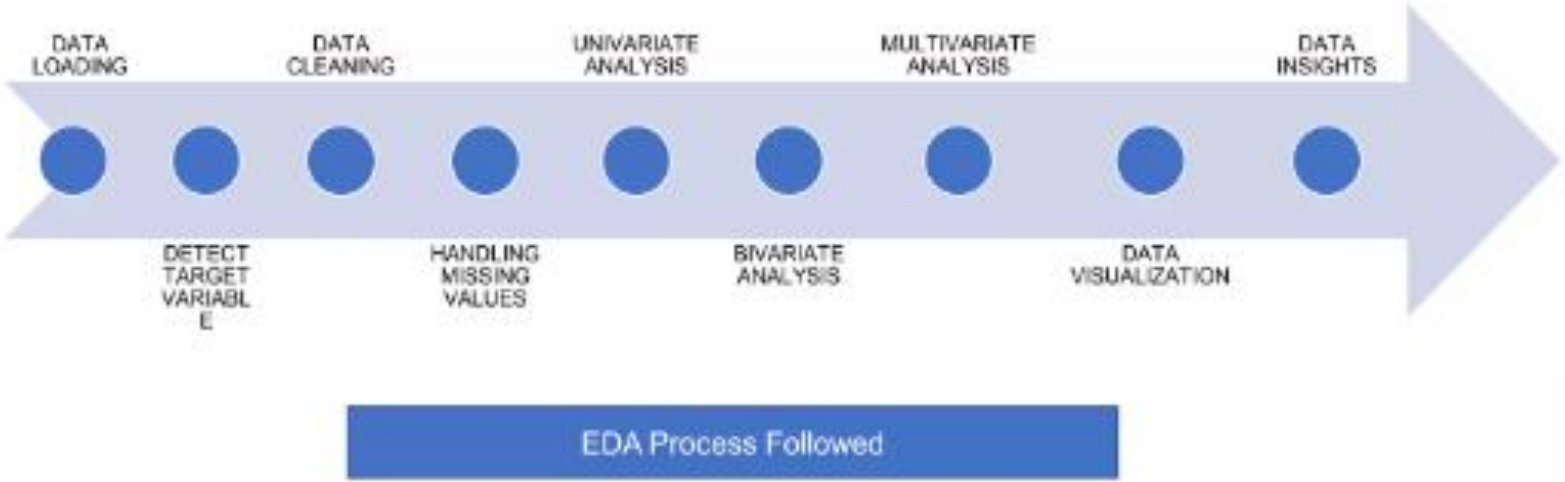
What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is to analyse the patterns present in the data which in turn with respect to data set (loan) will make sure that the loans are not rejected for the applicants capable of repaying.Hence to provide business insights to the business.

➢ FLOW CHART


DATA LOADING
DETECT TARGET VARIABLE
DATA CLEANING
HANDLING MISSING VALUES
UNIVARIATE ANALYSIS
BIVARIATE ANALYSIS
MULTIVARIATE ANALYSIS
DATA VISUALIZATION
DATA INSIGHTS
EDA Process Followed

OBJECTIVE

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile

- 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 be a defaulter, then approving the loan may lead to a financial loss for the company.

Data Sourcing: Is the process of finding and loading the data into our system.

Classified in 2 categories

1) Public Data :

This type of Data is available to everyone. Eg:Public organisation (https://data.gov)

2) Private Data :

This data is given by private organizations. This is mainly used for organisation's internal analysis

1. ***Data Sourcing:*** Our Scope is Public Data: We have Loaded this data to our system :

   It contains the complete loan data for all loans issued through the time period 2007 t0 2011.

It contains 116 variables and 39718 observation for each loan list data during 2007 to 2011.

Importing Necessary Packages

- `import numpy as np`
- `import pandas as pd`
- `import matplotlib.pyplot as plt`
- `import seaborn as sns`

Important parameters from the data set

loan_status	Current status of the loan	
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.	
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.	
pub_rec_bankruptcies	Number of public record bankruptcies	No. of bankruptcy increases the chance of defaulters
tax_liens	Number of tax liens	These certificate are created when local govt. places liens on people's property due to unpaid property taxes
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	

loan_amt	Loan amount applied by the borrower
Interest rate	Interest rate in the loan
Grade	LC assigned loan grade
Sub grade	LC assigned loan sub grade
Annual Income	Income provided by the borrower during registration
Purpose of loan	Category provided by the borrower for the loan request
Emp_length	Employer length in years
Loan_date	
Home ownership	It indicates if the loan applicant has own house, rent, mortgage, other
Verification status	Indicates if income was verified or not

2. Data Cleaning:

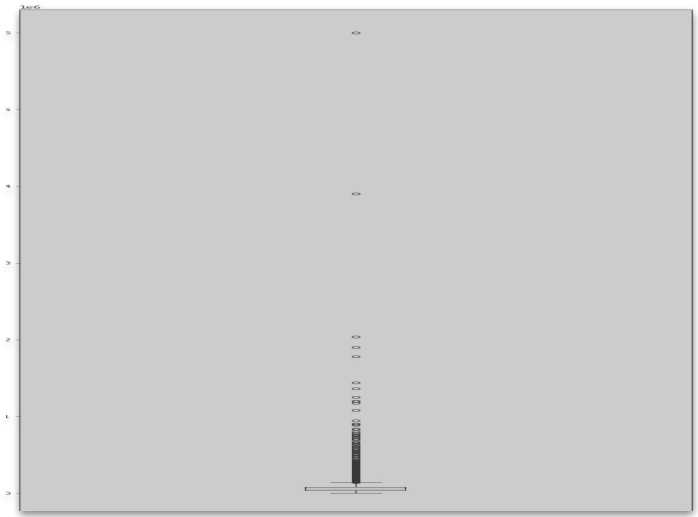
After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Below is the list of columns which are being deleted as it not not contributing much to the analysis:

('tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'mths_since_last_major_derog', 'emp_title', 'issue_d', 'pymnt_plan', 'url', 'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'earliest_cr_line', 'inq_last_6mths', 'mths_since_last_record', 'revol_util', 'initial_list_status', 'last_pymnt_d', 'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d', 'collections_12_mths_ex_med', 'policy_code', 'application_type', 'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens', 'delinq_2yrs', 'open_acc', 'pub_rec', 'revol_bal', 'total_acc', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee')

❖ ***Handling Outliers:*** *Outliers are the values that are far beyond the next nearest data points.*

This is a box plot of column 'annual_inc' to identify the outliers hence data of those are deleted from the data set



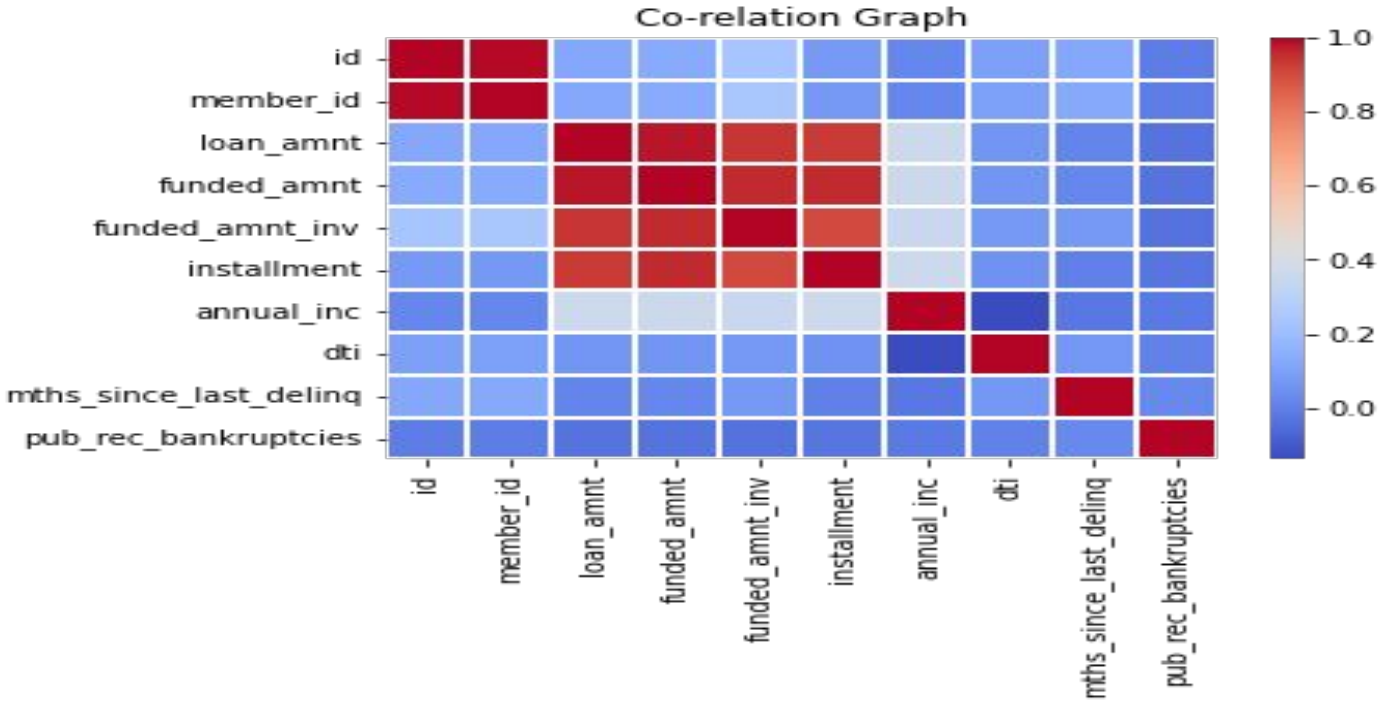
Removing the variable “current loan” from the data set

data[data['loan_status']=='Current'].count()

```
id                          1140
member_id                   1140
loan_amnt                   1140
funded_amnt                 1140
funded_amnt_inv             1140
term                        1140
int_rate                    1140
installment                 1140
grade                       1140
sub_grade                   1140
emp_length                  1098
home_ownership              1140
annual_inc                  1140
verification_status         1140
loan_status                 1140
dti                         1140
mths_since_last_delinq       363
pub_rec_bankruptcies        1140
dtype: int64
```

The correlation matrix:

Since we cannot use more than two variables as x-axis and y-axis in Scatter and Pair Plots, it is difficult to see the relation between three numerical variables in a single graph. In those cases, we'll use the correlation matrix.



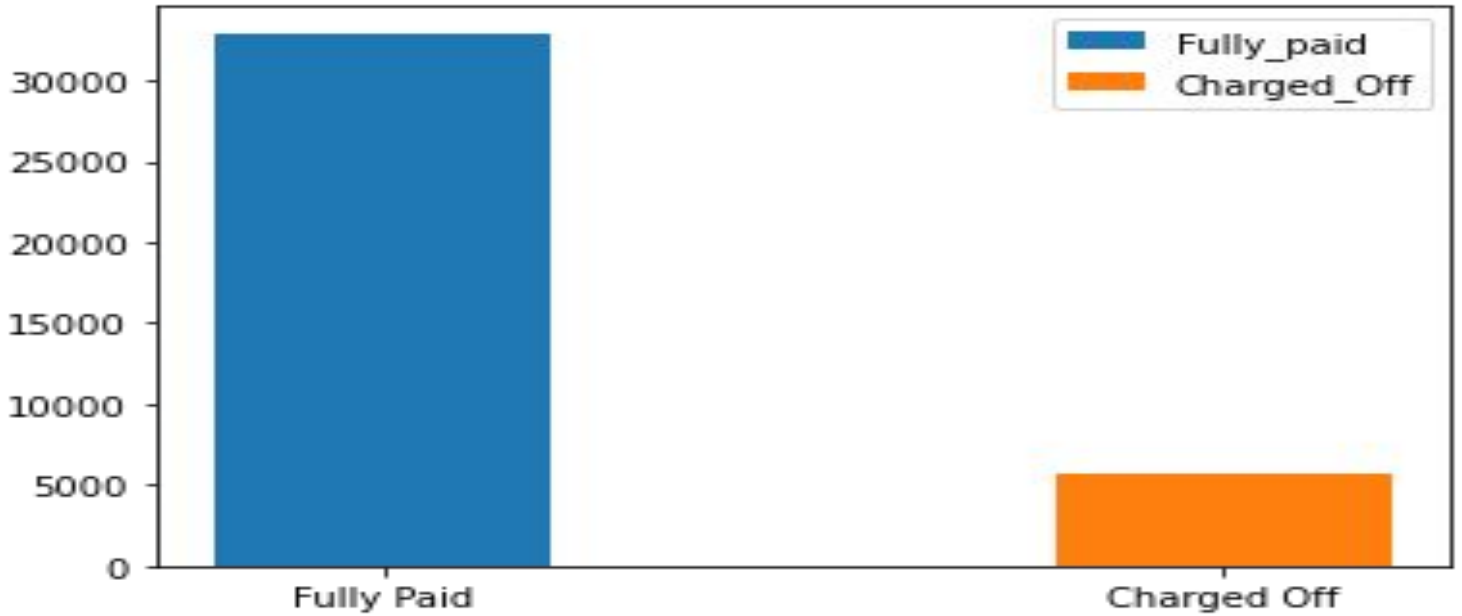
Heat Map observation:

From the above heat map we can conclude that:
There are 7 combination with high correlation and 6 combination with medium correlation,and only 1 combination with low correlation (dti vs annual income).

❖ ***Univariate analysis :*** Analyzing data over a single variable/column from a dataset, it is known as Univariate Analysis.

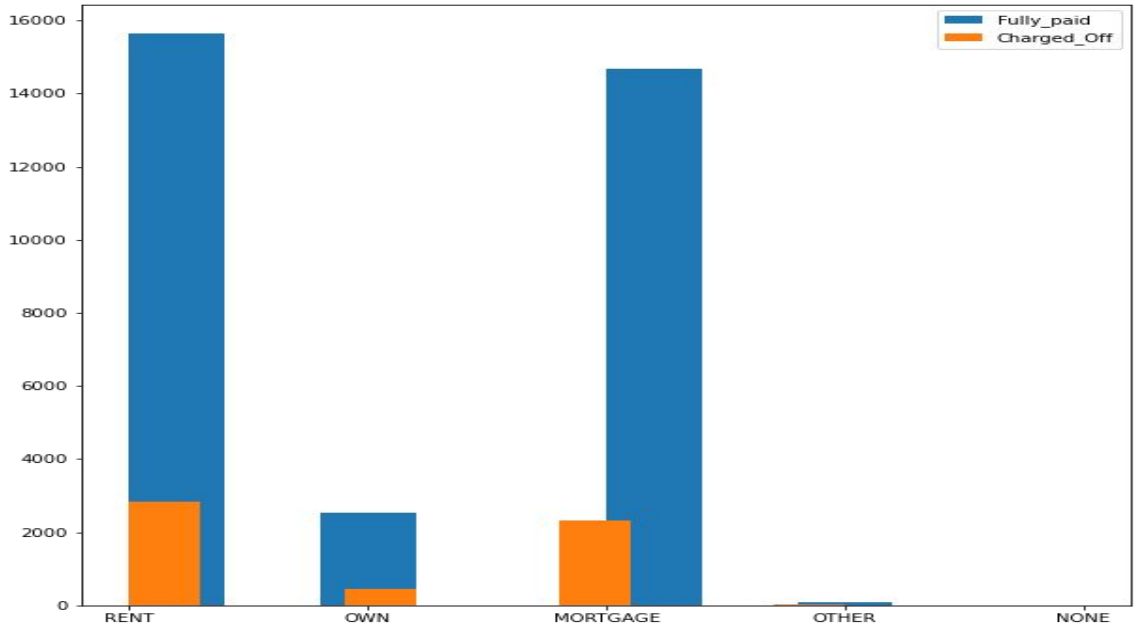
***Observation*: This graph shows the difference between the fully paid loan people and charged off people.**

Fully paid loan are more compared to charged off people



- ***Bivariate analysis :*** Analysing the data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

Categorical vs Categorical: We will plot a graph of Fully paid and charged off with respect to Home ownership and loan status



Observation:

By the above graph, we can infer that the Fully paid is more for people who have rented house in the data set.

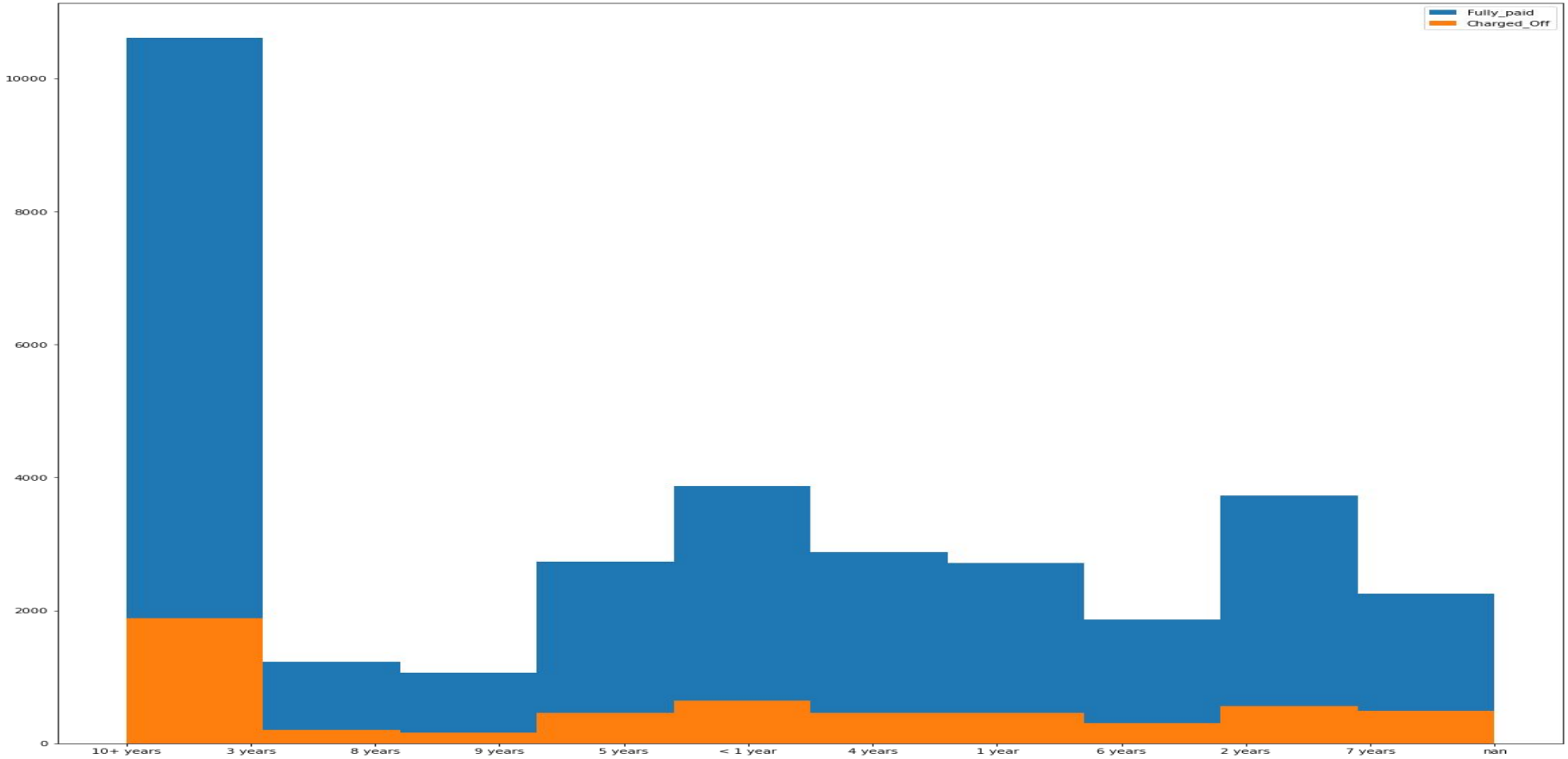
By the above graph, we can also infer that the Charged off people are less in people who have own house.

By the above graph, we can also infer that people with owned house have charged off percentage are less compared to fully paid loan amount

By the above graph, we can also infer that people staying rent and mortgage have almost similar no. of charged off people.

In all the three category charged off percentage is low than fully paid category.

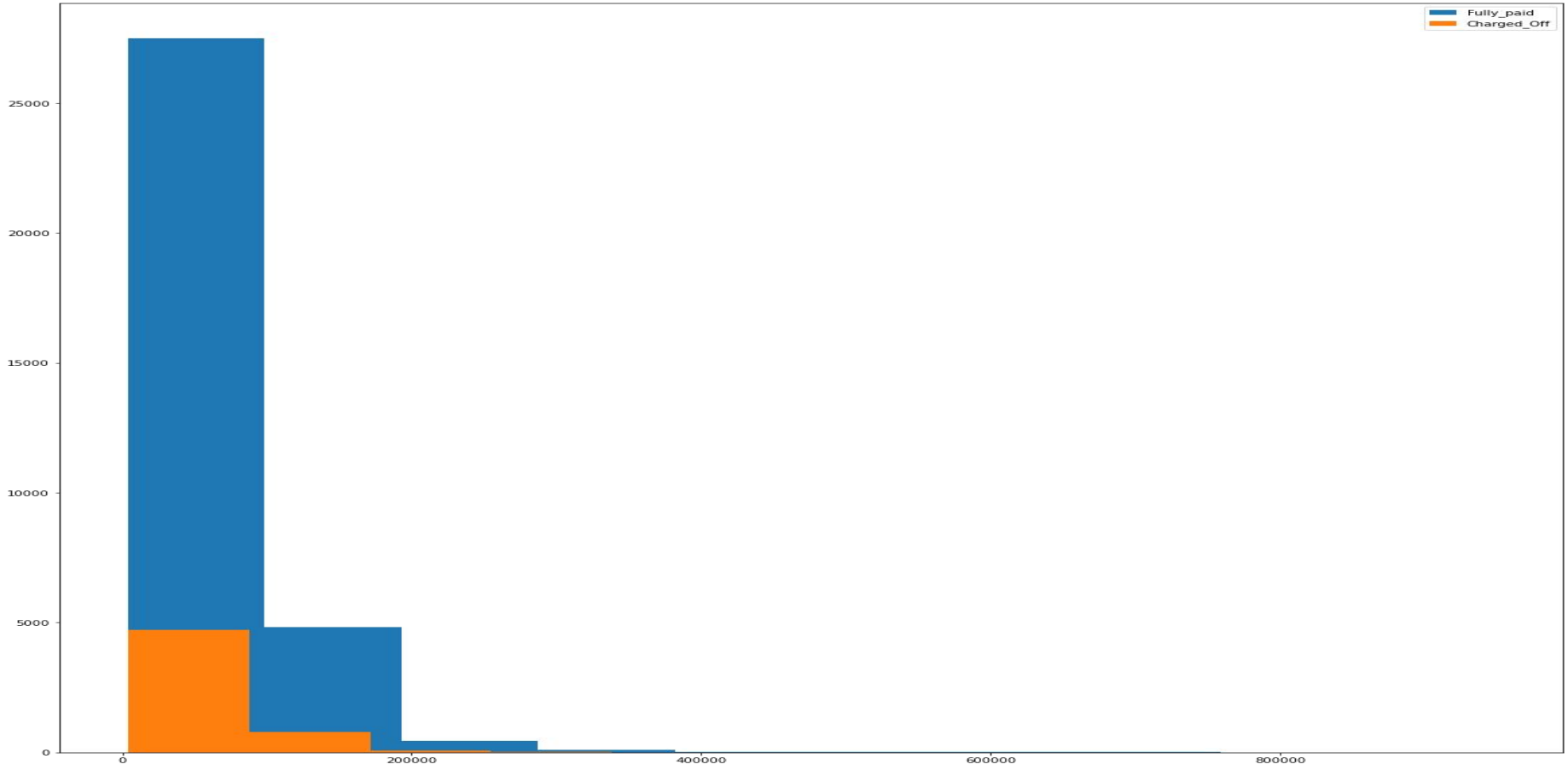
Categorical vs Numerical: We will plot a graph between Loan status V/s Employee length



Observation:

- By the above graph, we can infer that if employee length is more than 10+ yrs paid amount is more
- By the above graph, we can infer that people with employee length 8 years-9 years default less often compare to the other groups.
- People having 10+ years of experience have the highest no. of defaulters against the people who have less number of experience
- People with employee length between 1 year to 6 year have almost same no. of defaulters.
-

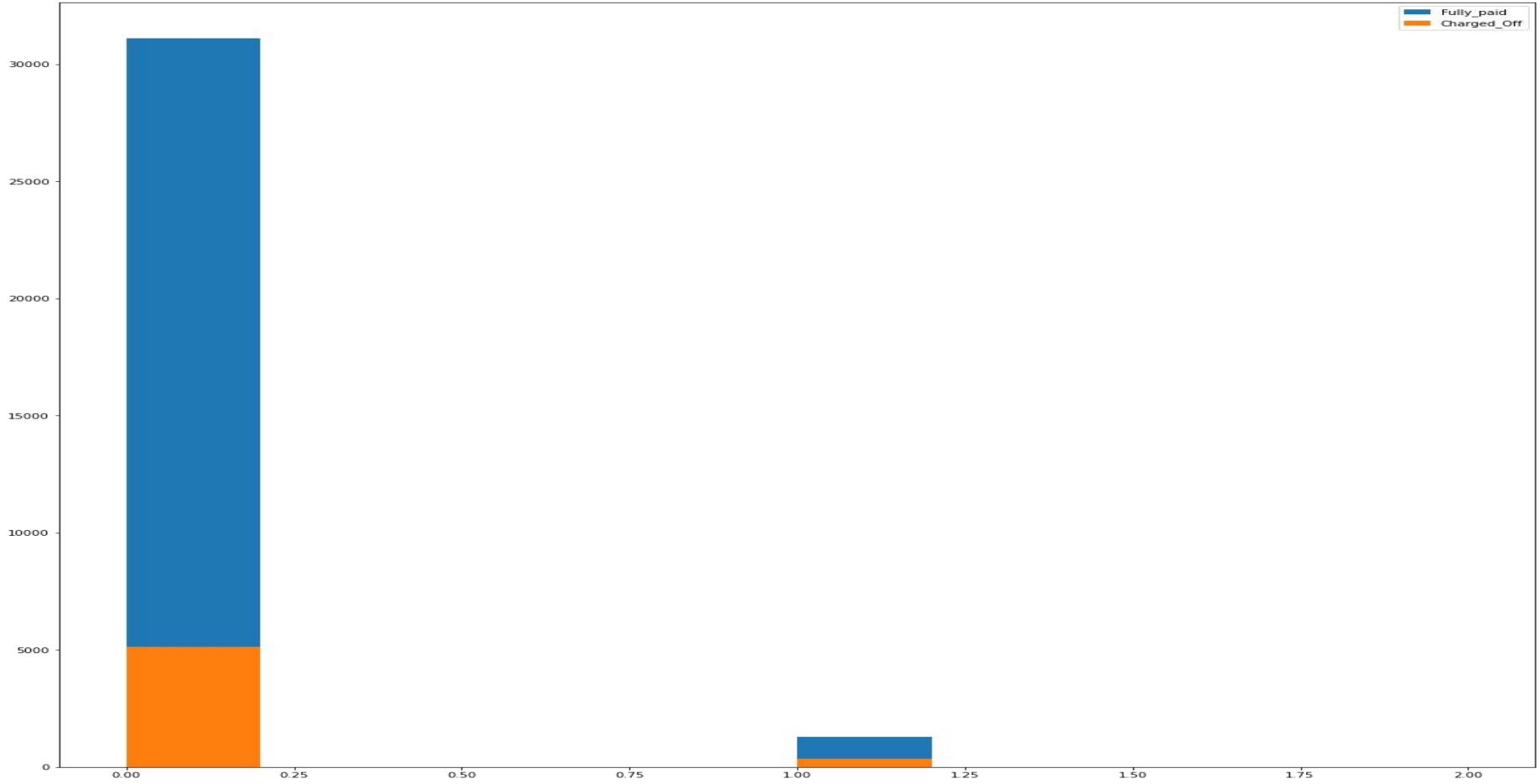
Categorical vs Numerical: Loan status Vs Annual income



Observation:

- By the above graph, we can infer that people with more annual income have more chance of paying the loan.
- By the above graph, we can infer that people with less annual income have less chance of paying the loan amount.
- By the above graph, we can infer that maximum loan is taken in the range of 0-400000 and with that defaulter count also decreases as we move from 0 - 400000
- We can also say that from 0 to close to 50000 we have more number of defaulters.
- As annual income increases loan count becomes low.

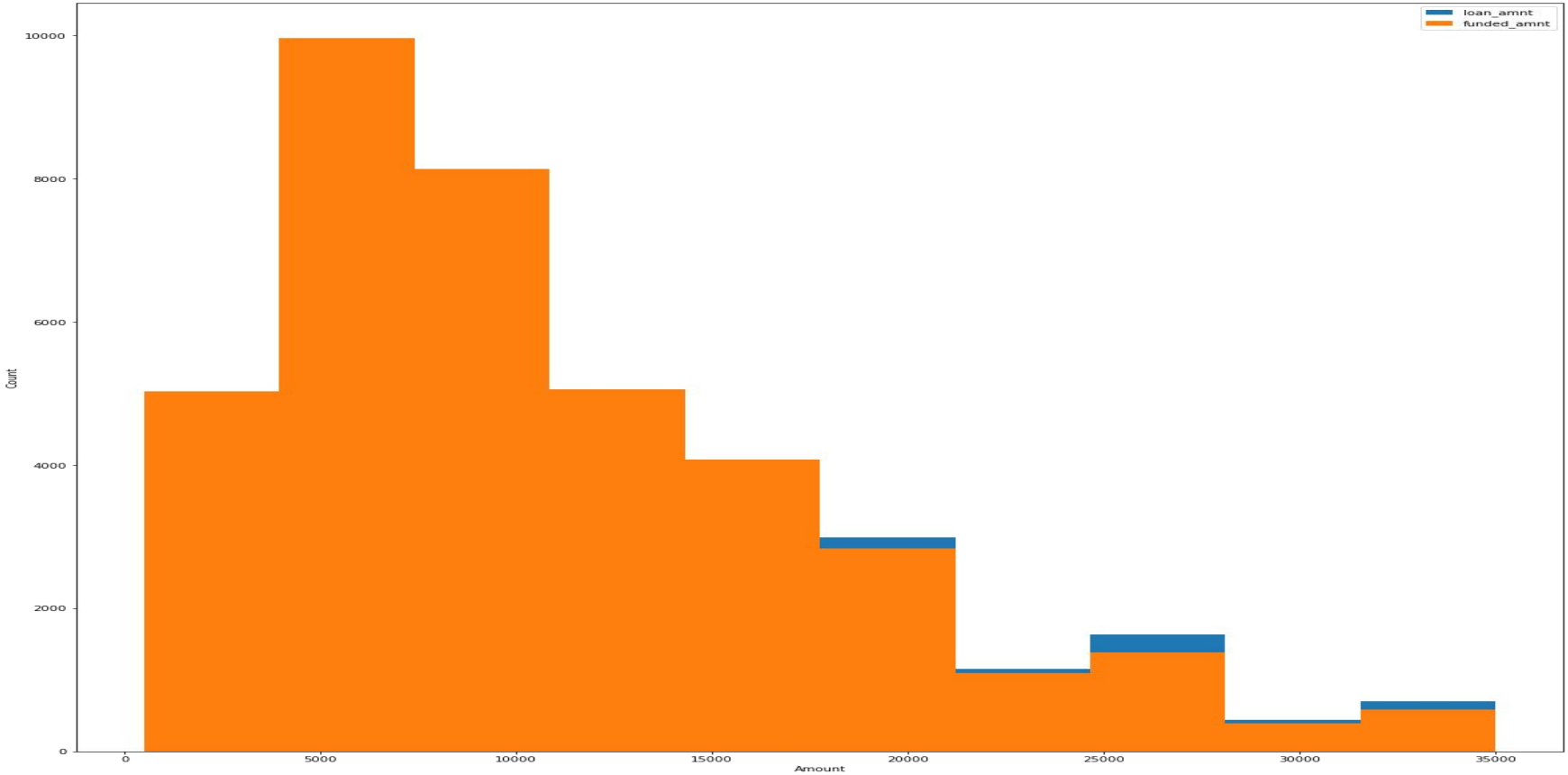
Categorical vs Categorical: Loan status vs pub_rec_bankruptices



Observation:

- By the above graph, we can infer that people form 0- 25% have 5000 defaulters and from 5000 to 300000 we have good loan which are fully paid.
- In the range of 1%to 1.25 % we have approx 1000 defaulters,and again here the fully paid loan number is nearly 2000

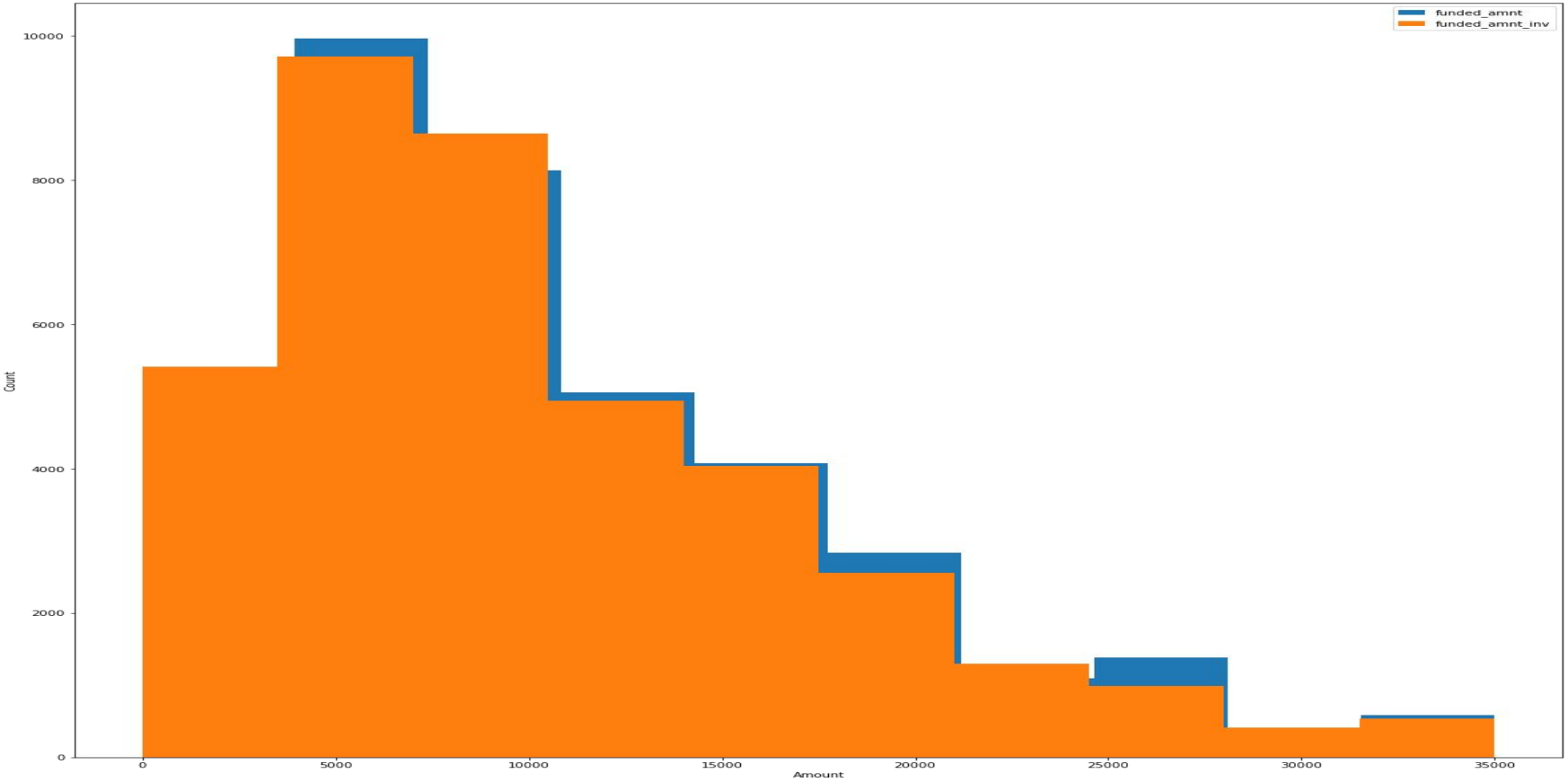
Numerical vs Numerical: Amount vs count



Observation:

- By the above graph, we can infer that 5000 people have got the full amount funded which they have applied.
- By the above graph, we can infer that the highest loan amount which was funded is on 10000
- On the scale of 17000 to 22000 there were few percentage of loan which was not funded hence we can mention them as defaulters.
- The highest number of defaulter were the amount was not funded are between 24000 to 28000.
- The overall average of loan funded is more than than the loan amount.

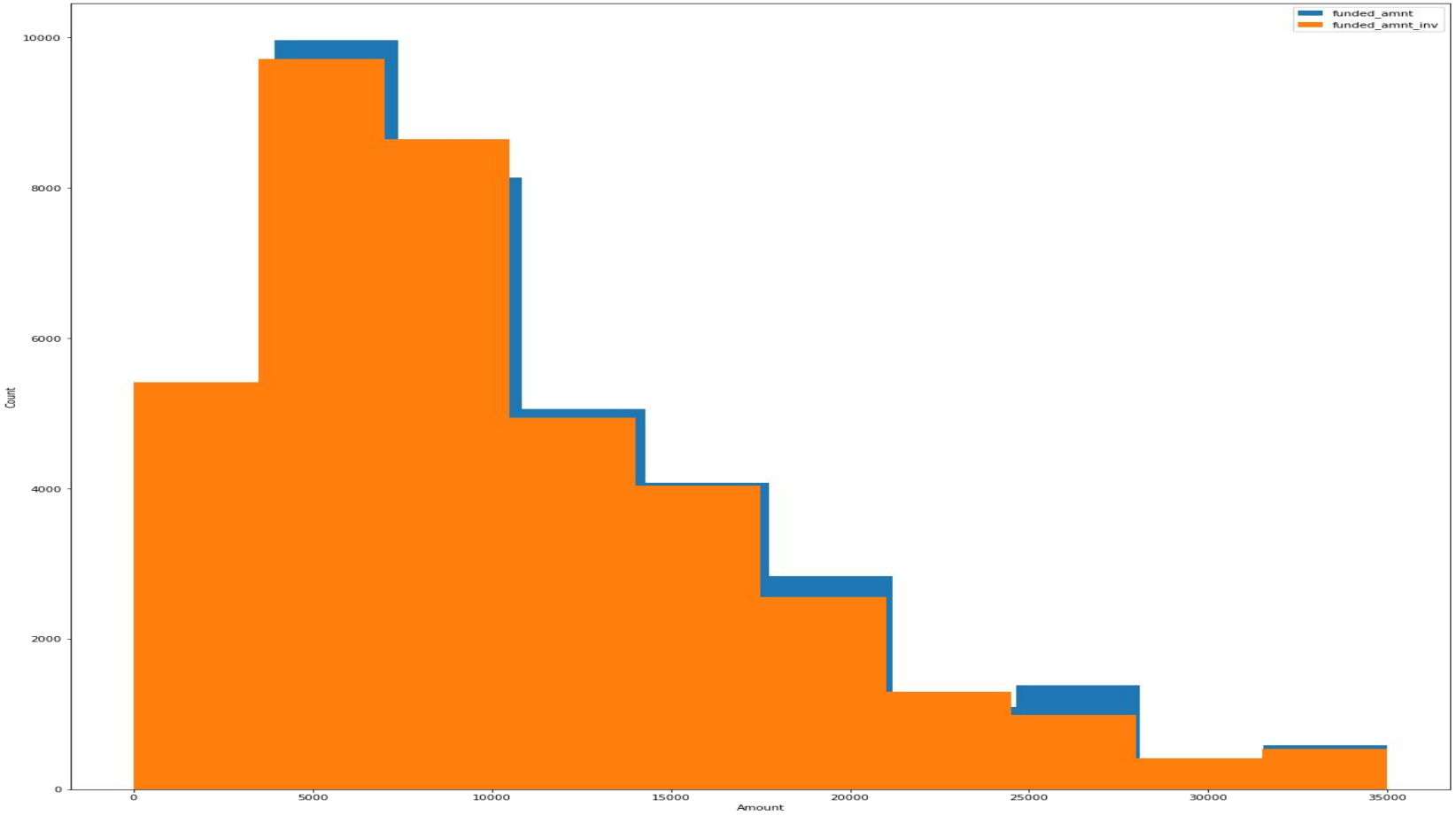
Numerical vs Numerical: Amount vs count



Observation:

- By the above graph, we can infer that out of 10000 people the highest loan was funded to approx 9000 people.
- By the above graph, we can infer that highest amount of loan was funded to very few number of people
- By the above graph, we can infer for amount 24000 to 28000 there there is a big margin between the funded amount and the funded amount which was invested.
- By the above graph, we can infer for amount 20000 to 24000 highest number of fund was invest with no defaulters.

Numerical vs Numerical: Amount vs count



Observation:

- By the above graph, we can infer that from 0-3000, on 5000 people funded amount was invested and there were no defaulters.
- By the above graph, we can infer that for amount 4000 -10000 the overall loan amount was more than the funded amount invested.
- By the above graph, we can infer that for amount 4000 -10000 there is a small margin of people for which the loan was not funded
- By the above graph, we can infer that for amount 24000 -28000 there is a overall huge difference between the loan amount and the funded amount invested.
- By the above graph, we can infer that overall percentage of funded loan invested is more than the loan amount.

Recommended group where loan can be credited.

1. *Person who have more number of service experience*
2. *People with high income category*
3. *People who stay in rented apartment/house are more likely to pay the loan.*

Risky groups (Defaulters)

1. *People with low annual income*
2. *People with more no. of bankruptcy*

Thank You