Loan Default Risk Analysis Case Study

1. Objective

To perform Exploratory Data Analysis (EDA) on the dataset to identify the factors influencing loan default risk. The goal is to analyze consumer and loan attributes to predict the likelihood of default.

2. Dataset Description

The dataset contains information about past loan applicants, including whether they defaulted on their loans.

3. Tasks Overview

1. **Load and clean the dataset**.
2. **Perform EDA** to uncover patterns related to loan default.
3. **Visualize the data** to gain insights.
4. **Summarize findings** and prepare the repository for submission.

```
import pandas as pd

# Load the dataset
file_path = 'path_to_your_file/loan.csv'
loan_data = pd.read_csv(file_path)

# Convert percentage strings to numeric values for 'int_rate'
loan_data['int_rate'] = loan_data['int_rate'].str.replace('%', '').astype(float)

# Drop columns with more than 50% missing data
threshold = len(loan_data) * 0.5
loan_data_cleaned = loan_data.dropna(thresh=threshold, axis=1)

# Drop remaining rows with missing values
loan_data_cleaned = loan_data_cleaned.dropna()

# Focus on loans that are either Fully Paid or Charged Off
loan_data_cleaned = loan_data_cleaned[loan_data_cleaned['loan_status'].isin(['Fully Paid',
'Charged Off'])]

# Convert the target variable into binary format (1 for Charged Off, 0 for Fully Paid)
loan_data_cleaned['loan_status_binary'] = loan_data_cleaned['loan_status'].apply(lambda x: 1 if
x == 'Charged Off' else 0)
```

```
loan_data_cleaned.head()
```

Step 2: Exploratory Data Analysis (EDA)

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot distribution of loan amounts
plt.figure(figsize=(10, 6))
sns.histplot(loan_data_cleaned['loan_amnt'], kde=True, bins=30)
plt.title('Distribution of Loan Amounts')
plt.xlabel('Loan Amount')
plt.ylabel('Frequency')
plt.show()

# Plot interest rate distribution
plt.figure(figsize=(10, 6))
sns.histplot(loan_data_cleaned['int_rate'], kde=True, bins=30)
plt.title('Distribution of Interest Rates')
plt.xlabel('Interest Rate (%)')
plt.ylabel('Frequency')
plt.show()

# Boxplot for loan amount vs loan status
plt.figure(figsize=(10, 6))
sns.boxplot(x='loan_status_binary', y='loan_amnt', data=loan_data_cleaned)
plt.title('Loan Amount by Loan Status')
plt.xlabel('Loan Status (0 = Fully Paid, 1 = Charged Off)')
plt.ylabel('Loan Amount')
plt.show()

# Boxplot for interest rate vs loan status
plt.figure(figsize=(10, 6))
sns.boxplot(x='loan_status_binary', y='int_rate', data=loan_data_cleaned)
plt.title('Interest Rate by Loan Status')
plt.xlabel('Loan Status (0 = Fully Paid, 1 = Charged Off)')
plt.ylabel('Interest Rate (%)')
plt.show()
```

Step 3: Insights and Analysis

Based on the visualizations:

- **Loan Amount**: Higher loan amounts tend to be associated with a higher likelihood of default.
- **Interest Rate**: Loans with higher interest rates show a greater tendency towards default.