**PROJECT DOCUMENTATION**

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |
| --- |
| **TITLE**: LOAN APPLICATION DATASET |
| **NAME**: KEERTHANA.D |
| **COURSE**: DA-DS (OFFLINE) |
| **BATCH**: JUNE 2025 |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| 1. | **Introduction** |
| 2. | **Aim** |
| 3. | **Business Problem / Problem Statement** |
| 4. | **Project Workflow** |
| 5. | **Data Understanding** |
| 6. | **Data Cleaning**   * Missing Values Imputation * Outlier Treatment * Handling Inconsistent Values |
| 7. | **Obtaining Derived Metrics** |
| 8. | **Filtering Data for Analysis** |
| 9. | **Statistical Analysis**   * Descriptive analysis * Test statistics and hypothesis testing |
| 10. | **Exploratory Data Analysis (EDA) – Univariate Analysis** |
| 11. | **Bivariate Analysis** |
| 12. | **Multivariate Analysis** |
| 13. | **Overall Insights from Analysis** |
| 14. | **Conclusion** |

1. **INTRODUCTION**

This dataset provides detailed records of loan applicants, combining demographic details, income information, loan parameters, and credit history. It contains both categorical attributes, such as **Gender**, **Marital Status**, **Education**, **Self-Employment status**, and **Property Area**, as well as numerical attributes like **Applicant Income**, **Co-applicant Income**, **Loan Amount**, and **Loan Term**.

The purpose of analyzing this dataset is to uncover trends, detect patterns, and prepare the data for further statistical analysis or predictive modeling. Such analysis can help financial institutions better understand applicant profiles and optimize their loan approval process.

1. **AIM OF THE PROJECT**

The aim of this project is to explore and analyse the loan dataset to gain meaningful insights into applicant profiles and loan characteristics. This involves cleaning and preparing the data, identifying key trends, and visualizing patterns that can help financial institutions understand their applicants better and make informed loan-related decisions.

1. **PROBLEM STATEMENT**

Deciding whether to approve a loan is challenging for banks and financial institutions. Without proper analysis, there is a risk of giving loans to applicants who may not be able to repay, or rejecting those who are capable. This project aims to analyse loan applicant data to find important factors that can help in making better and more accurate loan approval decisions.

1. **PROJECT WORKFLOW**

* **Data Collection** – import the loan dataset.
* **Data Understanding** – Check data shape, types, and sample records.
* **Data Cleaning** – Handle missing values, remove duplicates, drop unnecessary columns.
* **Data Transformation** – Create new columns (Total\_Income, EMI, Balance\_Income), convert data formats.
* **Outlier Detection** – Identify and treat extreme values.
* **EDA** – Perform univariate, bivariate, and multivariate analysis.
* **Visualization** – Create charts and graphs to understand patterns.
* **Insights** – Summarize main findings from the analysis.
* **Data Collection** – Load the loan dataset.
* **Data Understanding** – Check data shape, types, and sample records.
* **Data Cleaning** – Handle missing values, remove duplicates, drop unnecessary columns.
* **Data Transformation** – Create new columns (Total\_Income, EMI), convert data formats.
* **Outlier Detection** – Identify and treat extreme values.
* **EDA** – Perform univariate, bivariate, and multivariate analysis.
* **Visualization** – Create charts and graphs to understand patterns.
* **Insights** – Summarize main findings from the analysis.

1. **DATA UNDERSTANDING**

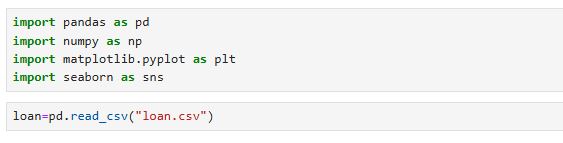
The dataset contains information about loan applications, including applicant details, income, loan amount, loan term, credit history, and property area. It provides a solid foundation to analyse applicant profiles, understand income–loan relationships, and explore factors that could influence loan approval decisions.

**Dataset Overview**

**Rows:** 367 records

**Columns:** 12 features

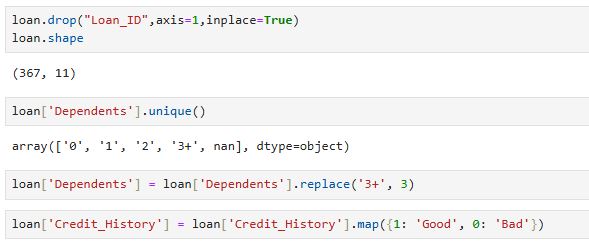
**Import libraries.**



* loan.shape – to find number of rows and columns.
* loan.head() – to print first 5 records.
* loan.tail() – to print last 5 records.
* loan.describe() – to get mean, median, min, max & standard deviation.
* loan.info() – to find null values and data types.

1. **DATA CLEANING**

“Data cleaning is the process of handling missing values, correcting inconsistencies, and preparing the dataset so that it is accurate and ready for analysis.”



* Drop the columns which does not contribute to analysis.
* .unique() is used to find all the unique values in a column.
* Standardize data formats (i.e.; converting “3+” as 3 in Dependents).
* mapping is used for binary categories (e.g., Credit\_History: 1 → "Good" , 0 → "Bad").
* loan.isnull().sum() - to check missing data in each column.

1. **Missing Values Imputation**

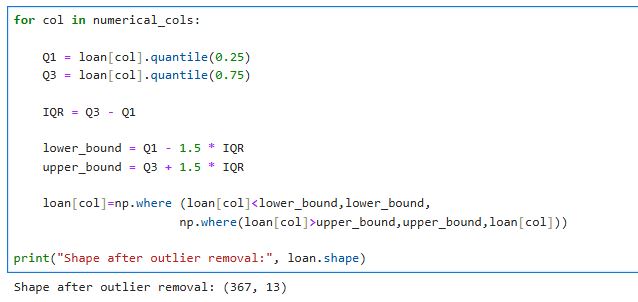


|  |  |
| --- | --- |
| * Handling missing values: * Fill categorical columns with the mode. * Fill numerical columns with the median. |  |

1. **Outlier Treatment**

* **Detection** – Boxplots and statistical methods (IQR – Interquartile Range) is used to identify extreme values in numerical variables like Applicant\_Income, *Co-*applicant\_Income, and Loan\_Amount.
* **Treatment** – Applying transformations (like log transformation) to reduce skewness caused by outliers.

Used Interquartile Range (IQR) method:



1. **Handling Inconsistent Values**

Inconsistent values occur when the same information is represented in different ways, which can cause errors in analysis. To ensure uniformity, values were standardized across the dataset.

* **Dependents**: "3+" was replaced with numeric 3 for consistency.
* **Gender**: All variations were standardized to "Male" and "Female".
* **Married**: All variations were standardized to "Yes" and "No".
* **Self\_Employed**: All variations were standardized to "Yes" and "No".
* **Property\_Area**: Ensured consistent format.

1. **OBTAINING DERIVED METRICES**

To enrich the dataset and improve analytical depth, the following new features (derived metrics) were created:

* **Total Income**
* Formula: Total\_Income = Applicant\_Income + Co-applicant\_Income
* Purpose: To understand the overall earning capacity of the household.
* **EMI (Equated Monthly Installment)**
* Formula: EMI = Loan\_Amount / Loan\_Amount\_Term.
* Purpose: To calculate the monthly loan repayment.

1. **FILTERING DATA FOR ANALYSIS**

**1.Missing/Invalid Records**

Rows with too many missing values were dropped after imputation.

Invalid values such as 0 income or unrealistic loan amounts were filtered out.

**2. Outlier Removal**

Very high Applicant\_Income or Loan\_Amount values (outliers) were removed to avoid bias.

**3. Loan Amount Restrictions**

Loans above a certain threshold (e.g., > 700) were excluded, as they were

unrealistic compared to most applicants.

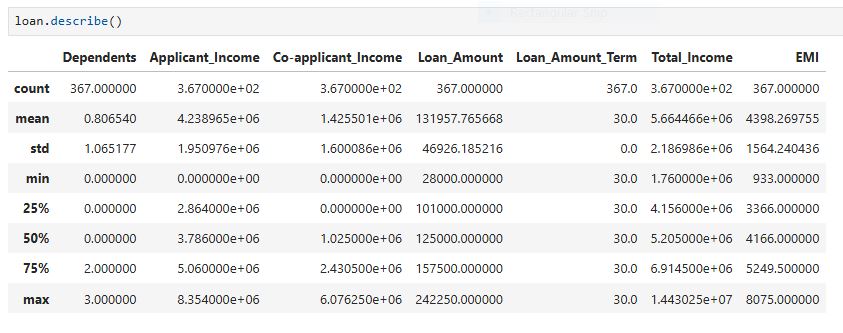
1. **Consistency Check**

Only records with proper Loan*\_*Amount*\_*Term and Credit*\_*History were retained for better insights**.**

1. **STATISTICAL ANALYSIS**

**1. Descriptive Analysis**

Used to summarize and understand the central tendency, spread, and distribution of data.

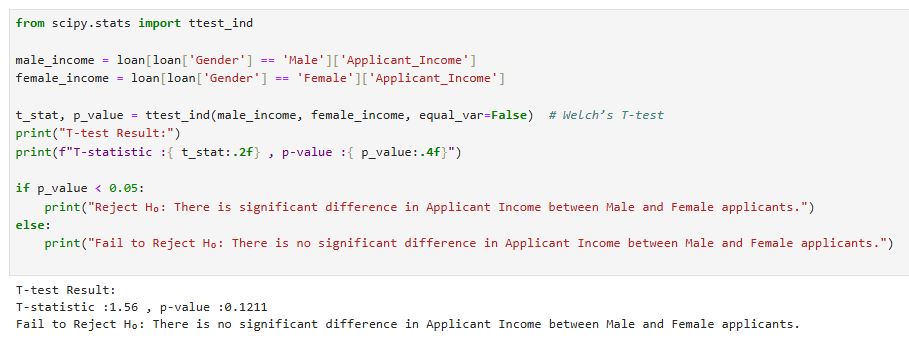
****

**2. Hypothesis testing**

Hypothesis 1: Independent two-sample t-test

**H₀ (Null Hypothesis):** Male and Female applicants have the same average income.

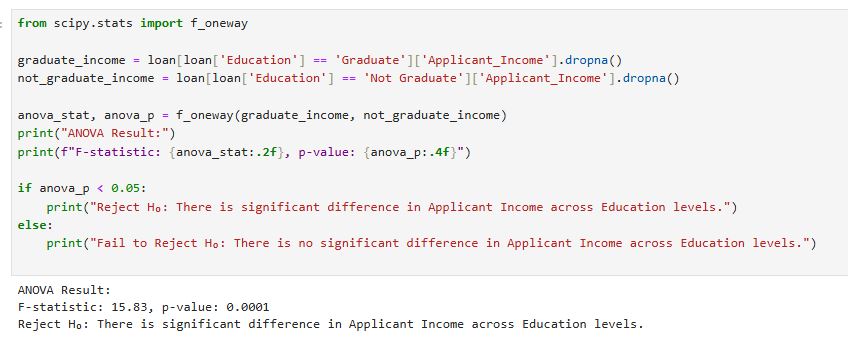
**H₁ (Alternate Hypothesis):** Male and Female applicants have different average incomes.



Hypothesis 2: One-way ANOVA test

**H₀ (Null Hypothesis):** Mean Applicant Income is the same for Graduates and Non-Graduates

**H₁ (Alternate Hypothesis):** Mean Applicant Income is not same for Graduates and Non-Graduates

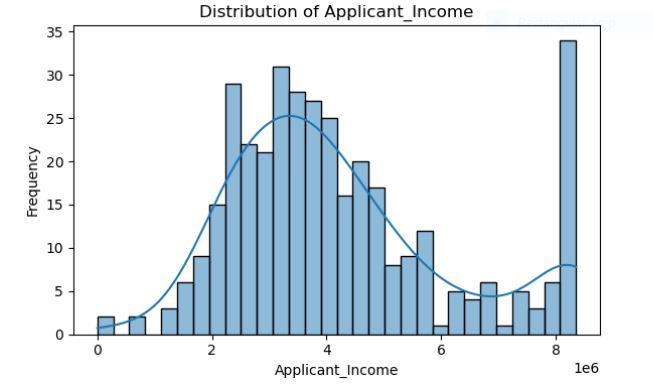


1. **EXPLORATORY DATA ANALYSIS (EDA)**

**UNIVARIATE ANALYSIS**

Univariate analysis means analysing **one variable at a time.**

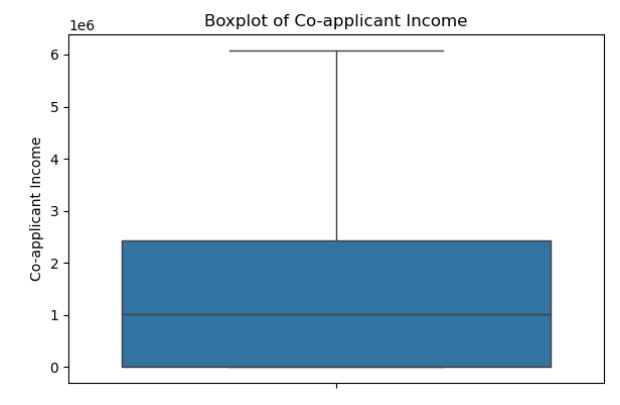
**HISTOGRAM**



**Insights Gained:**

* Q1 -> 25% of applicants fall below Q1
* Q2 -> 50% of applicants earn below this value and 50% earn above it.
* Q3 -> 75% of applicants fall below Q3 and the top 25% are high-income applicants.

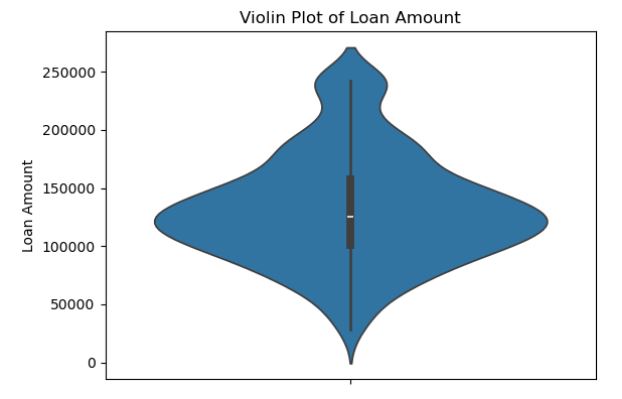
**Box plot**

****

**Insights Gained:**

* The middle line (median) of Co-applicant Income is around 1 million which means half of the co-applicants earn below this amount and half earn above.
* Most co-applicants have incomes between zero and around 2.5 million
* The longer whisker (upper side) suggests that smaller number of co-applicants have very high incomes compared to the rest.

**Violin plot**



**Insights Gained:**

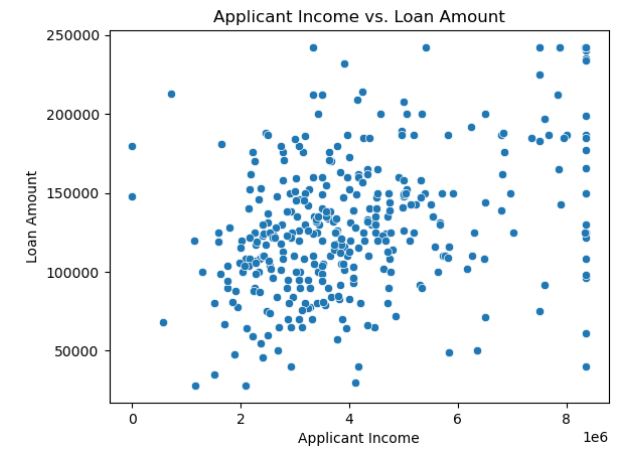
* The middle point is around 120,000 which means half of the applicants have loans below this value and half above.

* The violin is wider at the 100,000–150,000 range, showing that most applicants requested loan amounts in this range.
* Narrower regions indicate fewer applicants in those loan ranges.
* The distribution shows two bulges — one around 1,20,000 and another near 2,30,000.
* This suggests that there are two common loan sizes.
* The upper tail suggests a few applicants requested very high loan amounts.

1. **BIVARIATE ANALYSIS**

Bivariateanalysis is the statistical study of the relationship between twovariables.

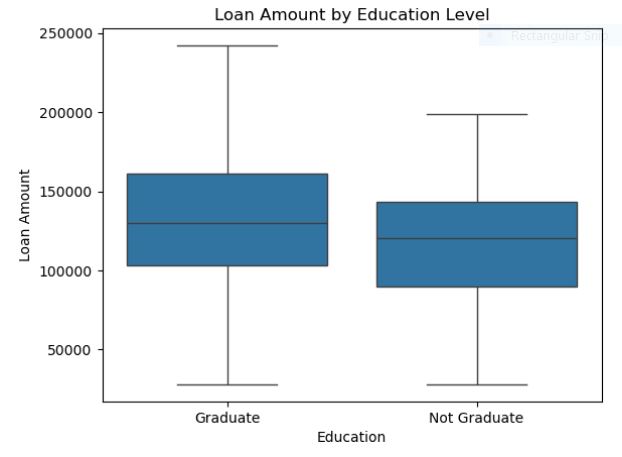
**Scatter plot**



**Insights Gained:**

* Most applicants earn between 1 million to 5 million and their loan amounts are usually between 75K and 150K.
* This indicates that the majority of loan seekers belong to the middle-income group.
* As applicant income increases, loan amount tends to increase slightly but not strongly.
* This could be because banks don’t give loans based only on income they also look at things like property value, repayment time and credit score.

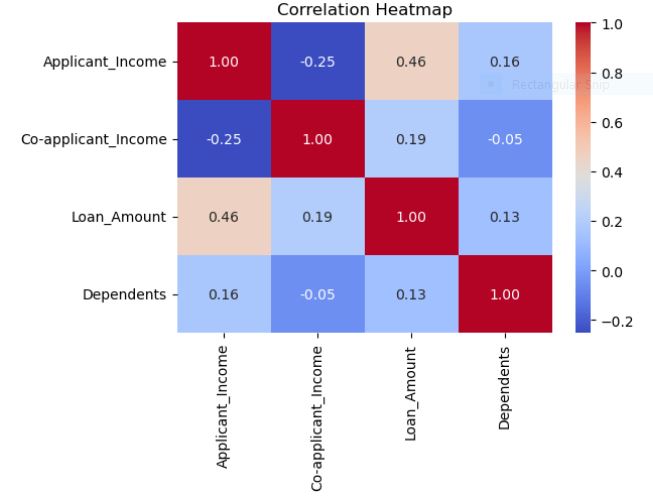
**Box plot**

****

**Insights Gained:**

* The chart shows that the median loan amount is similar for both graduates and non-graduates.
* However, the distribution of loan amount for graduates is more spread out as indicated by the larger box and more numerous outliers.
* This suggests that while the average loan amounts are similar, graduates tend to have a wider range of loan amounts compared to non-graduates.

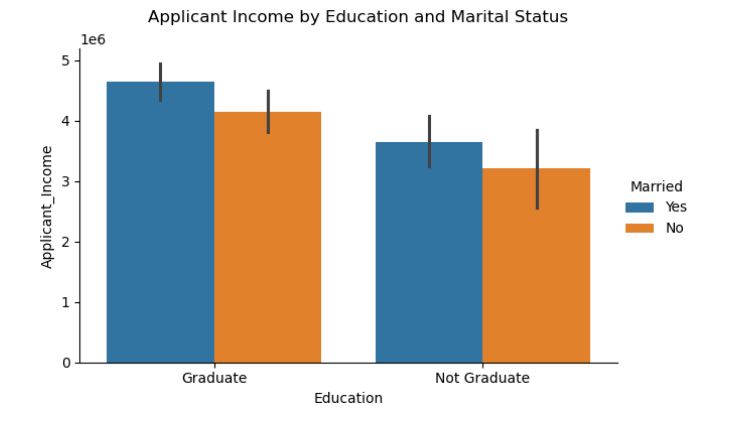
1. **MULTIVARAIATE ANALYSIS**



**Insights Gained:**

* Applicant's Income and Loan Amount: There is a strong positive link (0.46). This indicates that as an applicant's income increases, the requested loan amount also tends to increase.
* Co-applicant's Income and Loan Amount: Co-applicant Income has a positive correlation with loan amount (0.19), which is weaker than the applicant's income but still shows that a higher co-applicant income is associated with a larger loan.
* Dependents and Income: Correlation between Applicant Income and Dependents is 0.16. This suggests that applicants with more dependents tend to have a higher income.
* Co-applicant's Income and Dependents: A negative correlation exists between them (−0.05). This implies that as the number of dependents increases, the co-applicant's income tends to slightly decrease.

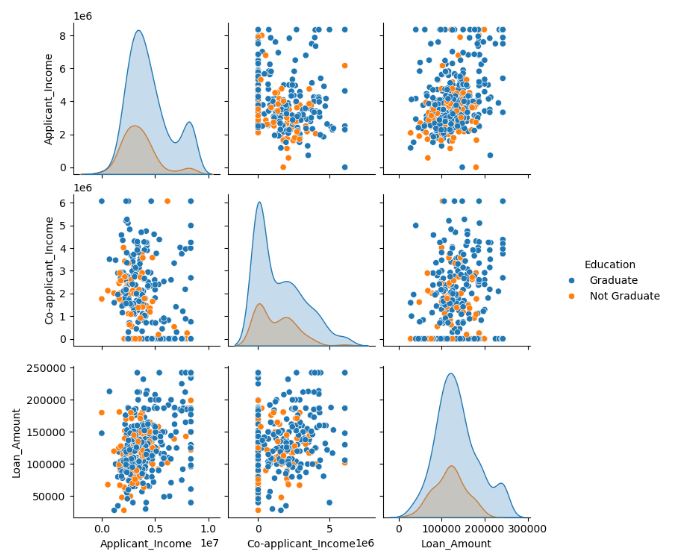
**STACKED BAR CHART**



**Insights Gained:**

* **Income and Education:** For both married and unmarried applicants, graduates have a slightly higher average income than non-graduates.
* **Marital Status and Income:**  Across both education levels, married applicants have a significantly higher average income than unmarried applicants.
* **Combined Insight:** The highest average income is found among married graduates, followed by married non-graduates, unmarried graduates, and finally unmarried non-graduates.

**PAIR PLOT**



**Insights Gained:**

* Education strongly affects applicant income – graduates earn more.
* Loan amounts are higher for graduates compared to non-graduates.
* Applicant income is positively related to loan amount, especially for graduates.
* Co-applicant income is less influenced by education and shows weaker relation to loan amount.
* Graduates have a financial advantage in loan approvals due to higher incomes and higher loan eligibility.

1. **OVERALL INSIGHTS FROM ANALYSIS**

**1. Data Quality Insights.**

* **Missing Values:** Some columns (like Loan Amount, Credit History, Co-applicant Income) have missing values that need imputation.
* **Inconsistent Entries:** Categorical fields (e.g., Gender, Education, Self\_Employed) may contain inconsistent spellings/cases ("Male"/"male", "Graduate"/"graduate").
* **Outliers:** Applicant and Co-applicant Income show extreme outliers that could distort analysis.
* **Skewed Distributions:** Loan Amount and Income variables are right-skewed, requiring transformations (e.g., log scale).
* **Duplicate Records:** Dataset should be checked for duplicate applicant IDs or repeated rows.

**2. Data Preparation Insights**

* **Handling Missing Values:** Impute numerical fields (median/mean) and categorical fields (mode).
* **Feature Scaling:** Standardize or normalize Applicant Income, Loan Amount, and Co-applicant Income to reduce scale impact.
* **Outlier Treatment:** Apply capping or log transformation for income and loan variables.
* **Derived Metrics:** Create new features such as **Total Income** and **EMI** for better predictive power.
* **Filtering Data:** Remove the rows which are not used for analysis.

**3. Univariate Analysis Insights**

* **Applicant Income:** Right-skewed distribution with a few very high-income outliers. Most applicants fall in the lower-to-mid income range.
* **Co-applicant Income:** Many applicants have zero co-applicant income, suggesting either single earners or dependent applicants.
* **Loan Amount:** Concentrated around 100,000–150,000, with a few very large loan requests indicating outliers.
* **Loan Amount Term:** Majority of applicants prefer the standard **360 months (30 years)** tenure.

**4. Bivariate Analysis Insights**

* **Education vs. Loan Amount:** Graduates tend to apply for slightly higher loan amounts compared to non-graduates.
* **Marital Status vs. Applicant Income:** Married applicants usually have higher household income than unmarried ones.
* **Property Area vs. Loan Amount:** Urban applicants often seek higher loan amounts compared to semi-urban and rural.

**5. Multivariate Analysis Insights**

* Applicants with **high income + good credit history** have the **highest loan approval rate**.
* Even low-income applicants often get approved if credit history is strong, showing credit score is more critical than income.
* **Graduate + Salaried applicants** tend to request and receive higher loan amounts than **non-graduate + Self-employed** applicants.
* Self-employed non-graduates often apply for smaller loans and face more rejections.
* **Married male applicants** dominate higher loan amounts, possibly due to dual-income households.

**6. Hypothesis Testing Results**

**T-Test: (** Applicant Income vs. Gender )

* **H₀:** There is no significant difference in applicant income between male and female applicants.
* **Result:** T-statistic ≈ 1.56, p-value ≈ 0.12

**Conclusion:** Fail to reject H₀ → No significant difference, male and female applicants have similar incomes statistically.

**ANOVA: (** Applicant Income across Education **)**

* **H₀:** There is significant difference in Applicant Income across Education levels
* **Result:** F-statistic ≈ 15.83, p-value ≈ 0.0001

**Conclusion:** Reject H₀ → Mean Applicant Income is not same for Graduates and Non-Graduates.

1. **CONCLUSION**

This loan dataset project provided a detailed exploration of the factors influencing loan applications and approvals. The analysis offers valuable insights for financial institutions and stakeholders to make more informed decisions.