Lending Club Case Study

Dillan Muthanna

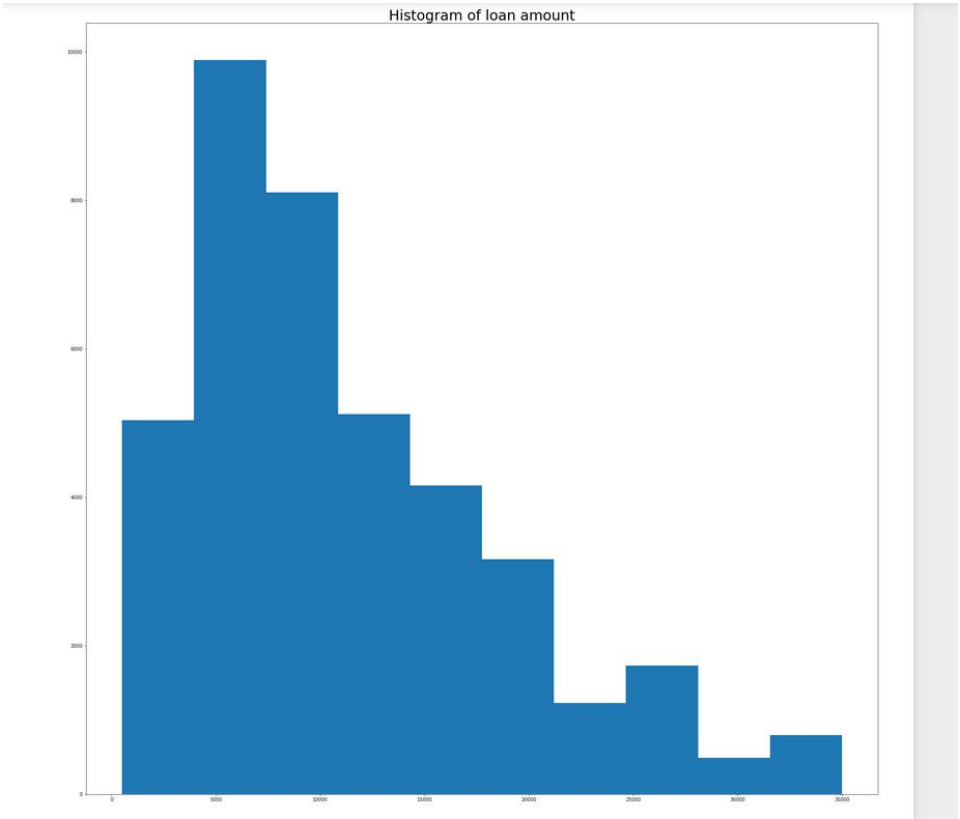
Cleaning columns

- About 60 columns out of 111 were full of null values.
- A few had redundant columns and had to be removed (eg initial_list_status, application_type, policy_code etc)

```
In [93]: clean_data_extra.columns

Out[93]: Index(['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv',
                'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp_title',
                'emp_length', 'home_ownership', 'annual_inc', 'verification_status',
                'issue_d', 'loan_status', 'desc', 'purpose', 'title', 'zip_code',
                'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line',
                'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util',
                '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',
                'last_pymnt_d', 'last_pymnt_amnt', 'last_credit_pull_d',
                'pub_rec_bankruptcies'],
               dtype='object')
```

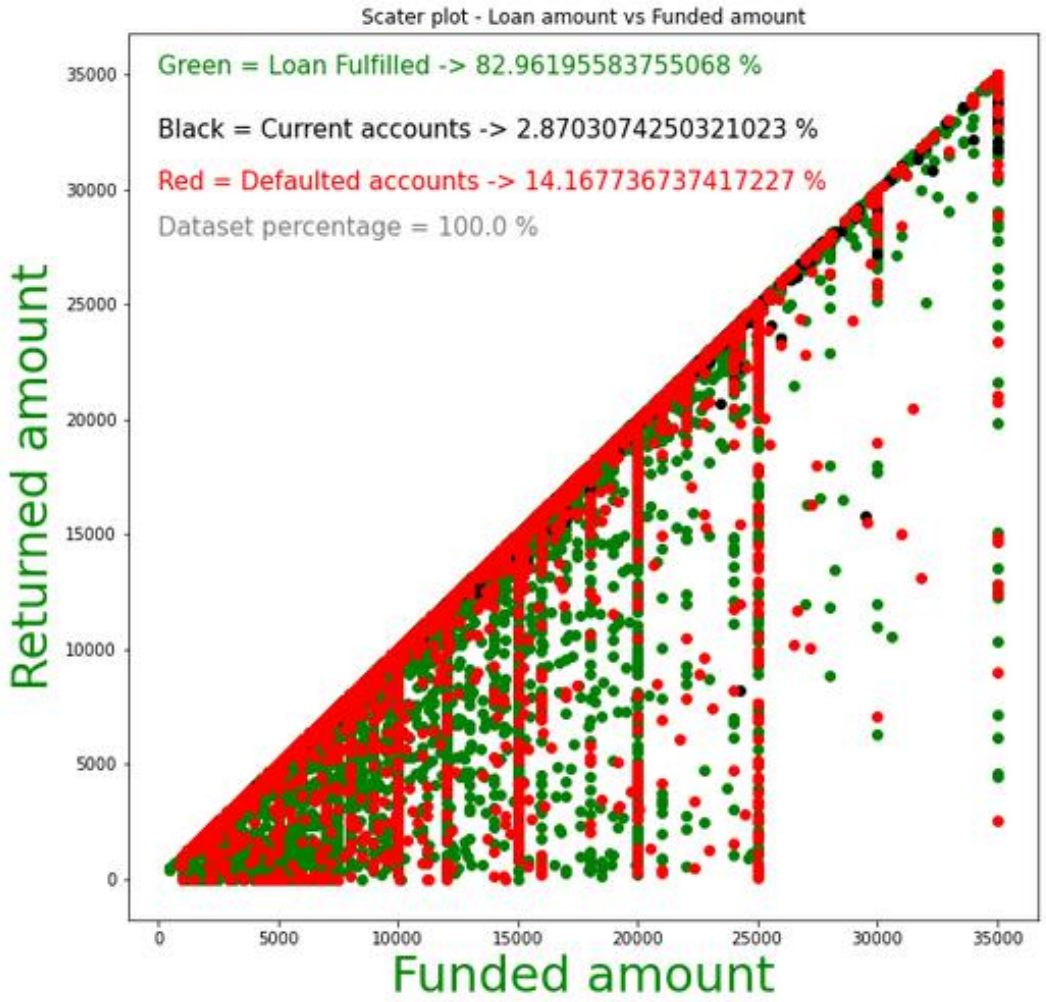
Lets Understand the distribution of loan amount




```
: clean_data_extra.funded_amnt.describe()

: count    39717.000000
  mean     10947.713196
  std       7187.238670
  min        500.000000
  25%       5400.000000
  50%       9600.000000
  75%      15000.000000
  max      35000.000000
  Name: funded_amnt, dtype: float64
```

Understanding trend trend in users

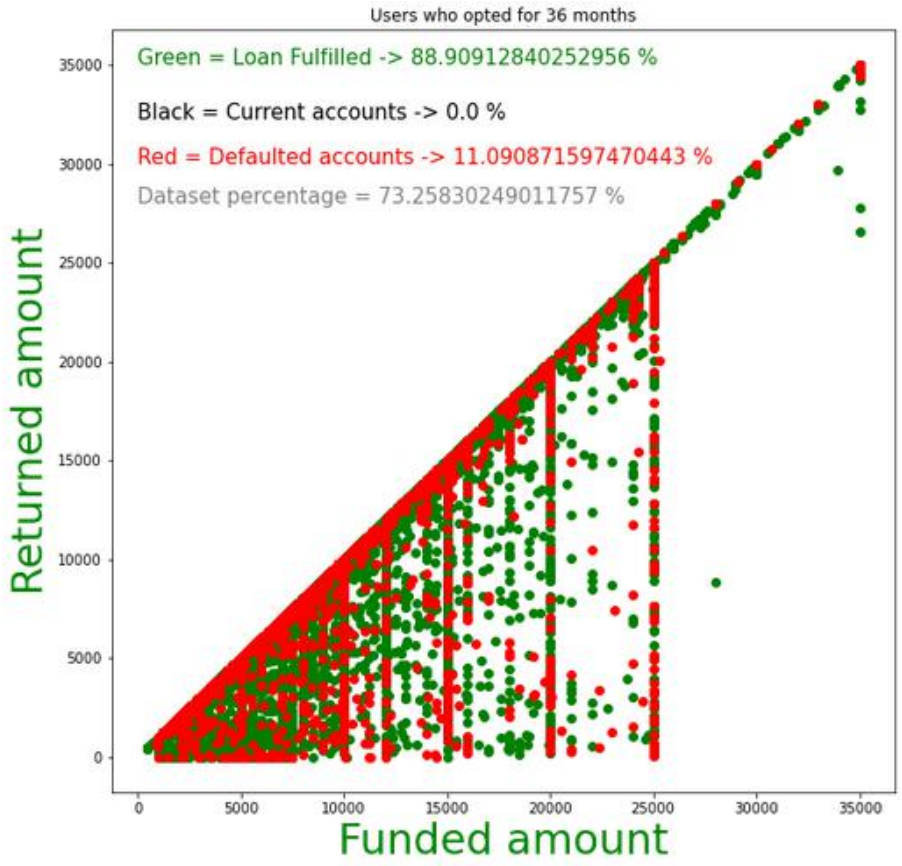


Insights

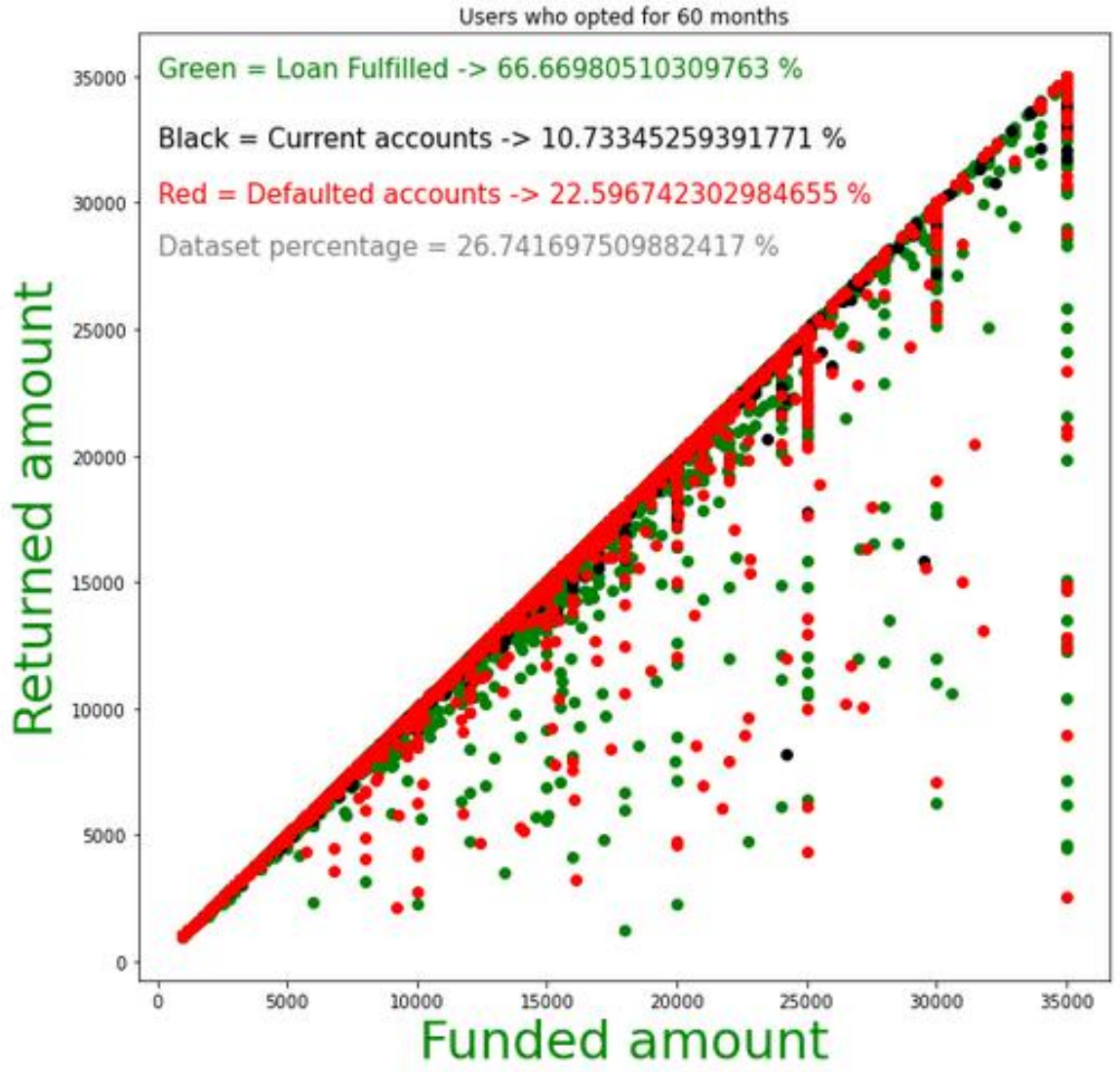
- Dataset is valid , there is no single instance where returned amount is greater than funded amount
- Funds are generally given around 500 range starting from 0 to 3500

- 82 percent of users are good users who repay on time , only 14 percent are defaulters

Understanding trend in different terms 36 months vs 60 months



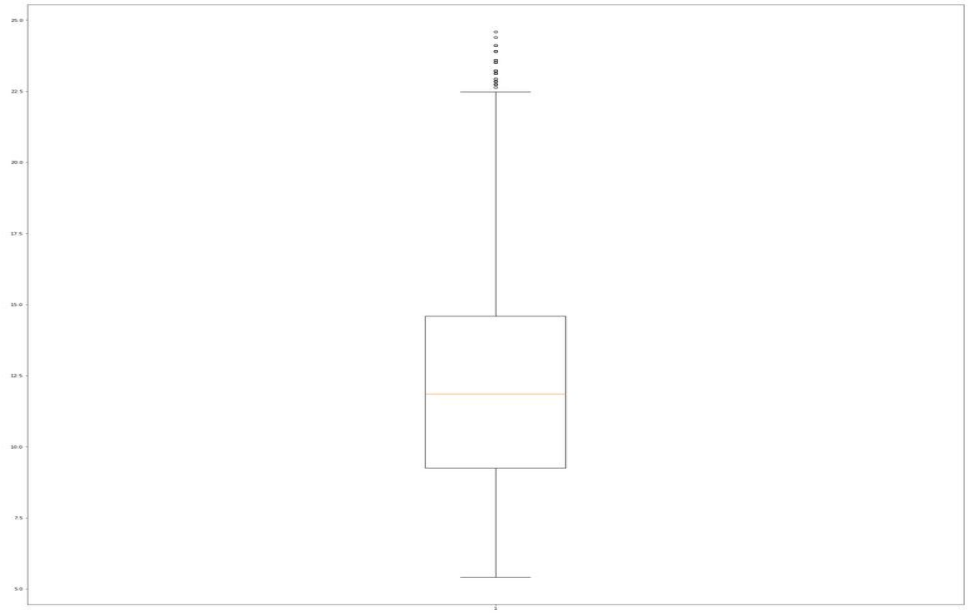
60 months


Users who opted for 60 months
Green = Loan Fulfilled -> 66.66980510309763 %
Black = Current accounts -> 10.73345259391771 %
Red = Defaulted accounts -> 22.596742302984655 %
Dataset percentage = 26.741697509882417 %
Returned amount
Funded amount
0
5000
10000
15000
20000
25000
30000
35000

Defaulters are more in 60 months as opposed to 36 months

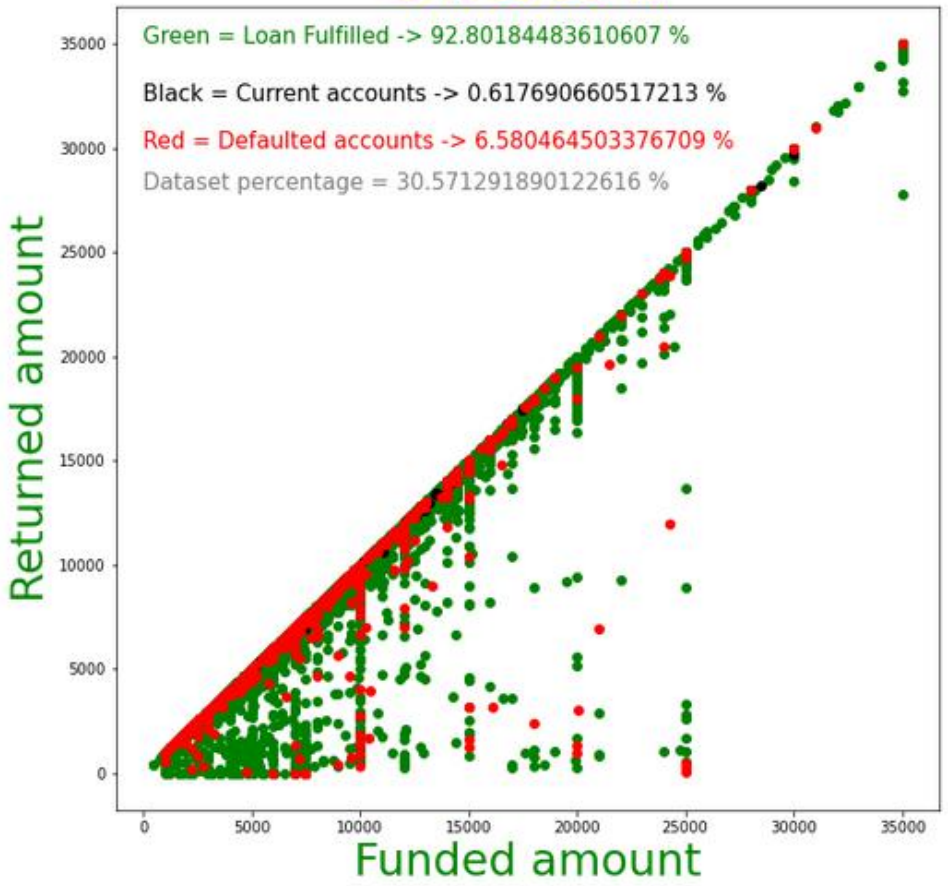
Maybe people get bored and lazy paying loans in the long term !

Interest rate

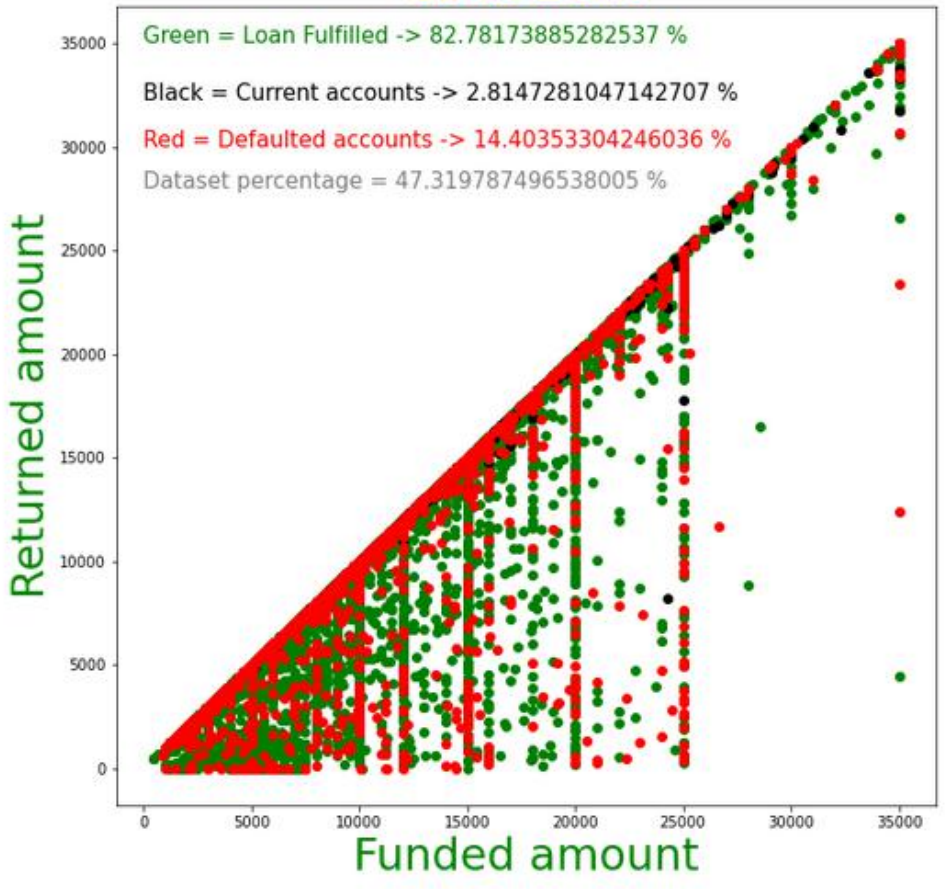


interest rate range between 5 and 24 also we have outliers above 22.5 , we need to get rid of those if we plan to use any classification algorithm)

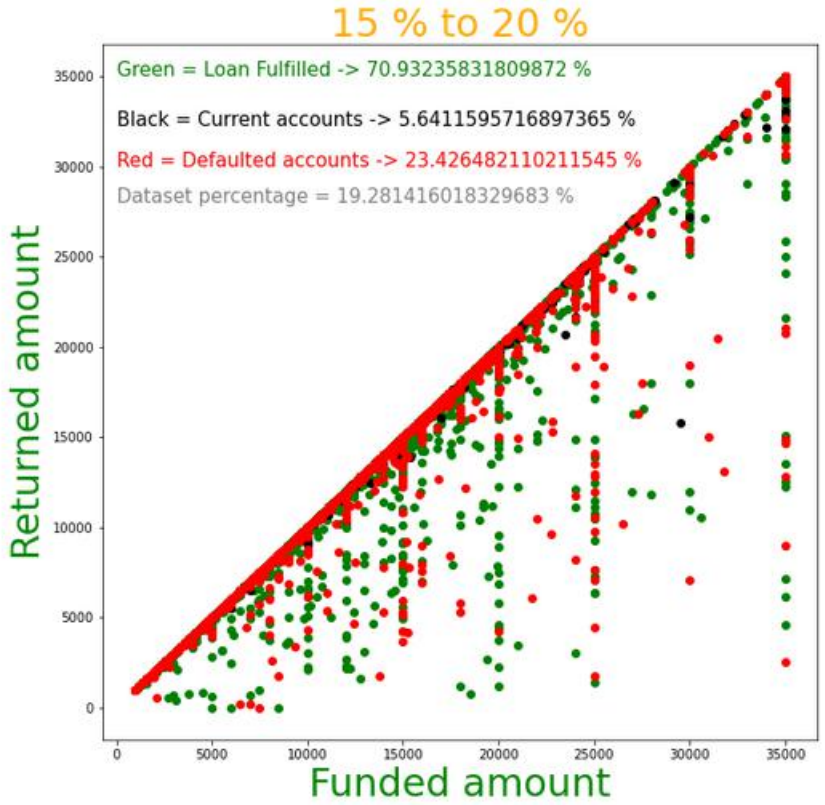
5 % to 10 %

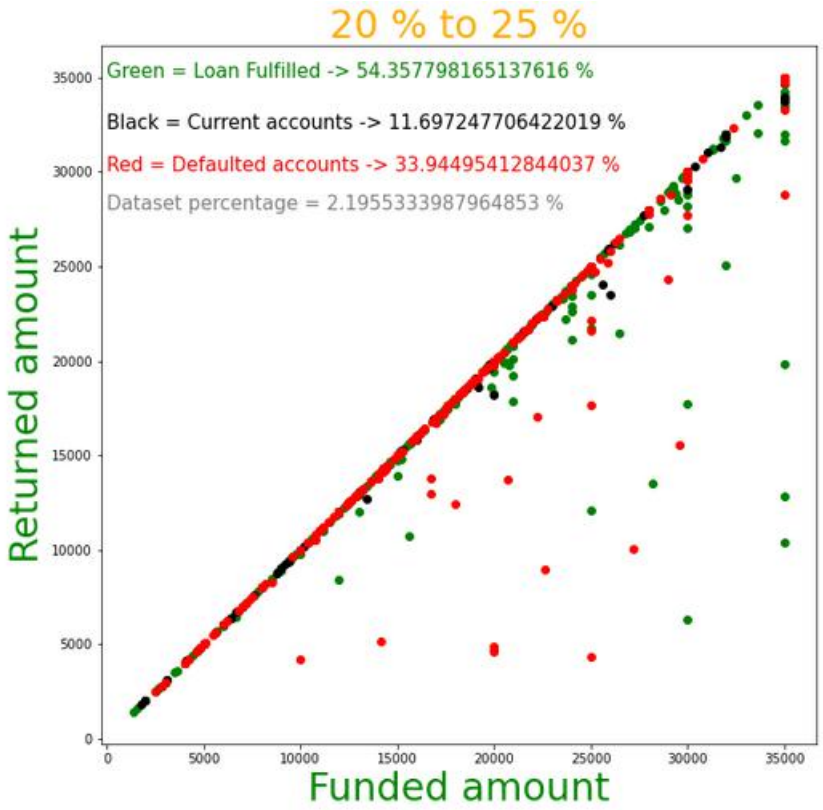

Green = Loan Fulfilled -> 92.80184483610607 %
Black = Current accounts -> 0.617690660517213 %
Red = Defaulted accounts -> 6.580464503376709 %
Dataset percentage = 30.571291890122616 %
Returned amount
Funded amount
0
5000
10000
15000
20000
25000
30000
35000

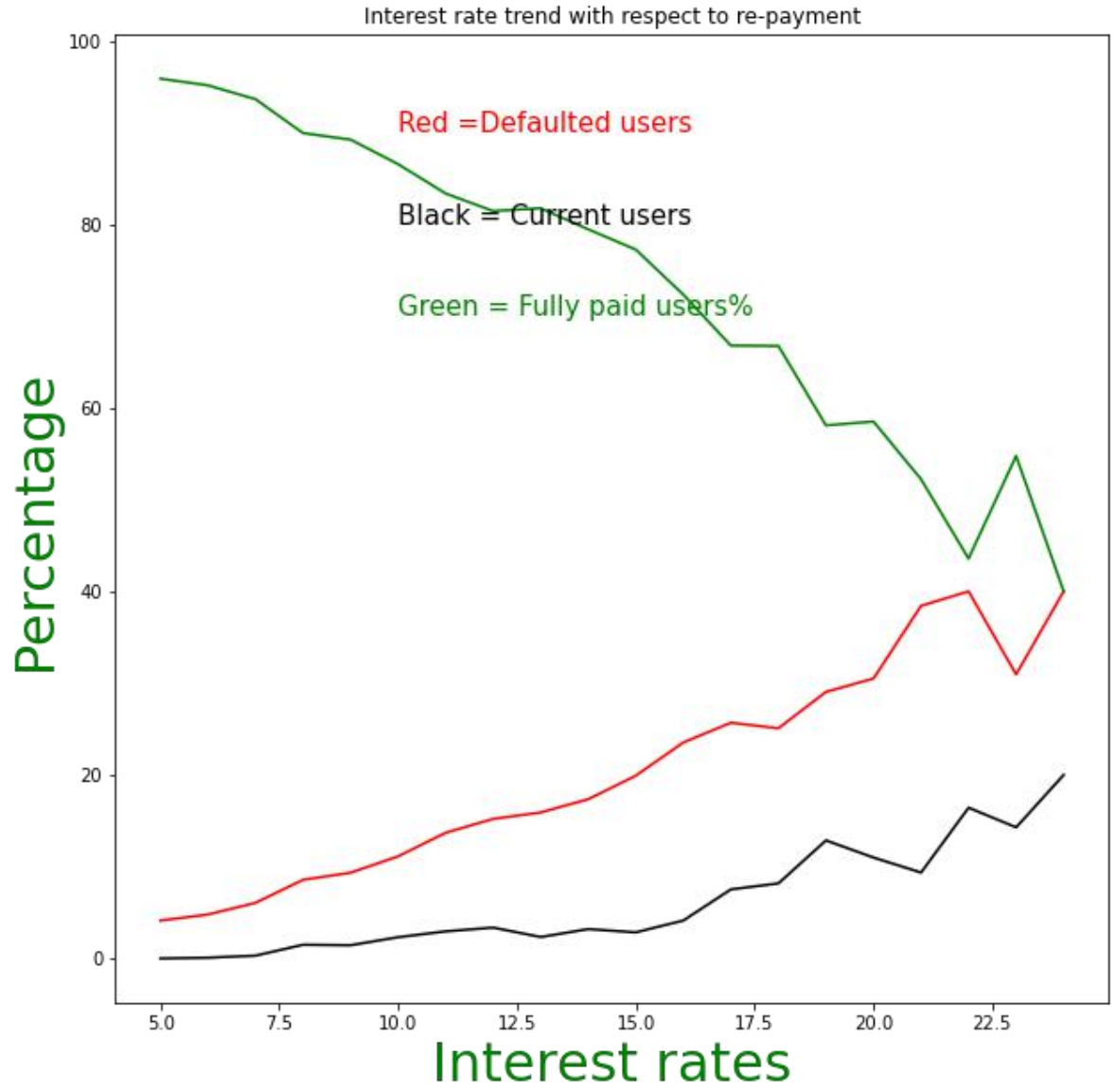
10 % 15 %


Green = Loan Fulfilled -> 82.78173885282537 %
Black = Current accounts -> 2.8147281047142707 %
Red = Defaulted accounts -> 14.40353304246036 %
Dataset percentage = 47.319787496538005 %
Returned amount
Funded amount
0
5000
10000
15000
20000
25000
30000
35000

As the interest rate increases, the number of defaulters is also increasing






Interest rate trend with respect to re-payment
Red =Defaulted users
Black = Current users
Green = Fully paid users%
Percentage
Interest rates
100
80
60
40
20
0
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

- interesting mirror like trend between fully paid accounts and defaulters. at low interest rates people are able to easily repay loans
- anomoly observed above 22.5 could be due to outliers in the dataset as found earlier

Interesting to see that people who usually describe in detail why they need the loan have high probability of repayment