Lending Club Assignment

Samuel Idaewor
Sofianne Amarouche

What are the driving factors that influence the tendency of a borrower to default?

The Exploratory Data Analysis (EDA)

Approach

Exploratory Data Analysis

- Perform Data cleaning and sanity checks on the data (data formatting, data validity checks)
- Create new derived columns and populate with useful data from other column(s)
- Univariate analysis
- Bivariate Analysis
- Draw conclusions based on results of analysis

Data Cleaning & Sanity checks

1. Removed columns that are completely null
2. Remove columns that contain same values for virtually all of the rows
3. Remove rows that are completely empty.
4. Format data and convert to proper type
5. Check validity of certain values e.g interest rates in percentage
6. Clean up categorization variables can merge or remove categories that are only a very insignificant part of the data set.

Create Derived columns

- Created column ‘loan_amnt_ranges’ which categorizes the loan amount into 7 ranges
- Added a column for ‘int_rate_ranges’ which categorizes interest rates into **low, moderate** or **high.**
- Created column ‘funded_amnt_ranges’ which categorizes the loan amount into 7 ranges
- Created a column for ‘percentage_closed_acc’ which represents the percentage of closed accounts derived by (total_acc - open_acc)/total_acc * 100.

Univariate Analysis

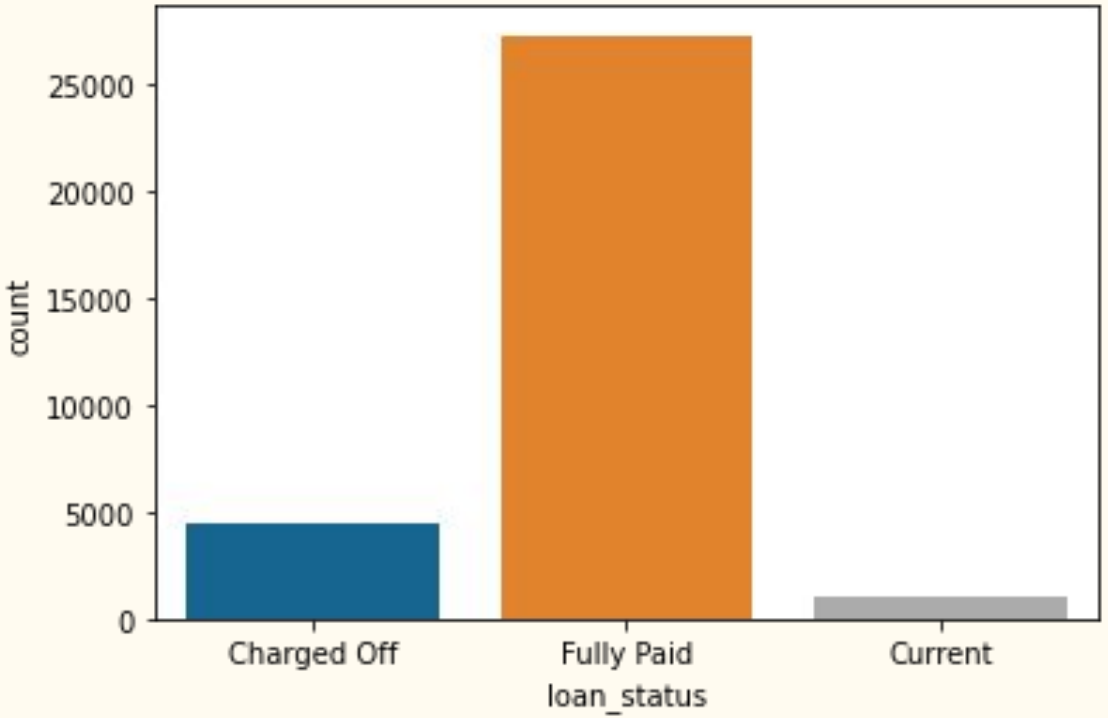
The table represent the variables that have been selected for both Univariate and Bivariate analysis.

S/N	Categorical	Quantitative
1	term	loan_amnt
2	grade	funded_amnt
3	emp_length	annual_inc
4	home_ownership	total_pymnt
5	verification_status	int_rate
6	issue_d	
7	loan_status	
8	purpose	
9	open_acc	
10	total_acc	

Loan Status

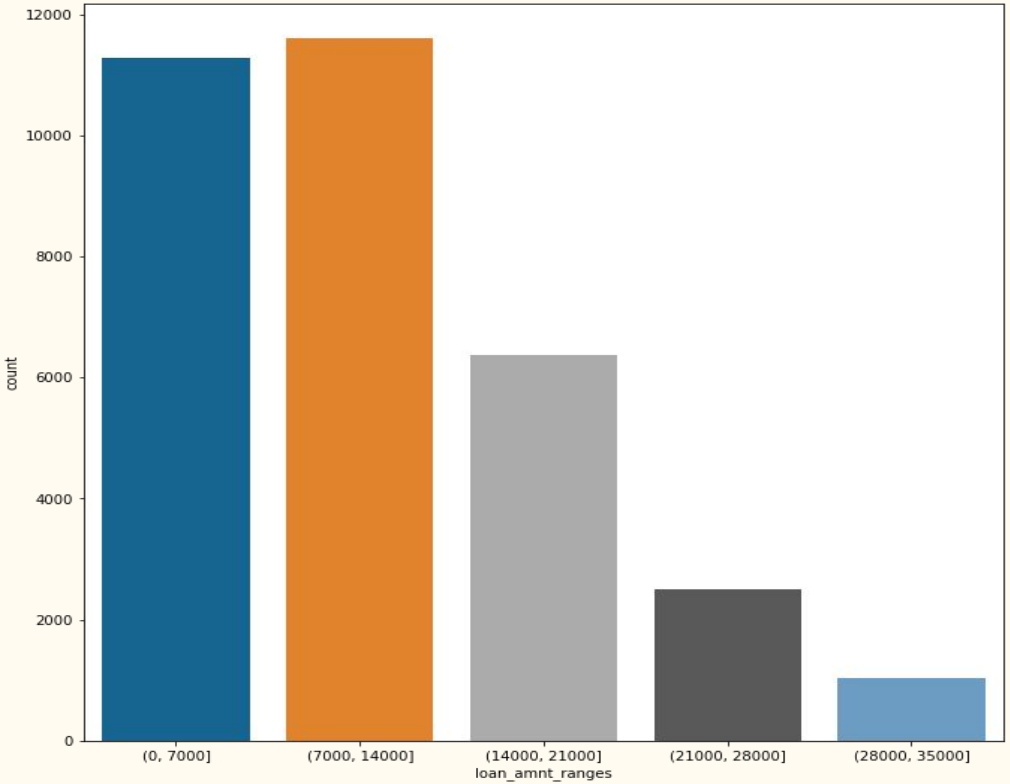
83% of the loans have been fully paid whereas approximately 14% of borrowers defaulted payment.

Hence the general idea would be to find how to further reduce the amount of defaulters.



Loan Amount Ranges

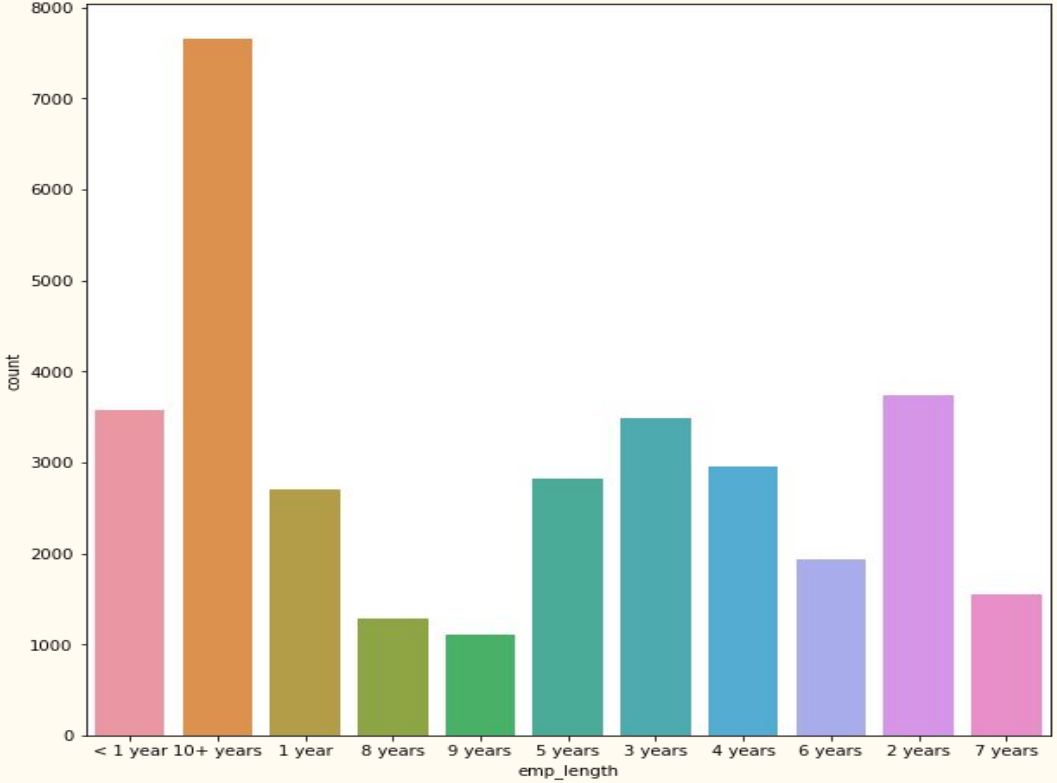
It can be seen that more that about 66% of the borrowers borrow up to a maximum of 14000.



Length of Employment

Here it is observed that people that has been employed for less than 6 year tend to take more loans than their counterparts who have been working for 6 years and more.

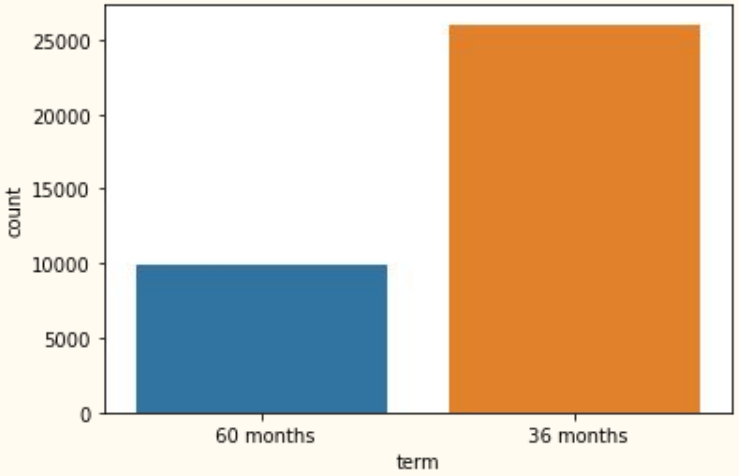
The spike in the 10 year plus category is probably because it contains a wider range of employees (10 year to retirement)

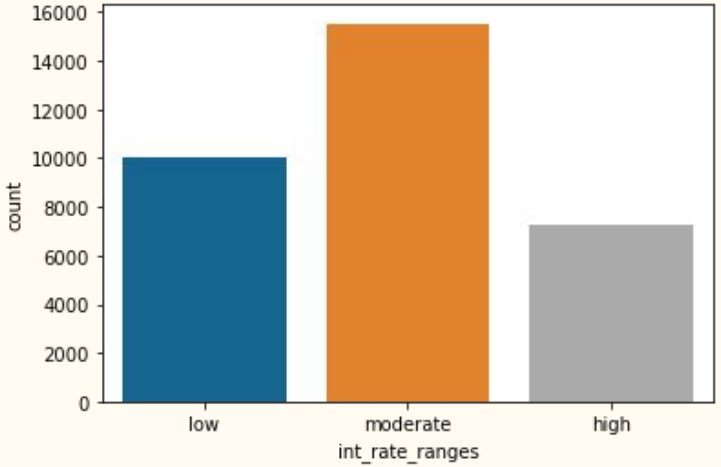


Interest rate ranges & Term

Most of the loans have a shorter term to pay back.

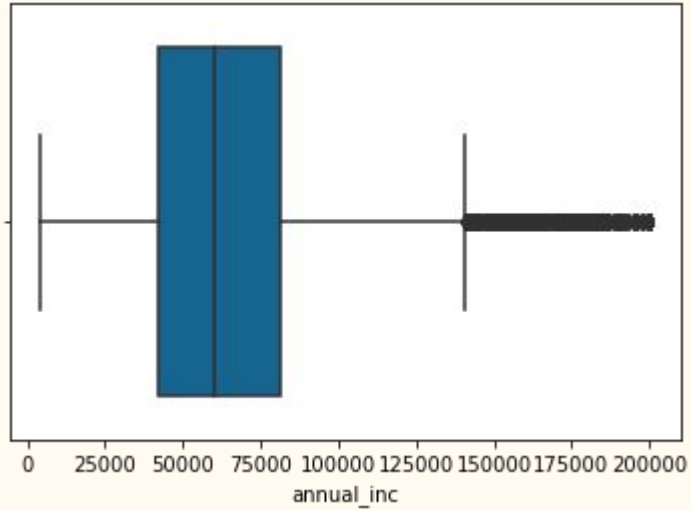
Loans with moderate(10-15%) interest rates are more common whereas there are more low(0-10%) interest loans than there are high(15-25%) interest loans.

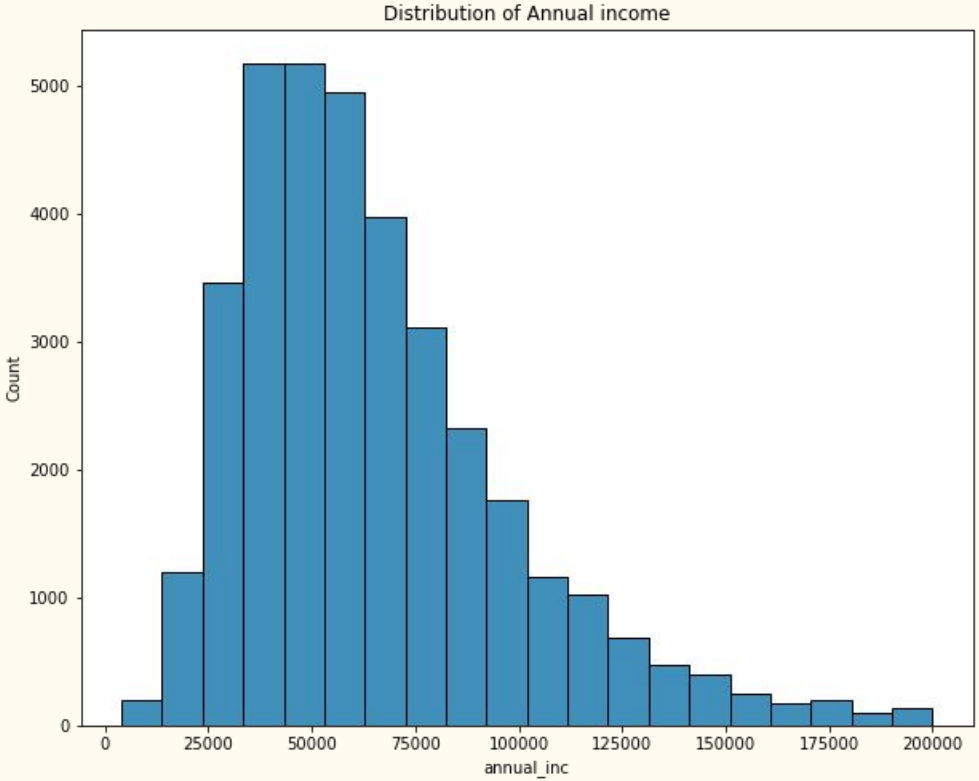




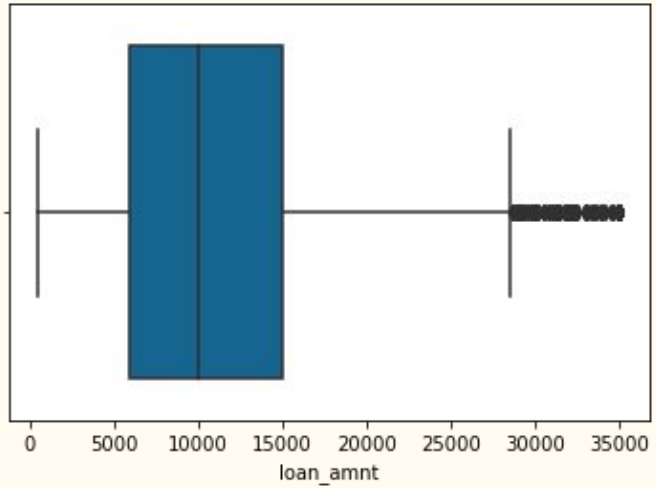
Annual Income

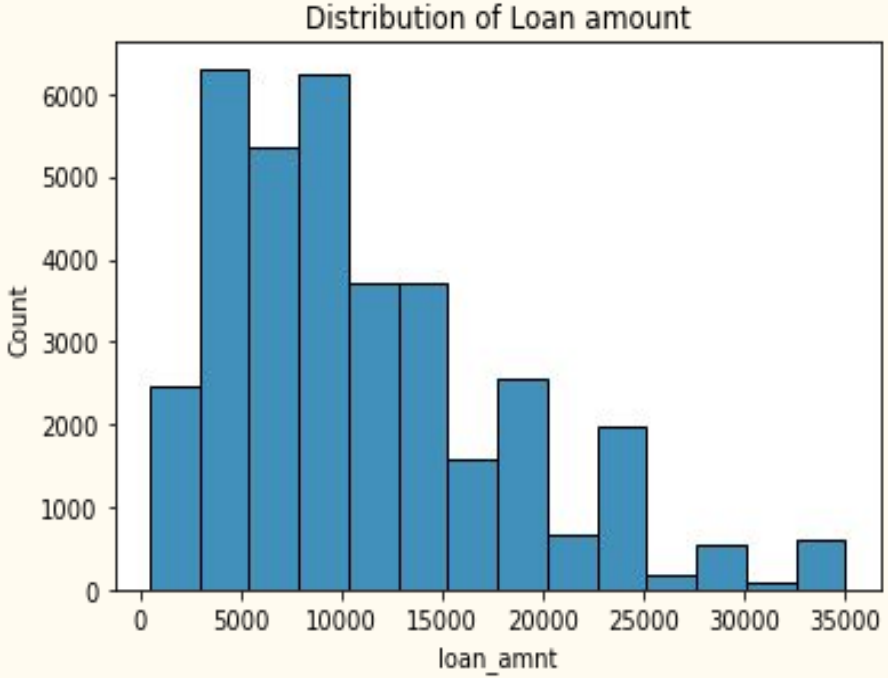
Average income is 60000. A large part of the borrowers earns between 40000 and 80000. There are more people in above the 75th percentile than beneath the 25th.



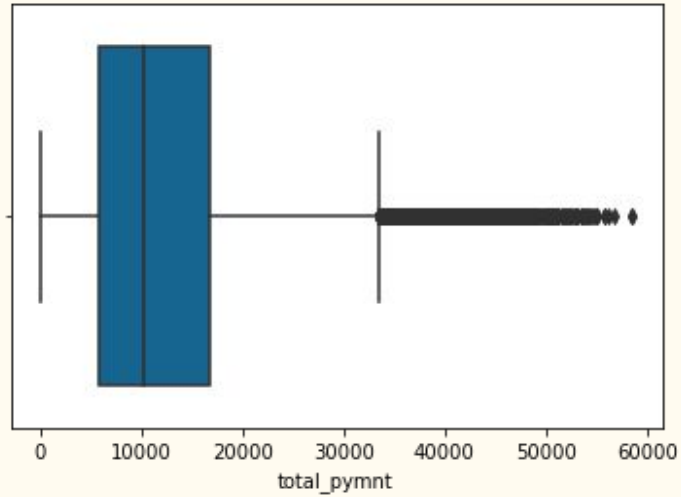


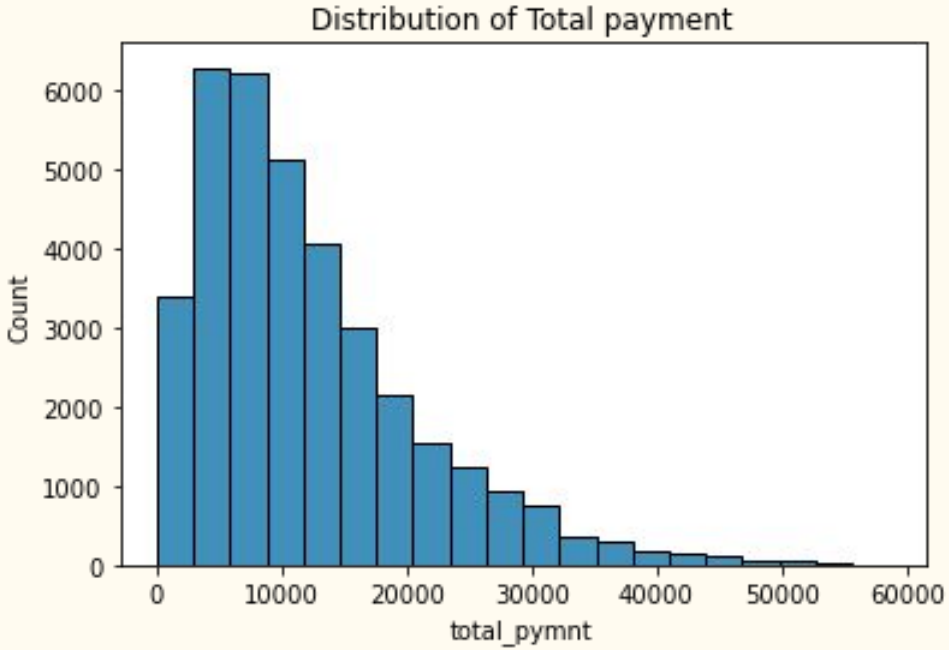
Loan Amount


0
5000
10000
15000
20000
25000
30000
35000
loan_amnt


Distribution of Loan amount
Count
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt

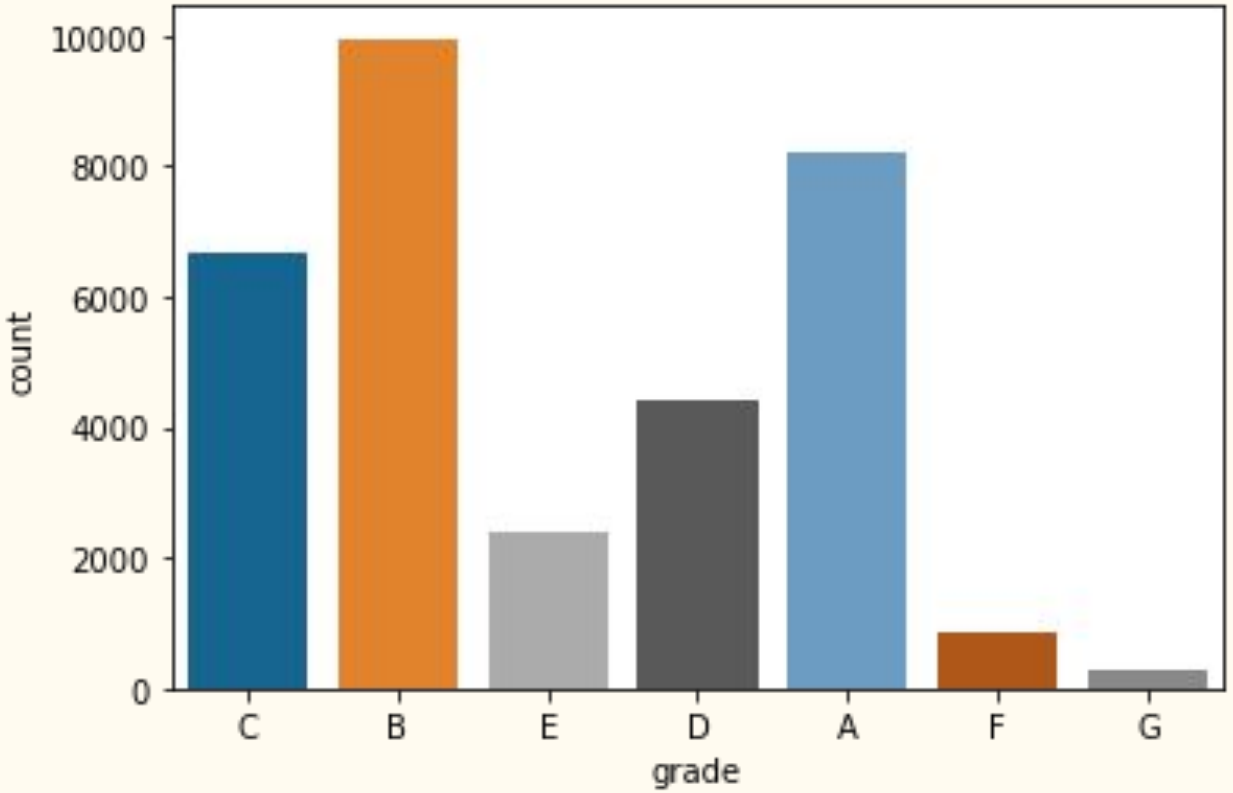
Total payment





Grade

About 30% of the borrowers belongs to grade B. Grades A-C contains more than 60% of the entire data set.



Bivariate Analysis

Here, two variables are compared to attempt to establish the relationship between them.

In most of the cases the loan status is one of the variables which is used in bivariate analysis since it is our target variable.

We want to see factors that affects the payment of default of a lender.

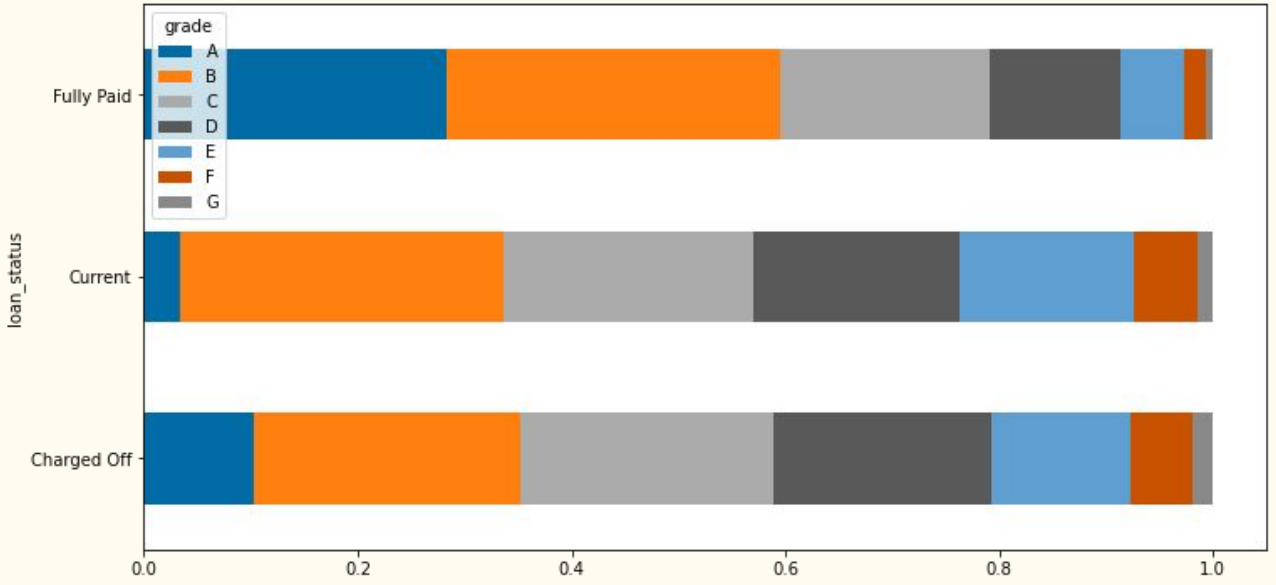
If a consistent pattern is observed then, we can further explain what may be responsible for it.

Grade vs Loan Status (Bar chart)

Borrowers belonging to grade A are more likely to pay off the loan than those belonging to grade B and C

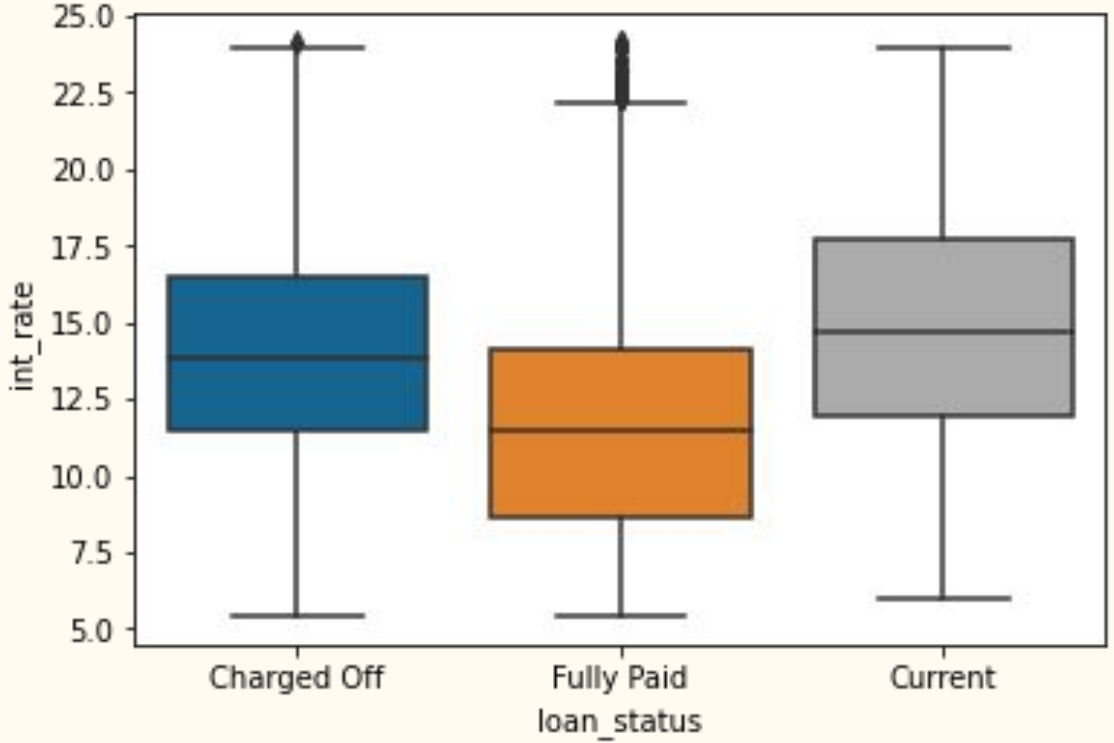
From the Univariate analysis we know that there are more borrowers in group B.

The stacked bar chart is based on the percentage of the number of borrowers in each grade by loan_status



Interest rates vs Loan Status (Box plot)

Here is can be seen that borrowers with low to moderate interest rates tend to pay off loans completely than those with higher interest rates.

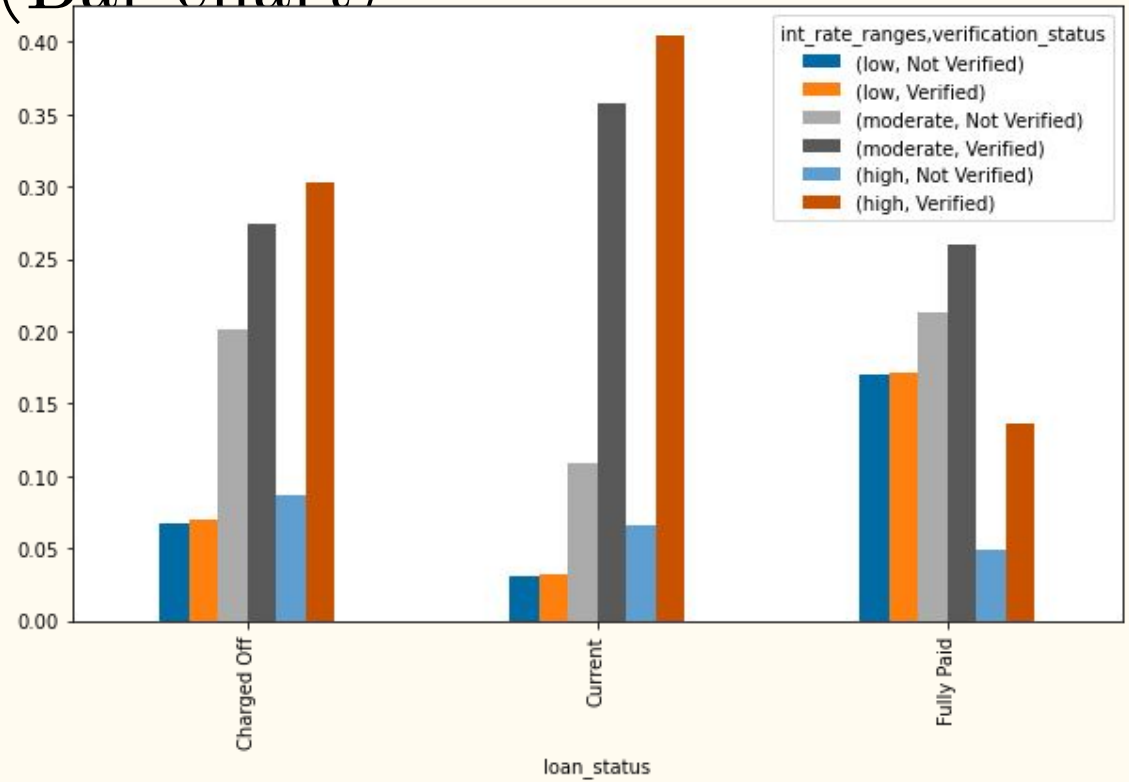


Loan status vs interest rate ranges and verification status (Bar chart)

Here we also include verification status to establish finding of previous slide.

Less than 15% of fully paid loans are have high interest rates and are verified.

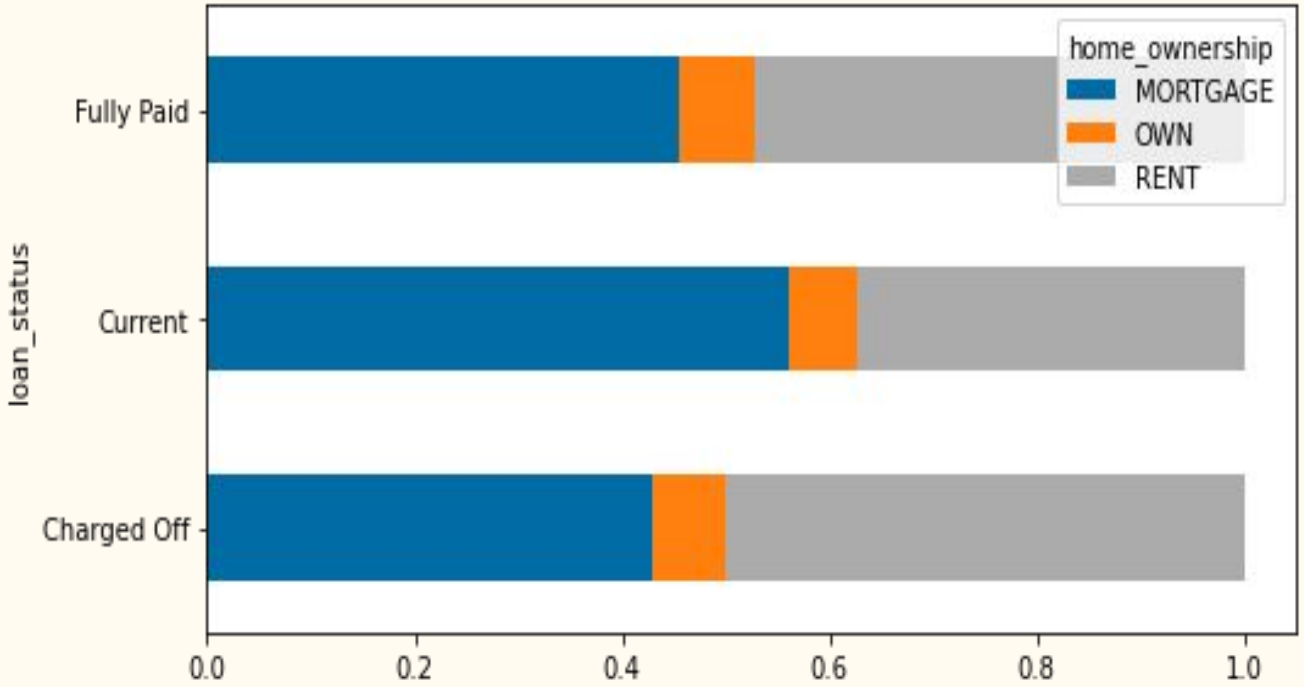
About 30% of verified high interest borrowers default on payment of loans.



**Open Account count vs loan status vs home ownership

People who are have mortgages are more likely to pay than those who rent houses.

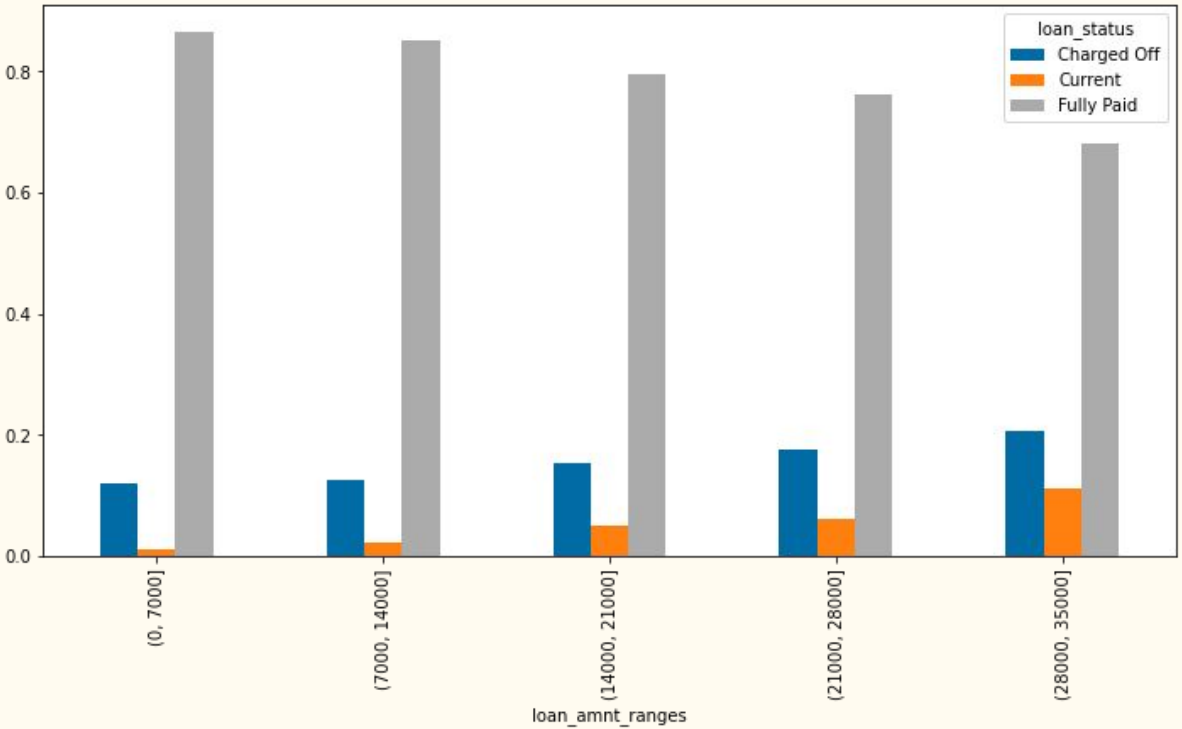
The margin is small, so we cannot really conclude that this is a factor.



Annual Income % count vs loan status and loan amount Ranges

It can be clearly seen that as the loan amount is increase the percentage of paid off loan within that loan amount range is reducing while that of defaulters are increasing.

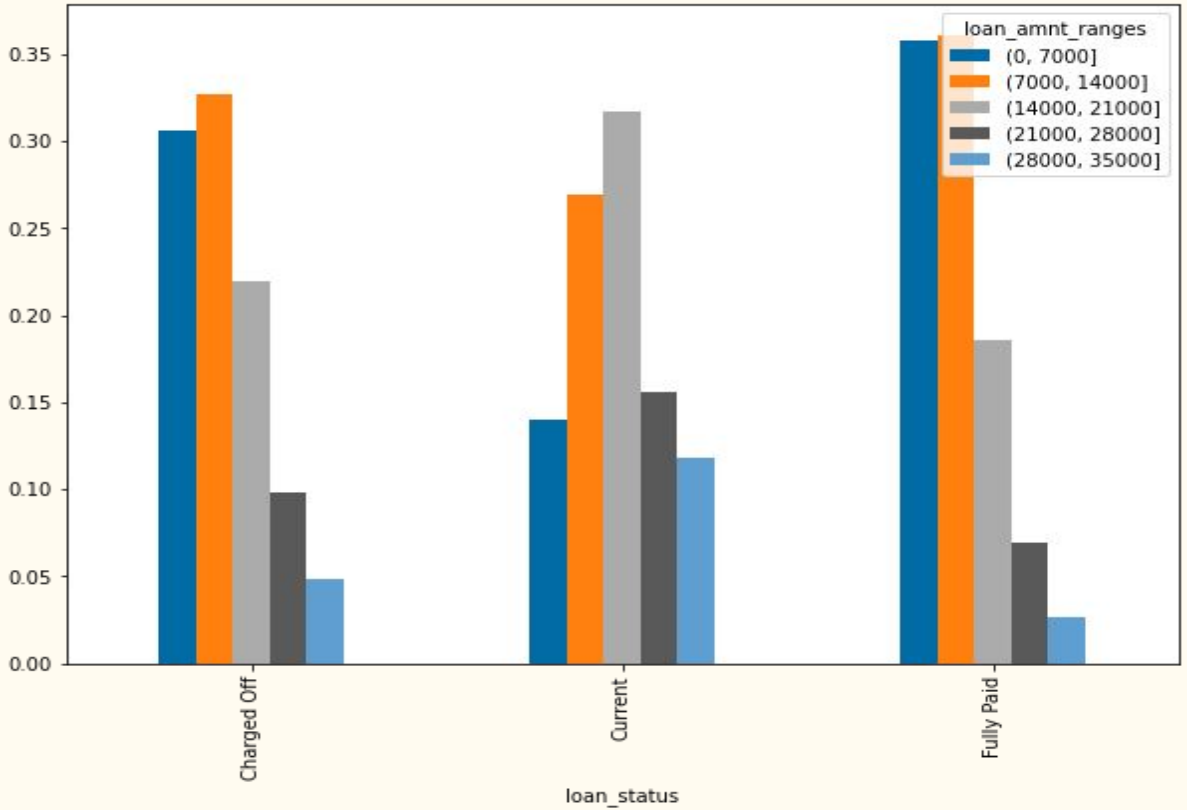
Hence loan amount and annual income are factors to consider together before lending to borrowers



Issued Loan count vs loan status vs Loan amount ranges

Comparing the percentages of defaulted and fully paid loans for the higher loan ranges, it can be observed that there is a higher percentage of defaulters than there are fully paid loans.

This further indicates that higher loan amounts more likely to default than lower amounts.

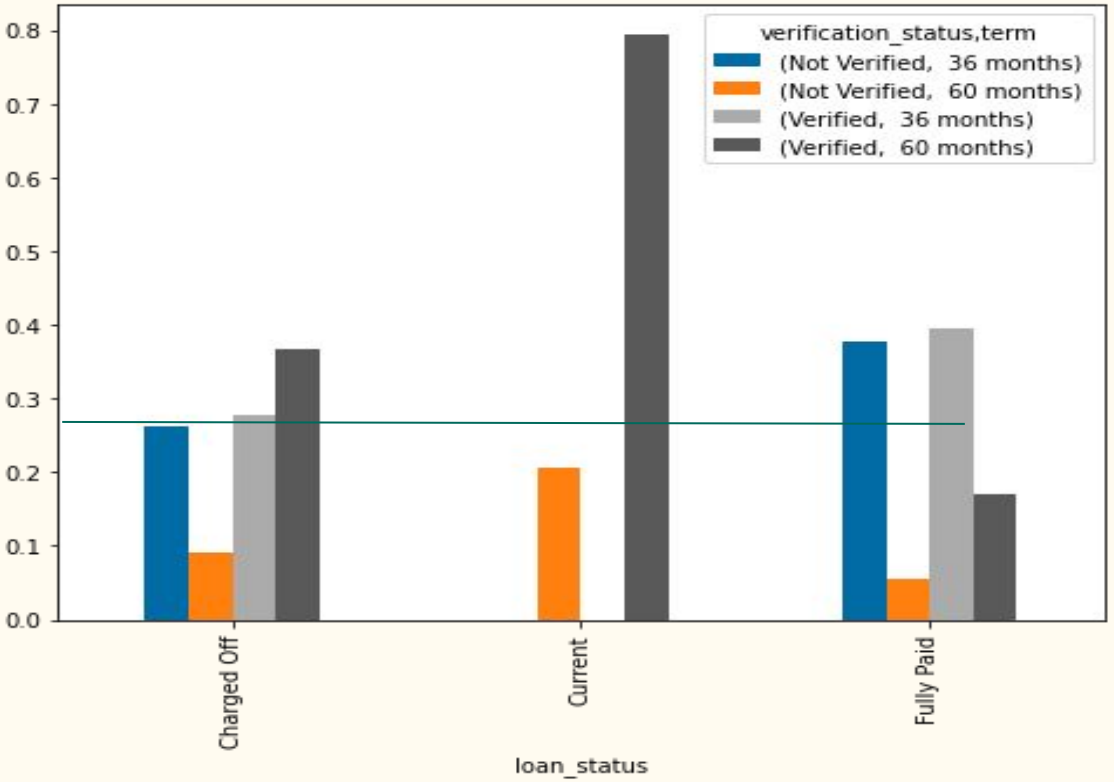


Loan status and term and verification status

More than 35% verified loan defaulters have a 60 month loan term compared to about 15% that is fully paid and verified.

There are also a higher percentage of verified fully paid loan with a term of 36 months than defaulters within same term.

Hence the longer term loans are more likely to default.



Conclusion

From the results of the analysis, the following variables have been found to be the main drivers of whether a borrower will default or not

Interest rate: Higher interest rate borrowers are more likely to default than low and moderate interest rate borrowers

Term: Shorter term loans are more likely to be fully paid than longer term loans

Grade: Grade A borrowers are most likely to pay than their grade C&D counterparts who are more likely to default on loan payment.

Loan Amount: The likelihood to default on a loan increases with increasing loan amount. Hence smaller loans amounts are more likely to be fully paid.