Capstone Project-1

EDA on Global Terrorism Dataset

- Soma Pavan Kumar(Pro Flex)
- spkumar1998@gmail.com

Contents

Problem Statement

On The Given Global Terrorism Dataset, Explore and analyze the data to discover key findings pertaining to terrorist activities.

To do Exploratory Data Analysis on the global Terrorism Data and key findings pertaining to Terrorist Activities.

Introduction

The Global Terrorism Database is an open Source Database that contains information in terrorist attacks around the world from 1970 through 2017. This Database includes systematic data on domestic as well as international terrorist attacks that have occurred during this time period and includes over 180,000 different entries and 135 different features describing the entries.

This 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.

Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- Our Dataset include more than 130 columns out of which more than 50 columns are empty and containing Data which are of no importance to our analysis.
- So, we have decided to remove these columns one by one from the DataFrame using the drop function

After cleaning the snippet of data looks like:

```
df.head()
```

	eventid	iyear	imonth	iday	country	country_txt	region	region_txt	provstate	city	latitude	longitude	summary	crit1	crit2	crit3	doubtterr	alternative	alternati
0	197000000001	1970	7	2	58	Dominican Republic	2	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	NaN	1	1	1	0.0	NaN	
1	197000000002	1970	0	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	NaN	1	1	1	0.0	NaN	
2	197001000001	1970	1	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	NaN	1	1	1	0.0	NaN	
3	197001000002	1970	1	0	78	Greece	8	Western Europe	Attica	Athens	37.997490	23.762728	NaN	1	1	1	0.0	NaN	
4	197001000003	1970	1	0	101	Japan	4	East Asia	Fukouka	Fukouka	33.580412	130.396361	NaN	1	1	1	-9.0	NaN	

The df.head() method shows us the first 5 rows of the Dataset.

Let us look at some statistics of the Data

- The Statistics of the data could be found out from an inbuilt function in pandas library called describe().

```
[5] df.describe()
```

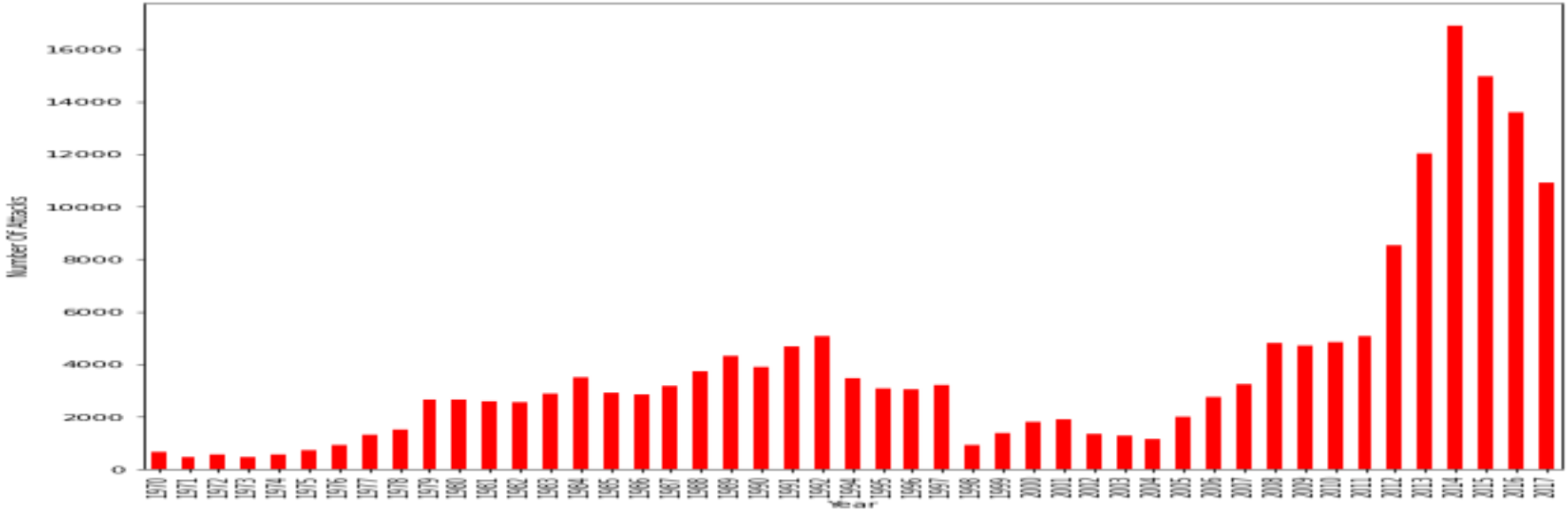
	eventid	iyear	imonth	iday	extended	country	region	latitude	longitude	specificity	vicinity	crit1	1816
count	1.816910e+05	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000	177135.000000	1.771340e+05	181685.000000	181691.000000	181691.000000	
mean	2.002705e+11	2002.638997	6.467277	15.505644	0.045346	131.968501	7.160938	23.498343	-4.586957e+02	1.451452	0.068297	0.988530	
std	1.325957e+09	13.259430	3.388303	8.814045	0.208063	112.414535	2.933408	18.569242	2.047790e+05	0.995430	0.284553	0.106483	
min	1.970000e+11	1970.000000	0.000000	0.000000	0.000000	4.000000	1.000000	-53.154613	-8.618590e+07	1.000000	-9.000000	0.000000	
25%	1.991021e+11	1991.000000	4.000000	8.000000	0.000000	78.000000	5.000000	11.510046	4.545640e+00	1.000000	0.000000	1.000000	
50%	2.009022e+11	2009.000000	6.000000	15.000000	0.000000	98.000000	6.000000	31.467463	4.324651e+01	1.000000	0.000000	1.000000	
75%	2.014081e+11	2014.000000	9.000000	23.000000	0.000000	160.000000	10.000000	34.685087	6.871033e+01	1.000000	0.000000	1.000000	
max	2.017123e+11	2017.000000	12.000000	31.000000	1.000000	1004.000000	12.000000	74.633553	1.793667e+02	5.000000	1.000000	1.000000	

- The above picture shows that the describe method gives median , mode, count , max and other functions. Also as median and mode are close enough , this dataset can be considered as a Normal distribution.

Let us begin with visualizations

1. Lets plot the most attacks according to their years

```
# we will be plotting the number of terrorist attack against the year of the attack, using a bar plot
df.groupby(['iyear']).size().plot(kind='bar',color='r')
plt.rcParams["figure.figsize"] = (15,10)
plt.xlabel('Year')
```

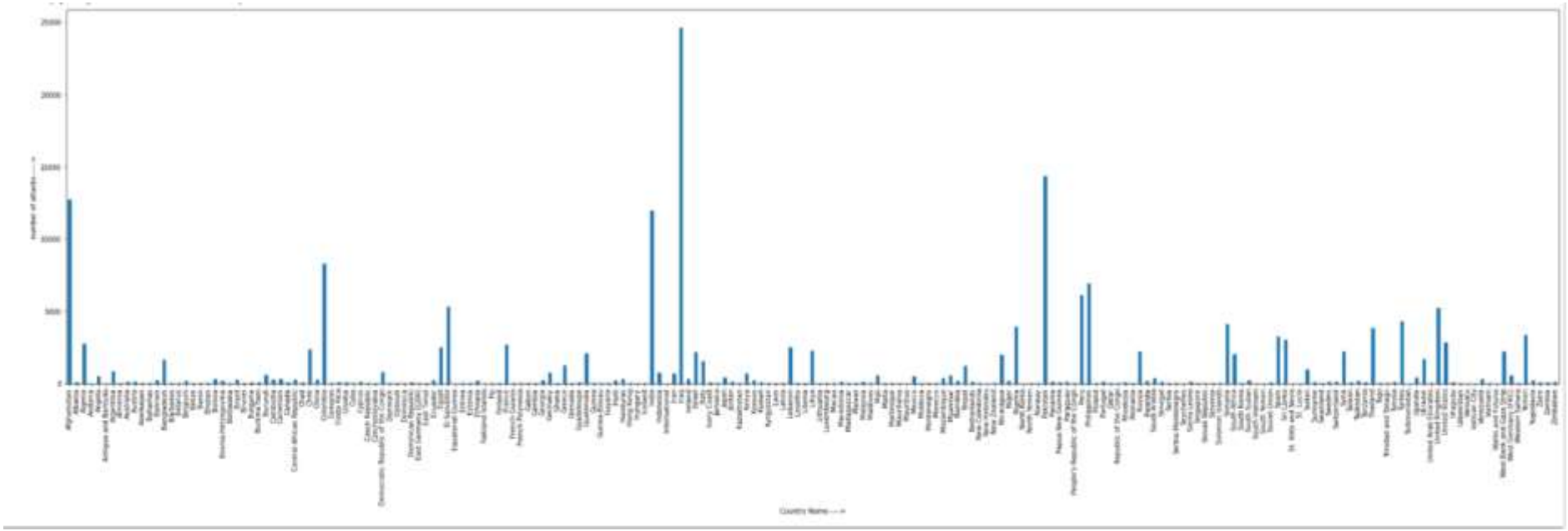


- The Visualizations are done using the inbuilt library matplotlib and the kind of plot used is “Bar Graph” .

- The Graph tells us that there has been a significant increase in the number of terror attacks from 1970 till 2014 where it reached the peak. But there was a sharp increase after 2007 till 2014 and there has been a decreasing trend since 2014 till 2017.

- In the year 2014 there has been over 16,000 attacks all over the world and the minimum number of attacks was in the year 1973 with less than 1000 attacks.

2. Plotting countries with number of attacks



- It is a bit hard to interpret the country names from the above graph. But, from the graph it can be made out that Iraq is at the top with approximately 25000 attacks, accounting for 14 % of all the attacks all over the world. And It is the country which has dealt with a lot of violence.

- Instead of analyzing the graph we can just print the countries with most number of attacks from the Data Frame by using the following code:

```
[ ] df_new=df.groupby(['country_txt']).count()
```

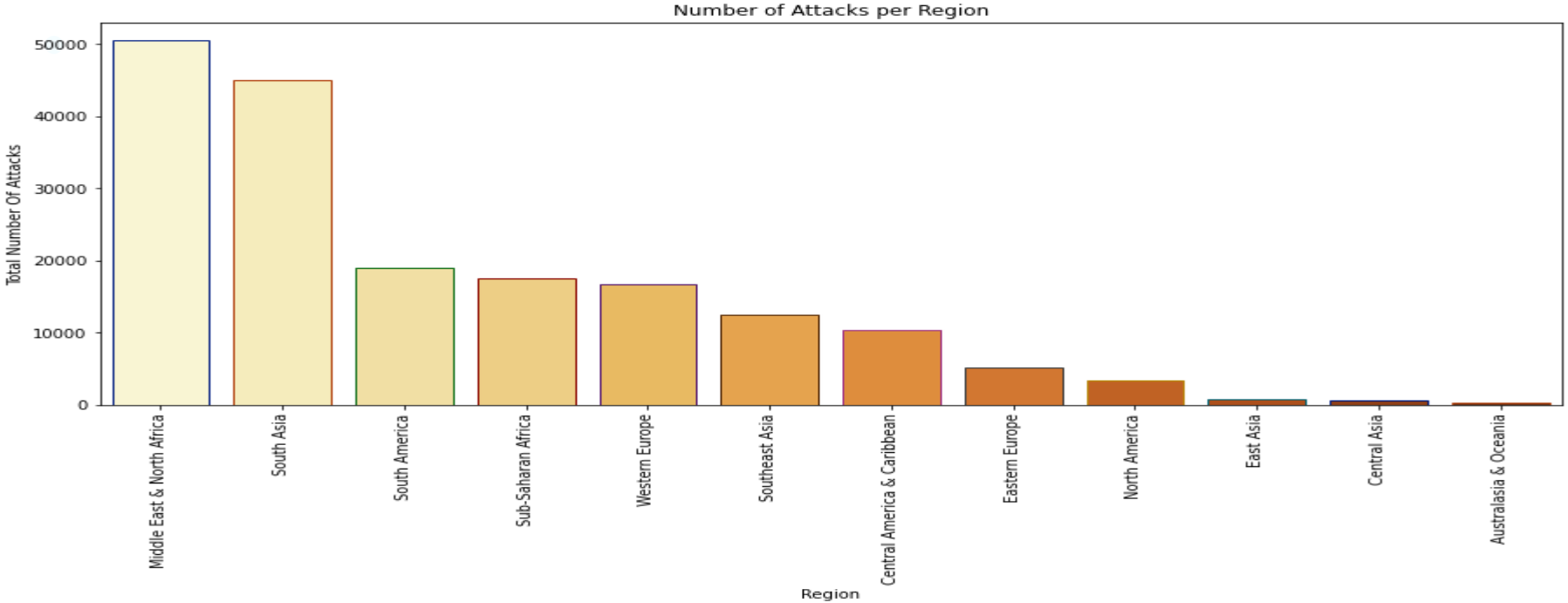
```
df_new=df_new.sort_values(['eventid'],ascending=False)
df_four=df_new['eventid']
df_four.head()
```

```
country_txt
Iraq            24636
Pakistan        14368
Afghanistan     12731
India           11960
Colombia         8306
Name: eventid, dtype: int64
```

- We have created a new Dataframe which is grouped by the variable ‘country_txt’ which holds the name of the countries and then sorted in descending order.

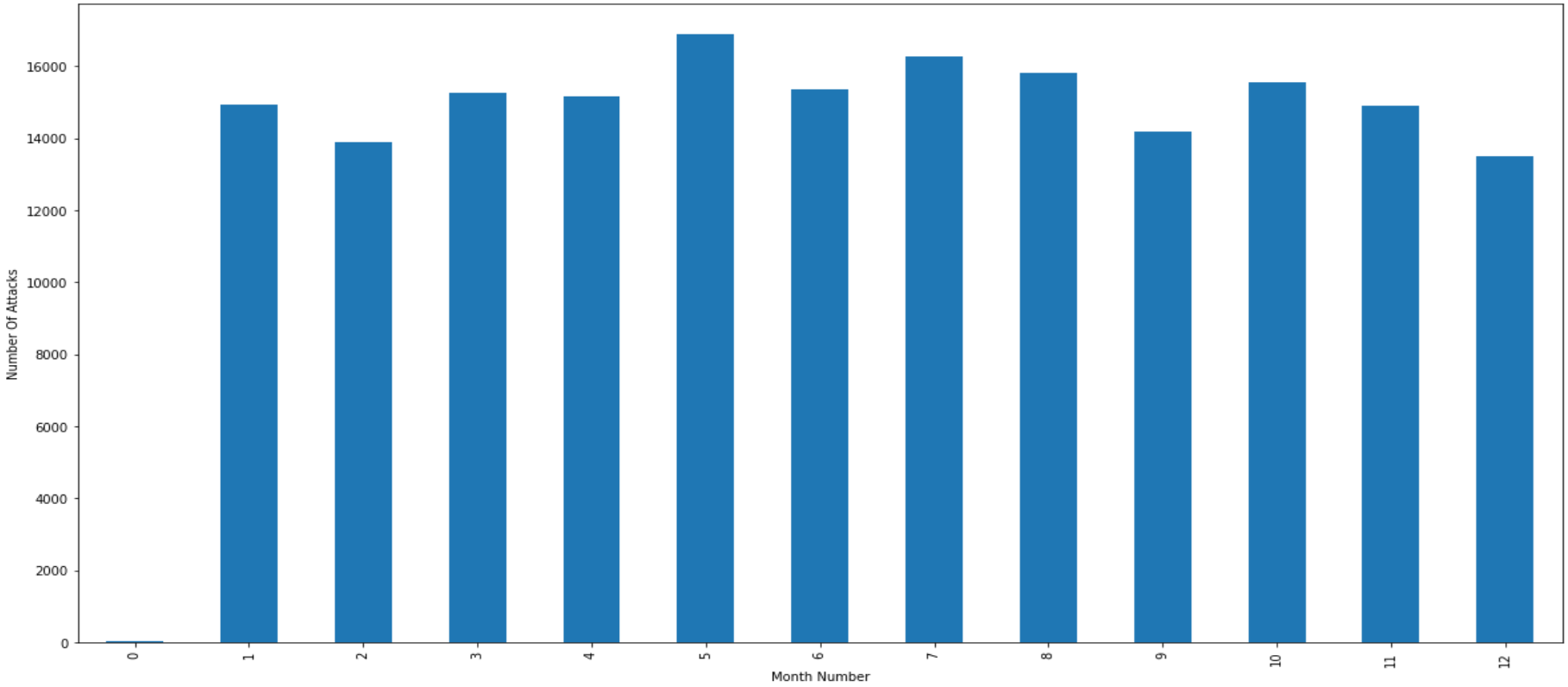
- The result of this cell is the top 5 countries listed by the number of terror attacks , India stands 4th in that list. The first place is for Iraq with approx. 25000 hits.

Let us visualize Region Wise Terror Attacks



The region that is attacked the most is Middle East and North Africa, because of Iraq and other North African nations. The second most attacked region is South Asia because of Pakistan and Afghanistan

3. Plotting frequency of attacks in their respective months



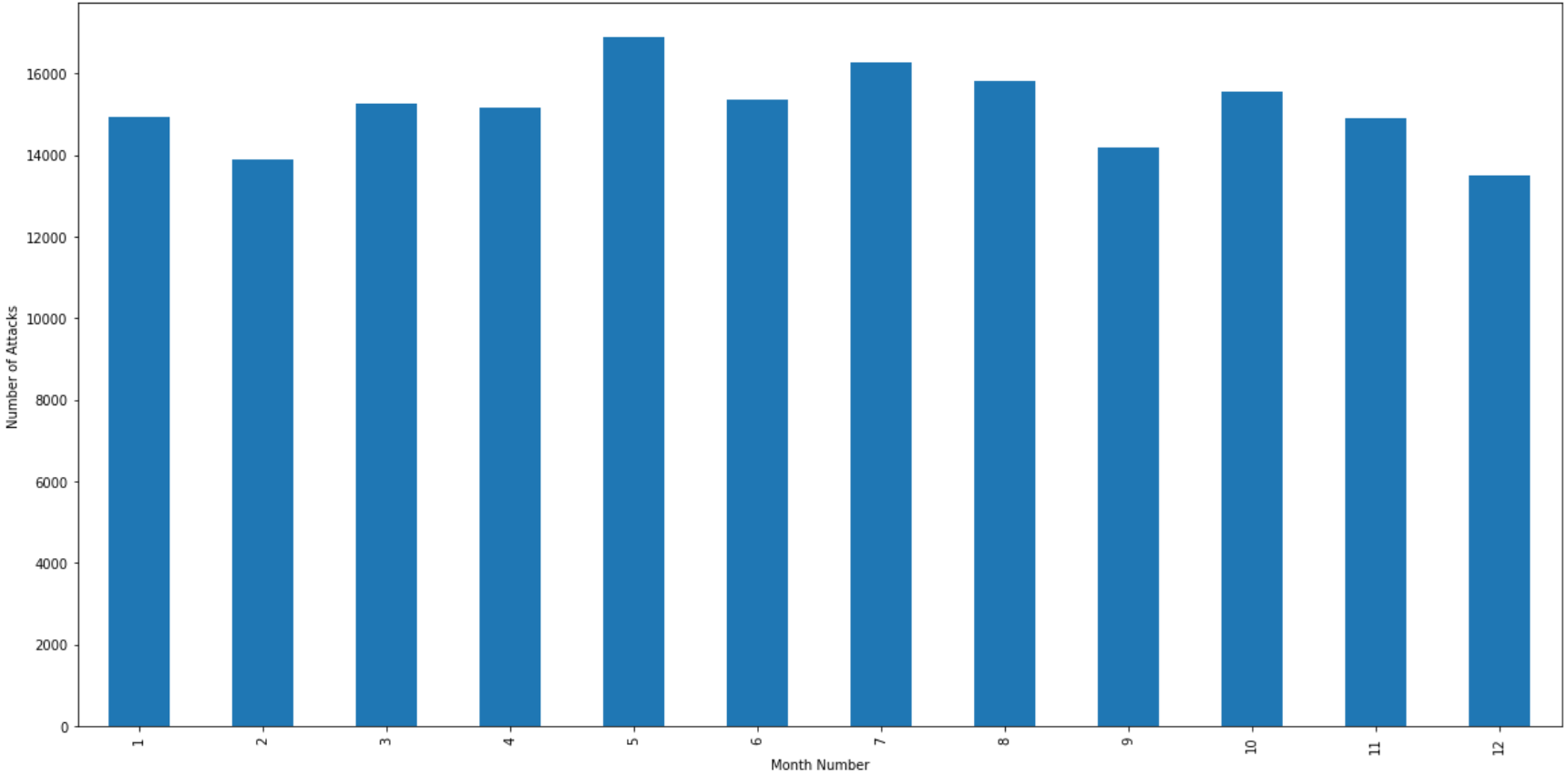
- The method used for plotting is the same as plotting for countries. But with a change in the variable name.

- We notice a slight discrepancy in the data as there is a value for 0^{th} month which is an error and can be rectified as :

```
[ ]  df_new_filtered=df[df['imonth']>=1]
```

- This line of code solves the discrepancy found earlier .

The rectified data can be plotted as:

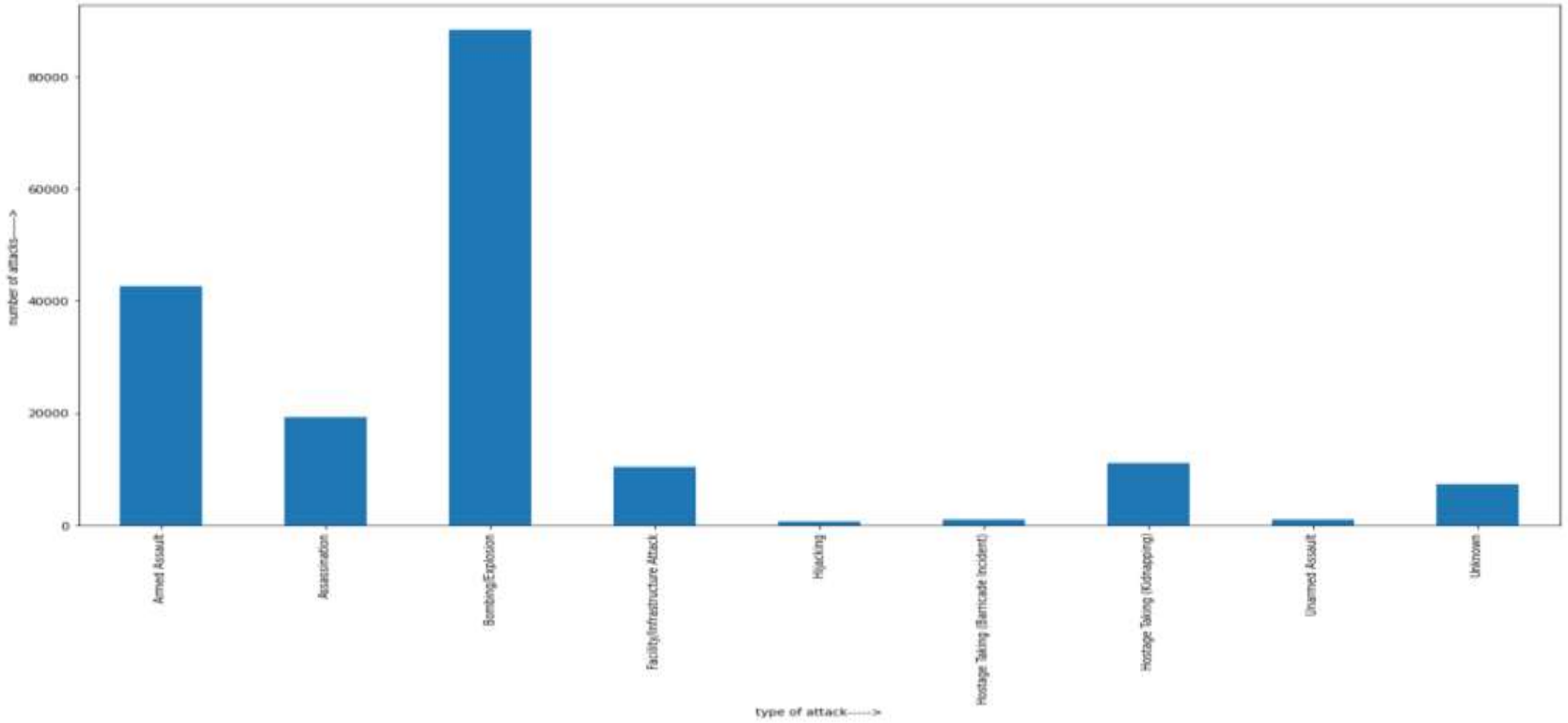


- Analysing the previous chart, we find out that there is no relation between the months and attacks happening

- It can be confirmed that Terrorists do not wait for any seasons or particular time or date for their attacks.

- Although the month of may has seen unusually large number of attacks surpassing the mark of 16,000 attacks a month.

4. Let us see what are the type of attacks used by terrorists and in what quantity

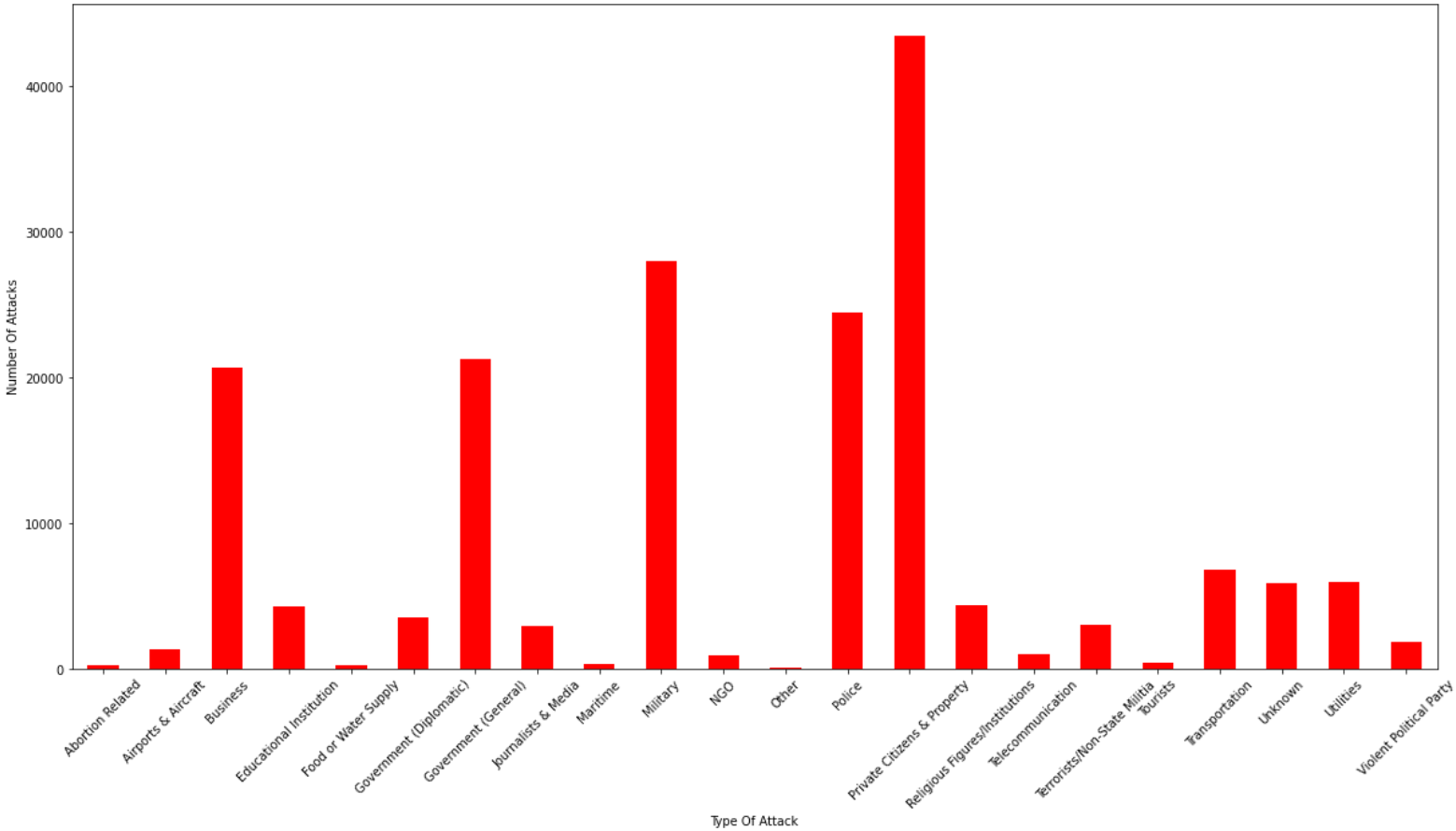


- We can see that there are various types of attacks used by the terrorists, they include Assassination , Bombing/Explosion, Hijacking, Taking Hostage and other types of attacks.

- The previous graph plot tells us that, Terrorists prefer Bombing/Explosion method for their attacks and next on their list is armed assaults(like Assault Rifles, Pistols) .

- Hijacking is least preferred by the terrorists.

Plotting the various targets of terrorists



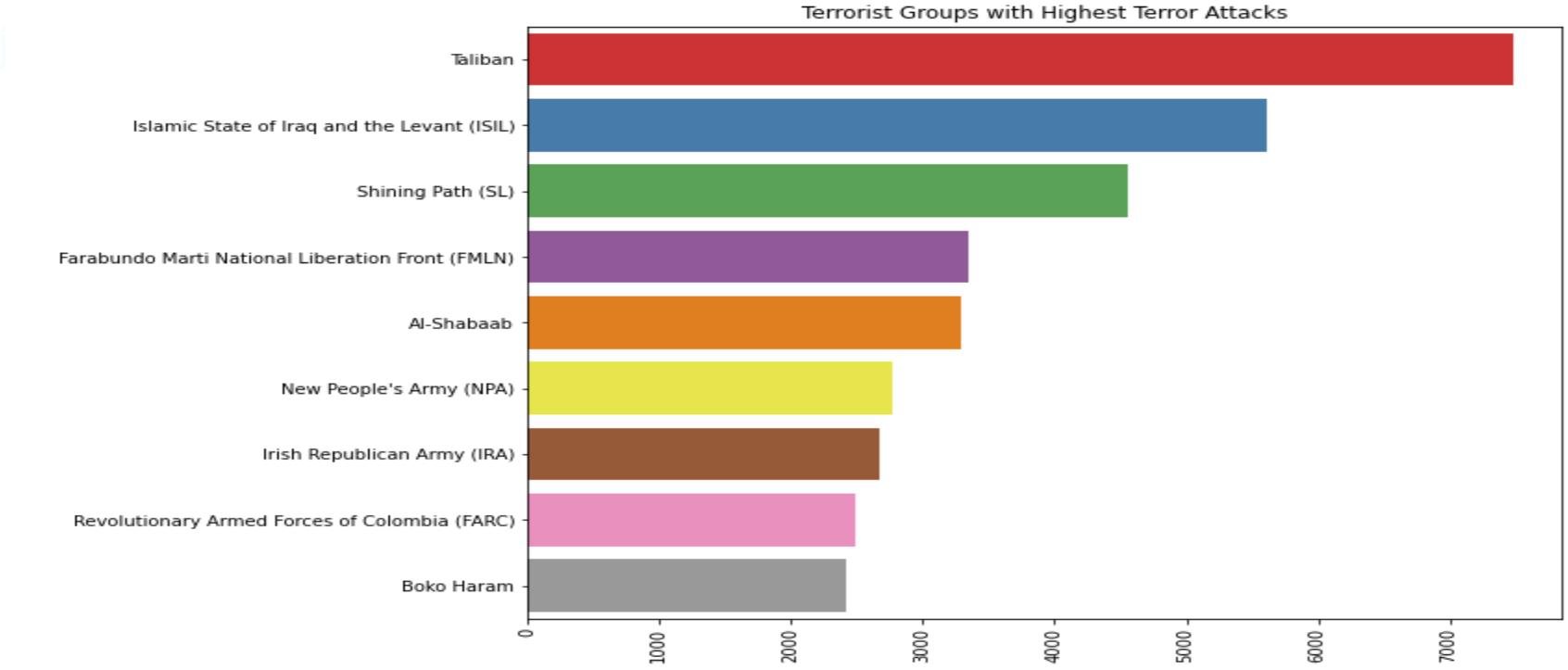
Here are some of the conclusions that we can make by looking at the above graph:

- Private Citizens are the most targeted by terrorists, sometimes private citizens are killed/assasinated in targeted knife attacks and also if that private citizen is an influentual person, then bombings are also carried out.

- Military and Police are second and third most targeted categories.

- Business and Diplomatic missions are the fourth and fifth most targeted categories.

Organizations that commit the most number of terror attacks



Word Cloud to check which city is affected the most

- The above two visualizations the first one is a graph which shows the terrorist organizations as per the number of attacks.
- As per the above graph, Taliban is the organization responsible for the most number of terrorist attacks.
- ISIL is at the second spot followed by The Shining Path.
- The next visualization is a Word Cloud. The word cloud displays the words in such a way that their size is in accordance with their frequency. The words that occur the most, have the largest size.
- This visualization is used to check which city is targeted the most. Baghdad is the most targeted city, followed by San Salvador.

Conclusion

- The Dataset chosen to Explore is the data on Global Terrorism. The entries in the data range from the year 1970 till late 2017.

- As seen from all the visualizations there has been an slight increasing trend in the number of terror attacks from 1970 onwards. But there was a slight dip from 1984 till 2004.

- After 2004 the number of attacks see a sharp rise till 2014 and we have seen a decreasing trend since 2014.

- The input parameter that does not affect the number of terror attacks is the month of the year, regardless of what month of the year it is the number of terror attacks remain almost same. Hence, the month of the year or seasonal changes cannot be used to predict the number of terror attacks.
- We also saw the most preferred style of attack for any terrorist is Bombing/Explosion, second most preferred style of attack being Armed Assault(rifles) followed by Assasination.
- Terrorists mostly attack Private citizens which can include any known public figure like a politician, influencer, social worker, etc.. They also mostly target businesses and government assets like government embassies and Military and Police

- One downside of this data is that most of the reasons why these attacks happen are political and cannot be determined by most of Machine Learning Algorithms, as the political situations can change drastically in a country or region, and it also influenced by a lot of vested interest groups that are not under the radar of many National Investigation Agencies.
- But none the less, a lot of conclusions can be made from the type of terror attacks to the targets of terror attaks and also about the most vulnerable regions in the world to terror attacks.