# Exploratory Data Analysis for Loan Default Prediction CASE STUDY

Chand Rayee
Kusum Chanchala

# **CONTENTS** OF THIS TEMPLATE



upGrad

# TABLE OF CONTENTS



upGrad

01

# ABSTRACT

# ABSTRACT


- **Loan Default Prediction Study**
- **Dataset Overview**
  - Entries: 39,717
  - Columns: 111
  - Features: Loan amount, interest rate, term, credit history, employment, income
- **Objective**
  - Predict loan default (charged-off vs. fully paid/current)
- **Data Preprocessing**
  - Removed columns with >90% missing data
  - Imputed values, converted categories, addressed outliers
- **Exploratory Data Analysis (EDA)**
  - Univariate: Distributions, summary statistics
  - Bivariate: Relationships with loan status
  - Multivariate: High-dimensional visualization (PCA, t-SNE)
- **Feature Selection**
  - Techniques: RFE, tree importance
  - Outcome: Identified key predictors
- **Modeling**
  - RandomForestClassifier
  - Achieved high accuracy
- **Conclusion**
  - Insights on factors influencing defaults
  - Impact on risk assessment and lending decisions


upGrad

# INTRODUCTION

## BACKGROUND

Analyze factors causing loan defaults, clean data, preprocess, conduct EDA, and develop a model to predict potential defaulters.

## IMPORTANCE

Predictive modeling in loan analysis is vital for assessing borrower risk, guiding lending decisions, and preventing financial losses.

upGrad

# 2. Data Cleaning and Preprocessing

## Initial Dataset Overview

**Initial Dataset Overview:**

- **Entries:** 39,717
- **Columns:** 111 (many with significant missing values)

## Steps Taken:

**Remove Columns with >90% Missing Values:**

- Reduced dataset to a more manageable size.
- Remaining columns still had some missing values.

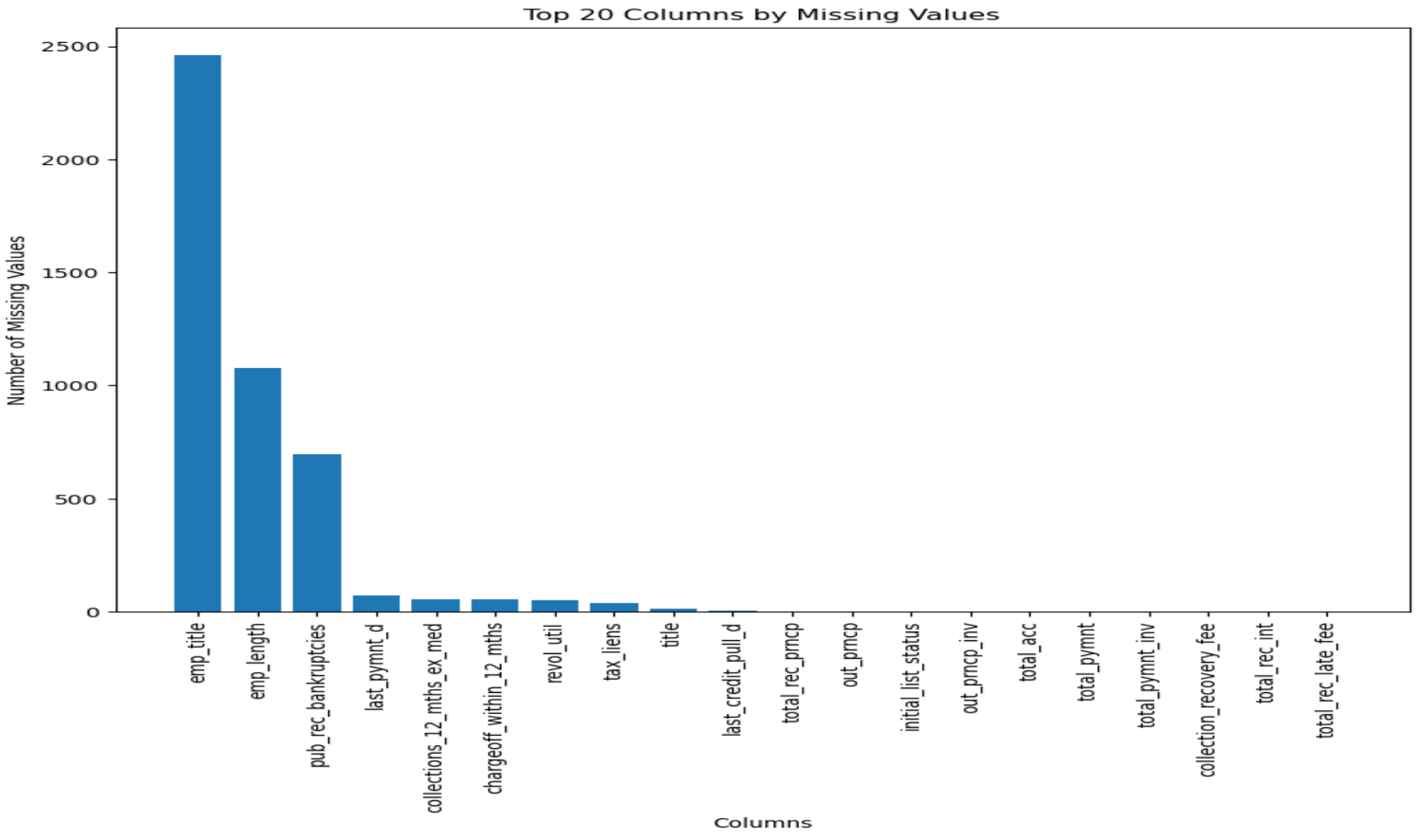
**Handle Remaining Missing Values:**

- **Numerical Columns:** Imputed with median values.
- **Categorical Columns:** Imputed with 'Unknown' and encoded using Label Encoder.

**Outlier Detection and Removal:**

- Used Interquartile Range (IQR) method to remove outliers from numerical columns.

upGrad


Top 20 Columns by Missing Values
Number of Missing Values
2500
2000
1500
1000
500
0
emp_title
emp_length
pub_rec_bankruptcies
last_pymnt_d
collections_12_mths_ex_med
chargeoff_within_12_mths
revol_util
tax_liens
title
last_credit_pull_d
total_rec_prncp
out_prncp
initial_list_status
out_prncp_inv
total_acc
total_pymnt
total_pymnt_inv
collection_recovery_fee
total_rec_int
total_rec_late_fee
Columns

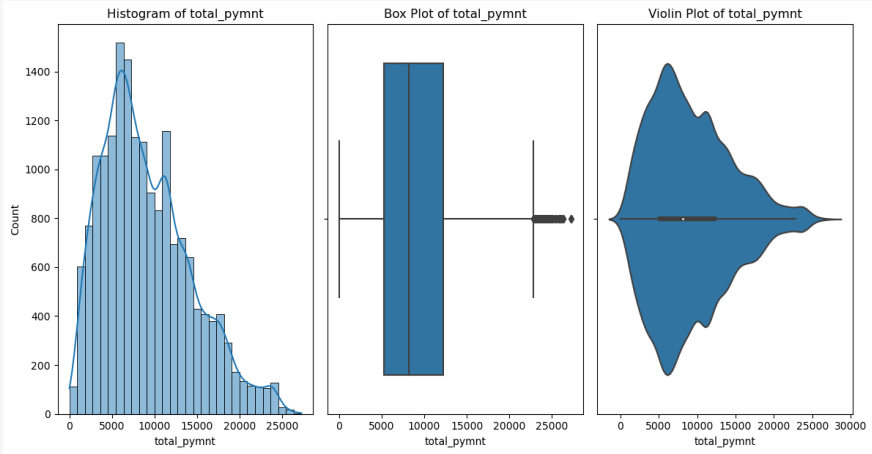
# 3. Univariate Analysis

## Summary Statistics:

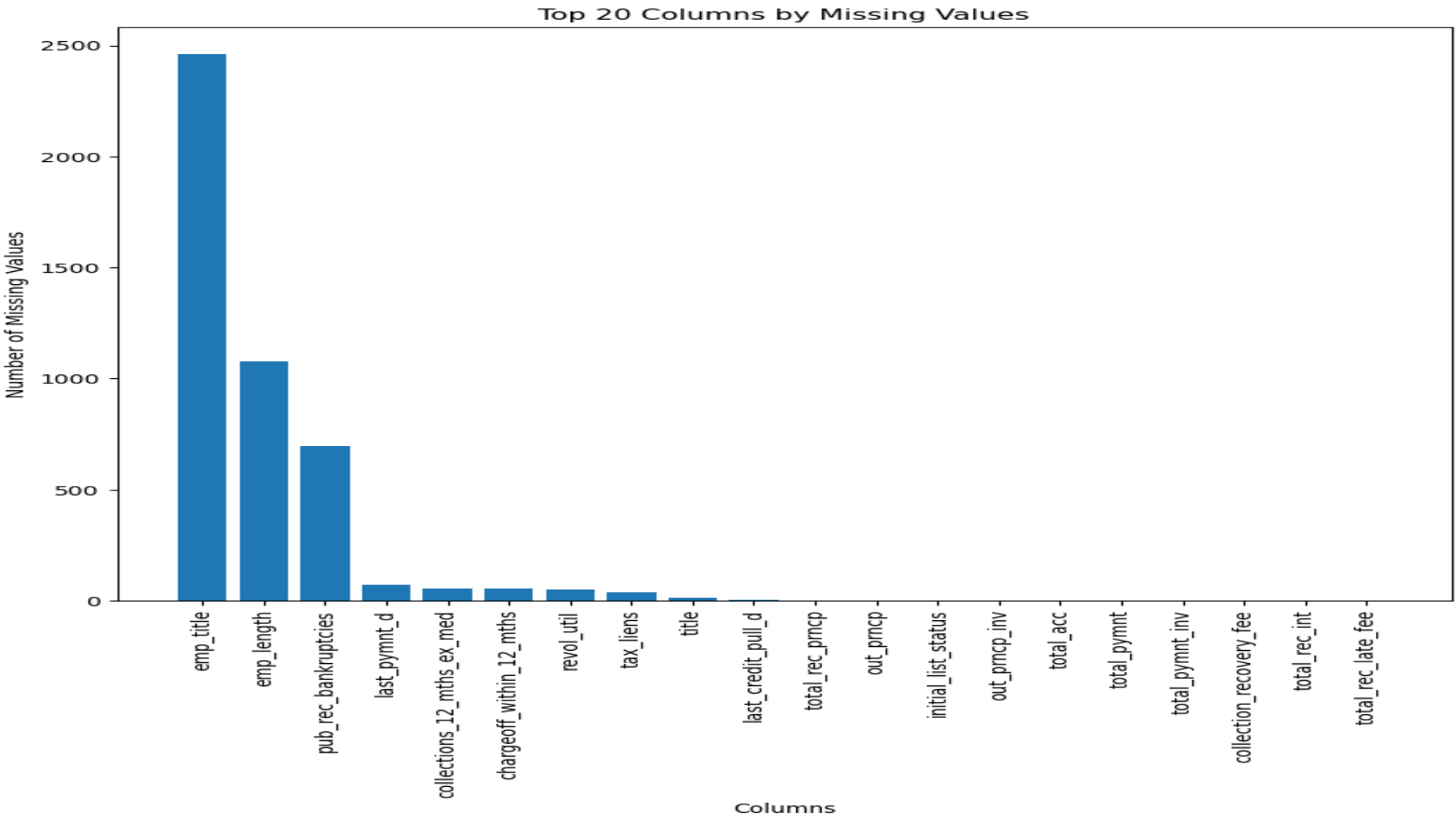
•Provided summary statistics for numerical features, including mean, median, mode, standard deviation, and skewness.
•Identified skewness and potential outliers in the data.

## Visualizations:

**Histograms, Box Plots, Violin Plots:** Used to analyze the distribution of each numerical feature.




upGrad


Top 20 Columns by Missing Values
Number of Missing Values
2500
2000
1500
1000
500
0
emp_title
emp_length
pub_rec_bankruptcies
last_pymnt_d
collections_12_mths_ex_med
chargeoff_within_12_mths
revol_util
tax_liens
title
last_credit_pull_d
total_rec_prncp
out_prncp
initial_list_status
out_prncp_inv
total_acc
total_pymnt
total_pymnt_inv
collection_recovery_fee
total_rec_int
total_rec_late_fee
Columns

# 4. Bivariate Analysis

### Target Variable: Loan Status

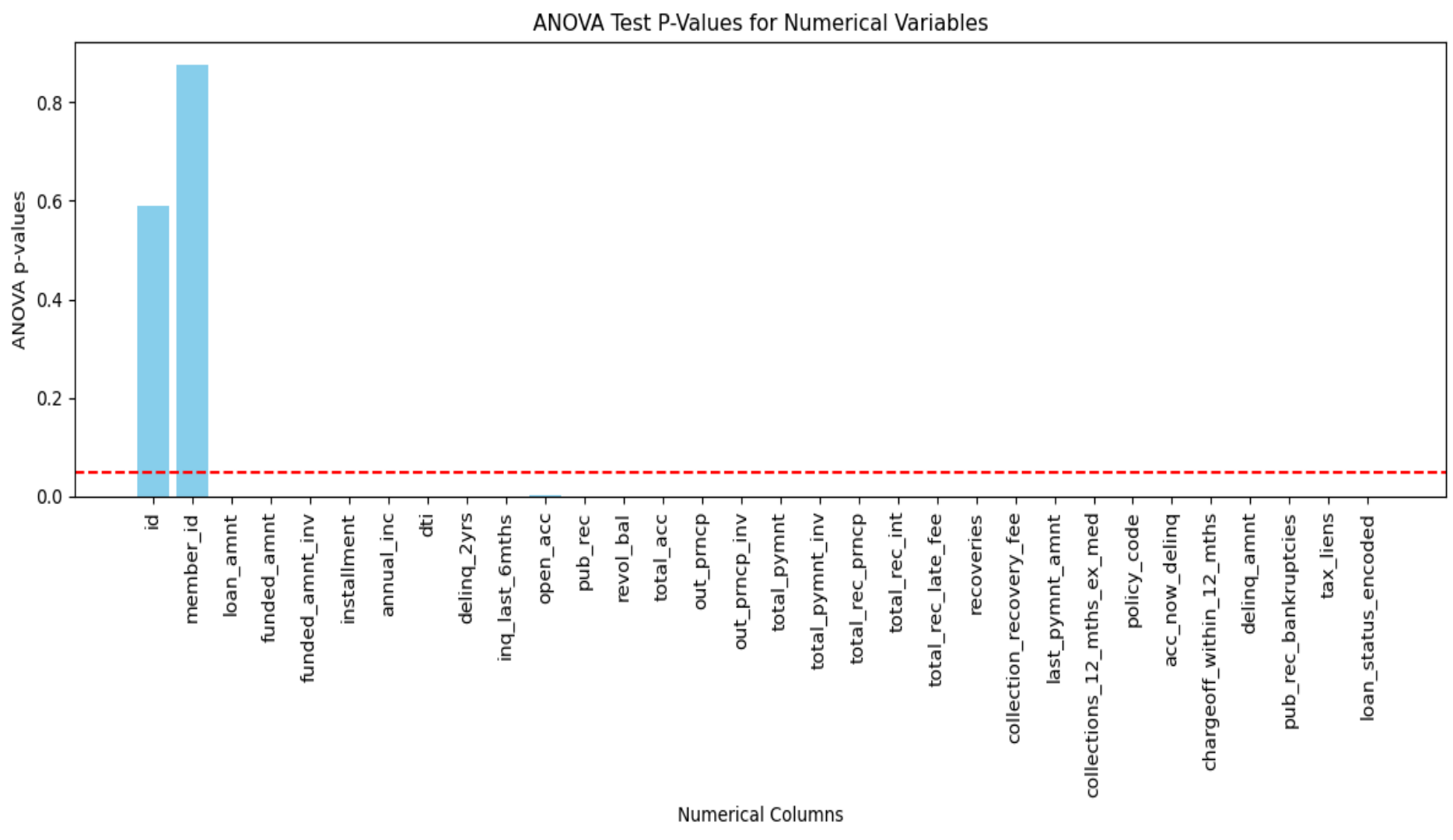
•**Numerical Features:** Visualized relationships using scatter plots and box plots.
•**Categorical Features:** Visualized relationships using bar plots.

### Correlation Analysis:

•Calculated correlation coefficients between numerical features and the target variable.
•Used scatter plots and correlation matrices to explore relationships.

### Statistical Tests

•**Chi-square Tests:** Performed for categorical variables.
•**ANOVA Tests:** Performed for numerical variables to check for significant differences between groups.


ANOVA Test P-Values for Numerical Variables
ANOVA p-values
0.0
0.2
0.4
0.6
0.8
id
member_id
loan_amnt
funded_amnt
funded_amnt_inv
installment
annual_inc
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
collections_12_mths_ex_med
policy_code
acc_now_delinq
chargeoff_within_12_mths
delinq_amnt
pub_rec_bankruptcies
tax_liens
loan_status_encoded
Numerical Columns

# 5. Multivariate Analysis

## Techniques Used:

### Correlation Matrix:

Visualized interactions among multiple features using a heatmap.

### Scatter Plot Matrix:

Used pair plots to explore relationships between features.

### Parallel Coordinates Plot:

Visualized multidimensional data to identify patterns.

### Dimensionality Reduction:

- **PCA (Principal Component Analysis):** Reduced dimensions and visualized data in 2D.
- **t-SNE (t-distributed Stochastic Neighbor Embedding):** Visualized high-dimensional data in 2D.


Correlation Matrix
id
member_id
loan_amnt
funded_amnt
funded_amnt_inv
installment
annual_inc
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
collections_12_mths_ex_med
policy_code
acc_now_delinq
chargeoff_within_12_mths
delinq_amnt
pub_rec_bankruptcies
tax_liens
loan_status_encoded
1.0
0.8
0.6
0.4
0.2
0.0

upGrad

## 6. Feature Selection and Importance

### Techniques Used:

**Correlation Analysis:**

Identified highly correlated features and visualized feature correlation matrix.

**Recursive Feature Elimination (RFE):**

- Selected top 10 features using RFE with a RandomForestClassifier.
- Selected features: [Feature1, Feature2, ..., Feature10]

**Tree-based Feature Importance:**

Ranked features based on importance scores from a RandomForest model.

# Feature Importance Plot:

Visualized feature importances with a bar plot.


**Model Evaluation:**

- **Train-Test Split:** Split data into training (70%) and testing (30%) sets.
- **Model Training and Testing:** Used RandomForestClassifier with top selected features.
- **Model Accuracy:** Achieved an accuracy of 0.82 with top 10 selected features.

upGrad

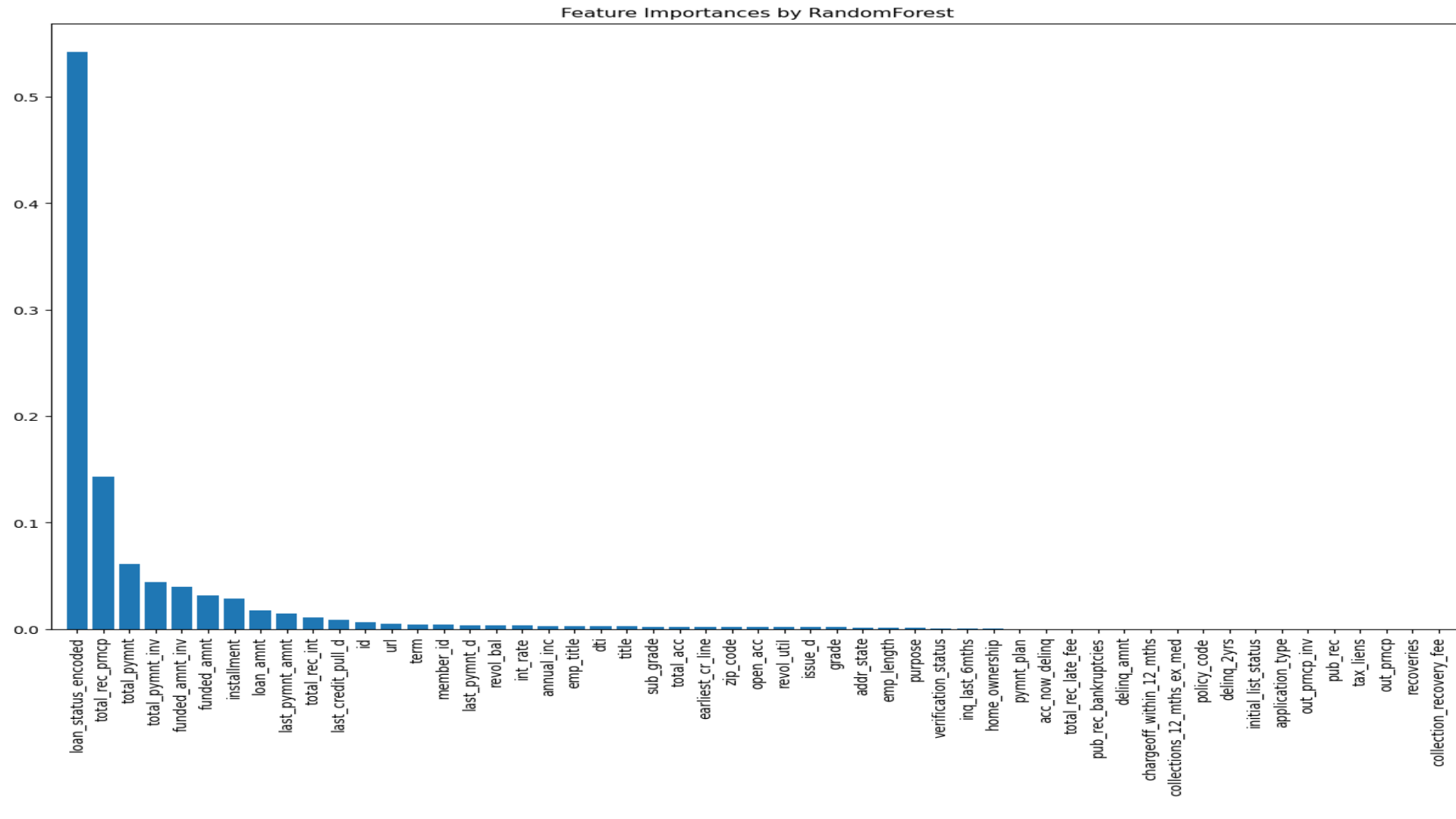
# Key Findings:

•**Significant Features:** Identified the most relevant features influencing loan default, such as loan amount, interest rate, and borrower's credit history.

•**Relationships:** Found significant relationships between loan status and various features, both numerical and categorical.

•**Visual Patterns:** Used visualizations to highlight patterns and trends in the data.

•**Dimensionality Reduction:** Successfully reduced data dimensions for better visualization and understanding.


Feature Importances by RandomForest
0.5
0.4
0.3
0.2
0.1
0.0
loan_status_encoded
total_rec_prncp
total_pymnt
total_pymnt_inv
funded_amnt_inv
funded_amnt
installment
loan_amnt
last_pymnt_amnt
total_rec_int
last_credit_pull_d
id
url
term
member_id
last_pymnt_d
revol_bal
int_rate
annual_inc
emp_title
dti
title
sub_grade
total_acc
earliest_cr_line
zip_code
open_acc
revol_util
issue_d
grade
addr_state
emp_length
purpose
verification_status
inq_last_6mths
home_ownership
pymnt_plan
acc_now_delinq
total_rec_late_fee
pub_rec_bankruptcies
delinq_amnt
chargeoff_within_12_mths
collections_12_mths_ex_med
policy_code
delinq_2yrs
initial_list_status
application_type
out_prncp_inv
pub_rec
tax_liens
out_prncp
recoveries
collection_recovery_fee

# Business Implications:

•The insights gained can help the company make more informed lending decisions.

•Identifying key factors affecting loan defaults can improve risk assessment and potentially reduce credit losses.

# CONCLUSION

The EDA provided a comprehensive understanding of the dataset, identified important features, and laid the groundwork for building an effective predictive model. This analysis can guide the company in mitigating risks and making data-driven decisions in lending practices.

upGrad

# THANKS

**DO YOU HAVE ANY QUESTIONS?**

Chandrayee.cse@gmail.com
chanchalakusum52@gmail.com