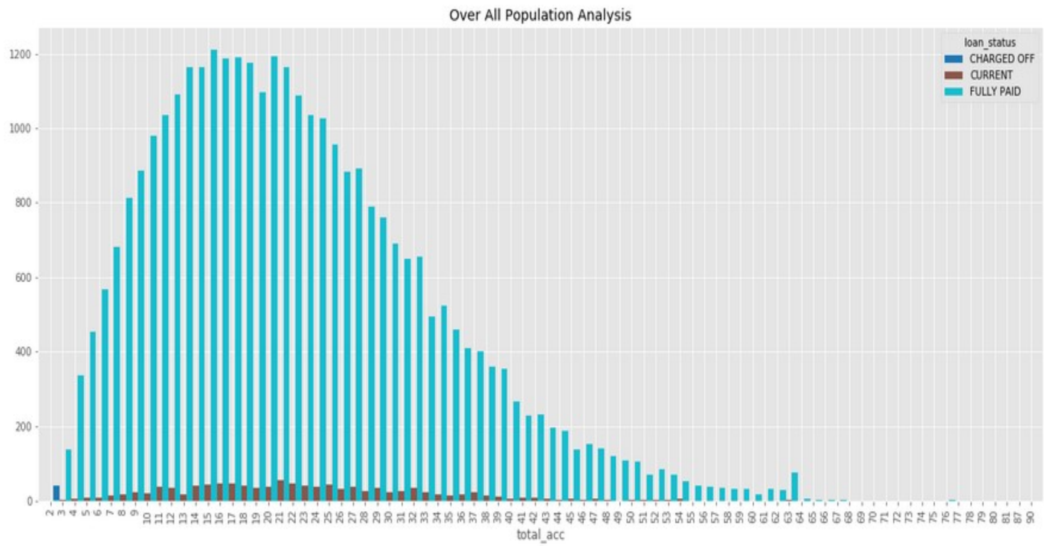


<u>Observations</u>

- Within the category Small Business is showing more charged off percentage i.e. 26% approx. (refer the above for percentage charge off calculation)
- For Over all population purpose=debt_consolidation have highest count of loan_status as 'Charged_Off'. This is because we are comapring the category against the whole population. However we need category wise information which gives more insights on which category is defaulted highly when compared to itself data.

Bivariate Analysis

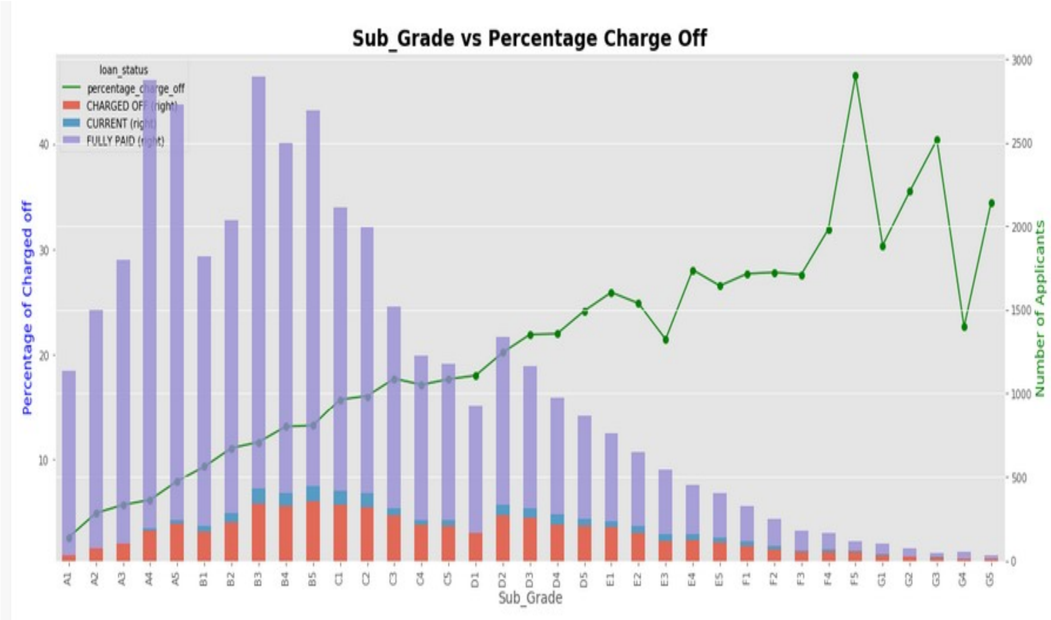


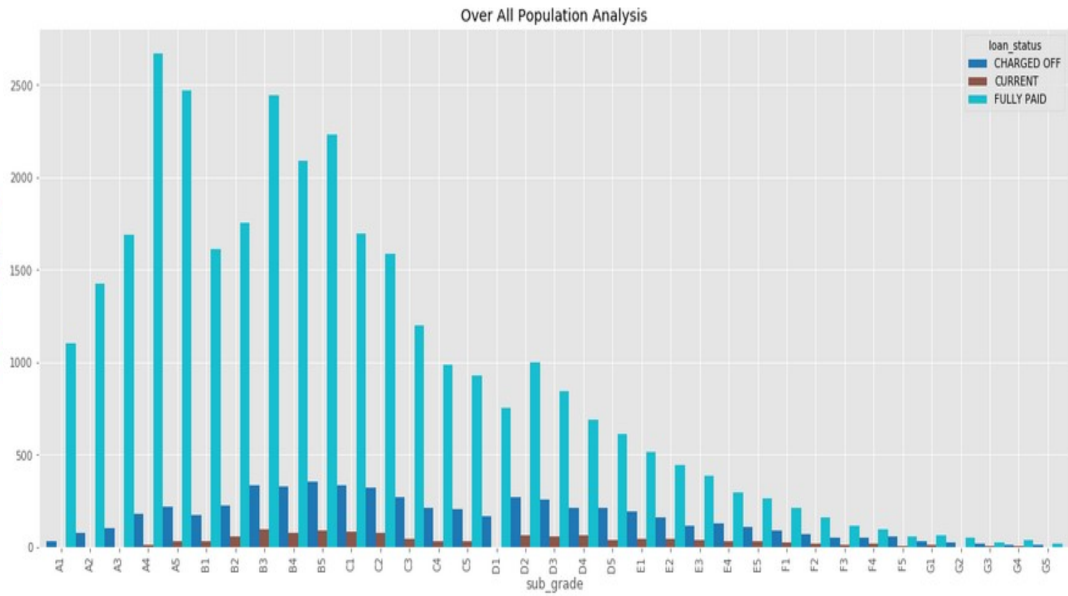


<u>Observations</u>

- Total Cedit line having 4 contribute to 21% charged off. We didn't consider credit line with 72 and 76 thoug its 100% because the sum of loan amount is less than 20k however for credit line 4 its $2M

Bivariate Analysis



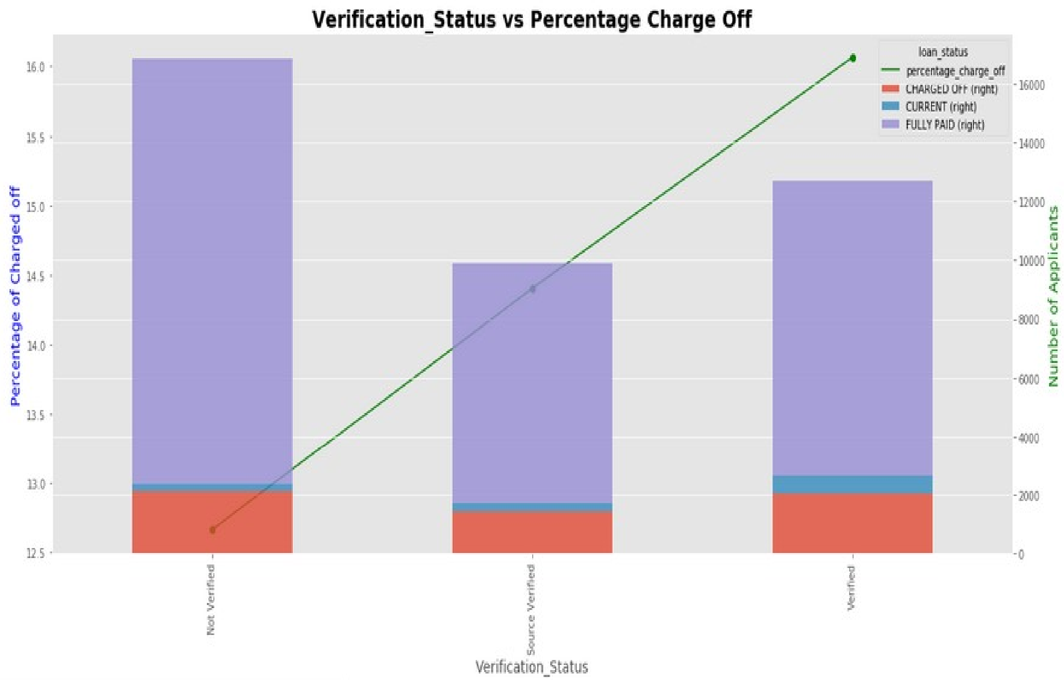


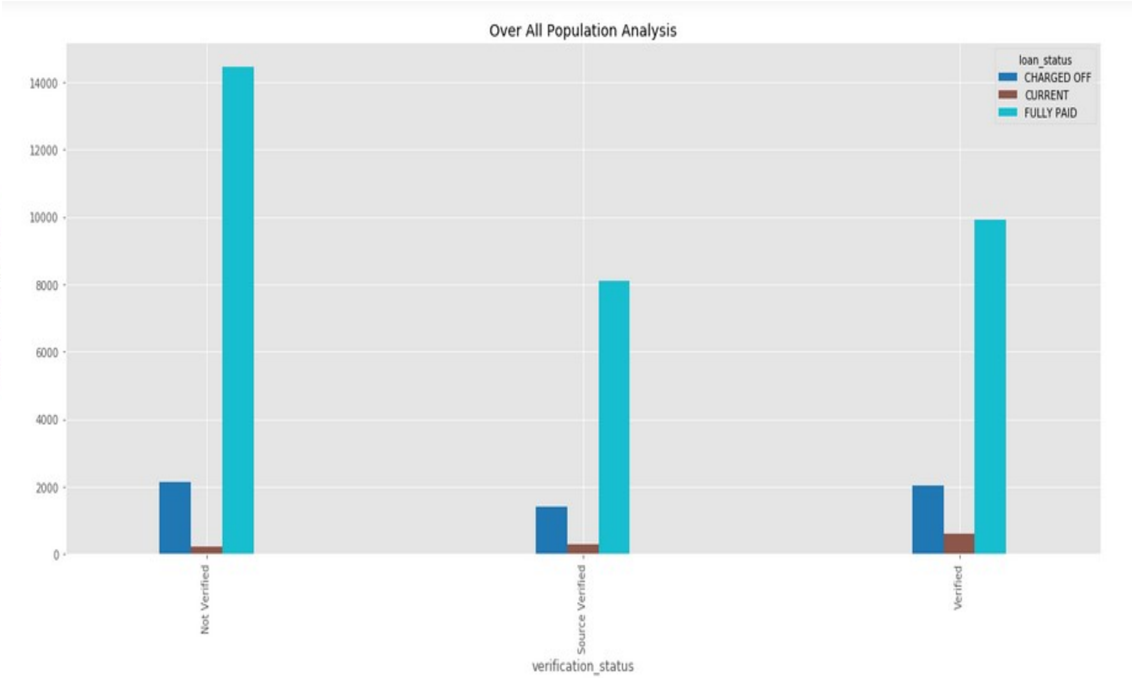
<u>Observations</u>

- Loan default percentage within category increase as sub-grade increases i.e. interest rate increases.
- On in overall population:
  1. People with sub_grade B5 have highest count of loan_status as 'Charged_Off'
  2. People with sub_grade B3 have second highest count of loan_status as 'Charged_Off'

Bivariate Analysis

UpGrad

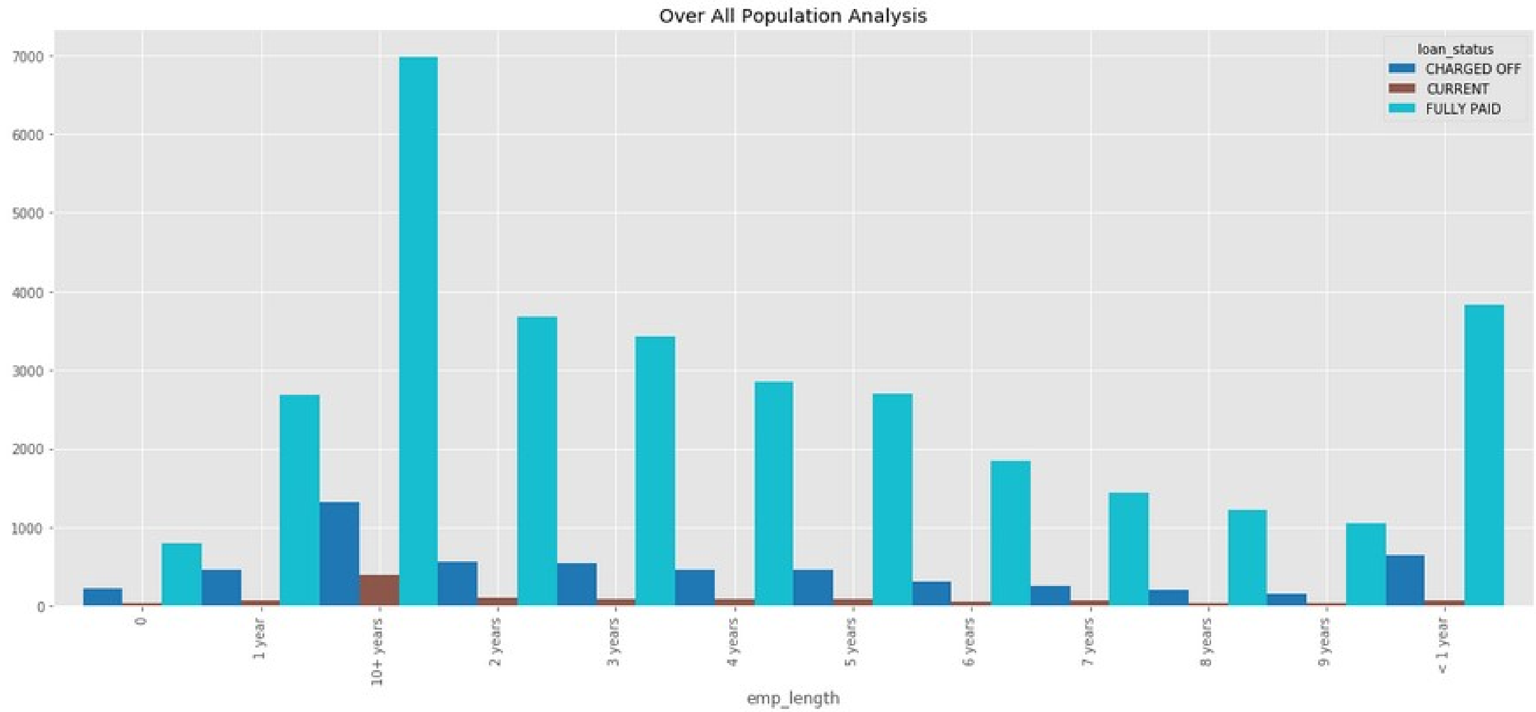




<u>Observations</u>

- Within a category LC verified applicants are defaulting more
- In over all population Not verified is contributing more defaults.

Bivariate Analysis



<u>Observations</u>

- People with emp_length 10+ years have highest count of loan_status as 'Charged_Off'
- People with emp_length <1 year have second highest count of loan_status as 'Charged_Off'

Conclusions

As Business would like to reject the loan application which has high risk of defaulting the loan for achieving the same following attributes are recommended to use while scrutinising an loan applicaiton

Attributes:

1. Emp_length (working experience of the applicant): Its observed that applicant of experience of 10+, 1 and 2 loans are charged off more.
2. Varification status : As analysis of verified category resulted more defaults than other categories.
3. Sub-grade/grade: As this attribute is equivalent presentation on interest. The analysis shown that increasing rates increases the default s.
4. Purpose: Analysis revealed that purpose contributes to the charged off.
5. Term: Analysis revealed that term will impact charged off rate positivly