LOAN DEFAULT PREDICTION

AIM: The "Loan Default Prediction" project aims to analyse and predict loan defaults by visualising various customer and loan metrics. The PowerBI dashboard facilitates data-driven decisions by displaying insights into customer demographics, loan characteristics, and payment histories. The dashboard helps identify key patterns and trends that contribute to loan defaults.

METRICS:

1.Customer ID

2.Loan Type

3.Loan Amount (USD)

4.Loan Term

8.Years Employed

9.Number of Dependents

10.Payment History

11.Loan-to-Value Ratio

12. Original Down Payment (USD)

METHODOLOGY:

1. **Data Generation**: Generate a raw, structured dataset using ChatGPT with missing values, outliers, and long decimal values.
2. **Data Collection**: Copy the generated data to an Excel sheet named "Loan Default Prediction Raw Data.xlsx".
3. **Data Cleaning**: Clean and trim the dataset using Python on Google Colab, handling missing values and outliers, and improving data quality.
4. **Data Structuring**: Save the cleaned and structured dataset as "Loan Default Prediction Structured Data.xlsx".
5. **Data Import**: Import the structured dataset into PowerBI.
6. **Dashboard Creation**: Create a PowerBI dashboard to visualise the data and derive insights.

METRICS DESCRIPTION:

1. **Customer ID**: A unique identifier for each customer, ensuring that each entry in the dataset is distinct and traceable.
2. **Loan Type**: The type of loan taken by the customer (e.g., mortgage, personal loan, auto loan), providing insights into the distribution and performance of different loan types.
3. **Loan Amount (USD)**: The total amount of money borrowed by the customer, expressed in US dollars. This metric helps analyse the correlation between loan amount and default rates.
4. **Loan Term**: The duration over which the loan is to be repaid, typically expressed in months or years. It provides an understanding of how loan duration impacts default probability.
   Ideal Value: Loan terms that match the industry standards for the specific loan type.

   Non-Ideal Value: Terms that are unusually short or long, which could indicate risk.

5. **Years Employed**: The number of years the customer has been employed. This metric is crucial in assessing the financial stability and reliability of the borrower.

   Longer employment durations, indicating job stability (e.g., 5+ years), are ideal. Short employment durations or frequent job changes, indicating instability (e.g., less than 1 year), are non-ideal.

6. **Number of Dependents**: The number of individuals financially dependent on the customer.

   This provides insights into the customer's financial obligations and potential strain on their repayment capacity. A lower number of dependents is ideal as it indicates fewer financial burdens.

7. **Payment History**: A record of the customer's past loan payments, indicating their payment behaviour and reliability.

   Consistent on-time payments, showing reliability, are ideal.

8. **Loan-to-Value Ratio**: The ratio of the loan amount to the appraised value of the asset purchased with the loan. This metric helps in assessing the risk associated with the loan.

   Ratios below 80%, indicating lower risk, are ideal. Ratios above 80%, indicating higher risk and lower equity in the asset, are non-ideal.

9. **Original Down Payment (USD)**: The initial amount paid by the customer towards the loan, expressed in US dollars.

   It reflects the customer’s initial financial commitment and potential risk. Higher down payments, indicating greater customer commitment

10. **Credit Score** is a numerical expression representing the customer’s creditworthiness.

    High credit scores (e.g., 700+), indicating low risk, are ideal. Low credit scores (e.g., below 600), indicating higher risk, are non-ideal.

11. **Annual Income (USD)** represents the total income of the customer per year, expressed in US dollars.

    Higher income, indicating better repayment capacity (e.g., $50,000+), is ideal. Lower income, which might indicate limited repayment capacity (e.g., less than $20,000), is non-ideal.

12. **Debt-to-Income Ratio** is the ratio of the customer’s total monthly debt payments to their gross monthly income.

    Ratios below 36%, indicating manageable debt levels, are ideal. Ratios above 43%, indicating a higher risk of default, are non-ideal.

CODE FOR DATA CLEANING:

```
import pandas as pd
import numpy as np
from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/My Drive/Raw Data Set.xlsx'
df = pd.read_excel(file_path)

def replace_missing_with_random(df, column):
    """Replace missing values with a random value within the column's range."""
    if df[column].dtype in [np.float64, np.int64]:
        min_val = df[column].min()
        max_val = df[column].max()
        df[column].fillna(np.random.uniform(min_val, max_val), inplace=True)
    elif df[column].dtype == object:
        df[column].fillna(df[column].mode()[0], inplace=True)
    return df
```

```
def replace_outliers_with_zero(df, column):
    """Replace outliers in the column with 0."""
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df.loc[(df[column] < lower_bound) | (df[column] > upper_bound), column] = 0
    return df

df['Customer ID'] = pd.to_numeric(df['Customer ID'], errors='coerce', downcast='integer')
df['Loan Amount (USD)'] = pd.to_numeric(df['Loan Amount (USD)'], errors='coerce')
df['Loan Term (Months)'] = pd.to_numeric(df['Loan Term (Months)'], errors='coerce')
df['Credit Score'] = pd.to_numeric(df['Credit Score'], errors='coerce')
df['Debt-to-Income Ratio'] = pd.to_numeric(df['Debt-to-Income Ratio'], errors='coerce')
df['Years Employed (Current Job)'] = pd.to_numeric(df['Years Employed (Current Job)'], errors='coerce')
df['Number of Dependents'] = pd.to_numeric(df['Number of Dependents'], errors='coerce')
df['Payment History (Past 24 Months)'] = pd.to_numeric(df['Payment History (Past 24 Months)'], errors='coerce')
df['Loan-to-Value Ratio'] = pd.to_numeric(df['Loan-to-Value Ratio'], errors='coerce')
df['Original Down Payment (USD)'] = pd.to_numeric(df['Original Down Payment (USD)'], errors='coerce')
```

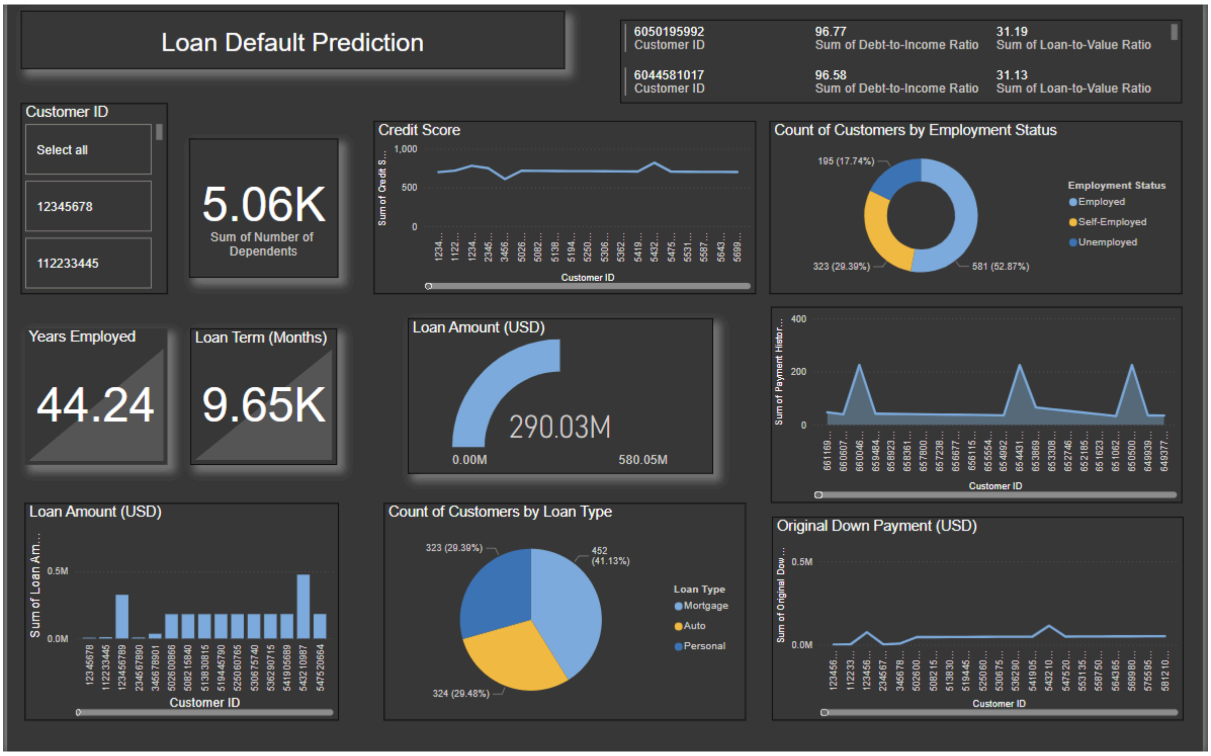
```
for col in df.columns:
    df = replace_missing_with_random(df, col)
```

```
columns_to_abs = [
    'Loan Amount (USD)',
    'Loan Term (Months)',
    'Credit Score',
    'Debt-to-Income Ratio',
    'Years Employed (Current Job)',
    'Number of Dependents',
    'Payment History (Past 24 Months)',
    'Loan-to-Value Ratio'
]
```

```
for col in columns_to_abs:
    df[col] = df[col].abs()
columns_to_round = [
    'Loan Amount (USD)',
    'Loan Term (Months)',
    'Credit Score',
    'Debt-to-Income Ratio',
    'Years Employed (Current Job)',
    'Number of Dependents',
    'Payment History (Past 24 Months)',
    'Loan-to-Value Ratio',
    'Original Down Payment (USD)'
]
for col in columns_to_round:
    df[col] = df[col].round(2)
df['Customer ID'] = df['Customer ID'].astype(int)
```

```
cleaned_file_path = '/content/Cleaned_Data_Set.xlsx'
df.to_excel(cleaned_file_path, index=False)
files.download(cleaned_file_path)
```

OUTPUT:



Loan Default Prediction

6050195992 Customer ID | 96.77 Sum of Debt-to-Income Ratio | 31.19 Sum of Loan-to-Value Ratio

6044581017 Customer ID | 96.58 Sum of Debt-to-Income Ratio | 31.13 Sum of Loan-to-Value Ratio

RESULT ANALYSIS:

Chart 1: Credit Score

The Credit Score line chart visualises the creditworthiness of each customer in the dataset, providing a clear view of how credit scores are distributed among different customers.

- X-Axis (Horizontal Axis): Represents the Customer ID. Each unique Customer ID is plotted along the X-axis, allowing for the identification of individual customers. This axis ensures that each data point corresponds to a specific customer, maintaining the integrity of the dataset and facilitating easy traceability.
- Y-Axis (Vertical Axis): Represents the Credit Score. Credit scores are numerical expressions that indicate a customer's creditworthiness, typically ranging from 300 to 850. Higher values suggest better creditworthiness and lower risk, while lower values indicate higher risk.
- High credit scores (e.g., 700+), indicating low risk, are ideal. Low credit scores (e.g., below 600), indicating higher risk, are non-ideal.

Chart 2: Donut Chart For Customer's Employment Status

The donut chart provides a clear visualisation of the employment status of 1099 customers, segmented into three categories: Employed, Self-Employed, and Unemployed.

- Employed:
    - Percentage: 52.87%
    - Number of Customers: 581
    - Description: This is the largest segment, indicating that the majority of customers are employed, suggesting a stable income source.
- Self-Employed:
    - Percentage: 29.39%
    - Number of Customers: 323
    - Description: This segment shows a significant portion of customers who are self-employed, indicating varied income stability.
- Unemployed:
    - Percentage: 17.74%
    - Number of Customers: 195
    - Description: This is the smallest segment, representing customers who are unemployed, indicating a higher potential risk for loan defaults.

Chart 3: Gauge Chart for Loan Amount

The gauge chart visualises the total loan amount for 1099 customers, with the total amounting to 290.03 million USD. This type of chart is effective for showing the proportion of a specific metric relative to a set target or maximum value.

Range: The gauge chart's scale represents the total loan amount, measured in USD.

Chart 4: Area Chart for Payment History (past 24 months)

The area chart visualises the sum of payment history over the past 24 months for 1099 customers, with a maximum value of 250.03. This chart is particularly useful for showing how payment histories vary significantly across different customers.

- X-Axis (Horizontal Axis): Represents the Customer ID. Each Customer ID is plotted along this axis, allowing for individual tracking of payment history.
- Y-Axis (Vertical Axis): Represents the sum of payments made by each customer over the past 24 months, measured in USD.
- Maximum Value: The highest sum of payments recorded is 250.03 USD.
- Variation: The area chart highlights the variability in payment histories, showing significant differences among customers.
- Peaks and Valleys: The chart will display peaks where customers have made higher total payments and valleys where payments are lower, indicating varying payment capacities and histories.

Chart 5: Column Chart for Loan Amount ($ USD)

The column chart visualises the loan amounts for 1099 customers, with each column representing a unique Customer ID and the height of the column indicating the loan amount in USD. The maximum loan amount in this dataset is $620,000 USD.

- X-Axis (Horizontal Axis): Represents the Customer ID. Each unique Customer ID is displayed along this axis, allowing for individual tracking of loan amounts.
- Y-Axis (Vertical Axis): Represents the loan amount in USD, with a range extending up to the maximum value of $620,000 USD.
- Maximum Value: The tallest column corresponds to the maximum loan amount of $620,000 USD.
- Variation: The chart showcases significant differences in loan amounts across different customers, with column heights varying greatly, indicating a wide range of loan amounts.

Chart 6: Pie Chart for Loan Type

The pie chart shows the distribution of loan types among 1099 customers:

- Mortgage: 41.13% (452 customers)
- Auto Loans: 29.48% (324 customers)
- Personal Loans: 29.39% (323 customers)

Inference

- Dominance of Mortgages: Most customers have mortgage loans, indicating a focus on home financing.
- Balanced Auto and Personal Loans: Auto and personal loans are nearly equally represented, showing diverse borrowing needs.
- Strategic Focus: Financial services could be tailored to emphasise mortgage products while also addressing the needs for auto and personal loans.

Chart 7: Line Chart for Original Down Payment (USD)

The line chart shows the original down payments made by 1099 customers, with each point representing a Customer ID. The chart ranges from a minimum down payment of $1,250 to a maximum of $491,881.

Inference

- High Variation: The chart displays significant differences in down payments, from very low to very high amounts.
- Financial Commitment: Peaks indicate high financial commitment, while troughs show lower down payments.
- Diverse Payments: The wide range reflects varying financial capacities among customers, useful for assessing financial stability and lending risk.

Cards:

These individual cards provide insights into each customer's employment stability, loan preferences, and financial responsibilities, which are crucial for assessing risk and tailoring financial services.

1. **Years Employed:**
    - ○ Varies widely among customers.
    - ○ Inference: Longer employment suggests stability and reliable income, while shorter employment may indicate potential instability.

2. **Loan Term (Months):**
   - Range: Shows variation from short to long-term loans.
   - Inference: Longer terms mean lower monthly payments but higher total interest, while shorter terms indicate quicker repayment but higher monthly payments.
3. **Number of Dependents:**
   - Range: Varies from none to several dependents.
   - Inference: More dependents suggest higher financial responsibilities, impacting loan repayment ability, whereas fewer dependents suggest greater financial flexibility.

Multi-Row Card

The multi-row card displays:

- Debt-to-Income Ratio (DTI): Ranges from 0.6 to 96.77. Higher values indicate more of a customer's income is spent on debt, signalling potential financial strain.
- Loan-to-Value Ratio (LTV): Ranges from 0.01 to 31.19. Higher values suggest a larger loan relative to the asset value, indicating higher risk.

Inference

- High DTI: May indicate financial stress; low DTI suggests manageable debt levels.
- High LTV: Indicates higher financial risk; low LTV reflects lower risk.

SUMMARY OF ANALYSIS:

The dashboard provides a comprehensive overview of the financial profiles and behaviours of 1099 customers, highlighting key aspects such as creditworthiness, employment status, loan amounts, and payment histories. The analysis reveals significant variability in financial metrics across customers, which is crucial for assessing loan risk and tailoring financial services.

1. **Credit Scores** vary widely, with higher scores indicating lower risk and better creditworthiness. The line chart illustrates a broad distribution of credit scores, reflecting diverse financial health among customers.
2. **Employment Status** shows that a majority are employed (52.87%), with significant portions self-employed (29.39%) and unemployed (17.74%). This distribution affects income stability and potential risk for loan defaults.

3. **Loan Amounts and Types** reveal a broad range of loan sizes and preferences, with a total loan amount of 290.03 million USD. Mortgages are the most common loan type (41.13%), followed by auto and personal loans, suggesting a focus on home financing but also diverse borrowing needs.

Key Points

1. **Diverse Financial Profiles**: The wide range of credit scores, loan amounts, and down payments highlights the varied financial situations and risk levels of customers.
2. **Employment and Financial Stability**: A large portion of customers are employed, which suggests relatively stable income sources. However, the presence of self-employed and unemployed individuals indicates varying levels of income stability and risk.
3. **Loan Preferences**: The predominance of mortgage loans reflects a focus on home financing, while the variation in loan amounts and types shows diverse customer needs and financial capacities.