Data Warehouses – Project02

Wroclaw University of Science and Technology, Date: April 27, 2022

Student:	--	Grade
Identifier	245784	
First name	Rahul	
Last name	Vijayvargiya	

This mini project assignment consists of 1 task.

Preliminary specification of the selected project topic

1. Title of the project

Global Terrorism Database
More than 180,000 terrorist attacks worldwide, 1970-2017

2. General description of the domain

The Global Terrorism Database (GTD) is the most comprehensive unclassified database of terrorist attacks in the world. The National Consortium for the Study of Terrorism and Responses to Terrorism (START) makes the GTD available via this site to improve understanding of terrorist violence, so that it can be more readily studied and defeated. The GTD is produced by a dedicated team of researchers and technical staff.

The GTD is an open-source database, which provides information on domestic and international terrorist attacks around the world since 1970, and now includes more than 200,000 events. For each event, a wide range of information is available, including the date and location of the incident, the weapons used, nature of the target, the number of casualties, and – when identifiable – the group or individual responsible

Data analysis is to enable, among others, to establish regions of the city and country in which the highest number of Terrorist attacks happened,
they also become more and more brutal., What kind of method or weapon they used,
and Which Org has done more damage to the world and its people, we can gain insight how many people has been killed till now in those attacks.

3. Description of the analysis area (selected fragment of the domain, intended for detailed analysis and development of the data warehouse)

We will be doing a detailed analysis of the GTD Dataset b/w 1970-2017.

It will be interesting to analyze the cities where the arrests took place, the time, which org has planned these attacks, what kind of weapon they used, how many people killed until 2017 in those attacks, we will complete exploratory data analysis of Global Terrorism Dataset.

2.1 Problems

The significant problems include:

1. Attacks Increased over the years
2. More and More People has been killed
3. Type of weapon or Type of attack
4. Finding most distructive terror org which planned alot of attacks
5. Finding the trend of the root cause

2.2 Project goal

2.2.1 Expectations and needs for decision-making support (research questions)

The analytical database should enable an in-depth analysis of the facts of GTD,

Type of attack, Weapon, Time, Terror Org,

1. Are there more suicide bombing, killing with guns, or hijacking aeroplane
2. At what place highest number of attacks took place
3. Which terror org has took place most attacks
4. Which attack killed most people
5. Which year has the most attacks took place

A proper analysis of historical data should provide factual answers to the above questions and provide information from which it will be possible to draw conclusions that constitute the basis for making correct decisions

2.2.2 Scope of analysis - Aspects examined

The Scope will be analyzed under fixed circumstances such as Year, Org Type, Killed, Type of attacks, place and so on,

We should analyze furthermore the ways to stop such attacks, Predictive Analysis could help us her, In depth analysis can be done in Tableau, Python

2.2.3 Potential data warehouse users

This analytical database will support the decision-making processes and understanding the root cause for such attacks will be used by a government entities and anti-terrorism squad of respective countries

4. Source of data

4.1. Data Sources

The source data needed for the creation of a data warehouse are presented in Table 1.

Table 1. Source data sets

No	File	Type	Rec.	Size [MB]	Description
1.	Global Terrorism Database	CSV	181619	155 MB	This Dataset is about Global Terrorist Attacks Between 1970 – 2017 Details

4.2. Location, availability of source data

The data comes from the websites:

We Use Kaggle to Download the Dataset,
it’s an Open-Source dataset,
I attached following link to the dataset, attached below

https://www.kaggle.com/datasets/START-UMD/gtd

Publications:

The GTD has been leveraged extensively in scholarly publications, reports, and media articles. Putting Terrorism in Context: Lessons from the Global Terrorism Database, by GTD principal investigators LaFree, Dugan, and Miller investigates patterns of terrorism and provides perspective on the challenges of data collection and analysis. The GTD's data collection manager, Michael Jensen, discusses important Benefits and Drawbacks of Methodological Advancements in Data Collection and Coding.

4.3. Data Dictionary – Interpretation, data semantics

I renamed the columns so they can be more readable to the end user and can understand what sort data it consists in more meaningful words, I Use Python for renaming the Columns and find out which column has the most null values and cannot use

Code below for renaming columns:

```
terror_df.rename(columns={'iyear':'Year','imonth':'Month','iday':'Day',
'country_txt':'Country','provstate':'state',

                 'region_txt':'Region','attacktype1_txt':'AttackType','target1':'Target','nkill':'Killed',

                'nwound':'Wounded','summary':'Summary','gname':'Group','targtype1_txt':'Target_type',

                'weaptype1_txt':'Weapon_type','motive':'Motive'}, inplace=True)
```

```
Column approxdate has been dropped since nulls percentage is 95 %
Column resolution has been dropped since nulls percentage is 99 %
Column location has been dropped since nulls percentage is 69 %
Column summary has been dropped since nulls percentage is 36 %
Column alternative has been dropped since nulls percentage is 84 %
Column alternative_txt has been dropped since nulls percentage is 84 %
Column attacktype2 has been dropped since nulls percentage is 97 %
Column attacktype2_txt has been dropped since nulls percentage is 97 %
Column attacktype3 has been dropped since nulls percentage is 100 %
Column attacktype3_txt has been dropped since nulls percentage is 100 %
Column targtype2 has been dropped since nulls percentage is 94 %
Column targtype2_txt has been dropped since nulls percentage is 94 %
Column targsubtype2 has been dropped since nulls percentage is 94 %
Column targsubtype2_txt has been dropped since nulls percentage is 94 %
Column corp2 has been dropped since nulls percentage is 94 %
Column target2 has been dropped since nulls percentage is 94 %
Column natlty2 has been dropped since nulls percentage is 94 %
Column natlty2_txt has been dropped since nulls percentage is 94 %
Column targtype3 has been dropped since nulls percentage is 99 %
Column targtype3_txt has been dropped since nulls percentage is 99 %
Column targsubtype3 has been dropped since nulls percentage is 99 %
Column targsubtype3_txt has been dropped since nulls percentage is 99 %
Column corp3 has been dropped since nulls percentage is 99 %
Column target3 has been dropped since nulls percentage is 99 %
Column natlty3 has been dropped since nulls percentage is 99 %
Column natlty3_txt has been dropped since nulls percentage is 99 %
Column gsubname has been dropped since nulls percentage is 97 %
Column gname2 has been dropped since nulls percentage is 99 %
Column gsubname2 has been dropped since nulls percentage is 100 %
Column gname3 has been dropped since nulls percentage is 100 %
Column gsubname3 has been dropped since nulls percentage is 100 %
Column motive has been dropped since nulls percentage is 72 %
Column guncertain2 has been dropped since nulls percentage is 99 %
Column guncertain3 has been dropped since nulls percentage is 100 %
Column nperps has been dropped since nulls percentage is 39 %
Column nperpcap has been dropped since nulls percentage is 38 %
Column claimed has been dropped since nulls percentage is 36 %
Column claimmode has been dropped since nulls percentage is 89 %
Column claimmode_txt has been dropped since nulls percentage is 89 %
Column claim2 has been dropped since nulls percentage is 99 %
Column claimmode2 has been dropped since nulls percentage is 100 %
Column claimmode2_txt has been dropped since nulls percentage is 100 %
Column claim3 has been dropped since nulls percentage is 100 %
Column claimmode3 has been dropped since nulls percentage is 100 %
Column claimmode3_txt has been dropped since nulls percentage is 100 %
Column compclaim has been dropped since nulls percentage is 97 %
Column weaptype2 has been dropped since nulls percentage is 93 %
Column weaptype2_txt has been dropped since nulls percentage is 93 %
Column weapsubtype2 has been dropped since nulls percentage is 94 %
Column weapsubtype2_txt has been dropped since nulls percentage is 94 %
Column weaptype3 has been dropped since nulls percentage is 99 %
Column weaptype3_txt has been dropped since nulls percentage is 99 %
Column weapsubtype3 has been dropped since nulls percentage is 99 %
Column weapsubtype3_txt has been dropped since nulls percentage is 99 %
Column weaptype4 has been dropped since nulls percentage is 100 %
Column weaptype4_txt has been dropped since nulls percentage is 100 %
Column weapsubtype4 has been dropped since nulls percentage is 100 %
Column weapsubtype4_txt has been dropped since nulls percentage is 100 %
Column weapdetail has been dropped since nulls percentage is 37 %
Column nkillus has been dropped since nulls percentage is 35 %
Column nkillter has been dropped since nulls percentage is 37 %
Column nwoundus has been dropped since nulls percentage is 36 %
Column nwoundte has been dropped since nulls percentage is 38 %
Column propextent has been dropped since nulls percentage is 65 %
Column propextent_txt has been dropped since nulls percentage is 65 %
Column propvalue has been dropped since nulls percentage is 79 %
Column propcomment has been dropped since nulls percentage is 68 %
Column nhostkid has been dropped since nulls percentage is 93 %
Column nhostkidus has been dropped since nulls percentage is 93 %
Column nhours has been dropped since nulls percentage is 98 %
Column ndays has been dropped since nulls percentage is 96 %
Column divert has been dropped since nulls percentage is 100 %
Column kidhijcountry has been dropped since nulls percentage is 98 %
Column ransom has been dropped since nulls percentage is 57 %
Column ransomamt has been dropped since nulls percentage is 99 %
Column ransomamtus has been dropped since nulls percentage is 100 %
Column ransompaid has been dropped since nulls percentage is 100 %
Column ransompaidus has been dropped since nulls percentage is 100 %
Column ransomnote has been dropped since nulls percentage is 100 %
Column hostkidoutcome has been dropped since nulls percentage is 94 %
Column hostkidoutcome_txt has been dropped since nulls percentage is 94 %
Column nreleased has been dropped since nulls percentage is 94 %
Column addnotes has been dropped since nulls percentage is 84 %
Column scite1 has been dropped since nulls percentage is 36 %
Column scite2 has been dropped since nulls percentage is 58 %
Column scite3 has been dropped since nulls percentage is 76 %
Column related has been dropped since nulls percentage is 86 %
```

4.4. Qualitative assessment of data:

The result of the qualitative analysis carried out with the use of Python, Library we used here is called Pandas,

First, We read the Csv file, convert it to a Pandas Data Frame, Later, I called Functions like Df.Describe, Df.dtypes, Df.Info and So on

To get the in-depth Qualitative Assessment of data

Data Source: globalterrorism.csv	Shape Of records (181691, 18)

Column	Dtype	Meaning	Remark
Year	int64	Year of attack	Integer Number of year
Month	int64	Month of attack	Contain Month in Integer
Day	int64	Day of attack	Contain Day in Int
Country	object	Country of attack	Country as String Object
state	object	State where it occurs	Store State as String Object
Region	object	Region of attack	Store Region as String
city	object	City Where it took place	Store City as String
latitude	float64	Latitude of the attack	Store Float Value of Latitude
longitude	float64	Longitude of the attack	Store Float Value of Longitude
AttackType	object	Attack type eg: bomb or gun attack	Attack type as String Object
Killed	float64	People killed in attack	Killed People store as a Float Data Type
Wounded	float64	People who are injured in attack	Wounded People Store as a Float data type
Target	object	Target of attack	String Object of taget
Summary	object	Summary or attack	String Object of Summary
Group	object	Terror org name	String object of Group
Target_type	object	Target type	String object of Target Type
Weapon_type	object	Weapon used	Weapon Type as a string object
Motive	Object	Motive of attack	Motive for the cause of attack is a string object

Columns	Null Values
Year	0
Month	0
Day	0
Country	0
state	421
Region	0
city	434
latitude	4556
longitude	4557
AttackType	0
Killed	10313
Wounded	16311
Target	636
Summary	66129
Group	0
Target_type	0
Weapon_type	0
Motive	131130

5. Multidimensional analytical models

5.1. Facts subject to analysis and their measures

Table 4. Facts subject to analysis

Lp.	Fakt	Measure(s)	Remarks
1.	Fact_Group	Group	Terrorist Group
2.			
...	...		

5.2. Facts analysis context

Table 5. Facts Analysis Dimensions

No.	Dimension	Properties, characteristics

1.	Time	Datetime Object, It Enable us to Time Series Analyze of Terror Attack
2.	Country	String Object, It Enable us to analyze the the country and Region of attack
3.	Attack	String Object, Type of attack,
4.	Damage	Int Object, Number of people killed
5.	Weapon	String Object, Type of Weapon Used in Attack
…	…	

5.3. Multidimensional Models (UML)

GENERAL CONCLUSIONS:

Use this section to provide your general conclusions:

I use Python for Analysis of dataset, getting insight regarding with dataset, this dataset is a time series,

We can identify data as per time and group which planned the attack, Which country has affected the most from the attacks

Identify the country with most attacks, most common way of attack, I perform Additional cleaning of the dataset, drop the columns with most null values, change the encoding of the dataset, find out the data types of each column and what data it stores, rename the columns so end use can easily understand what sort of data it consists of.

UML Diagram is Missing Just because my pc is slow, cannot open and use Visual paradigm,

REMARKS:

- *A report without final conclusions will not be checked and results in a negative score!*
- *The report file should be named **Proj01DW-StudentID-Last name-2022**, please use the PDF format*
- *You should use MS SQL SERVER 2019 (or 2017), Visual Studio and Tableau Prep (available at https://www.tableau.com/academic/students)*