Data Mining Project

"Analysis of Suicide and Drone Attacks in Pakistan"

By

Rahul Chandnani

24/03/2017

Objective

In this project, I have selected different data sets regarding the terrorism activities in my home country Pakistan and then I extracted data related to suicide bombings and drone attacks in Pakistan. After extraction, I merged the different data sets to work on.

The main purpose of this project is to do analysis on social problem like suicide bombings and drone attacks using Data Mining Techniques learned in the lab sessions and find out some interesting trends and information from the merged dataset.

I have used the following three data sets for merging:

- Global Terrorism Database https://www.start.umd.edu/gtd/
- Suicide Attacks in Pakistan http://www.pakistanbodycount.org/
- Drone Attacks in Pakistan http://www.pakistanbodycount.org/

Problem Understanding

As we all know Pakistan is fighting war against Terrorism and it is constantly being attacked by the anti-state elements, by doing suicide attacks in the whole country.

Suicide Bombing has wreaked ruin in Pakistan in the most recent decade or somewhere in the vicinity, it continued heightening after the US Operation Enduring Freedom in Afghanistan, indiscriminately killing many individuals every year, towering as a standout amongst the most unmistakable security dangers that each Pakistani confronts today.

Apart from suicide attacks, US has attacked tribal areas of Pakistan using drones, So the main aim of this case study in this project is to use data mining tools and techniques to discover some interesting information from the prepared data set that can be used by analysts and media personals to have overall statistics of these attacks (Suicide and Drone).

Some inspiring things which will be explored are given below:

- The total loss of human lives with these brutal attacks
- Which are the most attacked cities by the terrorists
- Which was the worst year
- Visualization of attacks
- In which American administration, there were most drone attacks

The above list is not the extensive list of analyses but these are the most important things to be explored. Complete work will be shown in corresponding sections.

Data Understanding and Preparation

I have used three different data sets and merged those data sets based on suicide and drone attacks. Following are the details for all the data sets.

Global Terrorism Dataset:

The Global Terrorism Database (GTD) is an open-source database including information on terrorist events around the world from 1970 through 2015. So, I will extract only those data points which are of type "suicide attack" in Pakistan.

Suicide Attacks in Pakistan:

Pakistan Body Count (www.PakistanBodyCount.org) is the oldest and most accurate running tally of suicide bombings in Pakistan. The given database has been populated by using majority of the data from Pakistan Body Count.

Drone Attacks in Pakistan:

This data set is also populated from Pakistan Body Count which gives us the data points related to drone attacks in Pakistan

Name of Data Set	*Number of Rows*	*Number of Attributes*
Global Terrorism Dataset	156772	137
Suicide Attacks in Pakistan	476	24
Drone Attacks in Pakistan	398	24

Data Preprocessing & merging:

Firstly, I loaded the Global Terrorism Dataset and extracted the desired data from it like given below:

```
myData=read.csv("globalterrorismdb.csv")
mydataPak_Suicide=myDataPak[which(myDataPak$suicide==1),]
```

All the suicide attacks in Pakistan

Then, I loaded my other two data sets and tried to check the common and uncommon attributes in both so that I can remove or add other necessary attributes in my final dataset. I also changed the names of some similar attributes and I have also added some new attributes and dropped some un necessary attributes for merging consistency, as shown in following figure.

```
drone=read.csv("DroneAttacks_Pakistan.csv")
suicide=read.csv("SuicideAttacks_Pakistan.csv")
intersect(names(drone),names(suicide))
 #checking the common attributes
setdiff(names(drone),names(suicide))
# checking the uncommon attributes
# adding aditional necessary attriutes in data sets,
#so that data sets can be merged
suicide$Is_Suicide=1
suicide$Is_Drone=0
drone$Is_Drone=1
drone$Is_Suicide=0
drone$Blast.Day.Type=NA
drone$References=NULL
drone$Comments=NULL
#changing column names of similar data
#for data integration from the above mentioned data sets
colnames(drone)[8]='Total.Died.Max'
colnames(suicide)[18]='Injured.Min'
colnames(suicide)[17]='Total.Died.Max'
colnames(suicide)[16]='Total.Died.Min'
```

After merging the data sets, I got my desired final data set in which I worked for analysis. Finally, I started cleaning my merged data set since it was messy and contained a lot of ambiguous things. I tried to do basic cleaning at this step, but some cleaning is done according to the tasks i.e. when it was needed while performing some task, as shown in following figure.

```
mergedData= apply(mergedData,2,function(x)gsub('\\s+', '',x))
mergedData=as.data.frame(mergedData)
levels(mergedData$City)= tolower(levels(mergedData$City))
levels(mergedData$Target.Type)= tolower(levels(mergedData$Target.Type))
mergedData$Total.Died.Max=as.numeric(mergedData$Total.Died.Max)
mergedData$Latitude=as.numeric(mergedData$Latitude)
mergedData$Longitude=as.numeric(mergedData$Longitude)
# Since date is in un organised format like "Wednesday-May 8-2002"
# So I'm extracting the last 4 characters from the date attribute
mergedData$Date= str_sub(mergedData$Date, start= -4)
head(mergedData$Date)
```

Analysis

1. *Plotting Location of Attacks:*

Firstly, I started my analysis by plotting the locations of attacks (suicide and drone both) on the world map by using the Latitude and Longitude attributes of the data set. Following are the plots.

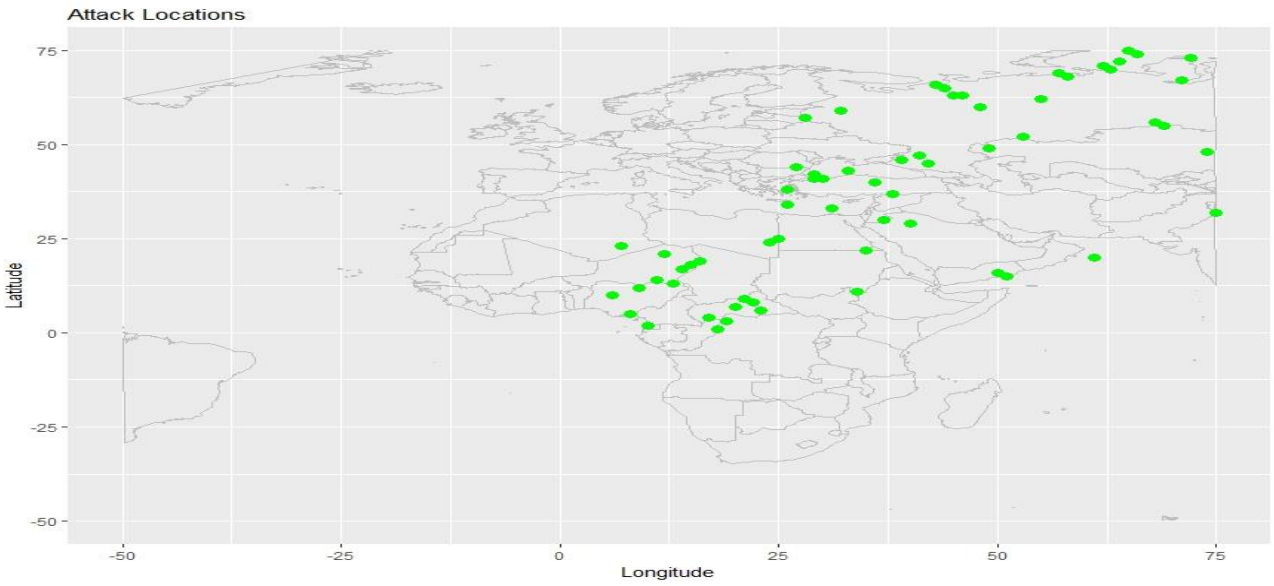


Figure 1 Location of Suicide Attacks

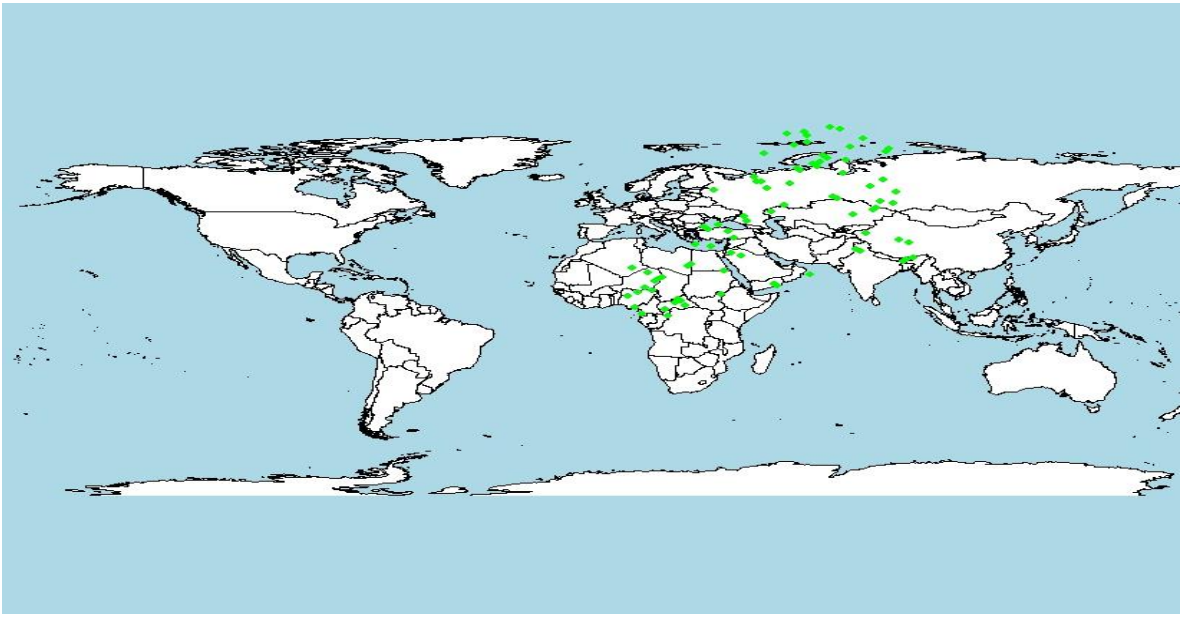
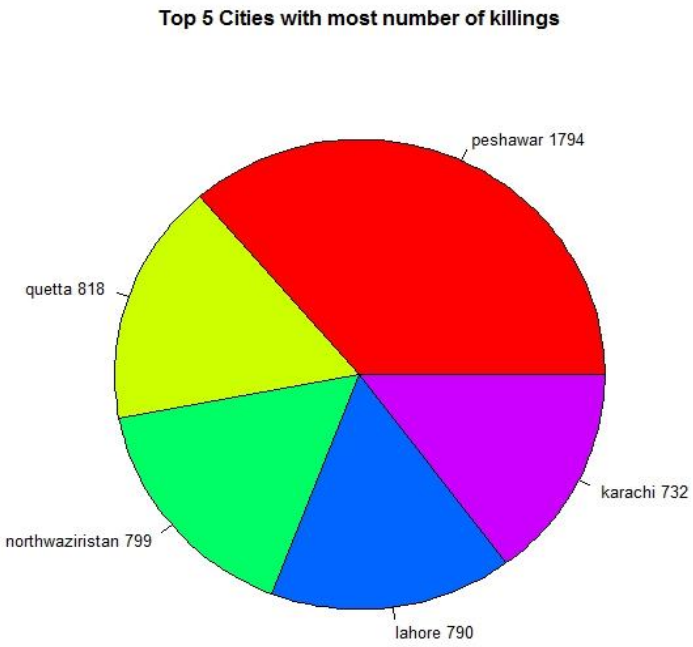


Figure 2 Locations of Drone Attacks

2. *Top five cities of Pakistan with most number of killings (Suicide Bombing)*

Then I tried to find out which are the most attacked cities of Pakistan by suicide bombings i.e. which cities are mostly targeted by the terrorists. It was clear from the analysis that two of the biggest cities (Karachi and Lahore) and economic hubs of Pakistan are at the target of terrorists. Following is the pie chart illustration of this finding.



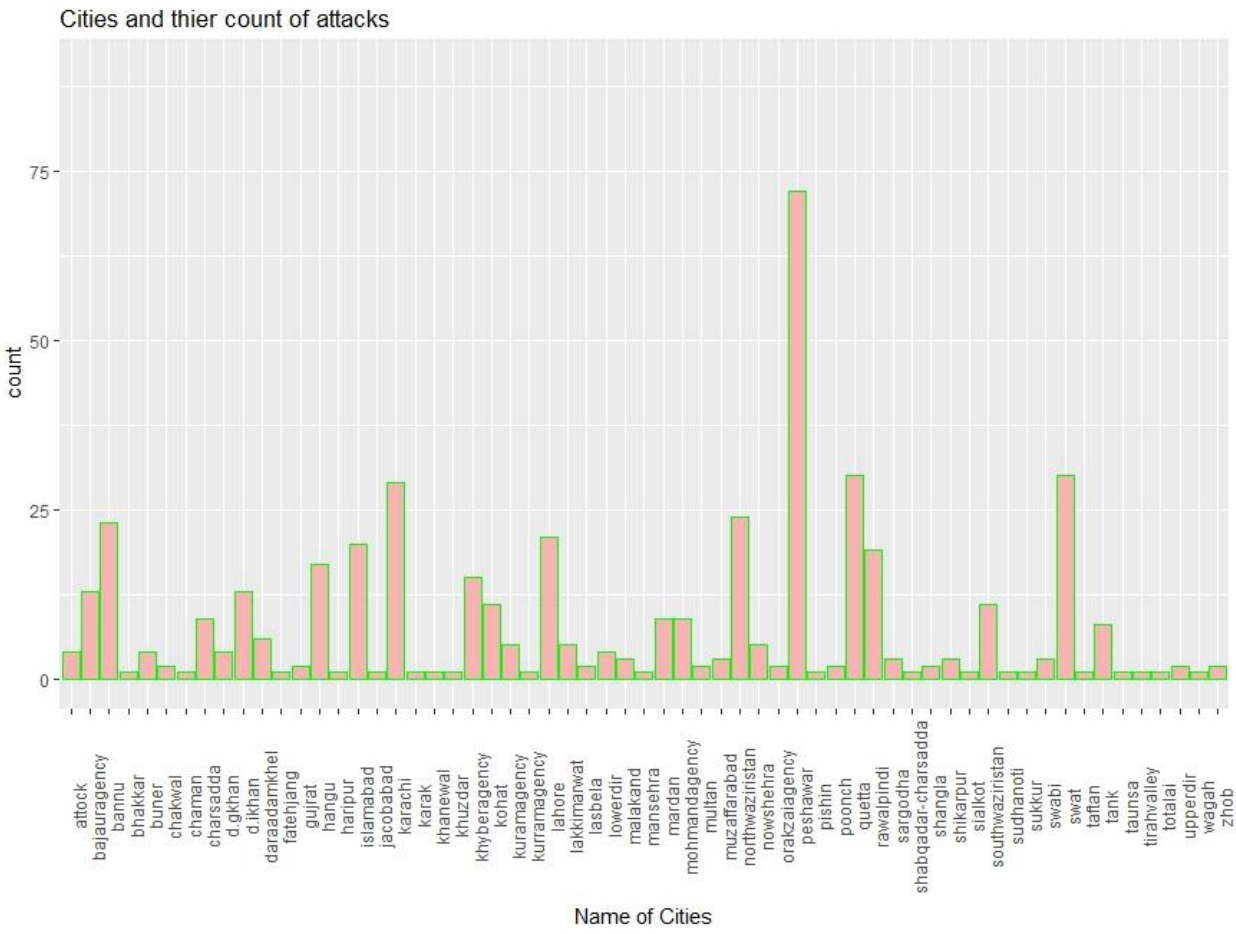
Top 5 cities with most number of killings

It is clear from the above pie chart that Peshawar is the most attacked city of Pakistan with highest number of killings i.e. 1794. It is sad that people of Peshawar have lost their loved ones with these senseless suicide bombings.

3. *Cities with the count of suicide attacks:*

So, after finding the top 5 cities with most number of killings, I tried to find the frequency i.e. count of attacks in each city of Pakistan. So, that this figure can be used for future safety measures by the Government and the law enforcing agencies.

The plot for the above information is given below:

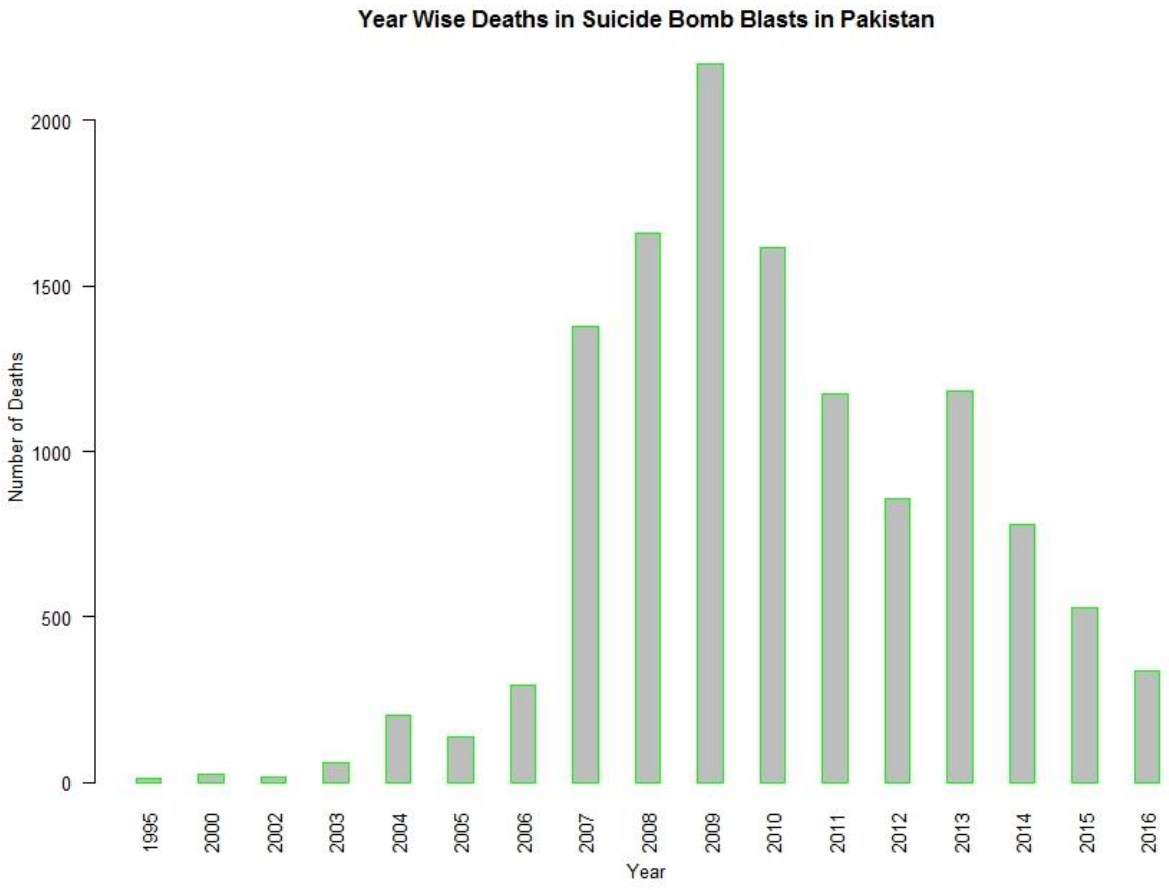


From the above plot, it is clear that the Peshawar stays at the top i.e. the most favorite place of attack for the terrorists. Furthermore, the above complete information can be used directly by the law enforcing agency for security measures to be taken in each specify city according to the sensitivity of the attacks.

4. Year Wise Deaths in Suicide Bomb Blasts in Pakistan:

This count will indicate which year was the worst year in terms of killings occurred due to the terrorism caused by suicide attacks.

Following is the graphical illustration:



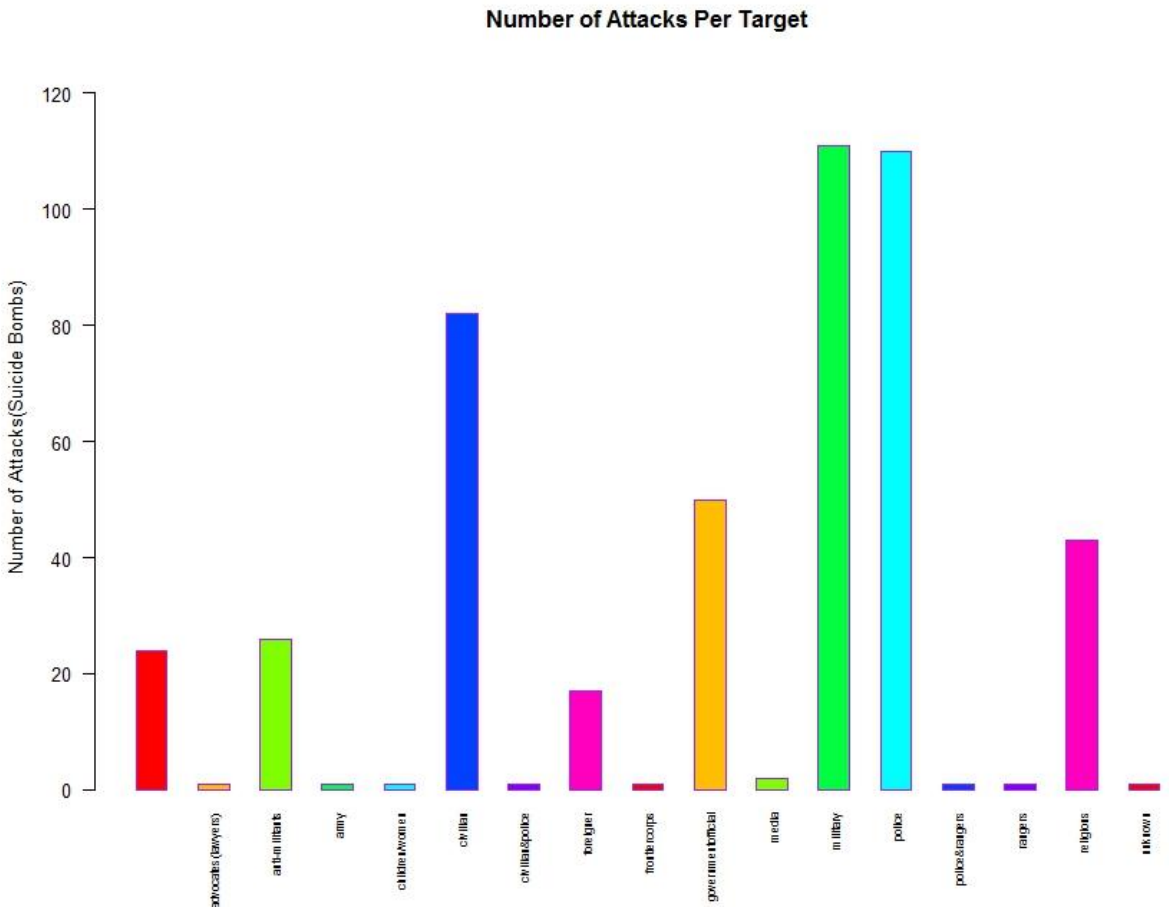
From the above plot, it is clear that the highest number of killings were in 2010 and this was the most difficult year for the families of those who lost the lives of their loved ones in suicide attacks.

5. *Number of Suicide Attacks Per Target:*

Now, the next interesting point is to find out that which category of people were targeted by the terrorists i.e. Military, Police etc.

So, this information again can be used by the controlling authorities to increase safety measures in those areas where the people of that category are highly populated.

Graphical representation of this information is given below:



From the above graphical illustration, it is found that the terrorists mostly targeted Military and Police by suicide bombing.

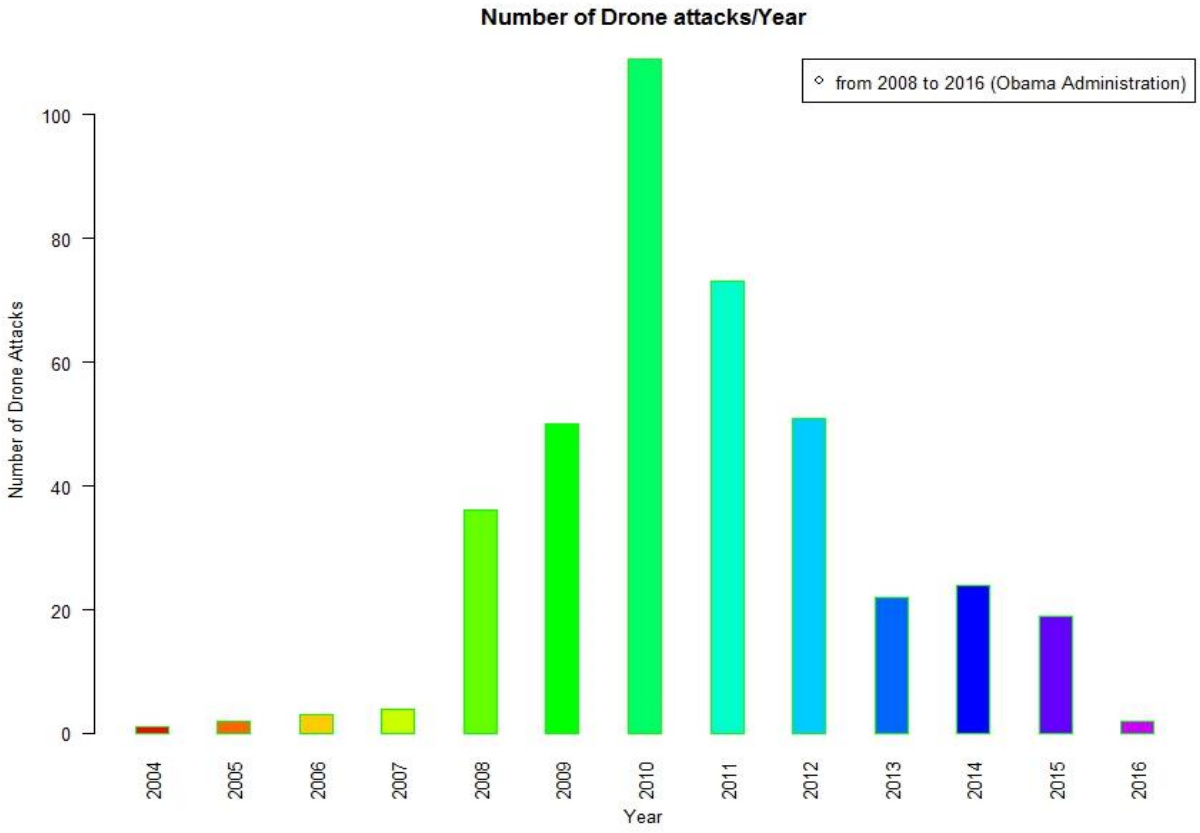
So, this conveys that normal people should avoid visiting places holed by military authorities because of the reason that terrorists are totally against the Military and Police.

Note: Now coming towards the analysis of Drone Attacks.

6. Number of Drone Attacks Per Year:

We are done with the separate analysis of suicide attacks, now coming towards the analysis of Drone attacks carried out by US in the tribal areas of Pakistan.

Graphical illustration is given below:

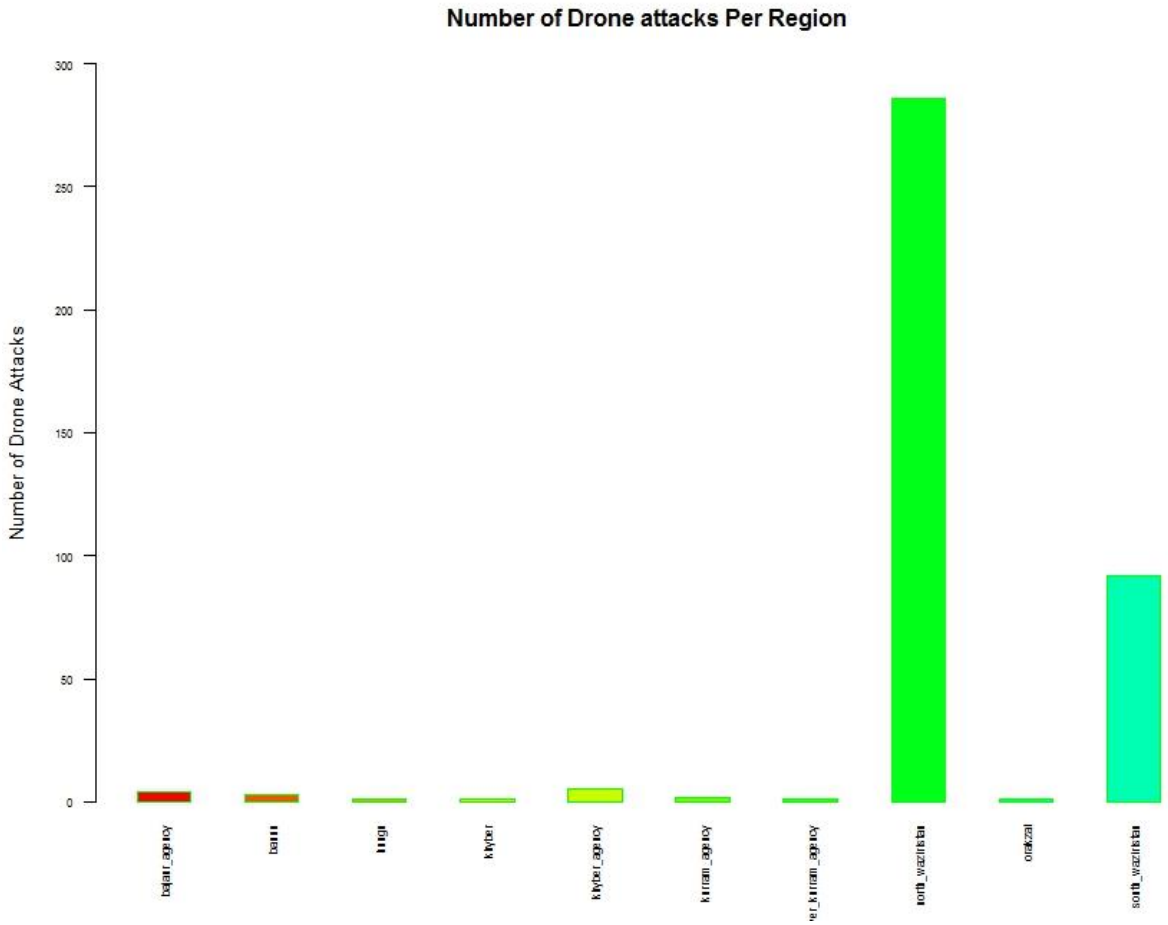


From the above plot, it can be seen that the 2010 was the year with most number of drone attacks in Pakistan.

This graph conveys one thing that the spike in drone attacks is from 2008 onwards and the one interesting point to note is that the Obama was the president of US from 2008-2016, so he was most favored to the drone attacks as compared to the administration of George Bush

7. Number of Drone Attacks Per Region:

The next interesting insight to find is that the which tribal regions were mostly affected by the drone attacks i.e. number of drone attacks per region.



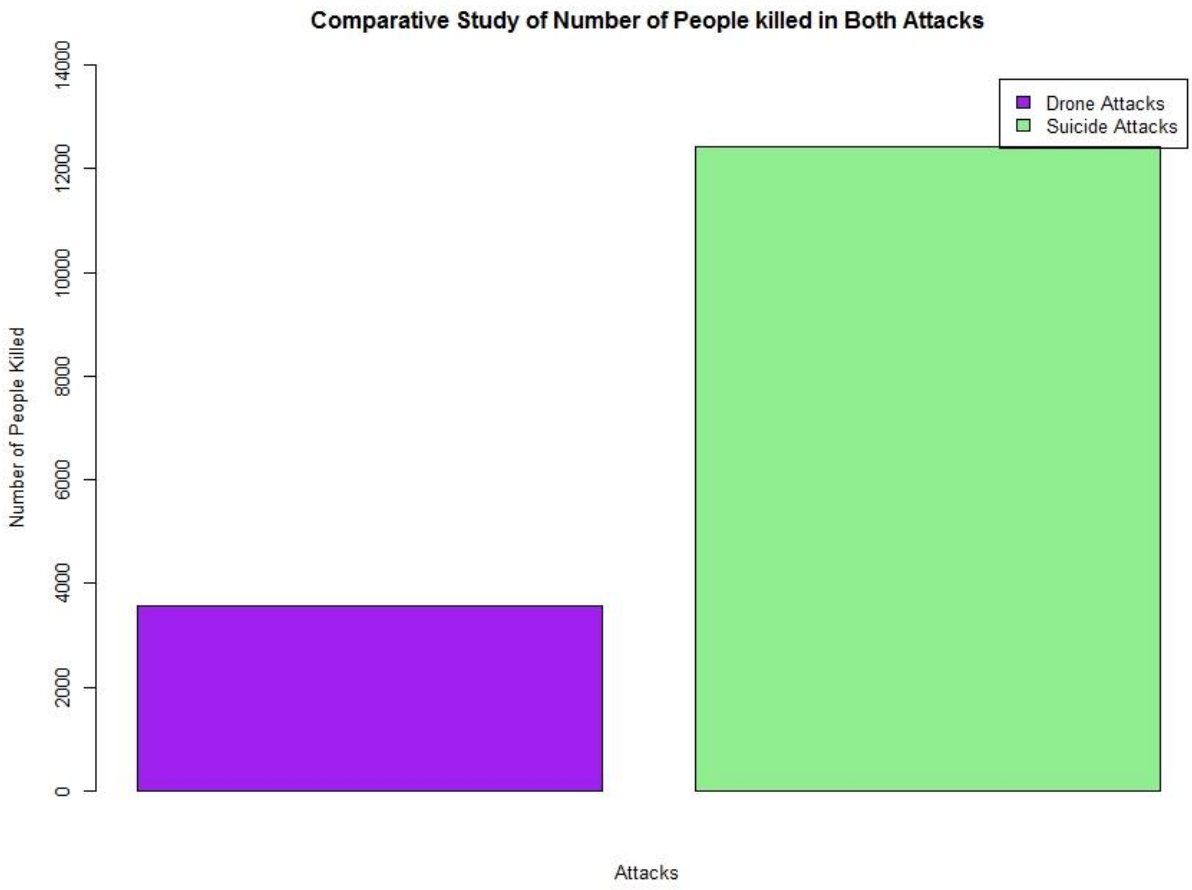
It can be seen around 280 drone attacks were done by US government on North Waziristan. It was the most vulnerable part according to the count of attacks.

Pakistan army should do operation in this tribal region to clean the area from terrorists.

8. *Comparative Study of Number of People Killed in both the Attacks:*

Now, finally one important thing to find out is the count of people who lost their life in both the attacks i.e. in suicide attacks as well as in the drone attacks.

Graphical illustration is given below:



Total number of innocent people who lost their lives in suicide bombings is: 12430

Total number of people who lost their lives in drone attacks is: 3556

Conclusion

From the analysis, we can conclude that 12430 innocent people have lost their lives in suicide bombings in Pakistan, being a Pakistani, this figure is very damaging for me and I really feel sad for the families of those people who lost their loved ones in this brutal terrorism. So, Government of Pakistan should take strong measures to stop the suicide bombings.

Now considering the drone attacks, the total count of people who lost their lives in drone attacks is 3556, but we are not sure whether all the people were innocent or terrorists but one thing is for sure that in these kinds of attacks there is always a collateral damage, so US should learn that this is not the right way to stop terrorism because if the people are innocent who are hit by the drone attacks then there is always a chance of increase in terrorism because of the fact that the families of those who were innocent and killed by drone attacks will again come hard to take revenge for the loss of their loved ones.

What I learned from this Project:

This project is the big learning curve for me because I followed the whole Data Mining Project process. I learned how to extract and merge different data sets. Another great thing I learned is that how to clean the messy data which I got after merging the data from three different sources. Apart from this, now I'm confident with R language and its visualization libraries.