PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

Department of Computer Science and Engineering

UE24MA242A - Mathematics for Computer Science Engineers

Session: August-December 2025

BANANA LEVEL PROBLEM

The **Banana Problem** is designed to assess students' ability to handle **end-to-end data acquisition and preprocessing**, a critical challenge across modern data-driven industries. While large volumes of data are available online, transforming unstructured web data into actionable insights remains a major bottleneck. This problem not only focuses on **scraping content from online sources**, but also emphasizes the importance of **data preprocessing**, **validation**, **descriptive analytics, and visualization**.

Objectives:

- To **scrape content** from one or more specified web sources.
- To **curate** the scraped content into a structured format (CSV).
- To apply **data cleaning and preprocessing techniques**, including:
  - Handling missing or inconsistent values
  - Checking for data correctness
- To perform **descriptive statistical analysis**
- To derive insights using appropriate **visualization techniques**

Tasks:

1. Accept **a website URL** as input.
2. **Scrape relevant content** from the web pages (e.g., headlines, product reviews, blog texts, articles).
3. Store the scraped content in a **CSV file** in a clean, columnar format.
4. Perform **data preprocessing**, which includes:
   - Detecting and handling **missing values**
   - Checking and correcting **data types**
   - **Removing duplicates** or noise if present
5. Conduct **descriptive statistical analysis** on the curated data.
6. Present the findings using suitable **graphs and plots** (e.g., Bar graph for comparison, histograms to detect skewness, boxplots for outlier detection).

Expected Output:

- A structured .csv file containing the cleaned and processed dataset.
- A script or notebook that performs:
  - Data preprocessing
  - Descriptive statistics
  - Visual analysis
- A summary of **insights** or patterns observed from the dataset.

Sample URL's from which web scraping can be done.

1. IMDb – Movies Dataset

- **URL:** https://www.imdb.com/chart/top/
- **What you get:** Movie titles, release year, ratings, number of votes.
- **Tasks:** Clean missing ratings, calculate average ratings by year/genre, visualize trends.

2. Books to Scrape (Sandbox for Practice)

- **URL:** http://books.toscrape.com/
- **What you get:** Book titles, prices, availability, ratings.
- **Tasks:** Handle missing prices/ratings, compute average price, visualize price distributions.

3. World Bank Data

- **URL:** https://data.worldbank.org/indicator
- **What you get:** GDP, population, literacy, CO_2 emissions.
- **Tasks:** Fill missing values for countries, compute growth rates, visualize comparisons with histograms

4. Wikipedia Tables

- **Example URL:** https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)
- **What you get:** Country GDP, GDP per capita, etc.
- **Tasks:** Clean inconsistent formats, handle "—" missing entries, plot GDP distributions.

5. Weather Data (Time Series)

- **Example URL:** https://www.timeanddate.com/weather/india/bangalore/historic
- **What you get:** Daily/hourly weather (temp, humidity, precipitation).
- **Tasks:** Clean N/A values, compute monthly averages, visualize seasonal trends.

6. GitHub Trending Repositories

- **URL:** https://github.com/trending
- **What you get:** Repository names, stars, programming language.
- **Tasks:** Handle missing descriptions, compute distribution of repos by language, visualize stars vs language.

7. Sports Statistics (ESPN / Cricket Info)

- **Example URL:** https://www.espncricinfo.com/records
- **What you get:** Player stats (runs, wickets, averages).
- **Tasks:** Handle incomplete stats, compute averages, visualize top players by runs/wickets.

8. UN Data – World Statistics

- **URL:** https://data.un.org/en/iso/
- **Data:** Population, life expectancy, health indicators.
- **Tasks:** Handle missing country values, compute averages per continent, visualize world maps.

9. Open Library Books

- **URL:** https://openlibrary.org/subjects/science
- **Data:** Book titles, authors, publication year, subject tags.
- **Tasks:** Clean missing authors/dates, analyze most common subjects, plot publications over time.

10. OECD Statistics

- **URL:** https://data.oecd.org/
- **Data:** Education levels, income inequality, unemployment rates.
- **Tasks:** Clean inconsistent "n/a" entries, compute descriptive stats, plot country comparisons.

11. Wikipedia – Olympic Medal Tables

- **URL:** https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table
- **Data:** Country-wise medal counts.
- **Tasks:** Handle missing medal entries, compute averages per country, visualize with bar charts.

12. Music Charts (Billboard)

- **URL:** https://www.billboard.com/charts/hot-100/
- **Data:** Song names, artists, rank, weeks on chart.
- **Tasks:** Clean missing weeks, compute frequency of artists, visualize rank vs duration.

13. NBA Player Stats

- **URL:** https://www.basketball-reference.com/leagues/NBA_2024_totals.html
- **Data:** Player stats like points, rebounds, assists.
- **Tasks:** Handle blank cells, compute averages, visualize distributions (histograms)

14. Airbnb Listings (InsideAirbnb Project)

- **URL:** http://insideairbnb.com/get-the-data/
- **Data:** Listings with price, location, number of reviews.
- **Tasks:** Clean missing reviews/price entries, compute average price per city, plot availability trends.

15 Glassdoor/Job Listings (via Kaggle dump)

- **Example Kaggle dataset:** https://www.kaggle.com/datasets/PromptCloudHQ/jobs-on-naukri
- **Data:** Job titles, salaries, locations.
- **Tasks:** Clean missing salaries, compute median salary, visualize by job role.

16. Covid-19 Data (Johns Hopkins University)

- **URL:** https://github.com/CSSEGISandData/COVID-19
- **Data:** Daily confirmed cases, deaths, recoveries by country.
- **Tasks:** Fill missing dates, compute moving averages, plot growth curves.

17. Indian Government Open Data

- **URL:** https://data.gov.in/
- **Data:** Agriculture, health, transport, education statistics.
- **Tasks:** Clean mixed formats, compute regional summaries, visualize trends with maps.

18. Eurostat

- **URL:** https://ec.europa.eu/eurostat/data/database
- **Data:** EU statistics (economy, population, trade).
- **Tasks:** Handle missing/misaligned years, compute averages, plot trends per country.

Good Practice Categories

- **Entertainment:** IMDb, Billboard, OpenLibrary
- **Sports:** Basketball Reference, ESPN, Olympic Medals
- **Economy:** World Bank, OECD, Eurostat, UN Data
- **Science/Health:** COVID-19 (Johns Hopkins), WHO data.
- **E-commerce:** BooksToScrape, Airbnb (InsideAirbnb).