

LENDING CLUB CASE STUDY

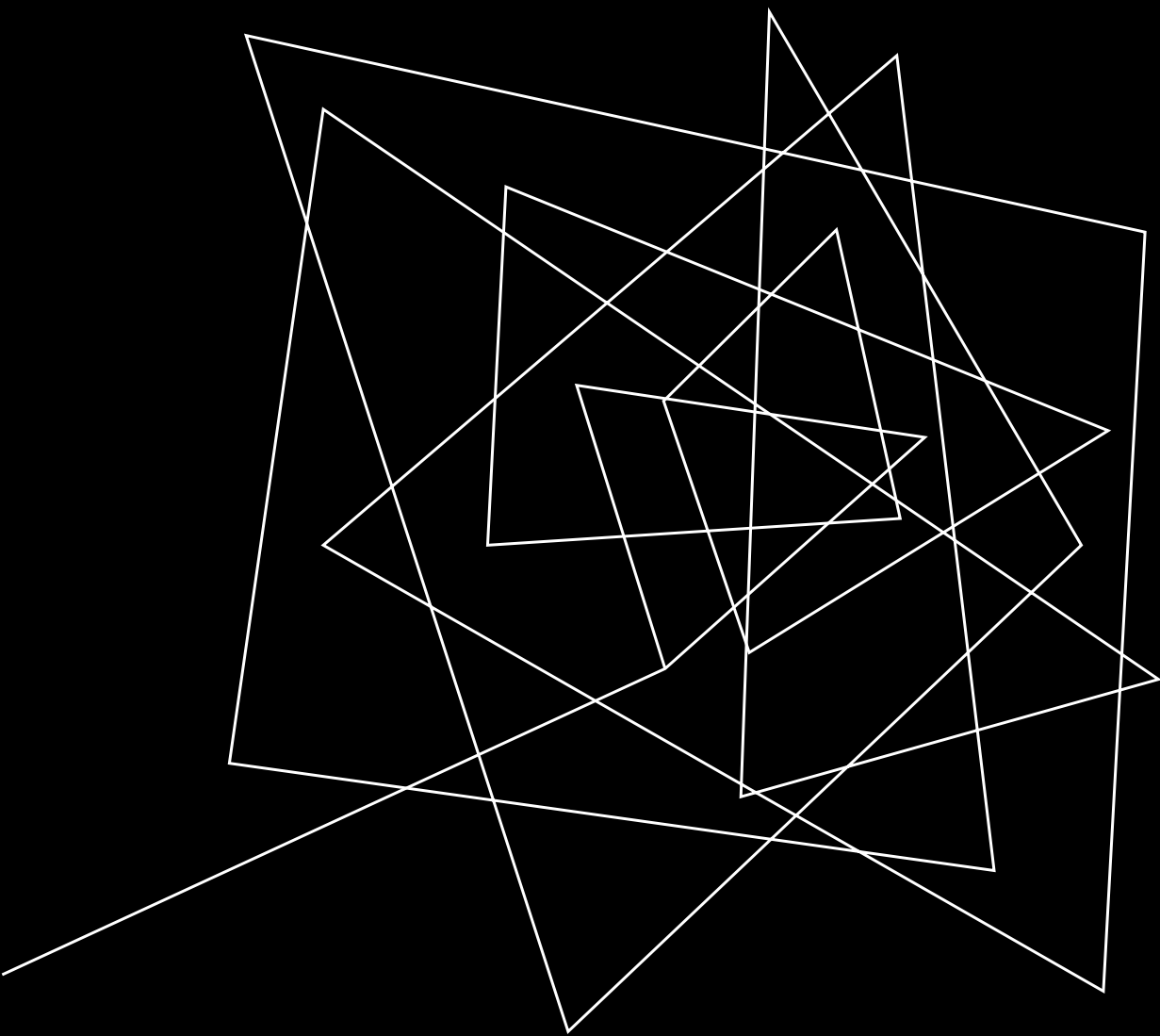
Snehal Zipare

Vijay Ranjane

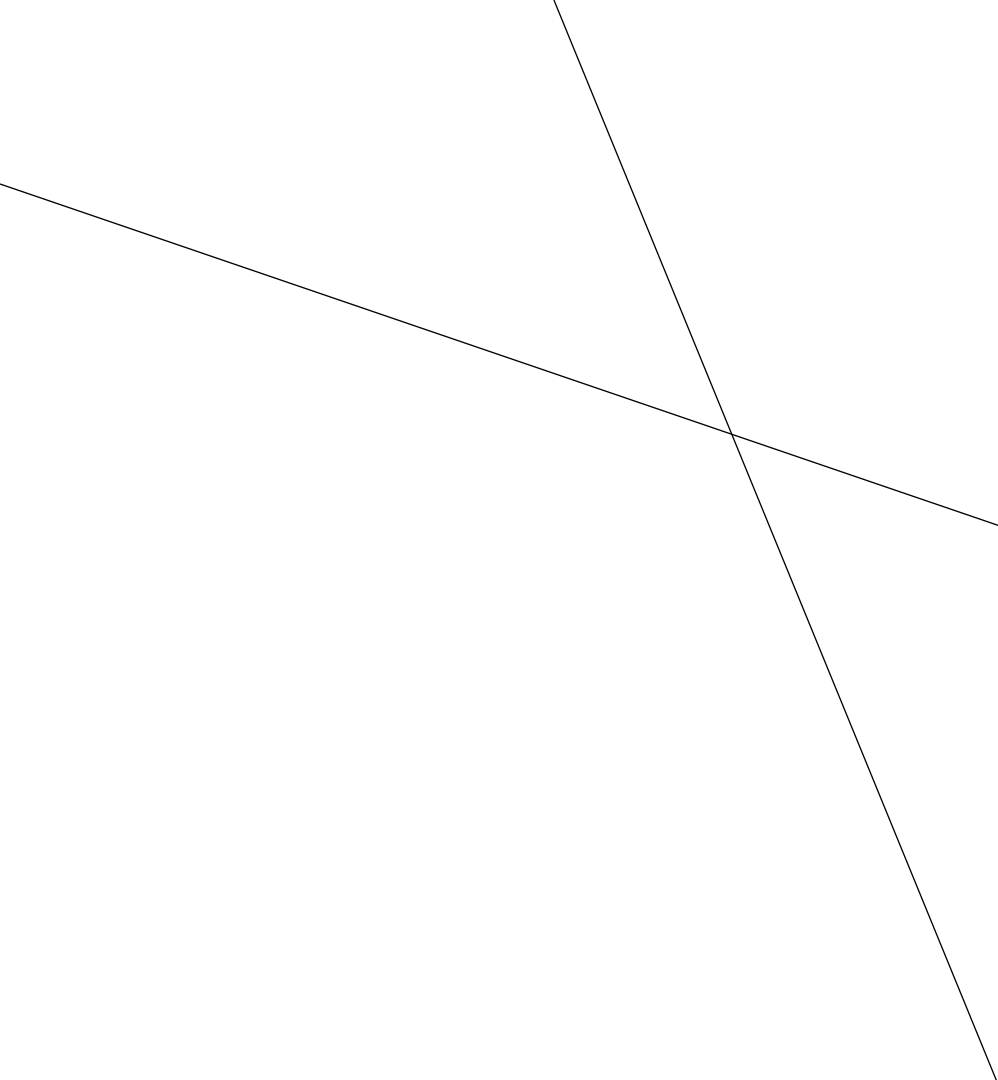
PROBLEM STATEMENT

Given a loan data set of a **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- 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



DATA CLEANING

- 
- We started with removing redundant or null value columns

```
loan_dataset.shape
```

- Before Cleaning

```
(39717, 111)
```

- After Cleaning, we have

```
loan_dataset.shape
```

```
(39717, 48)
```

Now we have 48 columns corresponding to loan

1. We will analyze the user details and factors that drive loan defaulters before approving loan. Then we can remove the columns
2. We will also remove unique value columns as well as, they don't contribute to the loan defaulters due to irrelevant information.
3. At the same time we will remove columns like "desc", "out_prncp_inv" , "total_pymnt_inv " as it has nothing to do with loan #### defaulting.

After data cleaning as above,

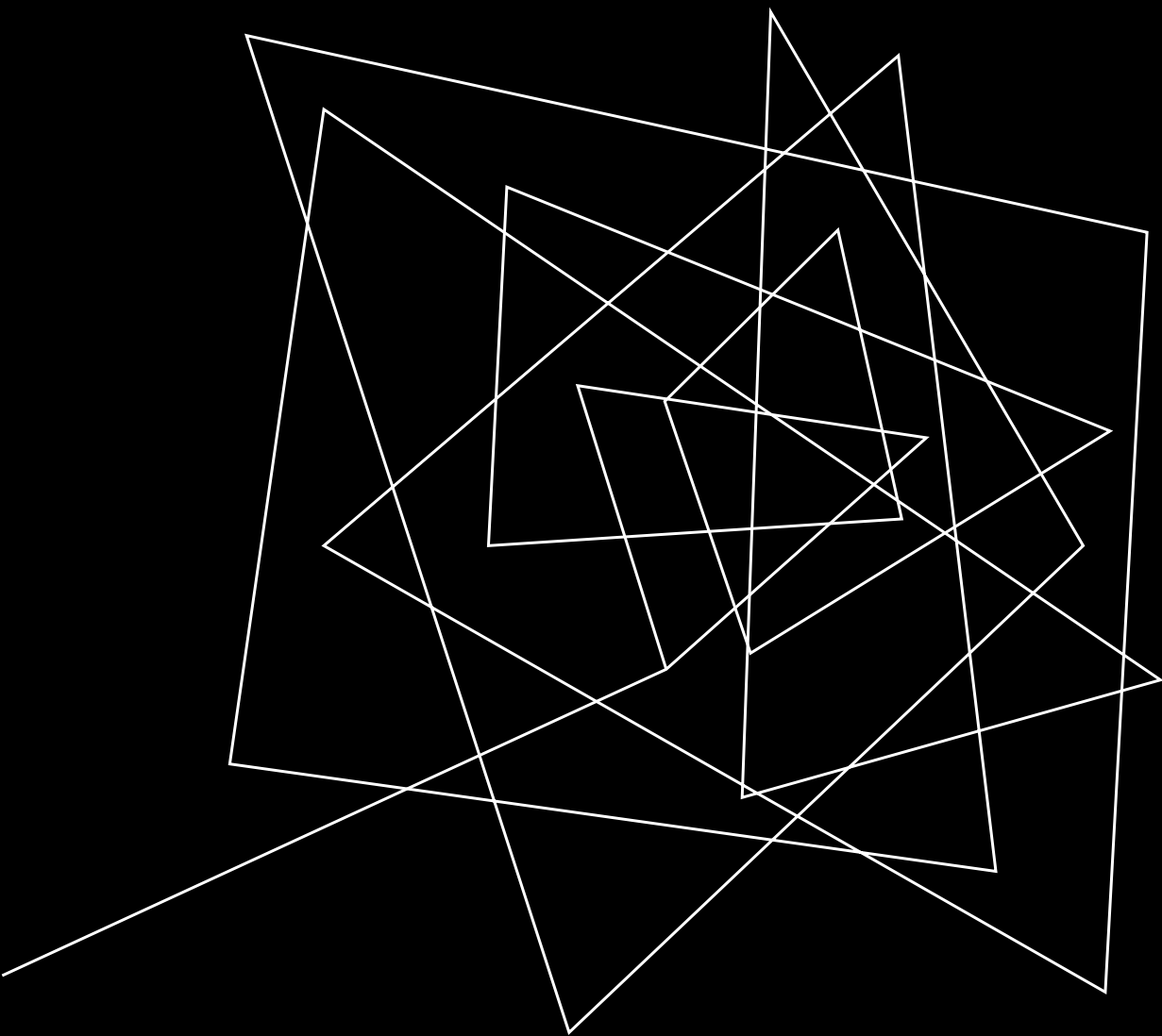
```
loan_dataset.shape
```

```
(39717, 21)
```

REPLACING MISSING VALUES

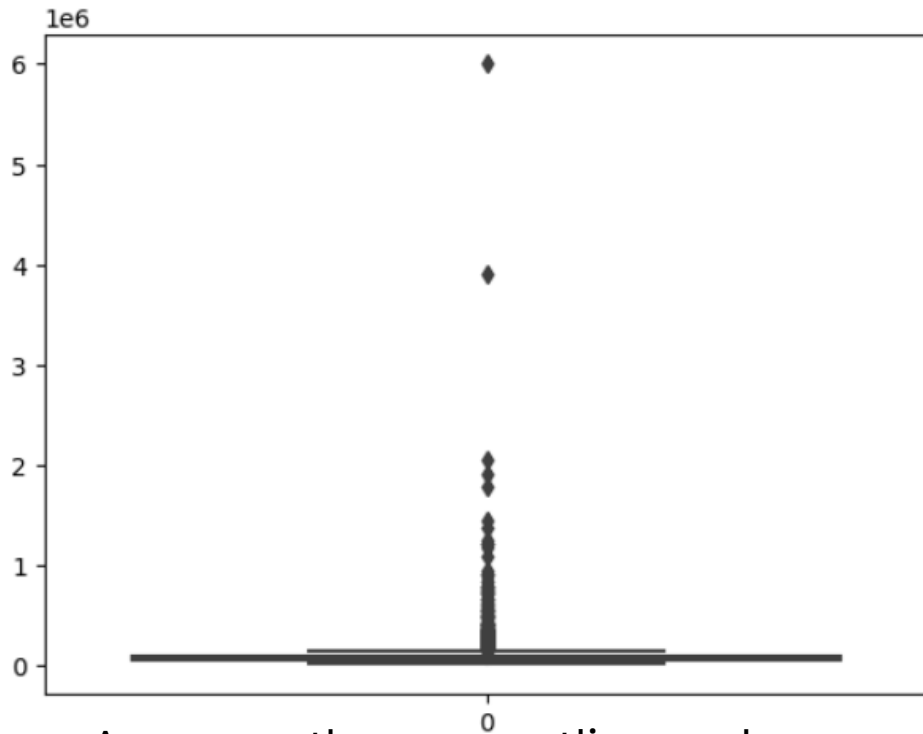
```
loan_amnt      0.000000
funded_amnt_inv 0.000000
term           0.000000
int_rate       0.000000
installment    0.000000
grade          0.000000
sub_grade      0.000000
emp_length     2.677761
home_ownership 0.000000
annual_inc     0.000000
verification_status 0.000000
issue_d        0.000000
loan_status    0.000000
purpose        0.000000
dti            0.000000
earliest_cr_line 0.000000
inq_last_6mths 0.000000
open_acc       0.000000
pub_rec        0.000000
revol_util     0.129611
total_acc      0.000000
dtype: float64
```

- Here we see "emp_length", "revol_util" have missing values and both are of type "object". So we will use mode and assign it for missing values
- We need to standardize the data for revol_util as it is object but has float values. We will remove rows with null data and then standardise revol_util
- Converting emp_length to numeric as it will be easy for our analysis. Setting < 1 year to 0 and 10+ years to 10

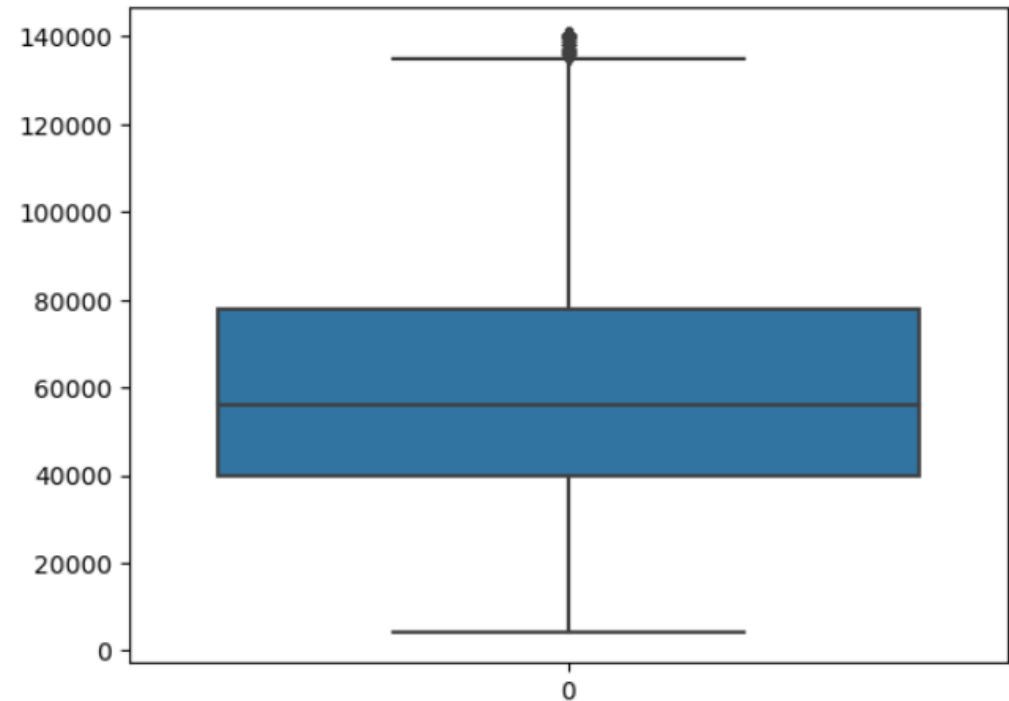


REMOVING
OUTLIERS

Annual income is of the major factors for loan approval and hence we need a proper and clean data for this. We will analyze the data and check for any outliers

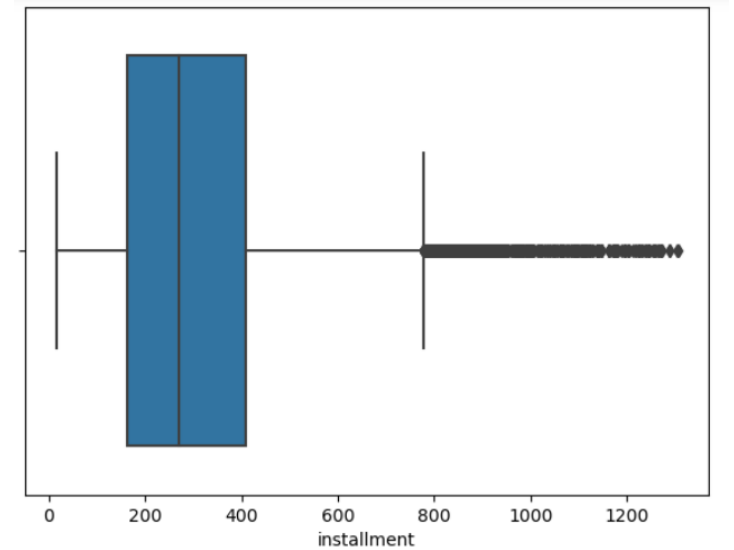
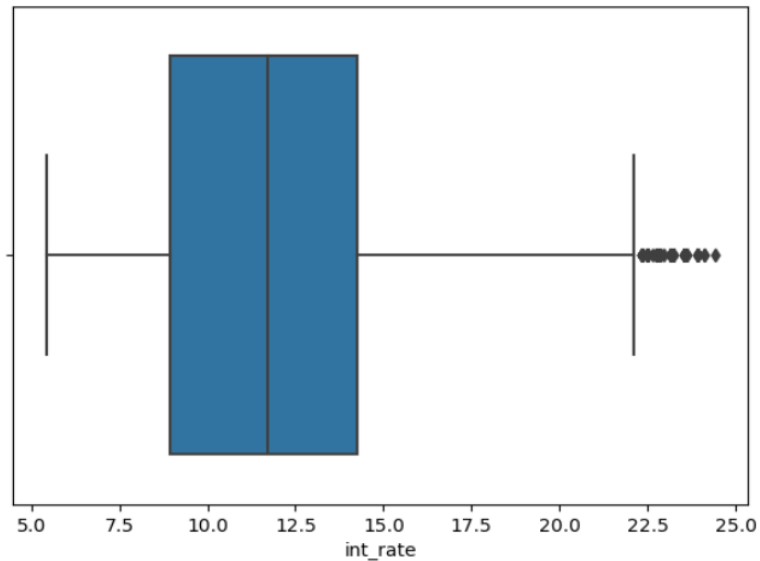
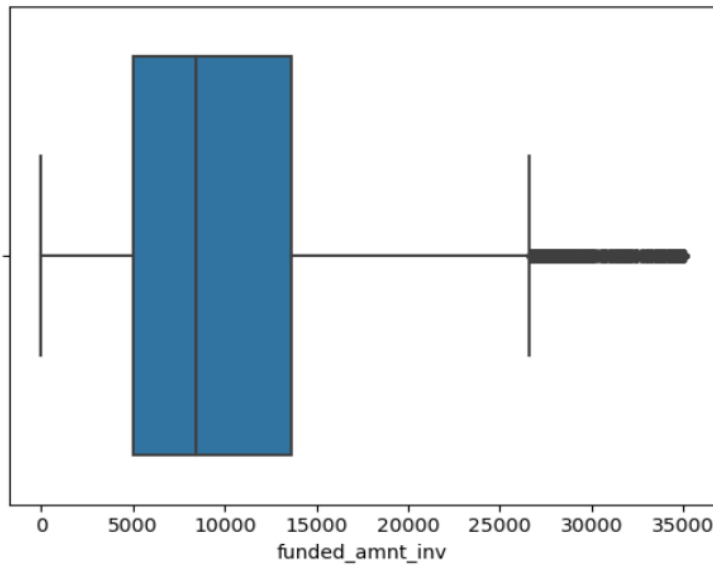
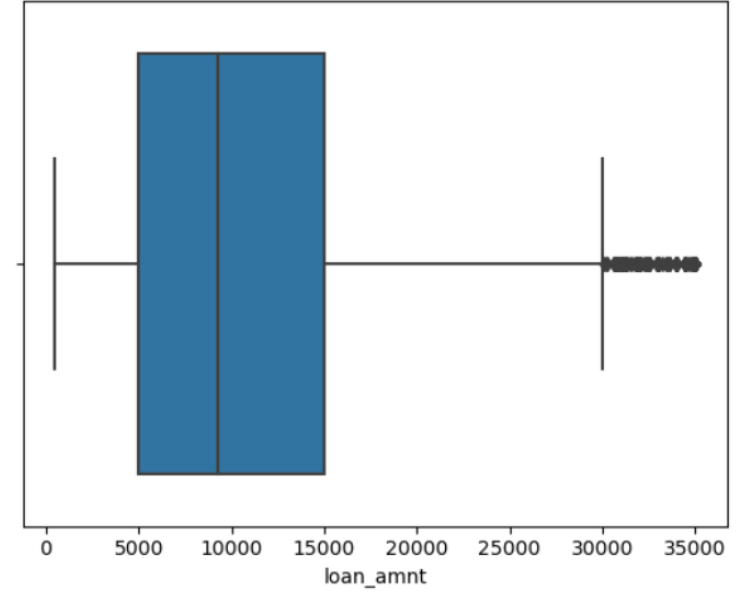
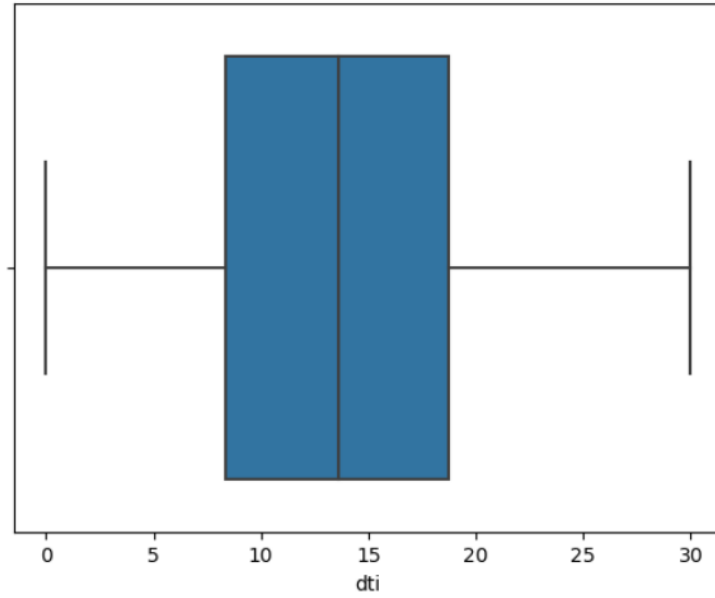
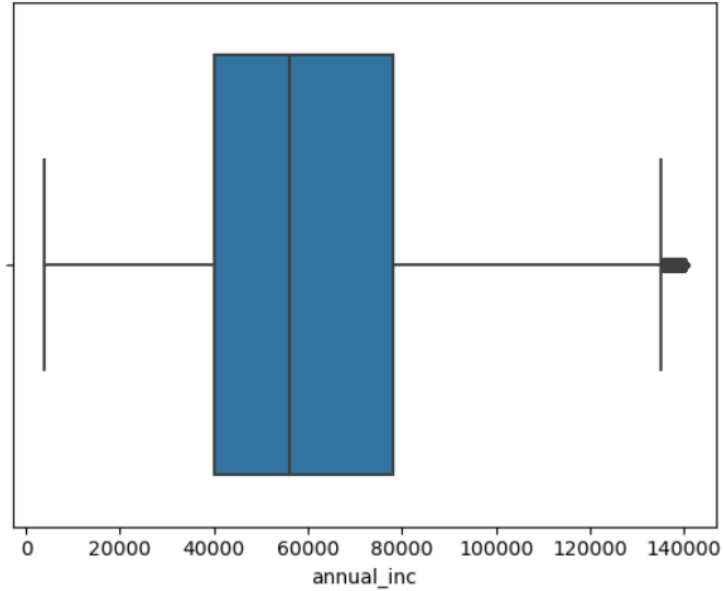


- As we see there are outliers and we need to remove them. After 95 percentile we see they are disconnected and we will remove outliers from 0.95 percentile



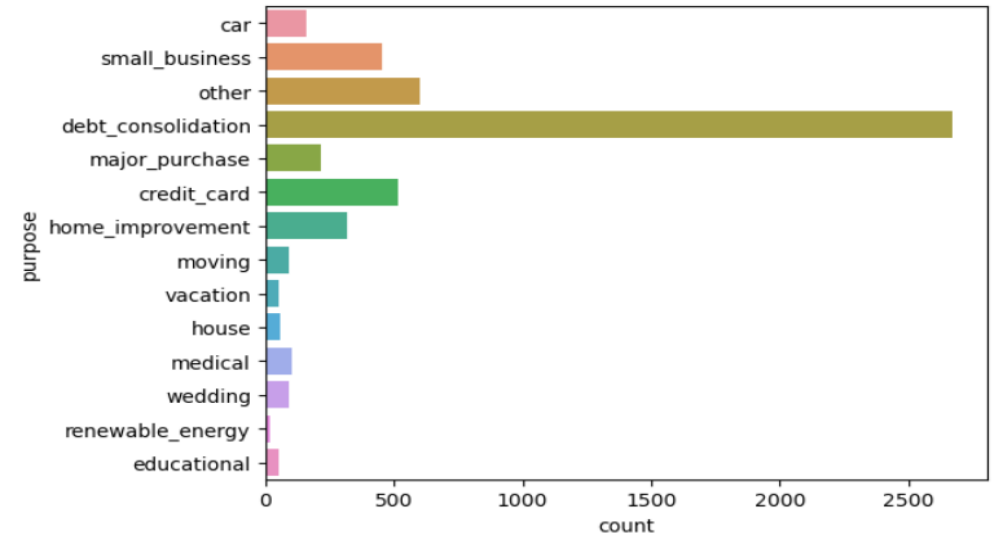
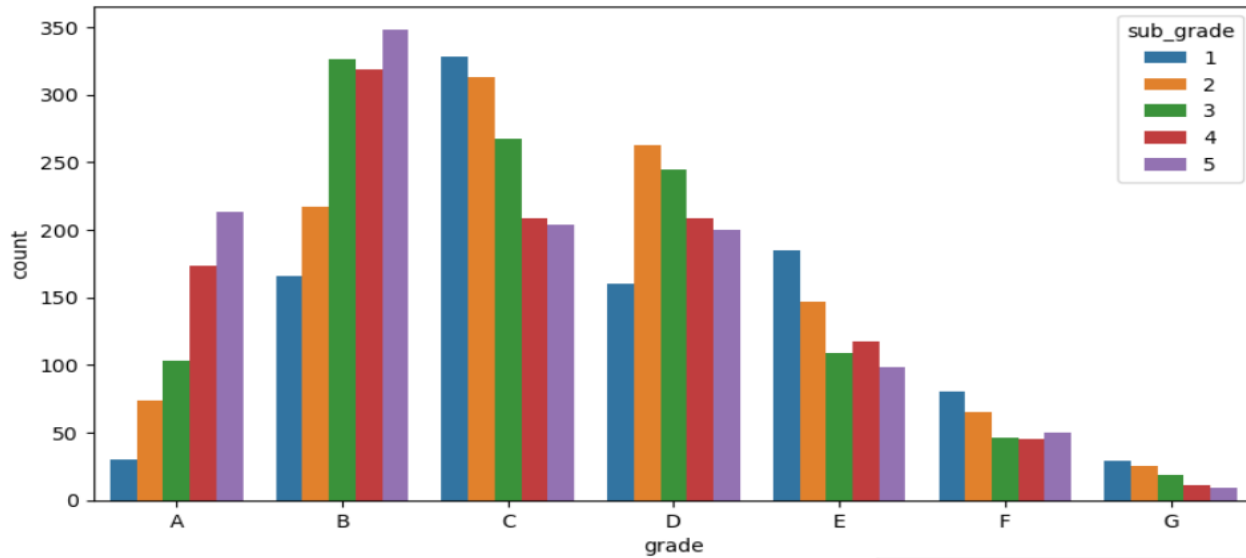
- Now we see there are no outliers for this columns.

Analysing Numeric Columns for outliers

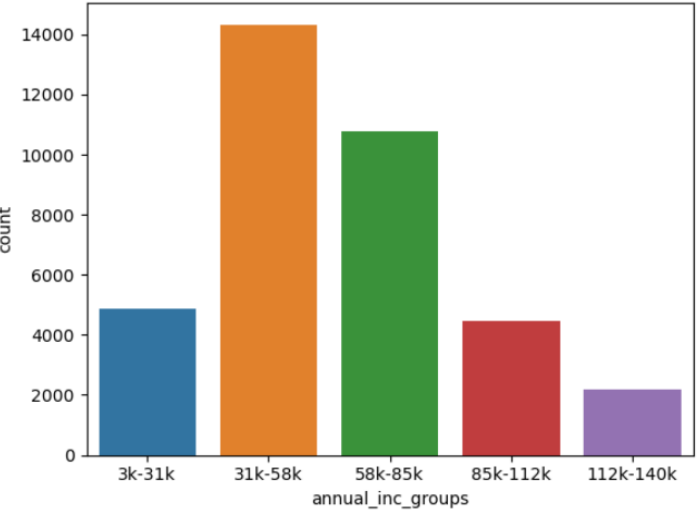
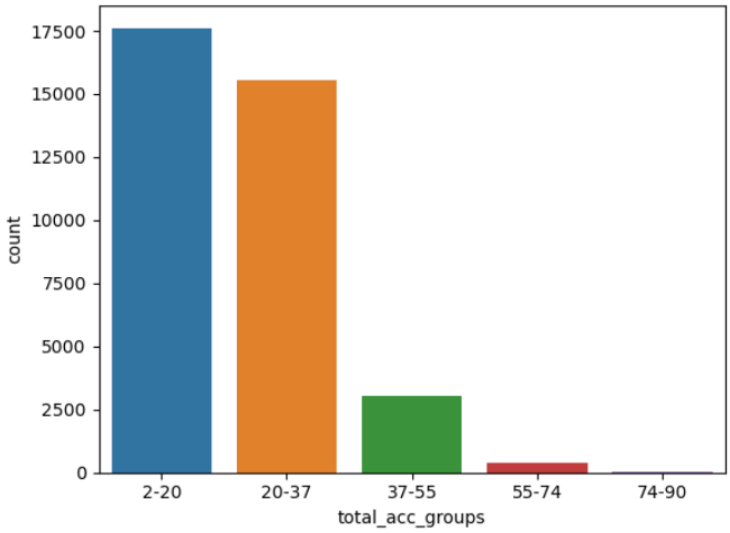
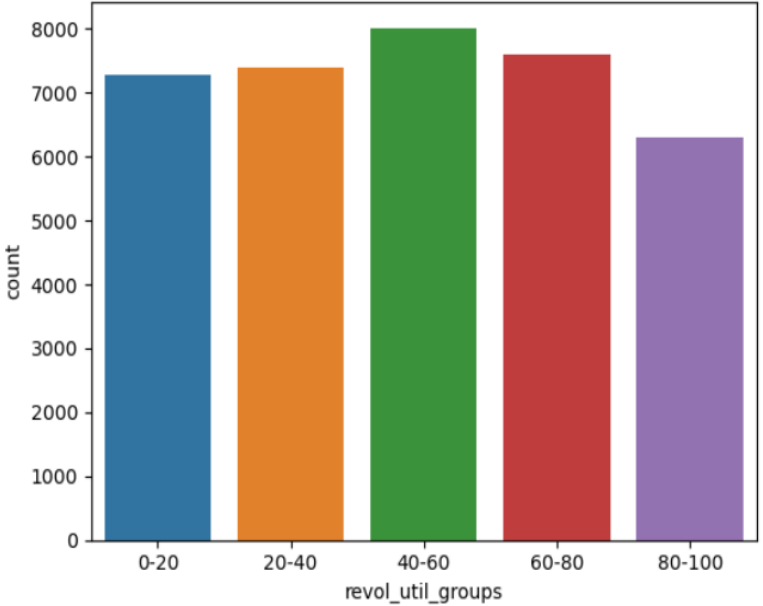
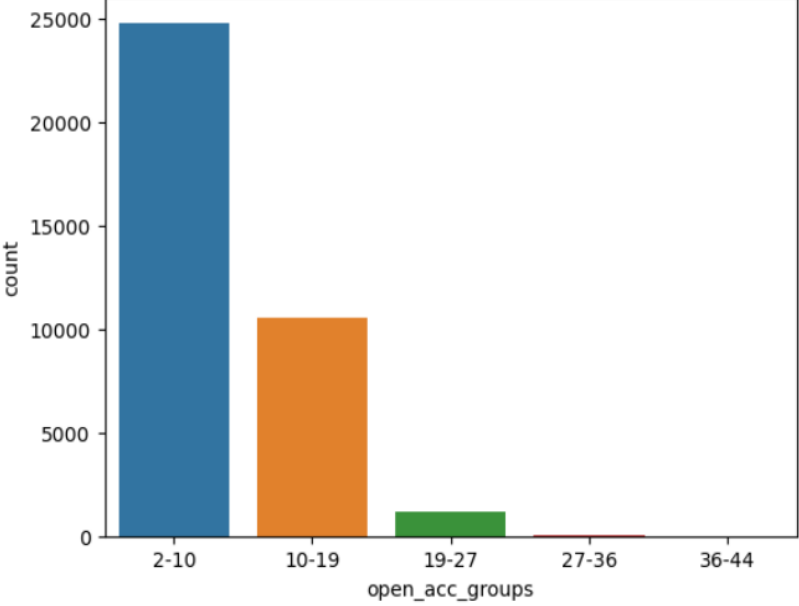
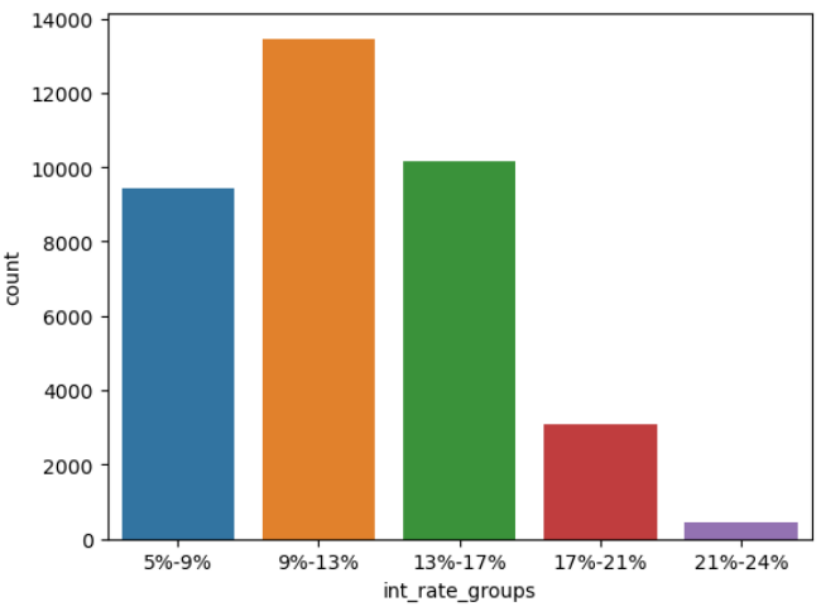


- We see outliers for loan_amount, funded_amnt_inv, int_rate and installment. But still the data is pretty continuous. Lets not remove it and keep as is

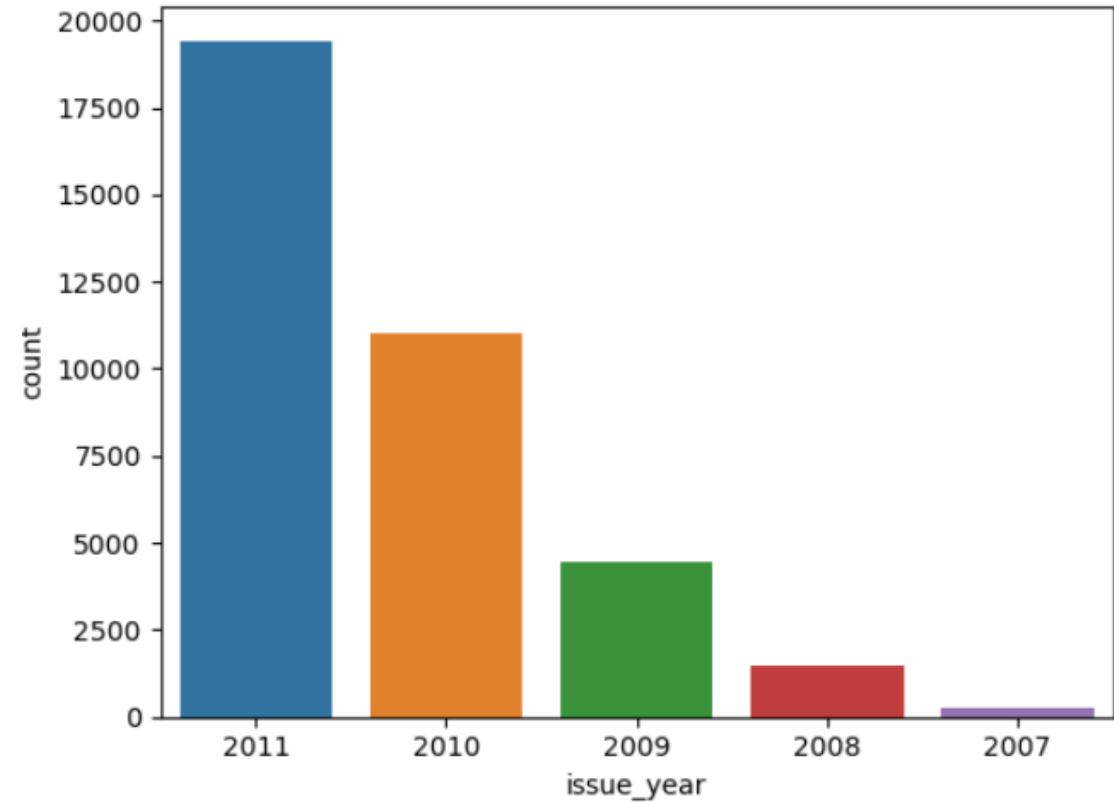
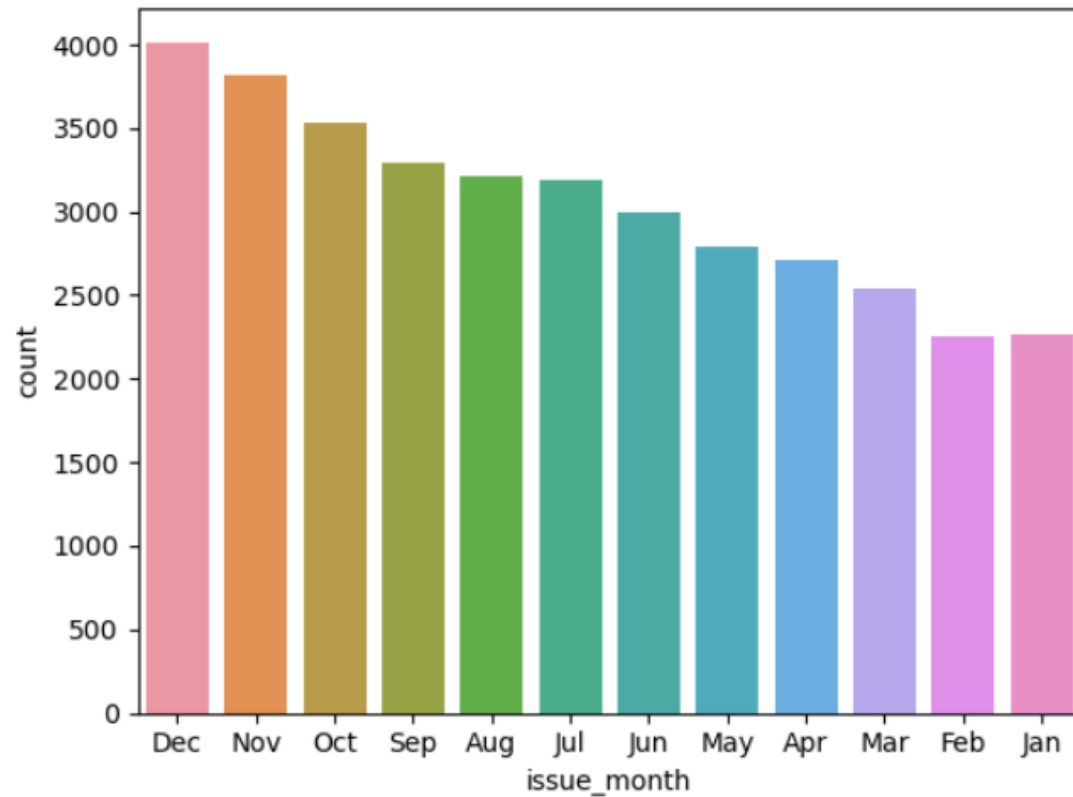
Analysing Categorical Columns for outliers



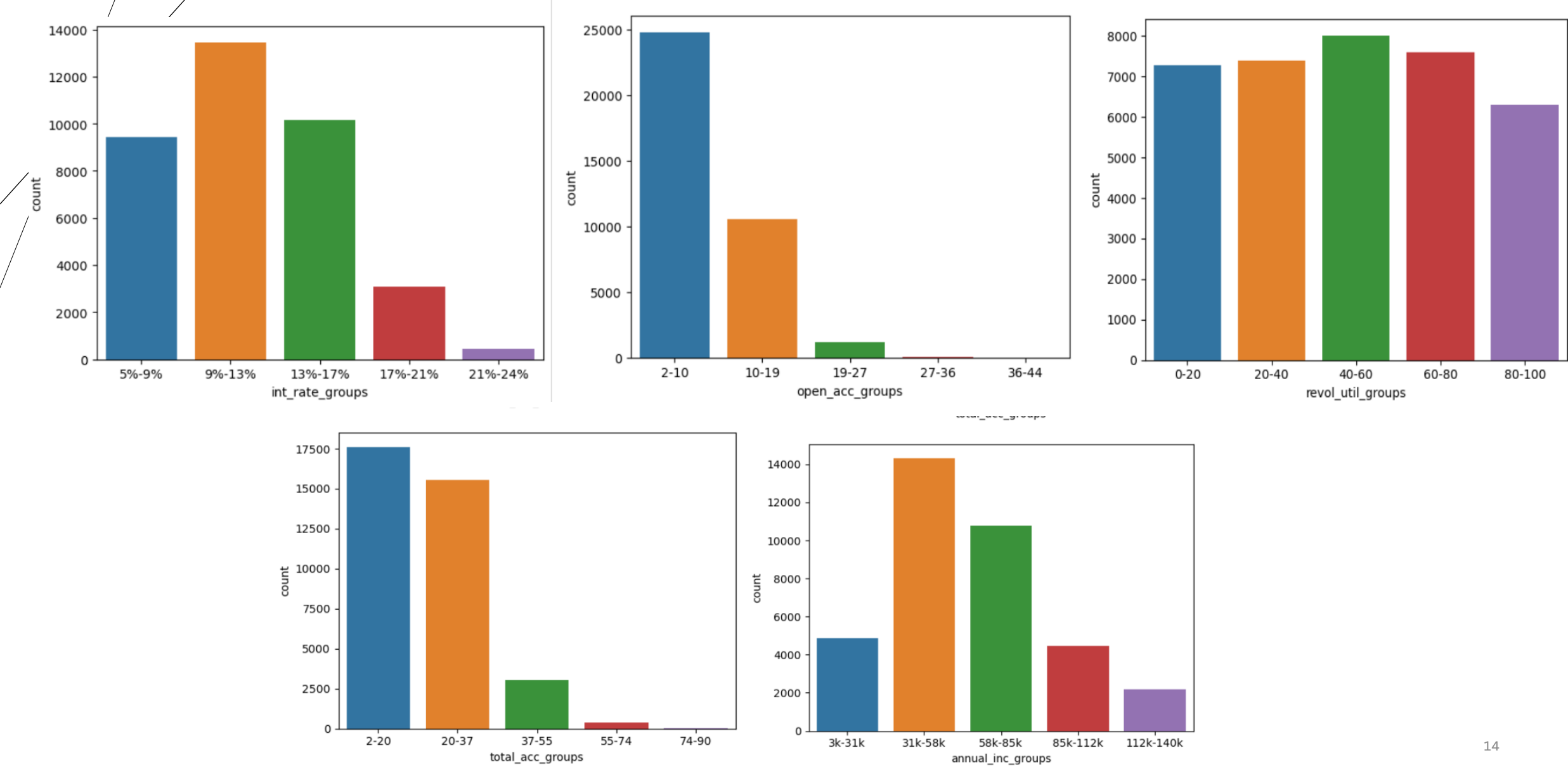
GROUPING CATEGORICAL DATA FOR COLUMNS



ANALYZING BY LOAN ISSUED MONTH AND YEAR



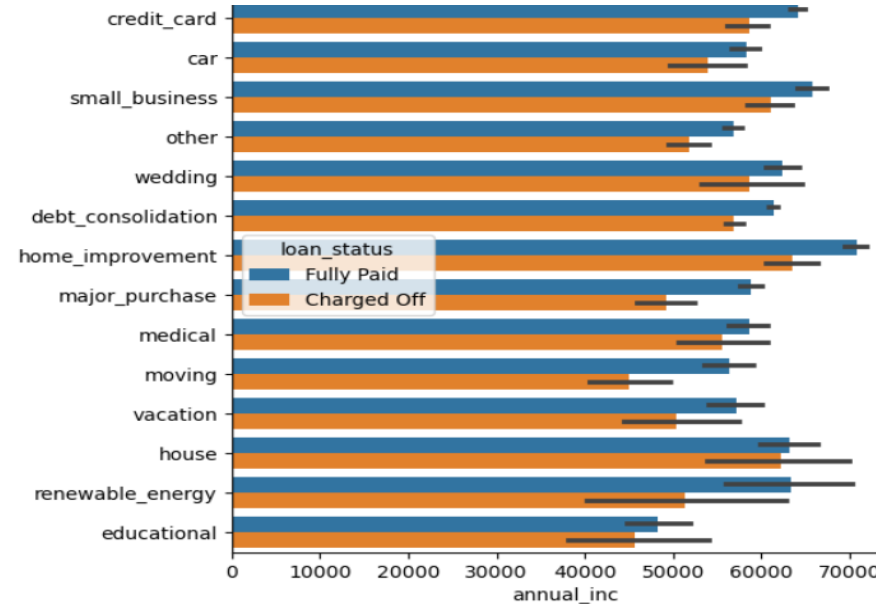
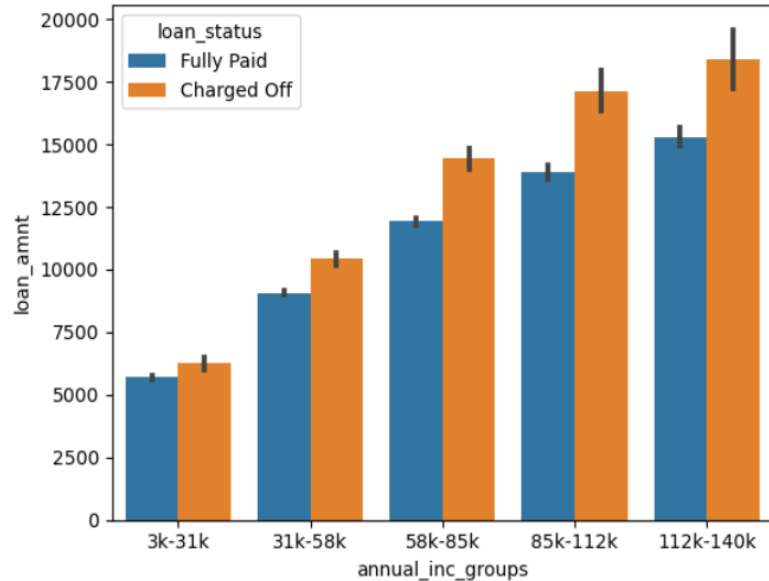
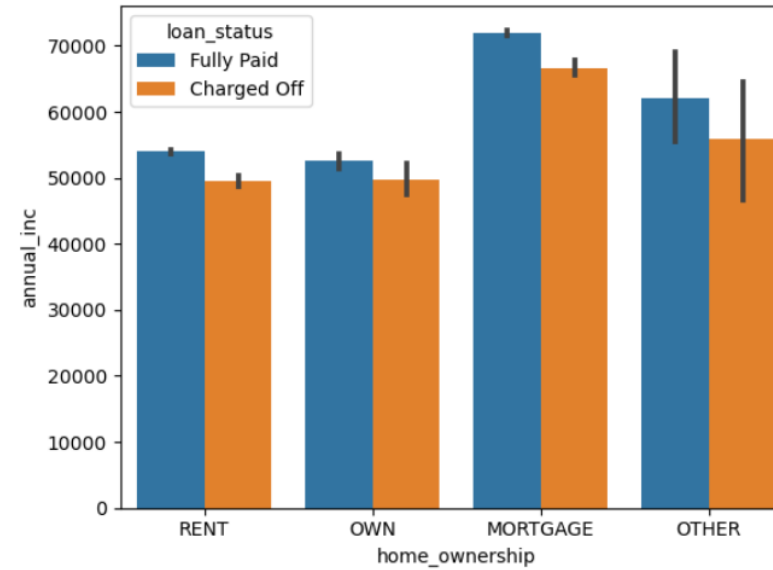
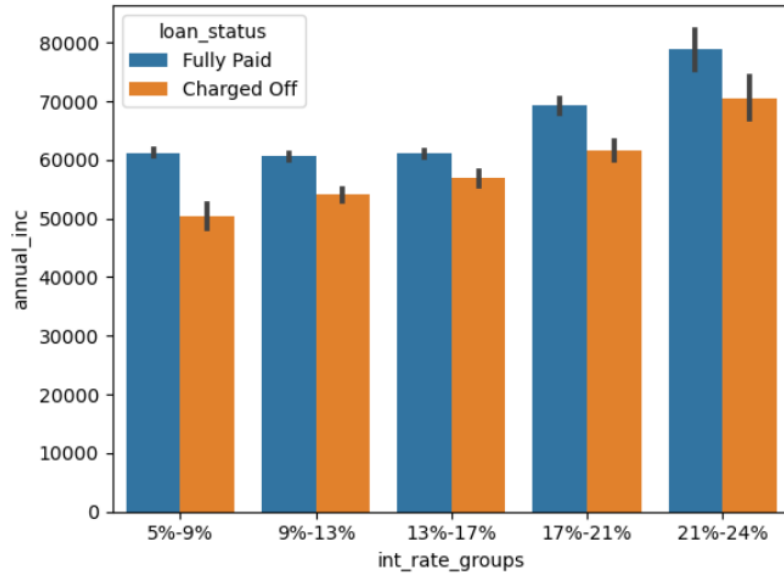
GROUPING CATEGORICAL DATA FOR COLUMNS



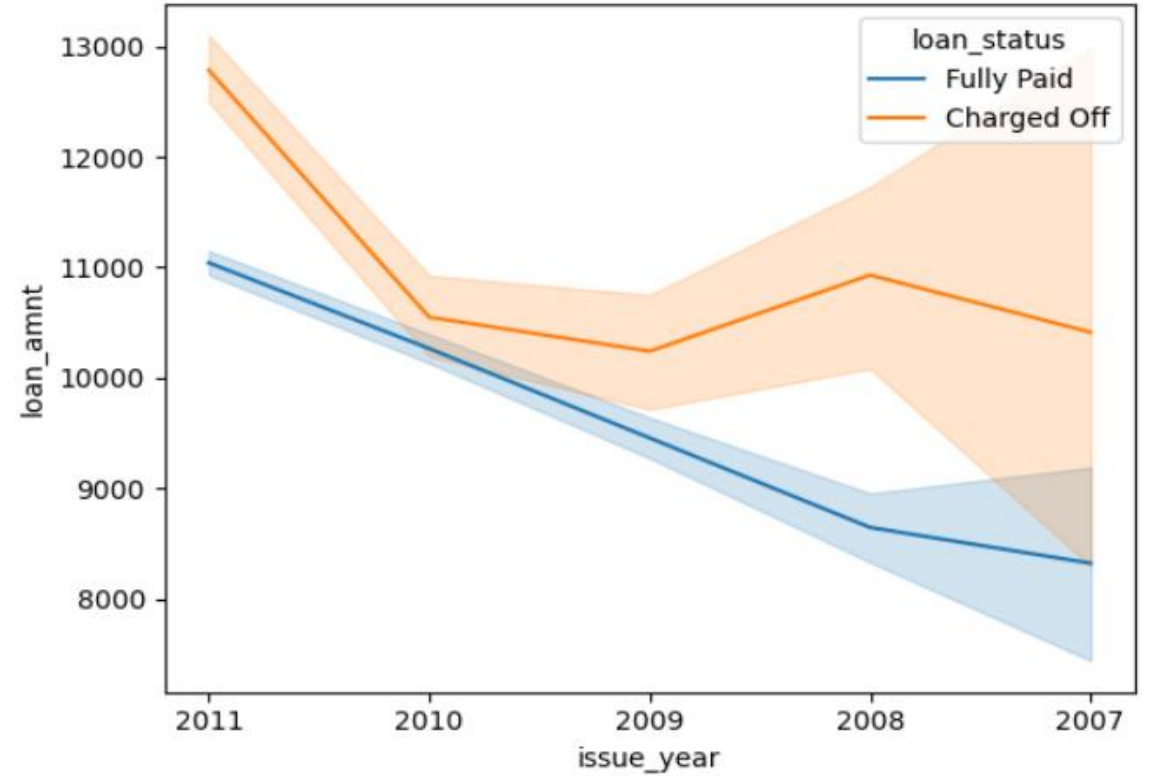
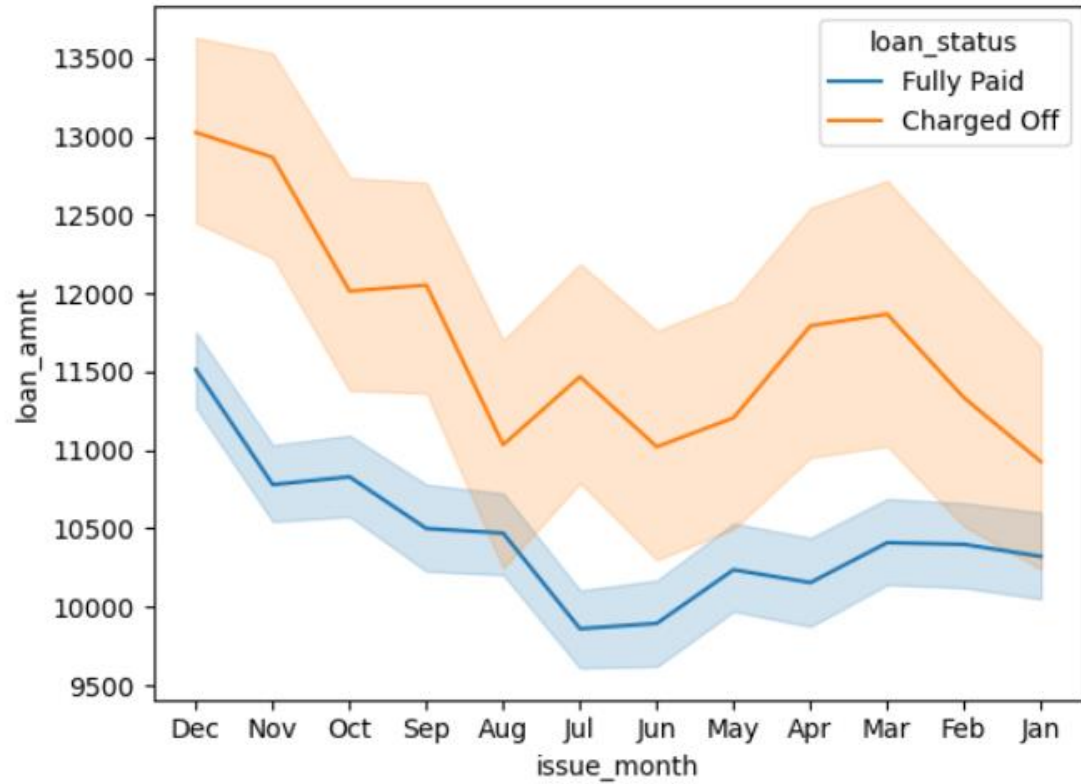
HERE WE CAN ANALYZE THE DATA AND WE CAN OBSERVE THAT

- Applicants who receive interest at the rate of 13-17%
- Applicants having house_ownership as 'RENT'
- Applicants who have 20-37 open_acc
- Applicants with employment length of 10
- Applicants who use the loan to clear other debts
- Applicants who have an income of range 31201 - 58402
- When funded amount by investor is between 5000-10000
- Loan amount is between 5429 - 10357
- When the number of derogatory public records is 0
- When monthly installments are between 145-274
- Term of 36 months
- When the loan status is Not verified
- Grade is 'B'
- When the no of enquiries in last 6 months is 0
- When the purpose is 'debt_consolidation'
- Dti is between 12-18
- total grade of 'B5' level.

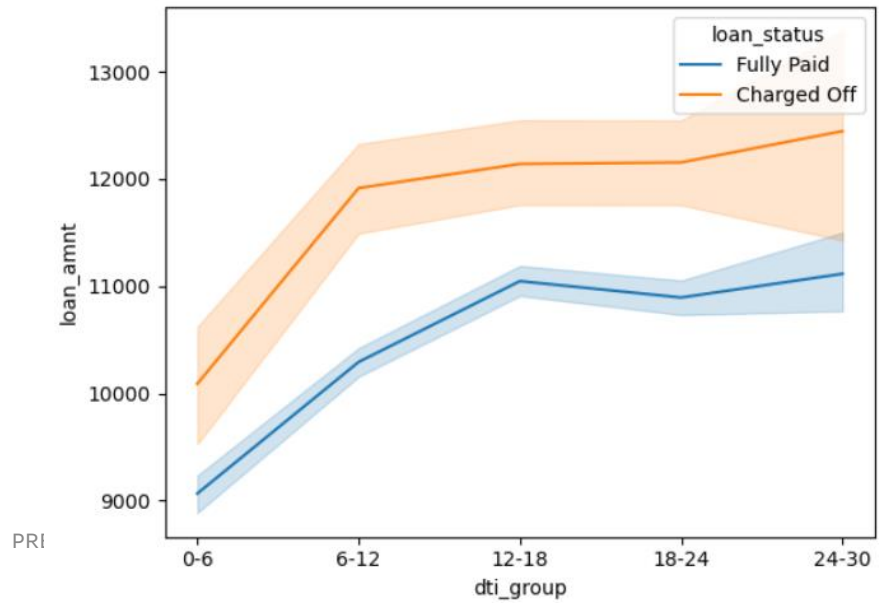
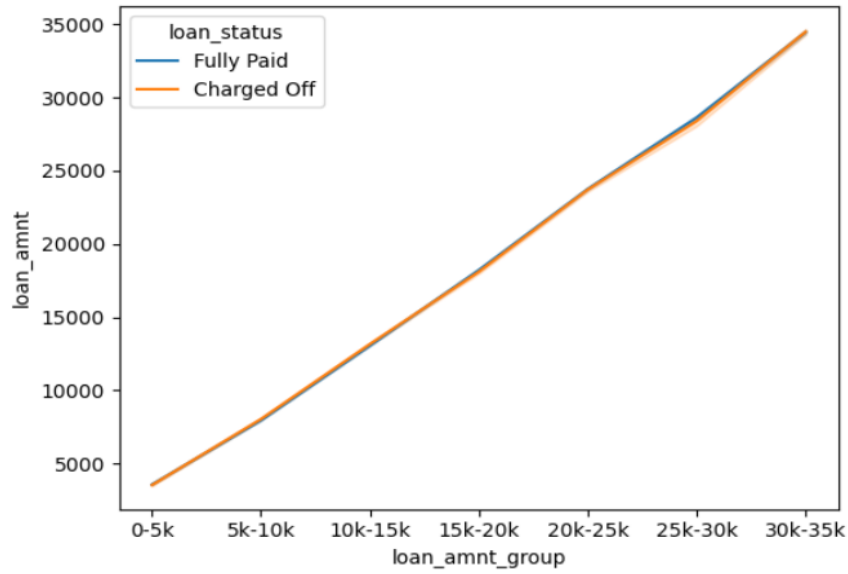
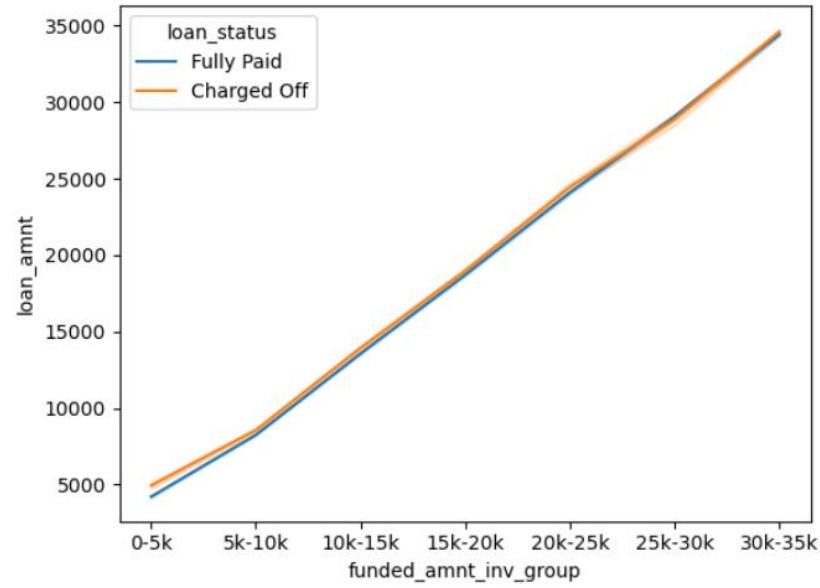
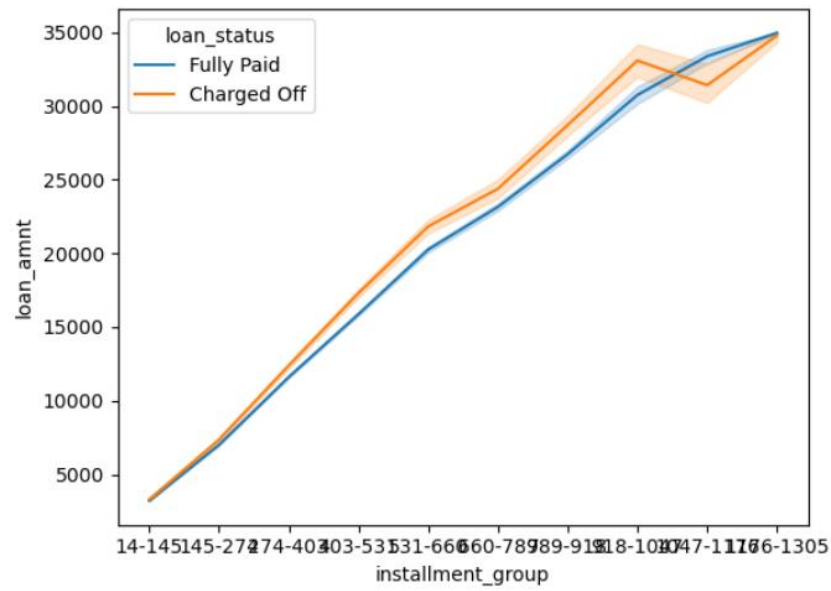
ANALYSING ANNUAL INCOME



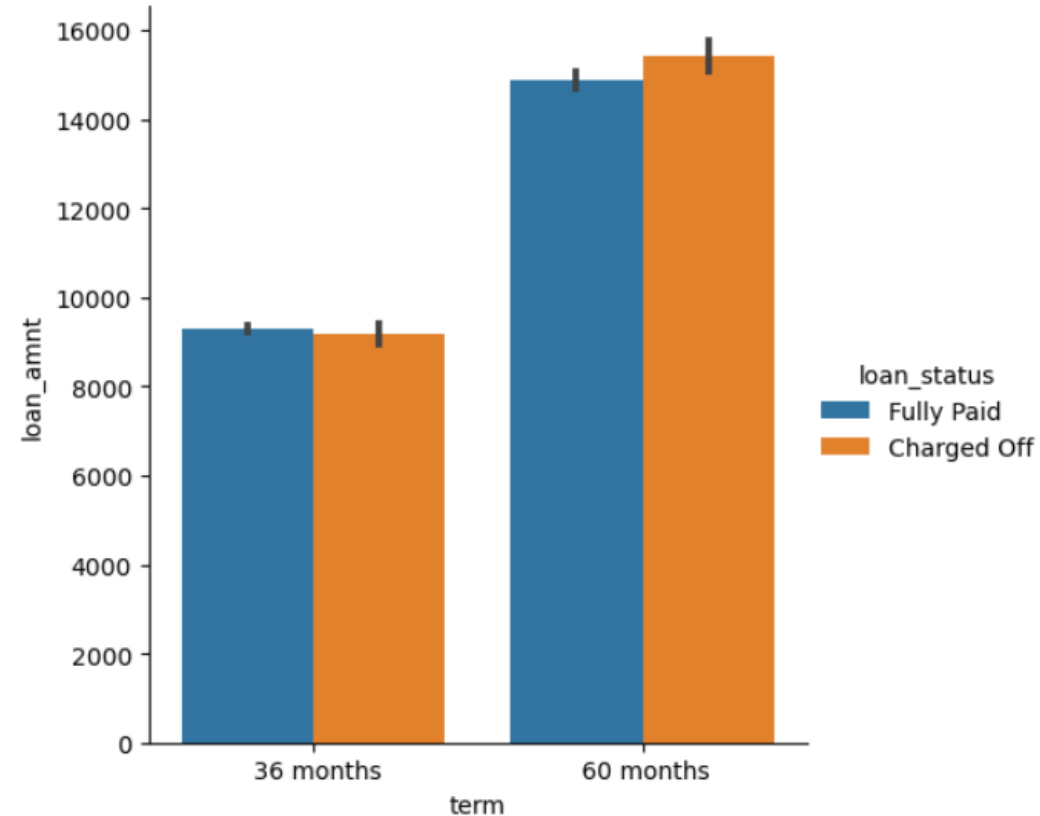
ANALYSING LOAN AMOUNT



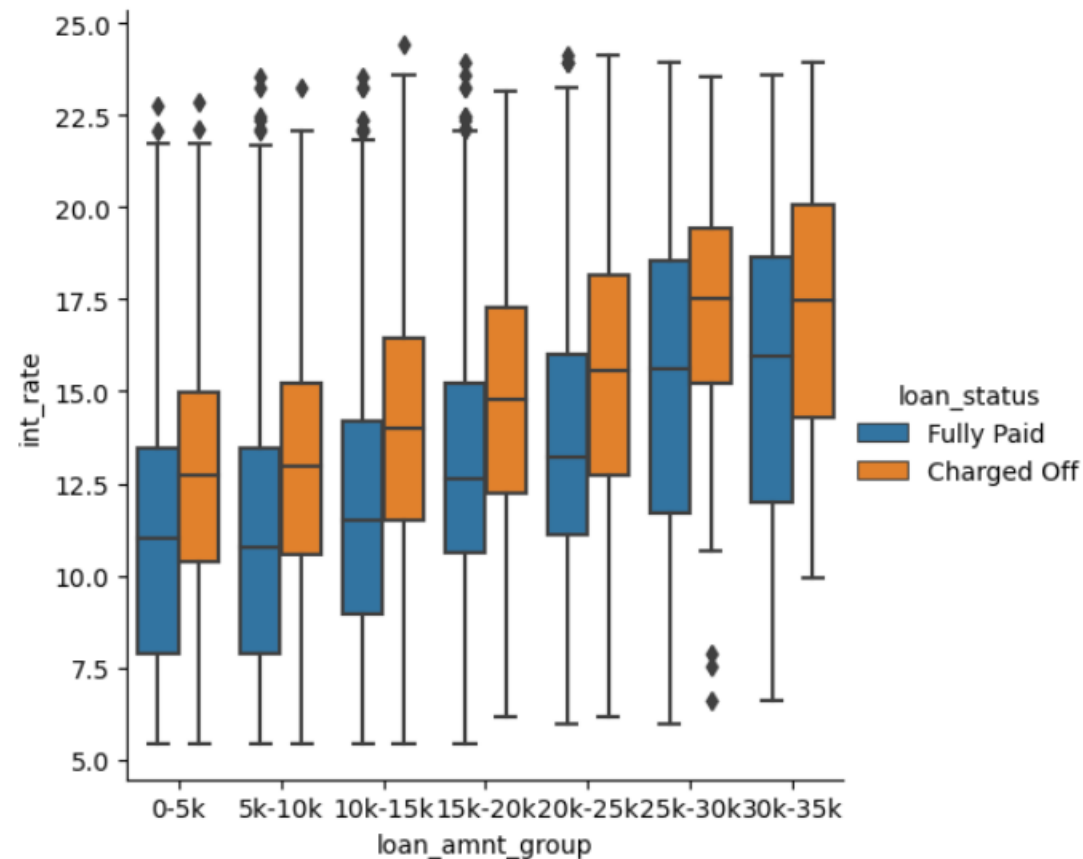
LOAN AMOUNT VS LOAN AMOUNT GROUPING



Here we can observe applicants who applied and defaulted have no significant difference in loan_amounts. Loan amount is not a factor to be considered for loan defaults



From Graph below we can see that for loan amount group the interest rate for charged off loans are higher than that of fully paid loans



OBSERVATION

- When the loan is verified and loan amount is above 16k and loan is not verified for loan amount below 8.5k
- Applicants taking loan for 'home improvement' and have income of 60k -70k and home ownership is 'MORTGAGE
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k and who receive interest at the rate of 5-9% and have an income of 50k-60k
- For grade G and interest rate above 20%
- When employment length is 10yrs and loan amount is 12k-14k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When home ownership is Rent, more than 2500 applicants are charged off
- Applicants with debt consolidation are loan defaulters
- 12k - 14k applicants have received interest rate of 9-13%
- Maximum number of defaulters occurred when the loan was sanctioned/issued in the month of december and issued in year 2011

A series of white, thin, overlapping geometric lines and polygons on a black background, located on the left side of the slide. The lines form various shapes, including triangles and quadrilaterals, some of which are nested or intersecting.

THANK YOU