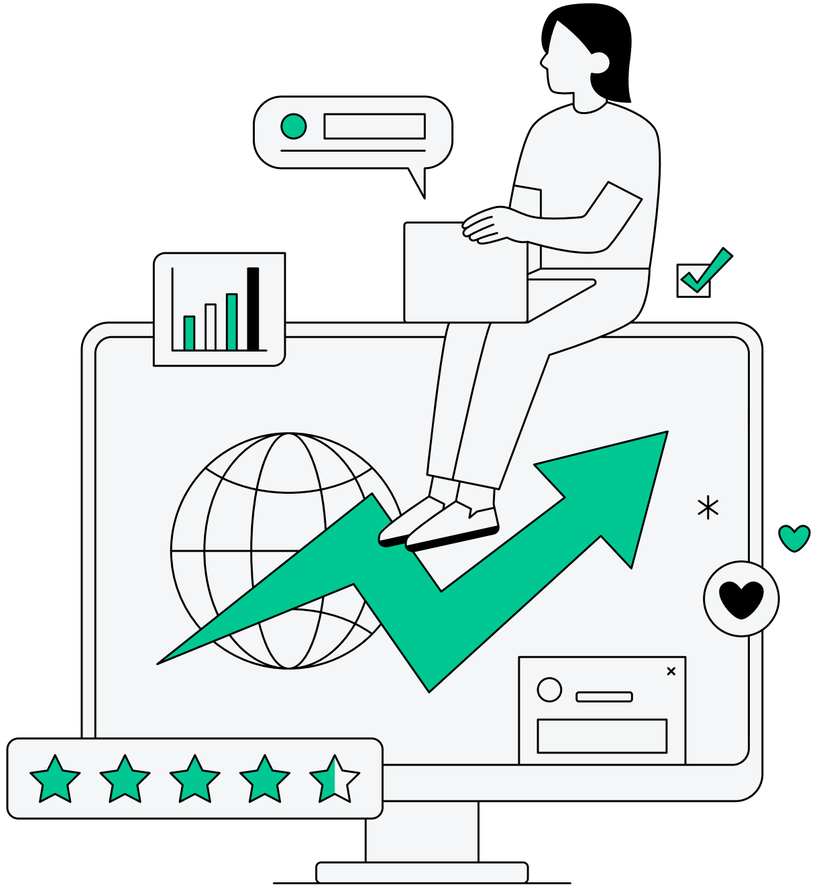
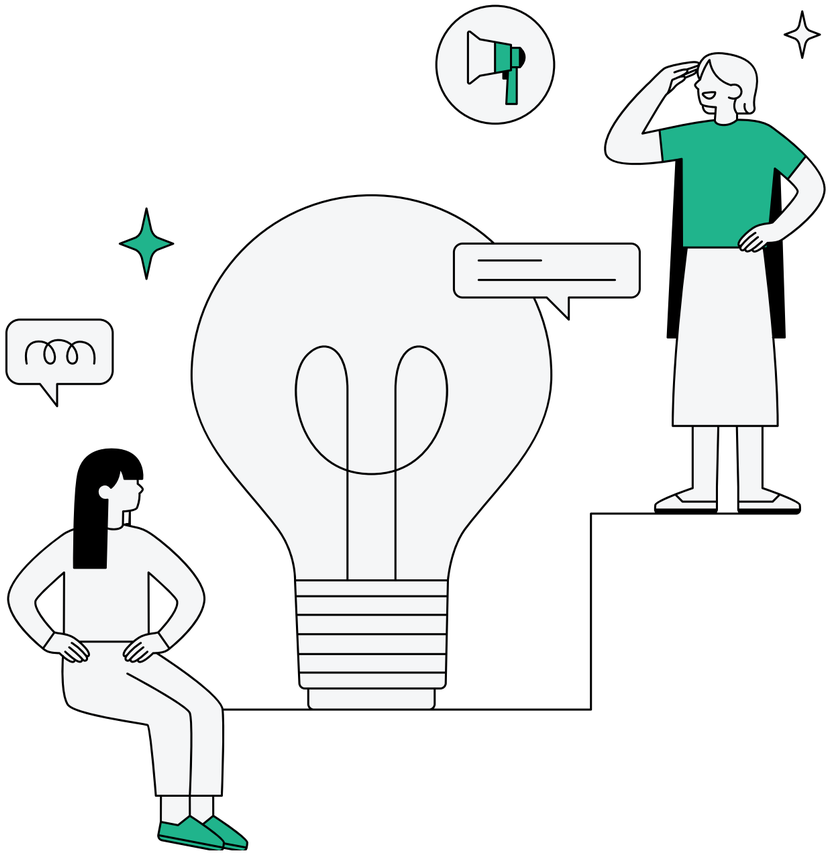
Presented by Dheeraj Salwadi & Devesh Khatri

Credit EDA Case Study

Finding defaulters





Introduction to Problem statement.

We have the loan data from a leading financial institution. The company needs to decide on loan approvals based on applicant profiles. The company aims to recognise defaulters as an incompetent system leads to the following risks:

- Risk 1: Loss of business if loans are not approved for applicants likely to repay.
- Risk 2: Financial loss if loans are approved for applicants likely to default.
-

Objective: Analyze past loan applicant data to identify patterns that predict default risks.

Methodology used in the analysis

01.

Data Understanding

Gathered historical data on loan applicants, including their attributes and understand it aspects.

02.

Data Cleaning

- Handled missing values and inconsistencies.
- Removed duplicate records to ensure data accuracy.

03.

Data Visualization

- Univarite Analysis: Analyzed individual variables
- Bivariate Analysis: Explored relationships between pairs of variables

Main challenges identified

04.
Risk Profiling

- Developed profiles of high-risk applicants based on significant variables.
- Identified patterns and characteristics common among defaulters.

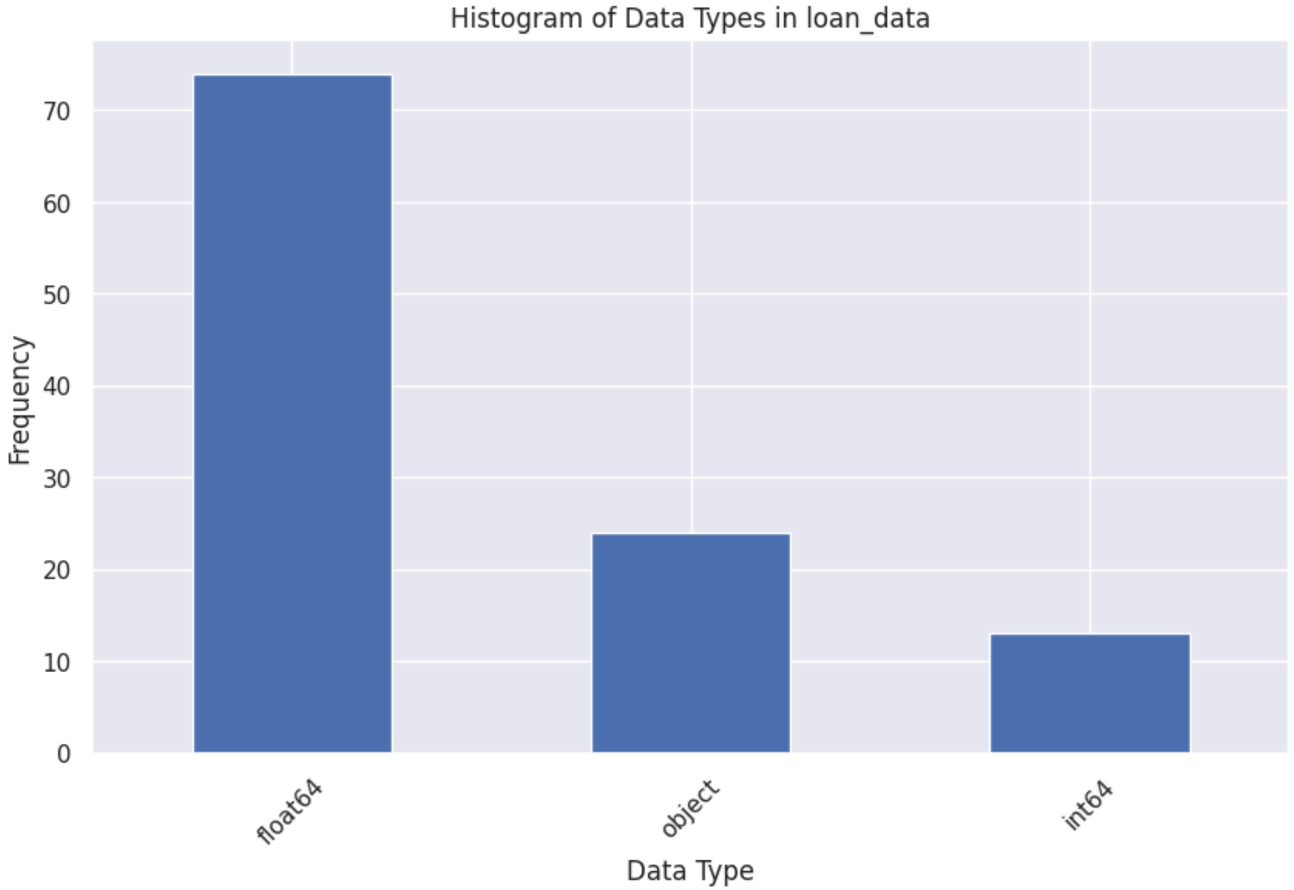
05.
Conclusion

- Summarized findings to draw actionable insights for risk assessment.
- Highlighted key factors driving loan defaults to inform lending decisions.

Challenges and proposed solutions

01. Data Understanding

Gathered historical data on loan applicants, including their attributes and understand it aspects



Observation:

The majority of the Data were present in the Data set were in **Float** type, and the rest of them were **object** and **Int**.

DATA CLEANING

02.

Handled missing values and inconsistencies. Removed duplicate records to ensure data accuracy.

STEP 1

Describe and Identify Null Values

=

```
: #checking the null values

loan_data_null = loan_data.isnull().sum()
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

loan_data_null
```

- Described the data to get an overview of the dataset, including the presence of **null** values.
- Identified columns with **null** values for further cleaning.

STEP 2

Handle High Null Value Columns

1

```
#calculating the null value percentage

loan_data_percentage = round((100*loan_data.isnull().sum()/len(loan_data)), 2)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

loan_data_percentage
```

- Cleaned columns with a significant number of null values by **removing** them.
- Ensured columns with more than **90%** missing data were **dropped**.

2

```
#checking the mean of the columns with null values
loan_data.isnull().mean()
```

Here we tried to check the null data in the data set

3

```
# removing null values more than 90%

loan_data = loan_data.loc[:, loan_data.isnull().mean() <= 0.9]
```

```
#Number of columns left after removal of the null values <= 90%
len(loan_data.columns)
```

```
55
```

Then we dropped all the columns which had more than 90% null data

4

```
#checking the data again

round((100*loan_data.isnull().sum()/len(loan_data)),2)
```

Then we checked the data again to see which columns had null data less 90%

5

```
# removing null values more than 30%, which is the column desc (doesn't relevant data needed) and
# mths_since_last_delinq (this is 64% null so this can be taken out of consideration)

loan_data = loan_data.loc[:, loan_data.isnull().mean() <= 0.3]
```

```
#Number of columns left after removal of the null values <= 30%
len(loan_data.columns)
```

DATA CLEANING

02.

Here we dropped two columns which had more than 30% null data, after investigatin we found that the data were not relevant

STEP 3

Impute Missing Values

- Imputed missing values for **numerical** columns with the **median**.
- Imputed missing values for **categorical** columns with '**unknown**'.

1
```
# List of columns with missing values
columns_with_missing_values = loan_data.columns[loan_data.isnull().any()]
columns_with_missing_values
```

2
```
# Identify categorical and numerical columns
categorical_columns = loan_data.select_dtypes(include=['object']).columns
numerical_columns = loan_data.select_dtypes(include=['float64', 'int64']).columns
```

here we wanted to Identify and find columns which categorical and numerical.

3
```
# Print the columns with missing values, separated by categorical and numerical
print("Categorical columns with missing values:", categorical_missing)
print("Numerical columns with missing values:", numerical_missing)
```

And printed the Data the columns according to the seperated criteria.

DATA CLEANING 02.

Here for an extra point we removed outlier in the Data and verified the Data

4
```
## Removing the percentage signs and convert to float
loan_data['int_rate'] = loan_data['int_rate'].str.rstrip('%').astype(float)

# Verify the changes
loan_data['int_rate'].head()
```

DATA CLEANING

02.

STEP 4

Drop Irrelevant Columns

- **Dropped columns** that were deemed **irrelevant** for **analysis** based on domain knowledge and exploratory analysis.

```
irrelevant_columns = [
    'next_pymnt_d',
    'mths_since_last_record',
    'mths_since_last_delinq',
    'desc',
    'title',
    'initial_list_status',
    'policy_code',
    'url',
    'zip_code',
    'member_id',
    'id'
]
```

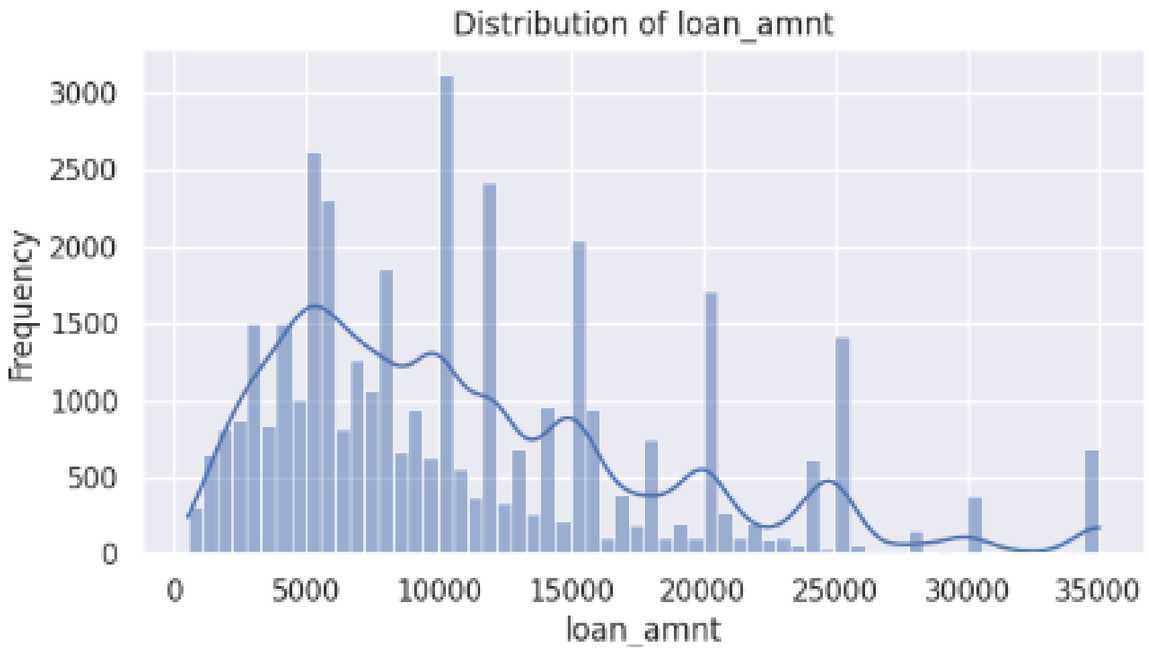
```
x= [cols for cols in irrelevant_columns if cols in loan_data.keys()]
loan_data.drop(columns=x, inplace=True)
```

```
len(loan_data.keys())
```

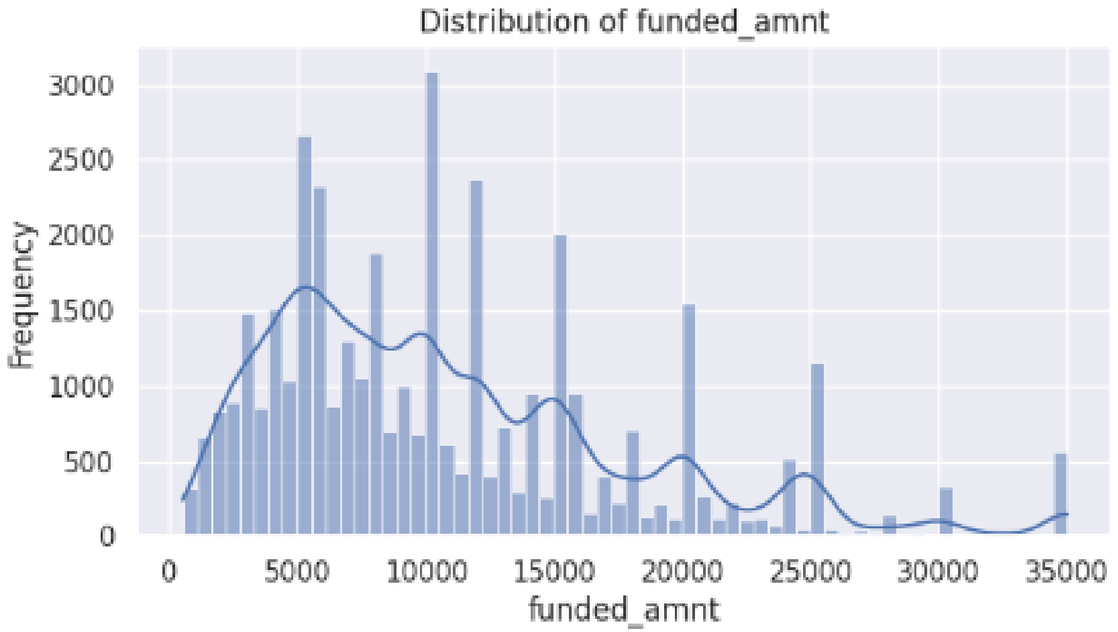
Challenges and proposed solutions

03. Data Visualization

- Univarite Analysis: Analyzed individual variables (numerical values)



Observation: Most of the loans are in the range of $2500 - $5000

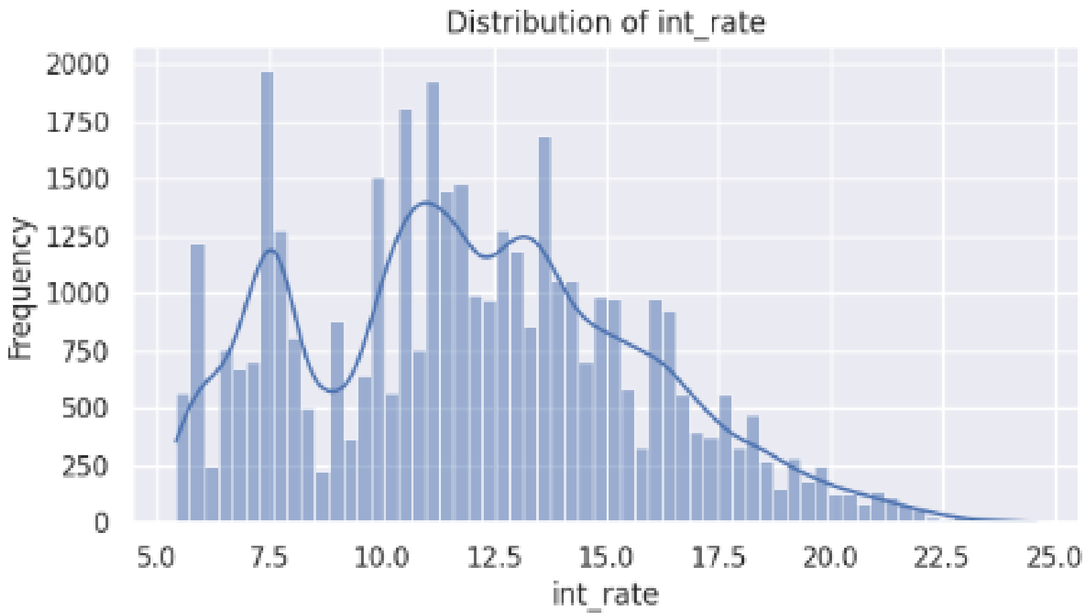


Observation: Most of the funded amounts are in the range of $2500 - $5000

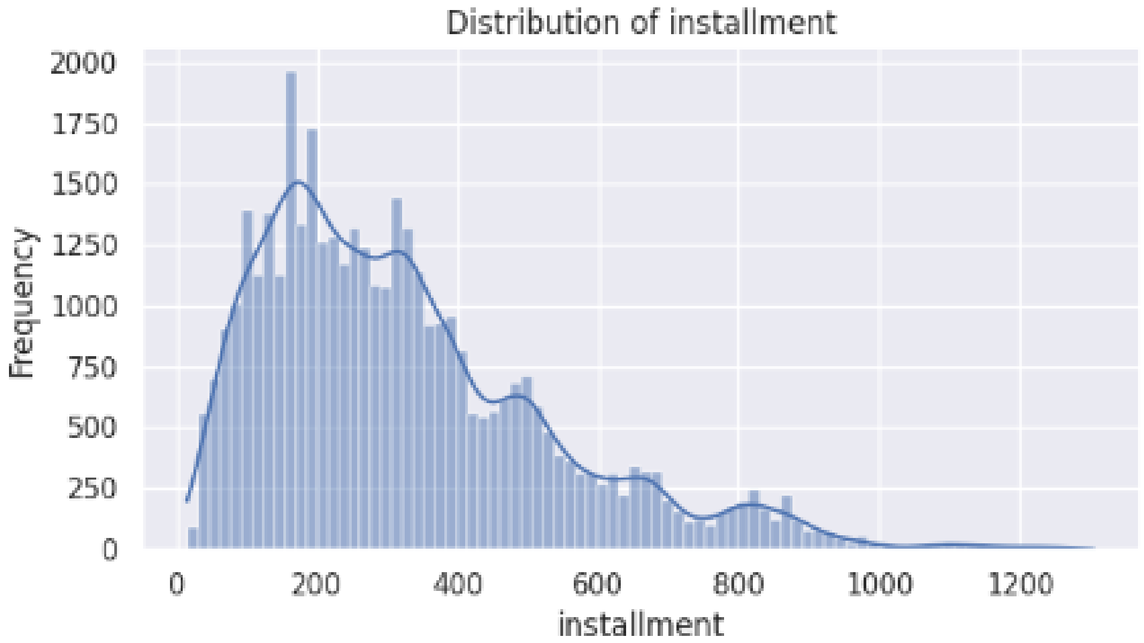
Challenges and proposed solutions

03. Data Visualization

- Univarite Analysis: Analyzed individual variables (numerical values)



Observation: Most of the interest rates are in the range of 10% - 12.5%.



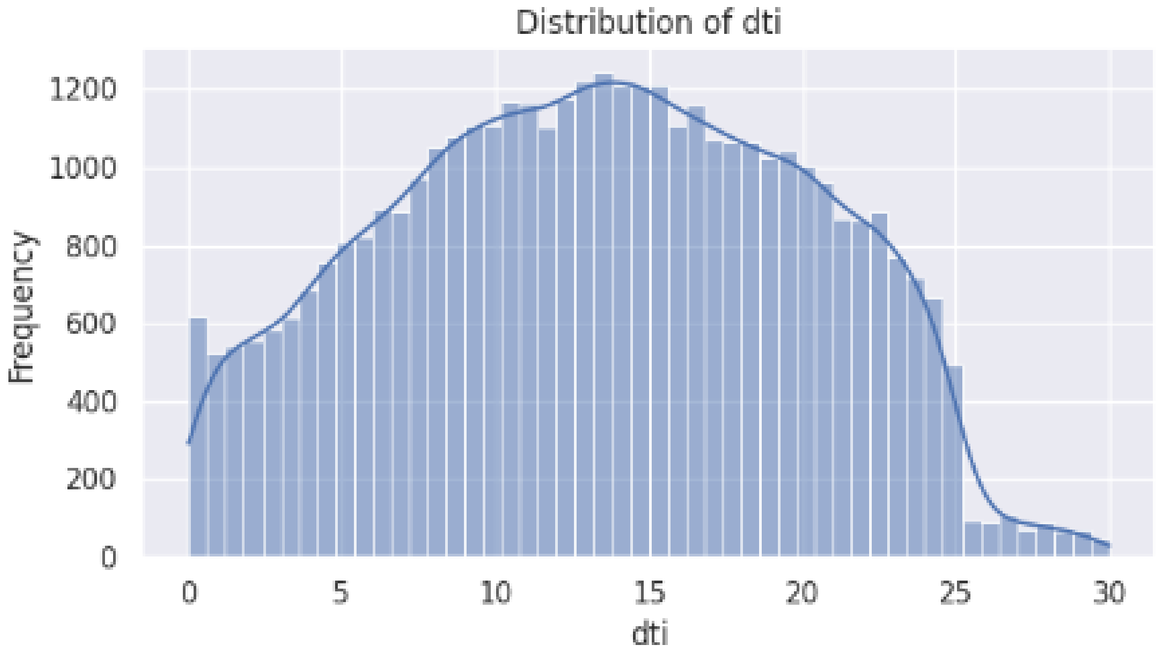
Observation: Most of the installments are in the range $100 - $200.

Challenges and proposed solutions

03. Data Visualization

- Univarite Analysis: Analyzed individual variables (numerical values)

Observation: Most of the annual_inc are distributed.

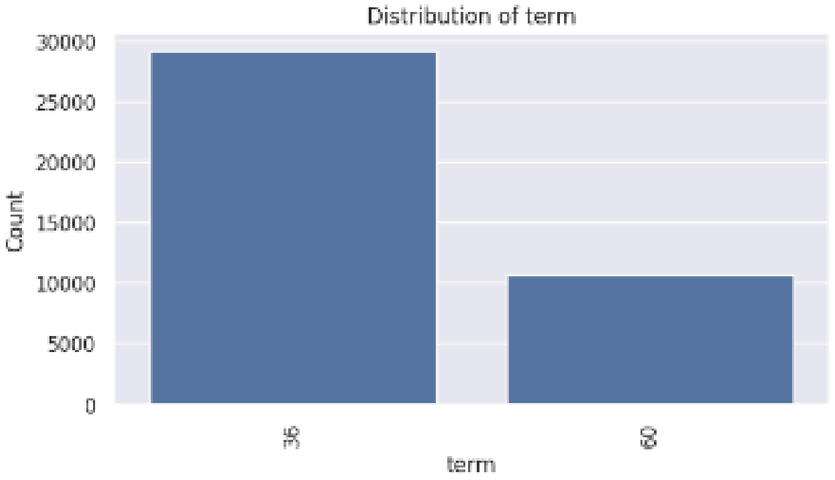


Observation: Most of the dti's are in the range 7.5 - 22.5.

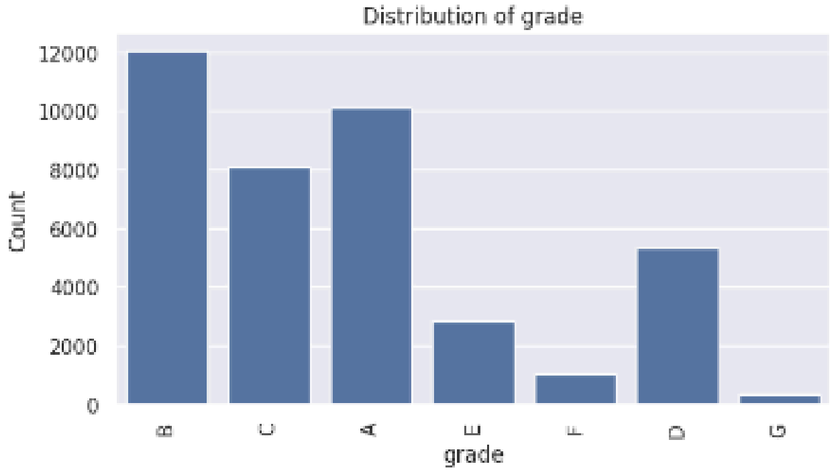
Challenges and proposed solutions

03. Data Visualization

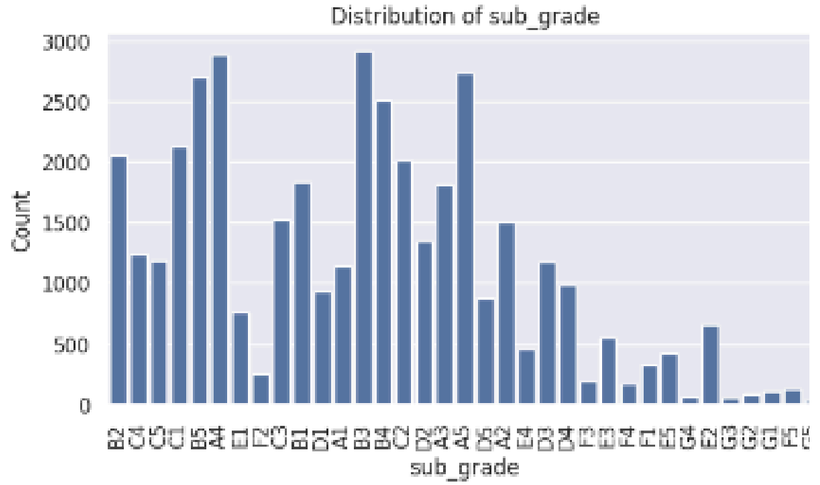
- Univarite Analysis: Analyzed individual variables (Categorical Values)



Observation: More loans are of 36 months duration.



Observation: Most loans belong to A,B and C category.

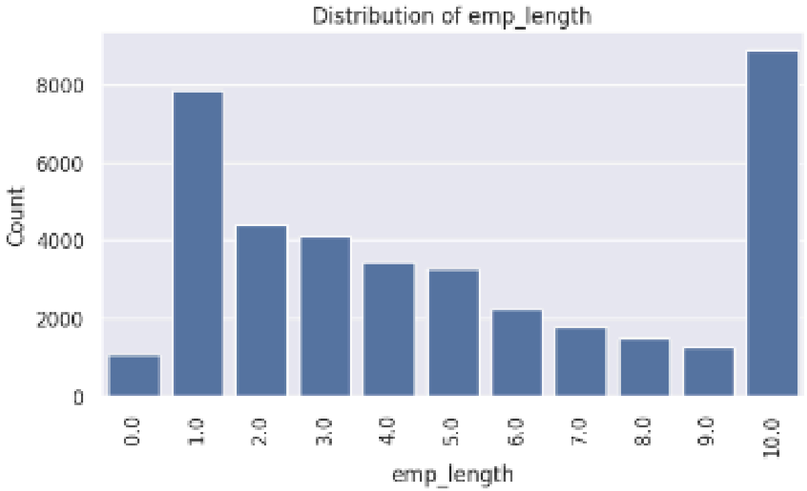


Observation: Loans have an irratic distribution for sub-grades.

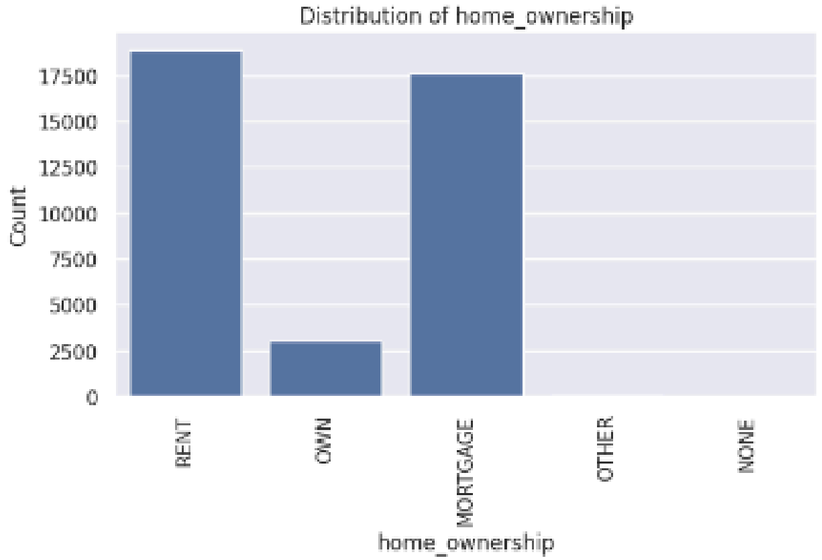
Challenges and proposed solutions

03. Data Visualization

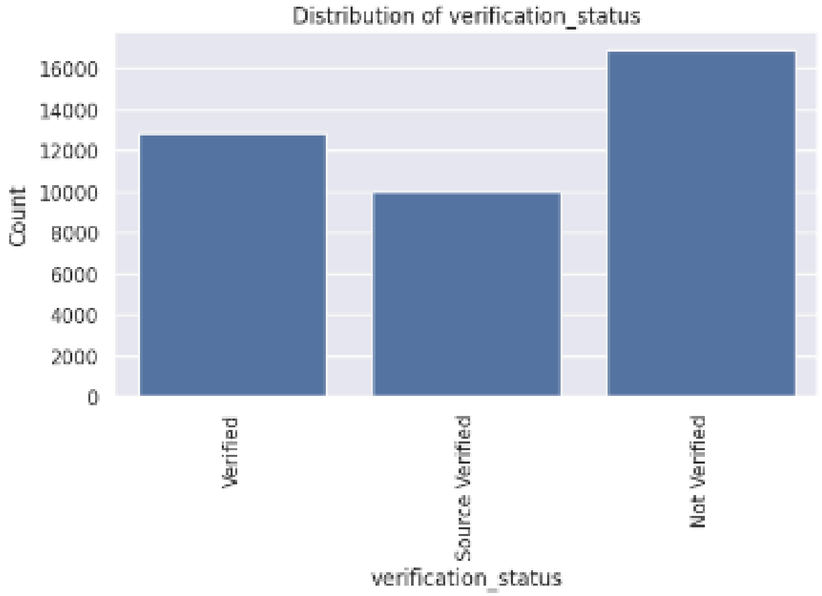
- Univarite Analysis: Analyzed individual variables (Categorical Values)



Observation: Most loans are by people who have more than 10 years of experience.



Observation: Most loans are of the RENT and MORTGAGE category.

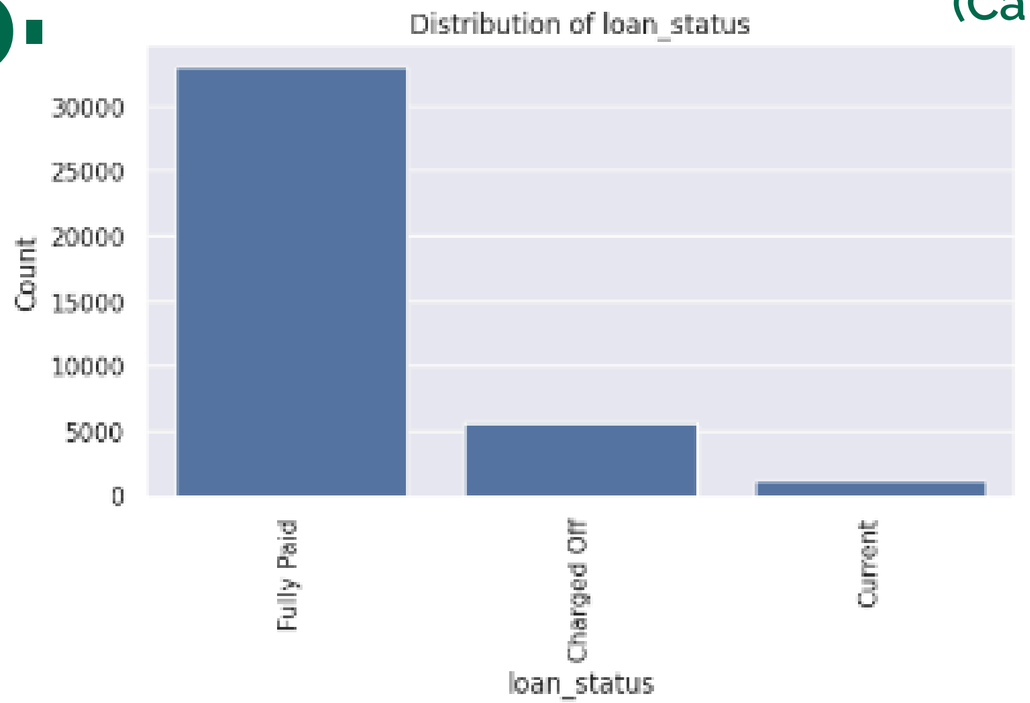


Observation: Loans are equally distributed in the status range.

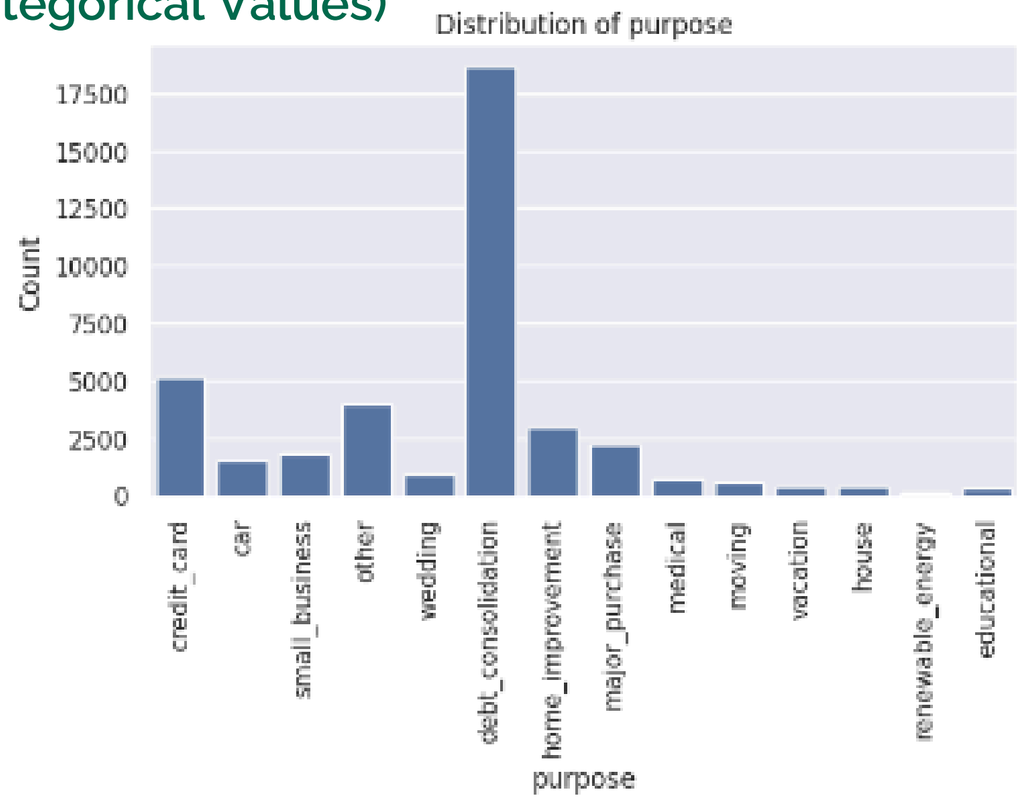
Challenges and proposed solutions

03. Data Visualization

- Univarite Analysis: Analyzed individual variables (Categorical Values)



Observation: About 15.28% of the loans get Charged Off.

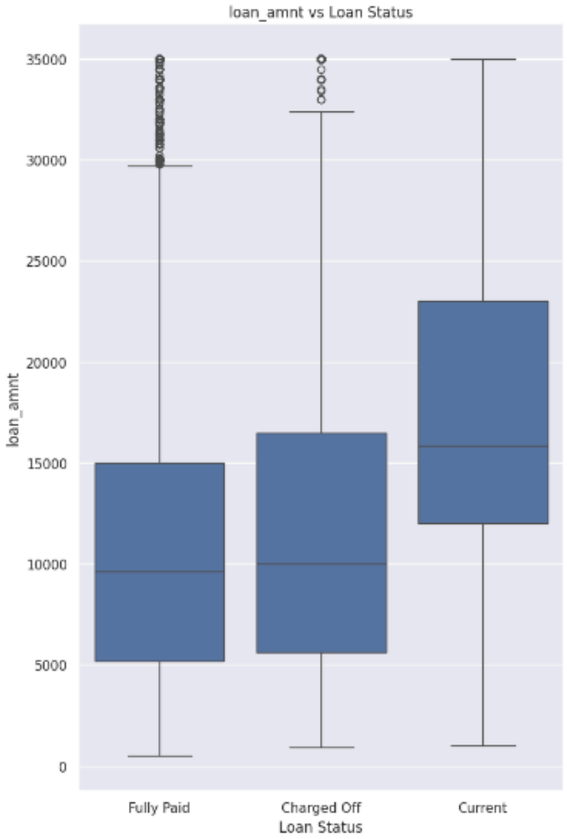


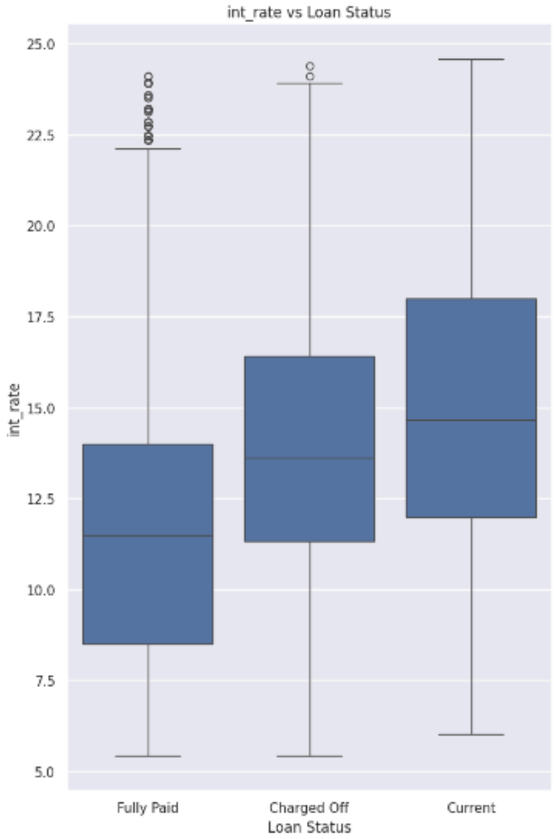
Observation: Most loans are for debt_consolidation.

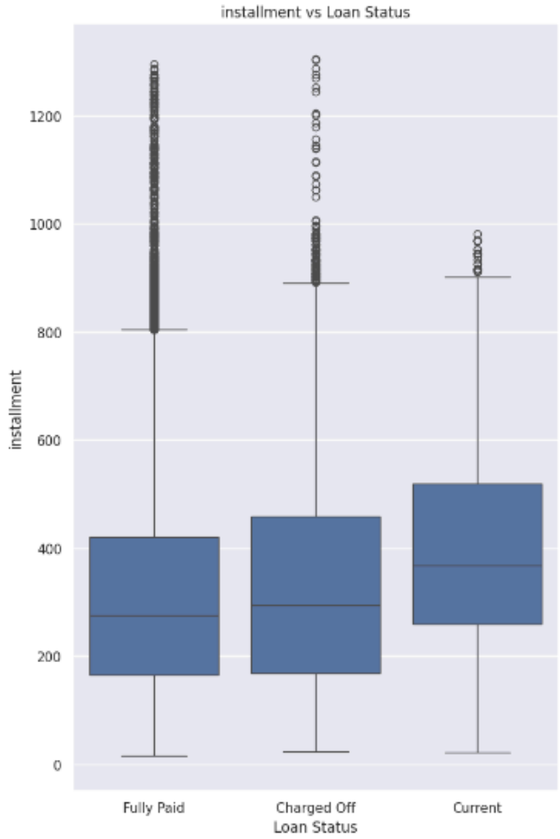
Challenges and proposed solutions

03. Data Visualization

Segmentated Univarite Analysis: Segmentated Analyzed individual variables







Observation: Loan amounts for charged off loans have similar median and mode.

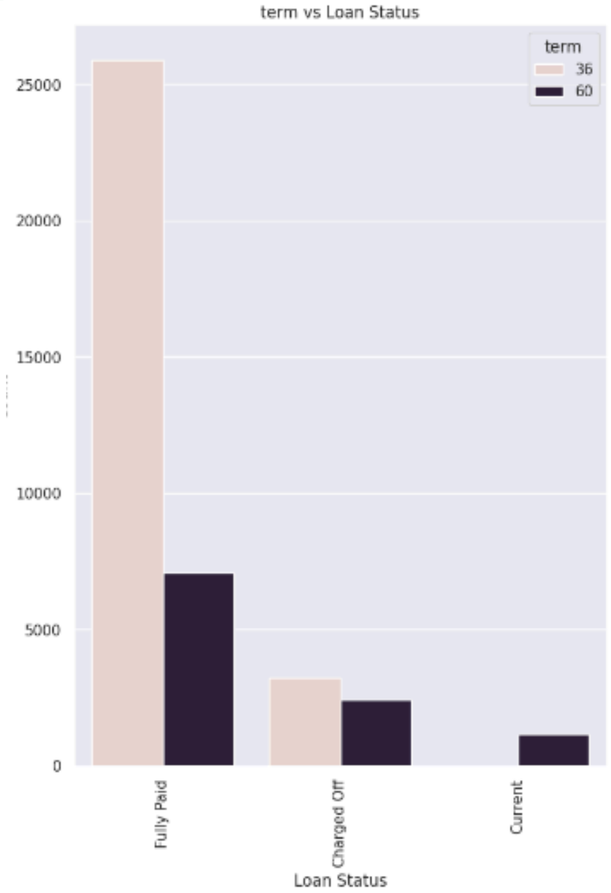
Observation: The data has similar metrices.

Observation: High interest loans have higher chances for failure.
Interest rate beyond 14% are most likely to get defaulted.

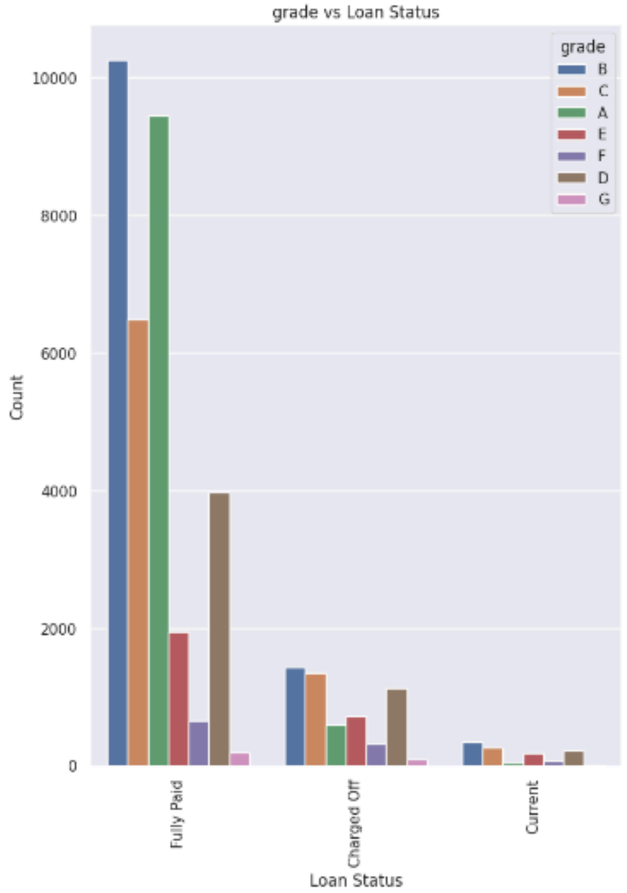
Challenges and proposed solutions

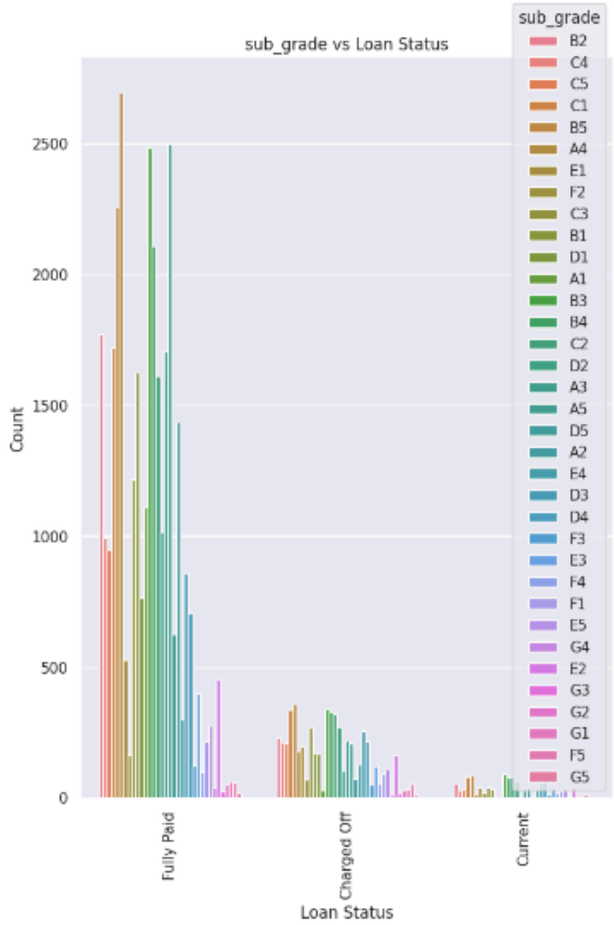
03. Data Visualization

Segmentated Univarite Analysis: Segmentated Analyzed individual variables



Observation: Loans of tenure 60 months are more likely to be defaulted.





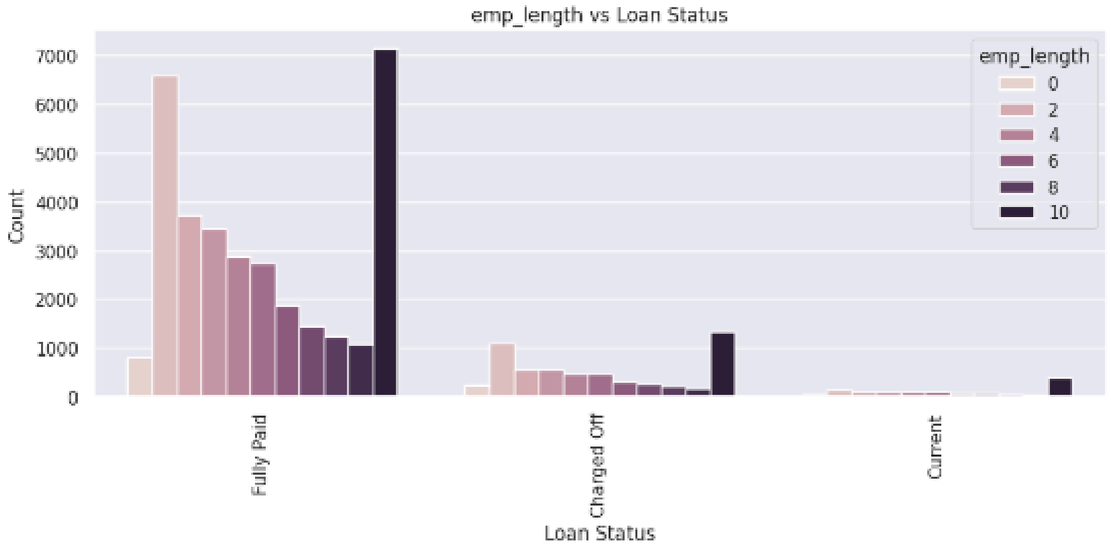
Observation: Loans are equally distributed in the status range.

Observation: Loans of grade B and E are more likely to be defaulted.

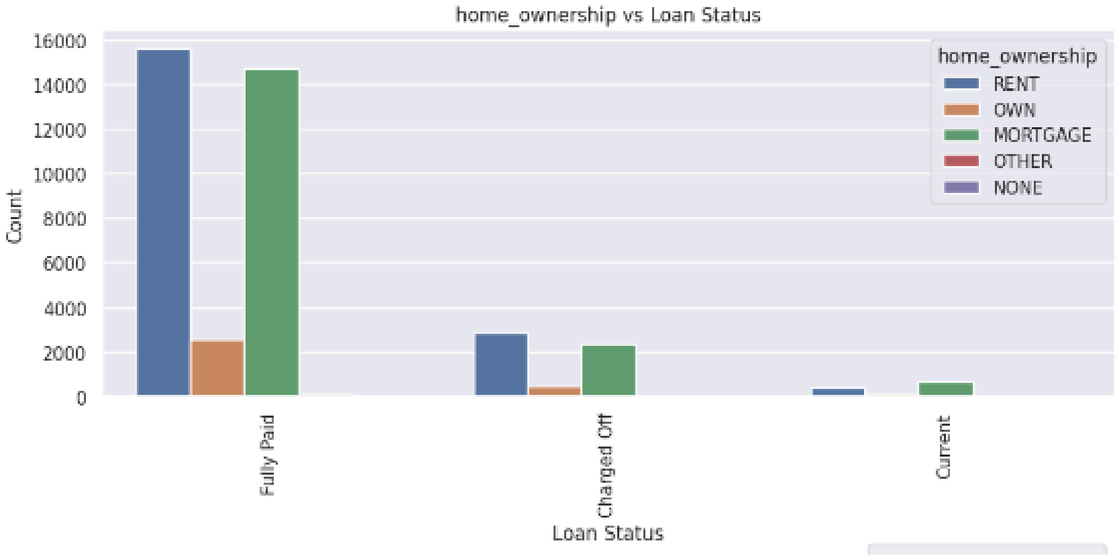
Challenges and proposed solutions

03. Data Visualization

Segmentated Univarite Analysis: Segmentated Analyzed individual variables



Observation: Highest loans are granted for the 10 years of experience.



Observation: Own category loans are more likely to be defaulted.

Challenges and proposed solutions

03. Data Visualization

Bivariate Analysis: Home Ownership and Verification Status vs. Charged Off Percentage

Home Ownership	Verification Status	Charged Off Percentage
MORTGAGE	Not Verified	11.89%
MORTGAGE	Source Verified	12.27%
MORTGAGE	Verified	15.08%
NONE	Not Verified	0.00%
OTHER	Not Verified	19.23%
OTHER	Source Verified	10.00%
OTHER	Verified	19.44%
OWN	Not Verified	12.62%
OWN	Source Verified	15.19%
OWN	Verified	17.25%
RENT	Not Verified	13.28%
RENT	Source Verified	15.83%
RENT	Verified	16.96%

This table displays the percentage of applicants that were charged off based on their house ownership and verification statuses. The charged off percentage is determined for each combination of property ownership and verification status, revealing how these variables interact to impact loan default rates.

Challenges and proposed solutions

04. Risk Profiling

Risk profiling identifies applicants who are more likely to default on their loans. By assessing crucial indicators, we may classify borrowers into risk tiers and make more educated lending decisions.

Risk Factors Considered:

- **Interest Rate (int_rate)**: Loans with interest rates more than **11%** are more likely to default, with rates larger than **14%** having a very high likelihood of default.
- **Loan Grade (grade)**: Loans with grades **E** and **F** have a higher default rate than loans with grades A, B, and C.
- **Term (term)**: Loans with a duration of **60 months** have a higher default rate than 36-month loans.

Challenges and proposed solutions

04. Risk Profiling

Risk Factor	Risk Level	Description
Interest Rate	High	Loans with interest rates > 14% have a very high probability of defaulting.
Interest Rate	Medium	Loans with interest rates between 11% and 14% are more likely to default.
Loan Grade	High	Loans with grades E and F are more likely to default.
Loan Grade	Low	Loans with grades A, B, and C are less likely to default.
Term	High	Loans with a term of 60 months are more likely to default.
Term	Low	Loans with a term of 36 months are less likely to default.
Verification Status	High	Non-verified loans have a higher risk of default.
Verification Status	Low	Verified and source-verified loans have a lower risk of default.
Home Ownership	Low	No significant impact on default risk identified from home ownership status.
Employment Length	Low	No significant impact on default risk identified from employment length.
Public Record	Low	No significant impact on default risk identified from public records.

Challenges and proposed solutions

04. Risk Profiling

Risk Factor	Risk Level	Description
Purpose	Low	No significant impact on default risk identified from loan purpose.
Application Type	Low	No significant impact on default risk identified from application type.
Loan Amount	Low	No significant impact on default risk identified from loan amount.
Installment	Low	No significant impact on default risk identified from monthly installment amount.
Funded Amount	Low	No significant impact on default risk identified from funded amount.
Payment Plan	Low	No significant impact on default risk identified from payment plan.
Sub Grade	Medium to High	Sub-grades provide finer granularity of risk within loan grades, with lower sub-grades within E and F exhibiting higher risk.
Delinquency (delinq_2yrs)	Low	No significant impact on default risk identified from the number of delinquencies in the past two years.

Summary

Using the risk indicators listed above, we can create a risk profile for each loan application, allowing the organization to make better informed lending decisions while minimizing possible financial losses.

By concentrating on high-risk criteria such as **interest rate**, **loan grade, duration**, and **verification status**, we can efficiently detect and **reduce** the danger of default.

Conclusion

We used Exploratory Data Analysis (EDA) to investigate numerous factors that impact the chance of loan defaults. Our main results are as follows:

Loan grades:
Loan grades **E** and **F** have a greater chance of default than grades **A**, **B**, and **C**. This implies that lower-quality loans are riskier investments for the lending institution.

Loan Terms:
Loans with a **60**-month duration have a higher default rate than **36**-month term loans. This suggests that longer-term loans carry a larger risk for the lender.

Loan Amount:
The majority of loans vary from **Rs. 2500** to **Rs. 5000**. This range has no significant influence on the chance of default, indicating that loan amount alone is not a reliable predictor of default risk.

Conclusion

We used Exploratory Data Analysis (EDA) to investigate numerous factors that impact the chance of loan defaults. Our main results are as follows:

Monthly installments:
Most monthly installments vary from **Rs.100** to **Rs.200**, which has no direct influence on default rates. This demonstrates that payment size is not an important factor in predicting loan default.

Interest Rates:
Loans with interest rates greater than **11%** are more likely to default, especially those with rates higher than **14%.** **Higher interest rates** dramatically raise the **danger of default.**

Charged-Off Loans:
Approximately **15%** of loans are charged off, resulting in significant financial losses for the organization. This emphasizes the significance of correctly calculating the risk of default in order to avoid damages.

Presented by Dheeraj Salwadi & Devesh Khatri

Thank you very much!