**Data Preparation/Feature Engineering**

**1. Overview**

Any machine learning project must include the crucial step of data preparation, which entails cleaning, organizing, and structuring raw data to produce a dataset that is appropriate for analysis and model training. Important actions in this preparing data consist of:

* Data Cleaning
  + Identify and handle missing values, outliers, and errors.
  + Address any issues that could affect the quality of the dataset.
* Data Transformation
  + Convert categorical variables into a format suitable for machine learning algorithms (one-hot encoding).
  + Normalize or scale numerical features to ensure uniformity.
* Data Integration
  + Combine multiple data sources if applicable to create a comprehensive dataset.
  + Ensure compatibility and consistency across integrated datasets.
* Data Splitting
  + Split the dataset into training and testing sets to assess model performance on unseen data.

**2. Data Collection**

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.

**3. Data Cleaning**

I used the Pandas library to retrieve only west African countries. I continue with preprocessing of data by drop columns(features) that have more than 25% of missing values. I also drop feature that I think are unnecessary for the analysis.

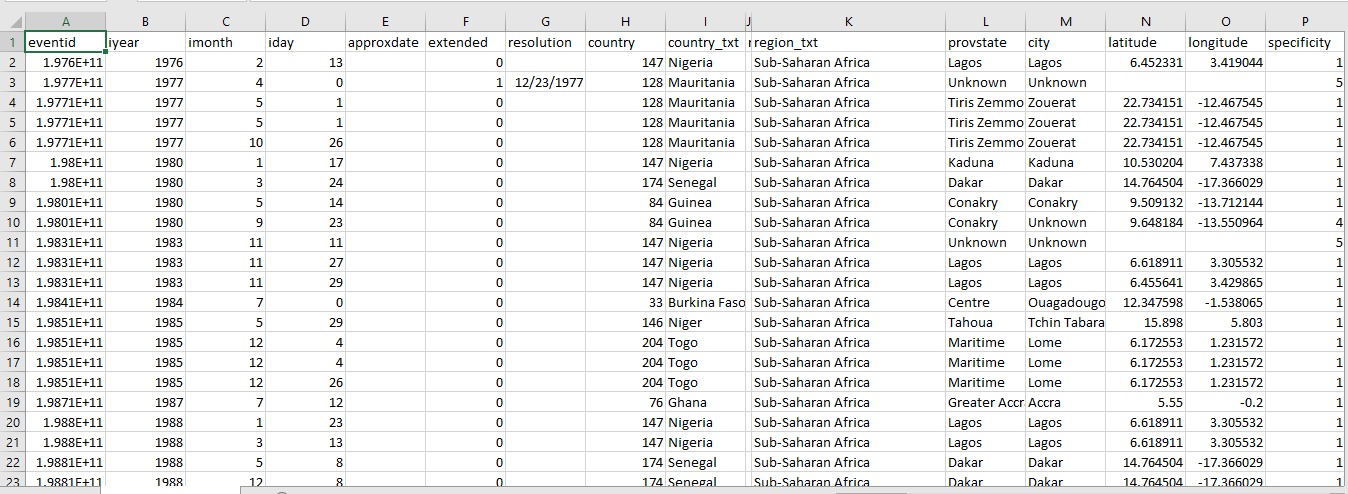
**4. Exploratory Data Analysis (EDA)**

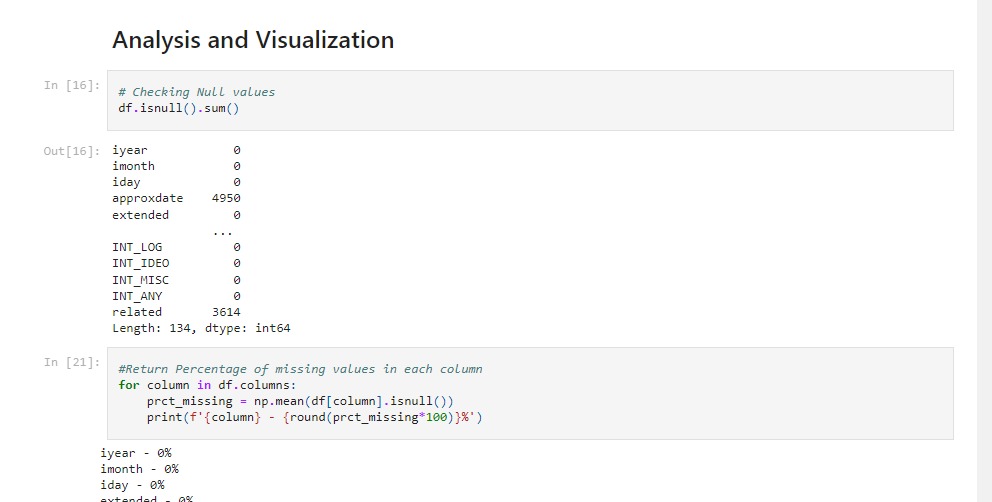
Multiple methods for dataframe has been used such: info,describe,sample,shape,head,drop….

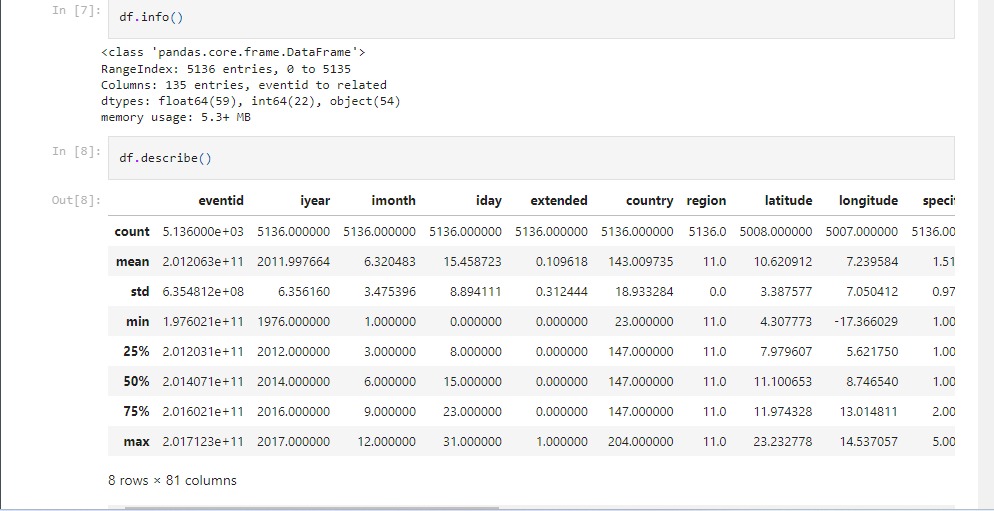
After that methods like fillna() has been used to fill in values.

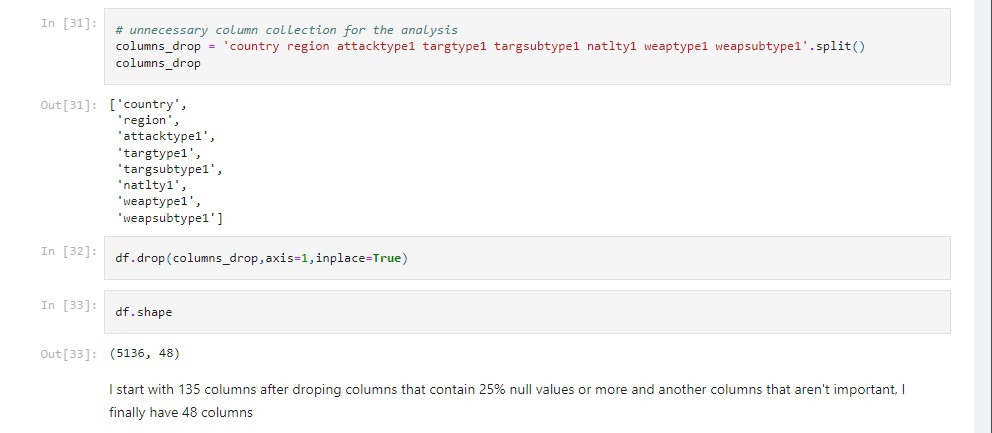
To see very good what I did I decided to uploaded the notebook on My Github so you can easily go through the visualization and codes. Click the link bellow:

<https://github.com/baafbass/Casptone_Project_For_SDG/blob/main/Terrorism_attack.ipynb>









**5. Feature Engineering**

During the analysis I tried to find relation between features based on the time (Years,days,months…) and based on the terrorist incidents such (most succeeded terrorist incident by countries and cities, and also the most group attacks)

**6. Data Transformation**

After dropping and filling in the unnecessary and missing values I used LabelEncoder for encoding the reminded categorical columns.

**Model Exploration**

**1. Model Selection**

In your project, you've considered several machine learning models, including Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost. The choice of models depends on various factors, including the nature of the problem, the characteristics of the dataset, and the specific goals of the project.

* Logistic Regression

Rationale:

Logistic Regression is a commonly used algorithm for binary classification problems.

It's interpretable and provides probabilities, making it suitable for understanding the impact of features on the likelihood of a terrorism incident.

Strengths:

Interpretable and easy to understand.

Efficient for linearly separable problems.

Weaknesses:

Assumes a linear relationship between features and the log-odds of the response.

* Random Forest

Rationale:

Random Forest is an ensemble method known for its robustness and ability to handle complex relationships in data.

It can capture non-linear patterns and interactions between features effectively.

Strengths:

Robust to overfitting.

Handles high-dimensional data well.

Provides feature importance scores.

Weaknesses:

May be computationally intensive for large datasets.

Can be challenging to interpret compared to simpler models.

* Decision Tree

Rationale:

Decision Trees are straightforward and can represent complex decision boundaries.

They are easy to interpret, providing transparency into the decision-making process.

Strengths:

Simple to understand and visualize.

No need for feature scaling.

Weaknesses:

Prone to overfitting, especially with deep trees.

Sensitive to small variations in the data.

* Gradient Boosting

Rationale:

Gradient Boosting is an ensemble technique that sequentially builds weak learners to improve overall predictive performance.

It's effective in capturing intricate relationships and achieving high accuracy.

Strengths:

High predictive power.

Handles complex interactions in the data.

Robust to outliers.

Weaknesses:

Can be computationally expensive.

Sensitive to hyperparameter tuning.

* XGBoost

Rationale:

XGBoost is an optimized implementation of gradient boosting with additional features.

It's known for its speed, efficiency, and scalability.

Strengths:

Fast and scalable.

Regularization to prevent overfitting.

Handles missing values well.

Weaknesses:

Requires careful parameter tuning.

Can be resource-intensive.

Considerations for Model Selection:

Dataset Size: Random Forest, Decision Tree, and Logistic Regression can be effective for smaller datasets, while Gradient Boosting and XGBoost often shine in larger datasets.

Interpretability: If interpretability is crucial, Logistic Regression and Decision Trees provide clearer insights into the model's decision-making process.

Complexity of Relationships: For complex, non-linear relationships, ensemble methods like Random Forest, Gradient Boosting, and XGBoost are generally more powerful.

**2. Model Training**

All the model has been trained by splitting the data into 80% of training and 20% of testing

**3. Model Evaluation**

Confusion Matrix,Accuracy Score,Precision Score,Recall Sccore,F1\_Score and Classification Report has been used for evaluation fo the model

**4. Code Implementation**

To see very good what I did I decided to uploaded the notebook on My Github so you can easily go through the visualization and codes. Click the link bellow:

<https://github.com/baafbass/Casptone_Project_For_SDG/blob/main/Terrorism_attack.ipynb>