9/19/2023

Visualisation for Data Analytics (CSC-40048)

Report on Case Study Analysis

Idongesit Chelly Okoko, Shibina Shajahan, Constance Jumbo
UNIVERSITY OF KEELE

Table of Contents

1. Part A

1.1 Renewable energy consumption and economic growth in OECD countries: A nonlinear panel data analysis

1.1.1 Motivation of the problem

This study sheds light on the intricate relationship between renewable energy usage and economic growth in OECD countries. The document is designed to serve as a reference for policy and government decisions, as well as for formulating informed investment plans to enhance long-term development and economic growth in the focal countries. The work also contributes to the greater scholarly discussion on energy and economics.

The motivation behind the paper is to analyse and demonstrate the non-linear positive relationship between renewable energy and economic growth. The paper provides a scientific foundation for developing renewable energy, optimising energy structures, and achieving the goal of low-cost energy transformation as well as providing lessons for renewable energy development in these countries.

Data analytics and visualization are crucial tools for understanding complex data in the fields of renewable energy and economic growth. Panel Data Analysis, Time Series Analysis, and Advanced Machine Learning Algorithms are some of the techniques used. These techniques help uncover trends, patterns, and seasonality while clustering group variables based on patterns. Geospatial visualizations, heatmaps, scatter plots, and interactive dashboards, are used to examine patterns, relationships, and trends. They provide an overview of distribution, illustrate correlations or causal relationships, depict relationships between variables and allow researchers to explore data in time and test hypotheses.

1.1.2 Dataset and Pre-processing Procedures

The research selected annual data from 34 OECD member nations from 2005 to 2016 as panel data samples to further investigate the relationship between renewable energy and economic growth. the attributes used are:

1. Explanatory variable: the total amount of Renewable Energy (RE) consumed.
2. Explained variable: the level of economic progress using each nation's actual GDP.

Threshold variables:

3. Non-renewable energy intensity (NEI): this demonstrates how dependent the country is on non-renewable energy.
4. Urbanization level (UL): This shows the process of and degree of population concentration in the city.
5. Per capita income (PCI): The ratio of the gross domestic product to the total population.

Control variables:

6. Industrial Structure (IS)
7. Population Structure (PS)
8. Energy Structure (ES)
9. Openness (OP)
10. Energy Price (CPI)

All ten variables of this dataset are numerical.

Pre-processing steps

The data was gathered and transformed to fit its purpose using the following steps (initial pre-processing was carried out in the data sources):

- **Panel Data**: Over 12 years, the data was arranged as a panel with 34 OECD countries. This panel data format enabled the use of panel regression models.
- **Logarithmic transformation**: The natural logarithm of each variable (aside from the control variables) was used to remove heteroscedasticity and unify the ten variables.
- **Differencing and Unit Root Test**: A Fisher-ADF and Fisher-PP two-unit root test was carried out to prevent spurious regression. This guarantees that there are no trends or seasonal components in the data, leading to false regressions. The test showed non-stationarity for ES, CPI, and LNNEI. The seven variables had to pass first-order difference; the first differentiation was used to achieve stationarity for the others. This transforms the data to reflect year-over-year changes rather than absolute values.
- **Threshold Determination**: different threshold models were used, and likelihood ratio tests were used to determine the optimal number of thresholds and estimate threshold values.

1.1.3 Analytical techniques and visualizations

1. *How does the impact change at different levels of Non-renewable energy intensity, Urbanisation level, and Per capita income? Using a Threshold Line Graph*

The line graph brings to light the impact of thresholds on the threshold variables, using a significance level of 5% and a narrow confidence interval. It demonstrates how renewable energy consumption affects economic development in a two-tiered manner, impacting both per capita income and urbanization level, while also having a single-tier effect on non-renewable energy intensity.

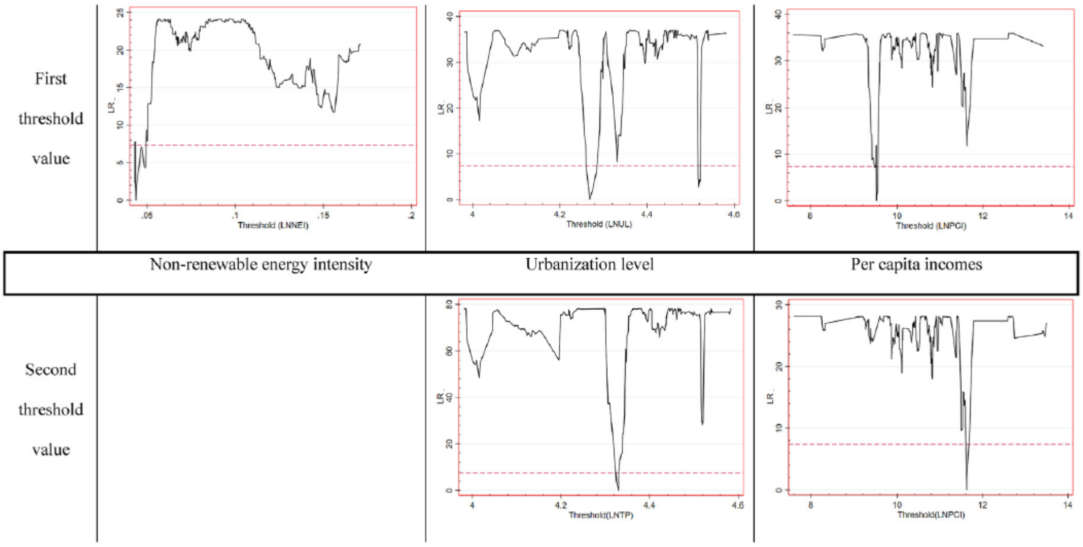


Fig.1.1.1 – threshold/line graph

To estimate the number of thresholds for the variables, a likelihood ratio test was conducted. After 500 iterations of sampling, images of the likelihood function findings were obtained depicting the estimated values and confidence intervals for the three threshold variables. The use of a threshold graph is quite handy in cases where the relationships between elements hinge on thresholds. It provides a versatile and efficient means to capture complex connection patterns across a wide range of fields.

2. *How has the composition of countries across threshold regimes changed over time? Using an Area Chart*

The Area Chart was used to visualize how the number of OECD countries falls into different threshold intervals. It's interesting because each country responds differently to these threshold levels over time.

As a result, the number of countries in these three intervals changes. These area charts are handy because they make it easier to grasp cumulative trends and make sense of complicated relationships.

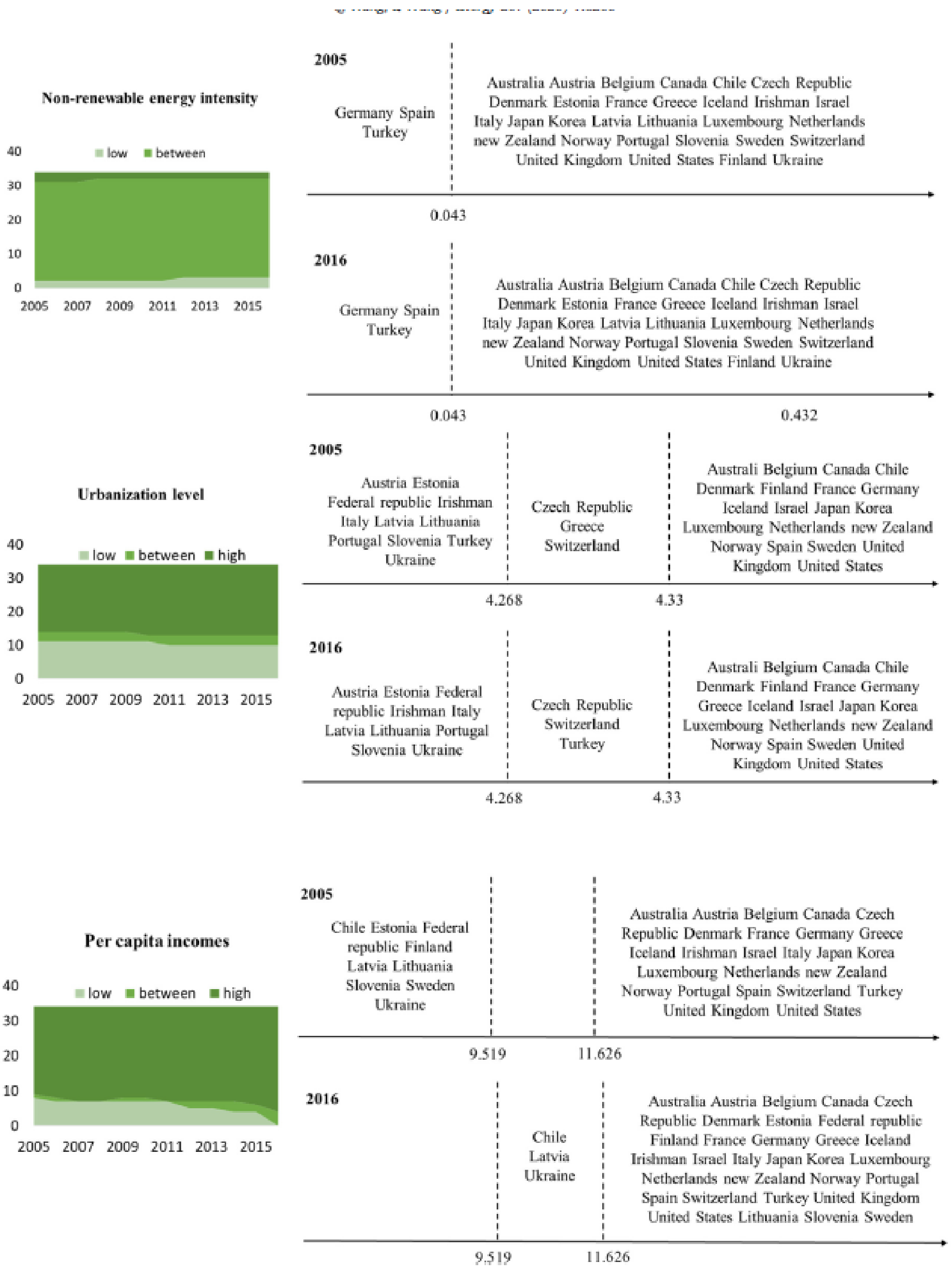


Fig 1.1.2 – area chart

Findings

- The charts reveal that most OECD nations have already crossed the threshold when it comes to non-renewable energy intensity. This suggests they heavily rely on non-renewable energy sources, which highlights the challenges involved in adopting renewable energy technologies and transforming this sector.
- Regarding urbanization, most countries have high urbanization levels. This shows that even though there's not much change in the overall urbanization rate, having a high level of urbanization can be advantageous for economic growth through renewable energy adoption.
- Lastly, for the per capita income threshold some OECD countries can overcome it under the right circumstances and with the necessary resources. This indicates the potential of renewable energy in supporting economic growth, even though it's not without its challenges for some member states.

1.2 Breast Cancer Data Analysis for Survivability Studies and Prediction

1.2.1 Motivation of the problem

This research paper focuses on predicting cancer survivability, particularly in breast cancer, using innovative machine-learning techniques. Motivated by cancer's global impact as the second leading cause of death, the study addresses key factors:

- *Improved Prognosis: Timely diagnosis and treatment significantly enhance cancer patient prognosis.*
- *Impact of Cancer Staging: The TNM classification system assesses cancer spread and stage, influencing treatment outcomes.*
- *Healthcare Disparities: Variations in survivability based on age, ethnicity, and race underscore healthcare disparities.*

Health care sector utilizes Self-Organizing Maps (SOMs), Density-Based Spatial Clustering (DBSCAN), and Multilayer Perceptron (MLP) for data analysis. Techniques include scatter plots for cluster identification, box plots for survivability distribution, codebook vectors for multi-dimensional mapping, and line graphs for cross-validation results. These tools empower researchers to analyse patient data, predict survivability, and tailor treatments for improved patient care. In the context of breast cancer, this comprehensive approach combines innovative analytics with visualization tools for enhanced insights and personalized treatment strategies, addressing a critical global health challenge.

1.2.2 Dataset and Preprocessing Steps

The dataset, originating from the SEER program, spans 1973 to 2012 and encompasses 740,506 records with 146 variables. It includes data on breast cancer incidence in the United States, offering insights into tumour features, cancer staging, patient demographics, survival outcomes, and AJCC cancer staging.

Preprocessing Steps

Preprocessing steps ensured data readiness for survivability prediction.

- Stringent inclusion criteria were applied to retain relevant cases with known survival history.
- Irrelevant variables like "FIRSTPRM," "BEHANAL," and "SRV_TIME_MON_FLAG" were removed.
- Due to post-2010 data limitations, HER2 status was excluded.
- The top 29 attributes essential for survivability analysis were chosen based on information gain.
- Continuous variables were scaled for consistency, while categorical ones underwent binary encoding.

- Effective handling of substantial missing data involved either exclusion or suitable imputation techniques.

Final Dataset

Post-preprocessing, the refined dataset comprises 85,189 cases with 26 key variables. It comprehensively captures critical aspects of breast cancer cases, including patient demographics, tumour characteristics, treatment methods, and survival-related data. This meticulously curated dataset serves as a robust foundation for subsequent analyses, with a specific focus on predicting breast cancer survivability.

1.2.3 Analytical techniques and visualizations:

In this study, Self-Organizing Maps (SOMs) were used in prioritizing clarity over complex Multilayer Perceptrons (MLPs). SOMs excel at understanding data, even with missing values. They are versatile, handling tasks like clustering, dimension reduction, and visualizing high-dimensional data. By preserving spatial connections, they unveil breast cancer survivability complexities. SOMs identified patient groups and explored influential factors.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identified patient cohorts with varying sizes and shapes, enhancing breast cancer survivability prediction accuracy. This complements the Self-Organizing Maps (SOMs) used for understanding patient clustering dynamics. The Multilayer Perceptrons (MLPs), known for their interpretability, were used to fine-tune predictions based on DBSCAN-identified clusters. This innovative strategy optimizes breast cancer survivability forecasts, prioritizing accuracy, and transparency in this methodology, providing valuable insights into patient outcomes.

1. *What can be learned about the patterns and trends among patients with shared characteristics*

Figure 1.2.1 scatter plot visually represents the distribution of patients among nine distinct clusters. Each point denotes a patient based on SOM and DBSCAN results, revealing how patients with shared characteristics form clusters, aiding trend, and pattern recognition. This plot simplifies identifying natural groupings and gauging the clustering algorithm's efficacy by highlighting distinct patient clusters.

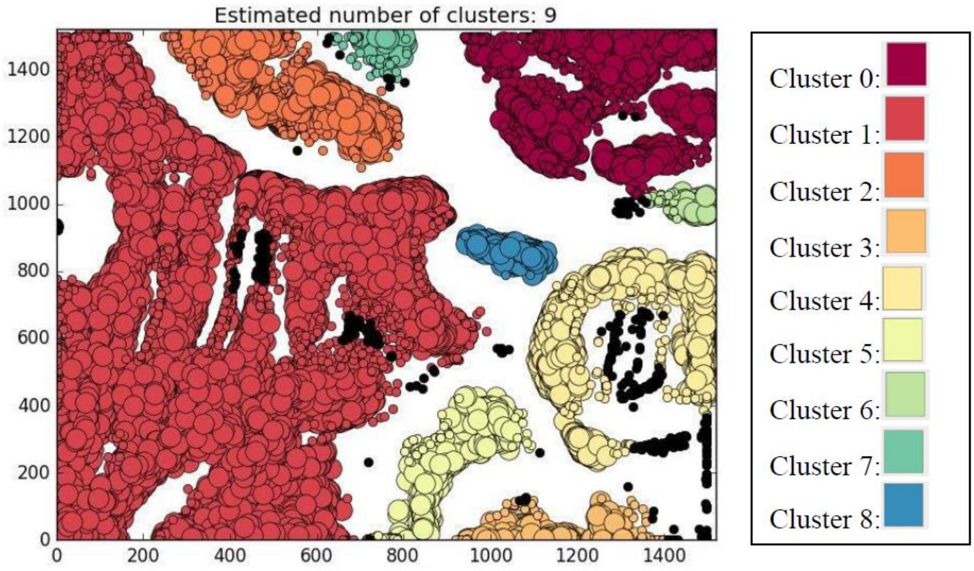


Fig 1.2.1 – scatter plot

2. *What are the impacts of different attributes on breast cancer patient outcomes within the clusters?*

Figure 1.2.2 employs a box plot to display the survivability distribution across nine patient clusters. This visualization highlights variations in patient outcomes, offering insights into attribute-related impacts. Box plots facilitate comparing survivability tendencies among clusters, aiding attribute significance assessment in breast cancer patient outcomes.

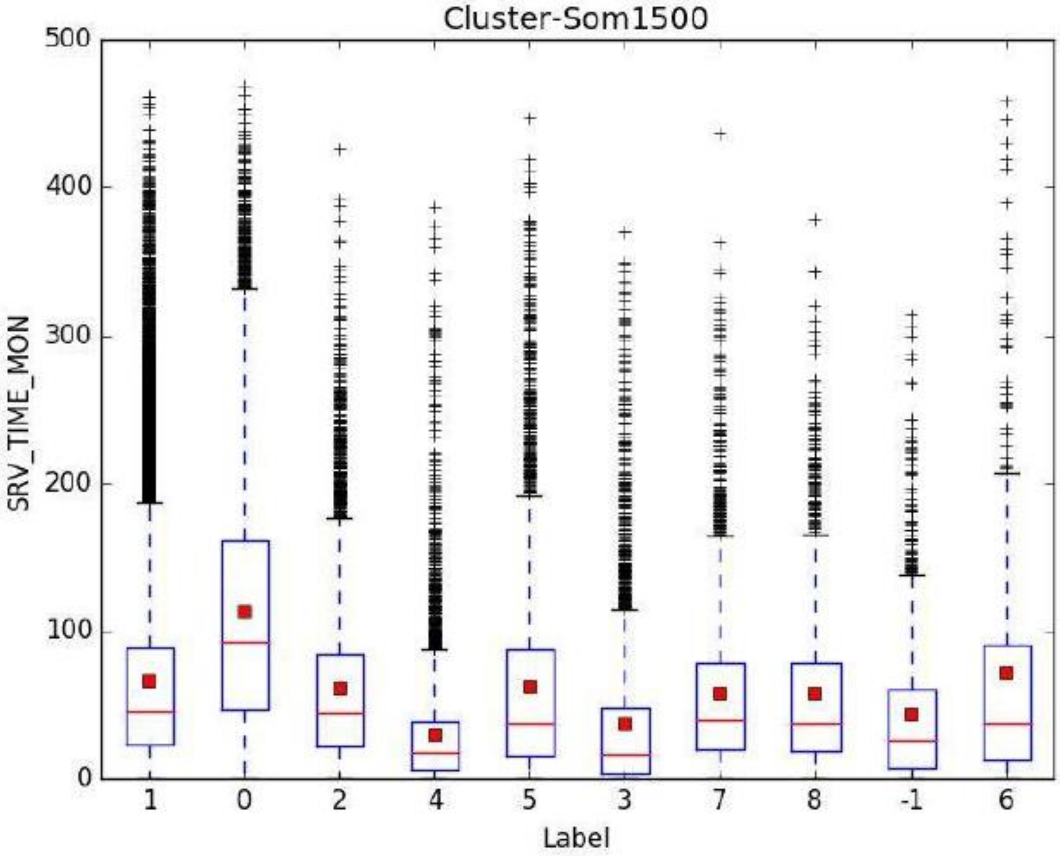
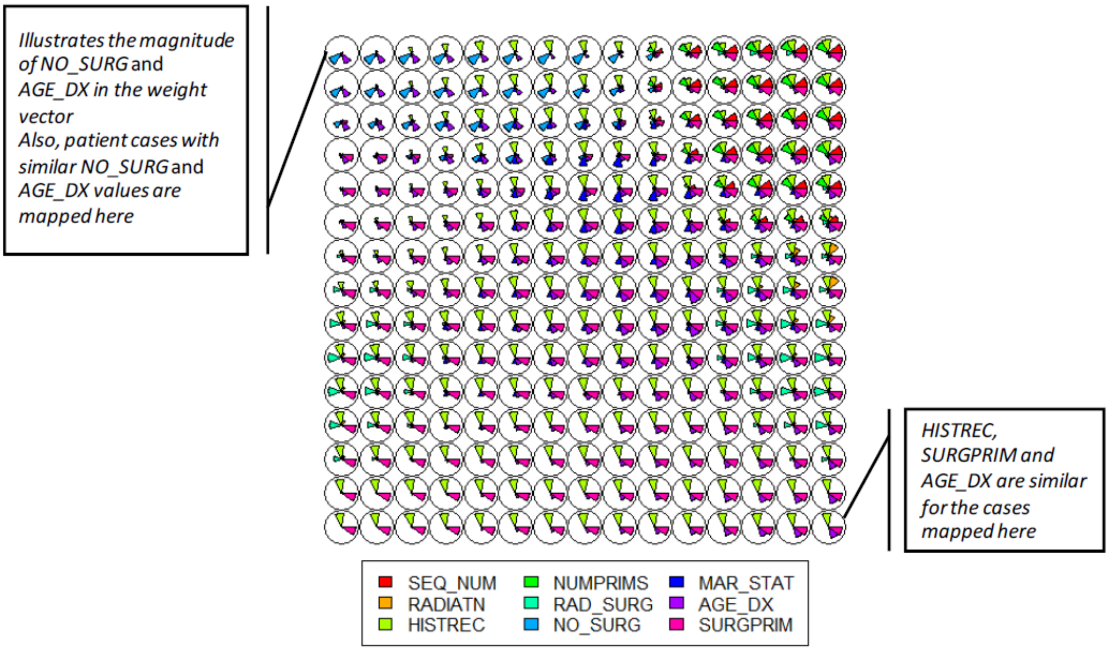


Fig 1.2.2 – box plot

3. *What insights can be gained regarding the attributes driving the clustering of breast cancer patients?*

Weight vectors are used in Figure 1.2.3 to illustrate how patients are grouped into clusters by the SOM based on their attribute values. These weight vectors visually depict the distribution of variables within node clusters, aiding in the identification of patterns and similarities among patients. Weight vectors help researchers understand which attributes are driving the clustering, facilitating the discovery of patient cohorts with shared characteristics.



Illustrates the magnitude of NO_SURG and AGE_DX in the weight vector
Also, patient cases with similar NO_SURG and AGE_DX values are mapped here

HISTREC, SURGPRIM and AGE_DX are similar for the cases mapped here

Fig 1.2.3 – weight vectors

4. *What conclusions can be drawn regarding the methodology's consistency and robustness in predicting breast cancer survivability?*

The line graph in Figure 1.2.4 is used to show the 10-fold cross-validation performances of MLP on different clusters, specifically cluster 1 (with high mean survivability) and cluster 4 (with low mean survivability). It serves to demonstrate the consistency and robustness of the proposed approach in predicting survivability across various patient clusters. The graph helps visualize how well the methodology performs across different subgroups of patients and highlights the improvement in prediction accuracy compared to the baseline MLP model.

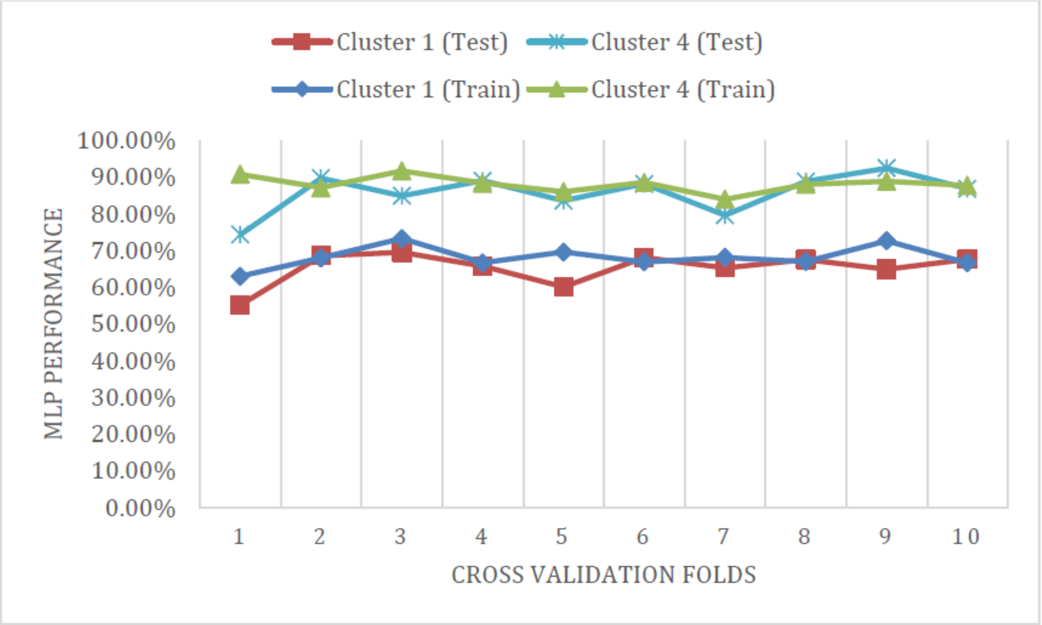


Fig 1.2.4

1.3 Anti-Money Laundering: Using data visualization to identify suspicious activity.

1.3.1 Motivation of the problem

As anti-money laundering regulations evolve, it is becoming increasingly difficult for financial institutions and regulators to stay ahead of the latest scams therefore, there is a need for the development and use of better technology in the management of anti-money laundering activities (Ray, 2015) as efficient data analytics can help regulators identify suspicious activities easily.

This paper explores the use of visualisation techniques for easy identification of patterns of money laundering activities. The timely interception of criminal activities and effective anti-money laundering control will also combat crimes such as drug trafficking and terrorist financing.

Analysts previously tracked financial transactions by using spreadsheets to review large amounts of data however, this was inefficient as they could not easily identify trends or correlations from the data (Chang et al., 2007). Data analytics techniques like descriptive analytics helped analysts gain insights into past trends, performance of investment history and fraud pattern detection. Machine learning techniques like time series and regression are also used to predict market trends and stock prices and the communication of insights to stakeholders is clearer with interactive dashboards, graphs, plots etc.

1.3.2 Dataset and Pre-processing Procedures

The data for visualisation in this paper is a sanitised dataset of bank transactions spanning two years (2014-2016) sourced from a large entity which is undisclosed for confidentiality reasons. To ensure privacy, the data was anonymised and stripped of client-identifying features like account numbers and other sensitive information. Although the preprocessing steps are not mentioned in the paper, the prototype (ink) system architecture shows that the dataset was pre-processed, and the following attributes were extracted:

- **Target_BSB & Target_Account** – Bank account of suspicious entity.
- **Dest_BSB & Dest_Account** – Bank account of customers or vendors.
- **Num** – Number of cashflows between target accounts
- **Sumamt** – Total sum of suspicious transactions for a given period.
- **DRCR** – This refers to credit and debit transactions.

After pre-processing, standardised data was inputted into an SQL database and the result was computed by the ink to generate the DOT code uploaded to GraphViz for visualisations.

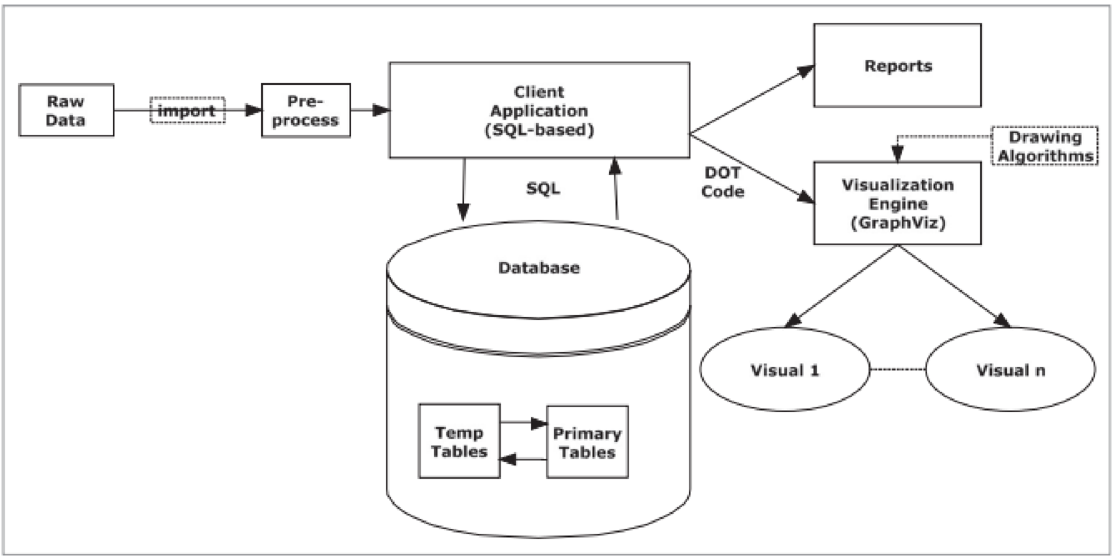


Figure 1.3.1-ink System Architecture

1.3.3 Visualisation and Analytics Technique

This paper proposes the use of Link Analysis in the detection of money laundering activities by visualising financial transactions between entities to identify suspicious patterns. To validate this strategy, a prototype application - ink was developed and demoed to show how the application would work in practice.

To achieve the objectives of the paper, the following questions were raised:

1. *How can visualization techniques assist in the efficient identification of patterns of money laundering activities?*

Since money laundering uses layers of transactions to disguise relationships, visualisation techniques can easily reveal these networks of relationships as they enable pattern recognition by revealing trends, prioritising suspicious alerts, and reducing false positives. Therefore, Link analysis is proposed as it demonstrates the degrees of separation between entities while revealing hidden trends and patterns (Singh and Best, 2016; Watkins et al., 2003).

2. *How can link analysis be applied in the detection of suspicious bank transactions?*

Link analysis was used to examine the relationship between networks and entities in this paper, the visualizations were outputted using GraphViz. GraphViz was the only visualization tool used in this study because it is affordable and open-source. Also, the authors of this paper wanted a static visualisation that was non-interactive. In this study, bank accounts are represented using nodes while transactions are represented using links.

To output visualisations, the data was imported into the ink, pre-processed, and standardised for SQL queries based on these money laundering indicators as provided by AUSTRAC, (2014):

- High volume of transactions within a short period or Structuring cash deposits or withdrawals.
- U-turn (reversed) transactions.
- Funds transfer involving banks in ‘interesting’ countries.
- Payments of loans by external parties.

The results of the queries are then stored in a temporary table as the source data for visualisation before being computed by the ink to generate the DOT code uploaded to GraphViz for the visualisations. Below is a sample of the outcome of the Link analysis.

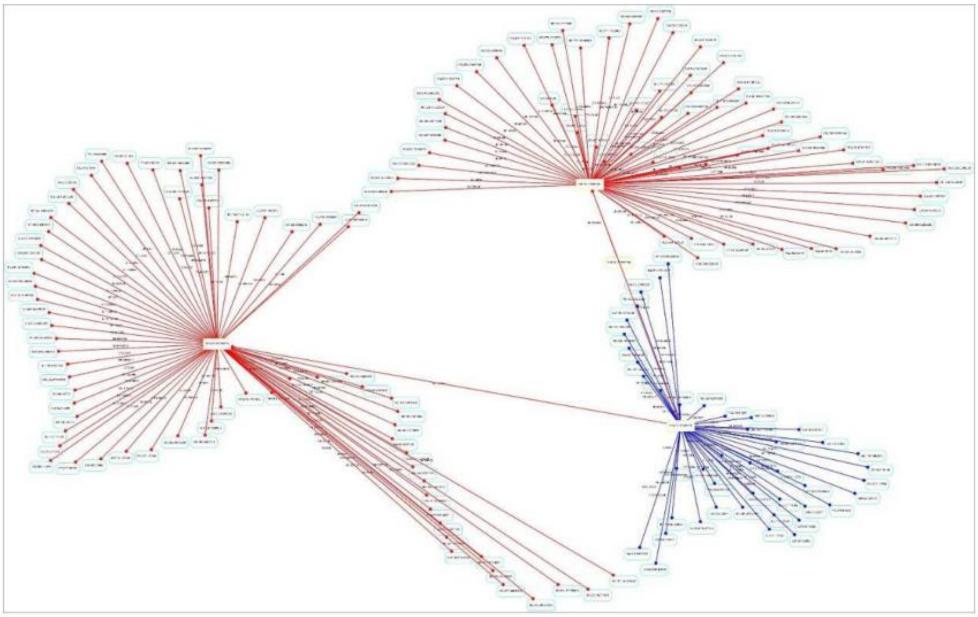


Fig 1.3.1 – Relationship between target accounts

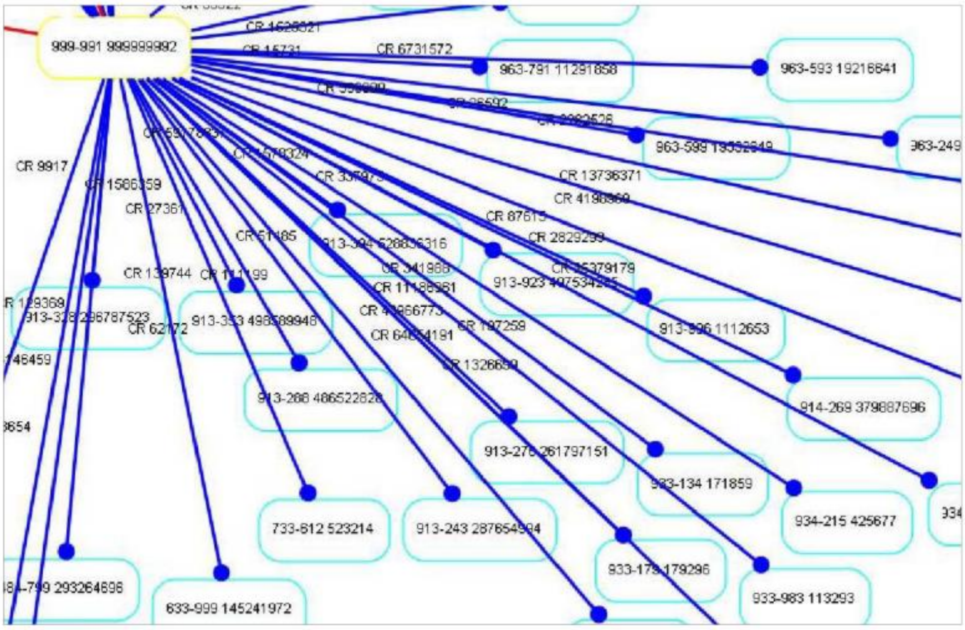


Fig 1.3.2 – Relationship between target accounts – expanded view

While Link analysis graphs are a great way to represent graphs and analyse complex relationships, overlapping nodes can easily cause problems of edge congestion and visual clutter especially when large datasets are being visualised (Von Landesberger et al., 2011). The data used in this study is small, showing only a high-level view of bank transactions which makes it more difficult to replicate for larger entities. This technique also makes it impossible to modify the positions of nodes to reduce visual clutter.

2. Part B

2.1 Description of Dataset

The dataset, titled "Violence Against Women and Girls," was sourced from Kaggle and encompasses data from 70 countries spanning 2005 to 2018. This dataset was part of the Demographic and Health Surveys (DHS) program. Its selection adheres to specified assessment criteria and stems from the dataset's societal significance. Given the rising global concern surrounding violence against women and girls, it was chosen as a pertinent subject for the analysis. Below are the details of the dataset.

- Number of rows – 12601 rows
- Number of columns – 8 columns
- Number Of data points – 100808

Table 2.1 Attribute Type and Description

Name of Attribute	Description of Attribute	Type of Attribute
RecordID	Numeric values unique to each question by country – 1 to 420 values	Discrete numerical data
Country	The country in which the survey was conducted – 70 different countries	Nominal categorical data
Gender	Male or Female – two unique values	Binary categorical data
Demographics Question	The different types of demographic groupings used to segment respondents. – marital status, education level, employment status, residence type, or age	Nominal categorical data
Demographic Response	Refers to demographic segment into which the respondent falls – 15 unique values	Ordinal categorical data
Question	Respondents were asked if they agreed with the 6 unique statements:	Nominal categorical data
Survey Year	The year in which the survey took place	Discrete numerical data
Value	% of people surveyed in the relevant group who agree with the question	Continuous numerical data

2.2 Data Pre-processing

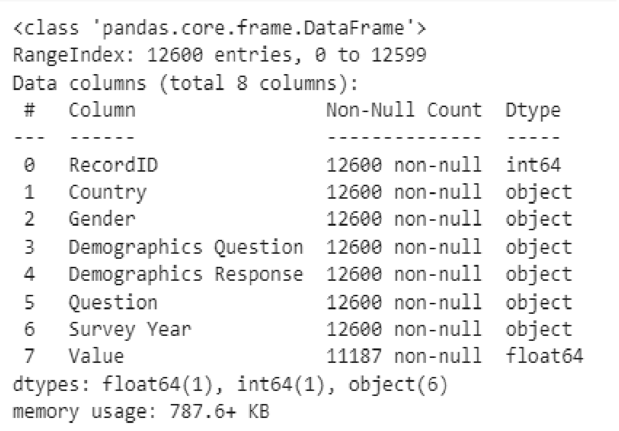
Table 2.2 Preprocessing Techniques Employed and Benefits

Pre-processing Technique Used	Benefits of Using Technique	Case Description for Dataset
Data Information	Provides a concise and informative summary of the dataframe including the memory space, dtype and non-null count.	The dataframe occupies 787.6+ KB of memory space and consists of 12600 entries and 8 columns. The datatype of each column and corresponding non-null count is obtained.
Missing values/null values	Missing data can reduce the statistical power of the analysis which may lead to misleading results and inaccurate conclusions, as such it is important to handle missing data appropriately.	The dataset showed that only the Value column had missing values. For the Value column, out of 12600 entries 1413 entries are missing. This implies that 11.2% of the Value column is missing, as a result, the null values were dropped from the dataset.

Check for duplicate values	Helps ensure data quality and accuracy, prevents errors and inconsistencies in analysis and avoids potential biases or misinterpretations that can arise from redundant information, leading to more reliable and valid results.	No duplicated row observed.
Variable selection	Simplifies the dataset and reduces the risk of data leakage.	‘RecordID’ column was removed due to its irrelevance.
Normalisation	Normalisation is required depending on the type of analysis to be carried out. Normalisation is used to ensure that features are on a consistent scale, and are easier to interpret and less affected by outliers.	The demographic response column was used to develop a 100% stacked bar chart, this necessitated normalising the column. This enabled better comparisons and pattern identification across the categories.

The results from running the code are provided below:

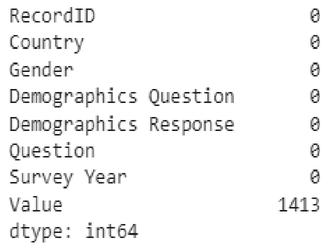
1. df.info(): in order to get a better understanding of the dataset, the python function was used to display the properties of the dataframe. This makes it effortless to observe the relationships between the features as well as identify missing values. This revealed that the DataFrame has 12,600 entries and 8 columns with mixed data types. Notably, the "Value" column contained missing data, with 11,187 non-null entries.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12600 entries, 0 to 12599
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   RecordID               12600 non-null  int64  
 1   Country                12600 non-null  object 
 2   Gender                 12600 non-null  object 
 3   Demographics Question  12600 non-null  object 
 4   Demographics Response  12600 non-null  object 
 5   Question               12600 non-null  object 
 6   Survey Year            12600 non-null  object 
 7   Value                  11187 non-null  float64
dtypes: float64(1), int64(1), object(6)
memory usage: 787.6+ KB
```

Fig 2.2.1- dataset info

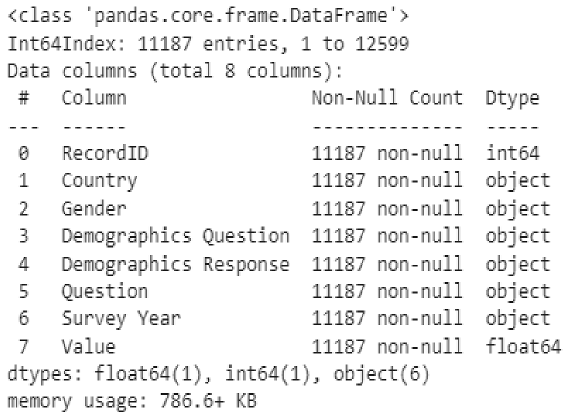
2. df.isnull().sum(): The code was used to identify and count missing values in each column of the DataFrame. The results show that only the "Value" column has missing values, with a total of 1,413 null entries.



```
RecordID                    0
Country                     0
Gender                      0
Demographics Question       0
Demographics Response       0
Question                    0
Survey Year                 0
Value                    1413
dtype: int64
```

Fig 2.2.2- dataset info

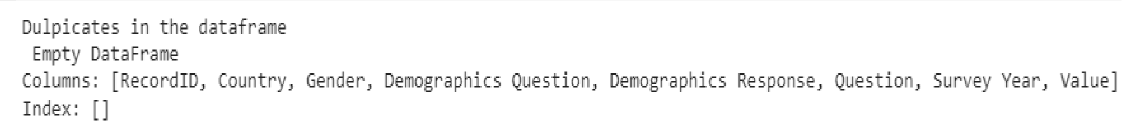
3. df.dropna(subset = [‘Value’]): The code was employed to eliminate rows with missing values in the "Value" column of the DataFrame, resulting in a DataFrame containing 11,187 non-null entries across all columns. This operation reduced the df ‘s memory usage to 786.6+ KB.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11187 entries, 1 to 12599
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   RecordID               11187 non-null  int64  
 1   Country                11187 non-null  object 
 2   Gender                 11187 non-null  object 
 3   Demographics Question  11187 non-null  object 
 4   Demographics Response  11187 non-null  object 
 5   Question               11187 non-null  object 
 6   Survey Year            11187 non-null  object 
 7   Value                  11187 non-null  float64
dtypes: float64(1), int64(1), object(6)
memory usage: 786.6+ KB
```

Fig 2.2.3 – dataset info

4. df[df.duplicated()]: The result, "Empty DataFrame," indicates that no duplicate rows were found in the DataFrame, meaning that all rows are unique based on the specified columns.



```
Dulpicates in the dataframe
 Empty DataFrame
Columns: [RecordID, Country, Gender, Demographics Question, Demographics Response, Question, Survey Year, Value]
Index: []
```

Fig 2.2.4 Dataset info

2.3 Visualisation for Decision-making

Domestic violence against women is a critical issue, with our study demonstrating its prevalence in 70 countries. Despite policy and advocacy efforts, domestic violence remains alarmingly high, affecting women across different demographics. This complex problem necessitates a comprehensive approach, including policy review, demographic analysis, and strategic interventions. This analysis emphasises the need to carry out a detailed investigation of the demographic elements at work, and the urgency of addressing domestic violence against women.

2.3.1 Seeing the Big Picture I

Visualisation tool: Chloropleth map

A choropleth map effectively visualizes the prevalence of violence against women and girls in 70 countries, providing a global perspective. This tool facilitates quick comparisons, using colour to highlight variations in prevalence rates across nations. Its visual clarity simplifies data comprehension, especially for those less familiar with statistics. The map's geographic context aids in understanding the issue's worldwide distribution, supporting interactive data access and awareness-building. This map type's colour shading indicates violence levels in different countries, enabling the identification of high and low-prevalence areas, making it an ideal choice for conveying this critical information.

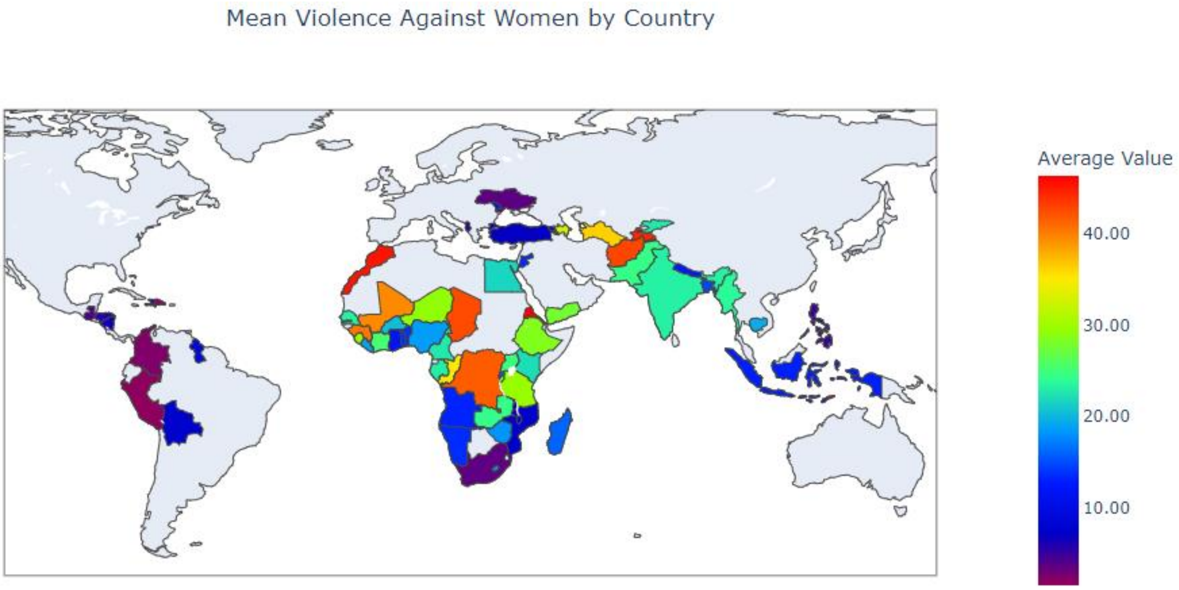


Fig 2.3.1 – Chloropleth map

Insights

This map offers valuable insights into identifying high-prevalence countries, enabling focused resource allocation and progress tracking over time. It enhances awareness by effectively conveying the issue's magnitude and urgency to a broader audience. Observing the disparity in values obtained in the countries raises questions like "what influences these disparate values" which is further enhanced by the colours of the map. Further research could be conducted to identify and analyse why some areas are more vulnerable to violence against women like in the case of Morocco with a bright red and South Africa with dark purple.

This visualization reveals regional disparities, prevalence hotspots, global scale, outliers, comparative data gaps, and temporal trends, empowering policymakers, and organizations to strategically address this global concern through targeted interventions.

2.3.2 Seeing the Big Picture II (top 10 countries)

Visualisation tool: Scatter plot

Scatter plots are utilized to depict the connection between violence rates. The country is marked on the x-axis and violence rates on the y-axis, aiding in trend identification and correlation assessment. Employing scatter plots for regional insights provides benefits such as geographical context, pattern recognition, and correlation analysis, making them valuable for understanding regional violence patterns and data-driven decision-making.

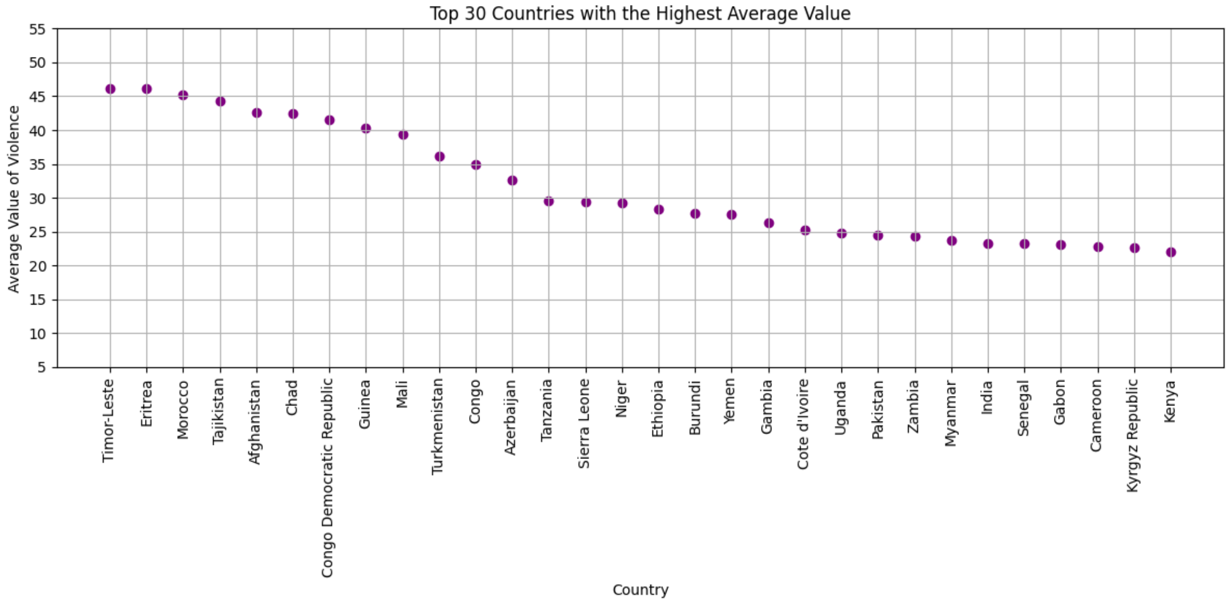


Fig 2.3.2 – Scatter plot

Insights

They help identify high-prevalence areas. They track trends, aiding in targeted interventions and policy assessment. This plots the average values against countries and helps identify outliers, compare country averages, reveal regional patterns, spot data quality issues, and uncover potential socio-economic or cultural trends.

2.3.3 The Different Faces of Violence

Visualisation tool: Horizontal bar charts

Horizontal bar charts offer excellent readability, catering to those accustomed to left-to-right reading. They effectively convey the frequency of categorical variables, aiding in identifying the most common reasons. Bar charts simplify complex violence data for broader comprehension, aiding those unfamiliar with statistics. They enable clear comparisons among violence types, displaying their prevalence. These charts are versatile, visually representing categories and values, facilitating comparisons. Additionally, they can depict temporal trends when necessary.

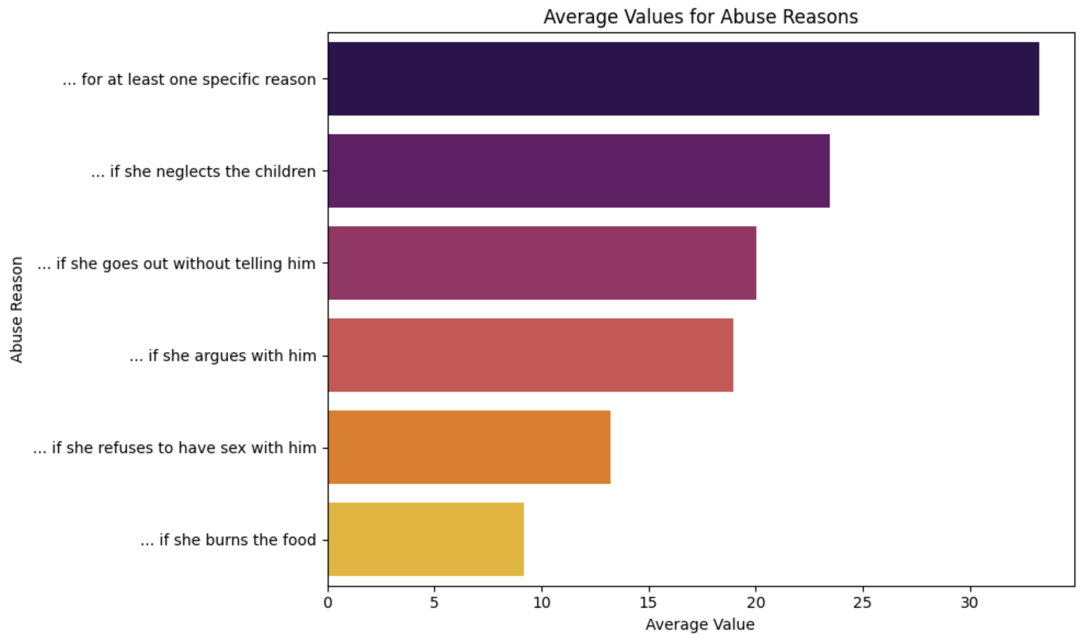


Fig 2.3.3 – Horizontal bar charts

Insights

Decision-makers can use the bar chart to identify and target the most common types of violence; this information can be used to prioritise initiatives to address the specific cause. For example, if one type of violence is more widespread than others, policymakers can prioritise implementing policies that have proven to be effective in addressing that type of violence. Comparing the incidence of diverse types of violence can assist in identifying areas in which additional work is required.

2.3.4 Who is Most Vulnerable

Visualisation tool: Heatmap

The heatmap provides a clear representation of the connections between demographic and socioeconomic variables and their influence on violence risk. By utilizing colour gradients, it effectively communicates the strength and patterns of these associations. Each cell in the heatmap corresponds to a demographic category, with darker colours signifying a more pronounced relationship between the demographic question and responses.

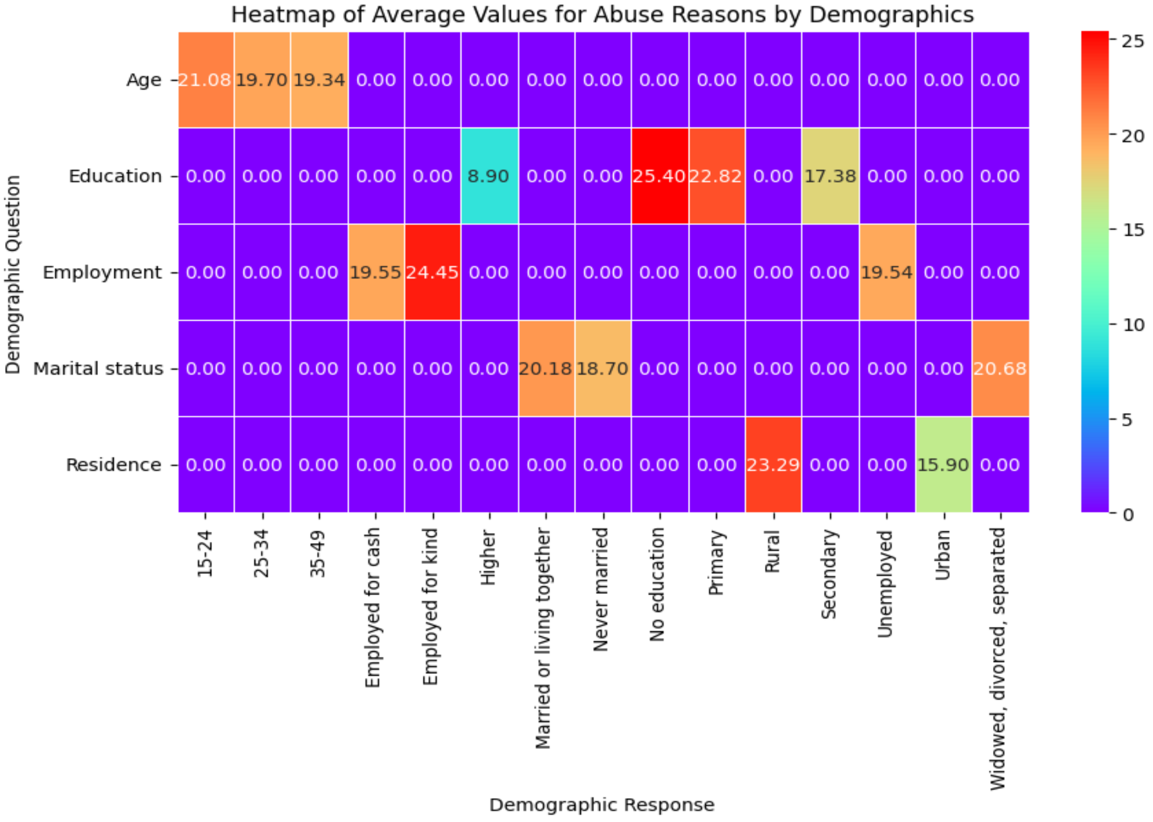


Fig 2.3.4 – Heat map

Insights

The visual aids in pinpointing areas for targeted interventions using demographic and socioeconomic factors. Decision-makers can use it to promote awareness and discussions regarding the connections between these factors and violence risk. The visual also reveals the strength and direction of correlations, identifies vulnerable groups, protective factors, and high-risk categories, addresses data gaps, and informs policy decisions. It effectively presents complex data relationships visually.

2.3.5 Considering Everyone's Story

Visualisation tool: Grouped bar charts

Grouped bar charts are a valuable visualization tool for exploring the correlation between male and female responses to specific questions. They excel in providing a side-by-side comparison of gender-based data, making it straightforward to identify disparities, similarities, or patterns in how males and females respond to various queries.

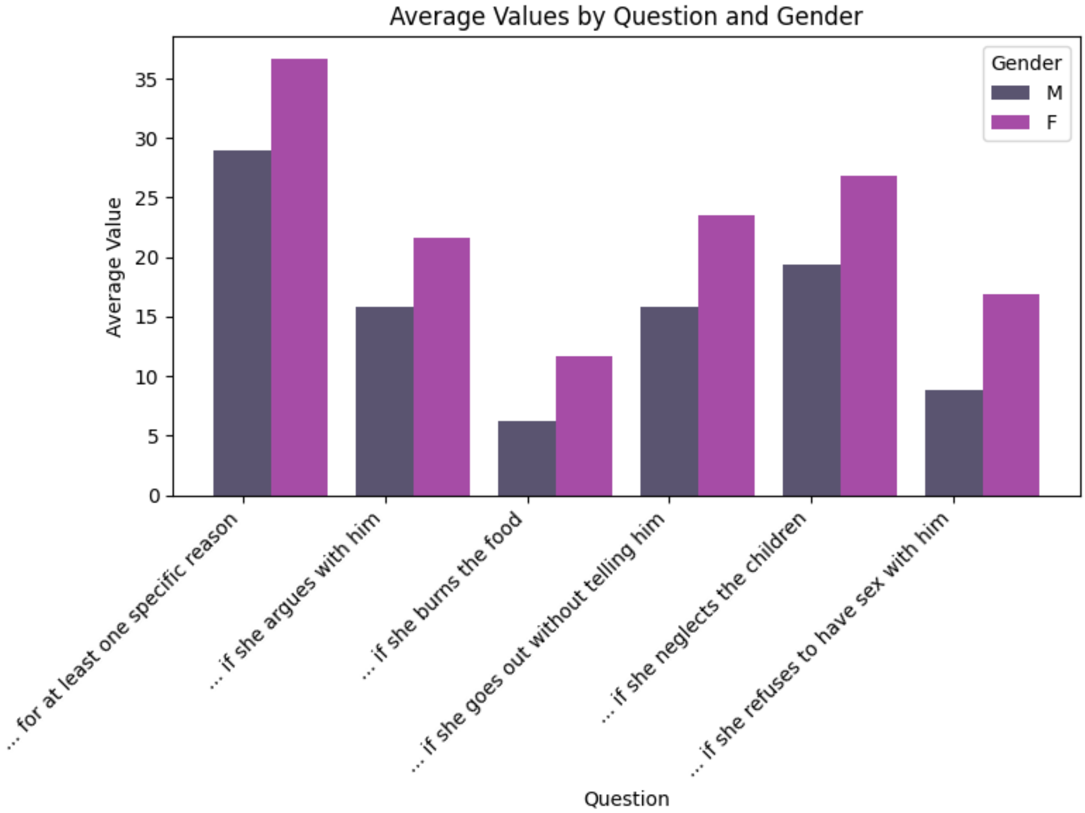


Fig 2.3.5 – Grouped bar chart.

Insights

This chart offers insights into gender-specific trends or differences within the dataset, shedding light on areas where responses diverge significantly or align closely. This gender-sensitive analysis is crucial for decision-makers and researchers aiming to address gender-related issues effectively, ensuring policies and interventions consider everyone's perspectives and needs, promoting inclusivity and equity. This graph clearly shows that the response from women in all the categories is higher than that of men, which implies that women experience greater level of violence than men are willing to admit.

2.3.6 Education and violence

Visualisation tool: Violin plots

Violin plots are versatile data visualization tools that blend aspects of box plots and kernel density plots. They offer a detailed view of data distributions, making them useful for comparing multiple categories, identifying trends, and displaying relationships. Violin plots are effective for exploring complex datasets, such as the comparison of education levels and violence values between male and female respondents, offering insights that inform policy and interventions in areas like gender-based violence and education.

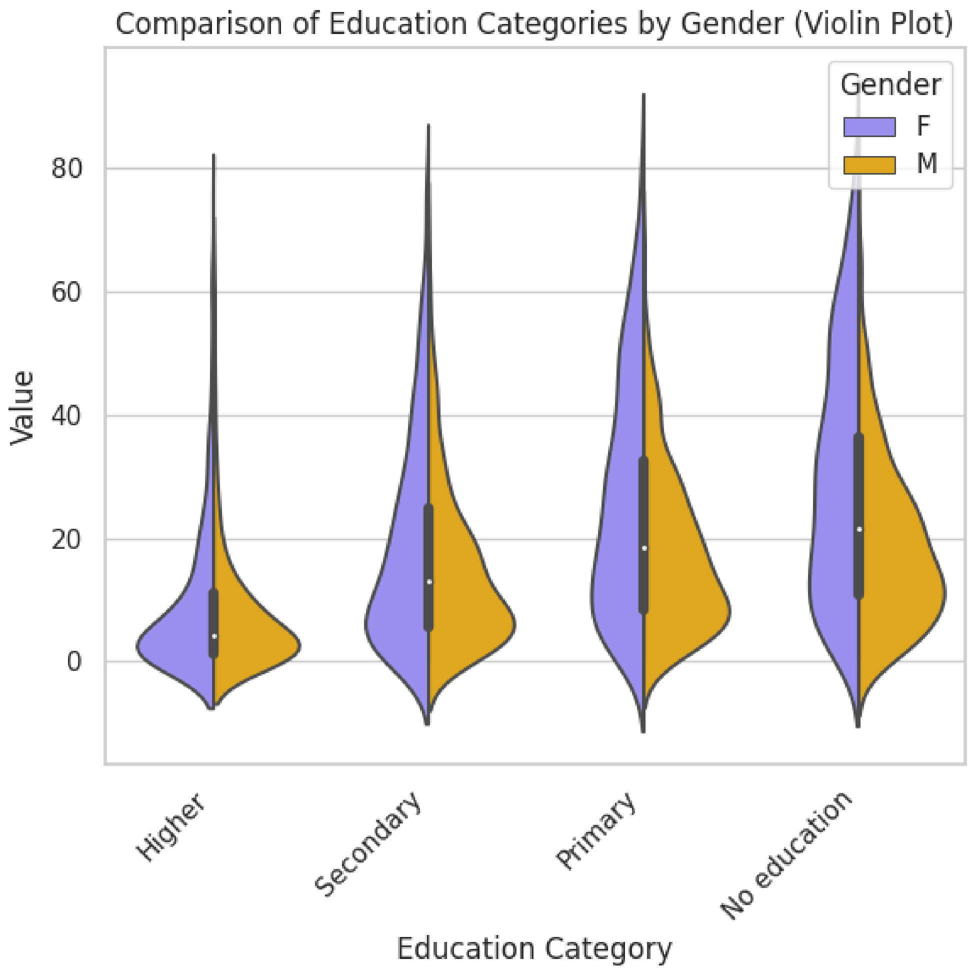


Fig 2.3.6 – Violin plot

Insights

In this context, violin plots help uncover disparities or commonalities in educational attainment between genders. Insights reveal as educational level increases, the value of violence associated with both male and female reduces. These visualizations empower decision-makers to formulate targeted strategies for educational equity, facilitating a nuanced understanding of gender-based disparities and the need for inclusive educational policies and interventions.

2.3.7 Regional Insights

Visualisation tool: stacked bar chart

A stacked bar chart is a useful tool for demonstrating how a bigger category is broken into smaller categories and what role each element plays in the total amount. The larger section in this instance is countries, and the sub-category is the demographic response that best describes the country. This provides a clear visualisation of the data, making it easy to see the countries' distinct demographics. This visualisation approach can be useful when attempting to depict trends over time, such as how demographics in countries change over time.

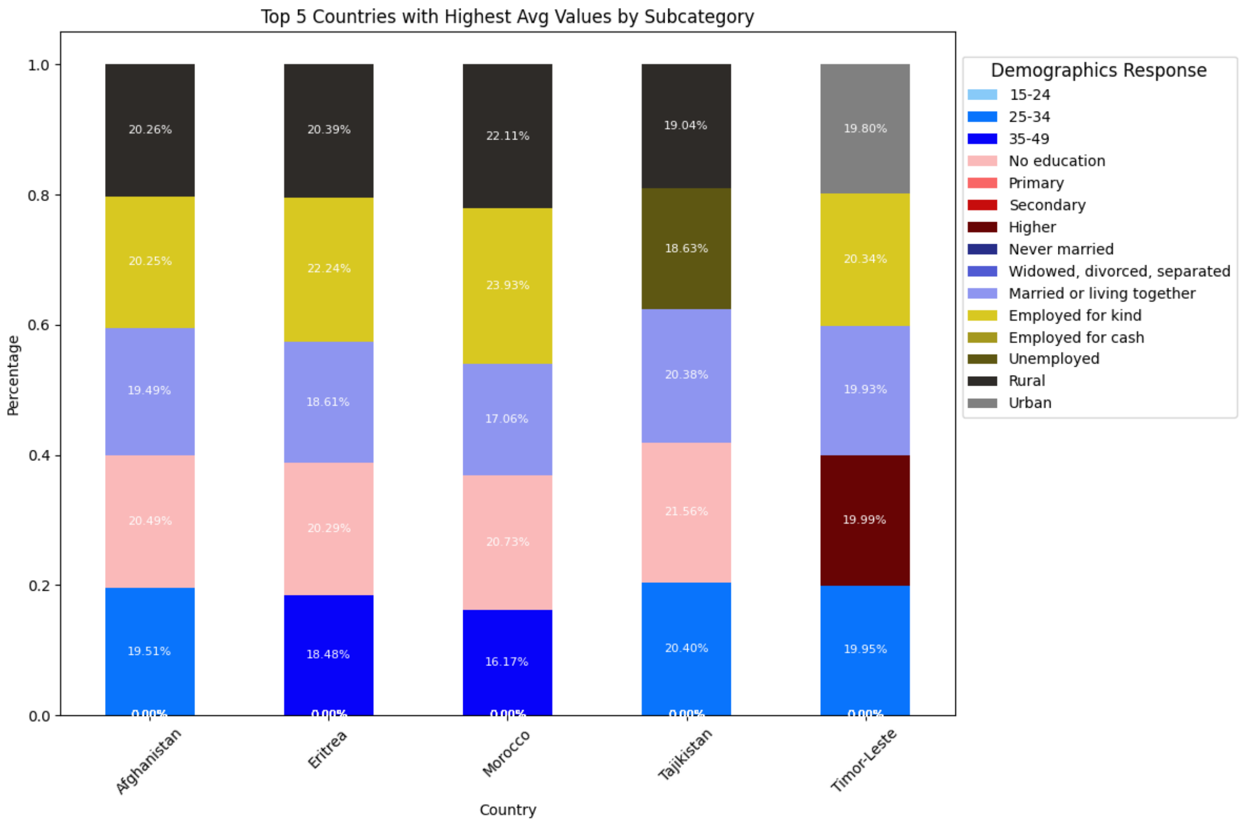


Fig 2.3.7 – Stacked bar chart

Insights

The top 5 countries with the highest occurrence of violence are shown in the stacked bar chart, which also identifies at-risk groups. It clearly shows the basic characteristics of both victims and potential victims. For instance, in Afghanistan, focused interventions should prioritise women of 25-34 years with no formal education, living in remote areas, and married with no cash income. In terms of intervention, community-based activities focusing on education, economic empowerment, and gender equality, with an emphasis on young girls in rural regions, should be undertaken. Collaborating with local organisations and governments to provide resources, support, and public awareness campaigns in response to violence against women and girls in this demographic.

2.3.8 Take Action

As observed, the top 5 countries where women experience the most violence are Afghanistan, Eritrea, Morocco, Tajikistan and Timor-Leste; and the most vulnerable women in these countries are married or living together in rural communities, employed for kind, have no education, and are aged 25– 34 years.

For further assessment, it is important to review the most recent policy changes in these countries that seek to end domestic violence against women. As well as carry out in-depth analysis of the demographic factors of the focus countries. With the knowledge from this process, the following steps are our proposed actions.

1. Identify all relevant stakeholders in the target regions.
2. Examine the cultural landscape to determine the best course of action.
3. Identify NGOs and support organisations currently working in target communities.
4. Select and partner with NGOs and local organisations to facilitate advocacy and support efforts.

The immediate goal of this partnership will be to create and execute a multilevel awareness campaign strategy through the following:

- Educational workshops and seminars.
- Collaborations with local schools and community organizations.
- Awareness-raising events during designated awareness months (e.g., International Day for the Elimination of Violence against Women).

This multilevel partnership hopes to be a long-term approach with the goal of advocating for stronger laws against domestic violence in these regions. We propose continuous monitoring of the impact of the awareness campaigns and use the data to create more robust support services in these regions.

2.4 Big Data

2.4.1 Introduction to Big Data

Big data (BD) is a complex, large dataset gathered from new data sources. Because of the volume and versatility of this data, traditional data processing software cannot manage, process, or analyze it. (Anon, n.d.). Big Data (BD) **volume** is quantified in terabytes, petabytes, and exabytes, distributed across numerous computers. Its **velocity** emphasizes real-time data collection, posing challenges in timely analysis and interpretation (Valchanov, 2023). BD's **variety** spans structured, semi-structured, and unstructured data, complicating traditional processing methods. Concepts like **veracity** (bias and abnormality in data) and **validity** (extracting meaningful insights) often accompany big data, influencing analysis and decision-making.

Because of its transformative nature, there is an ever-increasing demand for the usage of big data:

- Real-time or near real-time data-driven decision-making rather than intuition or experience.
- Improved customer insights that enable personalized marketing, product recommendation and improved customer service.
- Enhanced operational efficiency by identifying inefficiencies and bottlenecks.
- Cloud integration has made big data technologies more accessible and affordable for organizations of all sizes.
- Internet of Things and Sensor data management.
- Training and deploying machine learning and AI models.

2.4.2 Big Data Techniques:

Big Data has shifted the paradigm of conventional data analysis. Its technologies are designed to capture, store, process, analyse, and visualize large and complex datasets. These technologies are:

- BD Storage and Management: Hadoop Distributed File Services (HDFS); HBase.
- BD Processing and Analysis: Apache Hadoop (MapReduce); Apache Spark.
- BD Visualization: Tableau, Power BI, Apache Superset.
- BD Cloud Services: Amazon Web Services (AWS).

Apache Hadoop was the most popular before Apache Spark was introduced. Studies show that the combination of the two frameworks is the best approach. Below is a brief breakdown of these two systems.

Apache Hadoop

Hadoop is a top-level Apache project that is written in Java language. It is a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system (Khan et al., 2014). This is possible due to its parallel clusters and distributed file system. Hadoop ensures fault tolerance and data replication to prevent data loss.

Features in Hadoop

According to Khan et al., (2014), Hadoop is made up of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka and the power of the Hadoop platform is based on a data storage layer and data processing layer, the Hadoop Distributed File System (HDFS) and the MapReduce framework.

Hadoop Distributed File System (HDFS)

HDFS is a data storage system that supports hundreds of nodes in a cluster and provides a cost-effective and reliable storage capability. Although HDFS can handle structured and unstructured terabytes of data, it is not a general-purpose file system. HDFS enhances performance by providing storage across various software systems, reducing network congestion and ensuring fault tolerance through data replication.

Hadoop MapReduce framework

MapReduce is a programming paradigm that enables mass scalability across numerous servers in a Hadoop cluster. MapReduce corresponds to two efficient and cost-effective jobs:

1. The first is the map job function that divides the dataset into independent data partitions constituting key/value pairs(tuples). The function assigns each partition to a unique compute node. The Mapper outputs intermediate key-value pairs, which are then collected, sorted, and grouped by key for further analysis.

2. The reduction task receives inputs from map outputs and divides the data into smaller sets of tuples.

Apache Spark

Apache Spark is a potent big data technology for large datasets due to its in-memory processing, distributed computing, resilience through rich data processing libraries, caching, cluster management, data source connectors, streaming capabilities, community support, and a growing ecosystem (Salloum et al., 2016).

Features of Apache Spark:

- Polyglot: Spark code can be written in Java, Python, Scala, and R since it provides APIs in these languages.
- Speed: Spark is 100 times faster than Hadoop MapReduce for large-scale data processing. The speed is achieved through controlled partitioning which enables parallel distributed data with less network traffic.
- Multiple formats: Spark supports a wide range of data sources, including Parquet, JSON, Hive, and Cassandra, in addition to common formats like text files, CSV and databases.
- Real-time computation: Spark excels in real-time, low-latency computation due to its in-memory processing.
- Machine Learning: Spark's MLlib simplifies big data machine learning by combining data processing and ML capabilities in a single, efficient, user-friendly engine, eliminating the need for multiple tools.

Using Spark with Hadoop:

Hadoop components can be used with Spark in the following ways:

- HDFS: Spark has the capability to operate on HDFS, allowing it to take advantage of the distributed and redundant storage system that HDFS provides.
- MapReduce: we can use Spark alongside MapReduce in the same Hadoop cluster or independently as a processing framework.
- YARN: Spark applications can be configured to operate on YARN, which is part of Hadoop's NextGen ecosystem.

- Batch and real-time processing: MapReduce and Spark work in conjunction, with MapReduce handling batch processing tasks and Spark handling real-time processing tasks. (Dayananda, 2020).

2.4.3 Big Data: Violence against women and girls

Big data technologies are increasingly employed in this field to analyse and visualise large volumes of data. The research subject and dataset determine the choice of big data technology; for this case study various technologies can be employed for data processing, analysis, and visualisation.

Data gathering can be difficult, there are various publicly accessible online data repositories available. Hadoop Distributed File System (HDFS) should have been an outstanding tool for storage; however, it was designed to deal with larger datasets. Nonetheless, the data has the potential to expand as more surveys are conducted (currently, just 70 countries have been polled). HDSF or other big data storage solutions (such as a highly scalable and fault-tolerant NoSQL database and MongoDB, a document-oriented database) may be considered. (2023).

After successfully storing the data, it is pre-processed to guarantee that it is clean and accurate. This includes deleting duplicates, filling in missing values, and standardising the data format (if necessary). Apache Hadoop or Apache Spark can be used for these tasks. After cleaning and pre-processing the data, it can be analysed to observe patterns and trends for business insights. Apache Spark is also used for carrying out complex data analysis. It can be utilised for machine learning and neural network processing if necessary. Other examples of big data tools used for data analytics include Apache Flink, an open-source stream processing framework, and Apache Kafka, a distributed streaming platform (Anon, 2023).

Data visualisation makes data simpler to comprehend and interpretable. Big data technologies aid in the creation of dynamic and powerful visualisations of complex data. Tableau and D3 are excellent big data visualisation tools that could be employed in this case (Anonymous (n.d.)). These tools help in telling interesting stories and will allow the user to go deeper into the data. Good visuals are critical in raising awareness and implementing targeted interventions in this field. Leveraging these technologies to analyse and visualise complex data, valuable insights can be obtained to inform policy decisions and tailor interventions aimed at preventing and addressing violence against women and girls.

2.5 Team contribution statement

This dynamic team of three worked together seamlessly and tirelessly with each member bringing a unique strength to ensure the successful completion of the project. We worked collaboratively as a group for the entirety of the task, nonetheless, below is a breakdown of our individual contributions:

Tasks	Constance Jumbo	Idongesit Okoko	Shibina Shajahan
Literature Review	✓	✓	✓
Dataset Description	✓	✓	✓
Data Pre-processing	✓	✓	✓
Data Exploration and Visualizations	✓	✓	✓
Bigdata Analysis Techniques uses	✓	✓	✓
Contribution of each team member	✓	✓	✓

Our teamwork was defined by strong values and open and honest communication. We supported and encouraged each other through every step of the work. Our commitment to achieving the objectives of the task is admirable. It is safe to say that each member of the team brought something unique to the table which has led to the successful completion of the tasks.

References

1. AUSTRAC, 2014. Money Laundering Methodologies. AUSTRAC, Commonwealth of Australia http://www.austrac.gov.au/typologies-2008-methodologies Accessed: 30/ 1/2017.
2. Ray, A. (2015) *Emerging solutions in anti-money laundering technology*, *Celent*. Available at: https://www.celent.com/insights/501182631 (Accessed: 23 August 2023).
3. Shukla, N., Hagenbuchner, M., Win, K.T. & Yang, J. (2018). *Breast cancer data analysis for survivability studies and prediction*. [Online]. 1 March 2018. Computer Methods and Programs in Biomedicine. Available from: https://doi.org/10.1016/j.cmpb.2017.12.011.
4. Wang, Q. & Wang, L. (2020). *Renewable energy consumption and economic growth in OECD countries: A nonlinear panel data analysis*. [Online]. 1 September 2020. Energy. Available from: https://doi.org/10.1016/j.energy.2020.118200.
5. Singh, K. and Best, P. (2019) 'Anti-money laundering: Using data visualization to identify suspicious activity', *International Journal of Accounting Information Systems*, 34, p. 100418. doi:10.1016/j.accinf.2019.06.001.
6. Singh, K., Best, P., 2016. Interactive visual analysis of anomalous accounts payable transactions in SAP enterprise systems. Manag. Audit. J. 31 (1), 35–63.
7. Von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., Van Wijk, J.J., Fekete, J.D., Fellner, D.W., 2011. Visual analysis of large graphs: state-of-the-art and future research challenges. Comput. Graphics Forum 30 (6), 1719–1749.
8. Anon (n.d.). *What Is Big Data?* [Online]. Oracle United Kingdom. Available from: https://www.oracle.com/uk/big-data/what-is-big-data/.
9. Ellars, S. (2019). *Big data volume, variety, velocity and veracity*. [Online]. 19 July 2019. insideBIGDATA. Available from: https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.
10. Khan, N., Yaqoob, I., Hashem, M., Inayat, Z., Ali, W.K.M., Alam, M.M., Shiraz, M. & Gani, A. (2014). *Big Data: Survey, Technologies, Opportunities, and Challenges*. [Online]. 1 January 2014. The Scientific World Journal. Available from: https://doi.org/10.1155/2014/712826.
11. Valchanov, I. (2023). *Traditional Data and Big Data Processing Techniques*. [Online]. 28 April 2023. 365 Data Science. Available from: https://365datascience.com/trending/techniques-for-processing-traditional-and-big-data/.
12. Abdalla, H.B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*. [Online]. 9 (1). Available from: http://dx.doi.org/10.1186/s40537-022-00659-3.
13. Philip Chen, C.L. & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. [Online]. 275. p.pp. 314–347. Available from: http://dx.doi.org/10.1016/j.ins.2014.01.015.
14. Allam, Z. & Dhunny, Z.A. (2019). On big data, artificial intelligence and smart cities. *Cities*. [Online]. 89. p.pp. 80–91. Available from: http://dx.doi.org/10.1016/j.cities.2019.01.032.
15. Hariri, R.H., Fredericks, E.M. & Bowers, K.M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*. [Online]. 6 (1). Available from: http://dx.doi.org/10.1186/s40537-019-0206-3.
16. Sinha, D.S. (2020). BIG DATA ANALYSIS: CONCEPTS, CHALLENGES AND OPPORTUNITIES. *International Journal of Innovative Research in Computer Science & Technology*. [Online]. 8 (3). Available from: http://dx.doi.org/10.21276/ijircst.2020.8.3.29.
17. Coursera (2023). *4 Types of Big Data Technologies (+ Management Tools)*. [Online]. 16 June 2023. Coursera. Available from: https://www.coursera.org/articles/big-data-technologies.
18. Anon (2023). *Top Big Data Technologies You Must Know [2023]*. [Online]. 27 March 2023. InterviewBit. Available from: https://www.interviewbit.com/blog/big-data-technologies/?amp=1.

19. Anon (n.d.). *Big Data Technologies that Everyone Should Know in 2023*. [Online]. Available from: https://www.knowledgehut.com/blog/big-data/big-data-technologies#what-is-the-future-of-big-data?
20. Salloum, S., Dautov, R., Chen, X. et al. Big data analytics on Apache Spark. Int J Data Sci Anal 1, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9
21. Dayananda, S. (2020). *Spark Tutorial: Real Time Cluster Computing Framework*. [Online]. 25 November 2020. Edureka. Available from: https://www.edureka.co/blog/spark-tutorial/#Why_Spark_When_Hadoop_Exists.
22. Dayananda, S. (2020). Spark Tutorial: Real Time Cluster Computing Framework. *Edureka*. https://www.edureka.co/blog/spark-tutorial/#What_Is_Apache_Spark .

Appendix

```
################# Prep-processing ################################

import pandas as pd
df = pd.read_csv("/content/drive/MyDrive/Violence Against Women  Girls
Data.csv")

# Read dataset and view information about dataset
df = pd.read_csv("/content/drive/MyDrive/Data Visualization/Violence
Against Women  Girls Data.csv")
df.info()
print("\n")

# Check for missing values in the dataset
df.isnull().sum()

#drop null values
df = df.dropna(subset = ["Value"])
df.info()

# Check for missing values in the dataset
df.isnull().sum()

# Check for duplicate values
duplicates = df[df.duplicated()]
print('Dulpicates in the dataframe\n',duplicates)

#Delete RecordID column from the dataset
df = df.drop(columns=["RecordID"])
df


######################## VISUAL 1 #################################

import plotly.express as px

# Calculating the mean value for each country
mean_values = df.groupby('Country')['Value'].mean().reset_index()

# Creating an interactive choropleth map
fig = px.choropleth(
    mean_values,
    locations="Country",
    locationmode="country names",
    color="Value",
    hover_name="Country",
    color_continuous_scale="Rainbow",
    title="Mean Violence Against Women by Country",
)

# Adjusting the title's position
fig.update_layout(
    title=dict(
        text="Mean Violence Against Women by Country",
        x=0.5,  y=0.95, len=1.0

    )
)

# Customizing the legend
fig.update_coloraxes(colorbar_title="Average Value")
fig.update_layout(
    coloraxis_colorbar=dict(
        tickformat=".2f",
        x=0.75, y=0.45, len=1.0,
    )
)

# Showing the interactive map
fig.show()


############################# VISUAL 2 ##############################

import matplotlib.pyplot as plt
# Calculating the average value of the 'Value' column for each country
average_values = df.groupby('Country')['Value'].mean()

# Sorting the average values in descending order and select the top 20
top_20_values = average_values.sort_values(ascending=False).head(30)

# Creating a scatter plot to visualize the sorted average values
plt.figure(figsize=(12, 6))
plt.scatter(top_20_values.index, top_20_values, marker='o',
color='purple', label='Average Value')

# Customizing the plot
plt.title('Top 30 Countries with the Highest Average Value')
plt.xlabel('Country')
plt.ylabel('Average Value of Violence')
plt.xticks(rotation=90)
plt.grid(True)
plt.yticks(range(5, 60, 5))

# Showing the plot
plt.tight_layout()
plt.show()


##################### VISUAL 3 ######################################

import matplotlib.pyplot as plt
import seaborn as sns

# Group the DataFrame by 'AbuseReason' and calculate the mean of
'Value' for each reason
mean_values = df.groupby('Question')['Value'].mean().reset_index()

# Sort the DataFrame by the average value in descending order
mean_values = mean_values.sort_values(by='Value', ascending=False)

# Create a bar chart to visualize the average values for each reason
plt.figure(figsize=(10, 6))
sns.barplot(x='Value', y='Question', data=mean_values,
palette='inferno')
plt.title('Average Values for Abuse Reasons')
plt.xlabel('Average Value')
plt.ylabel('Abuse Reason')

# Show the plot
plt.tight_layout()
plt.show()


############################ VISUAL 4 ##############################

# Grouping the DataFrame by 'Demographics Question' and 'Demographics
Response' and calculate the mean of 'Value'
heatmap_data = df.groupby(['Demographics Question', 'Demographics
Response'])['Value'].mean().reset_index()

# Pivot the DataFrame to create a matrix for the heatmap
heatmap_data = heatmap_data.pivot_table(
    index='Demographics Question', columns='Demographics Response',
values='Value', fill_value=0
)

# Create the heatmap using Seaborn
plt.figure(figsize=(10, 5))
sns.heatmap(heatmap_data, annot=True, cmap='rainbow', fmt='.2f',
linewidths=0.5)
plt.title('Heatmap of Average Values for Abuse Reasons by
Demographics')
plt.xlabel('Demographic Response')
plt.ylabel('Demographic Question')


# Show the plot
plt.show()


######################## VISUAL 5 ####################################

# Calculate the average value of the 'Value' column for each question
and gender
average_values = df.groupby(['Question',
'Gender'])['Value'].mean().unstack()

# Define custom colors for male and female bars
colors = {'M': '#140b34', 'F': 'purple'}

# Create a grouped bar chart
plt.figure(figsize=(8, 6))

# Plot the grouped bar chart for male and female responses with custom
colors
bar_width = 0.4
index = range(len(average_values.index))
for gender, color in colors.items():
    plt.bar(
        [i + (bar_width if gender == 'M' else 2 * bar_width) for i in
index],
        average_values[gender],
        width=bar_width,
        label=gender,
        align='center',
        alpha=0.7,
        color=color
    )

# Customize the plot
plt.title('Average Values by Question and Gender')
plt.xlabel('Question')
plt.ylabel('Average Value')
plt.xticks([i + bar_width for i in index], average_values.index,
rotation=45, ha='right')
plt.legend(title='Gender')

# Show the plot
plt.tight_layout()
plt.show()
```

```
########################## VISUAL 6 ################################

education_data = df[df['Demographics Question'] == 'Education']
plt.figure(figsize=(6, 6))
sns.violinplot(
    data=education_data,
    x='Demographics Response',
    y='Value',
    hue='Gender',
    palette={'M': '#ffb400', 'F': '#9080ff'},
    split=True
)
plt.title('Comparison of Education Categories by Gender (Violin Plot)')
plt.xlabel('Education Category')
plt.ylabel('Value')
plt.xticks(rotation=45, ha='right')
plt.legend(title='Gender')
plt.tight_layout()
plt.show()


################################# VISUAL 7 ########################

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Assuming you have loaded your data into a DataFrame called 'df'

# Group by Country, Category, and Subcategory and calculate the average
value
grouped_data = df.groupby(['Country', 'Demographics Question',
'Demographics Response'])['Value'].mean().reset_index()

# Find the subcategory with the highest average value within each
category and country
top_subcategories = grouped_data.groupby(['Demographics Question',
'Country'])['Value'].idxmax()
top_subcategories_data = grouped_data.loc[top_subcategories]

# Select the top 10 countries with the highest average values
top_countries =
top_subcategories_data.groupby('Country')['Value'].mean().nlargest(5).i
ndex

# Filter the data to include only the top 10 countries
```

```
filtered_data =
top_subcategories_data[top_subcategories_data['Country'].isin(top_count
ries)]

# Pivot the data for plotting
pivot_data = filtered_data.pivot(index='Country', columns='Demographics
Response', values='Value')

# Normalize the data to create a 100% stacked bar
pivot_data = pivot_data.div(pivot_data.sum(axis=1), axis=0)
new_order = ['15-24', '25-34', '35-49', 'No education', 'Primary',
'Secondary',
             'Higher', 'Never married', 'Widowed, divorced, separated',
             'Married or living together', 'Employed for kind',
'Employed for cash', 'Unemployed',
             'Rural', 'Urban']
pivot_data = pivot_data.reindex(columns=new_order)

# Define a custom color mapping dictionary for our categories
category_colors = {
        '15-24': '#88CAF8',
        '25-34': '#0974FC',
        '35-49': '#0703F9',
        'No education': '#FAB9B9',
        'Primary': '#F96666',
        'Secondary': '#C70D0D',
        'Higher': '#680404',
        'Never married': '#272E89',
        'Widowed, divorced, separated': '#505AD3',
        'Married or living together': '#8E95F0',
        'Employed for kind': '#D8C821',
        'Employed for cash': '#A4981E',
        'Unemployed': '#5E5712',
        'Rural': '#2E2B28',
        'Urban': '#808080',

}

# Create a 100% stacked bar chart with custom colors
plt.figure(figsize=(12, 8))
ax = pivot_data.plot(kind='bar', ax=plt.gca(), stacked=True,
color=[category_colors.get(col, 'gray') for col in pivot_data.columns])

# Annotate the bars with subcategory labels
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate(f'{height:.2%}', (x + width/2, y + height/2),
ha='center', va='center', fontsize=8, color='white')

# Customize the plot
plt.title('Top 5 Countries with Highest Avg Values by Subcategory')
plt.xlabel('Country')
plt.ylabel('Percentage')
plt.xticks(rotation=45)

# Move the legend to the far right
plt.legend(
    title='Demographics Response',
    title_fontsize=12,
    loc='center left',
    bbox_to_anchor=(1, 0.70), fontsize=10,
)

# Show the plot
plt.tight_layout()
plt.show()
```