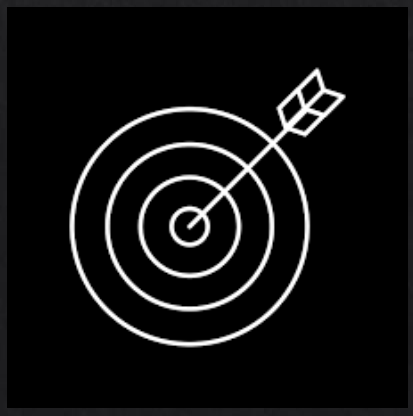
LOAN SANCTION DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

BY RITESH SINGH

PROJECT OBJECTIVE

- **Objective:**
  To analyze the **Loan Sanction Dataset** to understand applicant behavior, identify key factors influencing loan approval, and provide data-driven recommendations for improving decision-making.
- **Key Goals:**
  - Analyze **applicant demographics** and **loan characteristics**.
  - Identify relationships between **income and loan amount**.
  - Assess the impact of **credit history** on loan sanction likelihood.
  - Generate actionable insights for **data-driven lending policies**.



PROBLEM STATEMENT

- **Problem Context:**
  Financial institutions face challenges in determining loan eligibility due to multiple applicant-related factors such as income, employment status, dependents, and credit history.
- **Problem Statement:**
  The loan sanction process is influenced by various demographic and financial parameters.
  This project aims to **analyze how these factors collectively impact loan approval probability** and help identify **patterns in applicant eligibility**.
- **Key Challenge:**
  To uncover meaningful insights that can assist banks in making **accurate, fair, and data-driven loan decisions.**

DATASET OVERVIEW

- **Dataset Name:** loan_sanction_test.csv
  Total Records: 367
  Total Features: 12
- **Key Attributes:**
- **Applicant Information:**
  - Gender, Married, Dependents, Education, Self_Employed
- **Financial Information:**
  - ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term
- **Credit & Property Details:**
  - Credit_History, Property_Area
- **Note:**
  The dataset includes both categorical (e.g., Gender, Education, Property_Area) and numerical (e.g., LoanAmount, ApplicantIncome) variables — allowing for diverse analytical insights.

TOOLS & LIBRARIES USED

- **Programming Language:**
  - **Python → For data analysis, preprocessing, and visualization.**

Library	Purpose
pandas	Data loading, cleaning, and manipulation
numpy	Numerical operations and mathematical computations
matplotlib	Basic visualizations and plotting
seaborn	Advanced and aesthetic data visualizations

- Development Environment:
  - Jupyter Notebook / Google Colab → For code execution, documentation, and interactive analysis

DATA INSPECTION

- **Steps Performed:**

```
df = pd.read_csv("/content/loan_sanction_test.csv")
```

- **Loaded dataset** using pandas for initial exploration.
- **Checked data types** and structure of all columns.
- **Identified missing values** across categorical and numerical fields.
- **Classified variables** into **numerical** and **categorical** groups for analysis.

- **Observations:**

- Dataset contains a **mix of numeric and categorical attributes.**
- Missing values observed in:
  - **Gender**
  - **Self_Employed**
  - **LoanAmount**
  - **Credit_History**

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 367 entries, 0 to 366
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            367 non-null    object 
 1   Gender             356 non-null    object 
 2   Married            367 non-null    object 
 3   Dependents         357 non-null    object 
 4   Education          367 non-null    object 
 5   Self_Employed      344 non-null    object 
 6   ApplicantIncome    367 non-null    int64  
 7   CoapplicantIncome  367 non-null    int64  
 8   LoanAmount         362 non-null    float64
 9   Loan_Amount_Term   361 non-null    float64
 10  Credit_History     338 non-null    float64
 11  Property_Area      367 non-null    object 
dtypes: float64(3), int64(2), object(7)
memory usage: 34.5+ KB
```

MISSING VALUE TREATMENT

Column	Type	Method
Gender, Self_Employed	Categorical	Filled with Mode
LoanAmount, Loan_Amount_Term	Numeric	Filled with Median
Credit_History	Numeric	Filled with Median
Dependents	Categorical	Replaced ‘3+’ with 3

```
Missing Value Summary:
                   Missing Values  Missing %
Credit_History                 29   7.901907
Self_Employed                  23   6.267030
Gender                         11   2.997275
Dependents                     10   2.724796
Loan_Amount_Term                6   1.634877
LoanAmount                      5   1.362398
Married                         0   0.000000
Loan_ID                         0   0.000000
CoapplicantIncome               0   0.000000
ApplicantIncome                 0   0.000000
Education                       0   0.000000
Property_Area                   0   0.000000
```

DATA CLEANING SUMMARY

- **Cleaning Actions Performed:**
  - Replaced **'3+'** in *Dependents* column with numeric value **3**.
  - Converted **Dependents** column to **integer type** for uniformity.
  - Created a new derived column:
    ➤ **Total_Income = ApplicantIncome + CoapplicantIncome**
  - Verified and corrected **data formats** to ensure consistency.
  - Checked for **outliers and invalid ranges** in numerical columns.
- **Result:**
  Dataset became **clean, structured, and ready** for feature engineering and analysis with **no missing values** remaining.

```
#Cleaning categorical columns
#For categorical columns like Gender, Married, Dependents, etc., we fill missing values using the most frequent value (mode).
categorical_cols = ['Gender', 'Married', 'Dependents', 'Self_Employed']

for col in categorical_cols:
    df_clean[col].fillna(df_clean[col].mode()[0], inplace=True)
```

```
#Fixing 'Dependents' column
#The column sometimes has '3+'. Let's replace it with numeric 3 and convert it to an integer.
cell (Ctrl+Enter)
has not been executed in this session
df_clean['Dependents'].replace('3+', '3')
df_clean['Dependents'].astype(int)
```

```
#Handle missing values in numerical columns
#We will fill missing numeric values using the median, which is more robust against outliers
numeric_cols = ['LoanAmount', 'Loan_Amount_Term', 'Credit_History']

for col in numeric_cols:
    df_clean[col].fillna(df_clean[col].median(), inplace=True)
```

FEATURE ENGINEERING

- **New Features Created:**
- **Total_Income →**
  Combined applicant and coapplicant incomes to represent overall earning capacity.
  Total_Income = ApplicantIncome + CoapplicantIncome
- **LoanAmount_log →**
  Applied logarithmic transformation to reduce right skewness in loan amounts.
  LoanAmount_log = log(LoanAmount + 1)
- **EMI (Equated Monthly Installment) →**
  Estimated monthly repayment burden of the applicant.
  EMI = LoanAmount / Loan_Amount_Term
- **Balance_Income →**
  Represents income left after paying EMI — a measure of repayment ability.
  Balance_Income = Total_Income – (EMI × 1000)
- **Purpose:**
  To create **more meaningful variables** that capture financial strength, repayment potential, and income distribution for better analysis.

```
#Creating Total_Income feature
df_clean['Total_Income'] = df_clean['ApplicantIncome'] + df_clean['CoapplicantIncome']
```

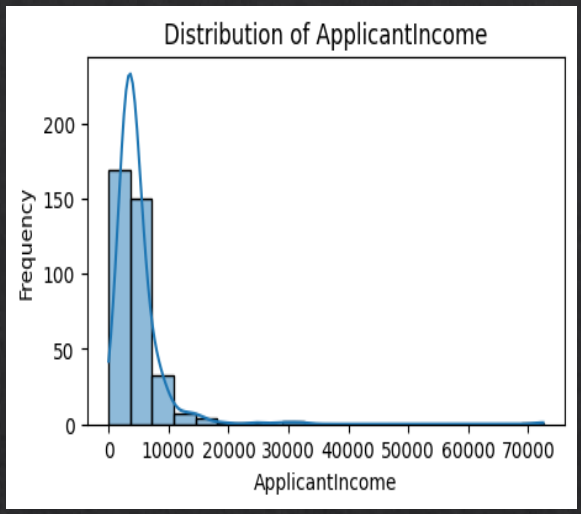
```
#Create useful new features
df_fe['EMI'] = df_fe['LoanAmount'] / df_fe['Loan_Amount_Term']   # EMI approximation
df_fe['Balance_Income'] = df_fe['Total_Income'] - (df_fe['EMI'] * 1000)  # residual income
```

ENCODING CATEGORICAL VARIABLES

- **Encoding Approach:**
- **Label Encoding (Binary Features):**
  - Applied to features with **two unique categories**
  - Examples: **Gender, Married, Education, Self_Employed**
  - Converts values like *Male/Female → 0/1*
- **One-Hot Encoding (Multi-Category Features):**
  - Applied to columns with **more than two categories**
  - Example: **Property_Area (Urban, Semiurban, Rural)**
  - Creates separate dummy variables for each category
- **Dropped Non-informative Column:**
  - **Loan_ID** removed as it does not contribute to analysis or prediction

- **Result:**
  All categorical features transformed into **numerical format**, making the dataset **ready for analysis and visualization.**

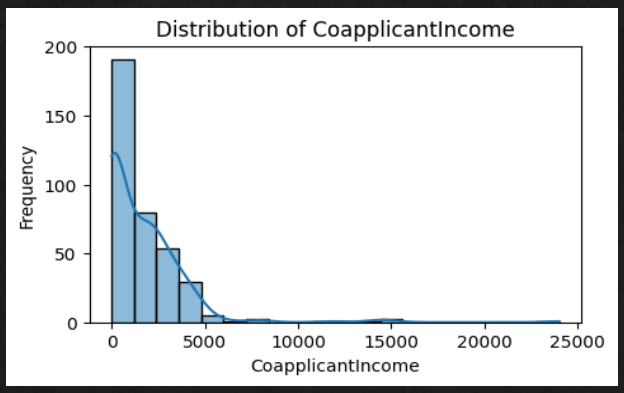
DATA DISTRIBUTION (NUMERICAL FEATURES)

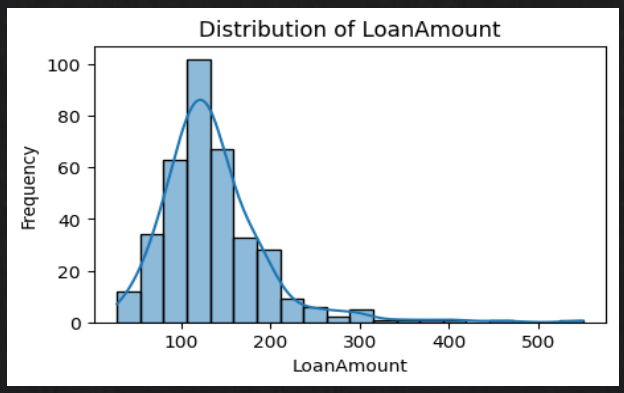


- **Numerical Columns Analyzed:**
  - **ApplicantIncome**
  - **CoapplicantIncome**
  - **LoanAmount**
  - **LoanAmount_log (transformed)**

- **Observations:**
  - **ApplicantIncome** → *Right-skewed*, with most applicants in lower–middle income range.
  - **CoapplicantIncome** → Concentrated in **low to medium range**; fewer high-income coapplicants.
  - **LoanAmount** → *Right-skewed* distribution, showing small loan amounts are more common.
  - **LoanAmount_log** → Normalized after transformation, suitable for analysis and modeling.

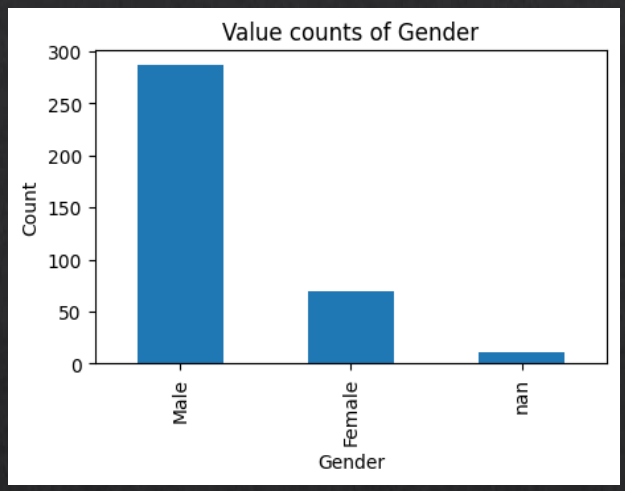
- **Insight:**
  Most applicants apply for **smaller loan amounts** relative to income, reflecting cautious borrowing behavior.



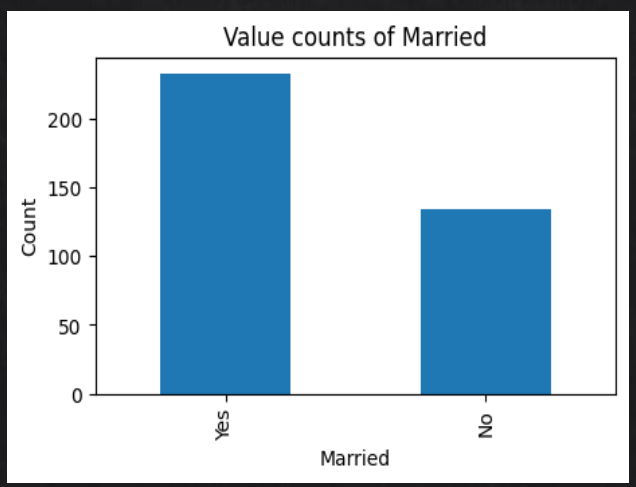


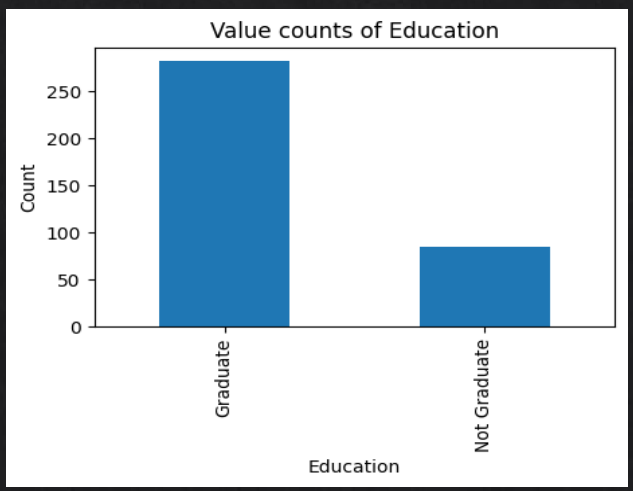
CATEGORICAL DISTRIBUTION

- **Categorical Columns Analyzed:**
  - **Gender**
  - **Married**
  - **Education**
  - **Self_Employed**
  - **Property_Area**

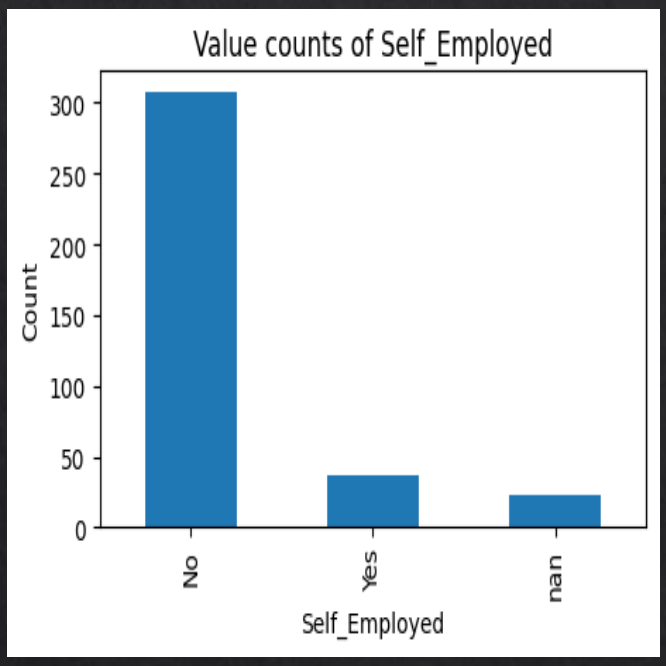


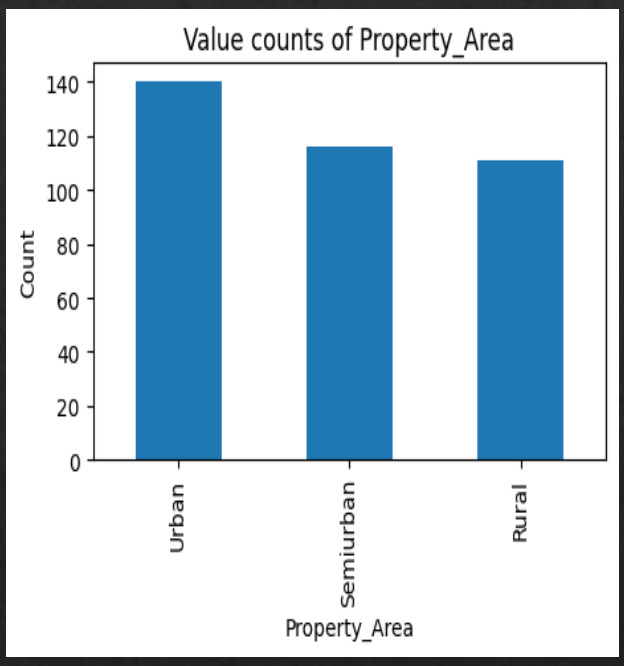
- **Key Observations:**
  - **Gender:** Majority of applicants are **Male**.
  - **Marital Status:** Most applicants are **Married**, indicating family-based loan applications.
  - **Education:** Predominantly **Graduates**, showing awareness of formal credit systems.





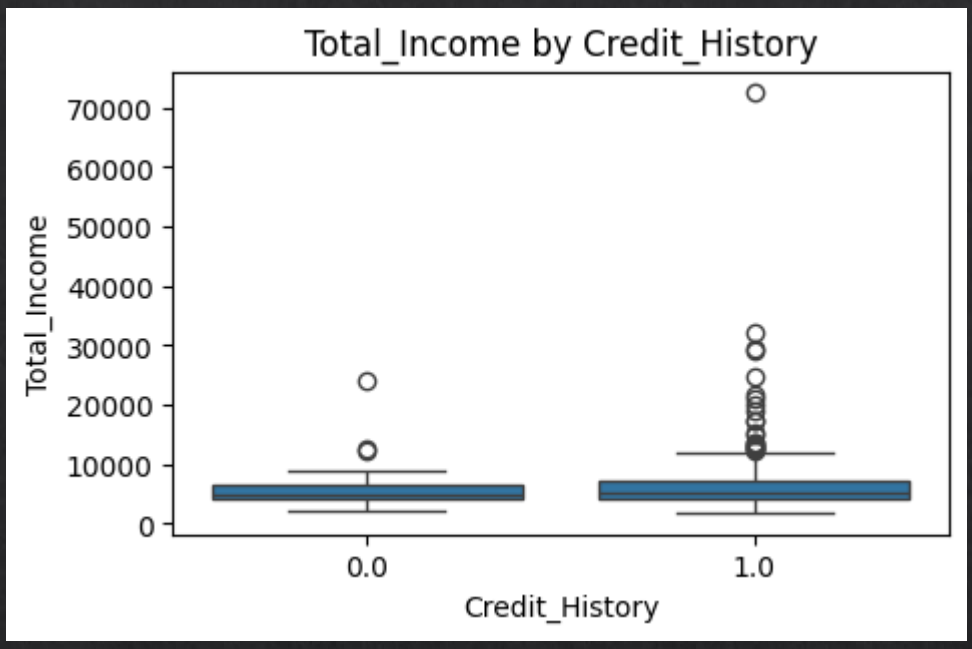
- **Self_Employed:** Smaller proportion compared to salaried applicants.
- **Property_Area:** Most applicants belong to **Urban** and **Semi-Urban** areas.





❖**Insight:**
Urban and educated segments dominate the loan applicant pool, reflecting better financial literacy and accessibility to banking services.

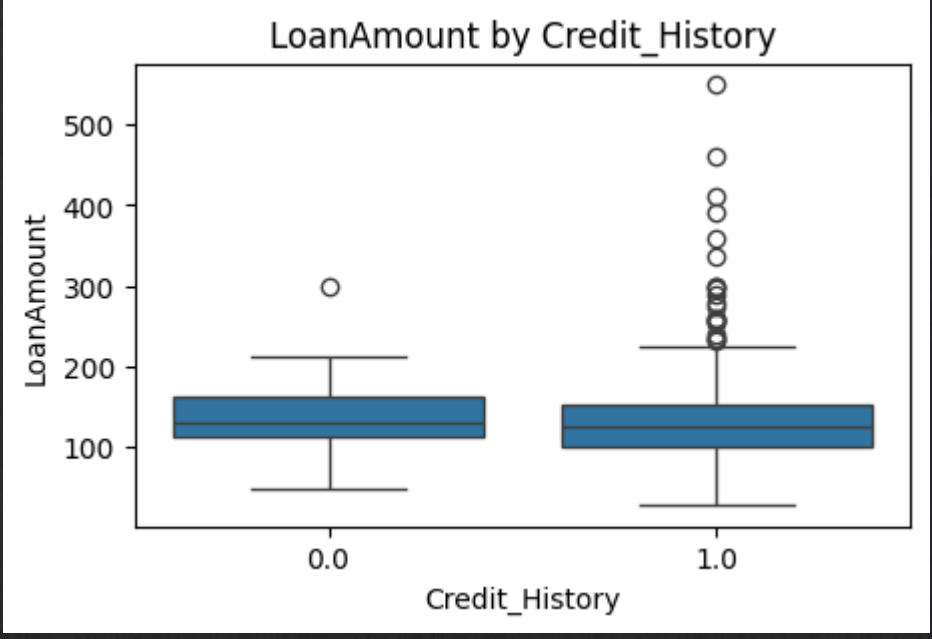
RELATIONSHIP — INCOME VS LOAN AMOUNT



❖ **Key Insights:**

- **Higher total income** generally corresponds to a **higher sanctioned loan amount**.
- Most applicants fall within the **low-to-moderate income and loan range.**
- A few **high-income applicants** have opted for **smaller loans**, reflecting **conservative borrowing** behavior.
- The relationship indicates that **income capacity strongly influences loan amount eligibility.**

CREDIT HISTORY ANALYSIS



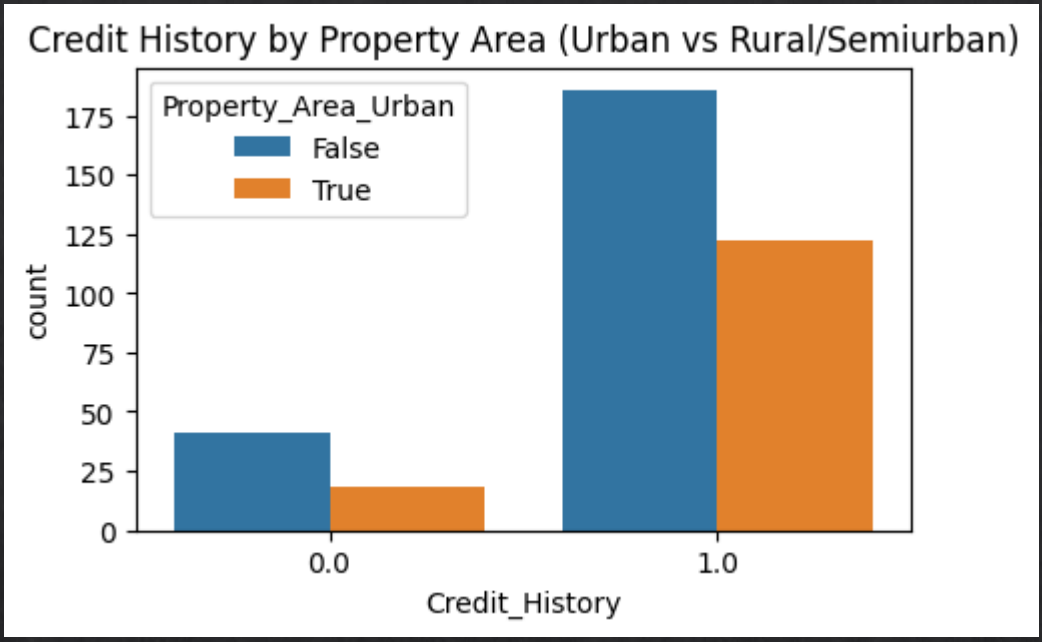
❖ **Key Observations:**

- Majority of applicants have **Credit_History = 1**, indicating a positive repayment record.
- Applicants with **good credit history (1)** are **significantly more likely** to have their loans sanctioned.
- A clear **positive correlation** exists between **Credit_History** and **loan eligibility**.
- Applicants without a credit history face a higher likelihood of rejection or smaller sanctioned amounts.

❖ **Insight:**

Credit history is a **critical factor** in the loan approval process, serving as a key indicator of applicant reliability and repayment behavior.

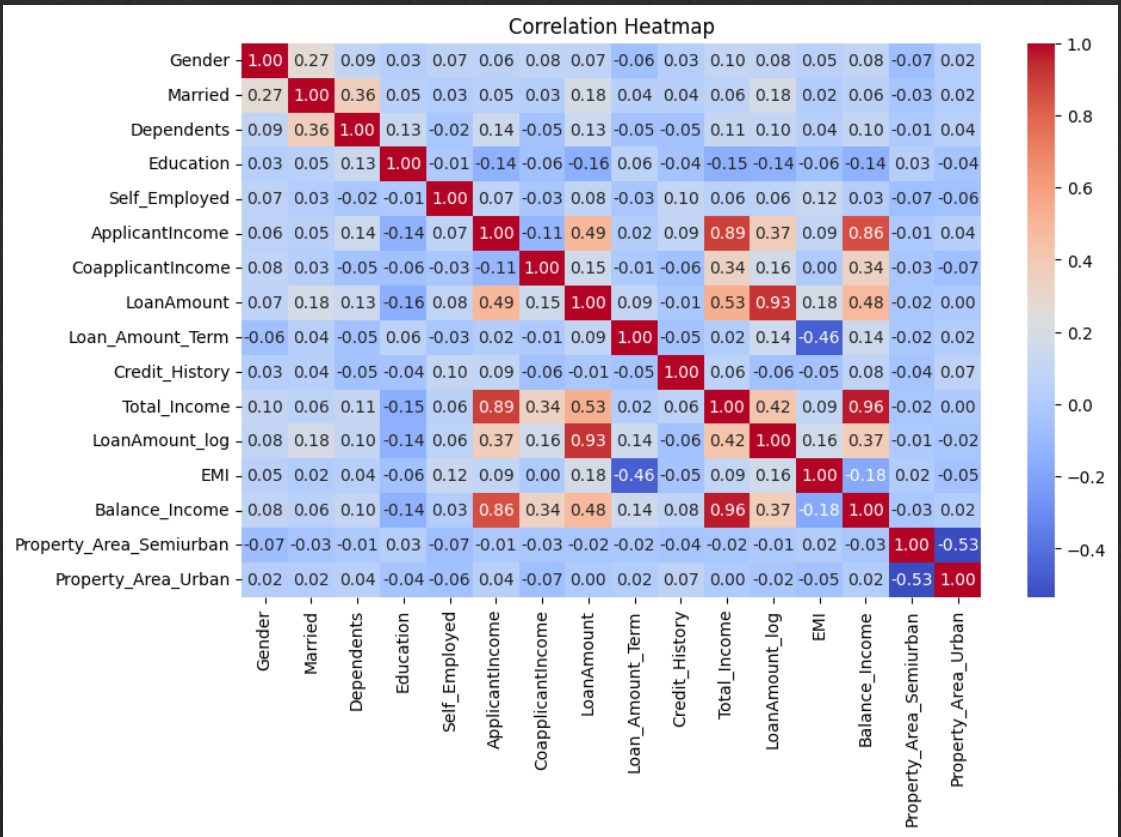
PROPERTY AREA & CREDIT HISTORY



- **Key Insights:**
- **Urban and Semiurban** areas show **higher loan approval rates**, with a greater proportion of applicants having **Credit_History = 1**.
- **Rural** applicants have **lower approval trends**, often linked to limited credit exposure or insufficient documentation.
- Credit awareness and accessibility appear **stronger in urban regions**, reflecting financial inclusion differences.

- **Observation:**
  Property location indirectly influences loan eligibility through **credit awareness and financial access** — highlighting the need for **targeted financial literacy programs** in rural areas.

CORRELATION HEATMAP



- **Key Observations:**
- **Total_Income** and **LoanAmount** show a **strong positive correlation**, indicating higher-income applicants tend to request higher loan amounts.
- **ApplicantIncome** is highly related to **Total_Income**, as expected.
- **Credit_History** demonstrates a **strong correlation** with other eligibility indicators, reaffirming its role as a key approval factor.
- Weak or near-zero correlations among unrelated variables suggest minimal redundancy.

- **Insight:**
  Correlation analysis helps identify the **most influential variables** in determining loan eligibility and ensures **efficient feature selection** for further analysis.

KEY INSIGHTS SUMMARY

- **Major Analytical Findings:**
  - **Credit History** is the **strongest determinant** of loan approval — applicants with good credit records are far more likely to be sanctioned.
  - **Applicant and Total Income** serve as **crucial indicators** of repayment capacity and financial stability.
  - **Urban Applicants** show **higher approval trends**, suggesting better financial inclusion and documentation support.
  - **Loan Amount** tends to **increase with income**, but overall distribution remains **right-skewed** — most applicants prefer smaller loans.
  - **Dependents and Education** have a **moderate influence**, reflecting partial impact on eligibility compared to income and credit history.
- **Overall Insight:**
  Loan approval decisions are primarily driven by **creditworthiness** and **income strength**, while demographic factors play a secondary role.

RECOMMENDATIONS

- **Strategic Recommendations:**
  - **Verify and prioritize credit history** during the loan approval process to minimize default risks.
  - **Adopt income-based risk scoring models** that evaluate applicant repayment capacity more accurately.
  - **Simplify loan documentation** and verification processes, especially for **rural and first-time borrowers**.
  - **Offer interest rate incentives** to applicants with **consistent repayment history** to promote responsible borrowing.
  - **Integrate predictive analytics models** to automate eligibility scoring and support faster, data-driven loan decisions.

- **Summary:**
  Implementing these recommendations can help financial institutions **reduce risk**, **enhance credit inclusion**, and **improve decision-making efficiency**.

CHALLENGES FACED

- **Key Challenges Encountered During Analysis:**
- **Handling Missing Data:**
  - Required careful imputation using median/mode to maintain data consistency without bias.
- **Dealing with Skewed Distributions:**
  - LoanAmount and Income variables showed strong right skewness, needing log transformations.
- **Encoding Categorical Variables:**
  - Choosing between **Label Encoding** and **One-Hot Encoding** based on feature type and cardinality.
- **Selecting Appropriate Transformations:**
  - Balancing data normalization without losing interpretability for business stakeholders.
- **Interpreting Correlations Meaningfully:**
  - Differentiating between statistical correlation and real-world causal relationships.

- **Learning Outcome:**
  Overcoming these challenges improved data understanding, ensuring **accuracy, interpretability, and analytical depth** in the final insights.

CONCLUSION

- **Overall Summary:**
  This exploratory data analysis (EDA) provided a comprehensive understanding of the **Loan Sanction Dataset** and the underlying factors influencing loan approval decisions.
  Through data cleaning, feature engineering, and visualization, the project successfully identified **patterns, relationships, and key indicators** that drive loan eligibility outcomes.
- **Key Takeaways:**
- **Credit History** is the **most critical determinant** of loan approval — applicants with a good repayment record have a significantly higher chance of loan sanction.
- **Applicant and Total Income** strongly influence loan eligibility, reflecting repayment potential and financial stability.
- **Property Area** plays an indirect role, with **Urban and Semiurban** applicants showing better access to credit facilities compared to Rural applicants.
- **Education and Dependents** have moderate impact, suggesting that demographic factors contribute to, but do not dominate, loan decisions.
- **Feature Engineering** (like Total_Income, EMI, and Balance_Income) helped in uncovering deeper insights about applicant affordability and financial behavior.

GOOGLE COLABORATORY LINK

https://colab.research.google.com/drive/1SNjcgRQ8b7UUTBS_SsusT24mitS0_lLQ?authuser=1#scrollTo=H1kdBJxoM4FK