

Data Project Ideas

AI and Data Analytics · Nayanta University

Course handout

September 2026

- [Reading the availability labels](#)
- [Cricket and the IPL](#)
- [Education and admissions](#)
- [Indian elections and politics](#)
- [Cinema and music](#)
- [Food and restaurants](#)
- [Air quality, weather, and the environment](#)
- [Demographics and districts](#)
- [Markets and the economy](#)
- [Popular culture and the internet](#)
- [Data you can collect yourselves](#)
- [Choosing a workable project](#)

Disclaimer: This handout was generated by Claude and has not been properly verified by Navin. Treat its suggestions, descriptions, and dataset links as starting points; independently check each source's licence, coverage, fields, and suitability before using the data.

A strong data project starts with a question, not with a technique. Choose a subject you genuinely care about, turn your curiosity into a question that data can answer, and then check whether the required data is available and usable.

This handout is a starting point, not a list of prescribed topics. You may adapt one of these questions, combine ideas, or propose a different investigation.

Reading the availability labels

- **Ready:** Public data is relatively easy to download and begin using.
- **Some work:** Useful data exists, but may require cleaning, combining files, extracting tables, or careful collection.

- **Difficult:** Access is restricted, incomplete, or likely to require permission.

Every named source below is linked and was checked by Claude on **2 September 2026**. **Official** means that the link is maintained by the organisation that publishes the data. **Public mirror** means a reusable copy maintained by somebody else; check its date, licence, and description before using it. **Explorer** means that the data can be viewed or queried online but may need extra work to download.

Cricket and the IPL

Possible data

- [Cricsheet IPL match data](#) — **Ready · official project**. Download IPL ball-by-ball files in CSV, JSON, YAML, or XML.
- [IPL auction and match compilation on Kaggle](#) — **Ready · public mirror**. It combines auction data with Cricsheet match data; read the data card and note the years covered.
- [ESPNcricinfo Statsguru](#) — **Some work · explorer**. It is useful for aggregate player and team records, but it is not a single clean download.

Questions you could investigate

- Does winning the toss actually affect the probability of winning?
- How large is home advantage in the IPL?
- Are expensive auction purchases associated with better performance?
- Did the impact-player rule change scoring patterns or team strategy?
- Which phases of an innings contribute most to differences between winning and losing teams?

Education and admissions

Possible data

- [JoSAA opening and closing rank archive](#) — **Ready · official**. Query by year, institute type, institute, programme, and seat category; save the tables you use.
- [MCC undergraduate counselling archive](#) — **Some work · official**. Download NEET-UG seat matrices, allotment results, and related documents by year and round.
- [NIRF 2025 overall ranking and institutional data](#) — **Ready · official**. Rankings, scores, and institution-level submissions are linked from the table.
- [CBSE examination statistics](#) — **Some work · official**. Year-wise Class X and XII aggregate results are in PDF reports; this is not student-level data.

Questions you could investigate

- How have closing ranks for particular programmes changed over time?
- Do institutes in large cities tend to close at different ranks from institutes elsewhere?
- How does the apparent branch-versus-college trade-off change across years?
- Do changes in rankings correspond to changes in student preferences?
- Which conclusions are robust to category, programme, and year—and which are not?

Indian elections and politics

Possible data

- [Lok Dhaba](#), from the Trivedi Centre for Political Data at Ashoka University — **Ready · research dataset**. It provides cleaned constituency- and candidate-level election data and a downloadable [codebook](#).
- [MyNeta](#), maintained by the Association for Democratic Reforms — **Ready · public-interest database**. It includes declared assets, education, and criminal cases from candidate affidavits.
- [Election Commission of India statistical reports](#) — **Some work · official**. Download election results and reports; formats vary across elections and years.

Questions you could investigate

- Are candidates' declared assets associated with their probability of winning?
- Does the number or type of declared criminal cases predict electoral success?
- Is voter turnout related to the winning margin?
- Are incumbents advantaged, and does that vary by state or election type?
- How have candidate education, wealth, or representation changed over time?

Take particular care here: association is not causation, affidavit data is self-declared, and politically sensitive claims need precise wording.

Cinema and music

Possible data

- [IMDb non-commercial datasets](#) — **Ready · official**. Daily compressed TSV files cover titles, genres, ratings, votes, and contributors; follow IMDb's non-commercial-use terms.
- [FilmDB's India box-office tables](#) — **Some work · explorer**. Tables include Indian gross, worldwide gross, budgets, and opening figures. They are estimates rather than a canonical industry dataset, so record the table, date, units, and cited sources you use.

- [Spotify Charts](#) — **Some work · official explorer**. It provides ranked streaming charts by place and date; sign-in or manual export may be required.

Questions you could investigate

- Is there evidence of a sequel advantage at the box office?
- Which genres receive high ratings but relatively small audiences?
- How do ratings change as the number of votes grows?
- Are films in different Indian languages gaining or losing relative visibility?
- Do star power, budget proxies, or release timing predict commercial performance?

Food and restaurants

Possible data

- [Zomato Bangalore Restaurants on Kaggle](#) — **Ready · public snapshot**. It has more than 50,000 records and fields for location, cuisines, cost, rating, votes, and service options. Treat it as a dated Bangalore snapshot, not a current national listing.
- Your own campus-food survey or observation log — **Ready to collect**. No pre-existing dataset is needed: define the variables, dates, locations, and sampling plan before collecting responses.

Questions you could investigate

- Does price predict rating?
- Do vegetarian-only restaurants receive different ratings after accounting for location and price?
- Which cuisines cluster in which neighbourhoods?
- Are ratings associated with the number of reviews?
- How strongly do food ratings vary by meal, day, location, or respondent?

Air quality, weather, and the environment

Possible data

- [CPCB's National Air Quality Index dashboard](#) — **Some work · official explorer**. Use it to inspect current station-level pollutant observations and verify station names and provenance.
- [Air Quality Data in India on Kaggle](#) — **Ready · public mirror**. It packages CPCB observations from 2015–2020 in analysis-friendly files; compare metadata with the official dashboard.
- [IMD Pune climate and rainfall archive](#) — **Some work · official**. It links daily rainfall, cumulative rainfall, monsoon, and all-India temperature series in several formats.

- Your own local measurements — **Ready to collect**. Use a fixed instrument, place, time, unit, and procedure so observations are comparable.

Questions you could investigate

- How large is the air-quality change around Diwali, and does it vary by city or year?
- How are rainfall, temperature, wind, and air pollution related?
- Which monitoring stations frequently have missing observations?
- Do weekday and weekend pollution patterns differ?
- How different are city-level conclusions from station-level conclusions?

Demographics and districts

Possible data

- [NFHS-5 district factsheet data on data.gov.in](#) — **Ready · official**. Download the spreadsheet or use the API for district indicators. A documented [cleaned CSV mirror](#) is convenient for starting analysis.
- [Census of India tables](#) — **Ready or some work · official**. Download Excel tables for 1991, 2001, and 2011, or use the [Census API documentation](#) for targeted extracts. Match geographic units carefully across years.
- [data.gov.in](#) — **Variable · official catalogue**. Search for a specific dataset, then check its publisher, update date, fields, licence, and documentation before choosing it.

Questions you could investigate

- Which districts perform much better or worse than their state average?
- How are female education, nutrition, sanitation, and health outcomes related?
- Which household assets spread fastest, and where?
- Do state averages hide large differences among districts?
- Which relationships remain after accounting for urbanisation or income proxies?

These datasets contain real social outcomes. Avoid deficit-based descriptions of people or places, and do not turn a correlation into a causal story.

Markets and the economy

Possible data

- [NSE historical reports](#) — **Ready or some work · official**. Download daily and monthly capital-market archives. Keep the adjusted/unadjusted-price choice explicit.
- [RBI Data Releases](#) and the [Database on Indian Economy](#) — **Ready or some work · official**. They cover inflation, policy and deposit rates, exchange rates, banking indicators, and downloadable time series.
- [RBI Bulletin data tables](#) — **Some work · official**. Use the monthly gold-price table alongside CPI and deposit-rate series from RBI; align frequency, units, and dates before comparing returns.

Questions you could investigate

- What would ₹10,000 invested ten years ago be worth under different choices?
- How much does the answer change when the starting date changes?
- Which investment had the largest drawdown or most volatile path?
- Did an apparent return beat inflation after costs and taxes?
- How do headlines about the economy compare with the underlying time series?

Historical performance is not investment advice. State all assumptions and avoid presenting a favourable time window as a universal result.

Popular culture and the internet

Possible data

- [Wikimedia Pageviews API](#) — **Ready · official API**. Retrieve daily or monthly views by article, project, access method, and agent from July 2015 onward.
- [MediaWiki Revisions API](#) — **Ready or some work · official API**. It exposes revision timestamps, users, sizes, comments, and other metadata; interpret edits cautiously.
- [YouTube Trending Video Dataset on Kaggle](#) — **Ready · public archive**. It contains daily trending records for India and ten other regions, collected through the YouTube API; check its coverage dates.
- [Google Trends](#) — **Some work · official explorer**. Export interest-over-time tables, remembering that values are relative indices rather than search counts.

Questions you could investigate

- Which events produce the largest spikes in attention to a person or topic?
- How seasonal is attention to festivals, examinations, films, or sports leagues?
- How long does attention persist after a major event?
- Which controversial pages experience bursts of editing?
- Do attention patterns differ across language editions of Wikipedia?

Data you can collect yourselves

Designing a dataset can be more instructive than downloading one. The hard part is turning a vague idea into measurable variables and collecting observations consistently. These projects intentionally have no dataset link: **you create the dataset**. Preserve the blank form or data-entry sheet, the completed data, and a short data dictionary so somebody else could understand what each row and column means.

Surveys

- Study routines, sleep, commuting, languages, reading, music, or media habits.
- Ratings of meals, common spaces, campus services, or events over several days.
- Estimates made before an event, later compared with actual outcomes to measure calibration.

Keep surveys short, voluntary, and anonymous. Avoid collecting sensitive personal information unless it is genuinely necessary and explicitly approved.

Physical measurement

- Height, arm span, or foot length, with a consistent measurement procedure.
- Walking time or step count along several campus routes.
- Plant height, leaf count, temperature, rainfall, noise, or bird observations at fixed places and times.
- Pendulum period versus length, or another controlled physical experiment.

Observation in the world

- Bus arrival times compared with the schedule.
- Prices for the same product across vendors or days.
- Counts of pedestrians, bicycles, two-wheelers, cars, or buses at a fixed point.
- Queue lengths and waiting times at different times of day.

Text and archival material

- Compare the front pages of several newspapers on the same dates.
- Compare word choice in a textbook, speech, novel, or news corpus.
- Examine the edit history of a Wikipedia article.
- Transcribe a short, well-defined sample of commentary or public speech and code repeated phrases.

Generated or simulated data

- Repeated dice or coin experiments.
- Monty Hall trials performed physically or simulated in code.
- Card-game outcomes recorded under a fixed set of rules.
- Synthetic data designed to test whether an analysis method can recover a known pattern.

Choosing a workable project

Before committing to a topic, write down:

1. **The question:** one sentence that can be answered with evidence.
2. **The unit of analysis:** a match, candidate, film, restaurant, district, day, person, observation, or something else.
3. **The variables:** what must be measured or obtained?
4. **The comparison:** between which groups, places, periods, or conditions?
5. **The data source:** who collected it, when, and for what purpose?
6. **The likely limitations:** what is missing, ambiguous, selected, or biased?
7. **The minimum viable result:** the smallest honest analysis that would still teach you something.

Start small. A well-answered narrow question is a stronger project than a sweeping claim supported by weak data. Use AI to help find sources, write and debug code, and explore alternatives—but verify the data, calculations, and claims yourselves.