Lending Club Case Study

Submitted by – Sidharth Rai (5760097)
Sourav Sirohi (5623940)

An assignment for upGrad and IIITB Machine Learning & AI Program.

CONTENTS

Problem statement

- **Business Understanding:**
  - **Lending Club,** a consumer finance company which specializes in lending various types of loans to urban customers.
  - There are two primary risks associated with loan approval decisions: l
    - **Loss of business** if loans aren't approved.
    - **Financial loss** if loans default.
  - The dataset provided contains information on past loan applicants and their loan statuses, aiming to identify patterns indicating loan default tendencies.
- **Business Objective:**
  - Minimize credit loss by identifying risky loan applicants.
  - Understand the driving factors behind loan default to optimize risk assessment and reduce credit loss.
  - By identifying risky loan applicants, the company aims to reduce credit loss and enhance risk assessment processes.
- **Data Understanding:**
  - The dataset contains complete loan data for all loans issued from 2007 to 2011, including information on loan statuses such as 'Fully Paid' and 'Charged Off'.
  - Referring to the provided data dictionary for a description of the variables included in the dataset.

Analysis Approach

Data Cleaning

- Dropping the null valued columns

Data Understanding

- Identifying the business use case of each column from the provided Data Dictionary*.

Univariate Data Analysis

- Observing each column one after the another, with describe, value count functions or visualization, dropping and selecting columns based upon business problem requirement.

Bivariate Data Analysis

- Relative analysis of columns with Loan Status column to determine the direct or indirect impact.

Conclusions

- Key factors include loan maturity, borrower grade, small business loans, and state-wise default rates, guiding informed lending decisions.

Recommendations

- Emphasize quality over quantity in long-term loans, exercise prudence with small business lending, and reassess operations in high-default states.

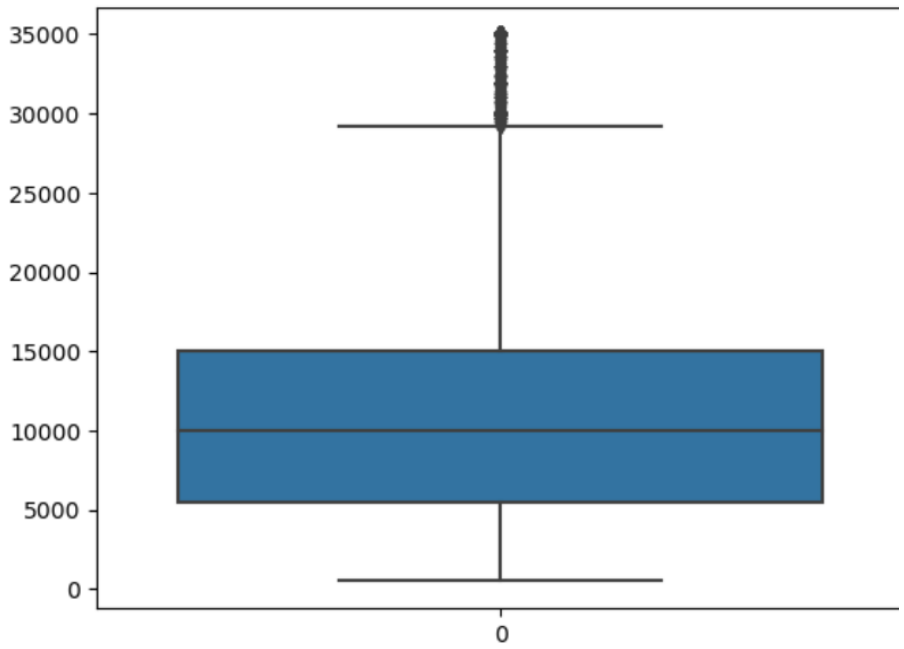
Univariate Analysis

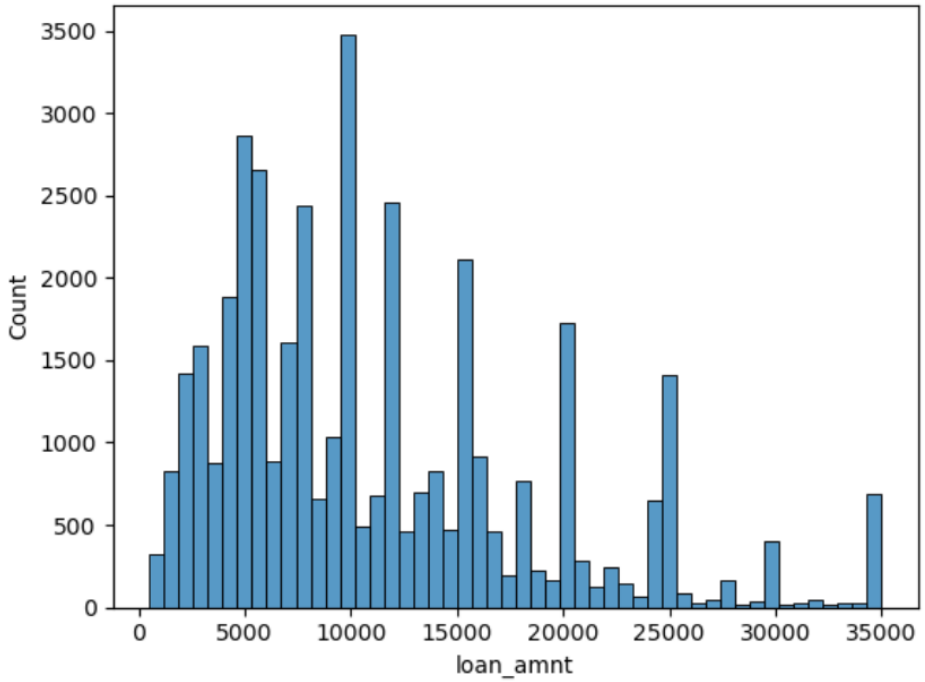
Unpacking loan dataset columns to understand their individual characteristics and relevance to loan default prediction.

Loan Amount

(loan_amnt)

The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value





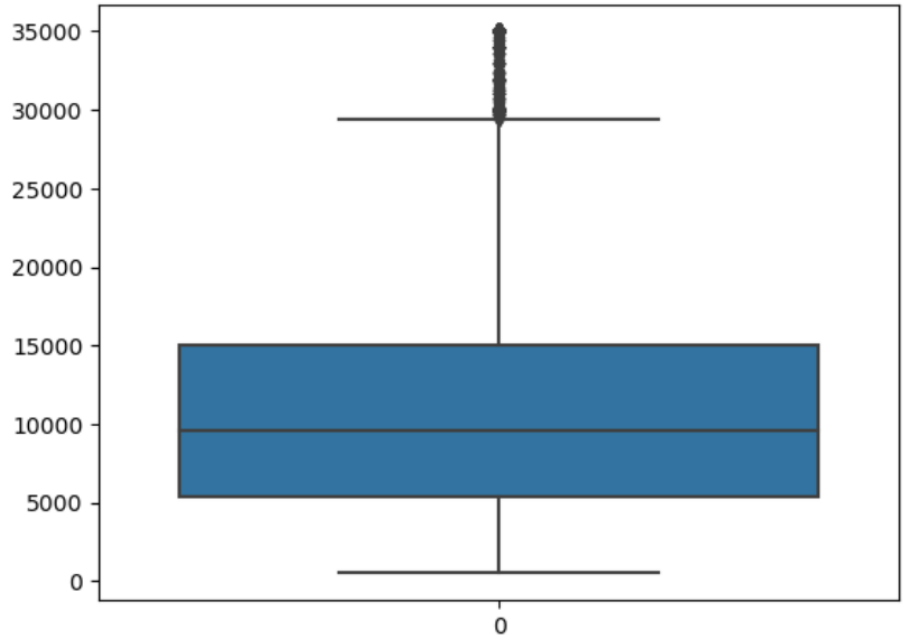
Observations

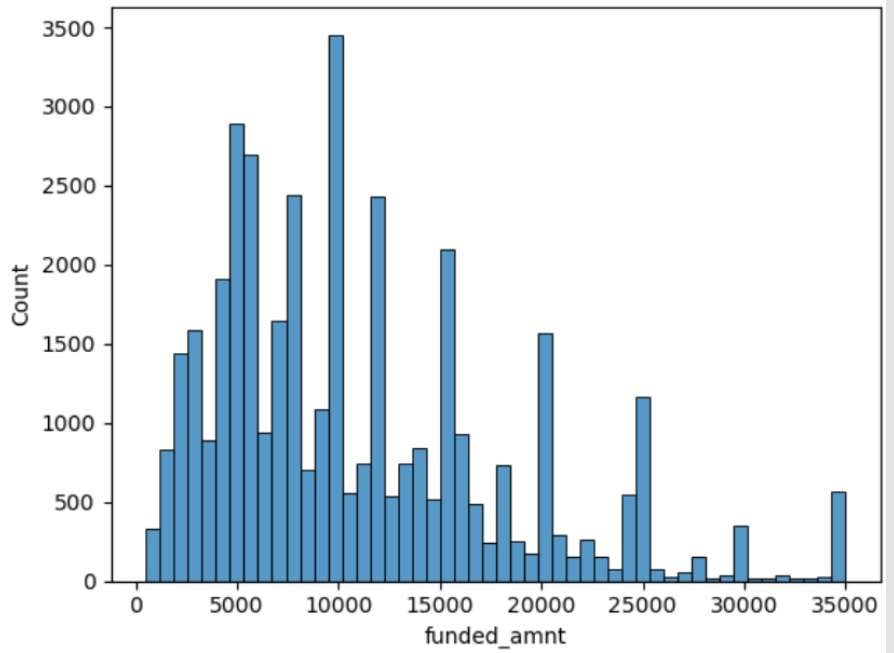
- Loan amount distribution is right-skewed, with higher values towards the upper range.
- Box plot indicates outliers at the higher end.
- Demand peaks for amounts in multiples of 5000 (e.g., 5000, 10000).

Funded Amount

(funded_amnt)

The total amount committed to that loan at that point in time.





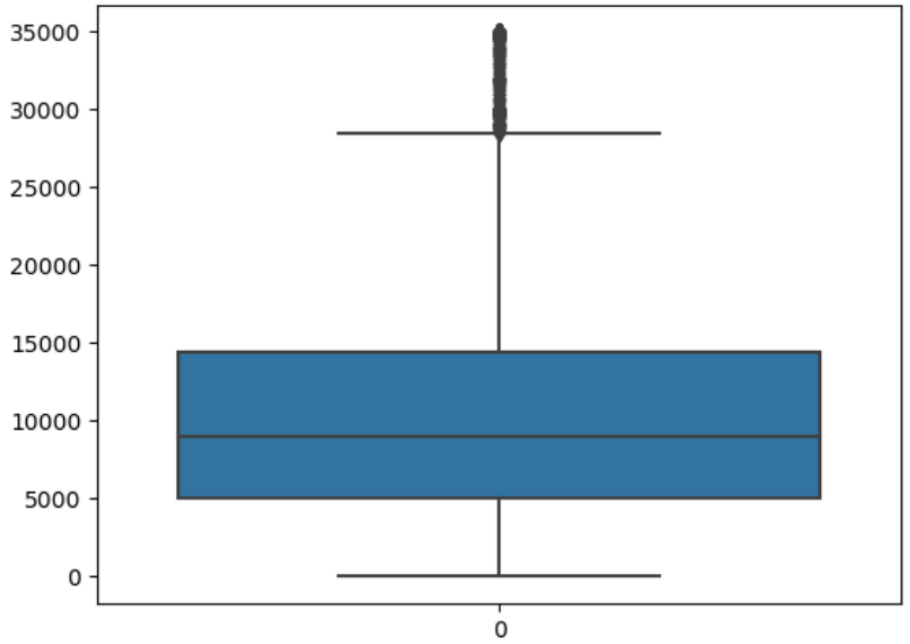
Observations

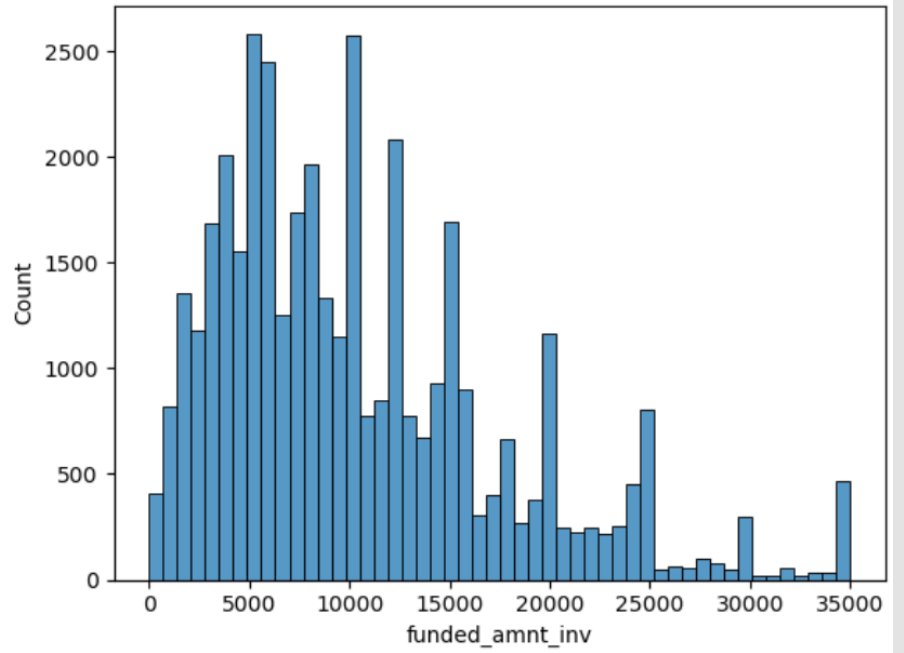
- Right-skewed distribution with outliers at higher values.
- Box plot confirms skewness and presence of outliers.
- Peaks observed for amounts in multiples of 5000.
- Distribution pattern mirrors loan requested amount distribution.

Funded Amount Inv

(funded_amnt_inv)

The total amount committed by investors for that loan at that point in time.





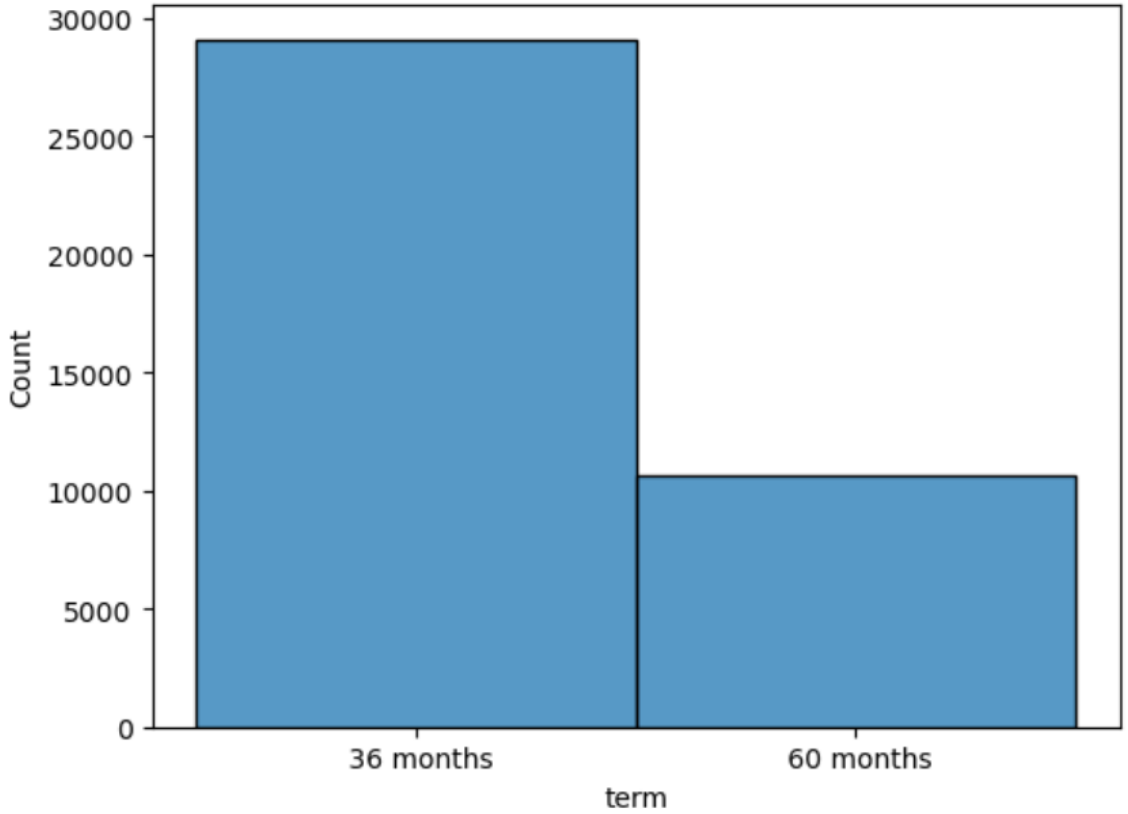
Observations

- Right-skewed distribution with outliers at higher values.
- Box plot confirms skewness and presence of outliers.
- Peaks observed for amounts in multiples of 5000.
- Distribution pattern mirrors loan requested amount and committed amount.

Term

(term)

The number of payments on the loan. Values are in months and can be either 36 or 60.



Observations

- Categorical attribute with 36-month or 60-month terms.
- Approximately 73% of loans have a 3-year tenure, while 27% have a 5-year tenure.

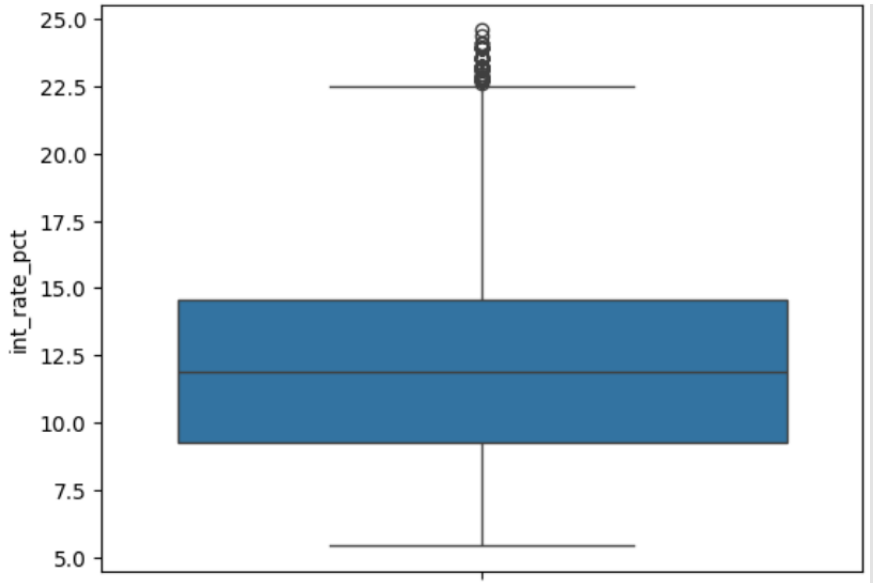
Interest Rate Percentage

(int_rate)

Interest Rate on the loan.

	int_rate
0	10.65%
1	15.27%
2	15.96%
3	13.49%
4	12.69%

	int_rate_cat	int_rate_pct
0	Medium	10.65
1	High	15.27
2	High	15.96
3	Medium High	13.49
5	Low	7.90

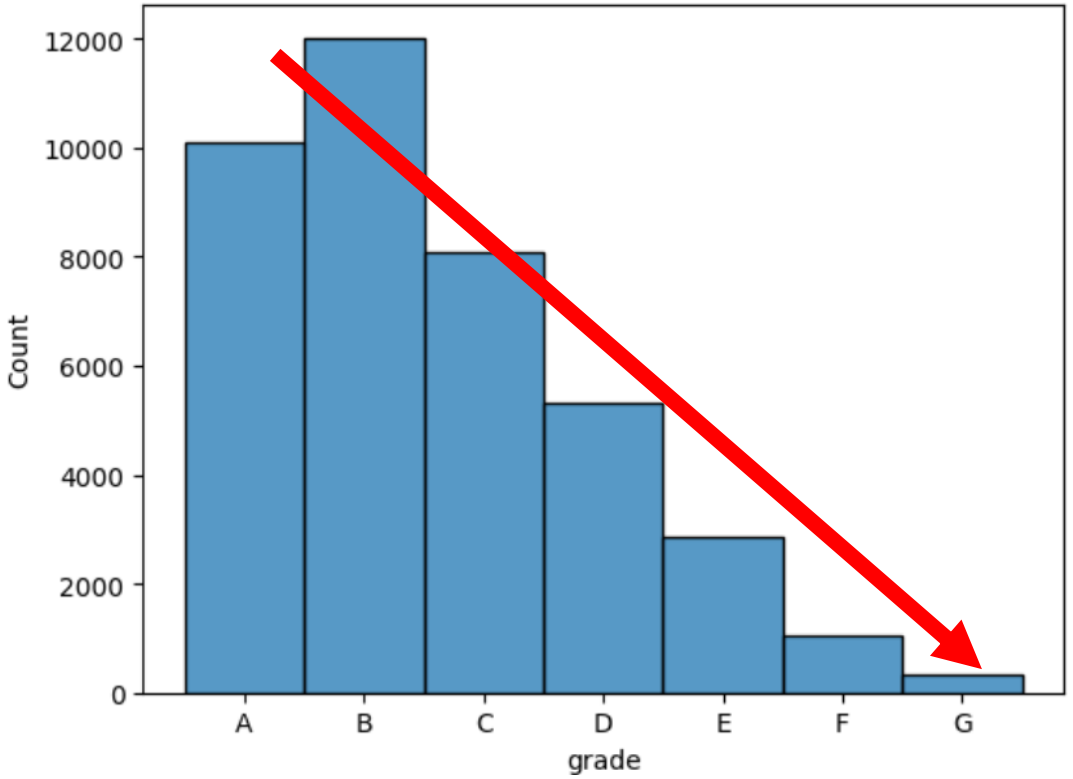


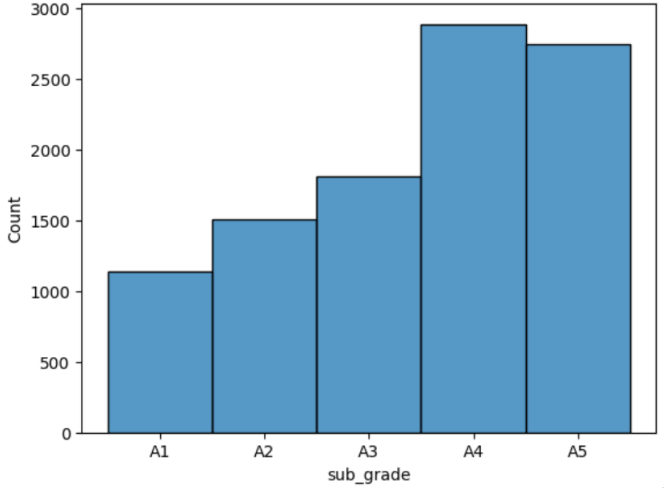
- Dropped '%' symbol, from object datatype to float datatype.
- Created a categorical variable based on interest rate.
  - Upto 1st quartile will be classified as 'Low Interest'.
  - From 1st quartile to median will be classified as 'Medium'.
  - From median to 75th quartile will be classified as 'Medium High'
  - From 75th quartile to 100th quartile will be classified as 'High'
  - Outlier will be classified as 'Extremely High'.
- **High Interest Rate may lead to high chances of defaulting.**

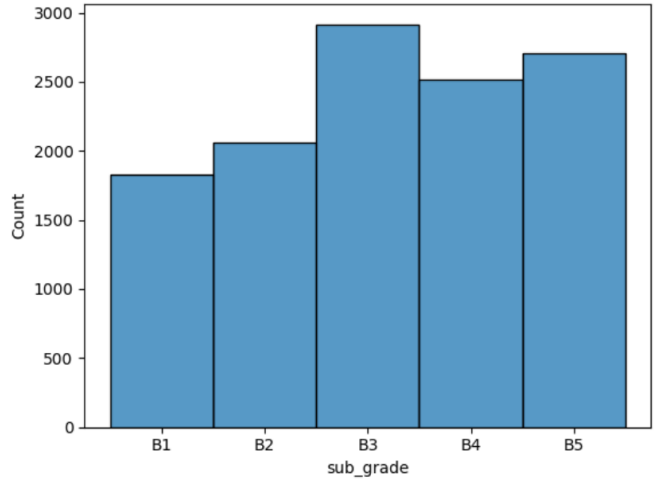
Grade and Sub Grade

(grade & sub_grade)

LC assigned loan grade and sub grade







Observations:

- Grade signifies borrower creditworthiness from A to G.
- Subgrade divides each grade into 5 categories, reflecting increasing risk.
- Majority of borrowers fall under Grade B, with decreasing numbers towards higher risk grades.

Employment Length

(emp_length)

Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

```
null count = 1075
Distinct count:
emp_length
10+ years    8879
< 1 year     4583
2 years      4388
3 years      4095
4 years      3436
5 years      3282
1 year       3240
6 years      2229
7 years      1773
8 years      1479
9 years      1258
Name: count, dtype: int64
```

```
null count = 0
Distinct count:
emp_length
10+ years    9954
< 1 year     4583
2 years      4388
3 years      4095
4 years      3436
5 years      3282
1 year       3240
6 years      2229
7 years      1773
8 years      1479
9 years      1258
Name: count, dtype: int64
```

Observations:

- 1075 observations have null values in employee length.
- These null values will be replaced with "10+ years", the most frequent category. (mode)

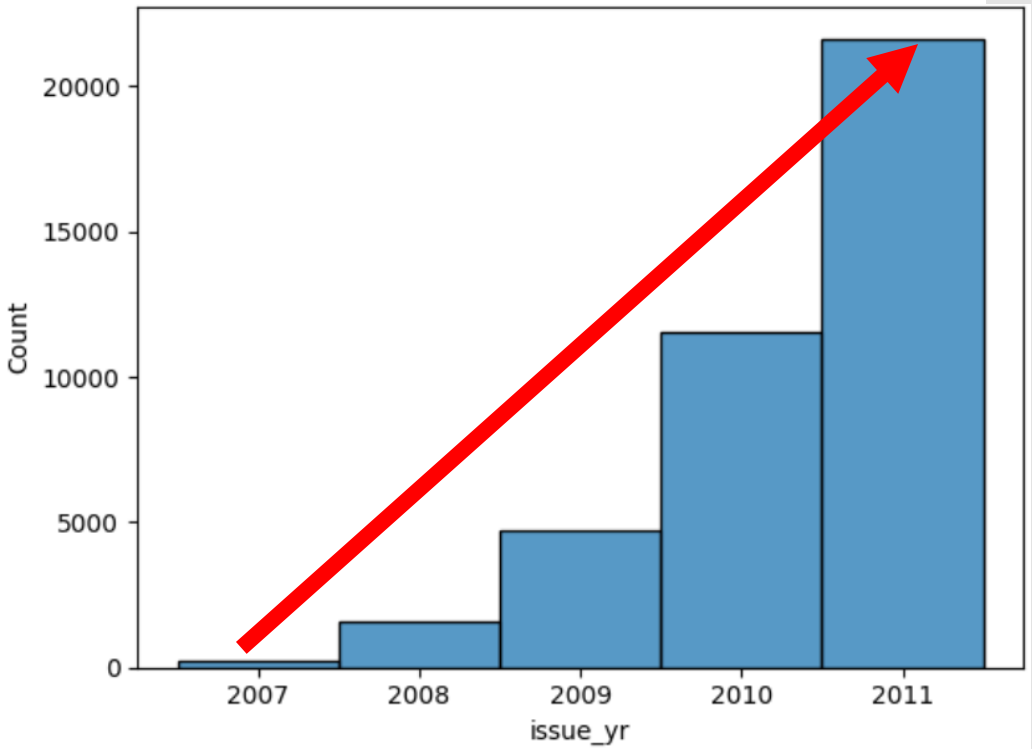
Issue Date

(issue_d)

The month & year which the loan was funded.

	issue_d
0	Dec-11
1	Dec-11
2	Dec-11
3	Dec-11
4	Dec-11

	issue_mnt	issue_yr
0	Dec	2011
1	Dec	2011
2	Dec	2011
3	Dec	2011
4	Dec	2011



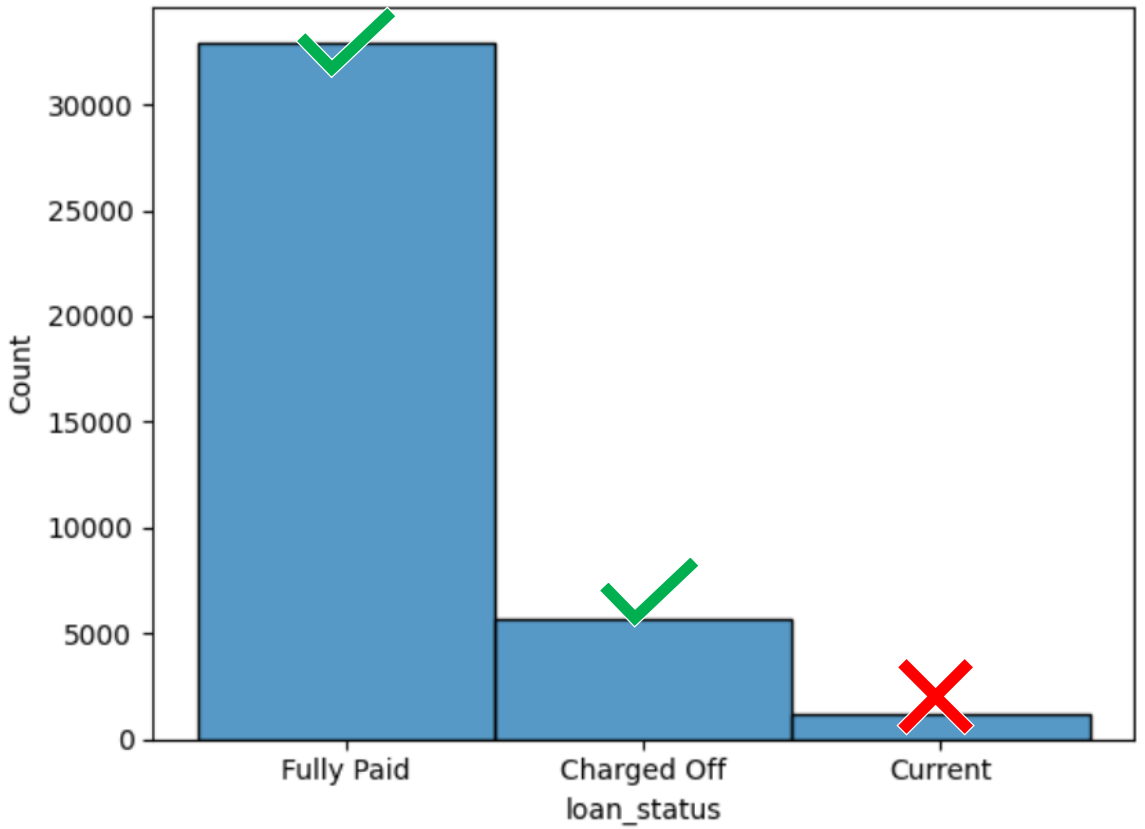
Observations:

- Month and year can be extracted into separate columns.
- Issue date spans from June 2007 to December 2011.
- Loan issuance increased annually from 2007 to 2011.
- Each year, loan issuance increased steadily from January to December, except for 2007 and 2008.

Loan Status

(loan_status)

Current Status of the loan



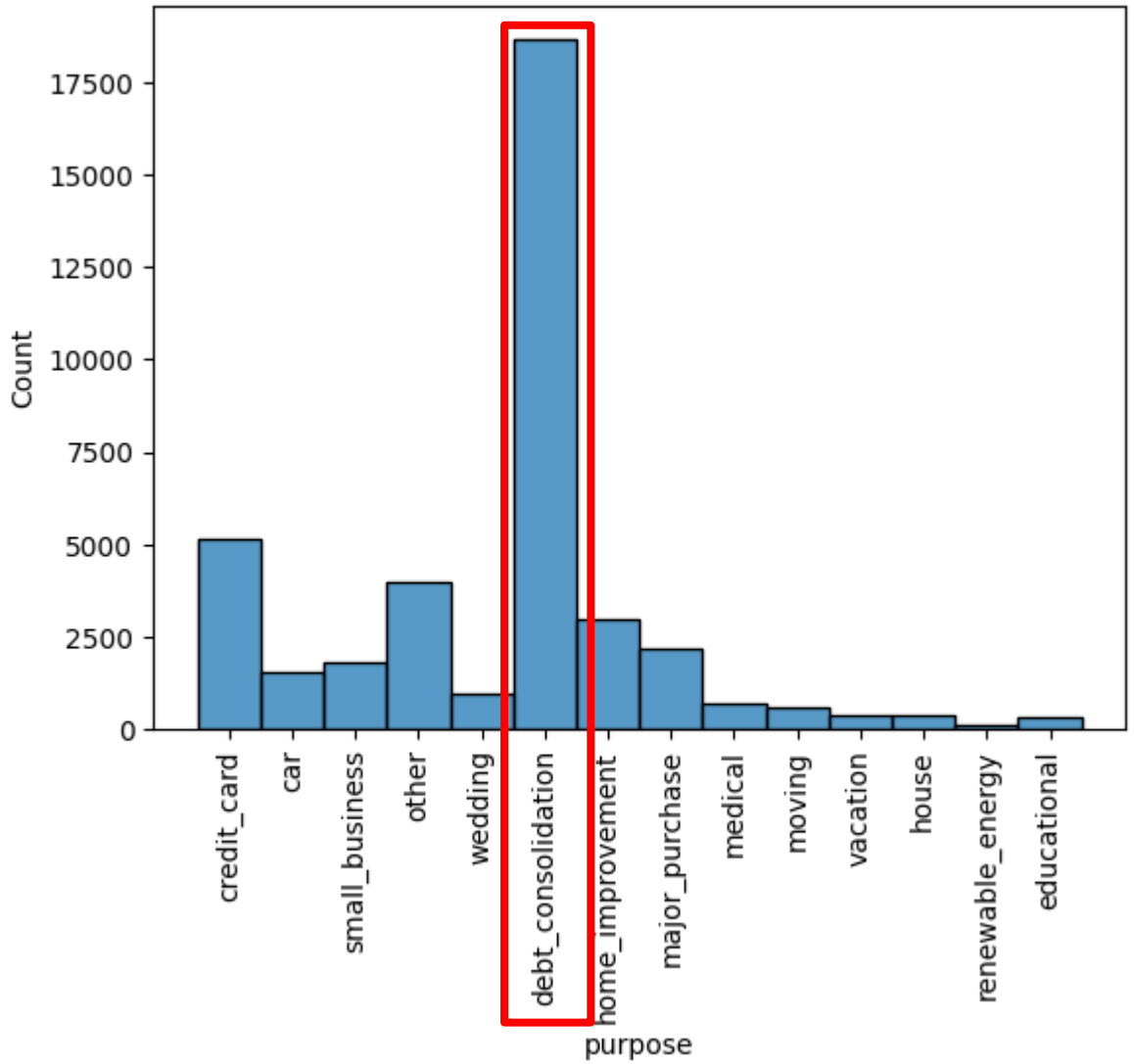
Observations:

1. 83% borrowers fully paid, 14% defaulted, 3% in currently active.
2. Active borrowers' data irrelevant, will be dropped

Purpose

(purpose)

A category provided by the borrower for the loan request.



Observations:

1. Over 60% loans for debt consolidation or credit card payment.

Earliest Credit Line

(earliest_cr_line)

The month the borrower's earliest reported credit line was opened.

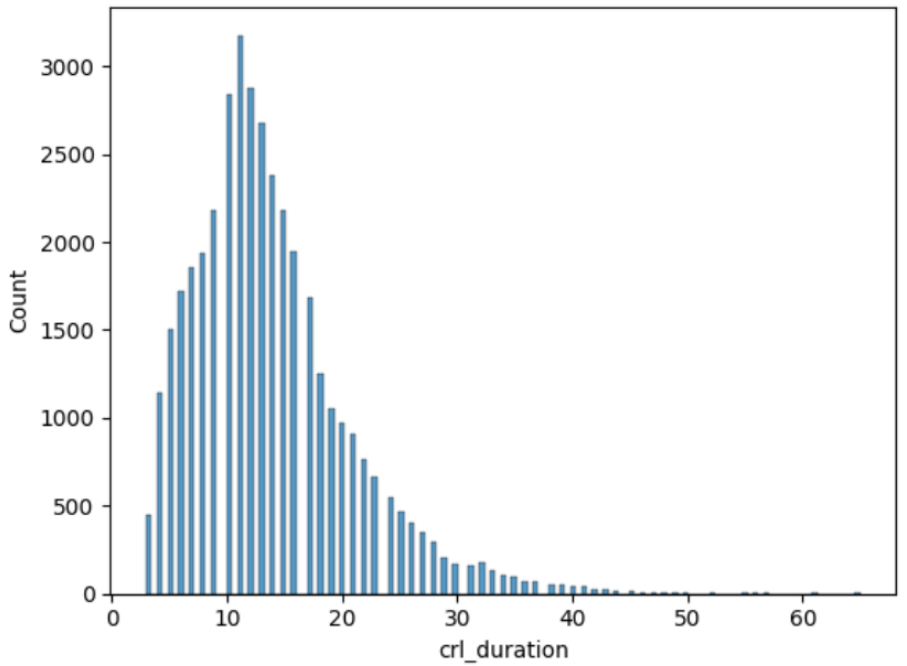
	earliest_cr_line
0	Jan-85
1	Apr-99
2	Nov-01
3	Feb-96
4	Jan-96

	earliest_crl_mnt	earliest_crl_yr
0	Jan	1985
1	Apr	1999
2	Nov	2001
3	Feb	1996
4	Jan	1996

	earliest_crl_yr	issue_yr	crl_duration
0	1985	2011	26
1	1999	2011	12
2	2001	2011	10
3	1996	2011	15
4	1996	2011	15

Observations:

1. Derive length of credit history (len_cr_hist) from issue month and year.
2. Create crl_duration by subtracting earliest credit line year from issue year.
3. Credit line duration distribution is highly skewed.



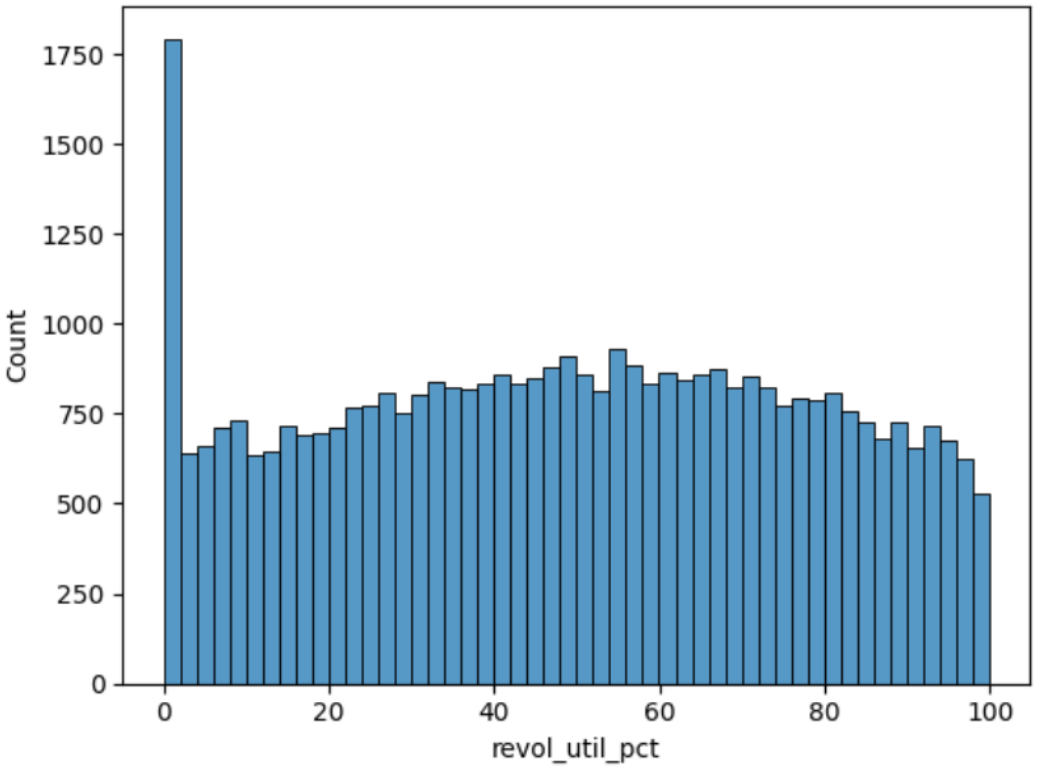
Revolving Utilization

(revol_util)

Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

	revol_util
0	83.70%
1	9.40%
2	98.50%
3	21%
4	53.90%

	revol_util_pct
0	83.70
1	9.40
2	98.50
3	21.00
4	53.90



Observations:

1. revol_util converted to floating point as revol_util_pct for analysis.
2. Distribution uniform except for 0% utilization cases.
3. 50 null occurrences filled with 0, the most frequent value.

Impactful Column Selection

Out of 111 Columns

Modified or Used

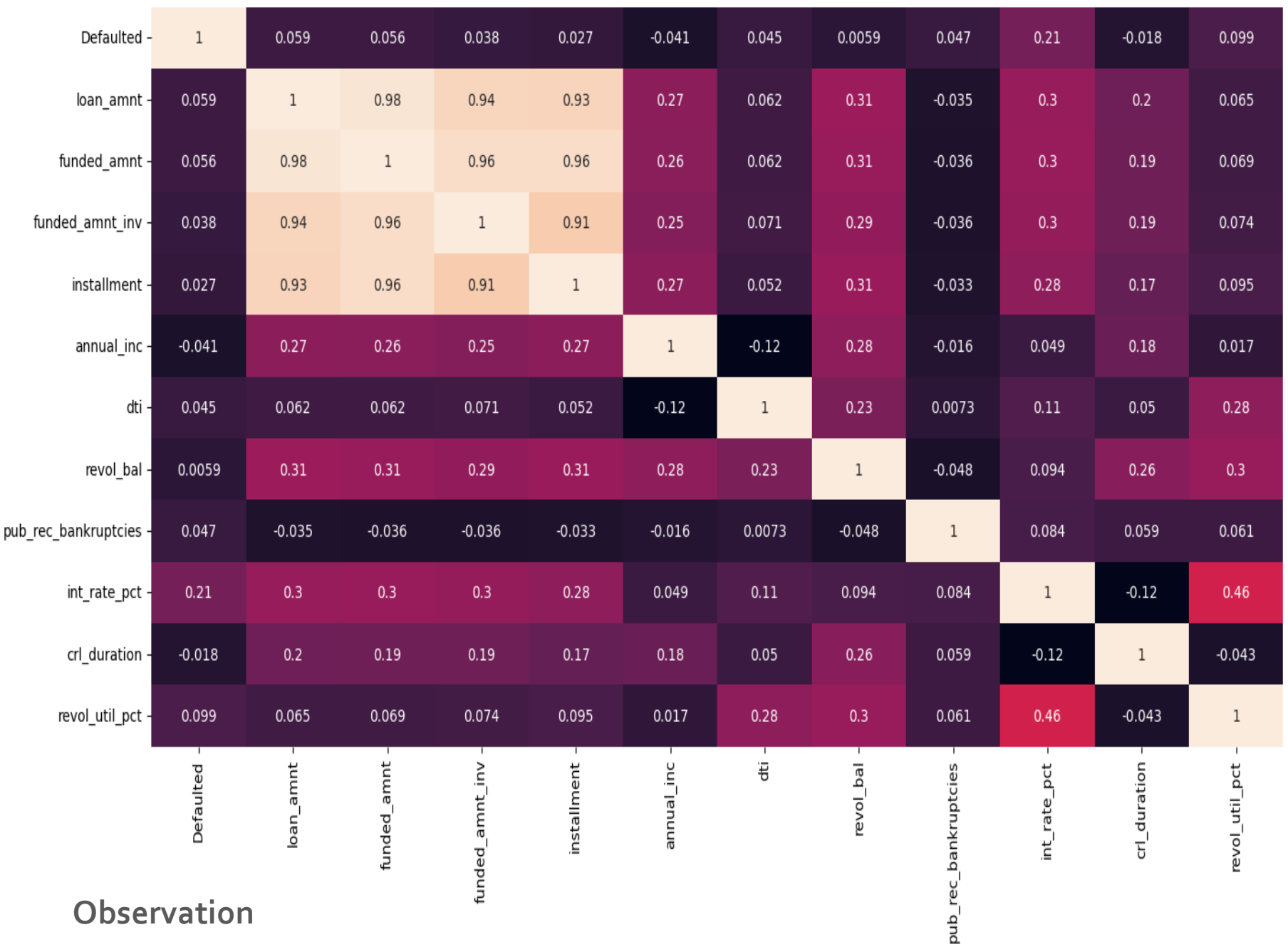
- loan_amnt
- funded_amnt
- funded_amnt_inv
- term
- int_rate_cat
- int_rate_pct
- installment
- grade
- subgrade
- emp_length
- home_ownership
- annual_inc
- verification_status
- issue_mnt
- issue_yr
- **loan_status**
- purpose
- addr_state
- dti
- delinq_2yrs
- earliest_crl_mnt
- earliest_crl_yr
- crl_duration
- inq_last_6mths
- open_acc
- pub_rec
- revol_bal
- revol_util_pct
- pub_rec_bankruptcies

Dropped or Not Used

- Dropped 54 Null Valued Columns.
- id
- member_id
- int_rate
- emp_title
- issue_d
- pymnt_plan
- url
- desc
- title
- zip_code
- earliest_cr_line
- mths_since_last_delinq
- nxt_pymnt_d
- mths_since_last_record
- initial_list_status
- collections_12_mths_ex_med
- policy_code
- application_type
- acc_now_delinq
- chargeoff_within_12_mths
- delinq_amnt
- tax_liens
- 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_amnt
- last_pymnt_d
- last_credit_pull_d
- revol_util

Correlation Matrix

(Heatmap)



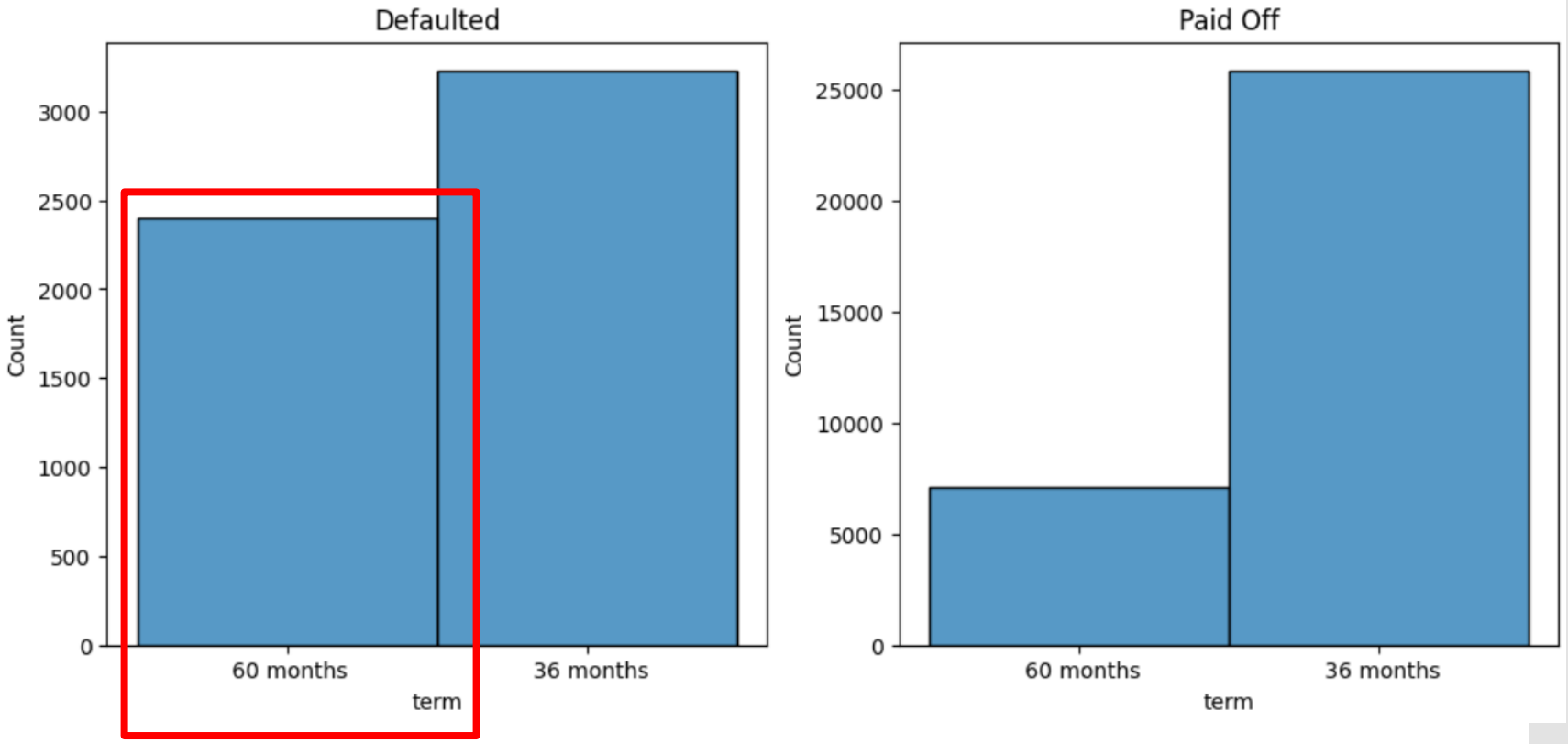
Observation

- No strong correlations observed with default in the heatmap.

Bivariate Analysis

Explore how attributes relate to loan default through paired analysis.

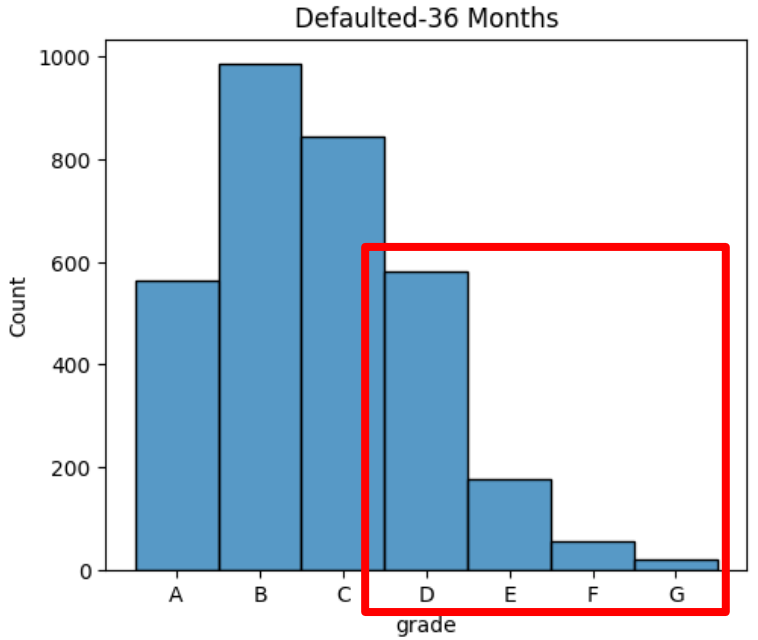
Term vs Loan Status

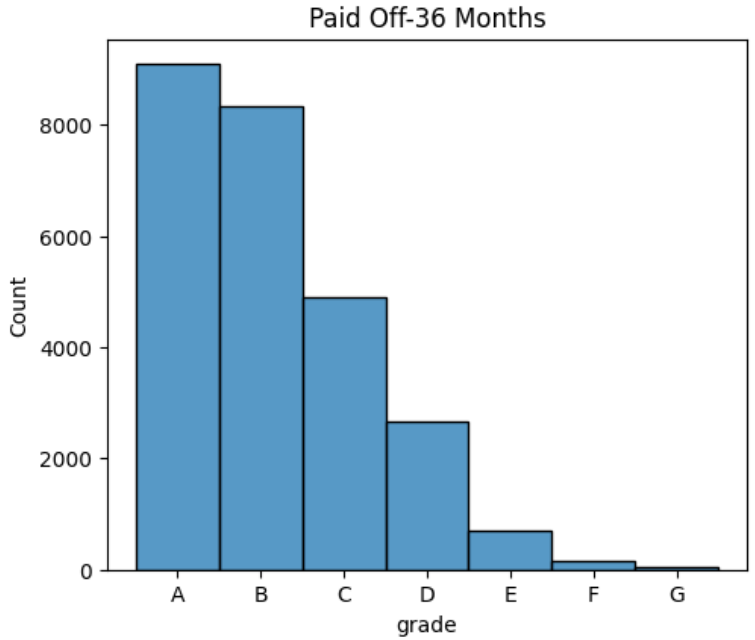


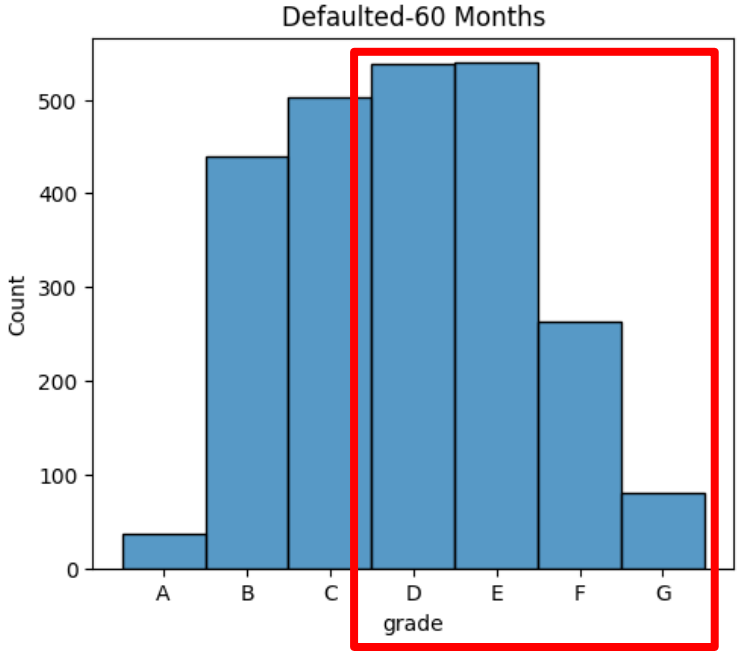
Observations:

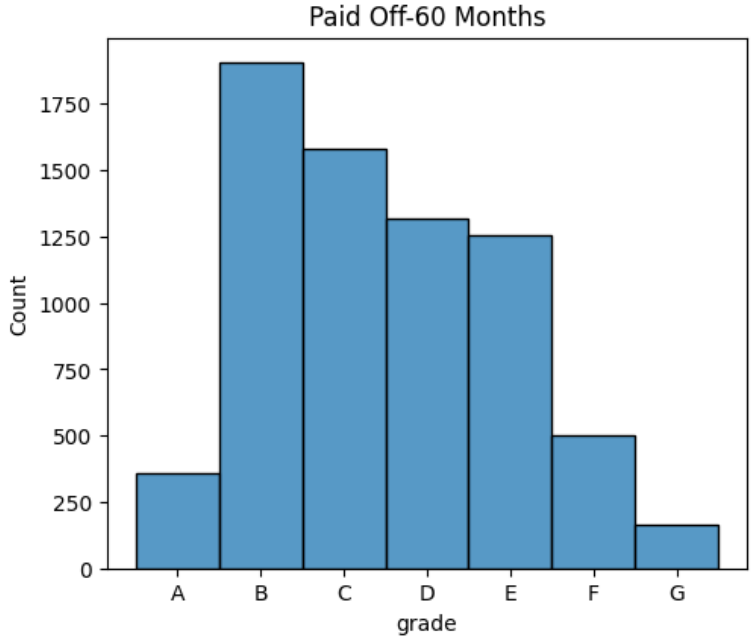
- Defaulters show higher incidence of long-term loans than fully paid borrowers.

Grade vs Term on Loan Status





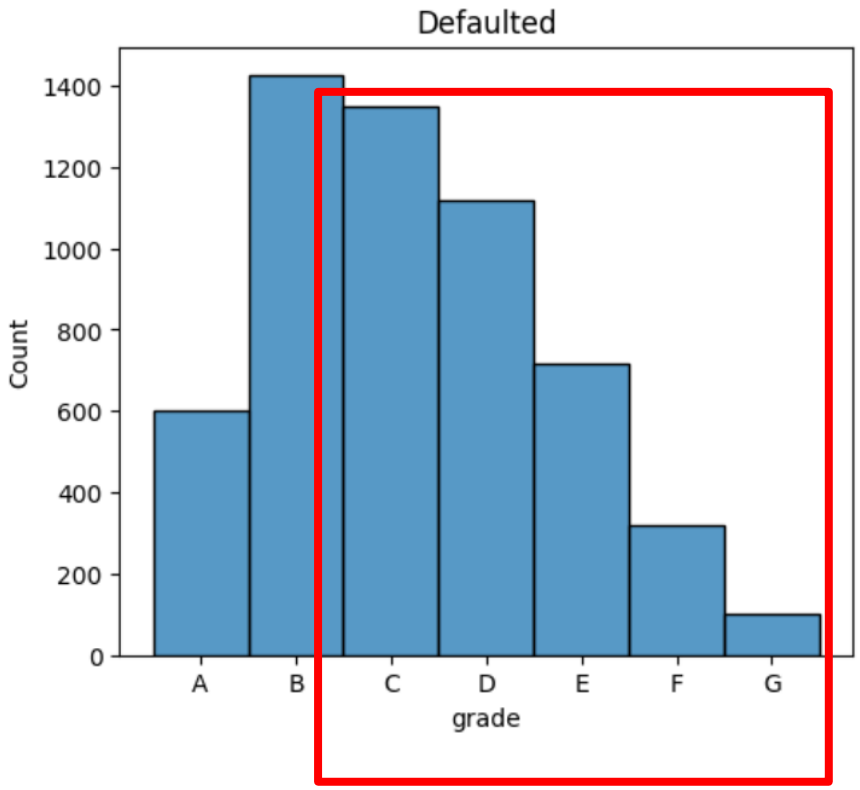


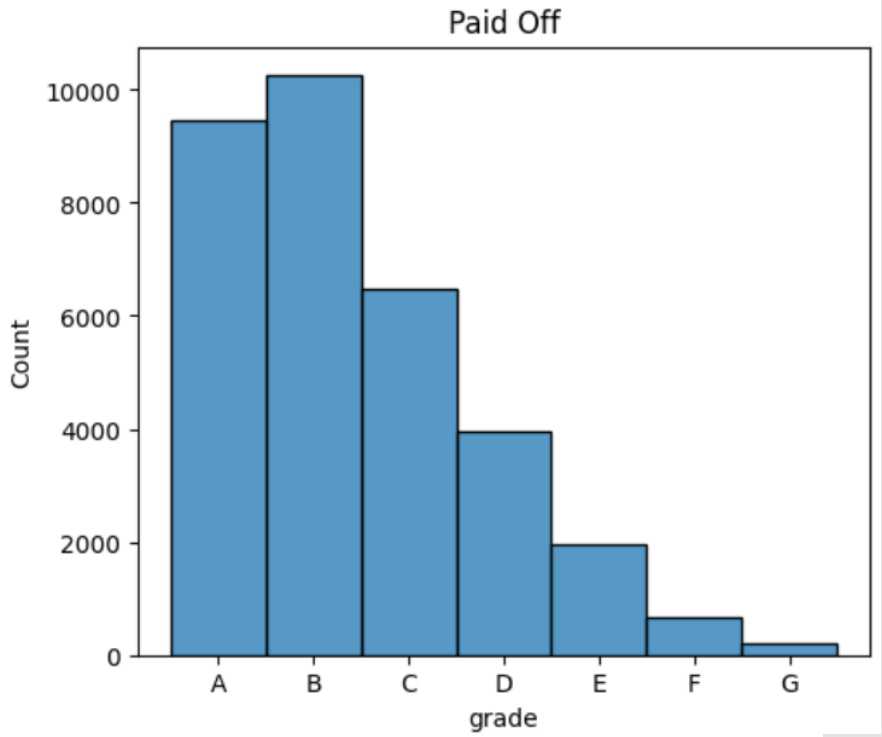


Observati

- Customers categorized as low graded opt for Long Term more and have eventually defaulted more as well.

Grade vs Loan Status

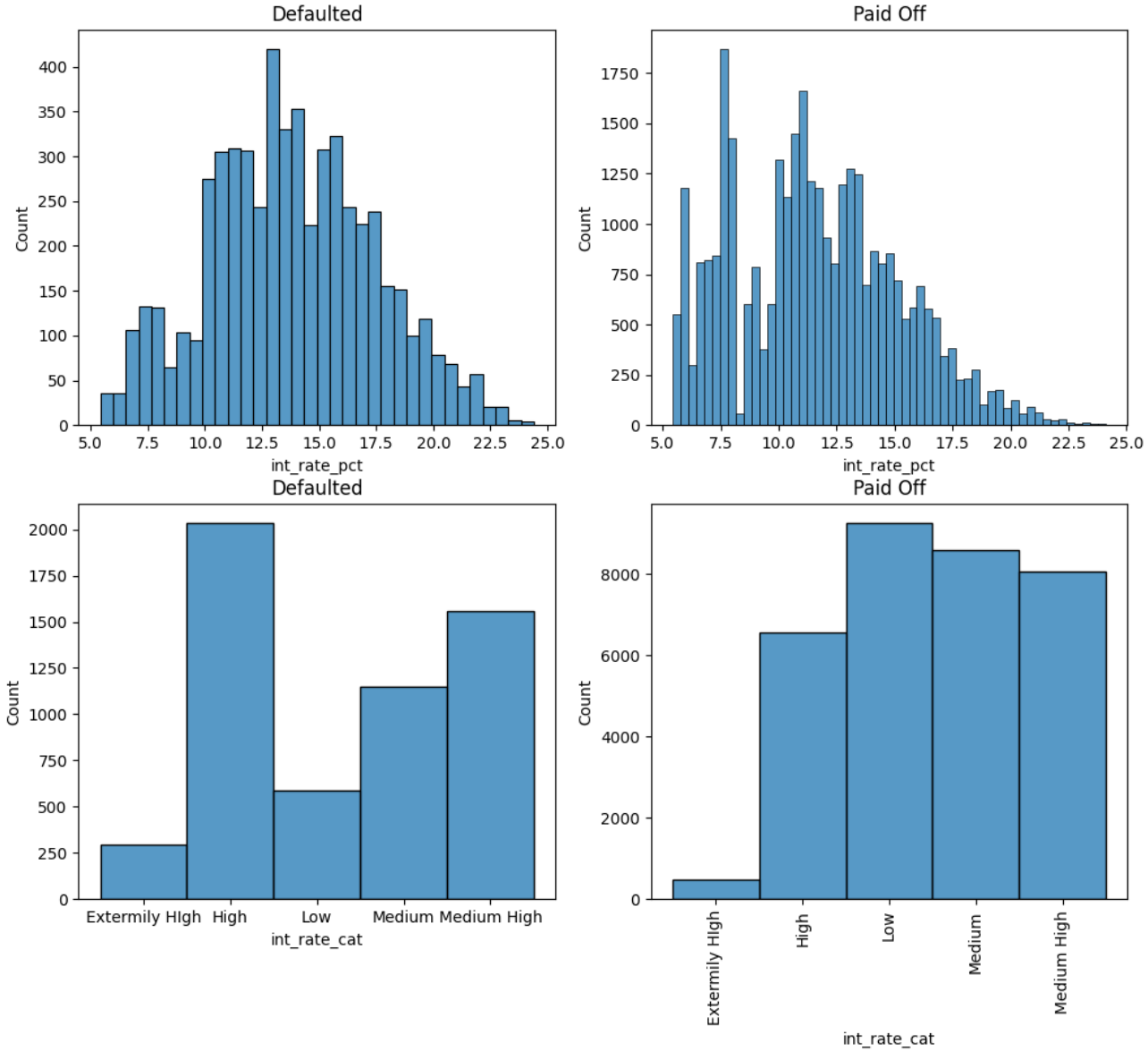




Observations:

- Lower the grades higher is the defaulted ratio.

Interest Rate vs Loan Status

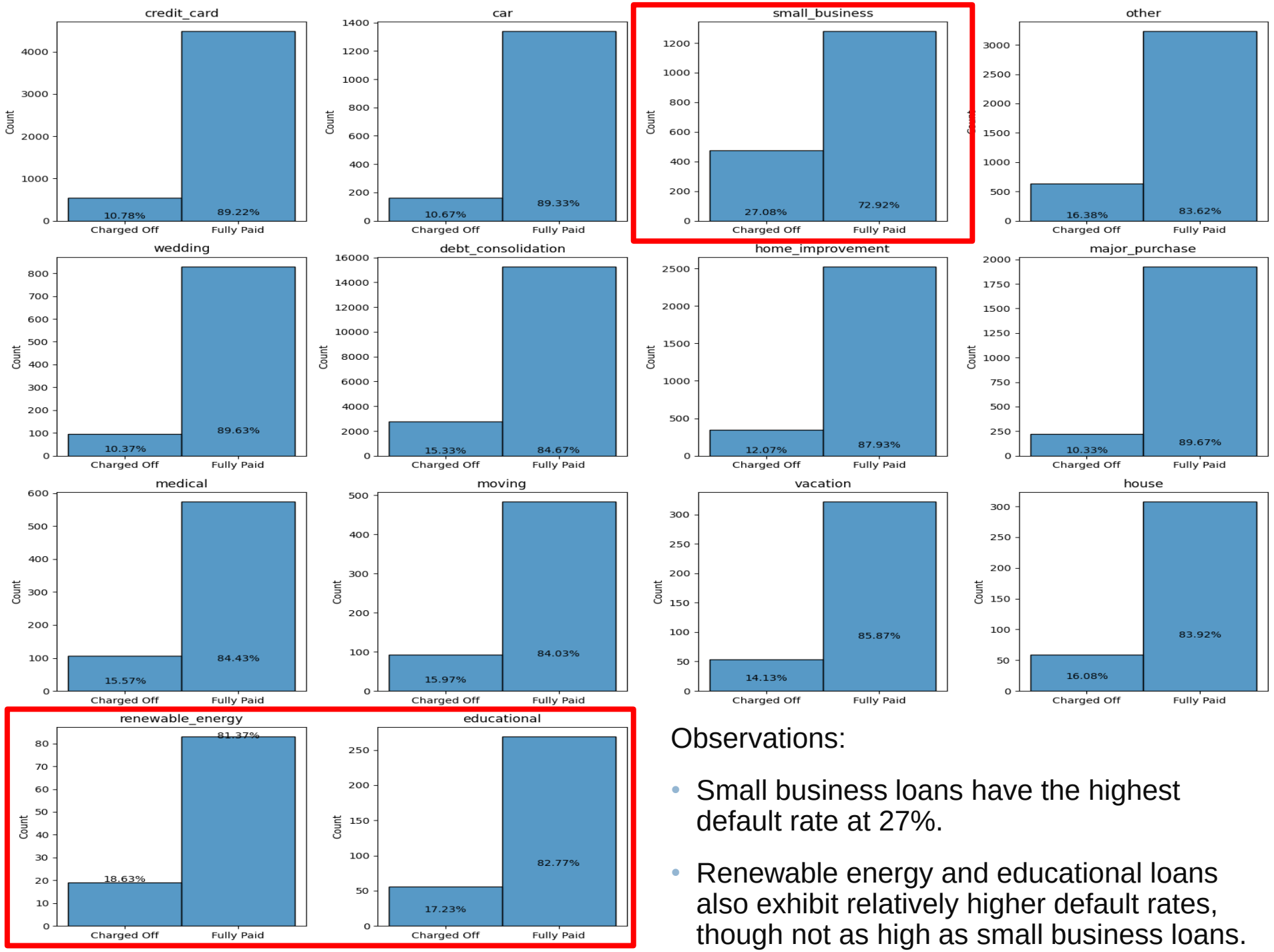


Observations:

Defaulted portfolio exhibits fatter tails in interest rate distribution, suggesting a potential link to higher default likelihood.

- Higher interest rates may indicate higher default risk, but financial institutions often charge more to compensate for risk.
- **Hence, Interest rates alone are not reliable predictors of default.**

Purpose vs Loan Status



Observations:

- Small business loans have the highest default rate at 27%.
- Renewable energy and educational loans also exhibit relatively higher default rates, though not as high as small business loans.

State vs Loan Status

State	Loan Status	Proportion (%)
AZ	Fully Paid	85.51
	Charged Off	14.49
GA	Fully Paid	84.18
	Charged Off	15.82
IL	Fully Paid	86.67
	Charged Off	13.33
CA	Fully Paid	83.81
	Charged Off	16.19
NC	Fully Paid	84.8
	Charged Off	15.2
TX	Fully Paid	88.12
	Charged Off	11.88
VA	Fully Paid	87.07
	Charged Off	12.93
MO	**Fully Paid**	**82.99**
	Charged Off	**17.01**
CT	Fully Paid	87.05
	Charged Off	12.95
UT	Fully Paid	84.13
	Charged Off	15.87
FL	**Fully Paid**	**81.88**
	Charged Off	**18.12**
NY	Fully Paid	86.61
	Charged Off	13.39
PA	Fully Paid	87.74
	Charged Off	12.26
MN	Fully Paid	86.61
	Charged Off	13.39
NJ	Fully Paid	84.47
	Charged Off	15.53

State	Loan Status	Proportion (%)
OR	Fully Paid	83.68
	Charged Off	16.32
KY	Fully Paid	85.53
	Charged Off	14.47
OH	Fully Paid	86.84
	Charged Off	13.16
SC	Fully Paid	85.62
	Charged Off	14.38
RI	Fully Paid	87.11
	Charged Off	12.89
LA	Fully Paid	87.59
	Charged Off	12.41
MA	Fully Paid	87.74
	Charged Off	12.26
WA	Fully Paid	84.47
	Charged Off	15.53
WI	Fully Paid	85.68
	Charged Off	14.32
AL	Fully Paid	87.59
	Charged Off	12.41
NV	**Fully Paid**	**77.45**
	Charged Off	**22.55**
AK	**Fully Paid**	**80.77**
	Charged Off	**19.23**
CO	Fully Paid	87.21
	Charged Off	12.79
MD	Fully Paid	84.16
	Charged Off	15.84
WV	Fully Paid	87.79
	Charged Off	12.21

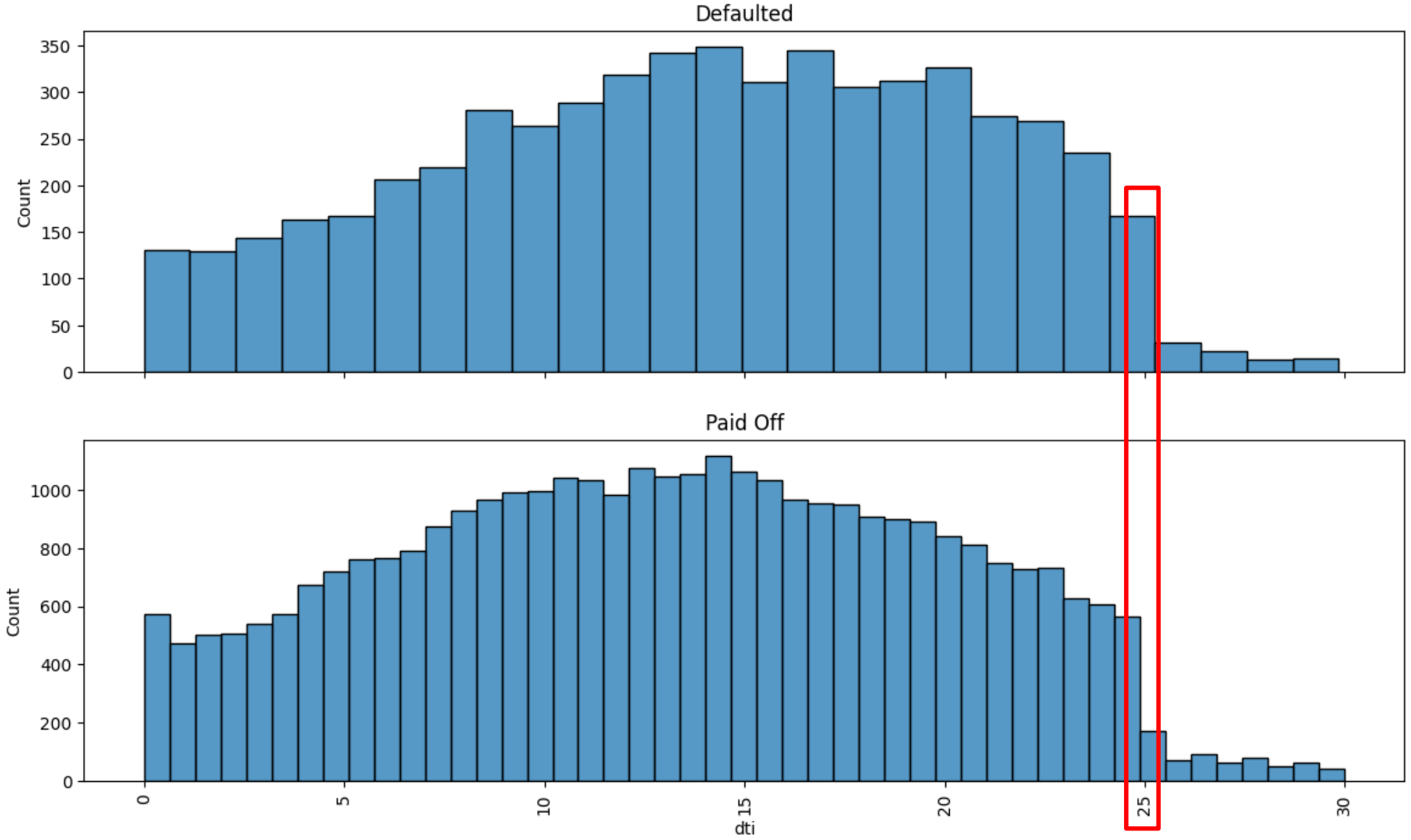
State	Loan Status	Proportion (%)
VT	Fully Paid	88.68
	Charged Off	11.32
MI	Fully Paid	85.37
	Charged Off	14.63
DC	Fully Paid	92.89
	Charged Off	7.11
SD	**Fully Paid**	**80.65**
	Charged Off	**19.35**
NH	Fully Paid	84.94
	Charged Off	15.06
AR	Fully Paid	88.51
	Charged Off	11.49
NM	Fully Paid	83.61
	Charged Off	16.39
KS	Fully Paid	87.84
	Charged Off	12.16
HI	Fully Paid	83.13
	Charged Off	16.87
OK	Fully Paid	86.06
	Charged Off	13.94
MT	Fully Paid	86.75
	Charged Off	13.25
WY	Fully Paid	95
	Charged Off	5
DE	Fully Paid	89.38
	Charged Off	10.62
MS	Fully Paid	89.47
	Charged Off	10.53
TN	Fully Paid	88.24
	Charged Off	11.76

State	Loan Status	Proportion (%)
IA	Fully Paid	100
NE	**Fully Paid**	**40**
	Charged Off	**60**
ID	Fully Paid	83.33
	Charged Off	16.67
IN	Fully Paid	100
ME	Fully Paid	100

Observations:

- MO, FL, NV, AK, SD, NE have higher default rates. NV and NE stand out with the highest default ratios: 22% and 40%, respectively.

DTI vs Loan Status



Observations:

- The DTI distribution for defaulted and paid-off portfolios is similar, suggesting it may not be a strong default indicator. However, an interesting pattern emerges: a cliff effect beyond a DTI of 25, indicating fewer loans given to borrowers with higher DTI.

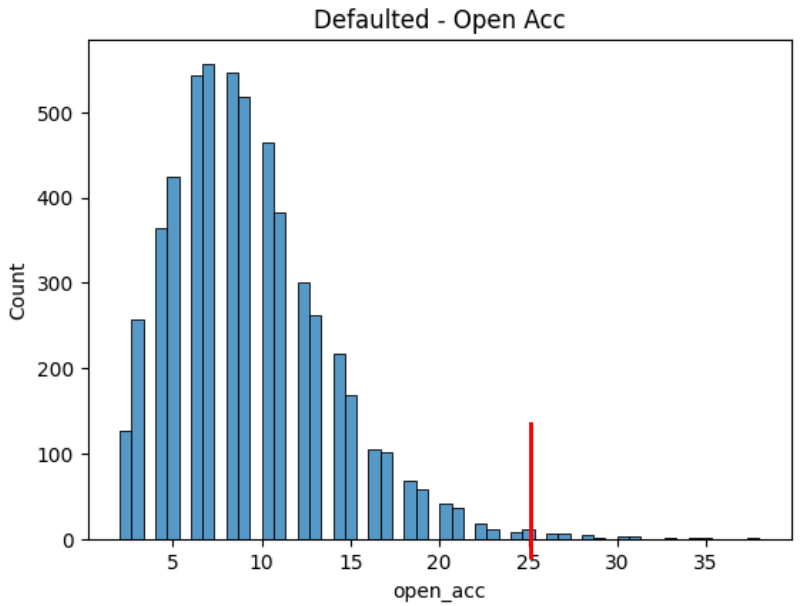
Number of Inquiries vs Loan Status

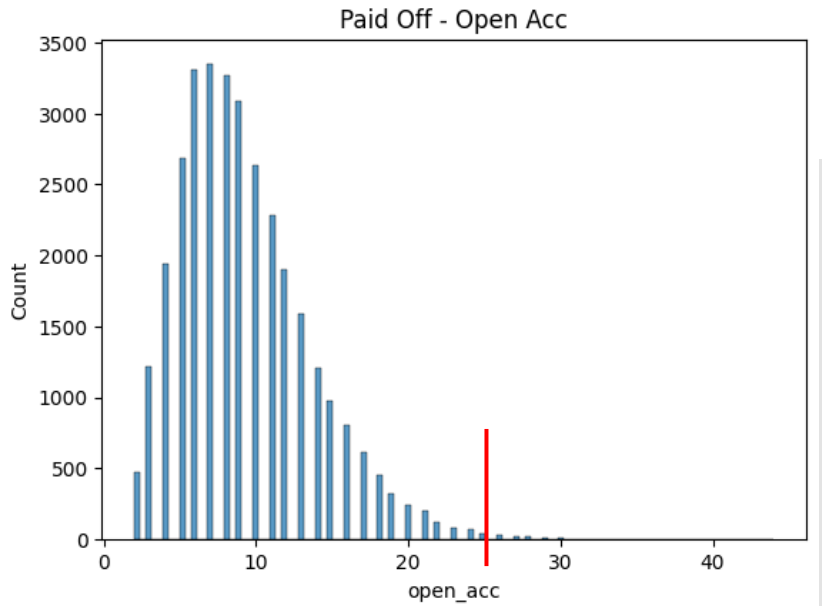
Number of Inquiries	Loan Status	Proportion (%)
0	Fully Paid	87.81
	Charged Off	12.19
1	Fully Paid	84.27
	Charged Off	15.73
2	Fully Paid	83.32
	Charged Off	16.68
3	Fully Paid	79.25
	Charged Off	20.75
4	Fully Paid	83.86
	Charged Off	16.14
5	Fully Paid	80.56
	Charged Off	19.44
6	**Fully Paid**	**74.6**
	Charged Off	**25.4**
7	Fully Paid	70.59
	Charged Off	29.41
8	**Fully Paid**	**78.57**
	Charged Off	**21.43**

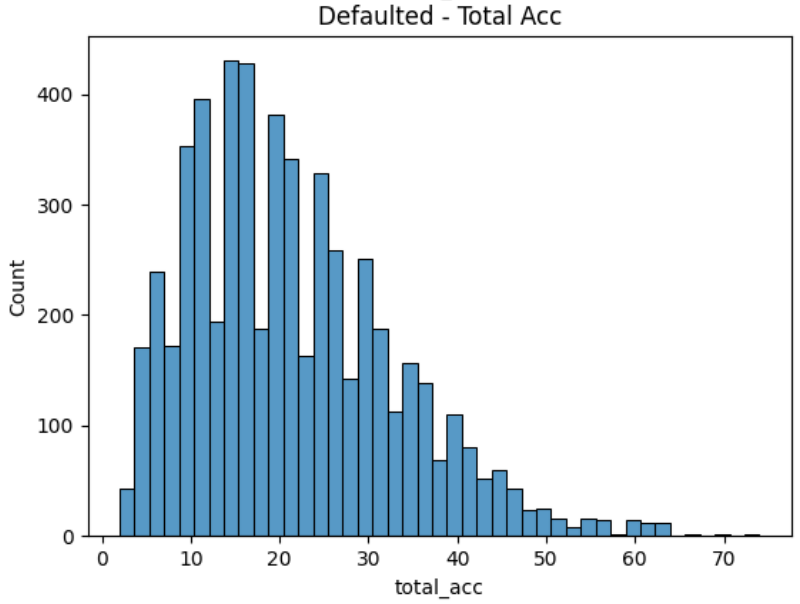
Observations:

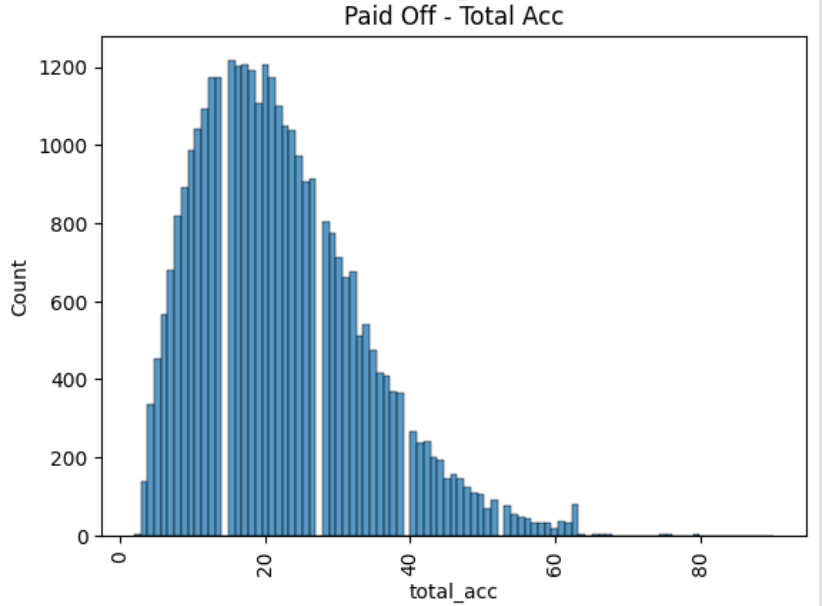
- Borrowers with 6 or more inquiries in the last 6 months show a significantly higher default percentage.

Open and Total Account vs Loan Status









Observations:

- The distribution of total accounts is similar for both defaulted and paid-off portfolios, indicating it may not be a reliable indicator for default.
- The distribution of open accounts displays fatter tails for defaulted portfolios compared to paid-off portfolios. Further analysis reveals a significant increase in default rates when the number of open accounts exceeds 25.

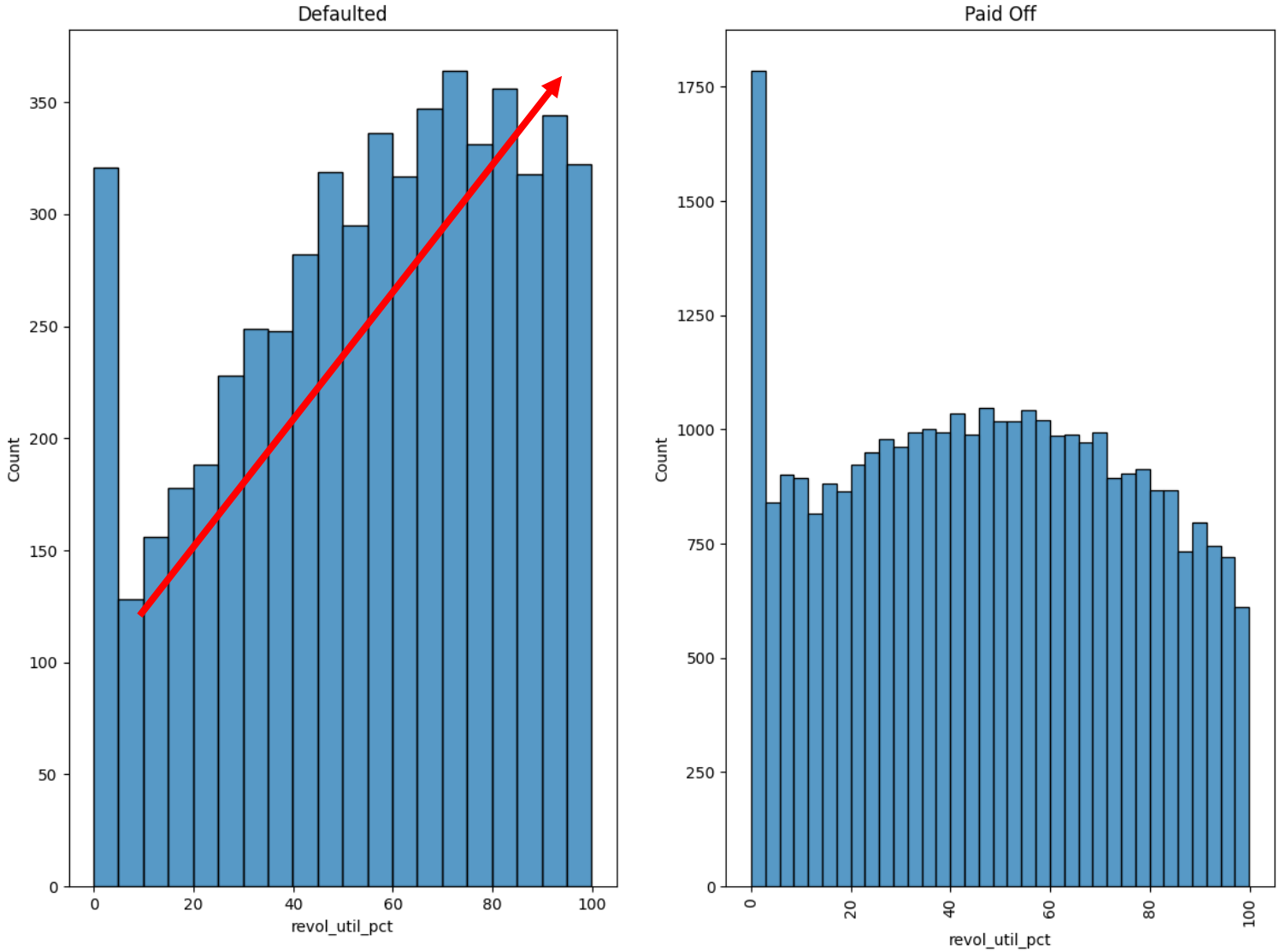
Public Record Bankruptcies vs Loan Status

Publicly Recorded Bankruptcy	Loan Status	Proportion (%)
0	Fully Paid	85.76
	Charged Off	14.24
1	Fully Paid	77.64
	Charged Off	22.36
2	Fully Paid	60
	Charged Off	40

Observations:

- The Publicly recorded bankruptcies serve as a reliable indicator for default, as the default rate demonstrates a direct correlation with the number of publicly recorded bankruptcies.

Revolving Utilization vs Loan Status



Observations:

- There is a significant increase in the default ratio when the revolving utilization exceeds 80%.

Conclusion

- Term of the loan is an important driving factor for default, shorter the maturity of the loan less risky it will be.
- Grade is also a driving factor for the default, lower the grade higher is the chance of default.
- Loan given for setting up small business carries the highest amount of risk followed by loan for renewable energies and education.
- Following states show higher default ratio:
  - MO, FL, NV, AK, SD, NE

  Among these NV and NE shows the highest default ratio.
- Borrower who made 6 or more inquires in last 6 months shows a higher tendency of default.
- Borrower who has 25 or more open account shows a higher tendency of default.
- Borrower with 90% or more revolving utilization shows a higher tendency of default.
- If a customer has any public record of bankruptcy then there is a much higher likelihood of default for that borrower.
- Interest rate may look like as if it is good factor to identify default, however one must not take the higher correlation between high interest rate and default rate for the causation of default. Often financial institutes charge higher interest rate to high risk customer to compensate for the added risk, and since most of the time it is these high risk customer that default, we see a positive relationship between high interest rate and default, however interest rate is not a leading indicator of default but rather a lagging one.

Recommendation

- The company should increase its long term loan exposure with better grade customer and at the same time should reduce its long term loan with the poor graded customer.
- The company should be extra cautious when approving loan for small business, if possible company should ask for collaterals if a poor graded borrower ask for business loan that too for long term.
- For few state like NV and NE where default rate is really high, company may consider closing its business there, given that these states gives a very small proportion of business to the company.
- Company should add additional checks based on higher number of open accounts, higher revolving utilization while approving a new loan.
- For customers with 2 or more publicly recorded bankruptcy the company should not approve the loan.