Mihir Pahwa

Lending Case Study - EDA

IIITB

EDA process

Define Objective

- Data understanding
- Data cleaning
- Highlight factors that affect potential defaults using EDA

Data Cleaning

- Handle columns with missing data
- Outlier handling
- Remove irrelevant columns
- Fixing incorrect data types

Univariate analysis

- Observe general trends in columns
- Filter out columns with no significant lift

Bivariate Analysis

- Converting numeric columns to categorical by binning , then observing lift.

Recommendations

- Features that affect default rate in a positive or negative manner
- Special patterns observed

Business Objective

Minimize Credit Loss

Identify factors influencing loan default risk.

Reduce the financial loss caused by defaulting borrowers.

Enhance Risk Assessment

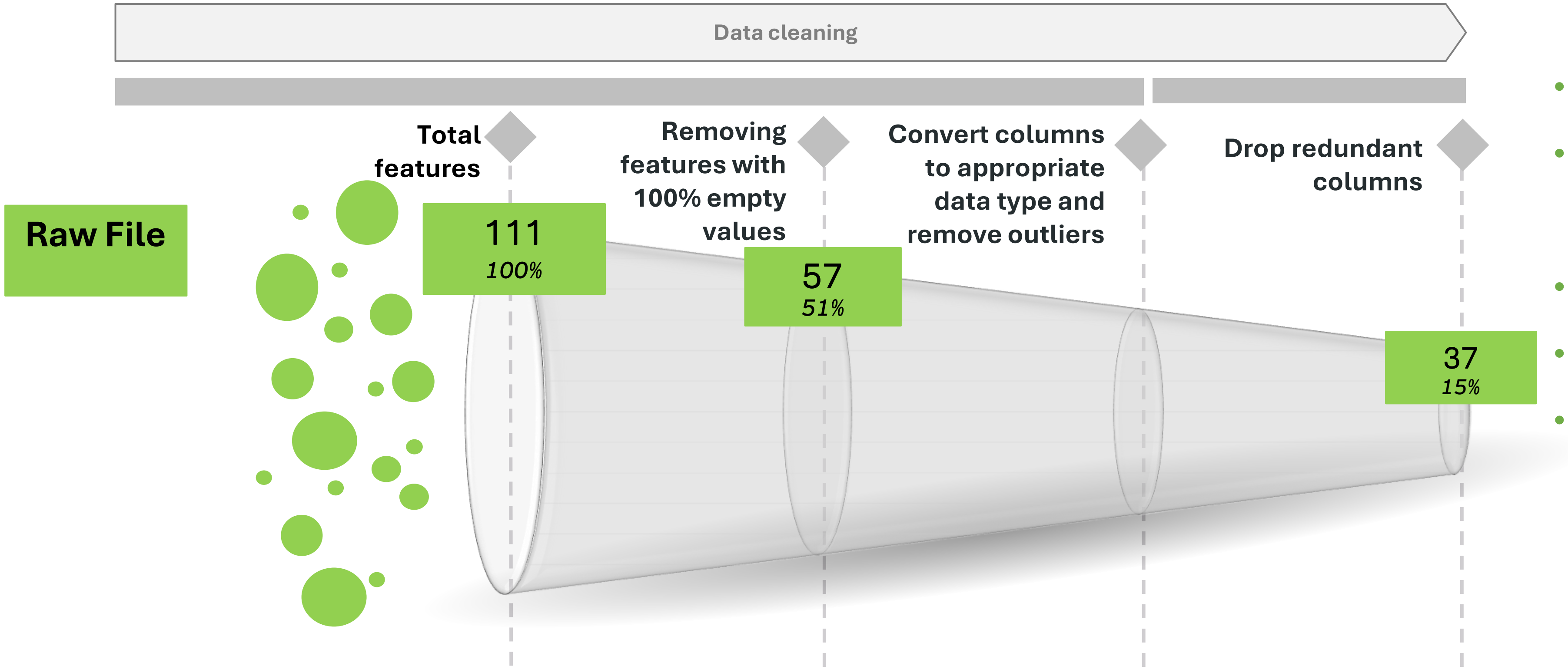
Understand driver variables associated with loan default.

Utilize this knowledge for portfolio and risk assessment.

Improve Decision-making Process

Utilize EDA to identify patterns indicating potential defaults.

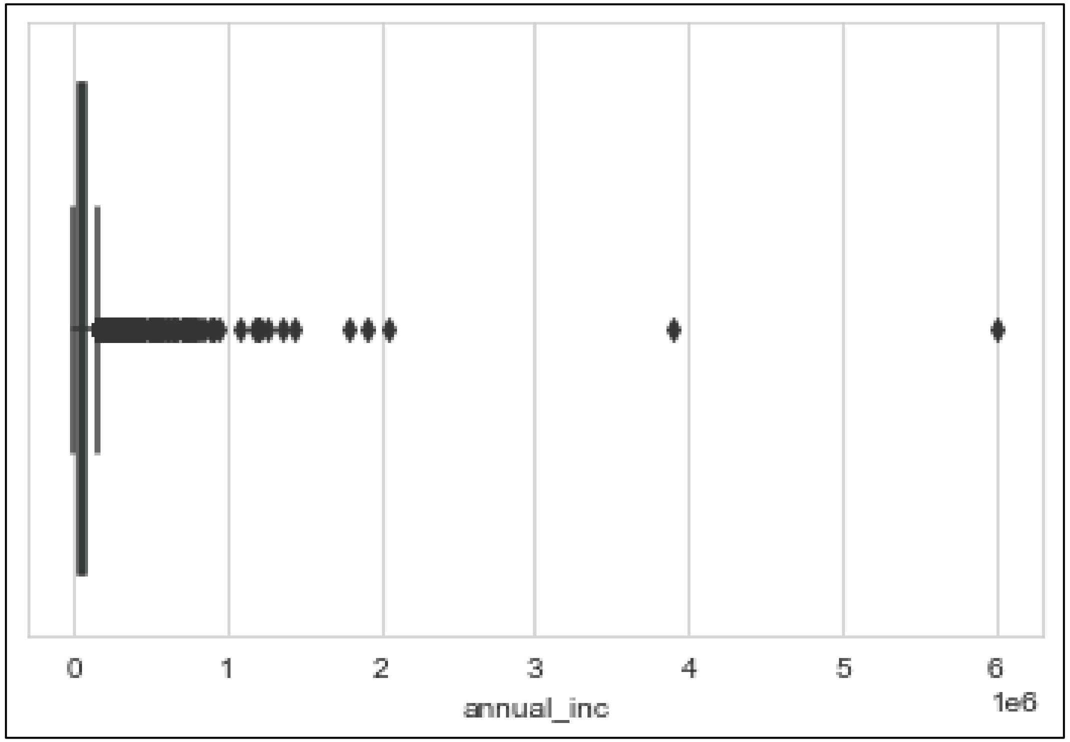
Data Cleaning

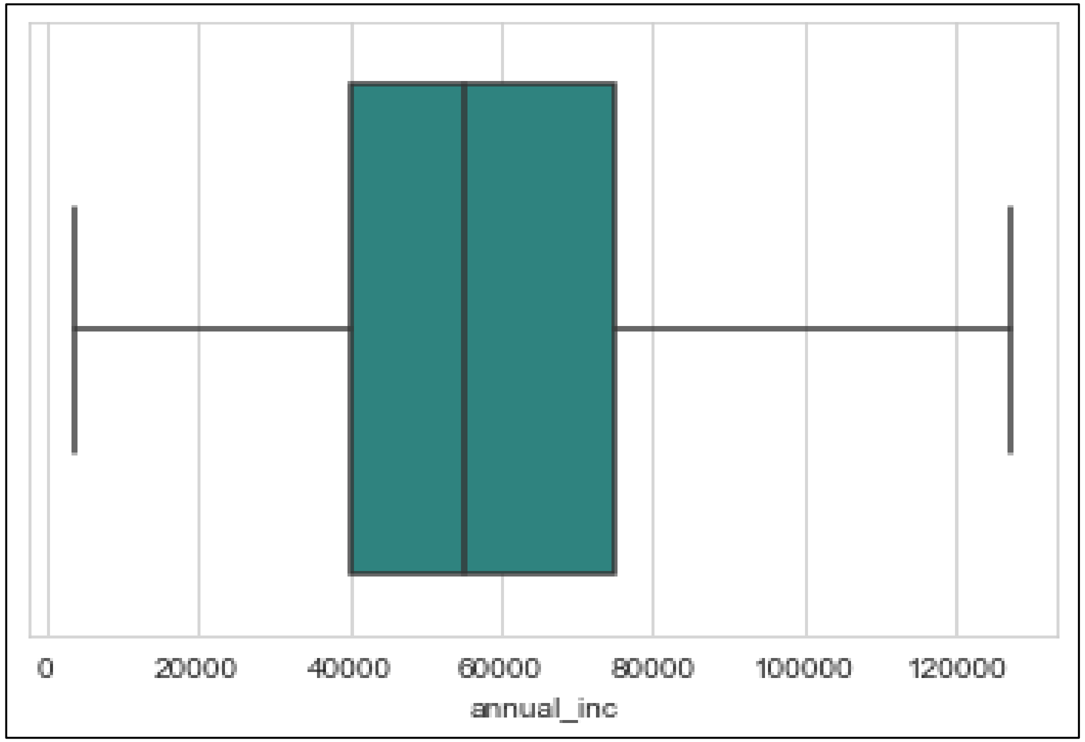


- The raw data contained 39,717 rows and 111 columns.
- The 'Loan_status' column was chosen as the target, mapping 'Fully paid' to 1 and 'Charged off' to 0.
- No loan ID duplicates were identified.
- 54 columns showed 100% null values.
- Approximately 15% of columns had incorrect data types

Outlier Removal

- Removing outliers from the annual income column helps refine the dataset, ensuring a more accurate representation of income distribution
- This process enhances the reliability of insights drawn from the data, particularly in understanding the relationship between income levels and default rates.

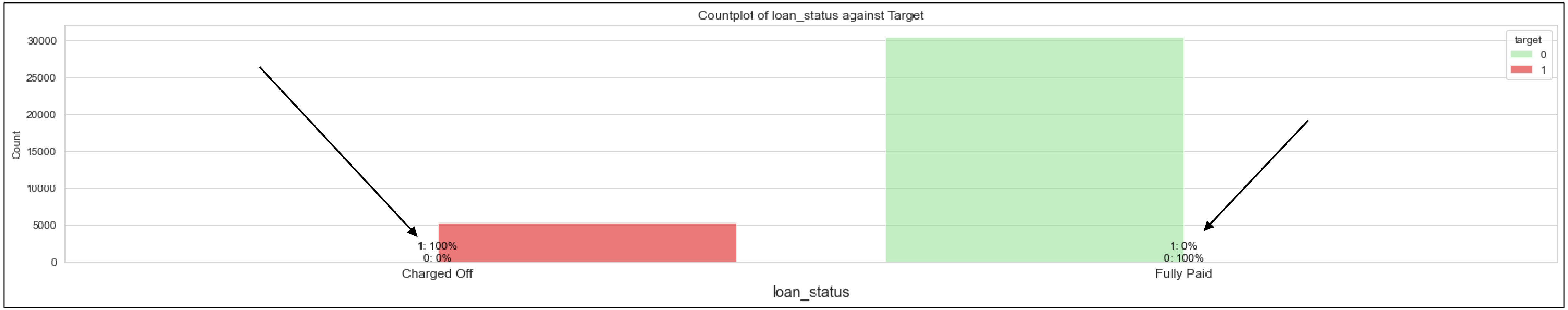




Influential Features Analysis

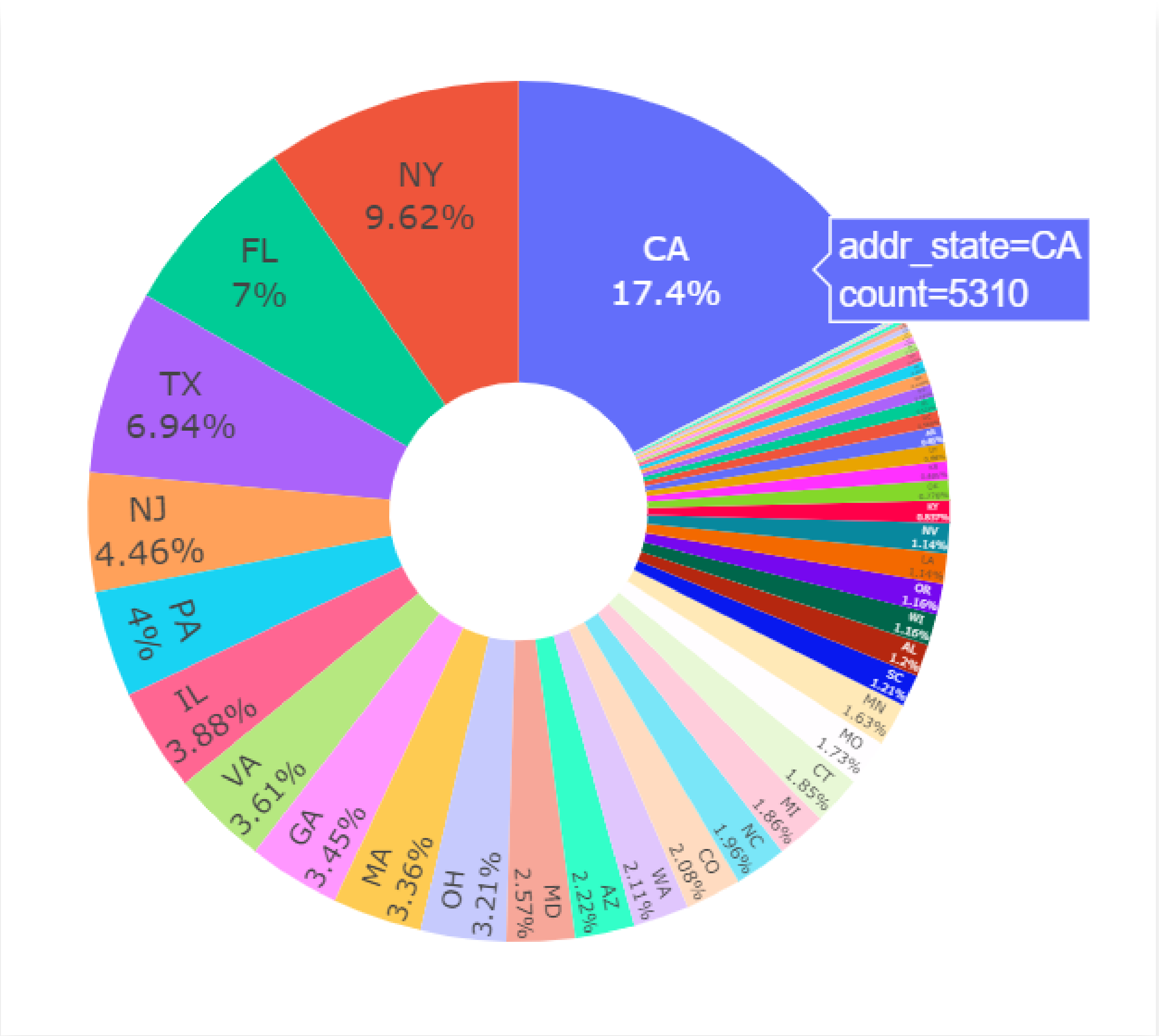
To pinpoint influential features, we utilize countplots displaying default rates within each feature category. These plots are annotated, indicating the count of defaults where '1' denotes defaults and '0' represents fully paid instances. This annotation provides a clear visual representation of default occurrences across different feature categories

1 : charged off
0 : Fully paid



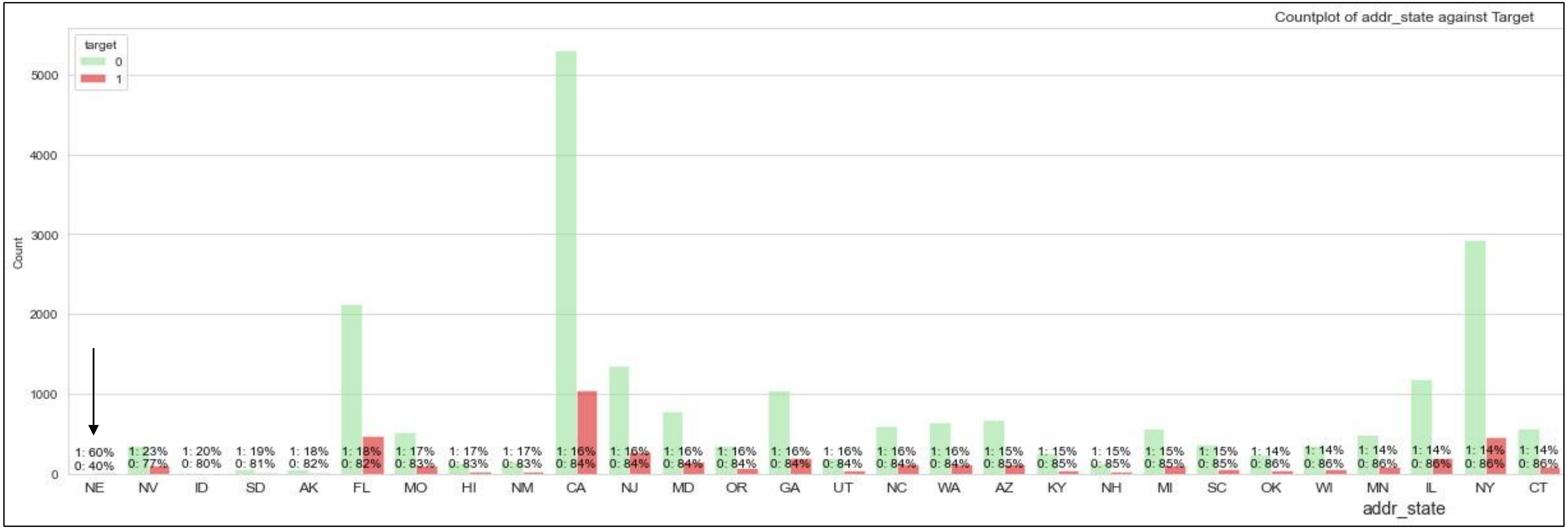
State-wise default rate

The pie chart shows the default rates across different states, with California displaying the highest default count. It's important to note that the higher default rate in California may be attributed to a higher volume of disbursals in the state.



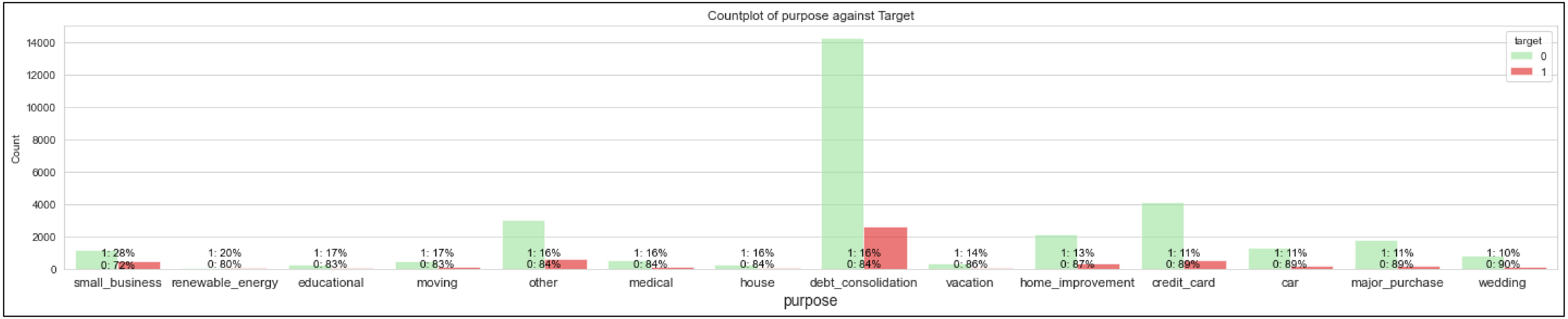
State-wise default

- The "addr_state" column reveals Nebraska (NE) with a 60% default rate, hinting at potential fraud occurrences.
- Also, states like Nevada (NV), Idaho (ID), South Dakota (SD), and Alaska (AK) exhibit notably high default rates.



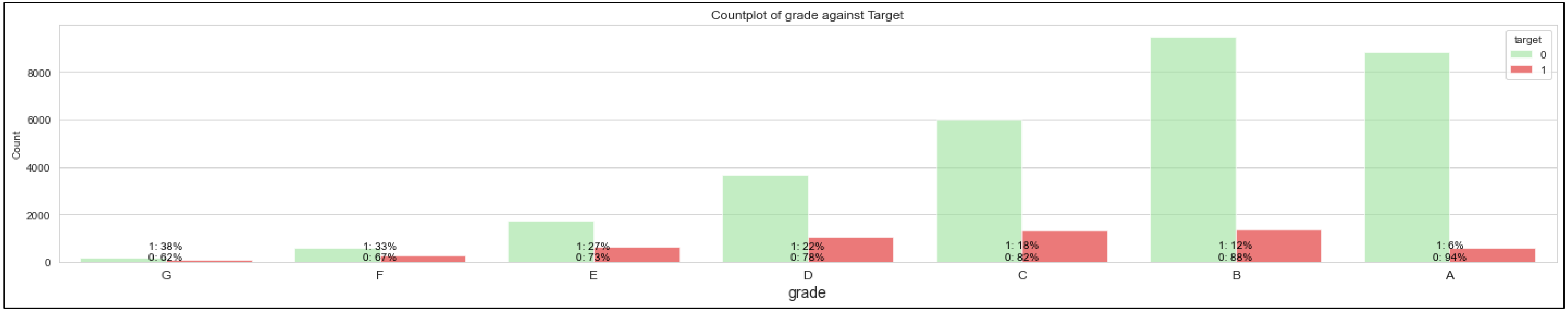
Loan Purpose

- "Small Business" category has the highest default rates, followed by "Renewable Energy."
- Implication: Consider adjusting interest rates to mitigate risk in high-default categories.



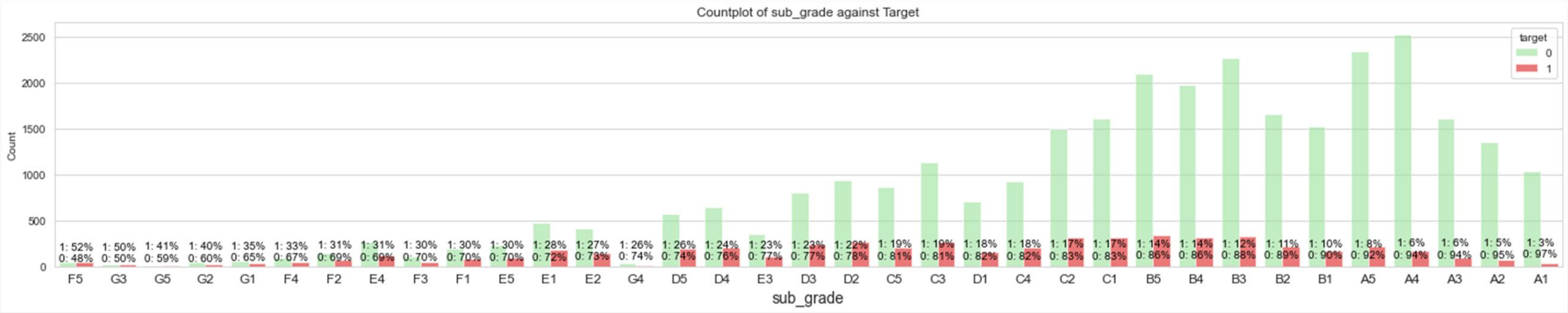
Grade

- The "Grade" column ('G', 'F', 'E’) categories display alarming default rates, exceeding 25%.
- As "Grade" decreases, defaults increase, making this column a robust indicator of potential defaults.



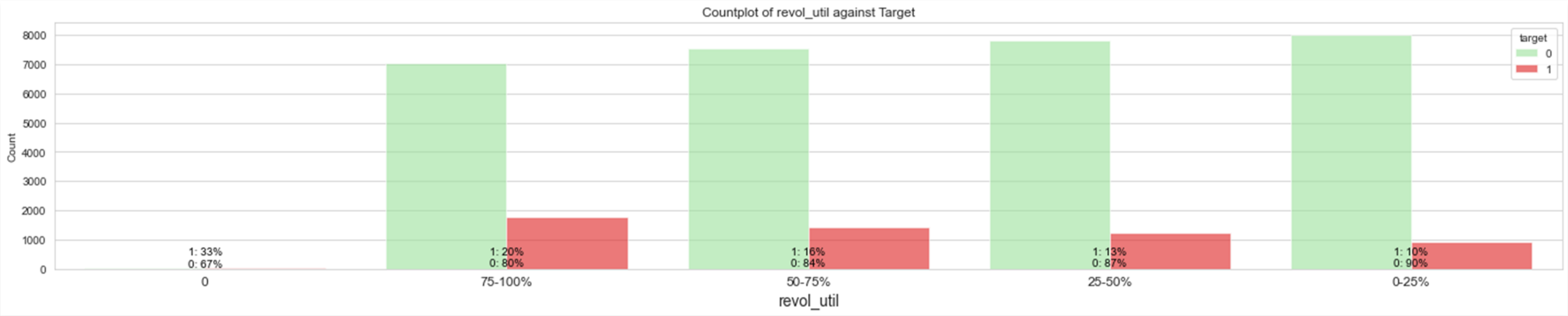
Sub-Grade

Similar trends in "Sub-Grade" column, with sub-grades (G, F, E) exhibiting high default rates.



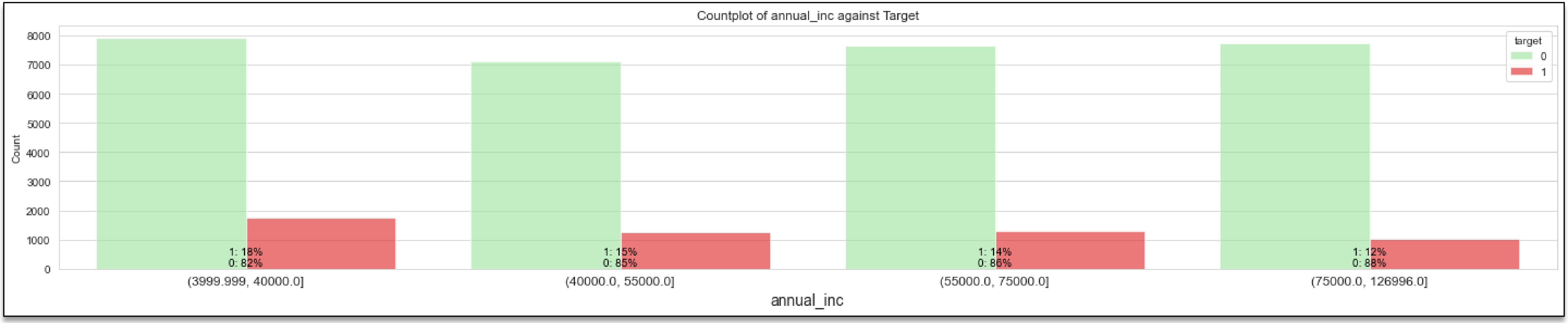
Utilization of credit

Higher "Revolving Utilization" indicates a higher default rate.



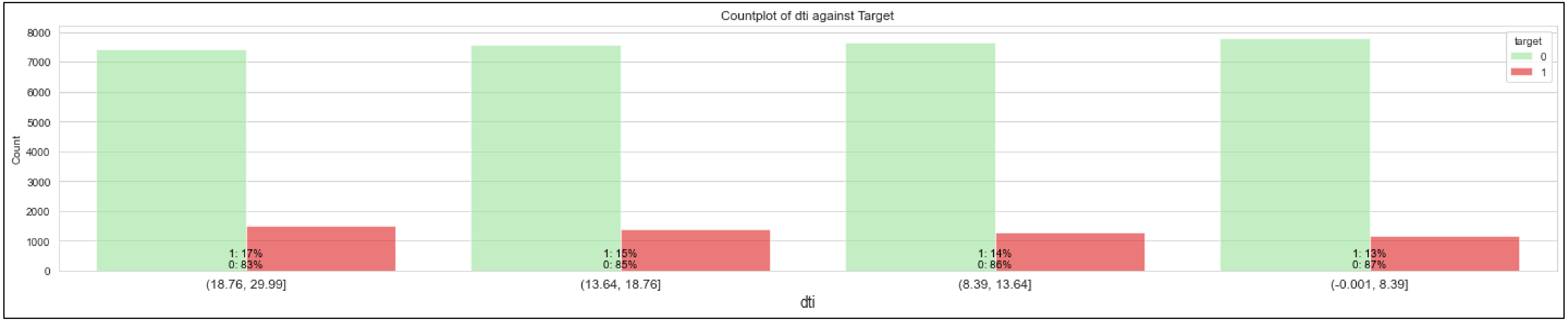
Annual Income

- Lower annual income correlates with higher default rates.
- Suggests a possible scenario of higher loan amounts disbursed among some members of lower-income groups.



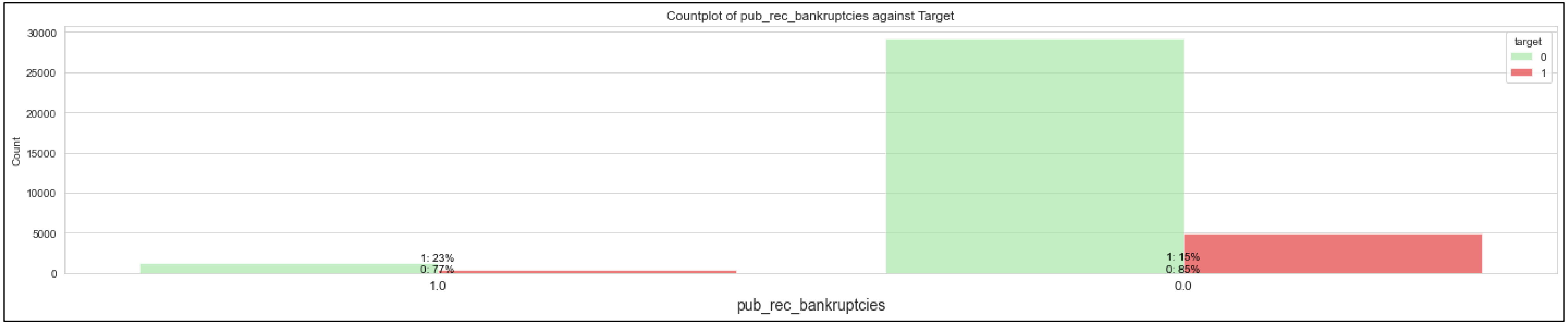
Debt-to-income Ratio

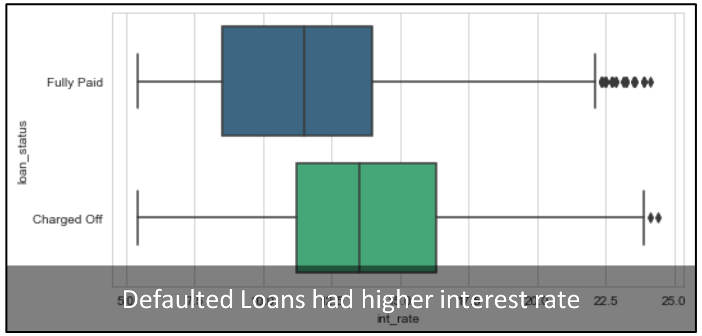
Higher debt-to-income ratio is associated with higher default rates.



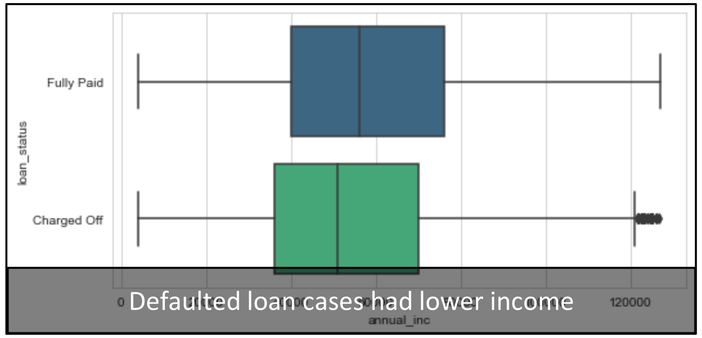
Public Bankruptcy records

Individuals with public bankruptcies have higher default rates.

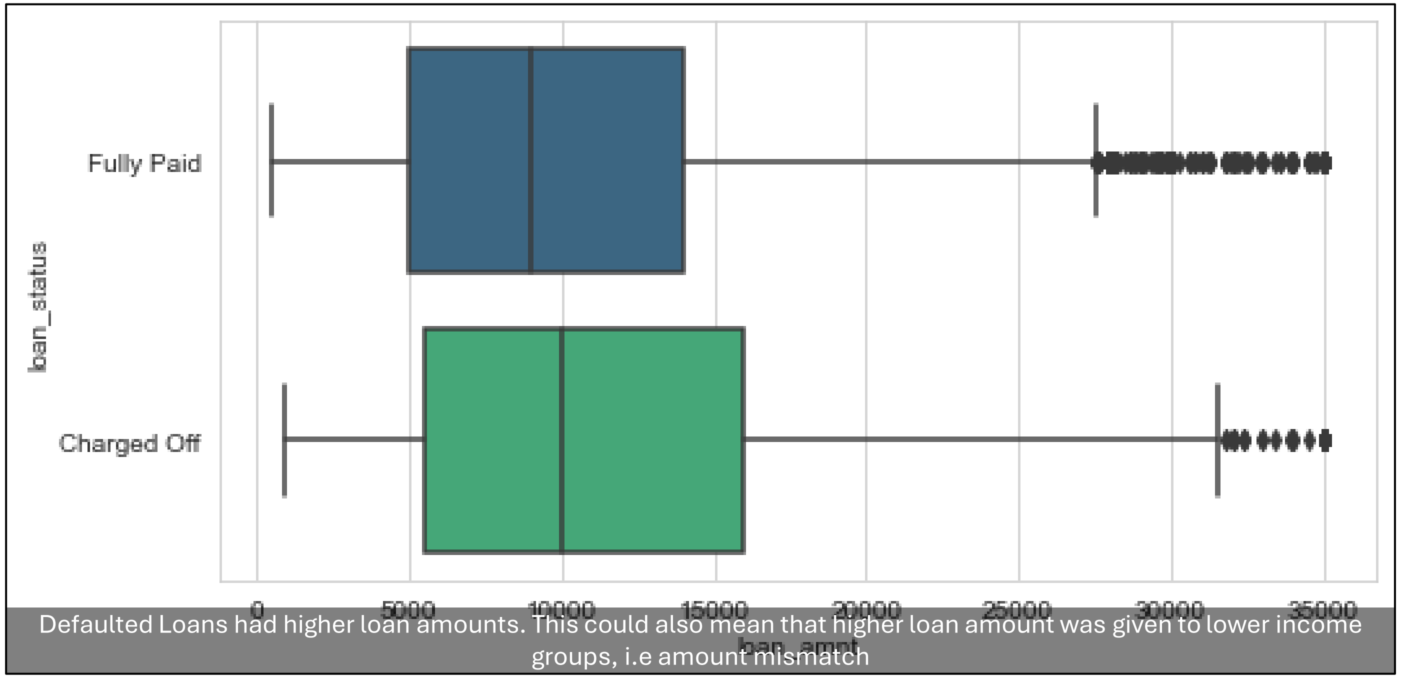




Defaulted Loans had higher interest rate



Defaulted loan cases had lower income



Defaulted Loans had higher loan amounts. This could also mean that higher loan amount was given to lower income groups, i.e amount mismatch

Created By : Mihir Pahwa