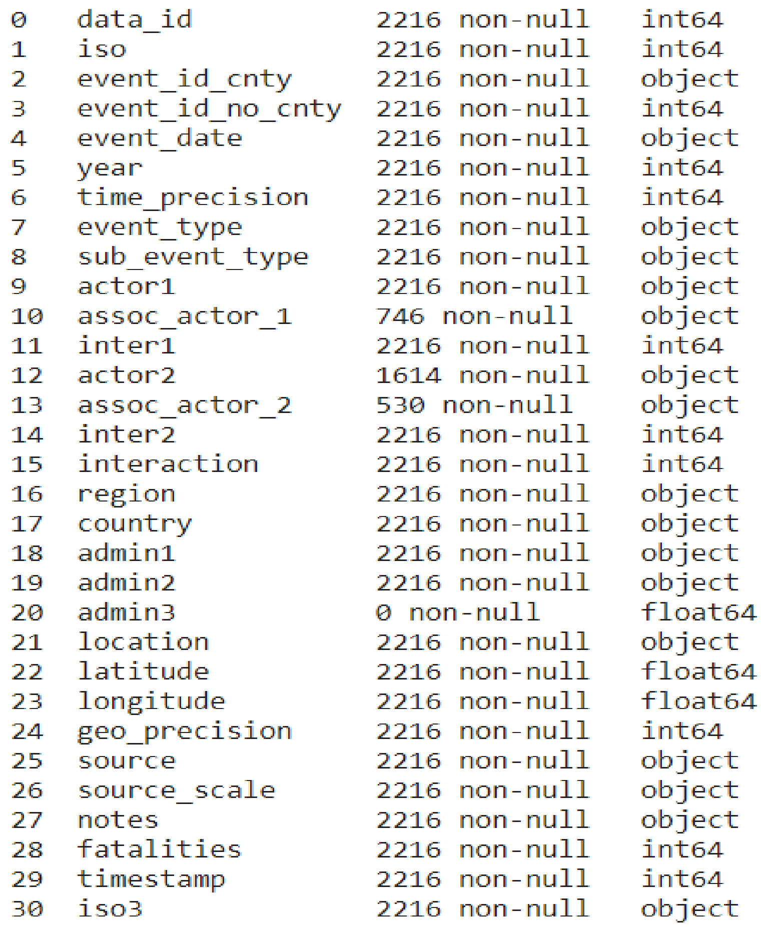
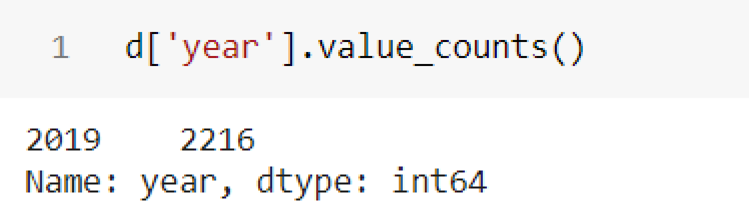
Question 1: What trends and patterns can you find when analyzing the ACLED data provided? How did you Analyze the Data? Provide a (very) brief description of the conflict dynamics for the time period / unit of analysis according to your findings.

The provided ACLED(Armed Conflict Location and Event Data) dataset contains the information about conflict events in Nigeria in 2019.Firstly,I need to get overall information about the data and which features contain missing values.



```
0   data_id          2216 non-null   int64
1   iso              2216 non-null   int64
2   event_id_cnty    2216 non-null   object
3   event_id_no_cnty 2216 non-null   int64
4   event_date       2216 non-null   object
5   year             2216 non-null   int64
6   time_precision   2216 non-null   int64
7   event_type       2216 non-null   object
8   sub_event_type   2216 non-null   object
9   actor1           2216 non-null   object
10  assoc_actor_1    746 non-null    object
11  inter1           2216 non-null   int64
12  actor2           1614 non-null   object
13  assoc_actor_2    530 non-null    object
14  inter2           2216 non-null   int64
15  interaction      2216 non-null   int64
16  region           2216 non-null   object
17  country          2216 non-null   object
18  admin1           2216 non-null   object
19  admin2           2216 non-null   object
20  admin3           0 non-null      float64
21  location         2216 non-null   object
22  latitude         2216 non-null   float64
23  longitude        2216 non-null   float64
24  geo_precision    2216 non-null   int64
25  source           2216 non-null   object
26  source_scale     2216 non-null   object
27  notes            2216 non-null   object
28  fatalities       2216 non-null   int64
29  timestamp        2216 non-null   int64
30  iso3             2216 non-null   object
```

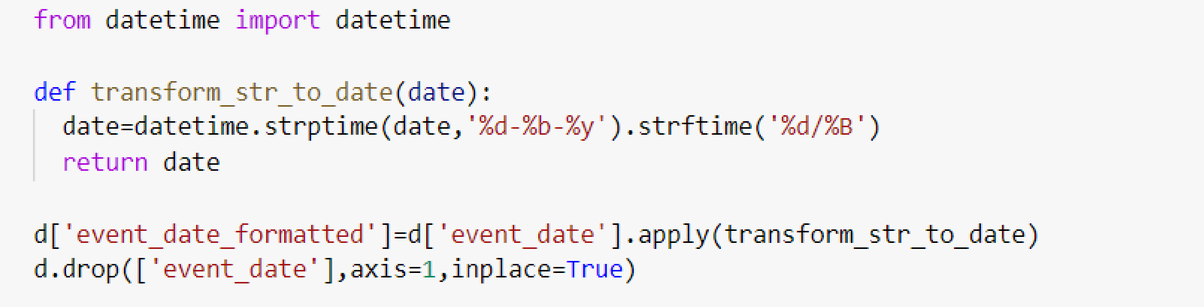
Assoc_actor_1,**actor2**,**assoc_actor_2** contain missing values,while **admin3** doesn’t contain any non-null value.It makes sense to get rid of **admin3** feature,as it doesn’t provide any informational value.After this,it needs to be checked that all events happened only in 2019,as ACLED PDF states.



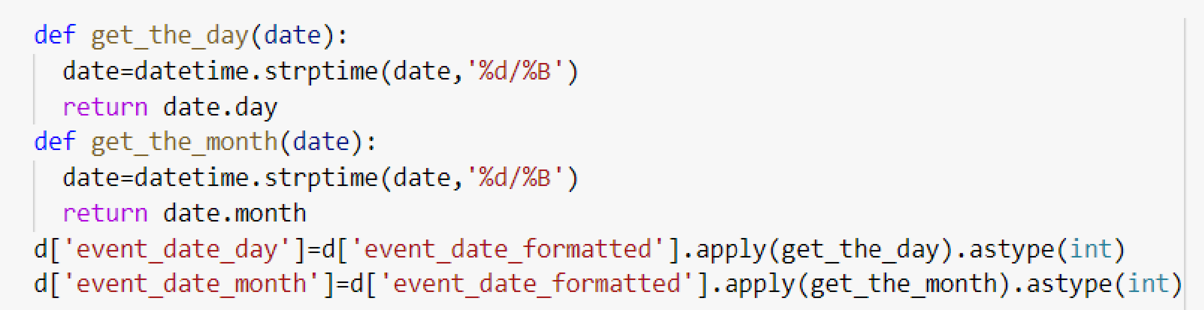
```
d['year'].value_counts()
```

```
2019    2216
Name: year, dtype: int64
```

Let’s do some feature engineering here.Event_date feature’s type is string,and not datetime.So,this feature needs to be transformed into datetime type,and then 2 new features should be created:event_date_day and event_date_month,that display only the day and the month,respectively.

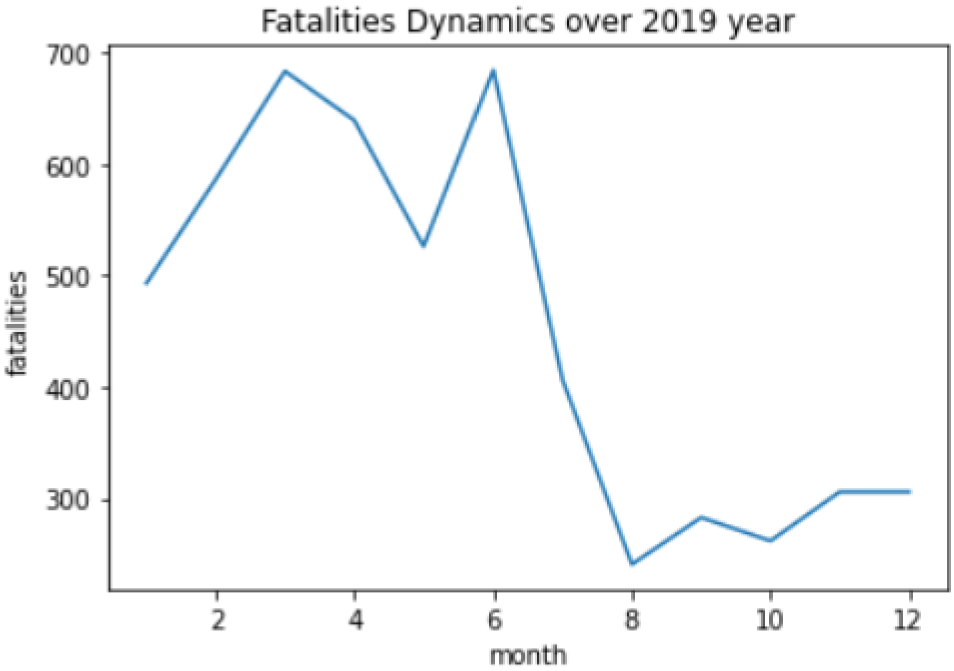


```
from datetime import datetime

def transform_str_to_date(date):
  date=datetime.strptime(date,'%d-%b-%y').strftime('%d/%B')
  return date

d['event_date_formatted']=d['event_date'].apply(transform_str_to_date)
d.drop(['event_date'],axis=1,inplace=True)
```



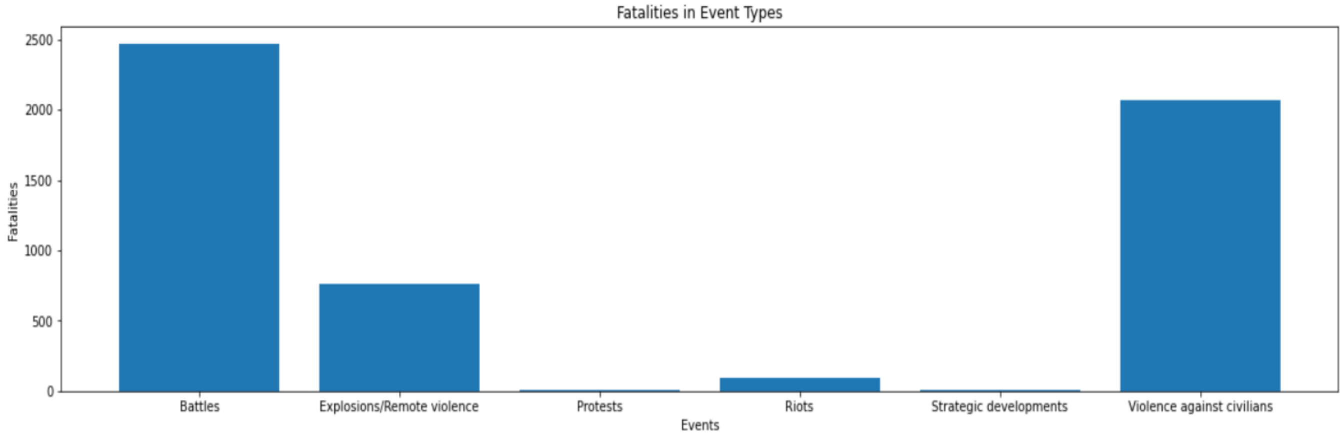
```
def get_the_day(date):
  date=datetime.strptime(date,'%d/%B')
  return date.day
def get_the_month(date):
  date=datetime.strptime(date,'%d/%B')
  return date.month
d['event_date_day']=d['event_date_formatted'].apply(get_the_day).astype(int)
d['event_date_month']=d['event_date_formatted'].apply(get_the_month).astype(int)
```

Let’s get the first pattern from our data by using a simple linear plot.Fatalities and the date are the first things that came to my eyes.



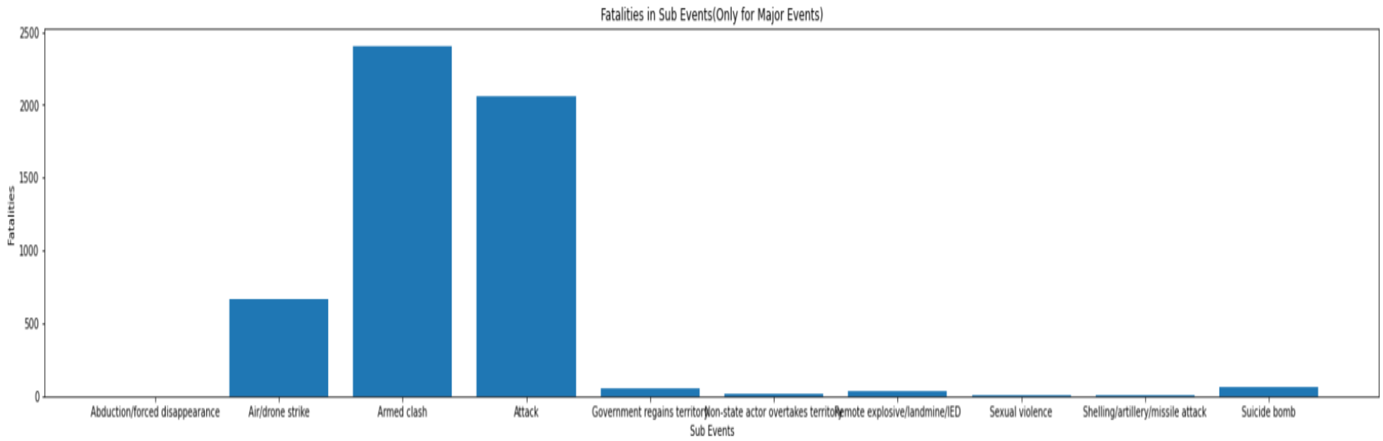
As it can be seen,the number of fatalities dramatically increases until March,and then slowly drops until it reaches its lowest position in August.The only exception is June,where fatalities suddenly increase,but then,the pattern of dropping continues.After August,fatalities start to increase,but not significantly.The plot seems simple,but we already gained a lot of information.

The next thing that needs to be known is the events and how they contribute to the fatalities.



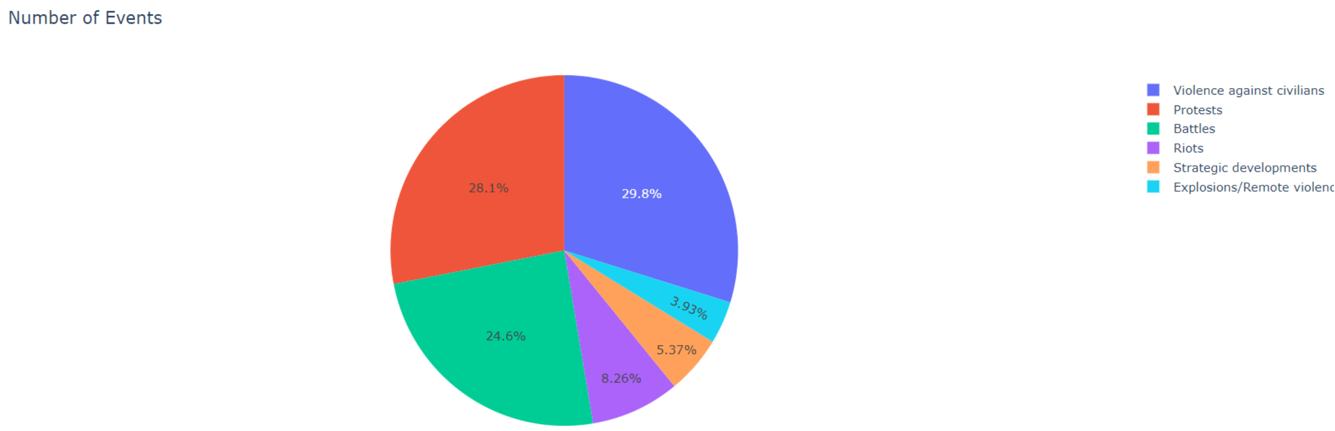
Battles are the most deadly event types with **violence against civilians** on the second place.**Protests** and **strategic developments** barely have any fatality.However,riots do have nearly 100 fatalities over a year.

Battles,Explosions/Remote Violence,Violence against civilians are the events,that are needed to be focused more,as they have significant amount of casualties.



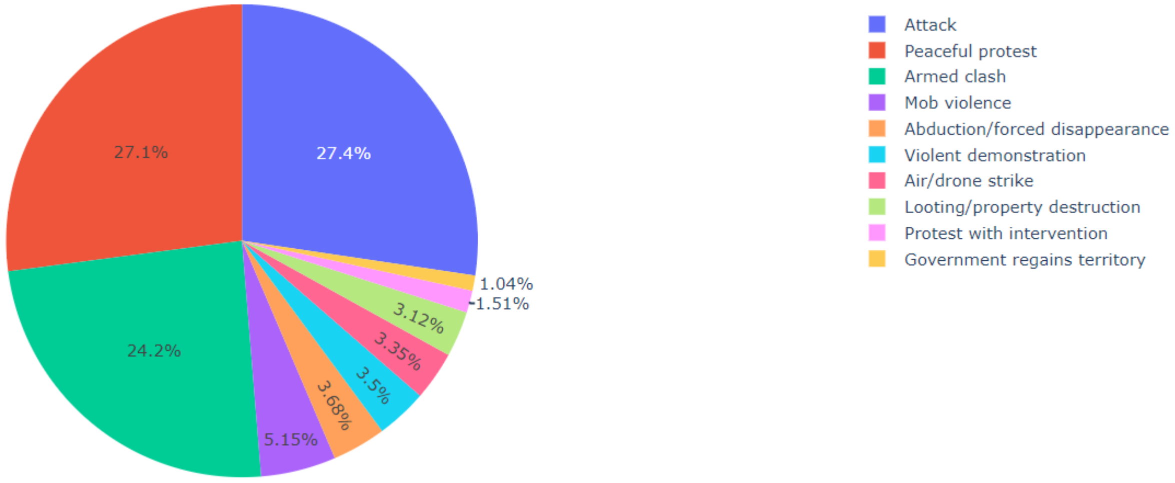
By analysing the sub-event types,that belong to the major events(**Battles,Explosions/Remote Violence,Violence against civilians**),it can be noticed,that **armed clashes** and **attacks** are the most dangerous types of sub-events(which is,of course,obvious).**Air/drone strikes** are also dangerous and they are responsible for more than 500 fatalities.

Let’s check the overall number of event types(their frequency)



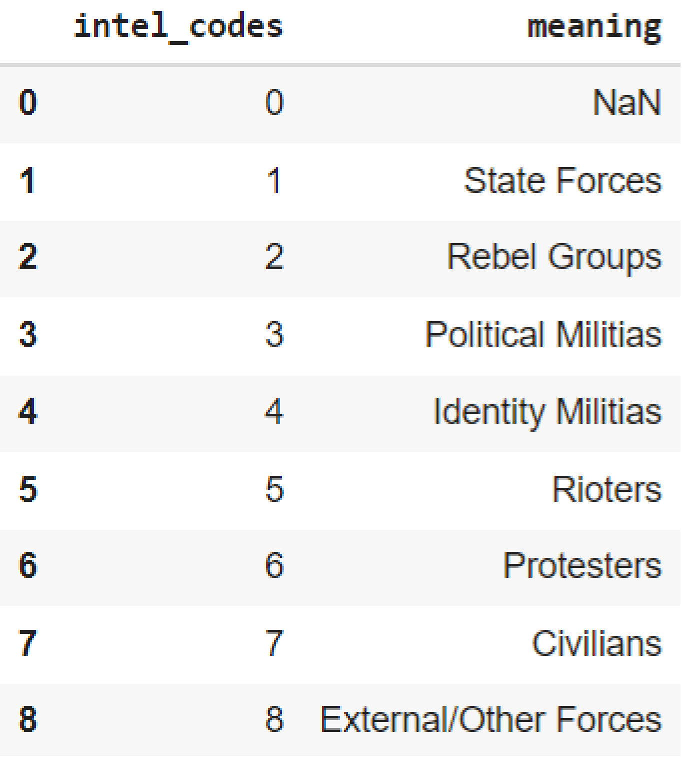
It can be stated that **protests** make up for 28.1% of the overall number of all event types.Even though there were a lot of protests going on,the number of fatalities was really low,compared to other event types.

It will be also reasonable to identify the frequency of sub-events.But,in this case,only top-10 most frequent sub-events will be included,as sub-events with really low frequency will corrupt the pie chart.



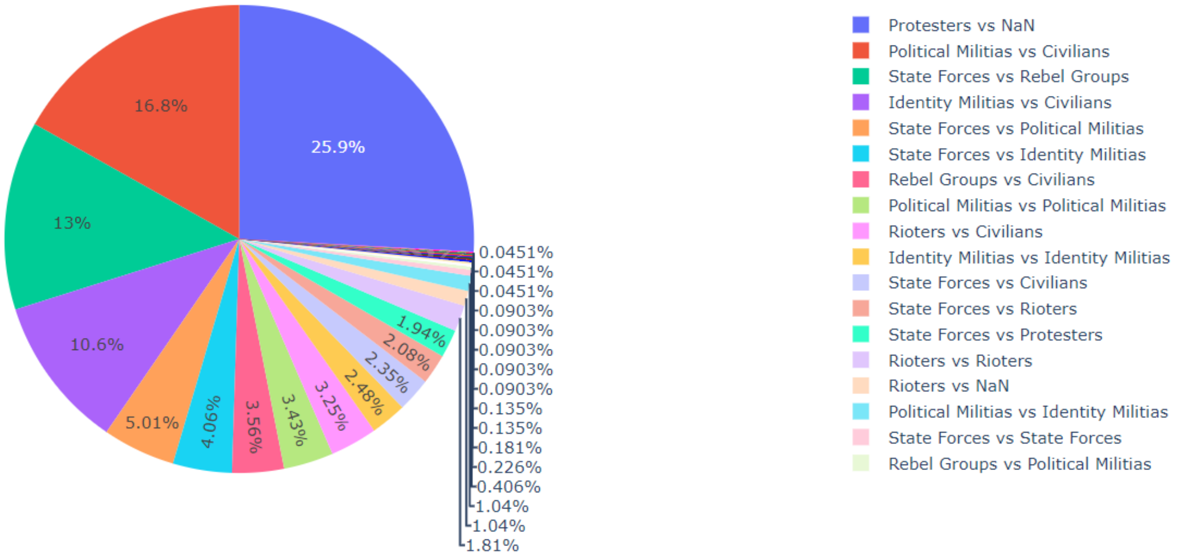
Attacks,peaceful protests and **armed clashes** are the most frequent sub-event types.

There is an interesting feature,which is called **interaction**.By looking at the PDF,I found the meaning of the Intel codes and I created a separate data frame,which contains the information for each Intel code.



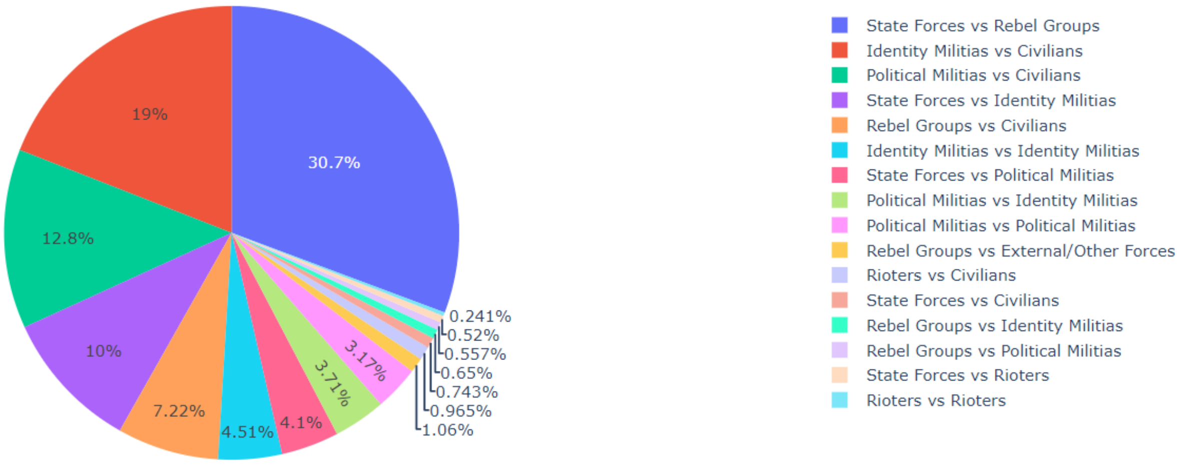
	intel_codes	meaning
0	0	NaN
1	1	State Forces
2	2	Rebel Groups
3	3	Political Militias
4	4	Identity Militias
5	5	Rioters
6	6	Protesters
7	7	Civilians
8	8	External/Other Forces

I decided to know which interactions are the most frequent and created a pie chart.



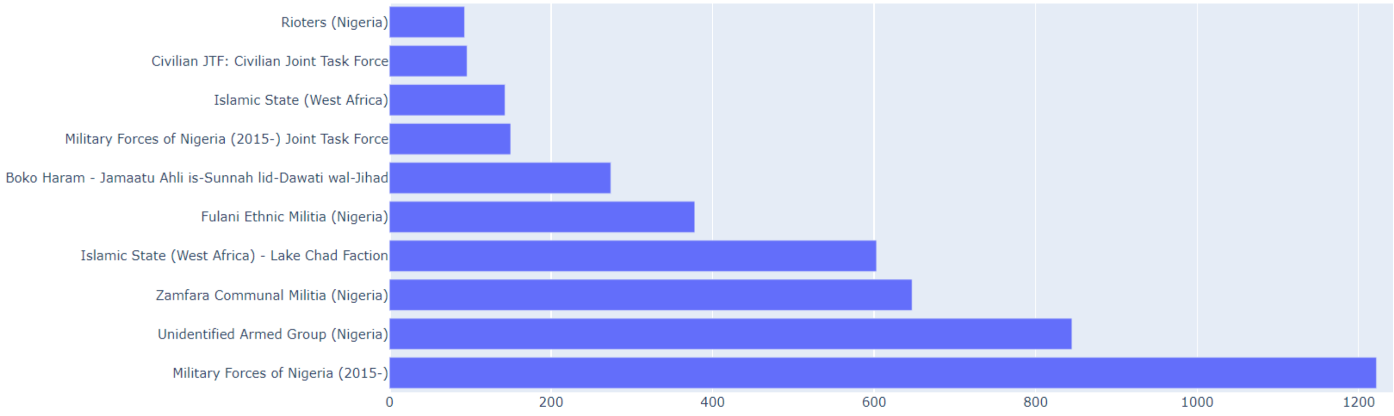
The most frequent fights were done by the protesters.16.8% of all fights were between **Political Militia and Civilians**.The interactions between **State Forces and Rebel Groups** accounted for 13% of all cases.

Let's do the same pie chart,but as a factor,let's use fatalities,not frequency.



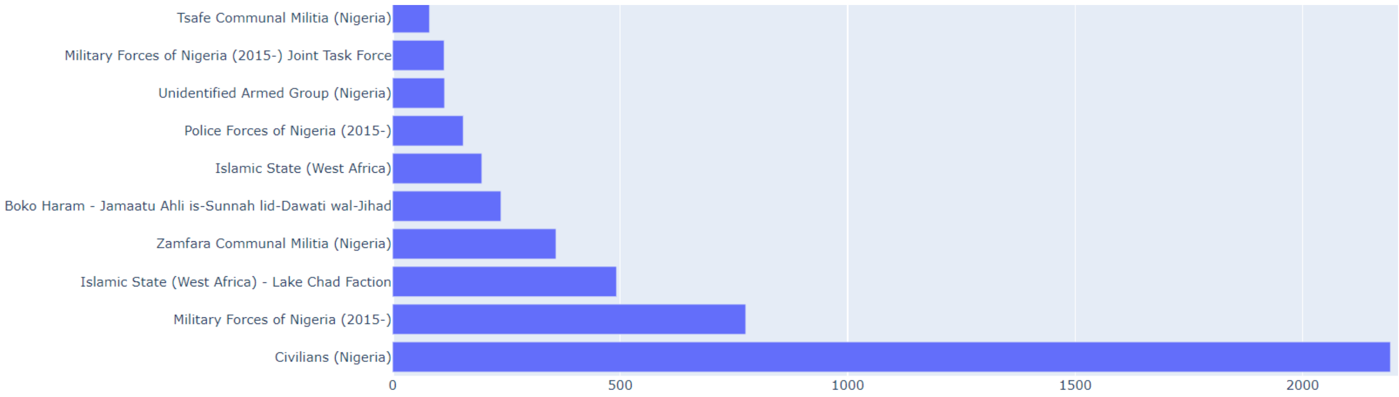
The interaction between **State Forces and Rebel Groups** is the most deadly fight.19% of all fatalities account for **Identity Militias vs Civilians**.in this case,**Identity Militias** are terrorist organizations.

Let’s view the organizations that have committed the highest number of fatalities.I will do it through analyzing **actor1** feature.



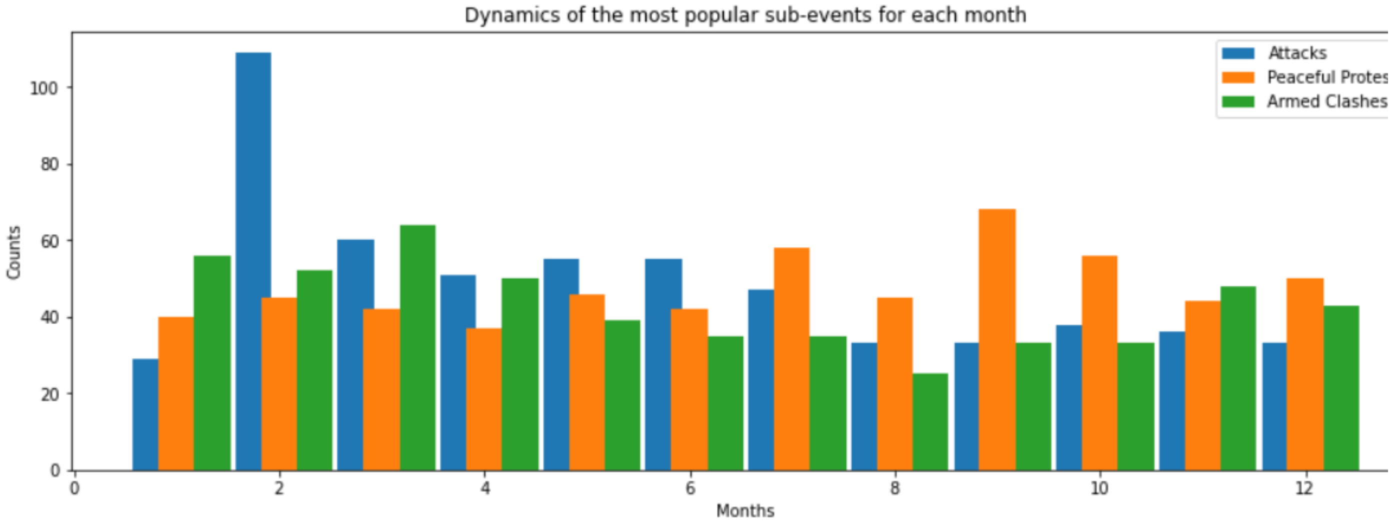
Military Forces of Nigeria committed more than 1200 casualties ,and both **Islamic State - Lake Chad Faction** and **Zamfara Communal Militia**,which are classified as **Identity Militias**,in summary,committed the same number of fatalities as **Military Forces of Nigeria**.

Let’s also consider the most suffered organizations by analyzing the **actor2** feature.

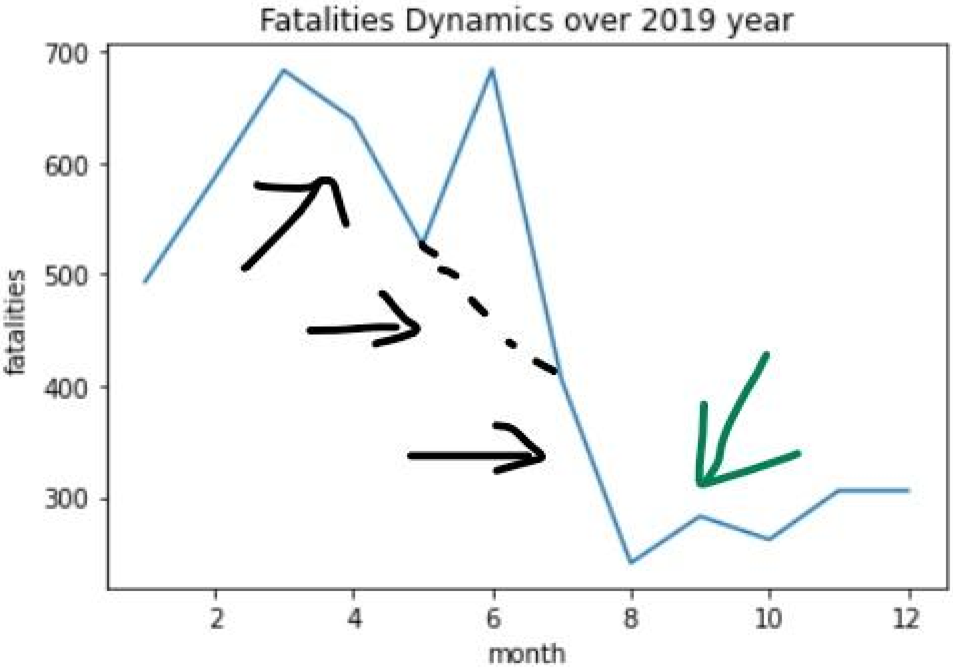


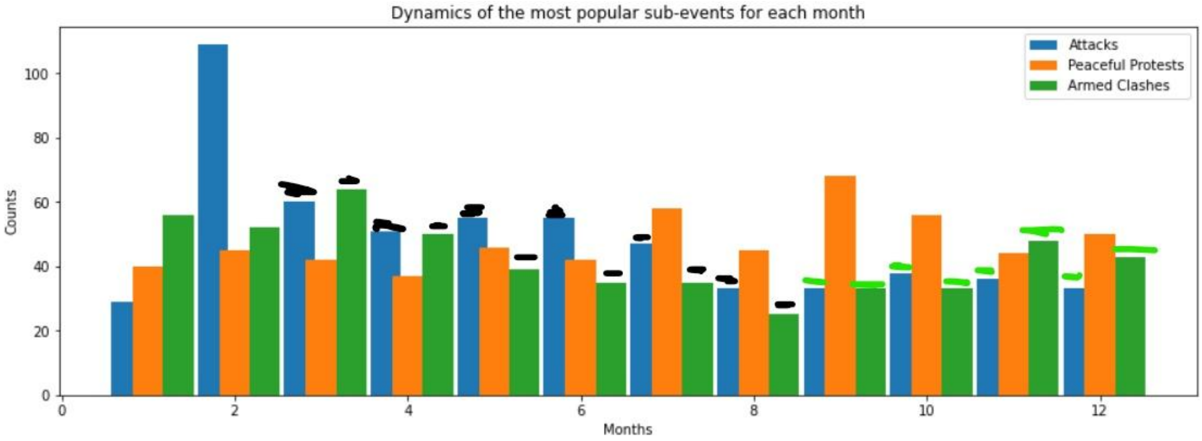
More than 2500 civilians have been killed during the 2019 year.**Military Forces of Nigeria** lost nearly 800 of its personnel.What's interesting is that **civilians** die more frequently than those who are actively engaged in military operations.

So,we already have useful information about the events,sub-events and the insights about actors.Thanks to the previous information about the most frequent sub-events,it will be really useful to know how the dynamics of these sub-events change over time.I will focus on the top-3 most frequent sub-events:**attacks,peaceful protests** and **armed clashes**.



As you can see,**February** was the most dangerous month,as it was the month with the highest number of attacks.Starting from March,**attacks** and **armed clashes** were decreasing,while the number of **peaceful protests** were increasing.After August,this pattern has been reversed.In our first linear plot,that was showing the fatalities dynamics,there was a pattern of fatalities going down after March and going up after August.

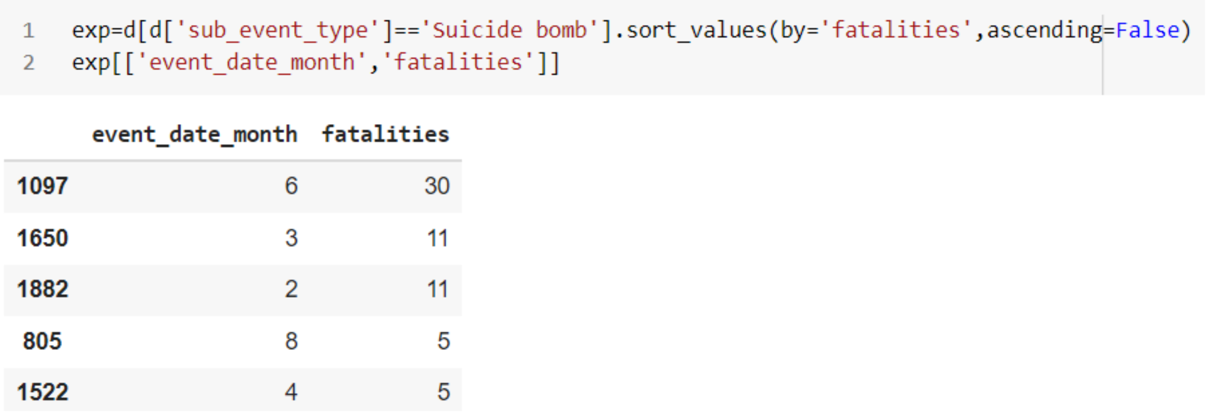




Black arrows in the first graph show how fatalities decrease(June is an exception) and the green arrow shows how fatalities start to increase.

In the second graph,black planes,which are located on the top of attacks and armed clashes,show the pattern of decreasing,while the green planes show the pattern of increasing.
There is a high correlation between these two types of information.However,in the case of June,it can be assumed that there was/were some outlier event/events that caused a lot of fatalities.

For example,a suicide bomb can be one of the outlier events,that made June to stand out from the pattern.

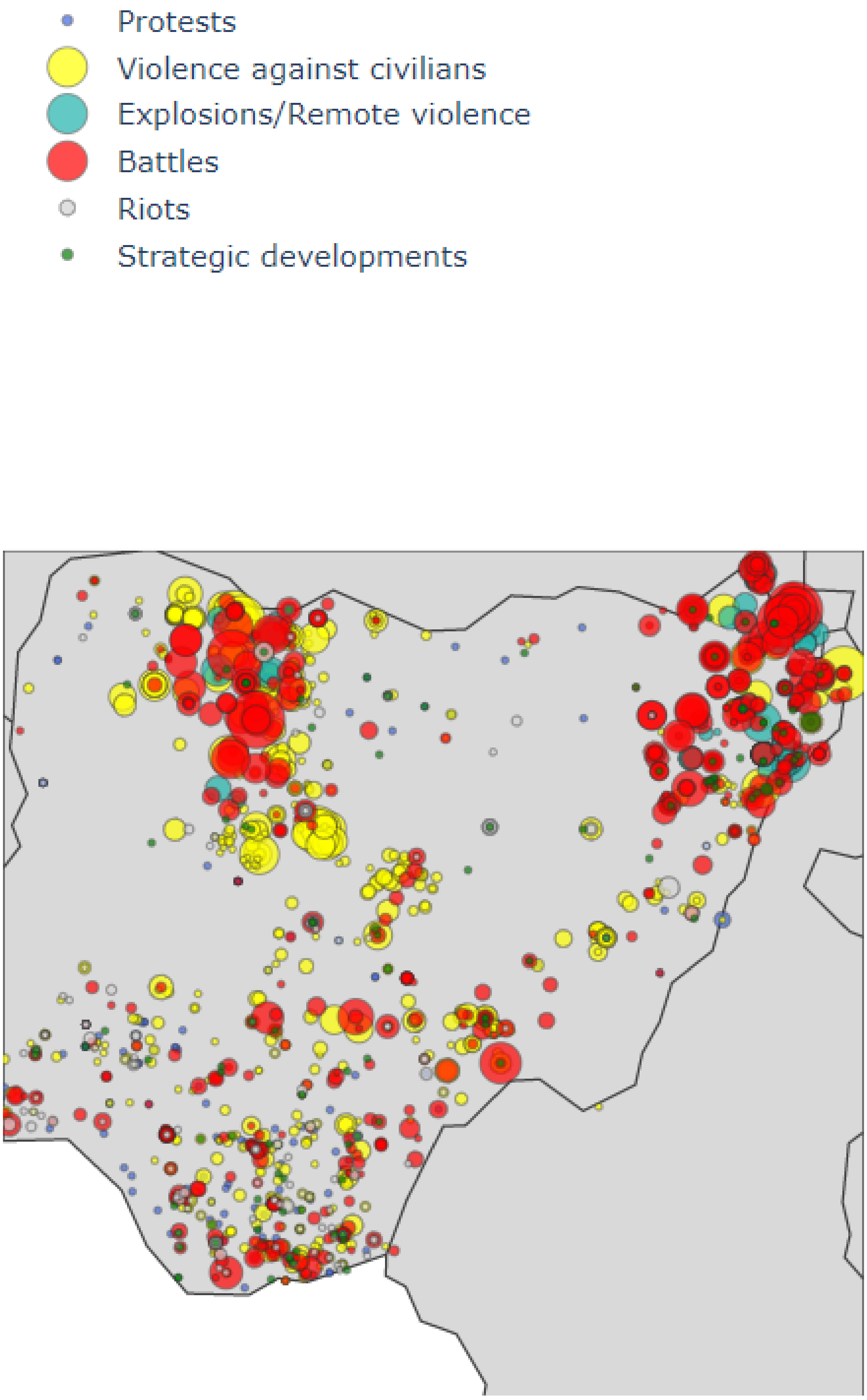


```
exp=d[d['sub_event_type']=='Suicide bomb'].sort_values(by='fatalities',ascending=False)
exp[['event_date_month','fatalities']]
```

	event_date_month	fatalities
1097	6	30
1650	3	11
1882	2	11
805	8	5
1522	4	5

As you see,30 people have been killed in June from suicide bomb,while in other months,this number was low.There maybe some other insignificant events,that became “significant” only in June.If there was more time,I will check every sub-event type to find out these outlier events.

We have **longitude** and **latitude** features,which will be very useful in constructing the map.This map will show the location of every event.Different colors mean different types of events,and the bigger the cluster is,the more fatalities have occurred.



INSIGHTS:

– Overwhelming majority of battles happened in the north of Nigeria

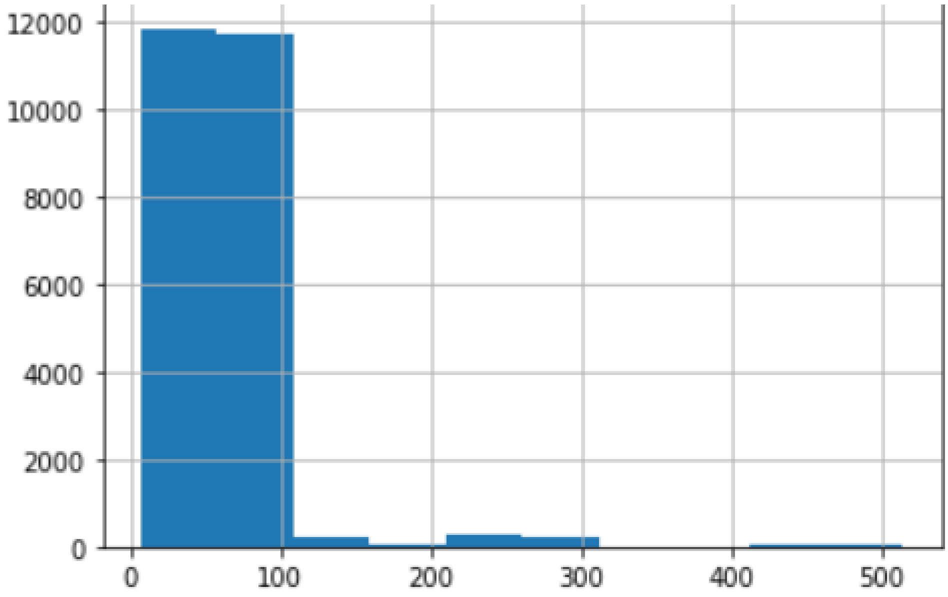
- The northern part of Nigeria is considered as a part of Nigeria,where events with high number of fatalities happen(just compare the size of clusters in between the north part and the south part)

- When looking to the north part,it can be seen that there are two major "dangerous" territories:first one is in the northwestern part,and another one is in a part of Nigeria,which is partially surrounded by Cameroon,Chad and Niger.

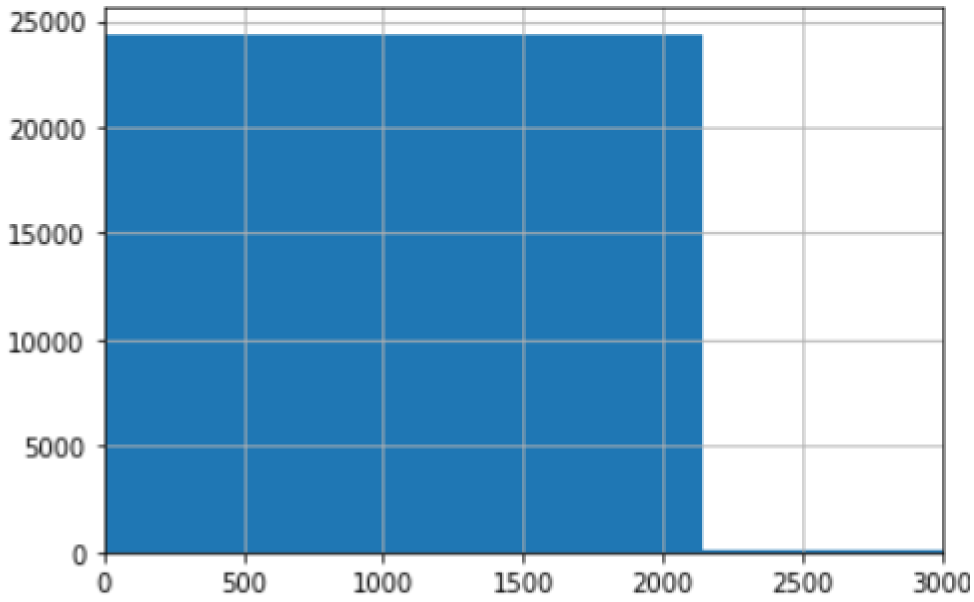
Question 2:What trends, patterns, and information can you find when analyzing the Newsfeed dataframe provided? How did you analyze the data? What was your methodology? What would you suggest pursuing if you had more time (e.g. a week or two weeks).

In the news feed dataset,there are two descriptions for each news article:**short** and **full**.Let’s analyze each type of description.

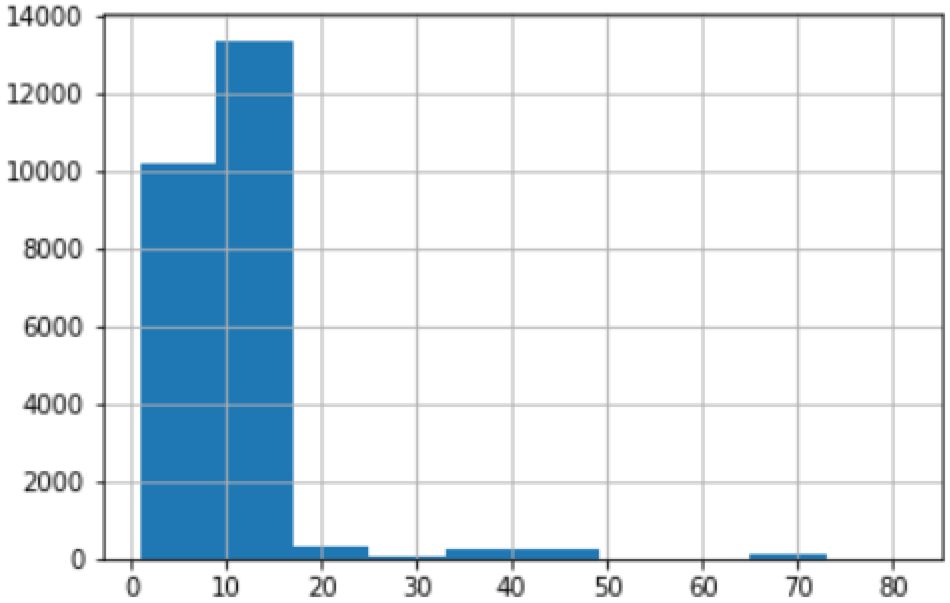
Firstly,we should know the average number of characters.



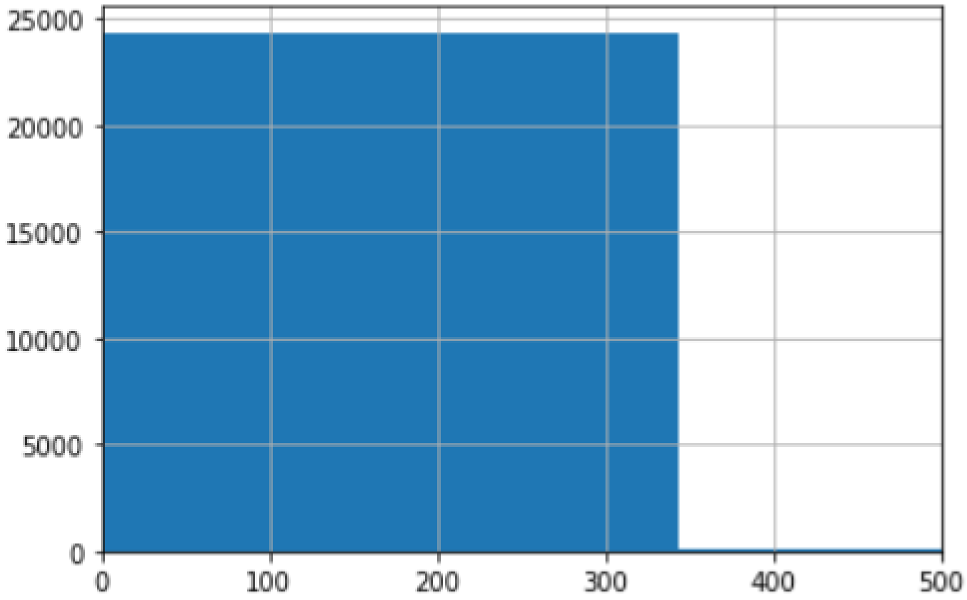
The number of characters in short descriptions reaches up to 100.



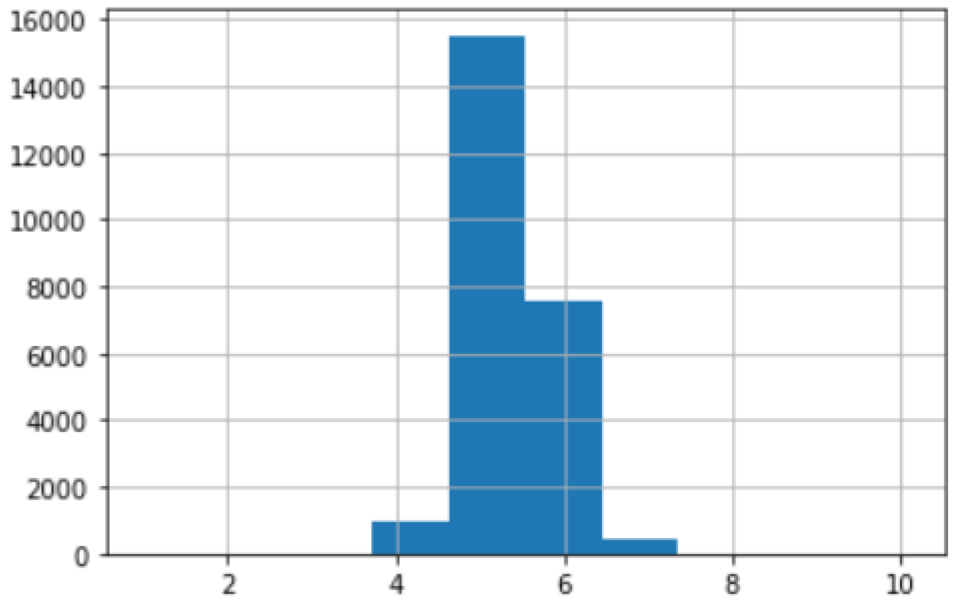
The number of characters in long descriptions normally reaches up to more than 2000.



Each short description normally contains 10-20 words.



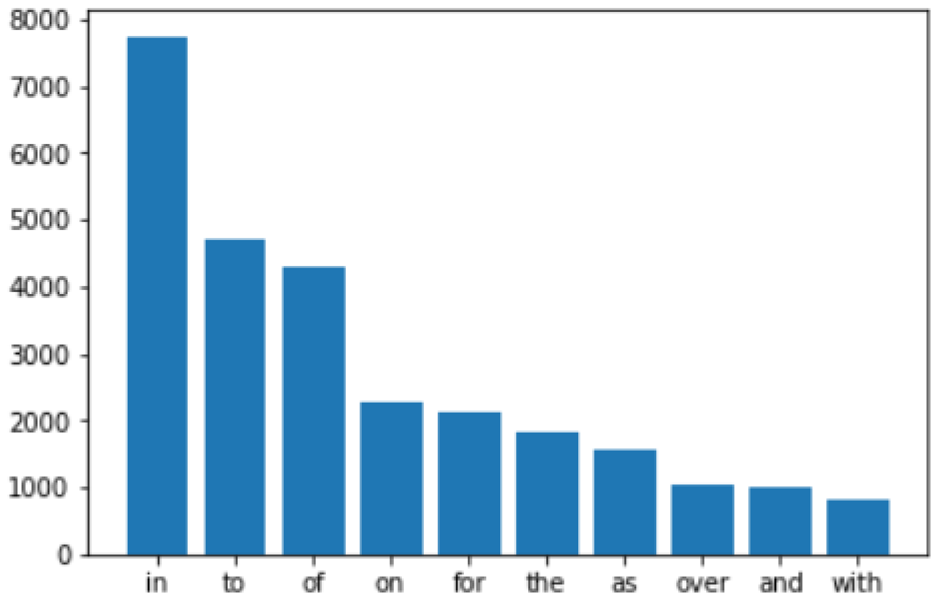
There are more than 300 words for each long description



The average word length for most of the words in full descriptions ranges between 5 and 6.

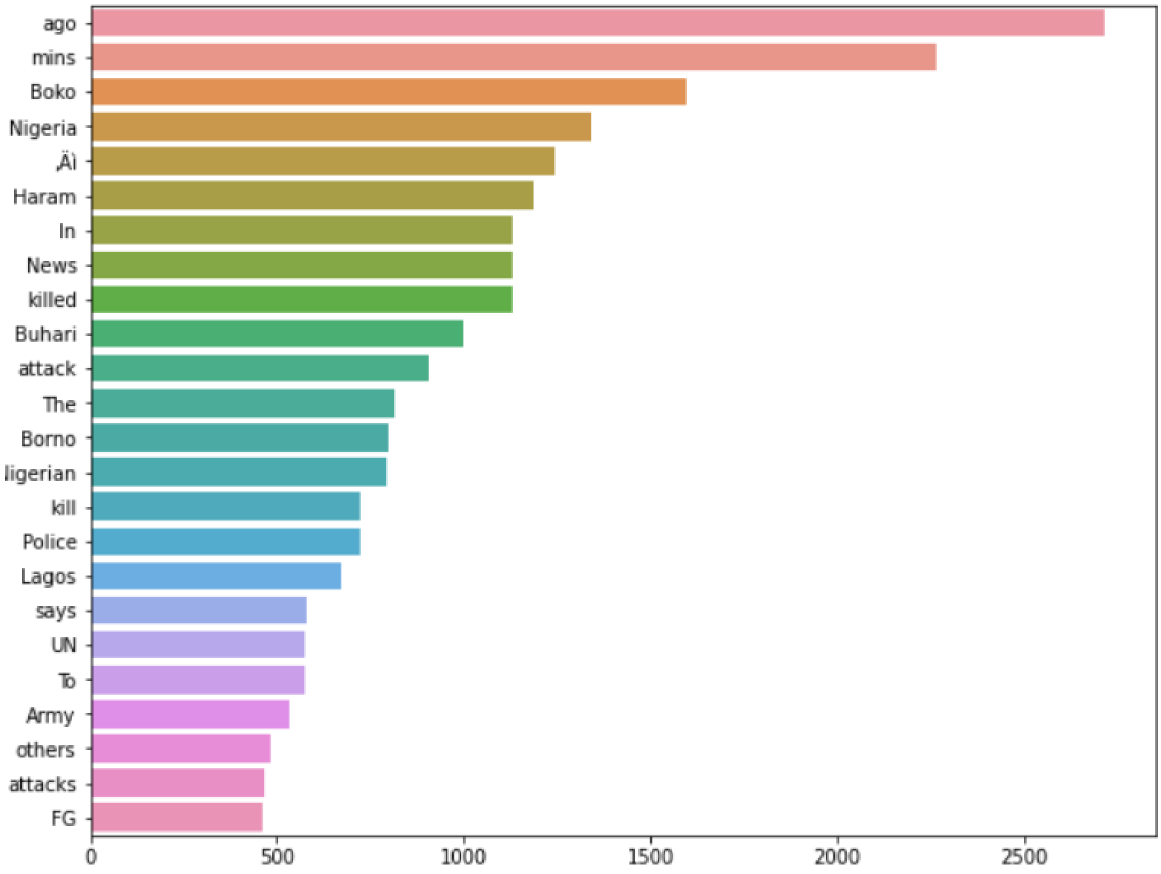
In this question,I will use NLP,as it will be impossible to analyze the text data without it.I will use the NLTK framework,as it has friendly-to-use syntax and lots of functionalities.

Firstly,it’s interesting to know which stopwords are the most frequent.



This is the bar chart that shows the most frequent stopwords in short descriptions.

We also need to know the most frequent words that appear in the news articles.In this case,stopwords should be excluded,as we don’t need them.

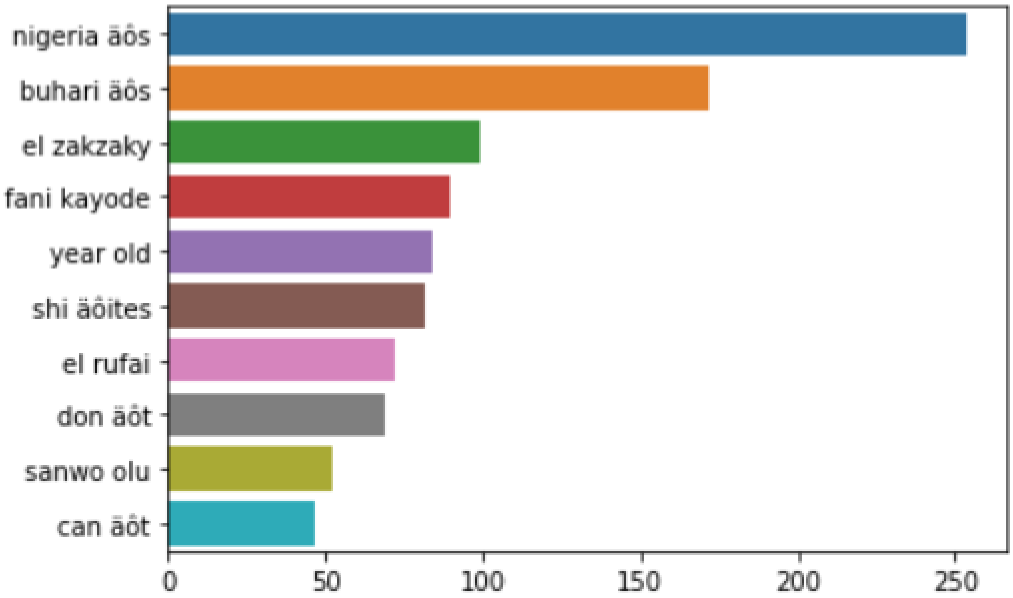


These are the most frequent words used in short descriptions.It is really interesting that words **Boko** and **Haram** are in the top-6.The word **Boko** has been used even more frequently,than the name of the country,**Nigeria**.So it means,that Boko Haram(terroristic organisation) is being mentioned frequently in the news.Basically,there are lots of reasons why they are mentioned by the news,but I think the main reasons are:

- They commit more crimes than any other organization(however,they killed far less people,than Zamfara Communal Militia,Islamic State-Lake Chad Faction and Fulani Ethnic Militia,based on our graph in Question 1)
- They commit more crimes(suicide bombs,strategic developments),that are specifically designated to kill less people,so that they can be mentioned more frequently in the news,than other organizations

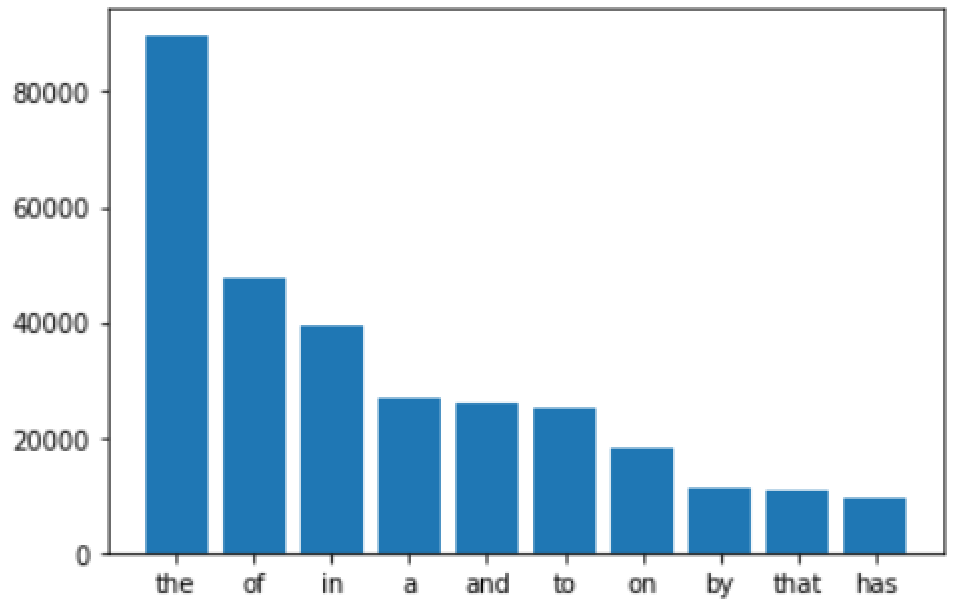
- Some news agencies specifically mention them for their own gain(very unlikely and sounds like a conspiracy theory)

It will be very useful to know the most frequent sequences of words.For implementing this,I will use N-gram technique.

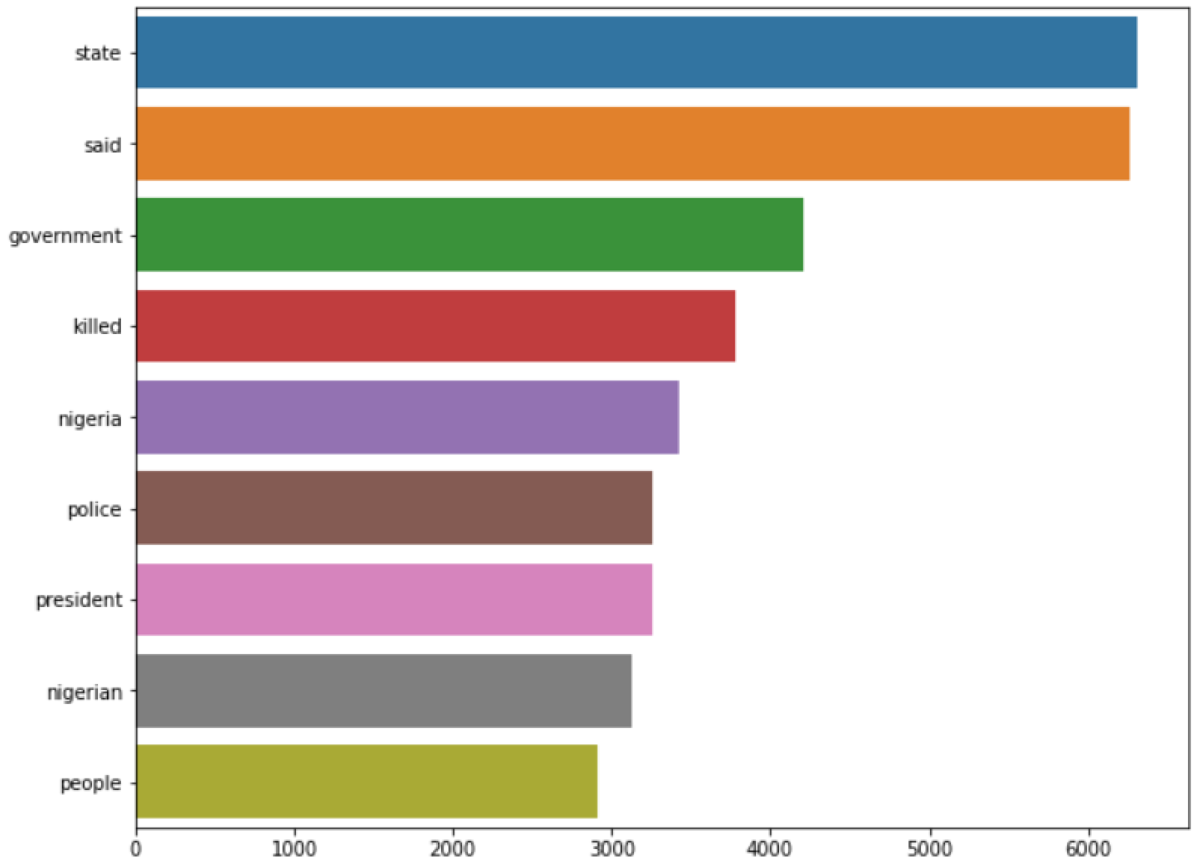


This graph shows the most frequent sequences of 2 words,or so called bigrams(n-grams technique has been used).El Zakzaky is a Shia leader and Fani Kayode is a renowned Nigerian politician.They are the most mentioned personas in the news articles.El Rufai is also in the top-10 and he is also a politician and a governor of the Kaduna State.

From now on,I will analyze only long descriptions as they are more informative.



the and **of** are used more often than **in**,which is the most used stopword in short descriptions.



This graph shows the most frequent words that appeared in full descriptions of the news articles.We don't see Boko Haram and other words that have been identified,when analysing short descriptions.It maybe because of the fact that words in the full descriptions don't include the name of terrorist organizations so much frequently,as "government","nigeria",and "president",due to limitations and regulations from the government.This pattern can also happen randomly,so there maybe no limitations at all.

Topic modeling will be reasonable to deploy here.Topic modeling can help me divide all the news articles into the specific topics(categories).In this case,I will use lemmatizer and tokenizer.I decided to get 5 topics and used the LDA model.These are the results:

1.**The,protest,group,attack,election,government** - this topic is about protests,that are likely to be caused as a consequence of the election

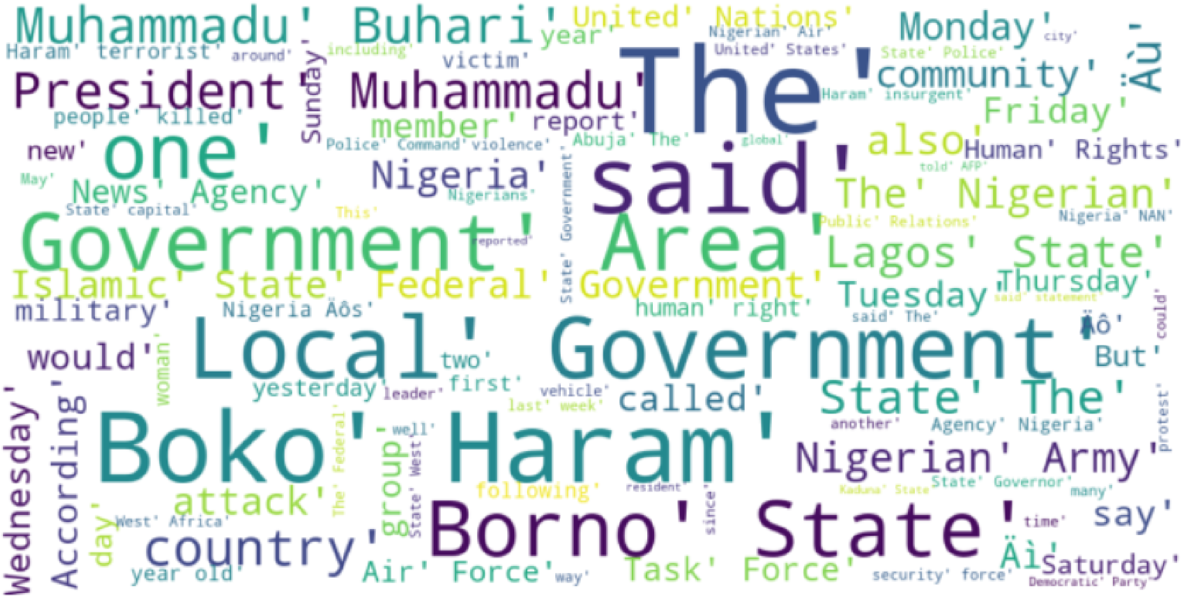
2.**The,State,Lagos,road,accident,people** - this topic is about man made disasters

3.**Nigeria,president,Buhari,Muhammadu,country,said,Abuja,National** - this topic is related mostly to the current president of Nigeria,Muhammadu Buhari

4.**Police,arrested,suspect,Lagos** - this topic is related to the events,happened in Lagos and the police investigations

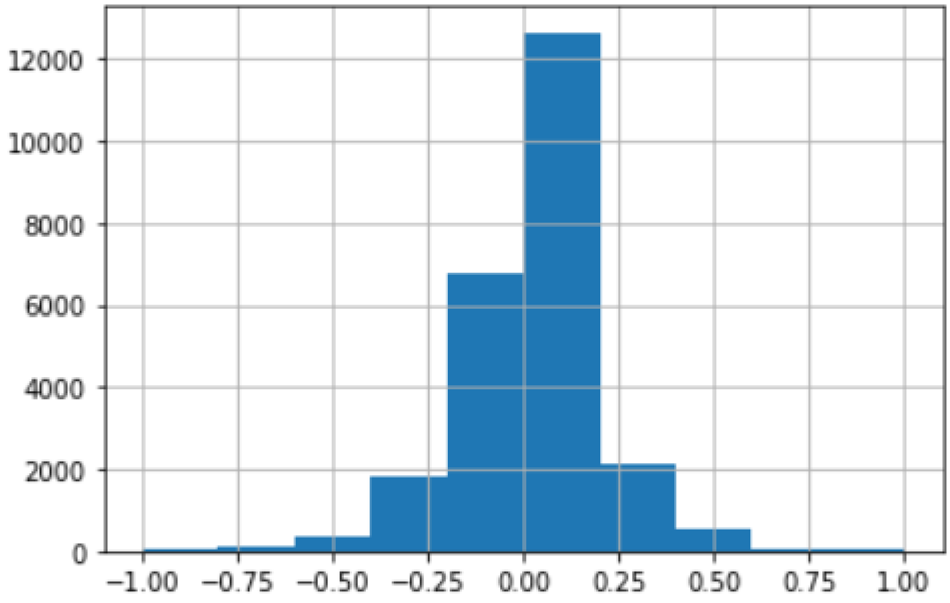
5.**The,State,attack,said,killed,Boko,Haram,Nigerian,terrorist** - this topic is solely related to Boko Haram and its terroristic acts

Let’s also use the most popular NLP technique - Word Cloud.



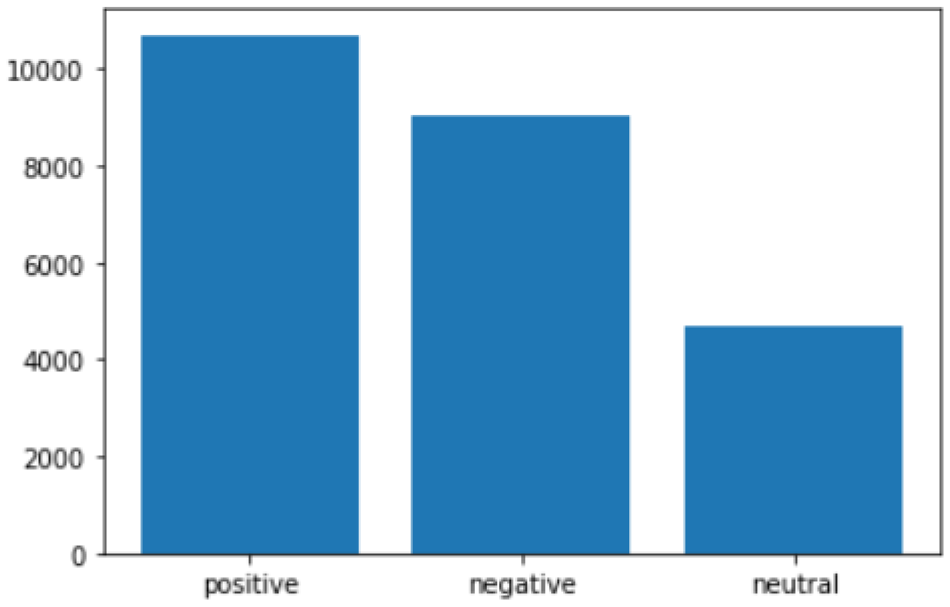
As expected,Boko Haram is mentioned here.Government,State,President,Local are also the keywords that appear frequently in the newsfeed.

It seems that Question 2 is Done,but not yet.Sentiment Analysis,in my opinion,is the most useful tool that can help us identify the emotional semantics of the news and,maybe,create a link with a Question 1.It is reasonable to conduct Sentiment Analysis,in order to know the emotional type of the news articles:positive,negative or neutral.In order to assess the "degree" of positivity or negativity,polarity scores are used.In this case I will use pre-trained TextBlob.

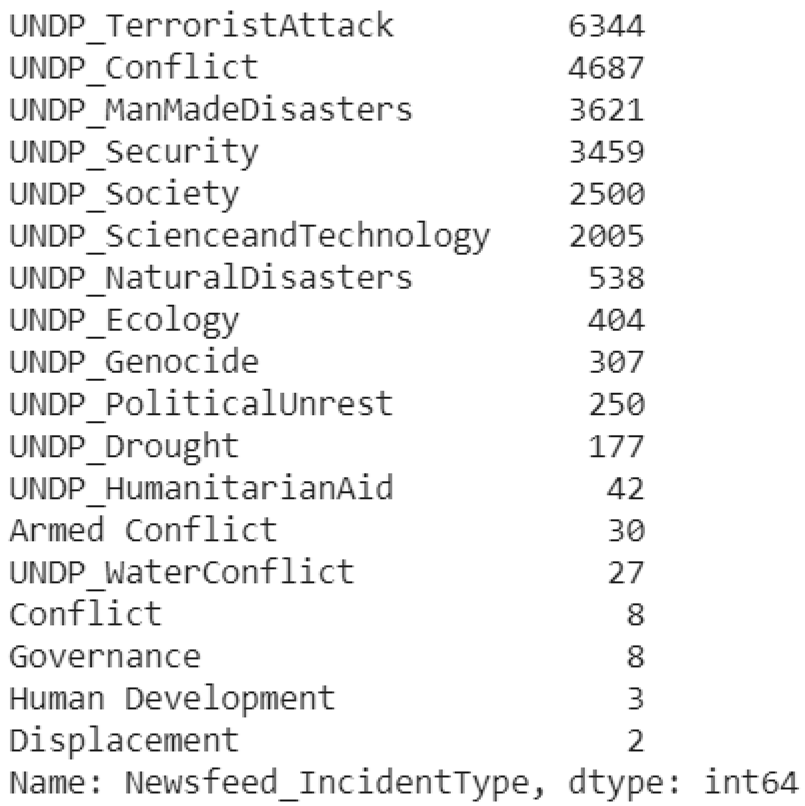


Most of the news articles have polarity scores ranging from 0 to 0.2.

If the news will be categorized as positive,if the polarity score is more than 0,neutral if polarity score is equal to 0, or negative,if the polarity score is less than 0,we can get pretty good insights.



The amount of positive news is bigger than negative ones.We can check whether the model was right by looking at the types of news articles.

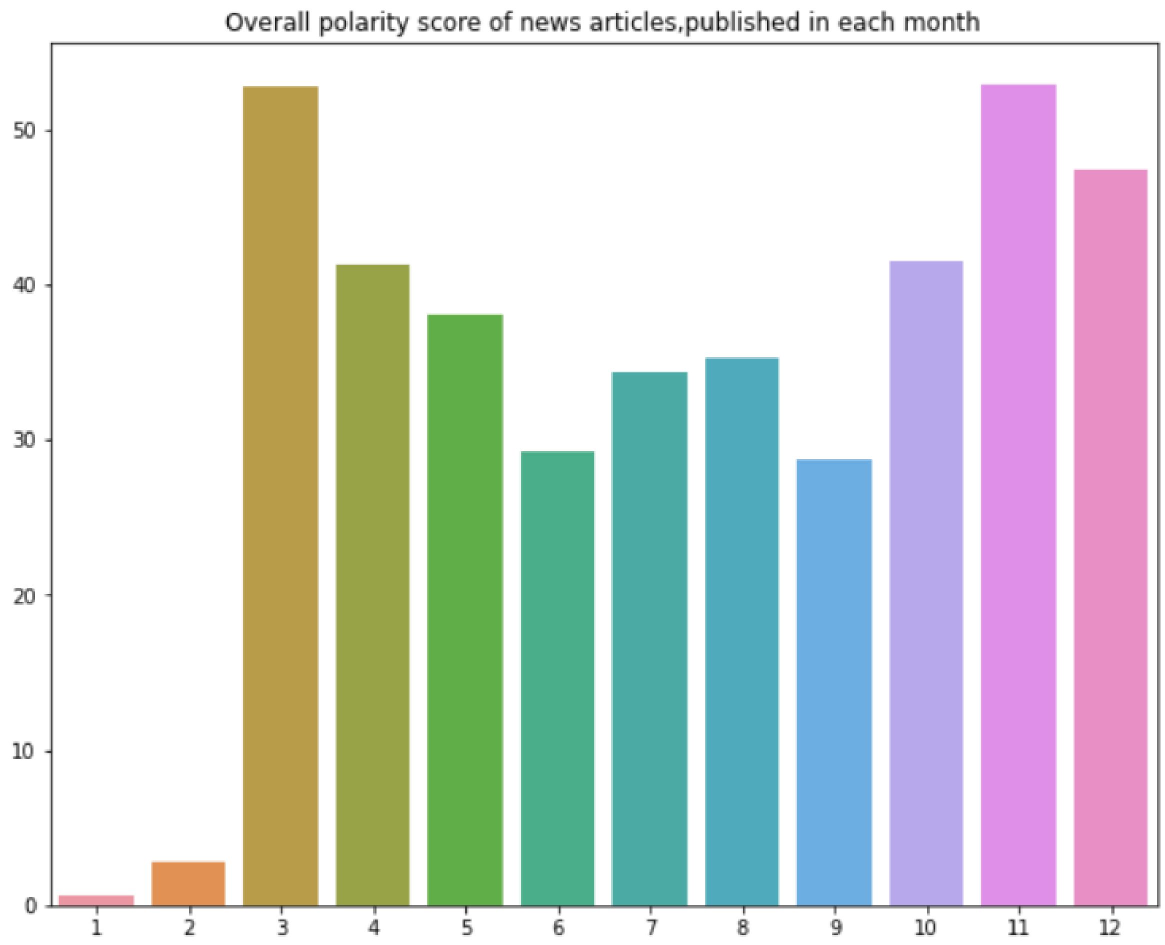


```
UNDP_TerroristAttack         6344
UNDP_Conflict                4687
UNDP_ManMadeDisasters        3621
UNDP_Security                3459
UNDP_Society                 2500
UNDP_ScienceandTechnology    2005
UNDP_NaturalDisasters         538
UNDP_Ecology                  404
UNDP_Genocide                 307
UNDP_PoliticalUnrest          250
UNDP_Drought                  177
UNDP_HumanitarianAid           42
Armed Conflict                 30
UNDP_WaterConflict             27
Conflict                        8
Governance                      8
Human Development               3
Displacement                    2
Name: Newsfeed_IncidentType, dtype: int64
```

Seems that our model was pretty accurate.However,I still believe that there is more negative news than positive.

If I had more time,I would create my own NLP model or use a pre-trained,but more complex model.Also,I would use the links of each article and scrape even more data.

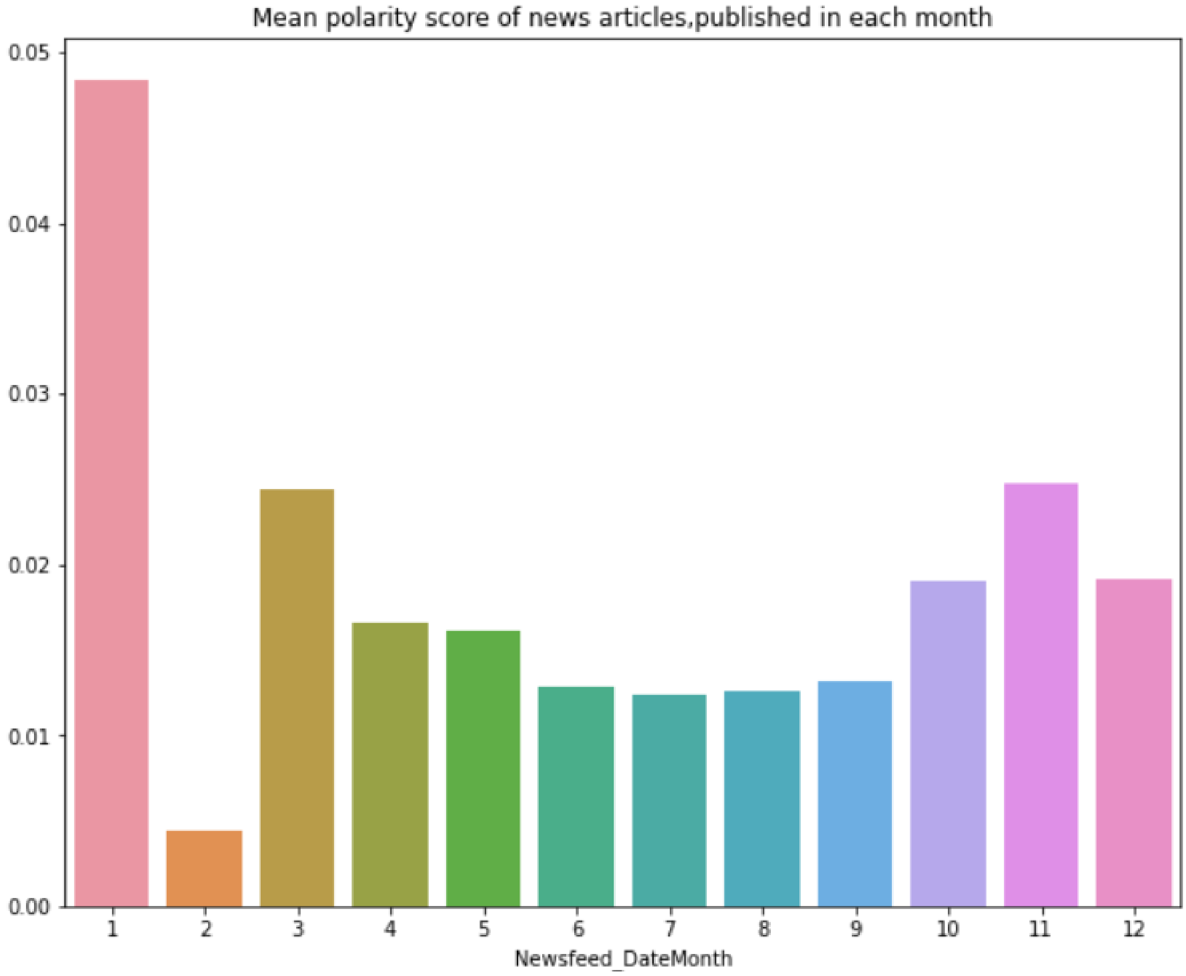
Last but not least,I decided to plot the polarity dynamics of the news articles.



QUESTION 3 : Can you identify any existing correlations or relationships between the two data frames?

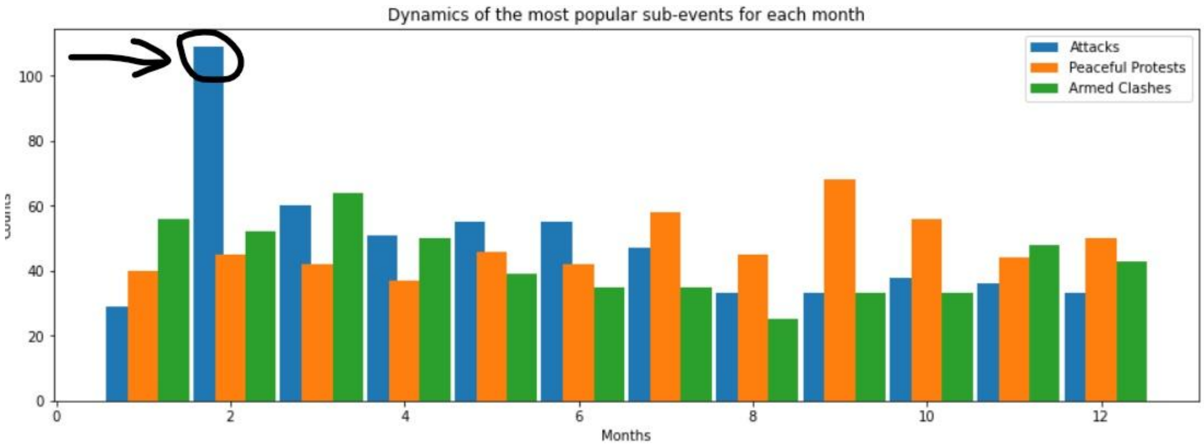
January has low polarity score,most probably due to the low amount of news articles,published this month.However,there is an interesting correlation.In the Question 1,it has been identified that February was the month when the highest numbers of attacks happened (in the grouped bar chart),and here,it can be seen,that February has low polarity score.Most probably,because majority of news articles was focusing on these attacks,which of course has negative semantics.

THIS GRAPH IS NOT VERY INFORMATIVE,as it does contain the number of published news articles into account.This factor should be eliminated in order to better assess the polarity scores for each month.In this case,I decided to compare means of polarity scores(sum of polarity scores/the number of news articles) rather than only the sum.

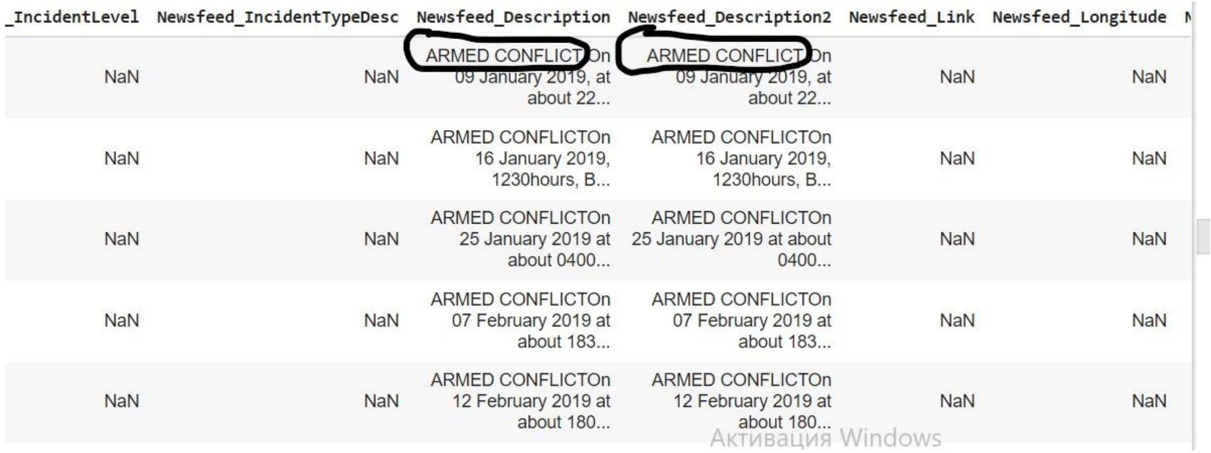


As it can be seen from the graph,the average news article that has been published in February has the lowest polarity score compared to other articles published in other months.

Also I want to note that news articles,published in summer,have identically low polarity scores(but not as low as in February),and it's mostly due to the large number of protests.It also should be said that summer has the perfect weather conditions for conducting the protests and strategic developments.



Another relationship can be found in the datasets.The event type has a separate feature in the ACLED dataset,but in the newsfeed data,it doesn’t have a feature,but it is shown at the beginning in both short and full description of the news articles.



_IncidentLevel	Newsfeed_IncidentTypeDesc	Newsfeed_Description	Newsfeed_Description2	Newsfeed_Link	Newsfeed_Longitude	N
NaN	NaN	ARMED CONFLICTOn 09 January 2019, at about 22...	ARMED CONFLICTOn 09 January 2019, at about 22...	NaN	NaN	
NaN	NaN	ARMED CONFLICTOn 16 January 2019, 1230hours, B...	ARMED CONFLICTOn 16 January 2019, 1230hours, B...	NaN	NaN	
NaN	NaN	ARMED CONFLICTOn 25 January 2019 at about 0400...	ARMED CONFLICTOn 25 January 2019 at about 0400...	NaN	NaN	
NaN	NaN	ARMED CONFLICTOn 07 February 2019 at about 183...	ARMED CONFLICTOn 07 February 2019 at about 183...	NaN	NaN	
NaN	NaN	ARMED CONFLICTOn 12 February 2019 at about 180...	ARMED CONFLICTOn 12 February 2019 at about 180...	NaN	NaN	

ACLED dataset contains **notes** feature,and I think,that these two datasets can be merged into one giant dataset.**Newsfeed_EventId** feature in Newsfeed data and **event_id_no_cnty** in ACLED data can be used as primary keys,as they represent the same information.