Loan Data Analysis

Presented by Gayatri Gautam & Harshita

Date-:2024-10-24

The intention of this project is to perform exploratory data analysis

CONTENT

Objective & Goal:

The objective of the organization is to minimize the level of credit losses by singling out certain categories of loan applicants who are most likely to default through exploratory data analysis (EDA) techniques. Having an understanding of the underlying reasons for loan defaults will enhance the company's risk evaluation practices and loan book management strategies

The loan data is used to analyze key metrics like loan amount, interest rate, employment length, and Total amount receive & Total payment. The goal is to identify patterns in the loan data to assist in decision-making for bank investments.

Data Loading and Cleaning

Code Insight:

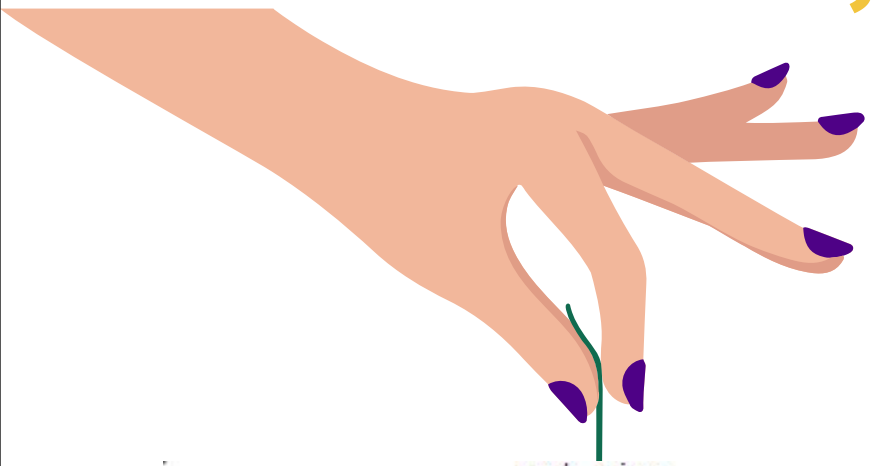
The dataset is loaded using pandas into a DataFrame, followed by removing missing values and duplicates.

Data Cleaning:

Transforming columns like int_rate, revol_util etc. into numeric types, and filling missing values where necessary.Read the data into a dataframe Dropping duplicates Rows & Columns. Check if the columns exist before dropping

Removing redundant information within a column. Check the shape of the DataFrame after removing redundancy

Mean, Standard Deviation, Minimum, Maximum, and various percentiles



Quantitative variables such as the sums of loans or the interest rates would have mean, minimum, maximum, and standard deviation values. Qualitative variables such as loan status or job title will include measures of unique values, the mode, and its frequency.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	39717.0	NaN	NaN	NaN	683131.91306	210694.132915	54734.0	516221.0	665665.0	837755.0	1077501.0
member_id	39717.0	NaN	NaN	NaN	850463.559408	265678.307421	70699.0	666780.0	850812.0	1047339.0	1314167.0
loan_amnt	39717.0	NaN	NaN	NaN	11219.443815	7456.670694	500.0	5500.0	10000.0	15000.0	35000.0
funded_amnt	39717.0	NaN	NaN	NaN	10947.713196	7187.23867	500.0	5400.0	9600.0	15000.0	35000.0
funded_amnt_inv	39717.0	NaN	NaN	NaN	10397.448868	7128.450439	0.0	5000.0	8975.0	14400.0	35000.0
term	39717.0	NaN	NaN	NaN	42.418007	10.622815	36.0	36.0	36.0	60.0	60.0
int_rate	39717.0	NaN	NaN	NaN	12.021177	3.724825	5.42	9.25	11.86	14.59	24.59
installment	39717.0	NaN	NaN	NaN	324.561922	208.874874	15.69	167.02	280.22	430.78	1305.19
grade	39717	7	B	12020	NaN	NaN	NaN	NaN	NaN	NaN	NaN
sub_grade	39717	35	B3	2917	NaN	NaN	NaN	NaN	NaN	NaN	NaN
emp_title	37258	28820	US Army	134	NaN	NaN	NaN	NaN	NaN	NaN	NaN
emp_length	39717.0	NaN	NaN	NaN	4.533223	4.000823	-1.0	2.0	4.0	9.0	10.0
home_ownership	39717	5	RENT	18899	NaN	NaN	NaN	NaN	NaN	NaN	NaN

About Outliers

An outlier is a single data point that goes far outside the average value of a group of statistics in another word an outlier is not a part of grouping

```python
# Removing outliers using the IQR method
def remove_outliers(loan, column):
    Q1 = loan[column].quantile(0.25)
    Q3 = loan[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return loan[(loan[column] >= lower_bound) & (loan[column] <= upper_bound)]

# Removing outliers from  numerical columns
columns_to_check = ['loan_amnt', 'int_rate','annual_inc', 'dti','revol_bal','total_pymnt']

# Applying the outlier removal function to each column
for col in columns_to_check:
    loan = remove_outliers(loan, col)

# Checking the shape of the data after outlier removal
loan.shape
```

```
(33769, 57)
```

Representation of Outlier Goal:

The code presents a technique for eliminating outlier observations in a data set employing the Interquartile Range method. Extremes are splits in data points that distort interpretations, and so their exclusion guarantees better analyses.
Sequential Description of the Implementation:
Specifying the remove_outliers method:

The primary aim of the remove_outliers(loan, column) function is to serve as an implementation of the IQR method for detecting and mitigating outliers in the relevant departments of the loan DataFrame.
Find Quartiles:

Q1 (24th percentile) with the Q3 (76th percentile) percentiles are worked out in the order of quantile(0.25) and quantile(0.75), which implies the data percentiles whose extremum lies within 25% and 75% of the data respectively.
The IQR (Interquartile Range) is the calculation of Q3 - Q1 as this represents the middle 50% spread of the data.
Explore outlier limits:

The lower bound is calculated as Q1 - 1.5 * IQR, and the upper limit is set at Q3 + 1.5 * IQR. Any data that is beyond these ranges is termed an outlier.
Jagged Edges Elimination:

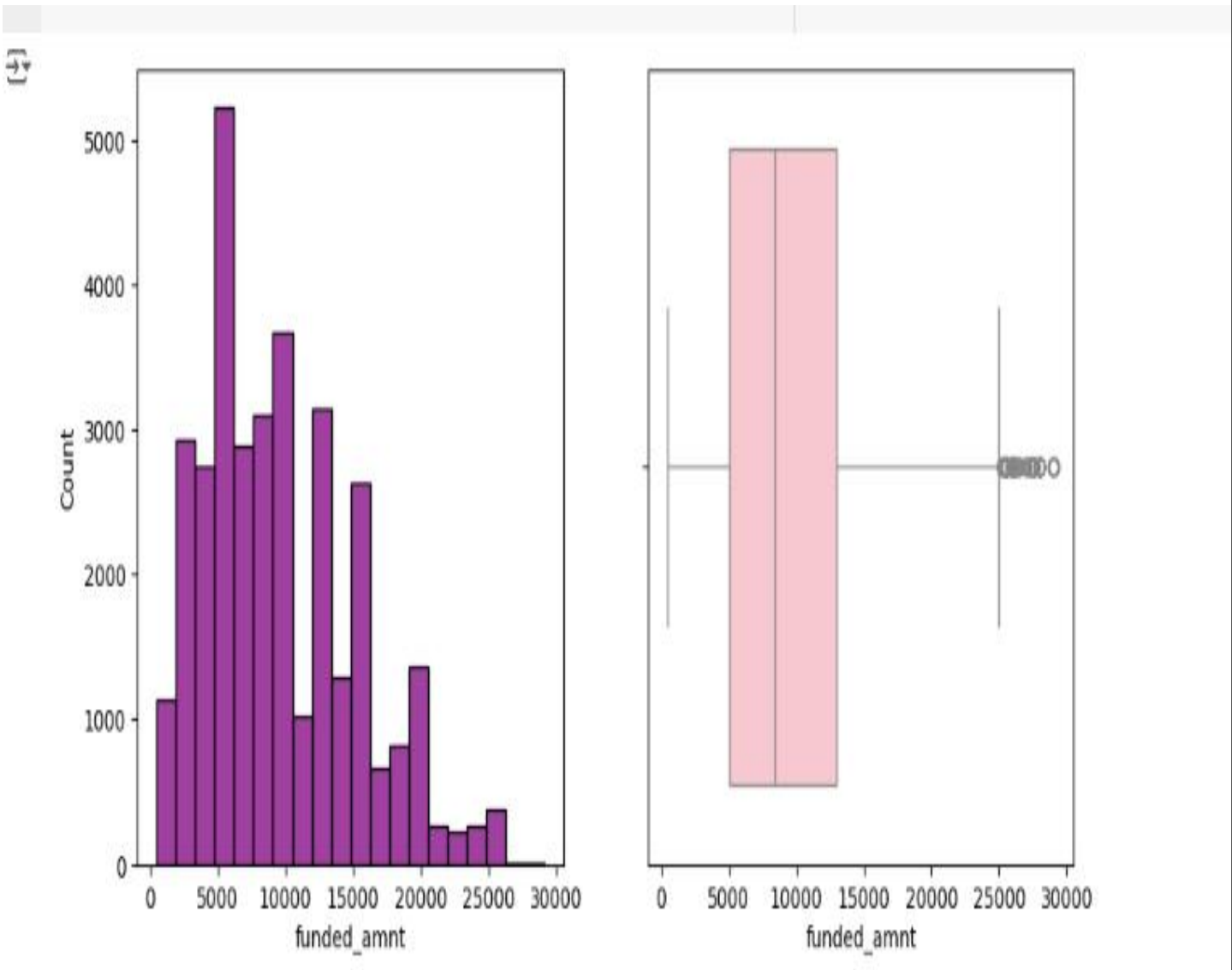
The function filters the data by using a condition which does not allow values to be above the upper limit and below the lower limit.

Univarient Analysis

The focus of the analysis is to examine only one variable in isolation (univariate) through the use of histograms and box plots for both "funded_amnt" and "int_rate".

First Plot:

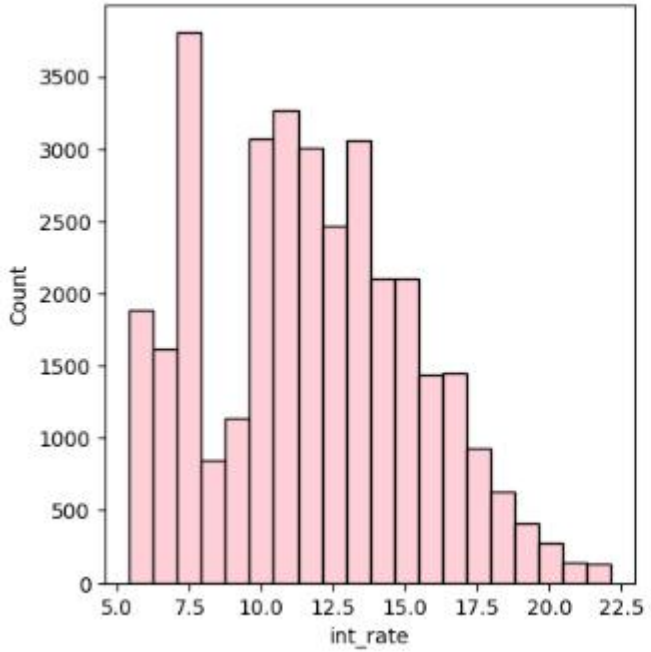
Histogram and Box Plot of the 'funded_amnt' (Bank Funded/Investment):
A purple-colored histogram is constructed for the attribute "funded_amnt" illustrating the frequency distribution of the funded amount.
A box plot is constructed to represent the dispersion, mean, and possible outliers in “funded_amnt”.

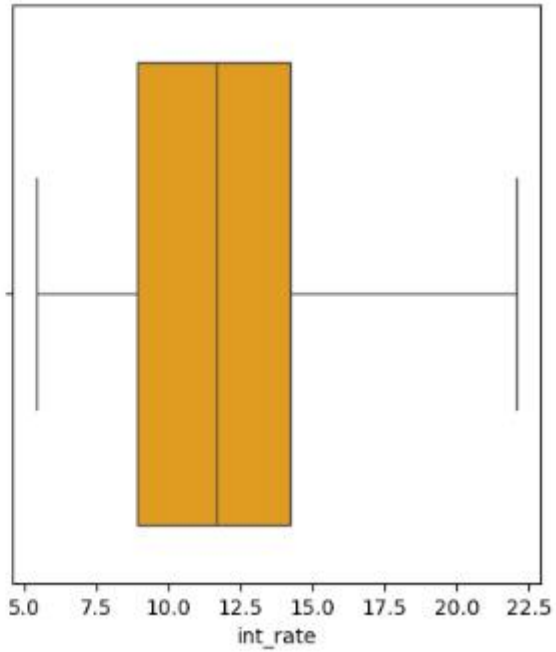


Univarient Analysis

Second Plot:

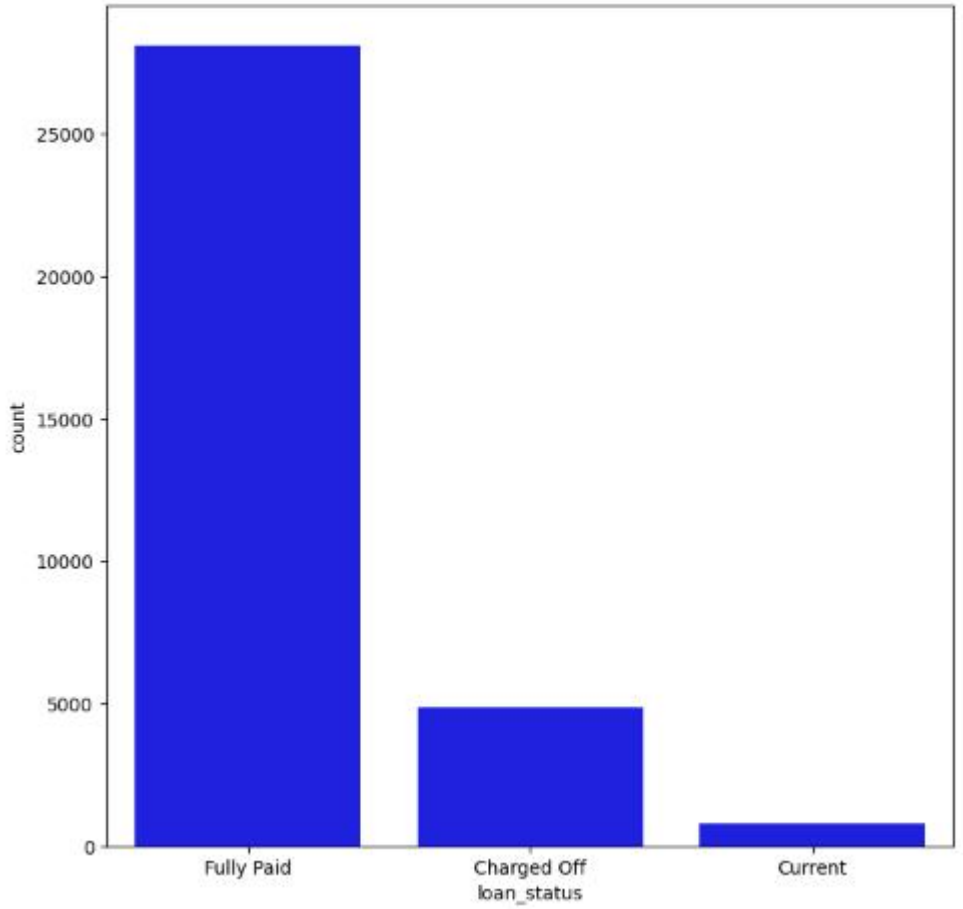
Histogram and Box Plot for "int_rate" (Bank Rate of Interest):
The histogram which is colored in pink illustrates the frequency distribution of the rates of interest.
The box plot shows the range of interest rates, and trends or extreme values if any.





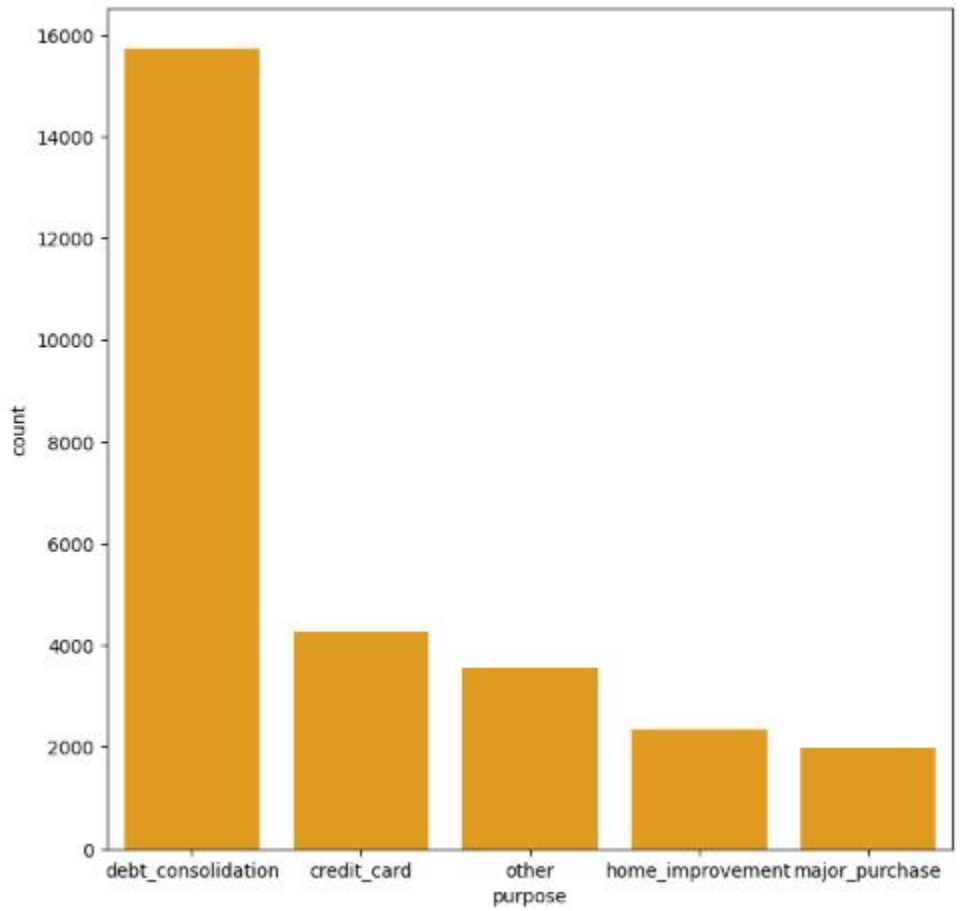
Univarient Analysis

This plot shows the frequency of different loan statuses (e.g., fully paid, charged off, etc.) using a blue-colored bar plot.



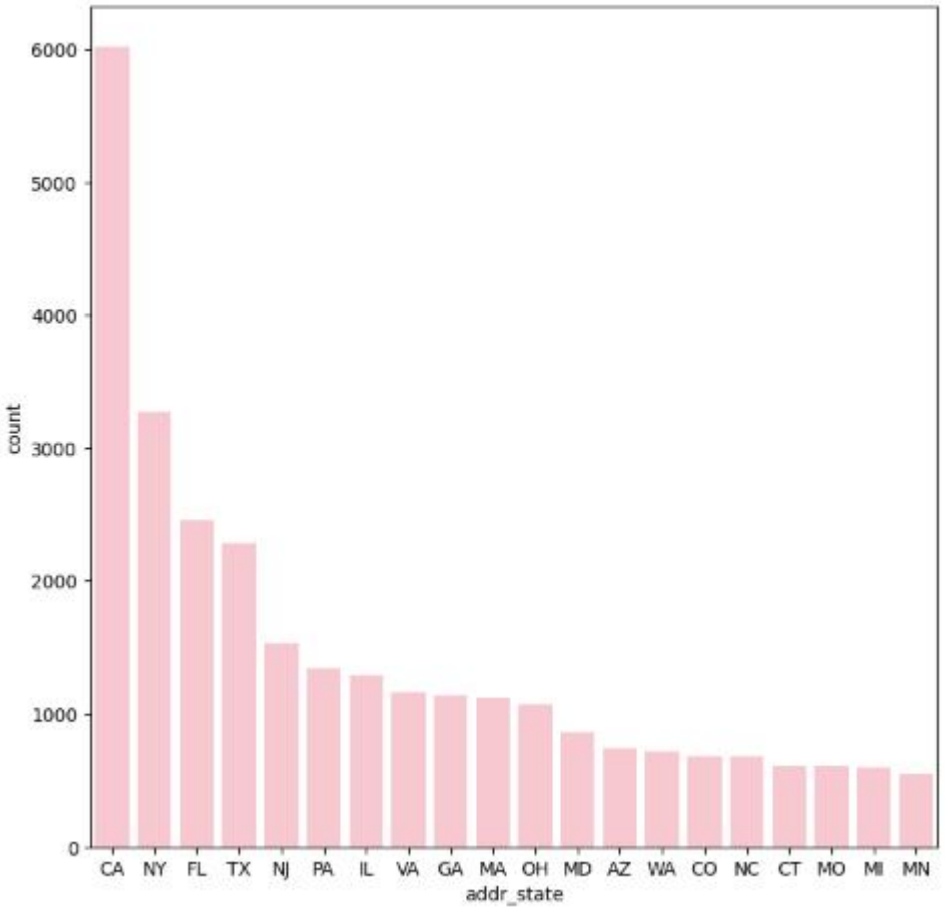
Univarient Analysis

This bar plot visualizes the frequency of the different purposes for which loans were taken (e.g., credit card, home improvement, etc.) using a Orange-colored bar plot



Univarient Analysis

This plot shows the distribution of loans across top 20 states



Bank Investment Source by Loan Attributes

The present analysis focuses on how the amounts funded by banks (bank investments) differ depending on various loan related attributes such as loan status, purpose, state and the duration of the loan. Each chart shows the average bank investment prevailing across these categories

Bivariate Analysis

Part 1

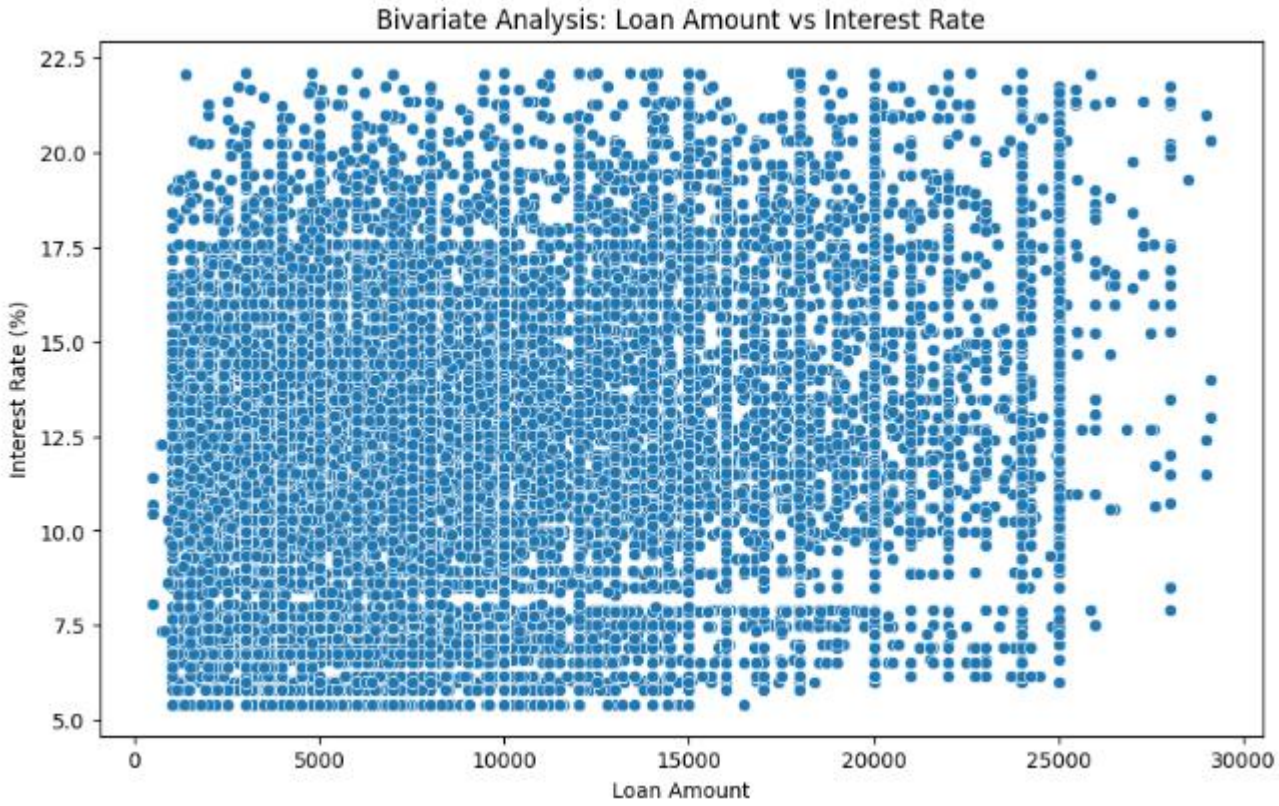
About-:

Through bivariate analysis a statistician can quantify possible an actual relationships between the elements of the dataset, which ultimately leads to better proof of how one element influences the other.

Plots Description:

Loan Amount v/s Interest Rate (Scatter Plot):

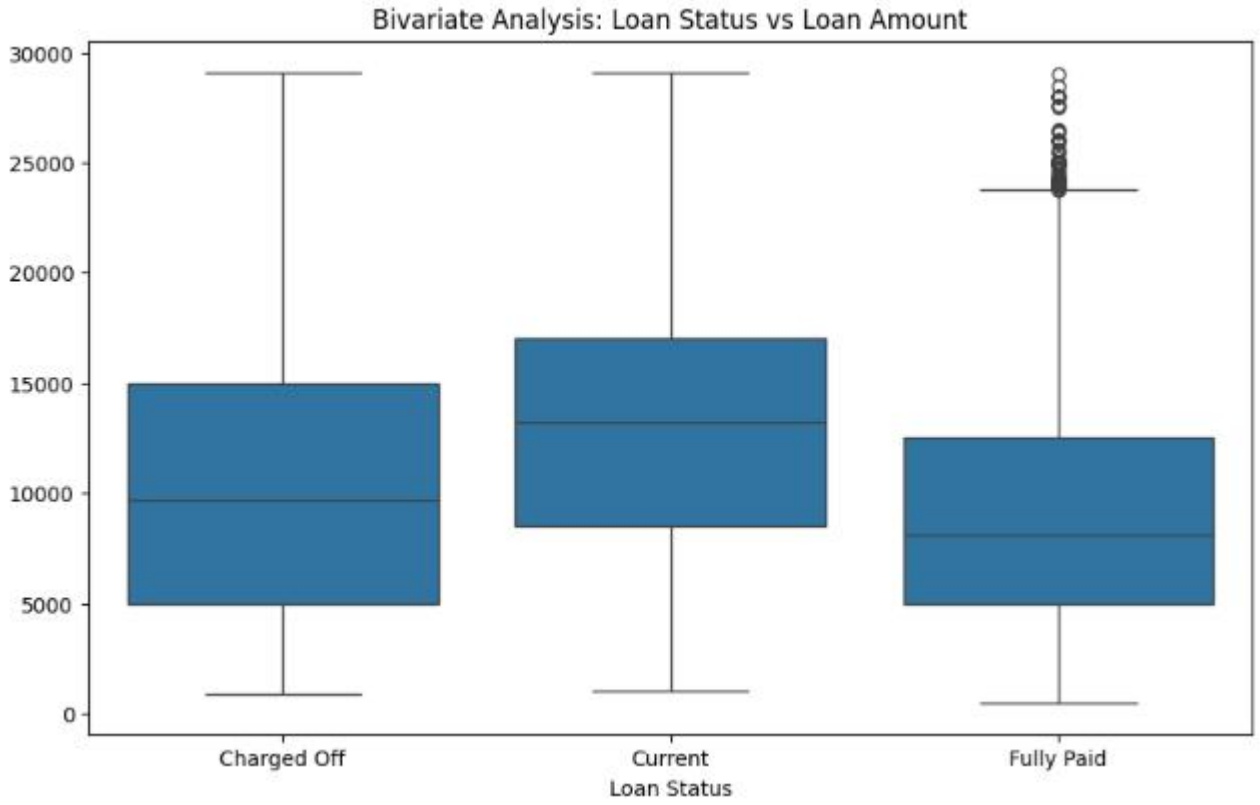
Type of Plot: Scatter plot.
Objective: To see the relationship between the loan amour and interest rate on the loan.
X-axis: “Loan Amount” (loan_amnt).
Y-axis: “Interest Rate (%)” (int_rate).
Insight: Scatter plots are useful in visualizing patterns or relationships such as the impacts of higher loan amounts revolving to higher or lower interest rates, or none.



Part 2

Loan Status versus Loan Amount (Box Plot):

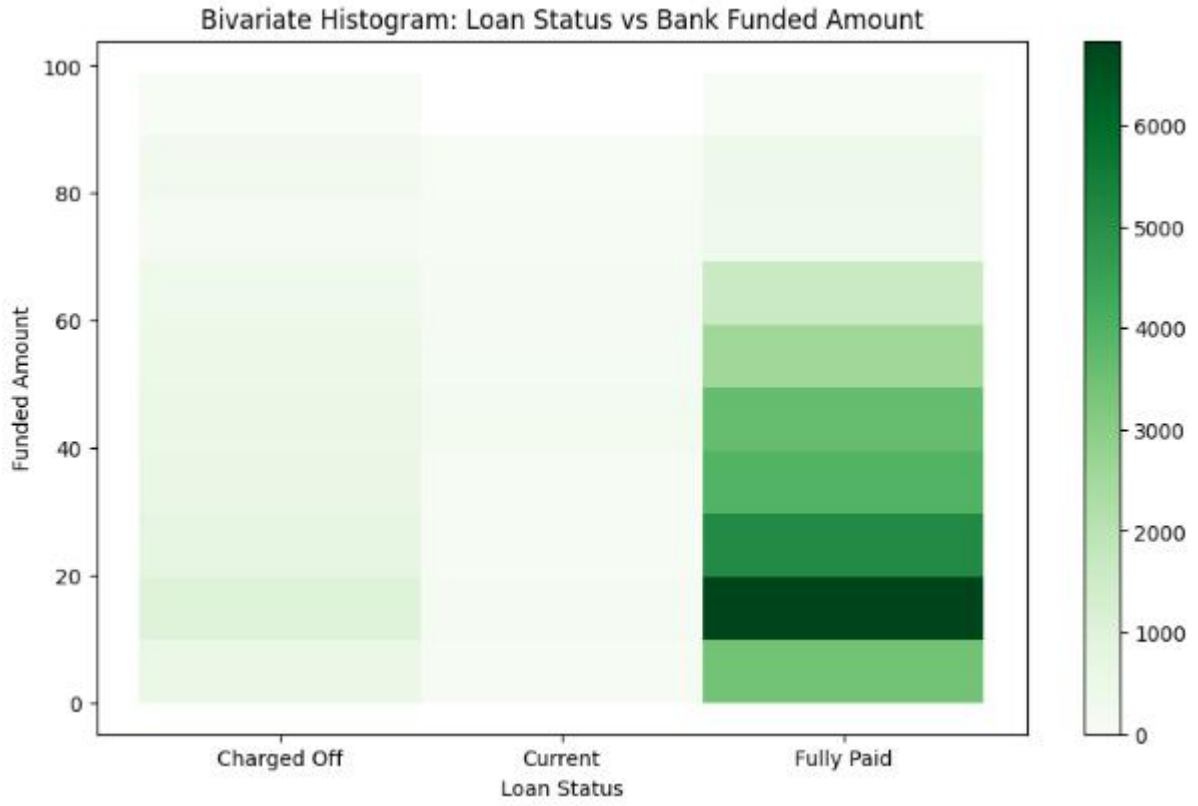
Type of Plot: Box plot.
Objective: To analyze the distribution of loan amounts within the various loan status groups.
X-axis: “Loan Status” (loan_status).
Y-axis: “Loan Amount” (loan_amnt).
Insight: Box plots clearly depict how the loan amounts are spread for every loan status without the concern of extreme values or dispersion.



Part 3

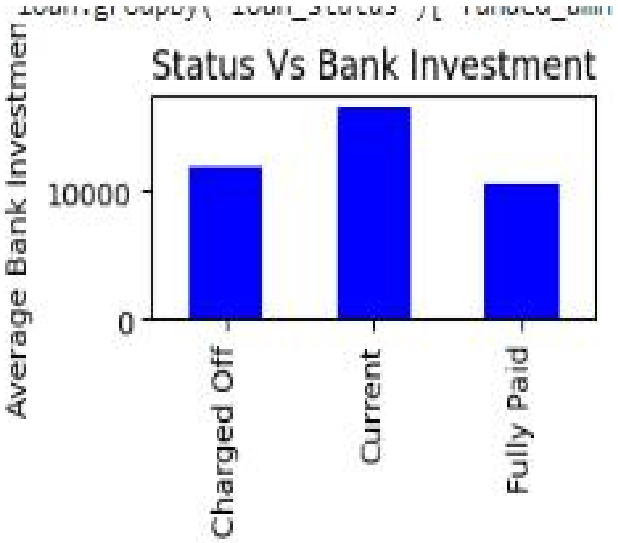
Loan Status versus Bank Funded Amount (Bivariate Histogram with Heat Map):

Type of Plot: Heat map with histogram.
Objective: To show the proportion of bank-funded amounts in relation to the loan statuses.
X-axis: “Loan Status” (loan_status).
Y-Axis: “Funded Amount” (funded_amnt).
Insight: It forms a heat map where the color varies within the categories according to the number of loans taken in that particular category facilitating easy marking of the highest combination.



Part 4

Graph 1-:
Loan Status against Average Bank Investments
Type: Bar plot
X-axis: Loan status (e.g., Fully Paid, Charged Off, etc.)
Y-axis: Funded amounts mean value
Insight:
The graph shows bank investment amounts for every loan status category.
It is moreover possible to observe the relationship between different loan statuses and the average investment made by the bank in those loans. For instance, if the average investment in “Fully Paid” loans is significantly higher than the rest, it may suggest that the bank aims to cater more to the low-risk borrowers.



Part 5

Graph 2-:

Type of Chart: Bar Cart

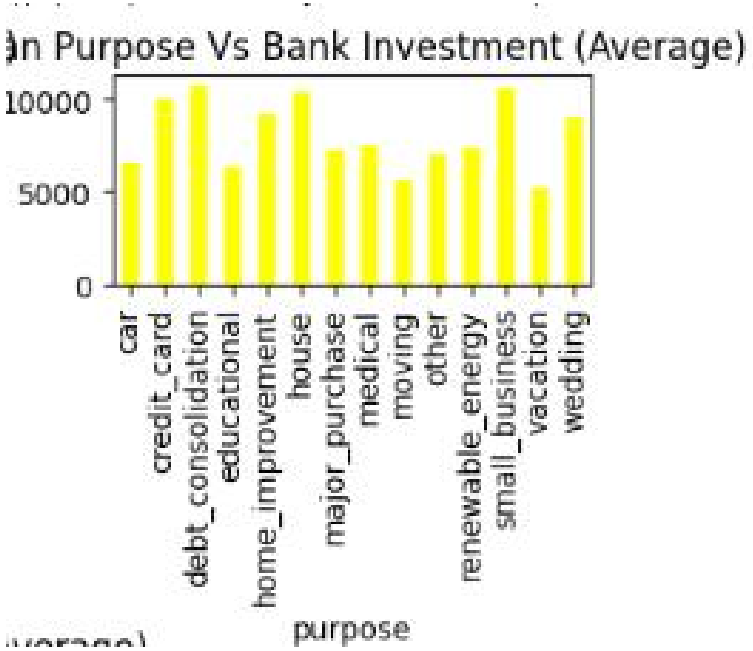
X-axis: Reasons for Loan Usage (For example – Credit Card, Home Improvement, Debt Consolidation etc.)

Y-axis: Average Bank Investment

Understanding:

This graph demonstrates the average amount that a bank invests in loans of different purposes. It can show which lending categories usually receive larger banks' investments. For instance, it is likely that loans for 'home improvement' or 'debt consolidation' will have larger averages than loans for 'vacation' or 'small business' purposes.

The findings suggest which loans the bank is likely to provide more funds to.



Part 6

Graph 3: Investment by State and Commercial Banks (Top 5 by Average Bank Investment)
Graph Type: Bar plot (Top 5 states by average bank investment)
X-Axis: States (CA, NY, TX and so on)
Y-Axis: Average bank investment
Analysis:
This graph depicts the 5 states in which the bank lends the highest amounts (on average).
It helps to pinpoint certain geographic areas that have banks capable of disbursing higher average loans.
For instance, some regions, for example, the state of California or Texas might record higher average investments than other territories because of the size of the population or the population's affluence.

Graph 4: Loan Term with Bank Investment Average
Type: Bar Chart
X-Axis: Loan term in months (for example, 36 months, 60 months)
Y-Axis: Bank investment average
Description:
This diagram depicts the relationship between the duration of the loan (term) and the mean bank investment made.
In this case, it is expected that longer-term loans (60 months) will have a greater average investment than shorter course loans (36 months), implying that a bank may tend to commit more resources in the longer duration loans.
This situation may also suggest that it is preferable for the bank to fund such type of loans at the longest tenor possible since the higher returns are expected after some considerable period.

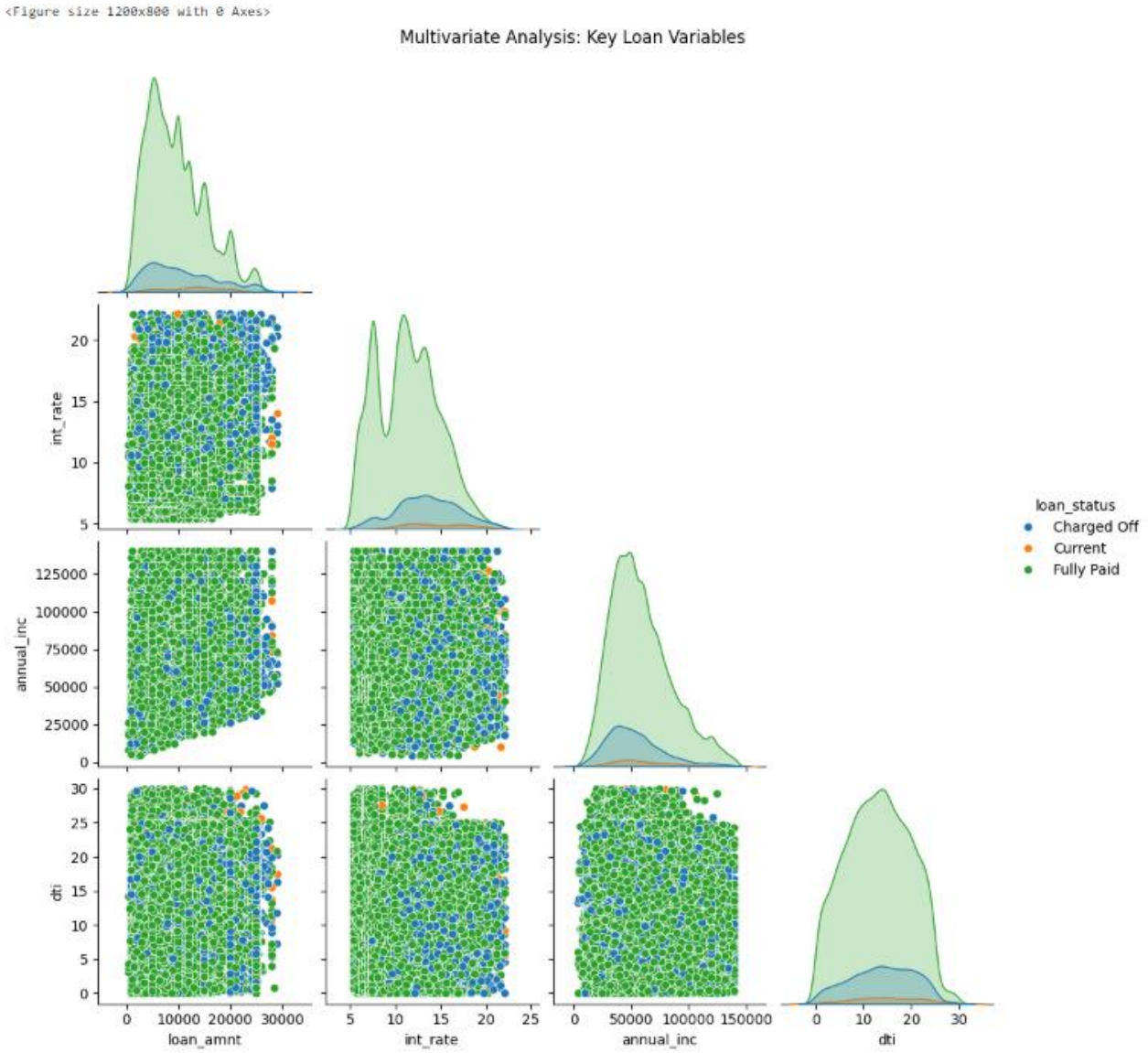
Multivariate Analysis

Examination of Connections through Visualisation: The pairplot is useful to quickly check the relationships that exist between several loan-related features and loan status as a dependent variable.

Loan Status Comparison:

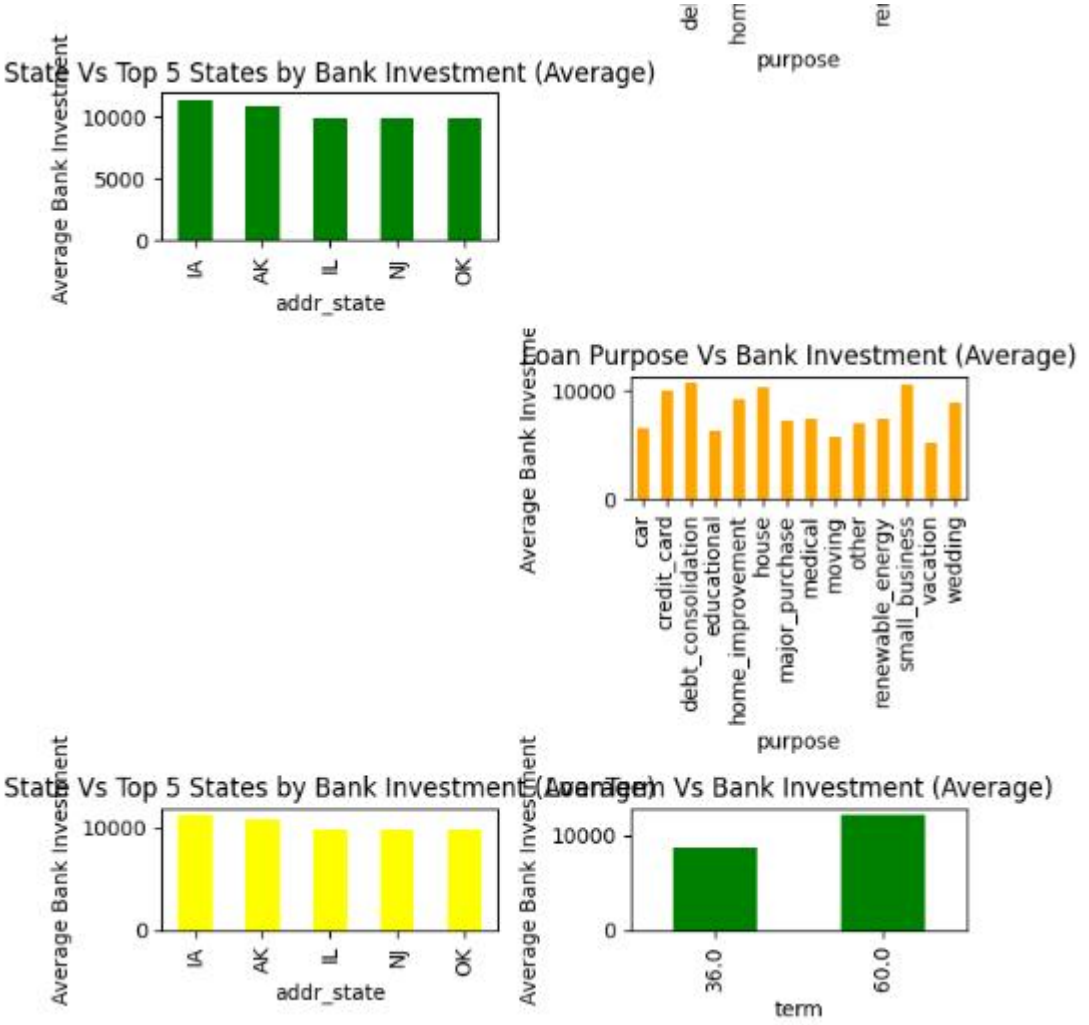
This also helps in differentiating clearly the different categories of loans e.g. Fully paid, Charged off amongst others.
For instance, it can be noted that the loans with higher interest rates are more likely to be charged off as opposed to those loans with lower rates which are paid off in full.
Distributions of variables:

The Circos view (displays on off-diagonal where appropriate) shows how the entire two-dimensional distributions of the variables and can show how some loan categories (e.g. low DTI ratio or high earning potential) are concentrated on value ranges of certain variables.



Conclusion of Bivariate Graph

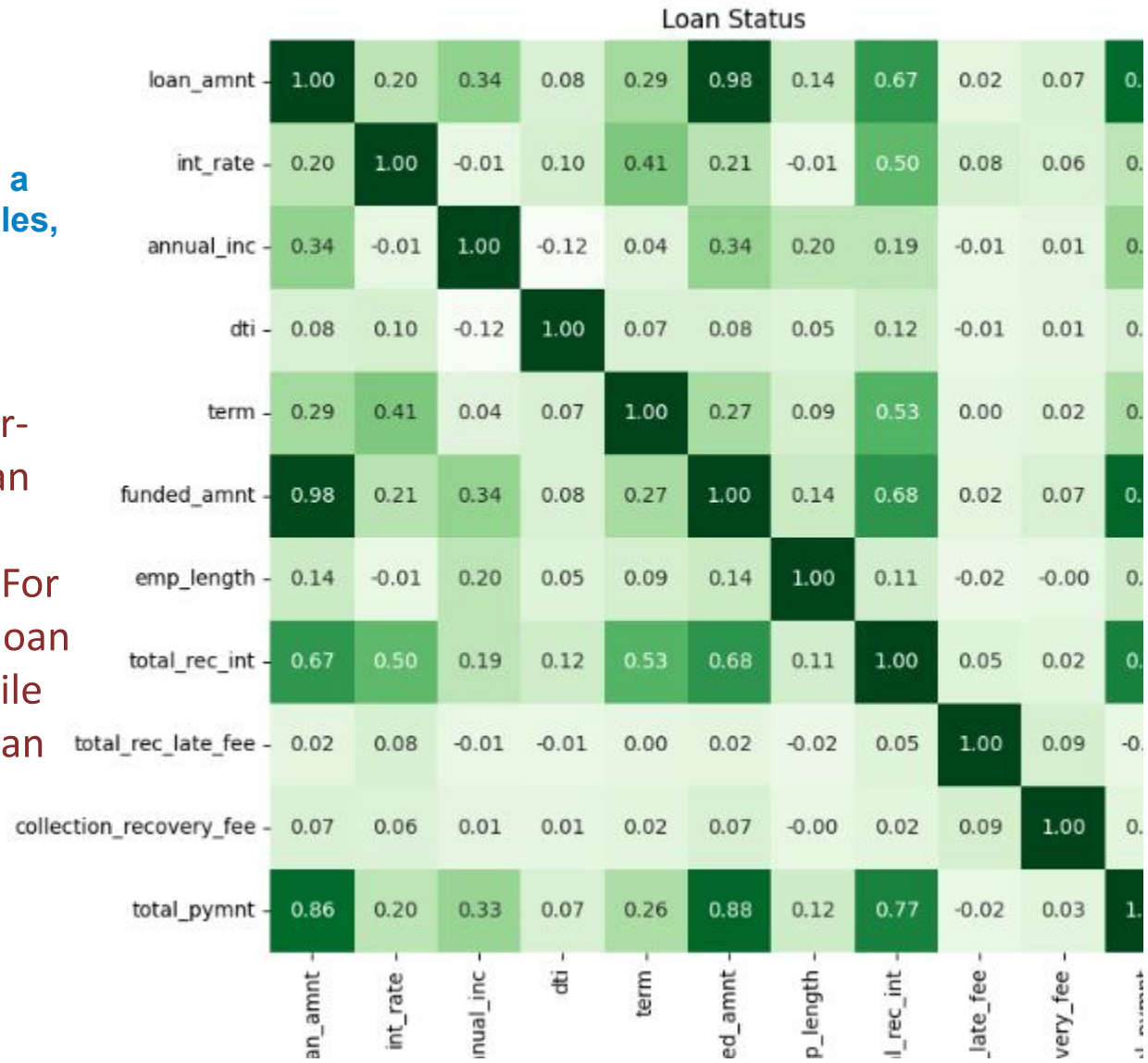
1) **Loan Status:** The mean amount of investment with respect to the loan status varies with the corresponding risk appetite of the bank (i.e. the bank tends to invest more money on the “Fully Paid” loans).
2) **Loan purpose:** The bank invests different amounts of money, depending on the purpose of the loan offered, indicating the most supported area by the bank’s investments.
3) **Top states:** it is possible that the average degree of the bank’s investments is difficult to achieve in average; it may be that the bank’s funding strategies are focused towards particular geographical regions with some states recording a higher average funding.
4) **Loan Term:** Considering loan term these determinants also include the loan structure, with the restriction of the term being possibly an investment motivation for the bank; higher term envelopes attracting more capital.



Correlation Heatmap of Loan Attributes

Objective:
This analysis aims to explore the relationships between key loan-related variables by calculating their correlations. The heatmap is a powerful visualization tool for understanding how different variables, such as loan amount, interest rate, and annual income, are interrelated.

The heat map helps in a graphical representation of the inter-relationships of attributes of the loan. Strong correlations can be helpful in advanced analysis, where weak or negative correlations may reveal some interesting or alarming areas. For instance, the positive correlation between the amount of a loan and the amount allocated to it is a reasonable assertion, while the negative correlation of the interest shed vs the size of loan taken out, brings sharply to focus certain aspects of lending behavior.



Recommendations When Disbursing the Loan

1) Check the credit score of the person.

2) Check the Job stability.

3) Provide lower interest rate to the customer.

4) provide Moderate Loan Amounts.

5) Focus on the purpose of Loan.

6) Check the past Loan status if Any

THANK YOU

Loan Data Analysis