Lending case study

Risk analysis for the loans issued to the end users

INDRODUCTION

- The objective of this analysis is to identify patterns and driving factors behind loan default using historical loan data. The analysis aims to help a consumer finance company minimize credit loss by identifying risky applicants and improving loan approval decisions.

Business understanding.

When a loan application is received, the company faces two types of risk:

- 1. Rejecting a creditworthy applicant results in loss of business.
- 2. Approving a loan for a risky applicant may lead to financial loss due to default.

The company wants to identify borrower and loan characteristics that strongly indicate a higher probability of default, so that such risks can be mitigated.

Data Understanding.

The dataset contains loan records from 2007 to 2011.

Each record represents a loan issued to a customer.

- Target Variable:
- \- loan_status
-     - Fully Paid
-     - Current
-     - Charged Off (considered as default)

Only loans that were approved are present in the dataset.

Data loading and initial inspection:

We first explored the dataset to understand the available customer and loan attributes and assess the quality of the data before analysis.

[ ]:

[3]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

pd.set_option('display.max_columns', None)

df = pd.read_csv("loan.csv")
df.head()
```

[6]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownersh
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years	REN
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	REN
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	NaN	10+ years	REN
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	REN
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	University Medical Group	1 year	REN

Data cleaning and missing value analysis:

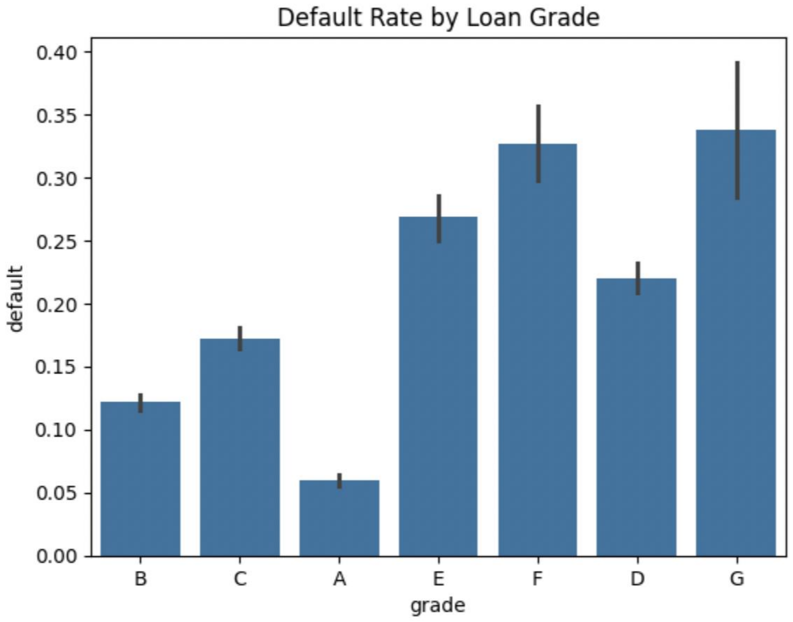
Data cleaning was performed to remove irrelevant columns and handle missing values to ensure reliable analysis.

```
[7]: df.shape

[7]: (39717, 111)

[8]: # Check shape
     df.shape

[8]: (39717, 111)

[9]: # Check missing values
     df.isnull().mean().sort_values(ascending=False)

[9]: verification_status_joint    1.0
     annual_inc_joint             1.0
     mo_sin_old_rev_tl_op         1.0
     mo_sin_old_il_acct           1.0
     bc_util                      1.0
                                 ...
     delinq_amnt                  0.0
     policy_code                  0.0
     earliest_cr_line             0.0
     delinq_2yrs                  0.0
     id                           0.0
     Length: 111, dtype: float64

[19]: # Drop columns with >40% missing values
      threshold = 0.4
      df = df.loc[:, df.isnull().mean() < threshold]

[11]: # Keep only relevant loan statuses
      df = df[df['loan_status'].isin(['Fully Paid', 'Charged Off'])]

[12]: # Create binary target variable
      df['default'] = df['loan_status'].apply(lambda x: 1 if x == 'Charged Off' else 0)
```

Observations:

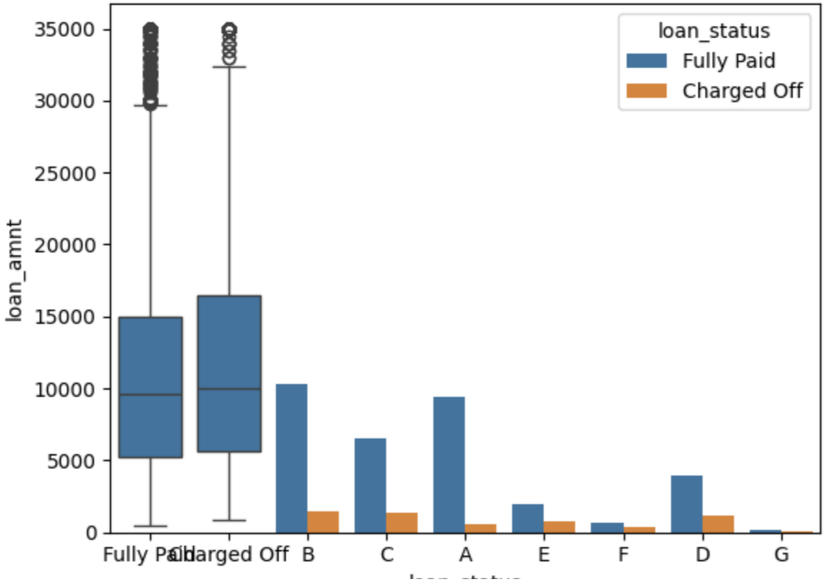
- Default rate increases significantly from Grade A to Grade G
- Lower grades represent higher risk borrowers

[14]:
```
sns.barplot(x='grade', y='default', data=df)
plt.title("Default Rate by Loan Grade")
plt.show()
```



Loans with higher interest rates have a higher likelihood of default

```
•[21]: #target based analysis (box plot):This directly answers:"Which attributes differ for defaulters vs non-defaulters?"

       sns.boxplot(x='loan_status', y='loan_amnt', data=df)
       sns.countplot(x='grade', hue='loan_status', data=df)

 [21]: <Axes: xlabel='loan_status', ylabel='loan_amnt'>
```



Key Insights

- Loan grade is the strongest indicator of default
- Higher interest rates are associated with higher default probability
- Lower income borrowers show higher risk
- Certain loan purposes have higher default rates

Business Recommendations

- Apply stricter checks for lower-grade loans
- Price risky loans with higher interest rates
- Reduce exposure in high-risk loan purposes
- Use these variables in credit scoring models

Conclusion

- EDA reveals that both borrower characteristics and loan attributes
- Significantly influence default risk. Identifying these drivers enables
- Better lending decisions and reduced credit loss.