Financial Risk Analysis Report

Yaw Jantuah Boateng

February 2026

1 Introduction

This report summarizes the cleaning and analysis process on a loan default dataset. A publicly available financial dataset was extracted from Kaggle for this exercise. The data was then cleaned and analysed with Python using Google Colab notebook. The process was highlighted on GitHub repositories to show a reproducible process. A report was also prepared with LaTeX using Overleaf editor.

2 Data Cleaning

Missing values were explored in the dataset. Fortunately, there were no missing values in any of the columns, implying no need to perform any missing data imputation. Duplicates were then successfully removed from the data.

```
# Check missing values
df.isnull().sum()

# Remove duplicates
df = df.drop_duplicates()
```

3 Initial Findings

Summary statistics indicate a total of 255,347 participants in the data, with an average age of 43 and a standard deviation of 14.99. Income averages about 82,500, and the loan amount is around 127,500. Loan default shows a remarkable average of 0.12 over an average loan term of 36 months. The average credit score was also recorded to be 574.26.

Table 1: Summary Statistics After Cleaning

	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditL
count	255347.000000	255347.000000	255347.000000	255347.000000	255347.000000	255347.00
mean	43.498306	82499.304597	127578.865512	574.264346	59.541976	2.50
std	14.990258	38963.013729	70840.706142	158.903867	34.643376	1.11
min	18.000000	15000.000000	5000.000000	300.000000	0.000000	1.00
25%	31.000000	48825.500000	66156.000000	437.000000	30.000000	2.00
50%	43.000000	82466.000000	127556.000000	574.000000	60.000000	2.00
75%	56.000000	116219.000000	188985.000000	712.000000	90.000000	3.00
max	69.000000	149999.000000	249999.000000	849.000000	119.000000	4.00

Basic distribution analysis highlights that income and loan amounts appear to be right-skewed, suggesting the presence of high-value outliers or a concentration of lower values with a tail extending to higher values.

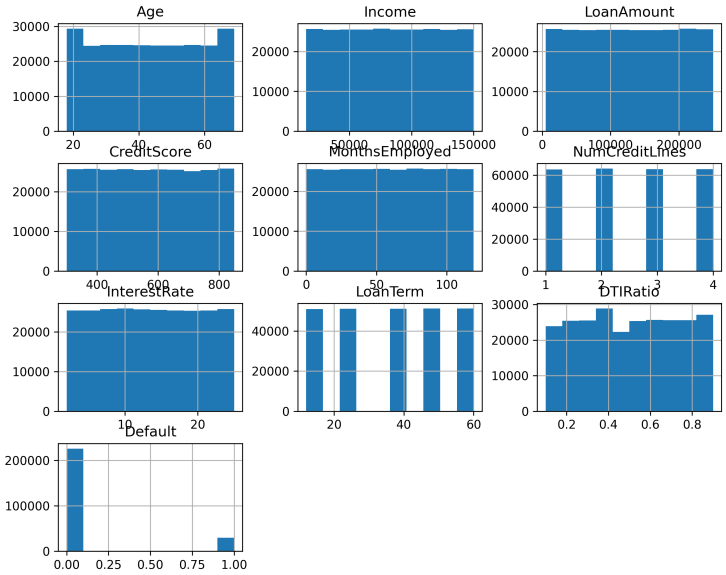

Figure 1: Basic Distribution Analysis

4 Conclusion

The dataset is now clean and ready for further modelling. The outputs of the basic analysis provide insights into the dataset's characteristics, such as the absence of missing values and the skewed distributions of income and loan amount.

A Full Python Code

```
# upload data
from google.colab import files
uploaded = files.upload()

import pandas as pd

# Load the dataset
df = pd.read_csv("raw_data.csv.csv")

# Show first 5 rows
df.head()

# Check structure
df.info()

# Check summary statistics
summary_table = df.describe()
print(summary_table.to_latex())

# Check missing values
df.isnull().sum()

# Remove duplicates
df = df.drop_duplicates()
print("done")

# Safe cleaned dataset:
df.to_csv('cleaned_data.csv', index=False)
files.download("cleaned_data.csv")
print("Saved")

# Basic Visualization
import matplotlib.pyplot as plt
df.hist(figsize=(10, 8))
plt.title("Distribution␣Analysis")
plt.savefig("distribution_analysis.png", dpi=300,
    bbox_inches='tight')
plt.show()
files.download("distribution_analysis.png")
```