Department of Computer Engineering
College of Engineering
Polytechnic University of the Philippines Sta. Mesa

CMPE 40163: Exploratory Data Analysis Final Requirement

Exploratory Data Analysis of Fatalities in the Israeli-Palestinian Conflict

Submitted by:

JAMCO, KENNETH A.
BSCOE 3-2

Submitted to:

EDCEL B. ARTIFICIO

I. Introduction and Purpose of the Analysis

Describe your chosen domain and your objective for this project. You can also provide an overview of your chosen field. For example, if you want to analyze banking data, you can give an overview of the banking industry and why manipulating its data is important in making business decisions. This is also the section where you talk about the questions you will answer at the conclusion.

Overview of the Israeli-Palestinian Conflict

A long and complicated struggle revolves around the land of Palestine, dating back to the late 19th century when the idea of creating a Jewish homeland in Palestine gained momentum through the Zionist movement. This resulted to issues about the borders, the status of Jerusalem, the rights of palestinian refugees, and the establishment of a Palestinian state.

In 1948, the United Nations made a decision to divide Palestine into two states, one for Jewish people and the other for Arabs. But this plan faced opposition from the Arab nations, leading to a war with Israel. This war caused the displacement of hundreds of thousands of Palestinians and led to the founding of the State of Israel. The conflict has been marked by a number of major events, including the Six-Day War in 1967, the First and Second Intifadas in the 1980s and 2000s, and the ongoing conflict in Gaza.

The chosen domain for this project is "Fatalities in the Israeli-Palestinian Conflict." The objective of this project is to provide a comprehensive and accurate overview of fatalities in the conflict starting from year 2000. And also to analyze and explore the patterns, trends, underlying factors related to the fatalities, the demographics of the victim, and circumtances of their death that have occurred in the Israeli-Palestinian conflict. Questions about this dataset will be answered such as the overall trends in fatal incidents over time, hotspots of fatal incidents, and if there are any evidence of conflict escalation or de-escalation done by the involved parites based on the dataset overtime. This is conducted to shed a light to the human lives that have been lost due to this conflict.

II. Data Dictionary

Outline of data variables and high-level information about the dataset (how many rows, columns? etc.)

This dataset currently have 11,124 rows and 16 columns which are "name", "date_of_event", "age", "citizenship", "event_location", "event_location_district", "event_location_region", "date_of_death", "gender", "took_part_in_the_hostilities", "place_of_residence", "place_of_residence_district", "type_of_injury", "ammunition", "killed_by", and "notes".

- name: The name of the deceased individual, including their full name.
- date_of_event: The date when the incident or event leading to the fatality occurred.
- age: The age of the deceased individual at the time of the incident.
- citizenship: The citizenship or nationality of the deceased person (e.g., Palestinian, Israeli).
- event_location: The specific location or place where the incident occurred.

- event_location_district: The district or region within which the incident location is situated.
- event_location_region: The broader region where the incident took place (e.g., West Bank).
- date_of_death: The date when the individual died as a result of the incident.
- gender: The gender of the deceased individual (e.g., M for male).
- took_part_in_the_hostilities: Information about whether the deceased person took part in hostilities or was involved in any conflict-related activities.
- place_of_residence: The place of residence of the deceased individual at the time of the incident.
- place_of_residence_district: The district or region where the deceased person resided.
- type_of_injury: The specific type of injury sustained by the deceased individual (e.g., "gunfire," "stabbing").
- ammunition: The type of ammunition used in the incident that resulted in the fatality.
- killed_by: The entity or party responsible for causing the death of the individual (e.g., "Israeli security forces," "Palestinian civilians").
- notes: Additional information or context about the specific incident, including details about the circumstances, triggers, or other relevant factors.

III. Analysis Process

Provide an overview of how your approach your EDA. What are the statistical/graphical techniques that you employed? Why did you approach your EDA the way you did? What informed your decision to choose a statistical/graphical technique?

Bar graph, time series plots, histogram,
Time series analysis, descriptive statistics, frequency analysis,

For the Exploratory Data Analysis of the dataset Fatalities in the Israeli-Palestinian Conflict, I used both statistical and graphical techniques to gain a deeper understanding about the data. The graphical techniques used are:

- Bar graphs – it is used to visualize the distribution of categorical data like gender, citizenship, and type of ammunition. In this way, I can quickly learn the frequency of each category within these variables. Also, it is the simplest yet effective way to present such types of data.
- Time series plots – time series graphs are used to examine the trends and patterns of the events and fatal incidents over the years. This technique helped me understand the evolution and curves of the conflict over the years.
- Histograms – histograms helped me study continuous data such as the age of the victims. This helped me understand if there is any disparity between the Israeli and Palestine. What age group are move vulnerable to the conflict.

For statistical analysis I used:

- Time series analysis - This statistical technique helps identify trends, seasonality, and patterns over time, providing more detailed insights into the temporal aspects of the data.

- Descriptive statistics – measurements like mean, median, mode to get a quick insight into the data's distribution.

Analyzing these time series data helps in identifying long-term trends and patterns over time, which can be crucial in understanding the evolution of fatal incidents. I opted to use bar graphs and histograms in my analysis for quick and intuitive visualization of both categorical and continuous data. These graphical techniques are particularly useful for the initial exploration of the dataset as it vividly illustrate the distribution of data.

I also employed time series plots and conducted time series analysis. These methods were particularly useful because the dataset includes information about when events and deaths occurred. This allowed me to thoroughly investigate long-term trends and patterns over time, which are crucial for understanding how fatal incidents have evolved. In order to gain a deeper insight into the dataset, I used descriptive statistics and frequency analysis. These approaches provided a quantitative summary of the data, making it simpler to comprehend the central tendencies and the distributions of different attributes.

My choice of techniques was based on the dataset's characteristics and the specific goal of examining temporal trends and patterns in fatal incidents. By combining these methods, I was able to comprehensively explore the data and extract valuable insights about the nature of these incidents, the characteristics of the victims, and the broader patterns that emerged over time.

IV. Analysis and Insights

What emerged from your EDA? Were you able to answer your questions? Did you uncover any new questions or areas for future explorations? How do you think your analyses and insights would contribute to the body of knowledge of your chosen topic/domain expertise.

Throughout the EDA process, I have learned more about the Israeli-Palestinian conflict. Those numbers have given more context by watching videos regarding the said conflict. And also vice versa because the dataset allowed me explore and understand this long conflict more than just watching videos online.

As we all know, it is men that are always being sent to the battlefield when there is war or conflict with other countries. Throughout 2000 to 2023, almost 10,000 men died compared to an estimated 1,500 female casualties, that is 87% of the total war that is not going home to their family. Also, most of the casualties are around the age of 19 to 65 years old leaving the children and elderly all by themselves.

About my questions before the EDA, it did answer ny questions. Fatal incidents trend usually happen when there is a war and exchange of artillery between the two countries. For example, the graph of Total Recorded Events and Casualties by Year shows the highest spike at 2014. At that time, there is a Gaza War happening from July 08, 2014 to Aug 26, 2014. Second highes spike was at 2009 when The Gaza War, which began in December 2008 and ended in January 2009 happened. It was one of the deadliest periods in the Israeli-Palestinian conflict in recent years. More than 1,400 Palestinians and 13 Israelis were killed in the war. But after every spike of activity, follows a deep dip which indicates that de-escalation are done which helps me understand its dynamics. The geographic exploration contributed to identifying geographical patterns and understanding where incidents were most concentrated. This information is crucial for conflict analysts and policymakers. Examining the "killed_by" column provided insights into

the entities or parties responsible for fatal incidents. Analyzing this information helped me understand the involvement and actions of various parties, such as Israeli security forces, Palestinian civilians, or others.

The EDA uncovered new questions are possible future explorations. Further research could be done on why there is a pattern of spikes from 2000 to present because it seems that the de-escalation efforts are nothing but a band-aid solution. Also, analyzing why such age are the most common casualties in this war and could explain why they are so vulnerable during this difficult times.

The analysis and insights of Palestinian-Israeli Conflict contributes to the conflict studies in this war and to other countries. This knowledge can be valuable for conflict analysts, researchers, and policymakers seeking to understand and potentially address conflict dynamics, and it can serve as a foundation for further research in this domain.

V. Conclusion and Recommendation

What are your conclusions and recommendations?

The Exploratory Data Analysis of the dataset Israeli-Palestinian Conflict has gave me valuable insights into the up-down trend of activity of this conflict. The analysis revealed clear temporal trends, hotspots of conflict, and clear involvement of various parties to escalate and de-escalate the on-going war.

The overall trends in fatal incidents over time are increasing with periods of relative peace and increased violence. With Palestinian suffering more with an estimated 10,000 casualties starting from 2000 to present. While Israeli have an estimated 1,000 casualties from the same timeframe. The graphs Total Recorded Events and Casualties by Year with 2 subgraphs of count of events and count of deaths shows the said circumstance. Also, the graph Casualties done by involved parties per year showed the same temporal trend with an emphasize to the Israeli security forces leading the cause of casualties throughout this conflict.

Hotspots of fatal incidents clearly indicates that the Gaza City are leading with an approximate of 2200 casualties, followed by Rafah with estimated 800 casualties. Comparing the two hotspts, Gaza City lead by a mile. One of the reasons are Gaza have 2.3 million people compared to 200,000 population of Rafah. This means that more people are at risk in Gaza but also, Gaza has been under the Israeli-Egyptian blockade for over 15 years now. This means that Palestinian peole trying to go to Gaza for food, medicine and other supplies are more vulnerable to attacks. Another hotspot for fatal incidents is the Rafah, which is a home for a large number of Palestinian refugee.

The de-escalation efforts from both Palestine and Israel are conducted through peace talks, ceasefires, and international mediation from United Nations. One of the notable ceasefire happened during 2014 called 2014 Gaza ceasefire. The data supports this because there is a big dip of violence in the graph in 2015 following a downward trend until 2017. This proves that peacetalks and peace agreements are possible only if they both honor it to reach a lasting peace. In terms of escalation, both parties are doing rocket attacks, airstrikes, and ground invasions. The advancement of this technology increased the usage of missiles, drones, and tanks which also increased the number of civilian casualties. An estimated 2,900 deaths are because of missile, followed by live ammunition with 1,500. Moreover, the illegal settlement construction of Israel in the West bank are making it more difficult for the two parties to reach a resolution to this conflict.

This EDA has laid the foundation for more in-depth research. Future studies could explore the underlying causes of temporal trends, including the impact of specific incidents and external factors. Research could also investigate the vulnerability of specific age groups and genders in fatal incidents. First, more in-depth research. This analysis serves as a starting point for more extensive research. Future studies should delve into the root causes of temporal trends, including specific incidents and external factors. Additionally, research can explore the vulnerability of specific demographic groups during fatal incidents.

Explore geography further. Expanding our analysis to consider additional factors like socioeconomic conditions and population density can offer a more complete understanding of conflict hotspots.

Creating a data informed policies. Policymakers and conflict resolution experts can utilize these insights to create strategies for managing and resolving the Israeli-Palestinian conflict effectively.

Transparent data. Ensuring that data remains transparent and accessible is crucial for future research. Collaborative efforts to maintain comprehensive datasets on conflict incidents will deepen our understanding of conflict dynamics.These recommendations provide guidance for further exploration and action to better comprehend and address the Israeli-Palestinian conflict.

VI. References

https://www.unrwa.org/2014-gaza-conflict

https://www.aljazeera.com/news/2023/10/9/whats-the-israel-palestine-conflict-about-a-simple-guide#:~:text=But%20what%20unfolds%20in%20the,as%20intractable%2C%20complicated%20and%20deadlocked.

https://mfa.gov.il/Jubilee-years/Pages/1947-UN-General-Assembly-Resolution-181-The-international-community-says-Yes-to-the-establishment-of-the-State-of-Israel.aspx#:~:text=Re%2Dbirth%20of%20a%20nation&text=On%20that%20day%20the%20UN,decision%20to%20terminate%20its%20Mandate.

https://world101.cfr.org/understanding-international-system/conflict/israeli-palestinian-conflict-timeline

https://www.unicef.org/mena/documents/gaza-strip-humanitarian-impact-15-years-blockade-june-2022

https://www.theguardian.com/world/2014/aug/26/gaza-ceasefire-israel-palestinians-halt-fighting#:~:text=But%20the%20terms%20of%20the,previous%20war%2021%20months%20ago.

https://press.un.org/en/2023/sc15424.doc.htm#:~:text=Pointing%20to%20ongoing%20expansion%20of,to%20return%20to%20peace%20negotiations.

VII. Annex

A. Scripts and codes

```python
# Display the number of rows and columns
num_rows, num_columns = df.shape
print(f"Number of Rows: {num_rows}")
print(f"Number of Columns: {num_columns}")

# Show column names and data types
print("Column Names and Data Types:")
print(df.dtypes)

# Display basic summary statistics
summary_statistics = df.describe()
print("\nSummary Statistics:")
print(summary_statistics)

# Check for missing values
missing_values = df.isnull().sum()
print("\nMissing Values:")
print(missing_values)
```

```
[8] Number of Rows: 11124
Number of Columns: 16
Column Names and Data Types:
name                              object
date_of_event                     object
age                              float64
citizenship                       object
event_location                    object
event_location_district           object
event_location_region             object
date_of_death                     object
gender                            object
took_part_in_the_hostilities      object
place_of_residence                object
place_of_residence_district       object
type_of_injury                    object
ammunition                        object
killed_by                         object
notes                             object
dtype: object

Summary Statistics:
                age
count  10995.000000
mean      26.745703
std       13.780548
min        1.000000
25%       19.000000
50%       23.000000
75%       31.000000
max      112.000000

Missing Values:
name                                 0
date_of_event                        0
age                                129
citizenship                          0
event_location                       0
event_location_district              0
event_location_region                0
date_of_death                        0
gender                              20
took_part_in_the_hostilities      1430
place_of_residence                  68
place_of_residence_district         68
type_of_injury                     291
ammunition                        5253
killed_by                            0
notes                              280
dtype: int64
```

```
import io
df = pd.read_csv(io.BytesIO(uploaded['dataset.csv']))
```

```
[83] import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from matplotlib.gridspec import GridSpec
```

Data Cleaning Code

```
[49] print(f'size of set before delete duplicates: {df.shape}')
     df.drop_duplicates(inplace=True)
     print(f'duplicates deleted, {df.shape} left')
```

```
[16] missing_values_count = df.isnull().sum()
     print(missing_values_count)
```

```
name                              0
date_of_event                     0
age                             122
citizenship                       0
event_location                    0
event_location_district           0
event_location_region             0
date_of_death                     0
gender                           14
took_part_in_the_hostilities    1430
place_of_residence                61
place_of_residence_district       61
type_of_injury                   290
ammunition                      5246
killed_by                         0
notes                           277
dtype: int64
```

```
[50] # quantity of women and man
     w_m=df.groupby('gender').gender.count().sort_values(ascending=False).index[0]
     # imputing missing values in gender column with the most frequent value
     df.gender.fillna(w_m, inplace=True)
```

```
[19] mean_man = df[df['gender'] == 'M']['age'].mean()
     print("Mean age of males:", mean_man)

     mean_woman = df[df['gender'] == 'F']['age'].mean()
     print("Mean age of females:", mean_woman)

     # Imputing missing values in age column with the each mean value
     df.loc[df['gender'] == 'M', 'age'] = df.loc[df['gender'] == 'M', 'age'].fillna(mean_man)
     df.loc[df['gender'] == 'F', 'age'] = df.loc[df['gender'] == 'F', 'age'].fillna(mean_woman)
```

```
[21] # imputing missing values in took_part_in_the_hostilities column with the most frequent value
     df.took_part_in_the_hostilities.fillna('No', inplace=True)
     missing_values_count = df.took_part_in_the_hostilities.isnull().sum()

     # imputing missing values in 'place_of_residence' column with the most frequent value
     max_place_of_residence=df.groupby('place_of_residence').place_of_residence.count().sort_values(ascending=False).index[0]
     df.place_of_residence.fillna(max_place_of_residence, inplace=True)

     # imputing missing values in 'place_of_residence_district ' column with the most frequent value
     residence_district=df.groupby('place_of_residence_district').place_of_residence_district.count().sort_values(ascending=False).index[0]
     df.place_of_residence_district.fillna(residence_district, inplace=True)

     injury=df.groupby('type_of_injury').type_of_injury.count().sort_values(ascending=False).index[0]
     df.type_of_injury.fillna(injury, inplace=True)

     max_ammunition=df.groupby('ammunition').ammunition.count().sort_values(ascending=False).index[0]
     df.ammunition.fillna(max_ammunition, inplace=True)
```

```
[22] missing_values_count = df.isnull().sum()
     print(missing_values_count)

name                              0
date_of_event                     0
age                               0
citizenship                       0
event_location                    0
event_location_district           0
event_location_region             0
date_of_death                     0
gender                            0
took_part_in_the_hostilities      0
place_of_residence                0
place_of_residence_district       0
type_of_injury                    0
ammunition                        0
killed_by                         0
notes                           277
dtype: int64
```

Visualization Code

```
[71]

     age_groups = ['Child (1-18 yrs old)', 'Young Adult (19-25 yrs old)', 'Adult (26-65 yrs old)', 'Elderly(66 yrs old and up)']
     df['age_groups'] = pd.cut(df['age'], bins=[0, 18, 25, 65, 112], labels=age_groups)

     plt.figure(figsize=(10, 6))
     ax = sns.countplot(x=df['age_groups'], palette='icefire')
     plt.title('Age Group Frequency')

     ax.set_ylim(0, max(df['age_groups'].value_counts()))

     plt.xlabel('Age groups')
     plt.ylabel('Count')
     plt.show()
```

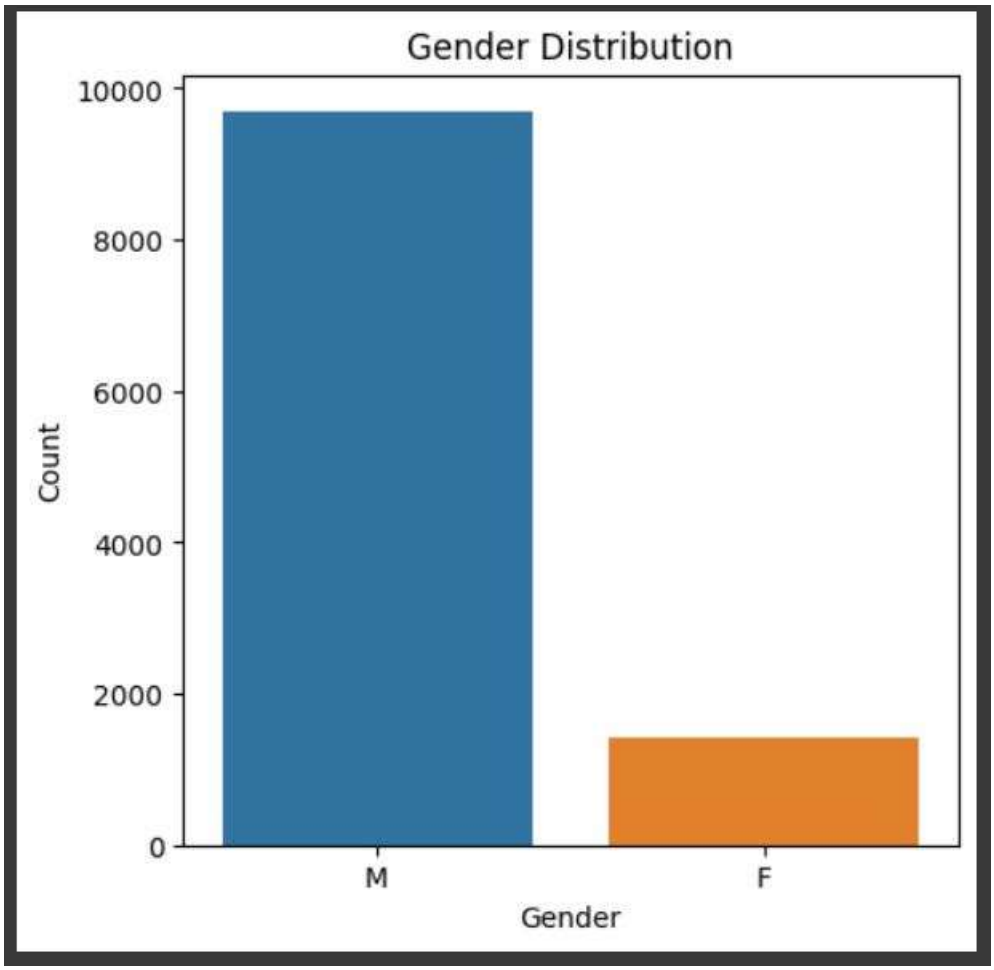
```
#Bar plot for 'gender'
gender_counts = df['gender'].value_counts()
plt.figure(figsize=(5,5))
sns.barplot(x=gender_counts.index, y=gender_counts.values)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

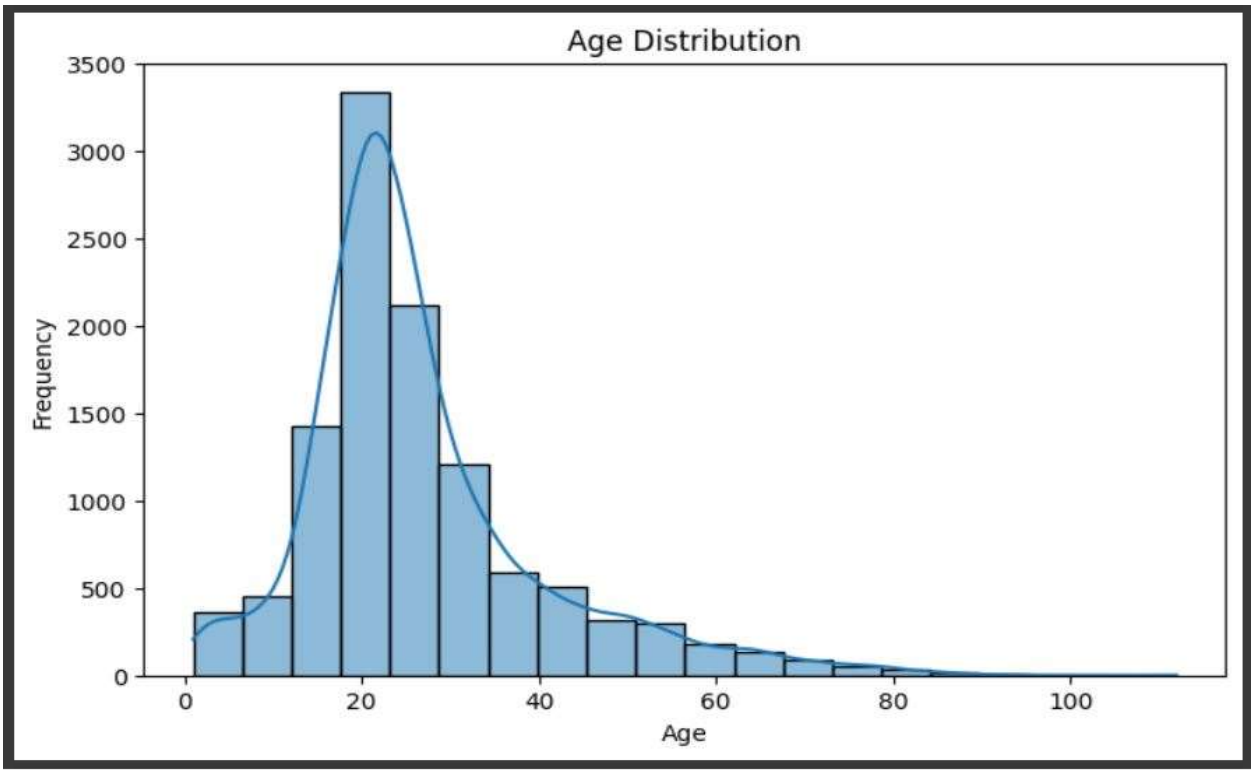
```
[69] citizenship_count = df['citizenship'].value_counts().sort_values(ascending=False)

     plt.figure(figsize=(10, 6))
     plt.bar(citizenship_count.index, citizenship_count.values)
     plt.title('Citizenship Distribution')
     plt.xlabel('Citizenship')
     plt.ylabel('Count')
     plt.xticks(rotation=90)
     plt.grid(True)

     # Set y-axis ticks at every 100
     plt.yticks(range(0, citizenship_count.max() + 1, 300))

     plt.show()
```

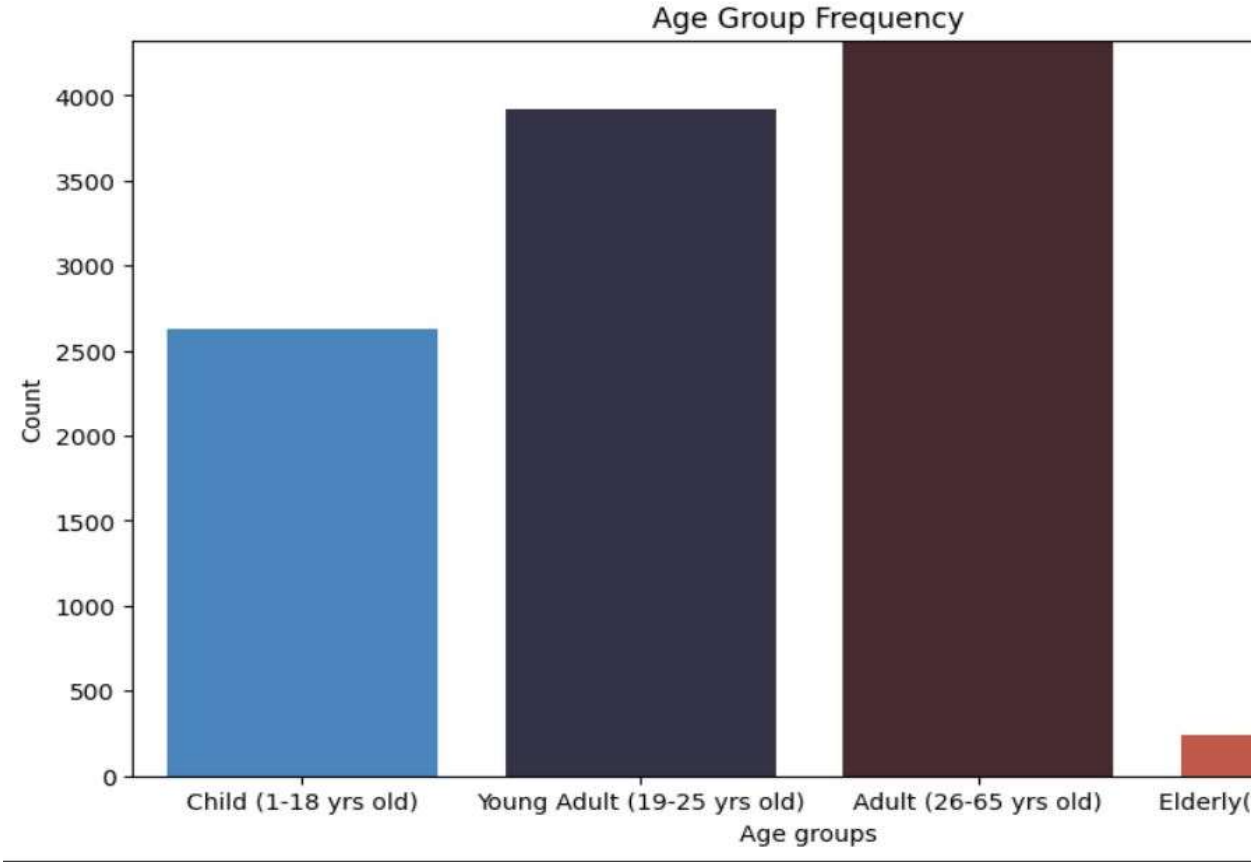
```
[31] # Histogram for the 'age' column
     plt.figure(figsize=(8, 5))
     sns.histplot(df['age'], bins=20, kde=True)
     plt.title('Age Distribution')
     plt.xlabel('Age')
     plt.ylabel('Frequency')
     plt.show()
```

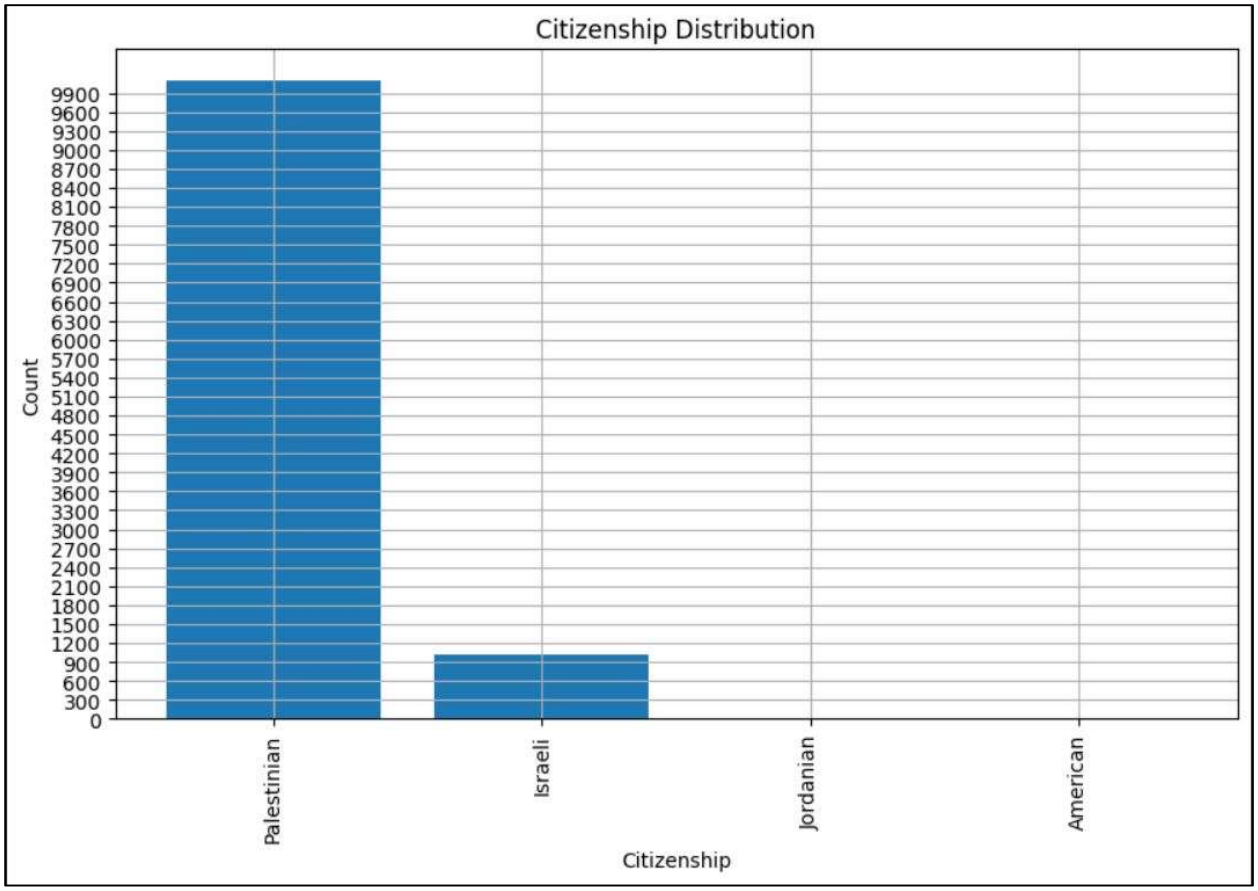
```
[88] import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Assuming 'date_of_event' and 'date_of_death' are in string format, convert them to datetime objects
     df['date_of_event'] = pd.to_datetime(df['date_of_event'])
     df['date_of_death'] = pd.to_datetime(df['date_of_death'])

     # Create a DataFrame with a DatetimeIndex using 'date_of_event'
     df_event = df.set_index('date_of_event')

     # Create a DataFrame with a DatetimeIndex using 'date_of_death'
     df_death = df.set_index('date_of_death')

     fig = plt.figure(figsize=(10, 8))

     # First subplot using 'date_of_event'
     ax1 = fig.add_subplot(211)
     df_event.resample('Y')['name'].count().plot(cmap='plasma', title='Total recorded Events and Casualties by year')
     ax1.set_xlabel('Year')
     ax1.set_ylabel('Count of Events')
     plt.grid()

     # Second subplot using 'date_of_death'
     ax2 = fig.add_subplot(212)
     df_death.resample('Y')['name'].count().plot(cmap='plasma')
     ax2.set_xlabel('Year')
     ax2.set_ylabel('Count of Deaths')
     plt.grid()

     plt.tight_layout()
     plt.show()
```

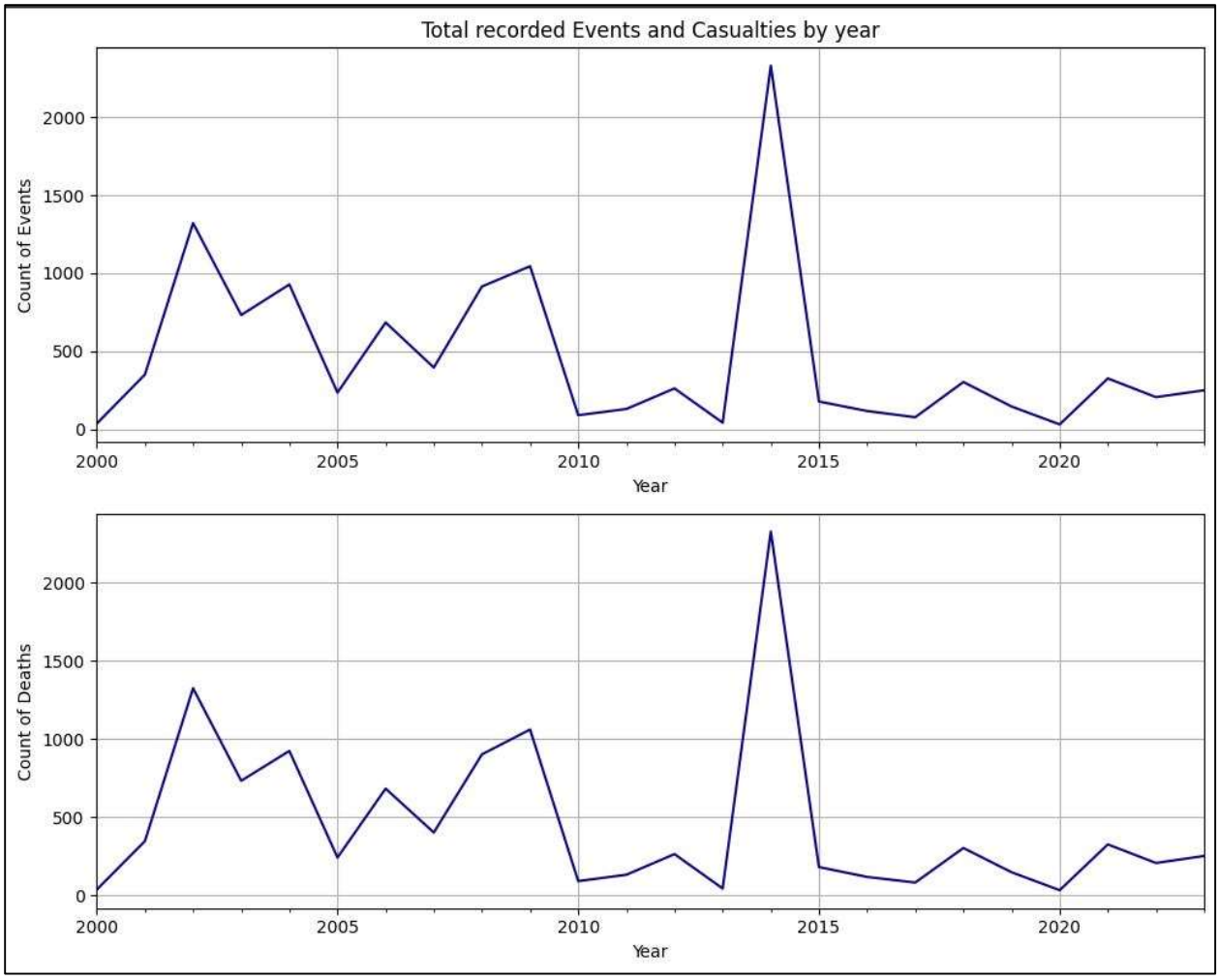
Visualizations

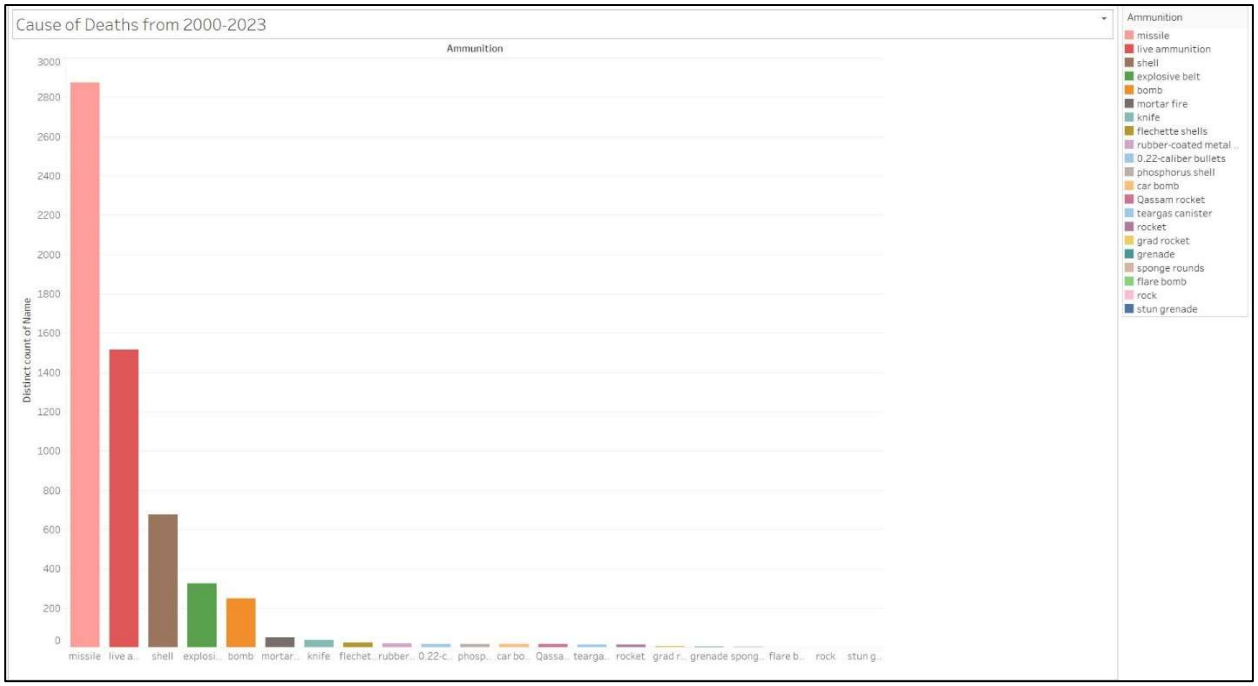


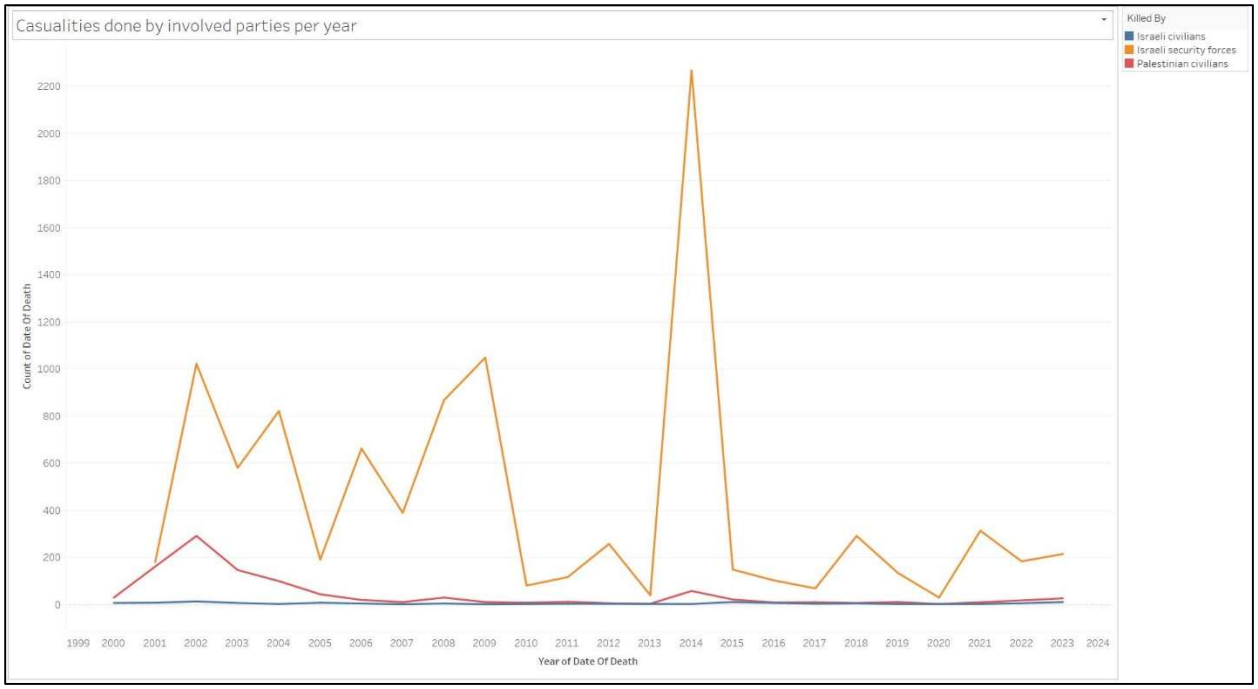

Age Distribution
Frequency
Age
0
500
1000
1500
2000
2500
3000
3500
0
20
40
60
80
100

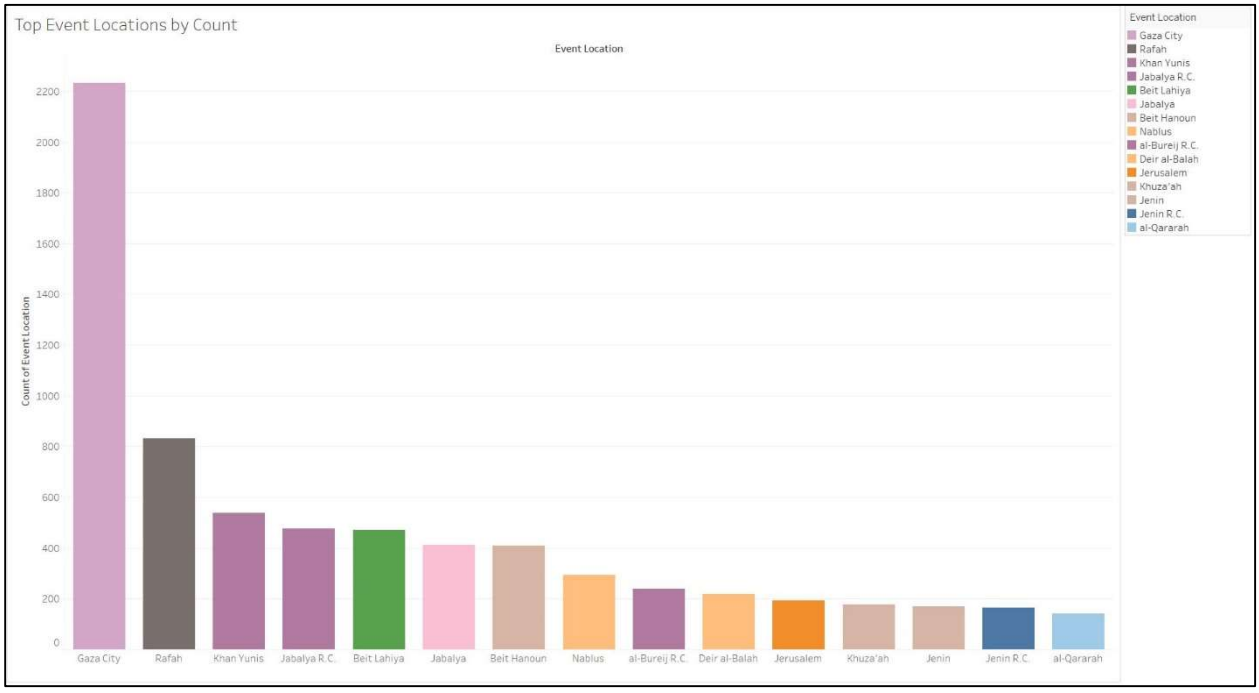

Age Group Frequency
Count
Age groups
0
500
1000
1500
2000
2500
3000
3500
4000
Child (1-18 yrs old)
Young Adult (19-25 yrs old)
Adult (26-65 yrs old)
Elderly(


Citizenship Distribution
Count
9900
9600
9300
9000
8700
8400
8100
7800
7500
7200
6900
6600
6300
6000
5700
5400
5100
4800
4500
4200
3900
3600
3300
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
Palestinian
Israeli
Jordanian
American
Citizenship


Total recorded Events and Casualties by year
Count of Events
2000
1500
1000
500
0
2000
2005
2010
2015
2020
Year
Count of Deaths
2000
1500
1000
500
0
2000
2005
2010
2015
2020
Year


Cause of Deaths from 2000-2023
Ammunition
Distinct count of Name
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
missile
live a..
shell
explosi..
bomb
mortar..
knife
flechet..
rubber..
0.22-c..
phosp..
car bo..
Qassa..
tearga..
rocket
grad r..
grenade
spong..
flare b..
rock
stun g..
Ammunition
missile
live ammunition
shell
explosive belt
bomb
mortar fire
knife
flechette shells
rubber-coated metal ..
0.22-caliber bullets
phosphorus shell
car bomb
Qassam rocket
teargas canister
rocket
grad rocket
grenade
sponge rounds
flare bomb
rock
stun grenade


Casualities done by involved parties per year
Killed By
Israeli civilians
Israeli security forces
Palestinian civilians
Count of Date Of Death
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024
Year of Date Of Death


Top Event Locations by Count
Event Location
Gaza City
Rafah
Khan Yunis
Jabalya R.C.
Beit Lahiya
Jabalya
Beit Hanoun
Nablus
al-Bureij R.C.
Deir al-Balah
Jerusalem
Khuza'ah
Jenin
Jenin R.C.
al-Qararah
Count of Event Location
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Gaza City
Rafah
Khan Yunis
Jabalya R.C.
Beit Lahiya
Jabalya
Beit Hanoun
Nablus
al-Bureij R.C.
Deir al-Balah
Jerusalem
Khuza'ah
Jenin
Jenin R.C.
al-Qararah