VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

B.Tech Final year (2022 Batch) Fall Semester 2025-26

BCSE207L – Programming for Data Science

Seminar Series

Domain	
Relevance	
Title	
Project Member(s):	Insert Photo (Member1) / REGNO: / NAME: / SIGNATURE: Insert Photo (Member2) / REGNO: / NAME: / SIGNATURE:

November 2025

1. Abstract (100–150 words)

Summarize the purpose, data sources, methods (R web scraping, preprocessing, database, web API, visualization), and key findings.

Example: “This project extracts unstructured data from PDF, blog, and Wikipedia pages using R web scraping libraries, converts them into structured format, stores them in SQLite, and visualizes dynamic insights via an interactive web API.”

2. Introduction

- Background and relevance of the chosen topic/domain
- Project objectives (in bullet points)
- Workflow overview (briefly describe data → processing → visualization steps)

(Include a simple flow diagram if possible)

3. Data Collection (Web Scraping Phase)

- List the 3 different sources (PDF, blog, wiki, etc.)
- Mention the R packages used:
    - pdftools, rvest, httr, XML, etc.
- Describe:
    - How data was extracted (methodology for each source)
    - Key challenges (encoding, noise, missing tags, etc.)

(Attach 1 sample raw data record or snippet)

4. Data Preprocessing and Transformation

- Cleaning steps (remove stopwords, punctuations, duplicates, NA values, etc.)
- Conversion from unstructured → structured → fused structured data
- Tools: tidyverse, dplyr, stringr, tidyr, etc.
- Describe how data was merged and formatted into a single dataframe.

(Show a before/after table — raw vs cleaned data)

5. Database Creation and Storage

- Type of database: SQLite / MySQL / PostgreSQL
- Method: using RSQLite / DBI package
- Structure: Table name, columns, data types
- Demonstrate storing and retrieving data from the database

(Include one CREATE TABLE or INSERT command snippet)

6. Exploratory Data Analysis (EDA)

- Tools: ggplot2, plotly, dplyr, summarytools
- Include:
    - Descriptive statistics (mean, median, frequency, etc.)
    - Graphs (bar, histogram, boxplot, etc.)
    - Insights (2–3 bullet points from EDA)

(Visuals should be clear and labeled)

7. Web API Development

- Describe how the API was created and used:
    - R packages: plumber, shiny, or RestRserve
- API endpoints:
    - /addData, /deleteData, /getData, /visualize
- Mention:
    - How the API dynamically updates the database
    - How the visualization responds to changes

(Include code snippet or URL of your local API endpoint)

8. Interactive Visualization

- Tools: plotly, shiny, rbokeh, highcharter, etc.
- Show how user interactions or API calls update the plots
- Include screenshots or embedded visuals

(Example: "When a record is deleted via API, the chart automatically updates to reflect the new totals.")

9. Results and Insights

- Discuss the findings and patterns revealed from the visualizations
- Interpret trends, correlations, or outliers
- Relate the results back to your project objective

10. Conclusion and Future Work

- Key outcomes achieved
- Technical learning outcomes (e.g., using plumber, working with databases, R visualization)
- Future enhancements (e.g., deploying API to the web, using authentication, or live dashboards)

11. References

Use APA or IEEE format for:

- Datasets
- Blog or wiki links
- R library references
  Example:
- [1] Hadley Wickham et al. (2019). *R for Data Science*. O'Reilly.
- [2] https://shiny.posit.co/
- [3] https://www.tidyverse.org/

Appendix

- Full code snippets
- API testing screenshots (e.g., Postman or curl)
- Link to GitHub repository

Report Formatting Guidelines

- Font: Times New Roman, Size 12
- Line spacing: 1.5
- Margins: 1 inch
- Page limit: 6–10 pages
- Include at least two visuals and one database screenshot