Lending Club Case Study

Manjiri jog
Chaitra R

Problem Statement

- To provide meaningful and useful insights to a company that specialises in lending various types of loans to urban customers. The decision of whether to lend money to customers is based on their profile and has certain risks, which if taken, may result into financial losses to company.
- The data about past loan applicants and whether they ‘defaulted’ or not, is given in CSV format.
- **The aim is to identify patterns which indicate if a customer is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, adjusting the interest rate, etc.**

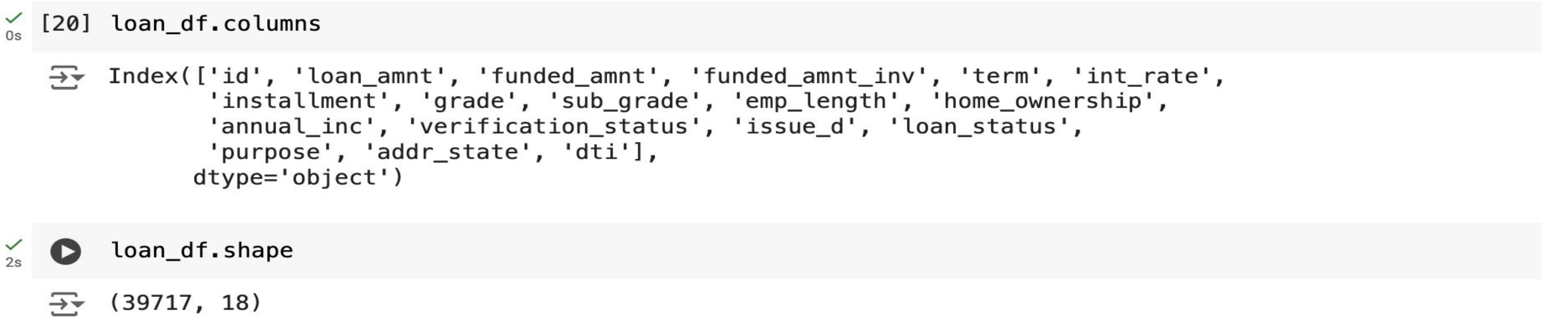
This is a near-real world scenario to be analyzed using the EDA. It involves study for some basic understanding of the domain of risk analysis, and it gives better understanding of the data observation, preparation and visualization

Data summary

- There are 39717 rows and 111 columns.

Data Observation and Cleaning

- There are no headers/footers/summary text.
- There are 54 columns with null/na values.
- There are around 9 columns having only 1 unique value.
- **Unique Identifiers**: Columns "id", "url", and "member_id" are all unique throughout the data.
- **Uninformative Job Titles**: Columns "emp_title" and "title" contain mostly unique values, but they represent job titles which likely don't significantly influence loan default analysis.
- There are around 25 columns which are demographic information , or information collected after loan_approval.
- We are dropping around 30 columns.
- Total of 93 columns are dropped and analysis is done using 18 columns.



```
loan_df.columns
```

```
Index(['id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate',
       'installment', 'grade', 'sub_grade', 'emp_length', 'home_ownership',
       'annual_inc', 'verification_status', 'issue_d', 'loan_status',
       'purpose', 'addr_state', 'dti'],
      dtype='object')
```

```
loan_df.shape
```

```
(39717, 18)
```

Data Preparation - imputing/cleaning

- Converting the term column into int by replacing months.
- Converting int_rate column to float to get precise values.
- In case of loan_status in current gives information of loan current still payed it does not give any valuable analysis for our case .Dropping the rows.
- Cleaning the emp_length column and converting to int.
- There are only 3 values with NONE type of ownership.Dropping them as it does not categorise into any value.
- Derived column issued_year and issued_month from issue_d.
- Derived a column loan_status_indic to convert ”Charged Off” to 0 and “Fully Paid ” to 1.
- In later sections derived bucket bins columns for annual_inc , int_rate,dti.

Univariate Analysis Primary Observations

Performed on each variable, to check distribution of the parameter

Some primary checks about Defaulters - which attributes clearly indicate higher chances of default cases.

1. loan_amnt: Most loans are in the range of 5,000 to 15,000.

2. funded_amnt: Similar distribution to loan_amnt.

3. funded_amnt_inv: Also similar distribution.

4. term: Most loans are either 3 years or 5 years (36 or 60 months).

5. int_rate: Most loans have interest rates between 10% and 15%.

6. installment: Most installments are between 200 and 400.

7. annual_inc: Most borrowers have annual incomes between 40,000 and 80,000.

8. dti: Most borrowers have debt-to-income ratios between 10% and 20%.

9. emp_length: Most borrowers have been employed for 10 years or less.

Univariate Analysis Primary Observations

Observations for chances of Default cases (loan_status = 'Charged Off') based on Comparison of data distribution:

1. Higher loan amount / funded amount might be linked to increased charge-offs.

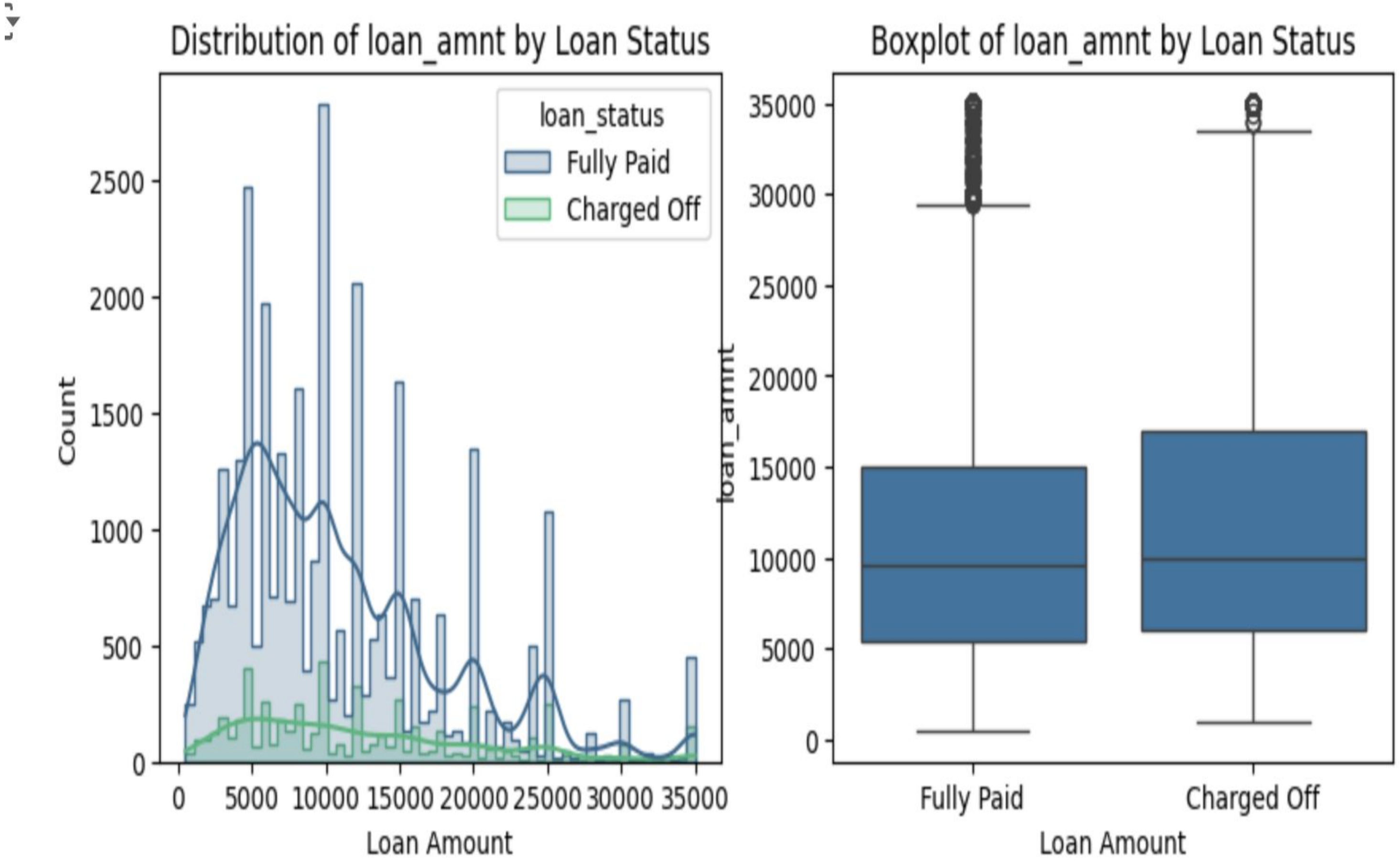
2. Loans with 60-month terms seem to have higher charge-off rates compared to 36-month term loans.

3. There is positive correlation between higher interest rates and charge-offs.

Univariate Analysis on Loan Amount

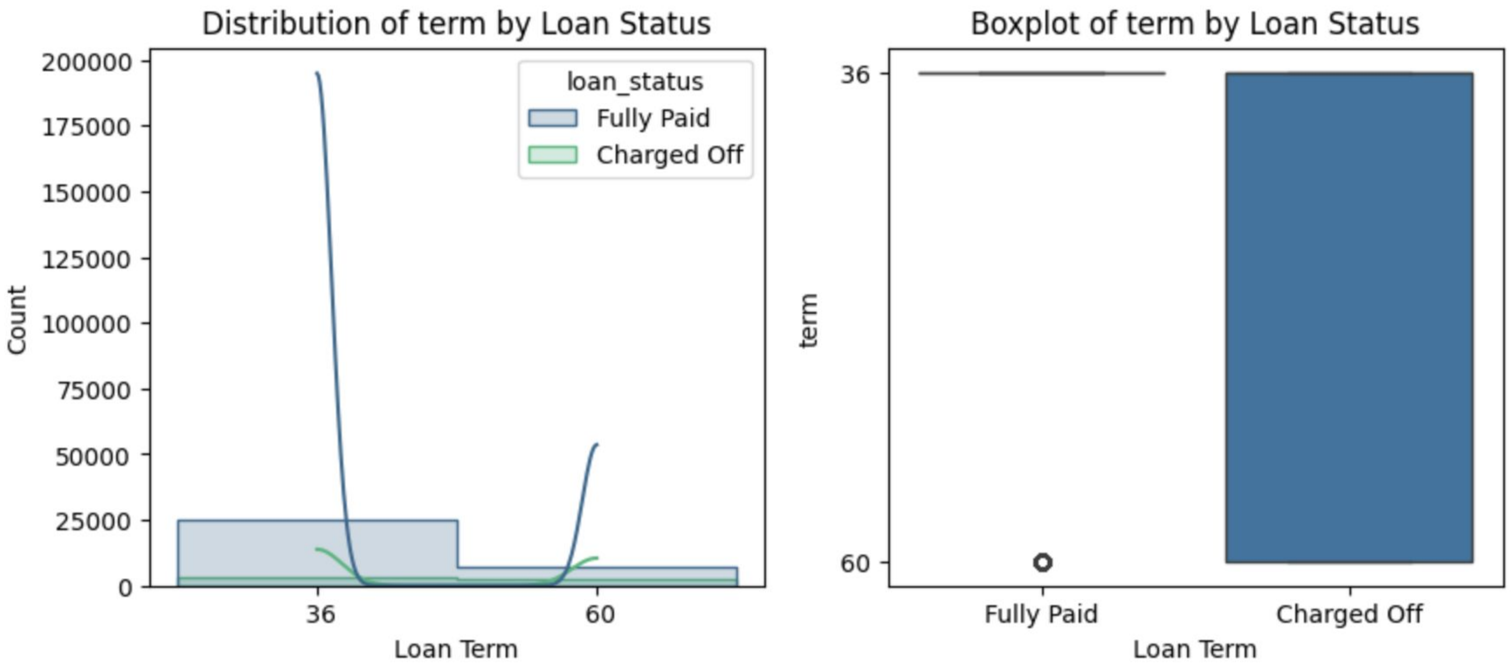
- Most loans are in the range of 5,000 to 15,000
- funded_amnt: Similar distribution to loan_amnt. funded_amnt_inv: Also similar distribution.

- Higher loan amount / funded amount might be linked to increased charge-offs.



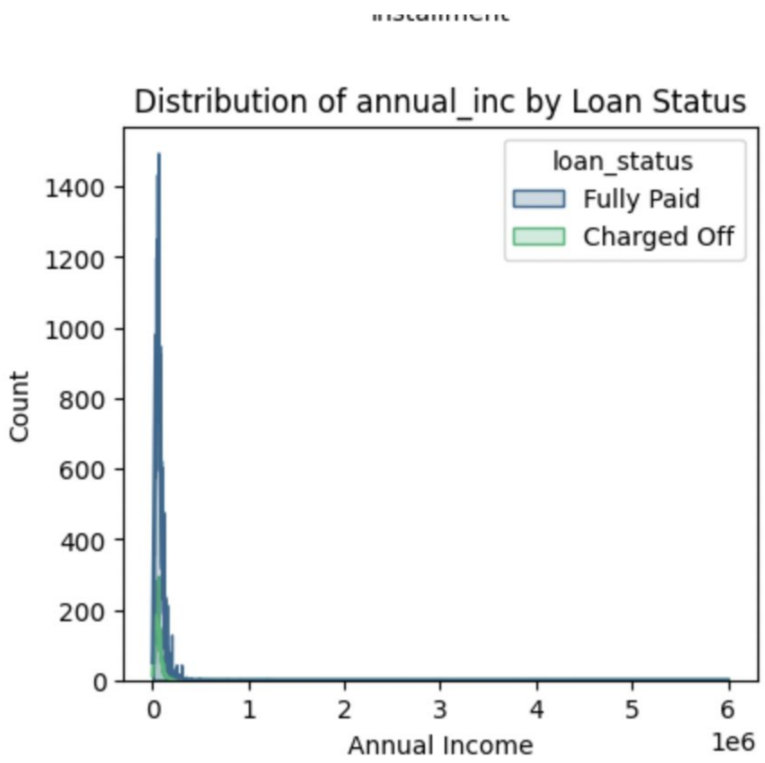
Univariate Analysis on Term

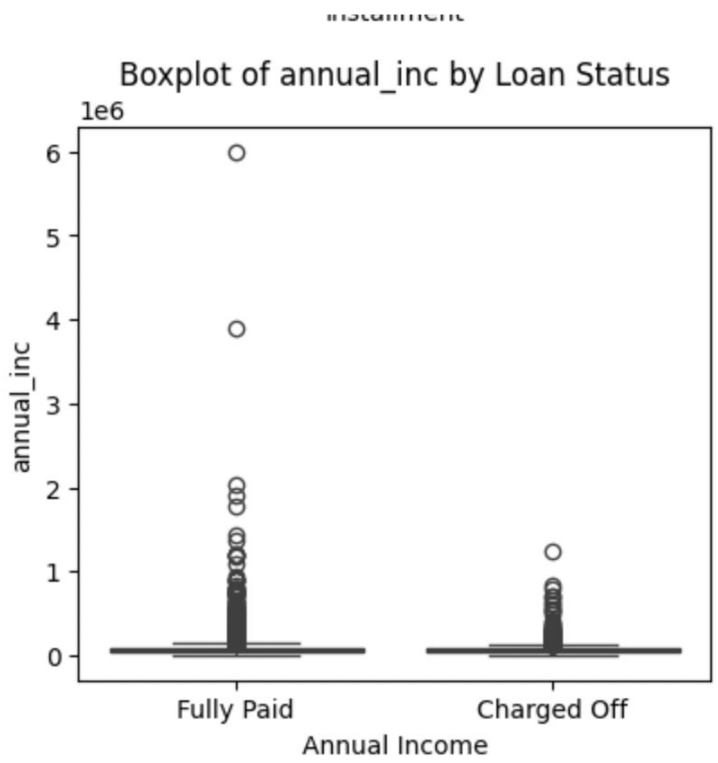
- Loans are given for either 36 or 60 months.



Univariate Analysis on Annual Income

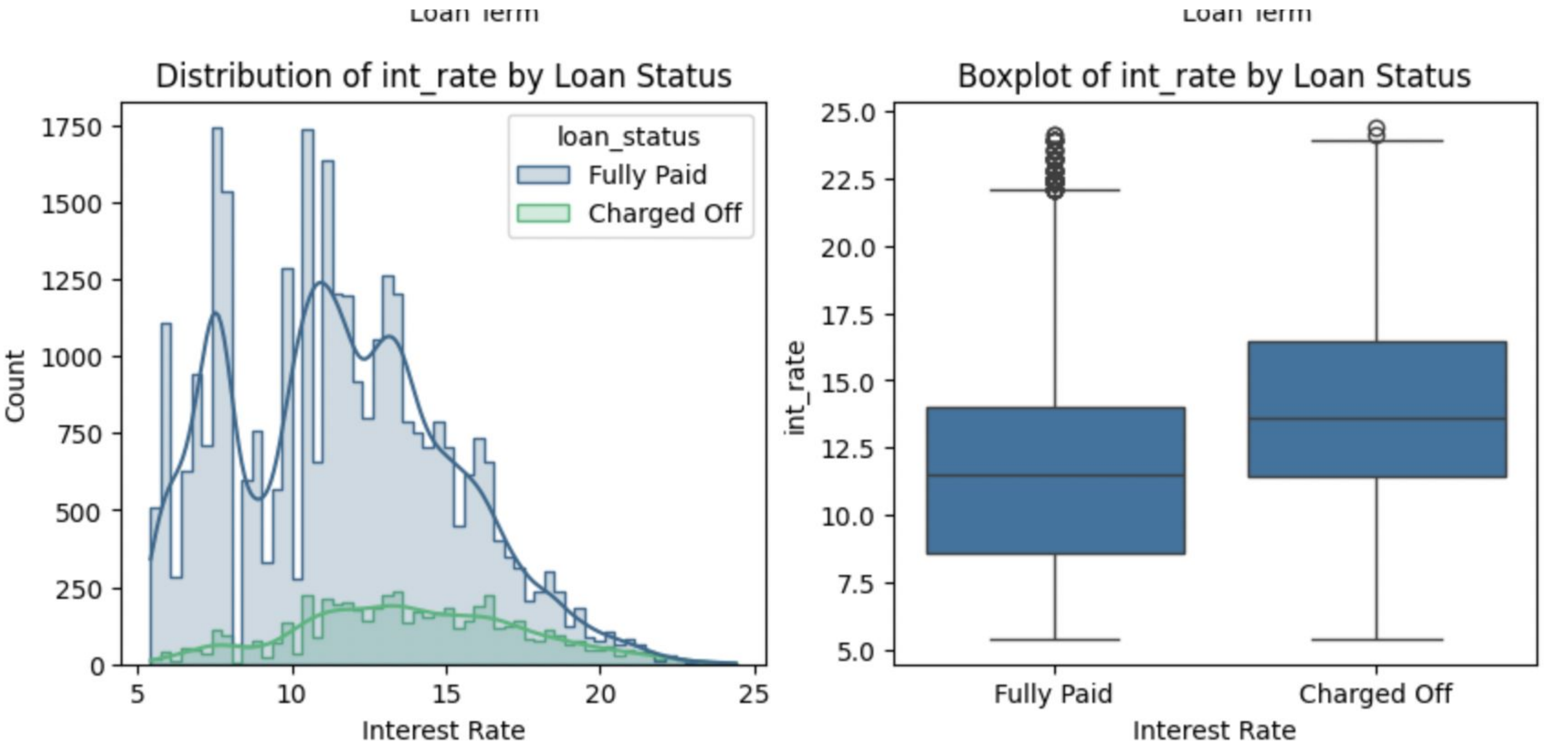
- Most loans are in the range of 5,000 to 15,000
- There seems to be outliers.It needs to be treated.





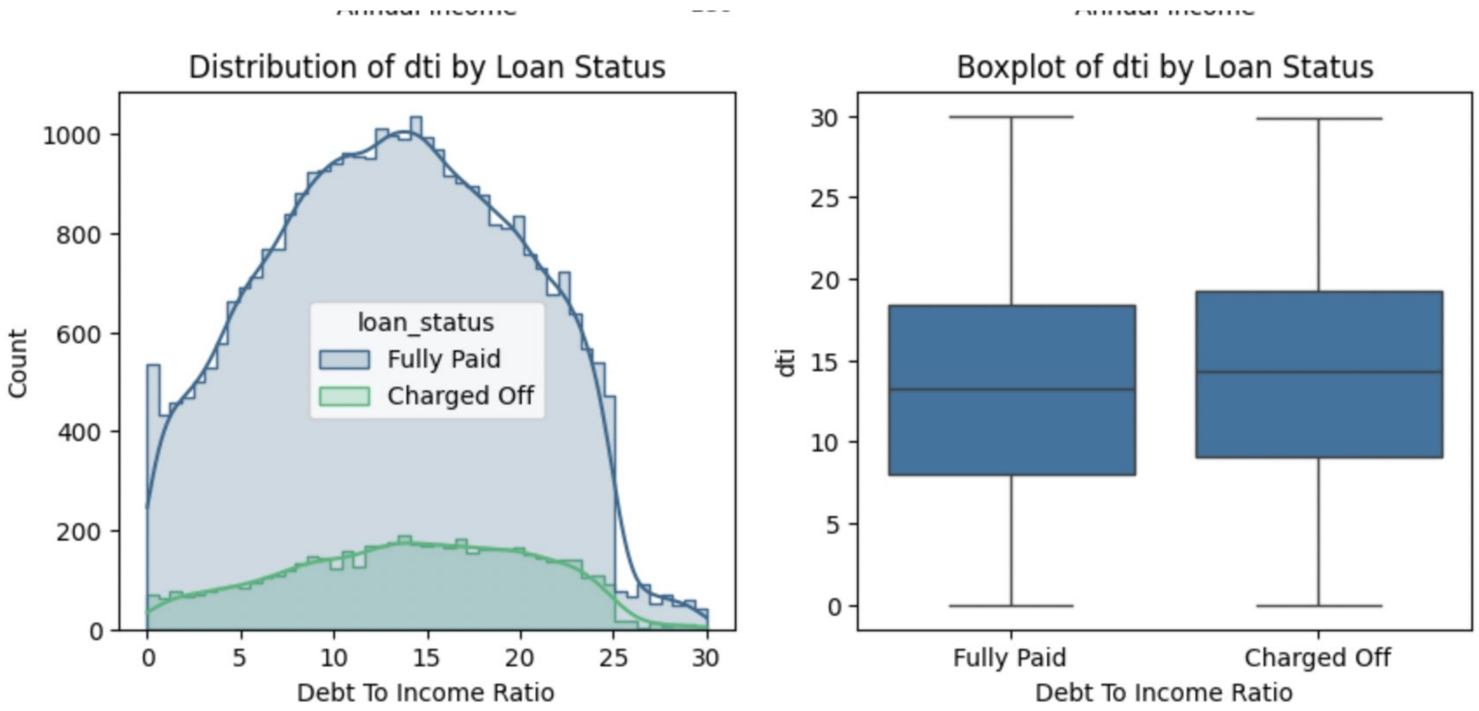
Univariate Analysis on Interest Rate

- Large number of loans charge interest rates between 10% and 15%.
- There is strong impact on the interest on higher percentage in case of charged-off



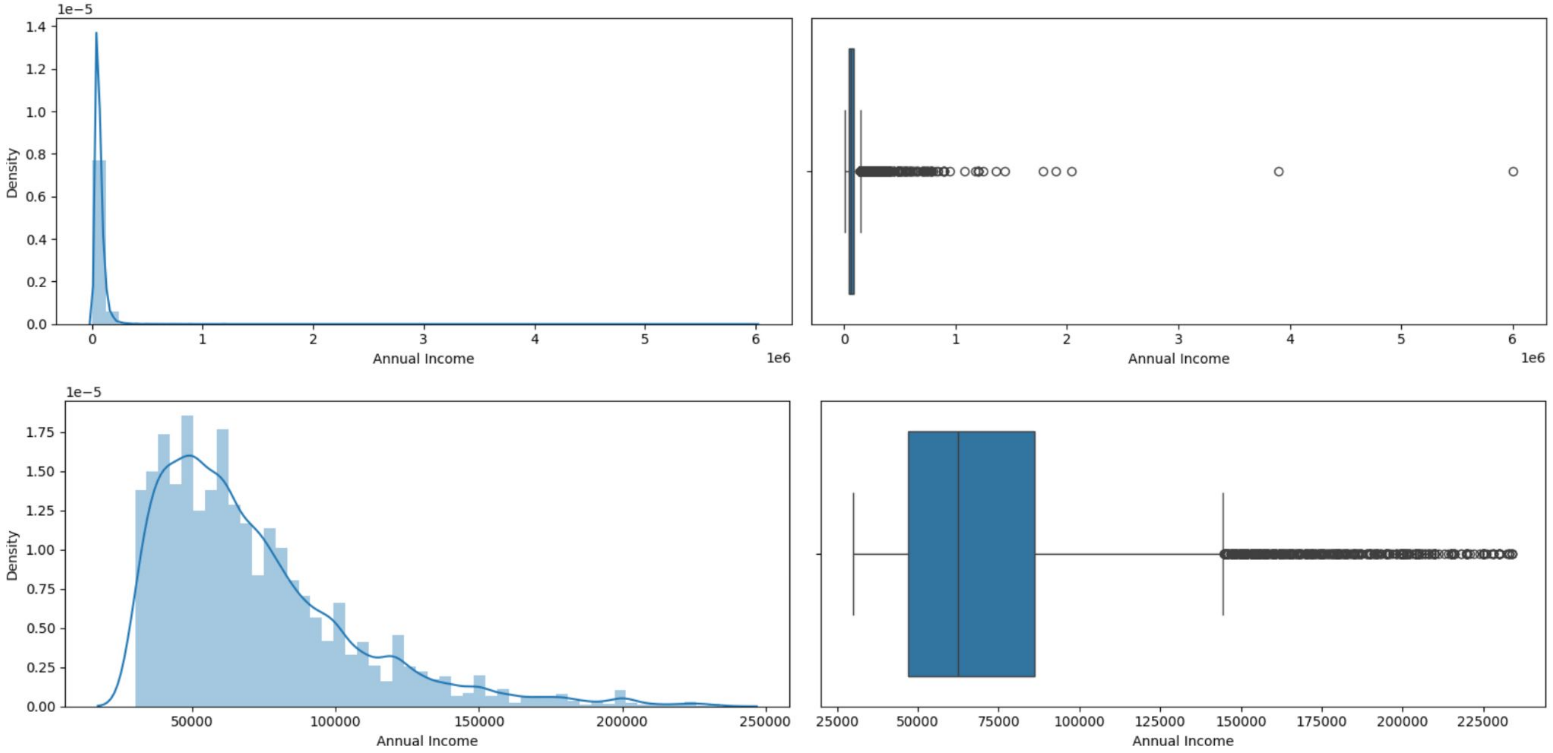
Univariate Analysis on DTI

- Most borrowers have debt to income ratio between 10% and 20%.



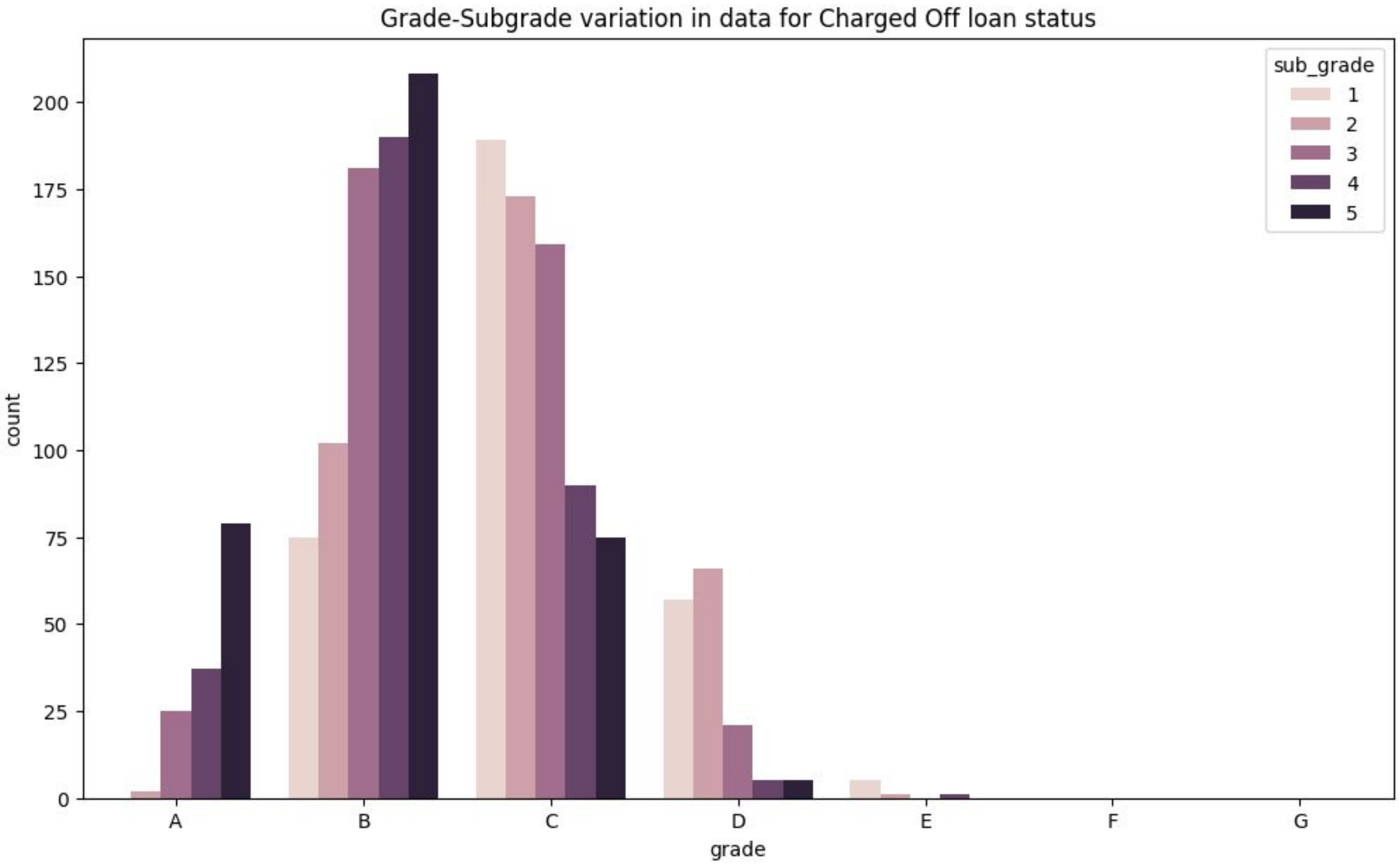
Outliers Treatment

1. Outlier treatment was done for loan amount, interest rate, debt-to-income ratio, and annual income.
2. Major impact was observed on the annual income.
3. The range of annual income chosen is 30000 to 234144



Univariate Analysis on Grades and Subgrades

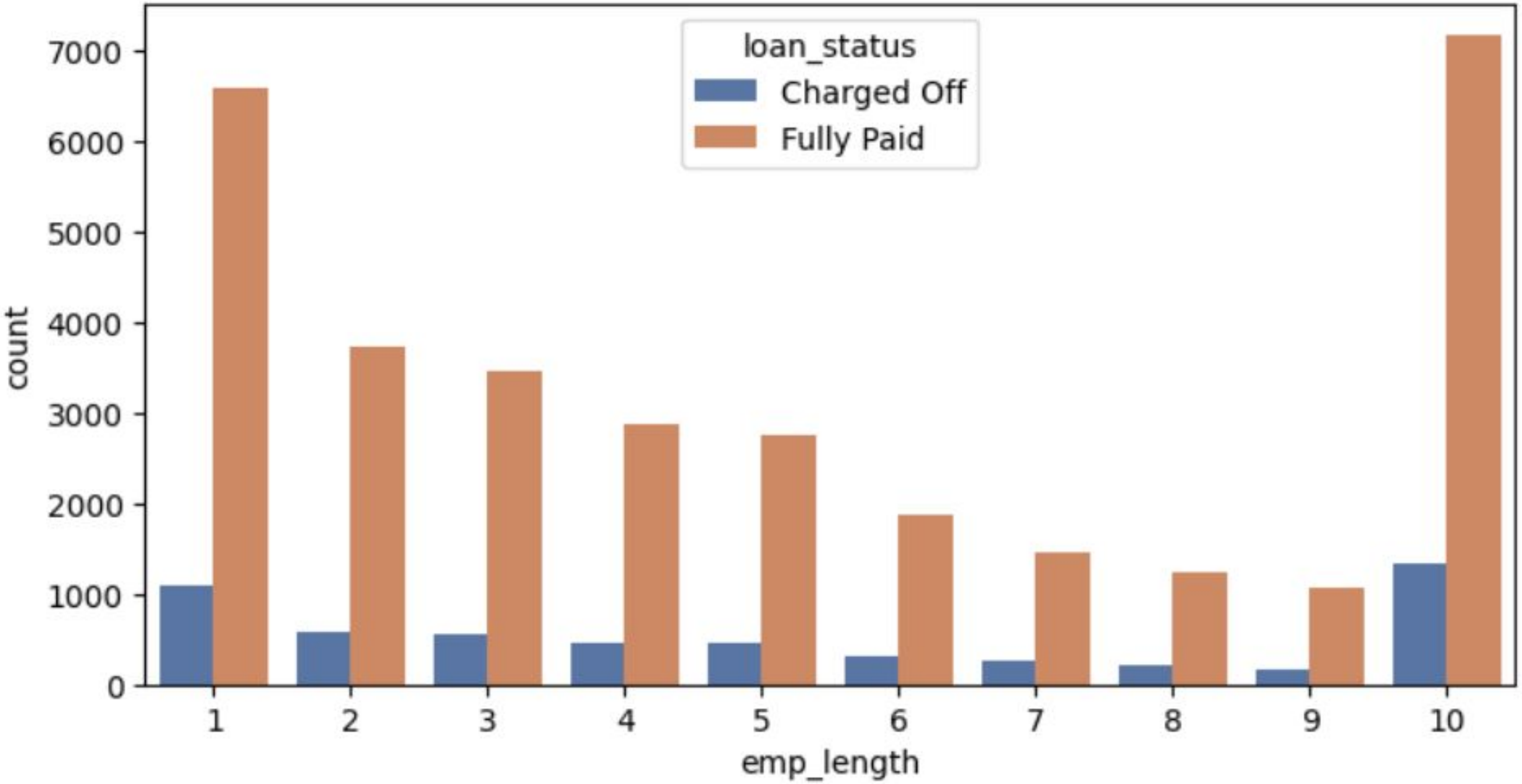
- Chances of Defaulting are high for Grade B5 and C1.



Univariate Analysis on Employment length

Customers with the least and the most employment experience are not defaulters

But, customers with 10+ years of experience are highest in numbers and have higher number of defaulters as well.



Observations about Defaulters

1. Purpose of loan amount is mostly for debt_consolidation in both cases - Fully Paid/Charged Off

2. Home ownership is Rent or Mortgage where there are more number of borrowers.

3. Chances of Defaulting are high for Grade B5 and C1.

4. Borrowers with Lower annual income have most number of defaulters.

5. Higher debt-to-income ratio indicates high chances of defaulting.

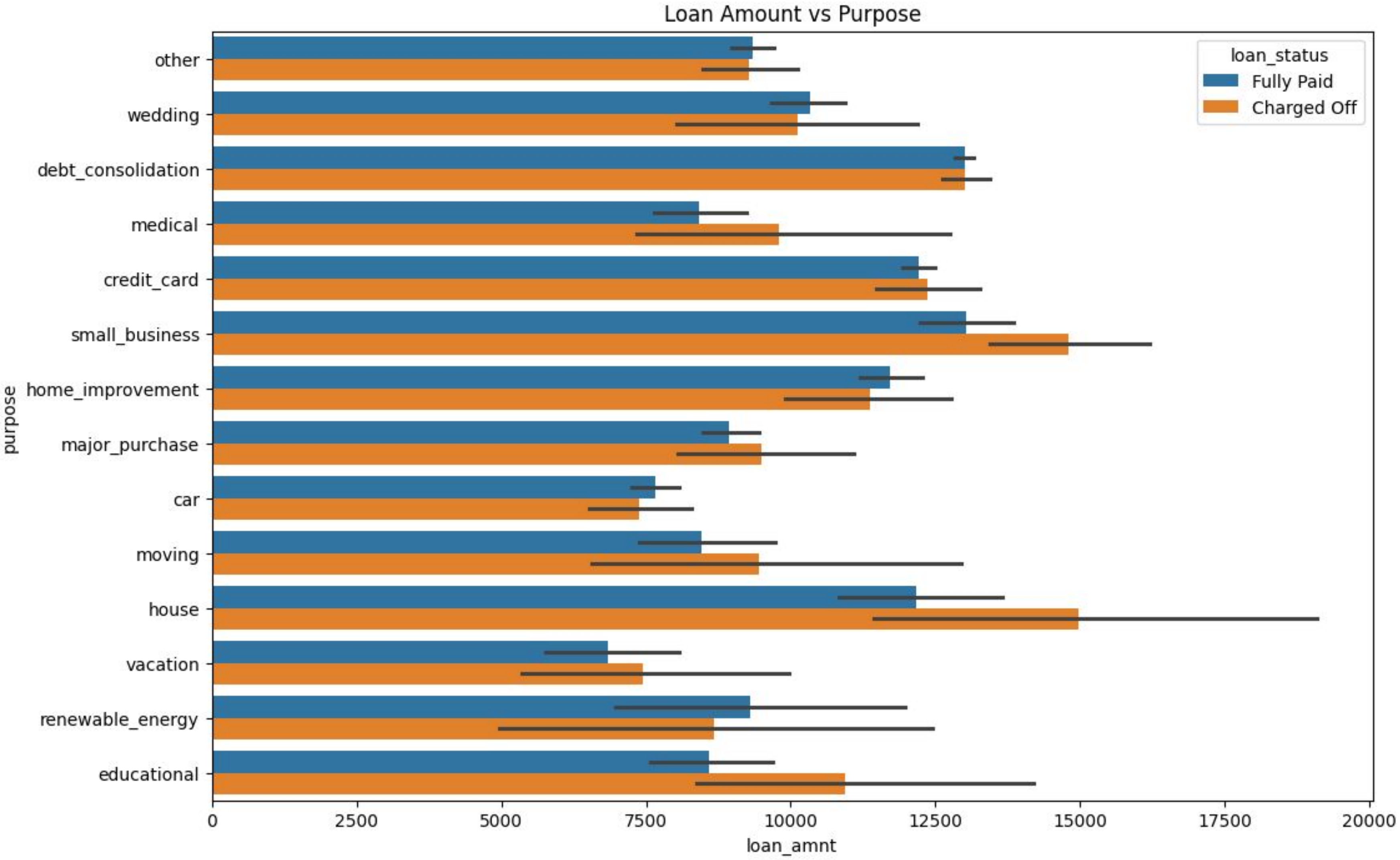
6. Higher interest rate on loan can lead to high chances of defaulting.

7. Customers with more than 10 years of employment are maximum in number and have highest number of defaulters

Bivariate Analysis

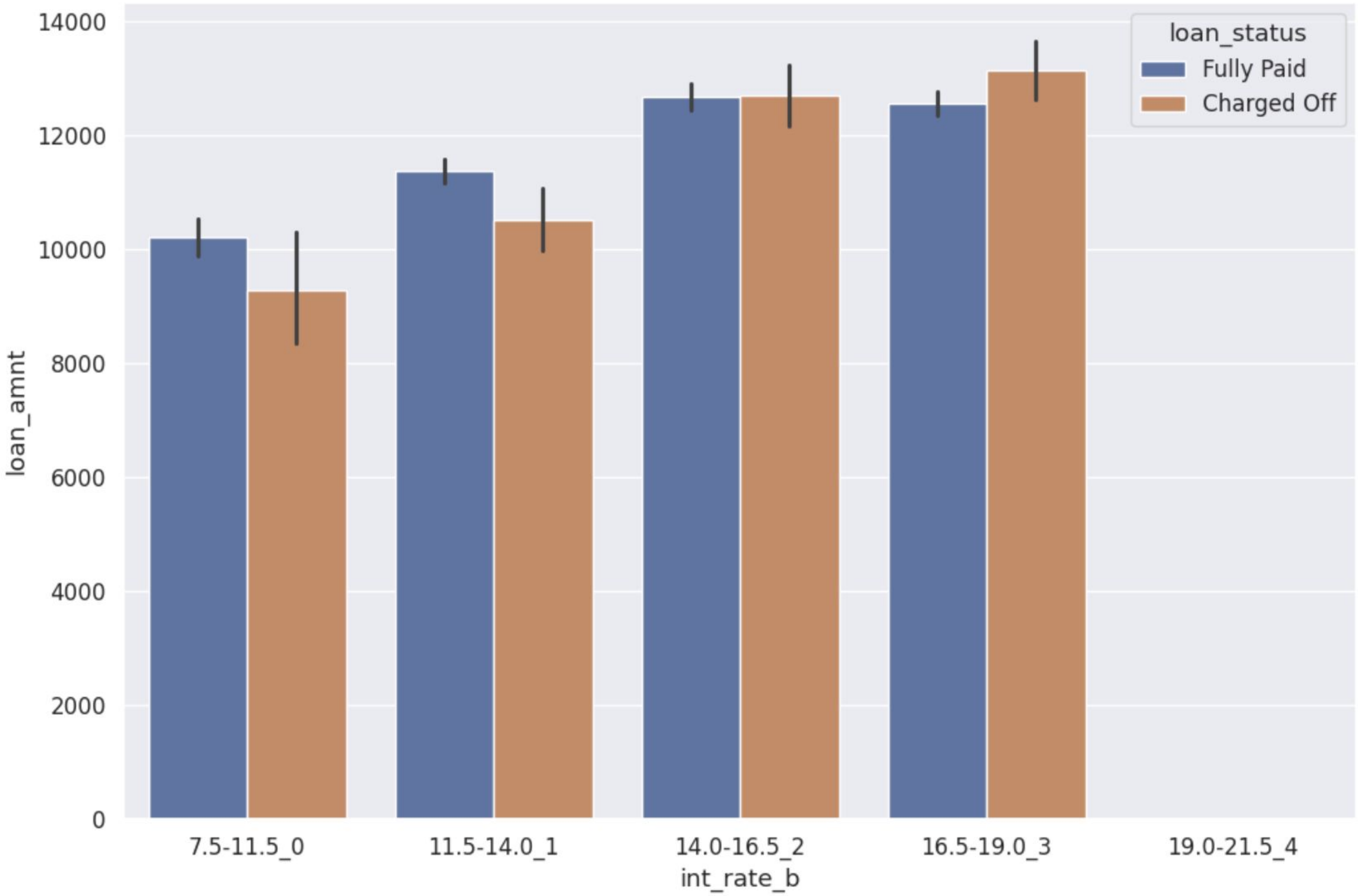
Loan Amount vs Purpose Analysis

- Loans amounts above 12500 and for small businesses or house are likely to default.
- Also, education loans above 10000 are likely to default.



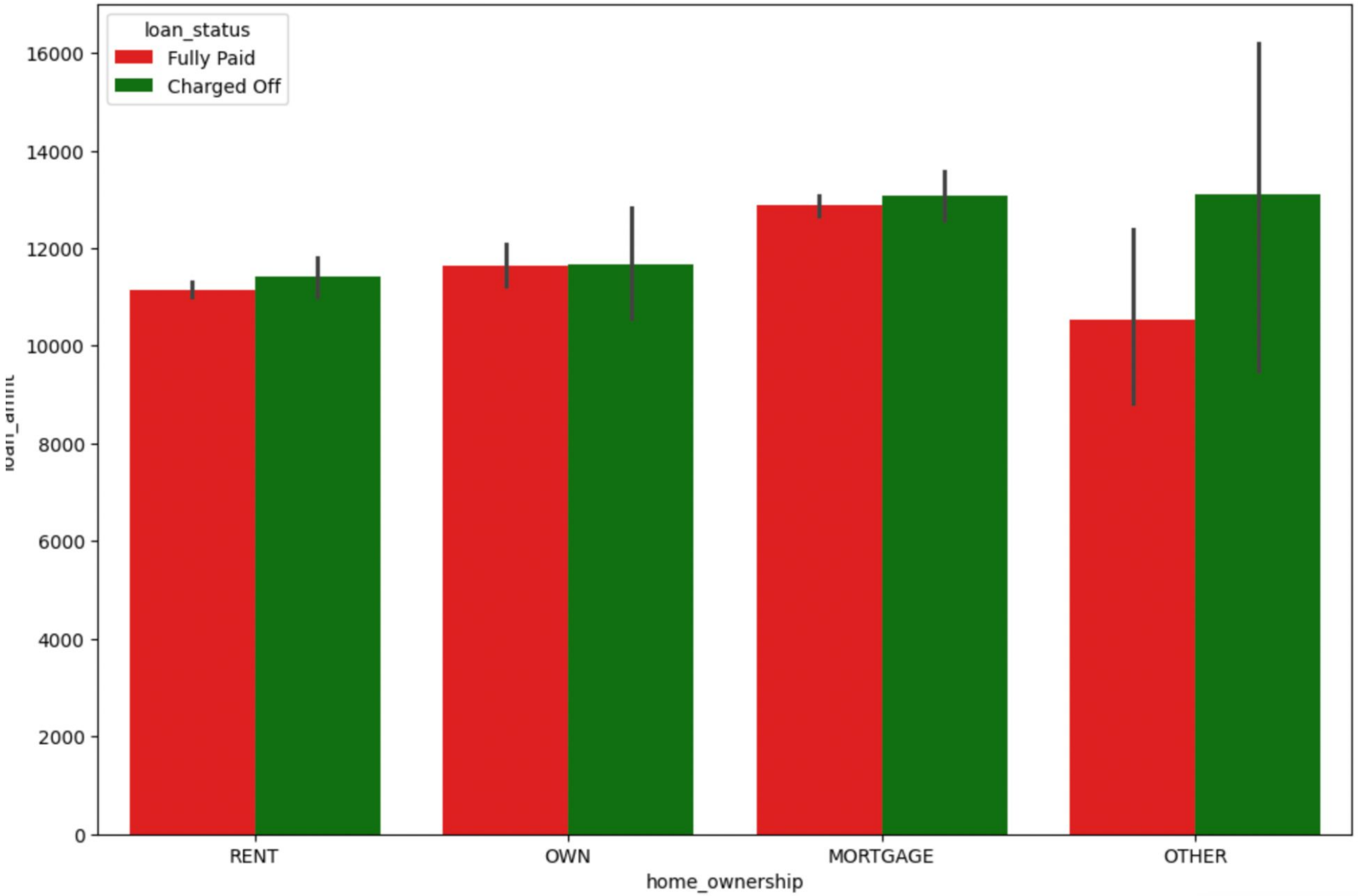
Interest Rate Amount and Loan Status

- Interest rate above 14% are most likely to default



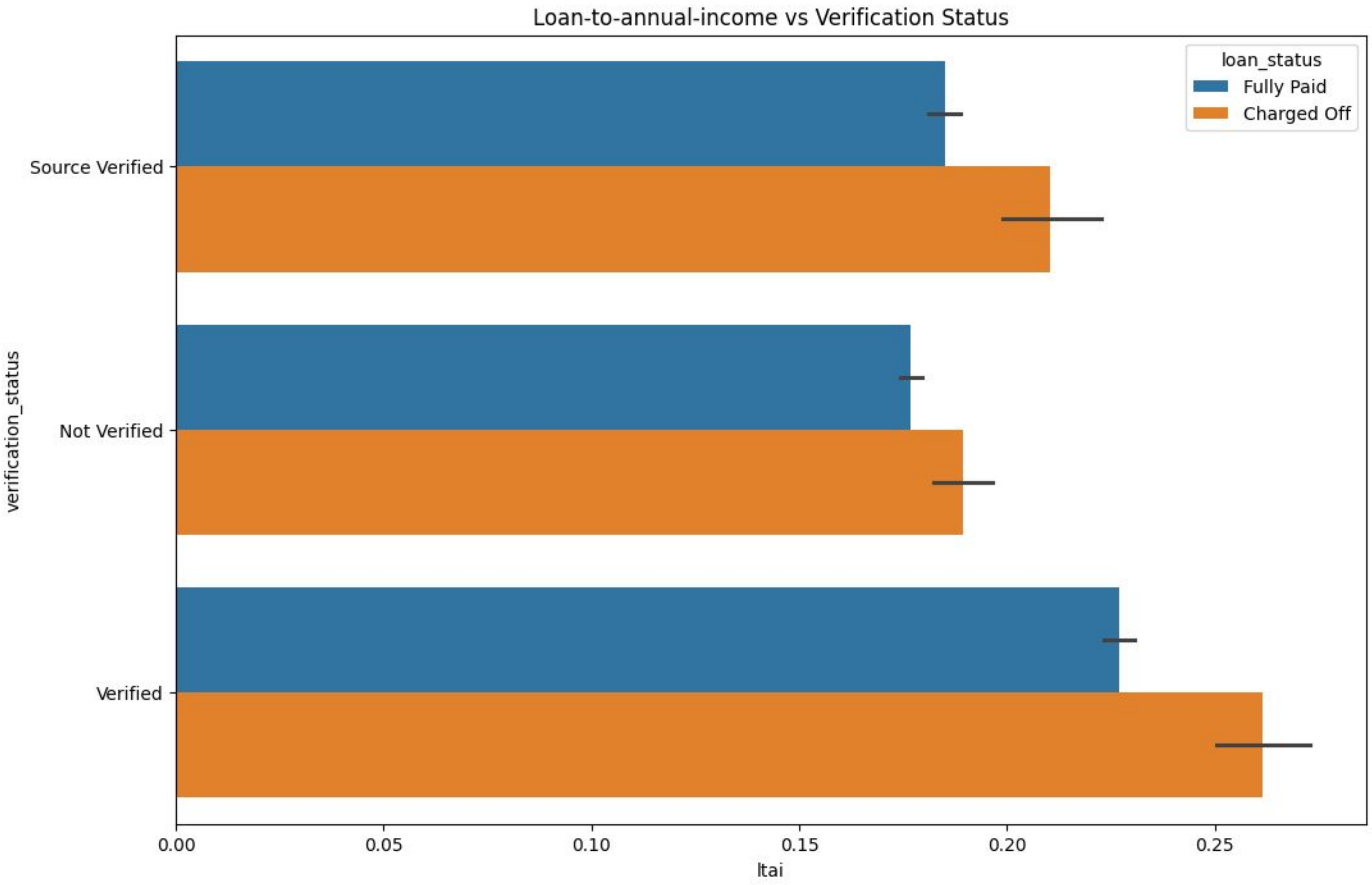
Home Ownership vs Loan Amount

- Loan owners in the category of “OTHER” are mostly defaulted.



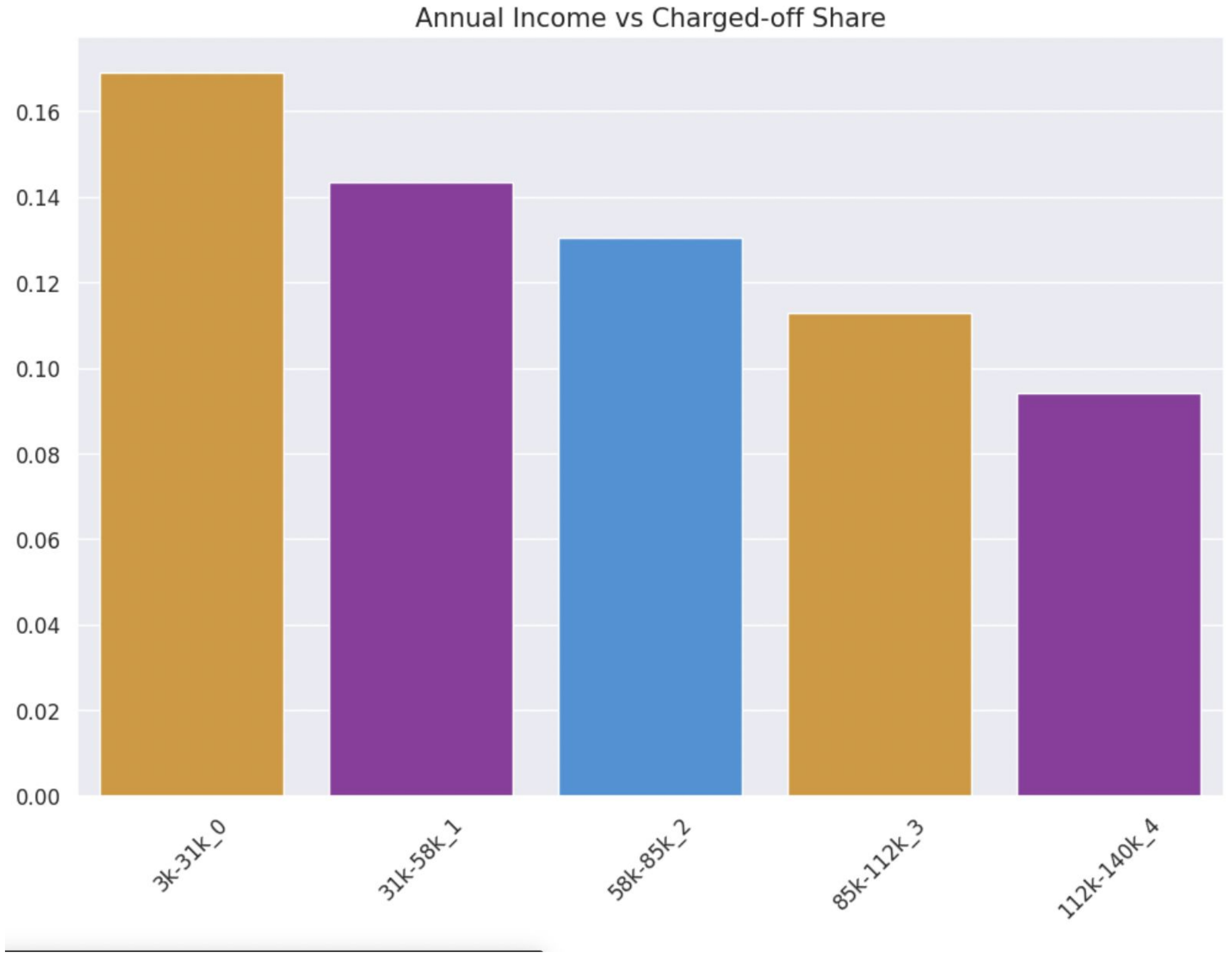
Loan-to-Annual-Income vs Verification Status

This is an interesting insight - despite the income was verified by lending club or at source, the Default cases are more if the loan-to-annual-income ratio is high.



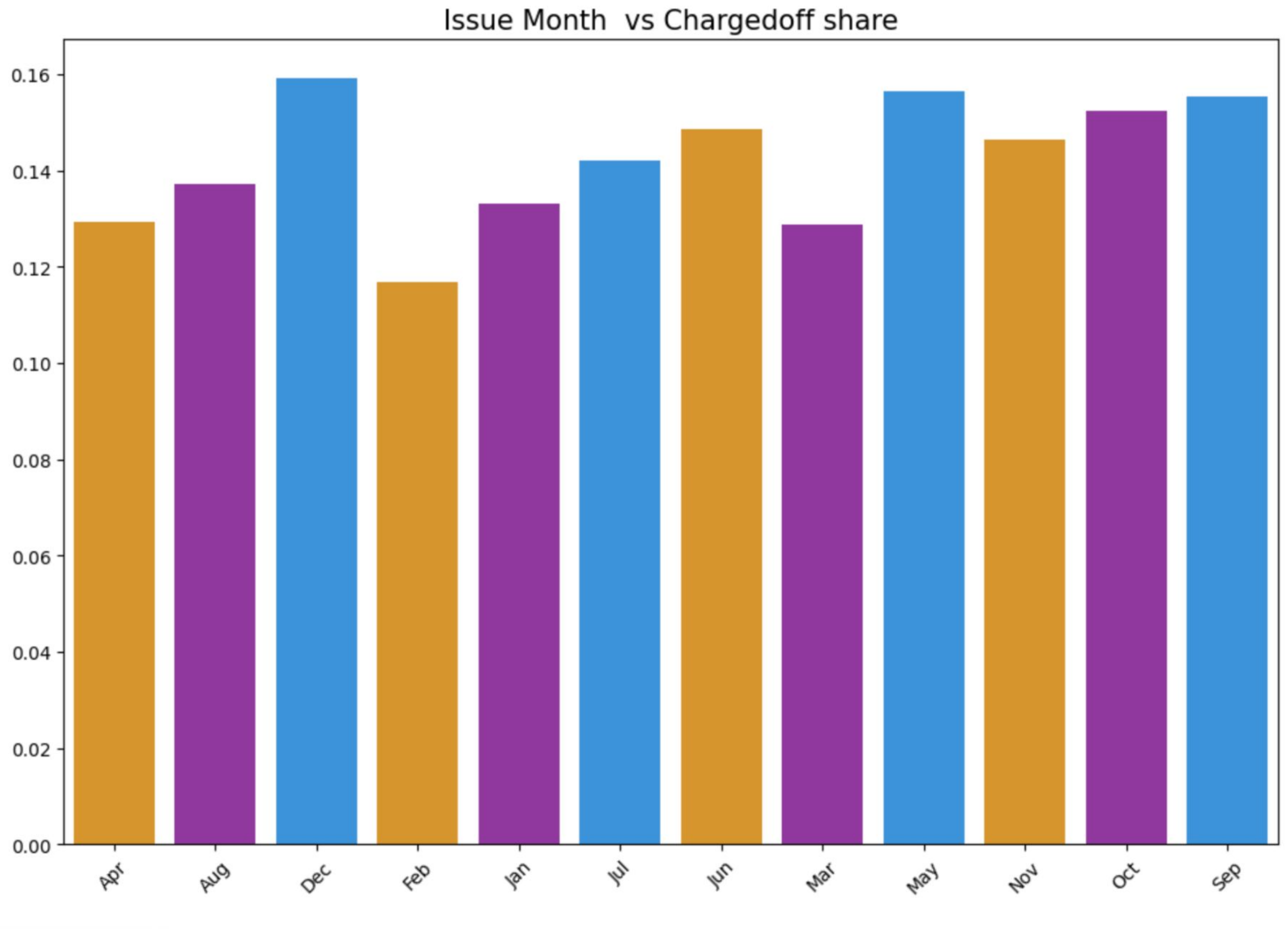
Annual income and charged off share

- Loan borrowers in range 3000-30000 are mostly to default.



Issue_month and charged off share

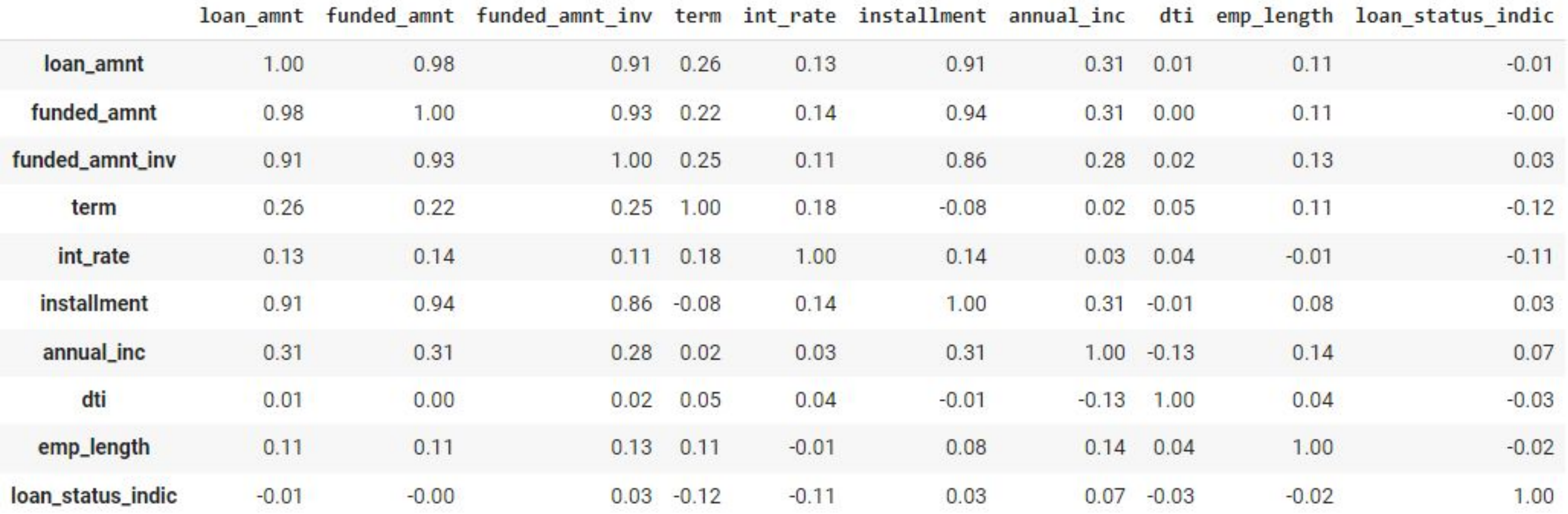
- Most of the loan bearers seemed to have a charged off share higher than 0.12 , but we see loans issued in Dec , May , september are likely to be charged off more



Bivariate Analysis

Plotted a Correlation Matrix to find the correlation among the data variables, including the converted Loan-Status-Indicator

Visual representation of the correlation matrix: Heatmap

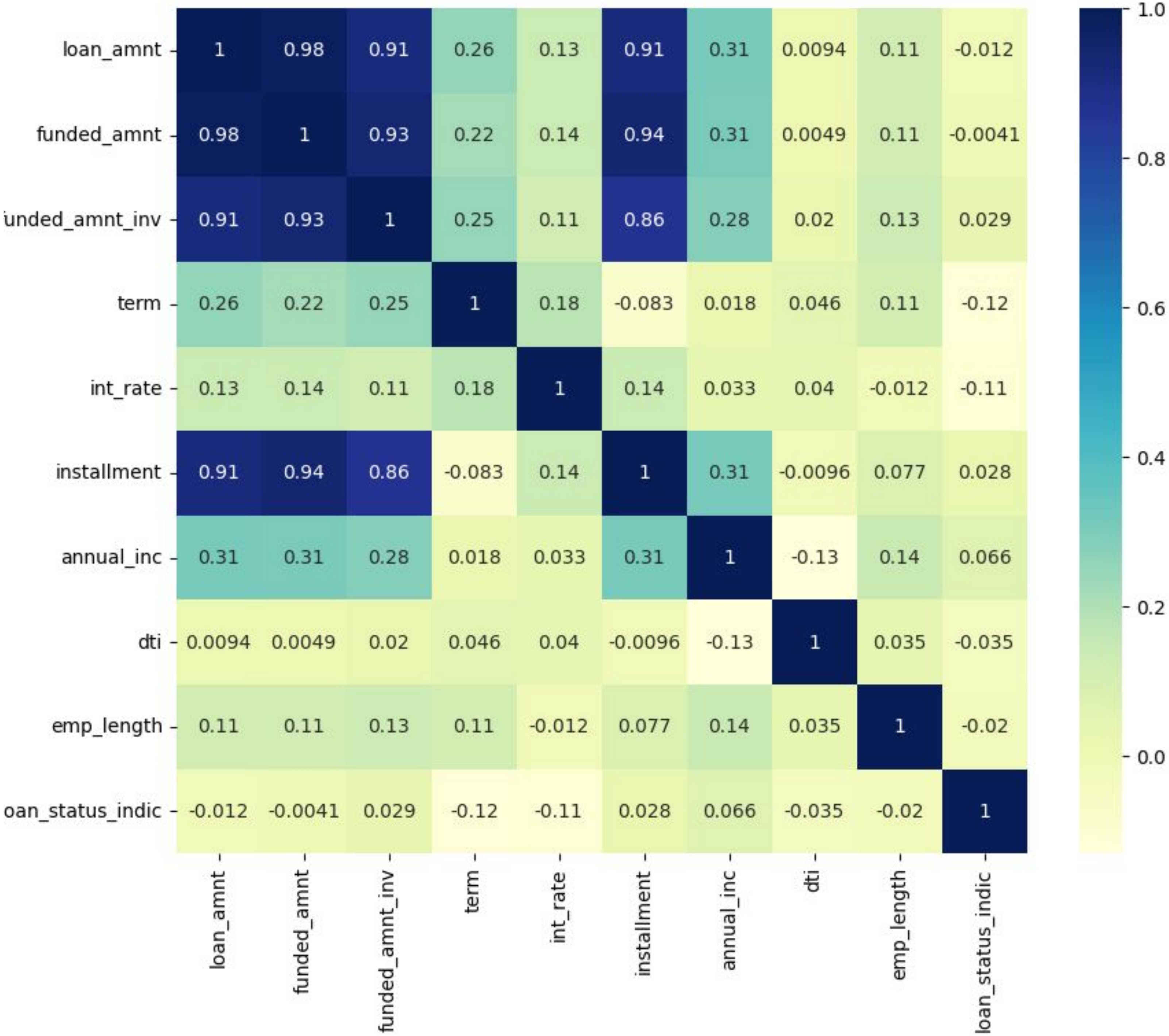


	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	annual_inc	dti	emp_length	loan_status_indic
loan_amnt	1.00	0.98	0.91	0.26	0.13	0.91	0.31	0.01	0.11	-0.01
funded_amnt	0.98	1.00	0.93	0.22	0.14	0.94	0.31	0.00	0.11	-0.00
funded_amnt_inv	0.91	0.93	1.00	0.25	0.11	0.86	0.28	0.02	0.13	0.03
term	0.26	0.22	0.25	1.00	0.18	-0.08	0.02	0.05	0.11	-0.12
int_rate	0.13	0.14	0.11	0.18	1.00	0.14	0.03	0.04	-0.01	-0.11
installment	0.91	0.94	0.86	-0.08	0.14	1.00	0.31	-0.01	0.08	0.03
annual_inc	0.31	0.31	0.28	0.02	0.03	0.31	1.00	-0.13	0.14	0.07
dti	0.01	0.00	0.02	0.05	0.04	-0.01	-0.13	1.00	0.04	-0.03
emp_length	0.11	0.11	0.13	0.11	-0.01	0.08	0.14	0.04	1.00	-0.02
loan_status_indic	-0.01	-0.00	0.03	-0.12	-0.11	0.03	0.07	-0.03	-0.02	1.00

Heatmap

High correlation exists among loan amount, funded amount, funded_amnt_inv and installment.

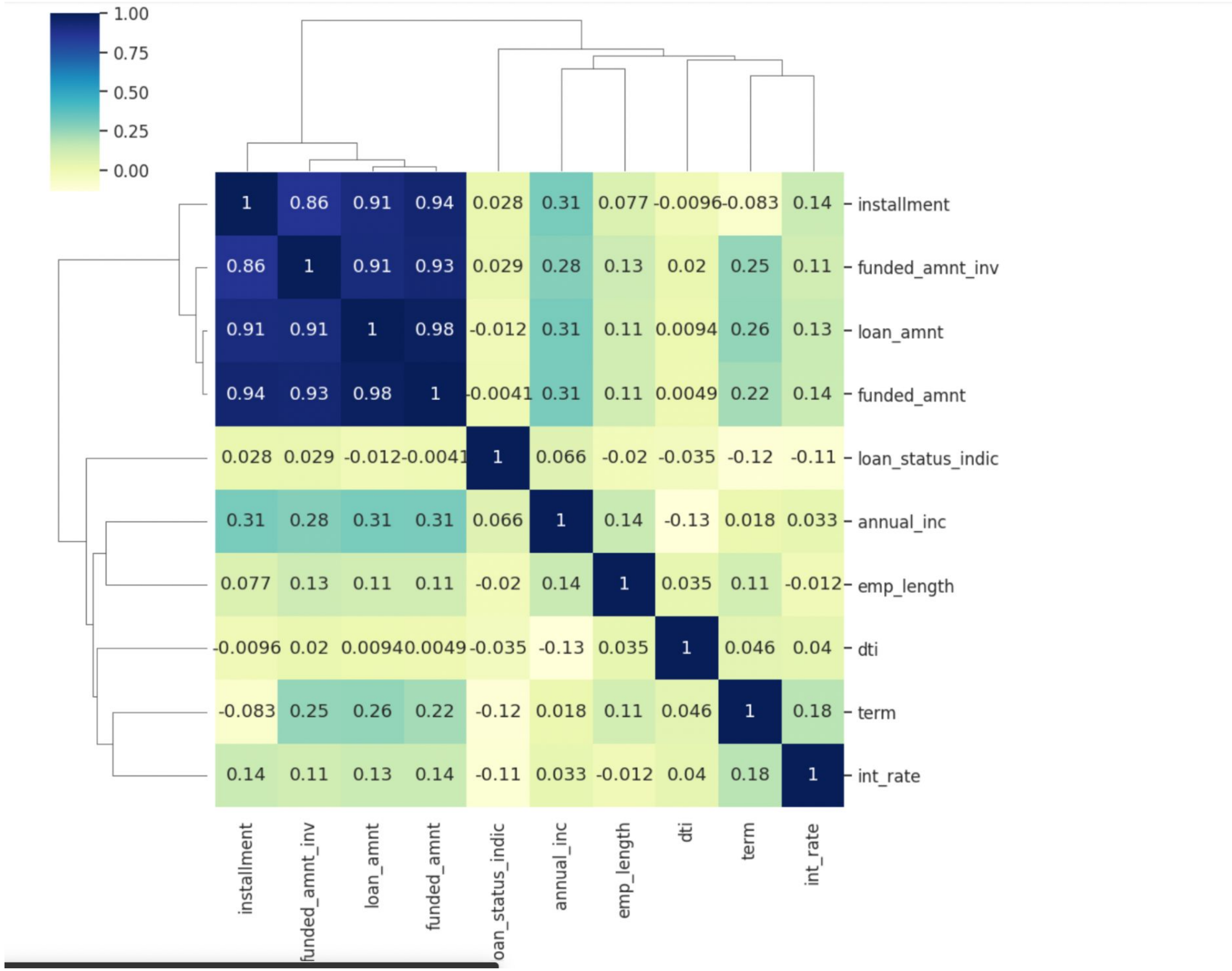
Medium correlation exists between Annual_inc and term.



	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	annual_inc	dti	emp_length	loan_status_indic
loan_amnt	1	0.98	0.91	0.26	0.13	0.91	0.31	0.0094	0.11	-0.012
funded_amnt	0.98	1	0.93	0.22	0.14	0.94	0.31	0.0049	0.11	-0.0041
unded_amnt_inv	0.91	0.93	1	0.25	0.11	0.86	0.28	0.02	0.13	0.029
term	0.26	0.22	0.25	1	0.18	-0.083	0.018	0.046	0.11	-0.12
int_rate	0.13	0.14	0.11	0.18	1	0.14	0.033	0.04	-0.012	-0.11
installment	0.91	0.94	0.86	-0.083	0.14	1	0.31	-0.0096	0.077	0.028
annual_inc	0.31	0.31	0.28	0.018	0.033	0.31	1	-0.13	0.14	0.066
dti	0.0094	0.0049	0.02	0.046	0.04	-0.0096	-0.13	1	0.035	-0.035
emp_length	0.11	0.11	0.13	0.11	-0.012	0.077	0.14	0.035	1	-0.02
oan_status_indic	-0.012	-0.0041	0.029	-0.12	-0.11	0.028	0.066	-0.035	-0.02	1

Clustermap

- Cluster 1:
  - - High correlation between loan_amnt, funded_amnt, funded_amnt_inv, and installment.
  - - These features are all related to the amount of money borrowed and the amount of money paid back.
- Cluster 2:
  - - High correlation between int_rate and dti.
  - - These features are both related to the risk of default.
- Cluster 3:
  - - High correlation between annual_inc and loan_status_indic.
  - - This indicates that borrowers with higher annual incomes are more likely to repay their loans.
- Cluster 4:
  - - High correlation between emp_length and loan_status_indic.
  - - This indicates that borrowers with longer employment histories are more likely to repay their loans.



	installment	funded_amnt_inv	loan_amnt	funded_amnt	oan_status_indic	annual_inc	emp_length	dti	term	int_rate
installment	1	0.86	0.91	0.94	0.028	0.31	0.077	-0.0096	-0.083	0.14
funded_amnt_inv	0.86	1	0.91	0.93	0.029	0.28	0.13	0.02	0.25	0.11
loan_amnt	0.91	0.91	1	0.98	-0.012	0.31	0.11	0.0094	0.26	0.13
funded_amnt	0.94	0.93	0.98	1	-0.0041	0.31	0.11	0.0049	0.22	0.14
loan_status_indic	0.028	0.029	-0.012	-0.0041	1	0.066	-0.02	-0.035	-0.12	-0.11
annual_inc	0.31	0.28	0.31	0.31	0.066	1	0.14	-0.13	0.018	0.033
emp_length	0.077	0.13	0.11	0.11	-0.02	0.14	1	0.035	0.11	-0.012
dti	-0.0096	0.02	0.0094	0.0049	-0.035	-0.13	0.035	1	0.046	0.04
term	-0.083	0.25	0.26	0.22	-0.12	0.018	0.11	0.046	1	0.18
int_rate	0.14	0.11	0.13	0.14	-0.11	0.033	-0.012	0.04	0.18	1

Observations and conclusions

- Higher the loan amount the more likely the person will default.
- In a annual income range below 31000 are most likely to default considering to range above 112000.
- Higher Loan amount in purpose of small Business , house are most likely to default.
- Loan borrowers from Grade B and subgrade 5 are to be defaulters. Similar with Grade C and subgrade 1.
- Loan borrowers who do not have a own home are most likely to default.
- Interest rate above 14% are most likely to default.
- Address state have an impact on the defaulters but further analysis is required in terms of geographic conditions and other factors

Thank you