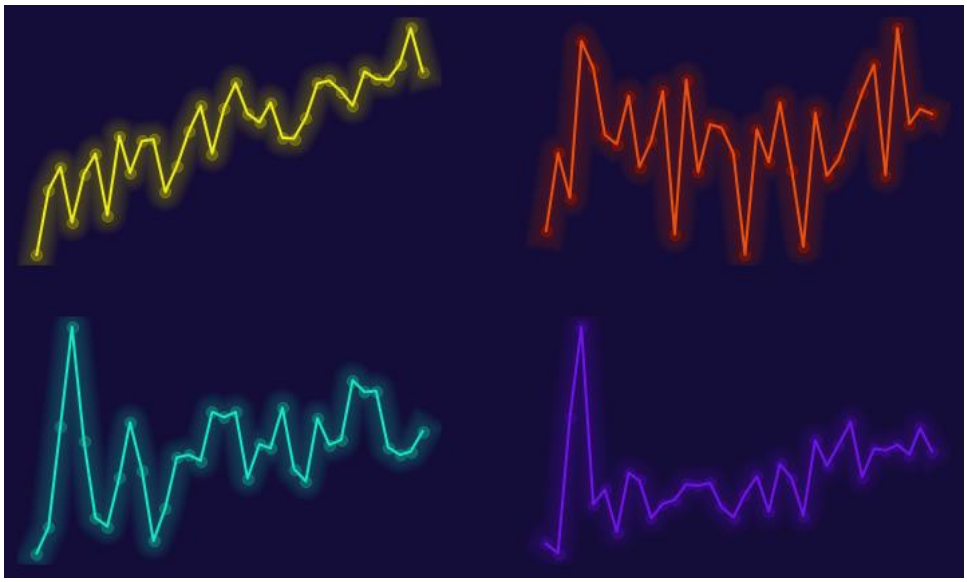
Social Data Science Student Hackathon 2023 Data handbook

Dept. Of Network and Data Science, Central European University, Vienna



CEU Department of Network and Data Science | CENTRAL EUROPEAN UNIVERSITY

SAI

Welcome to the Social Data Science Student Hackathon 2023 organized by DNDS CEU! This document contains everything you need to know about the hackathon's dataset and challenges. Below you will find an overview of the origins of the data, instructions how to access it, detailed description of each data file, and information about how to submit your solutions to the challenges and how your submissions will be assessed. If you have questions, you can contact us on Slack or via email at sdshackathon@ceu.edu.

Food security

The UN reports that world-wide over 700 million people faced hunger in 2022, and even more, 2.4 billion people or 29.6% of the total population, experienced moderate to severe food insecurity. The main forces driving food insecurity are identified as climate change, conflicts and social inequality. The World Food Programme (WFP) was funded by the UN with the mission to fight hunger by providing emergency food relief and working with communities to improve nutrition and build resilience. To better collect and allocate resources the WFP continuously monitors levels of food insecurity via representative household surveys. In addition to these direct measurements of food security, the WFP also collects secondary data such as market prices and conflicts or drought levels to understand the causes and to forecast food insecurity. In this hackathon, you will analyze data published by the WFP focusing on Yemen, a country torn by civil war with majority of its population in need of humanitarian aid.

Data description

The dataset consists of four data files about levels of food insecurity, market prices, rainfall, and conflicts in Yemen provided on a sub-national level covering Yemen's 22 administrative units. The data is real-world data with all its messiness: although we pre-processed the data for you, there are missing datapoints, the files do not cover exactly the same period and the order of the data might be different in the different files.

sdsh2023-food_security.csv

Each row provides the level of food insecurity for a given month and administrative unit in Yemen from July 2018 and April 2021. The column descriptions are the following:

- *date*: Date of the observation
- *admin*: Name of the administrative unit in Yemen
- *insuff_food*: The fraction of the population facing moderate to severe level of food insecurity based on household surveys

sdsh2023-market.csv

Each row provides the average market price of a commodity for a given month and administrative unit. The column descriptions are the following:

- *year, month*: Date of the observation
- *admin*: Name of the administrative unit in Yemen
- *category:* Category of the commodity, such as "cereals and tubers" or "vegetables and fruits"
- *commodity*: Name of the commodity such as "Wheat flour" or "Salt", contains also less conventional commodities such as price of labour or exchange rates.
- *price:* The price of one unit of the commodity in the local currency.
- *usdprice:* Same as *price* but given in US dollars.

sdsh2023-rainfall.csv

Each row provides indicators of rainfall and drought for a given month and administrative unit. The column descriptions are the following:

- *year, month*: Date of the observation
- *admin*: Name of the administrative unit in Yemen
- *average_rainfall:* The average rainfall for the given month, averaged over many years
- *rain_anomaly_1month:* One month observed rainfall compared to the average rainfall in percentage. For example, if *rain_anomaly_1month* is 150% then 50% more rain fell during the last month than the average.
- *rain_anomaly_3month:* Same as *rain_anomaly_1month* only considering the past 3-month period.
- *ndvi_anomaly:* NDVI stands for normalized difference vegetation index, it is a quantity derived from satellite imagery which measures the health of the vegetation of an area. NDVI anomaly compares the observed NDVI for the given month to its average.

sdsh2023-conflict.csv

Each row corresponds to a conflict event. The column descriptions are the following:

- *date*: Date of the event
- *admin*: Name of the administrative unit where the event happened
- *event:* Type of the event, e.g., “Battles” or “Riots”
- *fatalities*: Number of fatalities during the event
- latitude, longitude: Location of the event

Challenge 1 – Infographics

The goal of this challenge is to explore the dataset, identify patterns and create a data visualization or an infographic highlighting features of the data. The core data set provided by us must be the focus of the visualization; however, it can be combined with data that you obtained from other sources, for example, geographic data. And, as always, remember: correlation is not causation!

Winner selection: Submissions are ranked based on creativity of the analysis, clarity of presentation and aesthetics. A good visualization is clear about what question it poses and how it uses the data to answer it. Winning submissions are selected by a panel consisting of a faculty member, a postdoctoral researcher and a PhD student from the Department of Network and Data Science.

Submission: To ensure fairness, the panel will judge the submissions without knowing your name or affiliation; therefore, only include your team name in the graphic. Each team must submit an image file (pdf, png, jpeg) and a maximum 300-word caption via email to sdshackathon@ceu.edu. The subject of your email should include "infographics submission" and the name of your team. In case of multiple submissions, the last email before the deadline is considered final.

Challenge 2 – Prediction

The World Food Programme monitors levels of food security by conducting continuous surveys representative of the target population. Such efforts are costly, require a lot of resources and the target population may be difficult to reach. Secondary data such as market prices and weather statistics, on the other hand, are easier to access. In our challenge, we explore the possibility of predicting the level of food insecurity relying on such secondary data.

The file *sdsh2023-food_security.csv* contains the column *insuff_food* providing the fraction of the population facing moderate to severe level of food insecurity based on household surveys for each of the 22 administrative units in Yemen at a monthly resolution. The file *sdsh2023-food_security.csv* covers the period from July 2018 to April 2021, while the secondary data is provided beyond this period as well. Your task is to use this historical data and the secondary data to predict the *insuff_food* for the one-year-long period from May 2021 to April 2022. You are allowed to include publicly available secondary data from alternative sources.

Winner selection: Submissions are ranked based on their prediction's coefficient of determination compared to the actual data.

Submission: Each team must submit a csv file named *teamname-prediction.csv* with four columns: *year*, *month*, *admin*, *insuff_food*. Each row in the file must provide a prediction *insuff_food* for the month given by *year* and *month* and the administrative unit given by *admin*. Your submission, therefore, must contain 264=22 x 12 rows corresponding to the 22 administrative units and 12 months.

You do not need to submit your code; however, we do ask for a short maximum 200-word paragraph describing the main idea behind your solution. You must submit via email to sdshackathon@ceu.edu. The subject of the email should include "prediction submission" and the name of your team. In case of multiple submissions, the last email before the deadline is considered final.

Good luck and happy coding!
The organizers