1. Dataset "**Global Missing Migrants"**

A. Data Source Summary

- **Description**: This dataset provides a comprehensive record of missing migrants and their tragic journeys towards international destinations , collected by the Missing Migrants Project, an initiative implemented by **the International Organization for Migration (IOM**) since 2014. The dataset documents deaths and disappearances, shedding light on the challenges migrants face during their journeys. Please note that due to the complexities of data collection, the figures presented are likely an undercount. The dataset serves as a tribute to the individuals who lost their lives, as well as the families and communities impacted by their absence.

- **Source:** https://www.kaggle.com/datasets/nelgiriyewithana/global-missing-migrants-dataset?resource=download

- **Type of Data:** External Data

- **Owner:** International Organization for Migration (IOM) -Missing Migrants Project

- **Trustworthiness:** IOM, part of the UN System, is the leading inter-governmental organization for migration since 1951, with 175 member states and a presence in 171 countries, ensuring reliable and systematic data collection

B. **Data Collection Method**

Description: The data in this dataset was collected by the International Organization for Migration's (IOM) Missing Migrants Project. This project relies on a range of sources, including official records, media reports, and NGO reports, to track and document cases of missing migrants globally.

Time Period: The dataset covers incidents from 2014 -2023.

Data Collection Process:

- **Sources:** Data is aggregated from various sources, such as coast guard reports, testimonies from survivors, and media reports.
- **Verification:** Efforts are made to cross-verify information from different sources to ensure accuracy. However, the data is still likely an undercount due to the challenges in documenting migrant fatalities, particularly in remote or inaccessible regions.
- **Time Period:** The dataset covers incidents from 2014 onward, with regular updates to reflect new information as it becomes available.

C. Overview of Data Contents:

a. **Total Rows:** 13,020
b. **Columns:** 19

Columns:

- **Incident Type**: *Type of migration incident:*
  Incident:
  This typically refers to a single event or occurrence. For example, an incident might be a single reported case of a disease, a single accident, or a single crime.
  Cumulative Incident:
  This refers to the total number of incidents over a specified period. It is a running total that accumulates over time. For instance, if you are tracking the number of new cases of a disease each day, the cumulative incident would be the total number of cases up to that day.
  Split Incident:
  This term is less common and could refer to an incident that is divided into sub-categories or parts. For example, an incident that affects multiple categories or involves multiple types of events might be split into different parts for detailed analysis.
  Incident, Split Incident:
  This seems to be a combination of the above terms, indicating that the data point is categorized as both an individual incident and a split incident. This could be used to denote incidents that have been recorded individually but also need to be considered as part of a split or multi-faceted event.
- **Incident Year**: *Year when the incident occurred*
- **Reported Month**: *Month when the incident was reported*
- **Region of Origin**: *Geographical region where the migrants originated*
- **Region of Incident**: *Geographical region where the incident occurred*
- **Country of Origin**: *Country from which the migrants originated*
- **Number of Dead**: *Number of confirmed deceased migrants*
- **Minimum Estimated Number of Missing**: *Minimum estimated count of missing migrants*
- **Total Number of Dead and Missing**: *Total count of both deceased and missing migrants*
- **Number of Survivors**: *Number of migrants who survived the incident*
- **Number of Females**: *Number of female migrants involved*
- **Number of Males**: *Number of male migrants involved*
- **Number of Children**: *Number of children migrants involved*
- **Cause of Death**: *Cause of death for the migrants*
- **Migration Route**: *Route taken by migrants during their journey (if available)*
- **Location of Death**: *Approximate location where the incident occurred*
- **Information Source**: *Source of information about the incident*
- **Coordinates**: *Geographical coordinates of the incident location*
- **UNSD Geographical Grouping**: *Geographical grouping according to the United Nations Statistics Division*

D. Reason why I choose this Dataset:

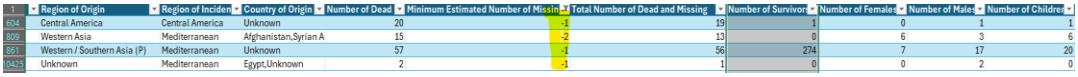
As an immigrant, I have a personal connection to the topic of migration. Fortunately, my experience was under different circumstances and by my own choice, which makes me deeply appreciative of the opportunities I've had. However, I also feel a strong sense of responsibility to recognize and understand the harsh realities faced by those who are not as fortunate, who are driven by the search for better opportunities. This search often comes with significant risks, including the possibility of death or disappearance.

This dataset provides crucial, concrete information that helps us understand the immense human cost of migration in these cases. It serves not only as a record but also as a solemn reminder of the dangers migrants face. By analyzing this data, we can foster greater empathy and encourage more institutions to take meaningful actions to address and mitigate these risks.

2. **Data Profile**

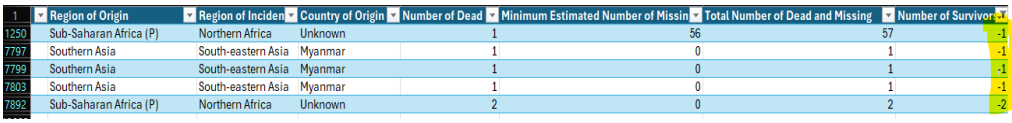
Missing Data:

- Region of Origin: 22 missing values, replaced with "Unknown".
- Country of Origin: 8 missing values, replaced with "Unknown".
- Number of Dead: 550 missing values, replaced with 0.
- Minimum Estimated Number of Missing: 4 rows with negative values (-1 and -2) were removed.



	Region of Origin	Region of Inciden	Country of Origin	Number of Dead	Minimum Estimated Number of Missin	Total Number of Dead and Missing	Number of Survivor	Number of Female	Number of Male	Number of Childre
604	Central America	Central America	Unknown	20	-1	19	1	0	1	1
809	Western Asia	Mediterranean	Afghanistan,Syrian A	15	-2	13	0	6	3	6
861	Western / Southern Asia (P)	Mediterranean	Unknown	57	-1	56	274	7	17	20
9425	Unknown	Mediterranean	Egypt,Unknown	2	-1	1	0	0	2	0

- Number of Survivors: Rows with negative values were removed.



	Region of Origin	Region of Inciden	Country of Origin	Number of Dead	Minimum Estimated Number of Missin	Total Number of Dead and Missing	Number of Survivor
1250	Sub-Saharan Africa (P)	Northern Africa	Unknown	1	56	57	-1
7797	Southern Asia	South-eastern Asia	Myanmar	1	0	1	-1
7799	Southern Asia	South-eastern Asia	Myanmar	1	0	1	-1
7803	Southern Asia	South-eastern Asia	Myanmar	1	0	1	-1
7892	Sub-Saharan Africa (P)	Northern Africa	Unknown	2	0	2	-2

- Migration Route: 3,021 missing values, replaced with "Unknown".
- Coordinates: 36 missing values, replaced with "Unknown".
- Information Source: 8 missing values, replaced with "Unknown".
- UNSD Geographical Grouping: 1 missing value, replaced with "Unknown"

Duplicate:

- 641 duplicate rows were identified and removed to ensure data integrity.

Data Types:

- Most columns are either int64 or object (string).
- **Number of Dead:** Originally stored as float64, this column was converted to int64 to reflect that the number of dead should be a whole number.

- The **"Coordinates" field** was split into latitude and longitude for potential future geographic visualizations.

- **Data shape after removing duplicates:** The original dataset had 13,020 rows, but after removing 641 duplicate rows, the dataset now contains fewer rows.
- **Split Coordinates column into latitude and longitude:** The original dataset had a single "Coordinates" column, which has now been split into two separate columns, "Latitude" and "Longitude." This operation added one additional column to the dataset.
- **Missing data handling:** This involves filling in or replacing missing values with "Unknown" or other appropriate placeholders and removing rows where negative values were found in certain columns.

So, after performing these operations, the dataset now has **12,371 rows** and **20 columns**, reflecting the removal of duplicates, the addition of the latitude and longitude columns, and the handling of missing data.

3. **Statistics**:

	Incident year	Number of Dead	Minimum Estimated Number of Missing	Total Number of Dead and Missing	Number of Survivors	Number of Females	Number of Males	Number of Children
count	12371.00	12371.00	12371.00	12371.00	12371.00	12371.00	12371.00	12371.00
mean	2019.02	2.58	2.00	4.57	6.88	0.48	1.13	0.25
std	2.43	9.76	16.04	20.44	43.32	2.92	3.91	2.47
min	2014.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
25%	2017.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
50%	2019.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
75%	2021.00	2.00	0.00	2.00	0.00	0.00	1.00	0.00
max	2023.00	750.00	750.00	1022.00	1950.00	94.00	135.00	250.00

Explanation of Key Statistics:

Number of Dead:

- Minimum (1): The minimum value in the "Number of Dead" column is 1, indicating incidents with a single reported death.
- Maximum (750): The maximum value is 750, reflecting large-scale disasters like shipwrecks involving many migrants.
- Mean (2.58): The mean indicates that most incidents involve a small number of deaths, but there are extreme cases (like the incident with 750 deaths) that skew the average.

Minimum Estimated Number of Missing:

- Minimum (0): After removing rows with negative values, the minimum number of missing migrants is 0, indicating incidents where no one was reported missing.
- Maximum (750): The maximum value is 750, representing significant incidents where many individuals were unaccounted for.
- Mean (2.00): The mean suggests that while some incidents involve large numbers of missing people, many involve only a few, which lowers the average.

4. **Limitations, potential bias and Ethical Considerations:**

Limitations:

- **Underreporting**: The dataset likely underreports the actual number of migrant deaths and disappearances due to the difficulty of obtaining accurate data in certain regions.
- **Inconsistent Reporting**: Variations in how incidents are reported and documented across different regions could introduce inconsistencies in the data.

Potential Bias:

- **Source Bias:** The reliance on media and NGO reports may introduce bias, as certain incidents may receive more attention than others, skewing the data.
- **Geographical Bias:** Some regions might have better reporting mechanisms, leading to an overrepresentation of incidents in those areas compared to more remote or conflict-prone regions.

Ethical Considerations:

- **Data Sensitivity**: The data deals with human tragedies, and care should be taken in its analysis and presentation to respect the individuals and communities affected.
- **Privacy:** While the dataset does not contain personal data, the sensitive nature of the incidents warrants careful consideration of how the data is used and shared.

5. Questions

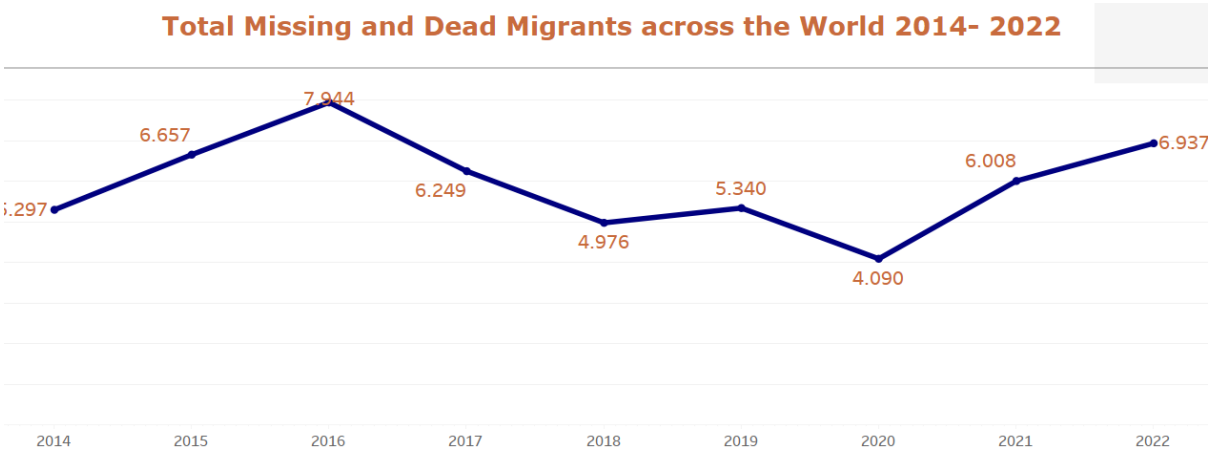
Demographic Analysis

- **Regions of Origin:** What are the most common regions of origin for missing migrants?
  - **Answer: Sub-Saharan Africa remains the dominant region of origin for missing and dead migrants, followed by Eastern Africa and Northern Africa.**
- **Gender Distribution:** How does the gender distribution of missing migrants vary across different regions or incidents?
  - **Answer: Male Dominance:** The number of men involved in incidents consistently exceeds that of females and children. This suggests that men may face disproportionate risks during migration.
  - **Fluctuating Trends:** The number of incidents involving each gender has fluctuated over the years, highlighting the dynamic nature of migration patterns and risks.
  - **Overall Increase:** While there have been fluctuations, the overall trend indicates an increase in incidents from 2014 to 2018, followed by a decline and then a slight increase in 2023.

Incident and Mortality Analysis

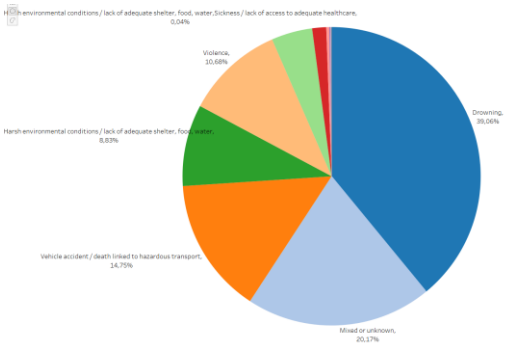
- **Dangerous Migration Routes:** Which migration routes are associated with the highest number of deaths and disappearances?
- **Answer:** The **Mediterranean** region, in particular, has a significantly higher proportion of casualties compared to other areas.

- **Trend Over Time:** How has the number of migrant deaths and missing persons changed over the years (from 2014 to 2023)? Data belong until July 2023.



Cause of Death and Incident Type Analysis

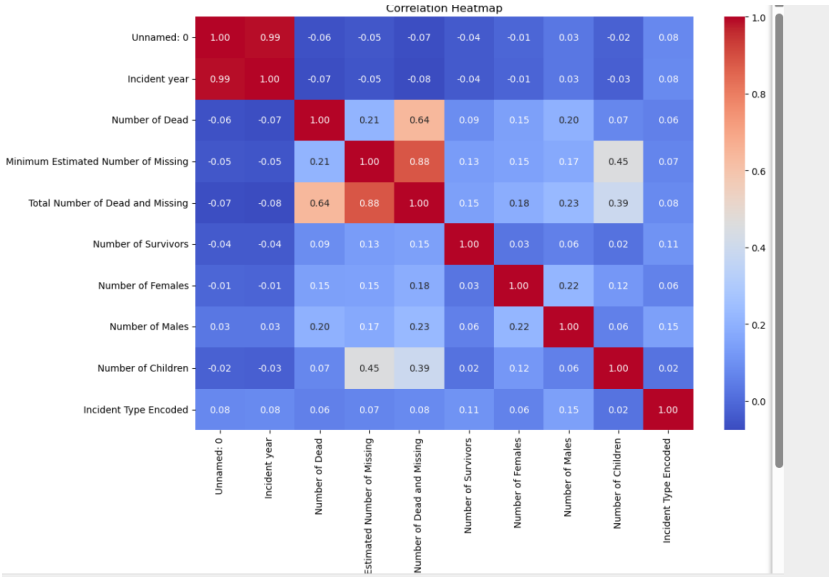
- **Common Causes of Death:** What are the most common causes of death among missing migrants?
  - **Answer: Drowning** (39%) is the most significant cause of death across all regions, highlighting the perilous nature of sea crossings.
  - **Mixed or Unknown Causes:** (20.17%) are attributed to "Mixed or Unknown" causes, indicating the need for further investigation and improve data collection.
  - **Vehicular Accidents:(**14%)of deaths, suggesting that land-based journeys also pose significant risks



Incident Type Severity: Are there certain incident types that are more likely to result in a higher number of deaths?

- **Answer: Drowning** (39%) is the most significant cause of death across all regions, highlighting the perilous nature of sea crossings.**Geographic and Environmental Analysis**

Correlation Analysis



- **Region of Origen and Casualties:** Is there a correlation between the number of migrants from a particular region of origin and the total number of dead and missing?
- This question was changed to **Region of Incident and Casualties:** Is there a correlation between the region of incident and the total number of dead and missing? because Shifting the analysis to focus on the "Region of Incident" rather than the "Region of Origin" can provide valuable insights. The "Region of Incident" likely has a more direct relationship with the number of dead and missing, as it points to where these tragic events are happening, which could be linked to dangerous routes, political instability, or other local factors.
    - Answer: Regions like the Mediterranean and South-eastern Asia tend to have higher overall numbers of casualties

- **Dead and Missing Correlation:** How strongly are the number of dead and the number of missing correlated?
    - **Answer: Overall, the box plot reveals that the data is highly skewed, with a concentration of incidents resulting in a lower number of casualties and a few outlier events with significantly higher numbers of casualties.**

- **Survivors vs. Casualties:** What is the correlation between the number of survivors and the number of dead and missing?

  The scatter plot shows a weak positive correlation between the number of survivors and the total number of dead and missing in migration incidents. This means that, generally, incidents with a higher number of survivors also tend to have a higher total number of dead and missing.

Hypothesis Testing: Regional Analysis

- **Proportion of Missing Migrants by Region:**

- **Null Hypothesis (H0):** The proportion of missing migrants is the same across different regions.
- **Alternative Hypothesis (H1):** Certain regions have a significantly higher proportion of missing migrants compared to others.

- Answer: The proportion of missing and dead migrants is **not** uniform across different regions. The number of casualties within each region has fluctuated over time, indicating varying levels of risk during different periods.