Introduction  
To  
Data Mining

by

*Rehan Ahmed  
Hassan Saeed*

[Pakistan Suicide Bombing Attacks](https://www.kaggle.com/zusmani/pakistansuicideattacks)

# Dataset

The dataset has been taken from [Kaggle](https://www.kaggle.com/zusmani/pakistansuicideattacks) and contains detailed information about 496 Suicide Bombings in Pakistan, and has been compiled by [Pakistan Body Count](http://www.pakistanbodycount.org). Suicide bombing is an operational method in which the very act of the attack is dependent upon the death of the perpetrator. Though only 3% of all terrorist attacks around the world can be classified as suicide bombing attacks these account for 48% of the casualties.

**Attributes**

* Date - the Gregorian date of the blast
* Islamic Date - the Islamic date of the blast
* Blast Day type - was it a Holiday or a Working Day?
* Holiday Type - if it was a Holiday, what type of Holiday was it?
* Time - the time of the blast, if available
* City - the city where the blast occurred
* Latitude - the Latitude coordinate of the blast
* Longitude - the Longitude coordinate of the blast
* Province - the Province where the blast occured
* Location - the exact location inside the City
* Location Category - the category of the location (total of 26 categories in the dataset)
* Location Sensitivity - from Low to Medium to High, based on how sensitive the location is
* Open/Closed - is the location open or closed for public
* Influencing event - if the bombing was connected to an event
* Target Type - if there was a target for the bombing
* Targeted sect (if any) - Targeted sect
* Killed min - Minimum number of people killed
* Killed max - Maximum number of people killed
* Injured min - Minimum number of people injured
* Injured max - Maximum number of people injured
* No. of suicide blasts - if there were more than 1 blasts at the same bombing
* Hospital name - The name of Hospital victims were taken to
* Temperature - Temperature on the day of the blast

Data Cleaning

The data was pretty inconsistent and dirty, containing a lot of duplicated values as well as a considerably high number of null values.

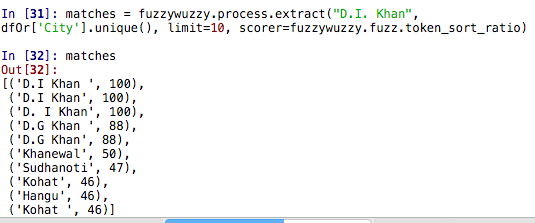
First, we tackled the inconsistent capitalization as well as the duplication of values.

* **Removing Inconsistencies**

In the **City** column, after applying the unique filter, numerous cities like “D.I. Khan” and “Kurram Agency” were being represented differently, taking the total number of cities to **93**.

First, we changed all the city names to lower-case and then we wrote a function using the **FuzzyWuzzy** library that matches each city’s name to every other city’s name and gives a similarity score to each name.

For example, for D.I.Khan,

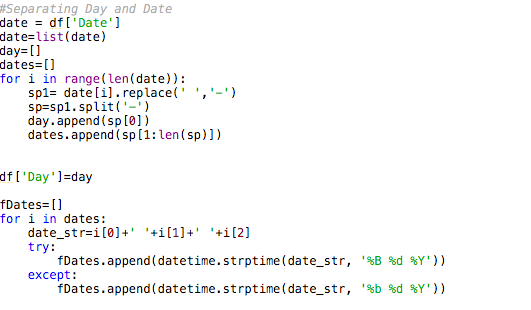
****

We then wrote a function to assign D.I.Khan to all other city names that had a similarity score of over 90, and then did the same for other city names that were being represented multiple times. We also applied **strip** and **lower** to city names to remove extra spaces and lower-case all the names.

This cleaning brought down the total number of cities in our dataset to 65. We also proceeded to remove the inconsistencies in the province names using the same method. We also applied **strip** and **lower** to province names to remove extra spaces and lower-case all the names. The same method was applied to location category, location sensitivity, Holiday Type.

* **Date Format**

The formatting of the date was pretty inconsistent too. Some dates recorded months as ‘January’ while others recorded months as ‘Jan’. So we wrote a function to separate dates and day from the date column and save it in a proper date format.



* **Null Values**

S# 0

Date 0

Islamic Date 0

Blast Day Type 10

Holiday Type 424

Time 211

City 0

Latitude 3

Longitude 3

Province 0

Location 3

Location Category 35

Location Sensitivity 36

Open/Closed Space 35

Influencing Event/Event 305

Target Type 26

Targeted Sect if any 48

Killed Min 146

Killed Max 16

Injured Min 131

Injured Max 32

No. of Suicide Blasts 82

Explosive Weight (max) 324

Hospital Names 199

Temperature(C) 5

Temperature(F) 7

Day 0

Date-old 0

Islamic Date Old 154

coordinates 0

There were a considerable amount of null values in our data, but for our initial analysis, we focused on fixing the important attributes. Given that suicide bombings are assumed to be religiously linked, Islamic Date was a pretty important attribute for our analysis. However, around 30% of values were missing.

To fix that, we used a Python library called **ummalqura** to convert the Gregorian dates to Islamic dates for all the missing variables.

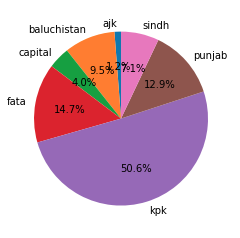
* **Line Breaks**

Lots of values in attributes like Location and Influencing Events contained Line Breaks, so when we tried to read the data in Knime, it read them as separate rows since Knime separates rows based on line breaks. We removed them using Python,

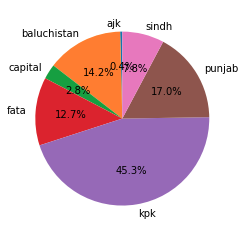
***df = df.replace('\n',' ', regex=True)***

# Analysis

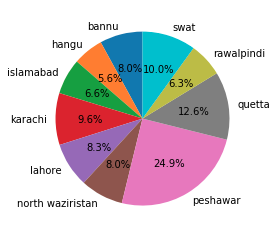
First, we analyzed the number of attacks for each province and given how unstable the law-and-order situation has been in KPK, it came as no surprise that over 50% of suicide bombing attacks had occurred there. FATA (Federally Administered Tribal Area) attracted the second largest number of attacks, followed closely by Punjab.



Secondly, we analyzed the maximum number of people killed for each province. Here too, KPK was the biggest victim. However, Punjab had a considerably larger number of people killed as compared to FATA, which is surprising since they had less number of attacks.

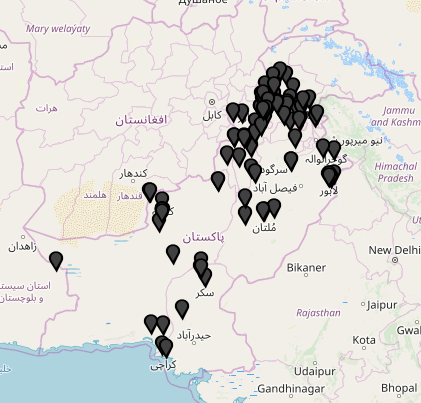
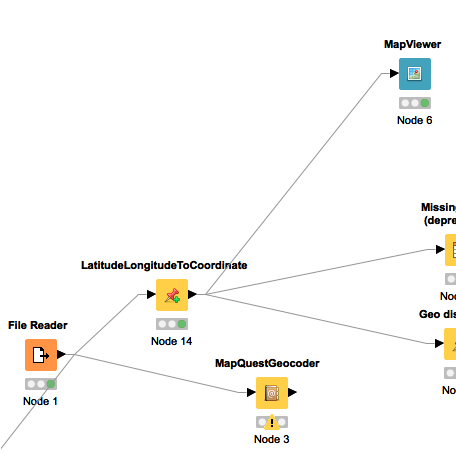


We also plotted the max number of attacks for each city, but since the total number of cities is 65, we only plotted the top 10 cities that attracted the largest number of attacks and **Peshawar** came out as the city that has suffered the most.



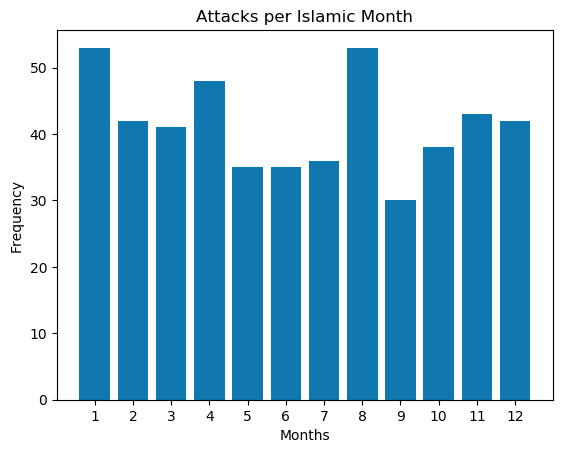
Here, we have plotted all the attacks that have happened on to the map of Pakistan, using the **Knime’s MapViewer** node. We used the Latitude and Longitudes given in the dataset. However, one issue with this is that the coordinates are not exact, most of them are just given for the city where the blast occurred in which means that many points on this map are superimposed on each other.

We tried using Google’s Address Geocoder but given the inconsistent nature of **Location** data, it didn’t work much.



Next, we performed a number of experiments to test some of our assumptions and extract insights from the data.

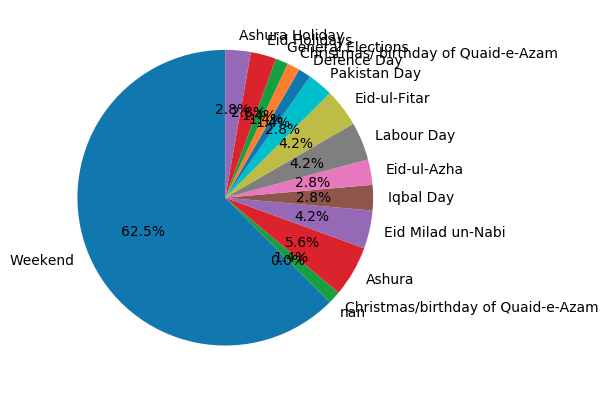
# Does the Islamic Month affect the number of attacks?



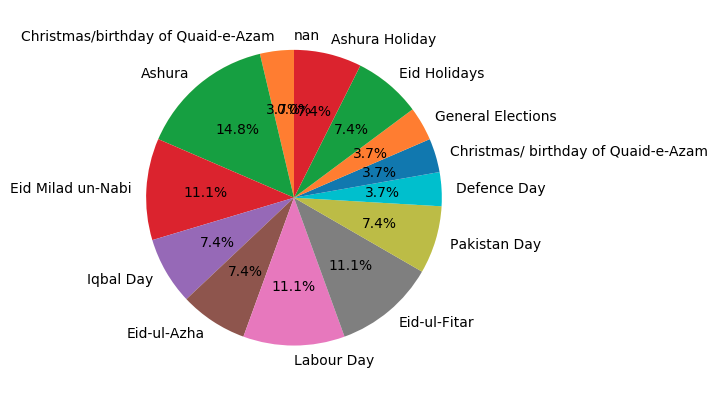
The biggest spikes can be seen in Muharram, which is considered a holy month for the Shia community, with a total of 53 attacks, and in Shaban, the month right before Ramadan. We can clearly see a big difference in the attacks depending on the Islamic Month.

# Does the type of Holiday affect the number of Attacks?

The most suicide attacks occurred on Weekends, which is expected as Weekend holidays are every week.

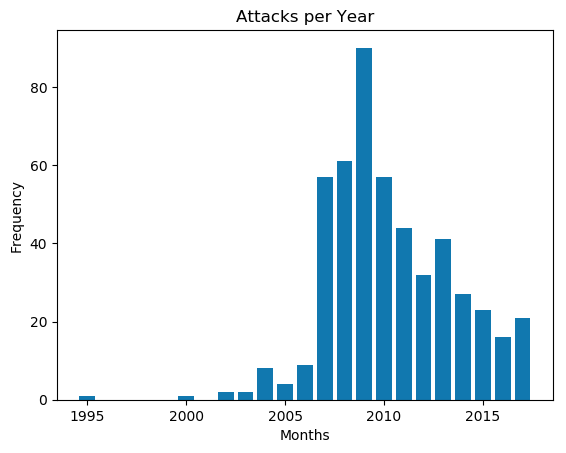


After removing ‘Weekend’, we can more clearly see the results,



The most attacks (14.8%) occurred on Ashura Holidays while the second-most attacks occured on Eid Milad-un-Nabi as well as Eid-ul-Fitar, showing a significant relationship between suicide bombings and religious holidays.

# Attacks Per Year

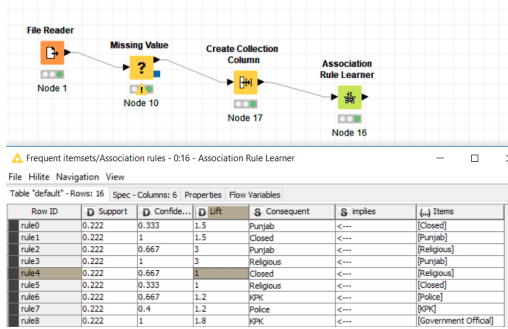


Overall, attacks have significantly increased over the years up till 2009, which was the peak year with over 80 suicide bombings, but since then, the trend has been mostly downwards.

# Dependency Between Province and Location of Attack

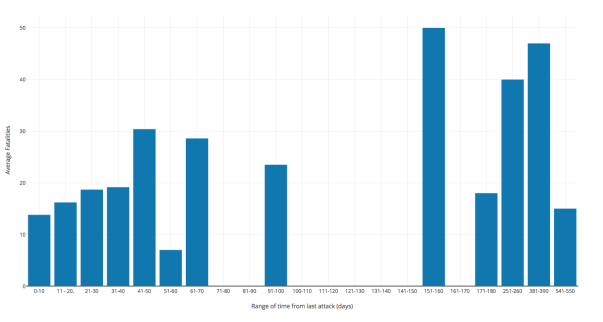
We tried to analyze an association between Province , Open/ Closed Space. We included Province, Open/Closed Space and Target Type. After exploring these using the Association Rule Learner, we observed that most of the blasts in Punjab had “Religion” as its target type and they were all in closed space.

Meanwhile,the majority of blasts that took place in KPK had Police or Government official as its target type.

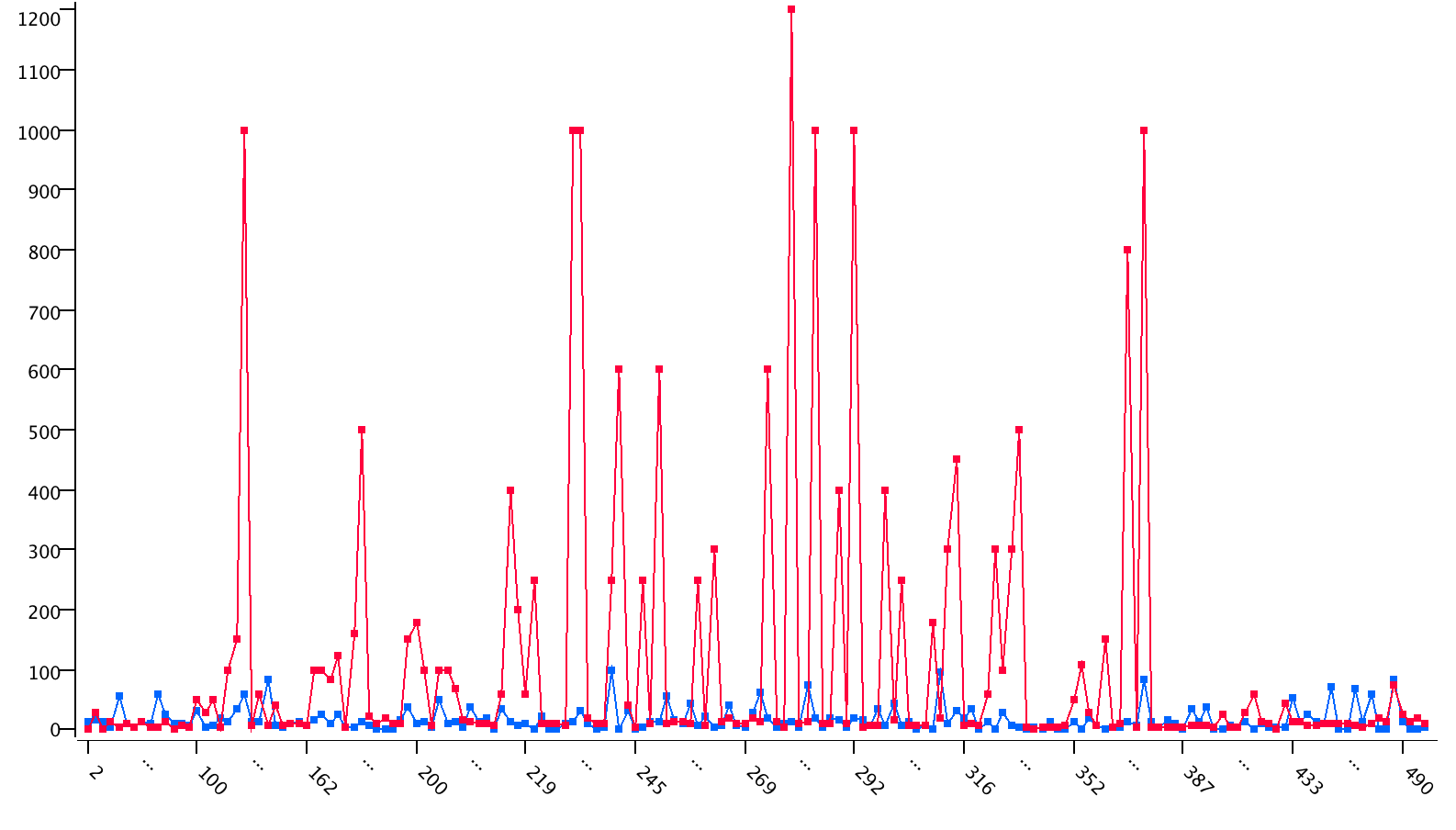


# Effect of delay between consecutive attacks

We also plotted the delay between consecutive attacks to see if the delay had any effect on the magnitude of the attack. The below chart shows how, the larger the delay between consecutive attacks, the more the number of average fatalities.

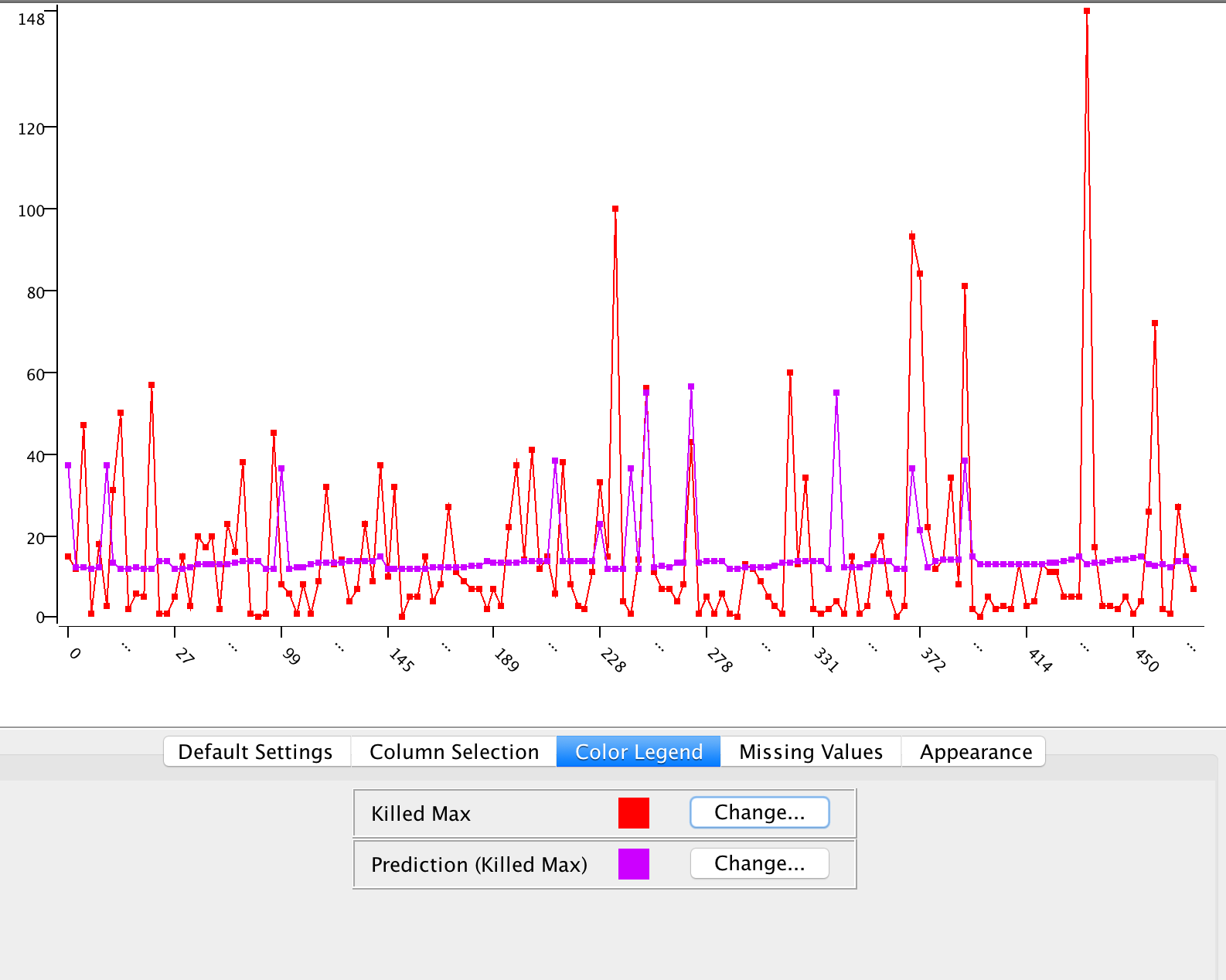


# Correlation between Explosive Weight and Killed Max

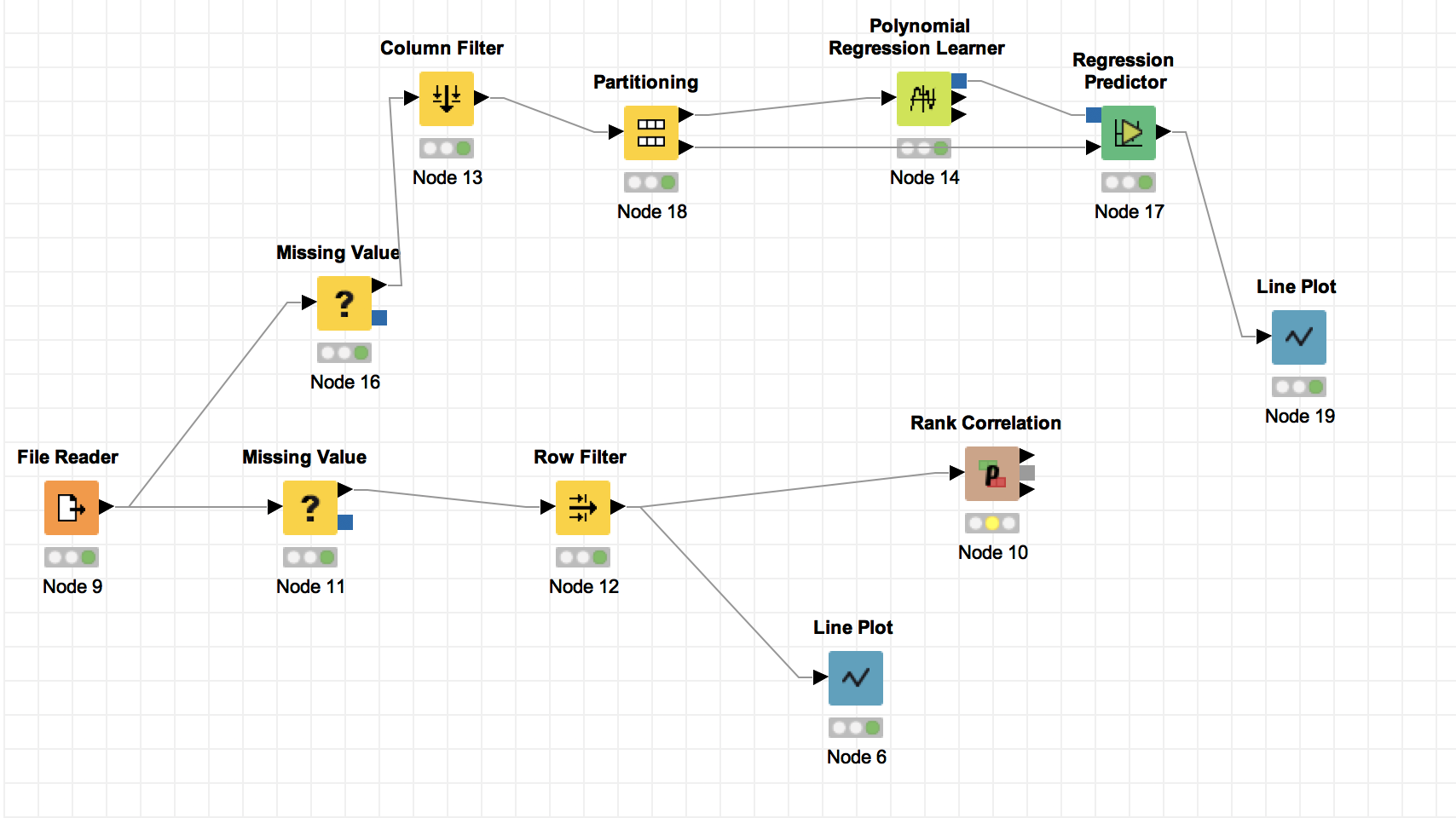


We can see a clear correlation between the both i.e. an increase in weights of explosive used resulted in more casualties

# Regression Model - Predict Max People Killed



We used the following attributes: Killed Max, No. of Suicide Blasts, Islamic Month, Weights to try to fit a regression model. We can see there is lots of variation between predicted and actual value. Hence we can conclude that there is not enough data to fit a regression model.



# Conclusion

The data we originally acquired was considerably inconsistent and dirty, with missing, duplicate, as well as unformatted values. After cleaning the data and using techniques learned in IDM, we managed to extract some very useful insights from the data. We observed how the North of Pakistan has faced the bulk of Suicide Bombings.

Some of the most surprising results include how the gap between consecutive bombings had a huge effect on the magnitude of the blasts that occured with the biggest gaps. Also, we noticed how religious holidays were the ones that attracted the most number of suicide bombings.