


CS 328 Writing Assignment – 2025

Expectations

First, watch this talk for inspiration –  The best stats you've ever seen | Hans Rosling

Also, poke around the videos and [case studies](#) linked here — <https://www.callingbullshit.org/>.

Can also check <https://www.data-wto-viz.com/> for exploring different types of graphs suitable for the data.

You will start from a (mostly) raw dataset; it might have various imperfections e.g., missing data. Also, you might get multiple datasets representing different aspects of a sector. Your task is to write a data-driven summary that draws certain conclusions and presents arguments for the conclusions either visually or via summarizing tables or statistical tests. You will be judged on the kind of insights, and how well you present your argument (e.g., visual illustrations, any number crunching e.g., correlations derived, statistical tests used to support your argument, etc.). The best format to submit is a Jupyter notebook, or you could make an interactive webpage.

Length – roughly 1000 words (excluding code).

Examples of such reports (much longer) –

https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/udise_201920.pdf

https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/aishe_eng.pdf

The structure of your investigation could be the following. Notice that the writeup need not follow this linear structure.

1. Give a summary of the data along with visual aids.
2. Posit a small set of hypothesis that you think is decidable using this data (e.g. state X does “better” than state Y in managing Covid).
3. Show how you will quantify your hypothesis.
4. Do analysis and settle the hypothesis, with suitable justification from the data.

Analysis Themes

The following are the analysis themes to choose from. For each, we have provided some data sources. Feel free to consult other ones. All code must be written by the group, no copying allowed.

1. Understanding our campus life

a. Food Wastage Dataset 2019-2020

Dataset Link:

<https://drive.google.com/file/d/1vGYG-J1lFYVI04j9V8TgR0bWUQpvcVv/view?usp=sharing>

The dataset is from Jaiswal Dining Hall for the year 2019-2020. For each day, there are four values, the amount of food (in kgs) wasted in Breakfast, Lunch, Snacks and Dinner. You can choose a specific time period for evaluation (whole dataset, one year, few months, etc.)

This dataset can be combined with the monthly mess menu and different events in the calendar (summer holidays, exams, etc.) to get insights about food habits. Further suggestions can be made for strategies/insights from the data to reduce food wastage.

b. Campus energy consumption data – [2020+2021](#) (monthly aggregate). For reference and possible use, here is the [energy consumption data](#) for all Indian states.

2. Telling the story of India

A lot of official datasets are available at – [Open Government Data \(OGD\) Platform India](#)

You need to decide on a sector and choose a dataset from there. As an example, for the education sector, you can navigate to the following page for the datasets.

<https://data.gov.in/dataset-group-name/school-education-statistics>

3. Charting the growth of AI in different countries

Can we take the top ML conferences in the last few years, get the author institutions and countries, and summarize the leaders / rising institutions/trends? E.g. this report – [The race to the top among the world's leaders in artificial intelligence](#)

Dataset for NeurIPS – [All NeurIPS \(NIPS\) Papers | Kaggle](#)

Other conferences – <https://github.com/martenlienen/icml-nips-iclr-dataset>

4. World inequality and how it has changed

World inequality dataset – <https://wid.world/>

5. How are different countries developing?

Various country-level statistics collected by UN. You can ask questions e.g. how well are the different countries doing on the various developmental goals and how are they making progress.

<https://unstats.un.org/home/>

<https://unstats-undesa.opendata.arcgis.com/>

6. Looking at Primary Education across multiple countries

PISA is a well-known test for measuring how well a particular country is doing in providing quality education. There are data available across years at –

<https://www.oecd.org/pisa/data/>

You can talk about how the indicators for specific countries have changed over time.

Participating countries are available at

<https://www.oecd.org/pisa/aboutpisa/pisa-participants.htm>

7. How are vaccination drives going + the inequalities underneath

Dataset on worldwide covid19 vaccinations

<https://www.nature.com/articles/s41562-021-01122-8>

A broader dataset for covid-19 (for current vaccination theme) can be found at:

<https://github.com/owid/covid-19-data/tree/master/public/data>

<https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

8. Effects of Economy, natural resources distribution, and government policies on CO2 emissions by countries

CO2 emissions dataset: <https://github.com/owid/co2-data>

9. Prevalence and reasons of childhood stunting in India (malnutrition, child diseases, sanitation, etc.)

World Bank datasets:

<https://data.worldbank.org/indicator/SH.STA.STNT.ZS?locations=IN>,

<https://data.worldbank.org/indicator/SH.STA.ORTH?locations=IN>

WHO datasets:

<https://www.who.int/data/gho/data/themes/theme-details/GHO/gho-nutrition>

10. Neglected Tropical Diseases in Africa and their eradication

WHO datasets: <https://www.who.int/data/gho/data/themes/neglected-tropical-diseases>

11. Maternal and Reproductive Health

WHO datasets:

<https://www.who.int/data/gho/data/themes/maternal-and-reproductive-health>

12. Road Safety

a. WHO datasets: <https://www.who.int/data/gho/data/themes/road-safety>

13. Covid-19 spread and how it affected different countries/regions.

a. State-wise data for India – WHO datasets: <https://www.mygov.in/covid-19>

b. Countrywise data –

<https://www.kaggle.com/code/abhinand05/covid-19-digging-a-bit-deeper>