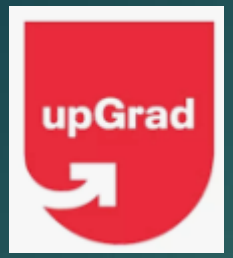


EXECUTIVE PG PROGRAMME IN MACHINE LEARNING & AI

Lending Club Case Study

PHANIRAJ CHAKILAM RAJESH KUMAR MALVIYA

Problem Statement

Understanding loan default factors using loan data to improve risk assessment and decision-making in lending.

Analysis approach

- **Data Understanding and Preprocessing:** Load and clean the loan dataset.
- **Exploratory Data Analysis (EDA):** Perform univariate, bivariate, and multivariate analyses.
- **Feature Selection and Engineering:** Identify key variables influencing loan default & with a Conclusion

Load the loan dataset

- Import the necessary libraries, ‘pandas’ for data handling and load the loan dataset into a data frame

  Loading Dataset (read_csv):

  Use **pd.read_csv()** to load a loan dataset from a CSV file into a DataFrame.

Results → Total 111 columns (snapshot below and more details refer python file)

Display the first few rows of the dataset to understand its structure

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term
0	1077501	1296599	5000	5000	4975.0	36 months
1	1077430	1314167	2500	2500	2500.0	60 months
2	1077175	1313524	2400	2400	2400.0	36 months
3	1076863	1277178	10000	10000	10000.0	36 months
4	1075358	1311748	3000	3000	3000.0	60 months

Data understanding

- Explore the dataset to understand its columns, data types and missing values

Data Understanding (info, describe, isnull):

info(): Displays information about columns, non-null counts, and data types.

loan_data.head(): Display the first few rows of the dataset to understand its structure

describe(): Provides summary statistics (mean, min, max) for numerical columns.

isnull().**sum**(): Counts missing values in each column.

```
loan_data.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past_12m	p
0	1077501	1296599	5000	5000	4975.00	36 months	10.65%	162.87	B	B2	...	NaN	NaN	
1	1077430	1314167	2500	2500	2500.00	60 months	15.27%	59.83	C	C4	...	NaN	NaN	
2	1077175	1313524	2400	2400	2400.00	36 months	15.96%	84.33	C	C5	...	NaN	NaN	
3	1076863	1277178	10000	10000	10000.00	36 months	13.49%	339.31	C	C1	...	NaN	NaN	
4	1075358	1311748	3000	3000	3000.00	60 months	12.69%	67.79	B	B5	...	NaN	NaN	

5 rows × 111 columns

Data understanding – Insights#1

The **loan amount** statistics provided summarizes the statistical distribution of loan amounts in a dataset

- 39,717 indicates the total number of loan records in the dataset
- Mean loan amount of $11,219.44 represents the typical or average loan size granted by the institution
- minimum loan amount of $500 represents the smallest loan granted by the institution
- 75th percentile (Q3) value of $15,000 indicates that 75% of loans are below this amount

Loan Amount Statistics

count	39717.00
mean	11219.44
std	7456.67
min	500.00
25%	5500.00
50%	10000.00
75%	15000.00
max	35000.00

Business Implications:

- Understanding Borrower Preferences:- These statistics provide insights into borrower preferences regarding loan sizes. They help in aligning product offerings (loan amounts) with customer demand.
- Risk Management:- Knowledge of loan amount variability (standard deviation) helps in assessing risk exposure. Higher variability may require different risk mitigation strategies.
- Customer Segmentation:- Loan amount statistics can aid in segmenting customers based on funding needs, enabling targeted marketing and personalized services.

Data understanding – Insights #2

The **loan default** statistics provided indicate the distribution of loan statuses within the dataset

```
Default Rate:
 loan_status
Fully Paid     82.961956
Charged Off    14.167737
Current         2.870307
```

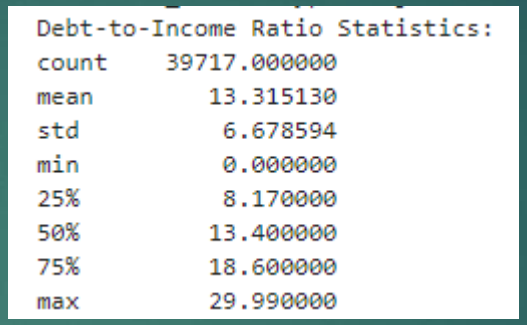
Business Implication: A high percentage of fully paid(82%) loans indicates that a majority of borrowers have successfully met their repayment obligations. This can be a positive indicator of borrower creditworthiness and loan performance within the portfolio.

Portfolio Performance:
Monitoring and managing the charged off(14%) loans is crucial for optimizing portfolio performance and minimizing financial losses.

Lending Strategy:
Lenders may adjust their lending strategies based on the observed default rates to mitigate risk and improve overall portfolio performance

Data understanding – Insights #3

The Debt-to-Income (**DTI**) Ratio statistics provided offer insights into the financial health and risk assessment of borrowers within the dataset.



```
Debt-to-Income Ratio Statistics:
count    39717.000000
mean        13.315130
std          6.678594
min          0.000000
25%          8.170000
50%         13.400000
75%         18.600000
max         29.990000
```

Business Implication:

1. Lower mean (13%) indicates – on average borrowers are managing their debt obligations within a reasonable portion of their income
2. Borrowers in 75th percentile may have a higher risk of financial strain compared to those in 25th percentile

Portfolio Performance:

Higher mean and maximum DTI ratios may indicate increased risk of default among borrowers with elevated debt levels relative to their income

Lending Strategy:

Identifying borrowers with excessively high DTI ratios can help lenders proactively address potential default risks and implement appropriate risk mitigation measures

Data Cleaning

- **Removing columns that are not required or not useful**

```
#List down unused columns
unused_Columns = ["member_id", "desc","zip_code", "mths_since_last_delinq","mths_since_last_record","next_pymnt_d",
                  "url", "emp_title","tax_liens", "title", "pymnt_plan", "initial_list_status", "collections_12_mths_ex_med",
                   "policy_code", "application_type", "application_type", "acc_now_delinq", "chargeoff_within_12_mths",
                   "delinq_amnt", "tax_liens"
                  ]

loan_data.drop(labels=unused_Columns, axis=1, inplace=True)
```

```
#dimensions of the DataFrame, first element indicates the number of rows (observations) in the DataFrame,
#second element indicates the number of columns (features) in the DataFrame
loan_data.shape
```

```
(39717, 39)
```

Data Cleaning

Missing Employment Length Data

loan_data.emp_length.isnull().sum():- provides a count of how many borrowers have missing employment length information. This is crucial for assessing data completeness and potential impacts on loan decision-making

```
**********************************************************************
Null count before data cleaning 1075
Null values after data cleaning 0
**********************************************************************
```

Missing Bankruptcy Records

loan_data.pub_rec_bankruptcies.isnull().sum() provides a count of how many borrowers have missing bankruptcy record information. This is important for assessing data completeness and potential impacts on credit risk assessment

```
**********************************************************************
Total null value in pub_rec_bankruptcies 697
Null values after data cleaning 0
**********************************************************************
```

Exploratory data analysis (EDA)

Performing Exploratory Data Analysis (EDA) involves examining and visualizing data to understand relationships, distributions, and patterns within a dataset. Here one can conduct univariate, bivariate, and multivariate analyses using Python with pandas, matplotlib, and seaborn libraries

Univariate Analysis
Univariate analysis focuses on exploring individual variables to understand their distribution and characteristics

Univariate Analysis (histplot, countplot): Visualizes the distribution of individual variables using histograms for numerical variables and bar charts for categorical variables.

Bivariate Analysis
Bivariate analysis explores relationships between pairs of variables to identify correlations or patterns.

Bivariate Analysis (scatterplot, boxplot): Examines relationships between pairs of variables using scatter plots to show correlation and box plots to compare distributions across categories.

Multivariate Analysis
Multivariate analysis examines relationships among multiple variables simultaneously.

Multivariate Analysis (heatmap): Displays correlations among multiple variables using a heatmap of the correlation matrix.

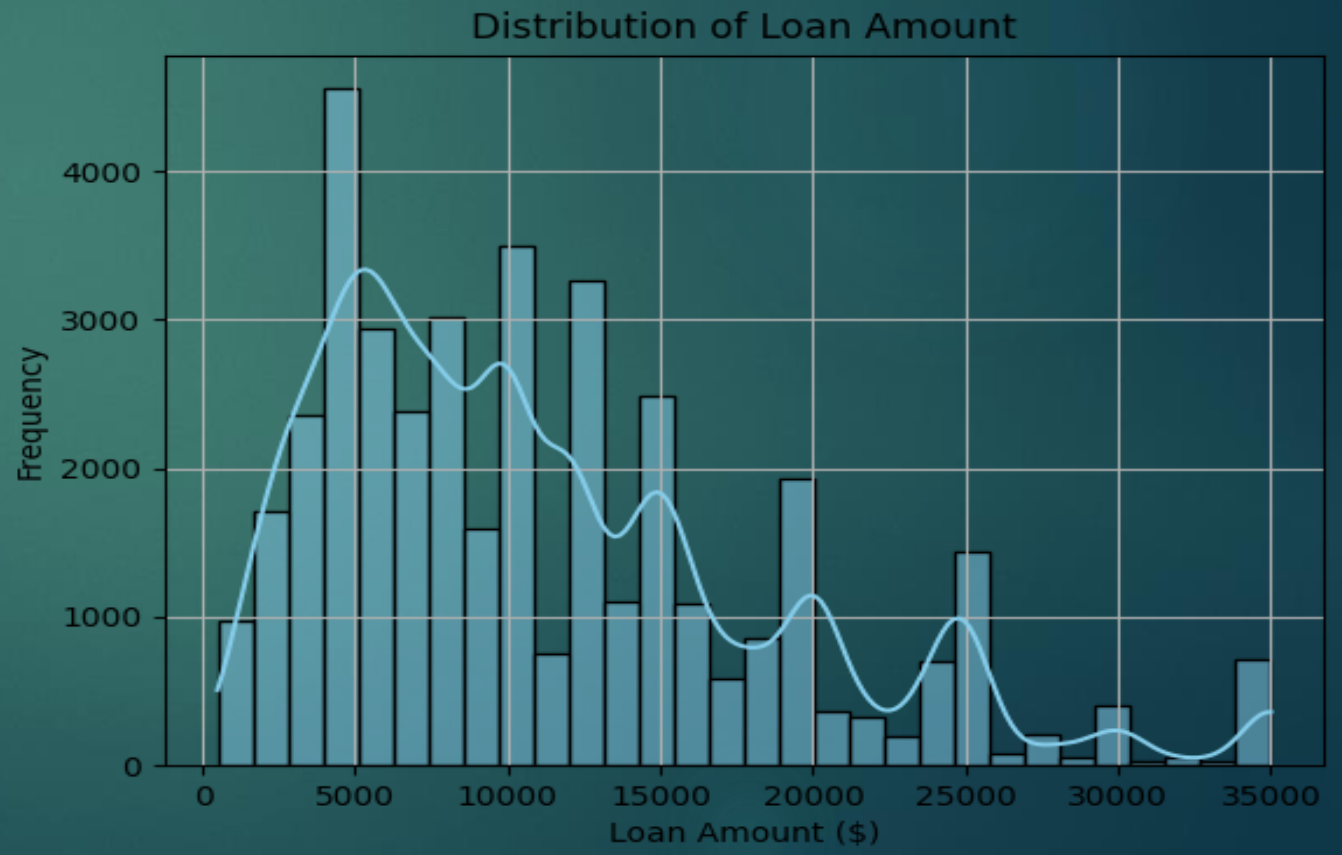
Univariate Analysis – Insights(#1)

Distribution of Loan Amount (Histogram)

Below represents how the loan amounts are distributed across different ranges
Most common amount range ($5000)
Spread/Diversity of Loan Amounts – between $5000 & $25000
Below distribution is left skewed indicates common loan amount preferences ($5000)

Business Implications:-

Understanding Borrower preferences → Distribution of loan amount revealing popular loan sizes preferred by borrowers, which can influence product offerings & Marketing strategies
Risk Assessment → Higher concentration in certain range indicate potential credit risk / market demand.
Product Strategy → Identifying popular loan sizes can guide product structuring & segmentation
Decision Making → All this data helps in loan approval criteria, risk assessment & customer targeting.



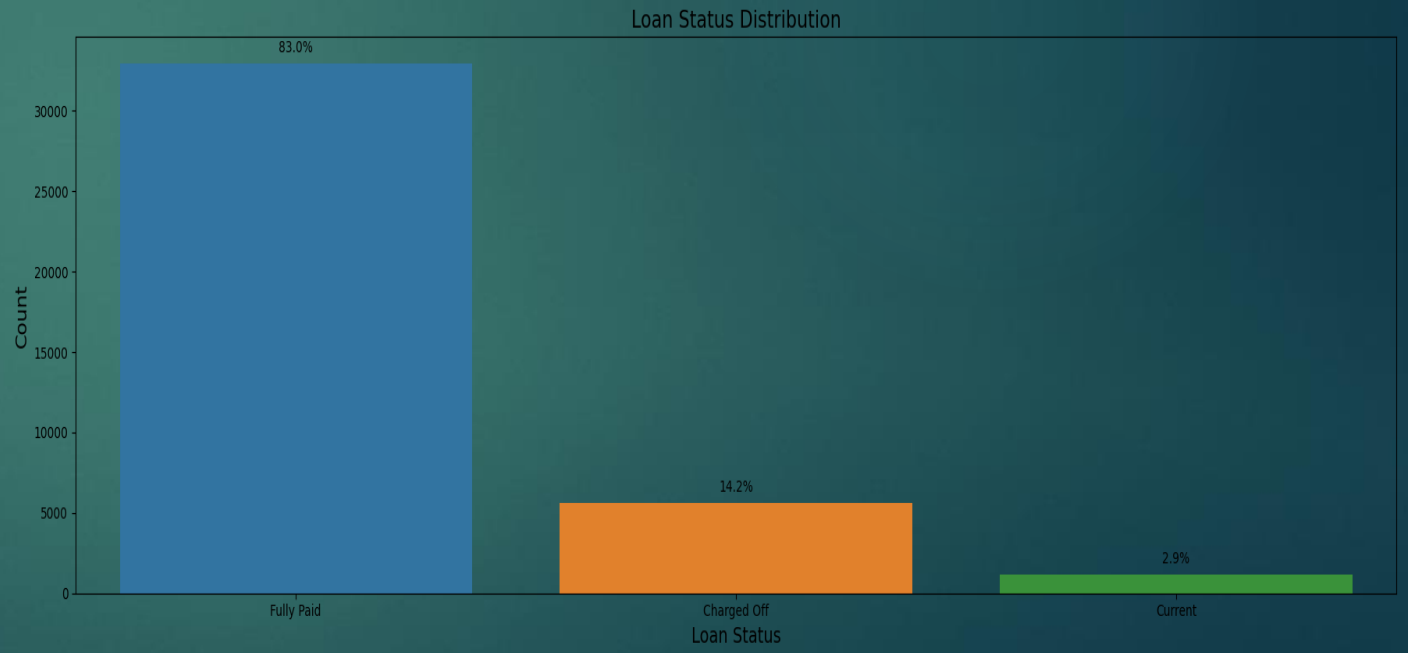
Univariate Analysis – Insights(#2)

Loan Status Distribution (Bar Chart)

Below chart represents frequency or count of different categories (loan statuses) within a categorical variable
Each bar represents a category (loan status), and the height of the bar corresponds to the number of occurrences (count) of that category in the dataset
Fully Paid (above ~30000), Charged Off(~5000) and Current – Performance of loans in terms of repayment

Business Implications:-

Portfolio Management → Loan status assists in areas that requires attention (example: Charged Off)
Risk Assessment → Potential default risk
Customer Engagement → customer engagement strategies based on repayment behavior



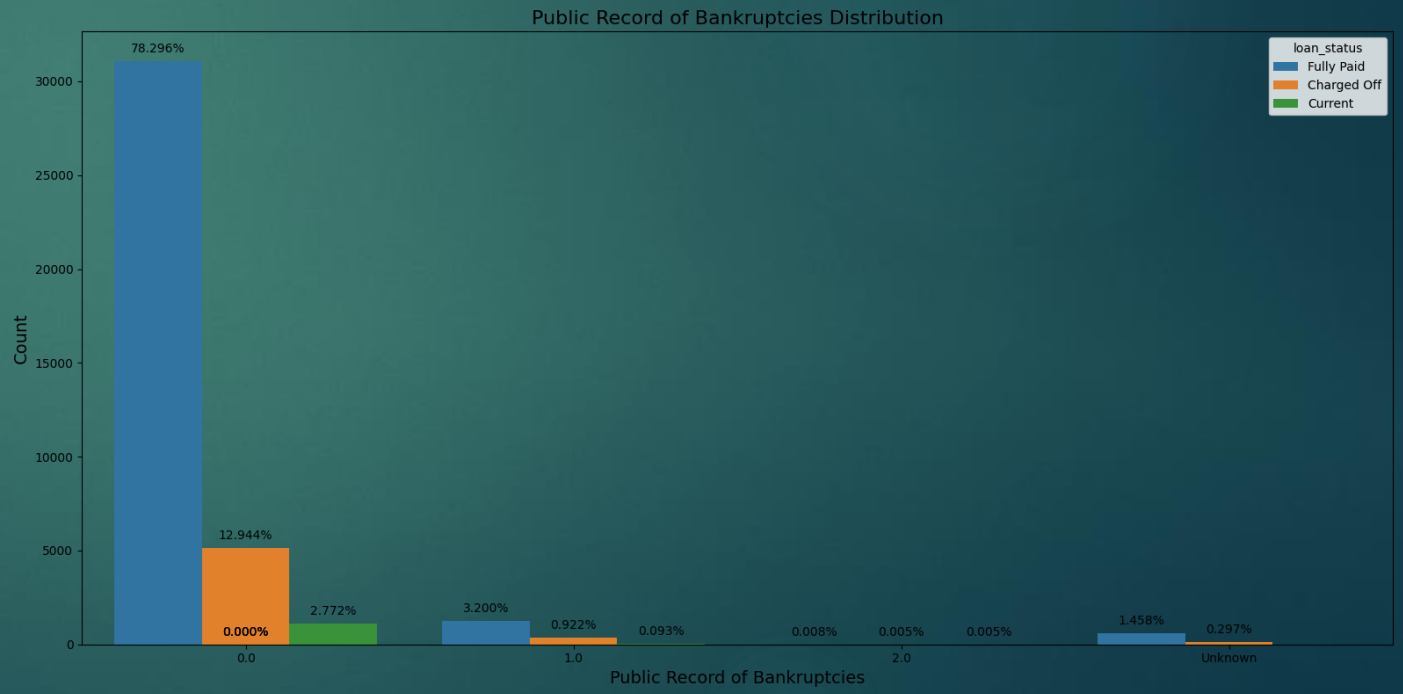
Univariate Analysis – Insights(#3)

Public Record of Bankruptcies Distribution (Bar Chart)

The chart below represents the frequency or count of different categories (Pubic Record of Bankruptcies) within a categorical variable
Each cluster of bars represents a category (Pubic Record of Bankruptcies), and the height of the bar corresponds to the number of occurrences (count) of that category in the dataset
#

Business Implications:-

Portfolio Management → Loan status assists in areas that require attention (example: Public record of bankruptcies: 1-2)
Risk Assessment → Potential default risk if applicant have previous bankruptcy record



Bivariate Analysis – Insights (#1)

Loan Amount vs. Debt-to-Income Ratio (Scatter Plot)

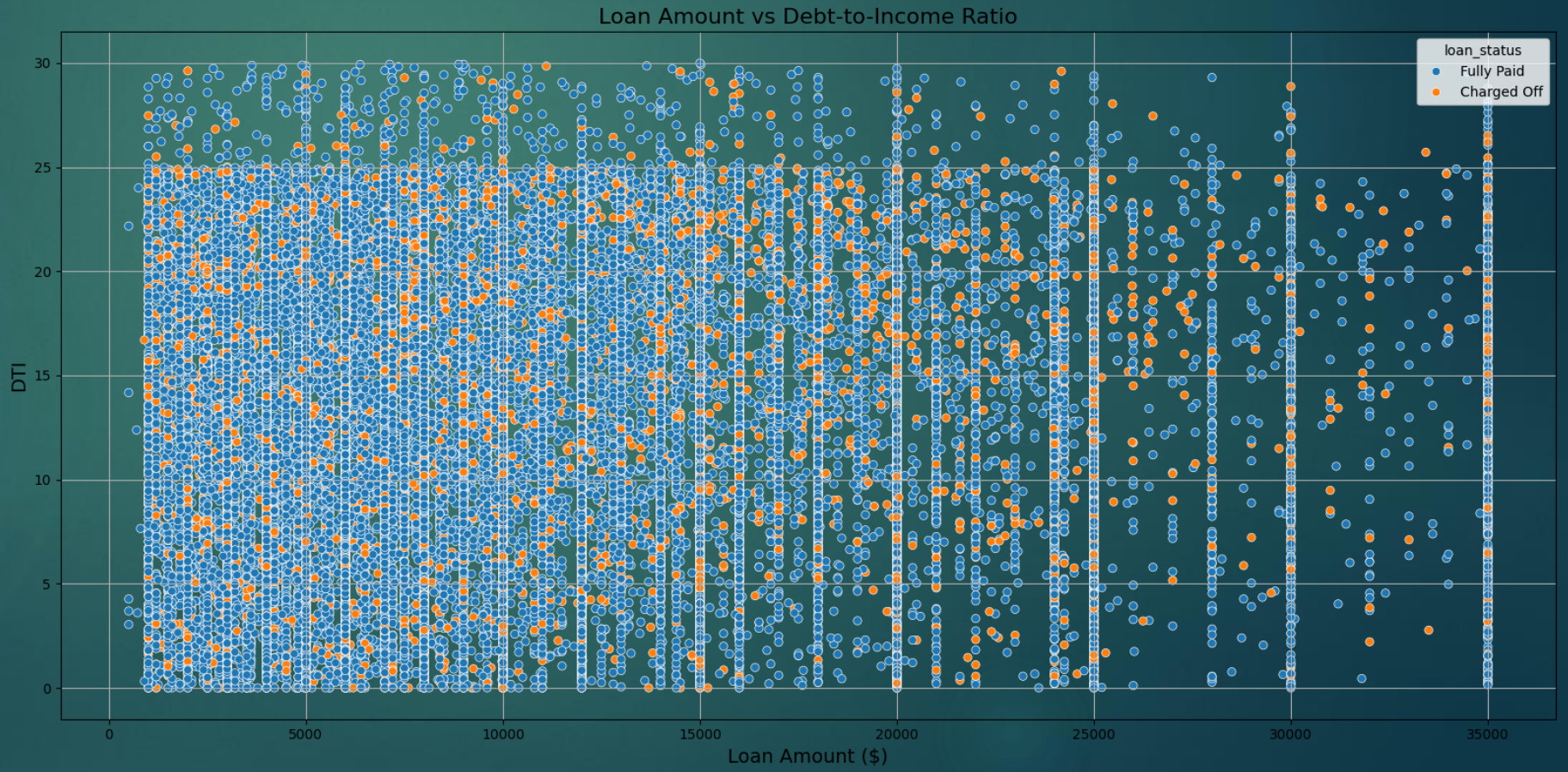
Below represents in X-axis is loan amount requested by borrowers. Each points position along the x-axis is size of loan.

And in Y-axis, represents the debt-to-income ratio, which is the ratio of a borrower's total monthly debt payments to their monthly gross income. Each point's position along this y-axis signifies the borrower's DTI ratio.

Business Implications:-

#Risk Assessment → **High Loan Amounts with High DTI →** Points clustered at high loan amounts and high DTI ratios may indicate borrowers with potentially higher financial risk.

from scatter plot – it indicates that high DTI ratio is 25 & till amounts ($20000) when compared between DTI ratio between 25 – 30 & Loan amounts between $20000 - $ 30000



Bivariate Analysis – Insights (#2)

Loan Status vs. Loan Amount (Box Plot)

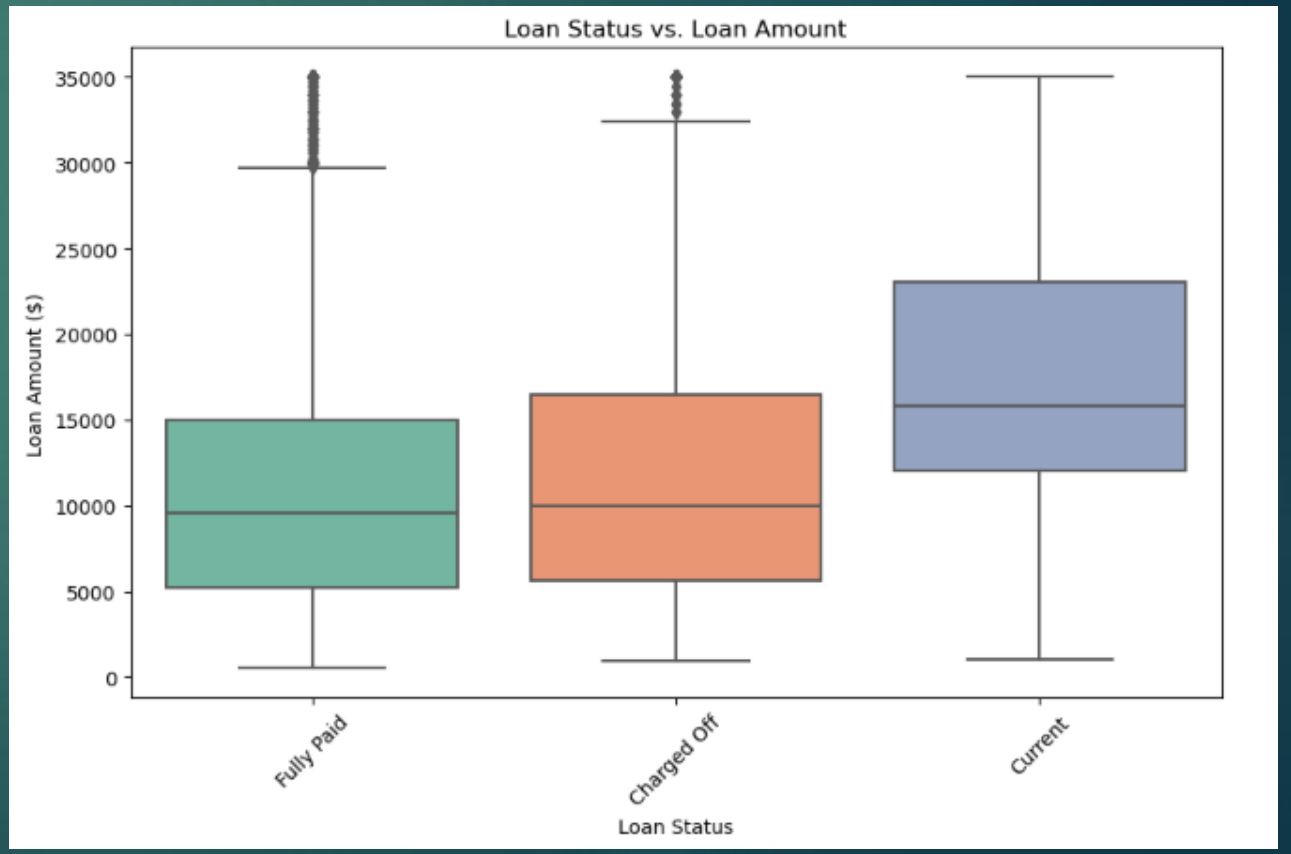
Below box plot shows the distribution of loan amounts for each loan status category (e.g., fully paid, charged off, current). It visually represents median, quartiles, and potential outliers for loan amounts associated with each loan status.

Business Implications:-

Risk Assessment → Charged Off having higher box plot indicates higher median compared to fully paid loans, suggests that higher loan amounts might be associated with increased default risk

Loan Amount thresholds → example as Current between loan amounts ~$12000 and ~$24000. Lenders adjust approval criteria based on thresholds

Outliers → Observed outliers between loan amounts $30000 and $35000. Lenders should assess whether these borrowers have the ability to repay such large amounts based on their income, debt-to-income ratio, credit history, and other factors.



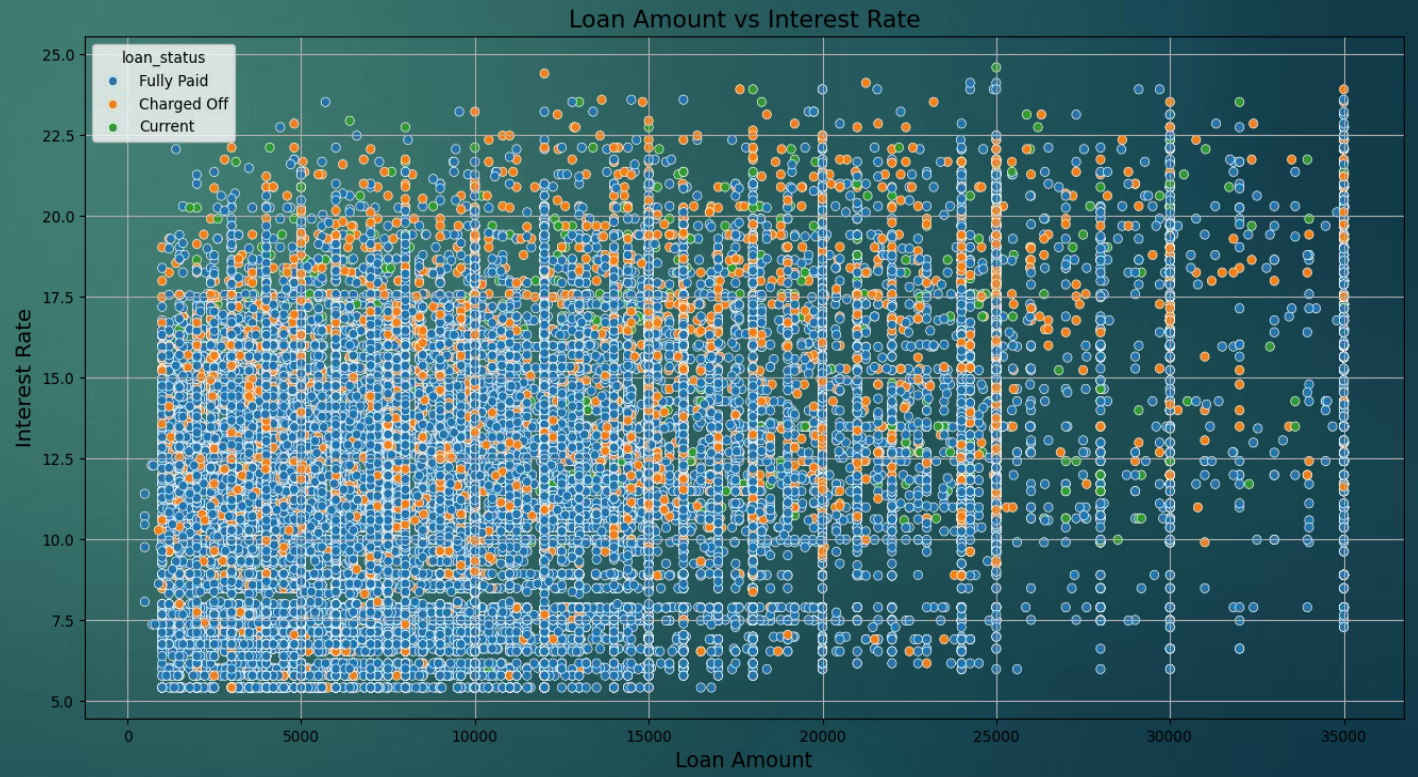
Bivariate Analysis – Insights (#3)

Interest Rate vs. Loan Amount (Box Plot)

Below represented in the X-axis is the loan amount requested by borrowers. Each point position along the x-axis is size of loan.
And in Y-axis, represents the interest rate.

Business Implications:-
#Risk Assessment → **High Loan Amounts with Interest Rate** → Points clustered at high loan amounts and interest rates may indicate borrowers with potential default risk.

The scatter plot – indicates that the highest interest rate is ~25 & till amounts ($35000)
If the interest rate is 12.5% and above there is more chance that the loan will default



Multivariate Analysis – Insights

Objective: Explore interactions among three or more variables in a 3D space for visualizing relationships between multiple variables simultaneously while using color to represent different categories of loan status.

Interpretation:

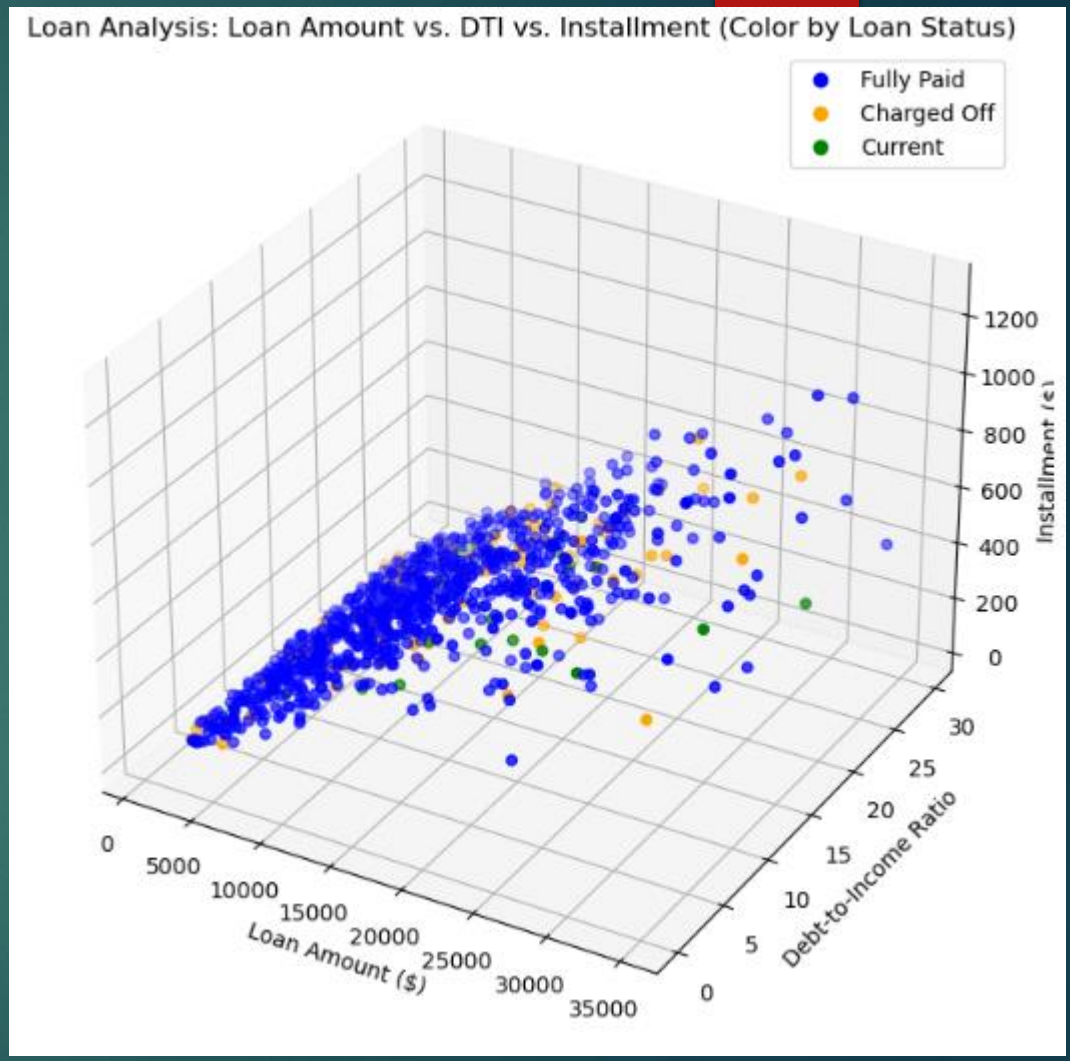
- **X-axis** (Loan Amount) represents total loan amount distributed to borrowers. **Y-Axis** (Debt-to-Income Ratio) indicates the borrower's debt-to-income ratio, which reflects their ability to manage debt based on income. Z-Axis (Installment) displays the monthly installment amount borrowers are required to pay.

Each data point (representing a loan) is color-coded based on its loan status (Fully Paid, Charged Off, Current).

- Fully Paid (Blue): Loans that have been fully paid off by borrowers.
- Charged Off (Orange): Loans that have defaulted or been charged off.
- Current (Green): Loans that are currently active and being paid off.

Business Implications:

- Risk Assessment → By examining where Charged Off loans cluster in relation to loan amount, debt-to-income ratio, and installment, lenders can identify riskier loan profiles. For example, higher loan amounts combined with high debt-to-income ratios might indicate a higher likelihood of default.
- Portfolio Analysis → Visualizing the distribution of loan statuses across different segments helps in understanding the composition of the loan portfolio. For instance, a higher concentration of Fully Paid loans in certain ranges might signify successful lending practices in those segments.



- Based on observed patterns, lenders can refine loan approval criteria to mitigate risks associated with high-risk loan profiles

Key variables influencing loan default

Firstly defined the problem statement & then identifying and understand the key variables that significantly influence loan default.

Data set used for analysis:- Loan data set with specific attributes - amount, interest rate, employment length, etc. Target variable Loan Status and its categories - 'Fully Paid', 'Charged Off', 'Current‘

Feature Selection Process:- Selecting key features like loan amount, annual income, debt-to-income ratio, loan term, grade, and employment length that are likely to impact whether a loan will be paid off or result in default.

Key Variables Identified :- Based on so far analysis, loan amount, annual income, and debt-to-income ratio are identified as the most influential factors affecting loan default.

Insights and Interpretations:- The scatter plot, as one example, shows how loan amount and debt-to-income ratio correlate with loan status, certain patterns indicated higher risk of default for certain borrower profiles

Recommendations:- Based on analysis, we recommend incorporating loan amount, income level, and employment length into credit scoring models to better predict loan default risk.

Summary - Key Insights

- Conclusion 1: Employment length appears to be inversely correlated with the likelihood of loan default, indicating that longer employment tenure might reduce the risk of default.
- Conclusion 2: Debt-to-income ratio (DTI) is positively correlated with loan default, suggesting that higher DTI ratios increase the risk of default.
- Conclusion 3: Loan grade and subgrade provided by Lending Club are strong predictors of default, with higher-grade loans exhibiting lower default rates.
- Conclusion 4: Homeownership status is another significant factor, with homeowners exhibiting lower default rates compared to renters.
- Conclusion 5: Loan purpose also influences default rates, with loans for debt consolidation having relatively lower default rates compared to other purposes such as small business or education.
- Conclusion 6: The length of the loan term is inversely related to default rates, with longer-term loans having higher default rates compared to shorter-term loans.
- Conclusion 7: Interest rates tend to be higher for loans that are charged off compared to fully paid loans, indicating a potential association between higher interest rates and increased risk of default.

Conclusion

- Through comprehensive data analysis, we've gained insights into the loan dataset, identifying important variables and their relationships with loan default.
- Key findings include the impact of loan amount, debt-to-income ratio, and borrower characteristics on loan status.
- Moving forward, leveraging these insights can help optimize lending strategies, improve risk assessment processes, and enhance decision-making in the loan approval process.