CMPS 360 L51

Data Science Fundamentals

Spring 2025

Dr. Tamer Elsayed

Project Part B: Data Science in Action

Genocide in Gaza: Telling the Story through Data Science

Ganna - 202205763

Atrab - 202207802

Non - 000000000

Fatma – 202008126

Contents

[Introduction 3](#_Toc197365182)

[Preprocessing 3](#_Toc197365183)

[Datasets 3](#_Toc197365184)

[Daily Casualties – Gaza 4](#_Toc197365185)

[Daily Casualties – West Bank 5](#_Toc197365186)

[Killed in Gaza 5](#_Toc197365187)

[Infrastructure Damaged 6](#_Toc197365188)

[Exploratory Analysis & Visualization 6](#_Toc197365189)

[1. Comparing Total Reported Casualties with Subcategory Totals 6](#_Toc197365190)

[2. plot 2 7](#_Toc197365191)

[3. plot 3 7](#_Toc197365192)

[4. plot 4 8](#_Toc197365193)

[5. plot 5 8](#_Toc197365194)

[6. plot 6 8](#_Toc197365195)

# Introduction

In this part of the project, we are going to explore and analyze the provided data about the ongoing genocide in Gaza. We will try to find interesting observations and explore potential associations. We will also perform hypothesis testing.

# Preprocessing

* Converted the PDF list in the "killed-in-gaza\_moh\_2024-09-21.pdf" file to CSV.
  + Used an online tool that converts pdf to xlsx then used that to create csv (“killed-in-gaza.csv”).
  + <https://www.ilovepdf.com/pdf_to_excel>
* Converted the JSON list in the "infrastructure-damaged.json" file to CSV.
  + Used an online tool that converts json to xlsx then used that to create csv (“infrastructure-damaged.csv”).
  + <https://data.page/json/csv>
* All tables are checked to see which columns are missing data from them and the following results are found

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Total # of Records | Columns with Missing Values | Missing Counts |
| Daily Casualties - Gaza | 551 | report\_source, killed, killed\_cum, injured\_cum, med\_killed\_cum, press\_killed\_cum | 12, 27, 7, 15, 417, 409 |
| Daily Casualties - West Bank | 551 | verified.killed, verified.killed\_cum, verified.injured, verified.injured\_cum, verified.killed\_children, verified.killed\_children\_cum, verified.injured\_children, verified.injured\_children\_cum | 148, 147, 162, 160, 148, 147, 162, 160 |
| Killed in Gaza | 50,020 | None (No missing values detected) | 0 |
| Infrastructure Damaged | 534 | civic\_buildings\_\_destroyed, educational\_buildings\_\_destroyed, educational\_buildings\_\_damaged, places\_of\_worship\_\_mosques\_destroyed, places\_of\_worship\_\_mosques\_damaged, places\_of\_worship\_\_ext\_mosques\_damaged, places\_of\_worship\_\_churches\_destroyed, residential\_\_destroyed | 451, 445, 444, 453, 464, 174, 451, 428 |

Each missing column is dealt with separately as described below:

* report\_source from **Daily Casualties – Gaza** table, they are just filled with “missing” as there are only 12 missing and we cannot make up a source
* killed, killed\_cum, injured\_cum, med\_killed\_cum, press\_killed\_cum are all deleted from **Daily Casualties – Gaza** table as there are already extrapolated versions of all these columns being provided from the source

# Datasets

We have 4 different data sets, providing different data related to the genocide:

* Daily Casualties - Gaza
* Daily Casualties – West Bank
* Killed in Gaza
* Infrastructure damaged

All are initially csv files that we convert to tables as a first step.

The table names and a description of the values of each field are given.

## Daily Casualties – Gaza

This dataset provides daily values for those killed and injured in the Gaza Strip since October 7th, 2023. There are 551 days of reports from 2023-10-07 to 2025-04-09.

Table name: **gaza\_daily\_casualties**

|  |  |
| --- | --- |
| **field name** | **value** |
|  |  |
| report\_date | date in YYYY-MM-DD |
| report\_source | one of mohtel, gmotel, unocha or missing |
| report\_period | one of 24, 48, or 0, representing the hours length of the reporting period |
| ext\_massacres\_cum | extrapolated cumulative number of multiple fatality events |
| killed | killed persons on the given report date |
| killed\_cum | cumulative number of killed persons to the report date |
| ext\_killed | same as killed but extrapolated |
| ext\_killed\_cum | same as killed\_cum but extrapolated |
| ext\_killed\_children\_cum | extrapolated cumulative number of children killed to the report date |
| ext\_killed\_women\_cum | extrapolated cumulative number of women killed to the report date |
| injured\_cum | cumulative number of injured persons to the report date |
| ext\_injured | injured persons on the given report date |
| ext\_injured\_cum | cumulative number of injured persons to the report date |
| ext\_civdef\_killed\_cum | extrapolated cumulative number of emergency services killed to date |
| med\_killed\_cum | cumulative number of medical personnel killed to the report date |
| ext\_med\_killed\_cum | extrapolated cumulative number of medical personnel killed to the report date |
| press\_killed\_cum | cumulative number of journalists killed to the report date |
| ext\_press\_killed\_cum | extrapolated cumulative number of journalists killed to the report date |

## Daily Casualties – West Bank

This dataset provides daily values for those killed and injured in the West Bank since October 7th, 2023. There are currently 551 days of reports from 2023-10-07 to 2025-04-09.

Table name: **west\_bank\_daiy\_casualties**

|  |  |
| --- | --- |
| **field name** | **value** |
| report\_date | date in YYYY-MM-DD |
| verified.killed | killed persons on the given report date |
| verified.killed\_cum | cumulative number of confirmed killed persons to the report date |
| verified.injured | injured persons on the given report date |
| verified.injured\_cum | cumulative number of injured persons to the report date |
| verified.killed\_children | number of children killed on the given report date |
| verified.killed\_children\_cum | cumulative number of children killed to the report date |
| verified.injured\_children | number of children killed on the given report date |
| verified.injured\_children\_cum | cumulative number of children injured to the report date |
| killed\_cum | same as verified.killed\_cum but yet to be independently verified |
| killed\_children\_cum | same as verified.killed\_children\_cum but yet to be independently verified |
| injured\_cum | same as verified.injured\_cum but yet to be independently verified (see below) |
| injured\_children\_cum | same as verified.injured\_children\_cum but yet to be independently verified |
| settler\_attacks\_cum | cumulative number of attacks by settlers on civilians to the report date |
| flash\_source | either un or fill, see [March 25 update](https://data.techforpalestine.org/updates/west-bank-frequency/) for more detail |

## Killed in Gaza

This dataset is a list of over 50,000 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, published March 26, 2025 by Iraq Body Count.

Table name: **killed\_in\_gaza**

|  |  |
| --- | --- |
| **field name** | **value** |
| Index | index of victim in the table |
| Name | english name translation |
| Arabic name | original arabic name from the source list |
| Born | date of birth: string in YYYY-MM-DD format, or empty string if not available |
| Age | age as a number |
| Sex | string of one of m for male or f for female |
| ID number | victim’s ID number |

## Infrastructure Damaged

This dataset provides daily values for the damaged infrastructure resulting from Israel's assault on Gaza since October 7th, 2023. The last record is on 2025-03-22.

Table name: **infrastructure\_damaged**

|  |  |
| --- | --- |
| **field name** | **value** |
| report\_date | date in YYYY-MM-DD |
| civic\_buildings.destroyed | cumulative number of government buildings destroyed to the date |
| civic\_buildings.ext\_destroyed | civic\_buildings.destroyed but extrapolated (see below) |
| educational\_buildings.destroyed | cumulative number of educational buildings destroyed to the date |
| educational\_buildings.ext\_destroyed | educational\_buildings.destroyed but extrapolated |
| educational\_buildings.damaged | cumulative no. of educational buildings partially damaged to the date |
| educational\_buildings.ext\_damaged | educational\_buildings.damaged but extrapolated |
| places\_of\_worship.mosques\_destroyed | cumulative number of mosques destroyed to the date |
| places\_of\_worship.ext\_mosques\_destroyed | places\_of\_worship.mosques\_destroyed but extrapolated |
| places\_of\_worship.mosques\_damaged\* | cumulative number of mosques partially damaged to the date |
| places\_of\_worship.ext\_mosques\_damaged | places\_of\_worship.mosques\_damaged but extrapolated |
| places\_of\_worship.churches\_destroyed | cumulative number of churches destroyed to the report date |
| places\_of\_worship.ext\_churches\_destroyed | places\_of\_worship.churches\_destroyed but extrapolated |
| residential.destroyed | cumulative number of residential units destroyed to the date |
| residential.ext\_destroyed | residential.destroyed but extrapolated |

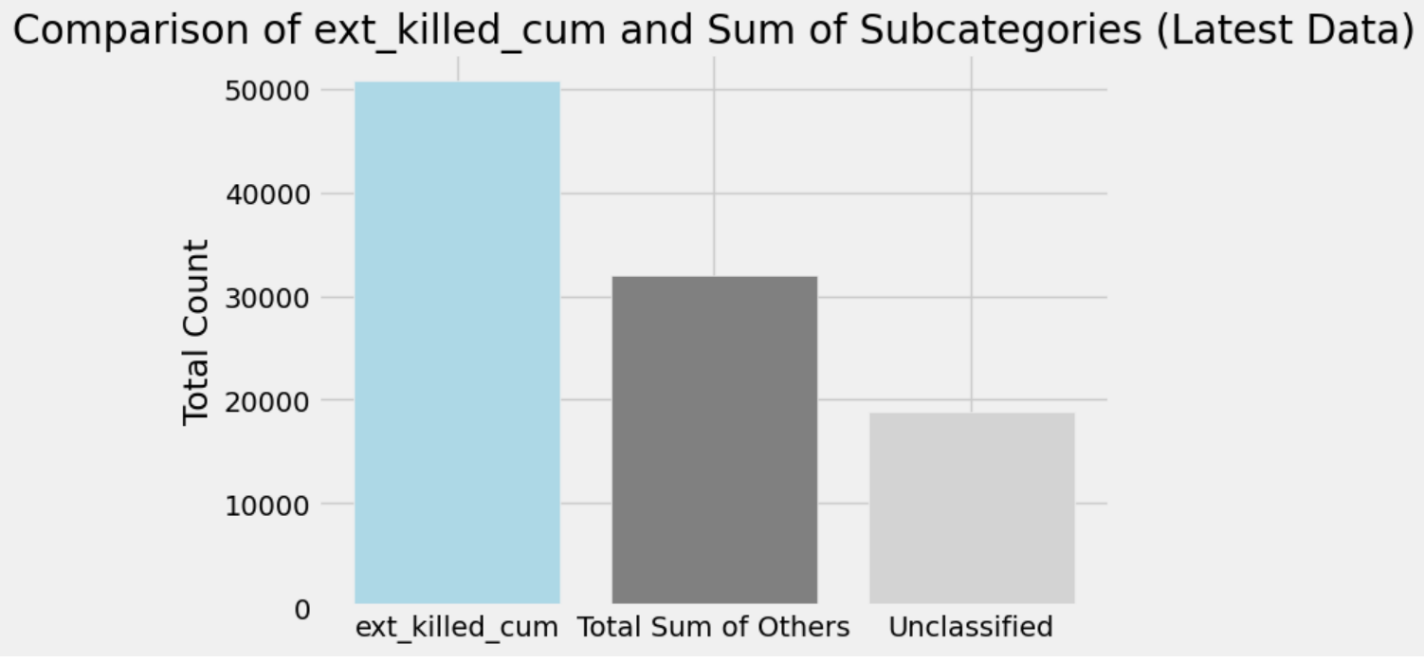
# Exploratory Analysis & Visualization

## 1. Comparing Total Reported Casualties with Subcategory Totals

comparing the total cumulative number of reported deaths with the cumulative counts of specific subcategories (children, women, civil defense, medical personnel, and press) as of the latest reported date in the "casualties\_daily\_gaza\_2025-04-12.csv" dataset. This comparison helps to visualize the proportion of deaths that have been categorized versus those that remain unclassified within these specific groups.

**Methodology:**

We first selected the relevant cumulative columns from the gaza\_daily\_casualties table: ext\_killed\_cum, ext\_killed\_children\_cum, ext\_killed\_women\_cum, ext\_civdef\_killed\_cum, ext\_med\_killed\_cum, and ext\_press\_killed\_cum. We then extracted the values from the last reported day to represent the cumulative totals up to that point. The "Total Sum of Others" was calculated by summing the cumulative counts of the subcategories. Finally, the "Unclassified" portion was determined by subtracting the "Total Sum of Others" from the ext\_killed\_cum. These values were then visualized using a bar chart to compare their magnitudes.



The substantial difference between the total externally reported deaths and the sum of the detailed externally reported subcategories suggests that a significant proportion of the casualties in Gaza, as of the latest report, have not been classified within these specific demographic or professional groups in the external reporting. This implies that the detailed breakdown by children, women, civil defense, medical personnel, and press, within the external reporting, accounts for only a fraction of this total.

## 2. plot 2

## 3. plot 3

## 4. plot 4

## 5. plot 5

## 6. plot 6