Project Part B: Data Science in Action!

Genocide in Gaza: Telling the Story through Data Science!

The genocide that is being committed in Gaza strip since October 2023 by the Israeli occupying forces is unparalleled in the recent history. More than 50k Palestinians were killed; 36% of them are children and 24% are women.[1] As data scientists, we are obliged to use our data science skills to tell the story of those who were murdered in our own way! In this project, you will be exploring and analyzing data related to that genocide in Gaza.

Important General Notes

- Read the description of the project **in full** and let me know if you have any questions right away!
- Start early!
- Only **one** member of the team submits, not all.
- You will submit both a report and full code.
- You are not allowed to change/add anything in/to the original datasets. When testing your code, I will use my own copy.
- You are not allowed to use external libraries for data science functionalities, such as pandas. **If you are in doubt, please ask me**.
- Discuss the project only with your teammates. Discussions that are across teams are not allowed.
- There will be a discussion with every group after submission of the project, in which the group members will present their work and answer questions about it.

Academic Dishonesty

Remember that "من غشنا فليس منا". I am very strict about this issue. You are not allowed to get any external help other than from your teammates. I will be checking your code against other submissions in the class for logical redundancy. If you copy someone else's code and submit it with minor changes, I will know. These cheat detectors are quite hard to fool, so please don't try. I trust you all to submit your own work only; please don't let me down. If you do, I will pursue the strongest consequences available according to QU policies.

Datasets

You are given four datasets in 5 files (3 CSV files + 1 JSON file + 1 PDF file):

- **Daily Casualties – Gaza**: This dataset provides daily values for those killed and injured in the Gaza Strip since October 7th, 2023. It includes number of people killed, injured, and number of events leading to multiple deaths. It also has breakdown by women, children and if medical personnel, first responder, or press. You can check the description of the attributes of each record here.
- **Daily Casualties - West Bank**: This dataset provides daily values for those killed and injured in the West Bank since October 7th, 2023. It includes number of people killed and injured and how many

[1] https://data.techforpalestine.org/

of those were children. Also includes number of Israeli settler attacks. You can check the description of the attributes of each record here.

- **Killed in Gaza**: List of known victims with their Arabic name, English name, birth date, gender, and ID number. It is derived from a list issued by health officials in Gaza. You can check the description of the attributes of each record here.
  Note: This data is last updated on September 21, 2024. A more updated list is provided, but in PDF format, in "killed-in-gaza_moh_2025-03-23.pdf". You will need to convert it to CSV (check preprocessing below).
- **Infrastructure Damaged**: This dataset has daily reports on the impact of IDF attacks on civilian infrastructure in Gaza, resulting in partial or total damage. You can check the description of the attributes of each record here. This data comes in a JSON file format.

The main source of the datasets is TechForPalestine data repository. The version you will use was downloaded on April 12th 2025 and it is attached to this project description.

1. Preprocessing [6 points]

In this first stage, you will get familiar with the data and clean it up by dealing with some special cases. Among that:

- Convert the PDF list in the "killed-in-gaza_moh_2024-09-21.pdf" file to CSV.
- Convert the JSON list in the "infrastructure-damaged.json" file to CSV.
- Discover missing information and deal with them *logically*.

2. Exploratory Analysis & Visualization [12 points]

In this stage, you will do some explorations over the data you have, trying to find interesting observations and explore potential associations.

To warm up, try initially to plot some raw data, such as the number of casualties per day or per month, etc. This can give you ideas about what you would like to explore further.

The goal in this stage is to show **six** **interesting** but **different**, plots out of the given datasets, from which you can draw some "interesting" insights (patterns, trends, associations, etc.). Interesting insights indicate insights that are **non-trivial**, useful to know, and well-motivated, that are typically drawn from plots that required more processing and analysis.

All types of visualization plots that we studied are considered (line, bar chart, histogram, overlaid chart, scatter, etc.). Plots generally become more interesting if data is analyzed in some way (for example, aggregated, grouped, etc.) compared to plotting raw data.

From the plots you will show in this part (and even others plots), you will come up with questions that you need to answer in the next part.

3. Hypothesis Testing [12 points]

In this stage, you will pose **three** **different** **interesting** questions (similar to the questions we studied in class during the course) and you will then answer each using hypothesis testing. Interesting questions

indicate questions that are non-trivial, useful to answer, and well-motivated. The questions can involve one sample of observed data (as we studied in session 7) or two samples (as we studied in session 8).

For each question, design (and report) a complete hypothesis testing procedure. For each question:

- State and explain the two hypotheses (including where the chance is).
- Choose the test statistic (including why you chose it and the meaning of its small/large values).
- Design a simulation experiment to get the empirical distribution of the test statistic under the null hypothesis.
- Perform the test and make a conclusion accordingly using the conventional significance levels.

When you think of a question, it is very important to think about where the *chance* (or randomness) is, and thus where the random *model* is.

When to Submit

You must submit your final project (check below) by **Sunday May 4th 11:59pm**.

What to Submit

For the final submission, you are required to submit two things:

a. **Report (PDF)**: ***documenting*** the process and results (including plots of course) for each part.
   - Motivation behind the plots or tests.
   - Summary of results (plots, small tables, etc.)
   - ***Role/contribution of each team member*****. "We all worked together" or "Everyone helped" or alike is not acceptable. There must be clear responsibilities and contributions.**

b. **Jupyter notebook(s)**: showing your ***full code*** for all the steps in all parts.
   - Keep all the code needed (to process the data, generate any plot, simulation etc.). You have to show the code that generates all intermediate tables/data/charts/etc.
   - Your code must be well-documented.
   - You must stick to the packages and the methods we used in the course, i.e., `numpy` (only functions we covered) and `datascience`. You can develop your own functions of course. Here are the list of references you need:
     - DataScience Package Documentation
     - Table Documentation
     - Python Reference Cheat sheet