

ID511M – Data Curation and Visualization
Laboratory Exercise: From the NIRF Website to a Working Database
Duration: 3 hours | Tools: Google Sheets, Python (pandas, Plotly)

1. Background and Task Overview

Real analysis rarely begins with a tidy CSV handed to you; it begins with a website. Since 2016, the National Institutional Ranking Framework (NIRF) has published yearly rankings of Indian engineering institutions on five parameters — Teaching, Learning and Resources (TLR), Research and Professional Practice (RPC), Graduation Outcomes (GO), Outreach and Inclusivity (OI), and Perception (PR) — together with an overall score and rank. The data lives as year-wise tables on nirfindia.org, with institute names, column layouts and list lengths that drift from year to year.

In this laboratory you will carry the data through its full life cycle: collect it from the NIRF website, organize it into a well-designed Google Sheet that acts as your database, connect that Sheet to Python, and only then curate and visualize with pandas and Plotly. Each stage feeds the next — a sloppy Sheet in Part A will punish you in Parts B and C.

2. Part A – Collection and Database Construction in Google Sheets (55 minutes)

- A1. Scope: the Engineering category, ranks 1–100, for the five most recent ranking years. Locate the year-wise ranking pages on nirfindia.org. You may collect the tables by any honest means — `pandas.read_html` on the ranking page, copy-paste, or the downloadable PDFs — but record in your notebook which method you used for each year and one problem you hit.
- A2. Design your database as one Google Sheet with: (i) a worksheet named *data* holding a single combined table — one row per institute per year with columns `institute_id`, `institute_name`, `city`, `state`, `year`, `tlr`, `rpc`, `go`, `oi`, `pr`, `overall_score`, `rank`; and (ii) a worksheet named *dictionary* describing every column, its type, its valid range and its source page. Do not keep one worksheet per year — a database table, not a pile of printouts.
- A3. Enforce quality inside the Sheet itself: apply data-validation rules so that scores must be numbers in `[0, 100]`, rank a positive integer, and year one of your five collected years. Add a conditional-formatting rule that highlights any cell violating these, and fix whatever lights up.
- A4. Set sharing to “Anyone with the link can view” and note the spreadsheet ID from the URL. You will submit this link; the timestamped version history of the Sheet is your evidence of your own collection work.