Exploring Global Terrorism Dataset (GTD) – Visual Exploration

Youssef Abdelmoamen

Abstract—In this paper, the focus was on analyzing the glob terrorism dataset. The data was analyzed to answer multiple research questions which would give useful indights on how terrorist incidents have evolved, changed and insights on the specific groups. For the first task, a time series analysis was carried out to check how terrorist activities have evolved. Next, for task two I used descriptive statistics to see the significant changes in the lethality of the terrorist attacks over time. Finally, a kmeans clustering algorithm was built to aid to find similar terrorist groups and how their activitiesdiffered in terms of geographical focus. Visualization techniques will be used throughout this analysis as well as computational methods to aid in refining our answers.

1 Problem Statement

Examining the patterns, trends, and features of international terrorist activity over the previous few decades is the main goal of this investigation. As a worldwide phenomenon, terrorism poses a difficult, multidimensional problem that affects international relations, policymaking, and security. By analyzing the GTD, the aim is to uncover insights that could contribute to a deeper understanding of the dynamics of terrorist incidents worldwide. This study attempts to provide answers to the following research questions:

- How have terrorist activities evolved over time in terms of frequency, severity, and geographical spread?
- Have there been significant changes in the lethality of terrorist attacks?
- Which terrorist groups have been the most active, and how do their activities differ in terms geographical focus?

The dataset is an extensive, publicly available database that contains thorough details on around 200,000 terrorist incidents that have occurred globally between 1970 and the present. Key data points include the date and location of the incident, attack type, weapons used, target type, perpetrator group, casualties, and more. The dataset is organized in a tabular format, with each row denoting a distinct terrorist occurrence and the columns listing different characteristics of every instance. The GTD's extensive coverage, in terms of historical span and global breadth, makes it ideally suited for this investigation. An extensive investigation of trends, patterns, and changes in terrorist activity is made possible by its comprehensive records.

2 State of the Art

With the introduction of visual analytics techniques in the 21st century, the field of terrorist analysis has undergone a paradigm shift. researchers have been innovating in the visual analysis of terrorism, with the GTD as a primary resource, changing the way these events are understood and predicted. These developments allow for the understanding of the terrorism patterns and allow for the enhancement of global security measures by predicting future occurrences. for visual analytics. Visual analytics and ensemble learning were combined in the first article which was created by Pan et al. while focusing on predictive models for classifying terrorist groups based on data from the GTD. They used Random Forest and decision trees to identify patterns that might be used in the future to forecast any terrorist attack, this methodology is very useful for my research, which looks for trends in terrorist conduct across time and in different regions [1].

In the other article that I read, Guo et al. focused on spatio-temporal characteristics of terrorism events and introduced six different visual analytics methods. They were able to use thematic maps and statistical charts to show the evolution of terrorism, which is similar to my goal where I aim to explore the patterns in terrorism and its global dispersion. Their use of “5W” theory for the data structuring is essential in directing the thematic focus of the visual analysis [2].

The work on visual analytics on terrorism in india presented at the IEEE conference is very interesting as it shows how unstructured data from online news and tweets can be integrated with the GTD to increase its richness [3].

After the analysis of these three papers, it is evident that despite visual analytics being able to provide valuable insights regarding terrorist trends, there are also some drawbacks. These drawbacks include that current models might not fully account for terrorist organizations, and important details might actually be missed due to the assumptions being solely based on the GTD. As a result, I will build on these approaches in my research where I will map terrorist activity clusters to improve spatial analysis and provide a geographical dimension o the patterns discovered.

In conclusion, the lesson learnt from these publications were invaluable. Not only do they show efficient visual analysis in the field of terrorism but also show me where the research could go: by combining comprehensive visual analytic with extensive, contextually-enriched datasets, a good understanding of the dynamics of global terrorism can be created.

3 Properties of the Data

The GTD is a large, publicly available dataset that contains details on around 200,000 terrorist attacks since 1970 till the present day. It is one of the most comprehensive datasets on the subject, and it was assembled by the National Consortium for the Study of Terrorism and Responses to Terrorism (START).

The data in the GTD was collected through a thorough process that involves reviewing news outlets, older databases and secondary source materials. To ensure the accuracy of the data, each incident from the dataset was cross-referenced from different resources [4For this analysis, I got the dataset from official website of the GTD [5](the dataset).

The dataset is in tabular format (csv) and it consists of discrete entries where each entry represents a unique incident

With field that include date, location, tactics, targets, perpetrators, casualties, and more. The types of these fields vary where some are categorical, numerical or text-based, encompassing a wide range in space and period of time. The spatial coverage of the database covers the whole globe while the temporal coverage of the database provides a nearly fifty-year longitudinal view of terrorist incidents.

After importing the dataset, looking at its shape (209706,135) and describing it to see the statistical values. Using the *isnull* method, I checked all the missing values but had to visualize them as well so the picture is clear, so the following heatmap was created:

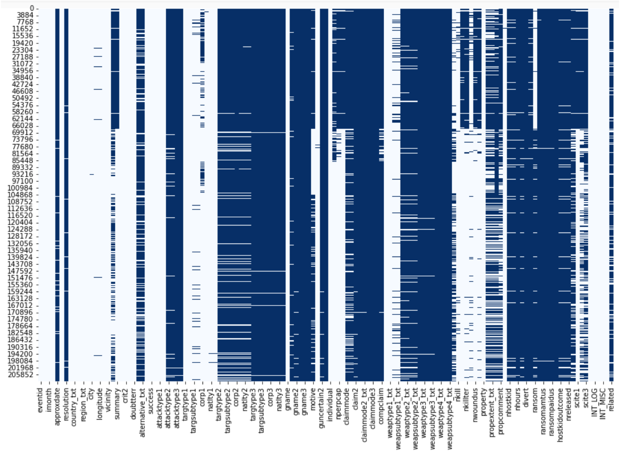


Figure 1: Heat map of missing values

This heatmap uses 2 shades of blue to indicate the presence or absence of data. Darker shades correspond to missing value while lighter colors correspond to non-missing values. The x-axis represents the different variables in the GTD while the y-axis represents individual records. Through observation of the heat map, it is evident that there are a lot of missing values in some variables more than others. For instance, variables such as attacktype2 and attacktype3 show very obvious gaps, indicating lack of information in these fields.

Starting the analysis with a lot of missing values could

problem, a strategic approach is necessary. The first solution is excluding/dropping some variables that are insignificant to the analysis, this way the integrity and reliability of the analysis will be maintained.

All in all, after initial analysis some data quality concerns were noted:

- **Missing values:** some fields including, including city, the amount of people killed and the amount of people wounded, would occasionally lack data. (heatmap)
- **Outliers**: a few of the numerical fields like casualty figures have outliers which represent particularly severe incidents. (boxplots were also create but will be shown in appendix)
- **Distributions**: The distribution of the incidents over time is not uniform, where some periods and regions show dense concentration of events.

To solve these data quality concerns, I listed the following solutions:

- **Normalization**: To account for non-uniform distribution across different regions and times.
- **Outlier analysis**: to check and see whether they are to be included in the analysis or not.

4 Analysis

4.1 Approach

In this following section, the tasks that were carried out to answer the research questions will be discussed. The goal is to prove that the repeated use of visuals and other techniques might help find the answer.

Before starting the analysis, the data was pre-processed, this is a very important step to insure higher computational accuracy which on the long run would result in better human judgement due to better visualizations.

Task 1: Evolution of Terrorist Activities over time.

Computational

To assess the frequency, severity, and spread of terrorist activities over time, a time series analysis was put in place. Which means that would be aggregated annually and across various geographical regions.

Visual

To show these trends and patterns, line charts will be generated. Also, Heatmaps will be created to show the severity of terrorist activities over different regions. This all should be interpreted by human judgement to understand the peaks and troughs correlating them to historical events.

Task 2: Changes in Tactics, Targets and lethality

Computational

To analyze the changes in tactics and targets, descriptive statistics and natural language processing (NLP) will be used. Casualty data will be used to assess lethality.

Visual

To illustrate these changes bar charts and a scatter plot with

Task 3: Analysis of Terrorist group activities

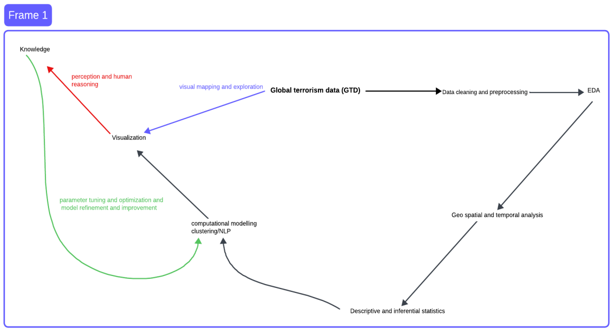
Computational

To understand the unique behavioral patterns of known terrorist groups, a cluster analysis will be applied to group terrorist incidents. Some feature engineering will be done beforehand as well.

Visual

Horizontal Bar chart will be used to show group profiles, how many times they committed an incident and compare different terrorist group profiles. In addition to that, bubble charts may also be used to show the scale of activities across a number of incidents and to show the geographical spread of each group.

Each task will be completed iteratively, and the visual results will be feeding back into computational analysis. Human reasoning will also be applied where the visualizations will be used to inform and refine computational methods. The end results will be combined to offer a thorough response to the research questions, utilizing both human understanding and computer power. (The following figure shows the analysis workflow plan).


Frame 1
Knowledge
perception and human reasoning
visual mapping and exploration
Global terrorism data (GTD)
Data cleaning and preprocessing
EDA
Visualization
parameter tuning and optimization and model refinement and improvement
computational modelling clustering/NLP
Geo spatial and temporal analysis
Descriptive and inferential statistics

4.2 Process

In this process section the goal is to show the 3 tasks that were carried out where each task aimed to answer one of the 3 research questions:

1. How have terrorist activities evolved over time in terms of frequency, severity, and geographical spread?
2. Have there been significant changes in the lethality of terrorist attacks?
3. Which terrorist groups have been the most active, and how do their activities differ in terms geographical focus?

Task 1

A time series and geospatial analysis were carried out to understand the evolution of terrorist incidents. After first reviewing the data, the focus was on fatalities ('nkill') and injuries ('nwound') as indicators for the severity of the incident. This step ensured the data's suitability for deeper temporal and spatial assessment.

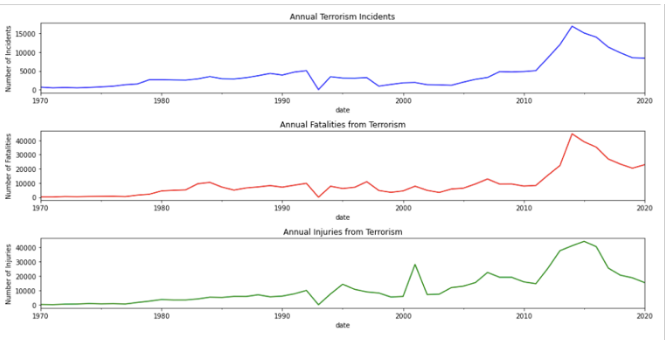


Figure 2: Line charts

The first line chart shows how frequent terrorist attacks happened annually. There is a rising trend in incidents over time, with very obvious peaks that might correlate with significant historical events. To understand the causative factors behind these peaks, like geopolitical changes or emergence of new terrorist groups, it requires very careful interpretation.

Moreover, the second line chart represents the annual fatalities resulting from terrorism, which indicates how severe the attacks were. It is a fluctuating trend with a lot of escalation, especially in recent years which shows that the severity of attacks is also increasing. This graph emphasizes the necessity for a more thorough examination of the underlying causes of the increase in severity.

The final line chart displays the number of injuries which is also an indicator of the severity of the attacks. Similar to the previous line chart, a detailed examination of the trends of these injuries can be an indicator for a change in weapons or tactics over time.

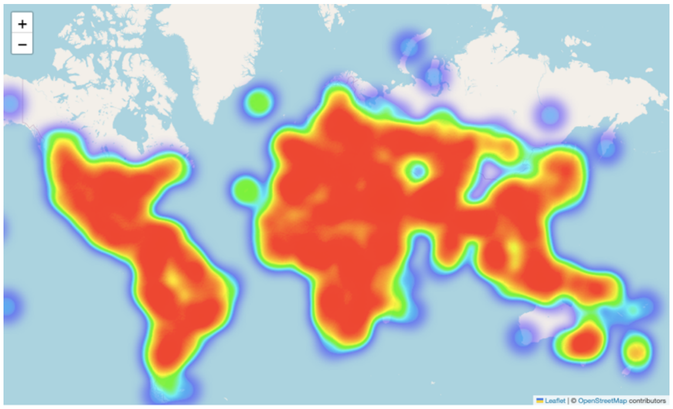


Figure 3: Heatmap of geographical spread of terrorism

The heatmap visualizes the geographical spread of the attacks and shows terrorism hotspots have expanded by highlighting regions with higher concentration of attacks (as the map is zoomed into, it gets more specific). This heatmap is essential for identifying areas that have been overly targeted which begs the questions about the changes in terrorist operations and regional conflicts.

Human judgement was used to analyze these figures and make observations. Patterns found in the visualizations provided guidance on how to proceed with the study, and the logic behind but also tools to refine the analysis and guide further inquiries. When additional analysis failed to show new insights, the iterative process was stopped which indicated that a comprehensive understanding was reached.

Task 2

After the dataset was curated, time-series computational methods were used to determine the baseline frequencies of terrorist attacks. In addition, a natural language processing (NLP) was also used to see how strategies and targets changed over time. The NLTK and scikit-learn libraries were utilized to convert the unstructured textual input into a structured format which made it possible to identify important terms that provided insight into the changing vocabulary of terrorism.

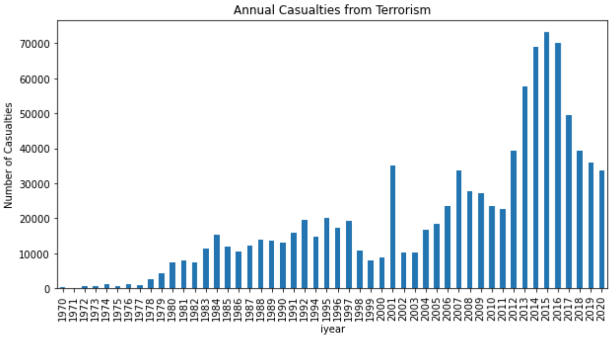


Figure 4: Bar charts

For this task, the first figure constructed was a bar chart which showed the casualties from terrorism year by year, providing a clear visual representation of lethality trends over the years. This bar chart served as a foundational visualization that allowed me to observe that while there were some fluctuations but there was an overall increase in the casualties, which is really noticeable in the latest years, this also indicated that the lethality of these attacks is increasing as well.

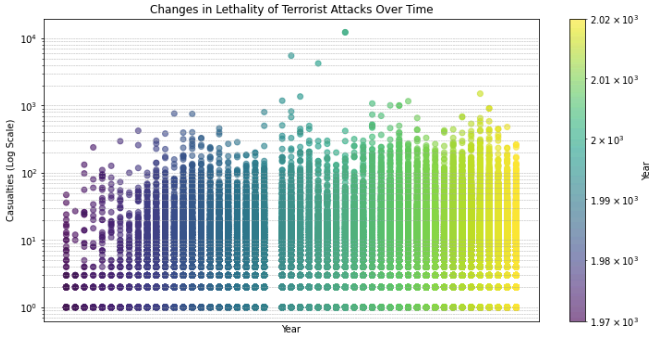


Figure 5: scatterplot

Moreover, I then progressed to a more granular analysis with the scatter plot, that uses a logarithmic scale to represent the casualties per incident over time. This approach allowed me to visualize the spread of data points and identifying outliers, which most of the time represent the most devastating attacks. The color representation by year gave me a visual cue of the temporal distribution.

indeed evolved and discernably increased. both the plots together illustrate a story on how terrorism evolved, from the increasing frequency and severity of the attacks to the change in nature of tactics. These finding allowed me to find a robust answer to the second research question and helped in bettering my understanding in the realm of terrorism.

Finally, the stopping condition was reached when no further refinement could have been applied to the model which would alter the overarching conclusions. Through an iterative process of testing and validation against known historical data, the use of parameters, such as the TF-IDF feature limit in the NLP model, was justified.

Task 3

The analysis method used to profile the activities of terrorist groups was multimodal, integrating human analysis with data-driven computational techniques. The goal of the entire procedure was to analyze terrorist organizations' activity, paying close attention to most common groups and their geospatial distribution.

The data's innate groupings were found through the use of KMeans clustering. The elbow method and silhouette analysis were used in a methodical manner to identify the ideal number of clusters. In order to provide unique group profiles, the procedure attempted to increase inter-cluster variance while decreasing intra-cluster variance. The Features that were standardized for the cluster analysis were the tactics which was represented by the column (‘attack1’), targets which was represented by column (‘targettype1’), and weapons which was represented by column (‘weaptype1’).

PCA was used to reduce dimensionality while maintaining the structure of the dataset in order to improve the clustering performance and visualization. The dataset was greatly reduced to principle components, which capture the majority of the variance.

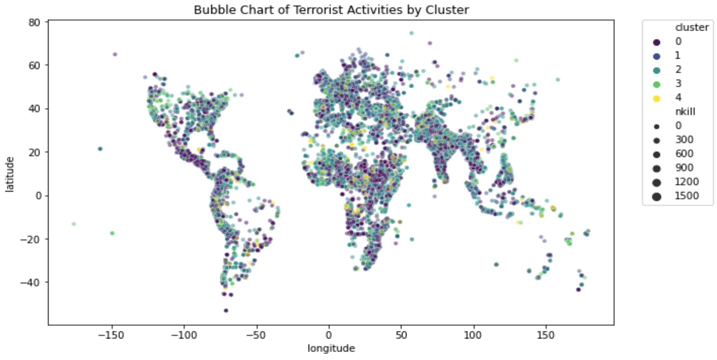


Figure 6: bubble chart of the different clusters

Using latitude and longitude as axes, a bubble chart was used to show the spatial distribution of terrorist activities. Each bubble size corresponds to the number of incidents, providing a clear visual representation of the concentration and impact of these activities. Each color represented a cluster from the kMeans clustering algorithm.

Clusters of larger bubbles in specific region may indicate

Differentiating the clusters by color aids in identifying patterns cross different terrorist groups. Distinct clusters in specific region may suggest that there is a presence of a local group that have specific operations and tactics.

Through examining he overlap of some bubbles, this may hint at potential alliances between terrorist groups or vice versa. This chart can help in making strategic decisions such as increasing security from international security agencies.

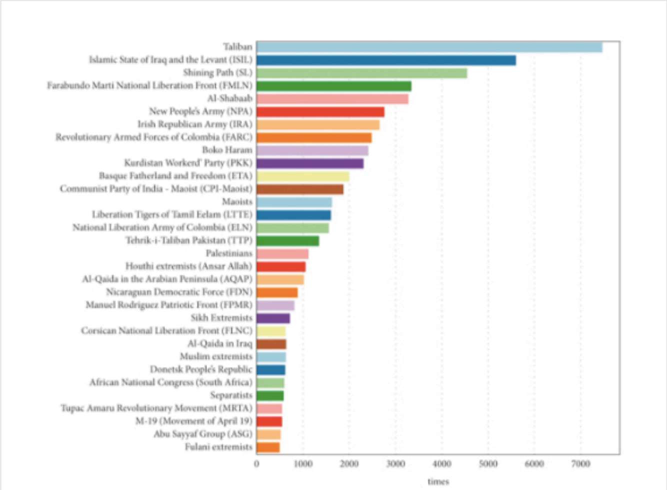


Figure 7: Bar chart of different terrorist groups

The bar chart shown in the previous figure represents each terrorist group by name and how often they were involved in an incident. Each group has a specific color with the most being the first at the top. As seen in the figure, the Taliban which are located in southern Afghanistan, were the highest terrorist group with over than 7000 incidents.

A thorough qualitative and visual understanding of terrorist group activities was the analysis's end product. The bubble map offered a broad perspective of the regional impact and terrorist groups, while the bar chart gave took a more informative approach to show the specific terrorist groups and the amount of incidents they have committed. This study's question could be effectively answered by the effective interpretation of complicated, multi-dimensional data, made possible by the analytical rigor that this thorough approach guaranteed.

4.3 Results

To examine the frequency and severity of terrorist acts, a time series analysis was used. Patterns in the distribution and intensity of terrorism were found through the use of geospatial mapping and annual aggregation of occurrences. Heatmaps showed regional hotspots that indicated the geographical spread of the incidents, while line charts made clear an increasing trend in the number of incidents. Through the combination of visual and algorithmic elements, we were able to answer the first research question with a yes they have increased in frequency, severity and spread over time.

In order to assess the lethality, statistical analysis combined with visual analytics revealed a trend of increasing death toll over time. An increase in attack severity was confirmed by bar charts that showed data on casualties annually as well as the scatterplot

This implied a concerning rise in assault lethality, probably due to improvements in armament or changes in tactics.

The most active terrorist organizations and their distinct operational patterns were found using clustering methods. Insights into each group's strategies and local impact were provided by bubble charts displaying the scope of activity and bar charts comparing group profiles. This investigation shed light on the different profiles of groups.

5 Critical reflection

An introspective assessment of the study's techniques, graphics, and cognitive processes is offered by this reflection of the analytical approach to comprehending terrorist activities. Human reasoning was crucial in determining the course of the investigation and navigating the complex nature of the data throughout the analysis. For example, the initial choice to do a time series analysis sprang from the realization that terrorism is a dynamic phenomenon whose evolution must be represented in a dynamic manner.

The visual representations created during this analysis were essential tools for reasoning rather than just outputs. They made it possible to convert incomprehensible numbers into comprehensible patterns, which improved human comprehension and made it possible to combine disparate data points into a compelling story. The dependence on images, however, also highlighted their drawbacks because they could only depict the dimensions that were intended to be shown, thereby hiding other important details that were not immediately apparent.

Some assumptions were made throughout the investigation, such as the idea that the number of occurrences that are reported is the same as the frequency at which terrorist actions actually occur. This assumption overlooks potential discrepancies in reporting and data collection across regions and time. Although informative, the cluster analysis was very dependent on the number of clusters and the variables used, which were subjective choices that may have a big impact on the result.

Using more sophisticated natural language processing (NLP) methods, including sentiment analysis or topic modeling, could have provided deeper understanding of the qualitative components of the data when considering what could have been done differently. Interactive visual analytics tools could have improved data exploration by enabling in-the-moment hypothesis testing and more in-depth analysis of certain incidents or trends.

The approach's limitations are apparent in its dependence on historical data, which might not take into consideration the dynamic nature of terrorism and its root causes. The socio-political circumstances of the attacks were also not thoroughly explored in the analysis, despite the fact that doing so could have shed light on the underlying causes and motivations of terrorist activity. If the data is sufficiently rich to allow for such in-depth analysis and the analysts are aware of the contextual elements that could affect the interpretation, there is hope that this approach can be applied to other datasets.

The lessons learned from this analysis emphasizes on the value of a well-rounded strategy that incorporates both human judgment flexible in my methodologies, be skeptical of easy conclusions, and always consider the broader context in which the data exists. Also, try adding more data sources and be ready to modify my strategy in reaction to new findings.

In summary, the study demonstrated the complexity of terrorism analysis while also successfully addressing the research questions. While data science can contribute to understanding in this field, human perspective is still invaluable. We should always look for innovative approaches and technological advancements that can further deepen study of this very complicated topic.

Table of word counts

Problem statement	228/250
State of the art	423/500
Properties of the data	500/500
Analysis: Approach	367/500
Analysis: Process	1251/1500
Analysis: Results	200/200
Critical reflection	499/500

References

[1] Pan, X. (2021). Quantitative Analysis and Prediction of Global Terrorist Attacks Based on Machine Learning. *Scientific Programming*, [online] 2021, p.e7890923. doi:https://doi.org/10.1155/2021/7890923.

[2] Guo, W., Liu, H., Yu, A. and Li, J. (2016). RESEARCH ON VISUAL ANALYSIS METHODS OF TERRORISM EVENTS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, [online] XLI-B2, pp.191–196. doi:https://doi.org/10.5194/isprs-archives-XLI-B2-191-2016.

[3] ieeexplore.ieee.org. (n.d.). *Visual Analytics of Terrorism Data | IEEE Conference Publication | IEEE Xplore*. [online] Available at: https://ieeexplore.ieee.org/document/7819677 [Accessed 1 Jan. 2024].

[4] www.start.umd.edu. (n.d.). *Data Collection Methodology*. [online] Available at: https://www.start.umd.edu/gtd/using-gtd/.

[5] Global Terrorism Database (2022). *Global Terrorism Database*. [online] University of Maryland. Available at: https://www.start.umd.edu/gtd/.

[6] safegraph.com. (n.d.). *Geospatial Data Analytics: What It Is, Benefits, and Top Use Cases | SafeGraph*. [online] Available at: https://www.safegraph.com/guides/geospatial-data-analytics#:~:text=Geospatial%20data%20analysis%20involves%20collecting.

[7] Sharma, P. (2019). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

[8] Tableau. (n.d.). *Time Series Analysis: Definition, Types, Techniques, and When It's Used*. [online] Available at: https://www.tableau.com/learn/articles/time-series-analysis#:~:text=What%20is%20time%20series%20analysis.