

LENDING CLUB CASE STUDY

SUBMISSION

Name : Arpitha Upadhyaya and Vivek Reddy Chinasani

Abstract

The idea of this presentation is to show the patterns which can help the firm to provide a loan to a person or not. The objective of the firm is to analyse the past data (between 2007 to 2011) of the customers and see if any patterns in the data which can help us find out if any customer can default a loan. Borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

The following presentation, the company wants to understand the driving factors 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.



Data Understanding

- Read the csv file and do the initial check for data.
- Check for datatypes
- Identify the target columns.

Data Cleanup

- Remove null columns
- Remove columns with high percentage of null values.
- Remove rows with null values.
- Remove outlier data.

Data Analysis

- .Identify driving factors using Univariate and bivariate analysis with various plots to identify predictors of defaults
- Get rid off Current variables.
- Ignore customer behavior Variables

Conclusions

- Identify predictors of default using the various analysis points.

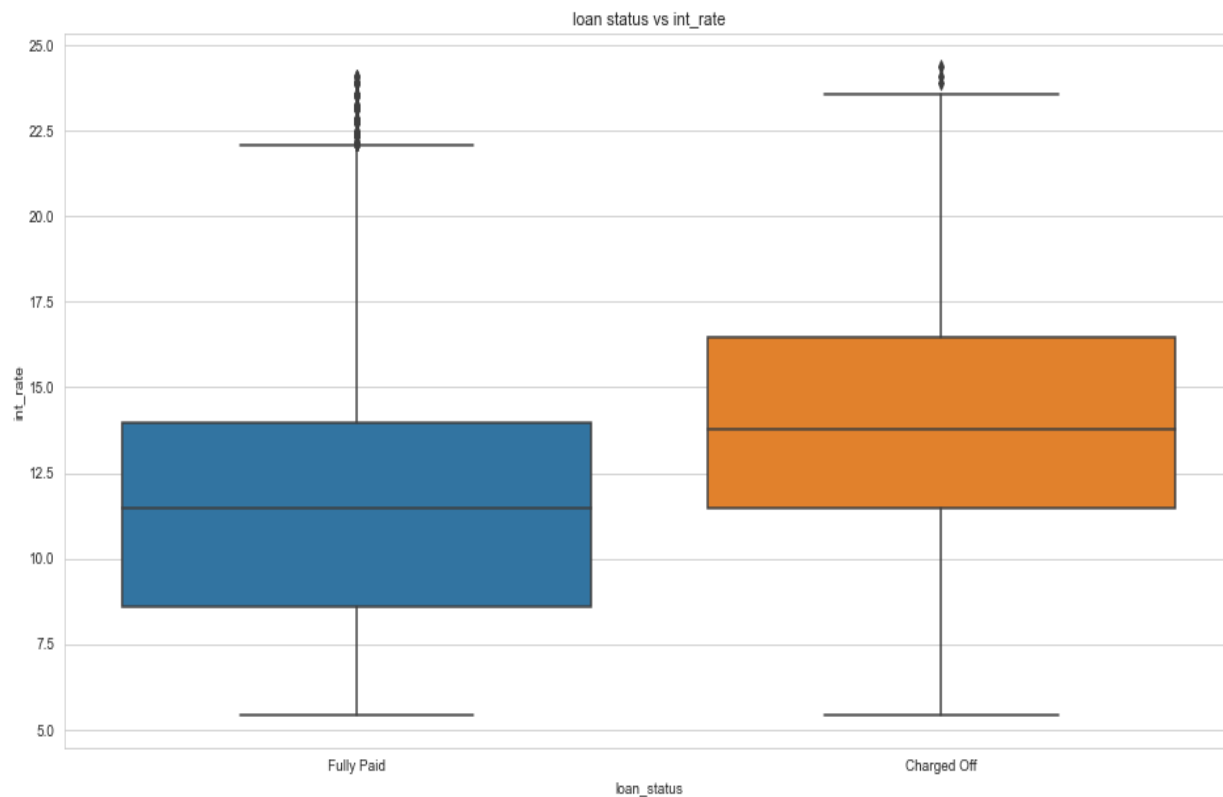
Co-relation matrix

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	annual_inc	dti	delinq_2yrs	inq_la
id	1.00	1.00	0.09	0.10	0.19	0.19	0.01	0.07	-0.03	0.04	0.07	-0.01	
member_id	1.00	1.00	0.09	0.10	0.18	0.18	0.01	0.07	-0.03	0.04	0.07	-0.01	
loan_amnt	0.09	0.09	1.00	0.99	0.95	0.32	0.24	0.96	0.23	0.42	0.07	-0.04	
funded_amnt	0.10	0.10	0.99	1.00	0.97	0.31	0.25	0.97	0.23	0.41	0.07	-0.04	
funded_amnt_inv	0.19	0.18	0.95	0.97	1.00	0.33	0.24	0.93	0.21	0.40	0.08	-0.04	
term	0.19	0.18	0.32	0.31	0.33	1.00	0.41	0.11	0.41	0.08	0.08	0.00	
int_rate	0.01	0.01	0.24	0.25	0.24	0.41	1.00	0.24	0.96	0.05	0.12	0.16	
installment	0.07	0.07	0.96	0.97	0.93	0.11	0.24	1.00	0.22	0.41	0.07	-0.03	
grade	-0.03	-0.03	0.23	0.23	0.21	0.41	0.96	0.22	1.00	0.04	0.11	0.17	
annual_inc	0.04	0.04	0.42	0.41	0.40	0.08	0.05	0.41	0.04	1.00	-0.10	0.03	
dti	0.07	0.07	0.07	0.07	0.08	0.08	0.12	0.07	0.11	-0.10	1.00	-0.04	
delinq_2yrs	-0.01	-0.01	-0.04	-0.04	-0.04	0.00	0.16	-0.03	0.17	0.03	-0.04	1.00	
inq_last_6mths	-0.04	-0.04	-0.01	-0.01	-0.02	0.05	0.17	-0.00	0.15	0.03	0.01	0.01	
open_acc	0.03	0.03	0.20	0.20	0.19	0.05	-0.03	0.19	-0.04	0.30	0.31	0.00	
pub_rec	-0.02	-0.02	-0.04	-0.04	-0.05	0.01	0.10	-0.04	0.11	-0.01	-0.00	0.01	
revol_bal	0.04	0.04	0.39	0.39	0.37	0.09	0.13	0.39	0.12	0.39	0.33	-0.09	
total_acc	0.04	0.04	0.27	0.26	0.26	0.10	-0.08	0.24	-0.08	0.43	0.24	0.07	
total_pymnt	0.08	0.08	0.88	0.89	0.87	0.26	0.21	0.87	0.20	0.40	0.06	-0.03	
total_pymnt_inv	0.15	0.15	0.85	0.86	0.90	0.27	0.20	0.84	0.18	0.38	0.06	-0.04	
total_rec_prncp	0.06	0.06	0.83	0.84	0.82	0.16	0.10	0.83	0.10	0.39	0.04	-0.05	
total_rec_int	0.07	0.07	0.77	0.78	0.76	0.45	0.50	0.73	0.49	0.29	0.12	0.02	
total_rec_late_fee	-0.05	-0.05	-0.02	-0.02	-0.03	0.02	0.10	-0.02	0.10	-0.03	-0.01	0.04	
recoveries	0.07	0.07	0.06	0.06	0.05	0.17	0.20	0.03	0.19	-0.05	0.03	0.02	
collection_recovery_fee	0.05	0.05	0.05	0.05	0.04	0.15	0.19	0.03	0.18	-0.05	0.02	0.02	
last_pymnt_amnt	0.14	0.14	0.45	0.46	0.45	0.18	0.09	0.44	0.08	0.24	0.02	-0.03	
pub_rec_bankruptcies	-0.00	-0.00	-0.02	-0.02	-0.03	0.02	0.08	-0.02	0.09	0.00	0.01	0.00	
loan_status num	0.05	0.05	0.05	0.05	0.04	0.18	0.21	0.03	0.20	-0.07	0.04	0.02	

The following factors has the highest co-relation against the Loan status:

- Grade
- Term
- Interest rate

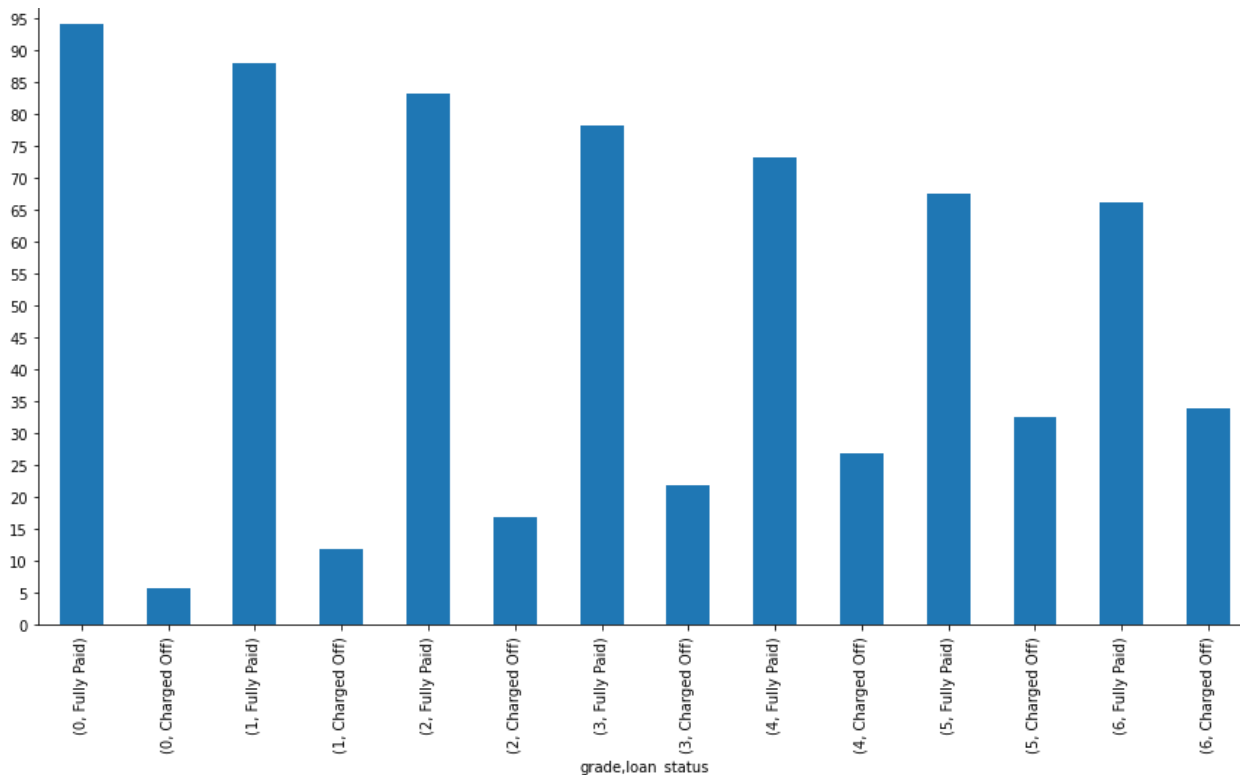
Analysis of Loan status vs Interest rate



Higher the interest rate, more people are getting charged off. Note median for

- Fully paid is about 11
- Charged off is about 14

Analysis of Grade vs Loan Status

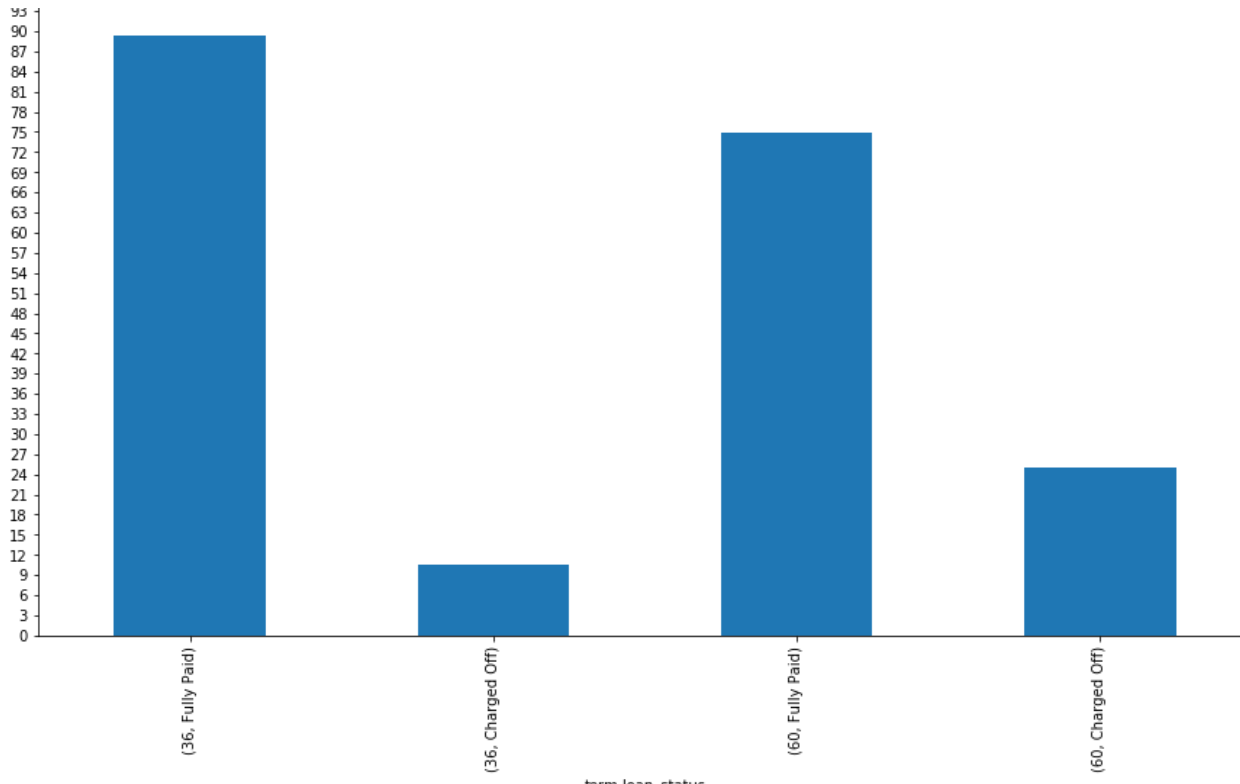


Note: Grade 0 is Grade A and 6 is Grade G

A is having least charged off people

F and **G** have highest charged off people.

Analysis of Term vs Loan status

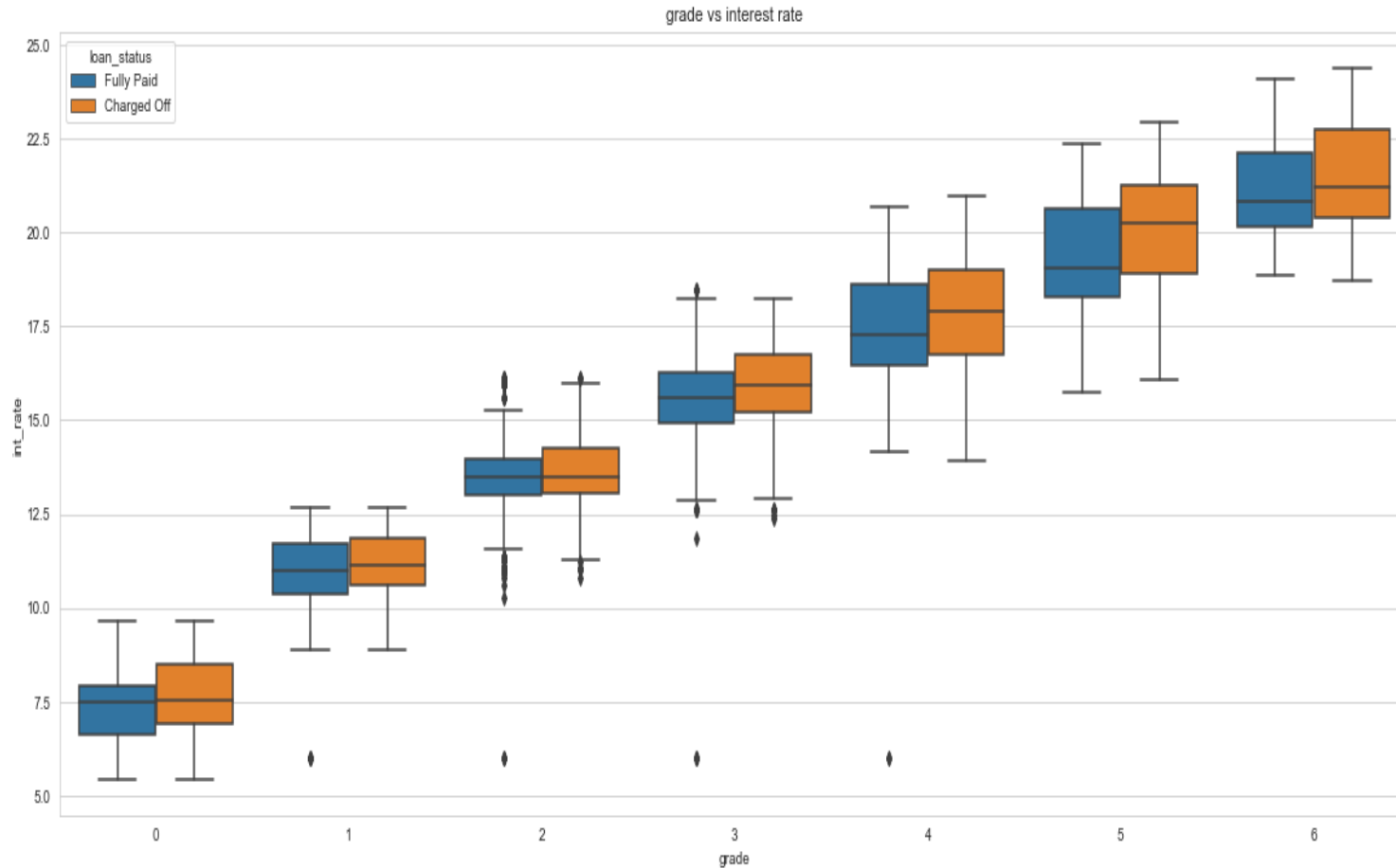


We just 2 terms i.e. 36 and 60 months. In that:

60 months have 25 % charged off.

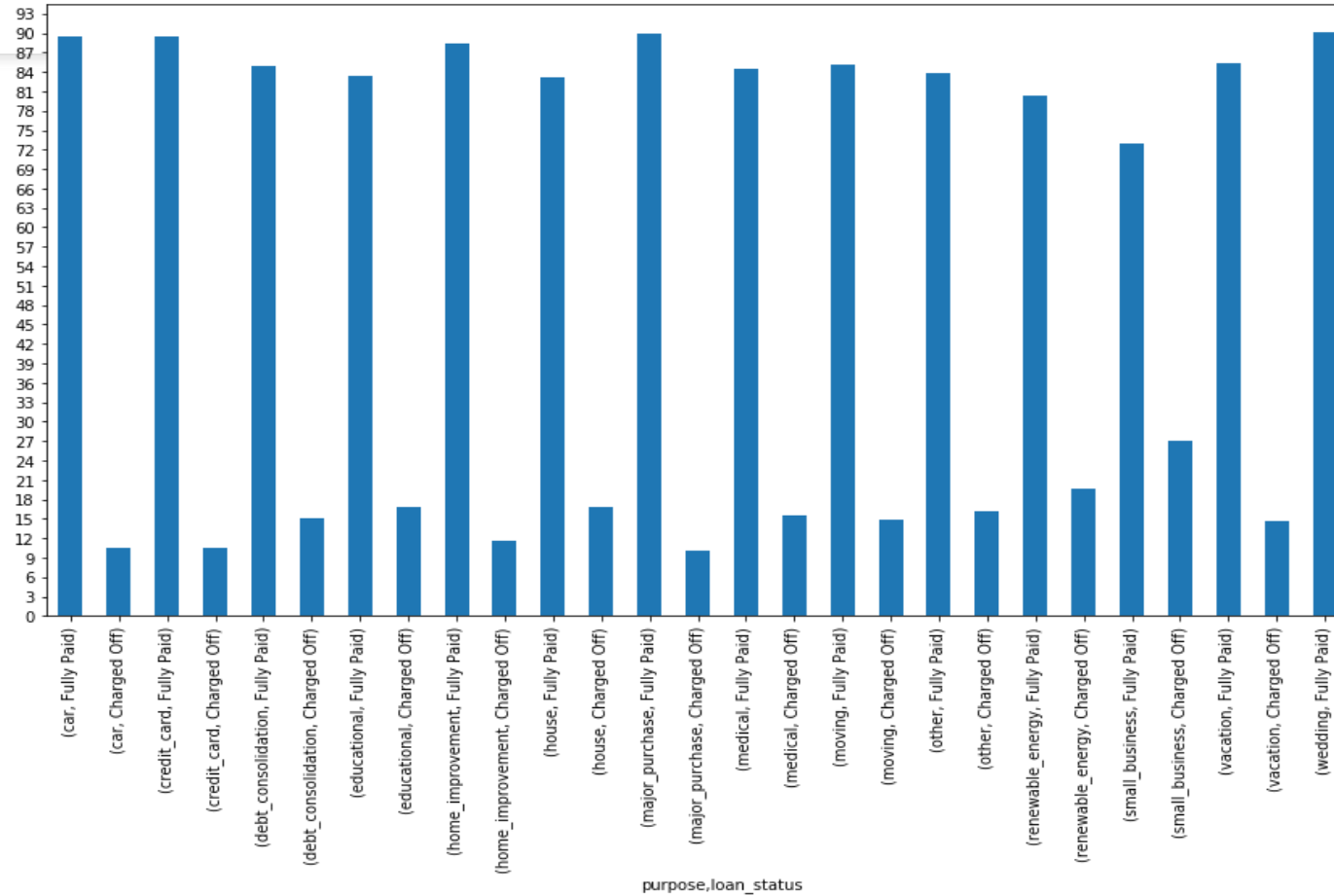
36 months have 10 % charged off.

Analysis of Grade vs Interest rate



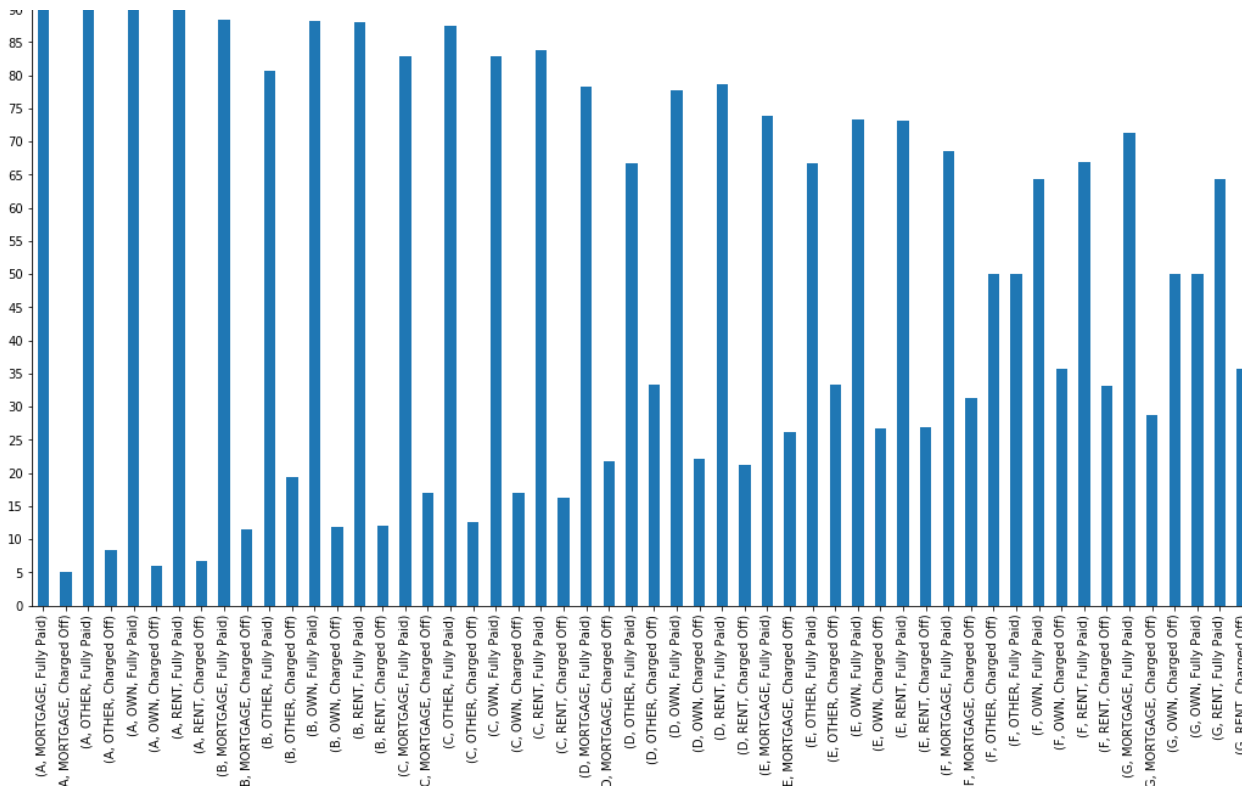
- Grade A has lowest interest rates and few charged off
- Grade G has highest interest rates and highest charged off
- Grade F is just below grade G and also has higher interest rates, charged off.

Analysis of Purpose vs Loan Status



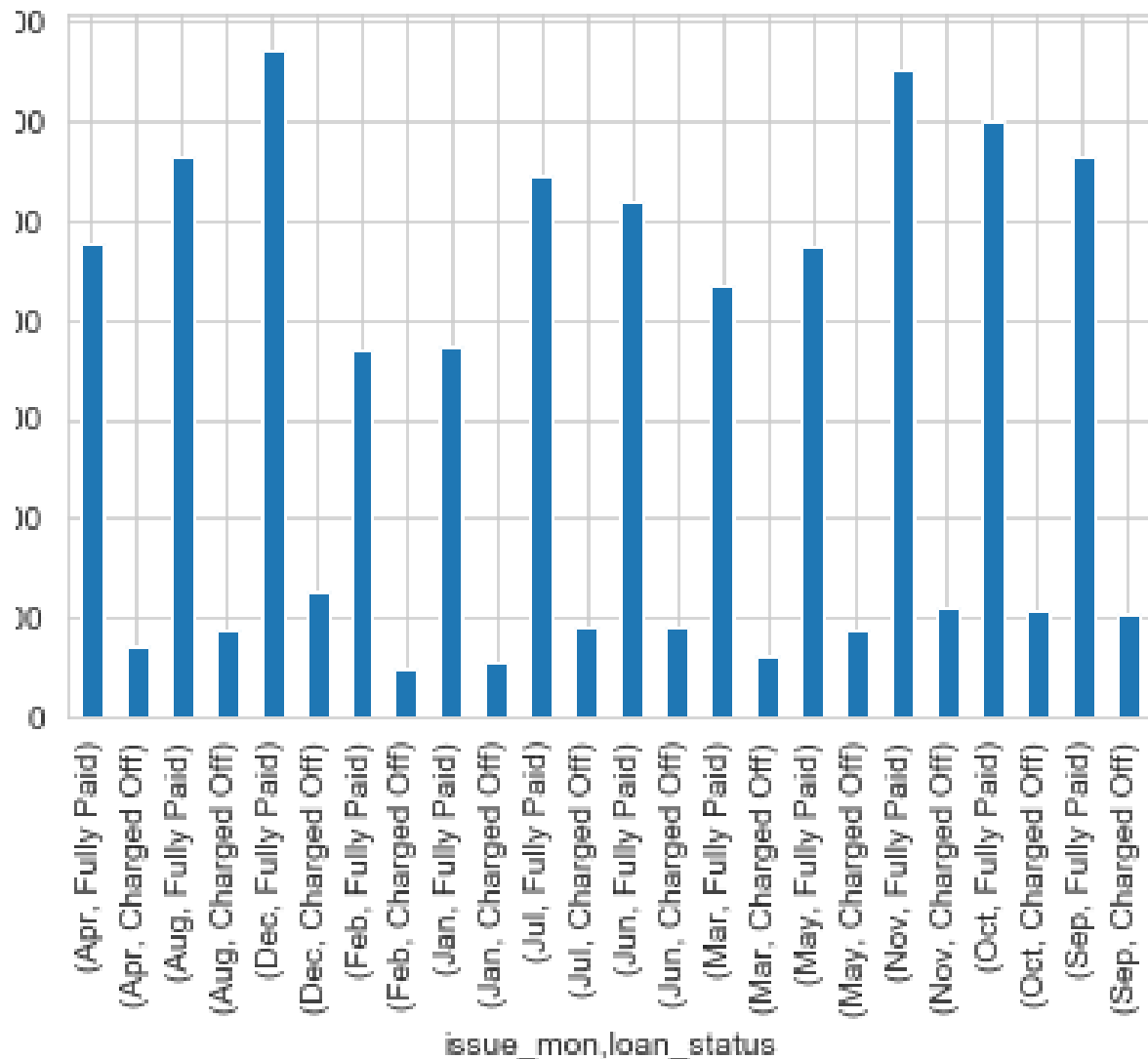
People with loan purpose of "Small business" have high percentage of defaulters.

Analysis of Grade, home ownership vs Loan status



F grade with "Other" home ownership have 50% charged off.

G with "Own home ownership" have 50% charged off.



Interesting note on when there is spike in loans over months

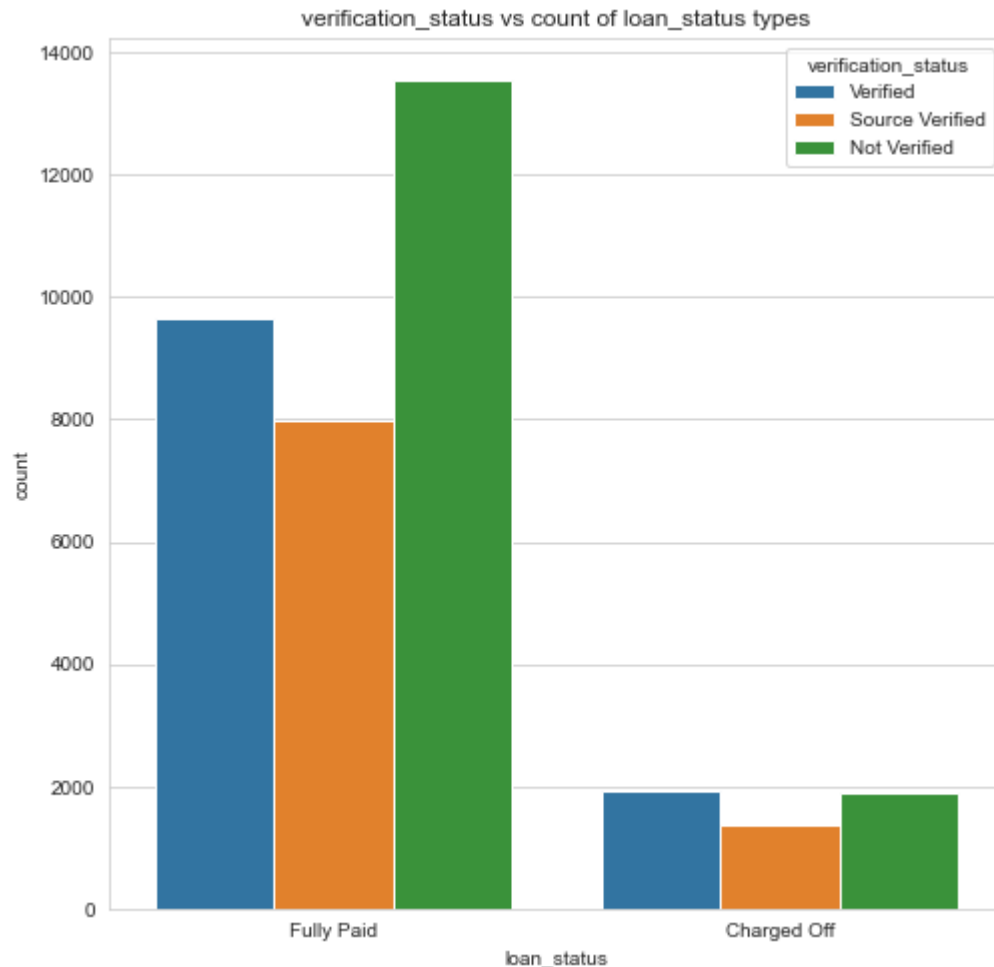
We have plotted Loan Issue month vs the number of loans

More loans were given in the months

October

November

December



Interesting note on when Verified data

We have plotted Loan status vs the number of loans for verification status. From this, we can determine that not percentage of Verified people getting charged off is marginally higher than not verified people which should not be the case.

May need to scrutinize the verification process for any wrong doing.

Conclusions

Top 3 Characters for a person who is likely default:

- **Interest rate** is the most important factor higher the interest rate, higher is the default rate.
- **Grade** is the most important factor. People from **F and G group** are most likely to default.
- **Term** is one other important factor. People with term of **60 months** have higher default rate.

Other important driving Characters:

- Loan purpose of **Small business** are more likely to default.
- People from **Grade A** are least likely to default.
- People from **Sub Grade A1** and with **Home ownership** as **OWN** has less than 1% chance of defaulting.
- People from **Grade F** with “**other**” home ownership have 50% charged off people.
- People from **Grade G** with “**Own home ownership**” have 50% charged off people.
- People from **Grade F, G** have around 40% chance of defaulting when interest rate is more than 20%.
- More loans are issued in the months of October, November and December.

Observations

- Employee length doesn't have any impact on default rate.
- Employees with 10+ years have taken more loans.
- 2010 and 2011 have more number of loans issued.
- Fully paid group people have median values of interest around 10-12.5% and charged off people have interest rates of around 12.5 to 15.
- From co-relation matrix, Recoveries and collection recovery fee is directly related to Charged off column. But these columns data is being ignored as it is known only after the loan is issued.
- From co-relation matrix, as total income and total accounts increases, the charged off people decrease but the margin is very small.
- California has highest number of loans from 2007 to 2011.

Note: All the plots are not presented here because of size .Kindly refer for more plots in the python file.

Recommendations

1. People in A grade have 95% chance to repay hence they are ideal candidates for loans. If they are of subtype A1 and with OWN home ownership they have more than 99% chance of repaying and can be considered as the cream layer set of group of people.
2. People with F and G grades have more chance to get charged off and they must be checked for higher income to avoid defaulting. Lesser funded amount is better to control the loss of money.
3. People should be recommended for shorter Terms to get the money sooner and avoid defaulting.
4. Verification process may need to be examined further to avoid any fraudulent activities.

