Lending Club Case Study

Group Members:
Jahana Shirin
Shasank Shah

Problem Statement

- To understand risk analytics in banking and financial services to minimise the risk of losing money while lending to customers.
- Two types of risks are associated with the bank's decision:
    I. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    II. 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
- The company which provided the data wants to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default so that the company can utilise this knowledge for its portfolio and risk assessment.

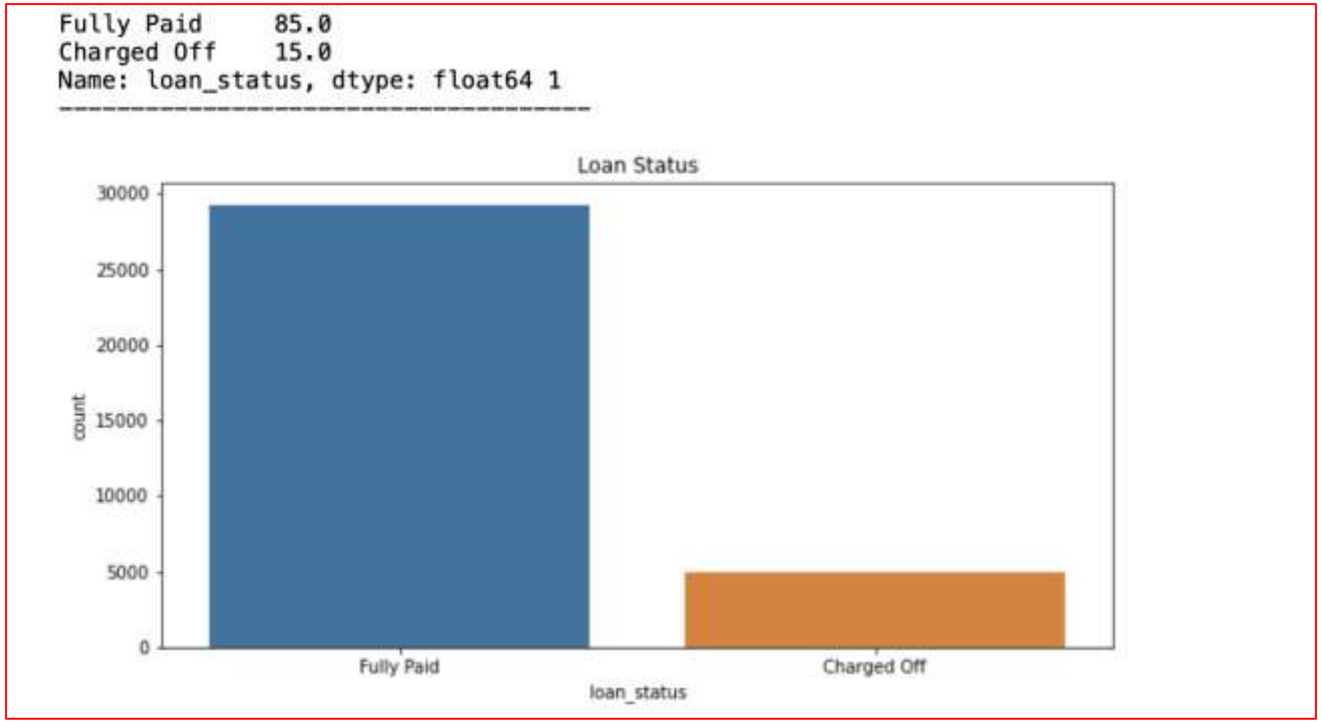
We will use Exploratory Data Analysis(EDA) to understand how consumer attributes and loan attributes influence the tendency of default.

Problem solving methodology

- Data Cleaning
  1. Removing NULL columns, unnecessary variables, special characters such as % from interest rate column, stripping whitespace for the entire data frame and finally considering only ‘Charged Off’ & ‘Fully Paid’ borrowers.
- Univariate Analysis
  1. Analyzing columns and plotting the distributions.
- Segmented Univariate analysis
  1. Analyzing continuous data columns with respect to categorical column
- Bivariate Analysis
  1. Analyzing two variables behavior like term & loan status
- Derived Metrics
  1. To create new variables using existing ones and get meaningful information by analysing them

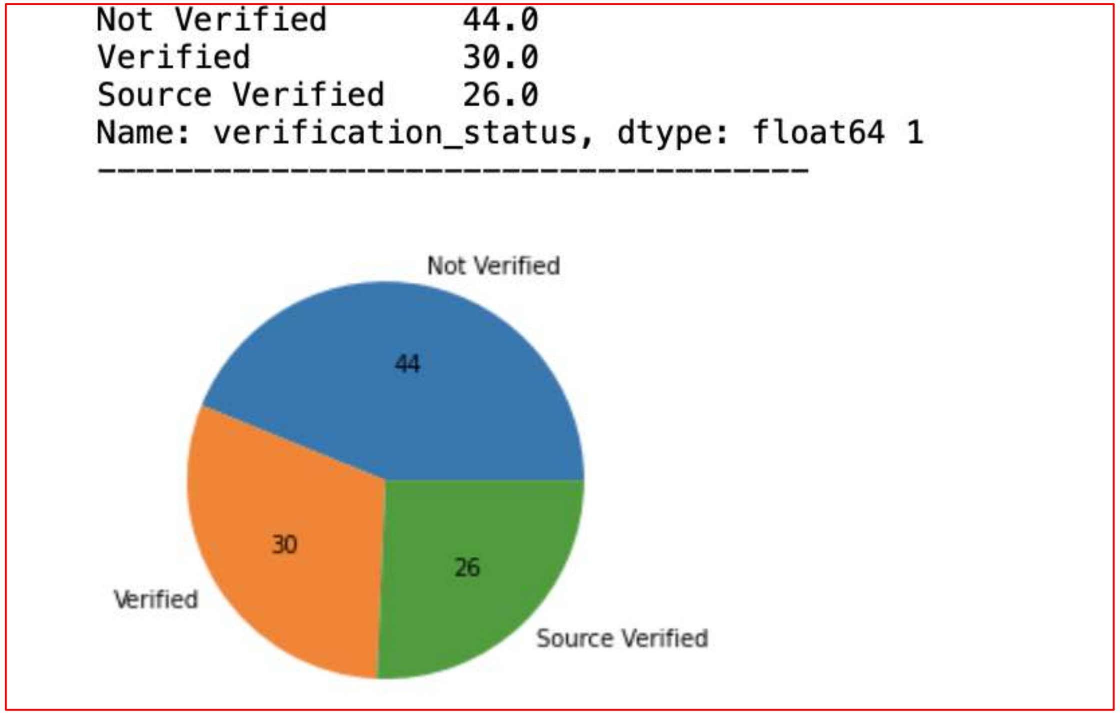
Univariate Analysis - Unordered Categorical Variable

It can be observed from the above that 85% of the applicants have paid off the loan whereas 15% have defaulted



```
Fully Paid     85.0
Charged Off    15.0
Name: loan_status, dtype: float64 1
-------------------------------------
```

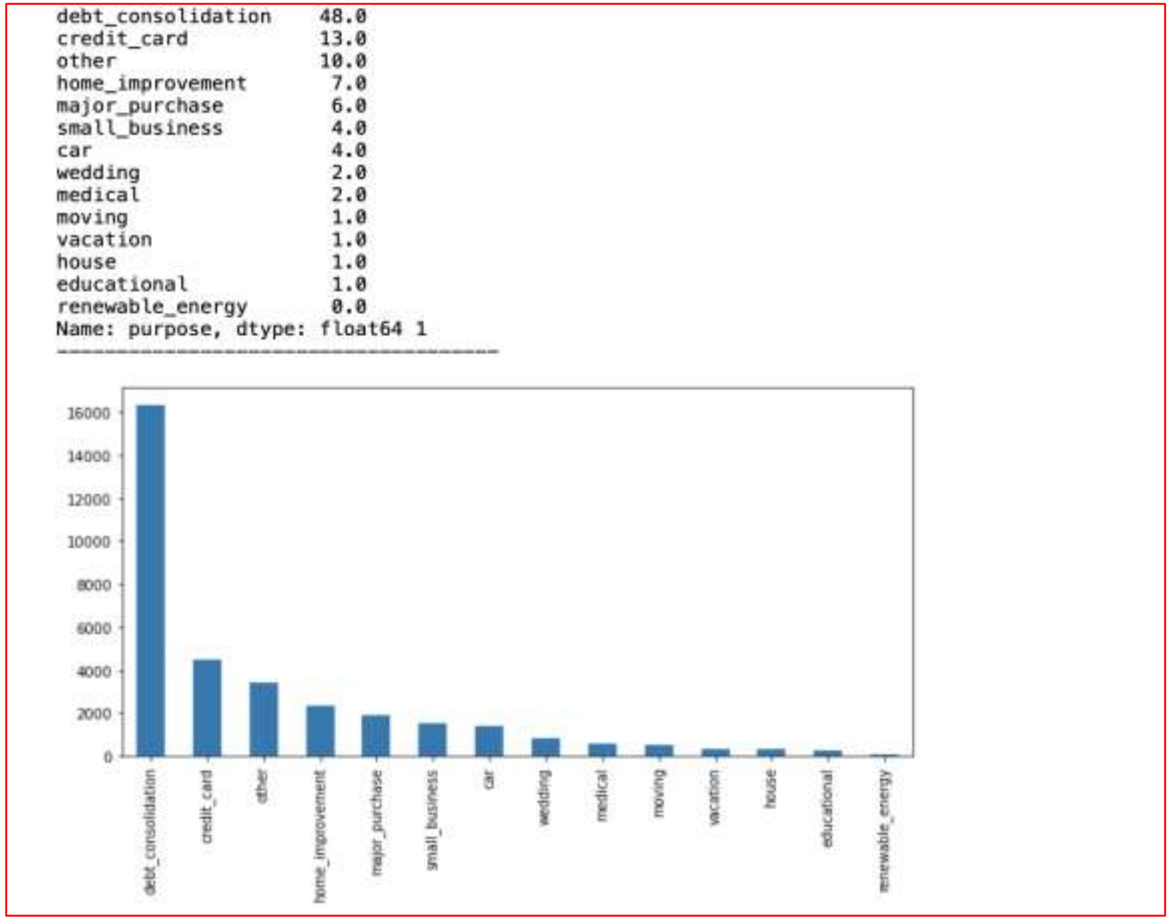
It's an interesting inference from the pie-plot we see here, 44% of the applicants are not verified before lending the loan amount. This can be one of the reasons for increasing defaulters.



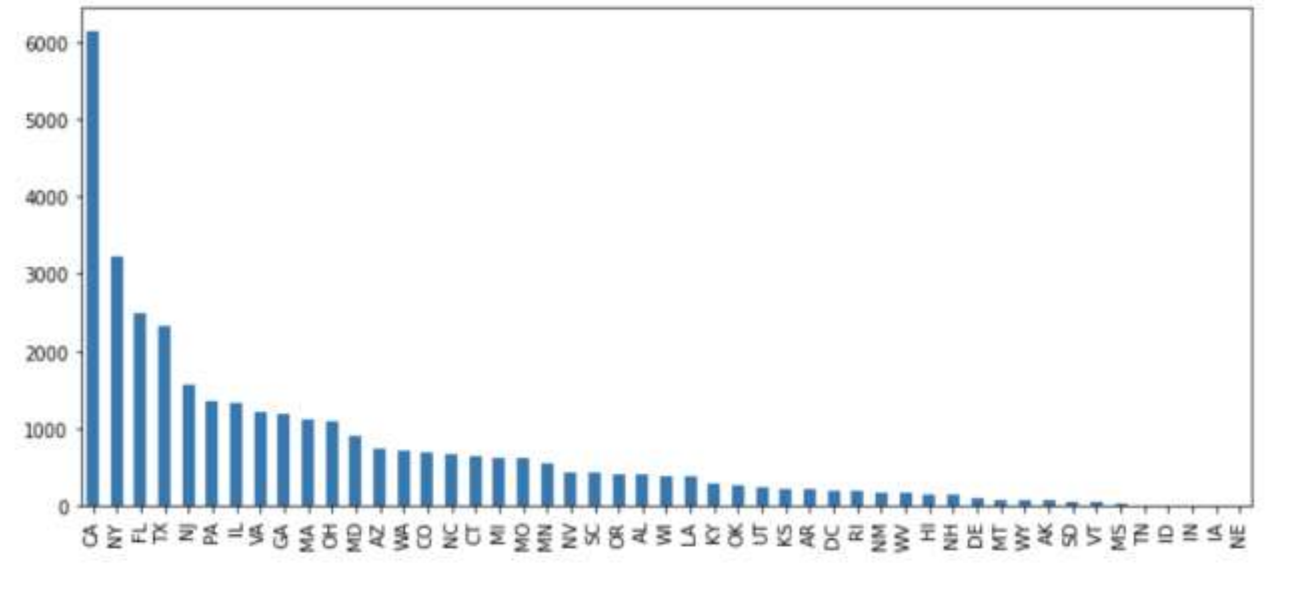
```
Not Verified       44.0
Verified           30.0
Source Verified    26.0
Name: verification_status, dtype: float64 1
--------------------------------------
```

Univariate Analysis - Ordered Categorical Variable

It can be observed from the graph that 48% of the applicants have taken the loan for debt consolidation.



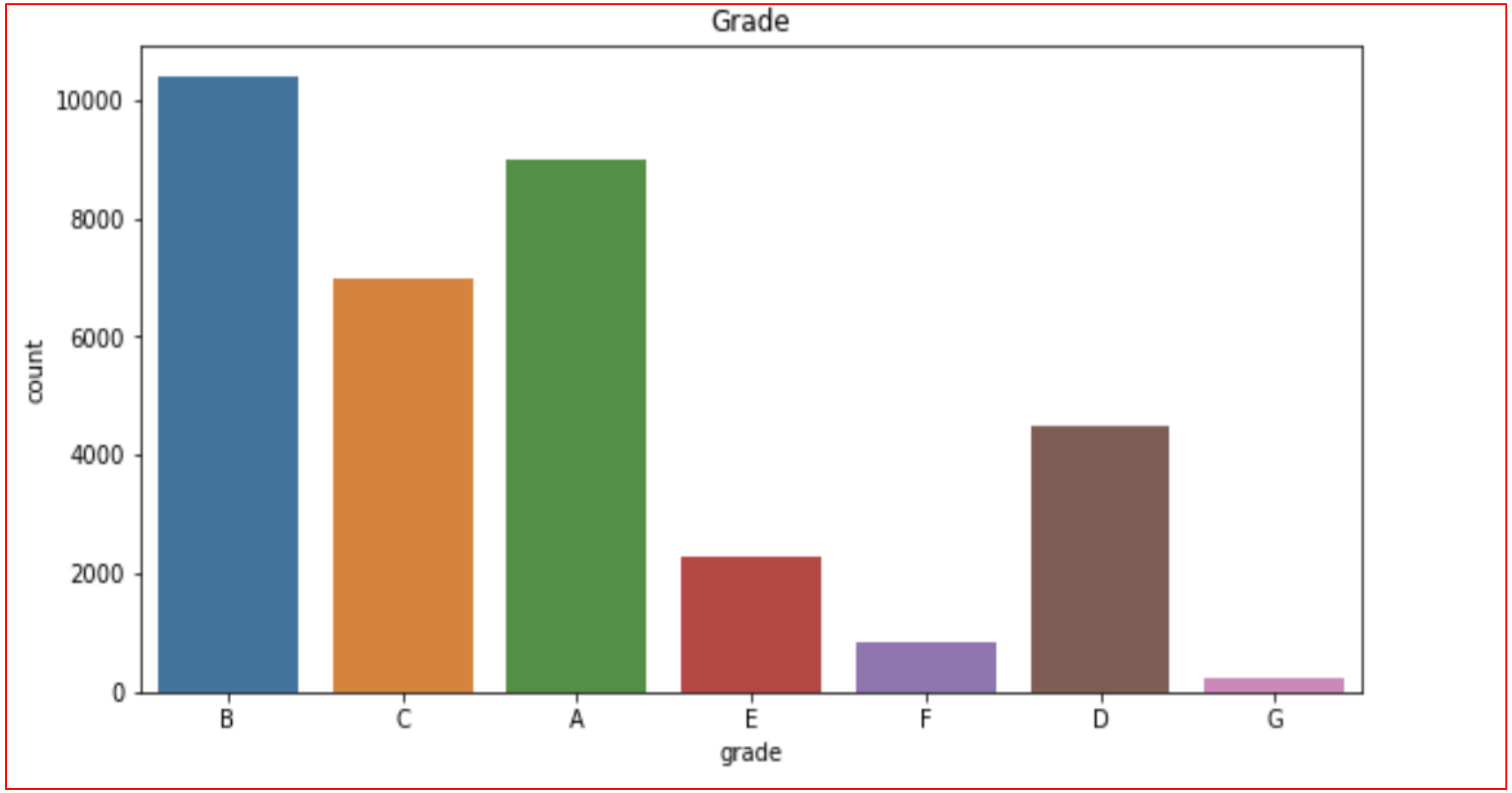
```
debt_consolidation    48.0
credit_card           13.0
other                 10.0
home_improvement       7.0
major_purchase         6.0
small_business         4.0
car                    4.0
wedding                2.0
medical                2.0
moving                 1.0
vacation               1.0
house                  1.0
educational            1.0
renewable_energy       0.0
Name: purpose, dtype: float64 1
----------------------------------------
```



California has largest count for loan applicants.

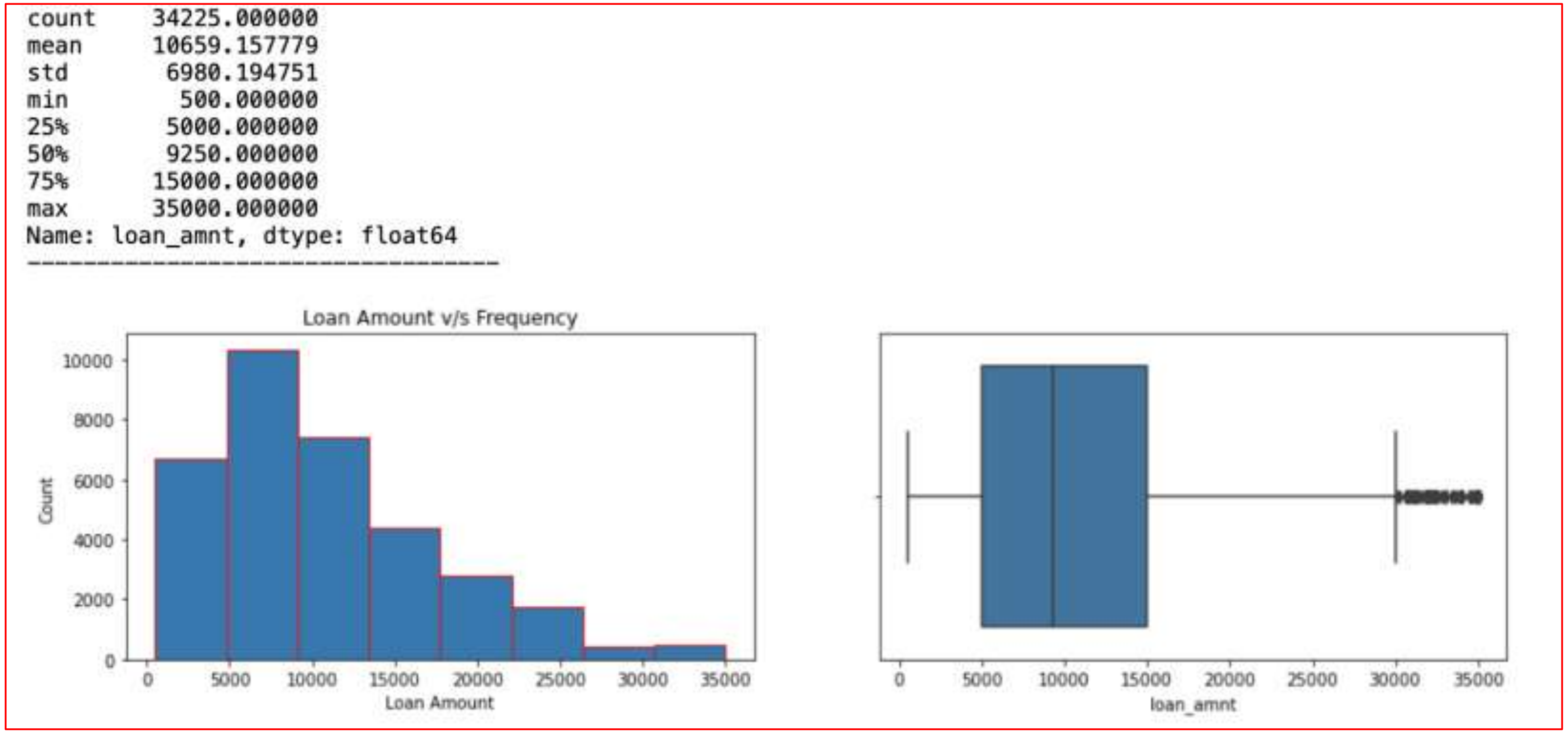
Univariate Analysis - Ordered Categorical Variable

Loan grading is a classification system that involves assigning a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest. Generally, a corresponds to High Quality while G is of the Least Quality. In the provided data, most of applications belong to A, B, C grades.



Univariate Analysis – Quantitative Variables

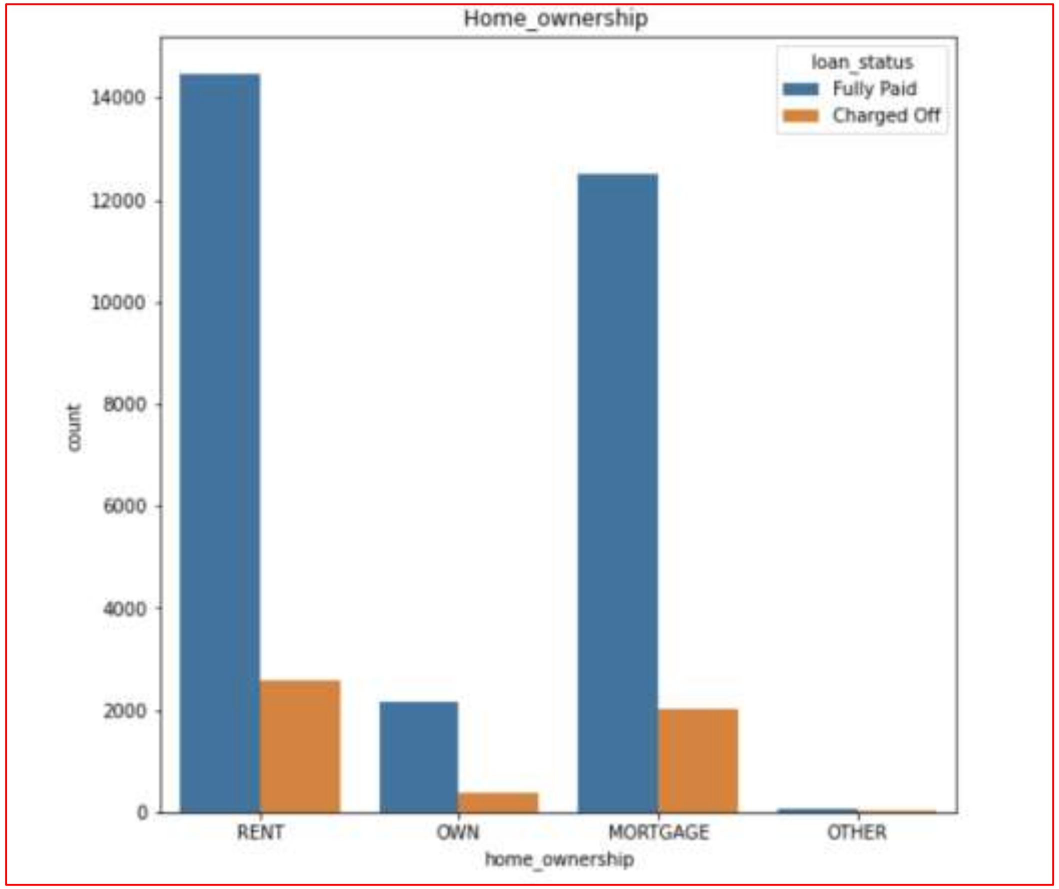
Most of the applicants have requested for a loan amount between 6000 to 15000 approximately.



```
count    34225.000000
mean     10659.157779
std       6980.194751
min        500.000000
25%       5000.000000
50%       9250.000000
75%      15000.000000
max      35000.000000
Name: loan_amnt, dtype: float64
-----------------------------------
```

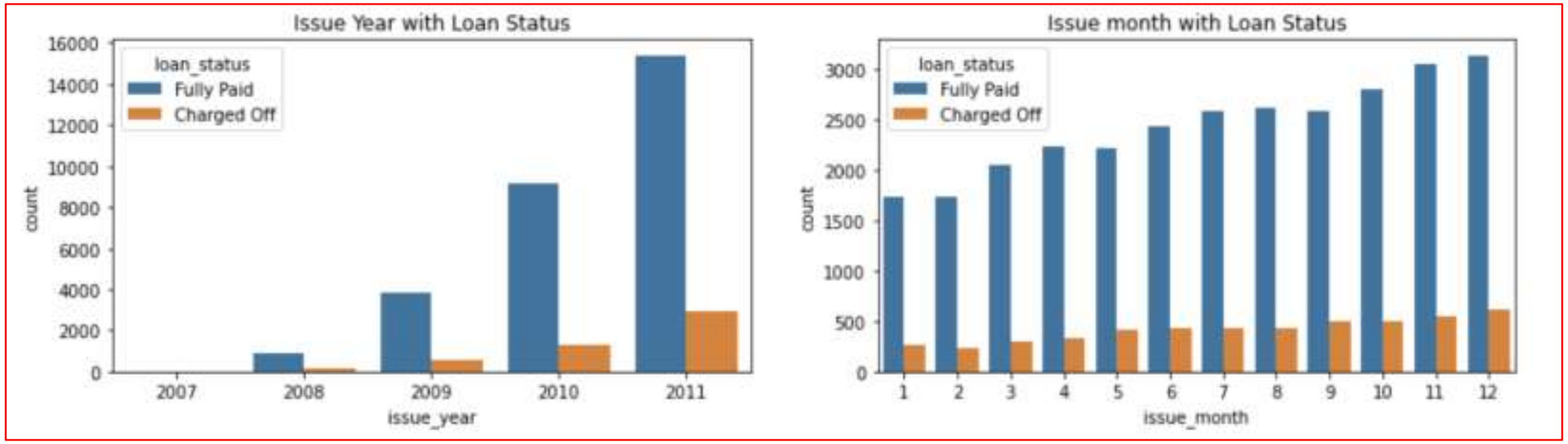
Segmented Univariate Analysis

Majority of the defaulters have rented or mortgaged homes.



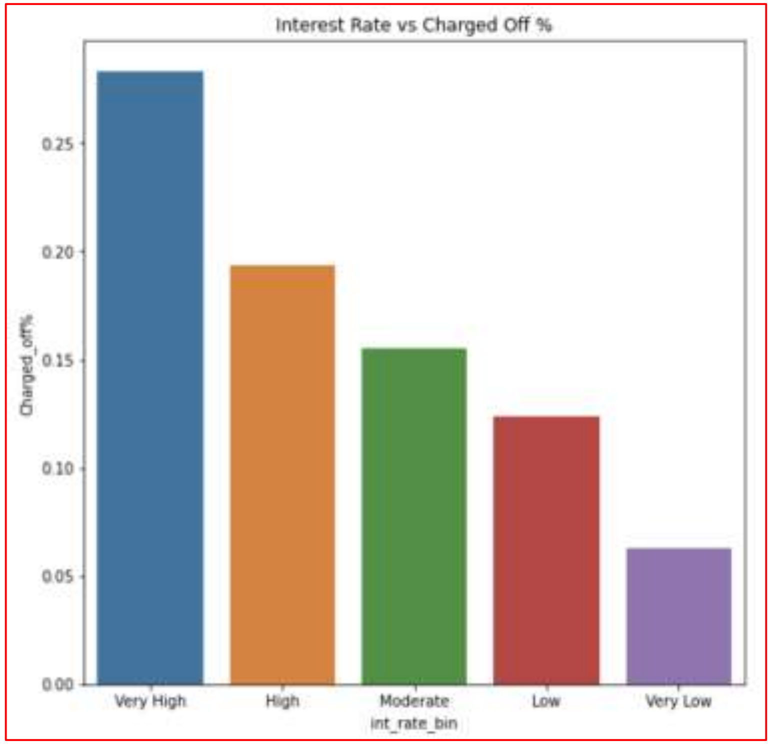
Derived Column Analysis

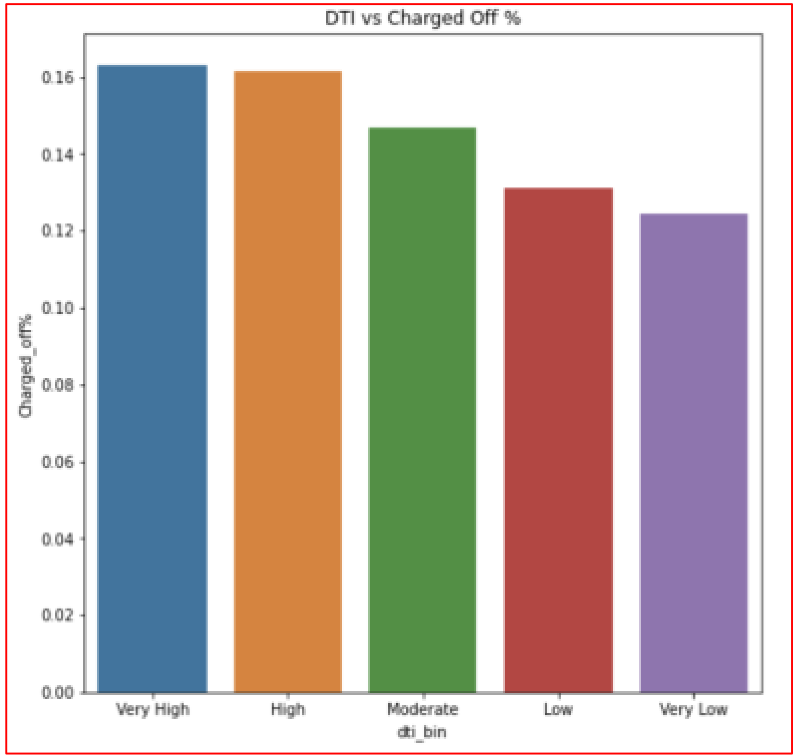
We can infer from the above plot that most of the defaulting have occurred in the year 2011. However, the number of loan applicants is also more in 2011. We can observe a gradual increase in the applicant count from 2007 t0 2011. Similar is the case for issues month as we observe that most of the issuance have occurred in the last few months i.e., October, November, December.

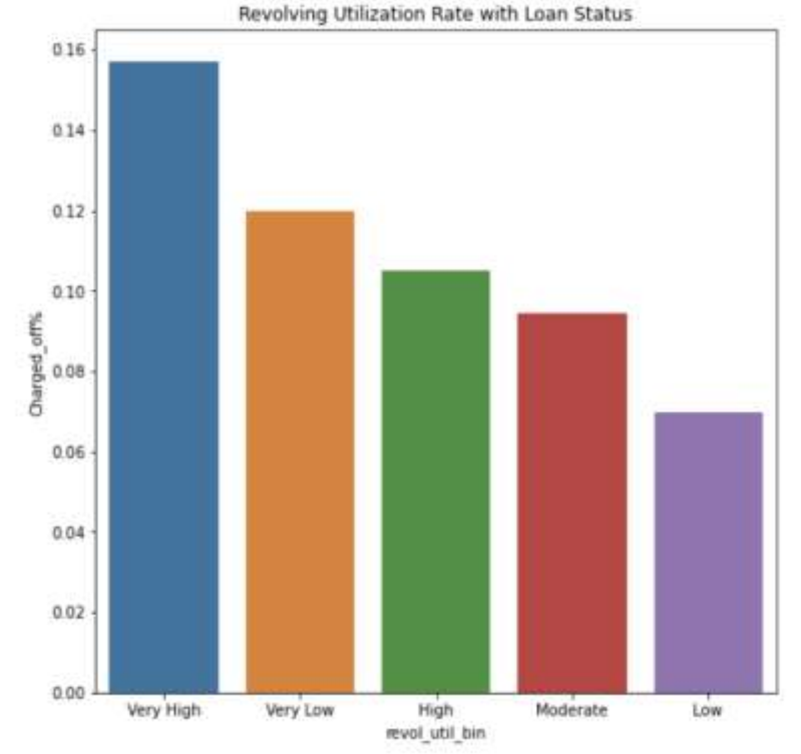


Bivariate Analysis

Defaulters had a high in Interest Rate(s), DTI & Revolving Utilization Rate



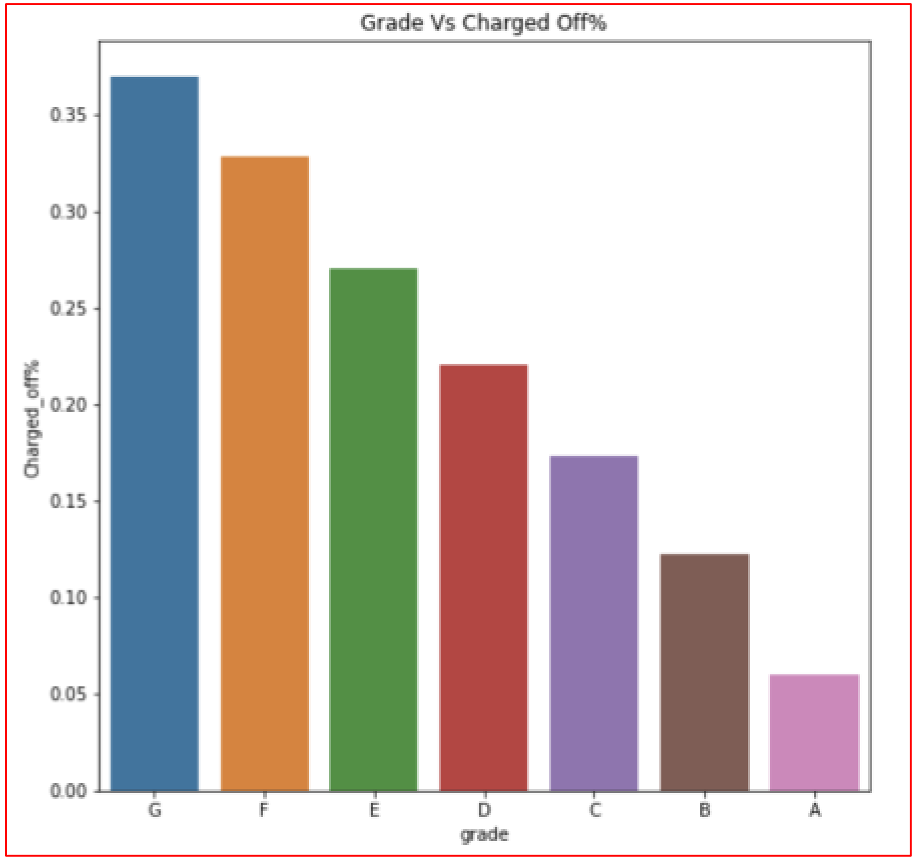


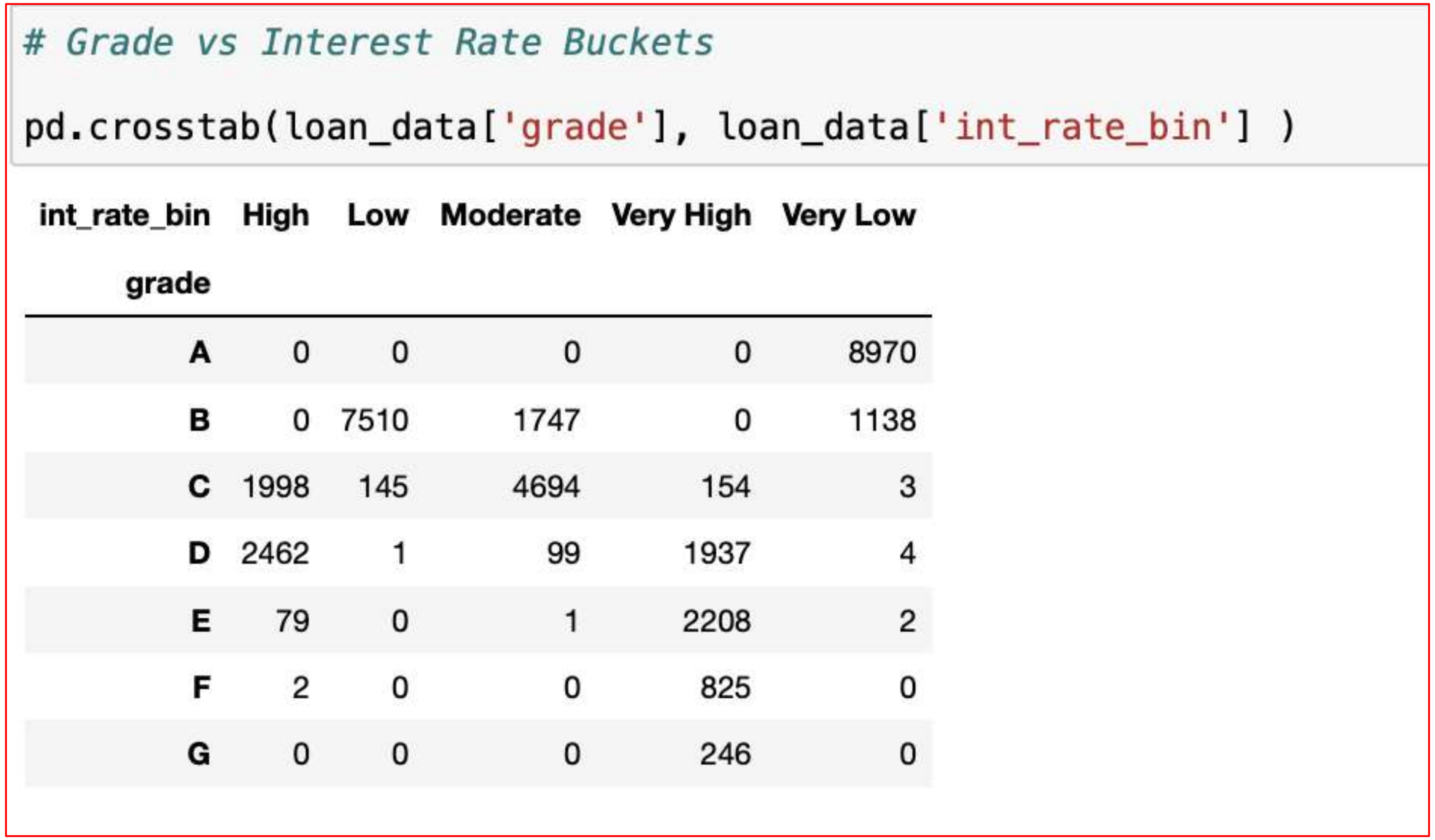


Bivariate Analysis

Lower grades have higher defaulting rate.

Very high interest rates have low grades. This indicates that the quality of the applications are low for very high interest rate. Defaulting rate is also high for very high interest rates as we have observed earlier.





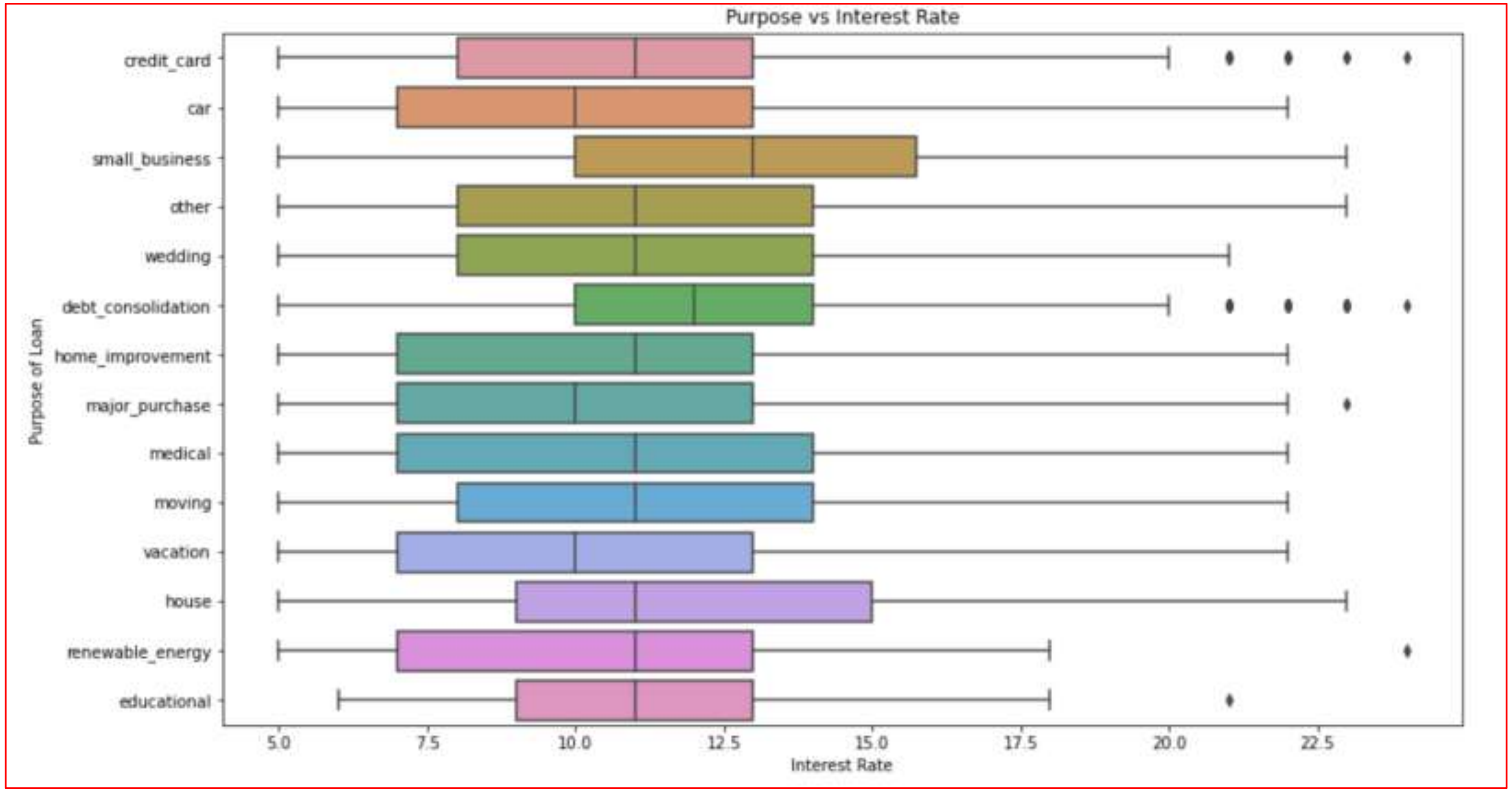
```
# Grade vs Interest Rate Buckets

pd.crosstab(loan_data['grade'], loan_data['int_rate_bin'] )
```

int_rate_bin	High	Low	Moderate	Very High	Very Low
grade					
A	0	0	0	0	8970
B	0	7510	1747	0	1138
C	1998	145	4694	154	3
D	2462	1	99	1937	4
E	79	0	1	2208	2
F	2	0	0	825	0
G	0	0	0	246	0

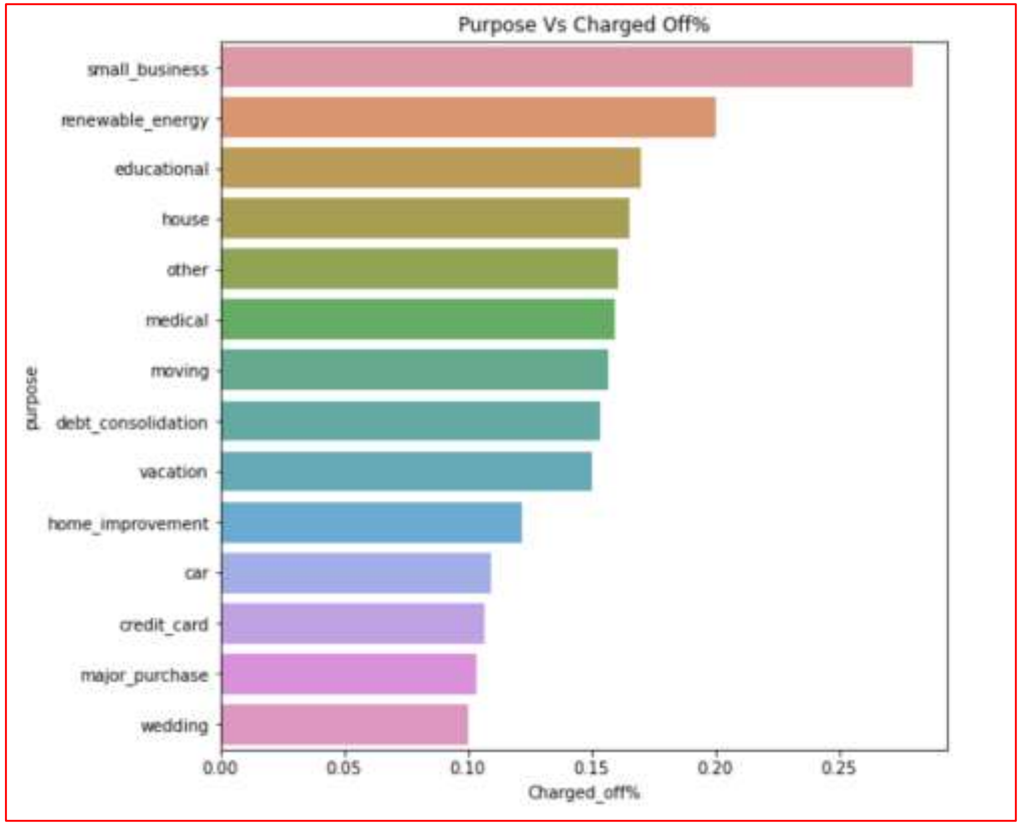
Bivariate Analysis

Small businesses have higher interest rate compared to others.



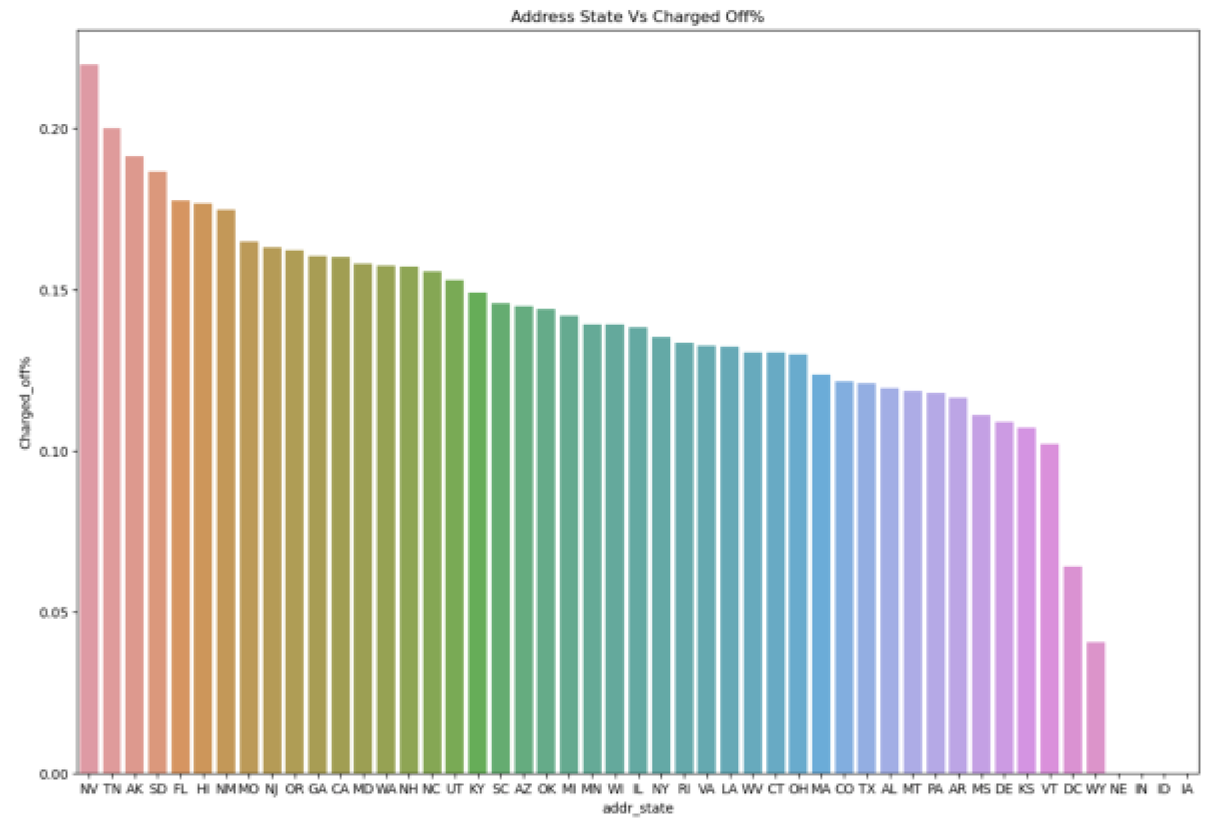
Bivariate Analysis

Highest percentage of defaulters belong to small business category followed by renewable energy. As we have already observed above, small businesses were also having higher interest rate.



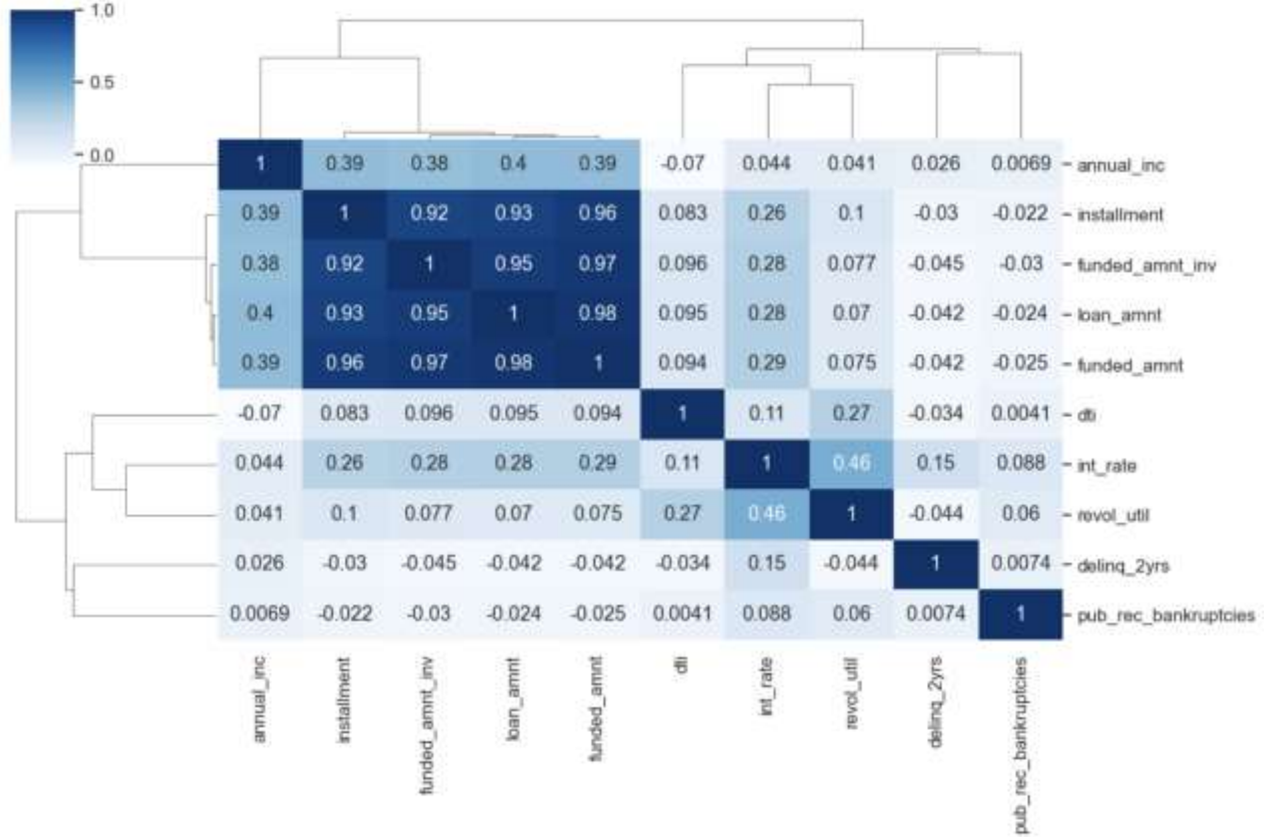
Bivariate Analysis

Charged off percentage is highest for Nevada.



Correlation Matrix

1. Loan amount, funded amount, investor amount & instalment has strong correlation.
2. Debt-to-income ratio with annual income is negative correlation. Increase in annual income decreases DTI.



	annual_inc	installment	funded_amnt_inv	loan_amnt	funded_amnt	dti	int_rate	revol_util	delinq_2yrs	pub_rec_bankruptcies
annual_inc	1	0.39	0.38	0.4	0.39	-0.07	0.044	0.041	0.026	0.0069
installment	0.39	1	0.92	0.93	0.96	0.083	0.26	0.1	-0.03	-0.022
funded_amnt_inv	0.38	0.92	1	0.95	0.97	0.096	0.28	0.077	-0.045	-0.03
loan_amnt	0.4	0.93	0.95	1	0.98	0.095	0.28	0.07	-0.042	-0.024
funded_amnt	0.39	0.96	0.97	0.98	1	0.094	0.29	0.075	-0.042	-0.025
dti	-0.07	0.083	0.096	0.095	0.094	1	0.11	0.27	-0.034	0.0041
int_rate	0.044	0.26	0.28	0.28	0.29	0.11	1	0.46	0.15	0.088
revol_util	0.041	0.1	0.077	0.07	0.075	0.27	0.46	1	-0.044	0.06
delinq_2yrs	0.026	-0.03	-0.045	-0.042	-0.042	-0.034	0.15	-0.044	1	0.0074
pub_rec_bankruptcies	0.0069	-0.022	-0.03	-0.024	-0.025	0.0041	0.088	0.06	0.0074	1

Conclusions

From the above analysis, we can conclude that below variables are the driving factors of default:

1. Higher Interest Rates
2. Low grades like F & G
3. Higher Debt-to-Income Ratio
4. Higher Revolving Utilization Rate
5. Small Business is the purpose
6. Home ownership is Rent or Mortgage
7. Lower annual income.

Other observations:

1. Nevada has highest default rate
2. 60 months term has higher Charged off percentage
3. Not verified applications have slightly higher default rate compared to verified and source verified.
4. 10+ years of employment length has most default rate followed by 7 years and 1 years.