MZUZU UNIVERSITY

FACULTY OF INFORMATION SCIENCE AND COMMUNICATIONS

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

TO	**:**	**MR. RUEBEN MOYO**
FROM	**:**	**JONES THUKUTA**
REG NUMBER	**:**	**BSDS2922**
LEVEL	**:**	**TWO**
COURSE TITLE	**:**	**DATA WRANGLING AND EXPLORATOTY DATA ANALYSIS**
PROGRAMME	**:**	**DATA SCIENCE**
COURSE CODE	**:**	**BICT2306**
SUBJECT	**:**	**PROJECT REPORT**
DUE DATE	**:**	**10TH JUNE, 2024.**

LOAN APPROVAL PREDICTION DATASET DATA WRANGLING AND EXPLORATORY DATA ANALYSIS

Lecturer Mr. Rueben Moyo,

By Jones Thukuta.

INTRODUCTION

Loan is one of the major requirements in the business realm of the modern world. With the increasing loan demands in the business industry, it is very cumbersome, time consuming and prone to human biases to process loan applications manually. The objective of this analysis is to perform data wrangling and exploratory data analysis on a loan approval prediction dataset. This dataset can be used to develop a reliable model that can assess the creditworthiness of loan applicants and predict the likelihood of loan approval based on the features provided by the applicants hence eliminating the need for manual loan processing.

1. DATASET DESCRIPTION

1.1 Meta Data

Dataset Title: Loan Prediction dataset

Source: Kaggle acquired on 18th April, 2024.

Dataset Description:

This dataset contains a collection of individuals that applied for a loan and it has their attributes that determined their loan status. The dataset can be used to determine whether a loan should be approved or not based on attributes like number of dependents, education, applicant’s income, credit history, property area and other attributes.

Metadata Characteristics:

Structure:

- The dataset is stored in comma-separated values (CSV) file format.

- It contains a header row with the column names.
- The data is organized in rows, each row representing a single applicant.

Granularity:

- The amounts ApplicantIncome, CoapplicantIncome and LoanAmount are per year which is very coarse.
- The Loan_Amount_Term is expressed in days which is very fine.

Accuracy:

- The dataset was acquired from Kaggle, which is one of the trusted sources for public datasets, and the fields are consistent.

Temporality:

- The dataset was acquired on 18th April, 2024.
- The dataset represents a snapshot of loan applications at a certain point in time and can become stale at any point in time.

Scope:

- The dataset covers 615 loan applications.

Dataset Columns:

- Loan_ID (object): Unique identifier for each loan applicant.
- Gender (object): The gender for each applicant.
- Married (object): Marital status of each loan applicant.
- Dependents (object): Number of dependents the loan applicant had.
- Education (object): The education level of the loan applicant.
- Self_Employed (object): Type of income source of the applicant.
- ApplicantIncome (integer): Amount of money the applicant was making annually.
- CoapplicantIncome (float): Amount of money the co-applicant of the applicant was making annually.
- LoanAmount (float): Amount of money the applicant was willing to lend.

- Loan_Amount_Term(float): The duration that the applicant was given to pay back the loan.
- Credit_History(float): The creditworthiness of the applicant based on previous loan taken.
- Property_Area(object): The setting that the applicant was residing.
- Loan_Status(object): Target variable that states whether the application was approved or not.

Data File:

- File: Loan Prediction dataset.csv
- Size: 38 KB
- Number of Rows: 615
- Number of Columns: 13

2. DATA CLEANING

2.1 Dealing With Missing Values

Checking the number of missing values:

- This was used to check if there are any missing values in the fields and it turned out that the fields had missing values.

Filling in missing values of object type:

- To deal with the missing values I started by filling in the object type fields using attribute mode imputation.

Filling in missing values of numeric types:

- Then I proceeded by filling in the numeric type fields using attribute mean imputation.

2.2 Correcting data types

- It was noted that ApplicantIncome was of int type and it was changed to float since money is expressed in float type

2.3 Checking for duplicates

- The other data cleaning aspect that was supposed to be addressed was presence of duplicates and it turned out that the dataset had no duplicates after checking for duplicates.

2.4 Addressing inconsistencies

Checking for outliers:

- There were 264 outliers that were found in the numeric type fields in total and since there are many outliers it was observed that dropping them would significantly reduce the dataset hence affecting the analysis.
- Therefore, the way forward was to reduce their impact during transformation rather than dropping them.

3. DATA TRANSFORMATION

3.1 Creating new variables

Feature combination:

- It involved combing the values of CoapplicantIncome and ApplicantIncome in oder to have the Total_Applicant_Income. Which in turn reduced the rendency of features and also helped in data reduction

3.2 Log transformation

Log tranform on Total_Applicant_Income:

- This involved taking the natural log of Total_Applicant_Income values to reduce them making them suitable for analysis.

Log Transform of LoanAmount:

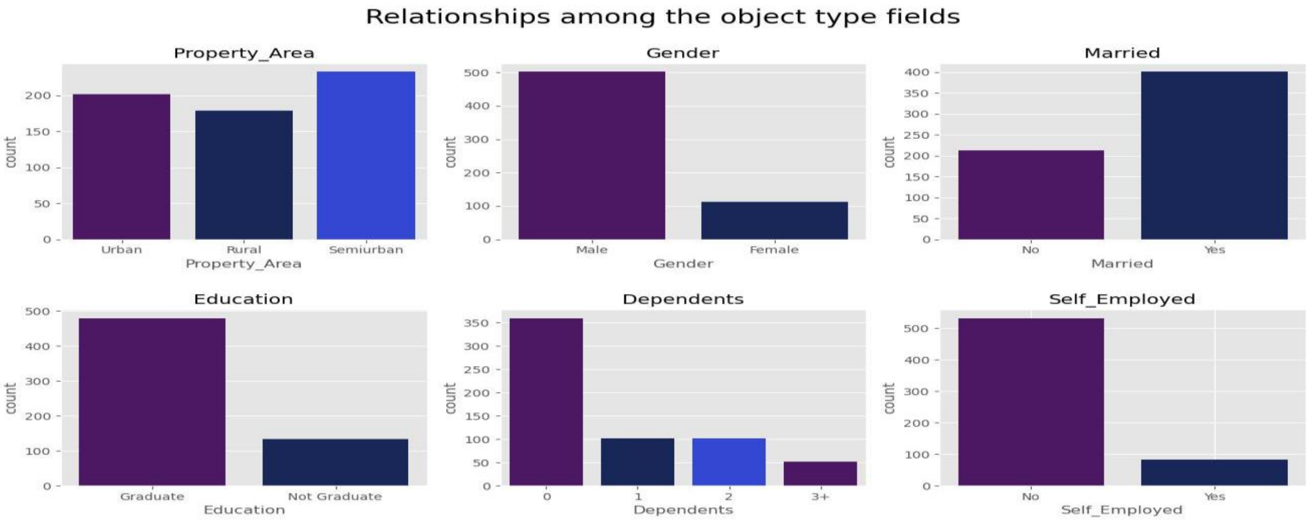
- This also involved taking the natural log if LoanAmount in order to also make it suitable for analysis.

Log Tranform for Loan_Amount_Term:

- Lastly, I also log transformed the Loan_Amount_Term field.

4. EPLORATORY DATA ANALYSIS

4.1 Using count plots to observe the relationships in the dataset



OBSERVATIONS AND INSIGHTS

Property Area:

- The count of properties in Urban and Semiurban areas is higher compared to Rural areas.
- This tells us that Urban and rural areas could have different loan approval patterns

Gender:

- There are significantly more males represented compared to females in the dataset.
- This may indicate that males are the predominant gender in loan applications, possibly due to socio-economic or cultural factors.
- This may affect the model's ability to generalize accross genders

Married:

- The count of married individuals is higher than unmarried individuals.
- This may indicate that married individuals often apply for loans more frequently than unmarried individuals due to financial stability and combined income, making it easier to apply and get approved.

Education:

- The count of graduates is much higher than the count of non-graduates.
- This may due to, graduates' better job prospects and higher incomes, making them more likely to qualify for loans

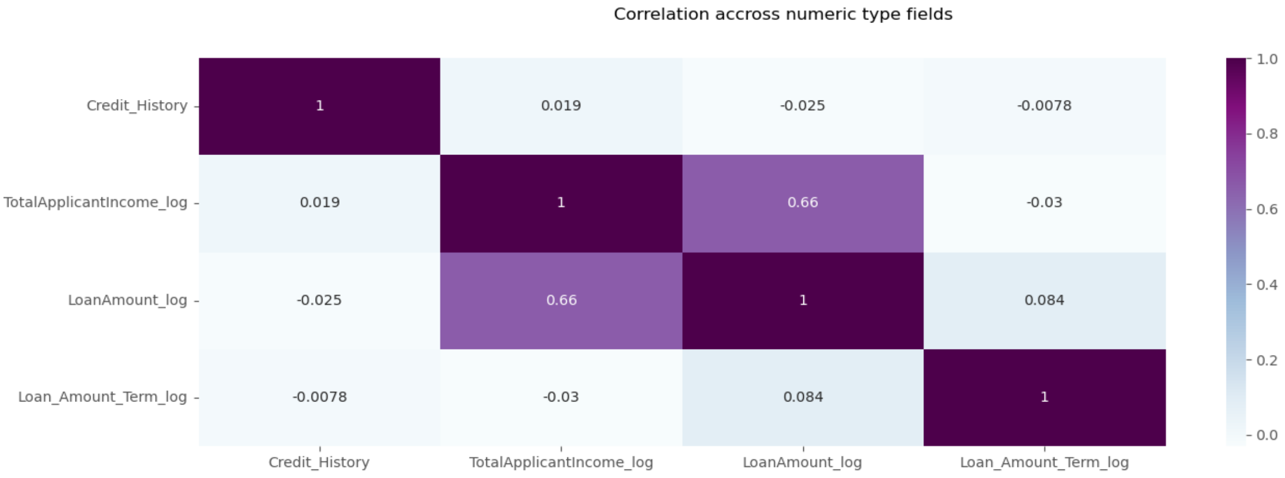
Dependents:

- The majority of individuals have 0 dependents, with a decreasing count as the number of dependents increases.
- This may due to financial burdens, while increasing dependents inversely correlate with loan applications.

Self_Employed:

- The count of individuals who are not self-employed is significantly higher than those who are self-employed.
- This may be so because, self-employed individuals often face more challenges in proving stable income for loan eligibility compared to regular salaried jobs.

4.2 Using Correlation heatmap to observe relationships accross numeric type fileds

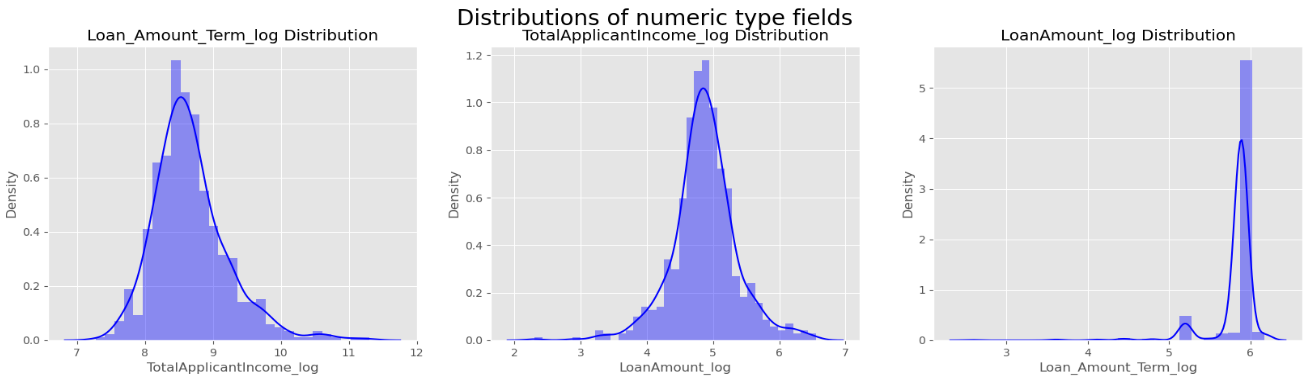


OBSERVATIONS AND INSIGHTS

- Credit_History and TotalApplicantIncome_log: Correlation coefficient is 0.019. indicating that an applicant’s credit history is independent to their income level.

- Credit_History and LoanAmount_log: Correlation coefficient is -0.025, implying that the credit history has little to no impact on the loan amount.
- Credit_History and Loan_Amount_Term_log: Correlation coefficient is -0.0078, indicating that the credit history does not significantly influence the term of the loan.
- TotalApplicantIncome_log and LoanAmount_log: Correlation coefficient is 0.66 indicating higher applicant incomes are associated with higher loan amounts.
- TotalApplicantIncome_log and Loan_Amount_Term_log: Correlation coefficient is -0.03 suggesting that income level has minimal influence on the length of the loan term.
- LoanAmount_log and Loan_Amount_Term_log: Correlation coefficient is 0.084, indicating that larger loan amounts tend to have slightly longer terms, but the relationship is not strong.

4.3 Using Distribution Plots to observe the distibution of numeric type fields



OBSERVATIONS AND INSIGHTS

Loan_Amount_Term_log Distribution

- The distribution of Loan Amount Term (log-transformed) is roughly normal with a peak around 9, having a slight positive skew.
- This indicates that there are more loans with shorter terms than longer ones.

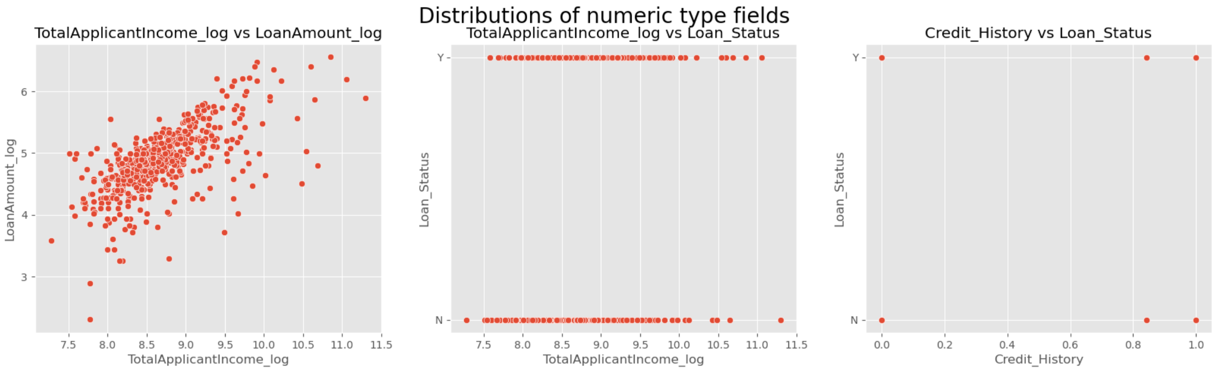
TotalApplicantIncome_log Distribution

- The distribution of Total Applicant Income (log-transformed) is slightly skewed to the right with a peak around 5.
- This suggests that there are fewer applicants with exceptionally high incomes.

LoanAmount_log Distribution

- The distribution of Loan Amount (log-transformed) is sharply peaked around 6 and is right-skewed
- This suggests that while most loans are of a similar amount, there are a few loans that are significantly larger.

4.4 Using Scatter Plots to observe the relationships between numeric variables



OBSERVATIONS AND INSIGHTS

TotalApplicantIncome_log vs LoanAmount_log

- There is a positive correlation between Total Applicant Income (log-transformed) and Loan Amount (log-transformed).
- This suggests that applicants with higher incomes tend to take larger loan amounts.

TotalApplicantIncome_log vs Loan_Status

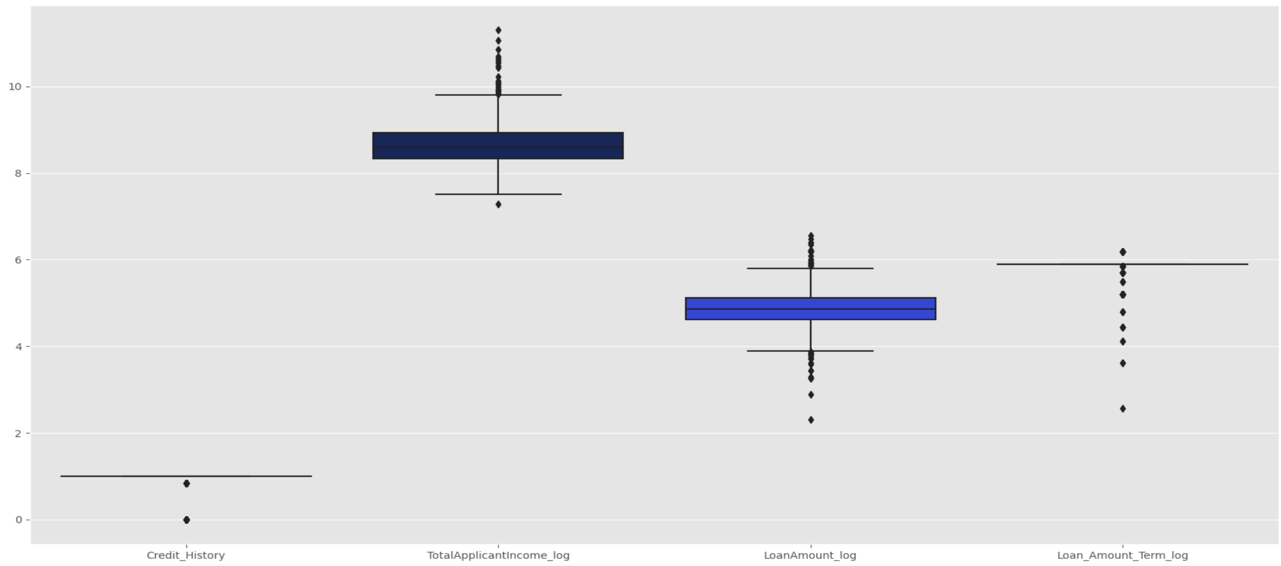
- There is no clear separation or trend indicating that higher incomes result in higher loan approval rates.
- This indicates that loan approval (Loan_Status) does not seem to be strongly dependent on the Total Applicant Income (log-transformed)

Credit_History vs Loan_Status

- There is distinction between applicants with Credit History values of 0 and 1.

- This indicates that credit History is a strong indicator of loan approval.

4.5 Using Box Plots to look for outliers



OBSERVATIONS AND INSIGHTS

Credit_History

- There are a few outliers below 0.5, but most of the data is concentrated at the 0 and 1 marks.
- This indicates that there may be a few erroneous or special cases in the data.

TotalApplicantIncome_log

- Shows a median around 8.25 with a relatively symmetrical distribution and several outliers on both the high and low ends
- This indicates that most applicants have a log-transformed income around 8.25
- The presence of outliers suggests there are a few applicants with exceptionally high or low incomes, which may need special consideration in loan evaluations.

LoanAmount_log

- Shows a median around 5.5 with a distribution skewed slightly to the higher values and numerous outliers on the higher end
- This indicates that most loan amounts are concentrated around the log value of 5.5, indicating a common loan amount among applicants.
- The outliers represent larger loans which may require special handling.

Loan_Amount_Term_log

- Shows a median around 6 with a slight skew towards higher values and numerous outliers on the higher end
- This suggests that the majority of loan terms are concentrated around the log value of 6, suggesting a common loan term duration
- The presence of outliers indicates that some loans have significantly longer terms.

CONCLUSION

In this project, we analyzed a loan prediction dataset to identify key factors influencing loan approval. Our process included data cleaning, transformation and exploratory data analysis (EDA). Here are the main takeaways:

Data Cleaning and Preparation:

- Addressed missing values in critical columns.
- Managed outliers in income and loan amounts.

Exploratory Data Analysis (EDA):

- Identified distributions in LoanAmount, ApplicantIncome, and Loan_Amount_Term.
- Found higher incomes and education levels correlated with loan approval.
- Examined interactions, such as ApplicantIncome and Credit_History, on loan approval.

Key Findings:

- Credit History: Positive credit history is crucial for loan approval.

- Income and Employment: Higher incomes and stable employment increase approval chances.
- Loan Amount and Term: Larger loan amounts decrease approval odds; loan terms have mixed impacts.
- Demographics: Male and married applicants have higher approval rates.
- Property Area: Semi-urban properties see higher approval rates.

Implications for Loan Prediction:

- Key predictors include Credit_History, ApplicantIncome, CoapplicantIncome and Property_Area.
- Insights support designing accurate and fair loan approval systems.

Future Work and recommendations:

- Developing and validating loan approval predictive models.
- Address potential biases in the loan approval process.
- Enhance the dataset with additional features.
- Test models in real-world scenarios and refine as needed.

In conclusion, our data preparation and analysis have set a solid foundation for predictive modeling, highlighting critical factors in loan approval and paving the way for more effective loan prediction systems.