Case Study: Lending Club Case Study

Submitted By: Utkarsh Chaturvedi

Problem Statement:

Business Understanding - Lending Club Case Study

We are working for a consumer finance company which specializes in lending various types of loans to urban customers. When we receive a loan application, we have to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with our 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

Business Objective

The objective is 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 need to research about risk analytics by understanding the types of variables and their significance should be enough.

Analysis Approach:

The following will the approach that will be followed to analyze the data and then provide a solution:

1. Understand the data properly
2. Clean and manipulate data as needed
3. Data analysis
4. Provide result of analysis in form of visualization

Analysis Results:

The first exercise is to analyze all the data that has been provided and only consider the data that is relevant. The objective is to find the applicants which have strong probability of defaulting and also to identify the applicants which can repay their loan.

The objective is that we want to know which loan applications are risky. So, the fields that that are created after a loan application is approved will not be part of our analysis. Also, I have considered some customer related fields that are relevant and will provide us with intelligence related to demographics.

Based upon that I have identified that the following data is relevant for our analysis:

- loan_amnt
- funded_amnt
- funded_amnt_inv
- term
- int_rate
- installment
- grade
- sub_grade

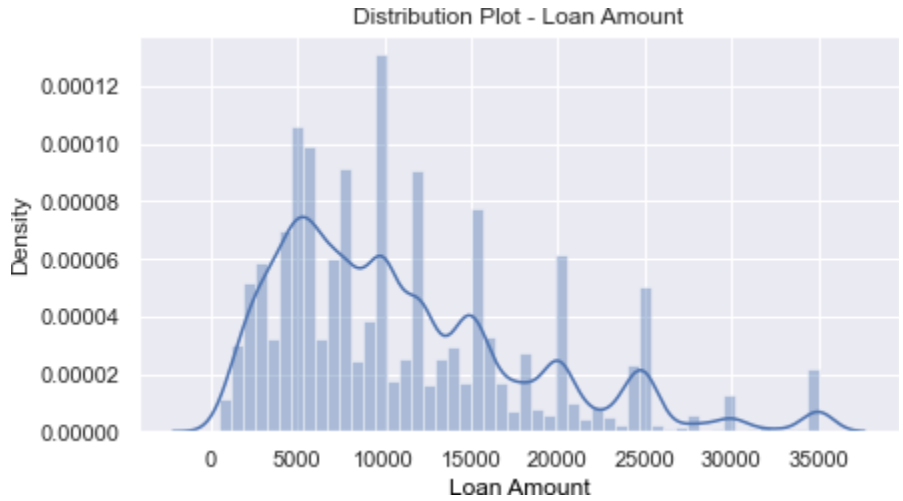
- emp_title
- emp_length
- home_ownership
- annual_inc
- verification_status
- issue_d
- loan_status
- url
- desc
- purpose
- title
- zip_code
- addr_state
- dti
- addr_state
- purpose
- loan_amnt
- funded_amnt
- funded_amnt_inv
- dti
- term

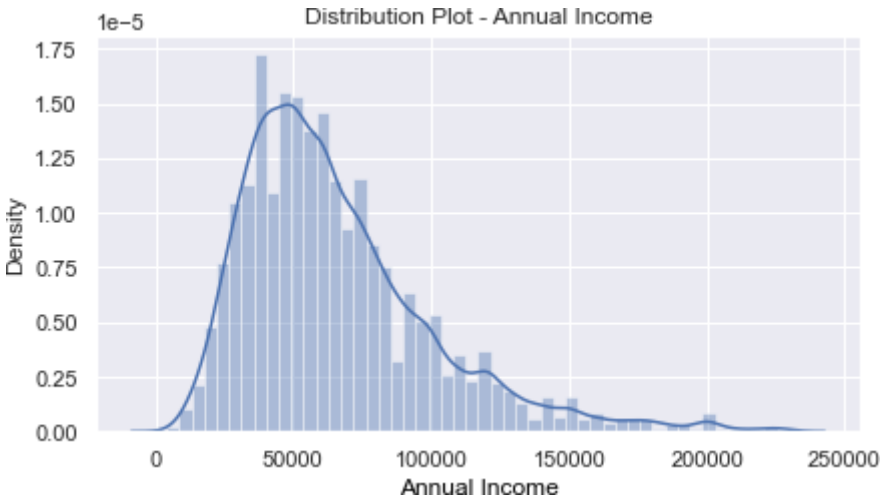
Data was cleaned:

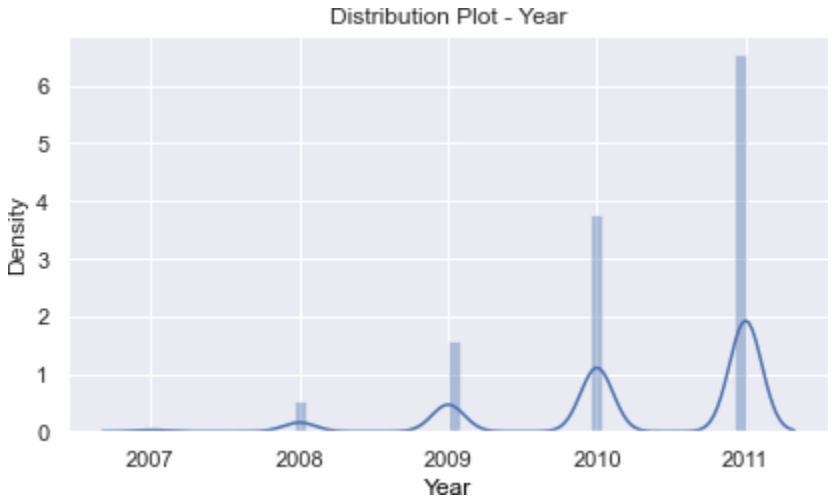
1. Removed blank entries
2. Did some manipulation on emp_length, int_rate, term and zip code fields
3. Created bins to be further used during bivariate analysis.

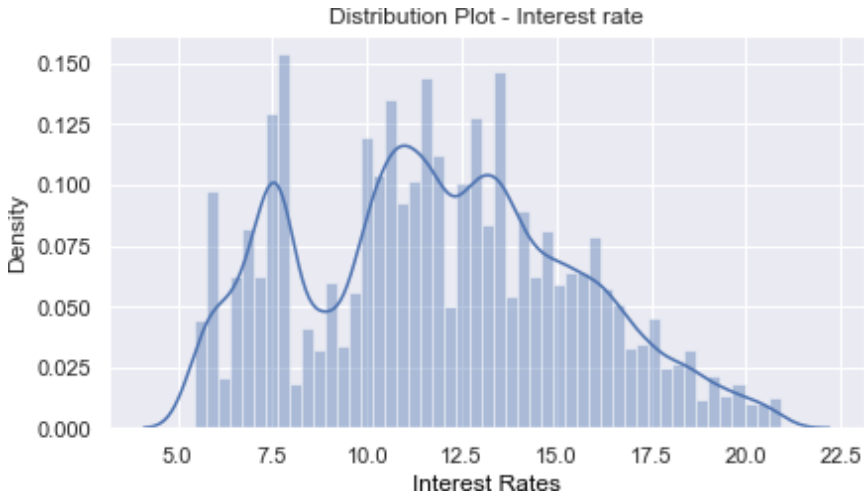
Univariant Analysis:

1. Outlier analysis was done using box plot on int_rate and annual_inc
2. Univariant Analysis on Loan Amount, Annual Income, Loan Issue Year (Derived metrics), Interest Rate, Loan Status, year(derived)




Distribution Plot - Annual Income
1e−5
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Density
0
50000
100000
150000
200000
250000
Annual Income


Distribution Plot - Year
6
5
4
3
2
1
0
Density
2007
2008
2009
2010
2011
Year


Distribution Plot - Interest rate
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Interest Rates

Based upon the above distribution plots the following are the conclusions based upon **Univariant analysis:**

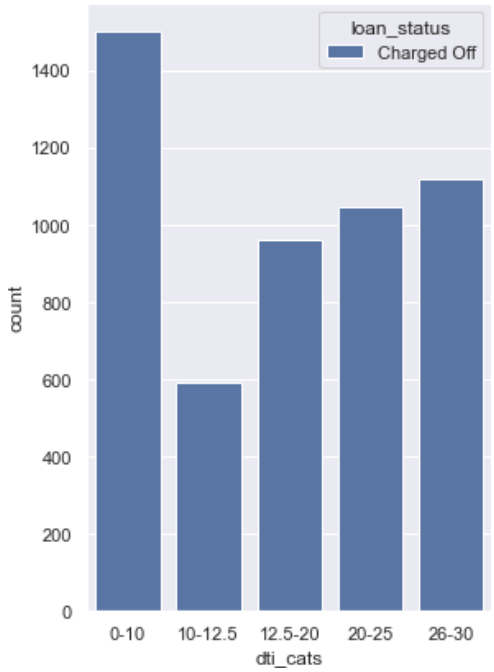
1. Most loans that are being taken are between $5000 - $10000
2. Annual income of customers that are taking loan is around $50000 range
3. Numbers of loans distributed are increasing by each year. Year 2011 had the greatest number of loans disbursed.
4. Interest rates are mostly between 10% - 15%

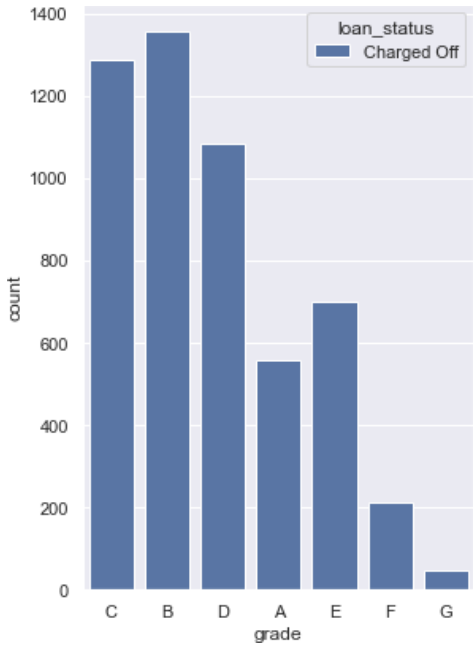
Bivariant Analysis:

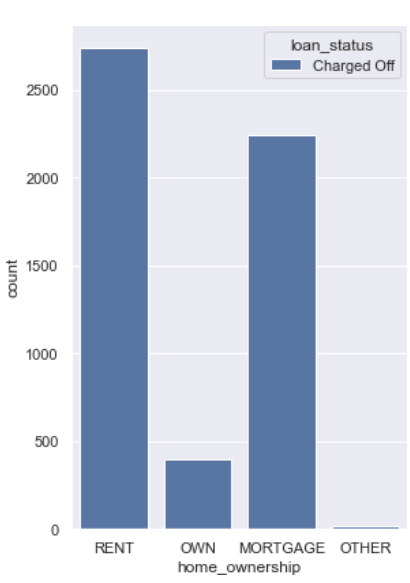
For Bivariant Analysis we will focus on the loan status of "Charged Off" so that we can analyze which variables have more impact. Based upon this analysis bank can make intelligent decision during loan approvals.

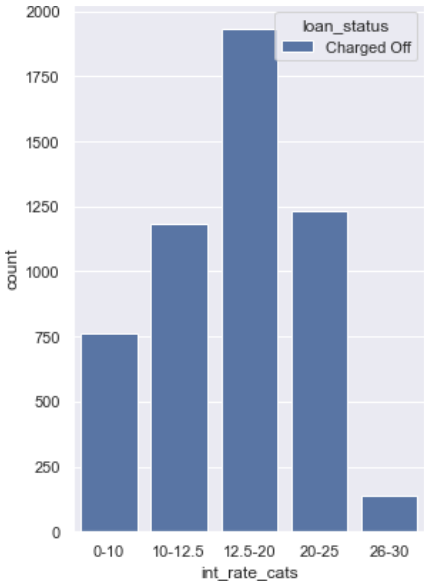
Bivariant analysis was done on the following variables:

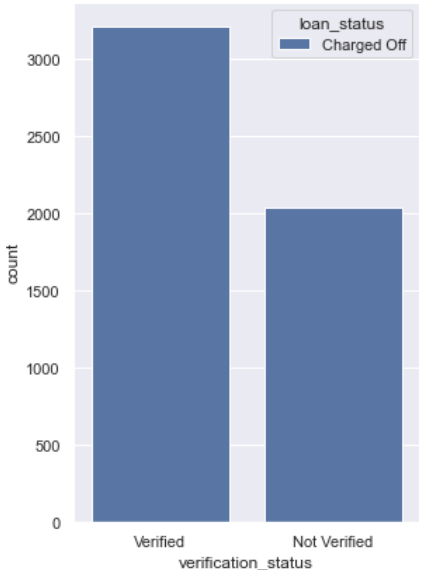
1. Loan Status (Charged off) with Grade
2. Loan Status (Charged off) with Verification Status
3. Loan Status (Charged off) with Sub Grade
4. Loan Status (Charged off) with Interest rate categories
5. Loan Status (Charged off) with dti categories
6. Loan Status (Charged off) with home ownership categories
7. Loan Status (Charged off) with state categories

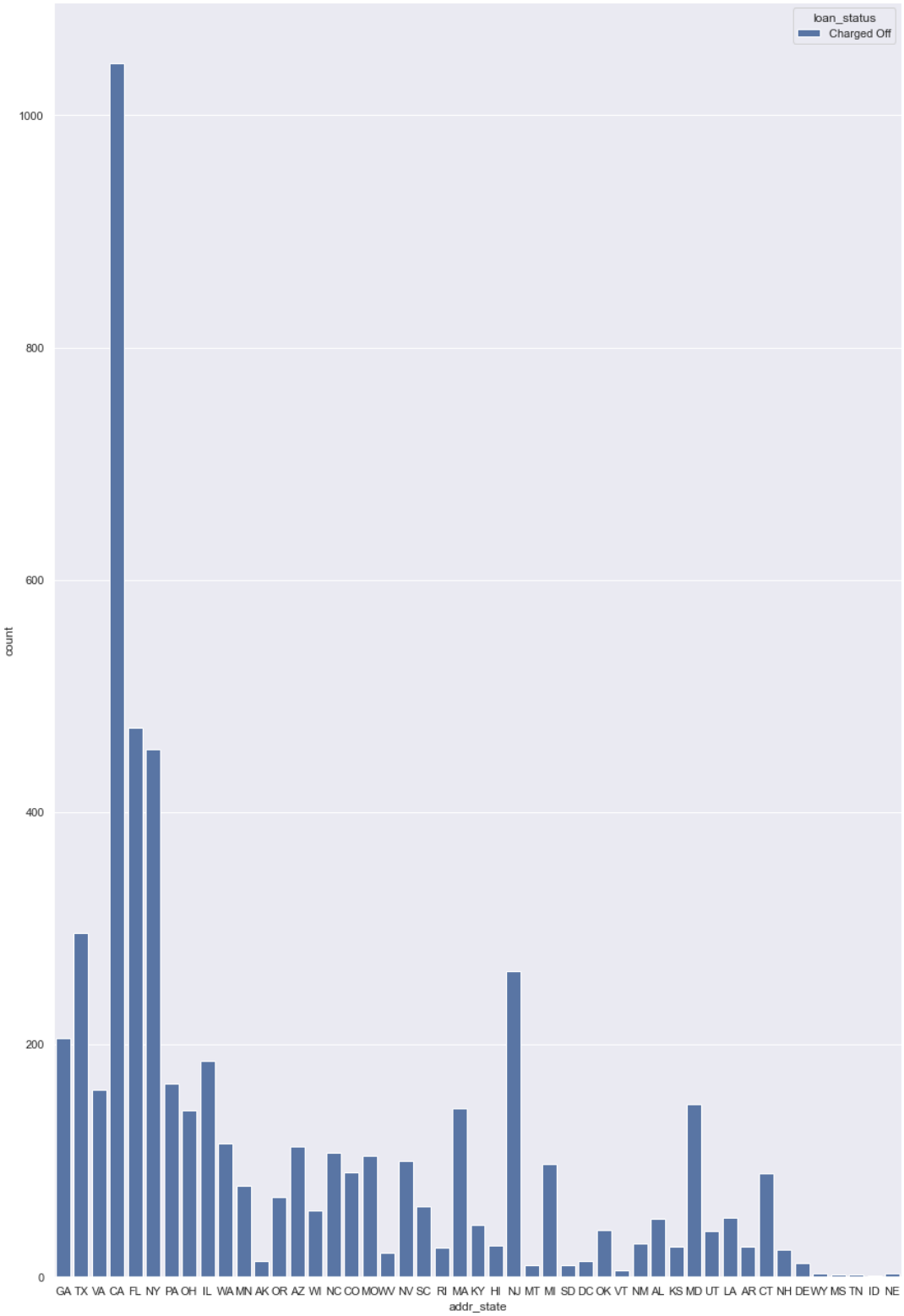

loan_status
Charged Off
count
dti_cats
0-10
10-12.5
12.5-20
20-25
26-30

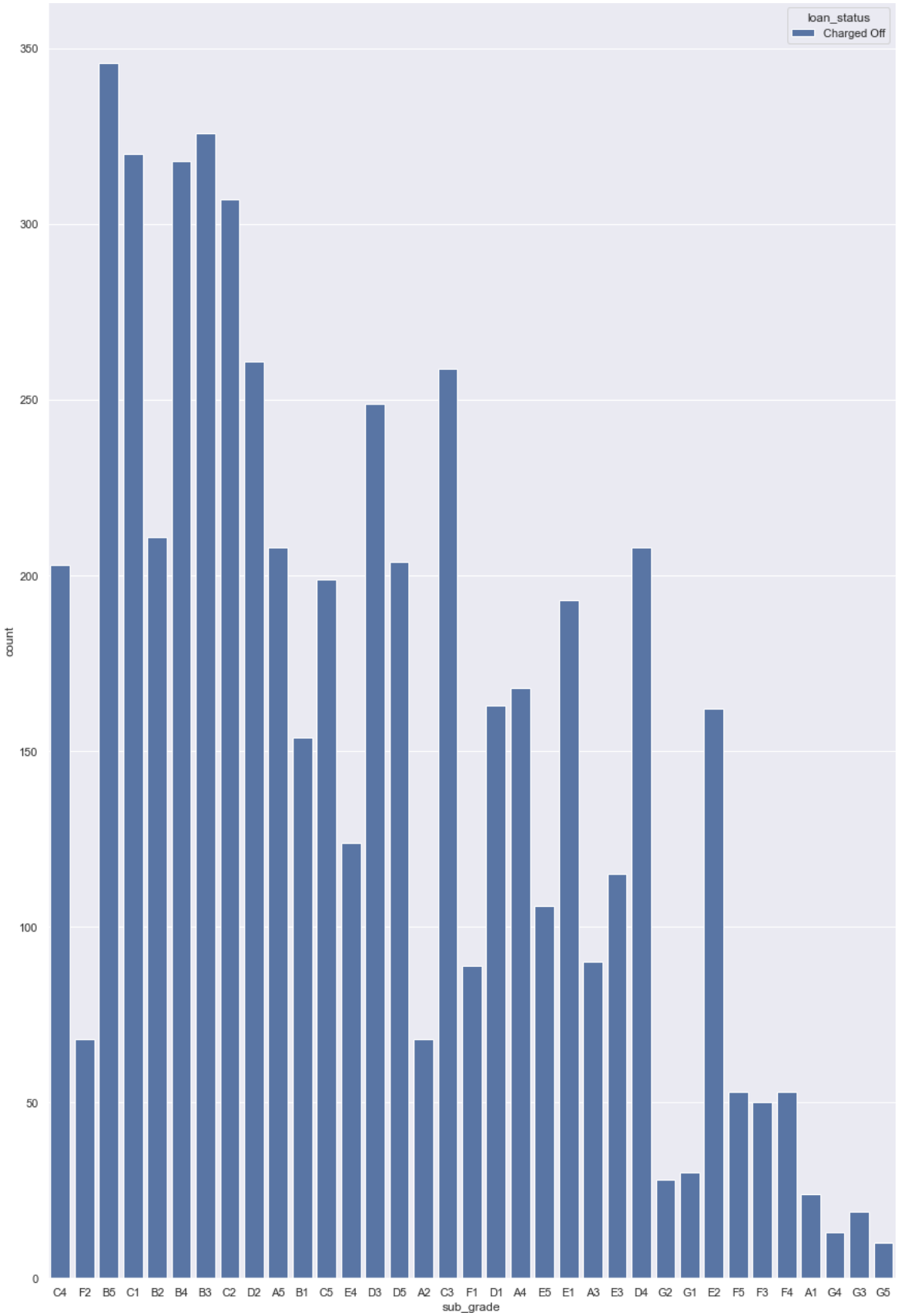

loan_status
Charged Off
count
grade
C
B
D
A
E
F
G


loan_status
Charged Off
count
home_ownership
RENT
OWN
MORTGAGE
OTHER


loan_status
Charged Off
count
int_rate_cats
0-10
10-12.5
12.5-20
20-25
26-30


loan_status
Charged Off
3000
2500
2000
1500
1000
500
0
count
Verified
Not Verified
verification_status


loan_status
Charged Off
1000
800
600
400
200
0
count
GA TX VA CA FL NY PA OH IL WA MN AK OR AZ WI NC CO MO WV NV SC RI MA KY HI NJ MT MI SD DC OK VT NM AL KS MD UT LA AR CT NH DE WY MS TN ID NE
addr_state

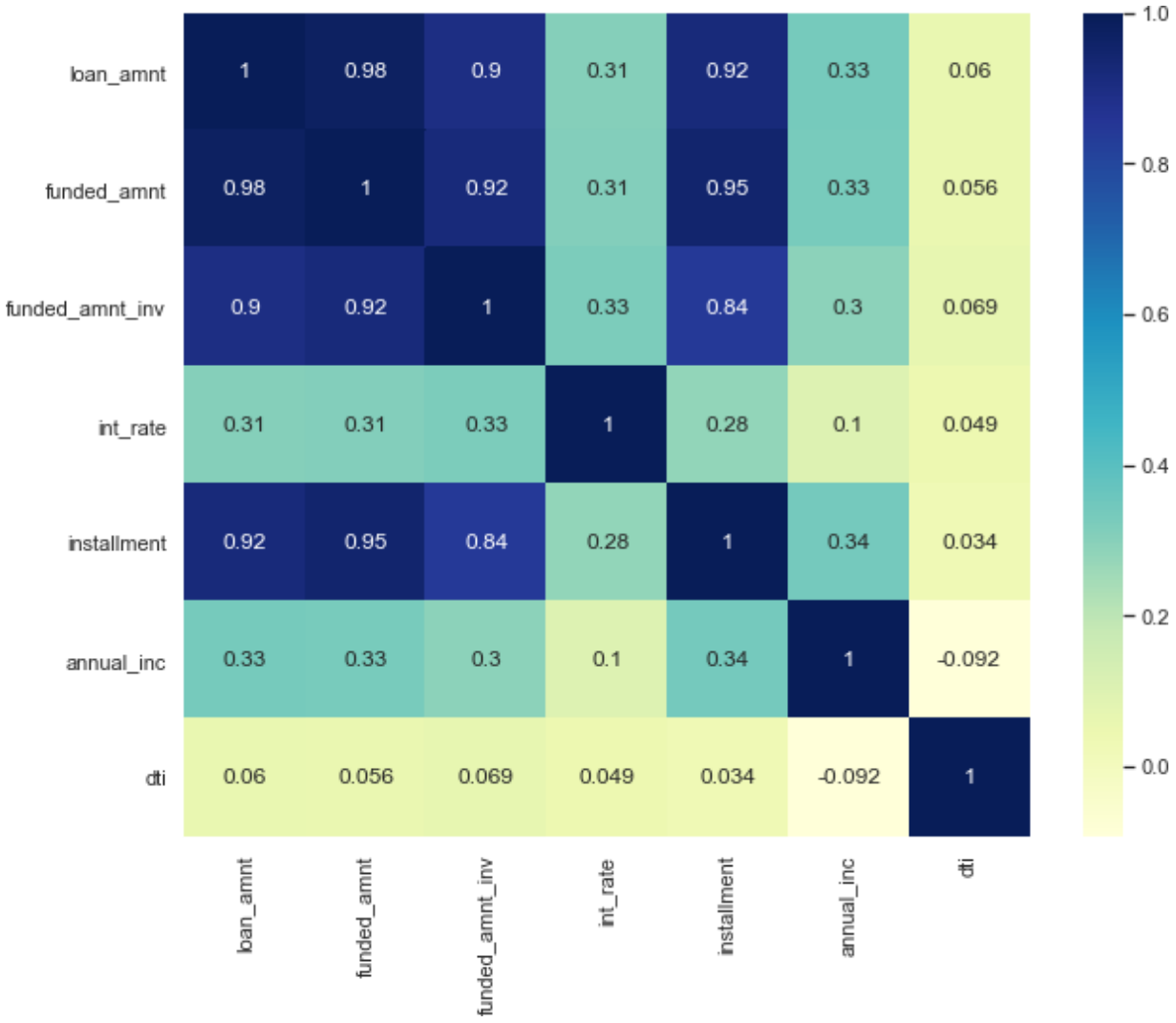

loan_status
Charged Off
350
300
250
200
150
100
50
0
count
C4 F2 B5 C1 B2 B4 B3 C2 D2 A5 B1 C5 E4 D3 D5 A2 C3 F1 D1 A4 E5 E1 A3 E3 D4 G2 G1 E2 F5 F3 F4 A1 G4 G3 G5
sub_grade

Based upon the above count plots the following are the conclusions based upon **Bivariant analysis:**

1. Customers that were verified have more chances of loan being charged off
2. Customers with sub grade B3, B4 and B5 have very more chances of loan being charged off
3. Customers with Grade G have less chances of loan being charged off
4. Customers with interest rates between 12.5% - 16% have high chances of loan being charged off
5. Customers with dti between less than 10% have high chances of loan being charged off
6. Customers who have rented accommodation have more chances of loan being charged off
7. Customers that are living in California have high chances of loan being charged off

Multivariant Analysis:

Multivariant analysis has been done using heatmap.



Based upon the above heat map the following are the conclusions based upon **Multivariant analysis:**

1. Loan amounts have high correlation with Funded amount, Funded amount committed by investor and installment
2. Funded amounts have high correlation with Loan amount, Funded amount committed by investor and installment
3. Funded amount committed by investor have high correlation with Loan amount, Funded amount and installment