UPGRAD EDA CASE STUDY

LENDING CLUB

Navita Goel
Priyanka Kumari

CONTENTS

- Business Problem
- EDA Insights & Recommendations
- Data Cleansing Approach
- Data analysis approach
- Visuals along with insights for each analysis points

BUSINESS PROBLEM

- For a consumer finance 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.

- To reduce the risks associated with the approval process, we are to identify patterns to indicate if a person is likely to default which may then be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

EDA INSIGHTS & RECOMMENDATIONS - I

1. **~14.2% of loans** have defaulted over last 4 years.
2. The total number of loans issued over the years suggest that the business is consistently growing over the years.
3. Most of the loans issued are in **5k-10K** range.
4. **>50%** of loans are in grade **A & B**, which have charge-off rates of **below 14.2**.
5. **>70%** of loans are of short duration (3 years).
6. The interest rates increase with duration of loans. Similarly, the interest rate increase as the loan grade lowers. This is to factor in riskiness of longer duration and low graded loans.

EDA INSIGHTS & RECOMMENDATIONS - II

7. Most of the customers have salary in 40 K to 80 K range.
8. More than 75% of customers have either rented homes or have mortgages
9. About 40% of customers come from California, New York and Florida.
10. Oddly enough, verified income loan applicants are doing poorly than unverified. **There is a need to investigate further to see if it is a data issue or evaluation issue.**
11. Grade C&D loans for customers having mortgaged or rented houses have relatively higher default rate. **More investigation is needed to identify root cause.**
12. No direct relation of loan amount issued to customer income or DTI. **Further investigation is needed to establish how customer's ability to pay is factored in loan approval process.**

DATA CLEANSING APPROACH

Data imputation

- Removed 74/111 attributes which were either null or had only one unique value.
- 2200 records were deleted to discard ongoing (Current status) loans and remove null values.
- Overall 68% of data fields were removed.

Data correction

- Changed data types of numerical fields(interest_rate, term etc.) and date fields(issue date, last payment date etc) to enable statistical operations during analysis.
- renamed column names(int_rate) for better readability

Outlier removal

- For few fields (loan amount, annual income), outliers were removed with threshold value of 1.5

Derived variables

- Introduced numeric column issue_year to represent year of loan issuance.

ANALYSIS APPROACH

Seaborn histograms, pie charts, bar plots, distribution, scatter plot, regplot, heatmap & box plots etc

Segmented *Univariate analysis*

For 12 (customer and loans) attributes we explored

1. distribution of individual attributes across paid and charged off loan segments.
2. The percentage of charged off loans for each values of that attribute.

Bivariate Analysis

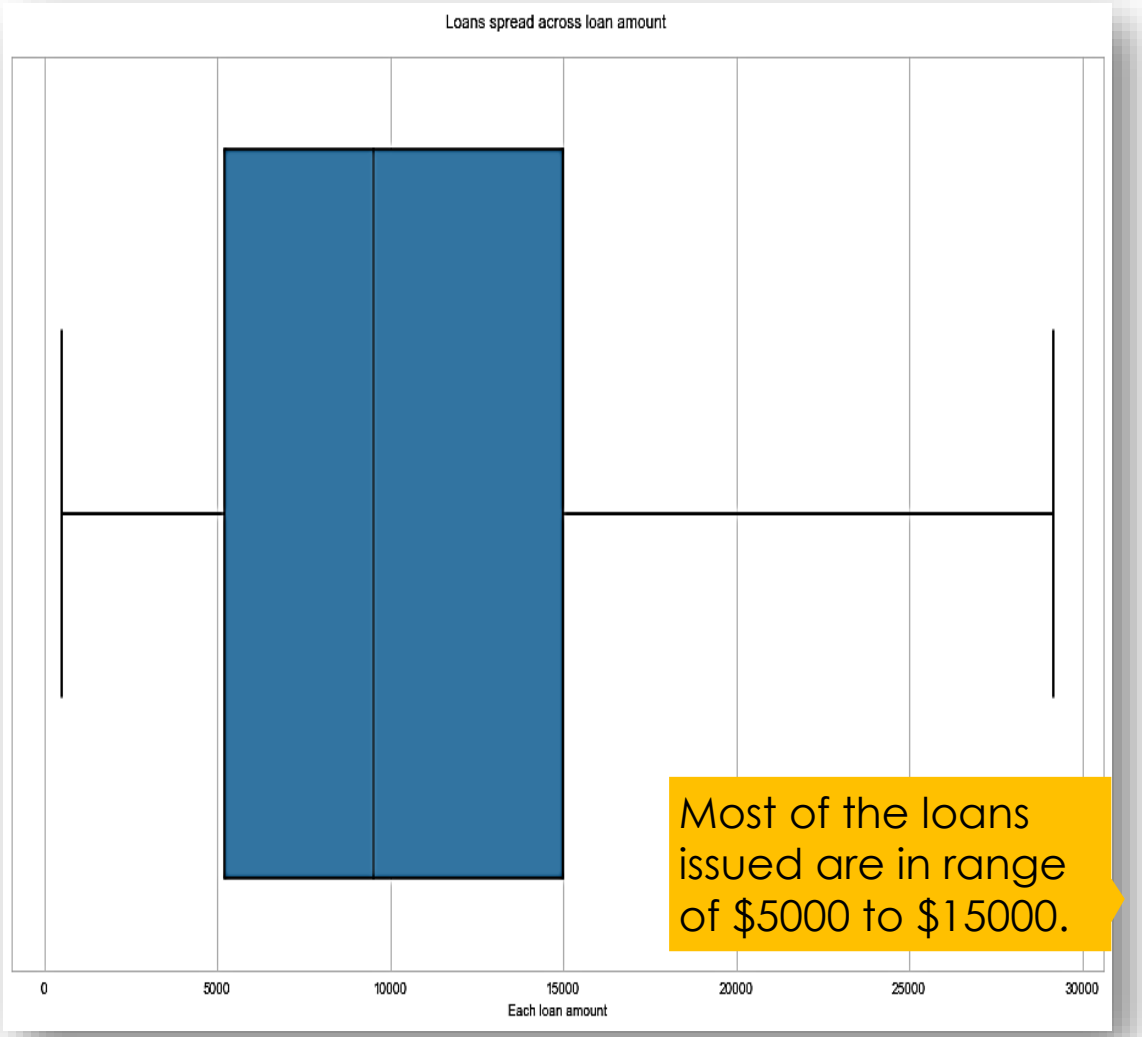
To explore 8 relationships between two attributes,

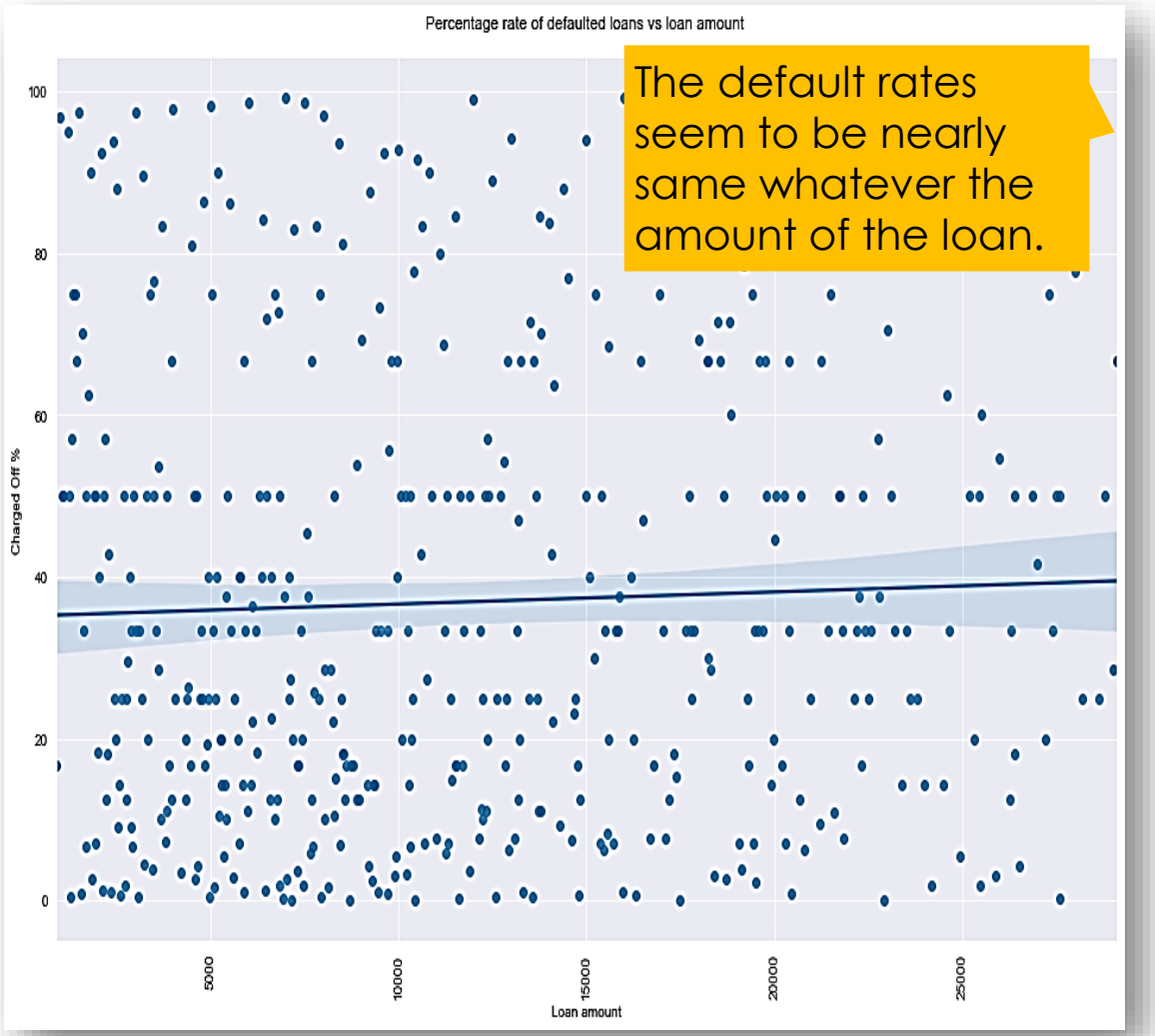
- Categorical vs. categorical(e.g. loan grade vs loan term)
- Numerical vs. numerical(e.g. loan amount vs annual income)
- Numerical vs categorical (e.g. housing status vs interest rates

Multivariate Analysis

- *Examined* the interaction between term, housing status & loan grade

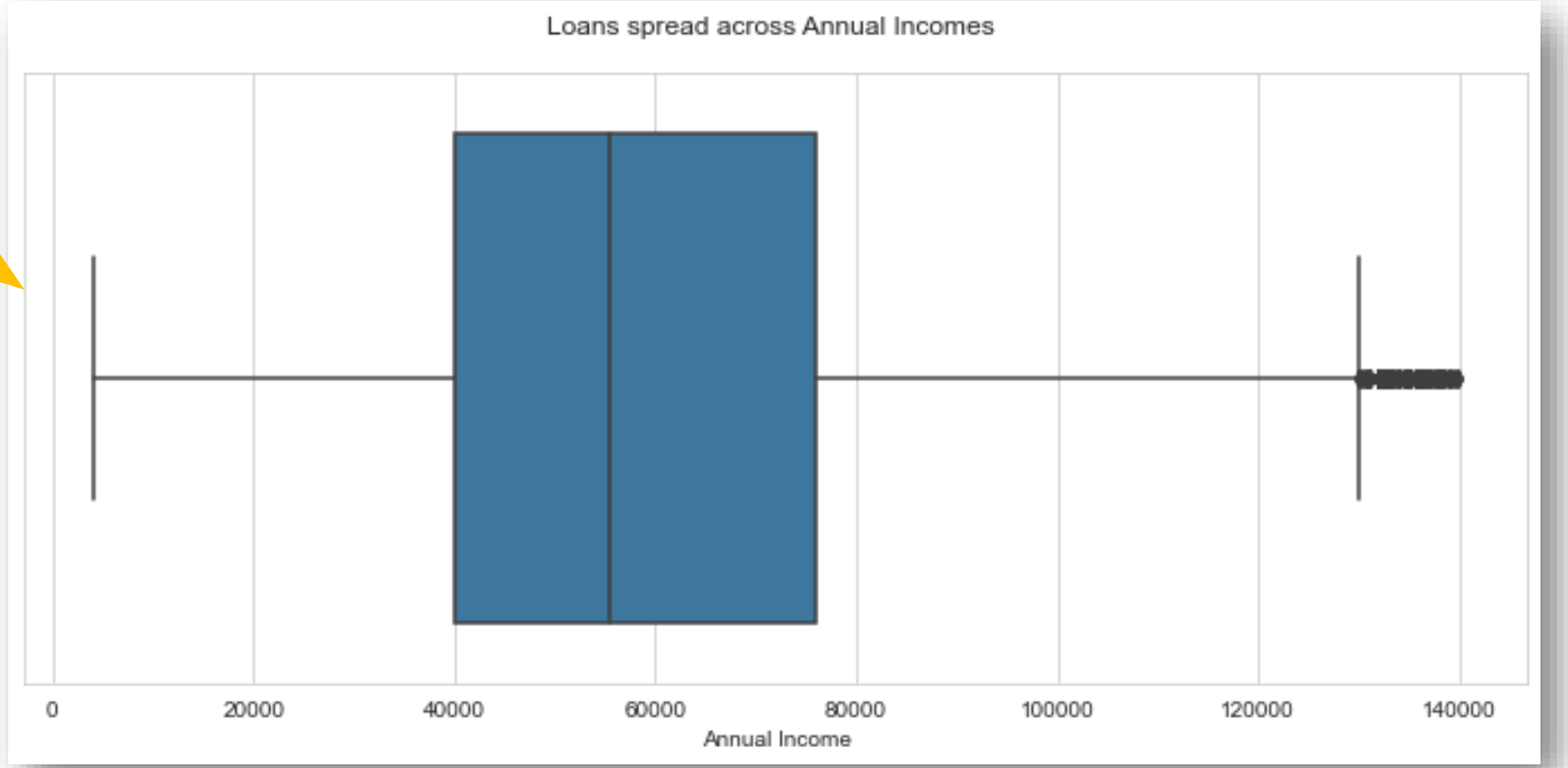
LOAN AMNT DISTRIBUTION & PERCENT DEFAULT RATES ACROSS LOAN AMOUNTS



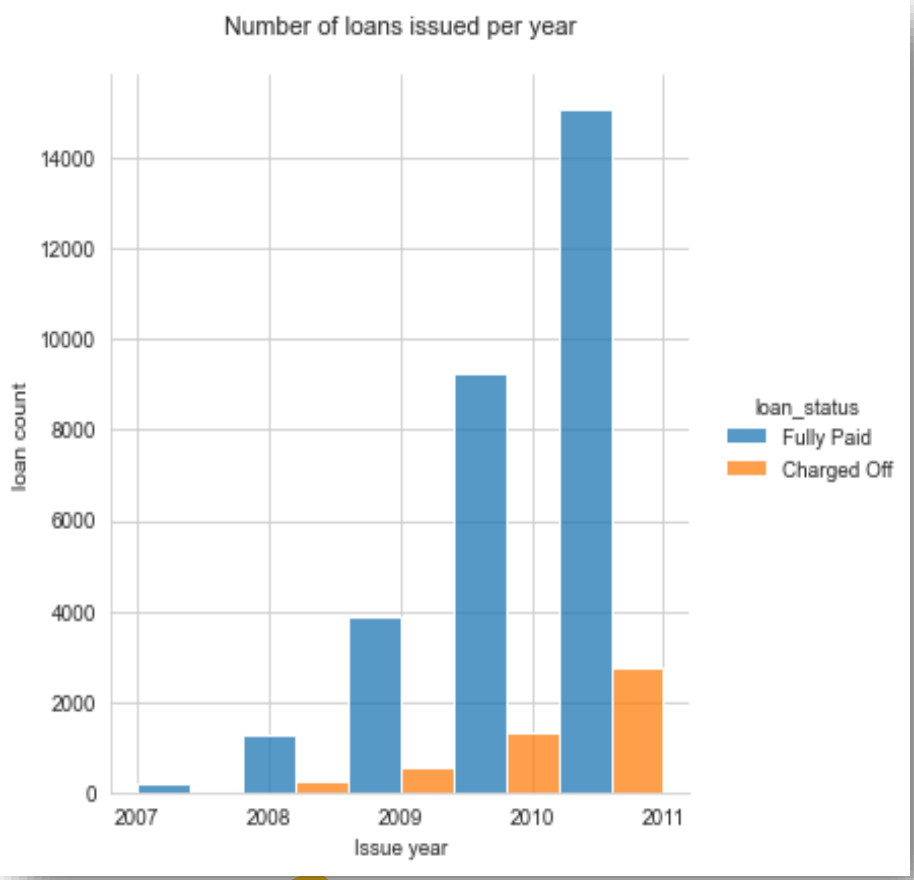


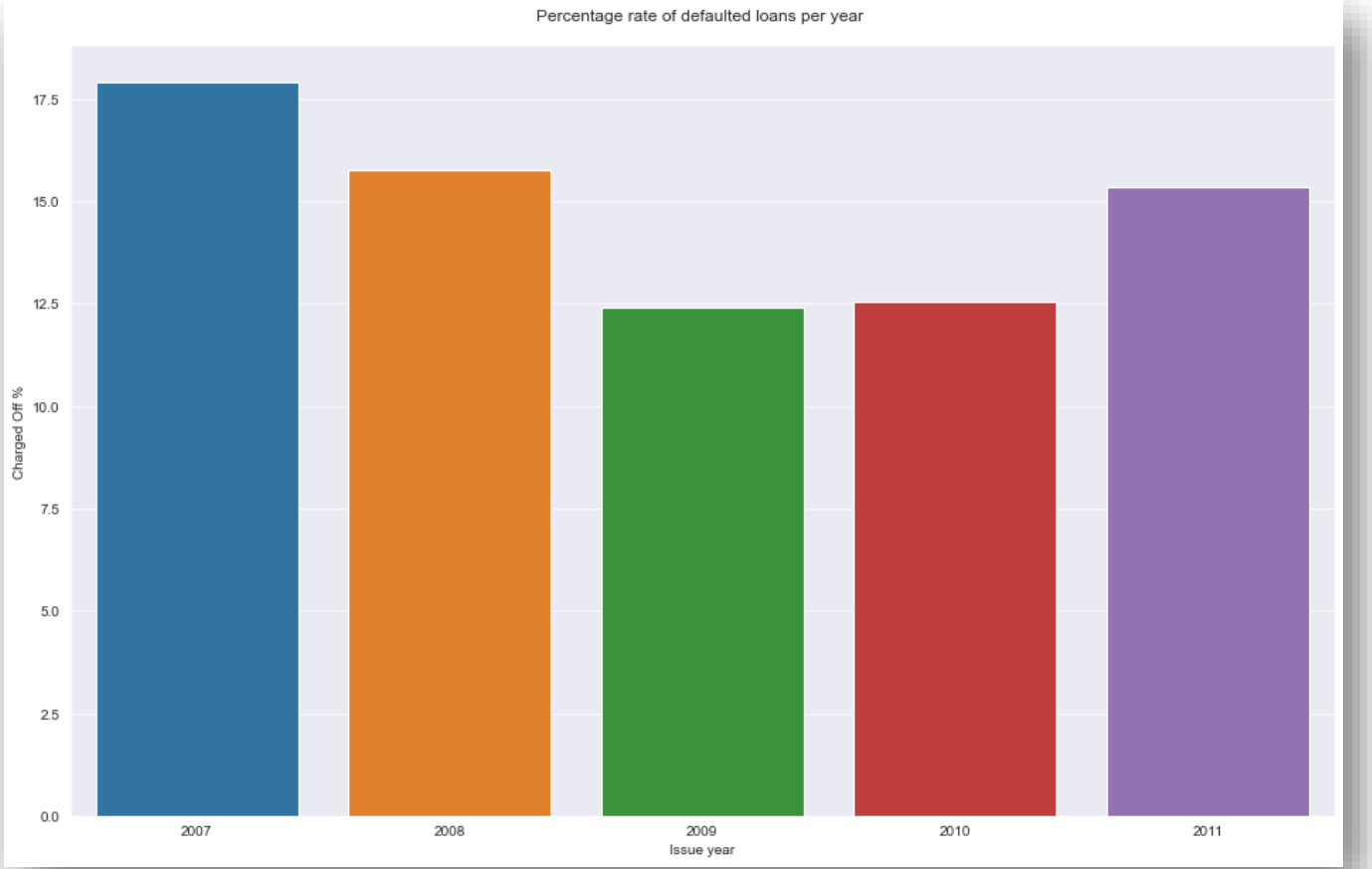
DISTRIBUTION OF LOANS ACROSS VARIOUS ANNUAL INCOMES

Most of the consumers have their salary in 40K to 80K range.



DISTRIBUTION OF LOANS ISSUED PER YEAR & PERCENT DEFAULT RATES OVER THE YEARS



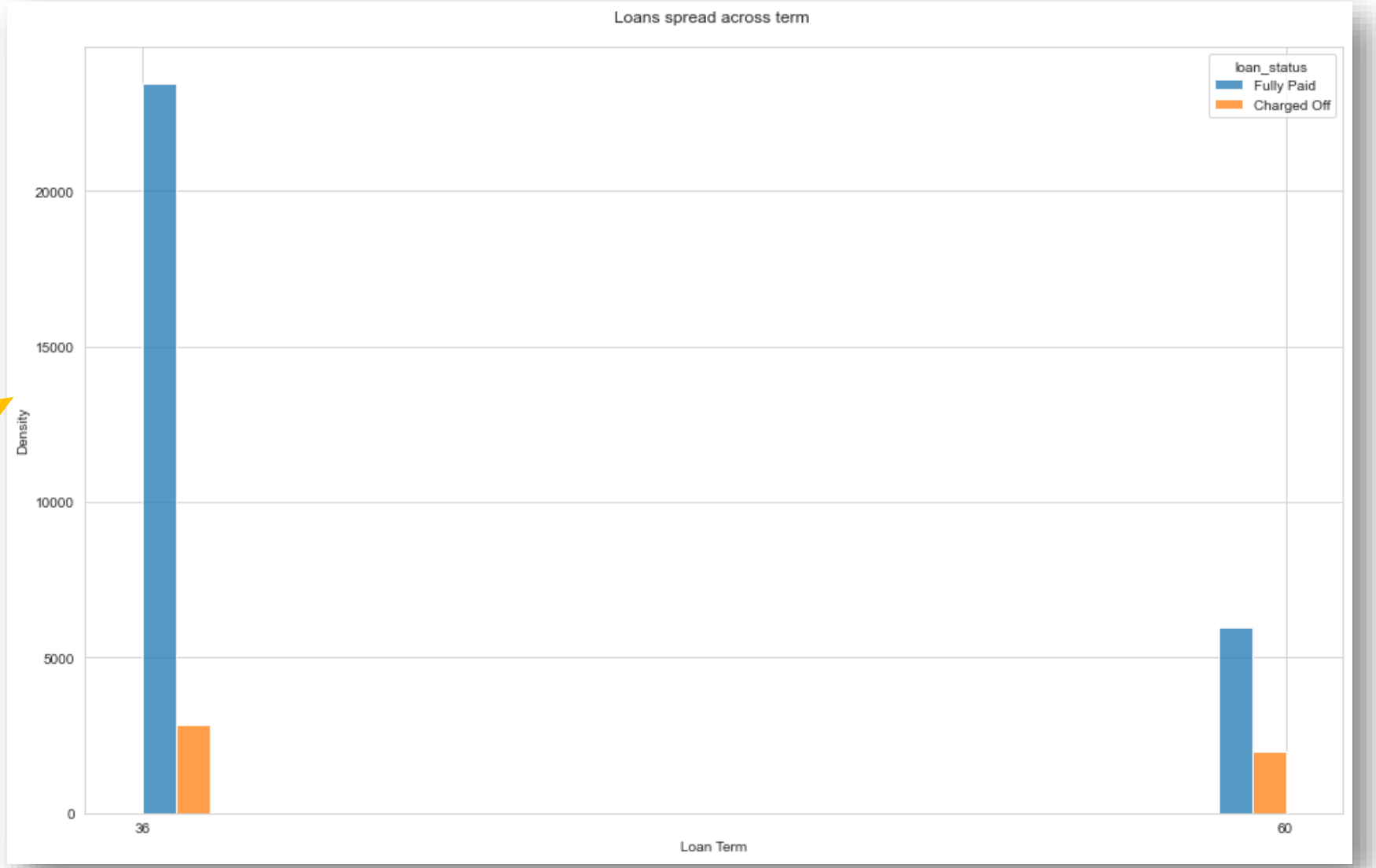


For both segments, the number of the issued loans increase steadily over the years. This indicates a growing business.

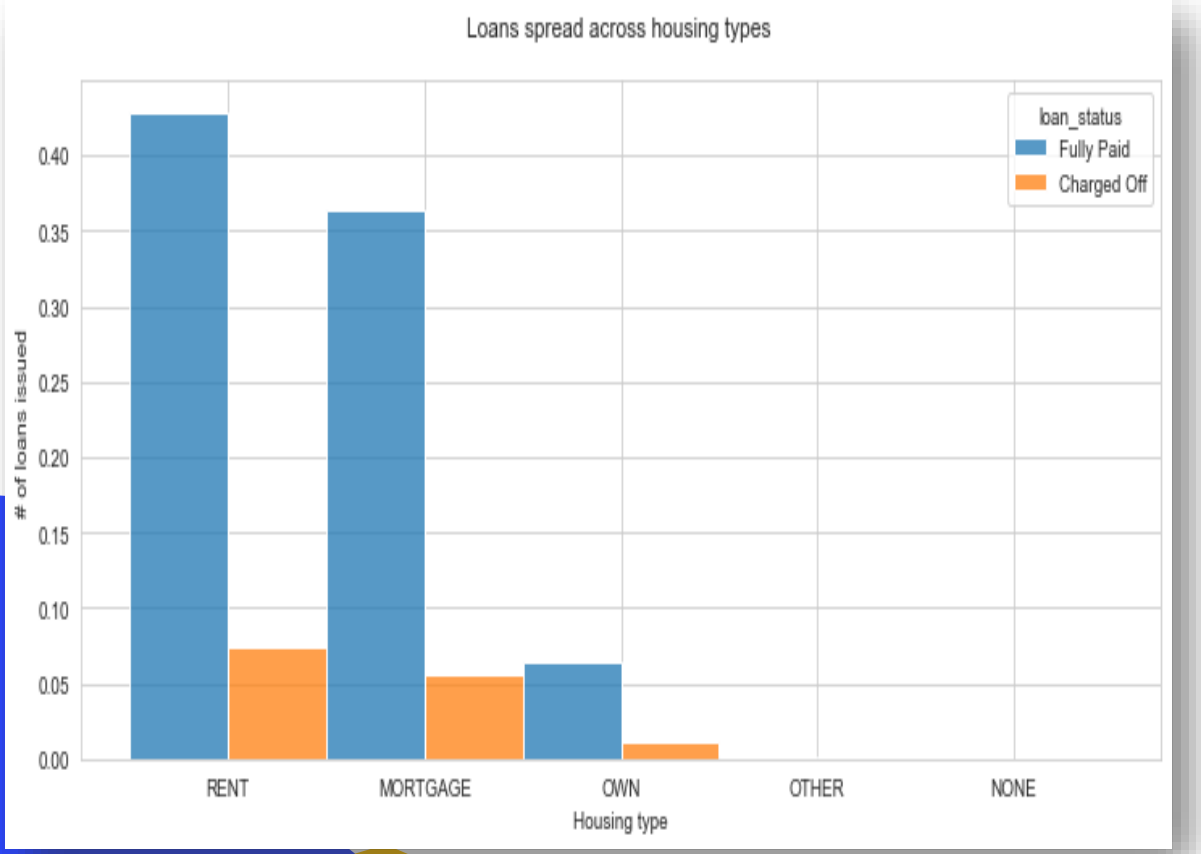
Considering small business in 1st two years, we ignore them. A relatively significant increase in last year default rate is concerning. Are we prepared for further business growth?

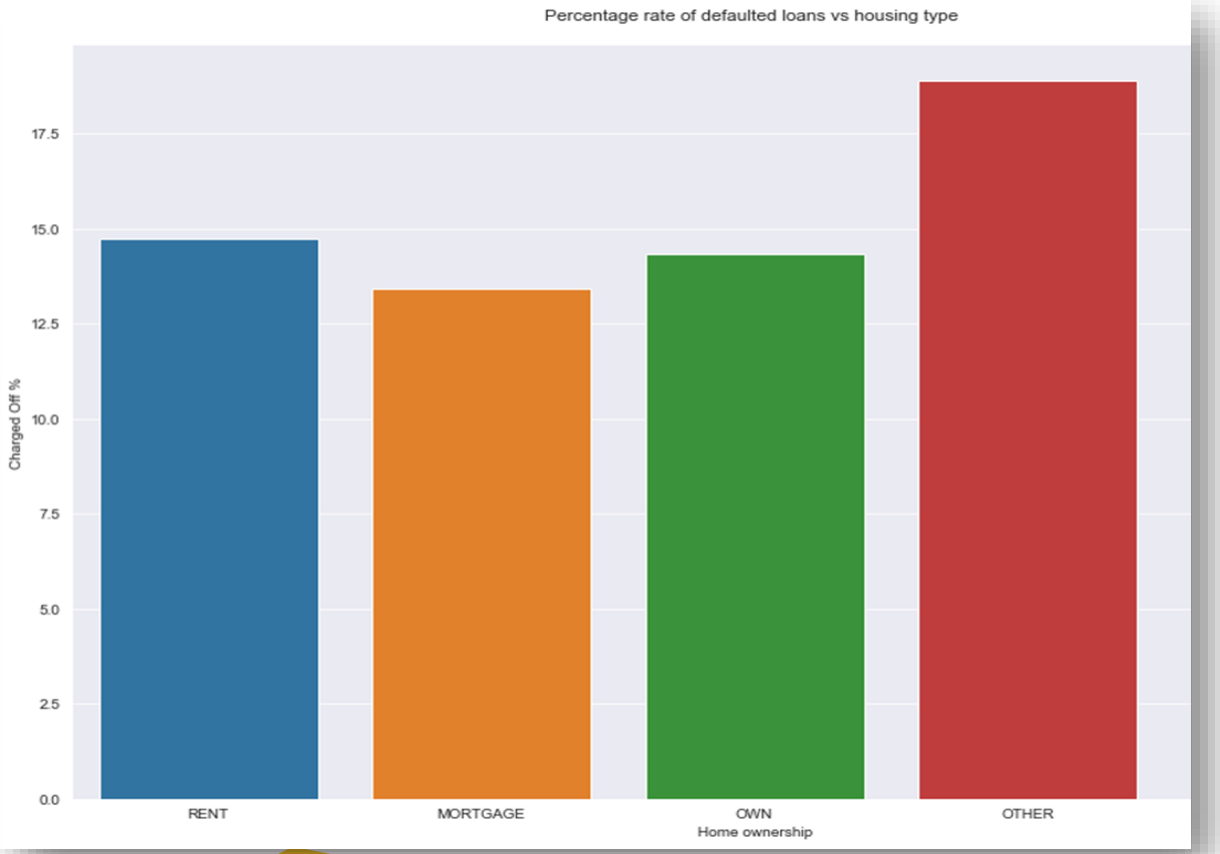
DISTRIBUTION OF LOANS ACROSS LOAN TERMS

>75% loans are for 36 months. More number of loans(both paid and defaulted segments) are issued for 3 years



DISTRIBUTION ACROSS HOUSING TYPE AND PERCENT CHARGED RATE OVER THEM



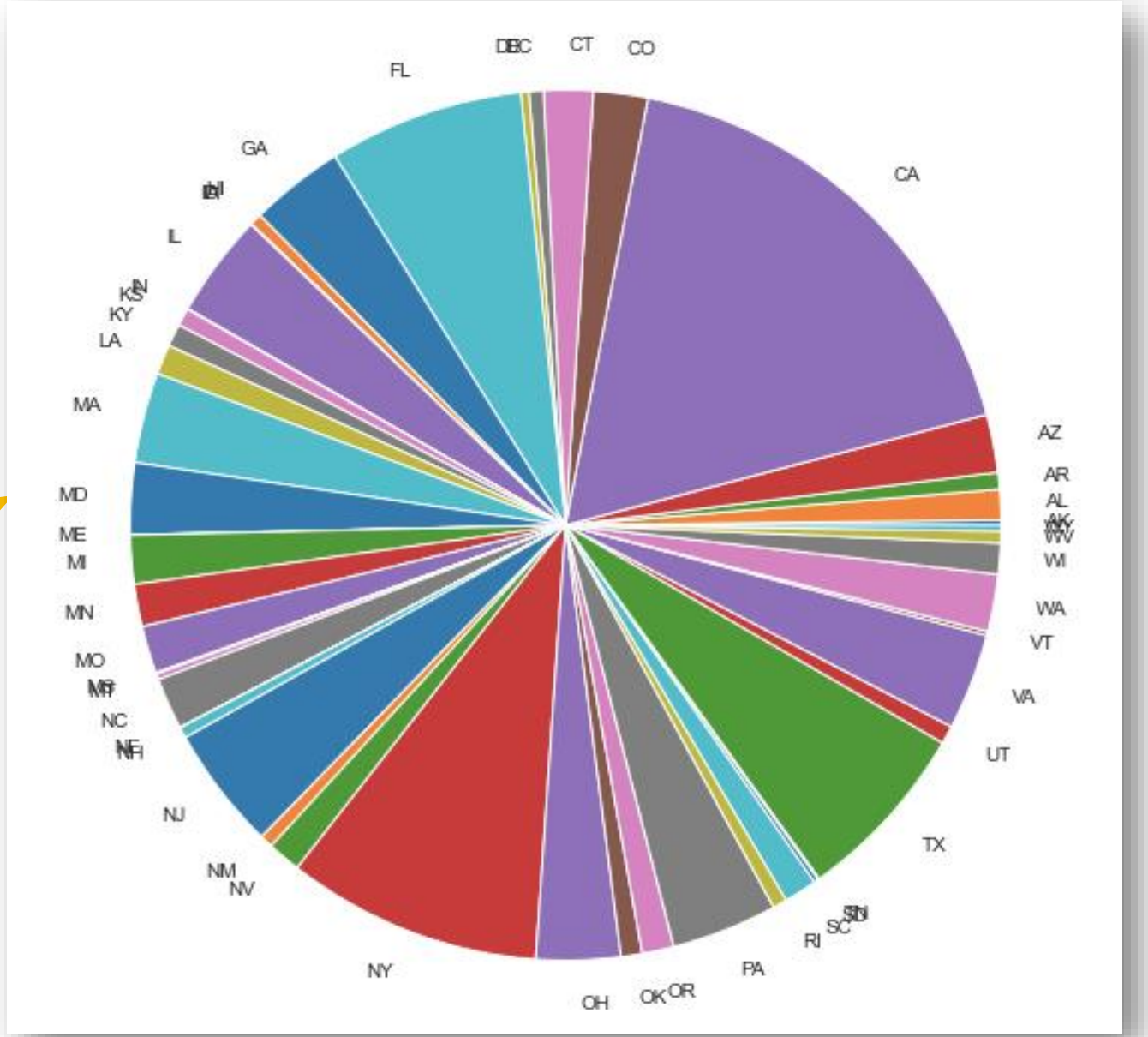


For both segments, most of the business comes from customers who have rented homes or are paying mortgages.

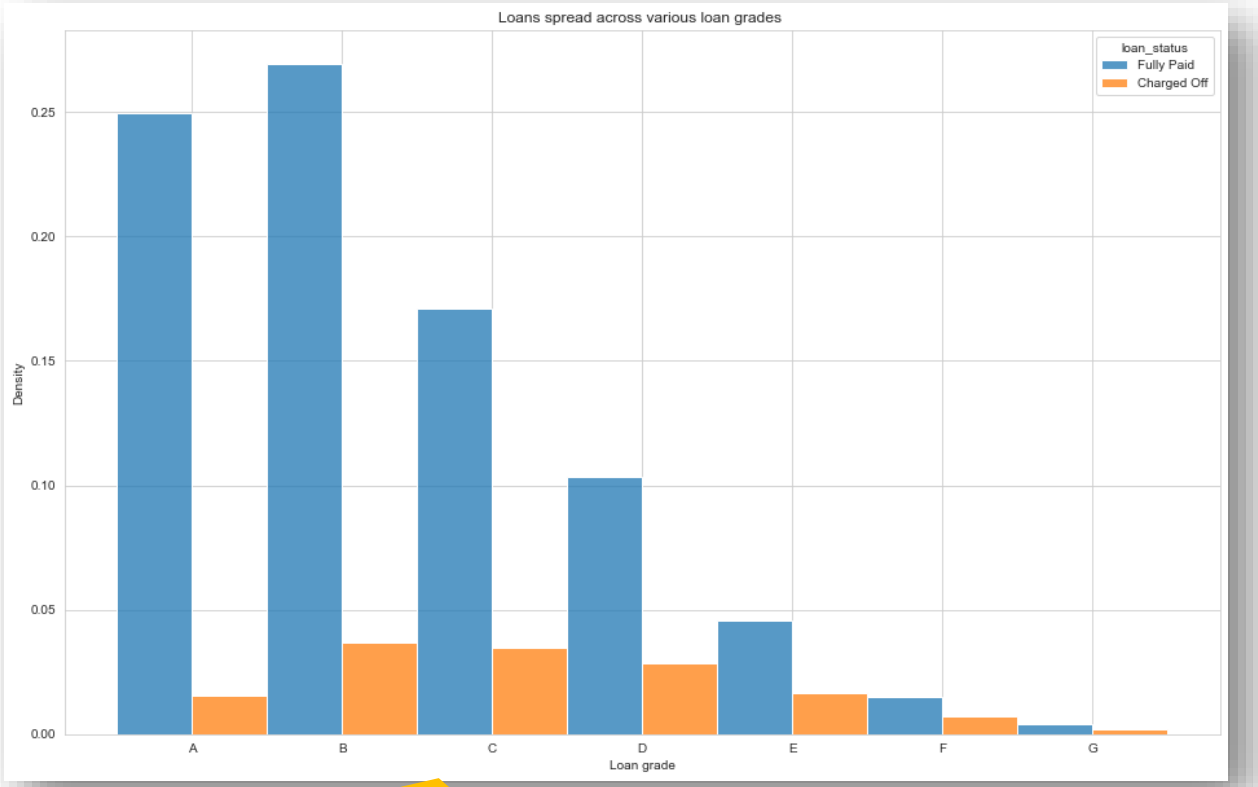
Ignoring OTHER(very few loans), the default rates for largest business category(RENT) are relatively higher than MORTGAGE. Need more stringent evaluation for customers with rented house.

DISTRIBUTION OF LOANS ACROSS STATES

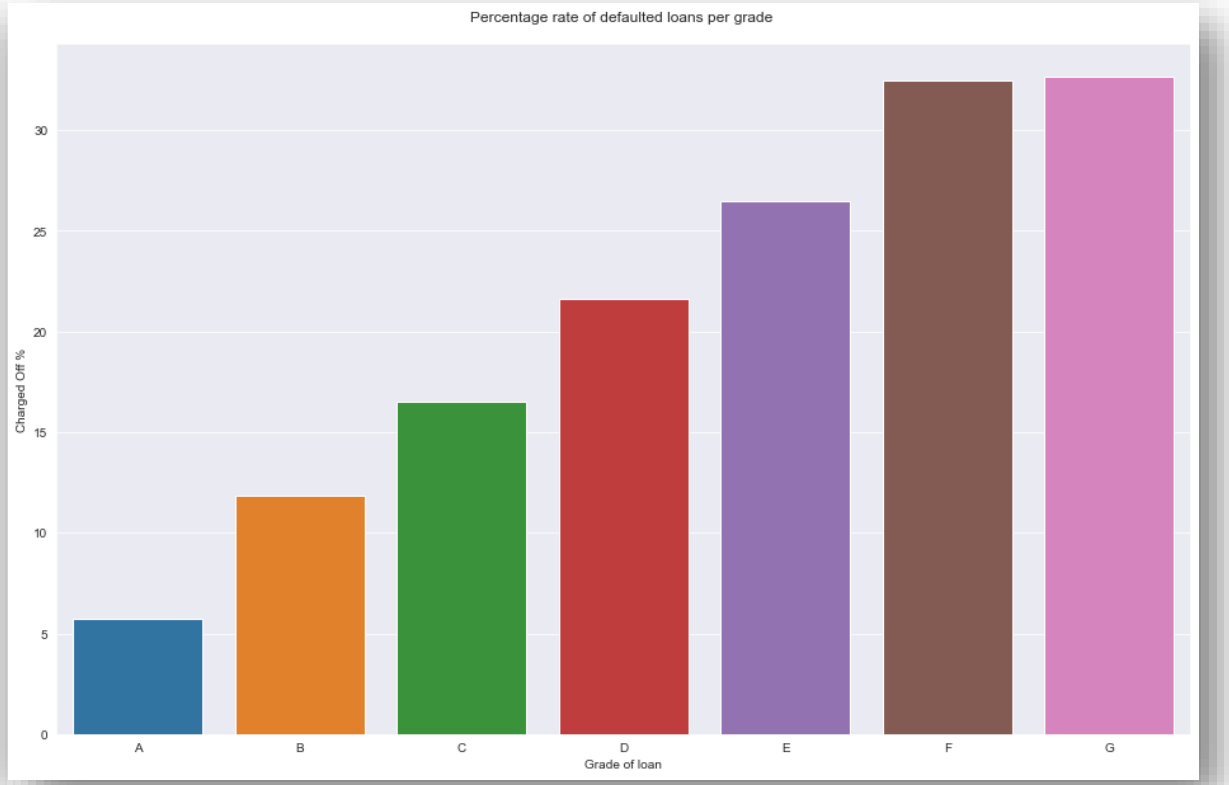
A major number of the loans are issued to customers in states of California, NY & Florida



DISTRIBUTION OF LOANS ACROSS LOAN GRADES & PERCENT DEFAULT RATES

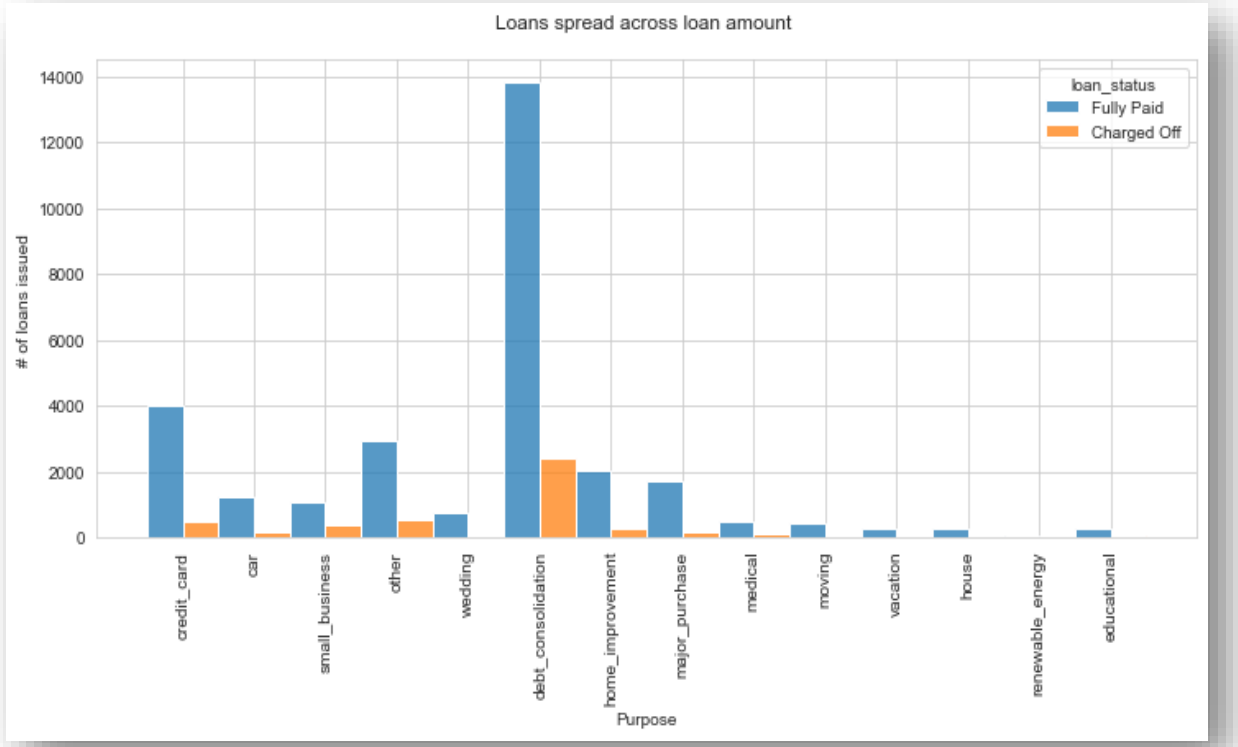


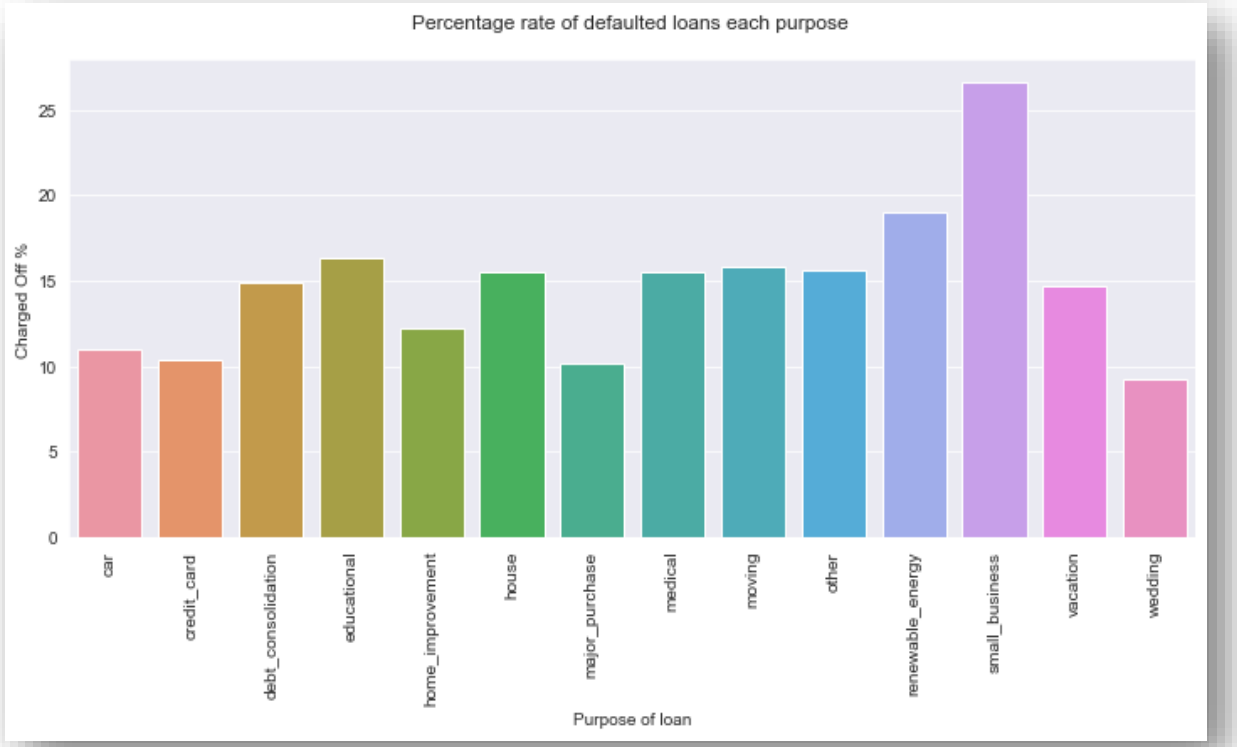
Majority of loans (both paid and defaulted segments) fall in good grades A,B & C. The number of loans issued reduce rapidly as grade goes down



As the loan grade lowers, loan default rate increases & become more risky.

DISTRIBUTION OF LOANS ACROSS PURPOSES OF LOAN & PERCENT CHARGE-OFF RATE

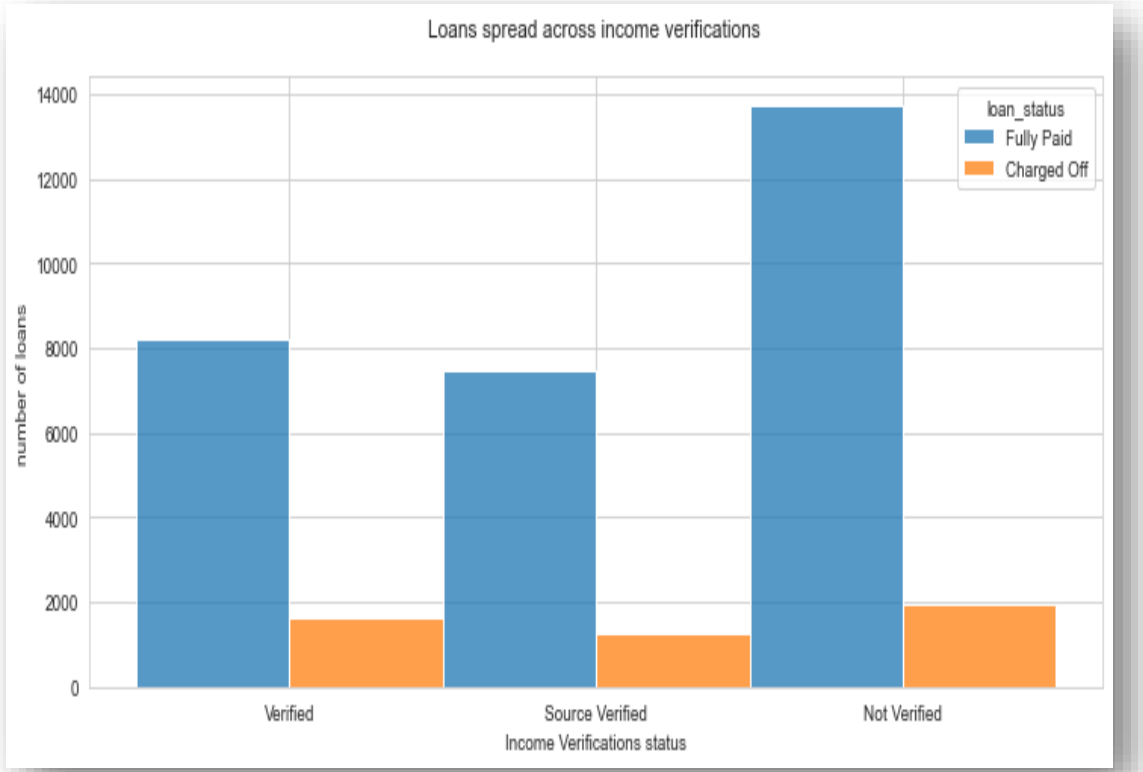




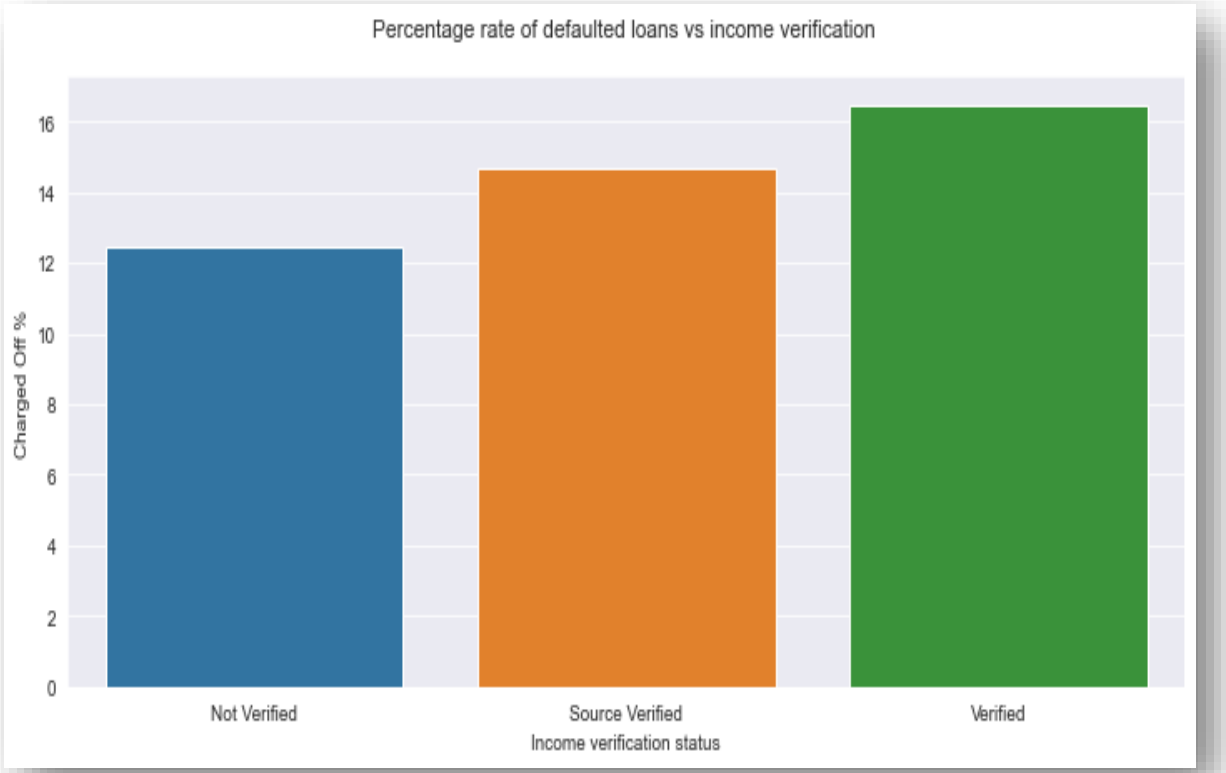
The majority of the loans are given for debt consolidation followed by credit cards & other

The loans issued to "Small Businesses" are most risky followed by “renewable energy”.

DISTRIBUTION OF LOANS ACROSS INCOME VERIFICATION STATUSES & PERCENT DEFAULT RATES

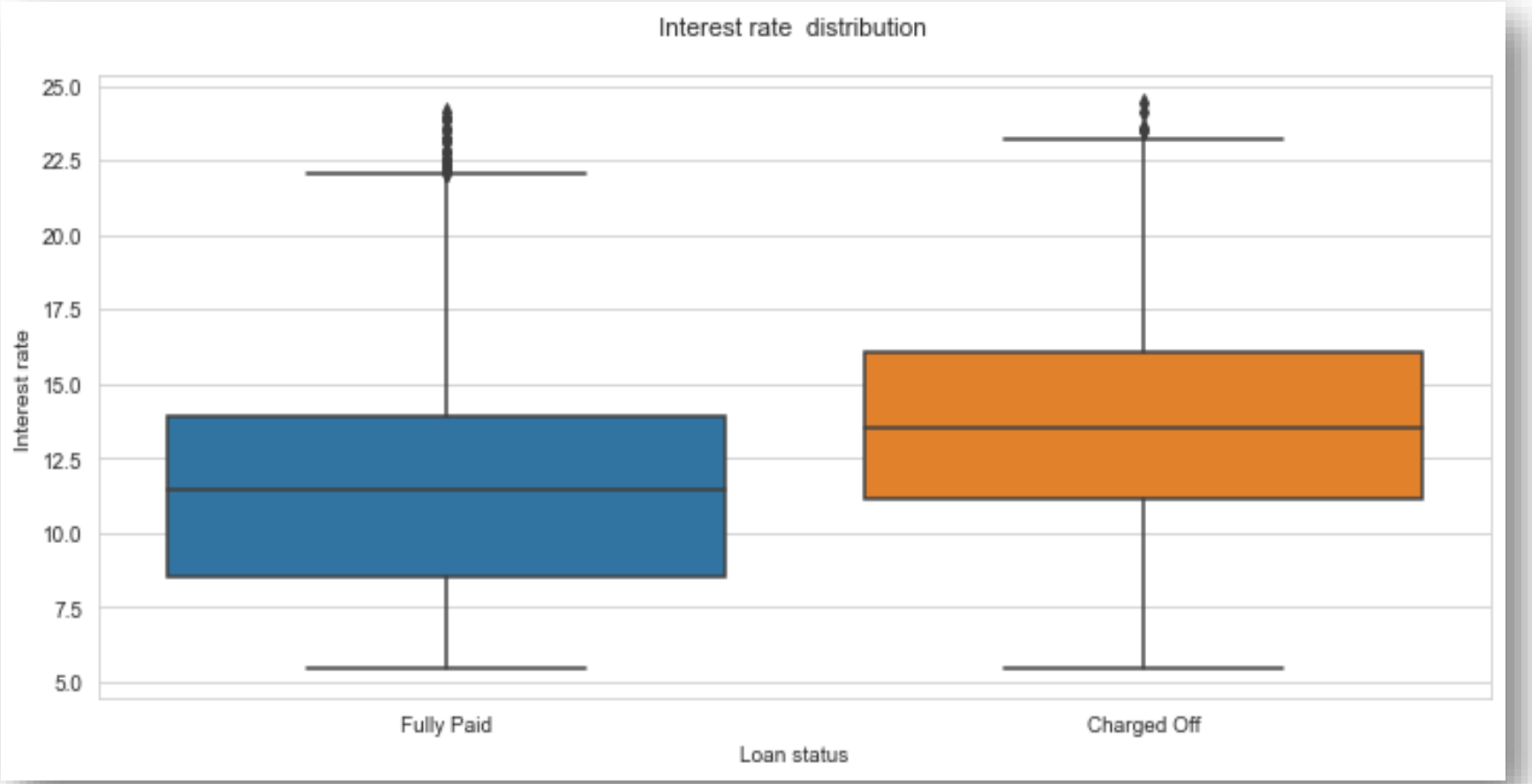


A lot more loans exist where customer information is not verified which is concerning.



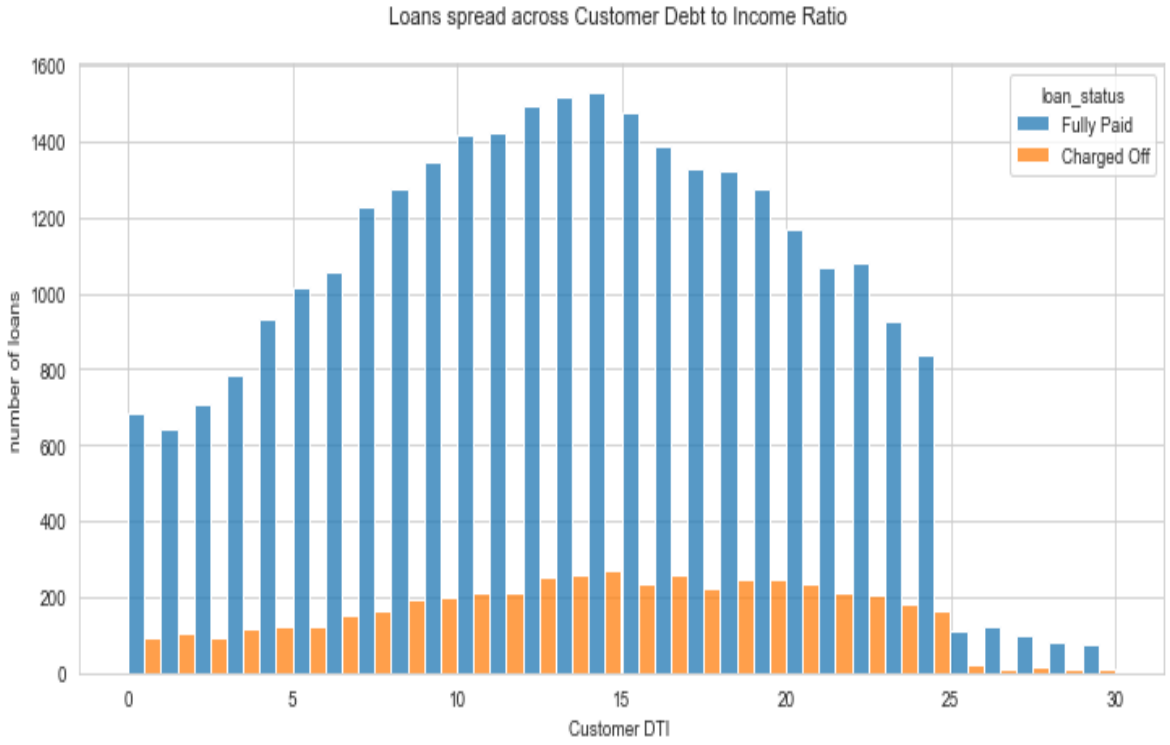
The default rate is highest for loan applications where income has been verified. This is counterintuitive. Data error?

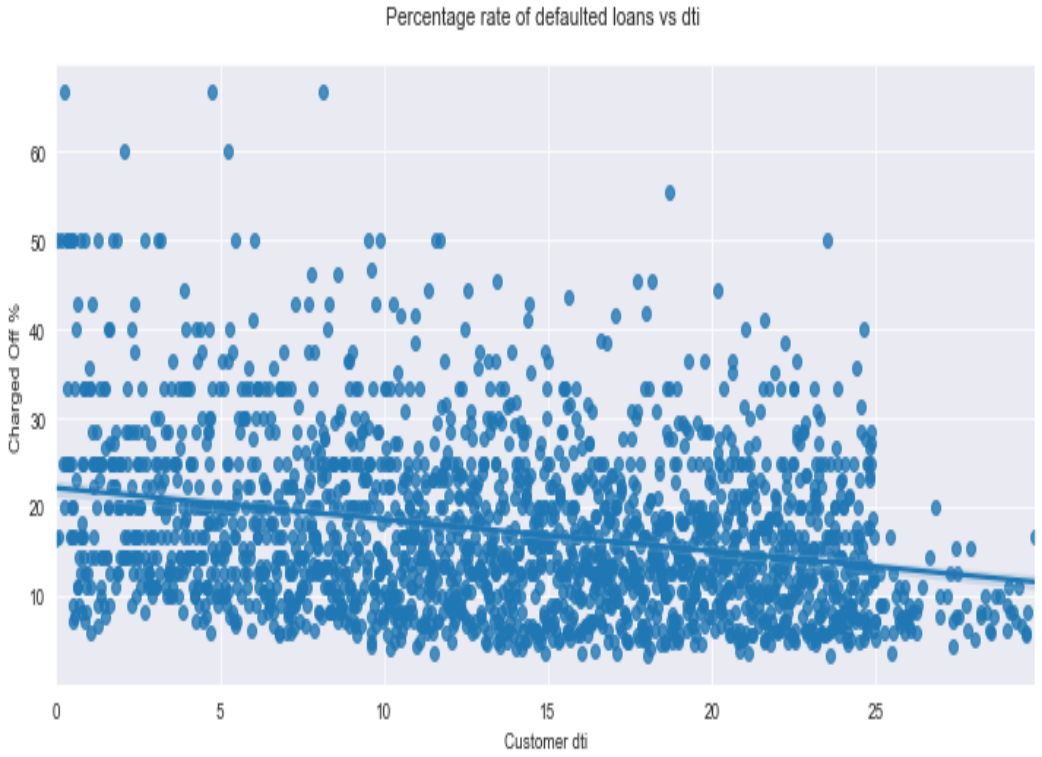
DISTRIBUTION OF INTEREST RATES ACROSS BOTH SEGMENTS



Interest rates are relatively higher for charged-off loans For fully paid loans, there are significantly higher number of outliers as compared to Charged off loans

DISTRIBUTION ACROSS DEBT-TO-INCOME-RATIO(DTI) AND PERCENT CHARGED OFF RATE

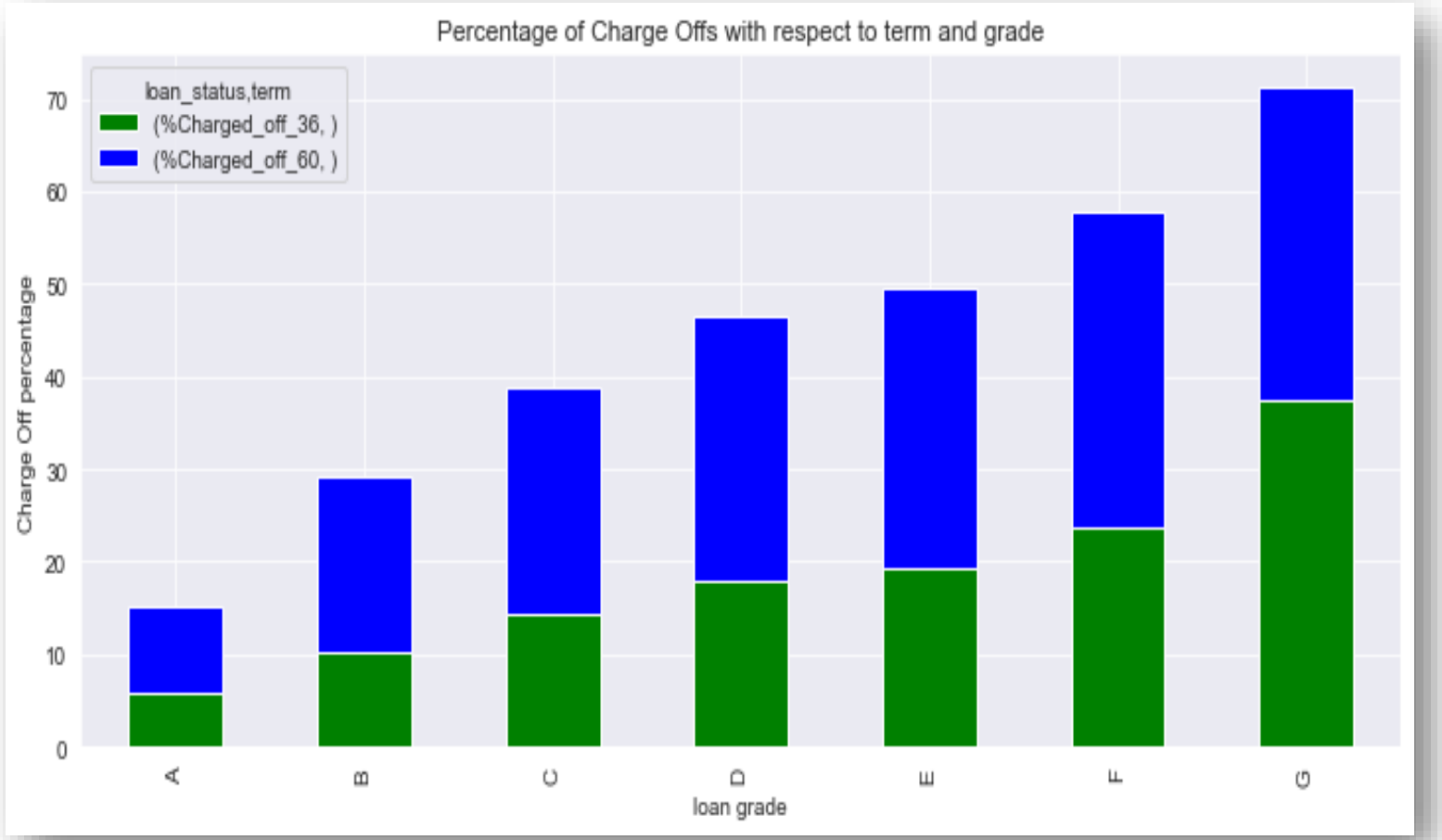




The majority of loans are issued to customers with average DTI.
More application rejections can explain less number of loans for high debt ratio customers

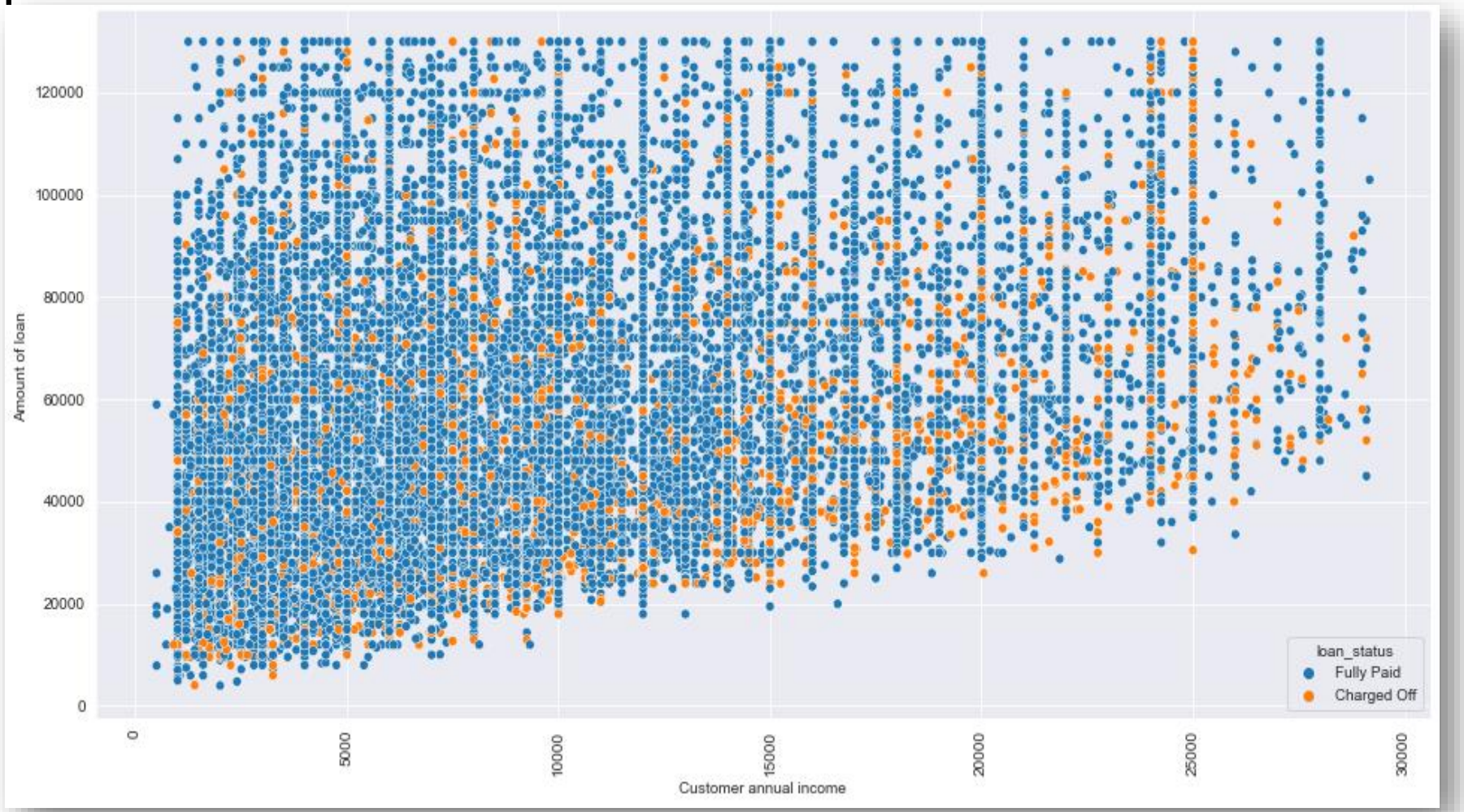
The loan default rate is higher for customers who have low debt to income ratio which is odd. We'll corelate it with other factors to see what might negatively impact loans when good customer having low DTI.

LOAN TERM VS LOAN GRADE



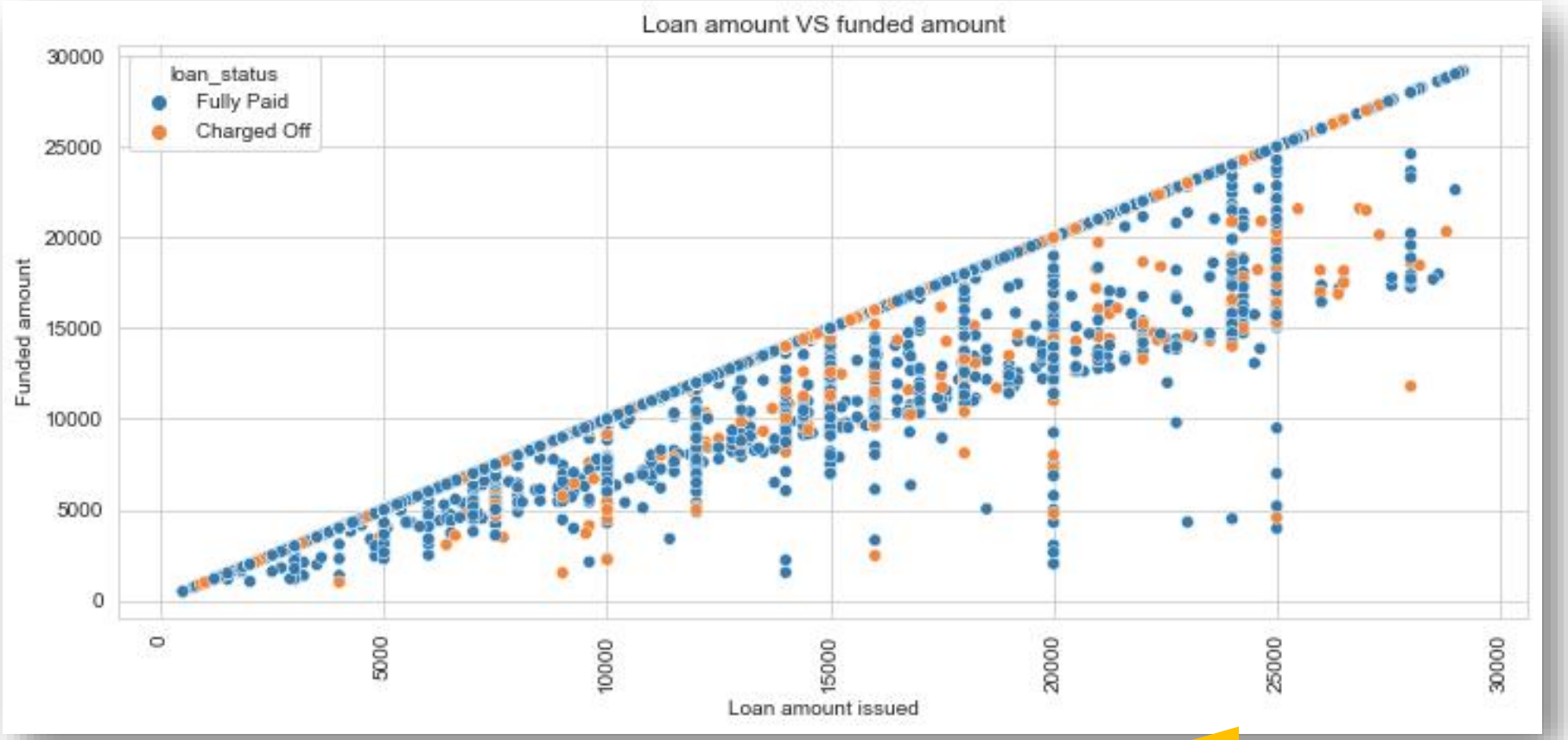
For each loan grade (Except for grade G), the charge off percentage in 60 months long loans is higher than 36 months loans. So higher term loans are doing consistently poor than lower term.

LOAN AMOUNT VS CUSTOMER ANNUAL INCOME



All ranges of loan amounts are issued to customers regardless of their annual incomes for both segments. This indicates there are more factors which decide the loan amount to be issued.

LOAN AMOUNT VS FUNDED AMOUNT

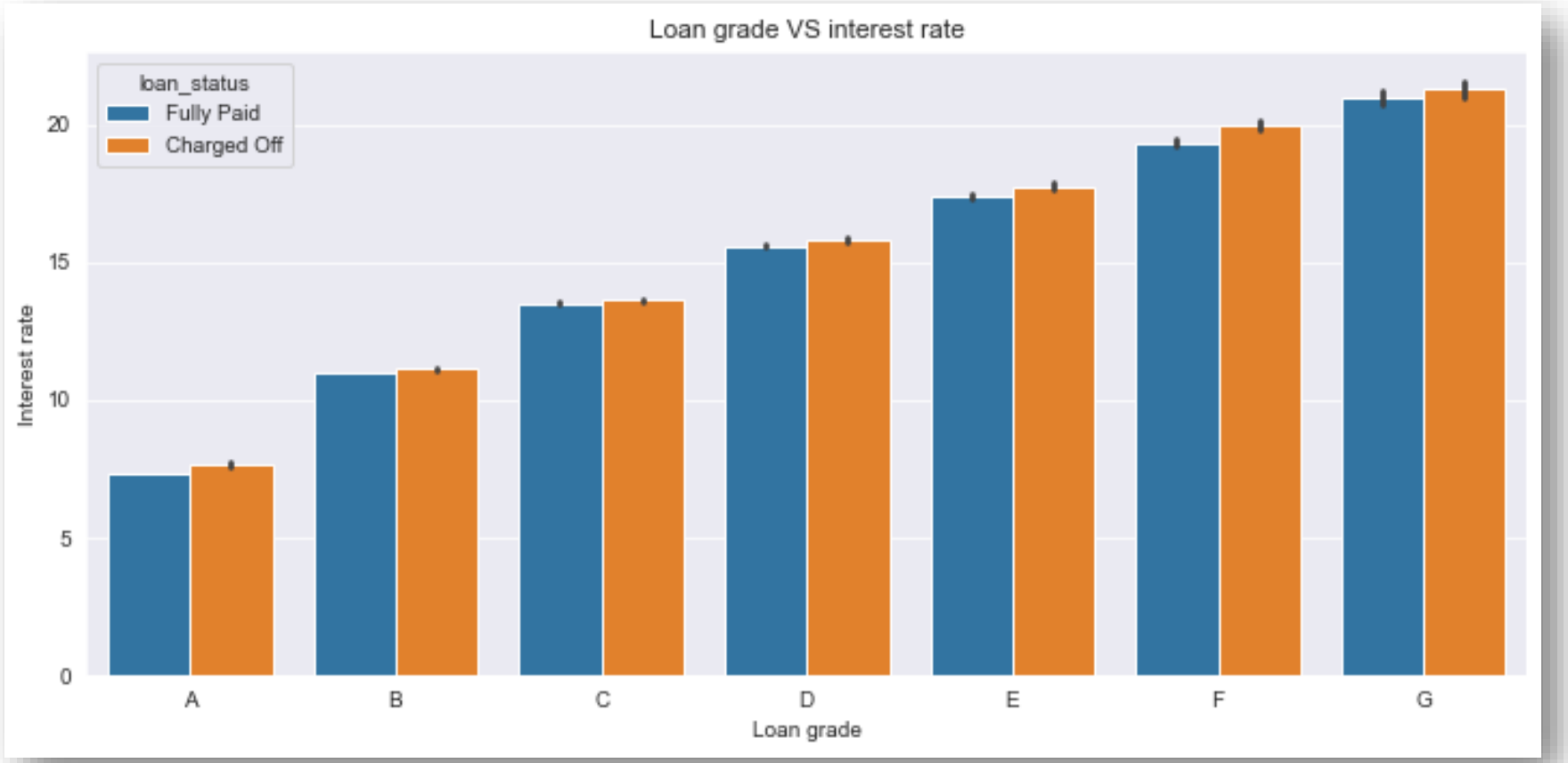


The loan amount is directly proportional to funded amount and is always less than funded amount.

LOAN AMOUNT VS INTEREST RATE

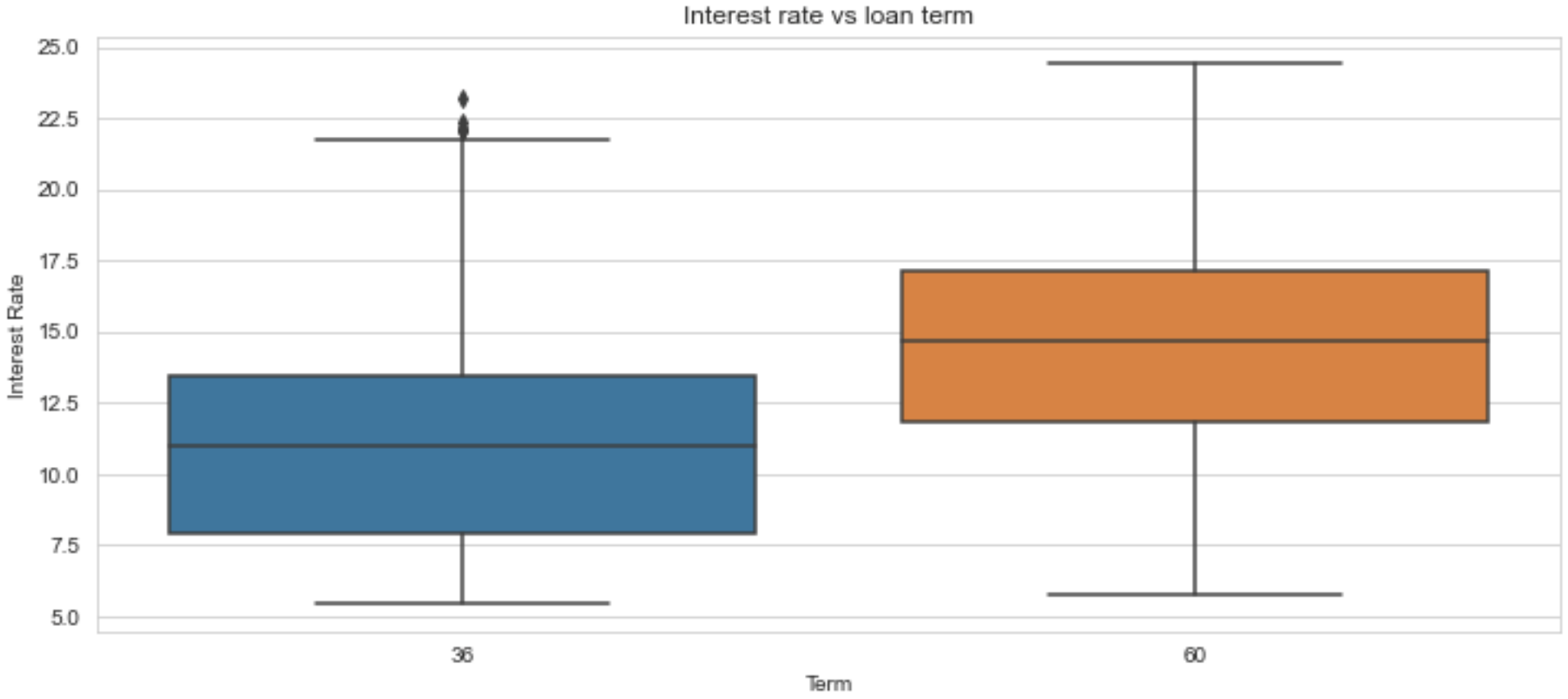
The interest rate of a loan is not directly related to loan amount in both segments

LOAN GRADE VS INTEREST RATE



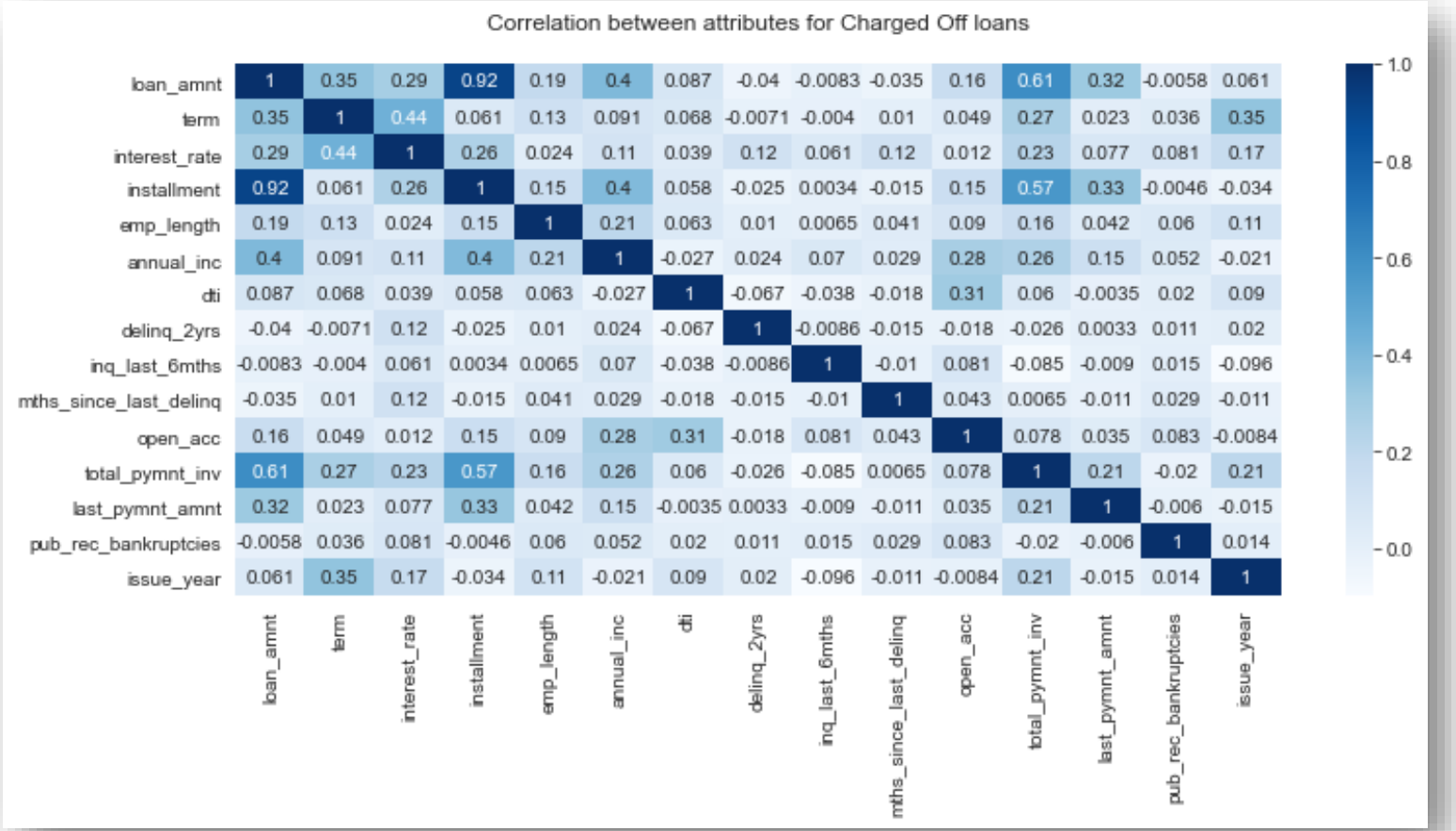
The interest rates increase as the loan grade lowers. this is to accommodate the risk associated with lower grade loans

INTEREST RATE VS LOAN TERM



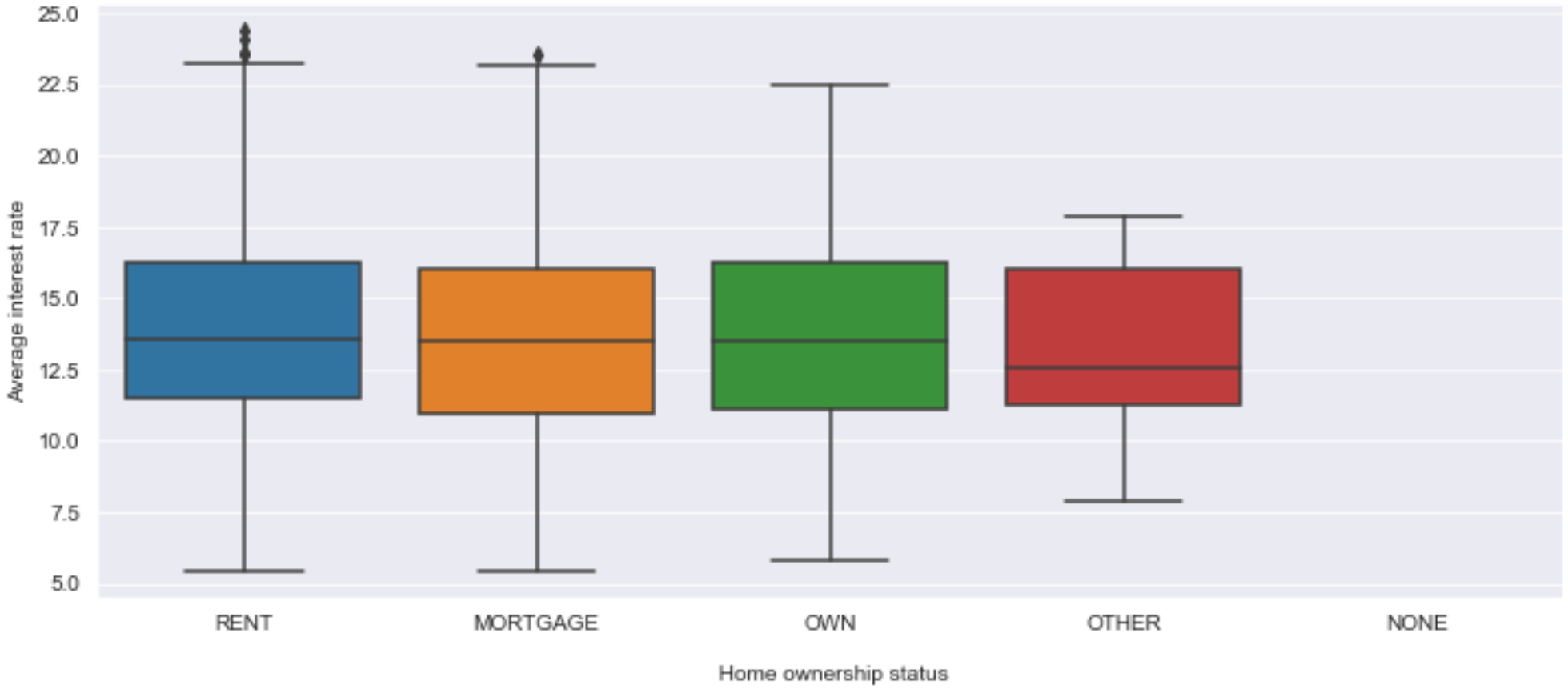
The rate of interest is higher loans of longer duration to accomodate the risk associated with higher duration loans

CORRELATION BETWEEN VARIOUS ATTRIBUTES FOR CHARGED OFF LOANS



1. The payment fields are related to loan amount.
2. The loan amount is negatively related to number of bankruptcy records of customers(pub_rec_bankruptcy).
3. Debt-to-Income (DTI) is negatively related to annual income.

HOUSING STATUS VS INTEREST RATES

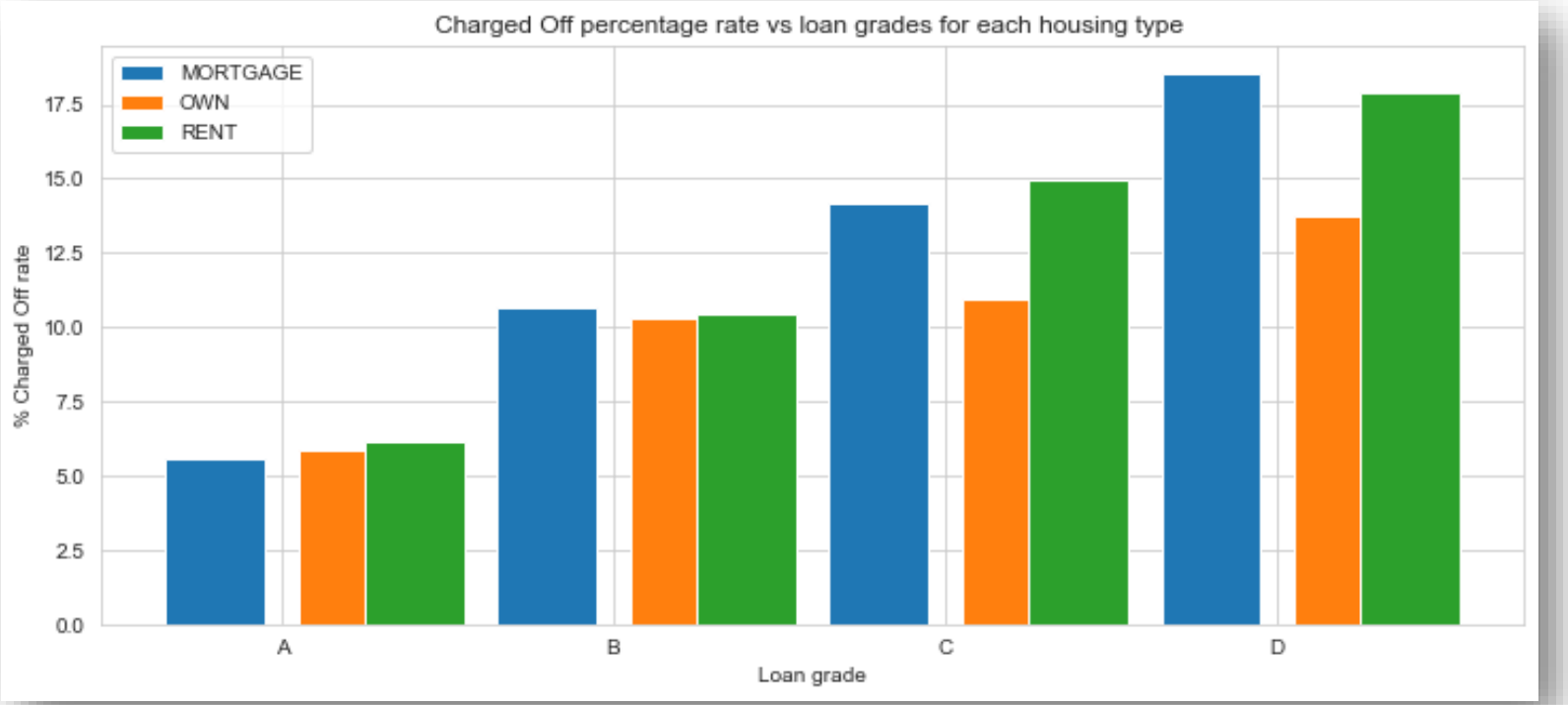


The median Interest rate is nearly similar for all housing status. When we relate it to previous analysis, where we concluded that OWNERs and OTHER housing customers have more tendency to default, what are other factors that are driving interest factors other than Housing status

CHARGED OFF % RATE VS LOAN GRADES FOR EACH HOUSING TYPE

HERE WE ARE GOING TO PICK ONE OF MOST POPULAR SEGMENT ACCORDING TO CRITERIA ON DTI, TERM AND GRADES & HOMEOWNER.

LOAN GRADES = A,B,C, D
HOUSING= RENT, MORTGAGE, OWN
DTI 10 TO 20
LOAN TERM 36



The grade A&B loans have similar default rates for all three categories. However, mortgage and rental customers default more for loan C & D grades. From previous analysis, these make up a significant number of loans.

#CONCLUSION: Grade C, D loans for mortgaged and rented customers needs to be further evaluated to understand the higher loan defaulted rates.

THANK YOU

Navita Goel
Priyanka Kumari

PS: refer to NavsGo/LendingClub_CaseStudy repo on github to access the code for data cleansing and analysis.