Loan Approval – Complete EDA Flow + Business Questions

Dataset context (typical columns):
income, credit_score, loan_amount, employment_status, dependents, education, age, experience, approval (target)

Step 1: Basic Understanding

- Load the data
- Check shape
- View first rows
- Use info() and describe()

Questions:

1. How many records and columns are present?
2. Which columns are numerical and which are categorical?
3. Is the dataset balanced between approved and rejected loans?
4. Are there any columns with wrong data types?

Step 2: Missing Value Analysis

- Check null values
- Percentage of missing data
- Visualize missing values (heatmap/bar)

Questions:
5. Which features have missing values?
6. What percentage of data is missing in each column?
7. Is missingness random or pattern-based?
8. Should we drop or impute these values? Why?

Step 3: Univariate Analysis (Single Column)

For Numerical:

- Histograms
- Boxplots

For Categorical:

- Countplots

Questions:
9. What is the distribution of income?
10. Are most applicants low-income or high-income?
11. Are there outliers in income or loan_amount?
12. How are credit scores distributed?
13. Which employment status is most common?
14. What is the most frequent education level?

Step 4: Bivariate Analysis (Feature vs Target)

Numerical vs Target:

- income vs approval
- credit_score vs approval
- loan_amount vs approval

Categorical vs Target:

- employment_status vs approval
- education vs approval

Questions:
15. Do approved applicants have higher income on average?
16. Is there a minimum credit score where approval increases sharply?
17. Are larger loan amounts rejected more often?
18. Which employment type has the highest approval rate?
19. Does education level affect approval?
20. Are applicants with more dependents rejected more?

Step 5: Multivariate Analysis

- Pairplot
- Correlation heatmap
- Groupby analysis

Questions:
21. Which numerical features are highly correlated?
22. Is income correlated with loan_amount?
23. Do people with high income but low credit still get rejected?
24. What combination gives highest approval:
(High income + high credit) or (Medium income + very high credit)?
25. Are there hidden patterns across 3 variables?

Step 6: Outlier Detection

- Boxplots
- IQR / Z-score

Questions:
26. Are extreme income values realistic or data errors?
27. Should outliers be capped or removed?
28. How do outliers affect approval trends?

Step 7: Business Insights from EDA

Students should answer:

29. What type of customer gets approved most?
30. What type of customer gets rejected most?
31. Which 3 features influence approval the most?
32. Can we define a “safe applicant profile”?
33. Can EDA alone help a bank reduce risk?
34. What bias can exist in this data?
35. Which features should go into the model?
36. Which features can be dropped?
37. What preprocessing is required before modeling?

These questions force students to:

- Think like analysts
- Reason like a bank
- Prepare for Logistic Regression
- Connect plots to real decisions

They won’t just “draw graphs”.
They will *understand how data controls real-world approvals*.