women

MDPI

Article

Enhancing Gender-Based Violence Research: Holistic Approaches to Data Collection and Analysis

Subeksha Shrestha *, Preeti Patel, Sentirenla Longchar and Aiswarya Francis Xavier

Computer Science and Applied Computing, London Metropolitan University, 166–220 Holloway Road, London N7 8DB, UK; p.patel@londonmet.ac.uk (P.P.); sel0217@my.londonmet.ac.uk (S.L.); aif0070@my.londonmet.ac.uk (A.F.X.)

* Correspondence: s.shrestha3@londonmet.ac.uk

Abstract: Gender-based violence (GBV) is a profound and pervasive societal issue, disproportionately affecting women across diverse settings, including homes, workplaces, and public spaces. Despite its prevalence, significant challenges impede research on GBV, particularly regarding data collection, analysis, and ethical handling. This study investigates the complexities inherent in GBV research, focusing on the obstacles posed by under-reporting, ethical considerations, data quality, and the need for cross-comparative standards. Using a combination of police records, web scraping, news reports, and survey data from USAID's Demographic and Health Surveys (DHS), our study examines strategies to work with sensitive GBV datasets, while maintaining data integrity. Our study advocates for improved demographic surveying and data integration methodologies that can enhance data accuracy and comparability. The findings suggest that while technological advancements, particularly generative AI and machine learning approaches, offer promising avenues for automating survey processes, reducing costs, and enhancing data collection efficiency, they present the limitations of secondary datasets, a lack of data disaggregation, and discrepancies in data coding systems, which highlight the necessity of refining global data standards.

Keywords: gender-based violence; data challenges; demographic survey; data collection

Academic Editor: Peter Schmidt

Received: 4 April 2025
Revised: 30 April 2025
Accepted: 27 May 2025
Published: 30 May 2025

Citation: Shrestha, S.; Patel, P.; Longchar, S.; Xavier, A.F. Enhancing Gender-Based Violence Research: Holistic Approaches to Data Collection and Analysis. *Women* **2025**, *5*, 19. https://doi.org/10.3390/women5020019

1. Introduction

Gender-based violence (GBV) remains a pervasive issue in contemporary society, with a significant disparity observed particularly against women. GBV refers to acts of physical, sexual, or emotional abuse committed against individuals based on their gender, often rooted in power imbalances and socially constructed norms regarding gender roles. Women are frequently victims of such violence in various settings, including homes, workplaces, schools, and public spaces. A survey conducted by the World Health Organization (WHO) indicates that one in three women globally has experienced physical or sexual violence at some point in their lives [1]. Several social factors, such as economic conditions, lifestyle, education, and employment status, influence the likelihood of women becoming targets of GBV [2]. Following the COVID-19 pandemic, there was an increase in gender-based violence, with many perpetrators, particularly young mothers' partners, conducting violence due to job loss, economic instability, and heightened stress levels.

Data collection, analysis, and sharing present significant challenges in GBV research, impeding efforts to provide necessary support. A persistent gender data gap, exacerbated by inadequate data collection methods, often fails to capture the experiences of women and girls. Areas particularly affected by this gap include workforce statistics, unpaid

care work, civic engagement, and the use of public services. Even in healthcare, where sex-disaggregated data are critical, the WHO only began disaggregating its Global Health Statistics by sex in 2019. Women's experiences, especially those related to violence, are frequently under-reported or inadequately documented. The importance of accurate and timely data collection has been highlighted by policymakers worldwide. For example, the UK Government's 2021 "Tackling Violence Against Women and Girls Strategy" [3] emphasizes the need for improved data to enhance the understanding of these crimes. The European Institute for Gender Equality (EIGE) has also stressed the importance of data collection that enables comparability across contexts. While law enforcement agencies routinely record data on violence against women, such administrative data are not typically collected for analytical purposes and, as a result, fail to capture the full extent of unreported GBV incidents.

Working with GBV data presents numerous concerns related to data integrity, reliability, sourcing appropriate data, and developing comprehensive models to uncover underlying patterns. This research seeks to navigate these complexities by emphasizing collaborative efforts and proposing measures to overcome obstacles in data collection and analysis. Research on GBV is particularly complex due to the sensitive nature of the subject, with privacy and confidentiality concerns for both victims and perpetrators often limiting access to data. When crucial data are inaccessible, balancing the need for comprehensive data with ethical considerations becomes a significant challenge. In addition to privacy concerns, ethical considerations in GBV research include the risk of re-traumatising victims and navigating cultural sensitivities, which may prevent participants from sharing their experiences. Methodological challenges include sampling biases across different nations and ensuring the accuracy of collected data. Researchers must also take care to avoid using personal or sensitive data in ways that could lead to misrepresentation or overgeneralization based on limited information. This research aims to explore holistic approaches to overcoming these barriers and proposes potential methods for working with sensitive datasets.

While this study focuses primarily on quantitative secondary data sources, we recognize the crucial role of qualitative methods, such as life histories and exploratory interviews, in providing deeper insights into the lived experiences of GBV survivors. Qualitative approaches capture dimensions of violence that are often overlooked in large-scale datasets and remain an important complementary area for future research. As this study concentrates on evaluating and comparing existing large-scale quantitative datasets for cross-country analysis, qualitative data collection was not incorporated.

To investigate GBV, we utilized multiple data sources, including police records, the web scraping of social media posts, published news reports on violence against women, and survey data. After careful evaluation, we opted to use survey data from USAID, specifically focusing on domestic violence [4]. This data, derived from Demographic and Health Surveys (DHS), encompasses information from over 90 countries and covers topics such as fertility, family planning, maternal and child health, nutrition, and migration. Our study will primarily focus on identifying and addressing limitations associated with sourcing and analysing GBV data.

2. Related Work

A study conducted by [5] critically examines the limitations of existing surveys and proposes improvements for data collection, particularly in relation to violence. Their analysis focuses on a range of surveys, including those from the Fundamental Rights Agency (FRA), the Office of National Statistics (ONS), the Survey of Violence Against Women, and the Crime Survey for England and Wales (CSEW). Despite the wide scope

of these surveys, they are found to fall short in terms of data quality, particularly in addressing violence-related issues. This deficiency in data quality leads to the generation of outcomes that are often unproductive or lacking in meaningful insight. To address these concerns, the study suggests improving survey questionnaires as a means to enhance data collection processes. Similarly, research by [6] addresses barriers associated with collecting data on gender-based violence, while prioritizing the well-being of victims. To prevent re-traumatization, the researchers developed a 14-item checklist designed to protect the privacy, dignity, and safety of participants during data collection. However, the study acknowledges limitations, particularly in terms of the risks posed by self-reported data, which may introduce biases and challenges in generalizing the findings. Another notable gap in GBV research is the tendency for studies to focus on specific regions, which complicates the comparison of global and regional data. For example, reference [7] conducted research on gender-based violence exclusively in Sub-Saharan Africa, limiting the broader applicability of their findings.

Survey data are a commonly employed method for researching GBV, as they allow for the collection of more comprehensive and inclusive data than that provided by under-reported police records. Surveys also offer greater detail about the nature of both victims and perpetrators. However, the COVID-19 pandemic disrupted many surveys, making data collection particularly challenging in certain countries, where reliance on secondary data became the only feasible option [8]. Ostadtaghizadeh et al.'s research highlights potential strategies for combining police reports, hotline data, and surveys from non-governmental organizations, though it is limited by the heavy reliance on secondary data.

In addition to common data challenges, such as under-reporting and geographic limitations, there are significant gaps in data concerning specific forms of violence against women. Methodological issues also arise when collecting data on certain types of violence, such as female genital mutilation, dowry-related violence, trafficking for sexual exploitation, honour-based crimes, and femicide, including intimate partner murders [9]. This United Nations report emphasizes the need for better methodological guidelines when designing and conducting surveys in a sustainable manner. It also underscores the challenge of restricted or censored data, which impedes researchers' ability to produce transparent and unbiased reports. Ethical concerns represent another significant barrier to research on gender-based violence. Legal frameworks surrounding violence vary across nations, influencing both the disclosure of victims' experiences and the outcomes of research. The authors of [10] highlight the emotional toll that studying victims' narratives can have on researchers, often leading to secondary trauma, which can complicate the research process and affect the quality of findings. While these constraints are well-identified, solutions to address them remain elusive, and the absence of comprehensive research guidelines continues to pose a dilemma.

As one of the primary challenges in researching gender-based violence involves data collection, particularly issues of data quality, geographic coverage, and under-reporting, our research utilizes the USAID Demographic and Health Surveys (DHS) dataset, which encompasses data collected from over 90 countries across multiple phases. This extensive dataset allows us to mitigate concerns over data scarcity and ensure the representation of diverse contexts.

Ethical considerations are also paramount in our research, particularly in avoiding the re-traumatisation of victims and ensuring the confidentiality of collected data. We adhere to strict data protection standards, anonymizing all data prior to analysis and dissemination, in line with DHS mandates. By carefully managing the use of these data, we aim to prevent misinterpretation, minimize public disclosure, and reduce participants' fears about engaging in future research.

3. Challenges in Data Selection and Tool Optimization

A primary challenge in researching GBV is the acquisition of high-quality and comprehensive data. A major barrier is the frequent under-reporting of cases, stemming from various factors, including the sensitive nature of GBV, prevailing social stigma, and concerns about personal safety or repercussions, particularly when the perpetrator is a known individual. These issues significantly complicate data collection efforts, making it challenging to access and apply secondary datasets effectively in research projects. The following sections will provide an in-depth examination of the obstacles associated with data sourcing, exploration, analysis, and visualization, with a focus on identifying and utilizing the appropriate tools to effectively tackle these limitations.

3.1. Exploration of Data Sources: Challenges, Comparison, and Selection Rationale

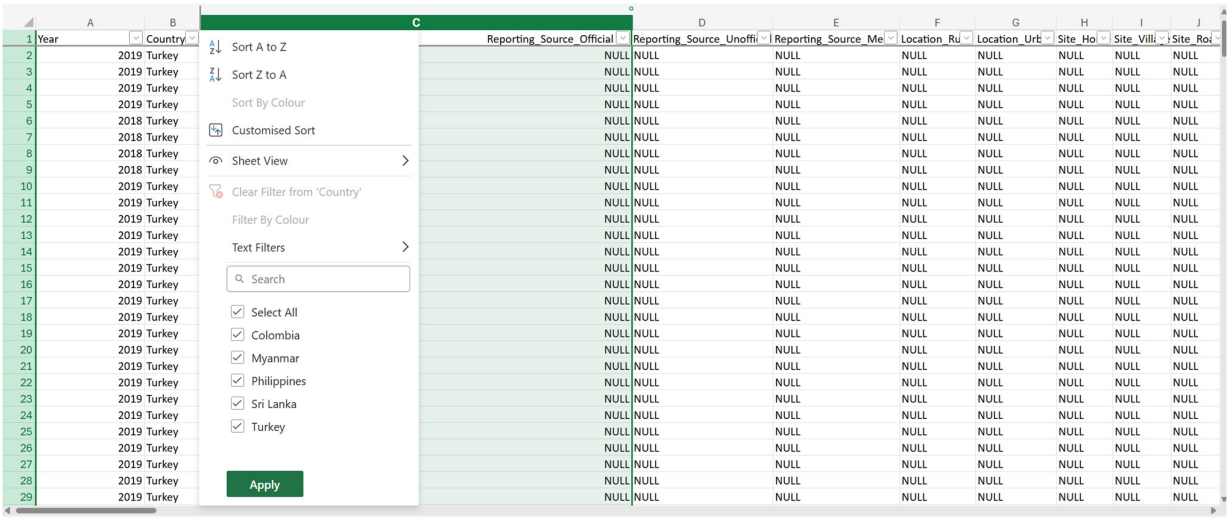
In our initial phase of data collection, the primary challenge was gaining access to firsthand reports of violence, particularly from law enforcement agencies, local governments, and third-sector organizations. As we aimed to capture a global perspective, we encountered fragmented datasets that were incompatible for comparative analysis. Although law enforcement agencies routinely record data on violence against women, these administrative records are primarily intended for internal monitoring rather than research and, thus, fail to capture the true extent of many unreported GBV incidents. For instance, single-country police records provided to us were limited to highly abstract and aggregated forms. Table 1 shows an excerpt from a dataset provided by the Nepalese police, which gives a broad overview of crime types and volumes over a five-year period. However, the lack of granular data presented a significant challenge, reflecting typical access restrictions associated with local and in-country law enforcement data on such a sensitive subject.

Table 1. Dataset from the Nepalese police.

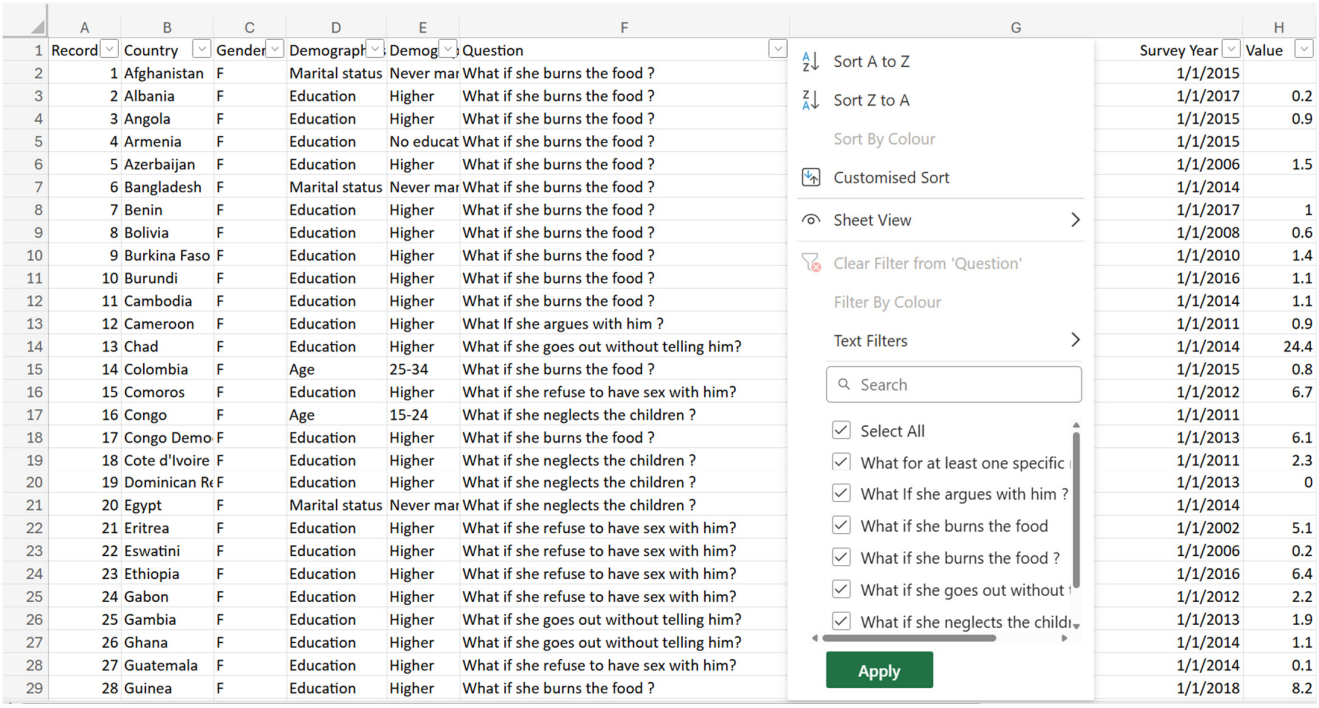
S. No.	Types of Violence	2019–2020	2020–2021	2021–2022	2022–2023	2023–2024	Total
1.	Rape	2230	2144	2532	2380	2387	11,673
2.	Attempt to rape	786	687	735	655	518	3381
3.	Polygamy	1001	734	852	809	723	4119
4.	Child marriage	86	64	84	52	52	338
5.	Accusations of witchcraft	46	34	61	49	43	233
6.	Illegal abortion	27	29	27	37	32	152
7.	Racial untouchability	43	30	39	15	27	154
8.	Unnatural intercourse	24	27	36	31	35	153
9.	Child sexual abuse	211	232	281	314	343	1381
10.	Human trafficking	15	1	10	23	10	59
11.	Abduction	47	34	67	72	59	279
12.	Domestic violence	14,774	11,738	14,232	17,000	16,519	74,263
13.	Acid attack	0	0	6	4	3	13
	Total	19,290	15,754	18,962	21,441	20,751	96,198

To broaden our dataset, we explored various online open-source platforms for gender-based violence data. A recurring issue was the availability of either sparse datasets or those heavily populated with null values. Figure 1a shows a dataset with records from only five countries, while other sources exhibited numerous missing values due to under-reporting or inconsistent data management practices. This lack of comprehensive and reliable data posed a substantial barrier to further analysis. Additional datasets, as shown in Figure 1b, provided more detailed information across several countries, including variables such as gender, marital status, education, and survey year. While these datasets offered valuable insights into factors contributing to domestic violence, they lacked essential information on

the severity of violence, the current domestic environment, and any details on perpetrators, thereby limiting the scope of GBV research.



(**a**)



(**b**)

Figure 1. Examples of limited datasets. (**a**) Reported GBV-related crimes across a limited selection of five countries, with a substantial number of missing values. (**b**) Another dataset covering more countries but lacking sufficient attributes.

We then explored the potential of using synthetic data generation to augment the dataset [11]. Our approach focused on creating synthetic entries that replicated the structure and attributes of the original data, as shown in Figure 2a. This method enabled us to simulate additional country-level data that were absent in the initial dataset. The synthetic values for each attribute were generated based on patterns identified within the original dataset, ensuring that the synthetic data retained realistic characteristics and was consistent with the existing data structure.

	RecordID	Country	Gender	Demographics Question	Demographics Response	Question	Survey Year	Value
0	1	Bangladesh	M	Education	Never married	...if she neglects the children	2010-12-31	4.03
1	2	Bangladesh	M	Education	Higher	...if she neglects the children	2018-12-31	1.82
2	3	Angola	F	Education	Never married	...if she argues with him	2015-12-31	4.33
3	4	Angola	F	Marital status	Never married	...if she argues with him	2010-12-31	3.59
4	5	Benin	F	Education	Higher	...if she burns the food	2019-12-31	3.64

(**a**)

	RecordID	Country	Gender	Demographics Question	Demographics Response	Question	Survey Year	Value	Perpetrator Relationship	Incident Severity
0	1	Cambodia	M	Education	No education	...if she argues with him	2014-12-31	3.01	Father	Moderate
1	2	Angola	M	Marital status	Never married	...if she burns the food	2019-12-31	2.82	Neighbor	High
2	3	Cambodia	M	Marital status	Never married	...if she burns the food	2011-12-31	4.24	Stranger	Critical
3	4	Burundi	F	Education	Higher	...if she argues with him	2010-12-31	3.72	Brother	Critical
4	5	Azerbaijan	M	Education	Higher	...if she argues with him	2013-12-31	4.83	Neighbor	Critical

(**b**)

Figure 2. Potential use of synthetic data. (**a**) Original dataset selected for generating synthetic data; where "…" indicates a placeholder for "what" (**b**) Resulting dataset after applying synthetic data techniques.

We also took the synthetic data generation a step further by introducing two new columns, as highlighted on Figure 2b. We added 'Perpetrator Relationship' and 'Incident Severity' that are derived from the existing attributes on the dataset. The 'Perpetrator Relationship' was derived synthetically from demographic responses and other related attributes, such as marital status, education, and family background, where relationships are categorized based on inferred assumptions rather than directly available data attributes. Meanwhile for the column 'Incident Severity', it was assigned based on the cause of the abuse, the value column, and then categorizing them as low, moderate, high, or critical. However, despite these enhancements, synthetic data still fall short in accurately capturing the complex variability present in real-world datasets. It also lacks the inherent noise and anomalies that are crucial for training robust models. While synthetic data can serve as a useful tool for augmenting datasets, its reliability and integrity remain limited when compared to actual data. For these reasons, we chose not to include synthetic data in our further analysis, as we were concerned about its limitations in reflecting the true patterns and nuances required for meaningful research in the context of GBV.

Next, we expanded our search for a suitable data source using web scraping applying beautiful soup, a library in python programming language for web scraping from online news articles. We used a website called 'News First' to extract information with the date of the crime, the murderer's name, and other details associated with the crime (Figure 3a) [12]. We explored another approach by web scraping data from Twitter. Specifically, we targeted tweets containing the hashtags #DomesticAbuse and #DomesticViolence to collect relevant data. We also considered including tweets with #MeToo but found that while it captured a broad spectrum of the gender-based violence, it lacked specific details about the victims and had more generalised information on various form of violence. Therefore, to maintain our focus and ensure data consistency, we narrowed down our scope to solely concentrate on domestic violence for thorough and detailed analysis (Figure 3b).

Name	Date	Murder Motive	Murderer Name	Protection Request	Way Of Killing	News Source 1	News Source 2	Age of Victim	Province	Perpetrator Status	Notes
TuÃŸÅ§e Baran	29/09/2019	Not Determined	Boyfriend	No	Firearm	http://www.milliyet.com.tr/gundem/evinde-basindan-vurulan-tugce-hayatini-kaybetti-6042935	NA	of age	Izmir	Prisoner	suspicious Death
Ebru Erdem	20/03/2019	Not Determined	Not Determined	Not Determined	Falling from high	http://www.milliyet.com.tr/istanbul-da-residansta-dehset-kan-gundem-galeri-2845198/	NA	of age	Ä°stanbul	Investigation Continues	suspicious Death
SongÃ¼l Ã–nemli	26/10/2019	Not Determined	Not Determined	No	Firearm	http://www.milliyet.com.tr/galeri/isir-dolu-olum-iki-genc-kadin-boyle-bulundu-6065264	NA	of age	Adiyaman	Not Determined	suspicious Death
Cansu GÃ¼ven	26/10/2019	Not Determined	Not Determined	No	Firearm	http://www.milliyet.com.tr/galeri/isir-dolu-olum-iki-genc-kadin-boyle-bulundu-6065264	NA	of age	Adiyaman	Not Determined	suspicious Death
Maiko Dzidziguri	23.02.2018	Unspecified	Unknown	No	drowning	http://www.milliyet.com.tr/gurcu-kadinin-esrarengiz-olumu--gundem-2615313/	NA	of age	Ä°stanbul	Unknown	suspicious Death
Alara Karademir	29/03/2018	Discussion	Someone familiar	No	Darpa	http://m.ilerihaber.org/icerik/yogun-bakimdaki-alara-karademir-hayatini-kaybetti-83550.html	NA	of age	Ankara	Investigation Continues	suspicious Death
Ganime Varsak	17/02/2018	Unknown	Unknown	No	burned	http://www.milliyet.com.tr/yanmis-cesedi-bulunan-kadinin-gundem-2611889/	NA	of age	KÃ±rÃ±kkale	Investigation	suspicious Death
Melahat Mersin	16/06/2018	Comprise from not detected	Somebody knows	No	Not Determined	https://www.yenisafak.com.tr/yasam/2018/06/17/insallah-annem-degildir	NA	of age	Izmir	Prisoner	suspicious Death
Bahar Akdemir	21/12/2019	Not Determined	Not Determined	No	Not Determined	https://www.cnnturk.com/turkiye/evinin-onunde-genc-kizin-cesedi-bulundu	NA	of age	Diyarbakir	Investigation Continues	suspicious Death
Derya TavÃ¼an	10/12/2019	Not Determined	Not Determined	No	Not Determined	http://www.nurnyet.com.tr/gundem/otel-odasinda-supheli-olum-413944206	NA	of age	Adana	Not Determined	suspicious Death
GÃ¼lperi Onur	31/10/2019	Not Determined	Boyfriend	No	Falling from high	https://www.sozcu.com.tr/2019/gundem/balkondan-duserek-olen-kadinin-sevgilisi-tutuklandi-5431437/	NA	of age	Antalya	Prisoner	suspicious Death
Adile GÃ¼ler	2/10/2019	Not Determined	Not Determined	No	Falling from high	http://www.milliyet.com.tr/gundem/konyada-supheli-olum-genc-kadin-3-kattan-dusup-oldu-6044972	NA	of age	Konya	Not Determined	suspicious Death
Alexandra KÃ¶ppel	15/10/2019	Not Determined	Not Determined	No	Suffocation	https://www.sozcu.com.tr/2019/gundem/iseci-kadinin-istanbulda-sir-olumu-kimligi-belli-oldu-5390892/	NA	of age	Ä°stanbul	Detection of	suspicious Death
Hatice KÃ¶se	23/09/2019	Not Determined	Not Determined	No	Not Determined	http://www.bursahayat.com.tr/haber/bursa-da-kiyiya-vuran-kadin-cesedinin-kimligi-aciklandi-232489.html	NA	of age	Bursa	Investigation Continues	suspicious Death
BirgÃ¼l K.	24/09/2019	Not Determined	Not Determined	No	Firearm	http://www.haberturk.com/ardahan-haberleri/71336413-genc-kadin-ahirda-olu-bulundu-esi-gozaltinda	NA	of age	Ardahan	Investigation Continues	suspicious Death
Ã‡afak Bozkurt	5/11/2019	Not Determined	Not Determined	No	Medicine	http://www.diyarbakirsoz.com/gundem/2-351supheli-olum-200877	NA	of age	Ergani	Not Determined	suspicious Death
Suriye SoydaÃŸ	6/11/2019	Not Determined	Not Determined	No	Not Determined	http://www.milliyet.com.tr/gundem/yuruyus-yapanlar-fark-etti-suyun-uzerinde-hareketsiz-yatan-6073547	NA	of age	Lapsike	Not Determined	suspicious Death
Melisa K.	6/11/2019	Not Determined	Not Determined	No	Firearm	https://gazetekarinca.com/2019/11/bilecikte-supheli-kadin-olumu-universite-ogrencisi-yasamina-son-verdi/	NA	of age	Bilecik	Not Determined	suspicious Death
Derya H.	4/6/2019	Not Determined	Not Determined	No	Not Determined	https://www.iha.com.tr/haber-gaziantepte-kadin-cesedi-bulundu-783459/	NA	of age	Gaziantep	Investigation Continues	suspicious Death
N.S.	16/06/2019	Not Determined	Not Determined	No	Firearm	https://www.haberler.com/sakarya-da-kazara-vurulugu-iddia-edilen-kadin-12150755-haberi/	NA	of age	Sakarya	Investigation Continues	suspicious Death

(**a**)

```
[ ] data
```

	text	date
0	Deaths caused by domestic abuse are so common...	Jan 14, 2024 · 2:08 PM UTC
1	#laurenboebert (notice how #X refuses to give ...	Jan 13, 2024 · 7:35 PM UTC
2	A year in review: GMP's response to tackling d...	Jan 13, 2024 · 11:00 AM UTC
3	@NestacUK speaking about hidden complex #traum...	Jan 9, 2024 · 1:16 PM UTC
4	Thoughts ??? #men #women #abuse #domestic viol...	Jan 6, 2024 · 8:55 AM UTC
...	...	...
3426	Thank you to @coastlandsnews for inviting Spe...	Mar 6, 2019 · 10:44 AM UTC
3427	#Domestic Violence and Abuse: Recognizing the ...	Mar 6, 2019 · 12:11 AM UTC
3428	#Domestic #abuse victims school #priority #edu...	Mar 5, 2019 · 10:01 PM UTC
3429	#DOMESTIC ABUSE	Mar 5, 2019 · 1:02 PM UTC
3430	We welcome inclusion of #economicabuse in the ...	Mar 5, 2019 · 11:30 AM UTC

3431 rows × 2 columns

(**b**)

Figure 3. Data derived from web scraping of media website (**a**) and Twitter (**b**), with tweets truncated and indicated by "…" to show continuation beyond displayed text.

We further enhanced our synthetic data generation by introducing two new columns, as illustrated in Figure 2b. Despite enhancements, synthetic data still fall short in reflecting the complex variability and inherent noise of real-world datasets, which are essential for building robust models. As we assessed Twitter data by collecting a sample of tweets using hashtags like #DomesticAbuse, #DomesticViolence, and #MeToo, while reviewing for patterns and relevance, we found most tweets lacked structure, verifiable details, and key metadata, making them unreliable for analysis. Due to these limitations, neither synthetic nor Twitter data were used in the final analysis.

To further expand our analysis on selecting datasets, we applied web scraping techniques using Python libraries, to extract information from online news sources. Specifically, we used the 'News First' website to gather details such as the date of the crime, the perpetrator's name, and other relevant information (Figure 3a). While experimenting on web scraping data from Twitter, we focused on tweets tagged with #DomesticAbuse and #DomesticViolence to collect pertinent data. Although we initially considered including #MeToo tweets to broaden the dataset, we found that this hashtag encompassed a wide range of gender-based violence topics without sufficient specificity about victims or details relevant to domestic violence. Therefore, to maintain data consistency and a focused scope,

we restricted our analysis to tweets specifically addressing domestic violence, as shown in Figure 3b.

To gain deeper insights from the data, we selectively extracted essential crime-related information into structured tables, with Table 2a derived from the web scraping of online media sources and Table 2b from Twitter data extraction. We organized and formatted these datasets to improve clarity, providing a structured view of data from both news articles and tweets. Through text mining, we filtered and extracted only the most relevant information, focusing on details such as types of abuse and victim–perpetrator relationships.

Table 2. Information extracted after web scraping: news website (a), Twitter (b).

(a)

Name	Age	Abuse Inflicted	Relation to Victim	Lead to Death?	Country	Year
null	61	null	husband	Y	Sri Lanka	2020
Shyamila Swapana	19	set on fire	husband	Y	Sri Lanka	2017
null	39	raped, verbally abused	husband	N	Sri Lanka	started in 2005, fled to NZ in 2017
null	29	killed using pole	husband	Y	Sri Lanka	2021

(b)

Name of Victim	Country	Type	Relation to Victim	Lead to Death	Date of Incident	info
null	Ireland	whipped	partner	N	null	null
null	United States	killed with hammer	husband	Y	null	null
Deborah Brandao	United States	stabbed	ex-boyfriend	Y	18 April 2021	null
null	Australia	house fire	son	Y	7 January 2024	null
Mandeep Kaur	United States	null	null	null	null	null

Despite these efforts, the dataset remained limited, particularly in capturing a broader range of case details. Additionally, the volume of extracted data was relatively low, as most articles from the news portal concentrated on extreme cases, likely selected for their newsworthiness, and tended to emphasize general awareness content. This focus restricted the data's relevance and limited its analytical potential.

The Demographic and Health Surveys (DHS) dataset was ultimately selected for our analysis after extensive evaluation of various data sources. Unlike administrative crime statistics or fragmented open-source datasets, DHS offered comprehensive, nationally representative data collected through standardized and internationally recognized survey methodologies. It provided a wide range of relevant variables, including detailed demographic information, socioeconomic indicators, and specific questionnaires that helped in recognizing patterns of perpetrators, allowing for a detailed analysis of gender-based violence patterns. Importantly, DHS ensured data consistency across multiple countries, facilitating comparative studies, while maintaining high data reliability and validity. While other sources such as police records, web-scraped news articles, and social media data contributed to preliminary exploration and contextual understanding, they lacked the depth, structure, and methodological rigor required for robust quantitative analysis as demonstrated in Table 3. Therefore, the DHS dataset was chosen as the most suitable foundation for our research, ensuring that our findings would be based on reliable, standardized, and analytically rich data.

Table 3. Comparison of explored datasets and justification for selected data.

Data Source	Data Type	Challenges/Limitations	Decision/Action
Nepalese police records	Administrative crime statistics	Highly aggregated data; lack of granular victim/perpetrator details; limited access to sensitive cases	Used for preliminary exploration; insufficient for detailed GBV research
Open-source online datasets	Survey data from few countries	Sparse data; heavy missing values; inconsistent formats	Explored but found insufficient for reliable cross-country analysis
Synthetic data (generated)	Simulated data entries	Lacked real-world variability, noise, and authenticity crucial for modelling GBV patterns	Not used for final analysis due to limitations in validity
Web scraping—news websites	Crime reports from articles	Focused on extreme/high-profile cases; limited data volume; biased toward newsworthy events	Supplemented understanding; not used as primary dataset
Web scraping—Twitter data	Social media posts (#DomesticAbuse, #DomesticViolence, #MeToo)	Incomplete metadata; lack of verification; inconsistent and generalized information	Supplemented understanding; not suitable for further analysis
Demographic and Health Surveys (DHS)	Large-scale, standardized surveys	Comprehensive, structured, multi-country data on domestic violence with demographic variables	Selected for primary analysis due to high-quality, consistency across regions, and relevance to GBV research

3.2. Tackling Dataset Challenges

After careful consideration, we ultimately chose to use the Demographic and Health Surveys (DHS) dataset provided by USAID. Working with this dataset presented several constraints, including inconsistencies, gaps in key variables, and limitations in regional and demographic specificity. Additionally, the dataset's large volume and abundance of attributes complicated the selection of relevant columns for analysis. The DHS data are collected in multiple phases, encompassing over 170 variables across numerous columns, making it extensive in scope.

Our focus centred on selecting data from five primary regions, each represented by multiple countries. However, the phases of data collection varied considerably by country and survey year, complicating our ability to establish consistent comparisons. To address this issue, we limited our analysis to recent data from phases 7 and 8, covering surveys conducted between 2015 and 2022 as shown in Table 4.

Accessing the DHS data via USAID was relatively straightforward, though our access was limited to datasets from a select number of countries. We received data from 19 countries along with supporting documentation. Throughout the project, we adhered rigorously to USAID's data privacy and sharing guidelines. No data, whether in raw or processed form, were distributed externally or in partial subsets. All processing was conducted within a secure environment, using verified credentials to ensure full compliance with institutional security standards.

To comply with DHS guidelines and address concerns related to anonymity, privacy, and ethical considerations, we implemented statistical disclosure controls. This involved aggregating data and grouping information into broader categories to safeguard individual identities. Additionally, we ensured that all reports and visualizations were based solely on specified metrics, further reinforcing data protection measures.

Table 4. Final DHS selected datasets based on country and phase.

Region	Country	DHS Phase	Year Selected	Years Available
Sub-Saharan Africa	Cameroon	7	2018	1991, 1998, 2004, 2011, 2018
	Ethiopia	7	2016	2000, 2005, 2011, 2016
	Liberia	7	2019–2020	1986, 2007, 2013, 2019–2020
	Nigeria	7	2018	1990, 2003, 2008, 2013, 2018
	Ghana	8	2022	1988, 1993, 1998, 2003, 2008, 2014, 2022
	Kenya	8	2022	1989, 1993, 1998, 2003, 2008–2009, 2014, 2022
	Tanzania	8	2022	1991–1992, 1996, 1999, 2004–2005, 2010, 2015–2016, 2022
Latin America and Caribbean	Colombia	7	2015	1986, 1990, 1995, 2000, 2005, 2010, 2015
	Guatemala	7	2014–2015	1987, 1995, 2014–2015
North Africa, West Asia, Europe	Albania	7	2017–2018	2008–2009, 2017–2018
	Turkey	7	2018	1993, 1998, 2003, 2008, 2013, 2018
	Armenia	7	2015–2016	2015, 2016
	Jordan	7	2017–2018	2017, 2018
Central Asia	Tajikistan	7	2017	2012, 2017
South and Southeast Asia	Afghanistan	7	2015	2015
	Bangladesh	7	2017–2018	1993–1994, 1996–1997, 1999–2000, 2004, 2007, 2011, 2014, 2017–2018
	India	7	2019–2021	1992–1993, 1998–1999, 2005–2006, 2015–2016, 2019–2021
	Myanmar	7	2015–16	2015–16
	Pakistan	7	2017–18	1990–91, 2006–07, 2012–13, 2017–18
	Philippines	7	2017	2017
	Nepal	7	2015	2015

During the initial stages of data extraction and processing, the STATA (.dta) format proved incompatible with our planned analysis tools, including Python, Power BI, and certain cloud services. One particularly large dataset, containing over 724,000 entries, presented significant challenges due to memory limitations and repeated system crashes during processing. Despite experimenting with various tools and cloud solutions, including options on Azure, the issues persisted. Ultimately, IBM SPSS emerged as the most effective

tool for managing and extracting this large dataset. To further streamline our workflow, we converted the survey data from STATA to .csv format, ensuring compatibility with our analysis tools and facilitating a smoother analysis process.

The initial dataset encompassed an extensive range of variables and respondents, with 5177 variables collected from 973,337 individuals across 19 countries. This comprehensive dataset offered a robust foundation for the study, covering diverse demographic characteristics and experiences. However, a notable challenge was the uneven distribution of sample sizes among countries. For example, India contributed a significantly large number of respondents (724,115); whereas, Ethiopia had only 3992 respondents, the smallest sample size. To mitigate this imbalance, we applied stratified sampling by country, reducing the sample size to 96,422. This approach ensured balanced representation across countries, while preserving the data's integrity.

A further challenge lay in managing the large number of variables within the dataset. After removing columns with null values, approximately 150 columns remained. To streamline our analysis, we first organized these columns into five primary categories. Within each module, we used color-coding to assist in the selection process, enabling us to efficiently identify and prioritize relevant attributes. This approach allowed us to reduce the dataset to 64 key columns for detailed analysis. Figure 4 presents the various colour-coded categories.

Categories	Colour code	Column selection and filtering strategy
Respondent's basic data		Essential columns for analysis
Reproduction		Selected/considered columns as an option
Marriage		Potential columns (not currently selected but can be considered later)
Partner's Characteristics & Respondent's Work		Unselected columns
Domestic violence		Collection of columns to be merged

Figure 4. Column selection and filtering strategy.

Since the USAID DHS surveys are conducted primarily in low- and middle-income countries (LMICs) undergoing demographic transitions, the dataset lacked representation from EU countries. The EU nations typically conduct their own surveys tailored to their unique health and demographic needs, creating a gap in data comparability within the DHS framework. To address this, we sought additional data from the UK Data Service. However, this presented two major challenges: first, the data available only extended up to 2012, limiting access to more recent records; second, the structure of the UK Data Service dataset differed significantly from that of DHS. While the DHS data focuses primarily on domestic violence and intimate partner violence (IPV), the UK Data Service encompasses a broader range of gender-based violence contexts. Due to these disparities, we ultimately decided not to integrate the EU data [13].

3.3. Environment Challenges

Selecting the right tools for data collection, analysis, and interpretation is essential in addressing the complexities of gender-based violence research. Using inappropriate or suboptimal tools provides misleading results, causing misinterpretation and an incomplete understanding of the issue. To ensure robust methods that accurately capture the nuances of the data, we strategically selected tools that were best aligned with our research objectives and analytical needs.

For the analysis, we employed Python within Jupyter Notebook, leveraging its flexibility and powerful libraries. This approach enabled us to import, clean, and convert data into a more manageable format, creating an environment for efficient exploration. Python's broad ecosystem of frameworks allowed for the identification of complex patterns, the execution of detailed analysis, and the application of advanced models, such as network graphs, to uncover deeper insights into the data. Despite Python's capabilities, one of the datasets we were working with, which contained over 720,000 rows, presented significant constraints, leading to system crashes and errors due to memory limitations. To resolve this, we explored alternative tools and found that IBM SPSS provided an effective solution for handling large datasets in STATA format, ensuring the smoother processing and management of the data.

Additionally, we turned to cloud-based solutions to enhance our data processing infrastructure. On Azure Cloud platform, we worked on the Azure Data Factory, facilitating the automation of data workflows, integration, and management. Azure Databricks supported advanced analytics for large-scale data processing, while Azure Blob Storage provided the necessary storage for handling large datasets. However, maintaining these services over time proved to be costly, which ultimately limited the duration and scope of our analysis, highlighting the obstacles associated with scaling up computational resources for large-scale research projects.

3.4. Detecting and Addressing Data Inconsistencies Through Power BI

For data visualization, we utilized Power BI, which proved to be an essential, efficient, and user-friendly tool for quickly extracting meaningful insights from the gender-based violence data. Power BI allowed us to present complex data through clear visuals by laying the groundwork for deeper analysis. Additionally, our decision to use Power BI was influenced by its ability to create tailored dashboards, which were designed to effectively communicate findings to policymakers and stakeholders focused on addressing and preventing gender-based violence. Figure 5 is a representative dashboard developed to provide the key GBV influential factors at a country-wide level.

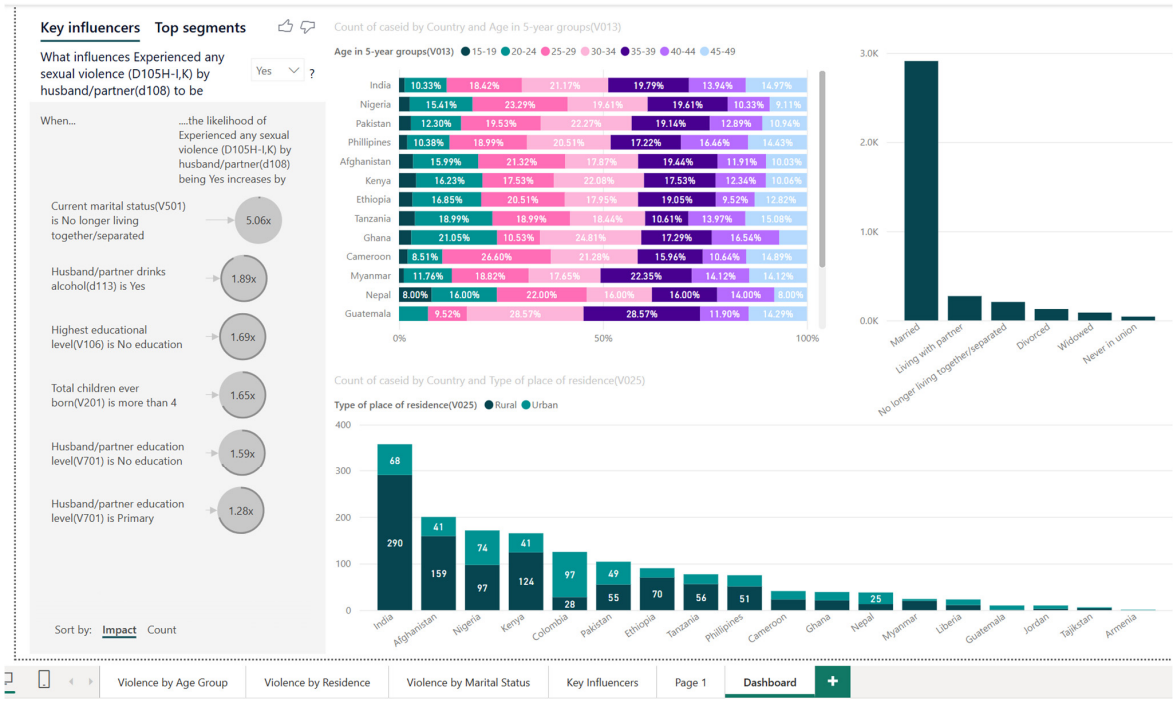
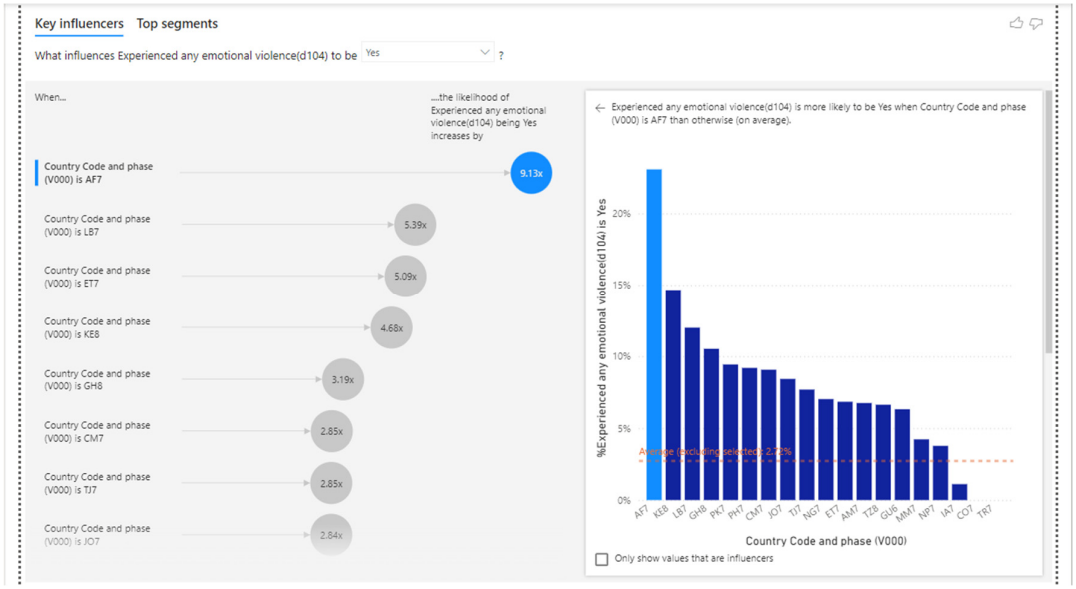
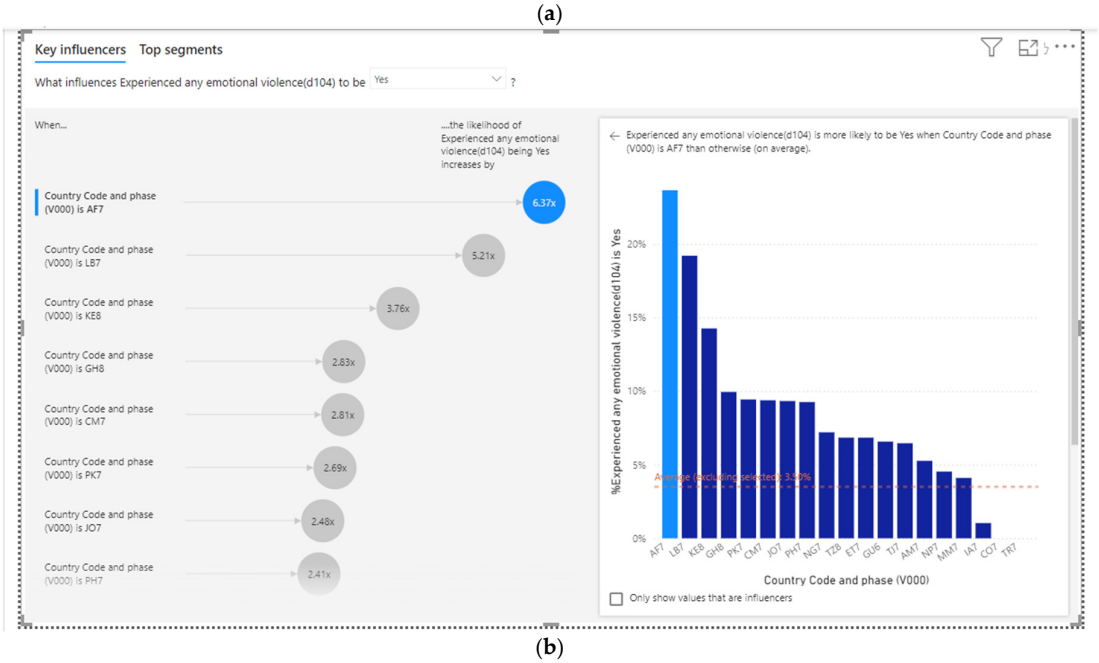


Figure 5. Power BI dashboard depicting GBV key influencers.

A particular anomaly we encountered was the uneven representation of populations of certain countries, which, if left unchecked, could lead to potential biases in the data. To address this, we implemented stratified sampling to ensure that key subgroups were adequately represented. Using Power BI, we identified the factors contributing to this bias and used the insights to guide our adjustments. Through this approach, we reduced the bias in the sample and also enhanced the overall generalizability and representativeness of the data across different countries. Figure 6a,b shows two examples of how a particular country's (Afghanistan) level of emotional violence can be calibrated in the context of all 19 countries.



(**a**)



(**b**)

Figure 6. Understanding impact of bias and post-adjustments on the data. (**a**) Before bias adjustment. (**b**) After bias adjustment.

3.5. Visualization Through Network Graphs

Network graphs provide a visual breakdown of the interconnections between various factors in the data, particularly focusing on elements related to domestic violence against women [14]. Given the broad scope of domestic violence categories, we concentrated

specifically on IPV (intimate partner violence). In the graph (Figure 7), the central node represents IPV that connects with various factors, such as attitude, history, and demographic details, with lines indicating how they are inter-related. The colours of the nodes and edges suggest different categories, such as demographic factors, a history of previous abuse, and attitudes toward violence, guiding to an understanding of the multiple influences on IPV.

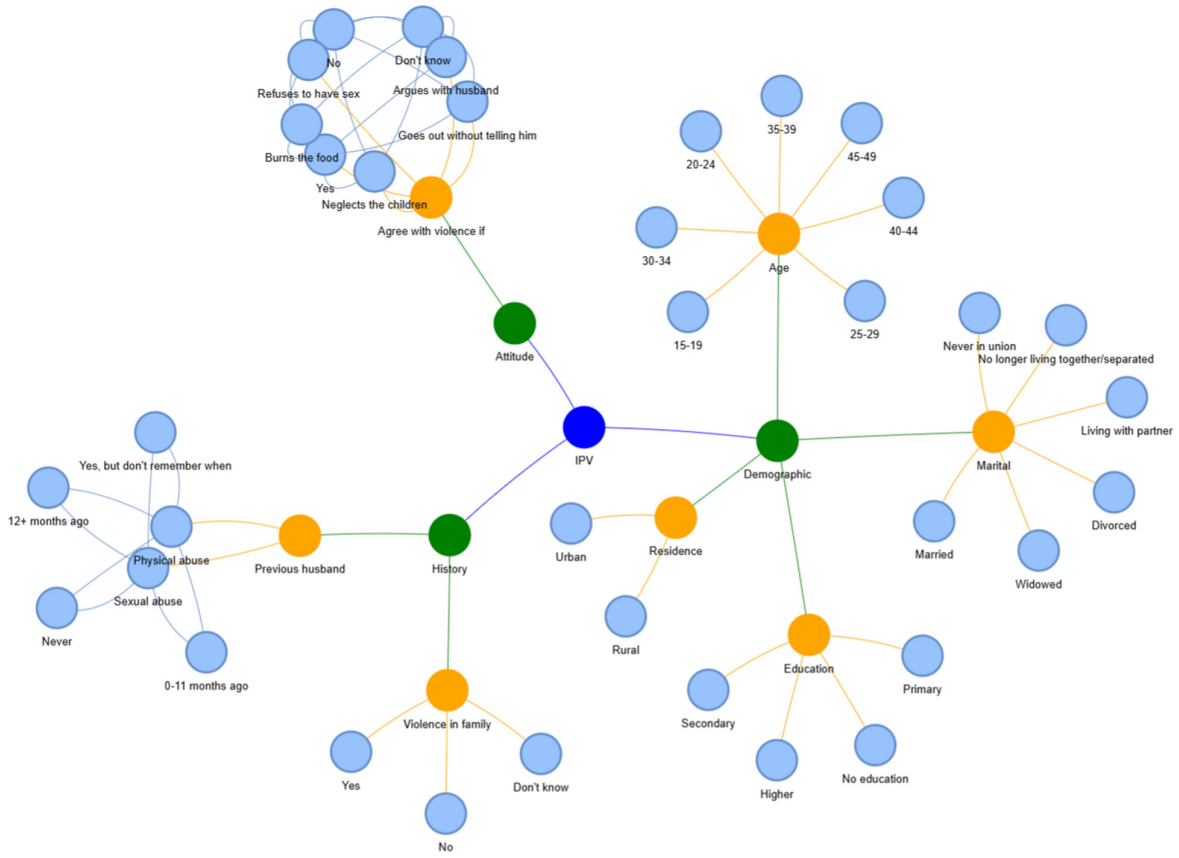


Figure 7. Network graph for intimate partner violence.

Network graphs are significantly relevant in the context of IPV, as they visually demonstrate the interconnectedness of various factors, such as a history of violence, demographics, and attitudes towards violence. For instance, they show how attitudes towards violence can link directly to an individual's likelihood to experience or perpetrate IPV, which is crucial for understanding the root causes of gender-based violence. Additionally, demographic factors, such as marital status, i.e., whether someone is married, divorced, or living with a partner, and their education level, i.e., primary, secondary, or higher education, are often key predictors of vulnerability to IPV. This approach helped us to identify patterns and relationships that may not be immediately apparent, thus informing targeted interventions, policies, and support systems for those affected by IPV.

Moreover, by visualizing the data in this way, it becomes easier to identify at-risk populations, such as individuals with certain educational backgrounds or marital statuses, and understand how these elements interact with other factors, such as a history of violence or urban/rural residence, offering a comprehensive view of the dynamics across different populations. Although the network graph provides valuable insights, its complexity can hinder usability, as the intricate relationships may lead to the misinterpretation of outcomes. We adopted simpler visualizations that were more effective in highlighting the key characteristics of both victims and perpetrators. However, these visualizations should be approached with caution, as the complexity of multiple nodes and interconnections

can make interpretation challenging and increase the potential for misrepresenting the relationships depicted.

4. Discussion

Large-scale GBV data collected through national surveys remains one of the most effective methods for capturing the experiences and perceptions of victims and, to a lesser extent, understanding the behaviours and characteristics of perpetrators. Systematic demographic surveys, such as the DHS, conducted over multiple years, enable longitudinal analysis and insights. However, this approach presents certain limitations; it is often costly and requires trained personnel to conduct in-person interviews and private discussions with respondents. These challenges may be alleviated by leveraging advances in technology, such as large language models (LLMs), which could automate elements of the survey process, potentially reducing costs and enhancing data collection efficiency.

The established protocols for collecting data on gender-based violence have become relatively well-defined; however, significant constraints persist in obtaining accurate data, particularly in low-resource and complex humanitarian contexts, such as those involving conflict, war, and asylum. In such environments, the collection of reliable GBV data is constrained by limited resources and logistical obstacles. Contemporary approaches increasingly rely on digital technologies, introducing new ethical and procedural considerations. These considerations encompass issues related to data ownership, the requirement for fully informed and ongoing consent, and the secure storage and use of sensitive data [15].

Remote data collection has been adopted as a strategy to enhance equity and inclusivity for marginalized populations in research [16]. However, this method presents its own set of ethical dilemmas, particularly in relation to informed consent and the provision of referral services for participants. Additionally, reference [17] examined studies that explore the use of digital technologies—such as mobile phone cameras, mobile applications, social media, web platforms, and videos—to facilitate self-reported data collection by women. While these technologies offer opportunities for greater autonomy in reporting, they also necessitate careful attention to issues of privacy, consent, and data security, especially in sensitive contexts. The under-reporting of GBV incidents remains a persistent challenge, particularly in relation to under-researched areas that capture diverse women's experiences of abuse, including those of older women (50+), women with disabilities, and migrant and Indigenous women. The authors of [18] analyse several online reporting and documentation platforms that use open-ended questions, allowing respondents to narrate their experiences in their own words. The discrepancy between actual prevalence rates and disclosed or reported cases—often referred to as the "grey zone"—may be significantly underestimated.

Recent advances in big data and machine learning have further extended their application to the analysis and prediction of GBV behaviours. The authors of ref. [19] focus on the use of machine learning to study instances of violence as reported in news media, while refs. [20–22] investigate the forecasting and predicting of such behaviours. Large language models (LLMs) also hold the potential to transform survey methodologies. Traditional survey methods, such as those employed by the Demographic and Health Surveys (DHS), are often resource-intensive and require trained personnel for administration. The integration of LLM interfaces in data collection could enable the capture of rich, qualitative responses to open-ended questions, which may be particularly beneficial given the sensitive and emotional nature of GBV. However, while LLMs may improve the efficiency of data collection, ensuring the elimination of biases and the accuracy of findings remains crucial. Notably, some researchers have explored the creation of synthetic survey responses [23], suggesting that future developments could include the design of artificial human personas and their corresponding responses to survey questions.

Challenges surrounding data integration and sharing continue to hinder the comparability of GBV data. For instance, the approach to counting multiple offences—whether all incidents should be recorded or only the most severe—remains contentious. Data disaggregation is critical to enhancing data quality, yet sex-disaggregated data are often lacking, as is detailed information on victims and perpetrators, particularly in police and justice datasets. The decentralization of GBV data collection, coordination, and compilation further complicates comparability, as does the absence of a standardized coding system across agencies for registering such data. These limitations impede efforts toward the harmonization of GBV data across many contexts.

5. Conclusions

In working with gender-based violence data, challenges in data collection, analysis, and sharing represent significant barriers to providing effective support for victims. Our study has highlighted key issues researchers may encounter when dealing with GBV datasets, offering insights to assist others undertaking similar research. One of the primary obstacles is the limited value of law enforcement records, particularly in low- and middle-income countries, where reporting mechanisms may be inconsistent or influenced by social and structural constraints. Media reports, too, present limitations; journalistic boundaries often mean that only the most severe or publicly notable cases are reported, leading to a skewed perspective on the prevalence and types of GBV incidents.

The tools and environments used for data analysis also significantly shape the quality and depth of insights generated. Access to appropriate software and secure, compliant storage solutions is essential to uphold data integrity and ensure that findings are robust. Moreover, the quality of data available for analysis is often inconsistent, with significant gaps in areas such as victim demographics, the nature of the violence, and perpetrator characteristics. These gaps highlight the need for consistent and systematic approaches to demographic population surveys, which, with a heightened focus on data quality, could yield better outcomes for victims.

Emerging technologies, particularly generative AI and machine learning, have the potential to play a transformative role in addressing these barriers. By enabling the assimilation of sensitive and diverse information sources, AI can help researchers overcome some data limitations. These technologies can assist in identifying patterns within complex datasets, predicting trends, and even creating synthetic data to augment sparse areas without compromising the confidentiality of real individuals. However, the use of these tools must be carefully managed to ensure ethical standards are maintained, especially in dealing with sensitive information.

While the current study focused on evaluating existing datasets, we acknowledge the importance of developing shared key indicators and standardized data registration strategies for more consistent and comparable GBV research. We plan to address this in future work by proposing a set of common indicators and data collection practices to strengthen the consistency and quality of GBV data across diverse contexts.

Ultimately, ongoing investment in advanced data collection and analysis methodologies, coupled with technological innovation, is essential to improve the quality and usability of GBV data, ensuring that it can effectively inform interventions and support strategies for those affected by GBV.

Author Contributions: Conceptualization: P.P.; methodology: P.P. and S.S.; software: S.L., A.F.X. and S.S.; validation: S.S. and P.P.; data curation: S.L., A.F.X. and S.S.; writing—original draft preparation: S.S. and P.P.; review and editing: all authors; supervision: P.P. All authors have read and agreed to the published version of the manuscript.

Funding: No external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data contained within the study are available from DHS.

Acknowledgments: The authors thank USAID for the DHS data provision.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

GBV	Gender-Based Violence
DHS	Demographic and Health Surveys
IPV	Intimate Partner Violence
USAID	United States Agency for International Development
EIGE	European Institute for Gender Equality
FRA	Fundamental Rights Agency
ONS	Office of National Statistics
CSEW	Crime Survey for England and Wales
LMICs	Low- and Middle-Income Countries

References

1. WHO. Violence Against Women. 25 March 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/violence-against-women#:~:text=Estimates%20published%20by%20WHO%20indicate,violence%20is%20intimate%20partner%20violence (accessed on 15 April 2024).
2. Eurostat. Every Woman Deserves to be Safe. But One in Three Women Still Experience Violence in the EU. 2024. Available online: https://eige.europa.eu/ (accessed on 25 November 2024).
3. Gov.uk. *Tackling Violence Against Women and Girls Strategy*; Home Office: London, UK, 2021.
4. USAID. In *United States Agency for International Development*; 2024. Available online: https://www.usaid.gov/ (accessed on 15 November 2024).
5. Walby, S.; Towers, J. Measuring violence to end violence: Mainstreaming gender. *J. Gender-Based Violence* **2017**, *1*, 11–31. [CrossRef]
6. Peterman, A.; Devries, K.; Guedes, A.; Chandan, J.S.; Minhas, S.; Lim, R.Q.H.; Gennari, F.; Bhatia, A. Ethical reporting of research on violence against women and children: A review of current practice and recommendations for future guidelines. *BMJ Glob. Health* **2023**, *8*, e011882. [CrossRef] [PubMed]
7. Beyene, A.S.; Chojenta, C.; Roba, H.S.; Melka, A.S.; Loxton, D. Gender-based violence among female youths in educational institutions of Sub-Saharan Africa: A systematic review and meta-analysis. *Syst. Rev.* **2019**, *8*, 1–14. [CrossRef] [PubMed]
8. Ostadtaghizadeh, A.; Zarei, M.; Saniee, N.; Rasouli, M.A. Gender-based violence against women during the COVID-19 pandemic: Recommendations for future. *BMC Women's Health* **2023**, *23*, 219. [CrossRef] [PubMed]
9. Division for the advancement of Women. *Violence Against Women: A Statistical Overview, Challenges and Gaps in Data*; Economic Commission for Europe and World Health Organization: Geneva, Switzerland, 2005.
10. Fraga, S. Methodological and ethical challenges in violence research. *Porto Biomed. J.* **2016**, *1*, 77–80. [CrossRef] [PubMed]
11. Patel, P. Synthetic data. *Bus. Inf. Rev.* **2024**, *41*, 48–52. [CrossRef]
12. NewsFirst. NewsFirst. 2024. Available online: https://english.newsfirst.lk/ (accessed on 12 November 2024).
13. UK Data Service. UK Data Service. 2024. Available online: https://ukdataservice.ac.uk/ (accessed on 29 October 2024).
14. Dgraph Labs, Inc. Graphs and Networks for Beginners. 2024. Available online: https://dgraph.io/blog/post/graphs-and-networks/ (accessed on 22 October 2024).
15. O'Brein, S.R.; Piay-Fernandez, N. *Global Symposium on Technology-Facilitated Gender-Based Violence Results: Building a Common Pathway*; Wilson Center: Washington, DC, USA, 2022.
16. Vahedi, L.; Qushua, N.; Seff, I.; Doering, M.; Stoll, C.; Bartels, S.A.; Stark, L. Methodological and Ethical Implications of Using Remote Data Collection Tools to Measure Sexual and Reproductive Health and Gender-Based Violence Outcomes among Women and Girls in Humanitarian and Fragile Settings: A Mixed Methods Systematic Review of Peer-Reviewed Research. *Natl. Libr. Med. Trauma Violence Abus.* **2023**, *24*, 2498–2529. [CrossRef]

17. Iyawa, G.E.; Akinmoyeje, B.A.; Kays, R.; Mutelo, S.; Ipinge, R. Women as Citizen Scientists: Identifying Potential Digital Solutions for Addressing Gender-Based Challenges. In Proceedings of the IST-Africa Conference, Virtual, South Africa, 10–14 May 2021; IEEE: Piscataway, NJ, USA, 2021.
18. Stevens, L.M.; Bennett, T.C.; Cotton, J.; Rockowitz, S.; Flowe, H.D. A critical analysis of gender-based violence reporting and evidence building applications (GBVxTech) for capturing memory reports. *Front. Psychol.* **2024**, *14*, 1289817. [CrossRef] [PubMed]
19. Bello, H.J.; Palomar-Ciria, N.; Gallego, E.; Navascués, L.J.; Lozano, C. Big Data Techniques to Study the Impact of Gender-Based Violence in the Spanish News Media. *Central Eur. J. Commun.* **2023**, *16*, 101–116. [CrossRef] [PubMed]
20. Rodríguez-Rodríguez, I.; Rodríguez, J.-V.; Pardo-Quiles, D.-J.; Heras-González, P.; Chatzigiannakis, I. Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 8244. [CrossRef]
21. Rahman, R.; Alam Khan, N.; Sara, S.S.; Rahman, A.; Khan, Z.I. A comparative study of machine learning algorithms for predicting domestic violence vulnerability in Liberian women. *BMC Women's Health* **2023**, *23*, 542. [CrossRef] [PubMed]
22. Shifidi, P.P.; Stanley, C.; Azeta, A.A. Machine Learning-Based Analytical Process for Predicting the Occurrence of Gender-Based Violence. In Proceedings of the 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 16–18 August 2023; pp. 1–8.
23. Hämäläinen, P.; Tavast, M.; Kunnari, A. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In Proceedings of the CHI '23: CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–19.