# 📊 Loan Application Analysis & Prediction — Data Analyst Portfolio

Prepared by: Ugene Sanjit

Date: [Insert Date]

## 🔍 Project Summary

This project focused on analyzing and modeling loan application data to extract meaningful insights and predict loan approvals. The core objectives were to identify key patterns, handle data quality issues, apply statistical tests, and build a predictive model to support data-driven loan decisions.

## 🧹 1. Data Cleaning & Preprocessing

Initial Inspection:

- Reviewed dataset structure, column data types, and value distributions.

- Identified missing values and duplicate entries.

Missing Values Handling:

- Married, Num Dependants, Num Units → Filled with median values.

- Gender → Filled using mode (most frequent value).

Duplicates:

- Removed 1 duplicate row to maintain dataset integrity.

Data Type Fixes:

- Ensured columns were appropriately set (categorical, numeric, etc.).

Data Normalization:

- Standardized categorical entries for consistency.

- Normalized numerical fields where required.

## 📌 2. Key Data Insights

|  |  |
| --- | --- |
| Column | Missing Values |
| Married | 3 |
| Num Units | 4 |
| Num Dependants | 3 |
| Gender | 14 |

Duplicates Removed: 1

## 📈 3. Statistical Analyses

✅ Q1: Loan Status by Income, Marital Status & Bankruptcy

Exploratory grouping of loan statuses helped identify distribution patterns by applicant income, marital status, and bankruptcy status.

📊 Q2: Income vs Approved Loan Amount (Paired T-Test)

T-Statistic: -29.5753

P-Value: 0.0000 (Significant)

Interpretation: Applicant income and loan amounts differ significantly. Loan approvals likely consider multiple criteria beyond income.

👥 Q3: Loan Approval by Race (Chi-Square Test)

Chi-square Statistic: 102.6351

P-Value: 0.0000 (Significant)

|  |  |
| --- | --- |
| Race | Approval Rate |
| White | 90.83% |
| Hispanic | 76.58% |
| Black | 67.51% |

Conclusion: Race significantly affects loan approval likelihood, indicating potential systemic bias.

🚻 Q4: Loan Approval by Gender (Chi-Square Test)

Chi-square Statistic: 0.4654

P-Value: 0.4951 (Not Significant)

|  |  |
| --- | --- |
| Gender | Approval Rate |
| Male | 87.91% |
| Female | 86.45% |

Conclusion: No statistically significant difference in approval rates by gender.

## 🤖 4. Loan Approval Prediction Model

Model Type: Logistic Regression

Features Used: Marital status, race, income, gender, number of dependents, loan amount

## 🧮 Model Evaluation

Accuracy: Reasonable accuracy for binary classification.

Residuals Analysis:

- Mean Residual: 0.1156 (close to zero → unbiased model)

- Some skewness noted, indicating potential edge-case prediction errors.

## 📌 Conclusion

The model performs well, with room for refinement.

Residual patterns suggest some scope for feature enhancement or model tuning.

A promising baseline model for decision-making in loan approvals.

## ✅ Final Takeaways

- Significant demographic disparities in approval patterns.

- Strong evidence that income alone doesn't dictate loan approvals.

- Logistic Regression serves as a good starting point for modeling loan outcomes.

- Ethical concerns around race-based approval differences should be investigated further.