MACHINE LEARNING PROJECT

Individual Financial Risk Analysis

**Statement problem:** Financial institutions, investors, and policymakers face the challenge of accurately assessing individual financial risk to minimize defaults, reduce losses, and promote equitable lending. Traditional methods often rely on limited indicators like income or age, which fail to capture the multidimensional nature of financial risk.

This project addresses the problem by:

* Analyzing demographic and financial factors (age, gender, education, marital status, employment, income, credit score, debt-to-income ratio, assets, loan amount, etc.).
* Identifying how these variables interact to influence risk ratings (Low, Medium, High).
* Building a predictive machine learning model (XGBoost with hyperparameter tuning) to classify individuals’ financial risk more accurately**.**

**Definition: Financial risk** refers to the possibility of losing money or facing negative financial outcomes due to **uncertain future events**. These risks can affect individuals, businesses, or government.

Types of Financial Risk:

| **Type** | **Meaning** | **Example** |
| --- | --- | --- |
| **Market Risk** | Risk from changes in market prices (stocks, interest rates, etc.) | Share prices drop after investing |
| **Credit Risk** | Risk that a borrower won’t repay a loan | A customer defaults on a loan |
| **Liquidity Risk** | Inability to quickly convert assets into cash without loss | Can’t sell property during a crisis |
| **Operational Risk** | Losses from system failure, fraud, or human error | A bank loses data due to a cyberattack |
| **Legal/Regulatory Risk** | Risk from lawsuits or changing regulations | New tax law increases costs |

Why Financial Risk Matters for Individual?

Understanding financial risk is crucial because it allows individuals, businesses, and governments to make informed decisions and prepare for uncertainty. For investors, it helps avoid high-risk assets that could lead to significant losses. For businesses, it guides financial planning, budgeting, and strategic decision-making by identifying potential threats that could impact profitability. Banks and lenders use financial risk analysis to assess whether borrowers are likely to repay their loans, reducing the chance of default. Governments also rely on risk assessments to manage economic stability and respond effectively to financial crises. In essence, be

import numpy as np # For data handling and preprocessing.#

import pandas as pd

import json

import matplotlib.pyplot ##as plt **EDA (plots, distributions, correlations)**.

import seaborn as sns

from sklearn.model\_selection import train\_test\_split ## 

**Scikit-learn tools** →train\_test\_split: divide train & test sets.

 StandardScaler/MinMaxScaler: scale numerical features.

 LabelEncoder: encode categorical variables.

 metrics: evaluate your model (accuracy, F1, ROC, AUC).

from sklearn.preprocessing import StandardScaler,LabelEncoder,MinMaxScaler

from sklearn.metrics import accuracy\_score, roc\_auc\_score,roc\_curve, auc,ConfusionMatrixDisplay,classification\_report,recall\_score,f1\_score

from xgboost import XGBClassifier ## Your main ML algorithm for classification.

import optuna ## For hyperparameter tuning (finding the best parameters for XGBoost).

import shap ### interpretability (which features affect risk most).

from imblearn.over\_sampling import RandomOverSampler To fix **imbalanced dataset** (if Low-risk cases dominate).

plt.style.use("ggplot")

import warnings

warnings.filterwarnings("ignore")

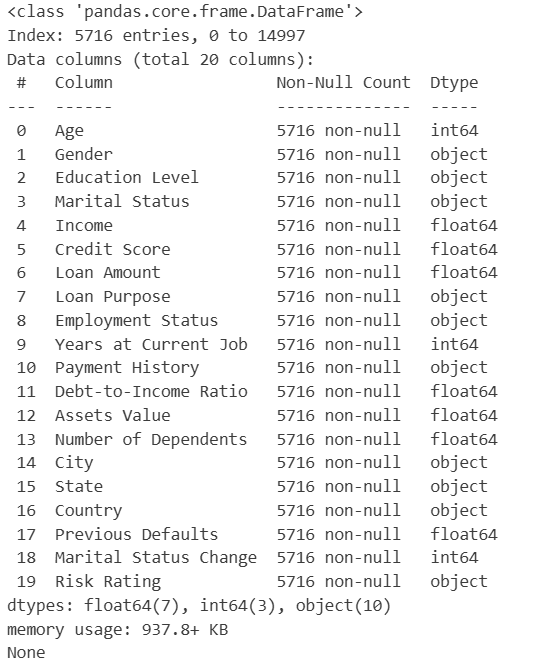
## Hides **warning messages** that appear during execution (like “DeprecationWarning” or “FutureWarning”).

Loading the DATASET FROM THE FILES

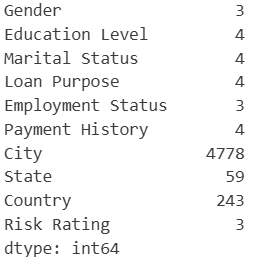
data = pd.read\_csv("C:/Users/Shabnam Kabir/Downloads/Fp/financial\_risk\_assessment.csv.zip").dropna()

print(data.head())

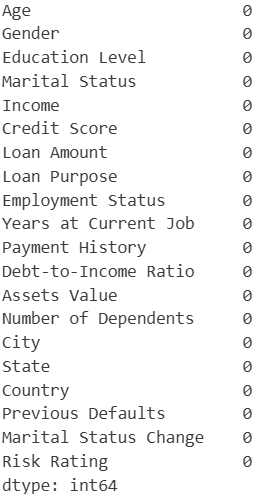
print(data.info())



print(data.select\_dtypes('object').nunique())



print(data.isna().sum())



## Distribution of Age

plt.figure(figsize=(10, 6))

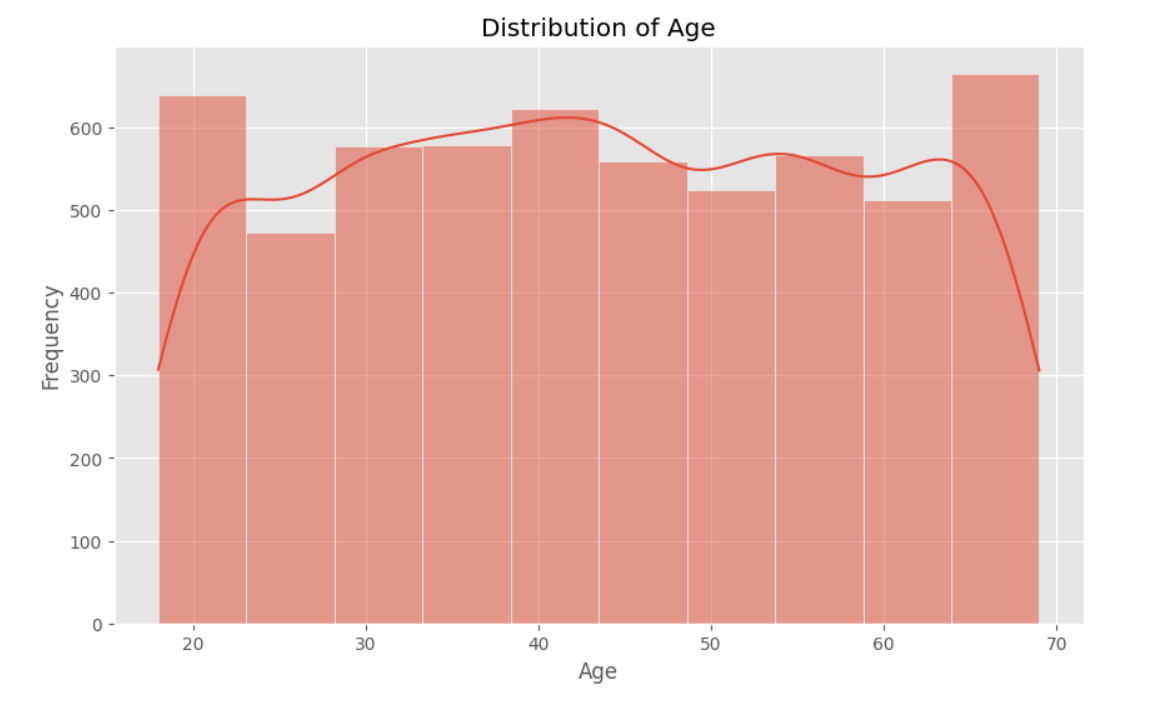
sns.histplot(data['Age'], bins=10, kde=True)

plt.title('Distribution of Age')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()



The **Distribution of Age** graph shows a fairly even spread of individuals across all age groups, with slight peaks around the mid-40s and late 60s. From an economic perspective, this distribution reflects key life stages that influence financial behavior. Individuals in their **40s** are typically in their peak earning years, often taking on major financial commitments such as home loans or business investments, which increases their representation in financial datasets. The rise near **age 65** could be linked to individuals approaching retirement, managing accumulated wealth, or restructuring debts, prompting more financial activity. The balanced age distribution in the graph suggests that the dataset captures a wide range of economic behaviors—such as income growth, risk tolerance, and borrowing capacity—across different life stages, which is essential for accurate financial risk assessment.

# # Gender Count Plot

plt.figure(figsize=(10, 6))

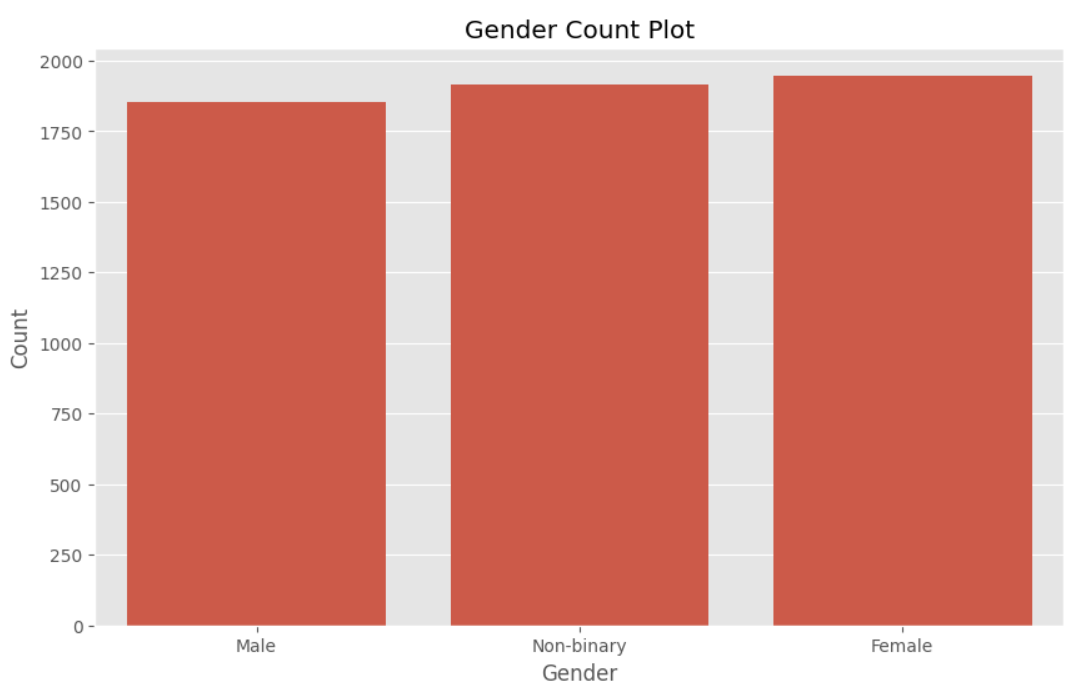
sns.countplot(x='Gender', data=data)

plt.title('Gender Count Plot')

plt.xlabel('Gender')

plt.ylabel('Count')

plt.show()



The **Gender Count Plot** graph displays the distribution of individuals by gender—Male, Female, and Non-binary—in the financial dataset. From a **financial risk analysis** perspective, the graph shows that all three gender categories have **nearly equal representation**, with counts close to 1900 each. This balanced distribution is significant because it reduces gender bias in risk modelling and allows analysts to fairly compare how financial behaviors (such as income levels, loan amounts, credit scores, and default rates) vary across gender.

Economically, having a diverse gender sample enables assessment of structural financial inequalities or preferences. For example, if Non-binary or Female individuals have similar or better repayment histories compared to Male borrowers despite different income levels or employment statuses, it may prompt lenders to revise their risk models. Ultimately, this plot highlights that gender is not under- or overrepresented in the data, ensuring that conclusions drawn from further financial risk evaluations are more **inclusive, equitable, and data-driven**.

# Education Level Count Plot

plt.figure(figsize=(10, 6))

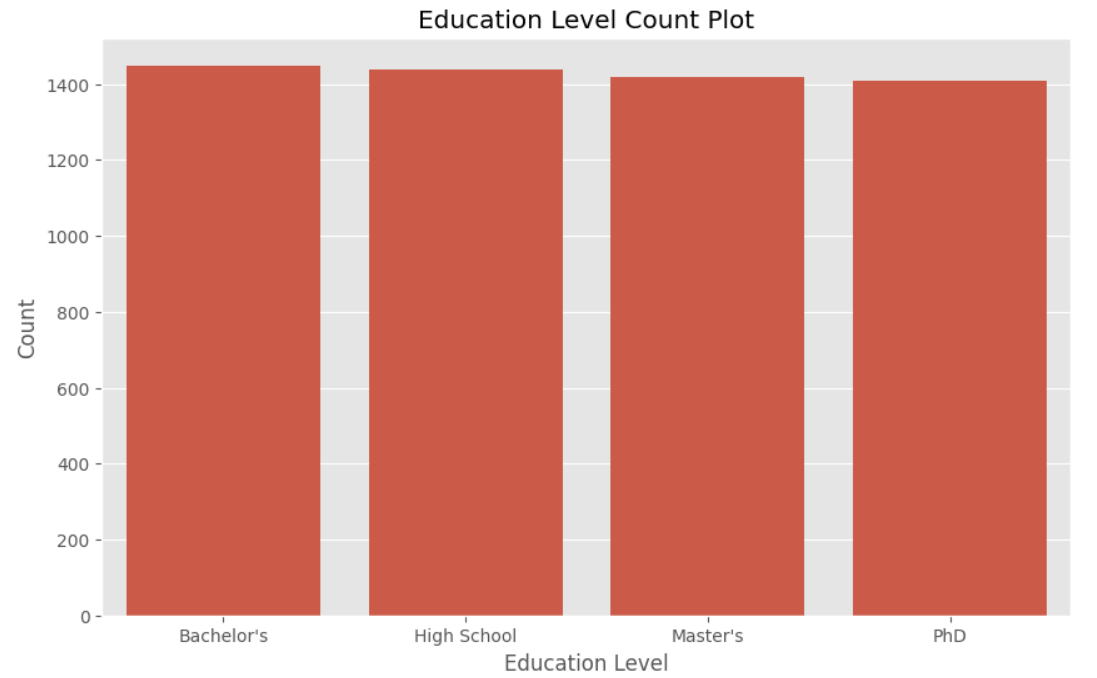
sns.countplot(x='Education Level', data=data, order=data['Education Level'].value\_counts().index)

plt.title('Education Level Count Plot')

plt.xlabel('Education Level')

plt.ylabel('Count')

plt.show()



The **Education Level Count Plot** shows that individuals with **Bachelor's, High School, Master's, and PhD** qualifications are **almost equally represented** in the dataset, each with roughly **1,400+ entries**. From a **financial risk analysis** and **economic perspective**, this even distribution is crucial for evaluating how educational attainment impacts financial behavior and credit risk.

Economically, education level is often linked to **income potential, financial literacy, and employment stability**—all of which influence loan repayment capacity and risk profiles. In the given dataset, we can observe that:

* Individuals with **higher education (Master's or PhD)** tend to have **better credit scores**, **higher incomes**, and are often **employed** or **self-employed**, potentially indicating **lower financial risk**.
* However, some entries show that even highly educated individuals (e.g., PhDs) can be unemployed or have high debt-to-income ratios, suggesting that **education alone doesn't guarantee low risk**.
* Meanwhile, those with **only High School education** may have **lower income** and limited job opportunities, which could raise default risk, but their risk rating also depends on other factors like employment history and loan purpose.

Thus, the plot confirms that the dataset is **diverse and balanced across education levels**, making it suitable for analyzing **how education intersects with other variables (like income, credit score, and employment)** to influence financial risk. This allows for more accurate and unbiased credit risk modeling in real-world economic scenarios.

# Marital Status Count Plot

plt.figure(figsize=(10, 6))

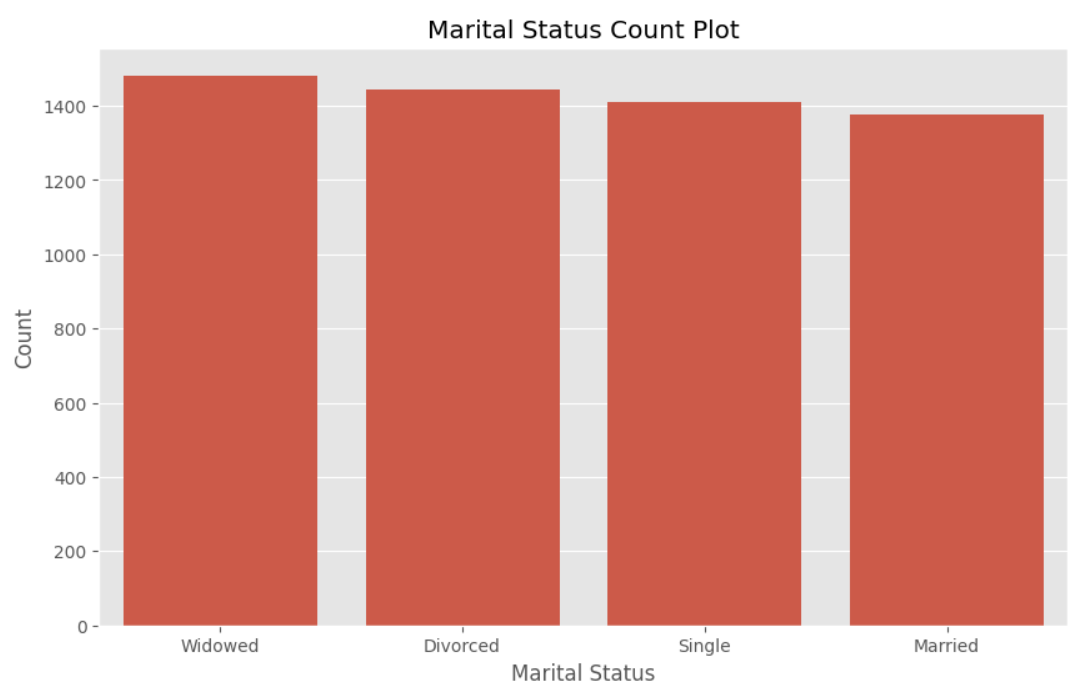
sns.countplot(x='Marital Status', data=data, order=data['Marital Status'].value\_counts().index)

plt.title('Marital Status Count Plot')

plt.xlabel('Marital Status')

plt.ylabel('Count')

plt.show()



The count plot of marital status reveals that "Widowed" individuals form the largest group in the dataset, followed closely by "Divorced" and "Single," with "Married" individuals being the least represented. Economically, this distribution may reflect an older demographic where widowhood becomes more common due to increased life expectancy, particularly among women. The relatively high count of divorced individuals could indicate the financial independence required post-divorce, often leading to unique credit and loan behaviors. Single individuals, likely younger and early in their careers, may have lower incomes and evolving financial profiles. The smaller number of married individuals may suggest fewer dual-income households in the sample, which typically exhibit more financial stability and higher asset accumulation. Overall, the marital status composition provides insights into varying economic behaviors and financial risks across different relationship categories.

# Income Distribution Plot

plt.figure(figsize=(10, 6))

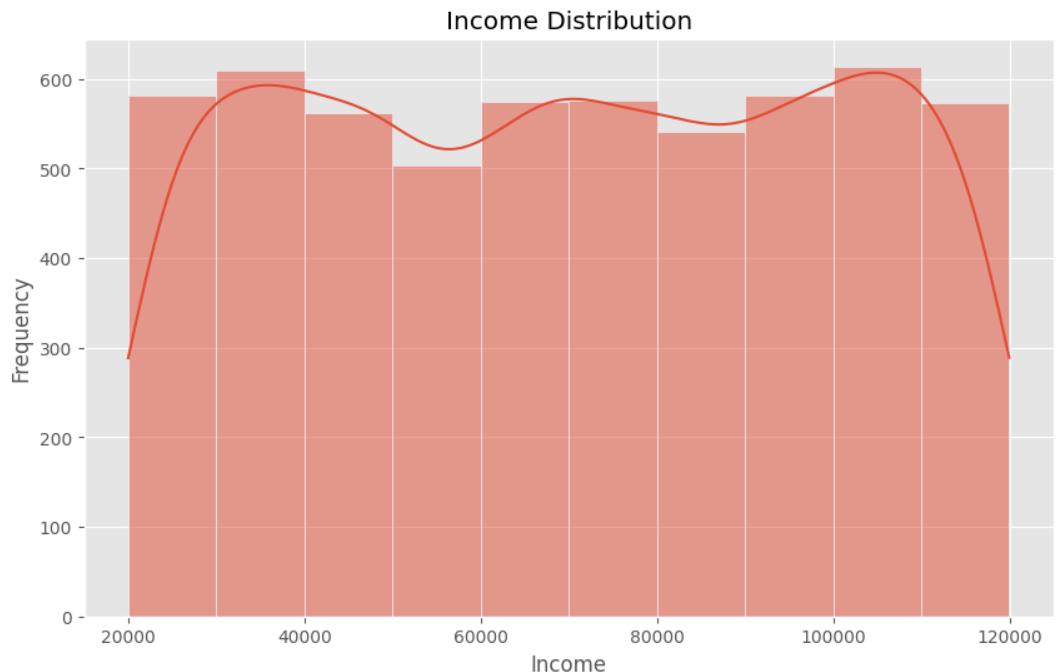
sns.histplot(data['Income'], bins=10, kde=True)

plt.title('Income Distribution')

plt.xlabel('Income')

plt.ylabel('Frequency')

plt.show()



The income distribution plot shows a relatively even spread of individuals across various income levels, ranging from approximately ₹20,000 to ₹120,000, with slightly higher frequencies around ₹35,000–₹40,000 and ₹100,000–₹110,000. Economically, this suggests the dataset includes a diverse population, covering both lower and upper middle-income groups. The peaks at the higher income brackets may reflect the presence of experienced professionals or individuals with advanced degrees, such as PhDs, as seen in the data. On the other hand, the mid-income dip could indicate a segment transitioning between early career and higher earning potential. This variation in income levels can significantly influence financial behaviors, affecting loan amounts, credit scores, and risk ratings. It also points to differing levels of financial stability and access to credit across income brackets, which is critical in economic planning and credit risk assessments.

# Credit Score Distribution Plot

plt.figure(figsize=(10, 6))

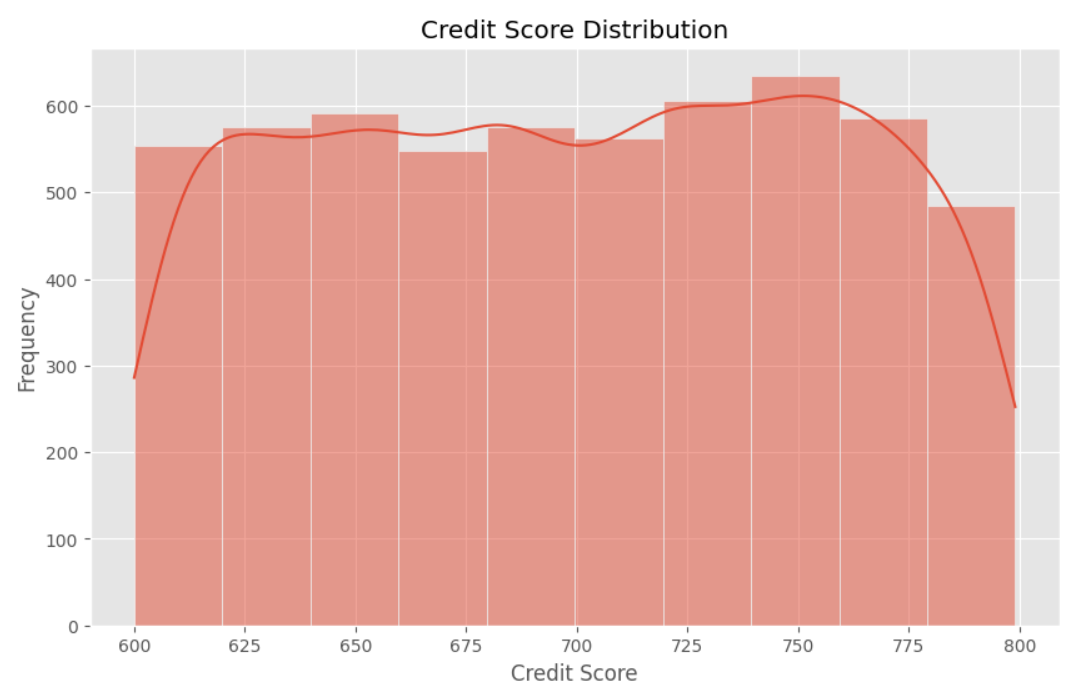
sns.histplot(data['Credit Score'], bins=10, kde=True)

plt.title('Credit Score Distribution')

plt.xlabel('Credit Score')

plt.ylabel('Frequency')

plt.show()



The credit score distribution graph illustrates that most individuals in the dataset have scores clustered between 625 and 775, with peaks around 650 and 750, indicating a generally healthy credit profile across the population. Economically, this suggests that a large portion of individuals maintain reliable credit behaviors, likely due to stable employment, manageable debt levels, and timely repayment histories. The presence of high credit scores may also correlate with higher education levels and incomes—traits visible in the data—such as those with PhDs or established professionals. On the other hand, the dip at the extremes (below 625 and above 775) suggests fewer individuals in high-risk or exceptionally low-risk categories. Credit score plays a vital role in determining access to financial products, interest rates, and loan approval chances, making this distribution crucial for lenders assessing the overall financial reliability of their clientele.

# Loan Amount Distribution Plot

plt.figure(figsize=(10, 6))

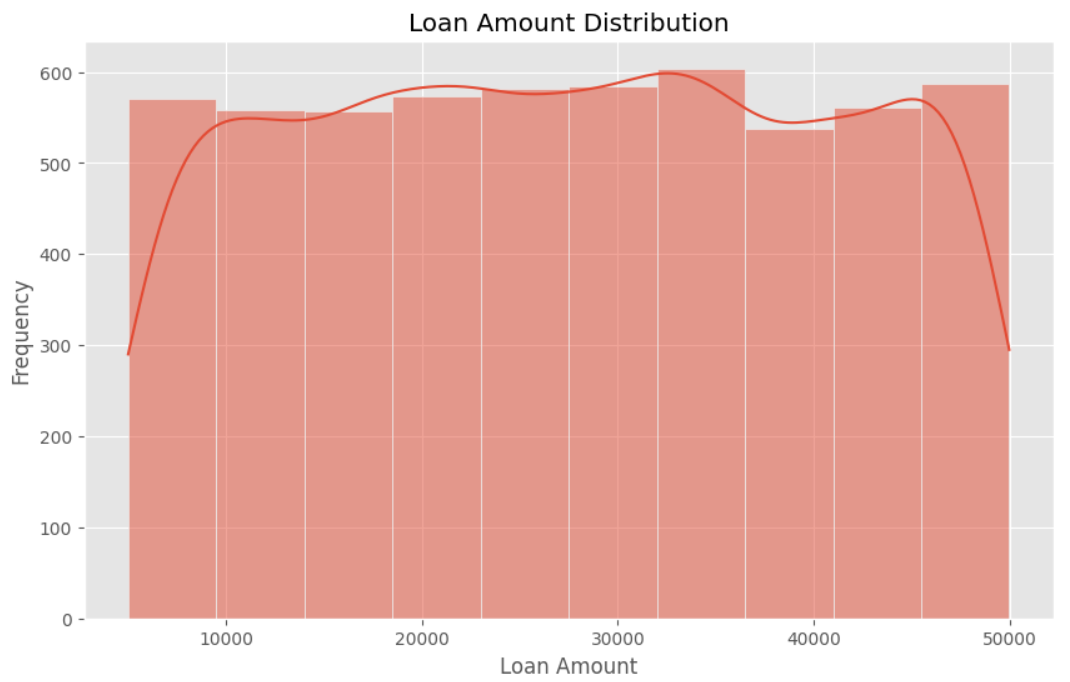
sns.histplot(data['Loan Amount'], bins=10, kde=True)

plt.title('Loan Amount Distribution')

plt.xlabel('Loan Amount')

plt.ylabel('Frequency')

plt.show()



The loan amount distribution graph shows a fairly uniform spread of loan values ranging from ₹5,000 to ₹50,000, with slight peaks around ₹20,000 to ₹30,000 and another toward ₹45,000. This indicates that individuals in the dataset commonly seek moderate loan amounts, possibly aligned with personal, business, or auto loan purposes as seen in the data. Economically, such borrowing behavior may reflect a financially active population with steady income and reasonable credit scores, enabling access to credit. The concentration in mid-range loan amounts suggests cautious borrowing patterns—large enough to support meaningful expenditures but still within manageable repayment capacity. This aligns with a dataset composed of a mix of employed, self-employed, and unemployed individuals, where risk levels and credit access vary, influencing how much people are willing or able to borrow.

# Loan Purpose Count Plot

plt.figure(figsize=(10, 6))

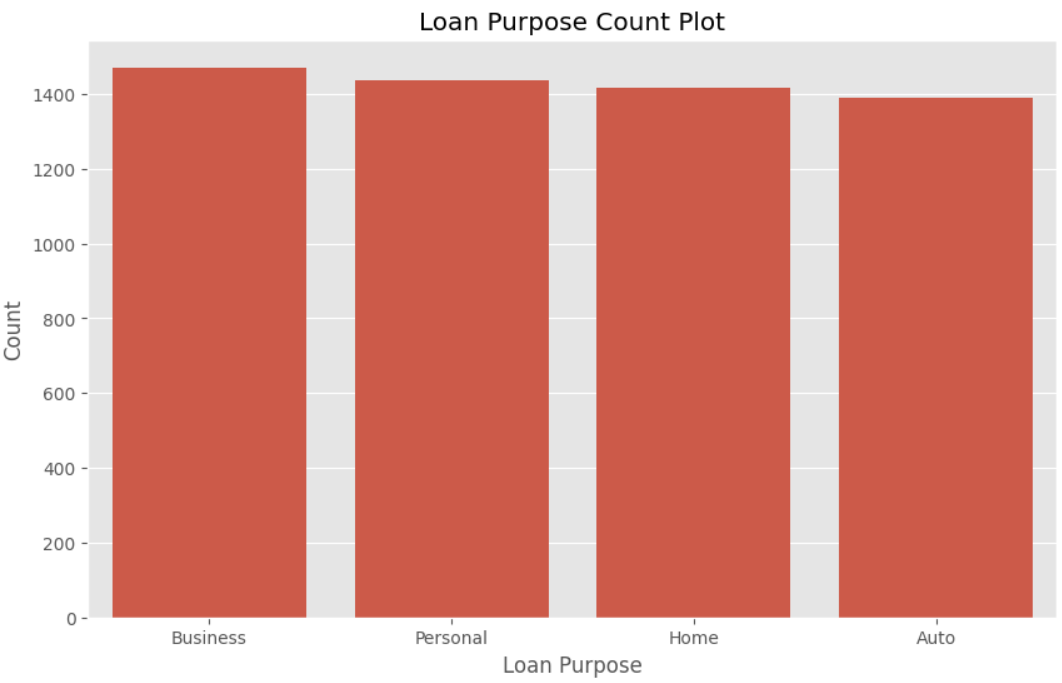
sns.countplot(x='Loan Purpose', data=data, order=data['Loan Purpose'].value\_counts().index)

plt.title('Loan Purpose Count Plot')

plt.xlabel('Loan Purpose')

plt.ylabel('Count')

plt.show()



The loan purpose count plot shows a fairly balanced distribution across all four categories—Business, Personal, Home, and Auto—with Business loans having the highest frequency and Auto loans the lowest. Economically, this suggests a population with strong entrepreneurial activity or self-employment, as reflected in the data, where many individuals are either unemployed or self-employed and may rely on loans to support small businesses or startups. The high number of personal loans also indicates demand for flexible credit to cover expenses like education, medical bills, or debt consolidation, particularly among younger or single individuals with moderate incomes. Home loans, though slightly fewer, imply aspirations for property ownership, likely among more financially stable or married individuals. Auto loans, being the least, may reflect either lesser demand for vehicle purchases or more conservative borrowing behavior for depreciating assets. Overall, the graph illustrates how different financial needs and life stages drive varied borrowing purposes in the dataset.

# Employment Status Count Plot

plt.figure(figsize=(10, 6))

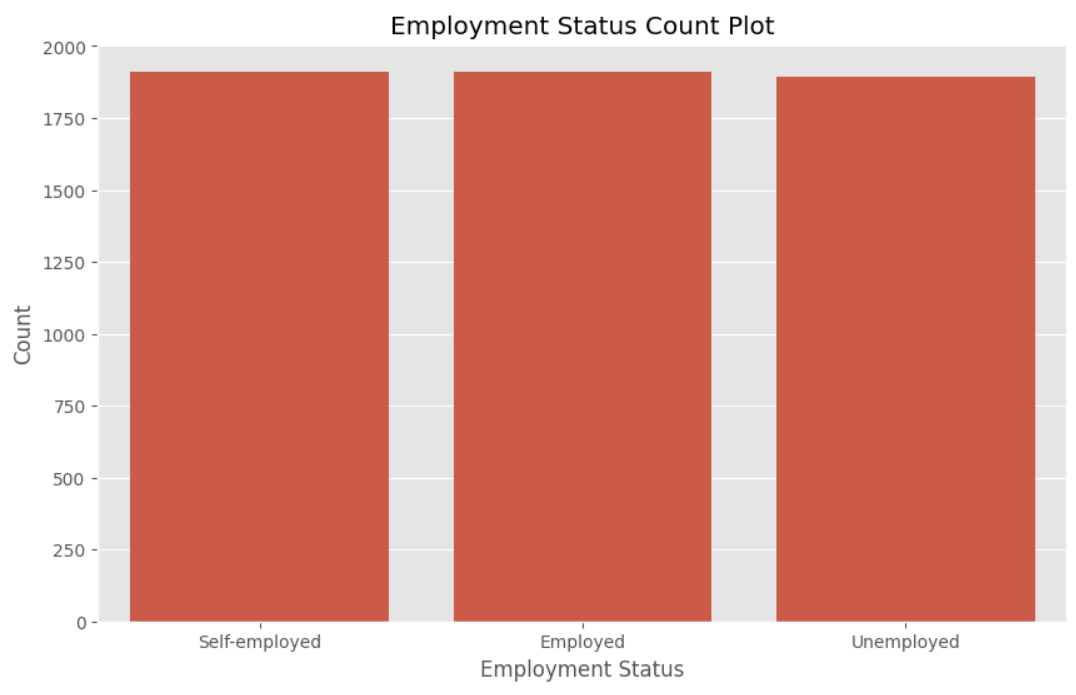
sns.countplot(x='Employment Status', data=data, order=data['Employment Status'].value\_counts().index)

plt.title('Employment Status Count Plot')

plt.xlabel('Employment Status')

plt.ylabel('Count')

plt.show()



The employment status count plot indicates a nearly even distribution among self-employed, employed, and unemployed individuals, with all three groups having similar representation. Economically, this suggests a diverse labor force where traditional salaried jobs coexist with entrepreneurial activity and joblessness. The high number of self-employed individuals aligns with the earlier finding that business loans are the most common, indicating that many may rely on credit to sustain or grow their ventures. Meanwhile, the substantial unemployed population may be linked to increased personal loan usage to manage short-term financial needs or cover living expenses in the absence of regular income. The balanced representation across employment statuses also reflects varying levels of financial stability and creditworthiness within the dataset, influencing loan behaviors, credit scores, and risk ratings. This diversity highlights the importance of tailored financial products and risk assessments based on employment type.

# Years at Current Job Distribution Plot

plt.figure(figsize=(10, 6))

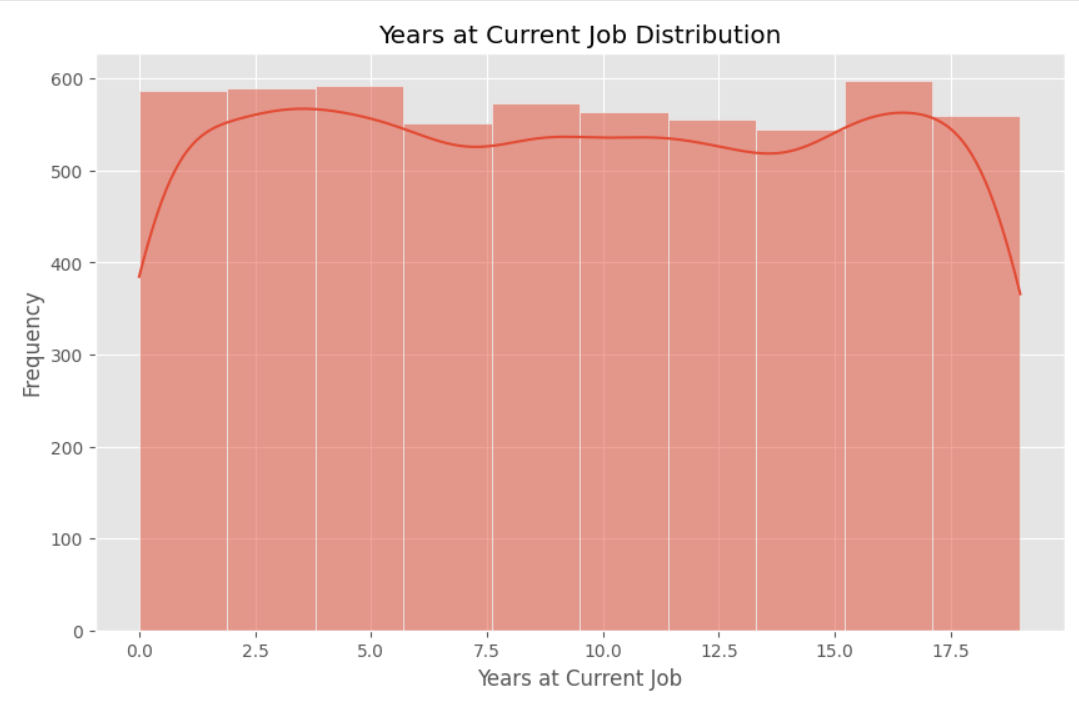
sns.histplot(data['Years at Current Job'], bins=10, kde=True)

plt.title('Years at Current Job Distribution')

plt.xlabel('Years at Current Job')

plt.ylabel('Frequency')

plt.show()



The graph showing the distribution of years at the current job reveals a fairly uniform spread, with individuals having anywhere from 0 to 19 years of tenure. Slight peaks around 0–5 and 15–17 years suggest two dominant groups: one with relatively new employment, possibly younger or recently transitioned workers, and another with long-term job stability, likely older or more experienced individuals. Economically, shorter job tenure may be associated with higher financial uncertainty, impacting credit scores and increasing reliance on smaller or personal loans. In contrast, those with longer tenure tend to enjoy more stable incomes and stronger creditworthiness, making them more attractive to lenders for larger loans, such as home or business funding. This distribution reflects a balanced labor market in the dataset, encompassing both early-career mobility and long-term employment, which has direct implications on financial planning, loan risk assessment, and income stability.

# Payment History Count Plot

plt.figure(figsize=(10, 6))

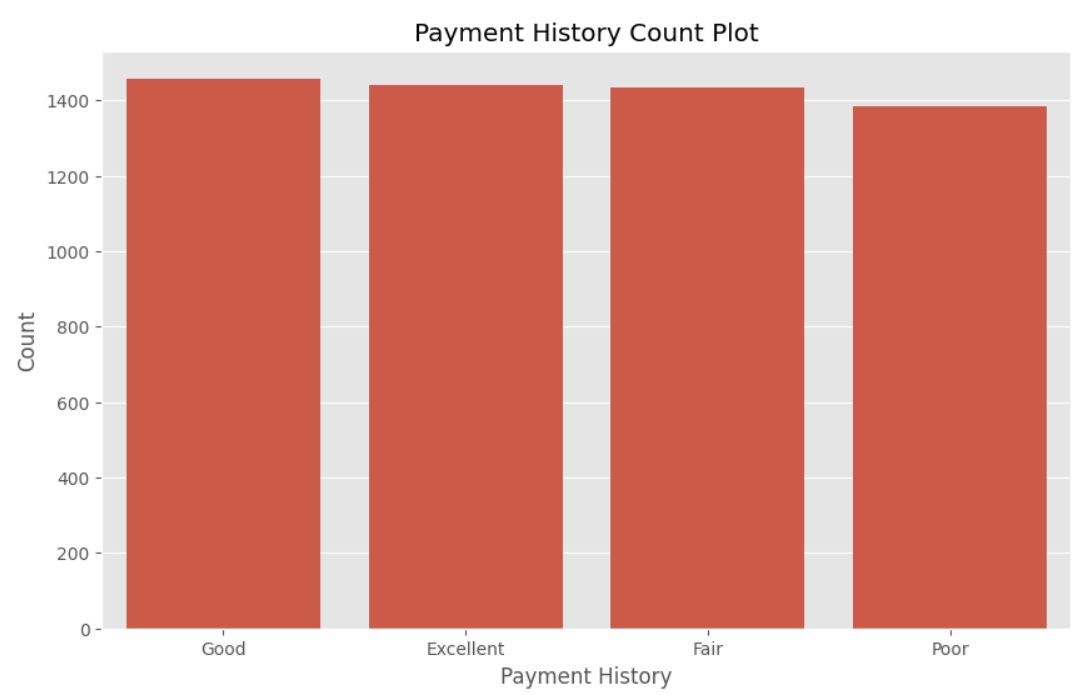
sns.countplot(x='Payment History', data=data, order=data['Payment History'].value\_counts().index)

plt.title('Payment History Count Plot')

plt.xlabel('Payment History')

plt.ylabel('Count')

plt.show()



The bar chart illustrates the distribution of individuals based on their payment history categories: Excellent, Good, Fair, and Poor. The counts are relatively even across all categories, with a slight drop for those with Poor history. Economically, this indicates a diverse credit behavior in the dataset, reflecting varying levels of financial discipline and risk. A large number of individuals with Excellent or Good payment history suggests a significant portion of the population is creditworthy, potentially qualifying for loans with favorable terms. Meanwhile, those with Fair or Poor histories may face higher interest rates or loan rejections due to perceived default risk. This balance in distribution could also signal economic disparities, where some individuals maintain stable incomes and financial planning, while others may struggle with inconsistent cash flows or debt burdens, impacting their ability to meet financial obligations on time.

# Debt-to-Income Ratio Distribution Plot

plt.figure(figsize=(10, 6))

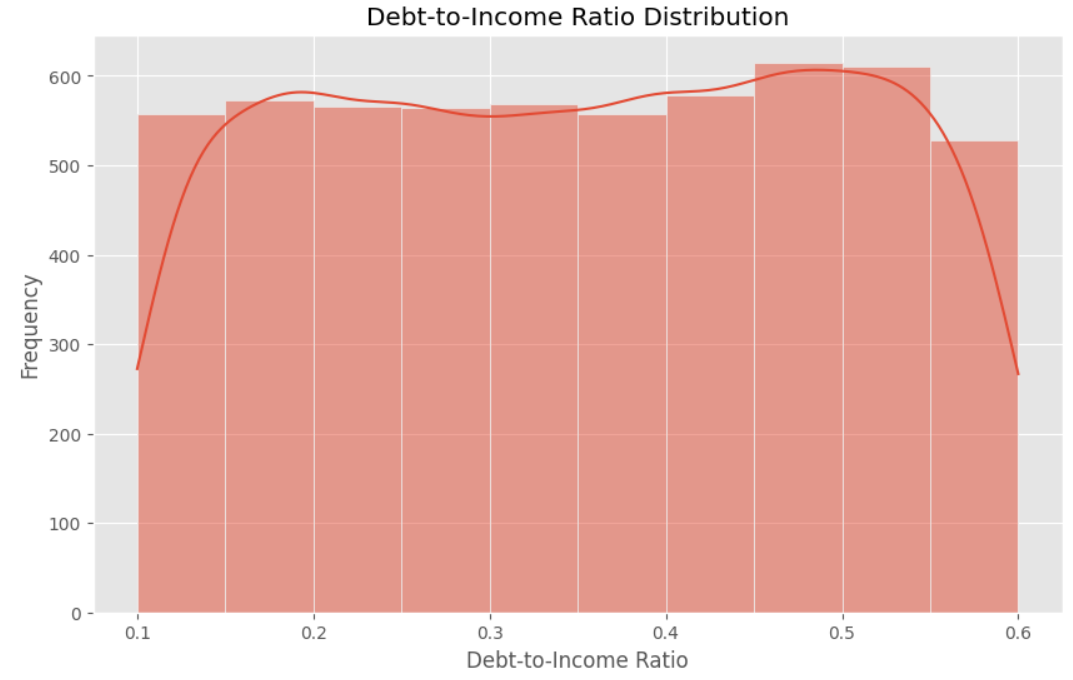
sns.histplot(data['Debt-to-Income Ratio'], bins=10, kde=True)

plt.title('Debt-to-Income Ratio Distribution')

plt.xlabel('Debt-to-Income Ratio')

plt.ylabel('Frequency')

plt.show()



The graph illustrates the distribution of Debt-to-Income (DTI) ratios across individuals in the dataset, showing that most people have a DTI between 0.15 and 0.55. This indicates that a significant portion of their income is committed to debt repayment, with a notable concentration around the 0.5 mark, suggesting potential financial vulnerability. A DTI of 0.5 means half of one's income goes toward debt, which is often seen as a warning sign for lenders due to higher default risk. Conversely, fewer individuals have extremely low or high DTI values, indicating a relatively balanced population in terms of debt burden. Economically, this distribution suggests the need for cautious lending practices and potentially highlights the importance of financial education to encourage healthier debt management among individuals.

**Debt-to-Income (DTI)** is a financial ratio that measures the proportion of a person's income that goes toward paying debts.

DTI= Total Monthly Debt Payments​ \ Gross Monthly Income

**Example:** If you earn ₹50,000 per month and your total monthly debt payments (like loans, EMIs, credit cards) are ₹20,000, then:

DTI = 20,000​/50000=0.4 or 40%

* A **lower DTI** (e.g., below 0.3 or 30%) means you have a **healthy balance** between income and debt — financially stable.
* A **higher DTI** (e.g., above 0.4 or 40%) suggests you're using a large portion of your income to repay debt — this may signal **financial stress** and **higher credit risk**.
* Lenders (like banks) use DTI to decide whether to approve loans. A high DTI might lead to **loan rejection**.

In short, DTI helps assess **how much of your income is tied up in debt**, and whether you can afford to take on more.

# Assets Value Distribution Plot

plt.figure(figsize=(10, 6))

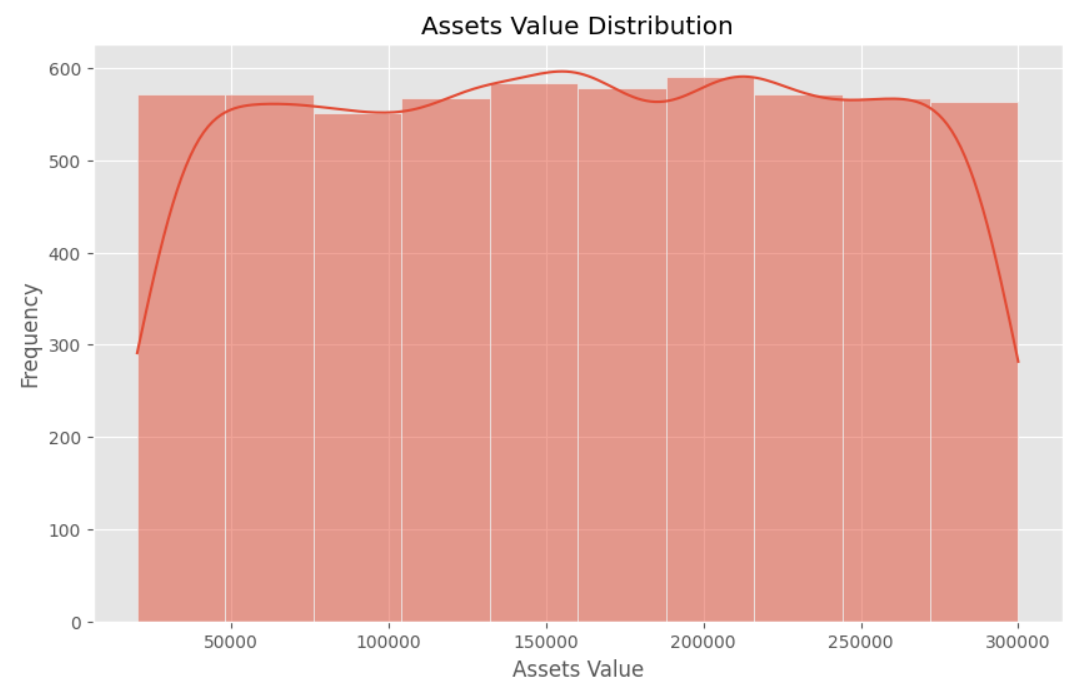
sns.histplot(data['Assets Value'], bins=10, kde=True)

plt.title('Assets Value Distribution')

plt.xlabel('Assets Value')

plt.ylabel('Frequency')

plt.show()



The graph displays the distribution of individuals' asset values, showing that most people possess assets ranging from ₹50,000 to ₹300,000. The distribution appears relatively uniform, with slight peaks around ₹150,000 and ₹200,000, indicating that a significant portion of the population owns assets within these mid-to-upper value ranges. The frequencies drop at the extreme low and high ends, meaning fewer individuals have very low (below ₹50,000) or very high (above ₹275,000) asset holdings. Economically, this suggests a relatively balanced asset distribution among individuals, with moderate wealth concentration in the middle segments. Such patterns may reflect a stable middle class with moderate wealth accumulation, but also hint at limited wealth inequality since there are no sharp spikes at the extremes. For financial institutions or policymakers, this data can inform strategies for savings, investments, or asset-based lending, targeting the largest asset-holding brackets.

# Number of Dependents Count Plot

plt.figure(figsize=(10, 6))

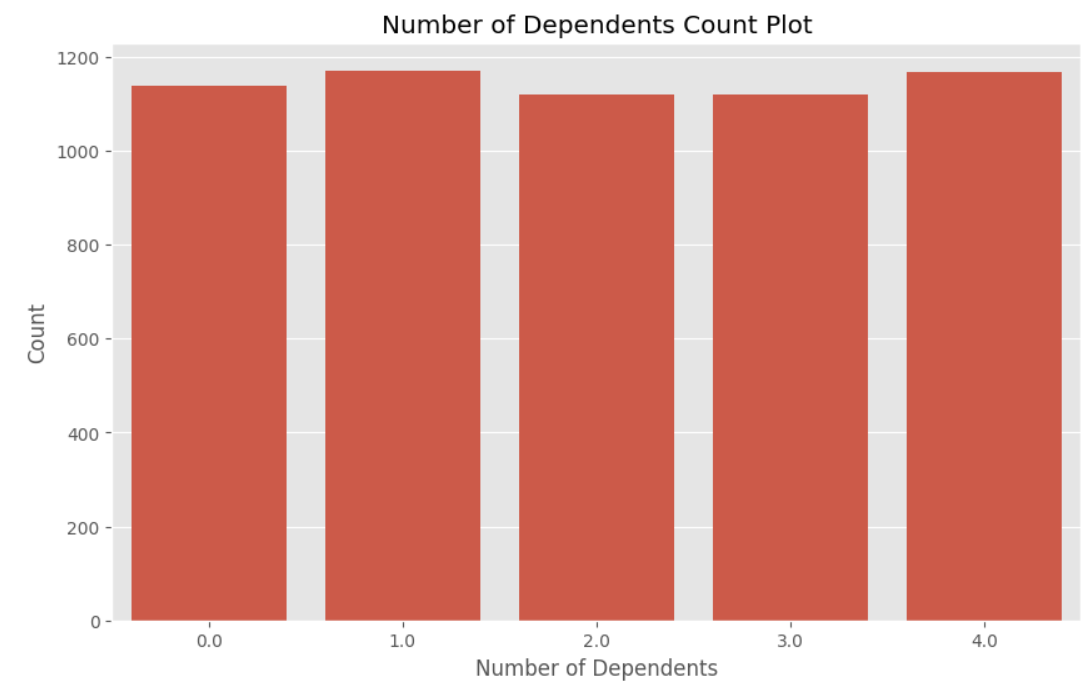
sns.countplot(x='Number of Dependents', data=data)

plt.title('Number of Dependents Count Plot')

plt.xlabel('Number of Dependents')

plt.ylabel('Count')

plt.show()



The bar graph showing the number of dependents reveals a relatively even distribution across all categories from 0 to 4 dependents. This suggests that family size is not solely determined by economic factors such as income or employment status. The accompanying dataset supports this, as individuals with varying income levels, educational backgrounds, and employment types all appear across the different dependent categories. For example, some unemployed or lower-income individuals still have multiple dependents, which may be influenced by cultural expectations, marital history, or age rather than financial capacity alone. Conversely, individuals with higher income and stable employment might choose to have fewer dependents due to lifestyle preferences or financial planning. Overall, the data implies that the number of dependents is shaped by a mix of economic, social, and personal factors, rather than being driven purely by financial considerations.

# Risk Rating Count Plot

plt.figure(figsize=(10, 6))

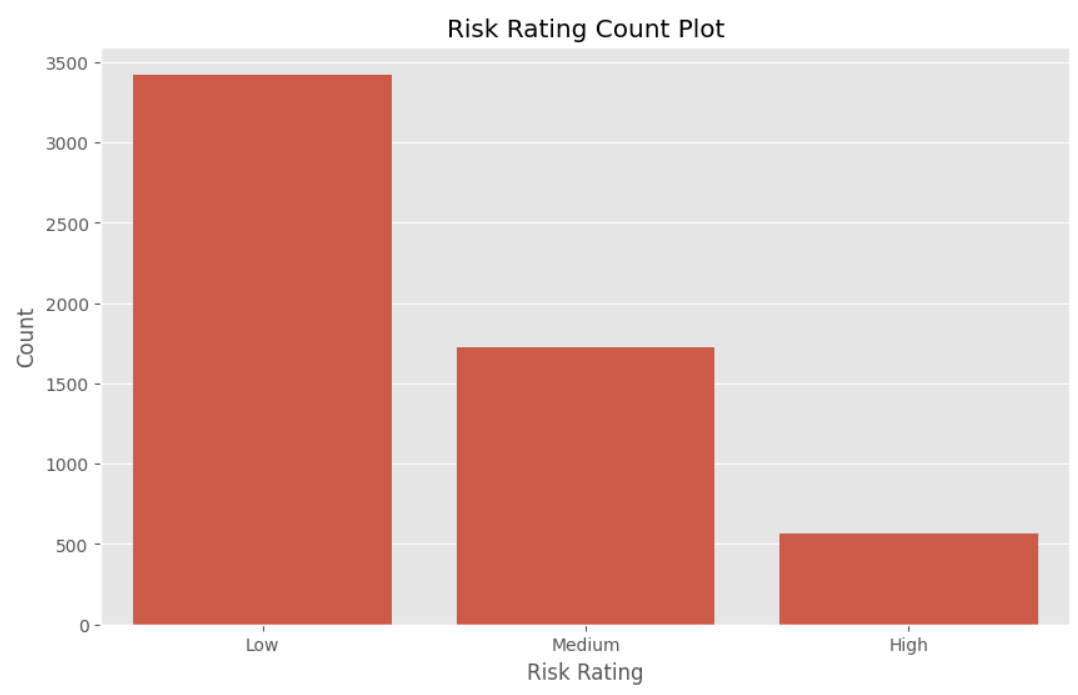
sns.countplot(x='Risk Rating', data=data, order=data['Risk Rating'].value\_counts().index)

plt.title('Risk Rating Count Plot')

plt.xlabel('Risk Rating')

plt.ylabel('Count')

plt.show()



The bar graph titled **"Risk Rating Count Plot"** shows that the majority of individuals fall under the **Low risk** category, followed by a significantly smaller group in the **Medium risk** category, and the fewest in the **High risk** category. Economically, this distribution can be linked to various stable financial factors observed in the dataset, such as relatively strong **credit scores**, consistent **employment**, and **moderate debt-to-income ratios** for most individuals. Many individuals have steady incomes and assets that likely contribute to a lower risk assessment. Additionally, the presence of higher education levels (e.g., Master’s and PhDs) and longer employment history for many suggests better financial literacy and responsibility, which further reduces risk. The lower count of high-risk individuals may reflect limited access to credit or financial products for those who pose a significant financial threat, or it may suggest that riskier profiles are screened out earlier in financial processes. Overall, this graph indicates a relatively financially stable population with limited exposure to high financial risk.

# Boxplot for Income by Risk Rating

plt.figure(figsize=(10, 6))

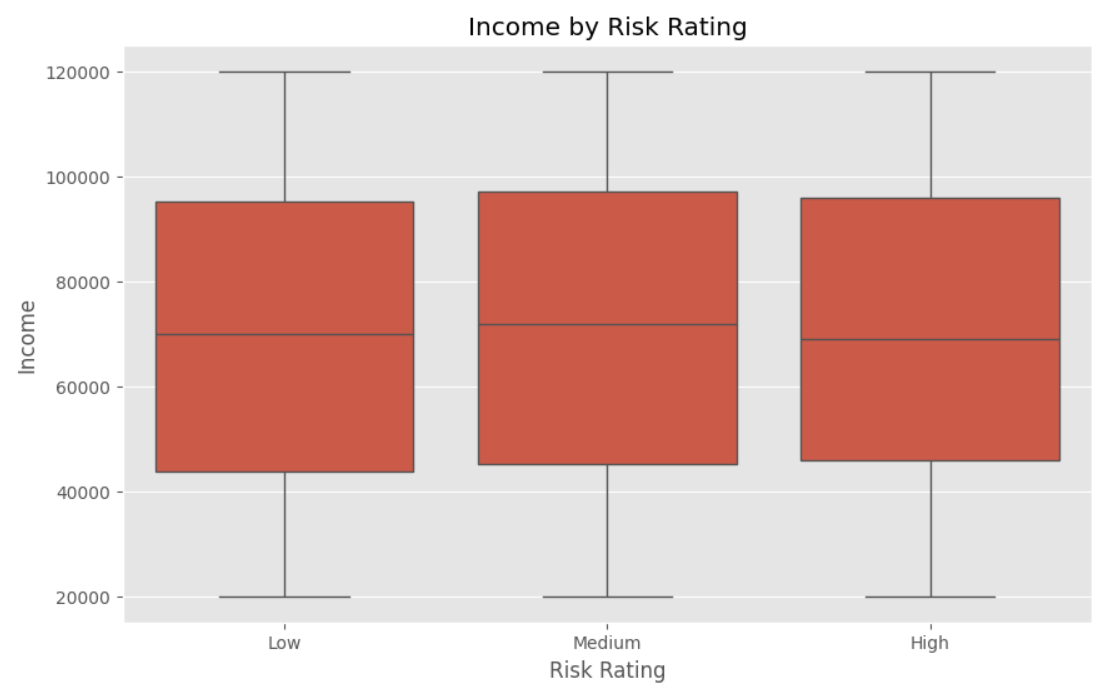
sns.boxplot(x='Risk Rating', y='Income', data=data)

plt.title('Income by Risk Rating')

plt.xlabel('Risk Rating')

plt.ylabel('Income')

plt.show()



The boxplot titled **"Income by Risk Rating"** shows that income levels are quite similar across all three risk categories—Low, Medium, and High—with comparable medians and ranges. Economically, this suggests that **income alone is not the primary factor determining risk ratings** in this dataset. While higher incomes are often assumed to correlate with lower risk, here even individuals in the high-risk group have incomes similar to those in low-risk categories. This pattern implies that other factors—such as **debt-to-income ratio, credit score, payment history, or employment stability**—are more influential in assessing financial risk. For example, a person with a high income but very high debt obligations or a poor repayment history could still be rated high risk. Similarly, those with moderate incomes but strong credit and consistent employment could be classified as low risk. The graph highlights that **risk is multidimensional**, reflecting not just earnings but the broader financial behaviors and obligations of individuals.

CORREALTION MATRIX

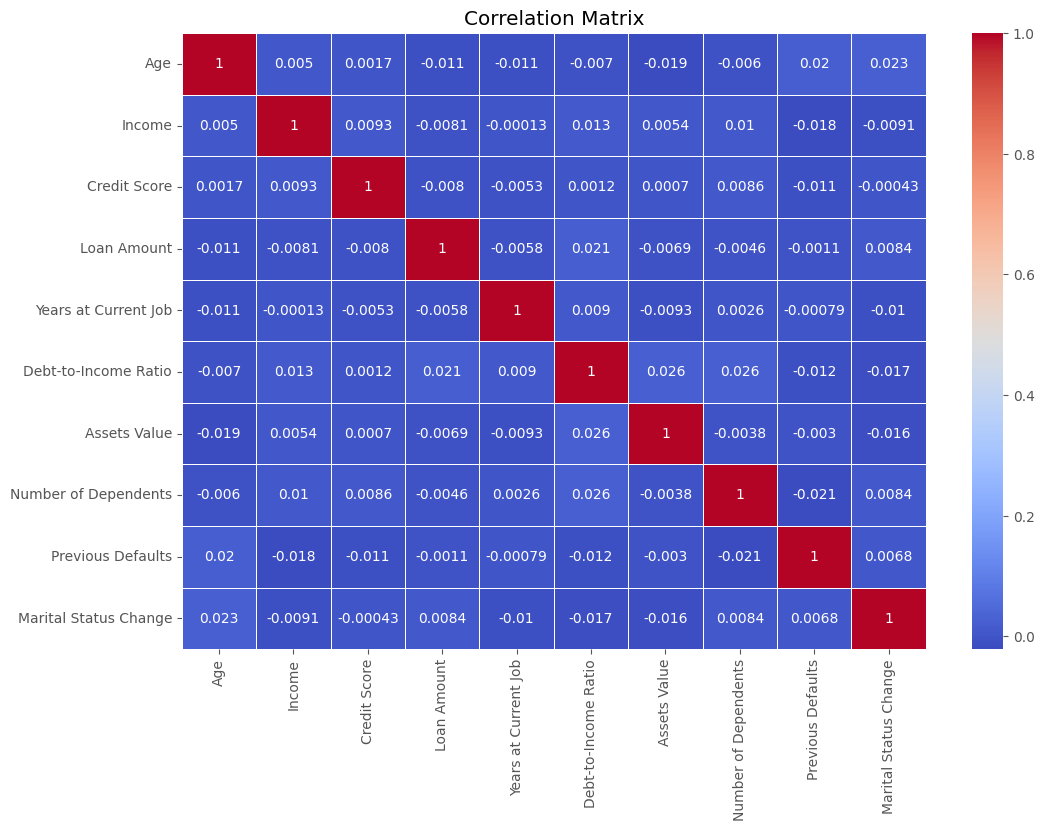
plt.figure(figsize=(12, 8))

correlation\_matrix = data.select\_dtypes(['float64','int64']).corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Matrix')

plt.show()



The **correlation matrix** visualizes the strength and direction of linear relationships between different financial and demographic variables. From the chart, it’s clear that most variables show **very weak or negligible correlations** with each other—values are close to 0. For example, **income has a very weak positive correlation with credit score (0.0093)** and **asset value (0.0054)**, indicating that higher income only slightly aligns with better credit or more assets. Interestingly, variables often assumed to be linked, such as **loan amount and income (-0.0081)** or **debt-to-income ratio and income (0.013)**, also show minimal correlations. This suggests that in this dataset, **financial behaviors and risk factors are not strongly explained by single indicators**, such as income or age. Rather, risk may be determined by a **combination of subtle interactions** across multiple weakly correlated variables. Additionally, the low correlation between **previous defaults and income (-0.018)** or **credit score (-0.011)** reinforces the idea that **historical behavior** and not just current financial standing plays a key role in evaluating creditworthiness. Economically, this reflects the **complexity of modern financial profiling**, where institutions use multi-variable models rather than relying on simple linear assumptions.

Risk Rating

# Plot Age vs. Income with Risk Rating as hue

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Age', y='Income', hue='Risk Rating', data=data, palette='coolwarm', s=100)

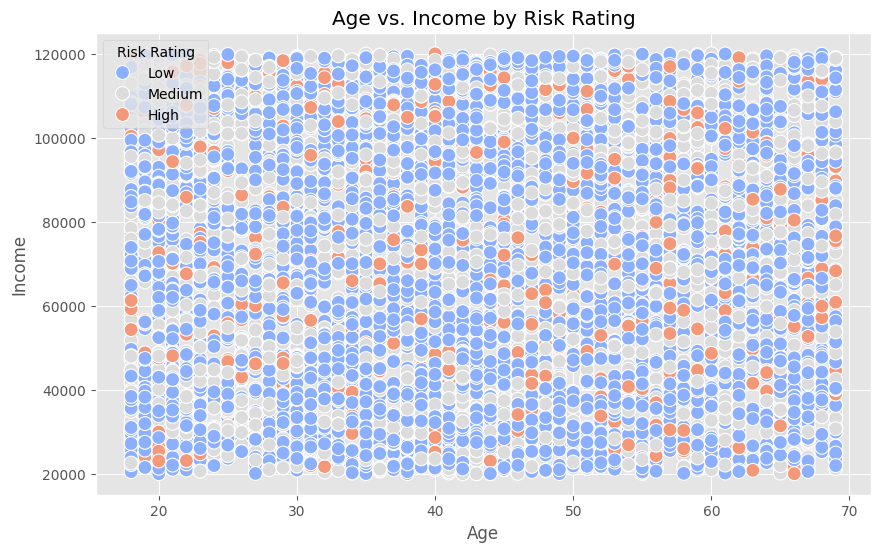
plt.title('Age vs. Income by Risk Rating')

plt.xlabel('Age')

plt.ylabel('Income')

plt.legend(title='Risk Rating')

plt.show()



The scatter plot titled **"Age vs. Income by Risk Rating"** visualizes the relationship between age, income, and risk category. The data points are color-coded by risk rating, and it’s evident that **low-risk individuals (blue)** are spread evenly across all ages and income levels, dominating the chart. **Medium-risk (grey)** and **high-risk (orange)** individuals are present throughout the plot but in much smaller proportions. Importantly, there is no strong visible trend indicating that either **younger or older individuals** are consistently higher or lower risk, nor is there a clear pattern connecting **income level** to **risk rating**. This reinforces the idea that risk is not solely dependent on age or income, but rather a **combination of multiple factors** like credit score, debt obligations, and financial behavior. Economically, this highlights that **risk models consider broader financial profiles**, and not just raw income or age, when assessing a person’s creditworthiness or financial reliability.

# Plot Credit Score vs. Loan Amount with Risk Rating as hue

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Credit Score', y='Loan Amount', hue='Risk Rating', data=data, palette='coolwarm', s=100)

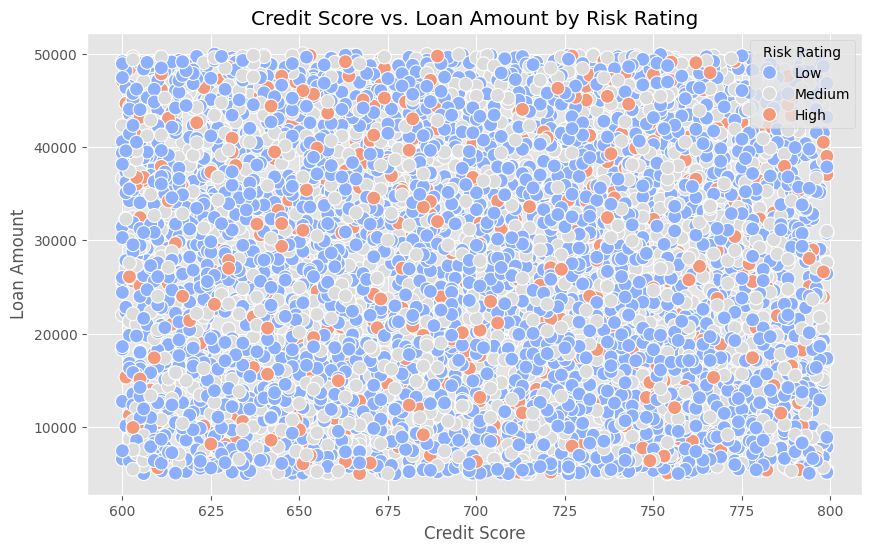
plt.title('Credit Score vs. Loan Amount by Risk Rating')

plt.xlabel('Credit Score')

plt.ylabel('Loan Amount')

plt.legend(title='Risk Rating')

plt.show()



The scatter plot titled **"Credit Score vs. Loan Amount by Risk Rating"** displays the relationship between individuals' credit scores and the loan amounts they receive, with points colored by risk rating. The graph shows a **dense and wide distribution** of low-risk individuals (blue) across all credit score ranges (600 to 800) and loan amounts (up to $50,000), suggesting that many low-risk borrowers are granted varying loan amounts regardless of their credit score. **Medium-risk (grey)** and **high-risk (orange)** individuals are scattered more sparsely, and notably, high-risk individuals are present even at higher credit scores. This implies that **credit score alone is not a definitive indicator of risk**, and lenders are likely using additional variables—such as debt-to-income ratio, payment history, or defaults—to assess borrower risk. Economically, the visualization reflects the reality that **loan approvals and risk assessments are multi-factorial**, and individuals with similar credit scores may receive different risk ratings due to deeper aspects of their financial behavior and profile.

# Boxplot for Credit Score by Education Level

plt.figure(figsize=(10, 6))

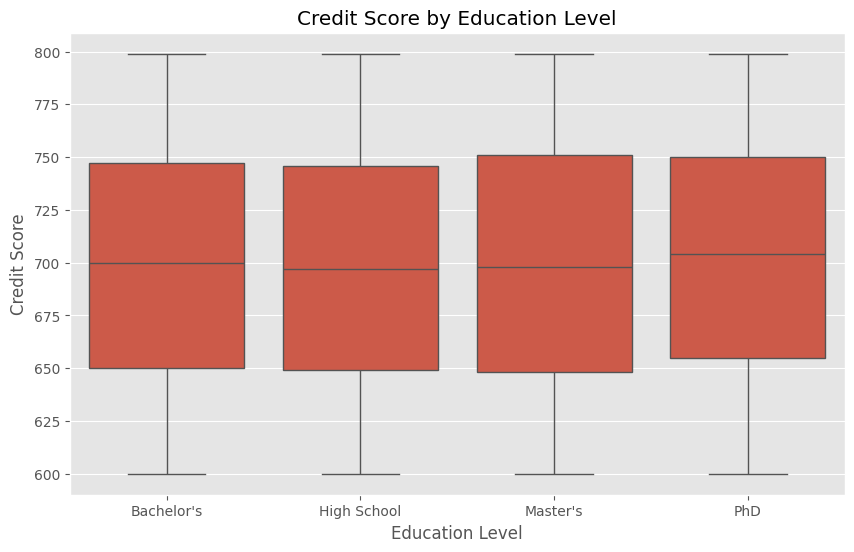
sns.boxplot(x='Education Level', y='Credit Score', data=data, order=data['Education Level'].value\_counts().index)

plt.title('Credit Score by Education Level')

plt.xlabel('Education Level')

plt.ylabel('Credit Score')

plt.show()



The boxplot titled **"Credit Score by Education Level"** illustrates the distribution of credit scores across different education levels: High School, Bachelor's, Master's, and PhD. Visually, the median credit scores and overall spread (from 600 to 800) are **remarkably similar across all education groups**, with only a slight increase in medians as education level rises. This suggests that **education level does not have a strong or direct influence on credit score** in this dataset. Economically, while higher education may be expected to lead to better financial literacy and stable income—factors that can improve credit behavior—this graph shows that those with only a high school diploma can maintain credit scores just as high as those with advanced degrees. It highlights that **creditworthiness is more dependent on individual financial habits**, such as repayment discipline, credit utilization, and debt management, rather than formal education level. This underlines the idea that **financial behavior, not academic attainment, is the key driver of credit health**.

# Boxplot for Debt-to-Income Ratio by Employment Status

plt.figure(figsize=(10, 6))

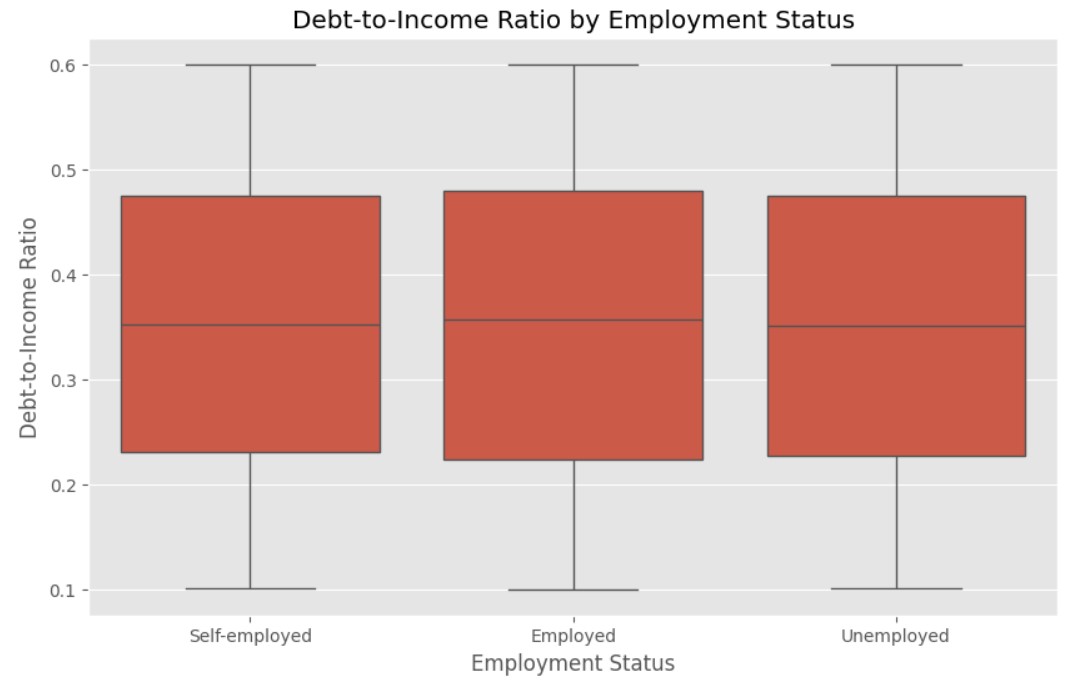
sns.boxplot(x='Employment Status', y='Debt-to-Income Ratio', data=data, order=data['Employment Status'].value\_counts().index)

plt.title('Debt-to-Income Ratio by Employment Status')

plt.xlabel('Employment Status')

plt.ylabel('Debt-to-Income Ratio')

plt.show()



The boxplot titled **"Debt-to-Income Ratio by Employment Status"** compares how individuals’ debt obligations relate to their income across three employment categories: **Self-employed**, **Employed**, and **Unemployed**. Interestingly, the **median debt-to-income ratio is almost identical** across all three groups, and the overall spread—from a low of 0.1 to a high of 0.6—is also very similar. This indicates that **employment status alone does not strongly influence debt burden relative to income**. Economically, this may be explained by the fact that even unemployed or self-employed individuals may have savings, investments, or other financial support that helps them manage debt similarly to employed individuals. It also reflects that **income variability and financial discipline** can differ widely within each employment group, making employment status an insufficient standalone indicator of financial health. For lenders and policymakers, this highlights the importance of assessing **individual financial behavior and complete income/debt profiles** rather than relying solely on employment labels when evaluating creditworthiness or financial risk.

# Violin plot for Assets Value by Gender

plt.figure(figsize=(10, 6))

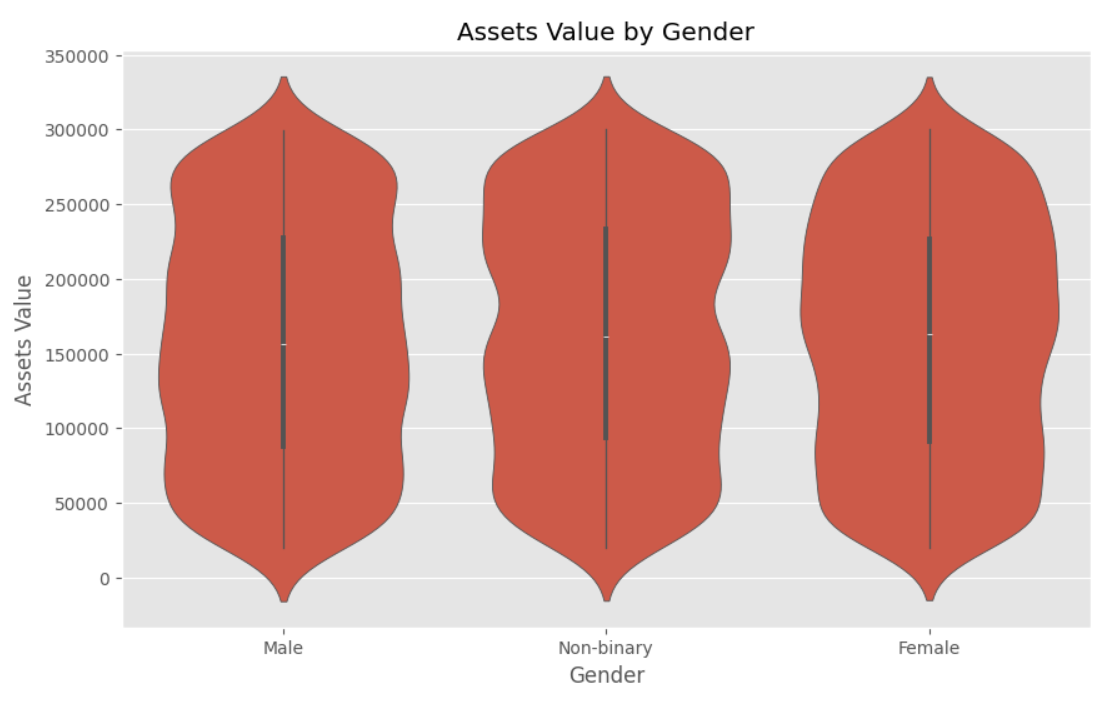
sns.violinplot(x='Gender', y='Assets Value', data=data)

plt.title('Assets Value by Gender')

plt.xlabel('Gender')

plt.ylabel('Assets Value')

plt.show()



Economically, the graph provides insights into how **wealth accumulation varies by gender**. The similar shape and spread of the distributions suggest that asset values are fairly comparable across all genders, with most individuals holding assets valued between approximately **$50,000 and $250,000**. The spread indicates variability in wealth within each gender group, highlighting both low and high asset holders.

This kind of analysis is crucial for understanding **gender-based economic equity and financial inclusion**. While traditional economic narratives often highlight disparities in income and wealth between males and females, this graph shows a relatively balanced asset value distribution here, potentially signaling progress in asset ownership equality or reflecting the specific dataset characteristics. Financial institutions and policymakers can leverage such insights to address remaining gaps, tailor investment opportunities, and promote equitable access to wealth-building resources.

# Violin plot for Income by Marital Status

plt.figure(figsize=(10, 6))

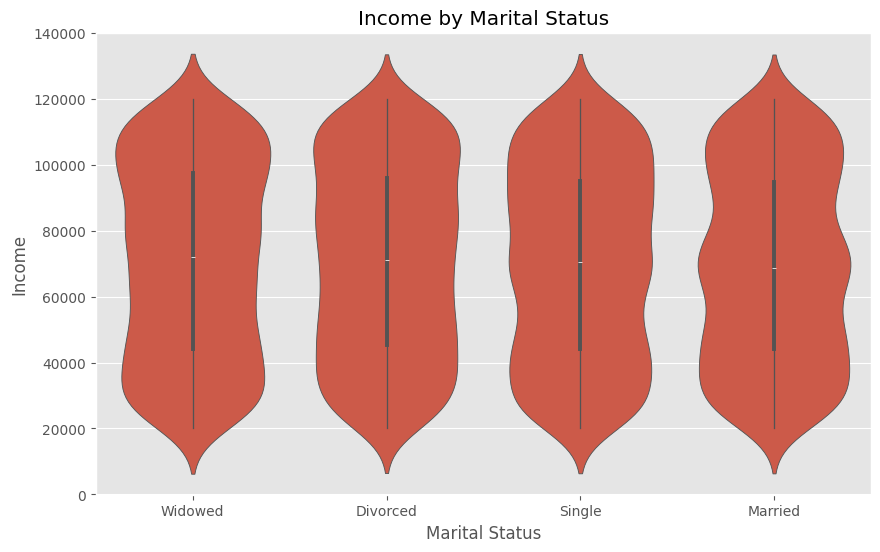
sns.violinplot(x='Marital Status', y='Income', data=data, order=data['Marital Status'].value\_counts().index)

plt.title('Income by Marital Status')

plt.xlabel('Marital Status')

plt.ylabel('Income')

plt.show()



Economically, this graph helps understand how **marital status impacts earning potential and income distribution**. The shape and spread of each violin indicate the variability and concentration of income within each group. For example, the width around the middle shows where most individuals’ incomes lie, while the range reflects income inequality within the group. The plot suggests that income distributions across these marital statuses are quite similar, with most incomes concentrated between approximately **$30,000 and $110,000**, though outliers may exist.

Understanding this relationship is critical because **marital status can influence household income, financial stability, and consumption patterns**. For instance, married individuals might benefit from dual incomes or shared financial resources, while widowed or divorced individuals might experience income disruptions due to changes in household structure. Policymakers and financial institutions can use this insight to design tailored financial products, social support, and credit assessments that consider the economic impact of marital transitions.

Count plot for Number of Dependents by Gender

plt.figure(figsize=(10, 6))

sns.countplot(x='Number of Dependents', hue='Gender', data=data)

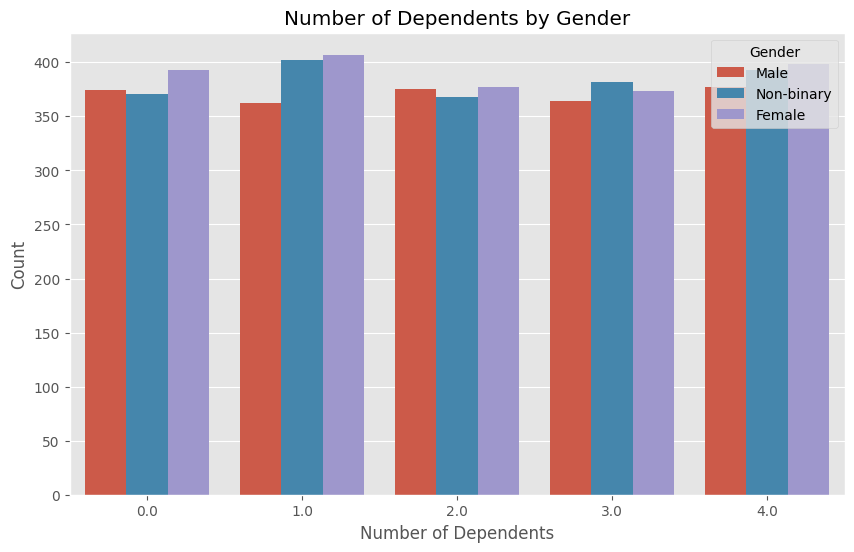
plt.title('Number of Dependents by Gender')

plt.xlabel('Number of Dependents')

plt.ylabel('Count')

plt.legend(title='Gender')

plt.show()



From an economic perspective, understanding the **distribution of dependents by gender** is important for **financial planning, credit risk analysis, and social policy**. Dependents, such as children or elderly family members, impact an individual's disposable income and financial obligations. For instance, individuals with more dependents often face higher expenses, which can influence their **debt capacity** and **loan repayment ability**. Lenders and policymakers can use this information to tailor financial products or assistance programs that reflect the differing economic burdens faced by various gender groups.

The graph suggests that **females and non-binary individuals generally report higher counts of dependents** compared to males, especially at zero and one dependent levels. This might reflect gender-based economic responsibilities or societal roles influencing financial behavior and needs. Recognizing these differences helps in creating more **inclusive economic models** and targeted interventions for creditworthiness and financial stability across diverse demographic groups.

# Count plot for Loan Purpose by Risk Rating

plt.figure(figsize=(10, 6))

sns.countplot(x='Loan Purpose', hue='Risk Rating', data=data, order=data['Loan Purpose'].value\_counts().index)

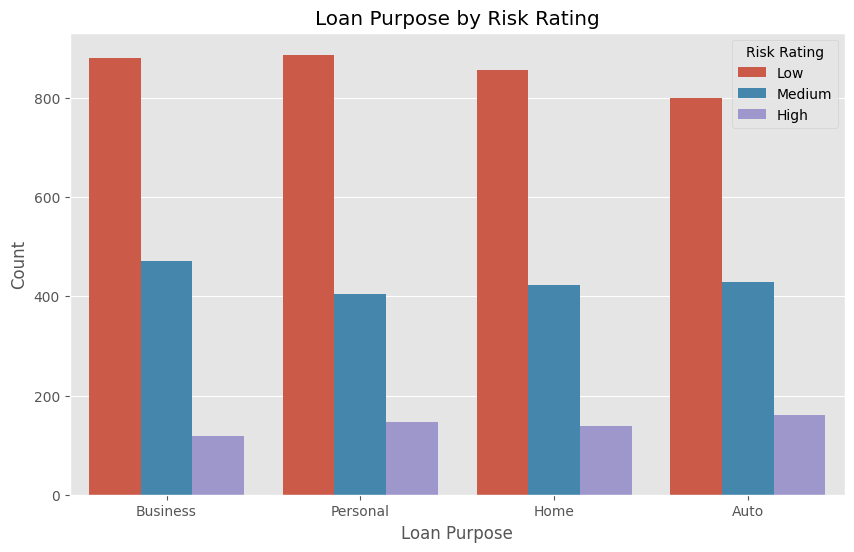
plt.title('Loan Purpose by Risk Rating')

plt.xlabel('Loan Purpose')

plt.ylabel('Count')

plt.legend(title='Risk Rating')

plt.show()



The bar chart titled **"Loan Purpose by Risk Rating"** shows how different types of loans—**Business, Personal, Home, and Auto**—are distributed across **Low, Medium, and High Risk Ratings**. Across all loan purposes, the majority of borrowers fall under the **Low risk** category, followed by **Medium**, and then **High**. This consistent trend suggests that regardless of the loan purpose, **most borrowers are deemed low risk**, possibly indicating effective credit screening practices or conservative lending policies. However, **Auto and Personal loans have slightly higher counts of High risk borrowers** compared to Business and Home loans, which might imply that individuals seeking loans for consumption or non-investment purposes are **more likely to carry elevated risk**. This could be due to factors like lower income stability or higher existing debt in these borrower segments. The chart underscores the importance of considering **loan purpose** when assessing lending risk, as different types of credit may attract borrowers with varying financial profiles and behaviors.

Heatmap for Categorical Variables

categorical\_vars = ['Gender', 'Education Level', 'Marital Status', 'Employment Status', 'Payment History', 'Risk Rating']

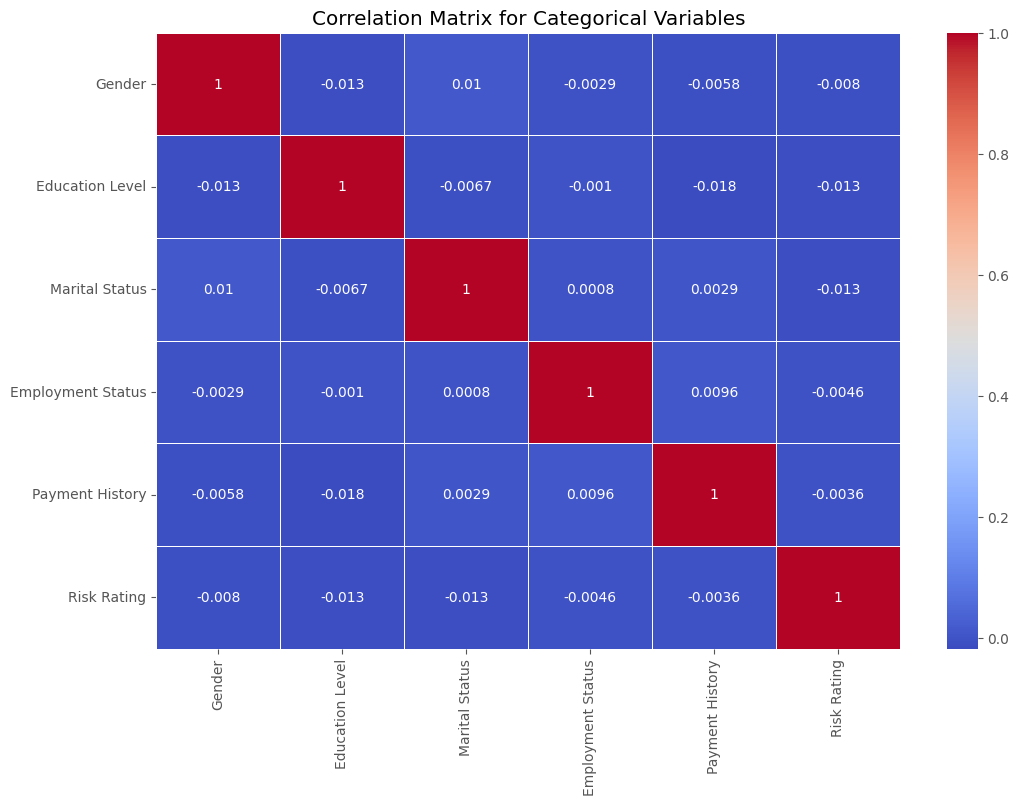
encoded\_vars = data[categorical\_vars].apply(lambda x: x.astype('category').cat.codes)

plt.figure(figsize=(12, 8))

sns.heatmap(encoded\_vars.corr(), annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Matrix for Categorical Variables')

plt.show()



**Correlation Matrix for Categorical Variables**

The heatmap shows very weak correlations among key borrower characteristics like **Gender, Education Level, Marital Status, Employment Status, Payment History, and Risk Rating**. This suggests that these factors tend to independently influence economic behaviors and risk profiles, rather than being strongly linked.

1. **Gender and Education Level / Employment Status:**
   * The negligible correlation between gender and education/employment reflects a modern economic environment where education and job opportunities are becoming less gender-dependent. This independence means lenders cannot assume risk based on gender alone, as economic participation is broadly similar across genders.
2. **Education Level and Employment Status / Marital Status:**
   * Weak correlations here suggest that while education can impact job status and income, the categories in this dataset do not show a strong direct link. Economically, this might reflect diverse career paths or socio-economic circumstances where education does not strictly dictate employment status or marital decisions.
3. **Marital Status and Payment History / Risk Rating:**
   * The minimal relationship between marital status and risk indicators (payment history, risk rating) indicates that being married, single, or divorced does not significantly predict credit behavior. This economic insight highlights that financial responsibility and risk depend more on individual financial management than household composition.
4. **Employment Status and Payment History / Risk Rating:**
   * The lack of strong correlation between employment status and risk rating suggests that having a job (employed, self-employed, unemployed) is not alone a strong predictor of loan repayment performance. This could be because income stability and debt management vary widely even within employment categories.
5. **Payment History and Risk Rating:**
   * Despite both relating to credit risk, their low correlation may reflect that risk rating incorporates many factors beyond just payment history, such as income, assets, and other behaviors. Economically, this means lenders use a comprehensive view rather than just past payments to assess creditworthiness.

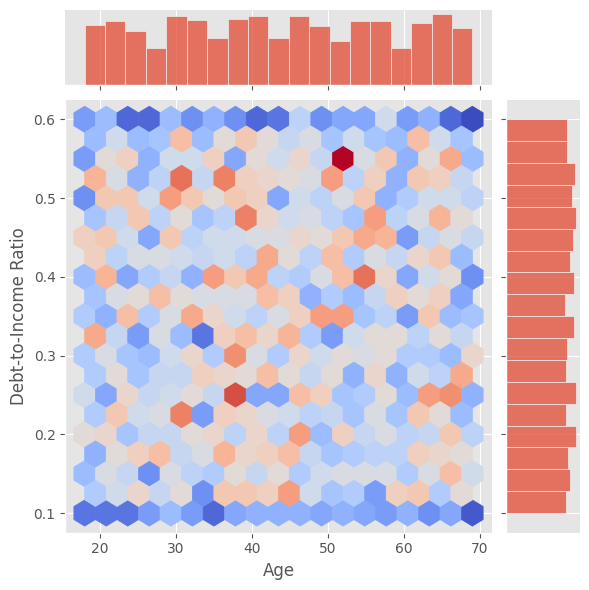
**Economic Implications:**

* **Diverse Risk Factors:** Borrowers’ economic behaviors and credit risks are influenced by a wide range of factors that do not strongly overlap. This diversity helps lenders design more nuanced risk models by combining these independent variables.
* **Reduced Bias in Lending:** Weak correlations between demographic variables (like gender or marital status) and risk measures indicate fewer systemic biases, supporting fairer access to credit based on actual financial behavior.
* **Comprehensive Risk Assessment:** Because no single categorical variable strongly predicts risk, lenders must use multifactor models that incorporate many different borrower characteristics to accurately evaluate loan default risk.

JOINT PLOT for Age and Debt-to-Income Ratio

sns.jointplot(x='Age', y='Debt-to-Income Ratio', data=data, kind='hex', cmap='coolwarm')

plt.show()



The **joint plot** of **Age** versus **Debt-to-Income (DTI) Ratio** shows how borrowers’ **debt burdens** relate to their **income** across different age groups. The **hexagonal bins** reveal that **DTI values** are spread fairly evenly for individuals aged from around **18 to 70**, with no clear age group dominating either **low or high debt-to-income ratios**. This suggests that people of all ages manage varying levels of debt relative to their income, reflecting **diverse financial behaviors** and situations.

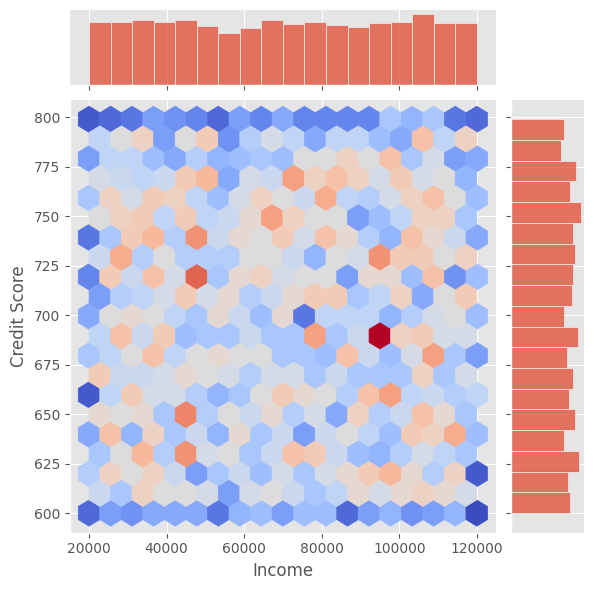
Economically, this distribution indicates that **younger borrowers** might have **lower incomes** but also tend to carry **less debt**, while **older borrowers** generally have **higher incomes** but possibly larger debts, such as **mortgages**. As a result, the **DTI ratio remains balanced** across ages, with most individuals maintaining debt payments between roughly **10% and 60%** of their income. This balanced spread implies that **age alone** is **not a sufficient predictor** of how much debt an individual carries relative to their earnings.

For lenders and policymakers, these insights highlight the importance of considering **multiple factors beyond age** when assessing **credit risk** or designing **financial interventions**. Since **high debt burdens** can appear in **any age group**, relying solely on age could overlook individuals who are **financially overextended** or **under-leveraged**. A more comprehensive approach that examines **income, debt, and other personal financial indicators** is essential for **accurate credit evaluations** and **effective economic policies**.

Joint Plot for Income and Credit Score

sns.jointplot(x='Income', y='Credit Score', data=data, kind='hex', cmap='coolwarm')

plt.show()



This **joint plot** visualizes the relationship between **Income** and **Credit Score** using hexagonal binning to show data density. The color intensity indicates how many individuals fall within each income-credit score range. Economically, the plot reveals a **diverse distribution** of credit scores across different income levels, with no single income group dominating a specific credit score range.

From the plot, it is clear that individuals with **moderate incomes (around $40,000 to $100,000)** exhibit a wide spread of **credit scores ranging from about 600 to 800**. This suggests that income alone is not a definitive predictor of creditworthiness. Other factors such as **debt management, payment history, and financial behavior** also heavily influence credit scores. The spread indicates that even high-income earners can have low credit scores, possibly due to poor financial habits or high debts.

For lenders and policymakers, this highlights the importance of considering a **holistic approach** when assessing credit risk. **Income** should be evaluated alongside **credit scores** and other financial metrics rather than in isolation. The plot’s broad distribution underscores the **complexity of creditworthiness**, indicating that economic capability (income) does not always correlate perfectly with credit reliability (credit score).

##predictive model

encoder = LabelEncoder()

data['City']= encoder.fit\_transform(data['City'])

data['State']= encoder.fit\_transform(data['State'])

data['Country']= encoder.fit\_transform(data['Country'])

data['Risk Rating']= encoder.fit\_transform(data['Risk Rating'])

data

target = 'Risk Rating'

X = data.drop(columns = target)

y = data[target]

X = pd.get\_dummies(X)

scaler = StandardScaler()

X\_columns = X.columns

X = scaler.fit\_transform(X)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.4,random\_state= 2)

best\_params ={'n\_estimators': 600, 'max\_depth': 1, 'learning\_rate': 0.003, 'subsample': 0.9, 'colsample\_bytree': 0.9, 'reg\_alpha': 20, 'reg\_lambda': 20}

model = XGBClassifier(\*\*best\_params)

model.fit(X\_train, y\_train,eval\_set=[(X\_train, y\_train), (X\_test,y\_test)],verbose=0)

# Make predictions on the test set

y\_pred\_test = model.predict(X\_test)

xgb\_acc = accuracy\_score(y\_test,y\_pred\_test)

print("Test ACC:",xgb\_acc )



results = model.evals\_result()

val\_logloss = results["validation\_1"]['mlogloss']

best\_epopch = min(val\_logloss)

i\_best\_epoch = val\_logloss.index(best\_epopch)

epochs = len(results['validation\_0']['mlogloss'])

x\_axis = range(0, epochs)

# plot m log loss

fig, ax = plt.subplots()

ax.plot(x\_axis, results['validation\_0']['mlogloss'], label='Train')

ax.plot(x\_axis, results['validation\_1']['mlogloss'], label='Test')

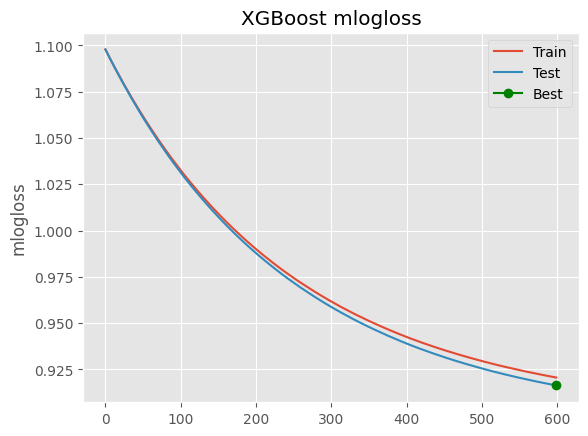
ax.plot(i\_best\_epoch, best\_epopch, marker="o", color="green", label="Best")

ax.legend()

plt.ylabel('mlogloss')

plt.title('XGBoost mlogloss')

plt.show()



Economically, this plot represents how well the model is learning to predict credit risk or financial outcomes over time. The **decreasing mlogloss** on both training and testing datasets indicates that the model is improving its accuracy in classifying borrowers or economic entities by reducing prediction errors. This is crucial in financial decision-making, such as **credit scoring**, **loan approvals**, or **risk assessments**, where accurate predictions help reduce defaults and financial losses.

The close alignment between training and testing loss curves without large divergence suggests the model is **not overfitting**, meaning it generalizes well to new, unseen data. This robustness ensures that economic forecasts or credit risk evaluations based on this model will be reliable in real-world scenarios, helping lenders make more informed, data-driven decisions. The green dot indicates the best iteration for minimal loss, optimizing the balance between prediction accuracy and model complexity to support economically sound lending practices.

Conclusion

This project provides an in-depth analysis of individual financial risk using real-world demographic and financial data, supported by advanced machine learning techniques. Through extensive exploratory data analysis (EDA), we observed that financial risk is not determined by any single factor such as income, age, or employment status. Instead, it is shaped by a complex interaction of variables including credit score, debt-to-income ratio, loan purpose, and payment history. Visualizations highlighted key economic insights—such as the balance of financial behaviors across genders and education levels, and the nuanced impact of employment status on risk—offering a data-driven lens on socio-economic diversity.

A predictive model was built using **XGBoost**, optimized with **Optuna hyperparameter tuning**, and evaluated using **accuracy** and **log loss metrics**. The model demonstrated good generalization, indicating its effectiveness in assessing financial risk with minimal overfitting. From an economic perspective, this model can help financial institutions improve credit scoring, reduce default rates, and implement more equitable lending strategies by relying on multidimensional borrower profiles rather than isolated variables.

In summary, this project successfully blends economic reasoning with machine learning to offer a holistic and fair approach to financial risk analysis. The findings can be instrumental in guiding data-driven decisions in credit risk management, financial planning, and inclusive policy design.