LENDING CLUB CASE STUDY

Chaitanya Kuchimanchi

AGENDA

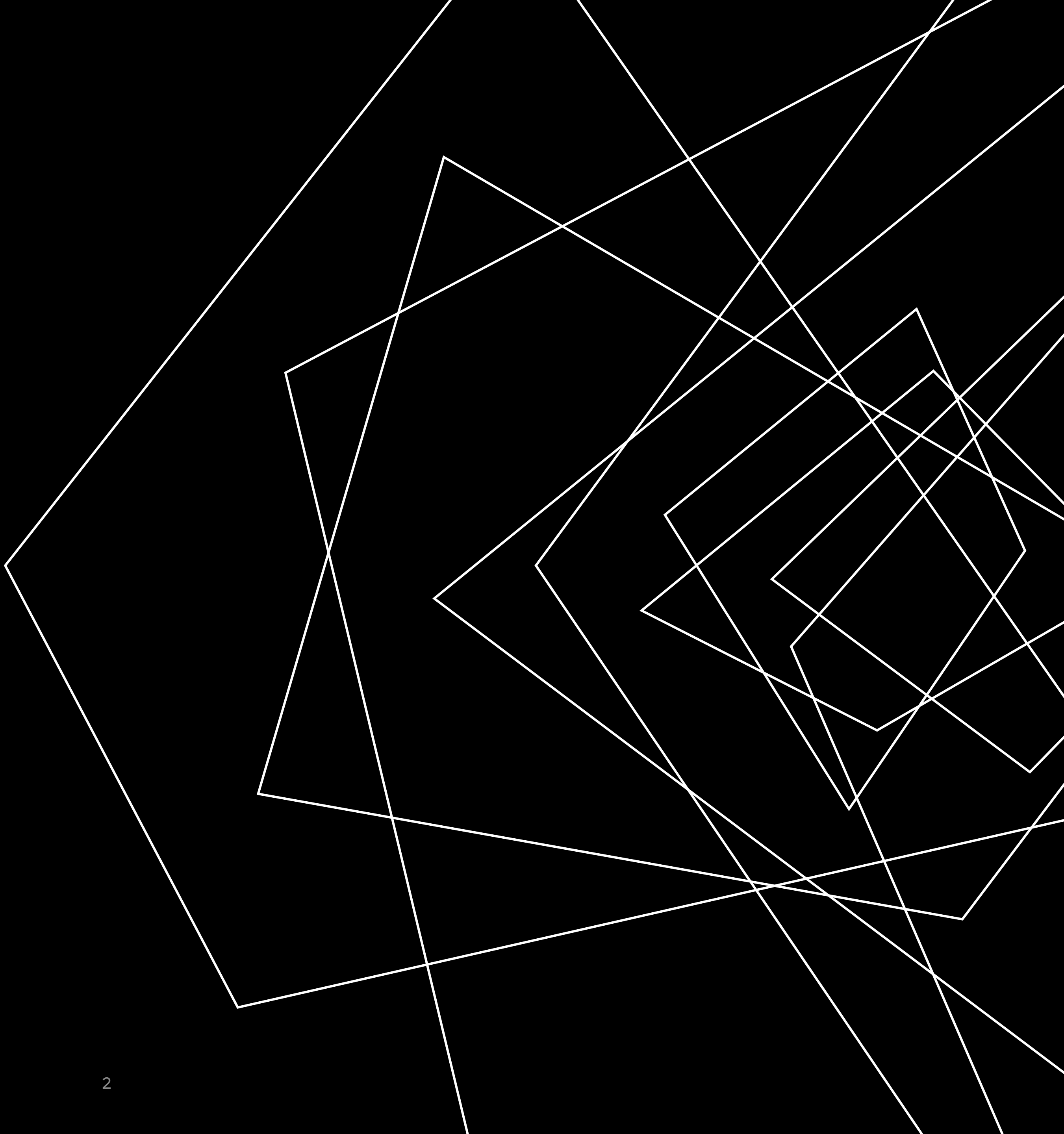
Problem Statement

Primary goals

Areas of growth

Timeline

Summary

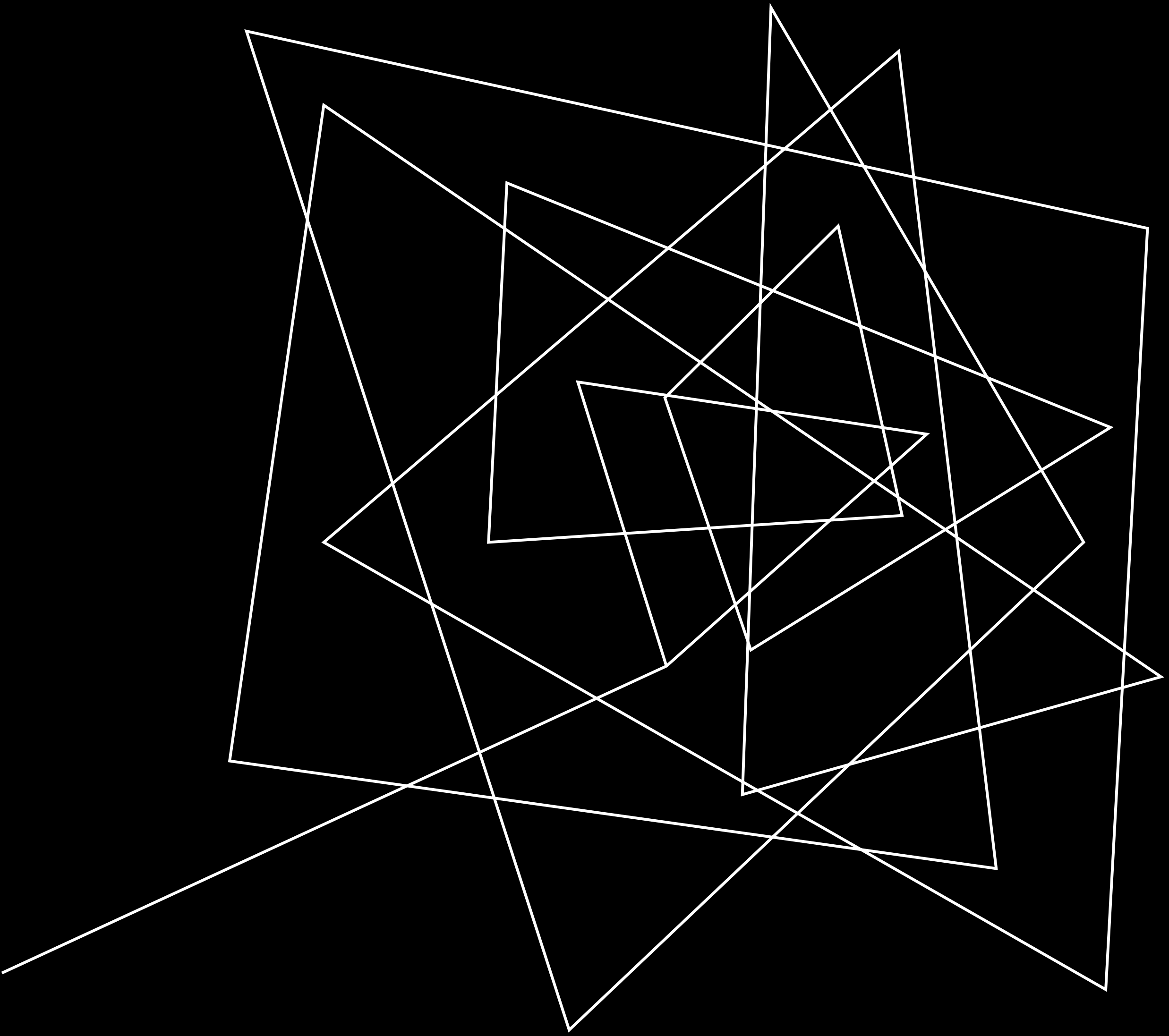


PROBLEM STATEMENT

The largest online loan marketplace provides a platform for borrowers to obtain personal, business, and medical procedure financing with lower interest rates through a quick and efficient online interface.

Lending loans to applicants with a higher risk of default is a significant financial loss for most lending companies, known as credit loss. The borrowers who default, also called 'charged-off' customers, cause the most significant losses to lenders. This case study aims to identify such risky loan applicants through EDA, leading to a reduction in credit loss.

In summary, the company seeks to determine the key factors or driver variables responsible for loan defaults. Understanding these factors would enable the company to assess portfolio and risk better.



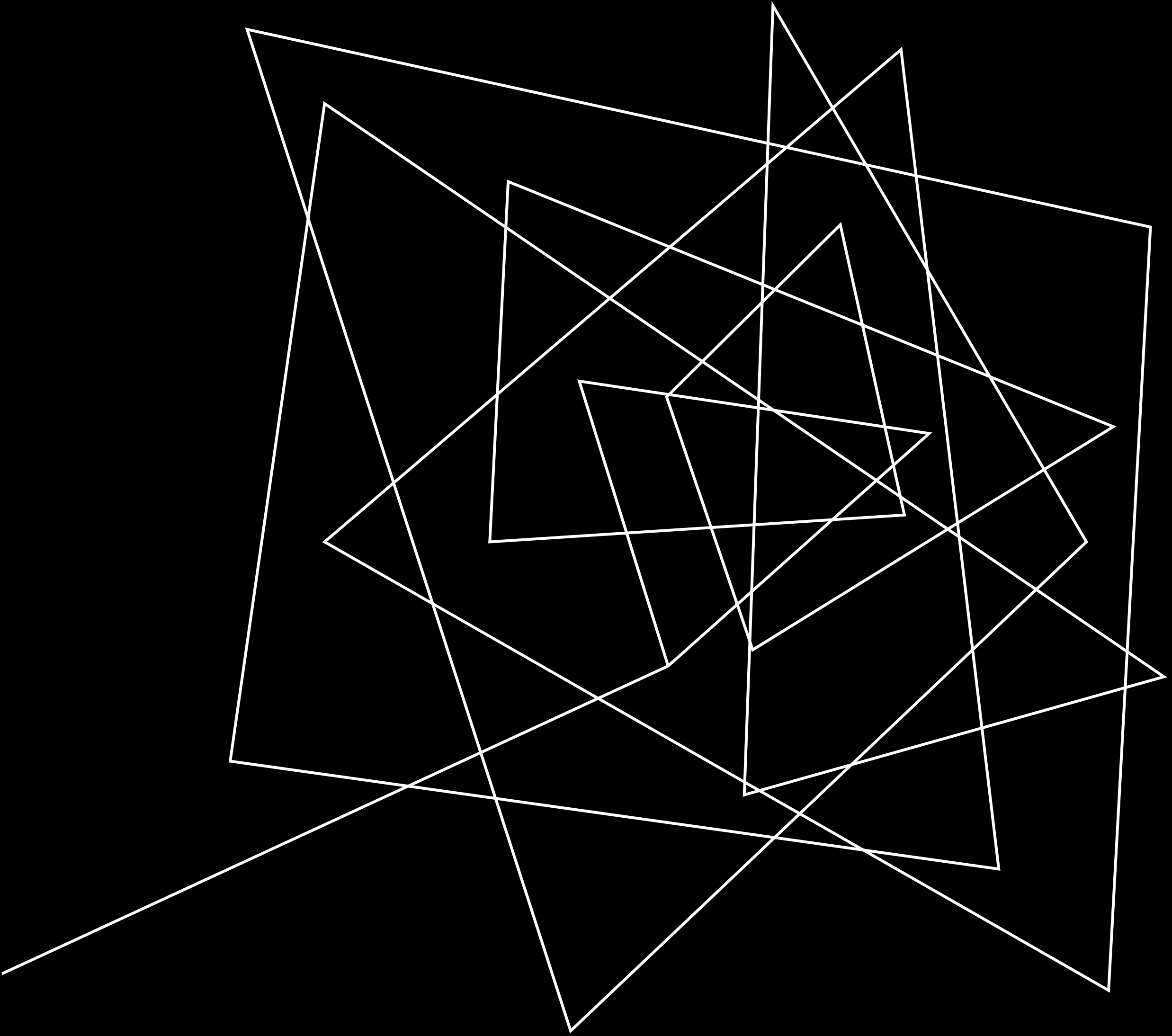
DATA UNDERSTANING

1. The data file consists of 39717 rows and 111 columns
2. There is no header or footer present in the data.
3. There are few columns with null values and few columns with inappropriate data types.
4. For better understanding segregated variables into 2 buckets
5. **Consumer_variables=** ['emp_length', 'home_ownership', 'annual_inc', 'addr_state’, 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'mths_since_last_delinq','open_acc', 'pub_rec’, 'revol_bal', 'revol_util', 'total_acc', 'last_credit_pull_d', 'pub_rec_bankruptcies’]
6. **loan_variables=** ['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries’, 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt']

1. The data file consists of 39717 rows and 111 columns
2. There is no header or footer present in the data.
3. There are few columns with null values and few columns with inappropriate data types.
4. For better understanding segregated variables into 2 buckets
5. **Consumer_variables=** ['emp_length', 'home_ownership', 'annual_inc', 'addr_state’, 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'mths_since_last_delinq','open_acc', 'pub_rec’, 'revol_bal', 'revol_util', 'total_acc', 'last_credit_pull_d', 'pub_rec_bankruptcies’]
6. **loan_variables=** ['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries’, 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt']

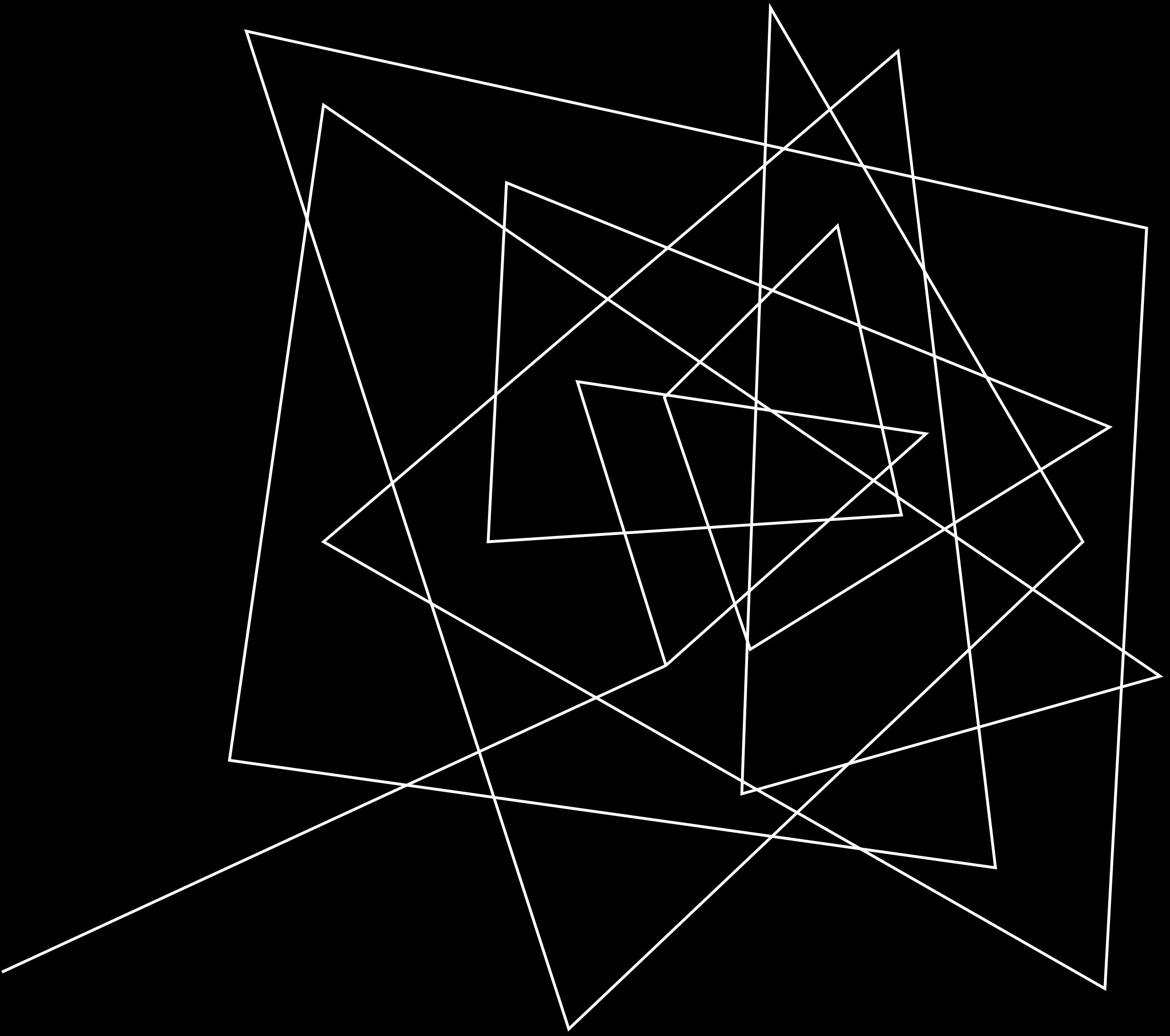
Identifying important variables

1. Identify important variables that are needed to find patterns for defaulters.
2. imp_variables= ['emp_length’, 'home_ownership’, 'annual_inc’, 'addr_state’, 'dti', 'pub_rec', 'total_acc’, 'pub_rec_bankruptcies’, 'loan_amnt’, 'funded_amnt’, 'funded_amnt_inv’, 'term', 'int_rate','grade', 'sub_grade','verification_status', 'issue_d’, 'loan_status', 'purpose’]
3. Now the data is filtered to have 39717 rows and 19 columns



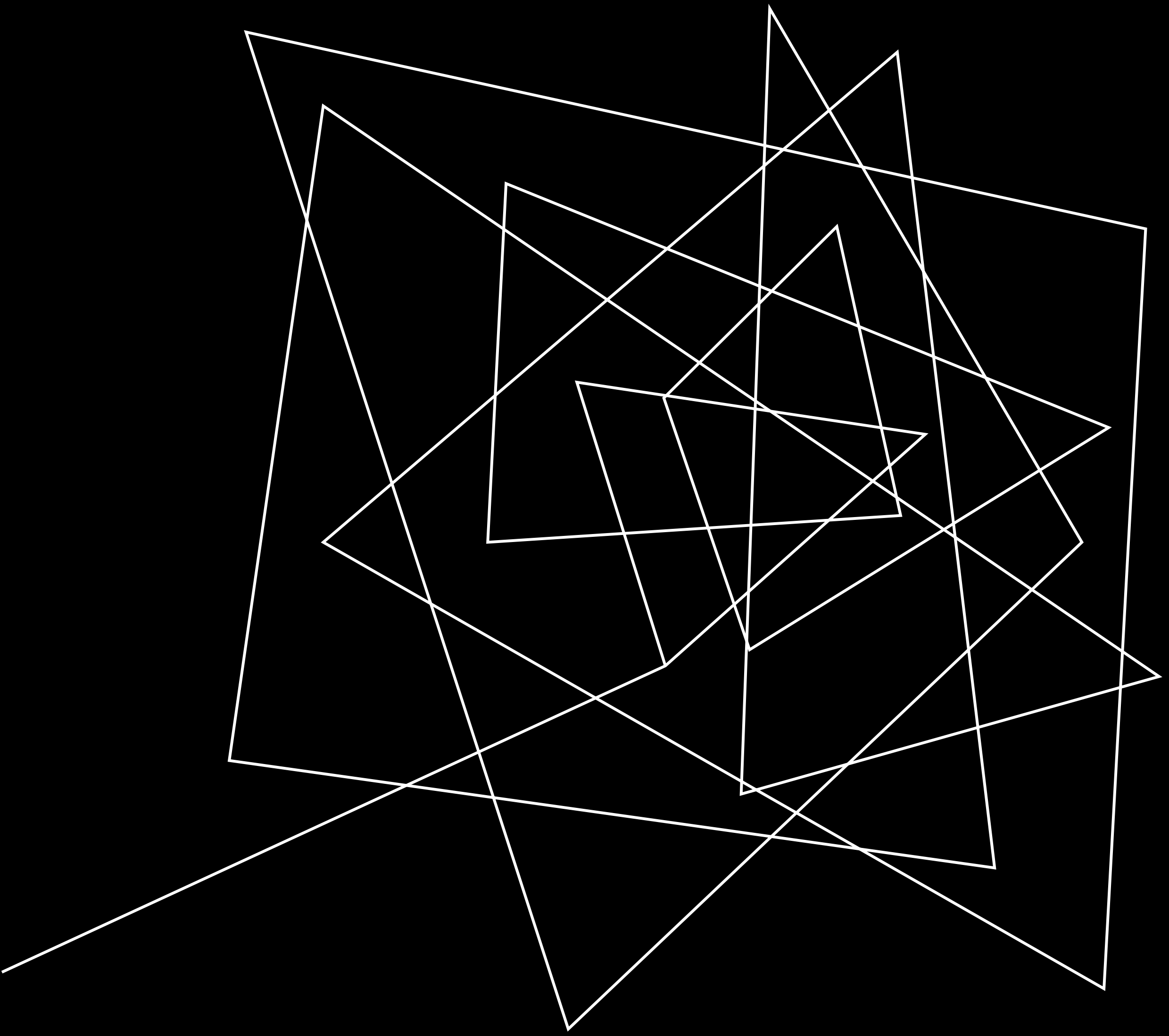
FIXING COLUMNS

1. Converted the below columns from Object to specific data type.
    1. **Int_rate → Int32**
    2. **Term → Int32**
    3. **Emp_length → Int 32**
    4. **Issue_d → month and year (2 separate columns)**
2. Separated issue_d (data format) column into 2 separate columns that contains Month and year respectively.
3. Now the data frame is modified to have 20 columns



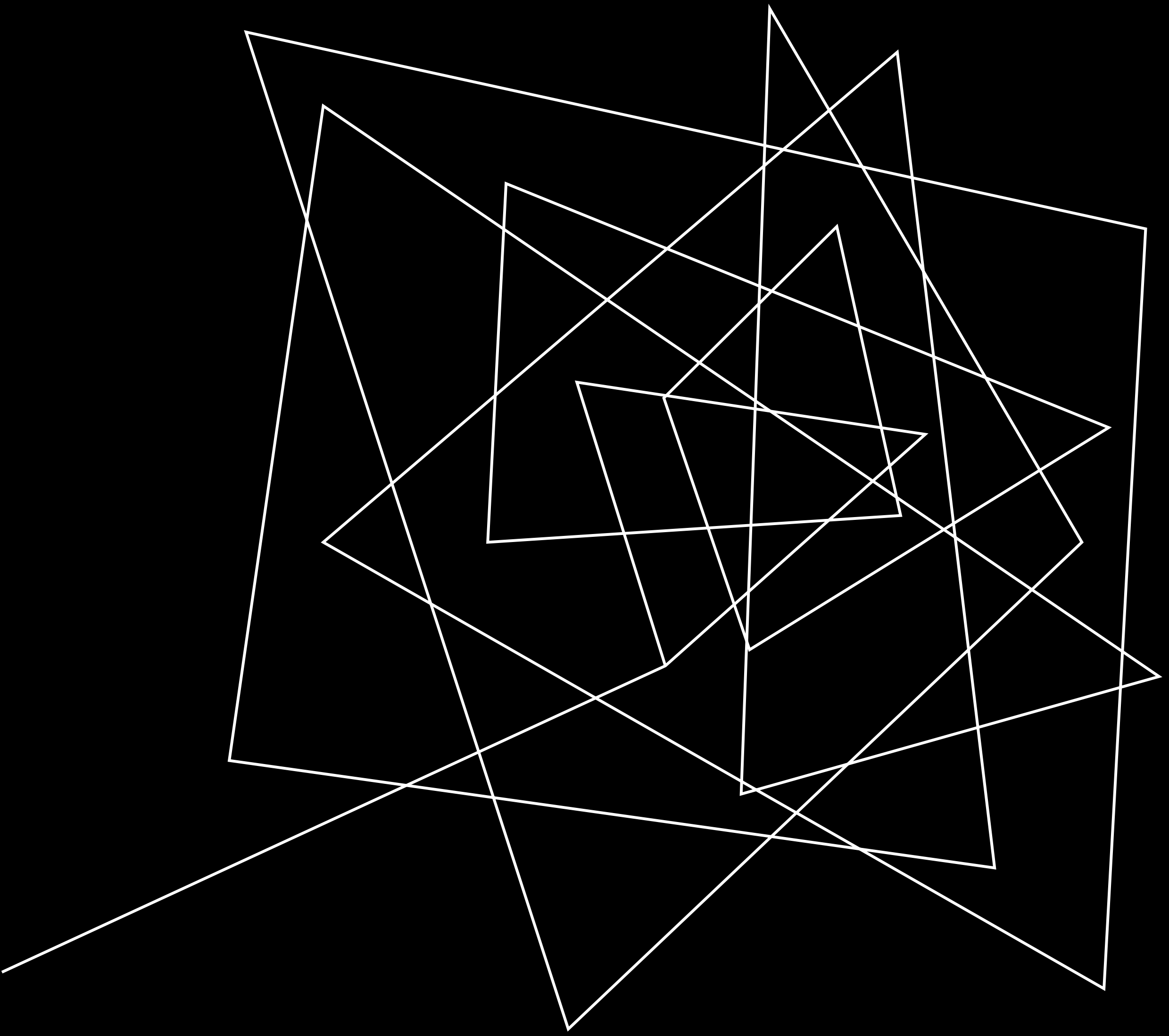
DATA IMPUTATION

1. Drop all the columns that have null values.
2. Drop all the rows that have null values and also drop duplicate rows.
3. Emp_length and pub_rec_bankruptcies are the two columns that have 1075 and 697 respectively.
4. Perform data imputation on emp_length
    1. **Check loan_amnt and annual_inc for all the records that have emp_length NULL.**
    2. **Taking the 75% percentile of both loan_amnt and annual_inc.**
    3. **Filter the data that has emp_length as not Null with the 75% percentile value taken in above steps and identify the most occurring emp_length.**
    4. **It is identified that 0 years to 5 years are repeated in 70% of the data.**
    5. **Now randomly replace all the null rows of emp_length with 0 – 5.**
5. Replace pub_rec_bankruptcies with most occurring value, 95% of the data has 0 in its records. Replace null values with 0.



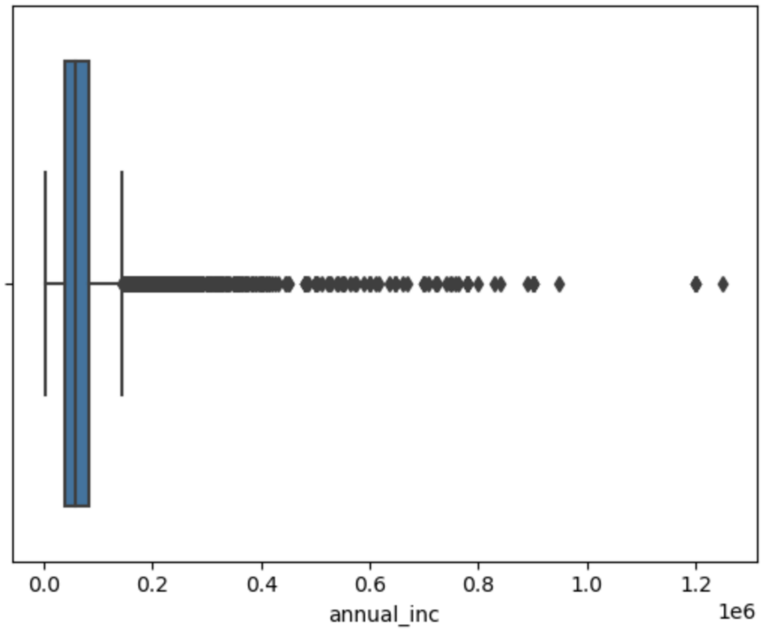
SANITY CHECKS

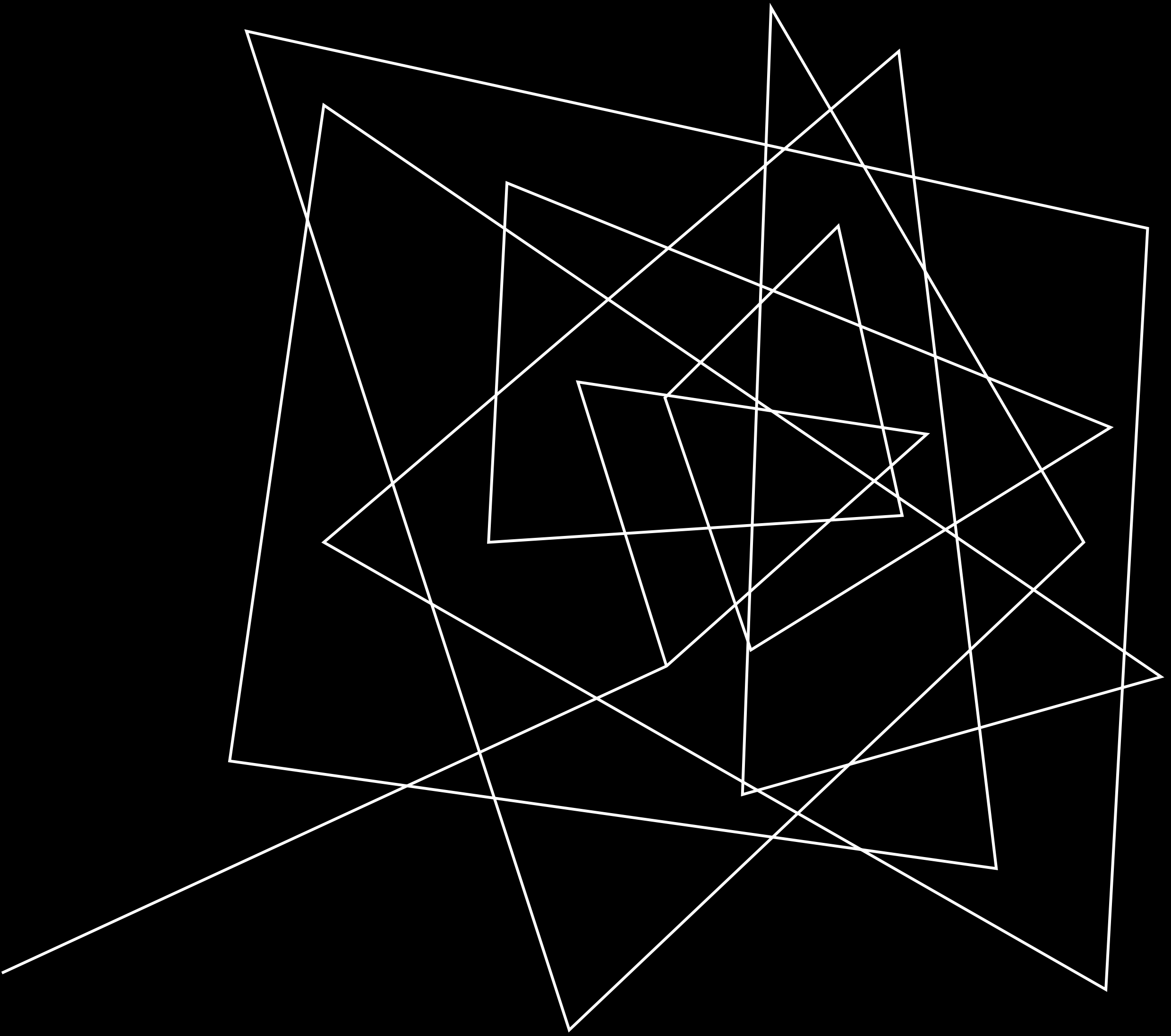
1. Check the data if that has any rows or columns that needs to be still addressed.
2. Check below data
    1. **Max and min values of emp_length should be between 0 and 10.**
    2. **Months column should have values between 1 and 12.**
    3. **Data in funded_amnt should be less than funded_amnt_inv**



IDENTIFYING OUTLIERS AND FIXING THEM

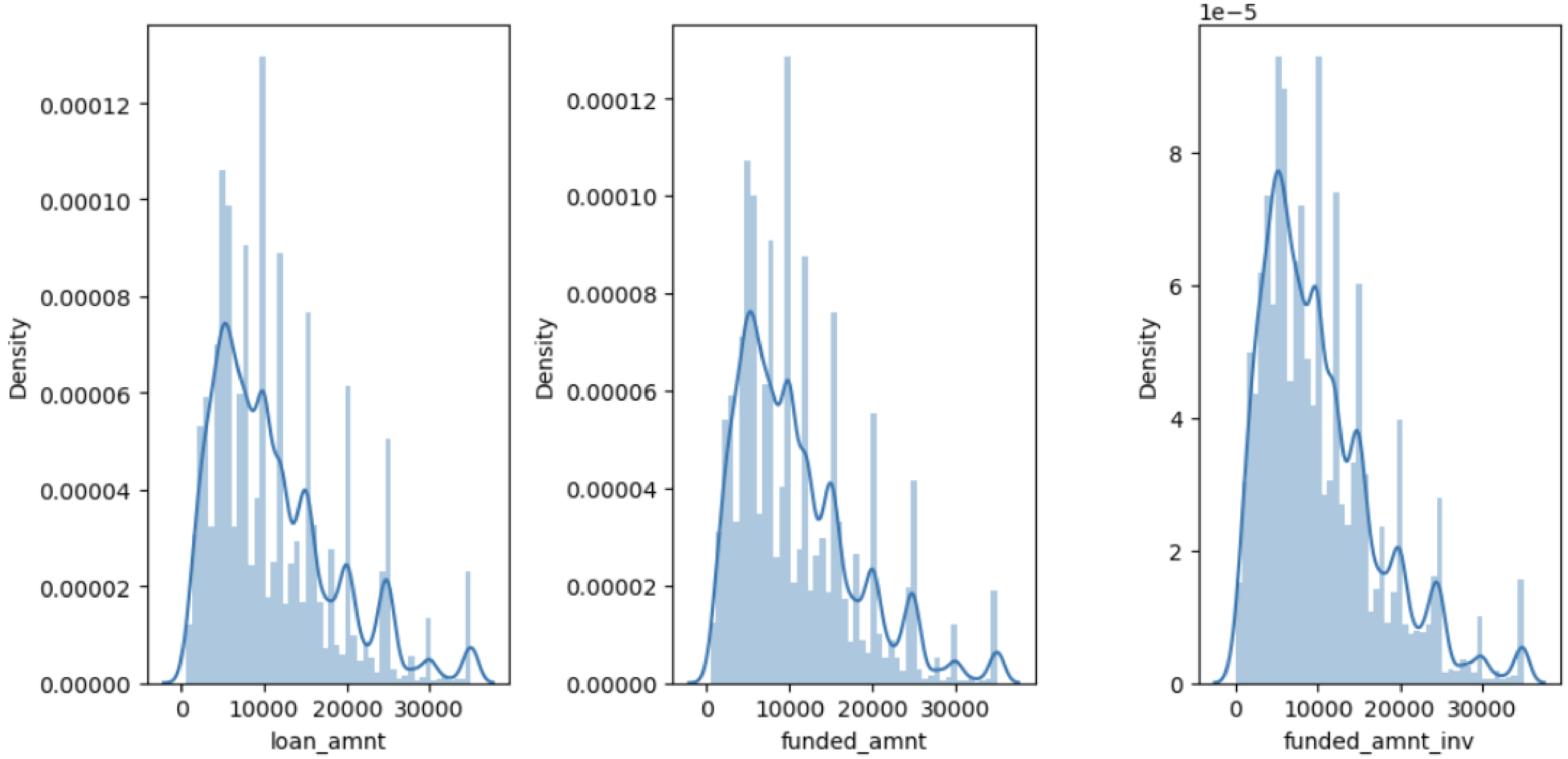
1. It is observed that annual_inc has few outliers. These outliers are observed by drawing a box plot.
2. Remove the outliers of annual_inc, it is observed that there are 7 records that has annual_inc > 1000000 and loan_amnt < 10000.
3. It is also further observed that annual_inc > 1500000 has outlier values. Filter the data to have annual_inc less than 1500000





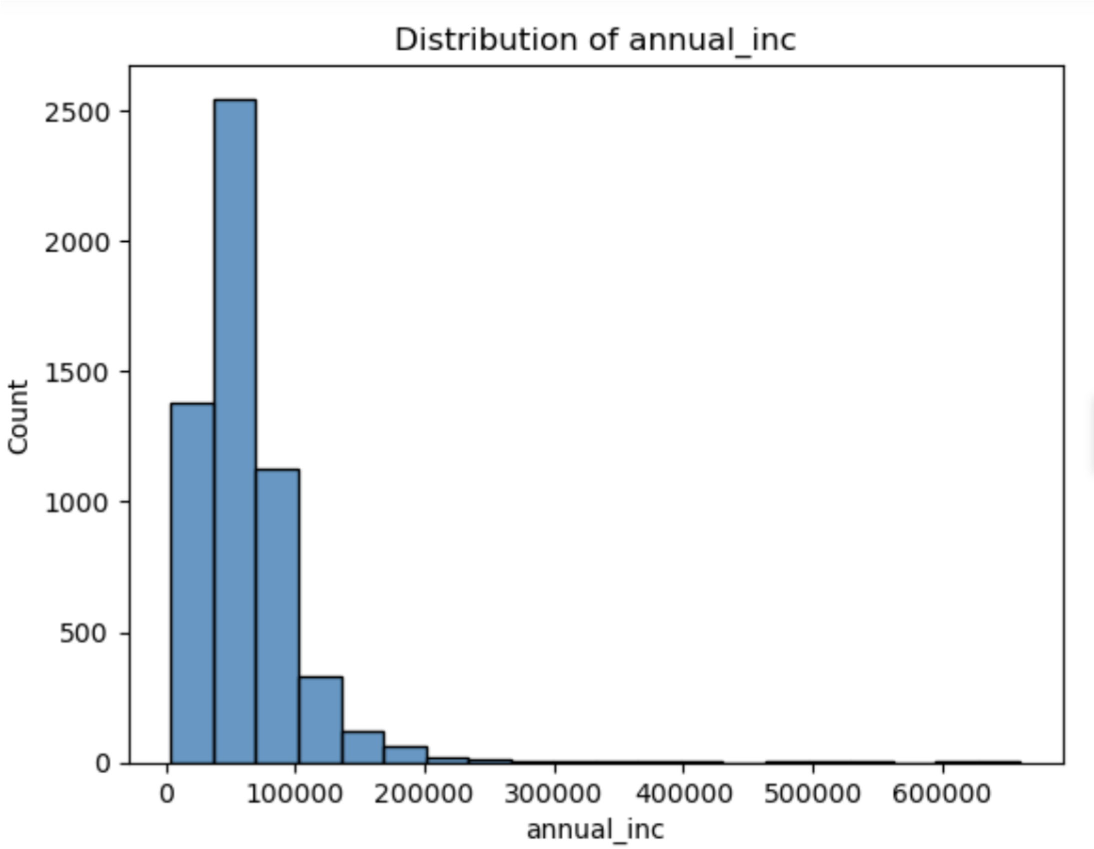
UNIVARIATE ANALYSIS

1. After looking at the columns loan_amnt, funded_amnt and funded_amnt_inv, it is observed that they all have the data distributed in a same way.
2. This can be identified by drawing a distplot.
3. For that reason we can include loan_amnt and eliminate other 2 columns.

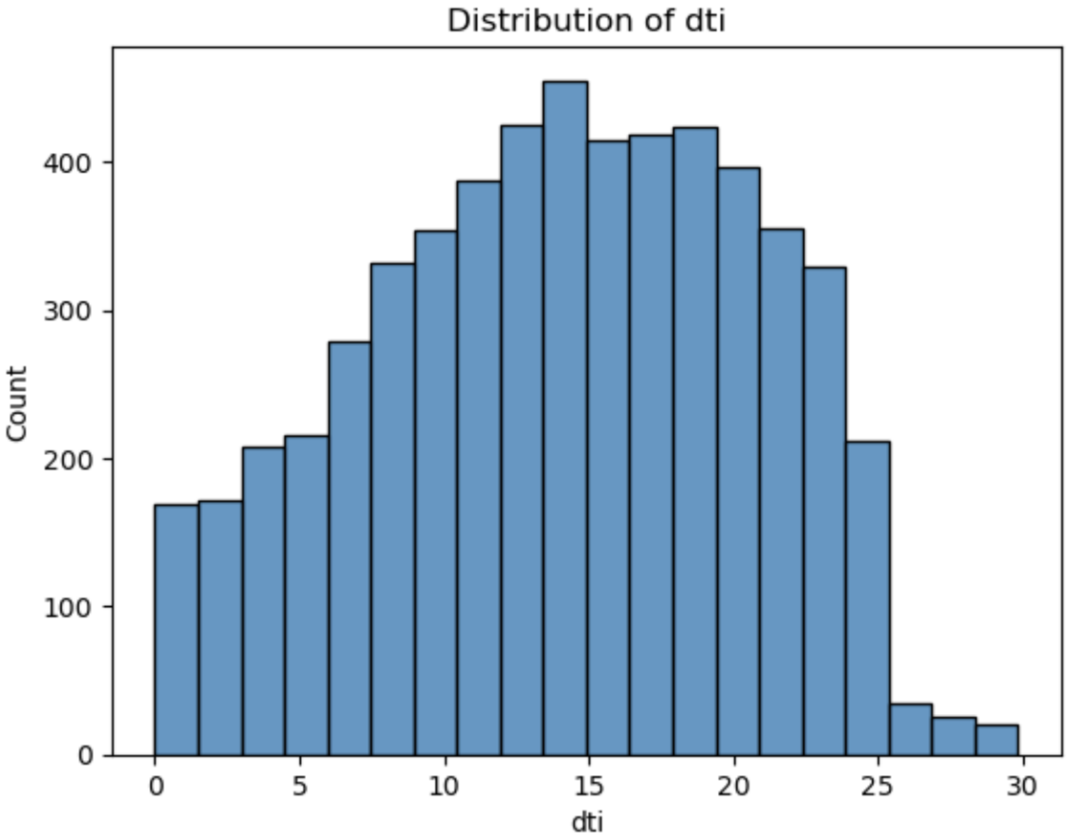


1. Categorizing the variables into below types
    1. Continuous variables = ['annual_inc', 'dti', 'total_acc', 'loan_amnt', 'int_rate']
    2. Ordered_categorical_variables = ['term', 'emp_length', 'pub_rec', 'pub_rec_bankruptcies', 'month', 'year']
    3. Unordered_categorical_variable = ['home_ownership', 'grade', 'sub_grade', 'verification_status', 'loan_status', 'addr_state', 'purpose']
2. Removing the rows that has loan_status = Active and only having the loan status that have values 'Fully Paid' and "Charged Off".
3. Further splitting the complete data set into 2 data sets that have loan_status "Fully Paid" and "Charged Off" as independent data frames.

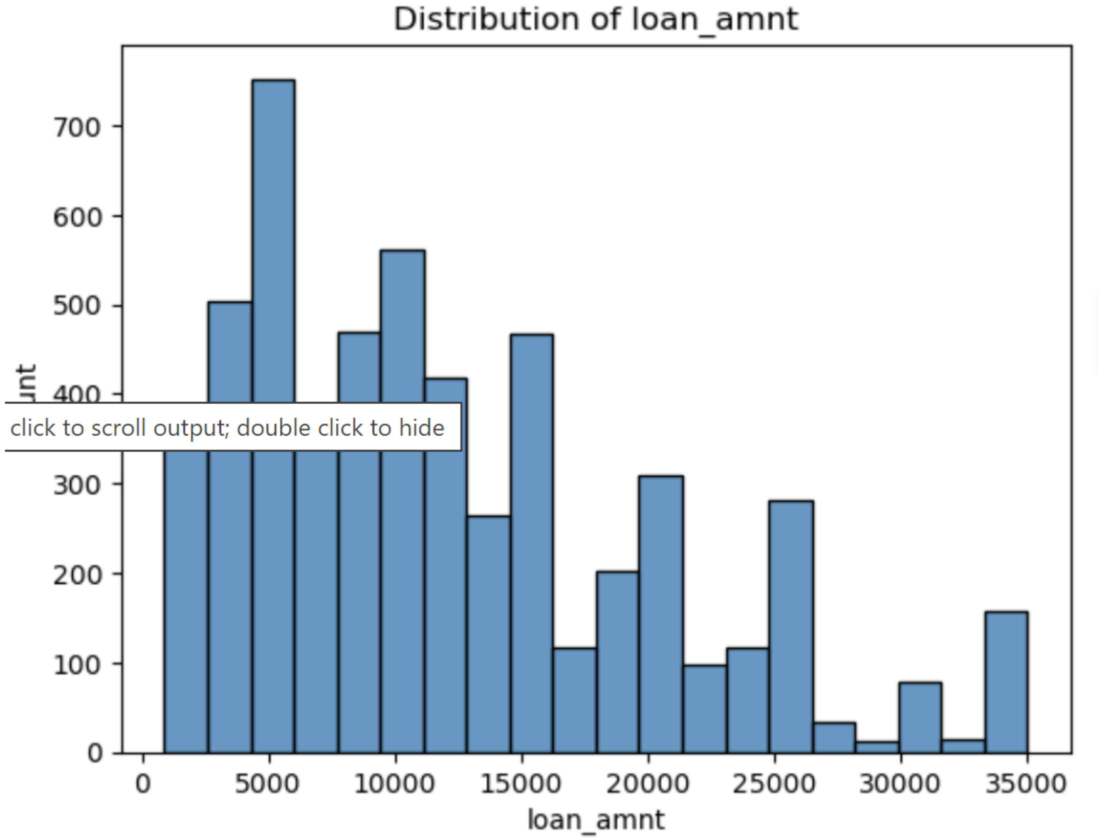
1. **Understanding distribution of annual_inc of defaulters**
2. **It is observed that the annual inc of the defaulters lies between 0 and 200000**

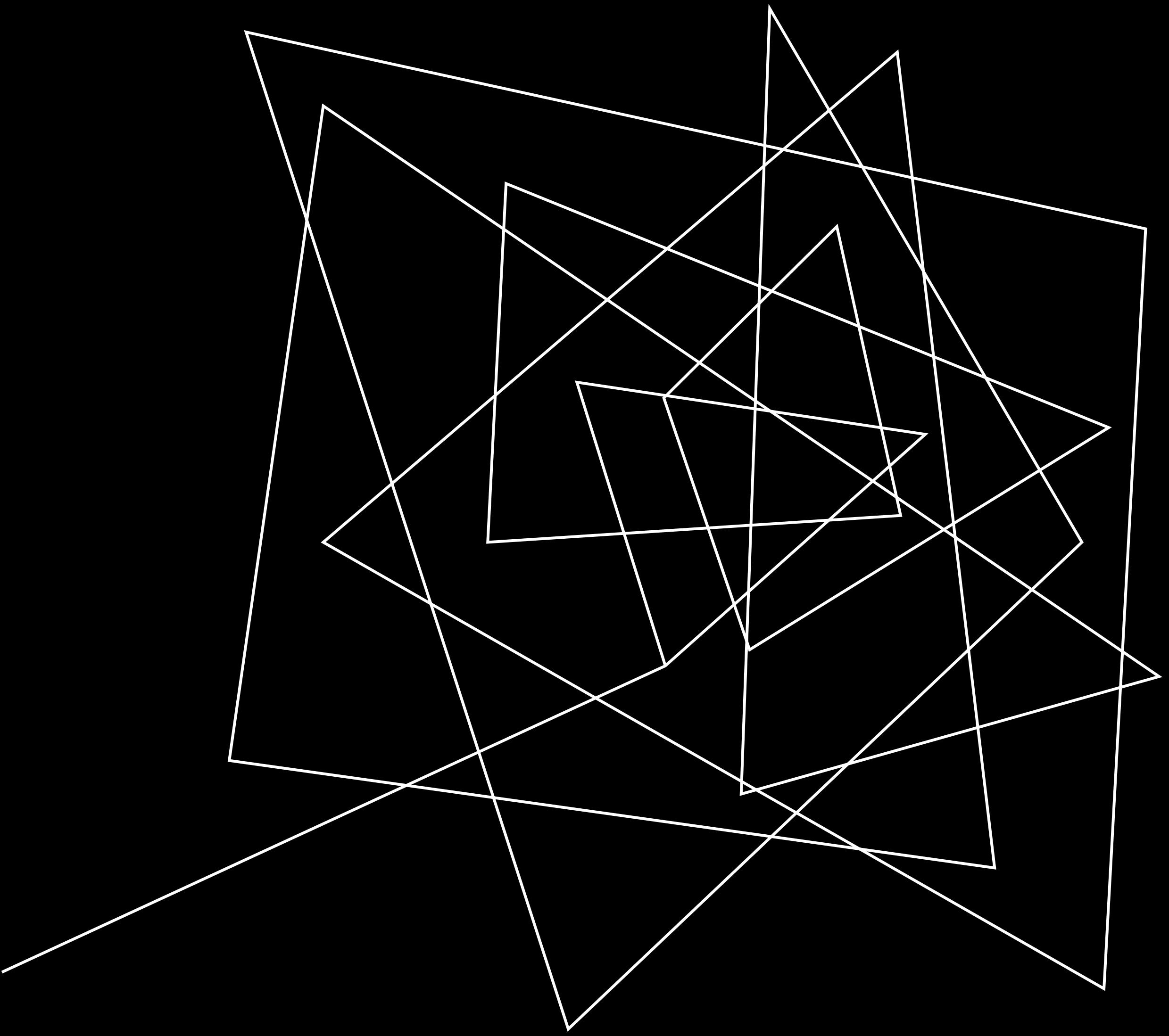


1. **Dti that lies beyond 25 value has very less number of defaulters.**



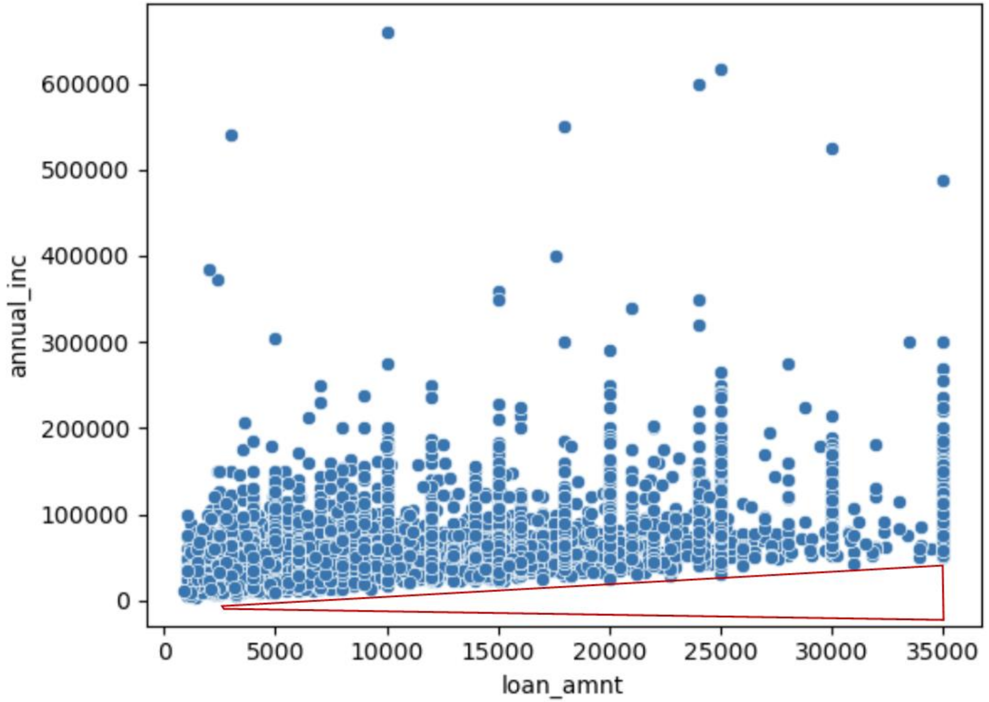
1. **It is observed that loan amount has spikes at round figures like 5000, 10000, 15000, 20000, 250000, 30000, 35000**

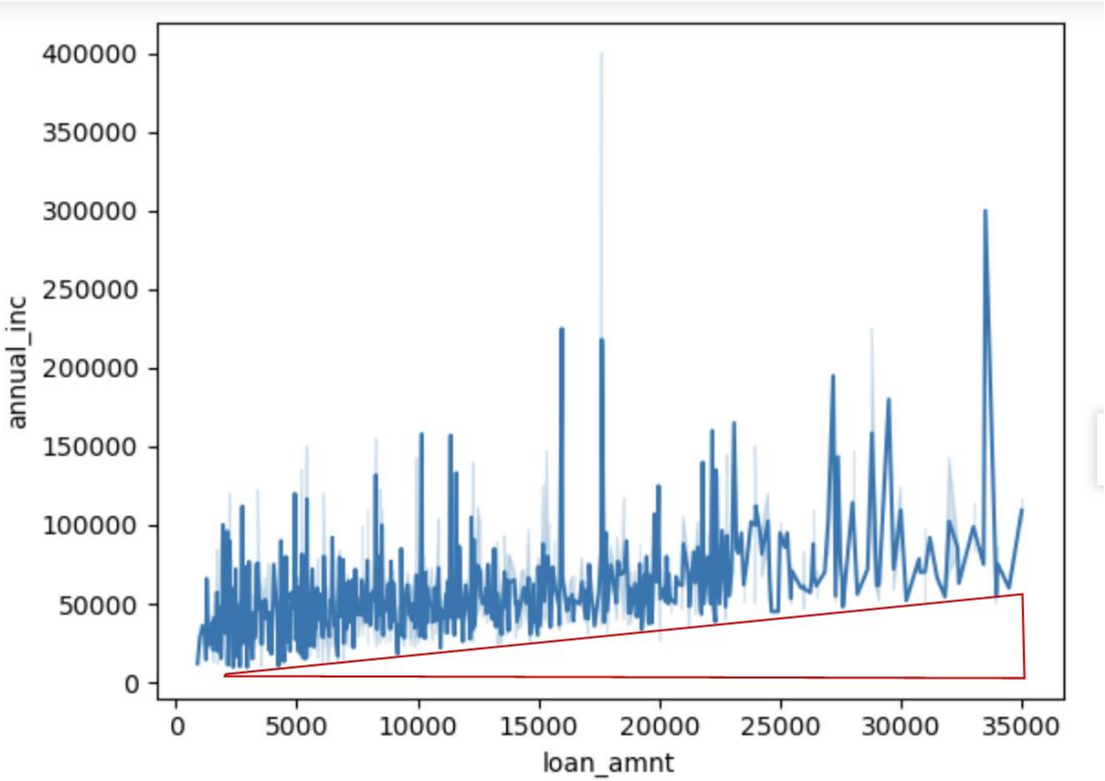




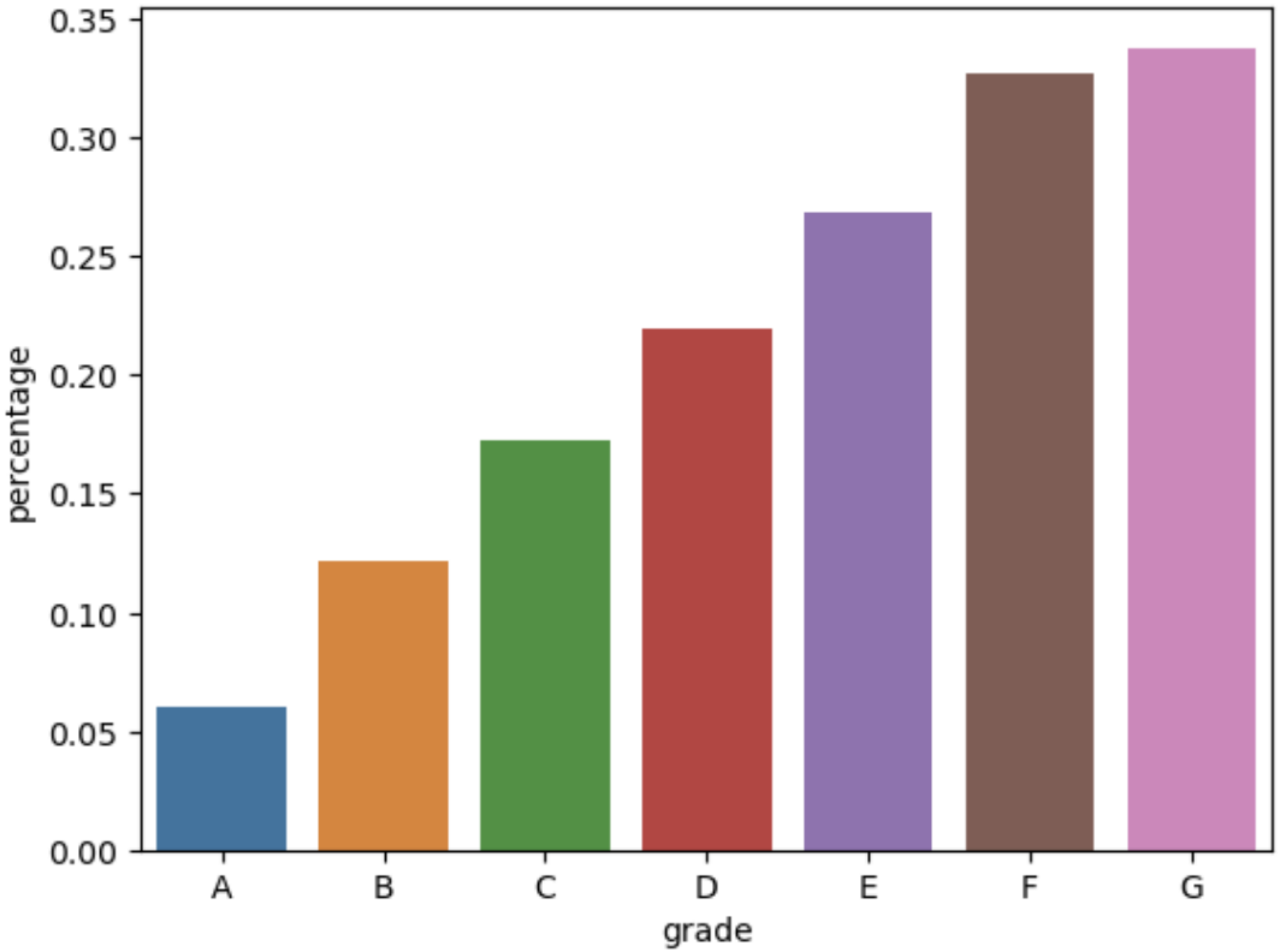
BIVARIATE ANALYSIS

1. **When checked loan amount against annual inc, it is observed that consumers with higher salary and with less loan amount has less defaulters.**

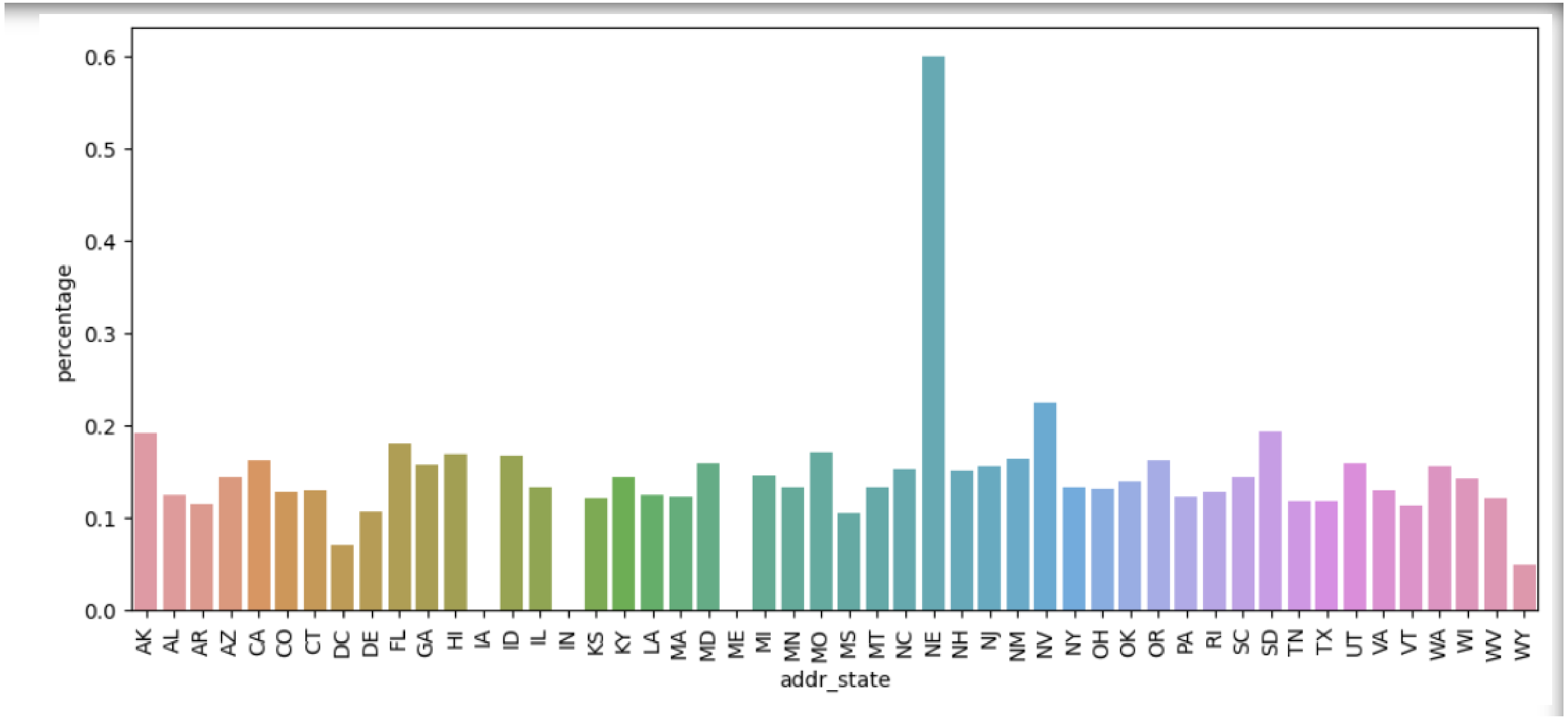




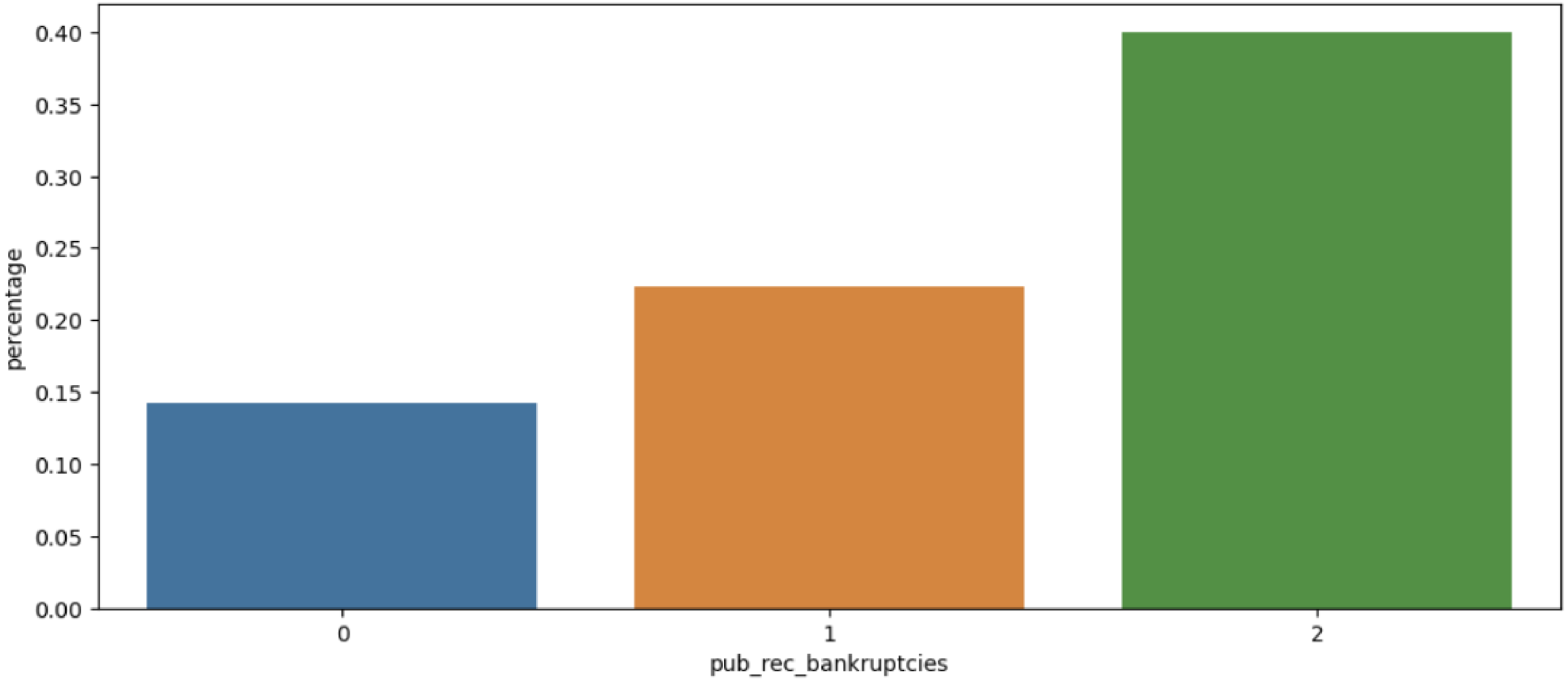
1. **When checked grade against the defaulter percentage, it is observed that grades with G, F and E has higher defualters.**

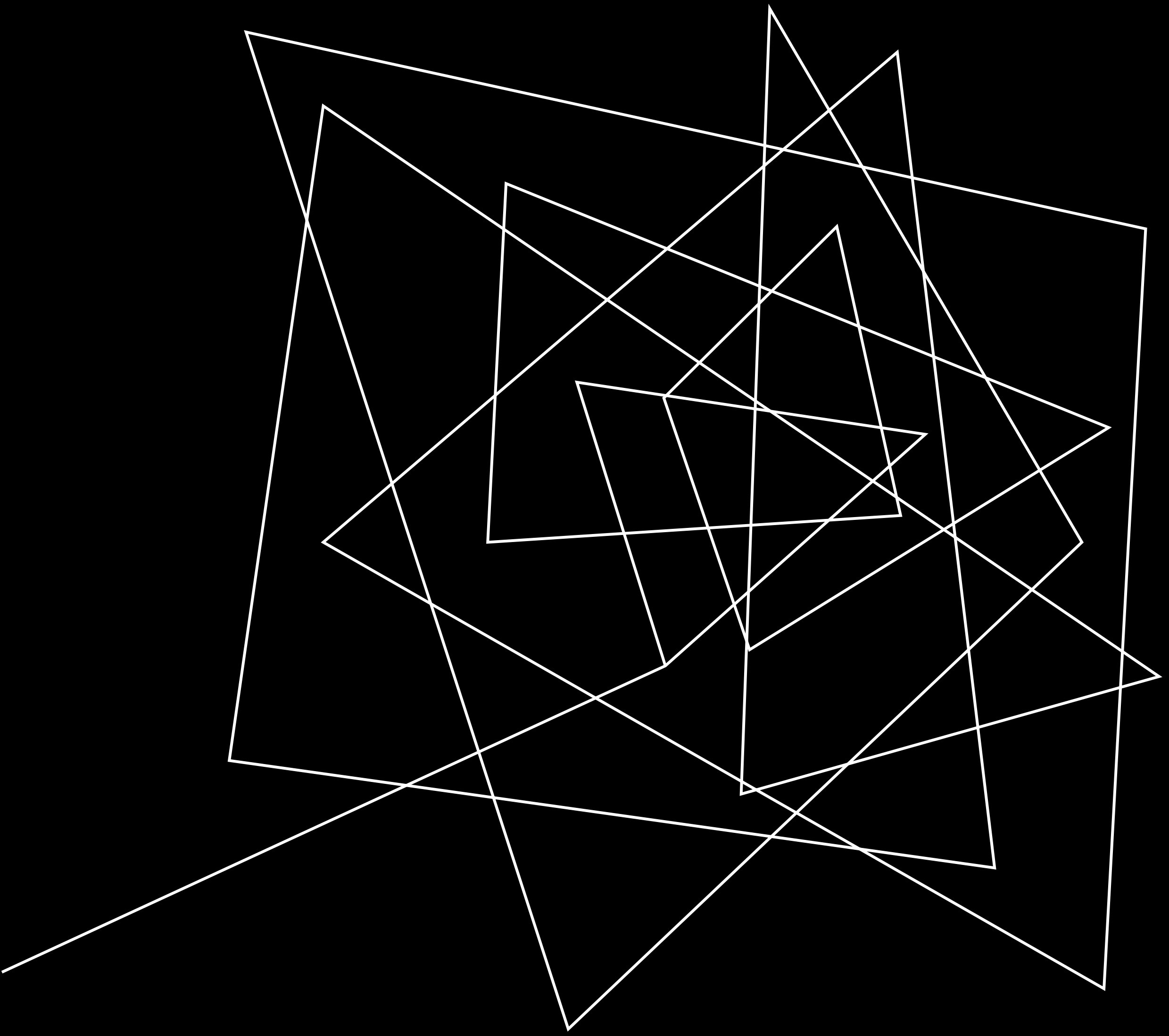


1. **When checked the states against the defaulter percentage, it is observed that NE, SD, NV states has higher defaulters.**



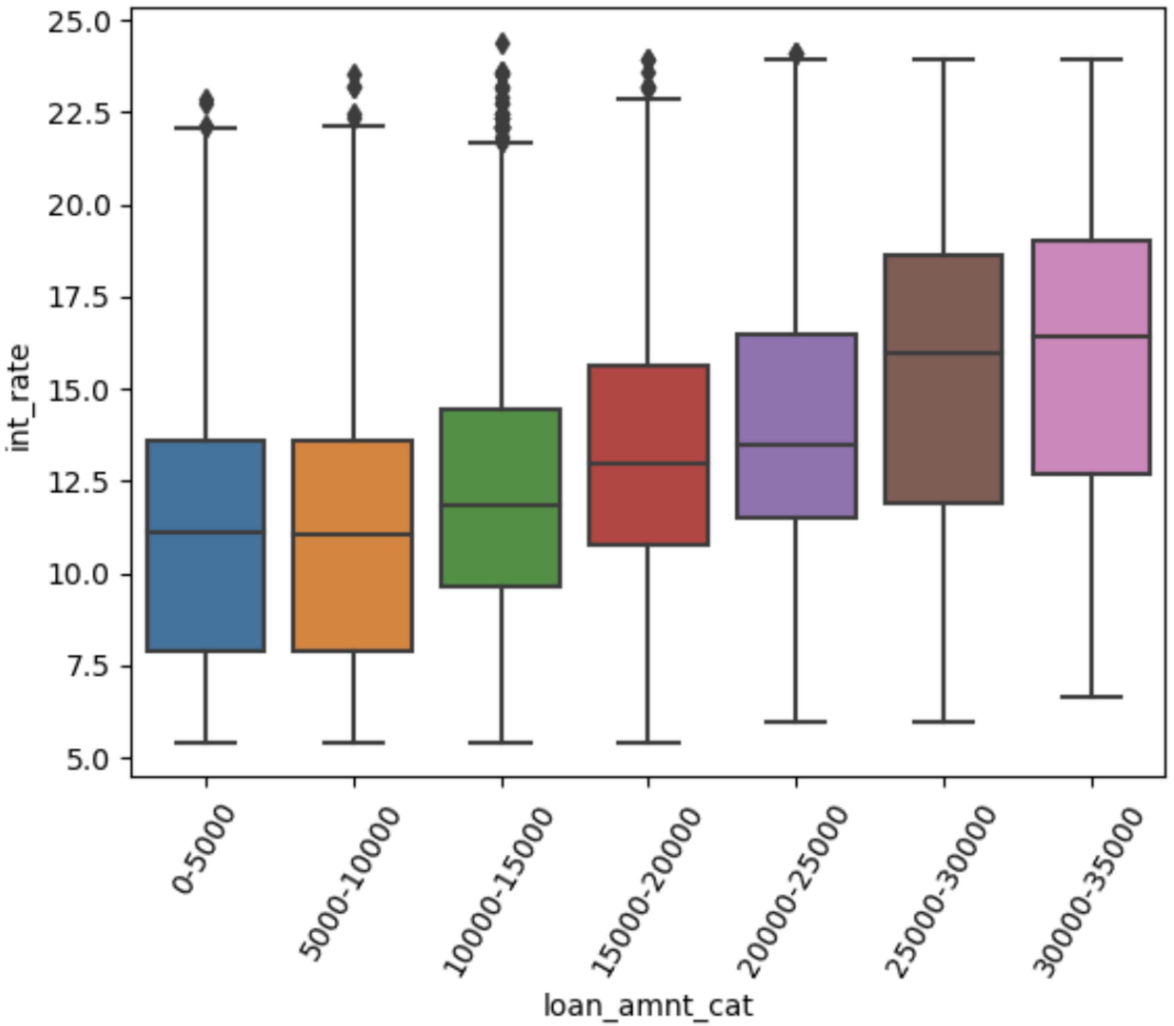
1. **When checked the pub_rec_bankruptcies against the defaulter percentage, it is observed that 2 and 1 have the highest defaulters.**





DERIVED METRICS

1. **When checked the loan amount against the interest rate and bucketing loan amount in 7 bins, it is observed persons with higher loan amount has high default rate.**



THANK YOU

Chaitanya Kuchimanchi