Lending Club Case Study

Problem Statement

- When a person applies for a loan, there are two types of decisions that could be taken by the company:

- Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
- 1. Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
- 2. Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- 3. Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

- Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements.

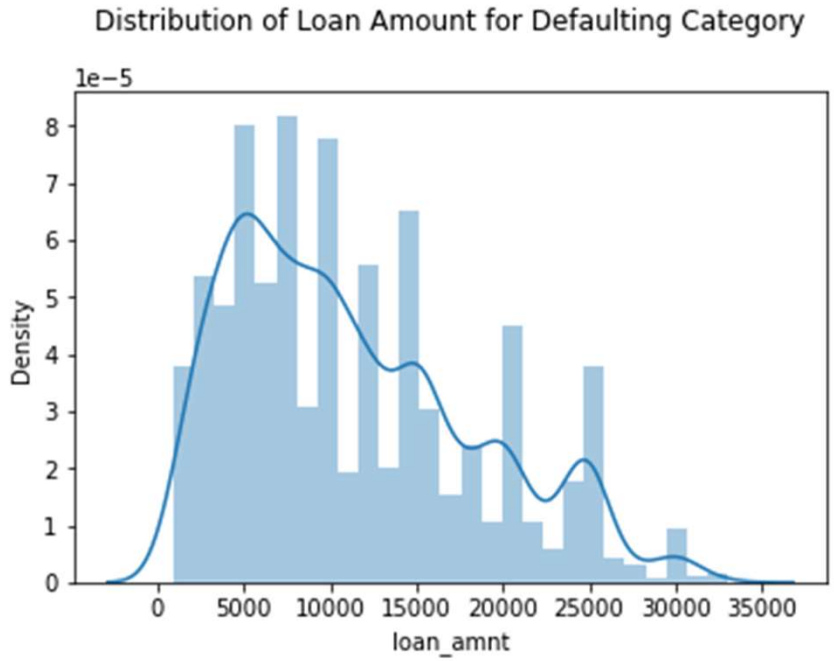
- Our Aim is to identify how various factors affect the defaulting of customer loan.

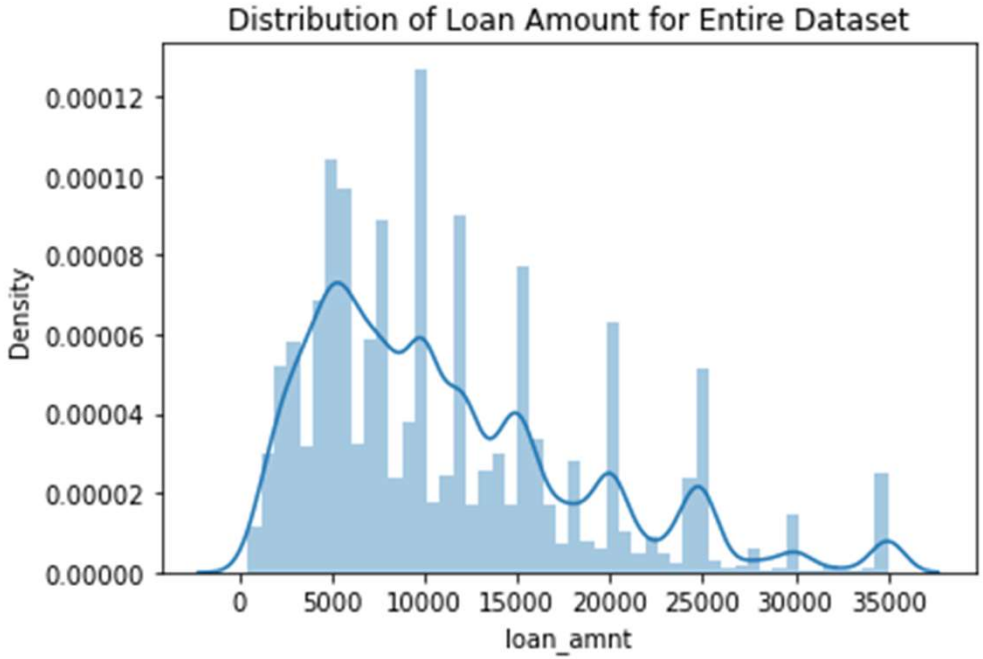
Data Cleaning

- The Dataset consisted of columns that were NULL for all index values, such columns were identified and removed
- All the features were not necessary for analysis, so only the one's required were picked
- Some columns out of the selected ones contained columns which have duplicate values for all the indexes, such columns were removed.
- Check for NULL values in selected columns and handling appropriately.
- Next step was to convert columns into appropriate formats for analysis, ex: converting of 'term' column to int format.
- Also new column named 'issue_d_month' was derived from existing column 'issue_d'
- Next step was to filter the data to include only charged of customers for analysis.
- Outliers were detected and removed from numerical columns.
- For Bivariate Analysis converting loan_status to 0's and 1's for Fully Paid and Charged Off customers.

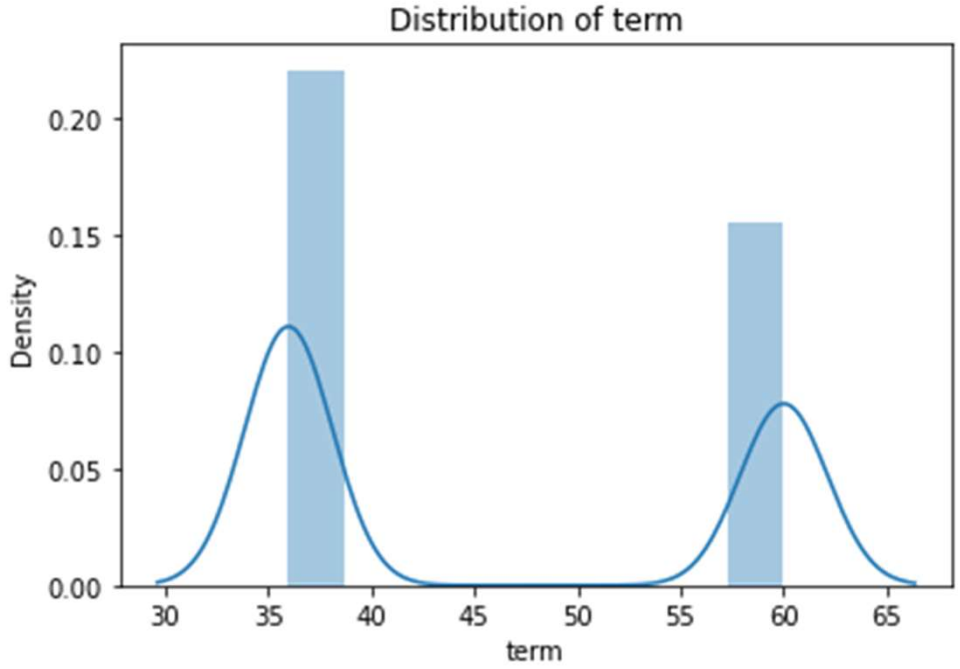
Univariate Analysis

- Plot of Distribution of loan amount for entire data set as well as for Charged off customers was plotted and from the plots we can conclude that loan is usually borrowed in 2.5k interval amounts by most number of people, spikes at 2.5k intervals say so.

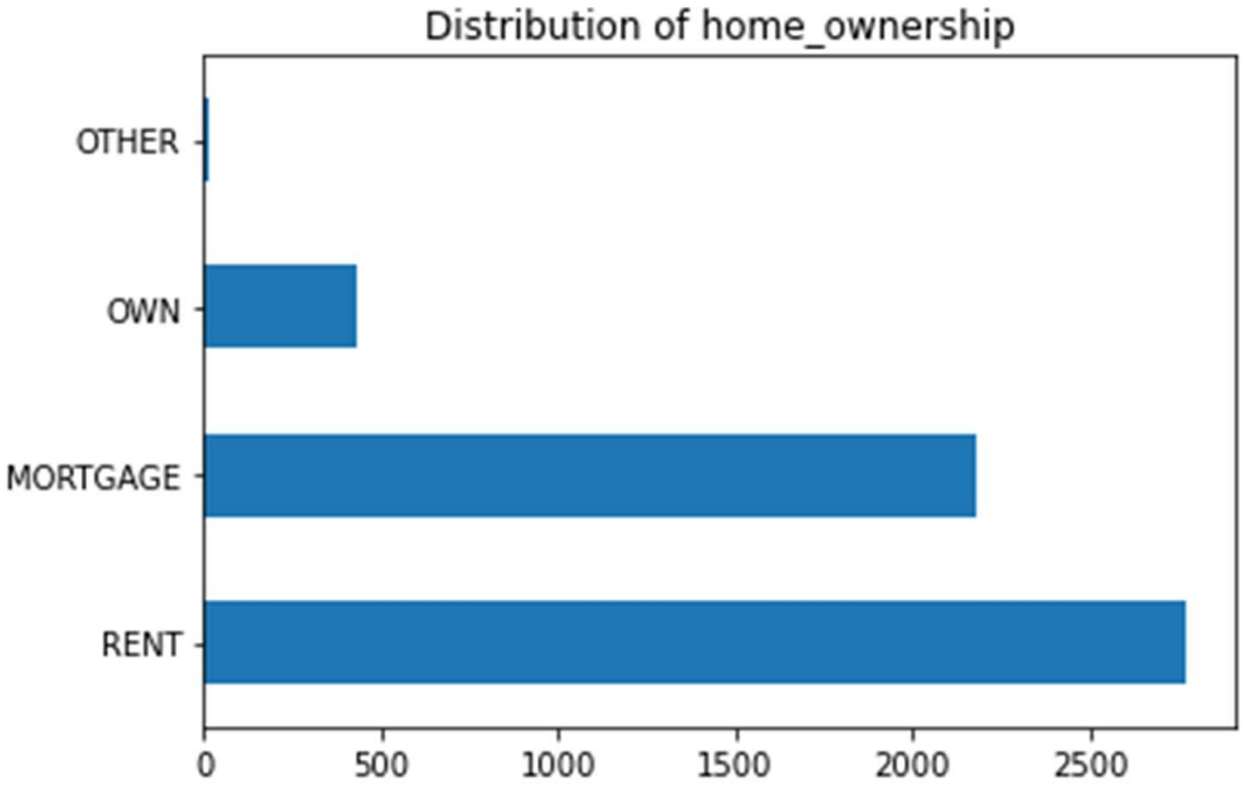




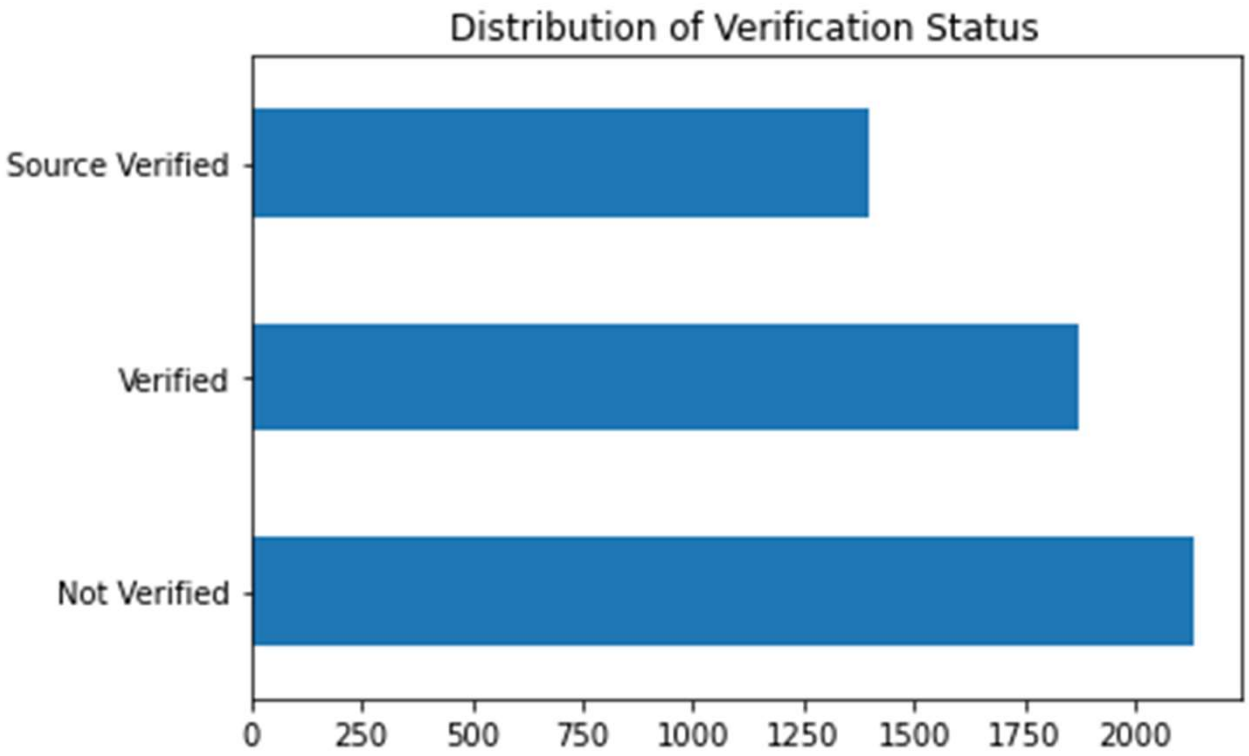
- Plot of distribution of term of loan for defaulting category shows that people whose term is 36 months are more likely to default than those who have term as 60 months. The difference is nearly 1.5 times. It might be because people are not able to arrange funds and as a result unable to pay the loans, however as interest rate keeps growing they are charged more and more.



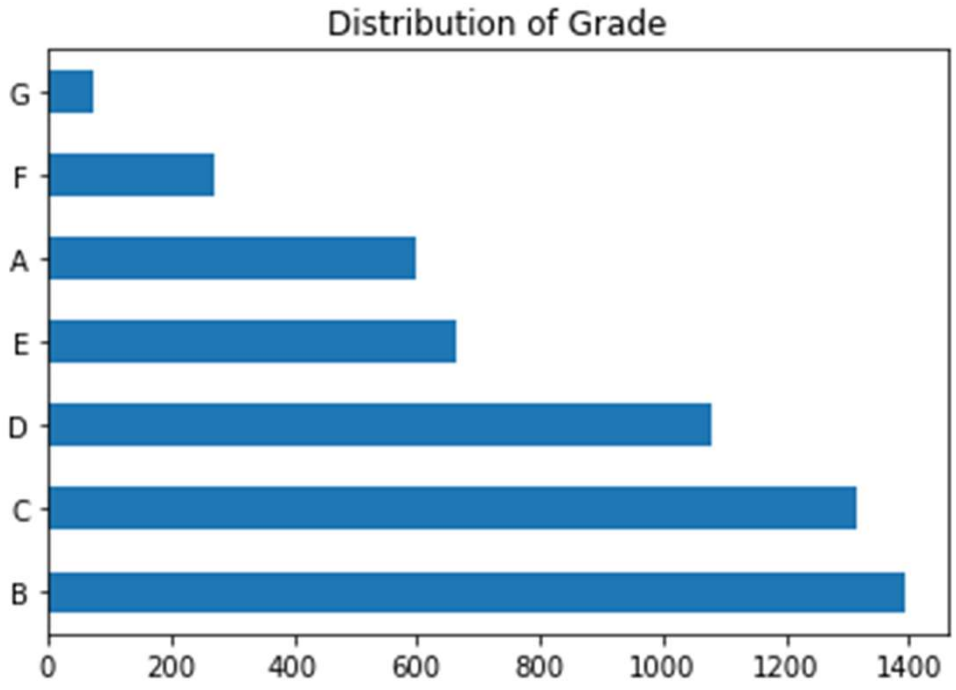
- Plot of distribution of ‘home_ownership’ shows that people who are in mortgage agreement or in rent are more likely to default than People who have Own Houses with Maximum number of defaults coming from people who live in rented homes and then from ones in mortgage homes.



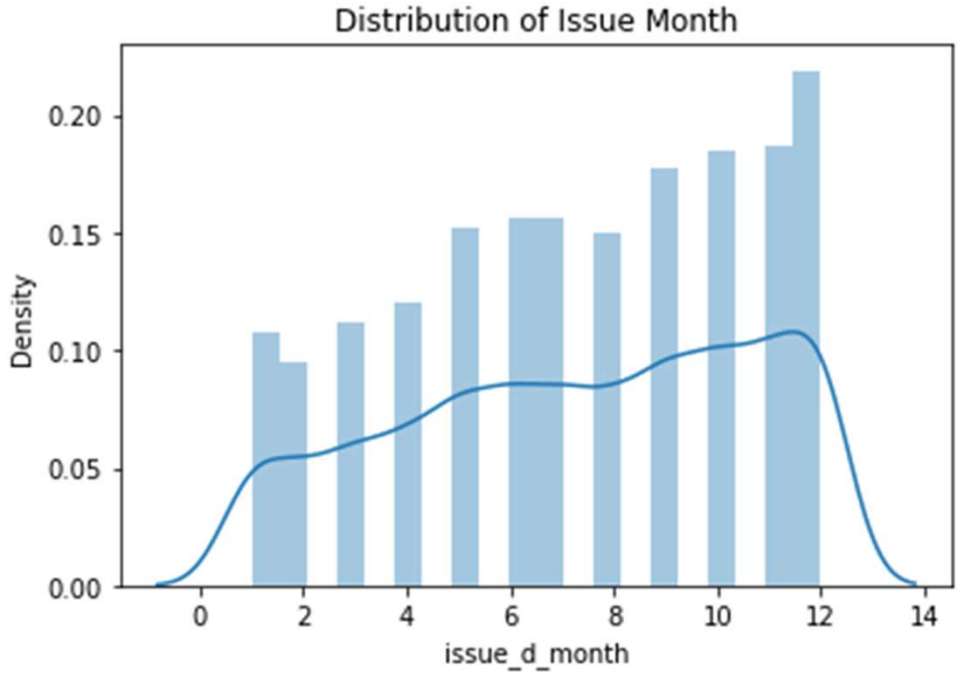
- Plot of distribution of verification status shows that not verified people are more likely to default than verified and verified ones are more likely to default than ones whose source of income is verified.



- Plot of distribution of grades shows that grades assigned as B,C and D has the highest default rate than all the other grades.

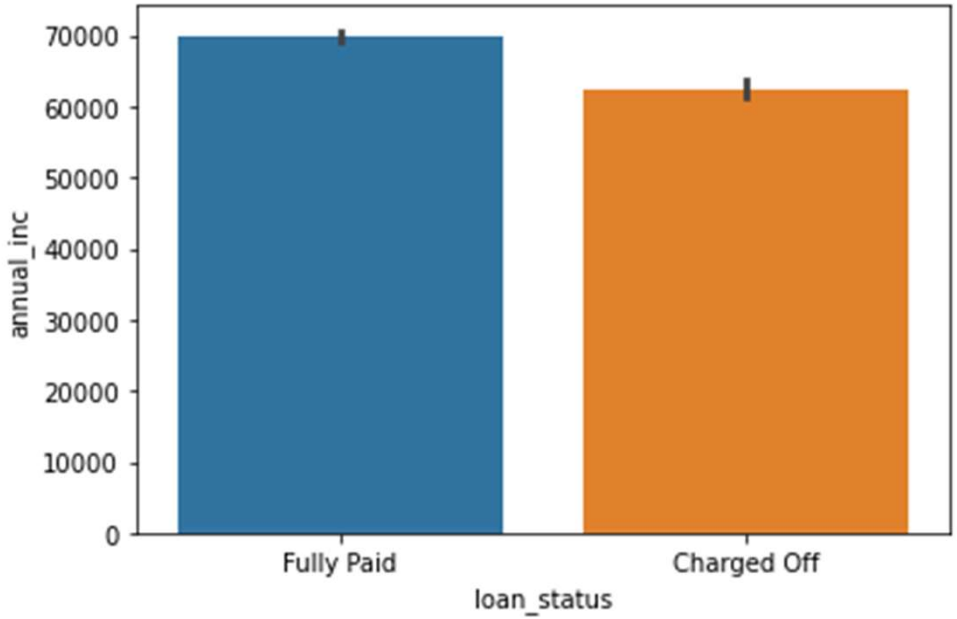


- Distribution of issue month of loan indicates that maximum loan defaults have happened for loans issued in month of December.

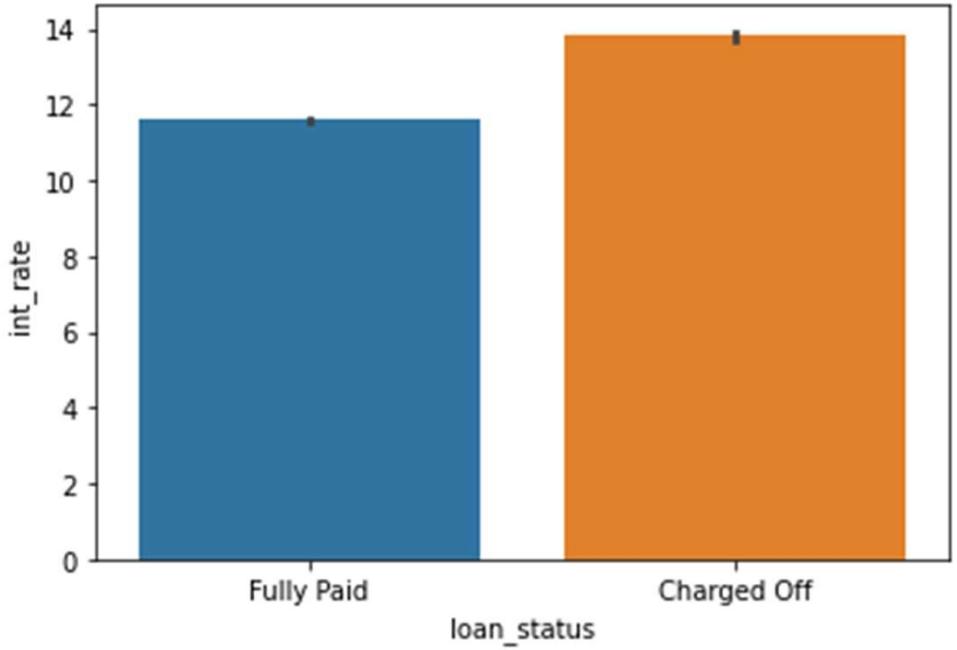


Segmented Univariate Analysis

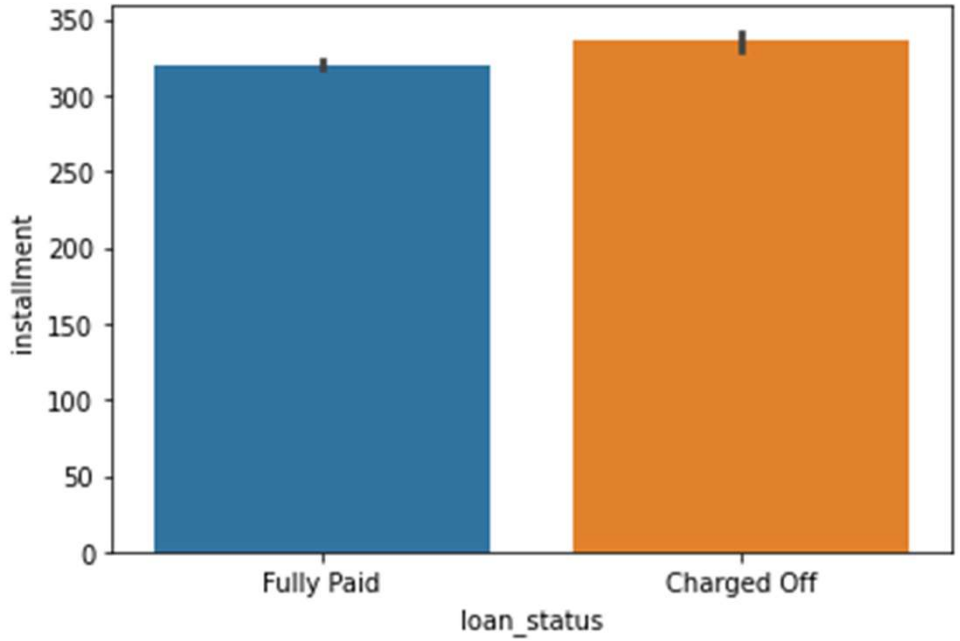
- The plot of ‘loan_status’ segmented over annual income indicates that people who are Charged Off have less overall annual income than ones who have Fully Paid the loan.



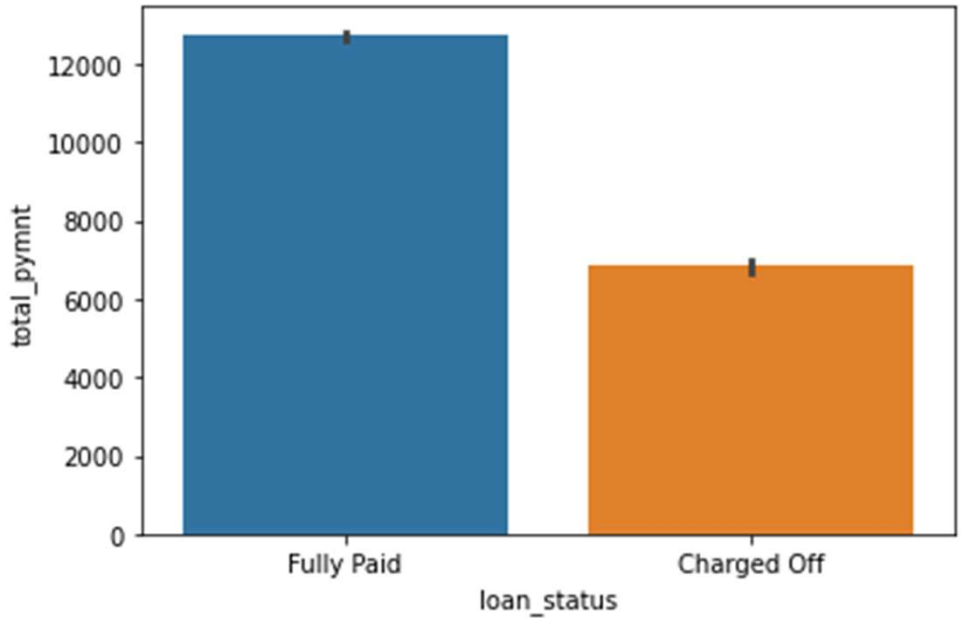
- Plot of ‘loan_status’ segmented over interest rate shows that the interest rate of Charged Off customers is more than that off Fully paid. Therefore it can be said that high interest rate loans are more likely to default.



- Plot of loan_status segmented over installments shows that average monthly installment paid by charged off customers is slightly more than that of Fully Paid customers.



- Plot of loan_status segmented over total payment received till date shows that default loans receive less amount of money in terms of loan repayments, nearly half as Fully paid.



Bivariate Analysis

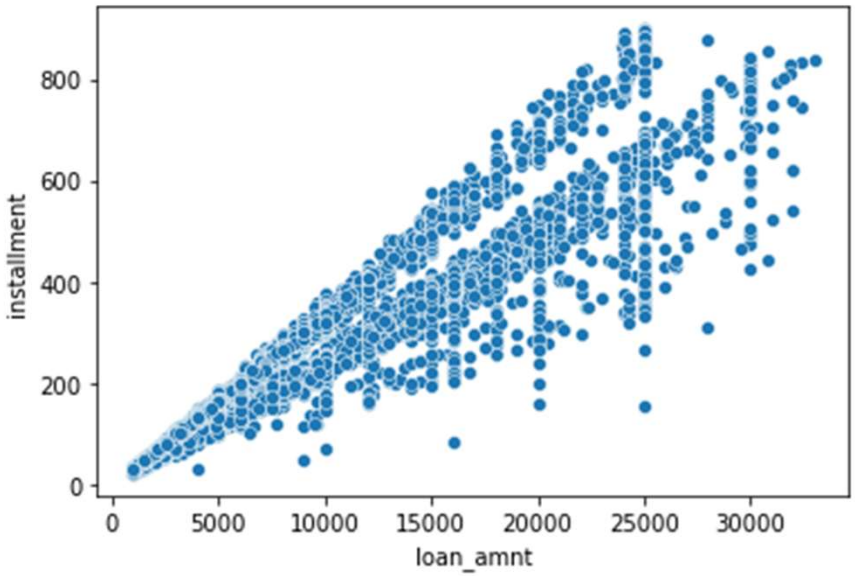
- The pivot table for home ownership vs verification status for loan_status as values shows that Mortgage-Verified and Rent-Not Verified are the most defaulting categories, though Mortgage-Verified is a peculiar category, which might indicate that the verification of mortgage home ownership category was not done properly.

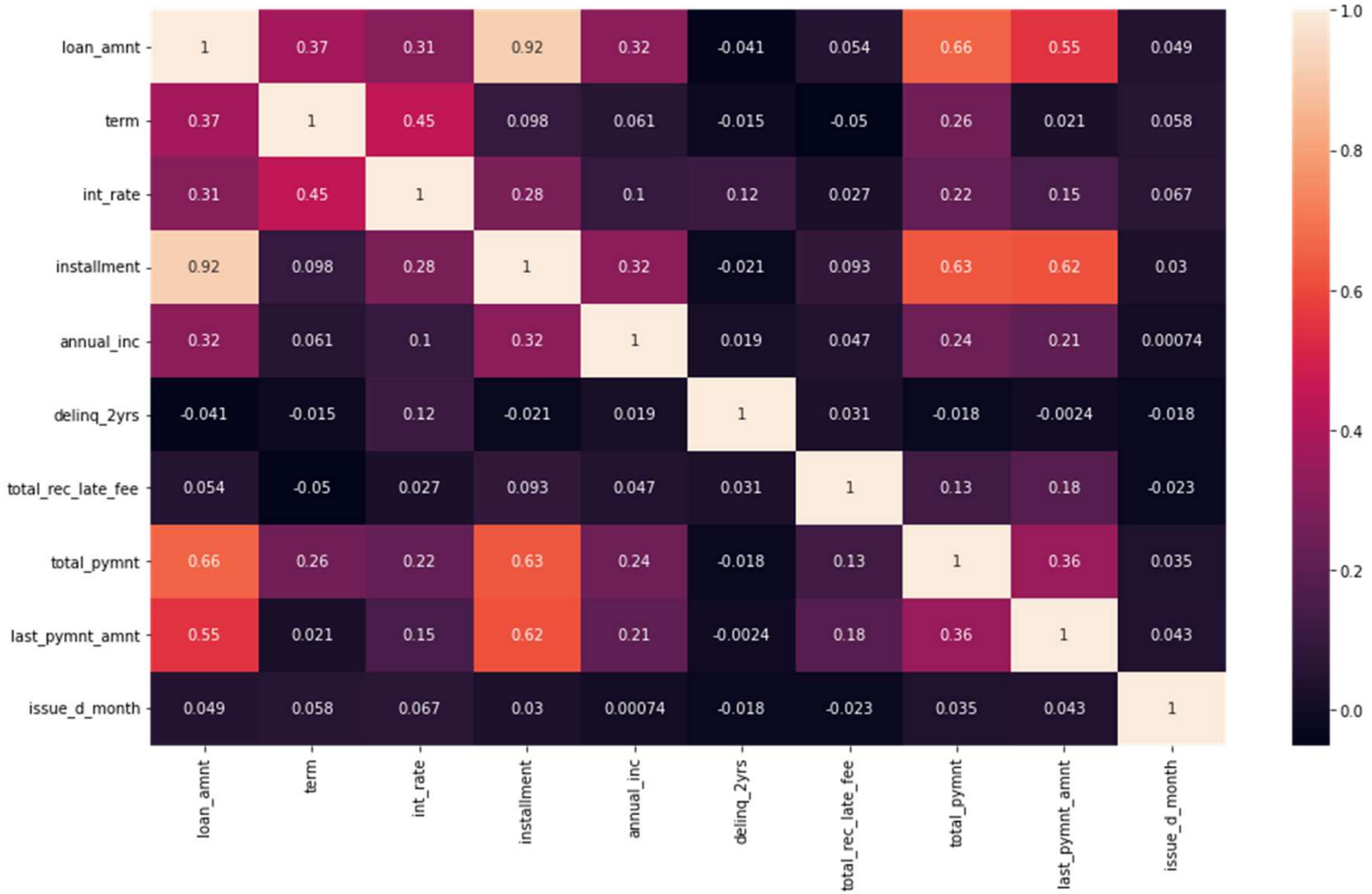
loan_status			
verification_status	Not Verified	Source Verified	Verified
home_ownership			
MORTGAGE	6916.0	3830.0	6275.0
NONE	3.0	NaN	NaN
OTHER	52.0	10.0	36.0
OWN	1459.0	742.0	774.0
RENT	8264.0	5095.0	5121.0

- The pivot table for home ownership vs grades for loan_status as values shows that people who are staying in rented or mortgage homes and have grades B, C and D are more likely to default.

loan_status							
grade	**A**	**B**	**C**	**D**	**E**	**F**	**G**
home_ownership							
MORTGAGE	5136.0	5003.0	3136.0	1957.0	1175.0	470.0	144.0
NONE	1.0	2.0	NaN	NaN	NaN	NaN	NaN
OTHER	24.0	32.0	16.0	17.0	7.0	2.0	NaN
OWN	863.0	902.0	579.0	363.0	186.0	61.0	21.0
RENT	4021.0	5736.0	4103.0	2748.0	1295.0	443.0	134.0

- By drawing a correlation matrix for all numerical variables we can see that highest correlation is between installment and loan_amnt. Therefore indicating the higher is the loan_amnt the higher is the average monthly installment.
- People often tend to take hefty loans and fail to pay their monthly installment and default on loan.




loan_amnt 1 0.37 0.31 0.92 0.32 -0.041 0.054 0.66 0.55 0.049
term 0.37 1 0.45 0.098 0.061 -0.015 -0.05 0.26 0.021 0.058
int_rate 0.31 0.45 1 0.28 0.1 0.12 0.027 0.22 0.15 0.067
installment 0.92 0.098 0.28 1 0.32 -0.021 0.093 0.63 0.62 0.03
annual_inc 0.32 0.061 0.1 0.32 1 0.019 0.047 0.24 0.21 0.00074
delinq_2yrs -0.041 -0.015 0.12 -0.021 0.019 1 0.031 -0.018 -0.0024 -0.018
total_rec_late_fee 0.054 -0.05 0.027 0.093 0.047 0.031 1 0.13 0.18 -0.023
total_pymnt 0.66 0.26 0.22 0.63 0.24 -0.018 0.13 1 0.36 0.035
last_pymnt_amnt 0.55 0.021 0.15 0.62 0.21 -0.0024 0.18 0.36 1 0.043
issue_d_month 0.049 0.058 0.067 0.03 0.00074 -0.018 -0.023 0.035 0.043 1
loan_amnt
term
int_rate
installment
annual_inc
delinq_2yrs
total_rec_late_fee
total_pymnt
last_pymnt_amnt
issue_d_month
1.0
0.8
0.6
0.4
0.2
0.0

Conclusions

- Loan is usually borrowed in 5k intervals by most number of people, spikes at 5k intervals say so.
- People whose term is 36 months are more likely to default than those who have term as 60 months.
- People who are in mortgage agreement or in rent are more likely to default than people who have own houses with maximum number of defaults coming from people who live in rented homes.
- Not verified people are more likely to default than verified and verified ones are more likely to default than ones whose source of income is verified
- Grades B,C and D account for most number of defaults.
- Mortgage-Verified and Rent-Not Verified are the most defaulting categories, though Mortgage-Verified is a peculiar category, indicating that it might be possible that verification is not done properly for Mortgage category of customers.
- People who are staying in rented homes and have grades B, C and D are more likely to default. Same applies for people staying in mortgage homes.
- The people who are Charged Off have least overall annual income.
- The average interest rate of Charged Off customers is more than that off Fully Paid customers(ignoring the Current Segment). Therefore it can be said that high interest rate loans are more likely to default. The same trend can be seen for monthly installment as well, i.e Charged Off people tend to have higher monthly installments.
- We see that as loan_amnt increases, the monthly installment also increases, and people often tend to take hefty loans and fail to pay their monthly installment and default on loan.

Variables affecting loan defaults

1. term
2. home ownership
3. verification status
4. grades assigned to loan
5. annual income
6. interest rate
7. Combination of home ownership and verification status

Thank You