Cabildo, Adan L. | BSCS-SS212

What variables might be relevant to the decision?

What are the hypotheses about how to improve the data? And what variables are needed to test the hypotheses?

- The frequency of pirate attacks exhibits chronological patterns, such as seasonal variations or trends over the years.
  - Variables required:
    1. Date (Year and Month)
    2. Attack Type
    3. Count of Pirate Attacks
- Certain geographic regions are more prone to pirate attacks.
  - Variables required:
    1. Latitude
    2. Longitude
    3. Nearest Country
    4. Location Description
- Specific ship characteristics, such as type, speed, or size, are correlated with a higher likelihood of being targeted in pirate attacks.
  - Variables required:
    1. Vessel Type
    2. Vessel Status
    3. Attack Type

What data sources can inform those variables?

What are the possible sources of the necessary information?

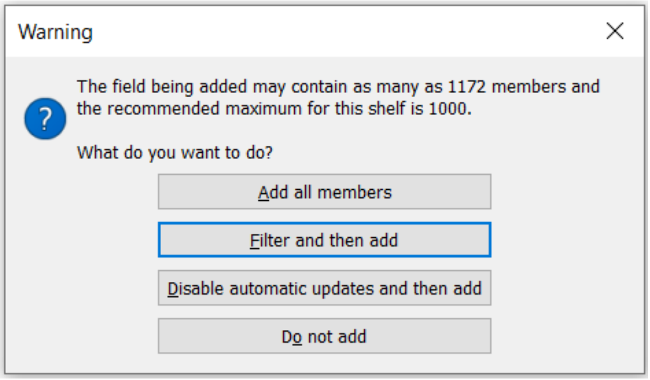
- International Maritime Bureau (**IMB**) which is under the International Chamber of Commerce – Commercial Crime Services (**ICC-CCS**)
  - Link: https://www.icc-ccs.org/

Are there any considerations that constrain which data sources can be used (timeliness, access, privacy, cost, standards, etc.)?

- Based on terms and conditions that ICC-CCS provides, where the **dataset** from Kaggle came from, the following can be considered:
  - The provided data is for general information and may change without notice.
  - The ICC-CCS cannot guarantee accuracy, completeness, and suitability of the data provided.
  - ICC-CCS provides live data of piracy (but this project's scope is for the years 1993 to 2020 only).

What does exploratory data analysis reveal about the available data?

When inserting Attack Description into tableau, a warning is prompted, proving transforming of data is required for this kind of variable.



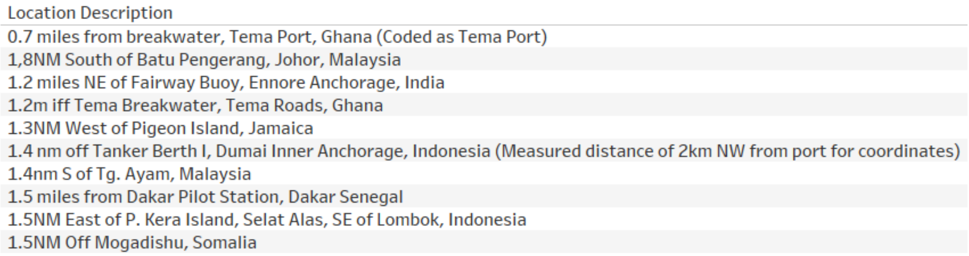
Warning

The field being added may contain as many as 1172 members and the recommended maximum for this shelf is 1000.

What do you want to do?

Figure 1. Warning Prompt for inserting Attack Description

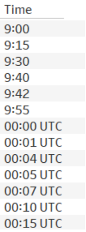
This transformation of data is also appropriate for Location Description because it contains a string data type that may be a problem for modelling and analysis.



Location Description
0.7 miles from breakwater, Tema Port, Ghana (Coded as Tema Port)
1,8NM South of Batu Pengerang, Johor, Malaysia
1.2 miles NE of Fairway Buoy, Ennore Anchorage, India
1.2m iff Tema Breakwater, Tema Roads, Ghana
1.3NM West of Pigeon Island, Jamaica
1.4 nm off Tanker Berth I, Dumai Inner Anchorage, Indonesia (Measured distance of 2km NW from port for coordinates)
1.4nm S of Tg. Ayam, Malaysia
1.5 miles from Dakar Pilot Station, Dakar Senegal
1.5NM East of P. Kera Island, Selat Alas, SE of Lombok, Indonesia
1.5NM Off Mogadishu, Somalia

Figure 2. Displaying 10 values of Location Description

Next, the time must be standardized into military time format.



Time
9:00
9:15
9:30
9:40
9:42
9:55
00:00 UTC
00:01 UTC
00:04 UTC
00:05 UTC
00:07 UTC
00:10 UTC
00:15 UTC

Figure 3. Displaying Time value differences

How will you transform the raw data into variables suitable for modeling?

What does the profile of each data source reveal about the quality of the available data?

- After analyzing the data and its source, which came from Kaggle that originates from ICC-CCS, the data follows a standardized format that can be easily used for machine learning and data analytics.
- The sources provide a csv file format for easier and compact dataset sharing.
- ICC-CCS also provides live or up-to-date data that can be used for future project improvement.

How should data from different sources be integrated? Is there a need to integrate?

- Integration is not required because this project has only one source which is Kaggle that originates from ICC-CCS. Kaggle uses the specific format from ICC-CCS.
- Should the project evolve into a bigger development, integration will be easier for Kaggle (which is the current source) and ICC-CCS (which the current source originates) uses the same format and contains csv file type.

What additional preparation steps are necessary on the integrated data to yield variables for analytic modeling?

- **Attack Description** values can be summarized and clustered by cross matching the values that contain 1 or more similar words, for easier analyzation and modelling.
- **Location Description** values can be summarized and clustered by cross matching the values that contain 1 or more similar words, for easier analyzation and modelling.
- **Not Applicable (NA) Attack Types** will be removed.
- **Time** will be converted to military time format.
- **Eez Country (Exclusive Economic Zone)** will be converted to its full country name.
- In conclusion, data that contains NA or missing values will be removed. Additionally, the values will be converted to standardized format such as, but not limited to, Military format for Time.