Configuration Manual

MSc Research Project
Msc FinTech

Adedoyin Fatobi
Student ID: x23111569

School of Computing
National College of Ireland

Supervisors: Noel Cosgrave and Faithful Chiagoziem Onwuegbuche

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. …………Adedoyin Fatobi……………………………………………

Student ID: …………x23111569……………………………………………………………..……

Programme: …………Msc in FinTech……………………… **Year:** ……2023/2024………..

Module: ………………Msc Research Project……………………………..………

Lecturer: ……………Noel Cosgrave and Faithful Chiagoziem Onwuegbuche …………

Submission Due Date: ……………………… September 16, 2024…………………………………………..………

Project Title: Machine Learning for Financial Inclusion and Safety: Empowering Women Against Violence……………….………

Word Count: …………2531……… **Page Count:** ……………19……….………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………………………Adedoyin Fatobi…………………………………………………

Date: ……………………September 16, 2024……………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Adedoyin Fatobi
Student ID: x23111569

1 Introduction

My research work on the topic "*Machine Learning for Financial Inclusion and Safety: Empowering Women Against Violence*" is submitted as part of the requirements for the completion of the MSc Fintech module. This configuration manual adds to the requirement for completion. The configuration manual will detail the steps involved in the execution of the study including the technologies and hardware configuration of the devices used. It would also aim to provide specifications for future research that may guide other research endeavours in the future ensuring reproducibility.

2 System Specification

2.1 Hardware

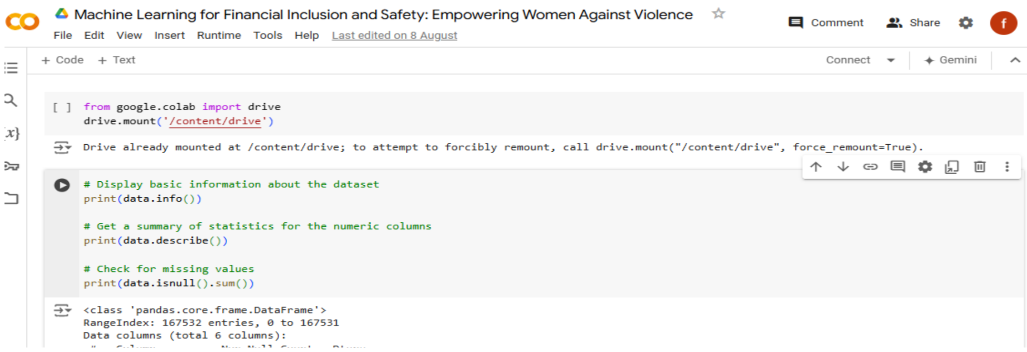
The research project was conducted on a Windows 11 pro computer:

- Operating System: Microsoft Windows 11 Home Single Language
- Computer Model: HP Pavilion x360
- Processor: 12th Gen Intel® Core (TM) i5-1235U, 1300 MHz, 10 Cores, 12 Logical Processors
- RAM: 16.0 GB

2.2 Software and Tools

- **Google Drive:** This is a cloud storage device that allows users to store files online, enabling access from any device.
- **Google Colab:** This is a cloud-based platform that allows for the writing and execution of codes written in Python. The platform is also interactive with access to GPUs and allows for note-sharing or collaboration. This platform also allows for seamless integration with Google Drive where raw data for analysis can be stored for analysis or where processed data in the course of the analysis can be stored for easy retrieval in the course of the analysis.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

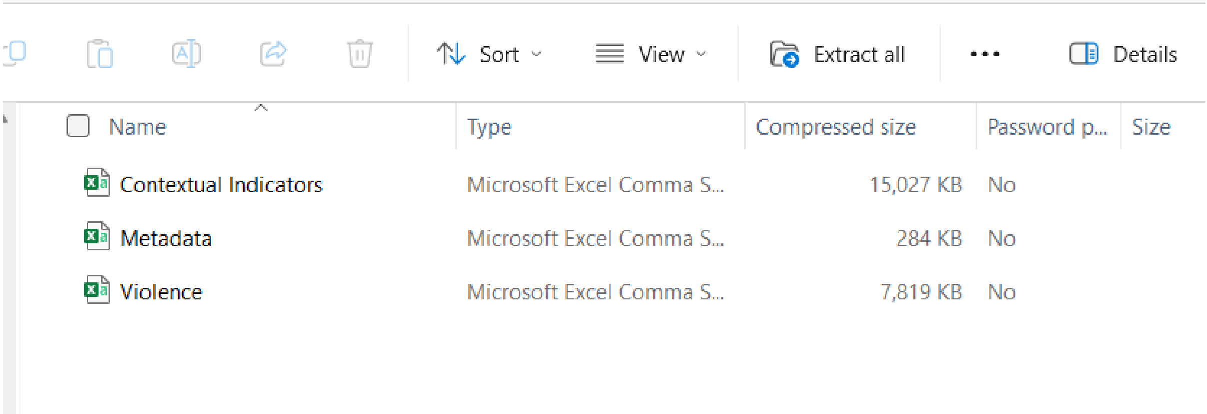

Machine Learning for Financial Inclusion and Safety: Empowering Women Against Violence
Comment
Share
File Edit View Insert Runtime Tools Help Last edited on 8 August
+ Code + Text
Connect
Gemini

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

    # Display basic information about the dataset
    print(data.info())

    # Get a summary of statistics for the numeric columns
    print(data.describe())

    # Check for missing values
    print(data.isnull().sum())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167532 entries, 0 to 167531
Data columns (total 6 columns):
```

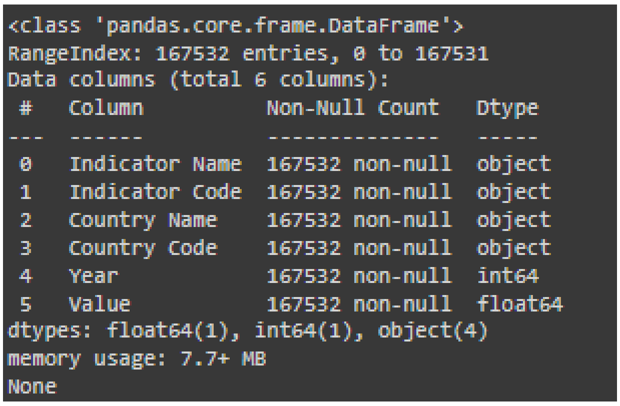
- **Microsoft Office LTSC:** This provided Excel which stores the raw data deployed for use in the Python environment on Google Colab and Word which is used for the typesetting of reports.

3 Data Source

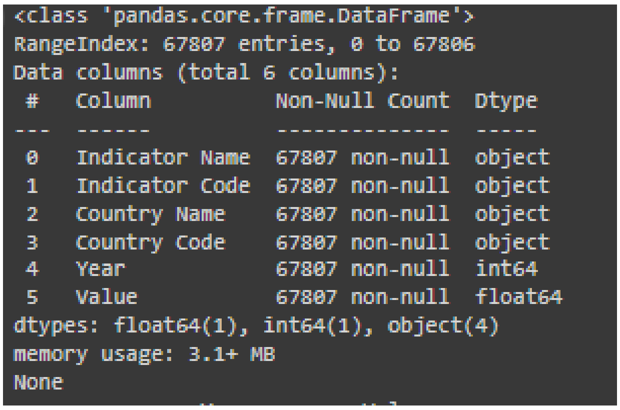
The data for the study was from the World Bank and Demographic and Health Survey. Two datasets were employed contextual_indicators.csv, and violence.csv. The contextual indicators dataset contains indicators on digital financial inclusion across the globe, while the violence dataset also contains violence indicators around the world. The description of the datasets are shown below:



Name	Type	Compressed size	Password p...	Size
Contextual Indicators	Microsoft Excel Comma S...	15,027 KB	No	
Metadata	Microsoft Excel Comma S...	284 KB	No	
Violence	Microsoft Excel Comma S...	7,819 KB	No	



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167532 entries, 0 to 167531
Data columns (total 6 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   Indicator Name  167532 non-null  object 
 1   Indicator Code  167532 non-null  object 
 2   Country Name    167532 non-null  object 
 3   Country Code    167532 non-null  object 
 4   Year            167532 non-null  int64  
 5   Value           167532 non-null  float64
dtypes: float64(1), int64(1), object(4)
memory usage: 7.7+ MB
None
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67807 entries, 0 to 67806
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Indicator Name  67807 non-null  object 
 1   Indicator Code  67807 non-null  object 
 2   Country Name    67807 non-null  object 
 3   Country Code    67807 non-null  object 
 4   Year            67807 non-null  int64  
 5   Value           67807 non-null  float64
dtypes: float64(1), int64(1), object(4)
memory usage: 3.1+ MB
None
```

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

4 Data Cleaning and Pre-processing

The data analysis was in two main stages – exploratory data analysis, and several other processes involving, statistical analysis, and machine learning (PCA) to derive meaningful insights from the datasets. The entire analysis was focused on understanding financial exclusion and vulnerability across countries using indicators derived from contextual and violence-related data.

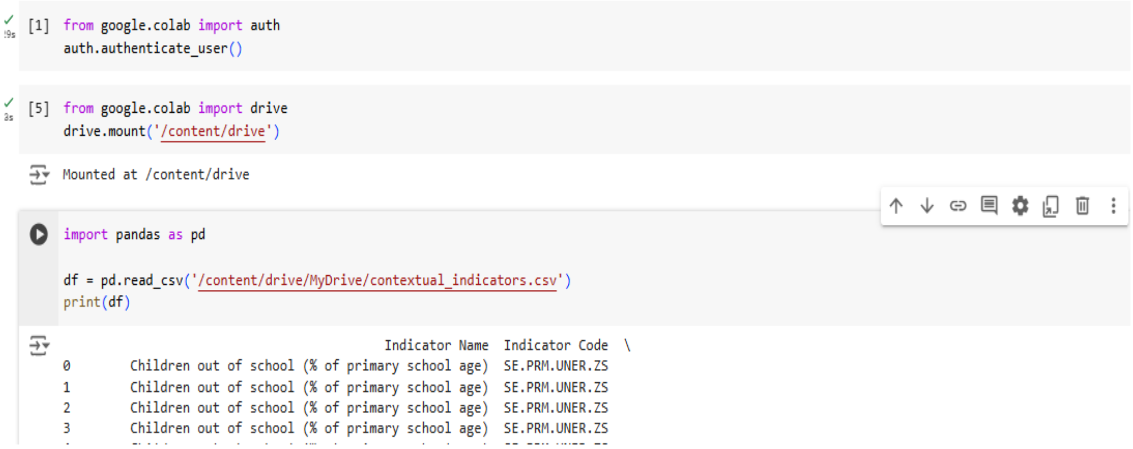
5 Data Techniques

5.1 Exploratory Data Analysis (Individual Analysis of Contextual and Violence-Related Data)

5.1.1 Financial Inclusion

This is the first part of the exploratory data analysis and it focusses on digital financial inclusion indices within the data.

From the start, we initiate the EDA process by mounting Google Drive to access the contextual_indicators.csv, which is loaded into a panda DataFrame. In the initial part, the data was verified with the display of the first few rows to ensure that the data has been read correctly.



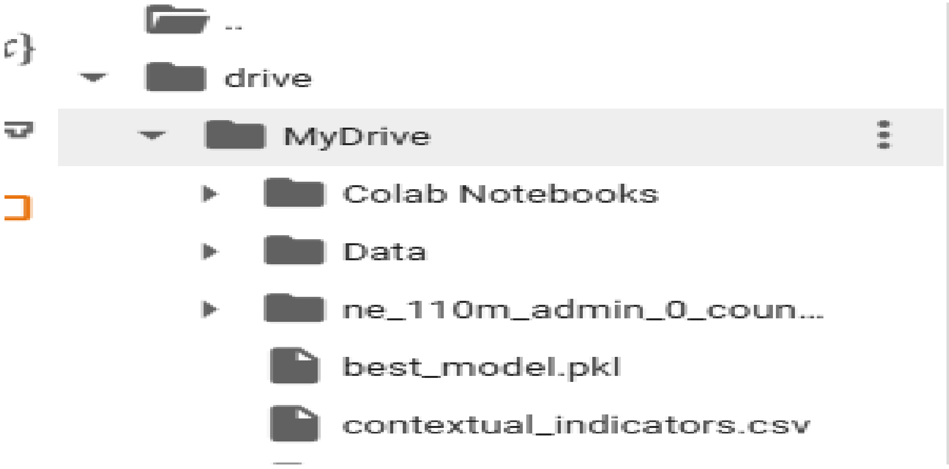
```
[1] from google.colab import auth
    auth.authenticate_user()
```

```
[5] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd

df = pd.read_csv('/content/drive/MyDrive/contextual_indicators.csv')
print(df)
```

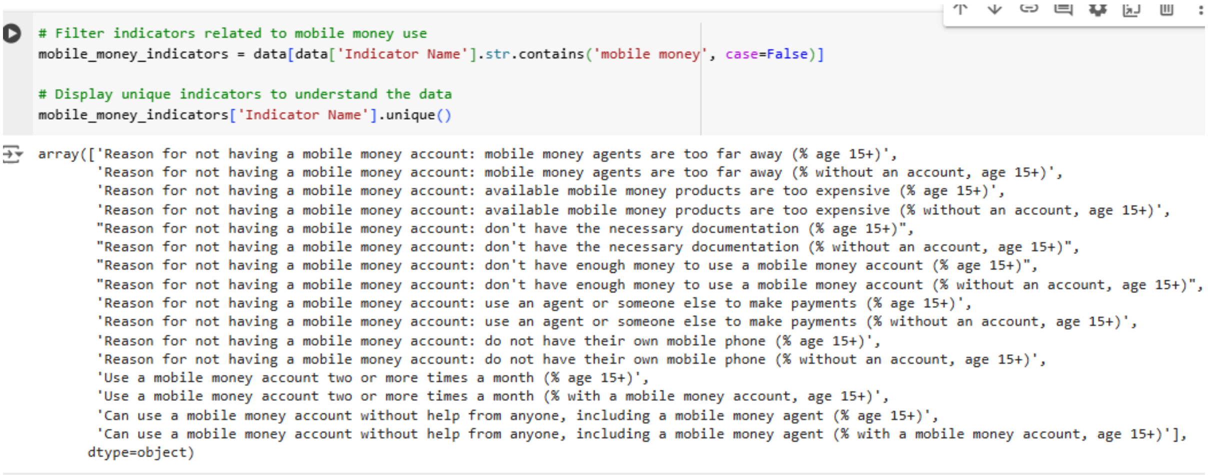
```
                              Indicator Name  Indicator Code  \
0       Children out of school (% of primary school age)  SE.PRM.UNER.ZS
1       Children out of school (% of primary school age)  SE.PRM.UNER.ZS
2       Children out of school (% of primary school age)  SE.PRM.UNER.ZS
3       Children out of school (% of primary school age)  SE.PRM.UNER.ZS
```



https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Focus on Mobile Money Indicators.

The analysis narrows down to focus on indicators related to mobile money usage. Using a filter for strings containing "mobile money," the script identifies and displays unique indicators that match this criterion. This step helps in understanding the range of mobile money-related data available within the dataset.

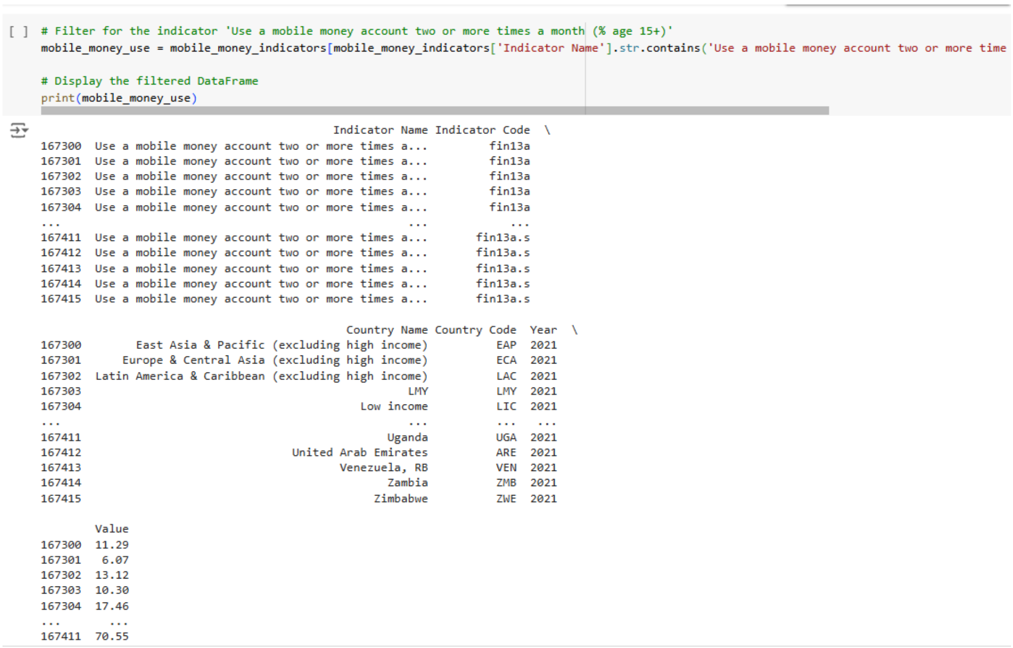


```
# Filter indicators related to mobile money use
mobile_money_indicators = data[data['Indicator Name'].str.contains('mobile money', case=False)]

# Display unique indicators to understand the data
mobile_money_indicators['Indicator Name'].unique()
```

```
array(['Reason for not having a mobile money account: mobile money agents are too far away (% age 15+)',
       'Reason for not having a mobile money account: mobile money agents are too far away (% without an account, age 15+)',
       'Reason for not having a mobile money account: available mobile money products are too expensive (% age 15+)',
       'Reason for not having a mobile money account: available mobile money products are too expensive (% without an account, age 15+)',
       "Reason for not having a mobile money account: don't have the necessary documentation (% age 15+)",
       "Reason for not having a mobile money account: don't have the necessary documentation (% without an account, age 15+)",
       "Reason for not having a mobile money account: don't have enough money to use a mobile money account (% age 15+)",
       "Reason for not having a mobile money account: don't have enough money to use a mobile money account (% without an account, age 15+)",
       'Reason for not having a mobile money account: use an agent or someone else to make payments (% age 15+)',
       'Reason for not having a mobile money account: use an agent or someone else to make payments (% without an account, age 15+)',
       'Reason for not having a mobile money account: do not have their own mobile phone (% age 15+)',
       'Reason for not having a mobile money account: do not have their own mobile phone (% without an account, age 15+)',
       'Use a mobile money account two or more times a month (% age 15+)',
       'Use a mobile money account two or more times a month (% with a mobile money account, age 15+)',
       'Can use a mobile money account without help from anyone, including a mobile money agent (% age 15+)',
       'Can use a mobile money account without help from anyone, including a mobile money agent (% with a mobile money account, age 15+)'],
      dtype=object)
```

Detailed Analysis on Specific Indicators.

Further, the script pinpoints a specific indicator – "Use a mobile money account two or more times a month (% age 15+)." It filters the dataset accordingly and converts the 'Year' column to an integer type for consistent filtering. This filtered data is then displayed to ensure accuracy.



```
# Filter for the indicator 'Use a mobile money account two or more times a month (% age 15+)'
mobile_money_use = mobile_money_indicators[mobile_money_indicators['Indicator Name'].str.contains('Use a mobile money account two or more time

# Display the filtered DataFrame
print(mobile_money_use)
```

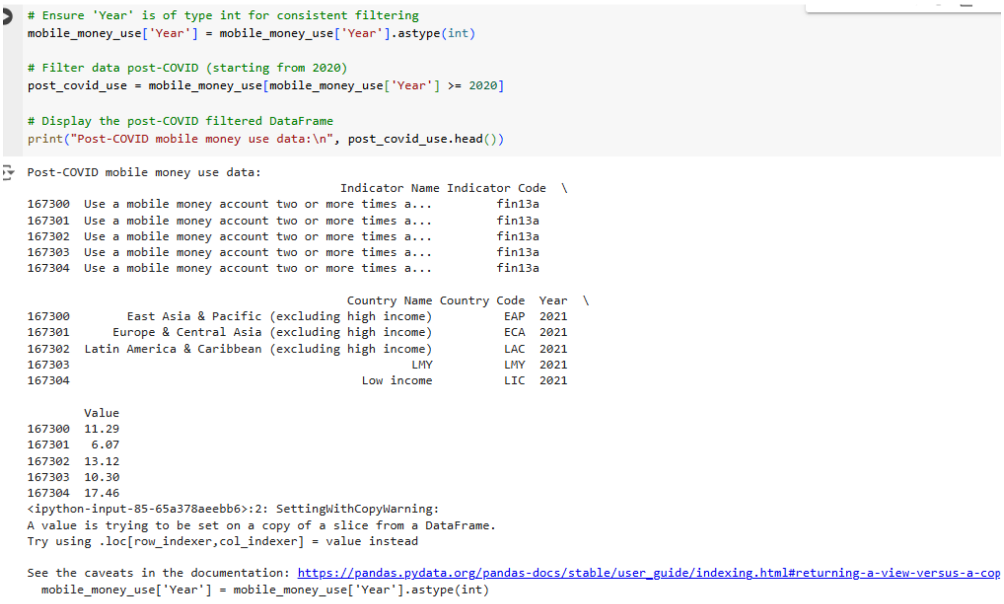
```
                                         Indicator Name Indicator Code  \
167300  Use a mobile money account two or more times a...        fin13a
167301  Use a mobile money account two or more times a...        fin13a
167302  Use a mobile money account two or more times a...        fin13a
167303  Use a mobile money account two or more times a...        fin13a
167304  Use a mobile money account two or more times a...        fin13a
...                                                   ...           ...
167411  Use a mobile money account two or more times a...      fin13a.s
167412  Use a mobile money account two or more times a...      fin13a.s
167413  Use a mobile money account two or more times a...      fin13a.s
167414  Use a mobile money account two or more times a...      fin13a.s
167415  Use a mobile money account two or more times a...      fin13a.s

                                         Country Name Country Code  Year  \
167300        East Asia & Pacific (excluding high income)          EAP  2021
167301      Europe & Central Asia (excluding high income)          ECA  2021
167302  Latin America & Caribbean (excluding high income)          LAC  2021
167303                                                LMY          LMY  2021
167304                                         Low income          LIC  2021
...                                                   ...          ...   ...
167411                                             Uganda          UGA  2021
167412                               United Arab Emirates          ARE  2021
167413                                      Venezuela, RB          VEN  2021
167414                                             Zambia          ZMB  2021
167415                                           Zimbabwe          ZWE  2021

        Value
167300  11.29
167301   6.07
167302  13.12
167303  10.30
167304  17.46
...       ...
167411  70.55
```

Post-COVID Trends Analysis.

Post-2020 data is extracted to analyze trends in mobile money usage after the onset of the COVID-19 pandemic. The script groups this data by country and calculates average values to observe country-specific trends in mobile money usage, which are displayed and visualized

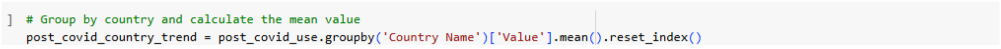
https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

using a bar graph. This visualization helps in quickly identifying countries with higher average usage rates.

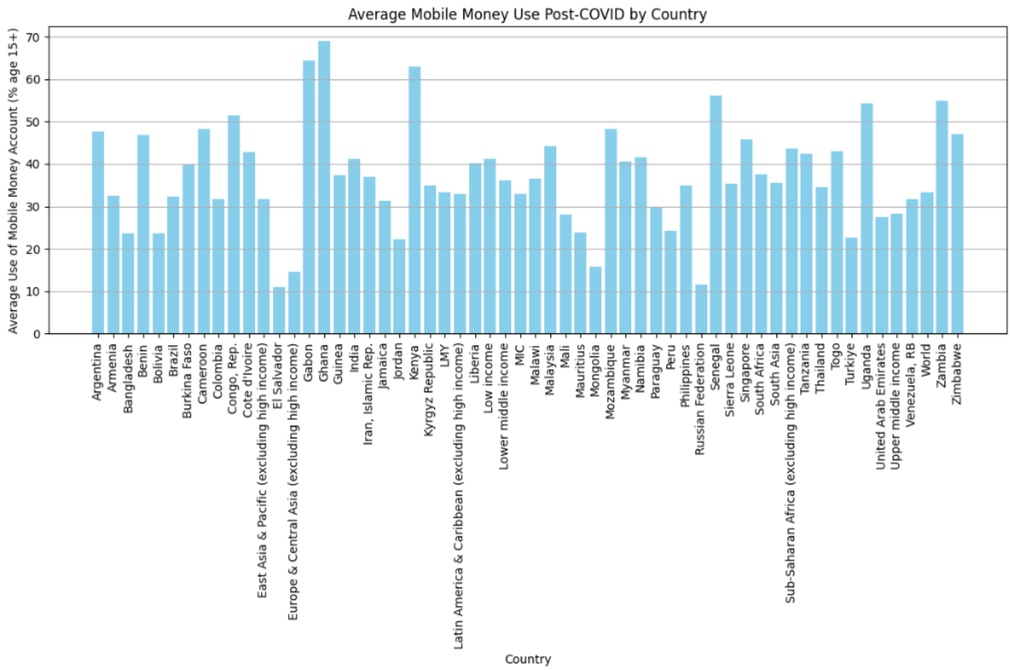


```
# Ensure 'Year' is of type int for consistent filtering
mobile_money_use['Year'] = mobile_money_use['Year'].astype(int)

# Filter data post-COVID (starting from 2020)
post_covid_use = mobile_money_use[mobile_money_use['Year'] >= 2020]

# Display the post-COVID filtered DataFrame
print("Post-COVID mobile money use data:\n", post_covid_use.head())
```

```
Post-COVID mobile money use data:
                                              Indicator Name Indicator Code  \
167300  Use a mobile money account two or more times a...         fin13a
167301  Use a mobile money account two or more times a...         fin13a
167302  Use a mobile money account two or more times a...         fin13a
167303  Use a mobile money account two or more times a...         fin13a
167304  Use a mobile money account two or more times a...         fin13a

                                        Country Name Country Code  Year  \
167300        East Asia & Pacific (excluding high income)          EAP  2021
167301      Europe & Central Asia (excluding high income)          ECA  2021
167302  Latin America & Caribbean (excluding high income)          LAC  2021
167303                                                LMY          LMY  2021
167304                                         Low income          LIC  2021

        Value
167300  11.29
167301   6.07
167302  13.12
167303  10.30
167304  17.46
<ipython-input-85-65a378aeebb6>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
  mobile_money_use['Year'] = mobile_money_use['Year'].astype(int)
```



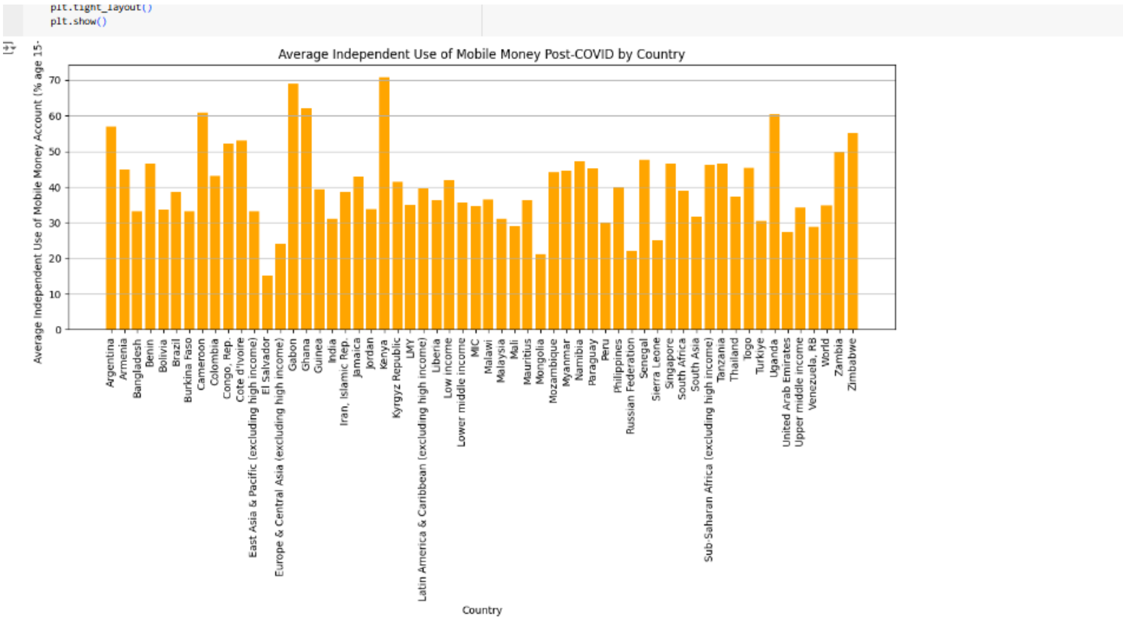
```
# Group by country and calculate the mean value
post_covid_country_trend = post_covid_use.groupby('Country Name')['Value'].mean().reset_index()
```



Independent Use of Mobile Money.

An additional layer of analysis focuses on the independent use of mobile money accounts, filtering for an indicator related to users operating their accounts without assistance. Similar to previous steps, the data is filtered post-2020, grouped by country, and visualized to highlight differences in independent usage across countries.

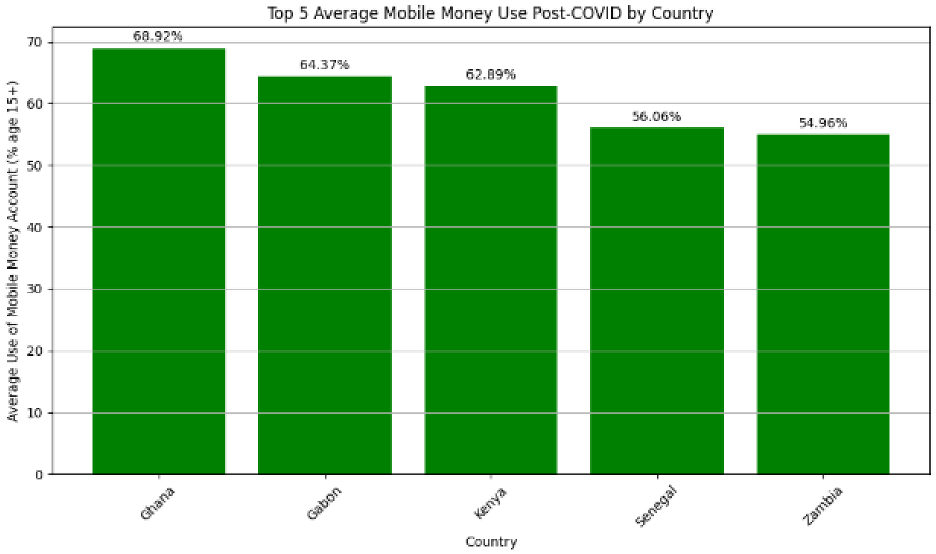
https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

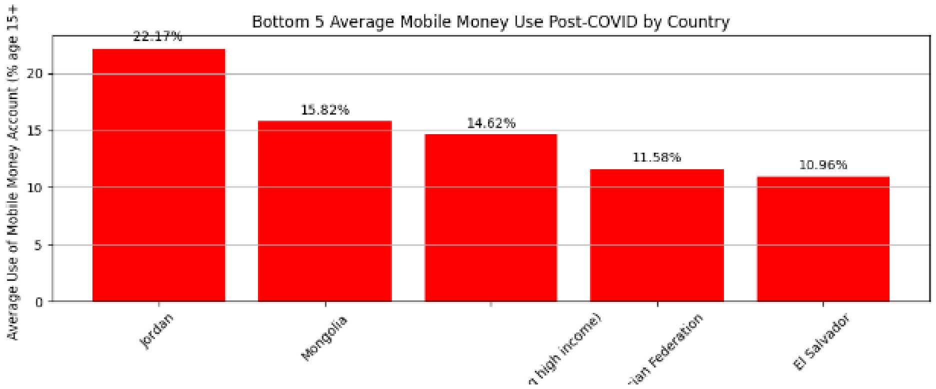


```
plt.tight_layout()
plt.show()
```

Comparative Analysis.

The script sorts the data on independent mobile money usage and identifies the top and bottom countries based on usage rates. It then visualizes these comparisons using bar charts, incorporating value labels for clarity. These visualizations are crucial for stakeholders to easily discern which countries are leading or lagging in mobile money adoption.





https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

5.1.2 Violence Indicators

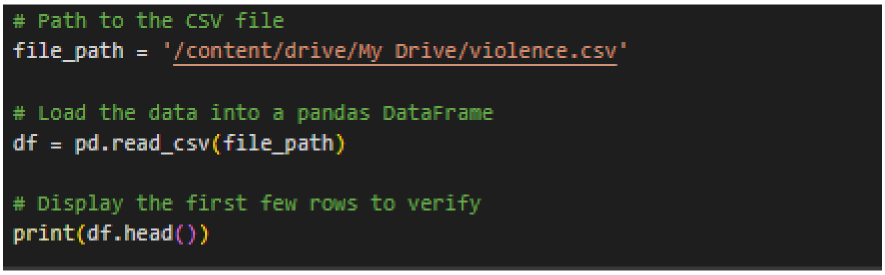
Here we, analyse and interpret violence-related data across countries. The analysis focuses specifically on indicators of various forms of violence across different countries. It uses libraries like pandas for data manipulation, seaborn and matplotlib for data visualization, and machine learning tools for predictive modeling.

Setup and Initial Data Loading

We import necessary libraries (pandas, matplotlib, seaborn) and sets the aesthetic style for seaborn plots.

Data Loading:

We read the CSV file containing violence-related data into a panda DataFrame and print out the first few rows to verify correct loading.



```
# Path to the CSV file
file_path = '/content/drive/My Drive/violence.csv'

# Load the data into a pandas DataFrame
df = pd.read_csv(file_path)

# Display the first few rows to verify
print(df.head())
```

Preliminary Data Exploration

Basic Information: this displays basic information and statistical summaries of the dataset, including checks for missing values.

Unique Values Exploration

Here, we print the number of unique values and samples of unique indicators and countries, which helps understand the diversity and scope of the dataset.

Data Filtering and Preparation

Keyword Filtering: we define a list of keywords related to various forms of violence and filter the DataFrame to include only those indicators that match the keywords.

Data Pivoting: here, we transform the filtered data into a pivot table format suitable for analysis, where each row represents a country and year combination, and columns represent different violence indicators.

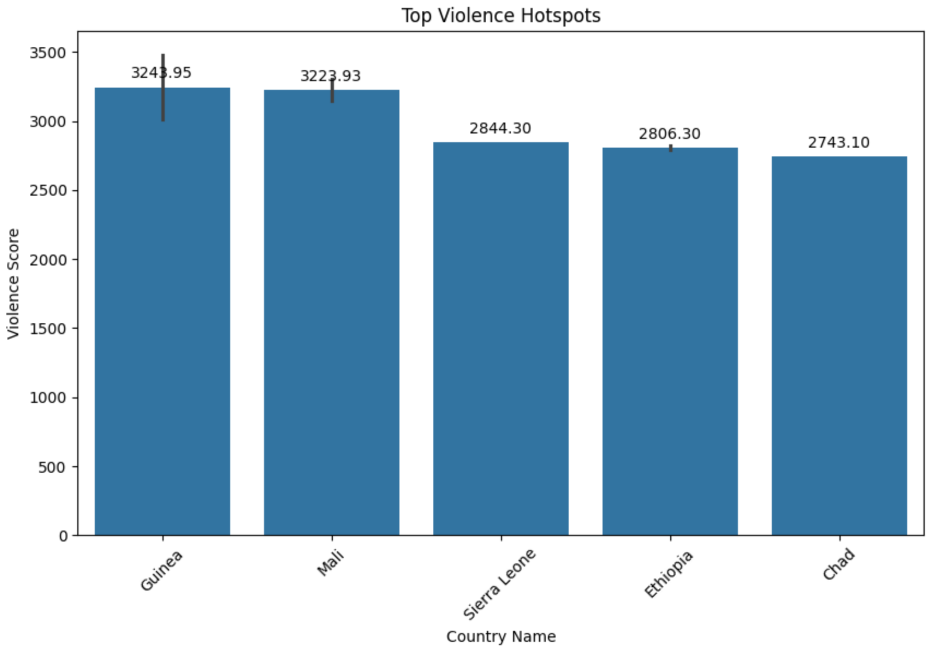
Feature Engineering

Violence Score Calculation: here, we create a new column called 'Violence Score' by summing up all indicator columns for each row, representing a cumulative measure of reported violence instances.

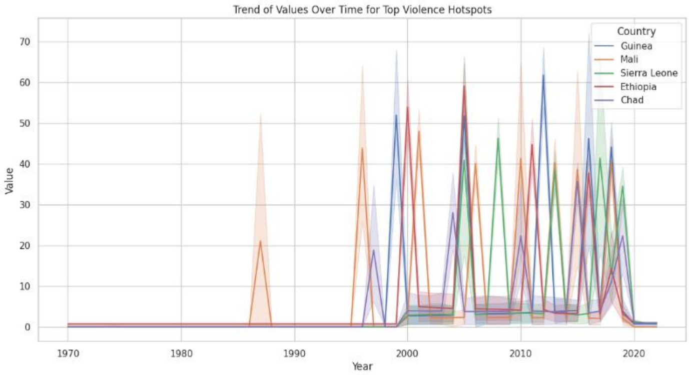
https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

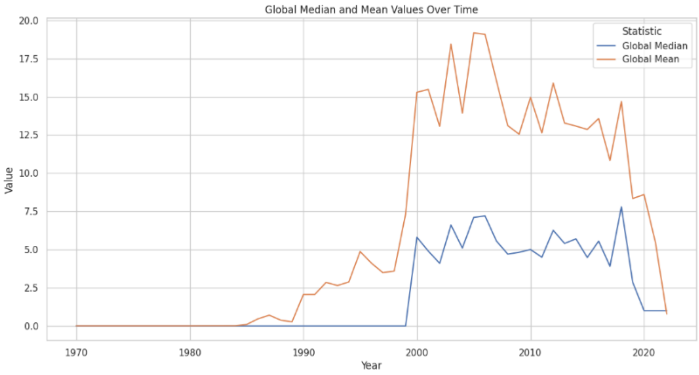
Exploratory Data Analysis (EDA)

Plotting Hotspots: we identify and visualise the top countries with the highest violence scores using a bar plot, helping to pinpoint regions with severe issues.



Trend Analysis: we generate line plots to display the global median and mean values of violence over years, and trends of violence metrics over time for specific countries identified as important.



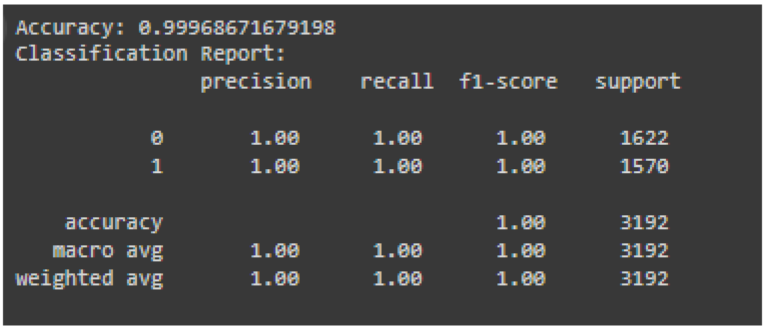


Predictive Modeling

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Model Building: we construct a predictive model to forecast levels of violence. It involves scaling the features, splitting the data into training and testing sets, fitting a RandomForest classifier, and evaluating the model performance using accuracy and a classification report.

The following results were obtained:



```
Accuracy: 0.99968671679198
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00      1622
           1       1.00      1.00      1.00      1570

    accuracy                           1.00      3192
   macro avg       1.00      1.00      1.00      3192
weighted avg       1.00      1.00      1.00      3192
```

5.2 Analysis of Integrated Contextual and Violence-Related Data

The procedures involved included data preparation and transformation steps and creating an analysis of the final indices (Financial Exclusion Index and Vulnerability Index) across countries and over time

5.2.1 Stage 1

This stage involves an integration of data handling, statistical analysis, and machine learning (PCA) to derive meaningful insights from the data. This was applied due to the complex nature of the data. By complex, it means that there was a broad dimensionality in the data. PCA was applied to reduce the dimensionality of the data in a way that would retain most of the variance. This process further broken down into the following steps:

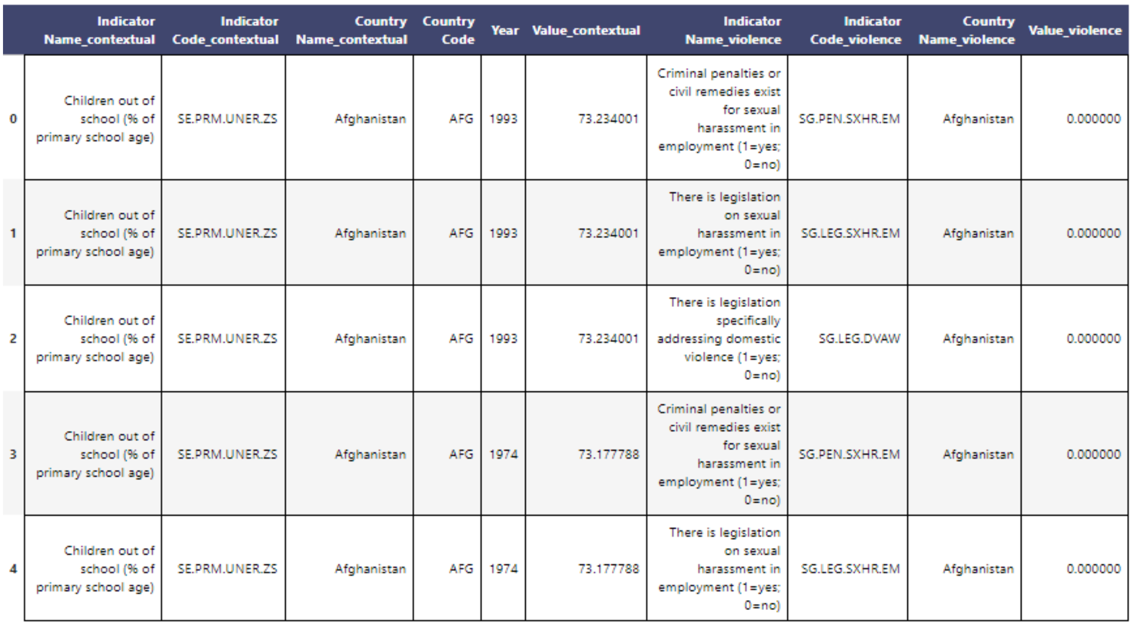
Data Loading

The CSV files containing contextual indicators and violence data are loaded into pandas DataFrame.

Data Exploration

Initial Display: The script prints the first few rows of both datasets to verify the correct loading and to provide a glimpse of the data structure.

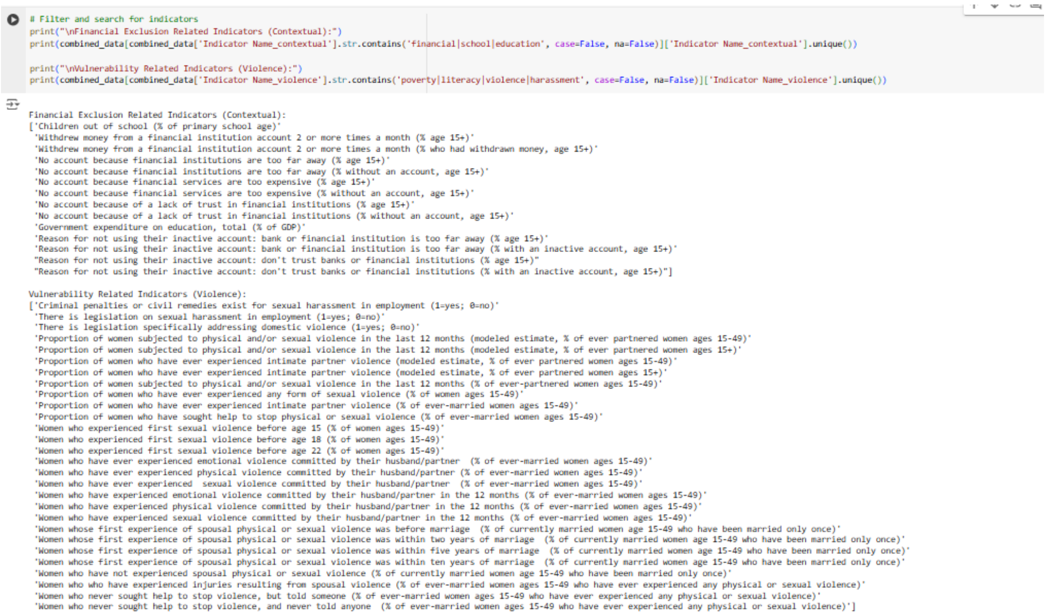
https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm



	Indicator Name_contextual	Indicator Code_contextual	Country Name_contextual	Country Code	Year	Value_contextual	Indicator Name_violence	Indicator Code_violence	Country Name_violence	Value_violence
0	Children out of school (% of primary school age)	SE.PRM.UNER.ZS	Afghanistan	AFG	1993	73.234001	Criminal penalties or civil remedies exist for sexual harassment in employment (1=yes; 0=no)	SG.PEN.SXHR.EM	Afghanistan	0.000000
1	Children out of school (% of primary school age)	SE.PRM.UNER.ZS	Afghanistan	AFG	1993	73.234001	There is legislation on sexual harassment in employment (1=yes; 0=no)	SG.LEG.SXHR.EM	Afghanistan	0.000000
2	Children out of school (% of primary school age)	SE.PRM.UNER.ZS	Afghanistan	AFG	1993	73.234001	There is legislation specifically addressing domestic violence (1=yes; 0=no)	SG.LEG.DVAW	Afghanistan	0.000000
3	Children out of school (% of primary school age)	SE.PRM.UNER.ZS	Afghanistan	AFG	1974	73.177788	Criminal penalties or civil remedies exist for sexual harassment in employment (1=yes; 0=no)	SG.PEN.SXHR.EM	Afghanistan	0.000000
4	Children out of school (% of primary school age)	SE.PRM.UNER.ZS	Afghanistan	AFG	1974	73.177788	There is legislation on sexual harassment in employment (1=yes; 0=no)	SG.LEG.SXHR.EM	Afghanistan	0.000000

Data Filtering

In this part, we search for specific indicators related to financial exclusion and vulnerability within the contextual and violence datasets, respectively. We filter indicators that contain keywords like 'financial', 'school', 'education', 'poverty', 'literacy', 'violence', and 'harassment'.



```
# Filter and search for indicators
print("\nFinancial Exclusion Related Indicators (Contextual):")
print(combined_data[combined_data['Indicator Name_contextual'].str.contains('financial|school|education', case=False, na=False)]['Indicator Name_contextual'].unique())

print("\nVulnerability Related Indicators (Violence):")
print(combined_data[combined_data['Indicator Name_violence'].str.contains('poverty|literacy|violence|harassment', case=False, na=False)]['Indicator Name_violence'].unique())
```

```
Financial Exclusion Related Indicators (Contextual):
['Children out of school (% of primary school age)'
 'Withdrew money from a financial institution account 2 or more times a month (% age 15+)'
 'Withdrew money from a financial institution account 2 or more times a month (% who had withdrawn money, age 15+)'
 'No account because financial institutions are too far away (% age 15+)'
 'No account because financial institutions are too far away (% without an account, age 15+)'
 'No account because financial services are too expensive (% age 15+)'
 'No account because financial services are too expensive (% without an account, age 15+)'
 'No account because of a lack of trust in financial institutions (% age 15+)'
 'No account because of a lack of trust in financial institutions (% without an account, age 15+)'
 'Government expenditure on education, total (% of GDP)'
 'Reason for not using their inactive account: bank or financial institution is too far away (% age 15+)'
 'Reason for not using their inactive account: bank or financial institution is too far away (% with an inactive account, age 15+)'
 "Reason for not using their inactive account: don't trust banks or financial institutions (% age 15+)"
 "Reason for not using their inactive account: don't trust banks or financial institutions (% with an inactive account, age 15+)"]

Vulnerability Related Indicators (Violence):
['Criminal penalties or civil remedies exist for sexual harassment in employment (1=yes; 0=no)'
 'There is legislation on sexual harassment in employment (1=yes; 0=no)'
 'There is legislation specifically addressing domestic violence (1=yes; 0=no)'
 'Proportion of women subjected to physical and/or sexual violence in the last 12 months (modeled estimate, % of ever partnered women ages 15-49)'
 'Proportion of women subjected to physical and/or sexual violence in the last 12 months (modeled estimate, % of ever partnered women ages 15+)'
 'Proportion of women who have ever experienced intimate partner violence (modeled estimate, % of ever partnered women ages 15-49)'
 'Proportion of women who have ever experienced intimate partner violence (modeled estimate, % of ever partnered women ages 15+)'
 'Proportion of women subjected to physical and/or sexual violence in the last 12 months (% of ever-partnered women ages 15-49)'
 'Proportion of women who have ever experienced any form of sexual violence (% of women ages 15-49)'
 'Proportion of women who have ever experienced intimate partner violence (% of ever-married women ages 15-49)'
 'Proportion of women who have sought help to stop physical or sexual violence (% of ever-married women ages 15-49)'
 'Women who experienced first sexual violence before age 15 (% of women ages 15-49)'
 'Women who experienced first sexual violence before age 18 (% of women ages 15-49)'
 'Women who experienced first sexual violence before age 22 (% of women ages 15-49)'
 'Women who have ever experienced emotional violence committed by their husband/partner  (% of ever-married women ages 15-49)'
 'Women who have ever experienced physical violence committed by their husband/partner (% of ever-married women ages 15-49)'
 'Women who have ever experienced  sexual violence committed by their husband/partner  (% of ever-married women ages 15-49)'
 'Women who have experienced emotional violence committed by their husband/partner in the 12 months (% of ever-married women ages 15-49)'
 'Women who have experienced physical violence committed by their husband/partner in the 12 months (% of ever-married women ages 15-49)'
 'Women who have experienced sexual violence committed by their husband/partner in the 12 months (% of ever-married women ages 15-49)'
 'Women whose first experience of spousal physical or sexual violence was before marriage  (% of currently married women age 15-49 who have been married only once)'
 'Women whose first experience of spousal physical or sexual violence was within two years of marriage  (% of currently married women age 15-49 who have been married only once)'
 'Women whose first experience of spousal physical or sexual violence was within five years of marriage  (% of currently married women age 15-49 who have been married only once)'
 'Women whose first experience of spousal physical or sexual violence was within ten years of marriage  (% of currently married women age 15-49 who have been married only once)'
 'Women who have not experienced spousal physical or sexual violence (% of currently married women age 15-49 who have been married only once)'
 'Women who who have experienced injuries resulting from spousal violence (% of ever-married women ages 15-49 who have ever experienced any physical or sexual violence)'
 'Women who never sought help to stop violence, but told someone (% of ever-married women ages 15-49 who have ever experienced any physical or sexual violence)'
 'Women who never sought help to stop violence, and never told anyone  (% of ever-married women ages 15-49 who have ever experienced any physical or sexual violence)']
```

Indicator Selection

Based on the filtered results, specific indicators are selected for further analysis.

Data Preparation

Data Combination and Further Filtering: The combined dataset is further filtered based on the earlier selected indicators.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Pivot Table Creation: A pivot table is created to transform the data into a format suitable for Principal Component Analysis (PCA), with each indicator as a column, and the 'Country Code' and 'Year' as the index.

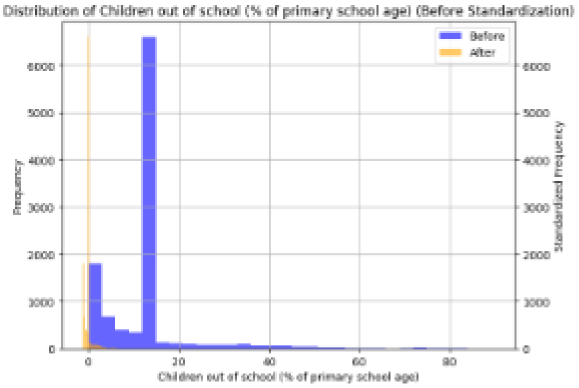
Data Cleaning

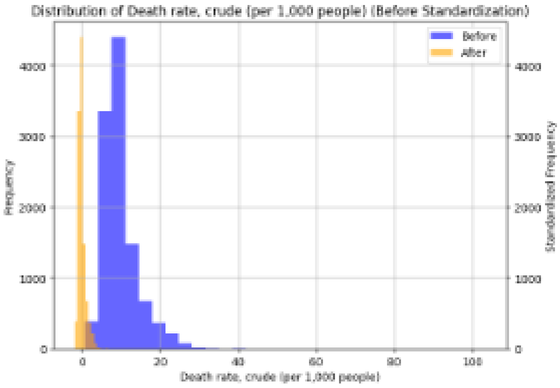
Missing Value Imputation: Missing values in the dataset are imputed using the mean of each column.

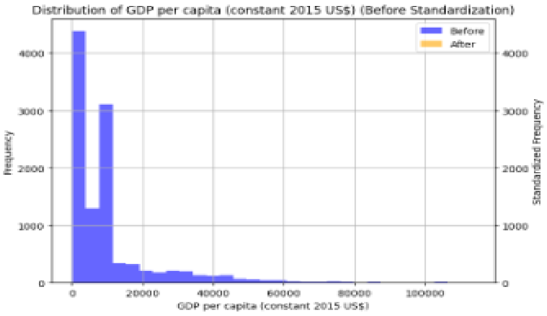
Data Standardization: The numeric data is standardised to have a mean of zero and a standard deviation of one, which is necessary for effective PCA.

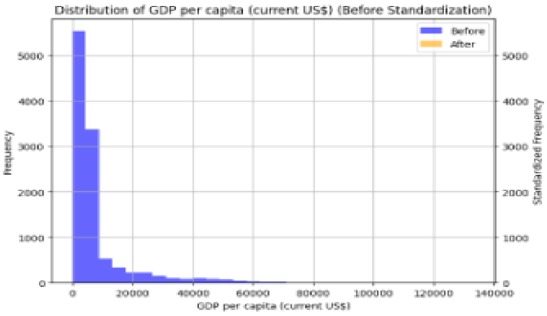
Visualization

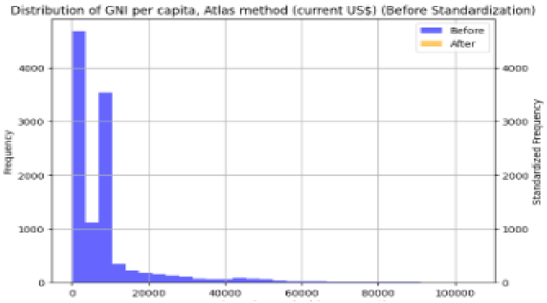
Histograms: Before and after standardization histograms are plotted for each indicator to visualize the effect of standardization. A few samples of these are shown below:

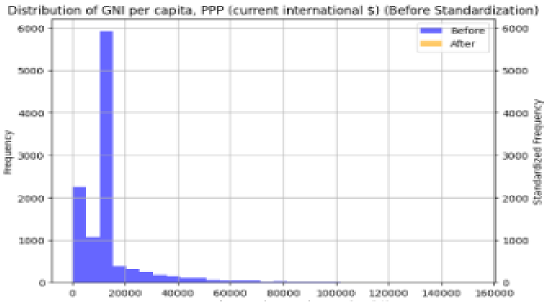

Distribution of Children out of school (% of primary school age) (Before Standardization)


Distribution of Death rate, crude (per 1,000 people) (Before Standardization)


Distribution of GDP per capita (constant 2015 US$) (Before Standardization)


Distribution of GDP per capita (current US$) (Before Standardization)


Distribution of GNI per capita, Atlas method (current US$) (Before Standardization)

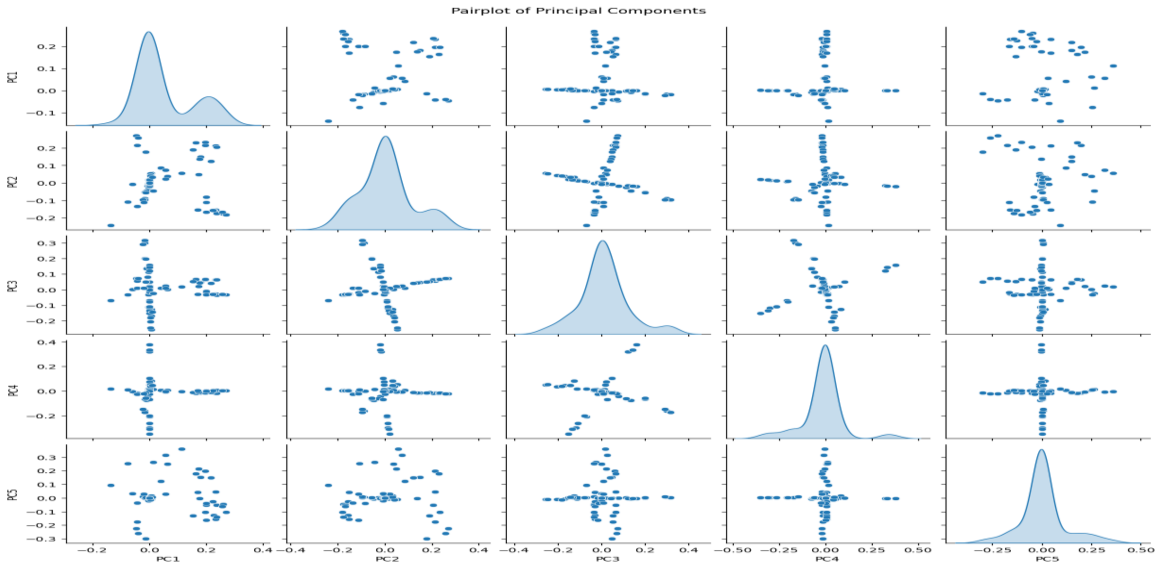

Distribution of GNI per capita, PPP (current international $) (Before Standardization)

Dimensionality Reduction

PCA Application: PCA is applied to reduce the dimensionality of the data, capturing the most significant variance in fewer dimensions – we used a total of 5 principal components.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Visualization and Analysis of PCA Results: we explore the explained variance ratio to understand the importance of each principal component.

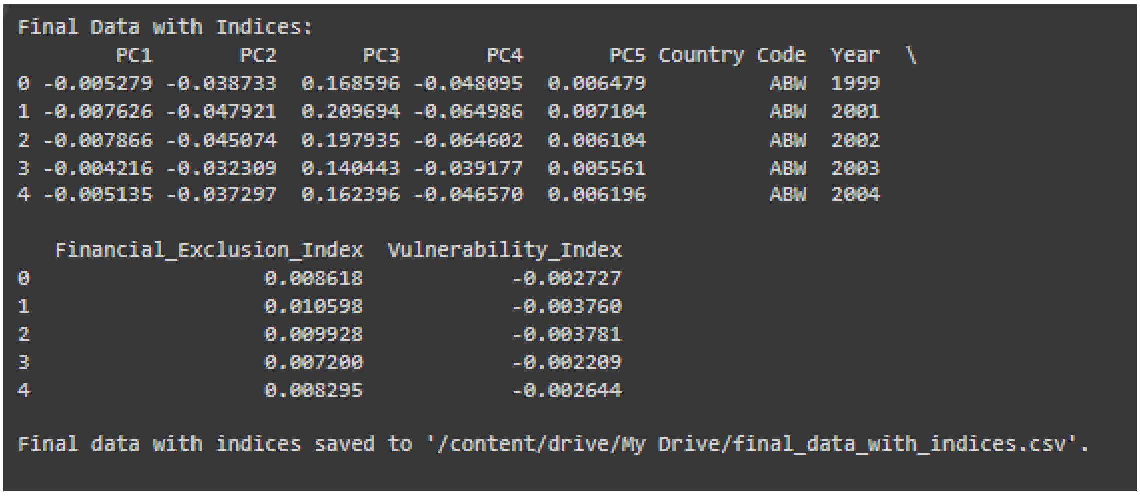


Index Creation

Index Calculation: we construct two indices, a financial exclusion index and a vulnerability index, based on selected principal components and their explained variances. These indices are aimed to quantify levels of financial exclusion and vulnerability for further analysis.

Data Saving and Output

Data Saving: The final DataFrame, including the indices, is saved back to a CSV file in Google Drive.
Loadings Display: The principal component loadings, which reflect the contribution of each original feature to the components, are also saved and displayed.



```
Final Data with Indices:
        PC1       PC2       PC3       PC4       PC5 Country Code  Year  \
0 -0.005279 -0.038733  0.168596 -0.048095  0.006479          ABW  1999
1 -0.007626 -0.047921  0.209694 -0.064986  0.007104          ABW  2001
2 -0.007866 -0.045074  0.197935 -0.064602  0.006104          ABW  2002
3 -0.004216 -0.032309  0.140443 -0.039177  0.005561          ABW  2003
4 -0.005135 -0.037297  0.162396 -0.046570  0.006196          ABW  2004

   Financial_Exclusion_Index  Vulnerability_Index
0                   0.008618            -0.002727
1                   0.010598            -0.003760
2                   0.009928            -0.003781
3                   0.007200            -0.002209
4                   0.008295            -0.002644

Final data with indices saved to '/content/drive/My Drive/final_data_with_indices.csv'.
```

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

5.2.2 Stage 2

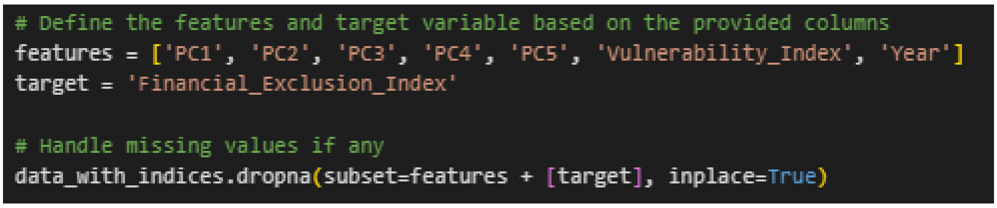
This stage involves applying machine learning techniques to the data prepared in the first stage. This is broken down into the following steps:

Machine Learning Part

Data Loading and Preprocessing

Load Data: we load the CSV file containing the saved data from the PCA analysis in Stage 1 (data_with_indices1.csv) from Google Drive. The data is also printed to show the first few rows to verify its loading.

Define Features and Target: The features selected for the machine learning model include the principal components (PC1 to PC5), the Vulnerability_Index, and the Year. The target variable to be predicted is the Financial_Exclusion_Index.



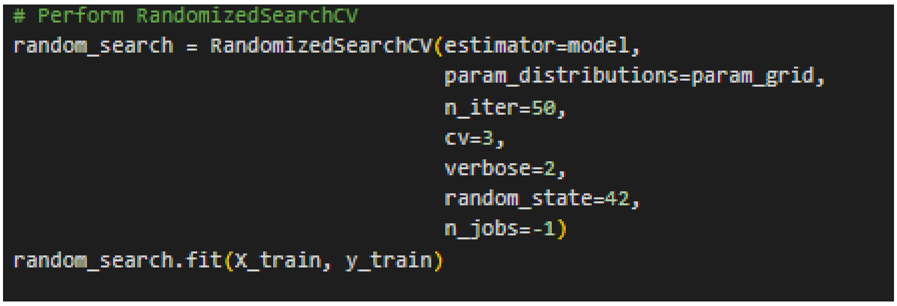
```
# Define the features and target variable based on the provided columns
features = ['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'Vulnerability_Index', 'Year']
target = 'Financial_Exclusion_Index'

# Handle missing values if any
data_with_indices.dropna(subset=features + [target], inplace=True)
```

Model Training and Evaluation

Split Data: The dataset is split into training (80%) and testing (20%) sets using sklearn's train_test_split.

Define Model: A RandomForestRegressor is chosen as the prediction model. This is useful for tasks involving regression.

Parameter Tuning: RandomizedSearchCV is used to optimize the model parameters (n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, bootstrap) over 50 iterations and 3-fold cross-validation. This helps in finding the best model configuration to improve prediction accuracy.



```
# Perform RandomizedSearchCV
random_search = RandomizedSearchCV(estimator=model,
                                   param_distributions=param_grid,
                                   n_iter=50,
                                   cv=3,
                                   verbose=2,
                                   random_state=42,
                                   n_jobs=-1)
random_search.fit(X_train, y_train)
```

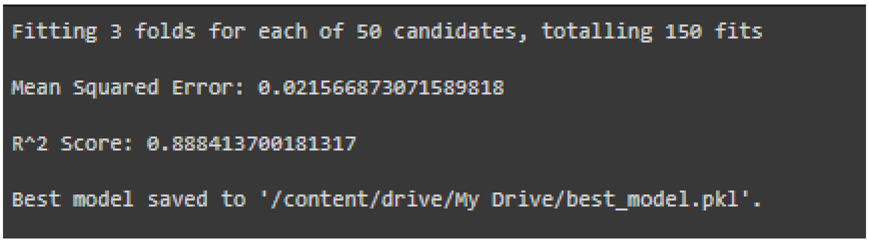
Fit Model: The best parameter set is used to train the model on the training data.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Predictions and Evaluation: Using the best model, predictions are made on the test set. The model's performance is evaluated using the mean squared error (MSE) and R² score, which are standard metrics for regression tasks. The results are shown below:

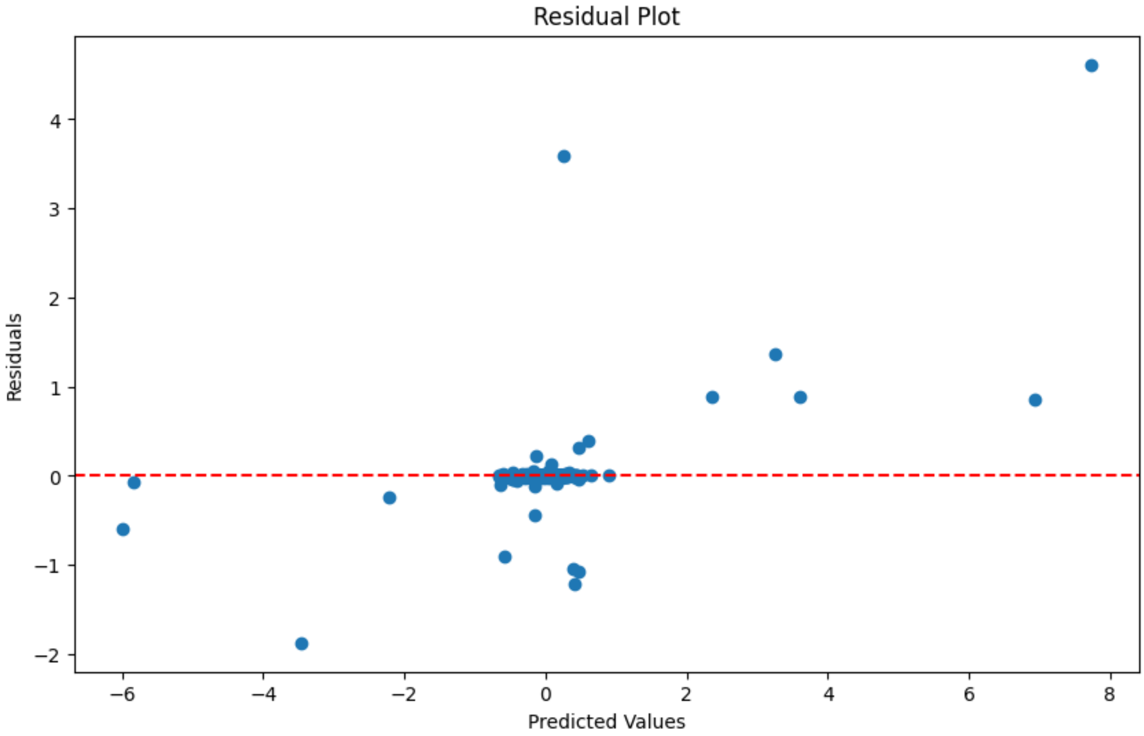
Model Output and Visualization

The best-performing model is saved to Google Drive, allowing for future use without retraining.



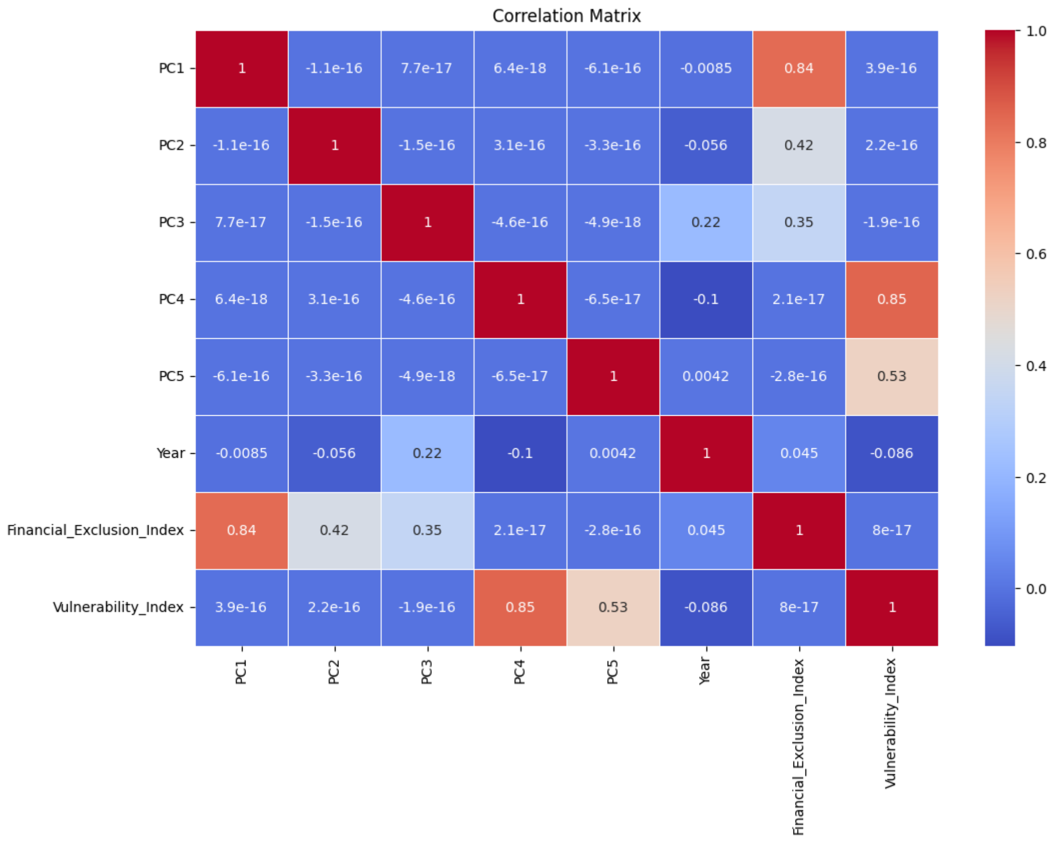
```
Fitting 3 folds for each of 50 candidates, totalling 150 fits

Mean Squared Error: 0.021566873071589818

R^2 Score: 0.888413700181317

Best model saved to '/content/drive/My Drive/best_model.pkl'.
```

Residual Plot: A residual plot (predicted vs. residuals) is generated to visually assess model performance. Ideally, residuals should be randomly distributed around the horizontal line at 0, indicating good model predictions. This was satisfied for the model appreciably as shown



Correlation Matrix: A correlation matrix for the numeric data is calculated and displayed using a heatmap. This visualization helps understand the interrelationships among the features, including the principal components and indices. The output is shown below:

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm



5.2.3 Stage 3

Here we integrate machine learning insights from the second stage with geospatial visualization to provide an even more comprehensive analysis of financial exclusion indices across different countries.

Partial Dependence Plots

Partial Dependence Plots: These plots are generated for selected features (PC1, PC2 – the strongest principal components, Vulnerability_Index) using the best model obtained from Stage 2. These plots help in understanding the marginal effect of these selected features on the predicted outcome, independent of other features.

Geospatial Data Handling

Load Shapefile: A shapefile containing geographic data (country borders, in this case) gotten from nature website is loaded.

Display GeoDataFrame Columns: we print the columns of the GeoDataFrame to verify the structure and available data after loading the shapefile.

Principal Component Analysis (PCA) Integration

Create PCA DataFrame: we create a new DataFrame to store our earlier PCA results, including principal components and additional data like country codes and years.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

Compute Indices: we compute the financial exclusion and vulnerability indices using a weighted sum of selected principal components, where weights are derived from the explained variance ratio provided by PCA done in Stage 1. This step quantifies the financial exclusion and vulnerability in numerical terms.

Add Indices to DataFrame and Display: The newly computed indices are added to the PCA DataFrame, and the first few rows are displayed to verify the data.

Saving PCA DataFrame: The final PCA DataFrame is saved to Google Drive, ensuring that the processed data is stored for future use.

Geospatial Visualisation

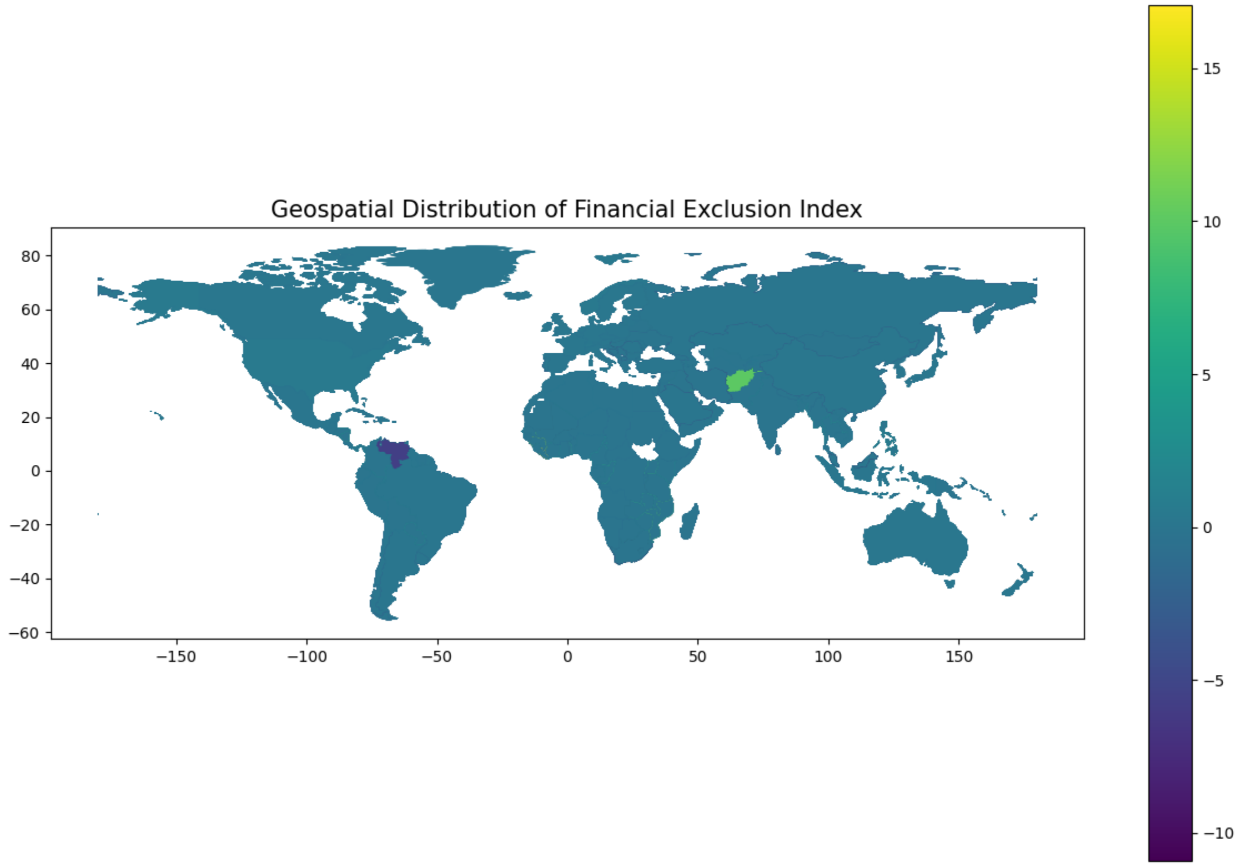
Data Preparation: An existing DataFrame (data_with_indices – data stored in the drive after PCA analysis in stage 1) is prepared to include only relevant columns for merging with the geospatial data.

Merge Data with GeoDataFrame: The GeoDataFrame containing country shapes is merged with the data_with_indices DataFrame on country codes to align financial exclusion data with corresponding geographic entities.

Verify Merged Data: Columns and a preview of the merged DataFrame are displayed to confirm the successful integration of financial indices with geospatial data.

Plot Geospatial Data: here we plot the financial exclusion index on a map using a distinct color map (viridis). The plot includes a legend and titles to enhance readability and interpretation. This visual representation is crucial for identifying geographical patterns and disparities in financial exclusion across countries. The geospatial plot is shown below

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm



The last part of the analysis implements a clustering analysis using the K-means algorithm.

Data Preparation: From the DataFrame pca_df, which is used to describe the data_with_indices – data on which PCA was applied in Stage 1. We select two columns – our financial exclusion and vulnerability indices – 'Financial_Exclusion_Index' and 'Vulnerability_Index', and drop any rows with missing values to prepare the data for clustering.

Clustering Process

We instantiate a Means object with 3 clusters and fit the K-Means model to the prepared data to find clusters based on the financial exclusion and vulnerability indices (Cui, et al., 2013).

Label Integration:

After fitting the model, we extract the cluster labels assigned to each data point and add them to the original DataFrame pca_df as a new column named 'Cluster'.

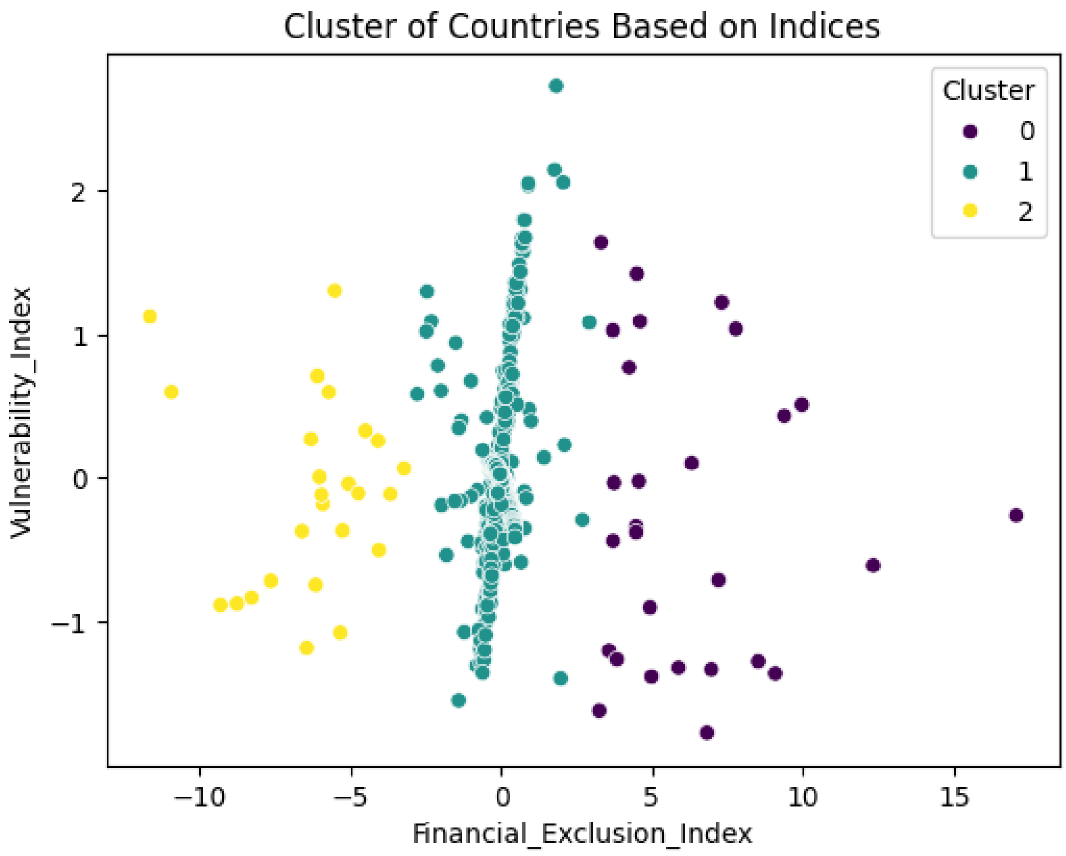
Visualization:

Using seaborn's scatterplot, we visualise the data points, plotting the 'Financial_Exclusion_Index' against the 'Vulnerability_Index'.

Each point is colored based on its cluster assignment, using the 'viridis' color palette.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

There is a final output of this code which helps to visualizing how countries are grouped based on their financial exclusion and vulnerability characteristics. The scatter plot is shown below:



6 Conclusion

This configuration manual provides a comprehensive overview of this research's key technologies, methodologies, and processes. It details the systematic approach to collecting, cleaning, processing, and analysing data using Python on Google Colab. By documenting the specific hardware and software configurations, data sources, and analytical techniques used, this manual serves as a valuable resource for replicating the research findings.

Ultimately, this manual not only support the reproducibility of this specific research project but also contributes to the broader academic and practical understanding of leveraging machine learning for financial inclusion and safety.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm

References

Bjorkegren, D. & Grissen, D., 2017. Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. *SSRN Electronic Journal*.

Cui, J., Liu, J. & Liao, Z., 2013. *Research on K-Means Clusitering Algorithm and its Implementation.* s.l., Atlantis Press.

Jedi, F. F., 2022. The Relationship between Financial Inclusion and Women's Empowerment: Evidence from Iraq. *Journal of Business and Management Studies* , 4(3), pp. 104-120.

Rahman , R. et al., 2023. A comparative study of machine learning algorithms for predicting domestic violence vulnerability in Liberian women. *BMC Women's Health*, 23(542), pp. 1-15.

Rodriguez-Rodriguez, I. et al., 2020. Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Applied Sciences* , 10(22), pp. 1-16.

https://liveprod.worldbank.org/en/topics/violence
https://dhsprogram.com/data/available-datasets.cfm