**KAUNAS UNIVERSITY OF TECHNOLOGY INFORMATICS FACULTY**

**INTRODUCTION TO ARTIFICIAL INTELLIGENCE**

**DATA ANALYSIS LAB WORK REPORT**

Report author: Orkun MANAP

Lecturer: doc.dr. Germanas Budnikas

Kaunas, 2021

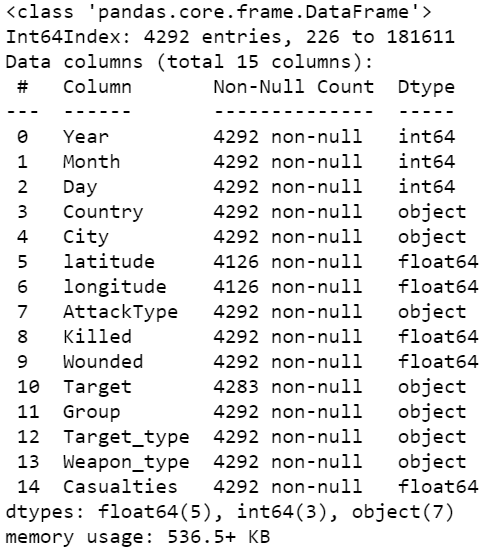
**1. Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor.**

**Selected Dataset:** Terrorism in Turkey

|  |  |
| --- | --- |
| **Link**: | https://www.kaggle.com/START-UMD/gtd |
| **Description**: | Statistics of actual terrorist attacks in Turkey |
| **Format**: | A data frame with 4292 observations on the following 15 variables. |

|  |  |
| --- | --- |
| Year  This field contains the year in which the incident occurred. | Month  This field contains the number of the month in which the incident occurred. |
| Day  This field contains the numeric day of the month on which the incident occurred. | Country  This field identifies the country where the incident occurred. |
| City  This field identifies the city where the incident occurred. | Latitude  The latitude of the city in which the event occurred |
| Longitude  The longitude of the city in which the event occurred. | Attack Type  The general method of attack and broad class of tactics used. |
| Killed  The number of total confirmed fatalities for the incident | Wounded  The number of total confirmed wounded for the incident |
| Target  The specific person, building, installation that was targeted and/or victimized | Weapon Type  General type of weapon used in the incident |
| Group  The name of the group that carried out the attack | Target Type  The general type of target/victim |
| Casualties  The number of sum of killed and wounded people. |  |

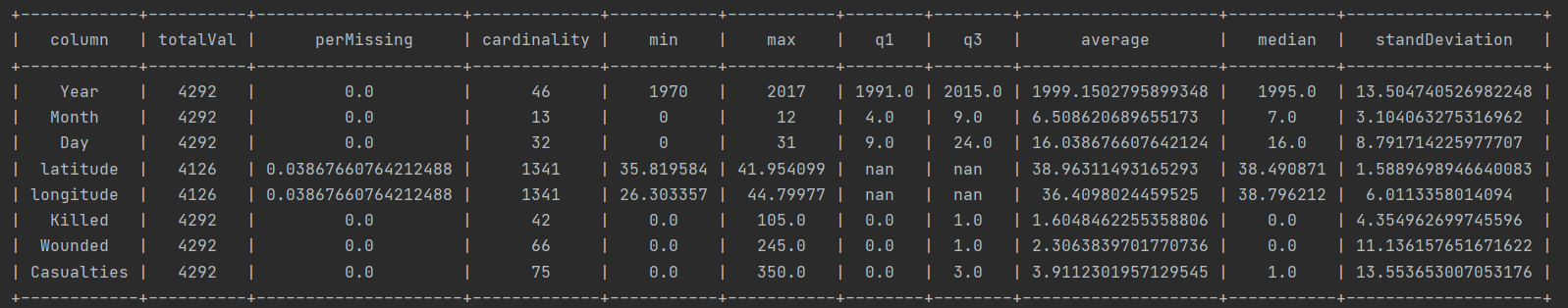
The columns have the following datatypes:



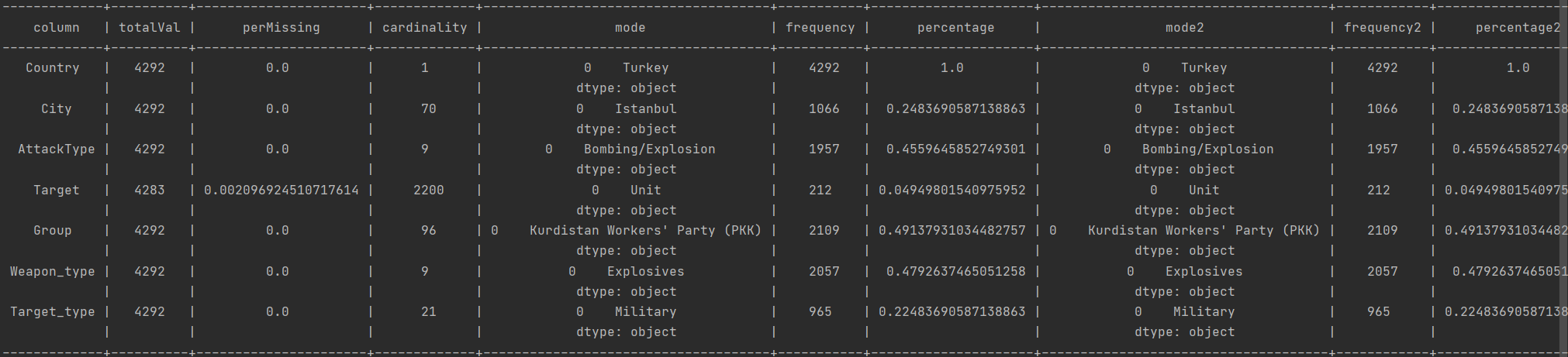
1. **For each numeric type attribute calculate:**

total number of values,

* + percentage of missing values,
  + cardinality,
  + minimum (min) and maximum (max) values,
  + 1st and 3rd quartiles,
  + average,
  + median,
  + Standard deviation.



1. **For each *category* type attribute calculate:** 
   * + total number of values,
     + percentage of missing values,
     + cardinality,
     + mode,
     + The frequency of the mode
     + Percentage value of the mode
     + Second mode value (mode 2),  Frequency value for Mode 2,  Percentage of Mode 2.



1. **Draw histograms of attributes. Provide descriptions of the distribution (eg, normal, exponential, etc.) and what conclusions can be drawn from it.**

|  |  |
| --- | --- |
| **Histogram** | **Description** |

|  |  |
| --- | --- |
| Year | Skewed left |
| Month | Bimodal distribution |
| Day | Uniform distribution |
| Country | Only one value  This attribute may not be used or used. It should be removed. |

|  |  |
| --- | --- |
| City | Right skewed |
| Latitude | Multimodal distribution however, there is a cut off at 42. |
| Longitude | Bimodal distribution however, there is a cut off at 45.0 |
| Attack Type | Skewed right |
| Killed | Skewed right |
| Wounded | Skewed right |
| Target | Multimodal distribution |
| Group | Multimodal distribution |
| Target Type | Multimodal distribution |
| Weapon Type | Skewed right |

1. **Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g., missing values for a categorical attribute based on the attribute estimate of the mode, extreme values being removed or corrected).**

There are 3 attributes that has missing values (Longitude, latitude, target). The highest missing value ratio is 0.0387. These values can be substitute with a value from a similar instance or can be ignored or can be treated all the missing values as equals. However, substituting is much more efficient. So, I’ve deleted rows that has missing values with terror.dropna().

There are no outliers that should be removed.

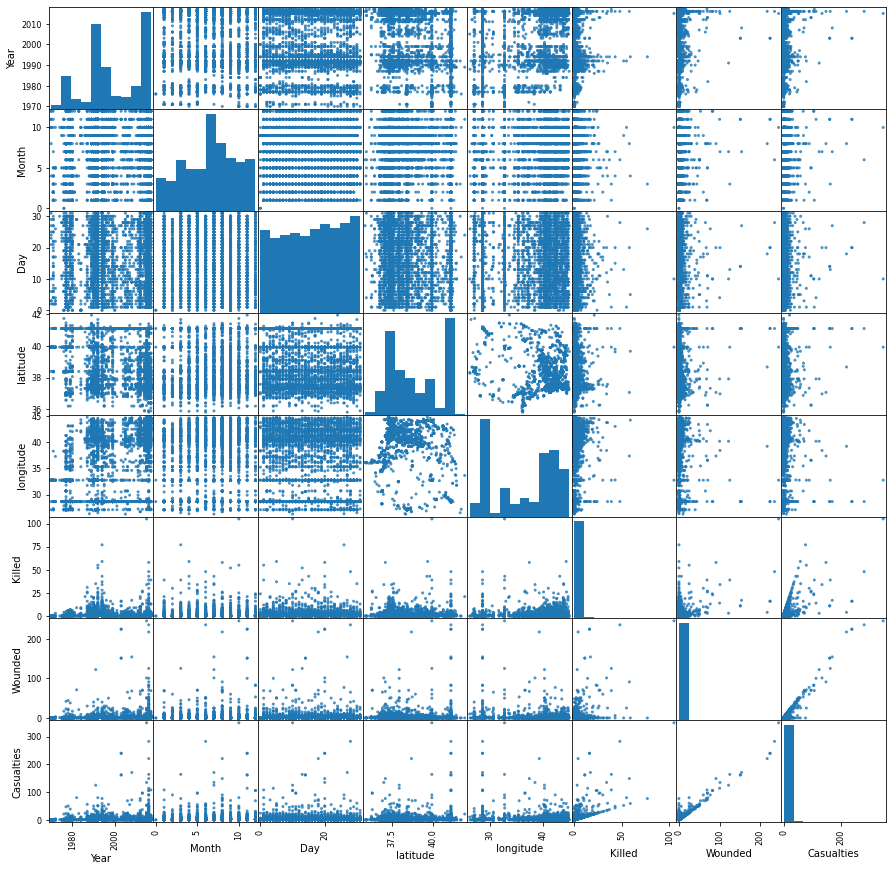
All following tasks are solved with this corrected dataset.

1. **Establish relationships between attributes using visualization techniques** 
   * **For numeric type attributes: Using a scatter plot type graph, provide multiple (2-3) examples with strong linear attribute dependency (direct or inverse correlation) and multiple examples with non-correlated (weakly correlated) attributes. Comment on results.**
   * **Provide an SPLOM diagram (Scatter Plot Matrix).**

|  |  |
| --- | --- |
| **Scatter Plot** | **Description** |
|  | No correlation:  Number of casualties increases while number of killed people increase. However, not all the time. |

|  |  |
| --- | --- |
|  | Strong Linear correlation:  Number of casualties increases while number of wounded people increase. Because of that, there is a strong linear correlation between casualties and wounded attribute. So, we can use only one of them, because they will do the similar thing while we are training the machine learning model. |
|  | No correlation between year and casualties. |
|  | No correlation between the latitude and longitude. |
|  | No correlation between the casualties and longitude. |

**SPLOM-Diagram:**



* + ***For categorical attributes:* Using the bar plot type diagram, give some (2-3) examples of attribute frequency and comment on the results.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Bar plots** | **Legend** | | **Comment** |
| Attack Type | Bombing  Armed Assault  Assassination  Facility/Infrastructure Attack  Hostage Taking  Unknown  Unarmed assault | | As can be seen from this graph, the most common terrorist attacks were bombings and armed attacks. The most preferred terror weapon of terrorists is bombing. Based on my own experience, they also used suicide bombing a lot. |
| City | Istanbul, Diyarbakır, Şırnak, Hakkari, Ankara… |  | Istanbul is Turkey's most important city. It has a population of almost 16 million. For this reason, it is understandable that terrorists prefer this city. Diyarbakir, Sirnak and Hakkari, are the largest city in Turkey's east. |
| Target | Military, Police, Private citizens/property, business, government… | | Army and police stations have been the most targeted places for terrorist attacks. |
| Weapon Type | Explosives, Firearms, Unknown, Incendiary, Chemical, Fake weapons, sabotage equipment, other. | | Explosives and firearms have been the most preferred weapons by terrorists. |

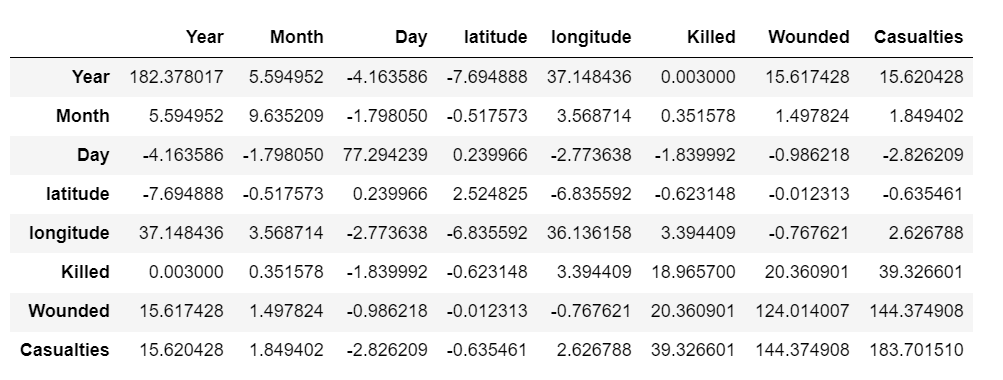
 **Provide some (2-3) examples of histograms and box plot diagrams depicting relationships between categorical and numeric type variables.**

|  |  |
| --- | --- |
| **Boxplot** | **Description** |
|  | Overlapping between, Casualties and Weapon Type attributes are big. Because of that, we can’t say there is any dependency between Casualties and Weapon Type. However, Explosives and Firearms have much casualties. |
|  | Overlapping between, Casualties and Target Type attributes are big. Because of that, we can’t say there is any dependency between Casualties and Weapon Type. |
|  | Overlapping between, Year and Target Type attributes are big. Because of that, we can’t say there is any dependency between Year and Weapon Type. |
|  | Overlapping between, Year and City attributes are big. Because of that, we can’t say there is any dependency between Year and City. |

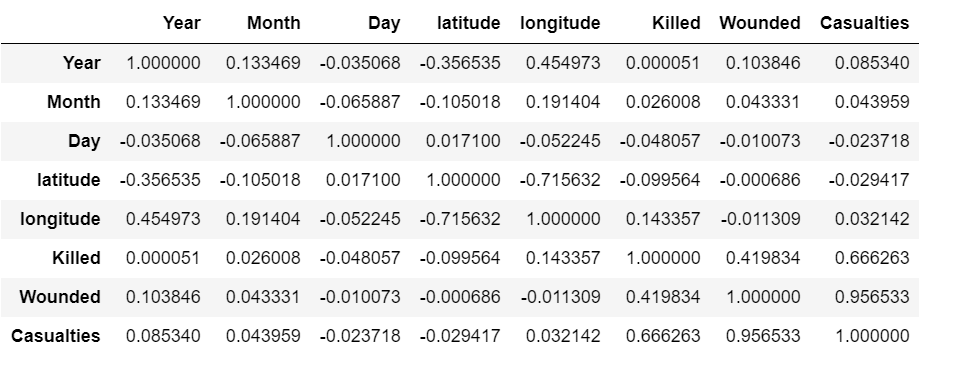
|  |  |
| --- | --- |
| **Histogram** | **Description** |
|  | Between 1986 and 2000, 1992 was the year with the most terrorist attacks. In 1992, the most armed terrorist attacks took place. Between 2000 and 2017, the most terrorist attacks occurred in 2016. Unlike 1992, the most common terrorist attack in 2016 was bombing. Likewise, the highest number of bombing attacks took place in 2015. There has been an increase in bombing attacks compared to previous years. It has been observed that the rate of robbery attacks has always remained the same. |
|  | Terrorist attack type bombing was the one with the least number of deaths. The terrorist attack with the most deaths was an armed terrorist attack. |
|  | Between 1970 and 2000, one of the most common target types, "Business", was found to be less preferred by terrorist organizations in the years after 2000. It was observed that terrorist attacks on the army increased in 2012 and the following years. The peak year of terrorist attacks on the army was 2016. |

1. **Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix. Comments on the results.**

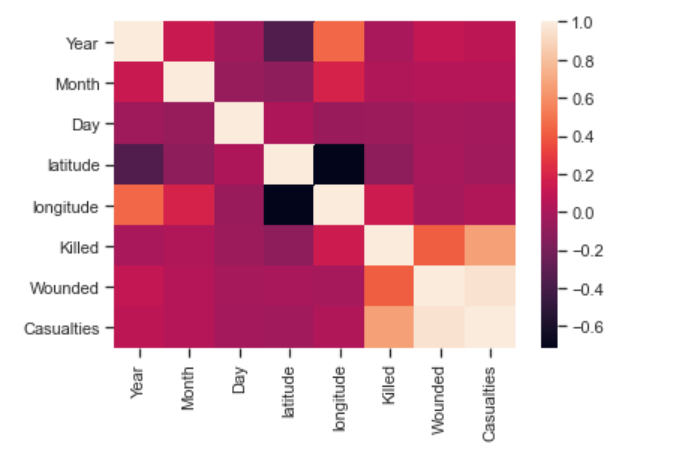
Covariance:



Correlation:

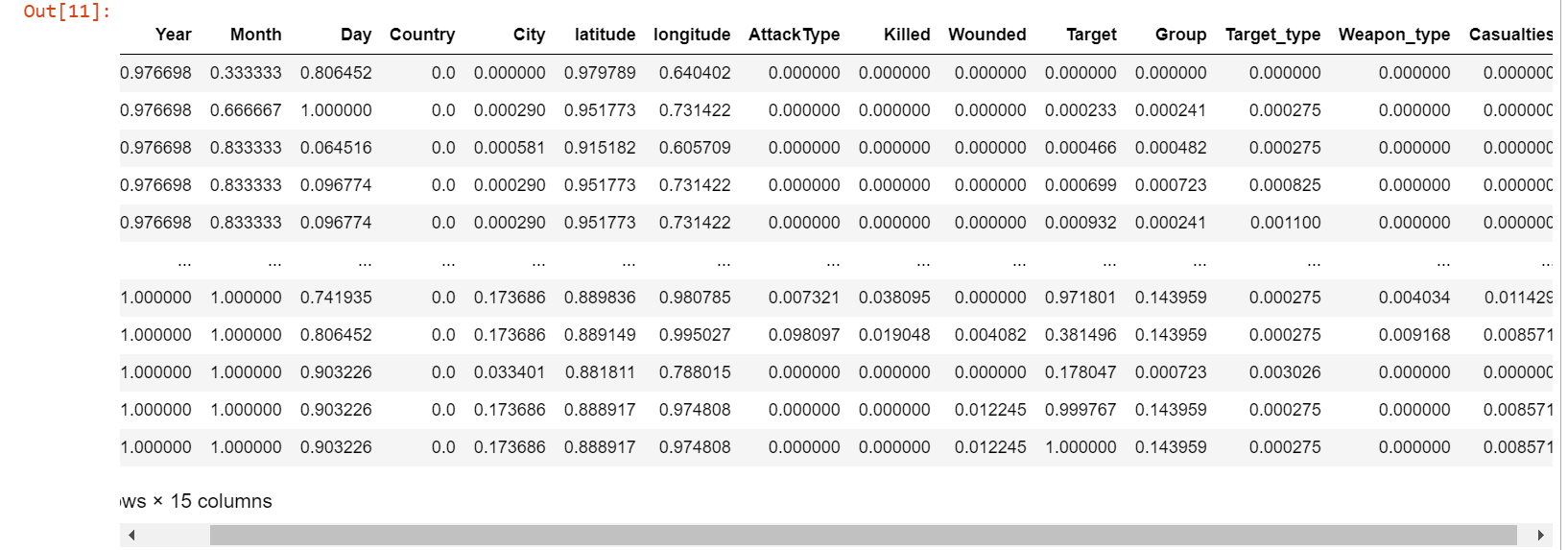


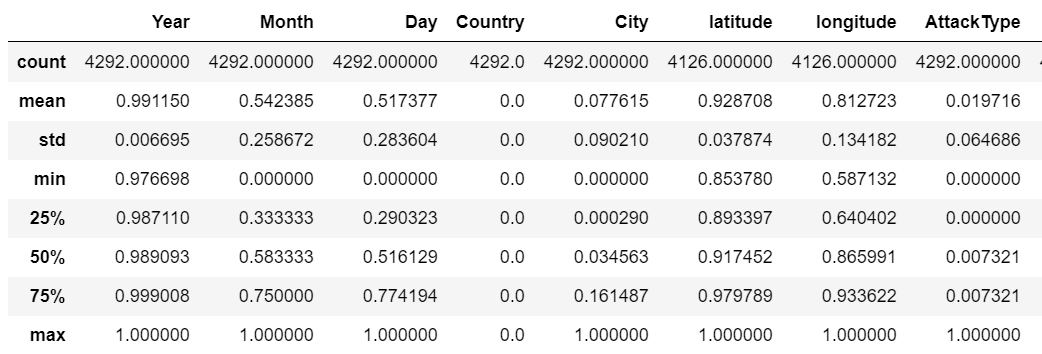
Correlation Matrix:



1. **Perform data normalization.**

Borders: 0-1

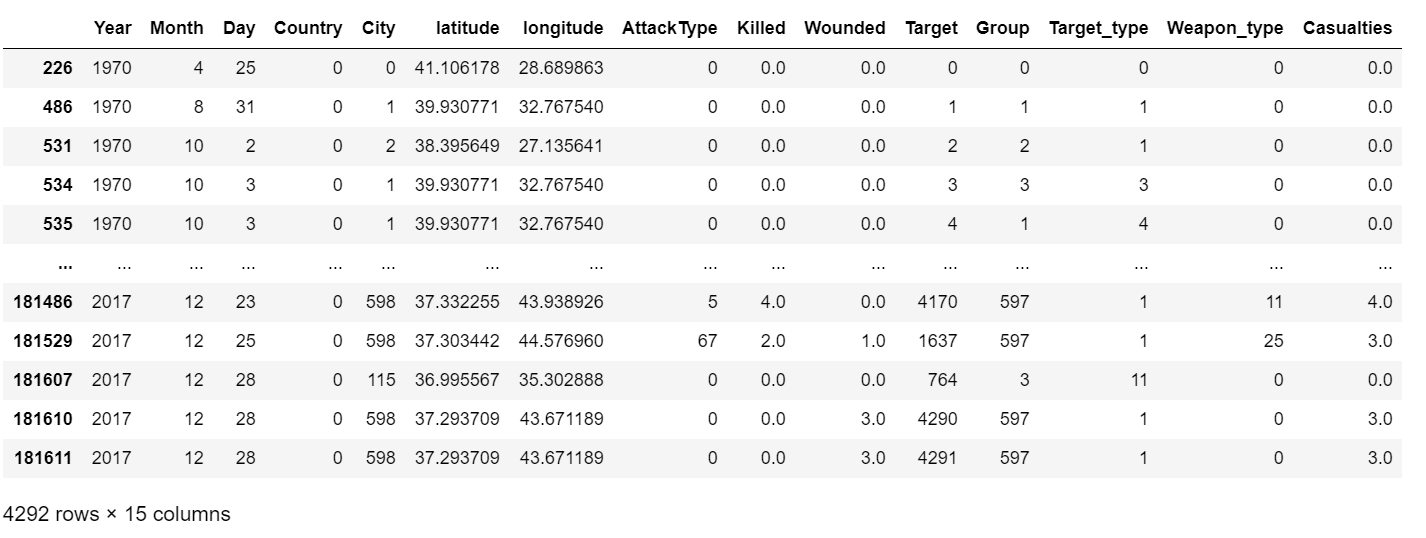




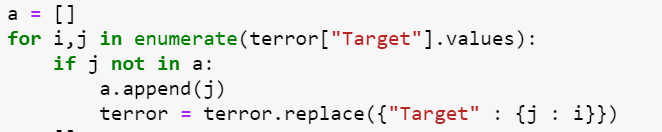


1. **Convert categorical variables to numeric type variables.**

**Before normalization:**



Example for implementation of converting categorical variable “Target” to numeric type:



Every categorical value has their own unique numeric value.