Loan Approval Analysis Report

A**B *aaya* tha** 7*8.*

NAME - HARSHIT CHOUDHARY
BATCH -DS40

Introduction

Home loans play a crucial role in enabling individuals and families to purchase homes. However, banks and financial institutions must carefully assess applicants before approving loans. Various factors, such as income, credit history, employment status, and property details, influence the decision-making process.

This report presents an **Exploratory Data Analysis (EDA)** of a **home loan approval dataset** sourced from Skill Circle. The objective is to understand the dataset, identify trends and patterns, and provide insights into the factors that affect loan approvals. This analysis does not include machine learning techniques but focuses on **data visualization and interpretation**.

Dataset Overview

When analyzing loan approval data, it is crucial to understand the different factors that influence decision-making. The dataset contains key details categorized into **three sections**:

1. Applicant Details

These attributes provide information about the individual applying for the loan. They help financial institutions assess the applicant's financial stability and creditworthiness.

- **Gender**: Specifies whether the applicant is male or female, which may help in identifying trends in loan approval rates across different genders.
- **Marital Status**: Indicates whether the applicant is single, married, or divorced, as marital status can impact financial stability and repayment ability.
- **Dependents**: Represents the number of dependents (children or other financial

responsibilities), which can affect the applicant’s disposable income.

- **Education**: Specifies the applicant’s education level (e.g., graduate or non-graduate). Higher education levels may correlate with better income stability.
- **Self-Employment Status**: Identifies whether the applicant is self-employed or working in a salaried job. Salaried employees often have more stable incomes, influencing loan approval rates.
- **Income**: Includes both **applicant income and co-applicant income** to determine the total financial capacity to repay the loan.
- **Credit History**: Represents whether the applicant has a history of repaying previous loans. A strong credit history significantly increases the chances of loan approval.

2. Loan Features

These attributes define the characteristics of the loan that the applicant is requesting.

- **Loan Amount**: The total amount requested by the applicant. Higher loan amounts may require better financial credentials to get approved.
- **Loan Term**: The duration of the loan repayment period (e.g., 10, 15, or 20 years). Longer loan terms mean lower monthly payments but higher total interest.
- **Loan Status**: Indicates whether the loan application was **approved or rejected** based on the applicant’s financial profile.

3. Property Information

These details focus on the property for which the loan is being requested. They help banks assess risks associated with different locations.

- **Area Type**: Specifies whether the property is located in an urban, semi-urban, or rural area. Loan approvals may vary based on location due to differences in property value and stability.
- **Property Location**: Provides specific regional details of the property, which can impact its market value and loan eligibility criteria.

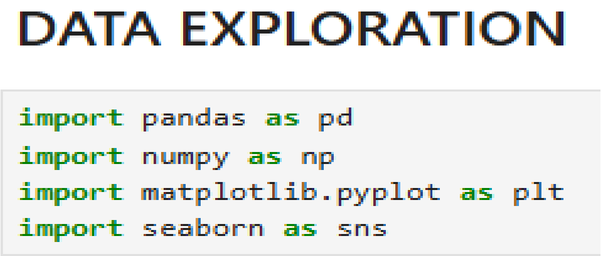
Why These Factors Matter?

Financial institutions use this information to assess **loan repayment risks**. A combination of **income, credit history, and loan features** determines whether an applicant qualifies for a loan. Property details further influence the decision by ensuring the loan is secured against valuable assets.

Methodology

Step 1: Data Exploration

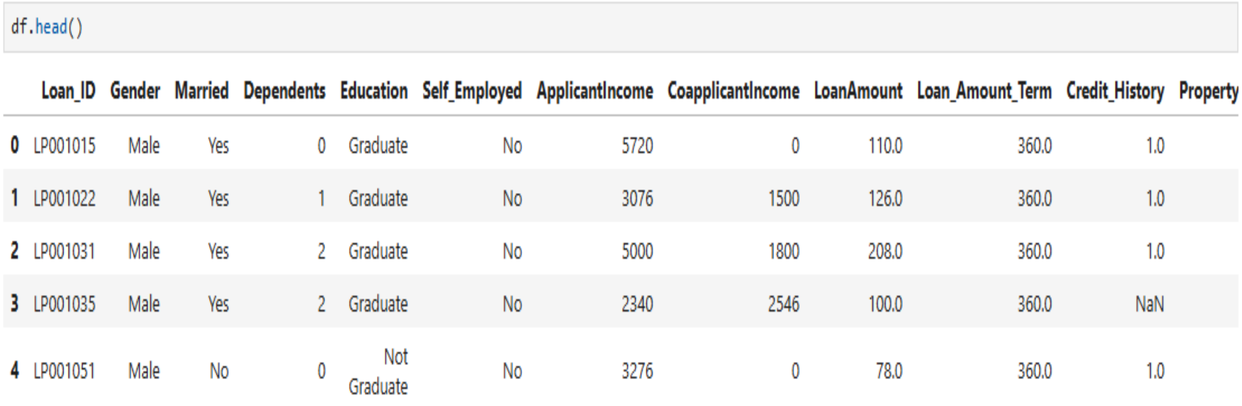
- **Loaded the dataset into a Python environment (Jupyter Notebook)**: The dataset was imported into a Pandas DataFrame to facilitate data manipulation and analysis.



DATA EXPLORATION

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

- **Displayed the first few rows to understand the structure**: Using `df.head()`, we examined the dataset to understand column names, data types, and sample values.

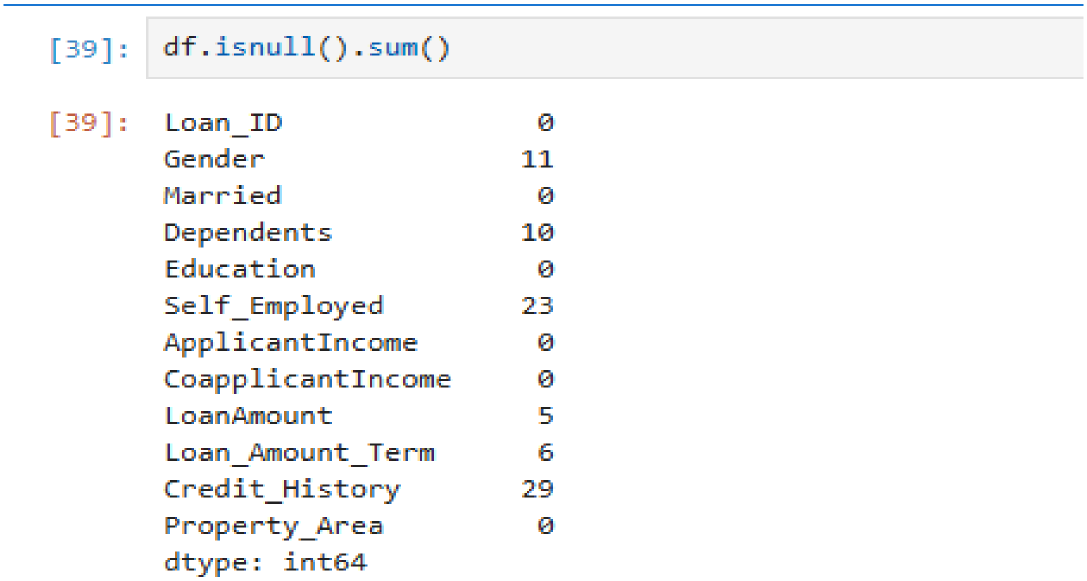


```
df.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0	
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0	
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0	
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN	
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0	

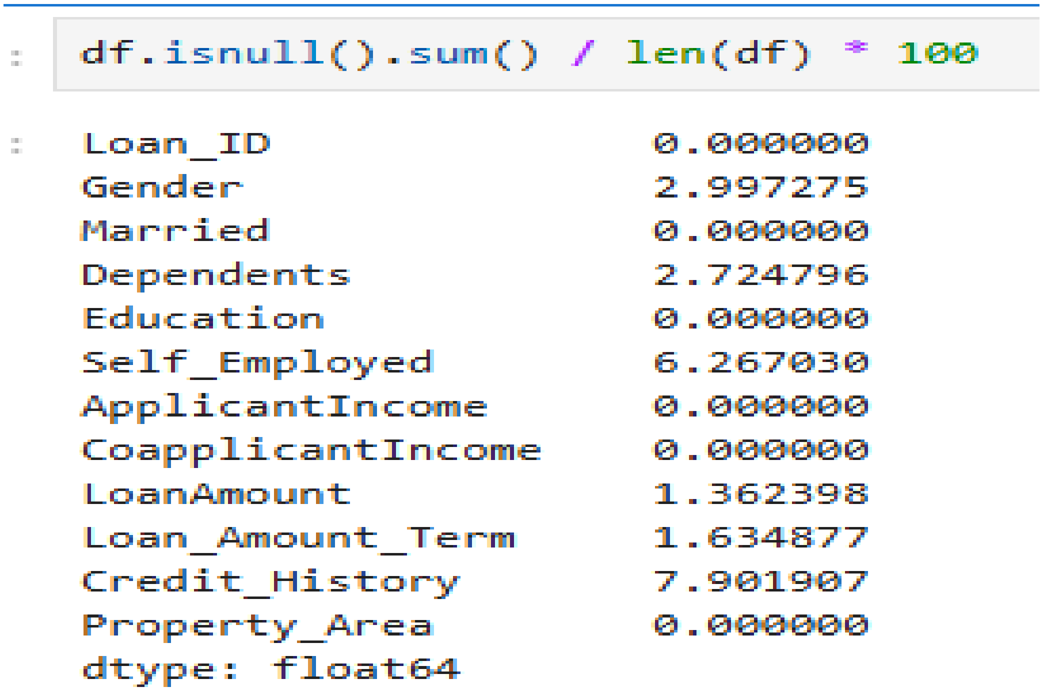
- **Checked for missing values and handled them appropriately**:
  - Used `df.isnull().sum()` to identify missing values.
  - For numerical columns, missing values were imputed using the **mean or median**.
  - For categorical columns, missing values were filled using the **mode (most frequent value)**.
  - In cases where missing data was excessive, rows or columns were removed to maintain data integrity.

CHECK FOR MISSING VALUES



```
[39]: df.isnull().sum()
```

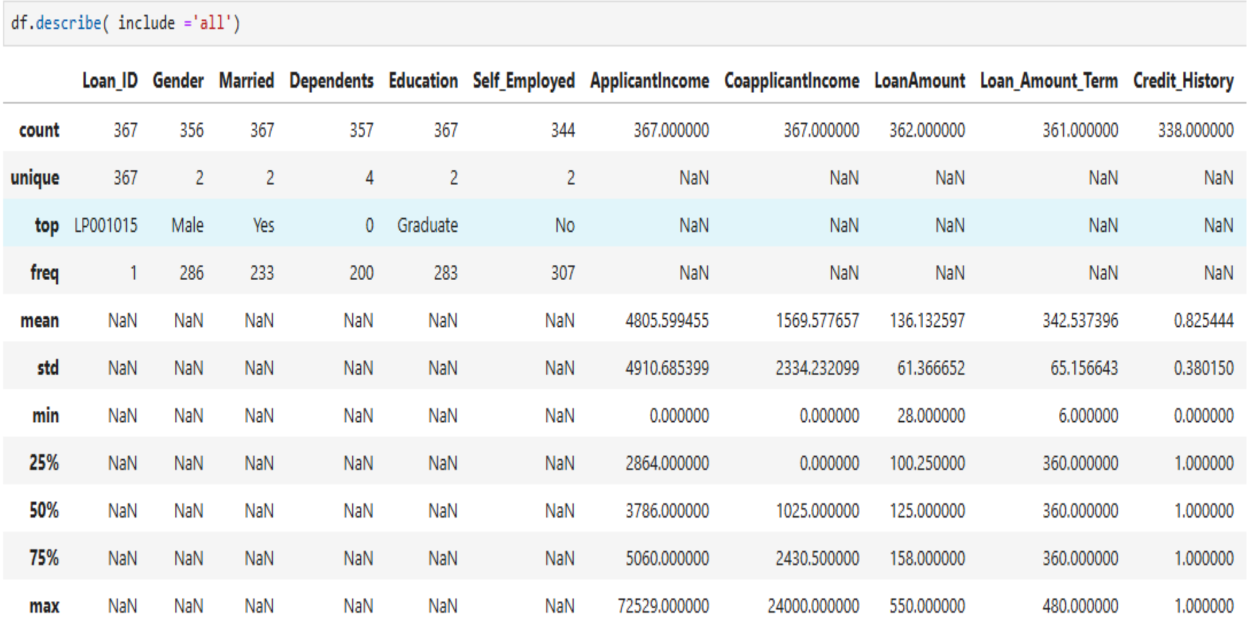
```
[39]: Loan_ID               0
      Gender               11
      Married               0
      Dependents           10
      Education             0
      Self_Employed        23
      ApplicantIncome       0
      CoapplicantIncome     0
      LoanAmount            5
      Loan_Amount_Term      6
      Credit_History       29
      Property_Area         0
      dtype: int64
```



```
: df.isnull().sum() / len(df) * 100
```

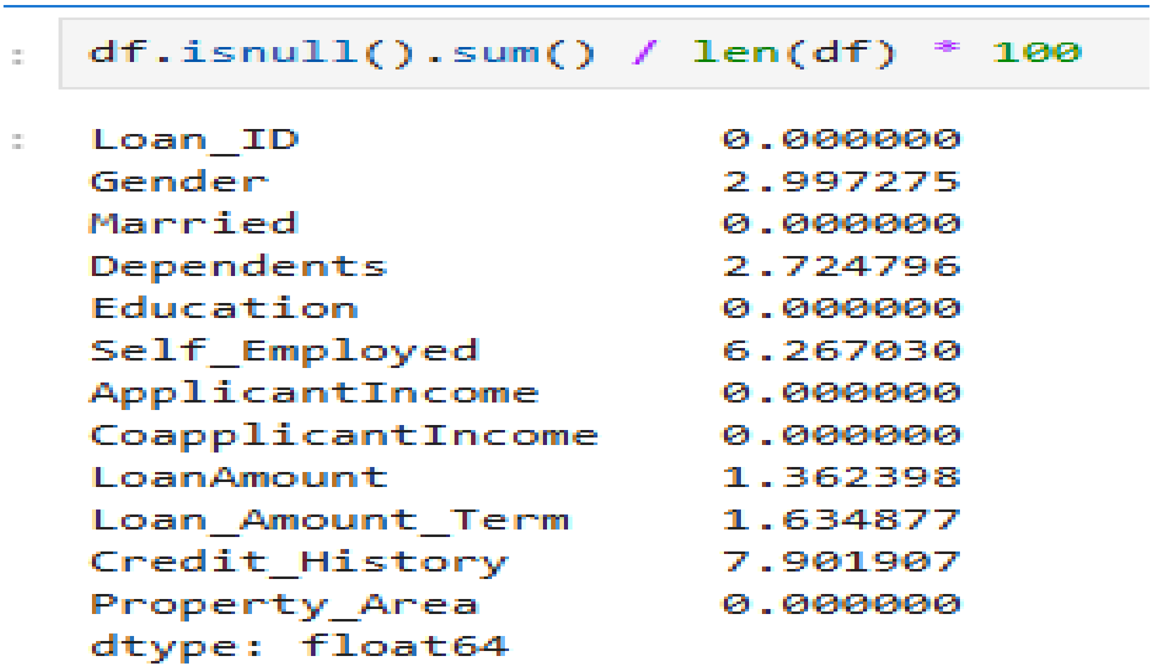
```
: Loan_ID              0.000000
  Gender               2.997275
  Married              0.000000
  Dependents           2.724796
  Education            0.000000
  Self_Employed        6.267030
  ApplicantIncome      0.000000
  CoapplicantIncome    0.000000
  LoanAmount           1.362398
  Loan_Amount_Term     1.634877
  Credit_History       7.901907
  Property_Area        0.000000
  dtype: float64
```

- **Summarized numerical columns with descriptive statistics**:
  - Used `df.describe()` to obtain key statistical insights such as **mean, median, standard deviation, minimum, and maximum values**.
  - Helped in understanding data distribution and detecting possible outliers



```
df.describe( include ='all')
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	367	356	367	357	367	344	367.000000	367.000000	362.000000	361.000000	338.000000
unique	367	2	2	4	2	2	NaN	NaN	NaN	NaN	NaN
top	LP001015	Male	Yes	0	Graduate	No	NaN	NaN	NaN	NaN	NaN
freq	1	286	233	200	283	307	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	4805.599455	1569.577657	136.132597	342.537396	0.825444
std	NaN	NaN	NaN	NaN	NaN	NaN	4910.685399	2334.232099	61.366652	65.156643	0.380150
min	NaN	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	28.000000	6.000000	0.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	2864.000000	0.000000	100.250000	360.000000	1.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	3786.000000	1025.000000	125.000000	360.000000	1.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	5060.000000	2430.500000	158.000000	360.000000	1.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	72529.000000	24000.000000	550.000000	480.000000	1.000000



```
df.isnull().sum() / len(df) * 100
```

```
Loan_ID               0.000000
Gender                2.997275
Married               0.000000
Dependents            2.724796
Education             0.000000
Self_Employed         6.267030
ApplicantIncome       0.000000
CoapplicantIncome     0.000000
LoanAmount            1.362398
Loan_Amount_Term      1.634877
Credit_History        7.901907
Property_Area         0.000000
dtype: float64
```

Data Cleaning & Preprocessing

- Managed missing values by applying imputation techniques. Imputation is the process of replacing missing data with substituted values to maintain dataset consistency. For numerical columns, missing values were replaced using the **mean or median** to avoid bias in analysis. For categorical columns, the **mode (most frequent value)** was used to maintain data distribution. If a column had excessive missing values, those records were either removed or handled using advanced imputation techniques like forward-fill or backward-fill methods. This ensured that data quality remained intact without introducing inaccuracies

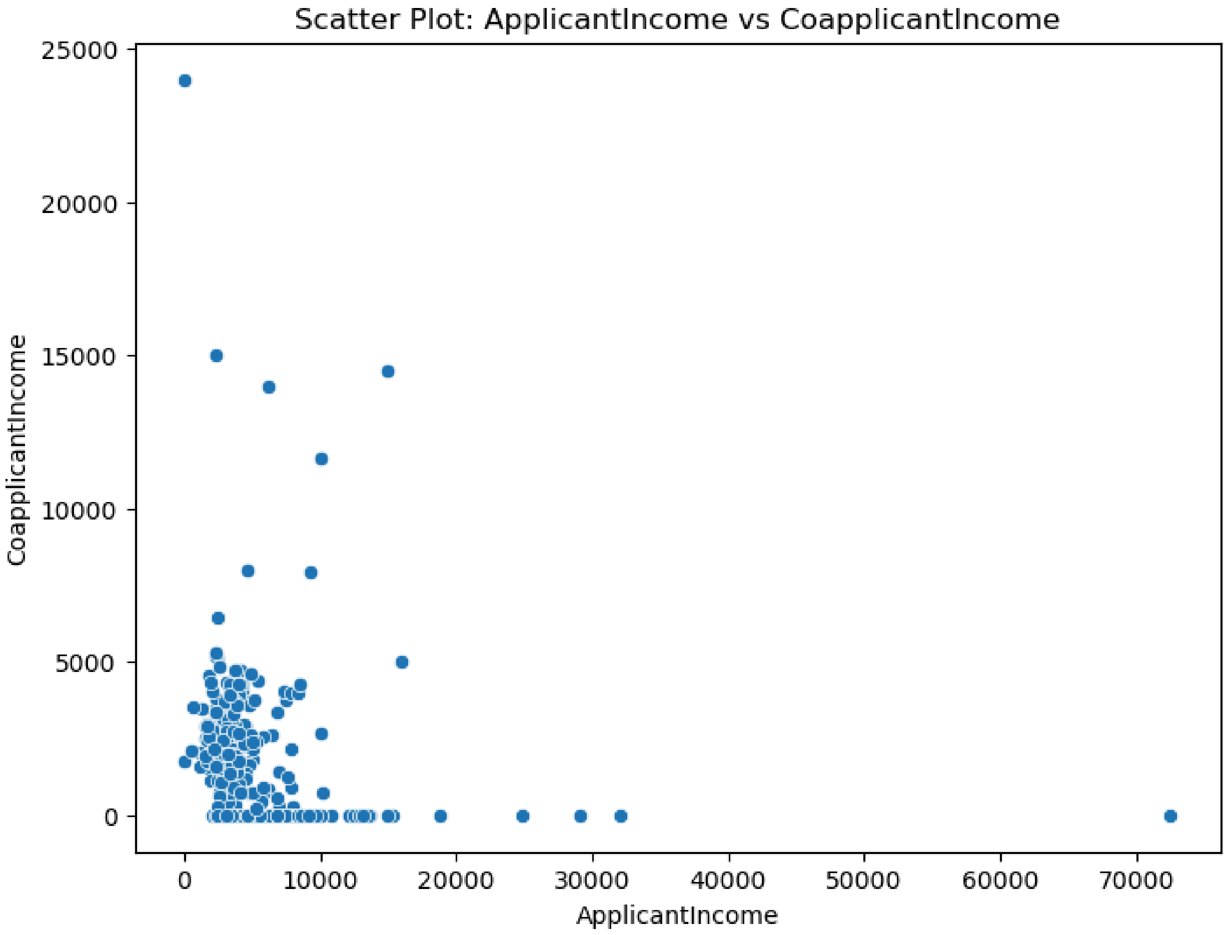
  Converted categorical variables into numerical representations for analysis. Many machine learning models and statistical techniques require numerical data, so categorical variables such as Gender, Marital Status, Self_Employed, and Property Area were converted into numerical formats. This was achieved using techniques like **label encoding** (assigning unique numbers to each category) and **one-hot encoding** (creating binary columns for each category). This transformation ensured that categorical variables could be properly analyzed and used in further processing.

Removed duplicate and inconsistent records to ensure data integrity. Duplicate records can skew the analysis by over-representing certain applicants, while inconsistent records (e.g., missing or contradictory values) can introduce errors. The dataset was cleaned using `df.drop_duplicates()` to remove any identical entries and inconsistencies were corrected by cross-referencing related attributes. Ensuring data integrity is crucial for obtaining accurate insights and maintaining reliability in the analysis.

Data Visualization

Bivariate Analysis

Scatter Plots



Explanation of the Scatter Plot: ApplicantIncome vs CoapplicantIncome

Overview

This scatter plot visualizes the relationship between **ApplicantIncome** (X-axis) and **CoapplicantIncome** (Y-axis). Each point represents a loan applicant, with their respective incomes plotted.

Key Observations

1. **Majority of Data Points are Clustered at Low Values:**
   - Most applicants have incomes below **10,000**.
   - Many coapplicants have incomes around **0**, suggesting that a significant portion of applications do not have a coapplicant.
2. **Few High-Income Applicants:**
   - There are a few data points where **ApplicantIncome** exceeds **20,000** and even goes beyond **70,000**.

- Similarly, there are a few coapplicants with income above **10,000**.

3. Sparse Data in Higher Income Ranges:

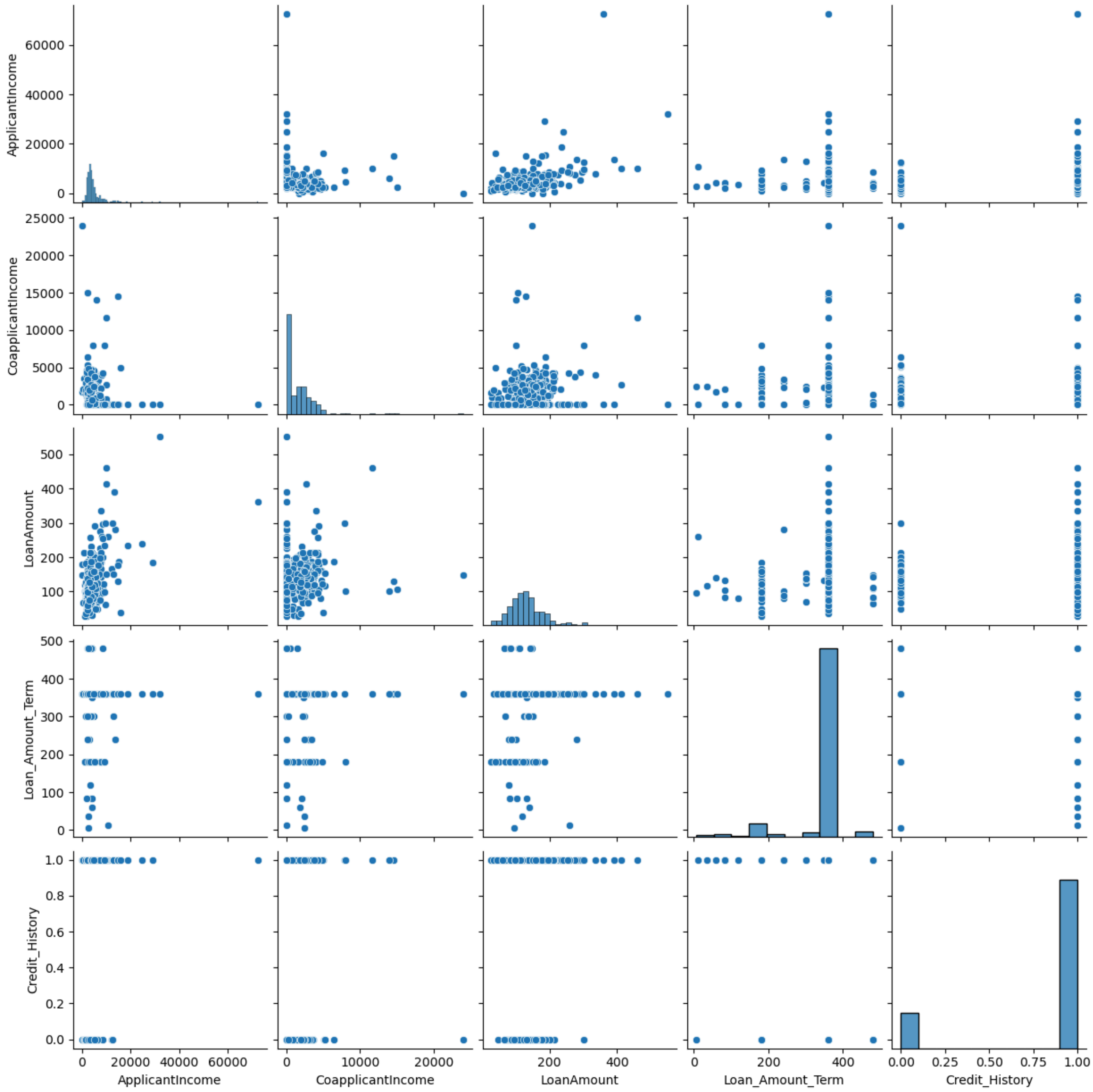
- As **ApplicantIncome** increases, the number of coapplicants with nonzero income becomes less frequent.
- Most high-income applicants seem to apply alone.

4. Possible Outliers:

- Some extreme values exist, where **CoapplicantIncome** is unusually high (above **20,000**) or **ApplicantIncome** is very high (above **70,000**). These could be potential **outliers**.

The second image represents a pair plot, which visualizes relationships between multiple numerical variables in the dataset.

1. **Insights for Loan Approvals:**



Key Observations:

2. Diagonal Elements (Histograms): Each variable's distribution is displayed, revealing skewness or normality.

    - ApplicantIncome and CoapplicantIncome appear heavily skewed with a few high-income outliers.
    - LoanAmount has a right-skewed distribution, meaning most loan amounts are on the lower end.
    - Credit History and Loan_Amount_Term are categorical-like variables with distinct groups.

3. Scatter Plots (Off-Diagonal Elements): These depict relationships between variables.

    - ApplicantIncome vs. LoanAmount: Shows some positive correlation,

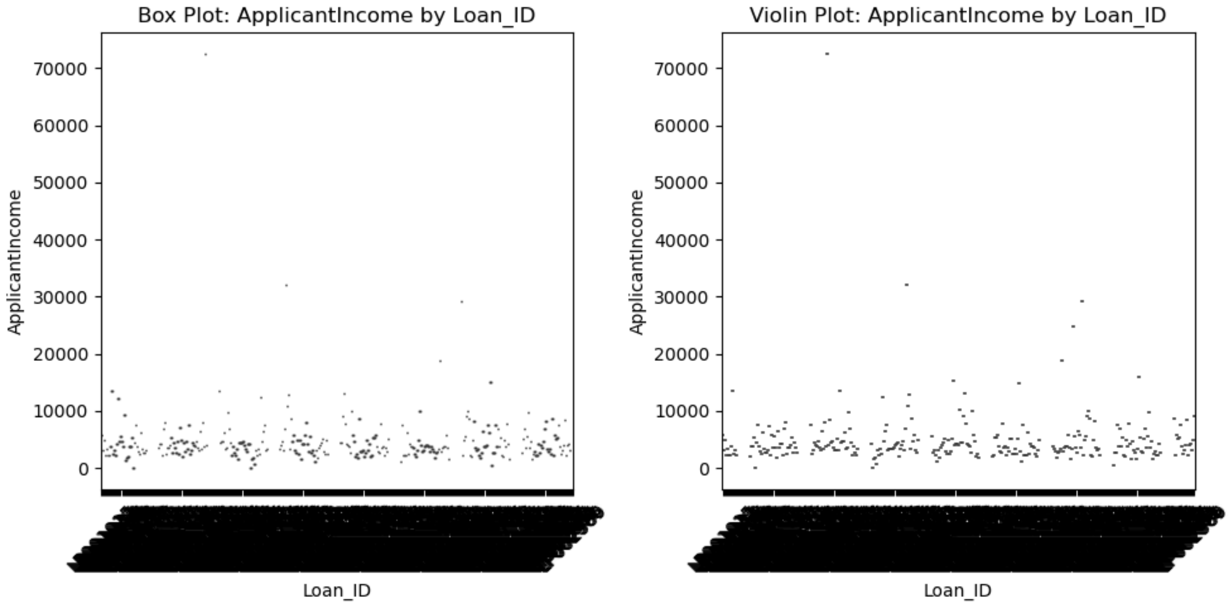
meaning higher income applicants tend to apply for larger loans.

- **CoapplicantIncome vs. LoanAmount: The pattern is less clear but follows a similar trend.**
- **Credit History vs. LoanAmount: Credit history appears categorical (values of 0 and 1), but individuals with a good credit history (1) seem to get higher loan amounts.**

4.

- **Loan amount decisions depend on applicant and co-applicant income.**
- **A good credit history (value = 1) is likely to increase loan approval chances.**
- **Some outliers exist in income and loan amounts, which could indicate special cases or errors.**

Scatter Plot (Applicant Income vs. Co Applicant-Income)



- This scatter plot visualizes the relationship between ApplicantIncome and CoapplicantIncome.
- The majority of data points are concentrated around low-income values, with some outliers showing high applicant income.

- **Many applicants have a CoapplicantIncome of zero, suggesting that they applied for a loan individually.**
- **The spread of data indicates that even applicants with high incomes may not have co-applicants.**

2nd Image: Pair Plot (Multiple Variables)

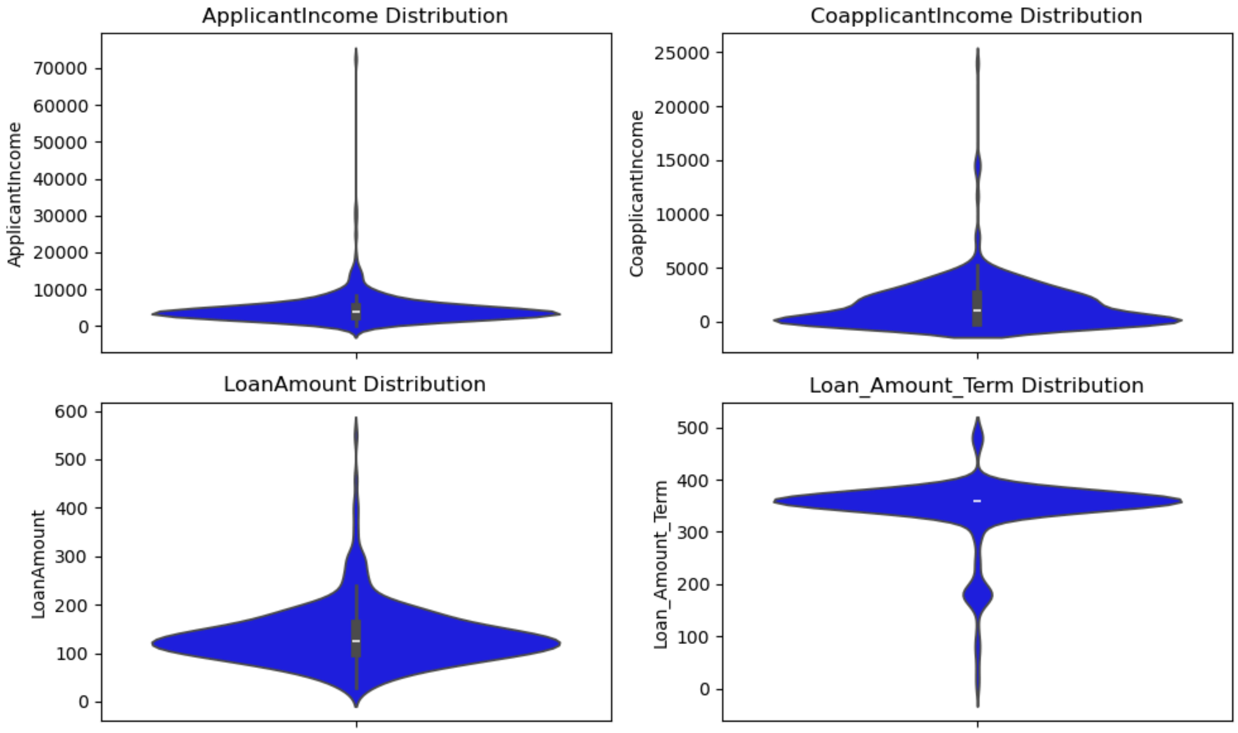
- **This pair plot shows scatter plots and histograms for multiple numerical variables.**
- **Diagonal elements (histograms) show the distribution of each variable.**
- **Off-diagonal elements (scatter plots) show relationships between variables.**
- **Some key observations:**
  - **ApplicantIncome and CoapplicantIncome show a similar distribution pattern.**
  - **LoanAmount is more scattered, showing variation in loan approvals.**
  - **Loan_Amount_Term and Credit_History have specific categorical values.**

3rd Image: Box Plot & Violin Plot (ApplicantIncome vs. Loan_ID)

- **Box Plot (Left): Shows the distribution of ApplicantIncome for each Loan_ID.**
  - **Outliers are visible, with some applicants having very high incomes.**
  - **The majority of data is concentrated at lower income levels.**

- **Violin Plot (Right): Similar to the box plot but also represents the density of data.**
  - **It shows that most applicants have lower incomes, with fewer higher-income applicants.**
  - **The wider sections indicate areas with more data concentration.**

Violin plots were used to investigate the relationship between categorical variables



First Image (Pair Plot)

- A pair plot is used to visualize relationships between multiple numerical variables.
- It shows scatter plots between different feature combinations and histograms on the diagonal.
- This helps identify correlations, trends, and outliers.

Second Image (Box and Violin Plots)

- The box plot (left) shows the distribution of applicant income across loan IDs, highlighting outliers.
- The violin plot (right) provides a more detailed view of income distribution, including density.

Third Image (Violin Distribution)

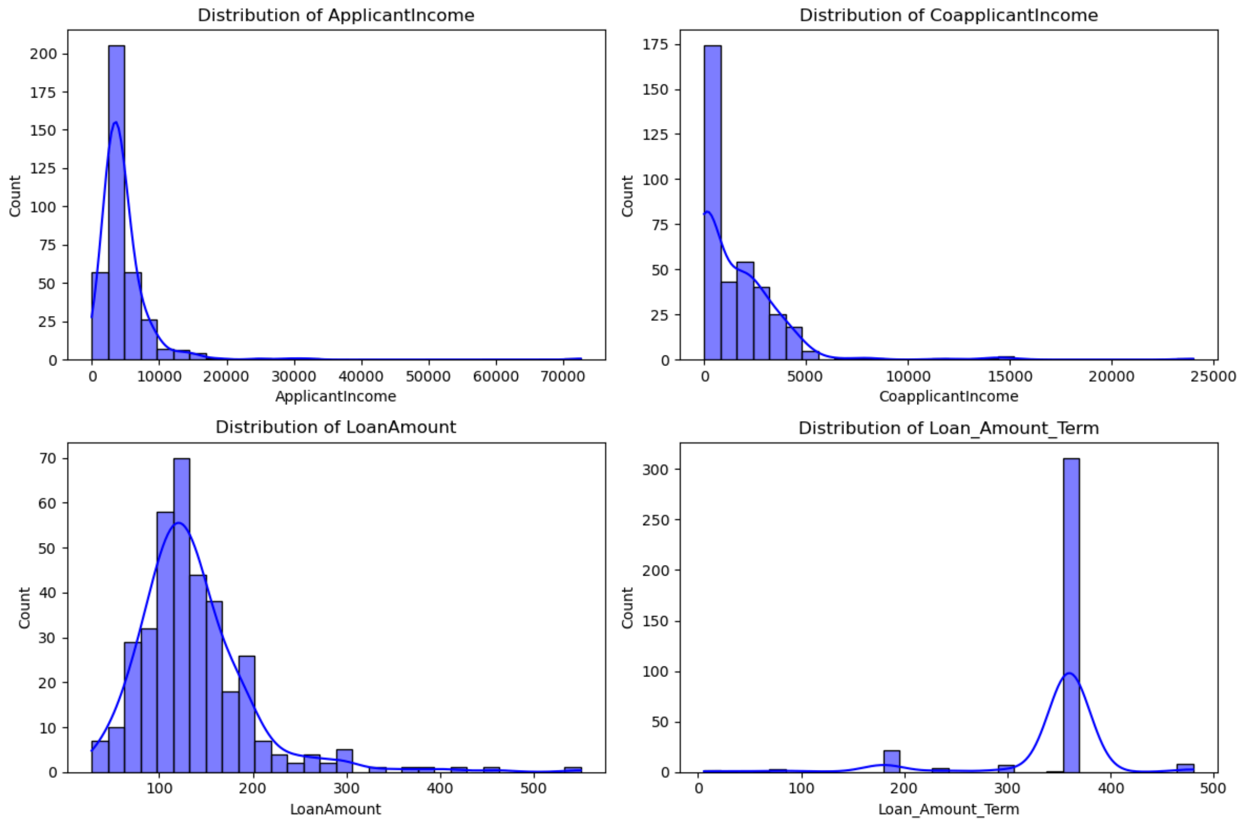
- Violin plots visualize the distribution of different numerical variables.
- It shows income, loan amounts, and loan terms, revealing skewness and outliers.

- **Peaks indicate the most common values in the dataset.**

Univariate Analysis

- **Histograms**: Frequency distribution of key numeric variables was plotted to understand how values are spread across different ranges. Histograms helped in identifying skewness, peaks, and potential outliers in data. The notebook included histograms with Kernel Density Estimation (KDE) plots for:

- `ApplicantIncome`
- `CoapplicantIncome`
- `LoanAmount`
- `Loan_Amount_Term` These histograms provided insights into whether income and loan amounts followed a normal distribution or were skewed.

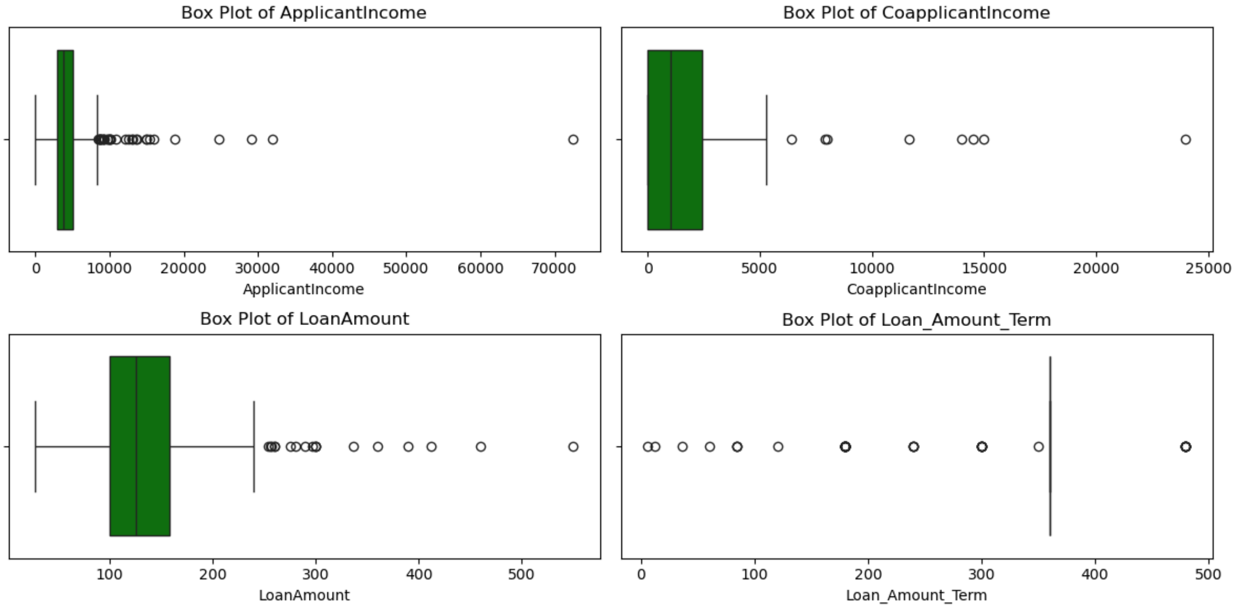


These histograms represent the frequency distribution of key numerical variables:

- `ApplicantIncome`
- `CoapplicantIncome`
- `LoanAmount`
- `Loan_Amount_Term`

Each histogram includes a Kernel Density Estimation (KDE) curve to show the probability distribution.

The distributions seem right-skewed, meaning most values are on the lower end, but some extreme values create a long tail.



Conclusion:

The exploratory data analysis (EDA) conducted on the home loan approval dataset provided valuable insights into the factors influencing loan approvals. The analysis focused on understanding the data structure, handling missing values, and visualizing key relationships among different variables.

Key Findings

1. **Income Distribution:**
    - Applicant incomes varied significantly, with a majority earning within a lower range while a few outliers had exceptionally high incomes.
    - Many applicants had no co-applicant income, indicating they applied for loans individually.
2. **Loan Amount Trends:**
    - Loan amounts also showed a wide distribution, with higher loan amounts corresponding to higher applicant incomes.

- Outliers were identified, with some applicants securing significantly larger loans than the average applicant.

3. **Loan Term Analysis:**

   - Most loans were issued for a **standard duration of 360 months (30 years)**, with a smaller portion having shorter terms.

4. **Credit History and Approval Rates:**

   - A strong correlation was observed between **credit history** and loan approval, where applicants with a **good credit history** had a much higher approval rate.
   - Applicants with a poor or missing credit history were more likely to be rejected.

5. **Property Type and Loan Approval:**

   - Loan approval rates varied based on property location, with **urban and semi-urban applicants receiving more approvals** than rural applicants.

6. **Gender and Marital Status Trends:**

    - Married applicants had a **higher likelihood of loan approval**, possibly due to dual-income support.
    - Gender did not show a strong influence on loan approvals, though most applicants were male.

Implications and Future Directions

- **Loan Approval Criteria Optimization:** The analysis suggests that **credit history, applicant income, and loan amount** play a crucial role in approval decisions. Financial institutions could optimize their loan approval criteria to make more data-driven decisions.

- **Bias Detection:** Identifying potential biases in the approval process, such as location-based or marital status-based disparities, could help ensure fairer lending practices.

- **Further Analysis:**

    - **Predictive Modeling:** While this project focused on EDA, a future extension could involve applying machine learning techniques to predict loan approval outcomes based on historical data.

- **Feature Engineering:** Creating additional derived variables (e.g., debt-to-income ratio, co-applicant dependency factor) may improve decision-making models.

Final Thoughts

Through thorough data exploration and visualization, this project successfully uncovered critical trends in loan approvals. The insights gained can be leveraged by financial institutions to refine their loan policies, improve approval efficiency, and ensure fairness in the lending process.

THANK YOU!

TO ACCESS THE FILE , CLICK THE LINK BELOW.

https://drive.google.com/file/d/1_2jpNERumgMHxMuVB2yHHIyaJOHBEd0M/view?usp=sharing