Data Science Fundamentals Project Report

Course Code: CMPS360

Semester: Spring 2025

Instructor: Dr. Tamer Elsayed

Project Title:

Genocide in Gaza: Telling the Story through Data Science!

Submitted by:

Fahrel Hidayat – 202206836

Mohd Muhtasim Bashar – 202205153

Zubair Bin Jashim – 202108715

Sheikh Hasin Ishrak – 202108209

Submission Date: 06/05/2025

Department of Computer Science & Engineering

Qatar University

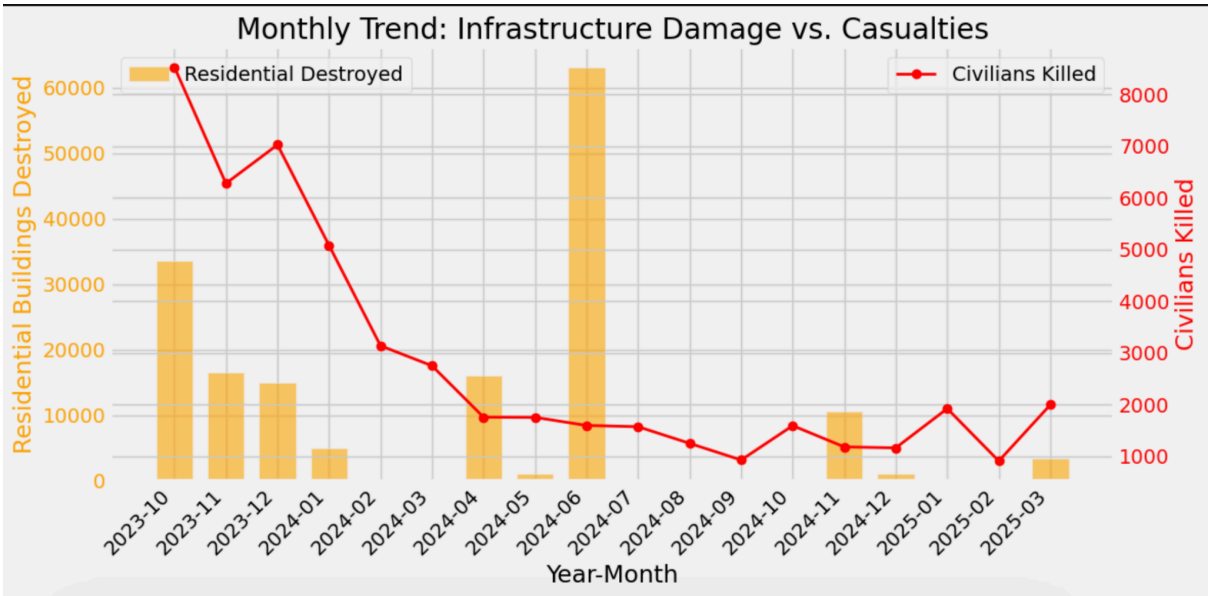
Doha, Qatar

1. Plots

Plot 1: Overlaid bar and line chart of monthly Civilian Infrastructure Damage vs Casualty Spikes.

Motivation Behind Plot 1:

The objective of this analysis is to examine the monthly correlation between civilian casualties and residential infrastructure damage in Gaza between October 2023 and March 2025. Placing these two metrics on each other, the plot attempts to look for patterns and connections that may show how conflict intensity and conflict nature evolve. Understanding whether spikes in infrastructure damage correlate with civilian casualties can provide insights into shifts in military tactics, the impact of ceasefires, and the broader humanitarian consequences of the war.



Plot Result Summary:

The overlaid bar and line chart illustrates the correlation between civilian infrastructure damage and casualty rates in Gaza from October 2023 to March 2025. October 2023 is marked by the highest civilian fatalities (around 8,000) and most residential building damage (over 30,000). This pattern of high casualties matching heavy damage persists through the last few months of 2023, reflecting the high-intensity initial phase of the conflict.

One notable point is the consistent decline in civilian casualties from early 2024 onwards, even as the destruction of buildings continues occasionally. Notably, February and March 2024 show a sharp drop in both casualties and infrastructure damage, coinciding with a ceasefire during that

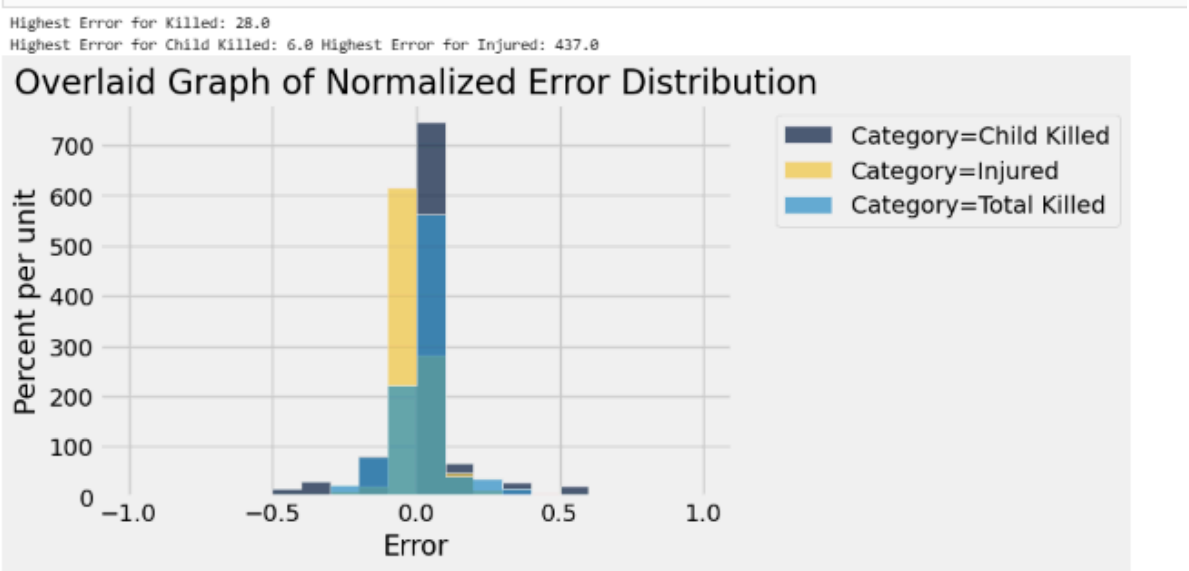
period. Similarly, early 2025 (January – February) also shows another ceasefire, which explains the absence of heavy damage during these months.

Most notably, in June 2024, infrastructure destruction rose significantly with no corresponding increase in civilian deaths to accompany it. This is a shift in the nature of attacks, rendered as targeted demolitions rather than high-casualty attacks.

Overall, since the initial period suggested a high correlation between damage and casualties, the following periods suggest that massive infrastructure destruction can occur independently, most likely due to strategic or political reasons, such as ceasefires and changing warfare tactics.

Plot 2: Normalized Error Distribution between data from flash and UN verified reports on cumulative numbers in West Bank

Motivation Behind Plot 2: Specifically, the goal of this plot is to assess the accuracy of flash reports, revealing the error magnitude and whether flash reports can act as a basis for immediate decision making in which humanitarian organizations can rely on. Furthermore, we have normalized the values for a better comparative insight between the errors of different categories.



Plot Result Summary: We normalized each value using the max value in each category so that they are all on the same scale with values between -1 to 1, so that we can generate an overlaid histogram. Below are the max values of each category:

1. Max Error Killed: 28.0
2. Max Error Child Killed: 6.0
3. Max Error Injured: 437.0

From the overlaid histogram, we can see that there are more spread for total and child killed errors reaching over .5 and -.5 but since their max error value is low at 28.0 and 6.0, respectively, the absolute error calculated will still be relatively low meaning that the flash reports are generally accurate with regards to deaths.

For example, if the normalized error is 0.5 for an error in total and child killed. Then the absolute error will be (0.5 * 28) = 14 and (0.5 * 6) = 3, which are relatively low error values.

Meanwhile, with injuries, even if they have less spread in the normalized histogram but due to having their max error value at 437.0, a slight spread away from the center (0 error) results in a larger change in absolute error.

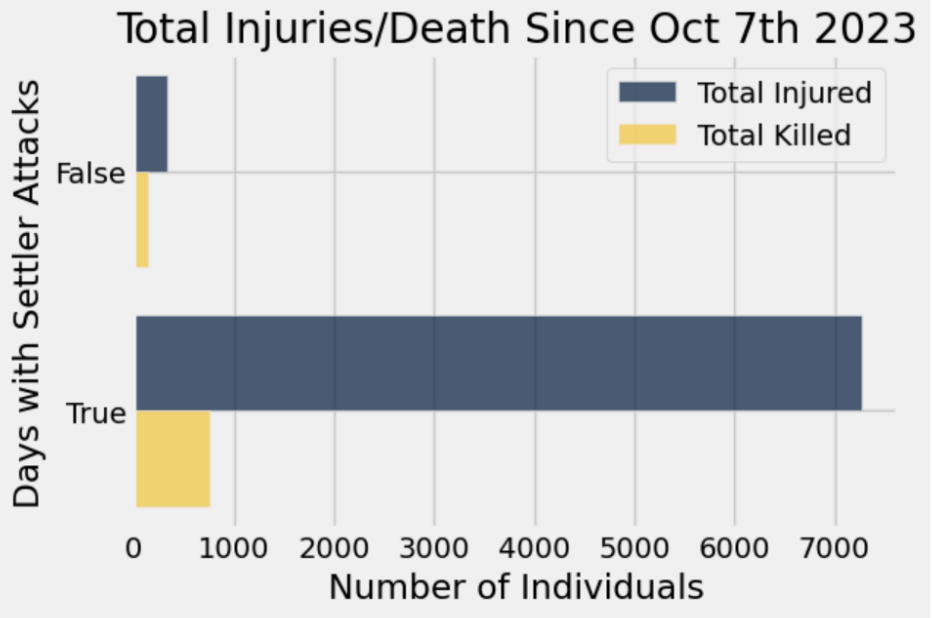
For instance, if the normalized error is 0.05, then (0.05 * 437) = 21.85. If the normalized error is 0.1, then (0.1 * 437) = 43.7. A change in 0.05 for normalized error resulted in a 21.85 change in absolute error. Therefore, the errors are wider for injury reports, making them less reliable compared to reports of deaths. Overall, most of the normalized errors of injuries rest between -0.1 and 0.1, so we can still conclude that the injury flash reports are still moderately accurate and not too far off from verified reports.

Flash reporting of fatalities in the West Bank is generally accurate and reliable with little to no bias, particularly for total and children deaths, with most errors clustering tightly around zero. However, injury reports show greater variability, suggesting that injuries are harder to document reliably in real-time. Overall, while flash reports provide a useful early estimate of fatalities, slight caution is needed when interpreting reported injury figures without later verification.

In the end, humanitarian organisations and political actors can rely on and use early data through flash reports for immediate decision-making and provide the necessary relief efforts for the Palestinians in a quick time without waiting for data to be verified.

Plot 3: Settler Attacks causes more Injuries/Killed in West Bank (Bar Chart).

Motivation behind Plot 3: This visualization aims to highlight the relationship between settler violence and casualtie**s** (both injuries and deaths) in the West Bank following October 7th, 2023.



Plot Result Summary:

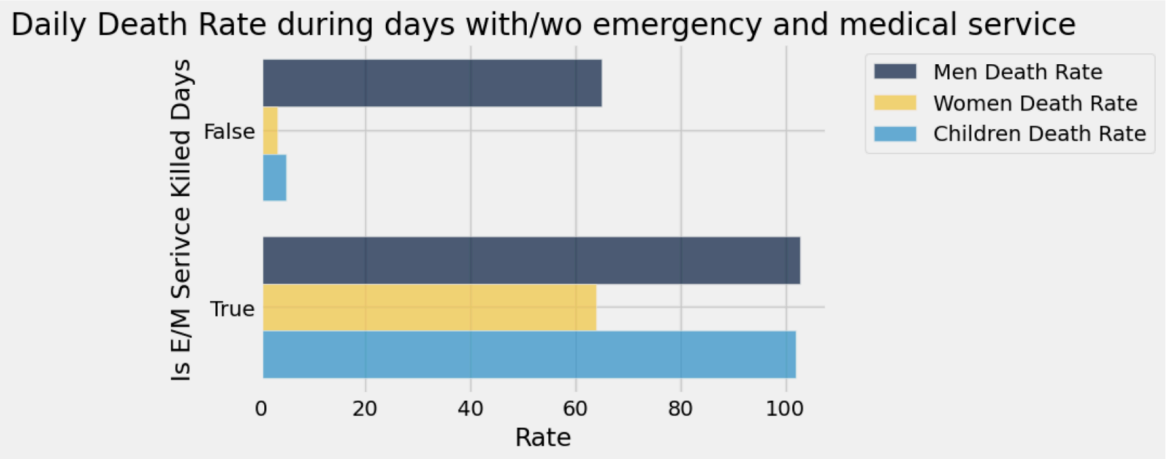
Most casualties occurred on days with Israeli settler attacks, with 750 dead, while on other days, it only amounted to 147 deaths. Injuries on settler-attack days far exceed those on days without, with values of 7276 and 337 deaths, respectively. This implies that Settler attacks are not isolated events; they occur alongside or escalate broader violence involving killings and injuries.

Casualties still occur without settler attacks, but at a lower scale. Even on days with no reported settler attacks, there are injuries and deaths. This shows the multi-dimensional nature of violence (e.g., military operations, clashes, raids).

The data shows a clear association between settler attacks and higher casualty counts. While violence also exists independently, settler-attack days coincide with disproportionate harm to Palestinians in the West Bank, highlighting that they are not just there to protest but also cause unnecessary harm to the Palestinians. This pattern is worth monitoring and provides evidence for prevention and accountability.

Plot 4: Daily Death Rate of different groups during days with/without emergency and medical services in Gaza.

Motivation Behind Plot 4: This bar chart aims to investigate the impact of the availability of emergency and medical (E/M) services on daily death rates, broken down by demographic groups (children, women, others).



Plot Result Summary:

The group averages indicate that days when emergency service personnel are killed are associated with a significantly higher number of civilian deaths.

Men's death rate is significantly higher on 'E/M Service Killed Days' from 65 to nearly 102/day. Overall, the total death rate essentially doubles when medics or civil defense teams are killed.

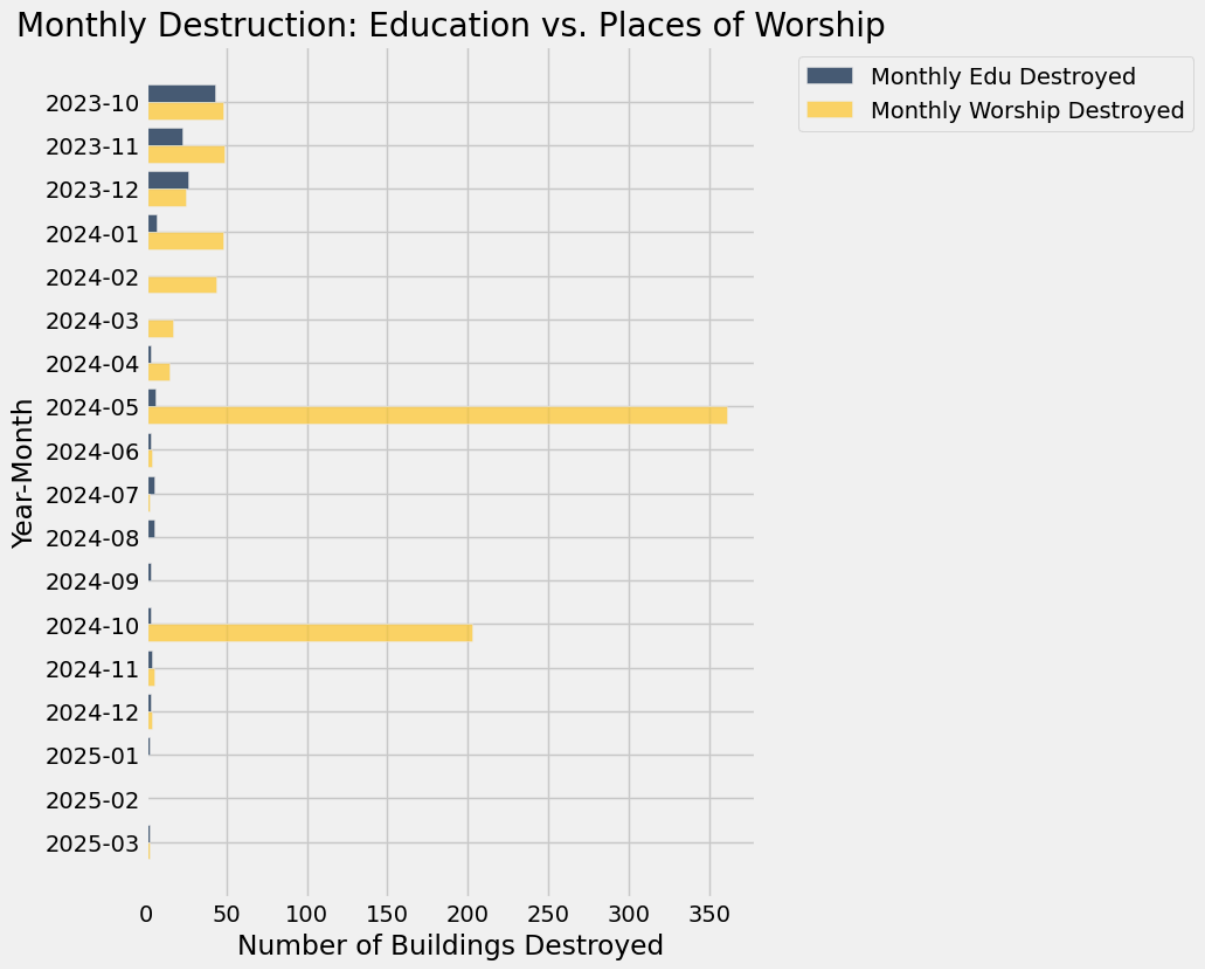
Child and women's death rates are 20+ times higher on days when emergency services are attacked. During "E/M Service Killed Days", child deaths spike to over 100 per day, while on non-service-killed days, child death rate is close to 5. Similarly, women killed on "E/M Service Killed Days" amount to 63 per day, while on other days, it is significantly lower at about 3 per day. This shows that when medical/rescue services are attacked, children and women are disproportionately affected. This could be because during times of conflict, children are not able to react when such a situation happens, and it is also very overwhelming for women, as both are much weaker physically in general. Therefore, any denied or delayed emergency care can greatly affect the number of children and women dying. Other highly potential reasons are the collapse of evacuation/rescue infrastructure from Israel's attacks, preventing children and women from escaping the conflict area.

It can be said that the presence of emergency response directly affects survivability, especially for children and women. Attacking those E/M teams increases the overall lethality of conflict events. Therefore, this strong association creates greater awareness of the importance of personnel aid, and when their number decreases after an attack, humanitarian aid and political people should rush and provide the necessary relief efforts and reinforce emergency services in such moments of crisis.

Plot 5: Monthly Destruction - Educational Buildings vs. Places of Worship

Motivation Behind Plot 5: This visualization aims to compare the monthly destruction trends of two specific types of critical civilian infrastructure: educational buildings and places of worship

(mosques and churches combined) in Gaza. The goal is to identify patterns over time, observe potential differences or similarities in how these infrastructures were affected, and compare these trends to the destruction of residential buildings shown in Plot 1.

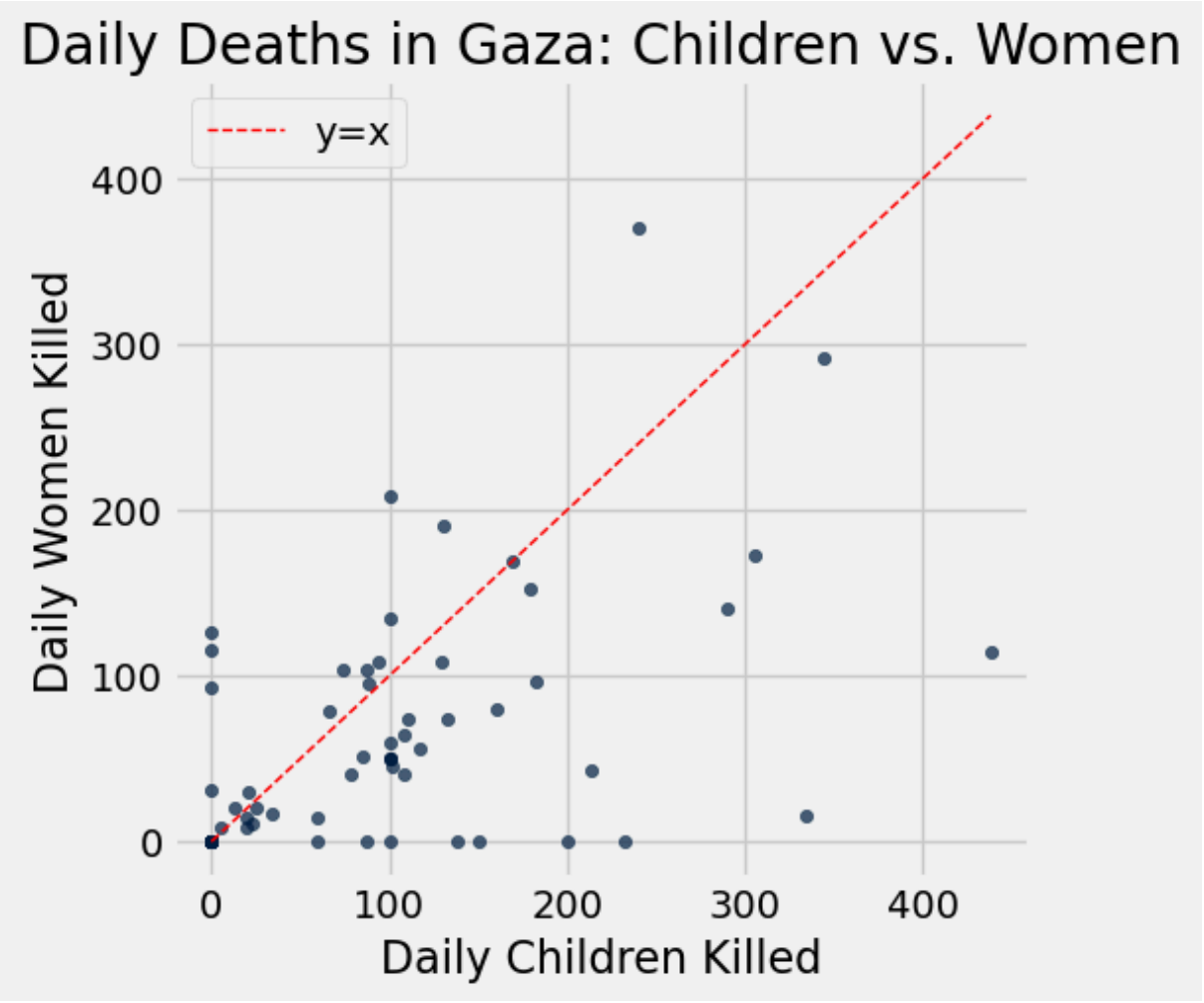


Plot Result Summary:

The bar chart displays the estimated number of educational buildings and places of worship destroyed each month. Similar to the trend observed for residential buildings in Plot 1, the initial months of the conflict (October-December 2023) witnessed substantial destruction in both categories, although the absolute numbers are lower than for residential units. In subsequent months, the destruction patterns diverge slightly: damage to educational facilities appears more sporadic, while the destruction of places of worship shows some later peaks (e.g., June 2024) that mirror the residential destruction trend. Periods corresponding to reported ceasefires (like early 2025) exhibit minimal destruction in both categories, reinforcing the connection between active conflict phases and infrastructure damage. Comparing this plot to Plot 1 reveals different scales of destruction but similar overall temporal patterns, suggesting a widespread impact across various types of civilian infrastructure, particularly during high-intensity phases of the conflict.

Plot 6: Daily Deaths in Gaza - Children vs. Women

Motivation Behind Plot 6: This scatter plot seeks to visually explore the relationship between the daily number of reported child deaths and women deaths in Gaza. The objective is to understand if fatalities among these two vulnerable groups tend to occur concurrently and to assess the relative scale of these casualties on a day-to-day basis.



Plot Result Summary:

The scatter plot compares the number of children killed versus women killed on the same day. A clear positive correlation is visible: days reporting higher numbers of child deaths generally also report higher numbers of women deaths. Many data points cluster near the origin (indicating days with low casualties for both groups) and along the y=x reference line, suggesting that on numerous days, the reported death tolls for children and women were of a similar magnitude. However, there is considerable scatter, with many points lying far from the y=x line, highlighting significant daily

variability where one group suffered disproportionately more than the other. The points appear slightly more dispersed above the y=x line, potentially indicating that days where child deaths exceeded women's deaths were more frequent or the difference was larger compared to the reverse. This visual observation aligns with the findings of Hypothesis Test 3, which concluded a statistically significant difference, with slightly more children killed on average per day than women.

2. Hypothesis Testing

Test 1: Does the day when emergency and medical services' individuals are killed result in higher civilian deaths?

Motivation Behind Test 1: To prove whether medical services' presence in Gaza is a major factor in the number of deaths and injuries, or are these injuries are just due to random chance.

Procedures:

Null Hypothesis: In the population, the distributions of deaths on days with/wo emergency services' deaths are the same. (They are different in the sample just due to chance.)

Alternative Hypothesis: In the population, on days with emergency services dead, death numbers are **higher**, on average, than on days without any emergency services individuals dead.

Keys: *E* = *Emergency*, *M* = *Medical*

Test Statistic Used: Average Deaths on E/M Service Killed Days − Average Deaths on E/M Service Killed Days

Reason: This shows whether there is an effect or not on the number of deaths between days when E/M services are killed and vice versa

Meaning:

1. Small Values (Close to Zero) = No real effect
2. Negative Values = On average, Less number of civilians are killed on days when E/M services are killed
3. Large Positive Values = On average, a larger number of civilians are killed on days when E/M services are killed.

Significance Level: We will conduct the hypothesis test at a significance level of **1%.**

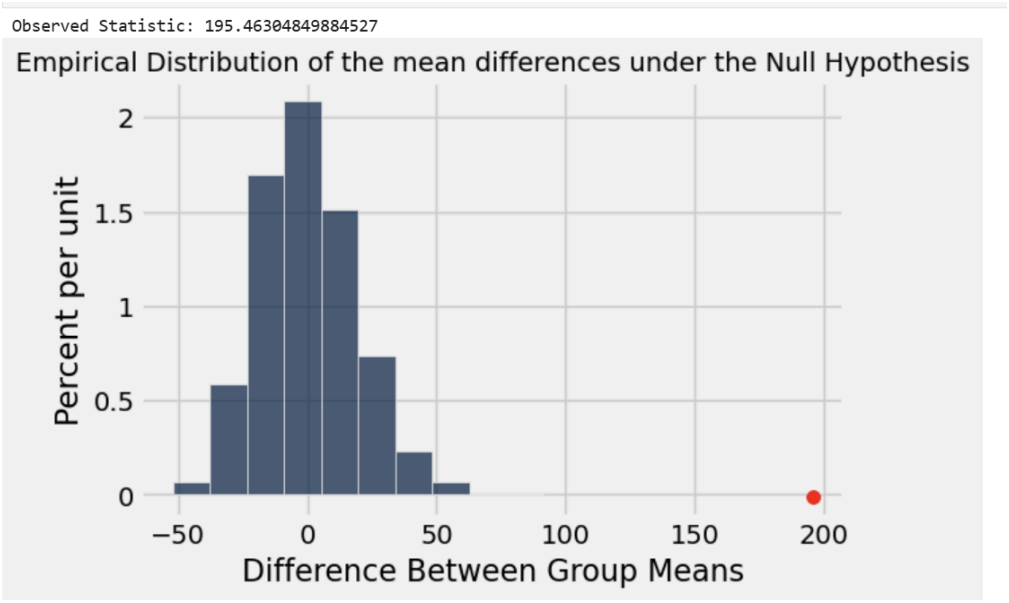
Simulation Steps:

Simulate by randomly shuffling the "Is Service Killed Days" labels among all days.

Every shuffle calculates the simulated test statistic.

Append the statistic to an array of test statistics.

Repeat 10000 times to form an empirical distribution of the test statistic under the null hypothesis.



Testing Conclusion: (p-value = 0.0)

Since the p-value is less than 0.01, we reject the null hypothesis. We conclude that the deaths of emergency and medical services personnel are significantly associated with an increase in the number of civilian deaths and are not just due to random chance.

Test 2: Is the age distribution of killed individuals different from Gaza's population distribution?

Motivation Behind Test 2:

According to demographic data, Gaza's population has the following approximate age structure:

* Ages 0–14: 39.75%
* Ages 15–64: 57.34%
* Ages 65 and over: 2.91%

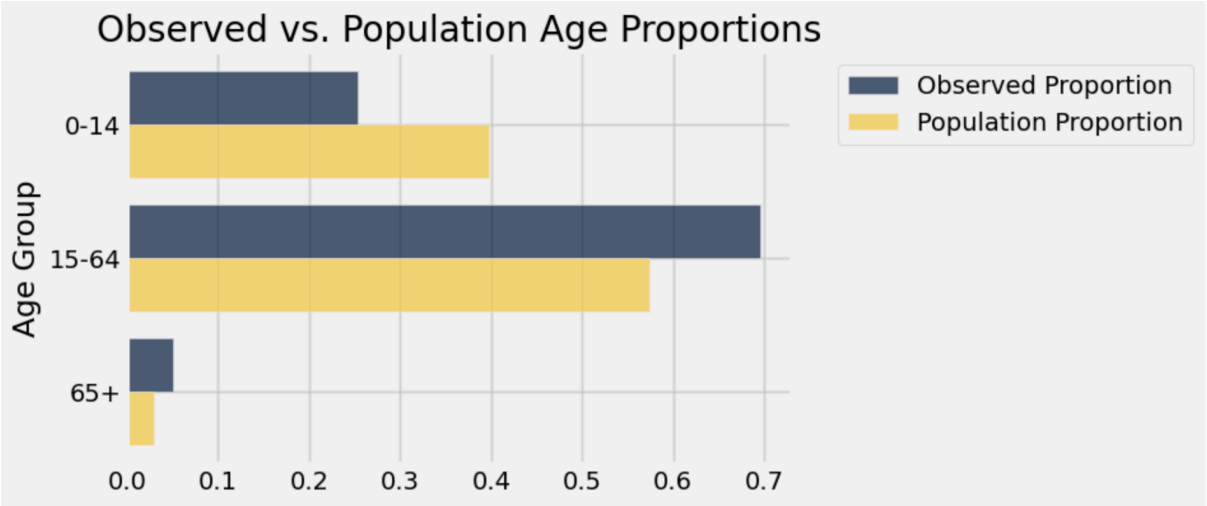
We want to test if the age distribution of the individuals recorded in the `killed_in_gaza` dataset significantly deviates from this known population structure.

Source:
https://www.wusf.org/2023-10-19/children-make-up-nearly-half-of-gazas-population-heres-what-it-means-for-the-war

Procedures:

Null Hypothesis:

The age distribution of individuals killed in Gaza (in our sample) follows the general population age distribution. Any observed difference between the sample distribution and the population distribution is due to random chance in sampling.



Alternative Hypothesis:

The age distribution of individuals killed in Gaza (in our sample) is different from the general population age distribution. The observed difference is not just due to random chance, suggesting certain age groups are disproportionately affected.

Test Statistic Used:

Total Variation Distance (TVD), which is calculated as:
TVD = sum(abs(observed proportions − population proportions)) / 2

Reason:

TVD measures the total difference between two categorical distributions. It quantifies how much one distribution needs to shift to match the other.

Meaning of the Test Statistic:

- Small values (close to 0): The observed distribution of killed individuals is very similar to the general population distribution.
- Large values (closer to 1): The observed distribution is very different from the general population distribution.

Significance Level:

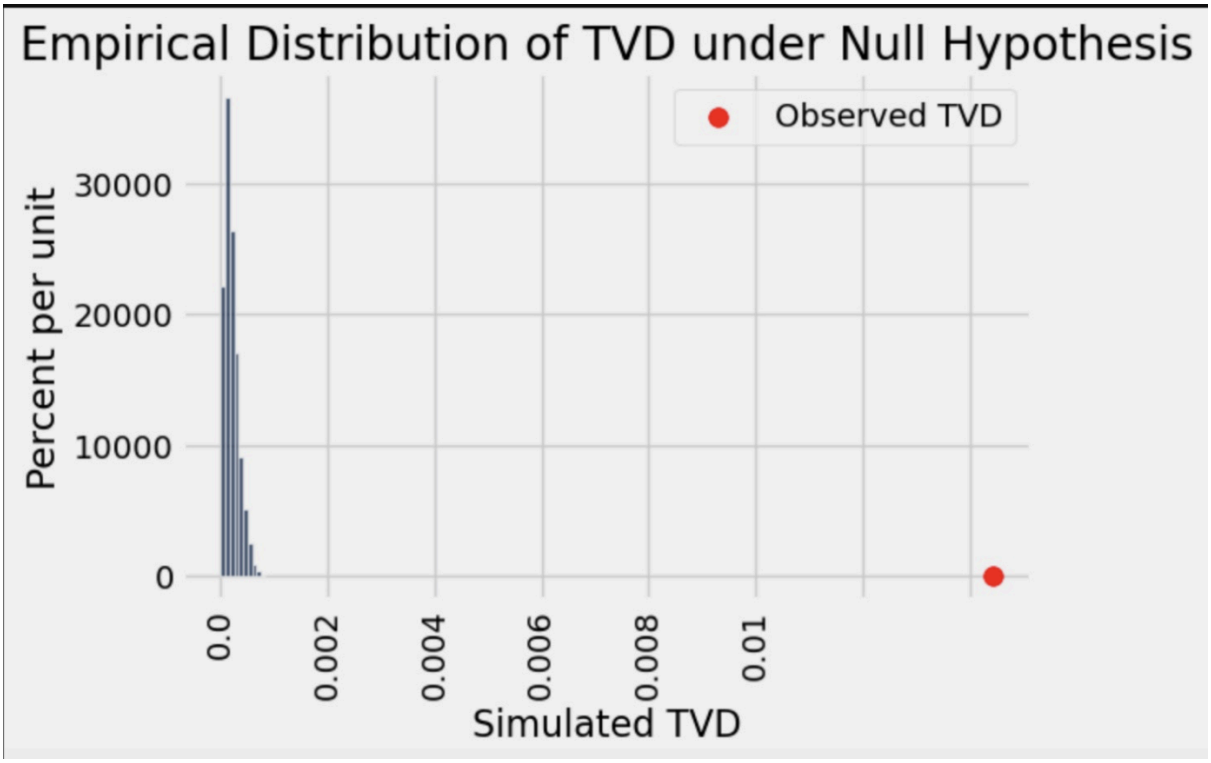
We will conduct the hypothesis test at a significance level of 1%.

Simulation Steps:

1. Define the Test Statistic: Use Total Variation Distance (TVD) to test expected vs. observed age ratios.
2. Compute Observed TVD: Calculate TVD between the sample actual age distribution and Gaza population ratios.
3. Simulate Under the Null Hypothesis:

   Repeat 10,000 times, Sample randomly with population ratios, and compute TVD between simulated and population ratios.

4. Visualize Simulation: Plot a histogram of simulated TVDs and plot the observed TVD.
5. Calculate P-Value: Determine the ratio of simulated TVDs with values equal to or higher than the observed TVD.



Testing Conclusion: (p-value = 0.00)

The calculated p-value is essentially 0, which is far less than our significance level of 1% (0.01). Therefore, we reject the null hypothesis.

The data provides extremely strong evidence that the age distribution of the individuals killed, as recorded in this dataset, is significantly different from the general age distribution of Gaza's population. The observed differences are highly unlikely to be due to random chance alone, indicating that certain age groups (particularly the 0-14 group, as seen in the bar chart) were disproportionately represented among the casualties compared to their proportion in the overall population.

Test 3: Is the average daily number of children killed significantly different from the average daily number of women killed in Gaza?

Motivation Behind Test 3:

Following the observation in Plot 6 that daily deaths of children and women seem correlated but not identical, this test aims to statistically determine if there is a significant difference in the average number of deaths per day between these two groups in Gaza, based on the available daily data.

Procedures:

Null Hypothesis:

In the population of days, the average daily number of children killed is the same as the average daily number of women killed. The average of the differences (Children Killed - Women Killed) is zero. Any observed average difference in the sample is due to random daily fluctuations.

Alternative Hypothesis:

In the population of days, the average daily number of children killed is different from the average daily number of women killed. The average of the differences (Children Killed - Women Killed) is not zero.

Test Statistic Used:

Average of the Daily Differences (Avg(Daily Children Killed - Daily Women Killed)).

Reason:

For paired data (daily counts for both groups), analyzing the difference within each pair (day) controls for factors affecting both groups on that day. We test if the mean of these differences is significantly different from zero.

Meaning:

Values Close to Zero: On average, the daily number of children and women killed is similar.

Large Positive Values: On average, more children are killed per day than women.

Large Negative Values: On average, more women are killed per day than children.

Significance Level: We will use a significance level of 5%.

Simulation Steps:

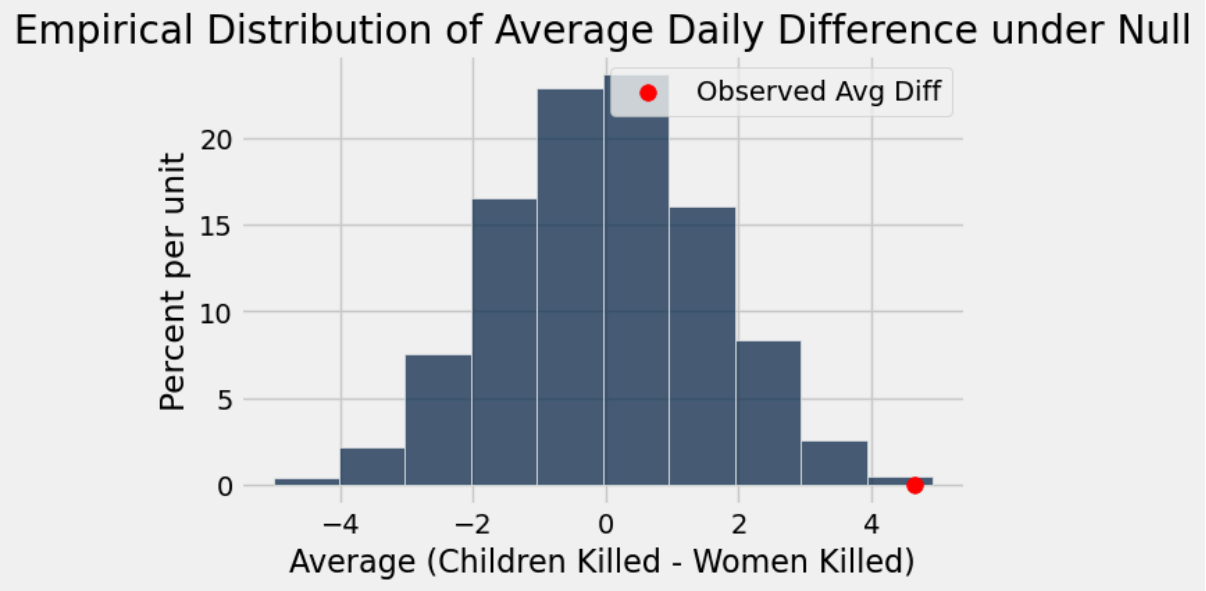
Calculate the difference between 'Daily Children Killed' and 'Daily Women Killed' for each day.

Simulate under the null hypothesis by randomly assigning a sign (+1 or -1) to each daily difference. This assumes that under the null, the sign of the difference on any given day is random.

Calculate the average of these randomly signed differences (the simulated test statistic).

Append the statistic to an array of test statistics.

Repeat steps 2-4 10,000 times to form an empirical distribution of the test statistic (average difference) under the null hypothesis.



Testing Conclusion: (p-value = 0.0007)

The calculated p-value (0.0007) is less than our significance level of 5% (0.05). Therefore, we reject the null hypothesis.

The data provides statistically significant evidence that the average daily number of children killed in Gaza is different from the average daily number of women killed during the period

covered by the data. Specifically, the positive observed average difference suggests that, on average, slightly more children were killed per day than women.

3. Individual Contributions

Task Distribution & Contribution		
Student Name	**Student QUID**	**Tasks Assigned**
Fahrel Hidayat	202206836	Plot 2, 3, 4, and Hypothesis Test 1
Zubair Bin Jashim	202108715	Preprocessing, Plot 5, 6, and Hypothesis Test 3
Mohd Muhtasim Bashar	202205153	Part A.2, Plot 1 and Hypothesis Test 2
Sheikh Hasin Ishrak	202108715	Part A.1 & Preprocessing