FOUNDATIONS OF DATA SCIENCE (CSD355)
LAB ASSIGNMENT - 4
STASTICS
(Deadline: 18-09-2025)

Download the loan dataset from the given link:-
https://www.kaggle.com/datasets/tanishaj225/loancsv

The **Loan dataset** is a well-known real-world dataset commonly used for teaching data analysis and predictive modeling. It contains records of loan applications with a mix of categorical and numerical variables such as applicant demographics (Gender, Marital Status, Education, Self-Employed), financial details (ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History), and contextual information (Property_Area). The target variable is Loan_Status, indicating whether a loan was approved (Y) or not (N). The dataset has around 600 entries, includes missing values, and outliers. The columns are as follows:

- Loan_ID
- Gender – (Male, Female).
- Married – Marital status (Yes, No).
- Dependents
- Education – (Graduate, Not Graduate).
- Self_Employed – Employment type (Yes = Self-employed, No = Salaried).
- ApplicantIncome
- CoapplicantIncome
- LoanAmount
- Loan_Amount_Term – Duration
- Credit_History – (1 = good history, 0 = no/poor history).
- Property_Area – (Urban, Semiurban, Rural).
- Loan_Status – (Y = Approved, N = Not Approved).

1. After downloading the .csv file, load its contents into a DataFrame in a Python notebook and perform preprocessing techniques as done in Lab 2, i.e., check for and handle null values if any (by using mean, median, and mode).
2. Compute the kurtosis of LoanAmount. Check whether the distribution is leptokurtic (peaked), platykurtic (flat), or mesokurtic. Visualize the distribution.
3. Analyze the distribution of LoanAmount separately for applicants whose loans were approved (Loan_Status = 'Y') and those whose loans were not approved (Loan_Status =

'N'). For each group, calculate the skewness and kurtosis, and determine which group shows a more extreme distribution.

4. Plot a histogram of ApplicantIncome with mean and median. What does the graph reveal about the income distribution of applicants?
5. Find the mean ApplicantIncome separately for Male and Female applicants. Which group has a higher central tendency of income?
6. Compute the mean and median of LoanAmount separately for approved and not approved loans. Which group tends to apply for higher loan amounts?
7. Using both ApplicantIncome and CoapplicantIncome, create a new column TotalIncome. Compute its mean, standard deviation, and IQR. Compare variability of TotalIncome with ApplicantIncome alone.
8. Create a new variable TotalIncome = ApplicantIncome + CoapplicantIncome. Compare the skewness, kurtosis, and coefficient of variation (CV = StdDev / Mean) of ApplicantIncome, CoapplicantIncome, and TotalIncome. Interpret whether combining incomes reduces skewness and variability in the data.
9. Calculate the mean, median, and interquartile range (IQR) of the ApplicantIncome variable to understand the central tendency and spread of income among applicants. Using the IQR method, identify and list all the outliers present in the income data. After detecting the outliers, remove them to create a cleaned dataset and then recalculate the measures of central tendency (mean and median). Compare the results before and after outlier removal to analyze how the presence of extreme values affects the central tendency of applicant incomes.
10. For Loan_Status = 'Y' and Loan_Status = 'N', compute the variance, standard deviation, skewness, and kurtosis of LoanAmount. Use both numerical measures and histograms to compare how the distributions differ between approved and rejected applicants.