

**Group – A4**

**Project Part B Report**

**CMPS 360\_L01: Data Science FundamentalsSpring 2025**

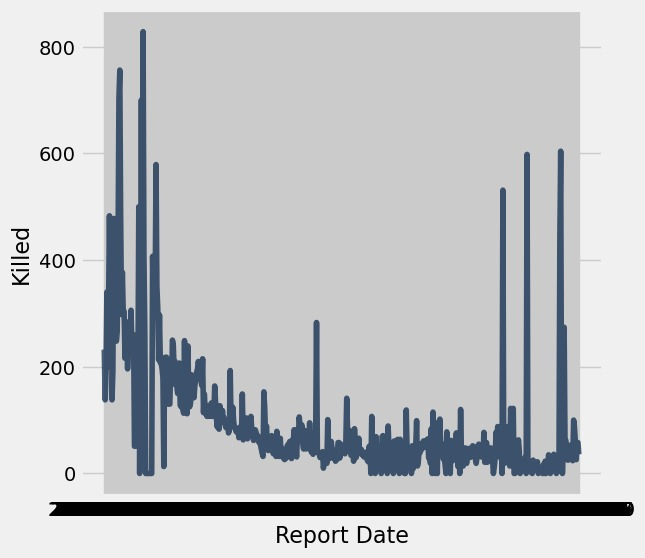
**Dr. Tamer Elsayed**

**06/05/2025**

**Part 1**

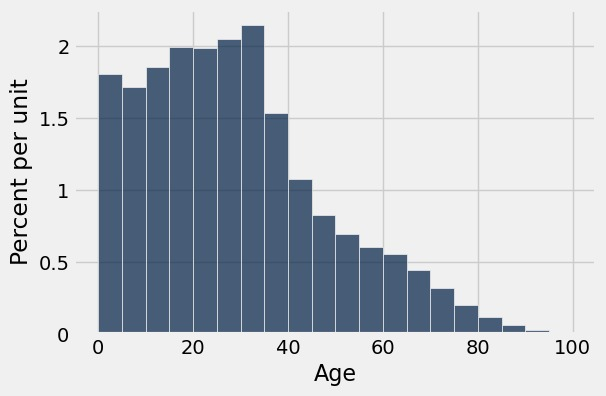
For part 1, we managed to clean 4 .csv files out of the 5 because we weren’t able to convert the .pdf into a csv after multiple and repeated efforts. Even asking for help didn’t improve our position so we left it. We were successful in converting the .json file into a csv. For cleaning, we discovered that there were repeating columns with the same name except there was an “ext” or “verified” in the attribute name. So we kept those columns and discarded the rest since these columns had no missing data. We also re-named the column headers to make it more readable. Overall we discard some columns where there were too many missing data to fill out or replace with 0 that it would affect our plots so we decided to discard them.

**Part 2**

**Plot 1**

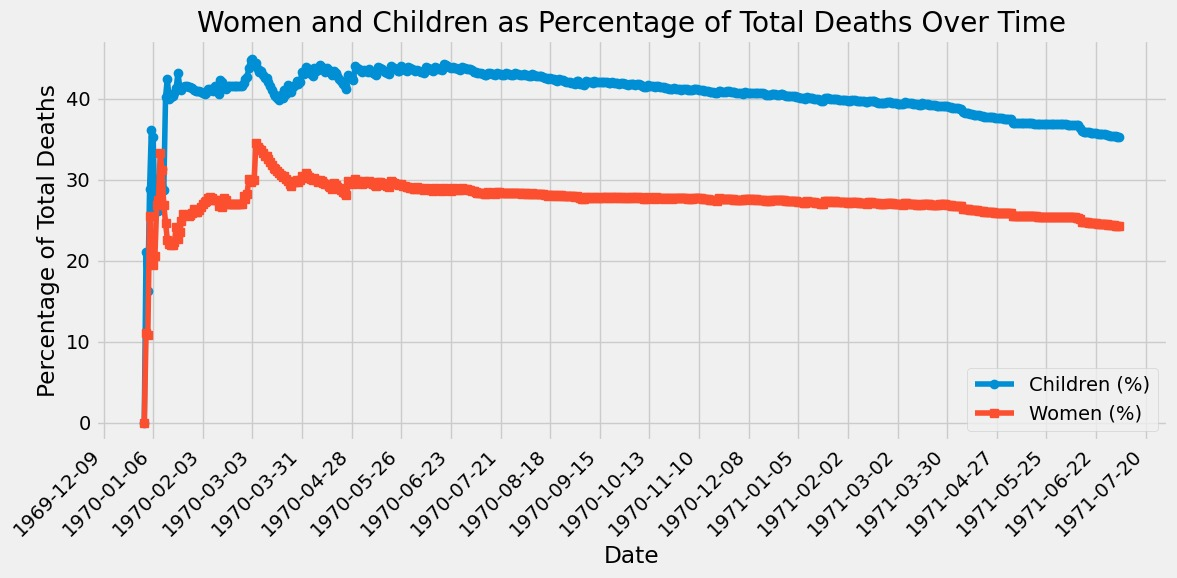
This plot was derived from *casualties\_daily\_gaza\_2025-04-12.csv* and it shows how much people are killed everyday from 10/7/2023 to 4/9/2025. This is a line graph which is done by using .plot(). As you can see from the graph that there are 6 spikes.

**Plot 2**



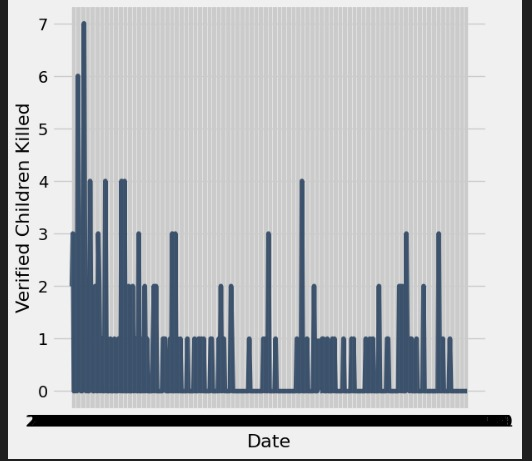
This plot was derived from *killed-in-gaza\_2024-09-21.csv* which shows the age distribution of the people who were killed in Gaza. This is a histogram where you can you .hist() function. This histogram shows which ages were the most to have died and this graphs tell us that the bulk of the people who died were between ages 0 – 35.

**Plot 3**



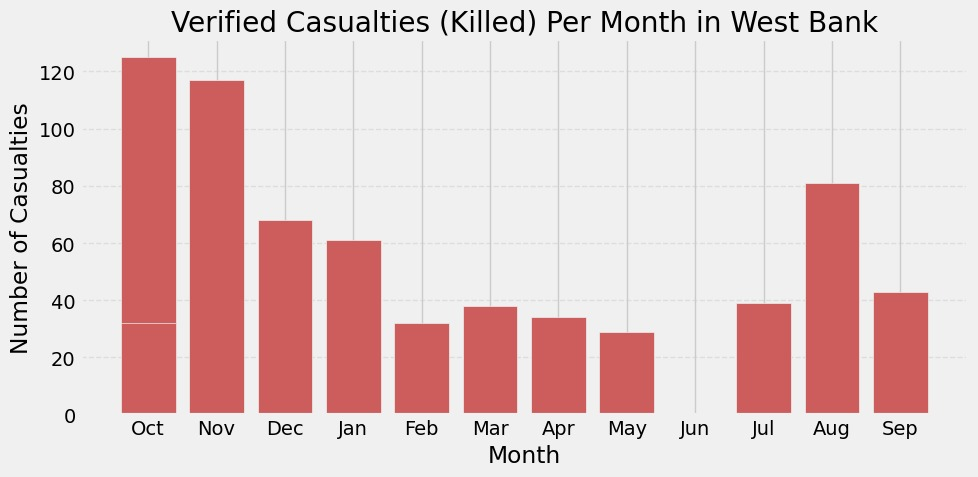
This graph was derived from *casualties\_daily\_gaza\_2025-04-12.csv* and it shows the percentages of both women and children out of the total killed. It is evident from the graph that more % of children have died than women and that the peak % of deaths for both women and children were the highest around 1970.

**Plot 4**



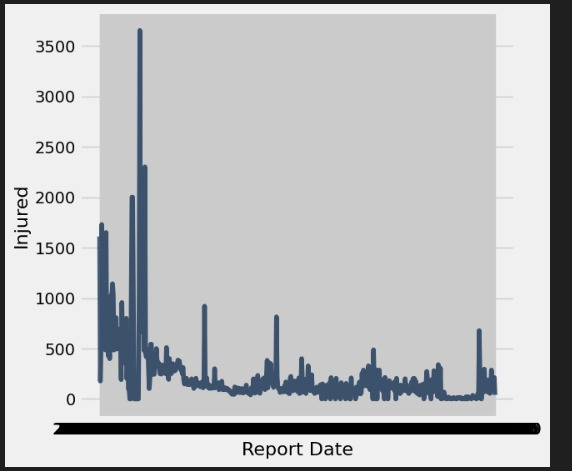
This plot was derived from *casualties\_daily\_west\_bank\_2025-04-12* and it shows the verified children that were killed. You can notice that these numbers are very low but this is because these kills were verified and confirmed rather that the other numbers which were reported but not verified for complete authenticity. This is a line graph which is verified killed over time.

**Plot 5**



This is a bar chart which was derived from *casualties\_daily\_west\_bank\_2025-04-12.csv* and it shows the number of casualties for each month of all the years in the West Bank. You can see from the graph that most of the casualties happened in October and November and the least we in May and June. It could be the possibility that the data may have been incomplete for June specifically.

**Plot 6**



This a line graph which was derived from *casualties\_daily\_gaza\_2025-04-12.csv* and it shows the amount of people who were injured in Gaza over time. You can see 1 huge spike at the beginning of the graph.

**Part 3**

**1. State the Hypotheses**

- Null Hypothesis (H₀):

"Each male in Gaza has a 65% chance of being killed in the war."

- Alternative Hypothesis (H₁):

"The true proportion of men killed is not 65%."

- Two-sided test: We test for any deviation (higher or lower).

2. Test Statistic

• Statistic: Proportion of men among total fatalities

3. Final Conclusion

• We accept the null hypothesis, The data does not support the alternative hypothesis that men were killed at a rate different from 65%. The red dot (63.98%) falls within the high-probability region of the null distribution, deviations this small are common under random sampling.



**Hypothesis 2 :**

Question : Are women and children being disproportionately targeted in Gaza, as evidenced by a higher proportion of their deaths compared to their share in the population?

1. Hypotheses

• Null Hypothesis:

Deaths in Gaza occur proportionally according to the general population distribution. That is, if 73% of the population are women and children, then approximately 73% of those killed should also be women and children. Any deviation from this proportion is due to random chance.

• Alternative Hypothesis:

The proportion of women and children killed is different from 73% either higher or lower suggesting a disproportionate targeting of this group.

• Where is the chance?

The randomness comes from the assumption that fatalities are a random sample from the general population. Under the null, each fatality independently has a 73% chance of being a woman or child.

2. Test Statistic

• Chosen statistic: Proportion of deaths that are women and children.

• Why this choice: It directly measures the percentage of women and children among all deaths and allows direct comparison with the expected proportion under the null hypothesis.

• Meaning:

o Small values (much less than 0.73) suggest underrepresentation — possibly that men are being killed more.

o Large values (much more than 0.73) suggest overrepresentation — possibly that women and children are being disproportionately killed.

o Values close to 0.73 support the null hypothesis.

3. Simulation Experiment

• Under the null hypothesis, we simulate 10,000 samples where each “death” is randomly labeled as either “Woman/Child” or “Other” with probabilities 0.73 and 0.27 respectively.

• In each simulation, we compute the proportion of women and children killed.

• This generates an empirical distribution of the test statistic under the null hypothesis.

• The observed statistic (actual proportion of women and children killed) is compared to this distribution.

4. Results

• Total deaths observed: 30319

• Women + children killed: 50846

• Observed proportion: 0.5963

• Population proportion under null: 0.73

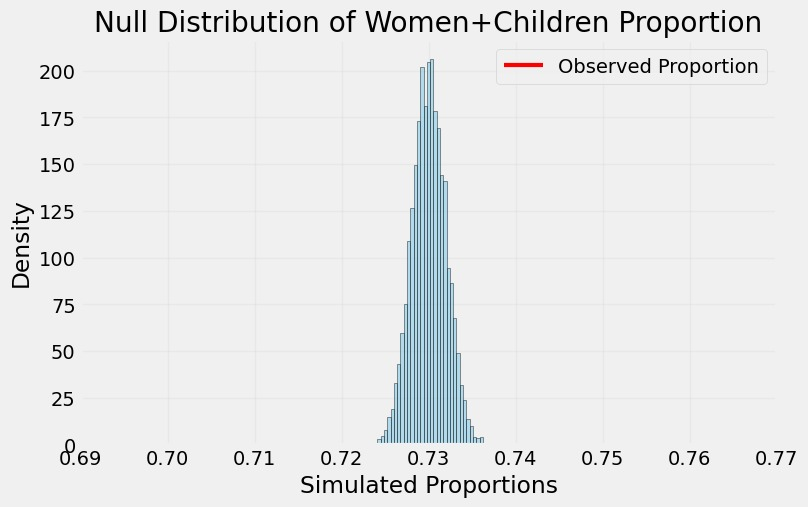
• Empirical p-value: 1.0

Interpretation & Conclusion

The observed proportion of women and children among the dead (59.63%) is significantly lower than the expected proportion under the null (73%). The p-value is approximately 1.0, meaning that a value as low or lower than 0.5963 is extremely likely under the null — in fact, not unusual at all.

Therefore, we fail to reject the null hypothesis.

Conclusion: Based on this test, there is no statistical evidence that women and children are being disproportionately targeted compared to their share in the population. In fact, they appear to be underrepresented among the fatalities — contrary to what the alternative hypothesis would suggest.



**Hypothesis 3**

1. State the Hypotheses

Null Hypothesis (H₀):

"Each child in the West Bank has a 20% chance of being killed in the war."

Alternative Hypothesis (H₁):

"The true proportion of children killed is greater than 20%."

One-sided test: We test specifically whether the proportion is higher than 20%.

2. Test Statistic

Statistic: Proportion of children among total fatalities in the West Bank.

3. Final Conclusion

We accept the null hypothesis. The data does not support the alternative hypothesis that children were killed at a rate higher than 20%.

The red dot (23.39%) falls within the high-probability region of the null distribution. Deviations of this size or larger occur often under random sampling, as indicated by a p-value of 0.0755 (which is above the 0.05 threshold).

Therefore, we conclude there is no statistically significant evidence that children were disproportionately killed in the West Bank.

