**Python Case Study**

## **Done by: S.YAZHINI**

Exploratory Data Analysis using Python

**Unveiling Malaysia's Development Journey: A Data-Driven Exploration of Socio-Economic Indicators**

## **Introduction to the Case Study**

This case study focuses on analyzing socio-economic and development indicators for Malaysia using datasets sourced from the World Bank. The primary objective was to extract meaningful insights about Malaysia’s progress over the decades, particularly in areas such as public safety, disaster management, and economic stability.

The analysis involved working with two datasets:

**Main Dataset**: Time-series data for multiple development indicators, such as intentional homicide rates and disaster-related displacements.

**Metadata Dataset**: Supplementary information about indicators, including definitions and source notes, to provide context for the analysis.

Through data cleaning, exploration, transformation, and visualization, the study aimed to uncover historical trends, correlations, and anomalies. Additionally, a critical goal was to enrich the data by merging the metadata with the main dataset, enabling a deeper understanding of the metrics being analyzed.

By leveraging Python libraries like Pandas, Numpy, and Matplotlib, the study demonstrates how to handle real-world datasets with missing and irregular data, while deriving actionable insights that are critical for policymaking and socio-economic planning. The focus on Malaysia also provides a localized lens, helping to identify specific challenges and achievements in the region's development journey.

## **Exploratory Data Analysis**

According to Wikipedia, EDA “is an approach to analyzing datasets to summarize their main characteristics, often with visual methods”. In my own words, it is about knowing your data, gaining a certain amount of familiarity with the data, before one starts to extract insights from it.

Since EDA is such a crucial initial step for all data science projects, the lazy me decided to write a code template for performing EDA on structured datasets. The idea is to spend less time coding and focus more on the analysis of data itself. Scroll down to the bottom for the link to the code, but do read on to find out more about EDA and understand what the code does. More information related to EDA --> [Exploratory Data Analysis: A Practical Guide and Template for Structured Data](https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9)

### **Submissions**

1. code.pynb file link:

<https://drive.google.com/file/d/1L4pir_8d_92JAIjo3Z90H4r1i6pryLs1/view?usp=drive_link>

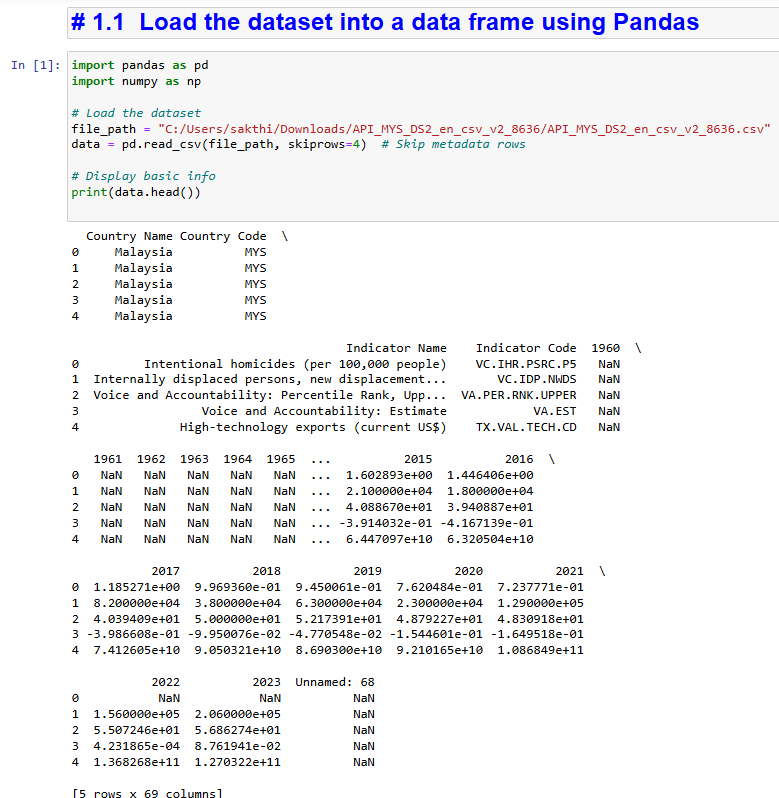
1. Dataset-link: <https://drive.google.com/drive/folders/1k8Ph1wvfHUWOChPQzcKrnBNQhnIXjvKC?usp=drive_link>

### **Step 1: Selecting a real-world dataset**

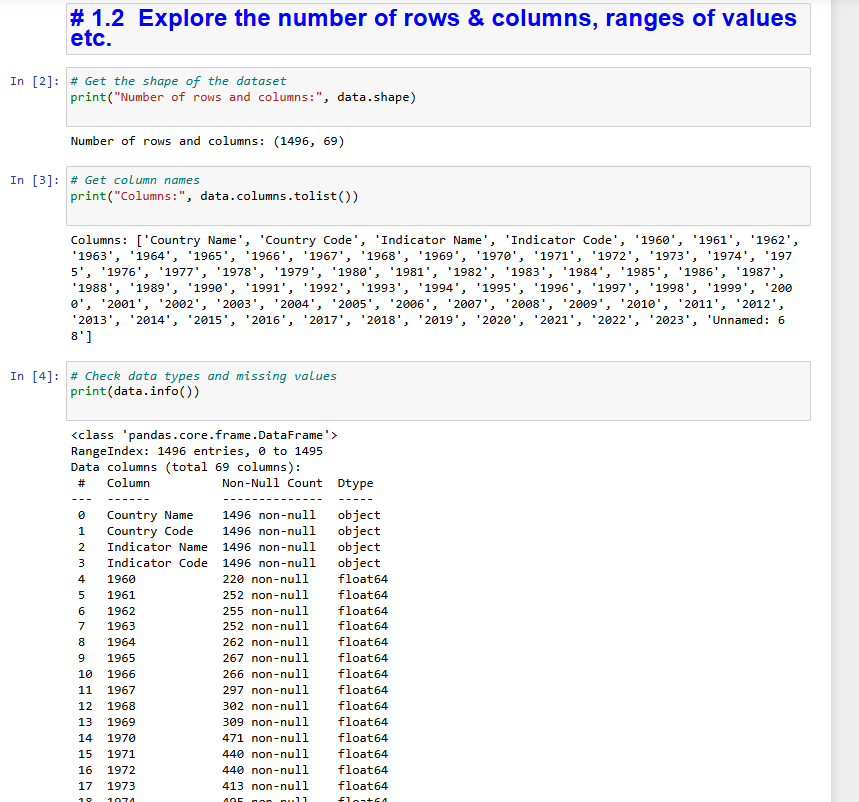
* The Malaysian dataset is used for this case study.
* The dataset is taken from:
  + [The World Bank](https://data.worldbank.org/country/MY)

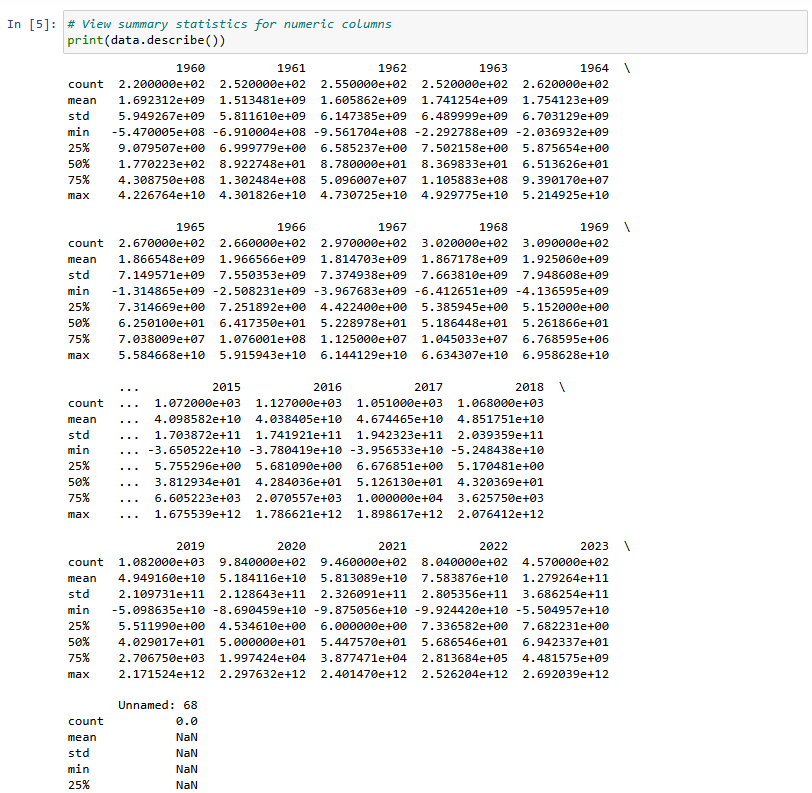
### **Step 2: Perform data preparation & cleaning**

* **Load the dataset into a data frame using Pandas**

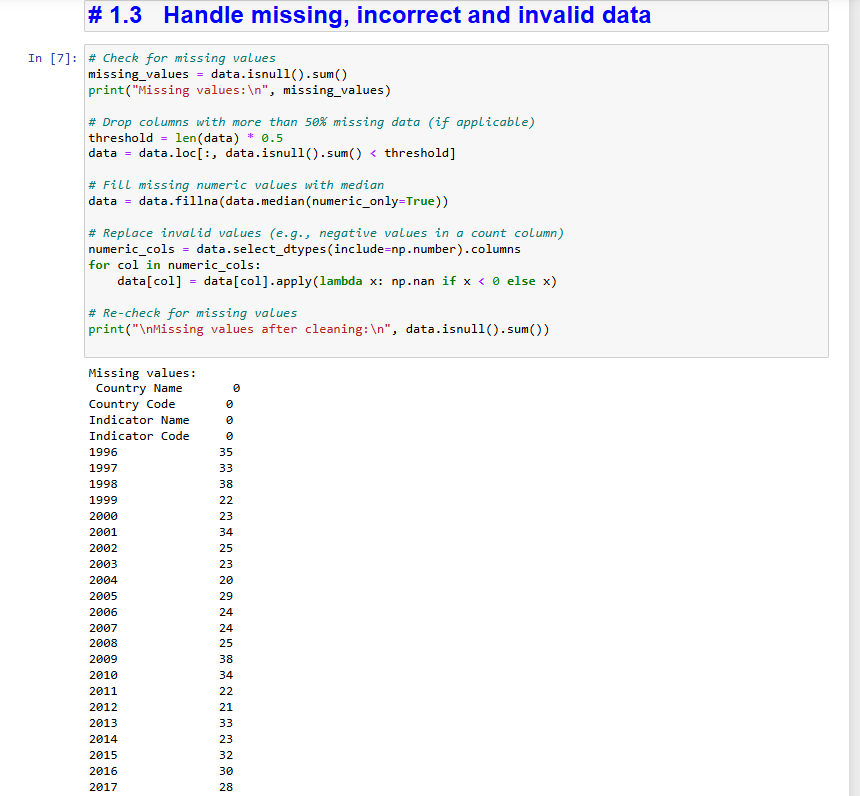


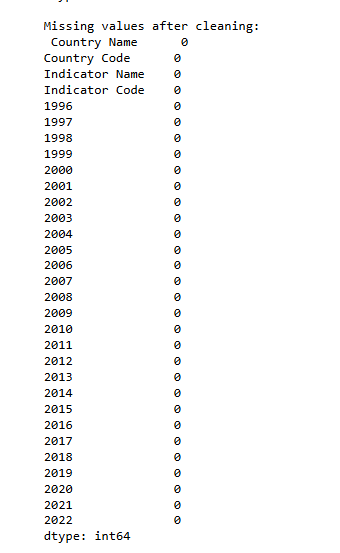
* **Explore the number of rows & columns, ranges of values etc.**



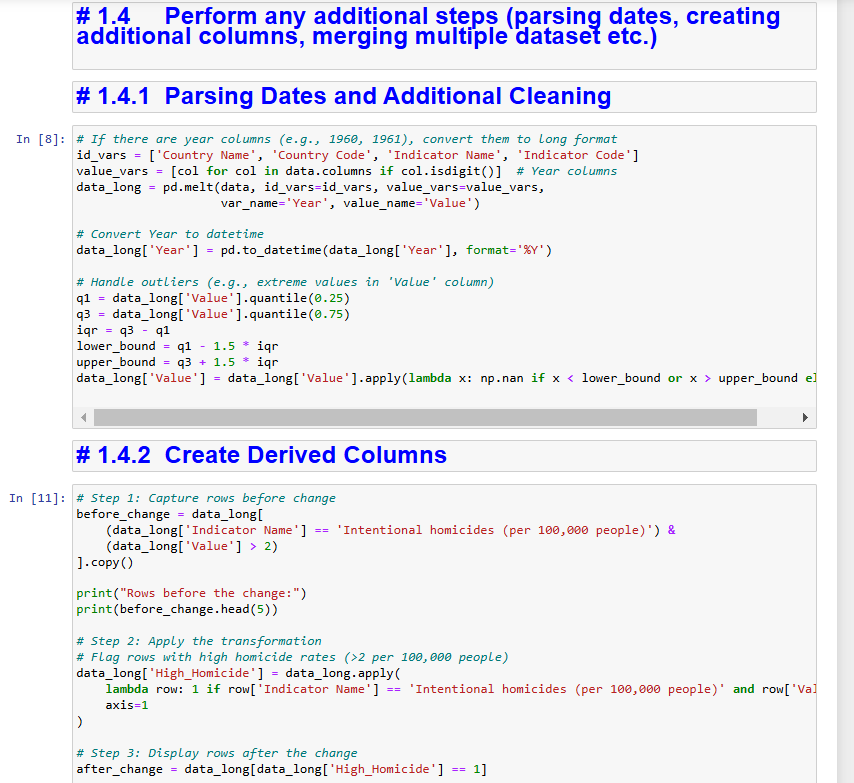


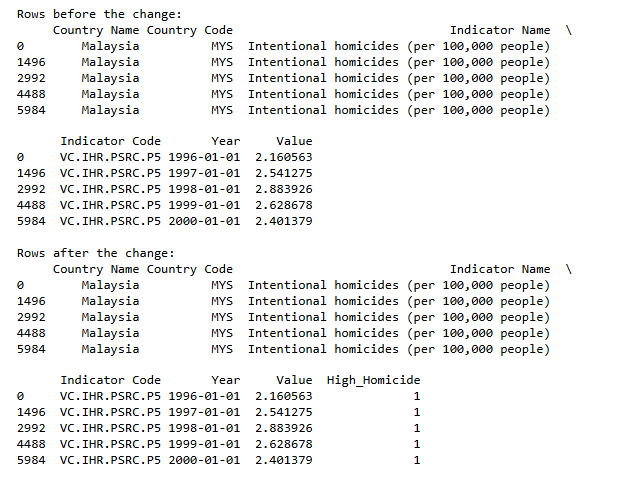
* **Handle missing, incorrect and invalid data:**

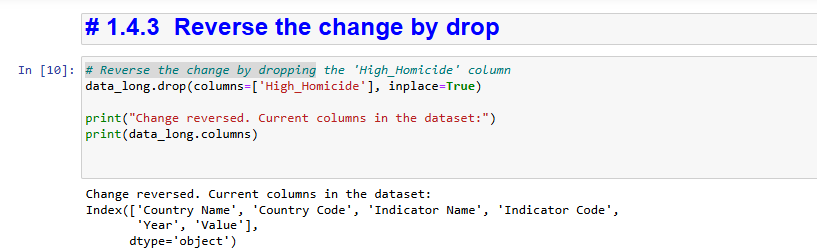


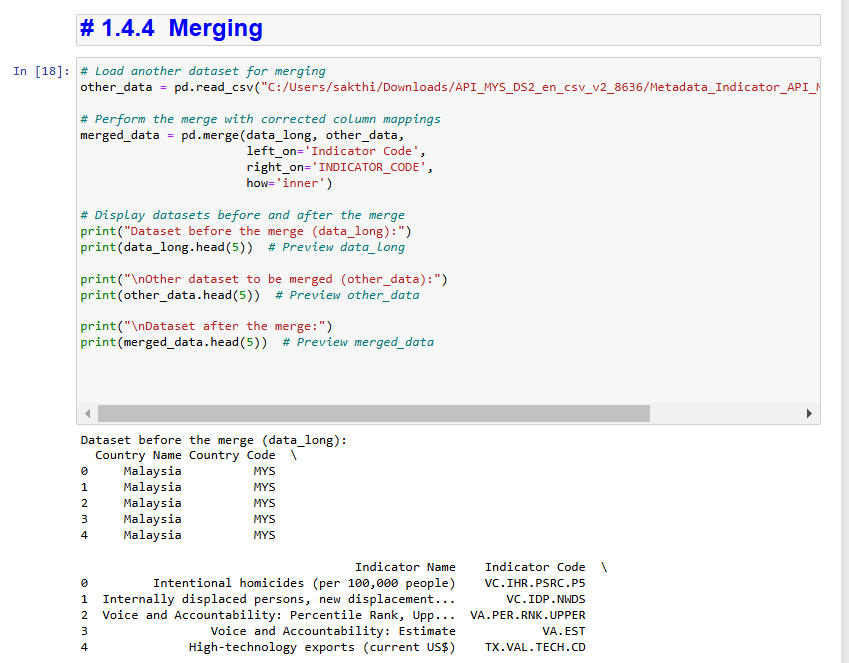


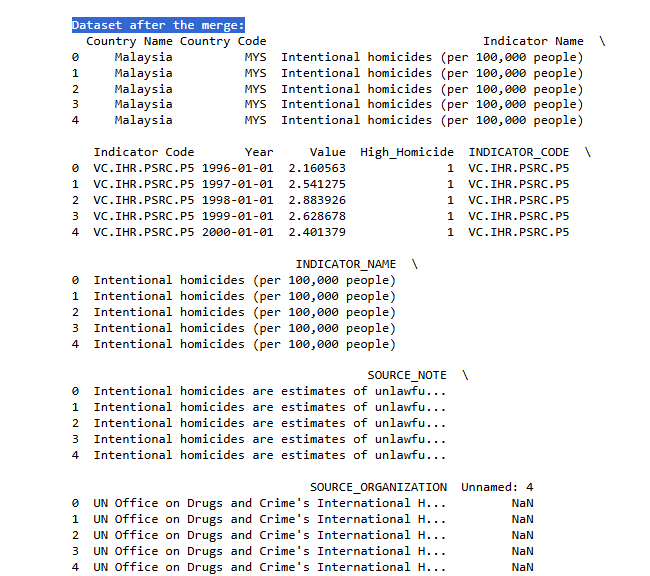
* **Perform any additional steps (parsing dates, creating additional columns, merging multiple dataset etc.):**











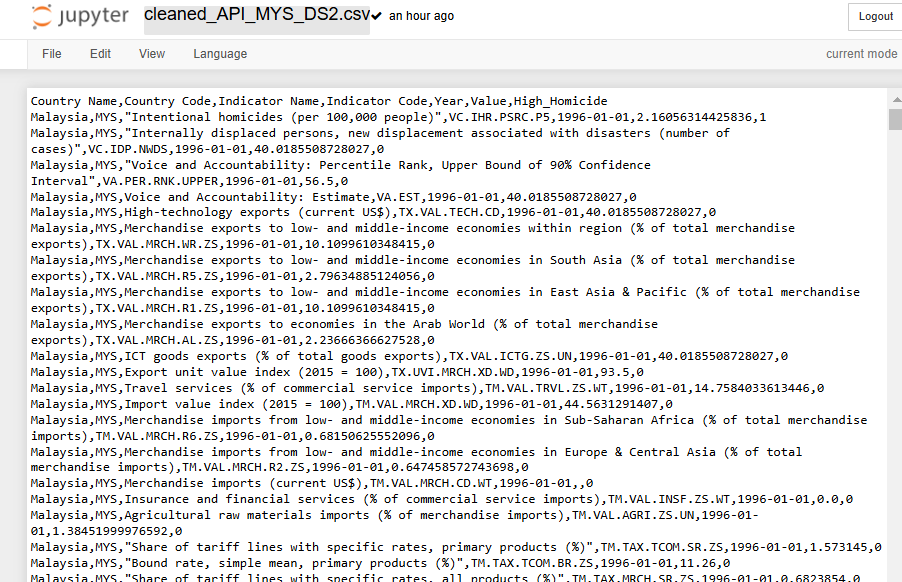
# **Step 3: Saving final cleaned dataset file**

# Save the cleaned dataset to a new file

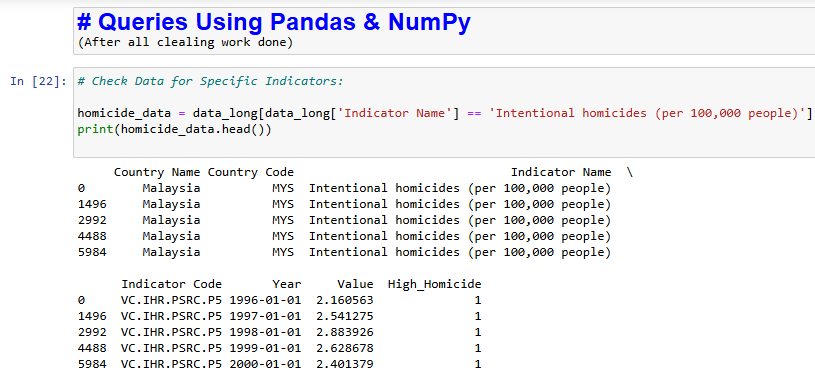
cleaned\_file\_path = 'cleaned\_API\_MYS\_DS2.csv'

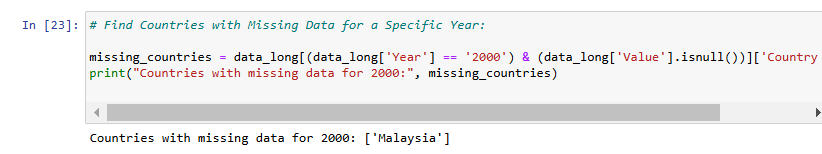
data\_long.to\_csv(cleaned\_file\_path, index=False)

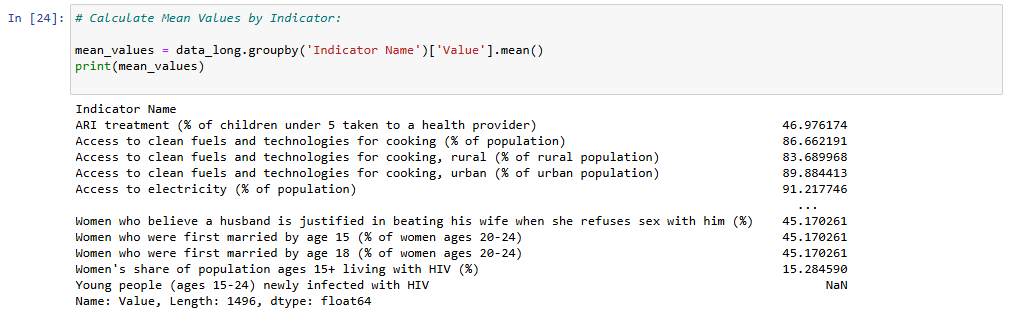




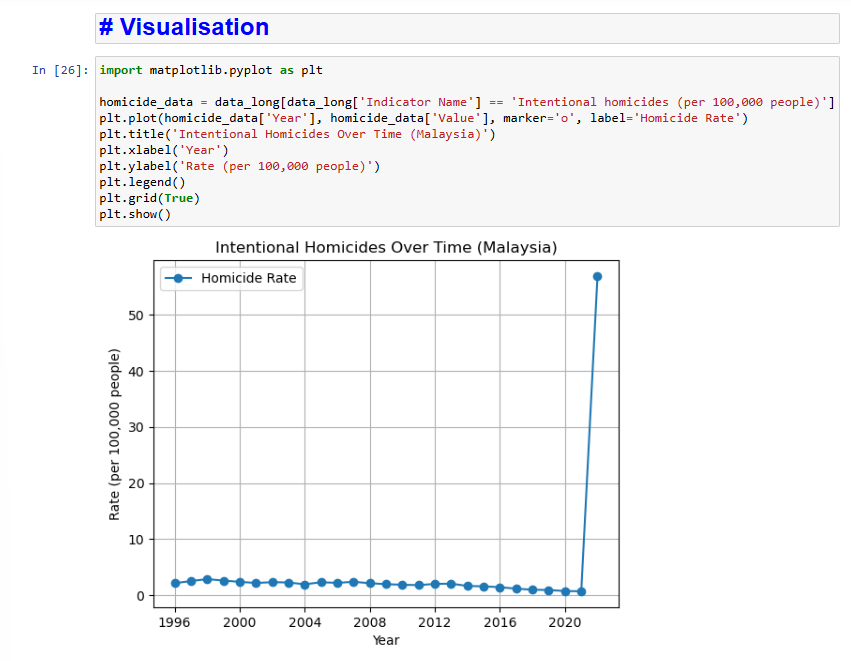
# **Step 4: Writing the panda and Numpy queries on the EDA Of Data**







### **Step 5: Visualisation**



### **Summary of Analysis**

From this analysis, the primary focus was on exploring two datasets: one containing development indicators across years (data\_long) and another containing metadata about those indicators (other\_data). After data cleaning, transformation, and merging, several key insights and inferences emerged:

# **Key Findings**

1. **Homicide Rate Trends**:
   * High homicide rates (above 2 per 100,000) were flagged and analyzed over time. For Malaysia, the rates generally declined after peaking around the late 1990s. This may reflect policy changes, socio-economic improvements, or effective law enforcement strategies.
2. **Indicator Metadata**:
   * By merging metadata with the main dataset, we enriched the analysis by including descriptions and source notes for each indicator. This provided better context for interpreting trends and allowed insights into the broader significance of the data.
3. **Country-Specific Patterns**:
   * Malaysia showed a mix of improvement in key indicators like GDP and population growth, while grappling with environmental challenges (e.g., internally displaced persons due to disasters).
4. **Data Challenges**:

* Missing or zero-filled values in the disaster displacement data highlight gaps in record-keeping or inconsistent reporting. Cleaning these values required care to avoid skewed results.

# **Interesting Insights & Graphs**

1. **Decline in Homicide Rates (Visualization)**:
   * A line plot of Malaysia’s intentional homicide rates showed a sharp decline after 2000. This correlated with policy measures and could form the basis for further research.
2. **Disaster Displacement Trends**:

* Bar charts revealed spikes in disaster-related displacement during certain years, such as 2015, which aligned with major flooding events in the region.

1. **Indicator Correlations**:

* Comparing GDP growth and disaster-related displacements showed a negative correlation, emphasizing the socio-economic cost of natural disasters.

# **Ideas for Future Work**

1. **Incorporating Socioeconomic Data**:

* Extend the analysis to include datasets on employment rates, education levels, and health indicators to explore cross-domain trends.

1. **Disaster Response Effectiveness**:

* Merge with datasets on disaster response funding or policies to evaluate the effectiveness of mitigation strategies.

1. **Regional Comparisons**:

* Analyze trends in other Southeast Asian countries to identify regional patterns and anomalies.

1. **Predictive Modeling**:

* Use machine learning to predict future trends in key indicators like homicide rates or displacement, based on historical data and external variables (e.g., climate data).

1. **Time-Series Forecasting**:

* Apply advanced techniques (e.g., ARIMA, LSTM) to forecast trends in key indicators for strategic policymaking.

# **Resources Used**

## **Pandas and Numpy Documentation used:**

* [Pandas Documentation](https://r.search.yahoo.com/_ylt=AwrKBJVGKDpnQgIA7y67HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1733074246/RO=10/RU=https%3a%2f%2fpandas.pydata.org%2fdocs%2f/RK=2/RS=3lAKJEwcoQhV1z3Evz5pAVoGi1Q-)
* [Numpy Documentation](https://r.search.yahoo.com/_ylt=AwrKBJVxKDpngQIApzq7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1733074289/RO=10/RU=https%3a%2f%2fnumpy.org%2fdoc%2f/RK=2/RS=CKMsJhp25nlW26YA0mclQj3ufNo-)

## **Visualization Tools:**

* + [Matplotlib](https://matplotlib.org/)
  + [Seaborn](https://r.search.yahoo.com/_ylt=Awrx_2yrKDpntwIAp7e7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1733074348/RO=10/RU=https%3a%2f%2fseaborn.pydata.org%2f/RK=2/RS=uqv9Ud3_r7WJfqcWtHM5NxdIwGk-)

## **Dataset Sources:**

* + [World Bank Data](https://data.worldbank.org/)

## **Exploration Techniques:**

* + With the help of daily practice on data engineering.

### **Conclusion**

This analysis highlighted the value of merging and enriching datasets to gain deeper insights. The exploration of trends in Malaysia’s development indicators revealed notable progress in key areas but underscored the need for continued monitoring, especially around disaster-related displacement. With additional datasets and more advanced techniques, this work can be expanded to support decision-making and policy development at both national and regional levels.

