LENDING CLUB CASE STUDY

-BY CA ARAVIND JAYARAMAN

AIMS AND OBJECTIVES

- The problem faced by the lending club is that if an applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- On the other hand, 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/ investor.
- The main aim of this case study is to identify the factors that most likely contribute to the default on loans by borrowers based on past data.
- This will help the approver in approving future loans in the most optimal way so as to maximize investor returns and minimize the risks associated.

SOURCE OF DATA AND METHOD OF ANALYSIS

- All analysis made in this case study is solely based on the data provided in the excel file loan.csv to us.
- The analysis is done using Python and EDA (exploratory data analysis techniques) using visualizations wherever applicable.
- All analysis is solely based on the data provided in the excel file. The predictive model is solely based on the data. Cause/ reasons are identified using business judgment wherever mentioned.
- Correlation does not indicate causation, so viewers are requested to exercise utmost caution and discretion for the use of data.

DEFINITION OF CERTAIN TERMS AS GIVEN IN THE INSTRUCTIONS

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

1. **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
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

PROCEDURE FOR ANALYSIS

1. First and foremost, we have identified columns with 100% NULL values and omitted them so as to simplify our data set.

2. We made a check for anomaly. For all rows,

funded_amnt_inv <= funded_amnt <=loan_amnt

This makes normal business sense and needs to be checked. No anomalies were found

3. We excluded certain columns as these were customer behavior variables. The customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.

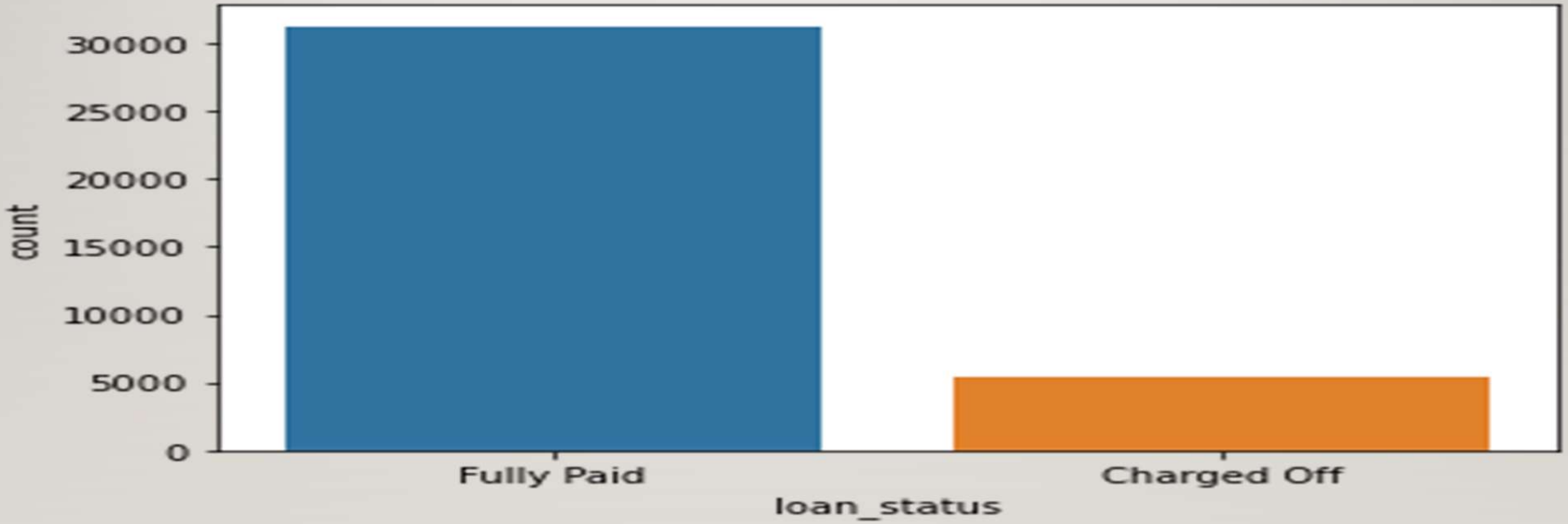
PROCEDURE FOR ANALYSIS

- 4. We also dropped some irrelevant columns that were of no value to the analysis whatsoever.
- 5.We also dropped records where the loan status was "Current".We made a new data in Pandas excluding records where loan status was current.This was done so that original data is preserved.
- 6.We made a check for missing values and decided to exclude columns where missing values >= 60% but did not find any.
- 7.We could have imputed the remaining missing values using methods such as replacing with Mode, K-nearest neighbour model etc but we chose not to do any as it wasn't expected.

PROCEDURE FOR ANALYSIS

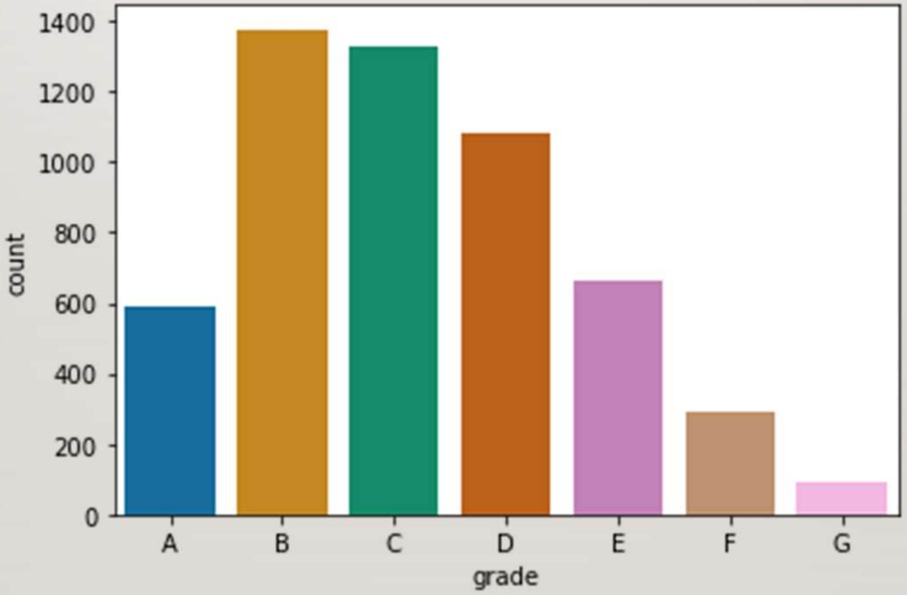
- 8. We dropped some single valued columns
- 9. We later checked for outliers using box plots and other EDA techniques.
- 10. Performed various visualizations

HISTOGRAM SHOWING COUNT OF LOANS CHARGED OFF



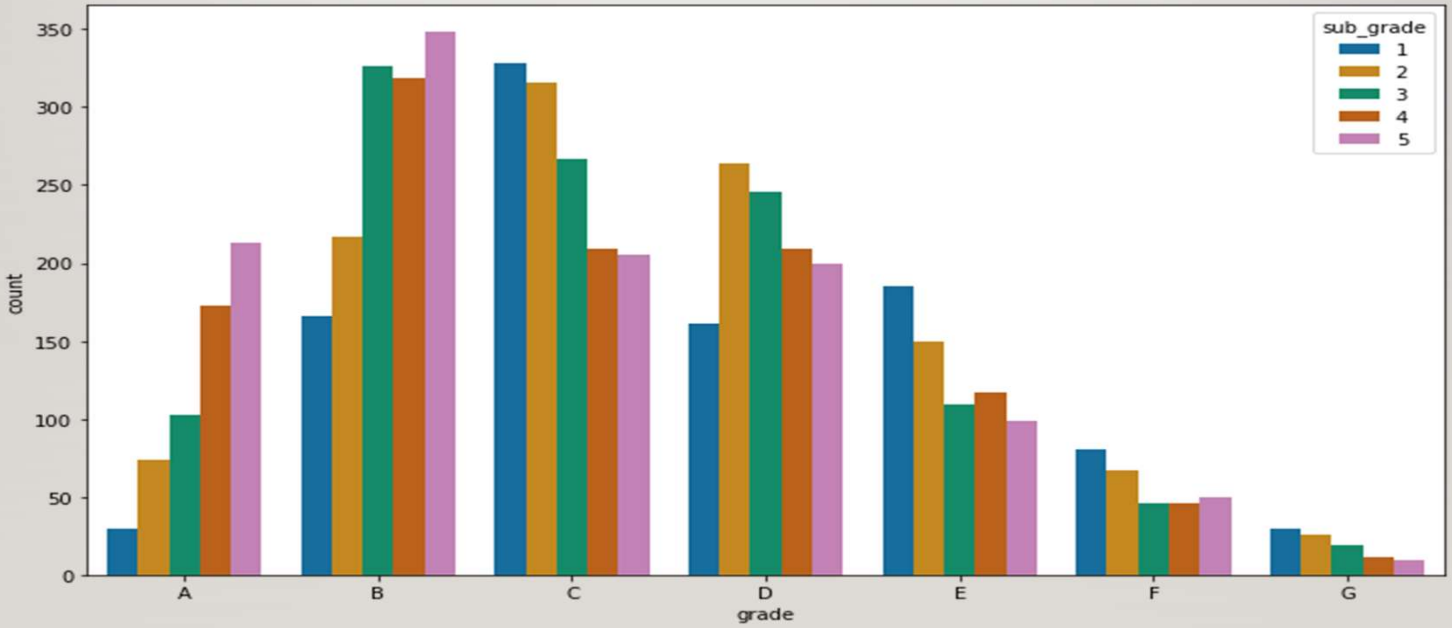
Going forward, we will analyze only those loans where there have been defaults

GRADEWISE ANALYSIS OF LOAN DEFAULTS



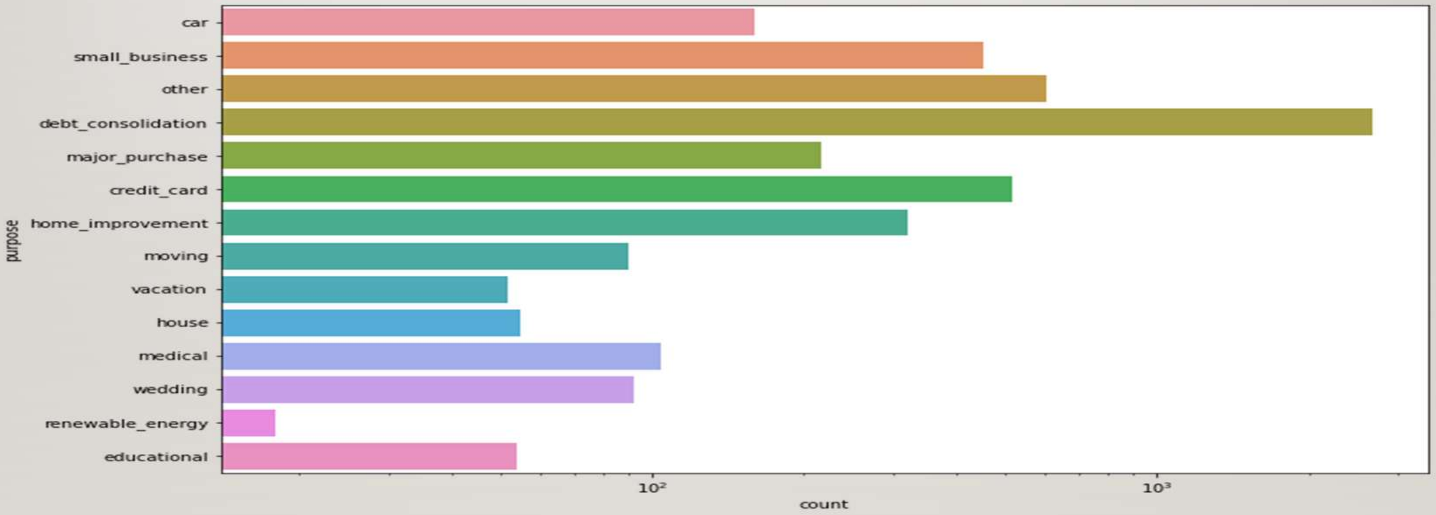
GRADE B HAS THE HIGHEST INSTANCES OF DEFAULTS

GRADEWISE ANALYSIS OF LOAN DEFAULTS



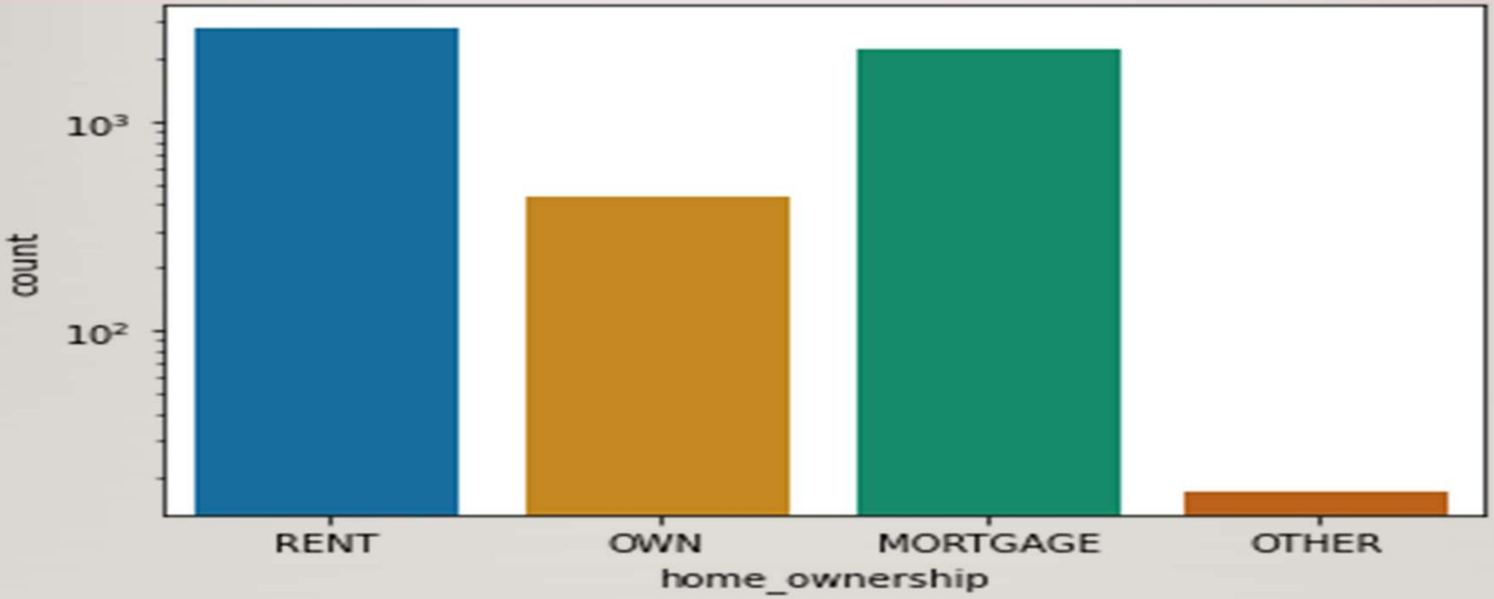
GRADE B AND SUBGRADE B5 HAS THE HIGHEST INSTANCES OF DEFAULTS

ANALYSIS OF DEAFULTS BASED ON THE PURPOSE OF LOAN



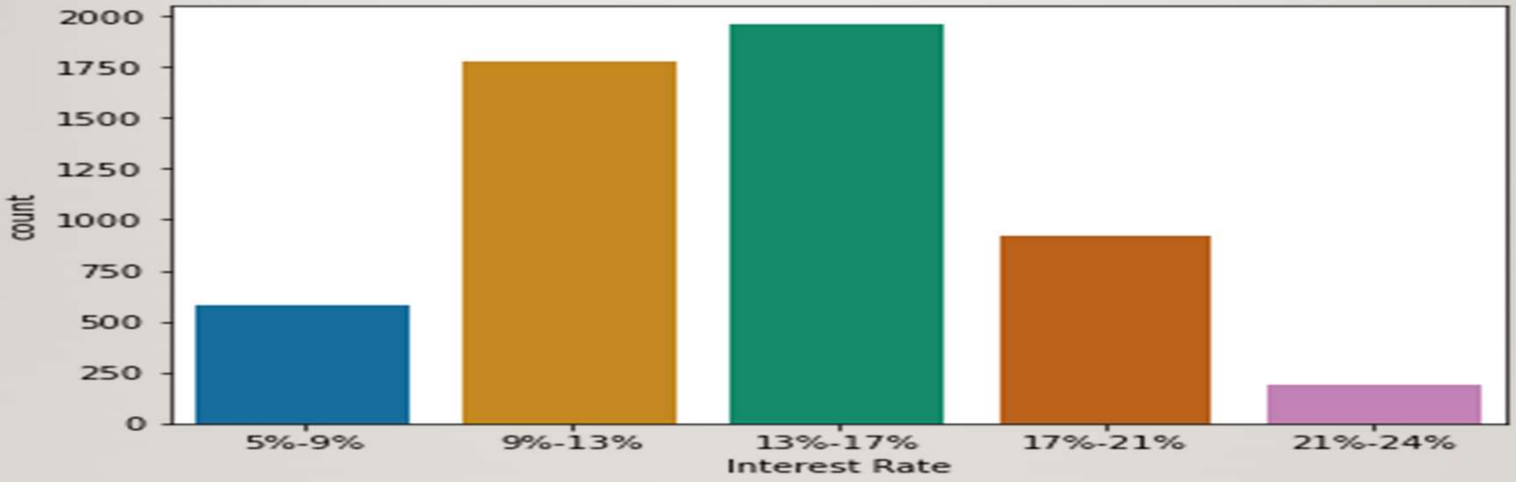
Above plot suggests that applicants whose main purpose of taking loan is to clear other loans or consolidate their debts are more likely to default. This can be due to poor financial management in general by such people.

ANALYSIS OF DEFAULTS BASED ON HOME OWNERSHIP



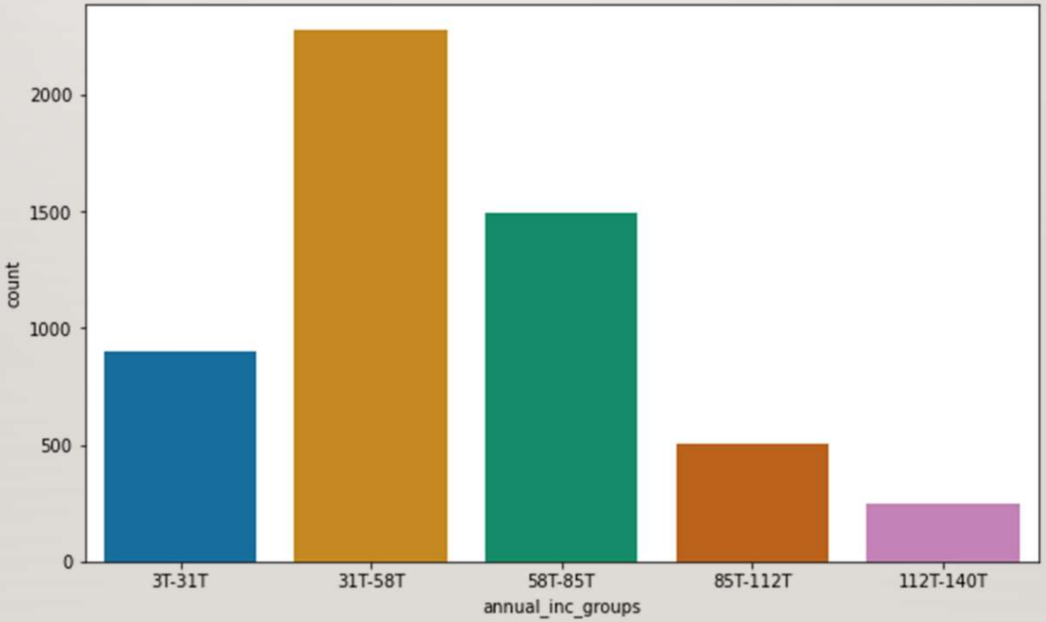
Above plot suggests that applicants living in Rented house are more likely to default than those having their own house. This maybe because those having their own house fear their house being mortgaged and attached for recovery.

ANALYSIS OF DEFAULTS BASED ON INTEREST RATE



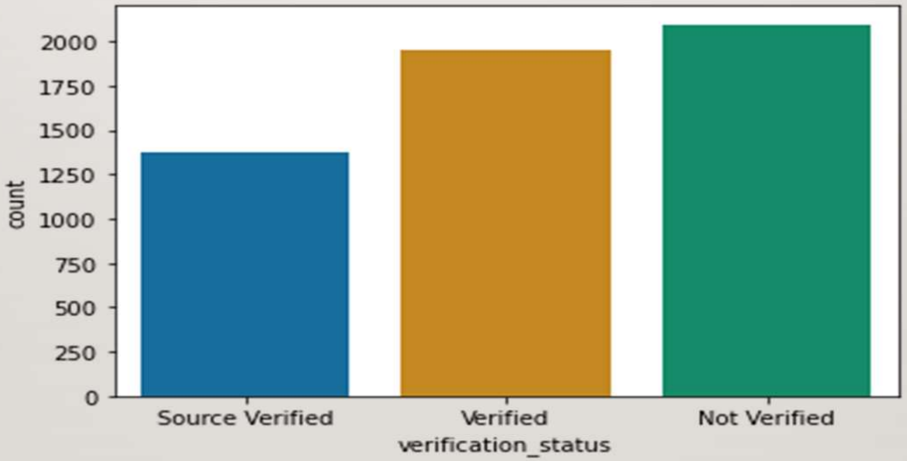
Applicants receiving interest at the rate of 13%-17% are more likely to default. This is the observation from the data.

ANALYSIS OF DEFAULTS BASED ON ANNUAL INCOME



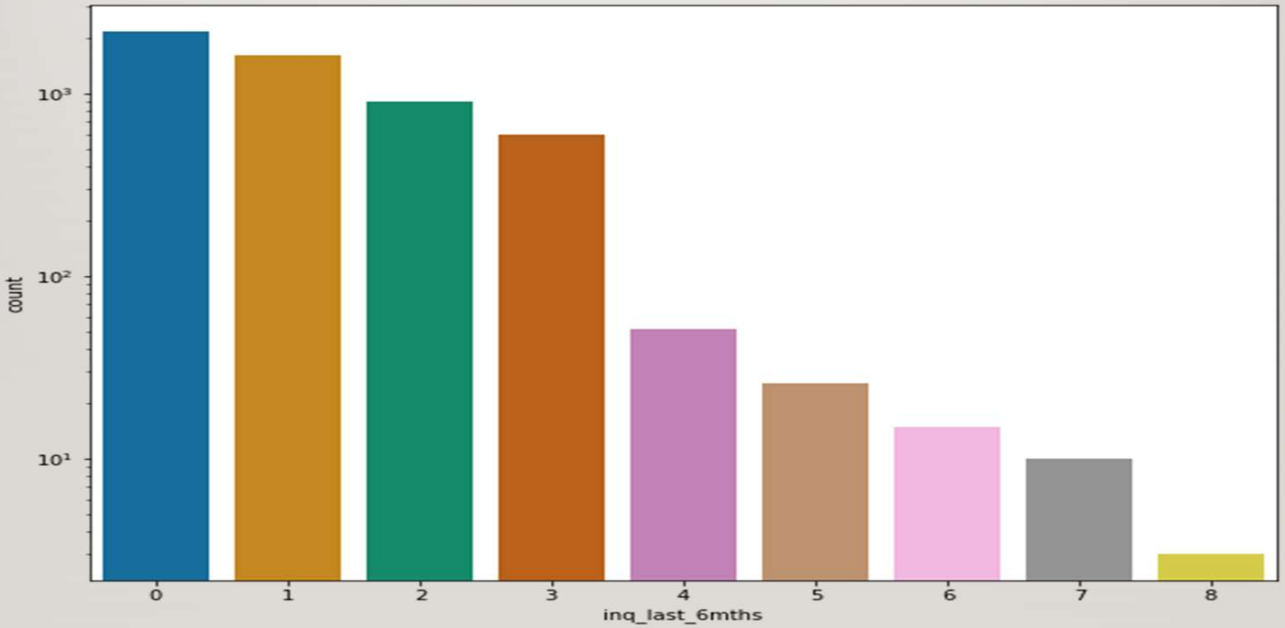
Based on the data, those earning between 31T-58T were most likely to default

ANALYSIS OF DEFAULTS BASED ON VERIFICATION STATUS

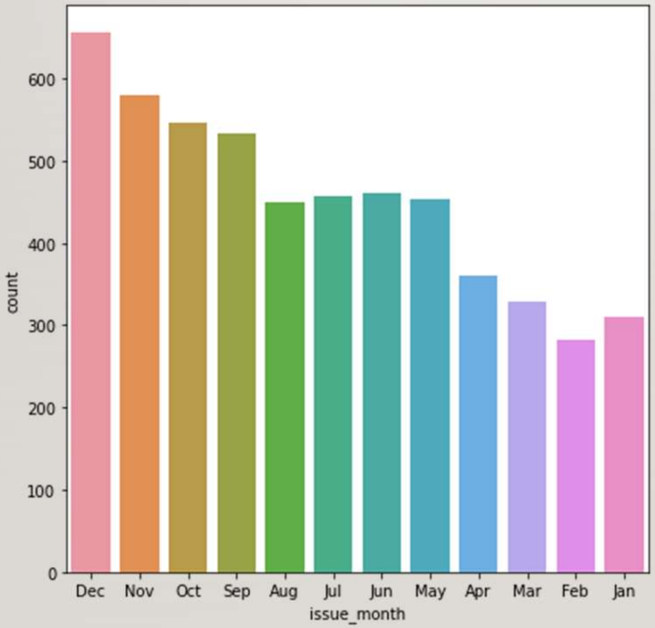


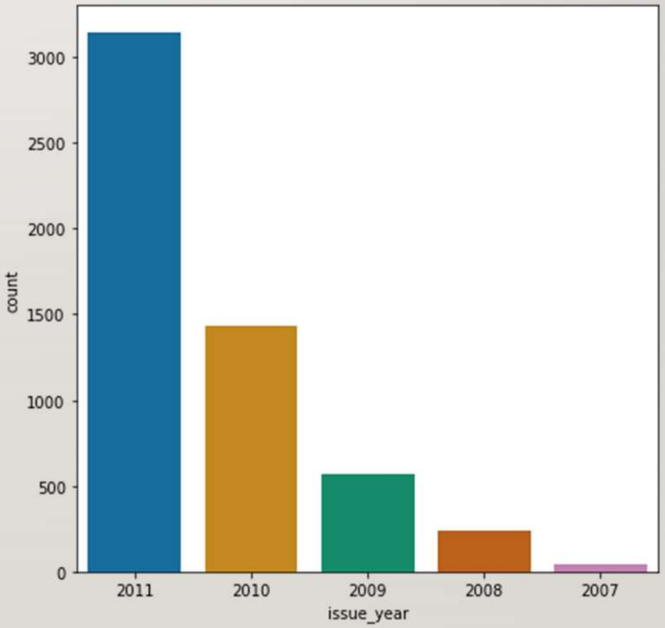
Those Loans which were not verified were most likely bad loans. This is natural as the chances of fraudulent applicants are higher

ANALYSIS OF DEFAULTS BASED ON ENQUIRIES IN LAST 6 MONTHS

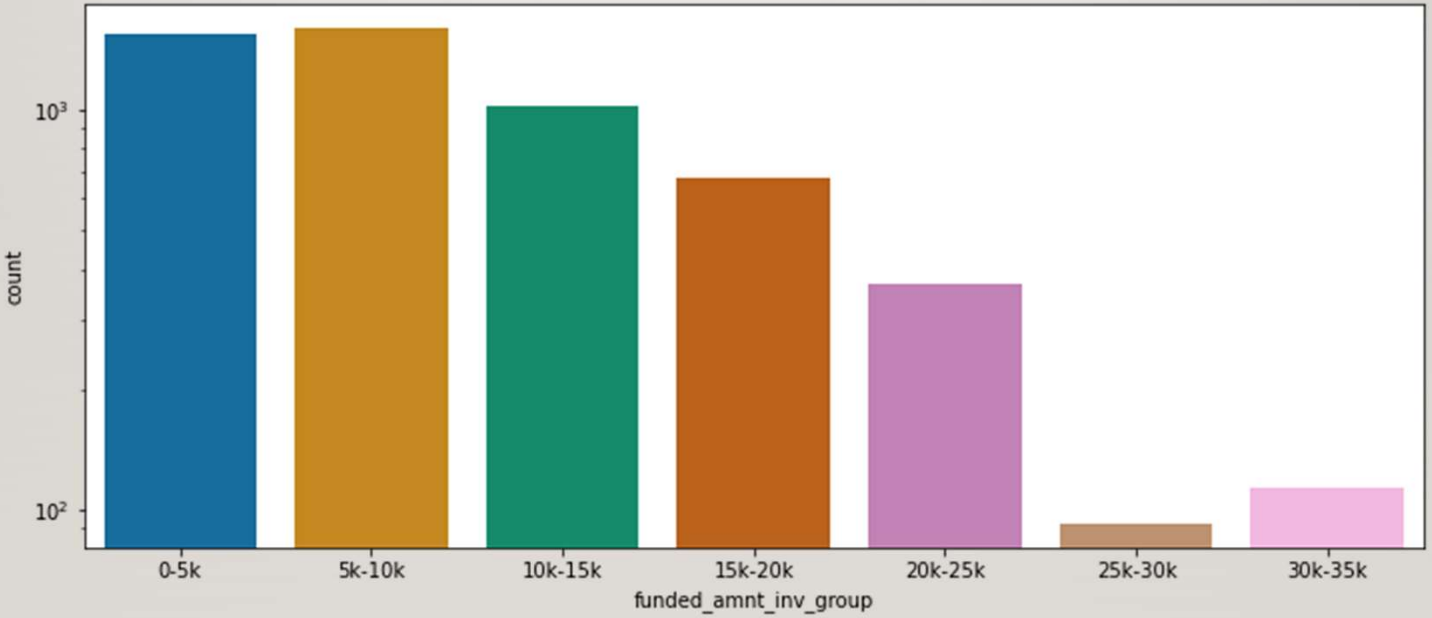


ANALYSIS OF DEFAULTS BASED ON DATE OF LOAN ISSUE

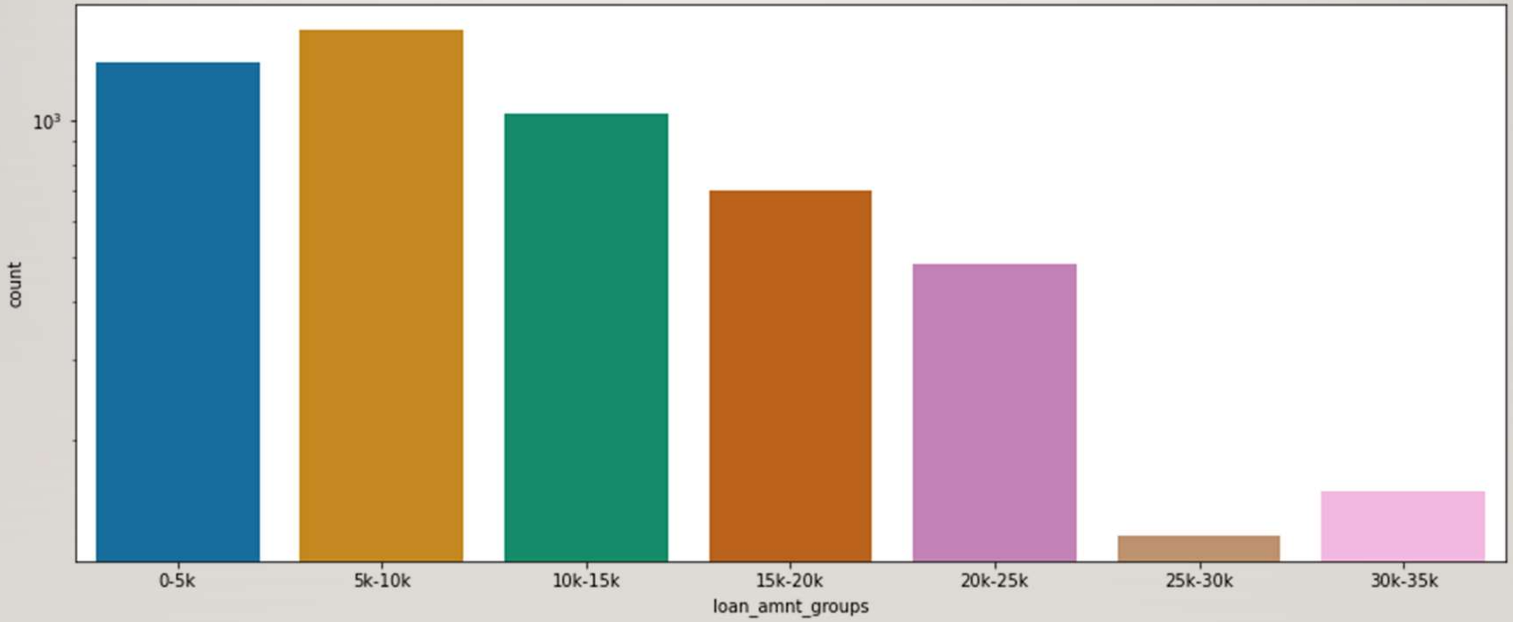




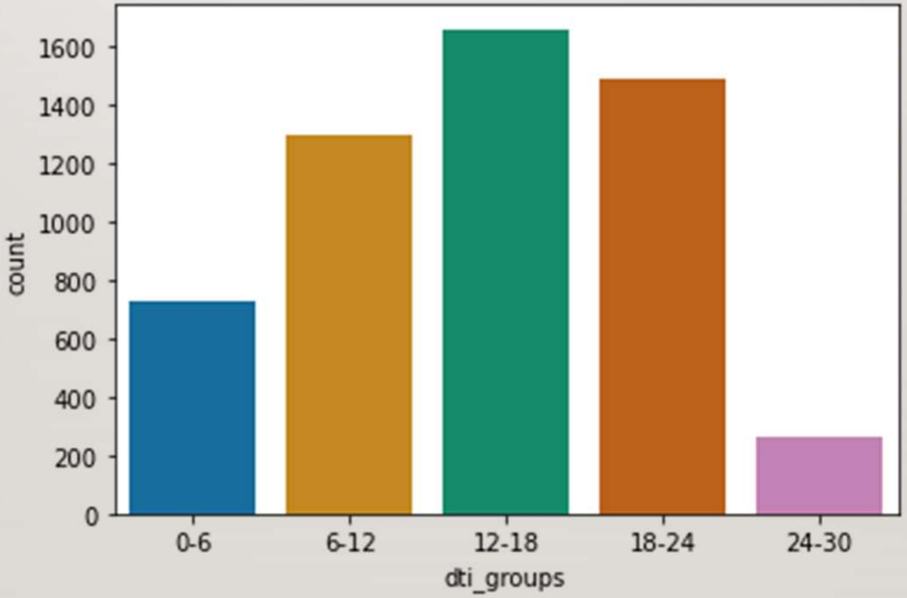
ANALYSIS OF DEFAULTS BASED ON FUNDED AMOUNT



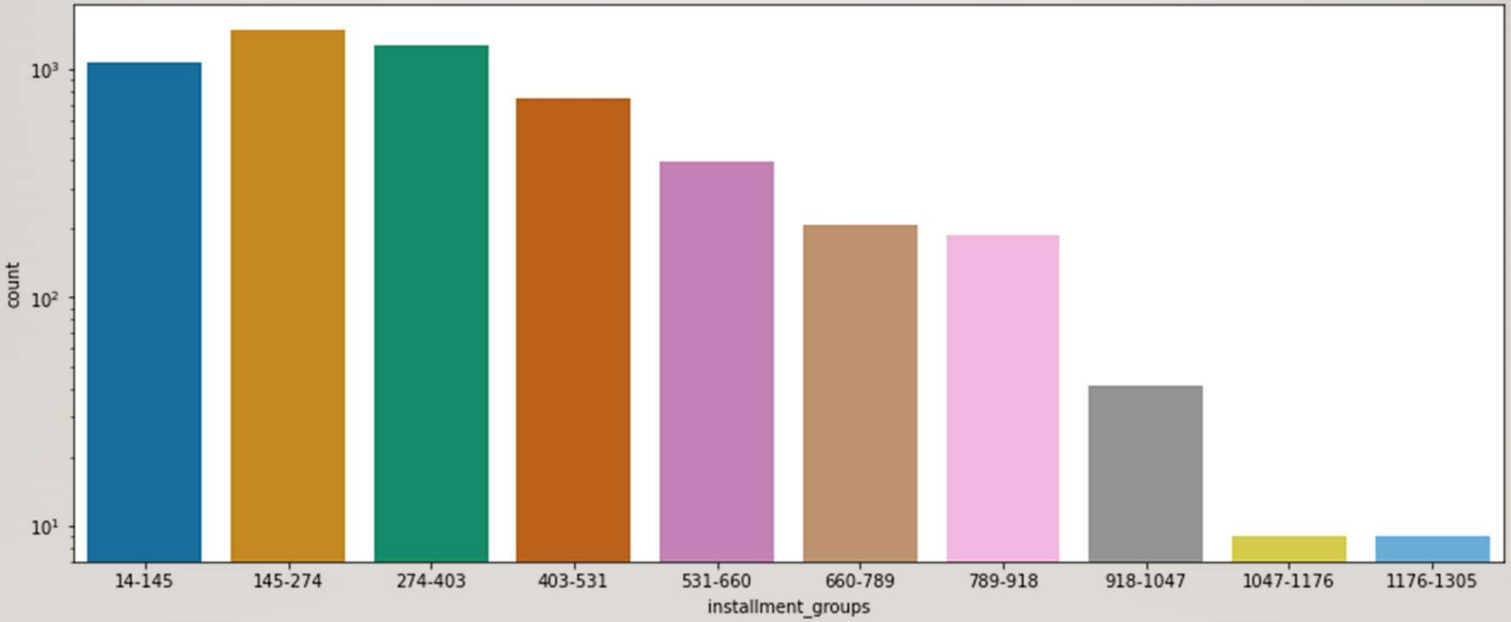
ANALYSIS OF DEFAULTS BASED ON LOAN AMOUNTS



ANALYSIS OF DEFAULTS BASED ON DTI GROUPS



ANALYSIS OF DEFAULTS BASED ON INSTALLMENTS



MAJOR OBSERVATIONS FROM OUR ANALYSIS

- MAJOR OBSERVATIONS Based on the various visualizations and plots above, we make the following conclusions which might be useful to predict which persons are more likely to default. Although correlation necessarily does not mean causation, here are some observations and likely reasoning in my judgment.

1. Applicants living in Rented house are more likely to default than those having their own house. This maybe because those having their own house fear their house being mortgaged and attached for recovery.
2. Applicants whose main purpose of taking loan is to clear other loans or consolidate their debts are more likely to default. This can be due to poor financial management techniques used.
3. Applicants receiving interest at the rate of 13%-17% are more likely to default. This is the observation from the data. Correlation doesn't indicate causation, so there is no particular cause as such.
4. Where installments are between 145-274, the chances of default are higher.

MAJOR OBSERVATIONS FROM OUR ANALYSIS

5. Applicants with DTi between 12-18 are more likely to make a default.

6. Where funded amt or loan amt is between 5000-10000, there is more likelihood of default.

7. Maximum defaults occur where verification status is 'Not verified'. Reason is quite obvious as it may have a higher level of fraudulent/ fake applications.

8. When the number of derogatory public records is 0, there the default is maximum.

9. When the no of enquiries in last 6 months is 0, there the default is maximum.

10. Applicants with income between 31T-58T are most likely to default. No particular cause identified.

11. Grade B and FINAL B5 grade applicants are the highest defaulters according to this data.

WHAT OTHER MODEL CAN BE USED?

- Based on the data and advanced regression techniques, we can build a model that indicates which factors are most likely to cause a default
- We can also test the significance of each of the categorical variables and drop off those explanatory variables that are not significant.
- This will help us simplify the model