**Data Processing Protocol for CATS Data Factory Importation**

1. Import raw GDELT data.
   1. Microsoft Server Name: **uniictdatdb01.database.windows.net**
   2. Microsoft SQL Database: **uni-oia-cats-datastgdb**
2. Merge **mentions** data with **exports** data.

|  |
| --- |
| **SELECT**  e.GLOBALEVENTID,  EventTimeDate,  MentionTimeDate,  Confidence,  MentionDocTone,  EventCode,  EventRootCode,  QuadClass,  GoldsteinScale,  ActionGeo\_Type,  ActionGeo\_CountryCode  **FROM**  **`gdelt-bq.gdeltv2.events` e**  **JOIN**  **`gdelt-bq.gdeltv2.eventmentions` m**  **ON** m.GLOBALEVENTID = e.GLOBALEVENTID |

1. Drop all rows where there are nulls present in any of the following columns:

|  |
| --- |
| “GLOBALEVENTID",  "EventTimeDate",  "MentionTimeDate",  "ActionGeo\_CountryCode",  "Confidence",  "MentionDocTone",  "EventRootCode",  "QuadClass",  "GoldsteinScale" |

1. Select **MentionTimeDate** rows within 15 days of associated **EventTimeDate**.
2. Convert **EventTimeDate** from datetime to date.
3. Select rows where **Confidence** is >= 40%.
4. Replace the following integer columns with their string counterparts:
   1. QuadClass

|  |
| --- |
| 1: 'Verbal Cooperation',  2: 'Material Cooperation',  3: 'Verbal Conflict',  4: 'Material Conflict' |

* 1. CAMEO EventRootCode

|  |
| --- |
| 1: 'MAKE PUBLIC STATEMENT',  2: 'APPEAL',  3: 'EXPRESS INTENT TO COOPERATE',  4: 'CONSULT',  5: 'ENGAGE IN DIPLOMATIC COOPERATION',  6: 'ENGAGE IN MATERIAL COOPERATION',  7: 'PROVIDE AID',  8: 'YIELD',  9: 'INVESTIGATE',  10: 'DEMAND',  11: 'DISAPPROVE',  12: 'REJECT',  13: 'THREATEN',  14: 'PROTEST',  15: 'EXHIBIT MILITARY POSTURE',  16: 'REDUCE RELATIONS',  17: 'COERCE',  18: 'ASSAULT',  19: 'FIGHT',  20: 'ENGAGE IN UNCONVENTIONAL MASS VIOLENCE' |

* 1. Country Name Strings
     1. [FIPS 10-4 Country Codes](https://en.wikipedia.org/wiki/List_of_FIPS_country_codes)

**Data Alert System Protocol for CATS Data Factory Importation**

**Chart, box and whisker chart

Description automatically generated**

24 Month IQR

12 Month IQR

**Purpose**: Since the distribution of each variable, at the CAMEO-code and country-level, cannot be confidently characterized as the normal distribution, therefore, we instead apply alternative approaches for outlier detection. In place of identifying outliers that exhibit discrepancies beyond three standard deviations, we detect abnormalities exceeding 1.5 and 3 times the Interquartile Range. Although, this approach appears to be a brute-force statistical method being applied to a complex problem of anomaly detection in the event type data, it is also a highly effective method short of clustering machine learning models (LSTM being the current industry-preferred model).

1. Create output variable by grouping the preprocessed data by Country, EventDate, and EventCode:
   1. **EventReportValue**
      1. The proportion of daily articles related to a country, per Event Root
   2. **GoldsteinReportValue**
      1. The median daily GoldsteinScale value of all articles related to a country, per Event Root
   3. **ToneReportValue**
      1. The median daily MentionDocTone value of all articles related to a country, per Event Root
2. Calculate Rolling Medians for Variables
   1. **EventReportValue**
      1. 1 Day Window
      2. 60 Day Window
   2. **GoldsteinReportValue** 
      1. 3 Day Window
      2. 60 Day Window
   3. **ToneReportValue**
      1. 3 Day Window
      2. 60 Day Window
3. Create the Interquartile Range for Each Variable
   1. 1 Day and 3 Day
      1. 12 Months
   2. 60 Day
      1. 24 Months
4. Create the Outlier Detection on Each IQR.
   1. **EventReportValue**
      1. Quantile75 + 1.5\*IQR = Mild Abnormality
      2. Quantile75 + 3\*IQR = Extreme Abnormality
   2. **GoldsteinReportValue**
      1. Quantile25 - 1.5\*IQR = Mild Abnormality
      2. Quantile25 - 3\*IQR = Extreme Abnormality
   3. **ToneReportValue**
      1. Quantile25 - 1.5\*IQR = Mild Abnormality
      2. Quantile25 - 3\*IQR = Extreme Abnormality

**Rationality:** Since we are specifically interested in CAMEO code that are related to Conflict, it stands to reason that an **increase** in the proportion of articles related to CAMEO Conflict codes should be flagged as a discrepancy in the spike or trend alerts for a given country. Similarly, since both Goldstein scores and article tones become more negative in value as Conflict of country-destabilizing events occur, a **decrease** in the Goldstein and Tone values should also be flagged for the spike and alert detection system.

1. Output the above variables in to Microsoft SQL (or other) Datatable for PowerBI output