MAKERERE UNIVERSITY

SEMESTER ONE 2024/2025 ACADEMIC YEAR

SCHOOL COMPUTING AND INFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF SCIENCE IN COMPUTER SCIENCE

MCS-7103 - MACHINE LEARNING

EXPLORATORY DATA ANALYSIS ASSIGNMENT

PATRICK ANAKU

2024/HD05/21916U

September 18, 2024

TABLE OF CONTENTS

1. INTRODUCTION

In this exploratory data analysis task, I investigated and analyzed the African crime dataset described in section 2.0. The main objective was to understand crime on the African continent. The specific objectives are to:

1. Use predictive modelling to predict trends, provide crime insights, and forecast potential crime hotspots. For instance, over time, which type of violent crime consistently has the largest rate of occurrences in the and in which region? Who are the actors, what are the fatality rates, et cetera?
2. Psychological Assessment: Understanding the mindset, particularly offenders.

2.0 DATA COLLECTION

2.1 How the dataset was obtained

The Armed Conflict Location & Event Dataset Project (ACLED) collected the dataset. ACLED collects reported information on the type, agents, location, date, and other characteristics of political violence events, demonstration events, and other select non-violent, politically relevant developments in every country and territory. ACLED focuses on tracking violent and non-violent actions by or affecting political agents, including governments, rebels, militias, identity groups, political parties, external forces, rioters, protesters, and civilians. The ACLED data are derived from a wide range of local, national, and international sources in over 75 languages.

3.0 DATA WRANGLING

3.1 Pre-Processing Techniques Used

Described in this section are the data pre-processing approach that I used to clean and organize the data. Python was chosen as the data analysis tool.

3.1.1 Loading the Dataset

The dataset was downloaded in Excel format from ACLED and changed to a CSV file. Then before loading the dataset into a Pandas Dataframe, I imported the necessary Python libraries namely Pandas, Matplotlib and Seaborn. Then, read the dataset into a Pandas Dataframe.

3.1.2 Understanding the Dataset

This was done by knowing the features each column stands for to avoid mistakes in data analysis and modelling. I created a data frame with the names of the columns, data types, the first and last few rows' values, unique column values and statistical summary from the data dictionary

3.1.3 Dataset Cleaning

The dataset cleaning was done by writing python code that importantly renamed some column names to reflect the actual meaning of the values in the column.

4.0 EXPLORATORY DATA ANALYSIS

4.1 Overall look into the data

This was just a continuation of the data-wrangling task. I took another detailed look at the dataset using various Pandas functions.

4.2 Exploring categorical features

4.2.1 Visualization for categorical features

These helped me to understand the distribution of different categories visually.

1. A bar chart to visualize the frequency of each category.
2. A pie chart to visualize the percentage of event types

4.3 Exploring numerical features

4.3.1 Visualization for categorical features

I used the plots below to visualize numerical features in my dataset bar chart to visualize the frequency of each feature. This helped me to understand the distribution of different categories visually.

1. A bar chart to visualize the frequency of each numerical feature.

2. Visualize the total number of fatalities by actor type. Here I used both a bar plot and pie chart.

3. I used the EVENT_DATE column to track changes in event occurrences over time. I basically extracted specific components like year for visualizing this temporal trend.
4. I analyzed patterns in latitude and longitude to identify hotspots of activity using geo features in the dataset. Here I used a scatter plot of fatalities by location.

4.4 Features Identified for Analysis

After the detailed feature exploration, I uncovered the essential features that could help my analysis. The features identified are event types, total fatalities, location, and the actors involved. I stated earlier that the reason for this EDA is to be able to perform predictive modelling, predict trends, provide crime insights, and forecast potential crime hotspots. For instance, over time, which type of violent crime consistently has the most significant rate of occurrences in the and in which region? Who are the actors, what are the fatality rates, et cetera?

5.0 FINDINGS AND CONCLUSIONS

Based on my selected features described in section 4.4 below are some insights and conclusions I have been able to get from the dataset.

1. Event Type Distribution: The most frequent event type is violence against civilians, which reveals the dominant forms of conflict and instability in Africa.

2. Fatality Trends: The fatality analysis showed that the highest rate of fatalities over time and by different factors, for instance, actor type and event type, is Angola, followed by the Democratic Republic of Congo.

3. Geographical Hotspots: The hotspot for events on the continent is Somalia, which has multiple conflicts.

4. Actor Involvement: Some relations exist between the actors involved in the conflict across the African continent.

5. Temporal Patterns: The trend analysis reveals that pattern conflicts are rising over time.

6. Dominant words used: Word clouds generated from event descriptions highlighted the most dominant word ‘Killed’ as used in describing the conflicts. This indicates the extent of the violence.