Lending Club Case Study

Exploratory Data Analysis

Supriya
Tamilarasan

Content

- Problem Statement
- Data Summary
- Data Cleaning
- Data Conversions
- Derived Columns
- Dropping or Imputing Rows
- Outliers
- Univariate Analysis
- Bivariate Analysis
- Conclusion

Problem Statement

- **Problem:** You work for a consumer finance company specializing in lending various types of loans to urban customers. When the company receives a loan application, it must decide whether to approve the loan based on the applicant's profile. There are two types of risks associated with the bank's decision:
- If the applicant is likely to repay the loan, not approving the loan results in a loss of business for the company.
- If the applicant is not likely to repay the loan (i.e., they are likely to default), approving the loan may lead to a financial loss for the company.
- **Objective:** Use Exploratory Data Analysis (EDA) to understand how consumer attributes and loan attributes influence the tendency of default.
- **Constraints:** When a person applies for a loan, there are two types of decisions that the company can make:

1. **Loan accepted:** If the company approves the loan, there are three possible scenarios:
    1. **Fully paid:** The applicant has fully paid the loan (both the principal and the interest).
    2. **Current:** The applicant is in the process of paying the installments, i.e., the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
    3. **Charged-off:** The applicant has not paid the installments in due time for a long period, i.e., they have defaulted on the loan.
2. **Loan rejected:** The company rejected the loan (because the candidate did not meet their requirements, etc.). Since the loan was rejected, there is no transactional history of those applicants with the company, and thus this data is not available in the dataset.

Data Understanding

Loan.csv file contains 39717 rows and 111 columns.

There are two types of attributes - Loan Attributes and Customer attributes.

Data Cleaning

Header, Footer, and Summary Rows: No header, footer, summary, or total rows were found in the dataset.

Duplicate Rows: No duplicate rows were detected.

Rows with 'loan_status' = 'current': 1,140 rows deleted, as they do not contribute to the analysis.

Columns with Null/Blank Values: 55 columns have been removed, as they do not participate in the analysis.

Unique Columns: 'url' and 'member_id' were unique and have been deleted. The 'id' column has been retained for potential future analysis.

Text/Description: 'desc' and 'title' columns, which contain text/description values, have been dropped.

Sub-group Analysis: The analysis is limited to the 'Group' level only; hence, the sub-group has been dropped.

Behavioral Data: 21 behavioral data columns have been deleted, as they were captured based on domain knowledge and are not available during loan approval, thus not contributing to the analysis.

Columns with Unique Values: 8 columns with values of 1, indicating uniqueness, have been dropped from the analysis.

Columns with High NA Values: Two columns with more than 50% missing data (NA) have been removed.

Final Dataset: After all data cleaning processes, 38,577 rows and 20 columns remain.

Data Manipulation: For the given dataset, No further data manipulation was required for the columns essential to Exploratory Data Analysis (EDA).

Data Conversions

'Term' : Additional string values were trimmed from the 'term' column, which has now been converted to an integer data type.

'Int Rate' : The 'int_rate' column was converted from a string to an integer data type, with the additional '%' symbol removed.

Loan and Funded Amount : The 'loan_funded_amnt' and 'funded_amnt' columns were converted to float data types.

Decimal Precision: The columns 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', and 'dti' were rounded to two decimal points for consistency.

Date Conversion: The 'issue_d' column has been converted to a date type.

Derived Columns

Derived Columns from 'issue_d': 'issue_year' and 'issue_month' were created to facilitate further analysis.

Range Columns: Derived columns—'loan_amnt_range', 'annual_inc_range', 'int_rate_range', 'installment_range', and 'dti_range'—were created by binning continuous data into multiple categories for enhanced analysis..

Dropping/Inputting the rows

The 'emp_length' and 'pub_rec_bankruptcies' columns contain 2.67% and 1.80% of rows with null values, respectively. Given the small percentage of missing data, these rows can be safely removed.

Total Rows Deleted: 4.48%.

Outlier Treatment

Outliers were identified and removed from the following columns: 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'installment', and 'annual_inc'.

Outlier Treatment: The quantile mechanism was applied to address outliers in the above fields.

Univariate Analysis Outcome

Loan Amount: Higher loan amounts are associated with a higher probability of the loan being charged off. Click here for Diagram

Interest Rates: Elevated interest rates correlate with an increased likelihood of loan charge-offs. Click here for Diagram

Installment Amount: Larger installment amounts result in a greater probability of loans being charged off. Click here for Diagram

Income: Higher income levels reduce the chances of a loan being charged off. Click here for Diagram

Debt-to-Income Ratio: An increased Debt-to-Income (DTI) ratio significantly raises the likelihood of loan charge-offs. Click here for Diagram

Loan Term: Loans with a 36-month term have a lower risk of being charged off. Hence, shorter loan terms reduce the probability of default. Click here for Diagram

Loan Grades: Loans with grades G, F, and E exhibit a higher risk of being charged off. Click here for Diagram

Employee Experience: No discernible pattern was identified in the chart based on years of employee experience. Click here for Diagram

Verification Process: Surprisingly, verified applicants exhibit a higher likelihood of their loans being charged off, indicating potential discrepancies in the verification process. Click here for Diagram

Purpose: Loans to small businesses present a higher risk of being charged off, with a charge-off rate exceeding 20%. Click here for Diagram

State-Specific Risk: The states of Nevada (NV), Tennessee (TN), and Alaska (AK) are identified as riskier, with charge-off rates exceeding 20%. Click here for Diagram

Bankruptcy: Applicants with a history of bankruptcy are more likely to default on their loans. Click here for Diagram

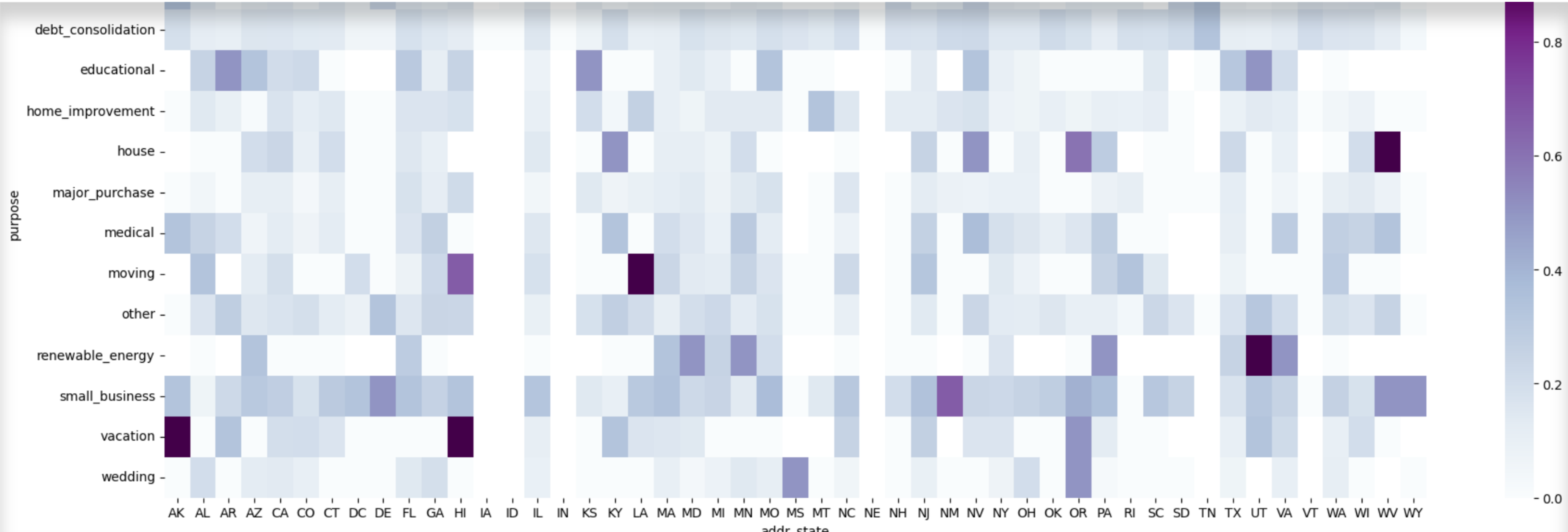
Loan Issuance by Month: Loans issued in December, May, September, and October have a higher likelihood of being charged off. Click here for Diagram

**Refer Annexture for detailed report: Annexture

Bivariate Analysis

State vs Loan Purpose

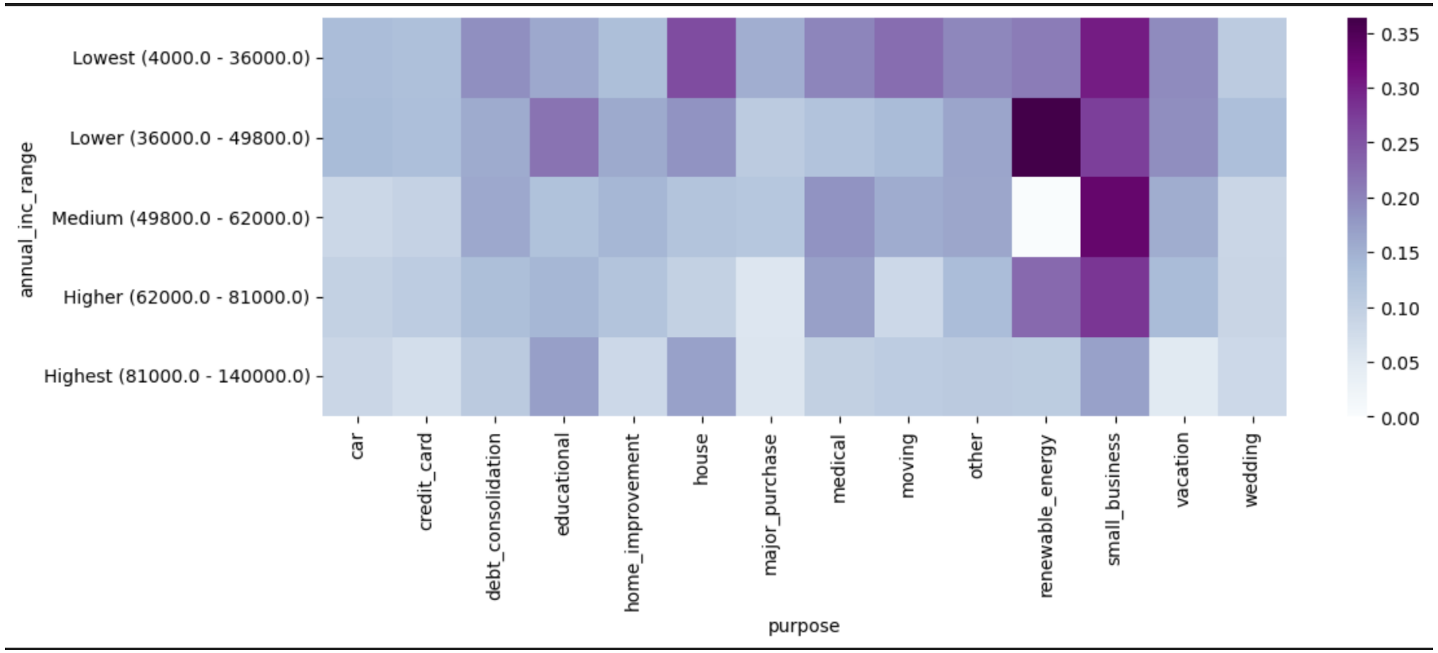
- The darker the intersection between a state's abbreviation (addr_state) and the purpose of the loan, the riskier the loan application. Key examples include:
  - Vacation Loans: High risk in AK, HI, OR
  - Education Loans: High risk in AR, KS, UT
  - Small Business Loans: High risk in DE, NM, WV
  - Renewable Energy Loans: High risk in UT, OR



Annual Income vs Purpose

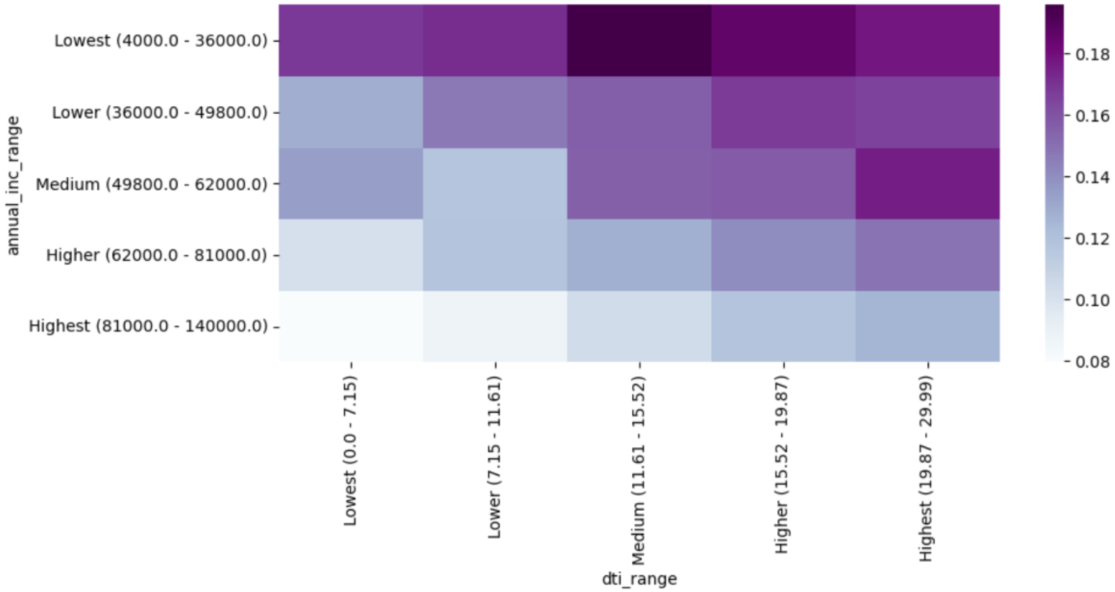
Some of the higher-risk loan applications include:

1. Small business loans for individuals in the lowest and middle-income brackets.
2. Renewable energy loans targeted at lower-income groups.



Annual Income vs debt-to-income

- Medium debt-to-income group in the lowest income range poses the highest risk



Recommendations

Addressing Loan Defaults: Implement measures to mitigate the risk of default among individuals with annual incomes between $4,000 and $36,000, as they exhibit a higher likelihood of defaulting on loans.

Improving Verification Processes: Reassess and enhance the verification process, since applicants who were verified have demonstrated higher rates of charge-offs.

Geographic Risk Assessment: Take note of the increased risk associated with loan applicants from Nevada (NV), Tennessee (TN), and Arkansas (AK), and adjust strategies accordingly.

Managing High Loan Amounts: Monitor and manage high loan amounts more effectively, as these are associated with increased instances of charge-offs.

Assessing Bankruptcy History: Exercise caution with applicants who have a history of bankruptcy, as they are identified as higher-risk individuals.

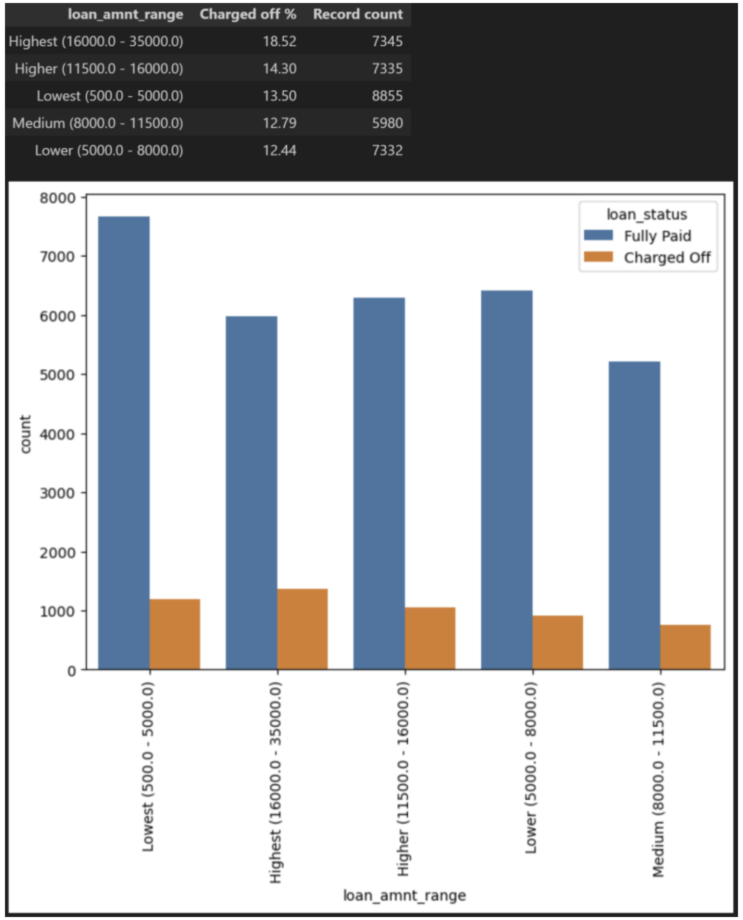
Evaluating Debt-to-Income Ratios: Pay closer attention to loan applicants with high debt-to-income ratios, as they have been shown to present a greater risk.

Focusing on Specific Loan Types: Be aware that small business loans for low and medium-income groups, as well as renewable energy loans for lower-income groups, have shown a tendency for higher charge-offs.

Thank you!

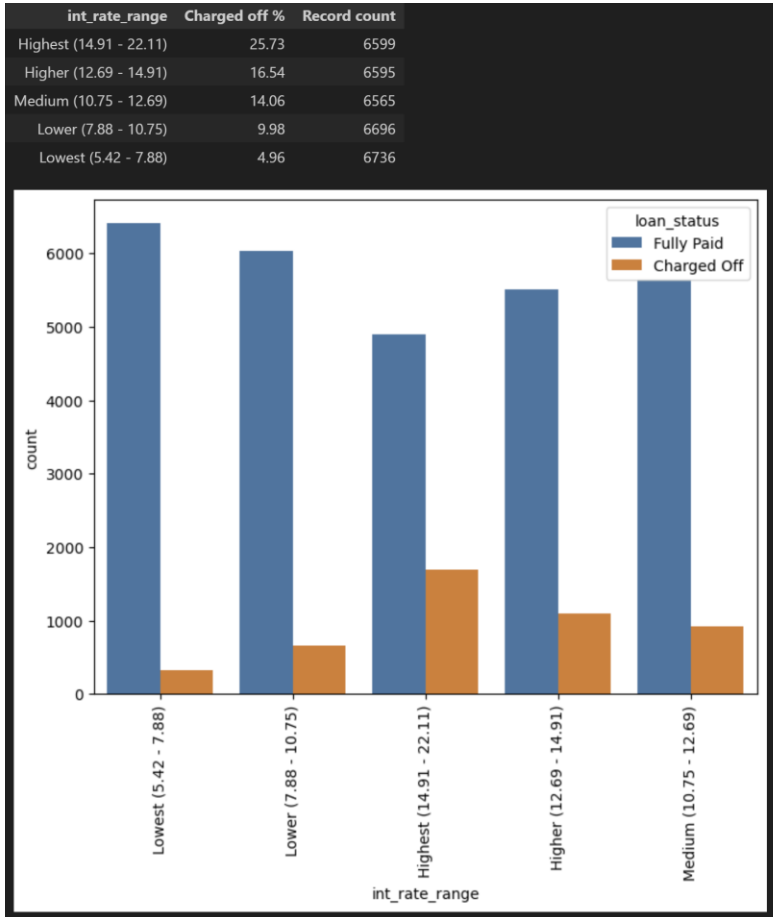
Annexture: Univariate Analysis Report

Loan Amount



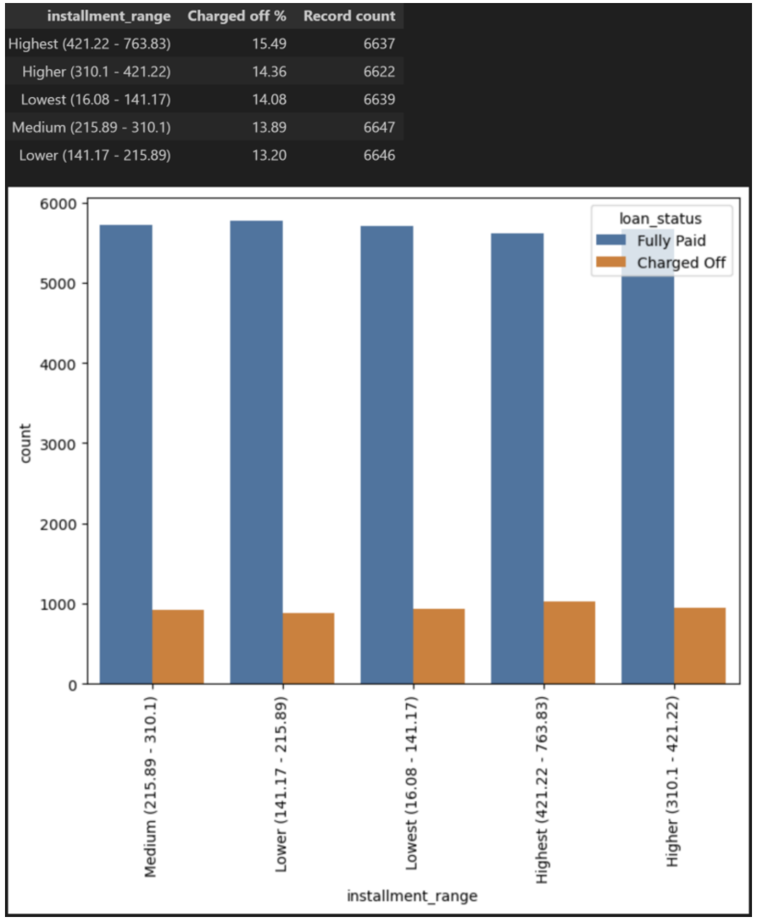
loan_amnt_range	Charged off %	Record count
Highest (16000.0 - 35000.0)	18.52	7345
Higher (11500.0 - 16000.0)	14.30	7335
Lowest (500.0 - 5000.0)	13.50	8855
Medium (8000.0 - 11500.0)	12.79	5980
Lower (5000.0 - 8000.0)	12.44	7332

Interest Rates



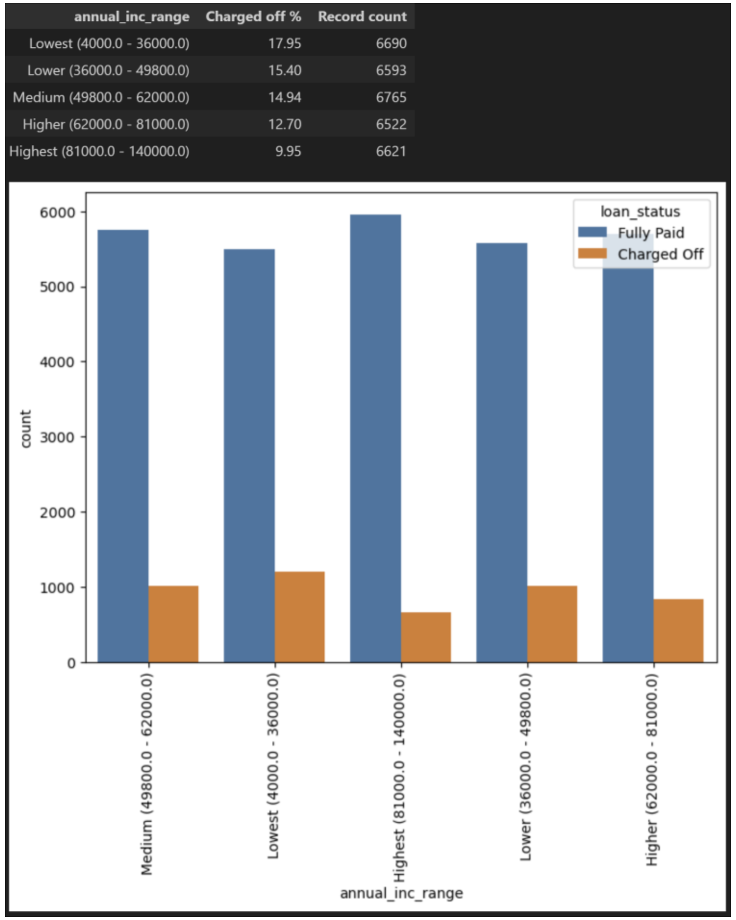
int_rate_range	Charged off %	Record count
Highest (14.91 - 22.11)	25.73	6599
Higher (12.69 - 14.91)	16.54	6595
Medium (10.75 - 12.69)	14.06	6565
Lower (7.88 - 10.75)	9.98	6696
Lowest (5.42 - 7.88)	4.96	6736

Installment Amount



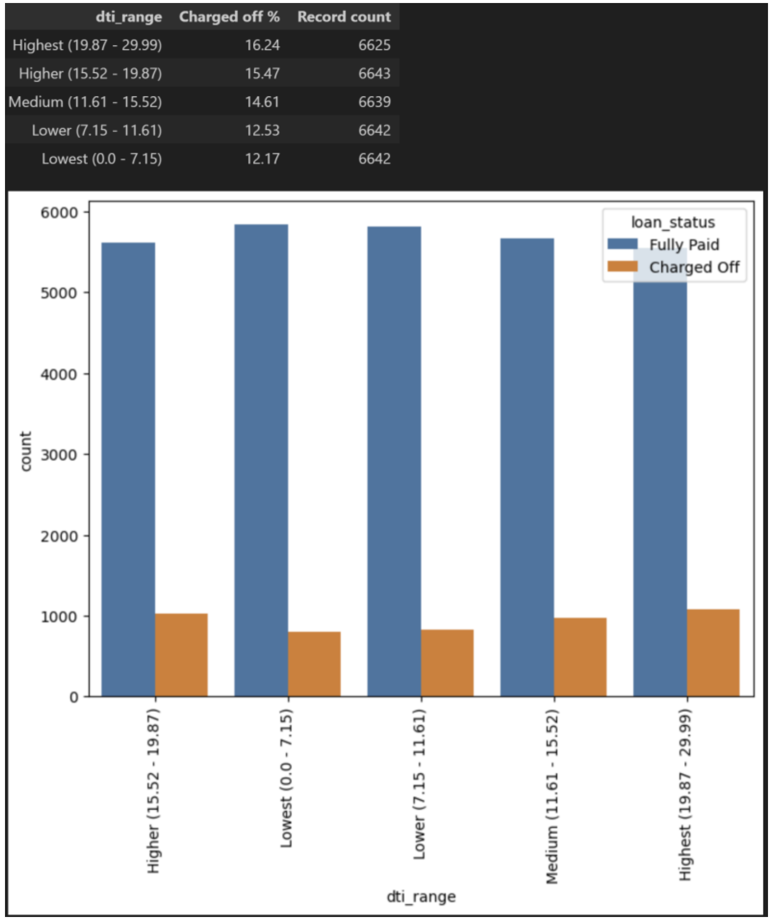
installment_range	Charged off %	Record count
Highest (421.22 - 763.83)	15.49	6637
Higher (310.1 - 421.22)	14.36	6622
Lowest (16.08 - 141.17)	14.08	6639
Medium (215.89 - 310.1)	13.89	6647
Lower (141.17 - 215.89)	13.20	6646

Income



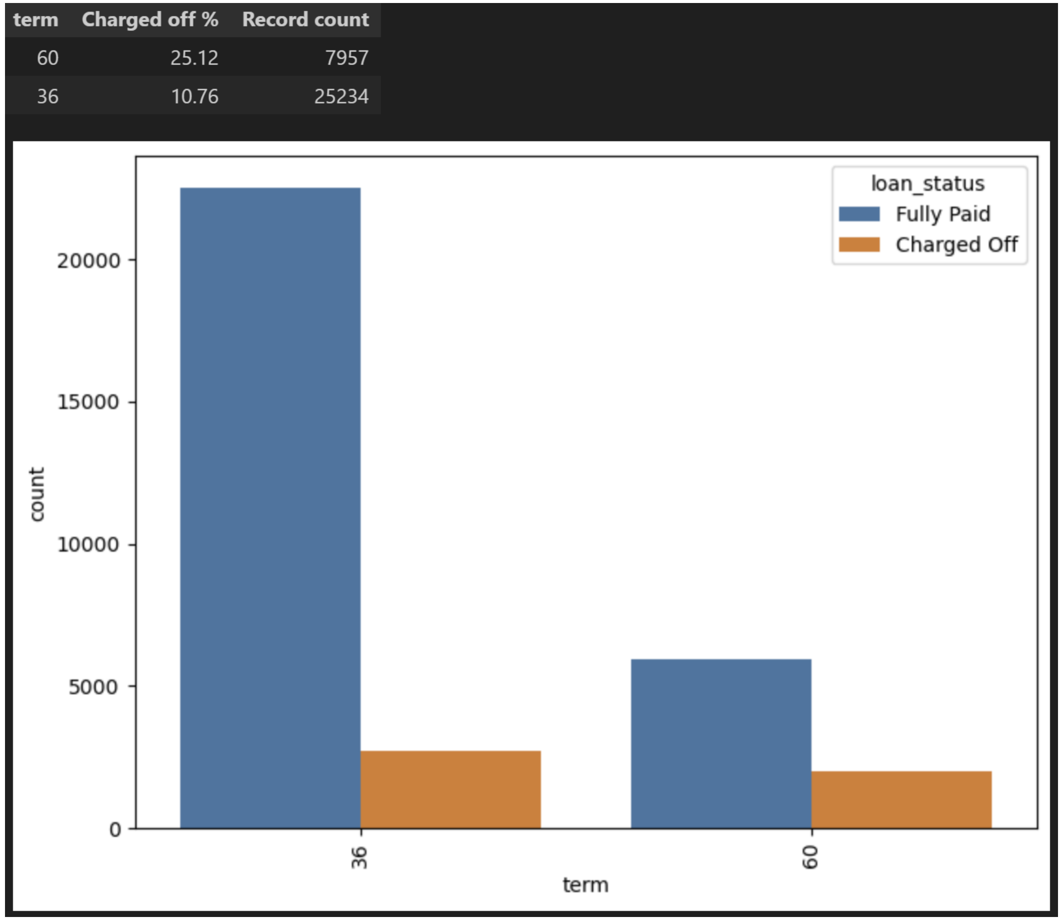
annual_inc_range	Charged off %	Record count
Lowest (4000.0 - 36000.0)	17.95	6690
Lower (36000.0 - 49800.0)	15.40	6593
Medium (49800.0 - 62000.0)	14.94	6765
Higher (62000.0 - 81000.0)	12.70	6522
Highest (81000.0 - 140000.0)	9.95	6621

Debt-to-Income Ratio



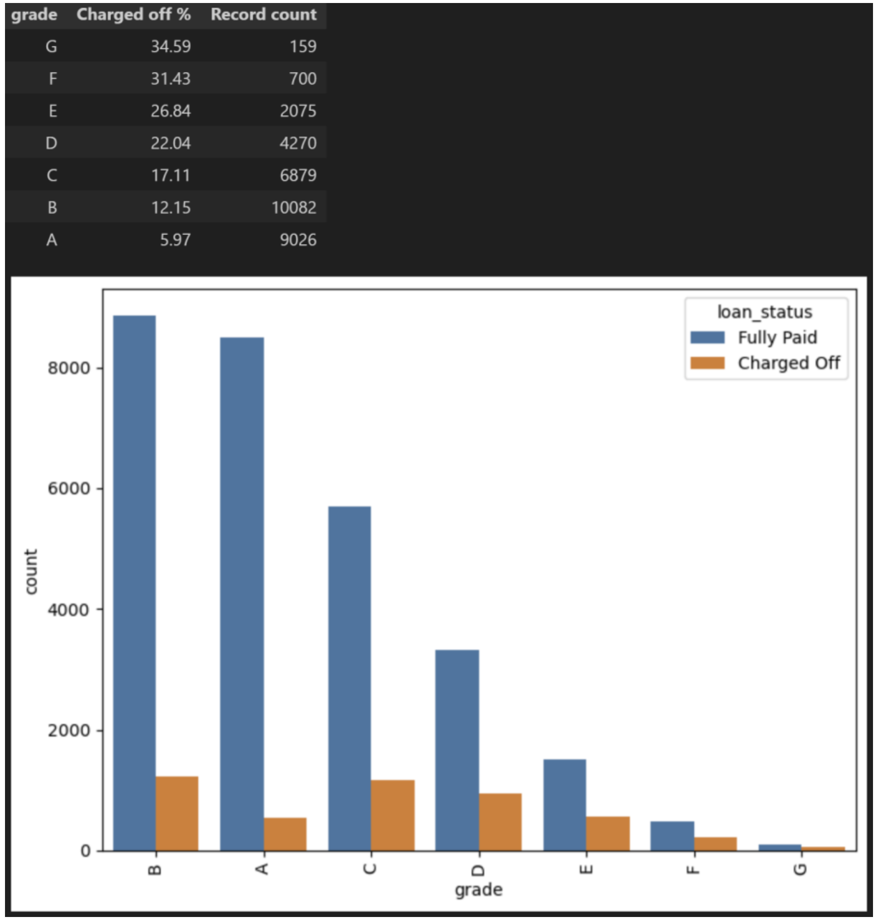
dti_range	Charged off %	Record count
Highest (19.87 - 29.99)	16.24	6625
Higher (15.52 - 19.87)	15.47	6643
Medium (11.61 - 15.52)	14.61	6639
Lower (7.15 - 11.61)	12.53	6642
Lowest (0.0 - 7.15)	12.17	6642

Loan Term



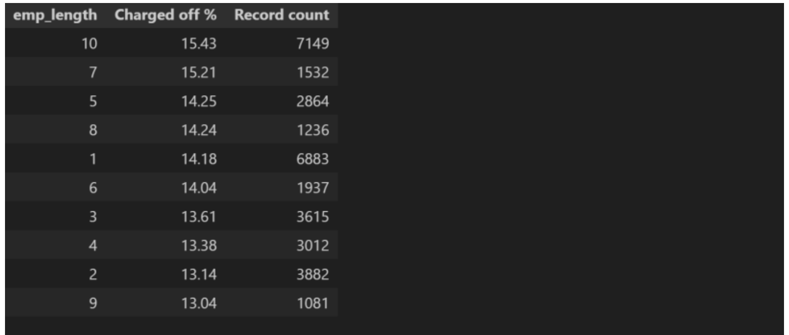
term	Charged off %	Record count
60	25.12	7957
36	10.76	25234

Loan Grades

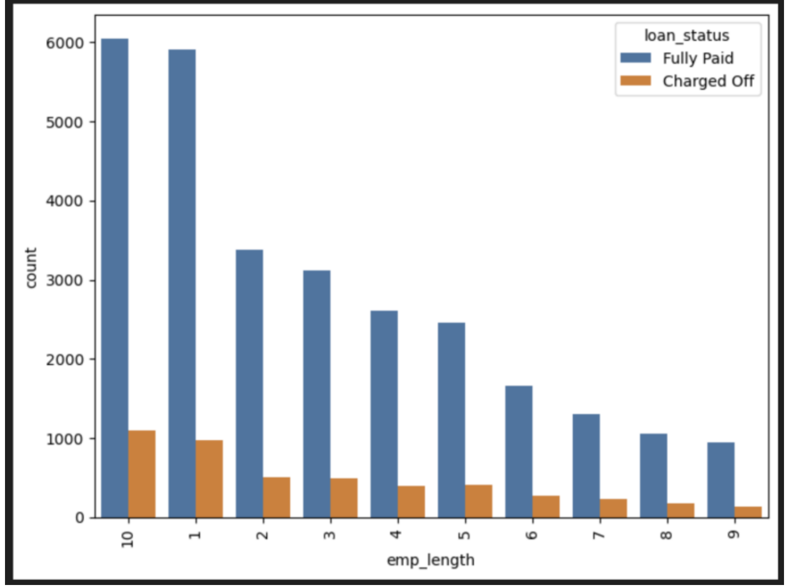


grade	Charged off %	Record count
G	34.59	159
F	31.43	700
E	26.84	2075
D	22.04	4270
C	17.11	6879
B	12.15	10082
A	5.97	9026

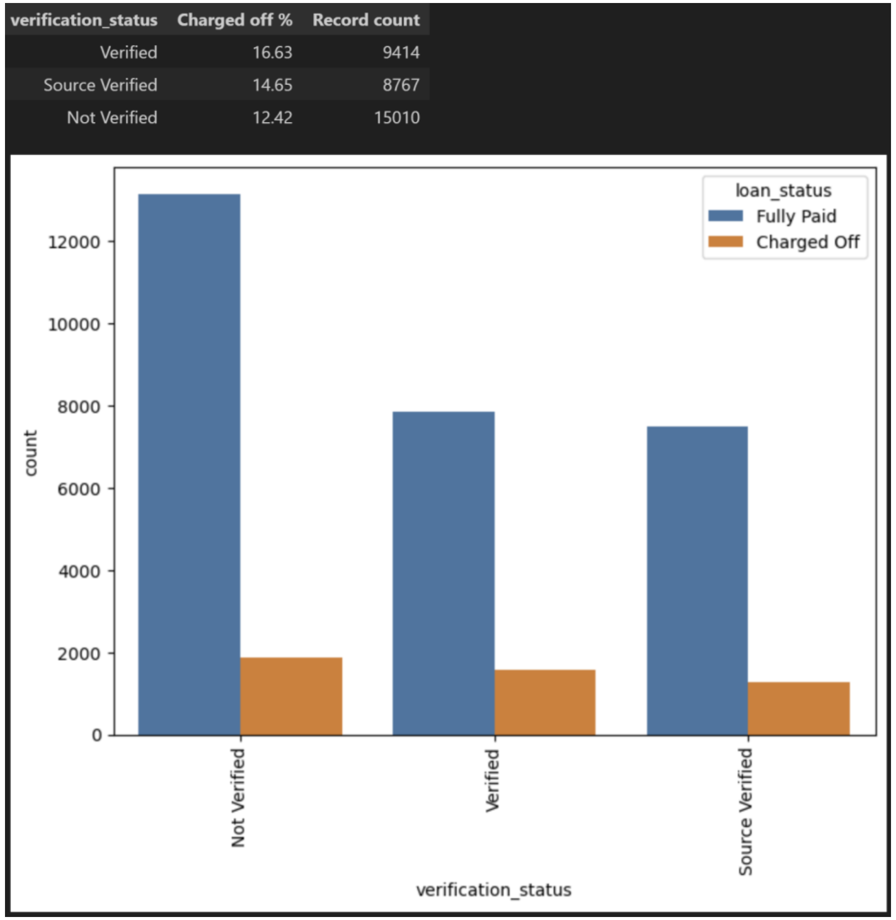
Employee Experience



emp_length	Charged off %	Record count
10	15.43	7149
7	15.21	1532
5	14.25	2864
8	14.24	1236
1	14.18	6883
6	14.04	1937
3	13.61	3615
4	13.38	3012
2	13.14	3882
9	13.04	1081

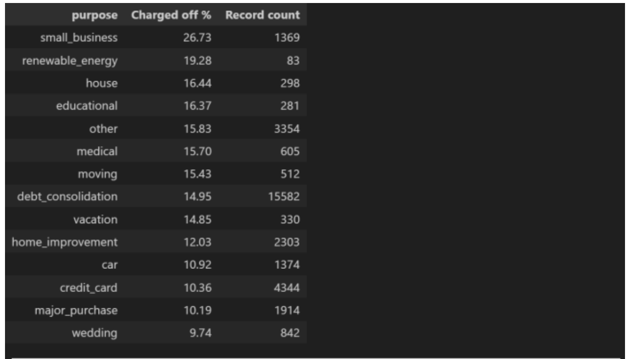


Verification Process

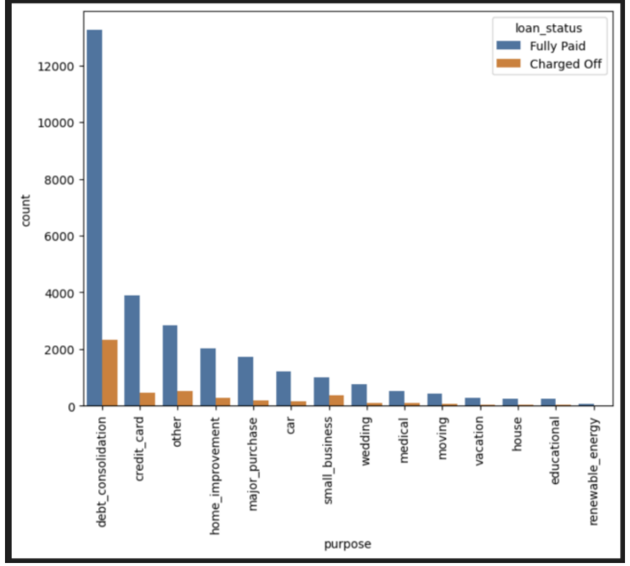


verification_status	Charged off %	Record count
Verified	16.63	9414
Source Verified	14.65	8767
Not Verified	12.42	15010

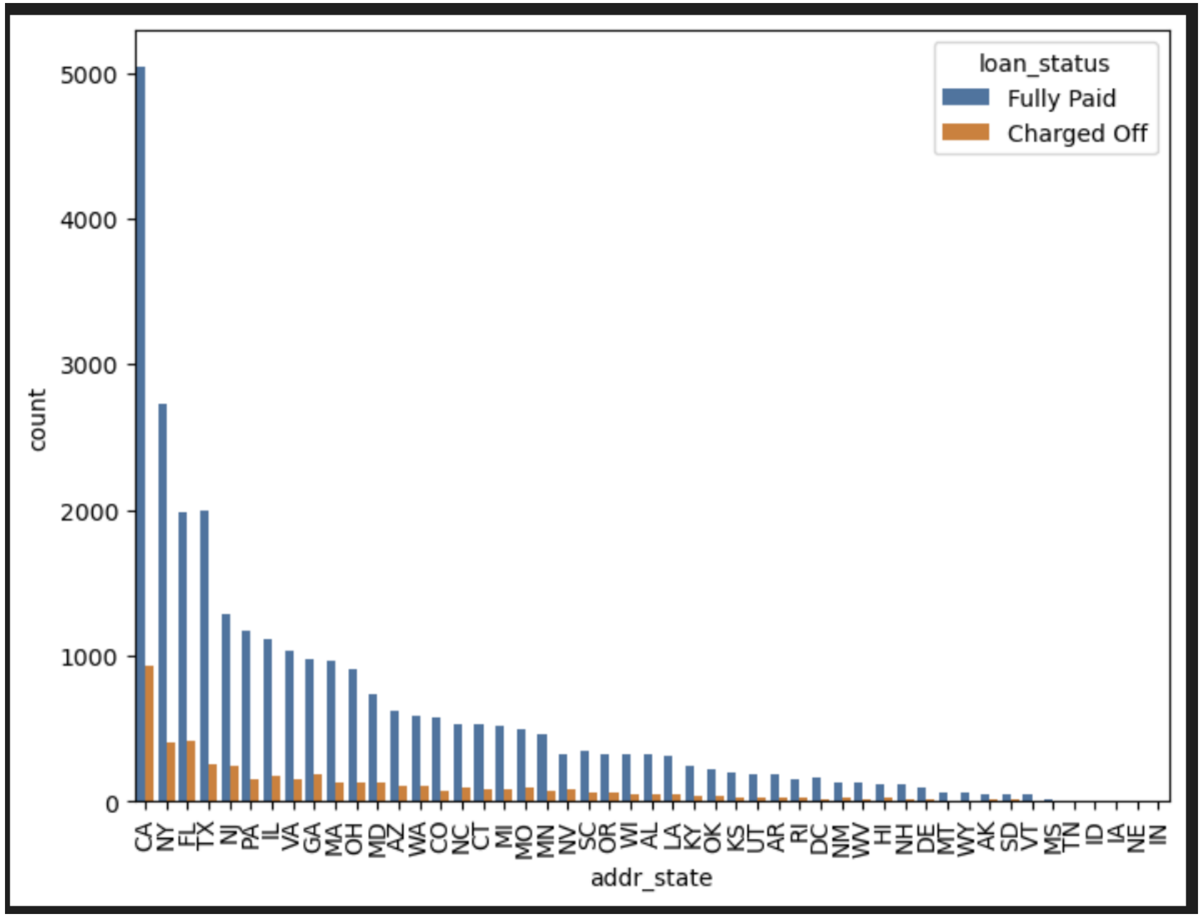
Purpose

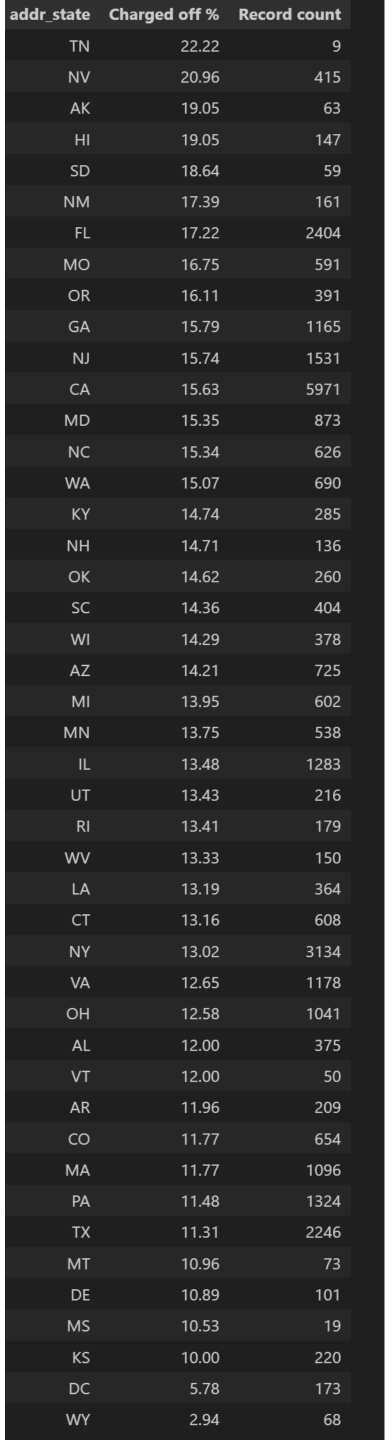


purpose	Charged off %	Record count
small_business	26.73	1369
renewable_energy	19.28	83
house	16.44	298
educational	16.37	281
other	15.83	3354
medical	15.70	605
moving	15.43	512
debt_consolidation	14.95	15582
vacation	14.85	330
home_improvement	12.03	2303
car	10.92	1374
credit_card	10.36	4344
major_purchase	10.19	1914
wedding	9.74	842



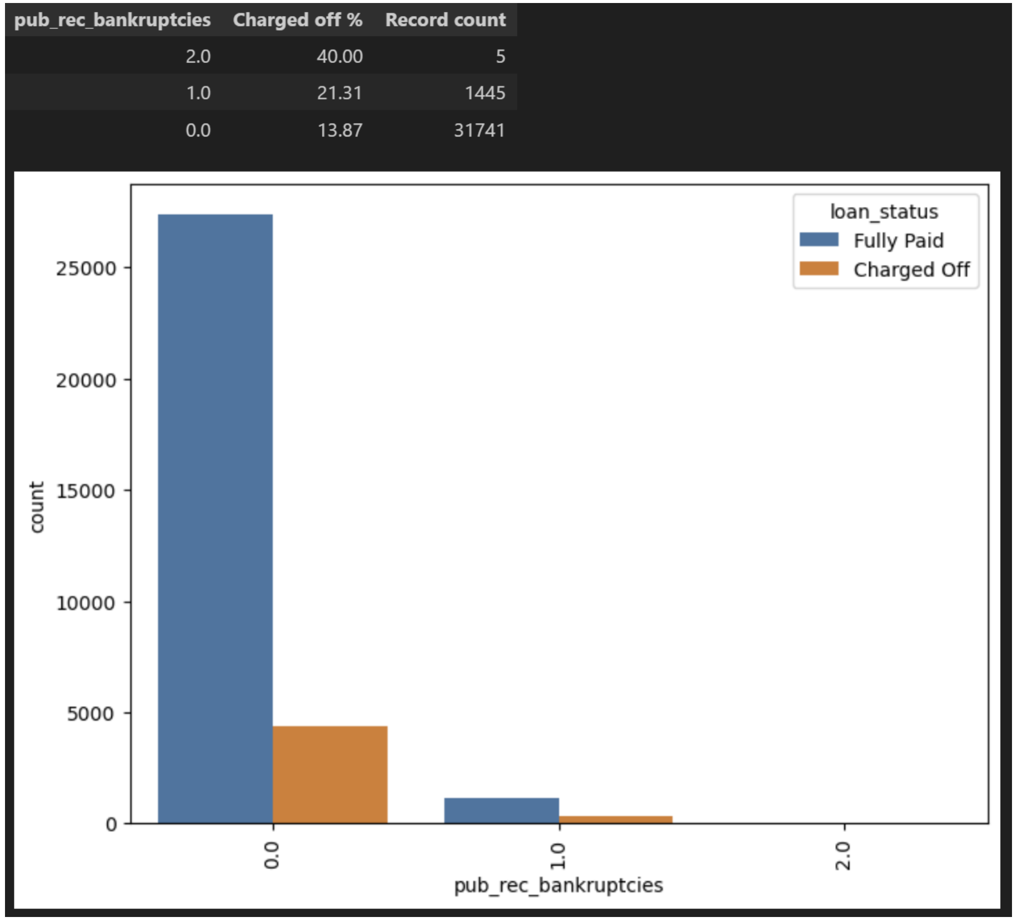
State-Specific Risk





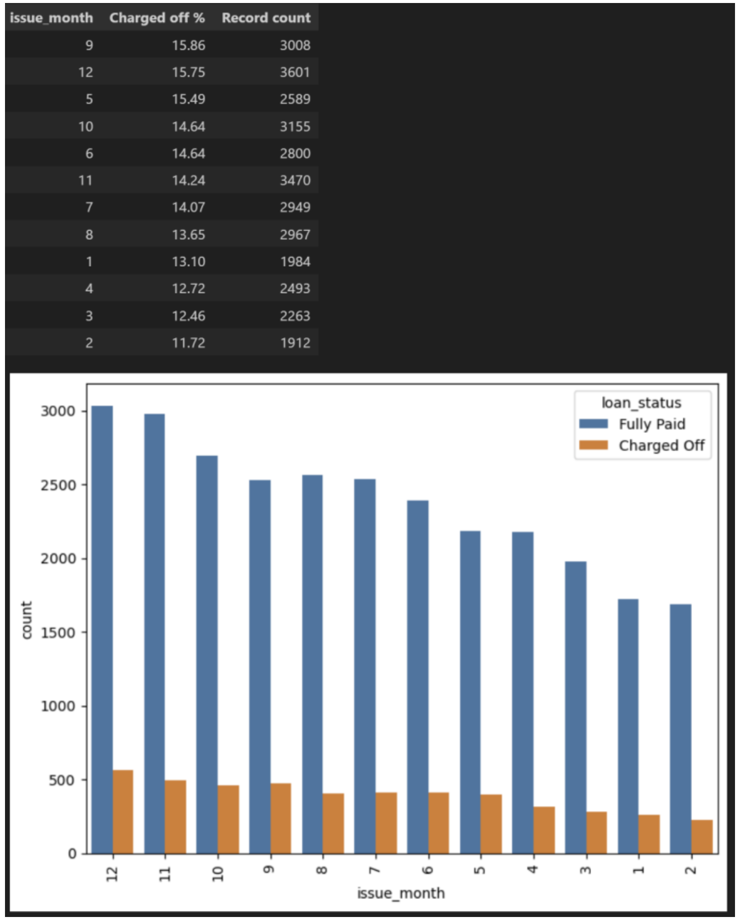
addr_state	Charged off %	Record count
TN	22.22	9
NV	20.96	415
AK	19.05	63
HI	19.05	147
SD	18.64	59
NM	17.39	161
FL	17.22	2404
MO	16.75	591
OR	16.11	391
GA	15.79	1165
NJ	15.74	1531
CA	15.63	5971
MD	15.35	873
NC	15.34	626
WA	15.07	690
KY	14.74	285
NH	14.71	136
OK	14.62	260
SC	14.36	404
WI	14.29	378
AZ	14.21	725
MI	13.95	602
MN	13.75	538
IL	13.48	1283
UT	13.43	216
RI	13.41	179
WV	13.33	150
LA	13.19	364
CT	13.16	608
NY	13.02	3134
VA	12.65	1178
OH	12.58	1041
AL	12.00	375
VT	12.00	50
AR	11.96	209
CO	11.77	654
MA	11.77	1096
PA	11.48	1324
TX	11.31	2246
MT	10.96	73
DE	10.89	101
MS	10.53	19
KS	10.00	220
DC	5.78	173
WY	2.94	68

Bankruptcy



pub_rec_bankruptcies	Charged off %	Record count
2.0	40.00	5
1.0	21.31	1445
0.0	13.87	31741

Loan Issuance by Month



issue_month	Charged off %	Record count
9	15.86	3008
12	15.75	3601
5	15.49	2589
10	14.64	3155
6	14.64	2800
11	14.24	3470
7	14.07	2949
8	13.65	2967
1	13.10	1984
4	12.72	2493
3	12.46	2263
2	11.72	1912